INTERNSHIP: INTERIM PROJECT REPORT

Internship Project Title	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	Mr. Himalaya Ashish
Name of the Institute	Indian Institute Of Information Technology and Management,Trivandrum

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
14/03/2021	12/06/2021	125	Jupyter Notebook	Python 3 (Numpy, Pandas, Matplotlib, Statsmodels, Seaborn, pmdarima, scikit-learn, fbprophet)

Acknowledgement:
I would like to express my deepest gratitude to Mr. Himalaya Ashish, my industry mentor, and Dr. Manoj Kumar T.K., 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:
Create a forecast model applying concepts of moving averages, forecasting methods, ARIMA models and time series forecasting. The model should be able to predict future sales by training on the historical sales data.

Introduction:
Here I have a superstore dataset that has sales data of different products from 2014-01-06 to 2017-12-30. Here the product categories sold in this superstore are furniture, technology, and office supplies. 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 this product category data. 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.
- Introduced myself 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 https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls
- 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
- Made new data frames furniture, office and technology which included the sales data of each product category.
- Reduced the dataset columns into order date and sales only.
- Restructured the data based on total sales occurring on each date.
- Made new data frames y_furniture, y_office, and y_technology which included the mean sales data of each product type of each month.
- Plotted the sales data of each of the product categories occurring on each day.
- Plotted the mean sales data of each of the product categories occurring on each month.
- Created boxplots based on sales of each product category.
- Performed ETS (Error Trend Seasonality) Decomposition on sales data of each product category.
- Conducted Augmented-Dickey-Fuller test to verify which all product category data are stationary.
- Found the best SARIMA time series forecasting models for each product category data 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 the 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 each product categories 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:

constant AR terms (lagged values of y)

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}$$

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

$$- \theta_1 e_{t-1} \ldots - \theta_q e_{t-q}$$

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.

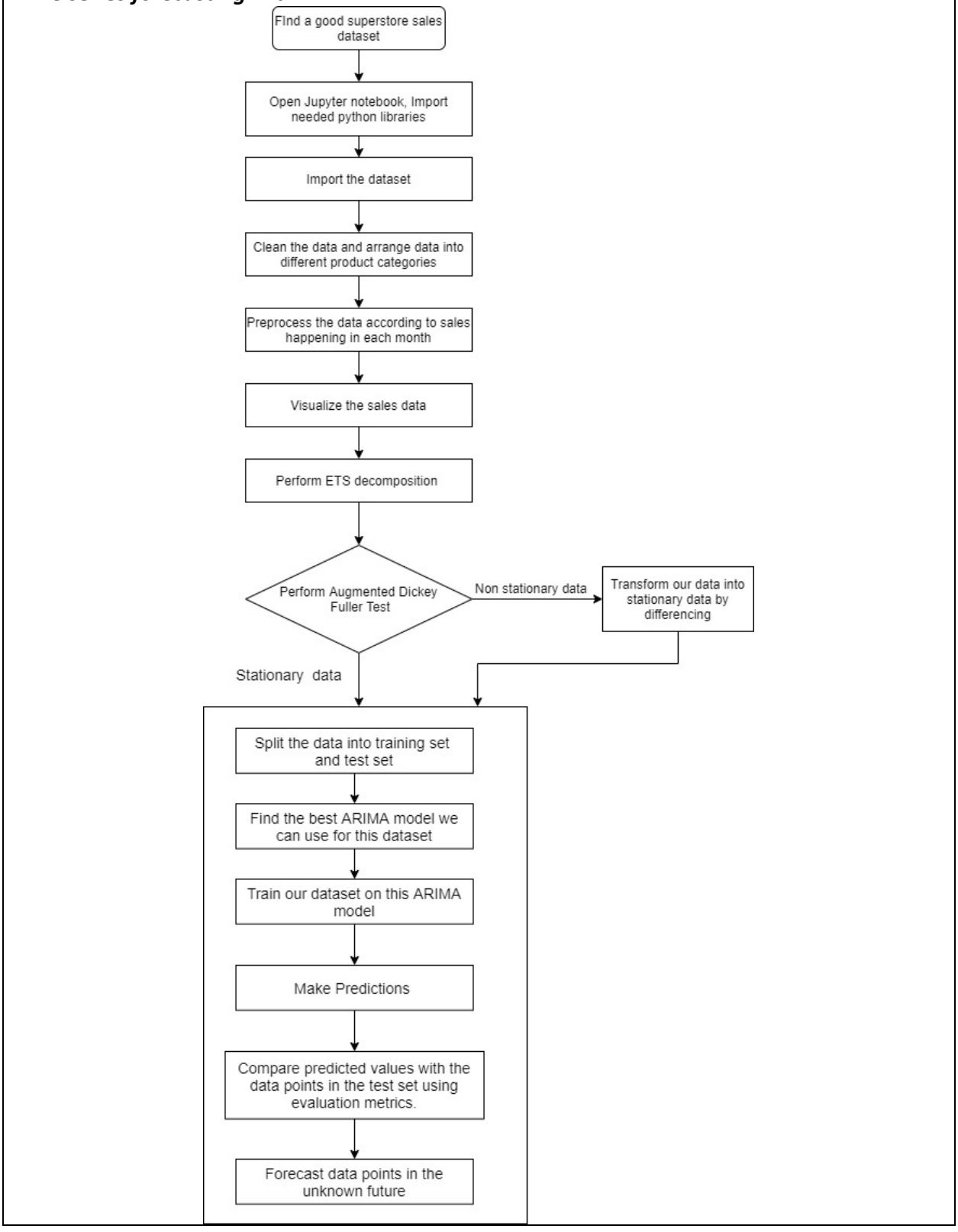
References:

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. Research paper on facebook prophet library

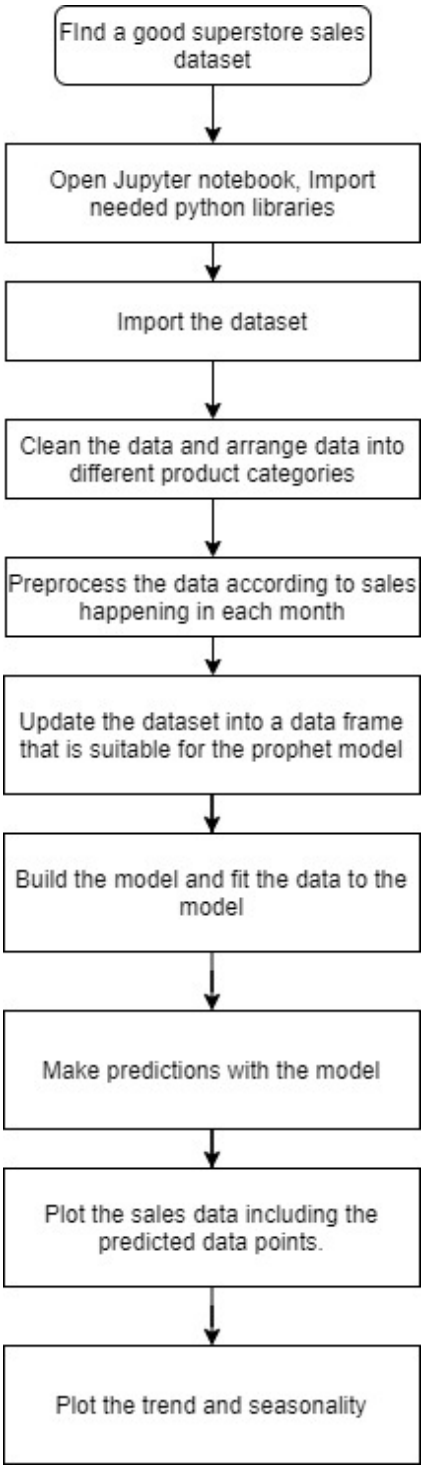
Charts, Table, Diagrams:

Project Workflow:

