NOVA IMS Information Management School

MEGI

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

Automated data analysis in quantitative research

Prototyping an automation software for obtaining fast and reliable insights in quantitative research settings

Gero Wahrenburg

Project report 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

AUTOMATED DATA ANALYSIS IN QUANTITATIVE RESEARCH. PROTOTYPING AN AUTOMATION SOFTWARE FOR OBTAINING FAST AND RELIABLE INSIGHTS IN QUANTITATIVE RESEARCH SETTINGS

by

Gero Wahrenburg

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

Advisor: Nuno Miguel da Conceição António

November 2022

ABSTRACT

This project results report introduces a work project in which a prototype was built as a “proof of concept” for a data analysis tool that automates data preparation, exploration and modeling tasks within empirical research settings.

Empirical research currently relies on manual processing and analysis of data. In order to enhance research efficiency and to support users, this work can be automated. In particular, tasks such as preprocessing, exploring and analyzing data with statistical methods can be simplified using an automated workflow.

The comprehensive tool envisioned should react flexibly to a variety of data input and incorporate a wide range of conventional analyses. It should produce a structured and formulated research report of Word (.docx) format to facilitate further user manipulation. It should communicate with the user through an intuitive web tool.

Exploring the possibility of such a tool, which is the scope of this work project, included the programming of a less comprehensive “proof of concept” tool. This prototype performs, upon the exploration and preprocessing of the data, linear regression with OLS estimation on a cross-sectional data set.

The prototype was successfully developed and tested on multiple data sets. It is ready to support users without programming skills or access to proprietary software in preprocessing and exploring their data and finding relationships between different variables. Every step of the analysis is explained to the user in the comprehensive automatically generated output report. The prototype provides a basis on which the functionality can be extended towards other use cases such as time series tasks as well as towards offering access to the application of more complex algorithms.

Keywords: Research automation; Automation; AutoML; Research process; Automated data mining

INDEX

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AutoML Automated Machine Learning, the automation of the application of machine learning to a variety of real-world problems

OHE One Hot Encoding, a data preprocessing method

CRISP-DM A Data Mining framework

Moderation The level to which the influence of one variable on another depends on the level of a third variable

Mediation The indirect influence or underlying relationship of one variable on another variable

1. INTRODUCTION

"Humans may be essential when it comes to formulating theories to explain results, but the rest of scientific writing—from a paper's introduction through its description of experiments, methods, and results—would likely benefit from automation."

Daniel Engber (Journalist), 2017

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Most methodological tasks in quantitative empirical research projects are related to the exploration, modification, modelling and interpretation of data. In practice, this is usually performed manually by the researcher. The analysis is often conducted with statistical analysis software such as SPSS Statistics by IBM, which is a popular statistics software in social sciences research, but also the most used software in health science and particularly popular for observational and experimental studies (Masuadi et al., 2021). With SPSS, researchers use manuals and instructions to perform their own statistical analyses from pull-down menus. In addition, researchers can use the proprietary programming language 4GL, which helps with automating repeated tasks (Wheeler, 2014).

Manuals for statistics software are readily available via numerous websites and books (e.g., Uedufy, 2022 and Hemmerich, 2015). Furthermore, free software extensions or macros that can be added to software such as SPSS have made it easier to use software for data analysis and automate repetitive tasks in the process (Zou et al., 2019). However, the process associated with using manuals and working with statistics software still widely relies on manual work, which is a time-consuming factor for researchers. Despite the oftentimes repetitive nature of many of these tasks, few researchers tap into the automation potential that automated data exploration, preparation and analysis pipelines can provide (Engber, 2017).

The project's goal is to explore the possibility of a tool that automates statistical analyses in quantitative empirical research. The tool delivers, as output, an empirical research report that provides the figures, tables, explanations and interpretations that are commonly provided in academic publications. The report can be used for orientation, inspiration and quality checking in the research process.

To dissect the empirical research process and identify automation potential, the CRISP-DM framework is used (Chapman et al., 2000). This open standard process model is widely used in analytics and data mining and can, using some analogies, be extended to describe the empirical research process. It provides an "orderly partition of the often complex data mining processes to ensure a practical implementation of data analytics and machine learning models" (Tripathi et al., 2021). CRISP-DM splits the data mining process into the phases Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. In the research process, we can adapt these phases to obtain a sequence of *Domain* Understanding, Data Understanding, Data Preparation, Modeling and Evaluation (there is no Deployment).

The most significant automation potential lies in the Data Understanding, Data Preparation and Modeling phases, because numerous existing approaches exist that automate exploration, preprocessing and modelling of data (see the literature review chapter for an overview of automation approaches). In contrast, the Domain Understanding Phase is not automatable because it depends on the experience and professional background of the user of the framework. The Evaluation Phase is only partly automatable, since interpretations that relate to domain understanding (as opposed to statistical inference) cannot be automated.

1.2. STUDY OBJECTIVES

The purpose of this project is to create a pipeline that automates the phases Data Understanding, Data Preparation, Modeling and (partly) Evaluation. During the project, the potential for applications within a wide range of conventional analyses is to be established (see Figure 1).

The potential applications include a large range from classical statistical analysis to advanced machine learning algorithms and span categorical and quantitative dependent and independent variables. The automated approach can make applying and understanding these methods easier compared to purely manual implementation with statistical software.

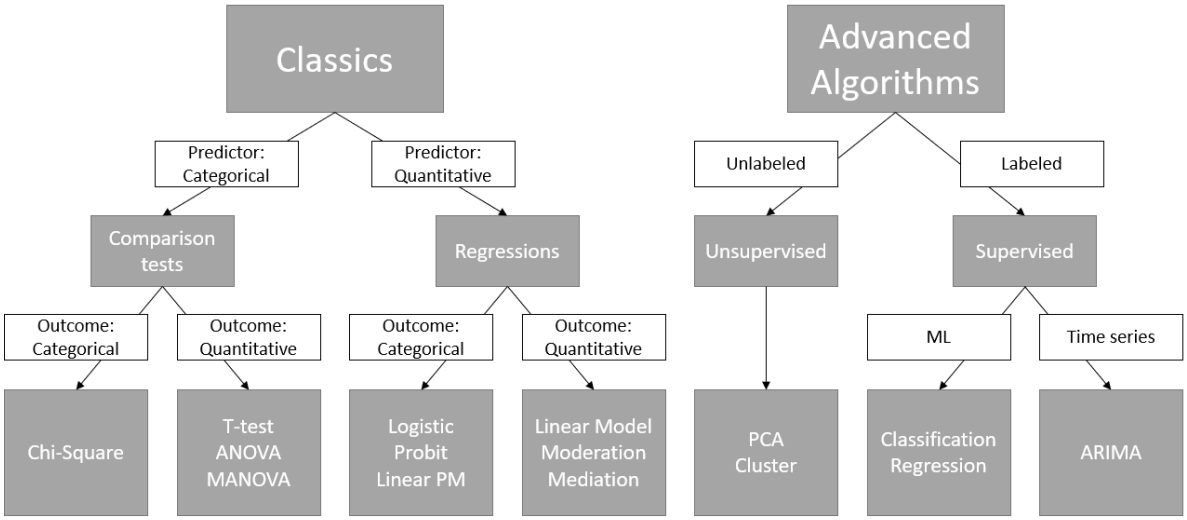


Figure 1.1 – Overview of Analyses with Automation Potential in Empirical Research

In principle, the pipeline can in the future be combined with an intuitive web tool that gathers the required information from the user via a sequence of web form elements and provides, by download of via email, a results folder. This folder contains all created tables and plots as well as the Word (.docx) research report and accompanying explanatory and support documents. With this approach, the software can be deployed as a web service. However, for the purpose of the prototype, the necessary information is directly provided to the prototype (manual entry in the form of a dictionary-structured input section of the first script of the automation sequence) and the results folder is locally created instead of being sent out or provided for download.

The scope of this project is limited to developing a prototype as “proof of concept”. Restrictions are introduced regarding the complexity of the automation pipeline:

Input: Restrictions are placed on the size and types of input data. In particular, very big data sets and non-tabular or time series data are out of scope. The comparatively easy case of a cross-sectional data set is assumed. This reduces the scope of potential structural issues with the data which need to be addressed (such as serial autocorrelation in time series). The restricted size excludes any performance-related issues, which would require a runtime optimization of the automation sequence.

Analyses: The range of models is restricted to estimating a linear regression with ordinary least squares (OLS) for the prototype. While the validity of OLS using tests for the relevant regression assumptions are required and biased results are flagged, no alternative estimators are used in order to simplify the range of potential result interpretations. Exploratory techniques (PCA and cluster analysis) have been successfully implemented as a first test before developing the prototype, because they are unsupervised and do not depend on user input (as regression does because the dependent and independent variables need to be determined by the user). However, the prototype as presented here is only concerned with linear regression using OLS estimation. Linear regression resembles a valid choice for the prototype portrayed here because the “multiple linear regression model and its estimation using ordinary least squares (OLS) is doubtless the most widely used tool in econometrics” (Fabra & Schmidheiny, 2010).

As specific objectives for the project of building a working prototype, the following functional requirements are determined:

1) The tool reliably performs analyses within the scope of the restrictions (Linear regression with OLS on cross-sectional datasets)
2) The tool follows a set of rules (default settings of the software) to automate the preprocessing of the data (including the treatment of missing values, transformation of categorical variables and treatment of duplicates and constants). In addition, the user can determine the precise preprocessing methods and parameters by providing additional input (advanced settings)
3) The tool produces quality output in the form of a Word (.docx) research report, including:
    a. Figures and Tables
    b. Statistical interpretations
    c. Explanation of the research methods applied
4) All figures and tables are, in addition, separately available in a results folder
5) The tool includes techniques for feature selection/reduction to deal with excessively large data sets
6) The user should be able to specify:
    a. Predictor and outcome variables in the data set
    b. Hypotheses to be tested and evaluated
    c. Advanced preprocessing settings (see point 2)
7) The user input is, for the prototype, directly provided to the prototype. In a complete version, this could be performed via a form-based web interface
8) The output report is formulated simple enough to be followed on the basis of high school mathematics and an introductory lecture in statistics

1.3. STUDY RELEVANCE AND IMPORTANCE

The low automation share prevalent in empirical research has advantages in terms of fine-grained control over the research process. Doing every step from data manipulation and preprocessing separately and manually allows the researcher to make statistical decisions along the whole research process, drawing on his expertise in quantitative methods as well as domain knowledge. However, this approach comes with significant drawbacks:

1) Inexperienced researchers, including students engaged in writing a thesis, spend a significant amount of time acquiring the prerequisite skills to perform tasks in statistics software like SPSS.
2) This time investment poses a barrier to entry in research. A number of students are deterred, often resorting to literature review projects.
3) Automation allows skilled researchers to save time and rapidly test ideas.

An additional drawback of manual research concerns the quality of the research output. Choueiry and Salameh (2019) find that "medical research suffers from statistical errors that could be otherwise prevented such as errors in choosing a hypothesis test and assumption checking of models". A large part of published research fails to report on the tests for assumptions underlying hypothesis tests (Hanif & Aymal, 2011). This contributes to the finding that most research findings are, actually, wrong (Ioannidis, 2015). In particular, assumption checks have the drawback that visual inspections are prone to the bias towards attaining significant results (Dwan, Gamble, Williamson & Kirkham, 2013), whereas tests have problems with small sample sizes (Barker & Shaw, 2015). Automating the analysis flow and applying rigorous statistical tests rather than allowing for benevolent and biased visual inspection can be hypothesized to improve the reliability and thereby the quality of research. This is especially relevant in the case of inexperienced, less skilled research such as the research conducted within the scope of a bachelor thesis, especially where the exposure to courses dedicated to statistics during the educational curriculum is limited.

Besides empirical research, a second domain of potential application is as an automated data analysis tool for companies, providing knowledge discovery in departments across the company where analytics expertise might be lacking. Here, the tool effectively takes the place of a fictitious research team that provides analyses on request and in the format of a comprehensible project report with a linear storyline and structure.

The tool fills a gap in the current market for data analytics software, especially within AutoML applications (Automated Machine Learning, the automation of the application of machine learning to a variety of real-world problems). Whereas broader data science is currently experiencing a wave of AutoML tools and competitions (e.g., Mamaev, 2019 and Malato, 2020), traditional data analysis still mostly relies on non-automatic handling. Bringing increased modelling automation to the functionality offered by a software such as SPSS can partly bridge the gap between traditional statistics and data science (Abdelfattah, 2020).

Moreover, current projects on no-code analytics (e.g., Chiechanowski et al., 2020) prove that opening advanced analytics to researchers without programming skills addresses an existing and growing demand in the context of an "advancing datafication of social sciences".

The following section will establish a brief review of competitor products and clarify the differentiation that this tool provides against these "competitors".

1. Analytics software packages

Examples: Alteryx Designer, IBM SPSS, SAS, STATA

Differentiation: The user has to get and install the software (a webtool is more accessible). Furthermore, the user has to build the required workflows himself, reading through manuals and documentation. Most decisions are made by the user. This makes the work time-intensive and error-prone. Drag& drop, canvas-based tools like Alteryx have gained traction in recent years. For instance, Alteryx became the default analytics software used by the global strategy consulting firm BCG, which seeks to enhance regular consultants' analytics profile (Miller, 2014). However, these tools are based on expensive subscriptions. On the other side, drop-down menu software such as SPSS has been criticized for feeling "odd and unmaintained" and has seen a stark decline in scholarly citations (Lindeløv, 2019). The availability of commonly used SPSS macros such as the PROCESS Macro by Hayes (Hayes, 2021) in Python libraries, in the case of PROCESS the PyProcessMacro (Andre, 2019) makes automation of tasks typically performed with SPSS, such as mediation and moderation analyses, with Python scripts feasible.

2. Multi-purpose programs based on cloud computing and AI

Examples: IBM Watson, Wolfram Alpha

These programs are, conveniently, cloud-based. The sophistication and variety of use cases is (by multiple margins) higher than in this project. For analytics, IBM helps end users obtain intelligent data analysis and visualization with its IBM Cognos Analytics application which was merged with Watson analytics. Here, interactive dashboards (similar to Tableau) and the application of machine learning algorithms are main components.

Differentiation: Despite making the analysis process easier, tools such as Cognos Analytics still require significant input from the user. They excel at agile, interactive data exploration. Yet they rely on user initiative and curiosity. In contrast, the tool envisioned in this project will require less user input, due to being tailored to quantitative research tasks, whereas Cognos Analytics has recently fostered concentration on broader Business Intelligence applications (Avidon, 2019).

3. Programming

Examples: Python, R

Writing code for data analysis is the most flexible way to perform data analysis. In particular, R is a widely used programming language in fields such as economics and features a variety of libraries that support most common analyses.

Differentiation: Coding imposes high barriers to entry on data analysis and is highly time-consuming. Because graphical user interfaces are commonly lacking, the availability of programming languages is restricted to "a minority of researchers who know how to code" (Choueiry & Salameh, 2019).

4. Automated Data Mining

Examples: Oracle Predictive Analytics, Microsoft Azure Machine Learning

These products offer "one-klick datamining" and use drag-and-drop interfaces to open datamining to non-experts.

Differentiation: Despite the simplicity of use, these programs are interactive and lack full automation (Leal, 2015). They are designed to make highly effective machine learning models available. The goal is not to make relatively simple analyses such as regressions more easily accessible.

5. Tutors and Consultants

Examples: Numerous analytics consulting companies and freelancers

In the field of academic research, paid statistic tutors offer to support inexperienced students in empirical projects. In corporate settings, consultants as well as internal analytics units are regularly tasked with data analysis projects that turn out to be less complex than imagined by the sponsor (such as simple clustering tasks), which leads to an overly resource-intensive treatment of simple tasks, leaving automation potential on the table.

Differentiation: The tool would be more cost efficient than any consultancy service and would thus help with getting simple data analysis tasks done with minimal resource and time investment.

2. LITERATURE REVIEW

The literature relevant to this project can be categorized into four streams. Firstly, the review will show the debate around the advantages and disadvantages of automated research. Secondly, the discussion on automatically generated and submitted research papers will be featured. Thirdly, previous ideas on making machine learning more accessible to unskilled users will be addressed. Fourthly, automated model selection approaches will be discussed. Fifthly, the combination of statistical programming languages and graphical user interfaces will be reviewed. Sixthly, the review will commence with a review of the Design Science Research methodology that was consulted in the design of the prototype.

2.1. AUTOMATED RESEARCH: ADVANTAGES AND DISADVANTAGES

The project is placed in the domain of research automation, which has recently been subject to a number of discussions and advances that will be briefly summarized here. One instance which lends itself to shedding light on the diverging opinions on this matter is the introduction of the LINEAR procedure by SPSS (Yang, 2013), a procedure that has added significant automation to linear regressions. Despite some reservations by researchers regarding the perceived reduction of the role of the researcher in the research process (Fields, 2013), a consensus exists that due to the need to generate insights in an environment of ever-increasing data availability, ways for automatic data analysis are necessary (Han & Kamber, 2006; Witten et al., 2011). Automatic data mining techniques are especially efficient in environments of comprehensive data storage, where the challenge is to synthesize knowledge from huge amounts of data (Asghar 2009).

Automated research has also been praised for the potential to make research output more readable by cutting back on jargon and aiding non-native English speakers (Engber, 2017).

However, human input remains critical, especially for inspiration from literature review, research question formulation, creation of a data sampling plan, evaluation of results and dealing with abnormal cases (Yang, 2013).

Legal and ethical issues have also been pointed out. For example, it has been brought forward that automatically scanning sources using Natural Language Processing algorithms could constitute plagiarism (Pells, 2017).

2.2. AUTOMATICALLY GENERATED RESEARCH PAPERS

In the past, computer programs have been utilized to generate “fake papers”, that have even been admitted to research conferences (van Noorden, 2014). However, these programs did not conduct any analyses on a user-provided data set. Instead, they generated a random paper on any plausible topic within a specific research field, such as computer science. Of relevance to this study are approaches that incorporate the analysis of user-provided data. In this area, significant progress has been made in recent years. This progress mainly incorporates Natural Language Processing (NLP) software to automate the generation of meta-analyses by working through the corpus of available research resources, including papers and books (Nichols, 2021; Futurism, 2020; Hvrinsum, 2021; Mikko, 2019). While impressive, this progress is also of limited relevance to this study because it concerns literature reviews and related research types. This study is about empirical research on own data. This specific field has been addressed by only few players. In particular, programs designed at writing manuscripts

of research papers have been developed in the past, including the software *Manuscript Writer* by the company sciNote (Pells, 2017; Engber, 2017). Such manuscript software resembles the highest similarity with the topic of this study. However, the manuscript prototyping software is primarily designed at organizing the results of experiments and allowing the collaboration between researchers. The situation that is faced by students during their theses, or professionals during analysis task, is simpler than that: Often, they have a data set to work from (obtained through surveys, data bases etc.) and need an automated analysis of this data. This study focuses on software that is tailor-made for this situation and neglects the broader research coordination addressed by *Manuscript Writer*.

2.3. FRAMEWORKS FOR ACCESSIBLE DATA MINING

In the evolving research on automated data mining, several contributions have provided solid foundations for the development of our project:

Marcos M. Campos of Oracle has conceptualized the requirements for successful automated data mining (Campos et al., 2009). Rather than "beating" the results of data mining experts, the goal of automation is allowing less skilled users to achieve good results with minimum effort. The key to unlocking the power of data mining for these users, according to Campos, is the use of a data-centric approach where users interact with the data rather than with abstract models. Campos also introduces a framework for building an automated model which consists of several phases and provides a helpful foundation for the structure of the tool planned in this project. These phases are: Computation of statistics, Sampling, Attribute data type identification, Attribute selection, Algorithm selection, Data transformation, Model selection and quality assessment and Output generation.

One of the key ideas of Campos is that the user applies intuitive query words such as EXPLAIN, PREDICT, GROUP, DETECT, MAP and PROFILE. As these words are referring to the data and not the models, users can apply data mining as similar to how they would work on a spreadsheet.

Bloom also put forward a helpful framework that spans the process from information receival until model output (Bloom et al., 2003).

Furthermore, several recent and relevant publications detail state-of-the art concepts in AutoML, including automated feature engineering, hyperparameter optimization and more. These publications provide guidance in identifying the methods implemented in this project (Waring et al., 2020, Hutter et al., 2019).

2.4. AUTOMATED MODEL SELECTION

Automated model selection tools such as the Autometrics algorithm have made finding appropriate variables easier by making many manual steps obsolete (Doornik, 2009). They are designed to support the selection of variables to optimize model selection across many conceivable combinations of predictors and interactions. This is very helpful because research has shown that they often outperform human researchers, even with domain expertise, when it comes to selecting the best combination of variables (Kamarudin & Ismail, 2017). However, the application designed in this project does not seek optimal results, but rather good results for unskilled users which are, in addition, easily interpretable. Since transformations performed during variable selection harm interpretability and some relationships which are potentially of interest to the user even become unobservable when the

corresponding variables are automatically deselected, model selection techniques and frameworks are more appropriate to advanced users and researchers.

2.5. AUTOMATED ANALYSIS PACKAGES WITH GUI (GRAPHICAL USER INTERFACE)

Choueiry and Salameh (2019) have created an automated data analysis package that builds on the R programming language. The user can interact with the R-based modules via a graphical user interface. The automation is designed to allow for "non-subjective" research by incorporating assumption checks and solutions for violated regression assumptions. The prototype includes methods to treat outliers and missing data, facilitating easy and fast preprocessing. This open-source approach is close to the concept of this prototype. Unlike the package created by Choueiry and Salameh, however, the core output of the prototype is not limited to tables and figures which are delivered during the interaction with the user input. Instead, a comprehensive report is written based on the initial user input.

2.6. DESIGN SCIENCE RESEARCH

Design Science Research (Hevner et al., 2004) is a research approach that stems from the demand for a structured framework facilitating the systematic creation of IT artifacts in research. Its value lies in the combination of engineering and research approaches. The framework can be used not only to structure the research work but also as a structure of the research report. Seven phases or sections are included: Problem Identification and Motivation, Objective of the solution, Design and Development, Demonstration, Evaluation, Communication and Contribution. These domains were used to structure this study.

3. METHODOLOGY

The Project was operationalized in Python, using the machine learning library scikit-learn (Pedregosa et al., 2011), the library for statistical computation statsmodels (Seabold & Perktold, 2010) and the specialized docx library for managing the Python code to MS Word report interface (Canny, 2013). Jupyter Notebook (Kluyver et al., 2016) was employed as programming environment for the initial prototyping using notebook interactivity. Afterwards, the integrated development environment PyCharm (JetBrains, 2017) was used to create an architecture of multiple scripts in order to orchestrate the end-to-end pipeline.

The pipeline was iteratively developed and tested with a set of cross-sectional data sets.

- The Smart Home Campaign dataset (Henriques, 2021) which provides categorical and numerical information on the characteristics and purchasing history of 2,500 customers of a fictitious company active in the smart home sector. This dataset was used for the sample output displayed in this report. An excerpt of the data is provided in the appendix.
- A survey dataset which provides population characteristics and numerical information on the response of consumers towards political statements that affect brand reputation (Burgold, 2022). An excerpt of the data is provided in the appendix.
- The Song Popularity dataset (Kaggle, 2022) which provides numerical information on the popularity as well as characteristics of different songs
- The Boston Housing dataset (Harrison & Rubinfeld, 1978) which provides numerical information on house prices and characteristics for 506 houses in the Boston area originally collected by the U.S. Census Service
- A dataset on peacekeeping missions (Doyle & Sambanis, 2000) which provides numerical as well as categorical and text information on peacekeeping missions and their outcomes from civil wars between 1944 and 1997

The report, in particular the outputs and findings, will be guided by the APA guidelines for reporting scientific analyses.

The project will be completed in several phases across the two task domains of programming and research (Figure 2).

The study is planned using Design Science principles with the intention of creating a resemble a purposeful and usable IT artifact in the form of a software tool. The design and evaluation of the prototype will be based on rigorous research methods and the results will be communicated in form of a final project report.

To be aligned with Design Science methodology this study needs to address the domains:

i) Problem Identification and Motivation
ii) Objective of the solution
iii) Design and Development
iv) Demonstration

v) Evaluation
vi) Communication
vii) Contribution

The study addressed these domains as follows:

Problem Identification and Motivation

In the Problem Identification and Motivation domains, this study summarized the current problems associated with manual data analysis as described in the Introduction chapter 1.1.

Objective of the solution

In the Objective of the solution domains, this study has detailed the overall requirements in chapter 1.2. The key idea is to create a proof-of-concept for a program that automates data analysis tasks in the phases Data Understanding, Data Preparation, Modeling and (partly) Evaluation. The detailed requirements concern the abilities of the program in the areas of data preprocessing, feature selection, data exploration and regression analyses.

Design and Development

The Design and Development domain is explained in detail as it comprises the core work or values creation of the study. The study structured the development work into multiple steps.

As a first step, a prototype for data exploration, preprocessing and PCA/Clustering was built. This task was chosen for prototyping because, apart from the user-provided dataset, no user input is needed. Regression analyses, in contrast, work from several user-specified hypotheses that are than tested using different models. This implies a higher complexity. The prototype was aimed at testing if the intended mechanics of the prototype worked, in particular writing from a Python script to a MS Word document to obtain the research report. To achieve this, the prototype has to fill a cloze text (a text with blanks) using variables obtained from processing the user-provided data set. Furthermore, the order and elements of the cloze text have to be flexible enough to show only in relevant situations. This process must be flexible to accommodate different data sets. In the prototype phase, a selection of 10 test data sets was collected, primarily from the kaggle datasets platform. These data sets were used for testing the prototype. One data set was selected for the original creation process of the prototype. To support an effective programming process, a data set with different kinds of variables (time, numerical, categorical, binary) was selected to incorporate processing solutions for the various data types expected in user-provided data sets.

Along with the creation of the prototype, a technical architecture was developed which splitted the total work which the program needs to perform in order to fill the MS Word output with the appropriate analyses into different modules. The architecture also determines the sequence in which they have to be run to perform different analyses. This architecture is shown in Figure 2. As shown, the simplified architecture shows three sequences of components that can be seen from the black, blue and orange numbered balls above each component. The three sequences each end with a research report as output: One report is created on data preprocessing (in the module *DATA_PREP_write*), one on data exploration including preprocessing (in the module *DISCOVER_write*)

and one including preprocessing, exploration and the regression analyses (in the module *EXPLAIN_write*).

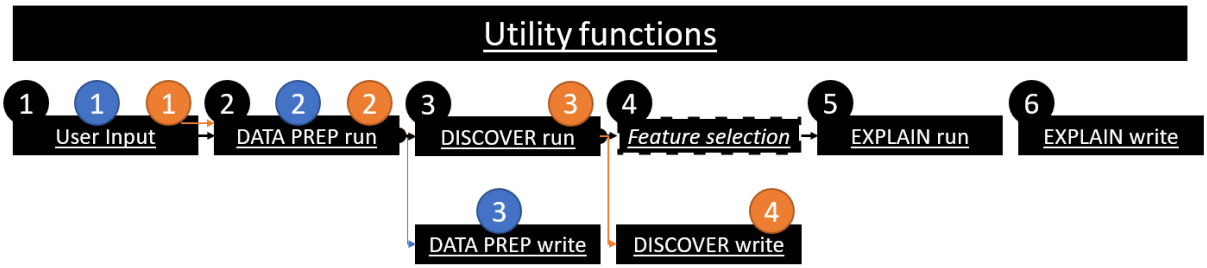


Figure 3.1 – Modules for the Data Analysis Program, Simplified Plot

The different modules are:

User input: Here, all user input is collected, including the data source and the analyses to be performed as well as information on specific requirements for feature selection and preprocessing (in case the user wants to deviate from the default settings). Whereas the input can be gathered separately via a web form in the envisioned comprehensive program, the prototype works with direct inputs into the DATA_PREP_run script.

DATA_PREP_run: Here, the preprocessing is performed and a preprocessed data frame object is returned. The preprocessing is detailed in Figure 3.

DATA_PREP_write: Here, a MS Word output is generated in the form of a report on the preprocessing conducted (this is not included in the prototype)

DISCOVER_run: Here, exploratory data analyses as well as PCA and Cluster analysis are conducted

DISCOVER_write: Here, a MS Word output is generated in the form of a report on the exploratory analysis and any detected patterns in the data is conducted (this is not included in the prototype)

Feature_selection: Here, features to be included in the regression analyses are added or selected and and a modified data frame object is returned

EXPLAIN_run: Here, hypotheses tests on user-provided hypotheses that were entered in the input module are conducted, using regression analyses including linear and logistic regression. In the prototype, only linear regression with OLS is performed. The module tests for regression assumptions using a series of tests as described in the results section. In particular, the linearity assumption (which is often validated when dealing with real-world data) is tested and a negative tests results in an automated attempt to log-transform the dependent and independent variables to make the relationship linear enough to warrant valid interpretation.

EXPLAIN_write: Here, a MS Word output is generated in the form of a report on the regression analyses in the data is conducted.

The _write modules are further divided into: 1. a layout module, which structures the addition of element in the Word (.docx) report from top to bottom by determing the tables and plots as well as

the text and the variables inserted into the text, 2. a texts module which contains all texts and 3. a helper function module which translates the steps defined in the layout module into the final code for the assembly of the document.

If the user wants to perform only data preprocessing, the program will run the sequence: User input – DATA_PREP_run – DATA_PREP_write

If the user wants to perform also exploratory analyses, it will run: User input – DATA_PREP_run – DISCOVER_run - DISCOVER_write

If the user wants to perform regression analyses, the sequence becomes: User input – DATA_PREP_run – DISCOVER_run - EXPLAIN_run – EXPLAIN_write. This sequence is producing the output shown in the results section.

This means that the regression analysis, representing the core output of the prototype, takes in the variables and plots generated in all different _run modules and combines all information in a comprehensive _write report. Therefore, this project report will only show and discuss the results of the regression analysis pipeline.

This architecture of the program was then detailed with architectures of the individual components, detailing the input and output of each. As a rule, all _run modules have as input a dataframe object and as output the variables and plots to be included in the report. _write modules have as input these variables and plots and as outputs the MS Word report. In addition*, DATA_PREP_run* and *Feature_selection* return modified versions of the provided dataframe (preprocessed or with selected features only).

The following figure provides an example by showing the simplified architecture for the DATA_PREP_run module (for preprocessing) module. Note that green font marks default settings that can be edited by the user. Thus, the user can deactivate certain steps within the preprocessing pipeline or select other thresholds or treatment algorithms. In the final program, this should only be possible if applicable, for example, duplicate removal should only be deselectable if duplicates are contained in the data set.

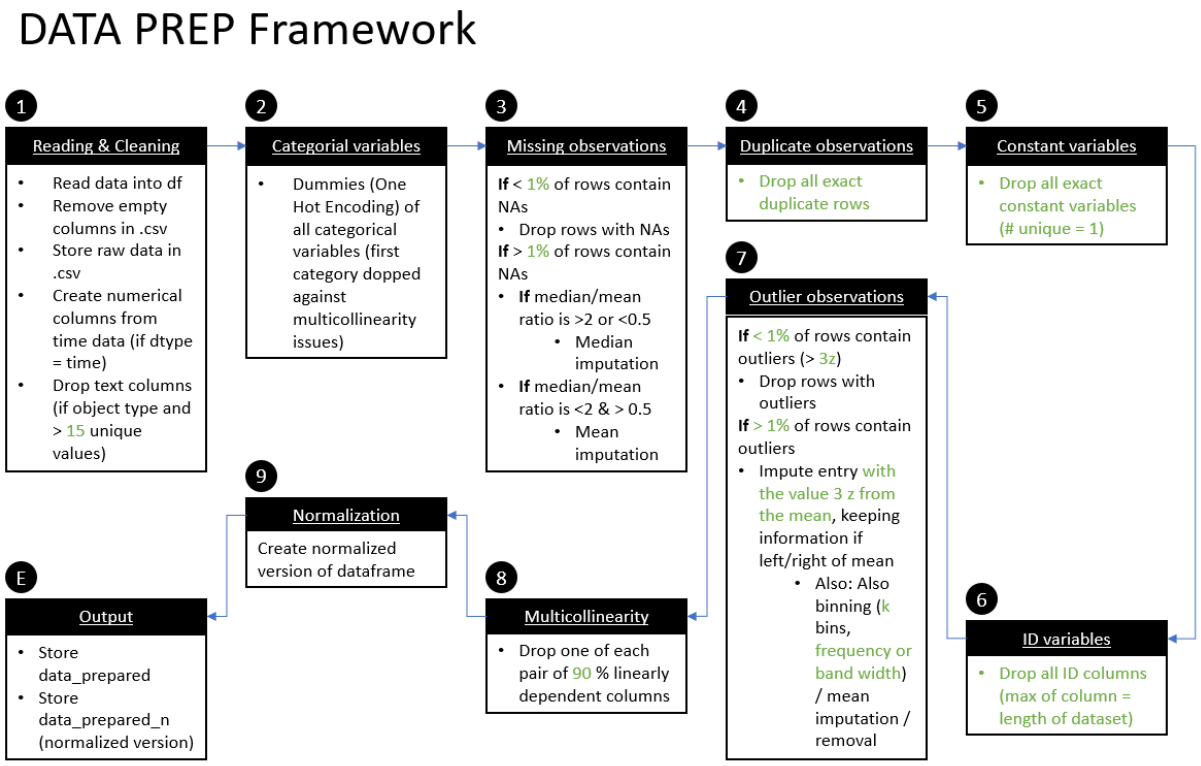


Figure 3.2 – Architecture for Data Preprocessing in Default Mode, Simplified

Please note that the data preprocessing framework as described above does only show the most important steps. Unmentioned steps include the automatic deletion of rows above the column header and the cleaning of column names to enforce alpha-numeric characters.

