INTERNSHIP: PROJECT REPORT

Internship Project Title	TCS iON RIO-125: Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical Data
Name of the Company	TCS ion
Name of the Industry Mentor	Debashis Roy
Name of the Institute	ICT Academy

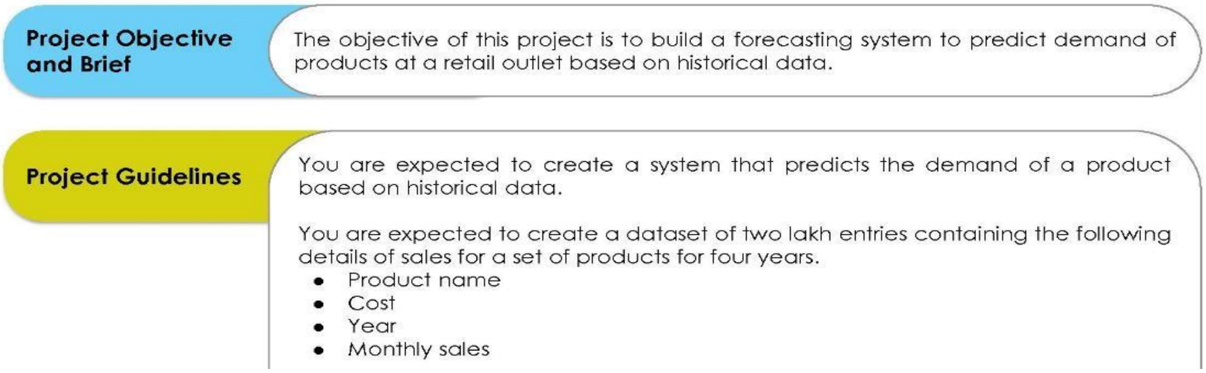
Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
31-3-2022	30-05-2022	125	Jupyter notebook	Python 3 (Numpy, Pandas, Matplotlib, Statsmodels, Seaborn, pmdarima, scikit-learn, fbprophet)

TABLE OF CONTENT

Acknowledgements

I would like to express my deepest gratitude to Mr. Debashis Roy, my industry mentor, and ICT Academy., my internal mentor, and TCS ion for providing me with the necessary facilities for the completion of this project. I am thankful for the valuable discussions I had at each phase of the project and for being a very supportive and encouraging project mentor. I would like to express my sincere thanks to all my friends who were actively part of the discussion room in this project and gave valuable suggestions.

Objective



Project Objective and Brief

The objective of this project is to build a forecasting system to predict demand of products at a retail outlet based on historical data.

Project Guidelines

You are expected to create a system that predicts the demand of a product based on historical data.

You are expected to create a dataset of two lakh entries containing the following details of sales for a set of products for four years.

- Product name
- Cost
- Year
- Monthly sales

Introduction

The dataset used for the problem statement is from a kaggle competition and the problem statement for the competition is as follows:

You are provided with daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

File descriptions:

- sales_train.csv - the training set. Daily historical data from January 2013 to October 2015.
- test.csv - the test set. You need to forecast the sales for these shops and products for November 2015.
- sample_submission.csv - a sample submission file in the correct format.
- items.csv - supplemental information about the items/products.
- item_categories.csv - supplemental information about the items categories.
- shops.csv- supplemental information about the shops.

For the purpose of doing our project we need only the dataset: sales_train.csv and items.csv constituting of various columns like:

- shop_id - unique identifier of a shop
- item_id - unique identifier of a product
- item_cnt_day - number of products sold. You are predicting a monthly amount of this measure
- item_price - current price of an item
- item_category_name - name of item category
- item_category_id - unique identifier of item category
- date - date in format dd/mm/yyyy
- date_block_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33

Here I have done cleaning and sanitizing the dataset and after that, I separate data of different product categories. Then completed some exploratory data analysis of the sales count and sales amount . Then I checked whether the data is stationary or not. Then I searched for a better SARIMA model for this dataset and found out the orders of the best model I can use.

Then built a SARIMA model and plotted my predictions based on that. Also evaluated the performance of my SARIMA models using evaluation metric root mean square error. Also made predictions based on unknown future data points. Then compared the sales happening in each product category by plotting data of them in the same plots.

Then I have done time series modeling of the same dataset with the help of Facebook's prophet library.

Internship Activities

- Gone through all the contents in welcome kit and day wise plan.
- Attempted RIO pre-assessment and passed it successfully in the first attempt itself.
- Interacted in the Digital discussion room.
- Attended both webinar 1 and 2
- Gone through youtube videos on 'Forecasting Methods Overview', 'Moving Averages', 'Time series forecasting', and 'ARIMA models' given in the project reference material.
- Downloaded the dataset from:Kaggle
  Link of the dataset:
  https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales
- Started working on the dataset with Jupyter.
- Imported the dataset to jupyter notebook
- Imported the needed libraries.
- Made sure that the dataset doesn't contain any missing values
- Since the dataset is huge as it contain the sales data of multiple stores. I particularly selected a shop that fit the criteria of the problem statement.
- Then I tried to split the data into different data frames and try to do the model separately based on their category of the item which they

belong to. But this step was failed as the data has more than 60 categories.

- Restructured the data based on total sales occurring on each date.
- Made new data frames x_sa(sales amount),y_sc(sales count) which included the mean sales data of each product type of each month.
- Plotted the sales count and sales amount occurring on each day.
- Plotted the sales count and sales amount occurring on each month.
- Created boxplots based on sales count and amount.
- Performed ETS (Error Trend Seasonality) Decomposition on sales count and amount.
- Conducted Augmented-Dickey-Fuller test to verify sales count and sales amount are stationary.
- Found the best SARIMA time series forecasting models for sales count and sales amount with the help of auto_arima function of pmdarima library in python.
- Created time series models of this order with the help of SARIMAX function.
- Compared predicted results with the test set data points and evaluated performance of my model with the help root mean square function.
- Done predictions of the data points in the unknown future.
- Built some deep learning models and compared their performance with SARIMA models and found that SARIMA models are better.
- Compared the sales happening in sales count and sales amount and plotted them.
- Done time series modeling with Facebook’s Prophet library.

Methodology

ARIMA and SARIMA models can be used for time series modeling tasks like this.

• ARIMA

(Auto Regressive Integrated Moving Average)

ARIMA performs well when working with a time series where the data is directly related to the time stamp. ARIMA model won't be able to understand any outside factors which weren't already present in the current data. ARIMA is fitted to time series data to better understand the data or to predict future points in the series (forecasting). ARIMA models can be applied when data is stationary and can be applied to non-stationary data after making it stationary through steps like differencing.

In autoregression model, we forecast using a linear combination of past values of the variable. The term autoregression describes a regression of the variable against itself. An autoregression is run against a set of lagged values of order p. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term).

