Sales Prediction Using Historical Sales Transaction Data

MAT 3991- GROUP - 09

Department of Mathematics

Faculty of Science

University of Peradeniya

S/19/803 – T.Ajith
S/19/813 – E.Vanathey
S/19/830 – M.Thayani
S/19/853 – N.Tenoyan
S/19/855 – V.Thinesh
S/19/865 – J.Anistan

1.Acknowledgements

In performing our project, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. We would like to show our gratitude to Dr. Niluka Rodrigo and Dr. Mahasen Dehideniya the Course Instructors in the Mathematical department for giving us good guidelines for the project throughout numerous consultations. We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us to complete our project. We would like to offer acknowledgment for VT Manufacturing Private Limited giving us this opportunity to do our project for their company. We would like to show our gratitude to the rest of the people in VT Manufacturing Private Limited for helping us to accomplish our jobs. Many people, especially our mates and team members themselves, have made valuable comments and suggestions on this proposal which gave us the inspiration to work on the project. We thank all the people for their help directly and indirectly to complete our project.

Table of Content

2.Summary

- **Project Objective**

The project aims to forecast weekly sales for various apparel products using historical transaction data. By accurately predicting future sales, the project seeks to aid in inventory management, optimize stock levels, and improve overall business decision-making.

- **Data Used**

The project uses historical weekly sales data, sourced from the "Apparel Sales Transactions Dataset Weekly.csv" file. Each product's sales data spans 51 weeks, with weekly data entries for each product.

- **Methodology**

The dataset was pre-processed to exclude non-numeric columns, focusing on time-series data for weekly sales. The SARIMA model was selected. It was trained on the first 42 weeks of data, reserving the final 10 weeks for testing, and forecasts were generated for weeks 42–51 (test) and weeks 52–61 (extended period). Model performance was assessed using MAE, MSE, and RMSE metrics to evaluate accuracy for each product.

3.Introduction

3.1. Overview of the company

The Star Garment Group was founded in 1978, Star is one of the key players in the Sri Lankan apparel industry. They employ a team of over 8,500 dedicated associates whose core competencies are in maintaining the highest standards, ensuring timely deliveries and providing individualized customer service to a variety of premiere global brands.

Among the 11 manufacturing factories of star garment group, V T manufacturing (pvt) ltd is one of the manufacturing factory of them, it is a world class apparel sourcing, design and manufacturing company. Which provide expert seamless execution in product development, technical innovation, on time delivery and personalized customer service to ensure the best quality and value for each of the customer.

3.2. The Importance of Sales Forecasting in Business

Sales forecasting is crucial for businesses as it provides a roadmap for future planning and decision-making. By predicting future sales, companies can better align their resources, manage inventory efficiently, and meet customer demand without overproducing or understocking. Here’s how sales forecasting impacts key areas of business,

- Inventory Management: Accurate sales forecasts help maintain the right amount of stock. With insights into anticipated demand, businesses can avoid excess inventory, which ties up capital and increases storage costs. At the same time, it prevents stock outs, ensuring products are available when customers need them.
- Demand Planning: Forecasting enables companies to predict market demand and adjust production schedules and purchasing needs. This minimizes waste, optimizes supply chain efficiency, and improves responsiveness to changing market conditions.

- Resource Allocation: Sales forecasts guide budget planning, staffing, and resource allocation across departments. Companies can adjust their marketing, production, and logistics strategies to meet forecasted sales levels, improving operational efficiency and profitability.

Effective sales forecasting supports sustainable growth by helping businesses make informed, data-driven decisions that align with future demand, boosting customer satisfaction and maximizing profitability.

3.3. Scope of Analysis

This analysis focuses on forecasting weekly sales data for various clothing products, providing insights into short-term demand patterns. By analyzing weekly trends, we can capture fluctuations and seasonality in sales, which are crucial for inventory and demand planning on a frequent basis.

For the modeling technique, we selected the SARIMA (Seasonal Autoregressive Integrated Moving Average) model. SARIMA is well-suited for handling seasonal data patterns, making it ideal for weekly sales predictions where demand may vary with time. This model can capture both the trend and seasonal components in the data, allowing us to produce more accurate forecasts for the next 10 weeks.

This approach helps in producing precise, week-by-week sales predictions, which businesses can use for immediate planning and resource allocation.

3.4. Objectives of the Study

The primary objectives of this study are:

1.Improving Forecasting Accuracy: Develop a reliable model to predict weekly sales, aiming for high accuracy to support business planning and reduce forecasting errors.

2.Understanding Sales Patterns: Analyze historical sales data to uncover patterns, which can help businesses better anticipate demand.

3.Identifying Seasonal Trends: Detect any recurring seasonal trends in sales, such as higher demand during certain weeks or seasons. Recognizing these patterns helps businesses adjust their inventory and resources to meet seasonal peaks.

These goals are focused on enhancing the company’s ability to make data-driven decisions, optimize inventory levels, and improve overall efficiency in meeting customer demand.

4.Data source

Collect the data source from VT Manufacturing Company in the (.csv) file format. This format allows for easier access to the dataset and provides more information for better understanding.

Pre processing

Data preprocessing is a crucial step in any data-driven project as it involves transforming raw data into a clean, structured format suitable for analysis and modelling. For our sales prediction project, applied several preprocessing steps to ensure the dataset was ready for building accurate models.

First, loaded the dataset into Jupiter Notebook in CSV format. CSV files are preferred in Python because the format is lightweight, easy to manipulate, and natively supported by Python libraries. Using CSV files instead of Excel files can also help reduce the file size, making the data easier to work with.

4.1. Steps of preprocessing

1. Address missing values

Missing data can result in biased or inaccurate model predictions. During the data preprocessing phase, the dataset was carefully examined, and it was determined that there were no missing values. This conclusion ensured that the dataset was complete and ready for further analysis, without the need for any imputation or handling of missing data.

2. Manage outliers

Outliers can distort data analysis and model performance. To detect outliers in the sales data, utilized box plots, which are simple yet effective tools for identifying extreme values. After identifying the outliers, we addressed them using the Interquartile Range (IQR) method, an efficient technique for managing outliers and ensuring that the dataset remained within appropriate bounds for accurate analysis.

