NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação
Master Program in Statistics and Information Management

Multivariate Data Analysis for Monitoring the Quality of the Commercialized Bottled Water in Bangladesh

K.M.Mostafa Anwar [M2013491]

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

MULTIVARIATE DATA ANALYSIS FOR MONITORING THE QUALITY OF THE COMMERCIALIZED BOTTLED WATER IN BANGLADESH

by

K.M.Mostafa Anwar [M2013491]

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Analysis and Information Management

Advisor : Prof.Dr.Paulo Jorge Mota Pinho Gomes

February 2018

DEDICATION

TO MY FATHER
WHO STARTED HIS ETERNAL JOURNEY TOWARDS THE HEAVEN
DURING THIS STUDY PERIOD

&

TO MY MOTHER

&

TO MY MOTHER AND FATHER-IN-LAW

&

TO MY AMAZING DAUGHTERS: TATINEE, TANUSHREE & TANVI

&

TO MY WIFE DR.MALA KHAN WHO REMAINS TO BE MY PERFECT FOIL IN ALL CRISES AND ADVERSITIES I HAVE BEEN FACING DURING MY STUDY STAYING AWAY FROM MY FAMILY

ACKNOWLEDGEMENTS

MY BELOVED WIFE AND PREMIER CHEMICAL METROLOGIST IN BANGLADESH DR.MALA KHAN WHO WAS SO KIND TO SUPPLY ALL THE VALIDATED DATA UPON CONDUCTING THE LABORATORY ANALYSIS AT HER LAB MEETING THE STRICT QUALITY CONTROL PROCEDURES, MY BRILLIANT FRIEND JOHN UTHRIZ WHO HAD HELEPED ME A LOT FOR PROGRAMING IN R USED IN THIS STUDY, MY MENTOR PROF.PAULO GOMES WHO SHOWED HIS TREMENDOUS AFFECTION TOWARDS ME IN SUPERVISING ME WITH HIS GREAT EXPERTISE IN THE SUBJECT, MY COURSE COORDINATOR PROF.ROBERTO HENRIQUES AND ALL MY COLLEAGUES AT NOVA IMS

ABSTRACT

Several multivariate statistical or chemometrics or pattern recognition techniques e.g. Principal Component Analysis, Factor Analysis, Hierarchical and Non-Hierarchical k-Mean Cluster Analysis have been applied to gain understanding about the quality of the packaged bottled drinking water in the market of Bangladesh. Twenty three (23) physico-chemical properties of total of 51 water samples have been investigated. The data set consists of 49 individuals from 11 Brands and 2 deionized ASTM TYPE-I water samples produced in the laboratory to be a technically pure water having Electrical Conductivity ~0.056 µS-cm-1. Descriptive statistics, analysis of variance, Non-Parametric Kruskal-Wallis tests have been conducted to detect statistical differences between the water types and different brands. Total of 23 attributes of water covering major ion contents: sodium, potassium, calcium, magnesium, iron, manganese, chloride, fluoride, sulphate, bicarbonate and nitrate and other features: pH, temperature, total dissolved solids, electrical conductivity, hardness, ammonium, nitrite, free cyanogen and chemical oxygen demand, total cation sum and total anion sum. Both the Principal Components Analysis and the Factor Analysis revealed that the differences between water individuals are best characterized by four Principal Components or Factors indicating material loadings, hardness or softness aesthetic acceptability and lightness/sutability for human consumption. Hierarchical and Non-Hierarchical k-means Cluster Analysis clearly identified the presence of four distinct clusters: A, B, C and D among the bottled water products in the market of Bangladesh. The profile features for each cluster have been defined as such the classification achieved to acquire improved and detailed understanding of the general properties of the products under study. We have observed that HCA using WARD algorithm provided us with more realistic classification solution in comparison with non-hierachical k-means as the Cluster members are truly reflecting their group pattern in line with their chemical compositions. HCA using WARD showed that BRAND05 and BRAND11 belonging to Cluster A products execssively loaded with materials and considered to be as hard waters. And BRAND09 and BRAND10 staying with DEIONIZEDWATER belonging to Cluster B are completely devoide of essential minerals as such seemed to be as ultra low mineral content type water or too soft in nature. The other folks BRAND03, BRAND04, BRAND06, BRAND07 and BRAND08 are also not having sufficient mineral contents so as to be very soft water indeed. Hence, waters belonging to Clusters A, B and C are not suitable for human consumption. Only two brands BRAND01 and BRAND02 staying in Cluster D appeared to be suitable for human consumption in every respect.The fact is the BRAND01 is produced by a foreign manufacturer. That means, all other local brands, except BRAND02 are essentially not having the appropriate quality to be drinking waters. From both PCA and FA these two brands BRAND01 and BRAND02 have been very well explained. These are the major outcomes of this study not immediately apparent from univariate approach or not appeared from the data set while looking through naked eyes. It is revealed that the multivariate data analytical techniques have potential to be useful complementary techniques to support the existing univariate practices for industrial quality assurance quality control, market surveillance, standardization process and or regulatory purposes and also seemed to be interesting to academic and scientific communities seeking advanced knowledge.

KEYWORDS

Chemometrics; Pattern Recognition; Principal Component Analysis; Factor Analysis; Cluster Analysis; Bottled Water Quality Monitoring; Market Surveillance.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

APHA	American Public Health Association
ASTM	American Society for Testing and Materials
AWWA	American Water Works Association
BDS	Bangladesh Standard
BQSP	Bangladesh Quality Support Programme
BSTI	Bangladesh Standards & Testing Institution
CA	Cluster Analysis
cGMP	current Good Manufacturing Practices
CODEX	FAO/CODEX Alimentarius Commission
CRM	Certified Reference Material
DNCRP	Directorate of National Consumers´ Rights Protection
DRiCM	Designated Reference Institute for Chemical Measurements
EC	European Commission
EU	European Union
FAO	United Nations Food & Agriculture Organization
GMP	Good Manufacturing Practice
HCA	Hierarchical Cluster Analysis
PCA	Principal Component Analysis
ISO	International Standards Organization, Geneva, Switzerland
QA/QC	Quality Assurance/Quality Control
SM	Standard methods for examination of water and wastewater
SPC	Statistical Process Control
TQM	Total Quality Management
UNIDO	United Nations Industrial Development Organization

USEPA United States Environment Protection Agency

USFDA United States Food and Drug Administration

WEF Water Environment Federation

WHO World Health Organization

1. INTRODUCTION

1.1. BACKGROUND

Water the most naturally abundant and the simplest molecule on the earth is not only essential for the existence of human beings but also critical for the existence of all flora and fauna on this planet (Cohn, 1999). It has been unequivocally demonstrated that water of good quality is crucial to sustainable socio-economic development (UNEP/WHO, 1996). Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection (WHO, 2008).

Like some other parts of the world Bangladesh is facing the reality that arsenic poisoning in groundwater now threatening millions (Smith, 2000), (Lewis, 1999) and (Tibbetts, 2000) of people using groundwater as their primary source of drinking water. In a separate study (Pedersen, 2003), (Khan, 2003), (Murshid, 2002) it has been evident that the history of drinking water quality management in Bangladesh is nothing but a systematic lack of monitoring and controlling quality of water thus having far reaching consequences on Bangladesh leading to a calamity like arsenic poisoning (Smith, 2000), (Lewis, 1999) and (Tibbetts, 2000).

In the midst of this crisis rapid and dramatic growth of the market of the bottled water popularly called “mineral water” has been seen from the late seventies or early eighties of the 20th Century (Khan, 2003), (Khan, 2008). Surprisingly, millions of liters of bottled water consumed daily by the people in Bangladesh without knowing almost nothing of the quality of the goods only relying on some declarations/specifications provided by the manufacturers on their product labels which are mostly (>53%) unreliable/ inconsistent (Khan, 2008). A more recent study (Rahman, 2017) also further confirmed this fact where the study states that “The data printed on the bottle labels are inconsistent and not informative enough and does not correspond to the real scenario of constituents in the packaged water.”.

It has primarily been appeared that Bangladesh Standards & Testing Institution (BSTI) having questionable, reportedly non-transparent and dysfunctional organizational system (UNIDO, 2005), (UNIDO, 2008) is issuing certification to the bottled water products prior to be marketed by the manufacturer under its Certification Mark scheme which is indeed regulatory in nature. Although Government of Bangladesh (GOB, 1997) defined standard for drinking water vide Bangladesh Gazzet Additional August 28, 1997, Tofcil-3, Rule-12: Allowable Limit of Drinking Water and Allowable Limit of Groundwater (GOB, 1997) and as a Certification Mark (CM) scheme implementing agency BSTI is authorized to enforce two technical regulations vide BDS 1240:1989 (BSTI, 1990) & BDS 1414: 2000 (BSTI, 2000) for bottled drinking water and natural mineral water respectively, in practice there is no credible and effective market surveillance system in operation (Khan, 2008). Neither BSTI nor the Directorate of National Consumers Right Protection (DNCRP) established as a regulatory authority under the Consumers Right Protection Act 2009 (GoB, 2009) is continuously monitoring and controlling the quality of these drinking water products with systematic, effective and appropriate methods, means and frequency of surveillance protocols to compare the quality criteria stated in BDS 1240 (BSTI, 1990) and or BDS 1414 (BSTI, 2000) or GoB drinking water regulations (GoB, 1997).

Assuming the above situation this author along with his other colleagues in Bangladesh conducted studies to accumulate, primarily, the general information on the quality and pertinence practices followed by the manufacturers in declaring specifications and other information on the labels (Khan, 2008) and secondly to determine the inorganic physico-chemical quality of the commercially available bottled water in Bangladesh (Khan, 2008). The quality and validity of the data declared on the product labels and the inorganic physico-chemical quality of the bottled water has been observed to be mostly (>53%) invalid (Khan, 2008) . The study (Khan, 2008) indeed applied univariate statistical techniques dictated in the prevailing industrial practices e.g. cGMP, (WHO, 2008), (USEPA, 2002), (CODEX, 1985) (CODEX, 2001a), (CODEX, 2001b) and (BSTI, 1990), (BSTI, 2000) and so on.

The univariate based and summarized study results has been documented (Khan, 2008) upon conducting a general survey as well as laboratory based various instrumental analyses on total of 23 physico-chemical properties (variables). The observation matrix comprising 51 samples or records from 11 individual Brands collected from the market of Bangladesh including 02 (two) laboratory reagent grade water samples namely DEIONIZED WATER01 (DIWa) and DEIONIZED WATER02 (DIWb) to be chemically pure and demineralized water as per ASTM Type-I. The details methods of analyses, quality control and quality assurance procedures, descriptive statistical treatment and data quality assessment and validation process have been explained in the literature (Khan, 2008). The laboratory analyses have been conducted as per the applicable international best practices and norms (APHA, 1995), (USEPA, 1996), (USEPA, 1997) to ensure the validity, reliability of the data.

Owners and producers of the primary data are this author and his colleague who conducted the laboratory analysis prior to publishing preliminary survey results (Khan, 2008). More technical resourecs and knowledge have been acquired from the stat-of-the-art designed apex national laboratory, namely, Designated Reference Institute for Chemical Measurements (DRiCM) within the Ministry of Science & Technology in Bangladesh. This access to the primary laboratory based data has opened the opportunity to this author for investigating these commercialized bottled water brands in Bangladeshi market applying multivariate techniques. Because the previous study (Khan, 2008) did not applied multivariate techniques to explore any underlying physico -chemical and the quality attributes for grouping, clustering or classification etc.

1.2. PROBLEM STATEMENT

By and large, as per the conventional national, regional, international water, environment and related technical regulations, standards, protocols and practices (BSTI, 1990), (BSTI, 2000), (CODEX, 1985), (CODEX, 2001a) (CODEX, 2001b), (CODEX, 2008), EC (1998), GOB (1997), (UNEP/WHO, 1996), (USEPA, 1996), (USEPA, 1997), (USEPA, 2002) and (WHO, 2008) the quality of water for drinking, agricultural and industrial purposes is being monitored through evaluating individual physico-chemical-biological parameters as well as determining the presence of organic-inorganic-microbiological contaminations. And these monitoring processes mostly utilize univariate statistical approach along with other empirical methods.

The water industries while practicing, say, QAQC within the framework of TQM or cGMP, using bare minimum univariate statistical approaches required by their respective standards decided by the industries themselves as well as the concerned regulatory enforcing agencies. Applying multivariate techniques are not in vogue, in general, with a few exceptions where Statistical Process Control (SPC), Lean Six-Sigma or similar approaches may be utilizing some multivariate techniques at a minimum level. Especially within the packaged/bottled drinking water industries, rigorous utilization of multivariate techniques has not been reported.

On the contrary state-of –the-art multivariate techniques, provide statistical methods for study the joint relationships of variables in data that contain intercorrelations and several variables can be considered simultaneously, interpretations can be made that are not possible with univariate statistics. From the thorough literature survey it has been evident that in the other field of knowledge, natural sciences and industrial QAQC the multivariate techniques have been applied widely which showed that multivariate data analysis may provide some different and better understanding about the system comprising a large number of variables and data sets intermingled among themselves and apparently seemed not be to understood trough univariate approaches or pattern are not immediately visualized from conventional univariate techniques. While univariate statistical analysis of a large amount of data seemed to be cumbersome and cause misunderstanding and error in the interpretation, multivariate statistical techniques are more robust and, thus, become more useful for data treatment and for identification of anomalous or other underlying patterns (Lourenço, 2010). To gain a further insight and interpretation of large data sets it needs to be rather invesrtigated and understood via applying more exploratory as well as state-of-the-art multivariate techniques.

Upon considering the above situation, this project investigated the applicability of various multivariate techniques in classifying and qualifying the commercially available bottled water brands in Bangladesh and further augmenting and improving the knowledge in this area to accumulate more scientific information with a potential to be appreciated by regulatory, standardization, consumers´ rights protection bodies or communities and academia. At this particular stage, the author made an assumption that in conjunction with the conventional univariate based Industrial QAQC practices these approaches could be considered to be an additional supplementary aid to understand further and or to “explore” the overall “state” or quality of the commercialed or marketed drinking waters with an ultimate aim to framing, building, testing new hypothesis or theory to be investigated at the further “confirmatory and or structural equation modeling” stage in future. This dissertation mostly covered the “exploratory analysis” to reach “unsupervised classification solution (s)”. Further investigation needs to be done via “confirmatory analysis” to reach any “structural equation(s)”.

It could be reasonably assumed that to enforce regulations, specifically, in monitoring and controlling the variables, or indicators, or attributes or pollutants in the water having potential health risks, univariate approach would be prevailing practices until it is convincingly proved through various rigorous investigations that there are some significant benefits and advantages could be attained from multivariate techniques.

Considering this particular context, the main research questions appeared before this study team are as below:

a. Is there any scope of applying multivariate statistical techniques to understand the physico-chemical quality attributes of bottled drinking water products available in the market?

b. Are these techniques able to provide plausible explanation of any physico-chemical phenomena of water under study?

c. Is there any possibility to define/identify any reduced number of latent, independent quality indicators or features synthesized from the physico-chemical properties of the bottled drinking water?

d. If yes, to what extent these new indicators or features are suitable to explain the drinking water quality?

e. Is it possible to develop, validate any model or classification mechanism to classify the bottled drinking water products with respect to their group profiles for the purposes of market surveillance?

f. What are the limitations and prospects of applying these models or classification mechanisms in water quality monitoring programme?

Considering the availability of time and access to resources as well as assuming the academic needs, during this research technically sufficient number of samples (51) or individual Brands have been investigated to address the limited number of research questions listed above. But to meet the scientific requirements it had been planned that at the very onset of the investigation a quite large and exhaustive number of variables at least 23 has been studied following the internationally accepted reliable, strict and validated laboratory analytical processes so that the chemistry and physics of the water under study are fully characterized and understood with respect to their physico-chemical nature. The data matrix has been constructed with utmost reliable, quality and valid physico-chemical analytical results as such the present multivariate exploratory as well as the future confirmatory analysis do not suffer from any shortfall in their data quality level.

1.3. AIMS & OBJECTIVES

General objective of this research is to augment the knowledge about the quality of the commercialized bottled water in Bangladesh from multivariate analytical point of view and exploring any underlying interactions among the physico-chemical properties of water which could not be discerned immediately from the existing conventional univariate statistical analysis.

To specific objectives are:

a. Determining the scope of applying multivariate techniques to understand the physico-chemical quality, suitability of bottled drinking water products available in Bangladeshi market.

b. Investigating the possibility of defining any reduced number of latent, independent quality indicators synthesized from the large number of physico-chemical properties of the bottled drinking water. Estimating the limits of applicability of these indicators and suitability for explaining the drinking water quality.

c. Determining and visualizing clusters, if any, among the brands and expressing as well as explaining them with respect to their group profiles.

1.4. RATIONALE

Understanding and monitoring the quality of the commercialized bottled drinking water are critical to protect the consumers' rights, to ensure the safety of the manufactured products, to enforce the regulatory regime (Cohn, 1999), (Khan, 2003), (Khan, 2008).

Following various industrial production processes manufactures are producing and placing their packaged/bottled water products in the market. Various declarations in relation to standards, certification, testing, inspections, compositions, quality are made in the product labels. It is imperative to know that to what extent these declarations are valid and reliable and in compliance with the applicable requirements and needs (Murshid, 2002a), Murshid, 2002b).

Considering the implications in public health (Smith, 2000) and consumers rights protection (Cohn, 1999) the above issues are more critical specially in the least developed country like Bangladesh where enforcement of regulatory and standards regime are not yet developed and efficient (Khan, 2003), (Khan, 2008).

Furthermore, it is also important for the manufacturers to know the quality features and the attributes of the commercialized products both from their own origins as well as from their competitors to be more efficient operators in the market place.

Public, regulatory, civil and consumers rights societies, academia and finally the consumers are interested to know the status of the quality and safety of the commercialized bottled drinking water available in the market.

Monitoring the quality of water is technically challenging as well as an ever evolving issue gaining more impetus due to significant advancement in the other fields of knowledge e.g. multivariate statistical techniques experienced tremendous growth due to increased computing capacities (Lebart, 2008), commendable improvement in laboratory instrumental techniques and increased awareness in consumers rights, public health and environment protection after globalization.

Besides, state-of-the-art multivariate techniques, underpinned by the easy access to advanced computing capacities, provides statistical methods for study of the joint relationships of variables in data that contain intercorrelations. As in these methods, several variables can be considered simultaneously, interpretations can be made that are not possible with univariate statistics and as such in the last more than five decades from early sixties applications became common in medicine, agriculture, geology, social sciences, environmental sciences, ecology and systematics and other disciplines (James, 1990). The opportunity for succinct summaries of large data sets, especially in the exploratory stages of an investigation, has contributed to an increasing interest in multivariate methods (James, 1990).

In the field of water quality assessment relatively recently in the last two decades multivariate techniques are being utilized and quite a large number of research published covering ground water (Belkhiri, 2010), (Chenini, 2009), (Kumar, 2010), (Mahmood, 2011), Silva (2008), (Singh, 2009), surface water (Ahmed, 2005), (Carlson, 2002), (Charkhabi, 2006), (Iscen, 2008), (Kaneene, 2007), (McNeil, 2005), (Mustapha, 2011), (Obeidat, 2011), (Swain, 2012), (Yusuf, 2013), spring water (Ragno, 2007), river water (Alam, 2010), (Adeogun, 2012), (Debels, 2005), (Ge, 2013), Kido, (Najafpour, 2008), (Samsudin, 2011), (Shrestha, 2007), (Shrestha, 2008), (Zhao, 2009), ocean water (Pati, 2012), (Saravi, 2011) of different parts of the world. These studies assessed the physico-chemical properties of the water with applications in hydrogeology, geo-chemistry, maritime research, environmental sciences mostly to monitor impact of chemicals, pollutants and waste loads due to industrial, agriculture and other anthropogenic practices and classify the waters based on their physico-chemical properties as well as to understand any spatial-temporal evolution.

Although a very few in numbers, some studies have been published where potable drinking waters either from tube wells (Hossain, 2013) or from public-municipal piped supply systems (Souza, 2005) and (Odagiu, 2011) or from natural springs (Šnuderl, 2007) have been assessed applying multivariate techniques.

Along with the descriptive statistical methods invariably all of the above mentioned studies applied various multivariate techniques, namely, Principal Component Analysis (PCA), Principal Factorial Analysis, Multiple Correspondence Analysis, Canonical Correspondence Analysis, Partial Least Square and Discriminant Analysis, Cluster Analysis etc. also to observe spatial-temporal pattern of the water under study.

But it is worth mentioning that in comparison with the above mentioned studies only a very few (Ghrefat, 2013), (Inam, 2010), (Lourenço, 2010) and (Van, 2012) investigations has been recorded so far which assessed the quality of the commercialized bottled drinking water through applying multivariate data techniques. From the brief literature review, detailed in the subsequent Section 2, it has also been evident that these studies are also not free from limitations.

But these studies showed the clear potential of applying these multivariate techniques in monitoring the quality of the bottled water in the markets. The studies have sufficiently raised impression that there is an opportunity to investigate more in this area to further amplify the applicability of these techniques to cover different manufacturing, market, regulatory and or standards and economic regimes.

Moreover, the pure natural water i.e. demineralized water has not yet been studied so far applying these multivariate techniques to understand its natural physic-chemical behavior. The DEIONIZED WATER is indeed a known pure water, theoretically, not having any mineral, anion, cation or dissolved solid, suspended solid or particulate materials. This version of water is produced through applying a very sophisticated technology removing all anion, cations, organics, microbes and particulate materials. The specifications defined in the ASTM TYPE-I standard clearly stated that Electrical Conductivity (EC) and Electrical Resistivity of this deionized water, called ASTM TYPE-I, must be around 0.056 uS/cm or 18.3 Mohm-cm respectively. It is expected that this very individual DEIONIZED WATER would show a very extreme behavior in terms of physics and chemistry in comparison to other water samples under study.

The proposed study would be investigating the scopes of application of multivariate techniques for developing, optimizing model for understanding, qualifying the commercialized bottled drinking water to be fit for the above mentioned areas of application. This study may pave the way of developing some classification for controlling and monitoring the bottled water market. This study would help improving the existing body of the knowledge specifically in application of multivariate techniques in drinking water quality monitoring.

Figure 1.1 – Bottled Drinking Water in Bangladesh

1.5. SCOPE

This study is aimed to:

1. Collect, collate and compile the data for various physico-chemical properties of different brands of bottled water from the market of Bangladesh which have been tested in a

national laboratory using valid and reliable techniques and ensuring appropriate QAQC measures as the applicable international norms and standards.

2. verify the applicability of multivariate techniques to understand the physico-chemistry of bottled drinking waters in the market of Bangladesh,

3. identify and or define some reduced number of latent, independent indicators synthesized from the original physico-chemical variables and verifying the capacity of interpretation of these new indicators applying PCA and related techniques,

4. determine and visualize clusters, if any, through Cluster Analysis (CA) based on the similarity behavior of the brands,

5. produce a dissertation and communicate the scientific results to share the knowledge among the peers.

2. LITERATURE REVIEW

As mentioned above a very few studies (Ghrefat, 2013), (Inam, 2010), (Lourenço, 2010) and (Van, 2012) are published in relation to application of multivariate data technique to understand bottled/packaged drinking water.

Ghrefat (2013), reported the study of 54 brands of bottled drinking waters in Saudi Arabia where eight selected major chemical ion variables : calcium, magnesium, sodium, potassium, chloride, sulfate, bicarbonate and nitrate were examined by correlation analysis, principal component analysis (PCA) and hierarchical cluster analysis. Hierarchical cluster analysis classified the brands into different groups and the products have a diverse character reflected by their chemical compositions and are dominated by Na-Ca-HCO3-Cl type water. Total hardness values identified as the influential parameters which dictated the classification of the brands from soft to moderately hard waters. The study reported that the constituents lie within the acceptable limits established by the national – international standards and guidelines. The main criticism about this report is that the investigator relied solely on the data declared on the product labels declared by the manufacturers. The investigator himself did not tested the water in any reference or credible laboratory. There is no indication about the quality: reliability, validity and chemical metrological traceability of the data he used. No report has been made about the number/type of the variables whether they are sufficiently representing the chemistry of the nature of the water under investigation or not.

Inam (2010) studied 20 trace heavy metals in 165 ground water and 8 commercial bottled brands in Akwa Ibom State, Nigeria. In addition to some major cations e.g. sodium, potassium, iron, manganese, ammonium it is essentially required that relevant major anions i.e. sulfate, chloride, bi-carbonate alkalinity, Floride, Nitrate, Nitrite including some other aggregate properties like hardness, turbidity, total dissolved solids Electrical conductivity must be tested to understand the main quality features of natural water as these are the main constituents which contribute to build up the total balanced ion sums. But without analyzing these anions the author wrongly claimed that the study has been conducted to ascertain the quality and suitability of the water for drinking. With this inherent major methodical flaw, based on only 20 trace metals the author classified the ground water and bottled drinking waters applying Correlation Analysis, PCA and Hierarchical Cluster Analysis by Ward method which is having obvious bias.

The study by (Van Hulle, 2012) Applied multivariate statistical methodologies to characterize the commercialized bottled water and tap water in the flemish market. In this study the physico-chemical composition reported on the label of 49 bottled still waters, 22 bottled sparkling waters and 13 tap waters were used to carry out a characterization study by means of multivariate techniques principal components analysis (PCA) and discriminant analysis (DA). A one-way analysis of variance (ANOVA) test (with known standard deviation) confirmed the difference among different water types and water brands. Principal components analysis revealed that the differences between water types are best characterized by components that indicate saltiness, hardness and pH. The component pH allowed discriminating between sparkling water and non-sparkling water. It was not possible to divide the different water types based on saltiness or hardness, but it could be demonstrated that different types of water exist (low mineral, high mineral and mineral). The main criticism about this

study is it has been conducted on the basis of the data declared on the product labels only, not from any laboratory based study by the authors themselves. It is not known to what extent these data declared on the labels are valid, reliable and chemical metrological traceable. It is also not investigated whether the number of variables are exhaustive or not and whether they are sufficiently explaining the physico-chemical properties or quality attributes of the water under study or not. The answers to these questions may be lying behind the fact that why the study could not divide or differentiate or classify the individuals reasonably based on the physico-chemical data used in this study (Van, 2012).

But the study by Lourenço et.al. 2010 came up with a very comprehensive and detailed investigation reports on 33 different types of brands bottled waters consisting of 18 natural mineral waters and 15 bottled water from springs at geographically different locations in Portugal. Same brand some time marketed carbonated and non-carbonated waters. As such 10 physico-chemical parameters of total of 39 bottled ware have been tested in 2009 by the Laboratory of Geochemistry for Resource Management and three new synthetic dimensions have been identified upon conducting PCA and efficiently able to distinguish and classify the bottled along three main Principal axes. The first Principal axis explains almost 50% of the total variance, which is a typical mineralisation axis. In particular, $HCO3^{-}$, Na^{+}, Ca^{2+}, Mg^{2+} and K^{+}, parameters with the highest loadings contributed significantly to give a plausible interpretation. The second axis covering 23% of the total variance, discriminates sulphate-chloride type waters. Finally, the 3rd principal component axis denoted a pollution index derived mainly by agricultural activities and has been confirmed by the high PC loading on NO3. In this study PCA been successfully applied to identify the main geotectonic interrelationships among physicochemical parameters and contributing to a new typology of bottled waters, based on their hydrochemical characteristics and geological occurrence.The study demonstrated that the first three Principal components are classifying the bottled (spring and natural mineral) water with respect to their hydro-geochemical properties. This study again depicted the potential of applying multivariate techniques in classifying the bottled water. But this study did not further explored other techniques e.g. Cluster Analysis or Discriminant Analysis and so on. Drawing inference about the groupings only based on PCA may sometime lead to a wrong interpretation. Care must be taken before drawing such inference, especially consideration and treatment must be done if there is any effect from outliers.

3. METHODOLOGY

3.1. THEORETICAL FRAMEWORK

3.1.1. Introduction

The codes for each brands from different manufactures have been assigned (Table 4.1). The codes for variables: physico-chemical properties also have been assigned. The units or the scales of measurements are also tabulated (Table 4.2) as they are different in scale, therefore, throughout this investigation we would be applying the multivariate techniques mostly on centered or standardized data matrix (Annexure 2) . This is indeed a very important assumption to be kept in mind from now on.

Initial Data Matrix X in Excel Speed Sheet (Annexure 1) has been constructed by further reducing and summarizing the original raw data. The Initial Data Matrix X has been tabulated in Excel Speed Sheet consisting of 23 variables for total of 49 individuals from 11 Brands plus two individuals DIW and DIWb from the controlled laboratory reagent grade water, namely, DEIONIZED WATER. Number of variables sufficient to explain the water chemistry are retained. For each variables, the normality test to be done to verify whether the data are from same or different population. Standardized or Centered Data Matrix X_CENTERED consists of the finally retained variables for total of 51 observations would be generated .

To understand the general trend of the variables the descriptive statistical analysis has been done and summary has be tabulated (Table 4.4). The mean, standard deviation, minimum and maximum values of each variable has been recorded. For each variable, non-parametric normality test (for 95% Confidence Interval) has been done. As an example, when the variable total dissolved iron (Fe) is considered, it is observed that this variable is behaving normally as per the Anderson-Darling Normality test where the Null Hypothesis (Ho: The Fe data came from a normal distribution) has NOT been rejected at 5% level (p-value:0.259). Histogram and Box Plot have been constructed to understand each parameter from univariate point of view as well.

Correlation matrix has been constructed to see the correlations among the variables. Summary of the descriptive statistics and correlation analysis through looking at the correlograme has been investigagted to see any underlying pattern to move forward for further multivariate data analysis.

3.1.2. Principal Component Analysis

It is possible to describe the pattern of relationships among the objects (individuals, sampling units) by reduction of a matrix of distances or similarities among the attributes or among the objects to one or a few dimensions or by cluster analysis (classification of the objects into hierarchical categories on the basis of a matrix of inter-object similarities). PCA reduces a large number of variables (e.g. measured physical properties, chemical properties, anions and cations) to a smaller number uncorreleted variables called Principal components. The principal components analysis (PCA), allowing to determine which factors (group of variables) account for the numerical variation of the clusters. Also, the definition of PCs helps to extract related variables, giving more information than

single indicators or variables and to infer the processes that control water chemistry. In most of the cases, PCA is applied to the linear correlation matrix to be constructed prior to running this analysis. As stated earlier the measurement scales or units of the variables are different hence instead of using the covariance matrix, in this study the correlation matrix has been used to extract the eigen values to construct the corresponding eigen vectors which generates Principal Axes (PAs) and the Principal Components. The MINITAB, R, SAS and similar statistical packages having procedure of PCA analysis essentially utilized the standardized data or use correlation matrix.

3.1.3. Factor analysis

FA is similar to principal components analysis where it uses the similarity properties intrinsic to the matrix under study where usually eigen values as well as the corresponding eigen vectors are generated from the correlation pattern underlying the attributes. It emphasizes the analysis of joint relationships among the attributes or variables. Canonical correlations among the attributes are evaluated to explore whether there are any latent factors existing. These Factors, surely reduced in number in comparision with that of original variables, could be assumed to be newer reduced dimensions or Axes. In case of FA, in brief, the joint relationship among the attributes are studied to project them towards a few number of axes, called Factors, for further augmenting the information about underlying interactions among these variables. The Factor Analysis process uses the correlation between variables in order to find the latent factors within them. While running the Factor Analysis in this study we used the software package SAS which essentially extracted the eigen values as well as the corresponding eigen vectors in constructing the Principal Factors from the Correlation Matrix. Hence it was not unlikely that the same results and outputs were obtaibed from both the PCA and the Factor Analysis.

It may be noted that success of using the Factor Analysis technique depends on the correlation structure presents in the input data. It was required to be confirmed that significant correlations among the variables were existing, otherwise the Factor Analysis might not provide additional useful information. This analysis involved several steps. The first was to analyze the correlation structure of the input data set. It was not unlikely that there were significant correlations existing among the majority of the bottled water quality variables. In the next step was to chose the method of extraction of eigen values. It is already mentioned above that our original input matrix contained data from the measurements in different scales hence instead of using the covariance matrix, the correlation matrix was used to extract the eigen values as well as to construct corresponding eigen vectors during this analysis. The third step was to take the decision on how many number of factors to be extracted and or retained for further interpretation. Interpretation of the factors to be made based on factor loadings and essentially this process was similar to that of Principal Component Analysis.

3.1.4. Cluster Analysis

With cluster analysis (CA), objects are placed in groups according to a similarity measure and then a grouping algorithm. The reduction in the data comes from forming g groups (g less than n) out of n objects.

After the Principal Component Analysis and confirmation through the factor analysis, it could be confirmed that there are example, four principal components or factors sufficient as the latent dimensions to explain the bottled water products under study.

Applying the similarity criteria using both factors and original variables the cluster analysis to be conducted to see whether there is any grouping exists among the bottled waters and or Brands under study. It would be investigated and visualized to understand the fact that which brand or product belongs to which group and what are the average and or overall behaviors of these groups or clusters.

The application of cluster analysis involves two main methods, either hierarchical or non-hierarchical. The methodology to be used for clustering based on factors and the original variables.

At the beginning a hierarchical procedure to be run to define the number of clusters to be extracted. Since in these unsupervised learning procedures the number of clusters depends on the data it is not necessary to define a priori how many clusters to be generated. The ultimate classification solution based on hierarchical procedures depends on the distance measurement and the aggregation algorithm used.

In particular, in this study the methods like Average, Centroid and Ward's to be used and the results would be verified to assess their suitability for further interpretation and applications.

The number of clusters decided thus in the previous steps would be further used prior to running the final non-hierarchical k-means algorithm in confirming the final clusters. Moreover, different distances would be used e.g. Euclidean distance, squared Euclidean distance etc.

All of these approaches may provide similar results or different results. But the final classification solution to be selected based on the performance of each approach i.e. based on the analysis of the R-square, SSE (sum of squared error (SSE) for a number of cluster solutions) and dendogram.

Then, the best combination of hierarchical procedures, which was in fact WARD minimum variance technique, to be used to generate the initial seeds of the non-hierarchical algorithm – k-means. It has been seen that WARD provided the best results. The number of factor or cluster would be finally retained from the WARD output. Following the generation of the clusters, classification among the individuals to be done based on a ''profiling analysis'' and creating profile plots both using original data and the centered data to have a better understanding and visualization. The general statistical properties of each cluster would be tabulated to understand their relative positions.

In brief, we may summarized that HCA unsupervised pattern detection method partitioned all cases into smaller groups or clusters of relatively similar cases that were dissimilar to the other groups. Squared Euclidean distances measures were chosen to measure similarity/dissimilarity among the variables while the Ward's linkage method was chosen to link initial clusters resulting from the initial clustering steps. The combined use of squared Euclidean distances as a similarity/dissimilarity measure and the Ward'smethod as a linkage algorithm was observed to produce very reliable clusters in HCA.

3.2. METHODOLOGICAL PROCEDURES

- The data had been collected from an appropriate advanced laboratory based analytical processes following internationally acceptable and application standards, norms, methods, practices and maintaining the appropriate quality assurance - quality control (QAQC) protocols (APHA, 1998), (APHA, 2012) to ensure the reliability, validity of the data. Following these applicable international guidance appropriate protocols and Standard Operating Procedures (SOPs) would be followed at every stage of laboratory processes from bottled water sampling, pretreatment, preservation, transfer to laboratories to sample coding, preparing test aliquots, conducting analysis, assessing-validating the quality of the test results and analytical methods and producing laboratory reports.

- The data matrices is constructed by recording 23 number of inorganic physico-chemical properties. Variables selected are exhaustive in number to explain the individual samples or objects (in this case the Brands) in full with respect to their chemistry, physics as well as quality features. Including two laboratory produced deionized water DIW, total of 51 individuals from the maximum number 11 brands available in the market was collected and analyzed to meet the sufficient conditions. Of course, while progressing on the way to conduct this study we have gone through the process of avoiding redundant variables and finally remaining with variables up to 18 excluding five variables, namely, TEMP-EC, NO2, SO4, FreeCN and COD as such not being used for PCA, FA and CA processes.

- As we have generated the data in the laboratory by our own there was no missing data.

- At the very onset of this study, basic descriptive statistical techniques have been applied to gain understanding on the overall data as well as to "have a general feel about the data" through descriptive, exploratory statistical analysis. This is particularly helpful to decide the pathway for the further advanced data analytical processes, to frame the hypothesis, to give a vision about the potential number of groups or clusters as well their probable physico-chemical nature.

- Investigation is conducted on the physico-chemical properties through constructing and understanding similarity-dissimilarity matrices, verifying correlations and covariances among the variables considering various brands of bottled water products.

- Understanding the scope of expressing the individual brands with respect to some reduced number of synthetic, independent latent dimensions or factors. Various multivariate techniques e.g. Principal Component Analysis (PCA); and then the Cluster

methodologies are applied to explore the quality and attributes of the bottled water so as to understand the overall physico-chemical phenomena and quality features.

- Investigation is conducted to check whether the data set for each variable are normally distributed or not. Through assessing the joint-probability distribution functions etc., multi-normal distribution properties have been verified applying hypothesis testing i.e. Kruskal-Wallis Test for Variable versus Brand.

- Limitations and scope of further improvement and research has been discussed on applying the multivariate techniques not only for monitoring the quality of the commercialized bottled drinking water in the market but also for industrial quality control quality assurance process during the production in the bottled water manufacturing plants. Further research opportunity has been checked for applying the techniques to discern the potable water matrices in terms of varying manufacturing, geographical, natural origins.

4. RESULTS AND DISCUSSION

4.1. ORIGIN OF DATA & VARIABLES

A team of researchers in Bangladesh produced and published a laboratory based study results (Khan, M., Anwar, K.M.M., Chowdhury, HQ, Bangladesh Academy of Sciences, 2008) upon conducting a general survey as well as running laboratory based various instrumental analyses on total of 23 physico-chemical properties (variables) for 51 samples from the total of 12 individuals consisting of 11 Brands collected from different departmental and grossary shops at different zones: Dhanmondi, Banani and Uttara of Dhaka City during a one-day sample collection campaign on 12 June 2001 and 01 controlled laboratory reagent grade water called DEIONIZED WATER. Total 51X23 = 1173 validated data structured in the Initial Data Matrix (Annexure 1: Table A1-1 and Table A1-2) upon generating through the physico-chemical analytical process.

The details methods of analyses, quality control and quality assurance procedures, data assessment and validation process have been explained in the published report (Khan, M., Anwar, K.M.M., Chowdhury, H., Bangladesh Academy of Sciences, 2008). The laboratory analyses have been conducted as per the applicable international best practices and norms to provide with a sound basis of the validity of the data (Annexure 1: Initial Data Matrix X Table A1-1 and Table A1-2) containing testing results from the laboratory are in Excel Spreadsheet.

In this present study the authors applied various multivariate data analytical techniques e.g. Principal Component Analysis, Factor Analysis and finally the Cluster Analysis: both hierarchical and non-hierarchical k-means method to explore and study further the quality and attributes of the commercialized bottled water as well as to investigate any interesting physico-chemical phenomena.

Prior to applying these advance multivariate techniques the authors also applied some other basis statistical techniques to gain overall understanding on the data as well as to "have a general feel about the data" through descriptive, exploratory statistical analysis. The resluts of these statistical analyses have been summarized in the Table 4.4 and Table 4-4. This approach particularly helped us to decide the pathway for the next advanced data analysis, to give a vision about the potential number of groups or clusters as well their probable physico-chemical nature.

It is worthwhile to note that the data and variables used here in this study are quasi exhasutive to explain the individuals in almost full with respect to macro elemental compositions. There are some variables specially the trace metals, organics substances, contaminants due to agricultural purposes and or residual antibiotics and or other industrial waste and substances may be required to explain or understand the bottled water matrix to its full.

But from the present analytical and academic point view, the number of variables are may considered to be as sufficient for this study as well as to explain in general the bottled water under study with respect to macro elemental compositions.

In this study through application of multivariate analyses the authors have indeed improved knowledge and understanding about the individual brands under study as well as assumed the

opportunity to apply these kind of multivariate statistical techniques in revealing the underlying physico-chemical and or quality phenomena having particular importance in industrial quality assurance quality control (QAQC), consumers rights protection, market surveilance, standardization and regulatory regime.

In the subsequent sections the authors described the outcomes of the various data analyses process and recorded the observations, possible interpretation, limitations and applicability of the techniques for this particular type of water quality assessment business. After a thorough discussions the authors have drawn a conclusion to summarize the study.

4.2. CODE FOR INDIVIDUAL, BRAND, VARIABLE, SCALE OR UNIT OF MEASUREMENT

The codes for individual brands from different manufactures has been assigned as per the Table 4.1 in the this study. Observations i.e. "Individuals" from each brand has been given relatively shorter codes to mark every single observation which could be seen in the first column of the Initial Data Matrix X (Annexure 1). To illustrate further let us consider a few examples: several observations or Individuals from BRAND02 has been denoted as B2a, B2b, B2c and so on. And Similarly, other Individuals from BRAND05 has been coded as B5a, B5b, B5c, B5d etc. For two Individuals or observations from DEIONIZEDWATER have been coded as DIWa and DIWb.

Sl.No.	Individual /Brand Code	Commercial Name/Brand
1	BRAND01 (B1)	Ampang
2	BRAND02 (B2a, B2b, b2c, B2d, B2e and B2f)	Aqua Mineral
3	BRAND03 (B3a, B3b, B3c, B3d, B3e and B3f)	Duncan's
4	BRAND04 (B4a and B4b)	Everest
5	BRAND05 (B5a, B5b, B5c, B5d and B5e)	Fresh
6	BRAND06 (B6a, B6b, B6c, B6d, B6e and B6f)	Mountain
7	BRAND07 (B7a, B7b, B7c, B7d and B7e)	Mum
8	BRAND08	Pran

Sl.No.	Individual /Brand Code	Commercial Name/Brand
	(B8a, B8b, B8c, B8d, B8e and B8f)	
9	BRAND09 (B9a, B9b and B9c)	Samurai
10	BRAND10 (B10a, B10b, B10c and B10d)	Yes
11	BRAND11 (B11a, B11b, B11c and B11d)	Trishna
12	DEIONIZEDWATER (DIWa and DIWb)	DEIONIZEDWATER
Total Individual: 51 of total Brand 12		

Table 4.1 – Code for Individual/Brand under study

The codes for variables: physico-chemical properties have been assigned and shown in the Table 4.2 where the units or the scales of measurements are also tabulated. As it has been observed that the units are different therefore, almost all the time most of the analysis under this study would be done on centered or standardized data matrix. This is indeed a very important assumption to be remember from now on throughout the rest of the study.

Sl.No.	Code	Variable Name	Scale/Unit of Measurement
1	TEMP	Temperature	Deg C
2	pH	pH	-
3	EC	Electrical Conductivity	uS/cm
4	NH4	Ammonium	mg/L
5	NO2	Nitrite Nitrogen	mg/L
6	NO3	Nitrate Nitrogen	mg/L
7	SO4	Sulphate	mg/L
8	Cl	Chloride	mg/L
9	HCO3	Bi-carbonate Alkalinity	mg CaCO3 /L
10	F	Floride	mg/L
11	HARD	Hardness	mg CaCO3/L
12	FreeCN	Free Cyanogen	mg/L
13	COD	Chemical oxygen Demand	mg O2/L
14	TDS	Total Dissolved Solids	mg/L

Sl.No.	Code	Variable Name	Scale/Unit of Measurement
15	Na	Sodium	mg/L
16	K	Potassium	mg/L
17	Ca	Calcium	mg/L
18	Mg	Magnesium	mg/L
19	Fe	Total Dissolved Iron	mg/L
20	Mn	Total Dissolved Manganese	mg/L
21	ANIONS_SUM	Total Anions	meq/L
22	CATIONS_SUM	Total Cations	meq/L

Total Variables: 22. Different Measurement Scales/Interval are mg/L: miligram per litre, CaCO3: Calcium Carbonate, uS/cm: microSiemens per cm, meq/L: miliequivalent per litre, O2/L: Oxygen per litre

Table 4.2 – Code and Measurement Unit for Variable

4.3. INITIAL DATA MATRIX X AND CENTERED DATA MATRIX X_CENTERED

Initial Data Matrix X in Excel Spreed Sheet (Annexure 1: Table A1-1 and Table A1-2) has been constructed by further reducing and summarizing the orginal raw data. The tabulated *Initial Data Matrix X* consists of a data set from analysis of 22 variables for total of 51 observations or Individuals including 49 Individuals from 11 Brands (BRAND01 to BRAND11) and 2 individuals from controlled laboratory reagent grade water coded as DEIONIZED WATER (DIWa and DIWb).

One variable TEMP-EC from the original raw data matrix has been droped because this parameter is coupled or paired with variable EC. And another TEMP (TEM-pH: coupled with pH) is retained for this study which is essentially related to the laboratory room temperature during the analysis and is coupled with the variable pH indeed. Considering the natural properties, to explain the water chemistry it is sufficient to study TEMP with pH instead of using TEMP with EC. As per the international standard practices the laboratory analysis is to be conducted at room temperature ranging from 20 Deg C to maximum 30 Deg C. In this case it has been observed that all the analysis has been conducted within the TEMP range 25.1 to 27.25 with an average 26.1 Deg C which are within the acceptable limits. Hence after this stage the authors conntinued study with 22 variables instead of original 23. To drop variable TEMP-EC (coupled with EC) at thi stage as well as in support of this initial inference we have conducted a single factor ANOVA analysis (Table 4.3) to test a Null Hypothesis at 5% significance level to see that the data or results of variables TEMP coupled with EC indeed came from the sample population of the results of the variable TEMP coupled with pH. Alternatively it could be reasonably assumed that there is no significant (at 5% level) evidence that they came from the different populations. Hence the authors' decision to drop the variable TEMP coupled with EC have sufficient justification and retaining only one variable TEMP with pH would be

sufficient to explain further the population under study. And instead of using code TEMP-pH we the code for this variable will remain to be as TEMP for the rest of the study for convenience.