Besides the default mode which does not require user input, the user can decide to provide specific information on how the data should be preprocessed. The available controls are:

- If to remove columns:
    - with unnecessary metadata. Survey services such as Qualtrix return a data file which includes metadata such as the IP address of the user as an automatically generated column. Such columns are automatically detected and deleted. In a webservice, the deletion can be suggested to the user who can then decide
    - with text rather than categories. Per default, columns of text type with more than 15 unique values are assumed to be text and deleted. In a webservice, the deletion can be suggested to the user who can then decide
    - that are constant. Constant columns carry no information, because they only feature one value. The user can decide if constant values should be deleted.
    - any columns that are not interesting to the analysis or reveal information on the dependent variable in a way that harms the analysis (leakage)
- If to remove duplicates

- If to map numerical columns to categories: Survey services return a data file with either all values numerical or text (e.g., “1,2,3” or “very good, good, ok” as answers to a question). The user will usually upload the numerical version of the data set in order to be able to conduct quantitative analyses. However, this leads to some columns being wrongly specified. For example, a “gender” column would feature 0, 1 and 2 rather than “male”, “female” and “other”. To avoid treating categorical information as numerical information, the user can translate all numerical values to categories which will then be encoded as described in the next paragraph
- How to encode categorical variables: The user can decide what share of observations are required to belong to a category to justify creating a dummy variable. Per default, all dummy variables are created and the category with the most observations is dropped to avoid perfect multicollinearity (the dummy variable trap)
- Which columns to rename and how. This is often required because survey software returns column headers such as “Q1_Age”. Simplifying variable names aids the readability of the analysis output
- How to treat missing values. Observations with missing values can be either dropped or imputed. The user can assign a method to each column. All unassigned columns are treated with the default method (the default mode architecture)
- How to identify outliers. The user can specify the IQR or the z-score approach to identify potential outliers. The IQR approach defines outliers as values with a distance of over 1.5 times the interquartile range above the third or below the first quartile. Note that both methods can only identify “extreme” values at one end of the range of values, but no otherwise odd-appearing values. The user can change the number of interquartile ranges or standard deviations used for both methods, the default is 1.5 IQR and 3 z
- How to treat outliers. The user can assign a method to each column. The methods include specifying a range of allowed values, imputing the mean, imputing the nearest non-extreme value (winsorizing) and binning (here, the user has to provide the number of bins and if the bins should be of equal width or of equal frequency)
- Which variables should be reverted. In surveys, some items are often asked reversely to maintain user attention throughout the survey (e.g., item 1: “I feel good today” and item 2: “I feel terrible today. Therefore, the user can decide to revert the scale of one item by subtracting the maximum of the column from each value
- Which columns should be combined into a new column. For example, the two items mentioned before could be replaced with one item named “well-being” that is the mean of the column for item 1 and the (reverted) column for item 2. In this case, the internal consistency of the generated variable is reported with Cronbach’s alpha
- If variables with high variance inflation factors (vif) should be dropped and how high the threshold should be
- If a subset of features should be selected. In this case, RFE (recursive feature elimination) is used, which returns the set of variables that achieve the highest fit for OLS regression on the dependent variable

After the overall architecture was designed, the individual components were programmed. The core component, EXPLAIN_run, which conducts all regression analyses, was then programmed using multiple steps. Firstly, literature research was conducted to provide best practices of regression

analyses in the targeted context. This context was found in publicly available academic theses. Secondly, an example output report template as Word (.docx) document was created, including all relevant information (variables and pictures) to be obtained and filled in by the program. Thirdly, the module was prototyped. Fourthly, the module was tested with multiple data sets and iteratively refined with expert input.

The literature review on regression analyses in a research context was also used to identify relevant hypotheses that the prototype needs to be able to test. The study identified four types of hypotheses: Linear, nonlinear (also called polynomial or quadratic), moderation and mediation effects. These are again splitted into different subcategories, which represent different curve specifications regarding the relationship between an independent variable and a dependent variable (as well as mediators and moderators) witin one of the four hypotheses types. The hypotheses are illustrated below.

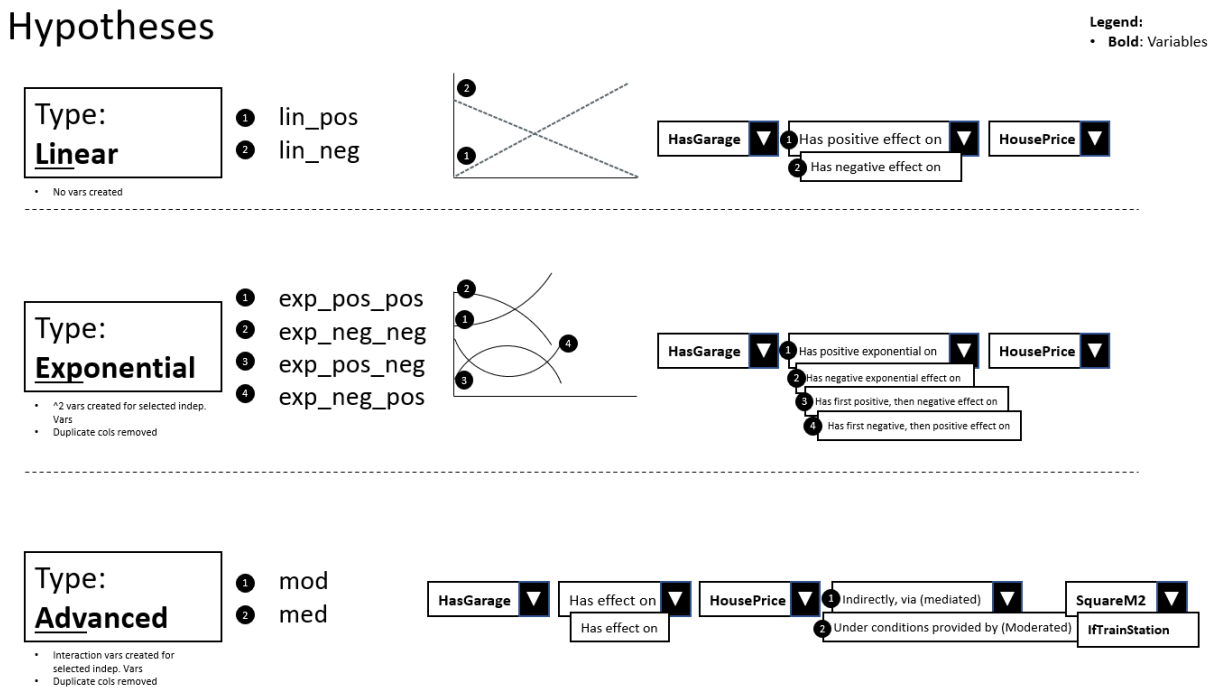
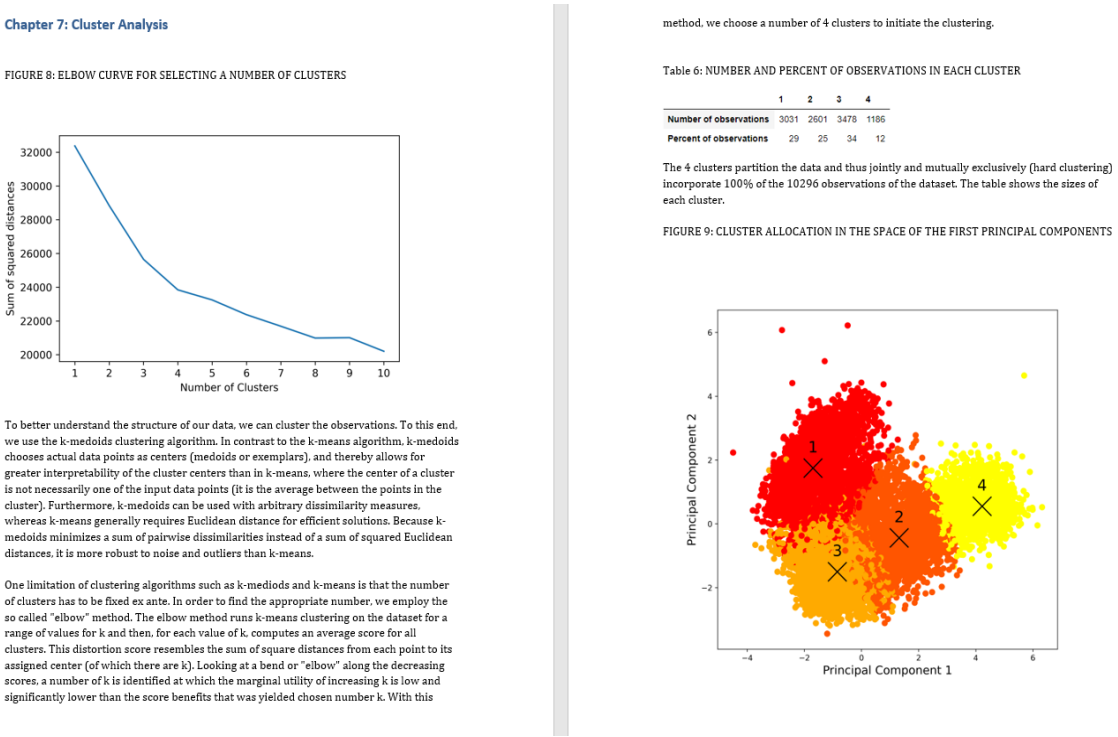


Figure 3.3 – Hypotheses Types and Curve Specifications

Demonstration

In the demonstration domain, this study considers screenshots from real Word (.docx) report outputs to be the most effective way of conveying the outcome of the project to a research audience. An example of this is provided below. This is a screenshot of the initial prototype which automated cluster analysis and principal component analysis. This report focuses, as stated before, on linear regression analysis.



Chapter 7: Cluster Analysis

FIGURE 8: ELBOW CURVE FOR SELECTING A NUMBER OF CLUSTERS

To better understand the structure of our data, we can cluster the observations. To this end, we use the k-medoids clustering algorithm. In contrast to the k-means algorithm, k-medoids chooses actual data points as centers (medoids or exemplars), and thereby allows for greater interpretability of the cluster centers than in k-means, where the center of a cluster is not necessarily one of the input data points (it is the average between the points in the cluster). Furthermore, k-medoids can be used with arbitrary dissimilarity measures, whereas k-means generally requires Euclidean distance for efficient solutions. Because k-medoids minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances, it is more robust to noise and outliers than k-means.

One limitation of clustering algorithms such as k-mediods and k-means is that the number of clusters has to be fixed ex ante. In order to find the appropriate number, we employ the so called "elbow" method. The elbow method runs k-means clustering on the dataset for a range of values for k and then, for each value of k, computes an average score for all clusters. This distortion score resembles the sum of square distances from each point to its assigned center (of which there are k). Looking at a bend or "elbow" along the decreasing scores, a number of k is identified at which the marginal utility of increasing k is low and significantly lower than the score benefits that was yielded chosen number k. With this method, we choose a number of 4 clusters to initiate the clustering.

Table 6: NUMBER AND PERCENT OF OBSERVATIONS IN EACH CLUSTER

	1	2	3	4
Number of observations	3031	2601	3478	1186
Percent of observations	29	25	34	12

The 4 clusters partition the data and thus jointly and mutually exclusively (hard clustering) incorporate 100% of the 10296 observations of the dataset. The table shows the sizes of each cluster.

FIGURE 9: CLUSTER ALLOCATION IN THE SPACE OF THE FIRST PRINCIPAL COMPONENTS

Figure 3.4 – Excerpt from the Data Exploration Report

Evaluation

In the Evaluation domain, this study determines an approach to testing the prototype with suitable data sets, some of which have already been identified during the prototyping phase. In addition, the study analyses the potential of implementing a web application to offer the program as an automated service to potential customers. The evaluation results are part of the results chapter of this report.

Communication

In the Communication domain, this project report will be the primary vehicle of communicating the research outcome. Additional means of communication such as blog posts will also be considered.

Contribution

In the Contribution domain, this study reflects on its relevance to society and in specific the data anlyses ecosystem. The contribution can be summarized as follows: The program prototyped in this study has the potential to take stress and time investment out of common data analysis tasks in educational and business contexts and to show that untapped potential lies in automated and guided analytics programs that aim at enabling unskilled users to use advanced data analysis.

4. RESULTS AND DISCUSSION

The core output of the prototype is the regression analyses report. It is structured in the way of a typical empirical research report. This structure is shown below:

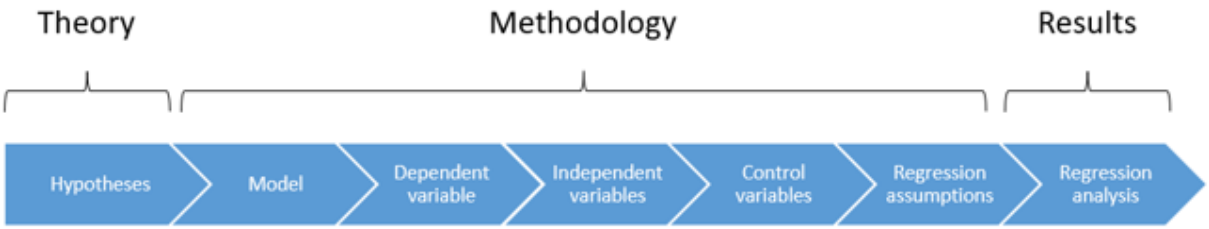


Figure 4.1 – Chapters of the Report

In the following section, the key elements of the report will be shown. They resemble the automated prototype output to a user-provided data set and are designed at helping students and professionals during the creation of research reports.

4.1. COMMENTED SCREENSHOTS OF THE RESULTS REPORT

The example software output (report) will be presented in the form of a screenshot of pages in the same order as they are arranged in the report document, accompanied with an explanatory section. The report shown is automatically generated based on the Smart Home Campaign dataset as example data set (Henriques, 2021). An excerpt of the dataset is provided in the appendix.
Each observation of the dataset represents a customer, and each column represents a piece of information on the customer, such as the amount spent on products within a certain product category.

The full report is provided in the appendix. In this section, the most relevant pages will be selected and discussed. Because the whole report is an automatically generated document that needs to be able to react flexibly to vastly different user input, formatting is not always perfect. For example, sometimes the title of a graph is not on the same page as the graph itself. Also, the sizing of individual tables and graphs is sometimes not ideal. The user has to adjust the formatting as required for his individual needs.

Because the automatically generated report is a prototype, academic citation is not yet incorporated rigorously. The automatic summary of sources and citations as well as the optimization of formatting are incorporated into future steps towards the final, comprehensive software.

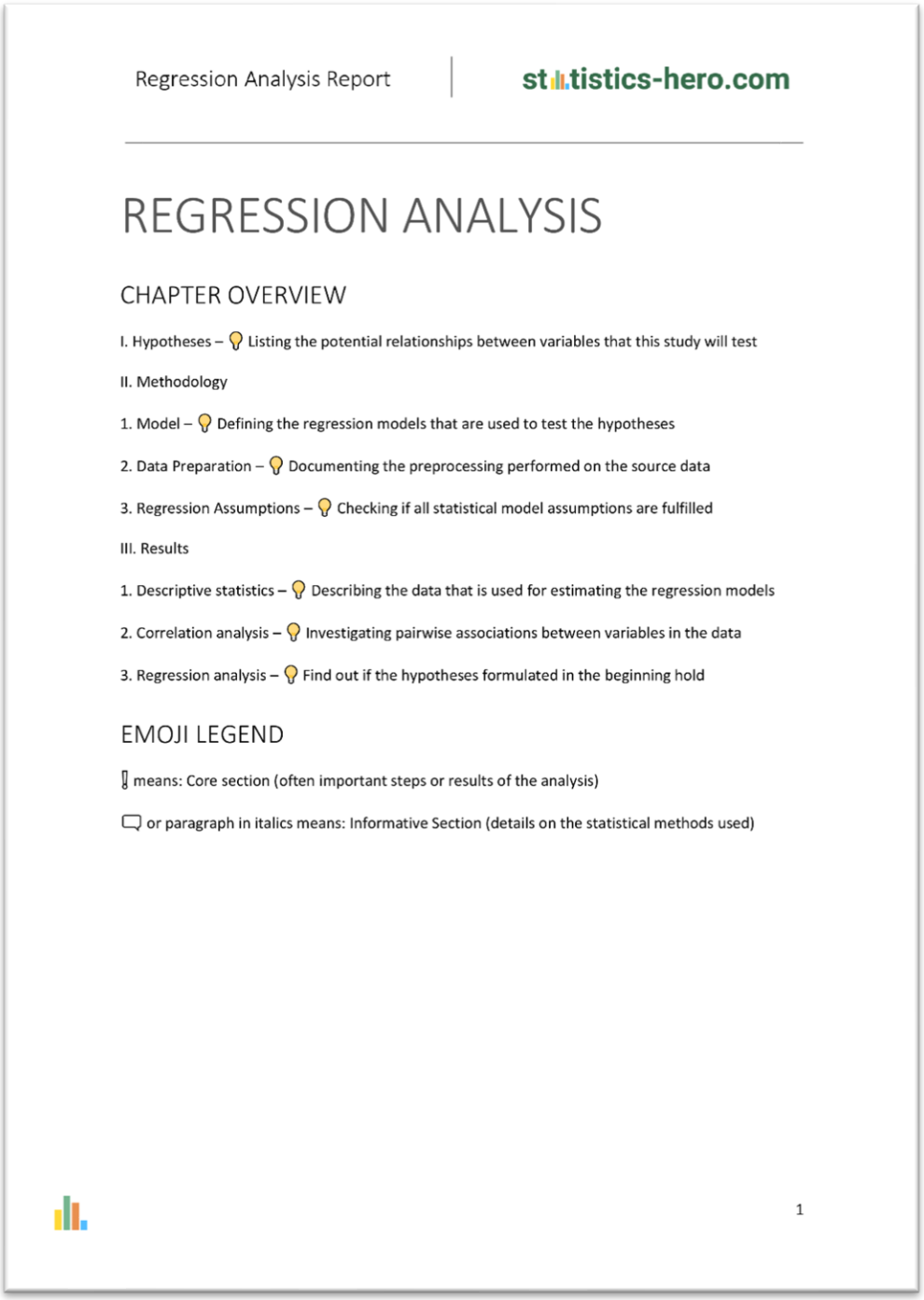
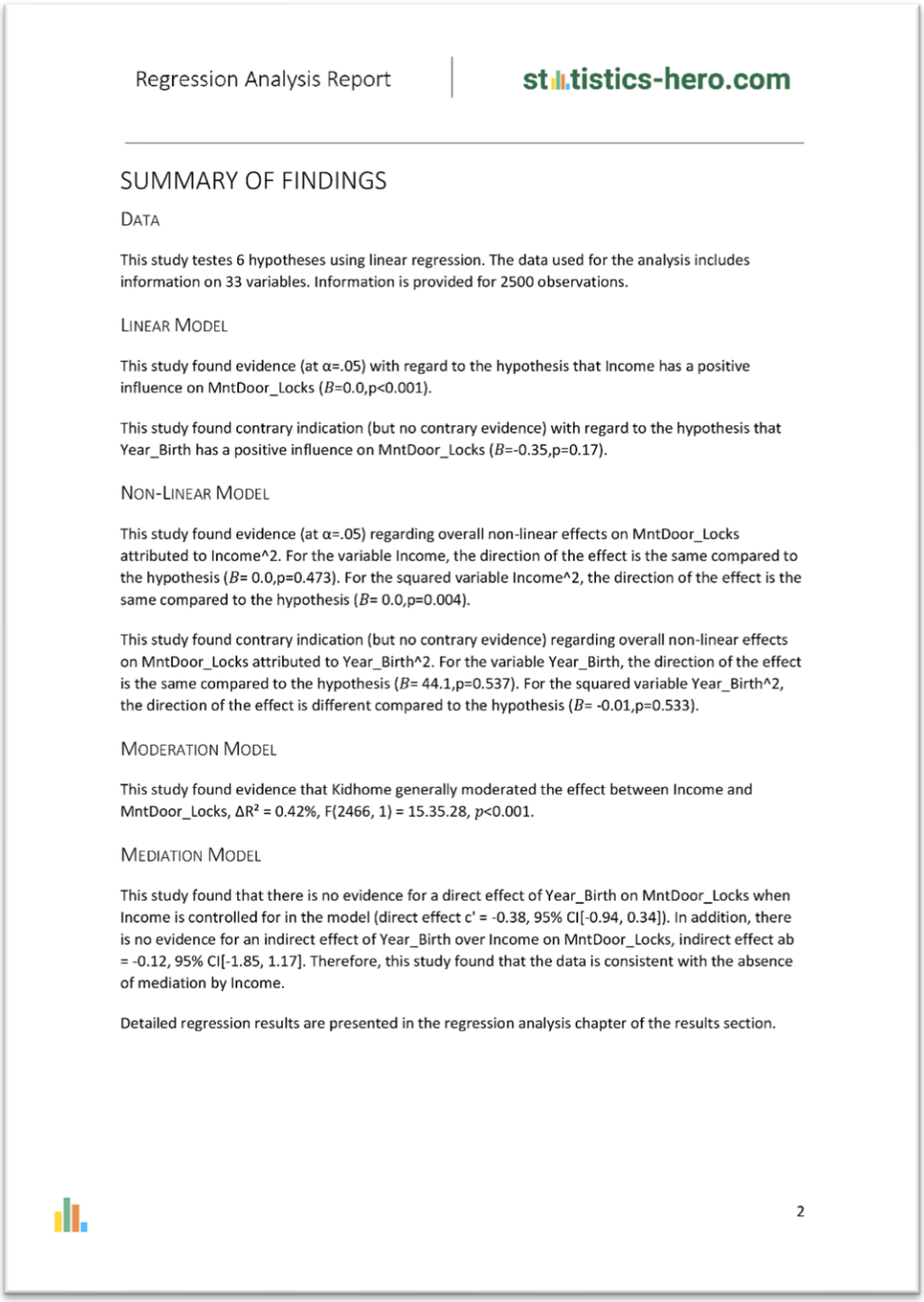


Figure 4.2 – Table of Contents - Screenshot

The first page of the report details the structure of the report and explains briefly what the contents of the individual chapters are. A page number is not provided because pagination is a function which is used by the Microsoft Word layout engine. Therefore, obtaining the page number within a document is a difficult task for automation tasks.

A brief legend explains the meaning of the two emoji characters which are used to point out particularly important text sections in order to help the user maintain a good overview. In order to help the user to identify explanatory sections, italic font is used.



Regression Analysis Report | statistics-hero.com

SUMMARY OF FINDINGS

DATA

This study testes 6 hypotheses using linear regression. The data used for the analysis includes information on 33 variables. Information is provided for 2500 observations.

LINEAR MODEL

This study found evidence (at α=.05) with regard to the hypothesis that Income has a positive influence on MntDoor_Locks (*B*=0.0,p<0.001).

This study found contrary indication (but no contrary evidence) with regard to the hypothesis that Year_Birth has a positive influence on MntDoor_Locks (*B*=-0.35,p=0.17).

NON-LINEAR MODEL

This study found evidence (at α=.05) regarding overall non-linear effects on MntDoor_Locks attributed to Income^2. For the variable Income, the direction of the effect is the same compared to the hypothesis (*B*= 0.0,p=0.473). For the squared variable Income^2, the direction of the effect is the same compared to the hypothesis (*B*= 0.0,p=0.004).

This study found contrary indication (but no contrary evidence) regarding overall non-linear effects on MntDoor_Locks attributed to Year_Birth^2. For the variable Year_Birth, the direction of the effect is the same compared to the hypothesis (*B*= 44.1,p=0.537). For the squared variable Year_Birth^2, the direction of the effect is different compared to the hypothesis (*B*= -0.01,p=0.533).

MODERATION MODEL

This study found evidence that Kidhome generally moderated the effect between Income and MntDoor_Locks, ΔR^2 = 0.42%, F(2466, 1) = 15.35.28, *p*<0.001.

MEDIATION MODEL

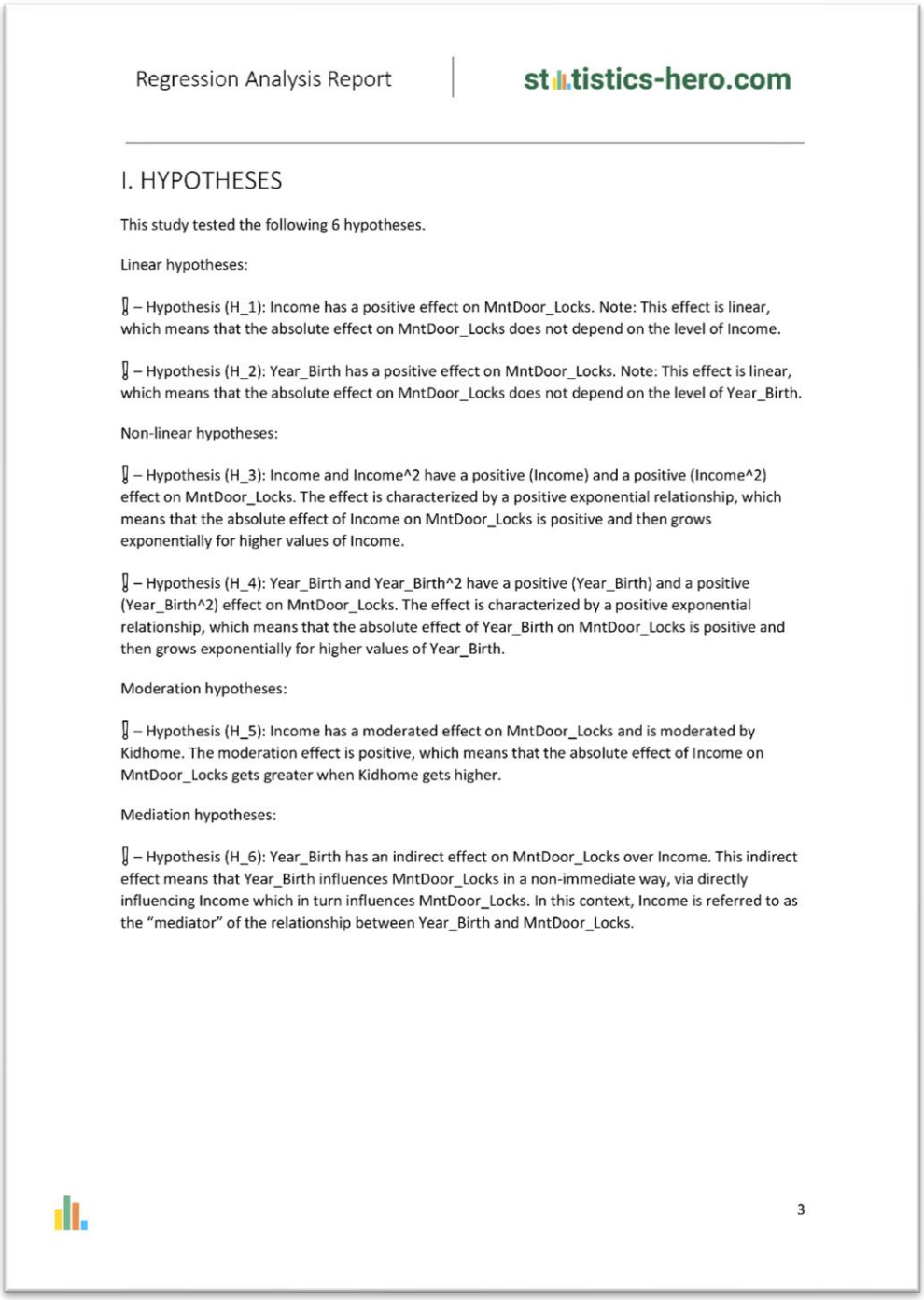
This study found that there is no evidence for a direct effect of Year_Birth on MntDoor_Locks when Income is controlled for in the model (direct effect c' = -0.38, 95% CI[-0.94, 0.34]). In addition, there is no evidence for an indirect effect of Year_Birth over Income on MntDoor_Locks, indirect effect ab = -0.12, 95% CI[-1.85, 1.17]. Therefore, this study found that the data is consistent with the absence of mediation by Income.

Detailed regression results are presented in the regression analysis chapter of the results section.

2

Figure 4.3 – Executive Summary – Screenshot

The executive summary gives a brief high-level overview over the findings of the different hypotheses (which the user already knows as he provided them). The four different types of hypotheses (linear, non-linear, moderation and mediation – see next section for details) are each tested with one or more (in the case of mediation) regression models. Therefore, one model tests for one type of hypothesis. The structural differentiation between the four types of hypotheses and the associated models will repeated to structure the report. Of course, if a user does not provide an instance of a hypothesis type, no hypotheses and models will be reported.



Regression Analysis Report | statistics-hero.com

I. HYPOTHESES

This study tested the following 6 hypotheses.

Linear hypotheses:

– Hypothesis (H_1): Income has a positive effect on MntDoor_Locks. Note: This effect is linear, which means that the absolute effect on MntDoor_Locks does not depend on the level of Income.

– Hypothesis (H_2): Year_Birth has a positive effect on MntDoor_Locks. Note: This effect is linear, which means that the absolute effect on MntDoor_Locks does not depend on the level of Year_Birth.

Non-linear hypotheses:

– Hypothesis (H_3): Income and Income^2 have a positive (Income) and a positive (Income^2) effect on MntDoor_Locks. The effect is characterized by a positive exponential relationship, which means that the absolute effect of Income on MntDoor_Locks is positive and then grows exponentially for higher values of Income.

– Hypothesis (H_4): Year_Birth and Year_Birth^2 have a positive (Year_Birth) and a positive (Year_Birth^2) effect on MntDoor_Locks. The effect is characterized by a positive exponential relationship, which means that the absolute effect of Year_Birth on MntDoor_Locks is positive and then grows exponentially for higher values of Year_Birth.

Moderation hypotheses:

– Hypothesis (H_5): Income has a moderated effect on MntDoor_Locks and is moderated by Kidhome. The moderation effect is positive, which means that the absolute effect of Income on MntDoor_Locks gets greater when Kidhome gets higher.

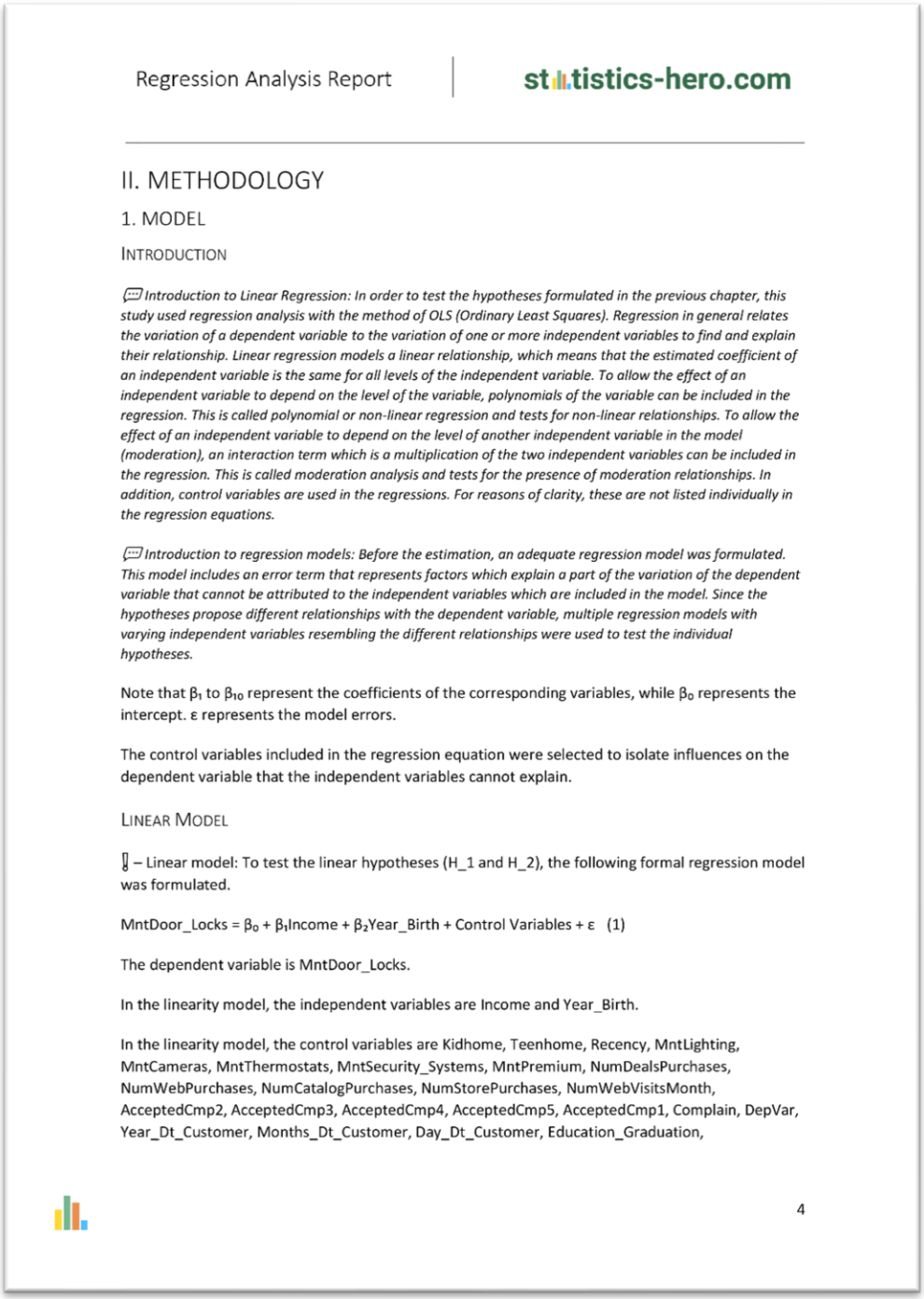
Mediation hypotheses:

– Hypothesis (H_6): Year_Birth has an indirect effect on MntDoor_Locks over Income. This indirect effect means that Year_Birth influences MntDoor_Locks in a non-immediate way, via directly influencing Income which in turn influences MntDoor_Locks. In this context, Income is referred to as the "mediator" of the relationship between Year_Birth and MntDoor_Locks.

3

Figure 4.4 – Hypotheses – Screenshot

The hypotheses chapter puts into text form the hypotheses that the user has provided to the prototype. In a web tool, these hypotheses would be collected from the user with a dropdown-menu type interface where the user selects the variables and relationships from a list. The hypotheses are sorted into linear hypotheses, non-linear hypotheses (where a quadratic term is used to detect non-linear relationships), moderation hypotheses and mediation hypotheses. Of the first two hypothesis types, an arbitrary number of hypotheses can be provided by the user. In the prototype, currently the number of moderation and mediation hypotheses is restricted to one each.



Regression Analysis Report | statistics-hero.com

II. METHODOLOGY

1. MODEL

INTRODUCTION

Introduction to Linear Regression: In order to test the hypotheses formulated in the previous chapter, this study used regression analysis with the method of OLS (Ordinary Least Squares). Regression in general relates the variation of a dependent variable to the variation of one or more independent variables to find and explain their relationship. Linear regression models a linear relationship, which means that the estimated coefficient of an independent variable is the same for all levels of the independent variable. To allow the effect of an independent variable to depend on the level of the variable, polynomials of the variable can be included in the regression. This is called polynomial or non-linear regression and tests for non-linear relationships. To allow the effect of an independent variable to depend on the level of another independent variable in the model (moderation), an interaction term which is a multiplication of the two independent variables can be included in the regression. This is called moderation analysis and tests for the presence of moderation relationships. In addition, control variables are used in the regressions. For reasons of clarity, these are not listed individually in the regression equations.

Introduction to regression models: Before the estimation, an adequate regression model was formulated. This model includes an error term that represents factors which explain a part of the variation of the dependent variable that cannot be attributed to the independent variables which are included in the model. Since the hypotheses propose different relationships with the dependent variable, multiple regression models with varying independent variables resembling the different relationships were used to test the individual hypotheses.

Note that β_1 to β_{10} represent the coefficients of the corresponding variables, while β_0 represents the intercept. ε represents the model errors.

The control variables included in the regression equation were selected to isolate influences on the dependent variable that the independent variables cannot explain.

LINEAR MODEL

– Linear model: To test the linear hypotheses (H_1 and H_2), the following formal regression model was formulated.

$$MntDoor\_Locks = \beta_0 + \beta_1 Income + \beta_2 Year\_Birth + Control\ Variables + \varepsilon \quad (1)$$

The dependent variable is MntDoor_Locks.

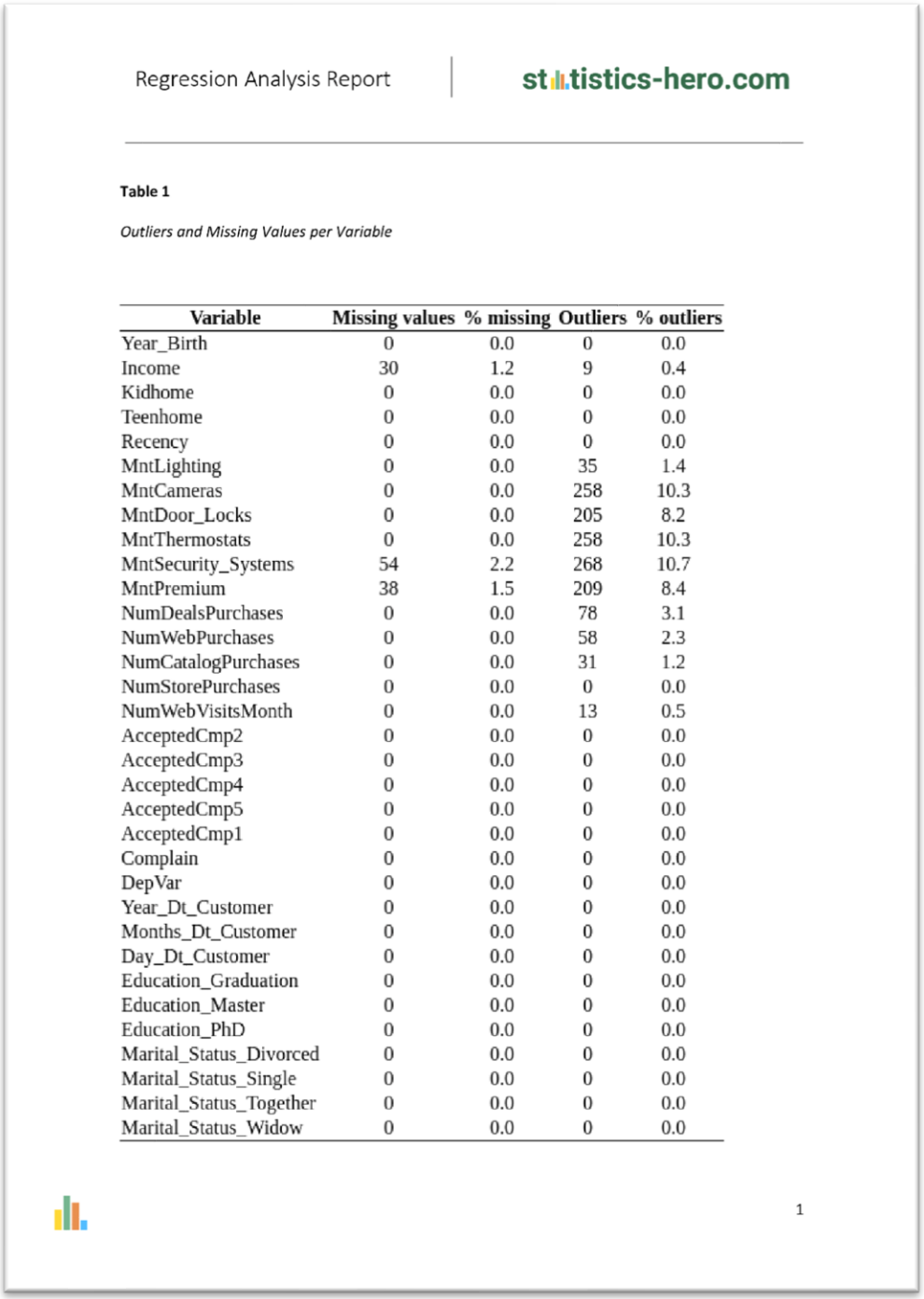
In the linearity model, the independent variables are Income and Year_Birth.

In the linearity model, the control variables are Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation,

4

Figure 4.5 – Models – Screenshot

For each type of hypothesis, a model is formulated and the required variables are listed. In the example listed above, the linear model is specified with all independent variables from the linear hypotheses provided by the used. The subsequent sections (not shown) repeat the process for the non-linear model (which includes quadratic terms for each independent variable provided), the moderation model (which includes an interaction term between the independent variable and the moderator variable provided) and the different mediation models (outcome and mediation model) required to estimate the different mediation steps. Of course, when no hypothesis concerning e.g., moderation is provided, the section is skipped and no model is built.



Regression Analysis Report | statistics-hero.com

Table 1

Outliers and Missing Values per Variable

Variable	Missing values	% missing	Outliers	% outliers
Year_Birth	0	0.0	0	0.0
Income	30	1.2	9	0.4
Kidhome	0	0.0	0	0.0
Teenhome	0	0.0	0	0.0
Recency	0	0.0	0	0.0
MntLighting	0	0.0	35	1.4
MntCameras	0	0.0	258	10.3
MntDoor_Locks	0	0.0	205	8.2
MntThermostats	0	0.0	258	10.3
MntSecurity_Systems	54	2.2	268	10.7
MntPremium	38	1.5	209	8.4
NumDealsPurchases	0	0.0	78	3.1
NumWebPurchases	0	0.0	58	2.3
NumCatalogPurchases	0	0.0	31	1.2
NumStorePurchases	0	0.0	0	0.0
NumWebVisitsMonth	0	0.0	13	0.5
AcceptedCmp2	0	0.0	0	0.0
AcceptedCmp3	0	0.0	0	0.0
AcceptedCmp4	0	0.0	0	0.0
AcceptedCmp5	0	0.0	0	0.0
AcceptedCmp1	0	0.0	0	0.0
Complain	0	0.0	0	0.0
DepVar	0	0.0	0	0.0
Year_Dt_Customer	0	0.0	0	0.0
Months_Dt_Customer	0	0.0	0	0.0
Day_Dt_Customer	0	0.0	0	0.0
Education_Graduation	0	0.0	0	0.0
Education_Master	0	0.0	0	0.0
Education_PhD	0	0.0	0	0.0
Marital_Status_Divorced	0	0.0	0	0.0
Marital_Status_Single	0	0.0	0	0.0
Marital_Status_Together	0	0.0	0	0.0
Marital_Status_Widow	0	0.0	0	0.0

1

Figure 4.6 – Data Preprocessing I – Screenshot

The preprocessing chapter begins with explaining the necessity of preprocessing (not shown) and then showing a table with information on missing values and identified potential outliers per variable (above).

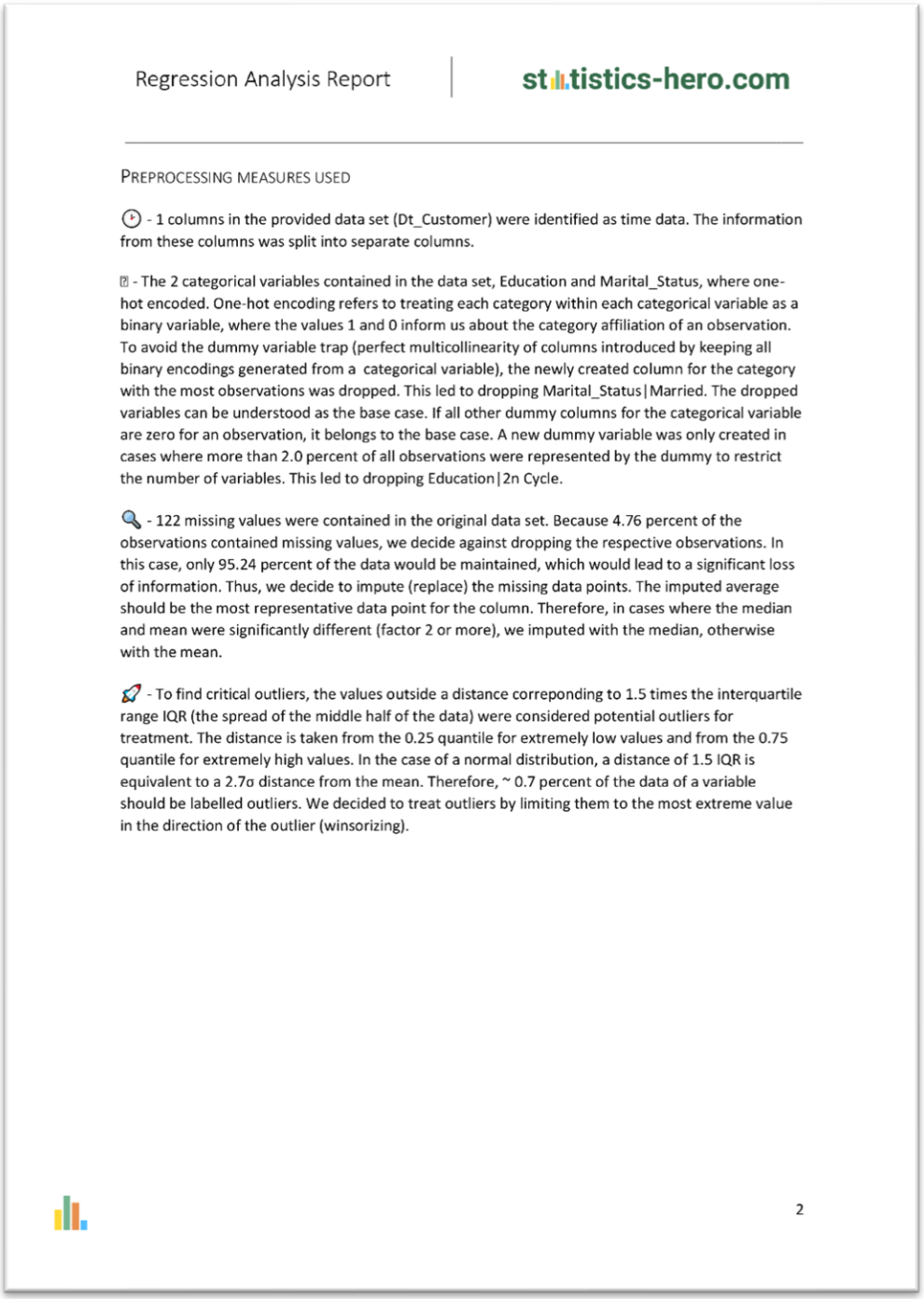
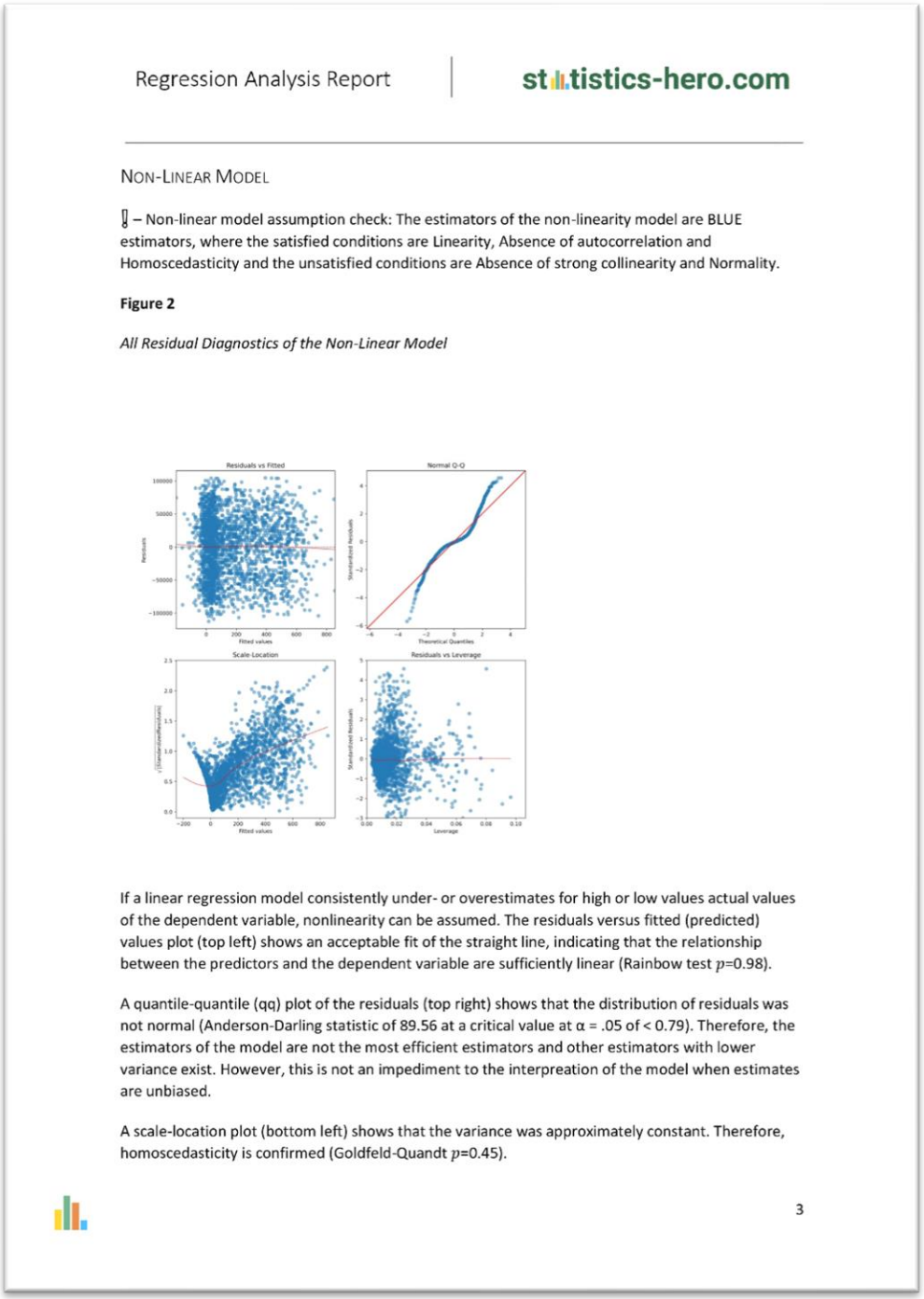


Figure 4.7 – Data Preprocessing II – Screenshot

The chapter on preprocessing continues with detailing important steps of the preprocessing process, including the treatment of missing values, potential outliers and categorical variables. For each step, default procedures are in place. However, the user is able to specify specific procedures if he prefers. If the data set would have contained any multi-item variables that have to be combined to a compound variable or any column with values to be reverted, this information would also be displayed here.



Regression Analysis Report | statistics-hero.com

NON-LINEAR MODEL

– Non-linear model assumption check: The estimators of the non-linearity model are BLUE estimators, where the satisfied conditions are Linearity, Absence of autocorrelation and Homoscedasticity and the unsatisfied conditions are Absence of strong collinearity and Normality.

Figure 2

All Residual Diagnostics of the Non-Linear Model

If a linear regression model consistently under- or overestimates for high or low values actual values of the dependent variable, nonlinearity can be assumed. The residuals versus fitted (predicted) values plot (top left) shows an acceptable fit of the straight line, indicating that the relationship between the predictors and the dependent variable are sufficiently linear (Rainbow test p=0.98).

A quantile-quantile (qq) plot of the residuals (top right) shows that the distribution of residuals was not normal (Anderson-Darling statistic of 89.56 at a critical value at α = .05 of < 0.79). Therefore, the estimators of the model are not the most efficient estimators and other estimators with lower variance exist. However, this is not an impediment to the interpreation of the model when estimates are unbiased.

A scale-location plot (bottom left) shows that the variance was approximately constant. Therefore, homoscedasticity is confirmed (Goldfeld-Quandt p=0.45).

3

Figure 4.8 – egression Assumptions I – Screenshot

After two pages which introduces the different Gauss-Markov assumptions as well as the different qualities of estimators, all models are checked to find out if they return biased estimators, which would make drawing conclusions from the results difficult. In the screenshot above, the non-linearity model is under investigation. The checks and diagnostics performed are visualized in a set of graphs.

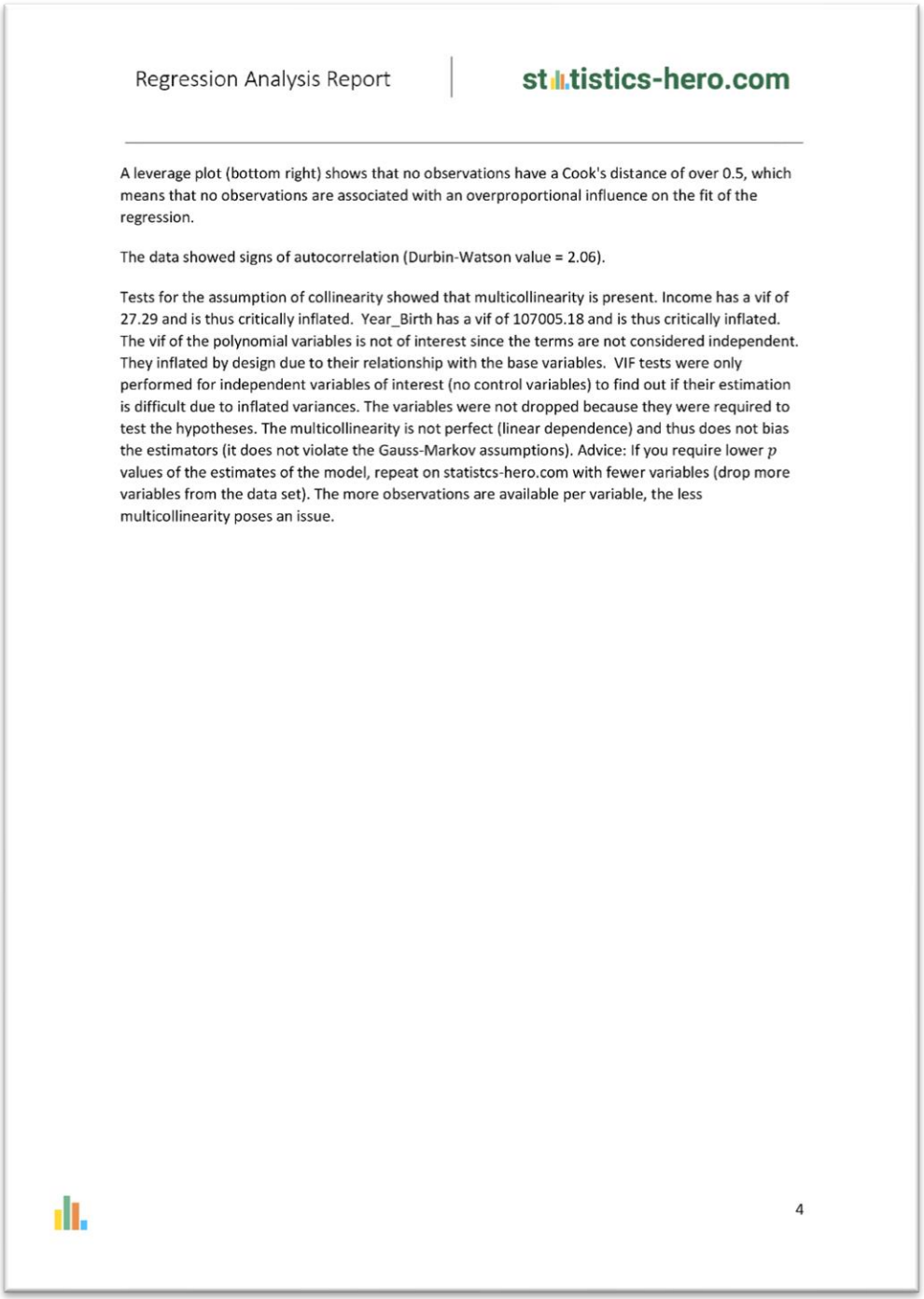
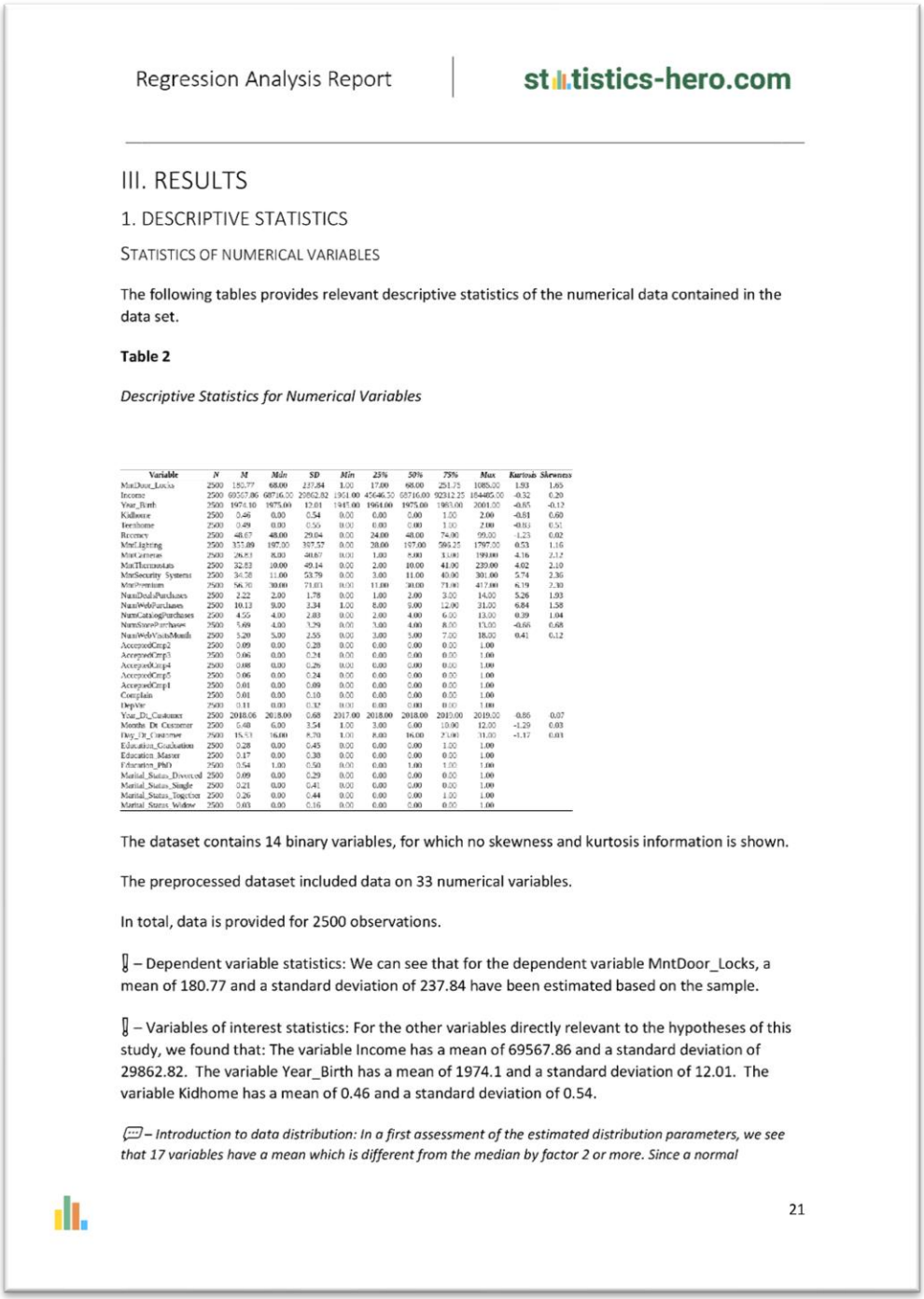


Figure 4.9 – Regression Assumptions II – Screenshot

The regression assumptions chapter also discusses potential multicollinearity problems by discussing the variance inflation factors (vif) of the relevant variables.



Regression Analysis Report | statistics-hero.com

III. RESULTS

1. DESCRIPTIVE STATISTICS

STATISTICS OF NUMERICAL VARIABLES

The following tables provides relevant descriptive statistics of the numerical data contained in the data set.

Table 2

Descriptive Statistics for Numerical Variables

Variable	N	M	Mdn	SD	Min	25%	50%	75%	Max	Kurtosis	Skewness
MntDoor_Locks	2500	180.77	68.00	237.84	1.00	17.00	68.00	251.75	1085.00	1.93	1.65
Income	2500	69567.86	68716.00	29862.82	1961.00	45646.50	68716.00	92312.25	164485.00	-0.32	0.20
Year_Birth	2500	1974.10	1975.00	12.01	1947.00	1964.00	1975.00	1985.00	2001.00	-0.65	-0.17
Kidhome	2500	0.46	0.00	0.54	0.00	0.00	0.00	1.00	2.00	-0.81	0.60
Teenhome	2500	0.49	0.00	0.56	0.00	0.00	0.00	1.00	2.00	-0.83	0.51
Recency	2500	48.67	48.00	29.04	0.00	24.00	48.00	74.00	99.00	-1.23	0.02
MntLighting	2500	357.09	197.00	357.57	0.00	20.00	197.00	566.25	1797.00	0.53	1.16
MntCameras	2500	26.83	8.00	40.87	0.00	1.00	8.00	33.00	199.00	4.16	2.12
MntThermostats	2500	32.83	10.00	49.14	0.00	2.00	10.00	41.00	230.00	4.02	2.10
MntSecurity_Systems	2500	34.58	11.00	53.79	0.00	3.00	11.00	40.00	301.00	5.74	2.36
MntPremium	2500	56.70	30.00	71.03	0.00	11.00	30.00	73.00	417.00	6.19	2.30
NumDeal-Purchases	2500	2.22	2.00	1.78	0.00	1.00	2.00	3.00	14.00	5.26	1.93
NumWebPurchases	2500	10.13	9.00	3.34	1.00	8.00	9.00	12.00	31.00	6.84	1.58
NumCatalogPurchases	2500	4.55	4.00	2.83	0.00	2.00	4.00	6.00	13.00	0.39	1.04
NumStorePurchases	2500	5.69	4.00	3.29	0.00	3.00	4.00	8.00	13.00	-0.66	0.68
NumWebVisitsMonth	2500	5.20	5.00	2.55	0.00	3.00	5.00	7.00	18.00	0.41	0.12
AcceptedCmp2	2500	0.09	0.00	0.28	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp3	2500	0.06	0.00	0.24	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp4	2500	0.08	0.00	0.26	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp5	2500	0.06	0.00	0.24	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp1	2500	0.01	0.00	0.09	0.00	0.00	0.00	0.00	1.00		
Complain	2500	0.01	0.00	0.10	0.00	0.00	0.00	0.00	1.00		
DepVar	2500	0.11	0.00	0.32	0.00	0.00	0.00	0.00	1.00		
Year_Dt_Customer	2500	2018.06	2018.00	0.68	2017.00	2018.00	2018.00	2019.00	2019.00	-0.86	0.07
Months_Dt_Customer	2500	6.48	6.00	3.54	1.00	3.00	6.00	10.00	12.00	-1.29	0.03
Day_Dt_Customer	2500	15.53	16.00	8.70	1.00	8.00	16.00	23.00	31.00	-1.17	0.03
Education_Graduation	2500	0.28	0.00	0.45	0.00	0.00	0.00	1.00	1.00		
Education_Master	2500	0.17	0.00	0.38	0.00	0.00	0.00	0.00	1.00		
Education_PhD	2500	0.54	1.00	0.50	0.00	0.00	1.00	1.00	1.00		
Marital_Status_Divorced	2500	0.09	0.00	0.29	0.00	0.00	0.00	0.00	1.00		
Marital_Status_Single	2500	0.21	0.00	0.41	0.00	0.00	0.00	0.00	1.00		
Marital_Status_Together	2500	0.26	0.00	0.44	0.00	0.00	0.00	1.00	1.00		
Marital_Status_Widow	2500	0.03	0.00	0.16	0.00	0.00	0.00	0.00	1.00		

The dataset contains 14 binary variables, for which no skewness and kurtosis information is shown.

The preprocessed dataset included data on 33 numerical variables.

In total, data is provided for 2500 observations.

– Dependent variable statistics: We can see that for the dependent variable MntDoor_Locks, a mean of 180.77 and a standard deviation of 237.84 have been estimated based on the sample.

– Variables of interest statistics: For the other variables directly relevant to the hypotheses of this study, we found that: The variable Income has a mean of 69567.86 and a standard deviation of 29862.82. The variable Year_Birth has a mean of 1974.1 and a standard deviation of 12.01. The variable Kidhome has a mean of 0.46 and a standard deviation of 0.54.

– Introduction to data distribution: In a first assessment of the estimated distribution parameters, we see that 17 variables have a mean which is different from the median by factor 2 or more. Since a normal

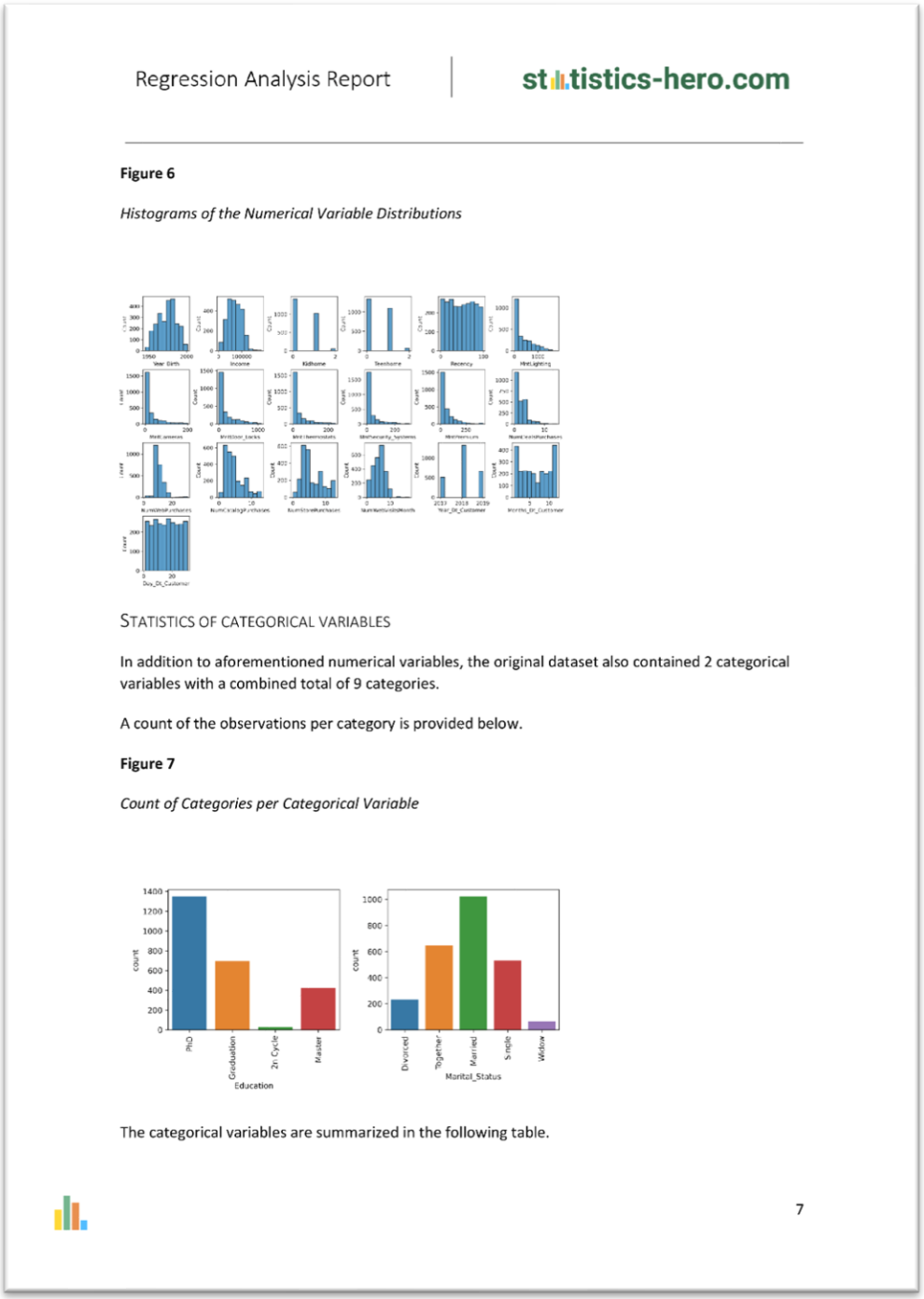
21

Figure 4.10 – Descriptive Statistics – Screenshot

The results chapter starts with the descriptive statistics of all variables. The dependent variable and the independent (and moderator/mediator) variables are elaborated on in the text section.

Figure 4.11 – Descriptive Statistics, Distributions – Screenshot

Information on the distributions of the dependent and independent variables is presented in text form.



Regression Analysis Report | statistics-hero.com

Figure 6

Histograms of the Numerical Variable Distributions

STATISTICS OF CATEGORICAL VARIABLES

In addition to aforementioned numerical variables, the original dataset also contained 2 categorical variables with a combined total of 9 categories.

A count of the observations per category is provided below.

