- **Business Problem:**

  Global superstore company sells many varieties of products in the Furniture, Office supplies and Technology categories. They cater to customers from various segments like individual Consumers, Corporates or Home offices. As a part of current business opportunity, company needs help in identifying the potential overall Sales for the rolling 10 days period so that it can prepare for fulfilling the customer orders / demands.

- **Background/History:**

  Global Superstore company offers and sells variety of products throughout the year. They need to keep up with the customer demands in terms of the quality and quantity of the products. At the same time, optimal amount of inventory needs to be maintained to keep up with all the delivery timelines. It also needs to be maintain adequate staff to serve all the customer needs, both in store and online demands. Hence, in order to fulfil the customer needs, it would be ideal to predict / know what the demand forecast looks like. This can be achieved by accurately predicting the Sales estimates for the products in near future, which in turn can help determine the optimal levels of product inventory to be maintained.

  As a part of this project, we are looking to help the global superstore with the US sales forecast for the upcoming / rolling 10 days, based on the sales data available for the years 2015 – 2018 based on Order dates. This will help them keep up with the customer demands, appropriate level of inventory maintenance, manage appropriate number of support staff and be prepared with the delivery arrangements.

- **Data Explanation (Data Prep/Data Dictionary/etc):**

Following features are present in the datasets:

- Row ID : Unique row identification
- Order ID : Order ID for customer orders
- Order Date : Date when order is placed
- Ship Date : Date when order is shipped
- Customer ID : unique ID for the Customer placing the order
- Customer Name : Customer Name
- Segment : Customer segment (Corporate / Consumer / Home office)
- Country : Country name (United States in this case for entire dataset)
- City : City name
- State : State name
- Postal Code : 5 digit zip code
- Region Code : Represents the Region where the customer resides
- Product ID : Unique ID of the product for which order is placed
- Category : Product category (Furniture / Office Supplies / Technology)
- Sub-Category : Sub-categories assigned to each of the products under specific categories
- Product Name : Product name
- Sales : Sales figures for the order

As a part of the final predictions, we will be forecasting the sales number for rolling 10 days in future.

Training dataset has 9800 records for Sales data. Out of these, Postal code column is the only one which has NULL values for 11 records in it. After reviewing those records as a part of data

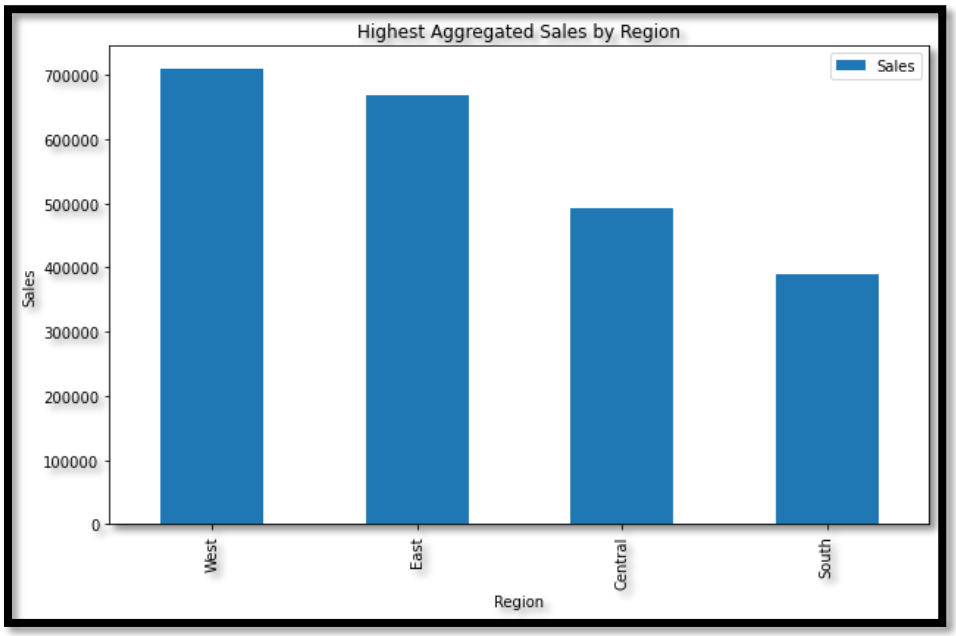
analysis and preparation, it was noticed that all of those records belonged to the **City** of **Burlington** in the **State** of **Vermont**. So, we will be filling in the Postal code value of '05401' as a default value for these records.

Couple of columns 'Order Date' and 'Shipping Date' are in DD/MM/YYYY format. We will convert these into YYYY-MM-DD format so as to be compatible with usage in Time Series analysis and processing.

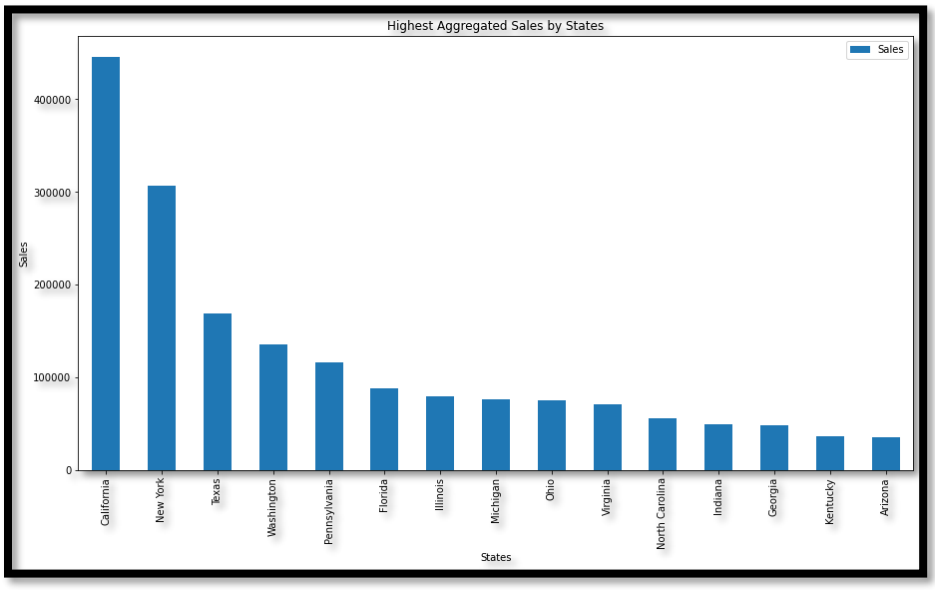
Also, the as we will be looking at the overall Sale data on a daily basis, we will need to aggregate the existing Sales data for each day. This will help us to predict the Sales for 10 days towards the end of the series.

