BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS AND PHYSICS

COMPARISON OF FORECASTING AND PREDICTION MODEL ACCURACY USING MACHINE LEARNING ALGORITHMS. A CASE STUDY OF STEER`S ZIMBABWE SALES DATA.

A RESEARCH SUBMITTED BY:

TAFADZWA RJ MHEUKA (B1747978)

TO

THE FACULTY OF SCIENCE

BINDURA UNVERSITY OF SCIENCE EDUCATION0

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE HONOURS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS (HBScSFM).

SUPERVISOR: MR. XXXXXX

DEC

APPROVAL FORM

The undersigned certify that they have read and recommend to the Bindura University of Science Education for acceptance of a dissertation entitled "**COMPARISON OF FORECASTING AND PREDICTION MODEL ACCURACY USING MACHINE LEARNING ALGORITHMS."**

Submitted by **TAFADZWA RJ MHEUKA**, Registration Number **B1747978** in partial fulfillment of the requirements for the Bachelor of Science Honours degree in Statistics and Financial Mathematics.

TAFADZWA RJ MHEUKA	…….	…….
B1747978	Signature	Date

Certified by

Ms. P.Hlupo	…….	…….
Supervisor	Signature	Date

Certified by

Mr.	…….	…….
Chairman of department	Signature	Date

DECLARATION OF AUTHORSHIP

I TAFADZWA RJ MHEUKA hereby declare that this research project herein is my own original work and has not be copied or extracted from previous sources without due acknowledgement of the sources.

…… ……..

Signature Date

DEDICATION

To the Mheuka Family.

ACKNOWLEDGEMENT

I would like to express my gratitude to the Almighty God, who got me this far, who blessed me with precise people to help me during the different stages of my study.

This research was partially supported by Simbisa Brands. I would like to thank my colleagues from the Bindura University of Science Education who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations of this paper. It gives me great pleasure to express my deepest respect and sincere thanks to my research supervisor, for her encouragement, corrections, valuable suggestions, guidance and discussion and during my studies. I would also like to show my gratitude to Ransom Mheuka my father for sharing his pearls of wisdom with me during the course of this report. He continually and convincingly conveyed a spirit of adventure concerning this research.

It is with enormous gratitude to thank my family for their love, helps, and supports, especially my parents Ransom and Joyce Mheuka for being supportive and helping me get all the annoying little things done, my great brothers, Takudzwa and Zvikomborero Mheuka and my sister Fadzai Chitima for supporting me in my pursuit of this degree. I am also grateful to all my friends and for their encouragement and help. I could not have achieved this without their help.

Lastly, I would like to thank all my lecturers at Bindura university who gave me knowledge which I have then applied into this research.

ABSTRACT

This study aims to compare the most accurate technique between ARIMA model to forecast sales and Linear regression model using the method of ordinary least squares (OLS) to predict sales data for the same forecasted period hence the accuracy measure from the two models will be used to compare accuracy between forecasting and prediction. Forecasts and Predictions form other models will be looked into in passing. The Research objectives include the following; Identify the most appropriate ARIMA model using Hyndman-Khandakar Algorithm, Forecast sales values (using the chosen ARIMA model) and measure the forecasting accuracy, Use Machine Learning algorithms to build a Linear regression (OLS) predictive model, To predict the response variables (sales) using the explanatory variables (number of customers), Build other Machine Learning Predictive Models for forecasting and Prediction. Compare RMSE values from forecasting and prediction models.

In modeling and forecasting sales for Simbisa Brands Ltd, a time series data is required and also for prediction of sales a linear regression model using the OLS approach, data with two variables (response and explanatory variables) is required. The secondary data was obtained from Simbisa Brands Ltd for the period January 2018 to March 2019 on daily basis. The data will be extracted from the GAAP system and exported into excel for manipulation and cleaning as it will contain some information that is not of use to this research. Methodologically the study will use the ARIMA model for time series forecasting and the Ordinary Least Squares (OLS) Regression model for prediction, the data set will be divided into training and testing data to correctly measure the accuracy of the two models. Most of the analysis in this research will be done used using the R and Python programming software. The R software will be used hand in hand with python programming to produce similar results for comparison of results the R software will mainly be used in the analysis of the time series data and the Python software for the OLS regression analysis part. The researcher will use scale dependent measures to access the accuracy of the models and based on the results obtained the RMSE value was slightly different by a small figure leaving a gap for further research on the matter at hand.

Key words: Ordinary Least Squares, ARIMA, Forecasting, Prediction

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

ACF	**Auto-correlation Function**
ADF	**Augmented Dickey-Fuller Test**
AIC	**Akaike Information Criterion**
AICc	**Akaike Information Criterion (bias corrected)**
ANOVA	**Analysis of Variance**
AR	**Autoregressive**
ARIMA	**Autoregressive Integrated Moving Average**
ARMA	**Autoregressive Moving Average**
ARCH	**Autoregressive Conditional Heteroscedastic**
BIC	**Bayesian Information Criterion**
CAPM	**Capital Asset Pricing Model**
EGARCH	**Exponential Generalized Autoregressive Conditional Heteroscedastic**
GAAP	**Generally Accepted Accounting Principles**
GARCH	**Generalized Autoregressive Conditional Heteroscedastic**
GLM	**Generalized Linear Model**
GLS	**Generalized Least Squares**
IGARCH	**Integrated Generalized Autoregressive Conditional Heteroscedastic**
IID	**Independent Identically Distributed**
IPR	**Intellectual Property Rights**
KPSS	**Kwiatkowski-Phillips-Schmidt-Shin**
LLS	**Linear Least Squares**

LMM	**Linear Mixed Models**
MA	**Moving Average**
MAD	**Mean Absolute Deviation**
MAPE	**Mean Absolute Percentage Error**
MCMC	**Monte Carlo Markov Chain**
MLE	**Maximum Likelihood Estimator**
MSE	**Mean Square Error**
MSLOF	**Mean Square Lack of Fit**
MSPE	**Mean Square Pure Error**
MSR	**Mean Square Residual**
OLS	**Ordinary Least Squares**
PPC	**Posterior Predictive Checks**
QSR	**Quick Service Restaurant**
RMSE	**Root Mean Square Error**
SARIMA	**Seasonal Autoregression Integrated Moving Average**
SPSS	**Statistical Package for Social Sciences**
SSE	**Sum of Squares Errors**
SSLOF	**Sum of Squares Lack of Fit**
SSPE	**Sum of Squares Pure Error**
SSR	**Sum of Squares Residual**
SVM	**Support Vector Machine**
SVR	**Support Vector Regression**

CHAPTER 1: INTRODUCTION

1.1.1 OVERVIEW

This chapter will explain the background of the study, the problem statement, aim of the study and objectives of the study. Likewise, it comprises of the research questions and hypothesis, the significance of the study, assumptions, the limitation and delimitations of the study. Lastly this chapter speaks of the ways the study is going to be carried out and it will conclude with a summary.

1.2.1 BACKGROUND OF THE STUDY

There's a well-known saying, signifying that" those who don't acquire knowledge from past events are fated to duplicate the same mistake". The finest and most accurate sales budgets or forecasts methods and prediction techniques endorse to such an historical proclivity. The best way to effectively forecast sales or profits is to partake a vivid insight of what transpired in the past, without clearly understanding the past it might be difficult to have a vivid insight of the future. Current history is packed with stories of companies and in some rare but unique cases the whole businesses that have made strategic errors because of inaccurate sales forecasts. In frequent scenarios, as a matter of fact, sales forecasting has been believed of as a sensitive practice in such organizations. Moreover, … company directors have worked hard to increase the practice by investing huge amounts of cash into improving sales forecasts. In the intervening time operators/users (management) of sales forecasts will always continue to question experts in the sales forecasting department as to why errors in sales forecasts continue to upsurge. Now and then different methods/techniques have been exploited in coming up with sales forecasts and these can be characterized into two groups namely qualitative and quantitative methods. Qualitative Forecasting techniques are used when there is no data existing or if the existing data is not relevant to forecasting, it`s not purely presumption or guesswork but they are well developed approaches that don't base on past data and are subjective grounded on the perception of consumers and experts. This technique is frequently applied to middle or long-range decisions Examples of qualitative forecasting methods are:

- informed opinion and judgment
- Delphi method
- market research and
- past life cycle analogy.

Quantitative forecasting techniques can be implemented when the following situations are satisfied:

- availability of numerical past information
- behavior of past sales pattern will continue in the future,

Cases of quantitative methods that have been used as forecasting tools include:

- weighted and simple N-Period moving averages
- simple exponential smoothing and
- Poisson process model-based forecasting and multiplicative seasonal indexes; only to mention a few.

Extensive study has been done in this area, though some methods have been abused comparative to the other methods. The method of Regression analysis has more than a few applications in economics and finance, for example, the statistical technique, is essential to the capital asset pricing model (CAPM) This model determines the association between the probable (expected) return of an asset and the risk premium. The method of ordinary least squares (OLS) can be used to predict demand basing on the sales made or future and this gave the writer the enthusiasm to explore more on the association between profits (sales) and customer buying fast-foods to see what the forthcoming events holds for the company. The Fast-food industry has witnessed an upsurge in sales and the numeral of customers in the prior years, the high demand of Fast-food products has made it to some extent cumbersome to make accurate sales forecasts. Simbisa brands (Innscor fast foods) is a multinational Company located in Zimbabwe. Simbisa Brands Limited (”Simbisa”) is a Public Corporation that possesses, runs and franchises Intellectual Property Rights (”IPR”) of a group of Quick Service Restaurant (”QSR”) brands. Simbisa is exceptional in that it not only possesses the IPR’s of the brands inside its portfolio but also owner-operates a widely held QSR brands. Simbisa is a Pan-African QSR operative that presently runs QSR restaurants in more than 10 countries across Africa with future

determinations of further extension across the region. The major events of the Group comprise of the provision of fast-food services, the manufacturing and marketing of biological assets and the manufacture and selling of household supplies. Although the Fast-food industry has been thriving to get recognition in the big industries, for Simbisa brands had a very poor performance in consumer demand for the years 2015 to 2016 due to quick failing economic environment and substantial strategic change Simbisa Brands. This made the firm to look into wider financial techniques to help control the above-mentioned problems (2015-2016 weak consumer demand). Simbisa brands takes budgets to be an important performance evaluation tool for the company, these budgets are calculated monthly and are spread all over the company so that everyone can work at the equal level of focus. Simbisa brands has a Statistics department under finance section which is dedicated to the formulation of budgets.

1.3.1 PROBLEM STATEMENT

Forecasting and prediction of future sales for a lot of the firms in Zimbabwe has become the most tedious job to perform. This is because of the current economic situation of the country and numerous factors have to be taken into account that are not only quantitative but also, qualitative. The goal of most business corporations is to maximize profits and have as many customers as they can have regardless if the economy is collapsing. Although sales forecasts or prediction maybe a tool that can be used to attain this, it is becoming all the time more cumbersome to be able to use the forecasting or regression analysis as a standard tool of measure, as the variables used for forecasting or prediction are no longer an explanation of the economic climate. The necessity to produce accurate results has suddenly mounted as most budgets are now an over estimation of reality. This is now resulting in rigorous methods being put in place to try and fix the irregularities between the forecasted figures and what is actually on the ground. As for Simbisa Brands the inconsistencies between the actual figures and the forecasted figures has led the company directors to suspect theft when actually the fact being that budgets are a clear picture of what may have taken place in the previous year but are not a clear representation of the current situation at hand. In most scenarios most firms may use budgets as a guideline to help them into the future but they can also be a cause for their down fall and imprecise assumptions do not stem from lack of sales forecasting techniques. Instead the quality of the information(data) and its sales profile play a crucial role for the forecast of the time series and

the construction of a regression model. Problems are going to be assessed when forecasting or predicting sales figures. In as much as the history of the sales maybe an important factor but history alone is not going to be the only factor that will influence but also the pricing of the products will have immense impact on the sales quantity and customer counts. From start, multivariate time series seem to be the most appropriate model for this research. In remembrance history can't always be repeated in referring to the past sales. There is only a single price for a product at any given time and this complicates the design of the multivariate time series. Therefore, in high note the need for accurate sales forecasts has gradually risen and this can only be attained by applying forecasting models such as the ARIMA (add more) model which will be used in this research and linear regression techniques such as the ordinary least squares (OLS)(addmore) for prediction.

1.4.1 AIM OF THE STUDY

This study aims to identify the most accurate technique between ARIMA model to forecast future sales and Linear regression model using the method of ordinary least squares (OLS) to predict sales data for the same forecasted period hence the accuracy measure from the two models will be used to compare accuracy between forecasting and prediction. Furthermore, it will look into other machine learning predictive models and compare the RMSE for the models.

1.5.1 OBJECTIVES OF THE STUDY

The Objectives of this study are:

A. Identify the most appropriate ARIMA model using Hyndman-Khandakar Algorithm
B. Forecast sales values (using the chosen ARIMA model) and measure the forecasting accuracy.
C. Use Machine Learning algorithms to build a Linear regression (OLS) predictive model
D. To predict the response variables (sales) using the explanatory variables (number of customers)

E. Build other Machine Learning Predictive Models for forecasting and Prediction.
F. Compare RMSE values from forecasting and prediction models.

1.6.1 RESEARCH HYPOTHESIS

H_0: Sales forecasted by the ARIMA model are more accurate as compared to sales predicted by an OLS Linear regression model.

H_1: Sales forecasted by the ARIMA model are less accurate as compared to sales predicted by An OLS Linear regression model.

RMSE – Root Mean Square Error a model accuracy measure.

$$Reject\ H_0\ if\ RMSE\ from\ the\ ARIMA\ forecasting\ model > RMSE\ from\ the\ Ordinary\ Least\ Squares\ regression\ model$$

1.7.1 SIGNIFICANCE OF THE STUDY

Numerous methods have been implemented over the years to forecast and predict demand. In the available literature relating to prediction and forecasting, it is well-known that no quantitative model would be perfect for all situations under any circumstances. Hence the motivation of this study is to build the most effective ARIMA model when forecasting demand for a highly volatile market, in this case the Fast food Industry and use the model to forecast future sales values and also use linear regression analysis to predict the sales for the same forecasted period and compare the accuracy for both models.

1.8.1 METHODOLOGY

In modeling and forecasting sales for Simbisa Brands Ltd, a time series data is required and also for prediction of sales, a linear regression model using the OLS approach, data with two variables (response and explanatory variables) is required. Therefore, secondary data will be obtained from Simbisa Brands Ltd January 2018 to March 2019. The data will be extracted from the GAAP system and exported into excel for manipulation and cleaning as it will contain some

information that is not of use to this research. The statistical tools include tables, graphs, Moving Average (MA), Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA) model and Regression Model. A time series is a set of data points, usually indexed at regular intervals, time series data occur naturally in many application areas. Time Series analysis methods pre-date those for Markov chain and general stochastic processes, the aims of time series analysis is to explain and summarize time series data, fit models (low-dimensional), and make forecasts.

The method of OLS is used for predicting the unknown parameters located in a linear regression model this is the most common estimation technique for linear models and as long as your model satisfies the OLS assumptions for linear regression, you can be sure that you're getting the best possible estimates. In gathering literature on the subject, the researcher consults the Bindura University's Library, Simbisa Bands Annual Reports, and the internet. Other Predictive models results will be done using python programming.

The R (version 4.0.2), SPSS (version 25.0), Microsoft Excel 2019 and Python editor Juypter Notebook (for python version 3.7) were used in the analysis.

1.9.1 SCOPE AND LIMITATIONS OF THE STUDY

The thesis is limited to the objectives. Research work is categorized by some restraints, some of these restraints include:

- time constraints
- the difficulties in obtaining relevant materials on the topic.
- the study is focused on a solely on one product which may prove insignificant to other product lines
- Qualitative aspect cannot be taken into account in coming up with the forecasts or the predictions, for instance customer preference.
- Data is collected from one brand only.

1.10.1 ORGANIZATION OF THE STUDY

Chapter 1 is made up of overview, which comprises the background of the study, problem statement and objective of the study. It also presents the aim of the study, the methodology, scope and limitations of the study and the significance of the study. Chapter 2 highlights related literature on the study topic with ideas of different writers whose findings have been defined in relation to the topic under study. Chapter 3 focuses on methodological review in the light of mathematical and statistical tools that are applicable to the analyses of the data. Basically, the study seeks to use time series models for forecasting and regression models for prediction. Chapter 4 deals with the data collection and analysis, and the results from the application of the ARIMA, the OLS regression models and other machine learning models for prediction and forecasting. Lastly, chapter 5 consists of summary, conclusion and recommendations.

1.11.1 CHAPTER SUMMARY

The chapter gave an introduction to the thesis report elaborating and painting a vivid picture on I the background of the study, problem statement and objectives guiding the study, methodology and significance of the study. In addition to these are scope and limitations of the study as well as thesis organization. The chapter concludes with this summary.

CHAPTER 2: LITERATURE REVIEW

2.1.1 INTRODUCTION

In this section, we provide a review of empirical studies on retail sales forecasting with a view to make clear the contribution of the study at hand. The section helps to compare and contrast with what other scholars have managed to cover in the field. The chapter also sheds light on the historical background of the methods and give a clear understanding of how the methods have evolved over time.

2.2.1 FORECASTING AND PREDICTION

Application of statistical models may be used with the purpose for predictions, according to Shmueli (2010), in fact, according to Gregor (2006) predictions depict "What will be and not why", a reference to the fact that one can predict without knowing the reasons for the prediction. Thus, accordingly, means the exclusion of any theoretical construct that may support the outcome of a prediction and that the occurrence of a future may happen if specific preconditions exist in the future. **According to Siegel (2013), predictions are not necessarily accurate to be significant for companies as they supersede any assumptions and support decisions with empirical data**. Shmueli (2010) further comes up with what she calls "predictive modelling", a reference to the process of using statistical data mining algorithm to the available sales information for the prediction of new or future observation (p29). Therefore, accordingly, any method that produces predictions becomes a predictive model and new observations will likely include observations that were not certain within the original data set including the observations of event to come. Shmueli and Kopp (2010) further emphasize that it is imperative to differentiate explanations from predictions by stating that predictions aim to explain casually what the future holds.

Siegel (2013) looks at forecasting as "a process that makes aggregate predictions on a microscopic level" such as estimation of the exact number for next month's Chicken Inn, Creamy Inn or any other fast-foods sales at Innscor fast-foods but this study will focus on Steer`s

sales. He goes on to define it as “an estimate of the probabilities of the possibilities for a key of variable at a future point in time” (p4). This applies favorably to companies such as Innscor fast-foods in that future sales should be forecasted with a view to proper budgeting and financial planning. A possible outcome of a forecast is that fried chicken sales grows with a 65% chance compared to 10% chance of no growth and perhaps 15% chance their appearance of a declining sales trend. Mahadevan (2010) views forecasting as distinguishable and based on short, medium- and long-term forecast. As a company therefore, following Mahadevan views, Innscor fast-foods should have a short-term forecast averaging 1-3 months, a medium-term forecast of about 10-18 months and a long-term forecast of 5-10 years.

It is worthwhile to note that the provision of future events that are associated with probabilities maybe seen as a difference between forecasting and predicting. Knaub (2015) sees a time series forecast as a more complex than a prediction by taking more variables such as seasonality, autocorrelations as well as smoothing techniques into account. The literature herein considers forecast, as more complex than prediction and accurate variables to investigate the future may prove complicated but are essential. A prediction that is justified by a theoretical construct potentially improves the quality of forecast because the insights derived from prediction can be applied in a forecasting model as well both concepts are useful to identify new observation but differ on the thesis to differentiate between forecast that foretell out of sample events for predictions that are based on data in addition to a supporting theory to observe new or future events that can also be detected within a sample. This will prove to be an obstacle to Innscor as a company. While predictions look into the future, they may do so in a short period without the provision of probabilities.

All in all; McCarthy, Davis, Golicic & Mentzer (2006) have come up with a number of factors to consider as regards forecasting and predictions. They have been shifts in the forecasting context for the past 20 years through the occurrence of the internet, globalization, competition as well as the increased number of sophisticated forecasting models with and without software support. Companies and their staff are often not too familiar with these new approaches due to the lack of training and poor commitment resources which result in unsatisfactory forecast performance. The low developments in forecasting and predicting call for new approaches that will improve the sales forecast and predictions. **Furthermore, a forecast includes one variable and the**

prediction in the underlying study includes the number of customers data as the independent variable, a factor that is difficult to encamps in the present Zimbabwean economic environment. The literature however has shown that both terms can be used interchangeably by many authors despite the emphasized difference and the advantages or disadvantages of each one of them.

2.3.1 TIME SERIES forecastig

Karapanagiotidis (2012) points out that, Modern time series forecasting methods are basically rooted in the knowledge history for-tell the future. Now the major issue to address is how we are going to interpreting the information given by the past events and most importantly how we are going to extrapolate future events relating to the past information, and this institute the main subject matter of time series analysis, the approach to forecasting time series is to first specify a model, though this need not be so. This model is a statistical design of the relationships between that which we observe and those variables we believe are related to that which we observe. The discussion of this topic is limited to the scope of those models which can be formulated parametrically. Currently there are at least 70 different forecast models among linear and nonlinear techniques for quantitative demand forecasting (Kerkkanen et al 2009). In the literature relating to forecasting, it is well-known that no quantitative model would be perfect for all situations (Pindyck, 2004). The models in general all have likenesses or somewhat have the same concepts, but in follow patterns from diverse areas, though numerous comparative studies have been defined in the existing literature relating to the study at hand, the discoveries do not propose what circumstances make a method better than the other (Veiga et al ,2012). Thus, circumstances of seasonality and perishability, as occur in the retail of food products, require analytical studies about the most suitable technique for each scenario (Veiga, 2009). Moreover, in order to make this corporate resource more robust, the article at hand aims to offer a straightforward analysis if time series models and regression models can properly work hand in hand to provide answers in fast food sales data, these model have extraordinary adaptive capabilities to deal with the linearity in problem solving. Several researches in the literature comparing linear and nonlinear forecasting methods have been performed with aggregate data of retail products internationally (Chu, 2003). This work show that any forecasting model can be regarded generally as appropriate models for sales data, from most presented papers no actual proof has convulsively

correlated the features which are pivotal for the choice of a precise forecasting method. As an Outcome, the accuracy of any precise method is supported up by the existence of established examples were the method would've been effectively used and the observed evaluations remain the appropriate method for solving specific situations, many theoretical and heuristic methods were developed and tested empirically in recent decades (Armstrong, 2000). Gauss (1974) discussed what can be categorized as the" classical" where the time series approach was derived from the regression analysis. Past context of enormous systems of equations models common among macro- econometrics forecasters of the 1950's case in point the Klein-Goldberg model (1955) it came vivid the forecasting models derived from "signal extraction" context forecasted at least as well as those based on complicated systems of economic relationship equations formulated as individual, yet interconnected, dynamic classical regression. Approaches such as ARIMA were able to extract trend, seasonal and cyclical components from a series by means of an iterated finite moving average process, far along methods improved further on this general method. Holt (1957) and Winter's (1960) further generalized this method to include a slope component in the forecast function. Brown (1963) reconstructed the problem in terms of a discounted least squares regression, termed "Linear exponential smoothing" it turns out that this method was just a unique extension of the Holt-Winters Method. These approaches developed due to their uncomplicatedness and ease of implementation for the general Expert. Nevertheless, these approaches are well-thought-out by others as to being improvised since they fail to unite hypothetical attention to the breakdown of cyclical components and they are formulated without recourse to a well specified statistical model. Makridakis and Hibbon (1979) seemed to propose that modest ad hoc approaches have a tendency to forecast at least as well as more complex approaches. Nevertheless, Newbold (1983) contended that it is not by arduously applying diverse approaches that we should be troubled with but somewhat, that we should take into account the related matters involved in forecasting any specific series in its own right. That is to say, the choice of method should inevitably come second as to the approach which depends on characteristics of the particular series and that most time would be well spent in considering case studies which assess how numerous approaches to forecasting complement the nature of the data, as opposed to "blind" horseraces between methods with this epiphany, it is vital to take into consideration more modern approaches which allow us greater flexibility in stipulating a distinct

statistical model, contrast to limiting ad hoc approaches. Likewise, this approach should free us to incorporate an economic foundation for our decision of a specific forecasting method.

