Time Series Forecasting

ARIMA Model:

For the ARIMA model, we need three hyperparameters that are d, p, and q. For estimating d AIC was optimized which can be found in the adfuller() function. Used by shifting the data values by different values of d and for each d AIC value was noted and the minimum value of AIC is used for differencing the time series to create a stationary time series.

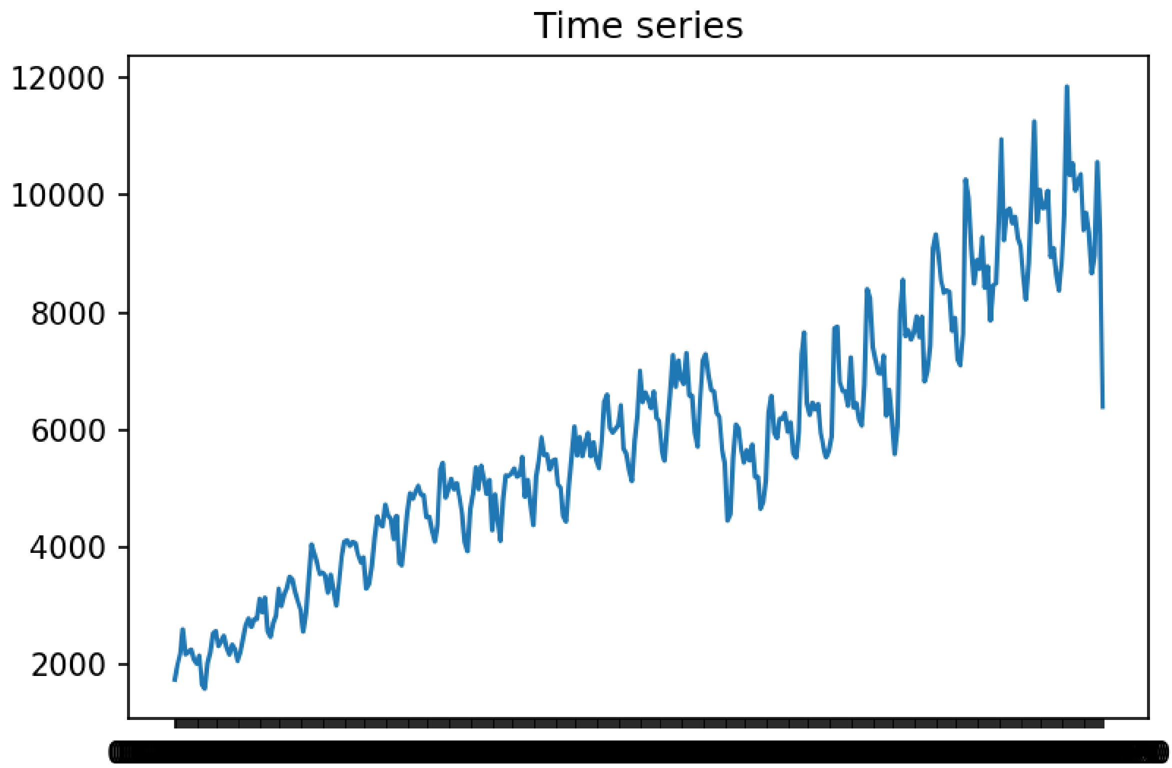
For estimating the values of p and q, the first autoregressive part was done using data and fitting a linear regression model for different values of p and then choosing the value of p which gives the least root mean squared error over the test data and predictions. Train test split was 70 to 30. Similarly, the Moving average part was done using the errors in the data and predictions and finally, q was obtained.

This whole procedure was done keeping in mind the ARIMA model which can be found as,

