INTERNSHIP: INTERIM PROJECT REPORT

Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

Internship Project Title	Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical Data
Name of the Company	ICT Academy of Kerala
Name of the Industry Mentor	Debashis Roy
Name of the Institute	ICT Academy of Kerala

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
30th April 2023	15rd July 2023	125	Jupyter Notebook, VSCode	Python Pakages: matplotlib, Pandas, Seaborn. MS Excel

---

TABLE OF CONTENT

Acknowledgements

I, Govind S, would like to express my sincere gratitude to my industry mentor, Debashis Roy, and TCS ioN for providing me with all the necessary resources and guidance throughout the completion of this project. I am grateful for their support and encouragement. I would also like to thank my colleagues for their knowledge sharing and feedback on my work. Additionally, I extend my appreciation to Kaggle user Pavan Kumar D for providing the dataset used in this project.

Objective

The objective of this internship project is to create a forecast model for predicting the demand of products in a retail store. The project involves applying concepts of time series forecasting, quantitative forecasting methods, and various forecasting models such as naive approach, ARIMA, auto-regressor, and moving average approaches. The aim is to develop accurate forecasting models that can assist in inventory management and meet customer needs.

Introduction

In today's fast-paced service industry, accurately predicting customer demand is crucial for effective inventory management. This internship project focuses on developing a forecasting model that can predict the demand of products in a retail store based on historical data from 2012-2014. By understanding and preprocessing the dataset, suitable forecasting models can be selected and trained to make predictions. The accuracy of these predictions can be validated using visualization techniques.

The dataset used in this project was obtained from the open-source Kaggle platform. It consists of data from a large Indian retail chain with stores in Maharashtra, Telangana, and Kerala. The dataset includes information such as product identifiers, department identifiers, categories of products, store identifiers, states, and sales figures. Additional auxiliary datasets, such as product prices and date-to-week ID mapping, were also utilized.

Internship Activities

- Familiarization with the project requirements and day-wise plan.
- Passing the RIO Pre-Assessment.
- Introduction in the DDR among peer groups.
- Learning about forecasting methods through endorsed YouTube videos.
- Data mining and extraction from open data sources, with the final dataset obtained from Kaggle.
- Working with the dataset in the VS code IDE.
- Importing the dataset using necessary Python libraries.
- Data cleaning and preprocessing, ensuring there are no missing values and converting data types as required.
- Exploratory data analysis (EDA) through various plots and visualization techniques.
- Feature engineering to include only the relevant attributes for time series analysis and forecasting.
- Grouping the data based on product categories and performing data analysis separately for each category.
- Decomposition of sales data to analyze seasonality and stationarity using seasonal decomposition plots.
- Checking data stationarity using the augmented Dickey-Fuller test.
- Applying differencing techniques to make non-stationary data stationary.
- Selecting appropriate forecasting models such as ARIMA and SARIMAX based on auto-correlation and partial correlation plots.
- Training the models and making predictions for future sales.
- Used auto ARIMA function to obtain the optimal set of parameters and values.
- Splitting the dataset into training and testing sets to evaluate the performance of different models with varying parameters.
- Trained the model on the entire dataset to predict the sales value of the future month.

Challenges & Opportunities

The most significant obstacle encountered was the extensive time and effort needed to discover an appropriate model and optimize it for optimal performance. However, this presented an opportunity to explore a wide range of forecasting techniques and understand the diverse areas where these models find extensive application.

Methodology

Source of data: The dataset was downloaded from kaggle.

About Dataset:

Anticipating the product demand in multiple stores of a retail chain

An expansive retail chain in India operates stores across three states: Maharashtra, Telangana, and Kerala. These stores maintain a wide range of products, including FMCG (fast moving consumer goods), perishables, and other categories. Effective inventory management is vital for ensuring a steady revenue stream for the retail chain. Meeting customer demand is crucial to avoid potential revenue loss, while excessive stocking of products can result in financial losses.

In this scenario, your objective is to develop a machine learning model that predicts product sales across the stores for a one-month period. These models will be utilized to enhance the inventory management software's recommendations at these stores.

The datasets are provided as cited below:

Columns in the dataset: "sales" (continuous)

train_data.csv:

● date : The date for which the observation was recorded

● product_identifier : The id for a product

● department_identifier : The id for a specific department in a store

● category_of_product : The category to which a product belongs

● outlet : The id for a store

● state : The name of the state

● sales : The number of sales for the product

Auxiliary Datasets:

● product_prices.csv : The prices of products at each store for each week

● date_to_week_id_map.csv : The mapping from a date to the week_id

● sample_submission.csv :The format for submissions

● The test_data.csv file has all the attributes of the train_data.csv file excluding the sales (target)column

Data Analysis: The process of data analysis consists of multiple stages, such as defining data needs, gathering and preparing the data, performing data cleaning and processing, analyzing the data, and finally communicating and documenting the findings.

