Minor 1
Final Report

For

Time series data analysis and forecasting for Supermarkets

Prepared by

Specialization	SAP ID	Name
BAO	500086235	Akshaj Agarwal
BAO	500083130	Aryan Thakur
BAO	500086928	Samriddh Goyal
BAO	500083263	Varnit Tomar

Department of Informatics
School Of Computer Science
UNIVERSITY OF PETROLEUM & ENERGY STUDIES,
DEHRADUN- 248007. Uttarakhand

Table of Contents

Acknowledgement

We would like to convey our sincere gratitude to our mentor, Dr. Bishwajeet Roy, for all the guidance, inspiration, and unwavering support he provided us with throughout the course of working on our project. Without his encouragement and helpful recommendations, this effort would not have been feasible.

We really appreciate Dr. TP Singh, the cluster chair, for all of his help with our study at SOCS.

Additionally, we are grateful to our kind Dean of SOCS, UPES Dr. Ravi S. Iyer for providing us with the tools we needed to complete our project work effectively.

We appreciate the assistance and helpful feedback provided by all UPES faculty members during the course of our project work. Finally, we can only thank our parents in the deepest sense for showing us the world and for all of the support they have given us.

Abstract

Forecasting is useful in various issue domains where forecasts are needed for exploitable business decisions or operational efficiency. A time sequence Because of their temporal character, datasets are easily employed for prediction and analysis. The supermarket sales prediction aids in the improvement of sales in a corporate setting. Predictions can be useful in making judgments about what to stock in the business and when to schedule promotions.Time series analysis is a method of analysing a set of data points collected over a period of time. Analysts use time series analysis to record data points at periodic intervals over a predetermined length of time rather than randomly. This method allows analysts to evaluate data points as a function of time, which can provide useful insights into patterns and correlations that may not be obvious when viewing data points in isolation. However, time series analysis can be complicated, and there are a variety of approaches that analysts might take. ARIMA is one of the time series forecasting approaches; this model has a significant impact on predicting future sales and managing enterprises. This model aids in improving your company's future predictions. Forecasting strategies have been increasingly popular in the business world in recent years. Forecasting is critical in making decisions and optimising corporate operations. In this study, we will look at how ARIMA time series forecasting is used in supermarket sales.

1. Introduction

1.1 Purpose of the project

The main purpose behind this project is to help the supermarkets to increase their sales and maximize their profit by knowing how much to stock for the upcoming season, and which items to mainly focus on. This all is possible using the past time series data of the stores, which reveals the buying behavior of the consumers.The time-series dataset comprises information about time that is useful for statistical analysis and forecasting. The supermarket sales forecast assists in improving sales in a professional context. The technique aids in problem-domain decision-making.

1.2 Target Beneficiary

The small supermarket owners are the main target beneficiaries of this project, it will help them to increase their sales and profit. Most of the small supermarket owners run their businesses without any prior knowledge about which items to stock which at times leads them to a severe loss. Supermarkets play an important role in the daily lives of many consumers. For this reason, it is important to understand how changes in the supermarket industry can affect consumers.

1.3 Project Scope

Time series analysis is an effective approach for forecasting future sales and trends in supermarkets. Time series research can help supermarkets identify prospective areas of sales growth or decline, as well as predict how changes in customer behavior would affect future sales. Time series analysis, for example, can be used to uncover patterns in client purchase habits that may be indicative of future trends. Furthermore, time series analysis can be used to forecast how economic events, such as inflation or recession, will affect future sales.

2. Literature Review

Many research works have been carried out for time series forecasting using the ARIMA model.

Renato Cesar Sato has proposed a study on disease management using the ARIMA model but hasn't used any mathematical details. To understand the trend in the curve we need to use certain mathematical tools also.[1]

Debadrita Banerjee analyzed the Indian stock market performance using the ARIMA model for six years with respect to time, he computed the Durbin – Watson value to check if the data had a positive or negative correlation. He also did statistical computations. [2]

DurduÖmer Faruk analyzed that the ARIMA model cannot deal with non-linear relationships, so he proposed a hybrid ARIMA model which was capable of exploiting the strengths of traditional time series approaches. This was tested using 108 months of observation of water quality data, water temperature, dissolved oxygen, etc. This hybrid model is capable of capturing the non-linear nature of complex time series data.[3]

3. Problem Statement

Grocery stores are constantly looking for innovative ways to boost sales and profitability. They have sales data, but they aren't always sure how to use it effectively. Supermarket owners must deal with the issue of stocking and selling items. They must figure out how to use their data to help them grow in the future.
As a result, a growing number of supermarket operators are turning to data analytics. Time series data analytics can assist supermarket owners in better understanding their customers and what they want. Supermarket owners can use this data to make judgments about what to stock and how to sell things. This can also assist supermarket operators in determining how to expand in the future.

