FACULTY OF BUSINESS, LAW AND DIGITAL TECHNOLOGIES

APPLIED ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

A DISSERTATION ON

USE OF MACHINE LEARNING FOR SALES FORECASTING

BY

OMAGE PELUMI SAMUEL

Dissertation submitted in partial fulfillment of the requirements for the award of Msc degree in Artificial Intelligence and Data Science at Solvent University

SEPTEMBER 2022

`

ACKNOWLEDGEMENT

I am grateful to my supervisor, Joe Appleton, for helping me focus on conducting the study and advancing the project. In addition to clarifying my ideas, he guided my dissertation route skillfully with his guidance.

Furthermore, I want to express my gratitude to everyone who assisted me with this research as well as to my family and close friends who have always been there for me. My gratitude also goes to my partner for her support at the worst moment.

`

ABSTRACT

A machine learning system learns from data to improve its performance through artificial intelligence (AI). Profitability is a priority for every business and as such sales forecasting is key to the business. A business can benefit from an accurate sales prediction by saving money on excess inventory, planning for the future, and increasing profit. Forecasting sales is primarily used to aid businesses in predicting their goals and quickly adjusting their strategies in order to boost productivity. This research used machine learning techniques to develop a machine learning model. Analysis was done and presented the use of the Seasonal Autoregressive Integrated Moving Average (SARIMA) method for developing a forecasting model that is able to support and provide a prediction of sales. According to the study, the dataset for model development was collected over time and model selected. As a result, it showed how well the model could accurately represent the historical data and make predictions based on it.

Keywords: Machine Learning, Sales forecasting; Time series; SARIMA.

`

TABLE OF CONTENTS

`

LIST OF FIGURES

`

LIST OF TABLES

`

`

INTRODUCTION

1.1 Background

The majority of companies want to be profitable. Typically, this is accomplished by increasing revenues while decreasing costs. Each and every organization depends on sales and sales forecasting. It is essential to running a successful firm. By increasing knowledge of the market, better forecasting aids in the development and improvement of company plans. In today's fiercely competitive economy and dynamic consumer landscape, accurate and timely forecasting of future income or sales can be tremendously beneficial for supermarket orders, planning, stock management, expansion, and decision-making. In today's contemporary world, Large shopping centers such as massive malls and marts are capturing data linked to sales of commodities or products with their many dependent or independent variables as a crucial step to be helpful in predicting future demands and inventory management. The dataset generated with different dependent and independent variables is a composite form of item characteristics, customer data, and data linked to inventory management in a data warehouse. Following that, the data is

`

adjusted in order to obtain accurate forecasts and collect fresh and intriguing outcomes that provide new insight on the understanding of the task's data. The primary objective of sales forecasting is to help businesses forecast their objectives and quickly alter their strategies to boost productivity. Sales forecasting is especially important because many household goods have a limited shelf life, which causes income loss in both shortage and surplus situations (Martne, 2019).

Typical sales forecast carefully examines the current settings or conditions, measures the financial inferences, acknowledges shortfalls and shortcomings before setting a budget, and considers strategic strategies for the upcoming year. In other terms, revenue forecasting is the process of estimating future sales based on historical data. An increased chance of success, regardless of external factors, results from a thorough understanding of prior tools and the capacity to forecast the company's future requirements. Businesses engaging forecasting continue to perform better than their non-prioritized counterparts (Punam et al., 2018).

Due to the increasing expansion of international stores and e-shopping, contests amongst numerous local small and medium-sized businesses (SMEs) and big supermarkets are only growing more heated and violent on a daily basis. In order to predict the sales volume for each item for the organization's inventory control, logistics, transport service, etc., the store or market tries to offer customized and limited-time promotions to attract more people depending on the day.

Due to a lack of information or the resources to do data analysis, small and medium-sized enterprises (SMEs) frequently have difficulty predicting

`

sales. Every business enterprise has time-sensitive responsibilities that must be done, which is what motivates forecasting sales. As an illustration, a nearby SME would like to forecast its daily sales in order to assign staff members to the appropriate shifts. The assumption is that a precise sales forecast will enable the SME to schedule its staff more affordably. It might be possible to remove an additional employee from the schedule if a day with low sales is anticipated, or vice versa if SME is running understaffed. Both circumstances are undesirable and result in inefficiencies inside the company.

Forecasting sales is often done at random by managers. As a result, qualified managers are becoming scarce and unreliable. Computerized solutions (that can fill in for qualified managers when they are unavailable or enable them to make the best choice by presenting prospective sales estimates) can help with sales forecasting. One way to put such a strategy into practice is to attempt to simulate professional managers' abilities in software applications (Grigorios, 2019).

As an alternative, machine learning can utilize its techniques to automatically generate accurate sales forecasting models using the dept of sales data and related information. Machine learning allows machines to acquire knowledge rather than being formally programmed (Stephen, 2015). When a computer program's performance at tasks in a class of activities, as measured, improves with learning, that program is said to have learned from experience (i.e., machine learning). Applications of machine learning that are widely used include the identification of phishing emails (Shahrivari, 2020), the theft of credit cards, financial predictions (Grigorios, 2019), personal assistants, product suggestions, automatic cars,

`

sentiment classification, etc. Data trends can be found using machine learning approaches, which are successful (Shahrivari, 2020).

This strategy is significantly easier. It is adaptable, so it can adjust to data changes, and it is not biased by the quirks of a single sales manager. However, it runs the risk of overestimating the human expert's prediction's precision, which is typically faulty. The highly developed algorithms for machine learning available today provide methods for anticipating or projecting a company's anticipated demand for income, which also aids in addressing the readily available yet low-cost computing and storage resources. SMEs selling electronics and phones can use dataset collected to predict their sales.

The communication (mobile phone) business has grown to be a significant contributor to national economic development as a result of advances in science and technology and the quick development of information technology worldwide (Lei, 2018). According to "Analysis Data," China alone sold 455.93 million smartphones in 2017, an increase of 1.1% over the previous year. By 2020, sales of smartphones in China are anticipated to be over 437.99 million units, demonstrating the huge market value of the mobile phone sector (smartphone market, 2018). When one of these links' experiences a difficulty, it may have an impact on the brand's overall sales as well as the long-term viability of the entire business. Selling involves more than just distributing goods; it also involves interacting with the market, learning about the industry's many mobile phone brands' development trends, and identifying the needs of the market's very large user base.

`

Nevertheless, more than a hundred distinct phone models (including those from Nokia, Samsung, Xiaomi, Infinix, iPhone, HTC, and Vivo, among others) are available, each with a unique set of specifications, some of which are better than others. The pricing and phone's features have a role in the decision a buyer makes about a certain phone device. The numerous factors, such as CPU, screen size, internal storage, etc., that customers consider when choosing a phone device. However, these phone products' prices vary depending on their specifications. The customer's final decision on a phone device is therefore impossible to predict. As a result, optimizing the phone sales process of SMEs is crucial to overcoming problems such delivery time, stock maintenance, promotion, and discounts, among others. This may be done by assessing sales of the phone items using forecasting. According to research and statistics, Samsung and Infinix smartphones will be the most popular in 2021. Since there are many different phone products, Samsung and Infinix products have been specially chosen for this investigation.

Application of time series forecasting in sales forecasting is more conventional but still highly attractive (James and Danielle 2017). In order for any process to operate over time based on previous data, time series forecasting is employed as the core component. Data from the past is used to make forecasts, and known future circumstances are taken into account (Mohit Gurnani et al. 2017). The creation and advancement of numerous forecasting models has received a lot of attention during the past few decades. Because it captures trends and seasonality, the time series machine learning method SARIMA known as Seasonal Autoregressive integrated moving average was utilized to anticipate or estimate sales

`

volume. In every retail setting, sales forecasting is essential since profitable sales are the foundation of every business. Additionally, a more precise forecast can be utilized to develop and enhance marketing strategies that help customers become more aware of products and services.

Phone related SMEs may be able to assess the sales of Samsung and Infinix phone products, keep more popular items in stock, and avoid spending money and time on items that have little to no demand or no demand at all in a given area by using the sales forecast results that were obtained using the machine learning method. The method used in collection of data in this thesis is made up of sales information for Samsung and Infinix products that was collected from SMEs.

1.2 Scope

This study's focus is solely on the item sales information (i.e Samsung and infinix phone product) from small and medium-scale enterprises (SMEs). The technique makes use of consumer purchasing patterns as well as aggregate information regarding SMEs' sales of goods (i.e Samsung and infinix phone product).

1.3 Research Objectives

1. Transforming data using a variety of preparatory methods in order to implement algorithms for machine learning.
2. Identifying key characteristics that will have the biggest impact on the product's sales.
3. To select the ideal algorithm for machine learning sales forecasting.

`

4. Choosing several metrics to evaluate how well the used machine learning algorithms operated.

1.4 Research Questions

To achieve the goal, two questions of inquiry have been established for this study. These are their definitions:

RQ: What is the most effective algorithm for machine learning-based SME sales prediction?

Reason: the goal of this research topic is to Utilize several machine learning techniques in developing a Machine Learning model. To determine which model best fits the data, these models are compared using a variety of criteria, including accuracy score, mean absolute error, and maximum error.

1.5 Research Hypothesis

H_0: The SARIMA approach did satisfactorily in the forecast of SME phone sales, notably for Samsung and Infinix phone products.

1.6 Dissertation Roadmap

This thesis follows the organizational framework below.

1. Introduction: The thesis' concept and requirements are briefly outlined in this introductory chapter. The project is set up in this chapter.
2. Literature review: This chapter gives a summary of the research that has been done on machine learning techniques and methodological comparisons. The purpose of this project is discussed in this chapter.

`

3. Methodology: This chapter discusses the many approaches and strategies that are employed throughout the project's design, along with how each step is completed.
4. Results: The findings and the interpretation of the results are provided in depth in this chapter.
5. Conclusion: The importance of the design results is discussed in this chapter along with system advantages and the model for accurate sales prediction.