The ARIMA model has been widely researched and applied in studies of forecast because of their theoretical properties and of the numerous empirical supporting evidence (Claudimar et al, 2010). Moreover, ARIMA model has similarity with most models of exponential smoothing, except for the multiplicative form of Holt-Winters (Makridakis, 1998). The ARIMA model was popularized by George Box and Gwilym Jenkins in the 1970's, with application in time series analysis and forecasting. The underlying theories described by Box and Jenkins (Box, 1970) and later by Box, are sophisticated but easy to understand and implement. For forecasting technique involves two steps (Delurgio, 1998); the analysis of time series and the selection of the forecasting model that best fits to the data set and ARIMA utilizes the same pattern of analysis and selection by decomposition methods and regression

2.4.1 LINEAR REGRESSION AND ORDINARY LEAST SQUARES (OLS)

- add data at regression more

Regression analysis is the part of statistics that examines the relationship between two or more variables related in a non-deterministic manner (Devore, 2009). Ordinary least squares (OLS) is a technique for estimating the unknown parameters in a linear regression model. This technique minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation. Thus, the OLS estimates are identical to the maximum likelihood estimates provided it satisfies the standard assumptions and this is done by the technique of generalized least-squares the application of GLS, Aitken (1934) has recommended another approach that is the transformation of the variables of the original model to produce a new one which satisfies the standard assumptions of the random variable and to which OLS method can be applied. One practical example was OLS can be used is the capital asset pricing model (CAPM). Numerous scholars have examined the CAPM and their approaches generally involve ordinary least square regression (OLS). CAPM assumes that equity returns are normally distributed, because they are fat tailed (Fama,1976) buy using only OLS to test the CAPM. There are two different variable types associated with regression and OLS:

- the first is the class of independent variables. Independent variables are observed through research and study.

- The second type of variable is the dependent variable or response variable.

The determination of regression is to model and examine the relationship between the dependent variable and the independent variable at hand. some of the most tough tasks of a predictor is to determine which of these to use, for the writer determine the variables to use in a regression, the relationship between the dependent and independent variables must be well-known. The significant relationship for this study is correlation and is well-defined as the linear relationship between two variables. The OLS performs well when the error terms are well-behaved or the underlying assumptions hold (Adebanji, 2013). However, failure of these assumptions leads to high sensitivity of the OLS, a simple technique of fitting a straight line when both variables are subject to error was studied by Wald in (1940) and afterward Bartlett's paper (1949) shows an adjustment of Wald's technique having the advantage of greater accuracy. Linear associations between variables affected by errors have been discussed by Watson et al., (1966). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased is discussed by Mc Elroy (1967), In his journal, he presented that linear regression model ordinary least-squares estimators are best linear unbiased if and only if the errors have equal variance and the same non-negative coefficient of correlation between each pair. All-time series models revolve around the notion that standard regression model involves specifying a linear parametric relationship between a set of exogenous variables (explanatory variables) and endogenous variables (dependent variables) and helps to guide in coming up with the intended results. Regression mostly is about relationship between variables. (Michael Lewis-Beck 1980, p. 9) "What is the relationship between variable X and variable Y?". If the relationship is expected to be linear, bivariate regression is used to fit a straight line to a scatterplot of observations on variable X and variable Y.

2.5.1 FORECASTING AND PREDICTION IN THE FAST-FOOD SERVICES

Though forecasting in the fast food business is now vastly advanced that has not always been the case, Finley 1986 emphasized that the acceptance of management science techniques, including forecasting has not occurred extensively in the food service industry. Miller (1991) led a study of three foodservice venues in which she solicited the types of forecasting models that were used. Outcomes of the study showed that mathematical models were not commonly used in the commercial foodservice. The most regularly used methods are variations of the nave model.

Respondents in all studies indicated that forecasting is a vital concept and established that extra training and development is required. Forecasting is significant as it contributes to improved liability, cost control, efficiency, productivity profit, maximization, and customer and worker satisfaction. The conclusions stated by Miller (1991) support the necessity for and significance of continued study in the area of restaurant and foodservice forecasting. The influence of enhanced forecasting techniques on service production ought to be recognized. Forecasting affects not only material needs, i.e., food and labor, but also all costs, employee morale, and customer fulfilment. Forecasting the suitable amount of food to be purchased in the foodservice operation is a vital decision that affects costs and the efficiency of the processes. Perishability of food makes forecasting an exceptional problem in the foodservice operation (Miller Shanklin, 1988). If the forecast is brief, foodservice processes will run out of food. On the other hand, forecasting food in surplus will cause an upsurge in food purchased and food produced, resulting in an increase in costs. Miller 1992 emphasized that Foodservice processes need correct purchasing forecasts. Accurate purchasing forecasts will assist in monitoring the following; the amount of food purchased, the number of employees essential to run the operation, and the amount of food waste on serving lines. Accuracy in the forecast is reflected in customer satisfaction, resource utilization, and profitability (Bloss, 1991). Restaurant covers (dine in guests) were developed and elevated by Miller et al. (1991). The number of forecasts was a result of the demand for specific menu items. Information was gathered from two restaurants in mid-sized hotel entities and this resulted in simple mathematical forecasts being practical tools that are applicable in commercial food-provision operations. To some measure, the approaches are easy and software and hardware are in place for food service operators. In addiction results shows that simple mathematical models may increase efficiency and effectiveness of hospitality management (Miller et al 1991); for example, for optimization of staffing and inventory levels, a food and beverage manager a forecast including controls which offer potential for cost savings and revenue enhancement. Miller (1993) went further in his research on forecasting restaurant covers to determine if they were a difference in selection of mathematical techniques based on short term and long-term data sets. In both studies, daily seasonality difference accounted for large proportion of the variance in the covers. Conclusively, therefore, seasonality is crucial when forecasting (Miller et al 1993) (SMALL PRED PARA)

2.6.1 CHAPTER SUMMARY

From various writers referred to in the context of this chapter it is vivid that while forecasting and prediction are a broadly researched field of study, no model can be applied for all forecasts and predictions. Diverse approaches have to be well-thought-out conditional to the various influences that affect the proposed environment. Although demand forecasting and prediction has been extensively explored, not many writers have considered the fast food service industry and a lot is still yet to be uncovered.

CHAPTER 3: METHODOLOGY

3.1.1 INTRODUCTION

This chapter introduces some fundamentals (Basics) of time series and regression (OLS) analysis for forecasting and prediction, we will look at components of time series data and the ordinary least squares regression analysis. This Chapter is an overview of time series and regression analysis techniques. The writer will also go through some univariate time series model for forecasting the concepts from this chapter will help in the application for the Autoregressive (AR) models, Moving Averages (MA) models, Autoregressive Moving Averages (ARMA) models, the Autoregressive Moving Integrated Moving Averages (ARIMA) models and other statistical concepts, this chapter will also look and the general linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ and some assumptions of OLS. In many problems, there are two or more variables that are related and it is crucial to model and explore this relationship. In general, suppose that there is one dependent variable that depends on k independent. The relationship between these variables is categorized by a mathematical model called a regression equation and can be used for data prediction. The regression model is fit to a set of sample data. This chapter will also discuss other forecasting and prediction techniques.

3.2.1 RESEARCH METHODS AND DESIGN

Research strategy and methods

The research held with respect to this dissertation was an applied one, but not new. The specific research method for this dissertation comprises of quantitative techniques, the aim of a quantitative research method is to categorize features, sum them, and build a statistical model in an effort to elucidate what is observed. This form of research is entirely dependent on the historical data available for the study at hand. Another aspect that can be considered is Qualitative research methods, although these are not used by the author within this dissertation, they are a vital, and a congenital part of the method used by the company being observed. In most cases this research method is used on the launch of a new product this is in most cases due

to the lack of historical information, although other companies' case in point Simbisa Brands prefer to utilize the method throughout the products life cycle.

Research Design

Any forecasting and regression (predictive) models are highly dependent on the use of insurmountable amounts of data to increase the accuracy of the model. Simbisa brands collects data using the GAAP Accounting application, the sales and the number of customers for any specific branch are recorded as the sales occur, Hence the data available is relatively a clear representation of what is on the ground. The data sample used is for a period of 364 days, analyzing daily sales.

3.3.1 SOFTWARE PACKAGES

In this research project I will mainly use open-sourced statistical packages for my analysis, most of the analysis in this research will be done used using the R and Python programming software. The R software will be used hand in hand with python programming to produce similar results for comparison of results. The data collected will be in Microsoft excel format and will be imported to R and Python from excel. The forecasting (ARIMA model building) part will be done using the R software and the rest of the analysis will be done using Python programming to obtain results the analysis will be carried out independently.

3.4.1 FORECASTING AND PREDICTION

Forecasting refers to the process predetermining future trends and the impact on the organization this process takes into account the past and the current information to predict the future events. There are two methods of forecasting

- Quantitative forecasting is an explanatory method which attempts to correlate variables using past records and trends to make forecasts. This method uses time series analysis
- Qualitative forecasting is the method that relies on expert judgment rather than numerical figures to make the forecasts.

A prediction is a statement which tries to explain a possible outcome or future event it comes from the Latin term Pre which refers to before and dicer which means say in English. Companies use predictions determined to guide through uncertain projects despite their uncertainty.

Differences Between Forecasting and Prediction

- A forecast refers to a calculation or an estimation which uses data from previous events, combined with recent trends to come up a future event outcome. On the other hand, a prediction is an actual act of indicating that something will happen in the future with or without prior information.
- **A Forecast is more accurate compared to a prediction**. This is because forecasts are derived by analyzing a set of past data from the past and presents trends. The analysis helps in coming up with a model that is scientifically backed and the probability of it being wrong are minimal. On the other hand, a prediction can be right or wrong e.g. if you predict the sales for a particular date, the result depends on many factors such as customers, weather etc.
- **Forecast is more appropriate when one deals with time series data. it's nothing but a "time" based prediction. The other " Prediction" is not necessarily time based but based on factors that influence a dependent variable which may be causal factors or correlation factors.**
- Forecasting is done for future data taking into consideration the past data and in prediction we study the factors affecting the variable of our interest, and how they affect it. Also, prediction does not take time into consideration, whereas forecasting does.
- Forecasting has more to do with out of sample observations, whereas prediction deals more with in-sample observation. Predicted OLS values are calculated for observations in the sample used to estimate regression. It is worthwhile to note that forecast pertains more for some dates beyond the data used to estimate the regression, so the data on the actual volume of the forecasted variable are not in the sample used to estimate regression.

Taking time-series into consideration, forecasting appears to be an **estimate** of a future value given past values of a time series. In regression, prediction appears to mean an **estimate** of a value whether future, current or past with respect to the given data. In regression, it is possible to extrapolate existent regression model to new subjects not being in the training sample and predict the outcome or dependent variable.

It is worth noting that in forecasting a look in the subjects' historical data to build a model and then predict certain outcome in future based on the same model. Forecasting and prediction both relate to basically the same concept i.e. both are future oriented. All forecast are predictions but not all predictions are forecasts since when regression is used to explain the relationship between two variables, **"forecast" implies time series and the future while "prediction" does not.**

3.4.2 PREDICTIVE MODELLING

Predictive modeling is the process of applying a statistical models or data mining algorithm to data for the purpose of predicting future values. Non-stochastic prediction (Geisser, 1993, page 31), is where the goal is to predict the dependent variable Y for future observations given their input or independent variable X, this description also includes time-based forecasting, where observations until time t are used to forecast future values at time $t + k, k > 0$. In prediction they are interval prediction, point prediction, predictive distribution, prediction regions and rankings. Any technique that produces predictions irrespective of its fundamental approach is a predictive model such as Bayesian or frequentist, parametric or non-parametric, data mining algorithm, statistical models.

3.5.1 SOME DEFINITIONS IN TIME SERIES

3.5.1 TIME SERIES

A Time Series is a stochastic process in which T is a set of time point, usually T = {0,±1,±2,…}, {1,2,3…}, [0,∞),or (-∞,∞), the term Time Series is also used to refer to the realization of such a process (observed time series). Time series can also be defined as an ordered sequence of values of a variable at equally spaced time intervals. Examples of time series data include daily average temperature, monthly sales of a particular product and so on.

Time series data can be used to plan for the future, understanding past behavior, business forecasting and so on. Basically, time series analysis attempts to understand the behavior of data through use of models to forecast future behavior based on past data, time series models used include MA, AR, ARMA, ARIMA, ARCH, GARCH, IGARCH, EGARCH and other, but mainly on this study I shall focus on few models.

3.5.2 DIFFERENCING

The differenced series is the change between each observation in the original data series i.e. $y_t' = y_t - y_{t-1}$. The differenced series will have only $T - 1$ values since it is not possible to calculate a difference y_t' for the first observation. Occasionally the differenced data will not appear stationary and it may be necessary to difference the data a second time i.e.

$y_t'' = y_t' - y_{t-1}'$

$= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \dots\dots\dots\dots\dots (eq\ 3-1)$

$= y_t - 2y_{t-1} + y_{t-2} \ \dots\dots\dots\dots\dots (eq\ 3-2)$

y_t'' will have $T - 2$ values.

When a series is differenced once a linear trend is removed, when it is differenced again both linear and quadratic trend are removed. A seasonal difference is the difference between an observation from the previous year

3.6.1 COMPONENTS OF TIME SERIES

A vital step in choosing appropriate modeling and forecasting procedure is to consider the type of data patterns exhibited from the time series graphs of the time plots. The sources of variation in terms of patterns in time series data are mostly classified into:

Trend – when there is long term increase or decrease in the data

Seasonal – when a series is influenced by seasonal factors and recurs on a regular periodic basis.

Cyclic – when the data exhibit rises and falls that are not of a fixed period.

3.7.1 UNIVARIATE TIME SERIES

Univariate means that there is only one variable in the model. Univariate time series models rely on predicting a variable basing on its past observations. They are a number of ways to model univariate time series, some of the ways shall be discussed below.

3.7.2 AUTOREGRESSIVE (AR) MODELS

An autoregressive (AR) model is basically a linear regression of the present value of the series in contrast to one or more past values of the data points. The value of p is called the order of the AR model and AR models of order p abbreviated AR (p) are widely used un time series analysis.AR models can be evaluated with one of several approaches, including standard linear least squares approach. They also have a straightforward interpretation, for an AR model the current values depend on past values plus an error term or a disturbance term a white noise. A widely used approach for modeling univariate time series is the autoregressive (AR) model represented in the equation below

$$X_t = \phi_1 X_{t-1} + \phi_1 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-11)$$

Where: w_t is the white noise process with $\mu = 0$ and σ^2 $w_t \sim N(0, q)$, ϕ, ϕ_1, ... are parameters of the model and X_t is the time series.

The backward shift notation of an AR model is denoted by:

$$X_t = (\phi_1 B + \phi_1 B^2 + \cdots + \phi_p B^p) X_t + \varepsilon_t \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-12)$$

Where: $\phi_1(B) = \varepsilon_t$

$$\phi(L) = \left(1 - \phi_1 B + \phi_1 B^2 + \cdots + \phi_p B^p\right) \ldots\ldots\ldots\ldots\ldots (eq\ 3-13)$$

where B is the backshift operator.

We seek the coefficient ϕ so as to minimize the error function. X_{t-1} is related to X_{t-2} and any shock to X_t will gradually fade off. If $\phi = 0$ then X_t is a white noise, else X_t is auto-correlated and not stationary. Larger value of ϕ lead to greater auto-correlation and negative values of ϕ results in oscillatory time series.

3.7.3 MOVING AVERAGE (MA) MODELS

Moving average model is parsimonious time series model that uses a linear regression of the current value of the series against the white noise or random shocks of one or more previous values of the series and used to account for short-run autocorrelation. The random shocks at each point are assumed to come from a normal distribution. The distinction in this model is that these random shocks are propagated to future values of the time series and fitting the MA estimates is

more difficult than with AR models because the error terms are not noticeable. Thus, iterative non-linear fitting measures should be applied instead of LLS, MA models also have a less clear interpretation than AR models. Moving Average (MA) is another common approach for modeling univariate time series models and are presented in the form:

$$X_t = \mu + \theta_1 X_{t-1} + \cdots + \theta_q X_{t-q} + w_t \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-14)$$

Where: w_t are the white noise process, μ is the mean of the process and $\theta, \theta_1, \ldots$ are parameters of the model and X_t is the time series.

The backward shift notation of an MA model is denoted by:

$$X_t = \left(\theta_1 B + \theta_1 B^2 + \cdots + \theta_q B^q\right) X_t + \varepsilon_t \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-15)$$

Where: $\theta_1(B) = \varepsilon_t$

$$\theta(L) = (1 - \theta_1 B + \theta_1 B^2 + \cdots + \theta_q B^q) \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-16)$$

where B is the backshift operator.

We seek the coefficient θ so as to minimize the error function. If $\theta = 0$ then X_t is a white noise process, else X_t is auto correlated. Large values of θ lead larger correlation and negative values of θ result in oscillatory time series. Auto-correlation on MA is only for the first lag. It should be taken into account that the error terms after the model is fit should be independent and follow the assumptions for a univariate process. Box and Jenkins popularized an approach that combines both MA and AR approaches (Box, Jenkins, and Reinsel, 1994)

3.7.4 ARMA AND ARIMA MODELS

The process $\{X_t, t \in \mathbb{Z}\}$ is an ARMA (p, q) process if $\{X_t\}$ is stationary and if for every t

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-17)$$

where $Z_t \sim WN(0, \sigma^2)$

The first part of the ARMA equation above is the Auto Regression (AR model) the relationship between X_t and past or lagged observations $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$. The second part is the Moving Average (MA model) we can model that the time series at time t is not only affected by the shock at time t but also shocks that have taken place before time t. AR coefficients are denoted

with $\phi_1, \dots, \phi_p$ and the MA coefficients are denoted by $\theta_1, \dots, \theta_q$ these are unknown parameters, but if the process has a non-zero mean then mean (μ) the mean of the process is a parameter along with the variance of the white noise $\{Z_t\}$. The MA is a Linear combination of the White Noise. An AR model is of order p and an MA model is of order q. $\{X_t\}$is said to be an ARMA (p, q) process with mean μ if $\{X_{t-\mu}\}$ is an ARMA(p, q) process, practically if the MA order is zero then we only have an AR process inversely is true. The equation above is an ARMA process with mean = 0, if the mean = μ then practically $X_{t-\mu}$ is an ARMA process.

ARMA model in more compact form

$$\phi(\mathrm{B})X_t = \theta(\mathrm{B})Z_t \dots\dots\dots\dots\dots (eq\ 3-18)$$

Where $\phi(Z) = 1 - \phi_1 Z - \cdots - \phi_P Z^P$

and $\theta(Z) = 1 + \theta_1 Z + \cdots + \theta_q Z^q$

The polynomials are called autoregressive and moving average polynomials respectively, the operator B is a lag operator sometimes written as L. The ARMA models are the core of modelling time series, ARMA models forms an important family of stationary time series for a number of reasons

- For any autocovariance function $\gamma(\cdot)$ such that $\lim_{h\to\infty} \gamma(h) = 0$ and any integer k > 0, it is possible to find an ARMA process with autocovariance $\gamma_X(\cdot)$ such that $\gamma_x(h) = \gamma(h)$ for $h = 0,1,2, \dots, k$. The condition that the covariance function decreases as we decrease the lag of h means that the depended on the time series becomes smaller and smaller as we increase the lag between two time points of the time series if the condition hold then it is possible to find an ARMA process with the same autocovariance function that is there exist an ARMA process for any autocovariance function with this property.
- The Linear structure of ARMA models makes prediction easy to carry out. Importantly we estimate the parameters assuming fixed orders p and q but when we fit ARMA process it is uncommon to know the exact worse and thus we need to import the vigorously approach to identify the orders p and q. Not all formulation $\phi(\mathrm{B})\alpha_t = \theta(\mathrm{B})Z_t$ model stationary time series. A stationary solution to the ARMA equation exists and is unique if and only if $\phi(Z) \neq 0\ for\ all\ Z \in \mathbb{C}$ such that $|z| = 1$. That is, it exists and is unique if and only if no zeroes of $\phi(Z)$ lie on the unit circle.

Necessary and sufficient conditions for an ARMA model to have stationary solution should be established. The question to address is when X_t is stationary given its representation as an ARMA process or when the ARMA model generates stationary process, the condition only involves AR polynomial ϕ and insists that this polynomial does not have solutions on the unit circle or that the solutions of the polynomial $\phi(Z)$ are not equal to one in absolute value thus when checking whether an ARMA process is stationary we simply need to get the solution to the polynomial and check whether they are on the unit circle.

Causal Process:

An ARMA process $\{X_t\}$ is a causal function of $\{Z_t\}$ if there exists constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \; for\; all\; t \in \mathbb{Z}$. Suppose $\{X_t\}$ is an ARMA process for which $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeroes. Then $\{X_t\}$ is causal if and only if $\phi(Z) \neq 0 \; for\; all\; Z \in \mathbb{C}$ such that $|z| = 1$. The coefficient of the causal function is given by $\psi(Z) = \sum_{j=0}^{\infty} \psi_j Z^j = \frac{\theta(Z)}{\phi(Z)}$, $|Z| \leq 1$. Given a time series probabilistic model usually we find multiple ways to represent it which represent to choose depends on the context of our problem. However not all ARMA process can be transformed from one representation to another.