In this particular project, we imported the data into the Python environment and gathered essential information about its structure and content. Subsequently, we conducted data pre-processing tasks, which involved examining the presence of null values, missing data, and outliers. Fortunately, we discovered that the dataset contained no null values, which enabled us to proceed with further analysis steps.

Categorizing the data: The data is classified into three data frames based on the category of products by this way forecasting can be done on each product category.

Analyzing the data: Each data frame is separately analyzed for better understanding of each product category which gave an idea of the purchase trend of the product.

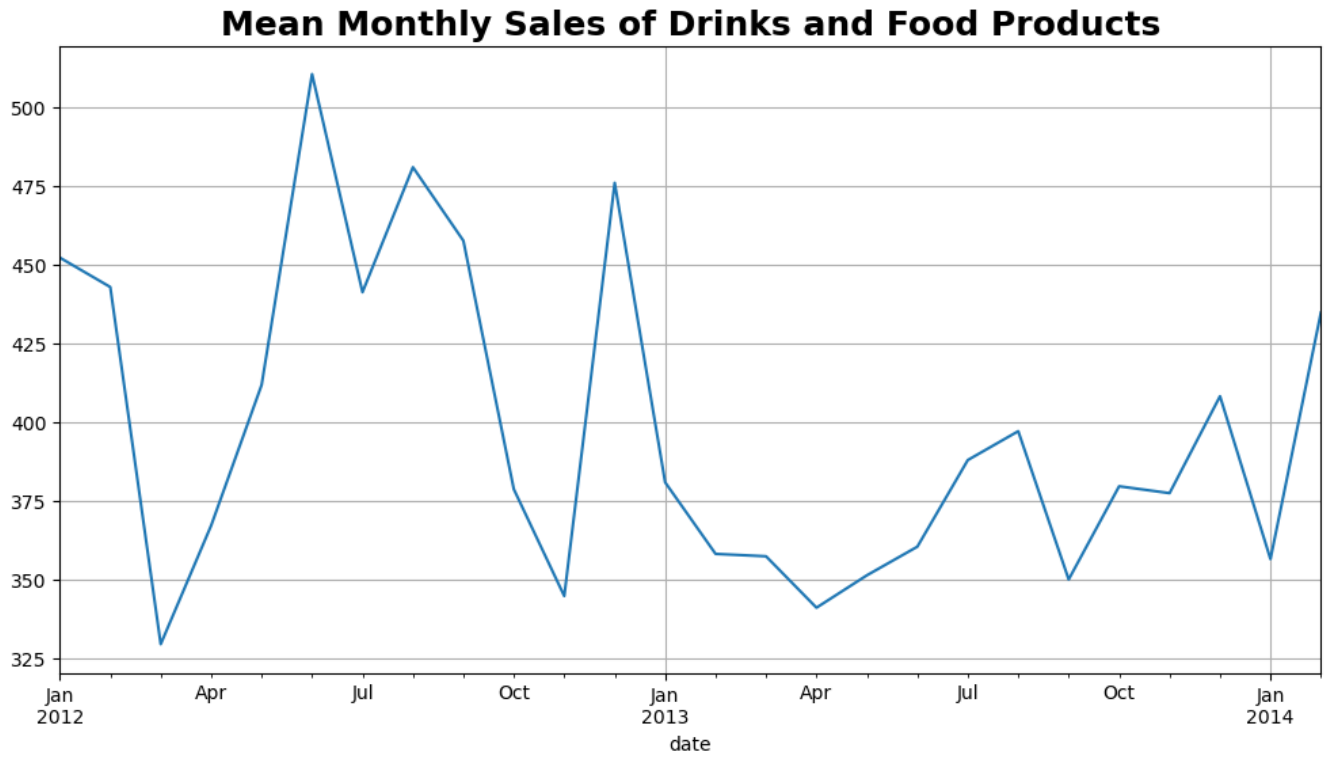

Fig 1

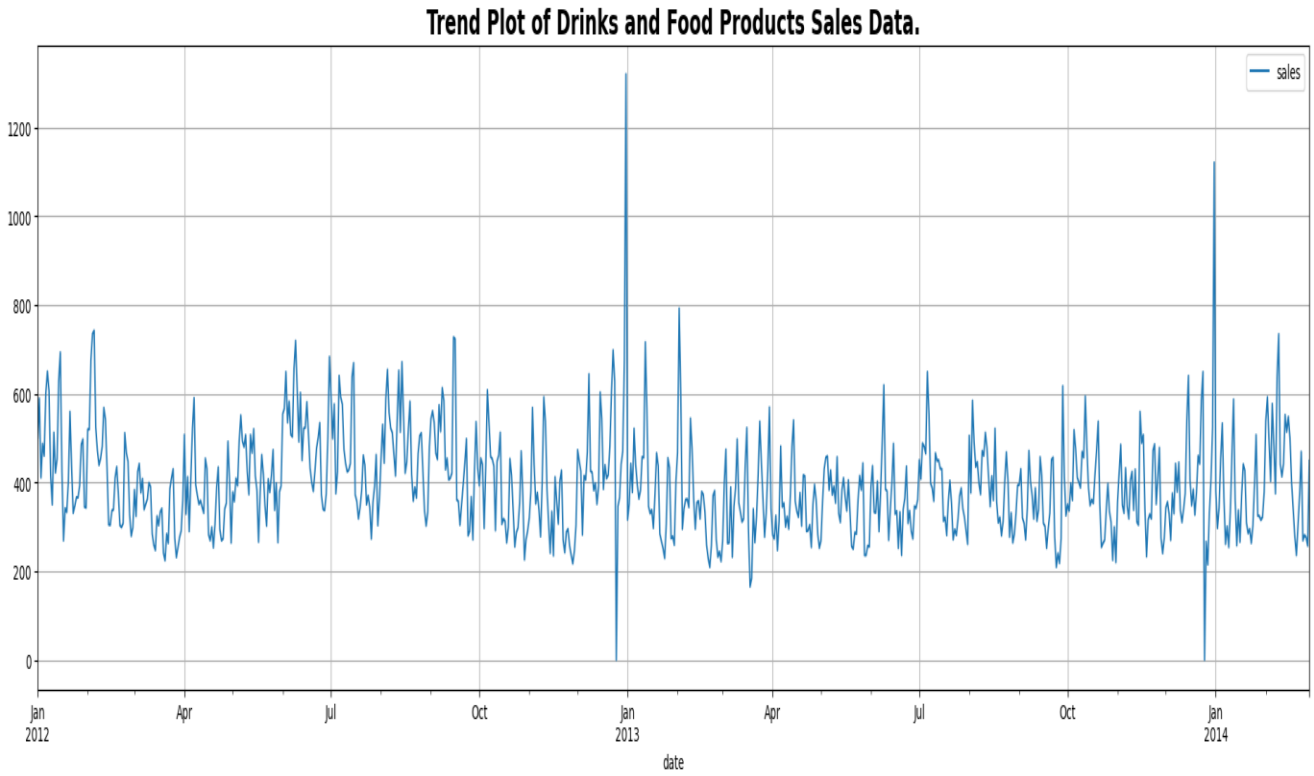

Fig 2

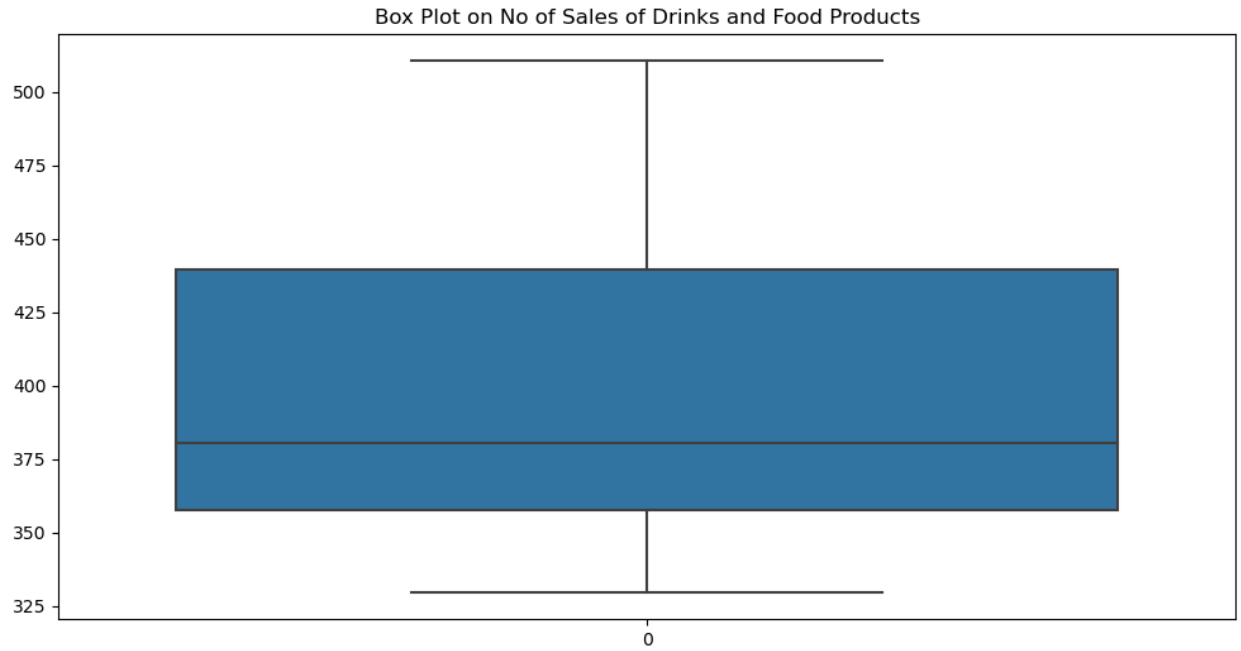

Fig 3

Similarly, data analysis is done for remaining two product categories.