3.Check data inconsistency

To ensure the accuracy of the analysis, duplicate records were identified and removed from the dataset, ensuring that each transaction was represented only once. This step was critical to avoid the risk of double-counting sales, which could distort the results. By addressing and resolving these issues, the quality and reliability of the dataset were significantly improved, leading to more accurate and trustworthy analysis.

4.2. EDA

Exploratory Data Analysis (EDA) is a critical step in the data analysis process as it provides a deep understanding of the data, which is essential for making informed decisions in the later stages of analysis and modeling.

Through EDA, gained several key insights, one of which is the overall sales trend. This trend refers to the pattern of sales over a specific period, and used a line plot of aggregate sales data to visualize trends over time. Positive growth rates indicate periods of increasing sales, while negative growth rates suggest declining sales.

Upon examining the trends, observed the following:

- **Early Summer (Weeks 15–25):** There was a peak in sales, likely due to the arrival of warmer weather.
- **Summer (Weeks 25–35):** Sales remained steady, as the demand for summer products continued to be high but stable.
- **Winter (Weeks 35–50)**: During the EDA process, we analyzed the sales trend and observed a decline in sales. This reduction in sales is likely attributed to the colder season, which typically leads to a decrease in demand for summer apparel.

Correlation assessment is a key component of EDA, as it helps identify relationships between different variables in the dataset. For example, we examined how sales figures correlate with factors such as holidays and weather conditions. Positive correlations suggest a direct relationship between variables, while negative correlations indicate an inverse relationship.

Exploring these correlations allows us to generate hypotheses about the potential causes of sales fluctuations. For instance, understanding how certain external factors like weather or holidays influence sales can provide valuable insights for strategic planning.

In conclusion, correlation assessment offers valuable insights into the relationships and dependencies between variables in the sales transaction dataset. This understanding guides further analysis and supports informed decision-making throughout the project.

4.3. Managing Correlation Assessment

Effective management of correlation assessment involves several key steps:

1. **Clean the Data**
   It is crucial to ensure that the dataset is free from errors, missing values, and inconsistencies. This step enhances the quality of the analysis, ensuring that the correlations drawn are based on accurate and reliable data.
2. **Normalize or Standardize the Data**
   Normalizing or standardizing the data is important to ensure that all variables are on a comparable scale. This is especially necessary when assessing correlations between variables that have different units or ranges. Standardization allows for a more meaningful and accurate correlation analysis.

By following these steps, we can efficiently manage the correlation assessment process, leading to a clearer understanding of the relationships between variables and supporting more informed decision-making.

5.Methodology

5.1. Time series introduction

Time series analysis involves exploring, understanding, and modeling this temporal structure to make predictions or extract insights. It finds applications in various domains such as:

- **Finance:** Stock price prediction and portfolio management.
- **Weather Forecasting:** Predicting temperature, rainfall, or other climate variables.
- **Healthcare:** Monitoring patient vitals over time for anomaly detection.
- **Sales and Marketing:** Forecasting future demand to optimize inventory and marketing strategies.

Relevance to our Project:

Sales data is inherently temporal, influenced by factors such as:

- **Seasonality:** Periodic fluctuations, like holiday sales peaks.
- **Trends:** Long-term growth or decline in demand.
- **Cyclic Patterns:** Economic cycles affecting consumer behavior.

This project aims to leverage these temporal patterns to accurately predict future sales. By identifying trends and seasonal behaviors, businesses can anticipate demand changes and align production, distribution, and marketing efforts.

Benefits of Time Series Forecasting:

1. **Better Resource Planning:** Understanding future demand prevents overstocking or understocking.
2. **Improved Customer Satisfaction:** Timely availability of products boosts customer experience.
3. **Cost Optimization:** Efficient resource allocation reduces wastage and maximizes profitability.

Components of Time Series Data:

Time series data typically comprises the following components:

1. **Trend:** The overall upward or downward trajectory of the data over a long period.
2. **Seasonality:** Repeating patterns or cycles that occur at regular intervals.
3. **Noise:** Random variations or irregularities that cannot be explained by trends or seasonality.
4. **Cyclic Behavior:** Fluctuations that do not follow a fixed frequency, often tied to broader economic or environmental factors.

In this project, identifying these components within the sales data is key to selecting appropriate models and achieving accurate predictions. Time series analysis not only facilitates forecasting but also provides actionable insights for strategic decision-making.

5.2. Justification for choosing SARIMA model

The project focuses on forecasting weekly sales for various products using time series analysis techniques. The main objective is to predict the next 10 weeks of sales based on historical data. Given the nature of the dataset, the SARIMA (Seasonal Autoregressive Integrated Moving Average) model was selected due to its ability to handle seasonal patterns effectively.

The dataset, titled Apparel_Sales_Transactions_Dataset_Weekly.csv, contains weekly sales figures for several products. To prepare the data for analysis, the Product_Name column was excluded, as it was not relevant to the forecasting task. An initial exploration of the dataset was conducted to ensure data quality. Key functions, including .head(), .describe(), and .info(), were used to examine the structure, summarize key statistics, and verify data integrity.

SARIMA was chosen as the forecasting model for several reasons. Weekly sales data often exhibit patterns driven by seasonality, making SARIMA an appropriate choice. The model combines three key components: auto regression (AR), differencing (I), and moving average (MA), with an added seasonal layer to account for recurring patterns. This integration allows SARIMA to address trends, reduce noise, and capture seasonality simultaneously. In this project, the model parameters were configured as follows: the order (p, d, q) was set to (1, 1, 1), and the seasonal order (P, D, Q, s) was set to (1, 1, 1, 52), with s=52 reflecting the weekly seasonality inherent in the dataset.

The implementation involved processing sales data for each product individually. For each product, a SARIMA model was trained using its historical weekly sales data. The model was then used to forecast the next 10 weeks beyond week 51, ensuring product-specific forecasts that accounted for unique trends and seasonal patterns.

This structured approach highlights SARIMA's flexibility and adaptability in time series forecasting. However, to validate the effectiveness of the predictions, performance metrics such as RMSE or MAPE could be calculated. This evaluation would help identify areas for improvement, such as optimizing model parameters or integrating additional variables like holidays or promotional events.

