# Dataset

The dataset used is the **RAPID dataset. RAPID is an abbreviation for Rapid Preliminary In-patient Data**. This dataset has the basic diagnostic information of a patient admitted in a hospital ward. The features in the dataset are as follows:

* Timestamp (GMT): The date and time at which the reading was taken.
* Lifetouch Heart Rate: The real-time heart rate of the patient.
* Lifetouch Respiration Rate: The real-time respiration rate of the patient.
* Oximeter SpO2: The real-time Oxygen levels of the patient.
* Oximeter Pulse: The real-time pulse of the patient.

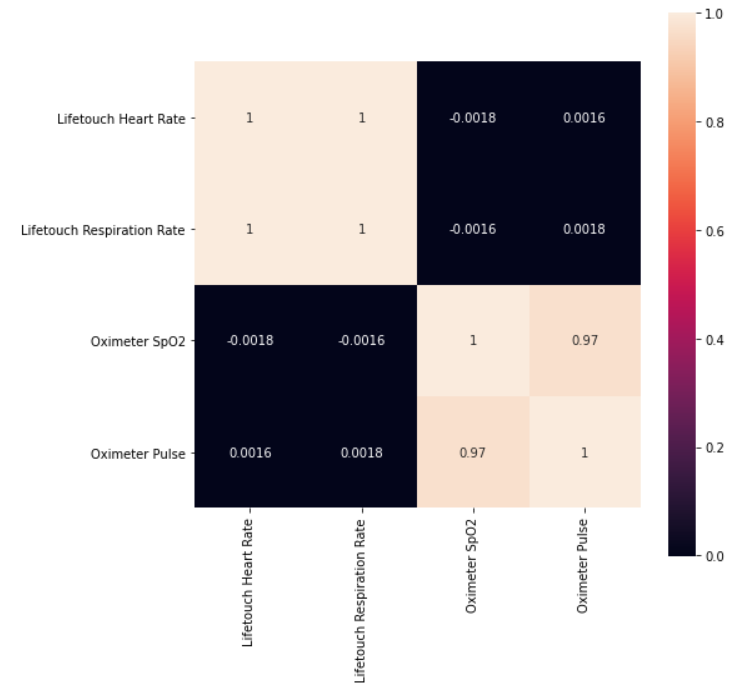
The source of the dataset was the student’s supervisor. The supervisor provided the dataset. Although there is limited information available on the dataset regarding the dataset. Since the problem requires regression, the data has not been split into training and test set, the complete dataset is used for regression since there are already very less instances in the dataset.

# Machine Learning

The algorithm used is the SARIMAX algorithm. The objective of this code is to predict separately using the **Regression** paradigm each of the features present in the dataset. To serve this purpose, the SARIMAX algorithm from statsmodels library for Python has been applied.