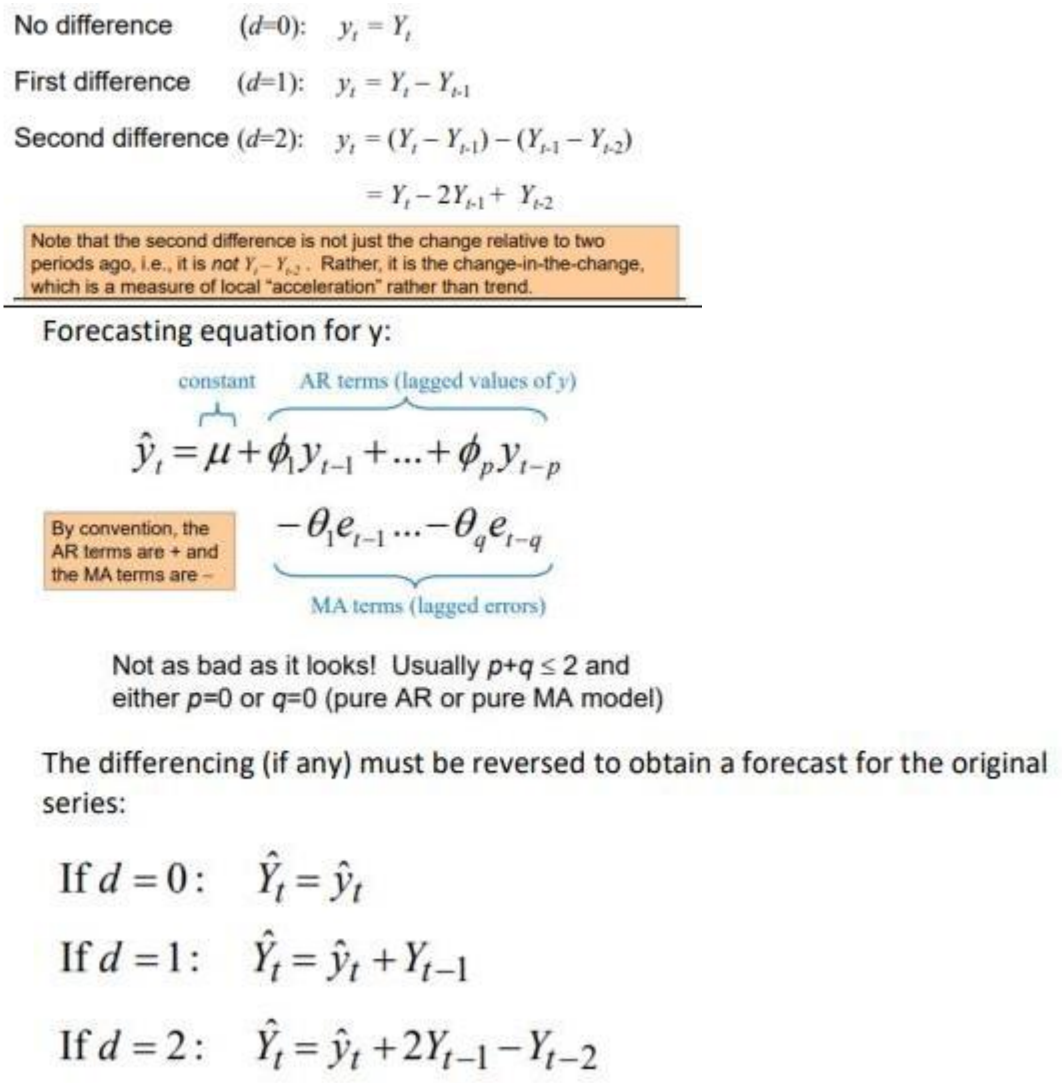
“Moving Average” (MA) Indicates the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

“Integrated” (I) Indicates that the data values have been replaced with the difference between their values and the previous values. This basically just means how many times did we have to difference the data to get it stationary so the AR and MA components could work.

A non-seasonal ARIMA model can be (almost) completely summarized by three numbers:
p = the number of autoregressive terms
d = the number of nonseasonal differences
q = the number of moving-average terms
This is called an “ARIMA(p,d,q)” model. The model may also include a constant term (or not).

- ARIMA forecasting equation

Let Y denote the original series.
Let y denote the differenced (stationarized) series.



No difference (*d*=0): $y_t = Y_t$

First difference (*d*=1): $y_t = Y_t - Y_{t-1}$

Second difference (*d*=2): $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$

$= Y_t - 2Y_{t-1} + Y_{t-2}$

Note that the second difference is not just the change relative to two periods ago, i.e., it is *not* $Y_t - Y_{t-2}$. Rather, it is the change-in-the-change, which is a measure of local "acceleration" rather than trend.

Forecasting equation for y:

$$\hat{y}_t = \overbrace{\mu}^{\text{constant}} + \overbrace{\phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}}^{\text{AR terms (lagged values of } y\text{)}}$$

By convention, the AR terms are + and the MA terms are –

$$\underbrace{-\theta_1 e_{t-1} \ldots - \theta_q e_{t-q}}_{\text{MA terms (lagged errors)}}$$

Not as bad as it looks! Usually $p+q \leq 2$ and either p=0 or q=0 (pure AR or pure MA model)

The differencing (if any) must be reversed to obtain a forecast for the original series:

$$\text{If } d = 0: \quad \hat{Y}_t = \hat{y}_t$$

$$\text{If } d = 1: \quad \hat{Y}_t = \hat{y}_t + Y_{t-1}$$

$$\text{If } d = 2: \quad \hat{Y}_t = \hat{y}_t + 2Y_{t-1} - Y_{t-2}$$

• SARIMA

The seasonal part of an ARIMA model is summarized by three additional numbers:

P = number of seasonal autoregressive terms

D = number of seasonal differences

Q = number of seasonal moving-average terms

The complete model is called an "ARIMA(p,d,q)X(P,D,Q)" model.

• Choosing best orders of ARIMA using pmdarima library.

The pmdarima (Pyramid ARIMA) is a separate library designed to perform grid searches across multiple combinations of p, d, q and P, D, Q. The pmdarima library utilizes the Akaike Information criterion (AIC) as a metric to compare the performance of various ARIMA based models. Then auto_arima function chooses the model with a minimum AIC value.

• Training the models using SARIMAX function
The statsmodels implementation of SARIMA is called SARIMAX. The "X" added to the name means that the function also supports exogenous regressor variables.

• Fbprophet library
Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

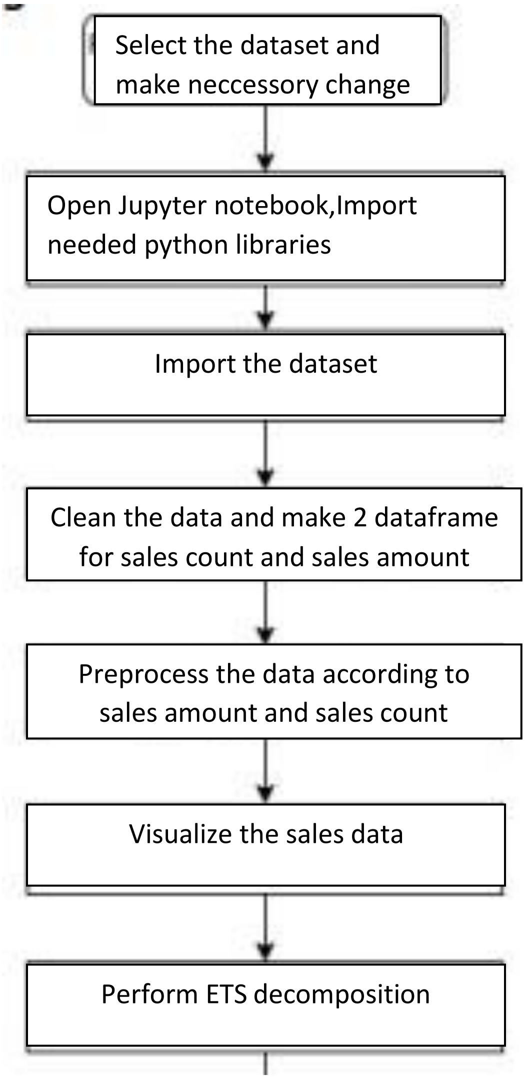
References:

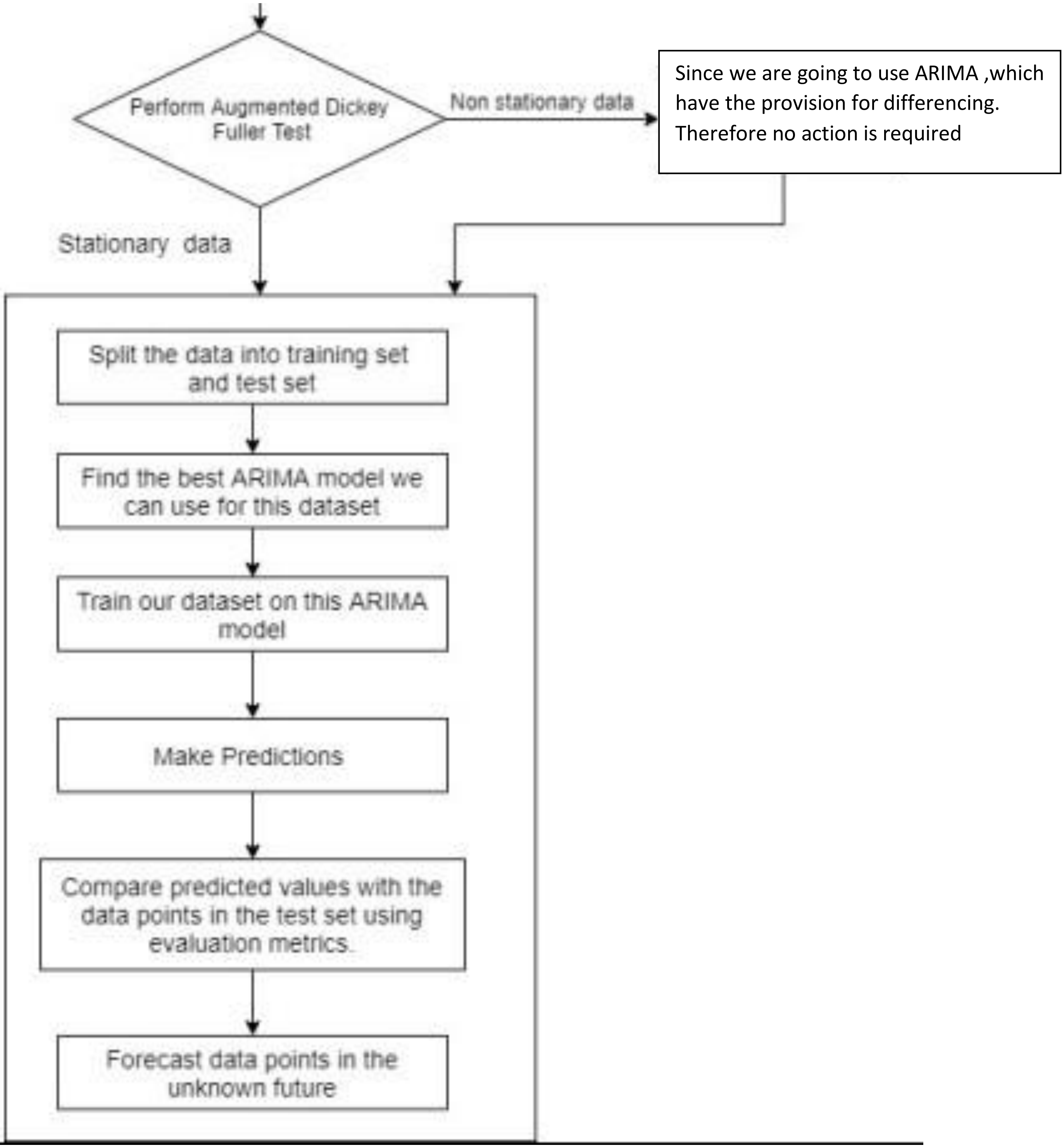
1. https://www.udemy.com/course/python-for-time-series-data-analysis
2. https://people.duke.edu/~rnau/Slides_on_ARIMA_models--Robert_Nau.pdf
3. https://courses.pieriandata.com/
4. https://www.youtube.com/watch?v=95-HMzxsghY
5. https://facebook.github.io/prophet/docs/quick_start.html#python-api
6. https://research.fb.com/prophet-forecasting-at-scale/
7. https://blog.exploratory.io/is-prophet-better-than-arima-for-forecasting-time-series-fa9ae08a5851

Charts, Table, Diagrams:

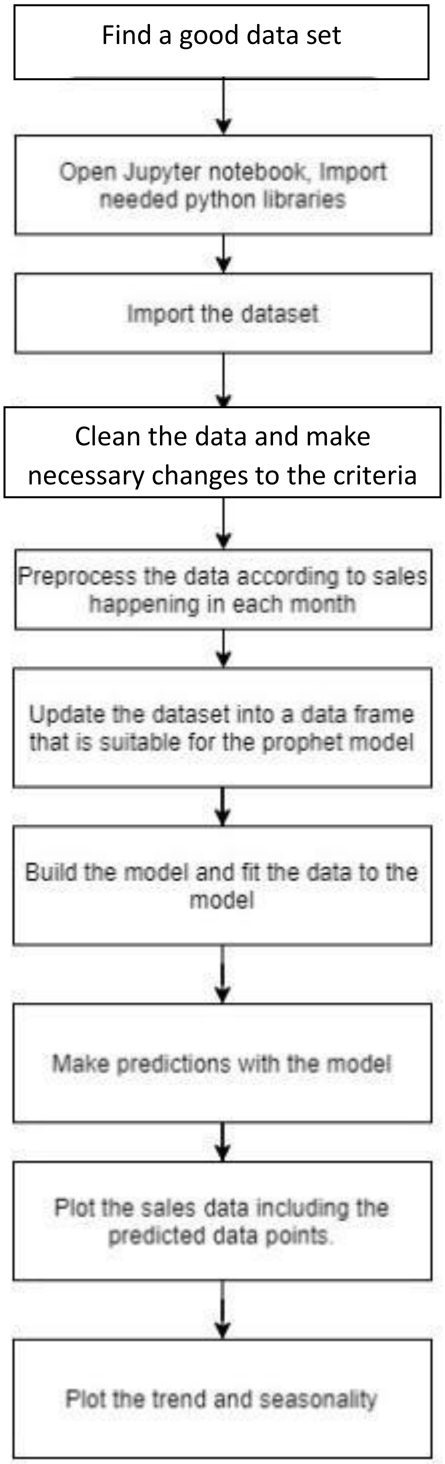
Project Workflow:

Time series forecasting with ARIMA

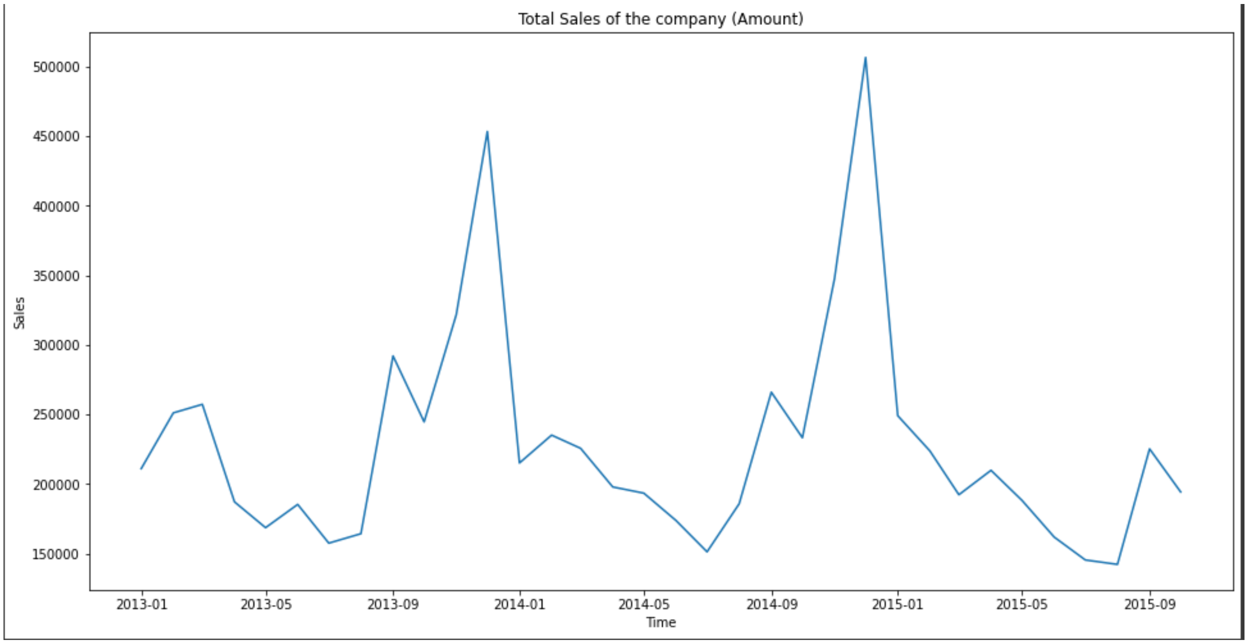

Select the dataset and make neccessory change
Open Jupyter notebook,Import needed python libraries
Import the dataset
Clean the data and make 2 dataframe for sales count and sales amount
Preprocess the data according to sales amount and sales count
Visualize the sales data
Perform ETS decomposition


Perform Augmented Dickey Fuller Test
Non stationary data
Since we are going to use ARIMA ,which have the provision for differencing. Therefore no action is required
Stationary data
Split the data into training set and test set
Find the best ARIMA model we can use for this dataset
Train our dataset on this ARIMA model
Make Predictions
Compare predicted values with the data points in the test set using evaluation metrics.
Forecast data points in the unknown future