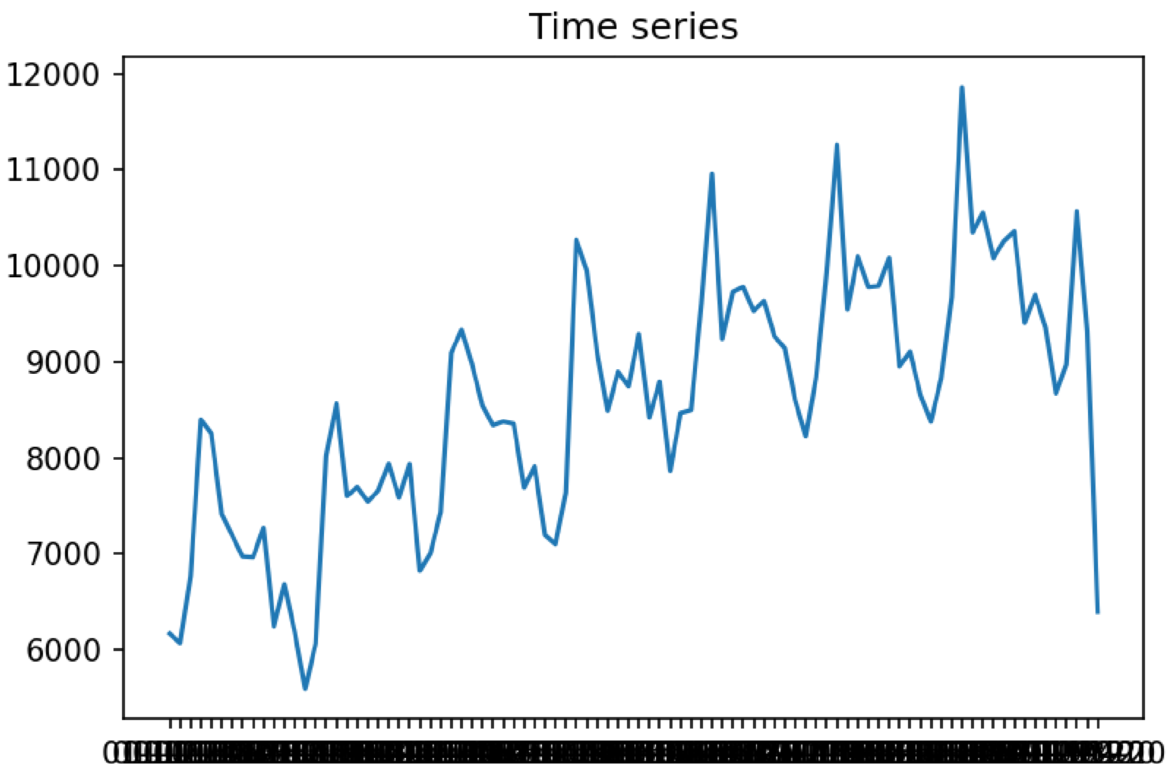
$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Observations and Results:

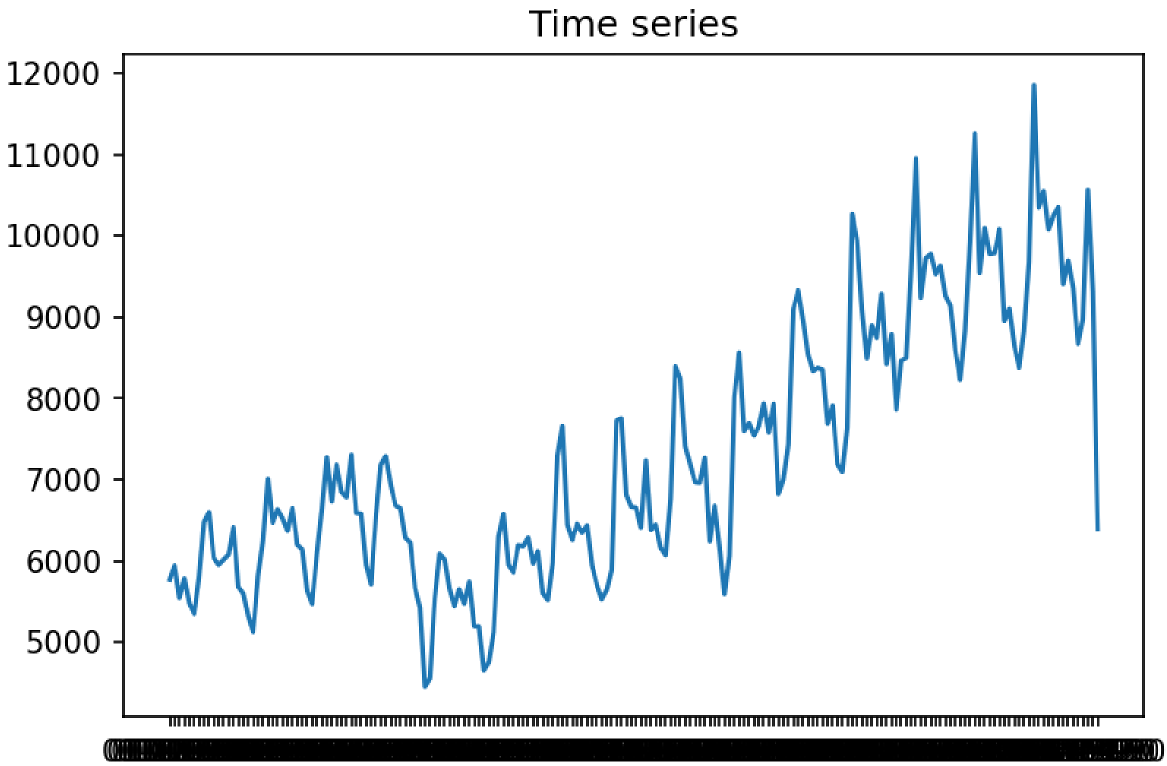
- For finding the value of d (differencing) AIC was optimised, it was observed that if the data has one seasonal period then d was equal to 1. Where as if the data has more then one seasonal period then value of d was greater than 1.
- For complete data d = 6



