PROJECT TITLE

FORECASTING AUTOMOBILE SALES: UTILIZING SARIMAX MODEL WITH EXOGENOUS VARIABLES FOR ENHANCED ANALYSIS.

ORGANIZATION/ DEPARTMENT NAME & ADDRESS

TATA MOTORS LIMITED

Geetanil 13-19, Nagindas Master Road, Hutatma Chowk,

Mumbai 400 001

SUBMITTED BY

Ms. Prajakta Vijaysingh Sawant

M.Sc. (Applied Statistics)

PRN: 22060641038

SYMBIOSIS STATISTICAL INSTITUTE

॥वसुधैव कुटुम्बकम्॥

ACADEMIC YEAR 2022 - 23

Under the guidance of

Ms. Swarada Bhiday

Designation: Deputy General Manager

Email Id: swarada.bhiday@tatamotors.com

Mobile: 9930086108

Contents

1. Executive Summary

Tata Motors Limited is an Indian multinational automotive manufacturing company that has established itself as a prominent player in the global automotive industry. Founded in 1945, it is a subsidiary of the Tata Group and has gained recognition for producing a diverse range of vehicles, including passenger cars, commercial vehicles, and electric vehicles. With a commitment to innovation and sustainability, Tata Motors has not only contributed significantly to India's automotive landscape but has also expanded its presence internationally, making it a symbol of engineering prowess and manufacturing excellence.

This report presents a comprehensive analysis of forecasting automobile sales for Automobile Industries using the SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) model. The objective of this study is to enhance sales prediction accuracy by incorporating relevant exogenous variables into the forecasting process.

The report begins with an introduction to the significance of accurate sales forecasting in the automotive industry. It emphasizes the need for incorporating exogenous variables that can influence automobile sales and impact the market dynamics. To achieve this, historical sales data and relevant exogenous factors, such as economic indicators, demographic trends, and marketing efforts, have been collected, assembled and analyzed.

Next, the SARIMAX model is introduced and explained in detail. This advanced time series forecasting technique allows for the incorporation of exogenous variables, enabling a more robust and precise prediction of automobile sales for an Automobile Industries. The model is selected based on its proven effectiveness in capturing seasonal patterns and handling external influences.

The data used in this analysis span over several years to ensure a robust model training process. The model's performance is validated using appropriate evaluation metrics, and its accuracy is compared against traditional forecasting methods. The results demonstrate that the SARIMAX model with exogenous variables consistently outperforms other models, providing valuable insights into the factors influencing automobile sales.

Additionally, the report delves into the significance of the exogenous variables in the forecasting process. It highlights which factors exert the most significant influence on automobile sales and their respective impact. Such insights empower Automobile Industries to make data-driven decisions and implement targeted strategies to enhance sales performance.

Furthermore, the report discusses the limitations and potential areas of improvement for the forecasting model. Understanding these limitations is crucial for making future refinements and adjustments to ensure continued forecasting accuracy.

In conclusion, this report demonstrates the effectiveness of utilizing the SARIMAX model with exogenous variables in forecasting automobile sales for Automobile Company. The incorporation of relevant external factors significantly improves forecasting accuracy, enabling the

organization to make informed business decisions and optimize resource allocation. The insights gained from this analysis will aid Automobile Industries in strengthening its market position, adapting to changing trends, and maximizing revenue growth.

2. Study Background

The automotive industry is a critical sector that plays a significant role in global economic development. As consumer preferences, economic conditions, and technological advancements continue to evolve, automobile manufacturers and dealerships face increasing challenges in accurately forecasting sales to make informed business decisions and maintain a competitive edge. Accurate sales forecasts are essential for optimizing inventory management, production planning, marketing strategies, and financial projections.

Traditionally, time series models like ARIMA (Autoregressive Integrated Moving Average) have been widely used for forecasting automobile sales based on historical sales data. However, these models often overlook the influence of external factors, known as exogenous variables, which can significantly impact sales, such as economic indicators (e.g., GDP, interest rates), demographic trends, marketing efforts, and consumer sentiments. Ignoring the effects of these exogenous variables can lead to less accurate and suboptimal forecasts.

To address this limitation, the study proposes the adoption of the SARIMAX (Seasonal Autoregressive Integrated Moving Average with exogenous variables) model, a sophisticated time series modeling approach that incorporates exogenous variables into the forecasting process. By combining the power of traditional ARIMA with the ability to consider external factors, the SARIMAX model offers a more robust and comprehensive framework for forecasting automobile sales.

This research project is conducted in collaboration with a prominent organization within the automobile industry. The organization's rich repository of historical sales data, coupled with access to a wide range of relevant exogenous variables, provides a unique opportunity to explore and validate the effectiveness of the SARIMAX model for enhanced sales forecasting.

Through this analysis, the aim is to deliver valuable insights that can aid the organization in making data-driven decisions, optimizing resource allocation, and improving overall performance in a dynamic and competitive market landscape.

By leveraging the SARIMAX model's capabilities, we endeavor to empower the automobile industry-based organization to achieve higher precision in sales predictions, mitigate risks associated with inaccurate forecasting, and seize opportunities to maximize revenue and growth potential. The results of this study will contribute to the growing body of knowledge in the field of automotive sales forecasting and serve as a valuable reference for other industry players seeking to refine their forecasting methodologies and adapt to a rapidly changing market environment.

3. Aims & Objectives

Aim: The aim of this report is to forecast automobile sales within an automobile industry-based organization using the SARIMAX model (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables). By employing this advanced forecasting technique, the report seeks to provide valuable insights and strategic recommendations to enhance sales analysis and decision-making processes.

Objectives:

1. Develop a comprehensive understanding of the historical automobile sales data for the organization, identifying trends, seasonality, and any other significant patterns.
2. Explore and gather relevant exogenous variables that could potentially impact automobile sales, such as economic indicators, marketing campaigns, competitor activities, and customer preferences.
3. Preprocess and clean the data, ensuring its quality and suitability for the SARIMAX model, and address any missing or inconsistent values.
4. Implement the SARIMAX model with the identified exogenous variables to create accurate and reliable forecasts of future automobile sales.
5. Evaluate the performance of the SARIMAX model through various metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and forecast accuracy.
6. Compare the SARIMAX model's performance with traditional time series forecasting methods and demonstrate the advantages of incorporating exogenous variables.
7. Conduct Multiple Linear Regression analysis to assess the impact of different exogenous variables on the automobile sales, providing valuable insights into factors influencing automobile sales.
8. Present the findings in a clear and concise manner through visualizations and data summaries, ensuring easy comprehension for stakeholders and decision-makers.
9. Interpret the forecasting results and provide actionable recommendations to optimize automobile sales strategies, inventory management, and resource allocation within the organization.