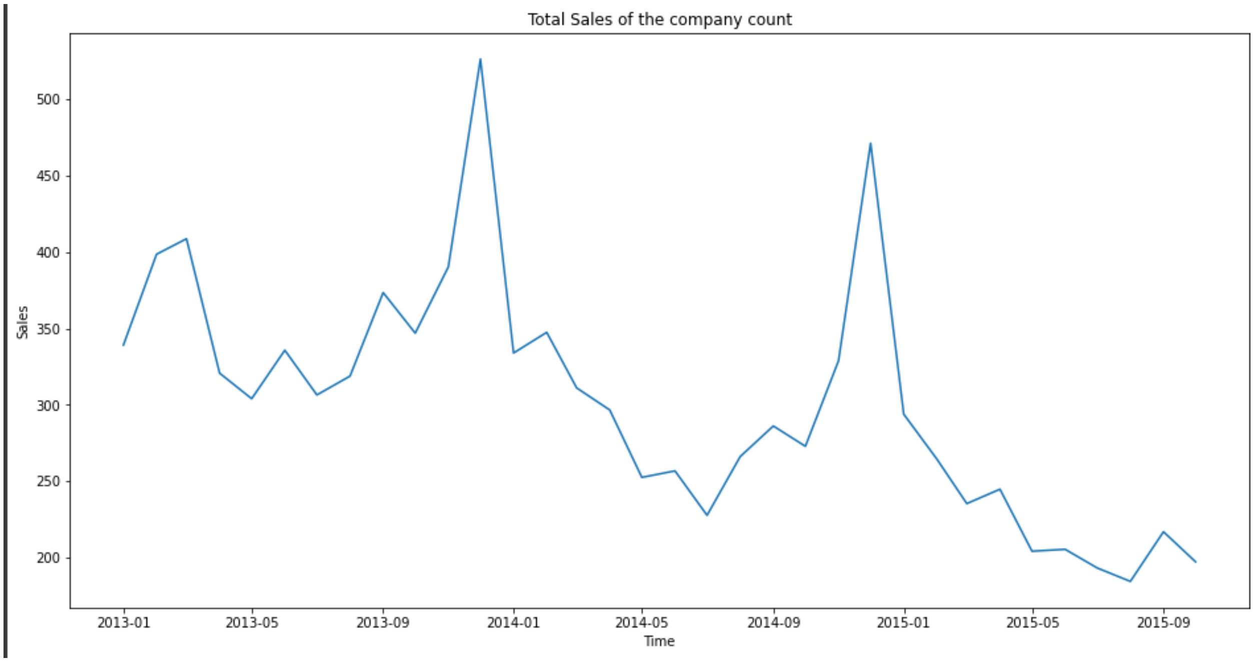
Time series modeling with prophet


Find a good data set
Open Jupyter notebook, Import needed python libraries
Import the dataset
Clean the data and make necessary changes to the criteria
Preprocess the data according to sales happening in each month
Update the dataset into a data frame that is suitable for the prophet model
Build the model and fit the data to the model
Make predictions with the model
Plot the sales data including the predicted data points.
Plot the trend and seasonality

Line plots of Sales based on the mean of monthly sales:

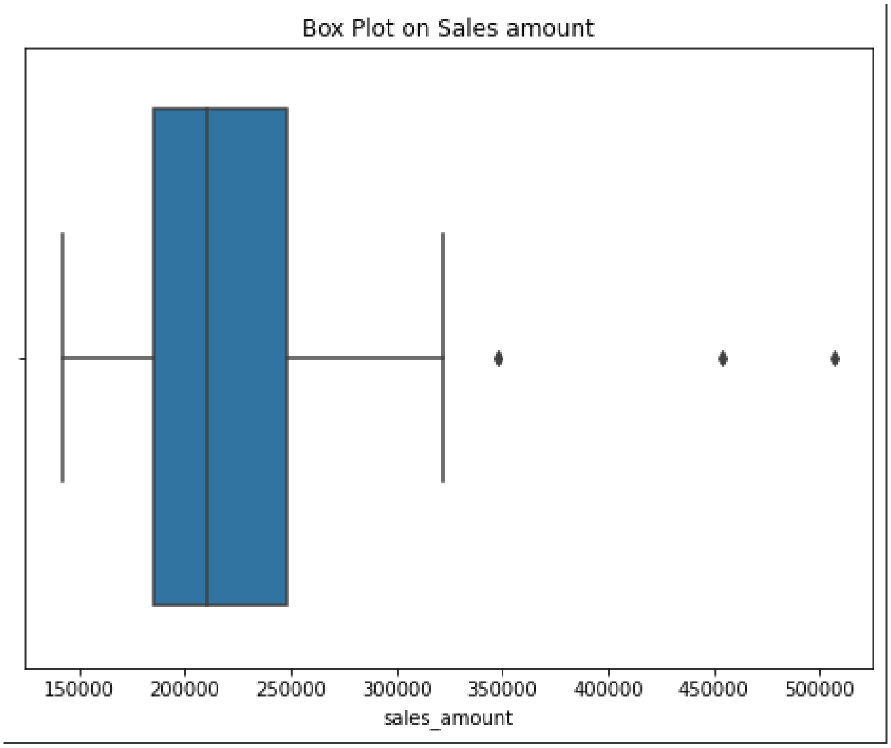


- In the sales(Amount) category, the highest sales occurred during December 2014 and the least sales occurred during July 2013

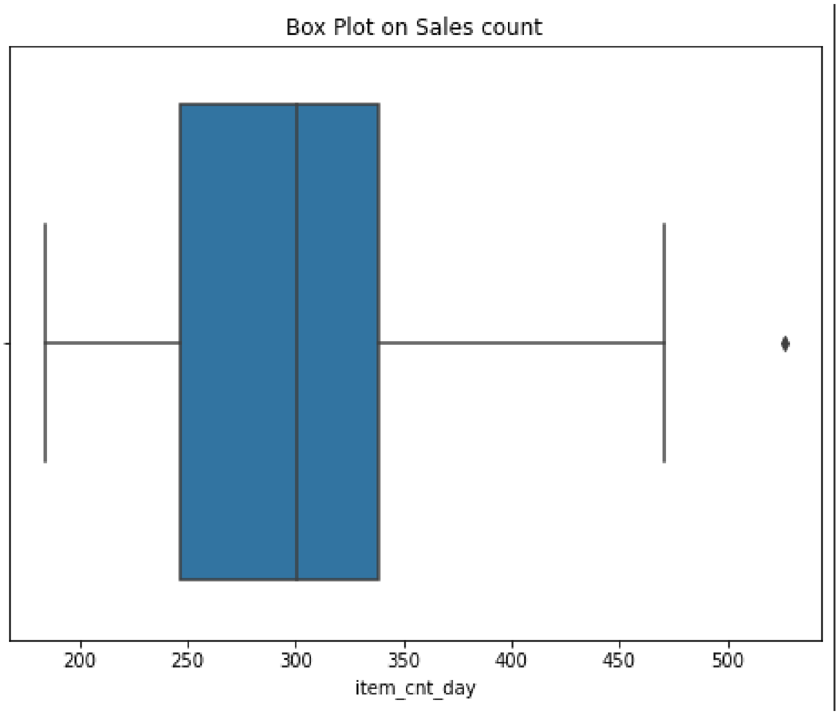


- In the sales(count) category, the highest sales occurred during December 2013 and the least sales occurred during July 2015

Boxplot on sales data of count and amount (Mean of monthly sales):