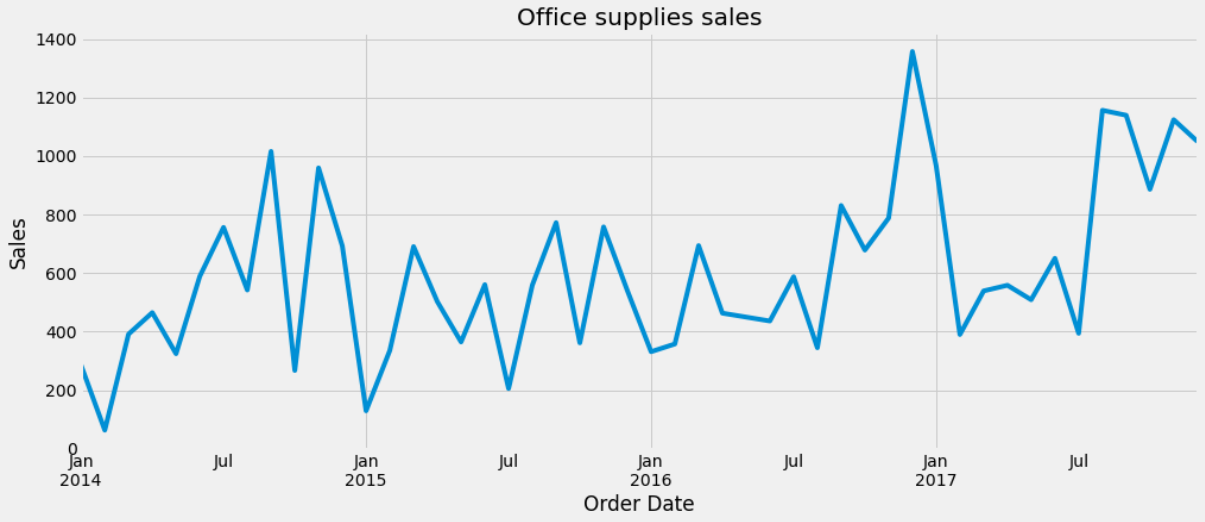
Time series forecasting with ARIMA



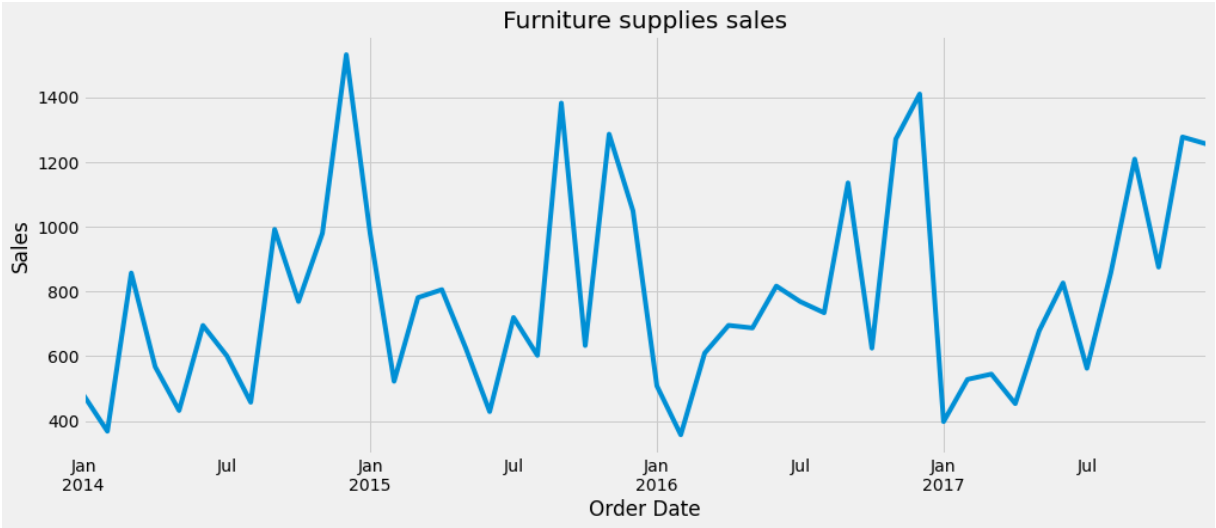
Time series modelling with prophet



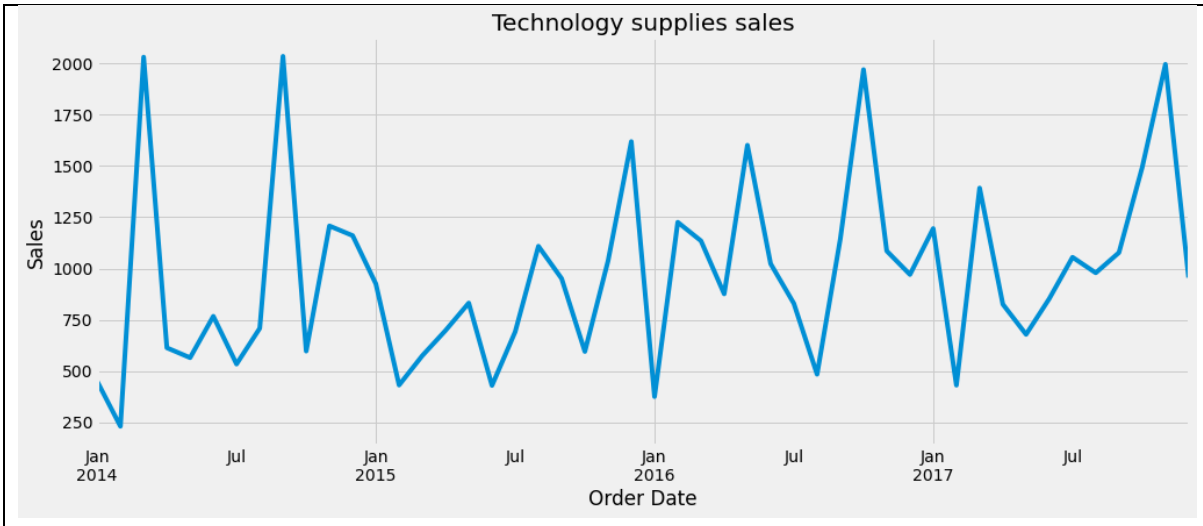
Line plots of product categories based on the mean of monthly sales:



- In the office supplies sales category, the highest sales occurred during December 2016 and the least sales occurred during February 2014.

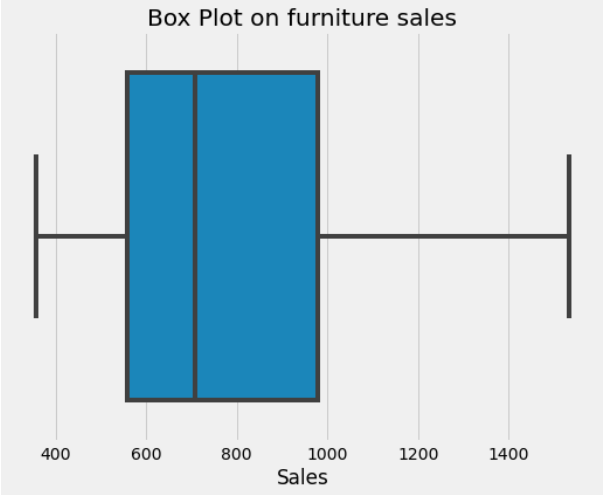


- In the furniture sales category, the highest sales occurred during December 2014 and the least sales occurred during February 2016.

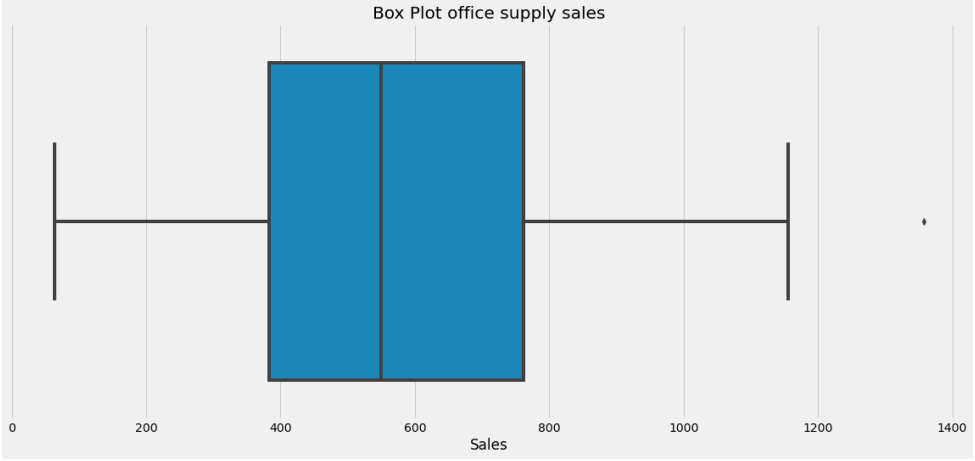


- In the furniture sales category, the highest sales occurred during September 2014 and the least sales occurred during February 2014.

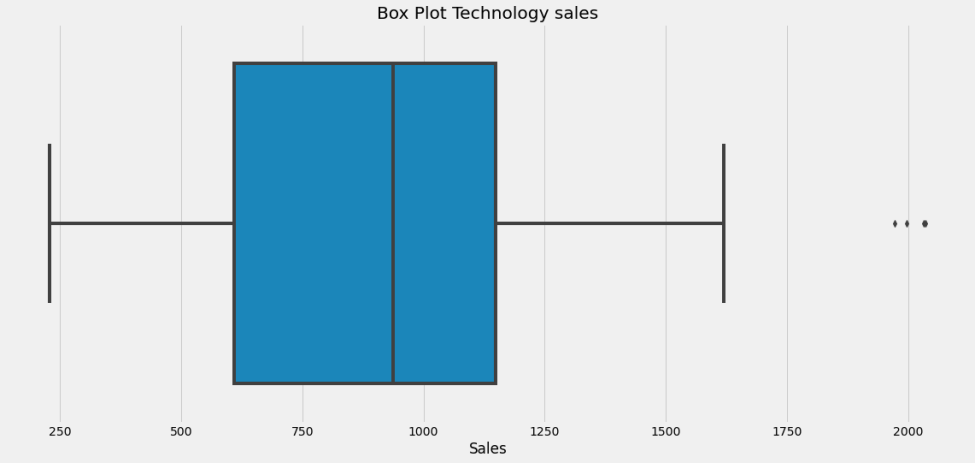
Boxplot on sales data of each product categories (Mean of monthly sales):



- In the furniture sales category, There are no outliers in the data.