Checking for seasonality: All the three category of data was checked for seasonality and some of them turned out to be seasonal which is observed from the seasonal decomposition plot, a rough idea of the sales can also be identified from this plot.

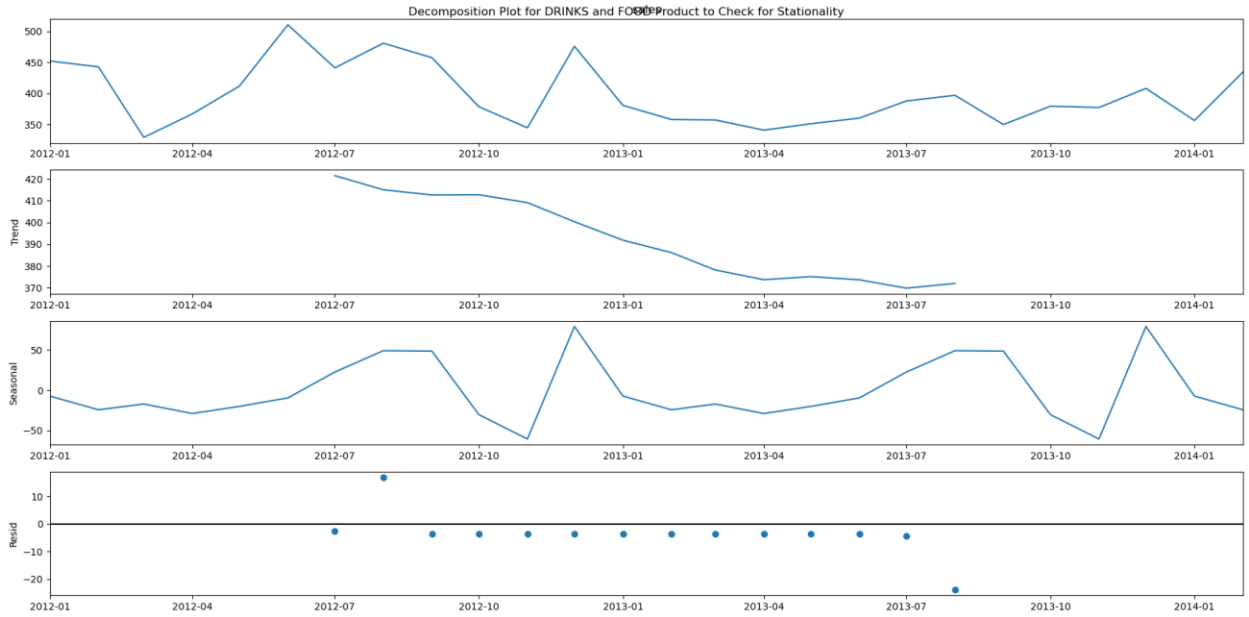

Fig 4 – Seasonal Decomposition for the category Drinks and Food Products.

From the above graph it can be observed that the trend shows a continuously fluctuating trend with a high and low sales values in every month.

