A Work Project, presented as part of the requirements for the Award of a International Master Degree in Finance from the the NOVA – School of Business and Economics.

# Forecasting the automotive market using Autoregressive Integrated Moving Average with Python

Alberto Bianchi, 33991

A Project carried out on the Big Data Analysis course, with the supervision of:

Carlos Santos

January 3rd,2020

## Abstract

The present document is based on an internship at TIPS 4Y and aims to forecast the growth of the automotive market for the next 5 years in Portugal through the Autoregressive Integrated Moving Average model. This report starts with a literature background focused on market development paying particular attention to the electric car market. Hereafter the used methodology is described, from which a detailed explication of the process used for gathering data is defined. Subsequently to this section, a presentation of the results of the tasks is done, followed by a critical opinion about them.



*Keywords:* Forecast; TIPS 4Y; ARIMA, Data


## INDEX

# 1. Introduction

During the last decade, we saw an incredible growth of data derived from the digitalization process. Understanding and analyzing all these data has therefore become strongly necessary for companies, especially for those that depend on the prospects of future growth for a sector. In fact, archiving an overall vision of the market is indispensable for being able to make informed choices in view of future changes. The purpose of this thesis is to analyze the data provided by the company TIPS 4Y in order to forecast the future trend of the car market for the next 5 years. In particular, we will try to define the trend of the main automotive sectors, (Gasolina, Diesel, and EV) in order to define which one will dominate the market. The first step will be to define the current market situation focusing on the regulations imposed by governments that can strongly influence the spread of a certain type of product compared to another. After that, we will move on to the analysis of the methods used for data extraction and their cleaning. In the last phase, machine learning model, such as Decision Tree, and ARIMA forecasting model will be used to define the growth of the automotive sale of cars in Portugal over the next 5 years, through the use of low implementation cost models developed in Python.

## 2. Literature review

Today, transport is almost exclusively dominated by internal combustion vehicles, with approximately 95 % of transport reliant upon liquid carbon fuels (Future transport Fuels).The fuel type now represents nearly 59.5% of all new passenger car registrations, while demand for diesel has continued its decline across the EU by 1.3 million units, with diesel's market share falling from 36.3% in the second quarter of 2018 to 31.3% this year. The remaining part, on the other hand, relating to the alternatively powered vehicles has increased by 27.5% compared to last year. In particular, the electrically chargeable vehicles, such as battery-electric and plug-in hybrid which account for 2.4% of sales, have seen an increase of 36.5% over the first six months of 2019. (Petrol vehicles increase domination of European sales, 2019) The last few decades have been characterized by strong growth in the electric car sector. However, to date, this market covers only a small percentage of cars in circulation in Portugal as well as in most European countries, even if the growing trends suggest an increase in market share in the coming years. (Kane, 2018). One of the causes of the increase in this market is undoubtedly due to the crisis related to the emissions produced by motor vehicles and to the high levels of air pollution in some cities, that have caused concern to citizens and encouraged the demand for cleaner vehicles. In order to solve these problems governments have begun to apply environmental regulations. As regards the field of the automotive sector, the European community response to this problem has been to prohibit the future sale of some combustion engine vehicles and, to introduce restrictions on diesel vehicles and bans in urban areas. Obviously, governments have not only limited themselves to applying restrictions on emissions but have also given economic incentives to purchase electric cars. These incentives mainly include tax reductions, free parking and charging. For example, in Portugal private citizens have up to 3,000 euros of incentive for purchasing an electric vehicle, meanwhile,

companies have exemptions from Motor Vehicle Tax [ISV] and Single Road Tax [IUC]. (Aguiar, 2019). These limitations and incentives obviously also pushed car manufacturers to review their commercial and investment strategies with the passage of production from diesel-powered vehicles to hybrid, electric and fuel cell vehicles. In fact, the number of companies that have invested and are planning to invest more resources in the development of battery-powered cars is growing. For example, in November 2019, VW's Supervisory Board announced that VW would invest $44 billion – one-third of the corporation's planned expenses—over the next five years in an *electric offensive* that will include developing the largest electric vehicle production network in Europe and carbon-neutral manufacturing. (What to make of Volkswagen's electric vehicle "offensive"?, 2019). And yet, by 2022, Mercedes-Benz is committed to "electrifying" the entire range - from Smart to SUVs - for a total of 130 versions, including 10 entirely electric models. (Plans for more than ten different all-electric vehicles by 2022, 2018). Toyota, which has always focused on the hybrid, has decided to tackle all-electric mobility, announcing the launch of more than 10 100% electric propulsion models from now until the early years of the 2020s. (Toyota to market over 10 battery EV models in early 2020s, 2018). Furthermore, again with reference to the economic benefits of these vehicles, we know that innovations and technology have led to a significant reduction in the price of the battery, whose cost has dropped from around $ 1,000 / kWh in 2010 at $ 176 / kWh of 2018. It is estimated that this figure will continue to decline progressively, as the expected price of an average battery pack will be around $94/kWh by 2024 and $62/kWh by 2030. (Goldie-Scot, 2019) For this reason, it is possible that consumers will opt more easily for a more environmentally conscious choice as early as 2024, thanks to the price parity between electric and non-electric cars. Another element that will also contribute to the widespread use of battery-powered cars, derives from the improvement of performance. Today, some models like the Tesla model S promise an autonomy of 525 km without having to recharge,

but already from 2021 with the Tesla Roadster this limit will reach 970km (Eletric Vehicle Database). So electric cars can already represent one valid alternative to ICE [1]vehicles, not only in urban areas but even extra-urban. The sensitivity towards environmental problems and the increasing accessibility of electric models has led governments and individuals to begin a path as a contribution to environmental sustainability, which has obviously been reflected in the economy with an increase in the number of sales of electric and hybrid cars. To understand the impact of these policies we analyzed the data provided to us by TIPS 4Y and built a time series model to identify future growth prospects.

# 3. Methodology and Data

## 3.1 Data Acquisition

The main objective of the research we carried out for TIPS 4Y was to obtain future forecasts on the growth prospects of the automotive market in order to give them an analysis of their data and the opportunity to compare the results with projections produced by other companies. In order to carry out this analysis, it is necessary to work on a multitude of different data. In the following chapter, I will list the main methods used for the extraction of the date we needed and the various problems we have encountered during this operation.

The company currently has 4 different databases containing most of the information related to Portuguese vehicles. For the purposes of our analysis, we only needed to work on two of these. Since not all databases were managed directly by TIPS 4Y to obtain the necessary data we had to

[1] *Internal Combustion Engine*: traditional engines, powered by gasoline, diesel, biofuels or even natural gas.

use different extraction procedures. The first extraction was definitely the most problematic, in fact, unlike the second one we did, it was not possible to use a method of direct extraction with SQL language or through the API of the servers. The data necessary for the analysis could in fact only be extracted from the VRC[2] website with a procedure that required the sequential iteration of the insertion of the individual license plates of the company; which meant repeating the same extraction procedure manually thousands of times. To overcome this limitation, it was necessary to develop an automated extraction program. After analyzing several options, the best alternative we found was to develop a program on Python using the *Selenium*[3] libraries. This library allows the extraction of data in an iterative and automatic way through *web scraping* [Appendix]. The purpose of the following operation was essential to find the maintenance cost of each individual car. Through the discovery of this information, it was, in fact, possible to find a correlation between the characteristics of a car and the relative maintenance cost, thereby allowing us to define which characteristic was more relevant. Once this extraction process was done, we obtained a database with the following features:

- Veh_Type: Identifies if the car is owned by a company or private individual.
- Plate_Nbr: vehicle registration plate.
- Plate_Date: Year in which the car was released.
- Make: It contains the following types of brands:: Alfaromeo, Audi, BMW, citroën, Dacia, fiat, ford, honda, Hyundai, Mercedes-Benz, Mitsubishi, Nissan, Opel, Peugeot, Renault,

[2] https://www.tecvrc.pt/

[3] Selenium is a portable framework for testing web applications. Selenium provides a playback tool for authoring functional tests without the need to learn a test scripting language (Selenium IDE). It also provides a test domain-specific language (Selenese) to write tests in a number of popular programming languages, including C#, Groovy, Java, Perl, PHP, Python, Ruby and Scala. The tests can then run against most modern web browsers. (Wikipedia, s.d.)

Seat, Smart, Toyota, Volkswagen, Volvo, Mazda, Kia, mini, land rover, Skoda, Jeep, Jaguar, Suzuki, Lexus.

- Model: It represents the specific model for each Brand
- Power_kW: Measure of the engine power in kilowatt-hour (kWh)
- Fuel_Type: Defines the type of fuel used (Diesel, Gasoline, Electric, and Hybrid)
- Gross_Weight: Total weight of the car
- Nbr_Doors: Number of doors of the car
- Gearbox: Indicates if the gear shift is manual or automatic
- Cost: Variable indicating the theoretical maintenance cost for each car

## 3.2 Data Cleaning

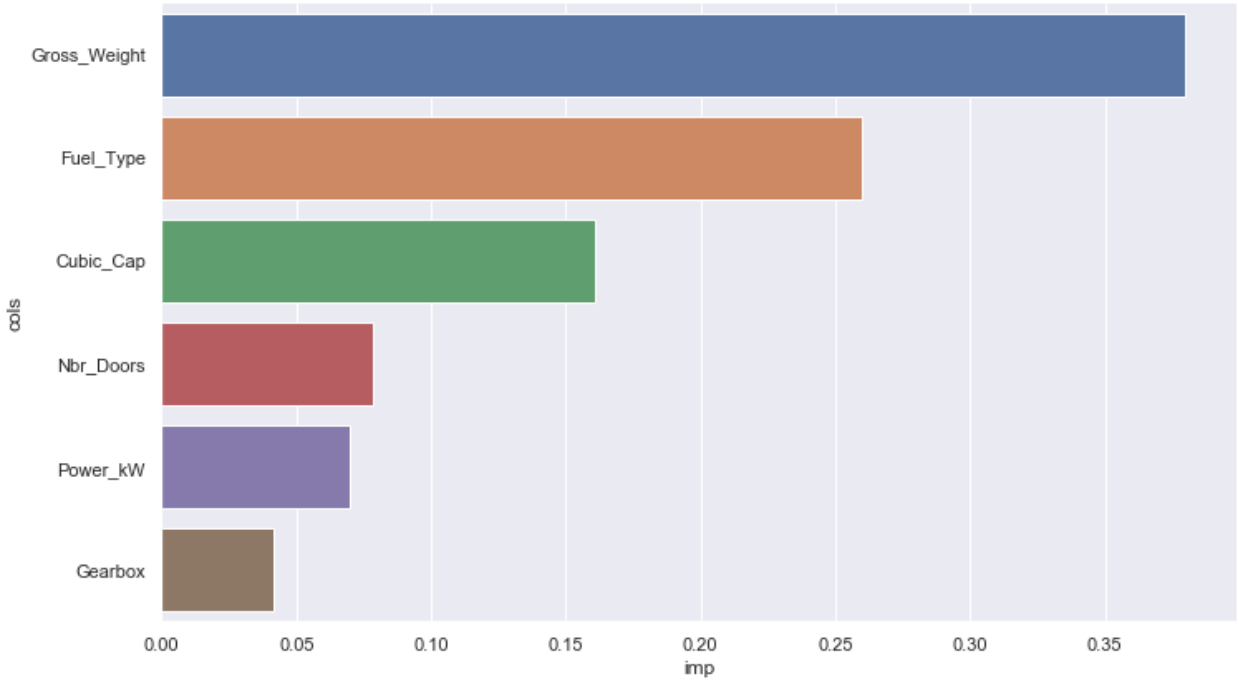
This session will cover the main problems we had within the extracted data and the types of methods applied to solve them. After obtaining the data, the next step was to perform Data Cleaning operations. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, due to misspellings during data entry, missing information or other invalid data. This process became necessary when is needed to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information (Erhard Rahm, 2000). The data cleaning procedure was carried out entirely using the Python language. There are two main reasons why we had to carry out data adjustment procedures. Firstly, since the goal of our research was to make a forecast of the car market in the coming years, there was a need to eliminate all the outliers. Some of the data obtained had in fact excessively and inexplicable values, the use of these data could

have excessively altered the results of the model, hence the need to remove them. The second problem instead derives from the need to transform the format of the variables from categorical to numeric. In fact, the model used in python was unable to treat categorical variables correctly.

