# TIME SERIES ANOMALY DETECTION FOR BIG DATA

**BY**

**SACHIN JANGONI**

**SATISH GOLI**

**ADEO2**

## CONTENT

## 1)INTRODUCTION:

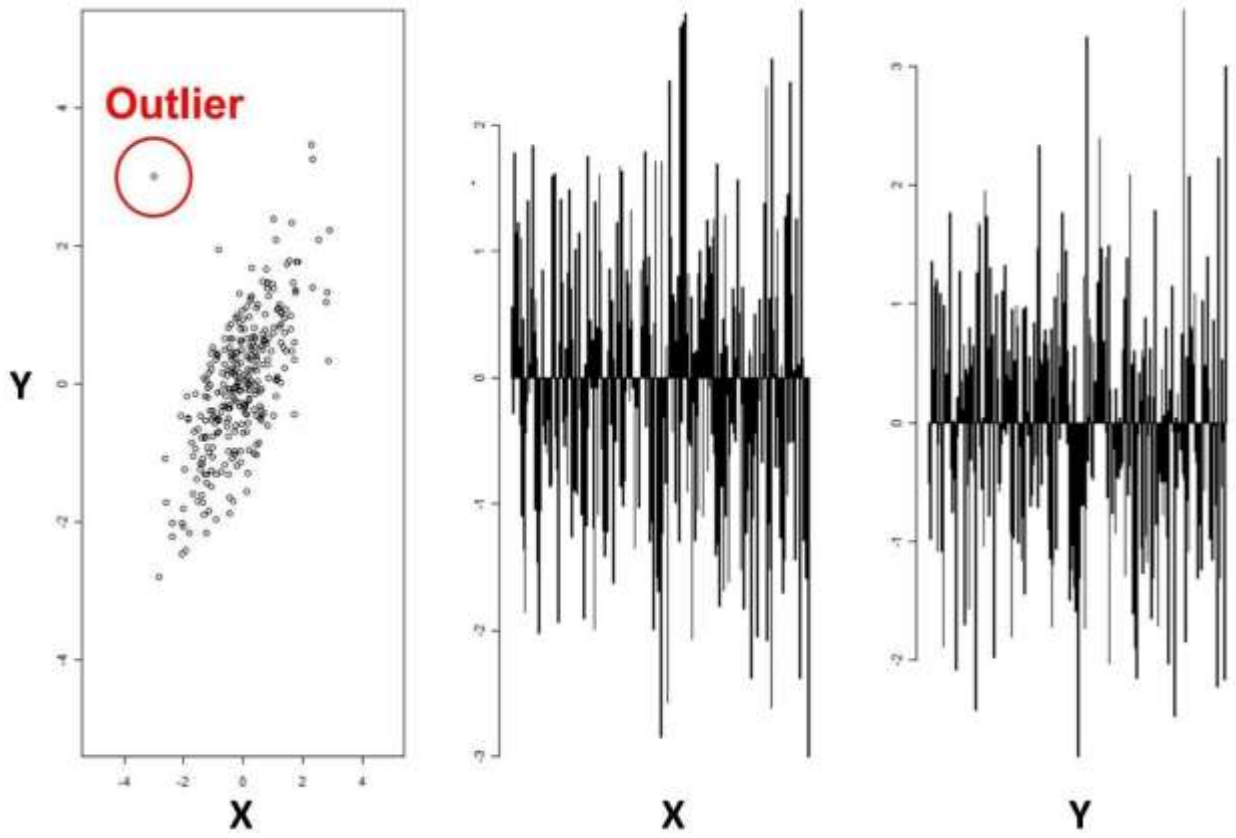
In today's world, anomaly detection can viably help in getting the misrepresentation, finding strange activity in huge and complex Big Data sets. This can end up being valuable in areas, for example, banking security, natural sciences, medication, and marketing, which are inclined to malignant activities. It is very essential to identify a normal trend with an abnormal trend in order to identify/fix issues quickly. The two machine learning algorithms that can enable two effective for anomaly detection are supervised and unsupervised algorithms. Data Pre-processing is one of the basic steps in the data mining process, which does the readiness and change of the dataset, to fill the missing values.

We are going to discuss the machine learning algorithms for anomaly detection namely, local outlier factor, isolation forest, k-means, Elliptic Envelope, one-class SVM and LSTM Autoencoders. The algorithms, which we were used for anomaly detection, are unsupervised machine learning algorithms. Because, for a large amount of unstructured data and also data is unlabelled. We are going to discuss about the anomaly detection, And machine leaning algorithms for anomaly detection.

## 2) DESCRIPTION:

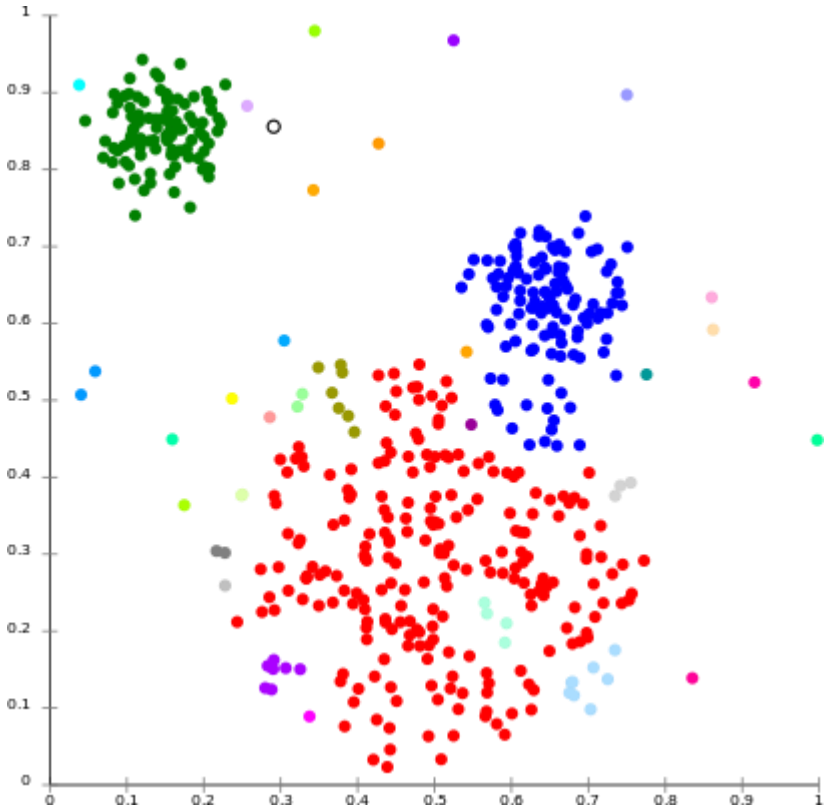
## 1) ANOMALY DETECTION:

Anomaly detection is a procedure used to distinguish abnormal patterns that do not comply with anticipated conduct. It tends to be viewed as the unusual procedure of figuring out what is normal and what is not. Anomalies are additionally known as anomalies, oddities, noise, and deviations

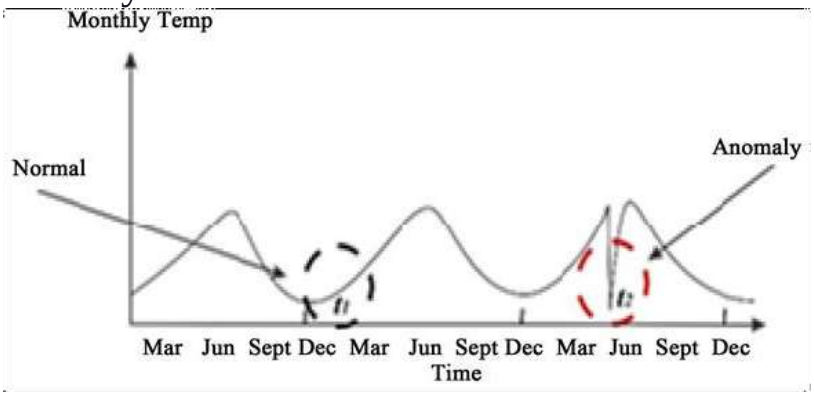



There are three types of anomalies they are:

I) POINT ANOMALIES: If one object is far away from the other objects as an anomaly, it is a point anomaly.

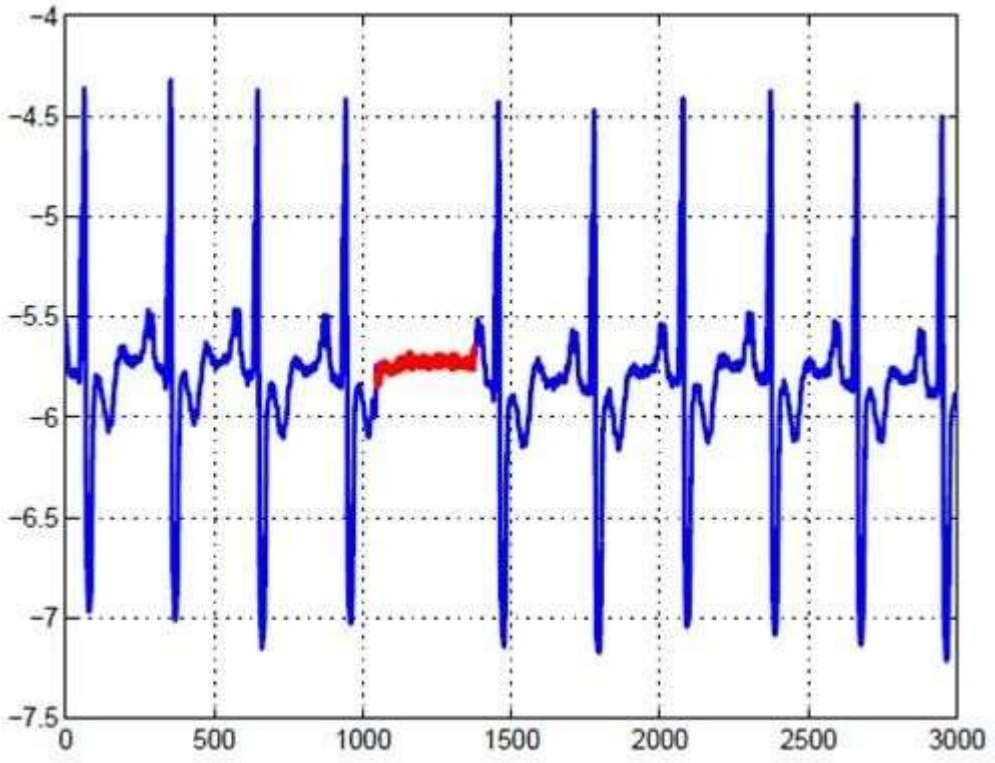


II) CONTEXTUAL ANOMALIES: If an object is deviating from what is normal in some defined context. Only in this case, it is a contextual anomaly.




III) COLLECTIVE ANOMALIES: If a collection of related data instances is anomalous with respect to the entire data set, it is termed as collective

anomalies.




## 3)DATASET:

the airline delays and cancellation dataset consists of 10 files from 2009 to 2018 with 28columns of time series insights. Which is available in the kaggle. from the dataset, we choose the columns of, Arrival delay, CRS elapsed time, Elapsed time. these columns give the data of the flight's delays and differences between the computer reservation time and actual flight delay time. we will detect the anomalies by using these columns. The dataset is huge it have so many missing values and the main thing the data is unlabeled.

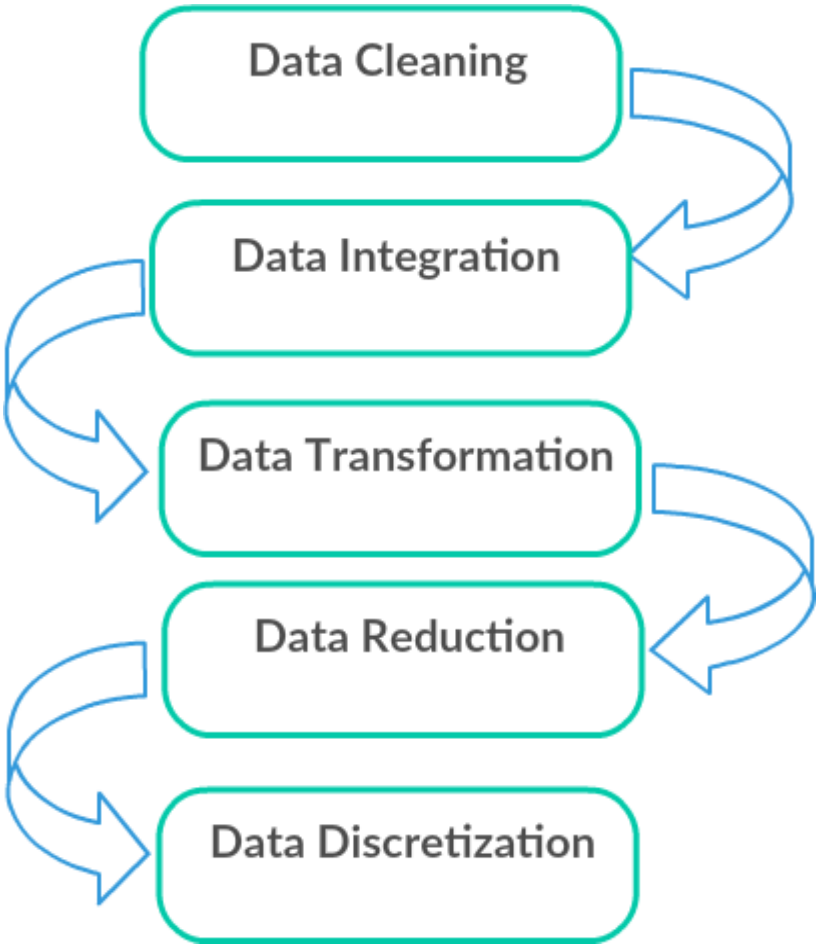
https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

## 4)METHODS:

### I)DATA PRE-PROCESSING:

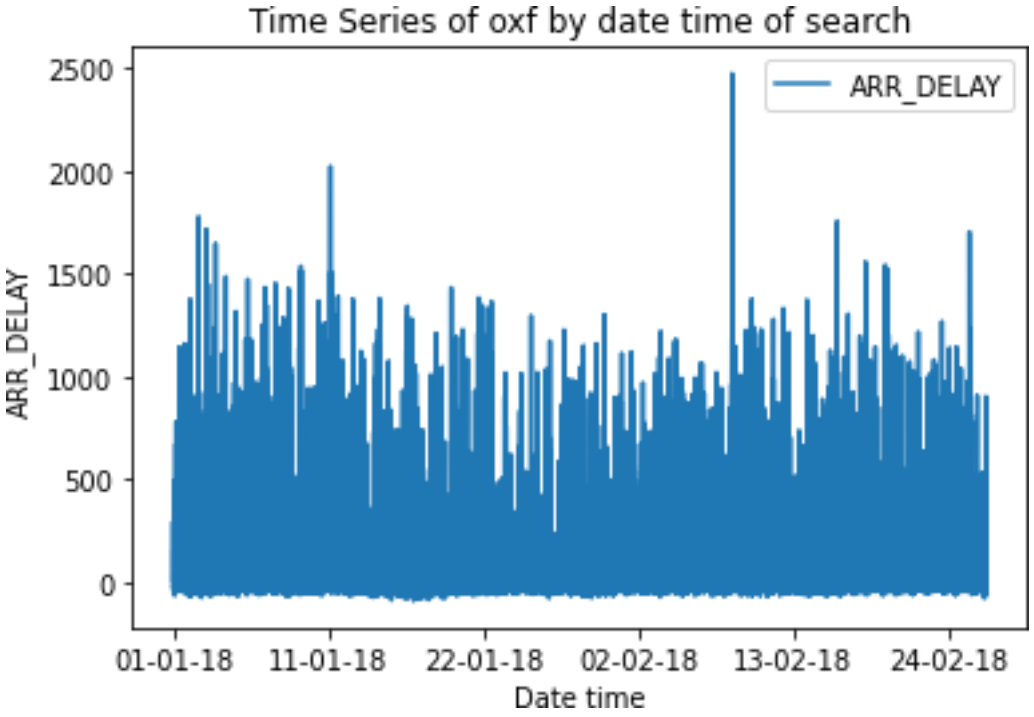
Data pre-processing is a data mining method that includes changing unstructured data into an understandable format. real-world data is regularly inadequate, conflicting, as well as ailing in specific behaviour's or trends, and is probably going to contain numerous anomalies. data pre-processing is a demonstrated technique for settling such issues. Firstly, we have to find the number of missing values in each column and then fill these missing values with the mean values.

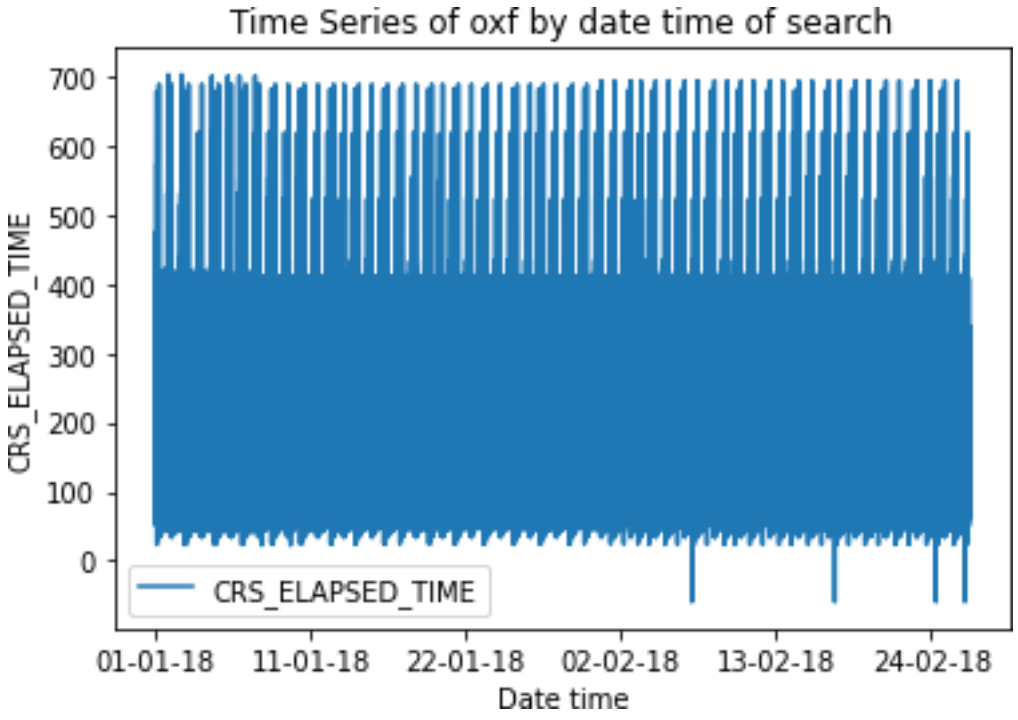
For the data we have some techniques to fill the missing values like interpolation, forward and backward values, mean and median values. For this data set there is less number of null values. So, I used the interpolation method values to fill the missing values. After filling the null values we have to check the previous data with null values and filled missing values with the interpolation method values.

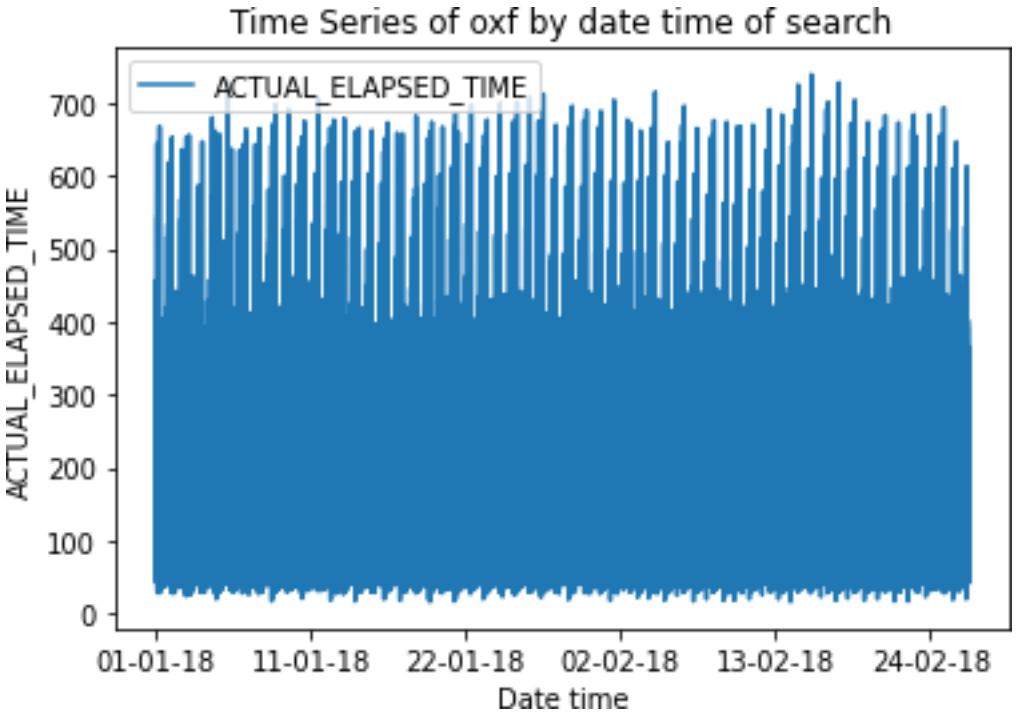


## II)EXPLORATORY DATA ANALYSIS(EDA):

It refers to the critical process of performing initial investigation on data so as to discover patterns, to spot anomalies , to test hypothesis and to check assumptions with the help of summary satistics and graphical representation.

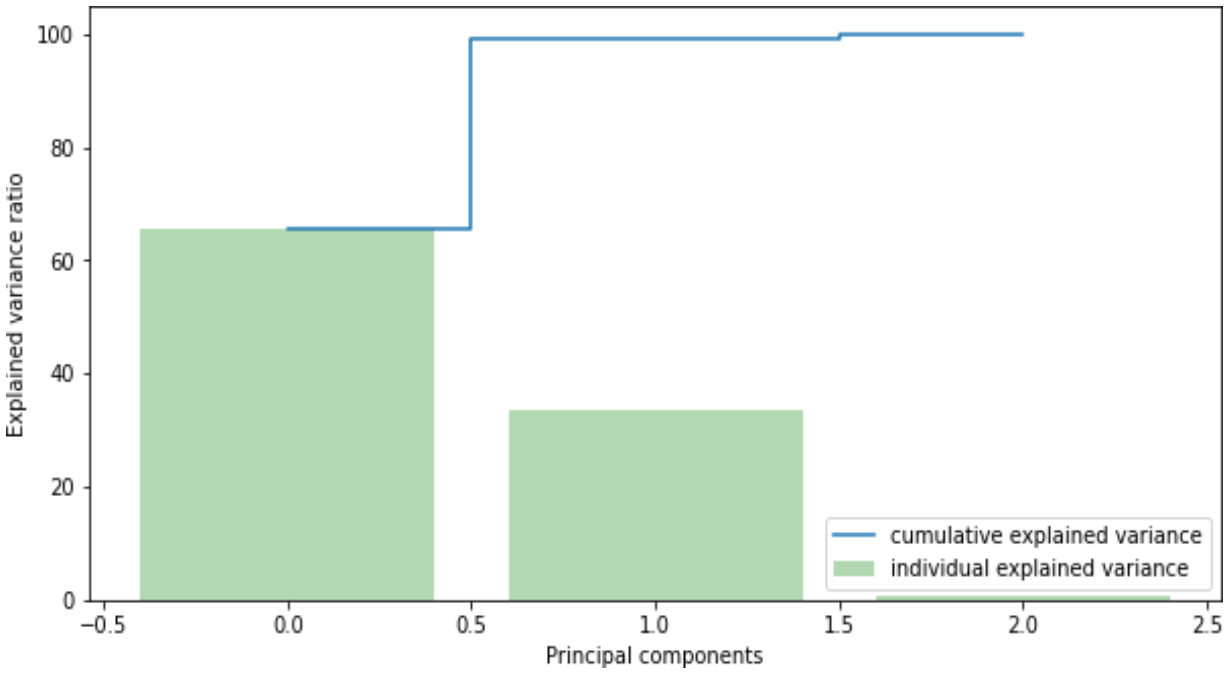





## III)PCA (principal component analysis):

Basically PCA is used for dimensionality reduction(Feature selection) and it will compresses maximum amount of data into two main features or principal components and also it will identify correlation between each metric.

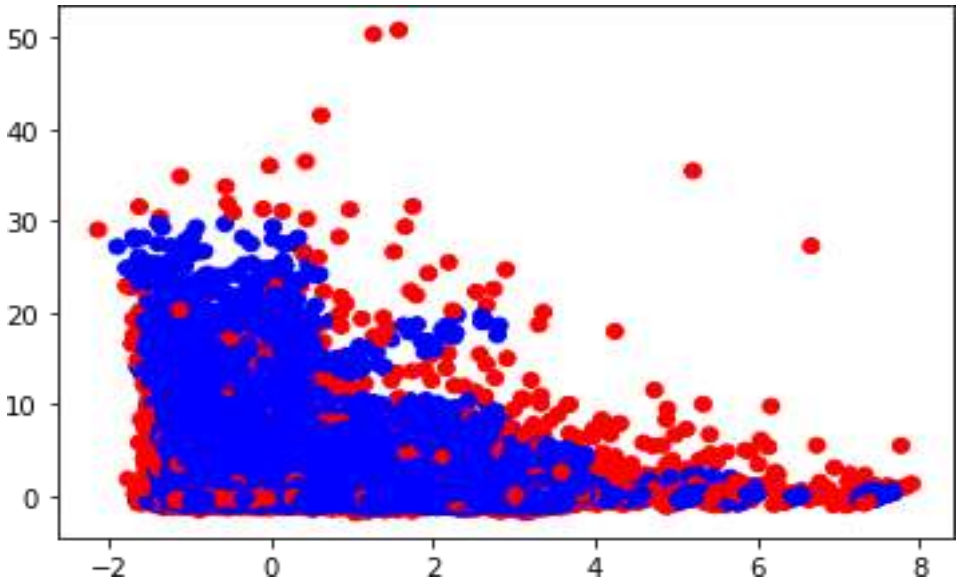


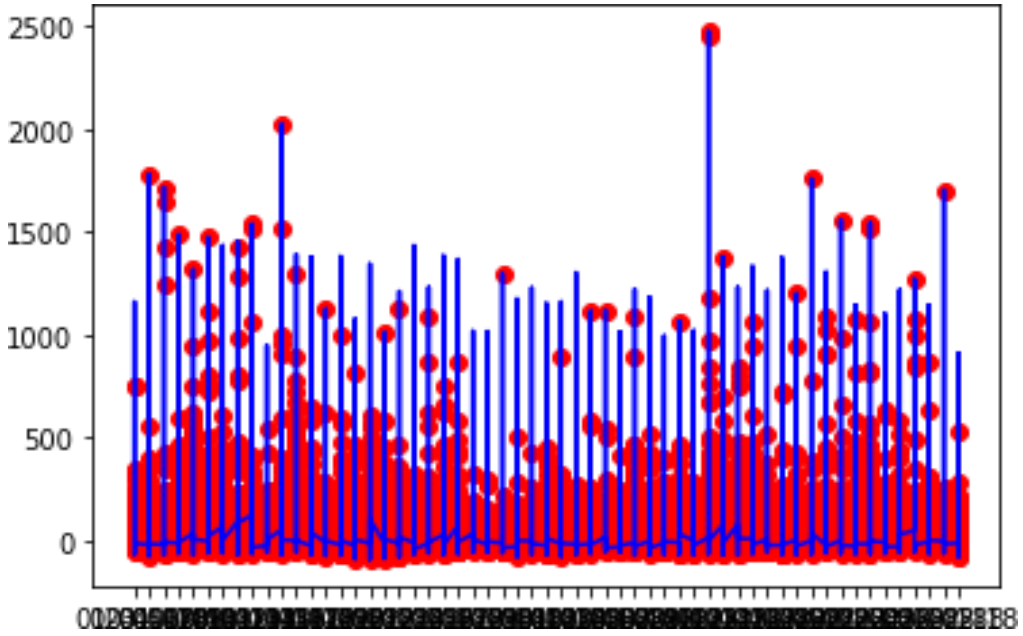
## MACHINE LEARNING ALGORITHMS:

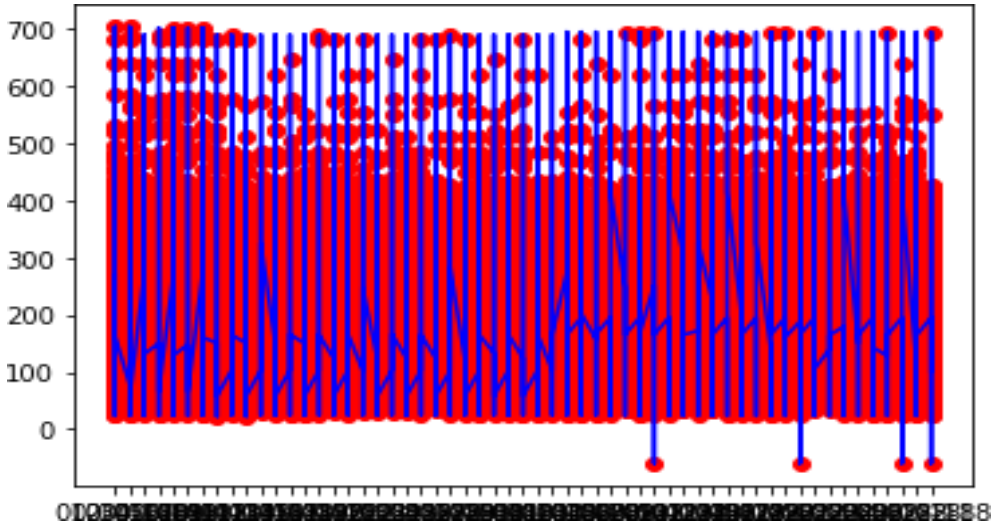
### IV)Local outline factor:

Local outlier factor into data mining of time series, wherein the lof is a neighbourhood level that relies upon how disengaged the object is as for the encompassing neighbourhoods. In addition, our last objective is to dole out an outlier factor. The level to which the object is distant. To every after effect in time series.

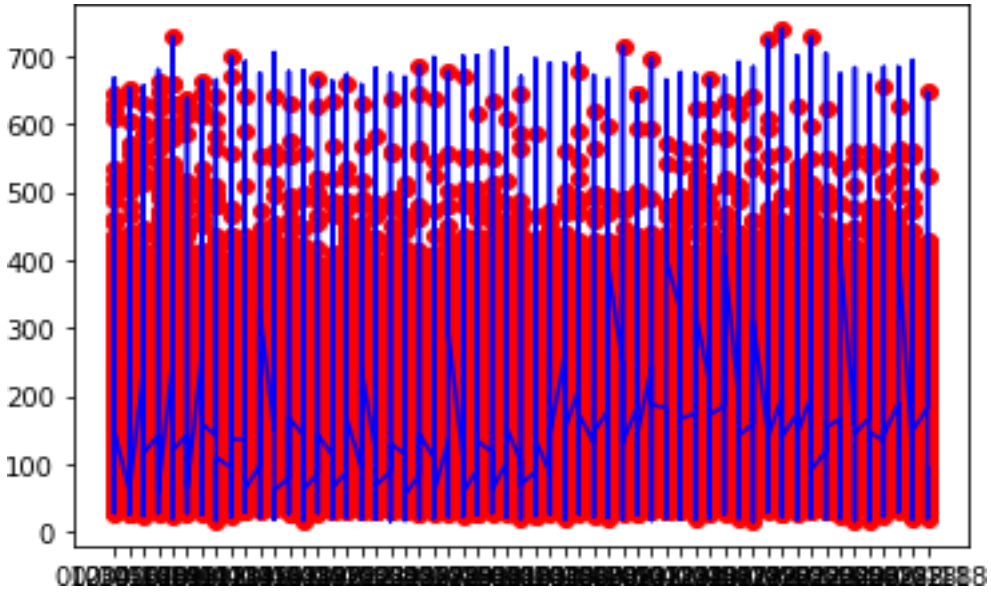
Local outlier factor calculates an anomaly score by using the local density of each sample point with respect to the points in its encompassing neighbourhood. The local density is inversely correlated with the average distance from a point to its






nearest neighbours. The anomaly score in LOF is known as the local outlier factor



score; its denominator is the local density of a sample point and its numerator is the average local density of the nearest neighbours of that sample point.

LOF assumes that anomalies are more isolated than normal data points such that anomalies have a lower local density, or equivalently, a higher local outlier factor score. LOF utilizes two hyper parameters: neighbourhood neighbourhood size and contamination. The contamination determines the proportion of the most isolated points to be predicted as anomalies.

Lof measures local variation of density of a sample vs its neighbours. Where k-nearest neighbours determine locality. Lower density samples are considered outliers. If is an ensemble extremely randomized tree-regressor that uses isolation to separate unusual data points. Compares local density of a point to the local density of its k neighbours. Both lof and if can be used in supervised and unsupervised settings

We have implemented SK learn library to use the local outlier factor .We given number of neighbours and contamination to get the number of anomalies.

## V)ISOLATION FOREST:

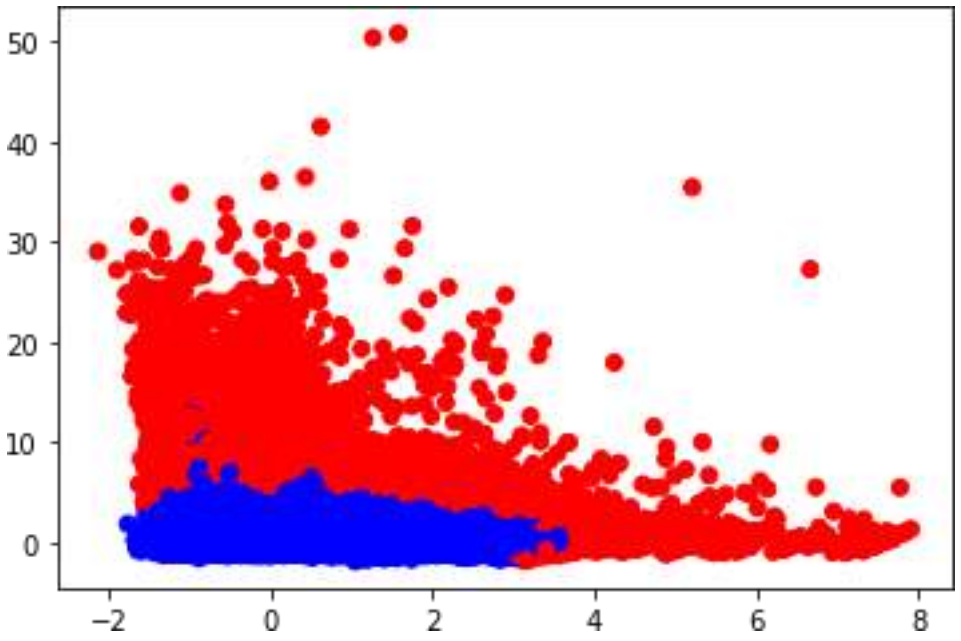
Isolation forest is an ensemble regressor, and it uses the concept of isolation to separate away anomalies.

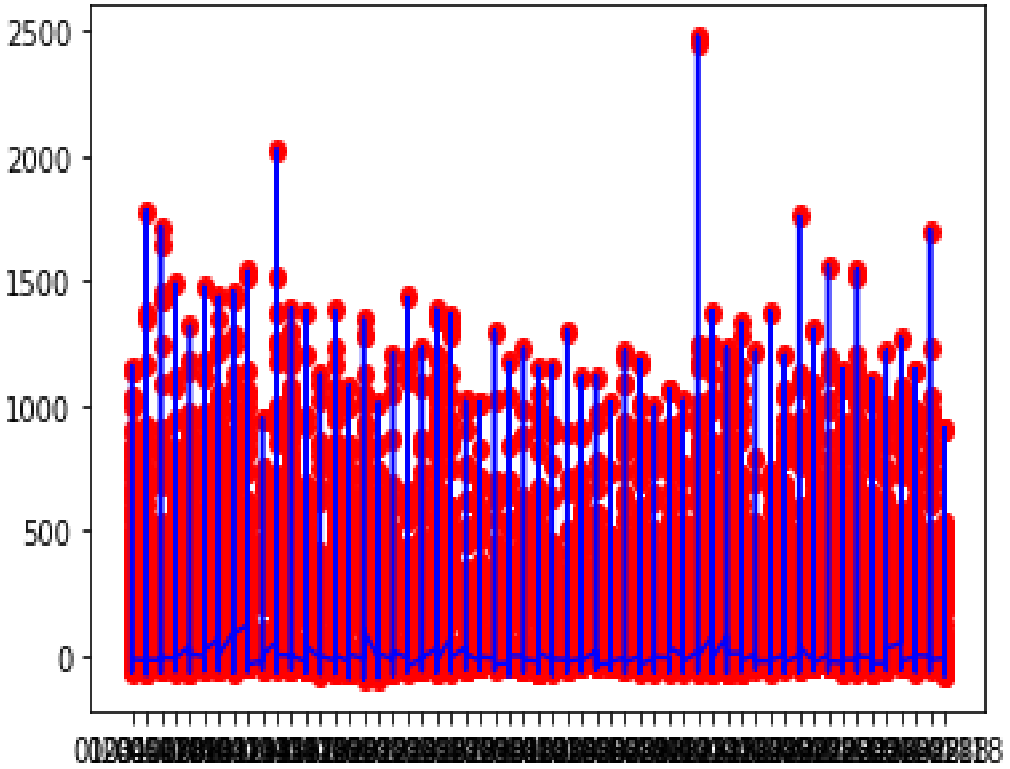
No profiling of typical occasions, and no point based separation computation. Rather, if assembles an ensemble of irregular trees for a given dataset, and inconsistencies are focuses with the briefest normal way length.

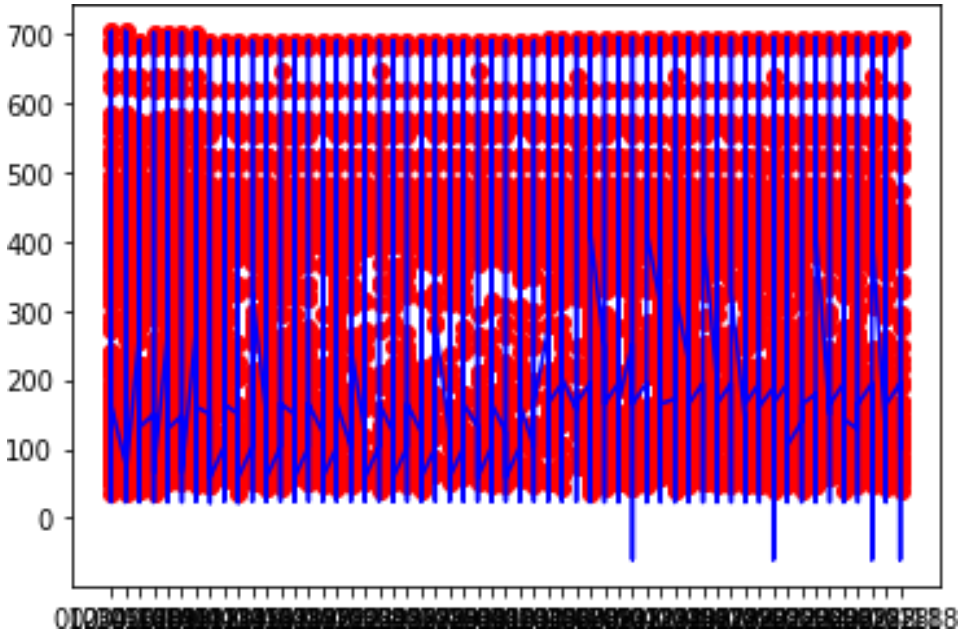
The Isolation Forest algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The logic argument goes: isolating anomaly observations is easier because only a few conditions are needed to separate those cases from the normal observations.

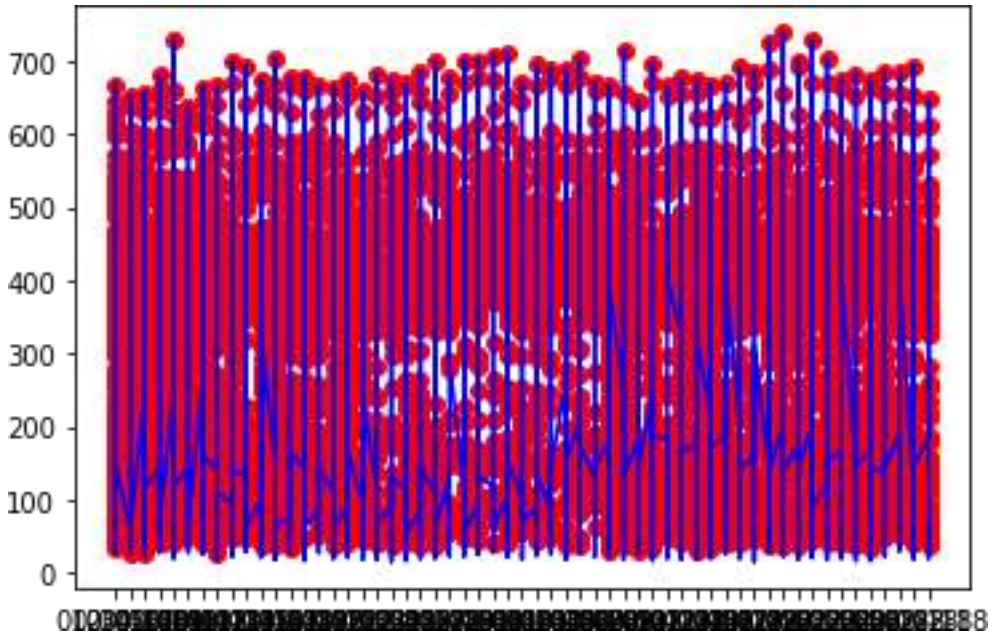
On the other hand, isolating normal observations require more conditions In this way, an anomaly score can be determined as the quantity of conditions required to isolate a given perception

The way that the algorithm constructs the separation is by first creating isolation trees, or random decision trees. Then, the score is calculated as the path length to isolate the observation.