Groups	Count	Sum	Average	Variance		
TEMP-pH	51	1325.5	25.99019608	0.354901961		
TEMP-EC	51	1320.2	25.88627451	0.374007843		
TEMP-pH: Temperature coupled with pH						
TEMP-EC: Temperature coupled with EC						
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.275392157	1	0.275392157	0.755627529	0.386781976	3.936142986
Within Groups	36.4454902	100	0.364454902			
Total	36.72088235	101				
Null Hypothesis Ho:	The TEMP-pH data coupled with pH have indeed came from the same population of the TEMP-EC data coupled with EC as such one set of TEMP couple with pH could be acceptable for further study and another TEM-EC could be dropped.					
Alternative Hypothesis Ha:	TEMP-pH and TEMP-EC came from different populations hence they are representing two different data sets as such no one should be droped					
Inference:	Since F (=0.755627529) is less than Fcrit (3.936142986) with p-Value: 0.386781976 Null Hypothesis Ho is NOT rejected at 5% significant level. Alternatively there is NO significant (5%) evidence to reject the null hypothesis that they are actually representing the same population, hence any one variable , say, TEMP-pH could be rertained and other one could be dropped in course of the further investigation					

Table 4.3 – Summary of Single-Factor ANOVA Analysis

The DEIONIZED WATER (DIWa & DIWb) are indeed two known pure water individuals, theoretically, not having any mineral, anion, cation or dissolved solid, suspended solid or particulate materials. This version of water is produced through applying a very sophisticated technolgy removing all anion, cations, organics, microbs and particulate materials. The specifications defined in the ASTM TYPE-I standard clearely stated that Electrical Conductivity (EC) and Electrical Resistivity of this deionized water, called ASTM TYPE-I, must be around 0.056 uS/cm or 18.3 Mohm-cm respectively. It is expected that this very individual DEIONIZED WATER (DIW) must show a very extreme behaviour in terms of physics and chemistry in comparison to other water samples under study.

Standardized or Centered Data Matrix *X* (Annexure 2: Table A2-1 and Table A2-2) consists of 22 variables for total of 51 Individuals is generated using *standardization* function of SAS Enterprise Guide 7.1(64-bit).

For conducting PCA and HCA and K-means, the Codes in Programing language R has been developed where at the very onset of running the algorithm, the initial data matrix (Annexure 1) in .csv format generated using MS Excel is applied. All outputs of PCA and CA analyses produced from this R Codes developed during this study have been organized in Tables attached in the Annexures.

4.4. DESCRIPTIVE STATISTICS AND GENERAL IMPRESSION ABOUT THE DATA

4.4.1. Descriptive Statistics

As mentioned earlier to understand the general trend of the variables the descriptive statistical analysis have been done and summary has been recorded in the Table 4-4 and 4-5. Further detailed results of this analysis have been tabulated and attached in the Annexure 3 (Table A3) where mean, median, standard deviation, minimum and maximum, 1st Quartile and 3rd Quartile values of each variable have been recorded.

VARIABLE	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD	FreeCN
MEAN	26.07	7.29	334.04	0.12	0.01	5.93	2.50	35.98	109.18	0.37	27.78	0.01
STDEV	0.55	0.89	316.04	0.04	0.00	10.82	0.00	57.48	92.88	0.17	30.90	0.00
Minimum	25.10	5.82	2.70	0.10	0.01	0.10	2.50	0.30	3.70	0.10	0.25	0.01
Maximum	27.25	9.16	996.37	0.24	0.02	36.10	2.50	185.17	301.60	0.55	87.09	0.01

Table 4.4 – Summary of Descriptive Satistics

VARIABLE	COD	TDS	Na	K	Ca	Mg	Fe	Mn	ANIONS_SUM	CATIONS_SUM
MEAN	20.00	217.96	36.07	1.84	21.10	6.57	0.19	0.11	2.97	3.23
STDEV	0.00	212.04	32.70	1.57	23.61	7.75	0.05	0.14	2.64	2.79
Minimum	20.00	1.73	0.00	0.05	0.01	0.03	0.10	0.04	0.13	0.10
Maximum	20.00	687.69	103.19	5.96	67.48	20.08	0.28	0.56	7.70	8.70

Table 4.5 – Summary of Descriptive Satistics

4.4.2. Important Observation: Non-consideration of Variables NO2, SO4, Free CN and COD

From the descriptive statistical results it has been revealed that four (04) variables NO2, *SO4, FreeCN* and *COD* having Standard Deviation=0 showing no variability hence they are not interesting for the further multivariate Principal Component Analysis (PCA) and Factor Analysis (FA) because these variables would not be able to contribute to PCs or Factors. Hence duirng the subsequent further analyses these four varibles would not be used. Hence finally 18 variables are retained for the rest of the study.

4.4.3. Plotting Histograms, Box Plots & Non-Parametric Tests for Normality

For each variable, non-parametric normality test (for 95% Confidence Interval) has been done. Some of the varibles (e.g. TEMP, pH, HCO3, Fe, Mn, Cations_Sum etc.) shown the behaviour Normal and others (e .g.EC, TDS, NH4, HARD, Cl, F, Na, K, Ca, Mg, Anions Sum etc.) were not Normally distributed.

As an example, when the variable total dissolved iron (Fe) is considered (Figure 4.1), it has been observed that this variable is behaving normally as per the Anderson-Darling Normality test where the Null Hypothesis (Ho: The Fe data came from a normal distribution) has NOT been rejected at 5% level (p-value:0.259).

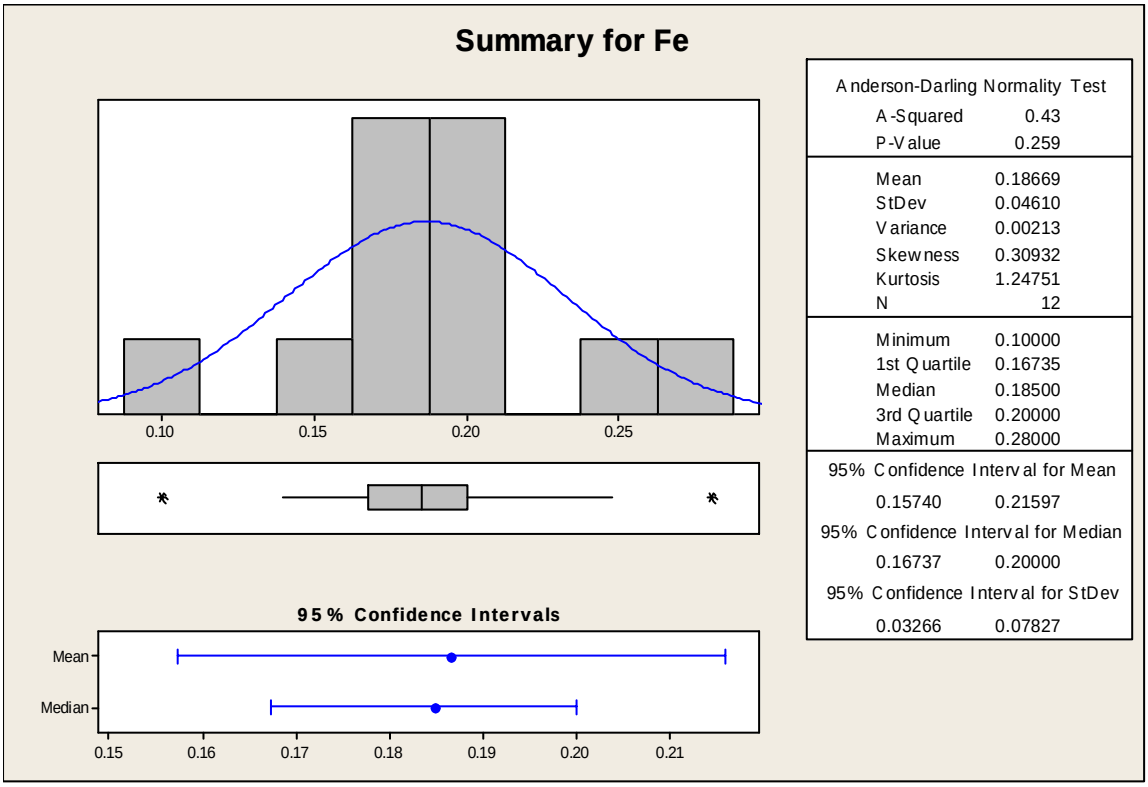

Figure 4.1 – Results of Anderson-darling Normality Test for Fe

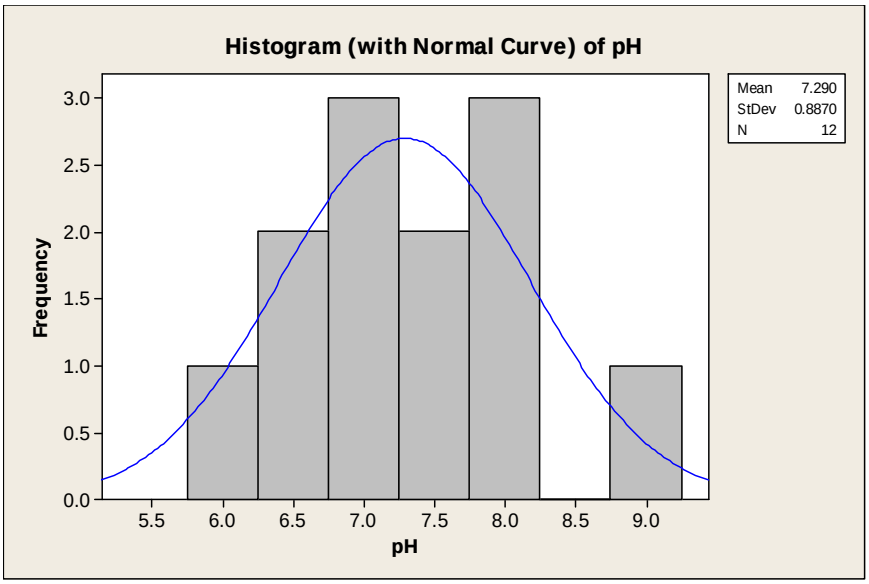

Figure 4.2 – Histogram of pH

But when the similar non-parametric test is performed for another variable, e.g. Mg the data for this variable seemed not to be Normal one (Null Hypothesis Ho: Mg Data comes from a Normal Population has been rejected at alpha=0.05 where p-value is = 0.005) (Figure 4.3). Similar situation happened for the data for variable Ca (Figure 4.4).

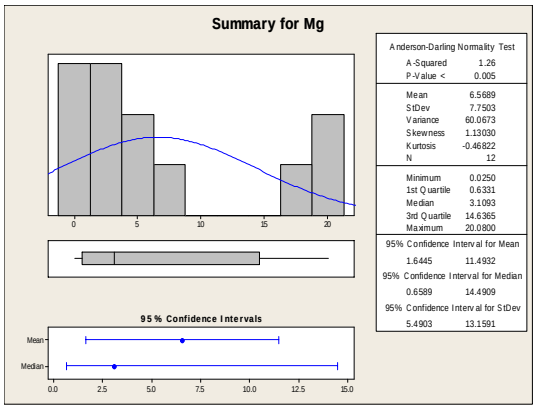

Figure 4.3 – Normality Test Results for Variable Mg

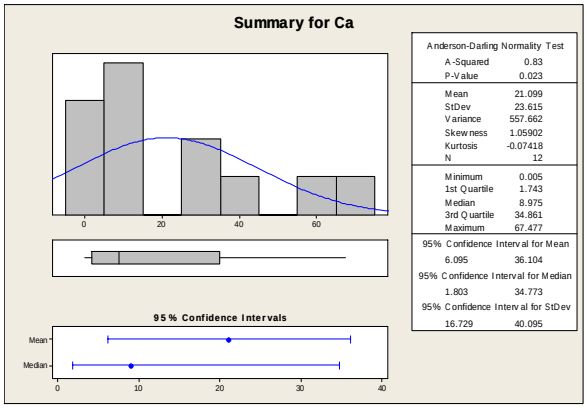

Figure 4.4 – Normality Test Results for Variable Ca

It is obvious that the sample botteld waters collected from the market are in fact produced by the various different manufacturers following various treatment process. These Brands are essentially not from the same population. Different manufacturers in fact are creating different populations or universes.

Moreovere, from the histogram of variable TEMP (Figure 4.5) it has been quite reasonably understood that the laboratory conditions or laboratroy temperatures were under controlled as such the room temperatures were varying normally around the estimated mean value.

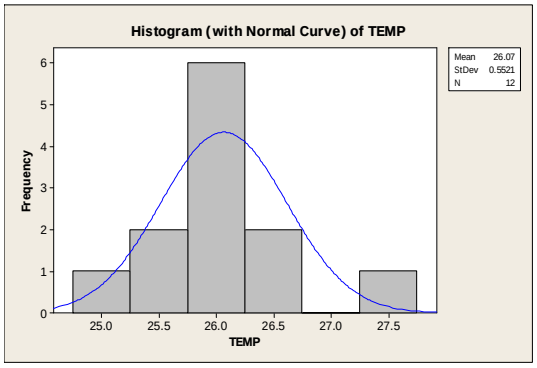

Figure 4.5 – Histogram for TEMP (Temperature)

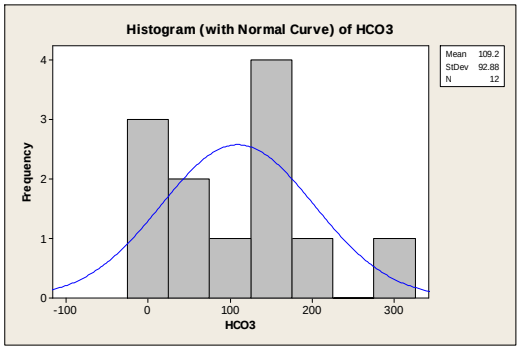

Figure 4.6 – Histogram of Bicarbonate Alkalinity HCO3

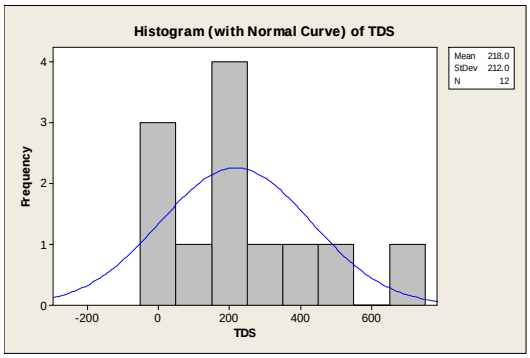

Figure 4.7 – Histogram & Normal curve for TDS

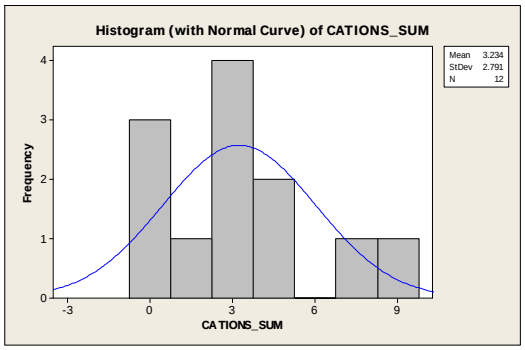

Figure 4.8 – Histogram & Normal Curve for CATIONS_SUM

Box plots have also been studied for all the variables under study to obtain knowledge on outliers and other spurious data. As examples the Box Plots for one variable Total Cations Sum and another variable Electrical Conductivity EC have been shown (Figure 4.9 and Figure 4.10).

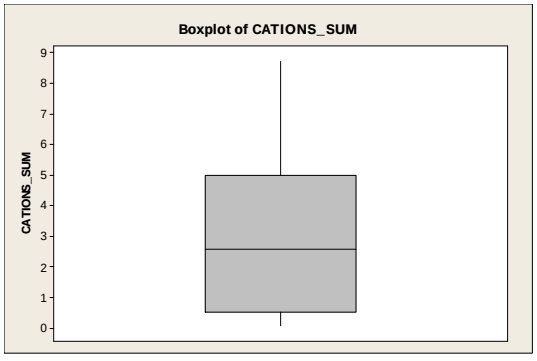

Figure 4.9 – Box Plot for CATIONS_SUM

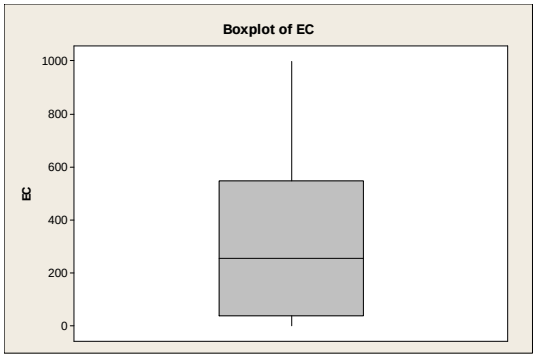

Figure 4.10 – Box Plot for EC

4.4.4. Kruskal-Wallis Tests for Variable vs Brand

In this study, a Kruskal-Wallis test was performed for each variable and brand (Table 4.5).
This statistical test was chosen to determine if production process at each factory (or each brand) has an influence on water quality parameters. The null hypothesis is that when looking at one particular bottled water variable, there is no difference of the median value in each of the various brands (H0 = M1 = M2 = M3…). The alternate hypothesis (Ha = medians are not all equal) is that there is a difference in the median value for each brand.
i.e.
Ho : M1 = M2 = M3… (there is no difference in median value of each of the various brands)
Ha: medians are not all equal or at least one median is different (there is a difference in the median value for each brand)

When using the $\alpha < .05$ significance level, every water quality parameter with the exception of iron demonstrate that for these parameters the hull hypothesis may be rejected. There is enough evidence to demonstrate that the measurements vary by brand. The results of this statistical test indicate that production process for differing brand from different factory is indeed a driving factor in the quantities of many of these water quality parameters under study. This is consistent with the concept that treatment process at factory is a prevailing factor in bottled water quality.

Usually, iron (Fe) concentrations should vary between different brands or producers as they are most commonly associated with varying production processes. However, results from this analysis show that the null hypothesis cannot be rejected which may be a result of the modification done to adjust the detection limit for the iron measurements. The Ryan-Joiner or Shapiro-Wilk non-parametric test for probability distribution (Figure 4.11) also shows the result is consistent with the above observation. From the graphical summary of the statistical analysis (Figure 4.1) it is evident that more than 25% samples are having the median values 0.20 mg/L which is actually the modified value for adjusting the detection limit of the analytical methods or measurement process and probably this most likely had an impact on the results of this statistical test on the iron concentrations. Another possibility is that as it has been assumed that the manufacturers are mostly utilizing ground water as their starting raw material either they do not installed iron removal system for pretreatment or their iron removal process may not fully effective and thus the overall Fe distribution is following natural distribution. But this particular issue could be further investigated in future research.

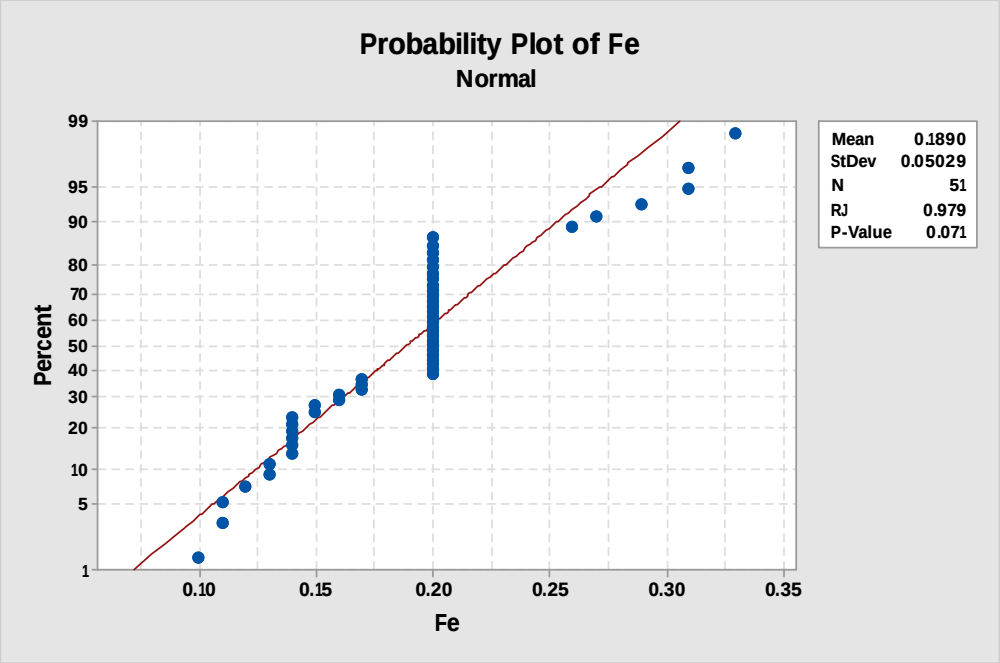

Figure 4.11 – Non-parametric Ruyan – Joiner (Shapiro-Wilk) Probability Test for Fe

Variable	H Statistcs (Chi-Square)	Degress of Freedom	Significance (p-Value)	Ho Accepted	Ho Rejected
TEMP	32.20	11	0.001 (adjusted for ties)		√
pH	47.28	11	0.000 (adjusted for ties)		√
EC	49.06	11	0.000		√
NH4	20.00	11	0.045 (adjusted for ties)		√
NO2				All values in column are identical	
NO3	46.91	11	0.000 (adjusted for ties)		
SO4		11		All values in column are identical	
Cl	48.57	11	0.000 (adjusted for ties)		√
HCO3	48.92	11	0.000 (adjusted for ties)		√
F	46.01	11	0.000 (adjusted for ties)		√
HARD	48.24	11	0.000		√
CN				All values in column are identical	

Variable	H Statistcs (Chi-Square)	Degress of Freedom	Significance (p-Value)	Ho Accepted	Ho Rejected
COD				All values in column are identical	
TDS	49.6	11	0.000		√
Na	47.85	11	0.000		√
K	47.14	11	0.000 (adjusted for ties)		√
Ca	48.56	11	0.000 (adjusted for ties)		√
Mg	47.77	11	0.000 (adjusted for ties)		√
Fn	18.92	11	0.062 (adjusted for ties)	√	
Mn	29.31	11	0.002 (adjusted for ties)		√
ANIONS_SUM	49.00	11	0.000 (adjusted for ties)		√
CATIONS_SUM	48.99	11	0.000 (adjusted for ties)		√

Table 4.6 – Kruskal-Wallis Test Results for Variable vs Brand

4.4.5. Correlogram & Correlation Matrix among the variables

From the Table 4.7 and 4.8 (correlation matrices) and the Figure 4.12 (correlogram) it has been oberved that the in general the correlation coefficient values are showing natural physical and chemical behaviour. From the established and existing physics and chemistry of water this correlation matrix could be explained. As an example Electrical Conductivity EC and Total Dissloved Solids (TDS) are strongly correlated (r ~1.00) which is quite obvious from the physical and chemical point of view, because dissolved substances have indeed contributed in building the electrical properties of water. By nature water H_2O, having slight electrical dipole moment and assymetry in charge distribution showing the property of an electrical non-conductor. Pure water in liquid state releases less cations and it is slighlty acidic in nature having relatively less H+ ions i.e. low pH ~6.4 or below. But when some minerals, substances and or ions are dissolved in the water then these substances are dissociated in the solution to assume ionized forms and as such they contribute in transforming the overall solution to be an electrical conductor as such giving rise to electrical conductivity value EC higher. Therefore, Total dissolved substances measured in TDS are positively contributing to increase in EC of the water solution. In practical terms, EC and TDS are so coupled that utilizing this natural properties of the water solution, invariably in several analytical techniques, measurement of EC are directly being used to estimate the TDS and turbidty values of the water in a solution phase only upon applying some slight temperature correction. This coupling between EC and TDS has also been further evident in this study as such the correlations among ANIONS_SUM, CATIONS_SUM, EC, TDS, are showing the highest correlation values ~1.0.

	TEMP	pH	EC	NH4	NO3	Cl	HCO3	F	HARD
TEMP	1								
pH	-0.42	1							
EC	-0.08	0.32	1						
NH4	0.17	-0.06	-0.02	1					
NO3	0.26	-0.14	0.32	0.18	1				
Cl	-0.06	0.21	0.89	-0.13	0.19	1			
HCO3	-0.02	0.25	0.73	0.26	0.37	0.46	1		
F	-0.15	0.58	0.61	0.2	0.38	0.39	0.71	1	
HARD	-0.37	0.53	0.85	-0.06	-0.05	0.76	0.65	0.53	1
TDS	-0.08	0.31	0.98	-0.03	0.28	0.87	0.69	0.58	0.83
Na	0.21	0.07	0.84	0.03	0.62	0.73	0.65	0.57	0.44
K	-0.36	0.41	0.72	0.02	0.32	0.57	0.73	0.69	0.81
Ca	-0.39	0.58	0.83	-0.08	-0.06	0.74	0.62	0.53	1
Mg	-0.25	0.31	0.88	0.04	-0.01	0.79	0.71	0.49	0.95
Fe	-0.06	0	-0.12	0.14	-0.24	-0.07	-0.05	0.04	0.03
Mn	0.16	-0.04	-0.01	0.61	-0.07	-0.15	0.33	0.1	0.02
ANIONS_SUM	-0.03	0.25	0.96	0.04	0.37	0.9	0.8	0.62	0.81
CATIONS_SUM	-0.11	0.35	0.99	-0.01	0.31	0.88	0.77	0.64	0.87

Table 4.7 – Correlations among Variables (to be continued...)

	TDS	Na	K	Ca	Mg	Fe	Mn	ANIONS_SUM	CATIONS_SUM
TDS	1								
Na	0.81	1							
K	0.68	0.47	1						
Ca	0.81	0.41	0.81	1					
Mg	0.86	0.53	0.78	0.93	1				
Fe	-0.11	-0.2	0.13	0.01	0.06	1			
Mn	-0.01	-0.02	-0.09	-0.02	0.12	0.02	1		
ANIONS_SUM	0.92	0.84	0.74	0.78	0.86	-0.09	0.05	1	
CATIONS_SUM	0.97	0.83	0.77	0.84	0.89	-0.09	0.01	0.97	1

Table 4.8 – Correlations among Variables

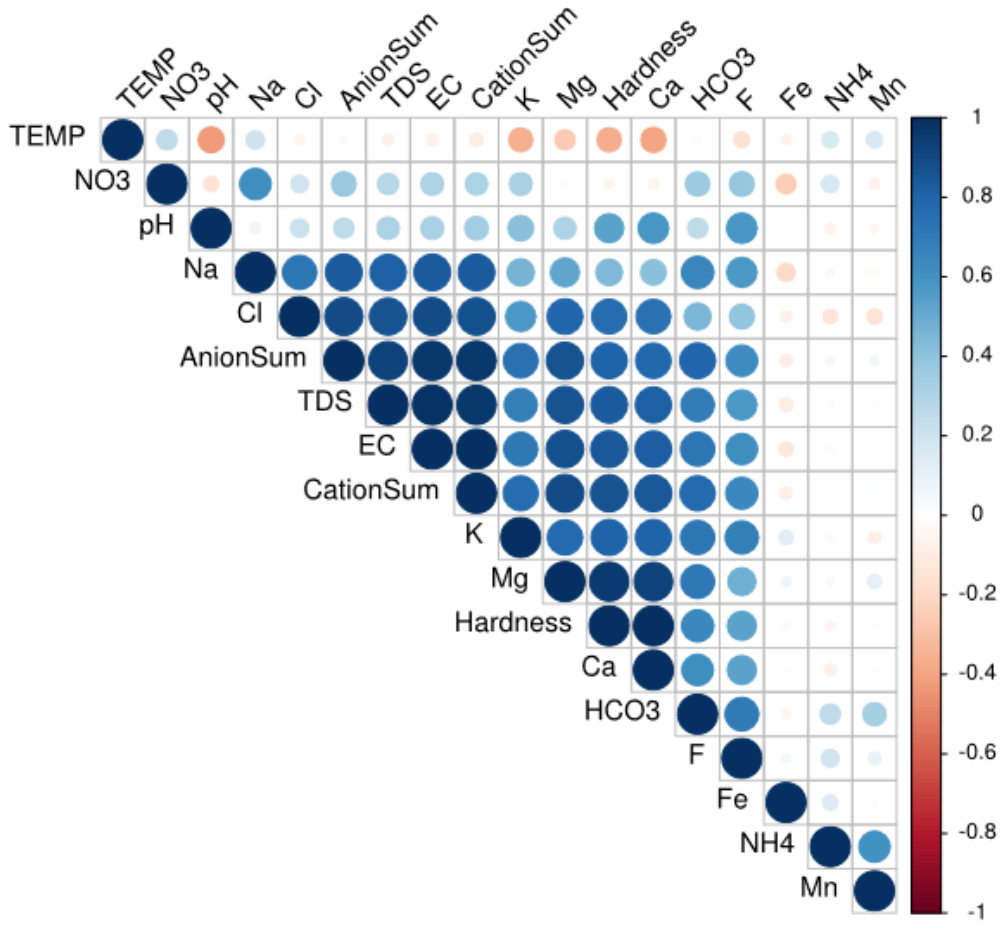

Figure 4.12 – Correlogram among the original variables used for further PCA Analysis

Ca, Mg, Fe, Hardness (HARD), HCO3, TDS, EC, ANIONS_SUM, CATIONS_SUM (r>0.7) are strongly correlated amomg themselves which is also clearly indicating that hardness as well as ion buildup has been done mostly by Ca, Mg, Fe based salts dissolved in the water. These variables are also giving mostly the “aesthetic nature” and or “softness”/”hardness”/”lightness” and aggregate properties of water. That means, less the values of these variables, softer or lighter the water. Higher values of these elements making the water more harder and making the water more unsuitable for human consumption with respect to the aestheic point of view. Alkalinity (HCO3) and Hardness are in general the measures of the “hardness” or “lightness” or “softness” of water.

Na and Cl are also strongly correlated (r=0.75) are giving indication that probably the water matrices are having sufficient amount of dissolved NaCl salts. Bangladesh being a country situated at the deltaic plane below The Himalayas and on the floodplane adjucent to the Indian Ocean it is very likely

that the manufacturers utilized groundwater as the raw materials mostly to produce these bottled water and are having higher values of NaCl originated from the deep acquifars.

Other variables are also showing some correlations which would be further used and explored to build the PC/Factors. It is obvious that Principal Component Analysis (PCA) as well as Factorial Analysis only could be meaningfully applied if there are significant correlations present among the variables. The eigen values or the inertia or the dispersions would be extracted from this correlation matrix to find out the Principal Components/Factors.

It would be the aim of this study to build the uncorrelated new synthetic variables .i.e. Principal Components/Factors to explain any latent behaviors of the individuals and or original variables.

4.4.6. Poor Correlations among pH, TEMP and other Variables

The variables pH and TEMP are showing very little correlations with other variables which is also seemed to be obvious in terms of physics and chemistry. It could be reasonably assumed that during the PCA, Factor and Cluster Analysis these variables may not be contributing much in building the PCs/Factors. Hence duirng the interpretation of PCA results these two variables may not be frequenty dicussed and or not incorporated. And they may not also be useful for further explaining the behaviour of the individuals in this particular context of analysis.

Altough the knowledge of chemistry and physics suggests that pH and TEMP are two variables useful to understand the nature of water but from this initial observation of the correlation pattern we may assume that PCA and or Cluster Analsyses or Factor Analysis may not fully be able to capture the effect of these two variables. This aspect could be further explored while we progress further on the way to our study.

4.4.7. Summary on Descriptive Statistics

In short, it is possible from the above exlporatory study to state that the important outcomes have been recorded that the bottled water data are not from the same popululation as they are produced by the different manufactures utilizing different treatment technologies at different times and in different batches.

And four variables NO2, SO4, Free CN and COD would not be further useful during the PCA, FA and Cluster analyses as such we are remaining with 18 variables.

Though PCA/FA/Cluster Analysis, we may not fully capture the contributions from pH and TEMP as well as their effects or contributions in building PCs/Factors as they are not significantly correlated with other variables. However, these two variables would be maintained throughout the study to understand further this aspect.

From the correlation matrix it has been evident that some variables e.g. EC, TDS, ANIONS_SUM, CATIONS_SUM, Na, Ca, Fe, HCO3, Cl, F may be sufficiently utilized to build PCs as well as to create explanation. But other variables like pH, TEMP may not fully be covered or explained as it has again been apparent from the correlation matrix.

It has always been admitted that all statistical techniques have their limitations as such the applicability and suitability must be considered for the particular field of application. In this study this issues and constraints have also been recognized.

4.5. PRINCIPAL COMPONENT ANALYSIS

4.5.1. Extraction of Eigen Values and Cumulative Percentage of Variance from Correlation Matrix

As stated earlier the measurement scales or units of the variables are different hence instead of using the covariance matrix, in this study the *correlation matrix* (Table 4.7, 4.8) has been used to extract the eigen values and or the Principal Components. This SAS or other statistical pakage having procedure of PCA analysis essentially utilizes the standardized data or use correlation matrix.

	Eigenvalue	Percentage of Inertia	Cumulative percentage of Inertia Explained
λ_1	9.64	53.53	53.53
λ_2	2.36	13.13	66.66
λ_3	1.83	10.18	76.84
λ_4	1.22	6.79	83.62
λ_5	0.97	5.41	89.03
λ_6	0.65	3.61	92.64
λ_7	0.45	2.49	95.13
λ_8	0.35	1.93	97.07
λ_{19}	0.21	1.19	98.25
λ_{10}	0.14	0.78	99.04
λ_{11}	0.1	0.58	99.62
λ_{12}	0.04	0.21	99.83
λ_{13}	0.02	0.11	99.94
λ_{14}	0.01	0.04	99.98
λ_{15}	0	0.02	100
λ_{16}	0	0	100
λ_{17}	0	0	100
λ_{18}	0	0	100
Total	18	100	

Table 4.9 – PCA Output Eigen Values & Cumulative Percentage of Variance/Inertia covered by Eigen values extracted from the Correlation Matrix

From the PCA analysis it is possible to evaluate the total intertia I_g and the total projected intertia I_g^* on the first few (say, 4) Principal Axes generated by first four eigen vectors of the correlation matrix.

Total Interia:

I_g = 9.6351+2.3630+1.8324+1.2215+0.9743+0.6489+0.4490+0.3478+0.2137
+0.1411+0.1040+0.0385+0.0191 +0.0080 +0.0032+0.0004+0.0000+0.0000
= 18

As expected the sum of these eigen values gives value 18 exactly equals to the number of variables.

The total projected inertia on the first four Principal axes CP1, CP2 and CP3 and CP4:

Ig* = 9.6351+2.3630+1.8324+1.2215
= 15.05

From the PCA output it is evident that the first four Principal Components, namely
PC1 (λ_1 =9.6351, Variance/Intertia explained: 53.53%),
PC2 (λ_2=2.63630, Variance/Intertia explained: 13.13%),
PC3(λ_3=1.8324 Variance/Intertia explained: 10.18%) and
PC4(λ_4=1.2215 Variance/Intetria explained: 6.79%)
are explaining ~83.62% *[*= $(I_g^*/I_g) * 100$*]* of the total inertia of the data.

From the Scree Plot (Figure 4.13), Eigen Value Plot (Figure 4.14) shown below as well as from the application of the Pearson's criteria as well as Kaisar's Criteria it has been possible to reach a decision that the first four (04) Principal Components are sufficiently explaining the intertia more than 83% of the total intertia I_g of the original data, as such it would not be unwise to consider these first four (04) Principal Components (i.e. q=4): PC1, PC2, PC3 and PC4 for further interpretation of the data as well as to proceed for further analysis.

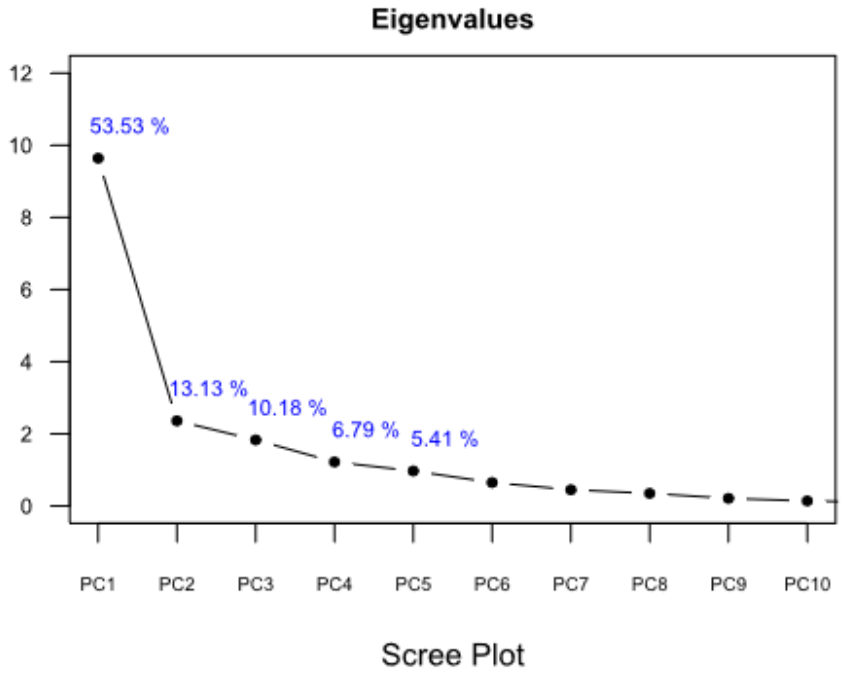

Figure 4.13 –Scree Plot

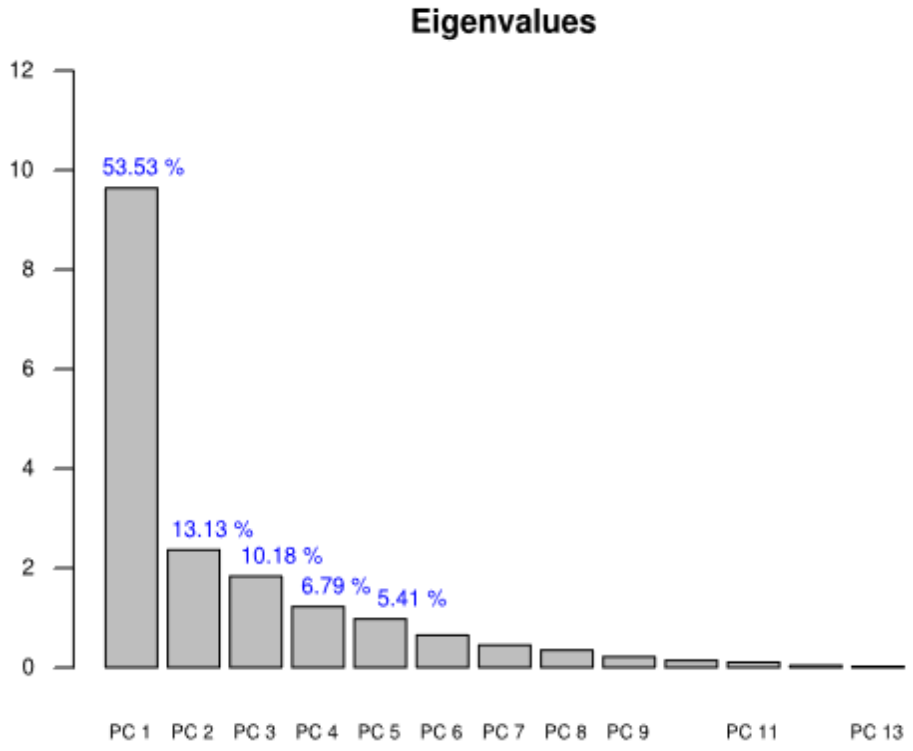

Figure 4.14 – Eigen Value Plot shows Percentage of Variability covered by each PC

Therefore, it could be reasonably stated that through this PCA technique the dimensions have been now reduced further from 18 to only 4 (orthogonal and uncorrelated ones) to explain sufficiently (83.62%) the given data set.

Rest of the orthogonal components are in fact constructing the residual matrix hence containing little inertia considered to be as noise which are not so interesting from this present study point of view. The Table 4.10 contains the Branwise Residual Matrix in Individual Space constructed upon running PCA which shows the coordinates for 12 Brands (11 Brands and DEIONIZEDWATER) along the next, say, seven PCs just noise.

In Annexure 4 (Table A4) we have also attached the Residual Matrix for Variables in the Variable Space where loadings along the rest of the 14 PCs for variables are tabulated. Surely this residuals are not containing any usable information leaving us with only noise.

	PC5	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**
BRAND01	0.05	-0.09	-0.09	0.15	-0.15	-0.12	0
BRAND02	-2.08	-0.11	-0.11	0.26	-0.05	0.06	0
BRAND03	-0.43	0.41	0.48	-0.26	-0.29	0.89	0.03
BRAND04	1.99	0.18	0.41	0.04	-0.31	-0.05	0
BRAND05	1.19	-1.52	0.69	-0.02	0.21	0	0.01
BRAND06	-0.22	-0.08	-0.22	0.13	1.28	0.1	0.03
BRAND07	-1.06	0.45	1.1	0.84	-0.2	-0.29	-0.03
BRAND08	-1	0.17	0.62	-1.22	0.01	-0.39	-0.02
BRAND09	0.06	-1	-0.89	0.02	-0.19	0.09	-0.2
BRAND10	-0.2	-0.89	-0.9	0.05	-0.41	-0.14	0.19
BRAND11	0.33	1.36	-1.4	-0.04	-0.02	-0.1	-0.01
DEIONIZEDWATER	1.35	1.11	0.31	0.06	0.12	-0.05	0.01

Table 4.10 – Residual Matrix

In the subsequent sections, the individuals here in our case the Brands as well as the original variables and their contributions, behaviours, significance would be explained in terms all the new synthetic dimensions i.e.PCs/Principal Factors.

As it is obvious that in this *new 4-D phase space* the first dimension *PC1 covers the maximum variability (53.53%)* of the data, and the second dimension PC covers the next highest variability (13.13%) and so on.

4.5.2. Correlations between the Original Variables and the First Four Principal Components denoted as PC1, PC2, PC3 and PC4

It is very imporant to see the correlations among the new synthetic dimensions i.e.PCs with the original variables. These correlation coefficients ultimatley would be utilized to estimate the absolute (CTA) and relative contributions (CTR) from each original variables in building the new synthetic dimensions or indices (PCs). All the individuals would further be explained, understood in terms of these new 4-D which is indeed a projection. This new coordinate system is, in another words" in fact a rotated one and shifted to an origin at *"the rigid body where the the centre of gravity of the originally located"*. The individuals are now away or close forming the Cloud around this origin or the center of gravity. Their distance could be estimated, say, by applying Euclidean Principal; square root of the sum of the squars of the coordinate values (discussed in the leter subsequent sections).

	PC1	PC2	PC3	PC4
TEMP	-0.19	0.72	-0.1	0.28
pH	0.43	-0.52	0.21	-0.46
EC	0.97	0.1	-0.13	0.1
NH4	0.02	0.42	0.76	0.03
NO3	0.29	0.7	-0.18	-0.49
Cl	0.85	0.01	-0.31	0.27
HCO3	0.8	0.26	0.32	-0.1
F	0.7	0.08	0.28	-0.52
HARD	0.91	-0.35	0.08	0.14
TDS	0.95	0.08	-0.13	0.13
Na	0.76	0.52	-0.25	-0.06
K	0.84	-0.16	0.12	-0.24
Ca	0.89	-0.39	0.06	0.09
Mg	0.91	-0.18	0.11	0.3
Fe	-0.06	-0.26	0.37	0.17
Mn	0.02	0.31	0.78	0.28
ANIONS_SUM	0.96	0.18	-0.07	0.1
CATIONS_SUM	0.99	0.08	-0.08	0.07

Table 4.11 – Correlations between original Variables and the PCs

- Principal Axis 1: PA1

From the correlation matrix (Table 4.11) it has been clearly evident that the first Principal Axis or Component (covering more than 53% of total inertia) has attracted most of the varibles including the major cations: Na, K, Ca, Mg which are probably associated with major anions: chloride (Cl), carbonate, fluride (F) as dissolved salts playing key roles in framing the main physical and chemical texture of the water under study. From looking at the Figure 4.15 it is immediately visible. This Axis is also sufficiently explaining other aggregate properties like EC, HCO3, HARDNESS, TDS, Anions_Sum and Cations_Sum. And from the chemistry we are fully awar of the fact that Ca, Mg together with carbonate mostly build up the hardness properties of the water. Accessive amount (higher positive

coordinates along PA1) of Hardness or HCO3 tells us that the water is "too hard" in nature and conversely, higer negative values along this first Principal Aix PA1 clearly tells us the story that the water individuals are most probably "too soft" or "ultra low mineral content" type. Thus PA1 in practical sense providing us with a factor or indicator in relation to the "nineral or material loading in the water". It explains the degree of substance dissolved in the water as well as the aggregate nature of individual.

It could be reasonably attributed to the fact that the individuals having *"higher coordinate along this PA1 towards positive direction"* containing essentially *"higher materials dissolved".* And oppositely *"having higher values towards the negative direction along PA1"* indicating the *"devoide of sufficient minerals or materials"* in the water individuals under study. This first Principal Axis could provisionally given a nomenclature to be as "Material Loading" although this aspect would be further understood and augmented later in the next section.

- Principal Axis 2: PA2

Also referring to the Figure 4.16 it may immediately assumed that Principal Axis PC 2 reflecting mainly the loading of some "Extrenous Materials" like ammonium (NH4), nitrate (NO3) etc. normally not welcomed in any drinking water as these may only appear due to some degree of contamination. But drawing this inference directy is also not straight forward because this Principal Axis captured only 13.13% of the total inertia of the original data set. This second Principal Component (Figure 4.16) also shows that pH and room temperature TEMP are opposing each other, that means, higher the temperature (coordinate towards PA2: +0.72) lower (coordinate along PA2: -0.52) the pH. Whereas Na is also attracted toward this PA2 significantly (with values: +0.52) with slight admixute of Mn (with values +0.31), the variables TEMP and pH have been playing major roles in formation of this second Principal Axis. As this very component PC2 mainly reflecting the presence of some kind of contamination we may give a provisional nomenclature to this factor to be as *"Extreneous Material Loading".* During Factor Analysis we will would be able to see again that we are not too far away of interpreting the nature of the water from this PC2, keeping also in mind that this Second Axis is not covering too much of the total inertial (only around 13%). Taking any discision about the nature of the individuals looking at this PA2 would not be so confirmatory.

- Principal Axis 3: PA3

Also refereing to the Figure 4.17 we may assume that this third principal component/axis PA3 has mostly been built from two major cations dissolved Iron (Fe), Dissolced Manganese (Mn) positively associated with carbonate anions (HCO3) most probably to form carbonate salts. Fe and Mn also decide mostly the overall taste, odor and visual appearance of the water. But it is the only Principal Component PC3 featuring soley iron (Fe) and manganease (Mn) together and it is very well known to any water chemist that these two variables are always studied in coupled as they decide the similar qualitative aspect of water jointly specially in relation to taste, odor and appearance of the water. And they also bear the information in relation to groundwater acquifers. We have not also forgetten the fact that this PC3 is only capturing the information around 10% of the original data set. Yet from the water chemistry point of view it is not unreasonable to give a provisional name of this PA3 to be

as “Aggregate Qualitative Feature” reflecting the qualitative (not quantitative) nature of the water attributed due to these two important elements iron and manganese together with carbonate anions which supposedly forming carbonate salts, typical in any ground water resource.

The variable ammonium (NH4) reflected slightly also along this third component PC3 may indicate the presence of some unacceptable materials in drinking water. But we may recall that NH4 along with NO3 has been reflected in a greater degree along the PC2. From the reality we know that ammonium NH4 and NO3 may appear in the water if there is any contamination due to, say, agricultural practices including application of nitrogen based fertilizer or pesticides as well as due to other anthropogenic activities including fecal discharge or urination in the near vicinity of the water source from where the manufacture extracted the raw water prior to treatment for packaging.

- Principal Axis 4: PA4

This component PC4 attracts (Figure 4.18) anions NO3, and F, and slightly cations iron (Fe) and manganese (Mn). As we have seen manganese together with iron contributes to the overall aesthetic features: taste, odor, appearance here we may imagine that manganease is probably associated with NO3 trigered from residual of fertilizers applied during agricultural practices in the near vicinity of the water sources used by the manufacture. The four variables: Fe Mn, NO3 and F contributing to releasing free +H ions in the solution as such help building pH. And as expected pH is here oppositing TEMP which suggests that higher the temperature resist in releasing +H ions in the water solution giving rise to lower pH value. But we have been again making this comment with a caution that this fourth Principal Component PC capturing only 6.79% of the total inertia as such there may be other underlying phenomena exist could not be fully discerned here.

Along this fourth Principal Axis we would like to associate a provisional behavioural pattern of the individuals or Brands representing or indicating a latent “ Aesthetic Acceptabiity” feature. We would be inclined to use this nomenclature while conducting the Factor Analysis at later stage along with PA4. As we indiacted this PA4 may also capture the presence of pollution although not yet fully capable of explaining this aspect.

4.5.3. Absolute (CTA) and Relative (CTRx1000) Contribution from Variables to build the Four (4) Principal Components

As mentioned above that the first Four PCs covered the variability 83.62%, it has also been observed that many variables have been sufficiently explained by this 4-Dimensions. At the extreme two right columns (Table 4.12) , the computation has been done to estimate how much inertia about a particular variable has been explained by the first four (04) PCs.To what extent these variables have been explained that have been expressed by defining some arbitrarily qualifications in the last column.

From the Table 4.12 we may observe that majority of the variables have either been “VERY WELL EXPLAINED” or been “WELL EXPLAINED” by these four Principal Axes (PAs) except Fe (Iron) which was not at all sufficiently captured (only ~23.6%) by the first four PCs. May be inclusion of another PC

e.g.PC5 may help capturing this variable but adding more variables may add more noise and adding more dimension ultimately defets our original goals of achieving parsimony and simplicity via reduction of dimension. That approach will ultimately not help us. Interestingly another variable TEMP is "FAIRLY EXPLAINED" and that is also expected because this variable indicates only the laboratory room conditions nothing to do with the composition or the chemical contents of the water under study. From this Table 4.12 it has again been sufficiently visible that EC, TDS, Na, Ca, Mg, Hardness (HARD), ANIONS_SUM and CATIONS_SUM best explained by this four PCs gave the confirmation that these variables are very important attributes and explaining the qualitative pattern of the water chemistry as a whole. As it has been noted before that dissolved solids are indeed given rise to electrical properties of water solution whereas water (pure) itself is an electrically poor performer. Total anions and total cations are again positively explaining the amount of materials loaded and their degree of presence decides the overall framework of the complex water matrix in solution form. Being a good solvant by nature water always tries to dissociate the disssolved materials into ioninc forms (either in cations and or in anions). When pH is 7.0 it is assumed that anion sums and cation sums would be equal theoretically to balance both the ions to make the water neutral in ionic sense. So if there are excessive materials loaded in the water or fortified with minerals either deliverately or due to any accedent the water gets TDS and EC values as well as Anions Sum and Cation Sum values higher. On the other hand, absence of minerals or presence of too low amount of substance again oppositely makes the water to be more "softer" or "lighter" and not suitable for consumption from the human health consideration. Ion free, ultralow mineralized or demineralized water is too risky for human health but best for industrial or laboratory purposes as they could be assumed as industry grade or laboratory water.

Variable	PA1	CTA	CTR	PA2	CTA	CTR	PA3	CTA	CTR	PA4	CTA	CTR	% of Variance Explained by First Four PCs	Remarks
TEMP	-0.19	4	35	0.72	220	520	-0.1	5	10	0.28	65	79	64.4	Fairly Explained
pH	0.43	19	184	-0.52	113	267	0.21	25	45	-0.46	174	213	70.9	Well Explained
EC	0.97	98	946	0.1	4	10	-0.13	9	16	0.1	8	9	98.1	Very Well Explained
NH4	0.02	0	0	0.42	75	177	0.76	319	584	0.03	1	1	76.2	Well Explained
NO3	0.29	9	84	0.7	210	496	-0.18	17	32	-0.49	199	243	85.5	Well Explained
Cl	0.85	75	720	0.01	0	0	-0.31	51	94	0.27	59	73	88.7	Well Explained
HCO3	0.8	66	640	0.26	28	67	0.32	57	104	-0.1	8	9	82	Well Explained
F	0.7	51	495	0.08	3	7	0.28	43	79	-0.52	225	274	85.5	Well Explained
HARD	0.91	85	824	-0.35	53	125	0.08	3	6	0.14	16	19	97.4	Very Well Explained
TDS	0.95	93	898	0.08	3	7	-0.13	10	18	0.13	13	16	93.9	Very Well Explained
Na	0.76	60	580	0.52	114	268	-0.25	34	62	-0.06	3	3	91.3	Very Well Explained
K	0.84	72	697	-0.16	11	26	0.12	7	14	-0.24	49	60	79.7	Well Explained
Ca	0.89	83	796	-0.39	66	155	0.06	2	4	0.09	7	8	96.3	Very Well

														Explained
Mg	0.91	86	833	-0.18	14	34	0.11	7	12	0.3	74	91	97	Very Well Explained
Fe	-0.06	0	3	-0.26	29	67	0.37	76	138	0.17	23	28	23.6	Poorly Explained
Mn	0.02	0	1	0.31	41	96	0.78	329	604	0.28	66	81	78.2	Well Explained
ANIONS_SUM	0.96	96	925	0.18	14	33	-0.07	3	5	0.1	8	10	97.3	Very Well Explained
CATIONS_SUM	0.99	101	974	0.08	3	6	-0.08	4	7	0.07	4	5	99.2	Very Well Explained

Table 4.12 – Absolute (CTA) and Relative (CTRX1000) Contributions from Variables to build Principal Components

4.5.4. Absolute (CTA) and Relative (CTRX1000) Contributions from Individuals to build the Four (04) Principal Components

In Individual space it is possible to reconstruct or express all Individuals i.e. BRANDS along a few reduced number of Principal Axes. Of course, all PAs extracted from the correlation matrix could explain each individual 100% but incorporating all the dimensions or expressing the individuals with the respect to all coordinates in Principal Axes space is meaningless hence prohibiting. The main aim to avoid the curse of dimensionality would be defeted then. Hence here comes the question of trade off. How many number of Principal Axes to be retained to explain the original data set under study. It has already been confirmed that only four dimensions would be sufficient to cover the 83.062% of the total original inertia. Hence along this four Principal Axes, the individuals need to be projected. The each individuals now having four coordinates along the Principal Axes would be plotted. Other way round it would be wise to study how much contribution has been dedicated to build the Principal Components.

It has also to be computed that through these four PAs how much (%) of inertia of any individual has been covered. In the Table 4.13 the absolute and relative contributions from the individuals have been explained. Through running PCA in individual space for all individuals or observations we have calculated CTA and CTR i.e. the relative contributions for all those 51 individuals in building the PCs are tabulated below Table 4.13. In the Annexure 5 (Table A5) we have also tabulated the overall Brandwise calculations to depict the contributions along the first four Principal Axes. To acquire a quicker and relatively simpler visualization and understanding we have tabulated this Brandwise outputs along the Principal Axes to get a relatively shorter version of this table easily comprehensible.

At this particular point in time , it is noted with care that PCA outputs or values in the Table 4.13 and Table A 5 (Annexure 5) apprently seemed to be slightly different (in terms of number and values) but that does not effect the overall interpretation or scenario of the water under study trying to understand through PCA, FA or CA as a whole. Specifcally during the Factor Analysis, in Section 4.6, we have reported all the outputs soley from BRANDWISE calculations for convenience. We have in fact find an opportunity to present both outputs 12 Brandwise and 51 Individual/Observationwise in tabulated forms as well as in graphical form to acquire more improved intuitive knowledge, visualization for easy interpretation. Our ultimate aim is to understand the quality of the BRANDs

commercialized in the market. Individuals or observations, however big in numbers (51) are utilized here to see the very finer detailed picture of the data set through numbers and values. And we have achieved these goals successfully. Hence, throughout the rest of this report we will present both the outputs side by side especially in visualization and graphical representation.

For interpretation purposes we have defined several qualifiers merely from some intuitive idea subjective indeed. If % of Inertia for any individual explained by the four PCs is $\geq$ 80%, then this individual is qualified as "Very Well Explained". Similary if the % of Inertia falls between 50-80% then the individual/Brand is rated as "Well Explained". If it falls within 40-50% the Brand is qualified as "Fairly Explained". When the % Inertia goes $<$ 40%, the Brand/Individual is qualified as "Poorly Explained". As it is evident that most of the individuals/Brands have been explained by these four PC/Factors except a few individuals from BRAND03, BRAND04 and BRAND09.

Individual	CP1	CTA	CTR	CP2	CTA	CTR	CP3	CTA	CTR	CP4	CTA	CTR	% Inertia Explained by First Four PCs	Remarks
DIWa	-4.02	33	831	0.07	0	0	-0.99	10	50	1.28	26	85	96.6	Very Well Explained
B1	2.19	10	61	2.7	60	93	7.39	584	693	2.39	91	72	91.9	Very Well Explained
B2a	1.27	3	71	3.85	123	655	-1.5	24	99	-1.75	49	135	96	Very Well Explained
B2b	0.99	2	78	3.01	75	727	-1.15	14	106	-0.79	10	50	96.1	Very Well Explained
B2c	1.31	3	69	4.2	146	706	-1.43	22	82	-1.53	37	94	95.1	Very Well Explained
B2d	1.03	2	75	3.11	80	688	-1.27	17	115	-0.53	5	20	89.8	Very Well Explained
B2e	1.45	4	70	4.3	153	621	0.87	8	26	-1.35	29	61	77.8	Well Explained
B2f	1.23	3	122	2.28	43	422	-1.02	11	84	-0.7	8	40	66.8	Well Explained
B3a	0.24	0	9	0.23	0	9	1.21	16	238	-0.43	3	30	28.6	Poorly Explained
B3b	0.19	0	16	-0.4	2	80	0.54	3	122	-0.51	4	110	32.8	Poorly Explained
B3c	0.26	0	19	-0.5	2	61	0.76	6	169	-0.53	5	83	33.2	Poorly Explained
B3d	-0.01	0	0	-1	8	222	0.82	7	155	-0.07	0	1	37.8	Poorly Explained
B3e	-0.04	0	1	-1	9	366	0.54	3	103	-0.38	2	50	52	Fairly Explained
B3f	-0.02	0	0	-1	9	373	0.53	3	99	-0.35	2	42	51.4	Fairly Explained
B4a	-0.73	1	117	0.26	1	16	0.68	5	101	-0.12	0	3	23.7	Poorly Explained
B4b	-0.87	2	97	0.51	2	34	0.21	0	6	-0.05	0	0	13.7	Poorly Explained
B5a	-1.93	8	314	-0.1	0	0	1.06	12	94	-0.04	0	0	40.8	Fairly Explained

Individual	CP1	CTA	CTR	CP2	CTA	CTR	CP3	CTA	CTR	CP4	CTA	CTR	% Inertia Explained by First Four PCs	Remarks
B5b	-2.28	11	254	0.02	0	0	2.44	64	293	0.06	0	0	54.7	Well Explained
B5c	-2.29	11	804	-0.2	0	7	-0.12	0	2	0.08	0	1	81.4	Very Well Explained
B5d	-1.81	7	644	-0.5	2	52	0.22	1	9	-0.86	12	146	85.1	Very Well Explained
B5e	-2.23	10	339	0.31	1	7	1.94	40	257	-0.43	3	13	61.6	Well Explained
B6a	-0.65	1	66	-1.7	24	456	0.4	2	25	-1.16	22	210	75.7	Well Explained
B6b	-0.57	1	43	-1.5	19	307	0.16	0	3	-1.4	31	257	61	Well Explained
B6c	-0.69	1	59	-1.9	29	435	0.32	1	13	-1.28	26	202	70.9	Well Explained
B6d	-0.73	1	68	-2	32	485	0.05	0	0	-1.31	27	215	76.8	Well Explained
B6e	-0.66	1	52	-2.1	38	545	0.14	0	3	-1.27	26	195	79.5	Well Explained
B6f	-0.58	1	40	-1.6	22	312	0.11	0	1	-1.39	31	230	58.3	Well Explained
B7a	-3.32	22	710	-0.5	2	13	-0.76	6	37	0.38	2	9	76.9	Well Explained
B7b	-3.64	27	786	-0.6	3	20	-0.56	3	18	0.82	11	40	86.4	Very Well Explained
B7c	-3.85	30	890	-0	0	0	-0.55	3	18	1.12	20	76	98.4	Very Well Explained
B7d	-3.96	32	775	0.68	4	23	-0.95	10	45	1.35	29	91	93.4	Very Well Explained
B7e	-3.87	30	895	-0.1	0	1	-0.55	3	18	1.09	19	71	98.5	Very Well Explained
B8a	5.18	55	689	-0.6	3	9	-0.16	0	1	2.04	67	107	80.6	Very Well Explained
B8b	5.55	63	805	-0.1	0	0	-1.5	24	59	1.75	49	80	94.4	Very Well Explained
B8c	6	73	777	0.11	0	0	-1.61	28	56	2.09	70	94	92.7	Very Well Explained
B8d	4.89	49	830	-0.7	4	15	-1.03	11	37	1.15	21	46	92.8	Very Well Explained
B8e	5.22	55	793	-0.7	4	13	-0.86	8	22	1.59	40	73	90.1	Very Well Explained
B8f	4.83	47	846	-0.7	5	20	-0.71	5	18	1.44	33	76	96	Very Well Explained
B9a	-0.32	0	19	0.18	0	6	0.09	0	1	-0.97	15	177	20.3	Poorly Explained
B9b	-0.93	2	209	-0	0	0	0.25	1	15	-0.75	9	135	35.9	Poorly Explained
B9c	-0.97	2	232	-0.2	0	11	0.28	1	19	-0.85	11	175	43.7	Fairly Explained

Individual	CP1	CTA	CTR	CP2	CTA	CTR	CP3	CTA	CTR	CP4	CTA	CTR	% Inertia Explained by First Four PCs	Remarks
B10a	-3.61	27	902	-0.4	1	12	-0.69	5	33	0.72	8	36	98.3	Very Well Explained
B10b	-3.55	26	733	-0.5	2	14	-0.95	10	52	0.37	2	8	80.7	Very Well Explained
B10c	-3.54	26	748	-0.2	0	1	-0.94	9	53	0.48	4	14	81.6	Very Well Explained
B10d	-3.6	26	874	-0.6	3	21	-0.44	2	13	0.74	9	37	94.5	Very Well Explained
B10e	-3.6	26	824	-0.8	5	38	-0.62	4	24	0.64	7	26	91.2	Very Well Explained
B11a	5.33	58	829	-1.1	11	38	-0.01	0	0	-0.74	9	16	88.3	Very Well Explained
B11b	5.35	58	781	-1.8	27	87	0.11	0	0	-1.15	21	36	90.4	Very Well Explained
B11c	5.36	59	827	-1.3	13	47	-0.01	0	0	-0.59	6	10	88.4	Very Well Explained
B11d	5.25	56	833	-1.2	12	43	0.24	1	2	-0.57	5	10	88.8	Very Well Explained
DIWb	-4.23	36	549	1.61	21	79	-0.99	10	30	2.23	80	152	81	Very Well Explained

Table 4.13 – Absolute (CTA) and Relative (CTRX1000) Contributions from Individuals to build Principal Components

- Principal Component 1: PC1

Observing the Figure 4. 15 we may clearly state that the First Principal Component/Axis PC1 attracted most Individuals from seven (07) Brands out of 12 Brands. And their relative contributions in building the Principal Axes have shown that they are mostly crowded along and around PC1 which is essentially explaining the "Material Loading" and or "total dissolved substances". It is not unreasonable to comprehend that this aggregate properties of the individuals are very well explained by PC1.

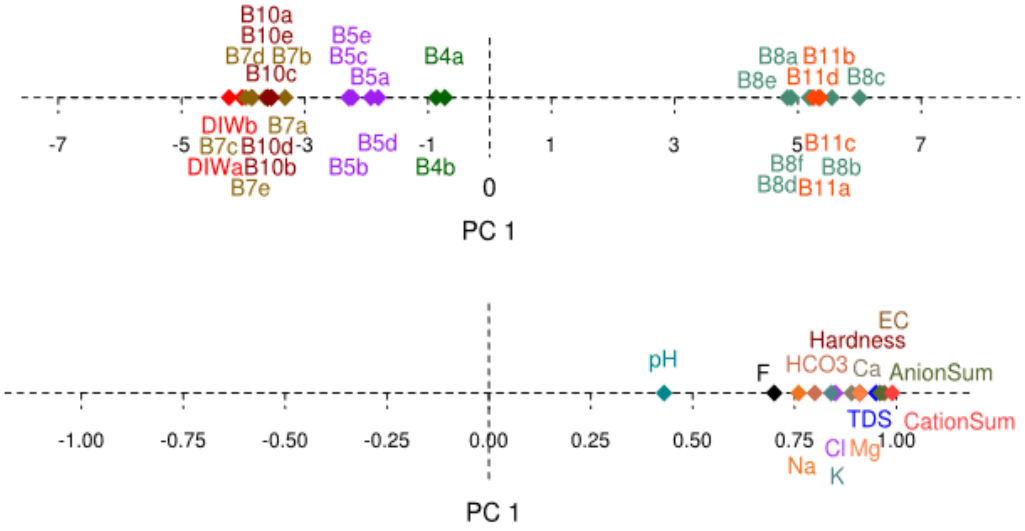

Figure 4.15 – Projection of Individuals and Variables along PC1

- Principal Component 2: PC2

From Figure 4.16 we may clearly observe that BRAND02 and BRAND06 are mostly attracted along this second PC2 which has been loaded mostly with ammonium (NH4), nitrate (NO3). This axis also captured slightly sodium Na a cation heavily detected in the ground water of Bangladesh as the country is indeed situated in a delta region at the Ganges-Bhramhaputra-Yamuna basin having it's a quite long coastline with Bay of Bengal as well as the Indean Ocean a huge natural acqua marine source of sodium chloride NaCl salt gets easily into the underground acquifers. It has also been explained above that this PC2 may indicate the presense of ammonium (NH4) and nitrate –nitrogen (NO3) occuring only due some unusual and unexpected reasons. Of course, there is a danger in making such a comment because in a natural water matrix it is very unlikely to have such compounds. But if there is any contamination, say, due to agricultural practices: application of fertilizers, pesticides etc. in the vicinity of the water source from where the manufacturers collected the raw water prior to treatment and bottling this is not impossible. Again it may be noted that this axix PC2 covers only 13.13 % of the total inertial of the original data. Hence this inference may not fully be reflecting the reality. There is a reason to have a doubt in drawing such easy conclusion.

But from futher exploration it would be clear that BRAND02 are more close to BRAND01 in respect to their nature and they are also lying far away from the general cluster formed by the others. But BRAND01 and BRAND 02 have almost same amount along PC1 (they are in very close proximity with respect to PC1), that means, they are also having similar amount of substances loaded or "materials loaded". So at this stage it is not unreasonable that PC2 mainly explains that between BRAND02 (with

individulas having +positive values) and BRAND06 (individuals attaining negative values) are opposing away each other along this axis (Figure 4.16) . But from our later FA and CA we will see that these two brands are belonging to two different clusters. Wheather they are good or bad in terms of quality or other physico-chemical attibutes that could not be stated in a broad ranging general term at this point in time.

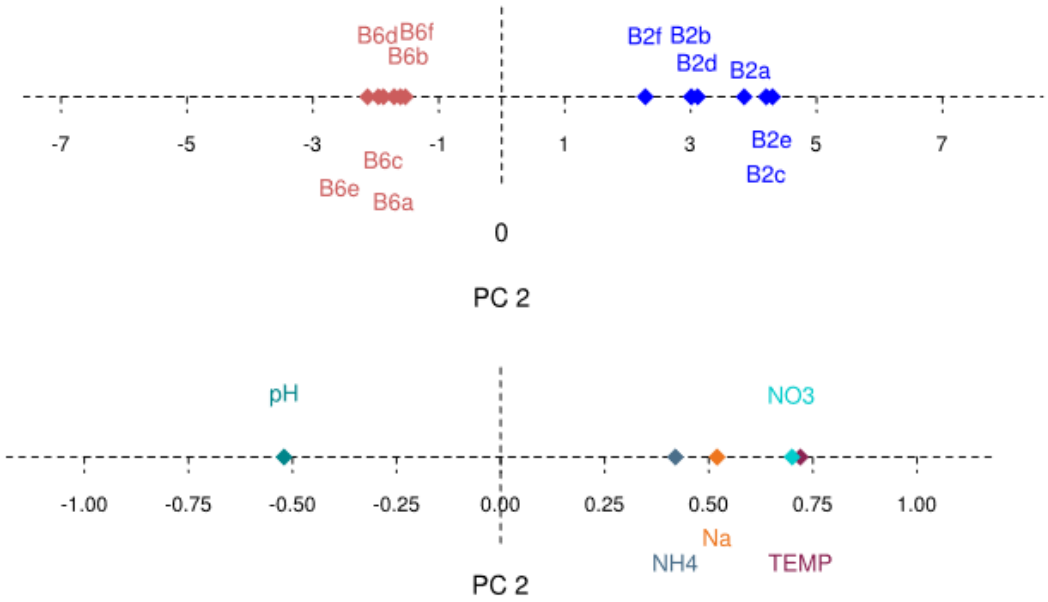

Figure 4.16 – Projection of Individuals and Variables along PC2

- Principal Component 3: PC3

Individual B1 belong to BRAND01 (having coordinate along PC3 (having coordinate +7.39) is a loan fighter brought out the new dimesion PC3 almost single handedly along with some significant contributions from individuals belong to Brands BRAND03 (Figure 4.17). Individuals from BRAND08 also contributed in building up this component although not so significant. Only two brands: B1 and B3 (Figure 4.17) have been attracted and explained very well along PC3 which is mostly built from ions: dissolved Iron (Fe), Dissolced Manganese (Mn) positively associated with carbonate ions (HCO3) most probably to form carbonate salts, typically present in any ground water resources. As we have assumed that iron (Fe) and manganese (Mn) in through formation of carbonate (HCO3) salt may reflect the *"Aggregate Qualitative Feature"* of the water as a whole but again we have not forgetten that this PC3 is only capturing the information around 10% of the original data set. Being the third Principal Component it is not completely meaningless to state that this PC3 featuring mainly iron (Fe) and manganease (Mn) jointly. In water chemistry these two variables are always studied in coupled as they provide the same qualitative aspect of water jointly as they decide the taste of the water as well as give information in relation to underground acquifers. The variable ammonium

(NH4) reflected slightly also along this third component PC3 which shows the presence of some unacceptable materials in drinking water. But we may recall that NH4 along with NO3 has been reflected in a greater degree along the PC2. From the reality we know that ammonium NH4 and NO3 may appear in the water if there is any contamination due to, say, agricultural practices including application of nitrogen based fertilizer or pesticides as well as due to other anthropogenic activities including fecal discharge or urination in the near vicinity of the water source from where the manufacture extracted the raw water prior to treatment for packaging.

As it has been explained in the above section 5.4 that PC3 is mostly associated with the qualitatively (not quantitatively) may represent overall "Aggregare Qualitative" feature associated with iron (Fe) and manages (Mn) and their carbonate. Whereas BRAND01 is staying almost at the extreme right towards the positive direftion BRAND03 stayig far below almost at the origin n this respect but both of them are withing positive range meaning that they may belong to an acceptable quality group with respect to this Fe-Mn-HCO3 created aggragate nature. Again we do not forget that we are speaking of PC3 which exlplain only around 10% of the total inertia which always leave us with some weak basis.

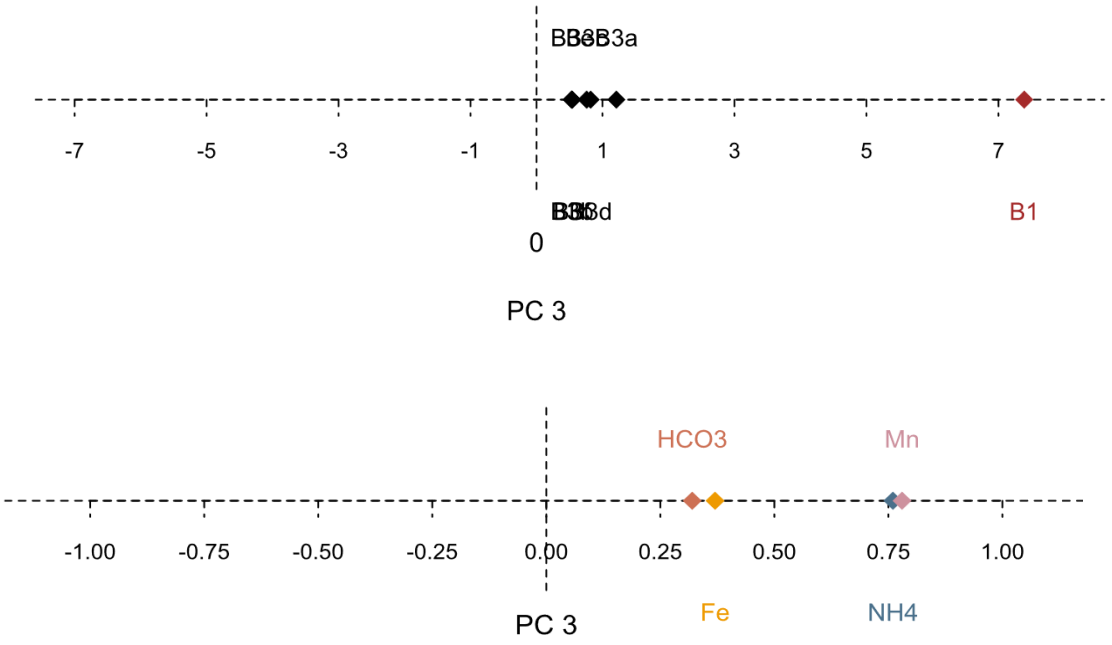

Figure 4.17 – Projection of Individuals i.e. Brands and Variables along PC3

- Principal Component 4: PC4

This very PC4 built up mostly by variables: NO3 and F probably associated with iron and manganease as we have mentioned in the section 4.5.2 may explain the overall "Aesthetic Acceptabiity" triggred due to again iron and manganease atoms formed in nitrate composition probably originated from resedues of agricultural chemicals to define its hardness pattern, taste, odor and appearance. Figure

4.18 shows that whereas BRAND06 and BRAND09 forming a cluster lying towards the negative side of the PC4 not far away from the origin, both BRAND08 and DEIONIZEDWATER residing toward the positive direction forming another cluster keeping themselves away from BRAND06 and BRAND09. As we know that DEIONIZEDWATER is fully devoide of substances and serving an extreme example of mineral free water or "ultralow mineral content water" we may imagine that BRAND06 and BRAND09 contains some minerals but they may conyain those unacceptable NO3 and F which are also lying tward the same negative portion of the PC4 (Figure 4.18). These two distinct groups appeared here will also remain in the two separate clusters which we will observe by the end of the Cluster Analysis process. Therefore, we have some reason to state that due to presence of nitrate bounded iron these two BRAND06 and BRAND09 kept themselves little away from demineralized type water. Note that these two brands still belong to low mineral content type water if not fully devoide of mineral to become ultimatley some "ultralow" or "too soft or "too light" water like DEMINERALIZEDWATER (DIW).

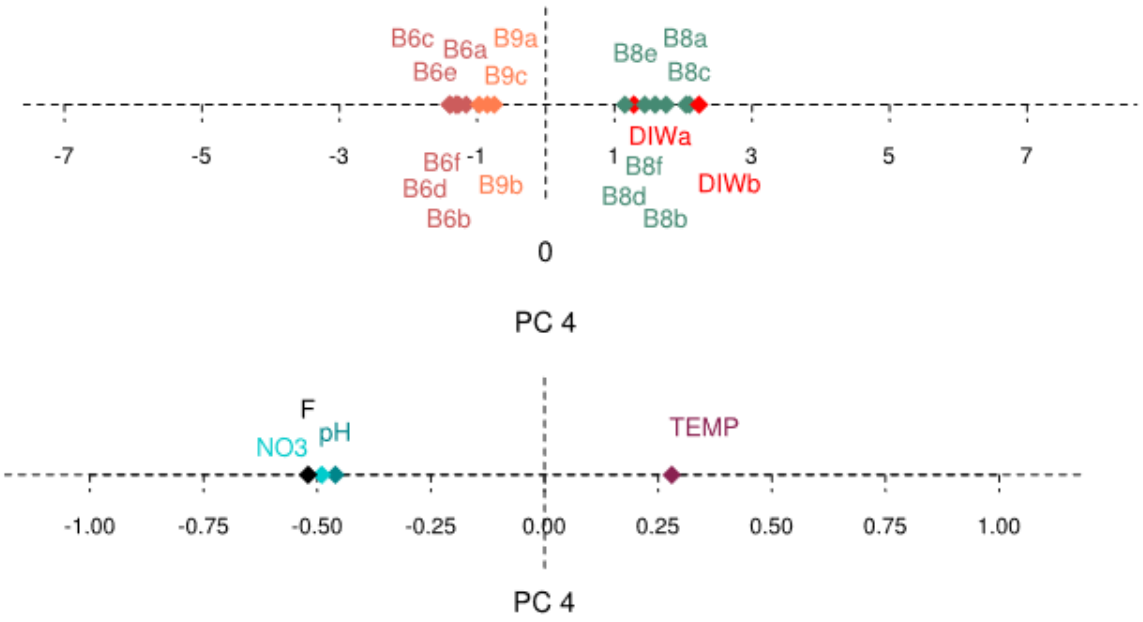

Figure 4.18 – Projection of Individuals i.e. Brands and Variables along PC4

4.5.5. Possible Explanation of Some Extreme Behavior of some individuals

The Euclidean distances have been estimated considering all PCs for all 51 individuals (j) from the center of the Cloud (at the origin of the Principal Plan) following the relation below:

$$d_j = \sqrt{\sum_{\nu} (Y_j^{\nu})^2}$$

Where, d_j = The Euclidean Distance of j-th Individual/BRAND from the origin of the Principal Axes

ν = ν-th Principal Axis

Y_j^{ν} = ν-th Principal Component/Coordinates along ν-th the Principal Axis j-th Individual/BRAND

The Euclidean Distances of all 51 individuals are tabulated in the Annexure 6 (Table A6) where one may see the very detailed nature of the in dividuals and how they have been dispersed around the center of the Cloud originated at the origin of all the 18 Principal Axes. The following spider wab like graphical representation (Figure 4.19) also provide a better undersranding about the distribution of the individual with respect to Euclidean Distance.

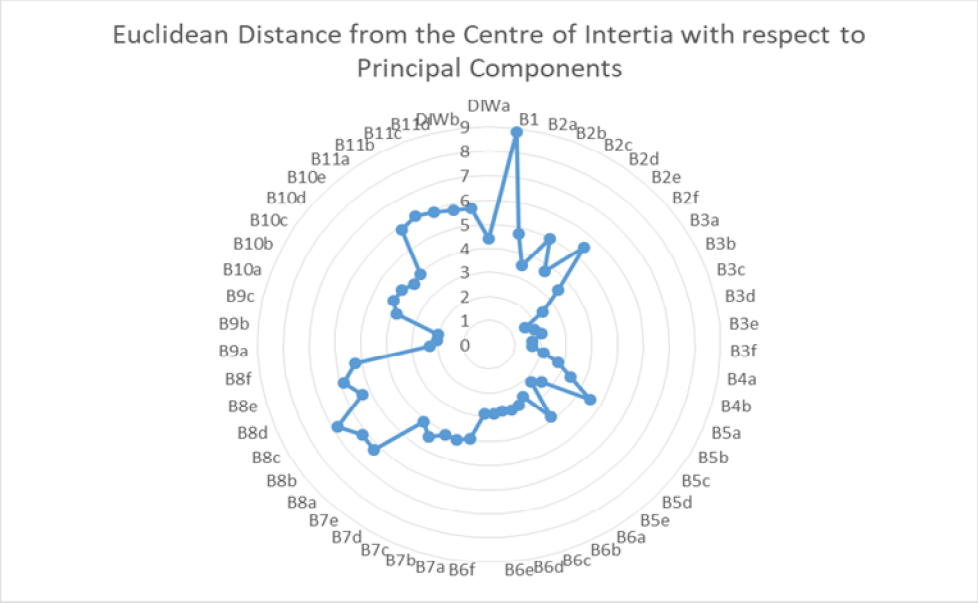

Figure 4.19 – Euclidean Distance of Individuals from the Center of the Cloud in Principal Components Space

To understand quickly as well as to obtain the Brandwise scenario as a whole we have estimated Euclidean Distances for 11 Brands plus one DEIONIZEDWATER and this has been tabulated and interpreted below (Table 4.14).

If the origin (0,0,0,0) of this new 4-D coordinate system with four PCs has been shifted to the centre of the gravity then it is good to assume that h=0. In other words if the origin has been shifted to "the center of the rigid rotating coordinate system" or at the centre of the Cloud formed by the individuals around this centre then it is mathematically can be considered that the Euclidean Distance from the Center (0,0,0,0) of the Cloud is nothing but the square root of the sum of the squares of all the coordinates (PCs) of the Individual (j).

BRAND	**EUCLIDEAN DISTANCE**
BRAND01	5.7
BRAND02	4.52
BRAND03	1.66
BRAND04	4.2
BRAND05	6.29
BRAND06	2.72
BRAND07	3.03
BRAND08	1.94
BRAND09	3.93
BRAND10	3.86
BRAND11	6.16
DEIONIZEDWATER	6.21

Table 4.14 –Euclidean Distances of BRANDS from the Center of the Cloud

The estimated Euclidean Distances of the individuals are tabulate in the Table 4.14 above. The interesting results came out to show that the four (04) individuals: BRAND05 (d_5=6.29), DEIONIZEDWATER (d_{12} = 6.21), BRAND11 (d_{11} =6.16), BRAND01 (d_1=5.7) and BRAND 02 (d_2=4.52) are relatively the most distant individuals in comparison with the other individulas staying very closely around the centre of the Cloud (rigid body origin of the PC Coordinate Space). These five individuals at distances 6.29, 6.21, 6.16, 5.7, 4.52 respectively are staying at the periphery of the cloud depicting that they are very different by natur from the rest seven individuals. This is indeed explaining the fact that these five individuals particularly either containing very high amount of dissolved substances, ions (anios and or cations) (BRAND05 and BRAND11) or they are not have any mineral content at all (DIONIZED WATER) or having other very different quality features (BRAND01, BRAND02) which dragged them out from the rest of the seven individuals.

Moreover, this interpretation or assumption has indeed a very sound basis because it has clearly been visible in the Table 4.14.

4.5.6. Principal Component Maps & Possible Interpretation

4.5.6.1. Variables Projected on The Principal Planes

Below the projection of the Variables have been depicted on Principal Planes (PC1&PC, PC2&PC3, PC1&PC4). If this maps are closely observed together with correlation matrix, the possible features and phenomena could be easily extracted. The interpretation for almost all vectors could reasonably be outlined. In this section the main and important vectors are discussed.

As water chemistry supports it has been clearly visible that angle between EC, TDS and angle between ANIONS_SUM, CATIONS_SUM and "angles" among all these four axes are "so small ~0" that it is clearly evident that they are strongly correlated. From the chemistry, as already discussed in the previous sections, it is very likely that the buildup of both anions and cations are due to the dissolved materials and dissociated or ionized due to the electronic charge distribution and their inherent nature impacted by the water molecule H_2O. Water molecule, having slight electric dipole moment arsing due to its assymetric electronic charge distribution at the outermost shells, tries to dissociate all the dissolved materials to ions (cations and anions). This phenomenon is ultimately contributing to have the water solution to be electrically active, although pure water molecule is electrically inactive in a sense that it does not conduct well the electricity. That means, when the materials are dissolved in the water then the electrical conductivity EC is gradually building up. Or other way round the EC is increased due to the presence of dissolved materials. That is why the EC, TDS, ANIONS_SUM and CATIONS_SUM are showing so strong correlations and in the vector plot they are almost along the same direction with almost same magnitude and could not be seen very separate from each other (Figure 4.20).