Figure 7

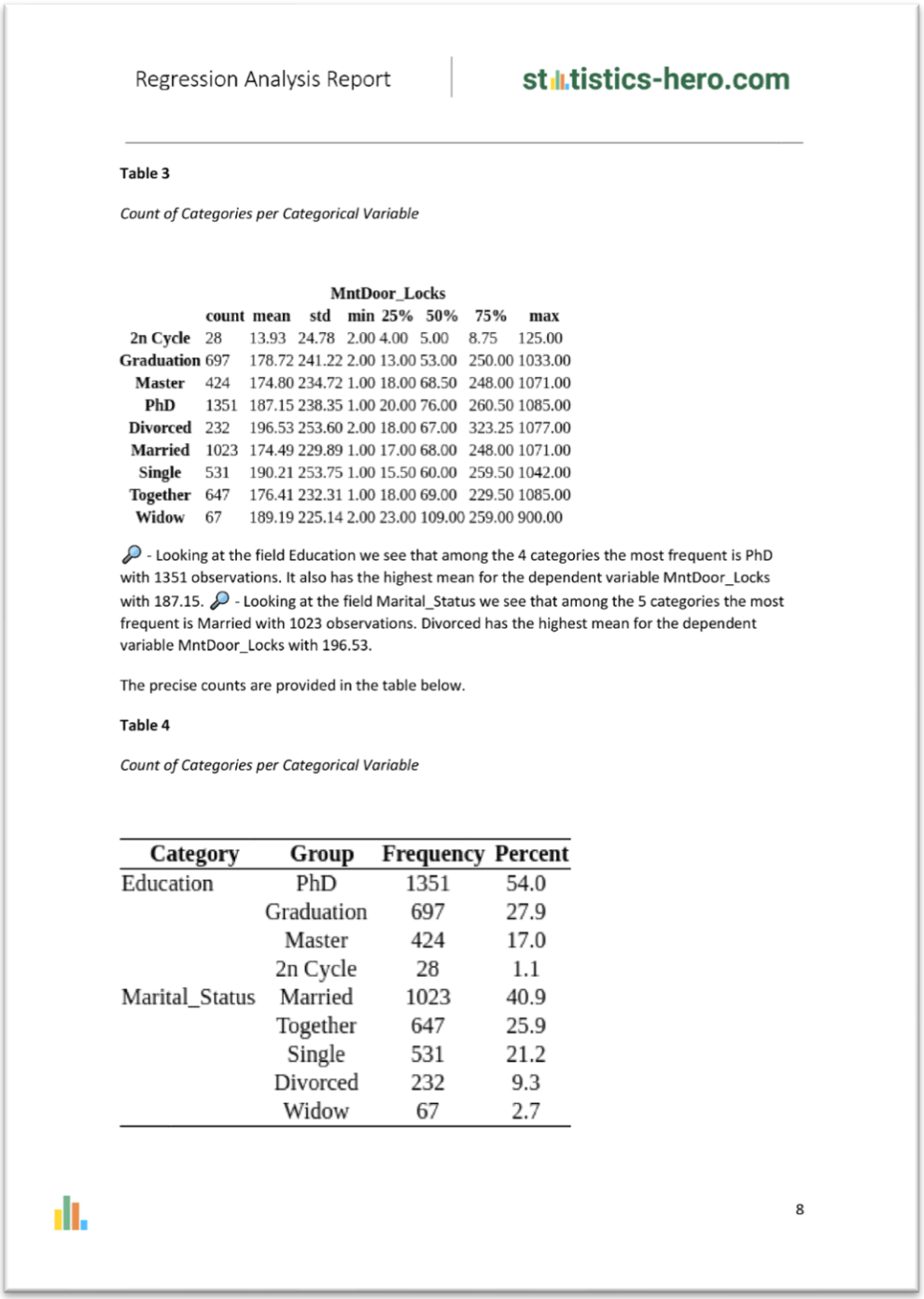
Count of Categories per Categorical Variable

The categorical variables are summarized in the following table.

7

Figure 4.12 – Descriptive statistics, Distribution Diagrams – Screenshot

The distributions of numerical and categorical variables is visualized to support the user in developing a sound understanding of the data set.



Regression Analysis Report | statistics-hero.com

Table 3

Count of Categories per Categorical Variable

	MntDoor_Locks						
	count	mean	std	min	25%	50%	75%	max
2n Cycle	28	13.93	24.78	2.00	4.00	5.00	8.75	125.00
Graduation	697	178.72	241.22	2.00	13.00	53.00	250.00	1033.00
Master	424	174.80	234.72	1.00	18.00	68.50	248.00	1071.00
PhD	1351	187.15	238.35	1.00	20.00	76.00	260.50	1085.00
Divorced	232	196.53	253.60	2.00	18.00	67.00	323.25	1077.00
Married	1023	174.49	229.89	1.00	17.00	68.00	248.00	1071.00
Single	531	190.21	253.75	1.00	15.50	60.00	259.50	1042.00
Together	647	176.41	232.31	1.00	18.00	69.00	229.50	1085.00
Widow	67	189.19	225.14	2.00	23.00	109.00	259.00	900.00

🔑 - Looking at the field Education we see that among the 4 categories the most frequent is PhD with 1351 observations. It also has the highest mean for the dependent variable MntDoor_Locks with 187.15. 🔑 - Looking at the field Marital_Status we see that among the 5 categories the most frequent is Married with 1023 observations. Divorced has the highest mean for the dependent variable MntDoor_Locks with 196.53.

The precise counts are provided in the table below.

Table 4

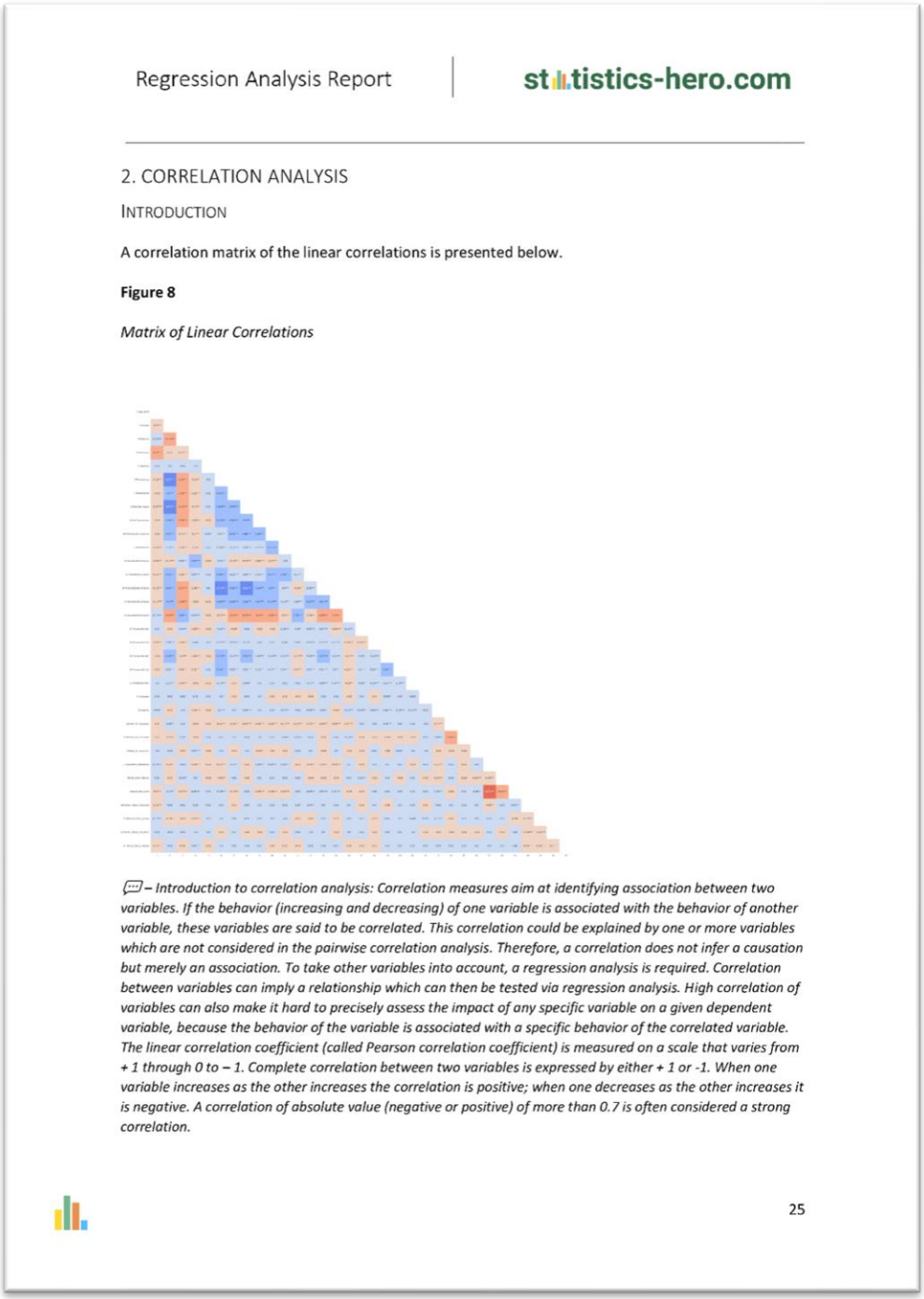
Count of Categories per Categorical Variable

Category	Group	Frequency	Percent
Education	PhD	1351	54.0
	Graduation	697	27.9
	Master	424	17.0
	2n Cycle	28	1.1
Marital_Status	Married	1023	40.9
	Together	647	25.9
	Single	531	21.2
	Divorced	232	9.3
	Widow	67	2.7

8

Figure 4.13 – Descriptive Statistics, Categories – Screenshot

Further information on the categories of each categorical variable is included in the last part of the descriptive statistics chapter.



Regression Analysis Report | statistics-hero.com

2. CORRELATION ANALYSIS

INTRODUCTION

A correlation matrix of the linear correlations is presented below.

Figure 8

Matrix of Linear Correlations

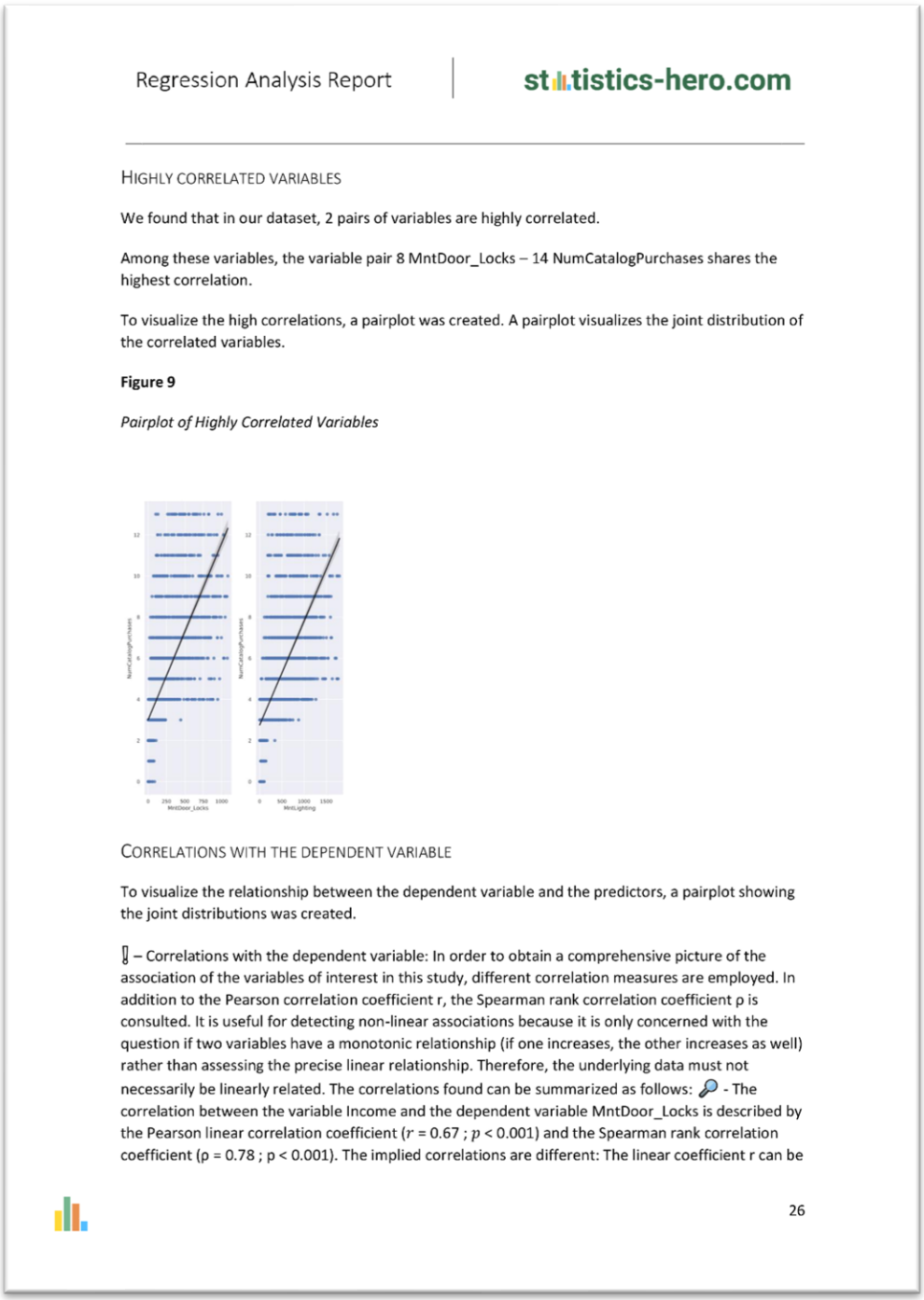
– Introduction to correlation analysis: Correlation measures aim at identifying association between two variables. If the behavior (increasing and decreasing) of one variable is associated with the behavior of another variable, these variables are said to be correlated. This correlation could be explained by one or more variables which are not considered in the pairwise correlation analysis. Therefore, a correlation does not infer a causation but merely an association. To take other variables into account, a regression analysis is required. Correlation between variables can imply a relationship which can then be tested via regression analysis. High correlation of variables can also make it hard to precisely assess the impact of any specific variable on a given dependent variable, because the behavior of the variable is associated with a specific behavior of the correlated variable. The linear correlation coefficient (called Pearson correlation coefficient) is measured on a scale that varies from + 1 through 0 to – 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. A correlation of absolute value (negative or positive) of more than 0.7 is often considered a strong correlation.

25

Figure 4.14 – Correlation Matrix – Screenshot

The correlations between all numerical variables are shown in a correlation matrix. In this instance, the matrix is very large due to the high number of variables. This represents a drawback of the automated program. The user can obtain the matrix as an image from the results folder (where all graphs are also placed in different color versions including a black and white version) to obtain a different image. He can also open the .csv file with the underlying data for the correlation matrix, which

is also included in the results folder. There, he can decide to exclude certain variables from the matrix and create a new correlation matrix using his own tool/software.



Regression Analysis Report | statistics-hero.com

HIGHLY CORRELATED VARIABLES

We found that in our dataset, 2 pairs of variables are highly correlated.

Among these variables, the variable pair 8 MntDoor_Locks – 14 NumCatalogPurchases shares the highest correlation.

To visualize the high correlations, a pairplot was created. A pairplot visualizes the joint distribution of the correlated variables.

Figure 9

Pairplot of Highly Correlated Variables

CORRELATIONS WITH THE DEPENDENT VARIABLE

To visualize the relationship between the dependent variable and the predictors, a pairplot showing the joint distributions was created.

– Correlations with the dependent variable: In order to obtain a comprehensive picture of the association of the variables of interest in this study, different correlation measures are employed. In addition to the Pearson correlation coefficient r, the Spearman rank correlation coefficient ρ is consulted. It is useful for detecting non-linear associations because it is only concerned with the question if two variables have a monotonic relationship (if one increases, the other increases as well) rather than assessing the precise linear relationship. Therefore, the underlying data must not necessarily be linearly related. The correlations found can be summarized as follows: - The correlation between the variable Income and the dependent variable MntDoor_Locks is described by the Pearson linear correlation coefficient ($r = 0.67$; $p < 0.001$) and the Spearman rank correlation coefficient ($\rho = 0.78$; $p < 0.001$). The implied correlations are different: The linear coefficient r can be

26

Figure 4.15 – High Correlations – Screenshot

The correlations between the most correlated variables of the dataset are visualized to help the user understand the structure of his data from a covariation perspective.

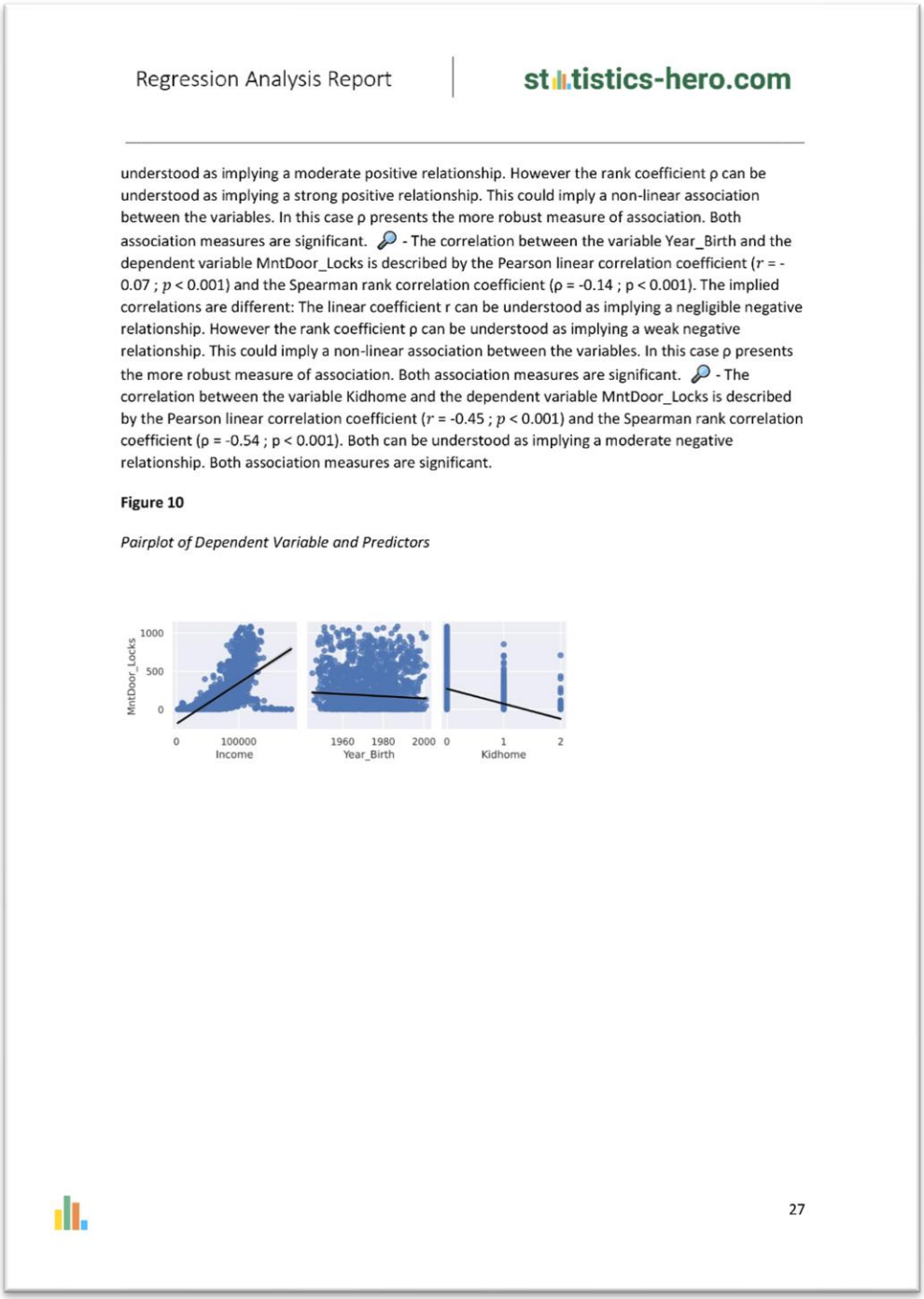
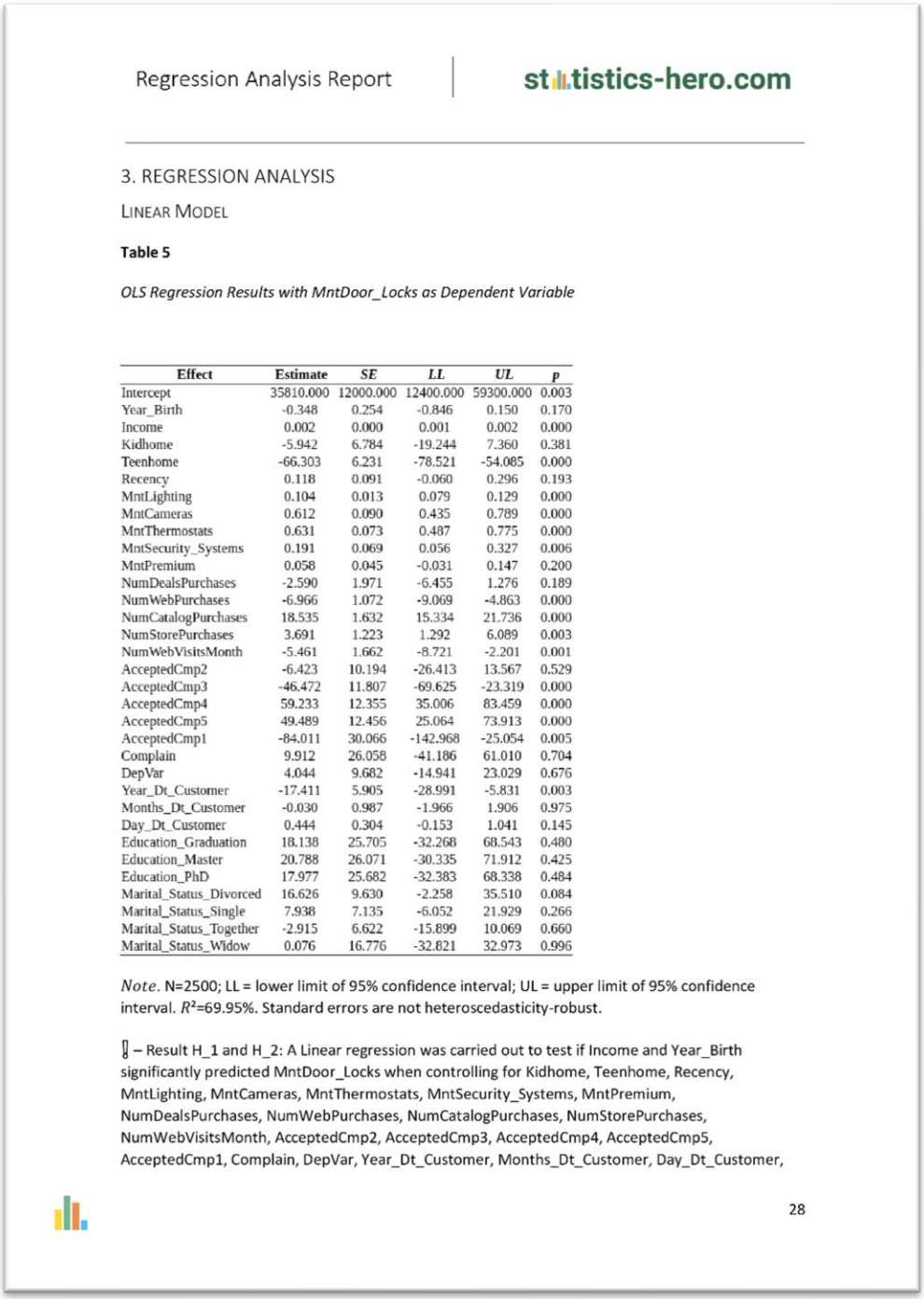


Figure 4.16 – Variable Correlations – Screenshot

The correlation between the dependent variable and the different independent variables and mediator/moderator variables are visualized to help the user detect any patterns that could either imply a non-linear relationship or that do not match the expectations behind the hypotheses formulated by the user.



Regression Analysis Report | statistics-hero.com

3. REGRESSION ANALYSIS

LINEAR MODEL

Table 5

OLS Regression Results with MntDoor_Locks as Dependent Variable

Effect	Estimate	SE	LL	UL	p
Intercept	35810.000	12000.000	12400.000	59300.000	0.003
Year_Birth	-0.348	0.254	-0.846	0.150	0.170
Income	0.002	0.000	0.001	0.002	0.000
Kidhome	-5.942	6.784	-19.244	7.360	0.381
Teenhome	-66.303	6.231	-78.521	-54.085	0.000
Recency	0.118	0.091	-0.060	0.296	0.193
MntLighting	0.104	0.013	0.079	0.129	0.000
MntCameras	0.612	0.090	0.435	0.789	0.000
MntThermostats	0.631	0.073	0.487	0.775	0.000
MntSecurity_Systems	0.191	0.069	0.056	0.327	0.006
MntPremium	0.058	0.045	-0.031	0.147	0.200
NumDealsPurchases	-2.590	1.971	-6.455	1.276	0.189
NumWebPurchases	-6.966	1.072	-9.069	-4.863	0.000
NumCatalogPurchases	18.535	1.632	15.334	21.736	0.000
NumStorePurchases	3.691	1.223	1.292	6.089	0.003
NumWebVisitsMonth	-5.461	1.662	-8.721	-2.201	0.001
AcceptedCmp2	-6.423	10.194	-26.413	13.567	0.529
AcceptedCmp3	-46.472	11.807	-69.625	-23.319	0.000
AcceptedCmp4	59.233	12.355	35.006	83.459	0.000
AcceptedCmp5	49.489	12.456	25.064	73.913	0.000
AcceptedCmp1	-84.011	30.066	-142.968	-25.054	0.005
Complain	9.912	26.058	-41.186	61.010	0.704
DepVar	4.044	9.682	-14.941	23.029	0.676
Year_Dt_Customer	-17.411	5.905	-28.991	-5.831	0.003
Months_Dt_Customer	-0.030	0.987	-1.966	1.906	0.975
Day_Dt_Customer	0.444	0.304	-0.153	1.041	0.145
Education_Graduation	18.138	25.705	-32.268	68.543	0.480
Education_Master	20.788	26.071	-30.335	71.912	0.425
Education_PhD	17.977	25.682	-32.383	68.338	0.484
Marital_Status_Divorced	16.626	9.630	-2.258	35.510	0.084
Marital_Status_Single	7.938	7.135	-6.052	21.929	0.266
Marital_Status_Together	-2.915	6.622	-15.899	10.069	0.660
Marital_Status_Widow	0.076	16.776	-32.821	32.973	0.996

Note. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=69.95%. Standard errors are not heteroscedasticity-robust.

– Result H_1 and H_2: A Linear regression was carried out to test if Income and Year_Birth significantly predicted MntDoor_Locks when controlling for Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer,

28

Figure 4.17 – Linear Model Results I – Screenshot

The linear chapter on the results of the regression analyses starts with the interpretation of the linear model which was estimated to draw conclusions on the linear hypotheses provided by the user. First, the regression table as well as the original hypotheses are shown.

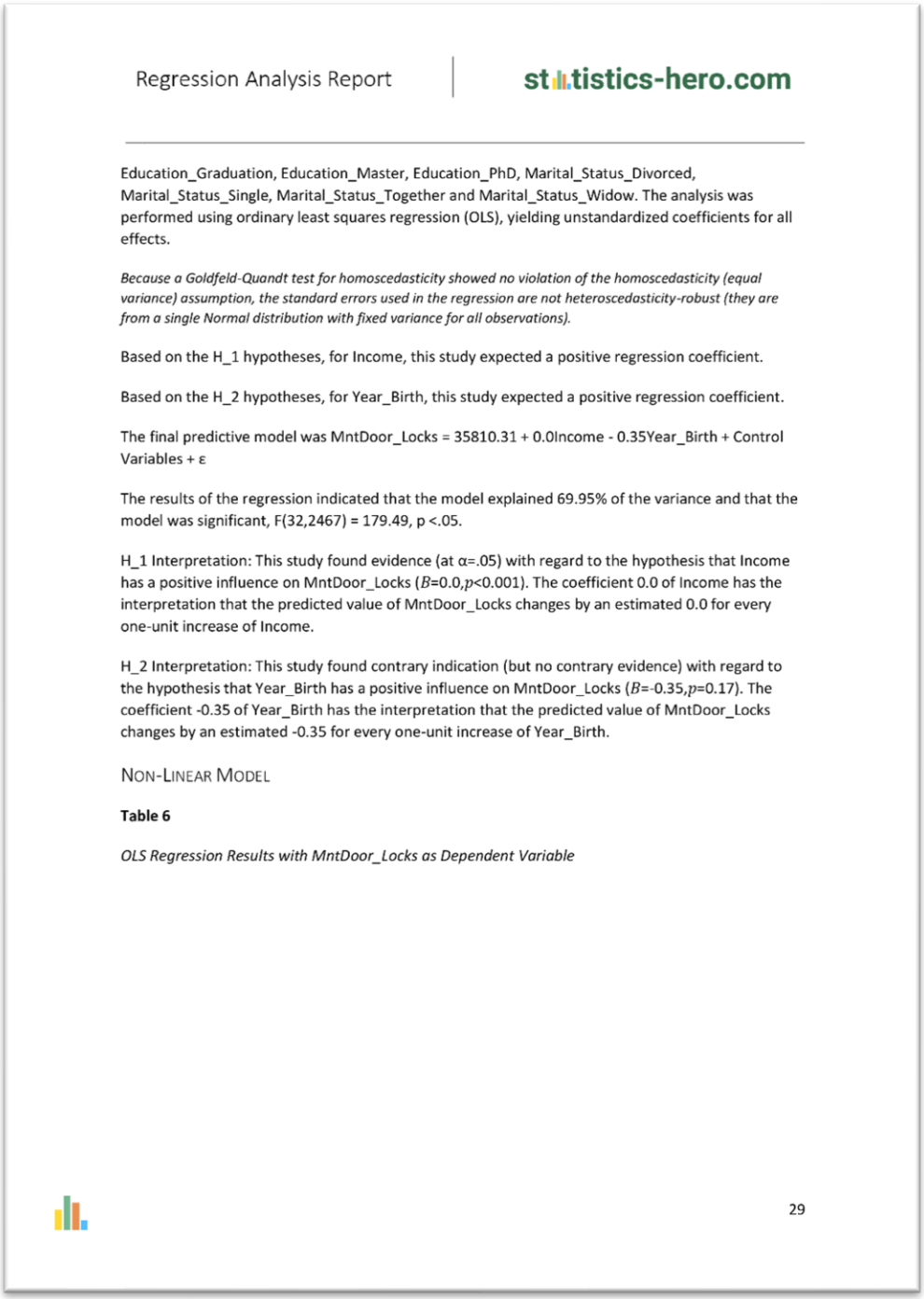
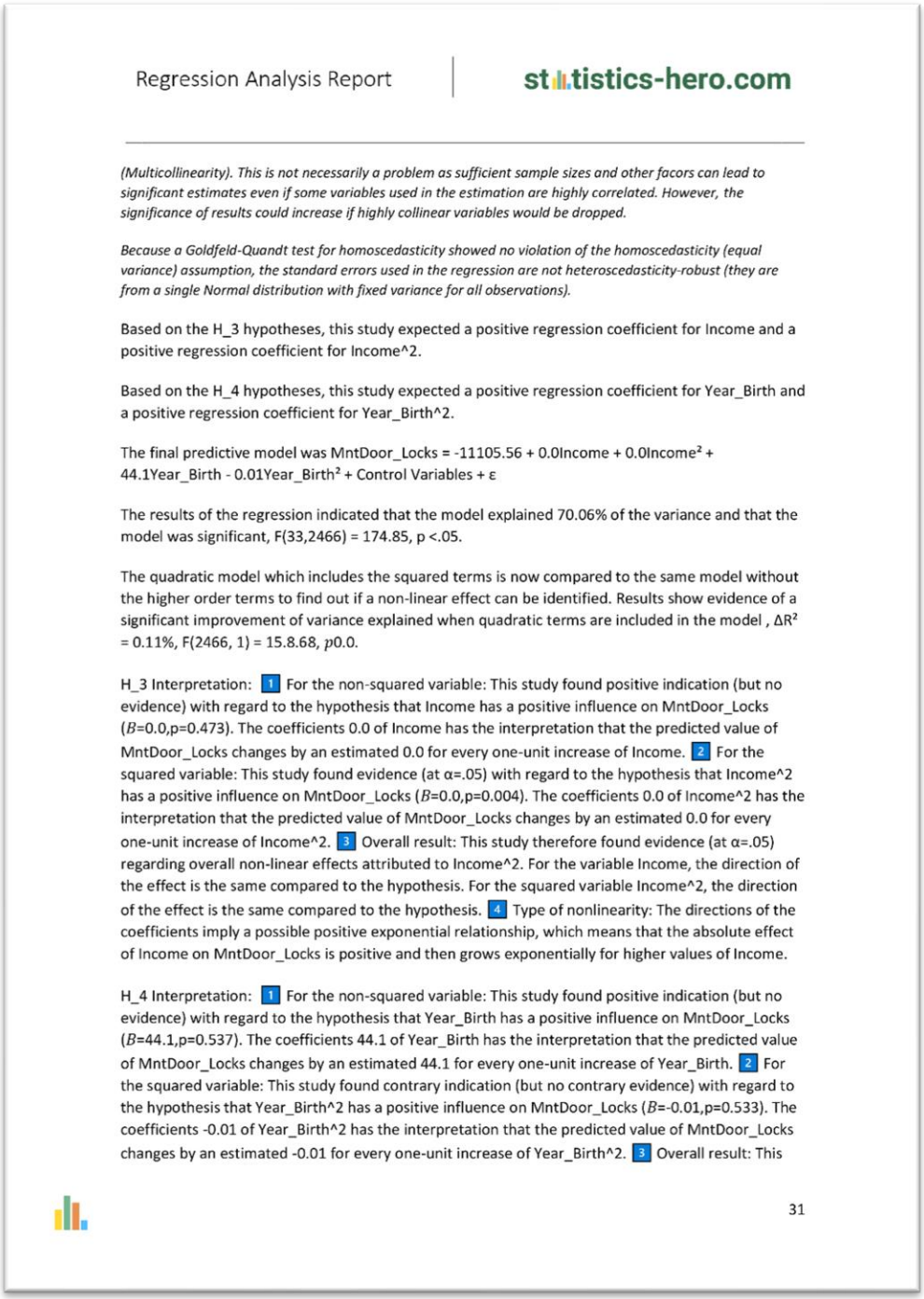


Figure 4.18 – Linear Model Results II – Screenshot

Afterwards, the expected coefficient direction, the estimated model formula as well as the explanatory power of the model and the interpretation of the individual hypotheses are explained.



Regression Analysis Report | statistics-hero.com

(Multicollinearity). This is not necessarily a problem as sufficient sample sizes and other facors can lead to significant estimates even if some variables used in the estimation are highly correlated. However, the significance of results could increase if highly collinear variables would be dropped.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Based on the H_3 hypotheses, this study expected a positive regression coefficient for Income and a positive regression coefficient for Income^2.

Based on the H_4 hypotheses, this study expected a positive regression coefficient for Year_Birth and a positive regression coefficient for Year_Birth^2.

The final predictive model was MntDoor_Locks = -11105.56 + 0.0Income + 0.0Income² + 44.1Year_Birth - 0.01Year_Birth² + Control Variables + ε

The results of the regression indicated that the model explained 70.06% of the variance and that the model was significant, F(33,2466) = 174.85, p <.05.

The quadratic model which includes the squared terms is now compared to the same model without the higher order terms to find out if a non-linear effect can be identified. Results show evidence of a significant improvement of variance explained when quadratic terms are included in the model , ΔR² = 0.11%, F(2466, 1) = 15.8.68, *p*0.0.