50
40
30
20
10
0
−2
0
2
4
6
8


2500
2000
1500
1000
500
0



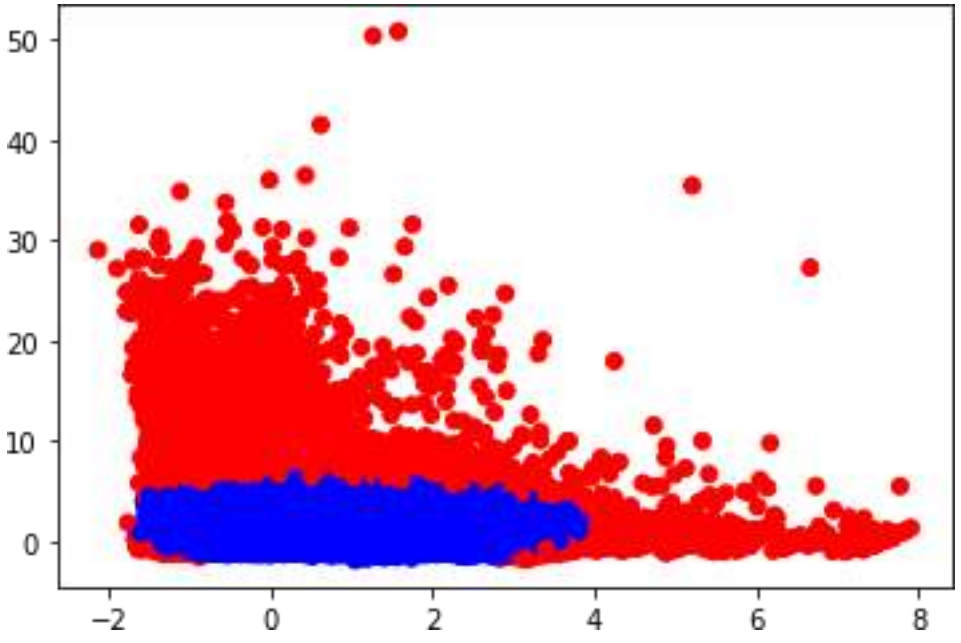



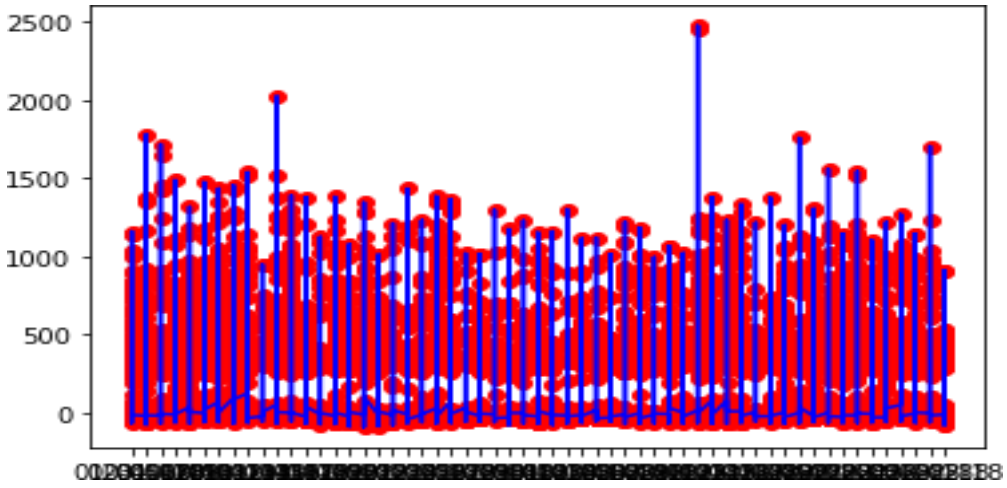

Isolation forest is implemented using sklearn, we need to give a maximum number of samples and contamination values to create isolation trees fo the dataset.

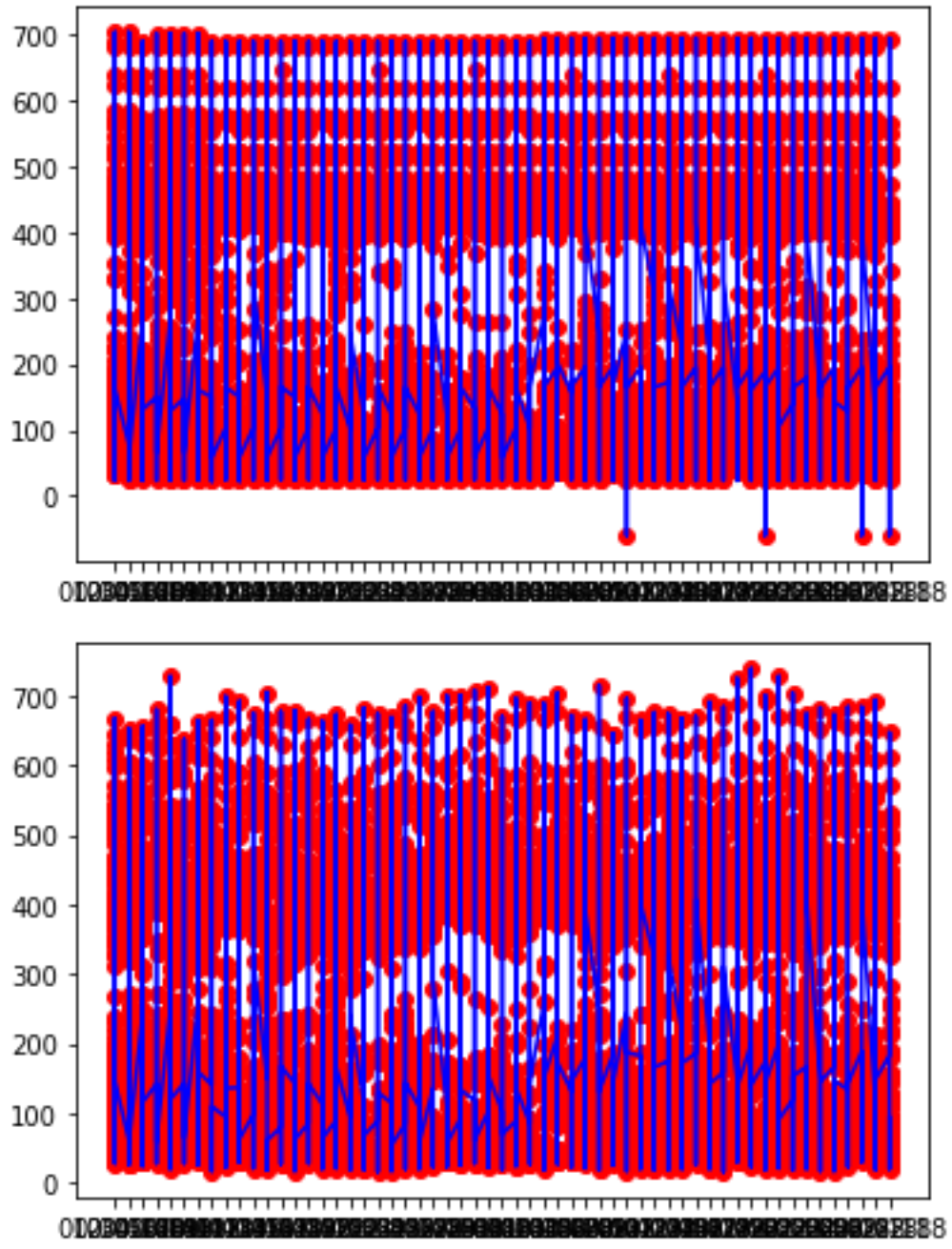
## VI)ONE-CLASS SVM:

One-class svm is an unsupervised machine learning algorithm that learn a decision function outlier detection. One-Class SVM is indeed unsupervised as there are no labelled data provided to the model. It can be considered as an outlier detection algorithm.

One-Class learn a decision boundary that achieves maximum separation between the samples of the known class and the origin. Only a small fraction of data points are allowed to lie on the other side of the decision boundary those data points are considered as outliers.






The training process is carried out only on samples representing normal behaviour, in this case, the functional margin is established to encompass all the training samples and separate them from the rest of the space.

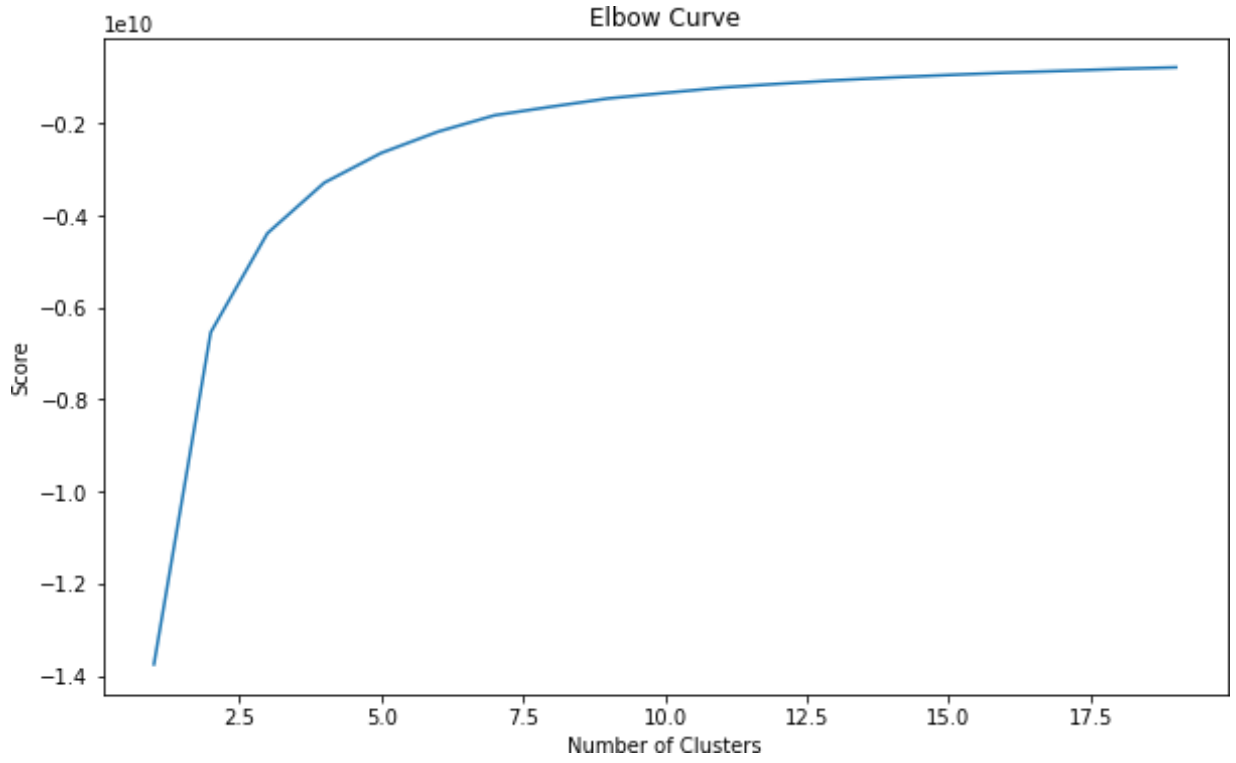
For the negative data points, it learns the boundaries of these data points and those points are able to classify any points that lie outside of the boundary. Those points are called outliers.

