DEPARTMENT OF
MECHANICAL AND INDUSTRIAL ENGINEERING

Manuel Romão dos Santos Gomes

BSc in Industrial Engineering and Management

Box-Jenkin’s Methodology in Python for Stock Managing

INTEGRATED MASTER IN INDUSTRIAL ENGINEERING AND MANAGEMENT

NOVA University Lisbon

March, 2022

DEPARTMENT OF
MECHANICAL AND INDUSTRIAL ENGINEERING

Box-Jenkin’s Methodology in Python for Stock Managing

MANUEL ROMÃO DOS SANTOS GOMES

BSc in Industrial Engineering and Management

Adviser: Ana Sofia Leonardo Vilela de Matos
Associate Professor, NOVA University Lisbon

Examination Committee:

Chair: Prof. Aneesh Zutshi,
Assistant Professor, FCT-NOVA

Rapporteurs: Prof. Ana Paula Ferreira Barroso,
Assistant Professor, FCT NOVA

Adviser: Prof. Ana Sofia Vilela Leonardo de Matos,
Associate Professor, FCT NOVA

Integrated Master in Industrial Engineering and Management

NOVA University Lisbon
March, 2022

Box-Jenkin’s Methodology in Python for Stock Managing

ACKNOWLEDGMENTS

I would like to thank all my professors, with a special thanks to Professor Ana Sofia and Professor Ana Paula Barroso for putting up with me not only during classes, but also for guiding me through this work.

I would also like to thank all my family for supporting me through this step of my life (fortunately a step that is about to be closed), specially to my grandparents, that raised and helped me as I was his son (and had to be very patient most of the times), my mother (and my aunt as well, I suppose…) for not letting me not finish the course and start learning and working based on Google (I am an autodidact, but it is always good to have a superior degree), my brothers and all of the rest of the family.

"On the turning away; from the pale and the downtrodden; And the words they say; Which we won't understand; Don't just accept that what's happening; Is just a case of other's suffering; Or you'll find that you're joining in; The turning away(...); No more turning away; From the weak and the weary; (...); From the coldness inside; Just a world that we all must share; It's not enough just to stand and stare; (...); No more turning away?" (Pink Floyd).

ABSTRACT

At the end of 2019, the world had shaken when social media communicated that a potential worldwide pandemic might be beginning. Early in 2020, most countries worldwide affected by the pandemic declared a state of emergency, announcing that people could not leave their houses.

When confronted with these security policies, many companies faced new management challenges regarding physical and technological resources.

Companies had to adapt their work style, allowing its employees to work remotely (some companies even adopted a hybrid work when the restrictions ended/ where on a break), on the other they had to adapt its technological resources for the information to be accessible for every employer with safety. For this purpose, large companies had to spend thousands or millions quickly adapting its information systems – both for acquiring more potent virtual private network and improve their capacity in terms of the online channel integration and invoice systems – as this was the only available channel to buy non-essential goods.

This thesis addressed the possibility of using Machine Learning (ML) to build a predictive model to forecast which will be the sales behavior over time, by analysing a time series.

That possibility consists in building a model for stock managing, that would be updated daily (with the sales till the previous day), and automatically predicts the future sales behavior, allowing an automated stock management process – not only without shortages but also without overstocking.

As a result, it was achieved a fully automated ML model, using a S3 bucket (from amazon web services) connected to a Databricks instance (launched through the S3 bucket), that has the capacity to receive the sales daily, treat the data and forecast the future data points of this sales time series.

Keywords: Time Series, Machine Learning, Predictive Models, Box-Jenkin’s Methodology, Stock Managing.

RESUMO

No final do ano de 2019 o mundo tremeu quando a comunicação social comunicou o possível começo duma pandemia mundial. No início do ano de 2020, a maior parte dos Países do Mundo, afetados pela pandemia, decretaram estados de calamidade e ordenaram que as suas populações não pudessem sair de casa.

Com estas medidas para contenção da pandemia, as empresas enfrentaram novos desafios em termos da sua gestão – tanto em termos da gestão dos recursos físicos, como tecnológicos.

Se por um lado as empresas tiveram que adaptar o seu modo de trabalho de modo que os seus colaboradores pudessem trabalhar remotamente (levando a que algumas adotassem mesmo o trabalho híbrido no pós-pandemia), por outro tiveram que adaptar os seus recursos tecnológicos para que a informação estivesse acessível a todos os trabalhadores, com segurança. Neste âmbito, grandes empresas tiveram que gastar milhares ou milhões na rápida adaptação dos seus sistemas de informação – tanto para terem *network* privada virtual mais potente, como para aumentarem a capacidade dos seus sistemas de integração e faturação do canal de vendas online – pois este era o único meio de venda possível para bens não essenciais.

Deste modo foi abordada a possibilidade de, através de um modelo de *Machine Learning*, contruir um modelo preditivo que analise o comportamento das vendas ao longo do tempo, analisando uma série temporal.

Essa possibilidade passa por desenvolver um modelo de gestão de stocks que seria atualizado todos os dias (com as vendas até ao dia anterior), e automaticamente prever o comportamento futuro, permitindo assim que haja uma gestão automatizada de stocks – sem ruturas e encomendas maiores do que o previsto.

Como resultado, foi desenvolvido um modelo de ML totalmente automatizado, tendo sido utilizado um *S3 bucket* (serviço da *Amazon Web Services*) conectado a uma instância de

Databricks, com a capacidade de ingerir dados diariamente, fazer o seu tratamento e prever as futuras vendas.

Palavas chave: Séries Temporais, Machine Learning, Modelos Preditivos, Metodologia de Box-Jenkin's, Gestão de Stocks.

CONTENTS

LIST OF FIGURES

ACRONYMS

ACF	Autocorrelation Function
AIC	Akaike's Information Criterion
API	Application Programming Interface
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
AWS	Amazon Web Services
BIC	Bayesian Information Criterion
CI	Confidence Intervals
d	Differencing Order
D	Seasonal Differencing order
EDA	Exploratory Data Analysis
ETL	Extract, Transform and Load
I	White Noise
IQR	Inter-Quartile Range
KPI	Key Performance Indicator
L	Lag Function
len	Length

LTA	Logical Transfer Architecture
MA	Moving Average
p	Autoregressive model order
P	Seasonal Autoregressive model order
PACF	Partial Autocorrelation Function
Prob (JQ)	Probability of Jarque-Bera
Prob (Q)	Probability of Ljung-Box
q	Moving Average model order
Q	Seasonal Moving Average model order
S	Seasonal Period/ Seasonality
SFTP	Secure file transfer protocol
sqrt	Square root
T	Trend

Symbols

ε	Random Error (white noise)
ω	Regression Model Parameters
μ	Mean
θ	Moving Average Model Parameters
H_0	Null Hypothesis
$y_{(t)}$	Time Series

1

INTRODUCTION

1.1 Context and Motivation

Information is the most nonviolent weapon globally, as evidenced by millions of dollars spent on intelligence by governments worldwide every year. Come to think about it, even any kind of research would be possible if there was no way to fetch, store, treat and visualize the retrieved information.

Also, companies worldwide increasingly prize data engineers, responsible for retrieving data from the systems (nowadays mostly cloud), and data scientists, responsible for treating data (this includes artificial intelligence) or data analysts, responsible for making the data available to all the stakeholders.

One known expected capability/ skill of an Industrial Engineer is inventory management. This skill can be used in a production environment, where the engineer decides the production capacity, or in a supply chain where the engineer decides the quantity to buy from a supplier. In both cases, this process is sensible, as the goal is not to lose sales or spend money on warehousing products that will not be sold in that period.

As an end-to-end developer, I establish the connections to the servers/ databases or cloud services, retrieve and treat the data, sometimes insert artificial intelligence and, finally, build dashboards with all the relevant Key Performance Indicators and information (with the best-suited information granularity).

The impact this kind of work has on an organization or company might be relevant – from raw data to dashboards with all relevant information and Key Performance Indicators (KPIs), and if needed even artificial intelligence.

This technique intends to project with precision the future sales, allowing a better stock management – that can have direct impacts on the economic results. For that reason, it is a priority concern for the business or company [1].

Even though a pandemic outbreak has not been considered probable for a long time, it has been known for more than 15 years to be a significant threat to businesses worldwide. For this reason, the utilization and implementation of virtual meetings and project management technologies for the last 15 years (or more) and the COVID-19 outbreak did not cause too many problems in the digital transformation for many companies [2].

There was, however, a demand for companies to adapt their labor regime in order to enable remote work and their technological resources to make information accessible while guaranteeing safety, which led to an increase spending on adapting their information systems – by acquiring more potent virtual private network (VPNs) and by improving their capacity in terms of online channel integration and invoice systems.

1.2 Problems and Objectives

This thesis addresses the possibility of developing an ARMA or ARIMA model that best fits the historical data and better forecasts the values of a time series, adopting the Box-Jenkin's methodology.

Artificial intelligence can be subset into distinct techniques, being the one used in this study Machine Learning, which applies algorithms of mathematics, statistics, optimization and knowledge discovery to extract patterns from the data [3].

The purpose of this thesis is to develop an ARMA or ARIMA model that best fits the historical data and better predicts the future on a time series, adopting the Box-Jenkins methodology.

It will be showed possible ways of importing information daily on a data lake, processing the information and forecasting daily. Data lakes can be seen as yesterday's unified storage [4].

According to this methodology, the first step is the model identification to test if the series is stationary and what kinds of differences will make it stationary (if it is not). If the series is not stationary, each data point will have distinct parameters, which may lead to more parameters than data points, making it impossible to work. The second step is the coefficient estimation to find the best parameters (autoregressive and moving average) that fit the data. The third and last step is the model diagnostics, which will give visibility to the model's performance. If the model performs well (by analysing the 4 model and residual diagnostic plots and both probability of Jarque-Bera and Ljung-Box values), the same can go to production (this was done with a S3 bucket as storage platform and a Databricks instance created with Amazon Web Services). Otherwise, identification, estimation and model and residual diagnostics must be repeated until the performance is not rejected.

Model identification is achieved by performing an Augmented Dickey-Fuller test for stationarity and finding the best differencing to turn it stationary and statistically significant.

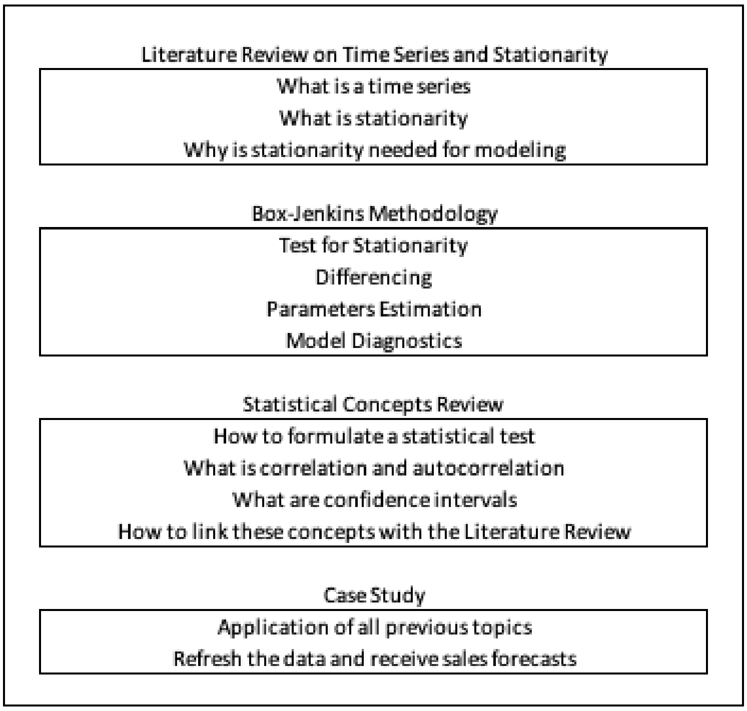
The second step, estimation and model checking, is finding the optimal model orders. This step can be achieved either by the visualization of the autocorrelation and partial autocorrelation plots or by calculating the Bayesian Information Criteria (BIC) and the Akaike's Information Criteria (AIC). This step also includes analysing the model's residuals – that must be uncorrelated and normally distributed.

If the second step is completed, the assumption is that the model will respond well to new data points that can be imported every day (or even every hour) and make predictions with the new data points. This is the third step of this methodology.

1.3 Approach and Contributions

This scientific research was thought as a way of linking technology and stock management. To elaborate this document, the first step was importing the data in a Jupyter Notebook. Using Python as the programing language.

Regarding the approach, there were developed innumerous models till the results passed all the requirements. The first model was developed with the original data points. As the residuals were not stationary and/ or uncorrelated, the development of this research included a back and forth of data transformations till the wanted output was achieved.



Literature Review on Time Series and Stationarity

What is a time series
What is stationarity
Why is stationarity needed for modeling

Box-Jenkins Methodology

Test for Stationarity
Differencing
Parameters Estimation
Model Diagnostics

Statistical Concepts Review

How to formulate a statistical test
What is correlation and autocorrelation
What are confidence intervals
How to link these concepts with the Literature Review

Case Study

Application of all previous topics
Refresh the data and receive sales forecasts

Figure 1.1 - Research and Development Approach

1.4 Document Organization

This thesis is divided in 5 chapters. First chapter introduces the context and motivation to develop this document, enounces the problems and objectives and describes the adopted approach and contributions.

The second and third chapters includes the literature review on time series, such as how it be decomposed, Box-Jenkin's Methodology, like the steps that need to be followed to implement this methodology, and some statistical tests needed for validations regarding Box-Jenkin's methodology for time series and predictive models.

Chapter four uses Box-Jenkin's methodology, aligned with the concepts explained in the third chapter, to develop and apply a machine learning predictive model to forecast sales, used for inventory managing.

In chapter five it is possible to find a summary of the analyses that was done, as well as the document conclusions. Also, conclusion includes further developments of the presented study.

2

TIME SERIES

Time series are generally defined as data points which are equally distanced in time. A time series is data collected and ordered through some time, where often the index (or independent variable) is the timestamp, and the dependent variable (that is aimed to be forecasted) is the data points [5]. The inputted data for this research has a frequency of one day.

