**OE3M31-Machine Learning**

**Machine Learning algorithms for forecasting blood demand data with Missing values and outliers**

**( Supply Chain : The Study of Demand Forecasting)**

horizontal line

# 

# Sakshi Palekar

# Under Dr. Sunil Agrawal

# 

# 

# Introduction

In this Machine learning (ML) Project , we were asked to experiment with real world blood demand data from Tema General Hospital of Ghana and to explore how machine learning algorithms can be used to Predict the future data by forecasting of blood demand series , after that we need to do time-series modeling , data pre-processing is thus vital for model selection, parameter estimation, and predictions . so After performing the required tasks on a dataset of my choice, here lies my final report.

We have the blood demand data of Patients with missing values and outliers so by using forecasting and Data Cleaning in machine learning will manage a future demand of blood supply and previous year’s data gaps.

Keywords : Machine Learning , Classification , Blood demand , Blood supply , Forecasting , Time-reversibility

## 

## 

## 

# 

# 

# Problem and Dataset description

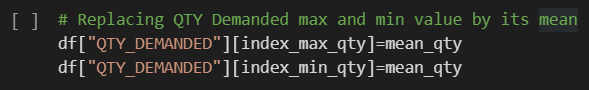
Monthly data collected from January 2013 to September 2020 on blood demand was obtained from Tema General Hospital for this ,The time-series data for this study is short in length with missing values and outliers study.

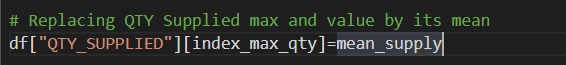
Visit the Data Set :- <https://colab.research.google.com/drive/1lFT_Xdh0zxEMZPirh-Z6IpieLV0RFO2M#scrollTo=3bca46e0>



**Removing Outlier**

Removing the highest and lowest value and replacing it with a mean for a graph with less low and highs.





**Data Cleaning**

The data cleaning or preprocessing was done in two stages using existing state-of-the-art algorithms. The blood delishows the time-series plot of the blood demand series with highlighted missing values n series had 14% missing values (n=13) and 11 gaps with an average gap size of 1.182. In the first preprocessing stage, missing value imputation was done by Panda’s Method (i.e Median )

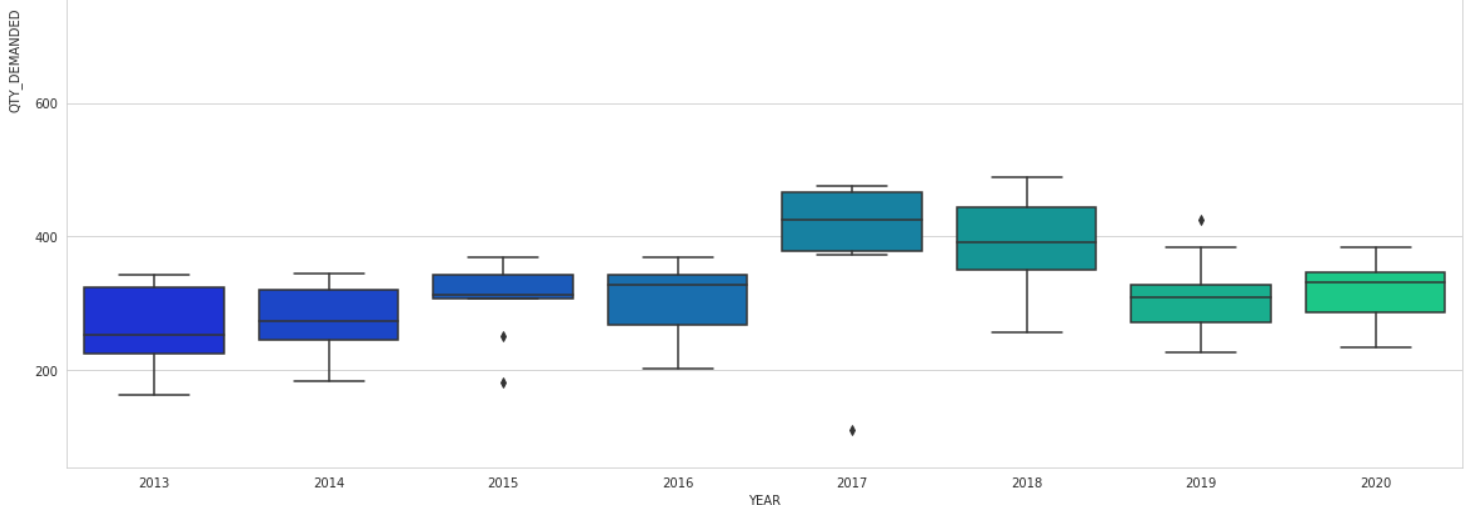
We have 4 features year, Month, QtyDemanded, QtySupplied

Out of which we have null data in QtyDemanded and QtySupplied.



For Quantity Demanded and Quantity Supplied

We have filled in null values with respect to the mean of the value of the respected year ranging from 2013 - 2019.



Replaced all null values as shown in heatmap



# Algorithm and technique description

ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values

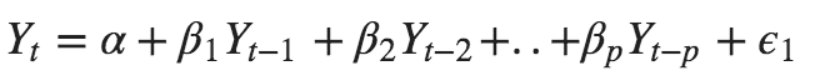
An ARIMA model is characterized by 3 terms: p, d, q

Where,

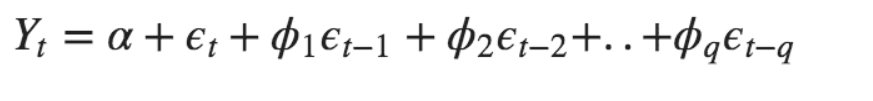
1. p is the order of the AR term
2. q is the order of the MA term
3. d is the number of differencing required to make the time series stationary

The first step to build an ARIMA model is to Make the time series stationary .There are two Models i.e AR and MA models.

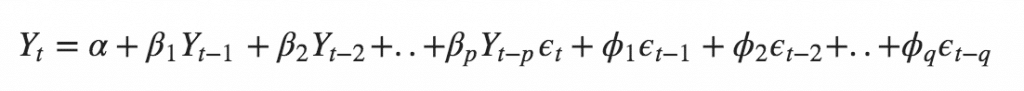
**Auto Regressive (AR only) model** :- It is one where Yt depends only on its own lags. That is, Yt is a function of the ‘lags of Yt’.



**Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.



An ARIMA model is one where the time series was different at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes.

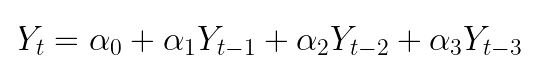


**Predicted Yt** = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

**ADF** :- This test is that the time series is non-stationary. So, if the p-value of the test is less than the significance level (0.05) then you reject the null hypothesis and infer that the time series is indeed stationary.

**ACF** :- It tells how many MA terms are required to remove any autocorrelation in the stationarized series

**PACF** :- Partial autocorrelation of lag (k) of a series is the coefficient of that lag in the autoregression equation of Y.