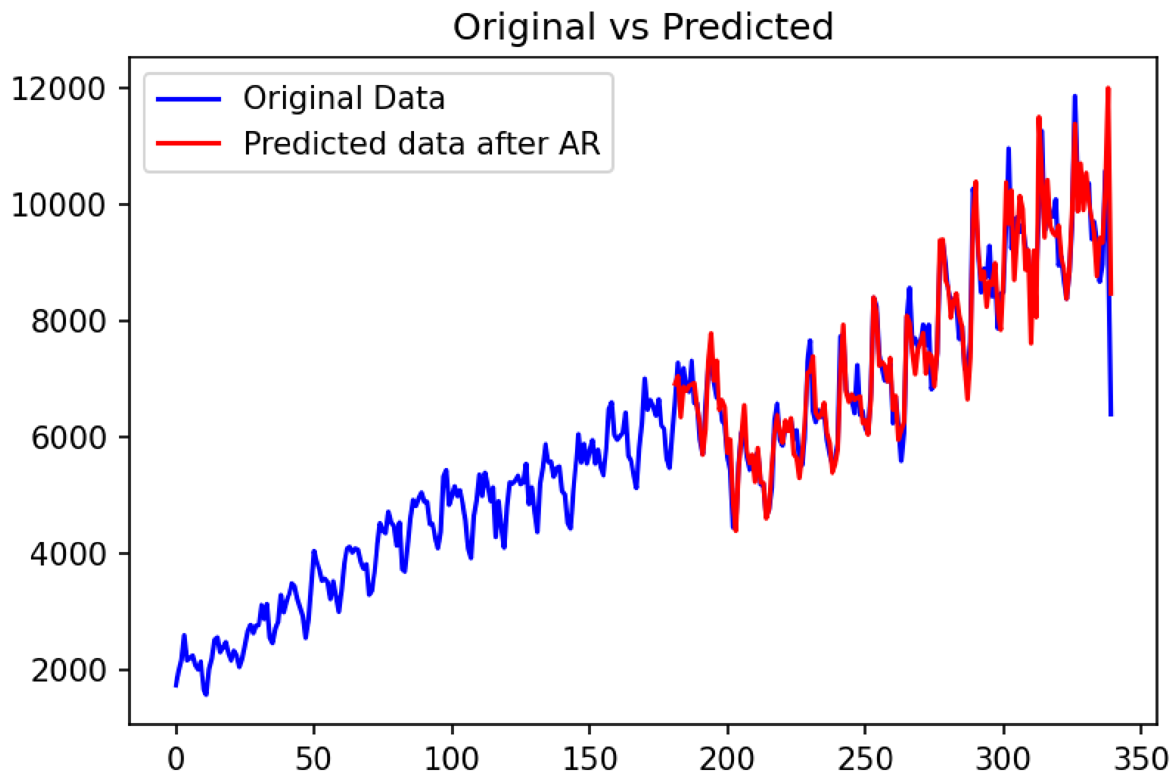
- For a segment of data which is nearly stationary d = 1



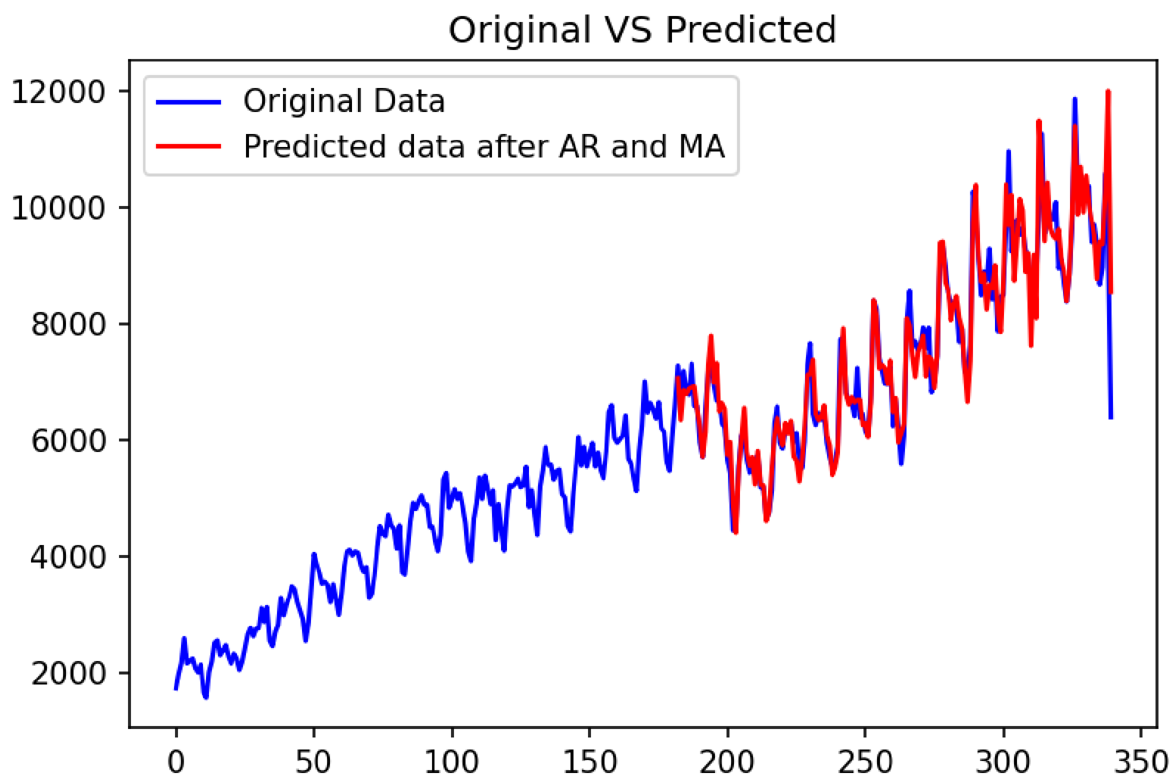
- For data having no seasonality like below d = 9 was observed.



- We can also use the p value returned by adfuller to check for stationarity using this the value of d was found different than while using AIC in the sense that using AIC gives more value of d for every case than using p value.
- While training the data for auto regressive model the data was shifted by p and then linear regression model was fitted on that data and values were predicted and root mean squared value was calculated. The p value giving the least value of root mean squared was selected as value of p.
- Like for below data p = 22 was observed.



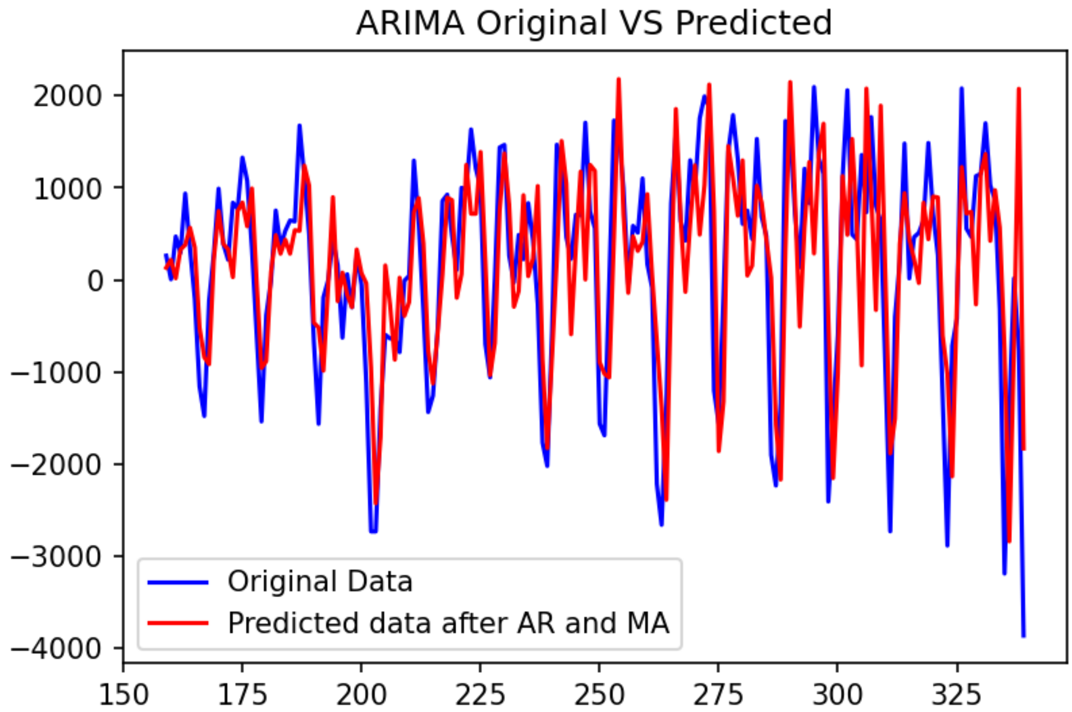
- While running the moving average part of ARIMA model similar to auto regressive part the liner regression model was used on the value of error obtained after the auto regressive part.
- For the same data as above the q value was obtained as 1.



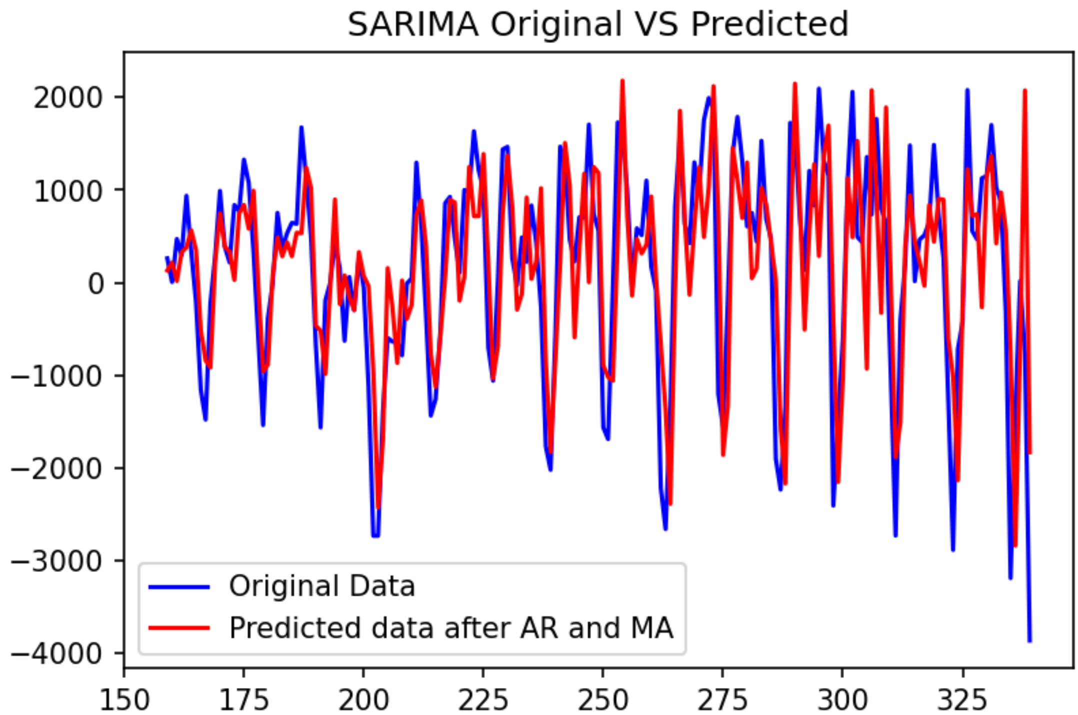
- The value in the **red are forecasted values** and the values in the **Blue are actual values**.

Comparison with inbuilt model:

- For comparison the inbuilt ARIMA and SARIMA model were aslo used the differencing for ARIMA was set as the original value which was observed above.
- The value of p and q were set to 1 and 2 respectively.



- For SARIMA value of p,q,d were set to 1, 1, 2 respectively and value of P, Q, D, S were set to 1, 1, 1, 12 respectively.



- As we can observe the ARIMA and SARIMA both fit the data with better accuracy and predictions are also better.

Holt Winter's Model:

For Holt winter's method of triple exponential smoothening we need to estimate the parameters alpha, beta, and gamma also with seasonality period. For estimating the seasonality period correlation can be used or the time series can be decomposed into particular seasonal periods.

For estimating the values of the alpha, beta, and gamma Scipy.optimise method was used where the cross validation function was used with initial values of parameters as 0, 0, 0 respectively and the values of parameters were bounded between 0 and 1. In cross validation the splits were minimum 3 and then the mean squared error loss was used to check the accuracy of the model. The values parameters giving minimum mean squared error was used for the final model.

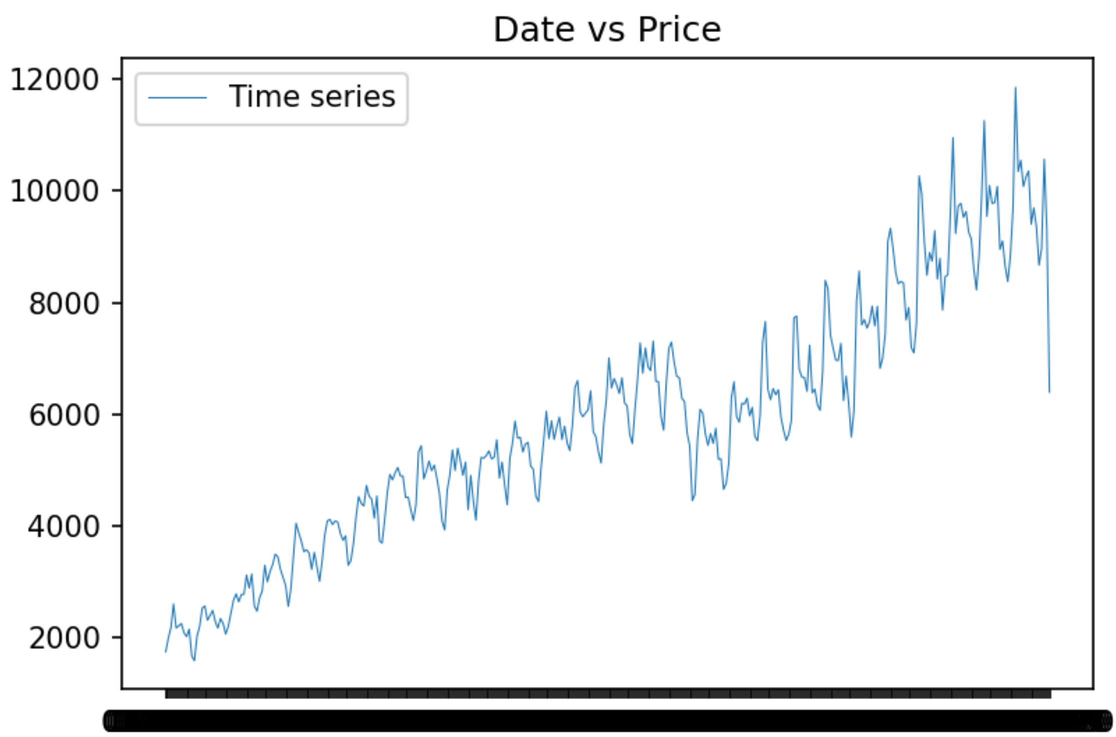
The Holt winter's method was used as:

For smoothness	Alpha	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},$
For Trend	Beta	
For Seasonality	Gamma	

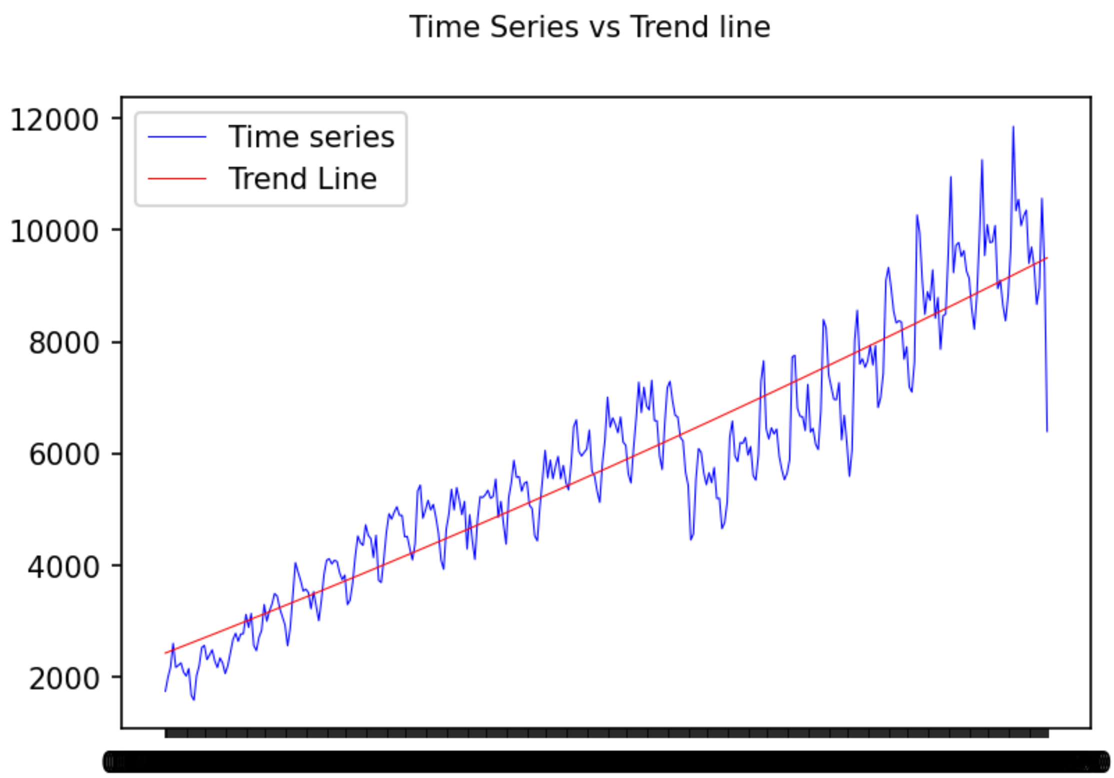
Here the value of y is the predicted or forecasted value and this model is additive model it can also be multiplicative but for this assignment additive model was used.

Obervations and Results:

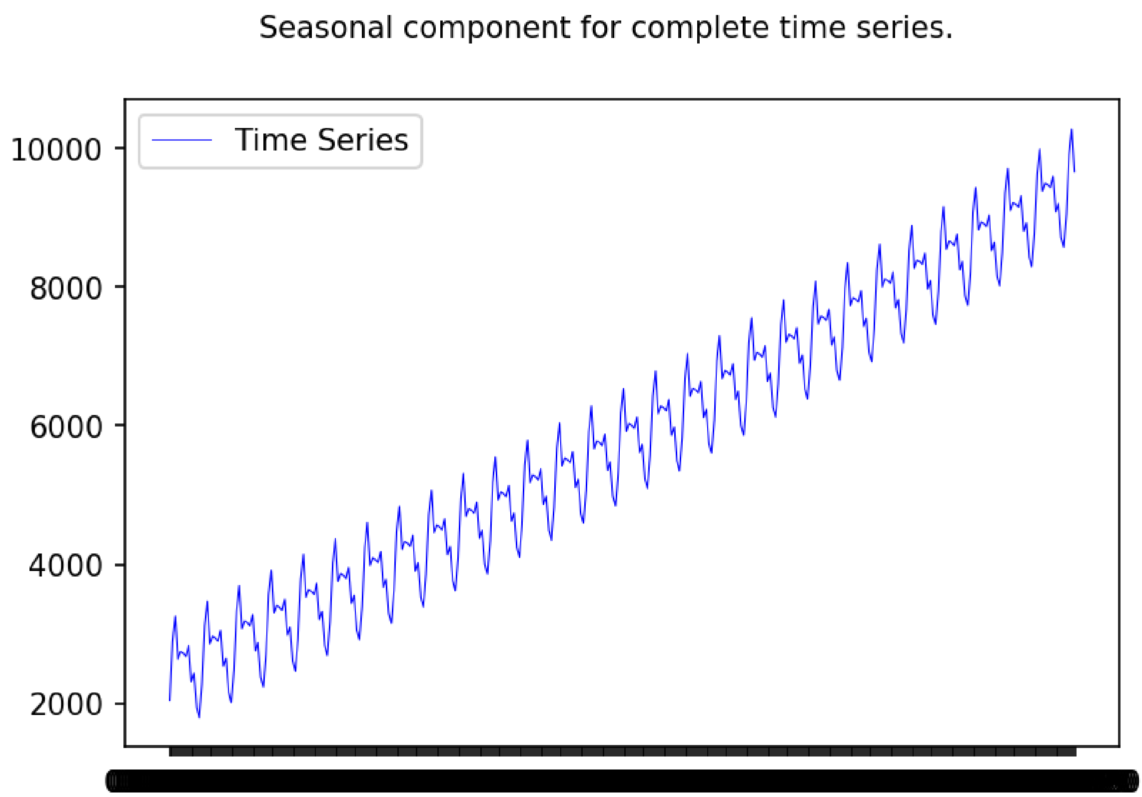
- Initial data can be shown as.



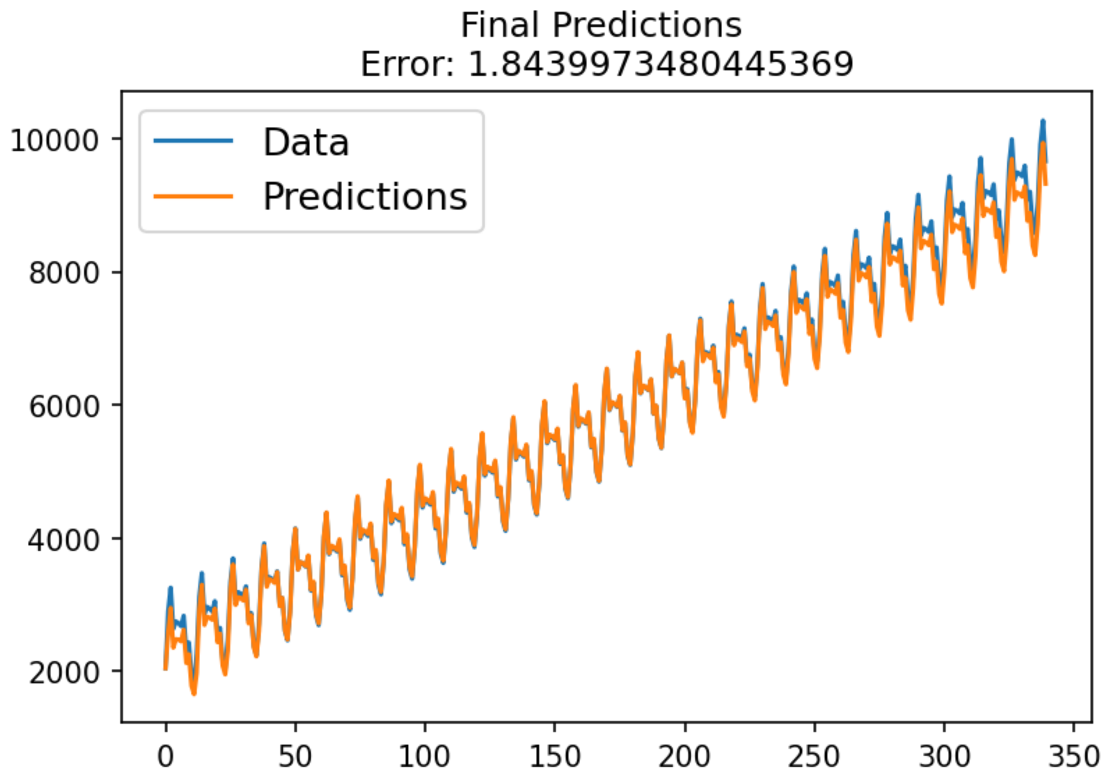
- For creating a seasonal data with 12 as period first the data was detrended by finding the trend. This was done by fitting a polynomial to the data as.



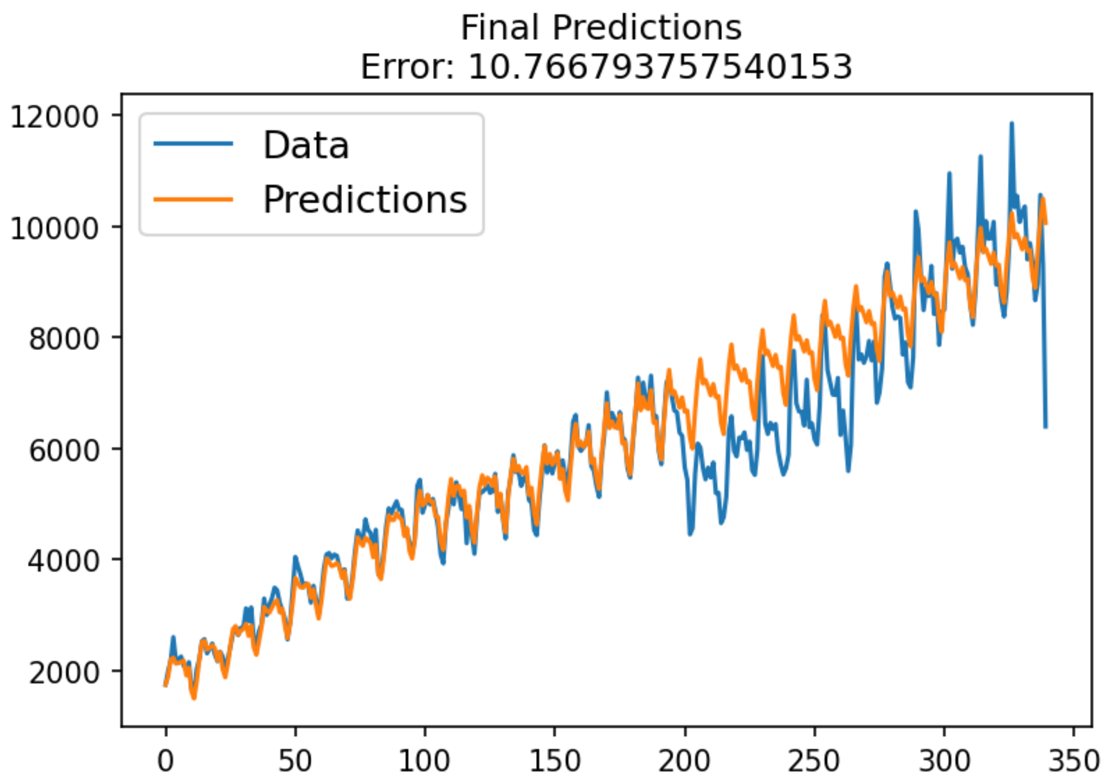
- After this the seasonal data was calculated as it was found out as.



- The model was used on both the seasonal data and the original data. While using the seasonal data and peroids as 12 the output has 0% error.



- While using the original data and 12 as periods 10% of error was observed.



- By seeing this we can observe that for a data having perfect seasonal period the model will be forecasting the values with very high accuracy where as if the data has different seasonality period like our original data the predictions will be less accurate.
- We can also say that using cross validations the training is very efficient than using other estimators like regression model etc.

- But one error was faced for small data that while fitting the model the value of data should be a whole number and should be a multiple of the seasonality otherwise frequent error were observed since we need this periodic data and when the values of periods are not present there would be error.

Dataused:

The data set used here for both models is an example of sales of a product in different years and months. The value of price can be considered as the mean or average price os sales in a month. The data is collected from 1992 to 2020. Upload the **Data.csv** to run the program.

Colab files: can be found at.

Link1

Link2

Conclusion:

Using both arima and holt winters we can conclude that these methods are very much accurate for forecasting a time series.