4. Project Description

4.1 Reference Algorithm

The model used in this project is a statistical model known as ARIMA(Autoregressive integrated moving average). An ARIMA model is a conventional statistical time series model that may be used to examine data. The model can be used to forecast future events and to analyze the relationships between different variables. The model is a linear model, which means it can predict future values of a variable but cannot explain the relationships between variables. Lagged moving averages are used in ARIMA to smooth time series data. The implicit premise of autoregressive models is that the future will resemble the past..[3]

4.2 Characteristic of Data

The secondary dataset, which has the characteristics of a temporal dataset, is collected from Kaggle.

- Data is saved on a daily basis.
- Number of transactions that take place on a daily basis.
- Transactions from various stores merged on the same day.

	A	B	C
1	date	store_nbr	transactions
2	01-01-2013	25	770
3	02-01-2013	1	2111
4	02-01-2013	2	2358
5	02-01-2013	3	3487
6	02-01-2013	4	1922
7	02-01-2013	5	1903
8	02-01-2013	6	2143
9	02-01-2013	7	1874
10	02-01-2013	8	3250

	A	B	C
1	date	transactions	
2	01-01-2013	770	
3	02-01-2013	93215	
4	03-01-2013	78504	
5	04-01-2013	78494	
6	05-01-2013	93573	
7	06-01-2013	90464	
8	07-01-2013	75597	
9	08-01-2013	72325	
10	09-01-2013	71971	

Fig1. Dataset

Statistical techniques were utilized to preprocess the dataset:-

- Data Cleaning
- Data Reduction
- Data Transformation
- Moving Average
- Rolling Mean
- Augmented Dickey-Fuller Test(To check stationarity)
- Log Transform

4.3 SWOT

Strengths

- Choosing Time Series Forecasting over other Forecasting Techniques for Time Series Data (Years, Days, Hours, etc.) provides advantages since we employ several models such as Moving Average, Weighted MA, Exponential Smoothing, HoltWinter, AR, ARMA, ARIMA, SARIMA.
- The ARIMA Model, which is used for Time Series Forecasting, can construct a model using Non-Stationary Datasets. Although our Dataset is 'Non-Stationary,' we may still use ARIMA to build a model on it.
- Dealing with Temporal Data, because the effects of time passing are significant, and Time Series Analysis is the best technique to deal with temporal data.

Weaknesses

- Accuracy of Predictions Using Time Series Long-term forecasting is unreliable because unexpected events are tough to predict and a variety of real-world factors affect the data.
- Every model won't work with every set of data or provide an answer to every question. Time series forecasting should be used when a data team is aware of a business concern and has the data and forecasting tools needed to address it.
- It is not always acceptable or useful to use time series forecasting, nor is it always appropriate or effective. Since there are no fixed rules for when to utilise forecasting and when to avoid it, it is up to analysts and data teams to be aware of these limitations and what their models can support

.

Opportunities

- Time series analysis allows you to clean your
- Forecasting is used in a variety of sectors. It has many practical uses, including weather forecasting, climate forecasting, economic forecasting, health forecasting, engineering forecasting, financial forecasting, retail forecasting, business forecasting, environmental studies forecasting, social studies forecasting, and more.
- A skilled forecaster can spot actual trends and patterns in past data by using clean, time-stamped data. Real insights can be distinguished from seasonal variations and random fluctuations, or "outliers," by analysts. A good forecast can identify the direction in which the data is changing by using time series analysis, which reveals how the data changes over time.

Threats

- The potential that the dependent variable was impacted by factors other than the study treatment at the same time the intervention was implemented.
- Dependence on the forecasts generated by various time series forecasting models is risky since errors might occur when using the predictions to make important decisions.
- Due to the large creation of time series data, analysis and forecasting of this data are becoming increasingly crucial. As continuous monitoring and collecting of these data become more widespread, the demand for more effective analysis and forecasting will only grow.
- It's likely that an analyst could surpass these algorithms given expertise with the subject for businesses that have significant volumes of historical data where more general heuristics and "laws" of their data could be developed.

4.4 Project Features

- Data Preprocessing - Making the data more relevant for our model.
- Rolling Window analysis - Calculate the rolling mean of the dataset, which will be useful for the moving average function.
- Data scaling
- Seasonal decompose - Checks for the trend followed by the dataset, any seasonality the data is giving.
- Autocorrelation analysis - ACF is used to determine which lags have sufficient correlations and better understand the pattern and properties of the time series.
- Test and train - Training our model to find the efficiency of some test data.
- Implementation - ARIMA model is implemented

- Forecast - Final results received are forecasted using some of the visualization techniques to search for the future trends possible.