Test for stationarity: To determine the whether the data is stationary on not, I have used Augmented dickey-fuller test which gives out the ADF test statistic, p-value, Number of observations used, No.of lags used, Critical Value from these we will be able to identify the data set for which differencing or any of the transformation needs to be done.

```
THE DRINKS AND FOOD PRODUCT SALES.

* ADF Test Statistics:  -3.600104895149909
* P-Value:  0.005759022697757557
* Number of Lags Used:  5
* Number of Observations used for ADF Regression and Critical Values:  108
* Critical Values:
        1% :  -3.4924012594942333
        5% :  -2.8886968193364835
        10% :  -2.5812552709190673

Inference on Stationarity:-
Strong evidence against H0 condition. Data has No Unit Root and is Stationary.
```

Fig 5 – Output of Augmented dickey-fuller test for drinks and food category

Converting non- stationary series into stationary: It is necessary to convert the data set to a stationary data in if not stationary in order to use ARIMA this can be done either by doing the transformation before feeding into ARIMA or can be included in the auto ARIMA function so that it will provide a “d” value which can be added as a parameter in the ARIMA model.

Autocorrelation and Partial Autocorrelation plots: ACP and PCP are used to determine the value of p, q, and d visually from a plot, best option is to use auto ARIMA function to determine these parameters.

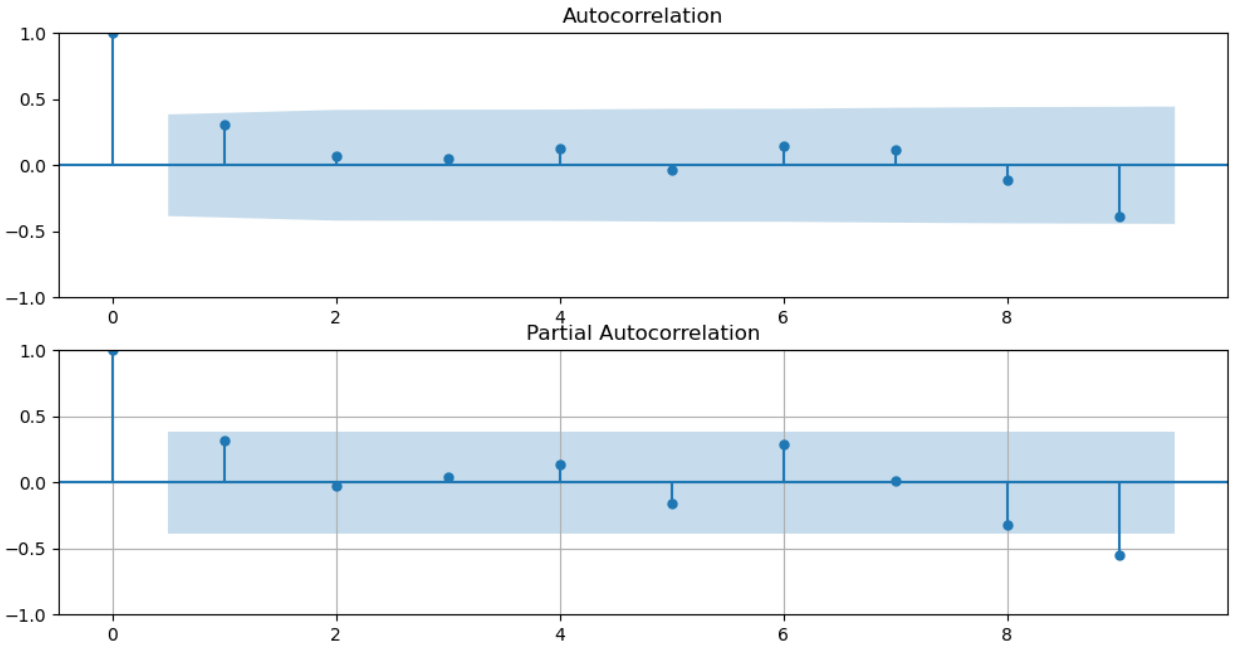

Fig 6 – ACP and PCP for the category Drinks and food products.

Using Auto ARIMA for choosing best parameters: Auto ARIMA is used to find the best set of parameters for the given dataset. This will calculate all the possible combination of the parameters and suggest the best model, which is then directly used to train and predict the value.