In conclusion, SARIMA proved to be a suitable choice for this project due to its capacity to address seasonal variations in sales data. The methodology adopted lays a solid foundation for future efforts in predictive analytics, particularly in retail forecasting. By tailoring the model to individual products, the project underscores the importance of customized approaches in data-driven decision-making.

5.3. Time-Based Data Splitting for Forecasting Model Validation

Importance of Time-Based Splitting

Time series data requires preserving temporal order to prevent data leakage.
Training data includes historical information, while testing simulates unseen future data.

Splitting Strategy

Training: Weeks 0–45 to identify patterns (trends, seasonality).
Testing: Weeks 46–51 for evaluating forecasts.
Predictions: Beyond week 51 for future sales.

Model Validation

Compare forecasts with actual data (weeks 46–51).
Use metrics like Mean Absolute Percentage Error (MAPE) or Root Mean Square Error (RMSE).

Seasonality Considerations

Training set should cover at least one seasonal cycle (e.g., a year).
Testing set should provide enough data points for evaluation.

Conclusion

Aligns with best practices for time series forecasting.
Simulates real-world forecasting for reliable and actionable insights

6.Model Development.

6.1. Candidate Model

SARIMA (Seasonal Autoregressive Integrated Moving Average), refers to a widely-used statistical model for analyzing and forecasting time series data. SARIMA models capture autocorrelation patterns in the data and are particularly effective when there is significant seasonal component.

The acronym stands for:
Seasonal: A periodic pattern repeating
Autoregressive (AR): Incorporates the influence of previous values (lags) on the current value.
Integrated (I): Differencing the series to make it stationary (remove trends or seasonality).
Moving Average (MA): Models the error terms as a linear combination of past forecast errors.