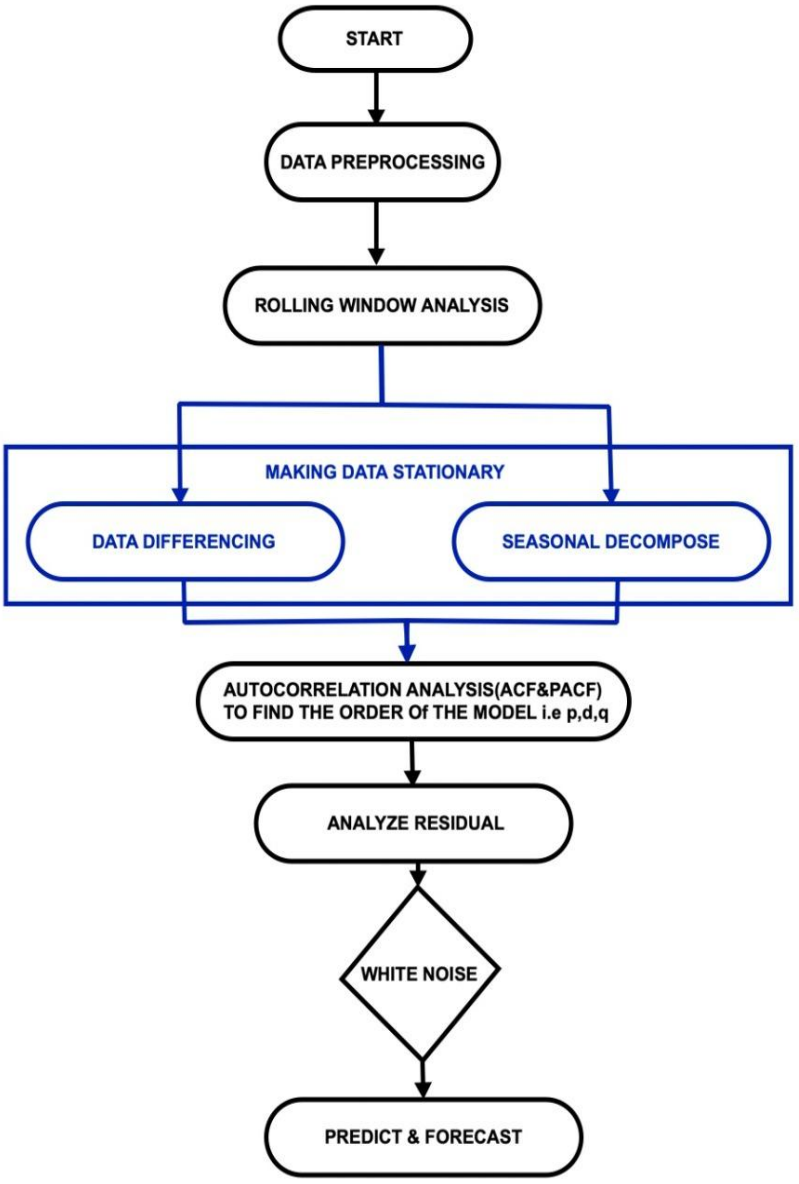


Fig 2. Level 0 Data flow diagram for ARIMA

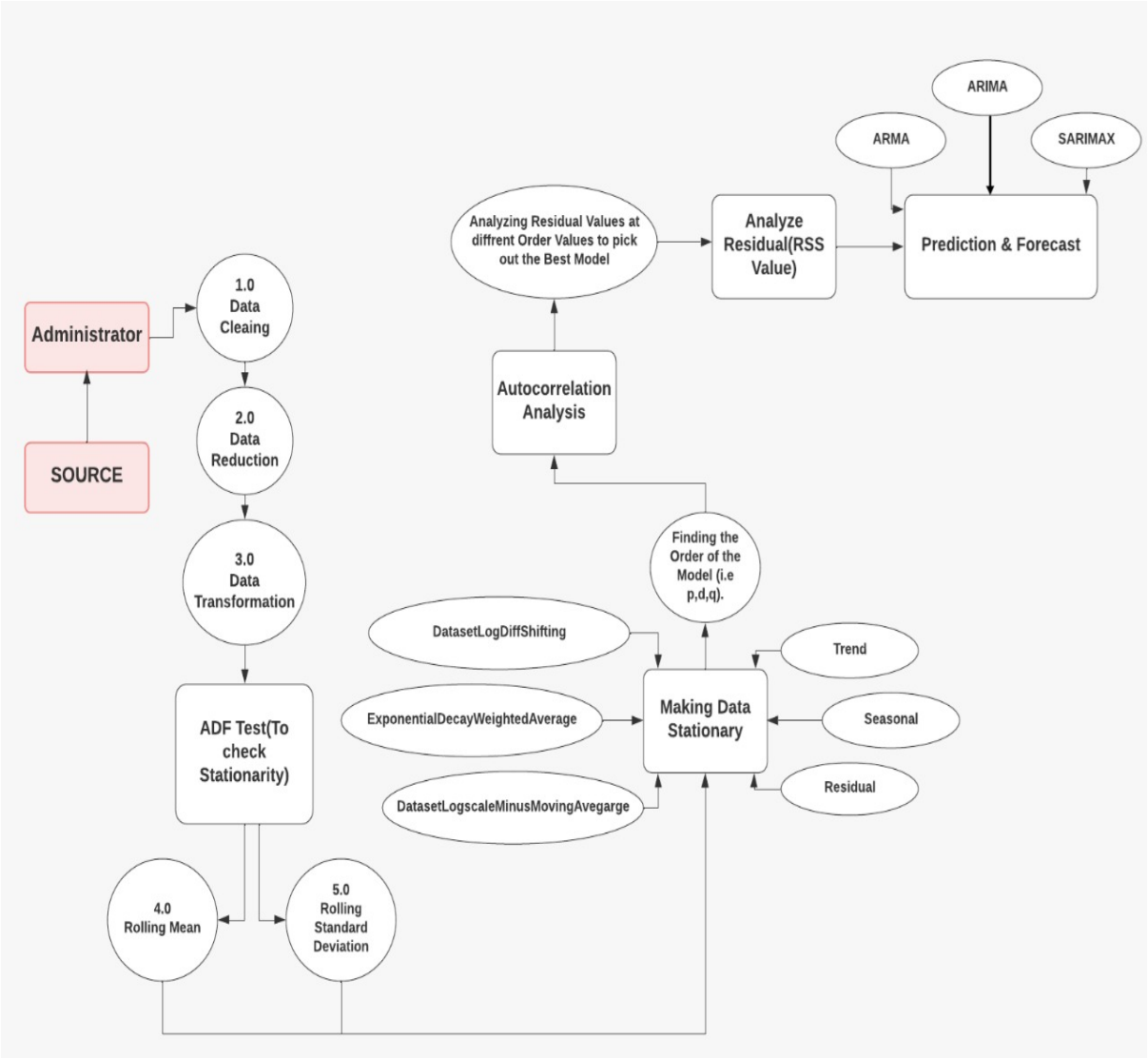


Fig 3. Level 1 Data flow diagram for ARIMA

First process is data preprocessing, in this process, the data is modified to make the dataset relevant to our model. Pre-processed data is then pushed forward for rolling window analysis where the rolling mean is calculated for the main attribute which is useful for the moving average function. For further process, data needs to be stationary which is done using the data differencing method. Seasonal decomposition is used to check for the seasonality in the dataset. To find future trends ACF is used, which makes a better understanding of the trend followed by our dataset. After all this process a test dataset is used to train the model and check for efficiency. Finally, the ARIMA model is implemented to make the final plot using some of the visualization techniques which gives us future predictions.