## 3.3 Data Analysis

After performing the extraction and manipulation, we started analyzing the data obtained. The main objective of the analysis was to create a time series model in order to predict the level of future growth of electric vehicles. To create this model, the first step was to perform market segmentation. Segmenting the market allowed us to divide the car market in Portugal into homogeneous groups, giving us a perspective on the vehicles preferred by consumers and on the types of fuels used. Thanks to this division of the market it was possible to carry out a specific analysis on each segment for better capturing trends and consequently understand more what future developments would have been. To perform market segmentation, we used the data collected through Web-Scraping. In particular, we decided to use the maintenance cost of each car to differentiate the market in different segments. To carry out this segmentation it was, therefore, necessary to use a method capable of graphically showing the division of the variables that determined the maintenance cost. We, therefore, decided to use the Machine Learning model called Decision Tree. To create this model in python, we used the *Sklearn* library. *Sklearn* contains the *tree* library, which has a built-in classes methods for various decision tree algorithms. To develop the model it was necessary to define the dependent and independent variables. The initial database contained a total of 15 available features, however, not all of these had useful information to perform market segmentation. For example, the variable called 'model' contained within it not only all the types of models available on the market but also customized cars with unique characteristics which therefore could not be

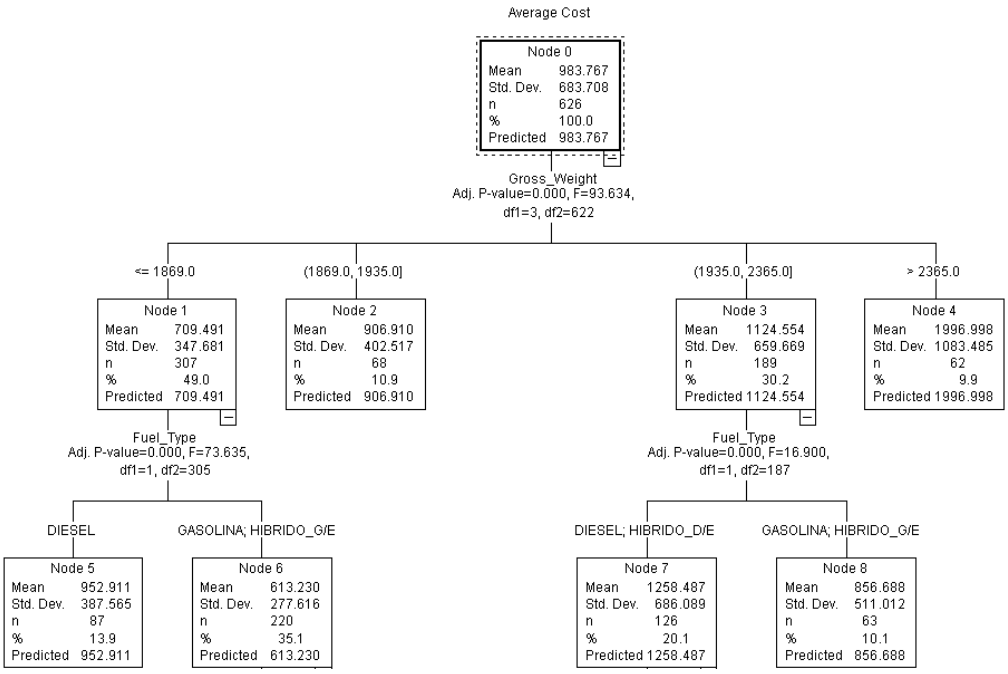
useful for segmentation. Other columns such as 'Category' and 'nbr_doors' had a single variable altogether and could, therefore, be excluded. To better understand what were the most useful features in explaining maintenance costs, we measured the level of significance of each variable:




4

Among the six variables selected, the highest significance level was associated with the Weight variable (38%), followed by Fuel_type (27%) and Cubic_Cap (17%); these three features consequently represented the main variables used by the *Sklearn* model to define the structure of the tree. Thus, the variables assigned to the independent variable X were only 6 out of 15 respectively: Cubic_Cap, Power_kW, Fuel_Type, Gross_Weight, Nbr_Doors, Gearbox while the dependent variable Y was assigned to the total maintenance cost, ie Total_Cost. After launching the model with the following features, we have obtained the following graphic representation:

[4] Significance of the features used by the Decision Tree



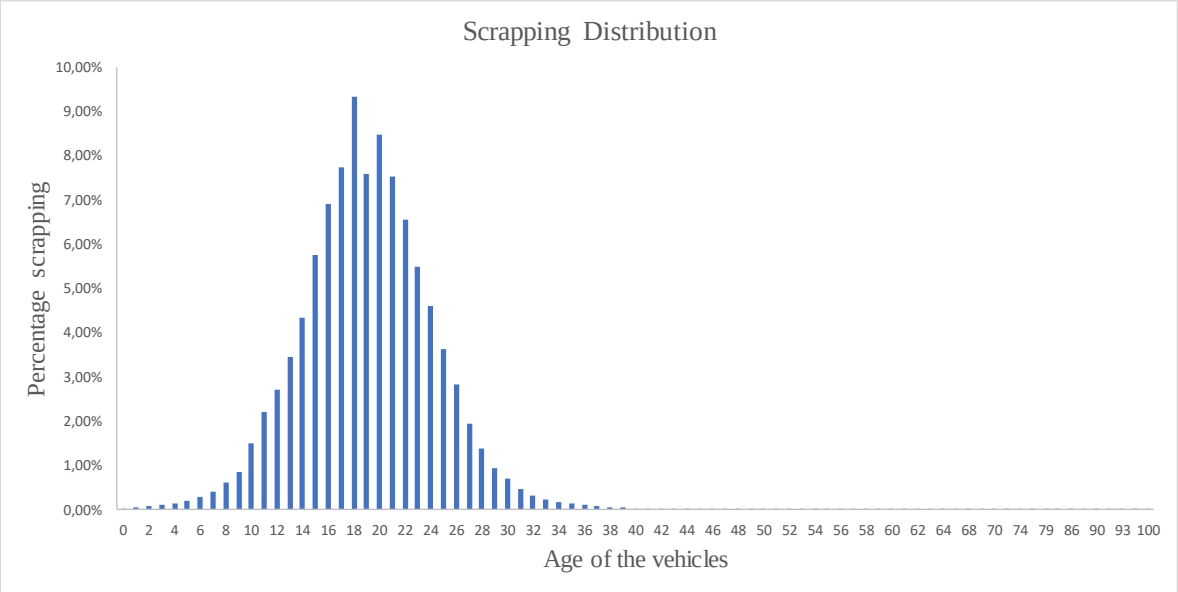

From this representation, we can see that the maintenance cost of the vehicles is divided in relation to the importance of the features. The first node, Node0, represents the entire population and this further gets divided into four homogeneous sets. The next four splits define the different weight ranges for which the price differs. In particular, Node2 and Node4 represent terminal nodes, while Nodes 1 and Node 3 split into further sub-nodes. From an analysis of the decision tree, we can see how the most significant variable to explain the maintenance cost is that relating to the weight of the cars. This observation is very curious as generally this type of variable would never be chosen by a person as an explanatory for maintenance costs. In fact, without analyzing the data, one would mainly opt for the choice of variables such as cubic cap, Power_km and Fuel_type, which would seem more appropriate in explaining the cost. The second most significant variable that determines the split of nodes 1 and 3 is that relating to Fuel_type. As can be seen, the separation by the type of fuel was used only with reference to these two nodes. A probable explanation for this decision is

that in these, the maintenance cost by type of weight differs considerably between gasoline and diesel, while for the other two nodes the average cost is relatively similar. In particular, it can be noted how the maintenance cost for diesel vehicles is generally higher for both types of weight, probably, this difference is to be associated with a more complicated or expensive type of processing. However, the following model did not give us information about electric cars. The main cause is to be associated with the few observations of electric cars we had available. Since these are relevant both in the representation of today's market and in the definition of a futures market, we decided to insert another segment that contains them. In conclusion, thanks to the decision tree we obtained 7 different segments in which 6 were represented by the tree's terminal nodes, respectively: Node5, Node6, Node2, Node7, Node8, Node4, and one by electric cars. Once we defined the different segments of the market we needed to identify all the cars that were part of it. To obtain them we performed an extraction procedure using the SQL. With this procedure, we obtained another dataset which contained the number of new cars placed on the market, differentiated for each of the following characteristics:

- Plate_Year: this feature contained the year of each vehicle starting from 1990 to 2019
- Make:  different type of brands available
- Fuel_type: description of the fuel type of the segment
- Gross_Weight: the weight of each car
- Gearbox: the type of gear of the car
- QTY: described the number of new vehicles placed on the market

Obviously for the purposes of our research, knowing only the number of new cars entering the market was not enough. To define the overall market, we also needed information relating to the exit rate of cars in relation to their age. Without this data, in fact, our forecast would not have been

accurate as it could also have counted the cars that were no longer part of the market. Obviously, the company did not have this type of information available and it was not possible to calculate it with the data available. For this reason, thanks to the help of TIPS 4Y, we contacted the ACAP[5] and requested data on the cars 'exit rate' per year. ACAP kindly provided us with their data and once we received it and made a series of changes, such as the transformation of the number of cars leaving from absolute value to the percentage value, we then obtained the following distribution:



[6]

Thanks to this last information we therefore had available both the data relating to the number of cars entering the market plus the data relating to the percentage of scrapped cars by age, which

[5] Associação Do Comércio Automóvel De Portugal

[6] The following graph represents the distribution of the percentage of cars that come out of the market according to the age of the vehicle. ie an 18-year-old car has a 9% chance of leaving the market. If we look at the percentage cumulatively we can conclude that an 18-year-old car has a total of about 46% chance of having left the market

allowed us to calculate the cars leaving the market. We could now continue our analysis and start the forecasting process.