The concept of causal process means that we can invert an AR or ARMA process into an MA process where the order can be up to ∞. An ARMA process $\{X_t\}$ is causal if we can express it as a linear process in $\{Z_t\}$ which is a white noise such that the sum of absolute values of the coefficients of the linear process is finite, the simplest example is an MA process of order q. When does an ARMA process causal? Similarly, to the property of stationary not all ARMA process are causal. First, we assume that the two polynomials ϕ and θ do not have any common zeroes that means the equation $\phi(Z)$ are not equal to the solutions polynomial $\theta(Z) = 0$ then a necessary and sufficient for X_t to be causal is that the zeroes of the AR polynomial are outside of the unit circle in other words if any Z which is a solution to the equation $\phi(Z) = 0$ then that Z is larger than 1 in absolute value. This condition implies that if the process is causal it is also stationary but not the reverse. Given the two polynomial of an ARMA we can then obtain the coefficient ψ of the representation using the relationship $\psi(Z) = \sum_{j=0}^{\infty} \psi_j Z^j = \frac{\theta(Z)}{\phi(Z)}$, $|Z| \leq 1$ which is an invertibility property of the AR polynomial. One simple example of a causal process

is the AR of order 1. The AR(1) process $X_t(1-\phi \mathrm{B}) = Z_t$ thus $\phi(Z) = 1-\phi(Z)$ this has only one zero at $Z = \frac{1}{\phi}$ a unique stationary solution to exists if and only if $|1/\phi| \neq 1$ or $|\phi| \neq 1$. $\{X_t\}$ is a causal process if and only if $\left|\frac{1}{\phi}\right| > 1$ or $|\phi| < 1$

Invertible Process:

An ARMA process $\{X_t\}$ is invertible if there exist constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty}|\pi_j| < \infty$ and $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}\ for\ all\ t\ \in \mathbb{Z}$. Suppose $\{X_t\}$ is an ARMA process for which $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeroes. Then $\{X_t\}$ is invertible if and only if $\theta(Z) = 0\ for\ all\ Z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficient for an invertible process is given by $\pi(Z) = \sum_{j=0}^{\infty} \pi_j Z^j = \frac{\phi(Z)}{\theta(Z)}$, $|Z| \leq 1$. We can also go from an MA or ARMA process to an AR process where the order can be ∞ meaning that the ARMA process is invertible, more specifically for processes to be invertible they exist constants $\{\pi_j\}$ where the sum of the absolute value of the constants is finite such that the linear combination of the past X_{t-j} weight by $\{\pi_j\}$is a white noise $\{Z_j\}$. The necessary and sufficient condition for an ARMA process to be invertible is that the zeroes of the MA polynomial $\theta(Z)$ are outside the unit circle that is the solutions to the equation $\theta(Z) = 0$. Given the two polynomials of an ARMA process we can then obtain the coefficients π of the linear representation using the relationship $\pi(Z) = \sum_{j=0}^{\infty} \pi_j Z^j = \frac{\phi(Z)}{\theta(Z)}$, $|Z| \leq 1$, which is an invertible property of the MA polynomial. In practice the invertible ARMA processes are preferred because if we can invert an MA process to an AR process, we can find the value of Z_t which is not observable basing on past values of X_t which are observable. If a process is non invertible then in order to find the value of Z_t we have to know all the future values of X_t. This is an acronym for Autoregressive Integrated Moving Average model. Estimates in this model are three items namely autoregressive (*p*), Integrated (trend - *d*) and Moving Average (*q*). The ARIMA model of a Time series is defined by three terms (*p, d, q).*

The process of finding integers usually very small (e.g. 0,1 or 2), values of *p, d* and *q* is actually the identification of a time series and this model the patterns in the data. When the value is zero, the element is not needed in the model. The middle element, *d,* is investigated before *p* and *q* the goal being to determine if the process is stationary and, if not, make it stationary before

determining the values of *p* and *q*. It is worth noting that a stationary process has a constant mean and variance over the time period of the study.

3.7.6 BOX JENKINS METHODOLOGY

The ARIMA model is the Generalization of the ARMA model in both statistics and econometrics, mainly in the Time Series Analysis. The Box-Jenkins methodology named after the statisticians George Box and Gwilym Jenkins uses ARIMA models to find the best fit of a time series to past values of this Time Series model to make forecast. It is applicable usually when data show evidence of non-stationarity where an initial differencing (Integrated) step can be applied to remove non-stationarity. The model is generally referred to as an ARIMA (*p, d, q*) model where *p, d* and *q* are non-negative integers representing the AR, MA and the integrated part of the model. The Box-Jenkins has three stages to consider:

1. Model identification and model selection: checking if the time series data is stationary that is to say they is absence of components such as trends and seasonality is so the data need to be differenced and if the variances are not constant over time the use of logarithmic function to stabilize the data may be applied, and plotting of acf and pacf graphs to identify orders of *p and q*.
2. Parameter estimation: The most common approaches to estimate parameters of an ARIMA model is to use the use maximum likelihood estimation or non-linear least-squares estimation.
3. Diagnostic checking: a number of tests are to be done to the model to check if it is adequate enough to make forecasts some of these tests include the Ljung-Box test and plotting acf and pacf of the residuals and if the model is inadequate, the process is repeated from step 1 to build a better model. Below is a diagram to summarize the Box Jenkins ARIMA process

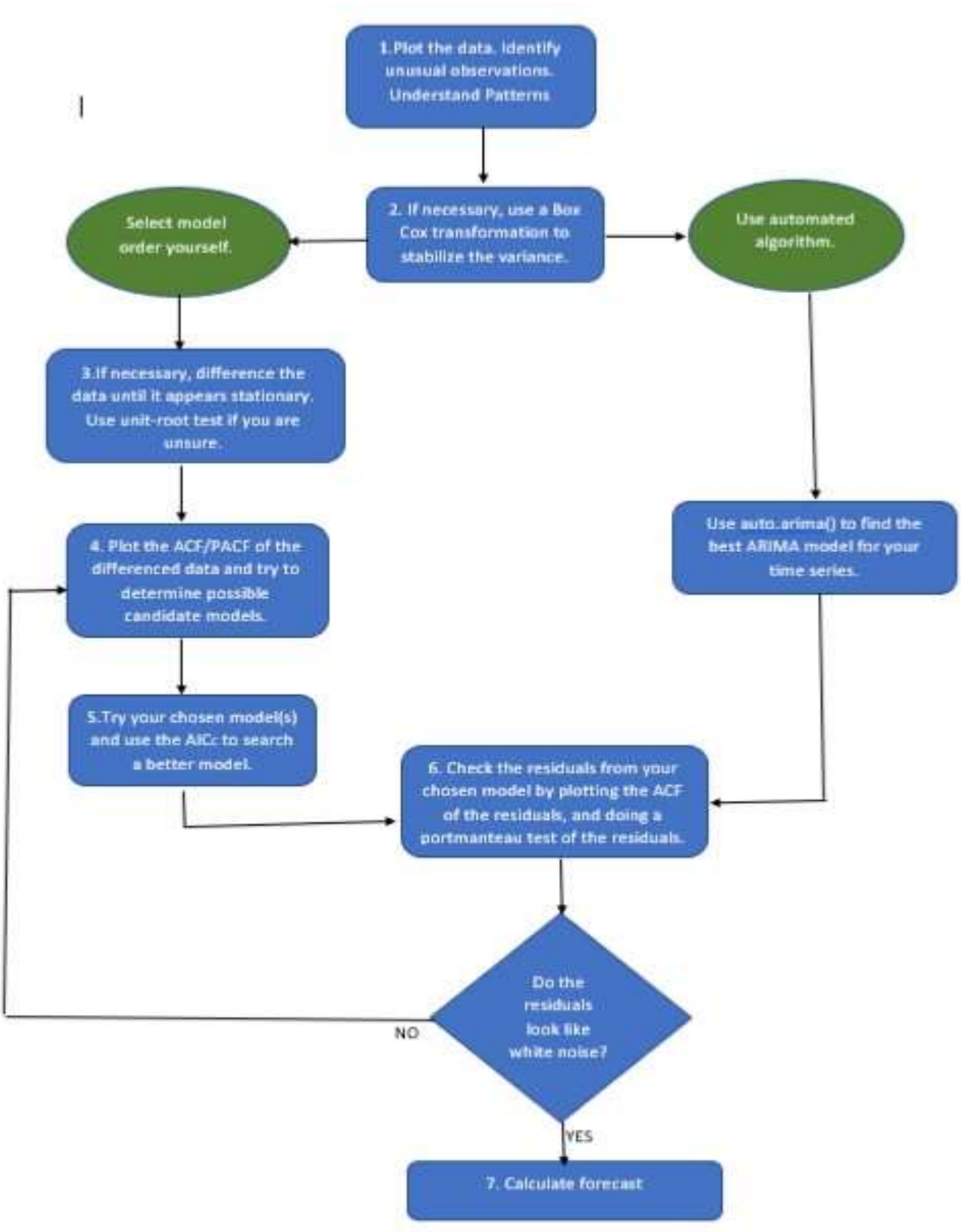

Figure 3.1 Box Jenkins Approach Summary

3.8.1 DIAGNOSTIC CHECKING

3.8.1 TEST FOR STATIONARITY

Augmented Dickey-Fuller Test

ARIMA models assumes that the time series is stationary. The Augmented Dickey-Fuller (ADF) test is then used to test to check if the time series is stationary or it is used to test the unit root in a time series data. The higher the negative value of the ADF becomes, the higher the probability of rejecting that there is a unit root at some level of confidence. Testing procedures for the ADF test are:

$$\Delta y_t = \gamma Y_{t-1} + \sum_{j=1}^{p} (\sigma_j \delta Y_{t-j}) + \varepsilon_t \ldots\ldots\ldots\ldots\ldots (eq\ 3-19)$$

$$\Delta y_t = \alpha + \gamma Y_{t-1} + \sum_{j=1}^{p} \left(\sigma_j \delta Y_{t-j}\right) + \varepsilon_t \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-20)$$

$$\Delta y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{j=1}^{p} (\sigma_j \delta Y_{t-j}) + \varepsilon_t \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-21)$$

A random walk may be imposed if we let α and β to become zeros specifically $\beta = 0$ ~ a random walk to have a drift. The following hypothesis guides the test:

H_0: Yt is a random walk, $\gamma = 0$

H_1 : Yt is a stationary process, $\gamma < 0$

H_0 : Yt is a random walk around a drift, $\gamma = 0, \alpha \neq 0$

H_1 : Yt is a level stationary process, $\gamma < 0, \alpha \neq 0$

H_0 : Yt is a random walk around a trend, $\gamma = 0, \beta \neq 0$

H_1 : Yt is a trend stationary process, $\gamma < 0, \beta \neq 0$

It has the following equation

$$DF_T = \frac{\hat{Y}}{SE(\hat{Y})} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-22)$$

Then DF_T is compared to the critical value of the Dickey-Fuller test and reject the null hypothesis if the DF_T value < critical value and conclude that there is no unit root.

3.8.2 KWIATKOWSKI-PHILLIPS-SCHMIDT-SHIN KPSS

To test a null hypothesis whether a time series is stationary around a deterministic trend we use the KPSS test and if the KPSS value is less than 0.05 we conclude that the times series is non-stationary, and if the p-value > 0.05, we fail to reject the null hypothesis.

3.8.3 GOODNESS OF FIT TEST

Akaike Information Criterion (AIC) considers the statistical goodness-of-fit and the number of parameters to be estimated to achieve a specific degree of fit, penalizing if number of parameters increase. Lower values of the index display the ideal model i.e. the one with lesser parameters that still provides an adequate/suitable fit to the data (Everitt, 1998).

$$AIC = log\hat{\sigma}_k^2 + \frac{n+2k}{n} \ldots\ldots\ldots\ldots\ldots (eq\ 3-23)$$

Where

$$\hat{\sigma}_k^2 = \frac{SSE_k}{n} \ldots\ldots\ldots\ldots\ldots (eq\ 3-24)$$

k is the number of parameters in the model.

The Hyndman-Khandakar Algorithm use the AIC bias corrected (AICc)

$$AICc = log\hat{\sigma}_k^2 + \frac{n+k}{n-k-2} \ldots\ldots\ldots\ldots\ldots (eq\ 3-25)$$

Where:

$\hat{\sigma}_k^2$ is given in the equation

k number of parameters in the model

n is the sample size

3.8.4 TESTING THE RESIDUALS

The Ljung-Box Test

The technique is equivalent to model validation of a non-linear squares fitting and it is under an assumption that the error term follows the supposition of a univariate process and that the residual is a white noise. If these assumptions are met then the model is a good model, however if it is not, another model is selected, Box-Ljung (1978) tests whether any of the group of the

autocorrelations of a time series are different from zero, it tests the complete randomness based on a number of lags, this is the portmanteau test.

H_0: $\rho = 0$ The data is independently distributed i.e. no correlations

H_1: $\rho \neq 0$ the data is not independently distributed

The Ljung-Box statistic

$$Q = n(n+2)\sum_{k=1}^{n}\frac{\hat{\rho}_k^2}{n-k} \dots\dots\dots\dots\dots (eq\ 3-26)$$

Where

n is the sample size (number of data points after any differencing done)

$\hat{\rho}_k^2$ is the sample autocorrelation at lag k

k is the number of lags being tested.

The choice of a plausible model depends on its p-value, if the p-value is above 0.05, signifying "non-significance." In other words, the bigger the p-value, the better the model.

3.9.1 FORECASTING PERFORMANCE CRITERION (MODEL ACCURACY)

Forecasting accuracy is a measure to assess the performance of forecasting models and it is a reverse value to the measure of forecasting error. They are more ways of calculating the measure of forecasting error and each measure produce a bit different information. The forecasting error is an expression of the deviation of forecasted values and actual values $\varepsilon_t = A_t - \hat{F}_t$.

Forecasting accuracy is measured in following ways:

- Mean Square Error (MSE)

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(A_t - F_t)^2 \dots\dots\dots\dots\dots (eq\ 3-27)$$

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{\sum_{i=1}^{n}|A_t - F_t|}{n^2} * 100 \dots\dots\dots\dots\dots (eq\ 3-28)$$

- Mean Absolute Deviation (MAD)

$$MAD = \frac{\sum_{i=1}^{n} |A_t - F_t|}{n} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-29)$$

- Root Mean Square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(A_t - F_t)^2} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-30)$$

The fitness of MSE, RMSE and MAD measures is relatively alike. There difference is not much take into account that strong errors are penalized by RMSE less than by others, MAE and RMSE signify a scale dependent measure, while others are not. All these measures are appropriate for evaluation of various forecasting models on the same test data.

MAPE is the widely used forecasting error measures when data sets are of different periods or have different scale, it expresses the percentage error, and makes it easy to interpret.

3.10.1 COVARIANCE AND CORRELATION

Covariance and correlation are two commonly applied terms both the fields of statistics and probability theory. The basic understanding of simple terms like mean, standard deviation, correlation, sample size and covariance are a characteristic of most reading articles on probability and statistics.

“Covariance” indicates the direction of the linear relationship between variables while “correlation” is essentially a measure of both the strength and direction of a linear relationship between two variables. In essence though; the two terms, covariance and correlation are highly comparable and work closely. Correlation is a function of variance, both terms measure the relationship and the dependency between two variables. One can obtain the correlation coefficient of the two variables by dividing the covariance of the two variables by the product of the standard deviation of the same values. Standard deviation actually measures the absolute variability of a data set distribution. When one divides the covariance values by the standard deviation, it significantly scales the value to a limited range of -1 to 1 which is essentially the range of correlation values. However, what sets them apart is the fact that correlation values are

standardize, whereas covariance values are not standardized. The formulas to calculate the covariance and the correlation for two variables X and Y are as follows:

$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \ldots\ldots\ldots\ldots\ldots (eq\ 3-31)$$

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{S_x S_y} \ldots\ldots\ldots\ldots\ldots (eq\ 3-32)$$

The Pearson correlation coefficient ranges from -1 to 1 and a correlation strength gets stronger as it approaches 1, and positive correlation indicates that variables move in the same direction i.e. if one variable increase the other one increase vice versa is true if one variable decrease and the other one decrease and a negative correlation means that variables moves in opposite directions one variable increase and the other decrease, they is no specific basis for interpreting the correlation coefficients. Akoglu (2018) suggest the following Table 3.1 below on how best to interpret the correlation coefficients

Table 3.1 Correlations Interpretations

Interpretation of the Pearson's and Spearman's correlation coefficients.				
Correlation Coefficient		**Dancey & Reidy (Psychology)**	**Quinnipiac University (Politics)**	**Chan YH (Medicine)**
+1	**−1**	Perfect	Perfect	Perfect
+0.9	**−0.9**	Strong	Very Strong	Very Strong
+0.8	**−0.8**	Strong	Very Strong	Very Strong
+0.7	**−0.7**	Strong	Very Strong	Moderate
+0.6	**−0.6**	Moderate	Strong	Moderate
+0.5	**−0.5**	Moderate	Strong	Fair
+0.4	**−0.4**	Moderate	Strong	Fair
+0.3	**−0.3**	Weak	Moderate	Fair
+0.2	**−0.2**	Weak	Weak	Poor
+0.1	**−0.1**	Weak	Negligible	Poor
0	**0**	Zero	None	None

3.11.1 REGRESSION ANALYSIS (ORDINARY LEAST SQUARES)

3.11.1 THE GENERAL LINEAR MODEL

General linear models (GLM) are extensively used in data analysis in almost areas of science. Some of the GLM are:

1. Simple linear regression: one response and one predictor
2. Multiple regression: multiple regression is not the same as multivariate regression.
    - Multiple regression: only one response and several predictors.
    - Multivariate regression: more than one response variable and predictors could be one or more.

3.11.2 ASSUMPTIONS OF LINEAR REGRESSION MODEL

Linearity: We draw a scatter plot of residuals and y values and If the scatter plot follows a linear pattern not a curvilinear pattern that shows that linearity assumption is met.
Independence: This assumption in mostly taken into account when we have longitudinal dataset that is the dataset is one where we collect observations from the same entity over time, for instance stock price data here we collect price info on the same over time. We generally have two types of data: cross sectional and longitudinal. Cross sectional datasets are those where we collect data on entities only once. Normality: we draw a histogram of the residuals, and then examine the normality of the residuals. If the residuals are not skewed, that means that the assumption is satisfied.
Equality of variance: We also use a scatter plot to check equality of variance. If the residuals do not fan out in a triangular manner that means that the equal variance assumption is met.

3.11.3 SIMPLE LINEAR REGRESSION MODEL AND ORDINARY LEAST SQUARES (OLS)

Simple Linear regression aim to determine the relationship between a single regressor variable X and a response variable Y. The regressor variable X is usually assumed to be a continuous variable and the expected value of Y for each value X is

$$E(Y/X) = \beta_0 + \beta_1 X \quad \ldots\ldots\ldots\ldots\ldots (eq\ 3-33)$$

where the parameters of the straight line β_0 and β_1 are unknown constants hence assume that each observation Y can be described by the model

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots (eq\ 3-34)$$

where ε is a random error term with mean zero and variance σ^2, $\varepsilon \sim N(0, \sigma^2)$

If we have n pairs of data (Y_1, X_1), (Y_2, X_2), $\ldots(Y_n, X_n)$ we may estimate the model parameters β_0 and β_1 by least squares. From equation (3-33), we have $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j, \quad j = 1,2, \ldots n$

And the least square function is

$$L = \sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} (Y_j - \beta_0 - \beta_1 X_j)^2 \quad \ldots\ldots\ldots\ldots\ldots (eq\ 3-35)$$

Minimizing the least squares function is simplified if we write the model equation (eq 3-33) as

$Y = \beta_0^1 + \beta_1 (X - \bar{X}) + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots (eq\ 3-36)$

Where:

$$\bar{X} = \left(\frac{1}{n}\right) \sum_{j=1}^{n} X_j \ldots\ldots\ldots\ldots\ldots (eq\ 3-37)$$

And $\quad \beta_0^1 = \beta_0 + \beta_1 \bar{X}$

Equation (eq 3-34) is frequently called the transformed simple linear regression model applying the transformed model, the least squares function becomes

$$L = \sum_{j=1}^{n} [Y_j - \beta_0^1 - \beta_1 (X_j - \bar{X})]^2 \quad \ldots\ldots\ldots\ldots\ldots (eq\ 3-38)$$

The least squares estimators of β_0^1 and β_1 say $\hat{\beta}_0^1$ and $\hat{\beta}_1$ must satisfy

$$\frac{\partial L}{\partial \beta_0^1}|_{\hat{\beta}_0^1 \hat{\beta}_1} = -2 \sum_{j=1}^{n} \left[Y_j - \hat{\beta}_0^1 - \hat{\beta}_1 \left(X_j - \bar{X}\right)\right] = 0 \quad \ldots\ldots\ldots\ldots\ldots (eq\ 3-39)$$

$$\frac{\partial L}{\partial \beta_1}|_{\hat{\beta}_0^1 \hat{\beta}_1} = -2 \sum_{j=1}^{n} \left[Y_j - \hat{\beta}_0^1 - \hat{\beta}_1 \left(X_j - \bar{X}\right)\right] = 0 \ldots\ldots\ldots\ldots\ldots (eq\ 3-40)$$

Simplifying the above equations gives

$$\beta_0^1 = \frac{1}{n}\sum_{j=1}^{n} Y_j = \bar{Y} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-41)$$

$$\hat{\beta}_1 \sum_{j=1}^{n} (X_j - \bar{X})^2 = \sum_{j=1}^{n} Y_j \left(X_j - \bar{X}\right) \ldots\ldots\ldots\ldots\ldots (eq\ 3-42)$$

Equations (eq 3-40) and (eq 3-41) are called the least squares normal equations and the solutions are:

$$\beta_0^1 = \frac{1}{n}\sum_{j=1}^{n} Y_j = \bar{Y} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-43)$$

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} Y_j\,(X_j - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-44)$$

The least squares estimators of the intercept and slope are $\hat{\beta}_0^1$ and $\hat{\beta}_1$ respectively, hence the fitted simple linear regression model is:

$$\hat{Y} = \hat{\beta}_0^1 + \hat{\beta}_1 (X - \bar{X}) \ldots\ldots\ldots\ldots\ldots (eq\ 3-45)$$

These are the Least squares estimators. They are other complex methods also called least square methods, the method applied above is denominated ordinary least square (*OLS*), due to its simplicity. In the precedent epigraphs $\hat{\beta}_1$ and $\hat{\beta}_0^1$ have been used to designate generic estimators and onwards, we will only designate *OLS* estimators with this notation.