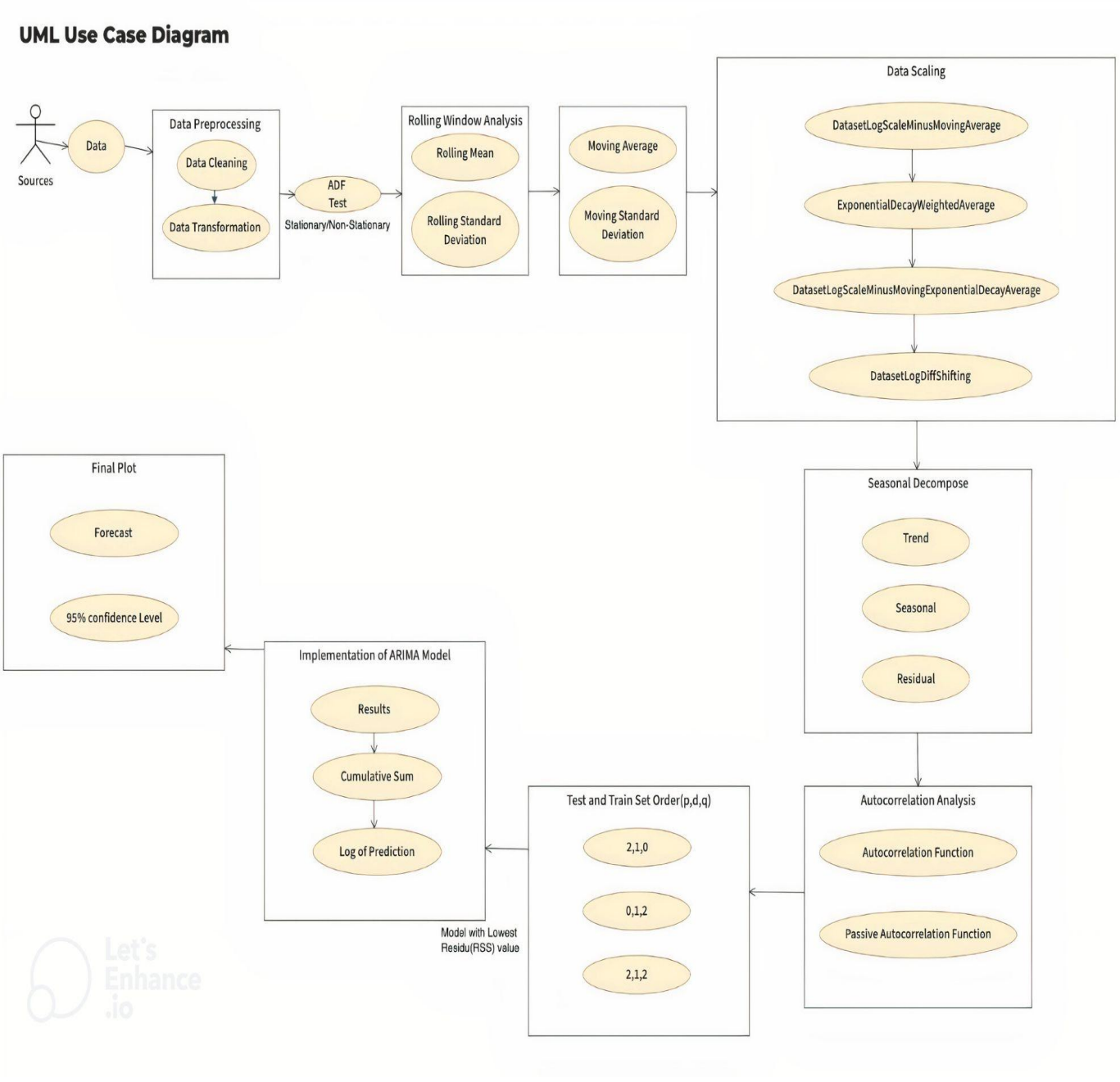


Fig4. Use case diagram for ARIMA

4.5 User Classes and Characteristics

One industry that can benefit from time series analysis is supermarkets. Analysts use ARIMA to model data like sales statistics or the number of customer visits in order to optimize stock and better forecast consumer behavior. A supermarket's recommended order volume from suppliers is one example of a prediction that may be made using time series analysis. Supermarkets may avoid being trapped with too much or too little of a commodity by comprehending historical trends. It is crucial to take into account all pertinent variables in order to model time series data effectively. Seasonality and autocorrelation fall under this.