In other words, a time series can thus be defined as a dataset collected over equal spaced time periods, ordered by date (ascendingly). Time series can be represented as a stochastic model, which can be decomposed into 3 main components: trend, seasonality period and noise (or residual) [5], [6]. It is also important to consider cyclicity when analysing a time series.

Figure 2.1 is an example of a time series and is a subset of the data used for this thesis.

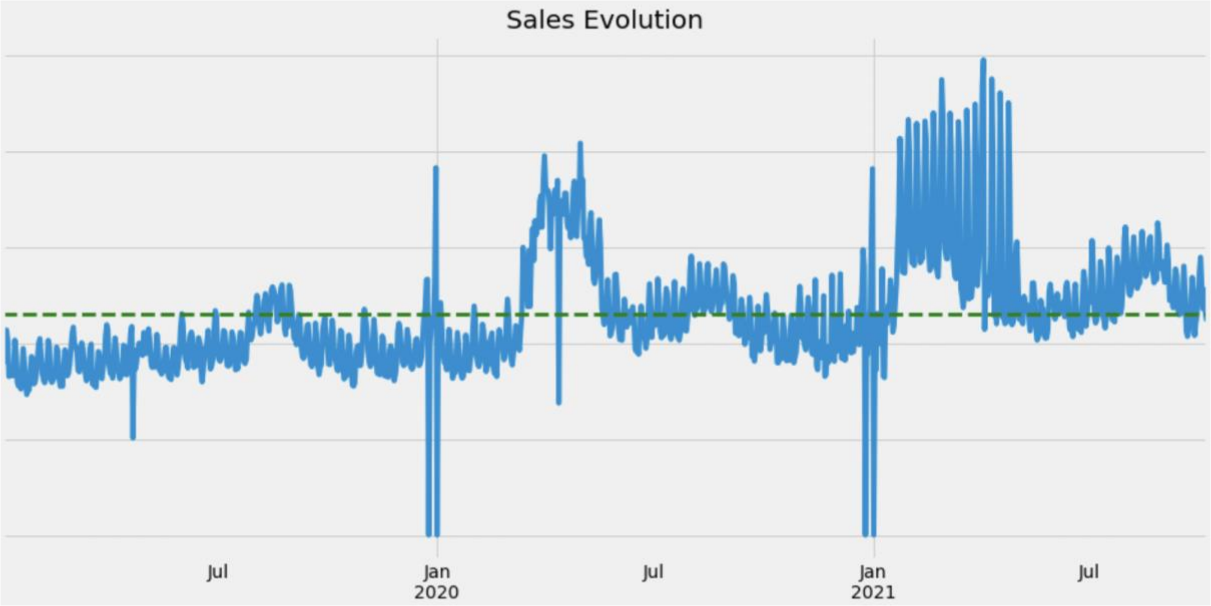


Figure 2.1 - Non-essential Product Sales Evolution

As it is possible to see in Figure 2.1, the data points are equally distributed in time, and these make an example of a time series. The independent variable, plotted on the x-axis, represents time (with the frequency of one day). The dependent variable, plotted on the y-axis, represents the quantity of non-essential consumer product sales. The green dotted line shows the median of the sum of all sales amounts (by day). This time series (as it is possible to visualize) seems to be a bit trendy with some seasonality and some noise.

Due to their complexity, these series are often challenging to work with. For this reason, and for being known for their accuracy and flexibility for distinct types of time series, Autoregressive Integrated Moving Average (ARIMA) models are widely chosen to model time series [7].

2.1 Decomposition

Decomposition is a key factor to decomplexify a time series, understating its components and better forecast its future values. A time series (y) can be decomposed into three characteristics: long term trend (T), seasonality (S) and residuals: $y_{(t)} = T_{(t)} + S_{(t)} + R_{(t)}$ or $y_{(t)} = T_{(t)} \times S_{(t)} \times R_{(t)}$ [8]. In resume, decomposing a time series into these characteristics can help understanding the best model to fit the data points.

This decomposition concept was performed using moving averages for the dataset used for this thesis and was achieved using a Python function called "seasonal_decompose" from the “statsmodels” library. The output is shown in Figure 2.3. This plot represents the decomposition the time series showed in Figure 2.1.

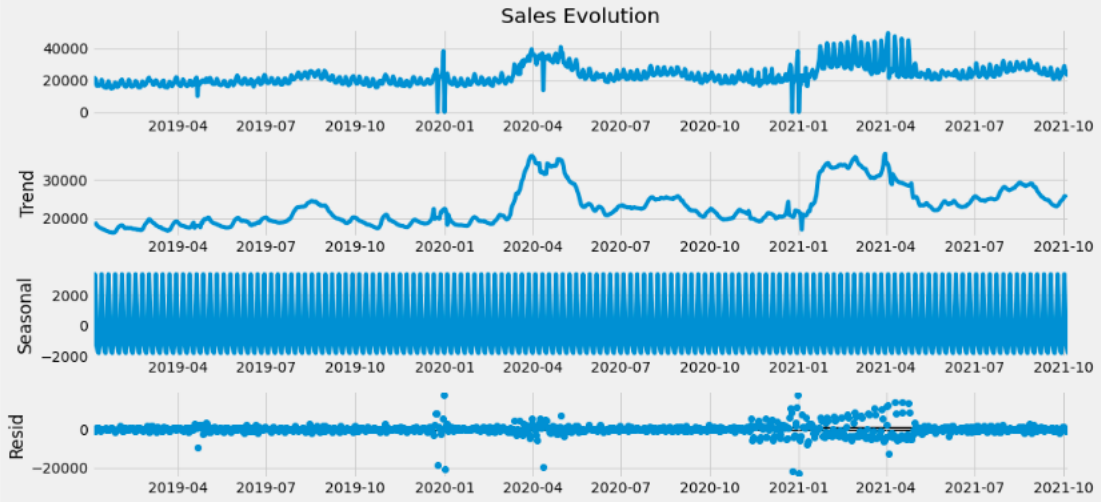


Figure 2.2 - Sales Decomposition in Trend, Seasonality and Residuals

The output of this this function (plotted in Figure 2.2) consists in 4 charts:

1. The first shows the chronologic evolution of the data points.
2. The second gives visibility on the series’ trend.
3. The third plots the seasonality. As it is possible to see, there is some severe seasonality.
4. The fourth, where the data points are not connected, display the residuals of the data points when compared to the possible data regression line (through the previously mentioned Python’s library).

2.2 Box-Jenkin’s Methodology

Box-Jenkins methodology is often chosen to find the optimal model parameters for ARIMA models. The reason is that these methodologies usually return the most accurate forecasting models for any different type of time series [9]. Box-Jenkin’s model can only be used if the time series is stationary – displays constant mean, variance and autocorrelation structure [10].

This methodology, acknowledged as a robust, computer-based iterative process that produces an autoregressive, integrated moving average model, has the capacity to adapt for seasonality and trends (that are components of a time series, as seen in section 2.1 Decomposition), transforms the data for the weight of parameters estimation, selects and tests the model and repeats the cycle as appropriate [9].

As shown in Figure 2.3, this methodology was chosen to develop a model that can both study the behavior of the past value and forecast future values. As previously mentioned, Box-Jenkin's methodology is implemented with simplicity. It can deal with robust and complex situations, adapting the model parameters to the dependent time series data through advanced mathematical and statistical processes, allowing risk and uncertainty analysis [9].

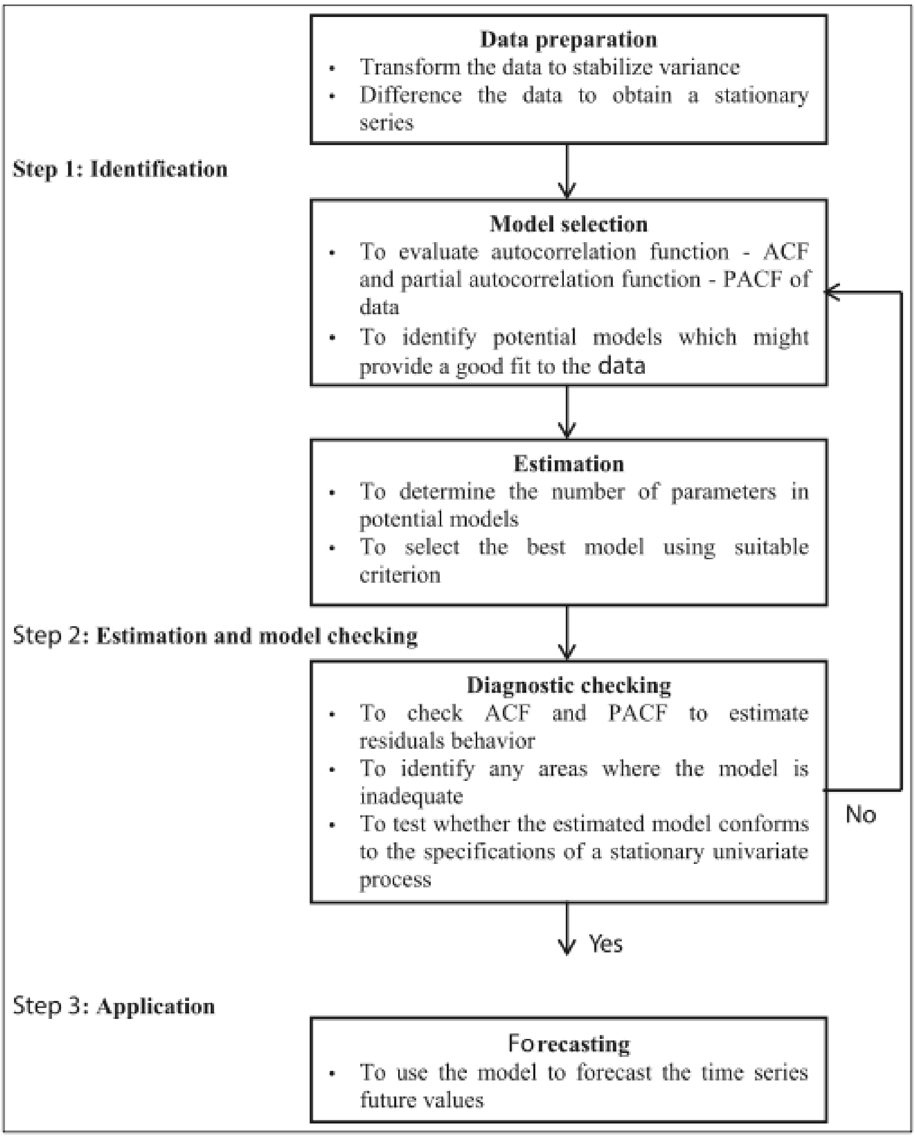

Data preparation
• Transform the data to stabilize variance
• Difference the data to obtain a stationary series
Step 1: Identification
Model selection
• To evaluate autocorrelation function - ACF and partial autocorrelation function - PACF of data
• To identify potential models which might provide a good fit to the data
Estimation
• To determine the number of parameters in potential models
• To select the best model using suitable criterion
Step 2: Estimation and model checking
Diagnostic checking
• To check ACF and PACF to estimate residuals behavior
• To identify any areas where the model is inadequate
• To test whether the estimated model conforms to the specifications of a stationary univariate process
No
Yes
Step 3: Application
Forecasting
• To use the model to forecast the time series future values

Figure 2.3 - Box-Jenkin's Methodology (Dritsakis & Klazoglou, 2018)

As Figure 2.3 shows, this methodology is composed of three steps. As with Maslow's pyramid, it is impossible to achieve the next step of the methodology if the previous assumptions steps are not validated.

The first step is Model Identification and includes:

- Data preparation: validating if the time series is stationary. If the series is not stationary, it is essential to understand what differencing order will make it stationary.
- Model Selection: identifying the model orders through the analysis of both the autocorrelation and partial autocorrelation plots. This step will allow discovering the best model to fit the data.
- Estimation: finding the best model orders for the chosen model (selected in the previous point).

The second step, coefficient estimation and model checking, is the step where the residuals of the model are analysed. They must be uncorrelated and normally distributed. If these conditions are proven incorrect, the assumption is that the model is not the good, and another model must be chosen. Otherwise, the assumption is that the model is ready for the third step, which includes:

- Check the autocorrelation function (ACF) and partial autocorrelation function (PACF): all the residuals must be uncorrelated for 95% confidence intervals.
- Identify if there are any areas where the model is inadequate.
- Test if the model is a stationary univariant process: the residuals must be uncorrelated and follow a standard Gaussian normal distribution (mean equal to zero).

The third and last step (Application) is using the model to forecast the future values of the time series.

2.3 Step 1 – Model Identification

The first step of this methodology intends to test the series values for stationarity and is divided in three sub-steps. The first (data preparation), tests the data for stationarity. If the series is not stationary, it must be transformed through differencing till the variance is stabilized [11].

This step includes (after stationarity is accomplished) relying on ACF and PACF to help identifying which model fits better the data - autoregressive (AR), moving average (MA), autoregressive integrated moving average (ARMA) or autoregressive integrated moving average (ARIMA) [11].

The third (estimation) intends to validate the number of parameters and select the best model. There are different approaches to estimating the residual behaviour, being the maximum likelihood estimation generally the preferred technique [10]. This sub-step can be validated using Akaike Information Criterion (AIC) value to verify the best fit model. The goal is to choose the model with minimum AIC values [12].

Estimation is a very important step of this methodology as non-stationary time series are not forecastable.

2.3.1 Autoregressive Models

Autoregressive (p model order) models forecast the series future values based on the series' previous values. The assumption is that the data is regressed on its own lagged values, and the prediction (y_t) can be formulated as [7]:

$$y_t = \sum_{y}^{p} w_i + y_{t-1} + b + \varepsilon_t$$

(1)

Where w_i are regression model parameters, b is a constant (bias) term and ε_t is the noise at each timestamp [7].

AR (p model order) is defined as the measure of how much a series and a lagged version of the same series vary together. The actual value of the series is equal to the sum of an autoregressive coefficient time of the previous value plus a shock term. This shock term corresponds to white noise, which is not forecastable. Statistically speaking, one can assume a series can be forecasted from its past values if there is any significant correlation (different from zero) at some own lag version – the concept of autocorrelation is precisely this. The series is correlated with a lagged version of the same series. Frequently, if a time series returns a negative autocorrelation, the same is denominated "mean reverted". On the contrary, if the series has a positive autocorrelation, it is denominated "trend-following".

When analysing the ACF and PACF, the assumption is that the AR is the best model to fit the data if the ACF tails down and the PACF cuts suddenly after p lags [10].

2.3.2 MA Models

MA (q) models, usually used to describe the series of past behavior, use residuals from the past as explanatory variables. For the model to be accepted, the assumption is that these residuals are independent and normally distributed (with zero mean), and the equation can be formulated as [7]:

$$y_t = \mu + \sum_{i}^{q} \theta \times \varepsilon_t + \varepsilon_t$$

(2)

MA (q) is calculated by regressing the values of a time series with the shock values of the past values of the same time series. The value of a time series in one period is equal to m_1 (moving average coefficient at lag one) times the previous value of the shock of the same time series plus a shock value of the current period.

Similarly, with the AR (p) models, when analysing the ACF and PACF, the assumption is that the MA is the best model to fit the data if the ACF cuts suddenly after q lags and the PACF tails down [10].

2.3.3 ARMA and ARIMA Models

ARIMA models are frequently used for time series forecasting, as these models result from the combination of autoregressive (AR), integrated (I) and moving average (MA) parts [13]. Developed by Box and Jenkins in 1970, these models combine two distinct processes: AR and MA [14].

In the ARMA (*p, q*) models, the model parameters refer to the autoregressive, differencing and moving average parts respectively.

As aforementioned, it is essential to ensure that the data is stationary to minimize seasonality, trends, or noise (it is easier to forecast the next data point if there is no change in the series' behavior). If the data is not stationary, a differencing order must be implemented to turn the series into stationary. With this transformation, the model changes from ARMA to ARIMA, which integrates a differencing (*i*) to make the series stationary [13]. In other words, ARIMA is an extension of the ARMA model that can also be applied on non-stationary data [7]. ARIMA model orders can be found by observing the ACF and PACF [13], and are the combination of AR (p) and MA (q) models.

Although ARIMA models are widely used to model time series, there are also some precautions. They are not indicated to deal with large numbers or MA or both AR and MA terms. It is preferred to use a model with the smallest set of parameters possible to minimize the probability of over-fitting [7].

2.3.4 SARIMA Models

Other Box-Jenkins models are called SARIMA (seasonal ARIMA). Considered an extension of the ARIMA models, they include a seasonal component that reflects the number of periods in each season [8].

These models are usually formulated as ARIMA (p, d, q) (P, D, Q) S. P is the seasonal autoregressive model order, D is the number of seasonal differences needed for stationarity, Q is the seasonal moving average model order and S is the seasonality period. [14].

2.3.5 Information Criteria

Another approach to finding the model orders is through the AIC and BIC.

Usually, AIC is used when the purpose of the study is to predict the time series' future values [15]. On the other hand, if the goal of the modeling is descriptive (have the perception on what features are the most meaningful to influence the output), the most commonly criterion used is the BIC [15].

The fact that as the data sample grows, the predictive accuracy improves due to subtle effects admitted to the model. This explains another difference between these two criteria: AIC will increasingly favor the inclusion of such effects, but BIC will not [15].

2.3.6 Outliers

Outliers (that sometimes can be presented as missing data points that have zero as value) can lead to models that do not perform well. These values can have impact on Pearson correlation test, sensible to bad data[16]. They can be detected using Tukey's test, which uses the quantile outlier detection method to detect the maximum, minimum, median, upper and lower quartiles of data and to draw the boxplot according to the formula: $IQR = Q_3 - Q_1$ [17].

After identifying these outliers (if there are any) using the interquartile range (IQR), it is possible to replace them with the median values [18].

Replacing these values with median values leads the model to fit the data, and to perform better [17]. So, one approach to have a good model is to replace them with the time series median.

In resume, in this thesis the outliers will be the data points (for the train data) where the absolute difference between the series values and the mean value is greater than 3 standard deviations. Also, these values will be replaced by the median of the series.

2.4 Step 2 – Coefficient Estimation and Model Checking

After finishing the first step of this methodology, as seen in Figure 2.3, the second step of this methodology is coefficient estimation and model checking.

2.4.1 Estimation

Estimation intends to find the best coefficient orders and inspect possible areas where the model is inadequate. This can be achieved by checking the ACF and PACF for residuals behaviour estimation [11].

2.4.2 Model Checking

Through model checking it is possible to check if the estimated model follows the assumptions of a stationary univariant process [11].

To test if the estimated model conforms of a stationary univariant process, the assumption is that the error term is a stationary unvaried process – the residuals should be white noise (independent and normally distributed) with constant mean and variance [10].

To study the model diagnostics, will be plotted distinct charts, and the output will consist in:

- Standardized residual over time – corresponds to the distribution of the residuals along the timeline. Should be centred around zero and not appear to have any trend.
- Normal Q-Q – another plot/ chart to check if the residuals are linearly aligned. Can be some exceptions, such as in both ends of the line.
- Histogram plus estimated density – histogram of the residuals with the theoretical line for a Normal Distribution. Can help to see if the residuals are normally distributed.
- Correlogram – a plot of the autocorrelation of the residuals at each lag. Autocorrelation must not be statistically significant for a 95% confidence interval.

It will also be used another way of diagnosing the model, through the probability of Ljung-Box (Q) and the probability of Jarque-Bera (JB). If any of these values is inferior to 5%, the null hypothesis must be rejected (residuals are uncorrelated and normally distributed). These tests will return:

- Prob (Q) – corresponds to p-value of the null hypothesis that the residuals have not any correlation.

- Prob (JB) – corresponds to the p-value of the null hypothesis that the residuals are Gaussian normally distributed.

This is the step where the model that best fits the data is chosen.

2.5 Step 3 – Application

Artificial intelligence can be subdivided into distinct techniques. One of these techniques is called machine learning, and it applies algorithms of mathematics, statistics, optimization and knowledge discovery to extract patterns from the data [3].

The 3rd (and last) step of this methodology consists in putting the model into application. This model has the capacity to import data, treat it and predict new data points.

So, model application is the same using the output model of Step 2 – Coefficient Estimation and Model Checking to forecast the series' future values [11].

There a lot of Logical Transfer Architectures (LTA) the end user can use to put the model into production (have the ETL and forecast processes automated). The ones shown in Figure 2.4 are just examples that might not be the best solution for every case (depending on the security, the number of users, if there are users outside the organization, etc.).

Figure 2.4 is a logical transfer architecture (LTA) example for the Amazon cloud. Some of the most relevant pieces for the project are the server piece, where the data will be loaded, the extract transform and load tool that will transform the data according to the model that passed step 2, the data lake, and the Power BI, that will import and allow to visualize the time series already with the forecasts calculated in the previous steps:

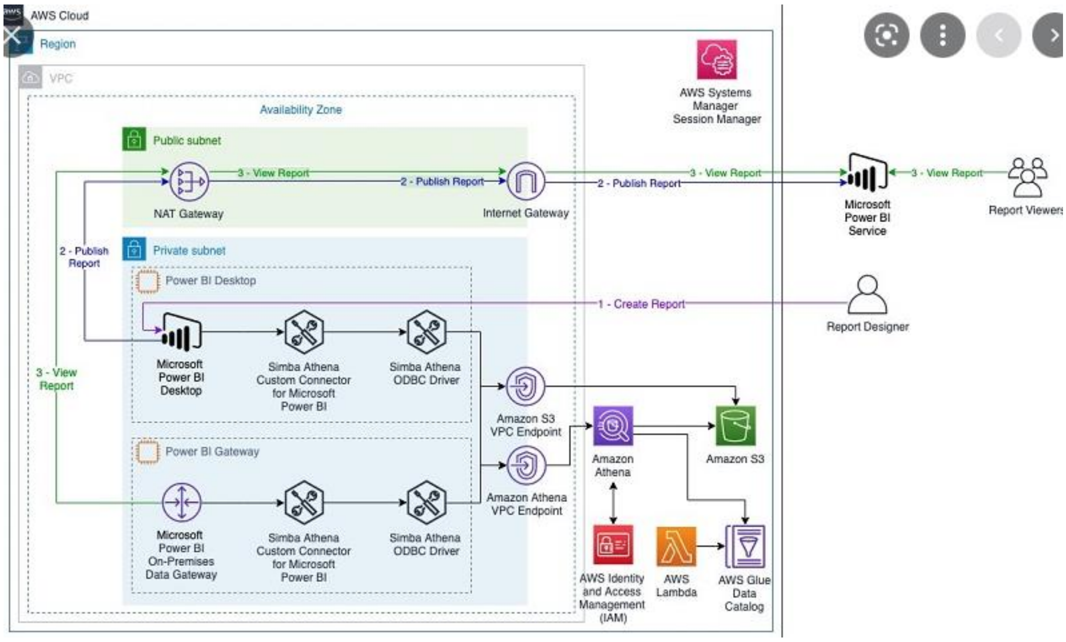


Figure 2.4 - AWS LTA Examples (Source: "https://aws.amazon.com/blogs/big-data/creating-dashboards-quickly-on-microsoft-power-bi-using-amazon-athena/", 2022-07-10)

Although this LTA has lots of pieces that are not talked about in this work, the purpose of Figure 2.4 is to show an example of the data transfer processes in the Amazon Cloud. Still, there are important services in this LTA (such as Amazon Web Services lambda), but they won’t be explained in this work.

Another LTA example is shown in Figure 2.5, this time using the Microsoft Cloud. Similarly with the Amazon Web Services example, the Microsoft Cloud (Azure) example is illustrative and might not be the best architecture for every case:

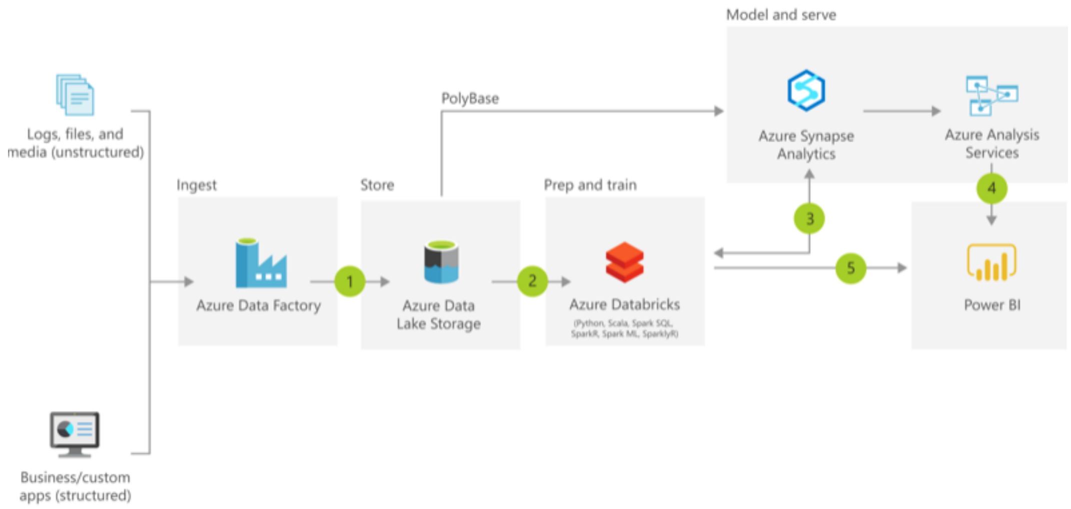


Figure 2.5 - Azure LTA Example (Source: "https://docs.microsoft.com/en-us/answers/questions/60902/what-is-the-difference-between-azure-synapsis-anal.html", 2022-07-10)

Figure 2.5 shows the data transfer process since raw data till Power BI visualizations. Once again, the purpose of Figure 2.5 is to show an example of an end-to-end data transfer process. Although this LTA has less services, there are still important services, such as Azure Synapse Analytics.

3 STATISTICAL REVIEW

Statistical tests will be used to study if there is any correlation between the time series and lagged values of the same time series. The one adopted for this research is the augmented Dickey-Fuller test, which is usually used to test for correlation between distinct time series and its null hypothesis that there is no correlation. The analyses are based on the p-value, calculated through regression surfaces [19].

This test can be done in Python using the “statsmodels.tsa.stattools.adfuller” library. The p-values are obtained through regression surface approximation [19]. If the p-value is close to significant, then the critical values should be used to judge whether to reject the null hypothesis.

When performing this test, the same will return:

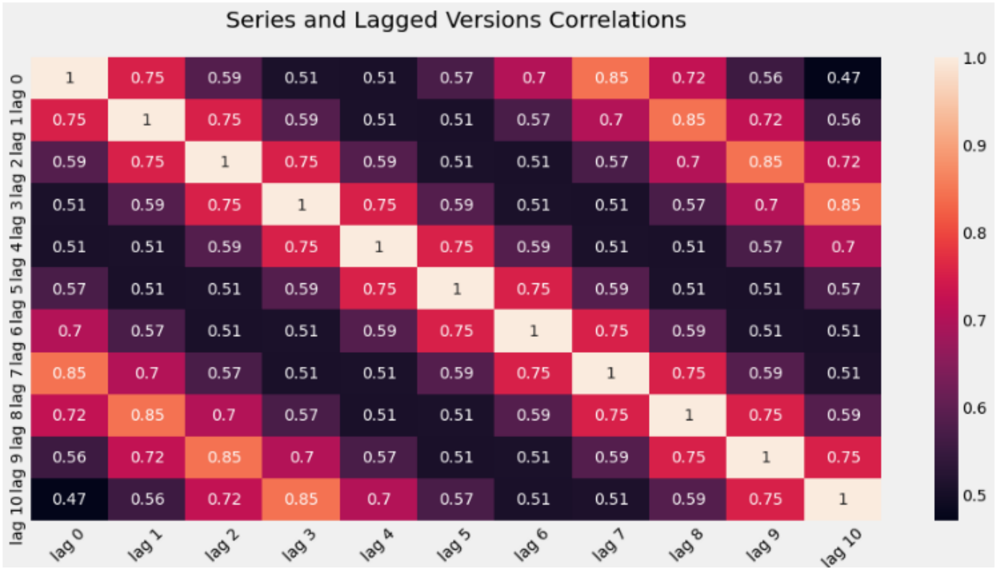
- Test statistic.
- Mackinnon’s approximate p-value based Mackinnon [19].
- The number of observations used for the ADF regression and calculation of the critical values. – this includes the critical values for {1, 5, 10}% [19].
- The maximized information criterion if auto lag is not none.

3.1 Correlation

There are different ways of validating the relationship between two time series, being one of the most common the correlation [20]. The most common correlation coefficients are Pearson’s, Spearman’s and Kendall’s. In this thesis it will be used the Pearson’s, as it has more advantages [20].

This correlation will be essential to use a lagged version of the time series to predict that same time series future values.

Figure 3.1 represents the correlation between a time series (daily frequency) and some lagged versions of the same time series:



	lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10
lag 0	1	0.75	0.59	0.51	0.51	0.57	0.7	0.85	0.72	0.56	0.47
lag 1	0.75	1	0.75	0.59	0.51	0.51	0.57	0.7	0.85	0.72	0.56
lag 2	0.59	0.75	1	0.75	0.59	0.51	0.51	0.57	0.7	0.85	0.72
lag 3	0.51	0.59	0.75	1	0.75	0.59	0.51	0.51	0.57	0.7	0.85
lag 4	0.51	0.51	0.59	0.75	1	0.75	0.59	0.51	0.51	0.57	0.7
lag 5	0.57	0.51	0.51	0.59	0.75	1	0.75	0.59	0.51	0.51	0.57
lag 6	0.7	0.57	0.51	0.51	0.59	0.75	1	0.75	0.59	0.51	0.51
lag 7	0.85	0.7	0.57	0.51	0.51	0.59	0.75	1	0.75	0.59	0.51
lag 8	0.72	0.85	0.7	0.57	0.51	0.51	0.59	0.75	1	0.75	0.59
lag 9	0.56	0.72	0.85	0.7	0.57	0.51	0.51	0.59	0.75	1	0.75
lag 10	0.47	0.56	0.72	0.85	0.7	0.57	0.51	0.51	0.59	0.75	1

Figure 3.1 - Non-Essential Consumer Good time series Autocorrelation (from lag 0 to 10)

Figure 3.1 shows that this time series' strongest correlation is with the lag equal to seven of the same time series. This can be explained by weekly seasonality.

This thesis aims to make conclusions based on the available data. It is called statistical inference – this statistical analysis is based on confidence intervals, where a distribution of results is compared to the theoretical values [21].

Recall that past values of the time series will be used for forecasting.

3.2 Confidence Intervals

Confidence intervals (CI) are defined as the confidence level in the statistical analysis results when those values are compared to the expected values [21].

For example, a 95% confidence interval is defined as a range with an upper and lower limit where the actual value is unknown (90% CIs or 99% CIs can be used, but 95% CIs are the most common). CIs can be calculated for all statistical tests (including odds ratios and percentages) [22].

Probability is related with sampling and estimating the possible/ probable values as output. If the sample was retrieved again and the 95% CI constructed (for the mean), it would mean that there was a 95% expectation that the values within the CI contain the true (and unknown) sample mean [22].

Figure 3.2 and Figure 3.3 represent the cumulative sales distribution and the sales density, respectively:

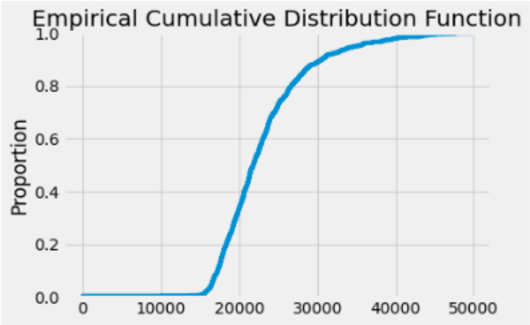


Figure 3.2 - Time series Values Empirical Cumulative Distribution Function

As Figure 3.2 shows, the probability of having the quantity of sales value bellow 15.000 per day is close to zero. On the other hand, the probability of selling more than 40.000 units per day is also near zero percent.

Other visualization that can be useful to understand the data behavior is the density plot. Figure 3.3 shows this distribution:

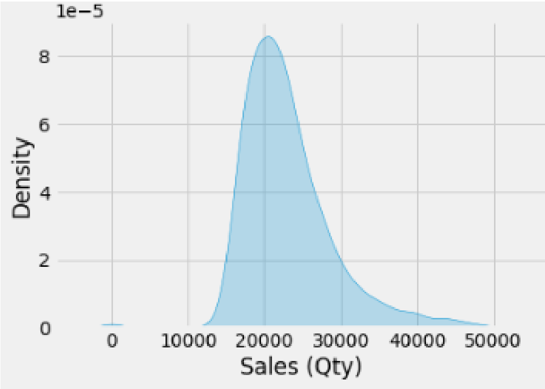


Figure 3.3 - Time series Values Density Plot

According to Figure 3.3, the most common (and expected) values of sales per day are 18.000 and 24.000 units per day.

The behavior of new data points is studied when it is possible to retrieve new data points with the same data (with replacement) to analyse if the behavior is what was statistically expected. Figure 3.4 represents the histogram of 10.000 times bootstrap replica of the time series data points with the percentile 2,5% and 97,5% (together they are the same as the 95% confidence interval).

Figure 3.4 represents a 95% confidence interval of the data points fetched if these data points were taken again (with replacement) 10.000 times.

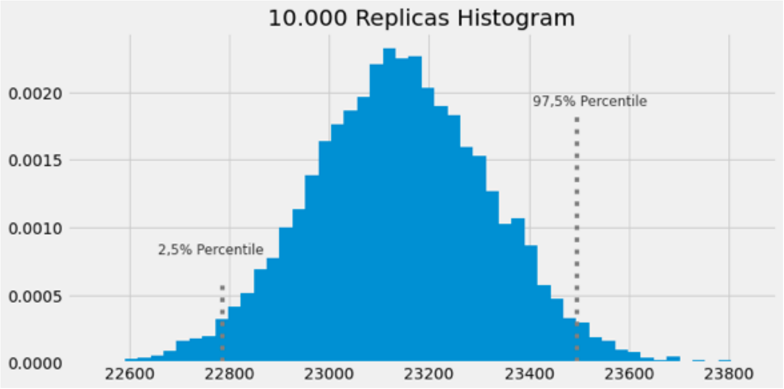


Figure 3.4 - 10.000 Replicas Sales Distribution with 95% Confidence Interval

Figure 3.4 displays the sales (quantity) distribution if the data was retrieved again (10.000 times) with replacement. As it possible to see, it can be estimated with 95% confidence that the sales values will be between +- 22.000 and +- 23.500 units per day. The dotted lines represent the {2.5, 97.5} confidence intervals.

3.3 White Noise

White noise can be thought of as a time series that does not contain information that can be used for estimation [23].

In other words, white noise can be thought as a series that has constant mean (zero) and variation (infinity) in time and zero autocorrelation at all lags. This means that the series is not forecastable, and it needs to have some modification in order to retrieve data that will allow to forecast the series' future values.

3.3.1 Augmented Dickey-Fuller test

A way to test if a series follows a random walk is the "Dickey-Fuller" test. In this case the reversion will occur on the difference of the values of the lagged series, and in this case the test checks if the slope is zero:

$$Sales_{day} - Sales_{day-1} = \alpha + \beta \, Sales_{day-1} + \epsilon_t$$

(3)

If the slop is zero, then the assumption is that the change in sales corresponds to some noise plus a mean value:

$$H_0: \beta = 0$$

(4)

If the p-value of the test is less than 5%, we can reject the null hypothesis (the series is a random walk) with 95% confidence. If the p-value is higher that 5%, we can't reject the null hypothesis.

This test was performed using a Python function called "adfuller" from the "statsmodels" library.

3.3.1.1 Stationarity

Stationarity may be defined as a probabilistic process in which both the mean and variance do not vary with time and the covariance between two periods is based only on the distance between these two periods [14].

A series is considered stationary if there are no trends, seasonality or any variance change trough the timeline. In resume, a series is stationary if it respects the following three criteria:

1. Trend must be constant.
2. Variance must be constant in time.
3. Autocorrelation must be constant.

A trend is defined as an increment (positive or negative) of the data points along the timeline.

Autocorrelation is the relation between a series and a lagged version of the same series.

Variance is defined as the wideness of the data points across time (their distance from the Linear Regression).

Random walks are examples of not stationary series. Seasonal series are also not stationary, as the mean increases and decreases over time.

This thesis will also focus on the steps to turn a non-stationary series into a stationary one (for not stationary series that cannot be modelled).

It is essential to ensure the series is stationary because each data point will have associated parameters when modelling the same. If each parameter(s) associated with a data point is different from the parameter(s) of the other data points, the model may end with more parameters than data points.

In this context, stationarity ensures that the modelling of the series does not have more parameters than it should be possible to model it.

4 CASE STUDY

In this chapter it will be boarded the data modeling of a time series. This chapter aims to use the studied methodology (Box-Jenkin's) to predict the sales in a quantity of a non-essential product during 2020 and 2021. For this research the data set (that was data between 1st of January 2019 and 7th of October 2021) will be divided into training data (year of 2019) and testing data (from January 1st on).

The methodology will also be tested to see if the output is as expected.

4.1 ARIMA

The first step that needs to be done is checking if the series is stationary. If it is not, it is necessary to define what differencing order will make the time series stationary.

After stationarity is accomplished, it is necessary to estimate the best model orders for the autoregressive and moving average parts. This will allow to fit the data into the best possible model to forecast future values.

The third and last step studies the residuals behavior. If the residuals are uncorrelated and follow a normal distribution, the assumption is that the model is good for forecasting. Otherwise, it is necessary to get back to identification and do some data transformation (such as aggregations) before looping through Box-Jenkins methodology again.

For this case study, the first approach will be a ARIMA model that will go through all the methodology steps.

4.1.1 Identification

The second step aims to of this methodology is identification. Identification is a critical step to ensure the good behavior of the data modeling. It is on this step of the analysis that it will be done some exploratory data analysis (EDA) to see if there are data points missing, if the series seems stationary, trendy, etc.

As these models are very sensitive to outliers, these values were studied and removed from the analysis. Figure 4.1 is representative of all the outliers in the all the dataset:

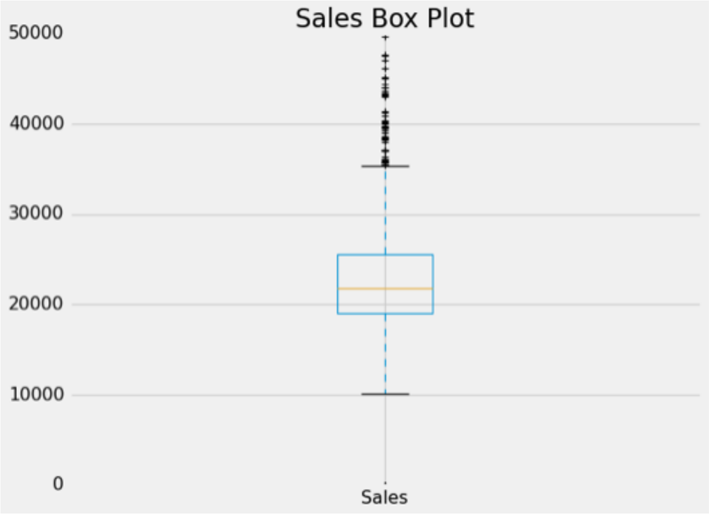


Figure 4.1 - Time Series Outliers using IQR = Q3 - Q1

As it is possible to verify through this box plot, most of the outliers correspond to values higher than the 3rd quartile plus 1.5 times the inter quartile range.

To remove the impact of these values in the model estimation, according to chapter Outliers, these values will be replaced by the series median.

As these models are very sensitive to outliers, these values were also studied. Figure 4.2 represents the evolution of the sales analysed values with no outliers (the green dotted line represents the median of sales). All the data points were subtracted with the training series mean to find these values. After this operation, if the difference's absolute value was greater than 3 times the series standard deviation, the data point is considered an outlier.

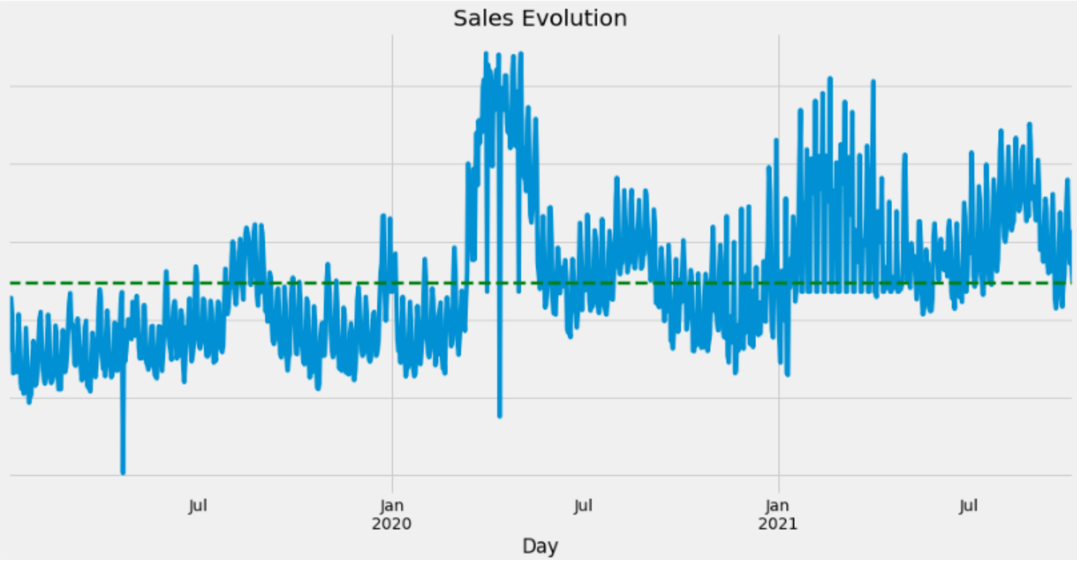


Figure 4.2 - Sales with outliers Removed and mean sales value (dotted)

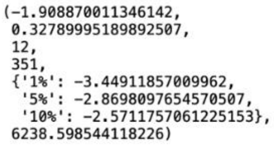
Although outliers have been removed, the time series still appears to be trendy and have some seasonality. These time series elements will be delt with later.

4.1.1.1 Stationarity

To ensure before starting to model the series, the first sub-step of estimation is that the data points are stationary – time series must be stationary. If this hypothesis can be disproven, it is necessary to find the model differencing that will make the series stationary.

The stationarity test will be performed according to the Augmented Dickey-Fuller test. Figure 4.3 has the needed results. The null hypothesis is formulated as the series being a random walk, and it can be disproven if the p-value of the ADF test is less than 5% (-2.86).

As previously mentioned on chapter 3, the first row is the test statistic, the second row is the p-value, the third row are the number of observations used for the Augmented Dickey-Fuller test, the 3 rows after are the critical for 1, 5 and 10 percent, and the last value corresponds to the maximized information criterion if the auto lag is not none:

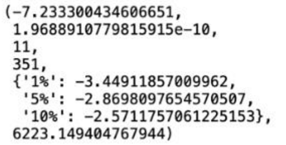


```
(-1.908870011346142,
 0.32789995189892507,
 12,
 351,
 {'1%': -3.44911857009962,
  '5%': -2.8698097654570507,
  '10%': -2.5711757061225153},
 6238.598544118226)
```

Figure 4.3 - Augmented Dickey Fuller Test for Original Time Series

The p-value is equal to -1.91, a value greater than 5% (-2.86). As this value is higher than the critical value for a 95% confidence interval, the null hypothesis (the series is a random walk - with positive trend and weekly and monthly seasonality) cannot be disproven, and the assumption is that the series is not stationary. Following the methodology, the next step is to find what differencing order will make the series stationary.

The differenced time series (the change between two consecutive observations) can be done to turn a non-stationary time series into a stationary one. For that purpose, this transformation was done and the ADF test output can be analysed in Figure 4.4 - Augmented Dickey Fuller Test for Time Series with Differencing Order = 1.



```
(-7.233300434606651,
 1.9688910779815915e-10,
 11,
 351,
 {'1%': -3.44911857009962,
  '5%': -2.8698097654570507,
  '10%': -2.5711757061225153},
 6223.149404767944)
```

Figure 4.4 - Augmented Dickey Fuller Test for Time Series with Differencing Order = 1

With the differencing equal to one transformation, the p-value is equal to 4.7e-17. The critical value for 5% confidence interval is –2.86. As the p-value is lower than the critical value, the null hypothesis is rejected, and the assumption is that the series is now stationary.

4.1.1.2 Autocorrelation and Partial Autocorrelation Plots

As seen in chapter 2.3 it is possible to find the autocorrelation and moving average orders when analysing ACF and PACF plots – done in Python through the "plot_acf" and "plot_pacf" functions from "statsmodels" library. If the ACF tails down and the PACF cuts of after p lags, the assumption is that the best model to fit the data is an autoregressive mode. If the PACF tails down and the ACF cuts down after q lags, the assumption is that the best model to fit the data is a moving average model. If neither of the functions cut down, the best model to fit the data is the combination of previous two model (ARMA or ARIMA).

The autocorrelation and partial autocorrelation of the time series with the differencing order equal to one were plotted to visualize the model order:

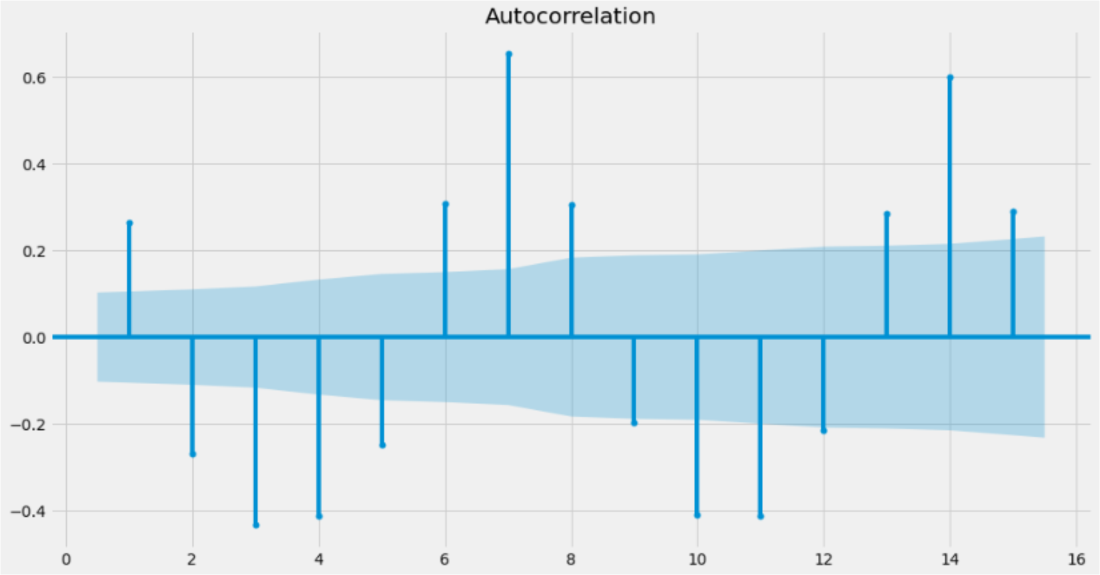


Figure 4.5 - Autocorrelation Plot with 15 Lags

As Figure 4.5, that plots the number of lags on the x-axis and autocorrelations on the y-axis suggests, the best model for this data set should be either AR or ARIMA. This deduction is based on the autocorrelation plot, that does not cut off suddenly. There are also 95 % confidence intervals bands, where the standard deviation is computed according to Bartlett's formula.

It is also necessary to plot Partial Autocorrelation of the time series. Figure 4.6 gives visibility of autocorrelation values.

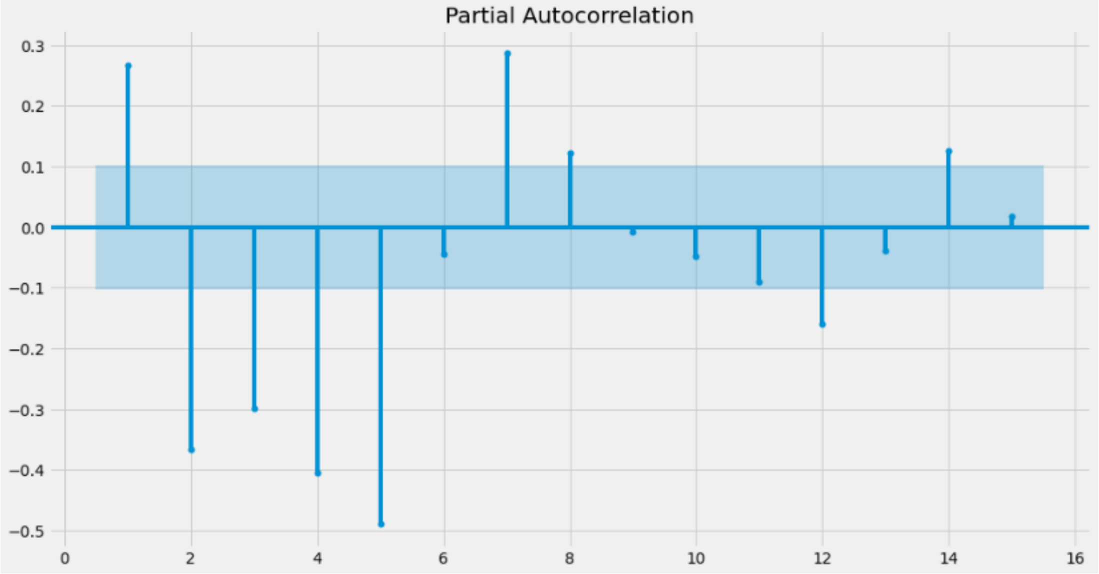


Figure 4.6 - Partial-Autocorrelation Plot with 15 Lags

Figure 4.6 plots the number of lags on the horizontal and the correlations on y-axis.

Also are 95% confidence intervals for the given level are returned according to $1/\sqrt{len(x)}$.

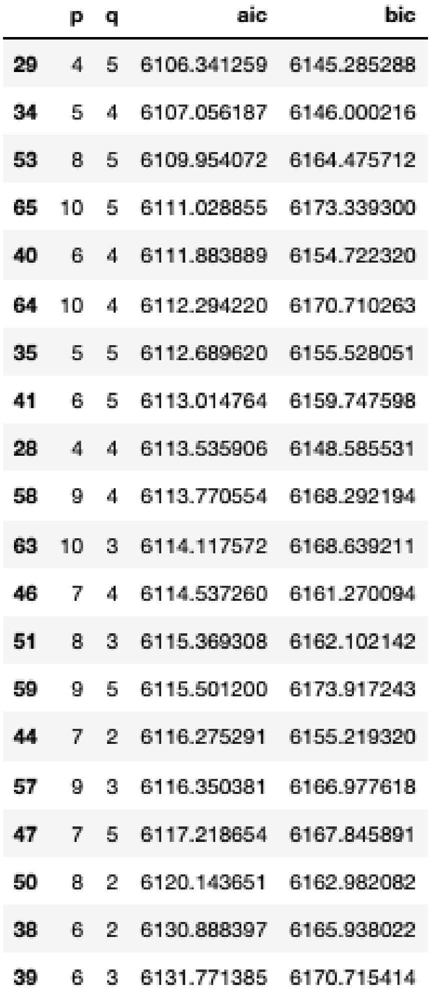
As neither of the charts cut off after all the 15 lags visualized in the plots, the assumption is that the best model to describe the data is an ARIMA model (the series is not stationary, and the differencing order is equal to one).

4.1.1.3 AIC and BIC

Deducting the model order for ARIMA models using only visualization is difficult. Since this deductive analysis is not objective, another approach was adopted to find the optimal model orders – the Akaike's information criterion and the Bayesian information criterion. As seen in chapter 2.3.5, another way to determine the model order is by analysing the AIC and BIC values. To accomplish this task, in Python, a for loop was built. The loop's content is to return the AIC and BIC values for each AR(*p*) and MA(*q*) model order.

The minimum is zero for the autoregressive model order, and the maximum is 10. The minimum is zero for the moving average model order, and the maximum is 5. This lower maximum in the MA order is to avoid false impressions of a good model, as seen in chapter 2.3.3.

The 20 first values ordered by AIC are shown in Figure 4.7 - AIC Ascendent Order, and the 20 first values ordered by BIC are shown in Figure 4.8 - BIC Ascendent Order.

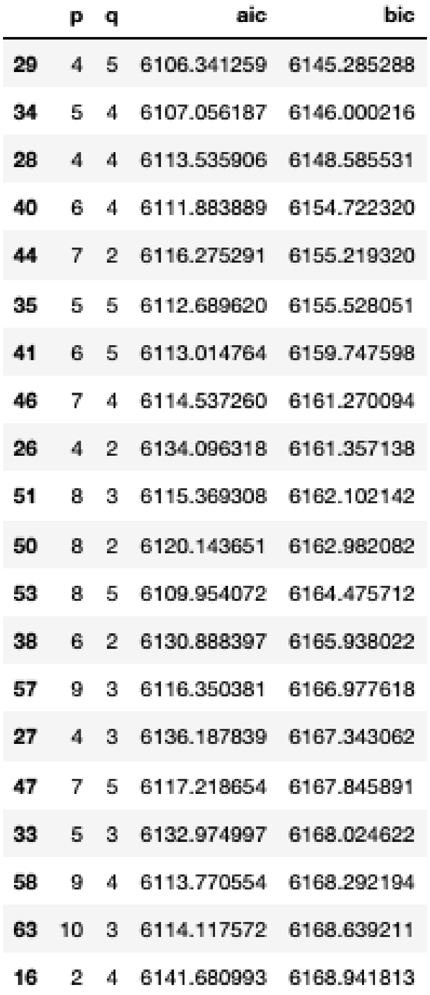


	p	q	aic	bic
29	4	5	6106.341259	6145.285288
34	5	4	6107.056187	6146.000216
53	8	5	6109.954072	6164.475712
65	10	5	6111.028855	6173.339300
40	6	4	6111.883889	6154.722320
64	10	4	6112.294220	6170.710263
35	5	5	6112.689620	6155.528051
41	6	5	6113.014764	6159.747598
28	4	4	6113.535906	6148.585531
58	9	4	6113.770554	6168.292194
63	10	3	6114.117572	6168.639211
46	7	4	6114.537260	6161.270094
51	8	3	6115.369308	6162.102142
59	9	5	6115.501200	6173.917243
44	7	2	6116.275291	6155.219320
57	9	3	6116.350381	6166.977618
47	7	5	6117.218654	6167.845891
50	8	2	6120.143651	6162.982082
38	6	2	6130.888397	6165.938022
39	6	3	6131.771385	6170.715414

Figure 4.7 - AIC Ascendent Order

As it is possible to see in Figure 4.7, the optimal model order considering the AIC is p = 4, d = 1 and q = 5. This means that, for this data, the best model orders are 4 for the autoregressive component, 1 for the differencing, and 5 for the moving average component.

The next step is to do the same analyses but ordered by BIC. After that, if the ordering is different, the model order must be chosen depending on the analyses purposed (to forecast or to study the time series behavior).



	p	q	aic	bic
29	4	5	6106.341259	6145.285288
34	5	4	6107.056187	6146.000216
28	4	4	6113.535906	6148.585531
40	6	4	6111.883889	6154.722320
44	7	2	6116.275291	6155.219320
35	5	5	6112.689620	6155.528051
41	6	5	6113.014764	6159.747598
46	7	4	6114.537260	6161.270094
26	4	2	6134.096318	6161.357138
51	8	3	6115.369308	6162.102142
50	8	2	6120.143651	6162.982082
53	8	5	6109.954072	6164.475712
38	6	2	6130.888397	6165.938022
57	9	3	6116.350381	6166.977618
27	4	3	6136.187839	6167.343062
47	7	5	6117.218654	6167.845891
33	5	3	6132.974997	6168.024622
58	9	4	6113.770554	6168.292194
63	10	3	6114.117572	6168.639211
16	2	4	6141.680993	6168.941813

Figure 4.8 - BIC Ascendent Order

Although not all results have the same order, for the first value this is true. If this didn't happen, a choice would have to be made – the model order according to the AIC or BIC order (depending on if the purpose is to predict or to study past values). In this case the assumption is that the best model order is p=4, d = 1 and q=5.

4.1.2 Estimation and Model Checking

With the model orders identified, the next step is to test them to see if they return the expected output. For this purpose, the model will be tested with new unseen data (data values for the year of 2020).

The first test that was made was the diagnostics plots, where it is possible to visualize the residuals behavior:

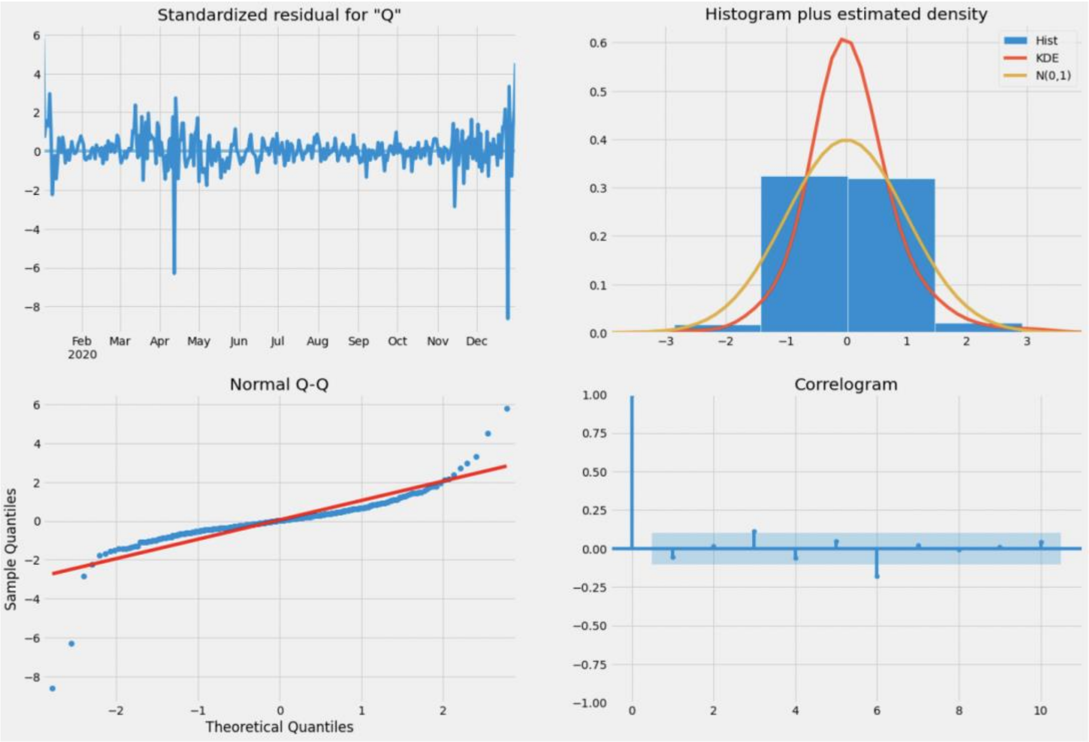
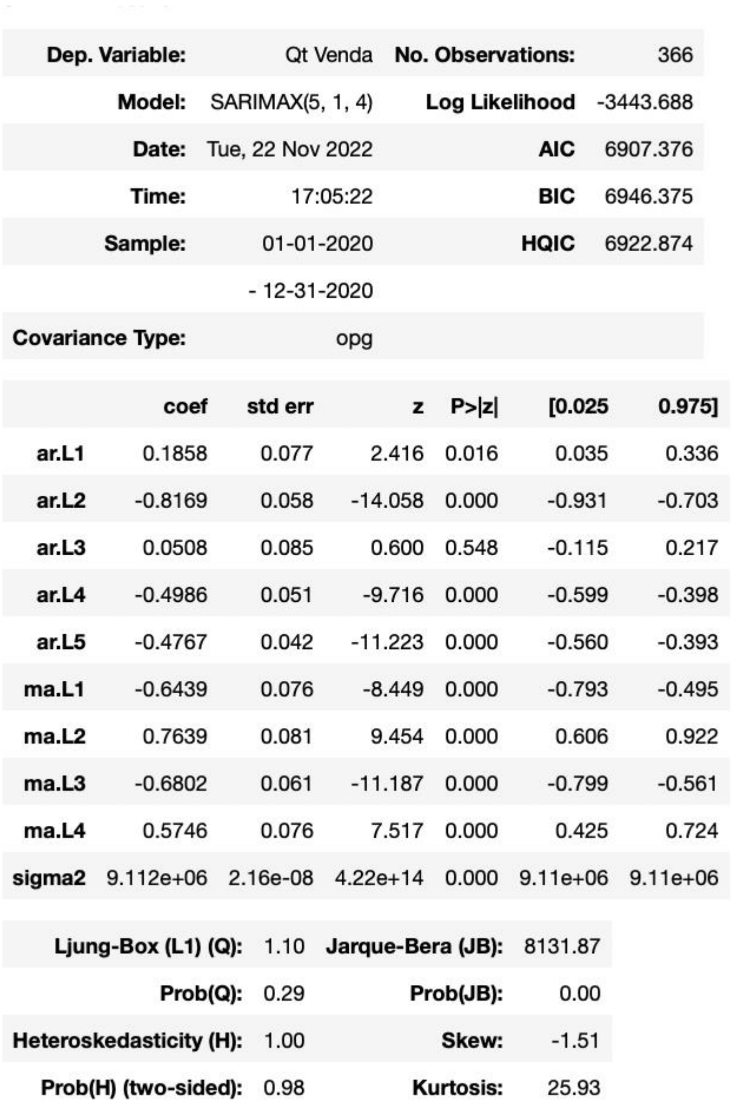


Figure 4.9 - Diagnostics Plot for ARIMA (p = 5, d = 1, q = 4)

As it is possible to visualize in Figure 4.9, the residuals do not look neither uncorrelated (in the correlogram there are lags different from zero statistically significant) nor normally distributed, as possible to see in the histogram plus estimated density. The next step, to be sure of the results, is to calculate Prob (Q) and Prob (JB). Figure 4.10 shows the output of these tests:



Dep. Variable:	Qt Venda	**No. Observations:**	366
Model:	SARIMAX(5, 1, 4)	**Log Likelihood**	-3443.688
Date:	Tue, 22 Nov 2022	**AIC**	6907.376
Time:	17:05:22	**BIC**	6946.375
Sample:	01-01-2020	**HQIC**	6922.874
	- 12-31-2020		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.1858	0.077	2.416	0.016	0.035	0.336
ar.L2	-0.8169	0.058	-14.058	0.000	-0.931	-0.703
ar.L3	0.0508	0.085	0.600	0.548	-0.115	0.217
ar.L4	-0.4986	0.051	-9.716	0.000	-0.599	-0.398
ar.L5	-0.4767	0.042	-11.223	0.000	-0.560	-0.393
ma.L1	-0.6439	0.076	-8.449	0.000	-0.793	-0.495
ma.L2	0.7639	0.081	9.454	0.000	0.606	0.922
ma.L3	-0.6802	0.061	-11.187	0.000	-0.799	-0.561
ma.L4	0.5746	0.076	7.517	0.000	0.425	0.724
sigma2	9.112e+06	2.16e-08	4.22e+14	0.000	9.11e+06	9.11e+06

Ljung-Box (L1) (Q):	1.10	**Jarque-Bera (JB):**	8131.87
Prob(Q):	0.29	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.00	**Skew:**	-1.51
Prob(H) (two-sided):	0.98	**Kurtosis:**	25.93

Figure 4.10 - Model Summary for ARIMA (p = 5, d = 1, q = 4)

Figure 4.10 is the output of the following Python ".summary" command. As it is possible to see, although the residuals are uncorrelated (Ljung-Box test statistic value is greater than 5%), they are not normally distributed, as Prob (JQ) is less than 5%. As the model checking revealed some flaws in the model, it is necessary to get back to the first step (estimation), transform the data, and loop through the methodology again.

4.2 Auto ARIMA

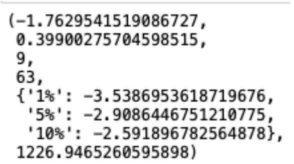
As the model didn't pass the test mentioned on section 2.4, the data must be modified so the residuals of the model are uncorrelated.

For this purpose and after (using a trial error), some changes were made in the data points:

- The data was agglomerated by the rolling 5 previous days – with this transformation each data point will contain the sum of the sales of the past 5 days.
- The data was shifted by 5 days – with this transformation each data point will contain the cumulative sum of the next 5 days sales.
- The data was modified so it shows in a 5 period (5 days) interval.

4.2.1 Identification

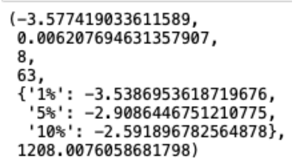
Once again, the first step is to test the data for stationarity. For this purpose and using 2019 data to train the model, the augmented Dickey-fuller test was performed (with no outliers):



```
(-1.7629541519086727,
 0.39900275704598515,
 9,
 63,
 {'1%': -3.5386953618719676,
  '5%': -2.9086446751210775,
  '10%': -2.591896782564878},
 1226.9465260595898)
```

Figure 4.11 - New dataset Augmented Dickey-Fuller test

Figure 4.11 shows that the series is not stationary (p-value is equal to -1,76, that is inferior to the 5% critical value of -2,91), so it cannot be modeled. To make it stationary, the differencing = 1 will be made. With this transformation, the new output of the test is:



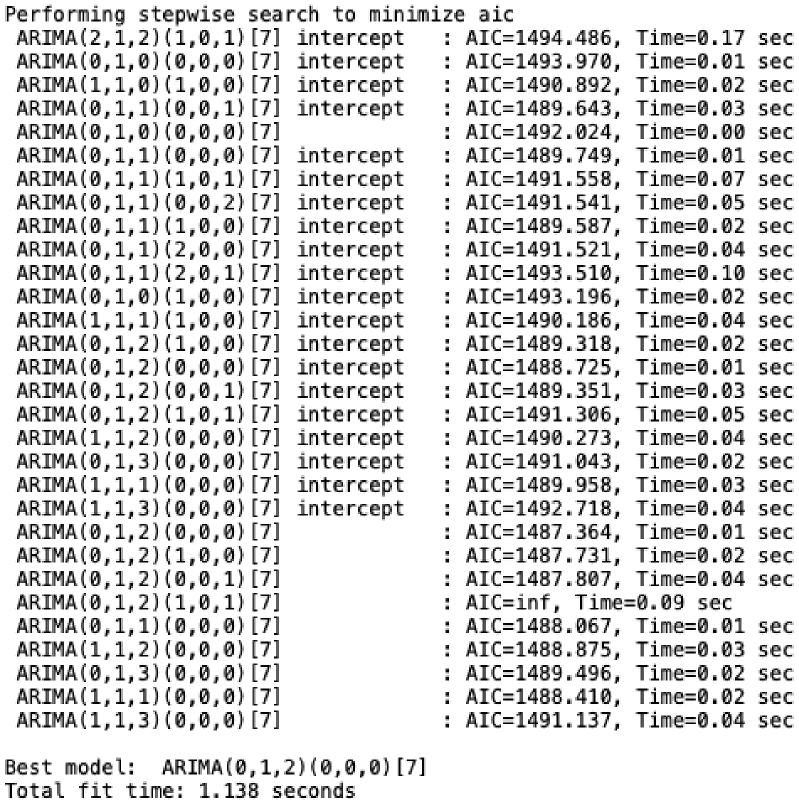
```
(-3.577419033611589,
 0.006207694631357907,
 8,
 63,
 {'1%': -3.5386953618719676,
  '5%': -2.9086446751210775,
  '10%': -2.591896782564878},
 1208.0076058681798)
```

Figure 4.12 - New dataset (with differencing = 1) Augmented Dickey-Fuller test

Figure 4.12 suggests that the series is now stationary. The next step is to calculate the optimal model orders.

4.2.2 Estimation and Model Checking

For this new dataset, another identification approach was adopted. To avoid unnecessary work, and as the hand work was already explained, it was used a Python module called "pmdarima.auto_arima" that will return the best model, ordered by BIC. This module requires the filling of some parameters, as the maximum differencing (for the series and seasonality) and model orders (*p, d, q, P, D, Q* and *s*). The output of this test is shown in Figure 4.13 and its parameters are visible in Python's documentation:



```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(1,0,1)[7] intercept   : AIC=1494.486, Time=0.17 sec
 ARIMA(0,1,0)(0,0,0)[7] intercept   : AIC=1493.970, Time=0.01 sec
 ARIMA(1,1,0)(1,0,0)[7] intercept   : AIC=1490.892, Time=0.02 sec
 ARIMA(0,1,1)(0,0,1)[7] intercept   : AIC=1489.643, Time=0.03 sec
 ARIMA(0,1,0)(0,0,0)[7]             : AIC=1492.024, Time=0.00 sec
 ARIMA(0,1,1)(0,0,0)[7] intercept   : AIC=1489.749, Time=0.01 sec
 ARIMA(0,1,1)(1,0,1)[7] intercept   : AIC=1491.558, Time=0.07 sec
 ARIMA(0,1,1)(0,0,2)[7] intercept   : AIC=1491.541, Time=0.05 sec
 ARIMA(0,1,1)(1,0,0)[7] intercept   : AIC=1489.587, Time=0.02 sec
 ARIMA(0,1,1)(2,0,0)[7] intercept   : AIC=1491.521, Time=0.04 sec
 ARIMA(0,1,1)(2,0,1)[7] intercept   : AIC=1493.510, Time=0.10 sec
 ARIMA(0,1,0)(1,0,0)[7] intercept   : AIC=1493.196, Time=0.02 sec
 ARIMA(1,1,1)(1,0,0)[7] intercept   : AIC=1490.186, Time=0.04 sec
 ARIMA(0,1,2)(1,0,0)[7] intercept   : AIC=1489.318, Time=0.02 sec
 ARIMA(0,1,2)(0,0,0)[7] intercept   : AIC=1488.725, Time=0.01 sec
 ARIMA(0,1,2)(0,0,1)[7] intercept   : AIC=1489.351, Time=0.03 sec
 ARIMA(0,1,2)(1,0,1)[7] intercept   : AIC=1491.306, Time=0.05 sec
 ARIMA(1,1,2)(0,0,0)[7] intercept   : AIC=1490.273, Time=0.04 sec
 ARIMA(0,1,3)(0,0,0)[7] intercept   : AIC=1491.043, Time=0.02 sec
 ARIMA(1,1,1)(0,0,0)[7] intercept   : AIC=1489.958, Time=0.03 sec
 ARIMA(1,1,3)(0,0,0)[7] intercept   : AIC=1492.718, Time=0.04 sec
 ARIMA(0,1,2)(0,0,0)[7]             : AIC=1487.364, Time=0.01 sec
 ARIMA(0,1,2)(1,0,0)[7]             : AIC=1487.731, Time=0.02 sec
 ARIMA(0,1,2)(0,0,1)[7]             : AIC=1487.807, Time=0.04 sec
 ARIMA(0,1,2)(1,0,1)[7]             : AIC=inf, Time=0.09 sec
 ARIMA(0,1,1)(0,0,0)[7]             : AIC=1488.067, Time=0.01 sec
 ARIMA(1,1,2)(0,0,0)[7]             : AIC=1488.875, Time=0.03 sec
 ARIMA(0,1,3)(0,0,0)[7]             : AIC=1489.496, Time=0.02 sec
 ARIMA(1,1,1)(0,0,0)[7]             : AIC=1488.410, Time=0.02 sec
 ARIMA(1,1,3)(0,0,0)[7]             : AIC=1491.137, Time=0.04 sec

Best model:  ARIMA(0,1,2)(0,0,0)[7]
Total fit time: 1.138 seconds
```

Figure 4.13 - Auto ARIMA Output

Figure 4.13 shows the model performance ordered by AIC. This is great, because it purposes of this this is to forecast (so the best criterion to decide the model order is Akaike's) and replaces the work of not only analysing the ACF and PACF, but also looping through the AIC and BIC results to discover the best model order.

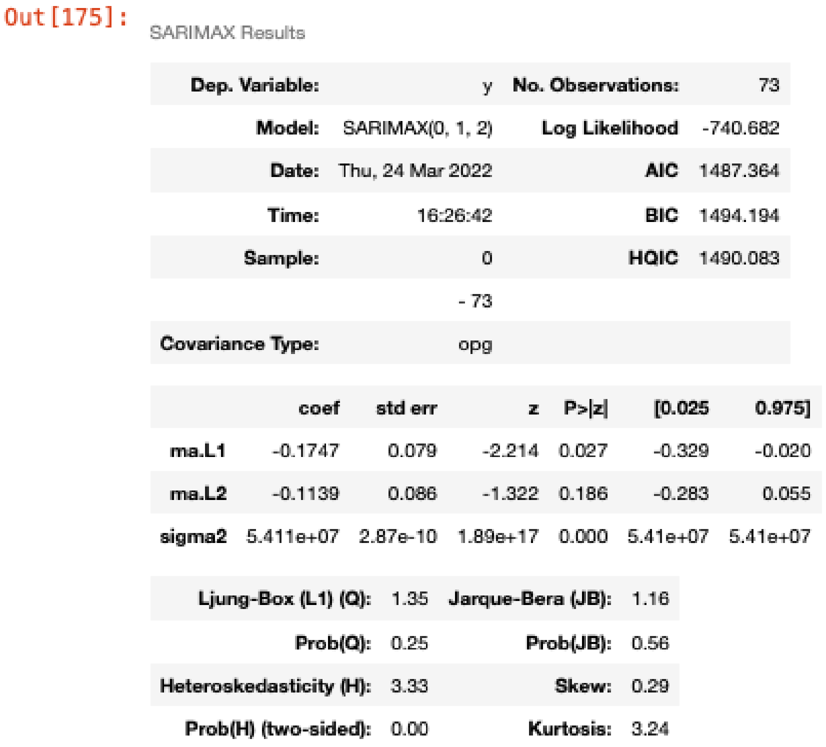
The best model orders for this transformed data are:

- p = 0.
- d = 1.
- q = 2.
- P = 0.
- D = 0.
- Q = 0.
- s = 7.

This is the final step – putting the model in production (automat data ingestion and treatment in the cloud in order to make it available to everyone for insights).

As the optimal model orders have been found, the next step consists in studying the residuals behavior.

Figure 4.14 displays the output of Prob (Q) and Prob (JB), that must be higher than 5% not to reject the null hypothesis - the residuals uncorrelated and normally distributed.



Out[175]:

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	73
Model:	SARIMAX(0, 1, 2)	**Log Likelihood**	-740.682
Date:	Thu, 24 Mar 2022	**AIC**	1487.364
Time:	16:26:42	**BIC**	1494.194
Sample:	0	**HQIC**	1490.083
	- 73		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ma.L1	-0.1747	0.079	-2.214	0.027	-0.329	-0.020
ma.L2	-0.1139	0.086	-1.322	0.186	-0.283	0.055
sigma2	5.411e+07	2.87e-10	1.89e+17	0.000	5.41e+07	5.41e+07

Ljung-Box (L1) (Q):	1.35	**Jarque-Bera (JB):**	1.16
Prob(Q):	0.25	**Prob(JB):**	0.56
Heteroskedasticity (H):	3.33	**Skew:**	0.29
Prob(H) (two-sided):	0.00	**Kurtosis:**	3.24

Figure 4.14 - ARIMA (*p* = 0, *d* = 1, *q* = 1, *m* = 7) Residuals Analyses

As both Prob (Q) and Prob (JB) are higher than the significance value (95%), the assumption is that the residuals are uncorrelated and normally distributed. Also, it is possible to visualize the residuals behavior. These plots are shown in Figure 4.15.

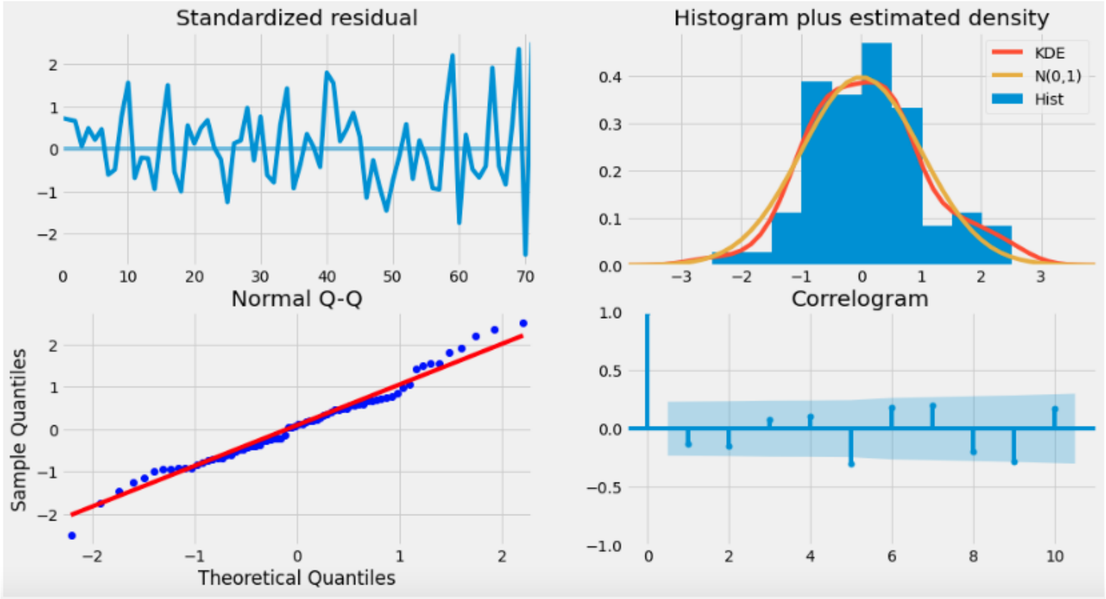


Figure 4.15 - ARIMA (p = 0, d = 1, q = 1, m = 7) Diagnostics Plots

The model is good for production, and the next step is to forecast the series' future values. The image bellow (Figure 4.16) shows the comparison between the forecasts and the real values:

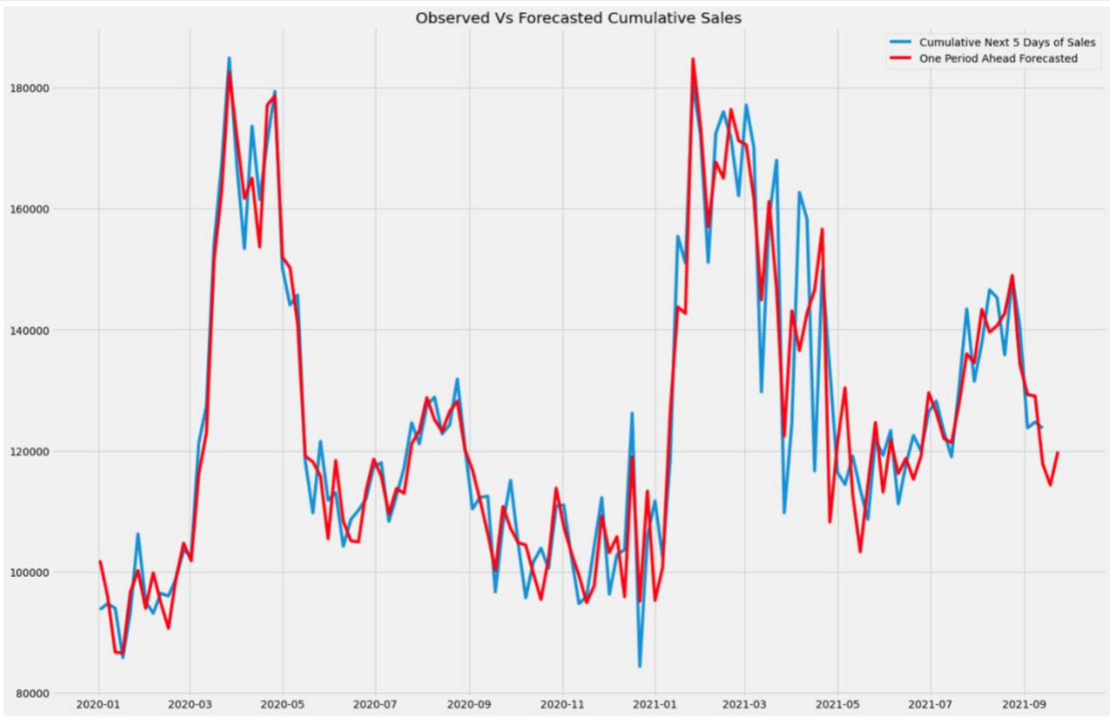


Figure 4.16 - ARIMA (p = 0, d = 1, q = 1, m = 7) One Step Ahead Forecast

It also possible to see the delta between the real values and the forecast in the image bellow. This was achieved by dividing the the difference between the real values and the forecast by the real values.

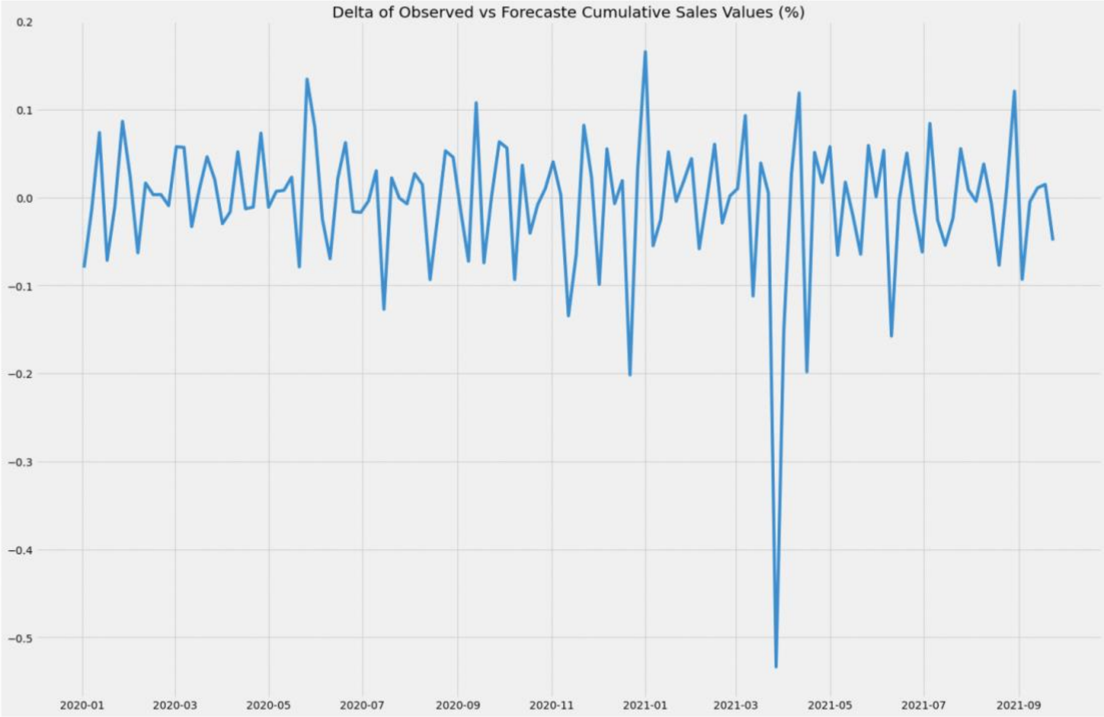


Figure 4.17 - ARIMA (p = 0, d = 1, q = 1, m = 7) Forecast Vs Real Values Delta

As Figure 4.17 shows, most of the deltas are between -10% and 10%. There are a few larger deltas, but further statistical tests are not the objective of this thesis.

Another way to investigate this comparison is by making a delta box plot. The image bellow represents a box plot of these delta (real values divided by forecasted) values:

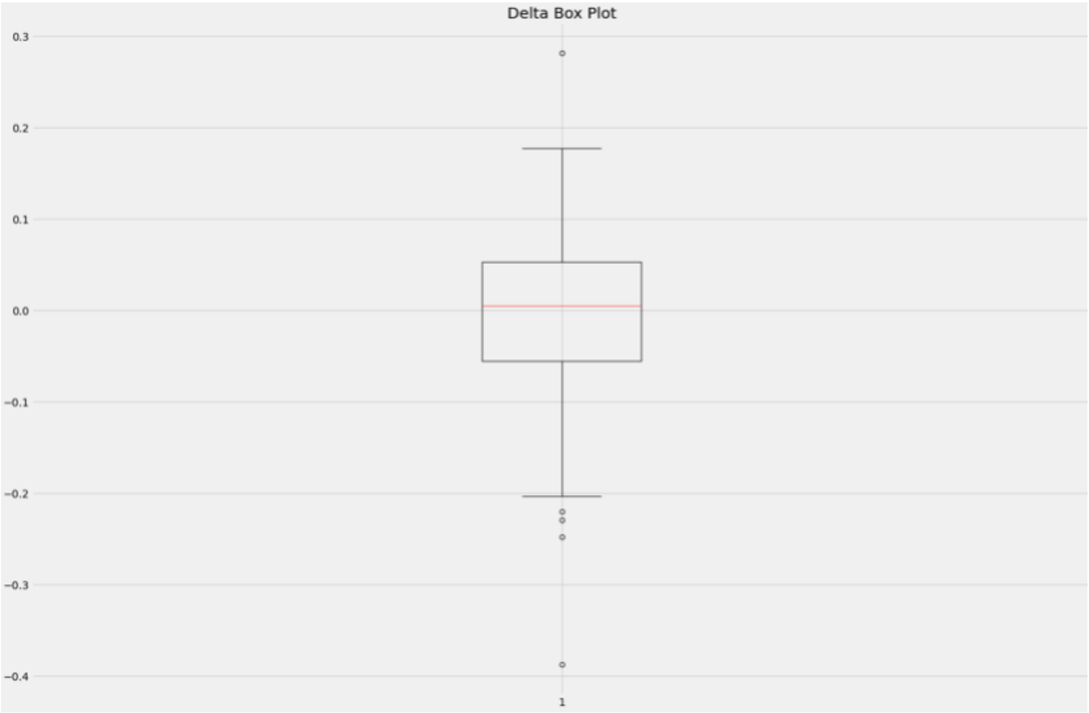


Figure 4.18 - ARIMA (p = 0, d = 1, q = 1, m = 7) One Step Ahead Forecast Delta Box Plot

Although there are some outliers, Figure 4.18 suggests that most of the observation have a [-5%; 5%] delta when compared to the real values. Recall that these predictions will fortune tell the next 5 days of sales.

As the author of this thesis thinks that visualizations can be of more use many times than data itself, it was also plotted a chart in which it is possible to see the real values and the forecast values with 95% confidence intervals. This output can be seen in Figure 4.19:

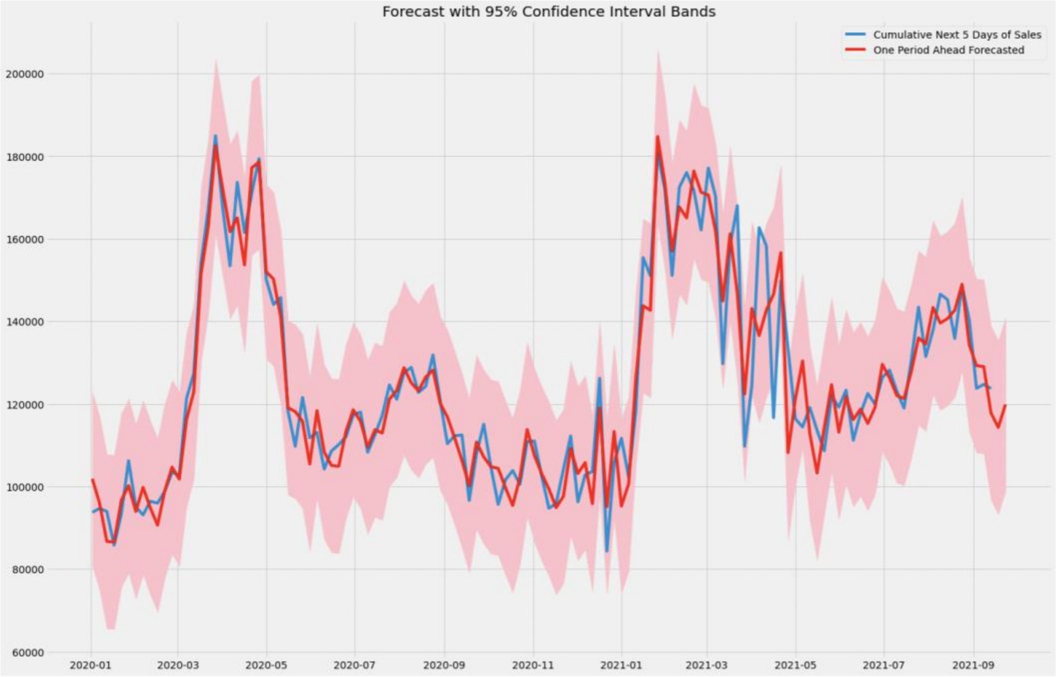


Figure 4.19 - ARIMA (p = 0, d = 1, q = 1, m = 7) Observed Values, Predicted Values and 95% Confidence Intervals Bands

As it is possible to see in Figure 4.19, the model was a success. Aside from a small period in which the real values are outside the confidence interval bands, that will be considered as outliers. The model is good to go for production

4.2.3 Application

Once the model has passed through the validation steps (step one and two), the next and final step is putting the model in application, with automated data imports and forecasts. This is the step where the data is dropped into a data lake, transformed by an API (can be anyone that has been showed in

To put the model into production, the daily sales would be transferred from the sales system via SFTP (secure file transfer protocol) into an API where they would be daily imported and treated. In the API all the machine learning calculations would be made and exported to a data lake. The next step would be importing them in a business intelligence tool, such as power bi, and make daily visualizations of the real values, deltas between real values and forecasted values, and future forecasts.

5
CONCLUSION

Efficient stock management is essential to avoid wastes and unnecessary warehousing costs. A poor stock management process can not only lead to increased storage costs, but also to a loss of profits due to shortages.

Adopting this forecasting model may allow it to work in small batches, adopting lean and agile methodologies to avoid waste (a thought that has been increasing not only in the stock managing processes, but also in the technological world with the project deliveries).

One of the goals of this research is to give visibility on the advantages of using technology and information technologies to add value to Industrial Engineering and Management. This objective was concluded with success as seen in chapter 4, as an automated model was developed, predicting with accuracy the next period sales in quantity.

More developments will be made in future research, giving more visibility to the cloud environments and other possible machine learning predictive models to forecast sales. Another interesting addition to the research is the concepts of Lean and Agile for Data Engineering and Data Science, through continuous integration and continuous delivery. Also, it will interesting to research about the difference between this thesis and real life stock managers (with no automated pipelines) accuracy.

One of the key benefits of adopting technology to manage stocks is the replacement of the periodic inventory managing techniques with fully automated continuous stock management algorithms. This kind of improvements would prevent the stock manager from the need of extracting, treating, analying and calculating buying quantities for each product (or stock category/ business unit, depending on the analysed granularity), ending up driving an increased time to bargain with the suppliers, finding secondary suppliers, validating minimum stocks for exposure (for instance in technological areas), planning budgets, etc.

All of the objectives of this thesis have been reached, leading to an automated analyses of a time series. This automated process includes turning the series into a stationary time series (key factor for forecasting the future values of a time series), fitting the daily imported data in the best model orders (defined through the Box-Jenkins methodology, using a SARIMA model) and forecasting the future value of sales.

BIBLIOGRAPHY

[1] D. Denis, M. St-Vincent, D. Imbeau, and R. Trudeau, "Stock management influence on manual materials handling in two warehouse superstores," *International Journal of Industrial Ergonomics*, vol. 36, no. 3, pp. 191–201, Mar. 2006, doi: 10.1016/j.ergon.2005.11.002.

[2] R. Y. Kim, "The Impact of COVID-19 on Consumers: Preparing for Digital Sales," *IEEE Engineering Management Review*, vol. 48, no. 3, pp. 212–218, Jul. 2020, doi: 10.1109/EMR.2020.2990115.

[3] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, "Evolutionary Machine Learning: A Survey," *ACM Computing Surveys*, vol. 54, no. 8. Association for Computing Machinery, Nov. 01, 2022. doi: 10.1145/3467477.

[4] N. Miloslavskaya and A. Tolstoy, "Big Data, Fast Data and Data Lake Concepts," in *Procedia Computer Science*, 2016, vol. 88, pp. 300–305. doi: 10.1016/j.procs.2016.07.439.

[5] Marco Peixeiro, "The Complete Guide to Time Series Analysis and Forecasting," Aug. 07, 2019.

[6] B. Zolghadr-Asli, M. Enayati, H. R. Pourghasemi, M. N. Jahromi, and J. P. Tiefenbacher, "A linear/non-linear hybrid time-series model to investigate the depletion of inland water bodies," *Environment, Development and Sustainability*, vol. 23, no. 7, pp. 10727–10742, Jul. 2021, doi: 10.1007/s10668-020-01081-6.

[7] A. H. Adineh, Z. Narimani, and S. C. Satapathy, "Importance of data preprocessing in time series prediction using SARIMA: A case study," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 24, no. 4, pp. 331–342, Jan. 2021, doi: 10.3233/kes-200065.

[8] S. Polwiang, "The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017)," *BMC Infectious Diseases*, vol. 20, no. 1, Mar. 2020, doi: 10.1186/s12879-020-4902-6.

[9] Y. Lu and S. M. AbouRizk, "Automated Box-Jenkins forecasting modelling," *Automation in Construction*, vol. 18, no. 5, pp. 547–558, Aug. 2009, doi: 10.1016/j.autcon.2008.11.007.

[10] I. Dobre and A. A. Alexandru, "MODELLING UNEMPLOYMENT RATE USING BOX-JENKINS PROCEDURE."

[11] N. Dritsakis and P. Klazoglou, "International Journal of Economics and Financial Issues Forecasting Unemployment Rates in USA Using Box-Jenkins Methodology," *International Journal of Economics and Financial Issues*, vol. 8, no. 1, pp. 9–20, 2018, [Online]. Available: http:www.econjournals.com

[12] W. Regis Anne and S. C. Jeeva, "ARIMA modelling of predicting COVID-19 infections", doi: 10.1101/2020.04.18.20070631.

[13] S. Makridakis, Ã. And, and M. Á. le Hibon, "ARMA Models and the Box±Jenkins Methodology," John Wiley & Sons, Ltd, 1997.

[14] C. Ateş, "Forecasting for the number of individual per dentist in Turkey; comparison of box-jenkins and brown exponential smoothing estimation methods," *Eastern Journal of Medicine*, vol. 25, no. 1, pp. 132–139, 2020, doi: 10.5505/ejm.2020.57805.

[15] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 3. Wiley-Blackwell, May 01, 2019. doi: 10.1002/wics.1460.

[16] N. Shong and C. Bs, "PEARSON'S VERSUS SPEARMAN'S AND KENDALL'S CORRELATION COEFFICIENTS FOR CONTINUOUS DATA," 2008.

[17] S. Ma, Q. Liu, and Y. Zhang, "A prediction method of fire frequency: Based on the optimization of SARIMA model," *PLoS ONE*, vol. 16, no. 8 August, Aug. 2021, doi: 10.1371/journal.pone.0255857.

[18] M. Maniruzzaman *et al.*, "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers," *Journal of Medical Systems*, vol. 42, no. 5, May 2018, doi: 10.1007/s10916-018-0940-7.

[19] J. G. Mackinnon, "Critical Values for Cointegration Tests," 2010. [Online]. Available: http://www.econ.queensu.ca/faculty/mackinnon/

[20] V. Bewick, L. Cheek, and J. Ball, "Statistics review 7: Correlation and regression," *Critical Care*, vol. 7, no. 6. pp. 451–459, Dec. 2003. doi: 10.1186/cc2401.

[21] J. Walsh PhD, "Statistical inference.," *Salem Press Encyclopedia*. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=90558475&site=eds-live

[22] S. F. O'Brien and Q. L. Yi, "How do I interpret a confidence interval?," *Transfusion (Paris)*, vol. 56, no. 7, pp. 1680–1683, Jul. 2016, doi: 10.1111/trf.13635.

[23] "THE NATURE OF TIME SERIES 1.2 WHITE NOISE," 2014. [Online]. Available: https://www.ebsco.com/terms-of-use

2022 Manuel Romão dos Santos Box-Jenkin’s Methodology in Python for Stock Managing

NOVA SCHOOL OF SCIENCE & TECHNOLOGY