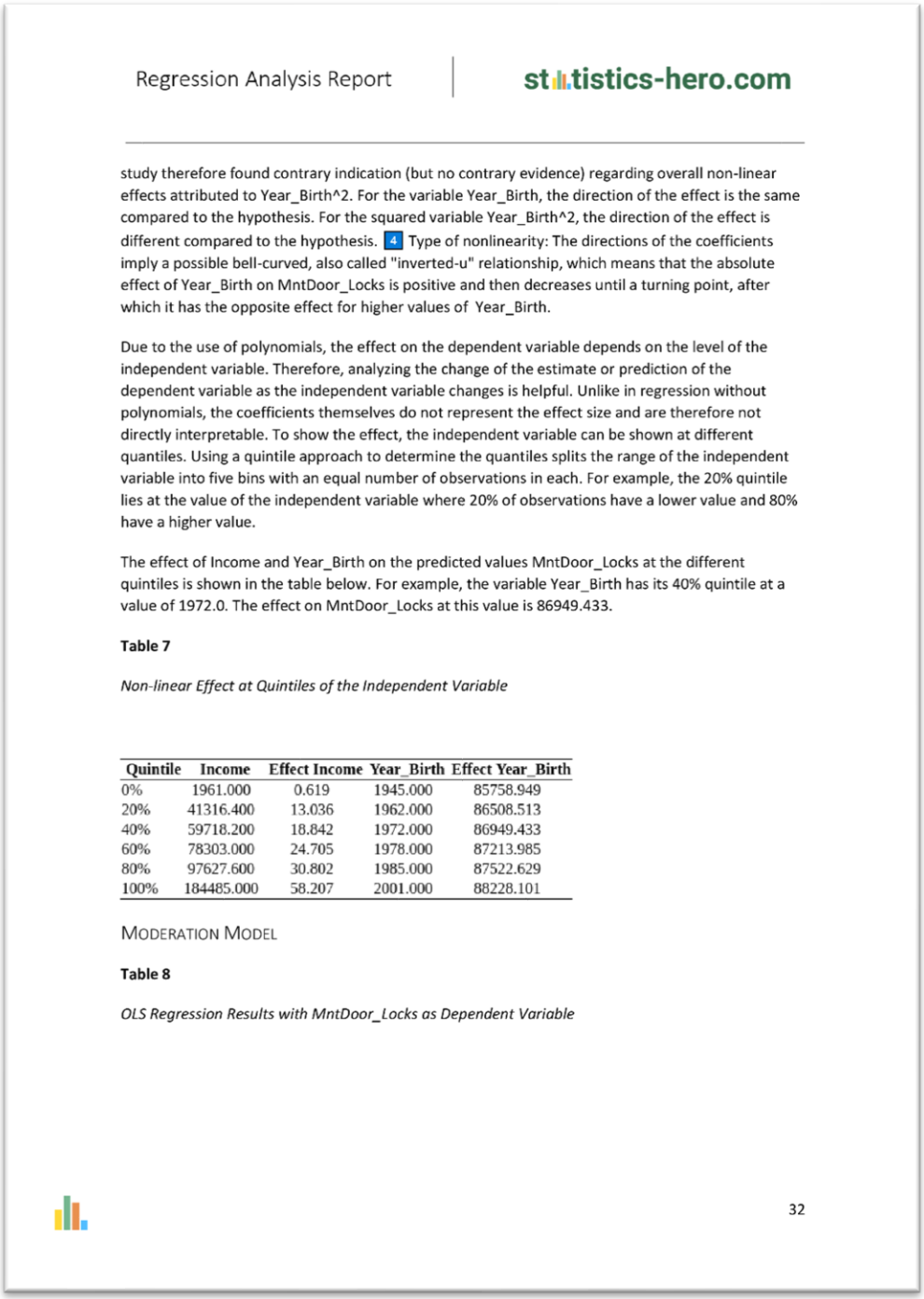
H_3 Interpretation: 1 For the non-squared variable: This study found positive indication (but no evidence) with regard to the hypothesis that Income has a positive influence on MntDoor_Locks (*B*=0.0,p=0.473). The coefficients 0.0 of Income has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 0.0 for every one-unit increase of Income. 2 For the squared variable: This study found evidence (at α=.05) with regard to the hypothesis that Income^2 has a positive influence on MntDoor_Locks (*B*=0.0,p=0.004). The coefficients 0.0 of Income^2 has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 0.0 for every one-unit increase of Income^2. 3 Overall result: This study therefore found evidence (at α=.05) regarding overall non-linear effects attributed to Income^2. For the variable Income, the direction of the effect is the same compared to the hypothesis. For the squared variable Income^2, the direction of the effect is the same compared to the hypothesis. 4 Type of nonlinearity: The directions of the coefficients imply a possible positive exponential relationship, which means that the absolute effect of Income on MntDoor_Locks is positive and then grows exponentially for higher values of Income.

H_4 Interpretation: 1 For the non-squared variable: This study found positive indication (but no evidence) with regard to the hypothesis that Year_Birth has a positive influence on MntDoor_Locks (*B*=44.1,p=0.537). The coefficients 44.1 of Year_Birth has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 44.1 for every one-unit increase of Year_Birth. 2 For the squared variable: This study found contrary indication (but no contrary evidence) with regard to the hypothesis that Year_Birth^2 has a positive influence on MntDoor_Locks (*B*=-0.01,p=0.533). The coefficients -0.01 of Year_Birth^2 has the interpretation that the predicted value of MntDoor_Locks changes by an estimated -0.01 for every one-unit increase of Year_Birth^2. 3 Overall result: This

31

Figure 4.19 – Non-Linear Model Results I – Screenshot

The non-linear model results start similar to the linear model results with the regression table (not shown). However, the interpretation of the individual hypotheses is more complex because different non-linear relationships can be found. Due to the presence of one base and one quadratic term which can each have different directions of the estimated coefficient, the user needs to be explained what shape the curve which resembles the non-linear relationship has. This explanation is provided on hypothesis level and the user is guided by numbered blue boxes to allow easier navigation.



Regression Analysis Report | statistics-hero.com

study therefore found contrary indication (but no contrary evidence) regarding overall non-linear effects attributed to Year_Birth^2. For the variable Year_Birth, the direction of the effect is the same compared to the hypothesis. For the squared variable Year_Birth^2, the direction of the effect is different compared to the hypothesis. 4 Type of nonlinearity: The directions of the coefficients imply a possible bell-curved, also called "inverted-u" relationship, which means that the absolute effect of Year_Birth on MntDoor_Locks is positive and then decreases until a turning point, after which it has the opposite effect for higher values of Year_Birth.

Due to the use of polynomials, the effect on the dependent variable depends on the level of the independent variable. Therefore, analyzing the change of the estimate or prediction of the dependent variable as the independent variable changes is helpful. Unlike in regression without polynomials, the coefficients themselves do not represent the effect size and are therefore not directly interpretable. To show the effect, the independent variable can be shown at different quantiles. Using a quintile approach to determine the quantiles splits the range of the independent variable into five bins with an equal number of observations in each. For example, the 20% quintile lies at the value of the independent variable where 20% of observations have a lower value and 80% have a higher value.

The effect of Income and Year_Birth on the predicted values MntDoor_Locks at the different quintiles is shown in the table below. For example, the variable Year_Birth has its 40% quintile at a value of 1972.0. The effect on MntDoor_Locks at this value is 86949.433.

Table 7

Non-linear Effect at Quintiles of the Independent Variable

Quintile	Income	Effect Income	Year_Birth	Effect Year_Birth
0%	1961.000	0.619	1945.000	85758.949
20%	41316.400	13.036	1962.000	86508.513
40%	59718.200	18.842	1972.000	86949.433
60%	78303.000	24.705	1978.000	87213.985
80%	97627.600	30.802	1985.000	87522.629
100%	184485.000	58.207	2001.000	88228.101

MODERATION MODEL

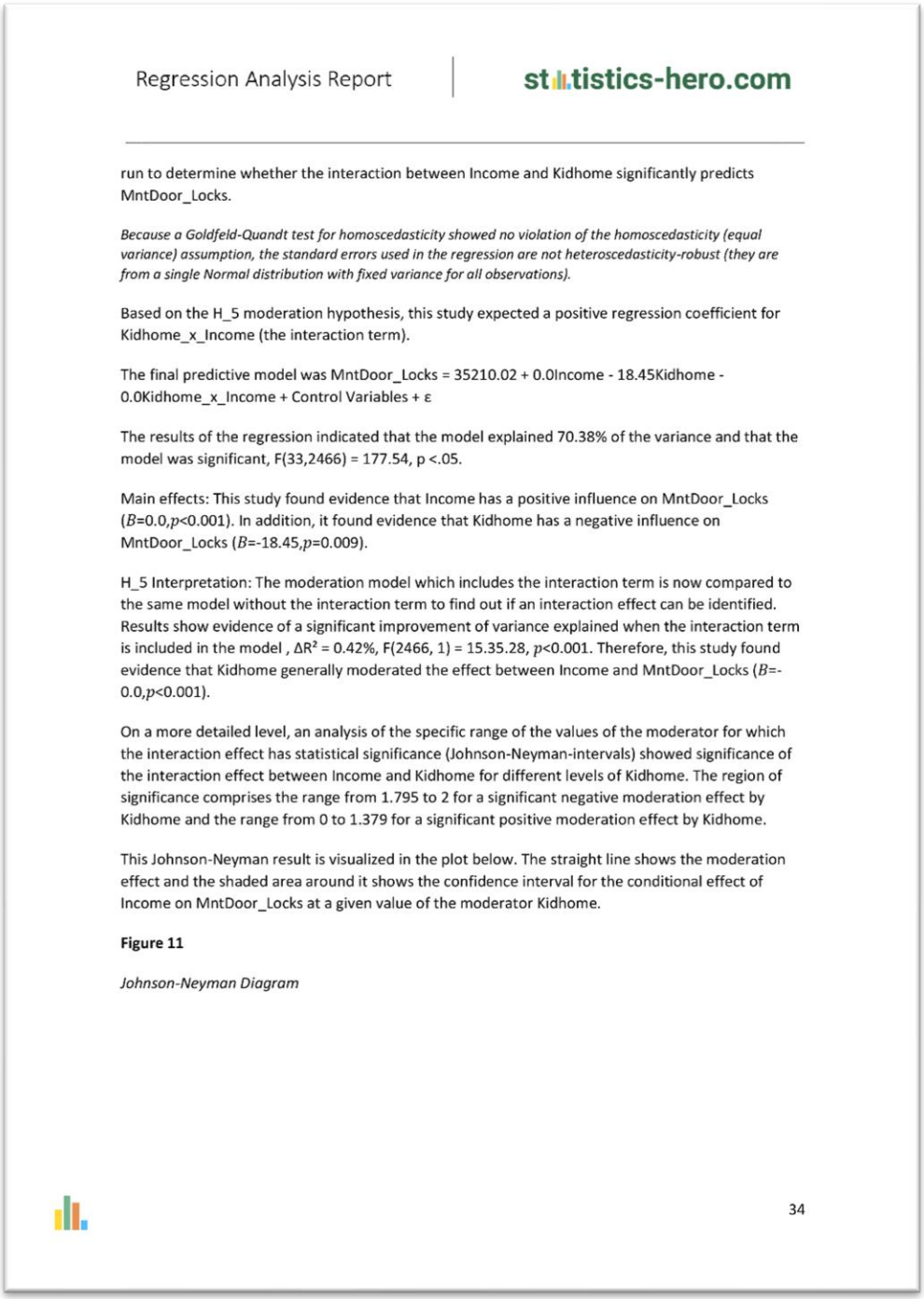
Table 8

OLS Regression Results with MntDoor_Locks as Dependent Variable

32

Figure 4.20 – Non-Linear Model Results II – Screenshot

Because the non-linear effect means that the effect of the independent variable on the dependent variable depends on the level of the independent variable, the effect is shown at each quintile of the independent variable. This is designed at helping the user understand the consequence of the estimated coefficients for the base and quadratic term in the actual estimation.



Regression Analysis Report | statistics-hero.com

run to determine whether the interaction between Income and Kidhome significantly predicts MntDoor_Locks.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Based on the H_5 moderation hypothesis, this study expected a positive regression coefficient for Kidhome_x_Income (the interaction term).

The final predictive model was MntDoor_Locks = 35210.02 + 0.0Income - 18.45Kidhome - 0.0Kidhome_x_Income + Control Variables + ε

The results of the regression indicated that the model explained 70.38% of the variance and that the model was significant, F(33,2466) = 177.54, p <.05.

Main effects: This study found evidence that Income has a positive influence on MntDoor_Locks (*B*=0.0,*p*<0.001). In addition, it found evidence that Kidhome has a negative influence on MntDoor_Locks (*B*=-18.45,*p*=0.009).

H_5 Interpretation: The moderation model which includes the interaction term is now compared to the same model without the interaction term to find out if an interaction effect can be identified. Results show evidence of a significant improvement of variance explained when the interaction term is included in the model , ΔR^2 = 0.42%, F(2466, 1) = 15.35.28, *p*<0.001. Therefore, this study found evidence that Kidhome generally moderated the effect between Income and MntDoor_Locks (*B*=-0.0,*p*<0.001).

On a more detailed level, an analysis of the specific range of the values of the moderator for which the interaction effect has statistical significance (Johnson-Neyman-intervals) showed significance of the interaction effect between Income and Kidhome for different levels of Kidhome. The region of significance comprises the range from 1.795 to 2 for a significant negative moderation effect by Kidhome and the range from 0 to 1.379 for a significant positive moderation effect by Kidhome.

This Johnson-Neyman result is visualized in the plot below. The straight line shows the moderation effect and the shaded area around it shows the confidence interval for the conditional effect of Income on MntDoor_Locks at a given value of the moderator Kidhome.

Figure 11

Johnson-Neyman Diagram

34

Figure 4.21 – Moderation Model Results I – Screenshot

The interpretation of the results of the moderation model follow the same pattern as the structure described before, including the regression table and the explanatory power of the model as well as the conclusion on the initial hypothesis. In addition, the regions of significance of the moderator are detailed.

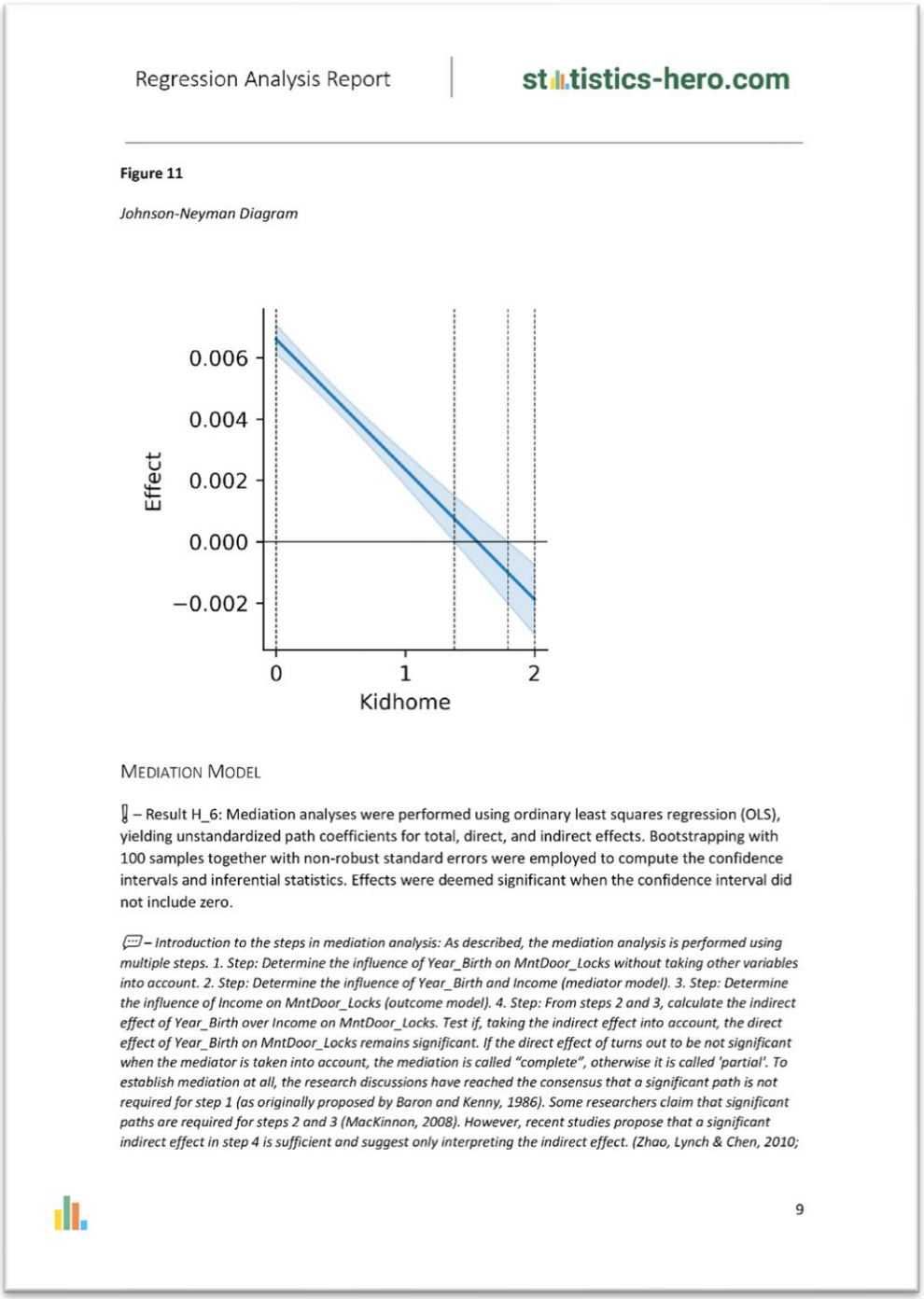
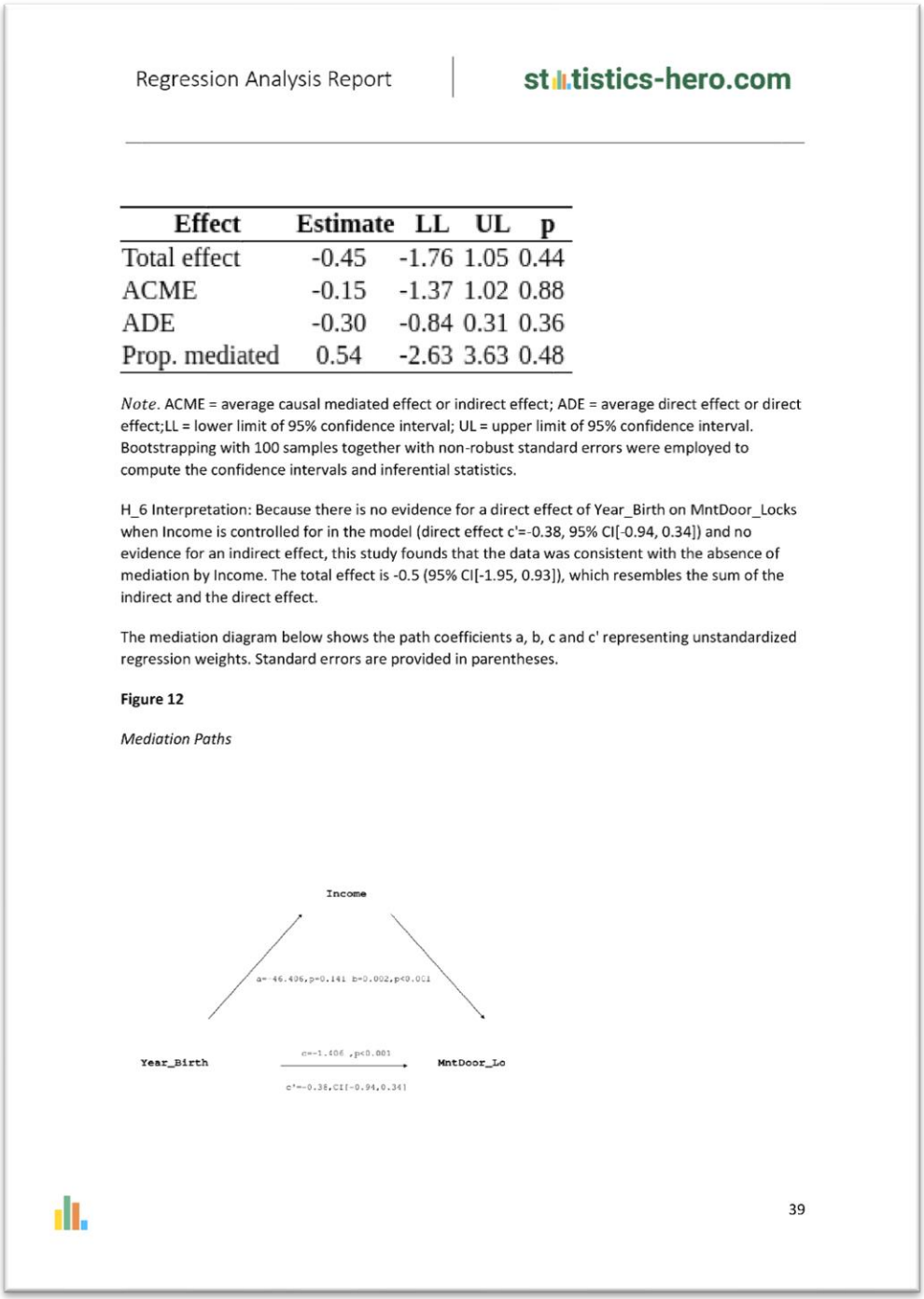


Figure 4.22 – Moderation Model Results II – Screenshot

A Johnson-Neyman plot shows the significance region of the effect across the levels of the moderator variable. Currently, the significance regions are only presented as separate bars with confidence intervals in case of two values of the moderator and as continuous plots in case of more than two values of the moderator.



Regression Analysis Report | statistics-hero.com

Effect	Estimate	LL	UL	p
Total effect	-0.45	-1.76	1.05	0.44
ACME	-0.15	-1.37	1.02	0.88
ADE	-0.30	-0.84	0.31	0.36
Prop. mediated	0.54	-2.63	3.63	0.48

Note. ACME = average causal mediated effect or indirect effect; ADE = average direct effect or direct effect;LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. Bootstrapping with 100 samples together with non-robust standard errors were employed to compute the confidence intervals and inferential statistics.

H_6 Interpretation: Because there is no evidence for a direct effect of Year_Birth on MntDoor_Locks when Income is controlled for in the model (direct effect c'=-0.38, 95% CI[-0.94, 0.34]) and no evidence for an indirect effect, this study founds that the data was consistent with the absence of mediation by Income. The total effect is -0.5 (95% CI[-1.95, 0.93]), which resembles the sum of the indirect and the direct effect.

The mediation diagram below shows the path coefficients a, b, c and c' representing unstandardized regression weights. Standard errors are provided in parentheses.

Figure 12

Mediation Paths

39

Figure 4.23 – Mediation Model Results I – Screenshot

The interpretation of the results of the mediation analysis starts with three regression analyses for the three mediation paths (independent variable to dependent variable, independent variable to mediator and mediator to dependent variable). Each step is shown with a regression table and the estimated path (not shown). Afterwards, the indirect effect is investigated and a mediation diagram is created which show the coefficients of each path together with the significance.

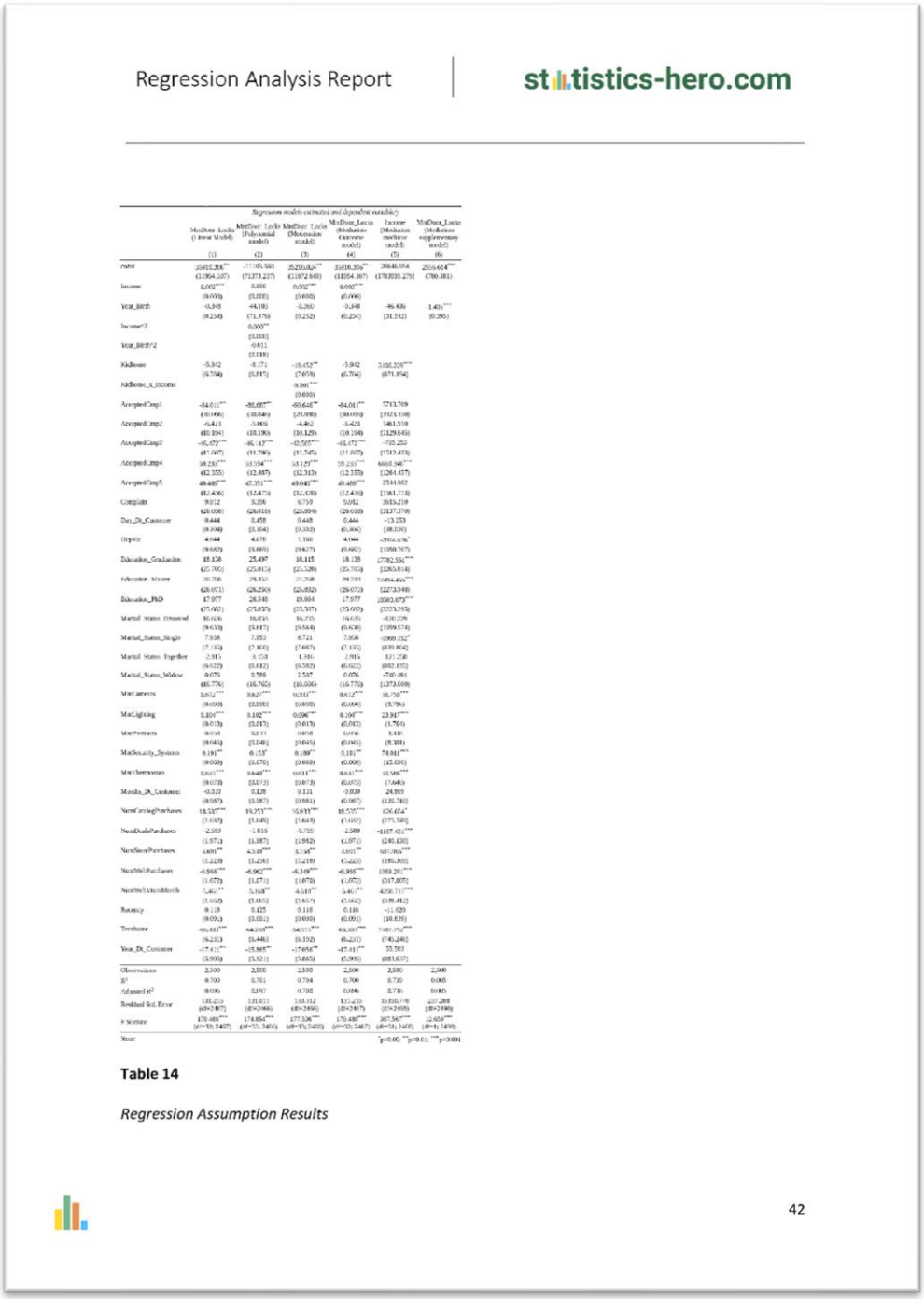


Figure 4.24 – Appendix: All Regressions Table – Screenshot

The appendix contains other visuals and tables, including boxplots and a summary table for the regression assumptions checks (not shown). In addition, a regression table for all regression models performed is included.

4.2. TESTING

The evaluation of the software proved that the presented prototype works reliably. As mentioned before, the prototype was developed and tested on the datasets as described in section 3.

The data sets were selected for their comprehensiveness which allowed testing most facets of the prototype on any datasets. For example, five of the datasets had categorical variables which require separate visuals and interpretations. In addition, the data sets allowed for including binary data in the analysis and for testing missing values and outlier treatments. The real-life survey data sets helped to work with “messy” data that included numerous items to be combined into single variables and unnecessary metadata such as IP address. In addition, the survey datasets had non-alphanumeric characters in the column names which helped make the preprocessing pipeline robust by enforcing alphanumeric column names for all variables. One data set had insufficient observations because to conduct the rainbow test which is performed to test for linearity, the data is split into two chunks and both chunks need to have more observations than columns to facilitate OLS regression. This helped make the preprocessing pipeline robust by enforcing feature selection in these cases (reduction to a number that allows OLS computation).

After these corrections, over 100 combinations of different hypotheses and dependent and independent variables were performed. Reports were successfully created for all combinations. At random, 10 runs of the prototype were compared to the corresponding SPSS output based on the same preprocessed data set. The outputs matched apart from minor deviations for variables with very high standard deviations and p-values over 95%, which are likely due to differences in OLS implementation between the Python Statsmodels module and SPSS.

5. CONCLUSIONS

The contribution offered by the prototype developed (and to a greater extend offered by a prospective comprehensive program built on its basis) concerns academia as well as business. In the area of empirical research, the prototype can help researchers accelerate the data preprocessing process using the comprehensive preprocessing options provided by the program. In addition, they can quickly investigate relationships between variables and obtain a comprehensive report detailing the results. In the area of business, many employees in fields such professional services (especially management consulting) face the challenge of having to quickly analyze data without sufficient education on data processing and analysis methods or access to data professionals. These employees can benefit from the prototype as it enables them to perform adequate preprocessing and produces a clearly explained output. For example, a sales manager could be tasked with an analysis on the influence of context information on customers on the expected purchase volume. Using the regression analysis offered by the prototype, the sales manager could quickly test his hypotheses on potential predictors of purchase volume and use the prototype output to prepare a report on his findings.

A large share of the process on automated data analysis in the recent years has focused on high tech applications such as complex machine learning models. Every-day analysis tasks such as simple regressions are still widely performed manually. This study has shown that untapped automation potential lies in explicitly addressing such tasks for the large target group of technically unskilled and unexperienced users. Besides making statistical analyses easier, the prototype also helps users understand and interpret the statistical issues and decisions because every output in the report is explained and additional background information is provided where necessary. This serves an educational purpose. Rather than blindly following a manual for statistics software and neglecting foundational parts of the analysis workflow (such as assumption checks, see Hanif & Aymal, 2011), the user is exposed to the statistical facets of the analysis and gains an enhanced understanding of statistical concepts due to the direct application to his data.

The study also found that data-centric design (Campos et al., 2009) provides a sound framework for designing such tools. The prototype developed in this study mirrors the elements included in the data-centric design frameworks and adopts the same easy-to-understand user-facing language, such as DISCOVER for data exploration and EXPLAIN for hypothesis tests and regression analyses. This helps unskilled users in using their intuition while preparing and conducting data analyses, making the process more accessible. However, the data-centric approach and abstracting from the models being built have the drawback that potential modeling issues such as the violation of statistical assumptions behind hypothesis tests are neglected. Therefore, incorporating automated checks is important. The core challenge is to create a robust analysis pipeline based on comparatively few user inputs because a higher number of inputs would increase the perceived complexity of the prototype and deter users without a quantitative background from using it.

Combined with advanced Natural Language Processing (NLP) technology that can identify research topics and aid with literature review and decision-making, automated data analysis could be embedded into a comprehensive approach to automated research which selects the right hypotheses and performs all statistical tests. The role of the human researcher would then be reduced, or at least more focused on areas where humans can add significant value, such as the evaluation and conclusion phase of research. Where this development leads remains to be seen.

The user input required for the prototype is currently provided in a dictionary format, where every preprocessing option can be specified by the user and, otherwise, a default method is used. This input can easily be gathered via a web form. Here, the user would first upload the data set and then provide the required preprocessing information in a sequence of form pages. In between the form pages, the prototype recalculates the preprocessed data or only the list of column names, depending on if the next step requires the updated data or only the updated column names. For example, when the user renames columns, the web form only collects the information and appends it to a dictionary with all entered preprocessing settings, which is delivered to the preprocessing module in the end. On the website, only the list of column names is changed and not the actual data set in order to accelerate the user input process. This means that the user can choose options based on the new column names on subsequent forms without the data set being updated at every step.

6. LIMITATIONS AND RECOMMENDATIONS

The Master project is only designed to be a prototype for a full automation program. Several opportunities for extending the functionalities of the prototype remain. These opportunities can be pursued in other research projects or commercial applications.

Firstly, more algorithms could be included in the model, especially for the regression component where currently only linear and regression with OLS is applied. The final program should incorporate other regression algorithms including logistic regression for binary dependent variables. In addition, more advanced and powerful regression methods such as decision trees or random forests could be employed to make advanced machine learning algorithms more accessible to unskilled users. A large variety of algorithms can be deployed and interpreted using the same structure and logic applied to this prototype. However, because the prototype produces a verbally expressed interpretation of the findings and the algorithms differ vastly in the parameters, coefficients and interpretations they take in or return, the structure of the prototype does not scale ideally towards the addition of additional algorithms. Whereas the between-scripts logic and the logic behind creating the word report can be used for applying other algorithms, the individual variables created and text sections printed will need to be adjusted to the specific algorithm used.

Secondly, the prototype aims at high automation and features many default values designed at making decisions for the user. While the preprocessing is flexible enough to cater to a variety of data sets, components such as the critical values for statistical tests are fixed and cannot be changed by the user in the current structure. This is partly due to the target to simplify the process and make decisions for the user where they can reasonably be made. Another approach would be a more guided approach where input from the user is repeatedly obtained throughout the process. This would allow for more complex analyses, where the direction of analysis has to be changed if new information is obtained. This approach would introduce more complexity, but also potentially yield more value because the specific domain expertise of the user and the broad range of potential analysis sequences could be included in the program. In addition, the approach would add educational value because the impact of different user decisions could be immediately addressed, leading to enhanced statistical understanding.

All in all, increasing the scope of analyses and enhancing the guided elements of the process could extend the scope of the project and result in a full-service automated data analysis solution for almost all circumstances where data analysis is required. The potential in this is very big and the value add for the end user is significant. This study expects that more (and more complex) automated data analysis programs will be introduced and successfully utilized in the future.

7. BIBLIOGRAPHY

Abdelfattah, E. (2020). Comment on article “What needs improvement with IBM SPSS Statistics?”. Retrieved from https://www.itcentralstation.com/questions/what-needs-improvement-with-ibm-spss-statistics

Ankerst, M. (2003). The perfect data mining tool: Interactive or automation – Report on the SIGKDD-2002 Panel. SIGKDD Explorations, 5, 110-111.

Andre, Q. (2019). PyProcessMacro, source code available at https://github.com/QuentinAndre/pyprocessmacro

Asghar, S. & Iqbal, K. (2009, April). Automated Data Mining Techniques: A Critical Literature Review. In *ICIME '09: Proceedings of the 2009 International Conference on Information Management and Engineering*. 2009 International Conference on Information Management and Engineering, Pages 75–79. IEEE Computer Society.

Avidon, E. (2019). IBM battling to change perception of Cognos Analytics BI platform. Retrieved from https://searchbusinessanalytics.techtarget.com/news/252466628/IBM-battling-to-keep-Cognos-Analytics-BI-platform-relevant

Barker, L. & Shaw, K. M. (2015). Best (but often forgotten) practices: checking assumptions concerning regression residuals. The American Journal of Clinical Nutrition, 102(3), 533-539. doi: 10.3945/ajcn.115.113498

Burgold, A. (2021). *The Impact of Political Statements as Advertising Messages on the Consumer Behaviour* [Unpublished Master dissertation]. University of Cologne

Bloom, B., Bhagwat, C., & Stengard, P. (2003). Automated model building and evaluation for data mining system. US Patent App. 10/383,641.

Canny, S. (2013). Python-docx. Retrieved from https://github.com/python-openxml/python-docx

Chapman, S., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R (2000): *CRISP-DM 1.0, step-by-step data mining guide*, SPSS Inc. Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Choueiry, George & Salameh, Pascale. (2019). Automating Data Analysis Methods in Epidemiology. Journal of data science: JDS. 17. 55-80. 10.6339/JDS.201901_17(1).0003.

