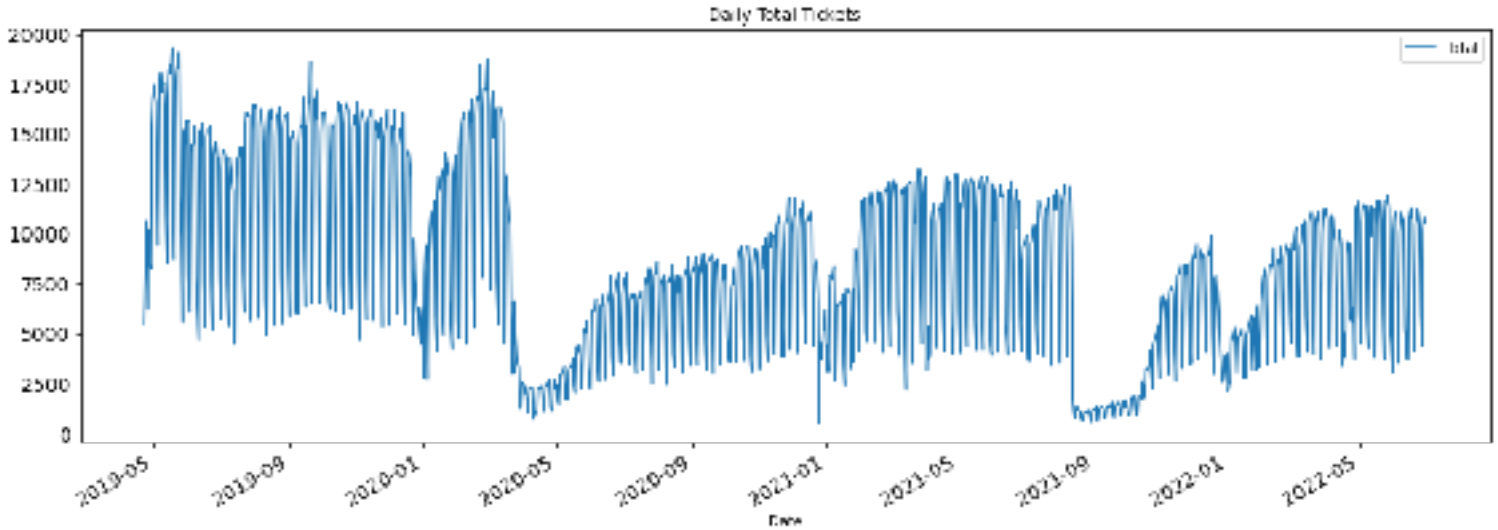
Passenger Prediction

EDA:

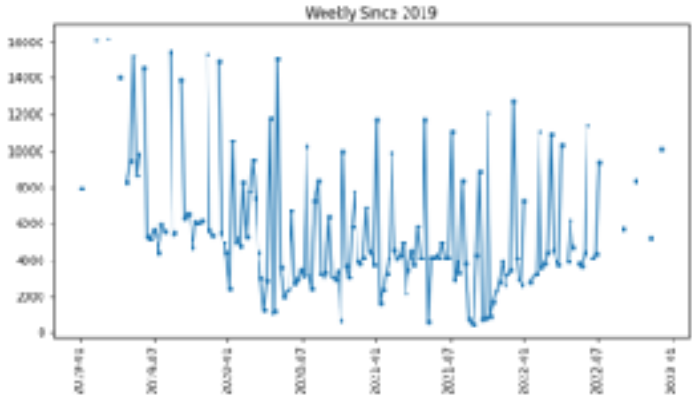
Time series data with Start : 22-04-19 , End : 30-06-22

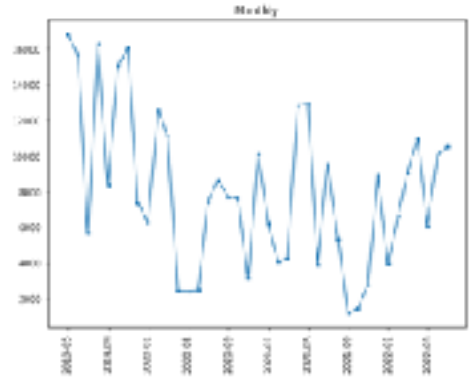
Date Cleaning : The data was sorted according to time. The total column is taken into account as the passenger count . The columns Myway and Paper Ticket were having very high discrepancies and imbalance.

Observations :



The data has no fine trend or seasonal lines and a clear drop in trend by mid 2020. It's highly volatile and was having high anomalies. On further investigation can see the data is having the dates of Corona Lockdowns. As this is a public transport system can see a clear dip in the lockdown period. This can affect the model as the model can pickup the anomaly as trend or seasonal patterns.





The weekly sales shows some peaks in the mid of the week rather than the weekends. This may be due to Corona restriction or the working community uses this rail more.

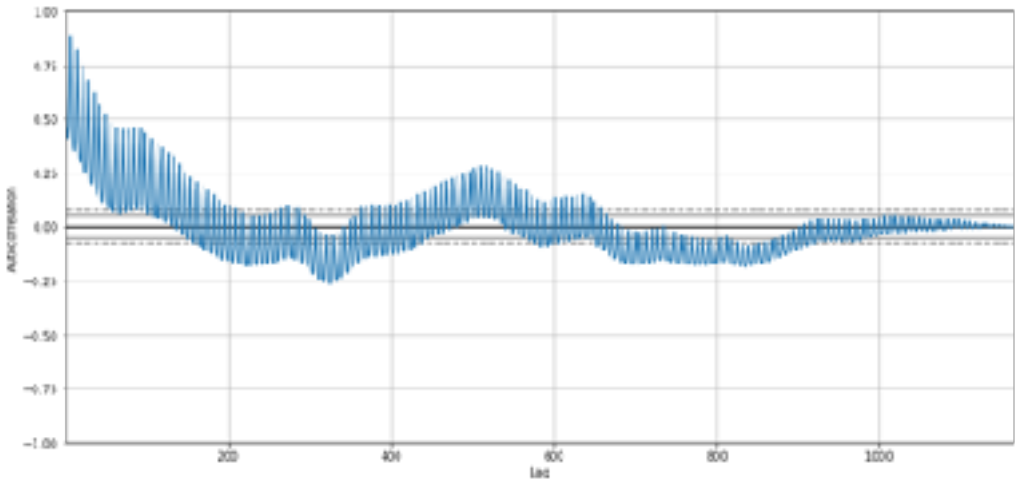
Monthly plot clearly shows there is very big dip in the month of 5-2020 and last half of 2020. There is another dip in the 9-2021 month which may have been accounted for the corona second wave

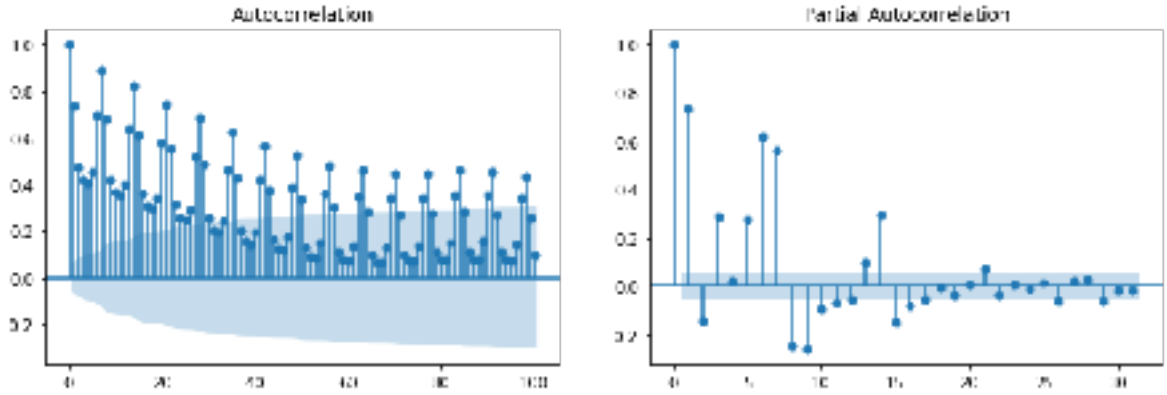
The time plot of the data shows all the characteristics of a stationary series. On doing an Augmented Dicky Fuller test the series is found to be a stationary series

Passenger Prediction

Model Building :

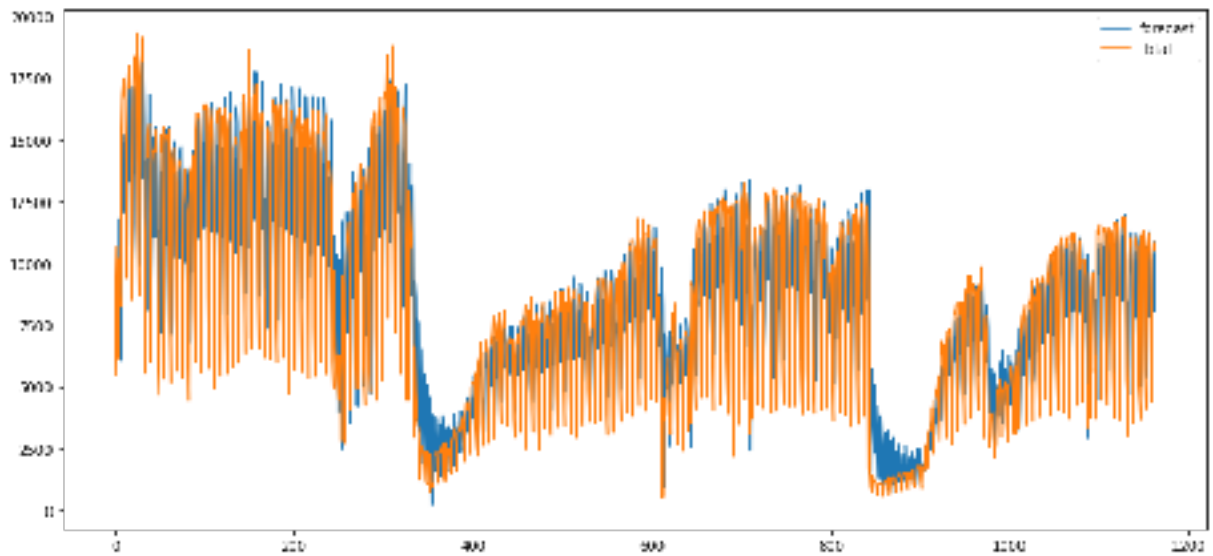
ARMA Models:



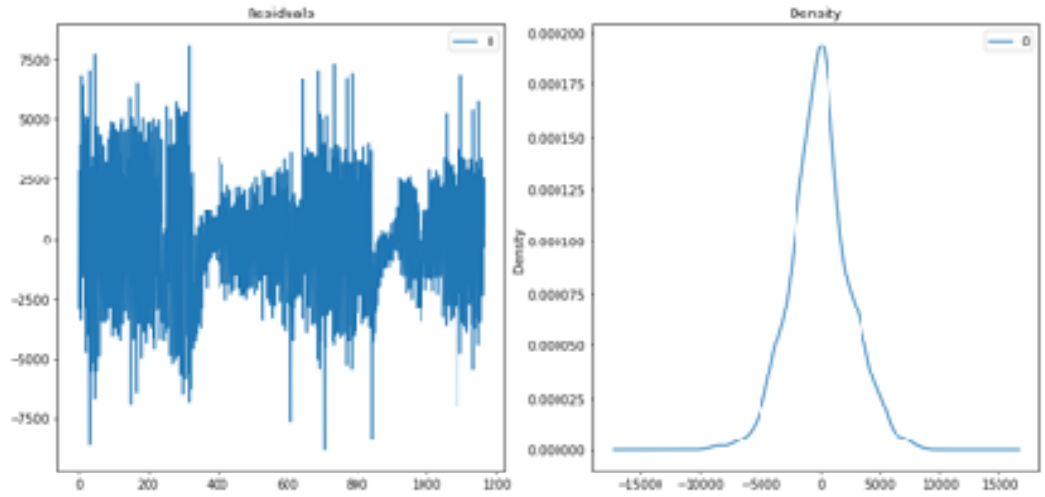


As the series is stationary its a good candidate for Auto Regression and Moving Average models.
Since this series is already stationary there is no differencing required.
This can be seen in the ACF and PACF plots. The current data is dependent on the lagged data.

Implemented a auto.arima which recommended (4,0,4) order ARMA model
Built first ARMA model with (p=4,d=0,q=4) (d=0 as no differencing required)

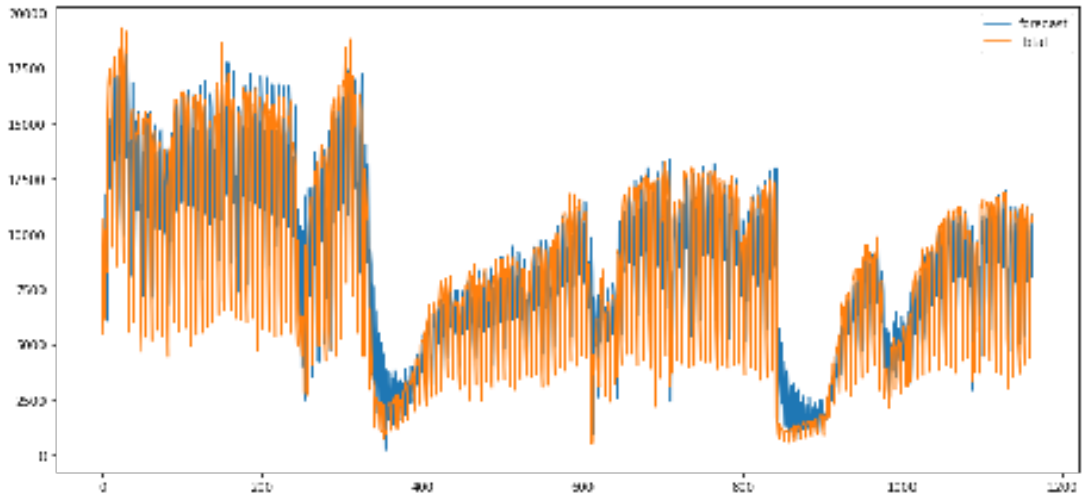


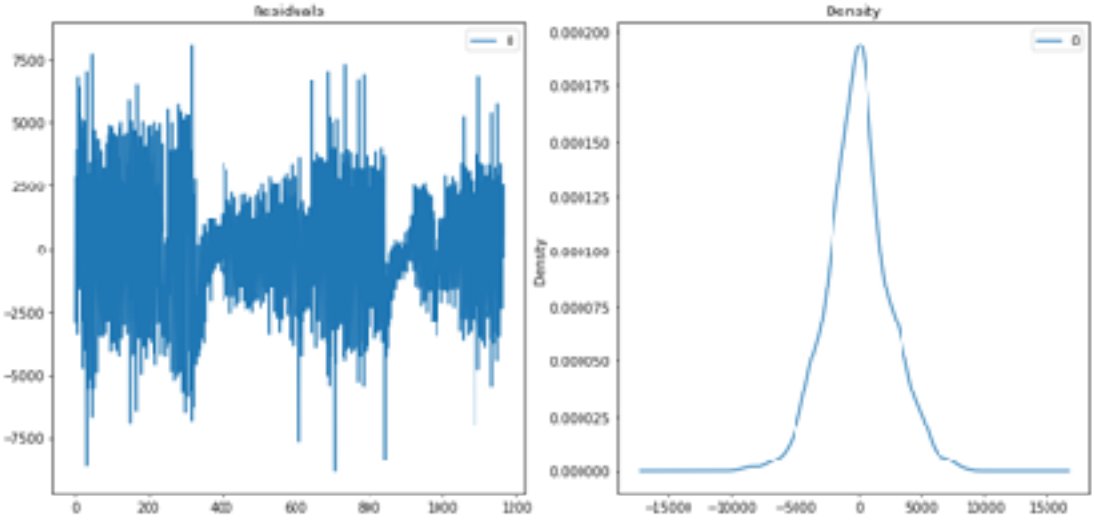
Passenger Prediction



On plotting the residual plot can see some movements which implements the model didn't pick the movements in the model.

Tried different ARMA models with different values of p,d,q but the residual plots were not satisfactory





As the series had more volatility tried GARCH but the model was not statistically significant

Passenger Prediction

FBProphet:

FBProphet is an open source project built by FaceBook for Time series analysis and forecasting. It provides open source python libraries for time series predictions.
FBProphet has many adaptable parameters for model building. This helps in controlling the trend and seasonal lines picked by the model. As well FBProphet has many inbuilt functionalities for plotting and analysing model.

Covid Handling : In prophet the covid anomaly can be handled in many ways by tweaking the model parameters. Below are the ways it can be handled:

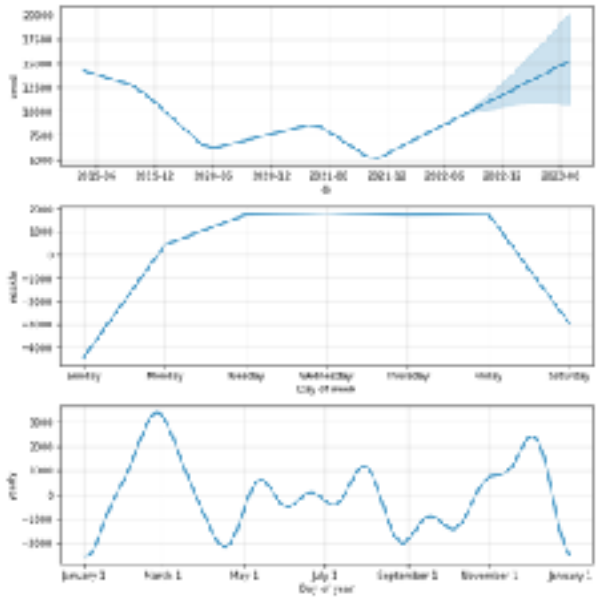
Holiday : Treating the covid dip as a one time holiday. In this way the model doesn't pick the covid trend through out the model. This can be achieved by supplying a holiday data frame to the holiday parameter during the model building. The covid lockdowns are supplied as holidays to the model and the future date will not have the holiday data to it during prediction, so the model treats the covid lockdown as one time holiday.

Change Points: Can manually specify change points during model building. The model by default picks the change point in the data and adjusts the prediction accordingly.

Regularisation Regressor: Can supply separate data frame as a regularisation regressor. This option is given as a parameter during model building. Can use index available in internet as a covid regressor and added it to the model.

Have applied the Holiday method and Change point method in this project.

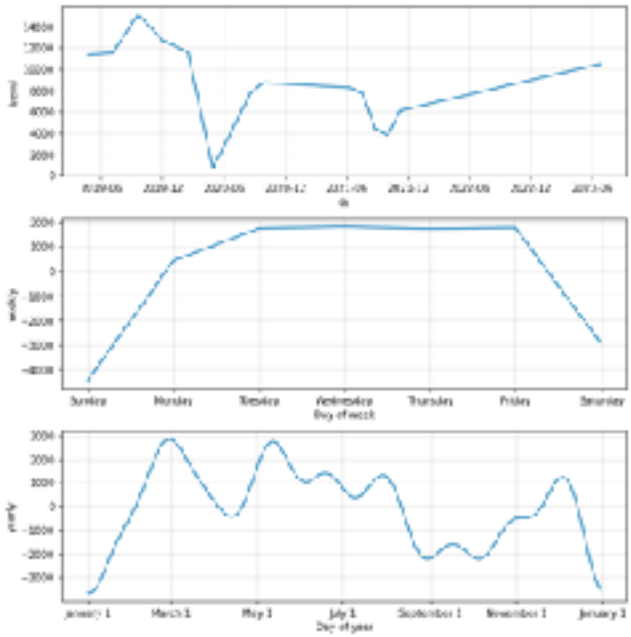
Model 1 - Initially a base line model with no parameters was built. The mode is also supplied with the future data set created 365 days ahead as input for prediction. The model is analysed with the inbuilt functions for cross validation and the metrics. MAPE is taken into account and is plotted. The method has many parameters which is also displayed in the metrics object



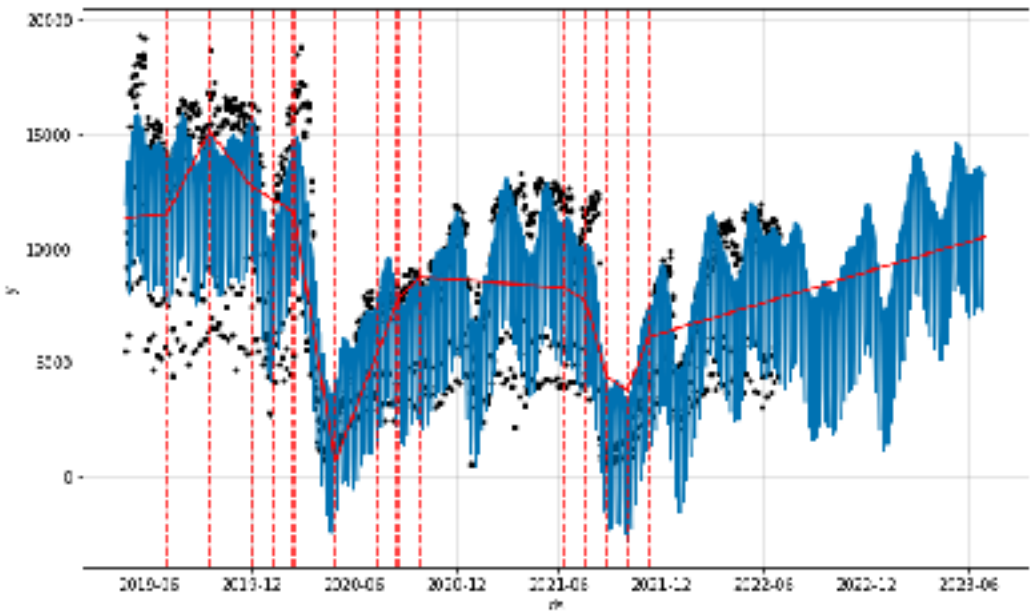
The components of the model is also plotted where the components are : weekly, yearly and trend movements. There is a notable change in pattern across week and the trend is also in upward movement. This can represent a normalisation after covid dip. There is a volatile yearly movement seen.

Passenger Prediction

Model 2 - Model with weekly seasonality given true and a change point value is given to control the trend after the covid dip as the trend can be due to shock recovery. As above the model is plotted and predicted with future 365 days.



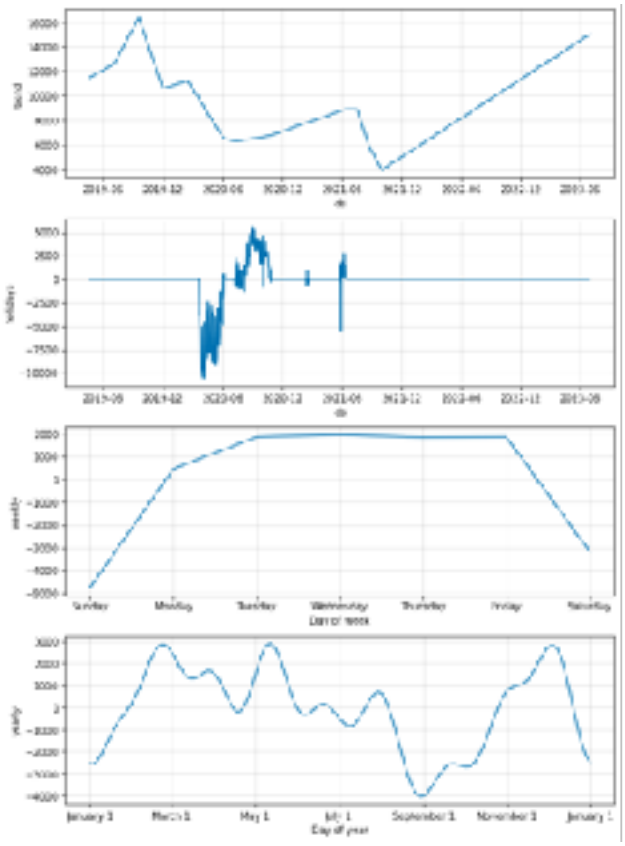
In this model the trend can be seen regularised because of the change point given and it follows the covid dip closely and the future trend is also regularise accordingly.

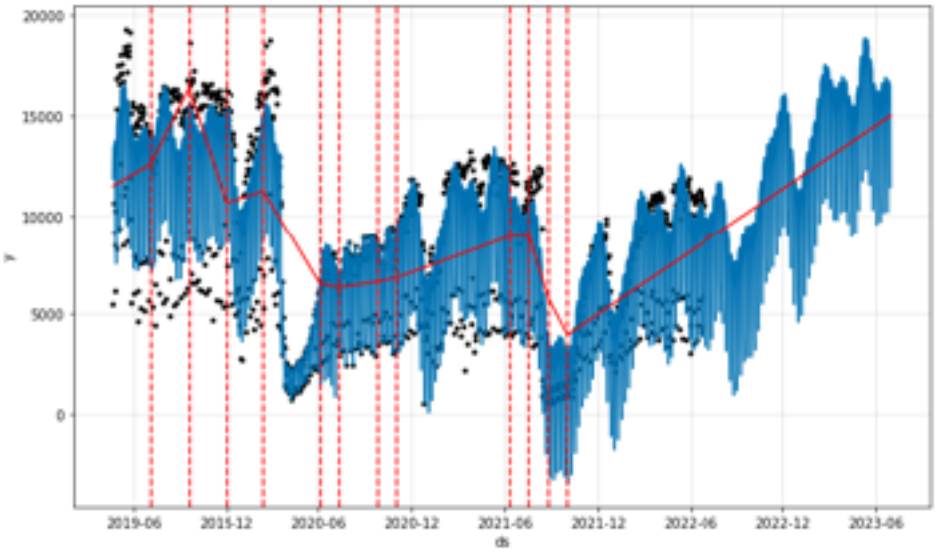


Another plot is plotted to show the change point picked by the model in red line . In this plot the can see that the model has picked the change point to a satisfactory manner according to the data.

Passenger Prediction

Model 3 - In this model the model is build by supplying holiday as a parameter. A holiday data frame is created for the covid lockdown dates and this is supplied as one time holiday to the model. In this way the model doesn't capture the covid shock for future predictions.





When plotting the model component there is another plot component specified for the holiday. In this we can clearly see the model dint pick the covid shock and treated it as single holiday. The trend line also has a considerable regularisation.

Model 4 - In this model change point parameter is tweaked higher for trend control across predictions.

Passenger Prediction