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 | 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 of Kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 11/04/2022 | 18/06/2022 | 125 hrs | Jupyter Notebook | Python 3 (Numpy, Pandas, Matplotlib, Statsmodels, Seaborn, pmdarima, scikit-learn, fbprophet) |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Challenges and Opportunities
* Reflections on the Internship
* Conclusions
* Link to code and executable file

**Acknowledgement**

I would like to express my deepest gratitude to Mr. Debashis Roy, my industry mentor, ICT Academy 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**

The dataset used for this project is a superstore dataset that has sales data of different products from 2011-01-07 to 2014-12-31. The product categories sold in this superstore dataset are furniture, technology, and office supplies. Project Process involves cleaning and sanitizing the dataset , stationarity check, non stationary to stationary transformation if needed, decomposition and finally Building forecasting system by 3 different approaches followed by its evaluation and predictions.

**Internship Activities**

* Gone through all the contents in welcome kit and day wise plan. .
* Attended webinar 1.
* Gone through YouTube videos provided as project reference materials such as ‘Forecasting Methods Overview’,’ Moving Averages’, ‘Time series forecasting’, and ‘ARIMA models’ given in the project reference material.
* Downloaded the dataset from : [Time Series Demo (Activity) | Coursera](https://www.coursera.org/learn/time-series-survival-analysis/supplement/yOmil/time-series-demo-activity)
* Started working on the dataset with Jupyter.
* Imported the dataset to jupyter notebook
* Imported necessary libraries.
* Visual inspection of raw data and brief summary of data is conducted including missing value detection
* Reduced the dataset into Order Date, Category and grouped by total Sales with Order Date made index of so reduced Data Frame.
* Extracted 3 different data frames of Furniture, Office Supplies and Technology wise total Sales with Order Date being index.
* Resampled as ‘Month Start’ and Visualized Furniture, Office Supplies and Technology separately
* Stationary Check process is done on respective Category
* Non stationary to stationary transformation is performed.
* Decomposition performed on sales data of each respective Category after stationary transformations.
* Created Forecast Model using 3 different approaches, its prediction and evaluation is done.

**Methodology**

**Data Understanding**

The dataset used for this project is a superstore dataset that has sales data of different products from 2011-01-07 to 2014-12-31([Time Series Demo (Activity) | Coursera](https://www.coursera.org/learn/time-series-survival-analysis/supplement/yOmil/time-series-demo-activity)). Deep understanding by visual inspection of raw data, dataset summary and missing value detection is done.

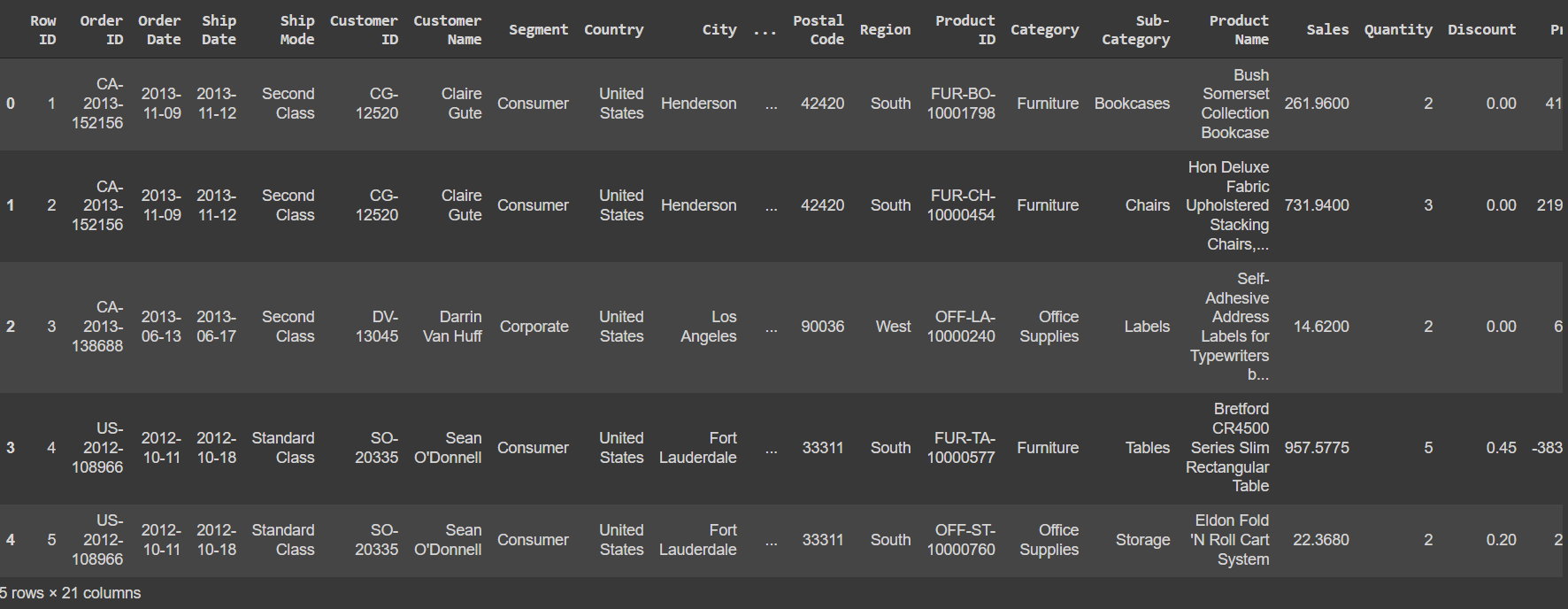


Fig: Raw Data

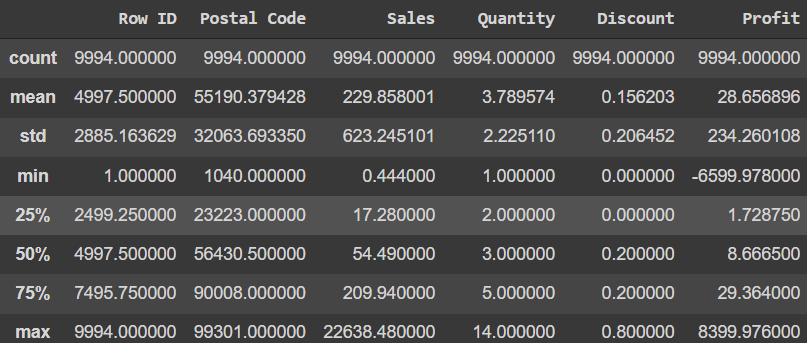


Fig: Summary of Data

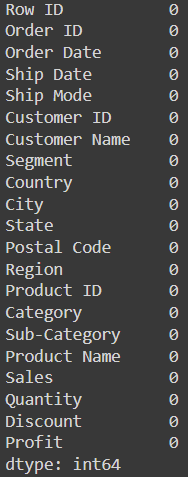


Fig: Missing Value detection

**Pre-Processing**

* **Simplifying Time series data:** We will start by simplifying the input data a bit to explore data types. To do so, we will look at Total Sales by Order Date and Category. This allows us to look a Time Series dataset with multiple time series.
* **Setting the index using existing variable:** We often want to set an Index explicitly, or manipulate an Index, for working with Time Series data. The Pandas DateTime Index is useful here, setting ‘Order Date’ as index.
* **Data Frame creation:** Individual Category wise Total Sales with Order Date as index is created**.**
* **Resampling:** The resample method generates a resampler object. To get to values, we need to specify an aggregation function if up sampling (moving to a lower frequency), or fill function if down sampling (moving to a higher frequency).Here Month Start denoted as ‘MS’ is given inside resampler object and visualized each category.

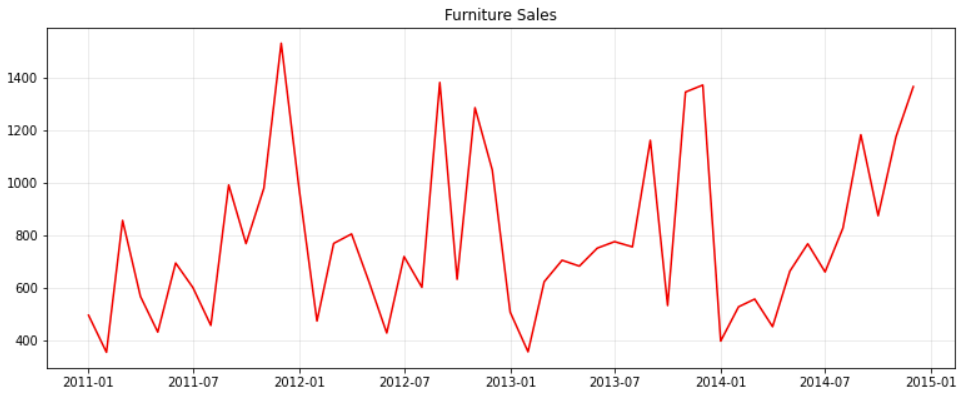
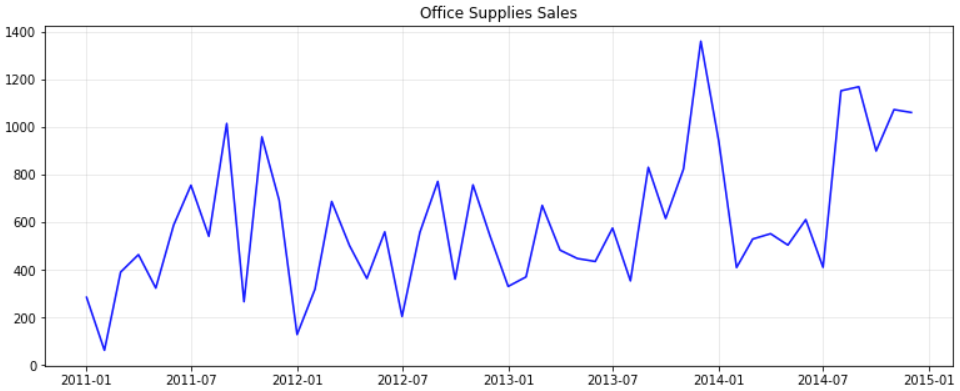


Fig: Furniture Sales

* Highest sales occurred between 2013-07 and 2014-01 and the lowest just after the month of 2011-01.

 Fig: Office Supplies Sales

* Highest sales occurred between last month after 2013-07 and 2014-01 and lowest just after month of 2011-01 and between 2011-07.

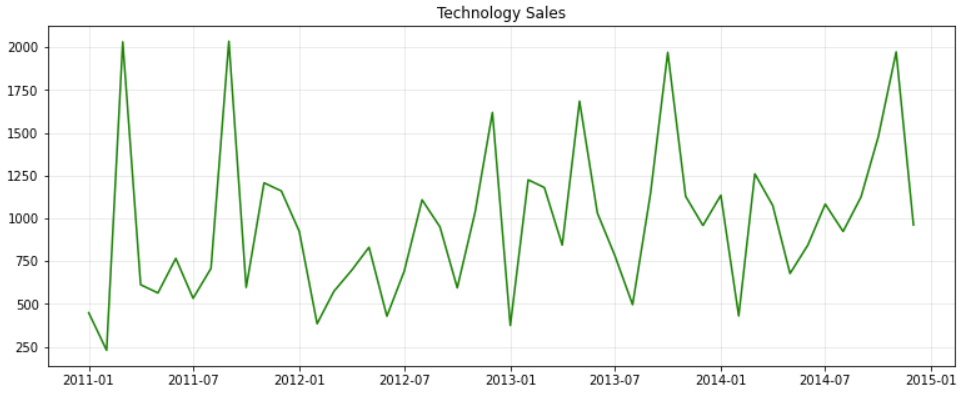
****

Fig: Technology Sales

* Highest sale occurred between last months after 2011-07 and 2012-01 and least just after month of 2011-01
* **Stationarity Check:** In order for time series data to be stationary, the data must exhibit four properties over time: constant mean, constant mean, constant autocorrelation structure, no periodic component. Three different approaches used for stationarity check are as follows.

1. Summary and statistics plots: The easiest way to check for constant mean and variance is to chop up the data into separate chunks, calculate statistics like mean and variance for each chunk, plot and evaluate.

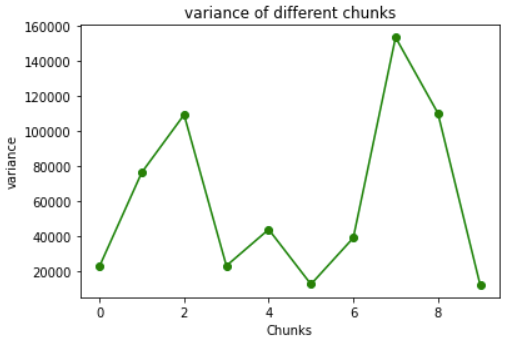
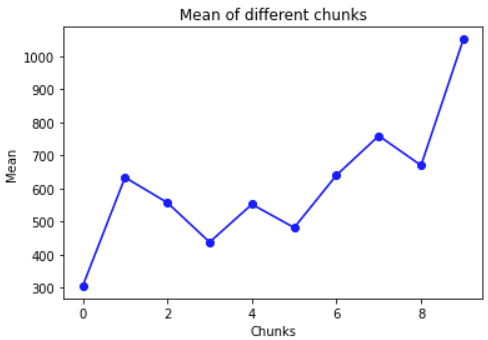


Fig1: mean and variance plot of Office Supplies Sales

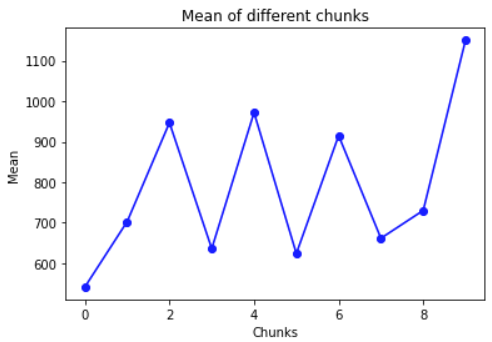
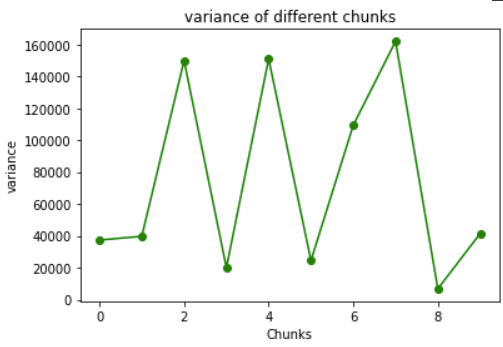
 

Fig2: mean and variance plot of Furniture Supplies

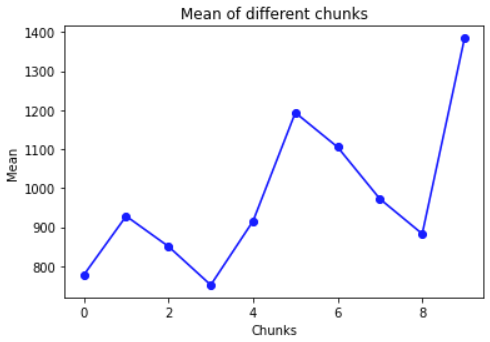
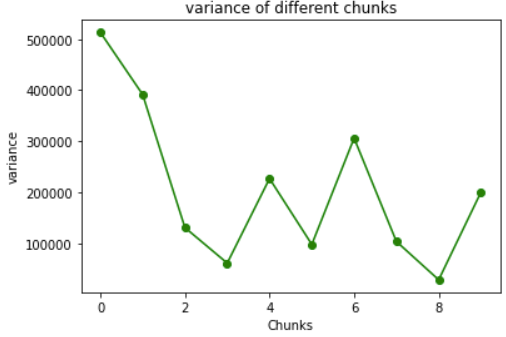
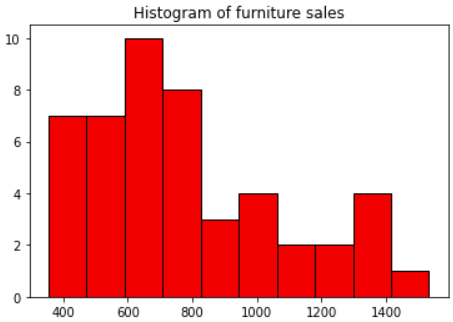
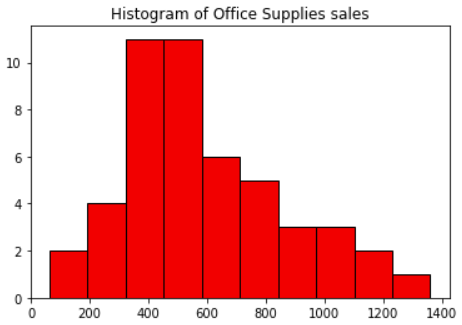
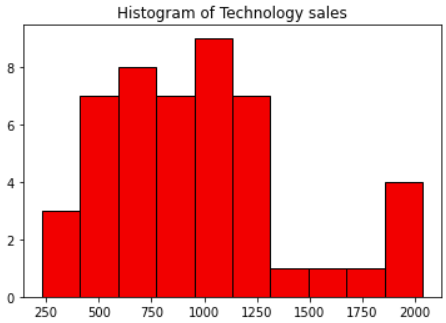
 

Fig3: mean and variance plot of Technology Sales

1. Histogram Plots: Plotting a histogram of the time series gives important clues into its underlying structure. A Normal distribution gives confidence that mean and variance are constant







1. Augmented Dicky-Fuller Test (ADF test): This is a statistical procedure to rule out whether a time series is stationary or not. Like any statistical test significance level of 0.05 is set.

**Null Hypothesis: Series is non stationary (p value > 0.05)**

**Alternate Hypothesis: Series is stationary**

ADF test for each Category Sales are as follows

* Furniture Sales: ADF: -5.6487, p-value: 9.9759x10^-07
* Office Supplies Sales: ADF: -4.5662, p-value:0.000149
* Technology Sales: ADF:-7.0638, p-value:5.1415x10^-10

**Conclusion of stationarity check:** Heteroscedasticity found in all Category’s Variance plot, No Normal distribution detected in any Category Distribution. Whereas ADF test shows stationarity for Furniture and Technology Sales and non stationarity in Office Supplies Sales. But Changing variation necessitated the need for non stationary to stationary transformation.

* **Non Stationary to Stationary Transformations:** by removing changing variation with log transformations. Natural log is applied to applied to all Category Sales data.
* **Decomposition:** After stationary transformation the data is decomposed into three components: trend, seasonal and residual. Here additive decomposition is applied to all Category Sales as seasonal component independent of trend is observed from resampled plot above. Stationary series is decomposed without trend and seasonality.

