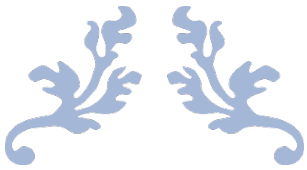


SALES FORECAST - TIME SERIES MODELING

SUBHASHINI NATARAJAN

JULY 31, 2022

SUBHASHINI NATARAJAN

Time Series Modeling objective

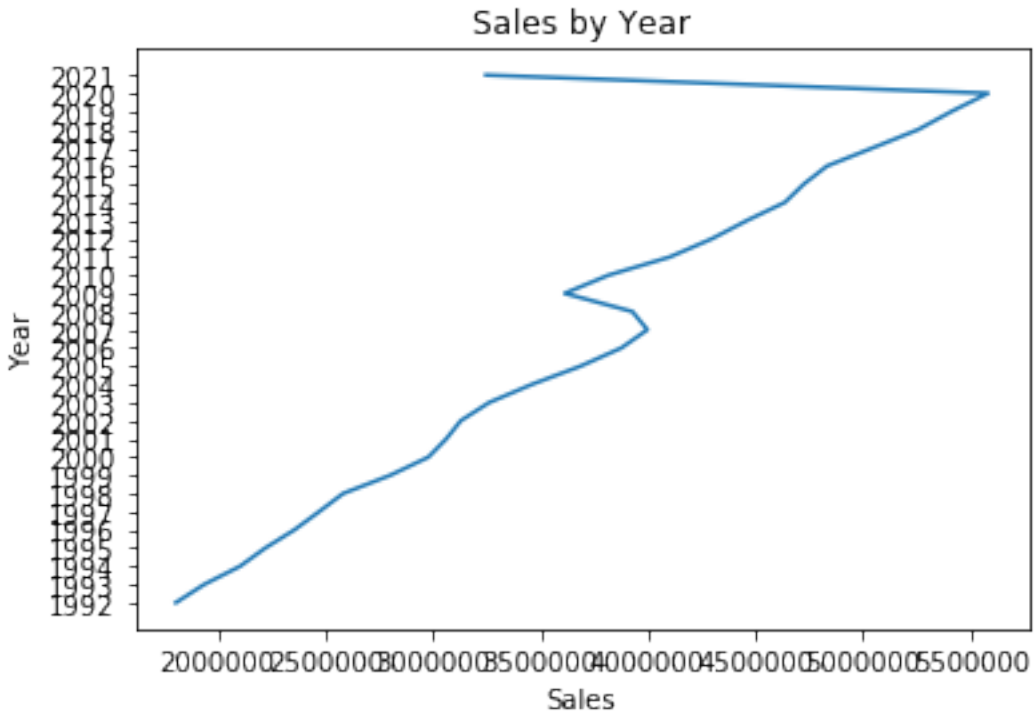
Time series analysis helps forecast data patterns by focusing on series of data points in time. This is one of the widely used modeling techniques used in several businesses like forecasting sales and stock prices. The current week assignment includes in performing the time series analysis with retail sales data. The document details the process followed for building the timeseries model for retail sales.