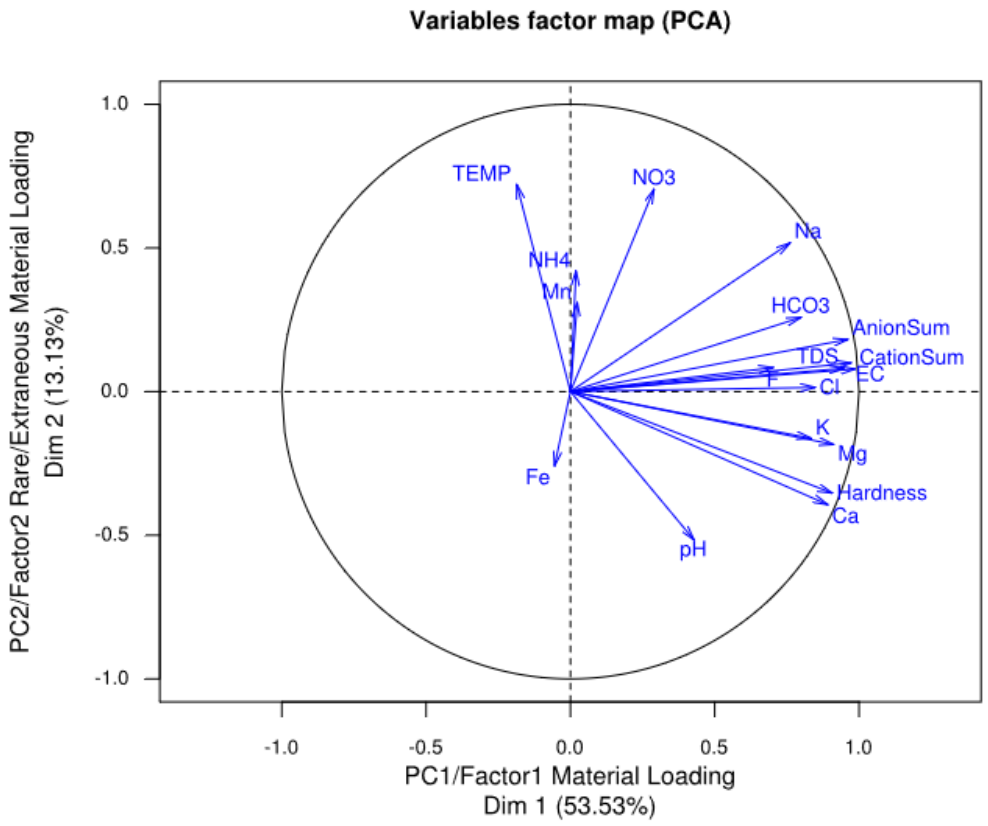

Figure 4.20 –Variables on Principal Axes 1 & 2

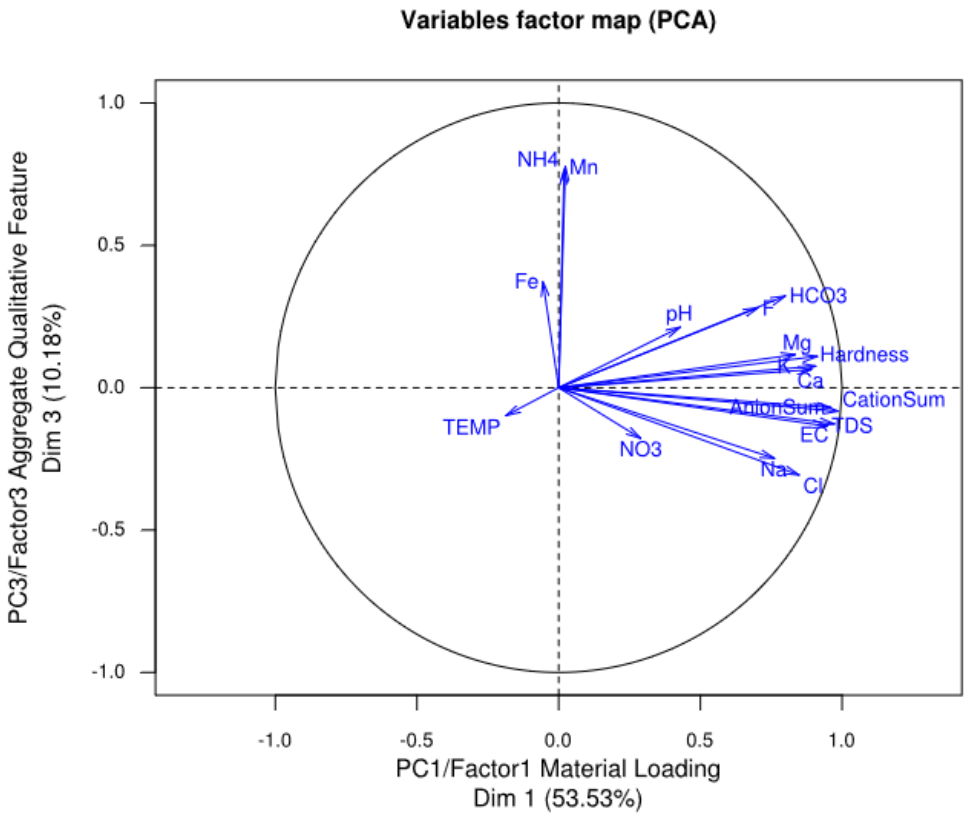

Figure 4.21 –Variables on Principal Axes 1 & 3

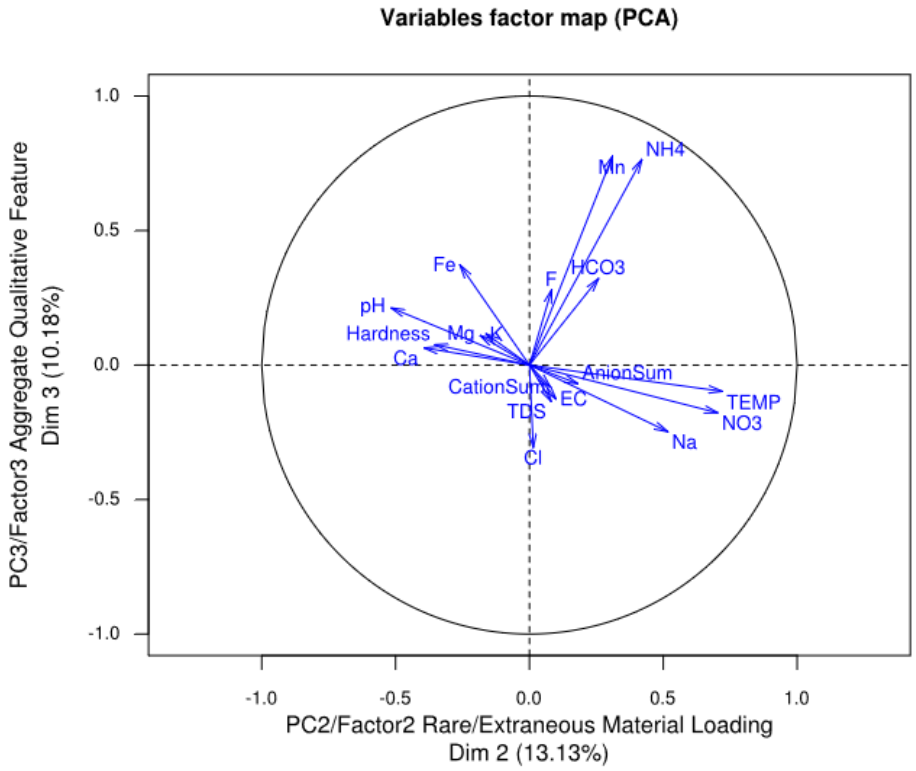

Figure 4.22 –Variables on Principal Axes 2 & 3

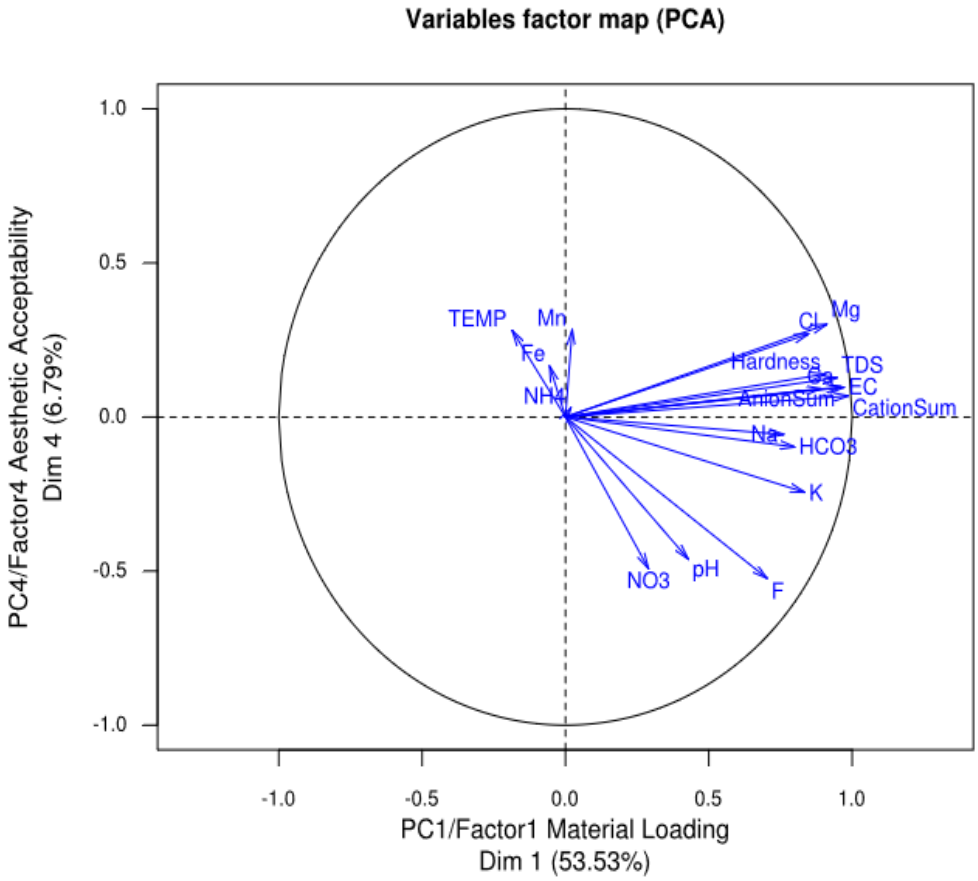

Figure 4.23 –Variables on Principal Axes 1 & 4

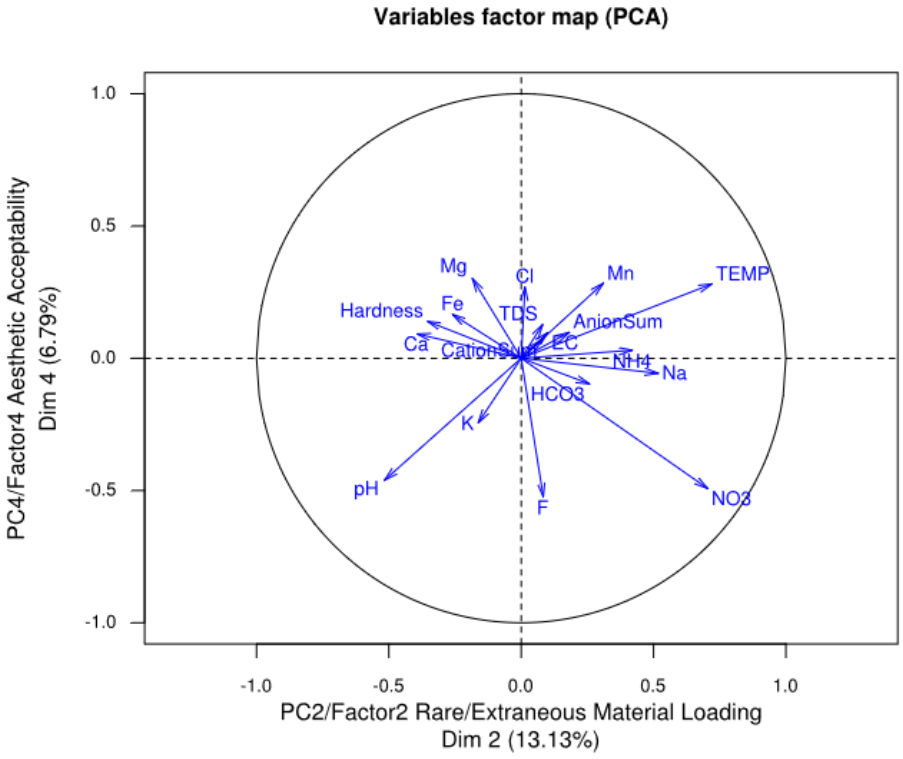

Figure 4.24 –Variables on Principal Axes 2 & 4

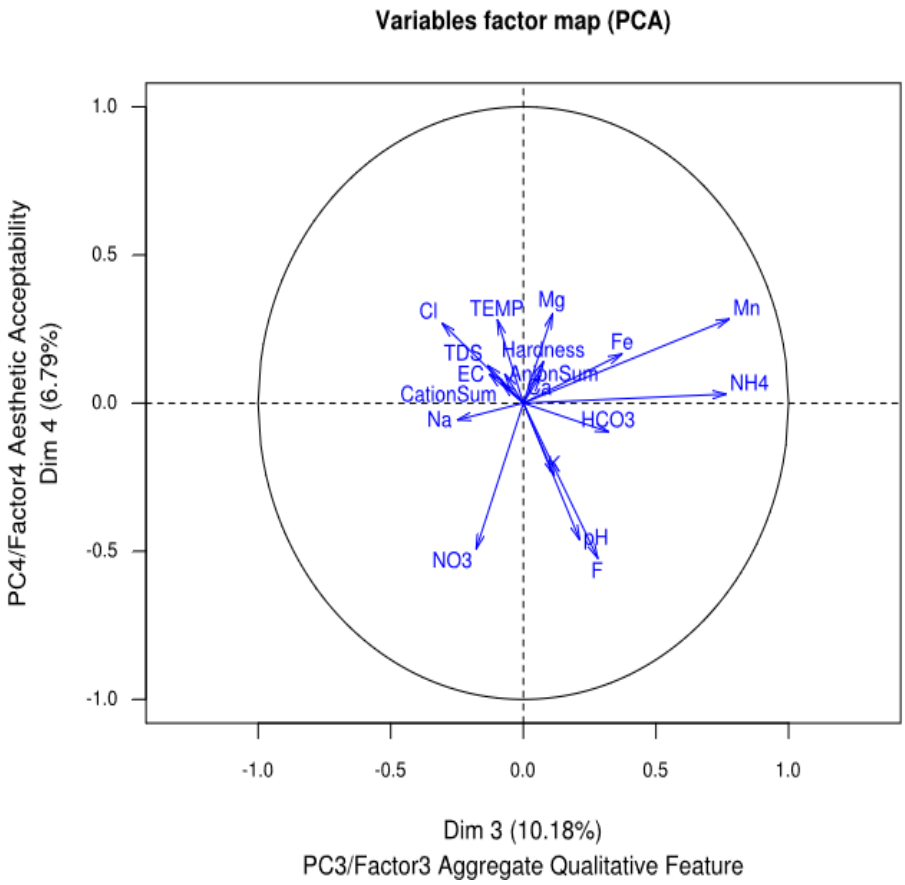

Figure 4.25 –Variables on Principal Axes 3 & 4

Whereas not fully visible by the naked eyes, the relation between pH and TEMP has been observed after a rotation and from the projection on Principal Axes 1 and 4 (Figure 4.23) which is also seemed to acceptable from the chemistry point of view. These two variables are almost oppositely directed, means having moderately negative (-ive) correlation (r=-0.63).

Further from the projection map on the Principal Axes 1 & 2 (Figure 4.20) it is also very much clear that majority of the variable vectors are towards the right side, are in close proximity, having relatively small angles among them as such building the First Principal Axis PC1/Factor1 which is explaining >54% of the total variability.

From the Plane 1&2 , Ca, Mg, Na, HARD are also very close to HCO3 which also explains the fact that the HCO3 alkilinity or hardness are builtup by these major ions, as such confirming the chemistry already known generally. Of course, they are also staying very close to ANIONS_SUM and CATIONS_SUM which is easily understandable that these atoms are also contributing to inonization activity. More the materials or molecules, higher the ions sums (cations and or anions).

From both the maps or projections it is visible that the IONS_SUMS and CATIONS_SUM are having vectors along the same direction with almost zero angle, means they are strongly correlated. This is again confirming the chemistry and physics that in a solution, if the pH is near to 7.0 (neutral) then both the number of anions and cations must be same. From this study this has again been confirmed that they are almost along the same line aligned towards the First Principal Axis PA1 indictaing the total "*Materials Load*"

Fe and NO2 are moderately correlated but is opposite direction (r=-0.59), means they are opposing each other. But Fe is not at all correlated or aligned with NO3 (r=-0.07), means, it is very likely that there is no molecule like say, FeNO3 and the like.

But watching the angle between NH4 and F (r=0.30), it could be only one possible explanation that they have some molecular bonding, because NH4 being a positive ion only could be forming molecule with another anion or so. There is not so other ion visible to have correlation with NH4 except F. Their contribution in building TDS, EC and Hardness, Anions_SUM and CATIONS_SUM is significant, means this assumption has some basis. Alternatively higher presence of NH4 in mainly in one BRAND01 (0.24 mg/L) may be an outlier. But from other information it is known that BRAND01 is showing some extreme behaviour in terms of all variables, means, this BRAND01 may not be an outlier at all rather may be an individual different from other general folks. This has also been confirmed further that this very BRAND01 is indeed the water having rather better acceptability for human consumption in comparison with others who are not fully suitable for consumption from human health realted consideration. Howevere, from watching the projections of variables on various Principal Planes (1&2, 2&3, 1&3, 1&4) sufficient information has been gained and it has been confirmed that the majority of the atoms or molecules are contributing in building the main four PCs as well as to creat the new 4-D space to explain the water water quality data at hand under this study.

4.5.6.2. Individuals Projected on The Plane of PC1/Factor1 & PC2/factor2 Axes

The graphs (Figure 4.26 and Figure 4.27) show the Individuals Brands' relative positions on the Plane of First Two Principal Axes/Factors. As it has been defined in the earlier sections that the First Axes clearly giving indicator about the qualitative effect associated with "total load of the dissolved substances" or the *"Materials Loading"* it has been very clear that BRAND05 and BRAND11 are highly loaded with the substance or they are may be called to be as "Excessive Mineralized" products. They are in fact "Heavily Loaded with Materials (not good for human consumption as they may contain substances like Ca, Mg, Fe, Na to give birth to a higher hardness values). Some times these excessively loaded products may be asthetically not acceptable as Fe content may give bad odour, blurry look or redish iron oxide precipitation etc. Essentially they may not be harmful for human health but not suitable for consumption.

Oppositly DEIONIZEDWATER is at the extreme left end indicating that this water does not contain mineral at all. As such BRAND06, BRAND09 and BRAND10 are also could be assumed to be as *"ultralow mineral content"* products. They are essentially "Devoid of Materials" which is also not desirable in respect of associated risk to human health.This kind of water could only be useful for

industrial purposes not for drinking. If consumed by human, this type of water not having mineral may attract the minerals contained in the human tissue. Because this deionized water is a highly strong solvent and substance hungry with low pH value ~6.4 or so, try to have its pH around 7.0 to be neutral one as such it would be trying to take out the mineral from human body and kedney would be trying to make the Urine pH at its natural level. This ultimately put eccessive load/risk on kedney and urinal system of human body. Therefore, consumption of this demineralized water is discouraged.

Other brands e.g. BRAND03, BRAND04, BRAND07 and BRAND08 are also around the left "low mineral contening region" along the Principal Axis 1 (PC1/Factor1). This is also clearly showing that they are *"Low Mineral Content"* products not suitable for human consumption, alarming.

At the middle range with respect to PC1 it is reasonably clarified that BRAND01 and BRAND 02 are having a 'Moderate or Balance Mineral Content' and such probably the candidate waters to be suitable for consumption. But it still needs more information to be sure about that and treated in the later part of this section.

Axis 2 or PC2 has separated mainly the BRAND01 and BRAND02 from other folks. Although at first look it appararently appears that the BRAND01 is an outlier but from othe various check and analysis it has been confirmed that this is not the case. This BRAND01 is rather expressing some different behaviours which kept this product away from other. As earlier it has been explained that the BRAND01 has major contribution in PC2 may be explaining some other phenomena. In the later part of this Section we will see that PC4 also reasonably seperates these two brands from other brands these Brands are indeed appeared to be the best quality water suitable for human consumption having aesthetically also acceptable explained by PC4.

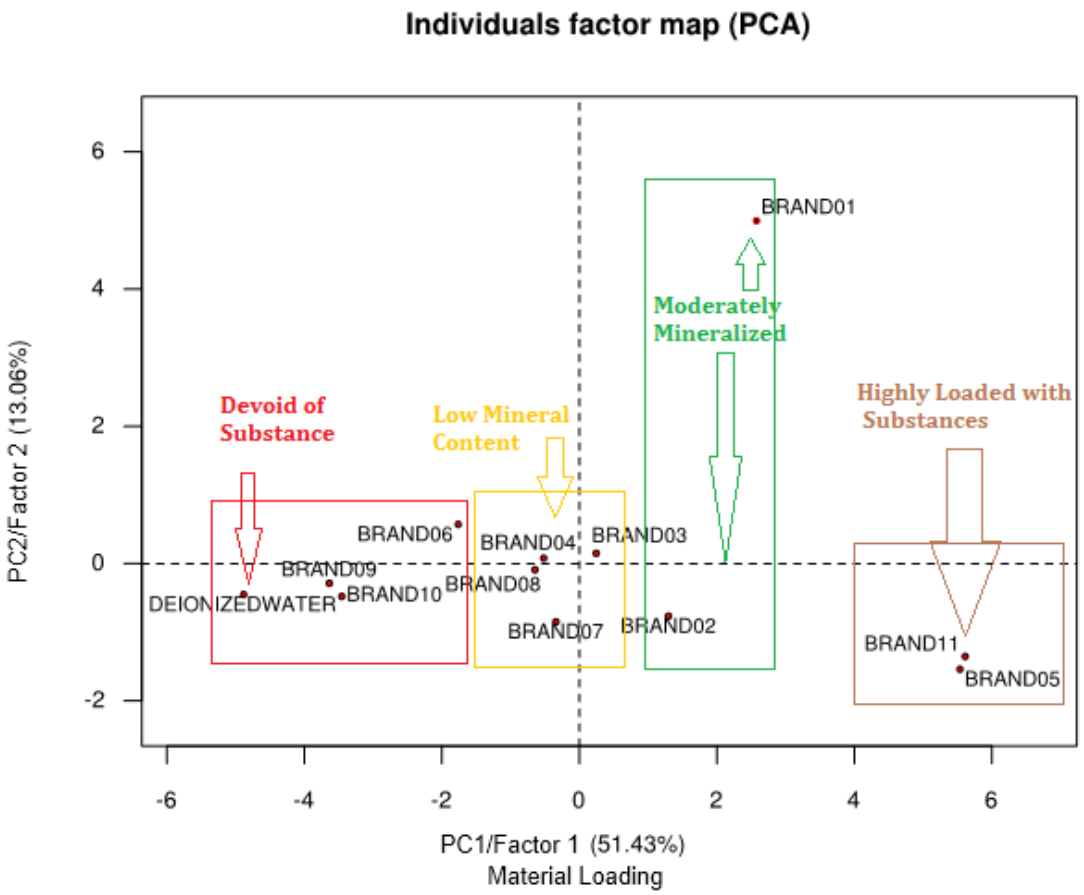

Figure 4.26 –Individuals or Brands on PC1/Factor 1

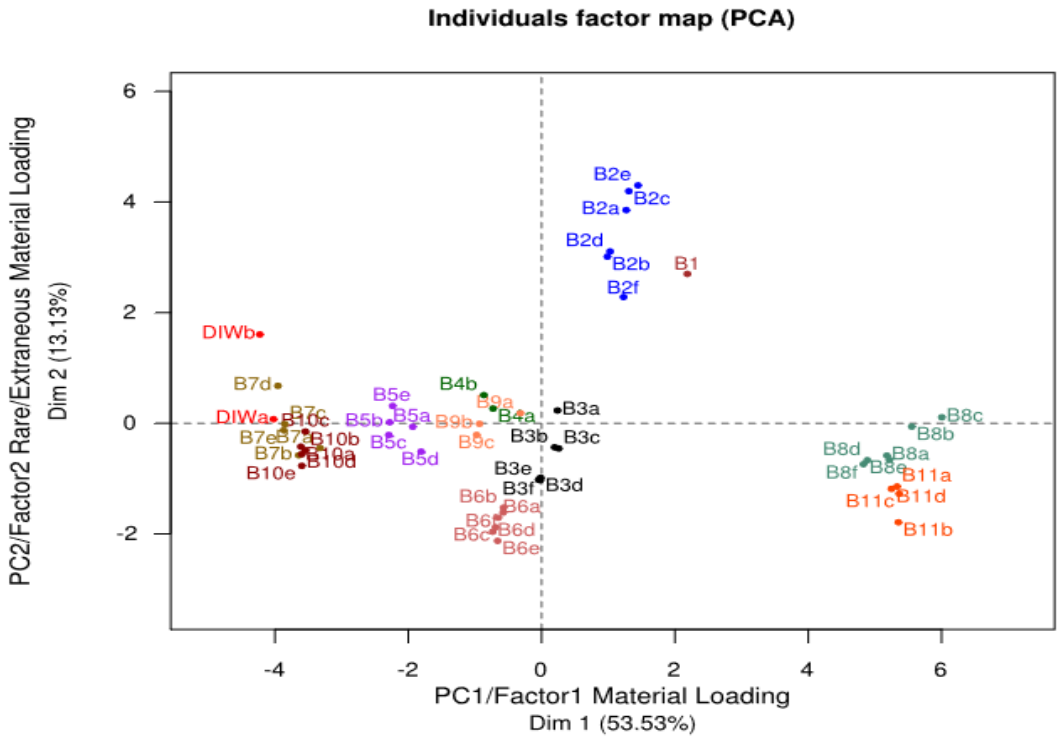

Figure 4.27 –Individuals projected on the Plane containing the First Two Principal Axes (PC1 & PC2)

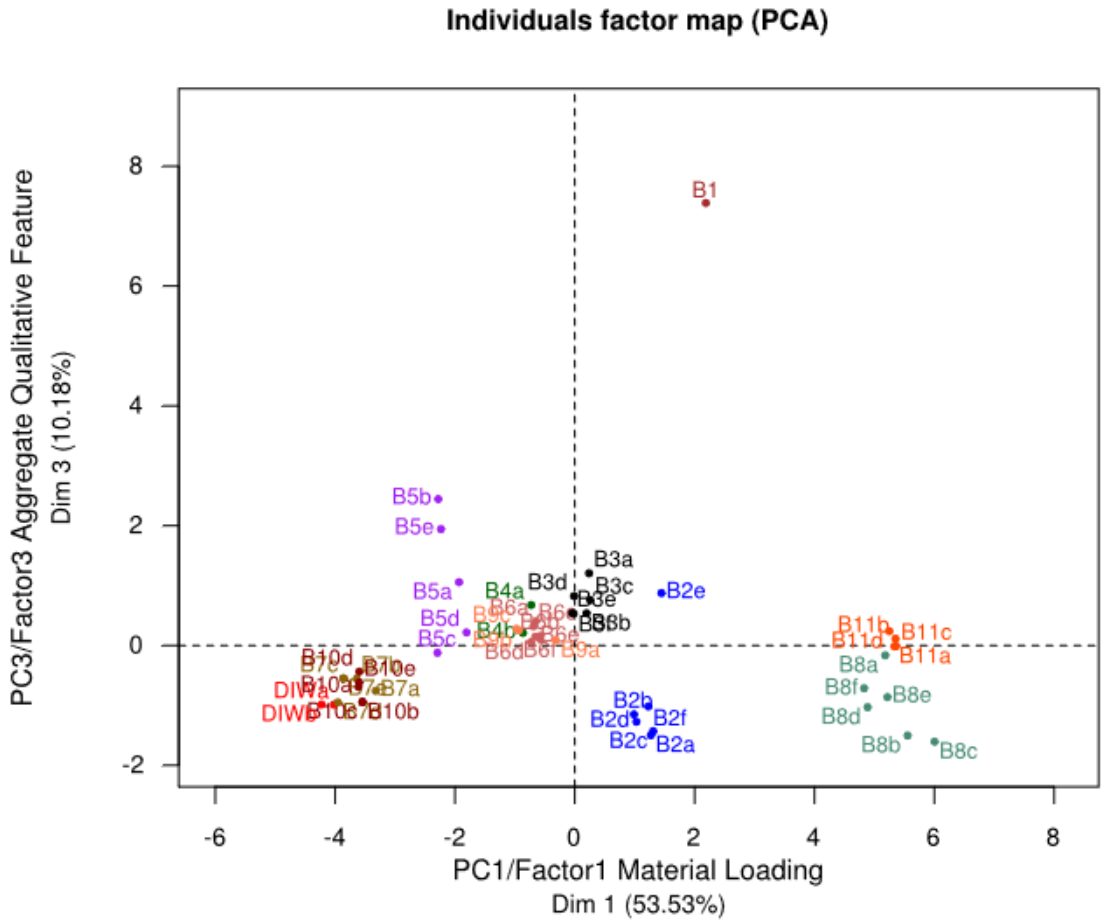

Figure 4.28 – Plot of Individuals on PC1/Factor1 vs PC3/Factor 3 Plane

From the projection map (Figure 4.29) on the PC2&PC3 Plane it has again been confirmed that anions associated with some cations are playing major role in building up this component which is also better explaining pH and TEMP. It is reiterated that anion quantity is not the main issue here but the presence and association with other cations like Na, K, NH4, Fe are playing the critical role. This is also probably the reason why only this component PC3 is explaining better the variables pH and TEMP. pH and TEMP are not fully explained by other PC/Factors. This axis soley attracted pH and TEMP, keeping in mind that this PC3 is covering only 10.18% of the total variances.

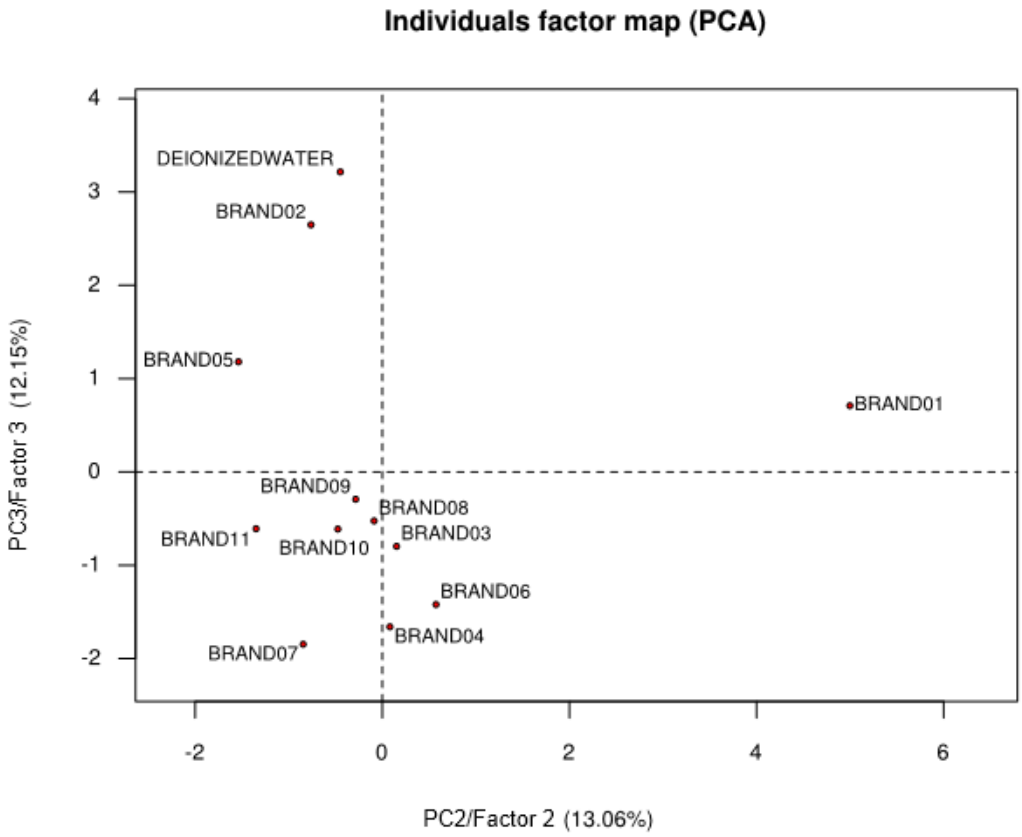

Figure 4.29 – Individuals projected on The Principal Plane PC2 & PC3

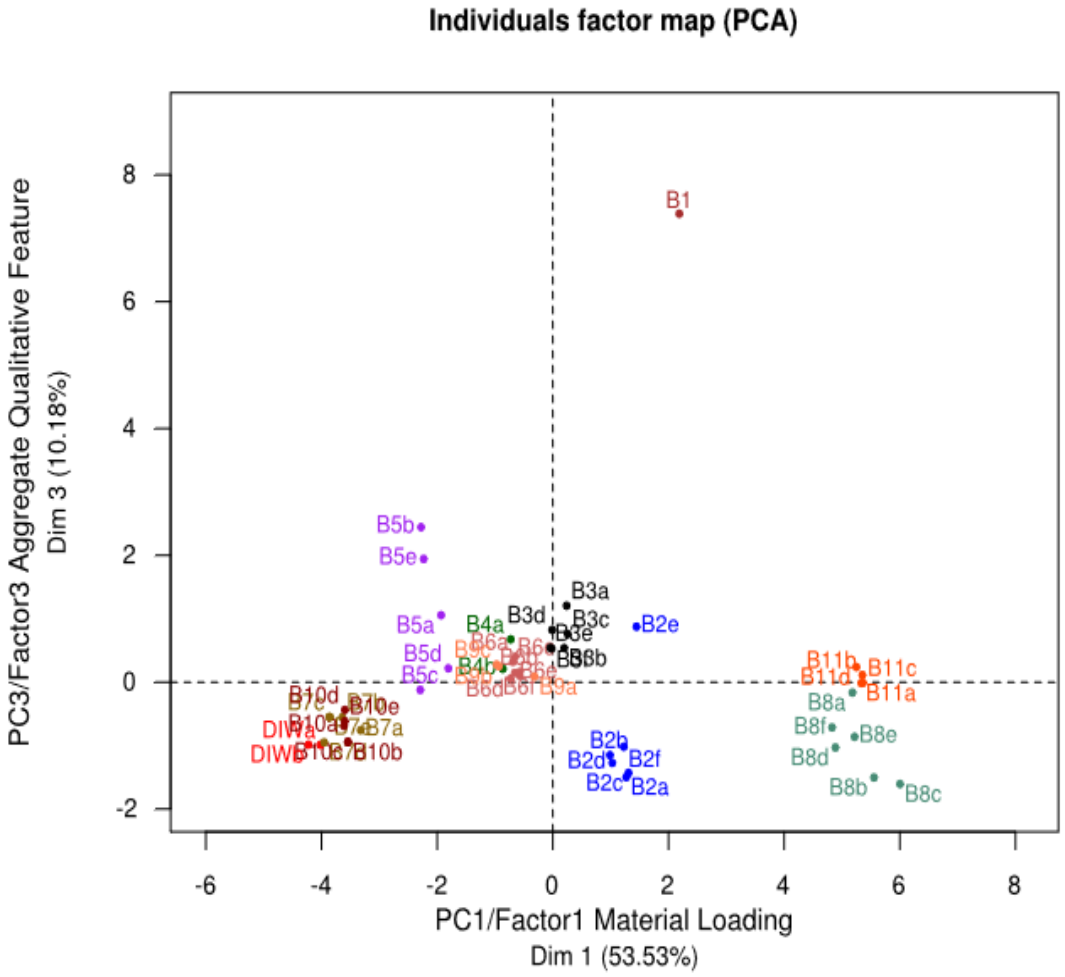

Figure 4.30 – Individuals projected on The Principal Plane PC1 & PC3

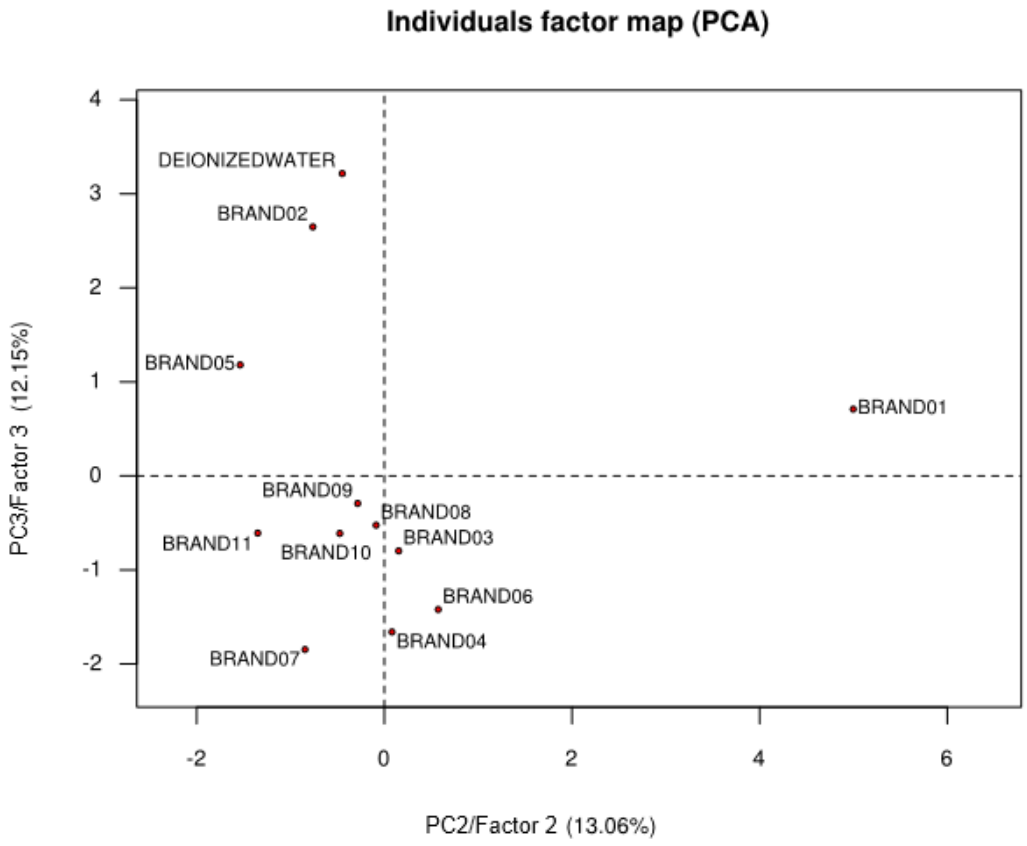

Figure 4.31 – Brands projected on The Principal Plane PC2 & PC3

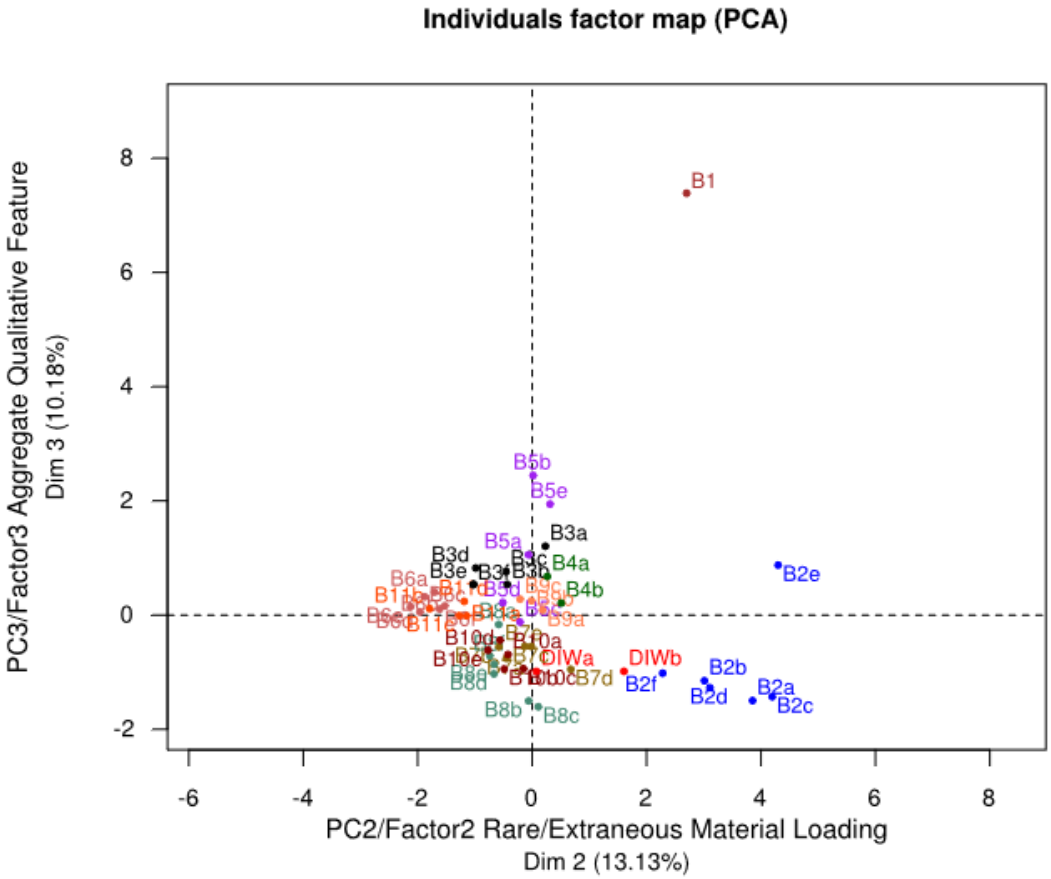

Figure 4.32 – Individuals projected on The Principal Plane PC2 & PC3

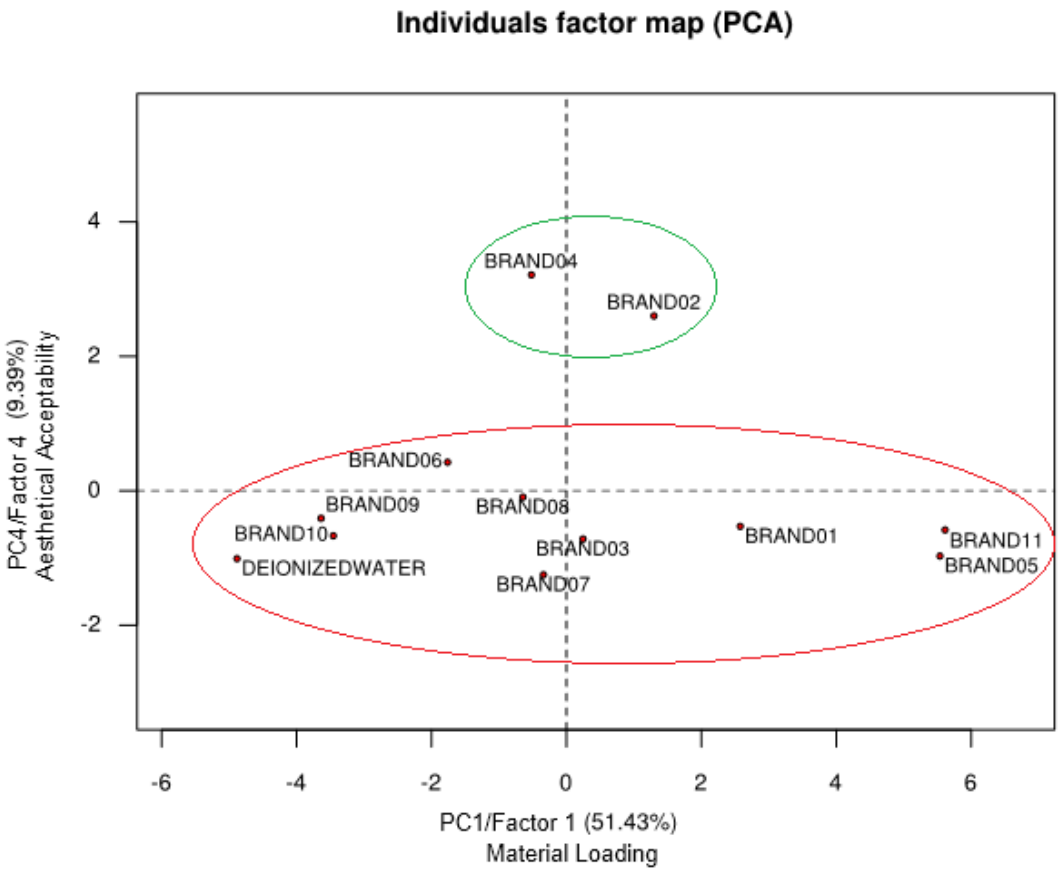

Figure 4.33 – Brands projected on The Principal Plane PC1 & PC4

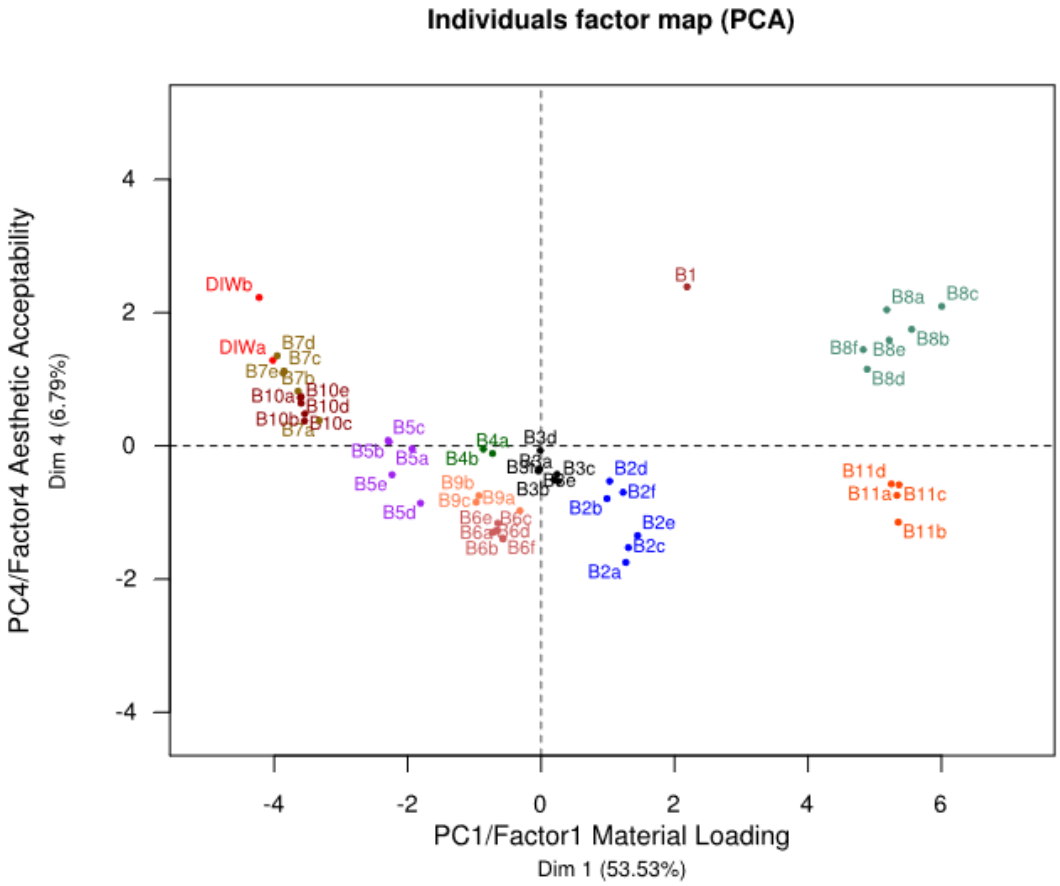

Figure 4.34 – Individuals projected on The Principal Plane PC1 & PC4

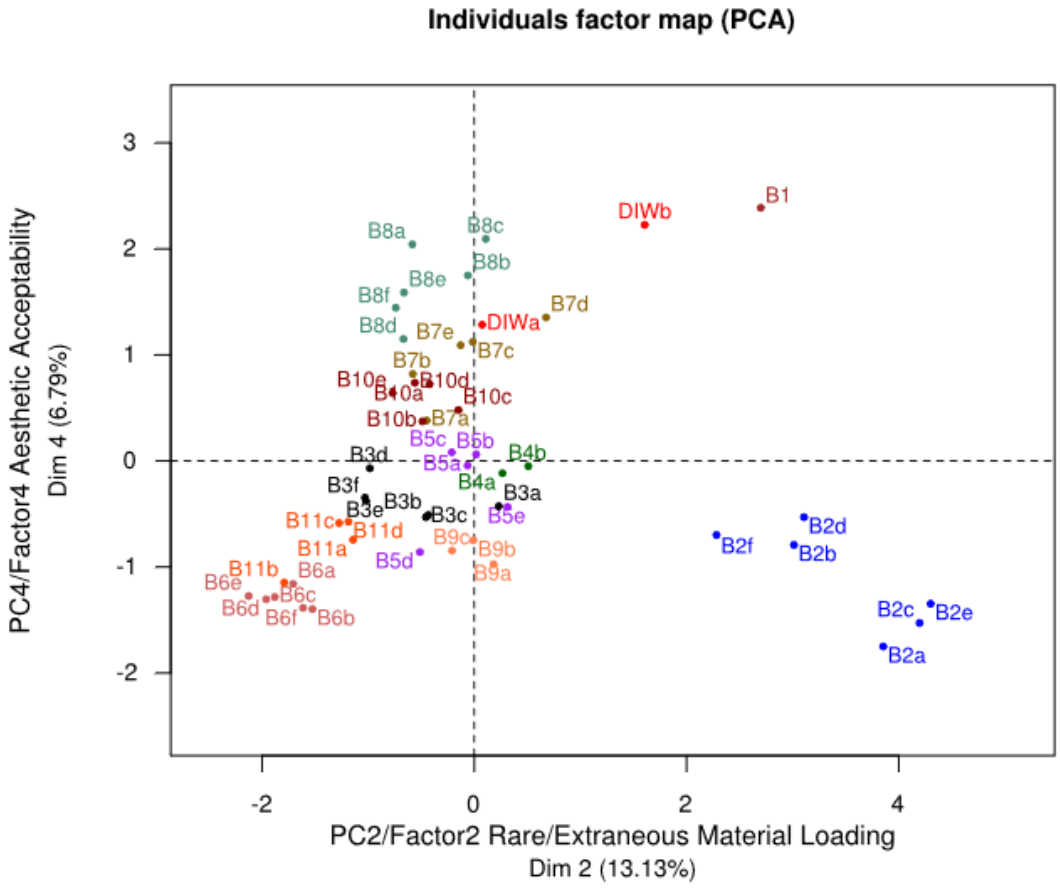

Figure 4.35 – Plot of Individuals on PC2/Factor2 vs PC4/Factor 4 Plane

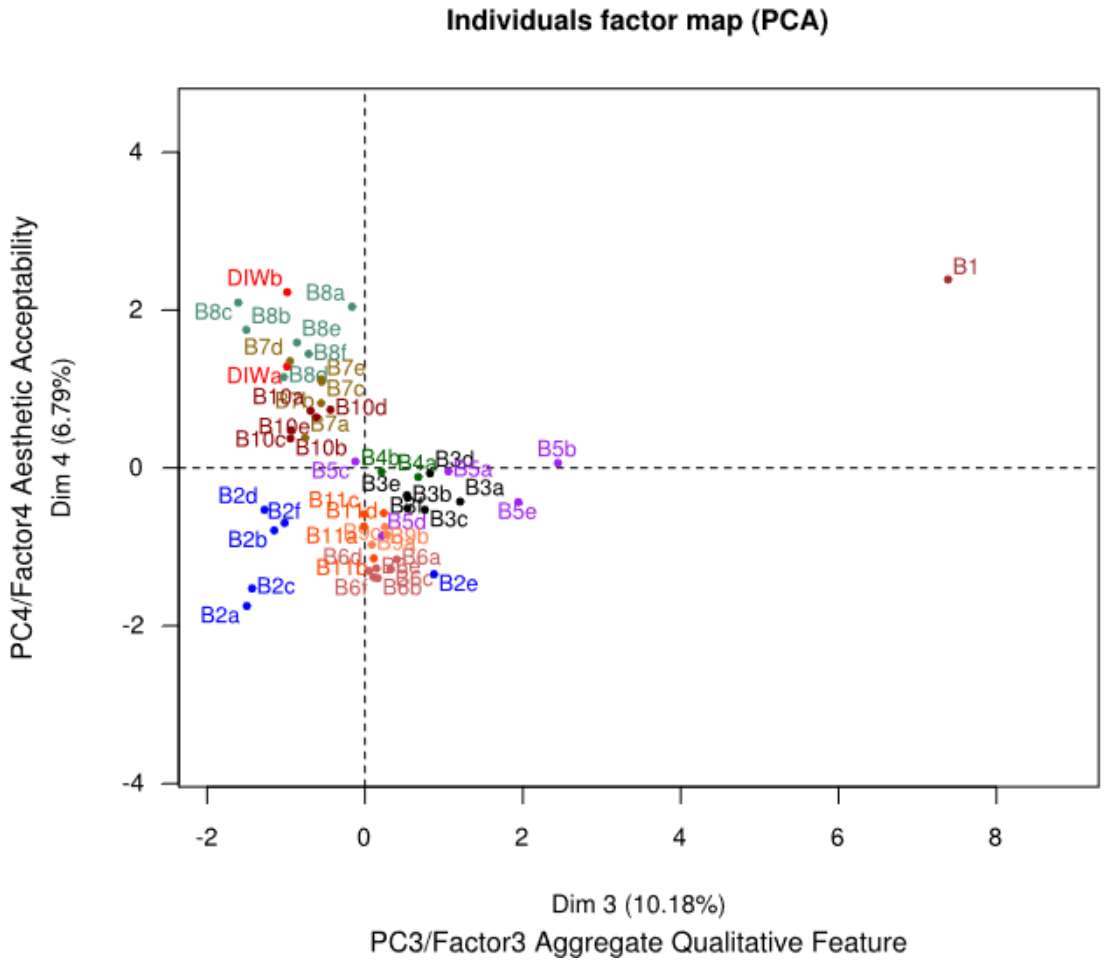

Figure 4.36 – Plot of Individuals on PC3/Factor3 vs PC4/Factor 4 Plane

From the projection on the PC1 and PC3 plane (Figure 4.28) the reasonable resolution or separation has been achieved. Whereas BRAND01 again seemed to be residing at the extreme locations alone towards the positive direction along PC3 it has been evident that BRAND03, BRAND04, BRAND05, BRAND06 showing very close proximity with respect to this axis PC3 and opposing the others e.g. BRAND02, BRAND07, BRAND08, BRAND09, BRAND10, BRAND11 and DEIONIZEDWATER. These later set of brands (BRAND02, BRAND07, BRAND08, BRAND09, BRAND10, BRAND11 and DEIONIZEDWATER) are in reality having very low values CATIONS_SUM and ANIONS_SUM in general as such these Brands are may be too “Soft” or too “light” not having mineral or having minerals with ultra low concentrations. Hence we may assume that this 3rd Principal Axis PC3 is to some extent explaining the degree of “Hardness” along positive direction or degree of “lightness” along the negative direction.