`

2.0 LITERATURE REVIEW

2.1 Sales and Prediction

It is well recognized in the Mobile phone industry that consumer desires are quite unstable. In actuality, consumers typically base their decisions on cost. In this instance, the managerial store attempted to solve this circumstance by lowering the price of manufacturing or by purchasing from the manufacturer directly or first party in an effort to lower the price. Nearly all of the firms previously mentioned are situated on an island far from store X. Therefore, the process of providing goods must be carried out with the proper timing and plan. In order to maintain sustainability throughout the supply chain, storage, and sales, it is intended that the goods not be stacked or hoarded in the warehouse. Prediction is an effort to foresee the future. evaluating current conditions against earlier ones. To forecast sales, one must establish the anticipated even deciding to manufacture, the volume of sales based on the outcomes, make judgments or implement policies revenue forecasts, prior to this scientific. The prospective sales and market are studied. market area will be in charge in the future. (Wibowo, 2018). This study allows for the prediction of phone sales that are doing well and those that aren't doing well based on the interests of the customers, so that the vendors can create a commodities inventory in the future. Based on the interests of the buyers and are predicted to significantly boost sales and lower the loss possibility as a result of unsold stock items well. Prediction is an effort to foresee the future. Evaluating current conditions against earlier ones. To forecast sales, one must establish the anticipated even deciding to manufacture, the volume of sales based on the outcomes, make judgments or implement

`

policies revenue forecasts, prior to this scientific. The prospective sales and market are studied. Market area will be in charge in the future (Wibowo, 2018). This study allows for the prediction of HP sales that are doing well and those that aren't doing well based on the interests of the customers, so that the vendors can create a commodities inventory in the future based on the interests of the buyers and are predicted to significantly boost sales and lower the loss possibility as a result of unsold stock items well. Accurate product sales forecasting can assist vendors in developing a rational replenishment strategy and cite a source for businesses' allocation of capital and inventory. At the moment, classical time series models are the primary means of predicting sales volume for instance, the Arima, and the combined prediction model. Ge Na employed the ARIMA model to accurately forecast an enterprise's footwear sales volume and generate positive outcomes (Ge Na et al., 2018). Change To forecast and model the sales volume of a medical device, Xiaohua used the random forest algorithm company (Chang and Xiong, 2018). Liu Lu created a multi-dimensional time series model based on the volume of cigarette sales on SVM and attained a high level of prediction accuracy.

2.1.1 Importance and factors influencing Sales Forecasting

In sales forecasting, supply and demand for products can be easily modified by countering momentary demand in light of the predicted estimate; and regular supply is also fostered; good inventory control is advantageously benefited by avoiding the weaknesses of under stocking and overstocking. Sales territory allocation and reallocation are made easier.

`

It is a forward planner because all other requirements such as raw materials, labor, plant layout, financial demands, warehousing, transportation, and so on are determined by the predicted sales volume. Sales prospects are sought for based on forecasts using sales forecasting, and so the discovery of selling effectiveness is achieved.

Forecasting benefits all involved in the process and is the greatest way to ensure flexibility to changing conditions. Collaboration among all parties involved results in a cohesive front, an awareness of the reasons behind actions, and a broader viewpoint; It regularizes productions by utilizing sales forecasting and avoiding extra time at high premium charges. It also decreases production inactive time. Sales representatives and sales trends are also regularized—increasing or decreasing based on the anticipated sales volume.

If the following criteria are carefully taken into consideration, an accurate sales prediction can be generated

General Economic Condition: It is critical to evaluate all economic variables affecting the customers and the company. The overall economic trend-inflation or deflation-and how it affects the firm positively or negatively must be considered. A detailed understanding of the business's economic, political, and general trends allows for more accurate forecasting. Market behavior in the past, national income, disposable personal income, consumer consumption patterns, and so on all have a significant impact on projection.

Industrial Operations: Markets are teeming with comparable items created by many businesses that compete with one another to improve sales. As a

`

result, related industries' pricing policies, design, advanced technology advancements, promotional efforts, and so on must be closely monitored. A new business may come out with items to the marketplaces and naturally alter the market share of the current firms. Unstable conditions—industrial discontent, government control through laws and regulations, insufficient raw material availability, and so on—have a direct impact on production, sales, and profitability.

2.2 Time Series Forecasting

In the aspect of confirmed historical data, time series models are used to predict outcomes. Moving average, smooth-based, and ARIMA are examples of common kinds. It's important to choose the model that works best depending on the unique time series because not all models will provide the same results for the same dataset. One of the most often used data science approaches in business, finance, supply chain management, manufacturing, and inventory planning is time series forecasting. A time component is often present in prediction difficulties, necessitating the extrapolation of time series data or time series forecasting. Time series forecasting is also an important topic of machine learning (ML) and may be framed as a supervised learning problem. It may be subjected to ML techniques including regression, neural networks, support vector machines, random forests, and XGBoost. .

An ordered collection of data points spaced out over time is referred to as a time series. While the other variable or variables continually changing values in this situation, time are typically an independent variable. Over fixed temporal intervals, the time series data is observed. This information might be in any quantitative and measurable parameter

`

relating to business, science, finance, etc. Analyzing time series data involves finding recurring patterns that appear in the data over an extended period of time. To do this, specialists use particular techniques to examine the properties of the data and extract insightful statistics that ultimately help with business forecasting. The creation of models that help anticipate business indicators and their future behavior is aided by specific characteristics of the time series that are presented. The quality of the forecasts will depend on how accurately the properties of the provided data are identified.

A collection of statistical approaches called time series forecasting methods can be extremely useful for estimating various variables and can be applied to any industry. You must look for three key characteristics in a time series to get reliable projections. Seasonality, stationarity, and autocorrelation are these.

2.2.1 Seasonality and autocorrelation

When a time series is presented and its delayed version is compared over a specific period of time, the degree of resemblance between the two is described mathematically as autocorrelation. A set of values for a variable or other thing are described by this time series. An entity's previous values and current values can be compared using autocorrelation to aid make this determination. Professionals are able to recognize and analyze data patterns, create relationships, and prepare for the future by using historical and present data. Seasonality can be measured when an entity displays similar values on a regular basis, that is, after a set period of time. As an illustration, every holiday season sees a comparable spike in the commercial sales of several products. Seasonality creates a

`

basis for the variable's predictability in relation to a specific hour of the day, day of the week, season, or event. Salespeople can create their plan in advance of that particular period with the aid of seasonal variation data.

2.2.2 Trends and Stationarity

A time series is considered stationary when its statistical qualities stay the same over time. In other words, the series' mean and variance remain unchanged. Stock prices, for example, are typically not static. KPSS tests, Dickey-Fuller tests, or expanded versions of these tests are implemented to decide whether a time series is stationary. Statistics-based techniques are the most common ones used to find stationarity. These tests essentially assess a null hypothesis in one direction or the other. A series' stationarity is thought to be very important; without it, a model showing the data would have varying degrees of accuracy at various time points. Therefore, before modeling, experts employ some strategies to convert a specified non-stationary time series into a stationary one. Trends are tracked over a long period of time. Its trend may diminish, increase, or remain stable based on the nature of the entity and any relevant influencing variables. Population, fertility rate, mortality rate, and other such phenomena, for instance, are some of the ones that primarily exhibit mobility and cannot be arranged into a stationary time series.

2.2.3 Time Series Data Modeling

Seasonal data can be patterned in a variety of ways. Moving averages, and Auto Regressive Integrated Moving Average (ARIMA) models are the three primary categories of time series models. The principal thing is to choose

`

the suitable forecasting approach depending on the properties of the time series data. Among all of the time series forecasting techniques, **the Moving Average (MA)** method is the very elementary and fundamental. For a univariate time series, this model is employed. In an MA model, it is presumable that the output (or future) variable will rely linearly on the present and past values. As a result, the average of the past data is used to form the new series. The MA model can be used to find and spotlight patterns and trend cycles. The frequently used seasonal data forecasting models is the exponential smoothing (ES) approach. In univariate series, the Exponential smoothing methodology is also utilized, just like the MA method. The weighted average of the previous values is used to calculate the new values in this case. The weight is given to a value decrease as it gets older. The straightforward (single) ES approach or the sophisticated (double or triple) ES time series model can be used depending on the trends and seasonality of the variable**.**

Time series data without a pattern or seasonality are smoothed using a straightforward exponential approach. In this method, the influence of historical data on the forecast is determined by a single smoothing coefficient, or alpha (). The forecast is more affected by the more recent values than the earlier values if it is closer to '1'. If the value is close to "0," then the opposite is true. Due to the existence of a trend in the data, the double exponential smoothing approach performs the smoothing operation twice. In addition to the factor-alpha, the parameter beta is also utilized to control how the series' trend changes. Depending on whether an additive or multiplicative damping effect is used, the trend may be linear or exponential. Due to trend and seasonality in the data,

`

Holt-Winters which is sometimes called the triple exponential technique comprises smoothing at three levels. So, in addition to the factors, this method also uses a parameter called gamma () to regulate the effect of seasonality on the series. For financial or economic institutions, exponential approaches are typically used.

Another popular forecasting approach that combines two or more time series models is the **autoregressive integrated moving average which is known as ARIMA model**. Multivariate non-stationary data work well with this model. The ARIMA approach is built on the ideas of autoregression, autocorrelation, and moving average. A model variant known as SARIMA is used in the case of seasonality data. Seasonal ARIMA which is also known as SARIMA is essentially an extension of ARIMA that takes into account the seasonal component of the time series. SARIMA support and facilitate with both trend and seasonality, whereas ARIMA can only analyze data with a trend. SARIMA takes into account the three seasonal parameters for the same as well as a fourth component for seasonal periods in addition to the three trend factors of autoregression, difference, and moving average. The flexibility of this model to include a wide range of factors and their combinations is one of its benefits.

2.2.4 Time Series Analysis using Python or R

Python and R are two popular computer languages for time series analysis. Python extends a more all-encompassing approach to data science, whereas R enables more specialized statistical computing. Python is simpler and easy to learn. The R statistical program, on the other hand, offers a larger ecosystem with built-in data analysis methods.

`

Python makes use of the Pandas software library, which was created especially for financial sector analysis and associated estimations. Time deltas (total length), time periods, and time stamps (specific points in time) are all used in this language (intervals). The built-in functionality includes these fundamental objects that include dates and times. For a number of tasks involving certain dates and times, data scientists occasionally employ a third-party module in addition to the built-in modules. The future expansion of the company can be predicted by experts using time series analysis.

2.3 SARIMA

Time uses the SARIMA (Seasonal Autoregressive Integrated Moving Average) approach. Seasonal data patterns and series predictions for stochastic model data.

There are four parts in the SARIMA modeling process:

(1) The model identification phase identifies the variables in the time-series' stationarity through analysis and verification; this also establishes the most pertinent

auto-regression and moving average combined; (2) The model estimate step examines the identifies the most effective model among those first-step models;

(3) model validation phase evaluates the accuracy of the selected model and determines any improvements that might be made.

(4) The model forecasting phase projects future data from the series that are provided with a interval of confidence.

`

2.4 Related works

Sales forecasting is a crucial need for company planning and wise decision-making, enabling businesses to forecast sales and make appropriate plans. Sales forecasting is crucial for offline organizations, and it's typically done using statistical techniques like regression or other models to forecast future sales and make appropriate plans for the business's marketing. The goal of sales prediction is to estimate future sales for businesses including supermarkets, grocers, eateries, bakeries, and patisseries. Sales forecasting assists the business in reducing the stock of items whose sales are expected to decline and increasing the stock of goods whose sales are anticipated to rise, which will result in an increase in the business's sales and the expression of the selling output variable.

Time series data have been the subject of statistical research and forecasting for a very long time. The monthly sunspot counts that Schuster looked at are one of the first series ever discovered. Time-domain approaches and frequency-domain methods are two categories into which time series investigations can be subdivided. The latter covers auto-correlation and cross-correlation analysis, while the former includes spectral analysis and, more recently, wavelet analysis. The moving average (MA) model and the autoregressive (AR) model are the two main components of a linear univariate time series model. An AR model's forecast depends on its prior observations, whereas an MA model's forecast depends on its prior errors. As the time for which forecasts are made is further in the future, the quality of the prediction’s declines. When it comes to creating approaches for univariate time series modeling, Box and Jenkins have been trailblazers. The SARIMA model can be defined by

`

include the potential for seasonal lags and differencing at a single lag in the ARMA model, as well as by allowing for seasonal components.

An additional traditional forecasting technique is seasonal ARIMA (SARIMA). This method has been utilized successfully in a variety of situations, including predicting vehicle traffic flow and demand for tourism. Box-Jenkins first proposed the seasonal time series ARIMA (SARIMA) model, which has been successfully applied to forecasting issues with the economy, the market, society, etc. The restriction of this model is that at least 50, and ideally 100 or more observations should be employed. This model has the advantage of reliable predictions over short periods. The data used in this model are accurate values without measurement errors, but it also incorporates the idea of measurement error to handle the differences between estimators and observations.

The seasonal Holt-Winter model and the seasonal autoregressive integrated moving average model were used by Akter and Rahman (2010) to study the milk supply of a dairy cooperative in the UK (SARIMA). The produced forecasts had an inaccuracy of less than 3%, according to their findings, and longer series give better forecasts than shorter series. In order to predict inflation rates in the Turkish economy, Saz (2011) examined the accuracy of SARIMA models. He put the stationarity through rigorous tests and shown that it is simultaneously deterministic and stochastic in character, with the latter form predominating the inflation process. This is true for both the seasonality and the Turkish inflation rate. He also offered the first analysis of fractional integration in a Turkish inflation series from 2003 to 2009. In Northern Thailand, Wongkoon et al. (2008) investigated the prevalence of dengue hemorrhagic fever (DHF). Data from

`

2003 to 2006 were analyzed using SARIMA models. When compared to data gathered from January 2007 to September 2007, the anticipated data were verified. Their findings demonstrated that the SARIMA model is useful for forecasting the frequency of DHF in Northern Thailand. Xiang (2008) utilized the SARIMA model to examine temperature data from Stockholm from 1756 to 2007 in order to analyze climate change. The outcome demonstrated that there is a steady framework in the temperature readings and that even the greatest outlier has a modest impact. Using increasing inflation data from July 1991 to December 2009, Aidoo (2010) suggested SARIMA (1,1,1) x (0,0,1)12 to predict Ghana's inflation rate. He predicted Ghana's inflation rates for the next seven months, which were in line with the actual inflation rate that was recorded from January to April and reported by the Ghana Statistical Service Department. Gallop et al. (2012) used SARIMA to estimate Vancouver bicycle traffic utilizing weather parameters and applied it to hourly bicycle count and temperature data. Hu et al. (2010) compared time series poisson regression and SARIMA models to evaluate the relationship between meteorological variables and the prevalence of cryptosporidiosis. In order to investigate the potential influence of meteorological variables on the transmission of cryptosporidiosis, they used time series Poisson regression and (SARIMA) models. The model evaluation revealed that SARIMA has greater predictive power than Poisson regression.

This study covered the topic of using SARIMA models to forecast and model Nigeria's inflation rates. One of the many uses of time series analysis and forecasting is in the domains of hydrology, environmental management, and economics and finance, to name a few. The Autoregressive Integrated

`

Moving Average (ARIMA), one of the best techniques for evaluating time series data, was used in this work. It was developed by Box and Jenkins. With a total of 120 data points, we built an ARIMA model for Nigeria's monthly inflation rates for the years November 2003 to October 2013 using the Box-Jenkins approach. The ARIMA (1, 1, 1) (0, 0, 1)12 model used in this study was created, and its parameters were calculated as 0.3587yt+0.6413yt-1-0.8840et-11 -0.7308912et-12+0.8268et. The monthly rate of inflation for the next year 2014 is predicted using this methodology. In order to battle the anticipated rise in inflation rates starting in the first quarter of 2014, policy makers will use the expected findings to gain insight into more suitable economic and monetary policies. (Otu et al., 2014)

According to a study conducted by Mustapha et al., (2021) The study focuses on trustworthy and advanced models of forecasting the exchange rate series due to the high and rising rate of uncertainty in the foreign exchange market in Nigeria. The goal of this study is to build a more accurate exchange rate model for Nigeria using more data points, while also taking into account the model's periodic seasonal component and using the estimated model to provide forecasts. The Central Bank of Nigeria (CBN) Statistical Bulletin was utilized for the study's monthly data for Nigeria from 1981:M1 to 2018:M12. The widely used Box-Jenkins approach, which extends the autoregressive (AR) and moving average (MA) processes, was used to analyze the data. The study produced SARIMA (0,1,1) x (1,1,1)12 from among the competing models using 456 data points, based on its AIC and BIC values. Using a sample of data for 2019, the estimated model is determined to be sufficient for generating forecasts. Using a sample of data for 2019, the estimated model is determined to be

`

sufficient for generating forecasts. The study recommends that policymakers should take the seasonal component into account when developing monetary policies aimed at foreign exchange in order to stabilize the economy since it has anticipated that the value of the Naira will decrease along a seasonal route.

3.0 METHODOLOGY

Tools used for this Project

Python

`

Programming languages like Python are used for creating software and websites, as well as performing analysis. The Python programming language has become one of the most used tools for data scientists, statisticians, and machine learning algorithm developers to analyze and compute data. There are several embedded libraries embedded in the program that make it easy for data scientists to accomplish specific tasks.

For this project, Pycharm, a Python integrated development environment, was used. Library used:

- Pandas

Pandas is a Python package, that is mostly used in data science, data analysis, and machine learning. In addition to its multidimensional array support, it relies on another package titled Numpy.

- Matplotlib

A Python library that permits the making of static, animated, and interactive visualizations is Matplotlib. In addition to making easy things easier, Matplotlib makes hard things possible.

- Seaborn

Data visualization is made easy with Seaborn's statistical graphics library for Python. Statistics plots can be enhanced with beautiful default styles and colour palettes. Built using the matplotlib library, it also integrates tightly with pandas data structures.

- Statsmodels

`

Statistical models and statistical data exploration can be performed using the statsmodels Python module, which provides classes, function calls, and tests for estimating many different statistical models.

Model

ARIMA is a model that statistically analysis the use of time series data to forecast future trends. ARIMA does the forecast based on previous values.

ARIMA, which stands for "Auto Regressive Integrated Moving Average," is essentially a class of models that "explains" a given time series based on its own previous values, that is, its inherent lags and the lagged prediction errors, such that equation can be used to anticipate future values. ARIMA models can be used to represent any "non-seasonal" time series that shows patterns and isn't just random noise. Three terms—p, d, and q—define an ARIMA model, where p denotes the phase of the AR term. The MA term's order is indicated by the letter "q." The amount of differencing needed to render the time series constant is denoted by the letter "d."

If a time series exhibits seasonal trends, seasonal terms must be included, and the result is SARIMA, an acronym for "Seasonal ARIMA". A linear regression model that employs its own lags as predictors is referred to as "auto-regressive" in the ARIMA algorithm. The predictors should be unrelated to one another and not correlated for linear regression models to be effective. The most typical method is to difference a series in order to make it stationary. Specifically, reduce old value from the current one. Multiple differencing's could be needed occasionally, depending on how complex the series is. Thus, the value of d is the smallest number of

`

differencing required to make the sequence stationary. Additionally, if the timeseries is stationary at this point, then d = 0.

3.1 Proposed research design

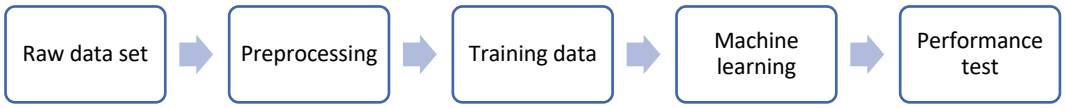


Figure 1: Research Design

The study suggests the above approach to address the research issue that was specified. SMEs raw data would undergo preprocessing, which would involve looking for missing values, outliers, and anomalies, exploring the data, and visualizing it. In order to create the prediction models, the algorithms would next be developed using the train data. The test data would be used to test the algorithms' performance in order to identify which model had the most accurate prediction performance.

3.2 Selection of Algorithm

It is not easy to choose an approach for every problem. Although there is no perfect algorithm that solves every problem, a select handful are well known for sometimes outperforming other algorithms. The systems' accuracy won't be the same for all sorts of data; it will vary depending on the type of data. In this thesis, algorithms for machine learning that was predicted to be effective on the problems is seasonal ARIMA (SARIMA).

`

CHAPTER FOUR

5.3 DATASET

This project utilizes a dataset which contains the sales for a mobile phone company from 2015 to 2019, the goal is to predict the sales for the coming month of the year 2021. The dataset consists of samples that were taken over a long period of time, from daily to weekly. The dataset is made up of a number of variables as shown in table 1. All of these elements are taken into account when creating a strong forecasting model to estimate how much of the sold products and their categories are consumed. In this scenario, the time series forecasting method can also be used to estimate how much of the product will be needed overall as it changes over time. In order to build a successful model to acquire the sale, a relationship between each variable's randomness must be established urgently. Simple

`

regression procedures may now be less to not at all successful in creating a trustworthy model as a result of this complex structure. As a result, plotting all the parameters against a time sequence can show us how their usage has changed over a particular period, and by taking into account these historical values, the prospective requirement can be accurately forecasted.

Columns	Description
Order Date	The day the sales was made
Delivery Date	The day the item was delivered
Customer ID	Unique number given to buyer
Customer Age	Age of buyer
Customer Gender	Male, Female
Location	Where sales were made to.
Zone	Zone 1,2,3 or 4
Delivery Type	Express, Shipped from abroad, Standard
Shipping Fee	Cost of shipping
Order Quantity	Quantity ordered
Sales	Sales made for each day (with shipping fee included)
Status	Delivered or returned
Rating	Rating gave for an item by customer (1-5)

Table 1: Variables of the dataset

`

5.4 Exploratory data analysis

Exploratory data analysis (EDA) is used examine and analyze data sets and relate their key properties, often using data visualization techniques. It makes it simpler to discover patterns, identify anomalies, test hypotheses, or verify assumptions by finding out how to best modify data sources to obtain the answers needed. A greater insight of the variables in the data set and their relationships to one another is provided by exploratory data analysis(EDA), which is firstly used to look into what data can disclose beyond the formal modeling or hypothesis testing task. It can also help in finding out if the statistical methods you are thinking about using for data analysis are appropriate.

After EDA on the dataset, the data set contains (2569 rows, 13 columns) contains no missing values and no duplicate records. The purpose of the exploratory data analysis is to give a few insights into the sales before preceding to build the time series model.

The first step when performing the exploratory data analysis was to import the necessary libraries needed, then the dataset was loaded into pandas data frame.

`

```
# import necessary libraries needed for analysis
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# using pandas to read the xlsx file
df = pd.read_excel("sam_desc.xlsx")


df = df.drop(['Delivery Date', 'CustomerID', 'Customer Age', 'Customer Gender', 'Status'], axis=1)

for cols in df.columns:
    print(cols)

 OrderDate
 Location
 Zone
 Delivery Type
 Shipping Fee
 Order Quantity
 Sales
 Rating
```

Figure 2: Importing

The code snippet shows how the dataset was loaded into the Dataframe and how some columns which will not be needed for the analysis. This part is the data pre-processing. Data pre-processing is a very important step in data analysis, this process involves taking raw data and cleaning the data into a format that can be easily analyzed by machine learning.

While cleaning the data,

- Missing values was checked
- The shape of the data was checked
- Duplicate values were checked
- And the columns data types were checked.

`

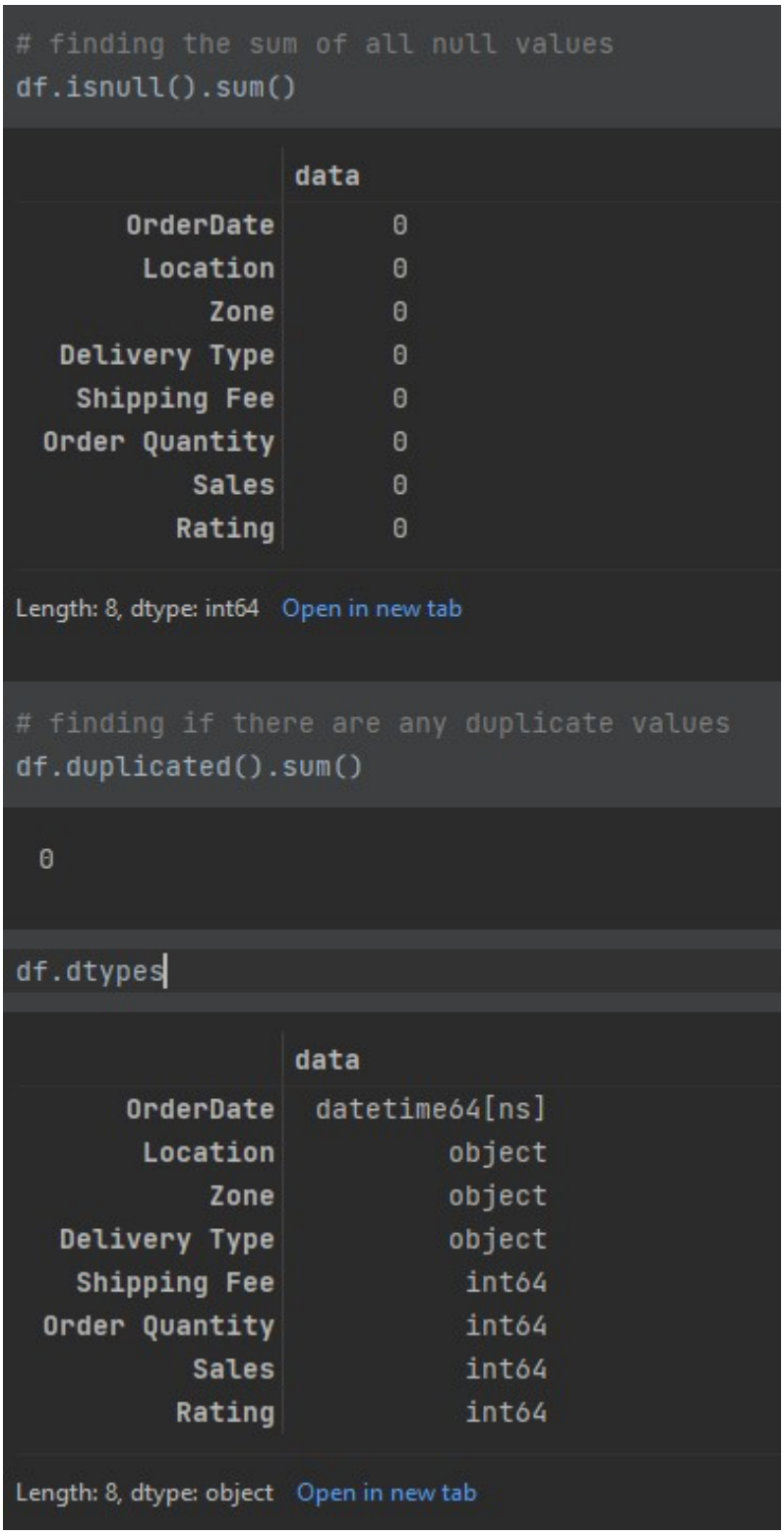


Figure 3: Image for data pre processing

`

This shows how the missing values were checked. As seen above, it shows that the dataset has no missing values as well as duplicate values. This proves that there's no need to perform any cleaning on the datasets.

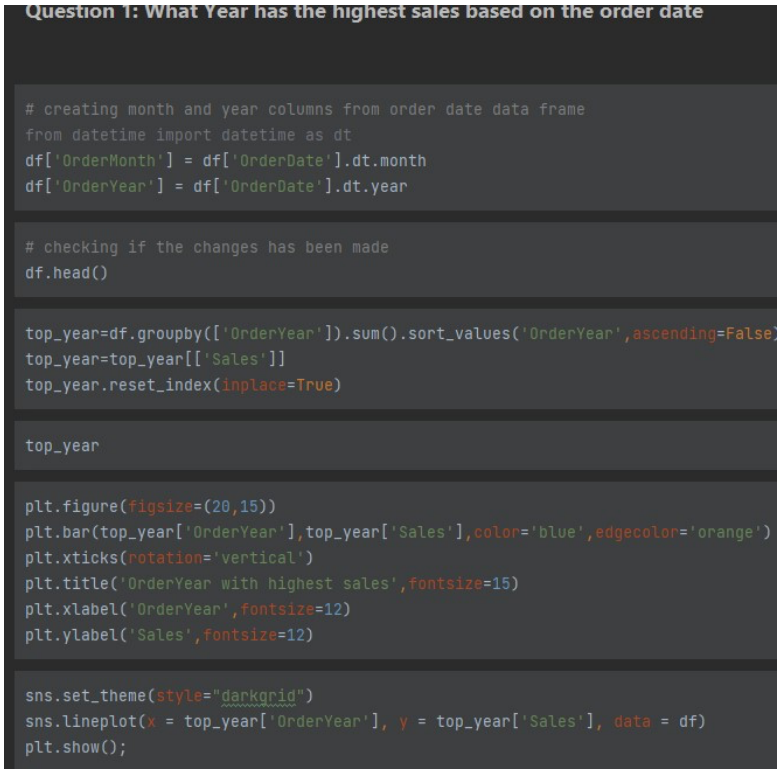


Figure 4: Basic description of the quantitative variable

This code snippet shows how the sales in the data was summed up by grouping the year, and also how the graph was generated to show which year generated the highest sales.

`

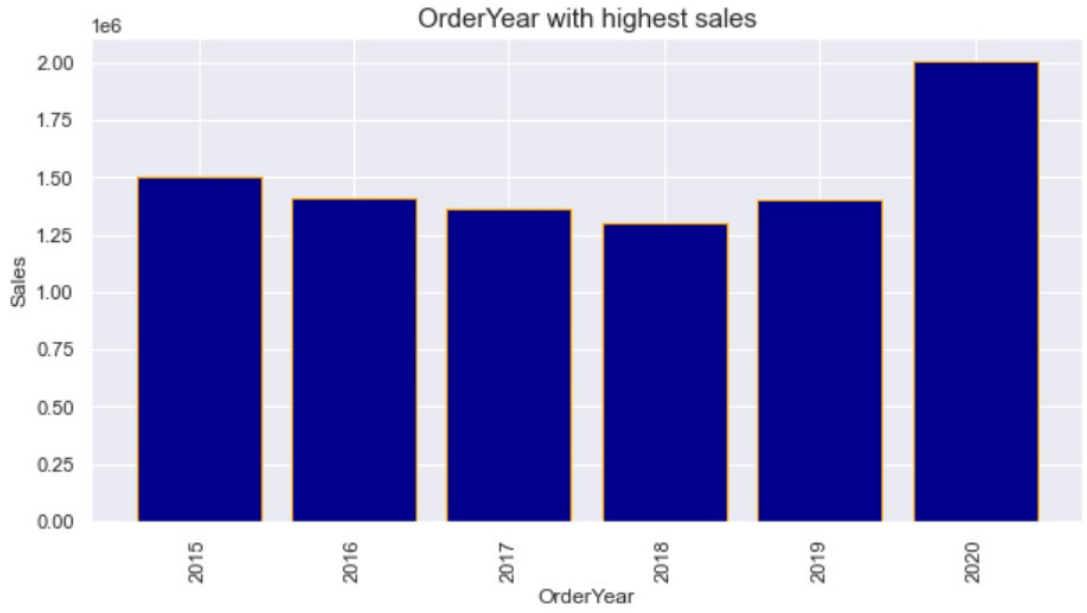


Figure 5: Order year with highest sale

It is evident that 2020 generated the highest sales while 2018 generated the lowest sales.

Question 2: Which location generates the highest sales.

```
top_location=df.groupby(['Location']).sum().sort_values('Sales',ascending=False).head(20)
top_location=top_location[['Sales']]
top_location.reset_index(inplace=True)
```

```
top_location
```

```
plt.figure(figsize=(20,15))
plt.bar(top_location['Location'],top_location['Sales'],color='#6a0dad',edgecolor='orange')
plt.xticks(rotation='vertical')
plt.title('Location with highest sales',fontsize=15)
plt.xlabel('Location',fontsize=12)
plt.ylabel('Sales',fontsize=12)
```

Figure 6: How Sales summed based on location

`

This code snippet shows how the sales in the data was summed based on location, and also how the graph was generated to show which location generated the highest sales.

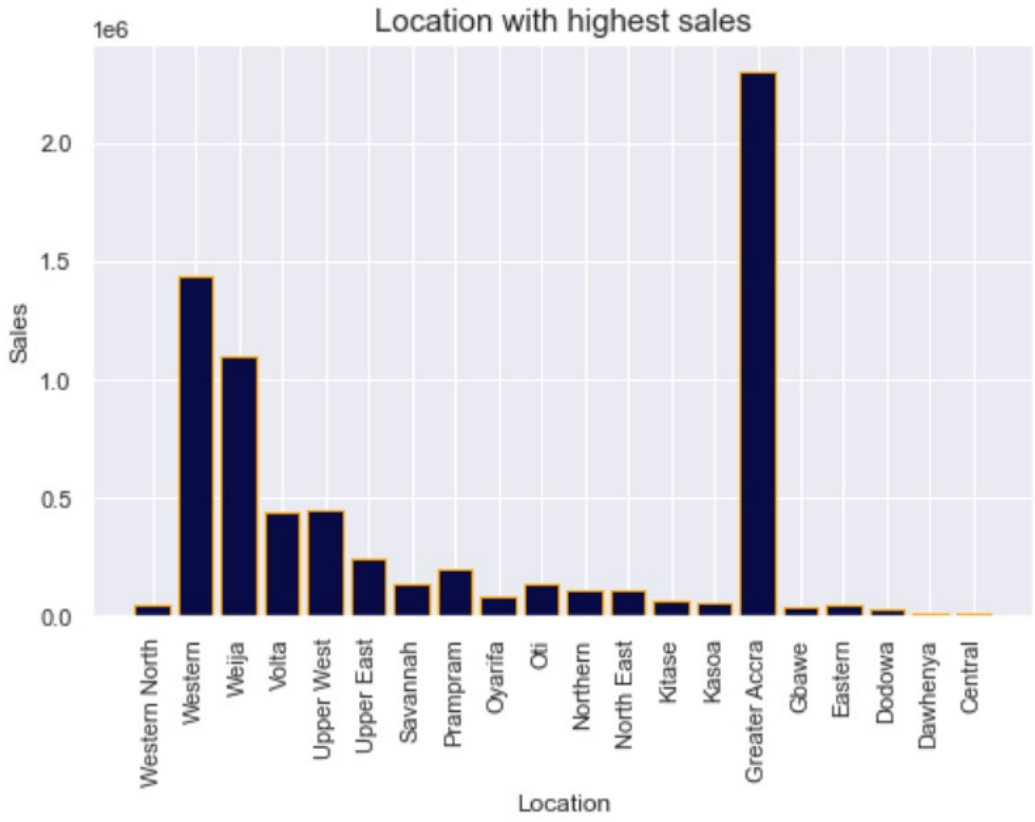


Figure 7: Location with highest sale

From the graph above, it is seen that greater Accra generated the highest sales while central and Dawhenya and Central generated the lowest or no sales.

`

Question 3: Which zone has the highest sales.

```
top_zone=df.groupby(['Zone']).sum().sort_values('Zone',ascending=False).head(20)
top_zone=top_zone[['Sales']].round()
top_zone.reset_index(inplace=True)
```

```
top_zone
```

	Zone	Sales
0	Zone 4	1176739
1	Zone 3	3934935
2	Zone 2	1582178
3	Zone 1	2302068

Figure 8: Zone with highest sale code

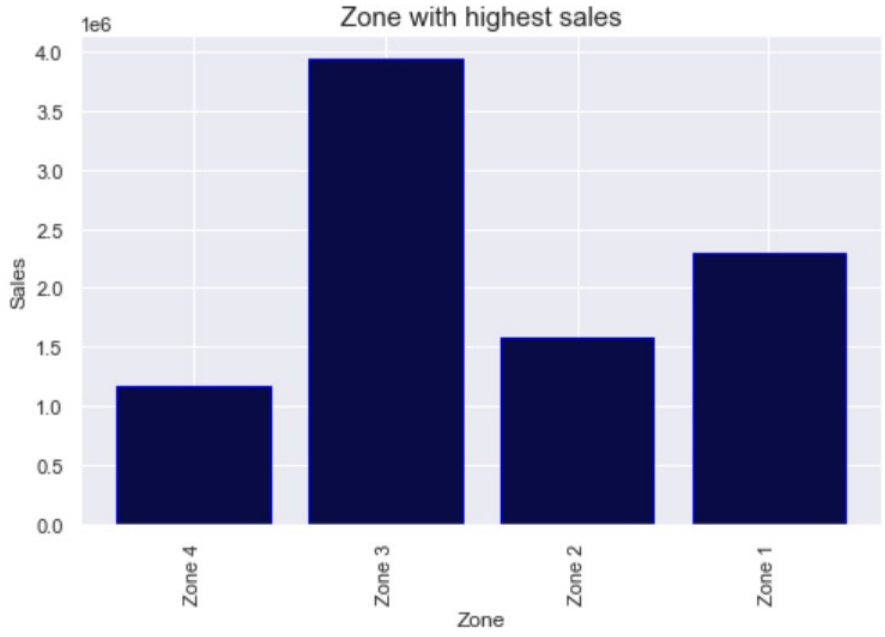


Figure 9: Zone with Highest sales

`

This code snippet shows how the sales in the data was summed based on zone, and also how the graph was generated to show which zone generated the highest sales.

From the exploratory data analysis, we could deduce that

- The Year 2020, has the highest sales, also the sales tend to increase yearly.
- Greater Accra generates the highest sales of while Brong Ahafo has the lowest sales.
- Zone 3 has the highest sales while zone 4 has the lowest sales
- Majority of the customers tend to use the “standard delivery” delivery type; therefore, the standard delivery generates the most sales
- Rating doesn’t have a significant implication on the sales, this is because orders with a rating of 1 tend to have higher sales than orders with a rating of 5.

5.5 FORECAST

The application of time series analysis in forecasting models are widespread in a number of industries, including the prediction of traffic, the economy, and energy demand. In fact, it is important to be precise when estimating sales because the outcome will influence future choices. The model was created using data from the reported monthly sales for the years 2015 through 2019 and then it was verified using data from 2020. The chosen model demonstrated that monthly instances could be predicted

`

using data from 1, 2, and 12 months earlier. Using seasonal autoregressive models, the projected monthly sales for 2021 were calculated. The use of the additive decomposition approach was made possible by the historical data's comparatively stable trend. The prediction values for the next 12 months were calculated using historical data spanning 14 years, from 2015 to 2020. When the time series displays seasonal fluctuation, seasonal ARIMA (SARIMA) is applied. The multiplicative process of SARIMA will be represented as (p, d, q) (P, D, Q) s, where (P) stands for seasonal autoregressive and (Q) for seasonal moving average. The length of the seasonal period is indicated by the letter "s" in subscript. As an illustration, s = 7 in an hourly data time series, s = 4 in a quarterly data and s = 12 in a monthly data. The backshift operator (B) is used to formalize the model. In order for Bkyt to equal yt-k, the time series observation backward in time by k periods is represented by Bk. The backshift operator was formerly used to offer a general stationarity transformation, according to which a time series is stationery if its statistical features (mean and variance) remain constant throughout time.

`

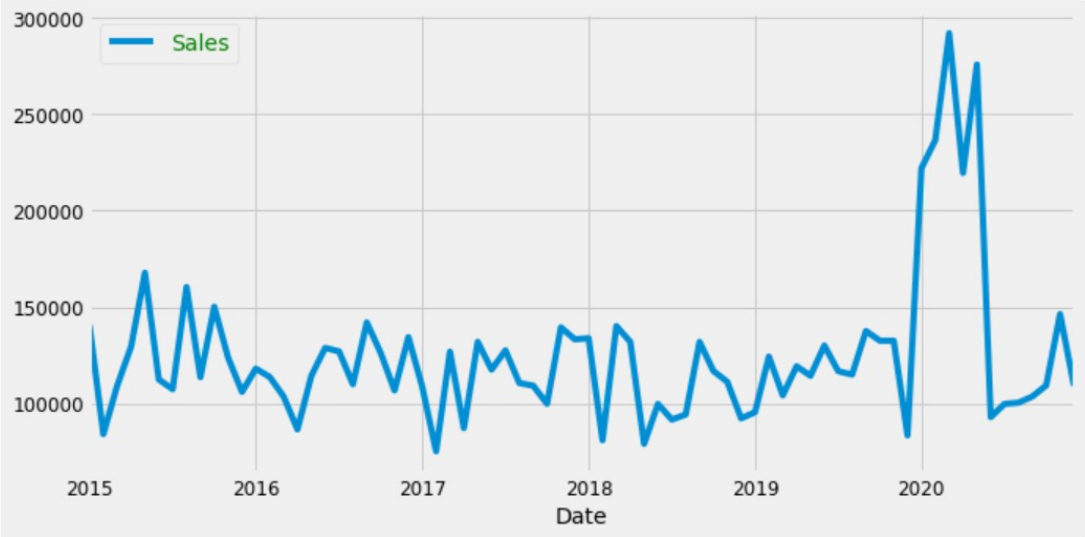


Figure 10: Original dataset

In figure 10, a line graph of historical data from 2015 to 2020 is displayed. The x axis shows the year, while the y axis shows the number of sales during that year. The graph shows the movement of sales from 2015 – 2020 ending.

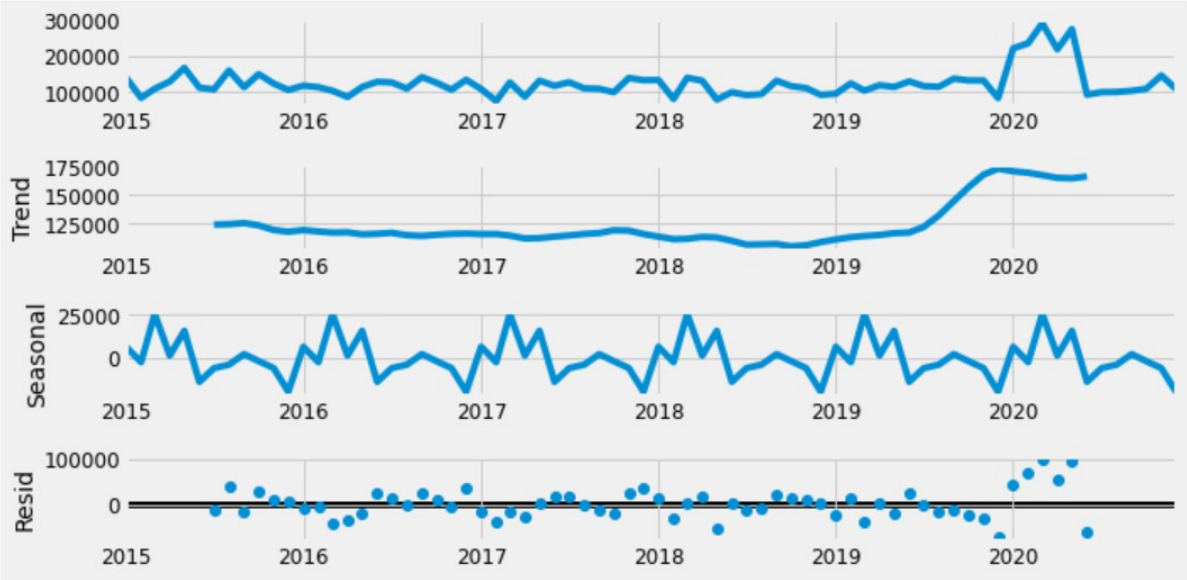


Figure 11: Dataset showing seasonality and trends

`

SARIMA was chosen as the best method to construct a model prediction because the time series plot of the historical data showed seasonal fluctuations that reveal a similar trend every year.

5.6 Time series Identification

The first step in model identification is time series identification. It is done by plotting sample autocorrelations (SAC) and then sample partial autocorrelations (SPAC) based on the original data. An efficient method for determining the randomness of a data set is to use autocorrelation plots. The autocorrelations of data values with various time lags are computed to determine the randomness. For all time-lag separations, if they were random, such autocorrelations should be close to zero. If not random, at least one of the autocorrelations will have a substantial non-zero value. For fitting ARIMA models, autocorrelation charts are also employed in the model identification phase. To be more precise, partial autocorrelations are helpful in determining the order of an autoregressive model. At lag p+1 and above, an AR(p) process has zero partial autocorrelation. If the sample autocorrelation plot suggests that an AR model would be appropriate, the sample partial autocorrelation plot is analysed to assist in determining the order.

`

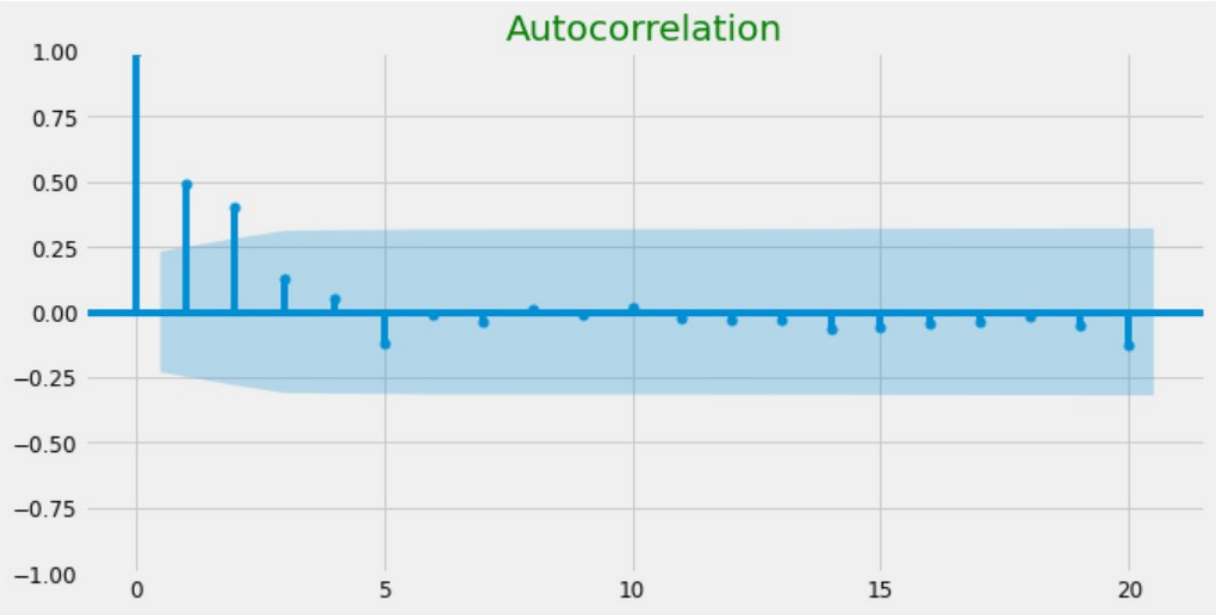


Figure 12: Autocorrelation graph

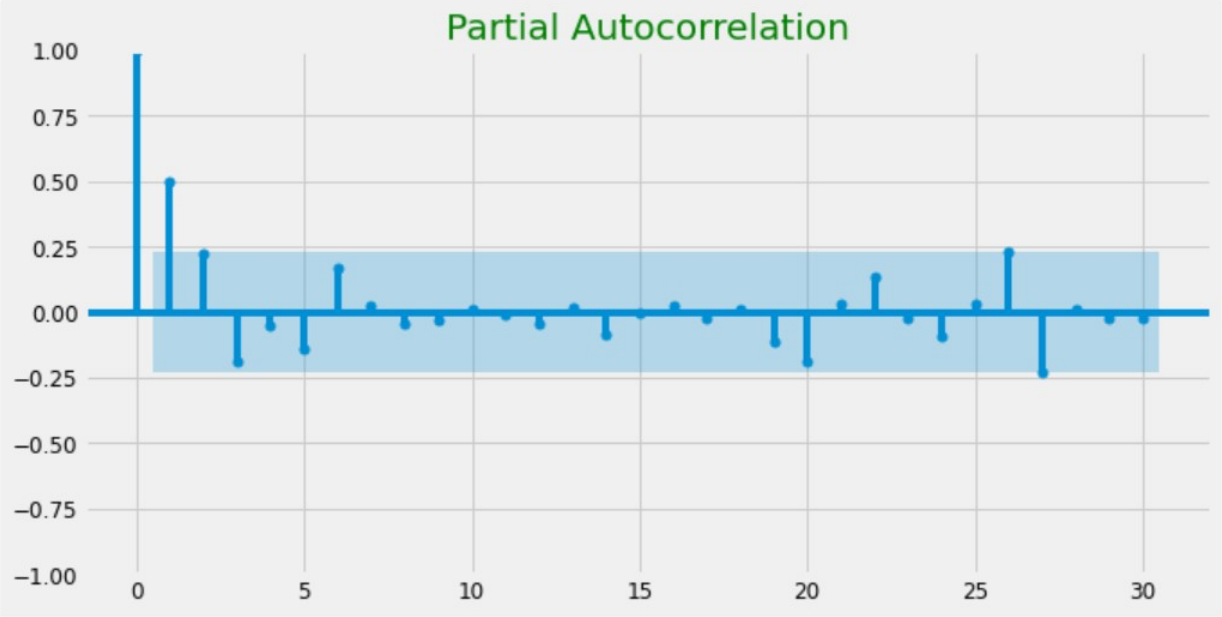


Figure 13: Partial autocorrelation graph

`

5.7 Forecast model

The seasonal ARIMA model is a multiplicative model that includes both seasonal and non-seasonal components. For the model, one quick notation is ARIMA (p, d, q) X (P, D, Q)s with P being the seasonal AR order, D being the seasonal differencing, Q being the seasonal MA order, S being the time span of the recurring seasonal pattern, and p being the non-seasonal AR order, d being the non-seasonal differencing, and q being the non-seasonal MA order.

For this study, The final SARIMA model (0,1,1)(1,1,1)12 was used to forecast the values of the 12 months-ahead.

```
mod = sm.tsa.statespace.SARIMAX(df,
                                order=(0, 1, 1),
                                seasonal_order=(1, 1, 1, 12),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results = mod.fit()
print(results.summary().tables[1])
```

Figure 14: Setting sarima parameters

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.3732      0.118      3.154      0.002       0.141       0.605
ar.S.L12      -0.2882      0.328     -0.878      0.380      -0.931       0.355
ma.S.L12      -0.2863      0.338     -0.847      0.397      -0.949       0.376
sigma2      2.21e+09   1.73e-10   1.28e+19      0.000    2.21e+09    2.21e+09
==============================================================================
```

Figure 15: Model summary statistics

`

5.3 Implementation

5.3.1 Test

The model performance was tested by checking it performance with the 2019 – 2020 sales data, before proceeding to make forecast for monthly sales in 2021.

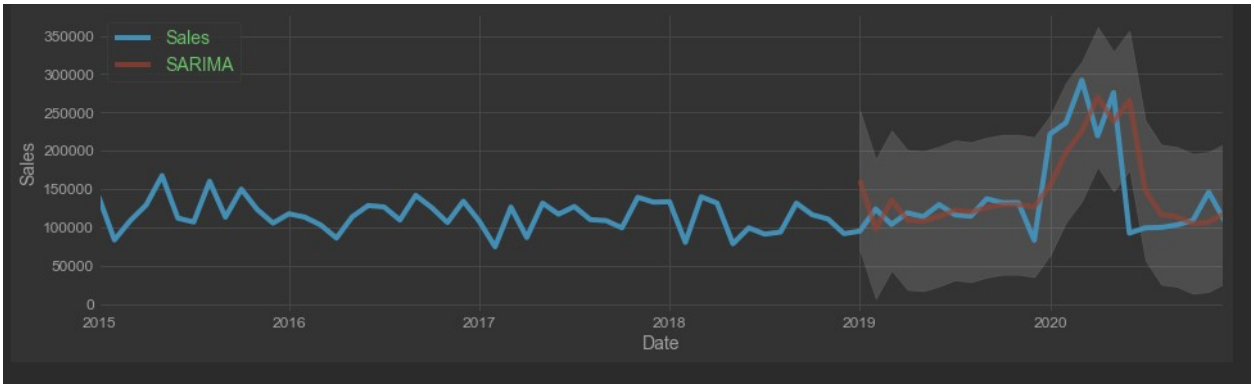


Figure 16: Model test performance on 2019-2020 sales data

The model was evaluated by comparing its performance from 2019 -2020.

From the above graph, we can see that the SARIMA line is close to the actual sales. This shows how effective the model is in predicting future prices.

`

5.7.2 Forecast

```
pred_uc = results.get_forecast(steps=12)
pred_ci = pred_uc.conf_int()
ax = df.plot(label='observed', figsize=(14, 4))
pred_uc.predicted_mean.plot(ax=ax, label='Forecast')
ax.fill_between(pred_ci.index,
                pred_ci.iloc[:, 0],
                pred_ci.iloc[:, 1], color='k', alpha=.25)
ax.set_xlabel('Date')
ax.set_ylabel('Sales')
plt.legend()
plt.show()
```

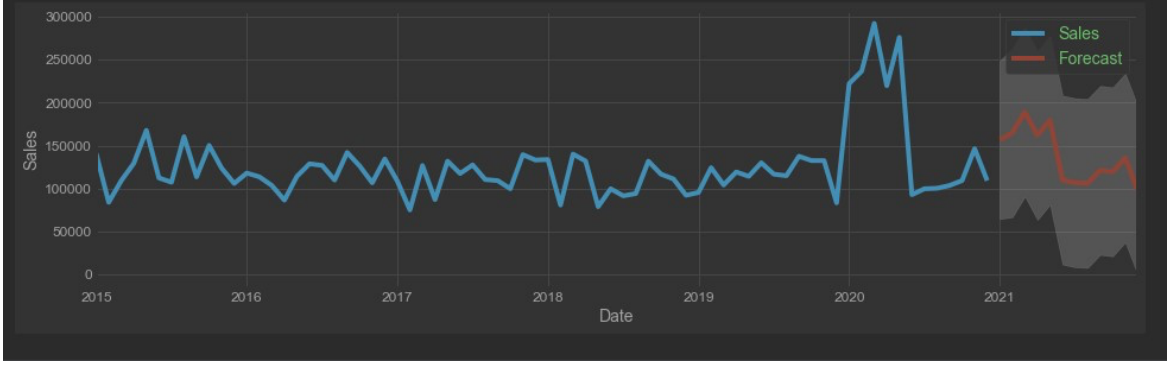


Figure 17: A zoom in into the Forecast result

This shows the forecast result for 2021 having evaluated the performance with previous years.

`

6.0 CONCLUSION

The prediction of the future sales is important to make a better policy. In this paper, the use of forecasting method was applied to predict monthly sales. The adjusted model prediction was developed by using SARIMA model based on the historical data. The result indicates that SARIMA (0,1,1) (1,1,1)12 was the fit model. The model was also be able to represent the historical data.

7.0 LIMITATIONS

Sets of data are crucial for training models. Since the trained model is not aware of future outcomes, employing outdated datasets could prevent it from making accurate predictions. The answer is to use existing datasets. When the datasets are tiny, the classifiers' modelling time is unknown because they will be trained more quickly. The modelling time will change depending on the amount of the datasets, thus when we utilize larger datasets, the modelling time will be well.

`

REFERENCES

Aidoo, E. (2010). Modelling and Forecasting Inflation Rates in Ghana: An Application of SARIMA Models. Master's Thesis. Högskolan Dalarna School of Technology and Business Studies. Sweden.

Akter, S. and Rahman, S. (2010). Agribusiness Forecasting with Univariate Time Series Modelling Techniques: The Case of a Dairy Cooperative in the UK. Journal of Farm Management. Vol. 13. No. 11. pp. 747-764.

Gallop, C., Tseand, C., Zhao, J., (2012), A Seasonal Autoregressive Model of Vancouver Bicycle Traffic Using Weather Variables.TRB 2012 Annual Meeting.

Grigorios Tsoumakas (2019). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, 52(1):441–447.

Hu, W., Tong, S., Mengersen, K., and Connel, D., (2010), Weather Variability and the Incidence of Cryptosporidiosis: Comparison of Time Series Poisson Regression and SARIMA Models. 51(1).

James J Pao and Danielle S Sullivan (2017). "Time Series Sales Forecasting". In: *Final Year Project*

Martne, A., Schmck, S., Pereere, C., Pirker, M., & Haltmeier (2018). A machine learning framework for customer purchase prediction in the non-contractual setting. European Journal of Operational Research, http://dio.org/ 10.1016/j.ejor.2018.04.034.

`

Mohit Gurnani et al., (2017). “Forecasting of sales by using fusion of machine learning techniques”. In: *2017 International Conference on Data*

Mustapha, A., Yola, A., Madaki, N., & Saad, U. (2021). Forecasting Nigeria's Exchange Rate Using SARIMA Modeling. *Dutse International Journal Of Social And Economic Research*, *6*(1). Retrieved 9 September 2022, from.

Otu, A., George A., O., Jude, O., Hope Ifeyinwa, M., & Andrew I., I. (2014). Application of Sarima Models in Modelling and Forecasting Nigeria’s Inflation Rates. *American Journal Of Applied Mathematics And Statistics*, *2*(1), 16-28. https://doi.org/10.12691/ajams-2-1-4

Punam, K., Pamula, R. and Jain, P.K. (2018). A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). pp. 617{620}.

Saz, S. (2011). The Efficacy of SARIMA Models for Forecasting Inflation Rates in Developing Countries: The Case for Turkey. International Research Journal of Finance and Economics. Vol. 62. pp. 112-142.

Wongkoon, S., Pollar. M., Jaroensutasinee, M. and Jaroensutasinee. K. (2008). Predicting DHF Incidence in Northern Thailand using Time Series Analysis Technique. International Journal of Biological and Life Sciences. Vol. 4. No.3. pp. 117-121.

Xiang, J. (2008). Applying ARIMA Model to the Analysis of Monthly Temperature of Stockholm. Master’s thesis in Statistics. Department of Economics and Society. Dalarna University. Sweden.

`

Shahrivari, V., (2020). Phishing Detection Using Machine Learning Techniques.

Stephen Marsland 2015. *Machine learning: an algorithmic perspective*. CRC press.

S Y Lei (2018) Research on Precision Marketing of OPPO Mobile Phone under the Background
of Big Data (Hubei University of Technology).
2018 The global smartphone market is saturated with Huawei's millet against the surge (Screen
printing industry) p60.
H Si and Y J Jiang 2018 Bank Personal Credit Evaluation Based on SVM Optimization Algorithm (Industry and Technology Forum) pp 70-71.
Wibowo, D. A. (2018). Prediksi Penjualan Obat Herbal Hp Pro Menggunakan Algoritma Neural Network. Technologia: Jurnal Ilmiah, 9(1), 33–41. https://doi.org/10.31602/tji.v9i1.1100

Ge Na, Sun Lianying, et al. (2018) Forecast and analysis of sales volume based on ARIMA time series model[J]. Journal of Beijing Union University,32(04):27-33(in Chinese).

Chang Xiaohua, Xiong Ao. (2018) Application of Adaboost based random forest algorithm in medical sales forecast[J]. Computer Systems & Applications, 27(02):202-206(in Chinese).

`

APPENDICE 1

Ethical Form

The ethics form was done now because sometime at July I lost my devices, I just figured out when checking my checklist I haven't done ethics form.

1.

`

Ethics release checklist (ERC)

Project details

Project name:	USE OF MACHINE LEARNING FOR SALES FORECASTING
Principal investigator:	Pelumi Omage
Faculty:	Faculty of Business, Law and Digital Technologies
Level:	Postgraduate
Course:	Artificial intelligence and data scien
Unit code:	MAIDS
Supervisor name:	Joe Appleton
Other investigators:	

`

Question	Yes	No
Q1. Will the project involve human participants other than the investigator(s)?	○	◉
Q1a. Will the project involve **vulnerable participants** such as children, young people, disabled people, the elderly, people with declared mental health issues, prisoners, people in health or social care settings, addicts, or those with learning difficulties or cognitive impairment either contacted directly or via a **gatekeeper** (for example a professional who runs an organisation through which participants are accessed; a service provider; a care-giver; a relative or a guardian)?	○	◉
Q1b. Will the project involve the use of **control groups** or the **use of deception**?	○	◉
Q1c. Will the project involve any **risk to the participants' health** (e.g. intrusive intervention such as the administration of drugs or other substances, or vigorous physical exercise), or involve psychological stress, anxiety, humiliation, physical pain or discomfort to the investigator(s) and/or the participants?	○	◉
Q1d. Will the project involve **financial inducement** offered to participants other than reasonable expenses and compensation for time?	○	◉
Q1e. Will the project be carried out by individuals unconnected with the University but who wish to use staff and/or students of the University as participants?	○	◉
Q2. Will the project involve sensitive materials or topics that might be considered offensive, distressing, politically or socially sensitive, deeply personal or in breach of the law (for example criminal activities, sexual behaviour, ethnic status, personal appearance, experience of violence, addiction, religion, or financial circumstances)?	○	◉
Q3. Will the project have detrimental impact on the environment, habitat or species?	○	◉
Q4. Will the project involve living animal subjects?	○	◉
Q5. Will the project involve the development for export of 'controlled' goods regulated by the Export Control Organisation (ECO)? (This specifically means military goods, so called dual-use goods (which are civilian goods but with a potential military use or application), products used for torture and repression, radioactive sources.) Further information from the Export Control Organisation	○	◉
Q6. Does your research involve: the storage of records on a computer, electronic transmissions, or visits to websites, which are associated with terrorist or extreme groups or other security sensitive material? Further information from the Information Commissioners Office	○	◉

`

☑ I/we have assessed the ethical considerations in relation to the project in line with the University Ethics Policy.

☑ I/we understand that the ethical considerations of the project will need to be re-assessed if there are any changes to it.

☑ I/we will endeavor to preserve the reputation of the University and protect the health and safety of all those involved when conducting this research/enterprise project.

☑ If personal data is to be collected as part of my project, I confirm that my project and I, as Principal Investigator, will adhere to the General Data Protection Regulation (GDPR) and the Data Protection Act 2018. I also confirm that I will seek advice on the DPA, as necessary, by referring to the Information Commissioner's Office further guidance on DPA and/or by contacting information.rights@solent.ac.uk. By Personal data, I understand any data that I will collect as part of my project that can identify an individual, whether in personal or family life, business or profession.

☑ I/we have read the prevent agenda.

`

APPEDICE II

LSTM Model

1.

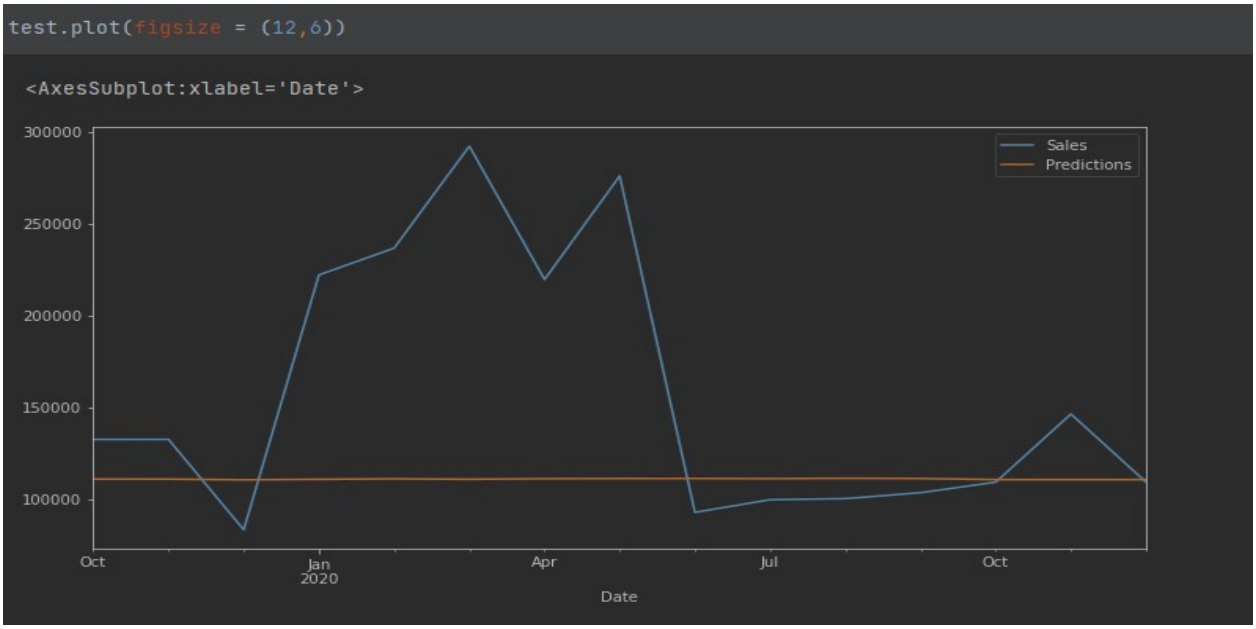
```
from keras.preprocessing.sequence import TimeseriesGenerator
#define generator
n_input = 12
n_features = 1
generator = TimeseriesGenerator(scaled_train, scaled_train, length=15, batch_size=1)
```

```
#defining the model
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import  LSTM

model_lstm = Sequential()
model_lstm.add(LSTM(units=100, activation='relu'
                    , input_shape=(n_input, n_features)))
model_lstm.add(Dense(1))
model_lstm.compile(optimizer='adam', loss='mse')
```

2.

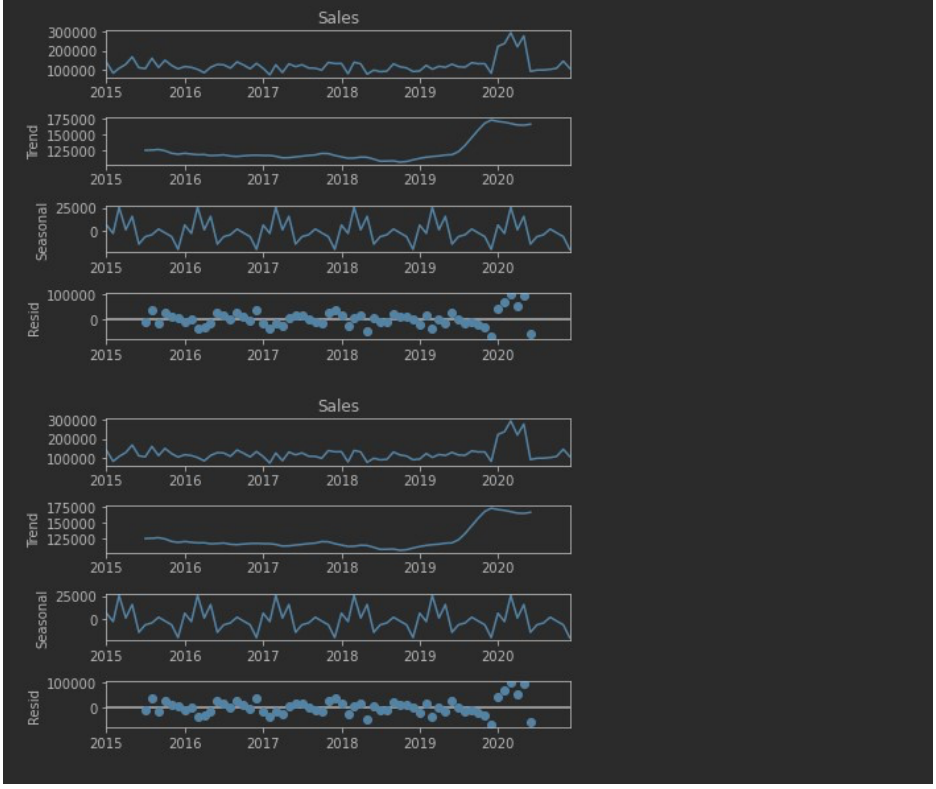
```
test.plot(figsize = (12,6))
```



`

3.

```
from statsmodels.tsa.seasonal import seasonal_decompose
results = seasonal_decompose(df['Sales'])
results.plot()
```



4.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaler.fit(train)
scaled_train = scaler.transform(train)
scaled_test = scaler.transform(test)
```

```
len(scaled_train)
```

```
57
```

5.

`

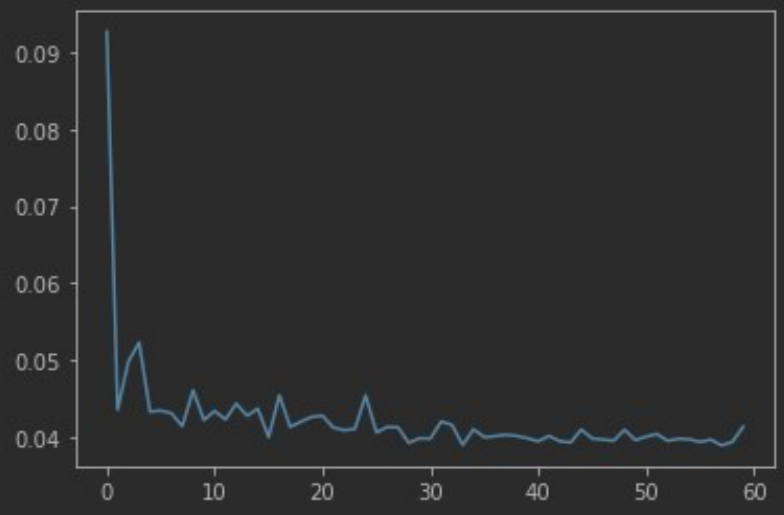
```
mod = sm.tsa.statespace.SARIMAX(df,
                                order=(0, 1, 1),
                                seasonal_order=(1, 1, 1, 12),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results = mod.fit()
print(results.summary().tables[1])
```

6.

```
model_lstm.fit(generator, epochs=60)
```

```
loss_per_epoch = model_lstm.history.history['loss']
plt.plot(range(len(loss_per_epoch)), loss_per_epoch)
```

```
[<matplotlib.lines.Line2D at 0x21e66d446a0>]
```



7.

`

```
import joblib
import streamlit as st
from streamlit_option_menu import option_menu
import pandas as pd
import matplotlib.pyplot as plt
from PIL import Image
import statsmodels.api as sm
import numpy as np

#Reading the xlsx file


df2 = pd.read_csv('samsung.csv')
df2['Date'] = pd.to_datetime(df2['Date'])
st.title('X Y Z SALES FORECASTING PLATFORM')


#sidebar menu


selected = option_menu(menu_title = None, options =["Products", "Analysis", "Forecast"],
                           icons=["phone", "house", "clock"], menu_icon="cast",
                           default_index=0,
                           orientation="horizontal",

                       styles={
                           "container": {"padding": "0!important"},
                           "icon": {"color": "orange", "font-size": "20px"},
                        "nav-link": {
                            "font-size": "20px", "text-align": "left",
                            "margin": "0px", "--hover-color": " #eee",
                        },
                        "nav-link-selected": {"background-color": "blue"},
                       },)
```

8.

```
def read_csv():
    df = pd.read_excel('sam_desc.xlsx')
    return df


def highest_sales():
    df = read_csv()
    df['OrderMonth'] = df['OrderDate'].dt.month
    df['OrderYear'] = df['OrderDate'].dt.month
    top_year = df.groupby(['OrderYear']).sum().sort_values('OrderYear', ascending=False)
    top_year = top_year[['Sales']]
    top_year.reset_index(inplace=True)

    return top_year

def highest_sales_graph():
    df = read_csv()
    df['OrderMonth'] = df['OrderDate'].dt.month
    df['OrderYear'] = df['OrderDate'].dt.month
    top_year = df.groupby(['OrderYear']).sum().sort_values('OrderYear', ascending=False)
    top_year = top_year[['Sales']]
    top_year.reset_index(inplace=True)

    fig = plt.figure(figsize=(10, 5))
    # plt.figure(figsize=(20, 15))
    plt.bar(top_year['OrderYear'], top_year['Sales'], color='blue', edgecolor='orange')
    plt.xticks(rotation='vertical')
    plt.title('OrderYear with highest sales', fontsize=15)
    plt.xlabel('OrderYear', fontsize=12)
    plt.ylabel('Sales', fontsize=12)
    plt.show()
    graph = st.pyplot(fig)
    return graph
```

9.

`

```
last_train_batch = scaled_train[-12:]
last_train_batch = last_train_batch.reshape((1, n_input, n_features))

model_lstm.predict(last_train_batch)

scaled_test[0]

test_prediction = []
first_eval = scaled_train[-n_input:]
present_batch = first_eval.reshape((1, n_input, n_features))

for i in range(len(test)):
    current_pred = model_lstm.predict(present_batch)[0]

    test_prediction.append(current_pred)

    present_batch = np.append(present_batch[:,1:,:],[[current_pred]], axis=1)
```