## 4. Time Series Forecast

With Time-series Forecasting we refer to all those methods used to predict future values, within a series. The predictions of a time series can be made with two different approaches: Univariate Time Series Forecasting and Multi-Variable Time Series Forecasting statistics. The Univariate approach involves the analysis of a single variable while the multivariate analysis examines two or more variables. In our model, we will focus mostly on univariate analysis. One of the main models of univariate statistics is ARIMA. ARIMA, short for 'AutoRegressive Integrated Moving Average', uses delays and shifts in historical data to discover trends and predict the future. This type of model can be used when the series does not show patterns and when it is not random white noise[7]. The first procedure to be carried out in this model is to understand whether the series analyzed is stationary. To check the stationary nature of the series, a graphical analysis can be performed to check the trend and identify any anomalous values. Understanding if the series is stationary is fundamental as this model uses a linear regression that uses its own lags as predictors.If the series is not stationary, so if the mean and variance are not constant over time, the series needs to be integrated. Integration is a process that consists of data transformation according to two different methods: the logarithmic transformation or the differential transformation. Logarithmic transformations help to stabilize the variance of a time series, meanwhile differencing can help stabilize the mean of a time series by removing changes in the level of a time series. If the series is already stationary, no data

[7] White Noise is a random signal with equal intensities at every frequency and is often defined in statistics as a signal whose samples are a sequence of unrelated, random variables with no mean and limited variance.

modification is necessary, and the order '*d*' takes the value of 0. After analyzing the stationary nature of the series, it is necessary to identify the values to associate to *p* and *q* by analyzing the partial and total autocorrelation functions. The equation of the AR is as follows:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

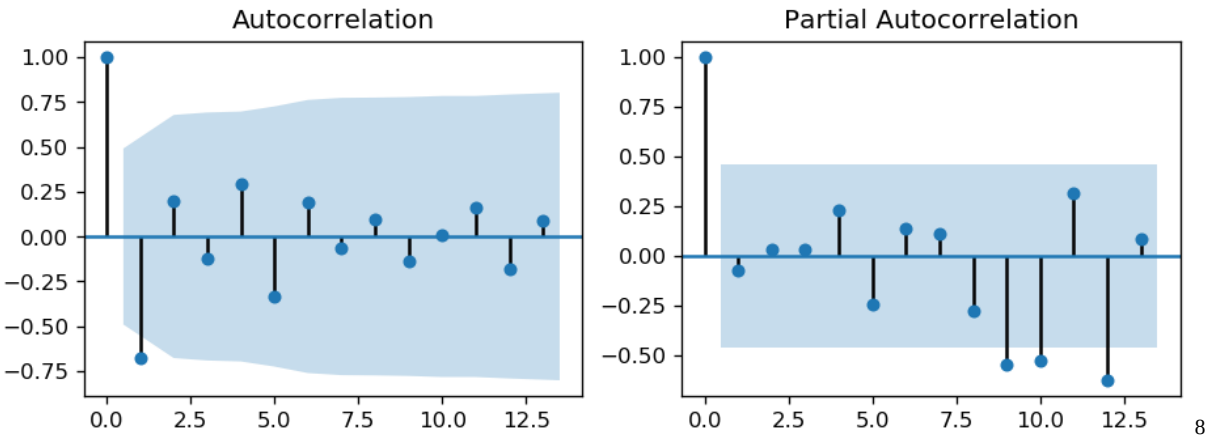
- $Y_t$: dependent variable at time *t*
- $Y_{t-1}$, $Y_{t-2}$… $Y_{t-p}$: responses in previous time periods, play the role of independent variables
- $\varepsilon_t$: white noise

Meanwhile the equation of the MA is the following:

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

- $Y_t$: dependent variable at time *t*
- $\varepsilon_{t-1}$, $\varepsilon_{t-2}$… $\varepsilon_{t-p}$: Errors in previous time periods

An intuitive and simple way to choose the model to adapt to the data consists in studying the partial and global autocorrelation function estimated on the historical series. These functions in fact measure the strength of the link between the observations of the historical series with the advantage of making different models comparable. In order to identify the values to be associated with the '*p*' and '*q'* variables of the AR and MA models, we used PACF and ACF plots.



[8]

*Figure 2: ACF and PACF of the fist segment*

The ACF and PACF plot shown show the auto dispersion of the residues. They include a 95% confidence interval band where anything outside the shaded band represents a statistically significant correlation. When one of the lags has a significant spike in the ACF, we can determine the number of moving average terms. In the same way, the spike in the PACF helps us to determine the number of autoregressive terms. From the analysis of the graph of the autocorrelation function, we observe that it has a significantly different zero spike at the first lag. We can assume, therefore, that the model that generated the series object of our analysis has a regular moving average component of order 1, which is q = 1. Analyzing the graph of the partial autocorrelation function we observe that it has a trend similar to that of the autocorrelation function. The generator model of the series will, therefore, have a regular autoregressive component of order 1, ie p = 1. After obtaining the order values ‘p’ and ‘q’ for each segment, it was necessary to verify the significance of the models. Checking the significance of the chosen model is fundamental to limit the side effects of overfitting or the practice of inserting too many parameters into the model to make it fit the data well. The fact that the model adapts well to the data does not imply also that it makes a

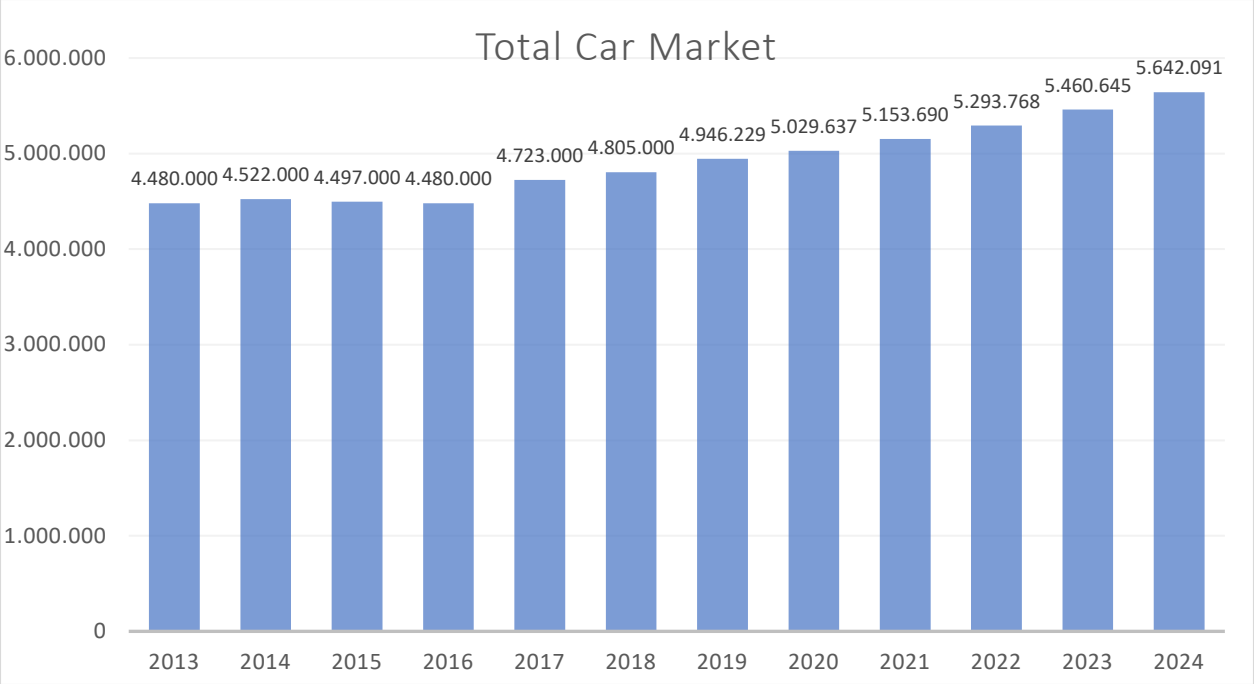
[8] Graph of the Autocorrelation function and of the Partial Autocorrelation Function of the series made stationary

good prediction, in fact, the more superfluous parameters are inserted the more the variability of the estimate increases. Generally, the main selection criterion used to determine the best combination of ARIMA model parameters is the AIC[9]. The AIC is an index that allows you to compare two models and to evaluate the variations in the adaptation of a model. AIC can only be interpreted from a comparative standpoint: the best model is always the one with the lowest AIC. So what we have to look after is not the value of the AIC itself, but the difference in AIC between the models. After trying different combinations and verifying that the significance of the coefficients was high on each segment, we launched the final ARIMA models. However, before arriving at the final results of the models it was necessary to make one last change. In fact, the results of these models would have only defined the number of new cars placed on the market for the next 5 years. However, as we know, every year a certain percentage of cars leaves the market. So after a few years, the actual number of cars will be necessarily lower than the one assumed. We, therefore, used the distribution provided by ACAP to remove a certain percentage of cars from the market based on the increase in their age.

## 5. Results

Through the analysis carried out with ARIMA, we obtained the forecast of each individual segment for the years from 2019 to 2020. The sum of these forecasts allowed us to analyze the overall car market, which we can observe in the following chart:

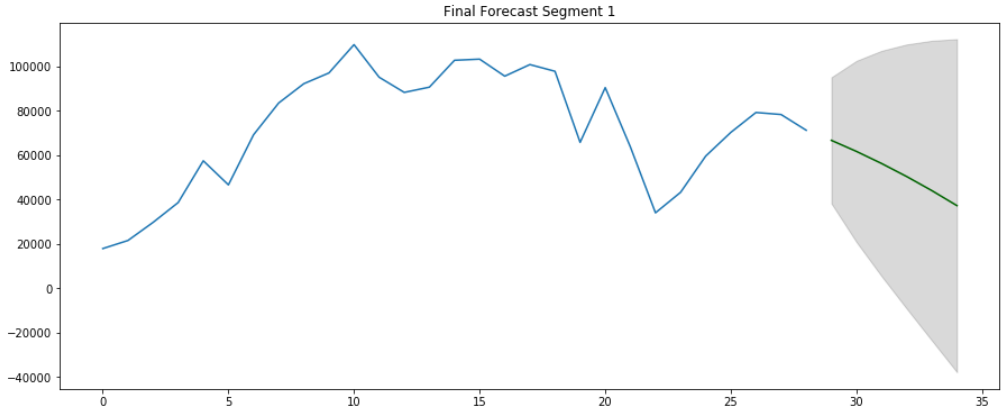
[9] Akaike's Information Criterion is usually calculated with software. The basic formula is defined as:
AIC = -2(log-likelihood) + 2K
Where: K is the number of model parameters (the number of variables in the model plus the intercept). Log-likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output. (Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media., 2003)



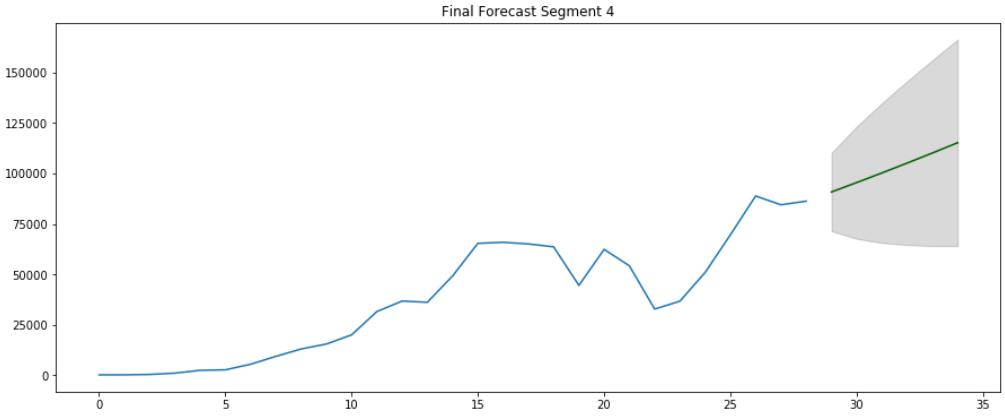

10

As you can see in the coming years we will see an increase in cars in Portugal from 4.9 million today to around 5.6 million. This perspective obviously does not take into account external elements that we could not include in our analysis, including additional restrictions on diesel/gasoline cars or the change in consumption of individuals. In fact, it is interesting to analyze the following segments individually as they already seem to reflect a change in consumer buying behavior. The first segment obtained with the decision tree, which refers to Diesel cars with a weight range up to 1869 kg, shows a sharp downward trend from the forecasts made between 2019 and 2024.

[10] Distribution of the total vehicles in Portugal from 2013 to 2024 based on our Data.