Finally from the Projection on the Principal Plane 1 & 4 (Figure 4.32), a reasonable understanding came out that this fourth component Factor PC4/ Factor 4 is probably explaining the OVERALL ACCEPTABILITY both in terms of ASESTHETICALLY as well as in terms of Suitable Amount of Minerals present in the products under study. This gives an idea that probably the only BRAND01 and BRAND02 have the balanced mineral contents as well as other aesthetically acceptable properties and hence they are the only QUALITY products and such are staying far away from others brands which do not meeting this Quality attribuite sufficiently. From the Table 4.13 it is evident that these two brands BRAND01 and BRAND02 have been very well explained, by the four PCs/Factors, respectively. This fourth component PC4/Factor 4 dragged these two Brands out separately from all other major Brands which are almost below the origin and around the bottom end (Figure 4.33), except BRAND06 who is slightly above the origin towards positive direction. It is observed that this very BRAND06 indeed staying at the borderline and sometimes it appears in one cluster and sometimes it appears in another cluster. This behaviour would be investigsted during the cluster analysis where we may visualize how this very BRAND06 migrates from one cluster (after analysis following WARD Hierarchical algorithm) and to another cluster (analysis following non-hierarchical k-Means technique). But BRAND01 and BRAND02 retain themselves together but far away from all other folks even after classification made upon applying different algorithms. This aspect has been treated later in the Section 4.7.

Hene the fourth Principal Component/Factor PC4 has been given a provisional name to indicate, say, "AESTHETICAL ACCEPTABIITY". During the Principal Factor Analysis and Cluster Analysis we would be able to see again that these two brands are in fact belong to a different cluster keeping themselves away from other brands belong to other clusters.

4.6. FACTOR ANALYSIS

4.6.1. Factor Analysis by Principal Component Analysis using Correlation Matrix

As mentioned earlier in Section 4.5.4, at the very onset of this treatise the Factor Analysis (FA) uses the correlation between variables in order to find latent factors within them and we have been reporting here the outputs of FA based solely on 12 BRANDS. As also noted before, in running the Factor Analysis using the software SAS Enterprise Guide 7.1 (64-bit), the standardized data matrix has always been used or in other words the eigen values for the Principal Factors have been extracted from the Correlation Matrix. Hence the authors have obtained the qualitatively similar results and outputs (if not exactly same in terms of values numerically) from this Factor Analyis as it has been obtained from the Principal Compenent Analysis explained above. In the light of the above it is worthy of mentioning with care that the outputs we have reported in this Section 4.6 and onward are based on the total of 11 Brands plus 01 DEIONIZED WATER for improved visualization purposes as well as to draw further inferences BRANDWISE. In actual terms both PCA based on 51 Individual and Brandwise FA came up with exactly the same interpretation. The rationale of this approache we have categorically mentioned earlier in the Section 4.5.4 which is nothing but to obtain a relatively quick, Brandwise, holistic features containing only a marginal differing numerical values throughout the whole process ahead.

Also we noted above that the success of using the Factor analysis technique depends on the correlation structure within the input data. Hence it was required to confirm that this correlation exists, otherwise the Factor Analysis may provide weak results. This analysis involved several steps. The first was to analyse the correlation structure of the data by using the correlation matrix and it has been discussed that there are significant correlations exist among majority: 16 out of 18 variables except pH and TEMP. As we already have observed that four variables, namely, NO2, SO4, Free CN and COD neither having any variability nor having any significant correlation with other variables hence were dropped from Factor Analyses as we have done during PCA.

In the next step the authors have chosen the method of extraction of eigen values. It has also been discussed that as the data matrix is having different scales hence the correlation matrix was used to get the eigen values out from the analysis. The fourth step was to take the decision on how many number of factors to be extracted and or retained for further interpretation. The correlation matrix is shown in Table 4.6 and Table 4.7. Interpretation of the factors have been made based on its loadings and it has already discussed in the sections above for Principal Component Analysis.

4.6.2. Selection of Principal Factors

There are three main criteria for defining the number of factors to retain; Pearson's, Kaiser's, and the Scree Plots. All of these methods were taken into consideration and all yielded the same solution: the optimal number of factors to be extracted is four. As shown in Table 4.15 , the percent of variance retained in these four factors is 86.03%. It has been confirmed that the First Four Principal Factors are covering the ~86.03% of the total variances or the total inertia of the original data. These four Principal Factors are also explainig the original variables and individuals in the same way what the authors have interpreted above. Hence throughout this discussion we have always used the phrases Principal Factors, Principal Axis at the same time when they were mentioning the phrases Principal Components, Principal Factors/Axes on the above sections and subsections.

Factor	Eigenvalue	Difference	Proportion	Cumulative
1	10.2856043	7.6745209	0.5143	0.5143
2	2.6110834	0.1802337	0.1306	0.6448
3	2.4308497	0.5519084	0.1215	0.7664
4	1.8789413	0.711303	0.0939	0.8603
5	1.1676383	0.5325424	0.0584	0.9187
6	0.6350959	0.1191375	0.0318	0.9505
7	0.5159584	0.3173939	0.0258	0.9763
8	0.1985645	0.0201598	0.0099	0.9862
9	0.1784047	0.0871401	0.0089	0.9951
10	0.0912646	0.0846696	0.0046	0.9997
11	0.006595	0.006595	0.0003	1

Table 4.15 – Output of the Factor Analysis and Eigenvalues of the Correlation Matrix

The scree plot and variance covered are depicted in the graphs below (Figure 4.37). They all are indeed showing the same results that have been obtained from the Principal Component Analysis. Hence no further extra information is available from these stage at this particular juncture.

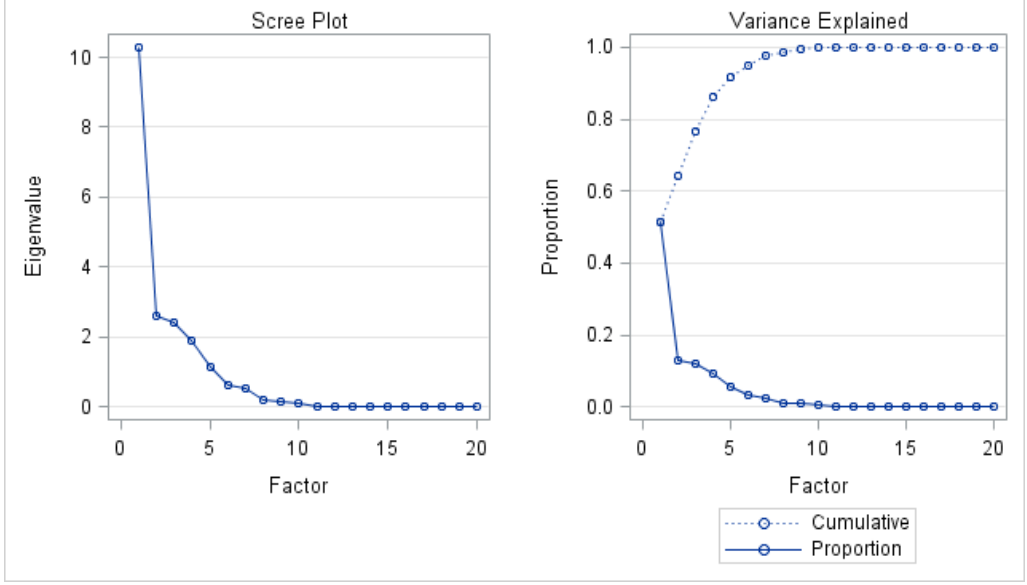

Figure 4.37 – Scree Plot & Variance Explained by the Factors from the Factor Analysis

In the Table 4.16 The Factor Pattern or the correlation among the variables and the factors are tabulated.

From the Factor Pattern Table 4.16 it has been quite clearly visible that the Factor 1 has received contributions from almost all variables except only a few like NH4, Fe, NO3 which are seemed to be obvious that these variables are generally showing very little correlations with other majority of the variables. Hence as it has been stated during the PCA interpretation that this Factor 1 could be explaining the overall "Material Loading" as it has been named for the PC1.

And interestingly it has again been confirmed that from the top of the Factor 1 pattern and order, the variables CATIONS_SUM, ANIONS_SUM, EC and TDS, HARD (hardness0 are playing the major roles building this factor and these variables are associated with the total composition of the water matrix, showing the materials load.

And as before have it been seen in PCA, here also Factor 2 has been incorporating mainly some loan variables (who are not fully addressed by the other Factors/Principal Comopnent) NH4, Mn, Cl, HCO3 and slightly brought in TEMP, Fe and NO3.

Factor 3 has again brought in the almost all ions, with slight amount from all of them, except a few. Hence as it has been stated earlier this component is probably explaining the overall anion effects qualitatively not quantitatively.

And finally the forth Factor 4 is probaly explaining the combined aesthetic acceptability as it has brought in the variables like HARD, Ca, Mg, F, Na, pH, TEMP, Fe, NO3 which are predominantly giving the "generel feeling of acceptance" of the products or not. This has also been observed in the PCA in the same way.

	Factor1	Factor2	Factor3	Factor4
stnd_CATIONS_SUM	0.98041	-0.07901	0.16169	0.02204
stnd_ANIONS_SUM	0.97093	0.00527	0.21897	0.05464
stnd_EC	0.96442	-0.11228	0.20743	0.0076
stnd_TDS	0.95521	-0.12664	0.21336	-0.00739
stnd_HARD	0.91919	0.0042	-0.04382	-0.3308
stnd_Mg	0.90734	0.19824	0.05874	-0.21135
stnd_Ca	0.90581	-0.05694	-0.07382	-0.35938
stnd_K	0.84594	-0.15347	-0.20802	0.10032
stnd_Cl	0.79986	-0.44852	0.218	-0.04428
stnd_HCO3	0.78718	0.51454	0.11306	0.06403
stnd_F	0.75539	0.16481	-0.22293	0.26245
stnd_Na	0.74874	-0.19595	0.34685	0.43594
stnd_pH	0.54416	-0.09333	-0.50781	-0.29409
stnd_Mn	0.22553	0.94668	0.09747	-0.06953
stnd_NH4	0.21543	0.9447	0.11701	0.05175
stnd_TEMP	-0.40118	0.25828	0.76306	0.26253
stnd_Fe	0.26613	0.21101	-0.73215	0.4113
stnd_NO3	0.32521	-0.25036	0.30795	0.66072

Table 4.16 – Factor Pattern

Factor Pattern Plots Initial (from Fig 4.38 to Figure 4.43) as well as Factor Rotation (Varimax) have been plotted (Figure 4.44 & Figure 4.45). The main interpretation is not different that has been described on the above previous sections and subsections merged with Principal Component Analysis results.

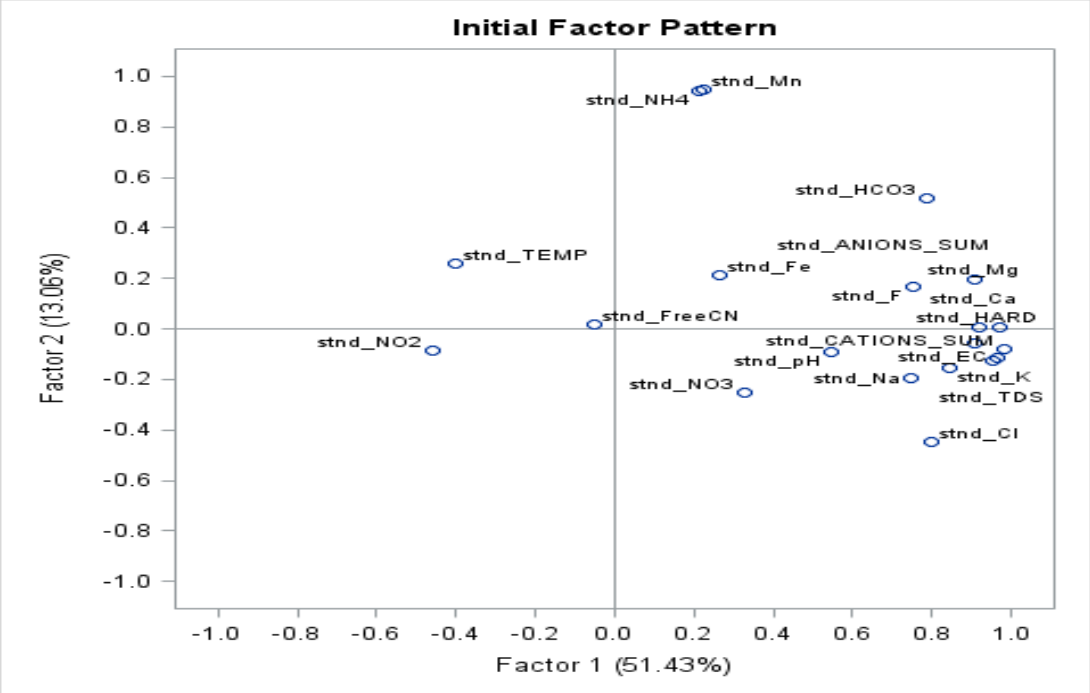

Figure 4.38 – Initial factor pattern on Plane 1 & 2

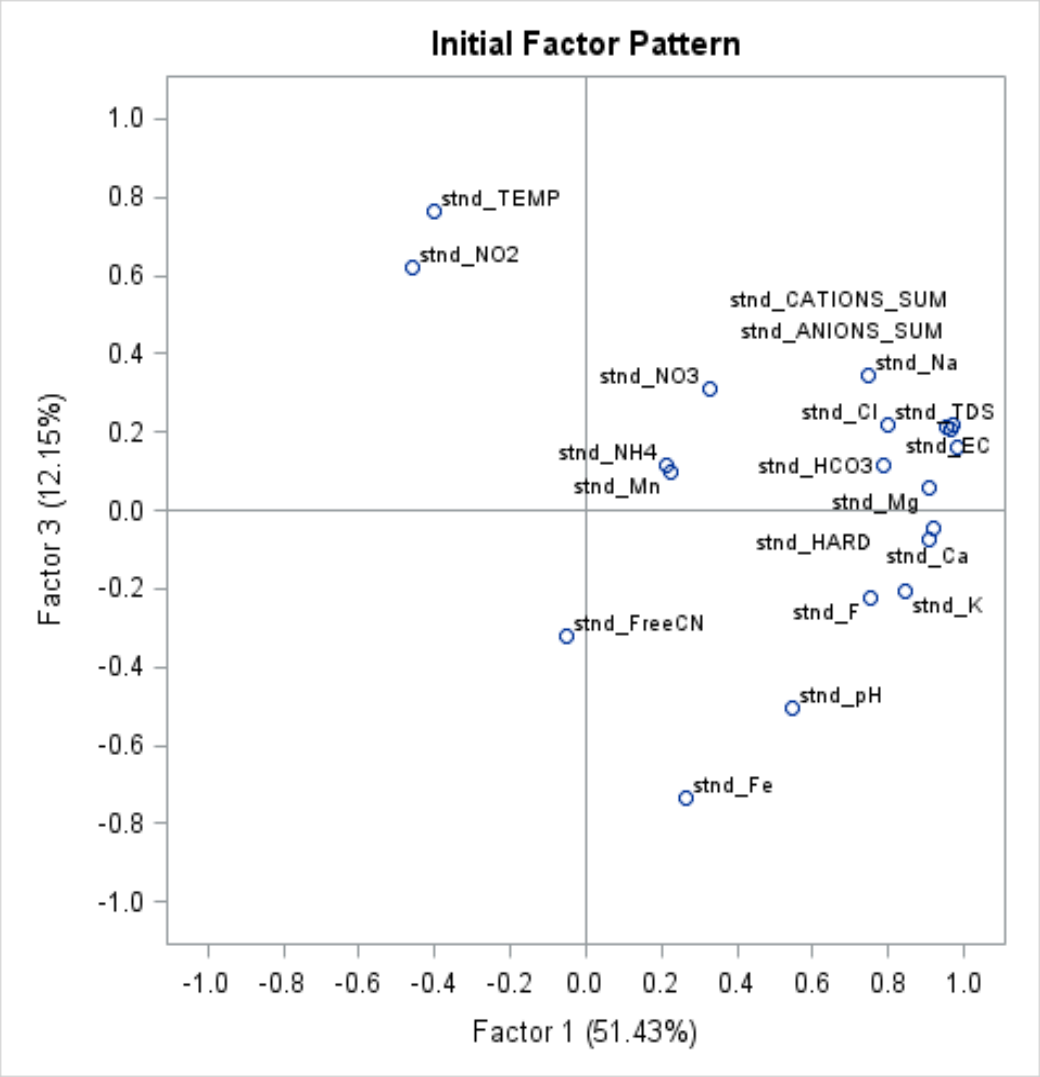

Figure 4.39 – Initial factor pattern on Plane 1 & 3

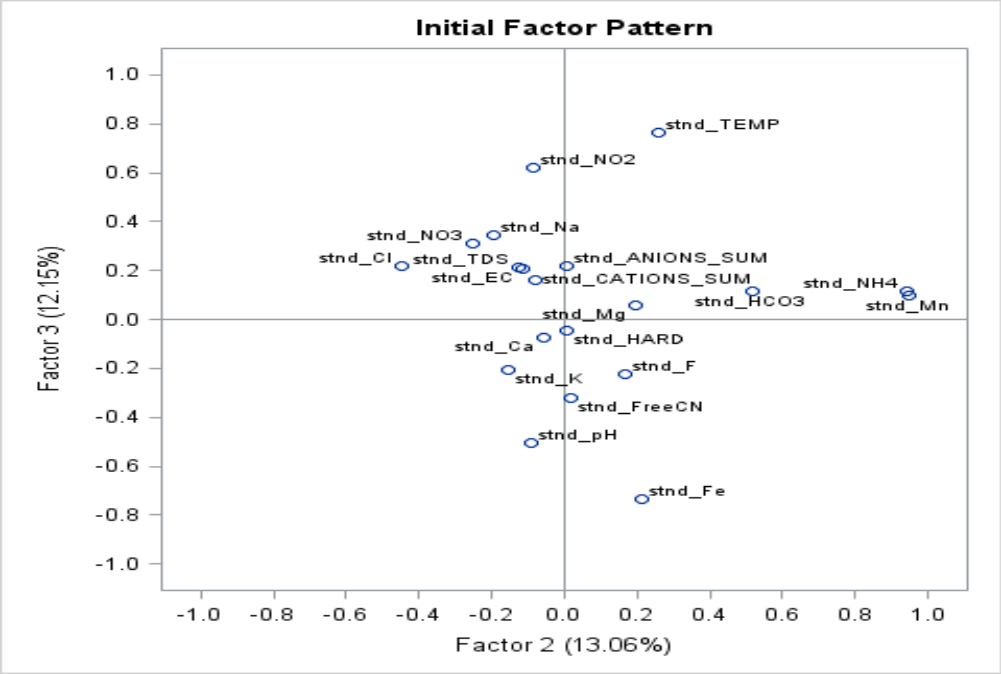

Figure 4.40 – Initial factor pattern on Plane 2 & 3

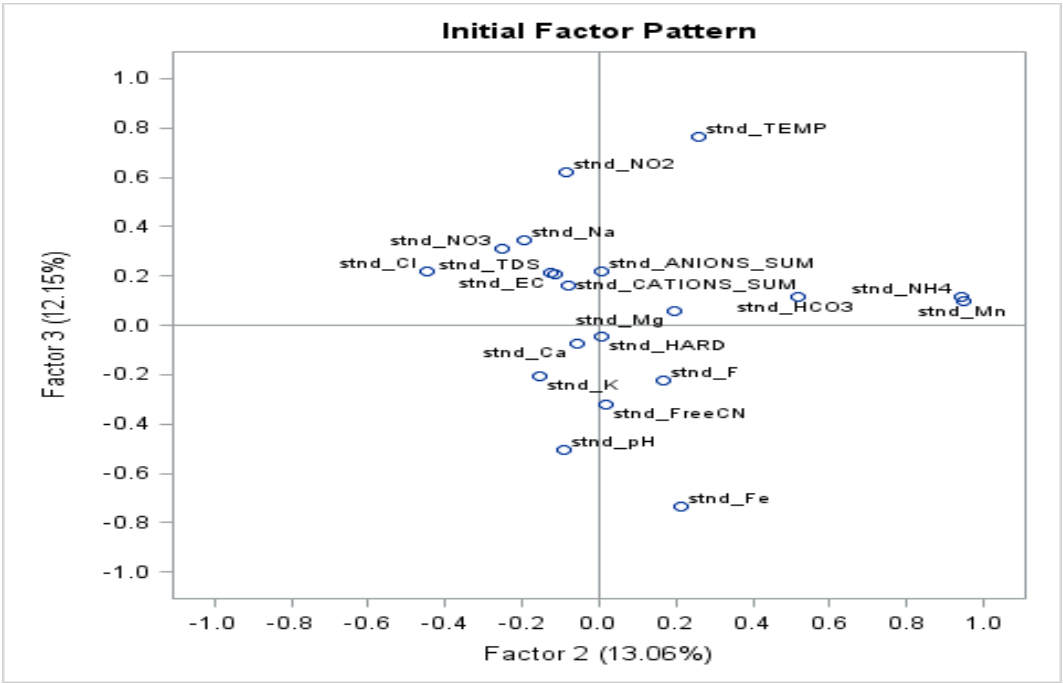

Figure 4.41 – Initial factor pattern on Plane 1 & 4

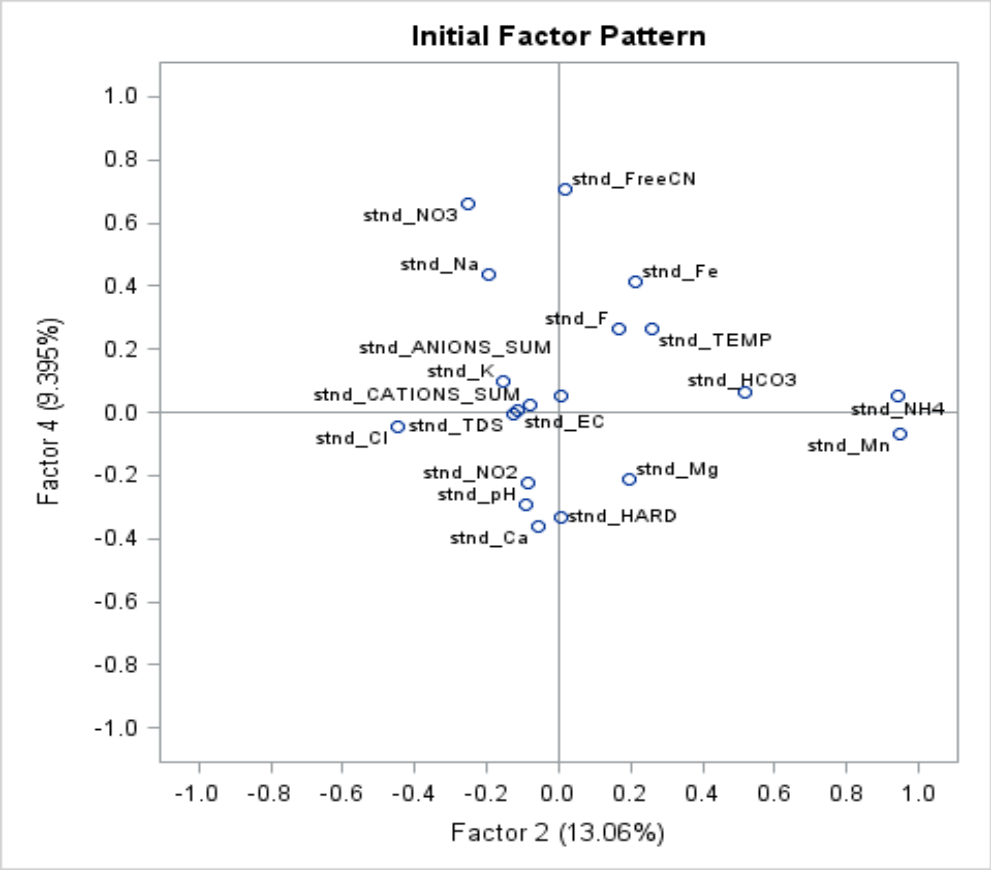

Figure 4.42 – Initial factor pattern on Plane 2 & 4

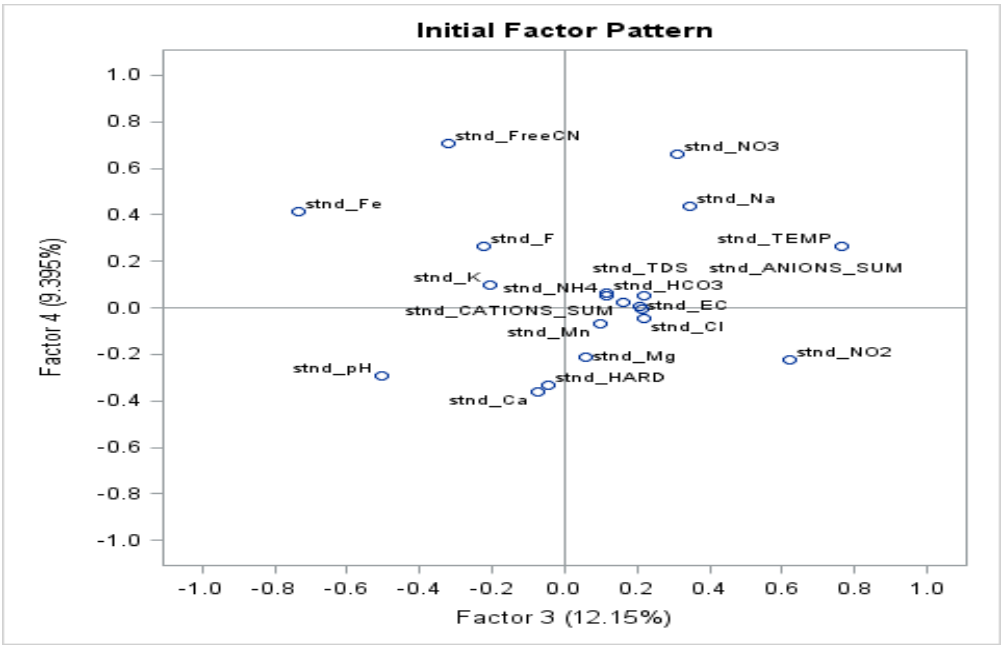

Figure 4.43 – Initial factor pattern on Plane 3 & 4

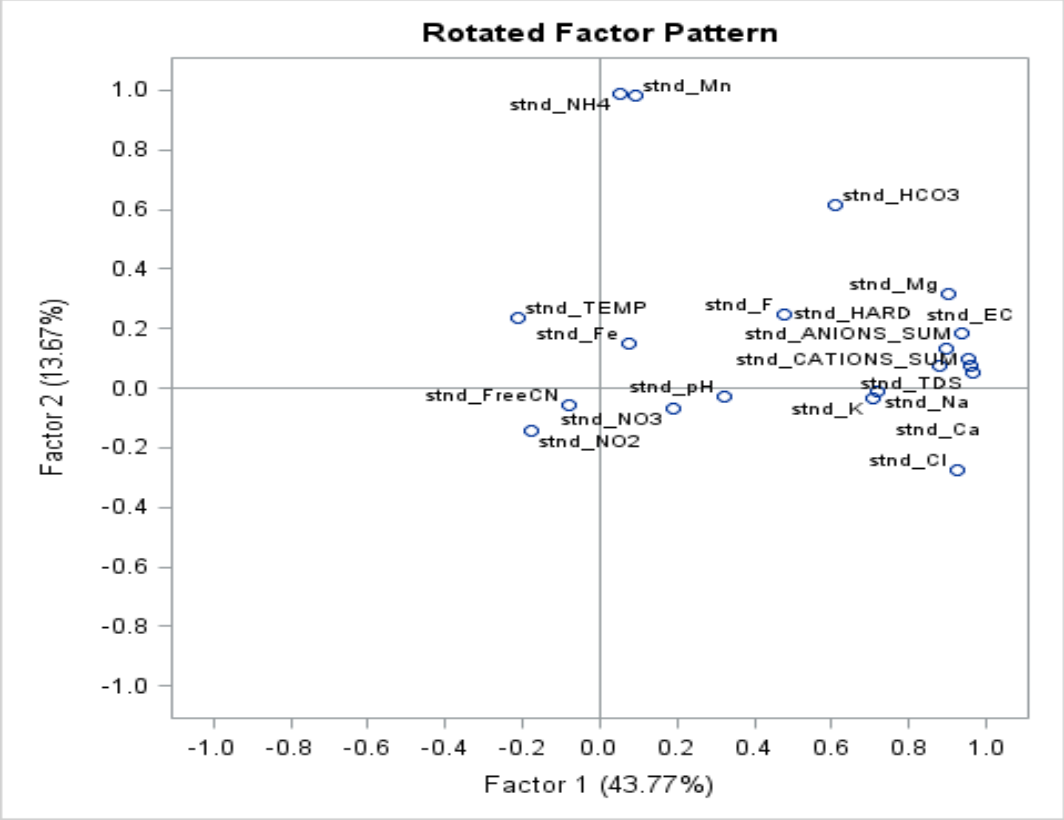

Figure 4.44 – Rotated (Varimax) Factor Pattern on F1 & F2 Plane

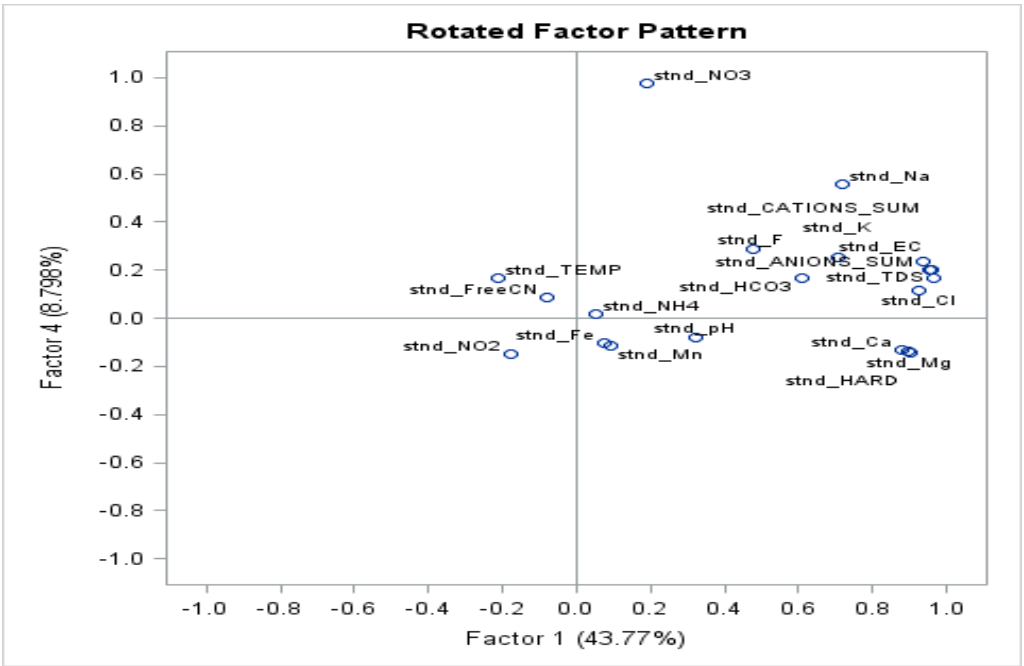

Figure 4.45 – Rotated (Varimax) Factor Pattern on F1 & F4 Plane

The results and interpretaion of the Factor Analysis are not different from that of Principal Component Analsys. The data are also available in Excel Spreed Sheets to have a further look at the data outputs and results.

4.7. CLUSTER ANALYSIS

4.7.1. Method

Latent 4 Dimensions from PCA and Factor Analysis: After the Principal Component Analysis or factor analysis, it has been concluded that there are four principal components or factors sufficient as the latent dimensions to explain the bottled water products under study.

Applying the similarity criteria using both factors and original variables the cluster analysis has been done to see whether there is any grouping exists among the bottled waters and or Brands. It has been investigated and visualized to understand the fact that which brand or product belongs to which group and what are the average and or overall behaviors of these groups or clusters.

The application of cluster analysis involves two main methods, either hierarchical or non-hierarchical. The methodology used for clustering based on factors and total 18 variables out of original 23 variables as 5 variables TEM-EC, NO2, SO4, Free CN and COD have been excluded in the previous stages.

At the begining a hierarchical procedure has been run to define the number of clusters to be extracted, since in these procedures the number of clusters depends on the data, which means that it is not neccessary to define a priori how many clusters to be generated. The ultimate solution based on hierarchical procedures depends on the distance measurement and the algorithm used.

In particular, in this study the methods like Average, Centroid and Ward's have been used and the results have been verified to assess their suitability for further interpretation and applications.

The number of clusters decided thus have been further used prior to runing the final non-hierarchical k-means algorithm in confirming the final clusters. Moreover, different distances were used. Euclidean distance, squared Euclidean distance etc.

All of these approaches returned similar results, and the solution was made based on the performance of each of the classification approaches, that is, based on the analysis of the R-square, SSE (sum of squared error (SSE) for a number of cluster solutions.

Then, the best combination of hierarchical procedures , which is in fact WARD minimum variance technique, was used to generate the initial seeds of the non-hierarchical algorithm – k-means. It has been seen that WARD provided the best results. The number of factor or cluster has been finally retained from the WARD output which was 4. This approach yielded better results. Following the

generation of the clusters, classification among the individuals has been done based on a ‘‘profiling analysis’’ and creating profile plots both using original data and the centered data to have a better understanding and visualiztion. The general statistical properties of each cluster have been tabulated to understand their relative positions.

4.7.2. Cluster Analysis using Hierarchical Approach

4.7.2.1. Average Linkage Method

The dendogram (Figure 4.46) shows the average distances between the clusters based on average linkage hierarchical cluster analysis. The dendogram shows the Brands grouped in different clusters vs the average distances between the clusters (Figure 4.47). Another Dendogram depicted the groups among the Individuals (Figure 4.48).

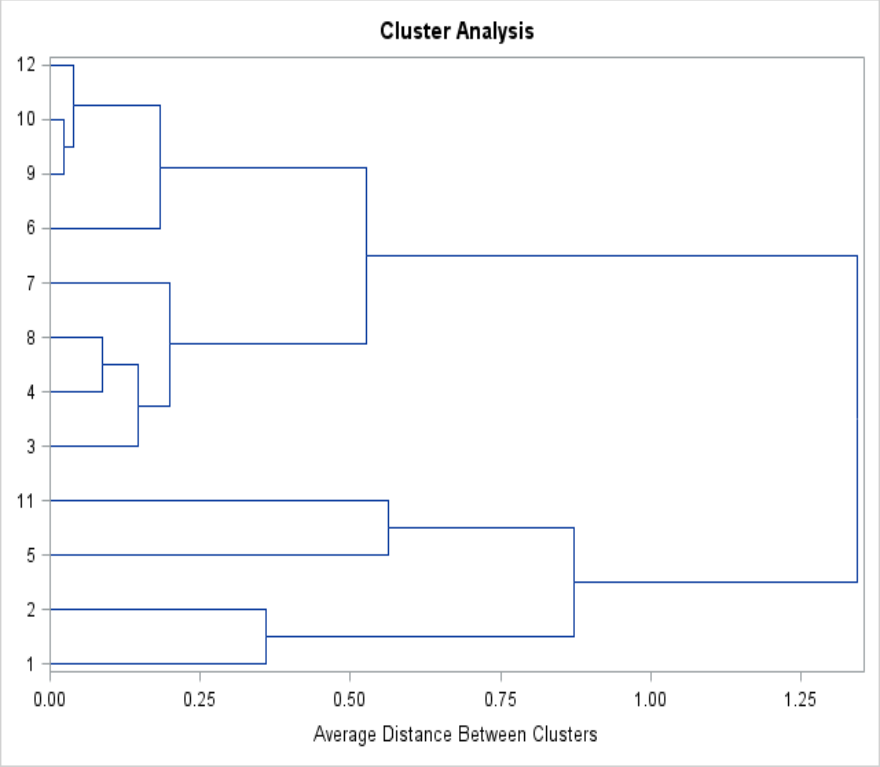

Figure 4.46 – Average Linkage Hierarchical

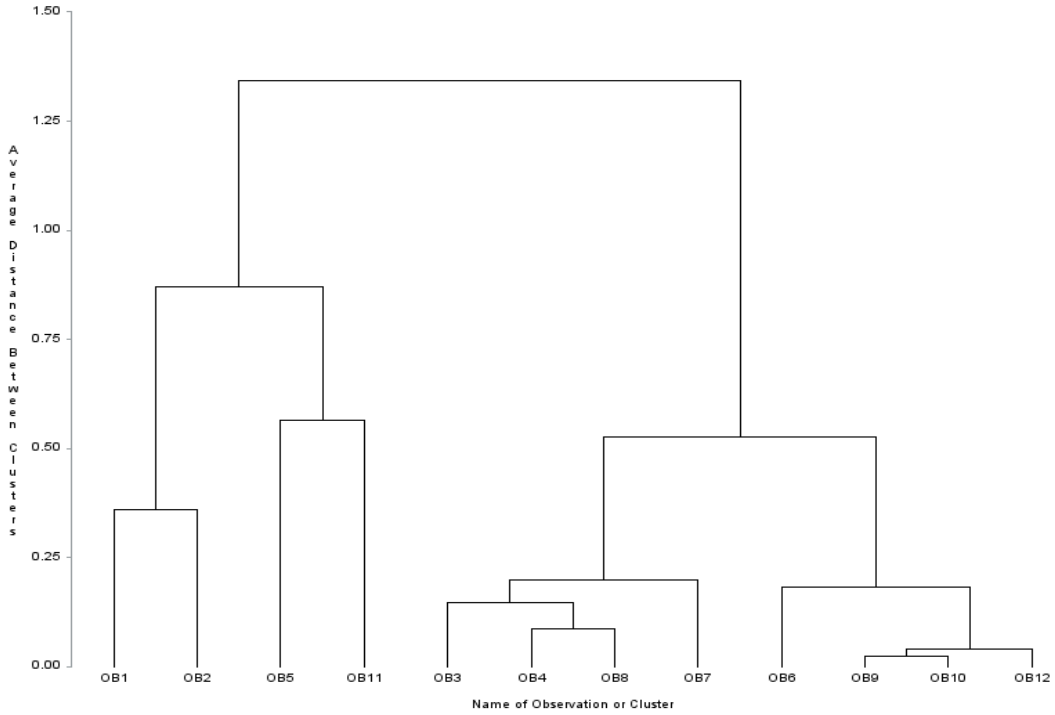

Figure 4.47 – Individuals vs Average Distances

B01: BRAND01, B02: BRAND02 and so on B12: DEIONIZEDWATER

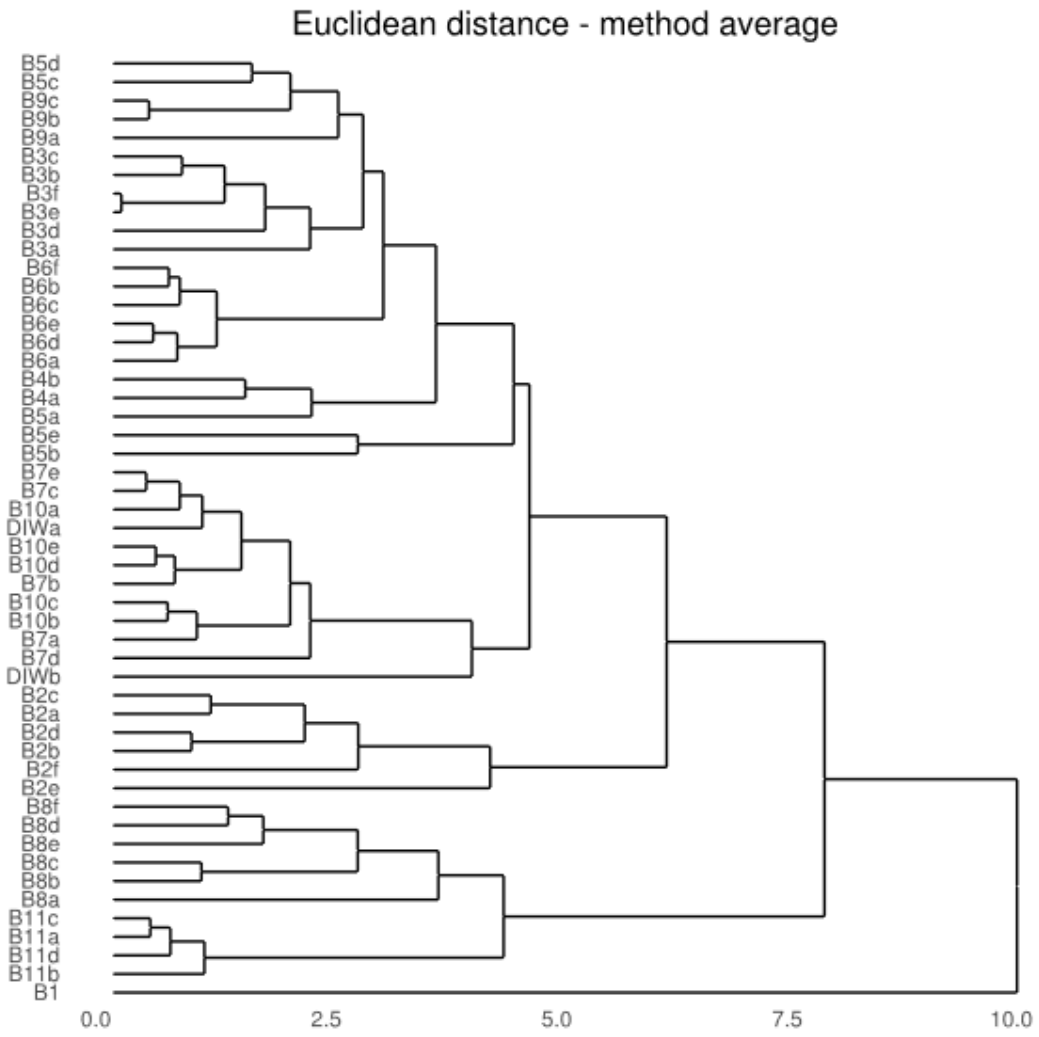

Figure 4.48 –Dendogram using Euclidean Distance with Average Linkage Method (Clusters of 51 Individuals)

The Table 4.17 below shows the cluster history, the R-Square values and other data for average linkage method.

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square
11	BRAND09	BRAND10	2	0	1
10	CL11	DEIONIZEDWATER (B12)	3	0.0002	1
9	BRAND04	BRAND08	2	0.0007	0.999
8	BRAND03	CL9	3	0.0024	0.997
7	BRAND06	CL10	4	0.0045	0.992
6	CL8	BRAND07	4	0.0046	0.988
5	BRAND01	BRAND02	2	0.0117	0.976
4	CL6	CL7	8	0.0948	0.881
3	OB5	OB11	2	0.0289	0.852
2	CL5	CL3	4	0.1177	0.735
1	CL2	CL4	12	0.7345	0

Table 4.17 – Cluster History & R-Square values (Average Linkage)

4.7.2.2. Centroid Method

The eigen values (Table 4.18) of the covariance matrix used in centroid method cluster analysis are similar, if not exactly same in terms of numerical terms, that has been obtained in the previous PCA (for 51 individuals) and Factor Analysis (for 12 BRANDS). This is because the centroid method uses the centered data (not the orginal data matrix) to extract the factors/PCs. First four factors covers the variances ~86.03%.

Factor	Eigenvalue	Difference	Proportion	Cumulative
1	10.2856	7.674521	0.5143	0.5143
2	2.611083	0.180234	0.1306	0.6448
3	2.43085	0.551908	0.1215	0.7664
4	1.878941	0.711303	0.0939	0.8603
5	1.167638	0.532542	0.0584	0.9187
6	0.635096	0.119138	0.0318	0.9505
7	0.515958	0.317394	0.0258	0.9763
8	0.198565	0.02016	0.0099	0.9862
9	0.178405	0.08714	0.0089	0.9951
10	0.091265	0.08467	0.0046	0.9997
11	0.006595	0.006595	0.0003	1

Table 4.18 – Eigenvalues of the Covariance Matrix: From Centroid Cluster Analysis Output

The dendogram 4.49 shows the distances of the centroids of the clusters based on centroid hierarchical cluster method, Figure 4.50 shows how the Brands are clustered and similarly Figure 4.51 shows how the Individuals are grouped into different clusters. And Figure 4.52 depicts the

Dendogram of 51 individuals forming the clusters based on Euclidean Distance with Complete Method.