- **Analysis & Methods:**

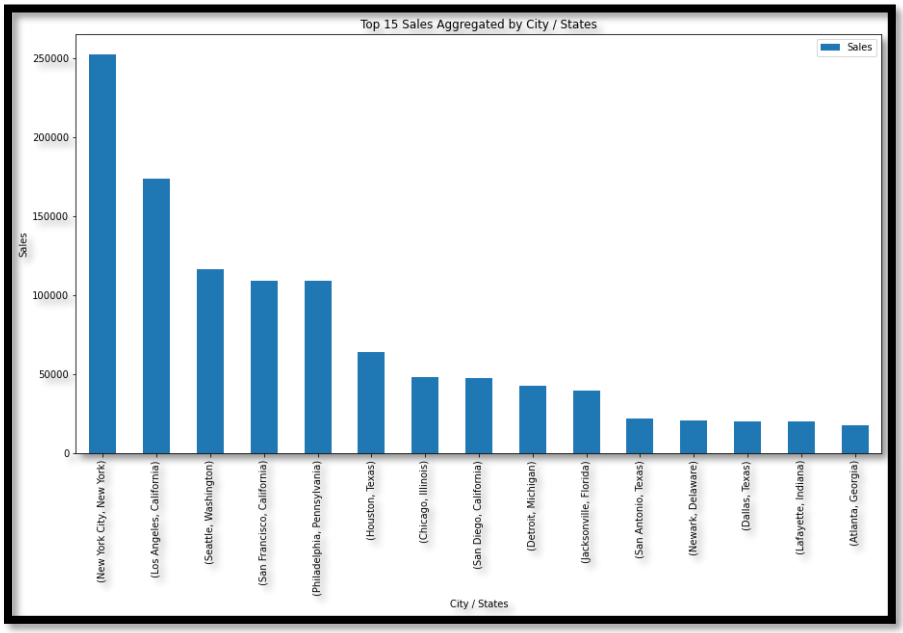
*Highest Sales by Region : West Region has highest Sales gross up ($700,000) with East region at close second. South region has lower Sales as compared to other regions. West and East region Sales demands are higher and company needs to be well prepared in terms of supplies, stocks and shipping to meet them all.*



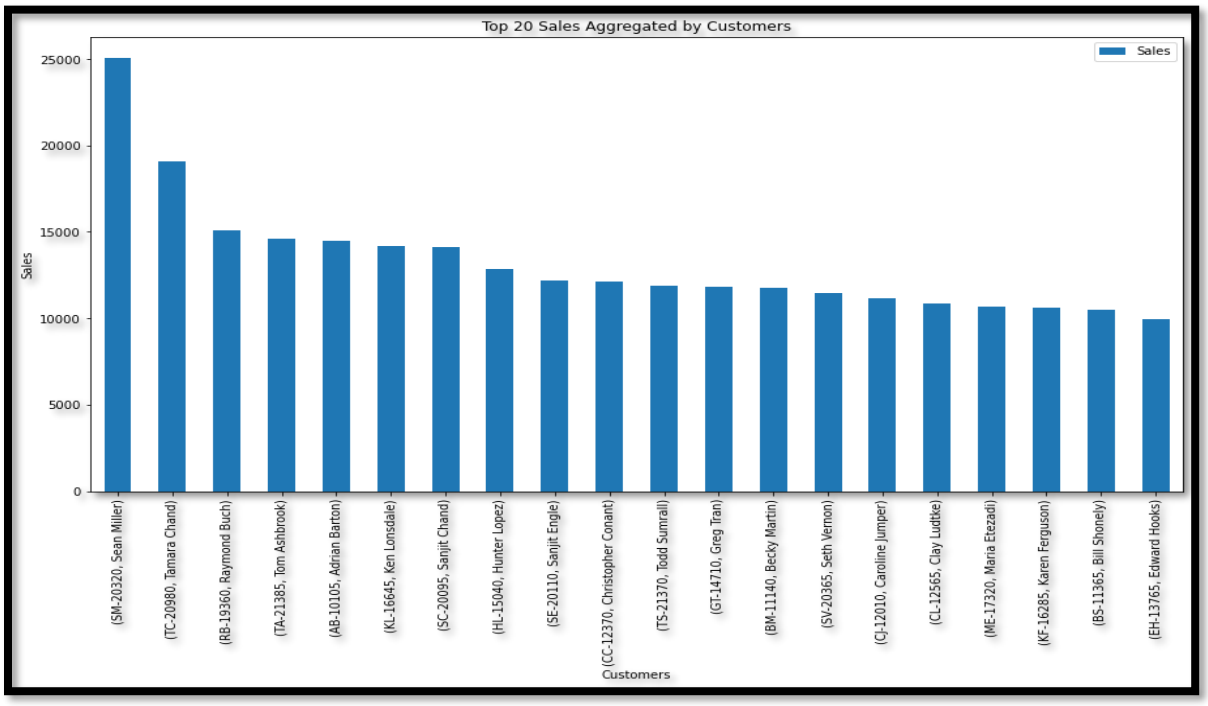
*Below is the snapshot of Top 15 States by Sales amounts. California being at the top followed by New York and Texas at 2nd and 3rd respectively. In particular, the states of California, New York, Texas and Washington contribute to higher sales. So, supplies, stocks and shipping needs to be prepared for these.*



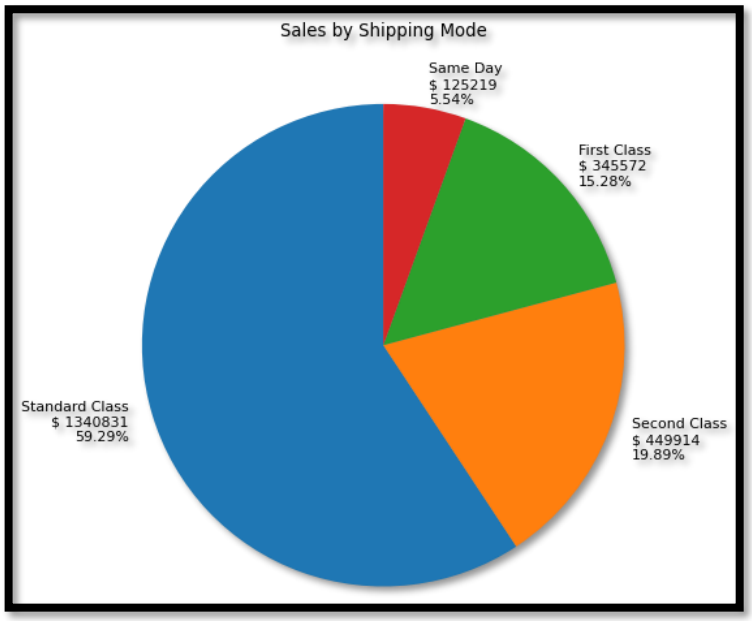

*Amongst the Top 15 Cities by Sales: New York City, NY leads the way followed by , Los Angeles, CA at 2nd and Seattle, WA at the 3rd place.*



*Below is a snapshot of Top 20 Customers with respect to Sales amounts.*

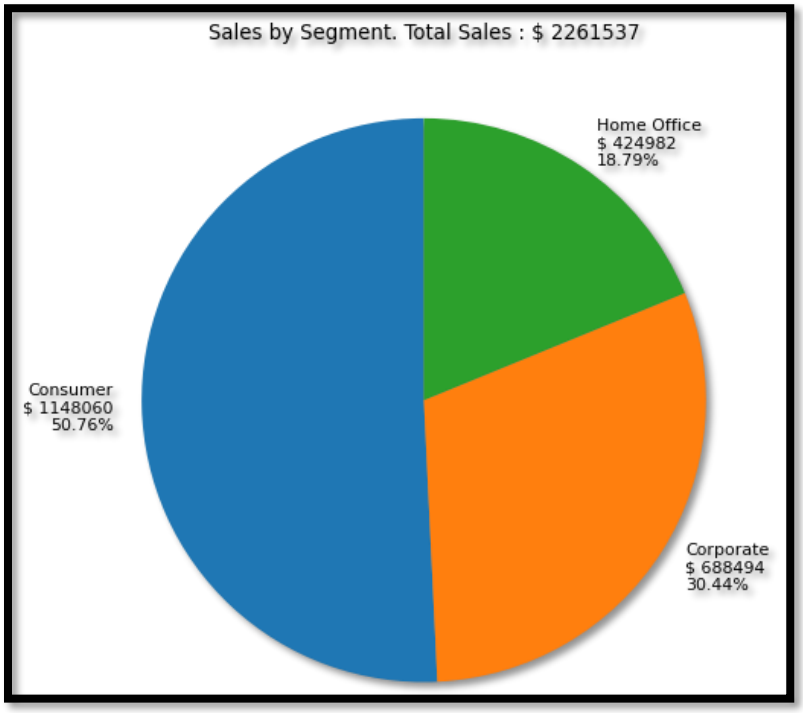



*Sales distribution by Shipping Modes : Standard Class being the major portion (around 59%) followed by Second Class (close to 20%). Standard class shipping mode and second class shipping mode demands will be higher and we need to be well prepared to accommodate the orders.*

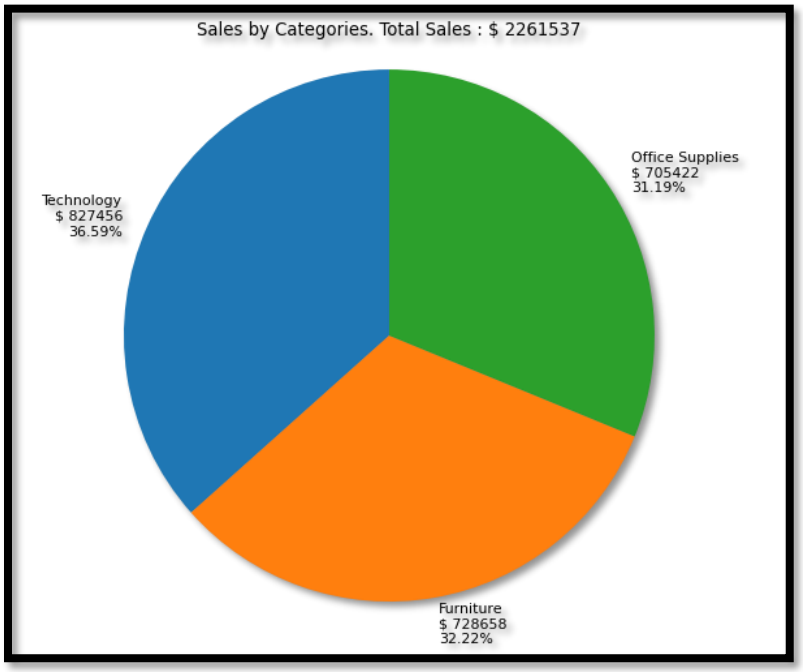



+

*When it comes to Segments, Consumers contribute to half the Sales (50%) followed by Corporate segment at 30%. Consumer business segment and corporate segment need to be supported well.*

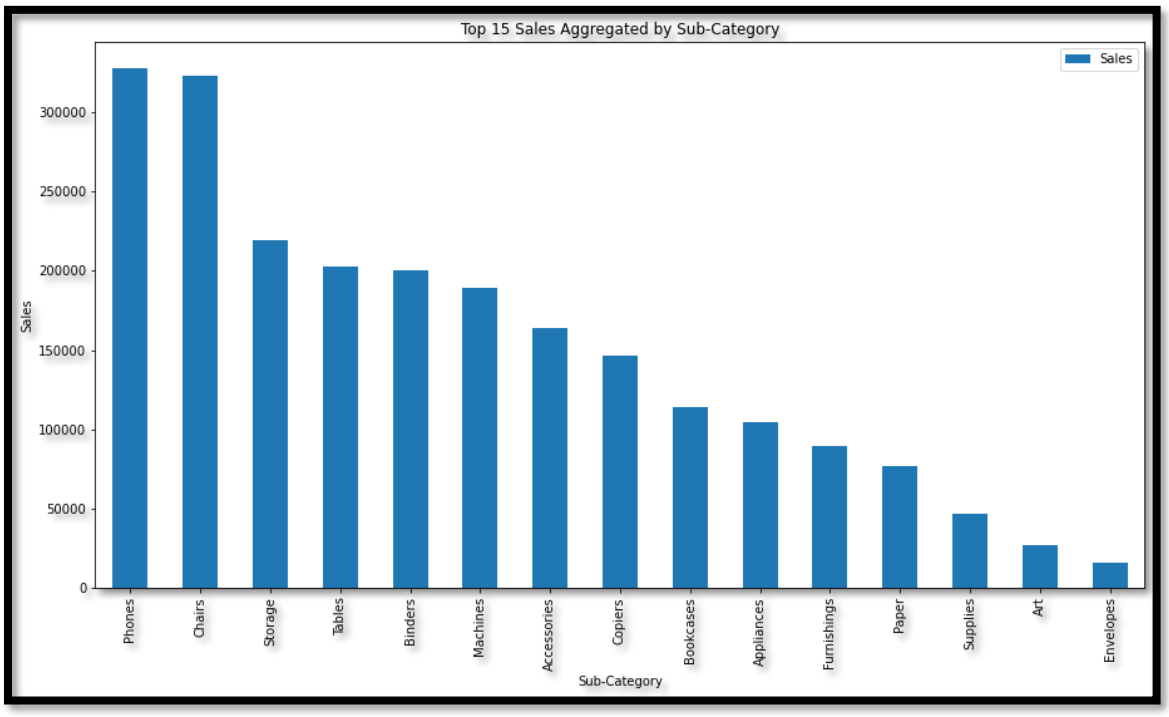



*Sales distribution by Categories : Technology at 36.5% followed by Furniture and Office supplies.*



*Top 15 Sales segregated by Sub-Category. Phones and Chairs are the top 2 followed by Storages.*

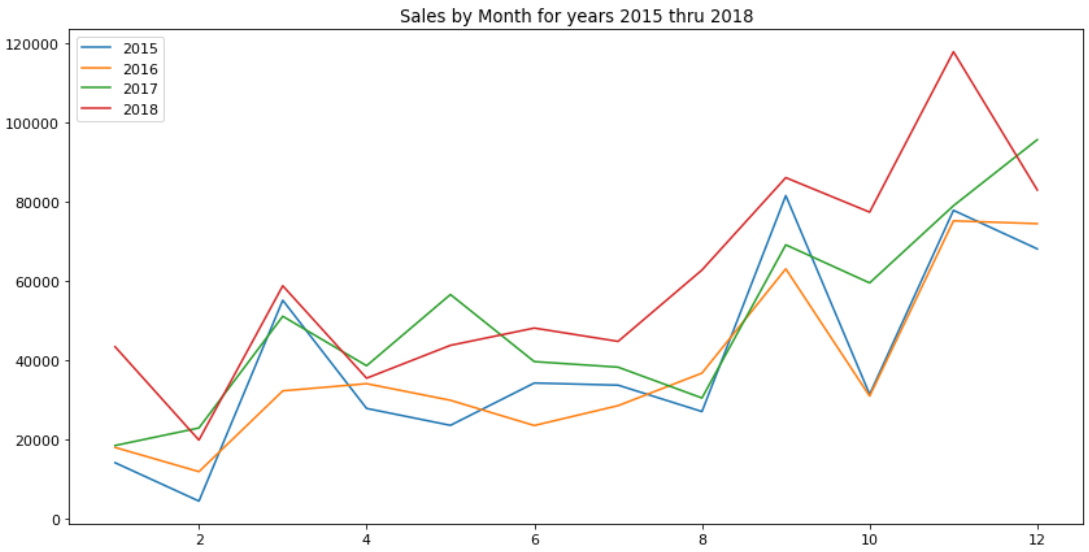
*Keep adequate stocks of items in Phones, Chairs, Storage units, Tables and Binders and machines to support the higher customer demands for these sub-categories.*



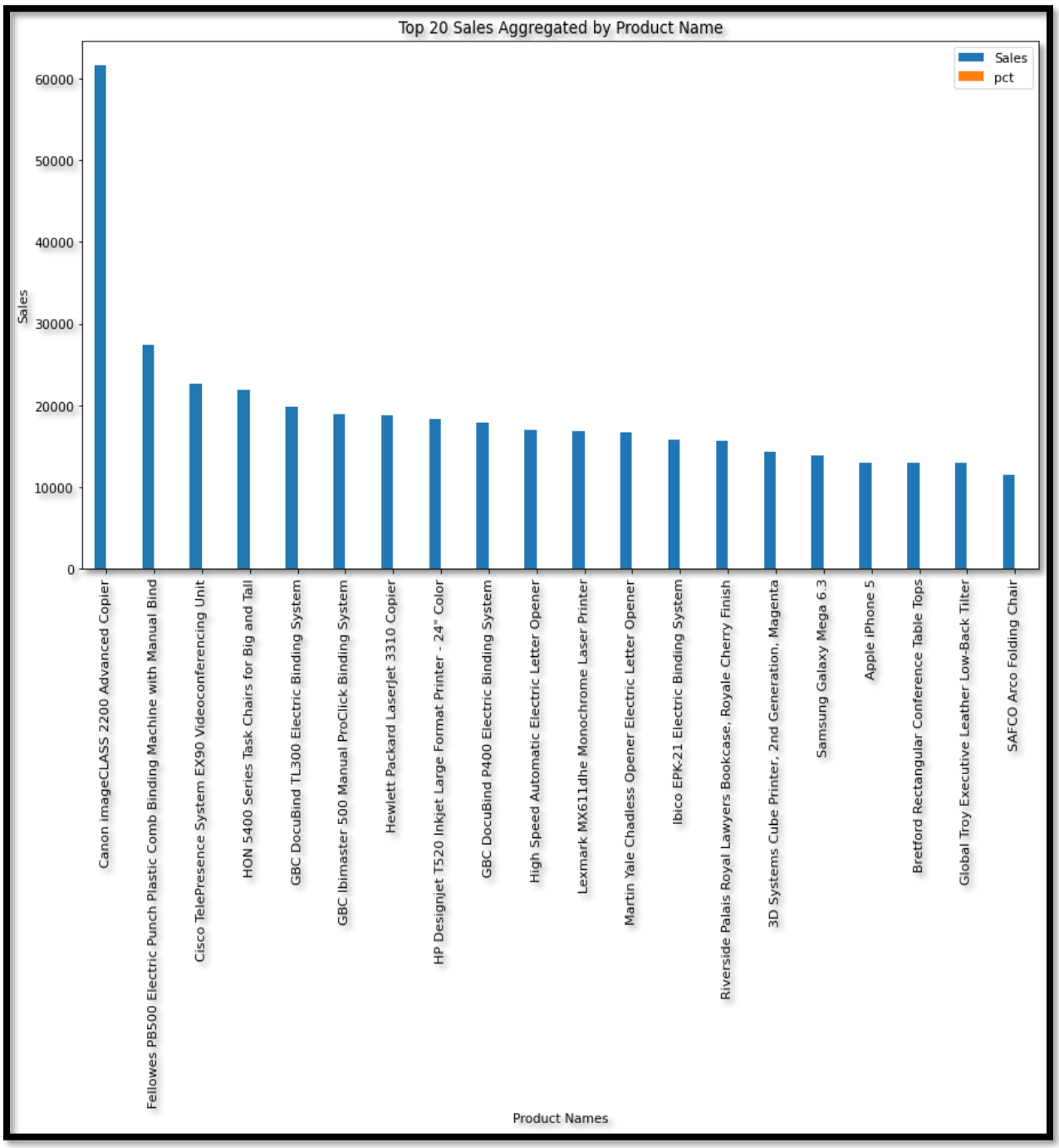

*Monthly Sales trends for each year : 2018 numbers apparently at the top followed by 2017*

*Maintain sufficient additional staff to support the higher Sales usually October – January periods*

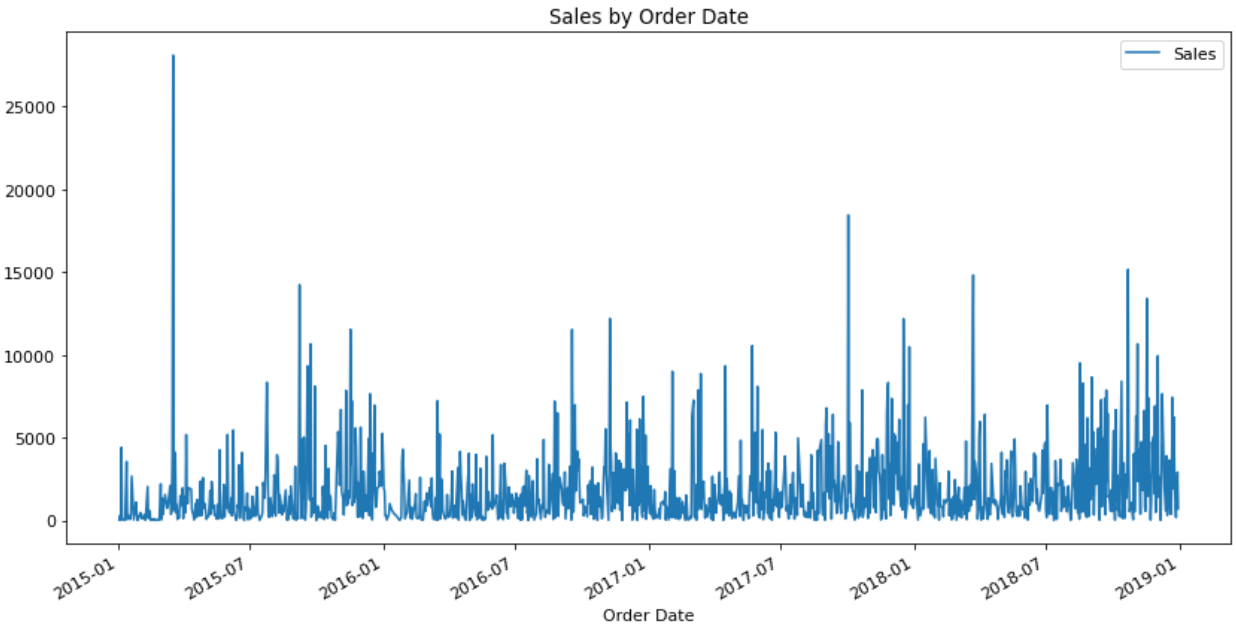
*Sales demands with some level of demands in the month of March*



**Top 20 Best Selling Products in terms of Sales : Canon image Class 2200 Advanced Copier is the topmost one. Ensure the high in demand Products are maintained in the stock / supplies.**

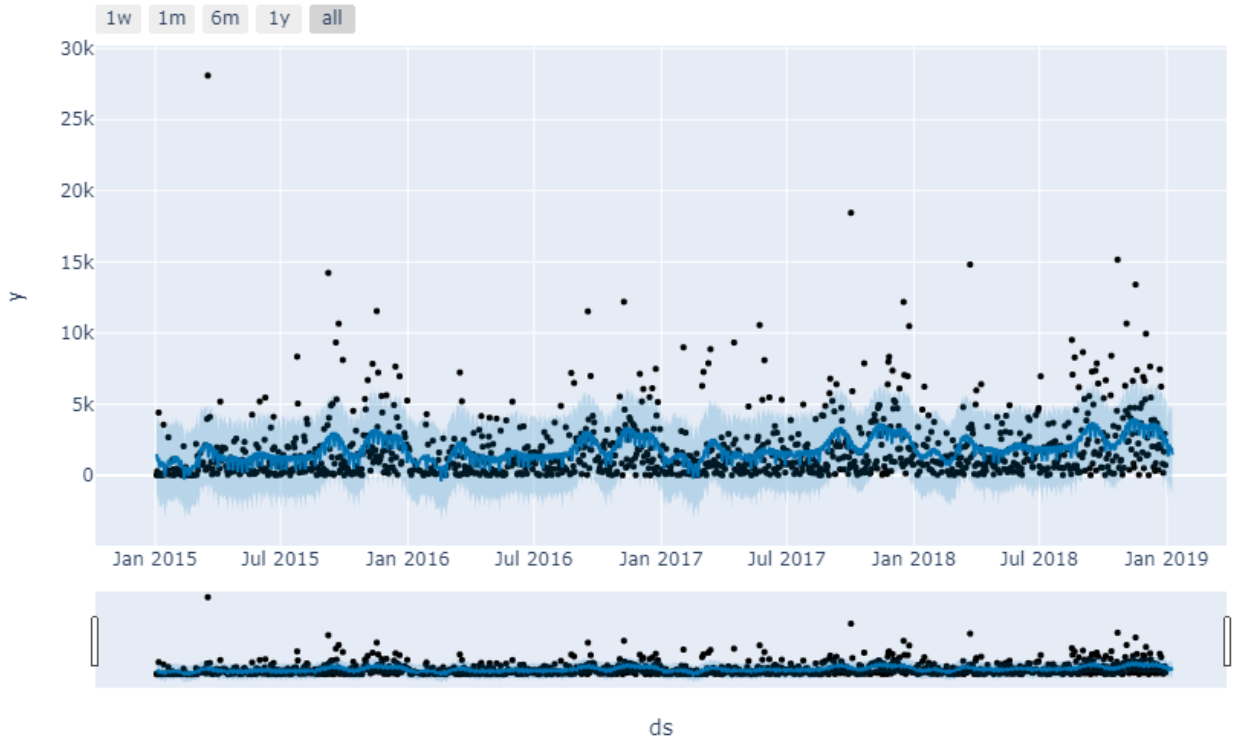


**Sales trends by Order Date**

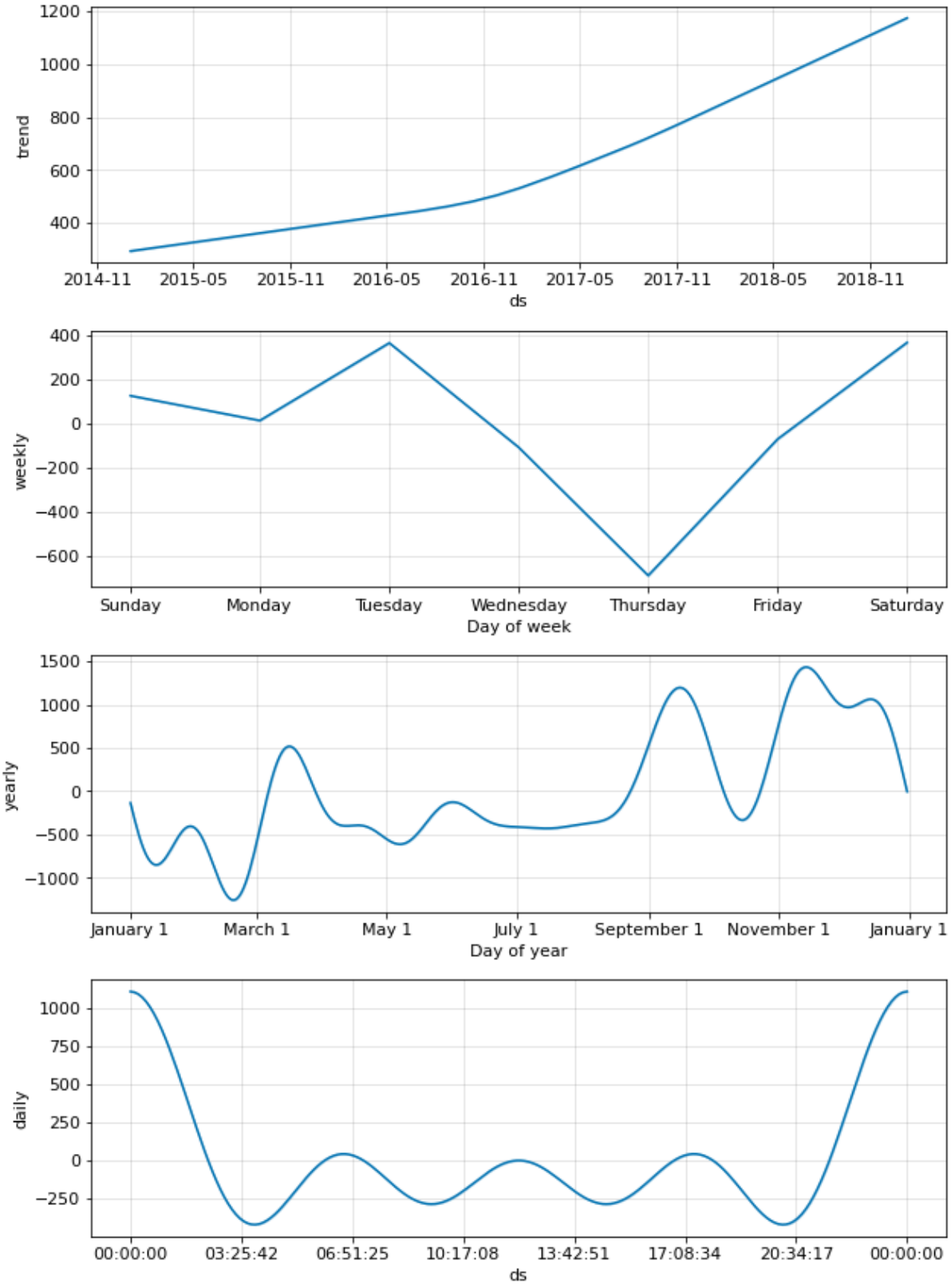



## Modeling and Evaluation:

**fbprophet model:**

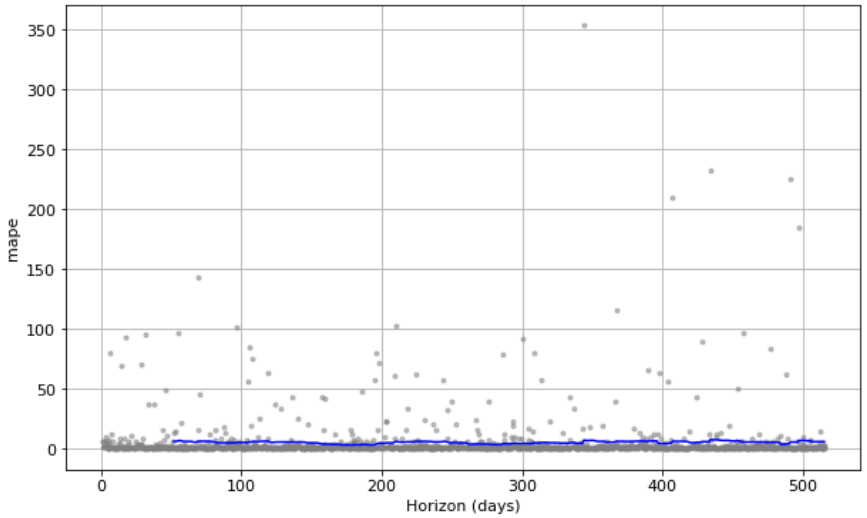


fbprohet model components:



## fbprophet diagnostics and validations:

| horizon | mse | rmse | mae | mape | mdape | coverage |
|---|---|---|---|---|---|---|
| 511 days | 6.879071e+06 | 2622.798236 | 1987.559586 | 5.580470 | 0.671296 | 0.742647 |
| 512 days | 6.956558e+06 | 2637.528676 | 2011.011200 | 5.640538 | 0.687591 | 0.742647 |
| 513 days | 6.903811e+06 | 2627.510453 | 1987.757728 | 5.631451 | 0.671296 | 0.742647 |
| 514 days | 6.792085e+06 | 2606.162840 | 1960.255295 | 5.661361 | 0.674211 | 0.750000 |
| 515 days | 7.019155e+06 | 2649.368852 | 1993.529684 | 5.664362 | 0.674211 | 0.742647 |

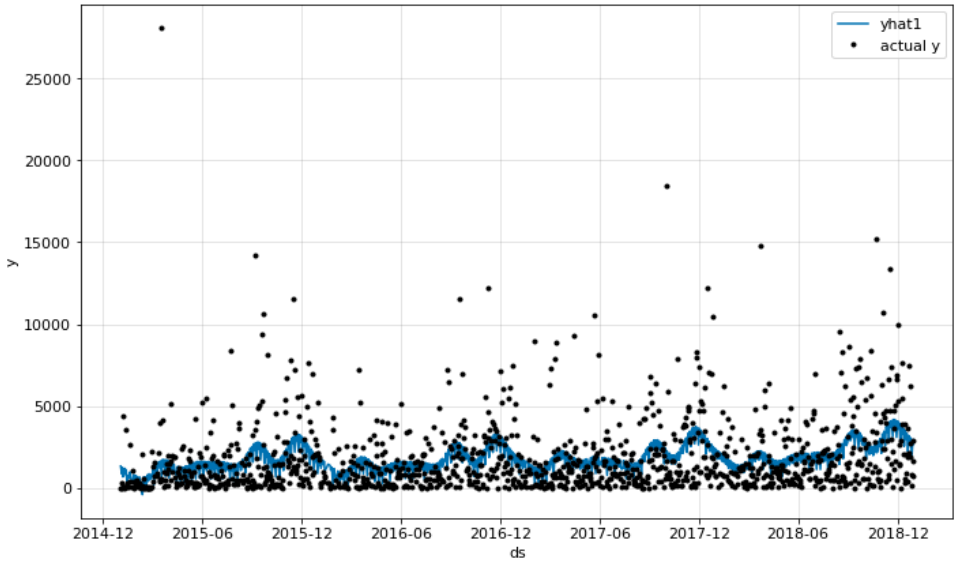



## Predictions using fbprophet

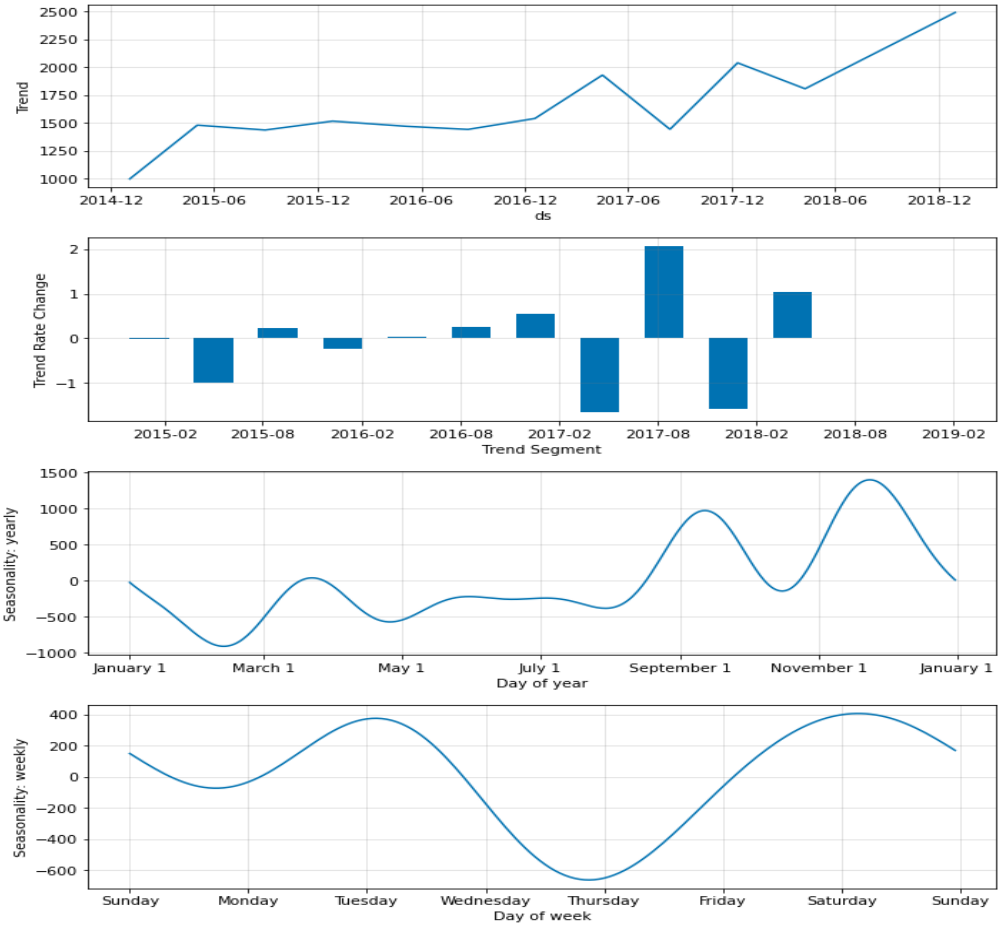
| ds | trend | yhat |
|---|---|---|
| 2018-12-31 | 1166.064301 | 2314.560165 |
| 2019-01-01 | 1166.993740 | 2559.996785 |
| 2019-01-02 | 1167.923180 | 1986.735887 |
| 2019-01-03 | 1168.852619 | 1310.160220 |
| 2019-01-04 | 1169.782059 | 1831.587080 |
| 2019-01-05 | 1170.711498 | 2179.070029 |
| 2019-01-06 | 1171.640938 | 1858.885180 |
| 2019-01-07 | 1172.570377 | 1675.237461 |
| 2019-01-08 | 1173.499817 | 1962.377954 |
| 2019-01-09 | 1174.429256 | 1438.779767 |
| 2019-01-10 | 1175.358695 | 818.861687 |

**Neural Prophet exercise:**

Neural Prophet forecast plot

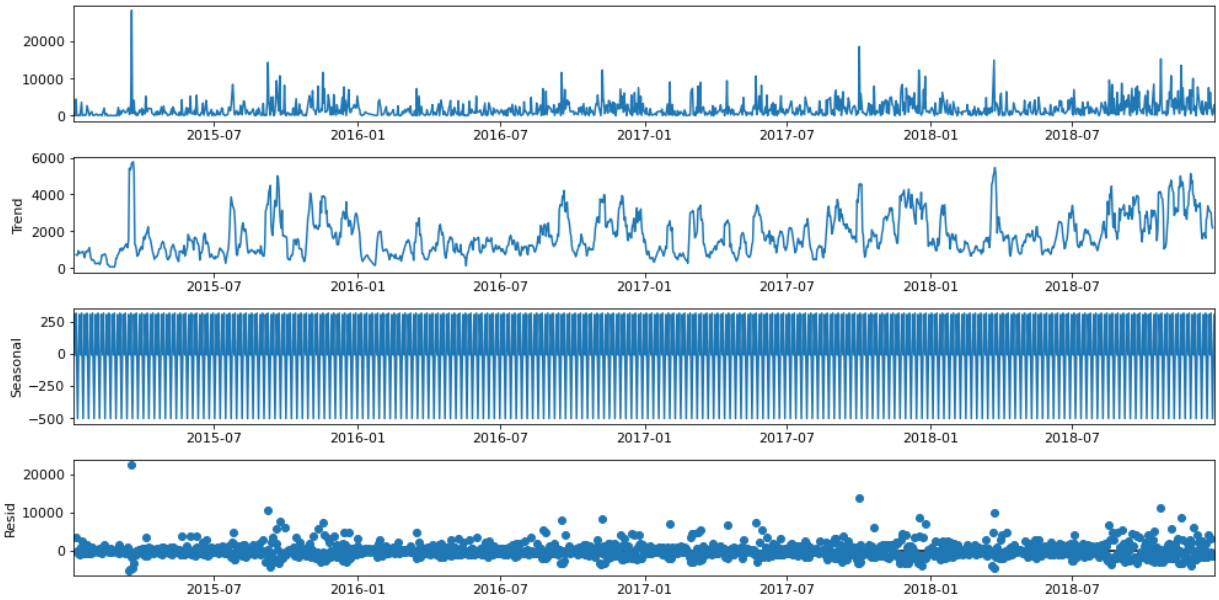



Neural prophet forecast components:



## Stationary / Seasonality Test for SARIMA modeling

***The given Sales Timeseries is found to be Stationary using Augmented Dickey Fuller Test***

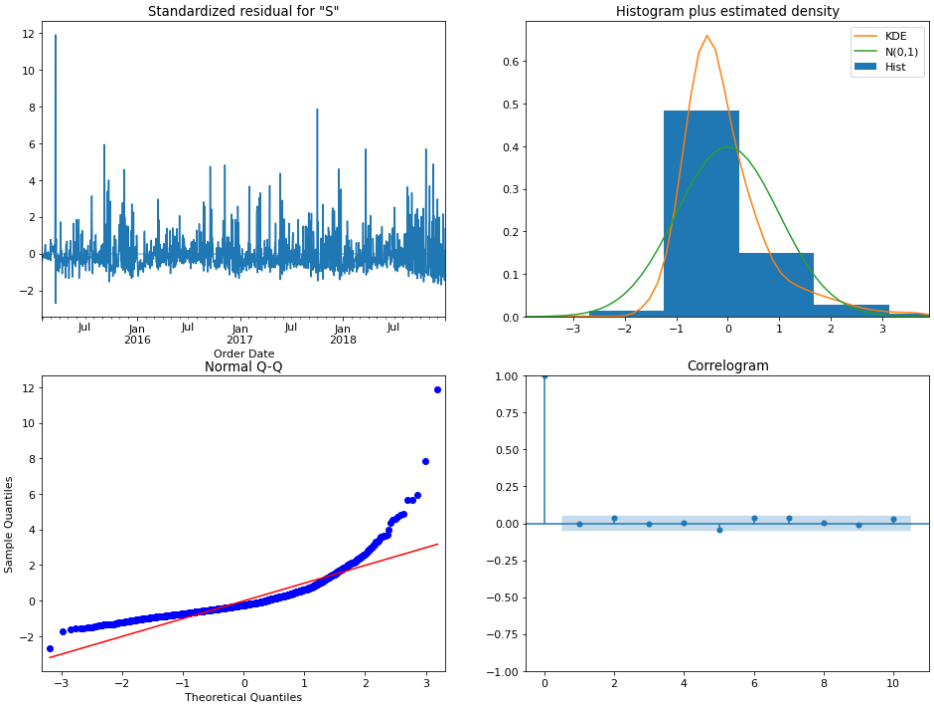



```
adf results:  (-6.444493081294046, 1.5781789978043213e-08, 15, 1442, {'1%': -3.4348929812602784, '5%': -2.863546418485167,
'10%': -2.5678382024888378}, 26015.358001313038)

ADF Statistic: -6.444493081294046
p-value: 1.5781789978043213e-08
```

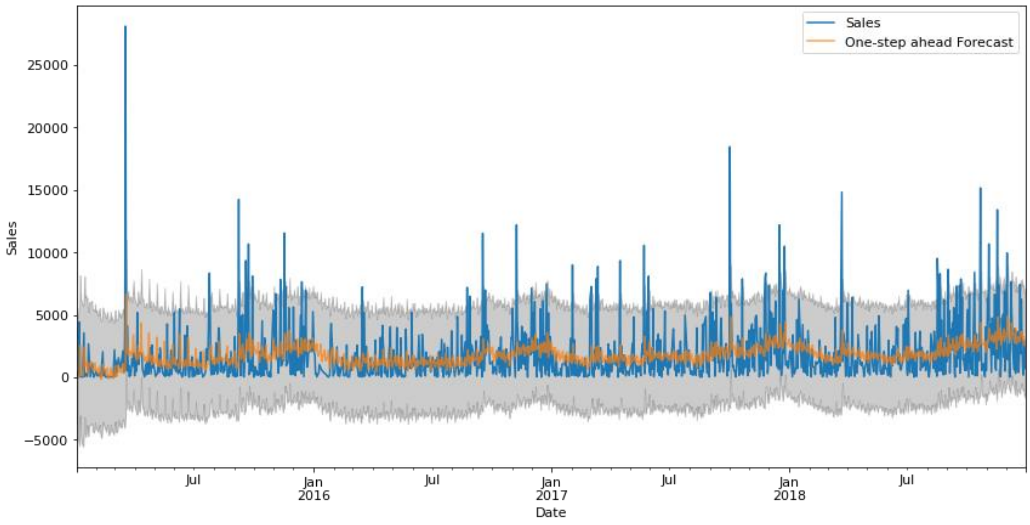
**Since the p-value is much lesser than 0.05 (very close to 0.0 in fact), we can conclude that this Sales data time series is Stationary series**

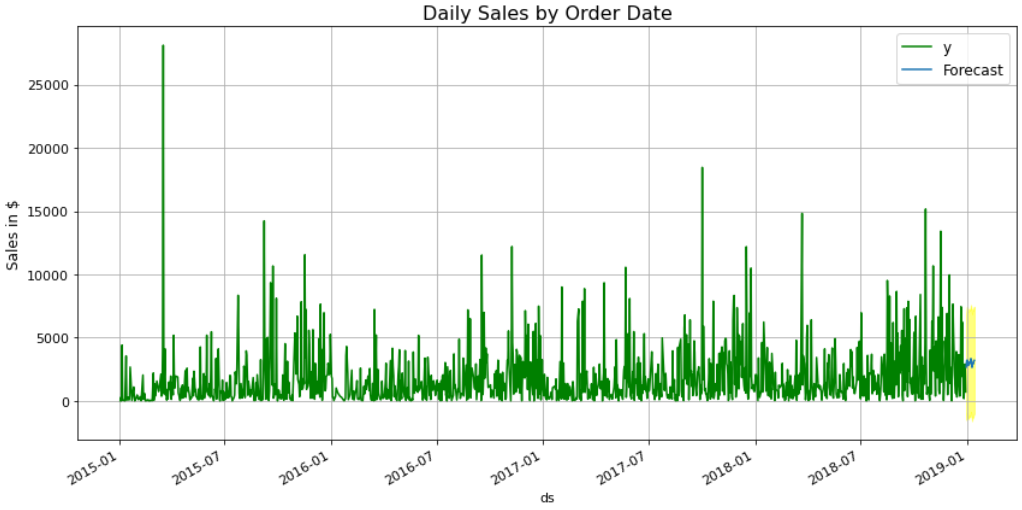
## SARIMA Model diagnostics



Looking at the correlogram, it appears that there is no autocorrelation in the residuals and it is potentially the white-noise in the data. Hence the residuals mean is zero not being correlated.

**Prediction Plot using SARIMA model**




**Sales Predictions using SARIMA model and RMSE / MAPE (Error) values:**

| | |
|---|---|
| 2018-12-31 | 2667.872280 |
| 2019-01-01 | 3200.266913 |
| 2019-01-02 | 3220.774239 |
| 2019-01-03 | 2796.500929 |
| 2019-01-04 | 2816.080403 |
| 2019-01-05 | 3045.837516 |
| 2019-01-06 | 2946.380407 |
| 2019-01-07 | 3008.356850 |
| 2019-01-08 | 3344.342459 |
| 2019-01-09 | 3066.914374 |

```
from sklearn.metrics import mean_squared_error
from math import sqrt

mse = mean_squared_error(y_forecast, y_actual)
rmse = sqrt(mse)

# root mean squared error of the forecast. rounded to 2 decimals
print(f'The Mean Squared Error of the forecasts is {round(rmse, 2)}')
```

```
The Mean Squared Error of the forecasts is 2130.24
```

```
# Get Mean Absolute Percentage Error (MAPE)
mape = np.mean(np.abs((y_actual - y_forecast) / y_actual)) * 100
mape
```

```
744.2796243651256
```

- **Conclusion:**

Looking at the above analysis and predictions, we can notice that the fbprophet model produces lower mape error value with sales forecast for 10 days between $1000 to $2600 range and neuralprophet model seems to provide similar forecast numbers. SARIMA model gives Sales predictions on a little higher side in the range of $ 2660 - $ 3350, with higher mape error of 744.

We can notice that higher level staff needs to be maintained during the months from October to January of following year and also for the month of March, owing to higher Sales orders. Also, West and East regions, consisting of States of California, New York, Washington, Pennsylvania constitute to higher Sales. So, adequate stock of supplies, staff and corresponding shipping arrangements need to be maintained accordingly. In particular, the sub-categories like Phones, Chairs, Storage Units and Tables are in higher demand and maintain the sufficient stocks of these.

- **Assumptions:**

Assuming that given data has no outliers or incorrect values present in it, since many of the features are texts (Customer Name, Address, City, State, Shipping Class etc.) in nature and numerical feature would be reflecting the actual information i.e. they do not contain any out of the range / unexpected values.

- **Limitations / Challenges:**

Ensuring that all the critical data elements hold valid Sales data points is important. With occasional spikes in the sales data for the given timeframe, causes the forecast range to be wide spread and it also causes some level of errors in these predictions as compared to the actuals.

- **Future Uses/Additional Applications / Recommendations:**

  The approach / models being created as a part of this opportunity can be partially utilized or concepts used for analyzing the category or sub-category or product level Sales forecasting with appropriate data collection / preparation. Even the predictions for City / State / region can be performed with necessary data preparation and cleaning activities.

- **Implementation Plan:**

  Since the outcome variable would be Sales amounts (numeric amounts) over a period of time – Time series forecasting methods like fbprophet, neuralprophet, SARIMA algorithms will be used for final predictions. We can train these models and then generate the final predictions. The results predictions can be cross verified with each other between these modeling methods. Also, the model diagnostics based on the Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) can be looked at to understand the predictions.

- **Ethical Assessment:**

  The current dataset contains Customer Name as personally identifiable attribute, especially when combined with Postal Codes, would identify the customers uniquely. We will ensure that Customer Name feature is not included in the Data Preparation phase along with any other non-essential features in the dataset (like row id, order id etc.)

- **References** :

  - Data source : https://www.kaggle.com/rohitsahoo/sales-forecasting

- Sharma, Rishabh (2021) “How to build SARIMA model in Python” https://medium.com/mlearning-ai/how-to-build-sarima-model-in-python-7ae83b14c884
- Brownlee, Jason (2020) “Timeseries forecasting with Prophet in Python” https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/
- https://xang1234.github.io/prophet/
- https://www.kaggle.com/prashant111/tutorial-time-series-forecasting-with-prophet
- https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/
- https://stackoverflow.com/questions/68142645/disable-info-fbprophet-disabling-daily-seasonality-from-printing
- https://www.kaggle.com/ohseokkim/predicting-future-by-lstm-prophet-neural-prophet
- https://facebook.github.io/prophet/docs/diagnostics.html
- https://nextjournal.com/fb-prophet/facebook-prophet-diagnostics
- Berk, Michael (2021) “Prophet vs. NeuralProphet” https://towardsdatascience.com/prophet-vs-neuralprophet-fc717ab7a9d8
- Alizadeh, Esmaeil (2020) “NeuralProphet: A Time-series modeling library based on Neural-Networks” https://towardsdatascience.com/neural-prophet-a-time-series-modeling-library-based-on-neural-networks-dd02dc8d868d