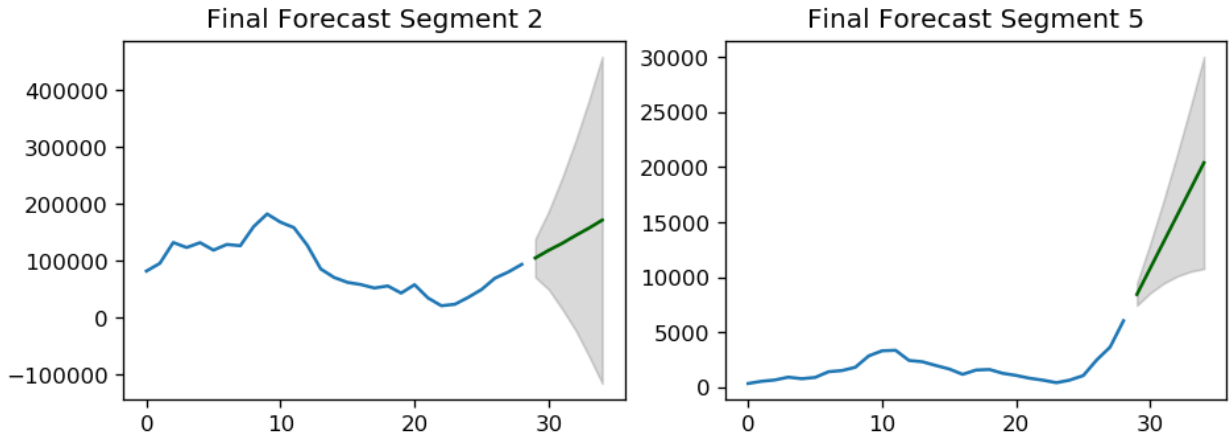



This would suggest that the market is slowly eliminating DIESEL vehicles, however, if we look at the forecasts of the fourth segment always relating to Diesel cars, but for a weight range between 1935 kg and 2365 kg, we can see how this is growing:




The reasons for this growth can mainly derive from the choice of consumers to use more capacious vehicles like SUVs or small vans. However, with a more careful analysis it can be seen that despite the fact that there is a growth on the part of vehicles with a higher weight range, overall, the number of diesel vehicles coming out of the market of the first segment is slightly higher than the increase of the fourth segment. Consequently, we can affirm that, for this sector, there will be a stationary trend in the number of cars purchased in the next 5 years and that therefore, substantially the share

percentage associated with diesel vehicles will decrease. The situation is different when we look at segments 2 and 5 belonging to the gasoline car category.

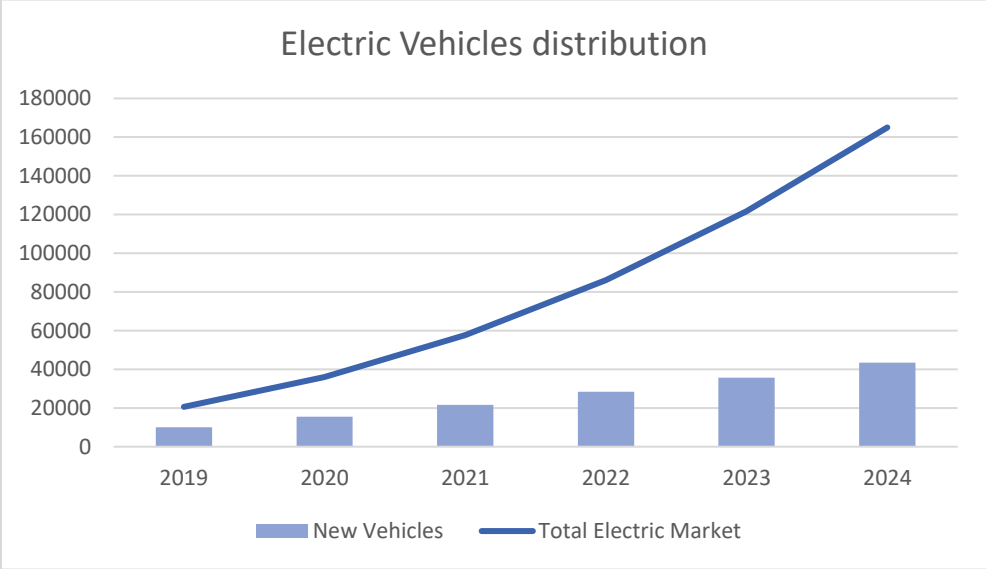



In both segments, we see a growing trend in future sales. In particular, for cars in segment 5 belonging to the weight category (1935,2365), we are seeing a sharp increase in the number of sales. The increase in gasoline cars may derive mainly from economic reasons since the average price of gasoline in Portugal has fallen significantly since 2014, so it is possible that this decrease in price has led consumers to switch from cars to diesel to cars gasoline both for economic and for pollution reduction reasons.

[11]

[11] Average price of Gasoline in Portugal

The last relevant segment for the purpose of understanding the forecast of the car market in Portugal is that relating to electric cars. Elements such as economic incentives and people's awareness of environmental sustainability have led the electric vehicle industry to grow exponentially in recent years. This trend was strongly confirmed by the distribution of the new electric cars which from 2013 to 2018 had a growth of 3144%. The ARIMA model also shows us that the growth expectation of this sector will continue uninterruptedly over the next few years, reaching a total of 165,000 electric cars in circulation in 2024 compared to the 20,000 we currently have on the market.




## 6. Conclusion

In the following thesis, we used statistical models to predict approximately the progress of the sale of new cars. In light of the results that emerged from our research, it is necessary to try to group and summarize the information gathered to elaborate a conclusive analysis that summarizes and gives a critical sense to the whole. From the results, we can see that it is the choice of consumers has moved from diesel to gasoline vehicles. Only the first five months of 2019 showed a 22.5%

reduction in sales and the reasons for this trend are different: traffic blocks for the older cars, campaigns in favor of ecological cars and greater incentives for Buying the latter is no surprise if motorists prefer to tack on alternative vehicles. There are two main alternatives chosen: gasoline and electric. Although both sectors are characterized by strong growth, it is good to dwell on the fact that the market that dominates is and will be for many more years to come that of gasoline vehicles. The electric car sector, in fact, still has certain adaptation problems. For example, we know that electric cars still lack the requisites of economic convenience and ease of use that instead characterize traditional cars, so strong consumer buying changes still remain unlikely. Furthermore, currently, we are not yet able to make electric vehicles a valid alternative to traditional vehicles that respect the trade-off between efficiency and eco-sustainability, but a change in the methods of energy production and the new opportunities offered by science will be able to favor the future transition to electric mobility. In conclusion, it is therefore highly possible that in the coming years we will see a sharp decrease in diesel vehicles. The market will mostly be dominated by petrol cars and we could potentially see an increase in the share of electric cars.

## 6.1 Limitations and future research

During the development of this study, we had several limitations. Although the observations obtained are relatively reliable, we must define all the problems that may have influenced the results of the forecasts. The ARIMA model is very efficient when working on time series especially when daily or monthly data available, this is because the greater number of data makes the model more accurate. However, all the data we had available was annual, so the model was trained on a relatively low amount of information, which therefore made the results highly variable, leading the confidence intervals of each forecast to be dispersive. The second problem relates to the independence of the segments. In fact, in creating each forecast we assumed that each segment was

independent of one another. However, this assumption is unrealistic since the correlation between segments is quite high. The application of other forecasting methods to confirm the results obtained could, therefore, be taken into consideration for future analyzes.

# Bibliography

(n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Selenium_(software)

Aguiar, C. (2019). PORTUGAL IS THE FOURTH EUROPEAN COUNTRY WITH THE HIGHEST NUMBER OF ELECTRIC VEHICLES SOLD. Retrieved from https://www.portugalms.com/en/jose-mendes-portugal-is-the-fourth-european-country-with-the-highest-number-of-electric-vehicles-sold/

Anderson, B. a. ( (2003) ). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer Science & Business Media. Retrieved from statisticshowto: https://www.statisticshowto.datasciencecentral.com/akaikes-information-criterion/

Anderson, B. a. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media.* Springer Science & Business Media.

Brid, R. S. (2018, October 26). *Introduction to Decision Trees*. Retrieved from Medium: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

*Eletric Vehicle Database.* (n.d.). Retrieved from ev-database.org: https://ev-database.org/compare/electric-vehicle-longest-range

Erhard Rahm, H. H. (2000). *Data Cleaning: Problems and Current Approaches.* University of Leipzig.

(n.d.). *Future transport Fuels.* European Expert Group.

Goldie-Scot, L. (2019, March 5). *A Behind the Scenes Take on Lithium-ion Battery Prices*. Retrieved from BloomberNEF: https://about.bnef.com/blog/behind-scenes-take-lithium-ion-battery-prices/

*Greenhouse gas.* (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Greenhouse_gas

Kane, M. (2018). *Portugal experiences really strong progress in EV acceptance*. Retrieved from insideevs: https://insideevs.com/news/342426/in-2018-electric-car-sales-in-portugal-exceeded-past-10-years-combined/

Meyer, R. P. (2014). *Climate Change 2014: Synthesis Report, Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* IPCC.

*Petrol vehicles increase domination of European sales*. (2019, September 4). Retrieved from autovistagroup: https://autovistagroup.com/news-and-insights/petrol-vehicles-increase-domination-european-sales

*Plans for more than ten different all-electric vehicles by 2022*. (2018). Retrieved from https://media.daimler.com/: https://media.daimler.com/marsMediaSite/en/instance/ko/Plans-for-more-than-ten-different-all-electric-vehicles-by-2022-All-systems-are-go.xhtml?oid=29779739

Russo, J. D. (n.d.). *Navigating The Hell of NaNs in Python*. Retrieved from https://towardsdatascience.com/: https://towardsdatascience.com/navigating-the-hell-of-nans-in-python-71b12558895b

Salvi, J. (n.d.). *Significance of ACF and PACF Plots In Time Series Analysis.* Retrieved from towardsdatascience: https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8

Salvi, J. (n.d.). *Significance of ACF and PACF Plots In Time Series Analysis.* Retrieved from towardsdatascience: https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8

*Toyota to market over 10 battery EV models in early 2020s*. (2018, December ). Retrieved from www.reuters.com: https://www.reuters.com/article/us-toyota-electric-vehicle/toyota-to-market-over-10-battery-ev-models-in-early-2020s-idUSKBN1EC0EB

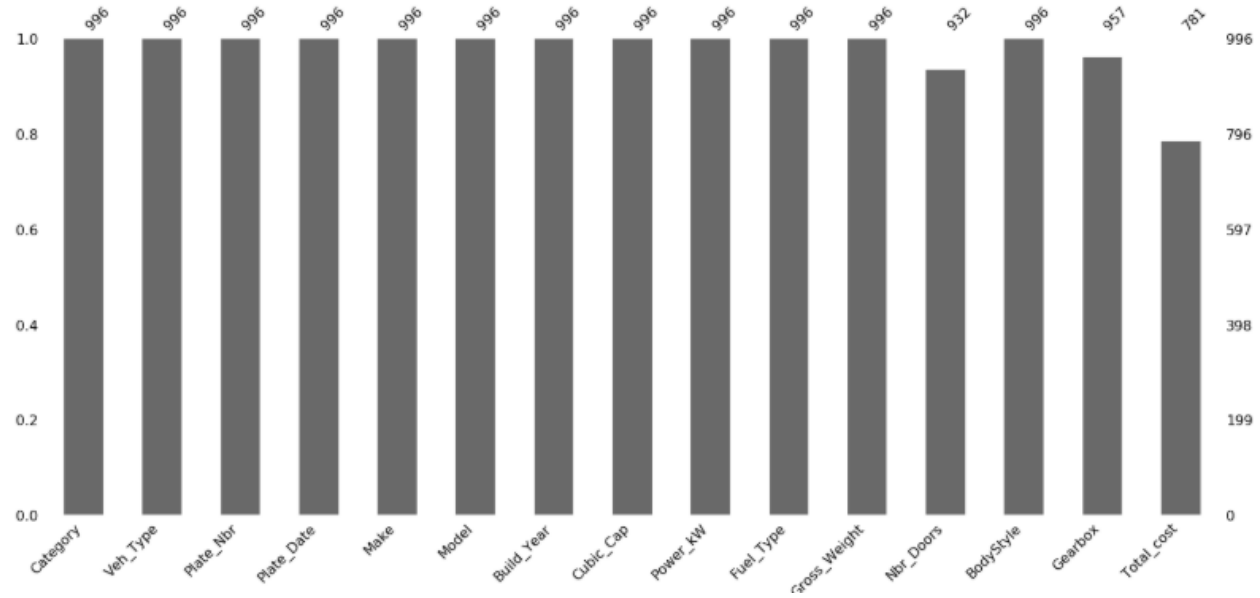
Trenberth, K. E. (2018). *Climate change caused by human activities is happening.* Journal of Energy & Natural Resources Law.