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square
11	BRAND09	BRAND10	2	0.0012	0.999
10	BRAND03	BRAND08	2	0.0091	0.99
9	CL10	BRAND07	3	0.0171	0.973
8	CL9	BRAND06	4	0.025	0.948
7	CL8	CL11	6	0.0729	0.875
6	BRAND05	BRAND011	2	0.0351	0.84
5	CL7	BRAND04	7	0.0774	0.762
4	BRAND02	CL5	8	0.1158	0.646
3	CL4	DEIONIZEDWATER (B12)	9	0.1299	0.516
2	BRAND01	CL3	10	0.1711	0.345
1	CL2	CL6	12	0.3454	0

Table 4.19 – History of Cluster (Centroid Cluster Analysis)

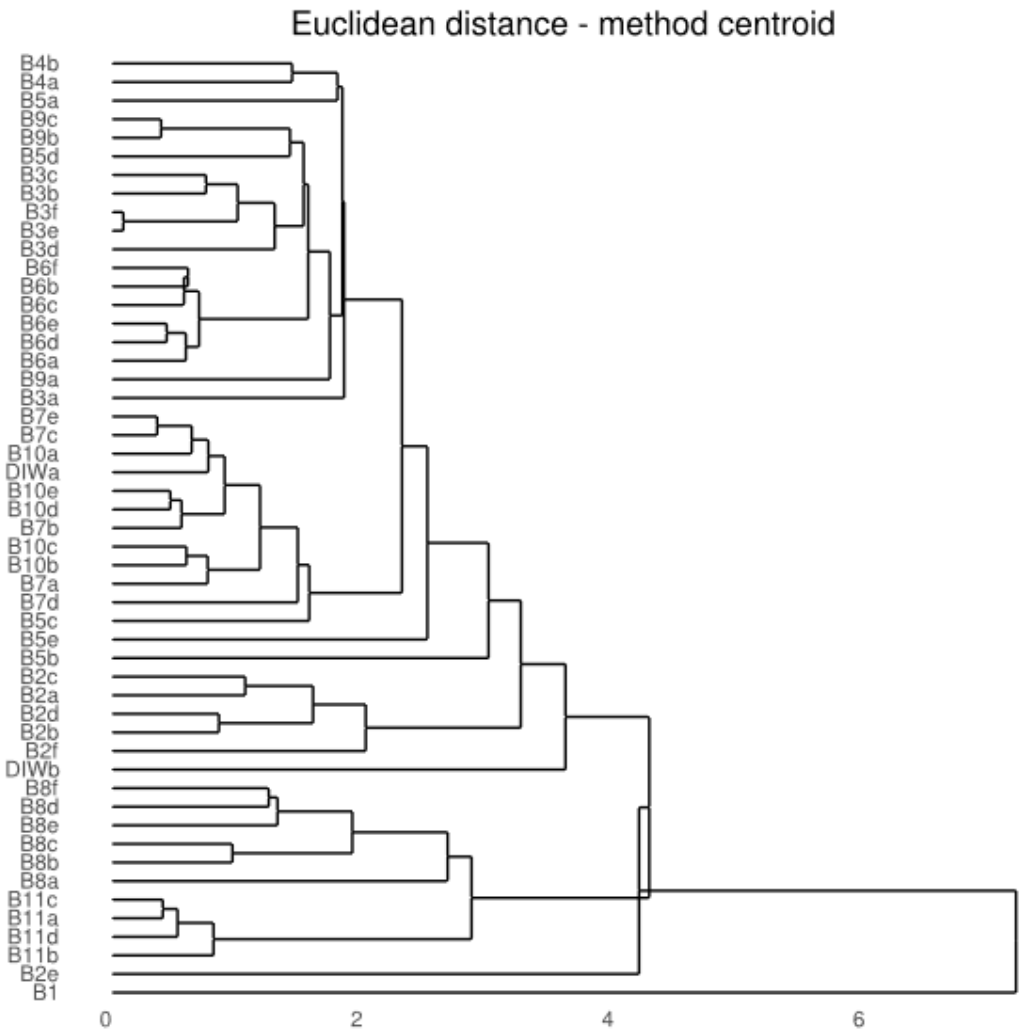

Figure 4.49 – Dendogram using Euclidean Distance with Centroid Method (Clusters of Individuals are revealed)

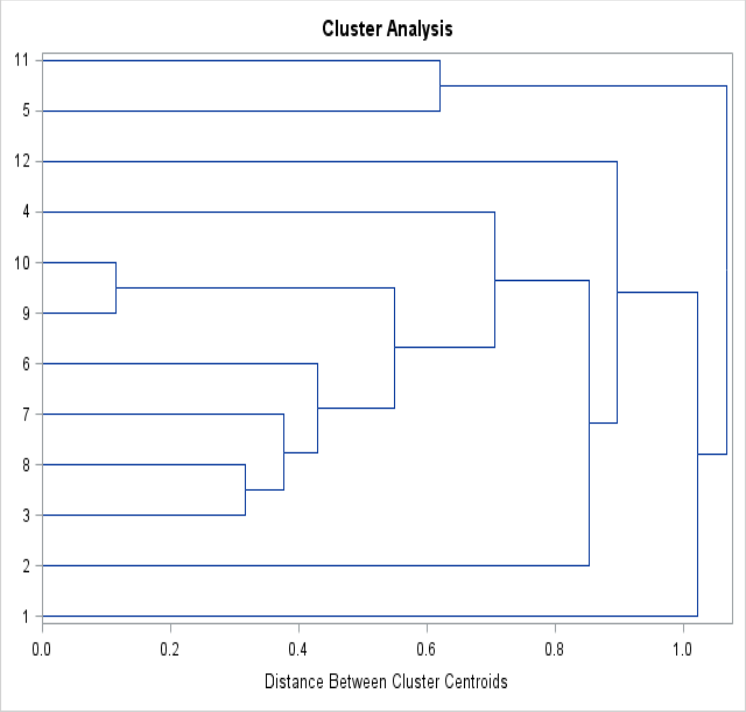

Figure 4.50 – Distance Between Cluster Centroids vs Cluster

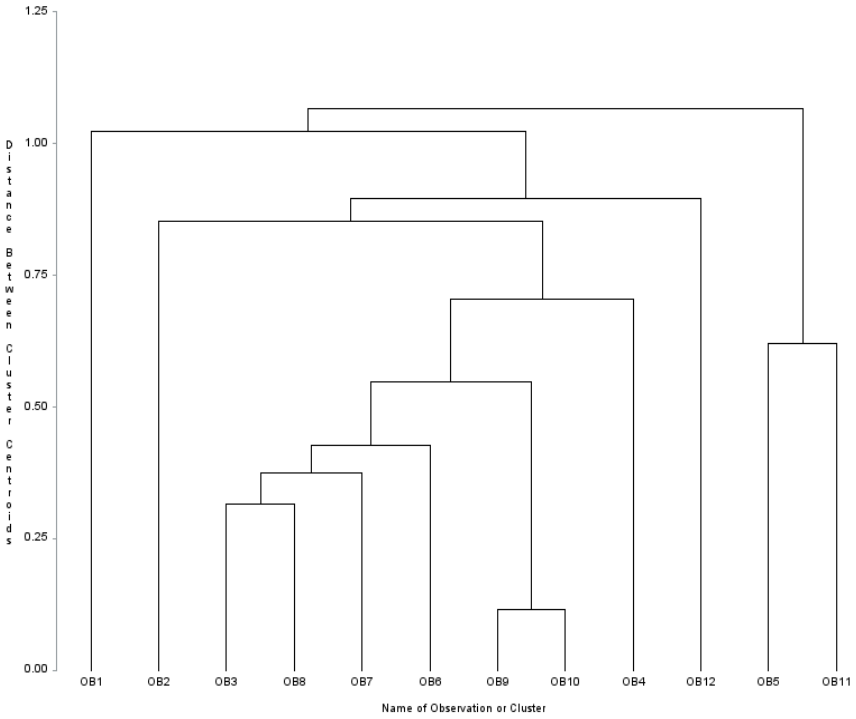

Figure 4.51 – Centroid Distances vs Brands

B01: BRAND01, B02: BRAND02 and so on B12: DEIONIZEDWATER

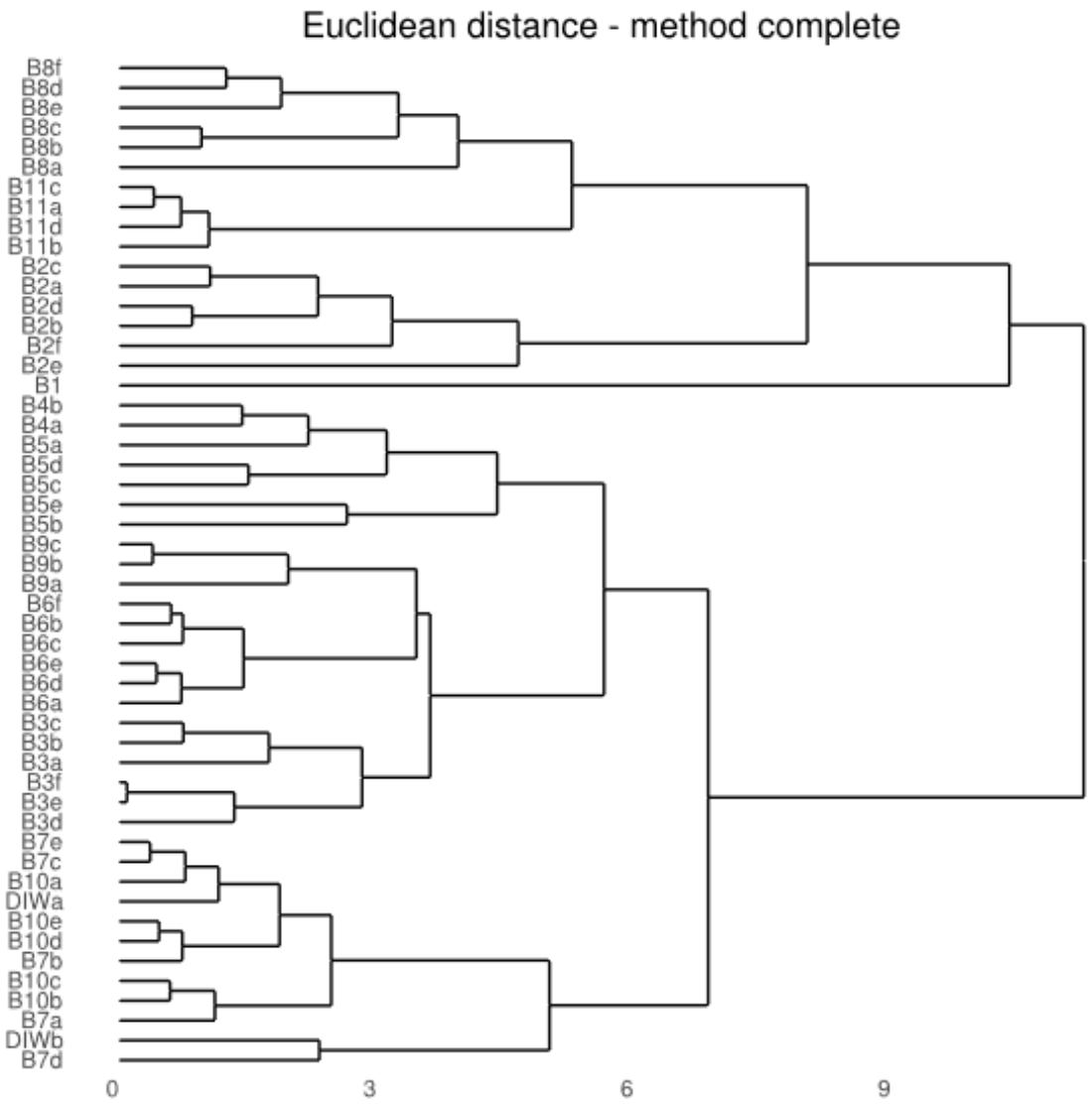

Figure 4.52 –Dendogram using Euclidean Distance with Complete Method

4.7.2.3. WARD's Minimum Variance Cluster Analysis

This WARD has particularly given the best results among all the hierachical techniques applied. The eigen values (Table 4.20) for the factors (essentially extracted from the centered data) are being the same and the first four factors are the principal factors to be considered further and covering 86.03% of the total interia or variance. The dendogram Figure 4.53 shows the distances of the centroids of the clusters based on WARD hierarchical cluster method and also shows the different Brands grouped in different clusters. In the Figure 4.54 the dendogram has been shown based on the same WARD hierarchical cluster method where the Individuals are grouped clearly.

Factor	Eigenvalue	Difference	Proportion	Cumulative
1	10.2856043	7.6745209	0.5143	0.5143
2	2.6110834	0.1802337	0.1306	0.6448
3	2.4308497	0.5519084	0.1215	0.7664
4	1.8789413	0.711303	0.0939	0.8603
5	1.1676383	0.5325424	0.0584	0.9187
6	0.6350959	0.1191375	0.0318	0.9505
7	0.5159584	0.3173939	0.0258	0.9763
8	0.1985645	0.0201598	0.0099	0.9862
9	0.1784047	0.0871401	0.0089	0.9951
10	0.0912646	0.0846696	0.0046	0.9997
11	0.006595	0.006595	0.0003	1

Table 4.20 – WARD Method: Eigenvalues of the Covariance Matrix (from the Centered Data)

The Table 4.21 shows the cluster history, the R-Square values and other data for WARD Cluster analysis method.

Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**
11	BRAND09	BRAND010	2	0.0012	0.999
10	BRAND03	BRAND08	2	0.0091	0.99
9	CL10	BRAND07	3	0.0171	0.973
8	CL9	BRAND06	4	0.025	0.948
7	OB5	BRAND011	2	0.0351	0.913
6	CL11	DEIONIZEDWATER B12	3	0.0647	0.848
5	CL8	BRAND04	5	0.0695	0.778
4	BRAND02	CL5	6	0.1004	0.678
3	BRAND01	CL4	7	0.1377	0.54
2	CL3	CL6	10	0.1948	0.345
1	CL2	CL7	12	0.3454	0

Table 4.21 – History of Cluster (WARD Cluster Analysis)

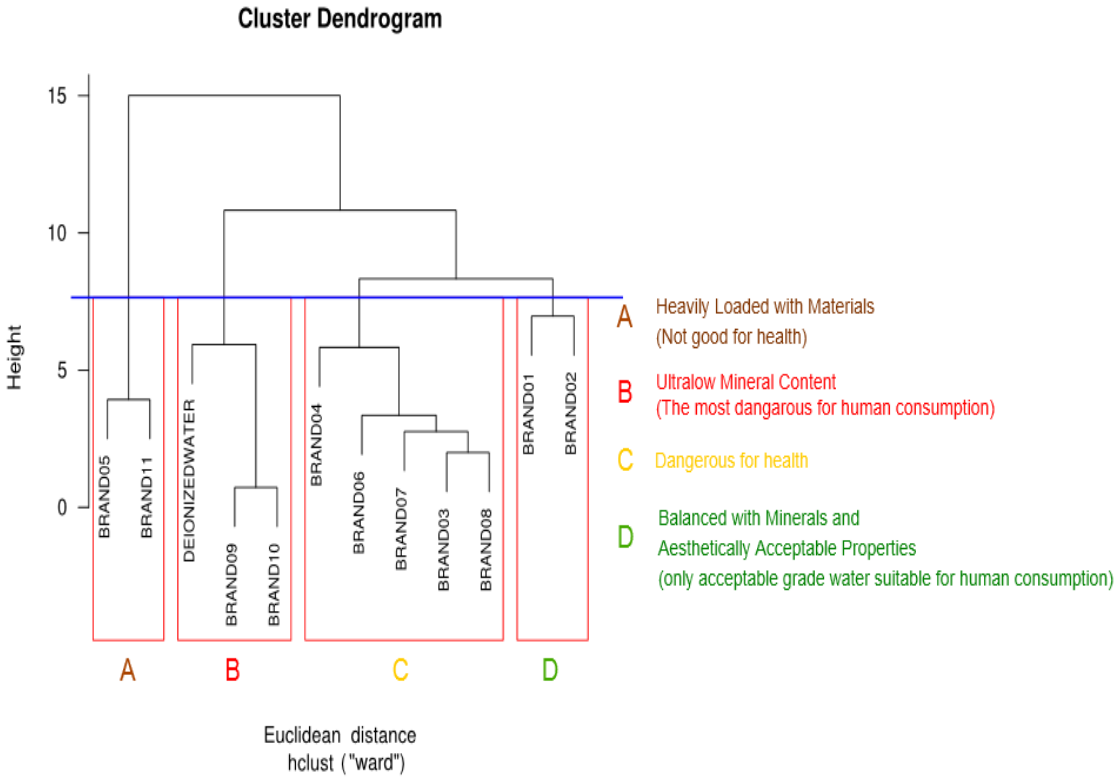

Figure 4.53 – Dendogram from WARD Cluster Analysis

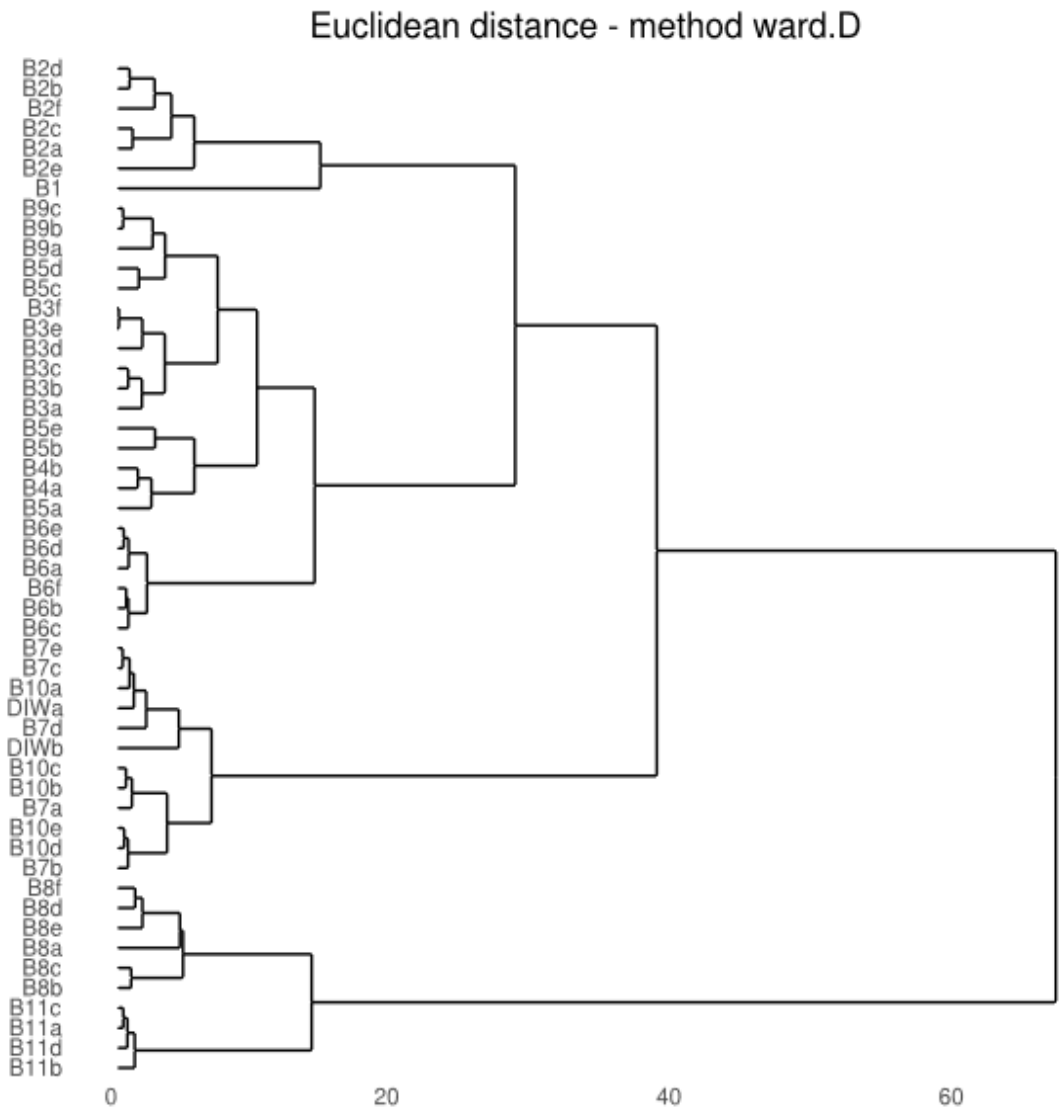

Figure 4.54 –Dendogram using Euclidean Distance with WARD Method

4.7.2.4. Cluster Analysis Result: Members from WARD Hierarchical approach

It has been evident from WARD's techniques four clusters (Figure 4.53) are present with the following individuals in each cluster.

Cluster A: BRAND05 - BRAND11: Heavily loaded with materials (may not be suitable for consumption).

Cluster B: BRAND09 - BRAND10 – DEIONIZEDWATER (Ultralow Mineral Content and the most dangerous for human health but good for industrial usage).

Cluster C: BRAND03 - BRAND04 – BRAND06-BRAND07 - BRAND08 (Also dangerous for health and having some risk for health if continuously consumed for a longer period).

Cluster D: BRAND01 - BRAND02 (Balanced with minerals and aesthetically acceptable products only suitable for human consumption).

4.7.2.5. Choosing the appropriate cluster solution: Performance of Hierarchical Approaches

One common method of choosing the appropriate cluster solution is to compare the sum of squared error (SSE) for a number of cluster solutions. SSE is defined as the sum of the squared distance between each member of a cluster and its cluster centroid. Thus, SSE can be seen as a global measure of error. In general, as the number of clusters increases, the SSE should decrease because clusters are, by definition, smaller. A plot of the SSE against a series of sequential cluster levels can provide a useful graphical way to choose an appropriate cluster level. Such a plot can be interpreted much like a scree plot used in factor analysis.

To decide the number of cluster or seeds to be used in final non-hierachical k-Means method, the plots for R-Square and SSE have been drawn against the number of clusters. In this study SSE has been utilized instead of R-Square from the plots Figure 4.55 & Figure 4.56 as SSE graphics helped to determine the appropriate numbers of clusters, which plotting sum of squares by number of clusters extracted (using K means) within groups. This has also confirmed that the performance and or effciency of WARD technique above other hierachical approaches. The authors observed the better results using SSE than the plot generated with R-square with hierarchical clusters. This after comparison number of cluster has been selected from the plot of SSE.

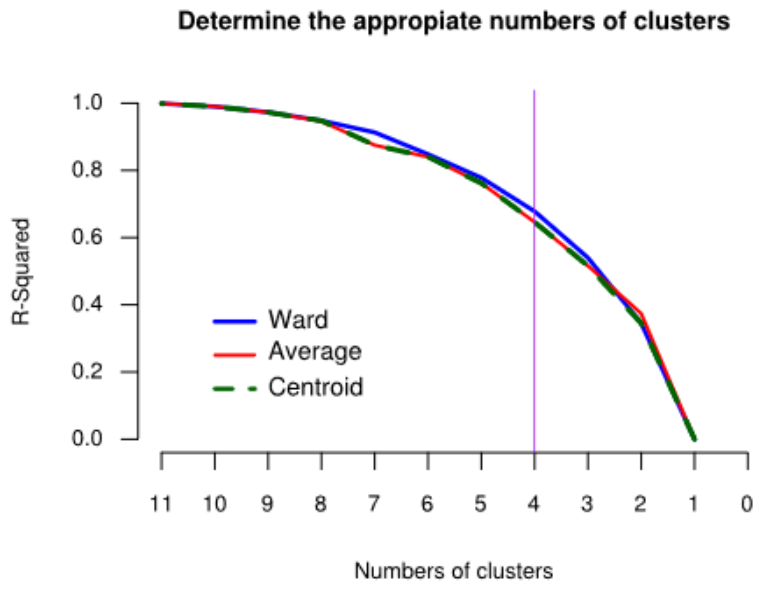

Figure 4.55 –R-Squared vs Number of Clusters from Different Hierarchical Cluster Analysis

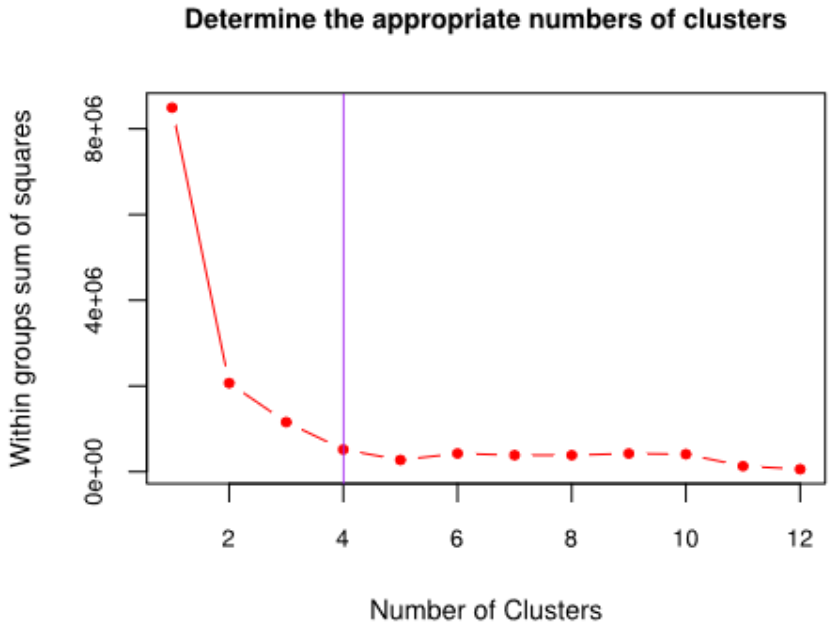

Figure 4.56 –SSE vs Number of Clusters from Different Hierarchical Cluster Analysis

4.7.2.6. Cluster Analysis: Non-hierarchical k-means approach

1. Results of K-Means Cluster Analysis

The Cluster analysis using K-means is clearly extracted the four distinct clusters. As it is a random process and depends from where it starts with the four seeds the authors ran five times K-means algorithm with 5 iterations and the authors reached always to the same solution.

Table 4.22 shows how all 51 individuals are attained membership among 4 clusters, namely, A, B, C and D.

Cluster A:	B8a	B8b	B8c	B8d	B8e	B8f	B11a	B11b	B11c	B11d							
Cluster B:	DIWa	B5a	B5b	B5c	B5d	B5e	B7a	B7b	B7c	B7d	B7e	B10a	B10b	B10c	B10d	B10e	DIWb
Cluster C:	B3a	B3b	B3c	B3d	B3e	B3f	B4a	B4b	B6a	B6b	B6c	B6d	B6e	B6f	B9a	B9b	B9c
Cluster D:	B1	B2a	B2b	B2c	B2d	B2e	B2f										

Table 4.22 – Individuals classified among 4 Cluster from final k-means HCA

The following is the Brandwise result of the k-mean cluster analysis:

Cluster A: BRAND08 – BRAND11 (Heavily loaded with materials and may not be suitable for consumption).

Cluster B: BRAND05-BRAND07-BRAND10 – DEIONIZEDWATER (Ultralow Mineral Content and the most dangerous for human health but good for industrial usage).

Cluster C: BRAND03-BRAND04-BRAND06-BRAND09 (Low mineral content dangerous for health and having some risk for health if continuously consumed for a longer period).

Cluster D: BRAND01 - BRAND02 (Balanced with minerals and aesthetically acceptable products only suitable for human consumption).

The major difference is: the BRAND06 which was in Cluster C as per hierarchical WARD's method has been now migrated to Cluster B after K-Means procedure in the class showing that it is indeed an ultralow mineral content products. This is quite understandable that this BRAND06 was staying at the borderline of the other cluster and this brand is indeed having very low mineral substances. Therefore this solution from K-Means is quite practical and acceptable in this particular context.

Cluster	WARD	K-Means	Classification
Cluster A	BRAND05-BRAND11	BRAND08-BRAND11	Heavily Loaded
Cluster B	BRAND09-BRAND10-DEIONIZEDWATER	BRAND05-BRAND07-BRAND10-DEIONIZEDWATER	Ultralow Content
Cluster C	BRAND03-BRAND04-BRAND06-BRAND07-BRAND08	BRAND03-BRAND04-BRAND06-BRAND09	Low Mineral Content
Cluster D	BRAND01-BRAND02	BRAND01-BRAND02	Balanced

Table 4.23 – Clusters Detection Comparison between WARD and K-Means

2. Cluster Statistics

Cluster	Cluster A					Cluster B				
Variable/Statistic	Average	St Dev	Minimun	Maximun	Skewness	Average	St Dev	Minimun	Maximun	Skewness
TEMP	25.64	0.51	24.8	26.3	-0.25	26.06	0.7	25.3	28.1	1.41
pH	7.81	0.17	7.7	8.2	1.29	6.62	0.52	5.7	7.7	0.44
EC	905.89	121.77	763	1031	-0.19	37.68	35.94	2.4	99.7	0.73
NH4	0.1	0	0.1	0.11	2.28	0.11	0.03	0.1	0.2	2.09
NO3	7.07	8.68	0.3	17.9	0.35	1.33	1.55	0.1	4.2	0.79
Cl	154.69	60.75	100.43	269.9	0.79	1.63	1.45	0.3	4.59	0.81
HCO3	173.9	35.57	145	222	0.36	23.05	20.51	3.7	55.5	0.61
F	0.47	0.07	0.41	0.57	0.34	0.19	0.14	0.09	0.45	0.91
HARD	83.2	4.33	73.19	87.64	-0.95	2.67	3.43	0	10.29	0.98
TDS	609.74	142.86	488.32	959.64	1.26	24.17	22.98	1.54	63.8	0.72
Na	75.25	20.14	51.3	97.04	-0.29	8.38	7.55	0.02	22.79	0.65
K	4.03	1.67	2.69	6.09	0.35	0.63	0.66	0.05	1.72	0.79
Ca	63.29	4.25	54.45	67.91	-0.54	1.89	2.31	0.25	7.39	1.07
Mg	19.85	0.62	18.74	20.54	-0.48	0.79	1.12	0.02	2.91	0.87
Fe	0.19	0.05	0.13	0.31	0.84	0.2	0.06	0.11	0.33	0.78
Mn	0.06	0.02	0.03	0.08	-0.39	0.07	0.01	0.04	0.08	-1.61
ANIONS_SUM	7.41	1.36	6.16	10.07	0.73	0.51	0.4	0.13	1.16	0.61
CATIONS_SUM	8.18	0.68	7.37	8.94	-0.27	0.56	0.53	0.03	1.64	0.82

Table 4.24 – Cluster Statistics for Cluster A & B in Orginal Measurement Scale

Cluster	Cluster C					Cluster D				
Variable/Statistic	Average	St Dev	Minimun	Maximun	Skewness	Average	St Dev	Minimun	Maximun	Skewness
TEMP	25.84	0.34	25.4	26.6	0.59	26.64	0.37	26	27.1	-0.39
pH	8.25	0.84	6.8	9.2	-0.27	6.97	0.17	6.8	7.3	0.9
EC	267.24	21.73	234	298.5	0.22	554.71	27.14	497.3	577	-1.21
NH4	0.11	0.01	0.1	0.15	2.08	0.14	0.06	0.1	0.24	0.76
NO3	2.22	2.99	0.2	9.9	1.76	30.97	18.16	0.1	51.6	-0.35
Cl	15.39	12.11	2.4	31.1	0.17	56.43	22.9	4.85	68.03	-1.58
HCO3	102.05	60.05	27.8	166.5	-0.18	174.73	56.04	148	301.6	1.61
F	0.44	0.06	0.35	0.56	0.57	0.48	0.02	0.45	0.5	-0.46
HARD	27.82	10.06	8.51	37.94	-0.9	14.92	17.38	6.42	53.95	1.56
TDS	171.05	13.9	149.76	191.04	0.21	354.98	17.35	318.27	369.28	-1.21
Na	27.24	11.19	9.26	48.17	-0.08	93.39	27.34	34.6	121.56	-1.25
K	1.83	0.5	1.08	3.03	0.42	1.91	0.38	1.48	2.7	1.1
Ca	23.29	9.56	5.21	30.24	-0.91	10.47	11.64	4.64	36.54	1.54
Mg	4.19	1.95	1.96	6.7	0.07	4.45	5.75	1.78	17.41	1.59
Fe	0.19	0.05	0.11	0.29	0.45	0.15	0.04	0.1	0.2	0.23
Mn	0.08	0.01	0.04	0.09	-1.49	0.15	0.18	0.08	0.56	1.62
ANIONS_SUM	2.22	0.66	1.26	2.98	-0.24	5.03	0.23	4.75	5.35	0.02
CATIONS_SUM	2.75	0.49	2.11	3.55	0.28	5.02	0.34	4.81	5.77	1.53

Table 4.25 – Cluster Statistics for Cluster C & D in Orginal Measurement Scale

	Cluster A		Cluster B		Cluster C		Cluster Four	
	Average	St Dev	Average	St Dev	Average	St Dev	Average	St Dev
TEMP	**-0.94**	**1.15**	**0.32**	**1.24**	**-0.29**	**0.53**	**0.88**	**0.40**
pH	**0.58**	**0.00**	**-0.88**	**0.64**	**0.69**	**1.14**	**-0.21**	**0.29**
EC	**1.74**	**0.51**	**-0.95**	**0.13**	**-0.23**	**0.06**	**0.62**	**0.15**
NH4	**-0.47**	**0.00**	**-0.22**	**0.50**	**-0.34**	**0.14**	**1.58**	**2.02**
NO2	**-0.29**	**0.00**	**0.58**	**1.73**	**-0.29**	**0.00**	**-0.29**	**0.00**
NO3	**0.26**	**1.09**	**-0.45**	**0.15**	**-0.24**	**0.40**	**1.12**	**2.35**
Cl	**1.93**	**0.94**	**-0.60**	**0.03**	**-0.36**	**0.22**	**-0.02**	**0.74**
HCO3	**0.77**	**0.52**	**-0.96**	**0.23**	**-0.06**	**0.62**	**1.27**	**1.13**
F	**0.67**	**0.54**	**-1.14**	**0.87**	**0.46**	**0.42**	**0.68**	**0.10**
HARD	**1.81**	**0.15**	**-0.82**	**0.12**	**-0.14**	**0.39**	**0.11**	**1.04**
FreeCN	**-0.29**	**0.00**	**-0.29**	**0.00**	**0.58**	**1.73**	**-0.29**	**0.00**
TDS	**1.76**	**0.65**	**-0.93**	**0.12**	**-0.24**	**0.06**	**0.57**	**0.14**
Na	**1.08**	**0.83**	**-0.89**	**0.22**	**-0.16**	**0.36**	**1.00**	**1.48**
K	**1.59**	**1.46**	**-0.83**	**0.46**	**0.01**	**0.38**	**0.04**	**0.00**
Ca	**1.82**	**0.21**	**-0.83**	**0.10**	**-0.09**	**0.50**	**0.01**	**0.91**
Mg	**1.71**	**0.05**	**-0.76**	**0.15**	**-0.30**	**0.23**	**0.42**	**1.38**
Fe	**0.18**	**0.15**	**-0.30**	**1.31**	**0.39**	**1.11**	**-0.36**	**0.92**
Mn	**-0.39**	**0.12**	**-0.30**	**0.13**	**-0.24**	**0.04**	**1.48**	**2.38**
ANIONS_SUM	**1.65**	**0.20**	**-0.96**	**0.16**	**-0.27**	**0.23**	**0.80**	**0.04**
CATIONS_SUM	**1.73**	**0.33**	**-0.98**	**0.19**	**-0.19**	**0.15**	**0.61**	**0.05**

Table 4.26 – Cluster Statistics with Centered Data

3. Profile Graphs of Clusters

The profile graphs from the results from the K-means cluster analysis have been shown below. The profile plotting (Figure 4.57) showing the average values of all the data for all the variables belong to the same cluster.

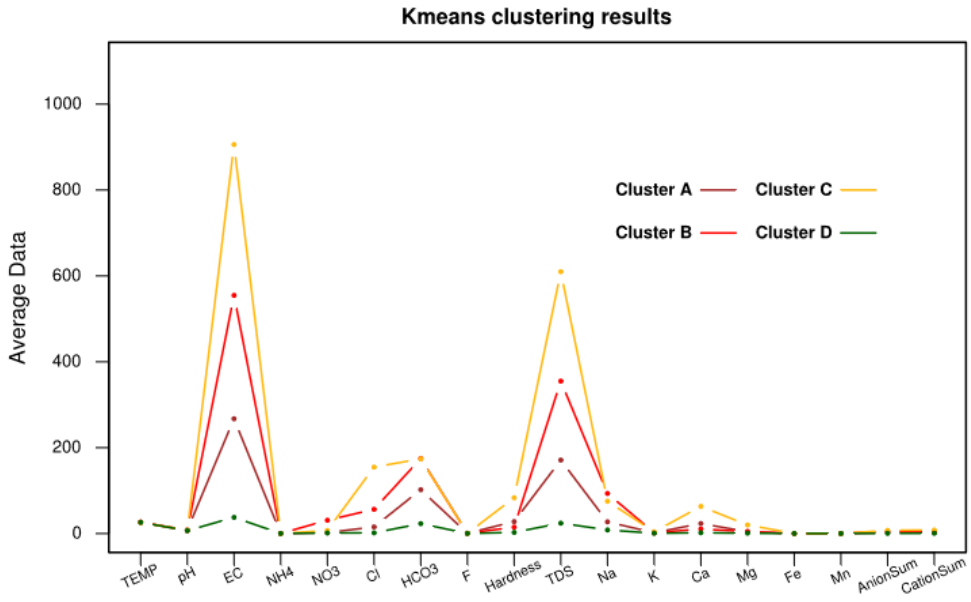

Figure 4.57 –K-Means Cluster Analysis Results Profile Plot with Original data

The profile plotting (Figure 4.57 and 4.58) showing the average values of all the original data and centered data for all the variables respectavly belong to the same cluster. Specifically profile from the centered data helped understanding the impact of each variable on each cluster. It is very interesting to see that the four clusters have very distinct behaviors. From the extreme right it is possible to observe the four clusters sequentially ordered and separated with respect to the variables say, CATIONS_SUM, ANIONS_SUM, then Mg, Ca, K, Na and so on up to HARD, TDS, EC which seemed to very logical that individual having higher values in this variables below to the cluster at the top and the having the lowest values in this variable belongs to the lowest echelon.

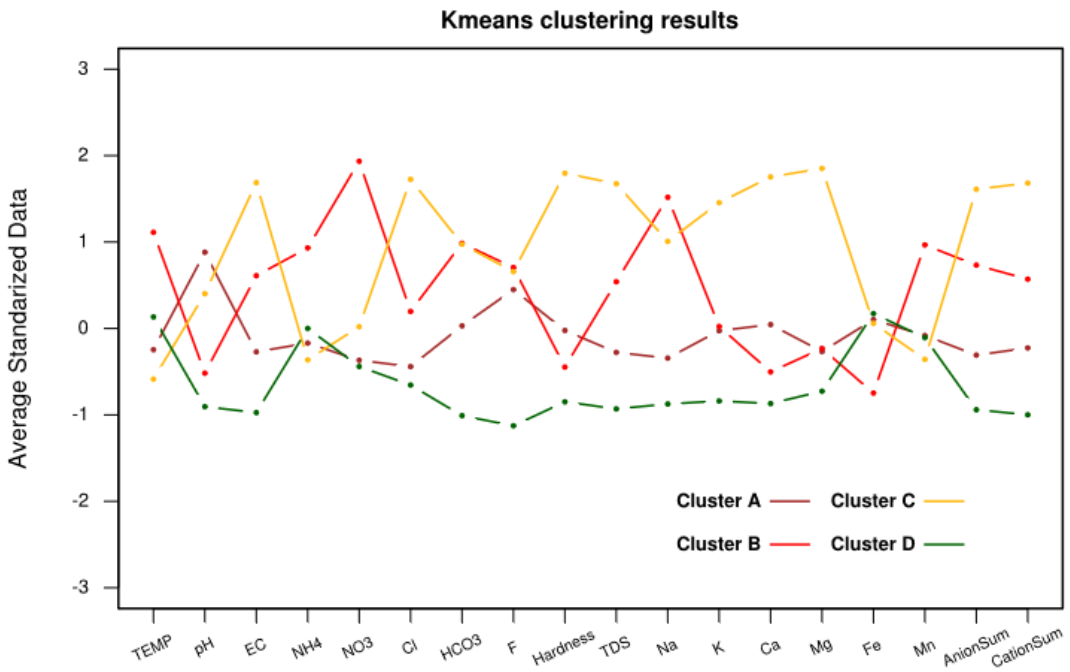

Figure 4.58 – K-Means Cluster Analysis Results Profile Plot with Centered Data

5. CONCLUSIONS

The study has clearly revealed the fact that this multivariate data analytical techniques: Principal Component Analysis, Principal Factor Analysis, Cluster Analysis have potential to be applied in acquiring improved understanding useful for industrial quality assurance quality control (QAQC), market surveillance, standardization process and or regulatory purposes as well as for interested academic and scientific communities seeking knowledge.

This study has clearly identified the presence of four distinct clusters among the bottled water products in the market of Bangladesh. And we have been able to create succesfully the profile for each cluster. From the cluster profiles as well as the classification achieved through this study we have acquired improved and detailed understanding of the general properties of the bottled water brands in the market.

From this study we may fairly draw a conclusion that this approach has a good potential to be utilized for any particular industrial in-house QAQC in setting, monitoring and controlling the product quality as an ongoing QAQC process monitoring purposes batch to batch basis after defining, confirming and validating their general quality profiles during the process validation. For any particular Brand the industry may set particular quality standards based on quality variables or indicators appropriately standardized. Then during the product development and process validation stage they may use these techniques preparing data sets for quite a big number of products from different batches and may define the profile map for each cluster for each brand or production line or production process. And then they may use this general cluster profiles and features later time during the real time production and ongoing QAQC process via matching – comparing with the general cluster profiles defined a priori.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This study did not observed presence of any contaminant. To identify any specific contaminant or any specific outlier among the marketted products further investigation may be required or application of other techniques may be explored. For detecting any contamination or spurious data or outlying product it may require some other Multivariate Techniques in addition to the existing univariate approaches. Application of various kinds of Discrimination Analytical (DA) techniques following Mahalanobis approaches should be explored in this regard.

As this study was limited to exploratory analysis and was primarily aimed to investiagte some unsupervised classification techniques for augmenting the existing knowledge to some extent as well as to gain some knowledge about the underlying physico-chemical interactions we did not yet investigated other approaches e.g. supervised classification techniques or other multivariate state-of-the-art chemoetrics techniques or pattern recognition processes including Multiple Linear Regression MLR, Cannonical Corresponding Analysis, Discriminant Analysis etc.

Ensuring utmost reliability and validity in conducting laboraty based study on physico-chemical properties of water on a very large sample size is always a costly scientific venture as well as time and resource intensive. We have used data set generated by our team in 2000-2003. Until the year 2003 in Bangladesh there were more than 37 manufacturers who had been marketing bottled water in single service containers (Khan, 2008) and in this study we have been able to analyse 51 Individuals from 12 Brands within our limted time and resource. Moreover, at the moment of sampling not all brands were available in the market. Whereas in this more than 12 years time period some of these manufacturers had gone out of production and at the same time some other new brands appeared in the market. To ensure the effective monitoring of the quality of the products as well as to gain further understanding about the temporal evolution of these bottled water brands it is essential to conduct further study with data set generated from more recent laboratory studies.

Therefore, considering the above it would be interesting to investigate further applying PCA, FA, CA, MLR, DA and other multivariate techniques to see any temporal evolution took place for any brand investigated under this study as well as to compare their status in the clusters. The results could be further improved incorporating new more brands introduced in the market in the mean time. This research team has already access to a new data set based on physico-chemical study conducted recently in 2017 on 14 bottled water brands including some new brands appeared in the market. This research team envisioned that these two data sets may jointly be used to attain a bigger sample size to gain improved understanding about the population including gaining knowledge on temporal pattern, if any.

As mentioned earlier this study was primarily aimed to be an exploratory one, and from this study we have identified four latent Principal Factors. Hence it is envisoned that we may conduct further study to develop some classification models based on Structural Equation Modeling SEM techniques upon running confirmatory factorial analysis CFA, MLR or Partial Least Square PLS technique and the like. As we came to know from other study and observation (Khan, 2008) that in Bangladesh the manufacturers are mostly applied various treatment to ground water in producing these packaged

bottled drinking water it is not unreasonable also to conduct a through investigation on ground water data set already available with this research team. This team has access to a data set consists of observations on more than 240 ground water and treated ground water samples which could also be investigated together with this bottled water brands for mining more knowledge and information about underlying physico-chemical phenomena, industrial practices, quality aspects as well as monitoring and surveillance features important for the communities be them regulatory or consumers' rights activists group or science and academia.

7. BIBLIOGRAPHY

Adeogun (2012). Adeogun, A.O., Babatunde, T.A., Chukwuka, A.V., Spatial and Temporal Variations in Water and Sediment Quality of Ona River, Ibadan, Southwest Nigeria.*European Journal of Scientific Research.ISSN 1450-216X Vol.74 No.2* (2012), pp. 186-204.

Ahmed (2005). Ahmed, S.M., Hussain, M. & Abderrahman,W. Using multivariate factor analysis to assess surface/logged water quality and source of contamination at a large irrigation project at Al-Fadhli, Eastern Province, Saudi Arabia.*Bull Eng Geol Environ (2005) 64*: 319–327. DOI 10.1007/s10064-005-0277-6.

Alam (2010). Alam, M. J. B., Ahmed, A.A.M., Ali, E., Ahmed, A. A. M.. Evaluation of Surface Water Quality of Surma River Using Factor Analysis. *Proc. of International Conference on Environmental Aspects of Bangladesh (ICEAB10)*, Japan, Sept. 2010. p186.

APHA (2012). APHA-AWWA-WEF (2012). Standard methods for examination of water and wastewater. 22th ed. Washington, District of Columbia, *American Public Health Association, American Water Works Association, Water Environment Federation*, 2012.

APHA (1998). APHA-AWWA-WPCF (1998). Standard methods for examination of water and wastewater. 20th ed. Washington, District of Columbia, *American Public Health Association, American Water Works Association, Water Environment Federation*, 1998.

Belkhiri (2010). Belkhiri, L., Boudoukha, A., Mouni,L. & Baouz, T., Multivariate statistical characterization of groundwater quality in Ain Azel plain, Algeria. *African Journal of Environmental Science and Technology Vol. 4(8),* pp. 526-534, August 2010. Available online at http://www.academicjournals.org/AJEST. ISSN 1991-637X © 2010 Academic Journals.

BSTI (1990). Bangladesh Standards & Testing Institution. Bangladesh Standard BDS 1240: 1989 Specification for Drinking Water. *Bangladesh Standards & Testing Institution (BSTI), Dhaka*, 1990.

BSTI (2000). Bangladesh Standards & Testing Institution. Bangladesh Standard BDS 1414: 2000 Specification for Natural Mineral water (First Revision). *Bangladesh Standards & Testing Institution (BSTI), Dhaka*, 2000.

Carlson (2002). Carlson, E. & Ecker, M.D. A Statistical Examination of Water Quality in Two Iowa Lakes. *American Journal of Undergraduate Research Vol.1* No.2 (2002).