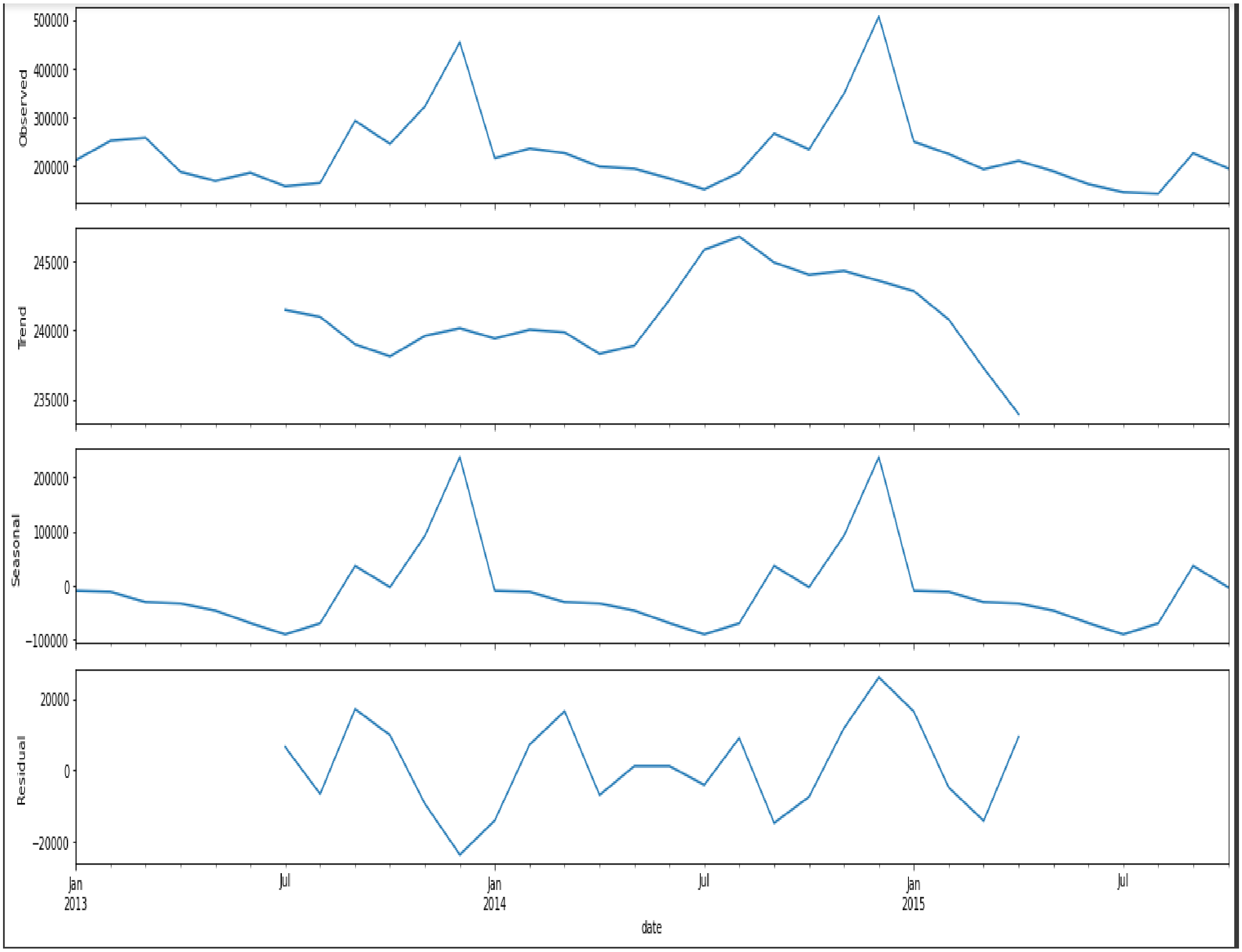
- In the sales amount category, there are 3 outliers in the data.



- In the sales count category, there is 1 outlier in the data.

ETS decomposition plot of sales(count and amount):

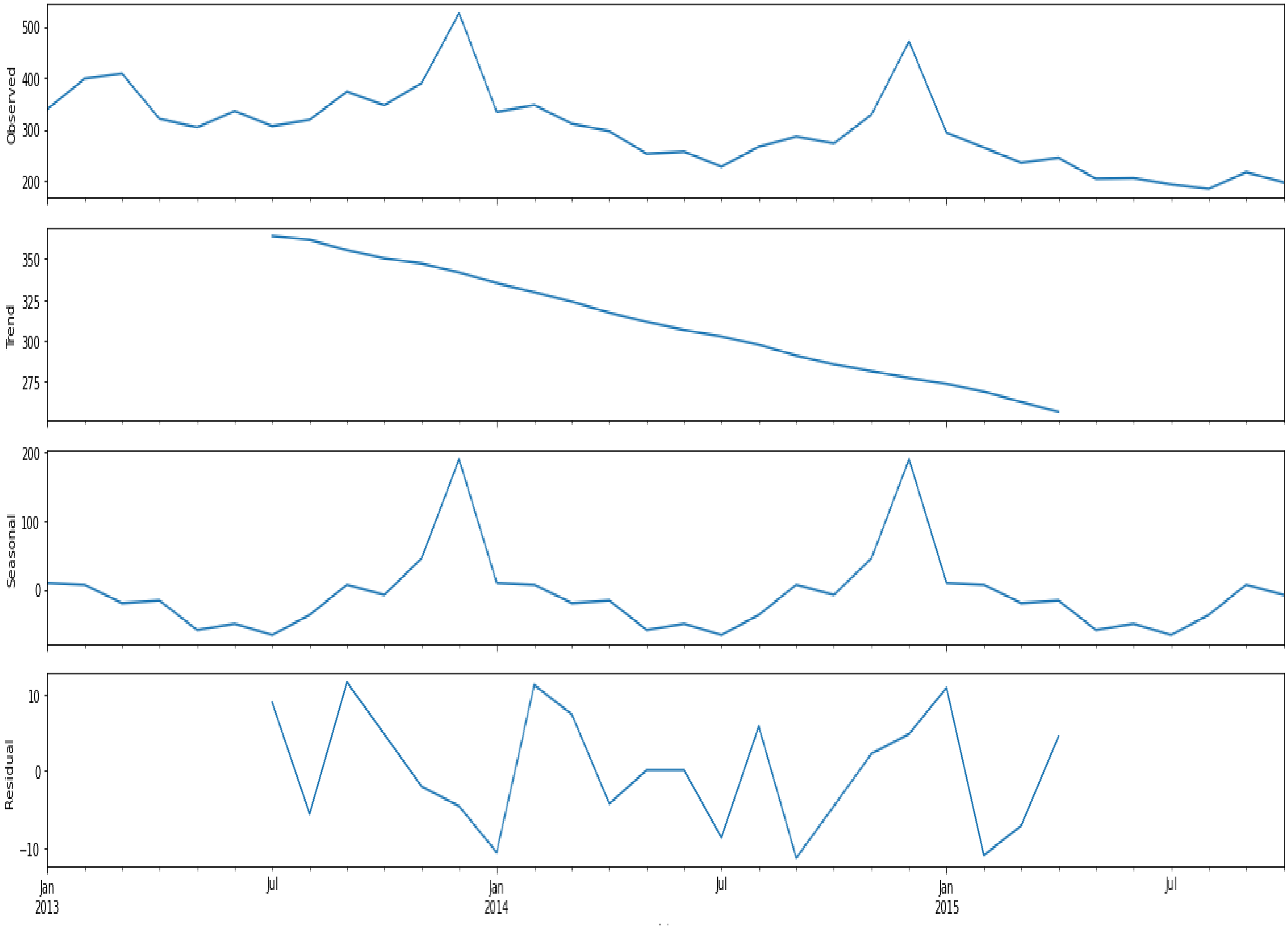
- **Sales amount**



Trend: A decreasing trend from the beginning of July 2013, then somewhat steady till may 2014 which make way to a rise till august 2014 from there on seeing a down steep till April 2015.

Seasonality: Sales is at its peak high in the month of November and January and least at the month of July

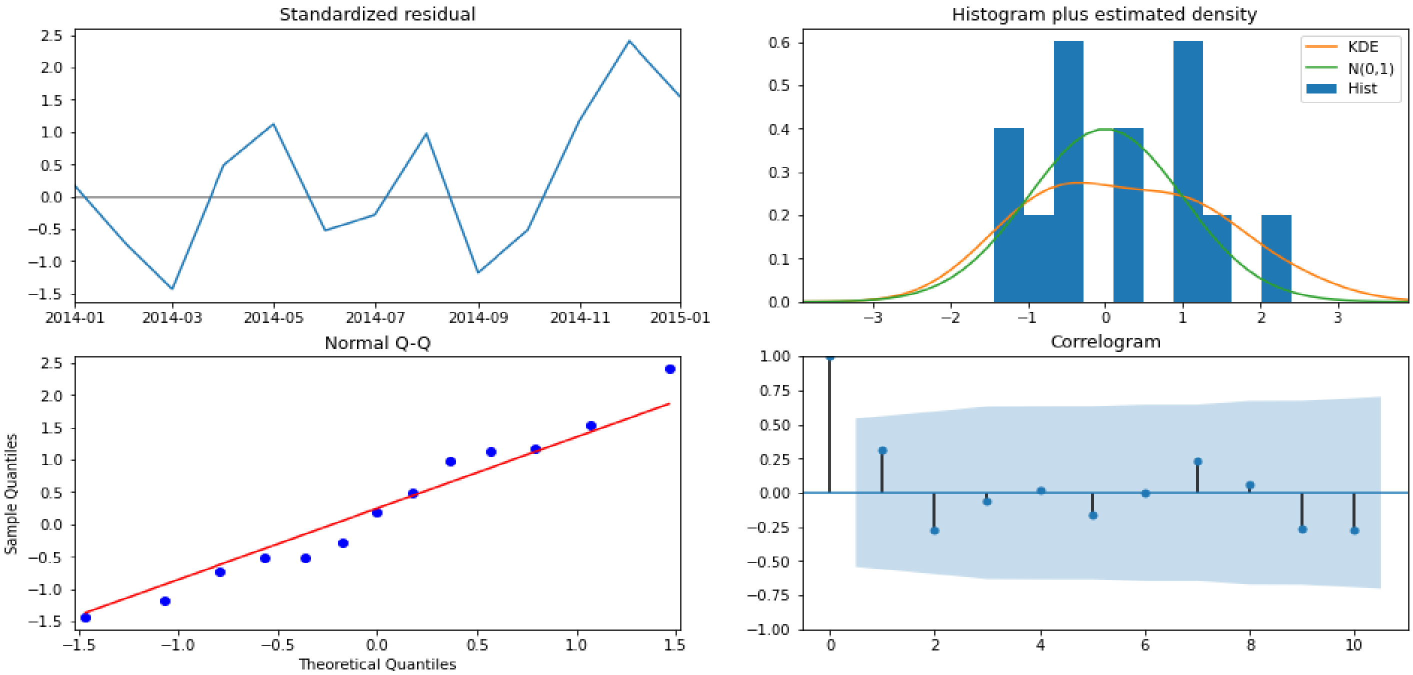
- Sales count



Trend: A steady downward trend almost linear from July 2013 to March 2015 .

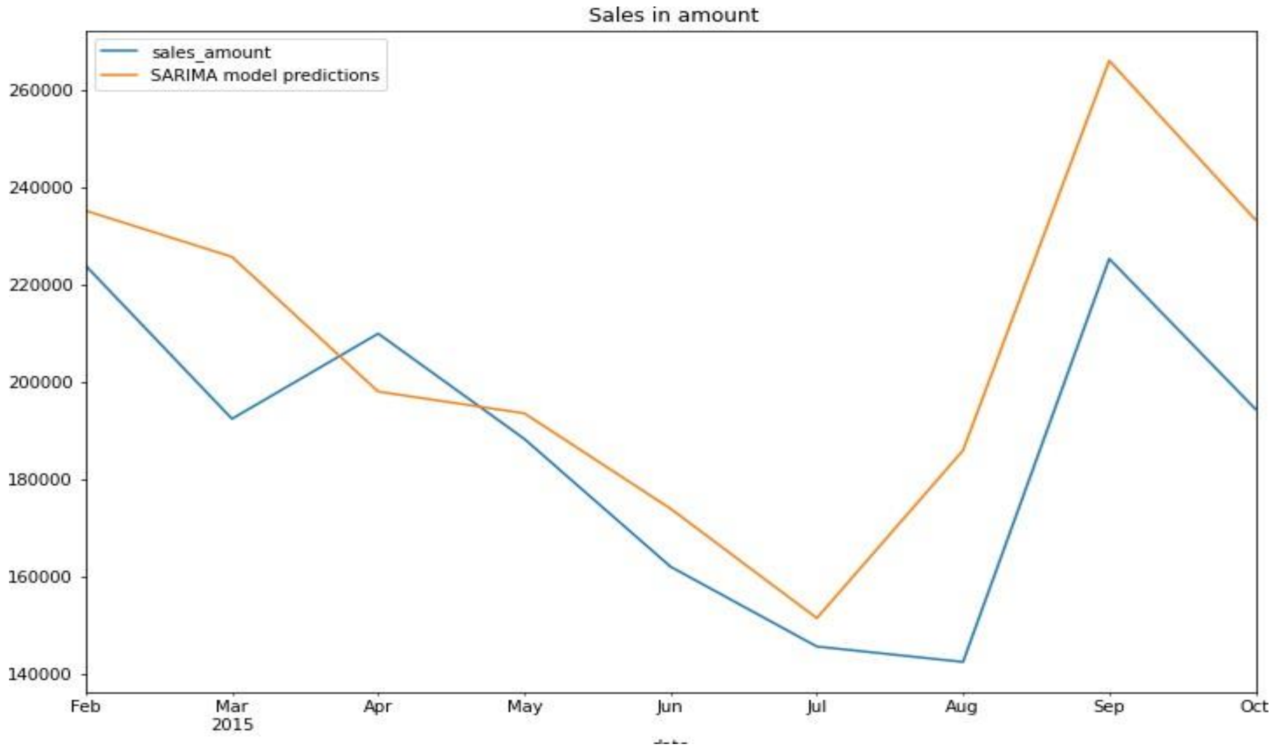
Seasonality: Sales is at its peak high in the month of November and January and least at the month of July.

Plot obtained after running model diagnostics test on the SARIMA model built for Sales amount data:



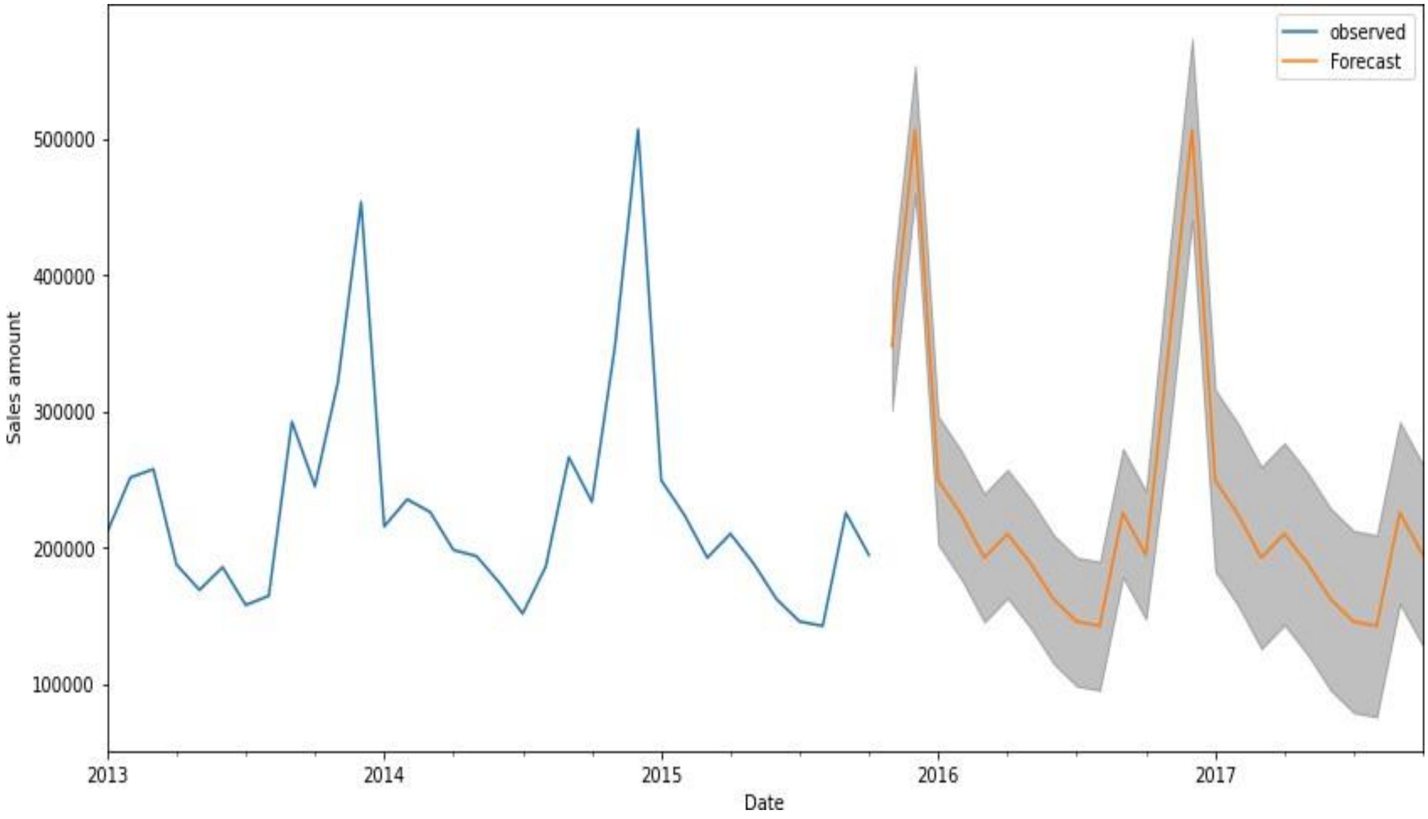
- From the above plots, we can say that our residuals are normally distributed

Plot obtained after making predictions about the known future and comparing it to the data points in the test set of Sales amount:



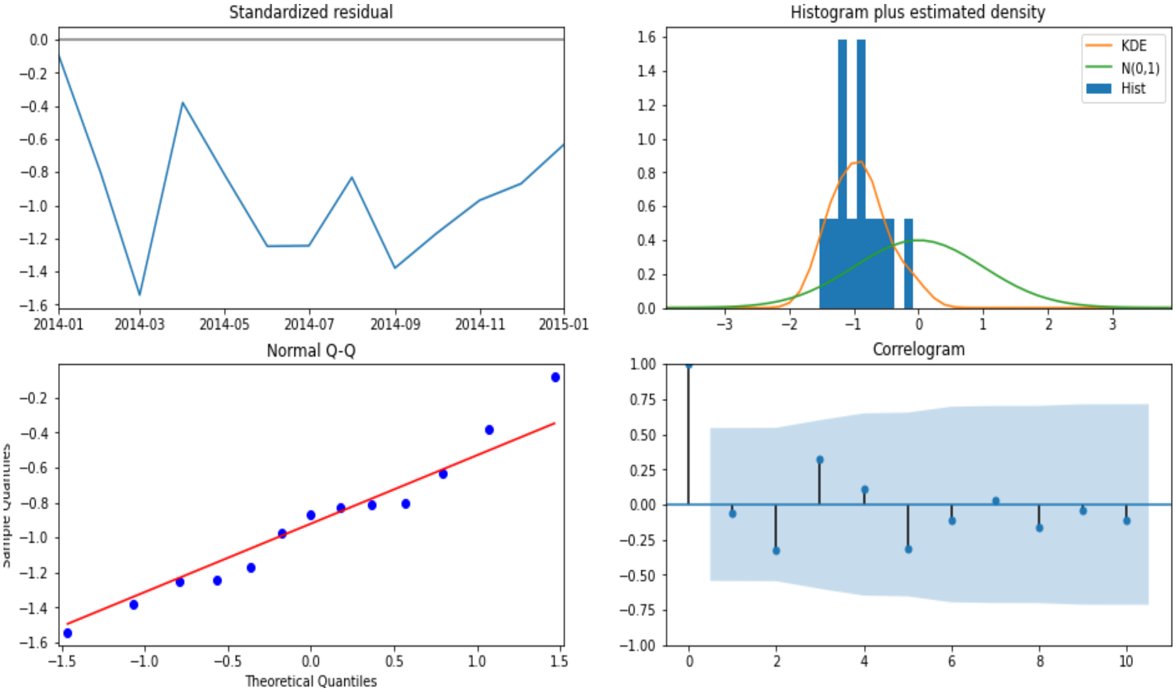
- From April to July model is predicting very well
- Even though our model predicted the seasonality very well the real trend that happened was higher than that happened in the previous years. As a result, the predicted results are less than the observed results

Plot obtained after making predictions to the unknown future for Sales amount (predicted values for the next two years):



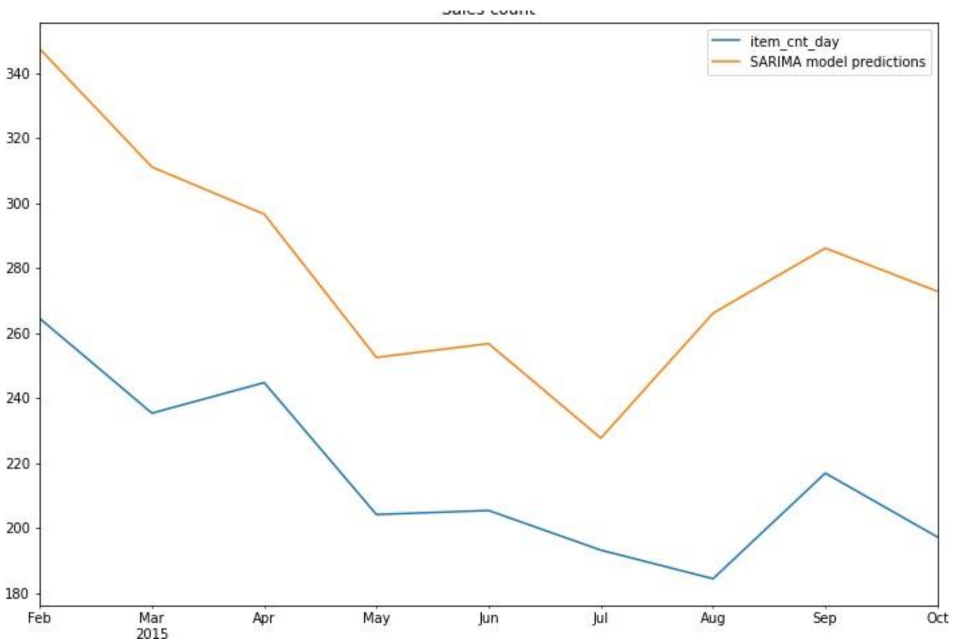
- There is a steady growth in the peak value around January of the coming 2 years

Plot obtained after running model diagnostics test on the SARIMA model built for Sales count data:

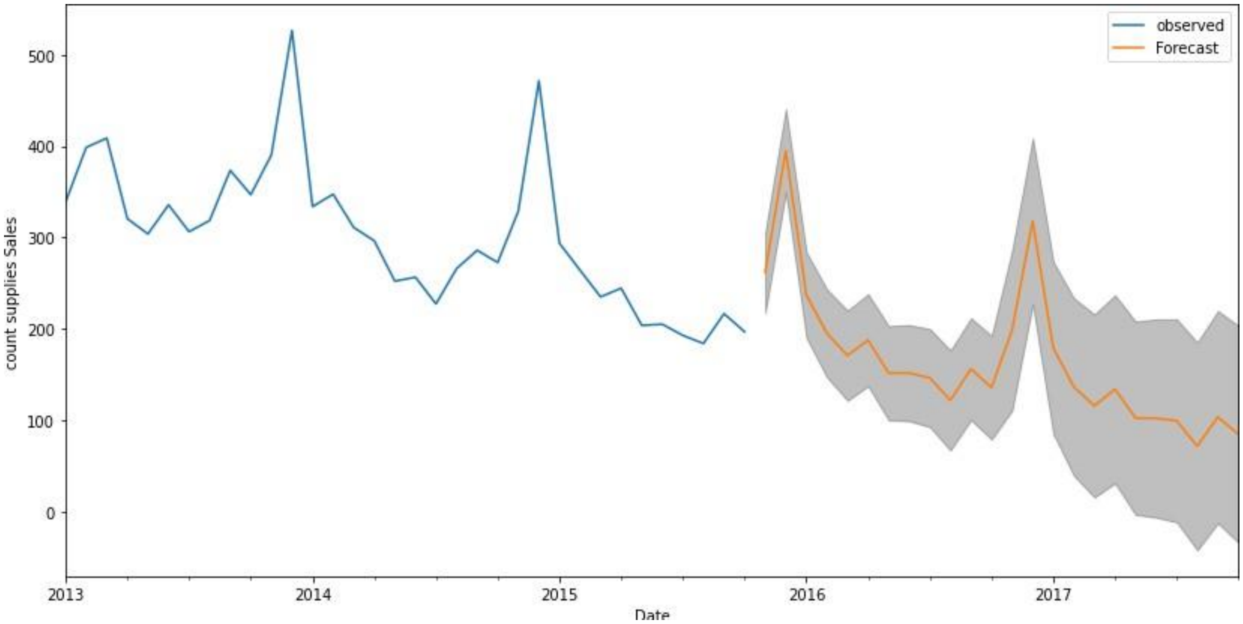


- From the above plots, we can say that our residuals are normally distributed

Plot obtained after making predictions about the known future and comparing it to the data points in the test set of Sales count:



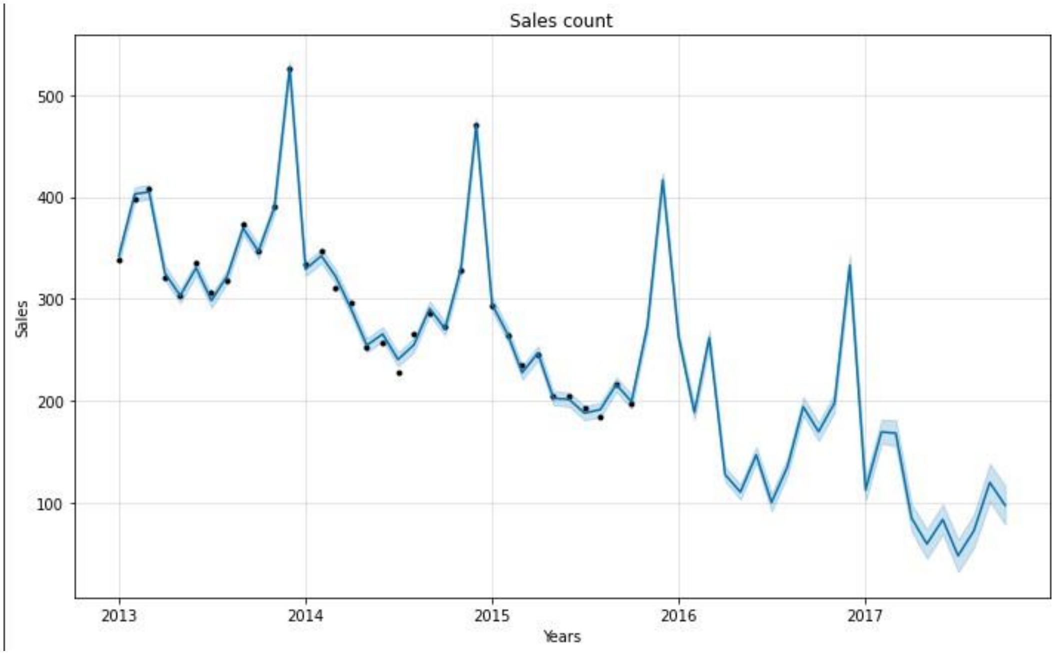
Plot obtained after making predictions to the unknown future for Sales count (predicted values for the next two years):



- It is clearly evident of the ongoing downward stats in the sales count in the coming 2 years

<u>Plots made with the help of fbprophet library:</u>

Plot made after forecasting sales count for next two years:

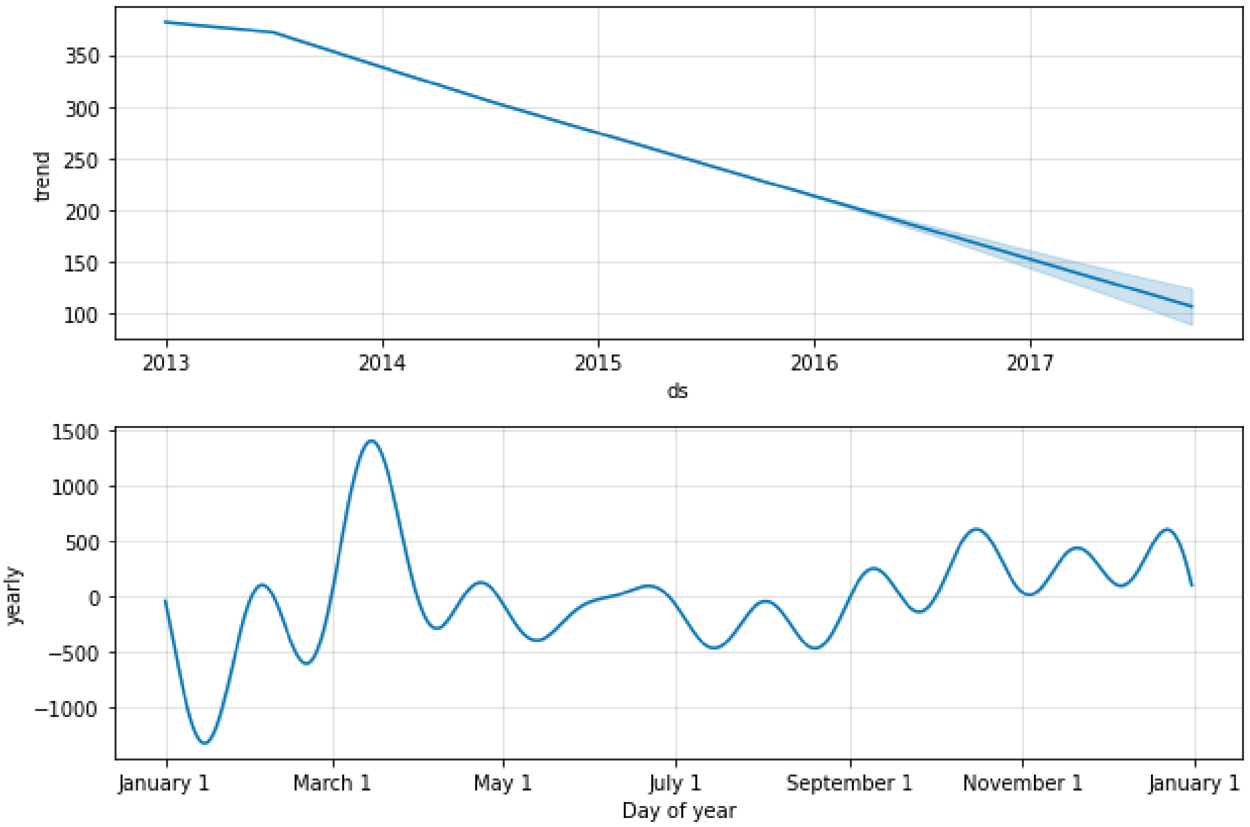


• In the predicted values, highest sales are predicted during the month of December 2015.

• In the predicted values lowest sales are predicted during the month of July 2017.

• We can see that in both the years 2016 and 2017, our model has predicted more sales during the months of December, January and March while comparing to other months.

• Our model has predicted lowest sales in the month of July during 2016 and 2017.



• From the plot of trends, it is clear that the trend is linearly dercreasing over the years. So, the sales of the products are decreasing over the years that means demand of the products are decreasing over the years.

• From the plot of yearly seasonality, it is clear that higher seasonality occurs during the first quarter of the year lowest seasonality occur during Second and third quarter of the year.

Reflections on the Internship:

- It was a great learning experience. It taught me how projects are done in the industry.
- Submitting daily activity reports helped me to keep track of what I was doing each day.
- The discussion forum helped me to connect with other learners and discuss topics related to this subject.

Conclusions:

- When considering sales count, more sales are happening in the month of December. So during this month demand for these products are higher.
- When considering sales count, fewer sales are happening in the month of July. So the manager should adopt any new business approaches to increase the sales during these months.
- When considering sales count, we can expect the demand for products will decrease in the next two years.
- When considering sales count, the sales may decrease up to 90 (Expected during (July 2017).
- So in conclusion there is an evident decreasing trend in the sales so there should be extra caution from the management side, and try to do some effort in the form of promotion of the products to increase the sales

Link to code

1. Google colab:

https://colab.research.google.com/drive/1Nn9rZGgbYwsT6q-V7qGhB_XfTq9lTEjw?usp=sharing

2 Github repository:

https://github.com/SHILPA-GRACE-V/TCS-IoN