1.Implementation:

To perform anomaly detection techniques by the help of K-Means and the One-Class Support Vector Machine learning algorithms.

1.1 Dataset preparation

1.1.1 Dataset cleaning up:

The first step is to import important libraries such as pandas to manipulate the data, Numpy to perform scaling and transformations on data, matplotlib and seaborn to visualize the data.

Secondly to visualize some observations and to make a brief overview such as data dimensionality, and it's found the data has 26553 observations and 20 features with no missing values.

1.1.2 Dataset features selection:

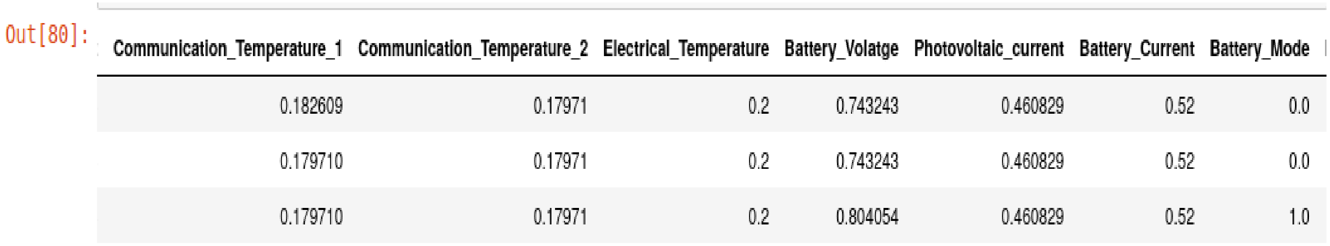
In Addition to the selection of features of interest to perform anomaly detection on, and to divide the whole dataset into a dataset of temperature and voltage. Hence we have three datasets each to be fed to learning algorithms.

1.1.3 Applying preprocessing for dataset

We utilized the MinMax scaler to convert the values to be within a range of [0,1].
here is the scaler equation:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

And the output dataset:



Out[80]:

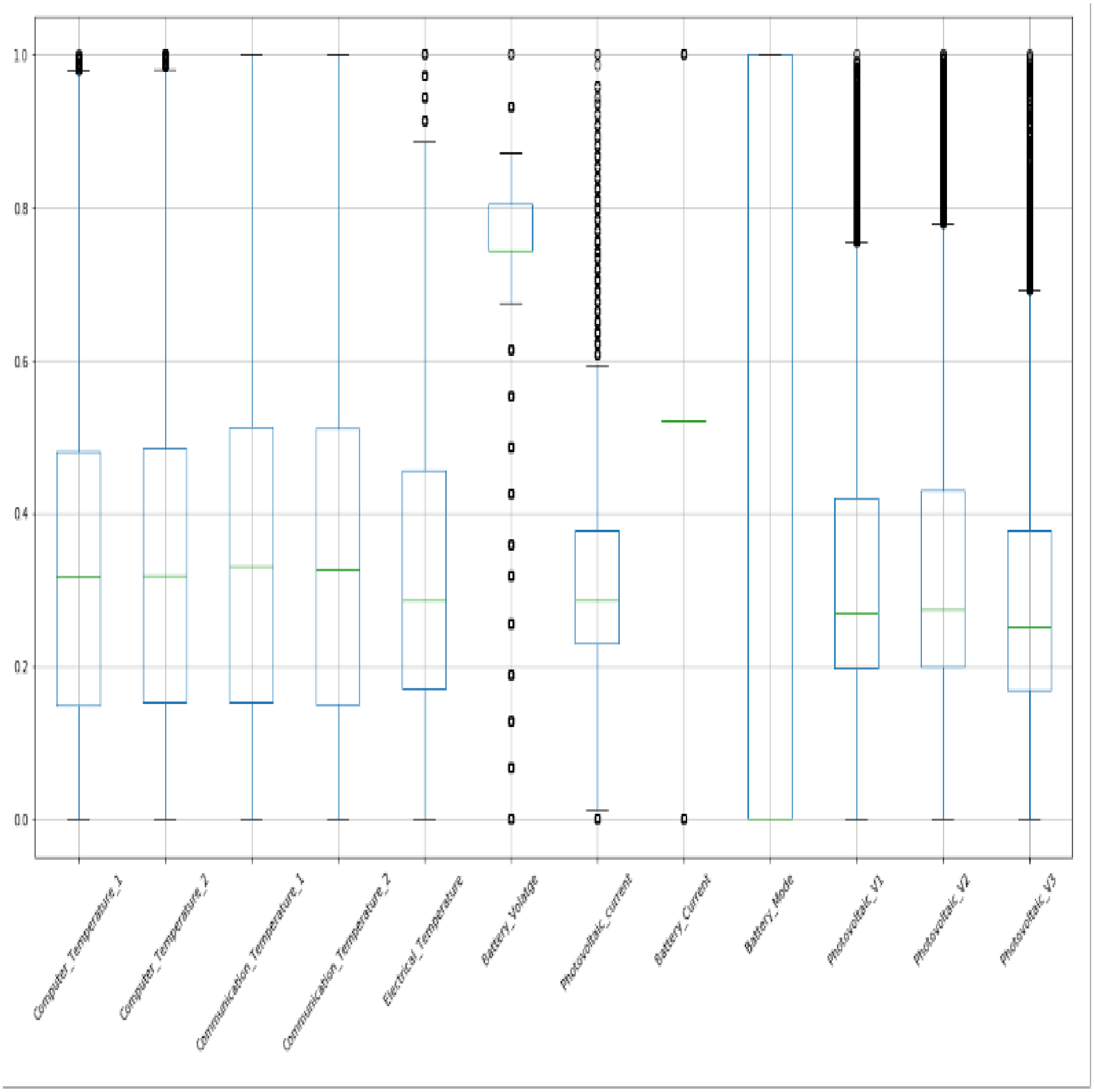
Communication_Temperature_1	Communication_Temperature_2	Electrical_Temperature	Battery_Volatge	Photovoltaic_current	Battery_Current	Battery_Mode
0.182609	0.17971	0.2	0.743243	0.460829	0.52	0.0
0.179710	0.17971	0.2	0.743243	0.460829	0.52	0.0
0.179710	0.17971	0.2	0.804054	0.460829	0.52	1.0