Campos, M., Milenova, B., & Stengard, P. (2009). Data-centric automated data mining. US Patent 7,627,620. Retrieved from https://www.oracle.com/technetwork/testcontent/automated-data-mining-paper-1205-128874.pdf

Ciechanowski, L., Jemielniak, D. & Gloor, P. (2020). TUTORIAL: AI research without coding: The art of fighting without fighting: Data science for qualitative researchers. Journal of Business Research, 117, retrieved from: https://www.sciencedirect.com/science/article/pii/S0148296320303854

Deisenroth, M. P., Faisal, A.A., Ong, C.S. (2020). Mathematics for Machine Learning. Cambridge University Press.

DeVoss, C. (2017). Artificial intelligence applications in scientific publishing. Retrieved from: https://blogs.biomedcentral.com/bmcblog/2017/05/03/artificial-intelligence-applications-in-scientific-publishing/

Doornik, J. A., 'Autometrics', in Jennifer Castle, and Neil Shephard (eds), The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry (Oxford, 2009; online edn, Oxford Academic, 1 Sept. 2009), https://doi.org/10.1093/acprof:oso/9780199237197.003.0004, accessed 24 Aug. 2022

Doyle, M., & Sambanis, N. (2000). International peacebuilding: A theoretical and quantitative analysis. *American Political Science Review* (94) pp. 779-801.

Dwan, K., Gamble, C., Williamson, P. R. & Kirkham, J. J.(2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLOS ONE, 8(7), e66844. doi: 10.1371/journal.pone.0066844

Engber, D. (2017). Humans Run Experiments, a Robot Writes the Paper. Slate Magazine. Retrieved from https://slate.com/technology/2017/12/science-papers-should-be-written-by-robots.html

Fabra, U.P., & Schmidheiny, K. (2010). The Multipe Linear Regression Model. In *Short Guides to Microeconometrics Fall 2010*. Retrieved from: https://www.schmidheiny.name/teaching/ols2up.pdf

Field, A. (2013). *SPSS Automatic Linear Modeling*. Retrieved from http://www.youtube.com/watch?v=pltr74IlxOg

Futurism (2020). This Grad Student Used a Neural Network to Write His Papers. Futurism Science and Technology News. Retrived from: https://futurism.com/grad-student-neural-network-write-papers

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publisher.

Hanif, A., & Ajmal, T. (2011). Statistical Errors in Medical Journals (A Critical Appraisal). Annals of King Edward Medical University, 17(2), 178-178. doi: 10.21649/akemu.v17i2.295

Harrison, D. & Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air, J. *Environ. Economics & Management (5)*, pp. 81-102

Hastie, T., Tibshirani, R., & Friedman, J. (2011). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York: Springer.

Hayes, A. F. (2021). Available at http://processmacro.org/index.html

Hayes, A.F. (2013). Mediation, Moderation, and Conditional Process Analysis. New York: Guilford Press.

Hayes, A.F. (2017). Hacking PROCESS for Estimation and Probing of Linear Moderation of Quadratic Effects and Quadratic Moderation of Linear Effects. Available at http://afhayes.com/public/quadratichack.pdf

Heller, P. (2021). Warum Maschinen besser abschreiben [Why machines cheat better]. Frankfurter Allgemeine Zeitung (FAZ). Retrieved from: https://www.faz.net/aktuell/wissen/computer-mathematik/ki-und-plagiate-warum-maschinen-besser-abschreiben-17461712.html

Hemmerich, W. (2015). StatistikGuru: SPSS Anleitungen [SPSS manuals]. Retrieved from https://statistikguru.de/spss

Henriques, R. (2021). [Smart Home Dataset] [Unpublished raw data from lecture "Predictive Methods of Data Mining"]. NOVA IMS – Information Management School

Hevner, A., March & S., Park, J. (2004). Design Science in Information Systems Research. *MIS Quarterly 28 (1)*, pp. 75-105.

Hutter, F., Kotthoff, L. & Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. Springer.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124

JetBrains, 2017. PyCharm. Available at: https://www.jetbrains.com/pycharm

Kaggle, 2022. Song Popularity Dataset. Retrieved from https://www.kaggle.com/datasets/yasserh/song-popularity-dataset

Kamarudin, N., Ismail, S. (2017). Manual and Automated Model Selection Procedures for Seemingly Unrelated Regression Equations with Different Estimation Methods. *Far East Journal of Mathematical Sciences 101(8),* pp. 1655-1670

Mamaev, A. (2019). How to build AutoML from scratch. Retrieved from https://alxmamaev.medium.com/how-to-build-automl-from-scratch-ce45a4b51e0f

Malato, G. (2020). How to build your own AutoML library in Python from scratch. Retrieved from https://towardsdatascience.com/how-to-build-your-own-automl-library-in-python-from-scratch-995940f3fa71

Masuadi, E., Mohamud. M., Almutairi, M., Alsunaidi, A., Alswayed. A., Aldhafeeri, O. (2021). Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis. Cureus. 2021 Jan 11;13(1):e12639. doi: 10.7759/cureus.12639. PMID: 33585125; PMCID: PMC7872865.

Mikko (2019). Automated Research and Beyond: The Evolution of Artificial Intelligence. Towardsdatascience. Retrieved from: https://towardsdatascience.com/automated-researcher-and-beyond-the-evolution-of-artificial-intelligence-5db4fdde6f1c

Miller, R. (2014). Alteryx Lands $60M To Boost Data Analytics App-Building Platform. Retrieved from: https://techcrunch.com/2014/10/06/alteryx-lands-60m-to-continue-building-data-analtyics-platform/

Nichols, G. (2021). AI can write a passing college paper in 20 minutes. ZDNet. Retrieved from: https://www.zdnet.com/article/ai-can-write-a-passing-college-paper-in-20-minutes/

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830

Peffers, K., Tuunanen, T., Rothenberger, M. & Chatterjee, S. (2007). A Design Science Methodology for Information Systems Research. *Journal of Management Information Systems 24(3)*, pp. 45-78.

Pells, R. (2017). Rise of the research-bots: AI software that writes your papers for you. Times Higher Education. Retrieved from: https://www.timeshighereducation.com/news/rise-research-bots-ai-software-writes-your-papers-you

Kluyver, T., Ragan-Kelley, B., Fernando Perez, Granger, B., Bussonnier, M., Frederic, J., Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90).

Leal, J. (2015). *Automated Data Mining and Automatic Visualization* (Master Thesis). Retrieved from https://eg.uc.pt/bitstream/10316/35524/1/Automated%20Data%20Mining%20and%20Automatic%20Visualization.pdf

Lindeløv, J. (2019). SPSS is dying. It's time to change. Blog entry, retrieved from https://lindeloev.net/spss-is-dying/

Rubenking, N. J. (2001). Hidden messages. Retrieved from http://www.pcmag.com/article2/0,2817,8637,00.asp

SAS Institute, Inc. (2011a). Getting started with SAS Enterprise Miner 7.1. Cary, NC

SAS Institute, Inc. (2011b). SAS Enterprise Miner 7.1 extension nodes: Developer's guide. Cary, NC

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Shearer C. (2000). The CRISP-DM model: the new blueprint for data mining, J Data Warehousing; 5:13—22.

StackExchange (2011). Is automatic linear modeling in SPSS a good or bad thing? Retrieved from http://stats.stackexchange.com/questions/7432/is-automatic-linear-modelling-in-spss-a-good-or-bad-thing

Tatalovic, M. AI writing bots are about to revolutionise science journalism: we must shape how this is done. Journal of Science Communication, 17, retrieved from https://jcom.sissa.it/archive/17/01/JCOM_1701_2018_E

Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing. Frontiers in artificial intelligence, 4, 576892. https://doi.org/10.3389/frai.2021.576892

Uedufy (August 26, 2022) How To Run Mediation Analysis in SPSS [2 Methods]. Retrieved from https://uedufy.com/how-to-run-mediation-analysis-in-spss/.

Van Noorden, R. (2014). Publishers withdraw more than 120 gibberish papers. Nature. Retrieved from: https://www.nature.com/articles/nature.2014.14763

Van Rinsum, H. (2021). Iris.ai: Wie Künstliche Intelligenz bei der Recherche hilft. Retrieved from: https://ki-marketing.com/iris-ai-wie-kuenstliche-intelligenz-bei-der-recherche-hilft/

Waring J, Lindvall C & Umeton R (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, 104, retrieved from https://www.sciencedirect.com/science/article/pii/S0933365719310437

Wheeler, A.P. (2014). Automating tasks in SPSS using production jobs. Retrieved from https://andrewpwheeler.com/2014/12/03/automating-tasks-in-spss-using-production-jobs/

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Elsevier Inc.

Woolridge, J.M. (2012). Introductory Econometrics : A Modern Approach. Mason, Ohio: South-Western Cengage Learning

Yang, H. (2013). The case for being automatic: Introducing the Automatic Linear Modeling (LINEAR) procedure in SPSS Statistics. *Multiple Linear Regression Viewpoints, 39(2)*, 27-37.

Zou, D., Lloyd, J. E. V., & Baumbusch, J. L. (2019). Using SPSS to analyze complex survey data: A primer. Journal of Modern Applied Statistical Methods, 18(1), eP3253. doi: 10.22237/jmasm/1556670300

8. APPENDIX

APPENDIX 1: SMART HOME DATASET (HENRIQUES, 2021)

Excerpt: 5 of 2,500 observations, ID column added for clarity

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer
1	1957	PhD	Divorced	22386	1	1	10.31.2018
2	1982	PhD	Together	68602	0	1	3.14.2019
3	1993	Graduation	Married	23846	1	0	12.18.2017
4	1981	Graduation	Together	40944	1	0	5.22.2019
5	1996	Graduation	Together	104758	0	0	2.20.2019

ID	Recency	MntLighting	MntCameras	MntDoor_Locks	MntThermostats	MntSecurity_Systems	MntPremium
1	65	32	1	29	2	0	20
2	44	41	0	5	0	0	2
3	53	4	10	4	2	0	2
4	67	18	6	26	5	2	8
5	14	670	79	685	191	210	114

ID	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp2	
1	5	9	3	3	9	1	
2	1	7	2	3	3	0	
3	1	6	2	3	7	0	
4	2	8	2	4	6	0	
5	1	10	8	4	1	0	

ID	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	Complain	DepVar	
1	0	0	0	0	0	0	
2	0	0	0	0	0	0	
3	0	0	0	0	0	0	
4	0	0	0	0	0	0	
5	0	1	0	0	0	0	

APPENDIX 2: SMART HOME DATASET VARIABLES (HENRIQUES, 2021)

Variable	Description
AcceptedCmp1	Flag indicating customer accepted offer in campaign 1
AcceptedCmp2	Flag indicating customer accepted offer in campaign 2
AcceptedCmp3	Flag indicating customer accepted offer in campaign 3
AcceptedCmp4	Flag indicating customer accepted offer in campaign 4
AcceptedCmp5	Flag indicating customer accepted offer in campaign 5
Complain	Flag indicating if customer has complained (last 18 months)
Dep Var	Target. Binary variable indicating if the Client accepted (1) or rejected (0) the marketing offer offer
Custid	Customer ID
Dt_Customer	Date of customer's enrolment with the company
Education	Level of education of Customer
Income	Yearly Income of household of Customer
Kidhome	Number of kids in household
Marital_Status	Marital Status of Customer
MntLighting	Amount spent on Lighting_related products (last 18 months)
MntDoor_Locks	Amount spent on Door_Locks and related products (last 18 months)
MntCameras	Amount spent on Cameras and related products (last 18 months)
MntThermostats	Amount spent on Thermostats and related products (last 18 months)
MntSecurity_Systems	Amount spent on Security_Systems and related products (last 18 months)
MntPremiumProds	Amount spent on premium products (last 18 months)
NumCatalogPurchases	Number of purchases made through catalog
NumDealsPurchases	Number of purchases made with discounts
NumStorePurchases	Number of purchases made through store
NumWebPurchases	Number of purchases made through web
NumWebVisitsMonth	Number of web visits a month to companies site
Recency	Days since last purchase
Teenhome	Number of teenagers in household
Year_Birth	Customer's Year of birth
Z_CostContact	Campaign's Cost per Contact
Z_Revenue	Campaign's positive answer revenue
ElementXX	Group Element. Reject in alaysis
Group	Group ID. Reject in analysis

APPENDIX 3: FULL AUTOMATED SOFTWARE OUTPUT ON EXAMPLE DATASET (REPORT)

REGRESSION ANALYSIS

CHAPTER OVERVIEW

I. Hypotheses – 💡 Listing the potential relationships between variables that this study will test

II. Methodology

1. Model – 💡 Defining the regression models that are used to test the hypotheses

2. Data Preparation – 💡 Documenting the preprocessing performed on the source data

3. Regression Assumptions – 💡 Checking if all statistical model assumptions are fulfilled

III. Results

1. Descriptive statistics – 💡 Describing the data that is used for estimating the regression models

2. Correlation analysis – 💡 Investigating pairwise associations between variables in the data

3. Regression analysis – 💡 Find out if the hypotheses formulated in the beginning hold

EMOJI LEGEND

❗ means: Core section (often important steps or results of the analysis)

🗨 or paragraph in italics means: Informative Section (details on the statistical methods used)

SUMMARY OF FINDINGS

Data

This study testes 6 hypotheses using linear regression. The data used for the analysis includes information on 33 variables. Information is provided for 2500 observations.

Linear Model

This study found evidence (at α=.05) with regard to the hypothesis that Income has a positive influence on MntDoor_Locks (B=0.0,p<0.001).

This study found contrary indication (but no contrary evidence) with regard to the hypothesis that Year_Birth has a positive influence on MntDoor_Locks (B=-0.35,p=0.17).

Non-Linear Model

This study found evidence (at α=.05) regarding overall non-linear effects on MntDoor_Locks attributed to Income^2. For the variable Income, the direction of the effect is the same compared to the hypothesis (B= 0.0,p=0.473). For the squared variable Income^2, the direction of the effect is the same compared to the hypothesis (B= 0.0,p=0.004).

This study found contrary indication (but no contrary evidence) regarding overall non-linear effects on MntDoor_Locks attributed to Year_Birth^2. For the variable Year_Birth, the direction of the effect is the same compared to the hypothesis (B= 44.1,p=0.537). For the squared variable Year_Birth^2, the direction of the effect is different compared to the hypothesis (B= -0.01,p=0.533).

Moderation Model

This study found evidence that Kidhome generally moderated the effect between Income and MntDoor_Locks, ΔR^2 = 0.42%, F(2466, 1) = 15.35.28, p<0.001.

Mediation Model

This study found that there is no evidence for a direct effect of Year_Birth on MntDoor_Locks when Income is controlled for in the model (direct effect c' = -0.38, 95% CI[-0.94, 0.34]). In addition, there is no evidence for an indirect effect of Year_Birth over Income on MntDoor_Locks, indirect effect ab = -0.12, 95% CI[-1.85, 1.17]. Therefore, this study found that the data is consistent with the absence of mediation by Income.

Detailed regression results are presented in the regression analysis chapter of the results section.

I. HYPOTHESES

This study tested the following 6 hypotheses.

Linear hypotheses:

– Hypothesis (H_1): Income has a positive effect on MntDoor_Locks. Note: This effect is linear, which means that the absolute effect on MntDoor_Locks does not depend on the level of Income.

– Hypothesis (H_2): Year_Birth has a positive effect on MntDoor_Locks. Note: This effect is linear, which means that the absolute effect on MntDoor_Locks does not depend on the level of Year_Birth.

Non-linear hypotheses:

– Hypothesis (H_3): Income and Income^2 have a positive (Income) and a positive (Income^2) effect on MntDoor_Locks. The effect is characterized by a positive exponential relationship, which means that the absolute effect of Income on MntDoor_Locks is positive and then grows exponentially for higher values of Income.

– Hypothesis (H_4): Year_Birth and Year_Birth^2 have a positive (Year_Birth) and a positive (Year_Birth^2) effect on MntDoor_Locks. The effect is characterized by a positive exponential relationship, which means that the absolute effect of Year_Birth on MntDoor_Locks is positive and then grows exponentially for higher values of Year_Birth.

Moderation hypotheses:

– Hypothesis (H_5): Income has a moderated effect on MntDoor_Locks and is moderated by Kidhome. The moderation effect is positive, which means that the absolute effect of Income on MntDoor_Locks gets greater when Kidhome gets higher.

Mediation hypotheses:

– Hypothesis (H_6): Year_Birth has an indirect effect on MntDoor_Locks over Income. This indirect effect means that Year_Birth influences MntDoor_Locks in a non-immediate way, via directly influencing Income which in turn influences MntDoor_Locks. In this context, Income is referred to as the “mediator” of the relationship between Year_Birth and MntDoor_Locks.

II. METHODOLOGY

1. MODEL

Introduction

Introduction to Linear Regression: In order to test the hypotheses formulated in the previous chapter, this study used regression analysis with the method of OLS (Ordinary Least Squares). Regression in general relates the variation of a dependent variable to the variation of one or more independent variables to find and explain their relationship. Linear regression models a linear relationship, which means that the estimated coefficient of an independent variable is the same for all levels of the independent variable. To allow the effect of an independent variable to depend on the level of the variable, polynomials of the variable can be included in the regression. This is called polynomial or non-linear regression and tests for non-linear relationships. To allow the effect of an independent variable to depend on the level of another independent variable in the model (moderation), an interaction term which is a multiplication of the two independent variables can be included in the regression. This is called moderation analysis and tests for the presence of moderation relationships. In addition, control variables are used in the regressions. For reasons of clarity, these are not listed individually in the regression equations.

Introduction to regression models: Before the estimation, an adequate regression model was formulated. This model includes an error term that represents factors which explain a part of the variation of the dependent variable that cannot be attributed to the independent variables which are included in the model. Since the hypotheses propose different relationships with the dependent variable, multiple regression models with varying independent variables resembling the different relationships were used to test the individual hypotheses.

Note that β_1 to β_{10} represent the coefficients of the corresponding variables, while β_0 represents the intercept. ε represents the model errors.

The control variables included in the regression equation were selected to isolate influences on the dependent variable that the independent variables cannot explain.

Linear Model

– Linear model: To test the linear hypotheses (H_1 and H_2), the following formal regression model was formulated.

$$\text{MntDoor\_Locks} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{Year\_Birth} + \text{Control Variables} + \varepsilon \quad (1)$$

The dependent variable is MntDoor_Locks.

In the linearity model, the independent variables are Income and Year_Birth.

In the linearity model, the control variables are Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow.

Non-Linear Model

– Non-linear model: To test the non-linear hypotheses (H_3 and H_4), the following formal regression model was formulated.

$$MntDoor\_Locks = \beta_0 + \beta_1 Income + \beta_2 Income^2 + \beta_3 Year\_Birth + \beta_4 Year\_Birth^2 + Control\ Variables + \varepsilon \quad (2)$$

The dependent variable is MntDoor_Locks.

In the non-linearity model, the independent variables are Income, Income^2, Year_Birth and Year_Birth^2.

The non-linearity model is also frequently referenced as "polynomial" regression model because polynomials of the same variable are included in the regression. The regression thus uses the base variable and the polynomial of order to in the estimation. In this study, only one polynomial of degree 2 was included to allow for the interpretation of the effect of the independent variable. This becomes very difficult when more polynomials are added. The goal of adding polynomials to the equation was to detect non-linear relationships between the dependent and independent variables while applying a linear regression approach and the associated methodology. This allows for an intuitive interpretation of the findings.

In the non-linearity model, the control variables are Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow.

Moderation Model

– Moderation model: To test the moderation hypothesis (H_5), the following formal regression model was formulated.

$$MntDoor\_Locks = \beta_0 + \beta_1 Income + \beta_2 Kidhome + \beta_3 Kidhome\_x\_Income + Control\ Variables + \varepsilon \quad (3)$$

The dependent variable is MntDoor_Locks.

In the moderation model, the independent variables are Income, Kidhome and Kidhome_x_Income, where the former is the moderator and the latter the interaction term.

The variables were centered in order to manage multicollinearity issues. This means that the mean of the variable was subtracted from each observed value of the two variable for the variables multiplied in the interaction term. The interaction term was then calculated based on the centered variables. This helps to avoid inflated variances due to the close relationship between the variables involved (base variables and interaction). These inflated variances are commonly detected because they exhibit high vif values (variance inflation factors) which shows a large correlation with the set of other independent variables used in the regression. However, the choice to center the variables is, while being very common, not strictly necessary because a multicollinearity problem can be considered to be absent in the first place. The reason is that the variables involved are not independent by design and thus low vif values can not be expected or required.

In the moderation model, the control variables are Year_Birth, Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow.

Mediation Model

– Mediation model: To test the mediation hypothesis (H_6), the following formal regression models were formulated.

– Introduction to mediation models: The mediation analysis (H_6) was performed using multiple models. Firstly, a simple linear regression model to determine the influence of Year_Birth on MntDoor_Locks was estimated as a supplementary model Secondly, the mediator model was estimated to test for a significant relationship between Year_Birth and Income. Thirdly, the outcome model was estimated to test for a significant relationship between Income and MntDoor_Locks. Fourthly, the final computation of the mediation paths to test for a significant indirect effect was performed. The significance of the indirect effect reveals if the data is consistent with a mediation effect. If the direct effect of is generally significant, but not significant in case the mediator is taken into account, the mediation is called 'complete', otherwise it is called 'partial'.

– Mediation, Outcome model: The outcome model to test the mediation hypothesis was formulated as follows.

$$\text{MntDoor\_Locks} = \beta_0 + \beta_1 \text{Year\_Birth} + \beta_2 \text{Income} + \text{Control Variables} + \varepsilon \quad (4)$$

The dependent variable is MntDoor_Locks.

In the outcome model of the mediation analysis, the independent variables are Year_Birth and Income, where the latter is the mediator.

In the outcome model of the mediation analysis, the control variables are Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow.

– Mediation, Mediator model: The mediation model to test the mediation hypotheses was formulated as follows.

$$\text{Income} = \beta_0 + \beta_1 \text{Year\_Birth} + \text{Control Variables} + \varepsilon \quad (5)$$

Within the mediatior model, the dependent variable is the mediator Income.

In the mediator model of the mediation analysis, the independent variables are Year_Birth.

In the mediator model of the mediation analysis, the control variables are Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow.

Additional background

💬 – Introduction to hypothesis tests: The estimation of the regression equation with OLS is the basis for the hypothesis tests. These tests concern the question if the null hypothesis can be rejected at a determined significance level. The null hypothesis states that the respective independent variables do not explain the dependent variable. The contrasting alternative hypothesis states that the independent variables do in fact contribute to the explanation of the variation of the dependent variable. The test procedure employs t-statistics and p values that, at a predetermined significance level, reveal if the hypothesis can or cannot be rejected. In this study, a significance level of 5% is used, meaning that the null hypothesis can be expected to be wrongly rejected in 5% of the cases. If the p value is below the threshold of 0.05, the null hypothesis can be rejected, which leads to an acceptance of the alternative hypothesis. If the p value is above the threshold, the null hypothesis cannot be rejected. Two-tailed tests are used to determine significance. Therefore, each side of the distribution refers to 2.5%.

2. DATA PREPARATION

INTRODUCTION

💬 *– Introduction to data preprocessing: Many algorithms, including those employed for linear and logistic regression, are sensible to outliers. The reason for this is that the fit of a curve (regression) is evaluated with distance measures. Thus, outliers that are not characteristic of the general population distort the results. This problem is further exacerbated in instances where squared distance is used, such as in ordinary least squares regression (OLS). Moreover, missing values and categorical data can often not be processed by statistical methods. To be able to apply algorithms for further analysis, the data set needs to undergo a data preparation or preprocessing phase. The data was therefore prepared in a multi-phased approach.*

The table below highlights the missing values and outliers per column.

Table 1

Outliers and Missing Values per Variable

Variable	Missing values	% missing	Outliers	% outliers
Year_Birth	0	0.0	0	0.0
Income	30	1.2	9	0.4
Kidhome	0	0.0	0	0.0
Teenhome	0	0.0	0	0.0
Recency	0	0.0	0	0.0
MntLighting	0	0.0	35	1.4
MntCameras	0	0.0	258	10.3
MntDoor_Locks	0	0.0	205	8.2
MntThermostats	0	0.0	258	10.3
MntSecurity_Systems	54	2.2	268	10.7
MntPremium	38	1.5	209	8.4
NumDealsPurchases	0	0.0	78	3.1
NumWebPurchases	0	0.0	58	2.3
NumCatalogPurchases	0	0.0	31	1.2
NumStorePurchases	0	0.0	0	0.0
NumWebVisitsMonth	0	0.0	13	0.5
AcceptedCmp2	0	0.0	0	0.0
AcceptedCmp3	0	0.0	0	0.0
AcceptedCmp4	0	0.0	0	0.0
AcceptedCmp5	0	0.0	0	0.0
AcceptedCmp1	0	0.0	0	0.0
Complain	0	0.0	0	0.0
DepVar	0	0.0	0	0.0
Year_Dt_Customer	0	0.0	0	0.0
Months_Dt_Customer	0	0.0	0	0.0
Day_Dt_Customer	0	0.0	0	0.0
Education_Graduation	0	0.0	0	0.0
Education_Master	0	0.0	0	0.0
Education_PhD	0	0.0	0	0.0
Marital_Status_Divorced	0	0.0	0	0.0
Marital_Status_Single	0	0.0	0	0.0
Marital_Status_Together	0	0.0	0	0.0
Marital_Status_Widow	0	0.0	0	0.0

9. PREPROCESSING MEASURES USED

🕑 - 1 columns in the provided data set (Dt_Customer) were identified as time data. The information from these columns was split into separate columns.

 - The 2 categorical variables contained in the data set, Education and Marital_Status, where one-hot encoded. One-hot encoding refers to treating each category within each categorical variable as a binary variable, where the values 1 and 0 inform us about the category affiliation of an observation. To avoid the dummy variable trap (perfect multicollinearity of columns introduced by keeping all binary encodings generated from a categorical variable), the newly created column for the category with the most observations was dropped. This led to dropping Marital_Status|Married. The dropped variables can be understood as the base case. If all other dummy columns for the categorical variable are zero for an observation, it belongs to the base case. A new dummy variable was only created in cases where more than 2.0 percent of all observations were represented by the dummy to restrict the number of variables. This led to dropping Education|2n Cycle.

🔍 - 122 missing values were contained in the original data set. Because 4.76 percent of the observations contained missing values, we decide against dropping the respective observations. In this case, only 95.24 percent of the data would be maintained, which would lead to a significant loss of information. Thus, we decide to impute (replace) the missing data points. The imputed average should be the most representative data point for the column. Therefore, in cases where the median and mean were significantly different (factor 2 or more), we imputed with the median, otherwise with the mean.

🚀 - To find critical outliers, the values outside a distance correponding to 1.5 times the interquartile range IQR (the spread of the middle half of the data) were considered potential outliers for treatment. The distance is taken from the 0.25 quantile for extremely low values and from the 0.75 quantile for extremely high values. In the case of a normal distribution, a distance of 1.5 IQR is equivalent to a 2.7σ distance from the mean. Therefore, ~ 0.7 percent of the data of a variable should be labelled outliers. We decided to treat outliers by limiting them to the most extreme value in the direction of the outlier (winsorizing).

3. REGRESSION ASSUMPTIONS

INTRODUCTION

💬 – Introduction to regression assumptions: Multiple Linear Regression with OLS requires several conditions to be fulfilled in order to return estimators that have desirable statistical properties, which guarantee the validity of the OLS estimation of the regression coefficients. These OLS estimator properties, sorted in ascending order in terms of their desirability, are I. unbiasedness, II. lowest sampling variance (= efficiency) among unbiased linear estimators (BLUE ~ best linear unbiased estimator), and III. lowest sampling variance (= efficiency) among all unbiased estimators (MLE ~ maximum likelihood estimator).

💬 – Introduction to 'BLUE' estimators: The conditions under which an OLS estimator is the BLUE estimator are described under the Gauss-Markov theorem and are also called Gauss-Markov or Multiple Linear Regression (MLR) assumptions or conditions. They are 1. Linearity, 2. Absence of autocorrelation, 3. Absence of perfect collinearity, 4. Exogeneity and 5. Homoscedasticity.

1. Linearity: The parameters estimated using the OLS method must be themselves linear. This means that if the true relationship between the dependent and independent variables is not linear, OLS can still be used if appropriate measures are taken to model the relationship as linear in the parameters. Such measures can include adding polynomials of the independent variable or using the log of a variable.

2. Absence of autocorrelation: The residuals of consecutive observations are independent. This is, for example, frequently violated when the observations are measurements within a time series.

3. Absence of strong collinearity: The independent variables are not explained by other independent variables. This problem stems from the fact that independent variables are usually not independent of each other. Therefore, they do not only explain a part of the variation of the dependent variable, but also a part of the variation of the other independent variables used in the model. To achieve a 'BLUE' estimator, perfect multicollinearity must not be present. Perfect multicollinearity is rare and refers to cases where different variables include the exact same information and are therefore linearly dependent (e.g., the value in one column is always twice the value in another). If multicollinearity is strong, but not perfect, the model can still return 'BLUE' estimators and can still be interpreted. However, the OLS algorithm has difficulties in isolating the influence of specific independent variables, which leads to broader confidence intervals and lower p values of variables that are largely explained by other variables in the model. Therefore, the estimate of the coefficient of those variables is more reactive to changes of the set of independent variables provided than in the case of largely independent variables.

4. Exogeneity: The independent variables are not correlated with the error term. This can be the case when not all relevant variables are included in the model and unobserved relationships between observed and unobserved variables underly the observed relationships.

5. Homoscedasticity:Nno matter what the values of our regressors might be, the error of the variance is constant. This means that the variance does not change for changing levels of the independent variables.

💬 – Special case unbiased estimator: If all assumptions excluding homoscedasticity are fulfilled, the OLS estimator is still unbiased, but another unbiased estimator with lower sampling variance exists. This case can be understood as the minimum requirement required for a valid interpretation of the regression results.

💬 – Special case efficient estimator: If all of the assumptions above and the additional assumption of (6.) normally distributed errors are fulfilled, the OLS estimators are equal to the maximum likelihood estimator (MLE) and thus asymptotically efficient, meaning they have the smallest sampling variance of all estimators as

the sample size grows. This case can be understood as the optimal basis for the interpretation of the regression results. Note that the assumption of a normal distribution does not refer to the variables themselves, but for the errors of the model which are the differences between the predicted and actual values of the dependent variable.

Methods used for assumption checks

💬 – Methods for assumption checks: To test the underlying OLS assumptions, this study used the following methods.

– Linearity is tested using the Rainbow test, which compares the fit of the model between different subsamples. This follows the rationale that a model that does not include relevant non-linear relationship will feature a worse fit for high or low values of the independent variables compared to the fit for the "middle" of the data. If the p value of the rainbow statistic is below $\alpha = .05$, the null hypothesis that the data is linear is rejected and non-linearity can be concluded.

– Absence of autocorrelation is tested using the Durbin-Watson-Test which tests for autocorrelation of order 1 between the residuals of neighboring observations. If the Durbin-Watson-statistic lies outside the corridor 1.5-2.5, autocorrelation can be concluded.

– Absence of strong collinearity is tested by computing the Variance Inflation Factor (VIF), which represents the extent to which a variable can be explained by another variable in the model. All variables with a VIF above 10 can be considered to exhibit collinearity. Because precisely assessing the specific impact of a control variable is not important, only variables which are relevant to a hypothesis to be tested in this study are tested for non-collinearity.

– Homoscedasticity is tested using the Goldfeld-Quandt-Test, which tests for constant error terms by comparing the variances at high and low values of the independent variables. If the p value of the Goldfeld-Quandt statistic is below $\alpha = .05$, the null hypothesis that the variance is homoscedastic is rejected and heteroscedasticity (changing variances) can be concluded.

– Normality of errors is tested using the Anderson-Darling-Test, which tests the assumption that the sample used comes from a specific (in this case normal) distribution. To test the assumption of normal residuals, the regression residuals (the differences between the actual and predicted values of the dependent variable) are calculated and used as sample. If the Anderson-Darling statistic lies above the critical value of the statistic at $\alpha = .05$, the null hypothesis that the residuals are normally distributed is rejected and not-normally distributed residuals are concluded.

– Exogeneity cannot be tested in observational studies using statistical tests. To be sure that the effects observed are causal means to make sure that the effects are not truly produced by an unobserved third variable that is related to the independent and dependent variable and thus confounds the measured effect between the observed variable. If the data set used is generated in a randomized control trial, the treatment is exogenous by design. If the data set used stems from an observational study, we assume that theoretical reasoning has led to the inclusion of all relevant variables in the data set (as control variables) so that no confounding effect can be expected.

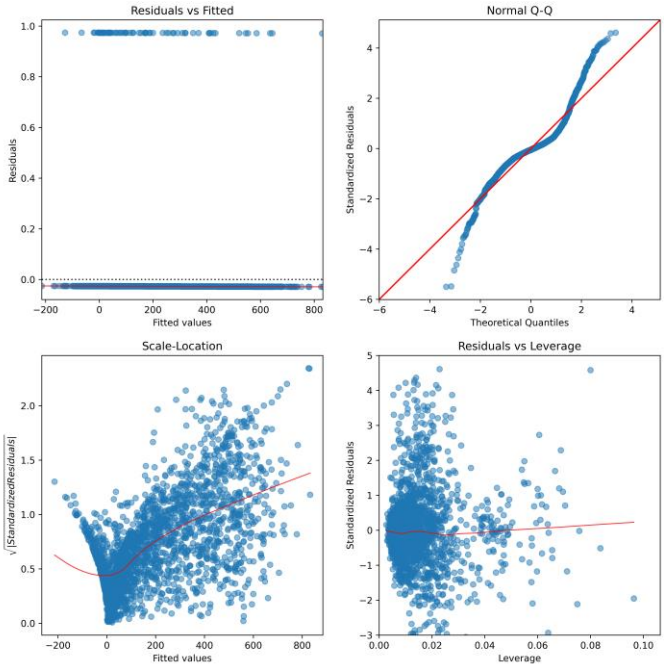
Results of assumption checks

Linear Model

– Linear model assumption check: The estimators of the linearity model are BLUE estimators, where the satisfied conditions are Linearity, Absence of autocorrelation, Absence of strong collinearity and Homoscedasticity and the unsatisfied conditions are Normality.

Figure 1

All Residual Diagnostics of the Linear Model



If a linear regression model consistently under- or overestimates for high or low values actual values of the dependent variable, nonlinearity can be assumed. The residuals versus fitted (predicted) values plot (top left) shows an acceptable fit of the straight line, indicating that the relationship between the predictors and the dependent variable are sufficiently linear (Rainbow test p=0.98).

A quantile-quantile (qq) plot of the residuals (top right) shows that the distribution of residuals was not normal (Anderson-Darling statistic of 84.88 at a critical value at α = .05 of < 0.79). Therefore, the estimators of the model are not the most efficient estimators and other estimators with lower variance exist. However, this is not an impediment to the interpreation of the model when estimates are unbiased.