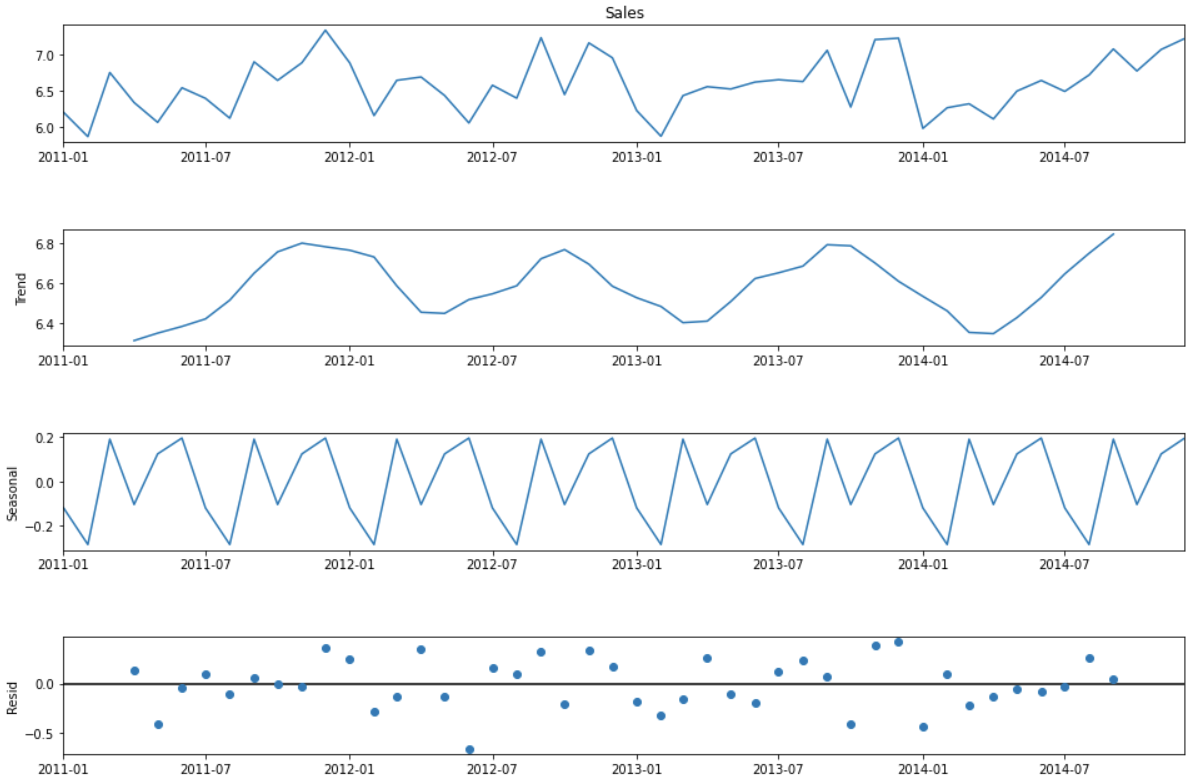
****

Fig1: Decomposition of Furniture Sales

* From the above plot of trend an significant increase in sales is observed after the month of July 2014 with sudden decrease in early months of 2014 until April
* From the plot of seasonality it has been observed that Highest sales occur during the month of June and December. Lowest sales are usually recorded over the month of February and August.

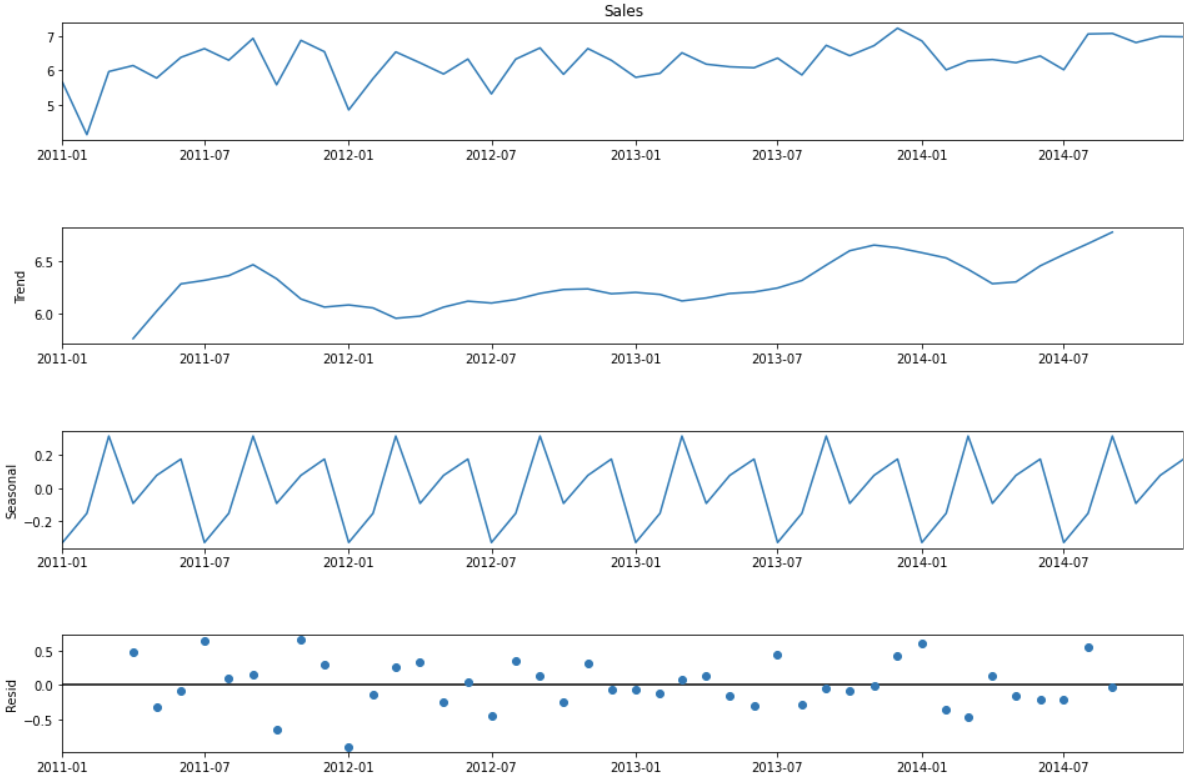
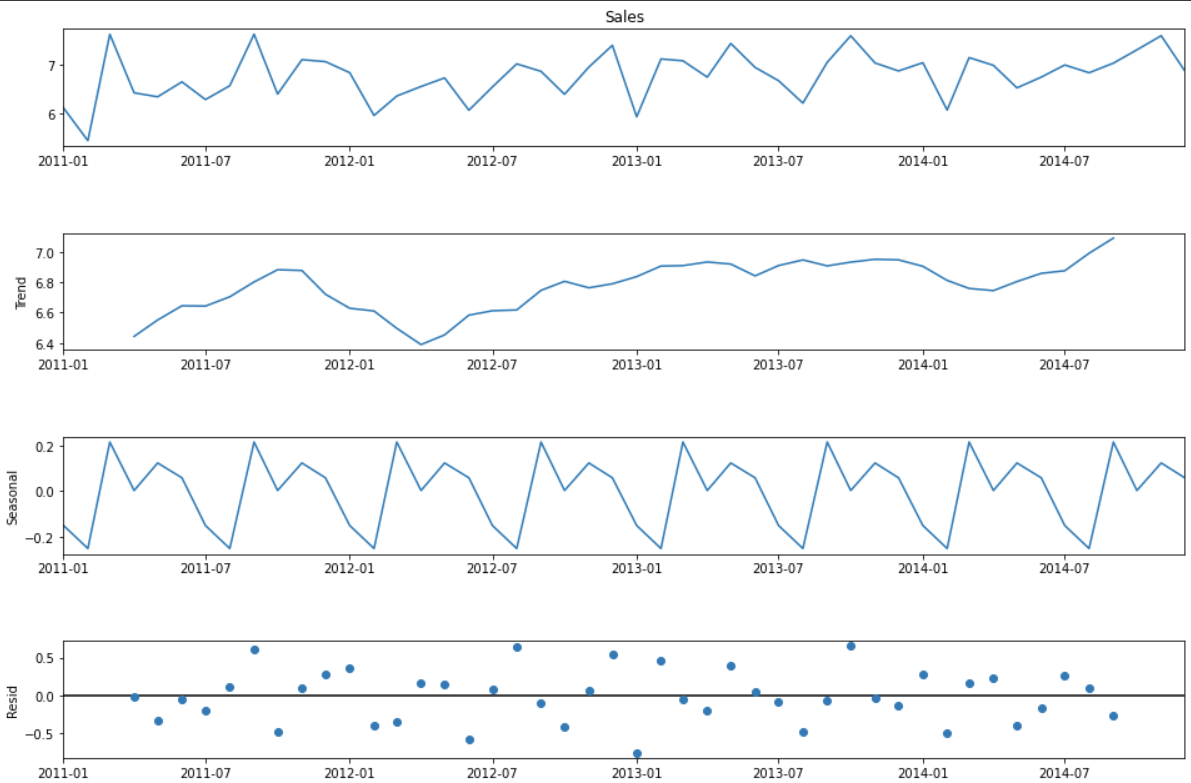


Fig2: Decomposition of Office Supplies Sales

* From the above plot of trend an significant increase in sales is observed after the month of July 2014 with sudden decrease in early months of 2014 until April.
* From the plot of seasonality, it has been observed that Highest sales occur during the month of February and August With least during the month of January and July.

Fig3: Decomposition of Technology Sales

* From the plot of trend, it is detected that significant increase in sales after the month of July 2014 with sudden decline during the months from February to April in 2012.
* From the Seasonality Plot, Highest Sales is observed during the months of March and September every year with least during the months of February and August.