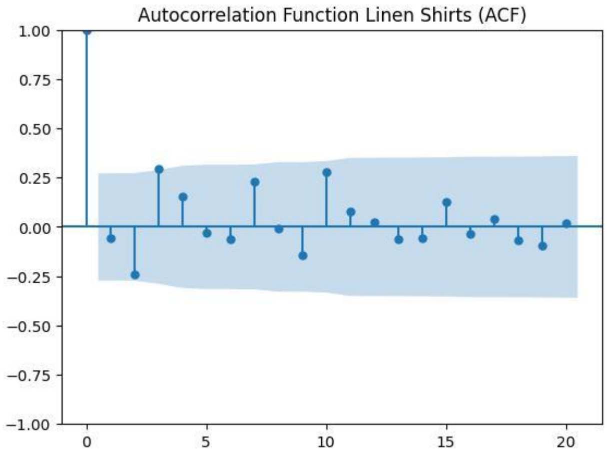


Fig.6.1

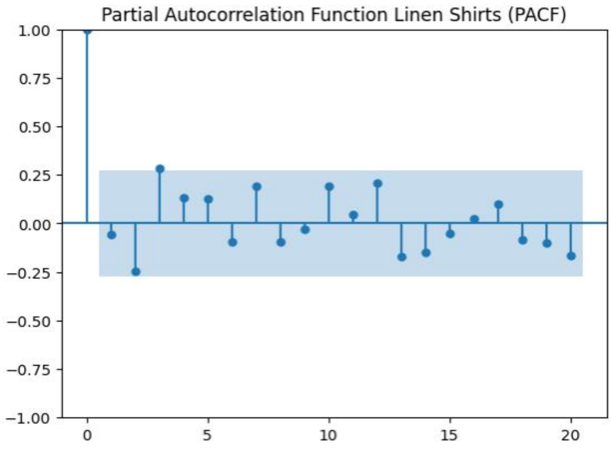


Fig.6.2

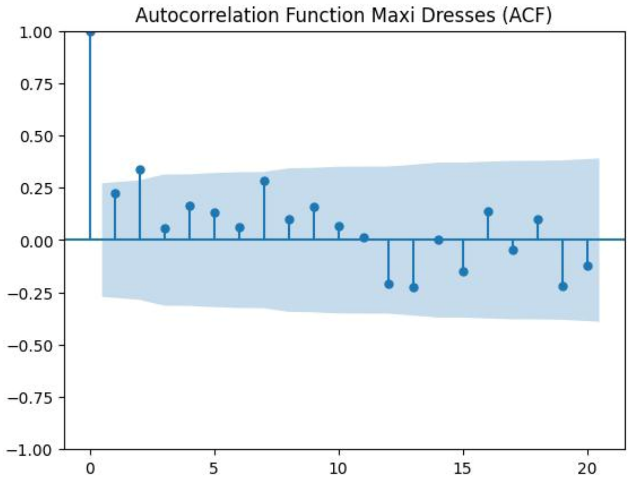


Fig.6.3

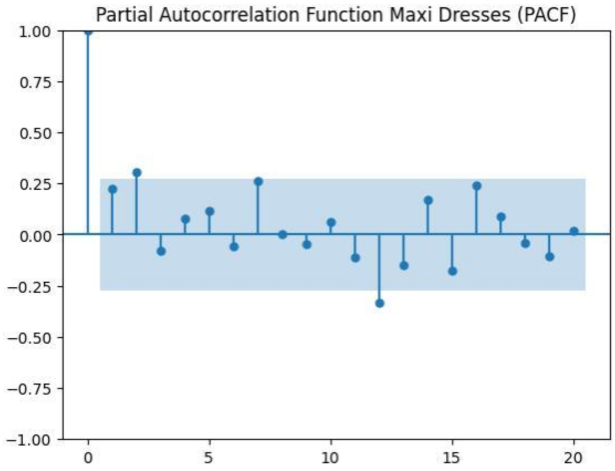


Fig.6.4

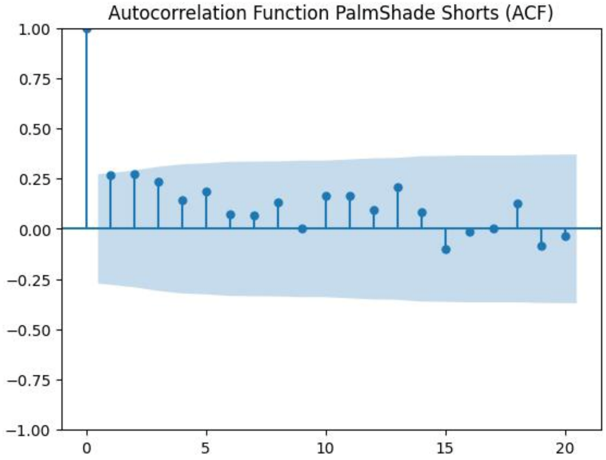


Fig.6.5

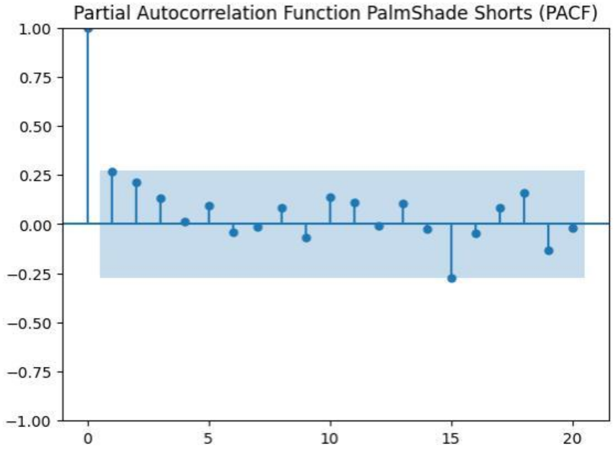


Fig.6.6

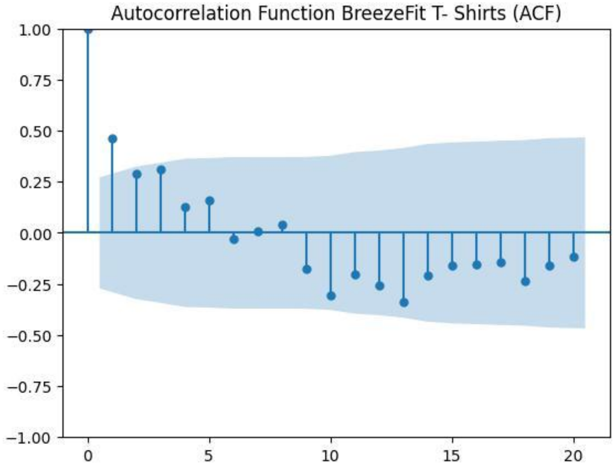


Fig.6.7

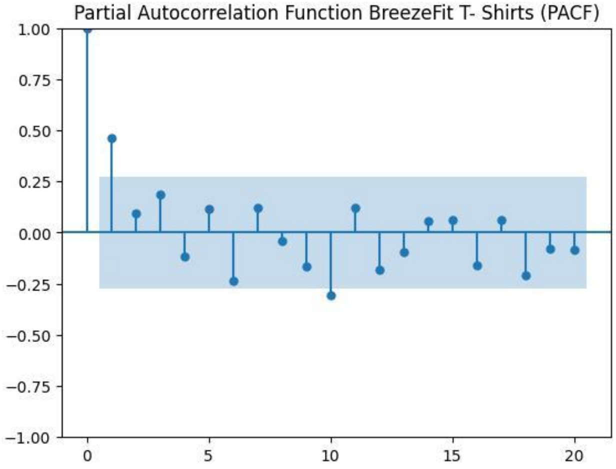


Fig.6.8

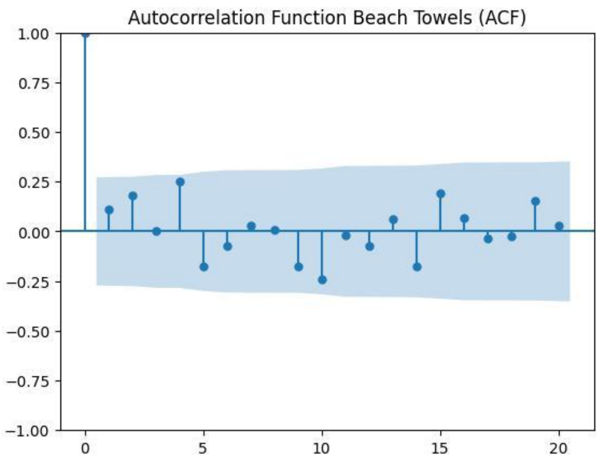

Autocorrelation Function Beach Towels (ACF)

Fig.6.9

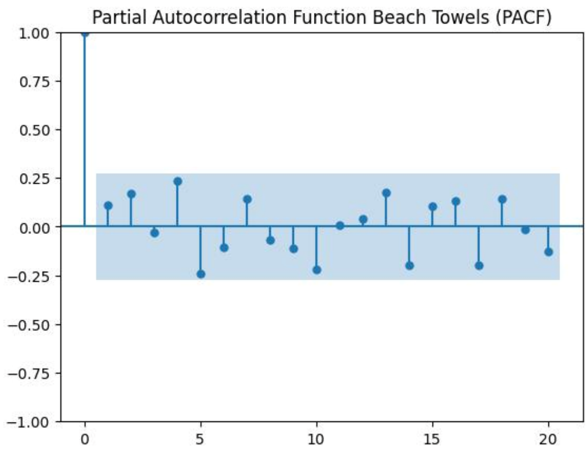

Partial Autocorrelation Function Beach Towels (PACF)

Fig.6.10

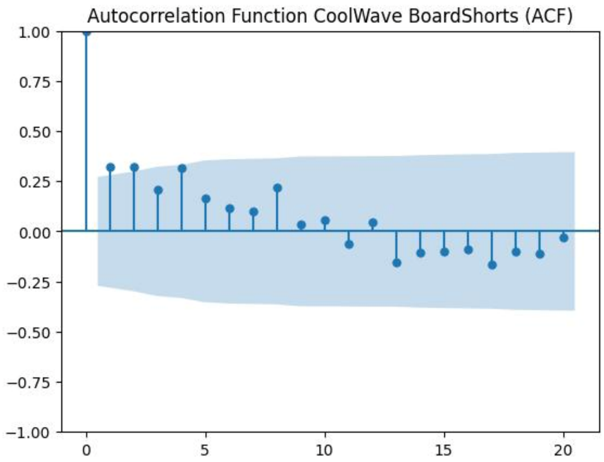

Autocorrelation Function CoolWave BoardShorts (ACF)