A scale-location plot (bottom left) shows that the variance was approximately constant. Therefore, homoscedasticity is confirmed (Goldfeld-Quandt p=0.46).

A leverage plot (bottom right) shows that no observations have a Cook's distance of over 0.5, which means that no observations are associated with an overproportional influence on the fit of the regression.

The data did not show signs of autocorrelation (Durbin-Watson value = 2.06).

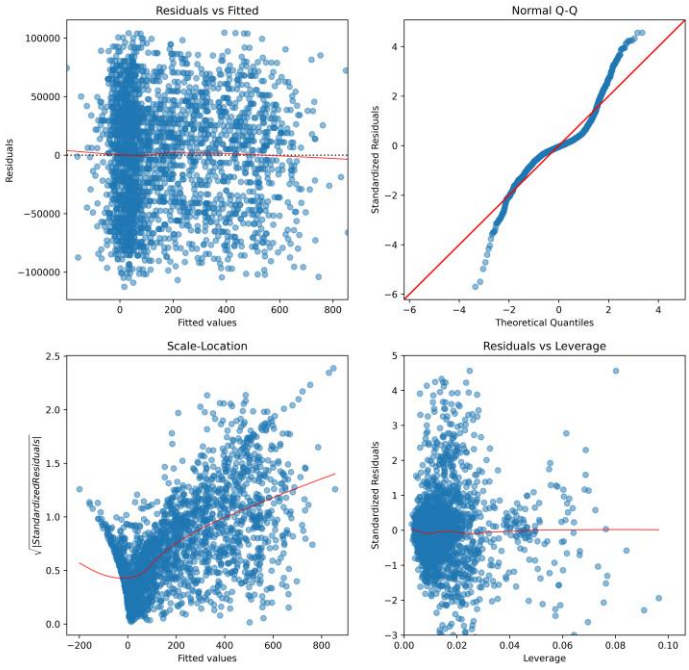
Tests for the assumption of collinearity showed that multicollinearity was not a concern. Income has a vif of 3.83 and is thus not critically inflated. Year_Birth has a vif of 1.35 and is thus not critically inflated. VIF tests were only performed for independent variables of interest (no control variables) to find out if their estimation is difficult due to inflated variances.

Non-Linear Model

– Non-linear model assumption check: The estimators of the non-linearity model are BLUE estimators, where the satisfied conditions are Linearity, Absence of autocorrelation and Homoscedasticity and the unsatisfied conditions are Absence of strong collinearity and Normality.

Figure 2

All Residual Diagnostics of the Non-Linear Model



If a linear regression model consistently under- or overestimates for high or low values actual values of the dependent variable, nonlinearity can be assumed. The residuals versus fitted (predicted) values plot (top left) shows an acceptable fit of the straight line, indicating that the relationship between the predictors and the dependent variable are sufficiently linear (Rainbow test p=0.98).

A quantile-quantile (qq) plot of the residuals (top right) shows that the distribution of residuals was not normal (Anderson-Darling statistic of 89.56 at a critical value at α = .05 of < 0.79). Therefore, the estimators of the model are not the most efficient estimators and other estimators with lower variance exist. However, this is not an impediment to the interpreation of the model when estimates are unbiased.

A scale-location plot (bottom left) shows that the variance was approximately constant. Therefore, homoscedasticity is confirmed (Goldfeld-Quandt p=0.45).

A leverage plot (bottom right) shows that no observations have a Cook's distance of over 0.5, which means that no observations are associated with an overproportional influence on the fit of the regression.

The data showed signs of autocorrelation (Durbin-Watson value = 2.06).

Tests for the assumption of collinearity showed that multicollinearity was not a concern. Income has a vif of 27.29 and is thus critically inflated. Year_Birth has a vif of 107005.18 and is thus critically inflated. The vif of the polynomial variables is not of interest since the terms are not considered independent. They inflated by design due to their relationship with the base variables. VIF tests were only performed for independent variables of interest (no control variables) to find out if their estimation is difficult due to inflated variances.

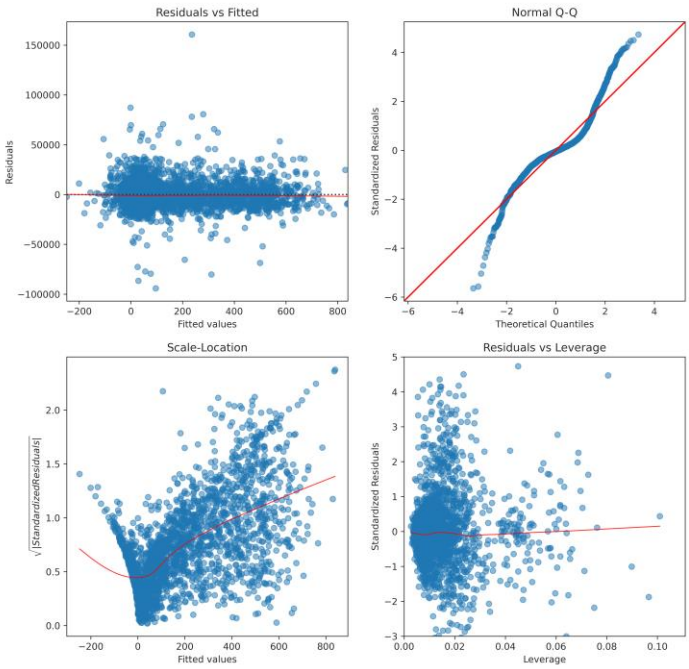
Tests for the assumption of collinearity showed that multicollinearity is present. Income has a vif of 27.29 and is thus critically inflated. Year_Birth has a vif of 107005.18 and is thus critically inflated. The vif of the polynomial variables is not of interest since the terms are not considered independent. They inflated by design due to their relationship with the base variables. VIF tests were only performed for independent variables of interest (no control variables) to find out if their estimation is difficult due to inflated variances. The variables were not dropped because they were required to test the hypotheses. The multicollinearity is not perfect (linear dependence) and thus does not bias the estimators (it does not violate the Gauss-Markov assumptions). Advice: If you require lower p values of the estimates of the model, repeat on statistcs-hero.com with fewer variables (drop more variables from the data set). The more observations are available per variable, the less multicollinearity poses an issue.

Moderation Model

– Moderation model assumption check: The estimators of the moderation model are BLUE estimators, where the satisfied conditions are Linearity, Absence of autocorrelation, Absence of strong collinearity and Homoscedasticity and the unsatisfied conditions are Normality.

Figure 3

All Residual Diagnostics of the Moderation Model



If a linear regression model consistently under- or overestimates for high or low values actual values of the dependent variable, nonlinearity can be assumed. The residuals versus fitted (predicted) values plot (top left) shows an acceptable fit of the straight line, indicating that the relationship between the predictors and the dependent variable are sufficiently linear (Rainbow test p=0.98).

A quantile-quantile (qq) plot of the residuals (top right) shows that the distribution of residuals was not normal (Anderson-Darling statistic of 85.61 at a critical value at α = .05 of < 0.79). Therefore, the estimators of the model are not the most efficient estimators and other estimators with lower variance exist. However, this is not an impediment to the interpreation of the model when estimates are unbiased.

A scale-location plot (bottom left) shows that the variance was approximately constant. Therefore, homoscedasticity is confirmed (Goldfeld-Quandt p=0.35).

A leverage plot (bottom right) shows that no observations have a Cook's distance of over 0.5, which means that no observations are associated with an overproportional influence on the fit of the regression.

The data did not show signs of autocorrelation (Durbin-Watson value = 2.05).

Tests for the assumption of collinearity showed that multicollinearity was not a concern. Income has a vif of 4.01 and is thus not critically inflated. Kidhome has a vif of 2.15 and is thus not critically inflated. The vif of the interaction term is not of interest since the term is not considered independent. It is inflated by design due to its relationship with the multiplied variables. VIF tests

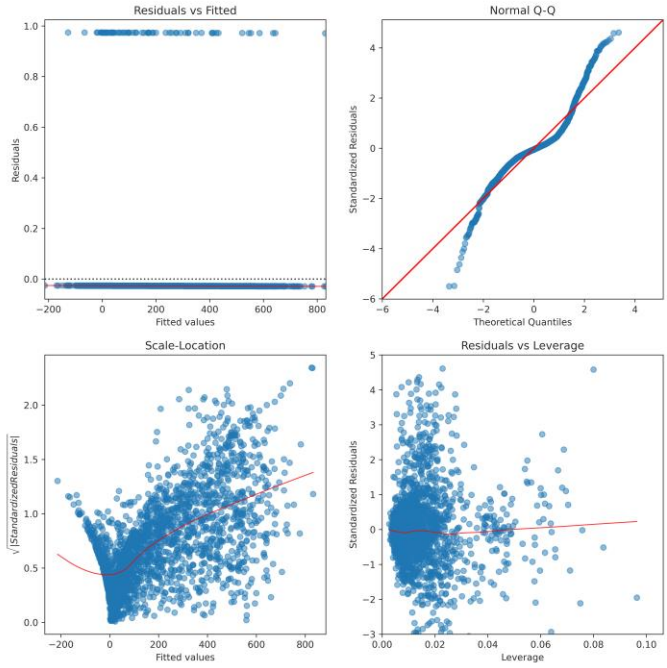
were only performed for independent variables of interest (no control variables) to find out if their estimation is difficult due to inflated variances.

Mediation Model

– Mediation Outcome model assumption check: The estimators of the outcome model within the mediation analysis are BLUE estimators, where the satisfied conditions are Linearity, Absence of autocorrelation, Absence of strong collinearity and Homoscedasticity and the unsatisfied conditions are Normality.

Figure 4

All Residual Diagnostics of the Mediation Outcome Model



If a linear regression model consistently under- or overestimates for high or low values actual values of the dependent variable, nonlinearity can be assumed. The residuals versus fitted (predicted) values plot (top left) shows an acceptable fit of the straight line, indicating that the relationship between the predictors and the dependent variable are sufficiently linear (Rainbow test p=0.98).

A quantile-quantile (qq) plot of the residuals (top right) shows that the distribution of residuals was not normal (Anderson-Darling statistic of 84.88 at a critical value at α = .05 of < 0.79). Therefore, the estimators of the model are not the most efficient estimators and other estimators with lower variance exist. However, this is not an impediment to the interpreation of the model when estimates are unbiased.

A scale-location plot (bottom left) shows that the variance was approximately constant. Therefore, homoscedasticity is confirmed (Goldfeld-Quandt p=0.46).

A leverage plot (bottom right) shows that no observations have a Cook's distance of over 0.5, which means that no observations are associated with an overproportional influence on the fit of the regression.

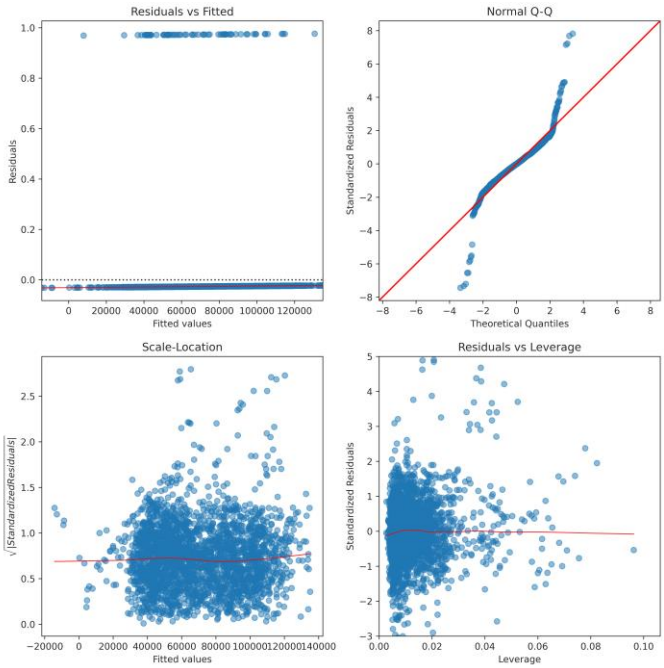
The data did not show signs of autocorrelation (Durbin-Watson value = 2.06).

Tests for the assumption of collinearity showed that multicollinearity was not a concern. Year_Birth has a vif of 1.35 and is thus not critically inflated. Income has a vif of 3.83 and is thus not critically inflated. VIF tests were only performed for independent variables of interest (no control variables) to find out if their estimation is difficult due to inflated variances.

[] – Mediation Mediator model assumption check: The estimators of the mediation model within the mediation analysis are Biased estimators, where the satisfied conditions are Absence of autocorrelation and Absence of strong collinearity and the unsatisfied conditions are Linearity, Homoscedasticity and Normality.

Figure 5

All Residual Diagnostics of the Mediation Mediator Model



If a linear regression model consistently under- or overestimates for high or low values actual values of the dependent variable, nonlinearity can be assumed. The residuals versus fitted (predicted) values plot (top left) shows a lack of fit of the straight line, indicating that the relationship between the predictors and the dependent variable is not linear (Rainbow test p<0.001). This means that the model is underfitting the data and the coefficients will not be unbiased. Advice: You can either repeat on statistics-hero.com with the adjustment of logging the dependent or independent variables (last step) or you can make a judgment call and consider the fit of the line linear enough to report that you have found linearity after visual inspection of the actual versus predicted values.

A quantile-quantile (qq) plot of the residuals (top right) shows that the distribution of residuals was not normal (Anderson-Darling statistic of 32.67 at a critical value at α = .05 of < 0.79). Therefore, the estimators of the model are not the most efficient estimators and other estimators with lower variance exist. However, this is not an impediment to the interpreation of the model when estimates are unbiased.

A scale-location plot (bottom left) shows that the variance was not approximately constant. Therefore, homoscedasticity cannot be confirmed (Goldfeld-Quandt p=0.05). This does not, however, lead to a bias of the estimators and the model is interpretable. To account for heteroscedasticity, robust errors werde used in the estimation.

A leverage plot (bottom right) shows that no observations have a Cook's distance of over 0.5, which means that no observations are associated with an overproportional influence on the fit of the regression.

The data did not show signs of autocorrelation (Durbin-Watson value = 2.0).

Tests for the assumption of collinearity showed that multicollinearity was not a concern. Year_Birth has a vif of 1.35 and is thus not critically inflated. VIF tests were only performed for independent variables of interest (no control variables) to find out if their estimation is difficult due to inflated variances.

Conclusion on assumption checks

 – Conclusion on assumption checks: Because the estimators of all regression models are at least unbiased, this study can interpret the results of all regression models.

Note that to establish an interpretation of the mediation model, the two underlying models (Outcome and Mediator model) must be interpretable.

III. RESULTS

1. DESCRIPTIVE STATISTICS

STATISTICS OF NUMERICAL VARIABLES

The following tables provides relevant descriptive statistics of the numerical data contained in the data set.

Table 2

Descriptive Statistics for Numerical Variables

Variable	N	M	Mdn	SD	Min	25%	50%	75%	Max	Kurtosis	Skewness
MntDoor_Locks	2500	180.77	68.00	237.84	1.00	17.00	68.00	251.75	1085.00	1.93	1.65
Income	2500	69567.86	68716.00	29862.82	1961.00	45646.50	68716.00	92312.25	184485.00	-0.32	0.20
Year_Birth	2500	1974.10	1975.00	12.01	1945.00	1964.00	1975.00	1983.00	2001.00	-0.85	-0.12
Kidhome	2500	0.46	0.00	0.54	0.00	0.00	0.00	1.00	2.00	-0.81	0.60
Teenhome	2500	0.49	0.00	0.55	0.00	0.00	0.00	1.00	2.00	-0.83	0.51
Recency	2500	48.67	48.00	29.04	0.00	24.00	48.00	74.00	99.00	-1.23	0.02
MntLighting	2500	355.89	197.00	397.57	0.00	28.00	197.00	596.25	1797.00	0.53	1.16
MntCameras	2500	26.83	8.00	40.67	0.00	1.00	8.00	33.00	199.00	4.16	2.12
MntThermostats	2500	32.83	10.00	49.14	0.00	2.00	10.00	41.00	239.00	4.02	2.10
MntSecurity_Systems	2500	34.58	11.00	53.79	0.00	3.00	11.00	40.00	301.00	5.74	2.36
MntPremium	2500	56.70	30.00	71.03	0.00	11.00	30.00	71.00	417.00	6.19	2.30
NumDealsPurchases	2500	2.22	2.00	1.78	0.00	1.00	2.00	3.00	14.00	5.26	1.93
NumWebPurchases	2500	10.13	9.00	3.34	1.00	8.00	9.00	12.00	31.00	6.84	1.58
NumCatalogPurchases	2500	4.55	4.00	2.83	0.00	2.00	4.00	6.00	13.00	0.39	1.04
NumStorePurchases	2500	5.69	4.00	3.29	0.00	3.00	4.00	8.00	13.00	-0.66	0.68
NumWebVisitsMonth	2500	5.20	5.00	2.55	0.00	3.00	5.00	7.00	18.00	0.41	0.12
AcceptedCmp2	2500	0.09	0.00	0.28	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp3	2500	0.06	0.00	0.24	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp4	2500	0.08	0.00	0.26	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp5	2500	0.06	0.00	0.24	0.00	0.00	0.00	0.00	1.00		
AcceptedCmp1	2500	0.01	0.00	0.09	0.00	0.00	0.00	0.00	1.00		
Complain	2500	0.01	0.00	0.10	0.00	0.00	0.00	0.00	1.00		
DepVar	2500	0.11	0.00	0.32	0.00	0.00	0.00	0.00	1.00		
Year_Dt_Customer	2500	2018.06	2018.00	0.68	2017.00	2018.00	2018.00	2019.00	2019.00	-0.86	-0.07
Months_Dt_Customer	2500	6.48	6.00	3.54	1.00	3.00	6.00	10.00	12.00	-1.29	0.03
Day_Dt_Customer	2500	15.53	16.00	8.70	1.00	8.00	16.00	23.00	31.00	-1.17	0.03
Education_Graduation	2500	0.28	0.00	0.45	0.00	0.00	0.00	1.00	1.00		
Education_Master	2500	0.17	0.00	0.38	0.00	0.00	0.00	0.00	1.00		
Education_PhD	2500	0.54	1.00	0.50	0.00	0.00	1.00	1.00	1.00		
Marital_Status_Divorced	2500	0.09	0.00	0.29	0.00	0.00	0.00	0.00	1.00		
Marital_Status_Single	2500	0.21	0.00	0.41	0.00	0.00	0.00	0.00	1.00		
Marital_Status_Together	2500	0.26	0.00	0.44	0.00	0.00	0.00	1.00	1.00		
Marital_Status_Widow	2500	0.03	0.00	0.16	0.00	0.00	0.00	0.00	1.00		

The dataset contains 14 binary variables, for which no skewness and kurtosis information is shown.

The preprocessed dataset included data on 33 numerical variables.

In total, data is provided for 2500 observations.

– Dependent variable statistics: We can see that for the dependent variable MntDoor_Locks, a mean of 180.77 and a standard deviation of 237.84 have been estimated based on the sample.

– Variables of interest statistics: For the other variables directly relevant to the hypotheses of this study, we found that: The variable Income has a mean of 69567.86 and a standard deviation of 29862.82. The variable Year_Birth has a mean of 1974.1 and a standard deviation of 12.01. The variable Kidhome has a mean of 0.46 and a standard deviation of 0.54.

– Introduction to data distribution: In a first assessment of the estimated distribution parameters, we see that 17 variables have a mean which is different from the median by factor 2 or more. Since a normal distribution is characterized by the identity of the mean and the median (as well as the mode), variables with

large differences between mean and median are expected to exhibit distributions that are skewed to the left or the right and thus differ significantly from the normal distribution.

💬 – Introduction to skewness: Overall, the dataset contains 9 variables (27% of all variables) with a skewness value of less than -2 or greater than 2. If the skewness is less than -2 or greater than 2, the data can be considered highly skewed. Skewness is the degree of asymmetry observed in a distribution. Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness. High skewness values thus imply that the median of variables is not centered above the mean. This can lead to problems in the application of statistical tests and procedures that require the assumption of normally distributed data.

💬 – Introduction to kurtosis: In addition, the dataset contains 0 variables (0% of all variables) with a kurtosis greater than 3 (also referred to as leptokurtic), which infers heavier tails (carrying more data) than the normal distribution. Since the normal distribution has a kurtosis of 3, a kurtosis greater than 3 is also referred to as a positive excess kurtosis.

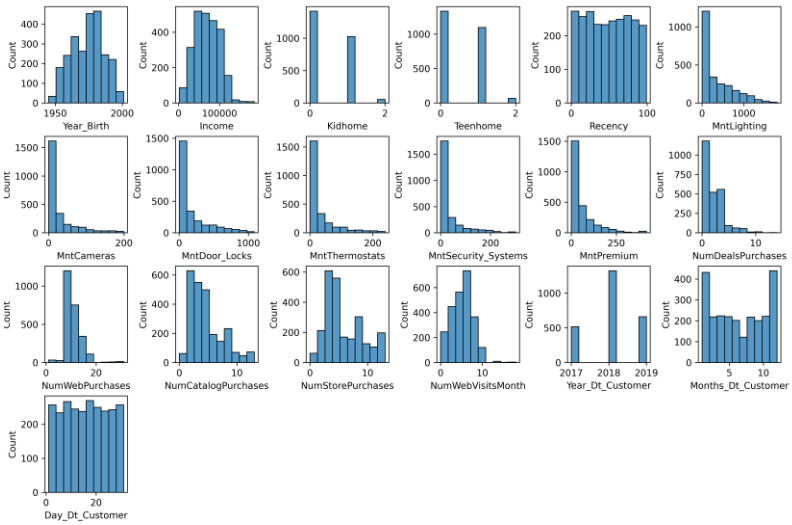
Also, the dataset contains 13 variables (39% of all variables) with a kurtosis less than 3 (also referred to as platykurtic), which infers lighter tails (carrying fewer data) than the normal distribution. Since the normal distribution has a kurtosis of 3, a kurtosis lower than 3 is also referred to as a negative excess kurtosis.

📝 – Variables of interest distribution: This study considers the data to be approximately normal for the range of skewness from -2 to +2 and excess kurtosis from -7 to +7 relating to a kurtosis range from -4 to 10 due to 3 being the normal value (Hair et al., 2010; Byrne, 2010). For the variables directly relevant to the hypotheses of this study, we found that: 🔎 - The variable MntDoor_Locks is not highly skewed because its skewness lies between 2 and -2. Its kurtosis is not outside the +/-7 range from 3 which means that the distribution is not substantially flatter or steeper than the normal distribution. In conclusion we can identify an approximately normal distribution for MntDoor_Locks as the data is not strongly skewed and is within acceptable kurtosis bounds. 🔎 - The variable Income is not highly skewed because its skewness lies between 2 and -2. Its kurtosis is not outside the +/-7 range from 3 which means that the distribution is not substantially flatter or steeper than the normal distribution. In conclusion we can identify an approximately normal distribution for Income as the data is not strongly skewed and is within acceptable kurtosis bounds. 🔎 - The variable Year_Birth is not highly skewed because its skewness lies between 2 and -2. Its kurtosis is not outside the +/-7 range from 3 which means that the distribution is not substantially flatter or steeper than the normal distribution. In conclusion we can identify an approximately normal distribution for Year_Birth as the data is not strongly skewed and is within acceptable kurtosis bounds. 🔎 - The variable Kidhome is not highly skewed because its skewness lies between 2 and -2. Its kurtosis is not outside the +/-7 range from 3 which means that the distribution is not substantially flatter or steeper than the normal distribution. In conclusion we can identify an approximately normal distribution for Kidhome as the data is not strongly skewed and is within acceptable kurtosis bounds.

A histogram buckets the range of values of each numeric variable into separate bins to visualize this distribution of the data.

Figure 6

Histograms of the Numerical Variable Distributions



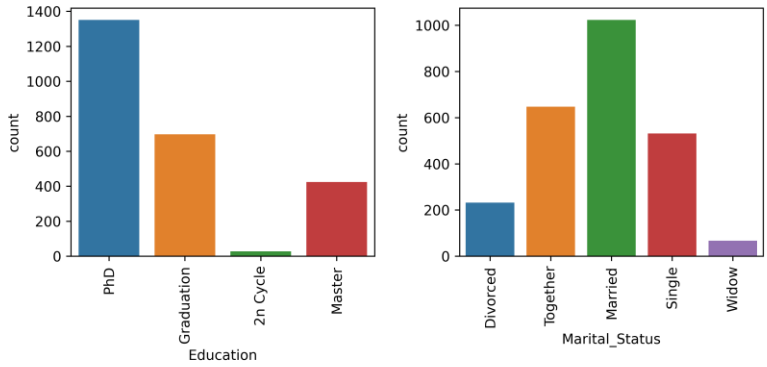
STATISTICS OF CATEGORICAL VARIABLES

In addition to aforementioned numerical variables, the original dataset also contained 2 categorical variables with a combined total of 9 categories.

A count of the observations per category is provided below.

Figure 7

Count of Categories per Categorical Variable



The categorical variables are summarized in the following table.

Table 3

Count of Categories per Categorical Variable

	MntDoor_Locks							
	count	mean	std	min	25%	50%	75%	max
2n Cycle	28	13.93	24.78	2.00	4.00	5.00	8.75	125.00
Graduation	697	178.72	241.22	2.00	13.00	53.00	250.00	1033.00
Master	424	174.80	234.72	1.00	18.00	68.50	248.00	1071.00
PhD	1351	187.15	238.35	1.00	20.00	76.00	260.50	1085.00
Divorced	232	196.53	253.60	2.00	18.00	67.00	323.25	1077.00
Married	1023	174.49	229.89	1.00	17.00	68.00	248.00	1071.00
Single	531	190.21	253.75	1.00	15.50	60.00	259.50	1042.00
Together	647	176.41	232.31	1.00	18.00	69.00	229.50	1085.00
Widow	67	189.19	225.14	2.00	23.00	109.00	259.00	900.00

🔎 - Looking at the field Education we see that among the 4 categories the most frequent is PhD with 1351 observations. It also has the highest mean for the dependent variable MntDoor_Locks with 187.15. 🔎 - Looking at the field Marital_Status we see that among the 5 categories the most frequent is Married with 1023 observations. Divorced has the highest mean for the dependent variable MntDoor_Locks with 196.53.

The precise counts are provided in the table below.

Table 4

Count of Categories per Categorical Variable

Category	Group	Frequency	Percent
Education	PhD	1351	54.0
	Graduation	697	27.9
	Master	424	17.0
	2n Cycle	28	1.1
Marital_Status	Married	1023	40.9
	Together	647	25.9
	Single	531	21.2
	Divorced	232	9.3
	Widow	67	2.7

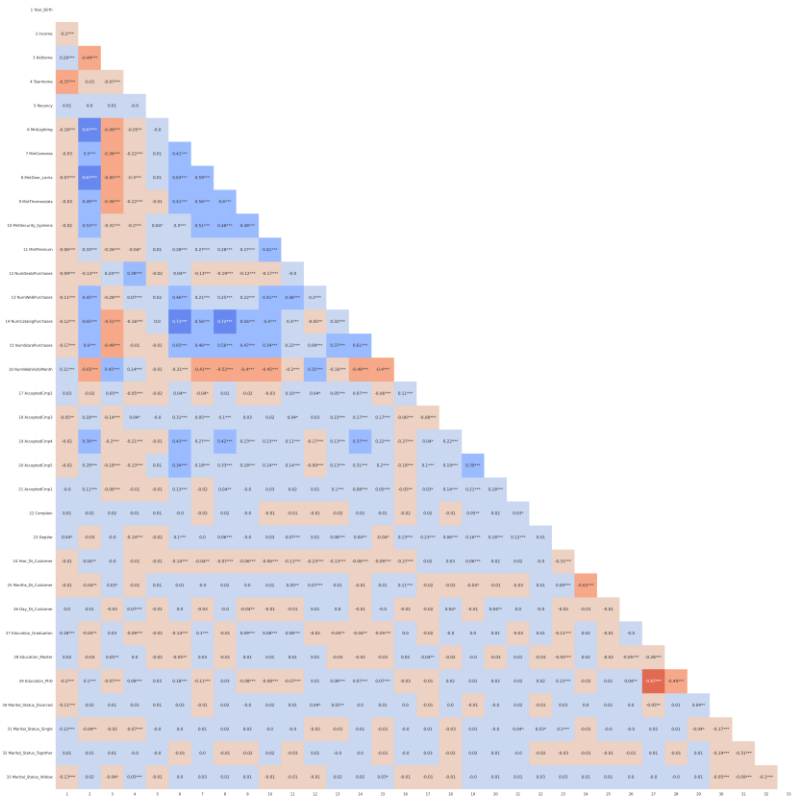
2. CORRELATION ANALYSIS

INTRODUCTION

A correlation matrix of the linear correlations is presented below.

Figure 8

Matrix of Linear Correlations



💬 *– Introduction to correlation analysis: Correlation measures aim at identifying association between two variables. If the behavior (increasing and decreasing) of one variable is associated with the behavior of another variable, these variables are said to be correlated. This correlation could be explained by one or more variables which are not considered in the pairwise correlation analysis. Therefore, a correlation does not infer a causation but merely an association. To take other variables into account, a regression analysis is required. Correlation between variables can imply a relationship which can then be tested via regression analysis. High correlation of variables can also make it hard to precisely assess the impact of any specific variable on a given dependent variable, because the behavior of the variable is associated with a specific behavior of the correlated variable. The linear correlation coefficient (called Pearson correlation coefficient) is measured on a scale that varies from + 1 through 0 to – 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. A correlation of absolute value (negative or positive) of more than 0.7 is often considered a strong correlation.*

HIGHLY CORRELATED VARIABLES

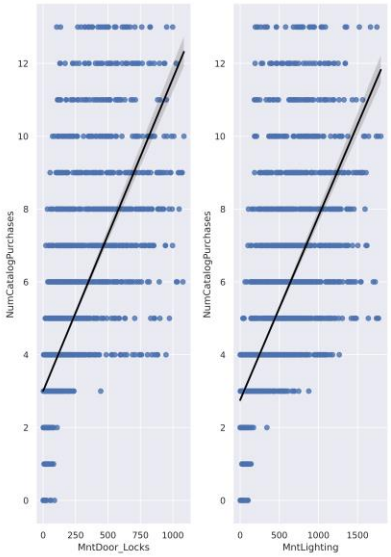
We found that in our dataset, 2 pairs of variables are highly correlated.

Among these variables, the variable pair 8 MntDoor_Locks – 14 NumCatalogPurchases shares the highest correlation.

To visualize the high correlations, a pairplot was created. A pairplot visualizes the joint distribution of the correlated variables.

Figure 9

Pairplot of Highly Correlated Variables



CORRELATIONS WITH THE DEPENDENT VARIABLE

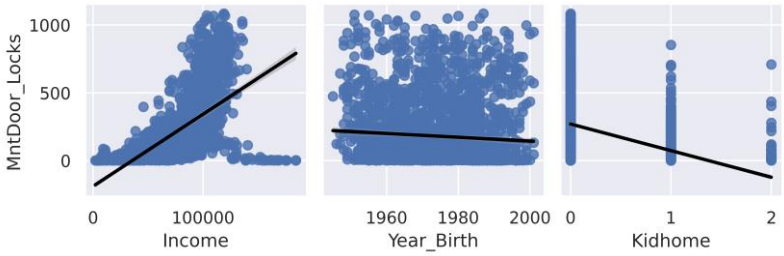
To visualize the relationship between the dependent variable and the predictors, a pairplot showing the joint distributions was created.

– Correlations with the dependent variable: In order to obtain a comprehensive picture of the association of the variables of interest in this study, different correlation measures are employed. In addition to the Pearson correlation coefficient r, the Spearman rank correlation coefficient ρ is consulted. It is useful for detecting non-linear associations because it is only concerned with the question if two variables have a monotonic relationship (if one increases, the other increases as well) rather than assessing the precise linear relationship. Therefore, the underlying data must not necessarily be linearly related. The correlations found can be summarized as follows: - The correlation between the variable Income and the dependent variable MntDoor_Locks is described by the Pearson linear correlation coefficient (r = 0.67 ; p < 0.001) and the Spearman rank correlation coefficient (ρ = 0.78 ; p < 0.001). The implied correlations are different: The linear coefficient r can be understood as implying a moderate positive relationship. However the rank coefficient ρ can be

understood as implying a strong positive relationship. This could imply a non-linear association between the variables. In this case ρ presents the more robust measure of association. Both association measures are significant. 🔎 - The correlation between the variable Year_Birth and the dependent variable MntDoor_Locks is described by the Pearson linear correlation coefficient (r = -0.07 ; p < 0.001) and the Spearman rank correlation coefficient (ρ = -0.14 ; p < 0.001). The implied correlations are different: The linear coefficient r can be understood as implying a negligible negative relationship. However the rank coefficient ρ can be understood as implying a weak negative relationship. This could imply a non-linear association between the variables. In this case ρ presents the more robust measure of association. Both association measures are significant. 🔎 - The correlation between the variable Kidhome and the dependent variable MntDoor_Locks is described by the Pearson linear correlation coefficient (r = -0.45 ; p < 0.001) and the Spearman rank correlation coefficient (ρ = -0.54 ; p < 0.001). Both can be understood as implying a moderate negative relationship. Both association measures are significant.

Figure 10

Pairplot of Dependent Variable and Predictors



3. REGRESSION ANALYSIS

LINEAR MODEL

Table 5

OLS Regression Results with MntDoor_Locks as Dependent Variable

Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	35810.000	12000.000	12400.000	59300.000	0.003
Year_Birth	-0.348	0.254	-0.846	0.150	0.170
Income	0.002	0.000	0.001	0.002	0.000
Kidhome	-5.942	6.784	-19.244	7.360	0.381
Teenhome	-66.303	6.231	-78.521	-54.085	0.000
Recency	0.118	0.091	-0.060	0.296	0.193
MntLighting	0.104	0.013	0.079	0.129	0.000
MntCameras	0.612	0.090	0.435	0.789	0.000
MntThermostats	0.631	0.073	0.487	0.775	0.000
MntSecurity_Systems	0.191	0.069	0.056	0.327	0.006
MntPremium	0.058	0.045	-0.031	0.147	0.200
NumDealsPurchases	-2.590	1.971	-6.455	1.276	0.189
NumWebPurchases	-6.966	1.072	-9.069	-4.863	0.000
NumCatalogPurchases	18.535	1.632	15.334	21.736	0.000
NumStorePurchases	3.691	1.223	1.292	6.089	0.003
NumWebVisitsMonth	-5.461	1.662	-8.721	-2.201	0.001
AcceptedCmp2	-6.423	10.194	-26.413	13.567	0.529
AcceptedCmp3	-46.472	11.807	-69.625	-23.319	0.000
AcceptedCmp4	59.233	12.355	35.006	83.459	0.000
AcceptedCmp5	49.489	12.456	25.064	73.913	0.000
AcceptedCmp1	-84.011	30.066	-142.968	-25.054	0.005
Complain	9.912	26.058	-41.186	61.010	0.704
DepVar	4.044	9.682	-14.941	23.029	0.676
Year_Dt_Customer	-17.411	5.905	-28.991	-5.831	0.003
Months_Dt_Customer	-0.030	0.987	-1.966	1.906	0.975
Day_Dt_Customer	0.444	0.304	-0.153	1.041	0.145
Education_Graduation	18.138	25.705	-32.268	68.543	0.480
Education_Master	20.788	26.071	-30.335	71.912	0.425
Education_PhD	17.977	25.682	-32.383	68.338	0.484
Marital_Status_Divorced	16.626	9.630	-2.258	35.510	0.084
Marital_Status_Single	7.938	7.135	-6.052	21.929	0.266
Marital_Status_Together	-2.915	6.622	-15.899	10.069	0.660
Marital_Status_Widow	0.076	16.776	-32.821	32.973	0.996

Note. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=69.95%. Standard errors are not heteroscedasticity-robust.