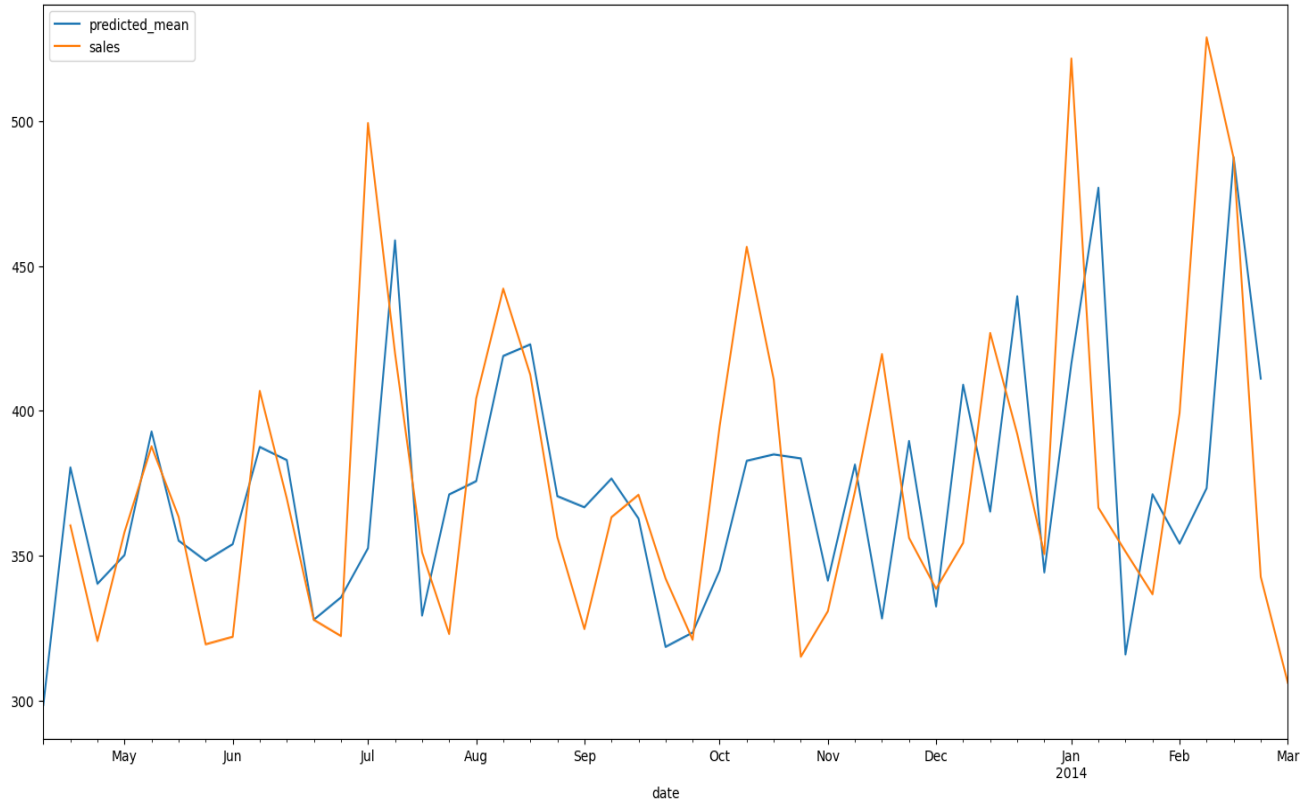

Fig 7 – Comparison between actual value and predicted value for the model SARIMAX

In Drinks and food category.

Fitting the model and predicting the values: The values are from the Auto ARIMA function is fed to the model and model is trained with a required dataset and prediction for the future month is done and these are plotted in a line graph.