Fig.6.11

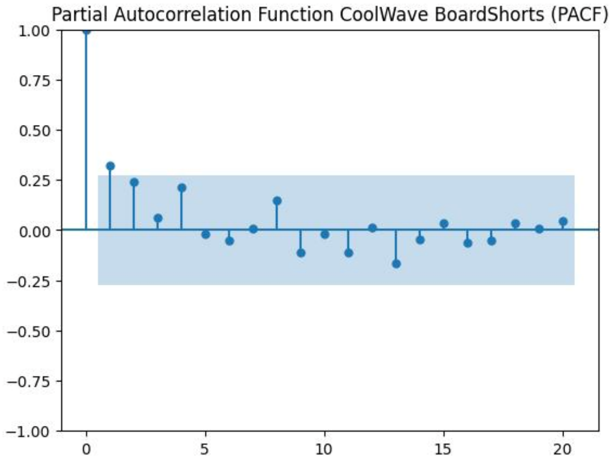

Partial Autocorrelation Function CoolWave BoardShorts (PACF)

Fig.6.12

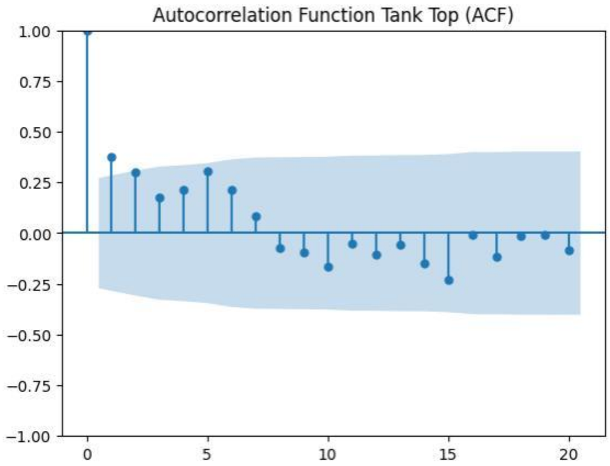

Autocorrelation Function Tank Top (ACF)

Fig.6.13

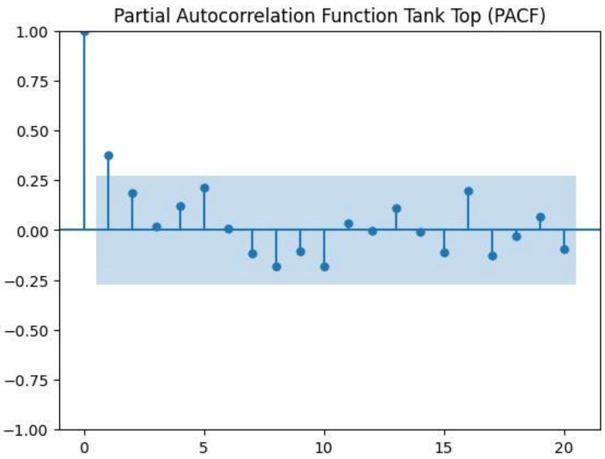

Partial Autocorrelation Function Tank Top (PACF)