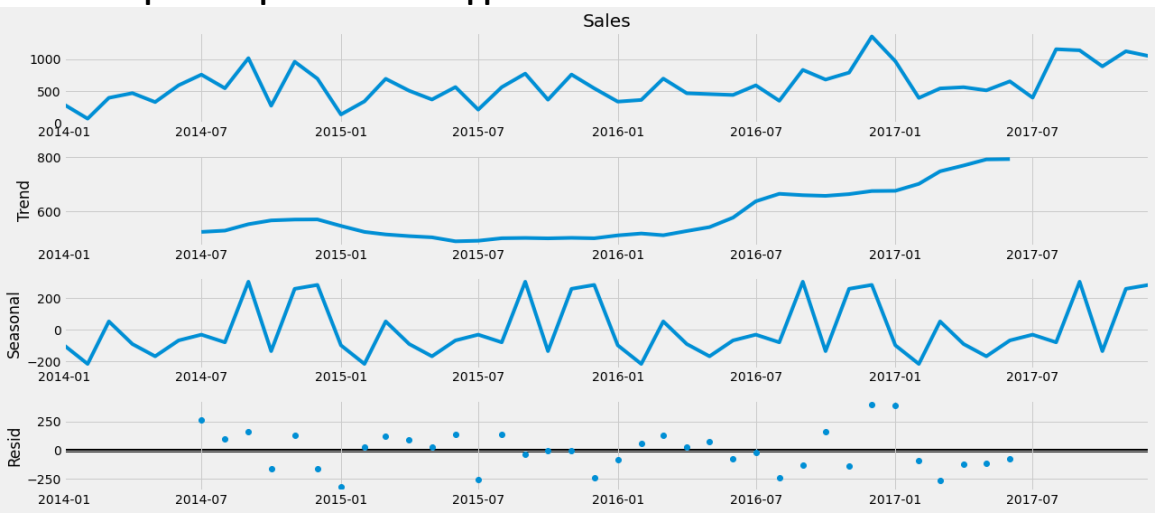


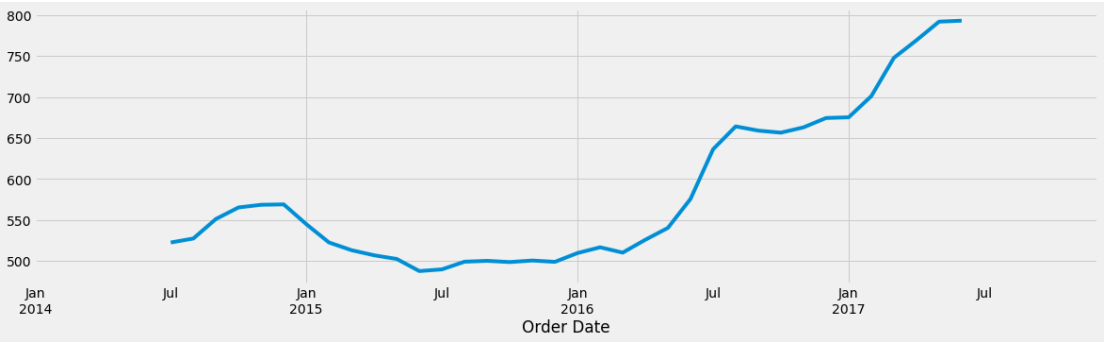
- In the office supplies sales category, There is one outlier in the data.



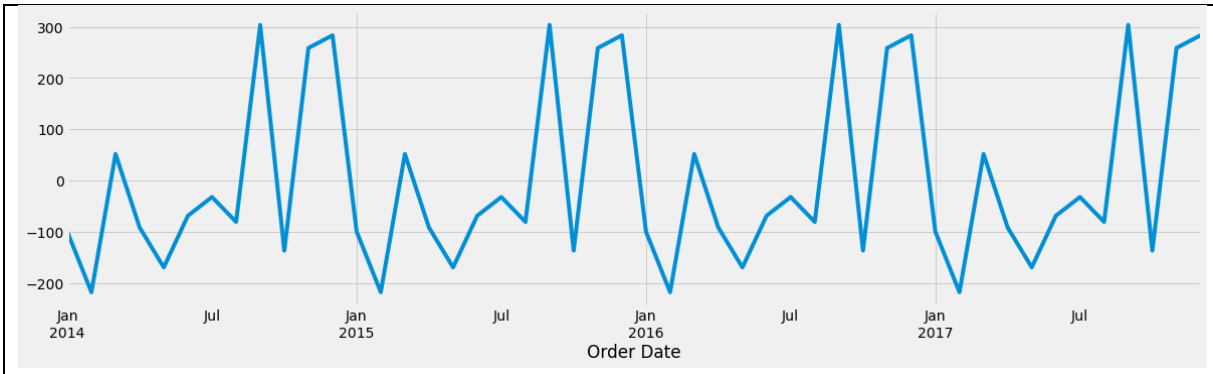
- In the technology products sales category, There are three outliers in the data.

ETS decomposition plot of office supplies sales:



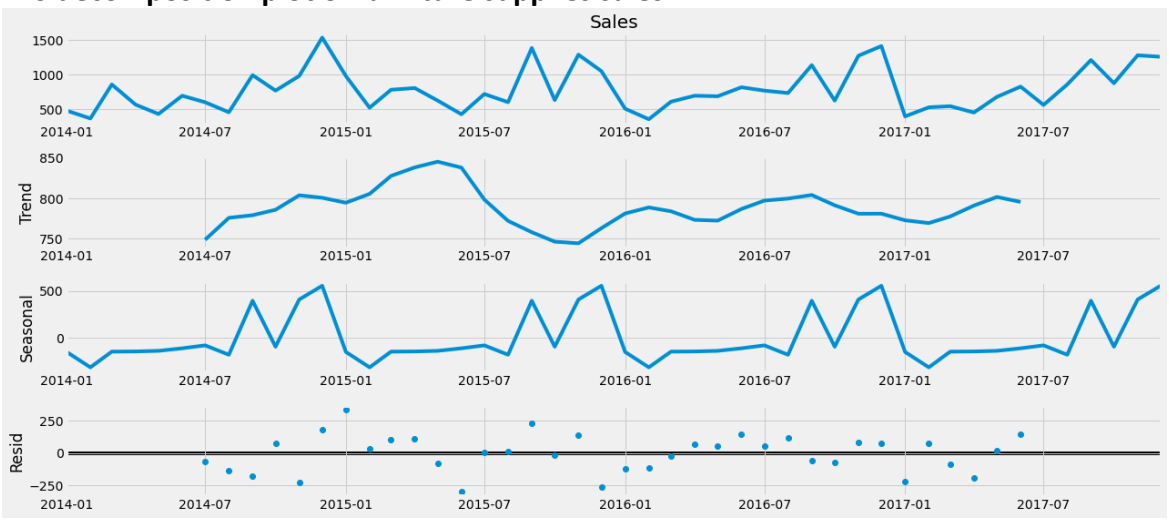


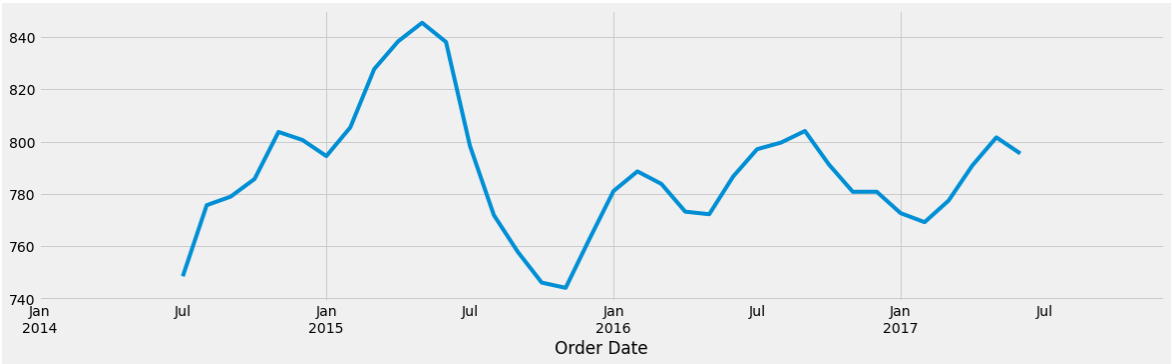
- In the above plot of trends, It can be seen that there is an increasing trend in the periods of July 2014 to December 2014. Then we can see a decreasing trend till June 2015. Then we can see an increasing trend till the end. The increasing trend from March 2016 till the end is very significant.



- In this seasonality plot of the office supplies sales category, we can see that more sales are occurring in the months of September, November and December. The least sales are occurring in the month of February.

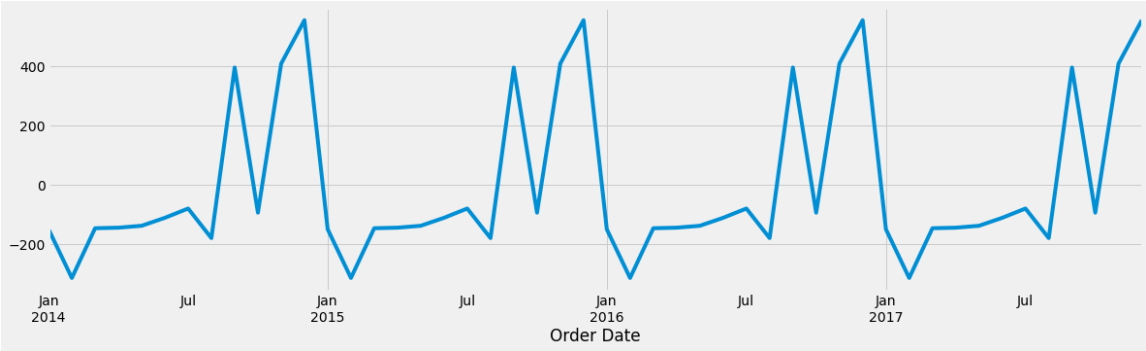
ETS decomposition plot of furniture supplies sales:





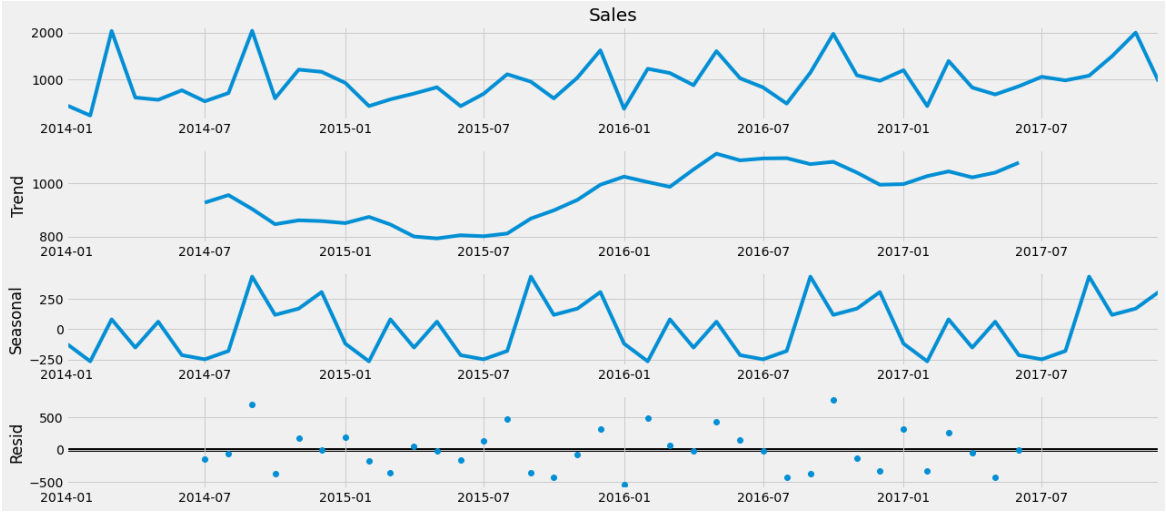
- In the above plot of trends, It can be seen that there is a significant increase in the sales of furniture supplies between July 2014 and November 2014. Then there is a small decrease till January 2015. Then there is a significant increase till May 2015. Then there is a sudden decrease till November 2015. Then there is an increasing trend till February 2016. Then there is a decreasing trend till May 2016. Then there is an increasing trend till September 2016. Then there is a decreasing trend till February

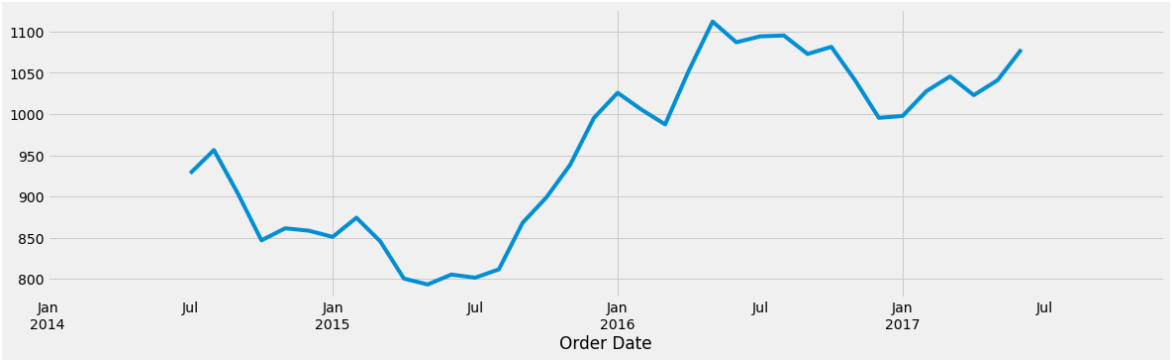
2017. Then the sales trend increases till June 2017. In the end, we can see a decreasing trend.



- From the above seasonality plot, we can say that the highest sales occur in the month of December and the lowest sales occur in the month of February. A stable number of sales are occurring between the months of March and July. Sales occurring in the months of September and October are also significantly higher.

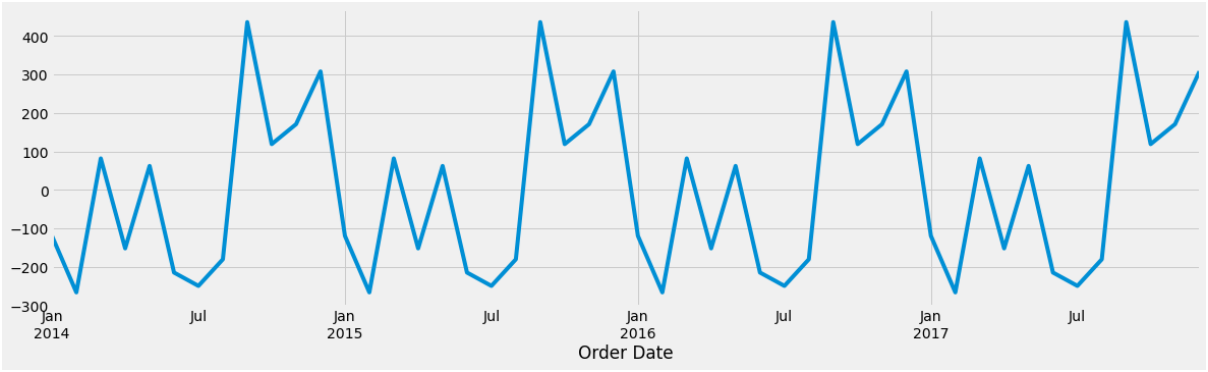
ETS decomposition plot of Technology products sales:





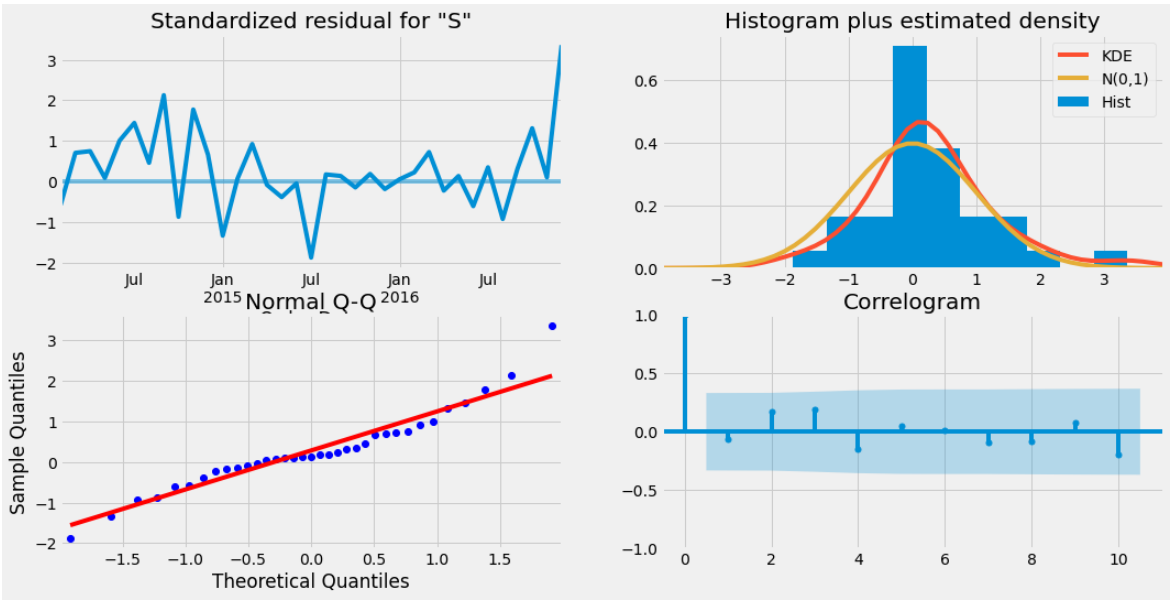
- From the above plot of trends, we can say that in the beginning one month there (July to August) there is a decreasing trend. Then there is a linearly decreasing trend till September 2014. Then we can see an almost neutral change till February 2015. Then we can see a decreasing trend till May 2015. Then there is an increasing trend till

January 2016. Then a sudden decreasing trend can be seen till March 2016. Then the trend increases linearly till May 2016. Then we can see a decreasing trend till December 2016. Then it slowly increases in the end.



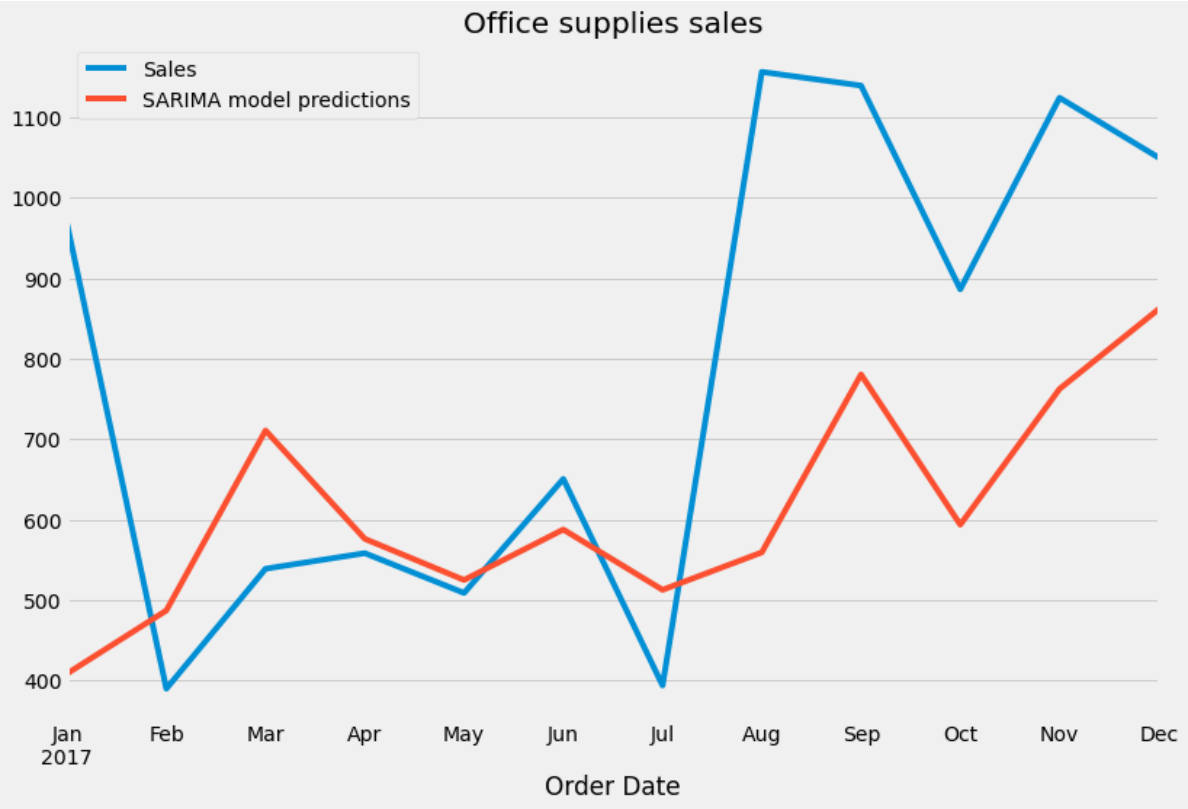
- In the above seasonality plot, we can see that there is a clear seasonality happening in the technology sales data.
- Usually, September marks the highest sales and in February happens the lowest sales. The period between September to December have better sales while comparing to sales happening in rest of the months.

