The Relationship between Crime Types in the City of Chicago.

Introduction

The City of Chicago contains 77 communities and data is available from the Chicago Police Department (CPD) from 2002 to 2015 on all reported incidents of crime at community level. This report will be looking at a subset of this data from 2002 and specifically the crimes of theft, narcotics, motor vehicle theft, battery and assault. These crimes will be used to investigate the question of interest: which communities are similar in terms of their levels of crime? Answering this question allows us to give more information to the Chicago Police Department for predicting what kind of crimes would be in a community by the rates of other crimes. Also, levels of crime have an impact on house prices. By clustering communities by similar crime rates we will be able to see what areas will be similar in terms of house prices. The question will be answered through clustering techniques, and the clustering techniques will be compared to see what one is most effective in answering the question.

Exploratory Analysis

Before any clustering could begin for the data set, exploratory analysis was conducted. Figure 1 shows a box plot illustrating the distribution of each data point, where each data point represents the ratio of number of committed crimes over the population of each community with the mean plotted as a blue dot. The variables theft, narcotics and battery all have outlying observations, illustrating that certain communities have a significantly high count of these types of crime. To overcome this issue, one method of clustering that can be used is hierarchical clustering. This 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.

From looking at Figure 2, all pairs of variables are plotted as scatterplots above and below the diagonal. We can see very strong correlations between the pairs Battery & Assault and moderately strong correlations between all the types of crime except Narcotics. This will help us understand that if a community had a high number of Battery then, that same community would likely contain high numbers of Assault.

Formal Analysis

Clustering 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. have a high intra-class similarity) while observations in different clusters to be as dissimilar as possible (i.e. have a low inter-class similarity). The methods of clustering that will be used to answer the question of interest are Hierarchical clustering and K Means clustering.

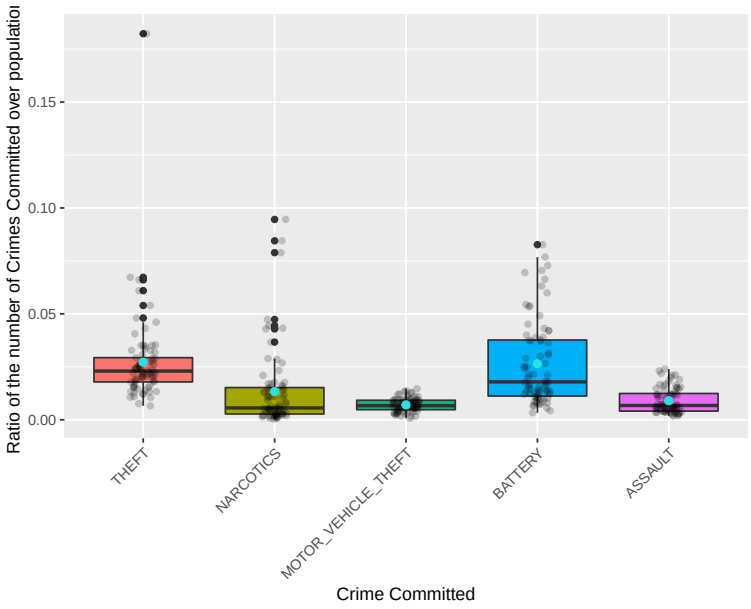


Figure 1: Boxplot showing the rate of selected crime committed over a number of communities. The blue dot illustrate the mean crime rate for each type of crime.

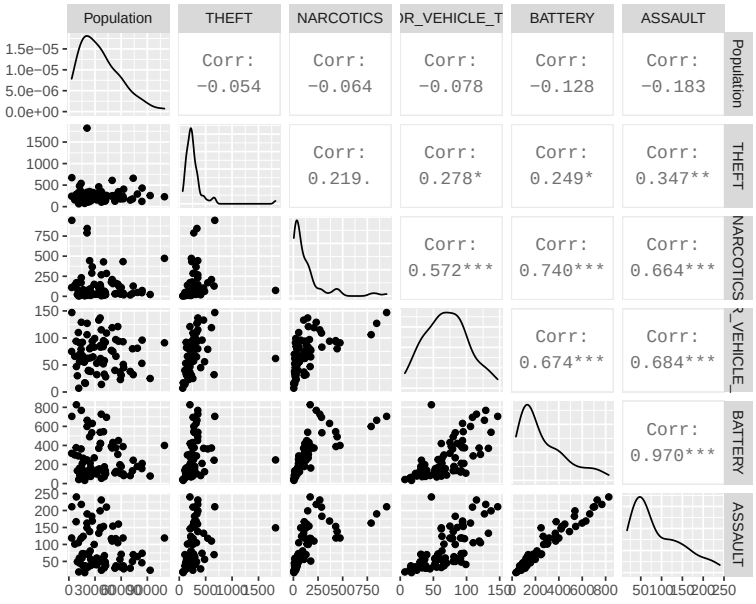
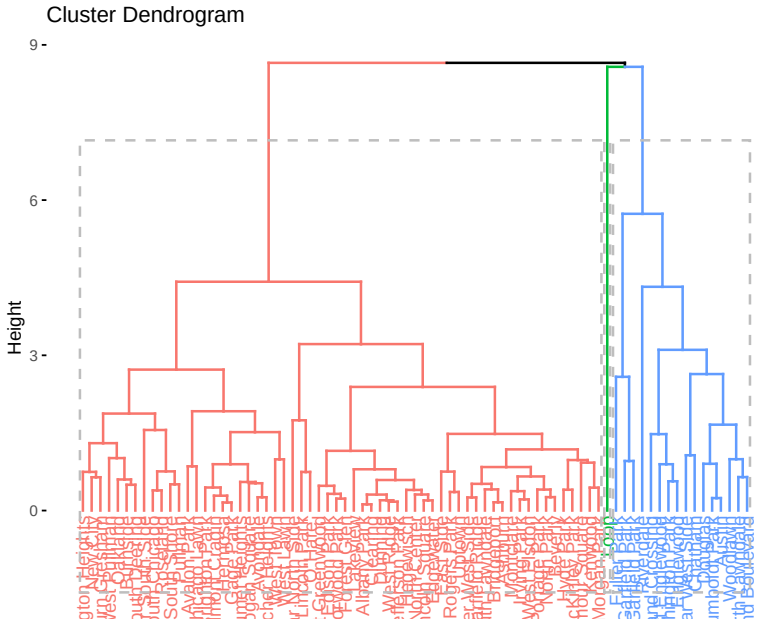


Figure 2: Pairs showing the correlation between the types of crime within the communities in Chicago

Hierarchical Clustering

Hierarchical clustering 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. Divisive clustering and agglomerative clustering. Agglomerative clustering is used for this analysis with complete linkage. This begins with every case being a cluster by itself. At each step, similar clusters are merged until all observations are in one cluster.

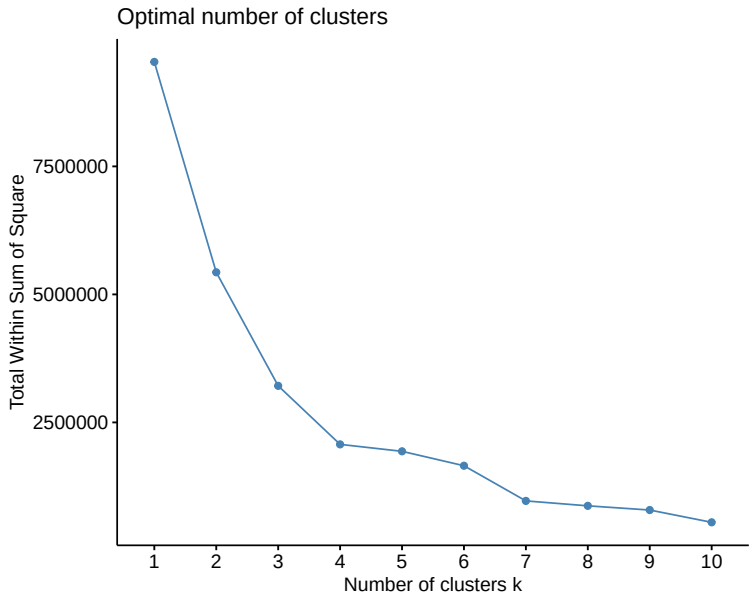
Hierarchical clustering can be visualised through a dendrogram which is a visualisation of a distance matrix. The main use of a dendrogram is to work out the best way to allocate objects to clusters. Figure... plots each cluster and distance. Looking at this figure, it is clear that there may be three clusters in the data which are coloured to represent each cluster. It is easy to see where the first cluster (red), the second cluster (green), and the third cluster (blue) begin.



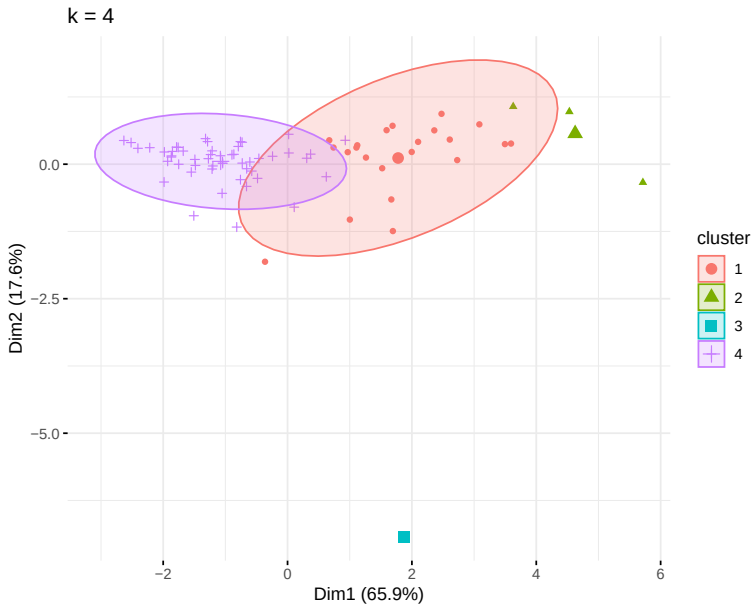
K-Means

The goal of K-means is to find k groups in the data. K-means attempts to find the assignment of observations to a fixed number of clusters K, that minimises the sum over all clusters of the sum of squares within clusters. 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. New centroids are then computed as the average of all observations in a cluster and then each observation is assigned to its closest centroid, and this is repeated until the observations are not reassigned or the maximum number of iterations is reached. Since there is no a response variable, the objective here is to try to find the number of clusters that minimises the total distance.

Since there was no prior knowledge on how many clusters would be our optimal numbers, an elbow plot was used , shown in Figure... to decide on the number 'k' of clusters to be used. The main elbow occurred at 4 suggesting k = 4 clusters would be suitable.



A plot of the cluster results was constructed to assess the choice of the number of clusters. Figure ... shows a visualization of the k-means clustering.



Discussion of Results

Silhouette plots were created to compare the two different methods of clustering, Hierarchical and K-means. The silhouette value measures how similar an object is to its own cluster by comparing to the other clusters. It ranges from -1 to 1, where a high value indicates an object has matched well to its own cluster and not to other clusters and thus means a "good" clustering fit. The width of the silhouette is defined as:

$$s_i = \frac{(b_i - a_i)}{\max\{a_i, b_i\}}$$

Table 1: Average Crime Numbers for each cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Population Average	40468.16909	15085.6267	20839	35690.28000
Average Theft count	333.36364	428.0000	1823	206.29412
Average Narcotics count	233.40909	860.0000	74	48.27451
Average Vehicle Theft Count	91.86364	126.6667	62	57.80392
Average Battery Count	489.04545	656.0000	248	145.50980
Average Assault Count	155.50000	188.0000	149	54.47059

where a_i is the average distance between observation i and the other observations in i's cluster, b_i is the minimum average distance between observation i and the observations in other clusters.

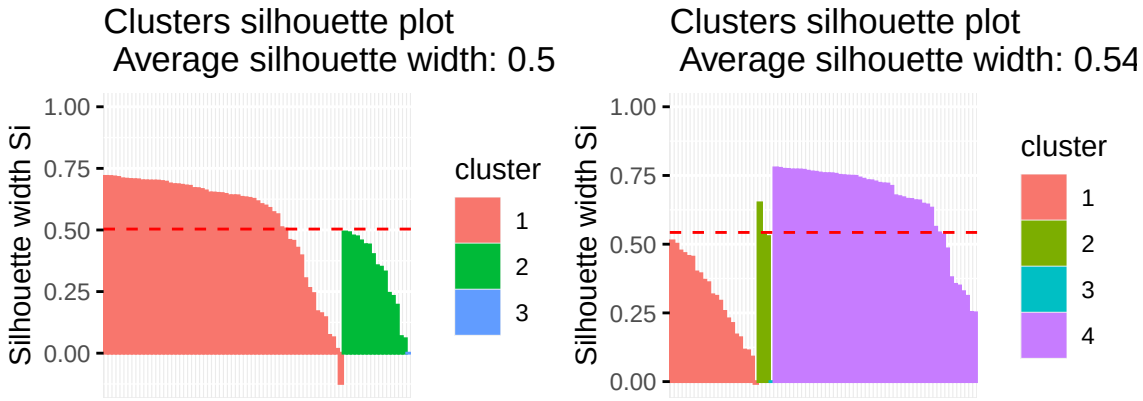


Figure 3: Silhouette Plot for Hierarchical and K-Means Clustering Techniques

From these figures and the average silhouette width, it can be said that for this dataset, the K-Means Clustering method is favorable over the Hierarchical Clustering method. The average width for Hierarchical was 0.51 and the K-Means value was 0.57. The difference in width isn't huge and so either could be said to be a good technique, however the one that will be used for further analysis will be the K-Means technique.

Using the K-Means clustering method, a results table above shows the average count data for each of the clusters as well as the average population.