A title I want to on two lines

Contents

```
#rticles
# https://stackoverflow.com/questions/29532214/add-author-affiliation-in-r-markdown-beamer-presentation
# https://stackoverflow.com/questions/29389149/add-image-in-title-page-of-rmarkdown-pdf
```

1 Data Exploratory

Before any clustering can begin for this dataset, Figure 1 shows a boxplot illustrating the distribution of each datapoint representing the count of crime committed in each community with the mean plotted as a light dot. We can see that the variables Theft, Narcotics and Battery all have outlier observations representing the different communities with significantly high counts of those types of crime. To overcome this issue, one method of clustering that will be used is hierarchical clustering; this will allow such data to be grouped and easily interpreted when the dendrogram plot is illustrated.

Although hierarchical clustering will help us illustrate correlations more easily, it does not always provide the best solutions. Using K-means alongside hierarchical will allow us to compute tighter clusters between observations and accepts clusters of different shapes and sizes such as elliptical clusters.

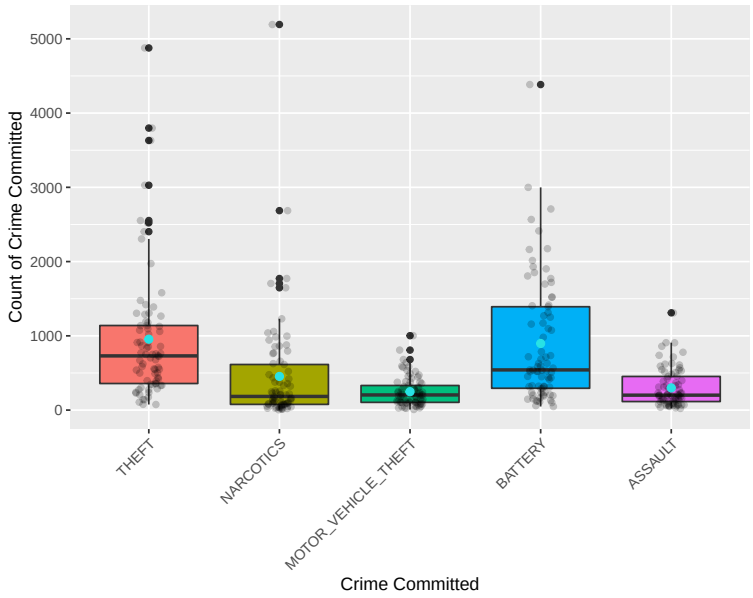
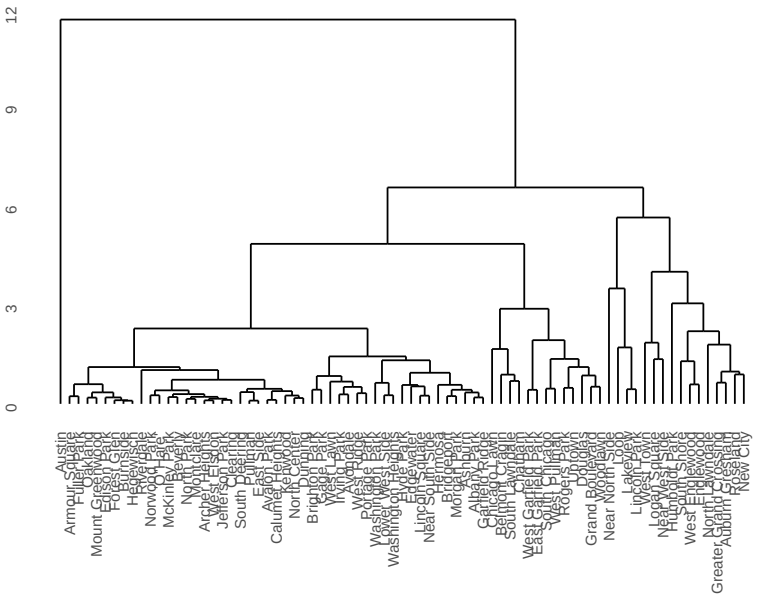


Figure 1: Boxplot showing the rate of selected crime committed over a number of communities. The blue dot illustrate the mean crime rate for each type of crime.

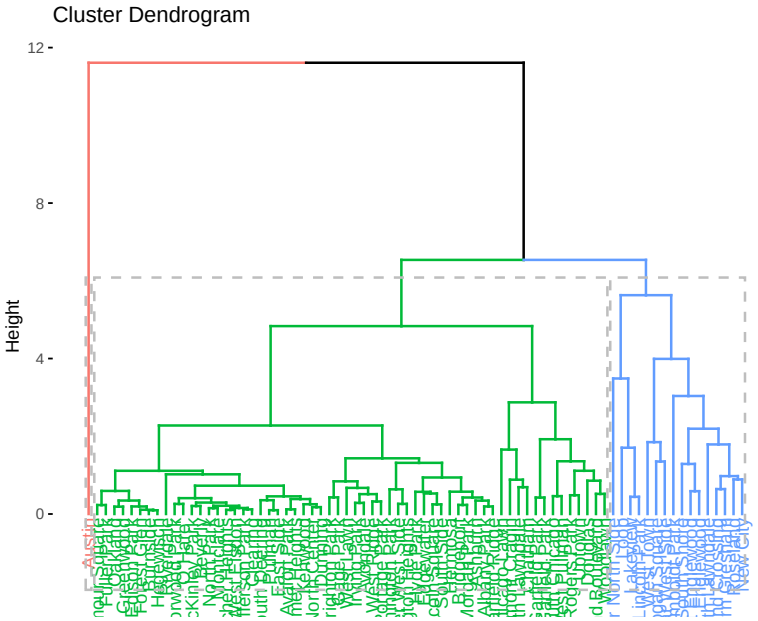
2 Hierarchical Cluster

Hierarchical clustering is also a kind of clustering algorithm, which creates a hierarchical nested clustering tree by calculating the similarity between data points of different categories. In a cluster tree, the original data points of different categories are the lowest level of the tree, and the top level of the tree is the root node of a cluster. There are two ways to create cluster trees: bottom-up aggregation and top-down splitting.

Here we are using cluster from the bottom up. Height in the figure refers to the Euclidean distance between two clusters. We find the two clusters with the shortest distance each time, and then condense them into a large cluster until they all condense into one cluster.



The tree diagram plots each cluster and distance. We can use the tree diagram to find the clustering of any element we choose. In the tree below, it is easy to see where the first cluster (red), the second cluster (green), and the third cluster (blue) begin.



3 K-Means

K-means algorithm is an unsupervised machine learning technique used to partition observations into subgroups or clusters. In theory, we want observations in the same cluster to be similar (i.e., high intra-class similarity) while observations in different clusters to be as dissimilar as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster. Since there is no a response variable, the objective here is to try to find the number of cluster that minimises total distance.

3.1 Optimal Clusters

Since there was no prior knowledge on how many clusters would be our optimal numbers. Elbow plot was applied and main elbow occured at 4 suggesting 4 clusters.

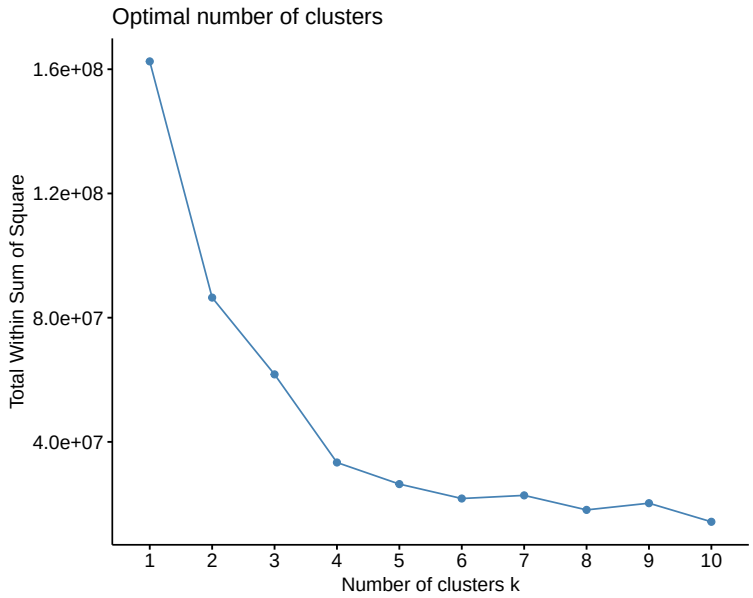


Figure 2: Elbow plot shows total within sum of square from 1 to 10 clusters

3.2 Visualising K-Means Clusters

It is a good idea to plot the cluster results because these can be used to assess the choice of the number of clusters. Now, let us visualise the data in a scatter plot with coloring each data point according to its cluster assignment.

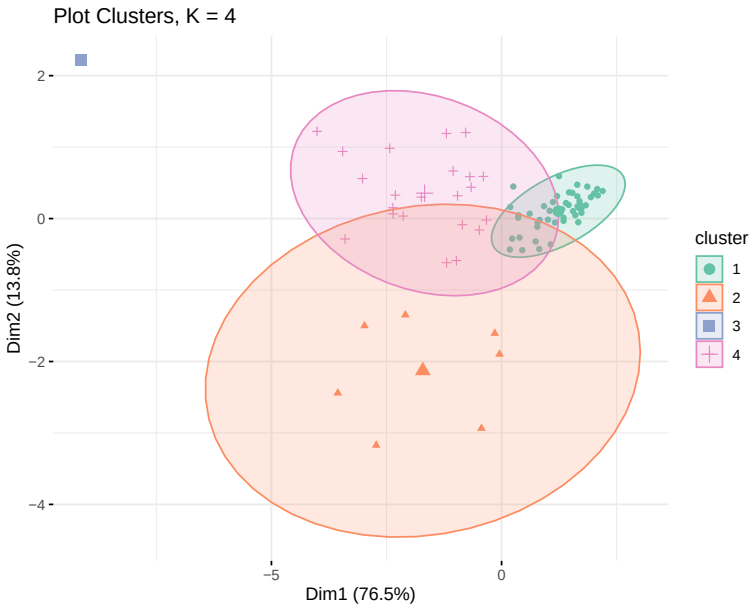


Figure 3: Visualising clusters, K=4