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



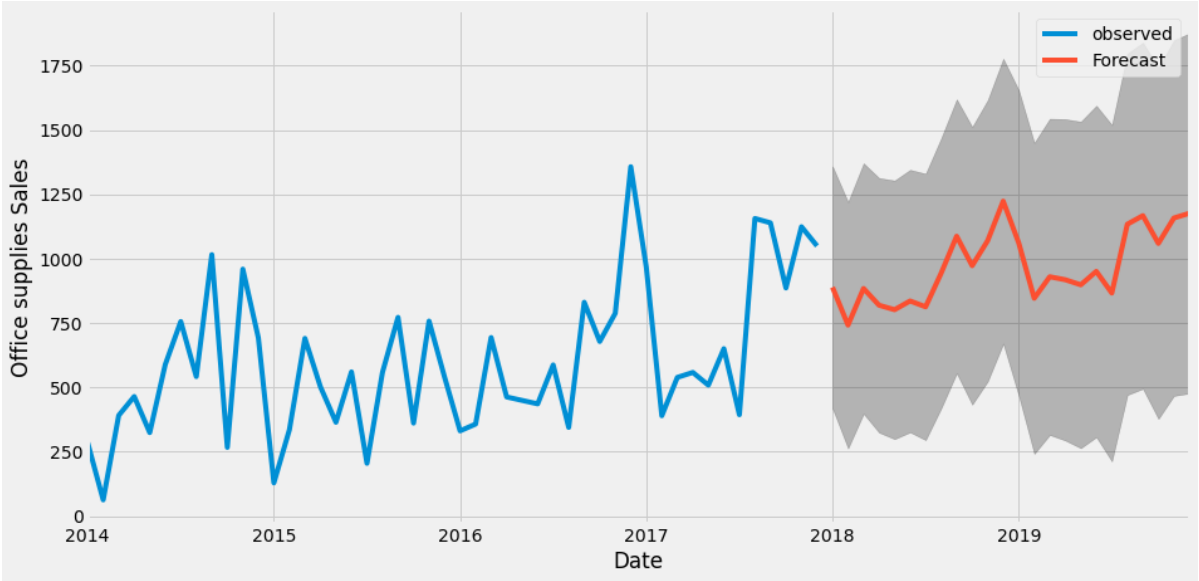
- From the above plots, we can say that our residuals are normally distributed.

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



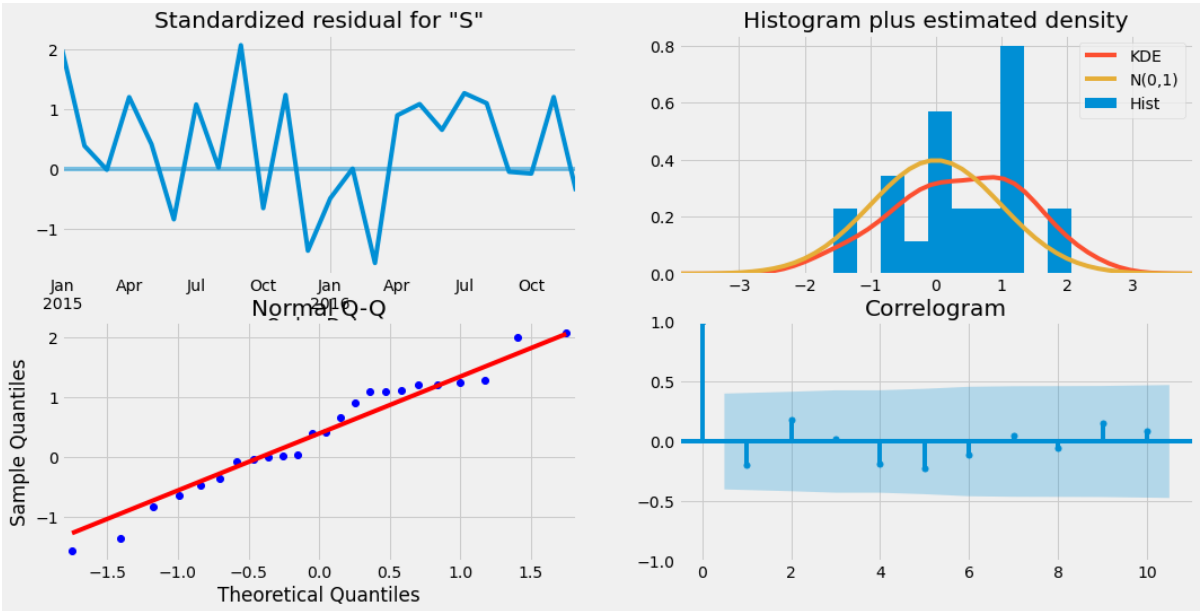
- From the above plot, we can say that we have a good model, which had predicted better in the months from February to July.
- For the months between August to December even though our model predicted the seasonality very well the real trend that happened was higher than that happened in the previous years. As a result, the predicted results are less than the observed results.

Plot obtained after making predictions to the unknown future for office supplies sales (predicted values for the next two years):



- In the above plot, I have plotted the forecasted values for the next two years (2018 and 2019).
- So, the plot tells us that the higher number of sales would be occurring in December for the year 2018 and in September and December for the year 2019.
- In the forecasting we can see that the sales are increasing better which tells us that the demand for office supplies will increase during these years.

Plot obtained after running model diagnostics test on the SARIMA model built for furniture sales data:

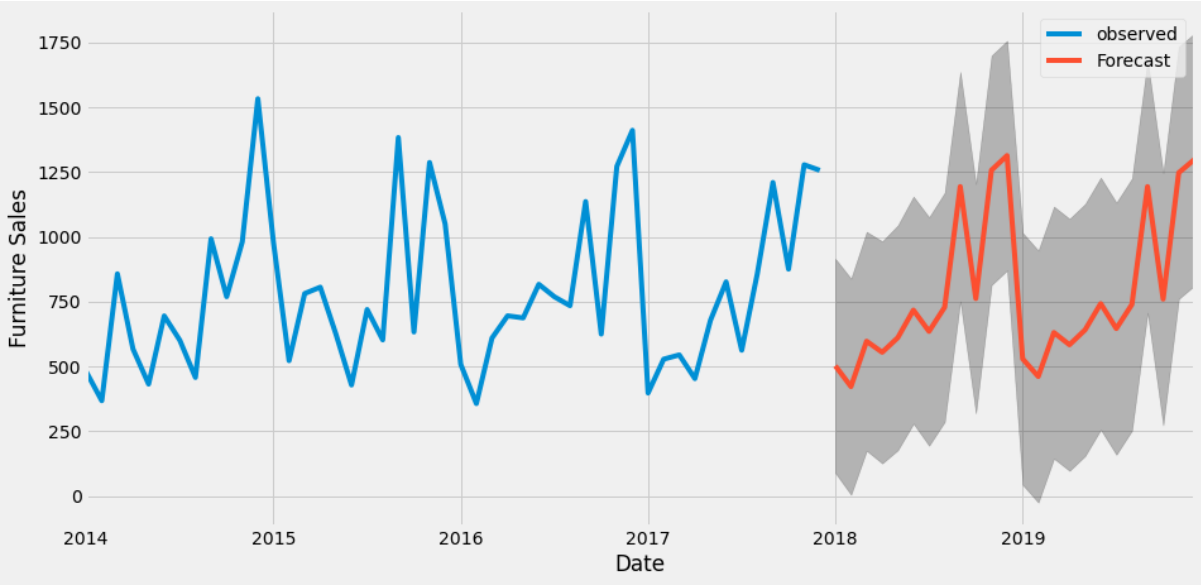


- From the above plots, we can say that our residuals are nearly normally distributed.

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

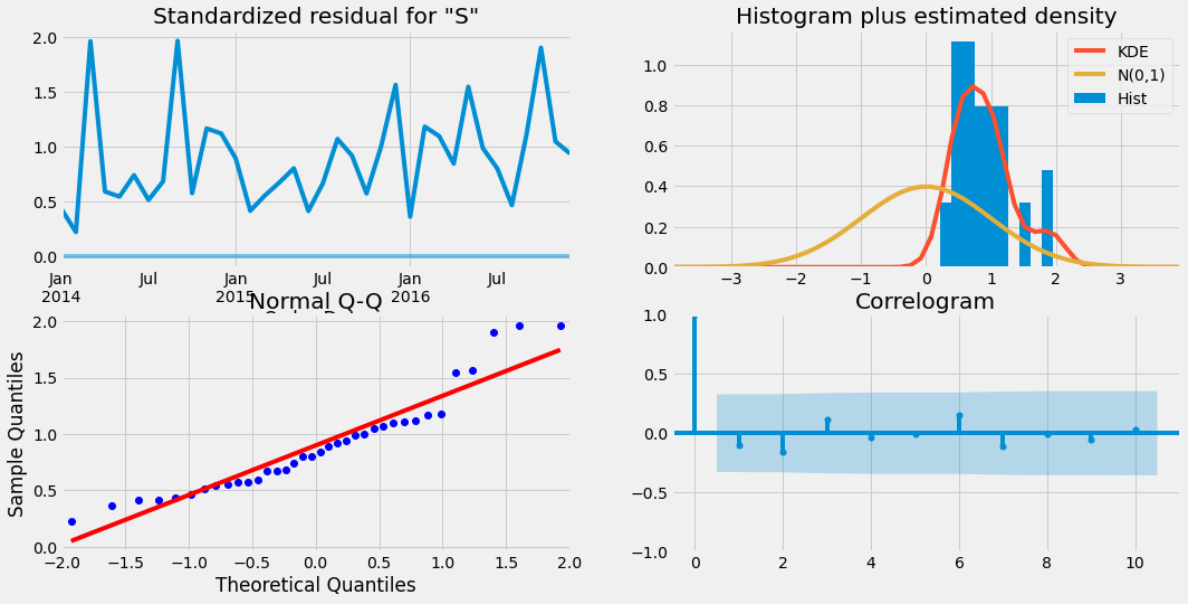
- From the above plot, we can say that our model had predicted the values really well.
- The predicted values are very near to the observed values in the months between July and December. So, we can say that we have a very good model.

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



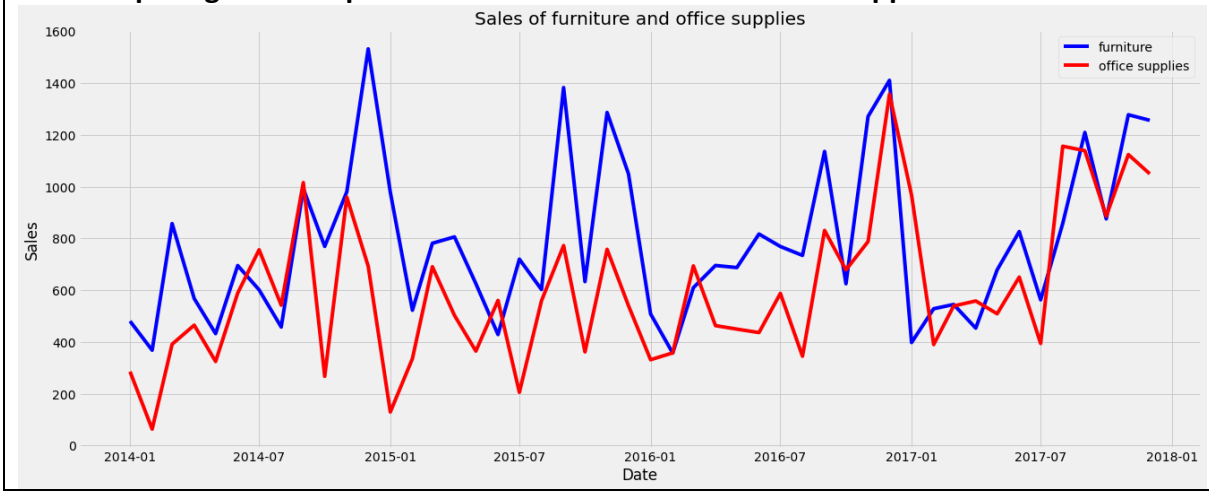
- Here I have plotted the data points including the forecasted values for the next two years also (2018 and 2019).
- From the plot, we can say that the sales will be higher during December of 2018 and 2019. So, during these months, demand for the furniture products will be higher.
- Also, we can say that sales during February is significantly lower and the demand for furniture products will be very less during this month.

Plot obtained after running model diagnostics test on the SARIMA model built for technology sales data:



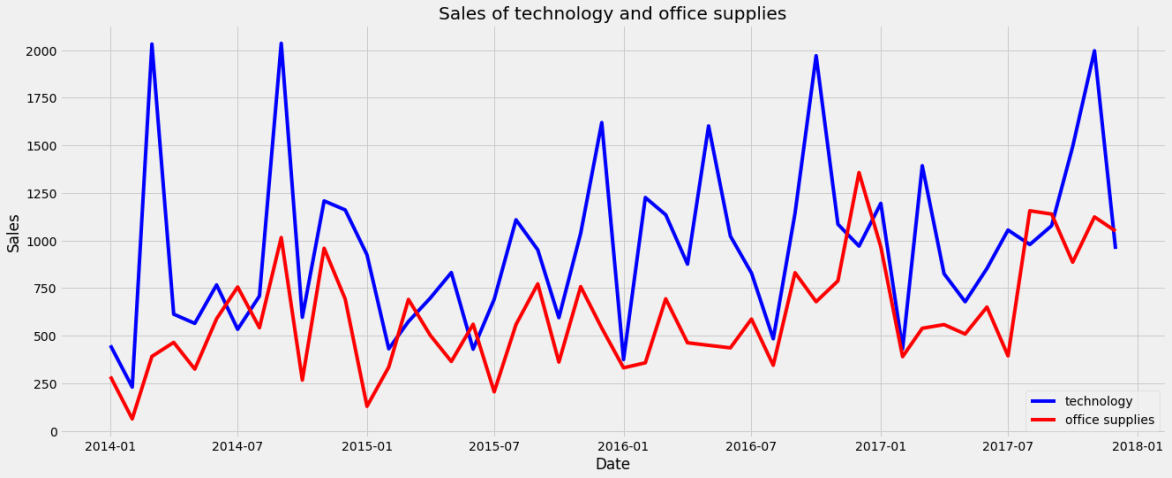
- From the above plots, we can say that our residuals are not normally distributed.
- Also, the residuals are very higher and they never reach zero.
- Also, from the correlogram plot, we can say that there is zero correlation.
- So here the time series model we got is white noise. So, we are not able to predict the values for the technology sales.

Plot comparing the data points in sales of furniture and office supplies sales:



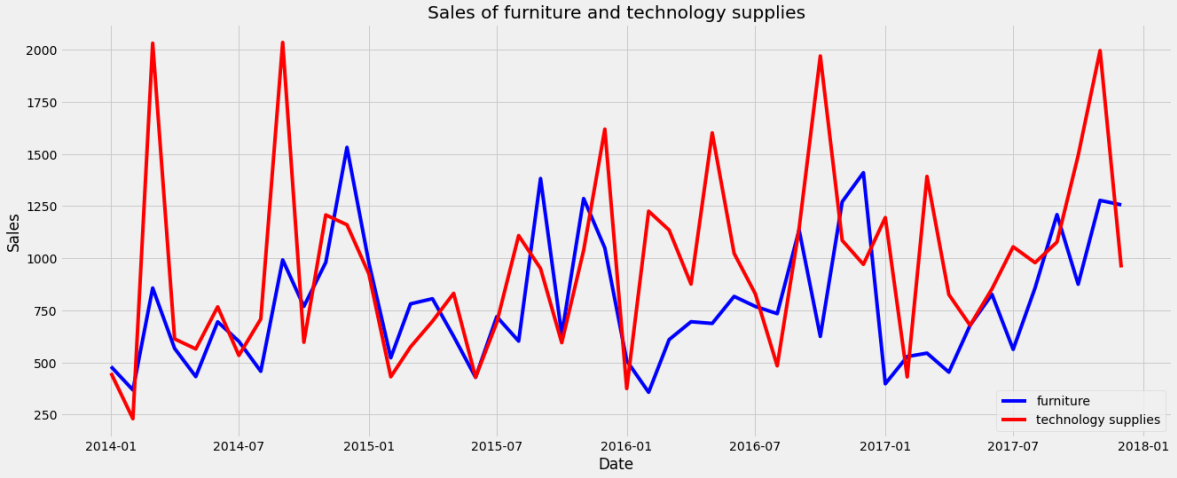
- From the above plot it is clear that for majority of the months, sales of the furniture products are higher than the office products.
- So, in this superstore, furniture products have higher demand than the office supplies.

Plot comparing the data points in sales of technology and office supplies sales:



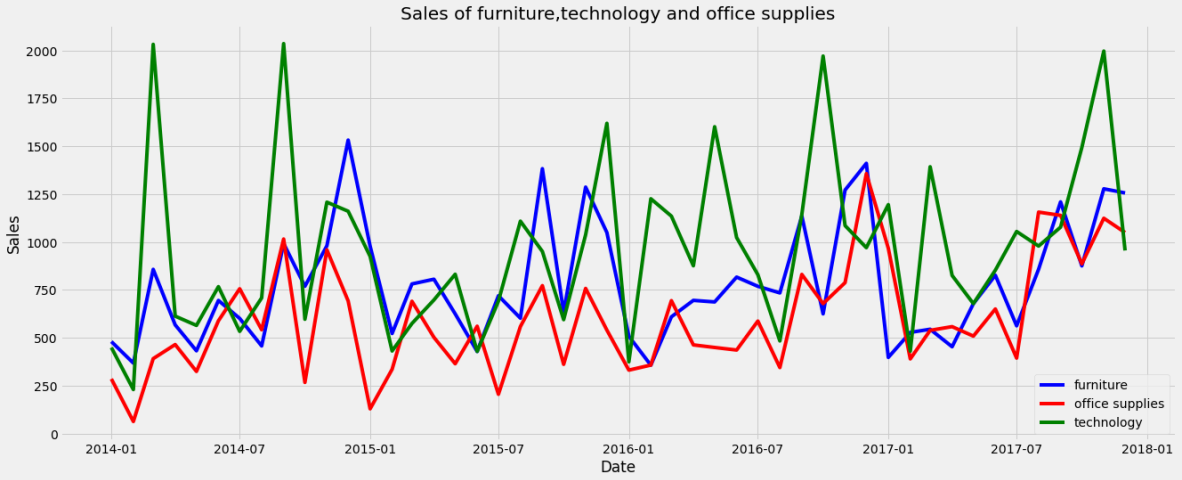
- From the above plot it is clear that for majority of the months, sales of the technology products are higher than the sales of the office supplies.
- So, in this superstore, technology products have higher demand than the office supplies.

Plot comparing the data points in sales of furniture and technology products sales:



- From the above plot it is clear that for majority of the months, sales of the technology products are higher than the furniture products.
- So, in this superstore, technology products have higher demand than the furniture products.

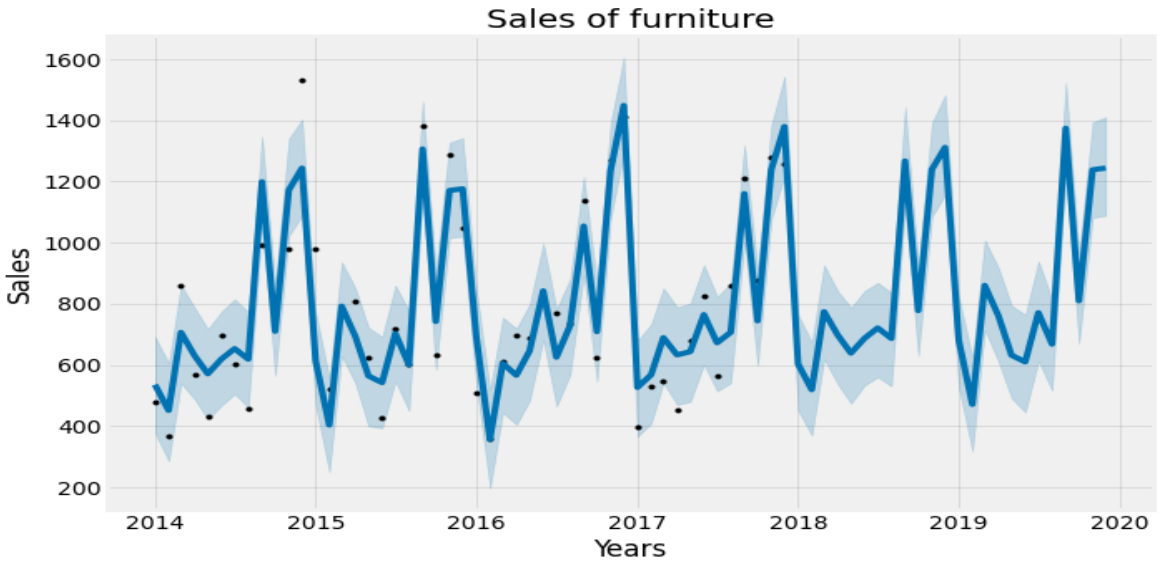
Plot comparing sales of all the three product categories:



- From the above plot it is clear that, in most of the times very higher sales are marked in the technology products sales. Then the next higher sales are occurring in the furniture products category. Sales of the office supplies category falls the last comparing to other two.
- So, in this superstore, more demand is for technology products than the other two product categories. Less demand is for office supplies.

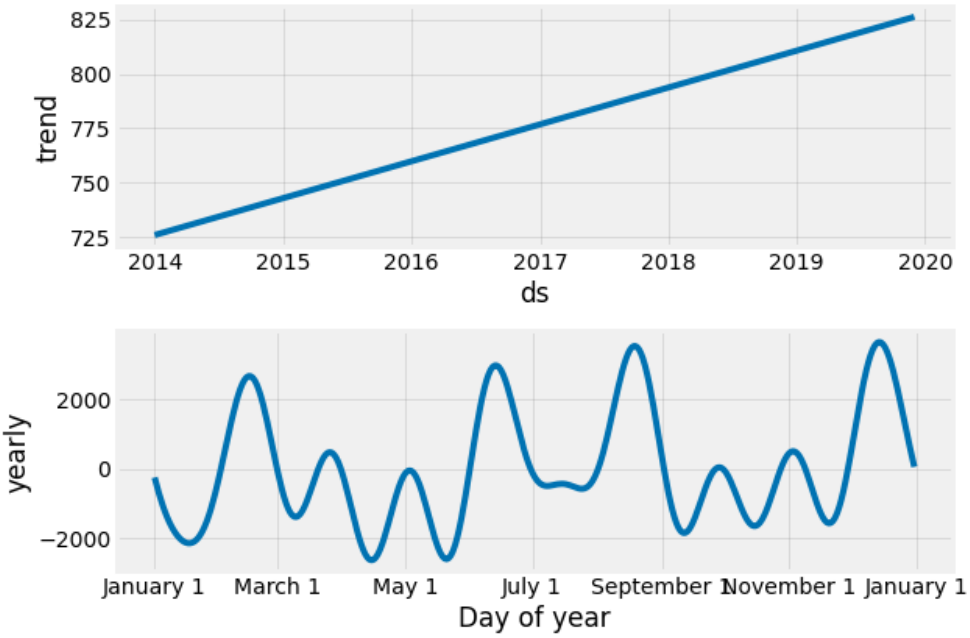
Plots made with the help of fbprophet library:

Plot made after forecasting sales values of furniture for next two years:



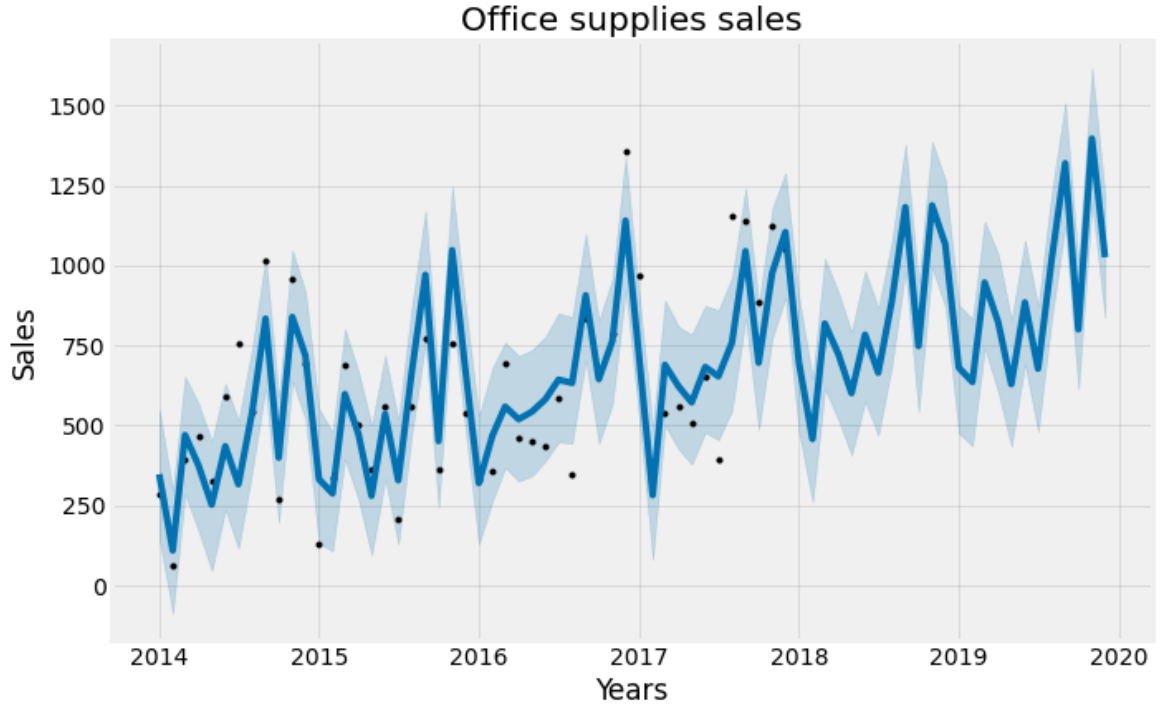
- In the predicted values, highest sales are predicted during the month of September 2019.
- In the predicted values lowest sales are predicted during the month of February 2019.
- We can see that in both the years 2018 and 2019, our model have predicted more sales during the months of September, November, and December while comparing to other months.

- Our model have predicted lowest sales in the month of February during 2018 and 2019.

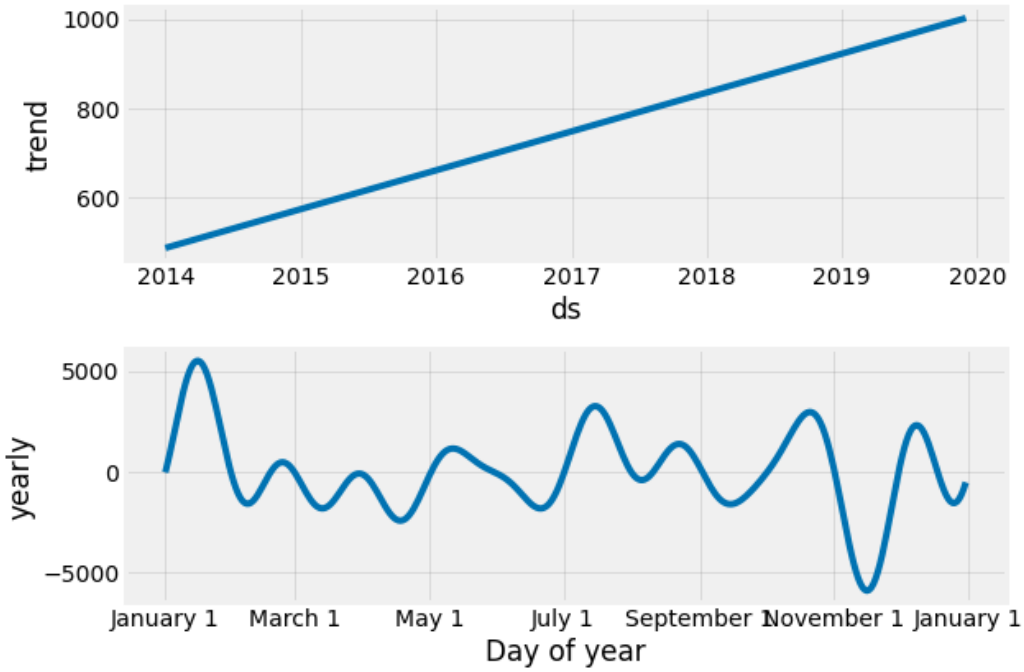


- From the plot of trends, it is clear that the trend is linearly increasing over the years. So, the sales of the furniture products are increasing over the years that means demand of the furniture products are increasing over the years.
- From the plot of yearly seasonality, it is clear that higher seasonality occurs during the middle part of December and in the middle part of August while the lowest seasonality occur during the middle part of January, middle part of April and the later part of May.