5. PERT Chart

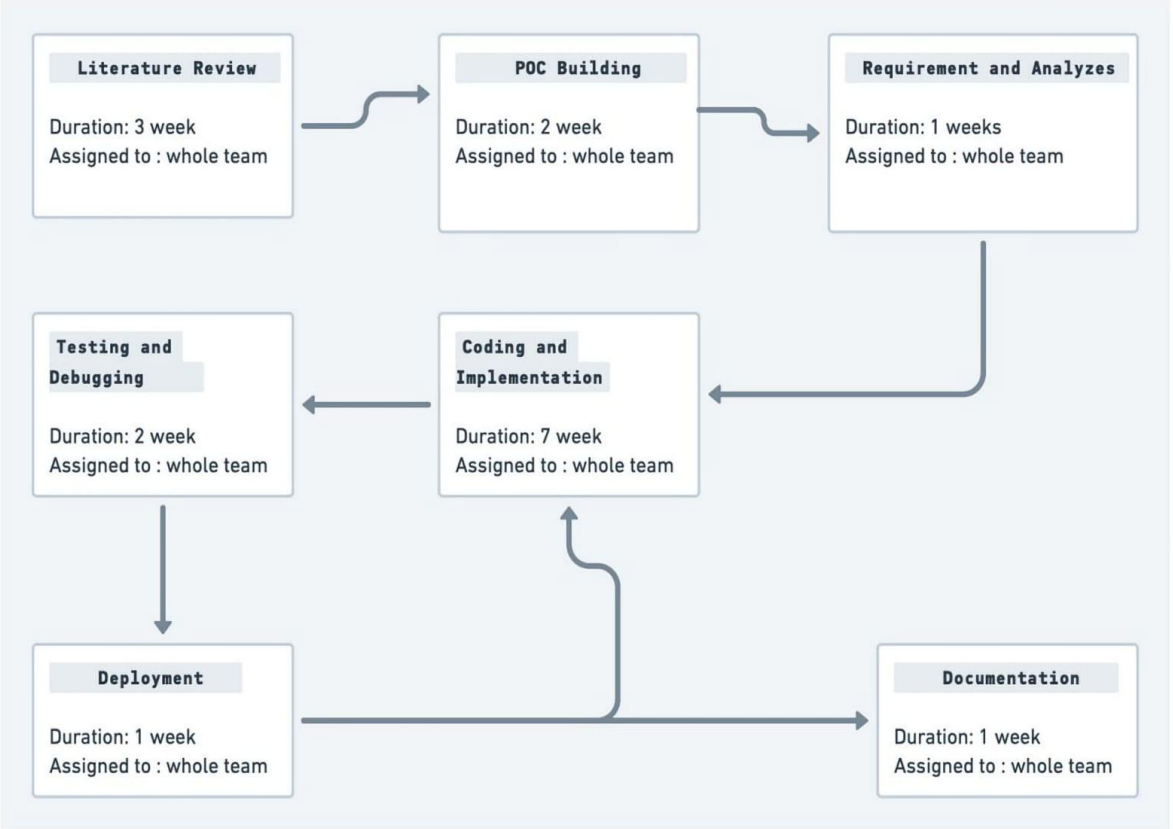


Fig5. Program Evaluation and Review Technique

6. Results

We analyzed several models, the AR(auto regressive) model, MA(moving Average) model and the ARIMA model and wondered which of these will be the best fit for our dataset. For this we checked for their RSS value(Residual sum of squares) which tells us about how much variance in the error terms will be there and the closer its value is to zero, the best fit your model will be. We got the lowest value for the ARIMA model itself, which made us to work ahead with this.

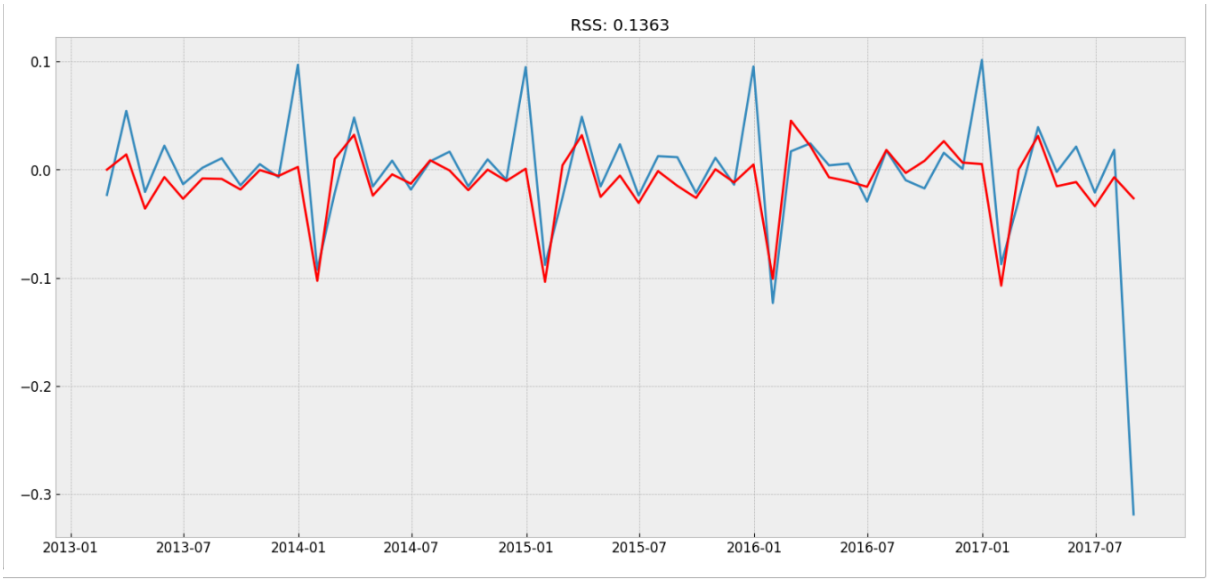


Fig6. Model Evaluation

Here we can see if we would have used the SARIMA model on the previous data then what results we would have got, it shows us the decline in the sales. Here the Root mean square value (RMSE) for our model is 0.0640 which shows that our model is good enough for the future predictions.

