



Project on Vehicle Dataset

**DATA MINING PROJECT REPORT**

Chandraprakash

Table of Contents

[Abstract 3](#_bookmark0)

[Introduction 3](#_bookmark1)

[Problem Statement 3](#_bookmark2)

[Supervised 3](#_bookmark3)

[Unsupervised 3](#_bookmark4)

[Methodology 4](#_bookmark5)

[Data Collection 4](#_bookmark6)

[Data Pre-Processing 4](#_bookmark7)

[Exploratory Analysis 10](#_bookmark8)

[MachineLearning 13](#_bookmark9)

[Supervised & Un -supervised Machine Learning 13](#_bookmark10)

[Supervised Task 13](#_bookmark11)

[Problem Statement 13](#_bookmark12)

[Unsupervised Task 17](#_bookmark13)

[Problem statement 17](#_bookmark14)

[Conclusion 20](#_bookmark15)

[Learning from the project 21](#_bookmark16)

[Referances 21](#_bookmark17)

# Abstract

In this Data Mining project, we got an opportunity to work on real world dataset and problem formulation. In which the dataset was provided which is based on vehicle information such as electric, hybrid &diesel mode vehicle use in different countries and there are total 74 features related to vehicle was present in the dataset. We have to analyze the dataset and generate challenging problems from it and then solve those challenging problem using machine learning techniques (Supervised and Unsupervised).

# Introduction

Data mining is a process to apply a set of methods for data analysis, exploration and discovering the problem solution in a large and complex dataset. In this project, we have provided dataset related to vehicle information that operates in electric, diesel and hybrid modes and countries or geographical area in which the vehicles are used. The names of the countries are encrypted with numbers. This dataset contains total 74 features and 861916 samples corresponding to 4031 vehicles, however, the measurement of each vehicle are not updated regularly. The challenging task is to analyze the dataset and finding the problems and then solved by performing data mining process and machine learning techniques. After analyzing the dataset, we have found the below two problems and applied the data mining methods in which data preprocessing is one of the most important and challenging tasks are performed which includes data cleaning, data exploration and etc, then applied machine learning models which are used for finding the best optimal results for those problems.

# Problem Statement

## Supervised

1. Predicting the Operation time using Regression Models.
2. Prediction of multiclass classification feature ‘Battery Supplier’.

## Unsupervised

1. Anomaly detection of vehicle Operation time with respect to vehicle operation distance using K means clustering .
2. Anomaly detection of vehicle Operation time with respect to vehicle operation distance using Isolation Forest.

# Methodology

The following steps are performed to achieve the goals which are listed below ;

## Data Collection :

Dataset was provided from the University and which was loaded and changed it into an appropriate format. Imported all the libraries which were needed to perform sampling, visualization, data modeling, data evaluation steps etc. Discovered the data to get insight on it, checked the data counts, statistical summary of attributes, break down the data based on target goal, and taken the features which are correlated to target variable for the prediction values.

## Data Pre-Processing :

This is one of the most important processes, i.e. cleaning the raw data by removing duplicate samples, treating missing values, outlier removing and prepared for modeling and evaluation of data. When the data is gathered, it is collected in a raw format and this data is imbalanced and not feasible for directly analysis purpose, so certain steps are executed to prepare the raw data into clean data which can be analyzed and used for modeling purpose.

Machine Learning algorithms could not give good result with missing features, so it was handled by inserting the median values and removing the null data.

After loading the raw dataset, found that there are multiple duplicate and wrong values present in the dataset like a particular country should be lie in specific geographical area but instead of this it lie in different geographical areas. For an instance, the ‘Country’ 3 lie in geographical areas ‘1,2,3,5,6’. Other features also have wrong values like city 4 lies in different countries ’27,26,22,21,1,6’ and also raw dataset contains lot of null values.

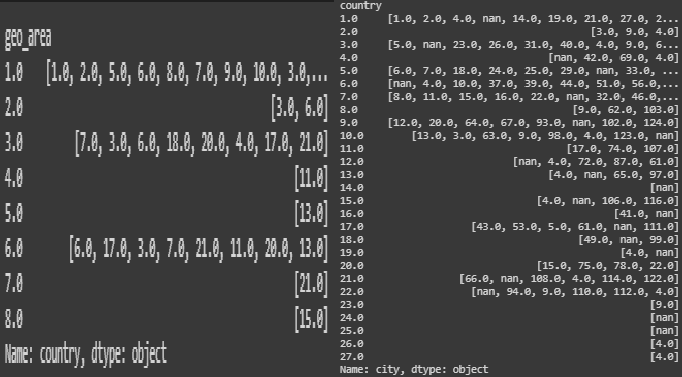


Figure 1 : wrong values in geo\_area and in country .

* 1. While doing the data cleansing, we have also found that feature “Sample\_Id ” is not unique, however, it should be unique for all recorded vehicle id. Lots of duplicate entries were there which are approximately 84449.

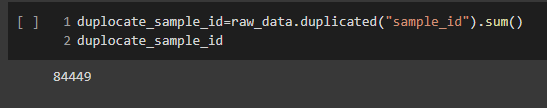


Figure 2: Duplicate values in sample\_id

* 1. For duplicate Sample\_id, some features have different values, for instance, “Battery\_Version”, “Battery\_replacement\_date”, “Battery\_Genreation” and “Battery\_Supplier” have different values, however, the sample id is same.

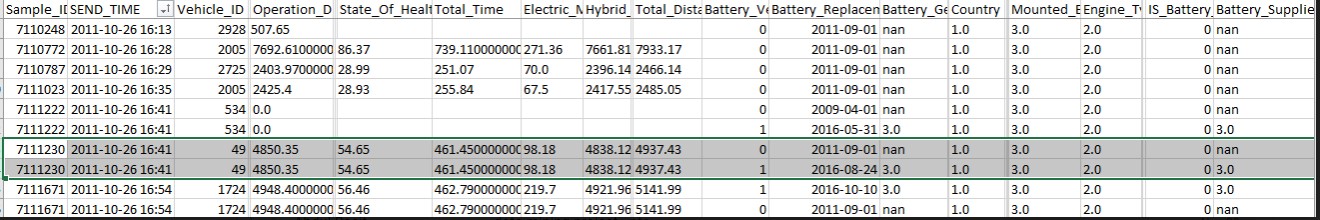


Figure 3: Different values in some features for duplicate sample\_id

* 1. The feature “Is\_Pattern\_available” gives the information of data logged or not logged. The numeric value ‘0’ shows the data is not logged whereas the value ‘1’ shows the data is logged, so taken the data which has value ‘1’ and got 40580 records has the value ‘1’.