Data Ingestion & Transformation

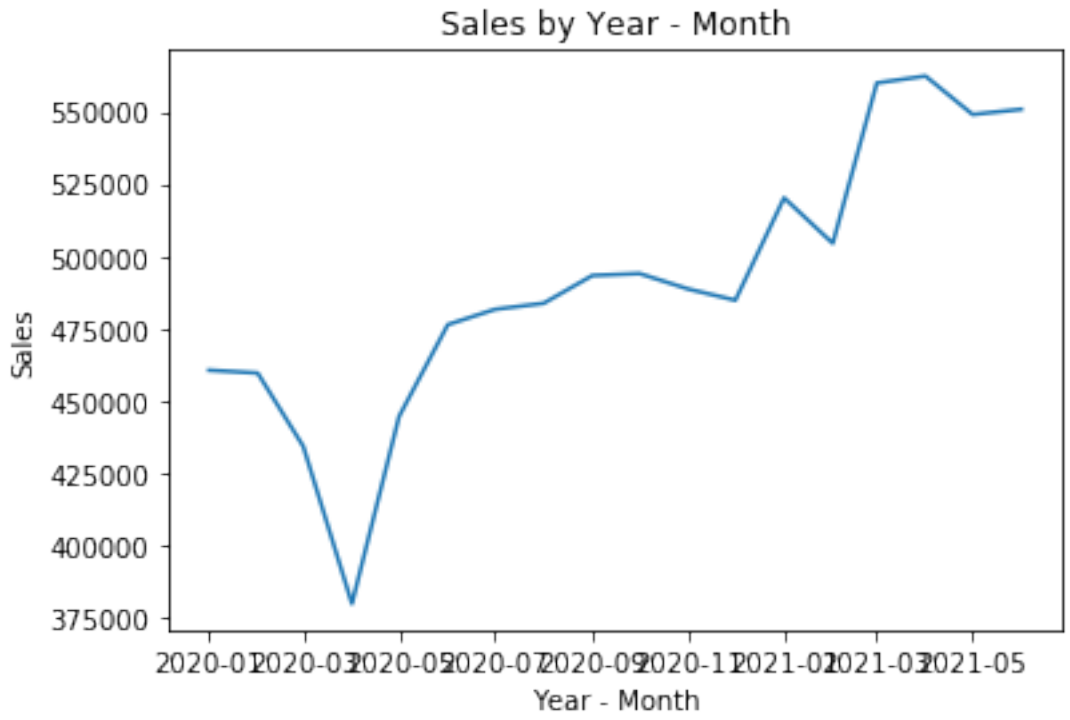
- The dataset provided for retail sales is loaded into a data frame.
- The data set is transposed to capture the monthly data as rows using melt method.
- Data type of the field, Year is changed from integer to string.
- A new field is created with Year-Month data and the field is converted into date.

Exploratory Data Analysis

1. In order to assess data trend, the sales volume is summarized by year and plotted against year. It can be observed from the graph that the sales have a steady increase until 2009. There is a drop-in sale in 2009 due to recession. There is a steady increase again until 2020 and there is a drop-in sale.

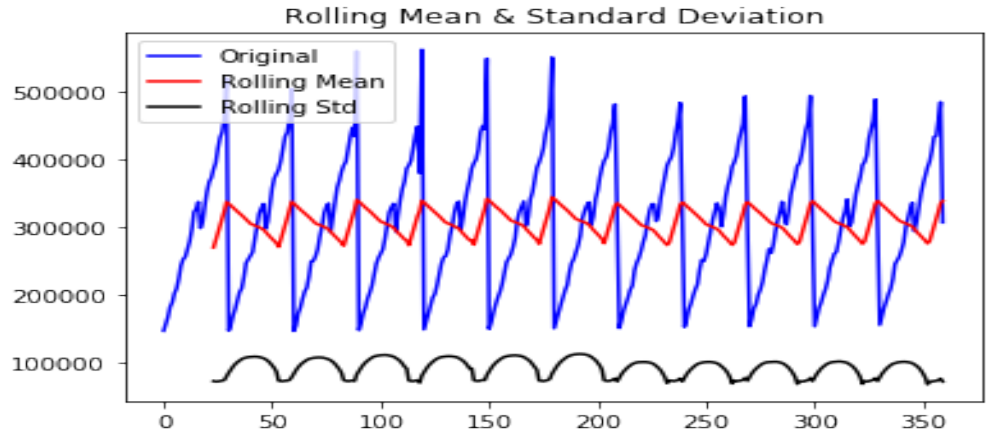


2. In order to assess the trend by month, the last two years of data is filtered into a new data frame and the data is plotted between Year Month and Sales. It can be observed that there is steep drop in sales by March 2020, possibly due to covid onset.



Evaluate the stationary nature of data

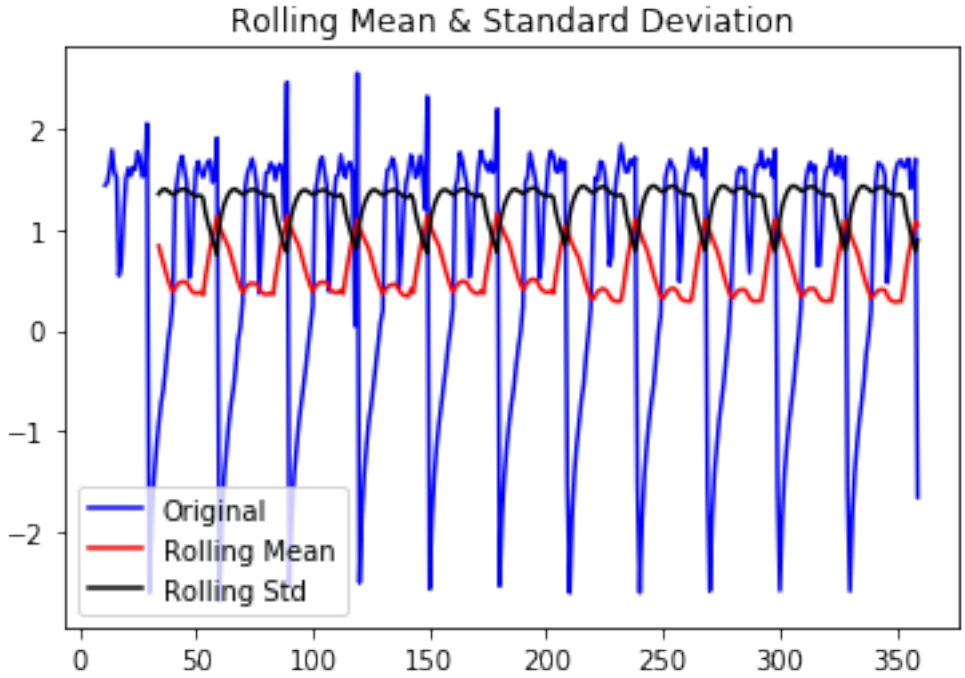
The most important pre-requisite for employing time series, is the stationary nature of data. The data is evaluated by observing rolling mean and standard deviation plotted with original data. It is also evaluated using Dickey – Fuller test.



Results of Dickey-Fuller Test:

```
Test Statistic                 -9.454956e+00
p-value                         4.511030e-16
#Lags Used                      1.700000e+01
Number of Observations Used     3.420000e+02
Critical Value (1%)            -3.449616e+00
Critical Value (5%)            -2.870028e+00
Critical Value (10%)           -2.571292e+00
dtype: float64
```

Dickey-Fuller test shows that test statistic is lesser than the critical value and hence the data is stationary. However, in order to smooth the rolling mean and standard deviation, will perform de-trending. Data after de-trending is observed to be smoother.



Choosing best parameter values

The parameters for SARIMA model are p, d and q.

P – Number of auto regressive terms – Lags of the response variable.
Q – Number of moving averages – Lagged forecast errors in the response variable
D – Number of differences – Order of differencing

The best parameter values are those that produces minimum AIC - Akaike Information Criterion. The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data.

For the given data set, the parameter values of 1,1,1 provide the lowest AIC -5867.7

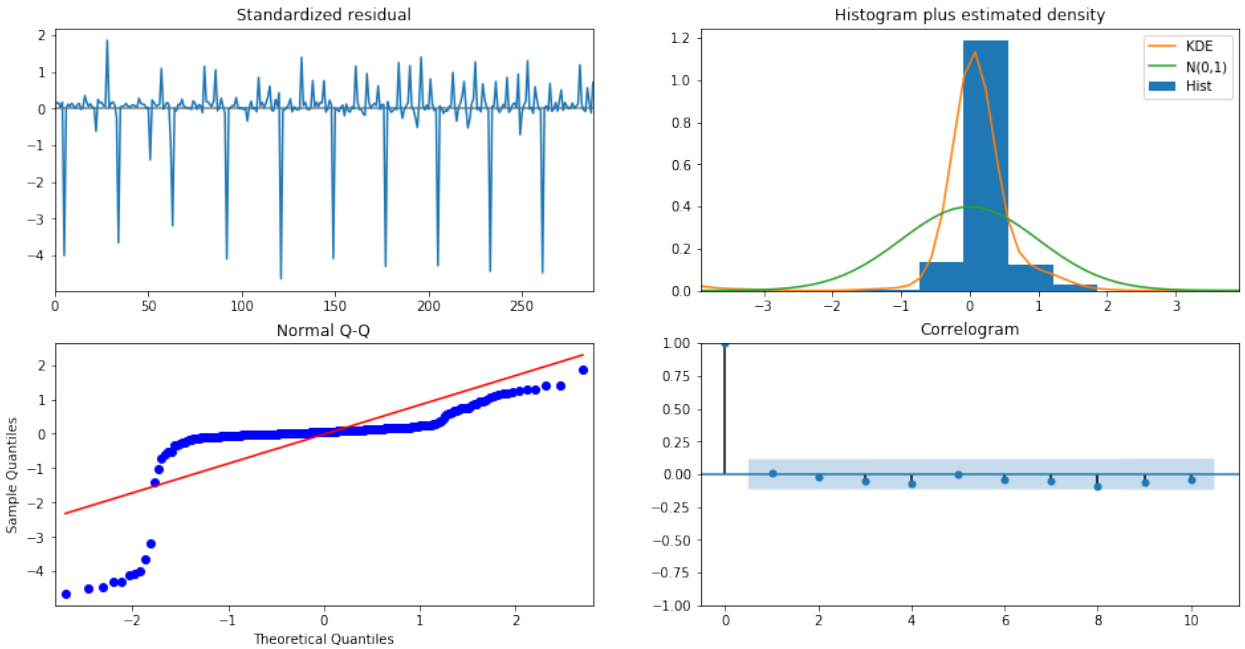
Model Building

The data is split into train and test sets. The last year of data between July-2021 and June 2022 is taken as test set and the rest of the data ae used for model training.

The SARIMA model is built and fit on training data. The diagnostic details are presented below –

```
==========================================================================
====
                 coef    std err          z      P>|z|      [0.025      0.
975]
--------------------------------------------------------------------------
----
ar.S.L52      -0.1524      0.093     -1.633      0.102      -0.335       0
.031
ma.S.L52      -0.7349      0.079     -9.300      0.000      -0.890      -0
.580
sigma2     4.999e+09   3.25e-11   1.54e+20      0.000       5e+09       5
e+09
==========================================================================
====
```

The diagnostics are plotted to validate the data assumptions.

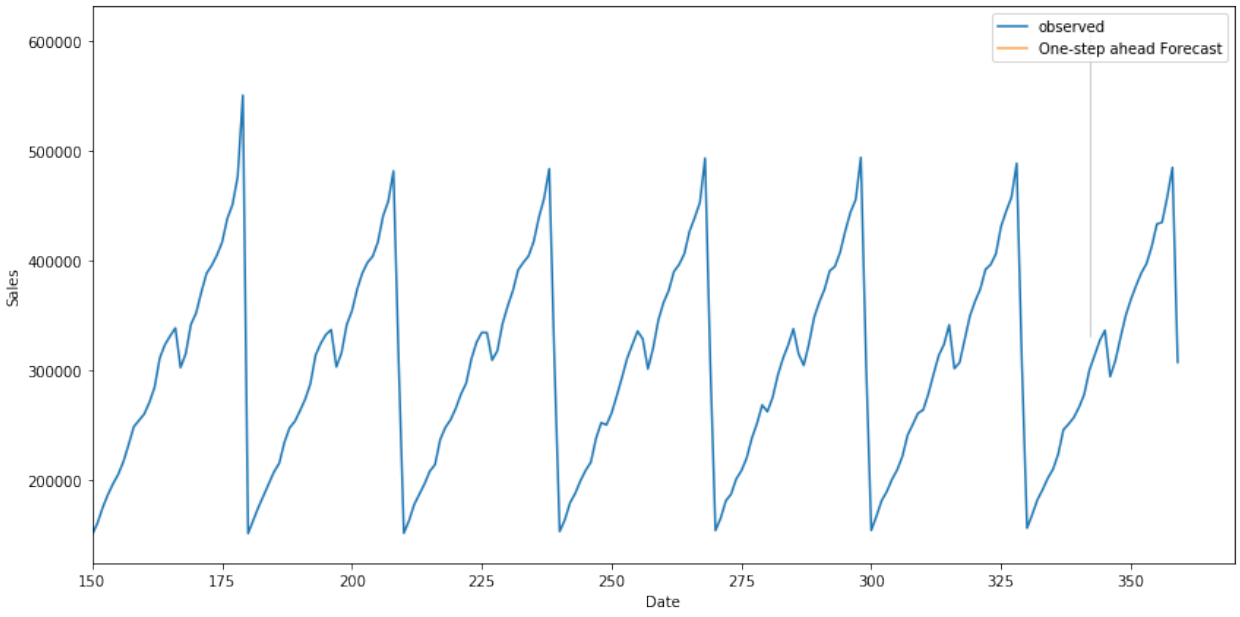


1. Standardized Residual plot indicates the seasonality in data.

2. In histogram, N(0,1) line is not as close to KDE line. This indicates that the data is less normal than optimal.
3. QQ Plot shows the distribution of residuals along the linear trend line.
4. Correlogram indicates that the residuals have low correlation with lagged versions.

Prediction using the model

The model is fit with test dataset and evaluated. The RMSE value for the model is 50238.43. The high value indicates that the model might not be a good fit for the dataset. However, based on the trending, pre-covid forecasting is applied onto post covid environment. Hence, likely result is to be obtained with other models as well.



References

The below listed links were used for the model building and evaluation.

https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/

https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced