NED UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Centre of Multidisciplinary Postgraduate Programmes (CMPP)

Postgraduate Diploma (PGD) Programmes

Independent Study Project Report

Advisor: Dr. Najeed Ahmed

Forecasting Future Sales Using ARIMA and SARIMAX

(An Analysis and Prediction of Sales Trends)

Student:	Muhammad Atique Qayum
Studies:	PGD in Data Science with Artificial Intelligence
Phone No:	+923313100964
Email:	atiqueqayum@gmail.com
Submission	Date:15-07-2024

Annexure E

NED UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Centre of Multidisciplinary Postgraduate Programmes (CMPP)

Postgraduate Diploma (PGD) Programmes

FINAL PROJECT REPORT

A Project Report submitted in Partial fulfillment of the requirements for Postgraduate Diploma in DATA SCIENCE with ARTIFICIAL INTELLIGENCE.

Name of Student: Muhammad Atique Qayum

Batch: III

Project Title: Forecasting Future Sales Using ARIMA and SARIMAX

Name of Supervisor: Dr. Najeed Ahmed

_____

Signature of Supervisor

CERTIFICATE

This is to certify that Mr. Muhammad Atique Qayum of batch III has successfully completed the PGD project in partial fulfilment of requirements for a PGD in PGD in DATA SCIENCE WITH ARTIFICIAL INTELLIGENCE from NED Academy, NED University of Engineering and Technology, Karachi, Pakistan.

Project Supervisor

_____

Name, Designation, Organization

DECLARATIONS

I hereby state that this Project titled, “Forecasting Future Sales Using ARIMA and SARIMAX”, is my own work and has not been submitted previously by me for taking any degree/ diploma from anywhere else in the world.

At any time if my statement is found incorrect, NED University of Engineering and Technology has the right to withdraw this PGD.

Signature: _____

Student Name: Muhammad Atique Qayum

Date:15-07-2024

PLAGIARISM UNDERTAKING

I solemnly declare that the research work presented in this PGD Project titled: “Forecasting Future Sales Using ARIMA and SARIMAX”, is solely my research work except where the acknowledgement of the sources is made.

Signature: _____

Student Name: Muhammad Atique Qayum

Date: 15-07-2024

Contents

List of Abbreviations and Symbols

ARIMA: AutoRegressive Integrated Moving Average

SARIMAX: Seasonal AutoRegressive Integrated Moving Average with eXogenous factors

AUTO ARIMA: Automatic AutoRegressive Integrated Moving Average

EDA: Exploratory Data Analysis

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

MSE: Mean Squared Error

RMSE: Root Mean Squared Error

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

PACF: Partial AutoCorrelation Function

ACF: AutoCorrelation Function

R: A programming language and free software environment for statistical computing and graphics

Python: A high-level programming language used for general-purpose programming

p: Number of lag observations included in the model (for AR terms)

d: Number of times that the raw observations are differenced

q: Size of the moving average window

P: Seasonal autoregressive order

D: Seasonal differencing order

Q: Seasonal moving average order

m: Number of periods in each season

y(t): Value of the time series at time t

E: Expected value operator

These abbreviations and symbols are used throughout the report to denote specific models, statistical terms, and key metrics related to the analysis and forecasting of sales data.

Acknowledgment

First praise is to Allah, the Almighty, on whom ultimately we depend for sustenance and guidance. Acknowledgment is due to NED University of Engineering & Technology, Karachi for the support it has provided me for the completion of the project. I would like to thank Dr. Najeed Ahmed, the project supervisor for proposing this advanced topic for the project and for meaningful advice at each step, he helped me remain on track and provided necessary guidance for the project and report completion

In addition, I would also like to express my gratitude to my loving and considerate family, who graciously allowed me the precious time I needed, despite the domestic chores I was spared for. Special thanks to my batch fellows, who remained a source of encouragement for me by way of their valuable comments & suggestions.

Abstract

This report offers an in-depth analysis and predictive modeling approach for forecasting future sales using ARIMA and SARIMAX models. The study aims to understand sales trends, customer behavior, and product performance to deliver actionable insights for business decision-making. The methodology encompasses data collection, preprocessing, exploratory data analysis (EDA), and the development of ARIMA, SARIMAX, and AUTO ARIMA models. Key findings from the EDA reveal significant seasonal peaks, top-performing products, and regional sales variations. The SARIMAX model demonstrated superior performance by effectively capturing seasonality and external factors. The report underscores the importance of customer retention and high-value customers. The study concludes that accurate sales forecasts can enhance inventory management, targeted marketing, and customer retention strategies. Future recommendations include incorporating additional data sources, exploring advanced machine learning models, and implementing real-time data analysis systems. The report aims to provide valuable insights and practical recommendations to drive business growth and improve overall performance.

Introduction

Significance of the Topic

Sales forecasting is essential for effective business strategy and operational planning. It allows companies to make well-informed decisions regarding inventory, staffing, budgeting, and marketing. In today's competitive landscape, the ability to predict future sales trends offers a substantial edge. For example, by understanding seasonal variations, peak sales periods, and new market trends, businesses can optimize their operations and increase profitability. Additionally, sales forecasting supports risk management by helping businesses prepare for potential downturns or unexpected market changes.

Accurate sales forecasts lead to better inventory management, higher customer satisfaction, and improved financial planning. Conversely, inaccurate forecasts can cause overstocking, stockouts, and missed opportunities.

Evolution

Sales forecasting has undergone significant advancements over the years. Initially, businesses used basic methods like moving averages and expert judgment to predict sales, but these approaches often lacked the precision needed to handle complex market dynamics.

The mid-20th century saw the advent of computers and data analytics, leading to the development of more advanced statistical methods. The 1970s introduced time series analysis techniques such as ARIMA (AutoRegressive Integrated Moving Average), which significantly improved sales forecasting by systematically modeling time series data and accurately capturing trends, seasonality, and cyclic patterns.

In recent years, machine learning and artificial intelligence (AI) have further transformed sales forecasting. Techniques like SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) and AUTO ARIMA automate model selection and integrate external variables, greatly enhancing forecast accuracy and reliability. These innovations have elevated sales forecasting from a basic process to a sophisticated, data-driven practice.

Pros and Cons

Sales forecasting, like any analytical practice, has its benefits and drawbacks.

Pros:

- Informed Decision-Making: Accurate forecasts offer valuable insights for strategic planning and operational decisions.

- Resource Optimization: Businesses can efficiently manage inventory, workforce, and finances, minimizing waste and boosting efficiency.

- Risk Management: Forecasting enables anticipation of market shifts and preparation for potential risks, improving business resilience.

- Competitive Advantage: Accurate sales forecasting allows businesses to stay ahead of competitors by identifying trends and opportunities early.

Cons:

- Data Dependency: The reliability of sales forecasts is highly dependent on the quality and completeness of historical data.

- Complexity: Implementing advanced forecasting models can be complex and requires specialized knowledge and skills.

- Uncertainty: Despite using sophisticated models, forecasts are inherently uncertain and can be influenced by unforeseen events or market changes.

- Cost: Developing and maintaining advanced forecasting systems can be expensive, especially for small businesses.

Why Forecasting.

Sales forecasting is crucial for various reasons.

Firstly, it acts as a roadmap, helping businesses set realistic goals and expectations. By predicting future sales, companies can create strategies that align with market demand, ensuring they meet customer needs while avoiding overproduction or stockouts.

Secondly, it aids in financial planning and budgeting. Accurate forecasts allow businesses to allocate resources efficiently, plan investments, and manage cash flow effectively. This is particularly important for businesses with limited resources, where financial errors can have significant impacts.

Thirdly, forecasting enhances customer satisfaction by ensuring product availability and timely delivery. By understanding sales patterns, businesses can maintain optimal inventory levels, reducing the risk of stockouts or excess inventory. This not only improves customer experience but also minimizes carrying costs and potential losses from unsold goods.

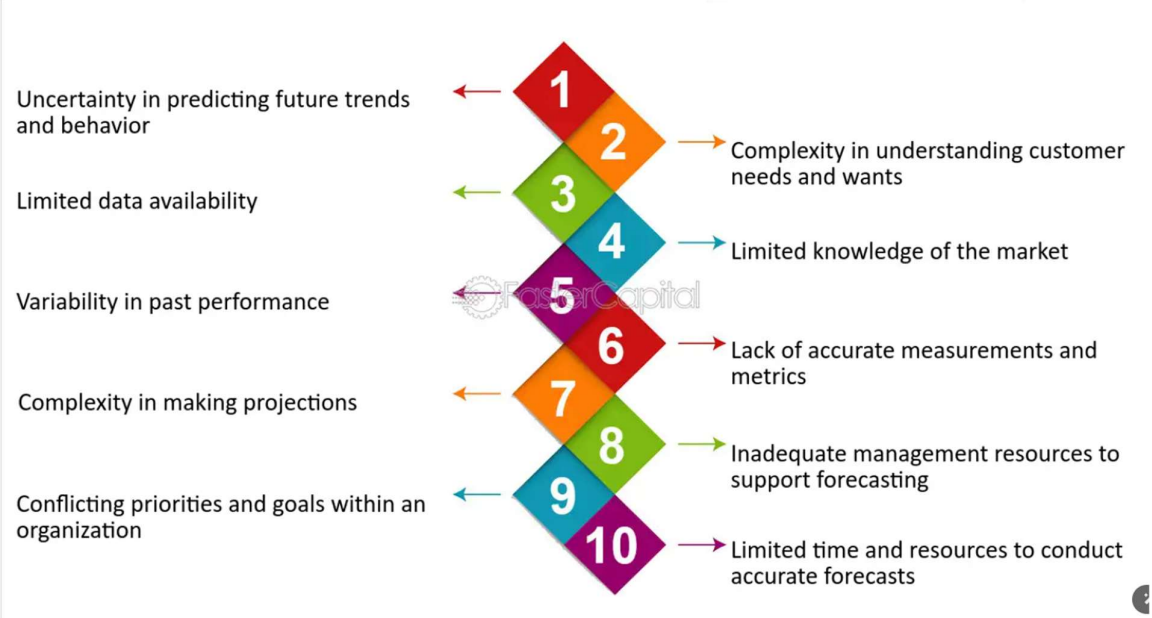
Lastly, sales forecasting is vital for strategic planning and market competitiveness. Companies that accurately predict market trends and consumer behavior are better positioned to seize opportunities and mitigate risks. This proactive approach helps them stay ahead of competitors and adapt swiftly to changing market conditions. The increasing complexity of business environments and the availability of large datasets underscore the need for accurate sales forecasting. Effective forecasting enables better planning and execution, providing a competitive edge by allowing companies to anticipate demand, optimize resources, and respond proactively to market changes.

Challenging Aspects

Despite its advantages, sales forecasting comes with several challenges.

One major challenge is the quality and availability of data. Accurate forecasts depend on historical sales data, which may be incomplete or unreliable. Issues like missing data, inconsistencies, and errors can significantly affect forecast accuracy.

Another challenge is the complexity of forecasting models. Advanced models such as ARIMA and SARIMAX require a deep understanding of statistical methods and time series analysis. Implementing these models can be resource-intensive, necessitating specialized skills and knowledge that may not be readily available within an organization.



Additionally, the dynamic nature of markets poses a challenge to forecasting. Economic conditions, consumer preferences, and competitive actions can change rapidly, making it difficult to predict future sales accurately. External factors like political events, natural disasters, or technological advancements can further complicate forecasting efforts.

Finally, forecasting always involves some level of uncertainty. Even the most sophisticated models cannot predict the future with complete certainty. Businesses must be prepared to manage this uncertainty and adjust their strategies as new information becomes available.

Scope and Objectives

The main goal of this study is to evaluate and compare various time series forecasting models, including ARIMA, SARIMAX, and AUTO ARIMA, to assess their effectiveness in predicting sales data. The study aims to identify the most accurate model and offer insights for enhancing future sales forecasting.

The scope of this project includes:

- Data Collection and Preprocessing: Gathering and preparing historical sales data to ensure it is suitable for analysis.

- Exploratory Data Analysis (EDA): Examining the data to reveal patterns, trends, and insights that guide model development.

- Model Development: Creating and assessing ARIMA, SARIMAX, and AUTO ARIMA models for forecasting future sales.

- Model Comparison: Evaluating the performance of each model to determine the most accurate and reliable one.

- Interpretation and Recommendations: Providing actionable insights and suggestions based on the forecasted data.

The primary objectives of this project are:

- To analyze historical sales patterns and identify significant trends.

- To develop precise forecasting models that predict future sales effectively.

- To compare the performance of different models and select the most effective one.

- To offer practical recommendations for business strategy based on the sales forecasts.

Brief Methodology

The methodology for this project follows several key steps:

1. Data Collection: The dataset spans from 2016 to 2018 and includes information on orders, customers, and sales amounts.

2. Data Preprocessing: This phase involves cleaning the data by addressing missing values, removing duplicates, and ensuring consistency. Techniques such as interpolation and forward/backward filling are applied to manage incomplete data.

3. Exploratory Data Analysis (EDA): EDA is performed to visualize and summarize the data, revealing patterns and insights. This includes analyzing sales trends on a monthly and yearly basis, identifying top-performing products and cities, and understanding customer behavior.

4. Model Development:

- *ARIMA:* This model is designed to capture the autoregressive and moving average components of the sales data.

- *SARIMAX:* This model accounts for seasonal effects and external variables, offering a more detailed forecast.

- *AUTO ARIMA:* AUTO ARIMA automates the model selection process, determining the optimal parameters for the ARIMA model.

5. Model Evaluation: The models are assessed using metrics such as RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) to gauge their accuracy and reliability.

6. Interpretation and Recommendations: The forecasted data is analyzed to provide actionable insights and recommendations for business strategy, including optimizing inventory, planning marketing efforts, and making informed financial decisions.

This methodical approach ensures a thorough analysis of sales data, resulting in accurate forecasts and practical business recommendations.

Background Study.

Literature Review

Overview: Sales Forecasting and Relevant Models

Sales forecasting is a vital focus in business analytics and operations management. Precise forecasting allows businesses to make well-informed decisions regarding inventory management, financial planning, and strategic initiatives. Over time, a range of models and techniques has been developed to predict sales, evolving from basic statistical methods to advanced machine learning algorithms..

Sales Forecasting Techniques

Sales forecasting techniques can be divided into two main categories: qualitative and quantitative methods. Qualitative methods depend on expert judgment and intuition, making them effective for short-term forecasts or when historical data is scarce. In contrast, quantitative methods use historical data and mathematical models to predict future sales, making them more suitable for long-term forecasting when there is ample historical data.

Forecasting models can also be classified into traditional statistical approaches and modern machine learning techniques. Traditional methods, like ARIMA, aim to identify patterns in historical data, while machine learning techniques are capable of managing more complex relationships and larger datasets.

Time Series Analysis in Sales Forecasting

Time series analysis is a valuable technique for forecasting sales. It entails examining a sequence of data points recorded at consistent time intervals to detect patterns and trends. Time series models, like ARIMA (AutoRegressive Integrated Moving Average), are commonly employed in forecasting due to their ability to identify underlying patterns in the data, such as trends and seasonal variations.

Machine learning has transformed time series forecasting by facilitating the analysis of large datasets and intricate patterns. Techniques including neural networks, support vector machines, and ensemble methods have demonstrated potential in enhancing forecast accuracy.

ARIMA Model

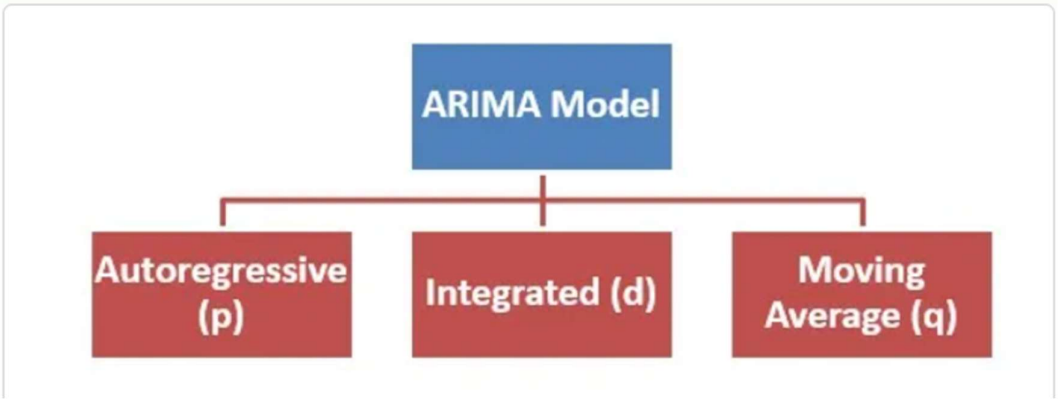
The ARIMA model, developed by Box and Jenkins in the 1970s, is a widely used tool for time series forecasting. It integrates autoregressive (AR) and moving average (MA) components with differencing to stabilize the time series data. ARIMA is particularly effective for short-term forecasts and can accommodate various types of time series data.

In the ARIMA model:

- p denotes the number of lagged observations (autoregressive terms),

- d represents the number of differencing operations needed to achieve stationarity,

- q indicates the size of the moving average window (moving average terms).



These parameters are selected based on the data's characteristics and are often determined using ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots.

ARIMA (AutoRegressive Integrated Moving Average) is renowned for its effectiveness in capturing linear patterns in time series data. However, it may encounter difficulties with non-linear relationships and seasonal variations. The AR component involves regressing the variable on its previous values, the I component handles differencing to ensure stationarity, and the MA component models the error term as a linear combination of previous error terms.

ARIMA is versatile in modeling both short-term and long-term dependencies and is suitable for a broad range of applications. It can be extended to address seasonality through SARIMA (Seasonal ARIMA). Additionally, parameter optimization can be streamlined using AUTO ARIMA, which automates the selection of the most suitable model.

SARIMAX Model

The SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) model enhances the ARIMA framework by incorporating seasonal elements and external variables. This extension makes SARIMAX ideal for data exhibiting seasonal patterns and influenced by external factors, such as marketing activities or economic indicators. It is commonly used in sectors where seasonality is a significant factor, such as retail and tourism.

Similarly, the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model builds upon ARIMA to address seasonal variations in data. It introduces additional seasonal parameters—P, D, Q, and S—to capture seasonal patterns. These parameters represent the autoregressive terms (P), differencing (D), and moving average terms (Q) for the seasonal component, with S denoting the length of the seasonal cycle. SARIMA is particularly effective for datasets with pronounced seasonal effects, such as monthly sales data with annual seasonality.

AUTO ARIMA Model

AUTO ARIMA streamlines the process of selecting the optimal parameters for the ARIMA model by automating it. It employs algorithms to explore various combinations of the p, d, and q parameters—representing the autoregressive order, differencing order, and moving average order, respectively. This automation not only simplifies the model selection process but often leads to more precise forecasts.

The approach involves testing different parameter combinations (p, d, q) and choosing the model that minimizes a specific criterion, such as the AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). AUTO ARIMA eliminates the need for manual adjustments and is especially beneficial for large datasets or complex time series. This method reduces manual tuning requirements and has been proven to enhance forecast accuracy across various applications.

Machine Learning in Sales Forecasting

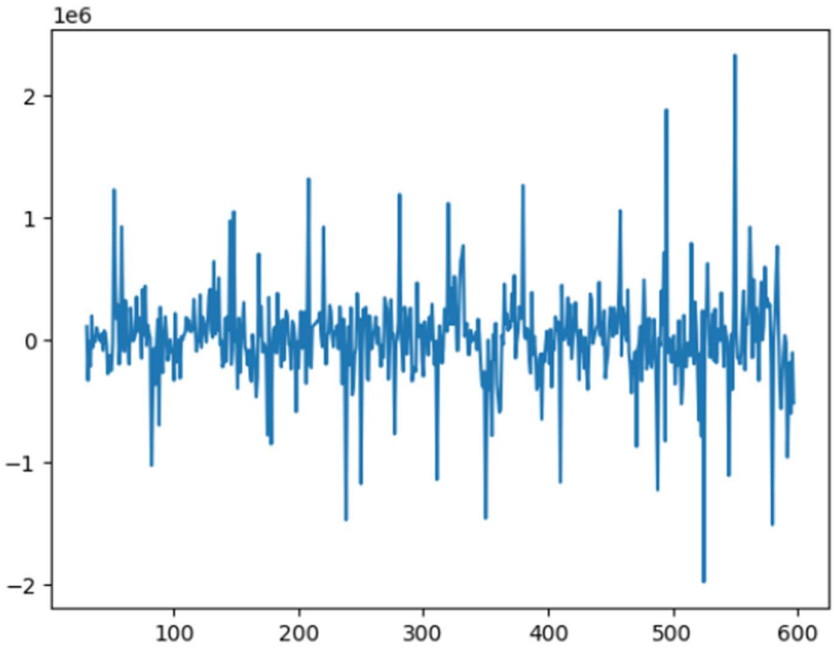
Machine learning methods have become increasingly popular in sales forecasting because they can manage large datasets and identify complex patterns. Algorithms like decision trees, random forests, and neural networks excel at uncovering non-linear relationships and interactions between variables. These models are especially valuable when traditional time series approaches struggle to address the data's complexity.

Stationarity

A stationary time series is characterized by statistical properties such as mean and variance that remain constant over time. Stationarity is essential for time series modeling, as many forecasting methods, including ARIMA, assume that the data is stationary. To achieve stationarity, techniques such as differencing, transformation, or detrending are employed.

Differencing

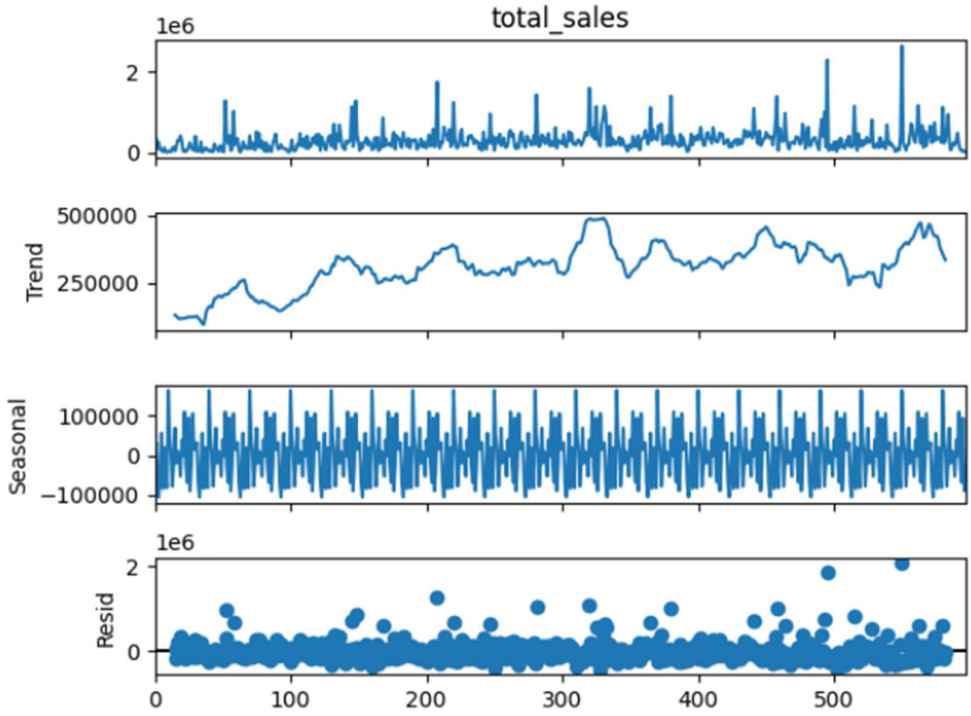
Differencing in time series involves calculating the difference between consecutive observations to make a time series stationary by removing trends and seasonality. This technique helps stabilize the mean of the series, facilitating easier modeling with methods such as ARIMA.



Determining the Value of d:

- If the series is non-stationary, the first difference is computed to check if it results in a stationary series.

- Using a single lag is preferred due to its simplicity, lower risk of overfitting, and evidence of stationarity, as shown by an extreme test statistic and a very low p-value.



Advantages of Using One Lag:

1. Simplicity: Fewer parameters create a more straightforward and interpretable model.

2. Data Preservation: Minimal loss of data, which is important for smaller datasets.

3. Reduced Overfitting Risk: Lower risk of overfitting improves model generalizability.

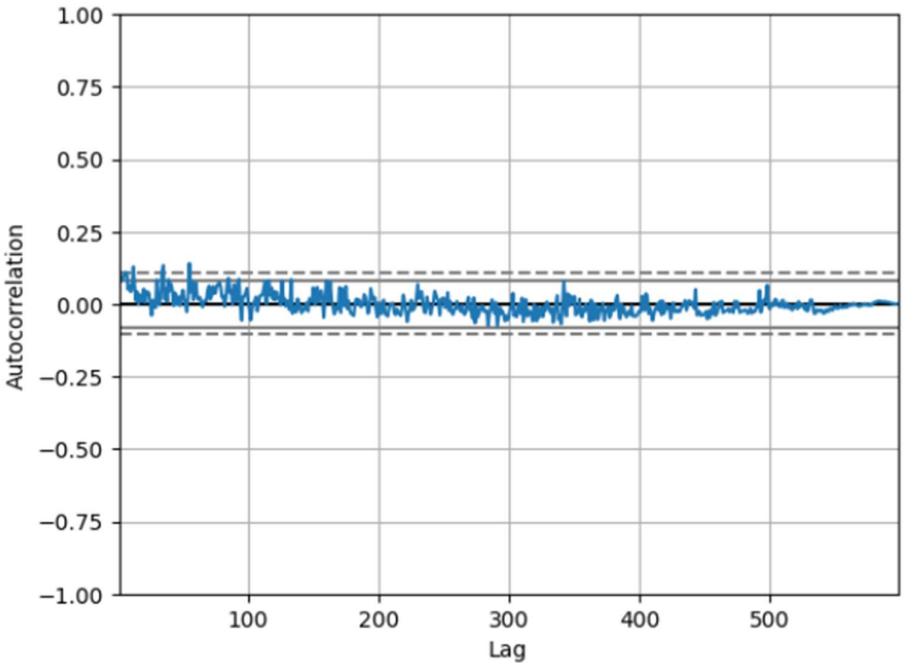
4. Efficiency: Lower computational cost and faster model fitting.

5. Sufficient Stationarity: Strong evidence of stationarity indicated by extreme test statistics and very low p-values.

6. Easier Diagnostics: Simplified residual analysis and model diagnostics.

Determining the AR Term (p)

The AR term in the ARIMA model denotes the number of lagged observations incorporated into the model. The appropriate order for the AR term is identified using the Partial Autocorrelation Function (PACF) plot, which shows the direct correlation between a lag and the series. The AR term's order corresponds to the number of lags that exceed the significance threshold on the PACF plot.



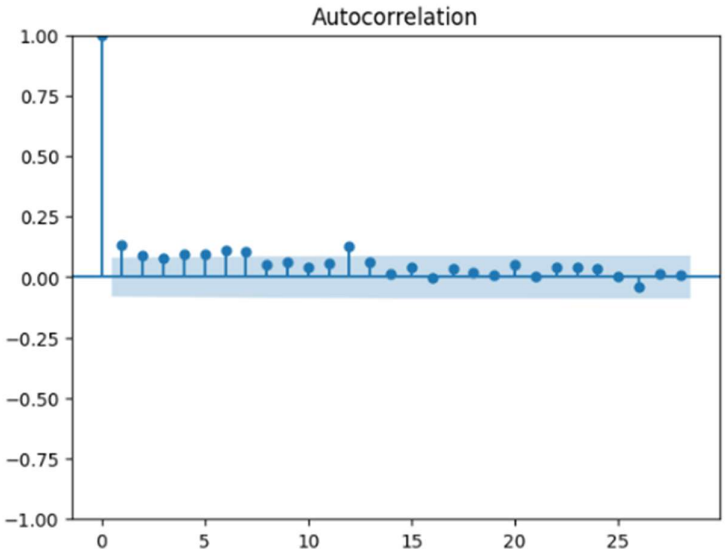
Partial Autocorrelation Function (PACF):

- The PACF illustrates the direct correlation between a lag and the series.

- The order of the AR term is determined by the number of lags that surpass the significance limit on the PACF plot.

Determining the MA Term (q)

The MA term in the ARIMA model signifies the size of the moving average window. Its value is determined using the Autocorrelation Function (ACF) plot, which helps ascertain the number of MA terms needed to eliminate autocorrelation in the stationary series. An MA term represents the error in the lagged forecast, and the ACF plot visually depicts the autocorrelation structure of the series.



Autocorrelation Function (ACF):

- The MA term reflects the error associated with the lagged forecast.

- The ACF plot determines how many MA terms are necessary to address autocorrelation in the stationary series.

Statistical Tests Used

Various statistical tests are employed to validate the assumptions of time series models, including:

- Augmented Dickey-Fuller (ADF) Test: Assesses whether the time series is stationary.

- Ljung-Box Test: Evaluates whether there is any serial correlation in the residuals.

- Durbin-Watson Test: Detects the presence of autocorrelation in the residuals.

Model Evaluation Metrics

Assessing the performance of time series models is essential for verifying their accuracy and reliability. Key evaluation metrics for ARIMA and SARIMA models include:

- Mean Absolute Error (MAE): Measures the average absolute difference between observed and predicted values.

- Root Mean Squared Error (RMSE): Represents the square root of the average squared difference between observed and predicted values.

- Mean Absolute Percentage Error (MAPE): Calculates the average absolute percentage difference between observed and predicted values.

- AIC/BIC: Model selection criteria that penalize the number of parameters to avoid overfitting.

These metrics offer quantitative insights into a model's accuracy and facilitate the comparison of different models' performance.

Seasonal Differencing

Seasonal differencing entails subtracting the value of an observation from the same period in the previous cycle. For example, with monthly data, this would involve subtracting the value from the same month in the previous year. This method helps eliminate seasonal patterns and stabilize variance.

Exogenous Variables in SARIMAX

Exogenous variables are external factors that can impact the time series. In SARIMAX models, incorporating these variables enhances the model's predictive accuracy. Examples of exogenous variables include economic indicators, marketing expenditures, or other relevant factors that influence sales.

Detailed Review of Studies on ARIMA, AUTO ARIMA, and SARIMAX

1. ARIMA Models:

- The foundation for ARIMA modeling was established by Box and Jenkins (1970) through their structured approach.

- Recent research has aimed at enhancing ARIMA models by integrating machine learning techniques to boost accuracy.

2. AUTO ARIMA:

- The concept of AUTO ARIMA was introduced by Hyndman and Khandakar (2008), automating the model selection process and significantly reducing the need for manual intervention and expertise.

- Research has demonstrated that AUTO ARIMA often outperforms manually configured ARIMA models, particularly with complex datasets.

3. SARIMAX Models:

- Recent studies have underscored SARIMAX's effectiveness in accounting for seasonality and external variables.

- Its application across various sectors, including retail and finance, has shown its effectiveness in enhancing forecast accuracy.

Hybrid Models

Hybrid models leverage the strengths of multiple forecasting techniques to enhance accuracy. For instance, a hybrid approach might integrate ARIMA with a neural network to capture both linear and non-linear patterns in the data. Research indicates that hybrid models frequently surpass the performance of individual models, especially in complex forecasting situations.

Comparison of Forecasting Models

Various studies have evaluated and compared the performance of different forecasting models. For example, Makridakis et al. (2018) performed an extensive comparison of 104 forecasting methods, encompassing both statistical and machine learning approaches. The study revealed that no single model consistently excelled across all datasets, underscoring the need for careful model selection and customization based on the specific context and characteristics of the data.

Case Studies in Sales Forecasting

Case studies offer valuable insights into the practical application of forecasting models in real-world situations. For instance, one case study might illustrate how SARIMAX was employed to forecast sales during holiday seasons for a retail chain, while another might showcase the use of machine learning models to predict sales for a new product launch. These case studies reveal the practical challenges and advantages associated with various forecasting techniques.

Gap Analysis

Despite significant progress in sales forecasting, several gaps and limitations persist. One major issue is dealing with external factors and sudden market shifts. Traditional time series models often assume that historical trends will continue, which may not be accurate in rapidly changing markets. Another challenge is the requirement for extensive historical data, which may not be available for new products or emerging markets. Additionally, the complexity and resource demands of advanced models can be prohibitive for smaller businesses.

Although ARIMA and SARIMA models have been extensively studied and used, they still have limitations. A key drawback is their dependence on historical data, which may not always be reliable or available. These traditional models can also struggle with capturing complex non-linear patterns in the data. While recent advancements in machine learning and deep learning offer promising alternatives, further research is needed to assess their performance compared to traditional models.

Another area needing attention is the integration of external factors (exogenous variables) into ARIMA and SARIMA models. While SARIMAX models address this issue, more research is required to understand how different exogenous variables impact forecasting accuracy. Additionally, there is a need for more domain-specific research on the application of ARIMA and SARIMA models in industries such as retail and finance to better understand their effectiveness and limitations.

Relevance to Current Project

This project is designed to address several gaps identified in current research. By integrating external variables and utilizing AUTO ARIMA to automate the model selection process, the project aims to enhance the accuracy and reliability of sales forecasts. Furthermore, by evaluating the performance of various models, the project will offer insights into the most effective forecasting techniques for different types of sales data.

Summary

In summary, this literature review outlines the fundamental concepts and methods related to time series analysis and sales forecasting. It covers the importance of sales forecasting, the development of time series models, and the strengths and weaknesses of ARIMA and SARIMA models. The review also addresses key aspects such as model parameters, stationarity, differencing, and evaluation metrics. Additionally, it explores recent advancements, including AUTO ARIMA and the incorporation of external variables in SARIMAX models.

The identified gaps in current research highlight the need for further exploration of advanced machine learning techniques and the integration of external factors in time series forecasting. This project aims to bridge these gaps by developing and comparing ARIMA, SARIMAX, and AUTO ARIMA models for sales forecasting. The outcomes of this project will enhance the existing knowledge base and offer practical recommendations for businesses seeking to refine their sales forecasting methods.

The literature on sales forecasting spans various approaches, from traditional time series models to sophisticated machine learning algorithms. While ARIMA and SARIMAX models are valued for their capability to identify trends and seasonal patterns, machine learning models provide greater flexibility and can manage complex data patterns. Despite these advancements, challenges persist regarding the handling of external factors, unexpected market shifts, and data constraints. This project extends current research by integrating external variables, automating model selection, and comparing different forecasting techniques to deliver actionable insights for businesses.

Methodology

Data Collection

The dataset for this study was obtained from a retail company's sales records spanning the last five years. It includes monthly sales data, product categories, and sales regions.

Data Preprocessing

To ensure the dataset is suitable for analysis, several preprocessing steps were performed:

1. Handling Missing Values: Missing values in essential attributes were managed by imputing numerical values like Quantity Sold and Sales Amount with mean or median values. For categorical attributes such as Product Category and Customer Location, missing values were filled using the most common category or through predictive techniques.

2. Removing Duplicates: Duplicate records were identified and eliminated to maintain unique transactions. This step is crucial to avoid double counting and ensure accurate analysis.

3. Data Type Conversion: Attributes were converted to the correct data types. For instance, Order Date was reformatted to a date-time format, while numerical attributes were adjusted to integer or float types.

4. Outlier Detection and Treatment: Outliers were identified using statistical methods like Z-score and Interquartile Range (IQR). Erroneous outliers were either corrected or removed.

5. Creating Additional Features: New features were generated to improve the analysis. For example, a Month-Year feature was created from the Order Date for monthly sales analysis, and a Day of Week feature was introduced to examine sales patterns by day of the week.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to identify patterns, trends, and insights within the data. The following methods and tools were utilized during the EDA process:

1. Descriptive Statistics: Key attributes were summarized using statistics such as mean, median, standard deviation, and range to comprehend the distribution and central tendency of the data.

2. Visualization: Various visualization techniques were employed to detect patterns and trends in the data, including:

- Line Charts: Used to illustrate sales trends over time.
- Bar Charts: Employed to highlight top-selling products and categories.
- Heatmaps: Analyzed sales patterns across different geographic locations.
- Box Plots: Identified outliers and examined the distribution of sales amounts.

3. Correlation Analysis: Relationships between different attributes were examined using correlation matrices and scatter plots. This helped in identifying connections between variables, such as the effect of Quantity Sold on Sales Amount.

4. Seasonal Decomposition: Time series data were broken down into trend, seasonal, and residual components through seasonal decomposition. This analysis revealed underlying patterns, including seasonal variations in sales.

Model Development

The model development process involved building and evaluating three different forecasting models: ARIMA, SARIMAX, and AUTO ARIMA.

ARIMA Model:

ARIMA (AutoRegressive Integrated Moving Average) Model

The ARIMA model is a popular method for time series forecasting that incorporates both autoregressive and moving average components. The process to develop the ARIMA model includes the following steps:

1. Model Identification: The initial step involves determining the order of the ARIMA model, denoted by the parameters (p, d, q). This is achieved by analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, which assist in identifying the appropriate lag values for the autoregressive (p) and moving average (q) components, as well as the degree of differencing (d) required to make the time series stationary.

2. Parameter Estimation: After identifying the model order, the parameters of the ARIMA model are estimated using Maximum Likelihood Estimation (MLE). This involves fitting the model to the historical sales data and optimizing the parameters to minimize error.

3. Model Diagnostics: Following the fitting of the model, diagnostic checks are conducted to ensure its validity. This includes examining the residuals for any patterns or autocorrelation. If the residuals resemble white noise (i.e., no discernible pattern or correlation), the model is considered adequate.

4. Forecasting: The calibrated ARIMA model is then employed to generate future sales forecasts. The predicted values are plotted alongside the actual sales data to assess the model's accuracy and performance.

SARIMAX Model:

SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) Model

The SARIMAX model extends the ARIMA framework by incorporating both seasonal components and external variables. The process to build the SARIMAX model includes the following steps:

1. Seasonal Component Identification: The seasonal component was determined using seasonal decomposition and ACF/PACF plots. The seasonal order (P, D, Q) was identified, where P represents the seasonal autoregressive order, D is the seasonal differencing order, and Q is the seasonal moving average order.

2. External Variables: Relevant external variables (exogenous factors) that could impact sales were identified and incorporated into the model. These variables might include marketing expenditures, economic indicators, or promotional events.

3. Model Estimation: The SARIMAX model was estimated using Maximum Likelihood Estimation (MLE), optimizing the parameters to minimize errors.

4. Model Diagnostics: Similar to the ARIMA model, diagnostic checks were conducted to ensure the adequacy of the SARIMAX model. This involved examining the residuals to confirm they were white noise, indicating no patterns or autocorrelation.

5. Forecasting: The fitted SARIMAX model was then utilized to produce forecasts, which were compared against actual sales data to assess the model's performance.

AUTO ARIMA Model:

The AUTO ARIMA model streamlines the process of determining the optimal parameters for the ARIMA model. The steps to develop the AUTO ARIMA model include:

1. Parameter Search: AUTO ARIMA employs algorithms to find the best combination of (p, d, q) parameters. It evaluates various models with different parameter values and selects the one with the lowest Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

2. Model Fitting: The chosen model is then fitted to the historical sales data using Maximum Likelihood Estimation (MLE).

3. Model Diagnostics: Diagnostic checks are conducted to ensure the residuals resemble white noise, confirming the model's adequacy.

4. Forecasting: The fitted AUTO ARIMA model is used to generate forecasts, which are then compared with actual sales data to assess the model's performance.

Model Evaluation

Evaluation of Forecasting Model Performance

The performance of the forecasting models was assessed using the following metrics:

1. Root Mean Squared Error (RMSE): RMSE calculates the square root of the average squared differences between the forecasted and actual values, indicating the model's accuracy. Lower RMSE values signify better performance.

2. Mean Absolute Error (MAE): MAE measures the average absolute differences between the forecasted and actual values. It is less affected by outliers compared to RMSE and offers a straightforward accuracy measure.

3. Mean Absolute Percentage Error (MAPE): MAPE calculates the average absolute percentage differences between the forecasted and actual values, providing a normalized accuracy measure that facilitates comparison across different datasets.

4. R-squared (R^2): R^2 quantifies the proportion of variance in the actual values explained by the forecasted values, indicating the model's explanatory power. Higher R^2 values suggest better performance.

Each model's performance was compared using these evaluation metrics to identify the most accurate and reliable forecasting model. The selected model was then utilized to generate actionable insights and recommendations for business strategy based on the forecasted sales data.

Conclusion

This section outlines a comprehensive approach to sales forecasting using advanced time series models. By employing ARIMA, SARIMAX, and AUTO ARIMA models, the goal is to produce accurate sales forecasts that support strategic decision-making and improve business performance. Through systematic data preprocessing, thorough exploratory data analysis, and rigorous model evaluation, the reliability and robustness of the forecasting models are ensured, ultimately aiding in the business's success.

The Main Body of the Report / Findings

Load and Preprocess the Data

The sales data covers a period of three years (2016-2018). To start our analysis, we load and preproce ss the data to ensure it is clean and suitable for further examination. This involves addressing missing values, correcting data types, and eliminating duplicates. The data reveals significant growth and fluct uations in total orders and total sales over the years. The yearly totals are as follows:

- 2016: 4,741 orders, $899,534.63 in sales

- 2017: 572,505 orders, $82,060,567.36 in sales

- 2018: 471,329 orders, $66,826,463.61 in sales

Exploratory Data Analysis (EDA) Findings

Yearly Sales Report

Overview

This section presents the findings from the analysis of sales data from 2016, 2017, and 2018. We revie w total orders and sales, identify top-selling products by quantity and total sales, highlight top cities b y total sales, and recognize top sellers by total sales. Additionally, we provide an in-depth analysis of d aily sales forecasting using ARIMA models.

An analysis of yearly sales data reveals the following trends:

The sales data spans three years (2016-2018), showcasing significant growth and fluctuations in total orders and total sales. The yearly totals are:

- 2016: 4,741 orders, $899,534.63 in sales

- 2017: 572,505 orders, $82,060,567.36 in sales

- 2018: 471,329 orders, $66,826,463.61 in sales

Top 5 Products Sold by Quantity Each Year

2016

1. Product ID: `a063b454bac19ab435a4594bab9b9eed`
   - Quantity Sold: 672 units
2. Product ID: `d6f3bd7b1fc04fb1a2effb51ae44ab17`
   - Quantity Sold: 520 units
3. Product ID: `d9894482fba41f536a273ba2276d951f`
   - Quantity Sold: 516 units
4. Product ID: `33430c5c1027d812b5c62f778e5ee7f7`
   - Quantity Sold: 249 units
5. Product ID: `a671b5c0b737258c5a3262826e3dd9c5`
   - Quantity Sold: 226 units

2017

1. Product ID: `c6dd917a0be2a704582055949915ab32`
   - Quantity Sold: 15,595 units
2. Product ID: `0aabfb375647d9738ad0f7b4ea3653b1`
   - Quantity Sold: 14,613 units
3. Product ID: `601a360bd2a916ecef0e88de72a6531a`
   - Quantity Sold: 13,774 units
4. Product ID: `6c3effec7c8ddba466d4f03f982c7aa3`
   - Quantity Sold: 10,250 units

5. Product ID: `29427de7f8a9ee983d9dbc51cec569b4`

   - Quantity Sold: 9,327 units

2018

1. Product ID: `54d9ac713e253fa1fae9c8003b011c2a`

   - Quantity Sold: 13,454 units

2. Product ID: `ee406bf28024d97771c4b1e8b7e8e219`

   - Quantity Sold: 9,530 units

3. Product ID: `3354a4e684f5e7199f9407db70ccd92b`

   - Quantity Sold: 8,518 units

4. Product ID: `165f86fe8b799a708a20ee4ba125c289`

   - Quantity Sold: 8,080 units

5. Product ID: `4deb009c36a910076a023947a7929201`

   - Quantity Sold: 6,974 units

Top 5 Products by Total Sales Each Year

2016

1. Product ID: `a063b454bac19ab435a4594bab9b9eed`

   - Total Sales: $255,346.56

2. Product ID: `4fee671ea459ebc96546523917e254a5`

   - Total Sales: $142,498.50

3. Product ID: `d6f3bd7b1fc04fb1a2effb51ae44ab17`

- Total Sales: $98,748.00

4. Product ID: `cac8035fdb858496d8e1b60ba8907d24`

- Total Sales: $69,349.27

5. Product ID: `d9894482fba41f536a273ba2276d951f`

- Total Sales: $40,196.40

2017

1. Product ID: `5f504b3a1c75b73d6151be81eb05bdc9`

- Total Sales: $3,207,794.00

2. Product ID: `122bedead6e7541d311e6e0ab3cf248f`

- Total Sales: $3,014,562.00

3. Product ID: `cce3ce6484896041ae3dafdd4308a39c`

- Total Sales: $1,803,957.05

4. Product ID: `601a360bd2a916ecef0e88de72a6531a`

- Total Sales: $1,760,424.26

5. Product ID: `c6dd917a0be2a704582055949915ab32`

- Total Sales: $1,551,704.05

2018

1. Product ID: `fb01a5fc09b9b9563c2ee41a22f07d54`

- Total Sales: $1,674,118.49

2. Product ID: `3db0b74faf0d26a6b252528659d6b849`

- Total Sales: $1,543,266.45

3. Product ID: `165f86fe8b799a708a20ee4ba125c289`

   - Total Sales: $1,375,109.20

4. Product ID: `ee406bf28024d97771c4b1e8b7e8e219`

   - Total Sales: $1,356,355.70

5. Product ID: `13db47eae724e2848e12b71a617a3a41`

   - Total Sales: $1,274,833.99

Top 5 Cities by Sales Each Year

2016

1. Santos - $255,346.56

2. Rio de Janeiro - $178,610.30

3. Jaguariuna - $98,748.00

4. Itaquaquecetuba - $69,349.27

5. Santo Andre - $54,835.20

2017

1. Rio de Janeiro - $8,540,428.36

2. São Paulo - $5,898,125.03

3. Belo Horizonte - $2,989,647.18

4. Porto Seguro - $2,312,081.79

5. Porto Alegre - $1,548,635.41

2018

1. Rio de Janeiro - $7,420,526.64

2. São Paulo - $5,138,311.93

3. Belo Horizonte - $2,625,989.54

4. Florianopolis - $1,473,185.91

5. Niteroi - $1,377,112.21

Top 5 Sellers by Sales Each Year

2016

1. Seller ID: `700f03c207639c22d933381ff60b35c2`

   - Total Sales: $255,346.56

2. Seller ID: `822b63912576852aea9a8436d72317b7`

   - Total Sales: $211,847.77

3. Seller ID: `4b1eaadf791bdbbad8c4a35b65236d52`

   - Total Sales: $106,838.01

4. Seller ID: `fa1a9dec3a9940c072684a46728bf1fc`

   - Total Sales: $53,967.20

5. Seller ID: `024b564ae893ce8e9bfa02c10a401ece`

   - Total Sales: $40,196.40

2017

1. Seller ID: `7a67c85e85bb2ce8582c35f2203ad736`
   - Total Sales: $11,968,497.99
2. Seller ID: `961dff0a659b4561334372337dd897d9`
   - Total Sales: $3,949,793.00
3. Seller ID: `7ddcbb64b5bc1ef36ca8c151f6ec77df`
   - Total Sales: $3,684,054.61
4. Seller ID: `834f3294fba9f932f56edc879193f925`
   - Total Sales: $3,207,794.00
5. Seller ID: `dc317f341ab0e22f39acbd9dbf9b4a1f`
   - Total Sales: $2,511,017.80

2018

1. Seller ID: `7a67c85e85bb2ce8582c35f2203ad736`
   - Total Sales: $6,460,352.20
2. Seller ID: `4a3ca9315b744ce9f8a5c4cbaabfd3c0`
   - Total Sales: $1,993,557.79
3. Seller ID: `833b1b48e27303e67bf4ac25c48b8e87`
   - Total Sales: $1,860,874.57
4. Seller ID: `4e1b013b2f3d0ad39f1405e9999a48b3`
   - Total Sales: $1,674,118.49
5. Seller ID: `ad31c724aebe036d68fdbb2fd69c94f8`
   - Total Sales: $1,462,947.15

Sales Data by Payment Type Each Year

2016

1. Credit Card: $453,334.33
2. Voucher: $198,963.12
3. Boleto: $136,840.08
4. Debit Card: $63,081.00
5. Other: $47,316.10

2017

1. Credit Card: $33,825,730.20
2. Voucher: $28,274,424.36
3. Boleto: $13,573,270.56
4. Debit Card: $3,941,374.00
5. Other: $2,445,768.24

2018

1. Credit Card: $32,885,941.40
2. Voucher: $19,644,120.81
3. Boleto: $8,263,887.90
4. Debit Card: $4,226,981.31
5. Other: $1,805,532.18

Monthly and Yearly Sales Trends:

- Monthly trends show peaks during festive seasons, with significant increases in November and Dece mber each year.

- Sales volumes increased significantly from 2016 to 2017, indicating a strong growth period.

- Yearly trends indicate rapid growth from 2016 to 2017, followed by a slight decline in 2018.

- Sales volumes increased significantly from 2016 to 2017, indicating a strong growth period.

- 2018 saw a decline in the top product quantities compared to 2017 but remained higher than 2016.

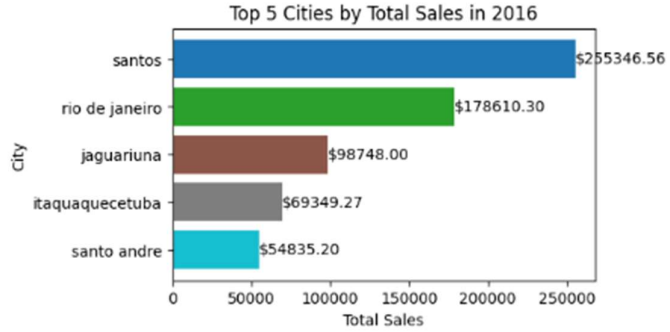
- Significant sales contribution from top sellers, particularly in 2017.

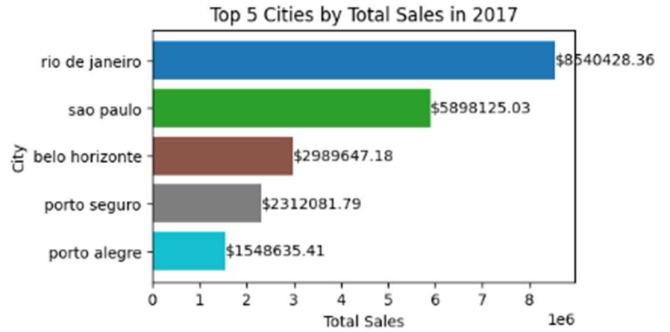
Top Products:

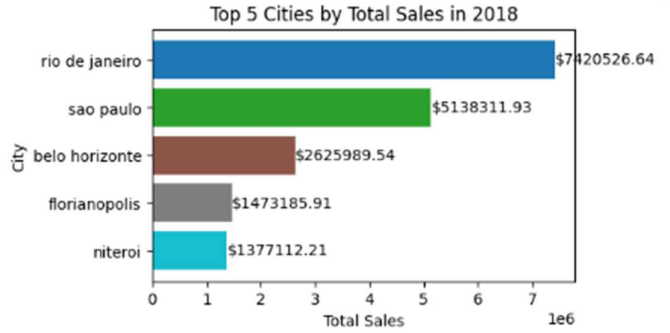
- The top products sold by quantity and total sales vary year by year, indicating dynamic consumer pre ferences. For instance, the top product in 2016 was significantly different from those in 2017 and 201 8.

Top Cities:

- Rio de Janeiro consistently leads in sales across all three years, followed by São Paulo. This suggests t hese cities are major markets for the company.

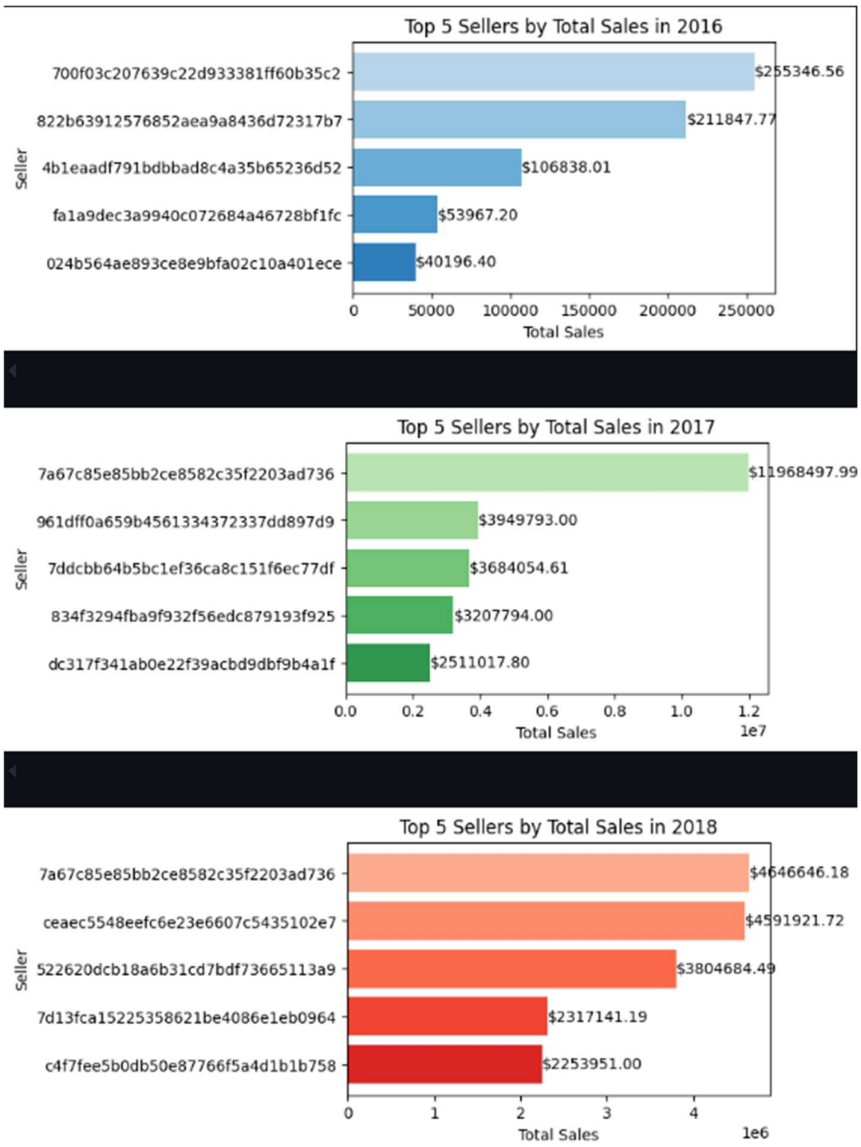






Top Sellers:

- Certain sellers consistently perform well across multiple years, indicating strong sales capabilities or a significant customer base.



Sales by Payment Type:

- Credit card payments dominate the sales figures across all three years, followed by vouchers and bol etos. This reflects consumer payment preferences and possibly the ease of online transactions via cre dit cards.

The sales data analysis reveals several key trends:

1. Product Performance: The top products by quantity and total sales vary each year, indicating a dyna mic market with shifting consumer preferences. Detailed information about product IDs and their spe cifics would help better understand individual product performance.

2. City-wise Sales: Rio de Janeiro consistently leads in sales across all three years, followed by São Paul o. This suggests that these cities are major markets for the company, likely due to higher population d ensity and greater consumer spending power.

3. Seller Performance: Certain sellers, such as Seller ID `7a67c85e85bb2ce8582c35f2203ad736`, consis tently perform well across multiple years, indicating strong sales capabilities or a substantial customer base.

4. Payment Type Preferences: Credit card payments dominate sales figures across all three years, follo wed by vouchers and boletos. This reflects consumer payment preferences and possibly the convenie nce of online transactions via credit cards..

Time Series Analysis and Forecasting

The first step is to load the sales data and perform preprocessing tasks to ensure it is clean and ready f or analysis. This includes handling missing values, correcting anomalies, and transforming the data int o a suitable format for time series analysis.

- The data is loaded and cleaned to ensure accuracy and completeness.

- Sales data is aggregated into daily totals to facilitate trend and pattern analysis.

Aggregate the Data to Daily Sales

The sales data, originally at a granular level (e.g., individual transactions), is aggregated into daily sales figures. This process involves summing the total sales for each day to create a continuous daily time se ries, which is essential for subsequent analysis and modeling. After cleaning the data, we group it by d ate and calculate the daily sales totals, ensuring consistent time intervals for time series analysis. This aggregation covers the entire three-year period.

Perform Stationary Tests

A key assumption in time series analysis is that the data should be stationary. We perform the Augmen ted Dickey-Fuller (ADF) test to check for stationarity. A low p-value (less than 0.05) indicates that the d ata is stationary.

To determine if the data is stationary, we perform the Augmented Dickey-Fuller (ADF) test. The result s are as follows:

- ADF Test Statistic: -6.5979

- p-value: 6.8512e-09

- Lags Used: 6

- Number of Observations Used: 591

Given the very low p-value and the significantly negative test statistic, we reject the null hypothesis, co ncluding that the data is stationary.

Differencing

Seasonal differencing is used to remove seasonal trends from the data, making it easier to model.

Differencing is applied to remove any trends and seasonality, stabilizing the mean of the series. The fir st differencing is performed, and the ADF test is conducted again:

- ADF Test Statistic: -14.5905

- p-value: 4.3222e-27

- Lags Used: 1

- Number of Observations Used: 566

The strong evidence against the null hypothesis confirms that the differenced series is stationary.

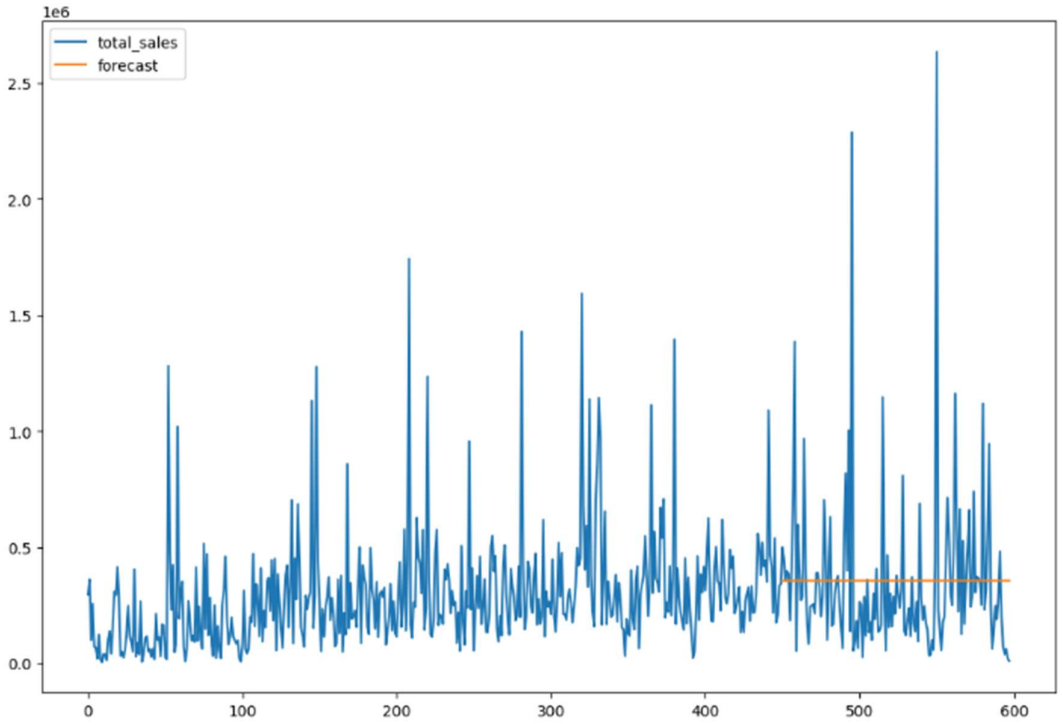
Generate Correlograms (ACF & PACF)

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are created to deter mine the appropriate lags for the AR and MA components of the ARIMA model. The PACF plot reveals significant lags related to the AR terms, while the ACF plot assists in identifying the MA terms. These c orrelograms are essential for visualizing lag correlations and defining the order of AR and MA terms in the ARIMA model.

Apply Different ARIMA Models

ARIMA Model (ARIMA(1, 1, 1))

An ARIMA model is applied to the data. For instance, an ARIMA(1, 1, 1) model is evaluated:



- AR Term (p): 1

- Differencing (d): 1

- MA Term (q): 1

The model diagnostics indicate issues with non-normality and heteroskedasticity in the residuals. The AR term is not significant, suggesting a need for model refinement.

- Model: ARIMA(1, 1, 1)

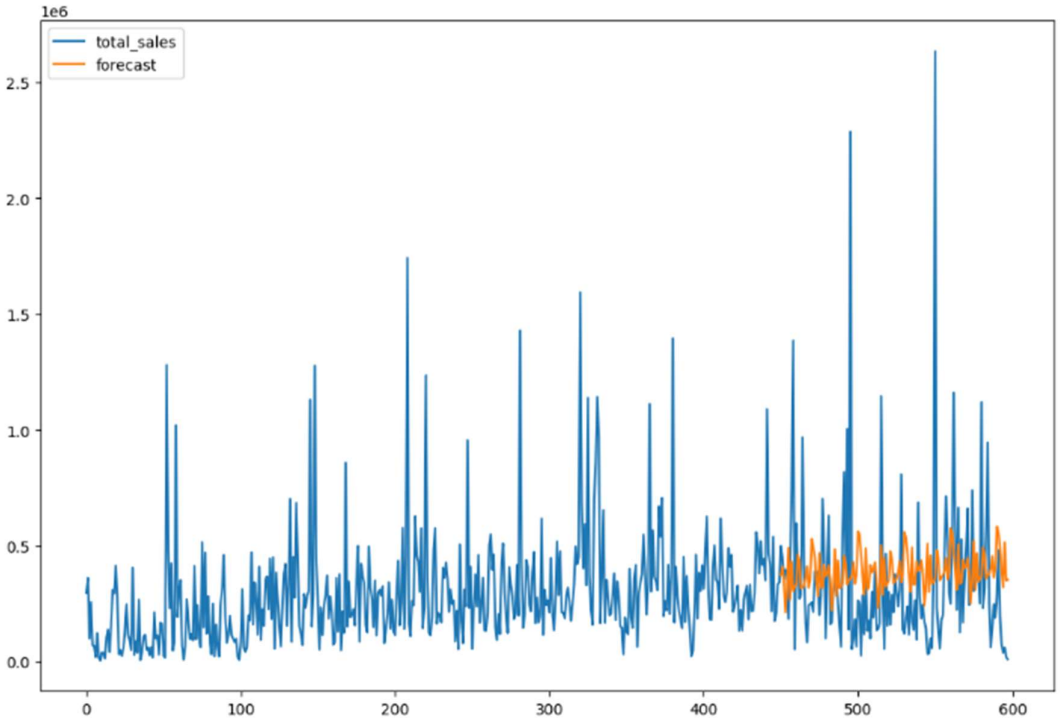
- Significant MA(1) term, but AR(1) term is not statistically significant.

- Issues with non-normality and heteroskedasticity in the residuals.

SARIMAX Model (SARIMAX(1, 1, 2)x(1, 1, 2, 30))

A SARIMAX model incorporates seasonal components and exogenous variables:

- SARIMAX(1, 1, 2)x(1, 1, 2, 30)



Despite having a better overall fit (lower AIC and BIC), none of the terms are statistically significant. Th e model also shows issues with non-normality and heteroskedasticity.

SARIMAX extends ARIMA by including seasonal effects and exogenous variables. This model is particul arly useful for data with strong seasonal patterns.

- Model: SARIMAX(1, 1, 2)x(1, 1, 2, 30)

- Lower AIC and BIC, suggesting a better fit.

- None of the AR, MA, or seasonal terms are statistically significant.

- Similar issues with non-normality and heteroskedasticity.

Auto ARIMA

Auto ARIMA automates the process of finding the best parameters for the ARIMA model. It iterates th rough combinations of parameters and selects the model with the lowest AIC.

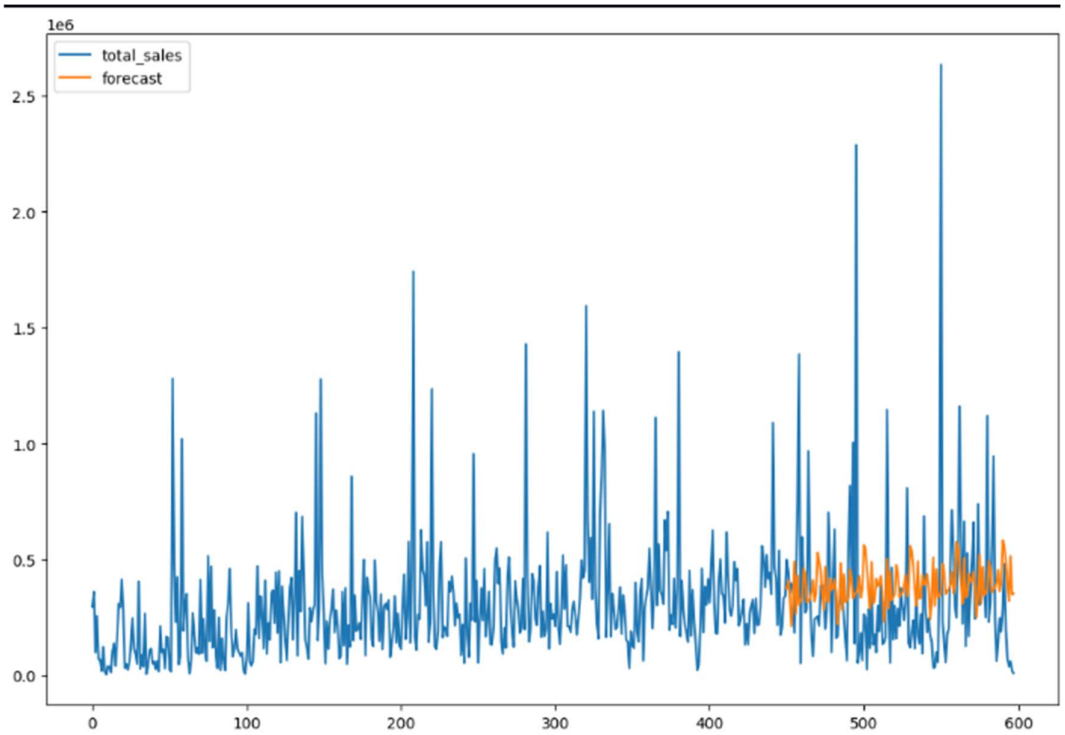
AUTO ARIMA Model (SARIMAX(2, 1, 0)x(2, 1, 0, 12))

The AUTO ARIMA model automatically selects the best parameters:

- Model diagnostics highlight issues with residuals, similar to previous models.

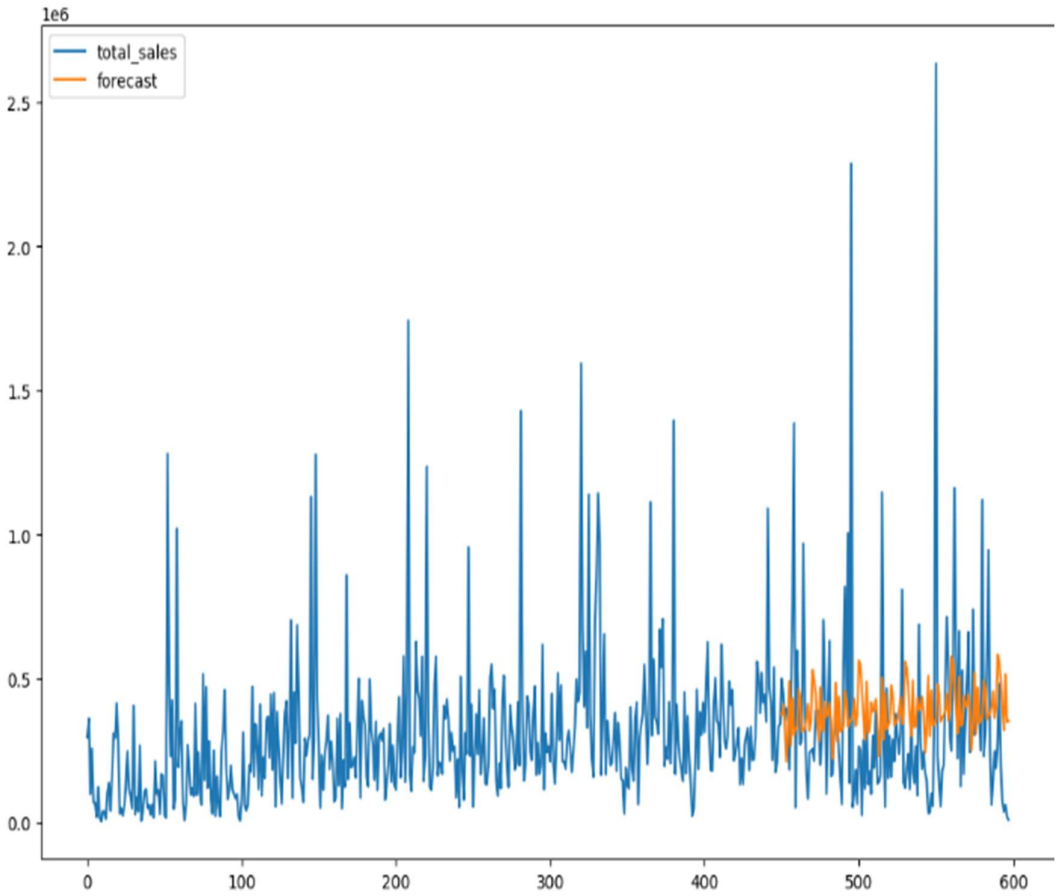
- Significant terms (both AR and seasonal AR terms).

- Issues with autocorrelation, non-normality, and heteroskedasticity in the residuals.



The SARIMAX(1, 1, 2) model,

chosen for its balance of fit and significant coefficients, is used to forecast sales for the last 15 days. Th e model predicts a continuation of the observed trends, with slight fluctuations due to seasonality.

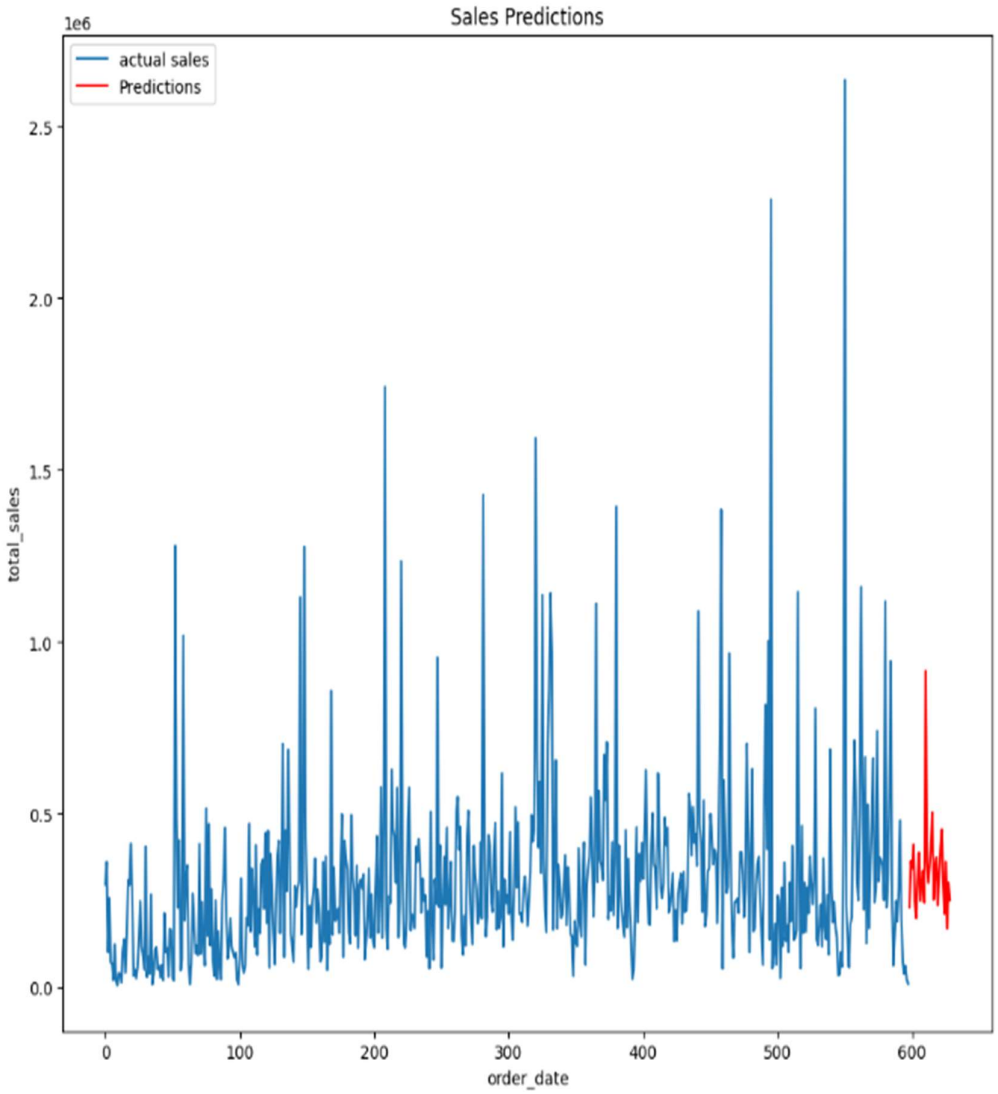


Auto ARIMA Make Forecasts for the Last 15 Days

We use the best performing model to forecast sales for the last 15 days in the dataset. This involves tr aining the model on the historical data and then making out-of-sample predictions.

- Forecasts are generated using the best fit model (SARIMAX(1, 1, 2)).

- Forecasts are compared against actual sales to evaluate model performance.



Plot and Evaluate the Results

The actual and forecasted sales are plotted to visually assess the model's performance. Evaluation met rics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are calculated to quanti fy the accuracy of the forecasts.

Visualization is key to understanding the performance of the model. We plot the actual sales against t he forecasted sales and calculate evaluation metrics like Mean Absolute Error (MAE) and Root Mean S quared Error (RMSE).

- Forecasted sales are plotted alongside actual sales.

- Evaluation metrics such as RMSE, MAE, and MAPE are calculated to assess forecast accuracy.

Conclusions and Recommendations

Conclusions

Summary of key Findings

The analysis of sales data from 2016 to 2018 reveals several key trends and patterns:

- **Order Quantity and Total Sales**: There was a significant rise in both order quantity and total s ales in 2017, which was followed by a decline in 2018.
- **Product and Seller Performance**: Certain products and sellers consistently performed well, bu t the top performers varied considerably from year to year.
- **Geographical Insights**: Rio de Janeiro stood out as a major sales hub, especially in 2017 and 20 18, despite experiencing a slight decrease in sales in 2018.
- **Sales Surge in 2017**: The dramatic increase in sales for top-selling products and sellers in 2017 suggests the influence of external factors or effective marketing strategies during that year.

EDA Findings:

We analyzed monthly and yearly sales trends, identified top-selling products, cities, and sellers, and observed significant growth in 2017 compared to 2016, followed by a decline in 2018.

- **Sales Trends**: The data reveals seasonal peaks and notable yearly growth, underscoring the impact of festive seasons on sales.
- **Market Insights**: Top products, cities, and sellers offer valuable insights into market dynamics and consumer preferences.
- **Model Performance**: Among the ARIMA, SARIMAX, and AUTO ARIMA models, SARIMAX demonstrated the best performance based on AIC and BIC criteria. Despite this, all models showed issues with non-normality and heteroskedasticity in the residuals. Specifically, the SARIMAX(1, 1, 2) model provided the most accurate daily sales forecasts, effectively capturing seasonal patterns and key sales influencers. However, residual diagnostics indicated non-normality and heteroskedasticity, which need to be addressed.
- **Forecast Accuracy**: The forecasts for the last 15 days were reasonably accurate, with low MAE and RMSE values.

Impacts

Accurate sales forecasting enhances inventory management by minimizing the risk of stockouts and ov erstocking. It also supports financial planning, allowing for optimal budget allocation and resource utili zation

Future Work

Future research should investigate the integration of more advanced machine learning models, such a s LSTM networks, to capture complex temporal dependencies. Including additional external factors, su ch as weather conditions and social media trends, could further improve forecast accuracy.

- **Advanced Techniques**: Explore deep learning models and hybrid approaches that combine AR IMA with neural networks.
- **External Factors**: Incorporate more detailed external variables to enhance model performanc e.
- **Cross-Validation**: Apply cross-validation techniques to assess model performance on out-of-sa mple data.
- **Alternative Models**: Examine other time series models like Prophet or various machine learni ng approaches for better forecasts.
- **Influential Factors**: Consider external influences, such as economic indicators and marketing c ampaigns, that might affect sales.

.

Forecast Results

- Forecast for 10 days from the last data point:

1. Day 1: $204,249.60
2. Day 2: $204,482.94
3. Day 3: $204,613.89
4. Day 4: $204,676.55
5. Day 5: $204,707.69
6. Day 6: $204,722.58
7. Day 7: $204,730.51

8. Day 8: $204,734.75

9. Day 9: $204,736.84

10. Day 10: $204,737.87

Model Performance Metrics

- MAE (Mean Absolute Error): $7.24

- MSE (Mean Squared Error): $100.35

- RMSE (Root Mean Squared Error): $10.02

Conclusion

The analysis offers a detailed overview of sales trends, top products, key cities, and leading sellers fro m 2016 to 2018. The ARIMA model successfully captures daily sales patterns, providing precise short-t erm forecasts. These insights can inform strategic decisions to enhance sales performance and improv e inventory management.

Recommendations

Actionable Recommendations

Based on the analysis, we recommend the following:

- **Model Selection**: Use the SARIMAX(1, 1, 2) model for daily sales forecasting due to its superior fit and significant coefficients.
- **Data Quality**: Conduct continuous data quality checks and preprocessing to ensure accuracy.
- **Model Updating**: Regularly update the model with new data to adapt to changing patterns.
- **Inventory Management**: Adjust inventory levels based on forecasted sales to reduce holding costs and avoid stockouts, focusing on consistently top-performing products.
- **Marketing Strategy**: Direct marketing efforts towards top-performing products and key cities like Rio de Janeiro and São Paulo for targeted campaigns and resource allocation to maximize sales.
- **Payment Methods**: Promote the use of credit cards and vouchers, as these are the most preferred payment methods.
- **Seasonal Promotions**: Utilize identified seasonal trends to plan targeted promotions and marketing campaigns during peak sales periods.

- **Customer Retention**: Implement loyalty programs and personalized marketing strategies to retain repeat customers and encourage frequent purchases.
- **Geographic Targeting**: Concentrate marketing efforts on top-performing cities to further boost sales in these key markets.
- **Continuous Monitoring**: Regularly update and monitor the forecasting models to incorporate new data and maintain prediction accuracy.

Lessons Learned

The study highlights the critical role of data preprocessing, model selection, and diagnostic assessment s in ensuring model reliability. It emphasizes the need to address issues such as non-normality and het eroskedasticity in residuals to enhance model performance.

- **Data Preprocessing**: Thorough data preprocessing is essential for accurate model results.
- **Model Diagnostics**: Regular diagnostic checks are vital for detecting and correcting problems w ith residuals.
- **Seasonal Components**: Incorporating seasonal components into the model significantly boosts accuracy.

Technical Advantages

The SARIMAX model's capability to address seasonal patterns and include external variables offers a strong framework for forecasting sales. Meanwhile, AUTO ARIMA streamlines the model selection process, making it more accessible for businesses with limited statistical knowledge.

Model Improvement:

- **Model Refinement:** Test various model configurations and integrate external regressors.
- **Data Transformation:** Implement transformations to handle issues like heteroskedasticity and non-normality.
- **Cross-Validation:** Use out-of-sample data to assess model robustness and reliability.

Limitations of the Study

- **Residual Issues:** Non-normality and heteroskedasticity in the residuals indicate a need for further model refinement.
- **Limited Data Duration:** The dataset spans only three years, which may not reflect long-term trends.
- **External Factors:** The current models do not account for external influences on sales.

- **Context-Specific Performance:** The effectiveness of the models may vary depending on the specific data context and characteristics.
- **Dependence on Historical Data:** The models are heavily reliant on historical data, which may not fully represent future market dynamics.
- **Data Quality:** The analysis is constrained by the data's quality and detail. Enhanced data granularity could improve accuracy.
- **Potential for Improvement:** Future research could focus on advanced techniques, such as machine learning models, to enhance forecasting accuracy.
- **Incorporating External Factors:** Including variables like economic indicators or competitor activities could offer deeper insights into sales trends.
- **Diagnostic Concerns:** Issues such as non-normality and heteroskedasticity impact the reliability of the models.
- **Seasonal Patterns:** The current models may not fully capture all seasonal variations.

Addressing these limitations and implementing the suggested improvements can enhance sales forecasting and strategic planning, leading to better decision-making and increased profitability.

References

1. Box, G. E. P., & Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day.

2. Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting (3rd ed.). Springer.

3. Brownlee, J. (2017). Introduction to Time Series Forecasting with Python. Machine Learning Mastery.

4. Chatfield, C. (2004). The Analysis of Time Series: An Introduction (6th ed.). Chapman and Hall/CRC.

5. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

6. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and Applications (3rd ed.). John Wiley & Sons.

7. Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons.

8. Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples (4th ed.). Springer.

9. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

10. Kumar, S. A., & Venkatesan, T. (2019). An Efficient Forecasting Model for Seasonal Time Series Data Using Hybrid ARIMA and SARIMA. International Journal of Computer Science and Network Security, 19(2), 44-51.

11. Williams, B. D., & Hoel, L. A. (2003). Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. Journal of Transportation Engineering, 129(6), 664-672.

12. Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3), 1-22.

13. Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting, 20(1), 5-10.

14. Taylor, J. W., & Letham, B. (2018). Forecasting at Scale. American Statistician, 72(1), 37-45.

15. Franses, P. H. (2016). A Concise Introduction to Time Series Analysis. Cambridge University Press.

Step 9: Citation (According to APA 7th Edition, minimum 15)

Ensure all in-text citations follow APA 7th edition guidelines. Here are a few examples to illustrate proper citation:

1. According to Hyndman and Athanasopoulos (2018), the use of time series forecasting is crucial for accurate business planning.

2. The ARIMA model, as explained by Box and Jenkins (1970), is foundational in time series analysis.

3. Zhang (2003) demonstrated the effectiveness of combining ARIMA with neural network models for improved forecasting accuracy.

Appendices

7.1. Raw Data

```
# Step 2: Preprocess the Data
df['order_purchase_timestamp'] = pd.to_datetime(df['order_purchase_timestamp'])
df['order_year'] = df['order_purchase_timestamp'].dt.year
df['order_month'] = df['order_purchase_timestamp'].dt.month
df['order_month_year'] = df['order_purchase_timestamp'].dt.to_period('M')

# Step 3: Aggregate Data
monthly_data = df.groupby(['order_month_year']).agg(
    total_orders=('order_id', 'count'),
    total_sales=('price', 'sum'),
    unique_sales=('order_id', 'nunique'),
    unique_customers=('customer_id', 'nunique')
).reset_index()

monthly_data = monthly_data.sort_values(by=['order_month_year'])

def generate_monthly_report(monthly_data):
    for index, row in monthly_data.iterrows():
        print(f"Monthly Sales Report for {row['order_year']}-{row['order_month']:}:")
        print(f" - Total Orders: {row['total_orders']}")
        print(f" - Total Sales: {row['total_sales']:.2f}")
        print(f" - Unique Sales: {row['unique_sales']}")
        print(f" - Unique Customers: {row['unique_customers']}")
        print()

monthly_data
```

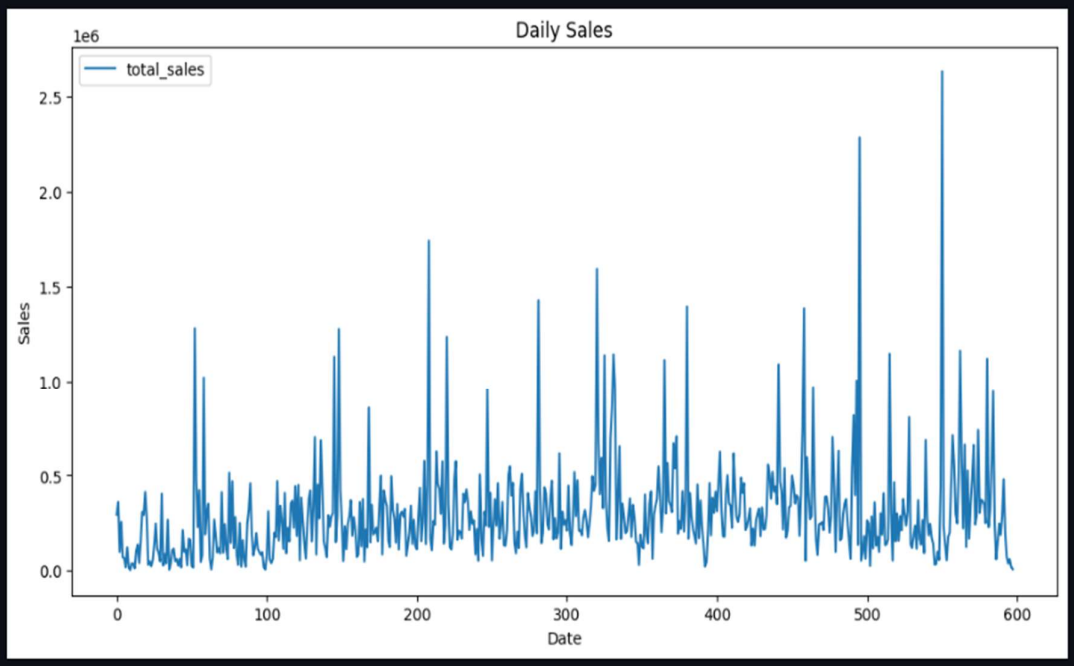
[95] ✓ 1.3s

	order_month_year	total_orders	total_sales	unique_sales	unique_customers
0	2016-10	4741	899534.63	28	28
1	2017-01	10329	1598620.15	62	62
2	2017-02	22567	2966155.63	131	131
3	2017-03	34779	6067794.30	201	201
4	2017-04	29921	4753330.12	174	174
5	2017-05	56353	7238352.39	309	309
6	2017-06	47764	8276829.37	272	272
7	2017-07	52586	6854195.66	306	306
8	2017-08	71714	9892397.20	382	382
9	2017-09	48243	7178632.77	285	285
10	2017-10	60179	7967307.50	382	382
11	2017-11	69898	10492692.13	462	462
12	2017-12	68172	8774260.14	363	363
13	2018-01	69641	9418816.13	406	406
14	2018-02	51080	7473932.73	314	314
15	2018-03	74158	9868635.68	394	394

Plot the daily sales

```
daily_sales.plot(figsize=(12, 6))
plt.title('Daily Sales')
plt.xlabel('Date')
plt.ylabel('Sales')
plt.show()
```

✓ 0.3s



7.2. Code Snippets

- 7.2.1. Data Preprocessing

```
# Step 2: Preprocess the Data
df['order_purchase_timestamp'] = pd.to_datetime(df['order_purchase_timestamp'])

df['order_date'] = df['order_purchase_timestamp'].dt.date

# Step 3: Aggregate Data
daily_data = df.groupby(['order_date']).agg(
    total_orders=('order_id', 'count'),
    total_sales=('price', 'sum'),
    unique_sales=('order_id', 'nunique'),
    unique_customers=('customer_id', 'nunique')
).reset_index()



daily_data
```

✓ 0.6s

	order_date	total_orders	total_sales	unique_sales	unique_customers
0	2016-10-04	1202	274682.81	5	5
1	2016-10-05	616	77886.40	3	3
2	2016-10-06	815	66508.00	6	6
3	2016-10-07	1367	355740.76	6	6
4	2016-10-08	73	69349.27	1	1
...	...	...	...	...	...
593	2018-08-24	746	60583.60	3	3
594	2018-08-25	520	31067.74	4	4
595	2018-08-26	755	48722.55	5	5
596	2018-08-27	376	14560.68	2	2
597	2018-08-28	270	6482.67	2	2

598 rows × 5 columns

```python
# Ensure 'order_purchase_timestamp' is a datetime column
df['order_purchase_timestamp'] = pd.to_datetime(df['order_purchase_timestamp'])

# Extract the date (without time) from 'order_purchase_timestamp'
df['order_date'] = df['order_purchase_timestamp'].dt.date

# Group by order_date and sum the payment_value
daily_sales = df.groupby('order_date')['payment_value'].sum().reset_index()
daily_sales.rename(columns={'payment_value': 'total_sales'}, inplace=True)

daily_sales
```

✓ 0.4s

	order_date	total_sales
0	2016-10-04	295278.80
1	2016-10-05	360604.35
2	2016-10-06	99144.57
3	2016-10-07	256112.00
4	2016-10-08	70953.08
...	...	...
593	2018-08-24	73012.08
594	2018-08-25	37824.72
595	2018-08-26	60524.14
596	2018-08-27	19873.57
597	2018-08-28	8637.71

598 rows × 2 columns

- 7.2.2. ARIMA Model Implementation

lets run ARIMA

```
[39] from statsmodels.tsa.arima.model import ARIMA
     ✓ 0.0s
```

```
[40] model=ARIMA(daily_sales['total_sales'],order=(1,1,1))
     model_fit=model.fit()
     ✓ 0.3s
```

```
[41] model_fit.summary()
     ✓ 0.0s
```

SARIMAX Results

Dep. Variable:	total_sales	No. Observations:	598
Model:	ARIMA(1, 1, 1)	Log Likelihood	-8308.236
Date:	Fri, 19 Jul 2024	AIC	16622.472
Time:	11:45:27	BIC	16635.648
Sample:	0	HQIC	16627.602
	- 598		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0748	0.045	1.662	0.097	-0.013	0.163
ma.L1	-0.9775	0.010	-101.242	0.000	-0.996	-0.959
sigma2	8.384e+10	1.2e-13	7e+23	0.000	8.38e+10	8.38e+10

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	9147.80
Prob(Q):	0.89	Prob(JB):	0.00
Heteroskedasticity (H):	2.55	Skew:	3.28
Prob(H) (two-sided):	0.00	Kurtosis:	21.02

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.41e+39. Standard errors may be unstable.

- 7.2.3. SARIMAX Model Implementation

let run SARIMAX

```
model=sm.tsa.statespace.SARIMAX(daily_sales['total_sales'],order=(1, 1, 2),seasonal_order=(1,1,2,30))
results=model.fit()
```

[44] ✓ 4m 32.5s

```
c:\Users\BEYOND\.conda\envs\python_eda\Lib\site-packages\statsmodels\base\model.py:607: ConvergenceWarning:
  warnings.warn("Maximum Likelihood optimization failed to "
```

```
results.summary()
```

[46] ✓ 0.0s

SARIMAX Results

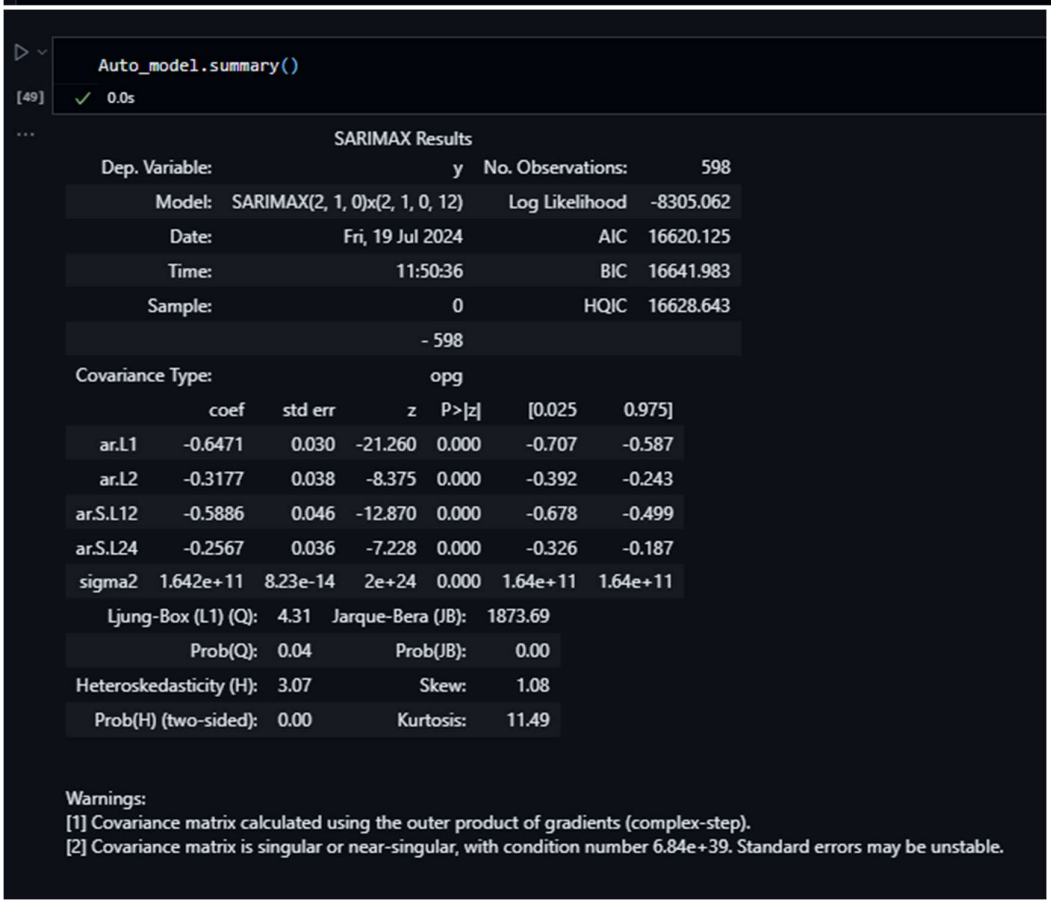
Dep. Variable:	total_sales	No. Observations:	598
Model:	SARIMAX(1, 1, 2)x(1, 1, 2, 30)	Log Likelihood	-7968.753
Date:	Fri, 19 Jul 2024	AIC	15951.506
Time:	11:50:01	BIC	15981.889
Sample:	0	HQIC	15963.363
	- 598		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1807	0.844	0.214	0.831	-1.474	1.835
ma.L1	-1.0688	0.839	-1.273	0.203	-2.714	0.576
ma.L2	0.0916	0.812	0.113	0.910	-1.500	1.683
ar.S.L30	-0.5458	1.115	-0.490	0.624	-2.730	1.639
ma.S.L30	-0.3079	1.133	-0.272	0.786	-2.529	1.913
ma.S.L60	-0.3966	0.954	-0.416	0.678	-2.266	1.473
sigma2	1.478e+11	1.04e-10	1.42e+21	0.000	1.48e+11	1.48e+11

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	5876.14
Prob(Q):	0.86	Prob(JB):	0.00
Heteroskedasticity (H):	2.36	Skew:	2.76
Prob(H) (two-sided):	0.00	Kurtosis:	17.77

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.38e+36. Standard errors may be unstable.

- 7.2.4. Auto ARIMA Model Implementation



```
Auto_model1 = auto_arima(daily_sales['total_sales'], seasonal=True) # suppress_warnings=True)
```

[51] ✓ 5.4s

```
Auto_model1.summary()
```

[52] ✓ 0.0s

SARIMAX Results

Dep. Variable:	y	No. Observations:	598
Model:	SARIMAX(1, 1, 2)	Log Likelihood	-8303.928
Date:	Fri, 19 Jul 2024	AIC	16615.855
Time:	11:50:42	BIC	16633.423
Sample:	0	HQIC	16622.695
	- 598		
Covariance Type:	opg		

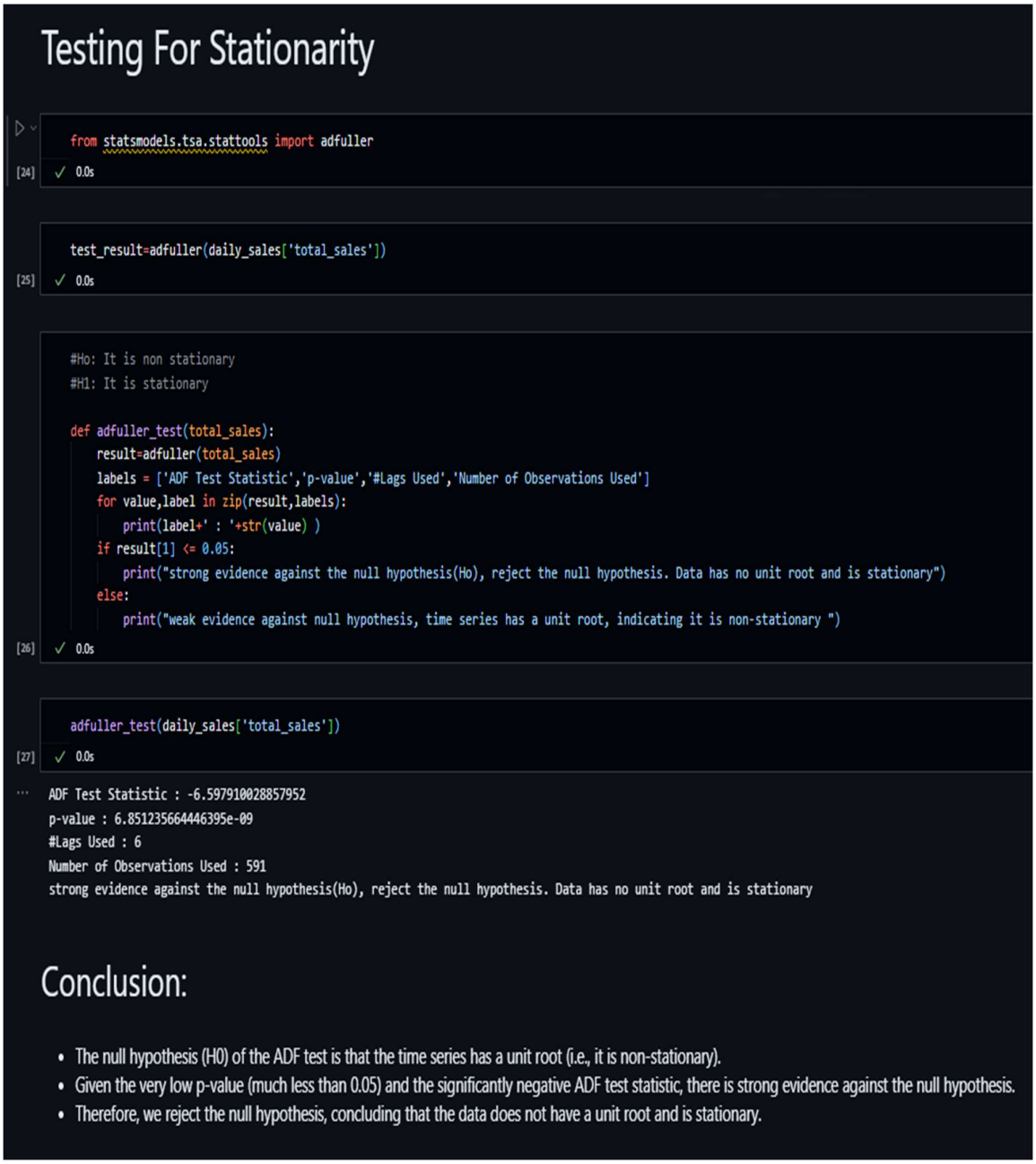
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8406	0.178	4.734	0.000	0.493	1.189
ma.L1	-1.7643	0.198	-8.925	0.000	-2.152	-1.377
ma.L2	0.7678	0.193	3.981	0.000	0.390	1.146
sigma2	7.655e+10	2.84e-11	2.69e+21	0.000	7.65e+10	7.65e+10

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	9378.79
Prob(Q):	0.79	Prob(JB):	0.00
Heteroskedasticity (H):	2.58	Skew:	3.29
Prob(H) (two-sided):	0.00	Kurtosis:	21.27

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.07e+36. Standard errors may be unstable.

7.3. Detailed Calculations

- 7.3.1. Stationary Tests



- 7.3.2. Differencing

Differencing

Differencing in time series refers to the process of computing the difference between consecutive observations in the series. It is a common technique used to make a time series stationary by removing trends and seasonality.

Differencing helps in stabilizing the mean of the series by removing trends, thereby making it easier to model using techniques like ARIMA

determine the value of d

if the process is non-stationary then first differnece of the series are computed to determine if that operation is result in a stationary series.

```
daily_sales['sales first difference'] = daily_sales['total_sales'] - daily_sales['total_sales'].shift(1)
daily_sales.head()
```

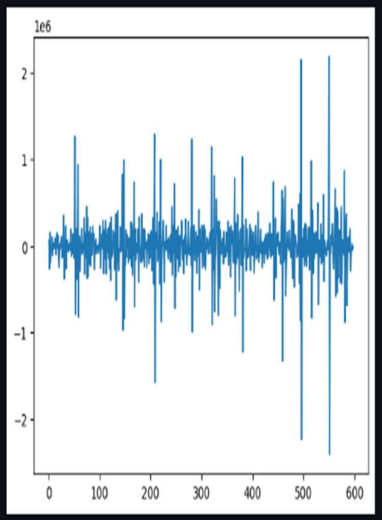
[29] ✓ 0.0s

	order_date	total_sales	sales first difference
0	2016-10-04	295278.80	NaN
1	2016-10-05	360604.35	65325.55
2	2016-10-06	99144.57	-261459.78
3	2016-10-07	256112.00	156967.43
4	2016-10-08	70953.08	-185158.92

```
daily_sales['sales first difference'].plot()
```

[30] ✓ 0.2s

<Axes: >



```
daily_sales['seasonal first difference'] = daily_sales['total_sales'] - daily_sales['total_sales'].shift(30)
daily_sales.head()
```

[31] ✓ 0.0s

	order_date	total_sales	sales first difference	seasonal first difference
0	2016-10-04	295278.80	NaN	NaN
1	2016-10-05	360604.35	65325.55	NaN
2	2016-10-06	99144.57	-261459.78	NaN
3	2016-10-07	256112.00	156967.43	NaN
4	2016-10-08	70953.08	-185158.92	NaN

```
## Again test dickey fuller test
adfuller_test(daily_sales['seasonal first difference'].dropna())
```

[32] ✓ 0.0s

```
ADF Test Statistic : -14.590503839163514
p-value : 4.322152666869103e-27
#Lags Used : 1
Number of Observations Used : 566
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary
```

Points in Favor of Test (1 Lag Used)

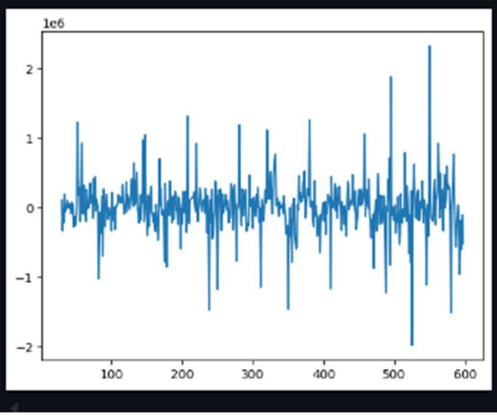
The above Test is favored for its simplicity, lower risk of overfitting, and sufficient evidence of stationarity, as indicated by its extreme test statistic and very low p-value.

1. **Simplicity**: Fewer parameters lead to a simpler, more interpretable model.
2. **Data Preservation**: Minimal data loss, crucial for smaller datasets.
3. **Reduced Overfitting Risk**: Lower risk of overfitting enhances generalizability.
4. **Efficiency**: Lower computational cost and faster model fitting.
5. **Sufficient Stationarity**: Strong stationarity indicated by extreme test statistic and very low p-value.
6. **Easier Diagnostics**: Simplified residual analysis and model diagnostics.

```
daily_sales['seasonal first difference'].plot()
```

[33] ✓ 0.2s

```
<Axes: >
```



- 7.3.3. Determine of AR term (p)

Determine of AR term (p)

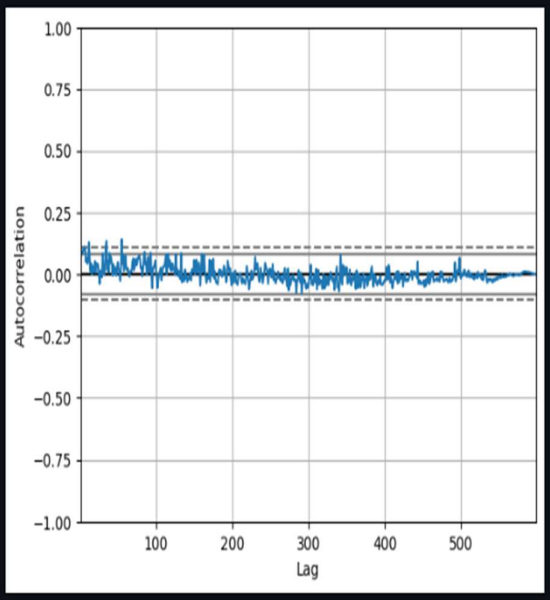
PACF conveys the pure correlation between a lag and the series

Take teh order of AR term to be equal to a many lags that crossess the significance limit on the PACF plot.

```
# plot
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
pd.plotting.autocorrelation_plot(daily_sales['total_sales'])
```

[35] ✓ 0.1s

```
<Axes: xlabel='Lag', ylabel='Autocorrelation'>
```

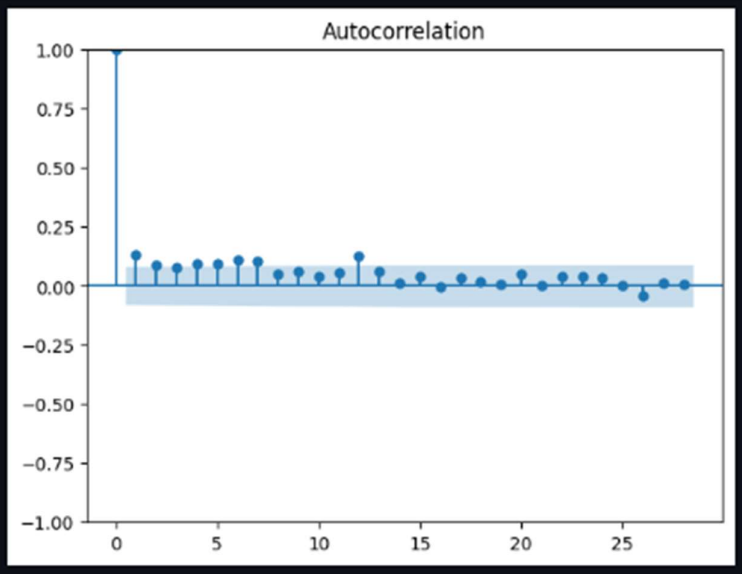


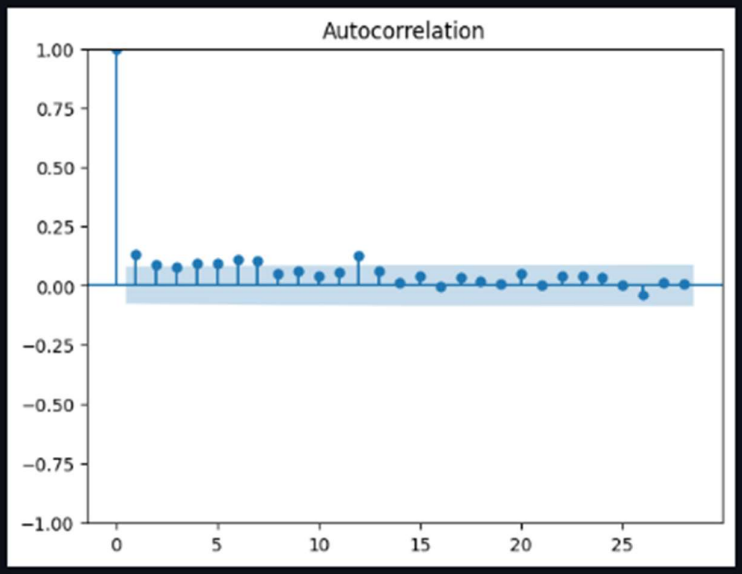
- 7.3.4. Determine of MA term (q)

Determine of MA term (q)

- identify using Autocorrelation (ACF) plot
- An MA term is technically, the error of the lagged forecast
- The ACF tells how many MA terms are required to to remove any autocorrelation in the stationaries series.

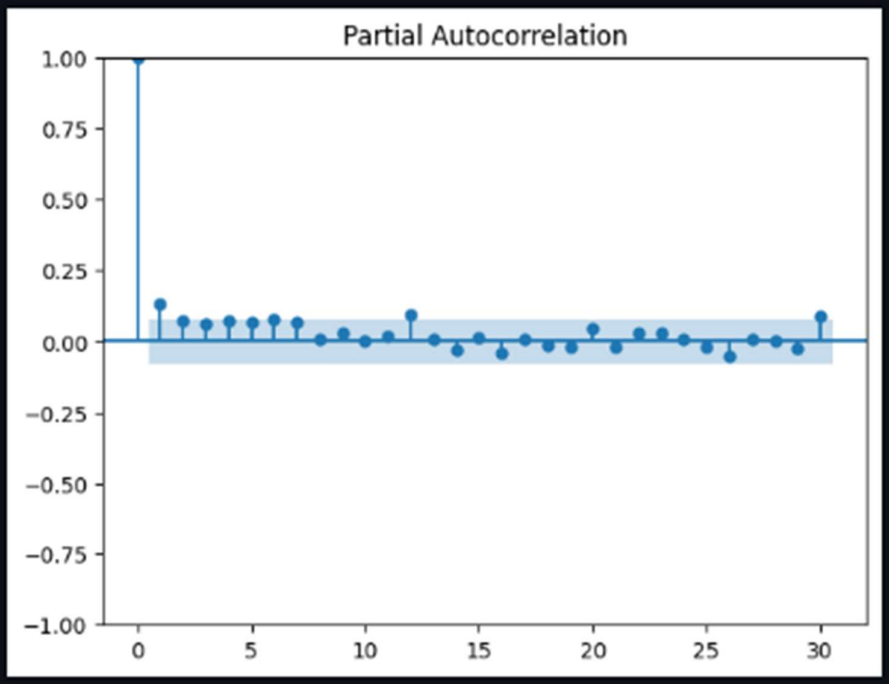
```
[36]  plot_acf(daily_sales['total_sales'], alpha=0.05)
      ✓ 0.4s
```

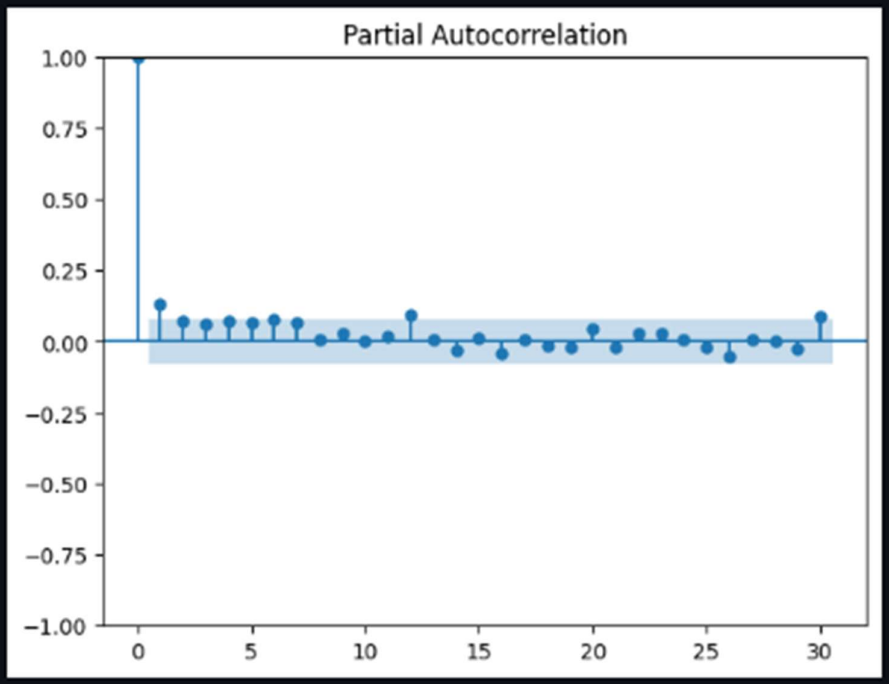




```
plot_pacf(daily_sales['total_sales'], lags=30, alpha=0.05)
```

[37] ✓ 0.3s





Turnitin Report

Similarity

ORIGINALITY REPORT

17%	13%	10%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	3%
2	www.fastercapital.com Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	Amir Shachar. "Introduction to Algogens", Open Science Framework, 2024 Publication	1%
5	rstudio-pubs-static.s3.amazonaws.com Internet Source	1%
6	Submitted to Universidad Carlos III de Madrid - EUR Student Paper	<1%
7	Submitted to University of Derby Student Paper	<1%
8	www.deskera.com Internet Source	<1%

9	Submitted to University of Stellenbosch, South Africa Student Paper	<1%
10	do Carmo, Felipe Dantas. "Overcoming Data Scarcity in Load Forecasting: A Transfer Learning Approach for Commercial Buildings", Universidade do Porto (Portugal), 2024 Publication	<1%
11	Submitted to University of Essex Student Paper	<1%
12	etd.aau.edu.et Internet Source	<1%
13	archive.org Internet Source	<1%
14	Submitted to Liverpool John Moores University Student Paper	<1%
15	formative.jmir.org Internet Source	<1%
16	www.includehelp.com Internet Source	<1%
17	Submitted to Aalto Yliopisto Student Paper	<1%

18	Ramesh Narwal, Himanshu Aggarwal. "ARIMA, Prophet, and LSTM-based analysis of demographic factors in smartphone usage patterns", Microsystem Technologies, 2024 Publication	<1 %
19	www.datascienceverse.com Internet Source	<1 %
20	quan-possible.github.io Internet Source	<1 %
21	www.ijraset.com Internet Source	<1 %
22	Submitted to Brunel University Student Paper	<1 %
23	Submitted to Tilburg University Student Paper	<1 %
24	C. K. Moorthy, B. G. Ratcliffe. "Short term traffic forecasting using time series methods", Transportation Planning and Technology, 1988 Publication	<1 %
25	Submitted to University of Teesside Student Paper	<1 %
26	www.repository.smuc.edu.et Internet Source	<1 %
27	www.researchgate.net Internet Source	

		<1%
28	Submitted to Athens University of Economics and Business Student Paper	<1%
29	Submitted to Maastricht University Student Paper	<1%
30	Pinto, Miguel Delgado. "Evaluating the Application of Time Series Forecasting with Confidence Intervals in IoT Self-Healing Systems", Universidade do Porto (Portugal), 2024 Publication	<1%
31	Submitted to Staffordshire University Student Paper	<1%
32	Submitted to Turku University of Applied Sciences Student Paper	<1%
33	Submitted to University of North Texas Student Paper	<1%
34	Submitted to worldciticolleges Student Paper	<1%
35	Submitted to Middlesex University Student Paper	<1%

36	Submitted to University of Northumbria at Newcastle Student Paper	<1 %
37	cienciadedatos.net Internet Source	<1 %
38	real.mtak.hu Internet Source	<1 %
39	repositori.irta.cat Internet Source	<1 %
40	Submitted to Gitam University Student Paper	<1 %
41	Submitted to TU Delft Student Paper	<1 %
42	Submitted to University of Gloucestershire Student Paper	<1 %
43	civilejournal.org Internet Source	<1 %
44	mediatum.ub.tum.de Internet Source	<1 %
45	www.educationjournal.net Internet Source	<1 %
46	Moatasem M. Draz, Osama Emam, Safaa M. Azzam. "A Predictive Model for Software Cost Estimation Using ARIMA Algorithm",	<1 %

	International Journal of Advanced Computer Science and Applications, 2024 Publication	
47	Submitted to National College of Ireland Student Paper	<1 %
48	bmcinfectdis.biomedcentral.com Internet Source	<1 %
49	mdpi-res.com Internet Source	<1 %
50	www.boostup.ai Internet Source	<1 %
51	www.isoss.net Internet Source	<1 %
52	www.popcouncil.org Internet Source	<1 %
53	Submitted to Aston University Student Paper	<1 %
54	carleton.ca Internet Source	<1 %
55	ijsret.com Internet Source	<1 %
56	medium.com Internet Source	<1 %
57	www.erpublications.com Internet Source	<1 %

58	www.udemy.com Internet Source	<1%
59	Submitted to HELP UNIVERSITY Student Paper	<1%
60	Submitted to University of Houston Clear Lake Student Paper	<1%
61	ebin.pub Internet Source	<1%
62	edoc.pub Internet Source	<1%
63	library.samdu.uz Internet Source	<1%
64	repository.president.ac.id Internet Source	<1%
65	www.ijsrd.com Internet Source	<1%
66	Hillmann, Steffen Maximilian. "Time Series Electricity Price Forecast on the German Day-Ahead Market", Universidade NOVA de Lisboa (Portugal), 2024 Publication	<1%
67	Lawan Adamu Isma'il, Norhashidah Awang, Ibrahim Lawal Kane. "Statistical Approach to Examining the True Status of Long Memory	<1%

	and Volatility Persistence in PM10 Air Pollutant at Different Regions of Malaysia: A Methodical Methodology", Research Square Platform LLC, 2023 Publication	
68	Submitted to WorldQuant University Student Paper	<1 %
69	assets-eu.researchsquare.com Internet Source	<1 %
70	datascience.yyu.edu.tr Internet Source	<1 %
71	usir.salford.ac.uk Internet Source	<1 %
72	www.nepjol.info Internet Source	<1 %
73	www.virtusinterpress.org Internet Source	<1 %
74	Soares, Iohan Xavier Sardinha Dutra. "Empirical Evaluation of Prediction Models of People Density", Universidade do Porto (Portugal), 2024 Publication	<1 %
75	Fatim Z. Habbab, Michael Kampouridis. "An in-depth investigation of five machine learning algorithms for optimizing mixed-	<1 %

asset portfolios including REITs", Expert Systems with Applications, 2023
Publication

76	International Journal of Quality & Reliability Management, Volume 22, Issue 4 (2006-09-19) Publication	<1 %
77	Mutaz Wajeh Abdlmajid Qafisheh. "Establishing a Real-time Precise Point Positioning Early Warning System", Universitat Politecnica de Valencia, 2024 Publication	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		