Charkhabi (2006). Charkhabi, A.H. & Sakizadeh, M. Assessment of Spatial Variation of Water Quality Parameters in the Most Polluted Branch of the Anzali Wetland, Northern Iran. *Polish J. of Environ. Stud. Vol. 15,* No. 3 (2006), 395-403.

Chenini (2009). Chenini, I.; Khemiri, S., (2009).Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int. J. Environ. Sci. Tech., 6 (3)* (2009), 509-519.

CODEX (1985). FAO/CODEX. CAC/RCP 33-1985 Code of Hygienic Practice for Collecting, *Processing and Marketing of Natural Mineral Waters,* FAO/CODEX, 1985.

CODEX, (2001a). FAO/CODEX. CODEX STAN 227-2001 General Standard for Bottled/Packaged Drinking Waters (Other than Natural Mineral Waters). *FAO/ CODEX*, 2001.

CODEX (2001b). FAO/CODEX. CAC/RCP 48-2001 Code of Hygienic Practice for Bottled/Packaged Drinking Waters (Other than Natural Mineral Waters. *FAO/ CODEX*, 2001

CODEX (2008). FAO/CODEX. CODEX STAN 108-1981 CODEX Standard for Natural Mineral Waters version 1981, Revision in 2008. *FAO/CODEX, Adopted 1981. Amendment 2001, 2011. Revisions 1997, 2008.*

Cohn (1999). Cohn, P. D., Cox, M., Berger, P.S., Health and Aesthetic Aspects of Water Quality. *Water Quality & Treatment A handbook of Community Water Supplies, 5th Edition*, Publisher AWWA, 1999, Chapter-2, pp 2.1- 2.86.

Debels (2005). Debels, P., Figueroa, R., Urrutia, R., Barra, R. & Niell,X. Evaluation of Water Quality in the Chillan River (central Chile) Using Physicochemical Parameters and a Modified Water Quality Index. *Environmental Monitoring and Assessment (2005) 110:* 301–322. DOI: 10.1007/s10661-005-8064-1, Springer 2005.

Ge (2013). Ge, J., Ran, G., Miao, W., Cao, H, Wu, S. & Cheng, L. Water Quality Assessment of Gufu River in Three Gorges Reservoir (China) Using Multivariable Statistical Methods. Advance *Journal of Food Science and Technology 5*(7): 908-920, 2013. ISSN: 2042-4868; e-ISSN: 2042-4876.

Ghrefat (2013). Ghrefat, H.A. Classification and Evaluation of Commercial Bottled Drinking Waters in Saudi Arabia. *Research Journal of Environmental and Earth Sciences 5*(4): 210-218, 2013. ISSN: 2041-0484; e-ISSN: 2041-0492. Maxwell Scientific Organization, 2013

GOB (1997). Government of Peoples Republic of Bangladesh (GOB), Bangladesh Gazzet Additional August 28, 1997, Tofcil-3, Rule-12: Allowable Limit of Drinking Water and Allowable Limit of Groundwater. *Government* of Peoples Republic of Bangladesh (GOB), Dhaka, Bangladesh, 1997.

GOB (2009), Government of the Peoples Republic of Bangladesh (GOB), Bangladesh Gazzet, April 28, 2009, Bangladesh Consumers´ Rights Protection Act 2009. *Government of the Peoples Republic of Bangladesh (GOB),* Bangladesh, April, 2009.

Hossain (2013). Hossain.M.G., Reza, A.H.M.S., Lutfun-nessa,M.& Ahmed, S.S. Factor and Cluster Analysis of Water Quality Data of the Groundwater Wells of Kushtia, Bangladesh: Implication for Arsenic Enrichment and Mobilization. *Journal Geology Society of India, Vol.81,* March 2013, pp.377-384.

Inam (2010). Inam,E., Kim, K.W., Ebong, G. & Eduok, U.. Trace Elements in Ground and Packaged Water in Akwa Ibom State,Nigeria. *Geosystem Engineering, 13*(2), June 2010, pp. 57-68.

Iscen (2008). Iscen, C.F., Emiroglu, Ö.,Ilhan, S.,Arslan, N.,Yilmaz,V. & Ahiska, S., Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. *Environ Monit Assess (2008) 144*:269–276. DOI 10.1007/s10661-007-9989-3.

James (1990). Frances C. James, F.C. & McCulloch, C.E.. Multivariate Analysis in Ecology and Systematics: Panacea or Pandora´s Box? *Annu. Rev. Ecol. Syst. 1990.21*:129-166. (Downloaded from arjournals.annualreviews.org by Florida State University on 05/05/09 for personal use only).

Kaneene (2007). Kaneene, J.B.,Miller, R.A. Sayah, R., Johnson,Y. J. , Gilliland, D. & Gardiner, J.C. Considerations When Using Discriminant Function Analysis of Antimicrobial Resistance Profiles To Identify Sources of Fecal Contamination of Surface Water in Michigan. *Applied and Environmental Microbiology, May 2007, p. 2878–2890 Vol. 73*, No. 9. 0099-2240/07/$08.00_0 doi:10.1128/AEM.02376-06.

Khan (2003). Khan, M., Anwar, K.M.M.,Commercialized Bottled Water (Part-I-V), *The Daily Bangladesh Today, 3-17,* April 2003, p5.

Khan (2008). Khan, M., Anwar, K.M.M., Chowdhury, H., Study on the Commercially Available Bottled water in Bangladesh: General Survey Results, Inorganic Physico-Chemical Quality and Related Issues. *Bangladesh Academy of Sciences,* 2008.

Khater (2014). Khater, A.E.M., Al-Jaloud, A. & El-Taher, A.Quality Level of Bottled Drinking Water Consumed in Saudi Arabia.*Journal of Environmental Science and Technology 7*(2):90-106, 2014. ISSN 1994-7887/DOI:10.3923/jest.2014.90.106

Kido, M., Yustiawati, M., Syawal, S.,Sulastri, Hosokawa, T., Tanaka, S., Saito, T., Iwakuma, T. & Kurasaki, M., Comparison of general water quality of rivers in Indonesia and Japan. *A Report by Division of Environmental Science Development, Graduate School of Environmental Science, Hokkaido University.*

Kumar (2010). Kumar, M. & Singh, Y.. Interpretation of Water Quality Parameters for Villages of Sanganer Tehsil, by Using Multivariate Statistical Analysis. *J. Water Resource and Protection, 2010, 2,* 860-863. doi:10.4236/jwarp.2010.210102 published Online October 2010 (http://www.SciRP.org/journal/jwarp).

Lebart (2008). Lebart, L.. Exploratory multivariate data analysis from its origins to 1980: Nine contributions. *Journal Electronique d'Histoire des Probabilités et de la Statistique Electronic Journal for History of Probability and Statistics Vol 4,* n°2; Décembre/ December 2008. www.jehps.net.

Lourenço (2010). Lourenço, C., Ribeiro, L. &Cruz, J.. Classification of natural mineral and spring bottled waters of Portugal using Principal Component Analysis. *Journal of Geochemical Exploration 107 (*2010) 362–372.

Lewis (1999). Lewis, D.R., Southwick, J.W., Oullet-Hellstrom, R., Rench,J., Calderon, R.L., Drinking Water Arsenic in Utah: A Cohort Mortality Study, *Environmental Health Perspectives, Volume 107,* No.: 5, May 1999, pp 359-365.

Mahmood (2011). Mahmood, A., Muqbool, W., Mumtaz, M.W. & Ahmad, F. Application of Multivariate Statistical Techniques for the Characterization of Ground Water Quality of Lahore, Gujranwala and Sialkot (Pakistan). *Pak. J. Anal. Environ. Chem. Vol. 12,* No. 1 & 2 (2011) 102-112.

McNeil (2005). McNeil, V.H., Cox,M.E., Preda, M. Assessment of chemical water types and their spatial variation using multi-stage cluster analysis, Queensland, Australia. *Journal of Hydrology 310* (2005), 181-200. Doi:10.1016/i.jhydol.2004.12.014.

Murshid (2002a). Murshid, S., Anwar, KM. M., Khan, M., Water Quality Monitoring: An Overview and Recommendation, *Paper presented at 4th International Conference on "Arsenic Contamination of Ground Water in Bangladesh: Cause, Effect & Remedy"*, Dhaka, January 12-13, 2002.

Murshid (2002b). Murshid, S., Anwar, KM. M., Khan, M., On the watch for lurking danger, *The Daily Star, Vol. XII* No.83, p-9, Dhaka, April 12, 2002.

Mustapha (2011). Mustapha, A. & ARIS, A.Z., Spatial Aspect of Surface Water Quality Using Chemometric Analysis.*Journal of Applied Sciences in Environmental Sanitation, Volume 6,* Number 4: 411-426, December, 2011.

Najafpour (2008). Najafpour, Sh., Alkarkhi, A. F. M., Kadir, M. O. A. & Najafpour, Gh. D., Evaluation of Spatial and Temporal Variation in River Water Quality, *Int. J. Environ. Res., 2(*4): 349-358, Autumn 2008.

Obeidat (2011). Obeidat, S.M, Sekhaneh, W., Momani, I.A. & Hamid, A.J. A.A. Assessment of Water Quality in Four Main Water Reservoirs in Northern Jordan. *International Journal of Chemistry Vol. 3,* No. 2; June 2011.

Odagiu (2011). Odagiu, A., Oroian, I.G., Mihăiescu, T., Covrig, I. & Vârban,D. Cluster Analyze Approach in Monitoring Some Physico-Chemical Parameters of Drinking Water from Municipal Network of Cluj-Napoca Town. *Bulletin UASVM Agriculture, 68*(2)/2011. Print ISSN 1843-5246; Electronic ISSN 1843-5386.

Pati (2012). Pati.S,Dash, M.K. & Mukherjeec,C.K.. Development of Water Quality Index for assessment of quality of water in the coastal water of Bay of Bengal at Visakhapatnam zone, India. The Macrotheme Review A multidisciplinary journal of global macro trends. October 2012 1(1).

Pedersen (2003). Pedersen, H.G., Jensen, V., Short-term Consultancy Input on Laboratory Capacity in Relation to Ground Water Quality Monitoring in Bangladesh, *End-of-Mission Report, Sector Programme Support, Water and Sanitation Sector,* DANIDA, Denmark, January, 2003 (personal communication).

Ragno (2007). Ragno, G., Luca M., Ioele, G. An application of cluster analysis and multivariate classification methods to spring water monitoring data. *Microchemical Journal 87* (2007) 119–127.

Rahman (2017). Rahman, I.M.M., Barua, S., Barua, R., Mutsuddi, R., Alamgir, M., Islam, F., Begum, A.A., Hasegawa, H. Quality assessment of the non-carbonated bottled drinking water marketed in Bangladesh and comparison with tap water. *Food Control 73* (2017) 1149-1158.

Samsudin (2011). Samsudin, M.S., Juahir, H., Zain, S.M, Adnan, N.H. Surface River Water Quality Interpretation Using Environmetric Techniques: Case Study at Perlis River Basin, Malaysia. *International Journal of Environmental Protection. IJEP Vol.1* No.5 2011 PP.1-8 www.ij-ep.org ○C World Academic Publishing. DOI 10.5963/IJEP0105001.

Saravi (2011). Saravi, H.N, Makhlough, A., Pourgholam, R, Din, Z.B. & Foong, S.Y.. Multivariate analysis of water quality parameters and phytoplankton composition in the southern of Caspian Sea. *International Aquatic Research, (2011) 3:* 205-216. ISSN 2008-4935.

Sghaier (2011). Sghaier, K., Barhoumi, H., Maaref,A, Siadat, M., Jaffrezic-Renault, N., Characterization and Classification of Groundwater from Wells Using an Electronic Tongue (Kairouan, Tunisia), *Journal of Water Resource and Protection, 2011, 3,* 531-539 doi:10.4236/jwarp.2011.37063 Published Online July 2011. (http://www.SciRP.org/journal/jwarp).

Shrestha (2007). Shrestha,S.& F. Kazama, F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software 22* (2007) 464-475.

Shrestha (2008). Shrestha,S., Kazama, F.& Nakamura, T. Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River. *Journal of Hydroinformatics, Vol.10*.No.1, 2008.

Silva (2008). Silva, M.C.R., Albuquerque, M.T.D. & Ribeiro, L.. Use of Water Quality Index to Evaluate the Influence of Anthropogenic Contamination on Groundwater Chemistry of a Shallow Aquifer, Loures Vally, Lisbon, Portugal. *Global Groundwater Resources and Management (Book: Chapter – 21), Editor: B.S. Paliwal Selected Papers from The 33rd International Geological Congress, General Symposium: Hydrogeology, Oslo (Norway)* Aug. 6-14, 2008, Scientific Publishers (India), Jodhpur, pp. 347-362.

Singh (2009). Singh, S.K., Singh, C.K., Kumar, K.S., Gupta, R. & Mukherjee, S. Spatial-Temporal Monitoring of Groundwater using Multivariate Statistical Techniques in Bareilly District of Uttar Pradesh, India. *J. Hydrol. Hydromech., 57,* 2009, 1, 45–54 DOI: 10.2478/V10098-009-0005-1.

Smith (2000). Smith, A.H., Lingas, E.O. and Rahman, M., Contamination of drinking - water by arsenic in Bangladesh: a public health emergency*, Bulletin of the World Health Organization, WHO, Vol.78,* No.9, 2000 pp.1093- 1103.

Šnuderl (2007). Šnuderl, K., Simoni, M., Mocak, J., Brodnjak, D.. Multivariate Data Analysis of Natural Mineral Waters. Von~ina, *Acta Chim. Slov. 2007, 54*, 33–39.

Souza (2005). Souza, A.M., Zanini1, R.R., Moraes, A.B.D. & Malavé, C.O. Quality of water using multivariate analysis. *Ciência e Natura, UFSM, 27* (1): 7 - 18, 2005.

Swain (2012). Swain, N.R., Application of Three Statistical Pattern Recognition Techniques for Temporal and Spatial Water Quality Analysis. *Thesis submitted for Master of Science Degree, Department of Civil and Environmental Engineering, Brigham Young University, USA,* December 2012.

Tibbetts (2000). Tibbetts, J., Water World 2000, *Environmental Health Perspectives, Vol.108* No.2, February 2000, pp A69-A72.

(UNEP/WHO, 1996). United Nations Environment Programme & World Health Organization, Water Quality Monitoring, Edited by Bartram, J. & Balance, R.. *Published on behalf of UNEP and WHO by Chapman & Hall, London,* 1996, p-1.

UNIDO (2005). United Nations Industrial Development Organization. PROJECT/PROGRAMME TF/RAS/RAS/03/001 – Market Access and Trade Facilitation Support for Asian LDCs, through Strengthening Institutional and National Capacities related to Standards, Metrology, Testing and Quality (SMTQ), Technical Mission Report Assessment of Bangladesh Standards and Testing Institution and provision of, recommendations, assistance and training for standards development and product certification based on the work of Rajinder Raj Sud, UNIDO Consultant, Standards Expert (TF/RAS/RAS/03/001/11-59). *Prepared for the Government of Bangladesh by the United Nations Industrial Development Organization (UNIDO*), October 2005.

UNIDO (2008). United Nations Industrial Development Organization. Bangladesh Quality Support Programme (EE/BGD/05/002) Strategic Planning for BSTI Mission Report based on the work of Mr. P. J Bonner (EE/BGD/05/002/11- 57-2007) International Expert in SMTQ, *Prepared for the Government of Bangladesh by the United Nations Industrial Development Organization (UNIDO),* January 2008.

USEPA (1996). USEPA, Compilation of EPA's Sampling and Analysis Methods, 2nd Ed., Edited by Keith, L.H. CRC Press , Inc. USA, 1996.

USEPA (1997). USEPA, EPA Methods and Guidance for Analysis of Water on CD-ROM, *USEPA Office of Water, Washington D.C.* 20460, EPA 821-C-97-001, April 1997.

USEPA (2002). USEPA, Code of Federal Regulations 40 CFR Parts 141 and 142 National Primary Drinking Water Regulations, *USEPA,* Monday, January 14, 2002.

Van Hulle (2012). Van Hulle, S.W. H.; Ciocci, M. C., Statistical evaluation and comparison of the chemical quality of bottled water and flemish tap water, *Desalination & Water Treatment; Feb 2012, Vol. 40* Issue 1-3, p183.

WHO (2008). World Health Organization, *WHO Guidelines for drinking-water quality. Vol 1, 3rd edition.* World Health Organization, Geneva, Switzerland, 2008.p-xv

Yusuf (2013). Yusuf, K.A., Oluwole, S.O., Abdusalam, I.O. & Majolagbe, A.O., Assessment of Spatial Variation of Surface Water Quality in Lagos, Using Multivariate Statistical Techniques. *Journal of Environment (2013), Vol. 02,* Issue 04, pp. 94-102.ISSN 2049-8373.

Zhao (2009). Zhao, Z.W. & Cui, F.Y. Multivariate statistical analysis for the surface water quality of the Luan River, China. *Journal of Zhejiang University SCIENCE A. 2009 10*(1):142-148. ISSN 1673-565X (Print); ISSN 1862-1775 (Online).www.zju.edu.cn/jzus; www.springerlink.com.E-mail: jzus@zju.edu.cn.

8. ANNEXES

Annexure 1: Initial Data Matrix X (Table A1-1 and Table A1-2)

Note: Individuals from each brand has been given relatively shorter codes to mark every single observation which could be seen in the first column of the Initial Data Matrix X. To illustrate further let us consider a few examples: several observations or Individuals from BRAND02 has been denoted as B2a, B2b, B2c and so on. And Similarly, other Individuals from BRAND05 has been coded as B5a, B5b, B5c, B5d etc. For two Individuals or observations from DEIONIZEDWATER have been coded as DIWa and DIWb.

Individual	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD
DIWa	26.4	5.9	2.4	0.1	0.01	0.1	2.5	0.3	3.7	0.1	0.50
B1	26.4	7.3	497.3	0.24	0.01	0.1	2.5	4.85	301.6	0.5	53.95
B2a	26.7	7.1	555.7	0.11	0.01	46.5	2.5	64.71	157.3	0.48	6.42
B2b	26.7	6.9	568.7	0.11	0.01	26.7	2.5	62.21	153.6	0.47	7.67
B2c	27	6.9	573.3	0.11	0.01	51.6	2.5	68.03	153.6	0.48	7.76
B2d	27.1	6.9	562	0.1	0.01	23.1	2.5	67.63	157.3	0.49	7.70
B2e	26.6	6.9	549	0.22	0.01	45.7	2.5	61.06	151.7	0.45	13.67
B2f	26	6.8	577	0.1	0.01	23.1	2.5	66.53	148	0.48	7.24
B3a	26.6	8.5	297	0.15	0.01	0.3	2.5	2.4	166.5	0.4	34.51
B3b	26	8	276.3	0.11	0.01	0.4	2.5	4.49	166.5	0.42	34.43
B3c	26.2	8.5	284	0.12	0.01	0.2	2.5	8.45	162.8	0.4	35.83
B3d	26	8	291	0.1	0.01	0.4	2.5	3.41	161	0.36	34.94
B3e	25.7	8.1	298.5	0.1	0.01	0.3	2.5	3.53	166.5	0.36	34.27
B3f	25.7	8.1	297.5	0.1	0.01	0.3	2.5	3.41	166.5	0.35	34.71
B4a	26	6.8	245.3	0.12	0.01	9.6	2.5	23.46	92.5	0.46	16.34
B4b	26.4	6.8	245	0.1	0.01	9.9	2.5	23.98	85.1	0.4	8.51
B5a	26.7	7.7	81.9	0.12	0.01	4.2	2.5	3.77	55.5	0.41	10.29
B5b	25.9	7	89.6	0.19	0.01	3.8	2.5	4.59	46.3	0.41	6.78
B5c	25.8	6.5	90.7	0.1	0.01	2.3	2.5	2.35	55.5	0.31	6.74
B5d	25.8	7.6	99.7	0.1	0.01	3.5	2.5	3.25	55.5	0.45	8.01
B5e	25.6	7	90.1	0.2	0.01	3.7	2.5	4.08	46.3	0.41	6.53

Individual	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD
B6a	25.5	9	246.3	0.11	0.01	2.5	2.5	28.23	27.8	0.43	31.66
B6b	25.7	9.2	259.3	0.1	0.01	2.5	2.5	24.22	28.3	0.48	32.16
B6c	25.4	9.2	251	0.1	0.01	2.3	2.5	31.1	31.5	0.43	31.81
B6d	25.6	9.2	259.5	0.1	0.01	2.3	2.5	31.1	31.5	0.44	31.76
B6e	25.4	9.2	260	0.1	0.01	2.1	2.5	31.1	31.5	0.43	37.94
B6f	25.5	9.2	252.3	0.1	0.01	2.3	2.5	27.33	29.6	0.44	32.29
B7a	25.3	6.5	22.6	0.1	0.01	2	2.5	1.73	33.3	0.1	1.11
B7b	25.4	6.2	11	0.1	0.01	0.5	2.5	0.68	5.6	0.1	0.54
B7c	26.4	6.5	9.5	0.1	0.01	0.5	2.5	0.3	5.6	0.1	0.55
B7d	26.9	6.2	11.5	0.1	0.01	0.3	2.5	0.3	5.6	0.1	0.39
B7e	26.2	6.4	10.6	0.1	0.01	0.5	2.5	0.3	5.6	0.1	0.31
B8a	26	8	907.7	0.11	0.01	0.5	2.5	226.97	148	0.41	73.19
B8b	26.2	7.7	992.7	0.1	0.01	0.3	2.5	221.67	148	0.43	82.25
B8c	26.3	7.7	1025.3	0.1	0.01	0.4	2.5	269.9	145	0.44	82.96
B8d	25.7	7.8	1002.5	0.1	0.01	0.3	2.5	130.6	146.2	0.41	81.96
B8e	25.9	7.8	1031	0.1	0.01	0.3	2.5	131.06	145.2	0.41	81.45
B8f	25.9	7.8	1019	0.1	0.01	0.3	2.5	130.83	146.2	0.41	81.88
B9a	25.8	7.5	290.3	0.11	0.01	1.7	2.5	6.29	140.6	0.55	20.17
B9b	26	7.4	255.7	0.1	0.01	0.3	2.5	4.45	122.7	0.56	10.99
B9c	25.8	7.5	234	0.1	0.01	0.4	2.5	4.63	124	0.54	10.69
B10a	26.1	6.8	23.5	0.1	0.01	0.1	2.5	1.31	14	0.11	1.06
B10b	25.5	6.7	26.9	0.1	0.01	0.2	2.5	0.85	14.8	0.1	0.52
B10c	25.8	6.6	23.8	0.1	0.01	0.2	2.5	1.36	14.8	0.12	0.53
B10d	25.7	6.6	22.8	0.1	0.01	0.3	2.5	1.09	13	0.11	0.71
B10e	25.5	6.6	21	0.1	0.01	0.3	2.5	1.09	13	0.09	0.81
B11a	25.3	7.7	765.3	0.1	0.01	17.9	2.5	106.78	222	0.57	87.64
B11b	24.8	8.2	780.7	0.1	0.01	17	2.5	100.43	212.8	0.55	86.65
B11c	25.2	7.7	771.7	0.1	0.01	16.5	2.5	117.78	212.8	0.54	87.49

Individual	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD
B11d	25.1	7.7	763	0.1	0.01	17.2	2.5	110.89	212.8	0.54	86.58
DIWb	28.1	5.7	3	0.1	0.01	0.1	2.5	0.3	3.7	0.1	0.01

Table A1-1: Initial Data Matrix X (Continued…)

Individual	CN	COD	TDS	Na	K	Ca	Mg	Fe	Mn	AnionSum	CationSum
DIWa	0.005	20	1.54	0.03	0.05	0.25	0.03	0.20	0.04	0.13	0.03
B1	0.005	20	318.27	34.60	1.91	36.54	17.41	0.20	0.56	5.16	4.85
B2a	0.005	20	355.65	102.52	1.86	4.64	1.78	0.10	0.08	5.23	4.9
B2b	0.005	20	363.68	98.81	1.73	5.59	2.08	0.13	0.08	4.78	4.81
B2c	0.005	20	366.91	101.74	1.81	5.67	2.10	0.14	0.08	5.35	4.94
B2d	0.005	20	359.68	100.52	1.85	5.66	2.04	0.14	0.08	4.94	4.88
B2e	0.005	20	351.36	93.98	2.70	10.08	3.59	0.15	0.08	5.02	4.98
B2f	0.005	20	369.28	121.56	1.48	5.11	2.13	0.20	0.08	4.75	5.77
B3a	0.005	20	190.08	35.53	2.30	28.19	6.33	0.16	0.08	2.88	3.55
B3b	0.005	20	176.83	35.67	2.18	28.09	6.35	0.17	0.08	2.94	3.55
B3c	0.005	20	181.76	31.98	2.21	29.13	6.70	0.17	0.08	2.98	3.47
B3d	0.005	20	186.24	25.70	2.32	28.54	6.40	0.26	0.08	2.81	3.15
B3e	0.005	20	191.04	23.55	1.88	28.33	5.94	0.20	0.08	2.9	2.99
B3f	0.005	20	190.40	24.29	1.95	28.55	6.16	0.20	0.08	2.9	3.05
B4a	0.005	20	156.99	36.36	1.97	10.32	6.02	0.27	0.08	2.41	2.66
B4b	0.005	20	156.80	48.17	3.03	5.21	3.30	0.29	0.08	2.3	2.72
B5a	0.005	20	52.42	22.24	1.69	7.39	2.91	0.31	0.08	1.16	1.64
B5b	0.005	20	57.34	9.90	1.52	4.51	2.27	0.33	0.08	1.02	0.91
B5c	0.005	20	58.05	19.59	1.72	4.50	2.24	0.20	0.08	1.08	1.32
B5d	0.005	20	63.80	22.79	1.70	5.43	2.58	0.20	0.08	1.13	1.53
B5e	0.005	20	57.66	9.71	1.40	4.23	2.29	0.20	0.08	1.01	0.88
B6a	0.005	20	157.63	20.34	1.80	29.39	2.27	0.20	0.07	1.37	2.6
B6b	0.005	20	165.95	20.41	1.66	29.91	2.25	0.16	0.08	1.26	2.62
B6c	0.005	20	160.62	10.21	1.69	29.70	2.11	0.17	0.09	1.51	2.16
B6d	0.005	20	166.58	9.26	1.59	29.80	1.96	0.20	0.04	1.51	2.11
B6e	0.005	20	166.40	11.17	1.48	29.90	2.05	0.20	0.05	1.5	2.2
B6f	0.005	20	161.37	20.14	1.73	30.24	2.04	0.14	0.09	1.37	2.61
B7a	0.005	20	14.46	13.00	0.40	0.92	0.19	0.14	0.08	0.68	0.65
B7b	0.005	20	7.04	13.05	0.22	0.49	0.05	0.20	0.08	0.18	0.62
B7c	0.005	20	6.08	0.95	0.20	0.51	0.04	0.20	0.08	0.17	0.09
B7d	0.005	20	7.36	1.25	0.08	0.38	0.03	0.15	0.08	0.16	0.09
B7e	0.005	20	6.78	0.87	0.13	0.27	0.04	0.20	0.08	0.17	0.07
B8a	0.005	20	580.93	97.04	2.75	54.45	18.74	0.31	0.08	8.91	8.57
B8b	0.005	20	635.33	90.93	2.78	61.89	20.36	0.13	0.06	8.76	8.8
B8c	0.005	20	656.19	93.35	2.72	62.52	20.44	0.14	0.07	10.07	8.94
B8d	0.005	20	641.90	90.21	2.77	61.27	20.42	0.14	0.06	6.16	8.75
B8e	0.005	20	959.64	85.88	2.69	61.49	19.97	0.20	0.06	6.16	8.53
B8f	0.005	20	652.16	86.52	2.70	61.34	20.54	0.20	0.06	6.17	8.6
B9a	0.005	20	185.79	38.75	1.19	15.01	5.16	0.11	0.08	2.59	2.9
B9b	0.005	20	163.65	35.88	1.09	7.83	3.15	0.20	0.07	2.22	2.25
B9c	0.005	20	149.76	35.66	1.08	7.73	2.96	0.20	0.07	2.25	2.22

Individual	CN	COD	TDS	Na	K	Ca	Mg	Fe	Mn	AnionSum	CationSum
B10a	0.005	20	15.04	3.98	0.33	0.86	0.20	0.20	0.05	0.33	0.26
B10b	0.005	20	17.22	6.65	0.23	0.41	0.11	0.12	0.07	0.33	0.34
B10c	0.005	20	15.23	6.56	0.33	0.45	0.08	0.11	0.08	0.34	0.34
B10d	0.005	20	15.59	5.09	0.35	0.56	0.15	0.20	0.08	0.31	0.29
B10e	0.005	20	13.44	6.70	0.31	0.66	0.16	0.20	0.06	0.31	0.36
B11a	0.005	20	489.79	51.30	5.90	67.87	19.77	0.20	0.03	7.02	7.41
B11b	0.005	20	499.65	51.54	6.03	66.95	19.70	0.20	0.03	6.68	7.37
B11c	0.005	20	493.44	51.63	6.09	67.91	19.57	0.20	0.04	7.16	7.42
B11d	0.005	20	488.32	54.12	5.84	67.17	19.01	0.20	0.07	6.97	7.44
DIWb	0.005	20	1.92	0.09	0.05	0.25	0.03	0.20	0.07	0.13	0.03

Table A1-2: Initial Data Matrix X

Annexure 2: Centered or Standardized Data Matrix X (Table A2-1 & Table A2-2)

Individual	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD
DIWa	0.41	-1.54	-353.00	-0.01	0.00	-6.72	0.00	-43.45	-96.08	-0.27	-28.03
B1	0.41	-0.14	141.90	0.13	0.00	-6.72	0.00	-38.90	201.82	0.13	25.43
B2a	0.71	-0.34	200.30	0.00	0.00	39.68	0.00	20.96	57.52	0.11	-22.11
B2b	0.71	-0.54	213.30	0.00	0.00	19.88	0.00	18.46	53.82	0.10	-20.85
B2c	1.01	-0.54	217.90	0.00	0.00	44.78	0.00	24.28	53.82	0.11	-20.76
B2d	1.11	-0.54	206.60	-0.01	0.00	16.28	0.00	23.88	57.52	0.12	-20.83
B2e	0.61	-0.54	193.60	0.11	0.00	38.88	0.00	17.31	51.92	0.08	-14.86
B2f	0.01	-0.64	221.60	-0.01	0.00	16.28	0.00	22.78	48.22	0.11	-21.28
B3a	0.61	1.06	-58.40	0.04	0.00	-6.52	0.00	-41.35	66.72	0.03	5.99
B3b	0.01	0.56	-79.10	0.00	0.00	-6.42	0.00	-39.26	66.72	0.05	5.91
B3c	0.21	1.06	-71.40	0.01	0.00	-6.62	0.00	-35.30	63.02	0.03	7.30
B3d	0.01	0.56	-64.40	-0.01	0.00	-6.42	0.00	-40.34	61.22	-0.01	6.41
B3e	-0.29	0.66	-56.90	-0.01	0.00	-6.52	0.00	-40.22	66.72	-0.01	5.74
B3f	-0.29	0.66	-57.90	-0.01	0.00	-6.52	0.00	-40.34	66.72	-0.02	6.18
B4a	0.01	-0.64	-110.10	0.01	0.00	2.78	0.00	-20.29	-7.28	0.09	-12.18
B4b	0.41	-0.64	-110.40	-0.01	0.00	3.08	0.00	-19.77	-14.68	0.03	-20.02
B5a	0.71	0.26	-273.50	0.01	0.00	-2.62	0.00	-39.98	-44.28	0.04	-18.23
B5b	-0.09	-0.44	-265.80	0.08	0.00	-3.02	0.00	-39.16	-53.48	0.04	-21.74
B5c	-0.19	-0.94	-264.70	-0.01	0.00	-4.52	0.00	-41.40	-44.28	-0.06	-21.78
B5d	-0.19	0.16	-255.70	-0.01	0.00	-3.32	0.00	-40.50	-44.28	0.08	-20.51
B5e	-0.39	-0.44	-265.30	0.09	0.00	-3.12	0.00	-39.67	-53.48	0.04	-22.00
B6a	-0.49	1.56	-109.10	0.00	0.00	-4.32	0.00	-15.52	-71.98	0.06	3.14
B6b	-0.29	1.76	-96.10	-0.01	0.00	-4.32	0.00	-19.53	-71.48	0.11	3.64
B6c	-0.59	1.76	-104.40	-0.01	0.00	-4.52	0.00	-12.65	-68.28	0.06	3.28
B6d	-0.39	1.76	-95.90	-0.01	0.00	-4.52	0.00	-12.65	-68.28	0.07	3.23
B6e	-0.59	1.76	-95.40	-0.01	0.00	-4.72	0.00	-12.65	-68.28	0.06	9.42
B6f	-0.49	1.76	-103.10	-0.01	0.00	-4.52	0.00	-16.42	-70.18	0.07	3.76
B7a	-0.69	-0.94	-332.80	-0.01	0.00	-4.82	0.00	-42.02	-66.48	-0.27	-27.42
B7b	-0.59	-1.24	-344.40	-0.01	0.00	-6.32	0.00	-43.07	-94.18	-0.27	-27.98
B7c	0.41	-0.94	-345.90	-0.01	0.00	-6.32	0.00	-43.45	-94.18	-0.27	-27.98
B7d	0.91	-1.24	-343.90	-0.01	0.00	-6.52	0.00	-43.45	-94.18	-0.27	-28.14
B7e	0.21	-1.04	-344.80	-0.01	0.00	-6.32	0.00	-43.45	-94.18	-0.27	-28.22
B8a	0.01	0.56	552.30	0.00	0.00	-6.32	0.00	183.22	48.22	0.04	44.66

Individual	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD
B8b	0.21	0.26	637.30	-0.01	0.00	-6.52	0.00	177.92	48.22	0.06	53.72
B8c	0.31	0.26	669.90	-0.01	0.00	-6.42	0.00	226.15	45.22	0.07	54.43
B8d	-0.29	0.36	647.10	-0.01	0.00	-6.52	0.00	86.85	46.42	0.04	53.44
B8e	-0.09	0.36	675.60	-0.01	0.00	-6.52	0.00	87.31	45.42	0.04	52.93
B8f	-0.09	0.36	663.60	-0.01	0.00	-6.52	0.00	87.08	46.42	0.04	53.35
B9a	-0.19	0.06	-65.10	0.00	0.00	-5.12	0.00	-37.46	40.82	0.18	-8.36
B9b	0.01	-0.04	-99.70	-0.01	0.00	-6.52	0.00	-39.30	22.92	0.19	-17.54
B9c	-0.19	0.06	-121.40	-0.01	0.00	-6.42	0.00	-39.12	24.22	0.17	-17.84
B10a	0.11	-0.64	-331.90	-0.01	0.00	-6.72	0.00	-42.44	-85.78	-0.26	-27.46
B10b	-0.49	-0.74	-328.50	-0.01	0.00	-6.62	0.00	-42.90	-84.98	-0.27	-28.00
B10c	-0.19	-0.84	-331.60	-0.01	0.00	-6.62	0.00	-42.39	-84.98	-0.25	-28.00
B10d	-0.29	-0.84	-332.60	-0.01	0.00	-6.52	0.00	-42.66	-86.78	-0.26	-27.82
B10e	-0.49	-0.84	-334.40	-0.01	0.00	-6.52	0.00	-42.66	-86.78	-0.28	-27.71
B11a	-0.69	0.26	409.90	-0.01	0.00	11.08	0.00	63.03	122.22	0.20	59.11
B11b	-1.19	0.76	425.30	-0.01	0.00	10.18	0.00	56.68	113.02	0.18	58.12
B11c	-0.79	0.26	416.30	-0.01	0.00	9.68	0.00	74.03	113.02	0.17	58.96
B11d	-0.89	0.26	407.60	-0.01	0.00	10.38	0.00	67.14	113.02	0.17	58.05
DIWb	2.11	-1.74	-352.40	-0.01	0.00	-6.72	0.00	-43.45	-96.08	-0.27	-28.52

Table A2-1: Centered Data Matrix X_CENTERED (Continued…)

Individual	CN	COD	TDS	Na	K	Ca	Mg	Fe	Mn	ANIONS_ SUM	CATIONS_ SUM
DIWa	0.00	0.00	-231.82	-39.42	-1.82	-21.99	-6.14	0.01	-0.04	-2.92	-3.37
B1	0.00	0.00	84.92	-4.85	0.04	14.31	11.25	0.01	0.48	2.11	1.45
B2a	0.00	0.00	122.29	63.08	-0.01	-17.60	-4.38	-0.09	0.00	2.18	1.50
B2b	0.00	0.00	130.33	59.37	-0.14	-16.65	-4.08	-0.06	0.00	1.73	1.41
B2c	0.00	0.00	133.56	62.30	-0.06	-16.57	-4.06	-0.05	0.00	2.30	1.54
B2d	0.00	0.00	126.33	61.08	-0.02	-16.58	-4.12	-0.05	0.00	1.89	1.48
B2e	0.00	0.00	118.01	54.54	0.83	-12.16	-2.58	-0.04	0.00	1.97	1.58
B2f	0.00	0.00	135.93	82.12	-0.40	-17.13	-4.03	0.01	0.00	1.70	2.37
B3a	0.00	0.00	-43.27	-3.92	0.43	5.95	0.17	-0.03	0.00	-0.17	0.15
B3b	0.00	0.00	-56.52	-3.78	0.30	5.85	0.19	-0.02	0.00	-0.11	0.15
B3c	0.00	0.00	-51.59	-7.47	0.34	6.89	0.53	-0.02	0.00	-0.07	0.07
B3d	0.00	0.00	-47.11	-13.74	0.45	6.30	0.24	0.07	0.00	-0.24	-0.25
B3e	0.00	0.00	-42.31	-15.90	0.01	6.09	-0.22	0.01	0.00	-0.15	-0.41
B3f	0.00	0.00	-42.95	-15.16	0.07	6.31	0.00	0.01	0.00	-0.15	-0.35
B4a	0.00	0.00	-76.36	-3.09	0.10	-11.91	-0.14	0.08	0.00	-0.64	-0.74
B4b	0.00	0.00	-76.55	8.72	1.16	-17.03	-2.86	0.10	0.00	-0.75	-0.68
B5a	0.00	0.00	-180.93	-17.21	-0.18	-14.85	-3.26	0.12	0.00	-1.89	-1.76
B5b	0.00	0.00	-176.01	-29.55	-0.36	-17.72	-3.89	0.14	0.00	-2.03	-2.49
B5c	0.00	0.00	-175.31	-19.86	-0.15	-17.74	-3.92	0.01	0.00	-1.97	-2.08
B5d	0.00	0.00	-169.55	-16.66	-0.18	-16.80	-3.58	0.01	0.00	-1.92	-1.87
B5e	0.00	0.00	-175.69	-29.74	-0.47	-18.01	-3.87	0.01	0.00	-2.04	-2.52
B6a	0.00	0.00	-75.72	-19.11	-0.08	7.15	-3.89	0.01	-0.01	-1.68	-0.80
B6b	0.00	0.00	-67.40	-19.04	-0.21	7.67	-3.91	-0.03	0.00	-1.79	-0.78
B6c	0.00	0.00	-72.73	-29.23	-0.18	7.46	-4.05	-0.02	0.01	-1.54	-1.24
B6d	0.00	0.00	-66.77	-30.19	-0.29	7.56	-4.20	0.01	-0.04	-1.54	-1.29
B6e	0.00	0.00	-66.95	-28.28	-0.39	7.66	-4.12	0.01	-0.03	-1.55	-1.20
B6f	0.00	0.00	-71.98	-19.30	-0.14	8.01	-4.12	-0.05	0.01	-1.68	-0.79

Individual	CN	COD	TDS	Na	K	Ca	Mg	Fe	Mn	ANIONS_ SUM	CATIONS_ SUM
B7a	0.00	0.00	-218.89	-26.44	-1.48	-21.32	-5.97	-0.05	0.00	-2.37	-2.75
B7b	0.00	0.00	-226.31	-26.39	-1.65	-21.75	-6.11	0.01	0.00	-2.87	-2.78
B7c	0.00	0.00	-227.27	-38.50	-1.67	-21.73	-6.12	0.01	0.00	-2.88	-3.31
B7d	0.00	0.00	-225.99	-38.19	-1.79	-21.86	-6.14	-0.04	0.00	-2.89	-3.31
B7e	0.00	0.00	-226.57	-38.58	-1.74	-21.97	-6.12	0.01	0.00	-2.88	-3.33
B8a	0.00	0.00	347.57	57.60	0.87	32.21	12.58	0.12	0.00	5.86	5.17
B8b	0.00	0.00	401.98	51.49	0.91	39.65	14.20	-0.06	-0.02	5.71	5.40
B8c	0.00	0.00	422.84	53.91	0.85	40.28	14.28	-0.05	-0.01	7.02	5.54
B8d	0.00	0.00	408.55	50.76	0.89	39.04	14.26	-0.05	-0.02	3.11	5.35
B8e	0.00	0.00	726.29	46.44	0.82	39.25	13.80	0.01	-0.02	3.11	5.13
B8f	0.00	0.00	418.81	47.07	0.83	39.10	14.38	0.01	-0.02	3.12	5.20
B9a	0.00	0.00	-47.56	-0.70	-0.68	-7.23	-1.00	-0.08	0.00	-0.46	-0.50
B9b	0.00	0.00	-69.70	-3.57	-0.79	-14.40	-3.01	0.01	-0.01	-0.83	-1.15
B9c	0.00	0.00	-83.59	-3.79	-0.79	-14.51	-3.20	0.01	-0.01	-0.80	-1.18
B10a	0.00	0.00	-218.31	-35.47	-1.54	-21.38	-5.96	0.01	-0.03	-2.72	-3.14
B10b	0.00	0.00	-216.14	-32.79	-1.64	-21.83	-6.05	-0.07	-0.01	-2.72	-3.06
B10c	0.00	0.00	-218.12	-32.89	-1.54	-21.79	-6.08	-0.08	0.00	-2.71	-3.06
B10d	0.00	0.00	-217.76	-34.36	-1.52	-21.68	-6.01	0.01	0.00	-2.74	-3.11
B10e	0.00	0.00	-219.91	-32.75	-1.56	-21.58	-6.00	0.01	-0.02	-2.74	-3.04
B11a	0.00	0.00	256.44	11.85	4.02	45.64	13.60	0.01	-0.05	3.97	4.01
B11b	0.00	0.00	266.29	12.09	4.16	44.71	13.54	0.01	-0.05	3.63	3.97
B11c	0.00	0.00	260.09	12.19	4.22	45.68	13.41	0.01	-0.04	4.11	4.02
B11d	0.00	0.00	254.97	14.68	3.97	44.94	12.85	0.01	-0.01	3.92	4.04
DIWb	0.00	0.00	-231.43	-39.36	-1.82	-21.99	-6.14	0.01	-0.01	-2.92	-3.37

Table A2-2: Centered Data Matrix X_CENTERED

Annexure 3: Results from Descriptive Statistical Analysis (Table A3-1 and Table A3-2)

	TEMP	pH	EC	NH4	NO2	NO3	SO4	Cl	HCO3	F	HARD
Minimun	24.8	5.7	2.4	0.1	0.01	0.1	2.5	0.3	3.7	0.09	0
First quartile	25.6	6.8	85.75	0.1	0.01	0.3	2.5	2.38	28.95	0.33	6.47
Median	25.9	7.5	259.5	0.1	0.01	0.5	2.5	6.29	122.7	0.41	13.67
Mean	25.99	7.44	355.4	0.11	0.01	6.82	2.5	43.75	99.78	0.37	28.53
Third quartile	26.35	8	565.4	0.11	0.01	4	2.5	65.62	155.5	0.48	35.38
Maximun	28.1	9.2	1031	0.24	0.01	51.6	2.5	269.9	301.6	0.57	87.64
STDEV	0.59	0.91	326.3	0.03	0	12.48	0	64.31	76.04	0.16	30.45

Table A3-1: Descriptive Statistics (Continued….)