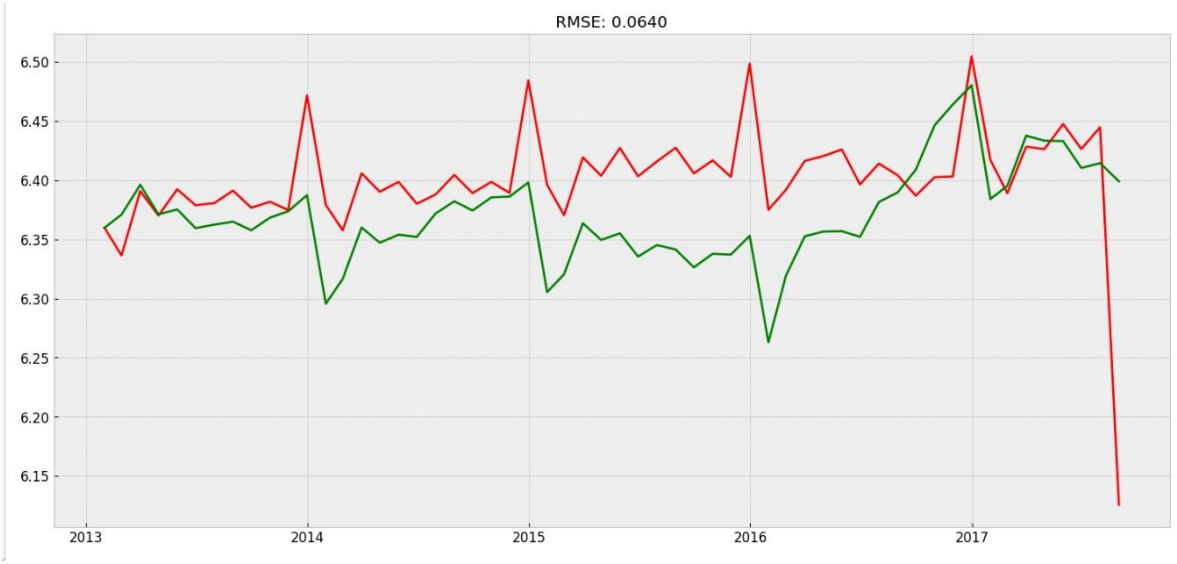


Fig6. Predictions

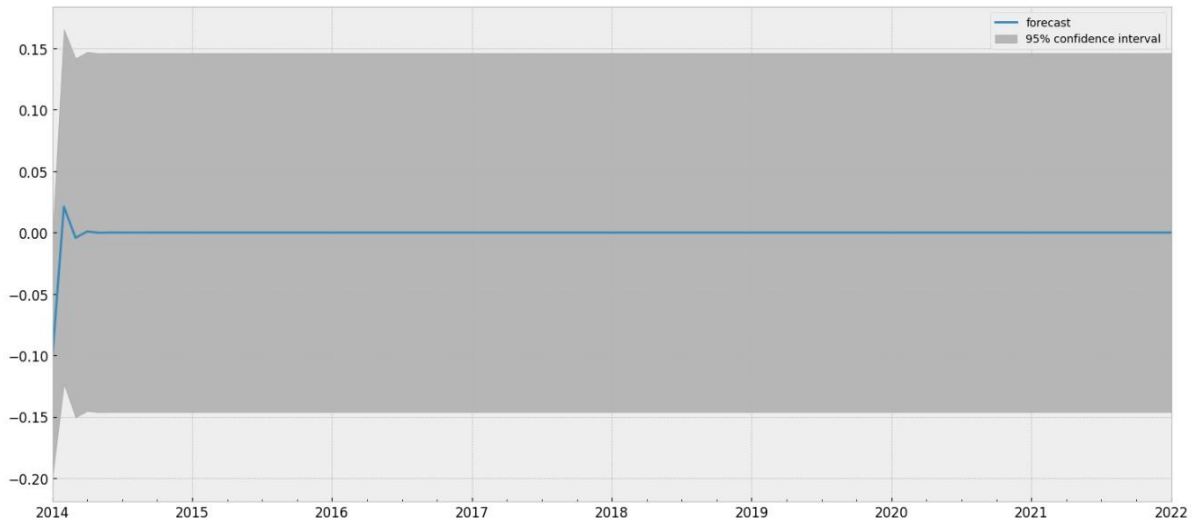


Fig7. Forecaste

7. Conclusion

We profitably studied a supermarket's time series data collection, which we analyzed using the ARIMA model. This has given us an idea of what future projections we can make based on this data. According to the forecasts, there is a chance that demand for the supermarket's items would rise in the near future. It is crucial to remember, however, that the data is not flawless. There is always the possibility of mistakes, thus the ARIMA model's projections should be taken with a grain of salt. Having said that, we believe our analysis is an excellent starting point for future research. With more data and model modification, it may be able to provide even more accurate estimates.

References

[1]Time Series Forecasting Model for Supermarket Sales using FB-Prophet by Bineet Kumar Jha and Shipla Pande
URL - https://ieeexplore.ieee.org/abstract/document/9418033

[2]Anomaly Detection with Time Series Forecasting by Adithya Krishnan
URL-https://towardsdatascience.com/anomaly-detection-with-time-series-forecasting-c34c6d04b24a

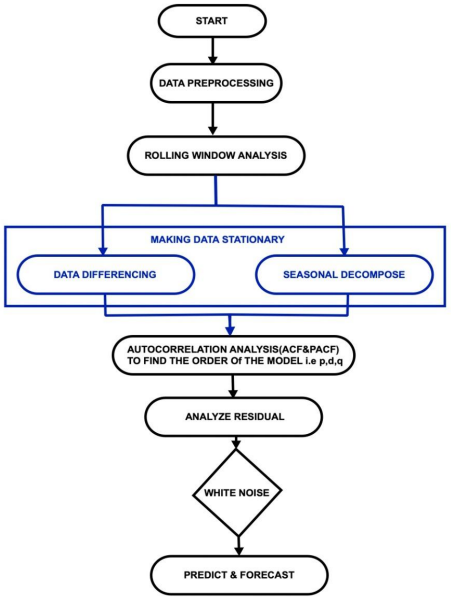
[3]Forecasting of Indian stock market using time-series ARIMA model by Debadrita Banerjee
URL - https://ieeexplore.ieee.org/abstract/document/6970973