The parameters nu for upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. And parameter gamma for Kernel

coefficient for rbf, poly and sigmoid. One-class svm is implemented in sklearn. Graph representing the anomalies.

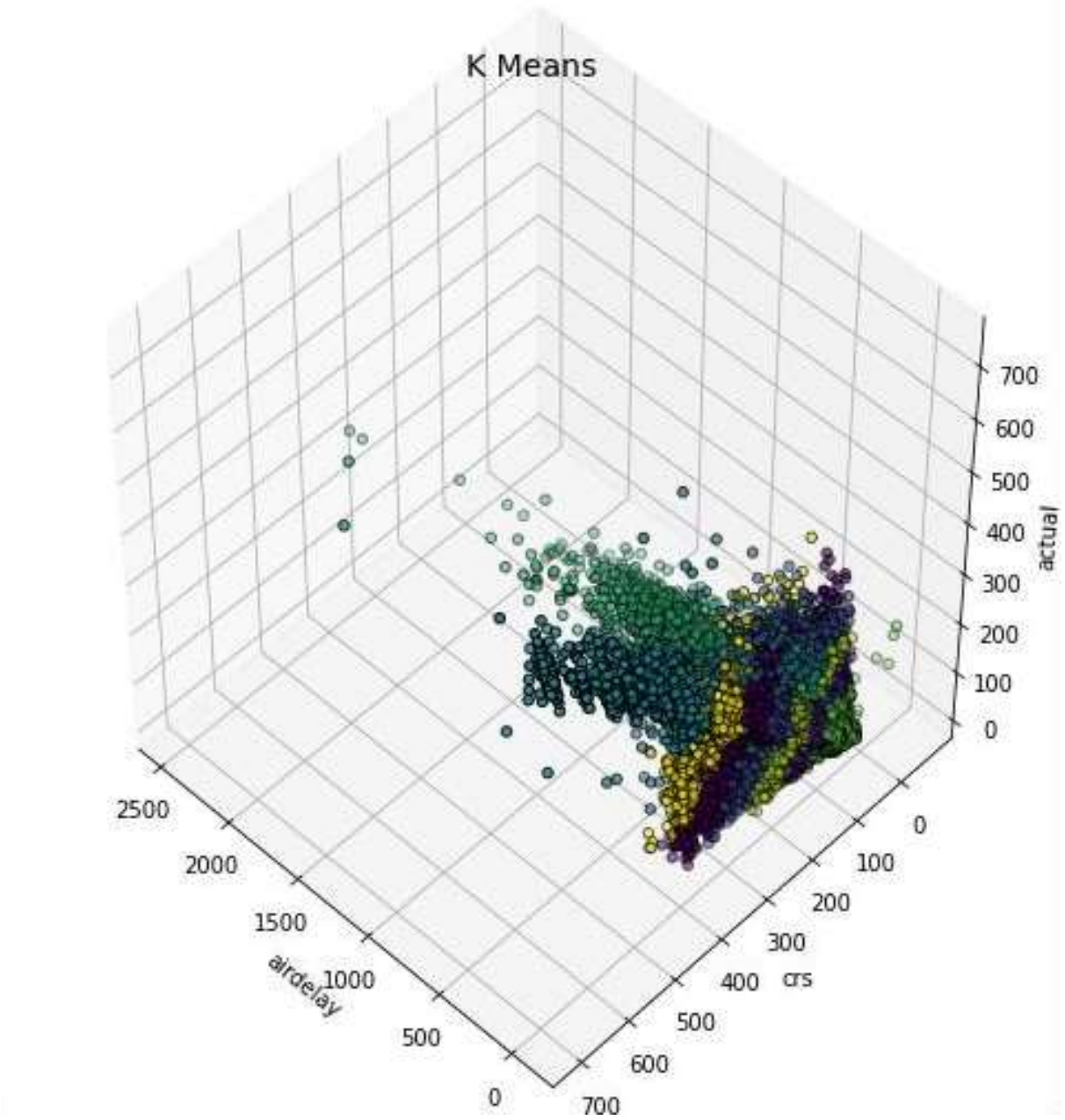
## VII)K-MEANS:

K-means is an unsupervised non -hierarchical clustering algorithm. It allows the observations of the data set to be grouped into separate clusters. Thus, similar data will be found in the same cluster. Furthermore, an observation can only be found in one cluster at a time. The same observation cannot therefore belong to two different clusters.




K-means is an iterative algorithm, which minimizes the sum of the distances between each individual and the centroid. The initial choice of centroids conditions the result .Admitting a cloud of a set of points; K-Means changes the points of each cluster until the sum can no longer decrease.

The result is a set of compact and clearly separated clusters, subject to choosing the correct value for the number of clusters. The user has to specify k (the number of clusters) in the beginning k-means can only handle numerical data k-means assumes that we deal with spherical clusters and that each cluster has roughly equal numbers of observations these points are the centres of the clusters. Assign each point (element of the data matrix) to the group of which it is closest to its



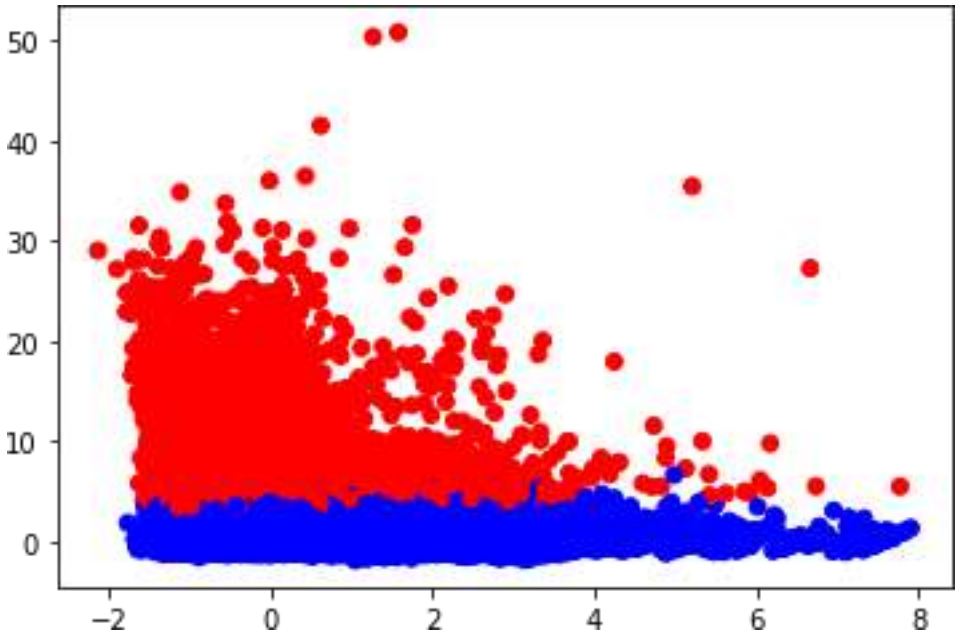

centre. Recalculate the centre of each cluster and modify the centroid

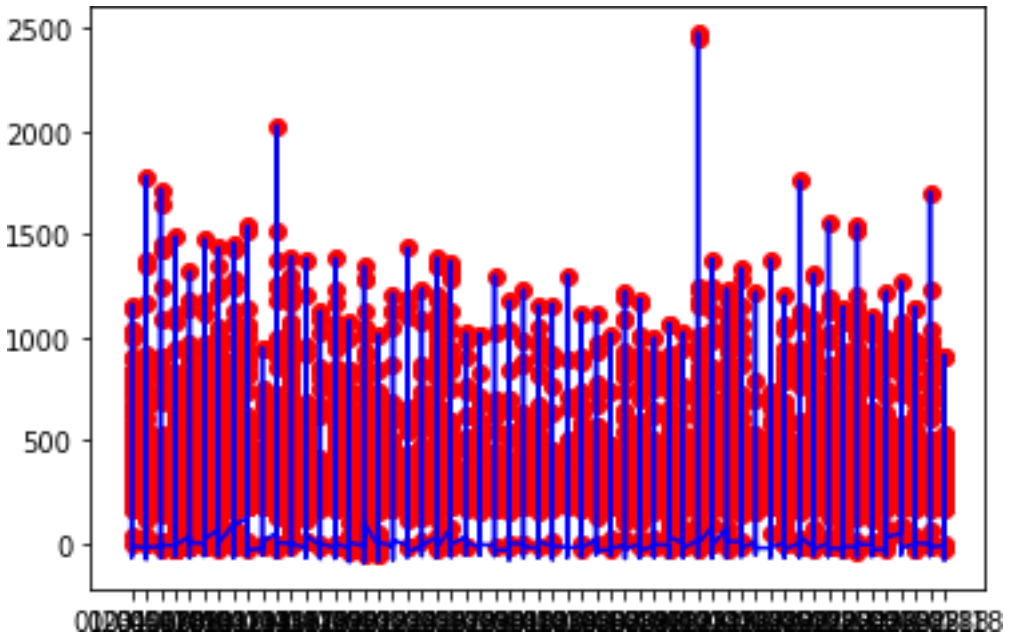
Using skit learn we implement k means algorithm Using skit learn, where we give only one parameter that is number of clusters. We obtain both centroid values and graph representing anomalies.

