"The Relationship between Crime Types in the City of Chicago."

Ali Abdelazim
Yuqing Bu
Chloe Trendall
Benjamin Wiriyapong
Olivia Wise

Contents

Introduction

Data from 2002-2015 from the Chicago Police Department (CPD) is available on all reported incidents of crime at from the 77 communities in Chicago. 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? This will be answered through clustering. From this we will explore the relationship between different types of crime in Chicago. Clustering communities by levels of crime can help answer questions regarding 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 the best method in answering the question.

Each community within Chicago had a different population size so before any analysis was conducted the counts of crime were converted to rates of crime per 10,000 people by dividing the crime counts for each crime by the population size for the community, and multiplying this by 10,000. Crimes between communities could then be compared on a common scale.

Exploratory Analysis

Before any clustering was carried out exploratory analysis was conducted. Figure 1 shows a box plot of crime rates for each crime. Each data point represents a community of Chicago and the mean crime rate across all the communities is plotted as a blue dot. Battery appears to have the widest range of crime rates and the highest mean crime rate, whilst crime rates for motor vehicle theft and assault are closer to the mean across the communities. This shows that communities within Chicago tend to have similar crime rates for motor vehicle theft and assault, whilst crime rates for battery vary more. Theft appears to have outlying observations which shows that certain communities have a significantly high count of this crime type compared to other communuties in Chicago.

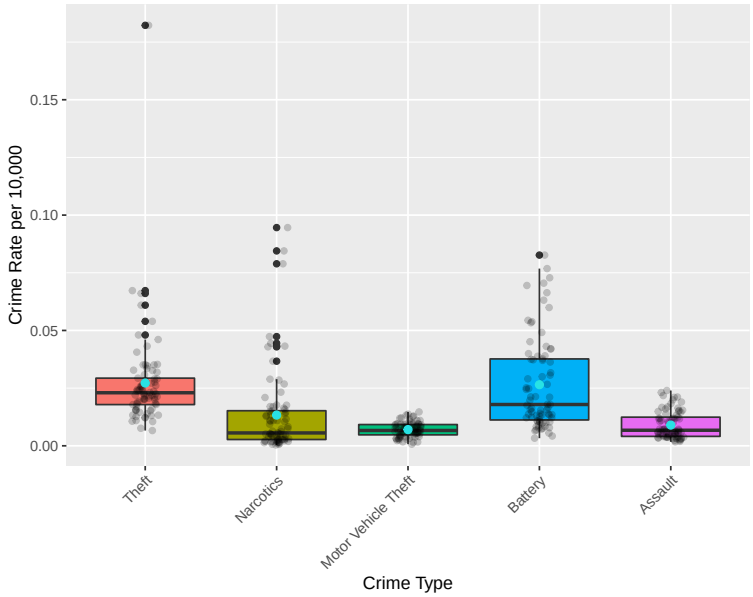

Figure 1: Boxplot of Crime Rate per 10,000 for Each Crime. The blue dot illustrate the mean crime rate for each type of crime.

A pairs plot was plotted to assess if crimes were correlated with each other, shown in Figure 2. There are

strong correlations between the crimes battery and assault, battery and narcotics, battery and motor vehicle theft and assault and motor vehicle theft. This helps us to understand the relationships between different crimes within the communities. For example, if a community had a high number of assaults then that same community would likely contain high numbers of battery related crimes.

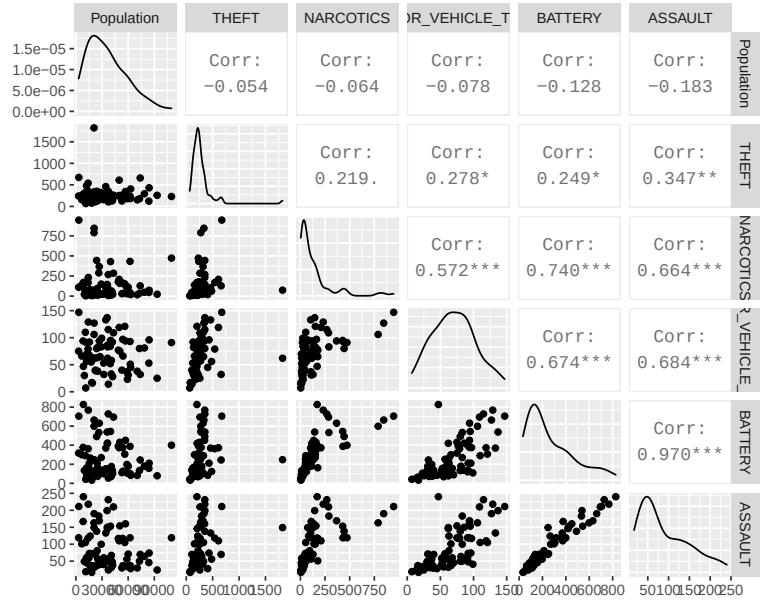

Figure 2: Pairs plot showing the correlation between the types of crime within the communities in Chicago

Formal Analysis

Clustering is an unsupervised machine learning technique used to partition observations into subgroups or clusters. By clustering, observations in the same groups are more similar to other observations in the same group (i.e. have a high intra-class similarity) and dissimilar to observations in other groups (i.e. have a low inter-class similarity). The group of similar data points is called a 'cluster'. The methods of clustering that will be used to answer the question of interest are hierarchical clustering and K-means clustering. These will be used to cluster the communities of the City of Chicago so communities in the same clusters will have similar crime rates to each other.

Hierarchical Clustering

Hierarchical clustering creates a hierarchical nested clustering tree by calculating the similarity between data points of different categories. Hierarchical clustering is divided into two types: Divisive and Agglomerative.

Agglomerative clustering is used for this analysis. This is where each data point begins as an individual cluster. Similar clusters are merged at each step until all observations are in one cluster. To decide which clusters should be combined a measure of dissimilarity between sets of observations is used, such as a measure of distance between pairs of observations. A linkage criterion is also used to specify the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Here, complete linkage is used. This is where two clusters are merged with the smallest maximum pairwise distance. The result of this method is a dendrogram.

A dendrogram is a visualisation of a distance matrix and a way to visualise hierarchical clustering. Figure 3 shows the resulting dendogram from agglomerative clustering using complete linkage on the data set. Looking at this figure it appears that there are 3 clusters in the data. Each colour represents a cluster. It is easy to

see where the first cluster (red), the second cluster (green), and the third cluster (blue) begin. It could be argued that 2 clusters would be appropriate as one community is in a cluster on its own.

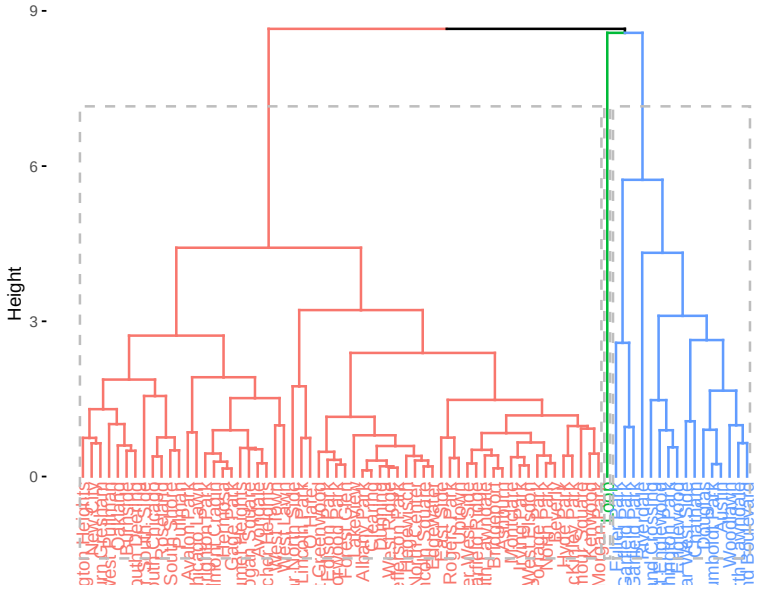

Figure 3: Dendogram of Complete Linkage

K-Means Clustering

The goal of k-means is to find k groups in the data. K-means clustering 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.

Firstly, k number of centroids are selected, then each observation is assigned to its closest centroid. New centroids are then computed as the average of all observations in a cluster and then each observation is assigned to its closest centroid. This is repeated until the observations are not reassigned or the maximum number of iterations is reached. So 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.

To determine the "optimal" number of clusters for k-means an elbow plot was used, shown in Figure 4. The main elbow occurred at 4 suggesting k = 4 clusters would be suitable for this data. This is the number of clusters used for the analysis.

Figure 5 shows a visualization of the k-means clustering to assess the choice of the number of clusters. It could be argued that 3 clusters could be appropriate for this data. From the plot it can be seen that there are very few observations in cluster 2, and these observations are close to the observations in cluster 1.

Discussion of Results

Comparing Hierarchical and K-Means Clustering Methods

To compare the clustering methods, silhouette plots were created, shown in Figure 5. 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:

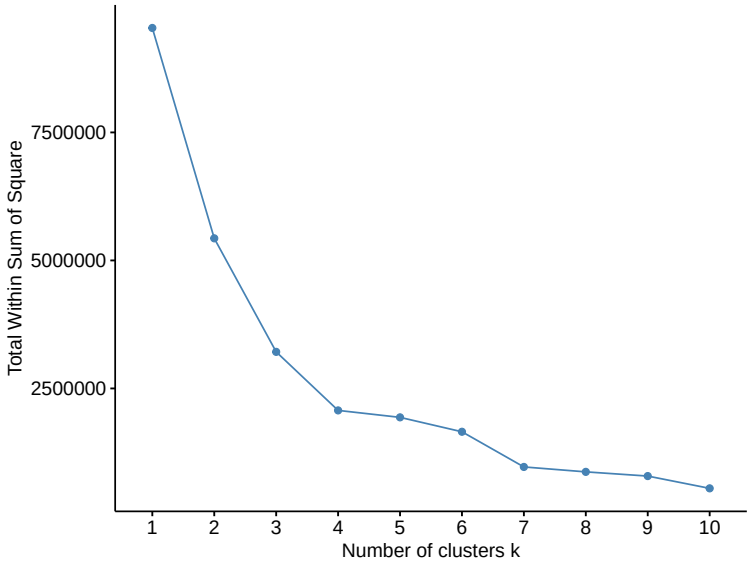

Figure 4: K-means elbow plot

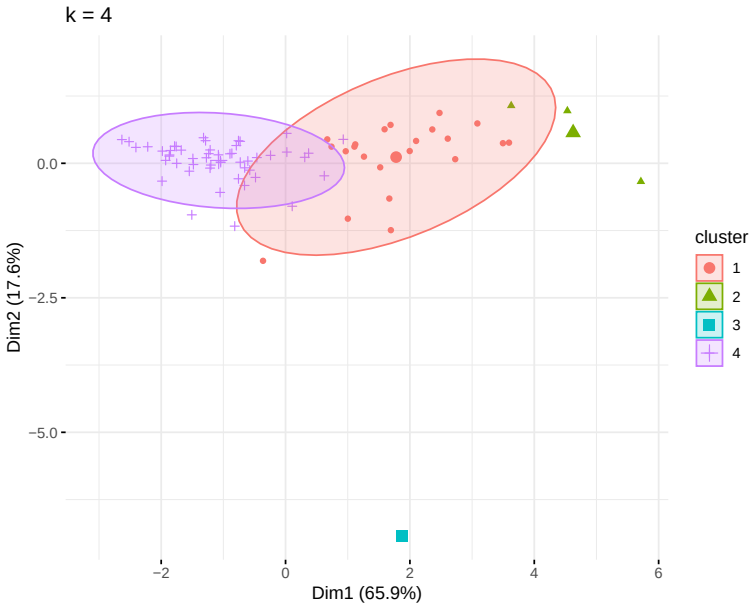

Figure 5: Visualisation of k-means clustering

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

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

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 6: Silhouette plot for hierarchical (left) and k-means (right) clustering

From Figure 5 and the average silhouette width it appears that for this data set that k-means clustering is favorable over the hierarchical clustering as it has a larger silhouette width. The average silhouette width for hierarchical clustering was 0.50 and the k-means value was 0.54. Also, the lack of negative Si values for k-means indicates that each observation is in the most suitable cluster given the clustering assignment. This is not true for hierarchical clustering, which suggests that some observations belong in alternative clusters under this clustering method. The difference in the silhouette widths isn't huge 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. In both methods of clustering the community Austin was the only community found by itself in one cluster as it didn't reflect the rest of the cluster trends.

Table 1 above shows the average count data for each of the clusters as well as the average population using k-means clustering. The crime count per 10,000 people was found first for each and then multiplied up to the average population for that cluster to illustrate the average expected number of reported criminal incidents for each cluster.

Limitations

Conclusion

In conclusion, two different clustering methods were applied to the Chicago Police Department data to try and capture the various relationships between crime types in the communities. Due to the variety of population sizes across the 77 communities, the number of crime incidents per 10,000 people was found for each community and crime type. This was to prevent any bias caused by difference in populations. The Hierarchical method found three distinct clusters with an average silhouette width of 0.5, while the K-Means method rose just above it with an average width of 0.54 and four distinct clusters. For the 3rd cluster in each method, Austin was the only community in that cluster due to it having a very different combination of crime counts as the others. After selecting K-Means to be the chosen clustering method for this data set, a table was made to help illustrate the general trends for each cluster with the average population of that cluster with the adjacent crime count for it. From this it is clear that there is no obvious relationships between all of them but that different combinations of all 5 of the crime types that were looked at is what clusters them together. In future, if there is a new community added to the data set, the Chicago Police Department might want to look at the number of crime incidents and fit it into one of the four clusters and make assumptions from there.

Appendix