Appendix A: Glossary

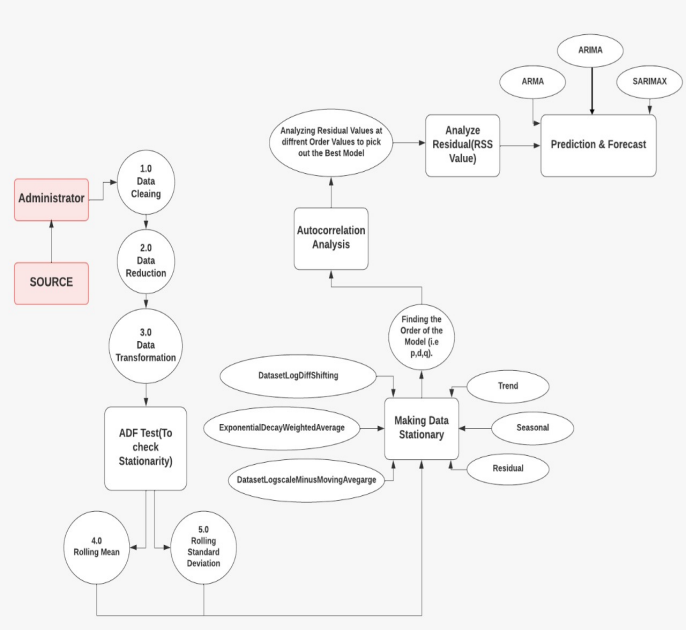
- ARIMA - Autoregressive Integrated Moving Average
- SARIMA - Seasonal Autoregressive Integrated Moving Average
- UML - Unified Modeling Language
- ACF - Autocorrelation Function

Appendix B: Analysis Model

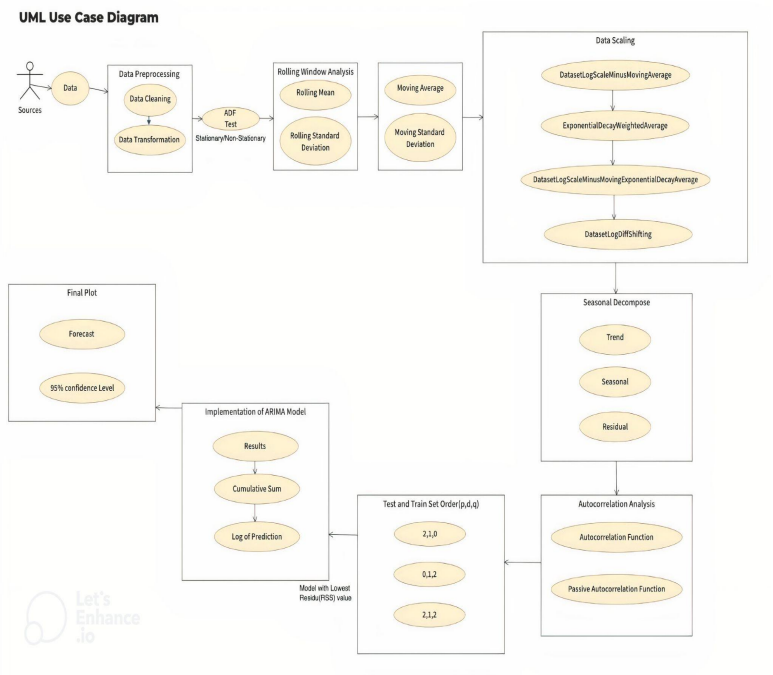
- Level 0 Data Flow Diagram(Fig2)



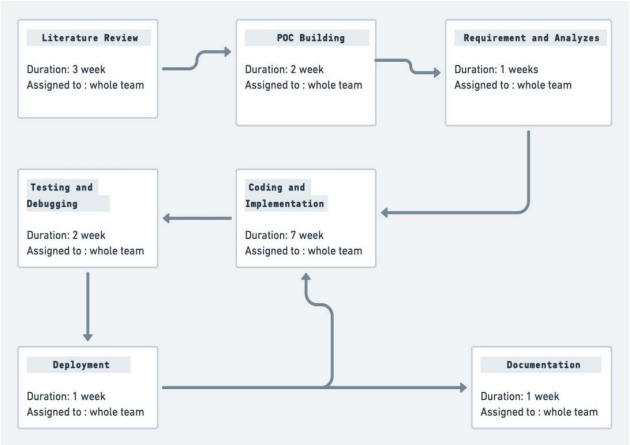
- Level 1 Data flow diagram(Fig3)



- Use case diagram(fig4)



- PERT Chart



Appendix C: Issues List

- Data has to be stationary
- Regression should be there in data