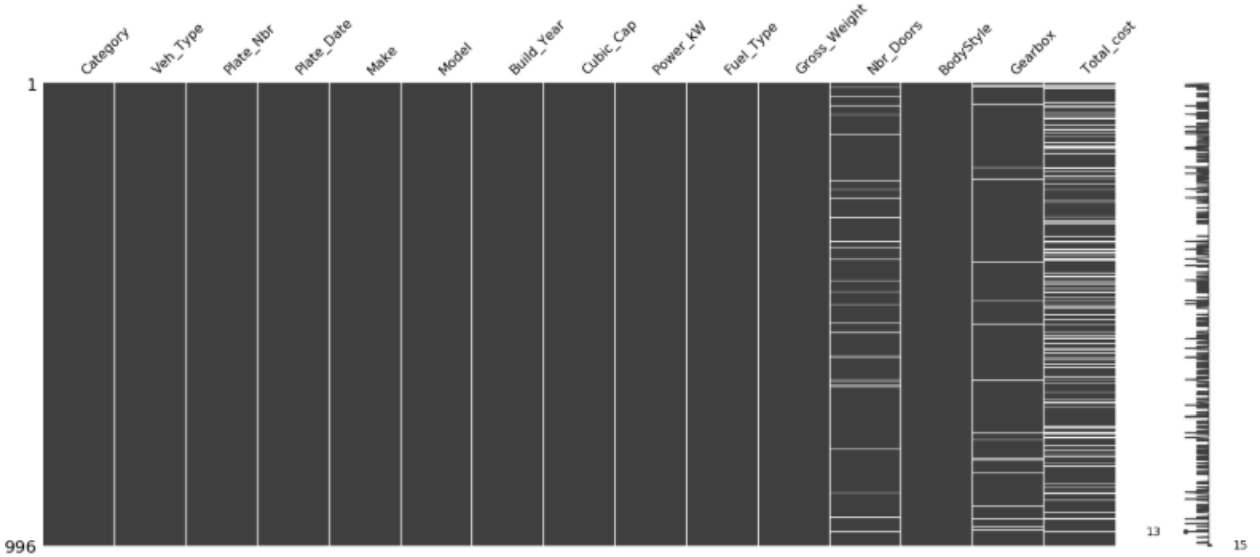
*What is web scraping?* (n.d.). Retrieved from scrapinghub: https://scrapinghub.com/what-is-web-scraping

*What to make of Volkswagen's electric vehicle "offensive"?* (2019, August 1). Retrieved from https://theicct.org/: https://theicct.org/blog/staff/vw-ev-offensive-20190801

# 7. Appendix

Appendix 1 – Missing Values of the First DataBase



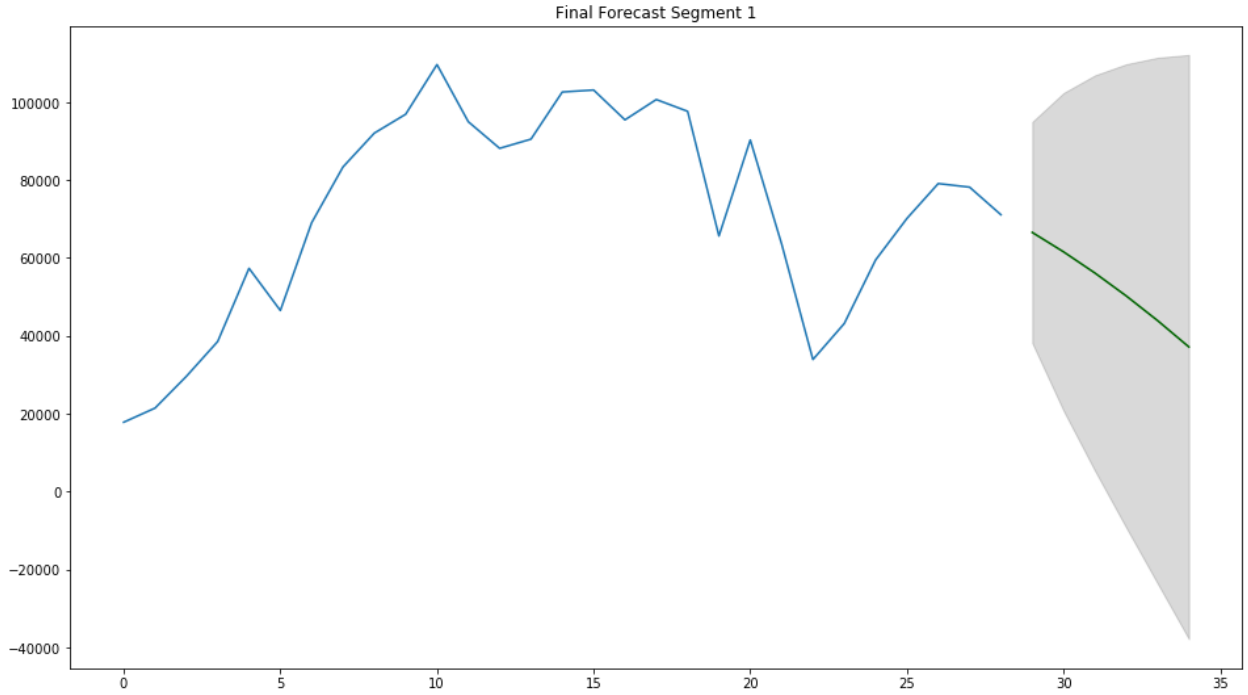
Appendix 2 – Summary of the First DataBase

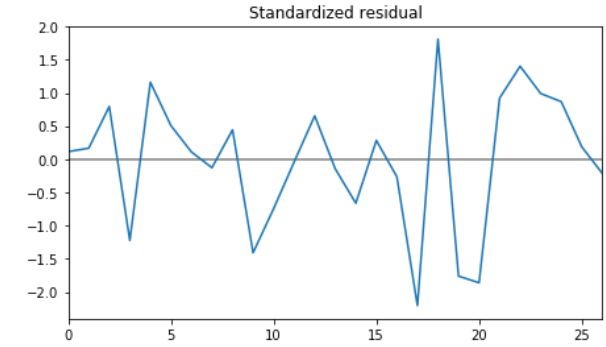
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 996 entries, 0 to 995
Data columns (total 15 columns):
Category        996 non-null object
Veh_Type        996 non-null object
Plate_Nbr       996 non-null object
Plate_Date      996 non-null int64
Make            996 non-null object
Model           996 non-null object
Build_Year      996 non-null int64
Cubic_Cap       996 non-null int64
Power_kW        996 non-null int64
Fuel_Type       996 non-null object
Gross_Weight    996 non-null int64
Nbr_Doors       932 non-null float64
BodyStyle       996 non-null object
Gearbox         957 non-null object
Total_cost      781 non-null float64
dtypes: float64(2), int64(5), object(8)
memory usage: 116.8+ KB
```

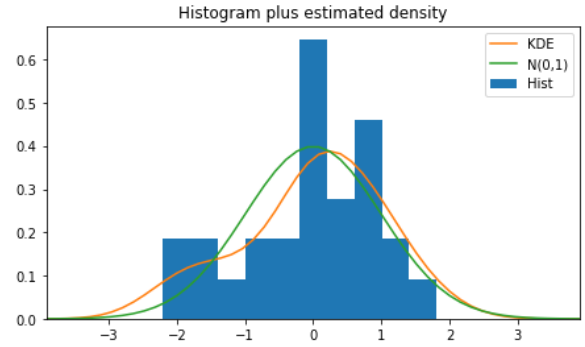
Appendix 3 – Dataframe after One-Hot Encoding

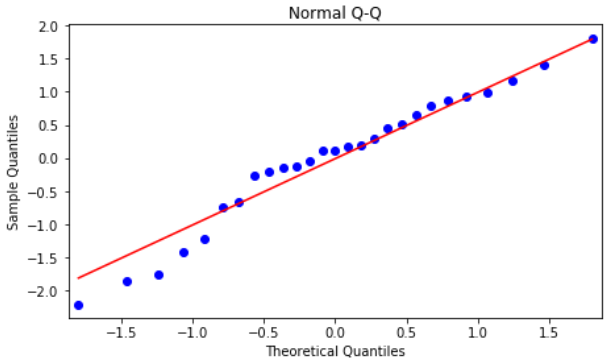
| Cubic_Cap | Power_kW | Fuel_Type | Gross_Weight | Nbr_Doors | Gearbox | Total_cost | ALFA ROMEO | AUDI | BMW | ... |
|---|---|---|---|---|---|---|---|---|---|---|
| 1461 | 81 | 1 | 1910 | 5 | 0 | 950 | 0.0 | 0.0 | 0.0 | ... |
| 1461 | 81 | 1 | 1914 | 5 | 0 | 915 | 0.0 | 0.0 | 0.0 | ... |
| 1496 | 85 | 1 | 1885 | 5 | 0 | 1013 | 0.0 | 0.0 | 1.0 | ... |
| 1461 | 66 | 1 | 1695 | 5 | 0 | 702 | 0.0 | 0.0 | 0.0 | ... |
| 1461 | 80 | 1 | 1960 | 5 | 0 | 2090 | 0.0 | 0.0 | 0.0 | ... |

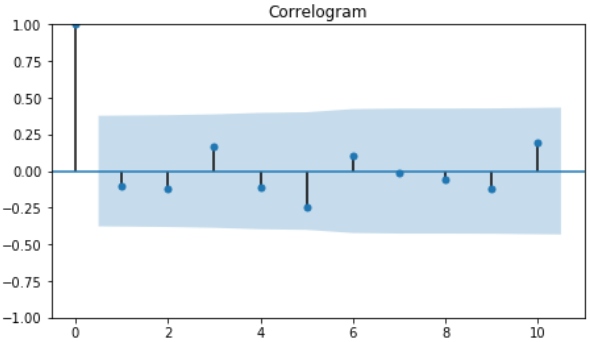
Appendix 4 – Segment 1 Diesel weight<1869



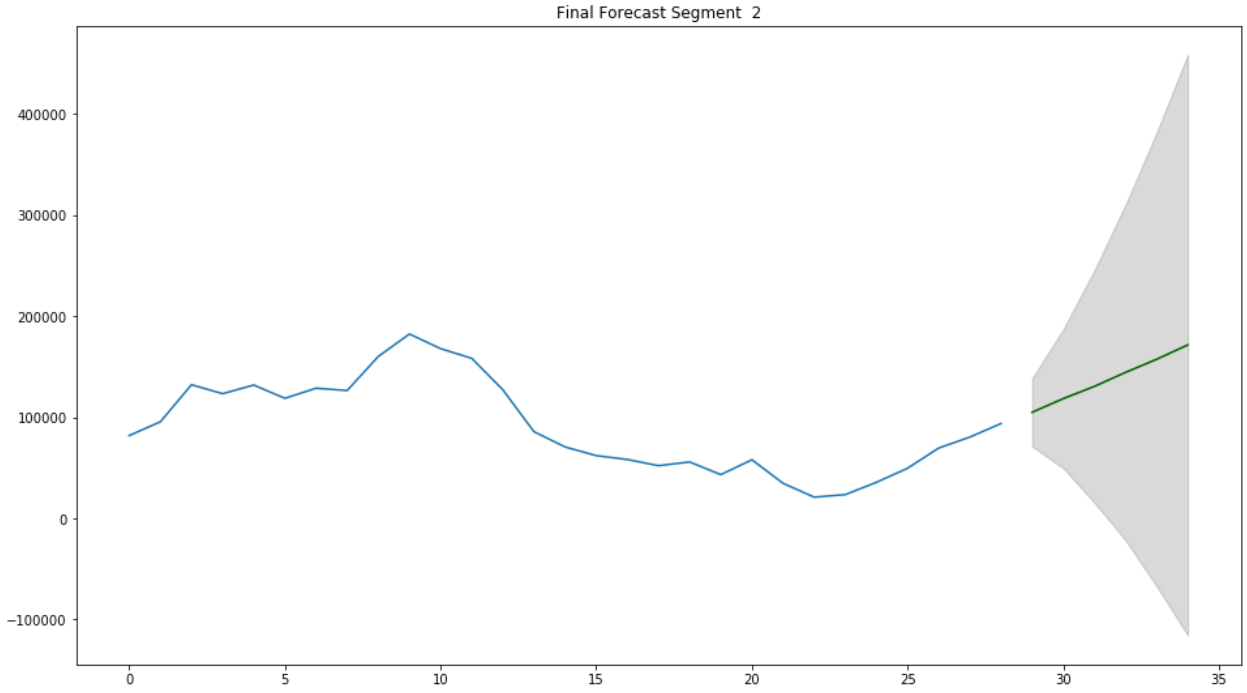



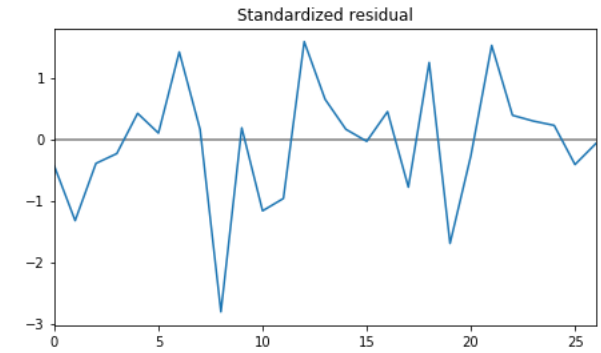



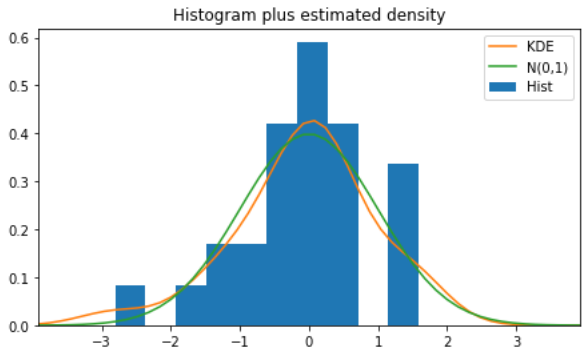



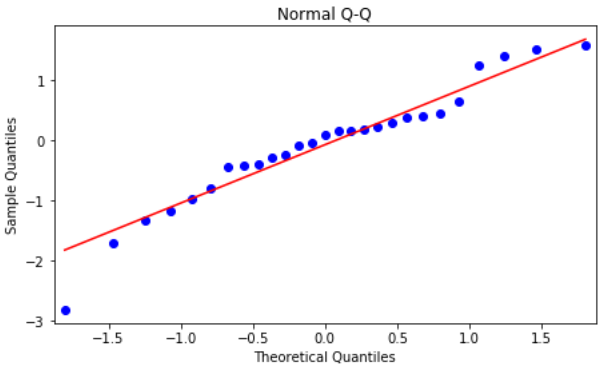



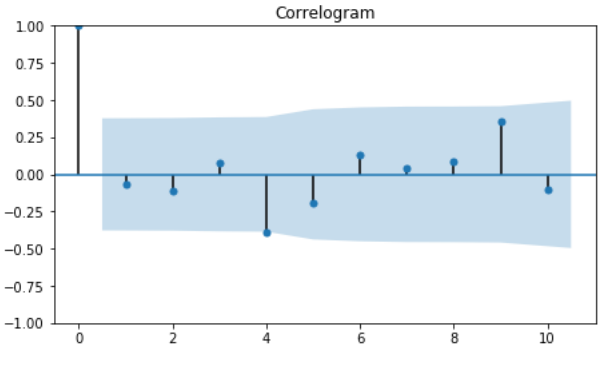
Appendix 5 – Segment 2



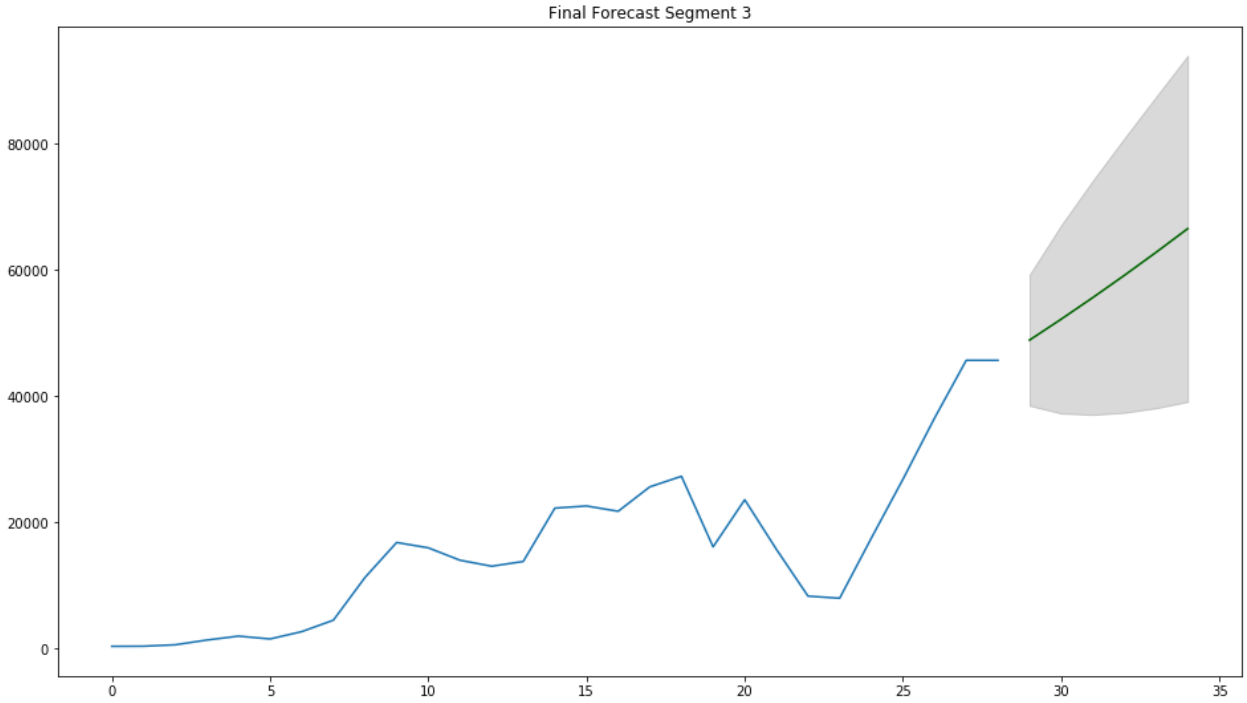



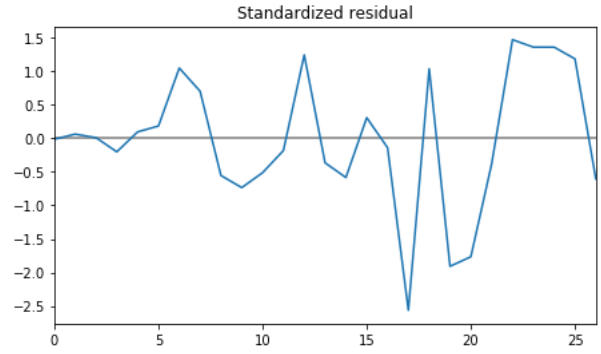



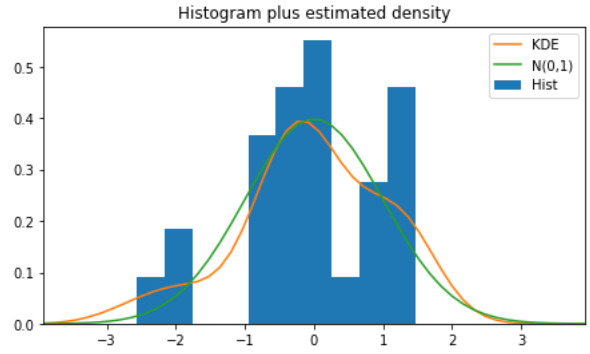



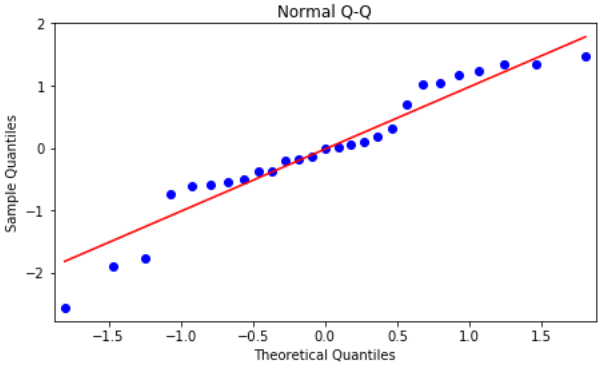



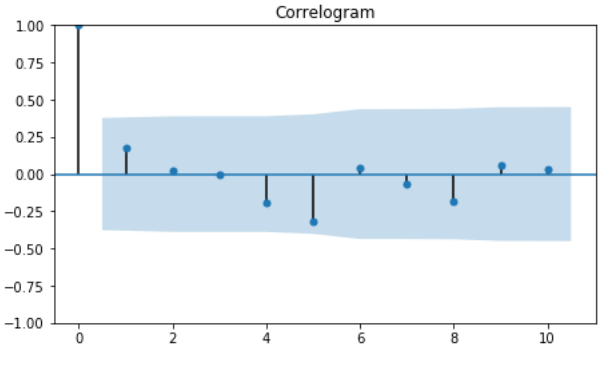
Appendix 6 – Segment 3



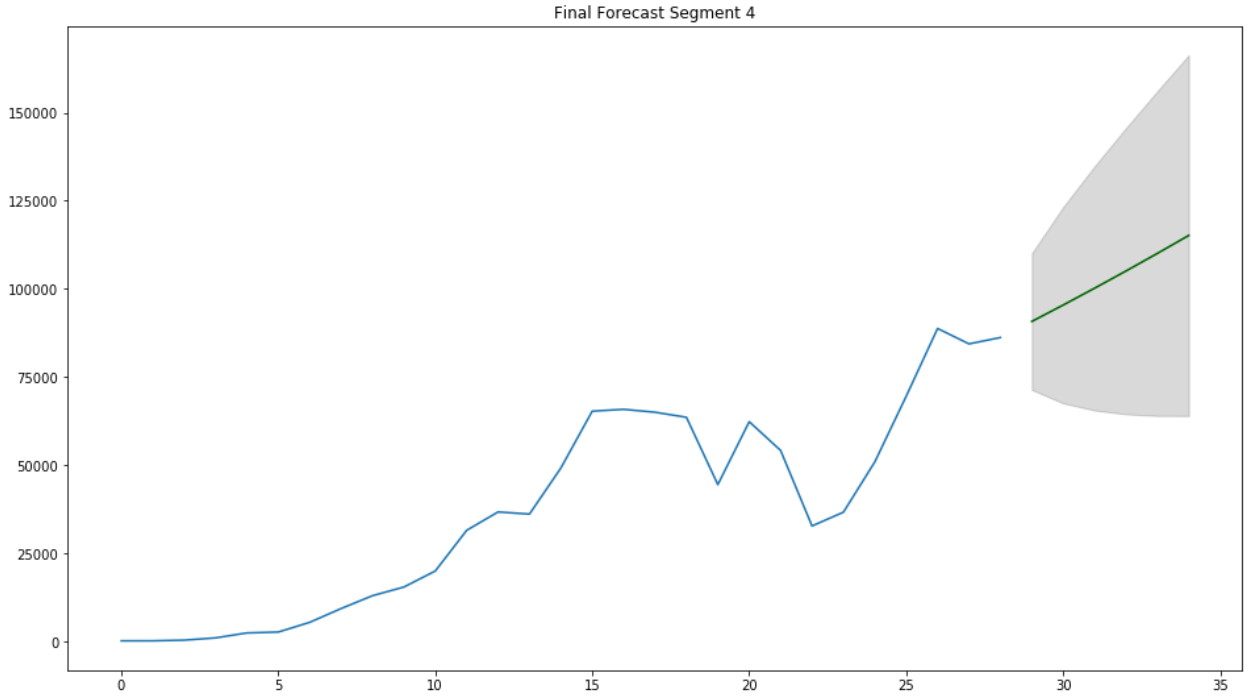



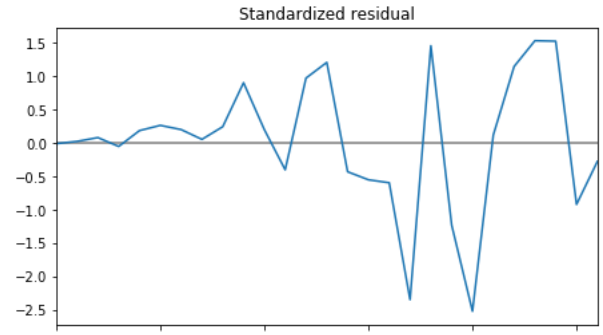



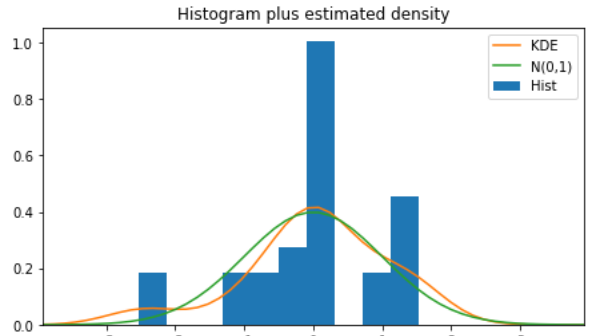



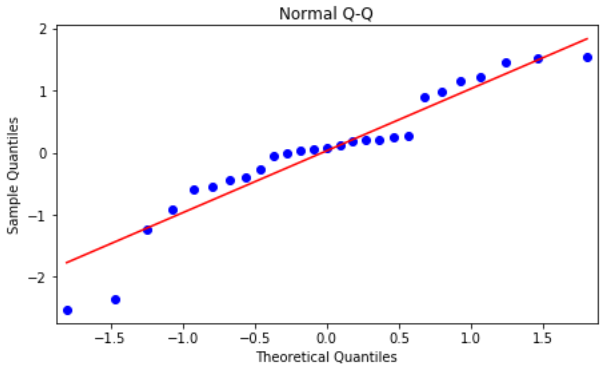



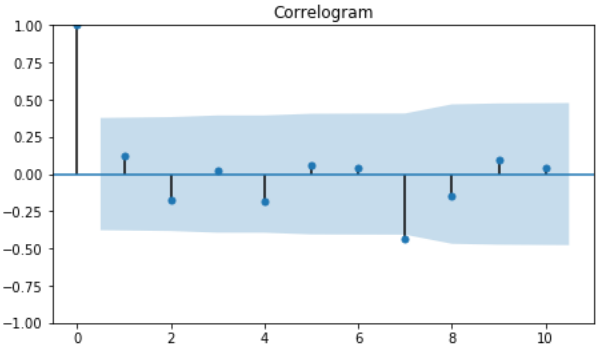
Appendix 6– Segment 4



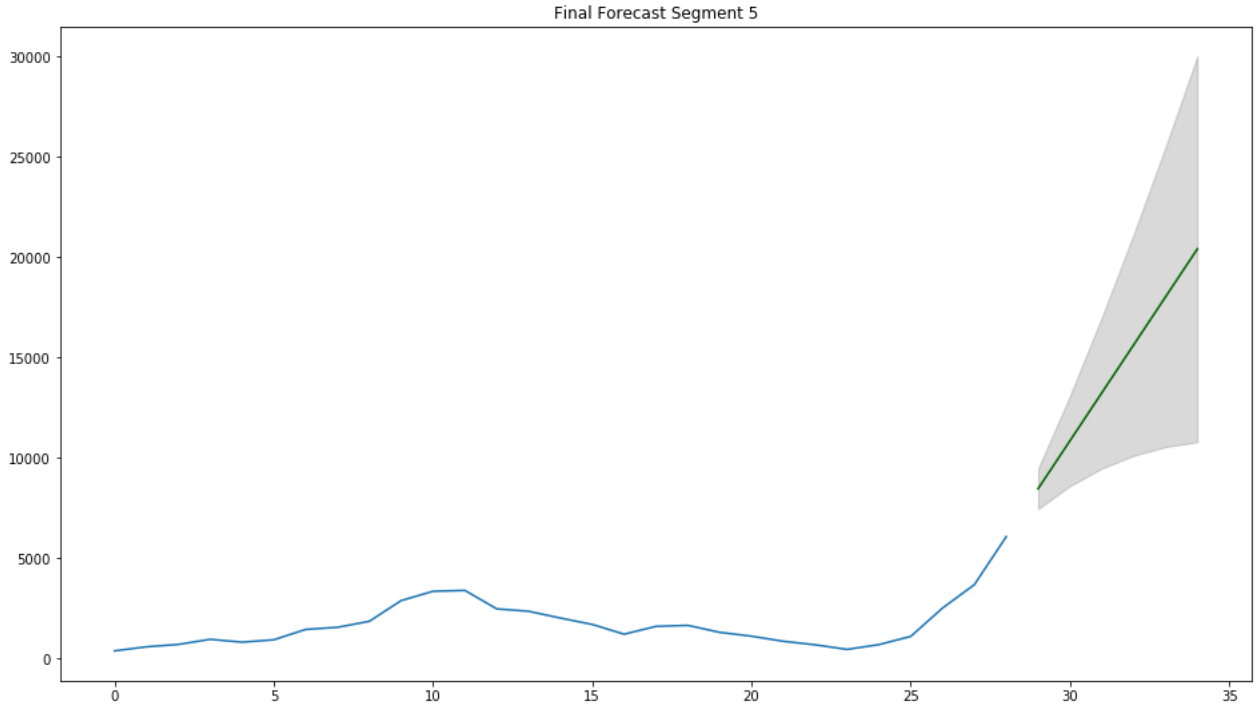



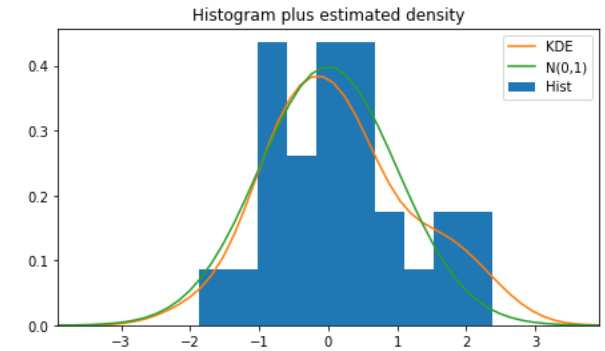



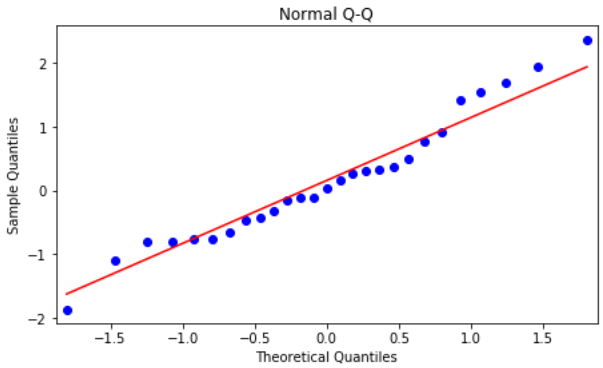



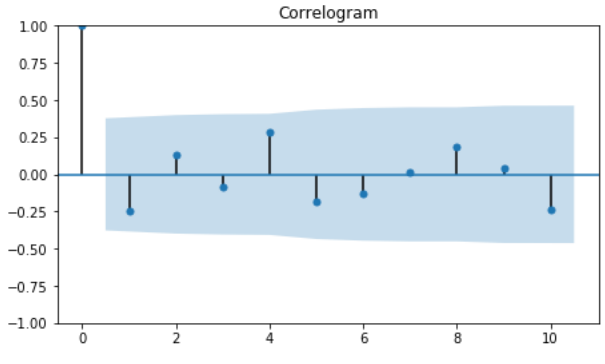
Appendix 7 – Segment 5



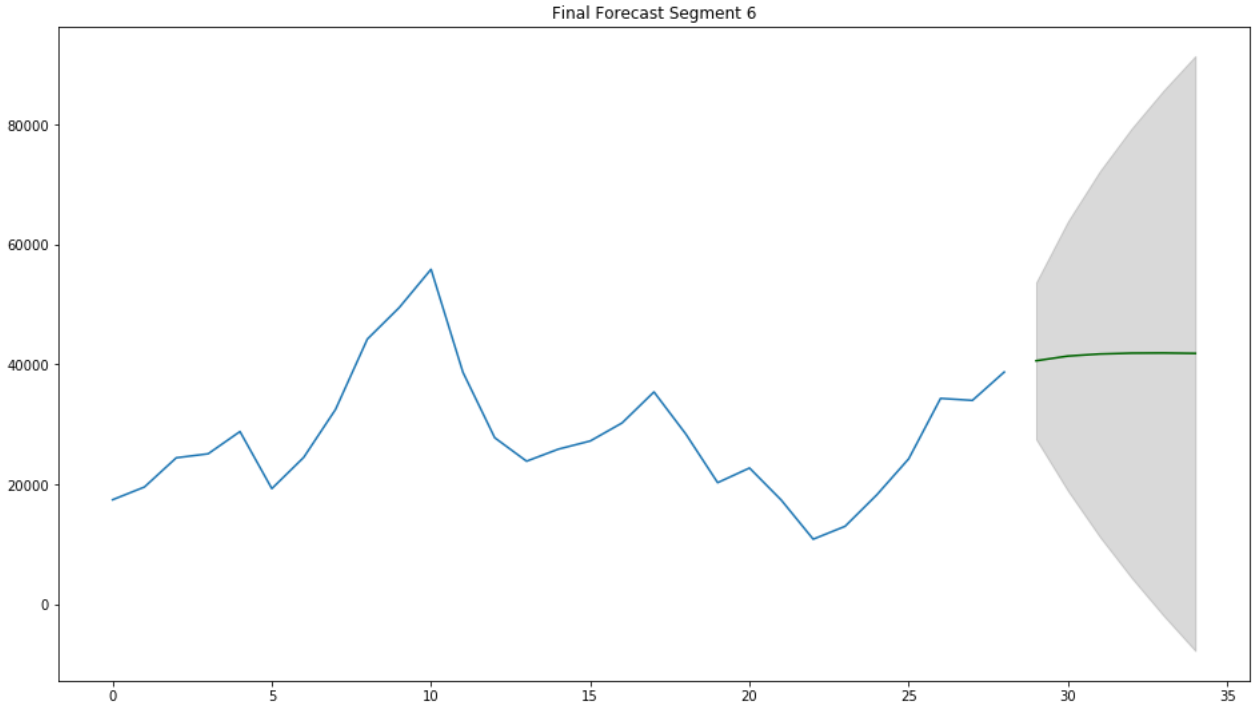


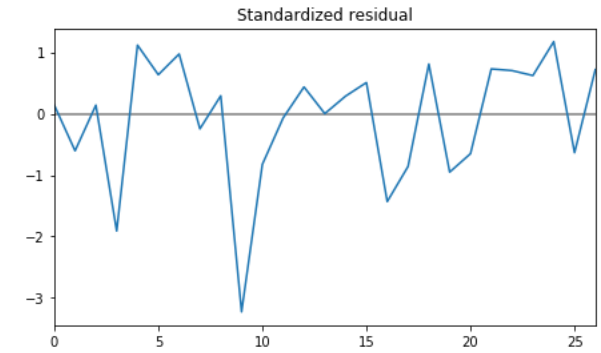


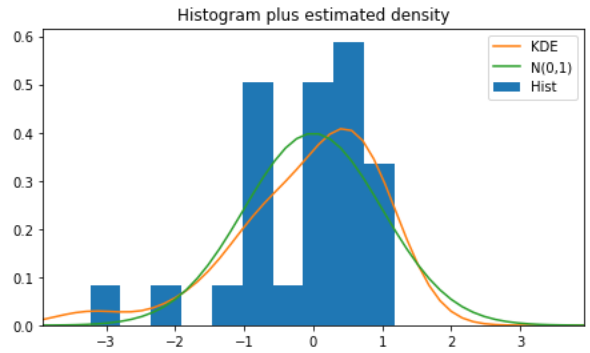



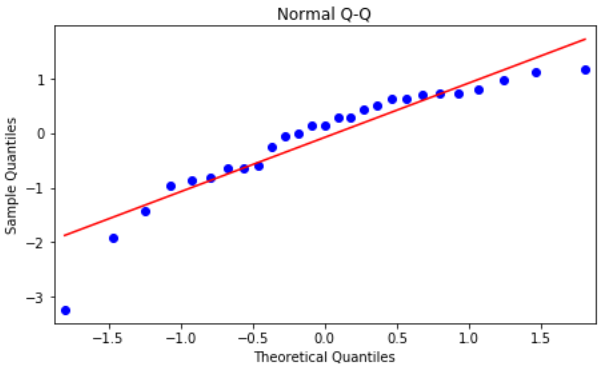



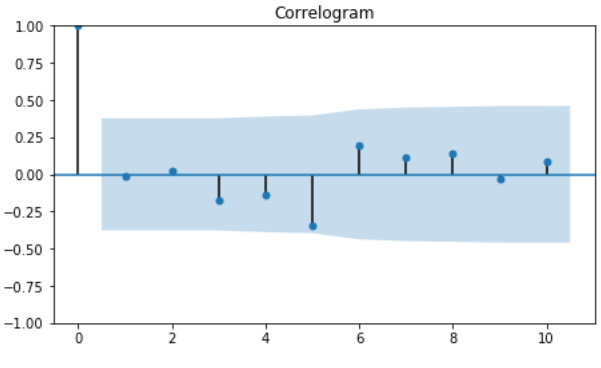
Appendix 8 – Segment 6



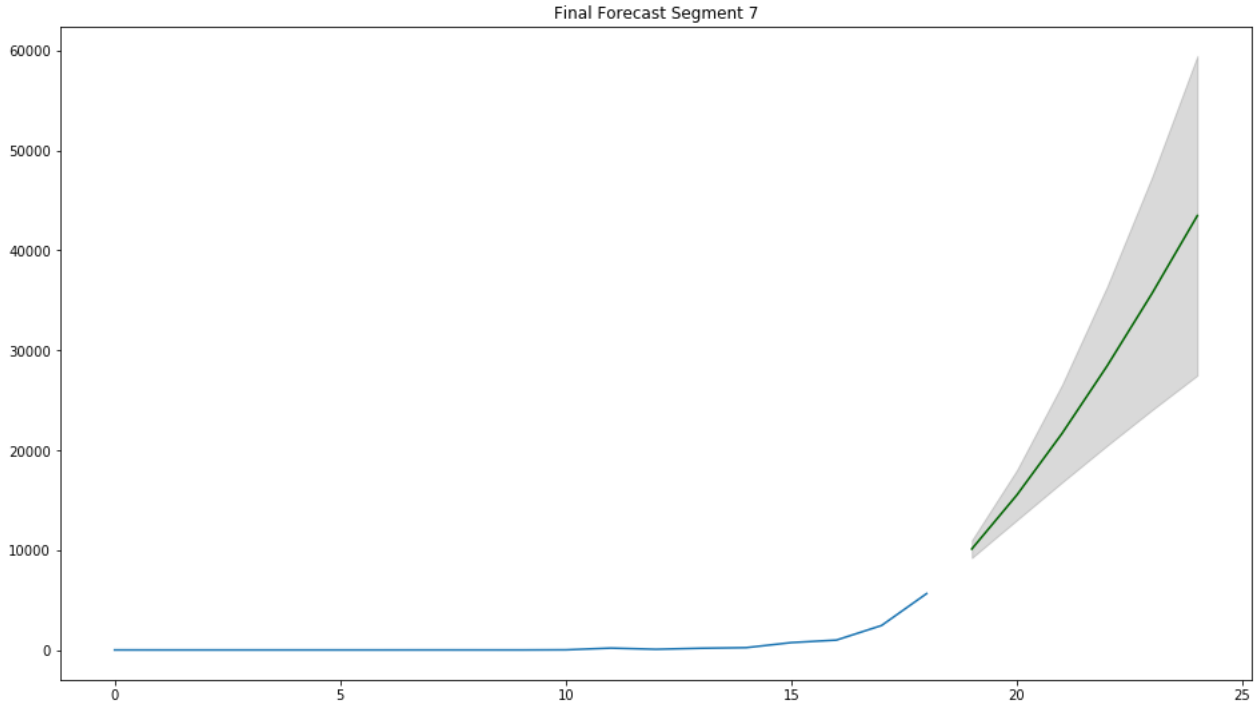



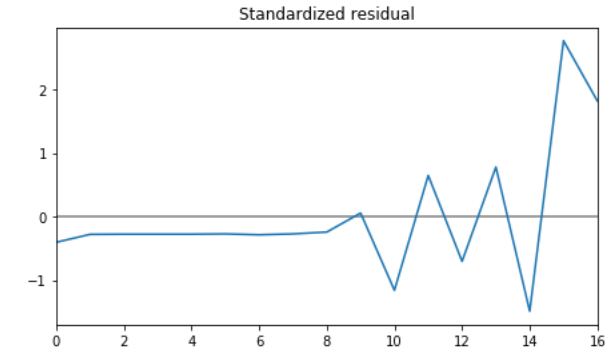



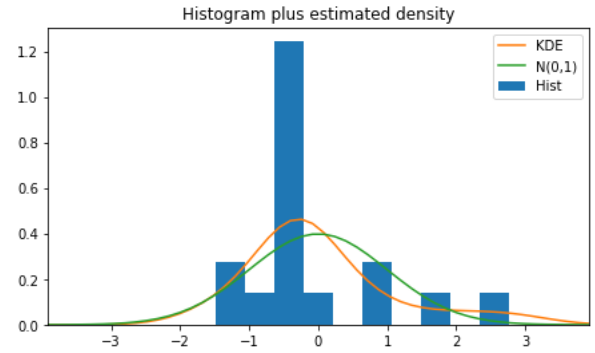



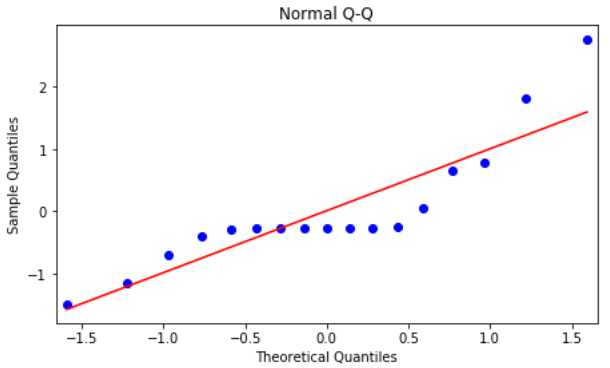



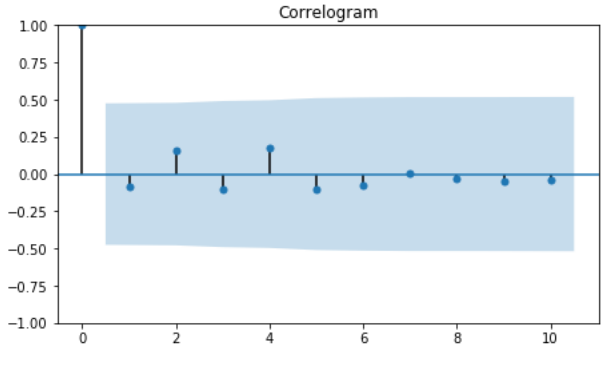
Appendix 9 – Segment 7



Appendix 10– Gasoline Average Price


**Data Cleaning**

The first step we have done was to check for duplicate values and any missing values in the data frame. The dataset did not contain duplicates, however, there were many missing values (NaN) [Appendix 1 Missing Values]. The term NaN stands for Not A Number and is a common missing data representation. It is a special floating-point value and cannot be converted to any other type than float (Russo, s.d.). The absence of values within the Database was mainly due both to the impossibility on the part of the extraction method used (Web Scraping) to always be able to save all the required data relating to maintenance costs and to the fact that some of these values weren't available in the databases. To overcome the problem of missing values different methods can be used, one of these is to replace the absent values with the average of all observations. However, since the objective of the analysis was to predict future trends, this method could have significantly altered the results, for this reason, it was decided to remove the cars of which there were no values.

Once this procedure was performed, the Database was mostly clean and consistent, however, for the purposes of subsequent analysis, it was necessary to make further changes to the features. After carrying out a descriptive analysis of the data, we find out that with the exception of some numerical variables, most of them were in the object format [Appendix 2: Summary of the Database]. Basically, this format is associated with all the features that contained categorical values. Since in order to carry out our first analysis we