1.2 Visualization of the anomalies in the Dataset:

In the figure above, we constructed a boxplot of all features in the dataset.
It seems that the data has many outliers shown in dots.

There are specific features that have outliers more than the others, it appears that the data contains voltage values are more likely to have more outliers than temperature.

Then we can build an intuition before feeding the data to the algorithms, hence we expect to get a higher ration of anomalies to be predicted from voltage data, and an overall significance ration of anomalies to whole dataset size.

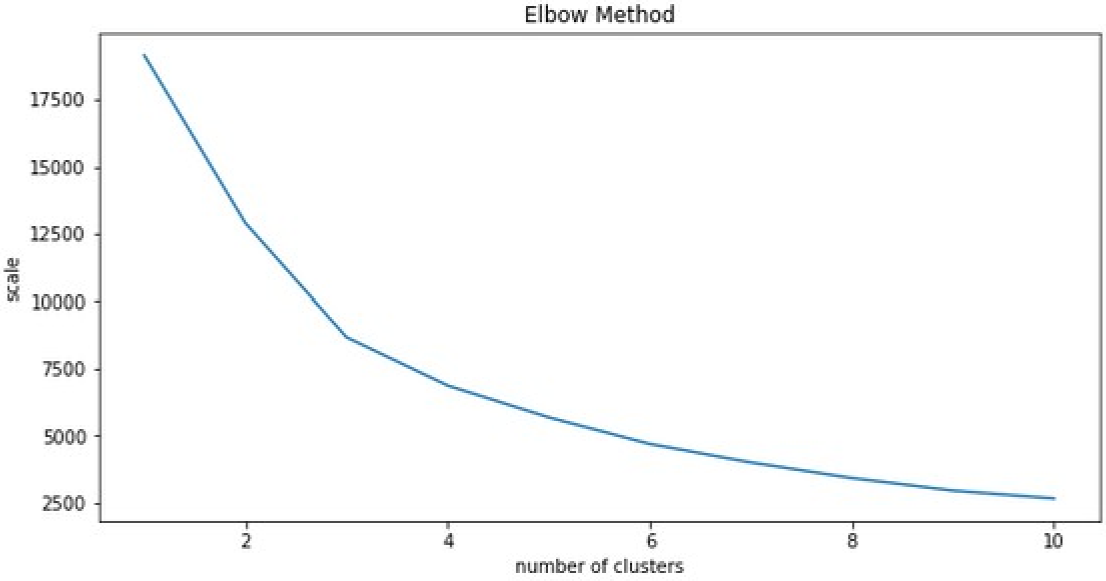


1.3 Learning and anomaly check using Kmeans algorithm:

It’s a common method called “Elbow Method” to be used before using the K-Means algorithm. Its functionality is to determine the number of clusters in a data set as the number the elbow of the curve represents.

