**METHODOLOGY**

**Models Used**

The three clustering techniques being used are K-means, K-modes and K-prototype. K-means is focusing more on the numerical data, while the K-modes techniques provides some insight into the categorical data. We also have the K-prototype technique which joins both the k-means and k-modes in order to deal with data that includes for numerical and categorical data.

The K-means algorithm which is unsupervised takes on a particular number of data points and assign them into *k* number of clusters. The data points are allotted to a scatter plot and there will be *k* number of clusters set. Which ultimately gives K number of centroids within the plot.

A computation will be achieved in order to find the data points which are closest to the centroids using the Euclidean distance.

Within the K-prototype algorithm a dissimilarity solution is used which is a way to measure how closely related the data samples are to each other. Which will consider both the numeric and categorical attributes. Let assume *sn* is the dissimilarity measure for numerical data types which is defined by the squared Euclidean distance and *sc* is the dissimilarity measure for all categorical data types defined as the number of mismatches of categories between two objects.

We consider the dissimilarity measure for the two objects as *sn +γ sc*, where *γ* acts as the weight to ensure there is a balance between the two to avoid any imbalance between either type of data. The k-prototypes algorithm has its similarities to the k-means algorithm with the only difference is that we include the k-modes approach for updating the categorical data values of cluster prototypes.

**Datasets**

The Data will be retrieved from the Newham parking database. The data will also be anonymised by ensuring registrations are not included and anything that could identify a particular individual or vehicle are redacted. Over 300,000 parking penalties data was generated.

The query used are all on-street parking penalties issued from 01/01/2019 to 31/12/2021. It also includes vehicle makes and colour, issue date, issue time, contravention code, location and civil enforcement officer (CEO) number. The data will be entered into jupyter notebook which is an application which is very well suited for creating and manipulating all types of data. All data will be entered as a CSV file.

**Pre-processing**

The first step would be to import all necessary libraries such as matplotlib, pandas and K-means. These tools are particularly useful for analysing datasets in python programming. Matplotlib is often used to generate visualisations and was particularly useful in displaying the scatter diagrams for K-means and K-modes. The next step was to redact all penalty case numbers as well as all vehicle registration marks to avoid identifying any individual and to protect the identities of all drivers affected.

As part of the data cleaning, the data was checked for any futile data and the same was done in order to check for duplicate. There were also a number of test cases which were removed and any contravention code starting with a W were considered warning notices which were also removed from the dataset.

**Model Architecture**

For the data to be visualised using the K-means and K-modes models, a scatter diagram would need to be created to visualise each cluster. K-means was imported from the sklearn cluster, which is a machine learning python tool.