	CN	COD	TDS	Na	K	Ca	Mg	Fe	Mn	ANIONS_ SUM	CATIONS_ SUM
Minimun	0	20	1.54	0.02	0.05	0.25	0.02	0.1	0.03	0.13	0.03
First quartile	0	20	54.88	10.05	1.08	4.37	1.87	0.15	0.07	1.02	0.9
Median	0	20	166.4	25.7	1.73	10.08	2.29	0.2	0.08	2.3	2.66
Mean	0	20	233.35	39.45	1.87	22.24	6.16	0.19	0.08	3.05	3.4
Third quartile	0	20	361.68	52.88	2.31	29.9	6.38	0.2	0.08	5.09	4.92
Maximun	0	20	959.64	121.6	6.09	67.91	20.54	0.33	0.56	10.07	8.94
STDEV	0	0	224.81	35.55	1.48	23.41	7.4	0.05	0.07	2.7	2.85

Table A3-2: Results from Descriptive Statistical Analysis

Annexure 4: Residual Matrix for Variables containing loadings of the variabkes along the rest 14 PCs

Variables	PA5	PA6	PA7	PA8	PA9	PA10	PA11	PA12	PA13	PA14	PA15	PA16	PA17	PA18
TEMP	0.09	0.46	-0.19	0.31	0.02	-0.02	-0.02	-0.01	0	0	0	0	0	0
pH	-0.22	0.45	0.1	0.05	0.12	0.09	0.08	-0.01	-0.01	0.02	0	0	0	0
EC	-0.03	0.04	0.06	-0.05	-0.01	0.05	-0.06	0.01	0.01	-0.01	-0.05	0	0	0
NH4	0.01	-0.09	0.41	0.19	-0.14	0.04	0.02	-0.01	0	0	0	0	0	0
NO3	0.15	-0.22	0.07	0.07	0.24	0.04	-0.02	0.06	0.02	0.01	0	0	0	0
Cl	0.04	0.07	0.23	-0.02	0.08	-0.18	0.11	0.01	-0.02	0	0	0	0	0
HCO3	-0.04	-0.11	-0.34	-0.03	-0.13	0.1	0.15	0.05	-0.02	0	0	0	0	0
F	0.08	0.22	-0.06	-0.1	-0.15	-0.2	-0.11	0.04	0.01	-0.01	0	0	0	0
HARD	-0.06	-0.03	-0.02	0.13	0.04	0.02	-0.03	0.02	0.03	-0.02	0.01	0.02	0	0
TDS	-0.03	0.06	0.07	-0.06	-0.02	0.12	-0.15	0.03	-0.09	0.01	0.01	0	0	0
Na	0.05	0.1	0.07	-0.22	-0.05	0.07	0.02	-0.1	0.03	0	0.01	0	0	0
K	0.23	-0.25	-0.15	0.2	0.06	-0.07	-0.02	-0.12	-0.04	0	-0.01	0	0	0
Ca	-0.08	0	-0.02	0.15	0.06	0.03	-0.02	0.02	0.03	-0.04	0.01	-0.01	0	0
Mg	0	-0.11	-0.05	0.06	-0.04	-0.02	-0.05	0.02	0.05	0.07	0	0	0	0
Fe	0.85	0.14	0	-0.13	0.08	0.05	0.01	0.02	0	0	0	0	0	0
Mn	-0.32	0.01	-0.12	-0.21	0.23	-0.06	-0.04	-0.02	-0.01	0	0	0	0	0
ANIONS_SUM	0.02	-0.02	0.01	-0.03	0.01	-0.07	0.14	0.03	-0.02	0	0	0	0	0
CATIONS_SUM	0	0.03	0.02	-0.04	-0.01	0.05	-0.01	-0.04	0.04	0	0.01	0	0	0

Table A4: Residual Matrix for Variables

Annexure 5: Brandwise Absolute (CTA) and Relative (CTRX1000) Contributions to build the Four (04) Principal Components

	PC1	CTA	CTR	PC2	CTA	CTR	PC3	CTA	CTR	PC4	CTA	CTR	% Inertia Explained by First Four PCs	Remarks
BRAND01	2.57	54	204	5	797	**769**	0.71	17	16	-0.53	12	9	99.8	Very Well Explained
BRAND02	1.29	14	82	-0.76	19	29	2.65	241	**343**	2.6	300	**331**	78.5	Well Explained
BRAND03	0.24	0	22	0.15	1	8	-0.8	22	**230**	-0.72	23	187	44.7	Fairly Explained
BRAND04	-0.52	2	15	0.08	0	0	-1.66	94	156	3.21	457	**585**	75.6	Well Explained
BRAND05	5.53	248	**774**	-1.54	76	60	1.18	48	35	-0.97	42	24	89.3	Very Well Explained
BRAND06	-1.76	25	**420**	0.57	10	44	-1.42	69	**272**	0.42	8	24	76	Well Explained
BRAND07	-0.34	1	13	-0.85	23	78	-1.85	117	**371**	-1.25	69	**170**	63.2	Well Explained
BRAND08	-0.65	3	111	-0.09	0	2	-0.52	9	73	-0.09	0	2	18.8	Poorly Explained
BRAND09	-3.64	107	**856**	-0.29	3	5	-0.29	3	6	-0.41	7	11	87.8	Very Well Explained
BRAND10	-3.46	97	**804**	-0.48	7	15	-0.61	13	25	-0.67	20	30	87.4	Very Well Explained
BRAND11	5.61	255	**830**	-1.35	58	48	-0.61	13	10	-0.58	15	9	89.7	Very Well Explained
DEIONIZEDWATER	-4.88	193	**618**	-0.45	6	5	3.22	354	**268**	-1.01	45	26	91.7	Very Well Explained

Table A5: Brandwise Absolute (CTA) and Relative (CTRX1000) Contributions to build the Four Principal Components

Annexure 6: Euclidean Distances of Individuals from the Centroid in the Principal Component Space

Individual	Euclidean Distance from the Centre of Intertia with respect to Principal Components
DIWa	4.4106
B1	8.87125
B2a	4.76117
B2b	3.53311
B2c	4.99225
B2d	3.74422
B2e	5.45813
B2f	3.51086
B3a	2.47146
B3b	1.53894
B3c	1.84507
B3d	2.09138

Individual	Euclidean Distance from the Centre of Intertia with respect to Principal Components
B3e	1.68637
B3f	1.68722
B4a	2.12767
B4b	2.77701
B5a	3.44825
B5b	4.51838
B5c	2.55682
B5d	2.25162
B5e	3.83524
B6a	2.52804
B6b	2.75475
B6c	2.85174
B6d	2.81621
B6e	2.88112
B6f	2.88939
B7a	3.94557
B7b	4.10851
B7c	4.08283
B7d	4.49305
B7e	4.08723
B8a	6.24085
B8b	6.18819
B8c	6.81172
B8d	5.36399
B8e	5.8582
B8f	5.24925
B9a	2.31586
B9b	2.03813
B9c	2.02271
B10a	3.80208
B10b	4.14167
B10c	4.09717
B10d	3.84673
B10e	3.96247
B11a	5.85633
B11b	6.05493
B11c	5.89797
B11d	5.74983
DlWb	5.70377

Table A6: Euclidean Distances of Individuals from the Centroid in the Principal Component Space

Annexure 7:

R Codes for Principal Component Analysis and Cluster Analysis (Developed by K.M.Mostafa Anwar)

==========

```
##  Multivariate exploratory data analysis using R packages "FactoMineR".

# Install and load required libraries

# factor extra is not available in CRAN, install as follow:

library("devtools")
install_github("kassambara/factoextra")

library(FactoMineR)
library(factoextra)
library("corrplot")
library(ecodist)
library(ggdendro)
library("e1071")


# If one of the previos lines provide a error, it must be checked that the packages are instaled in the machine
# To install the missing packages, it is required to run the function install.packages, something like this:

# install.packages("packages_name")

##  Read the data table with headers and brand in row names

qwater_data <- read.table("bottled_water_data.csv", sep=",", header=T, row.names=1)

# Calculate simple statistic (minimun, maximun, median, mean, standard deviation and first and third quartile).

stat_qwater <- data.frame(
  Min = apply(qwater_data, 2, min), # minimum
  Q1 = apply(qwater_data, 2, quantile, 1/4), # First quartile
  Med = apply(qwater_data, 2, median), # median
  Mean = apply(qwater_data, 2, mean), # mean
  Q3 = apply(qwater_data, 2, quantile, 3/4), # Third quartile
  Max = apply(qwater_data, 2, max), # Maximum
  sd_qw = apply(qwater_data, 2, sd)
  )

## format setting "round two decimals"

stat_qwater <- round(stat_qwater, 2)
```

```
print(stat_qwater)  # this R object "stat_qwater" stored the previous descriptive statistical analysis results.

write.csv(stat_qwater, "statatistic_qwater.csv") # write statistic results table in csv format

##  Important Note 1:
##  The standard deviation is zero for "NO2", "SO4", "CN" and "COD" for qwater_data.
##  As the value of these variables NO2, SO4, CN, COD is constant, the standard deviation is zero
##  for this razon we exclude it in the following analysis

qwater_data2  <- qwater_data[,c(1:4, 6, 8:11, 14:22)]  # generate a second data set excluding ("NO2", "SO4", "CN" and "COD" variables).

# correlation matrix

cor_wq2 <- cor(qwater_data2)

write.csv(cor_wq2, "correlation_matrix.csv")  # write correlation matrix table as CSV file

# Visualize the correlation matrix using a correlogram

svg("correlogram_correlation_matrix.svg", width= 6.5, height= 6.5)
corrplot(cor_wq2, type="upper", order="hclust",
       tl.col="black", tl.srt=45)
dev.off()

############
## Important Note 2: In PCA we should use standarized data because the measuring scale
## for each variable is different.

## Center data matrix

m_center_qw <- scale(qwater_data2, scale=FALSE)  # Here we used "centered parameter"

write.csv(m_center_qw, "center_matrix_data.csv") # write center data matrix


# Performing Principal Component Analysis (PCA)
# Here using the default parameter the data are scaled to unit variance

qwater.pca <- PCA(qwater_data2, ncp = 4, graph=FALSE)

## The previous command stores the PCA results in the object "qwater.pca".
## the output of the function is a list including:

print(qwater.pca)
```

```
#############################################################################################
## This is an addition to read the results from PCA analysis and improve the overall graphic
representation etc.
## for generating a color vector, all individuals from the same brand take the same color, it is to
immprove visualization results

vec.col <- c("red", "brown", rep("blue", 6), rep("black", 6), rep("darkgreen",2),  rep("purple",5),
             rep("indianred",6), rep("darkgoldenrod4",5), rep("aquamarine4",6), rep("coral",3),
             rep("darkred",5), rep("orangered", 4), "red")

#############################################################################################
### Producing Individual factor maps for all the possible axis combination. They are produced in the
# SVG format
#############################################################################################
##############################

svg("individuals_mapaxes12_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(1,2), cex=.8, choix="ind", cex.axis=0.9, las=1, cex.lab=0.9, cex.main=1,
         habillage="ind", col.hab=vec.col, autoLab="yes")
mtext("PC1/Factor1 Material Loading", side=1, line=2)
mtext("PC2/Factor2 Rare/Extraneous Material Loading", side=2, line=4.15)
dev.off()

svg("individuals_mapaxes13_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(1,3), cex=.8, choix="ind", cex.axis=0.9, las=1, cex.lab=0.9, cex.main=1,
          habillage="ind", col.hab=vec.col, autoLab="yes")
mtext("PC1/Factor1 Material Loading", side=1, line=2)
mtext("PC3/Factor3 Aggregate Qualitative Feature", side=2, line=4.15)
dev.off()

svg("individuals_mapaxes14_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(1,4), cex=.7, choix="ind", cex.axis=0.9, las=1, cex.lab=0.8, cex.main=1,
          habillage="ind", col.hab=vec.col, autoLab="yes")
mtext("PC1/Factor1 Material Loading", side=1, line=2)
mtext("PC4/Factor4 Aesthetic Acceptability", side=2, line=4.15)
dev.off()

svg("individuals_mapaxes23_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(2,3), cex=.8, choix="ind", cex.axis=0.9, las=1, cex.lab=0.9, cex.main=1,
          habillage="ind", col.hab=vec.col, autoLab="yes")
mtext("PC2/Factor2 Rare/Extraneous Material Loading", side=1, line=2)
mtext("PC3/Factor3 Aggregate Qualitative Feature", side=2, line=4.15)
dev.off()

svg("individuals_mapaxes24_nv.svg", width= 7, height= 6)
```

```
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(2,4), cex=.8, choix="ind", cex.axis=0.9, las=1, cex.lab=0.9, cex.main=1,
        habillage="ind", col.hab=vec.col, autoLab="yes")
mtext("PC2/Factor2 Rare/Extraneous Material Loading", side=1, line=2)
mtext("PC4/Factor4 Aesthetic Acceptability", side=2, line=4.15)
dev.off()

svg("individuals_mapaxes34_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(3,4), cex=.8, choix="ind", cex.axis=0.9, las=1, cex.lab=0.9, cex.main=1,
        habillage="ind", col.hab=vec.col, autoLab="yes")
mtext("PC3/Factor3 Aggregate Qualitative Feature", side=1, line=4.15)
mtext("PC4/Factor4 Aesthetic Acceptability", side=2, line=4.15)
dev.off()

### Important Note 3:In the previous Individual factor maps, as there are many inidviduals (51) it
was too hard to avoid
##  label overlap, hence we set the parameter autoLab "yes" which improved the graphics a little bit.
##  But to improve the result further maybe we need to move a little bit the labels in a vectorial
program(Inkskape, Illustrator)


##############################################################################################
### Producing Variable factor maps for all the possible axis combination. They are produced in the
#### SVG format
##############################################################################################

svg("variables_mapaxes12_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(1,2), cex=.85, choix="var", cex.axis=0.8, las=1, cex.lab=1,
        cex.main=1, autoLab="yes", col.var="blue")
mtext("PC1/Factor1 Material Loading", side=1, line=2)
mtext("PC2/Factor2 Rare/Extraneous Material Loading", side=2, line=4.15)
dev.off()

svg("variables_mapaxes13_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(1,3), cex=.85, choix="var", cex.axis=0.8, las=1, cex.lab=1,
        cex.main=1,  autoLab="yes", col.var="blue")
mtext("PC1/Factor1 Material Loading", side=1, line=2)
mtext("PC3/Factor3 Aggregate Qualitative Feature", side=2, line=4.15)
dev.off()

svg("variables_mapaxes14_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(1,4), cex=.85, choix="var", cex.axis=0.8, las=1, cex.lab=1,
        cex.main=1,  autoLab="yes", col.var="blue")
mtext("PC1/Factor1 Material Loading", side=1, line=2)
mtext("PC4/Factor4 Aesthetic Acceptability", side=2, line=4.15)
```

```
dev.off()

svg("variables_mapaxes23_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(2,3), cex=.85, choix="var", cex.axis=0.8, las=1, cex.lab=1,
         cex.main=1,  autoLab="yes", col.var="blue")
mtext("PC2/Factor2 Rare/Extraneous Material Loading", side=1, line=2)
mtext("PC3/Factor3 Aggregate Qualitative Feature", side=2, line=4.15)
dev.off()

svg("variables_mapaxes24_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(2,4), cex=.85, choix="var", cex.axis=0.8, las=1, cex.lab=1,
         cex.main=1,  autoLab="yes", col.var="blue")
mtext("PC2/Factor2 Rare/Extraneous Material Loading", side=1, line=2)
mtext("PC4/Factor4 Aesthetic Acceptability", side=2, line=4.15)
dev.off()

svg("variables_mapaxes34_nv.svg", width= 7, height= 6)
par(pch=20, mar=c(5.1, 6.1, 4.1, 1.5))
plot.PCA(qwater.pca, axes=c(3,4), cex=.85, choix="var", cex.axis=0.8, las=1, cex.lab=1,
         cex.main=1,  autoLab="yes", col.var="blue")
mtext("PC3/Factor3 Aggregate Qualitative Feature", side=1, line=4.15)
mtext("PC4/Factor4 Aesthetic Acceptability", side=2, line=4.15)
dev.off()

#############################################################
################## Generating all Tables ###################
#############################################################

## Correlation between variables and first Four components

round(qwater.pca$var$coord[,1:4], 2)

write.csv(round(qwater.pca$var$coord[,1:4], 2), "correlation_variables_four_comp.csv")  # write
table correlation variables for the first four components

##  Eigenvalues to make the scree plot

round(qwater.pca$eig, 2)
write.csv(round(qwater.pca$eig, 2), "eigenvalues_all_components.csv")  # write table eigenvalues for
all components

y <- round(qwater.pca$eig[1], 2)
y  <- as.numeric(unlist(y))

#  Scree plot
svg("scree_plot.svg", width= 6, height= 5)
par(las=1, mar=c(5.5,3.5,2,1))
```

```
plot(seq(1,length(y),1), y, ylim=c(0,12), xlim=c(1,10), pch=20, cex=1, type="b", xlab="Scree Plot",
        ylab="", main="Eigenvalues", cex.main=0.9, axes=FALSE)
box()
axis(2, cex.axis=0.7)
axis(1, at=seq(1,10,1), labels=c("PC1","PC2","PC3","PC4","PC5","PC6","PC7","PC8","PC9","PC10"),
cex.axis=0.6)
text(c(1.4,2.4,3.4,4.4, 5.4), round(qwater.pca$eig[1:5,1], 2) + 0.9,
        labels= paste(round(qwater.pca$eig[1:5,2],2), "%"), cex=0.7, col="blue")
dev.off()

# eigenvalues graph

svg("eigenvalues.svg", width= 5, height= 4)
par(mar=c(2.5,2,2.5,0.5))
barplot(qwater.pca$eig[1:13,1], width = 0.75, main="Eigenvalues", cex.main=1, ylim=c(0,12),
        names.arg=paste("PC", 1:13),
        cex.names=0.6, cex.axis=0.7, las=1)
text(c(0.7 ,1.7 ,2.7 ,3.5, 4.6), round(qwater.pca$eig[1:5,1], 2) + 0.35,
        labels= paste(round(qwater.pca$eig[1:5,2],2), "%"), cex=0.7, col="blue")
dev.off()

## The first component expresses 53.53 % of data variability
## The first component expresses four times more variability than the second, it affects four times
### more variables
## The data are represented by the first four components (53.53% + 13.13% + 10.18% + 6.79% =
## 83.63%)


## Detecting outliers

## Distance from the individuals to the centre of the Cloud

round(qwater.pca$ind$dis, 2)
write.csv(round(qwater.pca$ind$dis, 2), "distance_ind_centre.csv")  # write table distance individuals
to the centre of the cloud

## Important Note 4: These distances were calculated using Euclidean Distance measure
 # sqrt((table_ind[1,1]^2)+ (table_ind[1,4]^2) + (table_ind[1,7]^2) + (table_ind[1,10]^2))
 # See the notes in the notebook

# Example how was calculated distance..  the functionn did it automatically but this is one example
##### how it was calculated
##########

sqrt((qwater.pca$ind$coord[12,1]^2) + (qwater.pca$ind$coord[12,2]^2) +
(qwater.pca$ind$coord[12,3]^2)
        + (qwater.pca$ind$coord[12,4]^2))
```

```
########### Residual Matix Component 5 to 18

#### run again PCA analysis keeping all components to save the residual matrix

qwater.pcaRM <- PCA(qwater_data2, ncp=18, graph=FALSE)

round(qwater.pcaRM$var$coord[,5:18], 2)
write.csv(round(qwater.pcaRM$var$coord, 2), "residual_matrix.csv")  # write table residual matrix

sqrt((qwater.pcaRM$ind$coord[8,1]^2) + (qwater.pcaRM$ind$coord[8,2]^2) +
(qwater.pcaRM$ind$coord[8,3]^2)
        + (qwater.pcaRM$ind$coord[8,4]^2) + (qwater.pcaRM$ind$coord[8,5]^2) +
(qwater.pcaRM$ind$coord[8,6]^2)
        + (qwater.pcaRM$ind$coord[8,7]^2) + (qwater.pcaRM$ind$coord[8,8]^2) +
(qwater.pcaRM$ind$coord[8,9]^2)
        + (qwater.pcaRM$ind$coord[8,10]^2) + (qwater.pcaRM$ind$coord[8,11]^2) +
(qwater.pcaRM$ind$coord[8,12]^2)
        + (qwater.pcaRM$ind$coord[8,13]^2) + (qwater.pcaRM$ind$coord[8,14]^2) +
(qwater.pcaRM$ind$coord[8,15]^2)
        + (qwater.pcaRM$ind$coord[8,16]^2) + (qwater.pcaRM$ind$coord[8,17]^2) +
(qwater.pcaRM$ind$coord[8,18]^2))

####### Example: Distance was calculated using all dimentions (18) according to Euclidean Distance
############## measure


### Results presented in the graphical format using four dimensions as per the display
### demonstrated by Prof.Paulo Gomes during his lectures in the classroom.

table_ind <- cbind(
round(qwater.pca$ind$coord[,1], 2),
round(qwater.pca$ind$contrib[, 1], 1) * 10,
round(qwater.pca$ind$cos2[, 1], 3) * 1000,
round(qwater.pca$ind$coord[,2], 2),
round(qwater.pca$ind$contrib[, 2], 1) * 10,
round(qwater.pca$ind$cos2[, 2], 3) * 1000,
round(qwater.pca$ind$coord[,3], 2),
round(qwater.pca$ind$contrib[, 3], 1) * 10,
round(qwater.pca$ind$cos2[, 3], 3) * 1000,
round(qwater.pca$ind$coord[,4], 2),
round(qwater.pca$ind$contrib[, 4], 1) * 10,
round(qwater.pca$ind$cos2[, 4], 3) * 1000
)

colnames(table_ind)  <- c("CP1", "CTA", "CTR", "CP2", "CTA", "CTR", "CP3", "CTA", "CTR", "CP4",
"CTA", "CTR")

write.csv(table_ind, "table_individuals_CTA_CTR.csv")  # write table individuals
```

```
## variables

table_var <- cbind(
round(qwater.pca$var$coord[,1], 2),
round(qwater.pca$var$contrib[, 1], 1) * 10,
round(qwater.pca$var$cos2[, 1], 3) * 1000,
round(qwater.pca$var$coord[,2], 2),
round(qwater.pca$var$contrib[, 2], 1) * 10,
round(qwater.pca$var$cos2[, 2], 3) * 1000,
round(qwater.pca$var$coord[,3], 2),
round(qwater.pca$var$contrib[, 3], 1) * 10,
round(qwater.pca$var$cos2[, 3], 3) * 1000,
round(qwater.pca$var$coord[,4], 2),
round(qwater.pca$var$contrib[, 4], 1) * 10,
round(qwater.pca$var$cos2[, 4], 3) * 1000
)

colnames(table_var)  <- c("CP1", "CTA", "CTR", "CP2", "CTA", "CTR", "CP3", "CTA", "CTR", "CP4",
"CTA", "CTR")

write.csv(table_var, "table_variables_CTA_CTR.csv")  # write table variables

## Contribution of individuals to the construction of the components(CTA)

round(qwater.pca$ind$contrib[,1:2], 2)

## Contribution of variables to the construction of the components (CTA)

round(qwater.pca$var$contrib[,1:2], 2)

## CTR of individuals

round(qwater.pca$ind$cos2[, 1:2], 3)

## CTR of variables

round(qwater.pca$ind$cos2[, 1:2], 3)

## Description of the first dimension by the quantitative variables

lapply(dimdesc(qwater.pca), lapply, round,2)


### Graphics to represent the data

## table to assign color to represent individuals

pci.ind <- read.table(textConnection(
'CP ind value color
```

CP1 DlWa -4.02 red
CP1 B4a -0.73 darkgreen
CP1 B4b -0.87 darkgreen
CP1 B5a -1.93 purple
CP1 B5b -2.28 purple
CP1 B5c -2.29 purple
CP1 B5d -1.81 purple
CP1 B5e -2.23 purple
CP1 B7a -3.32 darkgoldenrod4
CP1 B7b -3.64 darkgoldenrod4
CP1 B7c -3.85 darkgoldenrod4
CP1 B7d -3.96 darkgoldenrod4
CP1 B7e -3.87 darkgoldenrod4
CP1 B8a 5.18 aquamarine4
CP1 B8b 5.55 aquamarine4
CP1 B8c 6 aquamarine4
CP1 B8d 4.89 aquamarine4
CP1 B8e 5.22 aquamarine4
CP1 B8f 4.83 aquamarine4
CP1 B10a -3.61 darkred
CP1 B10b -3.55 darkred
CP1 B10c -3.54 darkred
CP1 B10d -3.6 darkred
CP1 B10e -3.6 darkred
CP1 B11a 5.33 orangered
CP1 B11b 5.35 orangered
CP1 B11c 5.36 orangered
CP1 B11d 5.25 orangered
CP1 DlWb -4.23 red
CP2 B2a 3.85 blue
CP2 B2b 3.01 blue
CP2 B2c 4.2 blue
CP2 B2d 3.11 blue
CP2 B2e 4.3 blue
CP2 B2f 2.28 blue
CP2 B6a -1.71 indianred
CP2 B6b -1.53 indianred
CP2 B6c -1.88 indianred
CP2 B6d -1.96 indianred
CP2 B6e -2.13 indianred
CP2 B6f -1.61 indianred
CP3 B1 7.39 brown
CP3 B3a 1.21 black
CP3 B3b 0.54 black
CP3 B3c 0.76 black
CP3 B3d 0.82 black
CP3 B3e 0.54 black
CP3 B3f 0.53 black
CP4 DlWa 1.28 red

```
CP4 B6a -1.16 indianred
CP4 B6b -1.4 indianred
CP4 B6c -1.28 indianred
CP4 B6d -1.31 indianred
CP4 B6e -1.27 indianred
CP4 B6f -1.39 indianred
CP4 B8a 2.04 aquamarine4
CP4 B8b 1.75 aquamarine4
CP4 B8c 2.09 aquamarine4
CP4 B8d 1.15 aquamarine4
CP4 B8e 1.59 aquamarine4
CP4 B8f 1.44 aquamarine4
CP4 B9a -0.97 coral
CP4 B9b -0.75 coral
CP4 B9c -0.85 coral
CP4 DIWb 2.23 red
'), header=TRUE, stringsAsFactors=FALSE)

## table to assign color to represent variables

pci.var <- read.table(textConnection(
'CP var value color
CP1 pH 0.43 turquoise4
CP1 EC 0.97 tan4
CP1 Cl 0.85 darkorchid2
CP1 HCO3 0.8 salmon3
CP1 F 0.7 black
CP1 Hardness 0.91 darkred
CP1 TDS 0.95 blue
CP1 Na 0.76 chocolate2
CP1 K 0.84 cadetblue4
CP1 Ca 0.89 bisque4
CP1 Mg 0.91 sienna1
CP1 AnionSum 0.96 darkolivegreen
CP1 CationSum 0.99 brown1
CP2 TEMP 0.72 violetred4
CP2 pH -0.52 turquoise4
CP2 NH4 0.42 skyblue4
CP2 NO3 0.7 cyan3
CP2 Na 0.52 chocolate2
CP3 NH4 0.76 skyblue4
CP3 HCO3 0.32 salmon3
CP3 Fe 0.37 orange2
CP3 Mn 0.78 pink3
CP4 TEMP 0.28 violetred4
CP4 pH -0.46 turquoise4
CP4 NO3 -0.49 cyan3
CP4 F -0.52 black
'), header=TRUE, stringsAsFactors=FALSE)
```

```
## Here is the code to produce the graphics to represent individuals and variables over one axis
###############(example
## "pc1_ind_plot.svg", and "pc1_var_plot.svg") this code produce row graphics with overlapping
######## labels, we did manual edition
## in "Inkscape vectorial software" and the final graphics are presented with the names ######
########## "pc1_ind_var_plot_edit.svg".
## However we are able to run this chunk of code to get the row graphics previous edition.

### PC 1 individuals

x0 <- ceiling(max(table_ind[,1]) + 2)
x1 <- floor(min(table_ind[,1]) - 2)

svg("pc1_ind_plot.svg")
par(mar=c(1,1,1,1))
plot(rnorm(10), xlim=c(x1, x0), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
text(0, -1.15, "0")
axis(1, at=seq(-7, 7, 2), pos=0, lty=2, cex.axis=0.8)
text(0, -1.7, "PC 1")
segments(0,0.75,0,-0.75, lty=2)

points(pci.ind[pci.ind$CP == "CP1","value"],  rep(0, nrow(pci.ind[pci.ind$CP == "CP1",])),
      col= pci.ind[pci.ind$CP == "CP1","color"],
          pch=18, cex=1.5)
text(pci.ind[pci.ind$CP == "CP1","value"], rep(c(-0.9,0.5), nrow(pci.ind[pci.ind$CP == "CP1",])/2),
    pci.ind[pci.ind$CP == "CP1","ind"],
    col=pci.ind[pci.ind$CP == "CP1","color"], cex=0.9)
dev.off()

### PC 1 variables

svg("pc1_var_plot.svg")
par(mar=c(1,1,1,2))
plot(rnorm(10), xlim=c(-1.1, 1.1), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
axis(1, at=seq(-1, 1, 0.25), pos=0, lty=2, cex.axis=0.8)
text(0, -1, "PC 1")
segments(0,0.75,0,-0.75, lty=2)

points(pci.var[pci.var$CP == "CP1","value"],  rep(0, nrow(pci.var[pci.var$CP == "CP1",])),
      col= pci.var[pci.var$CP == "CP1","color"],
          pch=18, cex=1.5)
text(pci.var[pci.var$CP == "CP1","value"], rep(c(-0.9,0.5), nrow(pci.var[pci.var$CP == "CP1",])/2),
    pci.var[pci.var$CP == "CP1","var"],
    col=pci.var[pci.var$CP == "CP1","color"], cex=0.9)
dev.off()
```

```
######################

###  PC 2 individuals

svg("pc2_ind_plot.svg")

par(mar=c(1,1,1,1))
plot(rnorm(10), xlim=c(x1, x0), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
text(0, -1.15, "0")
axis(1, at=seq(-7, 7, 2), pos=0, lty=2, cex.axis=0.8)
text(0, -1.7, "PC 2")
segments(0,0.75,0,-0.75, lty=2)

points(pci.ind[pci.ind$CP == "CP2","value"],  rep(0, nrow(pci.ind[pci.ind$CP == "CP2",])),
      col= pci.ind[pci.ind$CP == "CP2","color"],
          pch=18, cex=1.5)
text(pci.ind[pci.ind$CP == "CP2","value"], rep(c(-0.9,0.5), nrow(pci.ind[pci.ind$CP == "CP2",])/2),
    pci.ind[pci.ind$CP == "CP2","ind"],
    col=pci.ind[pci.ind$CP == "CP2","color"], cex=0.9)

dev.off()

###  PC 2 variables

svg("pc2_var_plot.svg")

par(mar=c(1,1,1,2))
plot(rnorm(10), xlim=c(-1.1, 1.1), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
axis(1, at=seq(-1, 1, 0.25), pos=0, lty=2, cex.axis=0.8)
text(0, -1, "PC 2")
segments(0,0.75,0,-0.75, lty=2)

points(pci.var[pci.var$CP == "CP2","value"],  rep(0, nrow(pci.var[pci.var$CP == "CP2",])),
      col= pci.var[pci.var$CP == "CP2","color"],
          pch=18, cex=1.5)
text(pci.var[pci.var$CP == "CP2","value"], rep(c(-0.9,0.5), nrow(pci.var[pci.var$CP == "CP2",])/2),
    pci.var[pci.var$CP == "CP2","var"],
    col=pci.var[pci.var$CP == "CP2","color"], cex=0.9)
dev.off()

###  PC 3 individuals

svg("pc3_ind_plot.svg")

par(mar=c(1,1,1,1))
plot(rnorm(10), xlim=c(x1, x0), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
```

```
abline(h=0, lty=2)
text(0, -1.15, "0")
axis(1, at=seq(-7, 7, 2), pos=0, lty=2, cex.axis=0.8)
text(0, -1.7, "PC 3")
segments(0,0.75,0,-0.75, lty=2)

points(pci.ind[pci.ind$CP == "CP3","value"],  rep(0, nrow(pci.ind[pci.ind$CP == "CP3",])),
      col= pci.ind[pci.ind$CP == "CP3","color"],
          pch=18, cex=1.5)
text(pci.ind[pci.ind$CP == "CP3","value"], rep(c(-0.9,0.5), nrow(pci.ind[pci.ind$CP == "CP3",])/2),
    pci.ind[pci.ind$CP == "CP3","ind"],
    col=pci.ind[pci.ind$CP == "CP3","color"], cex=0.9)

dev.off()

###  PC 3 variables

svg("pc3_var_plot.svg")

par(mar=c(1,1,1,2))
plot(rnorm(10), xlim=c(-1.1, 1.1), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
axis(1, at=seq(-1, 1, 0.25), pos=0, lty=2, cex.axis=0.8)
text(0, -1, "PC 3")
segments(0,0.75,0,-0.75, lty=2)

points(pci.var[pci.var$CP == "CP3","value"],  rep(0, nrow(pci.var[pci.var$CP == "CP3",])),
      col= pci.var[pci.var$CP == "CP3","color"],
          pch=18, cex=1.5)
text(pci.var[pci.var$CP == "CP3","value"], rep(c(-0.9,0.5), nrow(pci.var[pci.var$CP == "CP3",])/2),
    pci.var[pci.var$CP == "CP3","var"],
    col=pci.var[pci.var$CP == "CP3","color"], cex=0.9)

dev.off()

###  PC 4 individuals

svg("pc4_ind_plot.svg")

par(mar=c(1,1,1,1))
plot(rnorm(10), xlim=c(x1, x0), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
text(0, -1.15, "0")
axis(1, at=seq(-7, 7, 2), pos=0, lty=2, cex.axis=0.8)
text(0, -1.7, "PC 4")
segments(0,0.75,0,-0.75, lty=2)

points(pci.ind[pci.ind$CP == "CP4","value"],  rep(0, nrow(pci.ind[pci.ind$CP == "CP4",])),
      col= pci.ind[pci.ind$CP == "CP4","color"],
```

```
          pch=18, cex=1.5)
text(pci.ind[pci.ind$CP == "CP4","value"], rep(c(-0.9,0.5), nrow(pci.ind[pci.ind$CP == "CP4",])/2),
   pci.ind[pci.ind$CP == "CP4","ind"],
   col=pci.ind[pci.ind$CP == "CP4","color"], cex=0.9)

dev.off()

###  PC 4 variables

svg("pc4_var_plot.svg")

par(mar=c(1,1,1,2))
plot(rnorm(10), xlim=c(-1.1, 1.1), ylim=c(-6, 2), type="n", axes=FALSE, xlab="", ylab="")
abline(h=0, lty=2)
axis(1, at=seq(-1, 1, 0.25), pos=0, lty=2, cex.axis=0.8)
text(0, -1, "PC 4")
segments(0,0.75,0,-0.75, lty=2)

points(pci.var[pci.var$CP == "CP4","value"],  rep(0, nrow(pci.var[pci.var$CP == "CP4",])),
      col= pci.var[pci.var$CP == "CP4","color"],
          pch=18, cex=1.5)
text(pci.var[pci.var$CP == "CP4","value"], rep(c(-0.9,0.5), nrow(pci.var[pci.var$CP == "CP4",])/2),
   pci.var[pci.var$CP == "CP4","var"],
   col=pci.var[pci.var$CP == "CP4","color"], cex=0.9)

dev.off()


##########################################
########  Cluster Analysis ###############
##########################################


### Read R_square data calculated from SAS Enerprise Guide 7.1 (64-bit)

RS_data  <- read.csv("R_square_cluster.csv")

svg("Rsquare_hclusters.svg", width= 5, height= 4)

par(las=1)
plot(RS_data$n_cluster, RS_data$ward, type="l", xlim=c(11, 0), ylim=c(0, 1),
          col="blue", main="Determine the appropiate numbers of clusters", axes=F, cex.main=0.9,
          xlab="Numbers of clusters", ylab="R-Squared", lwd=2.5, cex.lab=0.8)
axis(2, cex.axis=0.8)
axis(1, at=c(seq(1,11,1), pos=0), cex.axis=0.8)
points(RS_data$n_cluster, RS_data$average, type="l", col="red", lwd=2)
points(RS_data$n_cluster, RS_data$centroid, type="l", col="darkgreen", lwd=3, lty=2)
segments(10, 0.35, 9.25, 0.35, col="blue", lwd=2.5)
segments(10, 0.25, 9.25, 0.25, col="red", lwd=2)
```

```
segments(10, 0.15, 9.25, 0.15, col="darkgreen", lty=2, lwd=2.5)
text(9, 0.35, "Ward", cex=0.9, adj=0)
text(9, 0.25, "Average", cex=0.9, adj=0)
text(9, 0.15, "Centroid", cex=0.9, adj=0)
abline(v=4, col="purple")
dev.off()

### Hierarchical clustering using Euclidean Distance

Euclidean_distance = dist(scale(qwater_data2), method="euclidean")

hwar_eucli = hclust(Euclidean_distance, method="ward.D")

svg("eucli_hierarchical_cluster_ward.svg", width= 6, height= 6)
par(las=1)
plot(hwar_eucli, hang=-1, cex=0.8)
rect.hclust(hwar_eucli, 4)
dev.off()

hcentroid_eucli = hclust(Euclidean_distance, method="centroid")
svg("eucli_hierarchical_cluster_centroid.svg", width= 6, height= 6)
par(las=1)
plot(hcentroid_eucli, hang=-1, cex=0.8)
dev.off()

hcomplete_eucli = hclust(Euclidean_distance, method="complete")
svg("eucli_hierarchical_cluster_complete.svg", width= 6, height= 6)
par(las=1)
plot(hcomplete_eucli, hang=-1, cex=0.8)
dev.off()


haverage_eucli = hclust(Euclidean_distance, method="average")
svg("eucli_hierarchical_cluster_average.svg", width= 6, height= 6)
par(las=1)
plot(haverage_eucli, hang=-1, cex=0.8)
dev.off()

##############################################################################################
## The previous dendrograms are the simple one representation, there are others ways to improve
######### the visualization
## for example here we wrote another way, all the result graphs are called "versionB" at the end of
##### the name file.
##############################################################################################

##  Euclidean

svg("eucli_hierarchical_cluster_ward_versionB.svg", width= 6, height= 6)
```

```
ggdendrogram(hwar_eucli, rotate = TRUE, size = 2) + labs(title="Euclidean distance - method
ward.D")
dev.off()

svg("eucli_hierarchical_cluster_centroid_versionB.svg", width= 6, height= 6)
ggdendrogram(hcentroid_eucli, rotate = TRUE, size = 2) + labs(title="Euclidean distance - method
centroid")
dev.off()

svg("eucli_hierarchical_cluster_complete_versionB.svg", width= 6, height= 6)
ggdendrogram(hcomplete_eucli, rotate = TRUE, size = 2) + labs(title="Euclidean distance - method
complete")
dev.off()

svg("eucli_hierarchical_cluster_average_versionB.svg", width= 6, height= 6)
ggdendrogram(haverage_eucli, rotate = TRUE, size = 2) + labs(title="Euclidean distance - method
average")
dev.off()


##############################################################################################
####  K -means
##############################################################################################

## We applied a method that can help determine the appropriate numbers of clusters, it
## consist in plot the within groups sum of squares by number of clusters extracted.

# Choosing the appropriate cluster solution:

######## One common method of choosing the appropriate cluster solution is to compare the sum of
######### squared error (SSE)
######### for a number of cluster solutions. SSE is defined as the sum of the squared distance
######### between each member of a cluster and its cluster centroid. Thus, SSE can be seen as a
##### global measure of error. In general, as the number of clusters increases, the SSE should ###
####### decrease because clusters are, by definition, smaller.
## A plot of the SSE against a series of sequential cluster levels can provide a useful graphical way to
### choose an appropriate cluster level. Such a plot can be interpreted much like a scree plot used in
####### factor analysis.

wss <- (nrow(qwater_data2)-1)*sum(apply(qwater_data2,2,var))
for (i in 2:12) wss[i] <- sum(kmeans(qwater_data2, iter.max = 50, centers=i)$withinss)

svg("SSE_cluster_kmeans.svg", width= 5, height= 4)
plot(1:12, wss, type="b", main="Determine the appropriate numbers of clusters", cex.main=0.9,
        xlab="Number of Clusters", ylab="Within groups sum of squares", cex.lab=0.9, cex.axis=0.8,
pch=20, col="red")
abline(v=4, col="purple")
dev.off()
```

```
kc <- kmeans(qwater_data2, 4, iter.max = 5)

## Important Note: The object generated in the previous process called "kc" contains the cluster
##### results, we can check it
## calling the object in R console. just type "kc" whitout quotes
## we used these results to follow with the cluster anaysis. This note here for further understanding
##### if we are in need of on "kc" results



##  Result cluster using K-means, as is random process we did five times with 5 iterations we always
###### get the same answer.


## to profiling standarize data

## aggregate look by this function standarized data to graphic and table

qwater_scaled <- scale(qwater_data2)

mean_clusA <- apply(qwater_scaled[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b", "B6a",
"B6b", "B6c",
                                   "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, mean)
mean_clusB <- apply(qwater_scaled[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, mean)
mean_clusC <- apply(qwater_scaled[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b",
"B11c", "B11d"),], 2, mean)
mean_clusD <- apply(qwater_scaled[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b",
"B7c",
                                   "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, mean)

table_mean_kmeans <- cbind(mean_clusA, mean_clusB, mean_clusC, mean_clusD)

write.csv(table_mean_kmeans, "table_mean_kmeans.csv")

##  sd

sd_clusA <- apply(qwater_scaled[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b", "B6a",
"B6b", "B6c",
                                   "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, sd)
sd_clusB <- apply(qwater_scaled[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, sd)
sd_clusC <- apply(qwater_scaled[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b", "B11c",
"B11d"),], 2, sd)
sd_clusD <- apply(qwater_scaled[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b", "B7c",
                                   "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, sd)

table_sd_kmeans <- cbind(mean_clusA, sd_clusA, mean_clusB, sd_clusB, mean_clusC, sd_clusC,
mean_clusD, sd_clusD)
```

```
write.csv(table_sd_kmeans, "table_sd_kmeans.csv")

##  Profiling

svg("kmean_profile_standarized_nv.svg", width= 7.5, height= 5)

par(las=2, lwd=1.5, mar=c(3.5,4,2,0.5))

plot(1:18, table_mean_kmeans[,1], ylim=c(-3, 3), type="b",
        col="brown", pch=20, cex=0.5, axes=F, main= "Kmeans clustering results",
        ylab= "Average Standarized Data", xlab= "", cex.main=1)

points(1:18, table_mean_kmeans[,2], col="red", type="b", pch=20, cex=0.5)
points(1:18, table_mean_kmeans[,3], col="darkgoldenrod1", type="b", pch=20, cex=0.5)
points(1:18, table_mean_kmeans[,4], col="darkgreen", type="b", pch=20, cex=0.5)
axis(2, cex.axis=0.8)
axis(1, at=seq(1,18,1), labels=F)
text(rep(-3.7,18), rownames(table_mean_kmeans[1:18,]), srt=25, cex=0.65, xpd=TRUE)
box()

segments(x0=13, y0=-2, x1=13.75, y1=-2, col="brown")
text(12, -2, "Cluster A", font=2, cex=0.8)

segments(x0=13, y0=-2.5, x1=13.75, y1=-2.5, col="red")
text(12, -2.5, "Cluster B", font=2, cex=0.8)

segments(x0=16, y0=-2, x1=16.75, y1=-2, col="darkgoldenrod1")
text(15, -2, "Cluster C", font=2, cex=0.8)

segments(x0=16, y0=-2.5, x1=16.75, y1=-2.5, col="darkgreen")
text(15, -2.5, "Cluster D", font=2, cex=0.8)

dev.off()

## Profiling original data k-means results.

ord_mean_clusA <- apply(qwater_data2[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b",
"B6a", "B6b", "B6c",
                                       "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, mean)
ord_mean_clusB <- apply(qwater_data2[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, mean)
ord_mean_clusC <- apply(qwater_data2[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b",
"B11c", "B11d"),], 2, mean)
ord_mean_clusD <- apply(qwater_data2[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b",
"B7c",
                                       "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, mean)
```

```
table_ord_mean_kmeans <- cbind(ord_mean_clusA, ord_mean_clusB, ord_mean_clusC,
ord_mean_clusD)

write.csv(table_ord_mean_kmeans, "table_ord_mean_kmeans.csv")

##  Profiling

svg("kmean_profile_original_data_nv.svg", width= 7.5, height= 5)

par(las=2, lwd=1.5, mar=c(3.5,4,2,0.5))

plot(1:18, table_ord_mean_kmeans[,1], ylim=c(0, 1100), type="b",
        col="brown", pch=20, cex=0.5, axes=F, main= "Kmeans clustering results",
        ylab= "Average Data", xlab= "", cex.main=1)

points(1:18, table_ord_mean_kmeans[,2], col="red", type="b", pch=20, cex=0.5)
points(1:18, table_ord_mean_kmeans[,3], col="darkgoldenrod1", type="b", pch=20, cex=0.5)
points(1:18, table_ord_mean_kmeans[,4], col="darkgreen", type="b", pch=20, cex=0.5)
axis(2, cex.axis=0.8)
axis(1, at=seq(1,18,1), labels=F)
text(rep(-100,18), rownames(table_ord_mean_kmeans[1:18,]), srt=25, cex=0.65, xpd=TRUE)
box()

segments(x0=13, y0=800, x1=13.75, y1=800, col="brown")
text(12, 800, "Cluster A", font=2, cex=0.8)

segments(x0=13, y0=700, x1=13.75, y1=700, col="red")
text(12, 700, "Cluster B", font=2, cex=0.8)

segments(x0=16, y0=800, x1=16.75, y1=800, col="darkgoldenrod1")
text(15, 800, "Cluster C", font=2, cex=0.8)

segments(x0=16, y0=700, x1=16.75, y1=700, col="darkgreen")
text(15, 700, "Cluster D", font=2, cex=0.8)

dev.off()

##  sd

ord_sd_clusA <- apply(qwater_data2[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b", "B6a",
"B6b", "B6c",
                                    "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, sd)
ord_sd_clusB <- apply(qwater_data2[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, sd)
ord_sd_clusC <- apply(qwater_data2[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b",
"B11c", "B11d"),], 2, sd)
ord_sd_clusD <- apply(qwater_data2[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b",
"B7c",
                                    "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, sd)
```

```
## min

min_clusA <- apply(qwater_data2[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b", "B6a",
"B6b", "B6c",
                                  "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, min)
min_clusB <- apply(qwater_data2[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, min)
min_clusC <- apply(qwater_data2[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b", "B11c",
"B11d"),], 2, min)
min_clusD <- apply(qwater_data2[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b", "B7c",
                                  "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, min)

## max

max_clusA <- apply(qwater_data2[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b", "B6a",
"B6b", "B6c",
                                  "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, max)
max_clusB <- apply(qwater_data2[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, max)
max_clusC <- apply(qwater_data2[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b",
"B11c", "B11d"),], 2, max)
max_clusD <- apply(qwater_data2[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b", "B7c",
                                  "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, max)

## skewness

skewn_clusA <- apply(qwater_data2[c("B3a", "B3b", "B3c", "B3d", "B3e", "B3f", "B4a", "B4b", "B6a",
"B6b", "B6c",
                                  "B6d", "B6e", "B6f", "B9a", "B9b", "B9c"),], 2, skewness)
skewn_clusB <- apply(qwater_data2[c("B1", "B2a", "B2b", "B2c", "B2d", "B2e", "B2f"),], 2, skewness)
skewn_clusC <- apply(qwater_data2[c("B8a", "B8b", "B8c", "B8d", "B8e", "B8f", "B11a", "B11b",
"B11c", "B11d"),], 2, skewness)
skewn_clusD <- apply(qwater_data2[c("DIWa", "B5a", "B5b", "B5c", "B5d", "B5e", "B7a", "B7b",
"B7c",
                                  "B7d", "B7e", "B10a", "B10b", "B10c", "B10d", "B10e", "DIWb"),],
2, skewness)


table_original_stats_kmeans <- cbind(ord_mean_clusA, ord_sd_clusA, min_clusA, max_clusA,
skewn_clusA,
                                     ord_mean_clusB, ord_sd_clusB, min_clusB, max_clusB,
skewn_clusB,
                                     ord_mean_clusC, ord_sd_clusC, min_clusC, max_clusC,
skewn_clusC,
                                     ord_mean_clusD, ord_sd_clusD, min_clusD, max_clusD,
skewn_clusD)
```

[END OF THE DOCUMENT]