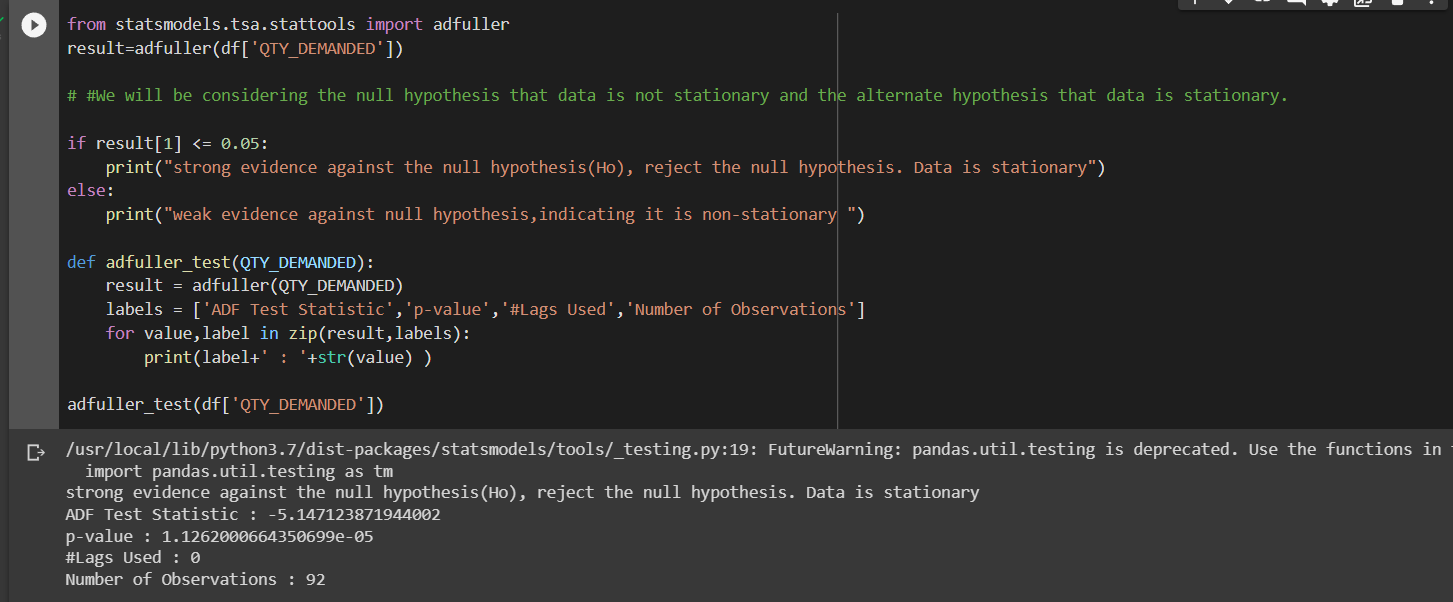
## Accuracy Metrics for Time Series Forecast

1. Mean Absolute Percentage Error (MAPE)
2. Mean Error (ME)
3. Mean Absolute Error (MAE)
4. Mean Percentage Error (MPE)
5. Root Mean Squared Error (RMSE)
6. Lag 1 Autocorrelation of Error (ACF1)
7. Correlation between the Actual and the Forecast (corr)
8. Min-Max Error (minmax)

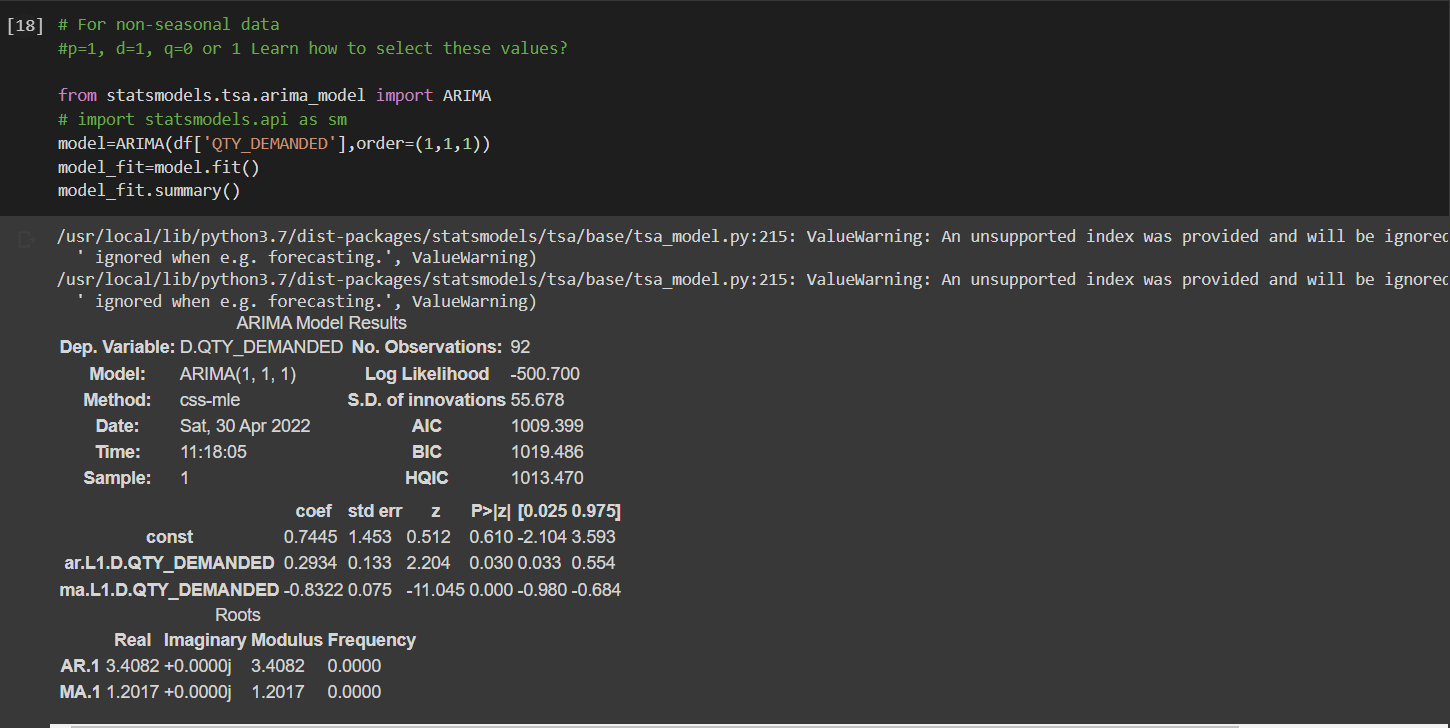
Typically, if you are comparing forecasts of two different series, the MAPE, Correlation and Min-Max Error can be used

# Model fitting, Results and Analysis

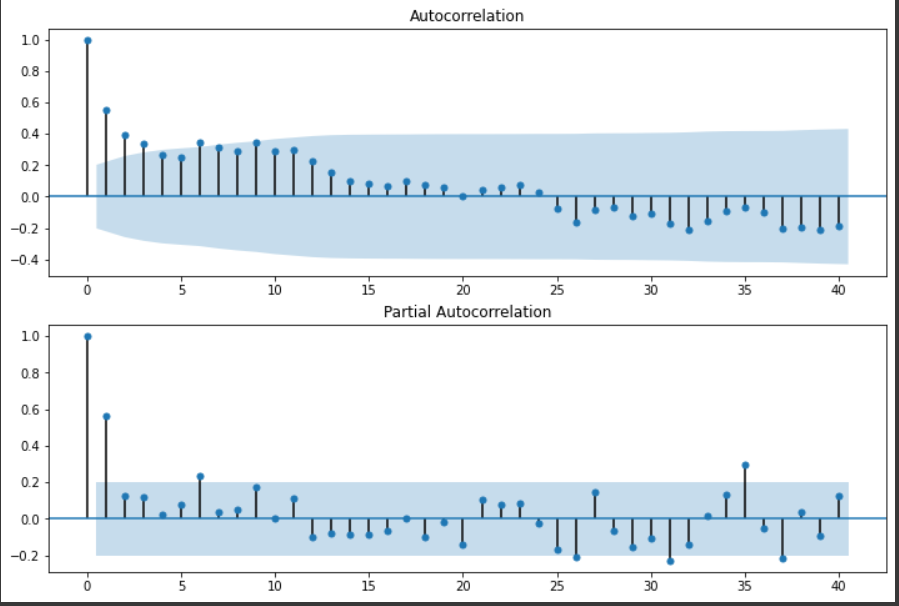
If P Value > 0.05 we go ahead with finding the order of differencing. Data is stationary , ADF Test Statistic : -5.147123871944002 ,p-value : 1.1262000664350699e-05

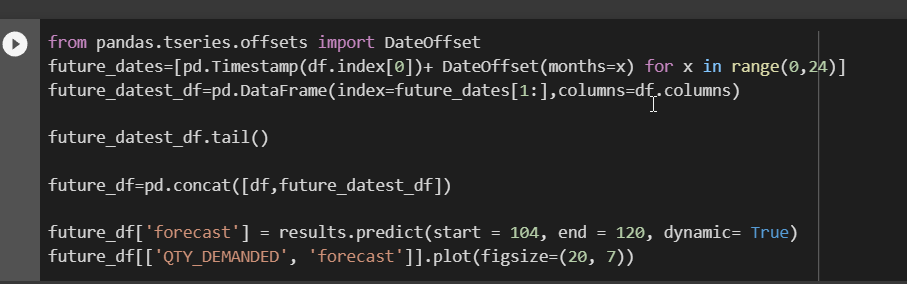


Since the P-value is greater than the significance level, let’s differentiate the series and see how the autocorrelation plot looks like.

l

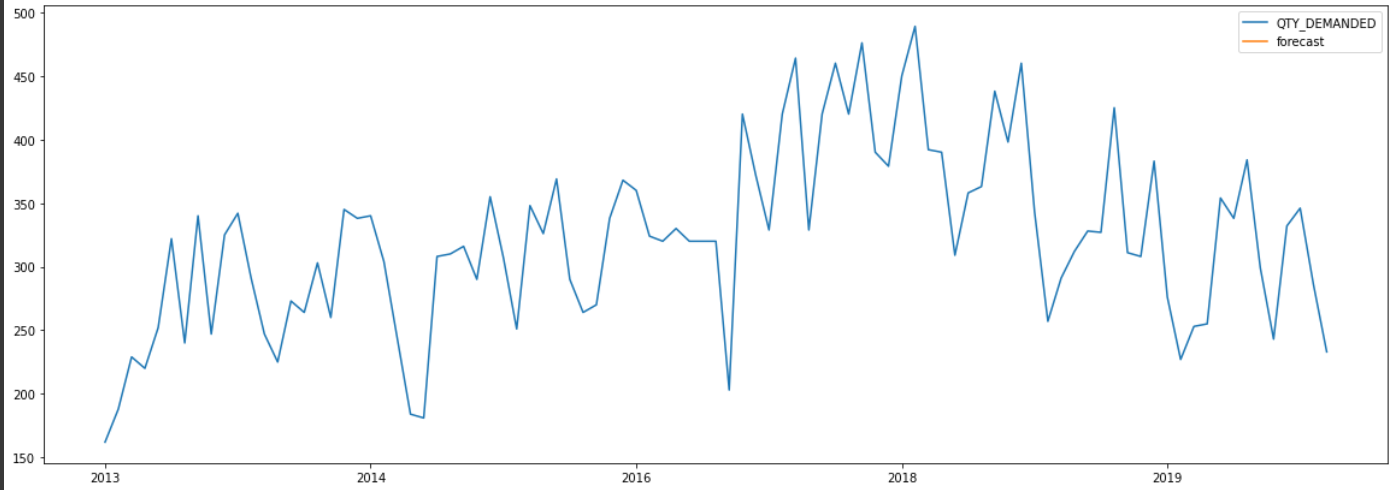
The time series reaches stationarity with two orders of differencing. But on looking at the autocorrelation plot for the 2nd differencing the lag goes into the far negative zone fairly quickly, which indicates, the series might have been over-differenced. We can observe that the PACF lag 1 is quite significant since it is well above the significance line. Lag 2 turns out to be significant as well, slightly managing to cross the significance limit (blue region). But I am going to be conservative and tentatively fix the p as 1 also Couple of lags are well above the significance line. So, let’s tentatively fix q as 2. When in doubt, go with the simpler model that sufficiently explains the Y.





The p-value is very less than the significance level of 0.05 and hence we can reject the null hypothesis and assume that the series is stationary.

Let’s visualize the series as well to confirm.



# 

# 

# Conclusion

ARIMA model in predicting future values based on the short time-series data .The data correction method could significantly affect the predictive outcome of ML algorithms and other classical time-series models. Hence, we recommend that future studies investigate the effects of different data preprocessing techniques on the time-series models’ predictive power for short or long series. Furthermore, the blood centers in Ghana should get proper database management systems to avoid data loss and outliers due to genuine recording errors. Future studies can also employ machine learning algorithms as a good alternative for backcasting past values of different time-series data with unavailable data of previous years.