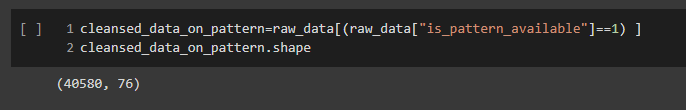


Figure 4: Data cleasingwrt “Is\_Pattern\_available”

* 1. After selecting the logged value which is equal to 1, found that there are 106 duplicate sample id still exists, which are removed by considering the features ‘Battery\_Version’and ‘Is\_Battery\_Changed’. We have taken these two features because the feature ‘Is\_Battery\_changed’ represent battery is changed or not if it turns to ‘1’ means the battery is changed and‘0’ mean not changed, therefore, if the ‘Is\_battery\_changed’ is equal to ‘1’ then ‘Battery\_Version’ should not be equaled to ‘0’ as ‘Battery\_Version’ represent number of times battery changed and vice versa. We have also found while narrow down the data that there are 23 records which

values are ‘0’ in feature ‘Battery\_Version’ and ‘1’ in ‘Is\_Battery\_Changed’ which are wrong therefore it was removed.

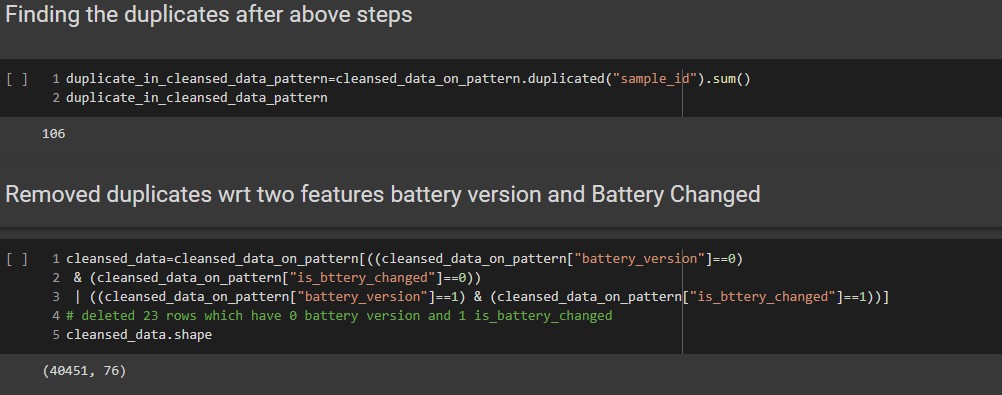


Figure 5: Duplicate data handling

* 1. The below graph shows the missing values present in all the features in which the maximum 93% values are missing for active\_charge feature. We have used only those features which have lesser than 40% missing values and removed the rest. After removing, the shape of the data set is (40451,42).

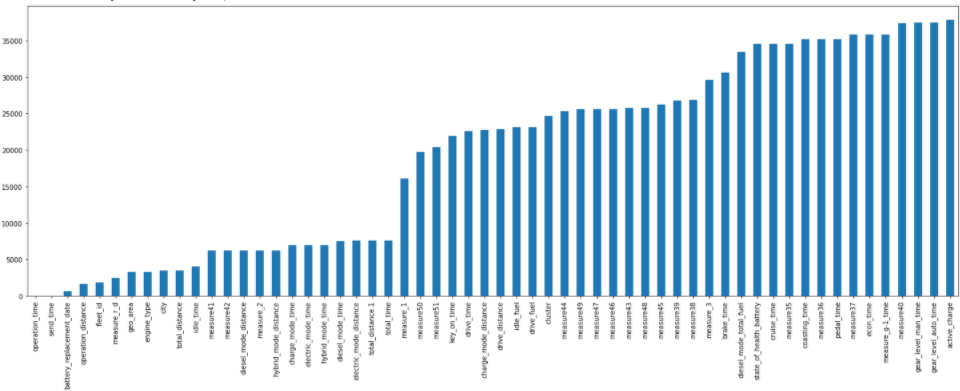


Figure 6: Missing values( %) in features

* 1. After cleaning the raw data, splitted the data into numeric and categorical features. After that, checked the outliers on numerical data, the below box plots show the outliers :

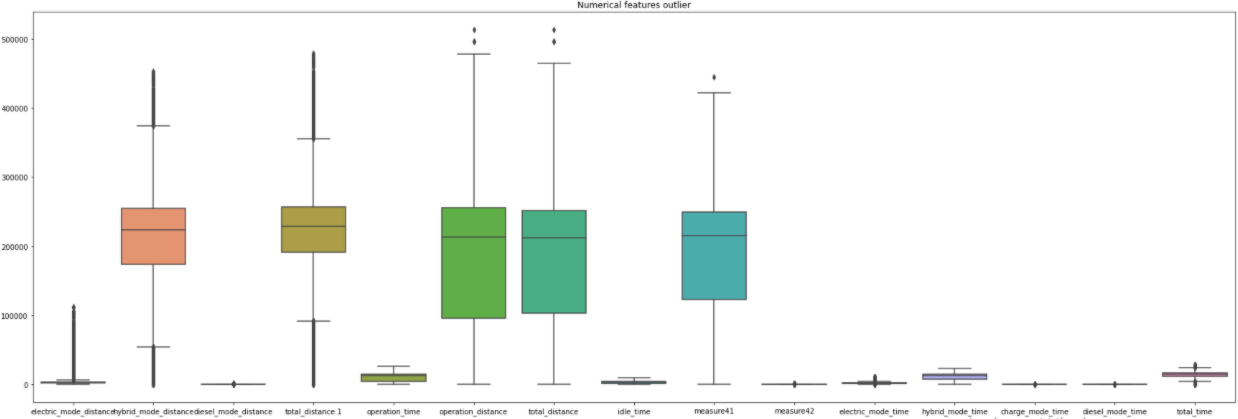
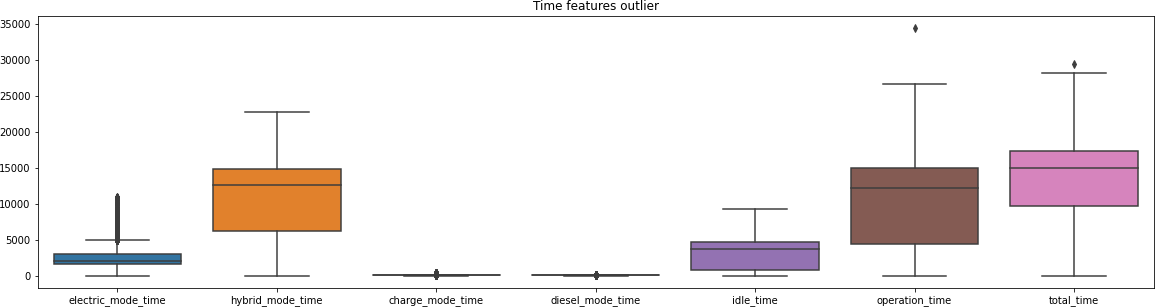


Figure 7: Numerical Feature Outlier

Here, we could see, the outlier is not showing in time features because the scale of the distance features is larger, so, we divide numeric dataset into further time or distance features for checking the outliers. Shown in below figures.



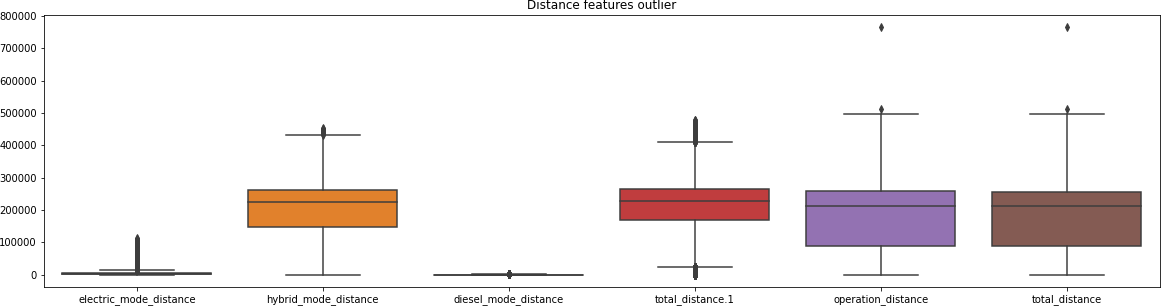


Figure 8: Time & Distance Feature Outlier

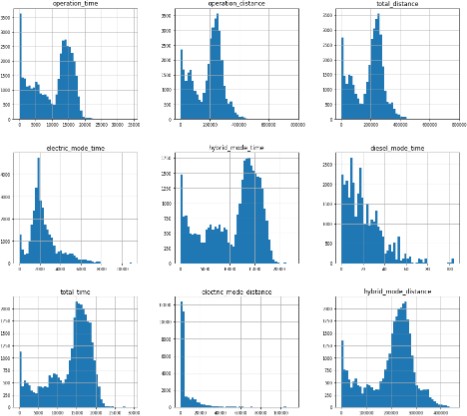
* 1. After removing the outliers from distance and time features, we checked the distribution of the numeric data, and verified whether it is normally distributed or skewed. The distribution says that which statistical method can be applied on dataset for the imputation. The below graph shows the numeric distribution

Figure 9: Features Distribution

From above figure, it shows that data is not normally distributed it skewed toward positive or negative. In this case, taking the mean for imputation is not a good idea because mean takes every data points. If we have a outliers than it effect the distribution of the dataset. It can drag the mean up and down. So, in this case, median would be the best choice for imputation.

Moreover, for categorical features used mode to fill null values.

1. For numerical feature selection, we perform Pearson Correlation by using heat map and analyze that features are highly correlate with each other, so, we apply PCA on features because PCA works well when we have a strongly correlated features. For categorical features, we also apply PCA. The below figure shows the numerical feature correlation heatmap.

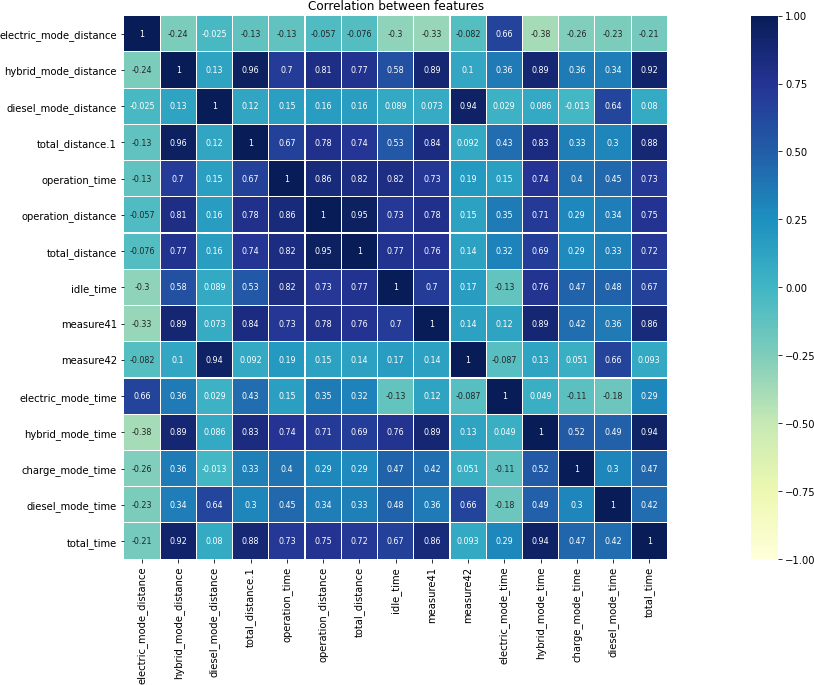


Figure 10: Correlation between features

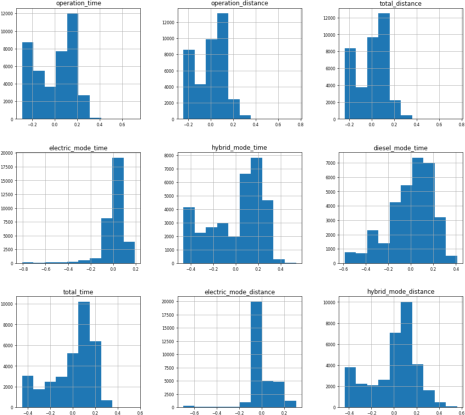
* 1. Normalizing the numeric features, which is then used when the features are in different ranges. If we have a feature with a large range, then it will affect the model. So, it is important to take all features in a same range to perform the model better therefore applied Min –Max approach for normalizing the data. The below figure shows the normalization on numeric features.

Figure 11: Normalization numeric features

* 1. After normalization of the data, we applied PCA (principle components analysis) on both numerical and categorical features after selecting the target feature.
  2. Used SMOTE( synthetic minority oversampling technique) for handling the class imbalanced data in classification task. We are doing oversampling on minority class to make the prediction better.

# Exploratory Analysis

1. Distance by Vehicle mode with respect to countries:

Mostly, Hybrid mode vehicles are used based on distance covered. For an example, country 2 covered distance 250000km in hybrid mode and in electric mode country 4 covered distance of 35000km. none of the countries switched to diesel mode for longer distances.

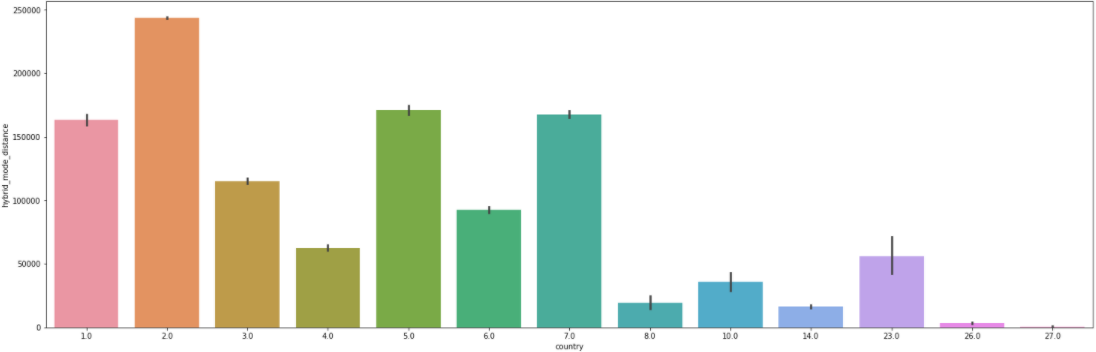


Fig 12: Hybrid\_mode\_distance

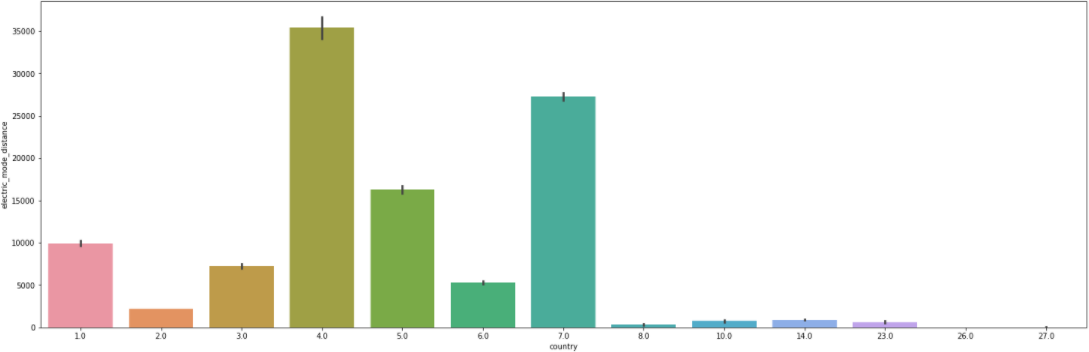


Fig 13: Electric\_mode\_distance

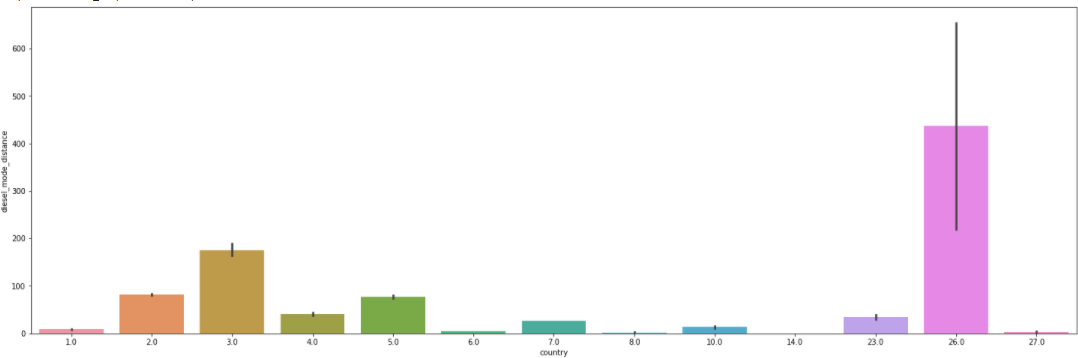


Fig 14: Diesel\_mode\_distance

1. Visualization of Engine type and Battery supplier relationship.

There are two engine types have in this dataset. Engine type 1 used batteries supplied by the supplier 1,2 and 4 whereas engine type 2 used batteries supplied by supplier 2.Before cleaning the data, we could see the engine type 3 & engine type 4 are exits, however, after cleaning, we could see, battery supplier 1,2& 4 data are available for engine 1 & 2.

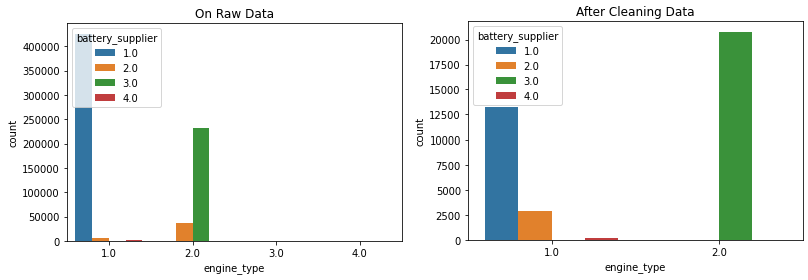


Fig 15: Engine Type

1. Similarly, In the battery supplier on raw data, battery generations are exits for all supplier, however, the counts are more for battery generation 1 and 3 in supplier 1 and 3 respectively. On other hand, after cleaning the data, the counts for battery generation 1 is decreased while other are nearly same counts before and after cleaning the data.

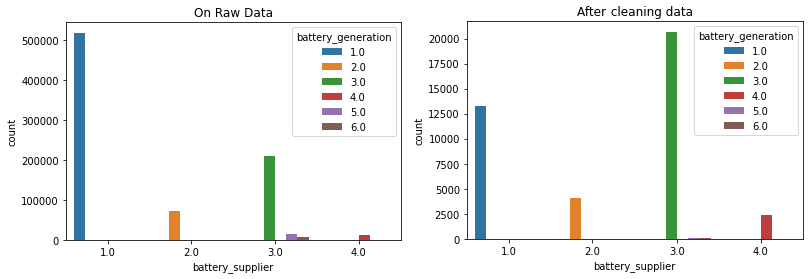


Fig 16: Buttery Supplier

1. The below graph shows the relation between Electric mode distance and country wise for ‘Battery\_Supplier’,‘Battery\_Generation’, ‘Battery\_Version’, ‘Engine\_Type’, ‘’Emission\_ Level’.

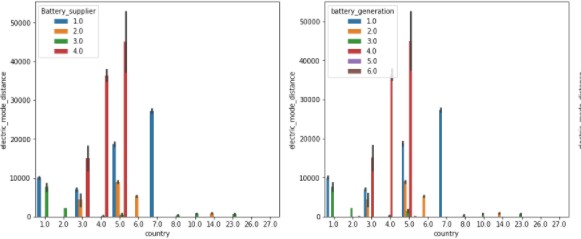


Fig 17: Country

In the first graph shows that the battery supplier 4 for country 5 is better than the rest of the other suppliers based on the distance covered by electric mode vehicles. In Electric mode distance, mostly battery generation 4 is used which is supplied by supplier 4 in countries 3, 4 & 5.

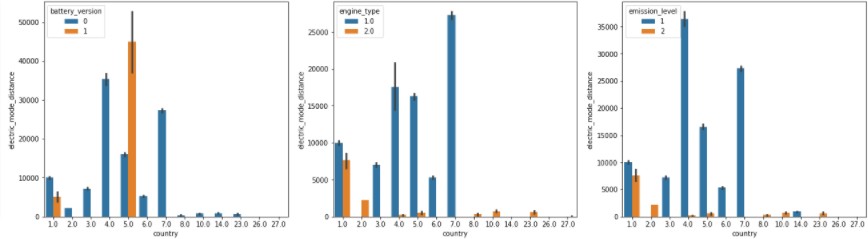


Fig 18: Relations between Country and Electric mode distance for Battery\_Version, Engine\_Type and Emission\_Level

In the above graph ‘battery version’, which gives the information of number of times battery changed, so the maximum number of battery changed was in country 5 and the engine type 1 is used in most of the countries and maximum for country 7, moreover, the emission level‘1’is majorly use in all the countries and highest for country 4.

# Machine Learning

## Supervised & Un -supervised Machine Learning

In Supervised Machine learning, we have used Classification algorithms i.e. Logistic Regression, Support Vector Classifier and in Regression algorithms, have used Linear Regression and Support Vector Regression. While, K- means clustering method is used in Un-supervised machine learning in the dataset for predicting the problem statements. The below methods are performed for data preprocessing, feature selection, data modeling and for finding the best classifier:

* Splitting the data into Training and Testing
* Used PCA
* SMOTE for oversampling
* Training and Testing the model
* Evaluation of Model

## Supervised Task

Problem Statement

1. Predicting the Operation time using Regression models
   * Finding correlation on all features with respect to ‘Operation time’ and applied threshold of greater than 0.25 and less than -0.25 .

* **Result without PCA ( on without Normalized dataset) :** We have checked by performing below algorithms with applying PCA on without normalized data.

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | **Model** | **MSE** | **R2** |
| 1 | Linear Regression | 3952533.01 | 0.89 |
| 2 | Support vector regression | 34138721.73 | 0.03 |

* **Result with PCA ( on without Normalized dataset) :** We have checked by performing below algorithms with applying PCA on without normalized data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | PCA | Linear Regression | | Support Vector Regression | |
| MSE | R2 | MSE | R2 |
| 1 | 2 | 9641150.33 | 0.73 | 34004617.35 | 0.04 |
|
| 2 | 5 | 8802133.71 | 0.75 | 34027684.80 | 0.04 |
| 3 | 9 | 5153919.61 | 0.85 | 34411262.73 | 0.02 |
| 4 | 13 | 5066738.60 | 0.86 | 34404611.06 | 0.02 |

* **Result without PCA:** We have checked by performing below algorithms without applying PCA and found that Support vector regression performed better than Linear Regression.

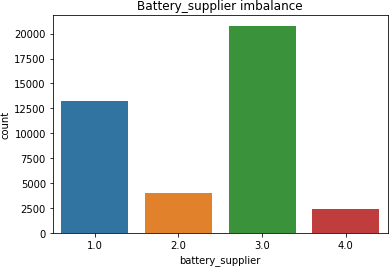
|  |  |  |  |
| --- | --- | --- | --- |
| S.No | **Model** | **MSE** | **R2** |
| 1 | Linear Regression | 0.01 | 0.83 |
| 2 | Support vector regression | 0.01 | 0.90 |

* **Result with PCA** : We have checked by performing below algorithms with different PCA values and found that Support vector regression gives better result than Linear Regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | PCA | Linear Regression | | Support Vector Regression | |
| MSE | R2 | MSE | R2 |
| 1 | 2 | 0.01 | 0.72 | 0.01 | 0.8 |
| 2 | 5 | 0.01 | 0.8 | 0.01 | 0.88 |
| 3 | 9 | 0.01 | 0.81 | 0.01 | 0.9 |
| 4 | 13 | 0.01 | 0.83 | 0.01 | 0.9 |

1. Prediction of multiclass classification feature ‘Battery Supplier’.

* Finding correlation on all features with respect to ‘Battery Supplier’ and applied threshold of greater than 0.25 and less than -0.25 .
* Applied SMOTE on Battery Supplier for balancing the data.
* Before SMOTE (Imbalanced)



* After SMOTE (Balanced)



* **Result without PCA and SMOTE**: After performing the below algorithms in this problem statement, found that Support Vector Machine has more accuracy than the Logistic Regression without PCA and Class Imbalance data.

|  |  |  |
| --- | --- | --- |
| S.no | **Model** | **Accuracy** |
| 1 | Logistic Regression | 90.52 % |
| 2 | Support Vector Machine | 95.31% |

* **Result with PCA and SMOTE** : After performing the below algorithms, found that Support Vector Machine has more accuracy than Logistic Regression with PCA and SMOTE.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **PCA** | **Logistic Regression Accuracy** | **Support Vector Machine Accuracy** |
| 1 | 2 | 63.66% | 89.12% |
| 2 | 5 | 66.11% | 89.86% |
| 3 | 10 | 71.0% | 92.22% |

## Unsupervised Task

Problem statement

1. Anomaly detection of vehicle Operation time with respect to vehicle operation distance using K means Cluster.
2. **Anomaly detection of State of Health of Battery and evaluation with different model on prediction of battery health.**

* In unsupervised algorithm, we have used the same cleaned dataset and applied different unsupervised clustering method i.e. K means clustering, HDBSCAN, Isolation forest , for finding outliers on State of Health of Battery and then used for prediction of battery health and checked the error before and after anomaly.

It is very important to identify the anomalies and solved it because it is one of the serious problems in machine learning as it can cause over fitting of the data and predict bad results.

* We have used PCA for dimensional reduction to identify the anomalies in the data. We have reduced the dimension from 76 to 8 and then plotted the data from which, got the maximum variance, also from the plot got the clusters according to the arrangement of the data and there are some data which easily detect as outliers.

**Isolation Forest for outlier detection**

* Isolation forest is a unsupervised, tree based algorithm, It is built on the theory of decision tree and random forest, and it’s a part of ensemble model class, available in sklearn.
* It is one of the fastest as well as consume less memory than the other anomaly algorithms.
* Isolation forest will train on the dataset and predict the data points into anomalies by marking -1 and 1. Also, we will find the scores, which indicates, lower values lesser than 0 observes anomalies and on other hand the values above 0 is considered as normal.. In the below fig, we could see that the negative scores indicates anomaly and the positive are normal. After checking the scores and anomaly for all dataset, it will print the predicated anomalies.
* we have used the same cleaned dataset and applied isolation forest algorithm.

In the below fig, we could see that the negative scores indicates anomaly and the positive are normal.

HDBSCAN Clustering for Hierarchical Density-Based Spatial Clustering of Applications with Noise.

* This clustering method is also one of the best method, It detect arbitrarily shaped cluster. It arranges the data points into hierarchies of cluster within clusters, It uses density of neighboring points to construct clusters and allowing noise points to be identified and excluded from clusters.
* It gives more core clusters based on the value of hyper parameters. We have used this method because it gives the expected clustering, however the other method such as K means performs poorly and fails to group the data into clusters and also we don’t have ground truth to analyze the clusters so in that case HDBSCAN performs best.
* After finding the cluster using HDBSCAN, we use cluster anomaly score to find the anomaly. More the values, the chances of anomaly is higher, on other hand, lesser the score the probability of being an anomaly is less.
* The below plotted graph shows the distribution of score for anomaly.

**Prediction of Battery Health after removal of anomaly detection using HDBSCAN**

**PCA and Feature Selection :**

K means Clustering

* We have used three clusters on data and the value of K(cluster) was found by applying Elbow method which is shown below(fig 19). After the implementation, the data was grouped(clustered) in three colors i.e. Green, yellow and Purple (refer fig 20).

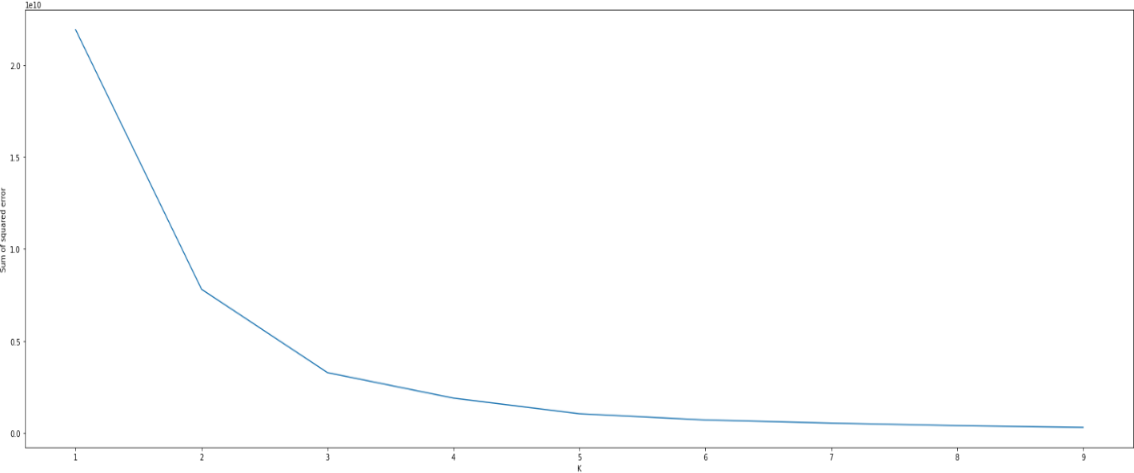


Figure 19 : Elbow curve

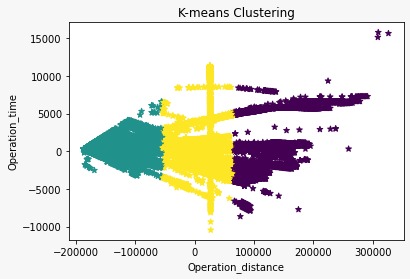


Figure 20 : K means clustering (K=3)

After that we found the centers of clusters and by using Euclidean distance for calculating the distances between each data point to the centers of its clusters. The center of cluster is represented as ‘Blue dot’ (refer fig 21,22,23).

For each cluster, the outliers are counted as data points stay out the threshold value, above the threshold values, the data points are counting as Outliers. We have used three thresholds values (50 : 80 :99) for finding outliers. The outliers are shown as “Red color” in below figures (fig 21,22,23).

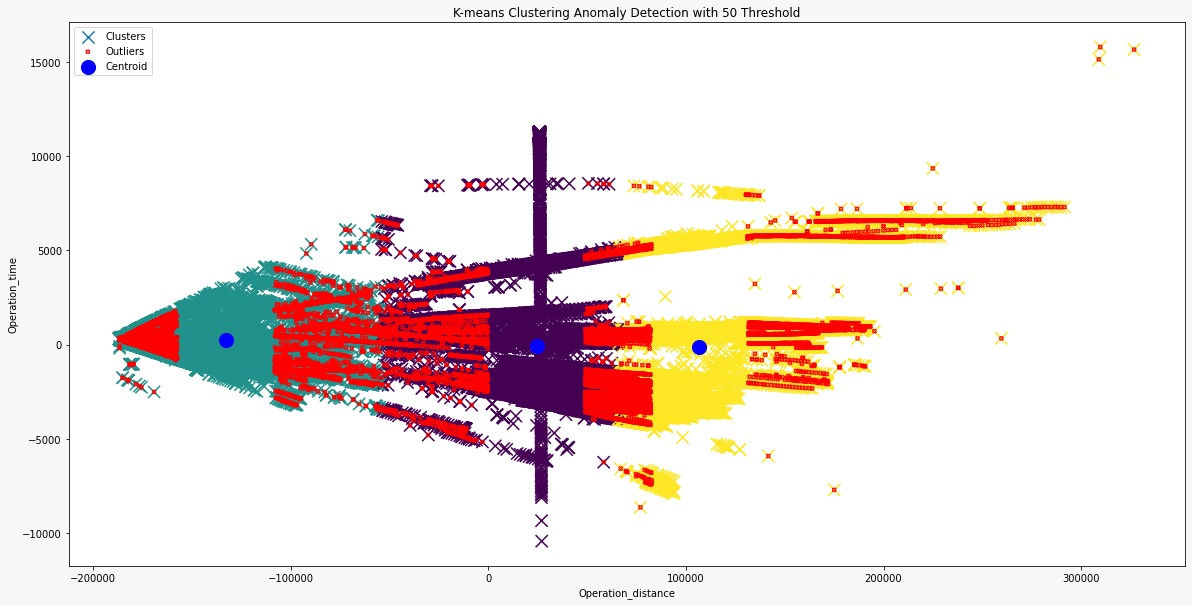


Figure 21: K means Clustering with 50 threshold

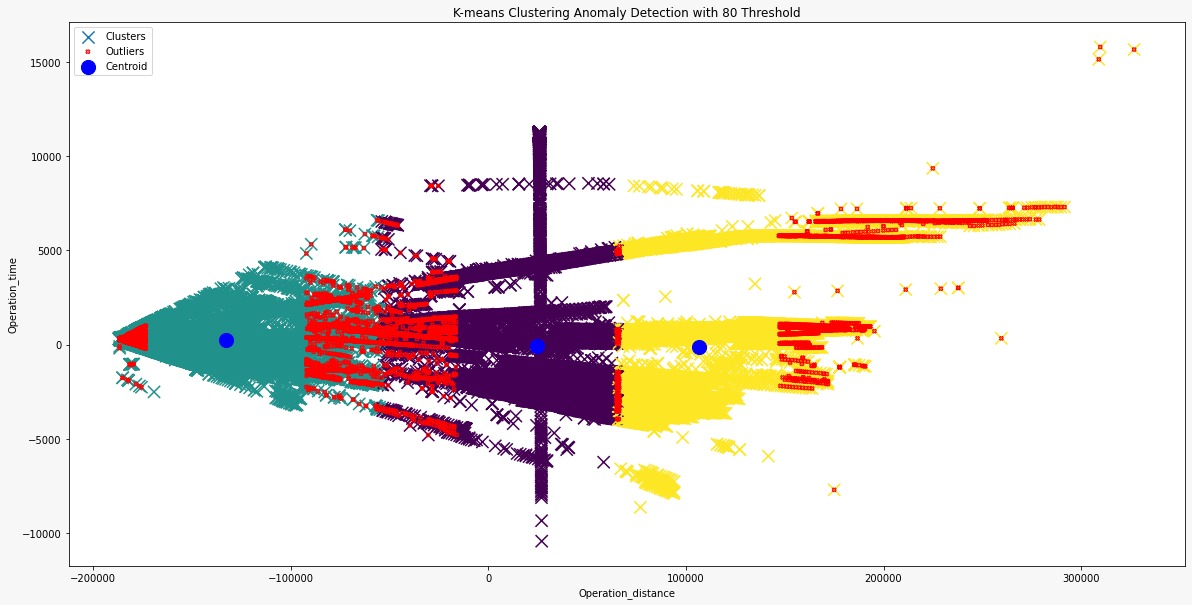


Figure 22: K means Clustering with 80 threshold

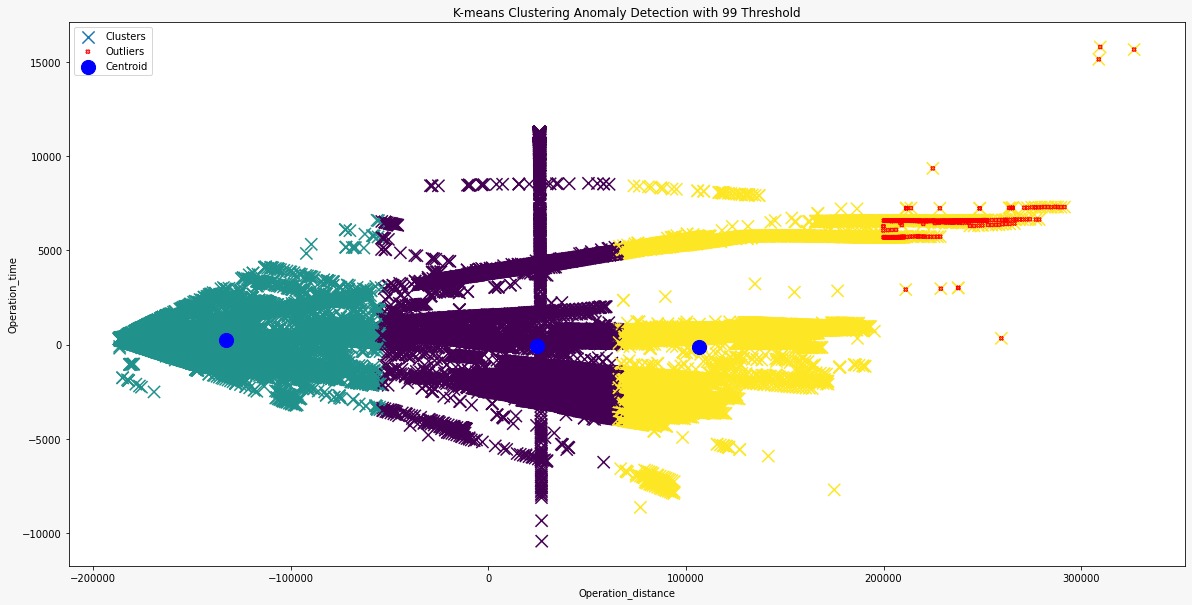


Figure 23: K means Clustering with 99 threshold

1. Anomaly detection of vehicle Operation time with respect to vehicle operation distance using Isolation Forest.

* Isolation forest is a unsupervised, tree based algorithm, It is built on the theory of decision tree and random forest, and it’s a part of ensemble model class, available in sklearn.
* It is one of the fastest as well as consume less memory than the other anomaly algorithms.
* we have used the same cleaned dataset and applied isolation forest algorithm for finding anomaly on vehicle operation time and vehicle operation distance.
* Also, we have found the scores, which indicates, lower values observes more anomalies. In the below fig, we could see that the negative scores indicates anomaly and the positive are normal.
* After checking the scores and anomaly for all dataset, it will print the predicated anomalies.

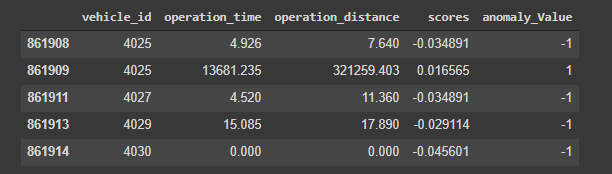


Figure 24: Score and predicated Anomalies

The below diagram shows the anomalies in red color and Normal data points in black color.

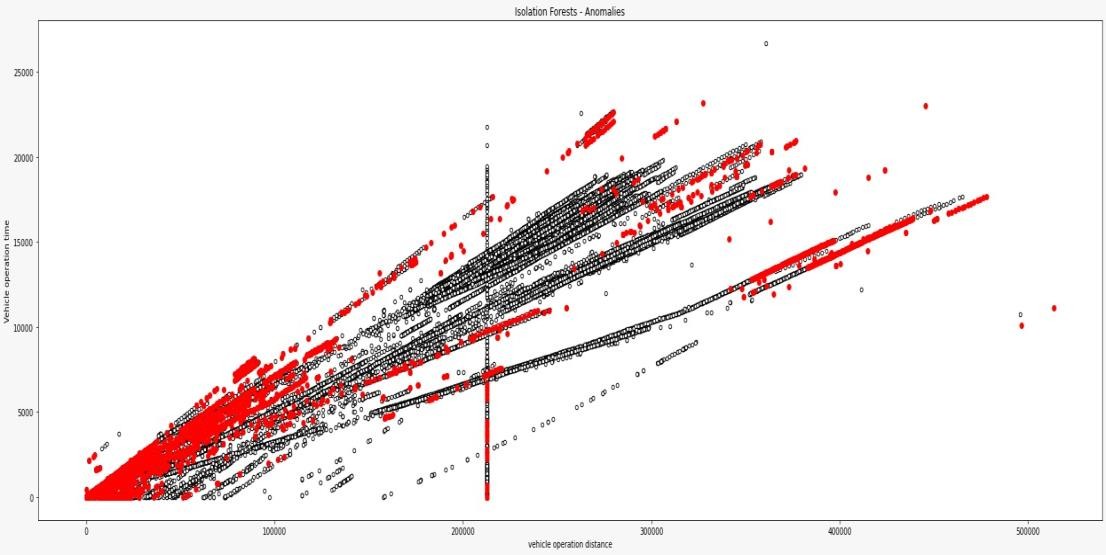


Figure 25: Scatter plot of Isolation forest Anomalies.

# Conclusion

In this project, we have analyzed the dataset, found the problems and solved using data mining and machine learning techniques. In the supervised machine learning algorithm, for the first problem, which is related to prediction of Operation time. In this problem Support Vector Regression model performed better than the other model with and without PCA on normalized and without normalized dataset .

In Second problem, which is related to prediction of multiclass classification feature ‘Battery Supplier’, Support Vector Machine gave better accuracy than Logistic regression with and without PCA and SMOTE method.

In un-supervised machine learning problem which is Anomaly detection of Vehicle Operation time with respect to vehicle operation distance Type, we have used K-means clustering algorithm for clustering the data points and finding the outlier. Also, we have checked the anomaly using Isolation Forest and find the outliers between vehicle operation distance and vehicle operation time.

# Learning from the project

It was good learning while doing this project, we have learnt many concepts while solving the problems and way to handle challenges. The interesting part was data preparation, while data cleaning and managing to handle missing values, balancing the data, feature extractions were challenging as well as interesting to find the way of handling these challenges.

We would like to share in brief one of the challenging part, moreover, we have detailed those in the methodology section:

Since the dataset had lots of incorrect and missing values, there were many duplicate samples in raw data. We have removed those duplicate values after understanding the features and relevant data. Data preprocessing part was most challenging and learning we had found during the project implementation.

# References:

1. <https://www.kaggle.com/c/data-mining-project>
2. <https://towardsdatascience.com/speed-up-your-algorithms-part-2-numba-293e554c5cc1>
3. <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02231/full>
4. <https://www.guru99.com/data-mining-tutorial.html>
5. [https://medium.datadriveninvestor.com/outlier-detection-with-k-means-clustering-in-python-](https://medium.datadriveninvestor.com/outlier-detection-with-k-means-clustering-in-python-ee3ac1826fb0) [ee3ac1826fb0](https://medium.datadriveninvestor.com/outlier-detection-with-k-means-clustering-in-python-ee3ac1826fb0)
6. <https://heartbeat.fritz.ai/isolation-forest-algorithm-for-anomaly-detection-2a4abd347a5>
7. [https://medium.com/learningdatascience/anomaly-detection-techniques-in-python-](https://medium.com/learningdatascience/anomaly-detection-techniques-in-python-50f650c75aaf) [50f650c75aaf](https://medium.com/learningdatascience/anomaly-detection-techniques-in-python-50f650c75aaf)
8. <https://petuum.medium.com/scalable-clustering-for-exploratory-data-analysis-60b27ea0fb06>