**AIT-614: Big Data Essentials**

**Forecast Analysis on Superstore Data**

**Project Report**

**By**

**K. Siva Naga Lakshmi(G01099587)**

****

**George Mason University, Fairfax VA.**

**Abstract**

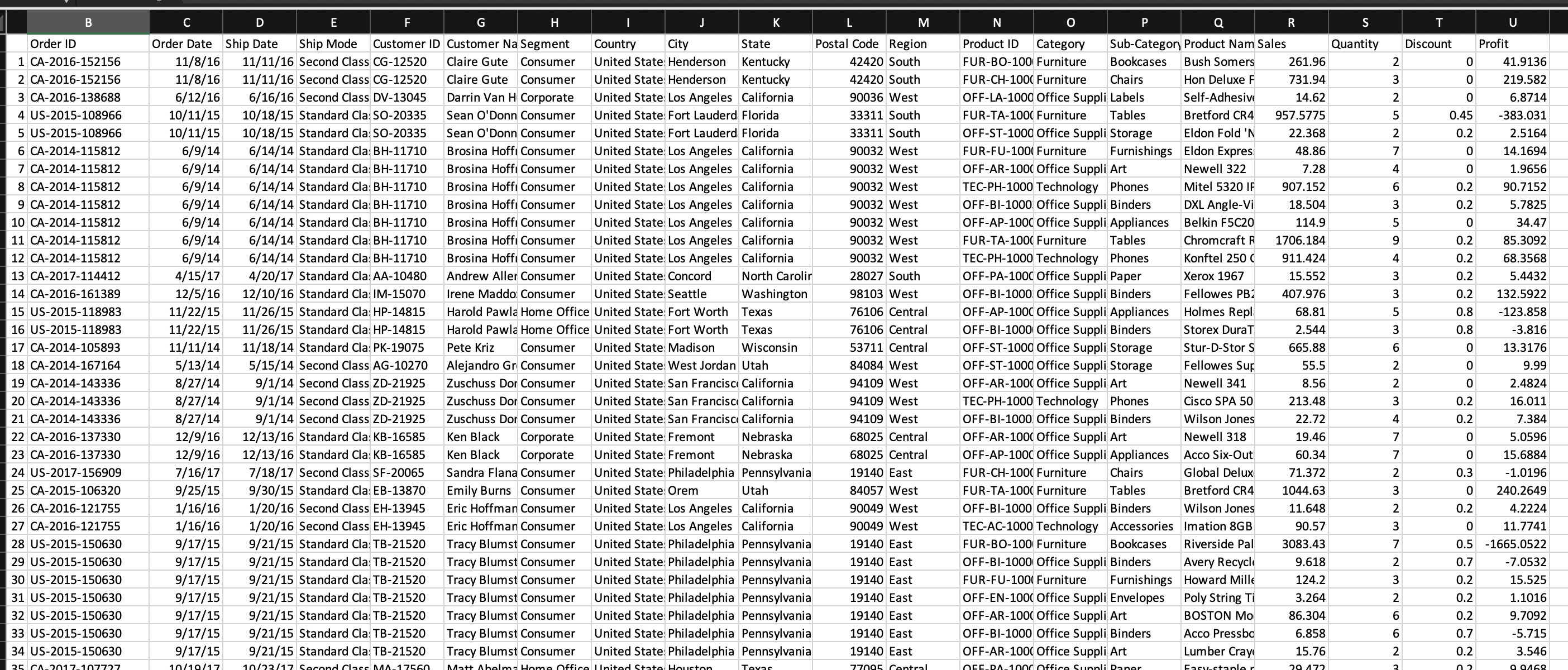
As the world is evolving with lots of data, we can observe the use of forecast analysis in stock markets, weather predictions and sales predictions in the industries. So, our aim is to predict the sales in a superstore data set by performing time series along with forecast analysis using ARIMA technique. We have performed analysis on the data set using Python. In our project we are performing on the furniture sales category and predictions performed on this category.

**Forecast Analysis:** it is the process of predicting the future values based on the past and present data, commonly by analyzing the trends.

**Time-Series:** series of data points indexed or listed over a period of time. Time series is a sequence which is taken successively with equal intervals of time.

**Data Set**

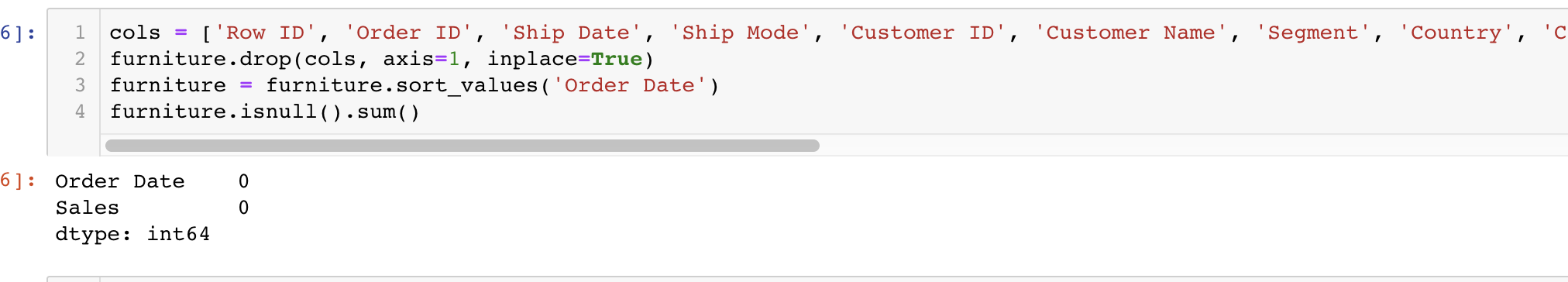
The dataset has been collected from the tableau website. The data set consists of around 10,000 records with 21 attributes like order ID, order date, shipping date, customer details, product details. The main attributes that we considered here are order date and category. In category we considered furniture related records for further analysis.



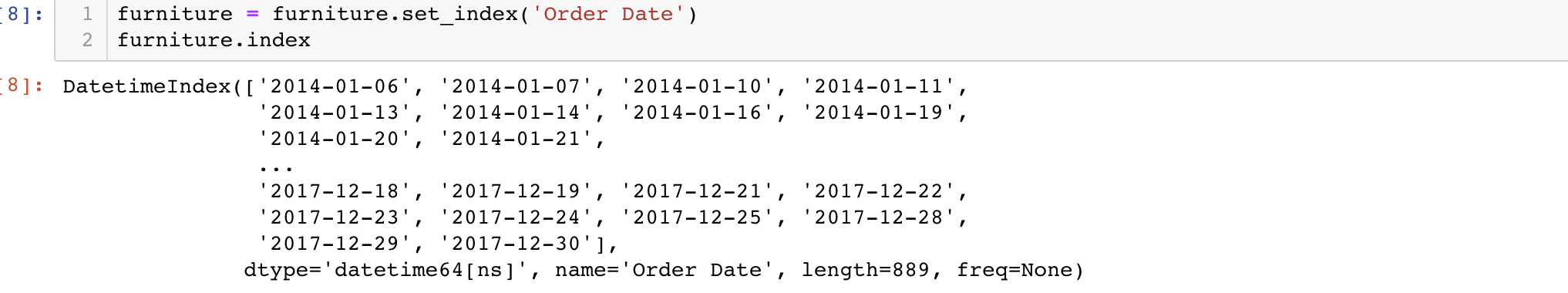
For the analysis we are considering 4-year furniture data.

**Data Preprocessing**

In this step we will check whether it has any null values and replace them. We will remove the unnecessary columns which are not required for the further analysis. After this Indexing will be performed for the order date column, it means that we cannot consider the current date time for the analysis as it will be tricky to consider, therefore we will use average sales of that month and start of each month as a time stamp.



The above image checks for the null values.