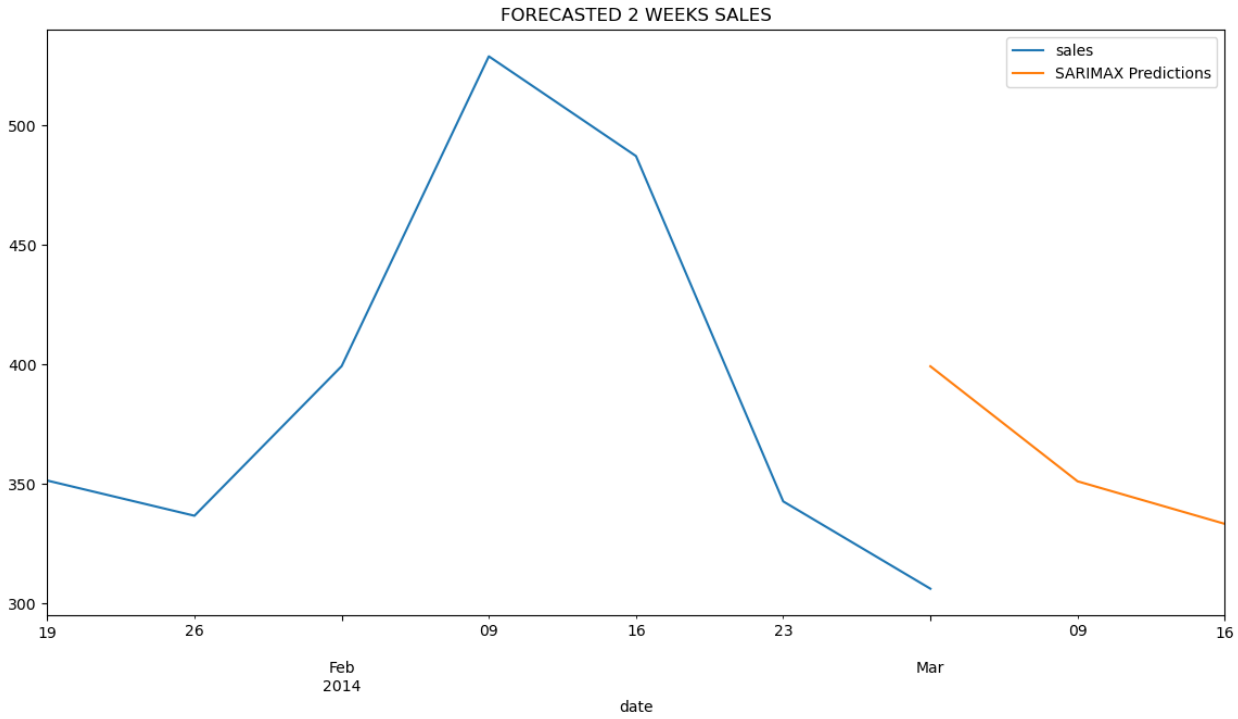

Fig 8 – Prediction for the coming month of February 2014 as per data.

** Please refer the code attached to refer the distribution and predictions of the remaining two categories.

Assumptions

It is assumed that the data from 2012 to 2014 is stable and there are no hidden anomalies in the data.

Charts, Table, Diagrams

The following are the charts and diagrams that describes the data, description of each is given below.

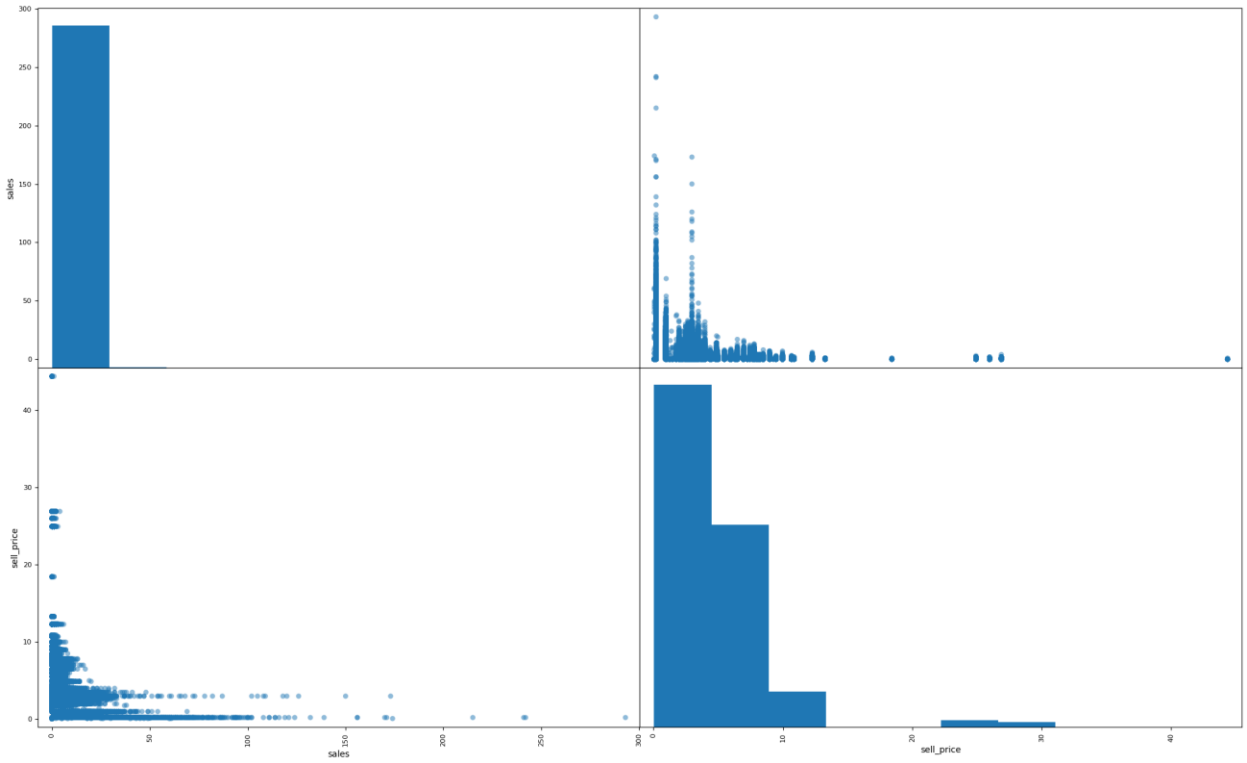

Fig 9 – Correlation plot of sales vs selling price.

hg

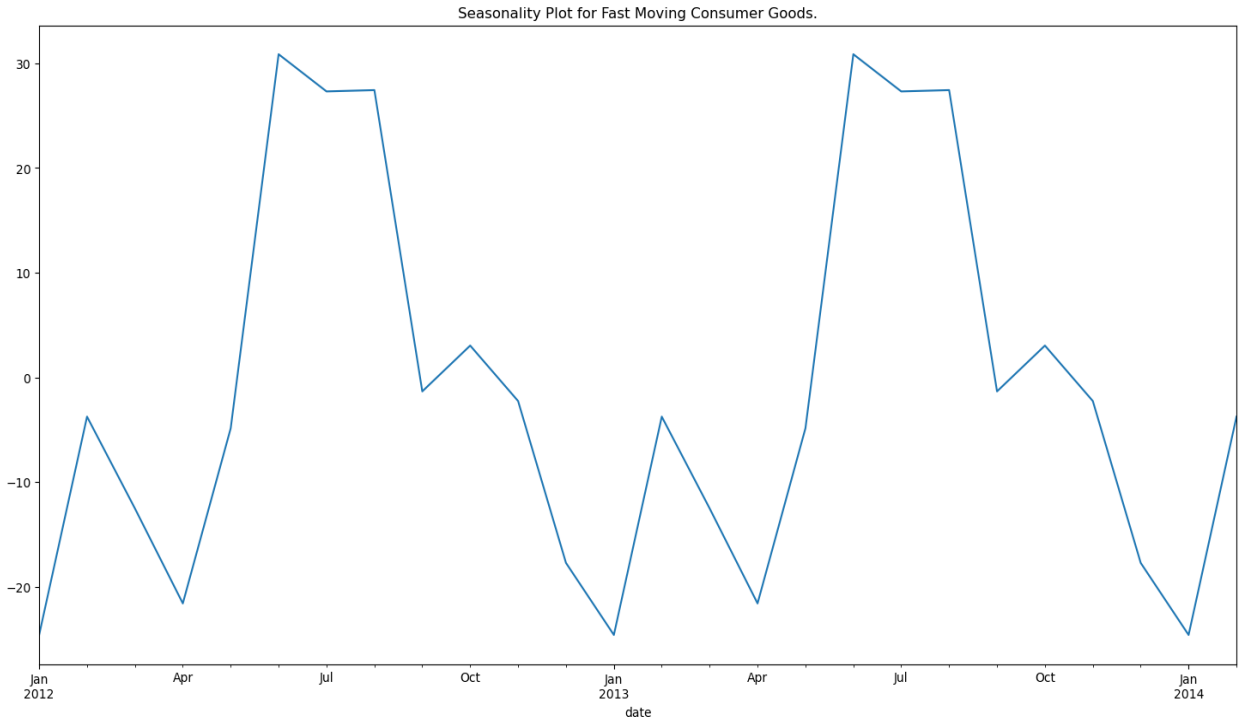

Fig 10 – Seasonality Plot for FMCG.

Reflections on the Internship

The main outcome of the internship is that we could predict the sales of the three category of products for the coming month which helps the sore to stock up the products accordingly.

Recommendations

It is recommended to use models like Fbprophet, Long Short-Term Memory (LSTM) Networks, Vector Auto regression (VAR) for better results and to accommodate holidays and exclusions from the data.

Outcome / Conclusion

From the given data it could be observed that the product category Drinks and foods are Trend graph shows a continuously fluctuating trend with a high and low sales values in every month.

The products under FMCG are showing the trend graph with gradual decrease in sales for nearly a year from July 2012 to July 2013.

For Other items the trend graph shows a gradual decrease in sales for nearly a year from July 2012 to July 2013 with seasonal changes during the year.

The outcome from the model is good for all the categories so that the same model is used to predict the sales for the coming month also.

Enhancement Scope

In this project enhancement can be made by opting different forecasting models and trying to include other external factors like recession and market trend which will affect the sales.

Link to code and executable file:

Demand-forecast-Time-series-analysis/Project_End_Predicting_the_sales_of_products_across_stores_of_a_retail_chain.ipynb at tcs-ioN---RIO-125-Internship-Work · crgs97/Demand-forecast-Time-series-analysis (github.com)