3.11.4 SOME CHARACTERISTICS OF OLS ESTIMATORS

The algebraic implications of the estimation are derived exclusively from the application of the OLS process to the simple linear regression model:

1. The sum of OLS residuals is equal to 0:

$$\sum_{t=1}^{n} \varepsilon_j = 0 \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-46)$$

Residual definition

$$\varepsilon_j = Y_j - \hat{Y}_j = Y_j - \hat{\beta}_1 - \hat{\beta}_0^1, \quad j = 1,2,\dots,n \ \dots\dots\dots\dots\dots (eq\ 3-47)$$

Summing up n observations:

$$\sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} \left(Y_j - \beta_0 - \beta_1 X_j\right) = 0 \ \dots\dots\dots\dots\dots (eq\ 3-48)$$

2. The OLS line at all times passes through the mean of the sample $(\bar{X}, \bar{Y})$

$$\bar{Y} = \beta_0^1 + \hat{\beta}_1 \bar{X} \ \dots\dots\dots\dots\dots (eq\ 3-49)$$

3. The sample cross product between each of the independent variables and the OLS residuals is zero

$$\sum_{j=1}^{n} X_j \varepsilon_j = 0 \ \dots\dots\dots\dots\dots (eq\ 3-50)$$

Which implies that

$$\sum_{j=1}^{n} X_j \varepsilon_j = \sum_{j=0}^{n} X_j \left(Y_J - \hat{\beta}_0^1 - \hat{\beta}_1 X_j\right) = 0 \ \dots\dots\dots\dots\dots (eq\ 3-49)$$

4. The cross product between the fitted values and the OLS residuals is zero

$$\sum_{j=1}^{n} \hat{Y}_j \varepsilon_j = 0 \ \dots\dots\dots\dots\dots (eq\ 3-50)$$

3.11.5 ASSUMPTIONS OF OLS

- The regression model is linear in the coefficients and error term
- The error term (ε) has a population mean (μ)=0
- No correlation between the error term (ε) and the independent variables
- No correlation amongst the errors (ε)
- The error term has a constant variance (no heteroscedasticity)
- No independent variable is a perfect linear function of other explanatory variables
- The error term is normally distributed

3.11.6 STATISTICAL PROPERTIES OF OLS ESTIMATORS

Some of the OLS assumptions possess some ideal properties, thus *OLS* are the best linear unbiased estimators.

Linearity and unbiasedness of the OLS

The *OLS* estimator $\hat{\beta}_1$ is unbiased. Similarly, one can show that the *OLS* estimator $\hat{\beta}_0^1$ is also unbiased. consider the fact, unbiasedness is a general property of the estimator, but we might "close" or "far-off" from the true parameter in a sample. Hence, its distribution will be centered at the population parameter in most cases.

Variances of the OLS estimators

Because the sampling distribution of our estimator is lies about the true parameter. Then, how spread out is this distribution? The variance of an estimator is a gauge of the accurateness of the estimator. To estimate the variance of $\hat{\beta}_0^1$ and $\hat{\beta}_1$ it must be assumed that they are no autocorrelation and the error term has a constant variance.

$$var(\hat{\beta}_0^1) = \frac{\sigma^2 n^{-1} \sum_{j=1}^{n} X_j^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \qquad var(\hat{\beta}_1)$$

$$= \frac{\sigma^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \ldots\ldots\ldots\ldots\ldots (eq\ 3-51)$$

3.11.7 HYPOTHESIS TESTING IN SIMPLE LINEAR REGRESSION

To test hypothesis about the slope and intercept of the regression model, assume additional assumption about the error term

$$\varepsilon_j \sim N(0, \sigma^2)$$

They are independently and normally distributed with mean zero and variance σ^2. To test the hypothesis that the slope equals some value say $\beta_{1,0}$ the hypothesis to test are

$$H_0: \beta_1 = \beta_{1,0}$$

$$H_0: \beta_1 \neq \beta_{1,0}$$

If $\varepsilon_j \sim N(0, \sigma^2)$ then $Y_j \sim N(\beta_0 + \beta_1 X_j, \sigma^2)$. Therefore $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, $\hat{\beta}_1$ is independent of MSE and MSE is the error or the residual mean square

Formulas to consider:

$$SSE = \sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} \left(Y_j - \hat{Y}_j\right)^2 \ldots\ldots\ldots\ldots\ldots (eq\ 3-52)$$

$$S_{yy} = \sum_{j=1}^{n} \left(Y_j - \bar{Y}_j\right)^2 \ldots\ldots\ldots\ldots\ldots (eq\ 3-53)$$

$$S_{xx} = \sum_{j=1}^{n} X_j^2 - \frac{(\sum X)^2}{n} \ldots\ldots\ldots\ldots\ldots (eq\ 3-54)$$

$$S_{xy} = \sum_{j=1}^{n} X_j Y_j - \frac{(\sum X)(\sum Y)}{n} \ldots\ldots\ldots\ldots\ldots (eq\ 3-55)$$

$$MSE = \frac{SSE}{n-2} \ldots\ldots\ldots\ldots\ldots (eq\ 3-56)$$

The statistic for normality assumption is given by

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{\frac{MSE}{S_{xx}}}} \ldots\ldots\ldots\ldots\ldots (eq\ 3-57)$$

And follows a t-distribution with $n-2$ degrees of freedom and we reject H_0 if

$$|t_0| > t_{\frac{\alpha}{2}, n-2} \ldots\ldots\ldots\ldots\ldots (eq\ 3-58)$$

To test hypothesis for the intercept

$$H_0: \beta_0 = \beta_{0,0}$$

$$H_0: \beta_0 \neq \beta_{0,0}$$

The test statistic is

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{MSE\ (\frac{1}{n} + \frac{\bar{X}}{S_{xx}})}} \ldots\ldots\ldots\ldots\ldots (eq\ 3-59)$$

Reject H_0 if:

$$|t_0| > t_{\frac{\alpha}{2}, n-2}$$

3.11.8 ANALYSIS OF VARIANCE (ANOVA) FOR TESTING SIGNIFICANCE OF REGRESSION

The regression sum of squares SSR is given by

$$SSR = \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y}_j)^2 \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-60)$$

$$S_{yy} = SSR + SSE \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-61)$$

And S_{yy} has $n-1$ degrees of freedom. SSR has 1 degree of freedom and SSE has $n-2$ degrees of freedom. Hypothesis:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

The F statistic:

$$F_0 = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} \sim F(\alpha, 1, n-2)\ distribution \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-62)$$

Table 3.2 The ANOVA Table

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SSR = \hat{\beta}_1 S_{xy}$	1	MSR	$\frac{MSR}{MSE}$
Error of Residual	$SSE = S_{yy} - \hat{\beta}_1 S_{xy}$	$n-2$	MSE	
Total	S_{yy}	$n-1$		

To test for significance of regression we use the test statistic

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-63)$$

Squaring the t_0 value

$$t_0^2 = \frac{(\hat{\beta}_1)^2 S_{xx}}{MSE} = \frac{\beta_1 \hat{S}_{xy}}{MSE} = \frac{MSR}{MSE} = F_0 \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-64)$$

Reject H_0 if $F_0 > F(\alpha, 1, n-2)$

3.11.9 MODEL ADEQUACY CHECKING: RESIDUAL ANALYSIS

In fitting any linear model, analysis of the residuals from a regression model
is necessary to determine the adequacy of the least squares fit. It is helpful to examine

- normal probability test (Test for normality): The plot must resemble a line and if this is the outcome, it is sufficient to test for the normality. Plotting of the histogram
- Residuals versus Fitted values (Test of independence): this is adequate to test for independence of residuals; the plot must be structureless.
- Test for constant mean and variance of the residuals: plotting residuals against (order of data) regressor variable can do the best for homogeneity of the mean and variance of the residuals. The plot of the residuals against the regressor should show that the mean varies closely to zero with a relatively constant variance

Regression models should never be used for extrapolation. Regression relationships are effective for values of the regressor variable inside the array of original data. As we move beyond the original range of X, we become less certain about the validity of the assumed model.

3.11.10 THE LACK-OF-FIT TEST

Regression models are to observe the unknown true function relationship of the data and mostly we would want to know if the model assumed is correct hence the goodness of fit of a regression model

Now we want to test the hypothesis that

H_0: The model adequately fits the data

H_0: The model does not fit the data

In this test we will partition the error into two parts

$$SSE = SSPE + SSLOF \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-65)$$

Where $SSPE$ is the sum of squares attributed to pure experimental error and $SSLOF$ is the sum of squares for the Lack of fit of the model

To find the $SSPE$ we need observations on Y for at least one level of X i.e. we need m distinct levels of X

$$SSPE = \sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju} - \bar{Y}_j)^2 \ \ldots\ldots\ldots\ldots\ldots.(eq\ 3-66)$$

There are $n - m$ degrees of freedom for $SSPE$

For Lack of fit the sum of squares is simply

$$SSLOF = SSE - SSPE \ \ldots\ldots\ldots\ldots\ldots.(eq\ 3-67)$$

With $m - 2$ degrees of freedom

The test statistic for lack of fit is

$$F_0 = \frac{MSLOF}{MSPE} \ \ldots\ldots\ldots\ldots\ldots.(eq\ 3-68)$$

We reject H_0 if $F_0 > F_{\alpha,m-2,n-m}$

This test procedure may be easily introduced into the analysis of variance conducted for the significance of regression and If the null hypothesis of model adequacy is rejected, then the model must be abandoned and attempts must be made to find a more appropriate model.

3.11.11 GOODNESS OF FIT: THE COEFFICIENT OF DETERMINATION

The quantity

$$R^2 = \frac{SS_R}{S_{yy}} = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^{n}(Y_j - \bar{Y})^2} \ \ldots\ldots\ldots\ldots\ldots.(eq\ 3-69)$$

is called the coefficient of determination and is often used to judge the adequacy of a regression model ($0 \leq R^2 \leq 1$).

In most cases to R^2 is referred to as the proportion of variability in the data explained or accounted for by the regression model. If the regressor x is a random variable so that y and x may be viewed as jointly distributed random variables, then R is just the simple correlation between y and x.

- If R^2 = 1, we say that the fitted model is perfect. That is all residuals are zero. What is the acceptable value of R^2?
- this depends on the scientific field from which the data is collected.
- Behavioral science may be collecting data reflecting human behavior would be very content to get an R^2 of 0.7

- Normally, values of $R^2 \geq 0.8$ are considered to show/indicate a good fit.

3.11.12 PREDICTION INTERVAL

A $100(1-\alpha)\%$ prediction interval on the mean of k future observations at X_0 is

$$\hat{y}_0 \mp t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{k}+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{xx}}\right)} \ldots\ldots\ldots\ldots\ldots (eq\ 3-70)$$

3.12.1 EVALUATING MODEL ACCURACY

Using the model build from the training data to predict the sales for the testing data and comparing to the actual sales values that are already available in the testing data (unexposed to the training data).

Model accuracy will be measured in the following ways.

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{\sum_{i=1}^{n}|A_t - P_t|}{n^2} * 100 \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-71)$$

- Mean Absolute Deviation (MAD)

$$MAD = \frac{\sum_{i=1}^{n}|A_t - P_t|}{n} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-72)$$

- Root Mean Square error

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(A_t - P_t)^2} \ \ldots\ldots\ldots\ldots\ldots (eq\ 3-73)$$

Where A_t is the actual values and P_t are the predicted values.

NOTE: Measures such as MAD and RMSE are scale dependent, if these values are huge, they may be misconstrued thus it is better to use relative measures such as the MAPE is the data sets are of different measures.

3.13.1 OTHER MACHINE LEARNING PREDICTIVE MODELS

Polynomial Regression

This implies a special case of linear regression. A polynomial equation is fitted on the data with a curvilinear relationship between the target variable and the independent variable. The field of statistics sees polynomial regression as a form of regression analysis, in fact the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. the value of the target value changes in a non-uniform manner in accordance with the predictor(s) in a curvilinear relationship. In polynomial regression we have the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n$$

β_0 is the bias while $\beta_1, \beta_2 \ldots \beta_n$ are the weights in the equation of the polynomial regression and n is the degree of the polynomial It is significant to note that the member of higher order terms increases with the increasing value of n, hence the equation becomes more complex.

Support Vector Machine (Regression)

This applies to machine learning where support vector machines (SVMs) also known as support vector networks, these are supervised learning models with associated learning algorithms which analyze data used for classification and regression analysis. This helps to find a hyperplane in an N-dimensional space (N the number of features that clearly classifies the data points). Support Vector classification is extended to solve regression problems, it becomes Support Vector Regression. The model products depend only on a subset of the training data since the cost function does not care about training points that lie beyond the margin analogously, the model produced depends entirely on a subset of the training data because the function ignores samples whose prediction is close to their target. Support Vector Regression has three different implications namely SVR, NuSVR, and LinearSVR. Linear SVR provides faster implementation

that SVR, but handles the linear kernel, whereas NuSVR deals with a slightly different formulation than SVR and LinearSVR.

Random Forest Regression

This is a supervised learning using ensemble learning method for regression. Ensemble learning method combines predictions from multiple machine learning algorithms to ensure accurate predictions than a single model. This is also one of the fastest machine learning algorithms giving accurate predictions for regression problems and works on a system that says a number of weakly predicted estimators when put together form a strong prediction and strong estimation. However, Random Forest Regression does not give precise continuous nature prediction and does not predict beyond the range in the training data and therefore may overfit datasets that are particularly noisy.

Decision Tree Regression

It is the simplest yet most powerful algorithm in machine learning and uses a flow chart like a tree structure to predict the output on the basis of input or situation described by a set of properties. It falls under the supervised learning in machine learning category and is used for the continuous output problem. Continuous output means the output of the result is not represented by just a discrete known set of numbers or values; therefore, it is not discrete. More so, it observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful, continuous output.

Holt`s Winters (Forecast)

Holt`s`-Winters is a model for time series behavior and models on three aspects of time series namely a typical value (average); a slope (trend) over time and a clinical repeating pattern (seasonality), the model uses exponential smoothing to encode lots of values for the present and the future.

SARIMA (Forecast)

Seasonal Autoregressive Integrated Moving Average (SARIMA) or Seasonal ARIMA is an extension of ARIMA that explicitly supports univariate time series with a seasonal component

and adds three hyper parameters to specify the Autoregression (AR), differencing or Integrated (I) and Moving Average (MA) for the seasonal component of the series as well as an additional parameter for the period of seasonality.

3.14.1 CHAPTER SUMMARY

This Chapter discussed the differentiation between forecasting and prediction, it also discussed on the approached to be used for prediction and forecasting. This chapter also discussed deeply the concept of time series ARIMA modelling and ordinary least squares regression modelling to be used both for forecasting and prediction.

CHAPTER 4: DATA ANALYSIS AND SIMULATIONS

4.1.1 INTRODUCTION

This chapter will address the research objectives as well as the research objective. This chapter is also divided into two sections namely:

SECTION A – This part will focus on results from the ARIMA forecasting model and all the analysis, outputs and graphs done on the section will be done in R Studio using R programming language.

SECTION B – This section will focus on results from the Regression prediction model and all the analysis, outputs and graphs done on this section will be done in the Jupyter Notebook using the python programming language.

ARIMA and OLS Regression models will be used to find the suitable forecasting and prediction models to forecast and predict sales for the first three months of 2019 and will be compared with the actual sales that occurred in the year of 2019. The main focus on this chapter is to try to compare the accuracy of the forecasting model and the prediction model. Other machine learning models will be build using the Jupyter notebook.

4.2.1 TRAINING AND TESTING SETS

Splitting data into training and test sets is a vital part of evaluating or checking any predictive or data mining models. In all the scenarios the training set is used to build the model and the testing set is for testing the data and the portion of the training set is always large than the portion testing set usually the defaults portioning if 80:20 (training to testing test) and the testing set is equal to or greater than 20% of the data set but not greater than the training set. The splitting of the date set will help us to measure the accuracy of the model how well our model behaves in the long run. In this study for both time series forecasting model and regression model the data will be divided into training and testing sets, the training set will range from 1 January 2018 to 31 December 2018 and the testing set will range from 1 January 2019 to 31 March 2019.

4.3.1 BEHAVIOR AND CHARACTERISTICS OF TIME SERIES DATA

4.3.2 PRELIMINARY ANALYSIS

For Univariate time series it is advisable to use a time series data recorded over a long period of time for forecasting, Meyler et al., (1988) suggested that at least 50 observations are sufficient enough for univariate time series forecasting. Problems would arise if few data points (observations) are used. Moreover, the time series may contain structural breaks when using a lengthy time series data which may be necessary to examining only a subset of the data i.e. splitting into training, validation and test sets or otherwise use dummy variables, only 60 sufficient degrees of freedom are need for statistical robustness and also a smaller data to evade structural breaks. A series plot is plotted against time to evaluate any structural breaks, data errors and outliers in the data set, this procedure also help to disclose if whether there is significant seasonal pattern in the time series. A dimension of the preliminary analysis for exploratory of the non-stationarity of the data is by taking into account the trend analysis plot of 364 days from 1 January 2018 to 31 December 2018 for Steer`s Zimbabwe sales as shown in Figure 4.1

4.3.3 TEST FOR STATIONARITY

Before building any Time series model it is first important to check for unusual behavior in the data hence, we first analyze the data to check for trends, seasonality or any random behavior. The study began with collected data of daily sales of fast-fast foods products from Steer`s Zimbabwe. To clearly visualize any patterns in the data a time series plot was used (figure 4.1). Figure 4.1 shows that the data exhibits an upward trend, the blue line is representing a constant mean line and the red line is the trend line and the upward trend can be due to the increased number of people buying fast-foods. From the time series plot it can be detected that high sales are experienced at the beginning and the end of each month and returns to normal the other days and Figure 4.1 clearly shows that the data is not stationary.

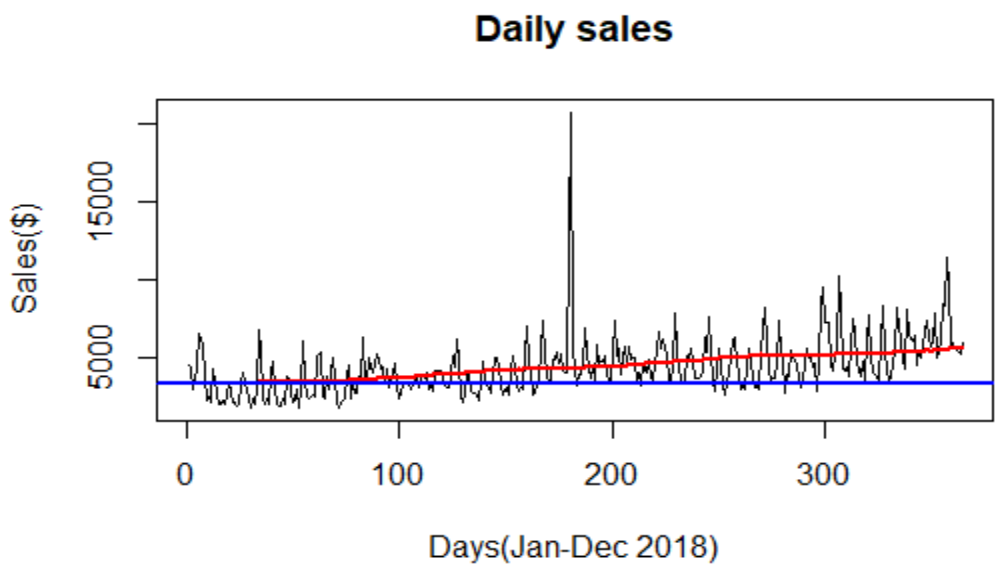

Figure 4.1 Time Series Plot

Figure 4.2 represents the box plot and the sales are plotted on monthly basis and it is clearly shown that the company sales spike at the end of the year and they are low at the beginning of the year and slightly rise from may onwards. The box plot can be explained because most people buy food during the December holiday and at the beginning of the year people will be spending mostly on schools opening hence low sales experience and high sales in May-June and august-September can be due to school holidays. The results obtained show an upward trend and the data can be interpreted as non-stationary.

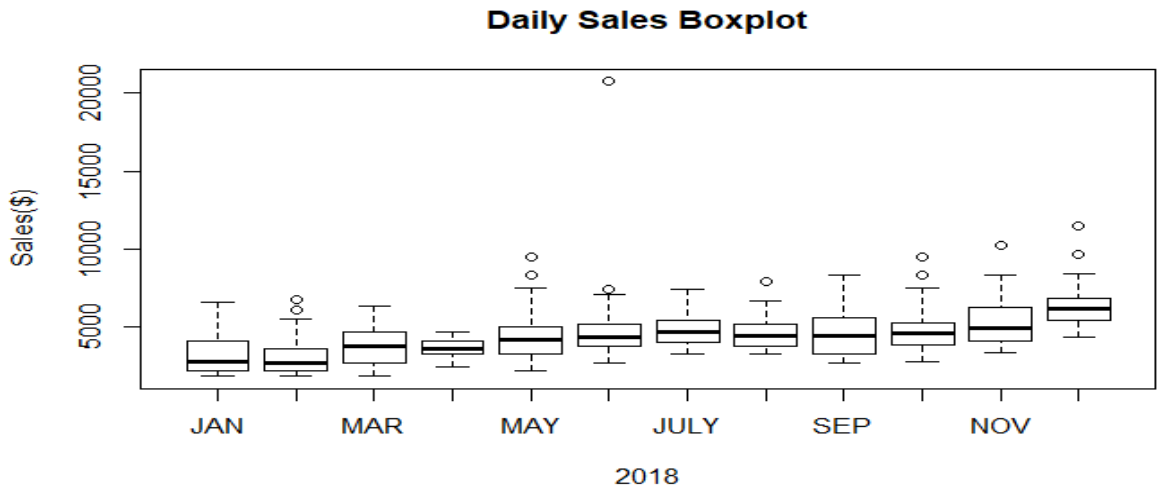

Figure 4.2 Box Plot (Sales)

Most researchers have contributed to the fact that visual checking is subjective, which tends to be true in most cases hence a Boxplot was plotted for a vivid picture on the non-stationarity of the data. The mean value in December is much higher than rest of the months. The Boxplot clearly indicates the non-stationarity of the data.

4.3.4 ADF AND KPSS TEST

Augmented Dickey-Fuller Test

Ho: Data is not Stationary

Dickey-Fuller = -4.938, Lag order = 7.1335, p-value = 0.01

alternative hypothesis: stationary

In adf.test (sales, alternative = c("s"), k = trunc(length(sales) -: p-value smaller than printed p-value. From the ADF test it can be concluded that the series is stationary

KPSS Test for Level Stationarity

Ho: Data are Stationary

KPSS Level = 3.5657, Truncation lag parameter = 5, p-value = 0.01

In kpss.test (sales): p-value smaller than printed p-value.

From the KPSS test it can be concluded that the series is not stationary.

4.3.5 AUTOCORRELATION FUNCTION (ACF) AND PARTIAL AUTOCORRELATION FUNCTION (PACF)

If an ACF plot fails to die off rapidly then it can be concluded that the data is non-stationary. Figure 4.3 shows that the acf plot shows correlation in most lags and the plot has failed to die off rapidly.

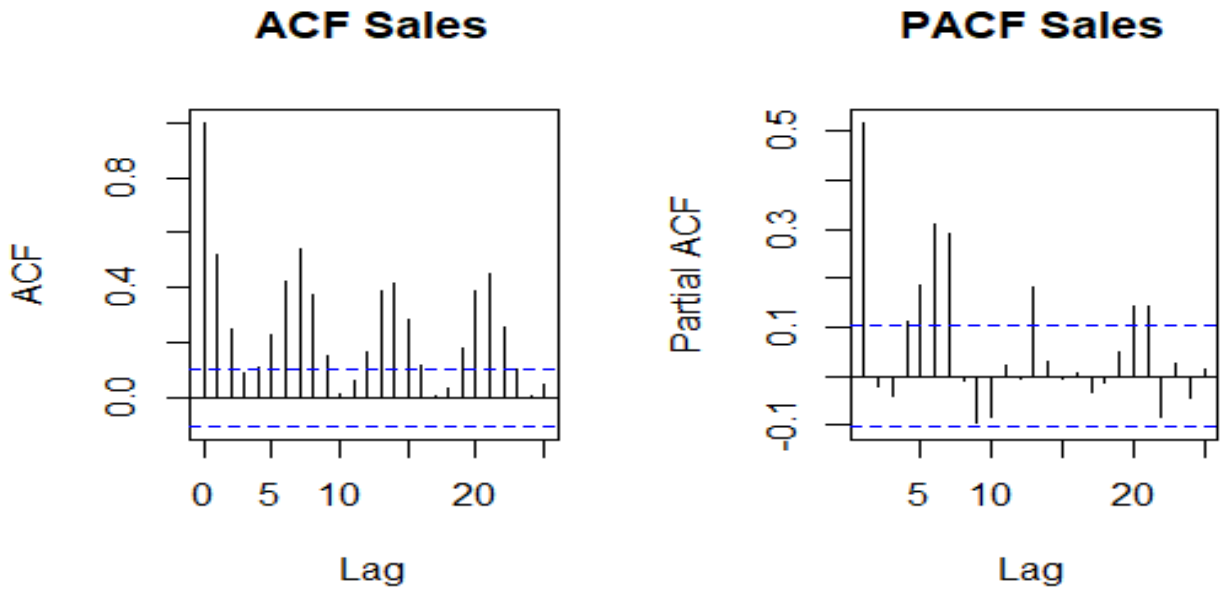

Figure 4.3 ACF and PACF Raw data

4.4.1 ARIMA MODELLING

Autoregressive Integrated Moving Average (ARIMA) models are of the form ARIMA (p, d, q). Where p is estimated from an AR model and q from an MA model, d is the Integrated part indicating the number of differenced times. To obtain the values of p and q the autocorrelation and partial autocorrelation function graphs of differenced data were used the following models from Table 4.1 were estimated using the "aicc", "aic" and the "bic" criterion:

Table 4.1 Estimated ARIMA models

Model	AIC
ARIMA(0,1,0)	: 6456.767
ARIMA(0,1,0) with drift	: 6458.787
ARIMA(0,1,1)	: 6376.001
ARIMA(0,1,1) with drift	: 6376.21
ARIMA(0,1,2)	: 6333.318
ARIMA(0,1,2) with drift	: Inf
ARIMA(0,1,3)	: 6330.195
ARIMA(0,1,3) with drift	: Inf
ARIMA(0,1,4)	: 6326.747

ARIMA(0,1,4) with drift	**: Inf**	
ARIMA(0,1,5)	**: 6324.536**	
ARIMA(0,1,5) with drift	**: 6324.355**	
ARIMA(1,1,0)	**: 6439.717**	
ARIMA(1,1,0) with drift	**: 6441.748**	
ARIMA(1,1,1)	**: 6334.699**	
ARIMA(1,1,1) with drift	**: Inf**	
ARIMA(1,1,2)	**: 6333.059**	
ARIMA(1,1,2) with drift	**: Inf**	
ARIMA(1,1,3)	**: 6330.807**	
ARIMA(1,1,3) with drift	**: Inf**	
ARIMA(1,1,4)	**: 6324.634**	
ARIMA(1,1,4) with drift	**: 6324.536**	
ARIMA(2,1,0)	**: 6431.223**	
ARIMA(2,1,0) with drift	**: 6433.264**	
ARIMA(2,1,1)	**: 6328.805**	
ARIMA(2,1,1) with drift	**: Inf**	
ARIMA(2,1,2)	**: 6319.158**	
ARIMA(2,1,2) with drift	**: 6319.838**	
ARIMA(2,1,3)	**: 6266.198**	*******
ARIMA(2,1,3) with drift	**: 6267.783**	
ARIMA(3,1,0)	**: 6405.819**	
ARIMA(3,1,0) with drift	**: 6407.867**	
ARIMA(3,1,1)	**: 6313.532**	
ARIMA(3,1,1) with drift	**: 6313.742**	
ARIMA(3,1,2)	**: 6310.213**	
ARIMA(3,1,2) with drift	**: 6312.185**	
ARIMA(4,1,0)	**: 6376.419**	
ARIMA(4,1,0) with drift	**: 6378.472**	
ARIMA(4,1,1)	**: 6308.76**	
ARIMA(4,1,1) with drift	**: 6310.272**	
ARIMA(5,1,0)	**: 6326.376**	
ARIMA(5,1,0) with drift	**: 6328.432**	
Best model: ARIMA(2,1,3)		

The AIC criterion was used to come up with the best model and the model with the smallest value of AIC was selected the AIC is used to test the goodness of fit of a model. From the above Table 4.1 estimated models the model with the least value of the AIC was selected, after estimation and checking of all the probable models, numeral potential models may sufficiently

represent the data but by seeing at the number of parameters the principle of parsimony is used to select a model that adequately fits the data. ARIMA (2,1,3) was considered the best model because it had the least AIC value.

4.5.1 MODEL STRUCTURE

Table 4.2 ARIMA model Structure

ARIMA(2,1,3)				
Coefficients:				
ar1	**ar2**	**ma1**	**ma2**	**ma3**
1.2188	**-0.9776**	**-1.9863**	**1.8184**	**-0.7281**
s.e. 0.0193	**0.0160**	**0.0616**	**0.0944**	**0.0486**
sigma^2 estimated as 1787176:		**log likelihood=-3126.98**		
AIC=6265.96	**AICc=6266.2**	**BIC=6289.33**		

As shown in the Table 4.2 above all the AR coefficients are significant this would indicate that the estimated coefficient is significantly different from zero. The estimated model shows that it has two AR models and three MA models and it is difference once. The estimated model is ARIMA (2,1,3)

4.6.1 DIAGNOSTIC CHECKING

The residuals were tested to check if the data follows the assumption of stationarity. A number of tests were performed including the Ljung-Box test, ACF and PACF plots, QQ plots, histogram and density charts.

The diagnostic analyses using the residual plot in Figure 4.4 shows that the residuals have a constant mean represented by the red line which is overshadowed by the blue constant mean line

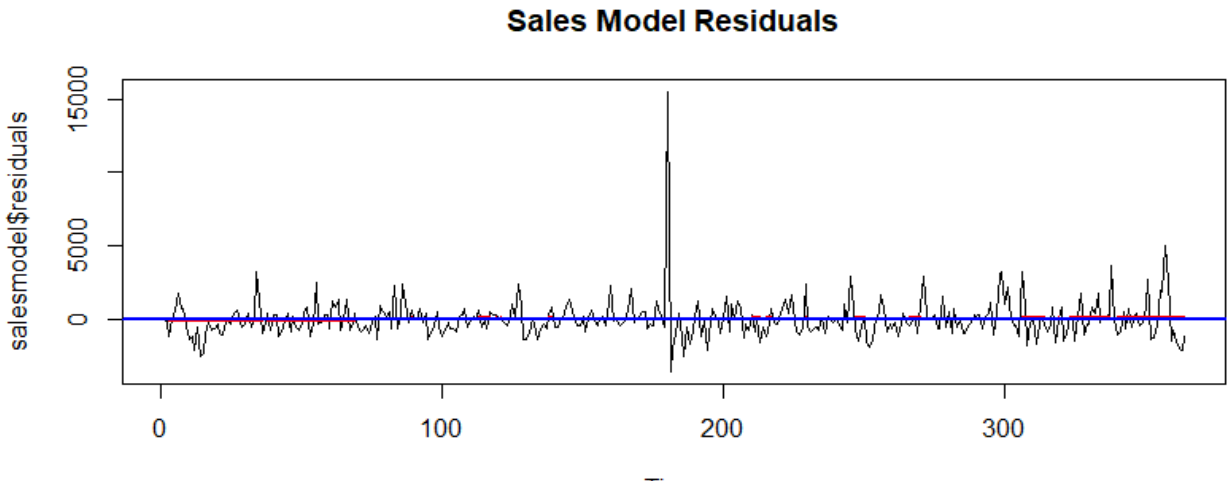

Figure 4.4 Residual Plot

and this indicates that the residuals are stationary and the series does not violate the assumption of constant mean and variance.

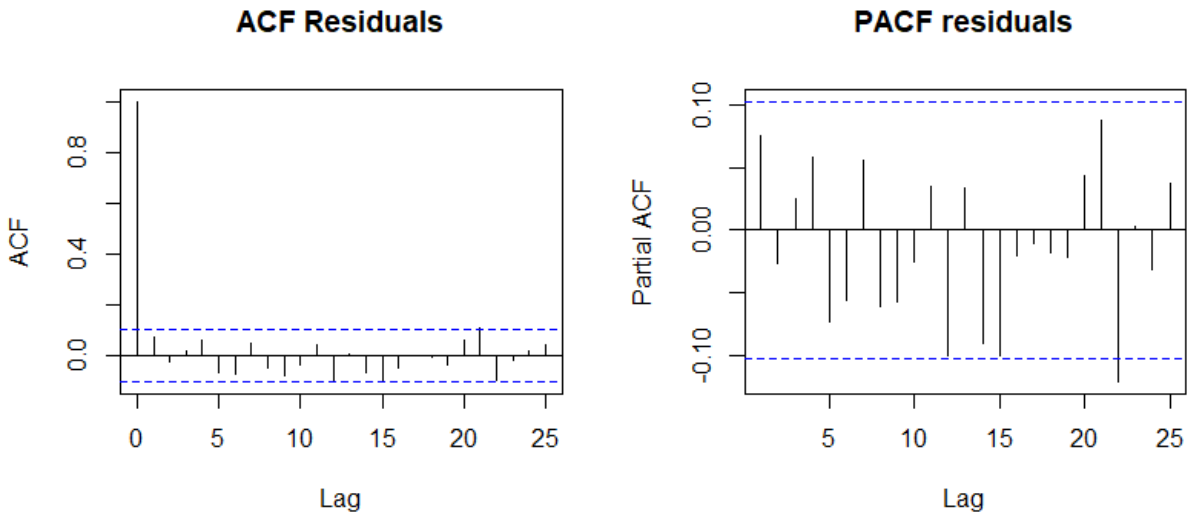

Figure 4.5 ACF and PACF Plots for Residuals

The autocorrelation (ACF) plot displays that for the first 25 lags, all sample autocorrelations fall inside the 95 % confidence bounds signifying the residuals appear to be random also the PACF except for lag 22. The ACF of the residuals illustrates that autocorrelation of the residuals are all zeros i.e. all the lags are uncorrelated thus concluded that the residuals have constant variance

the selected model (ARIMA(2,1,3)) and the true mean of the residuals is approximately zero. Hence, the selected model (ARIMA(2,1,3)) satisfies all the model assumptions.

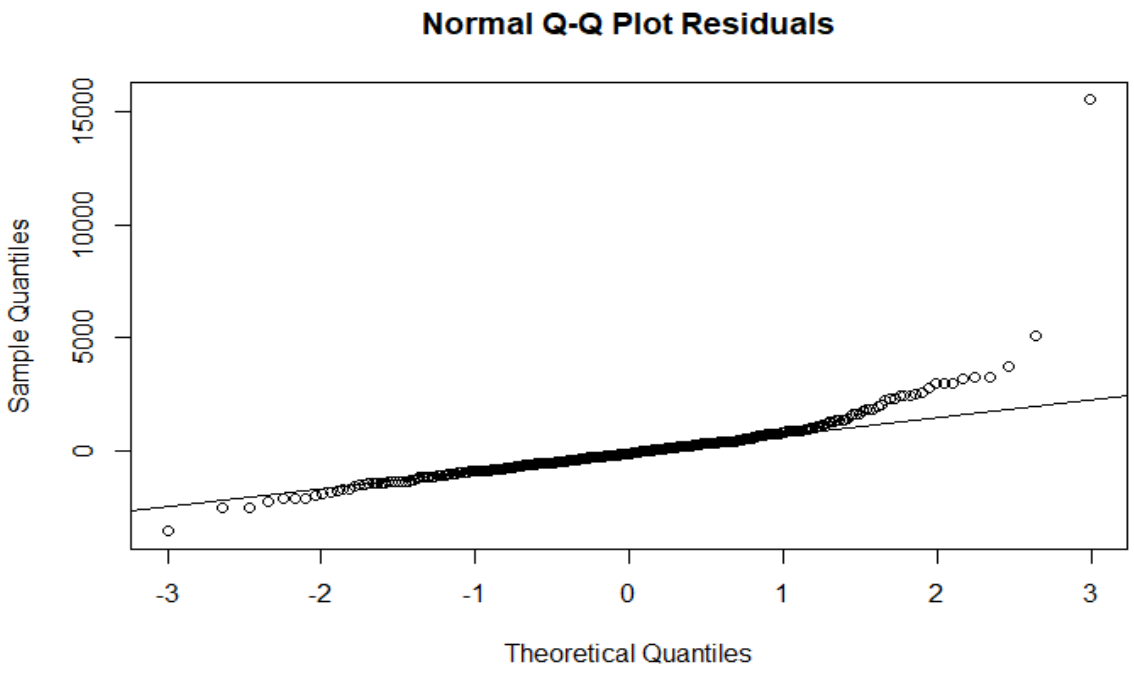

Figure 4.6 Normal QQ plot for Residuals

Figure 4.6 shows the normal Q-Q plot, the plot indicates that the normal distribution gives an adequate fit for this model though the extreme values slightly tail off most of the values are on the line showing that the residuals are normally distributed.

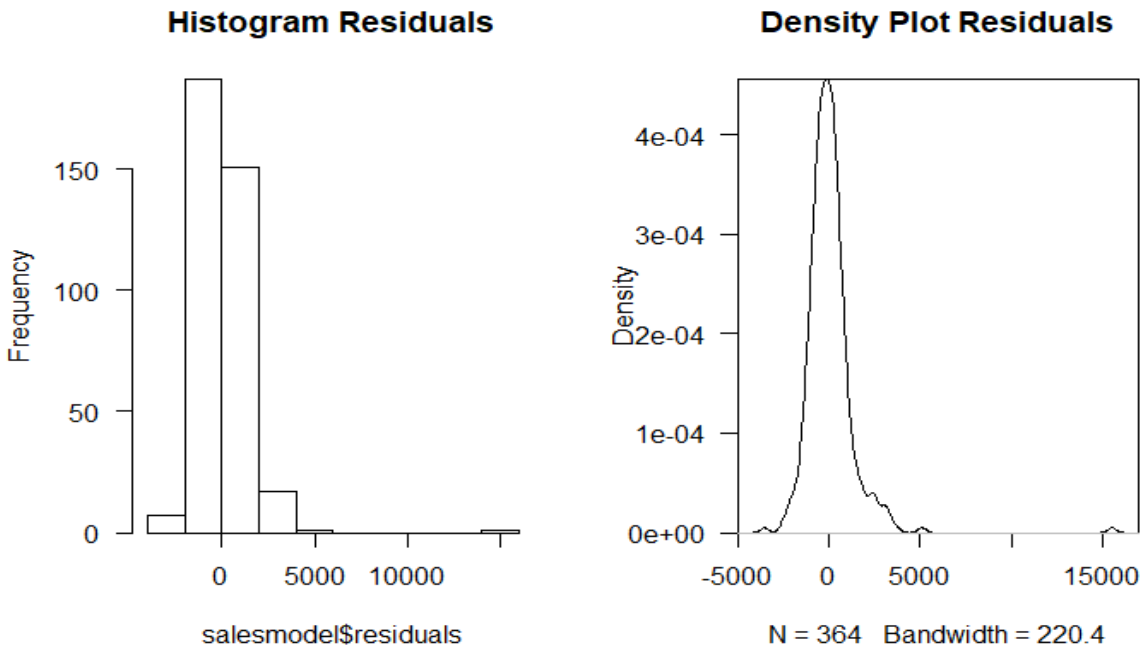

Figure 4.7 Histogram and Density Plots for Residuals

The histogram and the density chart on Figure 4.7 above show that the residuals have a bell like shape distribution this indicates that the residuals follow a normal distribution.

The Box-Ljung test for the residuals from the ARIMA (2,1,3) model was applied to determine if the residuals were random.

Box-Ljung test

Table 4.3 Box-Ljung Test for Residuals

X-squared = 2.0419, df = 1, p-value = 0.153
X-squared = 11.529, df = 10, p-value = 0.3178
X-squared = 21.454, df = 15, p-value = 0.1229
X-squared = 33.59, df = 25, p-value = 0.117
X-squared = 67.829, df = 58, p-value = 0.177
X-squared = 95.781, df = 95, p-value = 0.4583
X-squared = 155.21, df = 134, p-value = 0.1015
X-squared = 283.5, df = 261, p-value = 0.1618

The above Table 4.3 shows that all different lags tested the p-values is always greater than 5% indicating that the residuals are random and that the model gives an adequate fit to the sales data.

Meanwhile the ARIMA (2, 1, 3) satisfies all the essential assumptions, it can be concluded that the model provides an adequate fit of the fast-food sales data. Henceforth, the forecasting model would be formulated from the parameter estimates in Table 4.2.

4.7.1 FORECASTING

One of the key factors in decision making for some companies includes sales forecasting, thus it plays a vital role and it helps to have a vivid insight into the future basing on past and current behavior of the data. Box and Jenkins (1976) gives the basis for economic and business planning, optimization of industrial processes, inventory and production control. Earlier researches shows that in most scenarios the considered best model is not necessarily the one that proves the best forecast and this has widen the study to test for the accuracy of the model using test such as mean square error (MSE), Root Mean Square Error (RMSE) and others must be performed to the model.

4.7.2 FORECASTING MODEL

4.7.3 ARIMA FORECASTED SALES

Table 4.4 summarizes the forecasted values of Steer`s Zimbabwe over the period of 1 January 2019 to 31 March 2019 on daily basis with 95% confidence level using the ARIMA (2, 1, 3)

Table 4.4 ARIMA Forecasted Values

Date	Actual $	Point Forecast $	Lo 95 $	Hi 95 $
1/1/2019	5747.15	6665.216	4045.032	9285.399
2/1/2019	4455.15	6309.739	3619.651	8999.827
3/1/2019	4579.95	6035.116	3320.846	8749.386
4/1/2019	5040.3	6047.9	3331.083	8764.716
5/1/2019	5106.2	6331.945	3614.409	9049.482
6/1/2019	7595.95	6665.654	3938.073	9393.236

7/1/2019	9368.65	6794.717	4022.315	9567.119
8/1/2019	6255.95	6625.798	3778.415	9473.18
9/1/2019	3429.8	6293.744	3388.204	9199.284
10/1/2019	2175.4	6054.155	3128.435	8979.874
11/1/2019	2945.3	6086.741	3158.342	9015.14
12/1/2019	3193.25	6360.675	3430.895	9290.454
13/01/2019	3675.6	6662.701	3721.429	9603.973
14/01/2019	2591.85	6763.032	3779.345	9746.718
17/01/2019	811.15	6590.066	3540.659	9639.472
18/01/2019	1934.8	6281.167	3182.556	9379.779
19/01/2019	1926.05	6073.756	2957.893	9189.619
20/01/2019	2187.4	6122.927	3004.184	9241.67
21/01/2019	2811.35	6385.618	3264.724	9506.511
22/01/2019	2841.65	6657.728	3524.048	9791.408
23/01/2019	2941.6	6732.587	3558.628	9906.545
24/01/2019	2726.4	6557.82	3325.682	9789.957
25/01/2019	3940.4	6271.626	2997.296	9545.957
26/01/2019	6560.45	6093.65	2804.259	9383.042
27/01/2019	5349.45	6156.502	2863.976	9449.027
28/01/2019	4152	6407.092	3111.579	9702.604
29/01/2019	3735.95	6651.08	3341.65	9960.509
30/01/2019	3114.65	6703.491	3355.753	10051.23
31/01/2019	4060.4	6528.856	3129.225	9928.487
1/2/2019	4690.05	6264.768	2828.544	9700.991
2/2/2019	8287.4	6113.605	2663.983	9563.227
3/2/2019	7020.2	6187.528	2734.475	9640.58
4/2/2019	4607.95	6425.401	2968.492	9882.31
5/2/2019	3406.75	6643.065	3171.273	10114.86
6/2/2019	4333	6675.824	3167.579	10184.07
7/2/2019	3459.5	6502.969	2948.159	10057.78
8/2/2019	5699.5	6260.262	2673.394	9847.131
9/2/2019	9351	6133.42	2534.427	9732.413
10/2/2019	4074.15	6216.083	2613.328	9818.838
11/2/2019	3389.75	6440.835	2833.35	10048.32
12/2/2019	2260.8	6633.961	3010.789	10257.13
13/02/2019	3008	6649.64	2991.782	10307.5
14/02/2019	3927.85	6479.955	2780.097	10179.81
15/02/2019	4545.5	6257.808	2529.603	9986.012
16/02/2019	6429.3	6152.927	2413.579	9892.274
17/02/2019	5205.95	6242.259	2498.79	9985.727
18/02/2019	4357.55	6453.669	2704.61	10202.73

19/02/2019	4418.15	6624.016	2858.621	10389.41
20/02/2019	4011.3	6624.972	2826.581	10423.36
21/02/2019	4356.6	6459.611	2623.168	10296.05
22/02/2019	5320.85	6257.127	2395.4	10118.85
23/02/2019	8492.85	6171.985	2299.871	10044.1
24/02/2019	5133.3	6266.153	2389.538	10142.77
25/02/2019	5075.05	6464.162	2581.129	10347.19
26/02/2019	3059.45	6613.446	2713.573	10513.32
27/02/2019	2849.6	6601.832	2670.581	10533.08
28/02/2019	3406.4	6441.739	2475.869	10407.61
1/3/2019	5135.1	6257.965	2269.355	10246.58
2/3/2019	10261.1	6190.477	2192.062	10188.89
3/3/2019	4705.55	6287.872	2284.562	10291.18
4/3/2019	3041.45	6472.557	2462.043	10483.07
5/3/2019	2511.25	6602.446	2574.72	10630.17
6/3/2019	3057.65	6580.218	2522.663	10637.77
7/3/2019	2995.55	6426.148	2336.973	10515.32
8/3/2019	6267.45	6260.093	2150.301	10369.88
9/3/2019	8468.2	6208.311	2089.158	10327.46
10/3/2019	3759.65	6307.53	2183.08	10431.98
11/3/2019	3784.65	6479.082	2346.696	10611.47
12/3/2019	2882.2	6591.183	2441.333	10741.03
13/03/2019	2397.45	6560.111	2381.91	10738.31
14/03/2019	3116.25	6412.652	2205.46	10619.84
15/03/2019	6189.8	6263.299	2037.271	10489.33
16/03/2019	5916.9	6225.412	1990.358	10460.47
17/03/2019	4812.6	6325.239	2084.483	10566
18/03/2019	4411.6	6483.949	2234.582	10733.32
19/03/2019	2816.65	6579.803	2312.828	10846.78
20/03/2019	3227	6541.482	2247.563	10835.4
21/03/2019	2845.25	6401.071	2080.472	10721.67
22/03/2019	5523.25	6267.395	1929.459	10605.33
23/03/2019	7316.2	6241.726	1895.01	10588.44
24/03/2019	4424.9	6341.119	1988.298	10693.94
25/03/2019	3788.25	6487.356	2125.312	10849.4
26/03/2019	3406.5	6568.432	2188.73	10948.13
27/03/2019	3496.85	6524.292	2118.983	10929.6
28/03/2019	5286.55	6391.236	1961.281	10821.19
29/03/2019	5384.65	6272.211	1826.185	10718.24
30/03/2019	7866.2	6257.211	1802.58	10711.84
31/03/2019	7208.95	6355.284	1894.154	10816.42

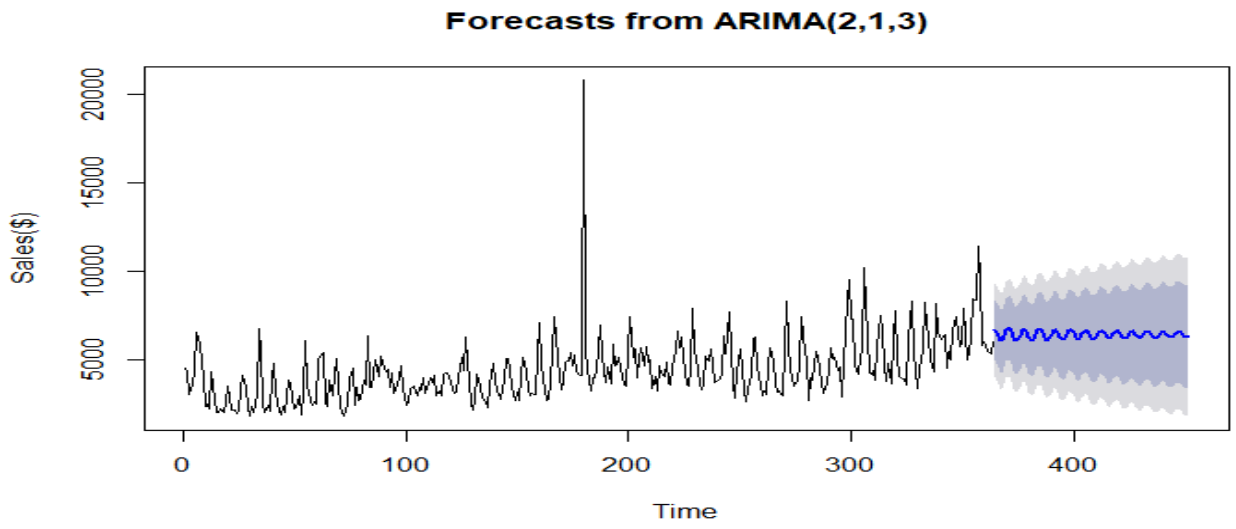

Figure 4.8 Forecasted Values Plot

4.8.1 MEASURING FORECASTING ACCURACY MEASUREMENT

One-step forecasts on test data

Fitting a model using training data has now become the most common and efficient practice, for evaluation of a model performance on a test data, the comparisons on the testing data use dissimilar forecast horizons. In our data we have used in this research, we have used the last 88 points for the test data, and estimated our forecasting model on the training data which run from 1 January 2018 to 31 December 2018 and the test data from 1 January 2019 to 31 March 2019, thus the forecast errors will be for 1-step, 2-steps, …, 88-steps ahead. The forecast horizon will usually increase with the forecast variances. We then obtain 1-step errors on the test data and use training data to estimate parameters, we use data preceding each other both on training and test data when computing forecast. 1,2,…,T-364 are the times for our training data we then compute our ARIMA model and compute Y_(T-60+h|T-61+h), for h = 1,..,T-1 the test set is not part of the parameter estimation, but still this provides a "fair" forecast. The Figure 4.9 below summarizes the results

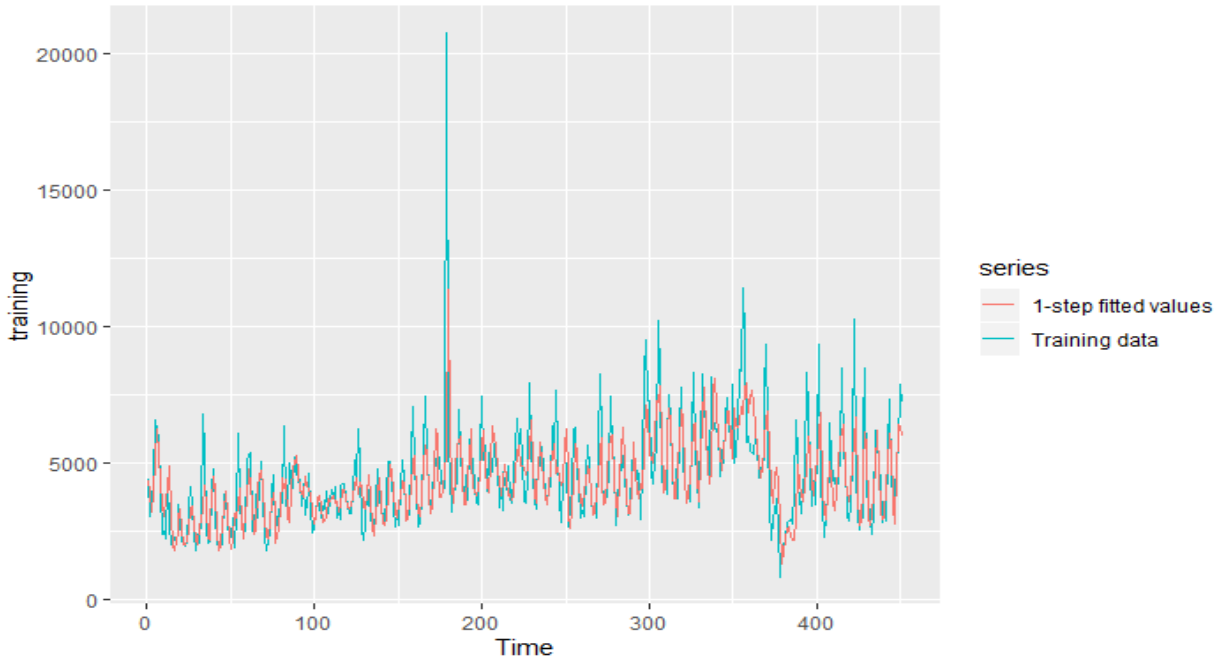

Figure 4.9 one-step fitted values

ACCURACY MEASURE

Table 4.5 ARIMA Model Accuracy measure

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	40.776	1325.789	804.934	-4.339	18.081	0.742	0.075

Table 4.5 represents the accuracy measurement for the ARIMA (2,1,3) model on Steer`s Zimbabwe sales data. Using the scale measurement unit RMSE it can be interpreted that the forecast can be ± $1325.79 away from the actual sales, using the MAE the forecast ± $804.93 away from the actual sales and using the MAPE it can be interpreted that the model has a percentage error of 18.08% meaning that the model forecasting accuracy percentage is 81.92% and can be concluded that this is a good model.

4.9.1 CORRELATION AND COVARIANCE ANALYSIS

Correlation Analysis is a measure of the strength of association between dependent and the independent variable together (Gujirati, 2004). In this research it is vital for us to study correlation between variables. The major objective of applying correlation is to test if multicollinearity is strong enough to nullify the concurrent presence of the explanatory variables in regression model and covariance test the direction of the linear relationship TableXX and XX below shows the correlation and the covariance matrix

Table 4.6 Correlation matrix

Correlation Matrix	Sales	Customers
Sales	**1.000000**	**0.771565**
Customers	**0.771565**	**1.000000**

The correlation matrix (Table 4.6) shows a Pearson Correlation coefficient value of $r = 0.77$ this indicates a strong positive relationship between the independent (Customers) and the dependent (Sales) variable at hand, the study then therefore tested the direction of the variable as shown in the covariances matrix below.

Table 4.7 Covariance Matrix

Correlation Matrix	Sales	Customers
Sales	**3.207400e+06**	**190396.813888**
Customers	**1.903968e+05**	**18985.462318**

The Table 4.7 above summarizes the covariance results and as mentioned earlier in chapter 3 covariance does not have a specific unit of measurement, hence only the sign of the figures in the matrix are considered for interpretation as shown in Table 4.7 above all the numbers are positive and this indicates that the variables are moving in the same direction i.e. if the sales increase the customers will also increase and vice versa is also true if the sales falls the customers will also fall. The movement and the direction of the variable will be displayed in a more visual method and Figure 4.10 will illustrate the representation.

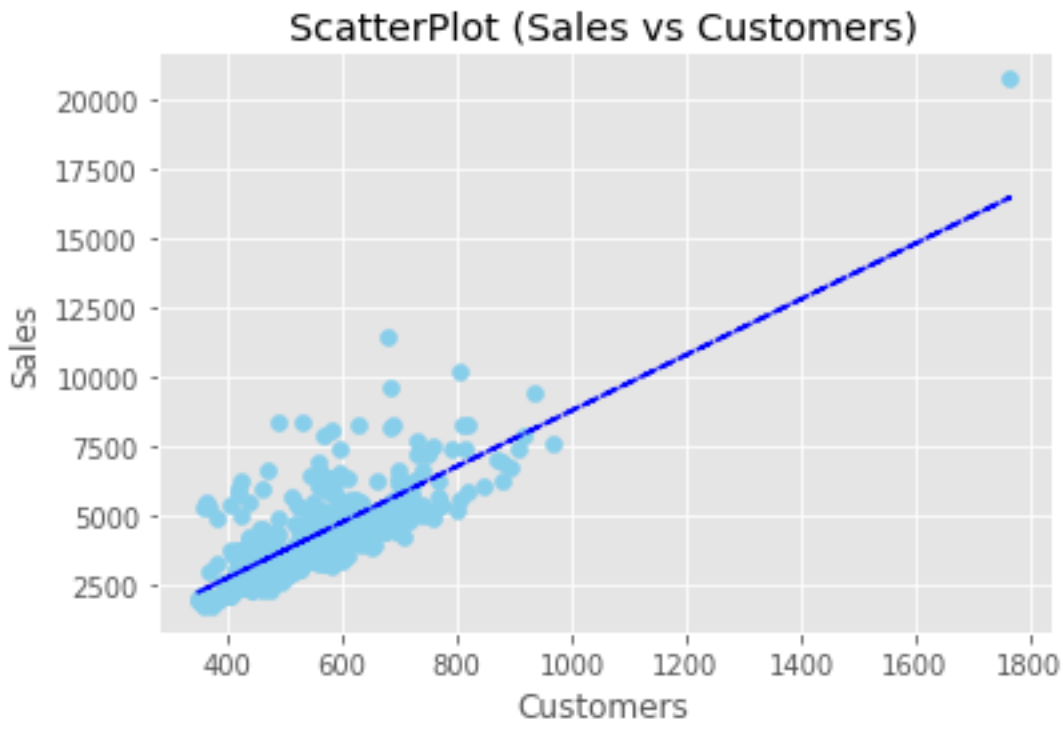

Figure 4.10 Scatter Plot (Sales vs Customers)

The blue dotted line is representing the line of best fit and it is clearly shown that the line is moving in a positive direction and the point are also scattered in the same direction as the line indicating positive relationship.

4.10.1 OLS REGRESSION ANALYSIS

A simple linear regression model is built and the Table 4.8 summarizes the model. The Juypter notebook using python programming language was used to build the model with the help of the following python libraries:

- NumPy – this package was used in the operations of arrays
- Pandas – this package was used in handling and importing the data frame into the notebook
- Matplotlib and seaborn – these packages were used in the construction of graphs
- SciPy – Used in some mathematical and probability computations such as the mean and the probplot for QQ plot
- Scikit learn – used in the splitting of the dataset and calculating the accuracy Metrix
- Stasmodels – used in the calculation of the OLS regression

The model will be in the form $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Where Y is the Profit, β_0 is the Intercept, $\beta_1$is the coefficient of the independent variable (Customers), X_1 is the Customers and ε is the standard error

Table 4.8 OLS Regression Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.595
Model:                            OLS   Adj. R-squared:                  0.594
Method:                 Least Squares   F-statistic:                     532.5
Date:                Sat, 11 Jul 2020   Prob (F-statistic):           4.20e-73
Time:                        06:59:06   Log-Likelihood:                -3077.9
No. Observations:                 364   AIC:                             6160.
Df Residuals:                     362   BIC:                             6168.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1223.1524    251.494     -4.864      0.000   -1717.725    -728.580
Customers     10.0286      0.435     23.076      0.000       9.174      10.883
==============================================================================
Omnibus:                      147.152   Durbin-Watson:                   0.438
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              470.730
Skew:                           1.878   Prob(JB):                     6.06e-103
Kurtosis:                       7.115   Cond. No.                     2.43e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.43e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

We need to test if the independent variable is relevant to our model or we need to drop them. In order to infer is a given variable is relevant or significant to the target variable we perform a t-test, the t-test behaves slightly different from a F-test its looks for the relationship between the target variable (dependent) and the predictor variable independently without taking into account other features if available, it does it one variable at a time. Now the t-test is performed between the customer variable and the target variable. The hypothesis will be represented as and the test is conducted at 95% confidence level

H_0: Variable constant value (customers) = 0

H_1: Variable constant value(customers) ≠ 0

Table 4.8 shows that the t-value = 23.076 is greater than the p value (P>| t | =0.000) we then reject the null hypothesis that the variable constant value is equal to 0 and we conclude that the

predictor variable (customers) is relevant in our linear regression model. obtained an R Square of .595 and a regression equation of SALES′ = −1223.1524 + 10.0286 (Customers). The ANOVA resulted in F = 532.5 with 1 and 362 degrees of freedom. The F is significant and a simple linear regression was calculated to predict Sales based on Customers. A significant regression equation was found (F(1,362) = 532.5, p < .001), with an R2 of .595. Sales predicted is equal to −1223.1524 + 10.0286 (Customers) when sales are measured in dollars. Customers average increased 10.0286 for each dollar of sale.

4.10.2 ANOVA TEST

The F-test is used for assessing the significance of the overall regression model In linear regression which is the case parent we only have one variable, it compares the model with one predictor in this case that will be the customer variable which is also referred to as intercept to the model in the case according to Table 4.9 will have the intercept value which is -1223.1524 with the specified model that we have with the variable Customer.

The hypothesis:

H_0: $\beta_1 = 0$ (the regression model is not significant)

H_1: $\beta_1 \neq 0$ (the regression model is significant)

We reject H_0 if F > PR(>F)

ANOVA For Testing Significance of Regression

Table 4.9 ANOVA Test for Significance

	Df	Sum_sq	Mean_sq	F	PR(>F)
Customers	1.0	6.931142e+08	6.931142e+08	532.51769	4.195273e-73
Residual	362.0	4.711718e+08	1.301580e+06	NaN	NaN

Since F (532.5) > PR(>F) (4.195e-73) we reject H_0 and conclude that $\beta_1 \neq 0$ (the regression model is significant)

4.10.3 MODEL ADEQUACY

Normal probability test (Test for normality):

The Quantile Quantile (QQ plot) plot is plotted and the plot must follow the QQ line and if this is the outcome, it is sufficient to test for the normality and we conclude that the residuals follow a normal distribution.

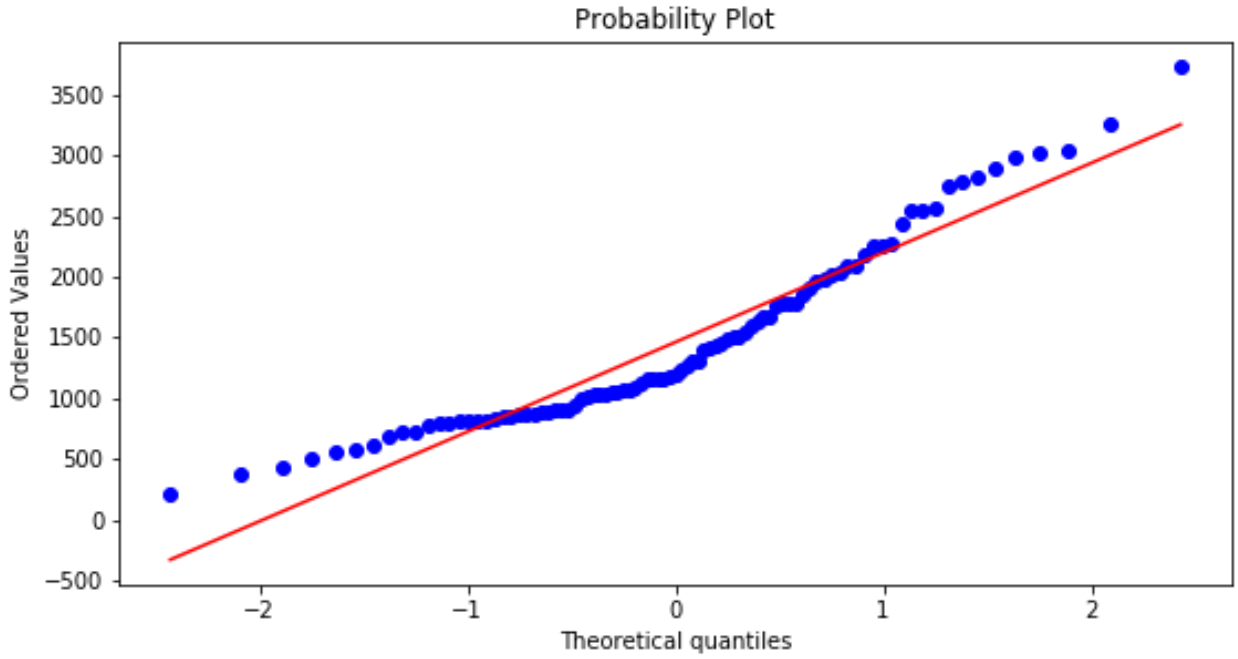

Figure 4.11 QQ Plots for regression Residuals

Figure 4.11 shows that, the Normal probability QQ plot almost follows a straight line as some values diverge from the expected value plot. Large sample data was used; thus, we can assume that the central limit theorem holds. This provides enough proof that the assumption of normality is not violated.

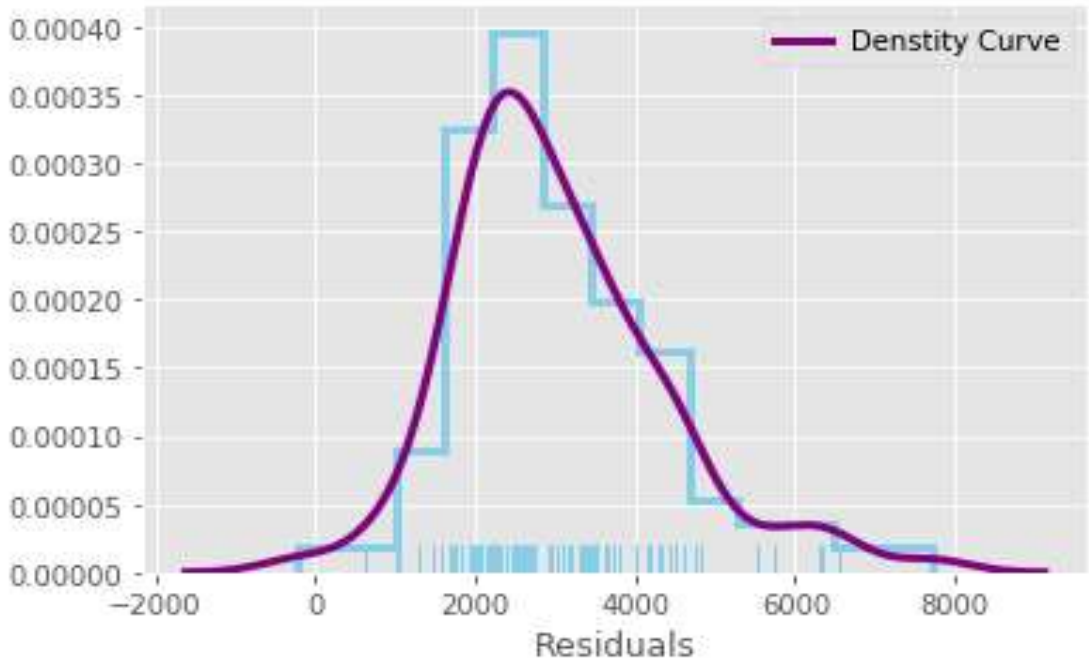

Figure 4.12 Histogram + Density Plot for Regression Residuals

The histogram above shows that the residuals are normally distributed and the density plot looks like a bell-shaped distribution providing further evidence that they are no violation of normality assumption but it may also be argued that the residuals are positively skewed basing on the visual inspection of the Figure 4.2 above. It's safe enough to say the whole distribution is near normal.

Homoscedasticity (equal variance) of residuals:

Plotting residuals against the regressor variable can do the best for homogeneousness of the mean and variance of the residuals. the plot of the residuals against the regressor should indicate that the mean diverges closely to zero with a relatively constant variance. If the assumption fails to hold the data will exhibit heteroscedasticity, this will make the determination of the true standard deviation of the prediction errors hard.

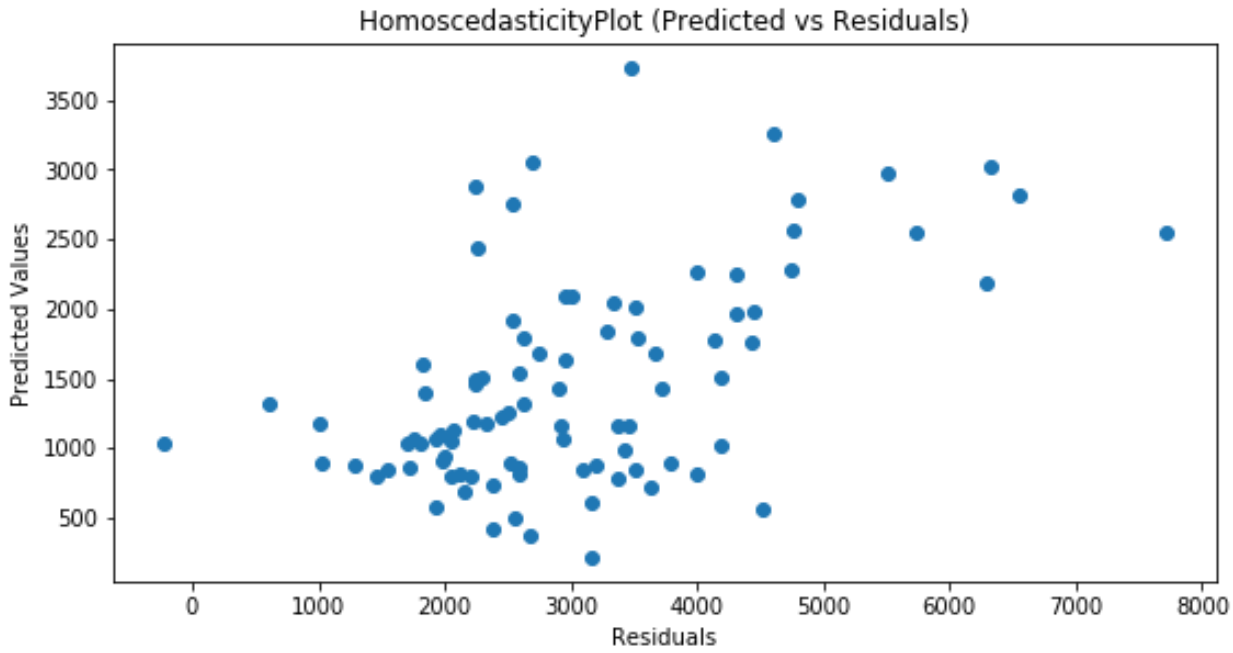

Figure 4.13 Residual vs Prediction Plot

We need to visualize to find out if the overall distribution is randomly sampled or does it increase in change of the residual values. Figure 4.13 shows that they are no patterns based on the residual and the prediction plot. There is no item that can be observed with an increasing variance with respect to increasing overall predicted values there is no patten observed

No autocorrelation of residuals:

This assumption is critically used in time series analysis and its states they must be no correlations amongst the residuals of the predicted model.

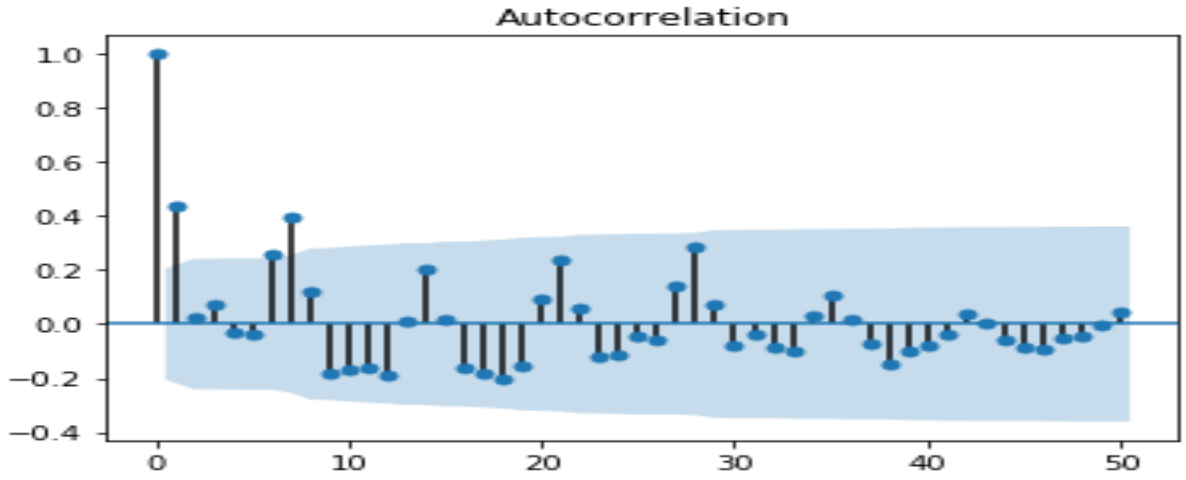

Figure 4.14 ACF plot

Figure 4.14 shows the first 50 lags of the ACF plot of the residuals and most the lags are not correlated as it is seen that most of the lags except 2 are below the 95% confidence bands.
All the above assumptions have concluded that our model adequately fits the data.

4.10.4 GOODNESS OF FIT (R^2)

The coefficient of determination denoted by R^2 is the proportion of variance in the dependent variable that is predictable from the independent variable. If the Linear model fits very well then, we have an R^2 value close to one there is also misconception that R^2 value cannot be negative. If we keep on adding variables to our model for better prediction accuracy then they are high chances that our R^2 will remain high close to one, however adjusted R^2 accounts for this by penalizing R^2 values that include non-useful predictors, so if we keep adding other variables which are of very little relevance to our overall model the adjusted R^2 value will go down, if the adjusted R^2 is much lesser than the R^2 value it's a sign that a variable may not be relevant to the overall model and we just have to find that variable and omit it. In the case of the underlying study the R^2 is 0.595 and the adjusted R^2 value is 0.594, so it's a very well fitted model.
Table 4.8 shows the R^2 and the adjusted R^2 values

4.11.1 PREDICTION

Regression Equation:

The prediction model will be in the form $Sales = -1223.152 + 10.029 * Customers + \varepsilon$

This model will be used to predict the sales on a 3-month period from 1 January 2019 to 31 March 2019

4.11.2 PREDICTED VALUES

Table 4.10 summarizes the predicted values of Steer`s Zimbabwe over the period of 1 January 2019 to 31 March 2019 on daily basis

Table 4.10 Predicted sales values

Date	Actual $	Predicted $	Date	Actual $	Predicted $
1/1/2019	5747.15	2698.01	**16/02/2019**	6429.3	4192.27
2/1/2019	4455.15	2537.56	**17/02/2019**	5205.95	3630.67
3/1/2019	4579.95	2948.73	**18/02/2019**	4357.55	2627.81
4/1/2019	5040.3	2948.73	**19/02/2019**	4418.15	2938.70
5/1/2019	5106.2	3008.90	**20/02/2019**	4011.3	3510.33
6/1/2019	7595.95	4804.01	**21/02/2019**	4356.6	3530.38
7/1/2019	9368.65	6548.98	**22/02/2019**	5320.85	5516.04
8/1/2019	6255.95	3991.70	**23/02/2019**	8492.85	3710.90
9/1/2019	3429.8	1825.53	**24/02/2019**	5133.3	4513.18
10/1/2019	2175.4	1294.02	**25/02/2019**	5075.05	2557.61
11/1/2019	2945.3	2126.39	**26/02/2019**	3059.45	2046.16
12/1/2019	3193.25	2066.21	**27/02/2019**	2849.6	2597.73
13/01/2019	3675.6	2447.30	**28/02/2019**	3406.4	3289.70
14/01/2019	2591.85	1725.24	**1/3/2019**	5135.1	7712.29
17/01/2019	811.15	220.30	**2/3/2019**	10261.1	2266.79
18/01/2019	1934.8	1033.27	**3/3/2019**	4705.55	2667.93
19/01/2019	1926.05	612.07	**4/3/2019**	3041.45	1935.84
20/01/2019	2187.4	1013.22	**5/3/2019**	2511.25	1965.93
21/01/2019	2811.35	2387.13	**6/3/2019**	3057.65	1935.84
22/01/2019	2841.65	2156.47	**7/3/2019**	2995.55	4302.58
23/01/2019	2941.6	2006.04	**8/3/2019**	6267.45	6288.24
24/01/2019	2726.4	1695.16	**9/3/2019**	8468.2	2497.44
25/01/2019	3940.4	2627.81	**10/3/2019**	3759.65	3169.36
26/01/2019	6560.45	4312.61	**11/3/2019**	3784.65	1975.96
27/01/2019	5349.45	3670.78	**12/3/2019**	2882.2	1554.76
28/01/2019	4152	3369.93	**13/03/2019**	2397.45	2387.13
29/01/2019	3735.95	2246.73	**14/03/2019**	3116.25	4432.95
30/01/2019	3114.65	2056.19	**15/03/2019**	6189.8	4142.13
31/01/2019	4060.4	3189.41	**16/03/2019**	5916.9	3991.70
1/2/2019	4690.05	3791.13	**17/03/2019**	4812.6	3420.07
2/2/2019	8287.4	5736.67	**18/03/2019**	4411.6	1755.33
3/2/2019	7020.2	4743.84	**19/03/2019**	2816.65	1835.56
4/2/2019	4607.95	3450.16	**20/03/2019**	3227	1815.50
5/2/2019	3406.75	2517.50	**21/03/2019**	2845.25	3510.33
6/2/2019	4333	2898.58	**22/03/2019**	5523.25	4753.87
7/2/2019	3459.5	2597.73	**23/03/2019**	7316.2	2748.16
8/2/2019	5699.5	4192.27	**24/03/2019**	4424.9	2286.84
9/2/2019	9351	2918.64	**25/03/2019**	3788.25	2216.64
10/2/2019	4074.15	3169.36	**26/03/2019**	3406.5	2326.96

11/2/2019	3389.75	1454.47	**27/03/2019**	3496.85	2537.56
12/2/2019	2260.8	2206.61	**28/03/2019**	5286.55	3339.84
13/02/2019	3008	3089.13	**29/03/2019**	5384.65	4613.47
14/02/2019	3927.85	3379.96	**30/03/2019**	7866.2	3480.24
15/02/2019	4545.5	4442.98	**31/03/2019**	7208.95	2597.73

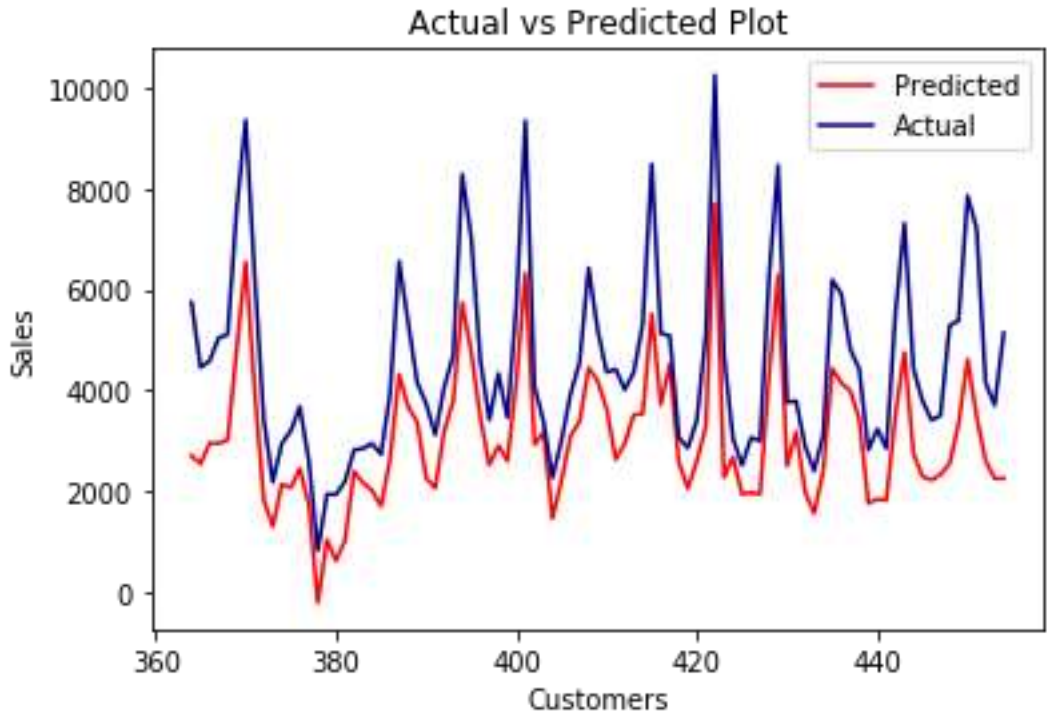

Figure 4.15 Actual vs Predicted

The plot above (Figure 4.15) shows the actual vs the predicted values and they is not big difference between the two.

4.11.3 EVALUATING PREDICTION MODEL ACCURACY

ACCURACY MEASURE

Table 4.11 OLS regression Accuracy measure

	MSE	RMSE	MAE	MPE	MAPE
	2699379.32	1642.979	1462.131	-4.339	16.067

Our model's MAE is $1462.13, which is impartially minor given our data's sales range. Table 4.11 shows that, MAPE states that our model's predictions are, 16.07% off from actual value meaning that our model has an accurate percentage of 83.93%. The RMSE measure shows

that the model is ±$1642.13 from the actual data which is a fairly small figure relating to our data range.

4.12.1 RESEARCH HYPOTHESIS

In this study the author mainly focused on RMSE and MAE as my measure of accuracy, as many researchers have different basis to measure and this depends on the researcher and this research has chosen RMSE and MAE because measures such as MAE and RMSE are scale dependent, and since the study is using same data set its more appropriate to use scale dependent measures, one of the two measures aims at the median and the other at the average.

RMSE method is more accurate, the MAE and the MAPE approaches were used before the RMSE and by squaring the errors we can get more accurate outcomes as the negative and positive errors don't cancel each other and stay in existence till the end of commutation, thus adding more accuracy to the model. RMSE is measured in the same units as the data, rather than in squared units and is representative of the size of a "typical" error.

Steps for the Hypothesis:

a) Stating the Hypothesis

H_0: Sales forecasted by the ARIMA model are more accurate as compared to sales predicted by an OLS Linear regression model.

H_1: Sales forecasted by the ARIMA model are less accurate as compared to sales predicted by an OLS Linear regression model.

b) Rejection Criterion

$$Reject\ H_0\ if\ RMSE\ from\ the\ ARIMA\ forecasting\ model > RMSE\ from\ the\ Ordinary\ Least\ Squares\ regression\ model$$

c) Test Statistic

ACCURACY MEASURE

Table 4.12 Comparison of ARIMA VS OLS Regression accuracy

	RMSE	MAE	MAPE
ARIMA	1325.789	804.934	18.081
OLS	1642.979	1462.131	16.067

d) Decision

Since RMSE value from the ARIMA forecasting model ($1325.79) < RMSE value from the Ordinary Least Squares regression model ($1642.97) we do not reject H_0 and conclude that Sales forecasted by the ARIMA model are more accurate as compared to sales predicted by an OLS Linear regression model.

4.13.1 OTHER MACHINE LEARNING MODELS

The results of these other models are not as clear as for the OLS Linear regression and the ARIMA Model as factors such as assumptions and model adequacy checking were ignored.

Prediction Models

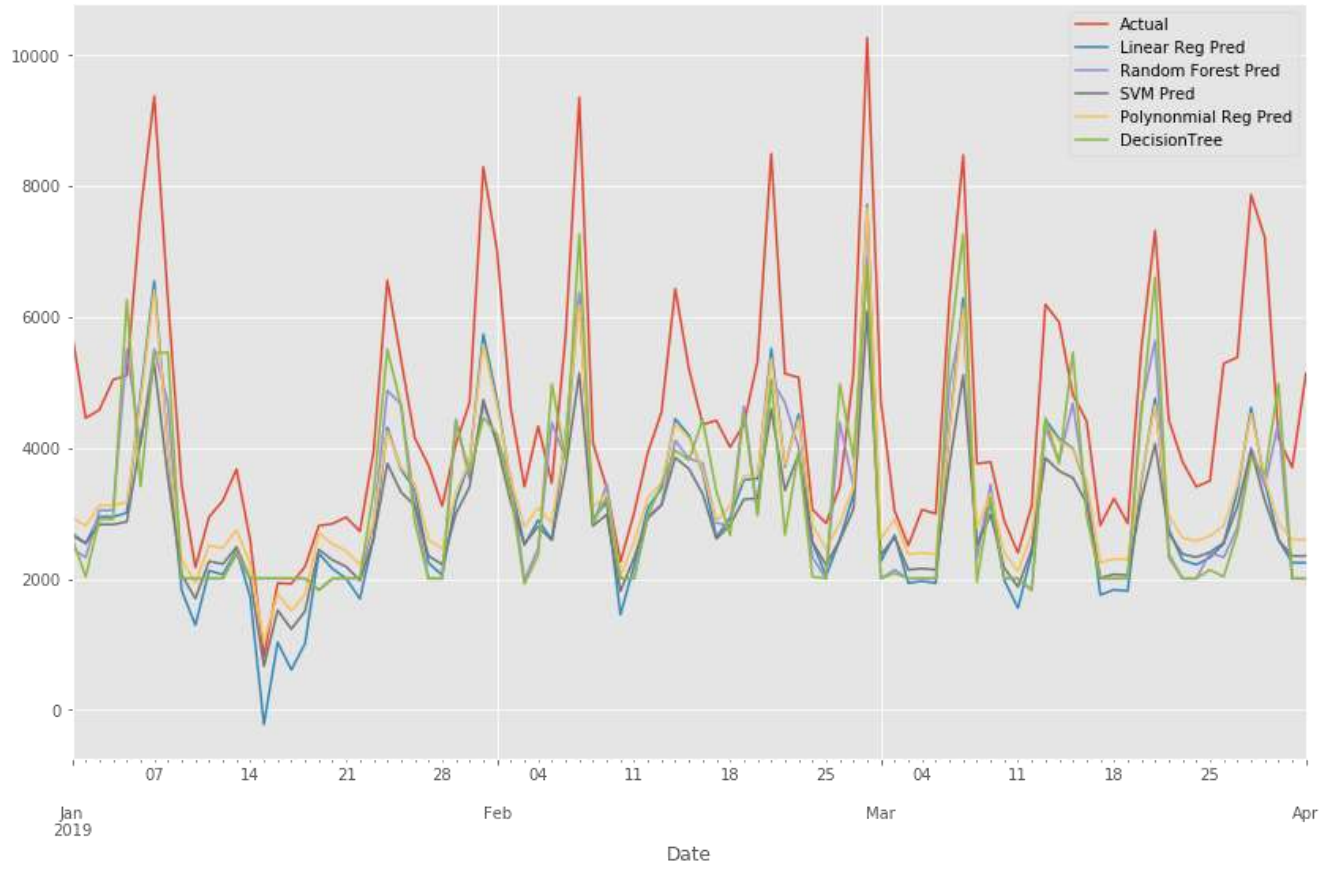

Figure 4.16 Prediction Models

From the Figure 4.16 it is clearly shown that the other prediction models almost follow the same pattern as the actual data, but also taking into account that the assumptions for the other models

were not considered but they were only used for predictions without checking them for adequacy. As for RMSE measure as shown in Table 4.13 it can be observed that the measures are relatively low indicating that the models are not bad for making predictions.

Table 4.13 RMSE for predictions

Model	RMSE
Linear Regression	1642.98
Random Forest	1709.03
Support Vector Machine	1873.25
Polynomial Regression	1526.02
Decision Tree	1758.62

Time Series Forecasting Models

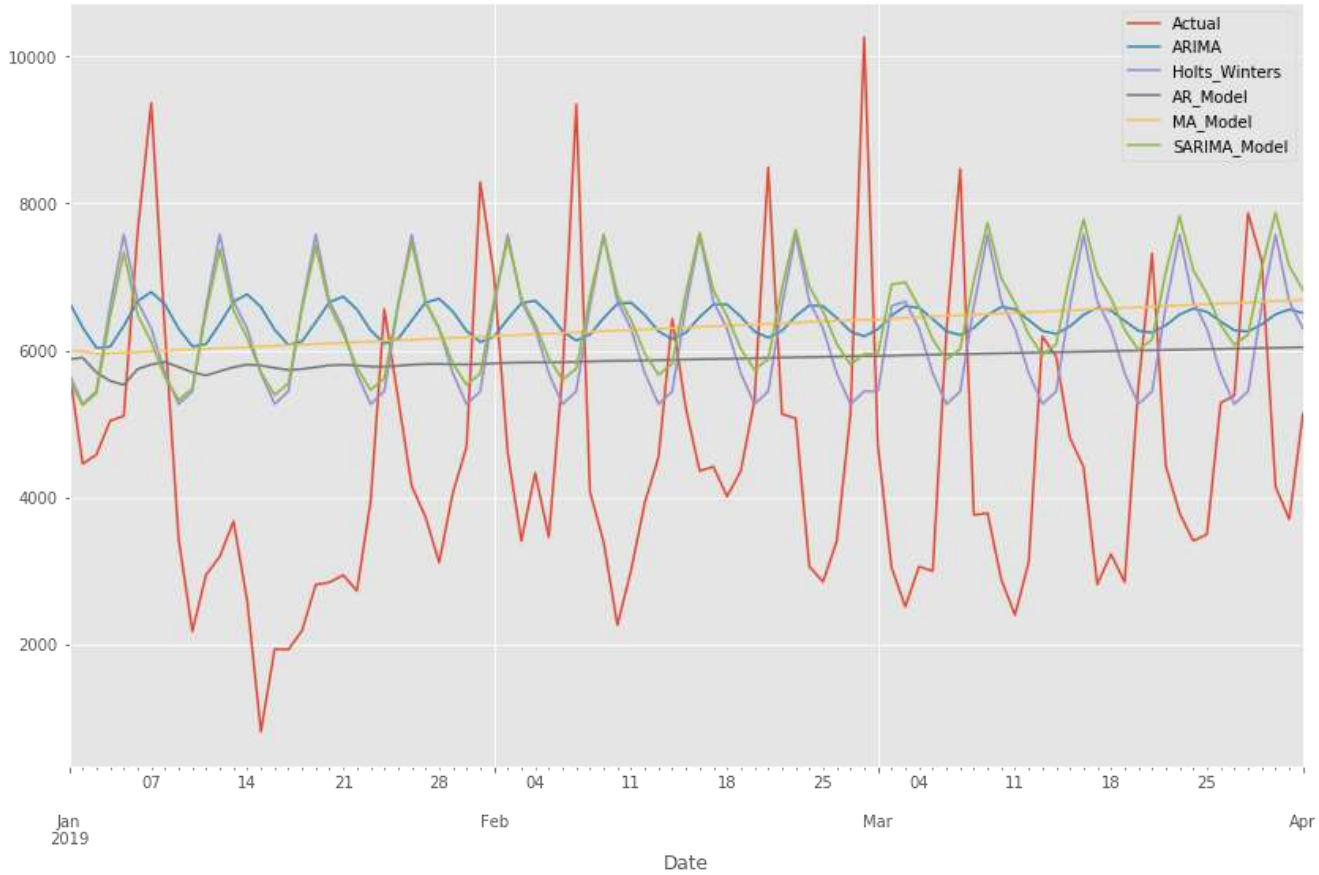

Figure 4.17 Time Series Forecasting Models

Table 4.14 RMSE for Forecasts

Model	RMSE
Holt`s Winters	2777.49
AR Model	2294.28
MA Model	2581.64
ARIMA	1325.79
SARIMA	2856.43

From the Figure 4.17 it is clearly shown that the other prediction models almost follow the same pattern as the actual data, but also taking into account that the assumptions for the other models were not considered but they were only used for forecasting without checking them for adequacy. As for RMSE measure as shown in Table 4.14 it can be observed that the RMSE measures for forecasts are slightly high than those for predictions. but conclusively it can be observed that out of the ten predictive models the ARIMA model had the smallest RMSE hence in case of Nando`s Zimbabwe it might be the best technique

See appendix C for a summary table for the forecasts and predictions values.

4.13.1 CHAPTER SUMMARY

The chapter focused on the analysis and interpretation of the data. Several laborious methods where performed in a quest to find the most accurate model between the ARIMA forecasting model and OLS regression model and minimize errors that may arise when performing forecasts and predictions. At the end the models ARIMA and OLS regression were fitted to the data and various test conducted to test the adequacy of the models and lastly the RMSE was used to compare the accuracy of the models. A vibrant picture of the conclusions of the analysis will be discussed in the next chapter.