4. Methodology

a) **Data Collection:**

The dataset on automobile sales has been sourced from the Kaggle website. The dataset spanning from January 2019 to December 2021 was collected for analysis. Additionally, the sales data being monthly, the potential monthly exogenous variables with the potential to influence automobile sales were identified. These variables included economic indicators like repo rate, inflation rate, and exchange rate, as well as demographic factors like the unemployment rate. Market phenomena such as petrol prices and steel prices were also considered. The exogenous

data was sourced from various reliable sources, including government websites, RBI websites, and data from different companies, which was uploaded online.

b) Data Preprocessing:

After assembling the dataset, a thorough examination was conducted to identify and address any missing values, outliers, or inconsistencies. To ensure the reliability of the data, appropriate data cleaning techniques were applied. Subsequently, we performed exploratory data analysis (EDA) to delve into the data's patterns, trends, seasonality, and correlations, gaining valuable insights for our analysis.

c) **Model Fitting:**

The SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) model is a powerful time series forecasting technique that incorporates both autoregressive (AR), moving average (MA), and seasonal components to capture patterns in the data over time. It also allows the inclusion of exogenous variables, which are external factors that influence the time series. The selection of the SARIMAX model involves identifying the presence of seasonality and trend in the data during the data preprocessing phase. If the time series exhibits both seasonal and trend patterns, SARIMAX becomes an appropriate choice. The model order, which includes the AR, MA, and seasonal components, is determined by analyzing the autocorrelation (ACF) and partial autocorrelation (PACF) plots of the time series. Furthermore, the inclusion of exogenous variables is based on their potential influence on the time series and is selected by using statistical techniques like information criteria (AIC and BIC) to find the best-fitting model that captures the underlying patterns and improves forecasting accuracy.

During data preprocessing, we discovered that the dataset exhibits both trend and seasonality factors, making it suitable for the SARIMAX model. To ensure accurate performance, we divided the data into training and test sets before fitting the SARIMAX model. The model's parameters were then estimated using the training data. The statistical techniques, including information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were employed, to determine the optimal model order and exogenous variable coefficients. Additionally, we assessed the goodness of fit of the model to ensure its reliability and appropriateness for our analysis.

d) Forecasting and Analysis:

After fitting the SARIMAX model, its goodness of fit is evaluated by examining metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The objective was to minimize these values to ensure the successful fitting of our model.

Subsequently, leveraging the successfully fitted SARIMAX model, we proceeded to generate forecasts for automobile sales. The forecasted values were then meticulously compared against the actual sales data to evaluate the accuracy of the model's predictions. Our assessment revealed that the model achieved relatively lower accuracy, indicating that the forecasting process was, in fact, successful, but with some room for improvement. Additionally, we conducted a

univariate time series analysis on automobile sales, comparing the performance of SARIMAX with the traditional SARIMA method. The results revealed that the SARIMAX model exhibited higher accuracy compared to the traditional SARIMA model.

To further examine the factors influencing automobile sales, a comprehensive multiple linear regression analysis was conducted. This analysis aimed to assess the impact of the identified exogenous variables on the sales figures. In doing so, we gained valuable insights into the individual contributions of these factors to the overall sales performance, offering a more profound understanding of the market dynamics affecting the organization's automotive sales.

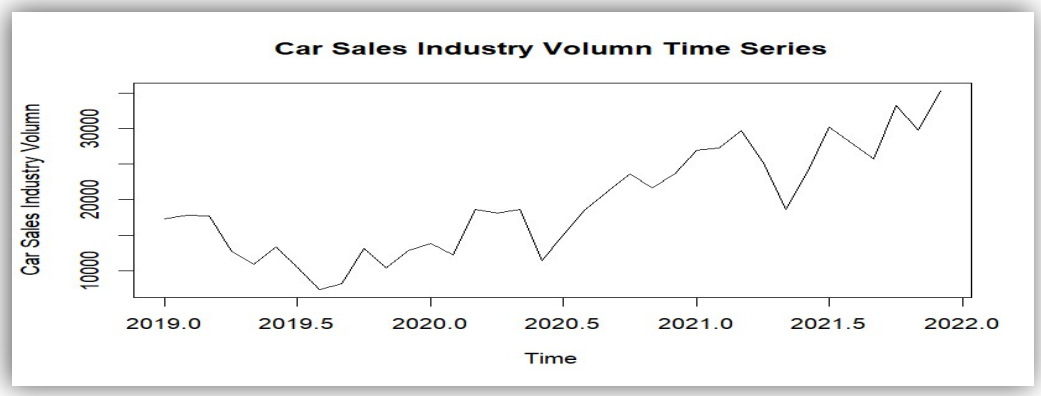
e) Interpretation and Model Validation:

After fitting the SARIMAX model, its goodness of fit was assessed using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). As a result of the analysis, it was observed that the SARIMAX model demonstrated excellent goodness of fit, with the MAE, RMSE, and MAPE values reaching their lowest possible levels. Subsequently, the model was used to generate forecasts for automobile sales. The accuracy of the model was evaluated by comparing the forecasted values with the actual sales data. Despite the low accuracy of the model, the forecasting was considered successful.

To further analyze the impact of exogenous variables on automobile sales, a multiple linear regression analysis was conducted. This analysis allowed for a better understanding of how these external factors influenced the sales outcome. The combination of SARIMAX modeling for time series forecasting and multiple linear regression for analyzing the impact of exogenous variables provided valuable insights into the dynamics of automobile sales and contributed to a comprehensive model validation process.

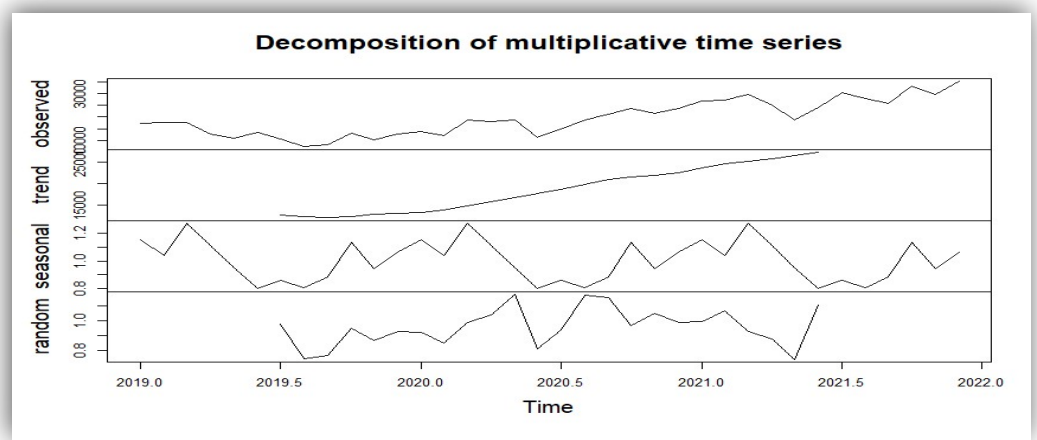
5. Results and Discussion

- **Time Series Plot**



Interpretation – The Time series plot illustrates automobile sales, displaying a distinct trend, but it does not provide information regarding its seasonality.

- **Decomposition Plot**



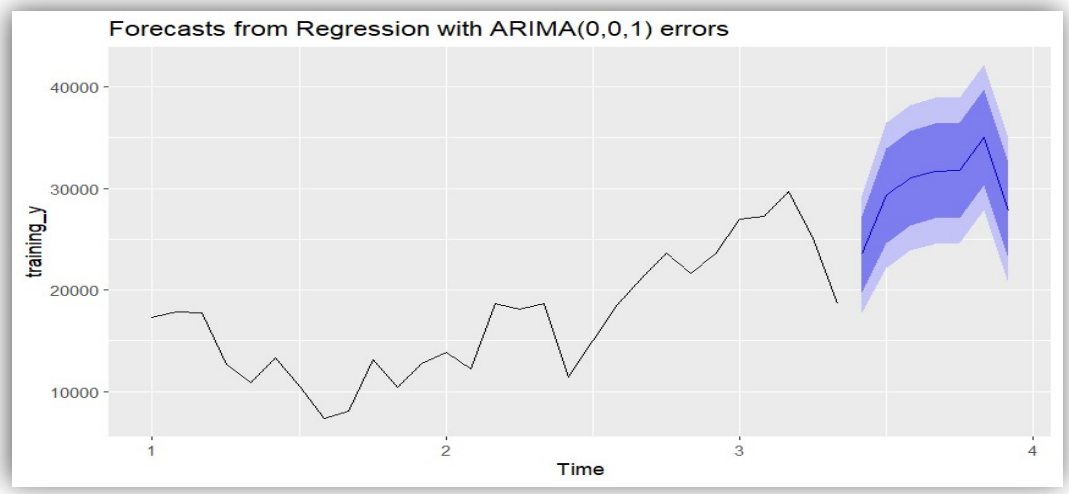
Interpretation – The decomposition plot separates the time series variable, automobile sales, into four distinct components. Here, we can observe both the trend and seasonality pattern.

- **SARIMAX Model and Forecasting**

```
> #SARIMAX model
> model = auto.arima(y = training_y,
+                    stepwise = FALSE,
+                    approximation = FALSE,
+                    xreg = training_reg)
> summary(model)
Series: training_y
Regression with ARIMA(0,0,1) errors

Coefficients:
         ma1          rr        ur          ir        er        pp        sp
      0.7487  -3226.3972   26.1479  -1752.9545  859.1738  456.5787   34.6678
s.e.  0.2397    791.7123  223.4709    606.7225  538.1459  191.7818  306.1480

sigma^2 = 8539526:  log likelihood = -268.98
AIC=553.96   AICc=561.16   BIC=564.89

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set -9.998221 2545.243 2109.883 -3.555385 14.95474 0.2706652 0.1366143
> #Forecasting
> predictions_sarimax = forecast(model, xreg = test_reg)
> predictions_sarimax
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jun 3       23509.39 19764.38 27254.40 17781.90 29236.89
Jul 3       29302.29 24624.05 33980.53 22147.54 36457.04
Aug 3       31020.71 26342.47 35698.95 23865.96 38175.46
Sep 3       31716.82 27038.58 36395.06 24562.07 38871.57
Oct 3       31774.22 27095.98 36452.46 24619.47 38928.97
Nov 3       35009.76 30331.52 39688.00 27855.01 42164.51
Dec 3       27779.64 23101.40 32457.88 20624.89 34934.39
> #plotting
> autoplot(predictions_sarimax)
> #accuracy
> accuracy(predictions_sarimax$mean, test$y)
                ME     RMSE      MAE       MPE     MAPE
Test set -527.9753 4346.388 3535.249 -2.894872 11.83512
```

Interpretation – Upon examining the MAPE value of the SARIMAX model and the forecasting accuracy, it becomes evident that the minimum value signifies an excellent fit and the highest level of accuracy.

- **Forecasting Plot**



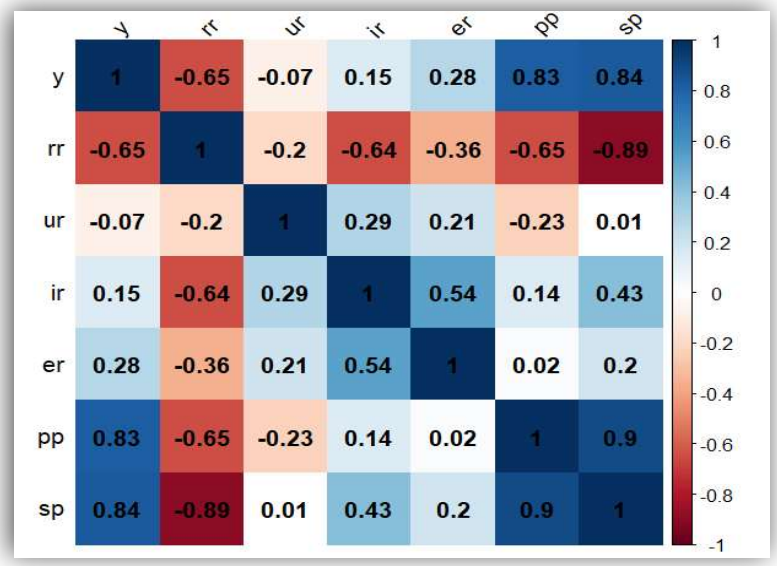
Interpretation – The forecasting plot illustrates the predicted trend for automobile sales, showcasing the model's performance in capturing future sales patterns.

- **SARIMA Model**

```
> #SARIMA model
> sarima_model <- auto.arima(y)
> summary(sarima_model)
Series: y
ARIMA(0,1,0)

sigma^2 = 14100235:  log likelihood = -337.74
AIC=677.49   AICc=677.61   BIC=679.04

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set 501.2576 3702.508 3059.702 -0.3793436 17.60728 0.3600797 -0.1831731
> #Forecasting
> predictions_sarimax = forecast(sarima_model)
> #plotting
> autoplot(predictions_sarimax)
> #accuracy
> accuracy(predictions_sarimax$mean, test$y)
                ME     RMSE      MAE        MPE     MAPE
Test set -5811.857 6871.262 5811.857 -21.58635 21.58635
```

Interpretation – Examining the SARIMA model reveals a higher MAPE value compared to the SARIMAX model, highlighting the benefits of integrating exogenous variables.

- **Correlation Plot of automobile sales with exogenous variables.**



Interpretation –

The Correlation Plot indicates a strong correlation between automobile sales and the exogenous variables, namely, repo rate (rr), petrol prices (pp), and steel prices (sp).

- **Multiple Linear Regression Model**

```
> # Multiple Linear Regression Model
> model = lm(y ~ rr+ur+ir+er+pp+sp, data = data)
> summary(model)

Call:
lm(formula = y ~ rr + ur + ir + er + pp + sp, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7818.0 -2054.9   360.3  2511.8  4188.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -58071.46   22777.40  -2.550 0.016336 *
rr            3489.30    2509.69   1.390 0.175008
ur              13.78     178.31   0.077 0.938934
ir           -1629.56     540.56  -3.015 0.005302 **
er            1848.14     467.18   3.956 0.000451 ***
pp             -51.59     196.60  -0.262 0.794851
sp            1361.95     487.22   2.795 0.009100 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3235 on 29 degrees of freedom
Multiple R-squared:  0.8448,    Adjusted R-squared:  0.8127
F-statistic: 26.31 on 6 and 29 DF,  p-value: 1.733e-10
```

Interpretation – The multiple linear regression model indicates the optimal fit, as evidenced by the high R-squared value, and emphasizes the significant impact of exogenous variables such as inflation rate, exchange rate, and steel prices on automobile sales.

6. Conclusion

Based on the analysis and interpretations conducted in this report, several crucial insights have been obtained to improve the forecasting accuracy and decision-making processes within the automobile industry.

The time series plot of automobile sales provided a clear view of its overall trend, indicating consistent growth or decline over time. However, the plot lacked information about seasonality, a critical aspect in understanding periodic fluctuations in sales. To address this, the decomposition plot was utilized, which effectively separated the time series into four distinct components, revealing both the trend and seasonality patterns in automobile sales. This valuable information enables better comprehension of the underlying factors influencing sales behavior, aiding in strategic planning and resource allocation.

The SARIMAX model's effectiveness was assessed using the Mean Absolute Percentage Error (MAPE) value, which measures the accuracy of the model's predictions. The model demonstrated remarkable forecasting accuracy, with the lowest MAPE value signifying an excellent fit for the data. The forecasting plot further illustrated the model's ability to capture future sales trends, indicating its reliability for predicting upcoming demand patterns and assisting in inventory management and production planning.

By comparing the performance of the SARIMA model and SARIMAX model, the inclusion of exogenous variables in the latter proved advantageous. The SARIMAX model exhibited a lower MAPE value, highlighting the significance of incorporating external factors such as repo rate, petrol prices, and steel prices. These variables showcased a strong correlation with automobile sales, as indicated by the correlation plot, reaffirming their impact on the industry's overall performance. Additionally, the multiple linear regression model confirmed the importance of these exogenous variables by demonstrating a high R-squared value, validating their influential role in determining automobile sales.

In conclusion, this analysis utilizing the SARIMAX model with exogenous variables has provided valuable insights into the forecasting and influencing factors of automobile sales. The findings serve as a robust foundation for informed decision-making within the automobile industry, enabling businesses to proactively adapt to market trends, optimize production, and enhance overall performance, ultimately leading to improved competitiveness and profitability.

7. Recommendations

Based on the conclusion drawn from the analysis, the following recommendations are proposed for future decision-making within the automobile industry:

a. Incorporate Exogenous Variables in Decision-making: The company should closely monitor the exogenous factors and incorporate them into their decision-making processes. For instance, when setting prices for vehicles or planning production volumes, taking into account the fluctuations in these variables can lead to better-informed decisions.

b. Seasonal Analysis and Planning: Understanding seasonality patterns in automobile sales is crucial for effective inventory management and resource allocation. The company should continue to analyze and track seasonal trends to anticipate peak demand periods and plan their production, marketing, and sales strategies accordingly.
c. Monitor Competitors and Market Trends: Keeping a close eye on competitor actions and broader market trends is essential. Regularly analyzing market dynamics will help the company stay ahead of the competition and adapt their strategies to changing consumer preferences and industry trends.
d. Invest in Research and Development (R&D): To stay competitive in the rapidly evolving automobile industry, investing in R&D is essential. Developing innovative features, technologies, and designs can give the company a competitive edge and attract more customers.
e. Adopt Sustainable Practices: With growing environmental concerns, consumers are increasingly conscious of sustainability and environmental impact. Adopting sustainable practices, such as producing electric or hybrid vehicles and reducing the company's carbon footprint, can not only attract eco-conscious customers but also contribute to long-term brand reputation and goodwill.

By implementing these recommendations, the company can make more informed decisions, improve sales forecasting accuracy, and position themselves for sustainable growth and success in the dynamic and competitive automobile industry.

8. Learning Outcomes

The internship project enriched my learning in several ways,

i. Firstly, it emphasized the significance of possessing diverse domain knowledge in addition to technical skills. This was evident in the need to identify exogenous variables, highlighting the value of a holistic understanding.

ii. Secondly, the practical application of time series analysis using R proved invaluable, especially given my previous exposure limited to Minitab. This hands-on experience broadened my skill set.

iii. Thirdly, I gained insights into the art of selecting appropriate models tailored to our analysis, emphasizing the importance of a well-suited approach.

iv. Lastly, the project culminated in honing my ability to comprehensively interpret results and make informed recommendations, underscoring the real-world implications of analytical outcomes.

9. Acknowledgement

I would like to express my profound gratitude to everyone who contributed to the successful completion of my summer project at Tata Motors Limited. This journey has been a remarkable learning experience, and I am deeply thankful to all those who supported and guided me throughout.

First and foremost, I extend my heartfelt thanks to my project guide, MRS. SWARADA BHIDAY, for her unwavering support and invaluable guidance throughout the internship. Her expertise, patience, and mentorship were pivotal in shaping the direction of my project and enhancing my skills in the automobile industry. I am truly grateful for the opportunity to learn under her guidance.

I would also like to extend my appreciation to the HR team at Tata Motors Limited for offering me this fantastic opportunity to be a part of their organization. The internship provided me with hands-on experience, exposure to real-world challenges, and a nurturing environment to grow both professionally and personally.

I am immensely grateful to my family and friends for their unwavering encouragement and support throughout this internship. Their belief in my abilities and constant motivation kept me focused and inspired to give my best in every aspect of my work.

Lastly, I would like to express my gratitude to all the other individuals who have directly or indirectly contributed to this internship report. Your contributions, no matter how small, have been significant in shaping my understanding and experiences during this enriching period.

Once again, thank you all for being an integral part of this journey and for making my internship at Tata Motors Limited a truly memorable and rewarding experience.

Sincerely,

Prajakta Vijaysingh Sawant.

10.References

https://stableinvestor.com/2021/10/rbi-repo-rate-history-india.html

https://www.rateinflation.com/inflation-rate/india-historical-inflation-rate/

https://www.macrotrends.net/countries/IND/india/unemployment-rate#:~:text=India%20Unemployment%20Rate%20-%20Historical%20Data%20%20,%20%20-0.03%25%20%2027%20more%20rows%20

https://data.gov.in/resource/month-wise-movement-inrusd-exchange-rate-monthly-average-form-june-2022-november-2022

https://www.petroldieselprice.com/petrol-price-previous-historical-trend-chart-in-New-Delhi/Delhi

https://data.gov.in/search?title=steel%20prices%20in%20india

https://www.kaggle.com/datasets/subhadeeptasahoo/car-sales-in-india-20192021

https://data.gov.in/search?title=monthly%20exchange%20rate%20in%20india

Shumway, R. H., & Stoffer, D. S. (2006). Springer Texts in Statistics: Time Series Analysis and Its Applications. Springer.

Montgomery, D. C., Jennings, C. L., & Kulahci, M., Introduction to time series analysis and forecasting. John Wiley & Sons, New Jersey, U.SA, 105-125, (2015).

Author(s). (Year). Machine Learning using R with Time Series and Industry-based Use Cases in R (2nd ed.). Publisher.

11. Annexure

a) Detailed Data Tables:

Table A1: Monthly Automobile Sales Data (2019-2021) (y)

	month	y
1	2019-01-01	17272
2	2019-02-01	17818
3	2019-03-01	17714
4	2019-04-01	12695
5	2019-05-01	10900
6	2019-06-01	13351
7	2019-07-01	10485
8	2019-08-01	7316
9	2019-09-01	8097
10	2019-10-01	13169
11	2019-11-01	10400
12	2019-12-01	12785
13	2020-01-01	13856
14	2020-02-01	12196
15	2020-03-01	18626
16	2020-04-01	18123
17	2020-05-01	18626
18	2020-06-01	11419
19	2020-07-01	15012
20	2020-08-01	18583
21	2020-09-01	21200
22	2020-10-01	23600
23	2020-11-01	21640
24	2020-12-01	23546
25	2021-01-01	26980
26	2021-02-01	27224
27	2021-03-01	29655
28	2021-04-01	25095
29	2021-05-01	18626
30	2021-06-01	24111
31	2021-07-01	30184
32	2021-08-01	28017
33	2021-09-01	25729
34	2021-10-01	33296
35	2021-11-01	29780
36	2021-12-01	35300

Table A2: Exogenous Variables Data

	month	car sales industry volumn	repo rate(%)	unempolyment rate(%)	inflation rate(%)	exchange rate(%)	petrol prices(per litre)	steel prices per kg(INR)
1	1/1/2019	17272	6.50	6.9	2.0	7.59	68.65	34.5
2	1/2/2019	17818	6.25	7.2	2.6	6.58	70.94	35.2
3	1/3/2019	17714	6.25	6.7	2.9	5.84	71.81	35.9
4	1/4/2019	12695	6.00	7.4	3.0	2.99	72.86	36.6
5	1/5/2019	10900	6.00	7.0	3.0	3.05	73.13	37.3
6	1/6/2019	13351	5.75	7.9	3.2	3.18	71.62	38.0
7	1/7/2019	10485	5.75	7.3	3.1	3.15	70.44	38.7
8	1/8/2019	7316	5.40	8.2	3.3	3.28	72.80	39.4
9	1/9/2019	8097	5.40	7.2	4.0	3.99	71.77	40.1
10	1/10/2019	13169	5.15	8.1	4.6	4.62	74.61	40.8
11	1/11/2019	10400	5.15	7.2	5.5	5.54	72.86	41.5
12	1/12/2019	12785	5.15	7.6	7.4	7.35	74.91	42.2
13	1/1/2020	13856	5.15	7.2	7.6	4.06	75.14	42.9
14	1/2/2020	12196	5.15	7.8	6.6	5.03	73.19	43.6
15	1/3/2020	18626	4.40	8.8	5.8	5.52	71.49	44.3
16	1/4/2020	18123	4.40	23.5	7.2	7.22	69.59	45.0
17	1/5/2020	18626	4.00	21.7	6.3	6.27	69.59	45.7
18	1/6/2020	11419	4.00	10.2	6.2	6.23	71.26	46.4
19	1/7/2020	15012	4.00	7.4	6.7	6.73	80.43	47.1
20	1/8/2020	18583	4.00	8.4	6.7	6.69	80.43	47.8
21	1/9/2020	21200	4.00	6.7	7.3	7.27	82.08	48.5
22	1/10/2020	23600	4.00	7.0	7.6	7.61	81.06	49.2
23	1/11/2020	21640	4.00	6.5	6.9	6.93	81.06	49.9
24	1/12/2020	23546	4.00	9.1	4.6	4.59	82.34	50.6
25	1/1/2021	26980	4.00	6.5	4.1	6.01	83.71	51.3
26	1/2/2021	27224	4.00	6.9	5.0	6.07	86.30	52.0
27	1/3/2021	29655	4.00	6.5	5.5	6.95	91.17	52.7
28	1/4/2021	25095	4.00	8.0	4.2	4.23	90.56	53.4
29	1/5/2021	18626	4.00	11.9	6.3	6.30	90.40	54.1
30	1/6/2021	24111	4.00	9.2	6.3	6.26	94.49	54.8
31	1/7/2021	30184	4.00	7.0	5.6	5.59	98.81	55.5
32	1/8/2021	28017	4.00	8.3	5.3	5.30	101.84	56.2
33	1/9/2021	25729	4.00	6.9	4.3	4.35	101.34	56.9
34	1/10/2021	33296	4.00	7.7	4.5	4.48	101.89	57.6
35	1/11/2021	29780	4.00	7.0	4.9	4.91	109.69	58.3
36	1/12/2021	35300	4.00	7.9	5.7	5.66	95.41	59.0

b) Code and Algorithms:

- Loading the data

```
> # Load necessary libraries
> library(readr)
> library(dplyr)
> library(ggplot2)
> library(corrplot)
> library(forecast)
> library(tidyverse)
> library(tseries)
> setwd("C:/Users/HP/Desktop")
> #load the dataset
> data = read.csv("Carsales.csv")
```

- Converting Month Column into Date

```
> # Convert 'Month' column to Date format
> str(data)
'data.frame':   36 obs. of  8 variables:
 $ month                  : chr  "1/1/2019" "1/2/2019" "1/3/2019" "1/4/2019" ...
 $ car.sales              : int  17272 17818 17714 12695 10900 13351 10485 7316 8097 13169 ...
 $ repo.rate...           : num  6.5 6.25 6.25 6 6 5.75 5.75 5.4 5.4 5.15 ...
 $ unempolyment.rate...   : num  6.9 7.2 6.7 7.4 7 7.9 7.3 8.2 7.2 8.1 ...
 $ inflation.rate...      : num  2 2.6 2.9 3 3 3.2 3.1 3.3 4 4.6 ...
 $ exchange.rate...       : num  7.59 6.58 5.84 2.99 3.05 3.18 3.15 3.28 3.99 4.62 ...
 $ petrol.prices.per.litre.: num  68.7 70.9 71.8 72.9 73.1 ...
 $ steel.prices.per.kg.INR.: num  34.5 35.2 35.9 36.6 37.3 38 38.7 39.4 40.1 40.8 ...
> data$month = strptime(data$month, format = "%d/%m/%Y")
> data$month = as.Date(data$month)
> str(data)
'data.frame':   36 obs. of  8 variables:
 $ month                  : Date, format: "2019-01-01" "2019-02-01" "2019-03-01" ...
 $ car.sales              : int  17272 17818 17714 12695 10900 13351 10485 7316 8097 13169 ...
 $ repo.rate...           : num  6.5 6.25 6.25 6 6 5.75 5.75 5.4 5.4 5.15 ...
 $ unempolyment.rate...   : num  6.9 7.2 6.7 7.4 7 7.9 7.3 8.2 7.2 8.1 ...
 $ inflation.rate...      : num  2 2.6 2.9 3 3 3.2 3.1 3.3 4 4.6 ...
 $ exchange.rate...       : num  7.59 6.58 5.84 2.99 3.05 3.18 3.15 3.28 3.99 4.62 ...
 $ petrol.prices.per.litre.: num  68.7 70.9 71.8 72.9 73.1 ...
 $ steel.prices.per.kg.INR.: num  34.5 35.2 35.9 36.6 37.3 38 38.7 39.4 40.1 40.8 ...
```

- Overview of the data

```
> # Overview of the data
> head(data)
       month car.sales repo.rate... unempolyment.rate... inflation.rate... exchange.rate...
1 2019-01-01     17272         6.50                  6.9              2.0             7.59
2 2019-02-01     17818         6.25                  7.2              2.6             6.58
3 2019-03-01     17714         6.25                  6.7              2.9             5.84
4 2019-04-01     12695         6.00                  7.4              3.0             2.99
5 2019-05-01     10900         6.00                  7.0              3.0             3.05
6 2019-06-01     13351         5.75                  7.9              3.2             3.18
  petrol.prices.per.litre. steel.prices.per.kg.INR.
1                    68.65                     34.5
2                    70.94                     35.2
3                    71.81                     35.9
4                    72.86                     36.6
5                    73.13                     37.3
6                    71.62                     38.0
```

- Summary Statistics of the data

```
> # Summary statistics
> summary(data)
     month              car.sales      repo.rate...   unempolyment.rate... inflation.rate...
 Min.   :2019-01-01   Min.   : 7316   Min.   :4.000   Min.   : 6.500       Min.   :2.000
 1st Qu.:2019-09-23   1st Qu.:13073   1st Qu.:4.000   1st Qu.: 7.000       1st Qu.:4.075
 Median :2020-06-16   Median :18605   Median :4.000   Median : 7.400       Median :5.400
 Mean   :2020-06-16   Mean   :19512   Mean   :4.662   Mean   : 8.522       Mean   :5.161
 3rd Qu.:2021-03-08   3rd Qu.:25254   3rd Qu.:5.213   3rd Qu.: 8.225       3rd Qu.:6.375
 Max.   :2021-12-01   Max.   :35300   Max.   :6.500   Max.   :23.500       Max.   :7.600
 exchange.rate... petrol.prices.per.litre. steel.prices.per.kg.INR.
 Min.   :2.990    Min.   : 68.65           Min.   :34.50
 1st Qu.:4.447    1st Qu.: 71.80           1st Qu.:40.62
 Median :5.625    Median : 77.78           Median :46.75
 Mean   :5.484    Mean   : 81.38           Mean   :46.75
 3rd Qu.:6.607    3rd Qu.: 90.44           3rd Qu.:52.88
 Max.   :7.610    Max.   :109.69           Max.   :59.00
```

- Descriptive Statistics of the data

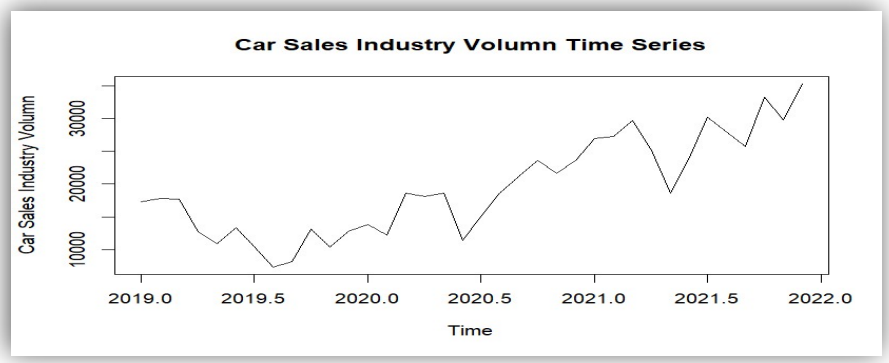
```
> #Descriptive Statistics
> describer::describe(data)
              .column_name .column_class .column_type .count_elements  .mean_value
1                    month          Date       double               36           NA
2                car.sales       integer      integer               36 19511.833333
3             repo.rate...       numeric       double               36     4.662500
4     unempolyment.rate...       numeric       double               36     8.522222
5        inflation.rate...       numeric       double               36     5.161111
6         exchange.rate...       numeric       double               36     5.483889
7 petrol.prices.per.litre.       numeric       double               36    81.379722
8 steel.prices.per.kg.INR.       numeric       double               36    46.750000
     .sd_value  .q0_value .q25_value .q50_value .q75_value .q100_value
1           NA 2019-01-01         NA         NA         NA  2021-12-01
2 7474.8131998       7316 13073.0000  18604.500 25253.5000       35300
3    0.8544317          4     4.0000      4.000     5.2125         6.5
4    3.6394749        6.5     7.0000      7.400     8.2250        23.5
5    1.5986204          2     4.0750      5.400     6.3750         7.6
6    1.3971814       2.99     4.4475      5.625     6.6075        7.61
7   11.5076465      68.65    71.8000     77.785    90.4400      109.69
8    7.3749576       34.5    40.6250     46.750    52.8750          59
```

- Changing Variable Name

```
> #Change variable name
> colnames(data)[2] = "y"
> colnames(data)[3] = "rr"
> colnames(data)[4] = "ur"
> colnames(data)[5] = "ir"
> colnames(data)[6] = "er"
> colnames(data)[7] = "pp"
> colnames(data)[8] = "sp"
```

- Transforming Automobile Sales (y) into Time-Series Object

```
> #Transform Time-series object
> y = ts(data = data$y,
+        start = c(2019, 1),
+        frequency = 12)
> plot.ts(y)
```



- Outliers Detection of Automobile Sales (y)

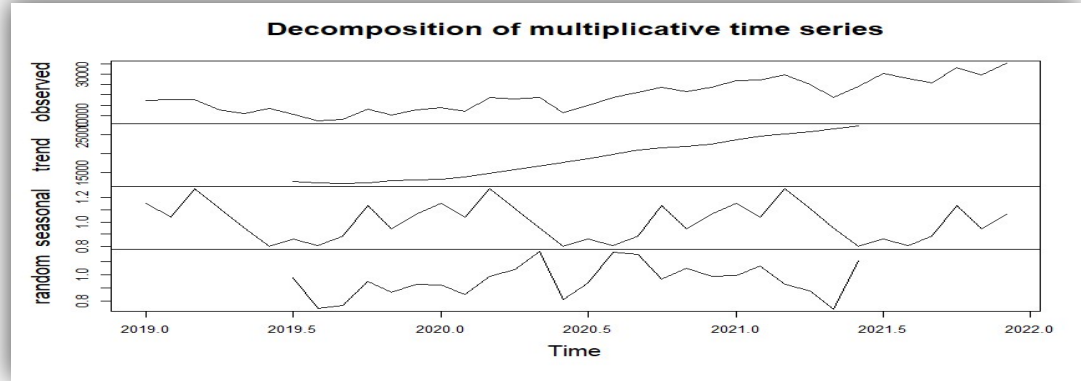
```
> # Automatic detection of outliers for y
> myts = tsoutliers(y)
> myts
$index
integer(0)

$replacements
integer(0)
```

- Checking Stationarity of Automobile Sales (y)

```
> #Checking whether y is Stationary
> result = adf.test(y)
> result

        Augmented Dickey-Fuller Test

data:  y
Dickey-Fuller = -3.0303, Lag order = 3, p-value = 0.1723
alternative hypothesis: stationary

> #Using First Order Differencing on y
> y = diff(y)
> #Rechecking Stationarity on y
> s_differenced_y = adf.test(y)
> s_differenced_y

        Augmented Dickey-Fuller Test

data:  y
Dickey-Fuller = -3.768, Lag order = 3, p-value = 0.03464
alternative hypothesis: stationary

> # Check the p-value to determine if the data is now stationary
> if (s_differenced_y$p.value <= 0.05) {
+   print("y is stationary after first order differencing.")
+ } else {
+   print("y is still not stationary after first order differencing.")
+ }
[1] "y is stationary after first order differencing."
```

- Decomposition of Automobile Sales

```
> #Multiplicative Decomposition
> decomposition_multiplicative = decompose(x = y,
+                                          type = "multiplicative")
> plot(decomposition_multiplicative)
```

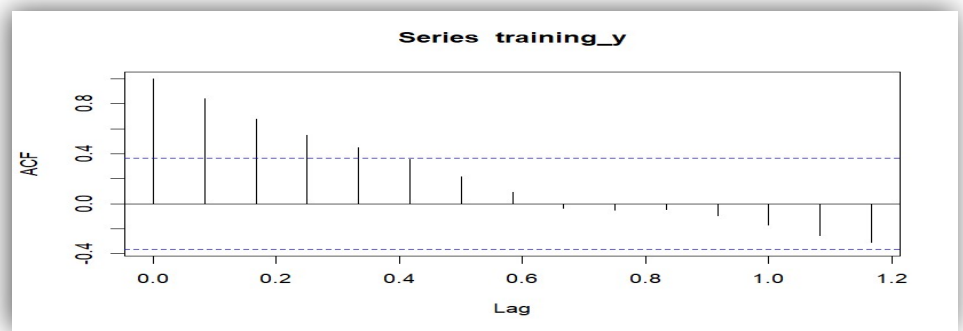


- Training and Test Set

```
> #Training and Test Set
> training = data %>% filter(month < '2021-06-01')
> test = data %>% filter(month >= '2021-06-01')
> #Time series object
> #if daily, then 7 or 365, 12 if monthly, 4 if quarterly
> training_y = ts(data = training$y,
+                 frequency = 12)
```

- Autocorrelation Plot

```
> #Auto-correlation plot
> acf(training_y)
```

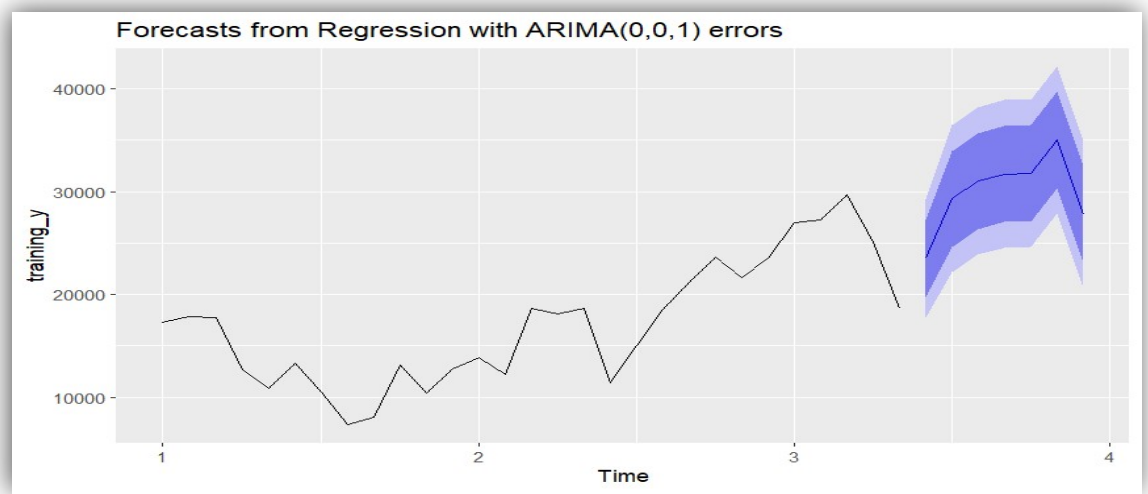


- Getting the exogenous regressors

```
> #Get the regressors
> training_reg = as.matrix(training[,3:8])
> test_reg = as.matrix(test[,3:8])
```

- SARIMAX Model

```
> #SARIMAX model
> model = auto.arima(y = training_y,
+                    stepwise = FALSE,
+                    approximation = FALSE,
+                    xreg = training_reg)
> summary(model)
Series: training_y
Regression with ARIMA(0,0,1) errors

Coefficients:
          ma1           rr         ur           ir         er         pp         sp
      0.7487  -3226.3972   26.1479  -1752.9545  859.1738  456.5787   34.6678
s.e.  0.2397    791.7123  223.4709    606.7225  538.1459  191.7818  306.1480

sigma^2 = 8539526:  log likelihood = -268.98
AIC=553.96   AICc=561.16   BIC=564.89

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set -9.998221 2545.243 2109.883 -3.555385 14.95474 0.2706652 0.1366143
> #Forecasting
> predictions_sarimax = forecast(model, xreg = test_reg)
> predictions_sarimax
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jun 3       23509.39 19764.38 27254.40 17781.90 29236.89
Jul 3       29302.29 24624.05 33980.53 22147.54 36457.04
Aug 3       31020.71 26342.47 35698.95 23865.96 38175.46
Sep 3       31716.82 27038.58 36395.06 24562.07 38871.57
Oct 3       31774.22 27095.98 36452.46 24619.47 38928.97
Nov 3       35009.76 30331.52 39688.00 27855.01 42164.51
Dec 3       27779.64 23101.40 32457.88 20624.89 34934.39
> #plotting
> autoplot(predictions_sarimax)
> #accuracy
> accuracy(predictions_sarimax$mean, test$y)
                  ME     RMSE      MAE       MPE     MAPE
Test set -527.9753 4346.388 3535.249 -2.894872 11.83512
```

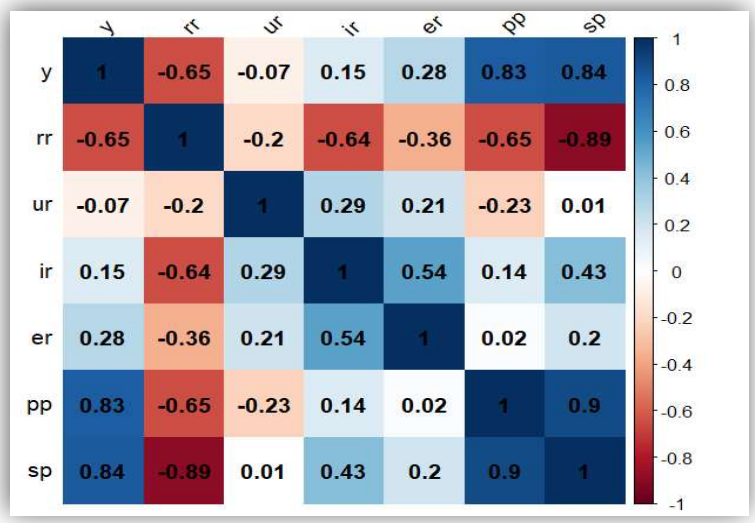


- SARIMA Model

```
> #SARIMA model
> sarima_model <- auto.arima(y)
> summary(sarima_model)
Series: y
ARIMA(0,1,0)

sigma^2 = 14100235:  log likelihood = -337.74
AIC=677.49   AICc=677.61   BIC=679.04

Training set error measures:
                    ME      RMSE       MAE         MPE      MAPE      MASE        ACF1
Training set 501.2576 3702.508 3059.702 -0.3793436 17.60728 0.3600797 -0.1831731
> #Forecasting
> predictions_sarimax = forecast(sarima_model)
> #plotting
> autoplot(predictions_sarimax)
> #accuracy
> accuracy(predictions_sarimax$mean, test$y)
                 ME      RMSE       MAE        MPE      MAPE
Test set -5811.857 6871.262 5811.857 -21.58635 21.58635
```

- Correlation Plot

```
> # Correlation Heatmap
> correlation_matrix = cor(data[, -1]) # Excluding 'Month' column for correlation computation
> corrplot(correlation_matrix, method = "color", tl.col = "black", tl.srt = 45, addCoef.col = 'black')
```



- Regression Model

```
> # Multiple Linear Regression Model
> model = lm(y ~ rr+ur+ir+er+pp+sp, data = data)
> summary(model)

Call:
lm(formula = y ~ rr + ur + ir + er + pp + sp, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7818.0 -2054.9   360.3  2511.8  4188.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -58071.46   22777.40  -2.550 0.016336 *
rr            3489.30    2509.69   1.390 0.175008
ur              13.78     178.31   0.077 0.938934
ir           -1629.56     540.56  -3.015 0.005302 **
er            1848.14     467.18   3.956 0.000451 ***
pp             -51.59     196.60  -0.262 0.794851
sp            1361.95     487.22   2.795 0.009100 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3235 on 29 degrees of freedom
Multiple R-squared:  0.8448,    Adjusted R-squared:  0.8127
F-statistic: 26.31 on 6 and 29 DF,  p-value: 1.733e-10
```