had to use a machine learning model through a python library called *sklearn* [12], which is unable to handle categorical values, we had to make further changes to the data. To be precise, all the categorical values contained in the column "Make", "Fuel_type", "Make", "Total_cost ", "Nb_doors", had to be converted from string format to integer. Initially, it was decided to assign a numerical value to each type of brand, for example: 'NISSAN': 0, 'RENAULT': 1. The problem with this type of encoding operation is that it only works when there is a natural order between the categories. However, in our case there wasn't an ordinal relationship and allowing the representation to lean on any such relationship could be damaging to solve the problem. In order to give the network more expressive power to learn a probability-like number for each possible label value, it was necessary to use a more complex encoding method: One-hot encoding. One-hot encoding is a binary style of categorizing, it works by converting every single categorical value into a sparse matrix in which each categorical value corresponds to a unique binary array. After one-hot encoding, we get a matrix with N-1 columns, and the matrix is full of zeros except for one 1 per row. After this last transformation, the data were finally accurate, consistent and ready to be analyzed. [Appendix 3 - OneHot Encoding]

## Web Scraping

[12] Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms.

Web scraping, also known as web data extraction, is the process of retrieving or "scraping" data from a website. The data is retrieved in HTML format, after which it is carefully parsed to extricate the raw data you want from the noise surrounding it, then the data is stored in the format and to the exact specifications of the project (What is web scraping?, s.d.)

## ACF and PACF

PACF is a partial autocorrelation function; it finds the correlation of the remains residuals with the next lag value hence 'partial' and not 'complete' as we remove already found variations before we find the next correlation. ACF is an auto-correlation function which gives us values of auto-correlation of any series with its lagged values. (Salvi, Significance of ACF and PACF Plots In Time Series Analysis)