**Forecasting Model**

Final stage involves Building Forecasting system of Category wise total Sales of Super store data. Predictive analysis is done by 3 different Model Approaches:

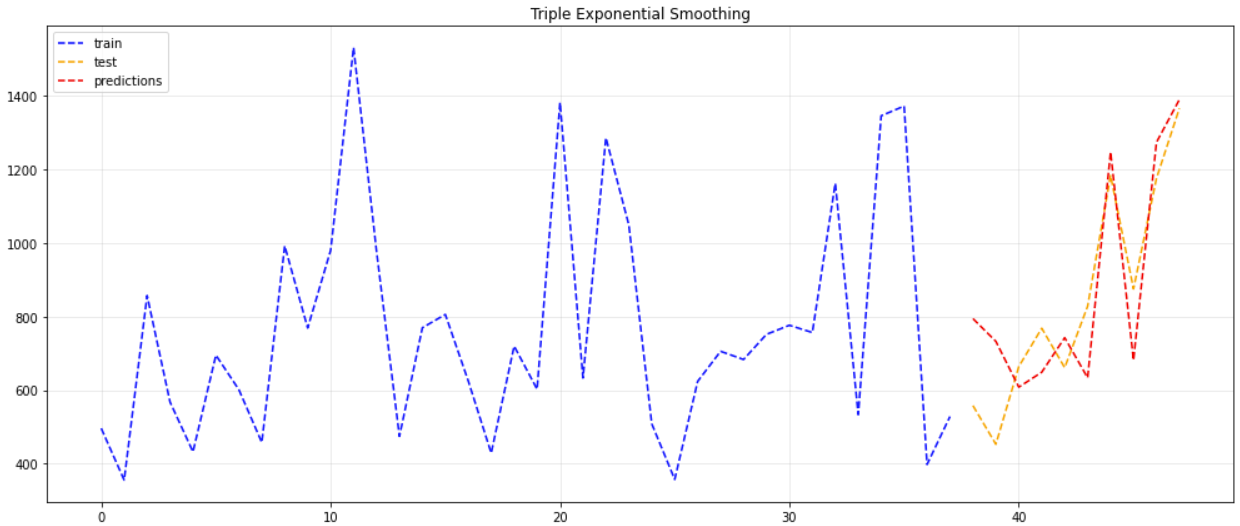
* Triple Exponential Smoothing
* SARIMA
* Facebook Prophet

**1.Triple Exponential Smoothing**

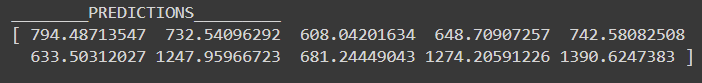
Exponential smoothing is a way to weight observations differently. Specifically, recent observations are weighted more than more distant ones. This makes intuitive sense. With three types of exponential smoothing Triple exponential smoothing is one that can pickup both trend and seasonality i.e. it has the benefits of double Exponential smoothing with ability to pickup on Seasonality.

**Procedure**:

* Create a time component as the length of Category Sales data
* Train-Test split: Test size is taken as 10 observation (20% of 48 instances)
* Model: Imported from library statsmodels with trend and seasonal components by additive decomposition
* Best possible seasonal period is applied in the exponential smoothing model and forecast is made.
* Plotted test, train and forecasted prediction
* The predictions are evaluated by the Means of Mean Squared Error (MSE) and Root Me
* **Furniture Sales**

Fig: Furniture sales; Triple Exponential Smoothing

* Nearly good predictions are made Good but its not exactly accurate.
* Evaluation: The predictions are made and the evaluation is done by the means of Mean Squared Error and Root Mean Squared Error



* MSE: 24843.3624
* RMSE: 157.6177
* **Office Supplies Sales**

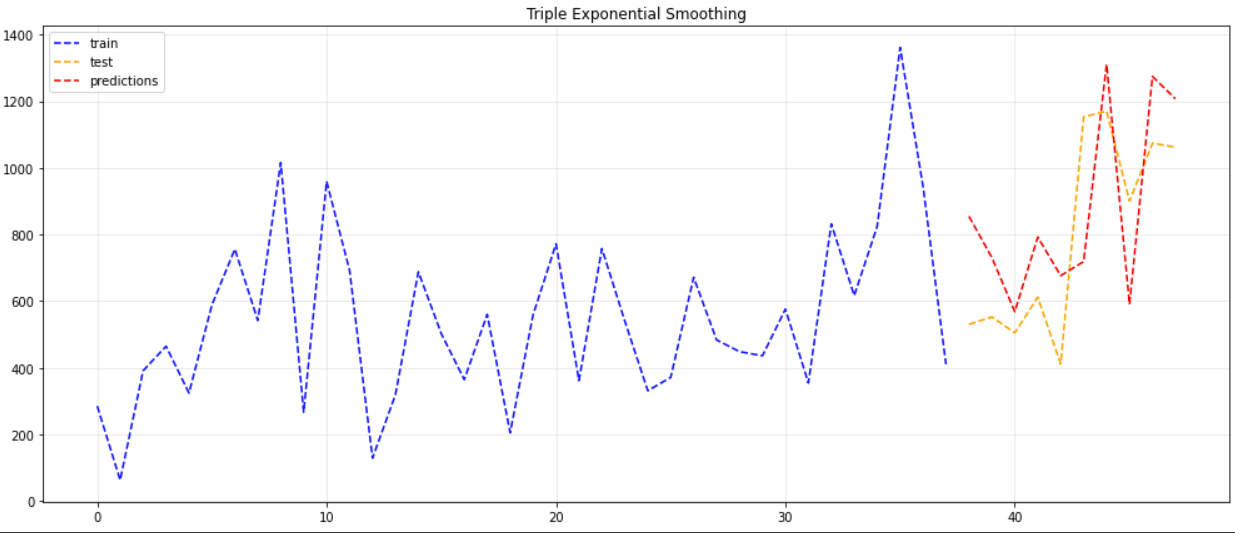
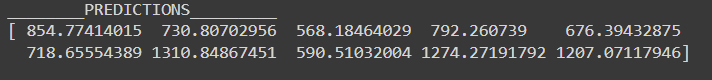


Fig: Triple Exponential Smoothing

* The Model is able to produce nearly good predictions but the model is unable to capture the volatility of Sales well enough
* Evaluation:



its done by the means of Mean Squared Error and Root Mean Squared Error

* MSE: 60917.16005
* RMSE: 246.8140
* **Technology Sales**

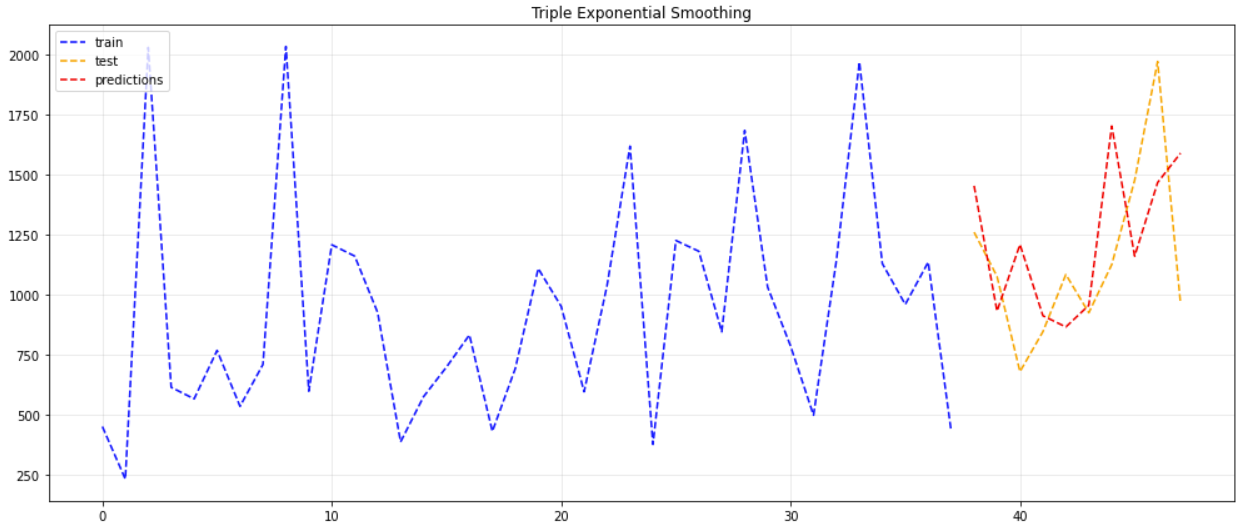
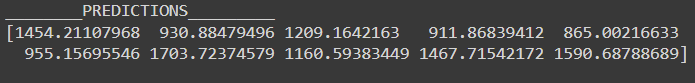


Fig: Technology Sales; Triple exponential smoothing

* The graph besides shows that model is unable to capture the volatility of Sales well enough.
* Evaluation: Evaluation is done by the means of Mean Squared Error and Root Mean Squared Error.



* MSE: 147965.3800
* RMSE: 384.36627

**2.SARIMA Model**

**MA models**: Specify that the current value of the series depends linearly on the series' mean and a set of prior (observed) white noise error terms.

**AR models**: Specify that the current value of the series depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term).

 In **ARIMA** models, the I stands for "Integrated" which just refers to the amount of differcing done on the data.

When we are determining our ARIMA model we will come across the following standard inputs:

* order(p,d,q):
  + p is number of AR terms
  + d is number of times that we would difference our data
  + q is number of MA terms

When we work with SARIMA models 'S' refers to 'seasonal' and we have the additional standard inputs:

* seasonal order(P,D,Q):
  + p is number of AR terms in regards to seasonal lag
  + d is number of times that we would difference our seasonal lag (as seen above)
  + q is number of MA terms in regards to seasonal lag
  + s is number of periods in a season

**Procedure**:

* **Auto Arima**: Python has an auto ARIMA that grid searches/optimizes our model hyperparameters for us. Over time, more of these goodies are porting to Python (e.g. pmdarima). This library contains an auto\_arima function that allows us to set a range of p,d,q,P,D,and Q values and then fit models for all the possible combinations. Then the model will keep the combination that reported back the best AIC value.
* Parameters with lowest AIC is then fitted to SARIMA Model.
* Plot is obtained after running model diagnostics test on the SARIMA model built for office data.
* Plot is obtained after making predictions about the known data and comparing it to the data points in the test set of Category sales
* Plot obtained after making predictions to the unknown future for Category sales.
* **Furniture Sales**

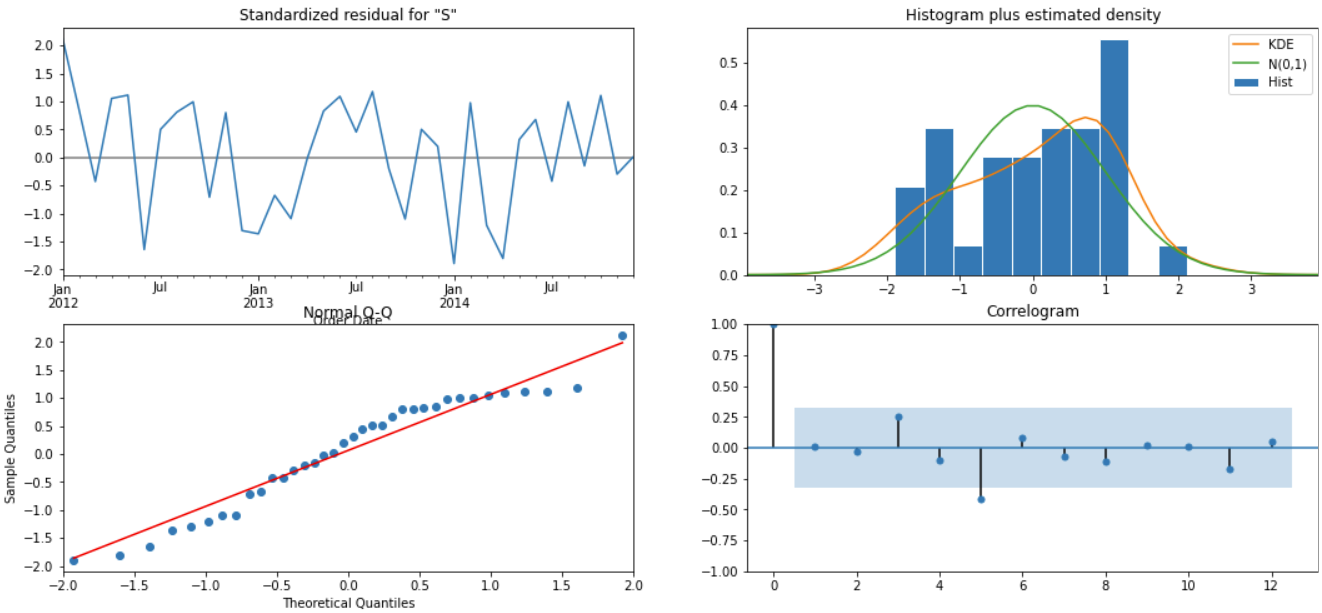


Fig: Diagnostics plot showing residual distribution

* Residuals are not normally distributed as observed from the diagnostics plot above.

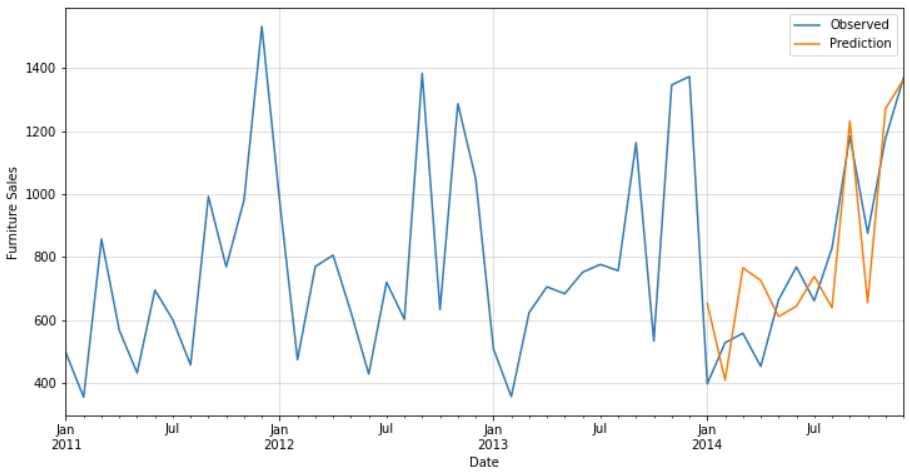


Fig: Prediction Plot

* Predictions are made along the test starting from January 2014 until the year end.
* Nearly a good model is created with some what near prediction during the month from May to September 2014.

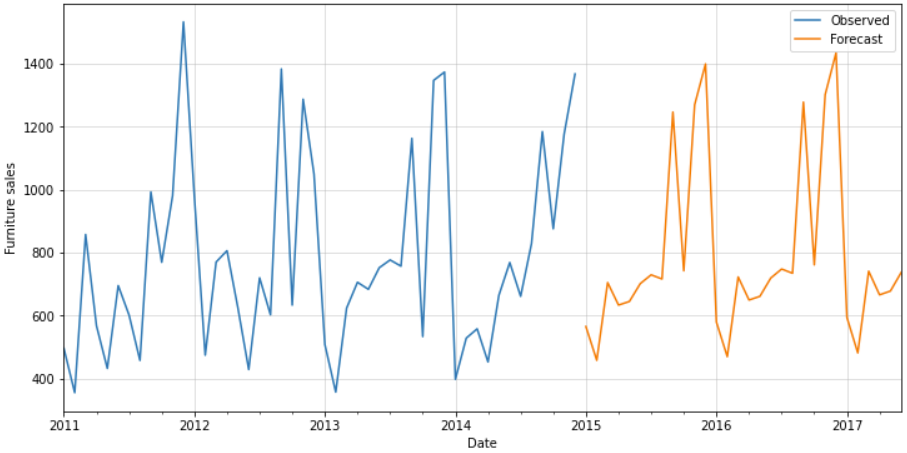


Fig: Forecast Plot

* The above plot shows the forecast that has been made for next 30 months i.e. two and half years
* The Forecast plot shows slight increase in sales over the years with Higher the demand for Furniture in the months of December and least in the months of February

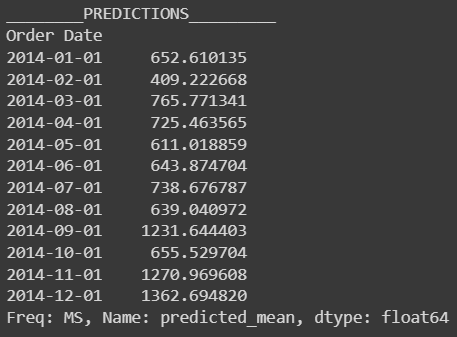
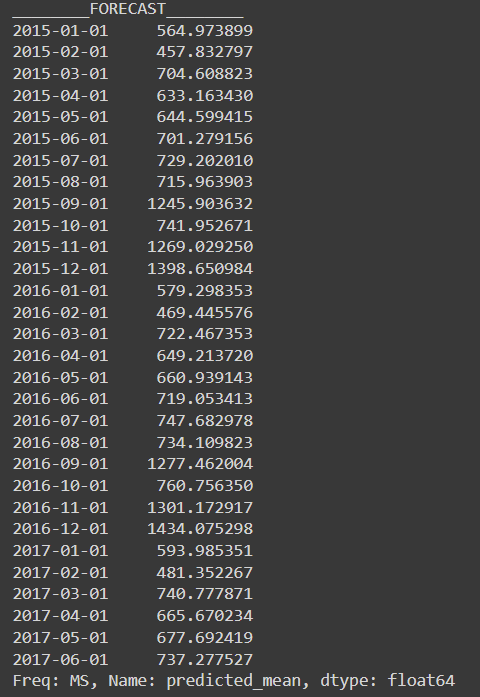
 

Fig: Predictions and Forecast

Evaluation:

Its done by the means of Mean Squared Error and Root Mean Squared Error

* MSE: 26390.2041
* RMSE: 162.4506
* **Office Supplies Sales**

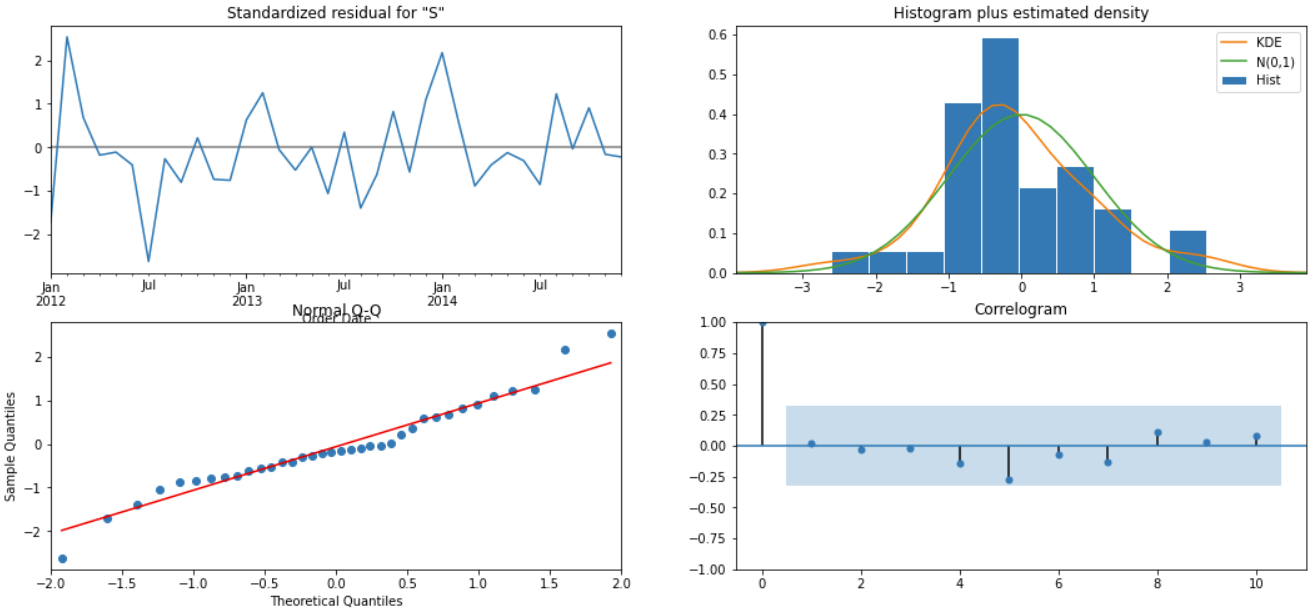


Fig: Diagnostic plot

* Near normal distribution of residual is observed from the diagnostic plot.

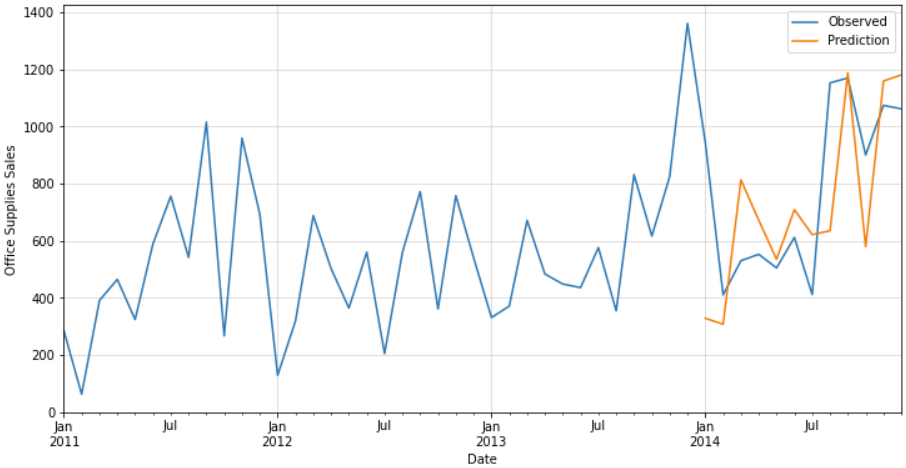


Fig: Prediction Plot

* Predictions are made along the test starting from the year 2014 until the year end.
* Model is able to capture the volatility of sales even if it does not produce exact accurate predictions.

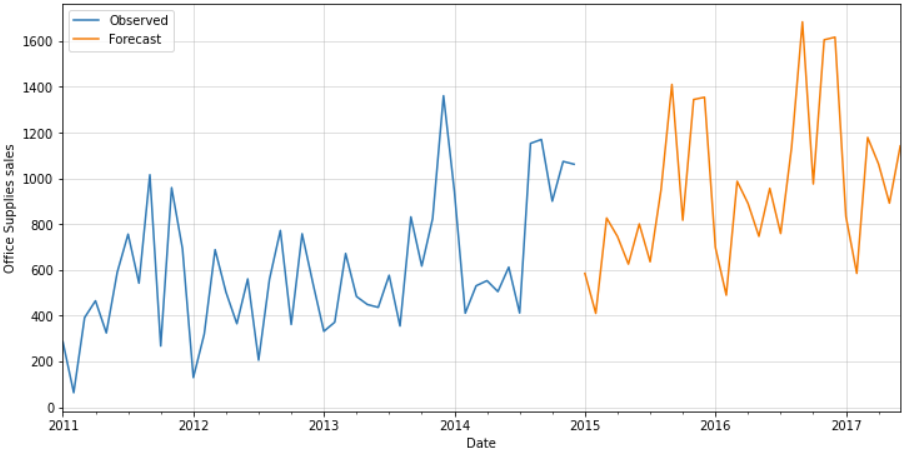


Fig: Forecast Plot

* The above plot shows the forecast that has been made for next 30 months i.e. two and half years
* The Forecast plot shows increase in sales of Office Supplies over the year(upward trend) over the years with Higher the demand for Office Supplies in the months of September and least in the months of February

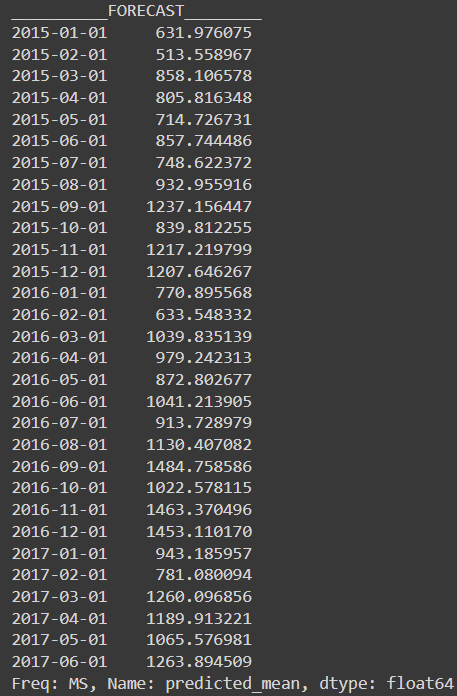
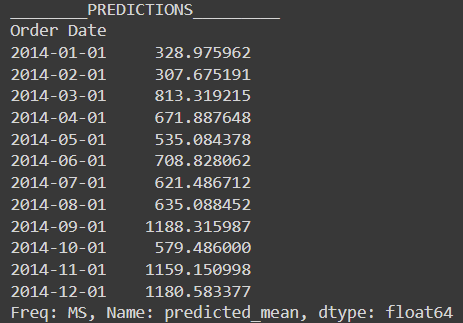


Fig: Predictions and Forecast

* Evaluation: Its done by the means of Mean Squared Error and Root Mean Squared Error
* MSE: 77472.6044
* RMSE: 278.3390
* **Technology Sales**

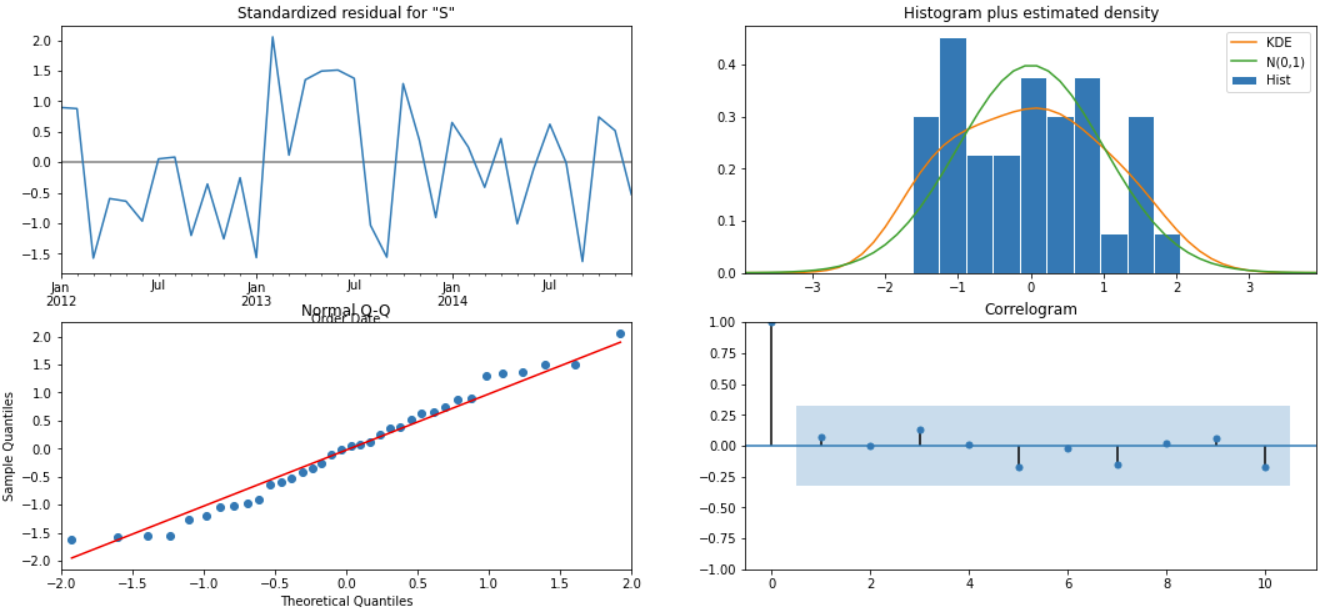


Fig: Diagnostic Plot showing residual distribution

* From the diagnostic plot residuals are not normal distribution is observed.

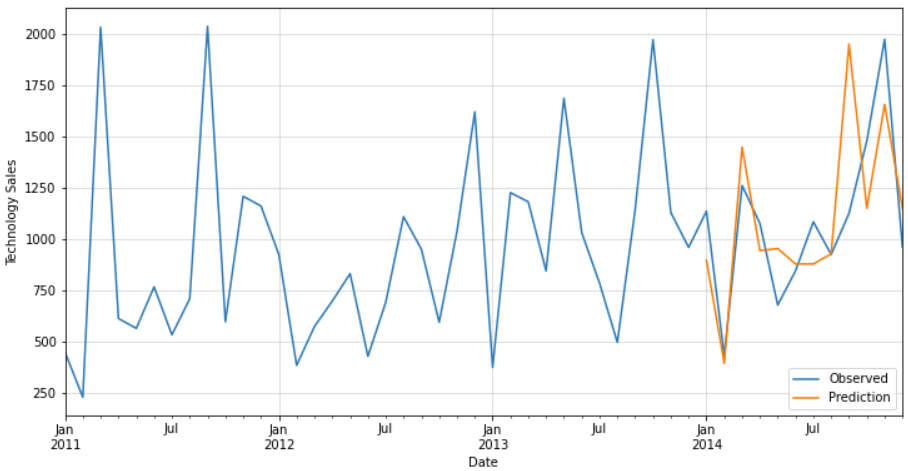


Fig: Prediction Plot

* Predictions are made from the month of January to December throughout the year 2014 along observation. Predictions are denoted by orange line
* The predictions some times failed to capture the volatility of Technology Sales although it’s a good model.

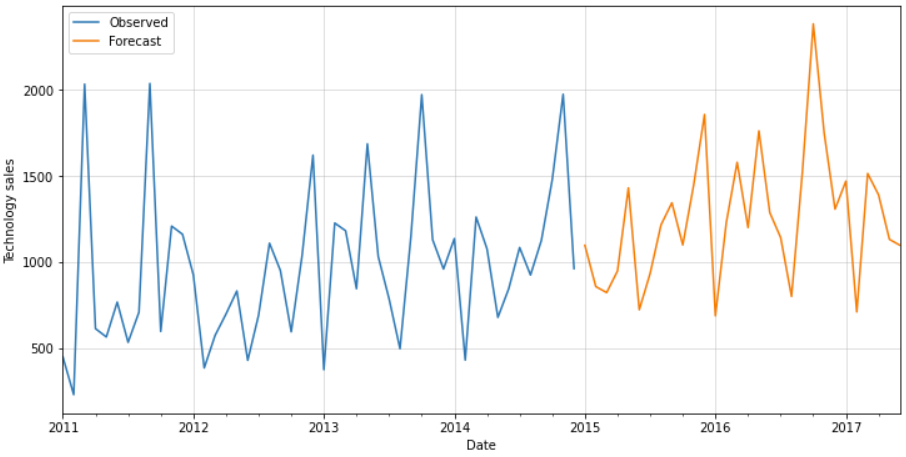
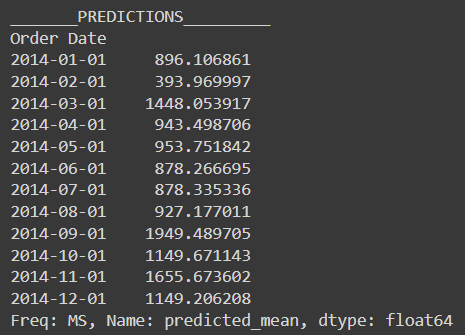


Fig: Forecast Plot

* Plotted the forecast the 30 months into the future up to 6 months in 2017.
* The Forecast plot shows slight increase in sales of Technology over the year(upward trend) over the years with Higher the demand for Technology in the months of December in 2015 and November in 2016. Least sales is observed in the months of June in 2015 and January in 2016



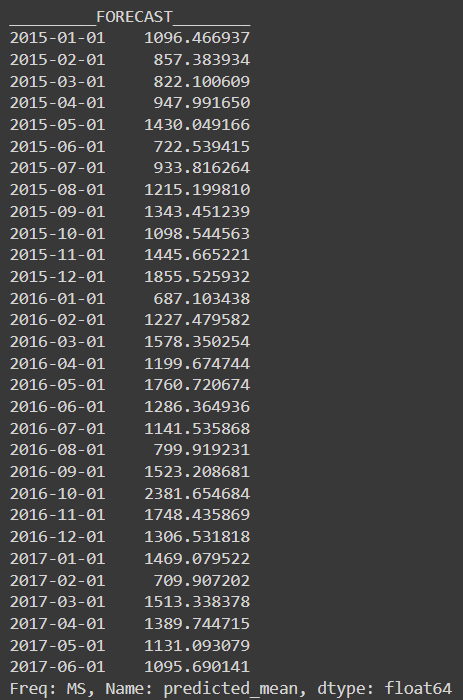


Fig: Predictions and Forecast

* Evaluation: Its done by the means of Mean Squared Error and Root Mean Squared Error
* MSE: 96115.5977
* RMSE: 310.0251

**3.Facebook Prophet**

Prophet is an general additive model that includes a number of highly advanced, intelligent Forecasting methods, including change point analysis:

y = g(t) + s(t) + h(t) + 𝜖𝑡ϵt

Here g(t) is the trend function which models non-periodic changes in the value of the time series, s(t) represents periodic changes (e.g., weekly and yearly seasonality), and h(t) represents the effects of holidays which occur on potentially irregular schedules over one or more days

* For trend, a piecewise linear or logistic growth curve trend is used.
  + Prophet automatically detects changes in trends by selecting changepoints from the data.
* For seasonality, different seasonality components are modeled using Fourier series.
* One can either use fb provided list or incorporate their own holidays into model.

**Procedure**:

* Importing prophet from fbprophet
* Creating data frame having ds column with type datetime and y column which is time series we are trying to predict
* Fit the model
* Creating a blank data frame to input predictions of periods, 36 months i.e. 3 years into future.
* Populating Forecast into blank data frame
* Plotted the Forecast
* Plotted individual components of forecast: trend, weekly/yearly seasonality,
* **Furniture Sales:**

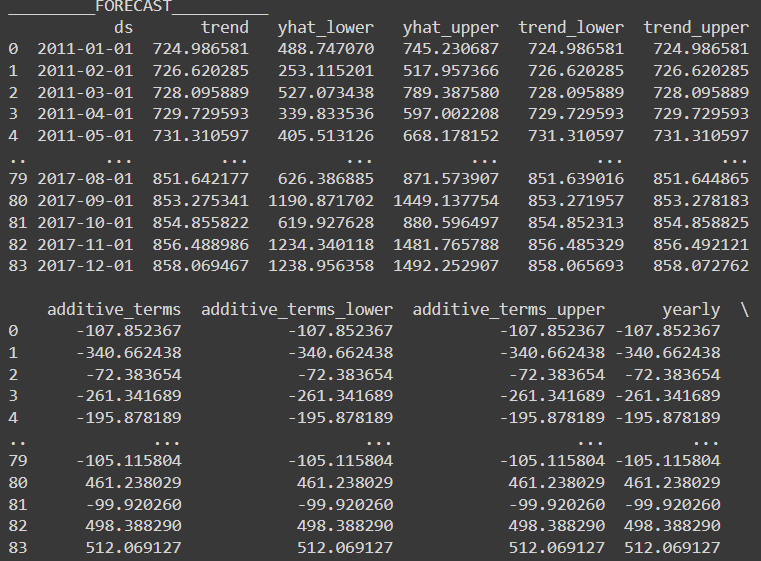


Fig: Populating Forecast into blank data frame

The above fig. represents Predictions from the prophet model is populated into the blank data frame created for up to 36 months

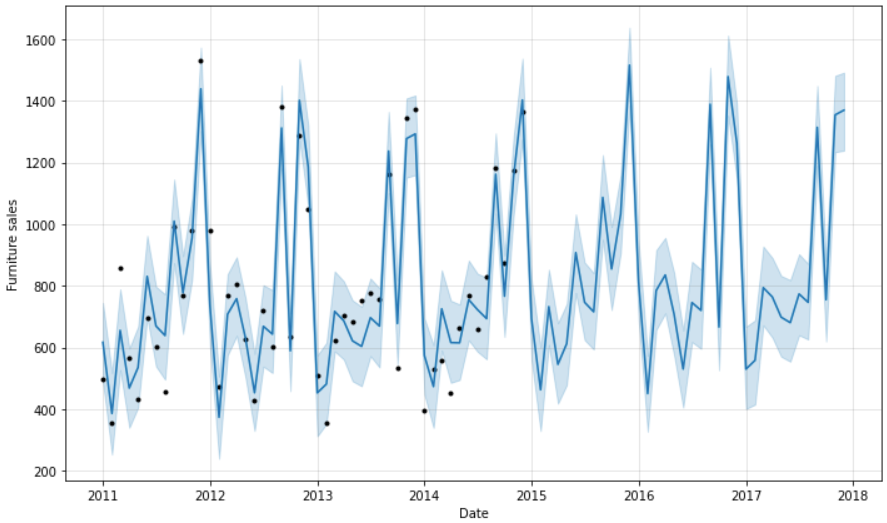


Fig: Forecast plot

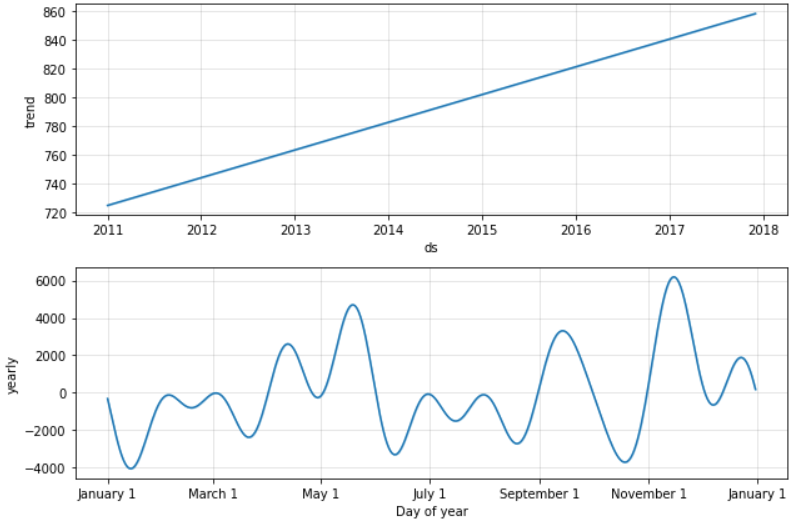


Fig: trend and Yearly Seasonality

* From the plots, trend is linearly increasing over the years i.e. demand of furniture is increasing over the years.
* Seasonality plots indicates Higher seasonality between the November and December and lowest between January and February.
* **Office Supplies Sales**

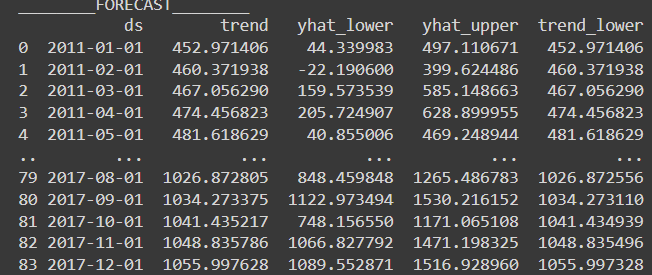


Fig: Populating forecast into blank data frame

The figure represents the predictions Populated into blank data frame.

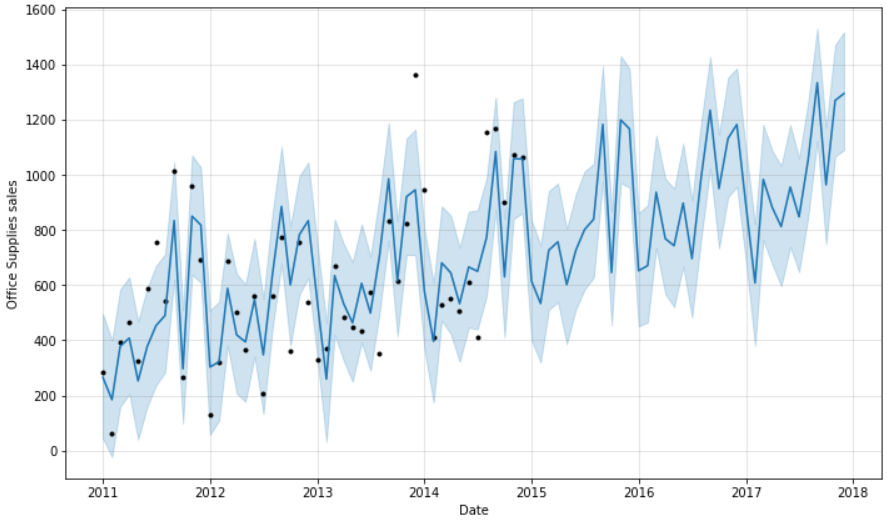


Fig: Forecast Plot

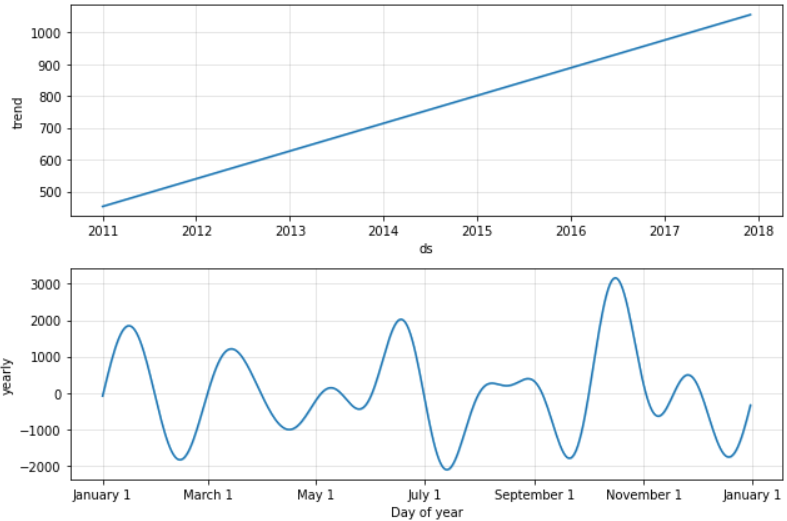


Fig: trend and Yearly Seasonality plots

* Plots of trend shows linear increase over the years which in turn insights to increasing demand of Office Supplies
* Higher seasonality observed between the months of October and November and lowest between July and August
* **Technology Supplies**

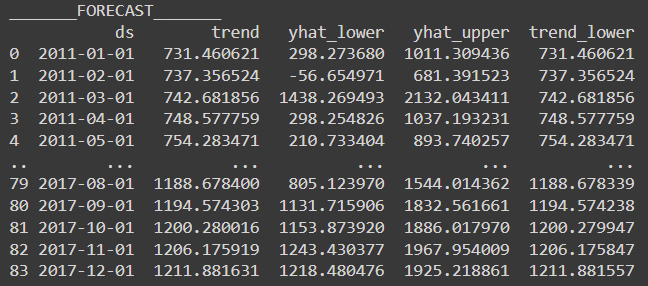


Fig: Populate forecast into blank data frame

The above figure represents the predictions from model populated into blank data frame.

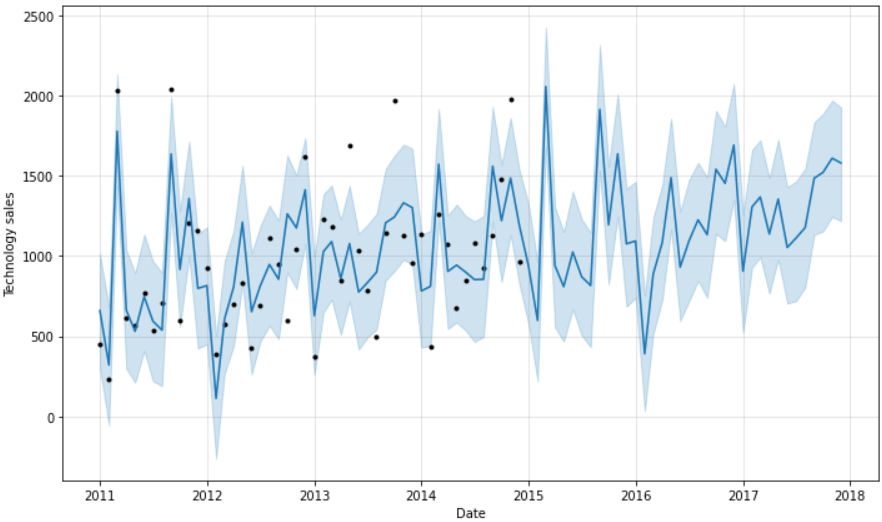


Fig: Forecast Plot

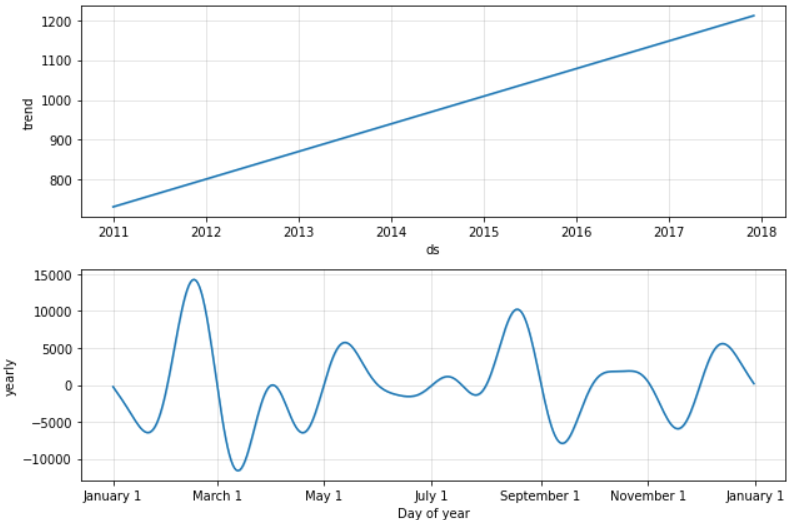


Fig: trend and year seasonality

* Linear increasing trend is observed from the plot denoting ever increasing demand for technology supplies
* Highest seasonality observed between February and march and lowest between march and April

**Challenges and Opportunities**

* I was able to create triple exponential smoothing and SARIMA with some transformations on data.
* The main challenge I have faced while executing the models was to obtain reduced root mean square error with SARIMA models being the least.

**Reflections on the Internship**

* Good learning Experience
* Showed how projects are being done in the industry
* Submitting daily activity helped me to keep track of my progress
* As a new learner of time series analysis, this internship opened up a whole new path for me to explore and learn

**Conclusions**

* Increasing trend is observed in all of the Categories Sales denoting yearly increase in demand of all Category.
* Technology is observed to have hit the peak total Sales compared to other Categories.
* Technology is observed to have most increasing demand among other Categories with Forecast being made.
* Office Supplies is observed to have least increasing demand among other Categories
* Category wise sales is generally lower during the month of February while higher sales are observed in the months of September to December
* Based on MSE and RMSE values Triple Exponential Smoothing model is better than SARIMA model as it gives more significance to recent observations but in case of Technology Sales MSE value is found to be lower for SARIMA model than Triple exponential smoothing.
* Prophet model is completely automatic forecasting techniques which can be brittle and they are often too inflexible to incorporate useful assumptions or heuristics.
* Prophet Model can produce high quality forecast compared to other two because of lack of specialized data science skill requiring substantial experience.

**Link to Code and Executable File**

* [Click here to open Colab Code](https://colab.research.google.com/drive/1AoPkQ-MZZ4M6nhYHEjuCA-vvY6MvamQh?usp=sharing)
* [Click here to open github repository](https://github.com/hashi4all/RIO-125-Forecasting-System---Project-Demand-of-Products-at-a-Retail-Outlet-Based-on-Historical-Data)
* [Click here to open my loom video](https://www.loom.com/share/d1dc3ae8205e467fae74994a8ba26cce)