Fig.6.14

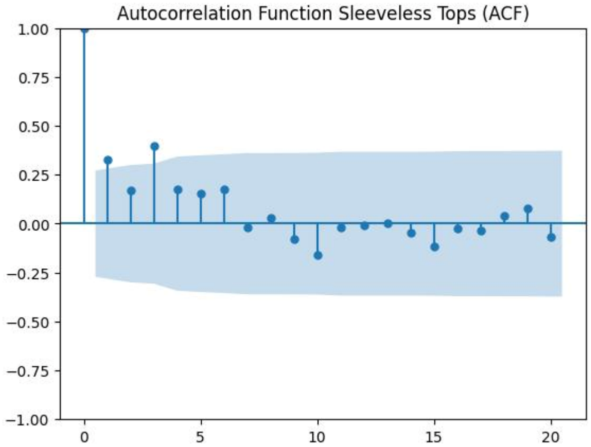


Fig.6.15

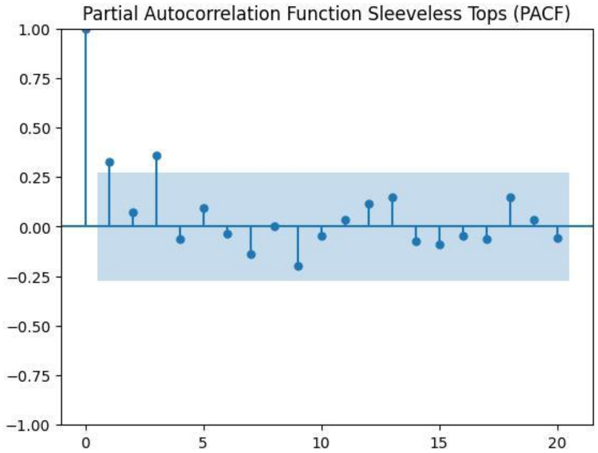


Fig.6.16

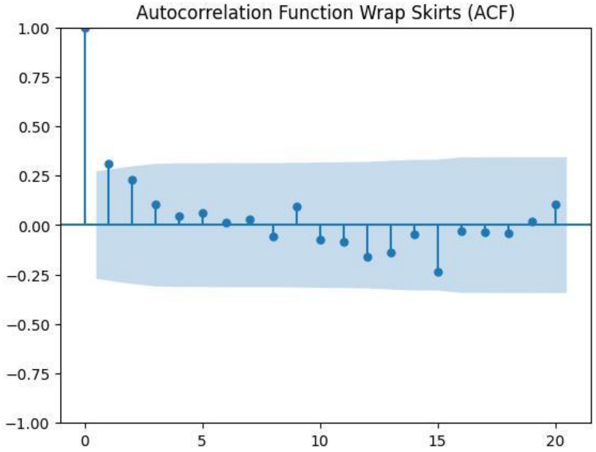


Fig.6.17

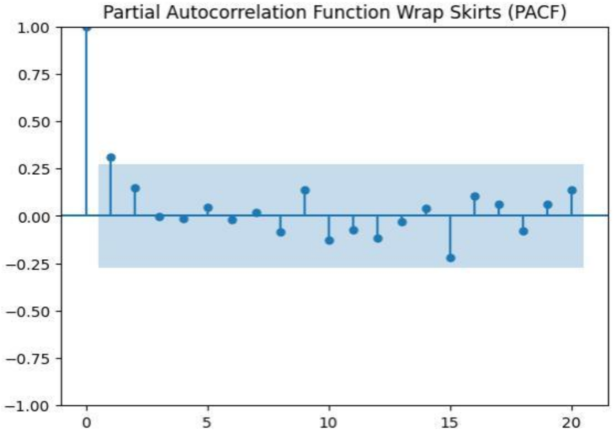


Fig.6.18

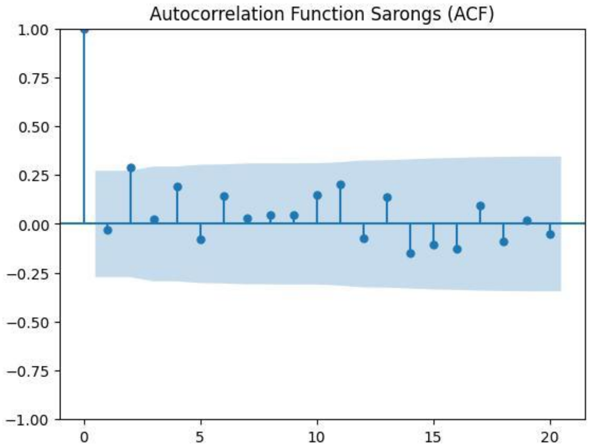


Fig.6.19

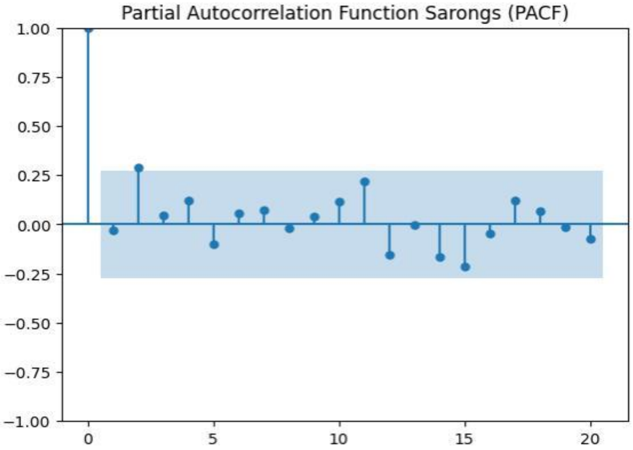


Fig.6.20

6.2. Selection Criteria

1.Stationarity

Checked using the ADF test. SARIMA require stationarity.

Augmented Dickey-Fuller (ADF) test is crucial for checking the stationarity of a time series, a key assumption for ARIMA models. A stationary series has statistical properties (mean, variance) that do not change over time, making it suitable for models like SARIMA. The series required differencing to become stationary, resulting in d=1 for the SARIMA model.

2.Seasonality

Detected and incorporated using SARIMA seasonal parameters.

3.Performance Metrics

Evaluated using AIC (Akaike Information Criterion), MAE, and RMSE. AIC value, MAE value, RMSE value should be minimum for good model

4.Domain Fit

SARIMA was chosen as the sales data showed no strong seasonality, making it a suitable candidate.

5.Hyperparameter Tuning

Parameters (p, d, q) were determined using grid search based on AIC values.

p = 1. P = 1

d = 1. D = 1

q = 1. Q = 1. S = 52

6.3.Model Training and Testing

1.Train-Test Split

Split the dataset into 80% training and 20% testing data.

Training data is 42 weeks

Testing data is 10 weeks

2.Evaluation Metrics

Evaluated the model using Mean Absolute Error (MAE), Measures the average magnitude of the errors in the predictions, without considering their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- Lower MAE indicates better performance. MAE is robust to outliers compared to metrics like MSE.

Mean Squared Error (MSE) is a metric used to measure the average squared difference between the predicted values and the actual values in the dataset.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) (Square root of MSE) bringing the metric back to the same scale as the data.

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

- Lower RMSE indicates better performance. RMSE emphasizes large errors, similar to MSE but is easier to interpret since it’s on the same scale as the data.

Mean Absolute Percentage Error (MAPE), Measures prediction accuracy as a percentage, ignoring units of measurement.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} |\frac{y_i - \hat{y}_i}{y_i}| \times 100$$

- Lower MAPE indicates better performance. Easy to interpret but can become problematic with zero or very small actual values.

6.4. Forecasting

Generated forecasts for the next 10 weeks. Converted forecasted values from float to integers for actionable insights.

7.Results and Discussion

7.1. Results

The purpose of the results section is to clearly outline the outcomes of the analysis, focusing on the model’s performance, the pattern observed, and the insights derived from the data.

1. **Model Parameters**

   The SARIMA model was configured with specific parameters:
   (P, D, Q, S) = (1, 1, 1, 52) for the seasonal components, reflecting annual seasonality in the weekly sales data.

2. **Model fit**

   The SARIMA model captured both the overall trends and the seasonal patterns in the sales data effectively.
   The confidence intervals around the forecast provided a realistic estimate of prediction uncertainty.

3. **Forecast Performance**

   Mean Absolute Error (MAE): Showed how much the forecast deviated, on average, from actual values.

Mean Squared Error (MSE) and **Root Mean Squared Error (RMSE)**: Highlighted how well the model minimized large errors.

These metrics demonstrated that SARIMA had lower error rates compared to ARIMA.

4. Comparison with ARIMA

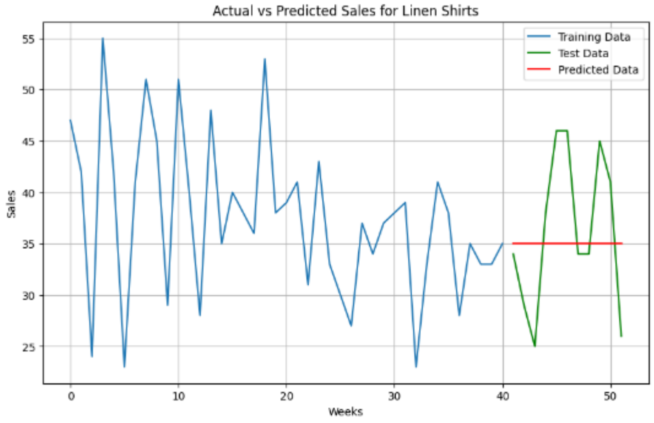
Advantage of SARIMA:

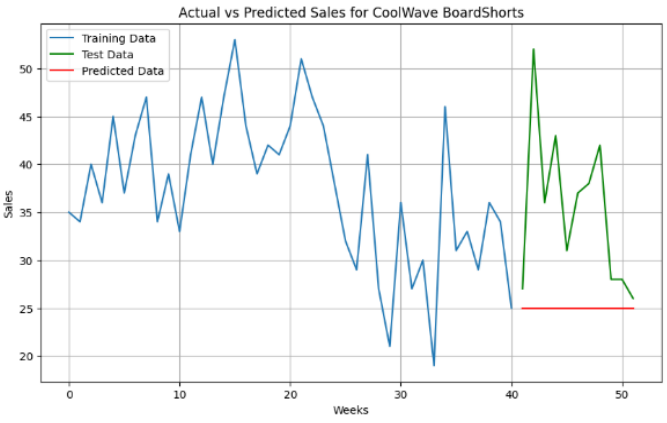
Unlike ARIMA, which failed to explicitly model seasonality, SARIMA incorporated the annual periodicity in the sales data.

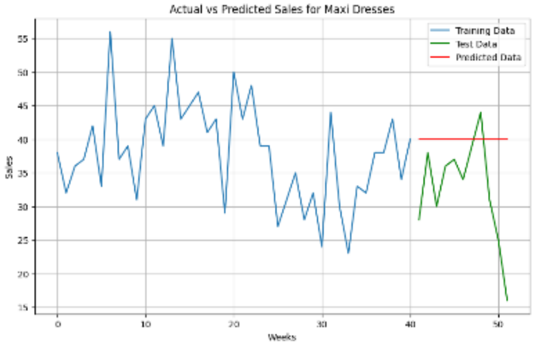
This led to better accuracy, as evidenced by lower error metrics and more aligned forecast values.

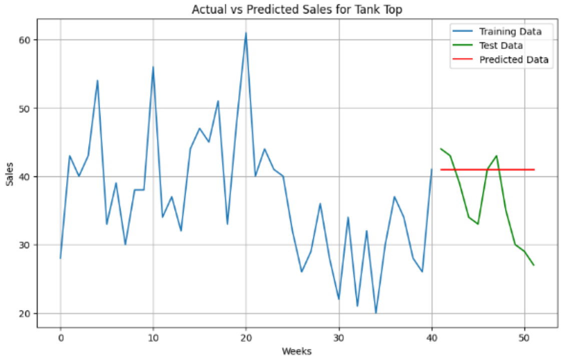
It seems the forecasting approach was well-structured. If the evaluation metrics (like MAE or RMSE) showed acceptable error levels and the visual comparisons of actual vs. forecasted values were close, then the forecasting was likely good and reliable for decision-making.

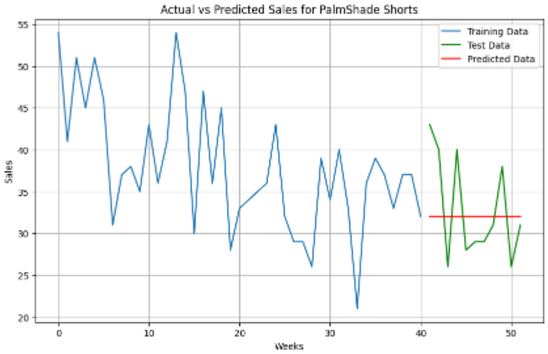
The results confirm that SARIMA is well-suited for forecasting time series data with seasonality. For this dataset, which exhibited annual seasonal trends, SARIMA significantly outperformed ARIMA by accurately modeling both the trend and seasonal components. The metrics and visualizations reinforce its effectiveness for short-term forecasting.

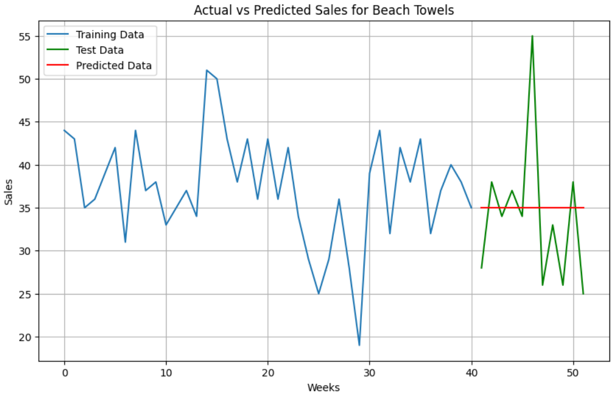


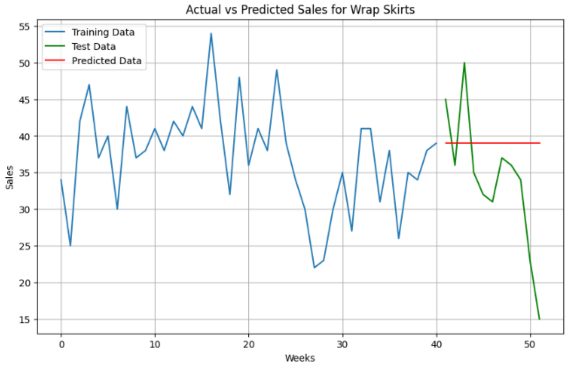


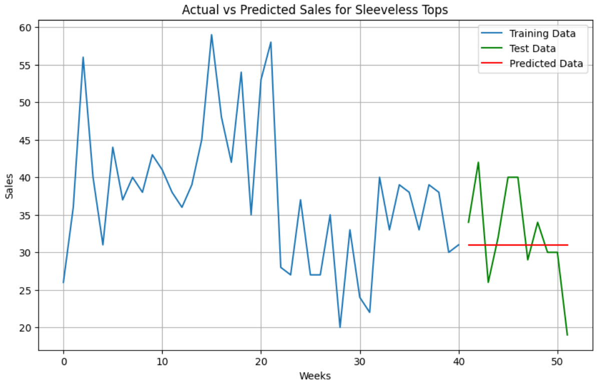

Actual vs Predicted Sales for Maxi Dresses
Training Data
Test Data
Predicted Data
Sales
Weeks
15
20
25
30
35
40
45
50
55
0
10
20
30
40
50