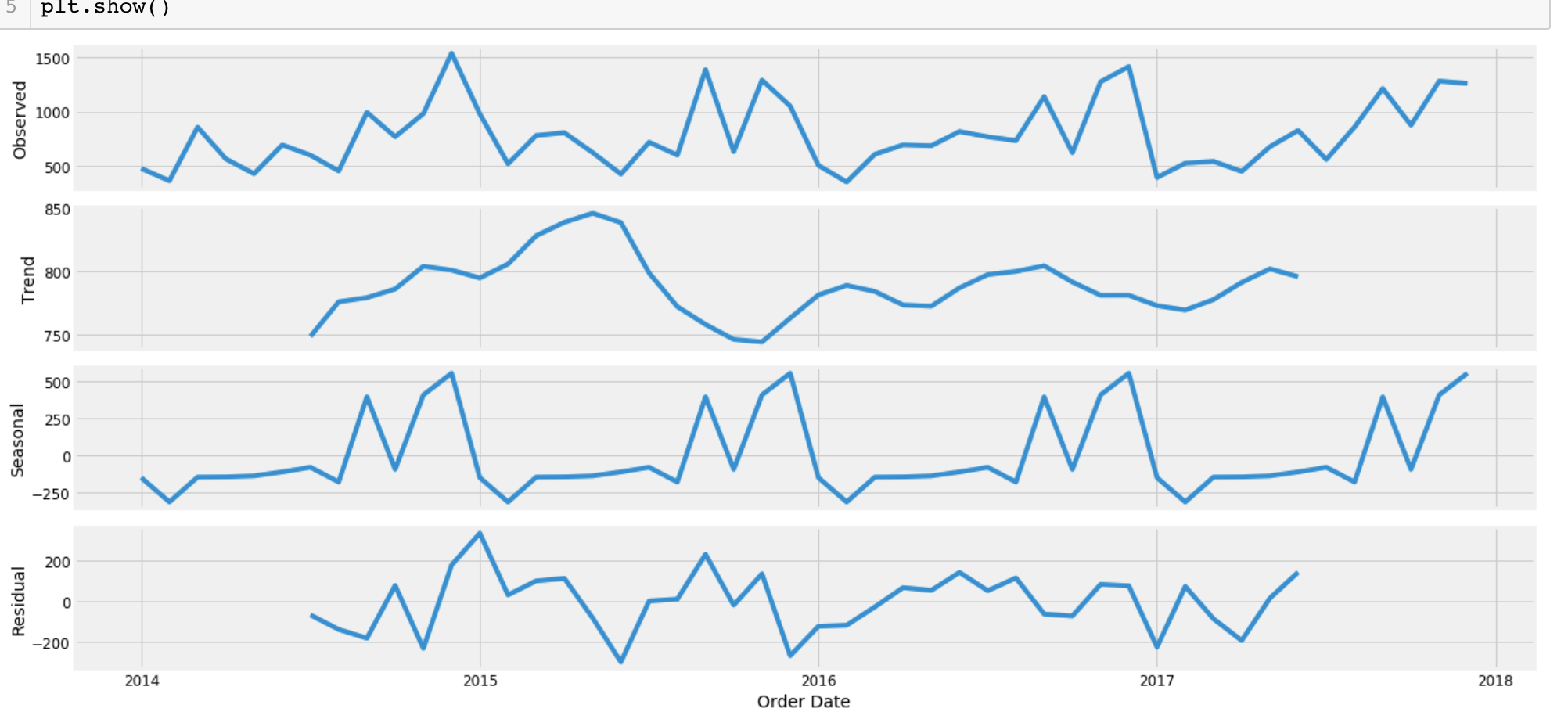
Above image depicts indexing the timeseries data.

**Visualizing the Furniture sales**

We will be visualizing the furniture sales data, and observe a different pattern, and it has time series seasonality pattern, where sales are less at the beginning of the year and they are increased gradually by the end of the year. We are using another visualization technique which decomposition time-series is, where it will decompose the components into three different which are seasonality, trend and noise. From the below plot we can say that they are unstable.



Above graph depicts the visualization for the furniture data.

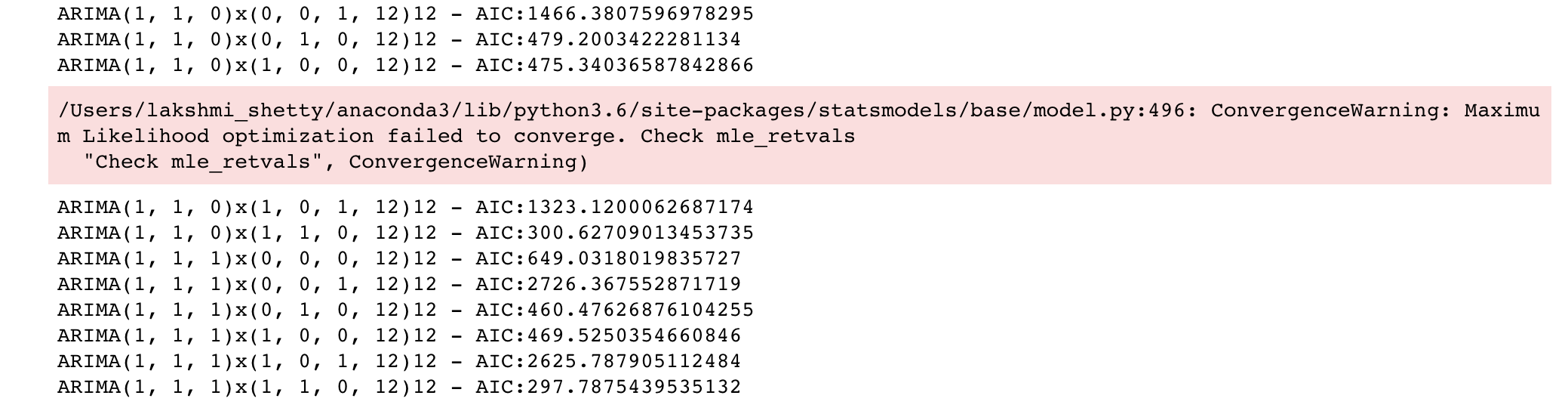


Above image depicts the graph of furniture for the seasonality, trend and noise.

**Time-Series Forecasting using ARIMA**

**ARIMA** the name stands for Autoregressive Integrated Moving Average technique. This technique is used for fitting the time-series either to understand the data and predict the future values (forecast analysis). This model can be applied on the dataset more than once when the data is non-stationary.

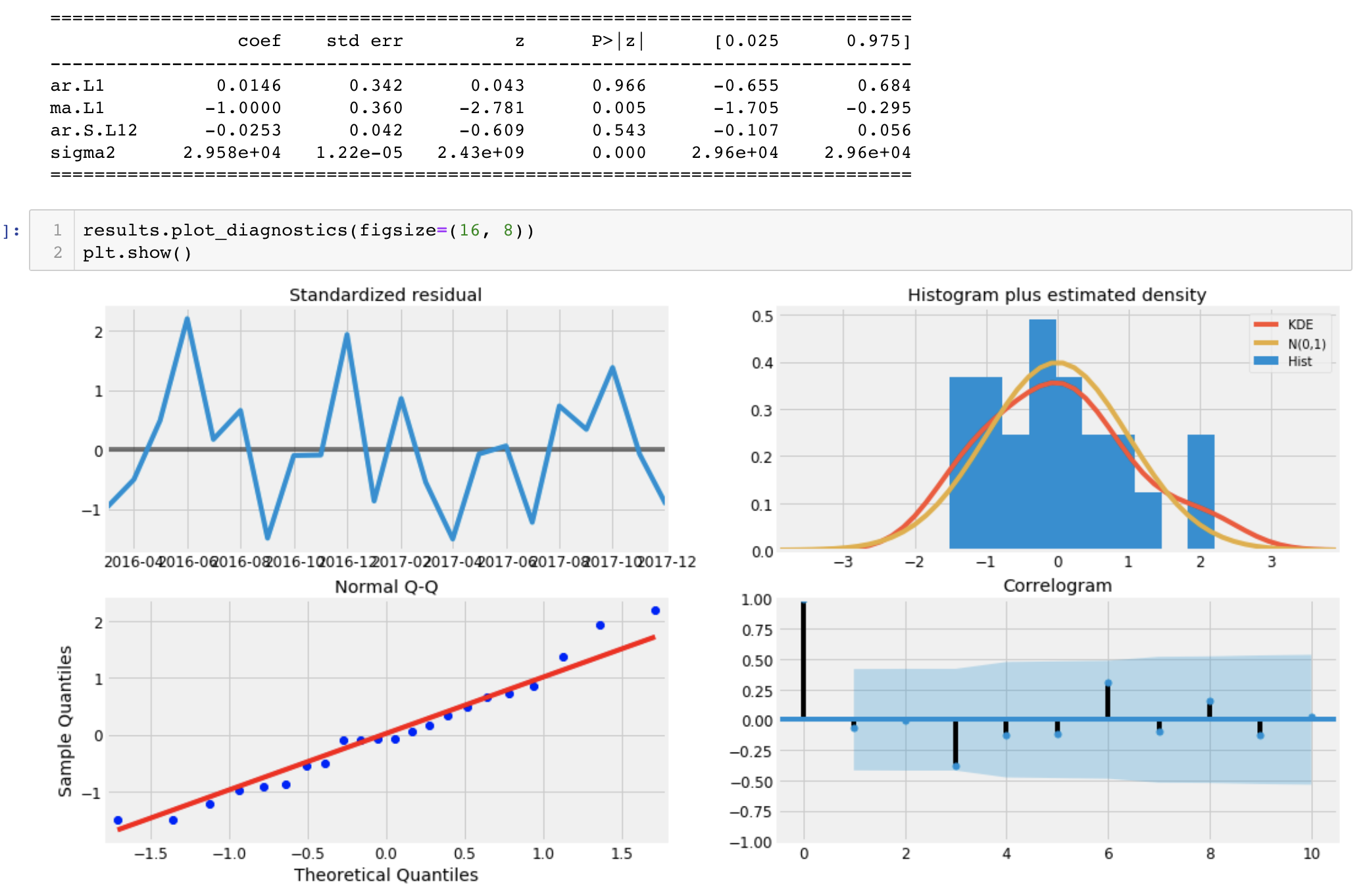
ARIMA model will be denoted with the notation of ARIMA (p, d, q). where those are the parameters and stands for seasonality, trend and noise in the data. In this step we will calculate the parameters for our model, for that we will use grid search technique to find the optimal parameters which gives better performance. From the below image we will consider the AIC value of 297.78 which is lowest for our model.



The parameters are calculated and few are mentioned in above screenshot, we consider the ARIMA (1, 1, 1)x(1, 1, 0, 12)12 – AIC297.78 as it is optimal.

**Fitting the ARIMA Model**

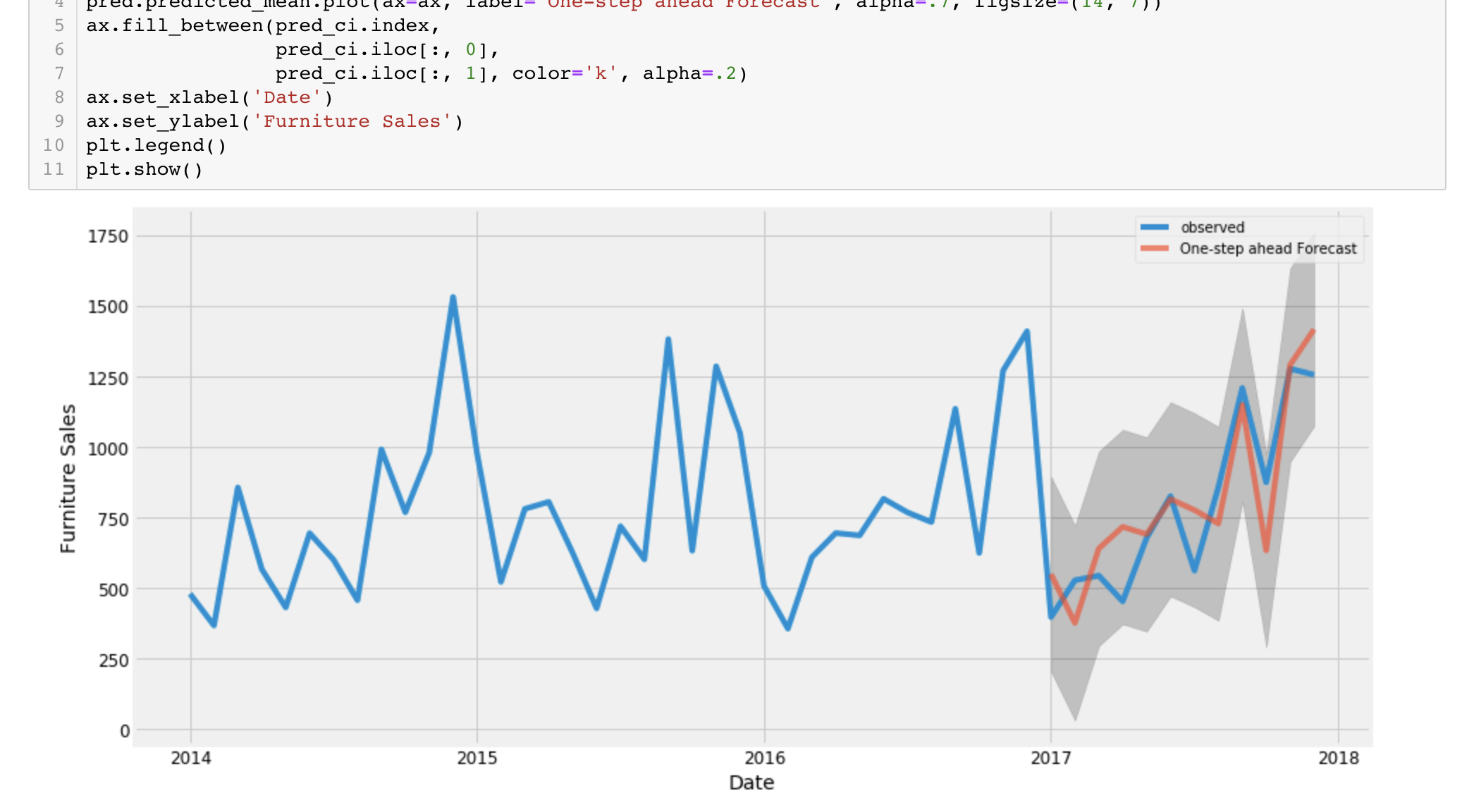
The model is fitted, and it will run the diagnostics to find out any unusual behavior, it is not perfect however the data is normally distributed.



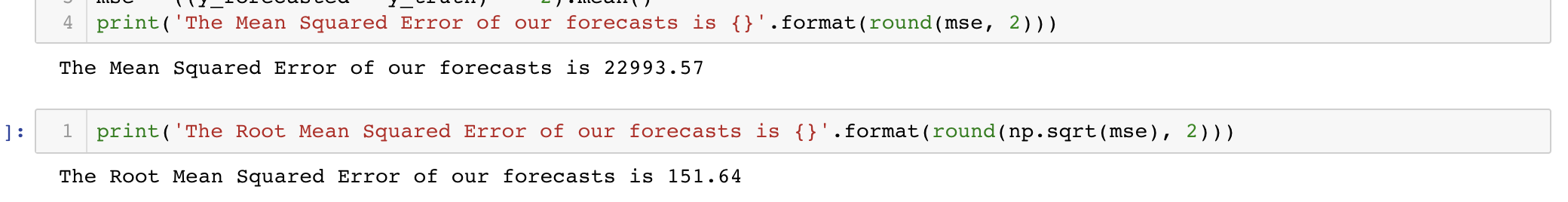
In the above model diagnostics is performed and to find the unusual behavior and residuals are normally distributed.

**Validating Forecast:**

To understand the accuracy of our model, we compare the predicted sales with real time sales, this we will do by setting the date stamp from 2017/01/01 to the end. From the graph we can observe that, observed values with the forecast values the trend is increased from beginning of the year to end of the year. Here Root mean square error and mean square are calculated.

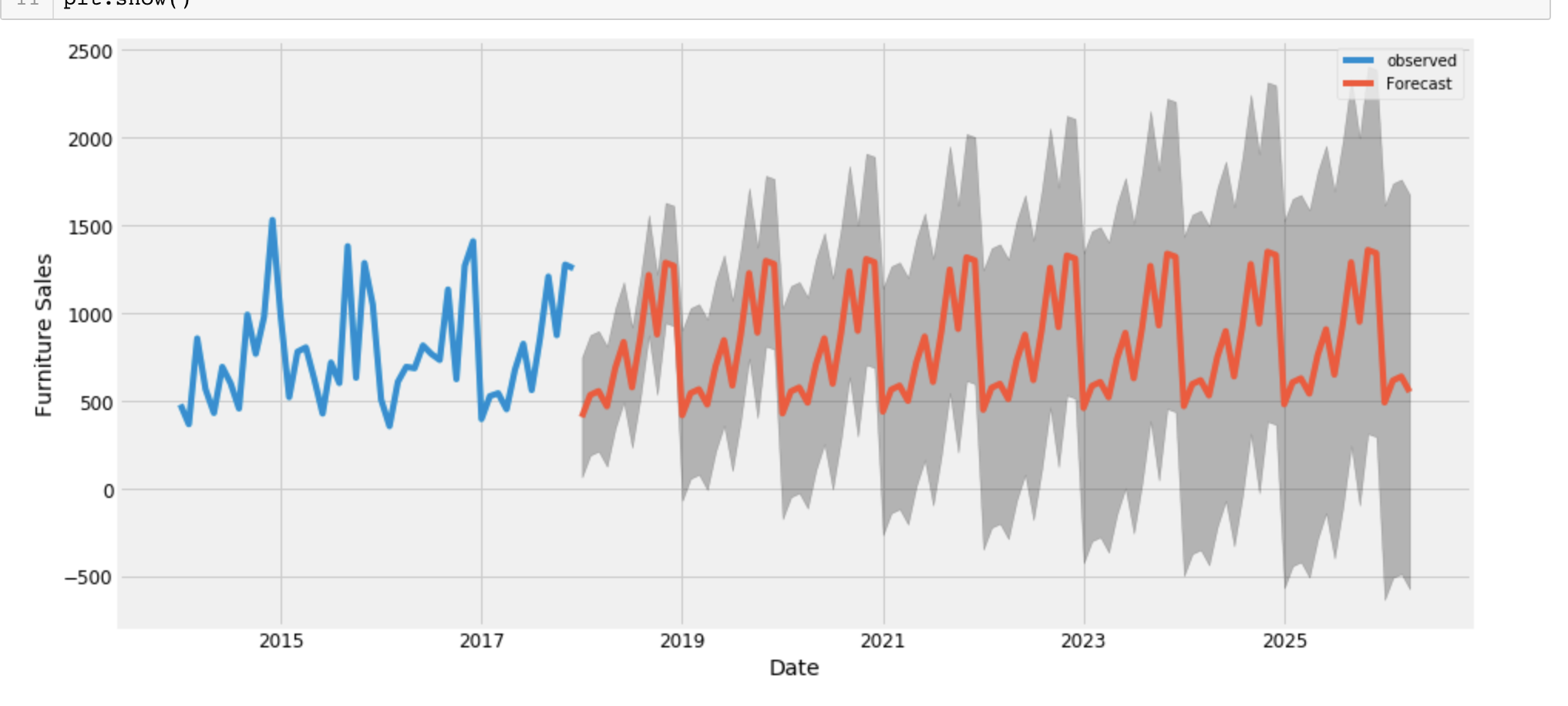


Mean square error estimates average square difference between the estimated value and what has been estimated. MSE value cannot be negative and the low MSE value close we will be finding the best fit. Our model was able to forecast the average furniture sales with 151.64



**Producing and Visualizing Forecast**

In this step we will predict the future sales of the furniture for the future years and the graph will predict them from past data and the graph gives better seasonality at the end of the year and in goes up and down in the middle of the year.



**Conclusion:**

From the graph we can mention that forecast has predicted the furniture sales data from the previous records, as the graph is increasing over the period of time in the year. And better performed for the seasonality. For the furniture sale went down or worst in the month of April. I would like to conclude the sales will not decrease for the furniture in future. In future we can compare the sales of furniture and office supplies and perform the time-series analysis on the combination of furniture and office supplies.

# Bibliography

*Analytics Vidhya*. (n.d.). Retrieved from https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

*Tableau Community*. (n.d.). Retrieved from Super store : https://community.tableau6.com/docs/DOC-123