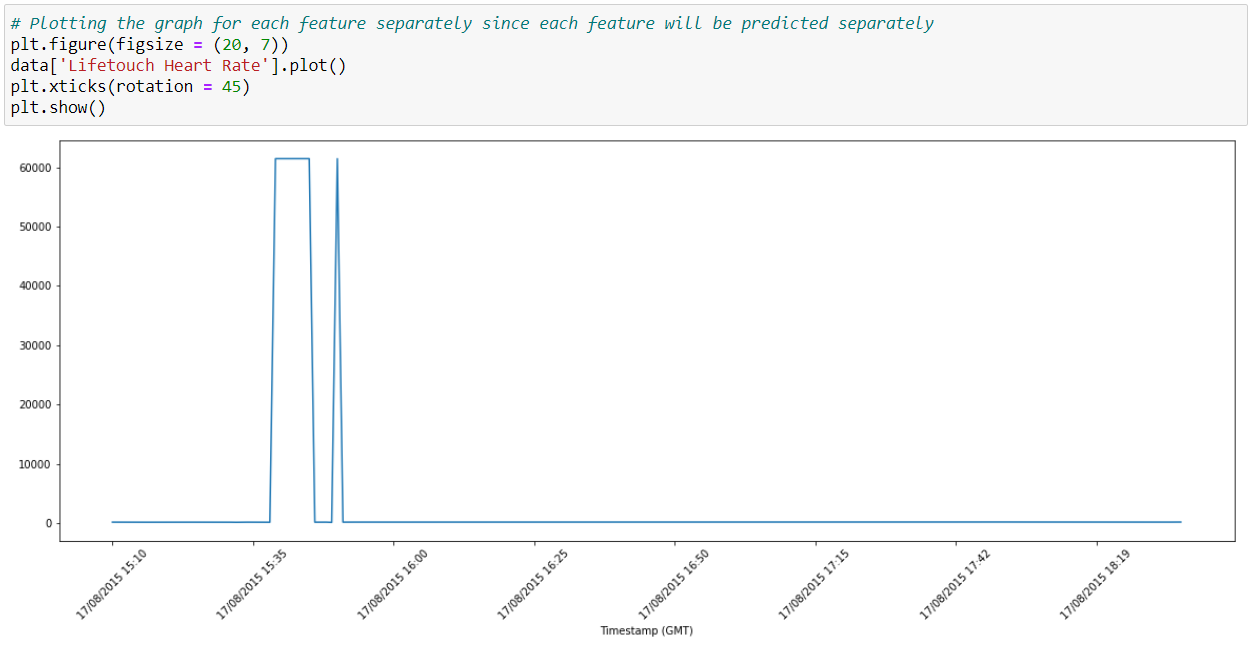
# Data Preprocessing

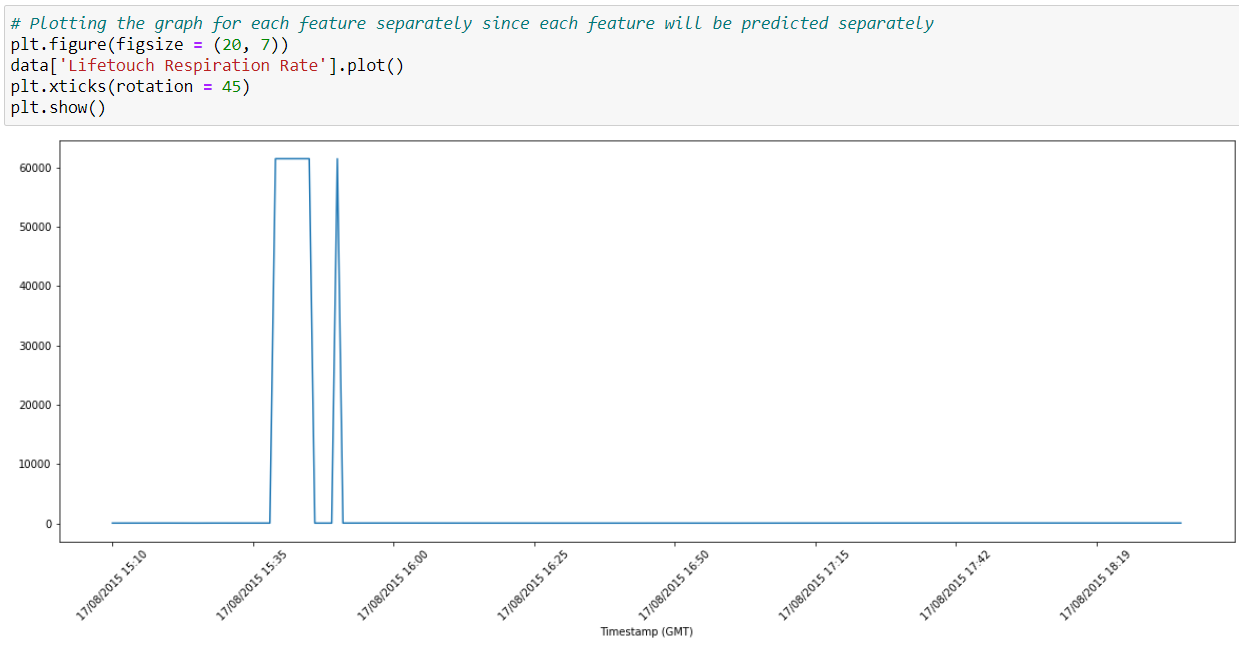
The code contains the following graphs that harbor the following purposes:

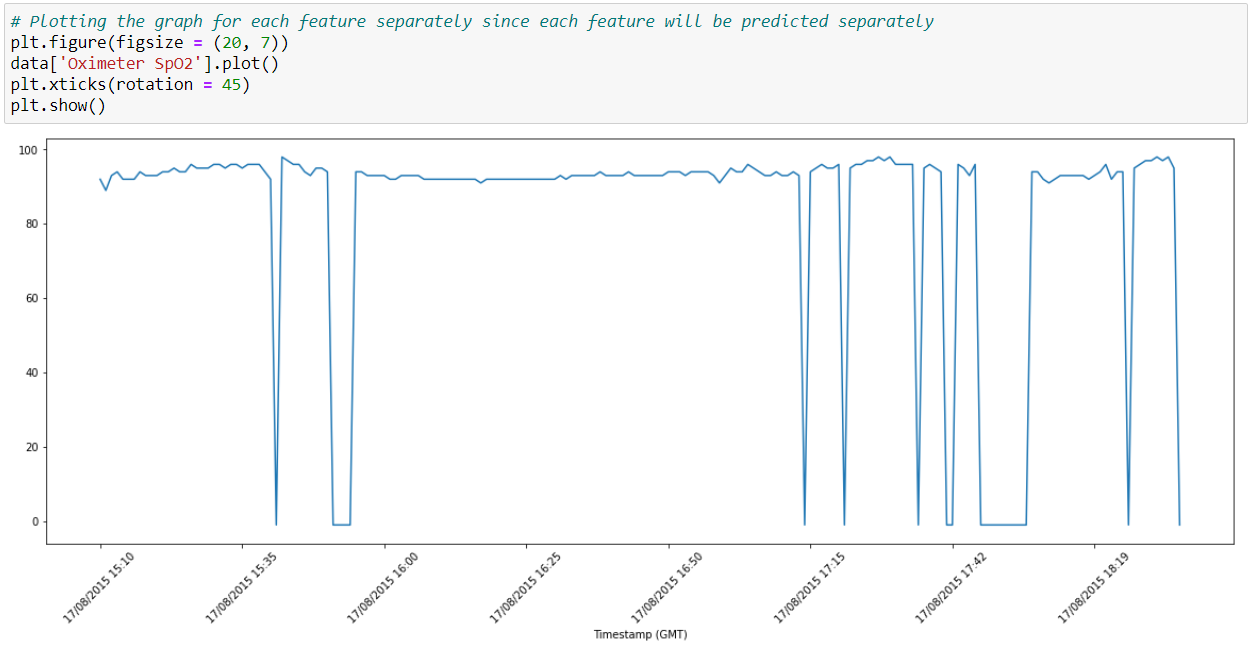


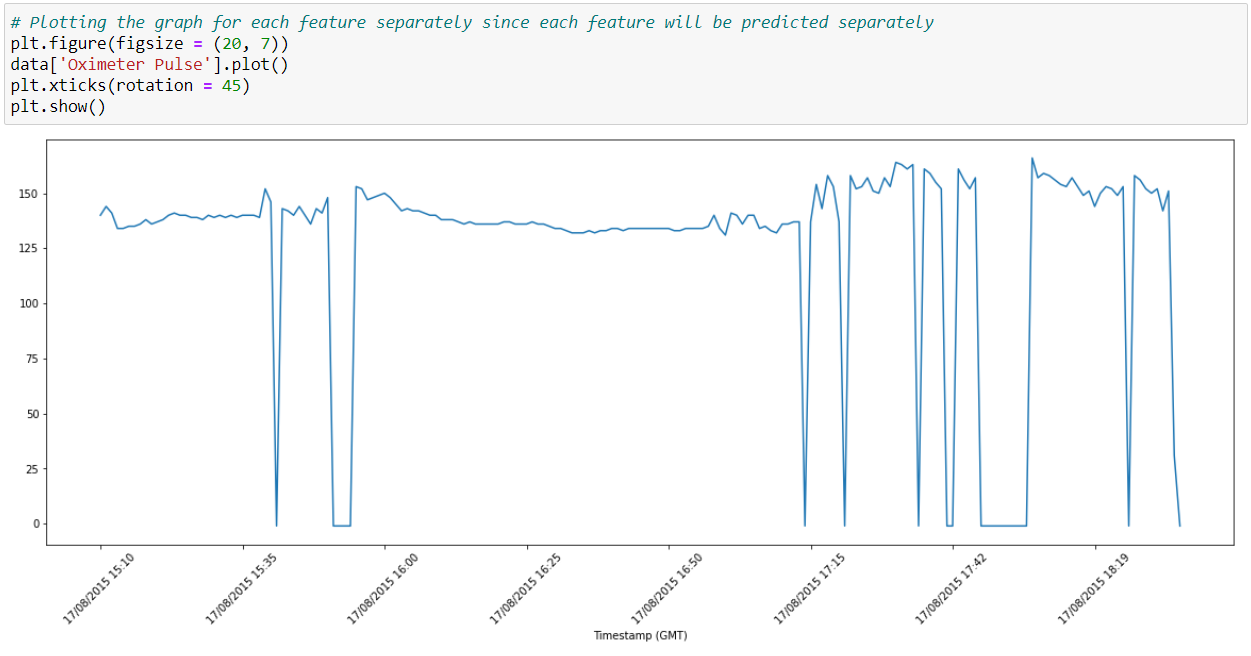
This is a correlation matrix that explains the correlation between the features in the dataset by calculating the correlation value for each of the features. The higher the correlation value, the better the correlation between the features in the dataset and higher the dataset quality.

The next four graphs are simply the data points plotted from the dataset to see the shape of the feature set.







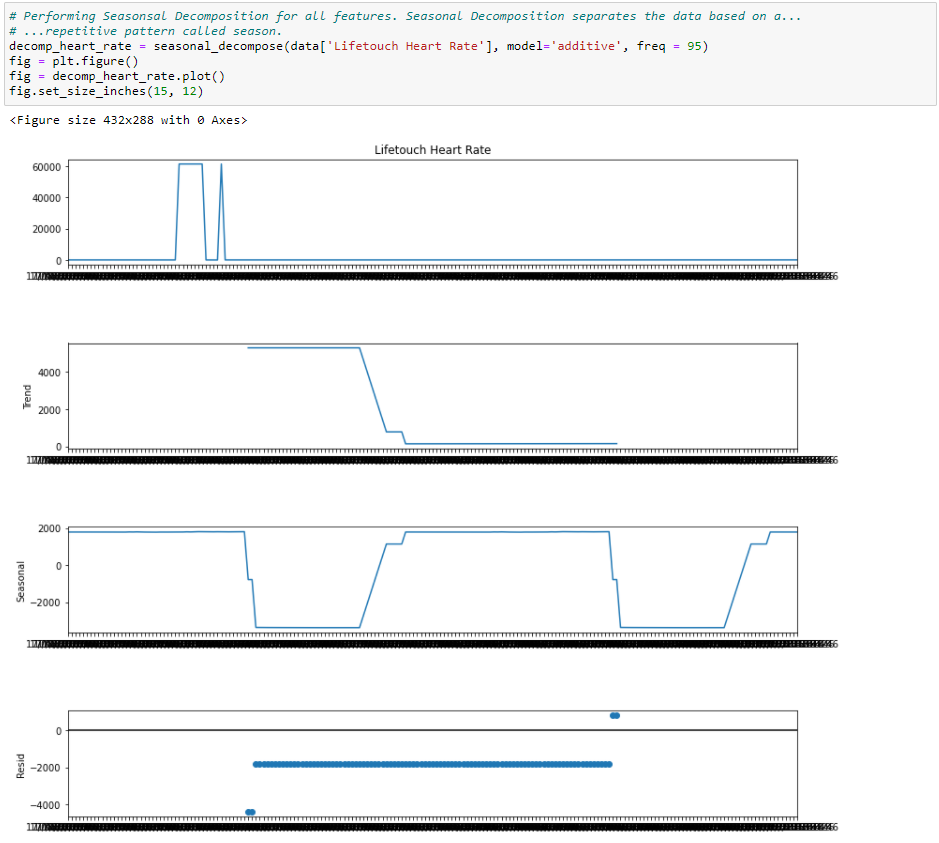


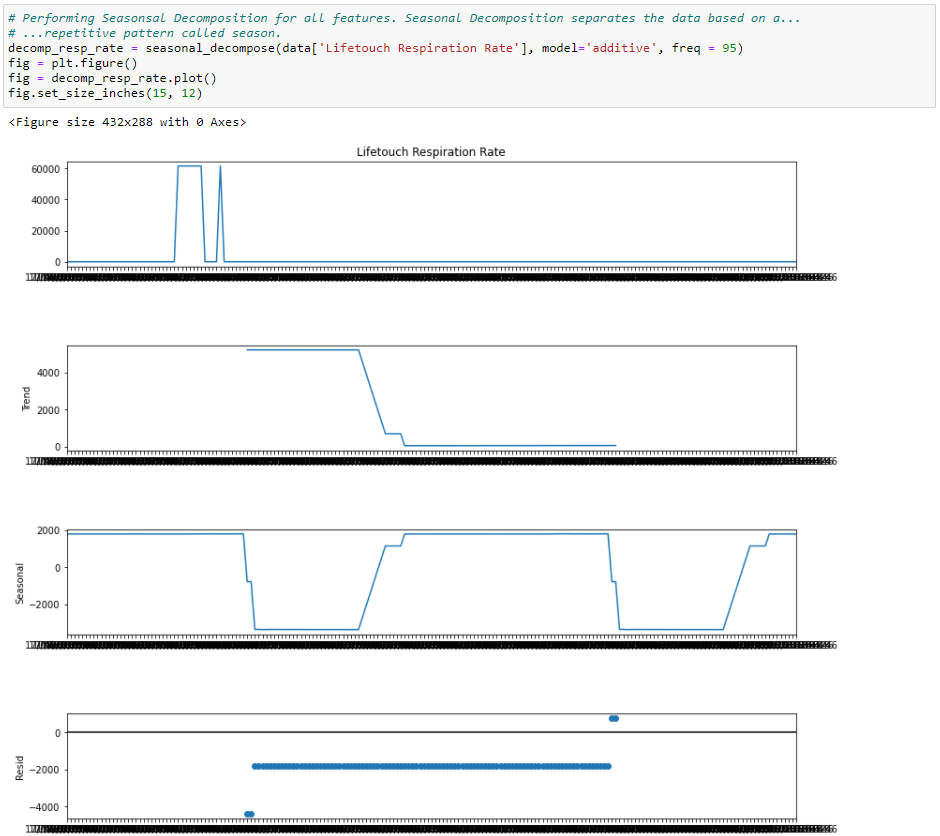
As you can see, the graphs for each feature look like the output on the screens of the machines that the hospitals use to measure the vitals of an admitted patient.

For the implementation of SARIMAX, there are a few conditions need to be met to perform Regression with it. These conditions are as follows:

* The data must be stationary
* The parameters of the SARIMAX algorithm including the seasonal ones must be determined

The first step in this regard is to handle the seasonality of the feature sets. The graphs below show the process.

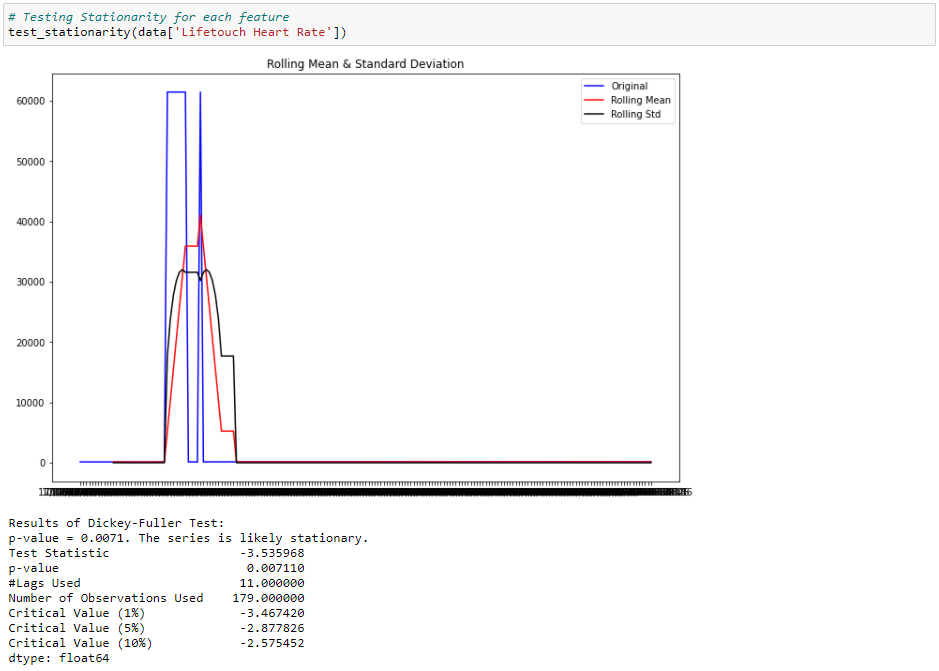


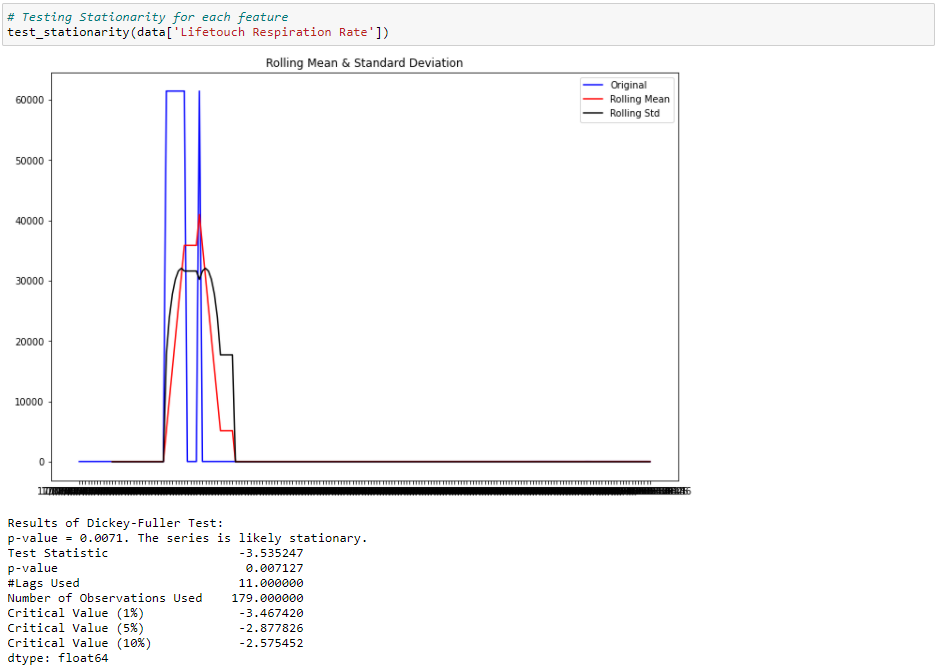


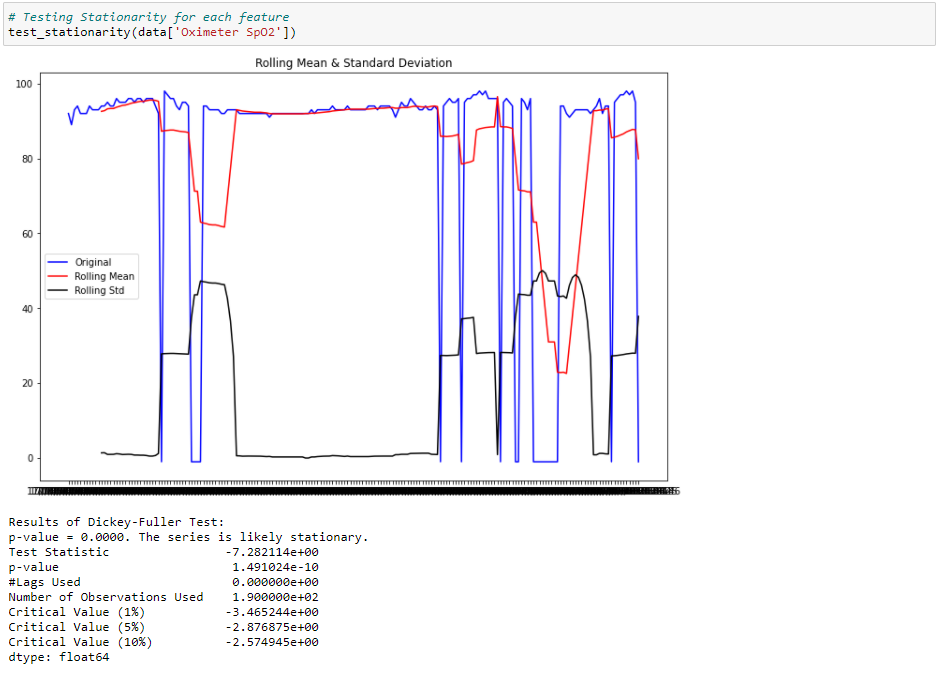


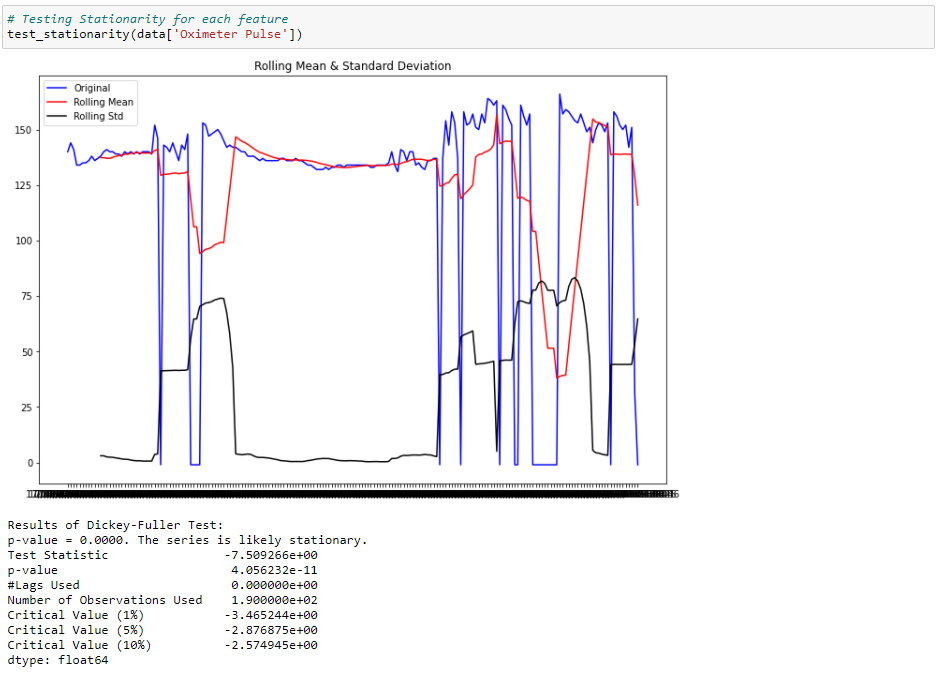


The following graphs are for testing the stationarity of each of the features in the dataset after applying the requisite data preprocessing on the features. These tests are performed using the Dickey-Fuller test.

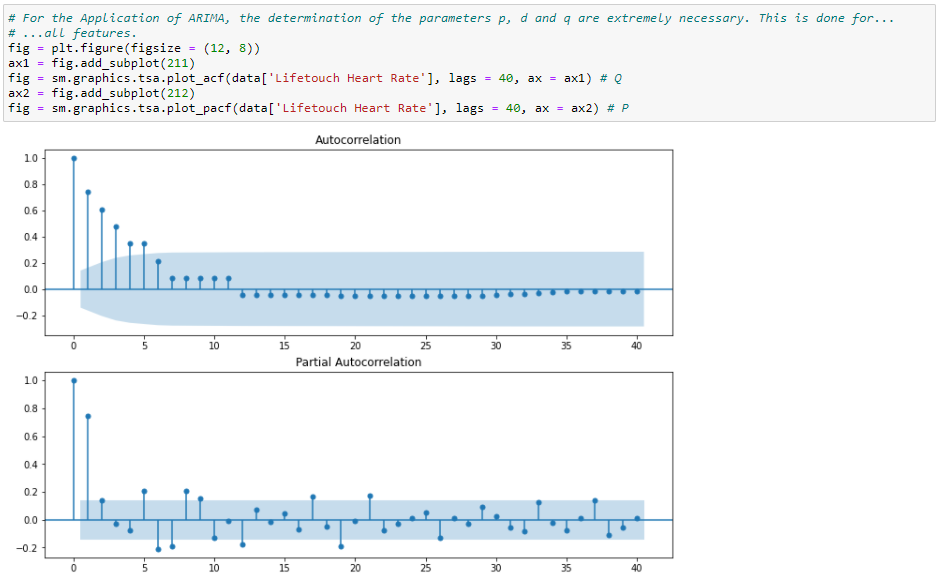


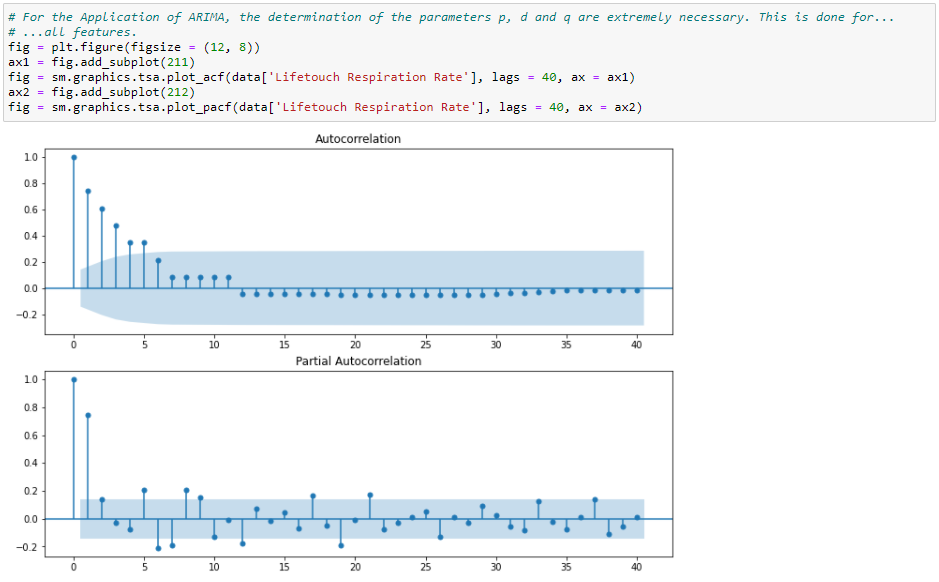


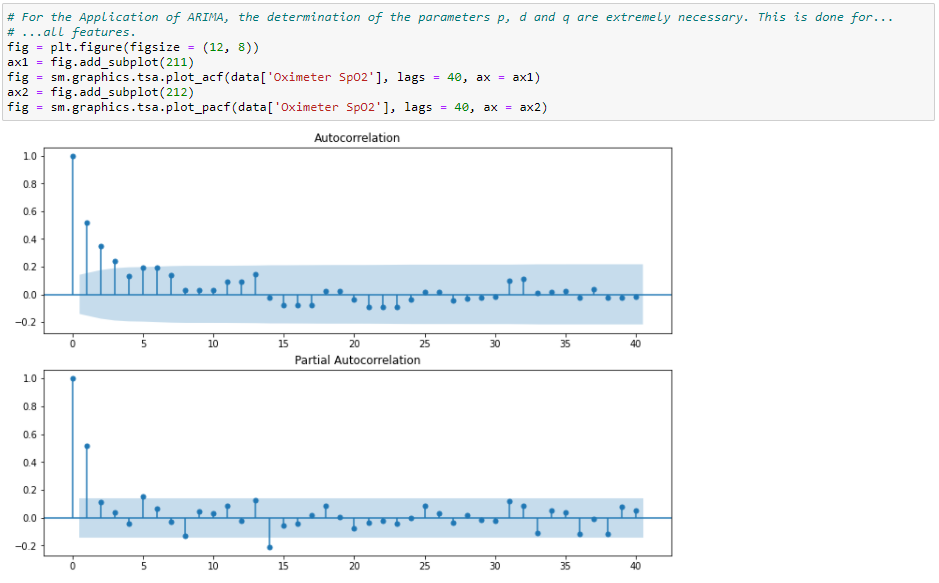


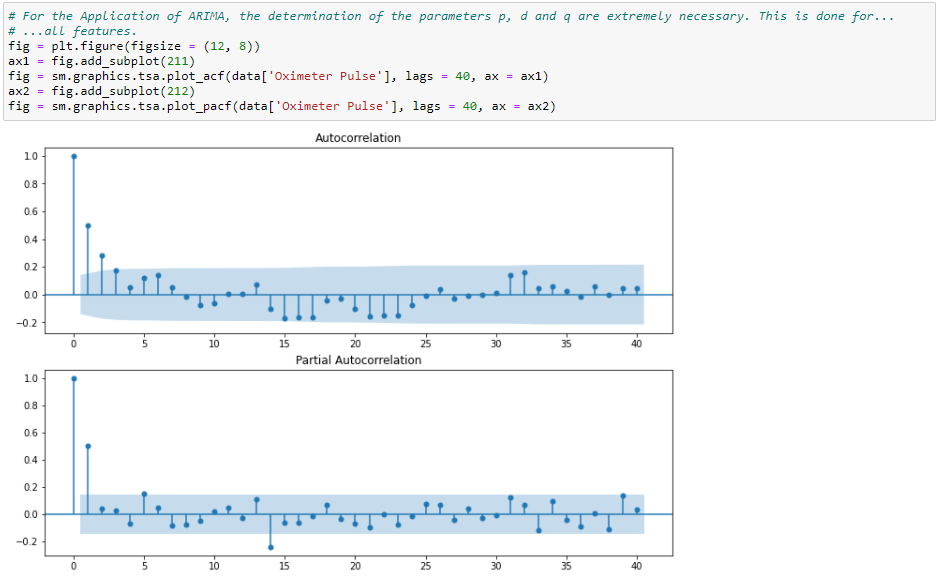


The next step is the determination of the hyperparameters of the SARIMAX models. I would like to make clear once again that the objective of this code is to predict using Regression the future values of each of the features in the dataset separately. Therefore, any procedure done is performed four times repeatedly to accommodate the four features in the dataset.

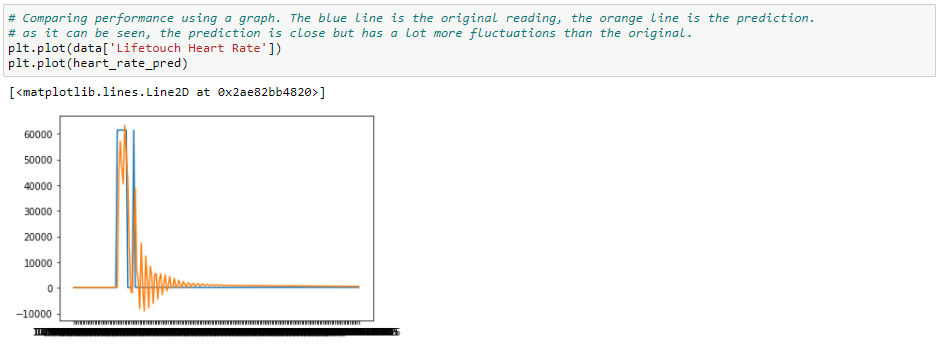


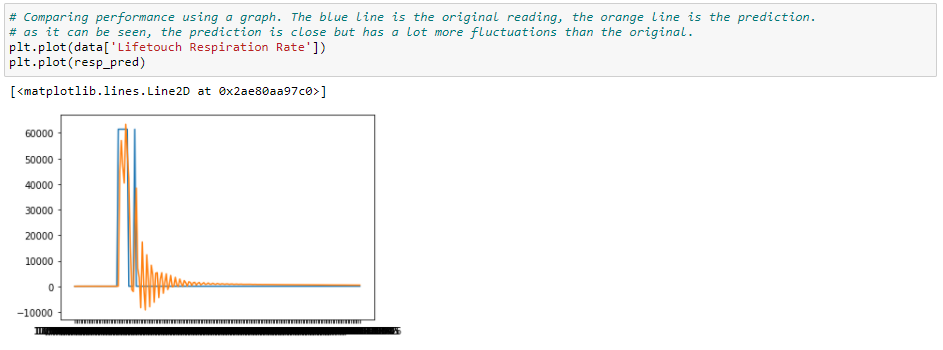


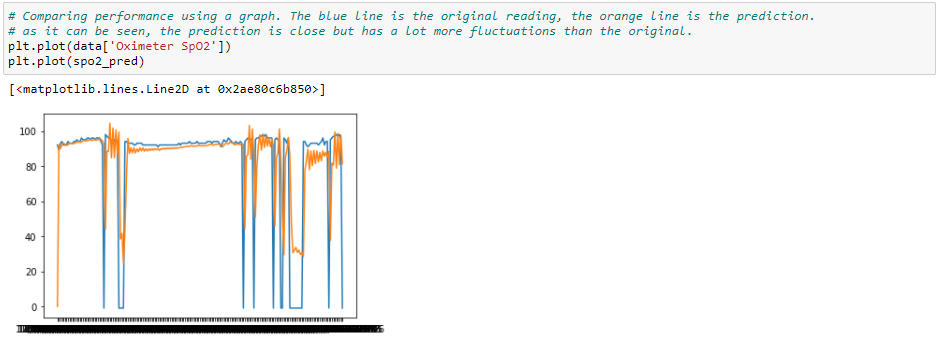


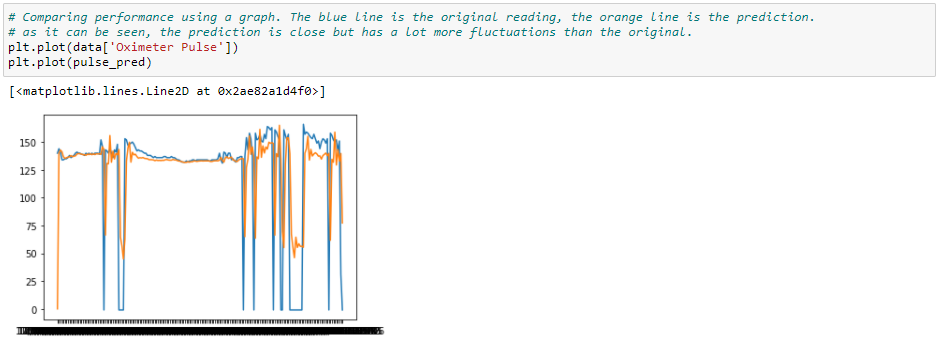


The data preprocessing is complete and the algorithm has been implemented. The following graphs are for the results generated for each feature in the dataset.









In the above graphs, the orange line represents the prediction from the algorithm and the blue line represents the feature set itself.