It works by computing the model inertia which is the mean square distance between the observations and their centroid and suggests the number of the cluster according to it, hence the lower the inertia is the better.

But once the number of clusters exceeds the actual number of groups in the data, the added information will drop sharply, because it is just subdividing the actual groups which are in our case one class which the elbow method suggest to subdivide it into almost 3 classes.
If we did so, overfitting is occurring due to the model would learn the detail and noise in the training data.

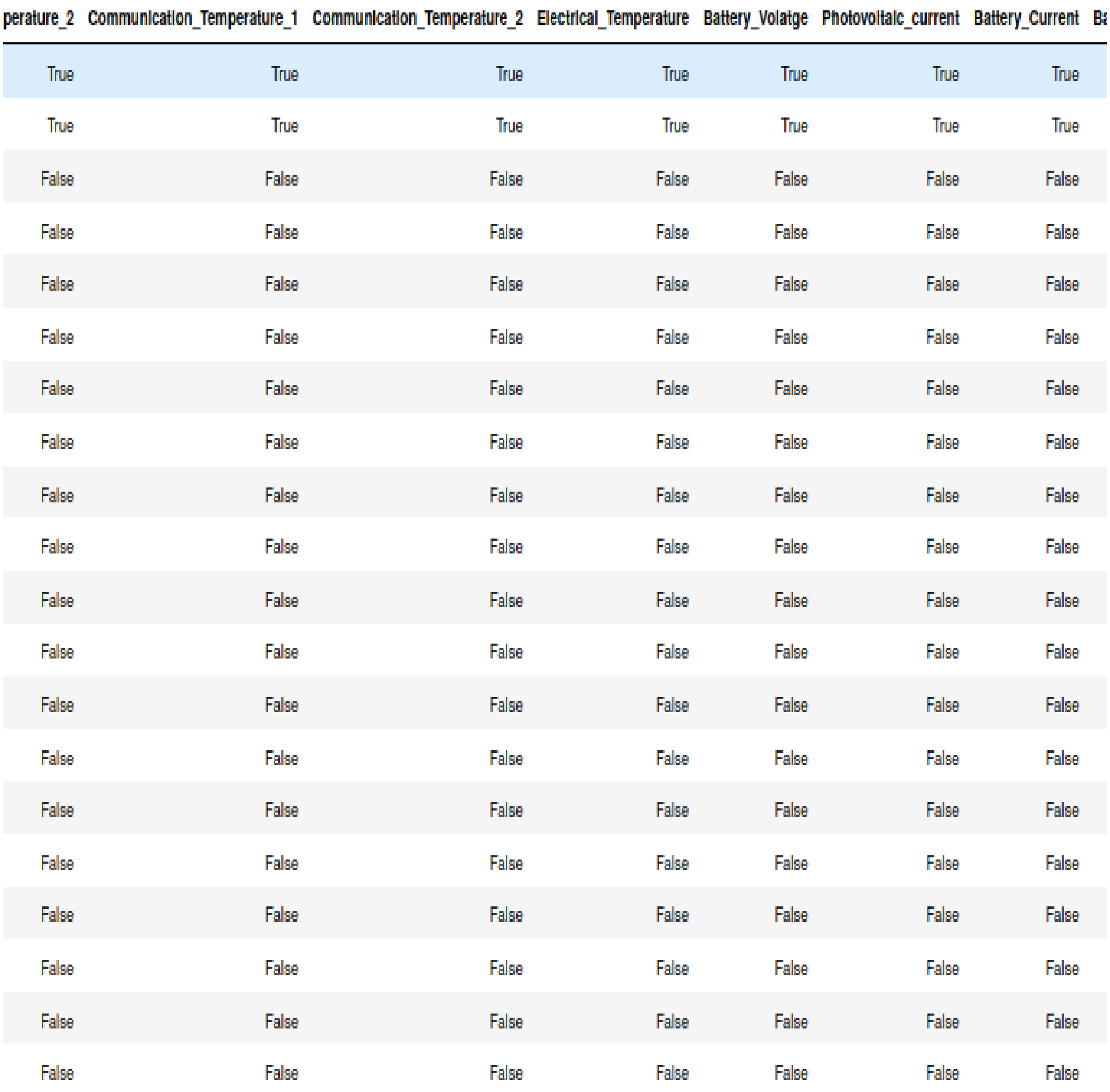


1.3.1 Fit the whole dataset:

we fit the data with one cluster, and in the figure below, we have plotted the average of inertia between the cluster centroid and the all the observations to construct a threshold by which we can classify data with certain interval as an anomaly. We make the anomaly threshold for above the value of 0.3 and under 0.1, and these data point is 2.7% of the whole dataset.

As the dataset does not have a label, we combine the two anomalies datasets predicted from both of the algorithms. First, we assure that the two predicted anomalies dataset from both of the algorithms are different from each other.

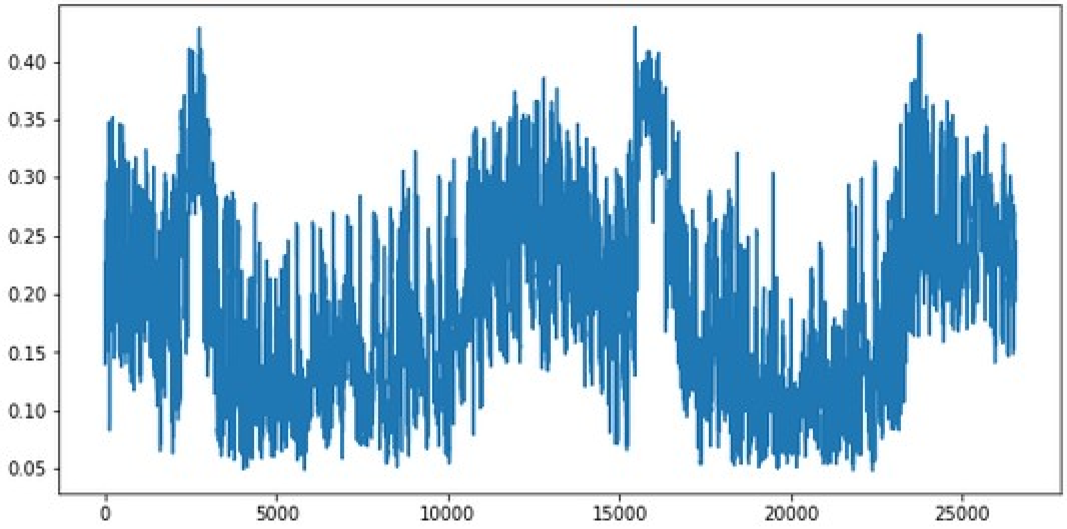
We simply made a check if the two anomalies dataset are different and the result is below:



perature_2	Communication_Temperature_1	Communication_Temperature_2	Electrical_Temperature	Battery_Volatge	Photovoltaic_current	Battery_Current	Ba
True	True	True	True	True	True	True	
True	True	True	True	True	True	True	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	

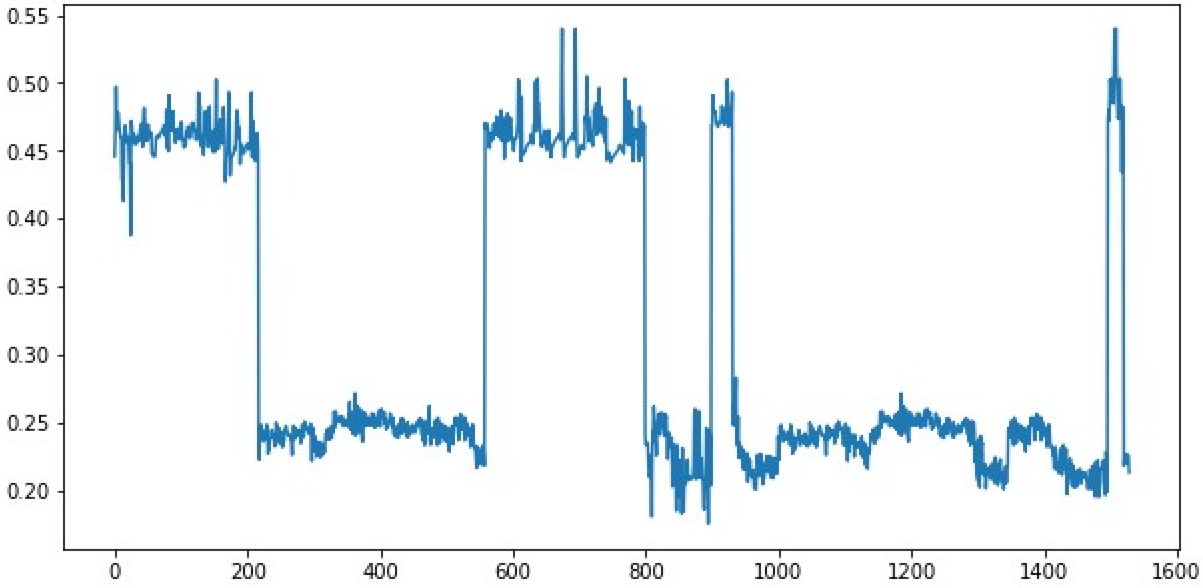
It seems that they have insignificant similarity, then we perform the combination between them safely. Hence both of the algorithms would be trained again with anomalies data that wasn't predicted by it.

Then these data points are to be saved to a dataset of anomalies as on further steps we fit again with these anomalies to test model accuracy.



We refeed the dataset of anomalies again for both of the K-Means and the One-Class SVM as this dataset to be our original data.

After feeding the dataset of anomalies to K-Means again and plot the average of inertia we got this figure:



Then to construct a threshold to determine anomalies above the value of 0.25.

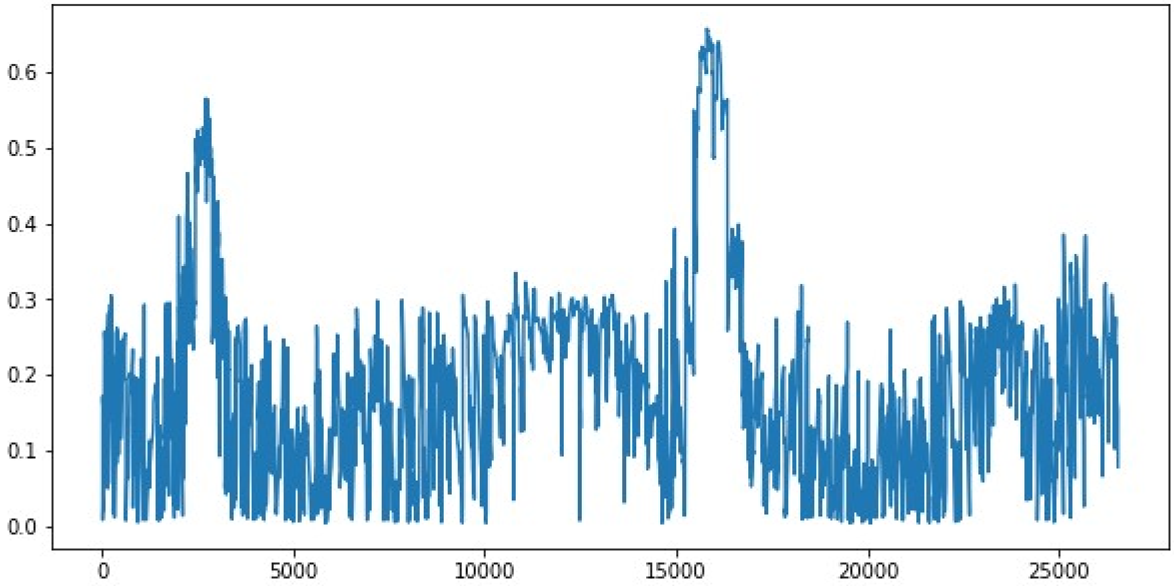
The proportion of anomalies predicted by K-Means to the anomalies dataset size is 44%.

1.3.2 Fit the temperature dataset:

We recall from boxplot above that the temperature data separated has low outliers or anomalies.

In the figure below, a plot of the average of inertia between temperature dataset and cluster centroid obtained from feeding temperature dataset to the K-Means algorithm.

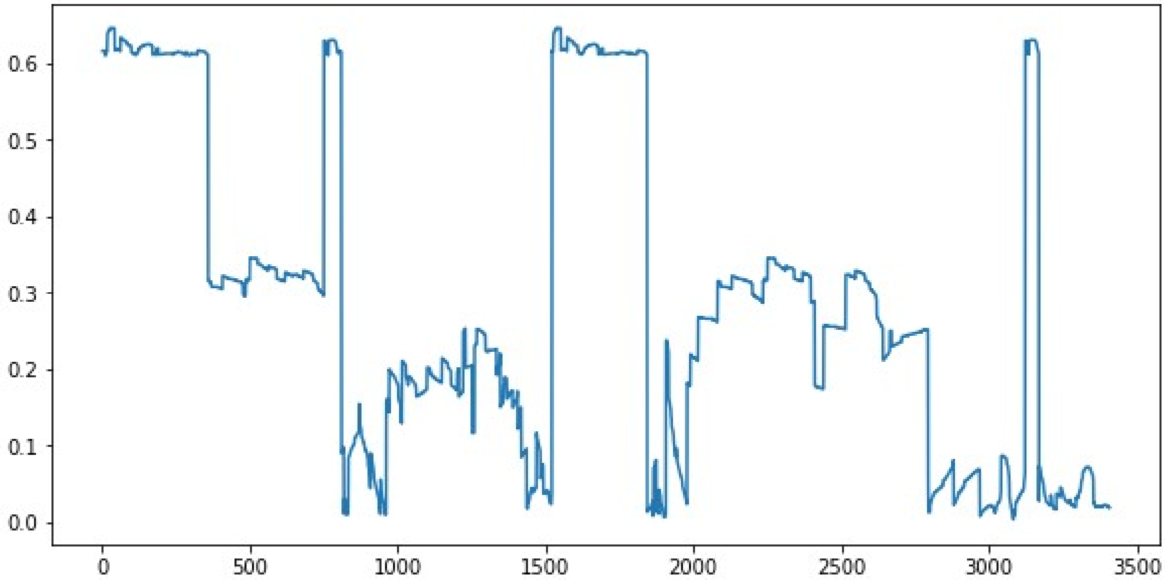
It appears to construct a threshold to specify anomaly data above the value of 0.3



Then we combine the two data frames predicted by both K-Means and One-Class SVM and construct a data frame of temperature anomalies.

The proportion of anomalies predicted by K-Means to the temperature dataset size is 9.8%.

Hence the average of inertia is plotted in the figure below:



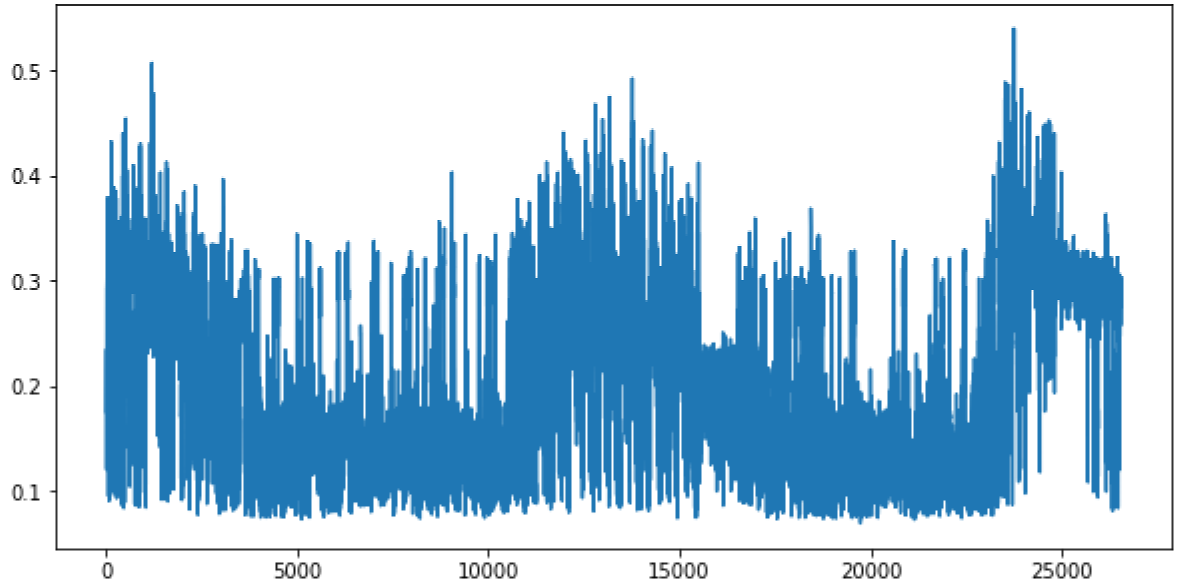
We construct a threshold of anomalies above the value of 0.3 and repeat the process of combining the predicted anomalies from One-Class SVM and this K-Means prediction to be fed again to both of the algorithms.

The proportion of anomalies predicted by K-Means to the anomalies temperature dataset size is 45%.

1.3.3 Fit the voltage dataset:

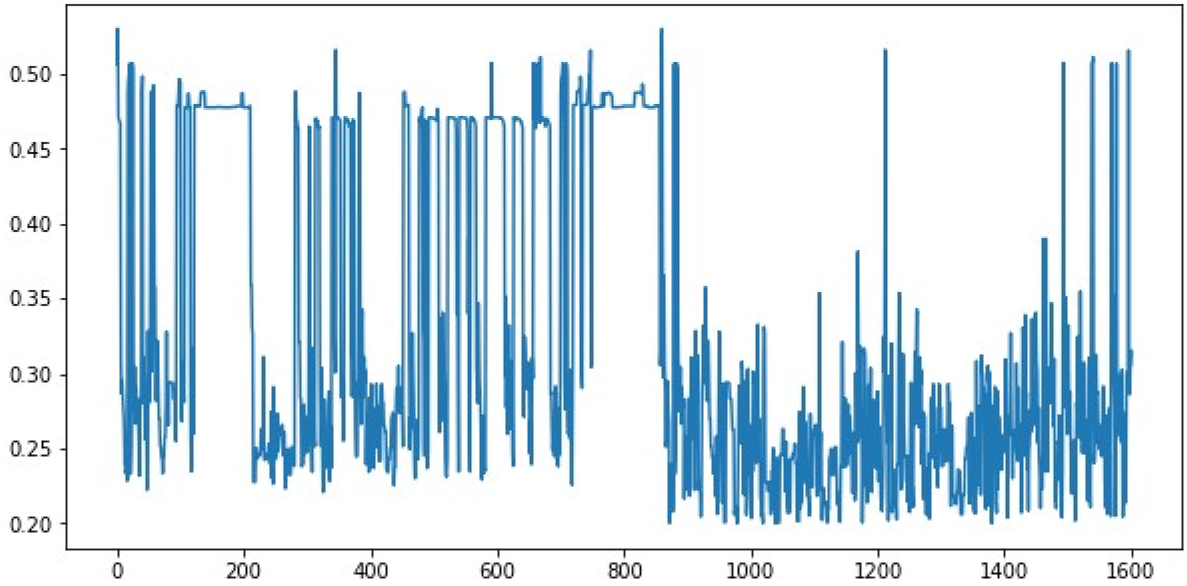
We recall from boxplot above that the voltage data has the majority of anomalies. Hence we expect the learning algorithms to predict more than temperature dataset and the whole dataset.

The voltage data is to be fed and the average of inertia is to plot in the figure below:



We set a threshold of anomalies to be the data above the value of 0.35.
The proportion of anomalies predicted by K-Means to the voltage dataset size is 2.7%.

By repeating the process of combining anomalies predicted from both of the algorithms to a dataset of anomalies, and replotting the average of inertia we obtained the figure below:



The proportion of anomalies predicted by K-Means to the anomalies voltage dataset size is 30%.

1.4 learning and anomaly check using One-Class SVM Algorithm:

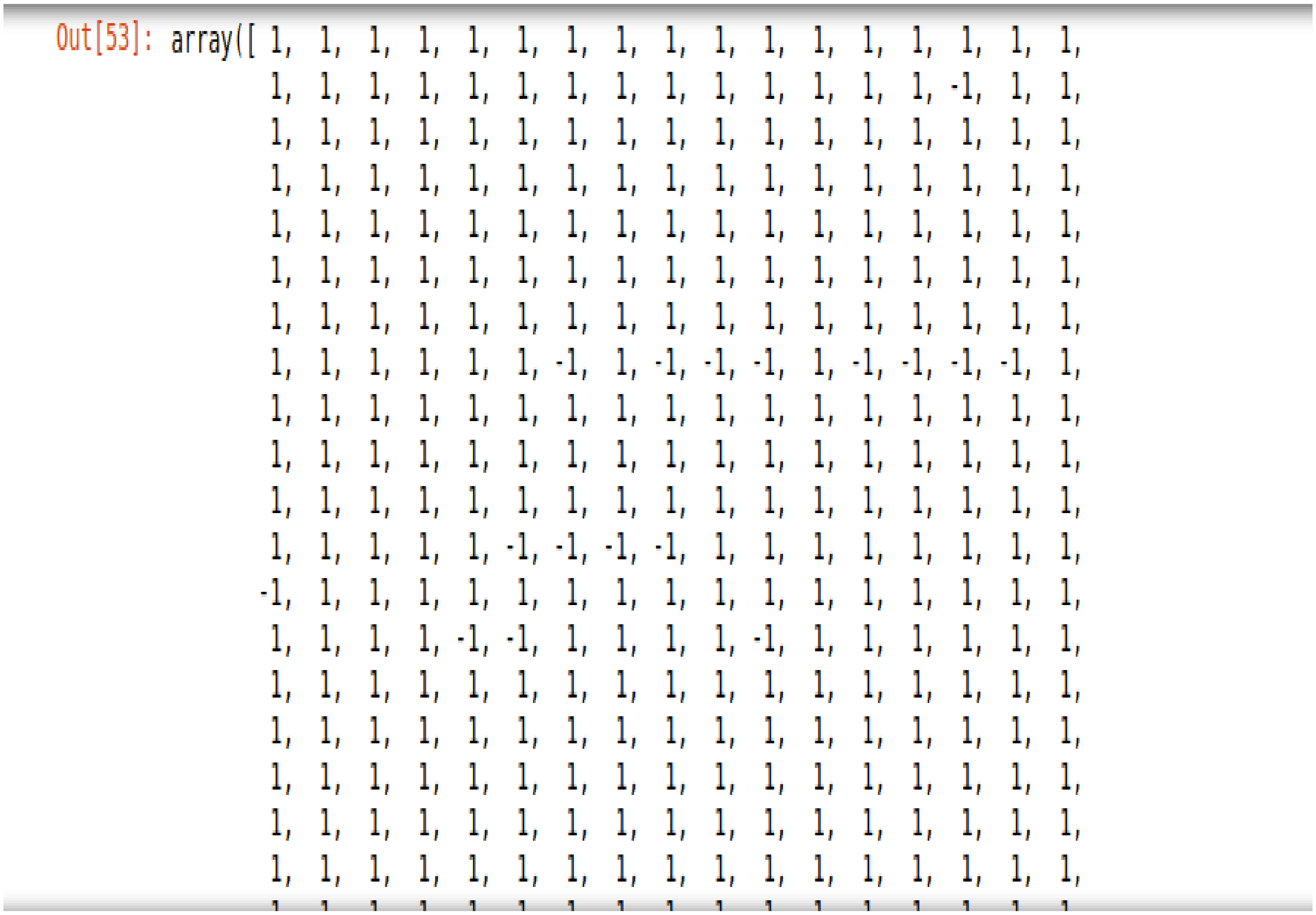
1.4.1 Fit the whole dataset:

We fed the whole dataset to One-Class SVM setting the parameters of RBF kernel function, gamma which is kernel coefficient to the value of 0.001, and nu which is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors to the value of 0.03.

Similar to K-Means setting a threshold based on the inertia, we used the radial basis function which calculates the squared euclidean distance between each two feature vectors.
Here is the RBF equation:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

After making a prediction, we got an array of ones and negative ones which represent anomalies predicted.
Here is the prediction output sample:



```
Out[53]: array([ 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1, -1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1, -1,  1, -1, -1, -1,  1, -1, -1, -1, -1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1, -1, -1, -1, -1,  1,  1,  1,  1,  1,  1,  1,  1,
                -1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1, -1, -1,  1,  1,  1,  1, -1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
                 1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,
```

The proportion of anomalies predicted by One-Class SVM to the whole dataset size is 3%.
After the combination of the two data frames of anomalies predicted by both of the algorithms, the proportion of anomalies predicted to the anomalies dataset size is 78%.

1.4.2 Fit the temperature dataset:

We repeat the process of feeding the dataset to the algorithm but this time to the temperature dataset.

The proportion of anomalies predicted to the temperature dataset size is 3%.
After the combination of the two data frames of anomalies predicted by both of the algorithms, the proportion of anomalies predicted to the anomalies temperature dataset size is 46%.

1.4.3 Fit the voltage dataset:

We repeat the process of feeding the dataset to the algorithm but this time to the voltage dataset.

The proportion of anomalies predicted to the voltage dataset size is 3.2%.
After the combination of the two data frames of anomalies predicted by both of the algorithms, the proportion of anomalies predicted to the anomalies temperature dataset size is 79.5%

1.5 Results and comparison:

We compare the two algorithms trained on anomalies datasets by proportion of anomalies predicted.

Algorithm/ Dataset	SVM	K-Means	Best Algorithm?
The whole	78%	44.2%	SVM
Temperature	46%	45.1%	They are very close but SVM is better by a little difference.
Voltage	79.5%	30%	SVM