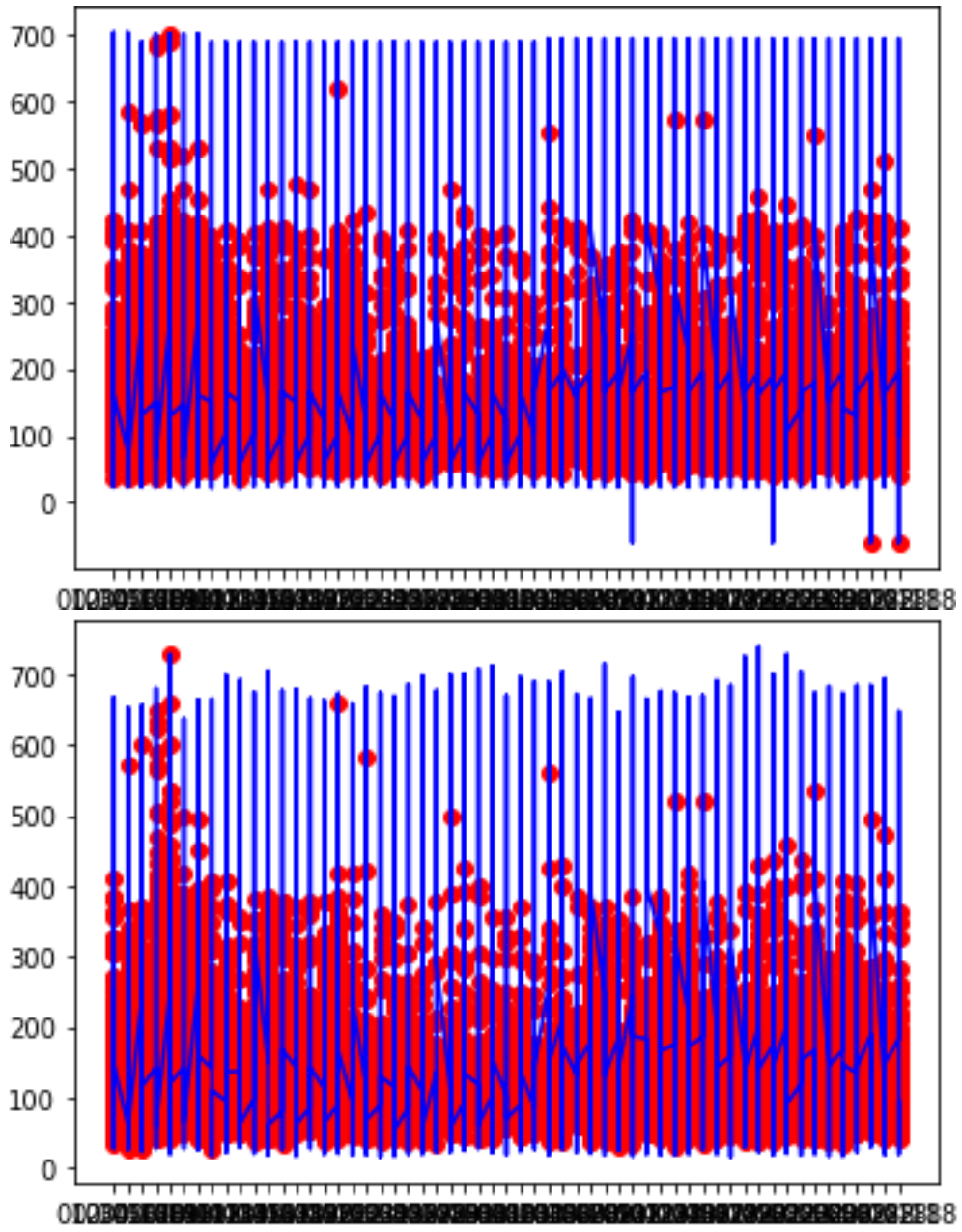
## VIII)ELLIPTIC ENVELOPE:

Elliptic Envelope is a method that tries to figure out the key parameters of our data's general distribution by assuming that our entire data is an expression of an underlying multivariate Gaussian distribution.

We set contamination parameter, which is the proportion of the outliers present in our data set. We use decision function to compute the decision function of the given observations. It is equal to the shifted Mahalanobis distances. The threshold for being an outlier is 0, which ensures a compatibility with other outlier detection algorithms.



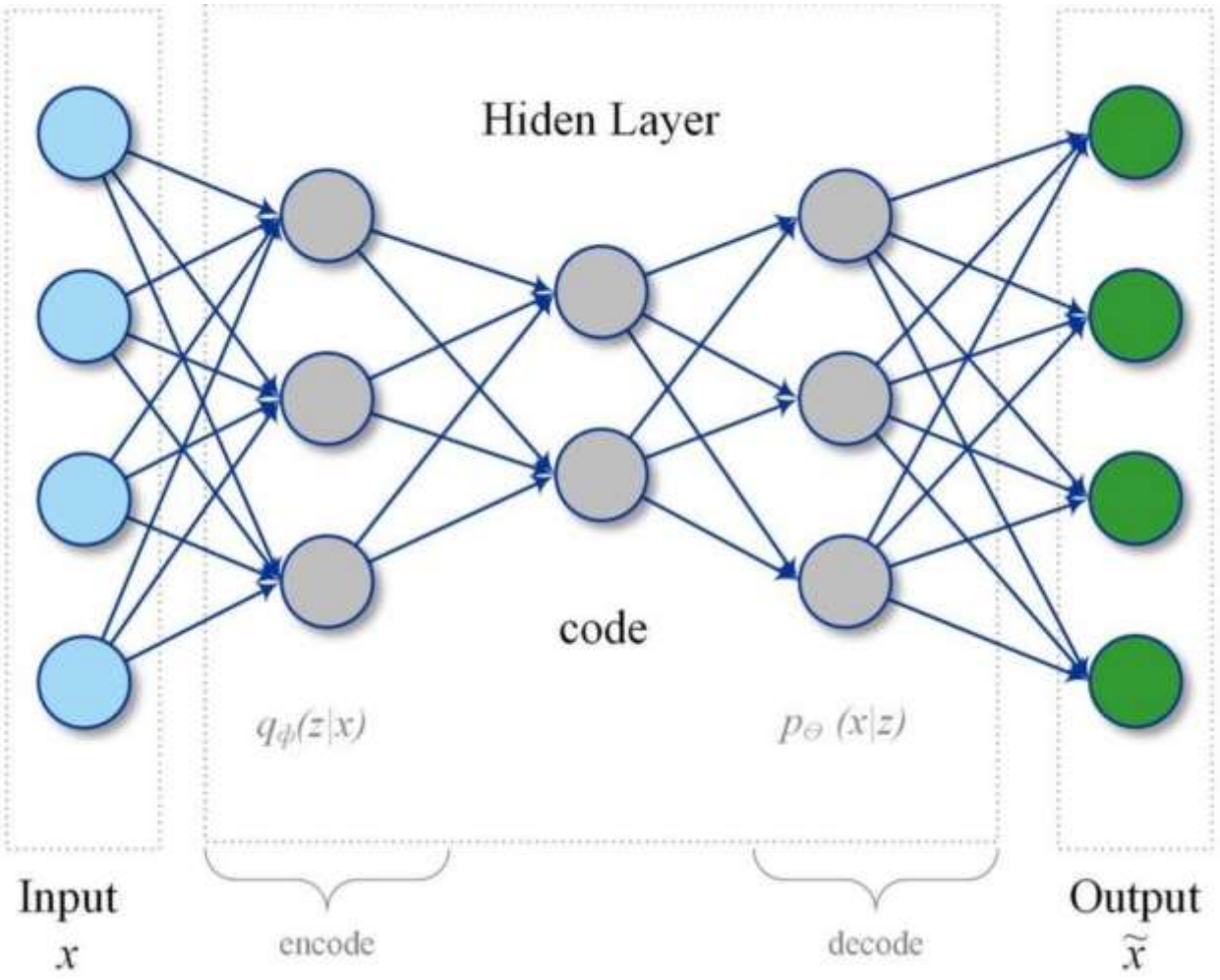






Elliptic envelope is implemented in SK learn. Graph representing the anomalies.

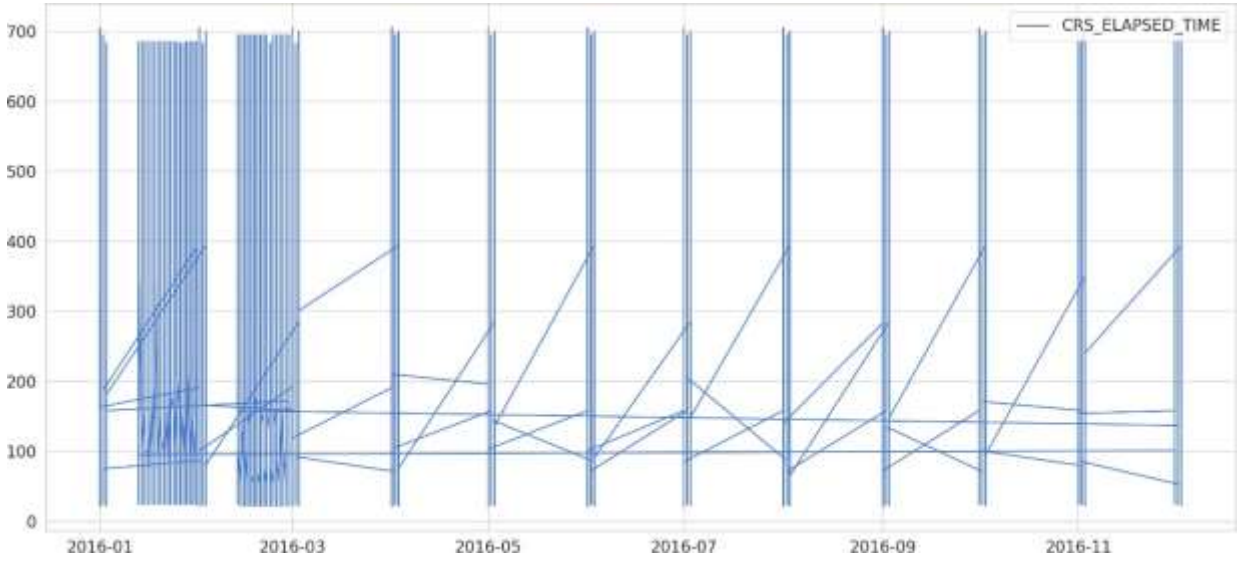
## XI)LSTM AUTOENCODERS:

The combination of two powerful concepts in deep learning LSTM and auto encoder. An LSTM Auto encoder is a usage of an auto encoder for sequence data utilizing an Encoder-Decoder LSTM architecture. The encoder some portion of the model can be utilized to encode or compress sequence data that in turn may be used in data visualizations or as a feature vector input to a supervised learning model.




Auto encoders are a type of self-supervised learning model that can learn a compressed representation of input data. To develop LSTM Auto encoder models in Python using the Keras deep learning library.