Plot made after forecasting sales values of office supplies for next two years:

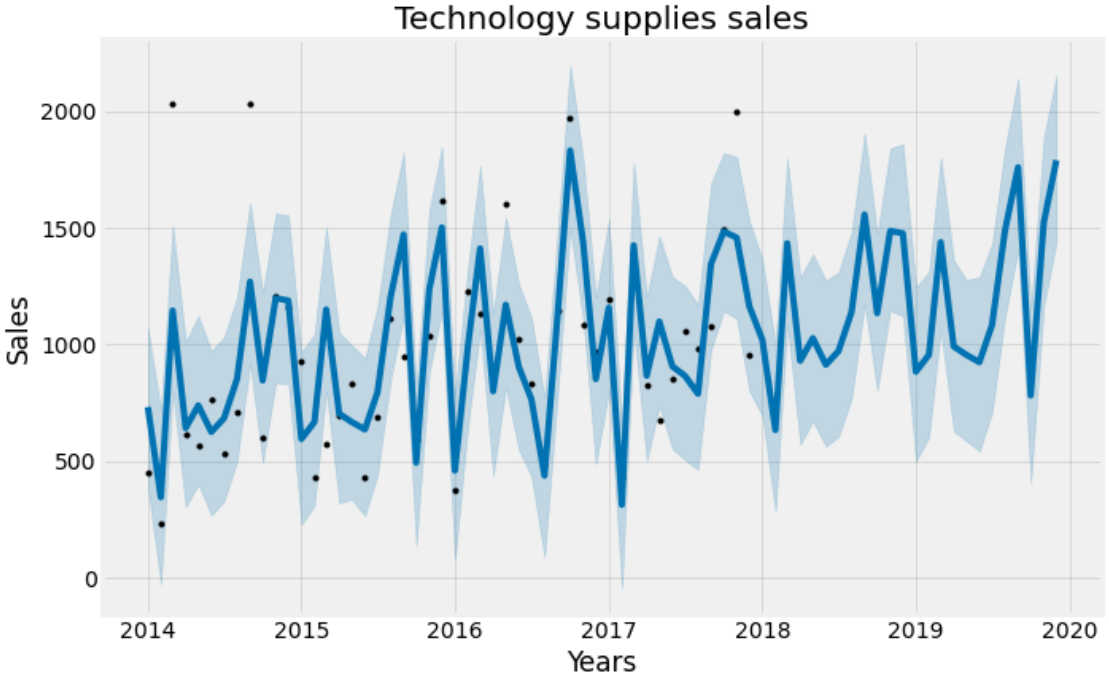


- In the predicted values highest sales are predicted during the month of November 2019.
- In the predicted values lowest sales are predicted during the month of February 2018.
- We can see that in both the years 2018 and 2019, our model have predicted more sales during the months of September and November while comparing to other months. But even in this case more sales are predicted during the year 2019.
- In 2018, lowest sales are predicted during February , while in 2019 lowest sales are predicted during May.



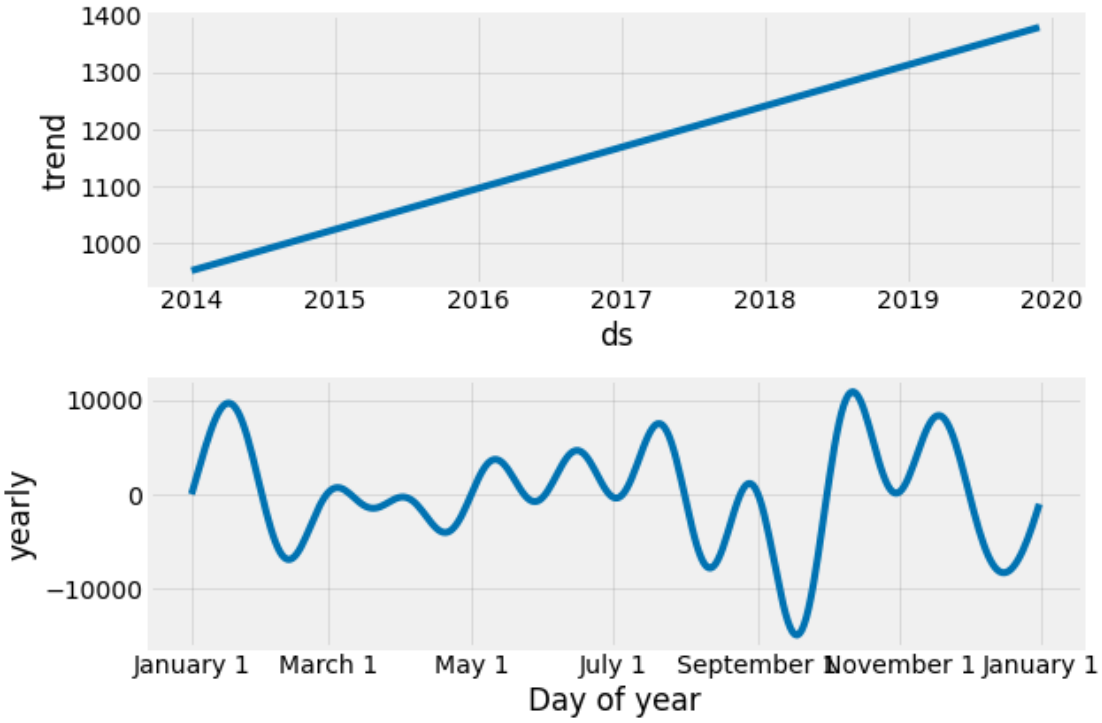
- From the plot of trends, it is clear that the trend is linearly increasing over the years. So, the sales of the office supplies are increasing over the years that means demand of the office supplies are increasing over the years.
- From the plot of yearly seasonality, it is clear that higher seasonality is during the middle part of January and lower seasonality is during the middle part of November.

Plot made after forecasting sales values of technology products for next two years:



- In the predicted values highest sales are predicted during the month of December 2019.

- In the predicted values lowest sales are predicted during the month of February 2018.
- We can see that in both the years 2018 and 2019, our model have predicted more sales during the months of March, September, November, and December while comparing to other months. But even in this case more sales are predicted during the year 2019.
- In 2018 the lowest sales are predicted during February while in 2019 lowest sales are predicted during October.



- From the plot of trends, it is clear that the trend is linearly increasing over the years. So, the sales of the technology products are increasing over the years that means demand of the technology products are increasing over the years.
- From the plot of yearly seasonality, it is clear that higher seasonality is during the beginning of October and the lower seasonality is during the middle part of September.

Challenges & Opportunities:

- Since the technology sales data is “white noise” I wasn’t able to create SARIMA based models on that.
- I was able to create SARIMA models for both office supplies sales data and furniture sales data.
- I also tried to build some deep learning models for this data, but later found SARIMA models have less root mean square error value than those deep learning models. So concluded that SARIMA models are better to apply in this dataset.

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:

- In the category of furniture sales, more sales are happening in the month of September. So during this month demand for these products are higher.
- In the category of furniture sales, fewer sales are happening in the month of February. So the manager should adopt any new business approaches to increase the sales during these months.
- In the category of furniture sales, we can expect the demand for furniture products will increase in the next two years.
- In the category of furniture sales, The sales may increase up to 1373 (expected during September 2019) and may decrease up to 471 (Expected during (February 2019) in the coming two years (2018 and 2019).
- In the category of office supplies sales, more sales are happening in the month of November. So during this month demand for these products are higher.
- In the category of office supplies sales, fewer sales are happening in the month of February. So the manager should adopt any new business approaches to increase the sales during these months.
- In the category of office supplies sales, we can expect the demand for furniture products will increase in the next two years.
- In the category of office supplies sales, The sales may increase up to 1396 (expected during November 2019) and may decrease up to 457 (Expected during (February 2018) in the coming two years (2018 and 2019).
- In the category of technology product sales, more sales are happening in the month of September. So during this month demand for these products are higher.
- In the category of technology product sales, fewer sales are happening in the month of February. So the manager should adopt any new business approaches to increase the sales during these months.
- In the category of technology product sales, we can expect the demand for technology products will increase in the next two years.
- In the category of technology product sales, The sales may increase up to 1790 (expected during December 2019) and may decrease up to 633 (Expected during (February 2018) in the coming two years (2018 and 2019).
- From all the models we can say that this superstore generally has low sales during the month of February. More sales usually happen in the months of September, November and, December.
- In this superstore, more demand is for technology products than the other two product categories. Less demand is for office supplies.

Link to code and executable file:

- Click here to open google colab file
- Click here to open my github repository
- Click here to open the loom video link