CHAPTER 5: CONCLUSIONS AND RECOMENDATION

5.1.1 INTRODUCTION

This chapter is focusing on the explanations, conclusions and summarizing the findings from the study at hand. Recommendations will also be made with the intention of helping the reader on certain conclusions that were not made clear in the previous chapters.

5.2.1 CONCLUSION

The aim of this study was to build a time series model and forecast daily sales for Steer`s Zimbabwe from 1 January 2019 to 31 March 2019 and then build a regression prediction model using the method of ordinary least squares and predict the sales for the same period as forecasted by the time series model and compare the accuracy of both models. The data used for this research was collected from the GAAP accounting software used by "Simbisa Brands". For the diagnostics checking of both models the adequately fits the data and for the ARIMA model, diagnostic checks for the model specifically the Ljung-box statistic showed that the model generated fits the data. Additional stimulating discovery was that the data might have appeared slightly stationary from a ADF test carried out and after visual inspection of the time series plot the data appeared to be non-stationary and the KPSS gave the results that the data was non-stationary, the variation in the stationarity test is a result of the lack of robust signals of seasonality and this is probable from fast food sales as customers prefer buying fast-food irrespective of the time of year. The Correlation test conducted showed that there is a strong relationship between the sales and the number of customers .Model adequacy checking and testing of model significance was done for the OLS regression model and the model passed all the test and no assumption was violated, the results from the ANOVA test showed that the model passed the significance test. Now that both the models ARIMA and OLS regression models passed the validations its was fair enough to use them for forecasting and prediction. After the

estimation of the parameters of selected models, a number of diagnostic and forecasting accuracy tests were performed. Since both models used the same data set it was fair enough to measure the accuracy using a scale dependent measure and the RSME was used as the bases of measure. However, from the findings the MAPE values showed that the OLS regression model performs slightly better than the ARIMA with a percentage difference of 2.01%, but since in this study the author focused on scale dependent measure the observation from MAPE were omitted. Using scale depended measures the MAE for the ARIMA model was $804.93 which was better compared to the one for the OLS regression model $1462.13 and the RSME indicates that the ARIMA model has an RSME value of $1325.79 which was slightly better than the one of the OSL regression $1642.98 though the ARIMA gave lesser values than the OSL the difference was not significantly big and both models are good.

With reference to the findings of the Study, it can be concluded that:

i. The most adequate model for the data was ARIMA (2, 1, 3)

ii. The ARIMA model passes all the validation test and was fit to model the data

iii. There is a strong relationship between the sales and the number of customers ..

iv. The regression equation $Sales = -1223.152 + 10.029 * Customers + \varepsilon$ fits the data well it passed all its validation test

v. According to scale dependent measures RMSE and MAE the ARIMA model is more accurate than the OLS regression model

5.3.1 RECOMMENDATIONS

i. Forthcoming studies would benefit from exploring the validity of other qualitative factors in forecasting and prediction using time series and regression techniques. A likely method would be to integrate a forecast or a prediction grounded on each day of the week as an alternative of just comparing index-based days i.e. comparing days of the same week, this could be of help since each week day has its own precise trend and factors affecting each sale.

ii. When conducting visual analysis for both ARIMA models and OLS regression Models it is of great importance to use various plots as experienced in this study since a single plot

may not be sufficient to yield satisfactory results and if possible, it is best to back up visual plots with associated test.

iii. ARIMA models have a habit of converging to the mean as the forecast horizon goes to infinity, thus it is of great importance to use large amounts of data when forecasting

5.4.1 AREAS OF FURTHER RESEARCH

After carrying out all the tests and building up with all the models, the author came to the conclusion that further investigation and researches can be done to upsurge the efficiency of forecasting and prediction. It may be valuable to explore Intervention Models. Intervention Models estimates the effect of an external or exogenous intervention (e.g. the introduction of new regulations) on a time-series. As the accuracy results show that they were a minor difference between the accuracy measures and this leaves this study without a good enough answer as which of the two models is be best and this leaves a wider room for further research using different data sets and case studies. Use of the Bayesian Models to access performance. Problems do arise when choosing a model which generally performs best and works in the future. By useful model checking and evaluation of visual methods for Bayesian methods models in both R and Python, it is useful to avoid chasing metrics, for example misclassification and RMSE. It is therefore paramount to understand how well a model suits the data and our personal know-how is also vital. Importantly, we should note the following:

- A Bayesian approach to Linear Mixed Models (LMM) in R/Python
- Checking model fit via posterior predictive checks (PPC)
- Evaluation of performance through cross validation (LOOCV in Bayesian Setting)

In Bayesian jargon, Posterior Predictive checks are a superficial way of checking effectiveness of the model since this originates from the methods used to assess the goodness-of-fit. They may be need to visually explore how well the model fits the data, once you have a fit model and the MCMC diagnostics has no “red flags”. A non-Bayesian may reflect some residual analysis or normal Q-Q plots (applicable, of course for Bayesian model). Since Bayesian Models are generative in nature, we can stimulate values under a model and check whether these resemble those in the original data. Since Bayesian models are generative in nature, this allows for simulation data sets under a model and comparisons can be made against observed one. If a model fits well, simulated values should look similar to those in our observed data. If this is not

the case then the model has been mis-specified and therefore not complaint to our parameter inference and this can be a great area to explore. Evaluation of Bayesian model is no different from evaluation of other models, since most follow the same goodness-of-fit testing and comparison steps, but then they are tools and concept particular for the Bayesian models. Some are quite complex and therefore difficult to apply since their output maybe unfamiliar which makes it an area of interest. In fact, it is distress to fail to understand a model output summary. Fortunately, though, there are visual ways of diagnosing model fit, evaluating performance and even depicting results from Bayesian models, this was identified as an area of further research.

REFERENCES

- Armstrong, J. S., Morwitz, V. G., Kumar, V.,*Sales forecasts existing consumer products and services: do purchase intentions contribute to accuracy*? International Journal of Forecasting, Vol. 16, No. 3, 2000, pp. 383-397.
- Berman, H. (n.d.). Residual Analysis in Regression. Stat Trek. Retrieved from http://stattrek.com/regression/residual-analysis.aspx
- BREIMAN, L. (2001b). *Statistical modeling: The two cultures. Statist*. Sci. 16 199–215. MR1874152
- BROWN, P. J., VANNUCCI, M. and FEARN, T. (2002). *Bayes model averaging with selection of regressors*. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 519–536. MR1924304
- Brockwell, P. J., and R. A. Davis. "*Introduction To Time Series And Forecasting*.".Biometrics 54.3 (2002): 1204. Web.
- Chu, C-W., Zhang, G. P., *A comparative study of linear and nonlinear models for aggregate retail sales forecasting, International Journal of Production Economics*, Vol. 86, N. 2, 2003, pp. 17-231.
- Chatfield, Christopher. *Time-Series Forecasting*. 1st ed. Boca Raton: Chapman Hall/CRC, 2002. Print.
- Claudimar.V DEMAND FORECAST IN RETAIL FOOD AS A TOOL FORSTRATEGIC SUSTAINABILITY IN A SMALL BRAZILIAN COMPANY". Future Studies Research Journal: Trends and Strategies 5.2 (2013): 113-133. Web.
- De Gooijer, Jan G., and Rob J. Hyndman. "25 Years Of Time Series Forecasting". International Journal of Forecasting 22.3 (2006): 443-473. Web.
- GEISSER, S. (1993). *Predictive Inference: An Introduction. Chapman and Hall, London*. MR1252174
- G.Kleinbaum, Lawrence.L.Kupper, Azhar.Nizam, & Keith.E.Muller. (2008). *Applied Regression Analysis and other Multivariable Methods*,(4th ed). USA: Brooks/ Cole Cengrage Learning
- George, Box. P. "*A Selection Of Time Series Models For Short- To Medium-Term Wind Power Forecasting*". Journal of Wind Engineering and Industrial Aerodynamics 136 (2013): 201-210. Web.
- Frost, J. (2013, May 30). *Regression Analysis*: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-doi-interpret-r-squared-and-assess-the-goodness-of-fit
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2009). *TheElements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR1851606
- Karapanagiotis.P . "*Multi-Variate Stochastic Volatility Modelling Using WishartAutoregressive Processes*". Journal of Time Series Analysis 33.1 (2012): 48-60.Web.

- Kerkkanen, A. Korpela, J .Huiskonen, J. *Demand forecasting errors in industrialcontext: measurement and impacts, International Journal Production Economics,*Vol.118, 2009, pp. 43-48
- Makridakis A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J.Newton,E. Parzen, R. Winkler, *The accuracy of extrapolation (time series) methods:results of a forecasting competition, Journal of Forecasting*, Vol.1, 1982, pp.111-153.
- Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. 3rd Edition, New York, New York: John Wiley & Sons.
- Veiga, C. P., Veiga, C. R. P., Vieira, G. E., *Financial impact of errors of Prediction: a comparative study between models of Linear prediction and neural networks applied in the business*, Online, Vol. 12, 2012, pp. 629-656
- Veiga, C. P., *Analysis of quantitative methods of demand forecasting: comparative and post-graduate studies*, Dissertation (Masters in Production Engineering and Systems)) - Pontifcia University of Catholic Brasil, 2009.
- Tayman, Han Lin." *Point and Interval Forecasts Of Age-Specific Fertility Rates: A Comparison Of Functional Principal Component Methods*". Journal of Population Research 29.3 (2012): 249-267. Web.
- WANG, S., JANK, W. and SHMUELI, G. (2008). *Explaining and forecasting online auction prices and their dynamics using functional data analysis*. J. Business Econ. Statist. 26 144–160.

APPENDICES

Appendix A (R Script)

```
getwd()

setwd("C:\\Users\\Ransom Junior\\Desktop\\Datasets")

library(forecast)

library(tseries)

salesdata <-round(read.csv("Steers Zim sales data.csv",header = FALSE),0)

head(salesdata)

sales <- ts(salesdata) #converting data to time series data

head(sales) #displaying few top cells of the data

##create equally spaced time points for fitting trends

time.pts <- c(1:length(sales))

time.pts <- c(time.pts-min(time.pts))/max(time.pts)

##fit a moving average

mav.fit <- ksmooth(time.pts,sales,kernel = "box")

sales.fit.mav <- ts(mav.fit$y,start = 1,frequency = 1)

lines(sales.fit.mav,lwd=2,col="red")

abline(sales.fit.mav[1],0,lw=2,col="blue")

boxplot(timedata2,ylab="Sales($)",xlab="2018", main="Daily Sales Boxplot")

decompose(sales,"multiplicative")

##ACF AND PACF test for stationarity

par(mfrow=c(1,2))

acf(sales,main="ACF Sales")

pacf(sales,main="PACF Sales")

##ADF Test for stationarity

adf.test(sales,alternative = c("s"),k=trunc(length(sales)-1)^(1/3))

adf.test(sales,k=1)

adf.test(sales,k=73)

##KPSS Test for Stationarity

kpss.test(sales)

kpss.test(sales,null = "Trend",lshort=F)

pp.test(sales)

## ARIMA modelling

salesmodel <-auto.arima(sales, ic=c("aicc","aic","bic"),approximation=F, trace = T,method = NULL,stepwise = F,test = c("kpss","adf","pp"))

print(salesmodel)

## Residual Analysis

par(mfrow=c(1,1))

re

plot.ts(salesmodel$residuals,main="Sales Model Residuals")

lines(residualss.fit.mav,lwd=2,col="red")

abline(residualss.fit.mav[1],0,lw=2,col="blue")

par(mfrow=c(1,2))

acf(salesmodel$residuals,main="ACF Residuals")

pacf(salesmodel$residuals,main="PACF residuals")

par(mfrow=c(1,1))
```

```
qqnorm(salesmodel$residuals,main = "Normal Q-Q Plot Residuals")

qqline(salesmodel$residuals)

par(mfrow=c(1,2))

hist(salesmodel$residuals,main = "Histogram Residuals")

plot(density(salesmodel$residuals),main = "Density Plot Residuals")

##Diagnostic checking Ljung-Box

Box.test(salesmodel$residuals,lag = 15,type = "Ljung-Box")

Box.test(salesmodel$residuals,lag = 25,type = "Ljung-Box")

Box.test(salesmodel$residuals,lag = 58,type = "Ljung-Box")

Box.test(salesmodel$residuals,lag = 261,type = "Ljung-Box")

Box.test(salesmodel$residuals,type = "Ljung-Box")

##forecasting

mysalesforecast2 <- forecast(salesmodel,h=88, level = c(95))

mysalesforecast2

par(mfrow=c(1,1))

plot(mysalesforecast2,xlab = "Time",ylab = "Sales($)")

##Forecast accuracy

sales2019 <- read.csv("steerstest.csv")

forecasted <- read.csv("forecasted1.csv")

testsales_data <- as.vector(sales2019)

forecasted_sales <- as.vector(mysalesforecast2)

accuracy(forecasted_sales,testsales_data)

print(summary(mysalesforecast2))

accuracy(mysalesforecast2)

training <- ts(subset(sales2019, end=length(sales2019)-88))

test <- ts(subset(sales2019, start=length(sales2019)-364))

cafe.train <- Arima(training, order=c(2,1,3), lambda = 0)

cafe.train %>%

 forecast(h=88) %>%

 autoplot() + autolayer(test)

autoplot(training, series="Training data") +

 autolayer(fitted(cafe.train, h=1),

        series="1-step fitted values")
```

Appendix B Jupyter Notebook (Python code)

```
import warnings

warnings.filterwarnings('ignore')

import pandas as pd

%matplotlib inline

import matplotlib

import ma

style.use("ggplot")

import seaborn as sns

import plotly.express as px

import plotly.graph_objects as go

from plotly.subplots import make_subplots

import numpy as np
```

```
import datetime as dt

from datetime import timedelta

from sklearn.model_selection import GridSearchCV

from sklearn.ensemble import RandomForestRegressor

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette_score,silhouette_samples

from sklearn.linear_model import LinearRegression,Ridge,Lasso,ElasticNet

from sklearn.svm import SVR

from sklearn.metrics import mean_squared_error,r2_score

import statsmodels.api as sm

from statsmodels.formula.api import ols

from statsmodels.tsa.api import Holt,SimpleExpSmoothing,ExponentialSmoothing

from statsmodels.stats.outliers_influence import variance_inflation_factor

from sklearn.preprocessing import PolynomialFeatures

std=StandardScaler()

pd.set_option('display.float_format', lambda x: '%.6f' % x)

#READING THE DATA

data = pd.read_csv("steers.csv")

#SPLITTING THE DATA

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=0,shuffle = False)

model_scores=[]

# CORRELATION AND COVARIANCE

from statistics import mean

ys = np.array(data[["Profit"]])

Xs = np.array(data[["Customers"]])

def best_fit(Xs,ys):

    xbar= sum(Xs)/len(Xs)

    ybar = sum(ys)/len(ys)

    n= len(ys)

    numer= sum([xi*yi for xi,yi in zip(Xs,ys)])-n*xbar*ybar

    denum= sum([xi**2 for xi in Xs])-n*xbar**2

    b=numer/denum

    a=ybar-b*xbar

    return a,b

a,b = best_fit(Xs,ys)

plt.scatter(Xs,ys, c="skyblue")

yfit=[a+b*xi for xi in Xs]

plt.plot(Xs,yfit,'r--',color="b")

plt.xlabel("Customers")

plt.ylabel("Sales")

plt.title("ScatterPlot (Sales vs Customers)")

plt.savefig("ScatterPlot.png")

plt.show()

print(data.corr())

print(data.cov())

# OLS REGRESSION USING STATS MODEL
```

```
X_with_constant = sm.add_constant(X_train)

model = sm.OLS(y_train, X_with_constant)

results=model.fit()

results.params

print(results.summary())

#ANOVA TEST

model1 = ols('Profit~Customers',data=data).fit()

anova_table = sm.stats.anova_lm(model1, typ=1)

anova_table

## Normality test

sns.distplot(residual,rug=True,axlabel="Residuals", rug_kws={"color":"skyblue"},


kde_kws={"color":"b","lw":3,"label":"Denstity Curve"},


hist_kws={"histtype":"step","linewidth":3,

                "alpha":1,"color":"skyblue"})

#QQ PLOT

fid,ax = plt.subplots(figsize=(9,4.5))

_,(_,__,r)=sp.probplot(residual, plot=ax, fit=True)

##Homoskedatscity test

import pylab

import scipy.stats as sp

fid,ax = plt.subplots(figsize=(9,4.5))

_=ax.scatter(y_pred,residual)

plt.ylabel("Predicted Values")

plt.xlabel("Residuals")

plt.title("HomoscedasticityPlot (Predicted vs Residuals)")

##Auto correlation test

acf = smt.graphics.plot_acf(residual, lags=50, alpha=0.05)

acf.show()

## Model Accuracy measure

from sklearn.metrics import mean_squared_error,mean_squared_log_error, mean_absolute_error

print("MSE:"+ str(mean_squared_error(y_test,y_pred)))

print("RMSE:"+ str(np.sqrt(mean_squared_error(y_test,y_pred))))

print("MAE:"+ str(mean_absolute_error(y_test,y_pred)))

print("MAPE:"+str(mean_absolute_error(y_test, y_pred)/len(y_pred)))

print("MSLE:"+ str(mean_squared_log_error(y_test,y_pred)))

print("RMSlE:"+ str(np.sqrt(mean_squared_log_error(y_test,y_pred))))

#PREDICTED VS ACTUAL

plt.plot(y_pred,"red",label="Predicted")

plt.plot(y_test,"darkblue",label="Actual")

plt.title("Actual vs Predicted Plot")

plt.ylabel("Sales")

plt.xla

redlinreg = linreg.predict(X_test)

# RANDOM FOREST REGRESSION

RandomForestModel = RandomForestRegressor()
```

```
RandomForestModel.fit(X_train,y_train)

y_predRandomForest = RandomForestModel.predict(X_test)

#SUPPORT VECTOR MACHINE

supportvectorRegmodel= SVR()

supportvectorRegmodel.fit(X_train,y_train)

y_predSupportVectorReg = supportvectorRegmodel.predict(X_test)

#POLYNOMIAL REGRESSION

poly = PolynomialFeatures()

train_poly=poly.fit_transform(X_train)

test_poly=poly.fit_transform(X_test)

linreg=LinearRegression(normalize=True)

linreg.fit(train_poly,y_train)

prediction_poly=linreg.predict(test_poly)

#RIDGE REGRESSION

RidgeRegModel= Ridge(alpha=0.01)

RidgeRegModel.fit(X_train,y_train)

y_predRidgeRegModel = RidgeRegModel.predict(X_test)

pd.set_option('display.float_format', lambda x: '%.6f' % x)redRandomForest)))

print("Root Mean Square Error for Random Forest: ",np.sqrt(mean_squared_error(y_test,y_predRandomForest)))

model_scores.append(np.sqrt(mean_squared_error(y_test,y_predSupportVectorReg)))

print("Root Mean Square Error for Support Vectore Machine: ",np.sqrt(mean_squared_error(y_test,y_predSupportVectorReg)))

model_scores.append(np.sqrt(mean_squared_error(y_test,prediction_poly)))

print("Root Mean Square Error for Polynomial Regression: ",np.sqrt(mean_squared_error(y_test,prediction_poly)))

model_scores.append(np.sqrt(mean_squared_error(y_test,y_predlinreg)))

print("Root Mean Square Error for Linear Regression Regression: ",np.sqrt(mean_squared_error(y_test,y_predlinreg)))

model_scores.append(np.sqrt(mean_squared_error(y_test,y_predsarimamodel)))

print("Root Mean Square Error for SARIMA Model: ",np.sqrt(mean_squared_error(y_test,y_predsarimamodel)))

predictions = data = pd.read_csv("predictions.csv")

forecast = data = pd.read_csv("forecast.csv")

#Converting "Observation Date" into Datetime format

forecast["Date"]=pd.to_datetime(forecast["Date"])

predictions["Date"]=pd.to_datetime(predictions["Date"])

forecast2 = forecast.set_index("Date")

predictions2 = predictions.set_index("Date")

pd.plotting.register_matplotlib_converters()

forecast2.plot(figsize=(14,9))

predictions2.plot(figsize=(14,9))
```

Appendix C (Other Models Forecast and Predictions)

FORECASTS						
Date	Actual	ARIMA	Holts_Winters	AR_Model	MA_Model	SARIMA_Model
1/1/2019	5747.15	6665.22	5686.87	5879.65	5998.93	5598.67
1/2/2019	4455.15	6309.74	5267.88	5899.4	5986.73	5255.25
1/3/2019	4579.95	6035.12	5442.44	5701.03	5952.82	5415.57
1/4/2019	5040.3	6047.9	6603.98	5582.93	5961.15	6459.22
1/5/2019	5106.2	6331.95	7575.53	5531.12	5969.48	7330.82
1/6/2019	7595.95	6665.65	6665.87	5744.57	5977.81	6488.51
1/7/2019	9368.65	6794.72	6295.9	5807.55	5986.13	6113.59
1/8/2019	6255.95	6625.8	5686.89	5841.67	5994.46	5632.62
1/9/2019	3429.8	6293.74	5267.9	5773.38	6002.79	5324.24
1/10/2019	2175.4	6054.16	5442.46	5701.39	6011.12	5482.73
...	...	...	...	...	...	...
3/27/2019	5384.65	6272.21	5268.06	6023.27	6644.17	6075.93
3/28/2019	7866.2	6257.21	5442.62	6026.98	6652.5	6215.72
3/29/2019	7208.95	6355.28	6604.16	6030.68	6660.83	7119.35
3/30/2019	4146.1	6489.52	7575.71	6034.39	6669.16	7874.2
3/31/2019	3702.45	6557.21	6666.05	6038.1	6677.49	7146.98

PREDICTIONS					
Date	Actual	Linear Reg Pred	Random Forest Pred	SVM Pred	Polynonmial Reg Pred
1/1/2019	5747.15	2698.01	2462.6	2659.788946	2931.78
1/2/2019	4455.15	2537.56	2328.26	2550.134144	2811.14
1/3/2019	4579.95	2948.73	3045.08	2831.124575	3124.14
1/4/2019	5040.3	2948.73	3045.08	2831.124575	3124.14
1/5/2019	5106.2	3008.9	5501.3	2872.245126	3171.01
1/6/2019	7595.95	4804.01	4043.63	4099.008228	4694.07
1/7/2019	9368.65	6548.98	5508.53	5291.504203	6406.11
1/8/2019	6255.95	3991.7	4681.8	3543.88079	3974.95
1/9/2019	3429.8	1825.53	2001.37	2063.540958	2299.08
1/10/2019	2175.4	1294.02	2001.37	1700.309426	1941.61
...	...	...	...	...	...
3/27/2019	5384.65	3339.84	2779.13	3098.408156	3433.64
3/28/2019	7866.2	4613.47	3998.48	3968.79315	4520.95
3/29/2019	7208.95	3480.24	3422	3194.356108	3547.54
3/30/2019	4146.1	2597.73	4388.82	2591.254695	2856.15
3/31/2019	3702.45	2246.73	2001.37	2351.384815	2597.4