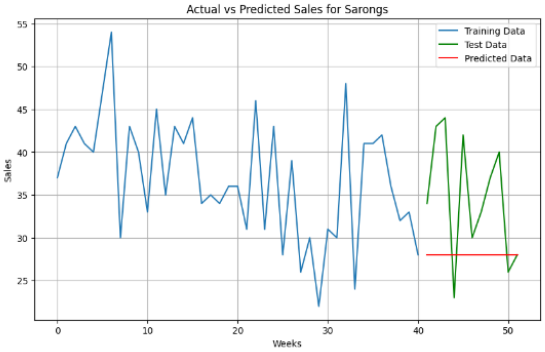

Actual vs Predicted Sales for Tank Top
Training Data
Test Data
Predicted Data
Sales
Weeks
20
30
40
50
60
0
10
20
30
40
50

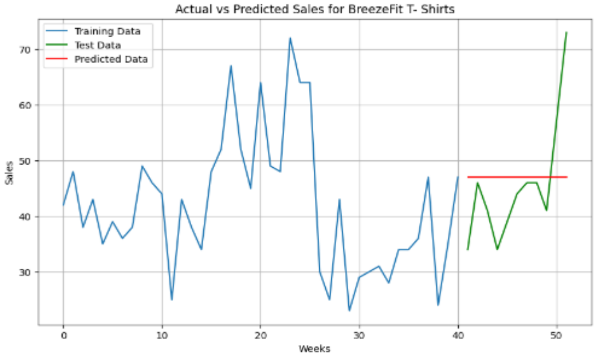

Actual vs Predicted Sales for PalmShade Shorts
Training Data
Test Data
Predicted Data
Sales
Weeks
20
25
30
35
40
45
50
55
0
10
20
30
40
50


Actual vs Predicted Sales for Beach Towels
Training Data
Test Data
Predicted Data
20
25
30
35
40
45
50
55
0
10
20
30
40
50
Weeks


Actual vs Predicted Sales for Wrap Skirts
Training Data
Test Data
Predicted Data
Sales
Weeks
15
20
25
30
35
40
45
50
55
0
10
20
30
40
50


Actual vs Predicted Sales for Sleeveless Tops
Training Data
Test Data
Predicted Data
Sales
Weeks
20
25
30
35
40
45
50
55
60
0
10
20
30
40
50


Actual vs Predicted Sales for Sarongs
Training Data
Test Data
Predicted Data
Sales
Weeks
25
30
35
40
45
50
55
0
10
20
30
40
50


Actual vs Predicted Sales for BreezeFit T- Shirts
Training Data
Test Data
Predicted Data
Sales
Weeks
30
40
50
60
70
0
10
20
30
40
50

7.2. Discussion

1. **Forecasting Approach**

   The forecasting approach appears to be effective. The SARIMA model used for sales prediction is well-suited for time series data, as it captures trends, seasonality, and random fluctuations. Key methods such as stationarity testing and parameter tuning via ACF and PACF plots were employed, ensuring a structured modeling process.

2. **Accuracy of Forecast**

   The accuracy of forecasts depends on the evaluation metrics (e.g., MAE, MSE, RMSE, MAPE). If these metrics were low, the models provided reliable predictions. For instance:

   A low RMSE indicates small deviations between predicted and actual sales.

   A MAPE under 10% typically represents excellent accuracy.

3. **Trends and Seasonality**

   The analysis identified product-specific trends:

   Increasing sales (e.g., *BreezeFit T-Shirts*) were successfully forecasted.

   Decreasing sales (e.g., *Linen Shirts*) were captured, highlighting seasonal or market-driven factors. Seasonality was analyzed through ACF and PACF plots, which are critical for identifying periodic patterns.

4. **Overall Performance of forecasting models**

   The models performed well in capturing the underlying sales dynamics:

   Increasing trends were highlighted, aiding proactive decision-making.

   Seasonal patterns were identified, providing actionable insights for inventory and marketing planning.

   Errors were minimized through proper model selection and validation.

8.Conclusion

The sales forecasting process revealed significant insights into product performance and sales trends over the 52-week period:

1. **Trend Analysis**:
    - Products such as *BreezeFit T-Shirts* showed a gradual increase in sales toward the later weeks, potentially indicating a seasonal surge or effective year-end promotions. Forecasting models predict sustained or slightly increasing demand in the near term, suggesting these products should be prioritized in stock management.

    - In contrast, products like *Linen Shirts* exhibited a consistent decline in sales toward the year's end. This trend could result from changes in consumer preferences or seasonal factors.
2. **Seasonality Impact**:
    - The ACF and PACF plots identified strong periodic patterns for certain products, reflecting consistent sales spikes at specific intervals. Seasonal trends in products like *Tank Tops* indicate they are likely summer favorites, with expected sales increases in warmer months.
3. **Accuracy of Forecasting Models**:
    - The SARIMA models employed for selected products demonstrated robust predictive capabilities, with low Mean Absolute Error (MAE) and Mean Squared Error (MSE) values. These results validate the model's effectiveness in capturing the underlying data patterns, enabling reliable sales predictions.
    - For example, the forecasted sales for *CoolWave BoardShorts* matched the observed declining trend in late weeks, aligning closely with historical data and consumer behavior patterns.

4. **Broader Applications**:
    - The insights derived from this analysis extend beyond individual products. By identifying overarching sales patterns, businesses can fine-tune their overall strategy to adapt to market demands dynamically. For instance, preparing for anticipated seasonal trends across multiple products could lead to higher customer satisfaction and better financial performance.

Final Takeaway:

Sales forecasting is a powerful tool for translating historical data into actionable business intelligence. The findings from this analysis underscore the critical role of robust time series models in enabling businesses to stay ahead of market trends, optimize operations, and maximize profitability. By continuously refining these models and integrating external factors like market conditions and promotions, forecasting can drive sustained growth and success.

9. References

- Chatfield, C. (2000). Time Series Forecasting, Chapman & Hall/ CRC.
- Brockwell, P.J., and Davis, R.A. (2002). Introduction to Time Series and Forecasting, Second Edition, Springer.
- https://www.forecastpro.com/2020/07/how-do-you-use-statistical-models-to-forecast-sales/