– Result H_1 and H_2: A Linear regression was carried out to test if Income and Year_Birth significantly predicted MntDoor_Locks when controlling for Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced,

Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow. The analysis was performed using ordinary least squares regression (OLS), yielding unstandardized coefficients for all effects.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Based on the H_1 hypotheses, for Income, this study expected a positive regression coefficient.

Based on the H_2 hypotheses, for Year_Birth, this study expected a positive regression coefficient.

The final predictive model was MntDoor_Locks = 35810.31 + 0.0Income - 0.35Year_Birth + Control Variables + ε

The results of the regression indicated that the model explained 69.95% of the variance and that the model was significant, F(32,2467) = 179.49, p <.05.

H_1 Interpretation: This study found evidence (at α=.05) with regard to the hypothesis that Income has a positive influence on MntDoor_Locks (B=0.0,p<0.001). The coefficient 0.0 of Income has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 0.0 for every one-unit increase of Income.

H_2 Interpretation: This study found contrary indication (but no contrary evidence) with regard to the hypothesis that Year_Birth has a positive influence on MntDoor_Locks (B=-0.35,p=0.17). The coefficient -0.35 of Year_Birth has the interpretation that the predicted value of MntDoor_Locks changes by an estimated -0.35 for every one-unit increase of Year_Birth.

Non-Linear Model

Table 6

OLS Regression Results with MntDoor_Locks as Dependent Variable

Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	-11110.000	71400.000	-151000.000	129000.000	0.876
Year_Birth	44.103	71.376	-95.860	184.066	0.537
Income	0.000	0.000	-0.001	0.001	0.491
Kidhome	-8.171	6.815	-21.535	5.194	0.231
Teenhome	-64.288	6.446	-76.928	-51.649	0.000
Recency	0.125	0.091	-0.053	0.303	0.169
MntLighting	0.102	0.013	0.077	0.128	0.000
MntCameras	0.627	0.090	0.450	0.804	0.000
MntThermostats	0.640	0.073	0.496	0.783	0.000
MntSecurity_Systems	0.153	0.070	0.016	0.291	0.029
MntPremium	0.033	0.046	-0.057	0.124	0.471
NumDealsPurchases	-1.816	1.987	-5.713	2.081	0.361
NumWebPurchases	-6.962	1.071	-9.061	-4.862	0.000
NumCatalogPurchases	19.253	1.649	16.018	22.487	0.000
NumStorePurchases	4.539	1.256	2.076	7.001	0.000
NumWebVisitsMonth	-5.168	1.665	-8.433	-1.904	0.002
AcceptedCmp2	-5.006	10.190	-24.987	14.975	0.623
AcceptedCmp3	-46.142	11.790	-69.262	-23.023	0.000
AcceptedCmp4	53.594	12.487	29.107	78.080	0.000
AcceptedCmp5	47.351	12.475	22.889	71.813	0.000
AcceptedCmp1	-86.688	30.046	-145.605	-27.770	0.004
Complain	9.396	26.018	-41.624	60.416	0.718
DepVar	4.678	9.669	-14.282	23.639	0.629
Year_Dt_Customer	-15.865	5.921	-27.476	-4.255	0.007
Months_Dt_Customer	0.138	0.987	-1.798	2.074	0.889
Day_Dt_Customer	0.458	0.304	-0.138	1.054	0.132
Education_Graduation	25.497	25.815	-25.124	76.117	0.323
Education_Master	29.352	26.250	-22.122	80.825	0.264
Education_PhD	26.548	25.855	-24.152	77.248	0.305
Marital_Status_Divorced	16.053	9.617	-2.805	34.912	0.095
Marital_Status_Single	7.954	7.160	-6.087	21.994	0.267
Marital_Status_Together	-3.150	6.612	-16.115	9.816	0.634
Marital_Status_Widow	0.589	16.765	-32.287	33.464	0.972
Income^2	0.000	0.000	0.000	0.000	0.004
Year_Birth^2	-0.011	0.018	-0.047	0.024	0.533

Note. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=70.06%. Standard errors are not heteroscedasticity-robust.

– Result H_3 and H_4: A Linear regression was carried out to test if Income, Income^2, Year_Birth and Year_Birth^2 significantly predicted MntDoor_Locks when controlling for Kidhome, Teenhome, Recency, MntLighting, MntCameras, MntThermostats, MntSecurity_Systems, MntPremium, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, Complain, DepVar, Year_Dt_Customer, Months_Dt_Customer, Day_Dt_Customer, Education_Graduation, Education_Master, Education_PhD, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Together and Marital_Status_Widow. The analysis was performed using ordinary least squares regression (OLS), yielding unstandardized coefficients for all effects.

– Interpretability warning: The assumption checks have shown that some variables in the model were highly correlated to an extent which could threaten the statistical significance of the estimated coefficients (Multicollinearity). This is not necessarily a problem as sufficient sample sizes and other facors can lead to

significant estimates even if some variables used in the estimation are highly correlated. However, the significance of results could increase if highly collinear variables would be dropped.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Based on the H_3 hypotheses, this study expected a positive regression coefficient for Income and a positive regression coefficient for Income^2.

Based on the H_4 hypotheses, this study expected a positive regression coefficient for Year_Birth and a positive regression coefficient for Year_Birth^2.

The final predictive model was MntDoor_Locks = -11105.56 + 0.0Income + 0.0Income² + 44.1Year_Birth - 0.01Year_Birth² + Control Variables + ε

The results of the regression indicated that the model explained 70.06% of the variance and that the model was significant, F(33,2466) = 174.85, p <.05.

The quadratic model which includes the squared terms is now compared to the same model without the higher order terms to find out if a non-linear effect can be identified. Results show evidence of a significant improvement of variance explained when quadratic terms are included in the model , ΔR² = 0.11%, F(2466, 1) = 15.8.68, p0.0.

H_3 Interpretation: 1 For the non-squared variable: This study found positive indication (but no evidence) with regard to the hypothesis that Income has a positive influence on MntDoor_Locks (B=0.0,p=0.473). The coefficients 0.0 of Income has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 0.0 for every one-unit increase of Income. 2 For the squared variable: This study found evidence (at α=.05) with regard to the hypothesis that Income^2 has a positive influence on MntDoor_Locks (B=0.0,p=0.004). The coefficients 0.0 of Income^2 has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 0.0 for every one-unit increase of Income^2. 3 Overall result: This study therefore found evidence (at α=.05) regarding overall non-linear effects attributed to Income^2. For the variable Income, the direction of the effect is the same compared to the hypothesis. For the squared variable Income^2, the direction of the effect is the same compared to the hypothesis. 4 Type of nonlinearity: The directions of the coefficients imply a possible positive exponential relationship, which means that the absolute effect of Income on MntDoor_Locks is positive and then grows exponentially for higher values of Income.

H_4 Interpretation: 1 For the non-squared variable: This study found positive indication (but no evidence) with regard to the hypothesis that Year_Birth has a positive influence on MntDoor_Locks (B=44.1,p=0.537). The coefficients 44.1 of Year_Birth has the interpretation that the predicted value of MntDoor_Locks changes by an estimated 44.1 for every one-unit increase of Year_Birth. 2 For the squared variable: This study found contrary indication (but no contrary evidence) with regard to the hypothesis that Year_Birth^2 has a positive influence on MntDoor_Locks (B=-0.01,p=0.533). The coefficients -0.01 of Year_Birth^2 has the interpretation that the predicted value of MntDoor_Locks changes by an estimated -0.01 for every one-unit increase of Year_Birth^2. 3 Overall result: This study therefore found contrary indication (but no contrary evidence) regarding overall non-linear effects attributed to Year_Birth^2. For the variable Year_Birth, the direction of the effect is the same

compared to the hypothesis. For the squared variable Year_Birth^2, the direction of the effect is different compared to the hypothesis. 4 Type of nonlinearity: The directions of the coefficients imply a possible bell-curved, also called "inverted-u" relationship, which means that the absolute effect of Year_Birth on MntDoor_Locks is positive and then decreases until a turning point, after which it has the opposite effect for higher values of Year_Birth.

Due to the use of polynomials, the effect on the dependent variable depends on the level of the independent variable. Therefore, analyzing the change of the estimate or prediction of the dependent variable as the independent variable changes is helpful. Unlike in regression without polynomials, the coefficients themselves do not represent the effect size and are therefore not directly interpretable. To show the effect, the independent variable can be shown at different quantiles. Using a quintile approach to determine the quantiles splits the range of the independent variable into five bins with an equal number of observations in each. For example, the 20% quintile lies at the value of the independent variable where 20% of observations have a lower value and 80% have a higher value.

The effect of Income and Year_Birth on the predicted values MntDoor_Locks at the different quintiles is shown in the table below. For example, the variable Year_Birth has its 40% quintile at a value of 1972.0. The effect on MntDoor_Locks at this value is 86949.433.

Table 7

Non-linear Effect at Quintiles of the Independent Variable

Quintile	Income	Effect Income	Year_Birth	Effect Year_Birth
0%	1961.000	0.619	1945.000	85758.949
20%	41316.400	13.036	1962.000	86508.513
40%	59718.200	18.842	1972.000	86949.433
60%	78303.000	24.705	1978.000	87213.985
80%	97627.600	30.802	1985.000	87522.629
100%	184485.000	58.207	2001.000	88228.101

10. Moderation Model

Table 8

OLS Regression Results with MntDoor_Locks as Dependent Variable

Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	35210.000	11900.000	11900.000	58500.000	0.003
Year_Birth	-0.360	0.252	-0.855	0.134	0.153
Income	0.002	0.000	0.001	0.002	0.000
Kidhome	-18.453	7.059	-32.294	-4.611	0.009
Teenhome	-64.915	6.192	-77.058	-52.772	0.000
Recency	0.116	0.090	-0.061	0.292	0.200
MntLighting	0.096	0.013	0.071	0.121	0.000
MntCameras	0.593	0.090	0.417	0.769	0.000
MntThermostats	0.611	0.073	0.468	0.754	0.000
MntSecurity_Systems	0.189	0.069	0.055	0.323	0.006
MntPremium	0.059	0.045	-0.030	0.147	0.195
NumDealsPurchases	-0.759	1.982	-4.645	3.127	0.702
NumWebPurchases	-6.349	1.070	-8.448	-4.251	0.000
NumCatalogPurchases	16.933	1.643	13.710	20.155	0.000
NumStorePurchases	3.158	1.218	0.769	5.547	0.010
NumWebVisitsMonth	-4.618	1.657	-7.867	-1.368	0.005
AcceptedCmp2	-4.462	10.129	-24.325	15.401	0.660
AcceptedCmp3	-42.565	11.745	-65.595	-19.534	0.000
AcceptedCmp4	53.129	12.313	28.984	77.273	0.000
AcceptedCmp5	49.040	12.370	24.783	73.297	0.000
AcceptedCmp1	-90.646	29.880	-149.239	-32.054	0.002
Complain	6.759	25.884	-43.998	57.516	0.794
DepVar	1.166	9.627	-17.713	20.044	0.904
Year_Dt_Customer	-17.056	5.865	-28.557	-5.555	0.004
Months_Dt_Customer	0.131	0.981	-1.793	2.054	0.894
Day_Dt_Customer	0.448	0.302	-0.145	1.041	0.139
Education_Graduation	18.114	25.528	-31.944	68.173	0.478
Education_Master	21.768	25.892	-29.005	72.541	0.401
Education_PhD	19.664	25.507	-30.354	69.681	0.441
Marital_Status_Divorced	16.255	9.564	-2.499	35.010	0.089
Marital_Status_Single	8.722	7.087	-5.175	22.618	0.219
Marital_Status_Together	-1.316	6.582	-14.222	11.590	0.842
Marital_Status_Widow	2.507	16.666	-30.174	35.188	0.880
Kidhome_x_Income	-0.001	0.000	-0.002	-0.001	0.000

Note. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=70.38%. Standard errors are not heteroscedasticity-robust.

– Result H_5: To test for moderation effects, moderation analysis was performed using ordinary least squares regression (OLS), yielding unstandardized coefficients for all effects. The analysis was

run to determine whether the interaction between Income and Kidhome significantly predicts MntDoor_Locks.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Based on the H_5 moderation hypothesis, this study expected a positive regression coefficient for Kidhome_x_Income (the interaction term).

The final predictive model was MntDoor_Locks = 35210.02 + 0.0Income - 18.45Kidhome - 0.0Kidhome_x_Income + Control Variables + ε

The results of the regression indicated that the model explained 70.38% of the variance and that the model was significant, F(33,2466) = 177.54, p <.05.

Main effects: This study found evidence that Income has a positive influence on MntDoor_Locks (B=0.0,p<0.001). In addition, it found evidence that Kidhome has a negative influence on MntDoor_Locks (B=-18.45,p=0.009).

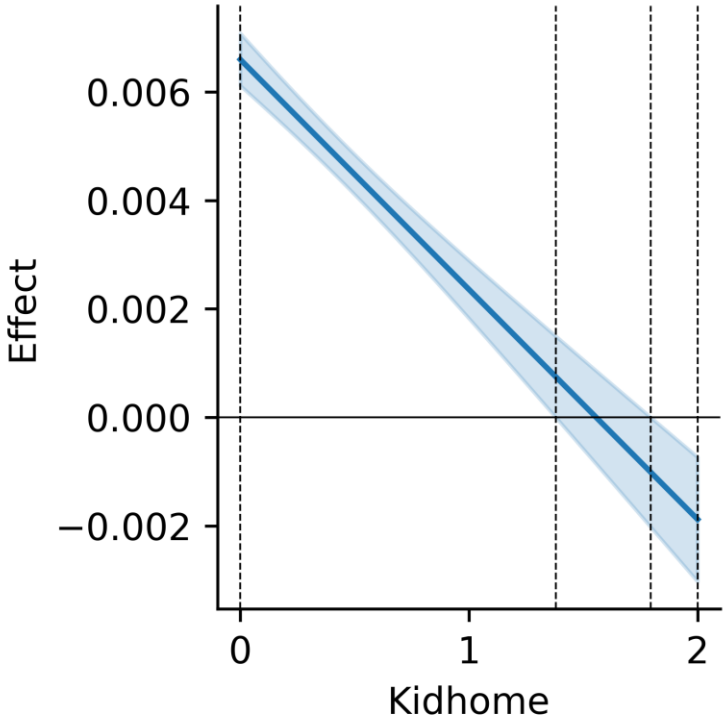
H_5 Interpretation: The moderation model which includes the interaction term is now compared to the same model without the interaction term to find out if an interaction effect can be identified. Results show evidence of a significant improvement of variance explained when the interaction term is included in the model , ΔR^2 = 0.42%, F(2466, 1) = 15.35.28, p<0.001. Therefore, this study found evidence that Kidhome generally moderated the effect between Income and MntDoor_Locks (B=-0.0,p<0.001).

On a more detailed level, an analysis of the specific range of the values of the moderator for which the interaction effect has statistical significance (Johnson-Neyman-intervals) showed significance of the interaction effect between Income and Kidhome for different levels of Kidhome. The region of significance comprises the range from 1.795 to 2 for a significant negative moderation effect by Kidhome and the range from 0 to 1.379 for a significant positive moderation effect by Kidhome.

This Johnson-Neyman result is visualized in the plot below. The straight line shows the moderation effect and the shaded area around it shows the confidence interval for the conditional effect of Income on MntDoor_Locks at a given value of the moderator Kidhome.

Figure 11

Johnson-Neyman Diagram



11. Mediation Model

– Result H_6: Mediation analyses were performed using ordinary least squares regression (OLS), yielding unstandardized path coefficients for total, direct, and indirect effects. Bootstrapping with 100 samples together with non-robust standard errors were employed to compute the confidence intervals and inferential statistics. Effects were deemed significant when the confidence interval did not include zero.

– Introduction to the steps in mediation analysis: As described, the mediation analysis is performed using multiple steps. 1. Step: Determine the influence of Year_Birth on MntDoor_Locks without taking other variables into account. 2. Step: Determine the influence of Year_Birth and Income (mediator model). 3. Step: Determine the influence of Income on MntDoor_Locks (outcome model). 4. Step: From steps 2 and 3, calculate the indirect effect of Year_Birth over Income on MntDoor_Locks. Test if, taking the indirect effect into account, the direct effect of Year_Birth on MntDoor_Locks remains significant. If the direct effect of turns out to be not significant when the mediator is taken into account, the mediation is called "complete", otherwise it is called 'partial'. To establish mediation at all, the research discussions have reached the consensus that a significant path is not required for step 1 (as originally proposed by Baron and Kenny, 1986). Some researchers claim that significant paths are required for steps 2 and 3 (MacKinnon, 2008). However, recent studies propose that a significant indirect effect in step 4 is sufficient and suggest only interpreting the indirect effect. (Zhao, Lynch & Chen, 2010; Rucker, Preacher, Tormala & Petty, 2011). This study follows this recommendation and employs the indirect effect as central measure of mediation and analysis result.

Step 1. This study found evidence (at α=.05) that, without taking other variables into account, Year_Birth has a negative influence on MntDoor_Locks, total effect c =-1.406,p<0.001. This is shown in the table below.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Table 9

(Simple) OLS Regression Results with MntDoor_Locks as Dependent Variable

Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	2956.614	780.181	1426.745	4486.483	0.000
Year_Birth	-1.406	0.395	-2.181	-0.631	0.000

$Note$. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=0.5%. Standard errors are not heteroscedasticity-robust.

Step 2. Using the mediator Income as dependent variable in the model, no evidence was found for an effect of Year_Birth on Income, path a =-46.406,p=0.141. This is shown in the table below.

Because a Goldfeld-Quandt test for homoscedasticity showed a violation of the homoscedasticity (equal variance) assumption, heteroscedasticity-robust standard errors were employed to compute the confidence intervals and inferential statistics. These accommodate for larger or smaller variances for some observations in order to avoid under-estimating the true uncertainty of the coefficient estimate (this study uses the HC1 version of Huber-White's robust standard errors).

Table 10

OLS Regression Results with Income as Dependent Variable

Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	20650.000	1780000.000	-3470000.000	3520000.000	0.991
Year_Birth	-46.406	31.542	-108.226	15.415	0.141
Kidhome	3106.229	871.194	1398.720	4813.739	0.000
Teenhome	7387.792	745.248	5927.133	8848.451	0.000
Recency	-11.620	10.639	-32.472	9.233	0.275
MntLighting	23.917	1.764	20.460	27.375	0.000
MntCameras	36.750	9.796	17.550	55.951	0.000
MntThermostats	30.586	7.640	15.612	45.560	0.000
MntSecurity_Systems	74.011	15.696	43.247	104.775	0.000
MntPremium	3.338	8.301	-12.933	19.608	0.688
NumDealsPurchases	-1167.431	240.130	-1638.076	-696.786	0.000
NumWebPurchases	1069.201	317.005	447.883	1690.520	0.001
NumCatalogPurchases	626.654	275.749	86.196	1167.113	0.023
NumStorePurchases	687.965	180.369	334.449	1041.481	0.000
NumWebVisitsMonth	-4298.717	338.412	-4961.993	-3635.441	0.000
AcceptedCmp2	1461.910	1129.645	-752.153	3675.973	0.196
AcceptedCmp3	-735.253	1512.433	-3699.568	2229.061	0.627
AcceptedCmp4	6660.346	1264.457	4182.057	9138.636	0.000
AcceptedCmp5	2544.982	1561.773	-516.036	5606.000	0.103
AcceptedCmp1	5713.709	3923.350	-1975.914	13400.000	0.145
Complain	3615.209	3137.370	-2533.922	9764.341	0.249
DepVar	-2074.676	1050.707	-4134.024	-15.329	0.048
Year_Dt_Customer	55.583	883.637	-1676.313	1787.479	0.950
Months_Dt_Customer	24.869	120.710	-211.717	261.456	0.837
Day_Dt_Customer	-13.253	38.020	-87.772	61.266	0.727
Education_Graduation	17780.000	2265.814	13300.000	22200.000	0.000
Education_Master	17490.000	2273.948	13000.000	22000.000	0.000
Education_PhD	18500.000	2223.295	14100.000	22900.000	0.000
Marital_Status_Divorced	-120.229	1059.574	-2196.956	1956.497	0.910
Marital_Status_Single	-1969.151	839.804	-3615.137	-323.166	0.019
Marital_Status_Together	-121.258	802.135	-1693.414	1450.898	0.880
Marital_Status_Widow	-740.491	1373.600	-3432.699	1951.716	0.590

$Note$. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=73.9%. Standard errors are heteroscedasticity-robust (HC1).

Step 3. In turn, the mediator Income shows no evidence regarding an effect on MntDoor_Locks, path b =0.002,p<0.001. This is shown in the table below.

Because a Goldfeld-Quandt test for homoscedasticity showed no violation of the homoscedasticity (equal variance) assumption, the standard errors used in the regression are not heteroscedasticity-robust (they are from a single Normal distribution with fixed variance for all observations).

Table 11

OLS Regression Results with MntDoor_Locks as Dependent Variable

Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	35810.000	12000.000	12400.000	59300.000	0.003
Year_Birth	-0.348	0.254	-0.846	0.150	0.170
Income	0.002	0.000	0.001	0.002	0.000
Kidhome	-5.942	6.784	-19.244	7.360	0.381
Teenhome	-66.303	6.231	-78.521	-54.085	0.000
Recency	0.118	0.091	-0.060	0.296	0.193
MntLighting	0.104	0.013	0.079	0.129	0.000
MntCameras	0.612	0.090	0.435	0.789	0.000
MntThermostats	0.631	0.073	0.487	0.775	0.000
MntSecurity_Systems	0.191	0.069	0.056	0.327	0.006
MntPremium	0.058	0.045	-0.031	0.147	0.200
NumDealsPurchases	-2.590	1.971	-6.455	1.276	0.189
NumWebPurchases	-6.966	1.072	-9.069	-4.863	0.000
NumCatalogPurchases	18.535	1.632	15.334	21.736	0.000
NumStorePurchases	3.691	1.223	1.292	6.089	0.003
NumWebVisitsMonth	-5.461	1.662	-8.721	-2.201	0.001
AcceptedCmp2	-6.423	10.194	-26.413	13.567	0.529
AcceptedCmp3	-46.472	11.807	-69.625	-23.319	0.000
AcceptedCmp4	59.233	12.355	35.006	83.459	0.000
AcceptedCmp5	49.489	12.456	25.064	73.913	0.000
AcceptedCmp1	-84.011	30.066	-142.968	-25.054	0.005
Complain	9.912	26.058	-41.186	61.010	0.704
DepVar	4.044	9.682	-14.941	23.029	0.676
Year_Dt_Customer	-17.411	5.905	-28.991	-5.831	0.003
Months_Dt_Customer	-0.030	0.987	-1.966	1.906	0.975
Day_Dt_Customer	0.444	0.304	-0.153	1.041	0.145
Education_Graduation	18.138	25.705	-32.268	68.543	0.480
Education_Master	20.788	26.071	-30.335	71.912	0.425
Education_PhD	17.977	25.682	-32.383	68.338	0.484
Marital_Status_Divorced	16.626	9.630	-2.258	35.510	0.084
Marital_Status_Single	7.938	7.135	-6.052	21.929	0.266
Marital_Status_Together	-2.915	6.622	-15.899	10.069	0.660
Marital_Status_Widow	0.076	16.776	-32.821	32.973	0.996

Note. N=2500; LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. R^2=69.95%. Standard errors are not heteroscedasticity-robust.

Step 4. We found no evidence of an indirect effect of Year_Birth over Income on MntDoor_Locks, indirect effect ab = -0.12, 95% CI[-1.85, 1.17]. This is shown in the table below.

Table 12

Indirect, Direct and Total Effects Estimated

Effect	Estimate	LL	UL	p
Total effect	-0.45	-1.76	1.05	0.44
ACME	-0.15	-1.37	1.02	0.88
ADE	-0.30	-0.84	0.31	0.36
Prop. mediated	0.54	-2.63	3.63	0.48

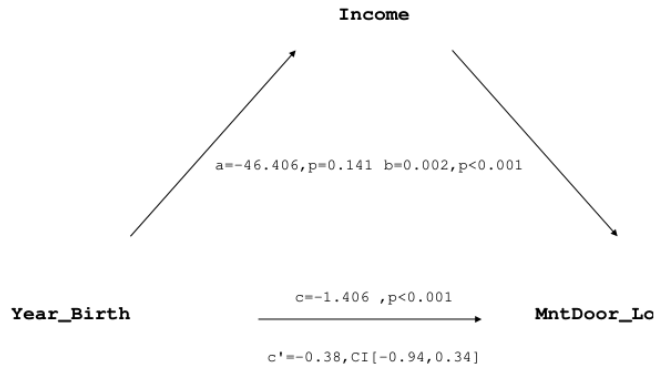
Note. ACME = average causal mediated effect or indirect effect; ADE = average direct effect or direct effect;LL = lower limit of 95% confidence interval; UL = upper limit of 95% confidence interval. Bootstrapping with 100 samples together with non-robust standard errors were employed to compute the confidence intervals and inferential statistics.

H_6 Interpretation: Because there is no evidence for a direct effect of Year_Birth on MntDoor_Locks when Income is controlled for in the model (direct effect c'=-0.38, 95% CI[-0.94, 0.34]) and no evidence for an indirect effect, this study founds that the data was consistent with the absence of mediation by Income. The total effect is -0.5 (95% CI[-1.95, 0.93]), which resembles the sum of the indirect and the direct effect.

The mediation diagram below shows the path coefficients a, b, c and c' representing unstandardized regression weights. Standard errors are provided in parentheses.

Figure 12

Mediation Paths



12. APPENDIX

Upon estimating the regression model as formulated in chapter 3.1., regression assumption checks were performed and reported in the table below.

Table 13

Regression Results

	Regression models estimated and dependent variable/y					
	MntDoor_Locks (Linear Model)	MntDoor_Locks (Polynomial model)	MntDoor_Locks (Moderation model)	MntDoor_Locks (Mediation Outcome model)	Income (Mediation mediator model)	MntDoor_Locks (Mediation supplementary model)
	(1)	(2)	(3)	(4)	(5)	(6)
const	35810.306**	-11105.560	35210.024**	35810.306**	20646.954	2956.614***
	(11954.107)	(71373.237)	(11872.649)	(11954.107)	(1783095.279)	(780.181)
Income	0.002***	0.000	0.002***	0.002***		
	(0.000)	(0.000)	(0.000)	(0.000)		
Year_Birth	-0.348	44.103	-0.360	-0.348	-46.406	-1.406***
	(0.254)	(71.376)	(0.252)	(0.254)	(31.542)	(0.395)
Income^2		0.000**				
		(0.000)				
Year_Birth^2		-0.011				
		(0.018)				
Kidhome	-5.942	-8.171	-18.452**	-5.942	3106.229***	
	(6.784)	(6.815)	(7.059)	(6.784)	(871.194)	
Kidhome_x_Income			-0.001***			
			(0.000)			
AcceptedCmp1	-84.011**	-86.687**	-90.646**	-84.011**	5713.709	
	(30.066)	(30.046)	(29.880)	(30.066)	(3923.350)	
AcceptedCmp2	-6.423	-5.006	-4.462	-6.423	1461.910	
	(10.194)	(10.190)	(10.129)	(10.194)	(1129.645)	
AcceptedCmp3	-46.472***	-46.142***	-42.565***	-46.472***	-735.253	
	(11.807)	(11.790)	(11.745)	(11.807)	(1512.433)	
AcceptedCmp4	59.233***	53.594***	53.129***	59.233***	6660.346***	
	(12.355)	(12.487)	(12.313)	(12.355)	(1264.457)	
AcceptedCmp5	49.489***	47.351***	49.040***	49.489***	2544.982	
	(12.456)	(12.475)	(12.370)	(12.456)	(1561.773)	
Complain	9.912	9.396	6.759	9.912	3615.210	
	(26.058)	(26.018)	(25.884)	(26.058)	(3137.370)	
Day_Dt_Customer	0.444	0.458	0.448	0.444	-13.253	
	(0.304)	(0.304)	(0.302)	(0.304)	(38.020)	
DepVar	4.044	4.678	1.166	4.044	-2074.676*	
	(9.682)	(9.669)	(9.627)	(9.682)	(1050.707)	
Education_Graduation	18.138	25.497	18.115	18.138	17782.551***	
	(25.705)	(25.815)	(25.528)	(25.705)	(2265.814)	
Education_Master	20.788	29.352	21.768	20.788	17494.455***	
	(26.071)	(26.250)	(25.892)	(26.071)	(2273.948)	
Education_PhD	17.977	26.548	19.664	17.977	18503.073***	
	(25.682)	(25.855)	(25.507)	(25.682)	(2223.295)	
Marital_Status_Divorced	16.626	16.053	16.255	16.626	-120.229	
	(9.630)	(9.617)	(9.564)	(9.630)	(1059.574)	
Marital_Status_Single	7.938	7.953	8.721	7.938	-1969.152*	
	(7.135)	(7.160)	(7.087)	(7.135)	(839.804)	
Marital_Status_Together	-2.915	-3.150	-1.316	-2.915	-121.258	
	(6.622)	(6.612)	(6.582)	(6.622)	(802.135)	
Marital_Status_Widow	0.076	0.589	2.507	0.076	-740.491	
	(16.776)	(16.765)	(16.666)	(16.776)	(1373.600)	
MntCameras	0.612***	0.627***	0.593***	0.612***	36.750***	
	(0.090)	(0.090)	(0.090)	(0.090)	(9.796)	
MntLighting	0.104***	0.102***	0.096***	0.104***	23.917***	
	(0.013)	(0.013)	(0.013)	(0.013)	(1.764)	
MntPremium	0.058	0.033	0.058	0.058	3.338	
	(0.045)	(0.046)	(0.045)	(0.045)	(8.301)	
MntSecurity_Systems	0.191**	0.153*	0.189**	0.191**	74.011***	
	(0.069)	(0.070)	(0.069)	(0.069)	(15.696)	
MntThermostats	0.631***	0.640***	0.611***	0.631***	30.586***	
	(0.073)	(0.073)	(0.073)	(0.073)	(7.640)	
Months_Dt_Customer	-0.030	0.138	0.131	-0.030	24.869	
	(0.987)	(0.987)	(0.981)	(0.987)	(120.710)	
NumCatalogPurchases	18.535***	19.253***	16.933***	18.535***	626.654*	
	(1.632)	(1.649)	(1.643)	(1.632)	(275.749)	
NumDealsPurchases	-2.589	-1.816	-0.759	-2.589	-1167.431***	
	(1.971)	(1.987)	(1.982)	(1.971)	(240.130)	
NumStorePurchases	3.691**	4.539***	3.158**	3.691**	687.965***	
	(1.223)	(1.256)	(1.218)	(1.223)	(180.369)	
NumWebPurchases	-6.966***	-6.962***	-6.349***	-6.966***	1069.201***	
	(1.072)	(1.071)	(1.070)	(1.072)	(317.005)	
NumWebVisitsMonth	-5.461**	-5.168**	-4.618**	-5.461**	-4298.717***	
	(1.662)	(1.665)	(1.657)	(1.662)	(338.412)	
Recency	0.118	0.125	0.116	0.118	-11.620	
	(0.091)	(0.091)	(0.090)	(0.091)	(10.639)	
Teenhome	-66.303***	-64.288***	-64.915***	-66.303***	7387.792***	
	(6.231)	(6.446)	(6.192)	(6.231)	(745.248)	
Year_Dt_Customer	-17.411**	-15.865**	-17.056**	-17.411**	55.583	
	(5.905)	(5.921)	(5.865)	(5.905)	(883.637)	
Observations	2,500	2,500	2,500	2,500	2,500	2,500
R^2	0.700	0.701	0.704	0.700	0.739	0.005
Adjusted R^2	0.696	0.697	0.700	0.696	0.736	0.005
Residual Std. Error	131.215 (df=2467)	131.011 (df=2466)	130.312 (df=2466)	131.215 (df=2467)	15350.778 (df=2468)	237.288 (df=2498)
F Statistic	179.488*** (df=32; 2467)	174.854*** (df=33; 2466)	177.536*** (df=33; 2466)	179.488*** (df=32; 2467)	367.567*** (df=31; 2468)	12.659*** (df=1; 2498)
Note:						*p<0.05; **p<0.01; ***p<0.001

Table 14

Regression Assumption Results

Topic	Statistic	Linear model	Non-linear model	Moderation model	Mediation (Mediator model)	Mediation (Outcome model)
Linearity	Rainbow p-value	0.98	0.98	0.98	0.00	0.98
	Rainbow critical p-value	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05
	Linearity result	Yes	Yes	Yes	No	Yes
Autocorrelation	Durbin Watson statistic (DW)	2.06	2.06	2.05	2.00	2.06
	No-autocorrelation range (DW)	1.5 to 2.5	1.5 to 2.5	1.5 to 2.5	1.5 to 2.5	1.5 to 2.5
	Absence of autocorrelation result	Yes	Yes	Yes	Yes	Yes
Multicollinearity	Multicollinearity: Variables affected	0	1	0	0	0
	Multicollinearity critical inflation	VIF > 10	VIF > 10	VIF > 10	VIF > 10	VIF > 10
	Multicollinearity absence result	Yes	No	Yes	Yes	Yes
Homoscedasticity	Goldfeld-Quandt p-value	0.46	0.45	0.35	0.05	0.46
	Goldfeld-Quandt p-value critical	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05
	Homoscedasticity result	Yes	Yes	Yes	No	Yes
Normal errors	Anderson-Darling statistic	84.88	89.56	85.61	32.67	84.88
	Anderson-Darling critical value	< 0.79	< 0.79	< 0.79	< 0.79	< 0.79
	Normality of errors result	No	No	No	No	No
Overall result	Estimator quality	BLUE	BLUE	BLUE	Biased	BLUE

Outlying values per variable as marked by a distance of more than 1.5 interquartile ranges from the 0.25 or 0.75 quantile are marked in the box and whisker diagram below.

Figure 13

Boxplot of Non-Binary Numerical Variables