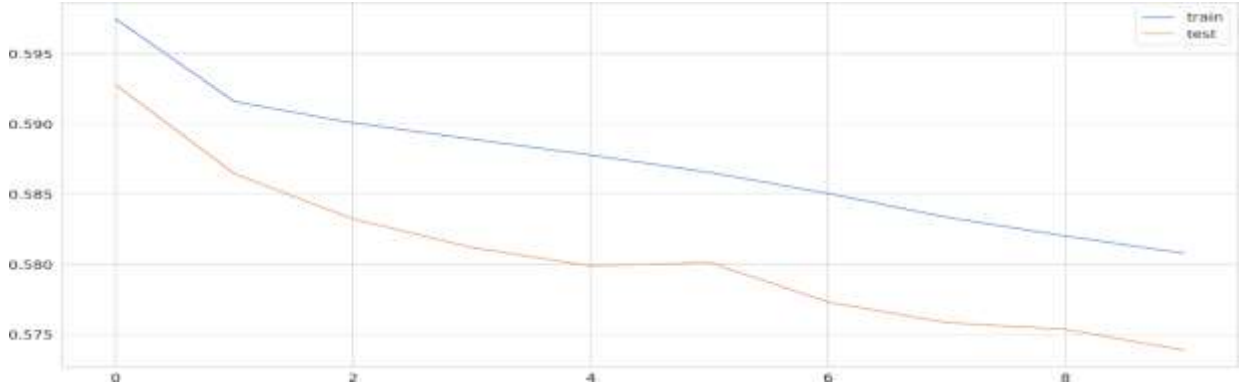
LSTM is configured to read the input sequence, encode it, decode it, and recreate it. The performance of the model is evaluated based on the model's ability to recreate the input sequence.



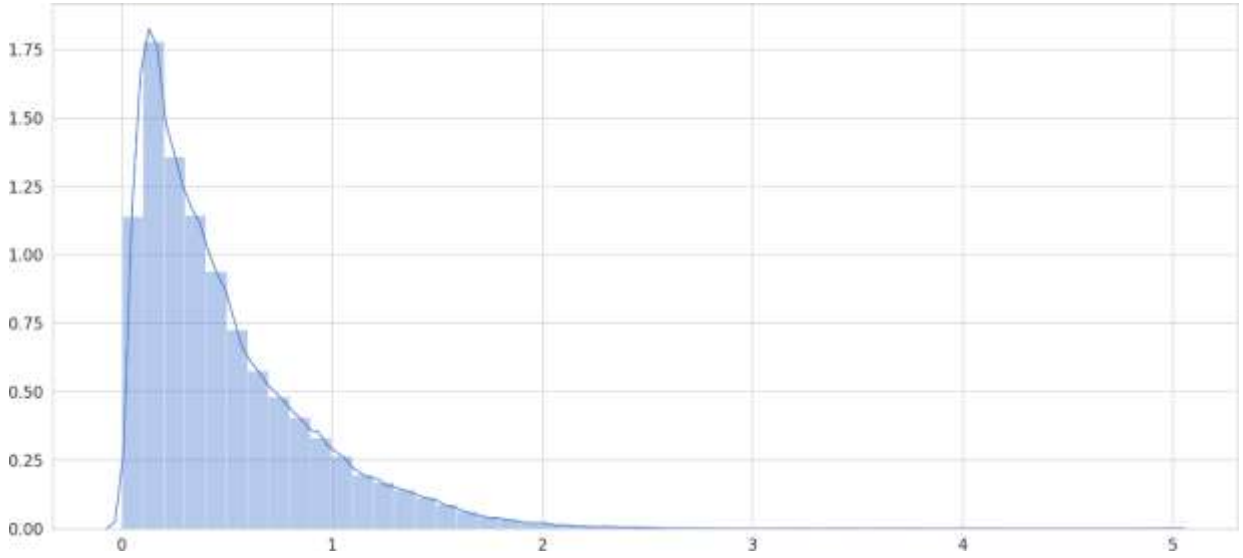

The above graphs show the results of the 2018 data with anomalies From the above graphs, the data contains only two columns flight date and actual elapsed time.

Lstm auto encoders are to learn an efficient encoding that uses fewer parameters.

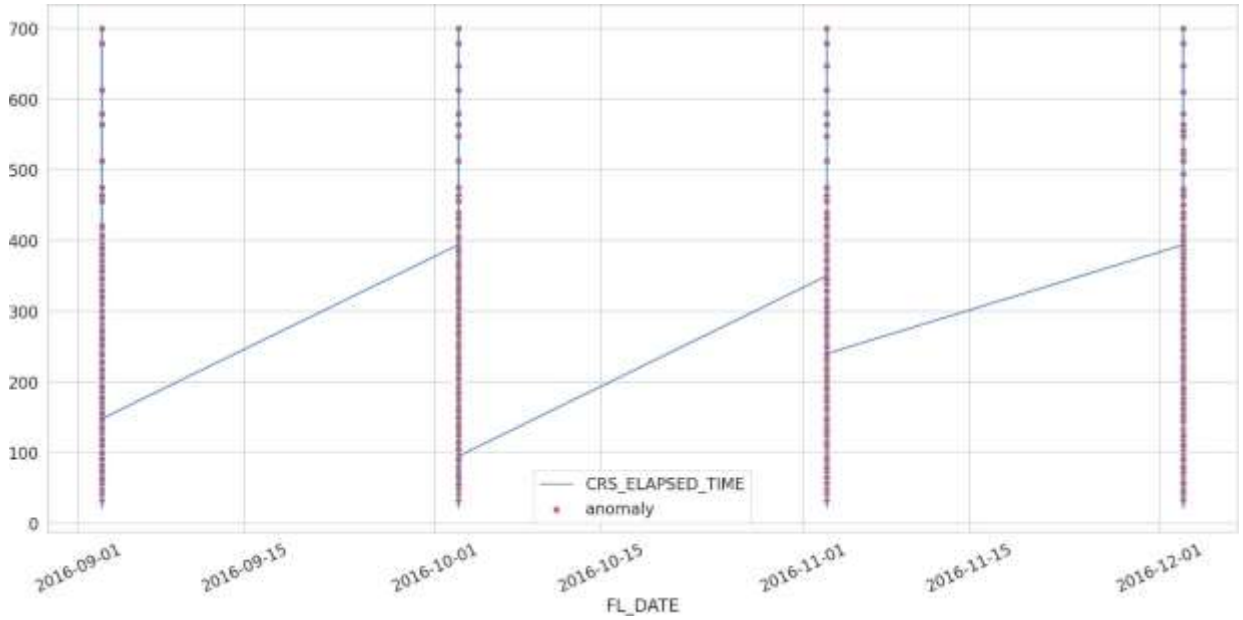
Firstly, train the auto encoder on normal data, after that take a new data point, if possible reconstruct it by using the auto encoder. If the data points are above the threshold values are considered as error values. Here, we used the 95 percentage of data to train the model. Rescale the data using the training data and apply the same transformation to the test data. Auto encoders should take a sequence as input and outs with a same sequence of the shape.



We trained the model with the 10 epochs and calculated the mean absolute error For both train and test data As we can see the graph above.




The value given for the threshold is 0.6, above this values containing the loss and the anomalies. From the graph we can see that the errors are high.




The anomalies are found in the test data are red color dots are occupied the most of the data points with steep changes to the delayed time.

Once the model achieves a desired level of performance recreating the sequence, the decoder part of the model may be removed, leaving just the encoder model.

This model can then be used to encode input sequences to a fixed-length vector.

We used the TensorFlow as our backend and Keras as our core model development library. We then set our random seed in order to create reproducible results. We define the datasets for training and testing our neural network.

## 5) TOOLS USED:

Google co-lab was used for the implantation of the code. the code is completed in python .google co-lab was used because of the size of the dataset and faster execution. Google co-lab has no configuration required, free access to gpu's and easy sharing. Google co-lab makes the execution faster with larger datasets.

## 6) EVALUTION:

Since I already told you the data is unlabeled so evaluating a model will be quite challenging so what I have done I calculated anomaly manually by using inter quatile range of 0.98 percent by using these I calculated confusion matric in order to evaluate my models.

| | isolated_forest | oneclasssvm | ellpticenvelop | lof |
|---|---|---|---|---|
| f1_score | 0.341552 | 0.212393 | 0.028874 | 0.052193 |
| accuracy | 0.978122 | 0.973830 | 0.967732 | 0.883204 |
| roc_score | 0.619937 | 0.572647 | 0.505453 | 0.519683 |
| sensitivity | 0.982270 | 0.980103 | 0.977023 | 0.977766 |
| specificity | 0.567423 | 0.352851 | 0.047969 | 0.032158 |

## 7) CONCLUSION:

To find the anomalies in the airline delays and cancellation dataset, we are used seven types of algorithms to detect the abnormality in the data. in these algorithms, the most popular algorithms like one-class SVM, isolation forest, K-means to identify the anomalous points.

Both one-class SVM and isolation forest are known to have good results when looking at accuracy and other metrics scores. However, one issue that I observed is that it could be time-consuming to get these results. Isolation forest and one-class SVM can identify the majority of anomalous points and isolation forest is for large datasets. K means the quickest yet it is less effective in distinguishing the inconsistencies. When the dataset is huge, the groups just recognize the extraordinary anomalies.

The elliptic envelope is fast and has a good accuracy score, but it takes more time to get the results. if the data had an ordinary distribution it envelops the majority of the anomalies. dBscan is good to detect the outliers, dbscan had a better performance than the local outlier factor. The local outlier factor calculates the outlier score. accuracy, f1 score are the important metrics to be consideredwhile evaluating anomaly detection algorithms.

In addition, LSTM Auto-Encoders is built using Keras and TensorFlow. LSTM Auto-Encoders is a model that can find anomalies in the delayed time, we can try to tune the model and the threshold to get even better results.

# 8)BIBLIOGRAPHY:


Data pre-processing:

https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data- imputation-with-examples-6022d9ca0779

https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825 https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Local Outlier Factor:

https://scikitlearn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html#:~:text=The%20Local%20Outlier%20Factor%20(LOF,lower%20density%20than%20their%20neighbors.

https://towardsdatascience.com/local-outlier-factor-for-anomaly-detection-cc0c770d2ebe

Isolation Forest:

https://towardsdatascience.com/anomaly-detection-with-isolation-forest-visualization- 23cd75c281e2

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

dbscan:

https://towardsdatascience.com/best-clustering-algorithms-for-anomaly-detection-d5b7412537c8 https://link.springer.com/article/10.1007/s11277-017-4961-1

One-class svm:

https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html

https://towardsdatascience.com/outlier-detection-with-one-class-svms-5403a1a1878c

kmeans:

https://medium.com/schkn/why-use-k-means-for-time-series-data-part-one-a8f19964f538 https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

elliptic envelope:

https://scikit-learn.org/stable/modules/generated/sklearn.covariance.EllipticEnvelope.html

https://www.programcreek.com/python/example/97374/sklearn.covariance.EllipticEnvelope

lstm autoencoders: