University of Glasgow
School of Mathematics and Statistics

Investigating the City of Chicago Crimes Dataset

By David Quinlan

Contents

1 Introduction

An investigation was conducted to examine the crime levels within the city of Chicago. The Chicago crime dataset was provided by the Chicago Police Department (CPD) and noted all incidences of crime whereby the CPD responded to and fully completed a case report on such an incident. This dataset was recorded over a 13 year period between 2002 and 2015 and contains all crime incidences recorded within the 77 different communities present in the city of Chicago. Each incident was given a classification at the time of the preliminary investigation, whereby the classifications are those that are defined by the FBI's national incident based reporting system [1].

The Chicago crime dataset is comprised of 1078 observations with 36 variables in total. As part of this assignment a subset of the dataset was to be selected. For a randomly assigned year, I was to choose a subset of the dataset that only included three crime variables of my choice. I decided to choose the variables Theft, Homicide and Narcotics, while I was randomly assigned the year 2005. More details about the variables contained within my crime dataset are given below:

- **Year:** The selected year of crime to be examined, in this case 2005.
- **Community Name:** The names of the 77 neighbourhoods present in the city of Chicago.
- **Community Number:** A number from 1 to 77 representing a specific neighbourhood.
- **Population:** The size of population of each community for the year 2005.
- **Theft:** The theft variable is defined by the CPD as "the unlawful taking, carrying, leading or riding away of property from the possession or constructive possession of another person". In this case the theft variable represents the number of occurrences of such events in each community.
- **Homicide:** According to the CPD the definition of this classification is "The killing of one human being by another". This variable records a count of the number of homicide 1st and 2nd degree murders for each community.
- **Narcotics:** The narcotics crime classification represents all crimes whereby someone was caught producing, distributing and/or using controlled substances as well as the equipment required in the production of such substances. This variable represents the number of those who were caught committing such crimes within each of the 77 communities in the city of Chicago.

Therefore, my chosen crime dataset contains the counts of the occurrences of theft, drug related activities and homicides for the 77 different communities within Chicago for the year 2005.

The main aims of this investigation are:

- To conduct some initial exploratory analysis on the newly created crime dataset.
- To perform cluster analysis using a variety of techniques to identify a possible number of similar (and dissimilar) groups present within the dataset.
- To explain and discuss the findings of the analysis performed.

1.1 Initial Exploratory Analysis

After subsetting the original Chicago crime dataset into a smaller dataset, it is important to examine this new dataset to obtain a deeper understanding of the data. From this moment on the subsetted crime dataset will, for simplicity, be referred to as the crime dataset.

Boxplots were fitted to give a graphical representation of the three variables Theft, Homicide and Narcotics. These boxplots are shown below in Figure 1.

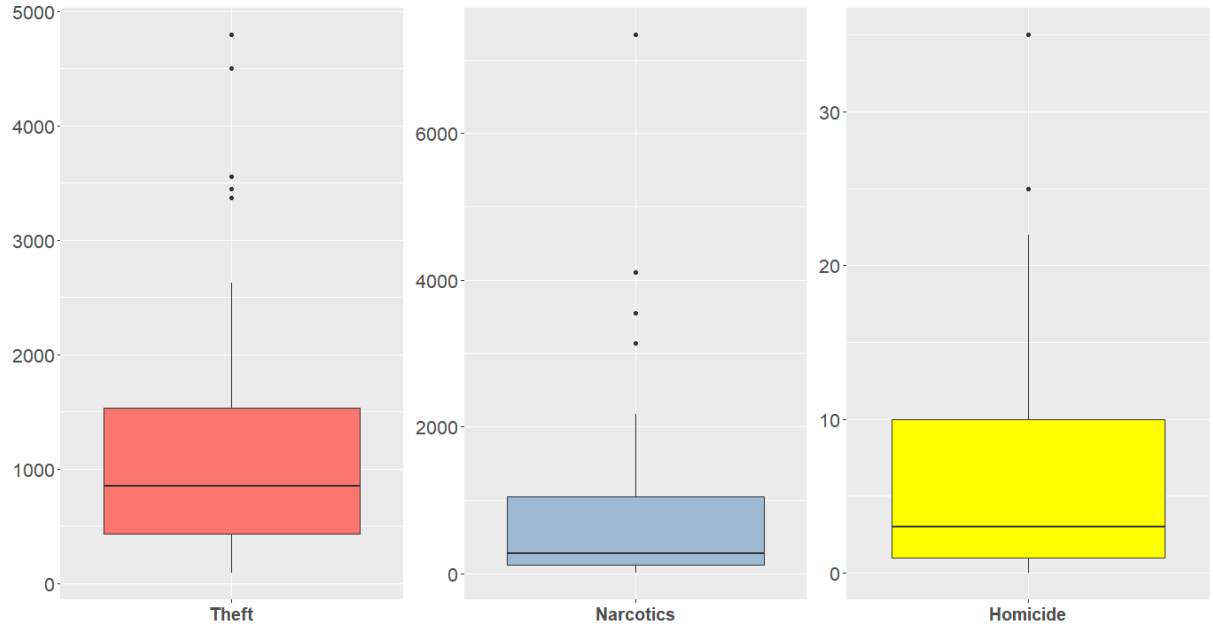

Figure 1: Boxplots of the variables Theft, Narcotics and Homicide contained within the crime dataset.

From the above boxplots one can see that the number of reported thefts ranges from the low hundreds to nearly 5000 for each of the different communities in Chicago. Similarly, the number of narcotic incidences ranges from 0 to approximately 7000 occurrences for the different regions present within the city of Chicago. Finally, the range for which homicides occurred within each of the 77 communities in Chicago was between 0 and 35. These boxplots also indicate possible outliers present within the dataset. In this case these outliers may also represent the areas where crime appears to be most frequently occurring.

Histograms were created to examine the frequency of occurrence for each of the three types of crime. These histograms are shown below in Figure 2.

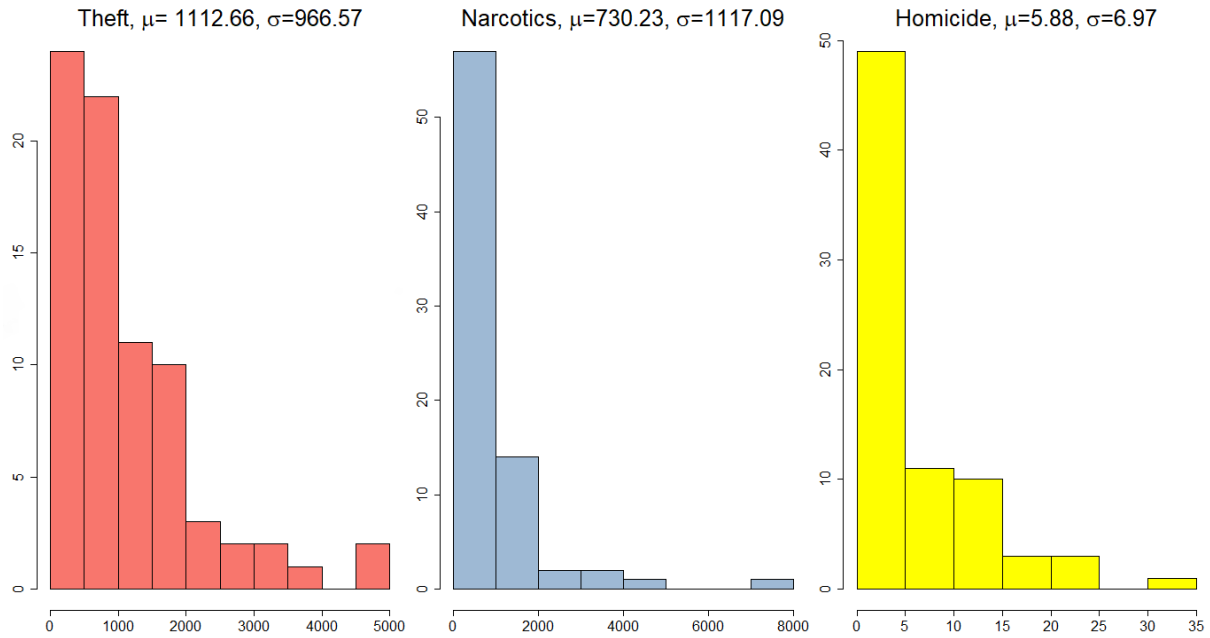

Figure 2: Histograms of the three variables Theft, Narcotics and Homicide contained within the crime dataset.

It is clear when examining the histograms that the number of occurrences of theft, homicide and drug related crimes is low for most of the communities. This is represented by the fact that each histogram is right skewed. The histograms also indicate how low occurrences of each crime occur most frequency with higher crime rates occurring less often. One can also tell from the histograms that the ranges on which the variables are measured are different from one another. The Theft and Narcotics variables being measured on larger scales when compared with the Homicide variable. Thus, when conducting cluster analysis it might be necessary to scale these variables.

Finally, it is important to examine the scatterplots for each pair of variables contained within the dataset. Figure 3 below, shows the pairs plot for the crime dataset. The scatterplots show that there is a strong positive relationship between the Narcotics and Homicide variables with a positive relationship also occurring between the Theft and Narcotics and Theft and Homicide variables, respectively.

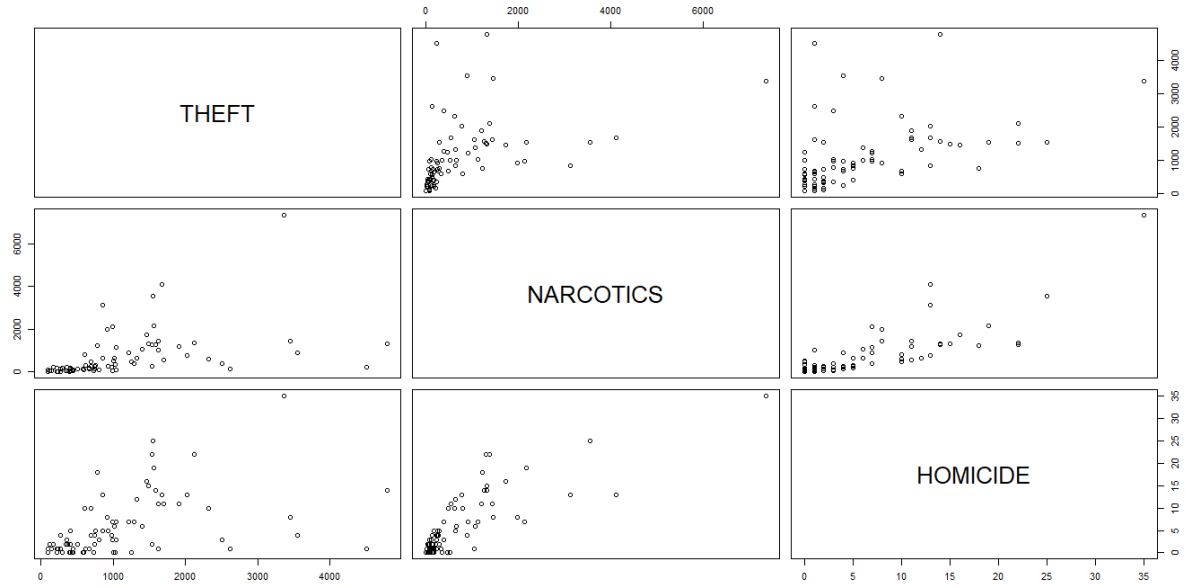

Figure 3: Scatterplots of the different pairwise combinations of the variables contained with the crime dataset.

2 Methods

2.1 Hierarchical Clustering Algorithm

Hierarchical clustering is an unsupervised learning technique which aims to group similar observations together and keep dissimilar observations apart. The hierarchical clustering algorithm is performed as follows:

1. Assign each observation to its own group.
2. A distance matrix is calculated between all observations using a specified distance measurement e.g. Euclidean, Manhattan or binary depending on the data type.
3. A specified linkage method is then used to measure the dissimilarity between two groups. Groups which are closest to each other are merged together to form one single group.
4. This process is repeated until only one group remains.

Examples of linkage methods include single, complete and average linkage as well as Ward's linkage method, each of which calculates the dissimilarity between groups in different ways.

In R, using the hclust function, it is possible to calculate the dissimilarity values for a pre-specified distance measurement and linkage method. Once the distance matrix has been obtained it is possible to represent the dissimilarities between clusters as a dendrogram. A dendrogram has a tree like structure from which it is possible to identify potential groups present in a dataset. The y-axis in a dendrogram represents a measure of closeness of the observations or groups. It is possible to cut a dendrogram into groups using the cutree function whereby the user specifies a certain number of groups to be extracted from the dendrogram.

2.2 K-means Clustering Algorithm

The k-means algorithm is a partition based clustering algorithm and is performed as follows:

1. Partition the data into K different clusters.
2. Calculate the centroid for each of the K clusters.
3. For each observation compute its distance from the centroid to which it is closest.
4. Reassign each observation to its nearest centroid.
5. Repeat steps 2-4 until the cluster memberships remain unchained or a maximum number of iterations is reached.

The k-means algorithm is prone to finding locally optimal solutions, therefore, it is important to use different starting values to help validate the solution. K-means is also sensitive to outliers.

3 Results and Discussion

3.1 K-means Clustering

In this section different clustering methods will be used to examine the possible number of similar groups present within this dataset. Before applying any clustering algorithms it was decided to divide the counts of each crime within each neighbourhood by the corresponding neighbourhood population to create a crime rate for which the different communities could be compared. Also, as mentioned above, since the Theft and Narcotics variables occur more frequency and have larger variances than the Homicide variable it was important to scale the data. Thus, the newly found crime rates were scaled to ensure that Theft and Narcotics did not have a larger influence than Homicide on the outcome of the analysis.

The k-means clustering algorithm was first used to examine the crimes dataset. The data was clustered for $k = 1, ..., 15$ where k is the number of clusters fitted to the dataset. For each clustering solution the within cluster sum of squares (WCSS) was calculated, with these values plotted against their corresponding number of clusters, k. This plot is called an "eblow" plot as it is believed the optimal number of clusters is found where an angle or "elbow" occurs in

the plot, i.e. where the reduction in the within cluster sum of square starts to plateau. The "elbow" plot for the k-means algorithm is shown below in Figure 4.

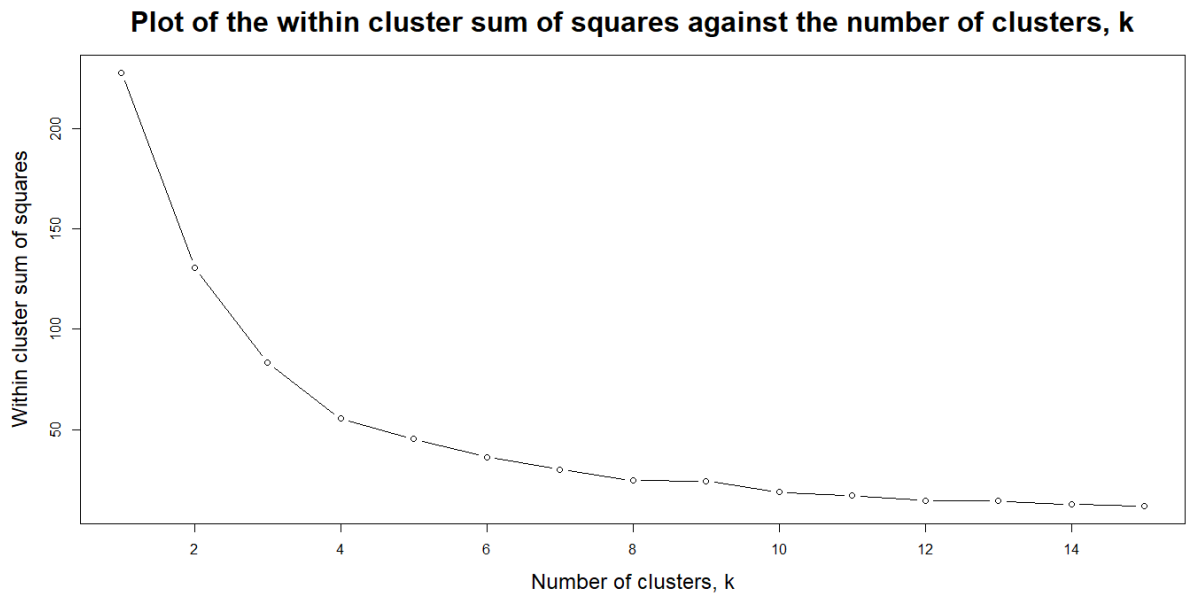

Figure 4: Plot of the within cluster sum of squares against its corresponding number of clusters, k.

There appears to be a small "elbow" present for the value $k = 4$, however more analysis will be conducted before concluding that this is the optimal number of clusters found by the k-means clustering algorithm.

It is now possible to fit scatterplots of the scaled Theft, Homicide and Narcotics rates against each other and colour these observations based on their cluster memberships found using the k-means clustering algorithm. In this case the 4 cluster solution will be used; these scatterplots are shown below in Figure 5.

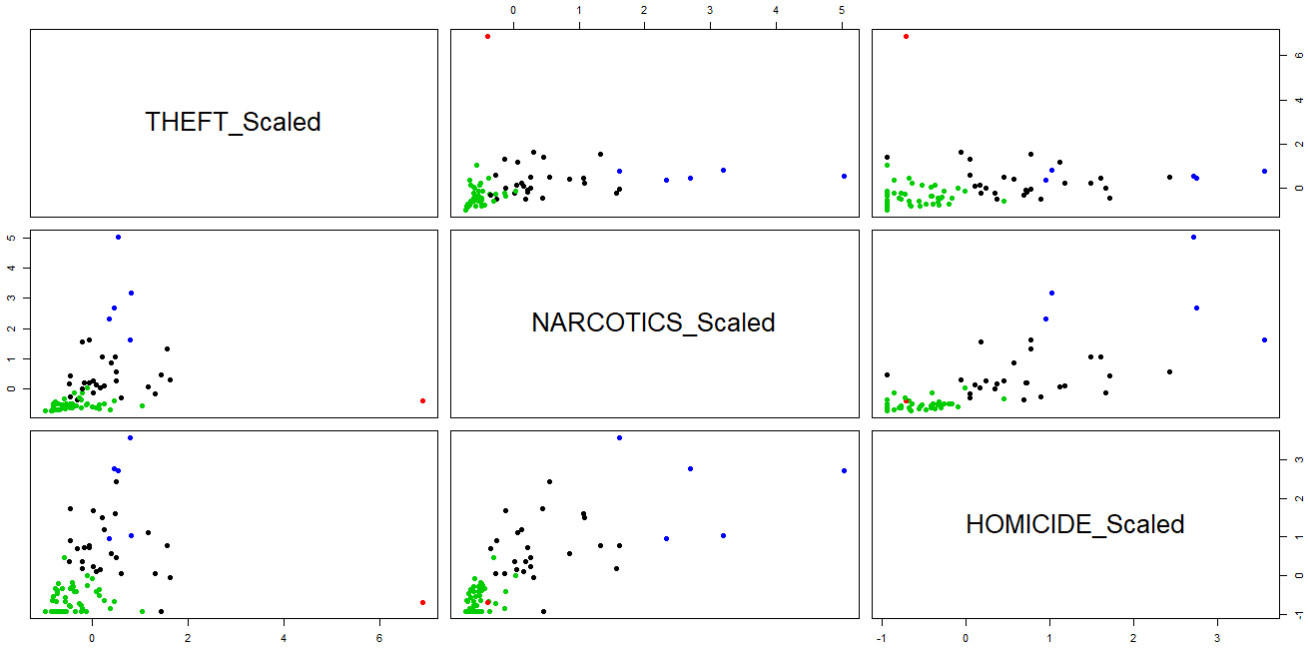

Figure 5: Scatterplots of each pair of scaled crime rate variables, grouped by the k-means cluster solution for k = 4.

It is clear from the scatterplots shown above that there are four groups present within the crime dataset. When examining the scatterplot of the scaled theft rate against the scaled narcotic rate, it is clear that there are 4 distinct groups of communities. Communities with low rates of thefts and drug related crimes appear to be clustered together. Similarly, communities with low theft rates appear to have been clustered with areas of low and medium drug related crime rates, respectively. Finally, the remaining cluster contains only one element with this observation representing the Loop community. In this case there is a really high theft rate but little to no drug related crimes. Since the area of Loop is also the location of Chicago's central business district these results make sense, given that the Loop is a rich neighbourhood which might be targeted by thefts and criminals. Also richer neighbourhood are not generally associated with drug related crimes and it might be for these reasons that the Loop community was clustered on its own.

Similarly, examining the scaled narcotic rates against the scaled homicide rates scatterplot, communities with low, medium and high rates of both Homicide and Narcotics appear to be clustered together.

3.2 Hierarchical Clustering

Next, hierarchical clustering was used to examine and help identify possible clusters present within the crime dataset. The distance matrix was calculated using the Euclidean distance function between all values contained within the scaled rates of the Theft, Narcotics and Homicide variables. Ward's linkage method was used as a measure of cluster similarity when combining clusters in the hierarchical clustering algorithm. Ward's linkage method treats cluster analysis problems as analysis of variance problems. Therefore, clusters are merged together when the error sum of squares is minimised. This produced a dendrogram which is shown below in Figure 6.

It is clear from the dendrogram that there may be some structure present within the dataset. From visual inspection of the dendrogram it appears that the optimal number of clusters present in the crime dataset is 4, as this is where there occurs a large range (vertical distance) over which the number of clusters in the solution does not change.

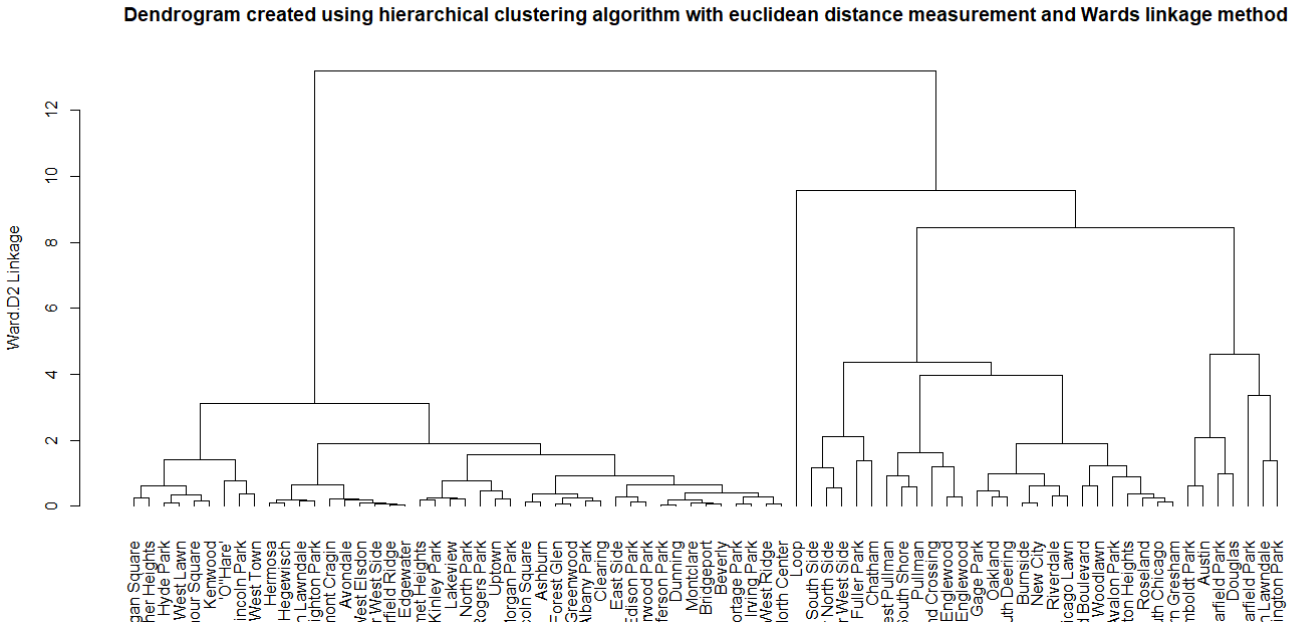

Figure 6: Dendrogram of the hierarchical clustering of the crime dataset using Euclidean distance and Ward's linkage method.

Hierarchical clustering using the Euclidean distance function and the single, complete and average linkage methods were also used to cluster the crime dataset, however, these linkage methods were unable to create adequate clustering solutions as their dendrograms contained little structure i.e. often the optimal cluster solution included 2 or 3 single observation clusters with the remaining values forming the final cluster.

3.3 Assessing Cluster Structure

In order to assess the clustering structure a silhouette plot was used. A silhouette plot can be used as a measure of how well a cluster solution fits the data. A silhouette value is a measure of how similar an observation is to its own cluster when compared to how similar that observation is to other clusters. A high silhouette value indicates that an appropriate number of clusters have been fitted to the data. The presence of a large amount of negative or low silhouette values may indicate how an inappropriate number of clusters has been fitted to the dataset. A silhouette plot was constructed for the believed optimal number of clusters found using the k-means clustering algorithm, i.e. 4 clusters. This silhouette plot is shown below in Figure 7.

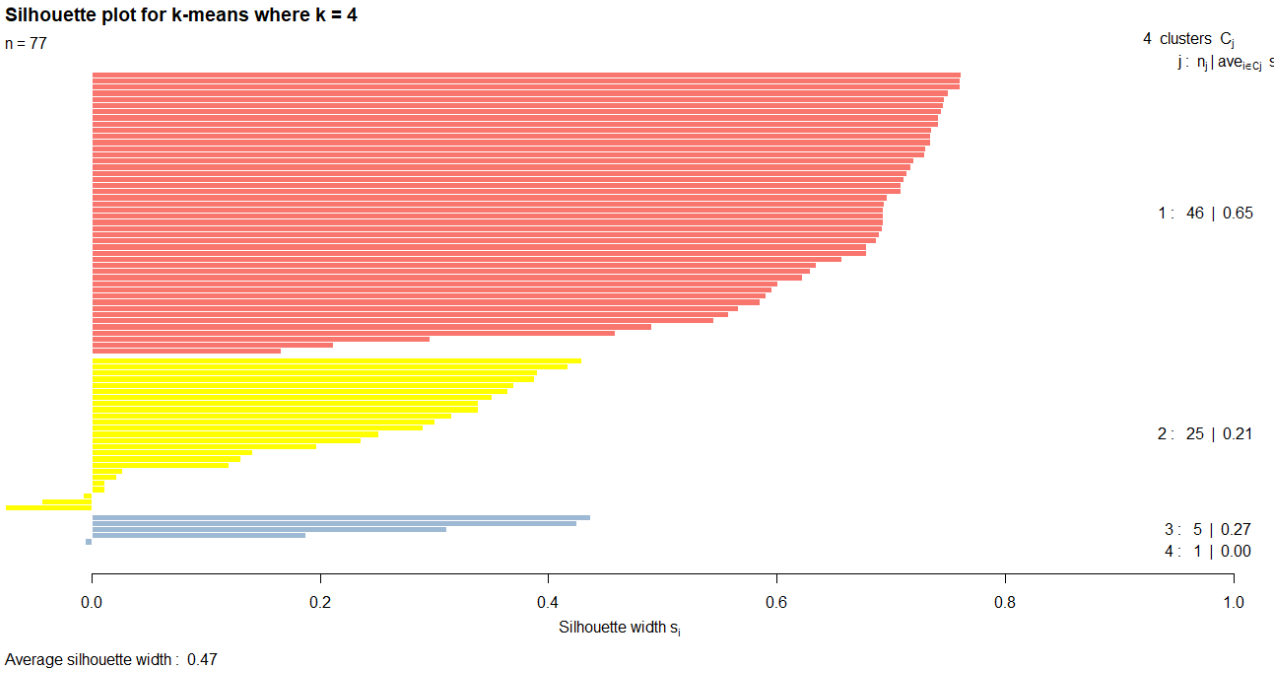

Figure 7: Silhouette plot for the 4 cluster k-means model.

The 4 cluster solution obtains an average silhouette width of 0.47, we can see from the plot that there are some negative values present which indicates how these communities might have been placed in the wrong cluster. Also the presence of a 1 observation cluster means that a silhouette width of 0 is assigned to this cluster. The low values of silhouette widths for the second and third cluster indicate the possible lack of structure present within these clusters.

Silhouette plots were also created for other k-means solutions, it was found that the 3 cluster solution obtained the highest average silhouette width of 0.56. This would suggest that a 3 cluster solution fits the crime dataset better than the 4-cluster model.

A silhouette plot was also fitted for the 4 cluster hierarchical clustering with Ward's linkage method. The silhouette width for this cluster solution was found to be 0.46. Thus, when comparing the average silhouette widths, the k-means clusterings for values $k = 3$ or 4 are found to be a slightly better fit to the data than the fit of the 4 cluster hierarchical clustering solution.

3.4 Cluster Solution Comparison

The Rand and adjusted Rand indexes can be used to compare two separate clustering solutions. Table 1 below is a table of the different Rand and Adjusted Rand index values obtained when comparing the different clustering solutions found within this investigation.

K-means	Hierarchical Clustering	Rand Index	Adjusted Rand Index
3 clusters	4 clusters	0.689	0.401
4 clusters	4 clusters	0.935	0.869

Table 1: Evaluation of the comparison between the k-means and hierarchical clustering solutions.

The adjusted Rand index is very similar to the Rand index, however, it has been adjusted for agreement due to chance that can occur between two cluster solutions. Thus, when examining the results above it is important to focus on the adjusted Rand index values.

It was found that there was some agreement between Ward's linkage hierarchical 4 cluster solution and the 3 cluster k-means clustering solution with an adjusted Rand index value of 0.401 obtained. However, there was found to be a large degree of similarity between the 4 cluster k-means and the 4 cluster Ward's linkage hierarchical clustering solutions, with an adjusted Rand index value of 0.869. Examining the relationship between the 4 cluster k-means and 4 cluster hierarchical clustering solutions in more detail a table was created to compare the cluster memberships of each solution. This table shown below in Table 2 demonstrates how there is a large amount of similarity between the two clustering methods. Therefore, 4 clusters seems like the optimal number of clusters present within the crime dataset.

		Hierarchical Clustering			
		1	2	3	4
K-means	1	0	0	5	0
	2	44	2	0	0
	3	0	0	0	1
	4	0	23	2	0

Table 2: Table comparing the 4 clusters from both the k-means and hierarchical clustering (with Ward's linkage method) algorithms.

3.5 Mapping the results of the clustering algorithms

It was possible to find online the shapefile of the Chicago region including the 77 individual communities which make up the city of Chicago. A map of Chicago was created which included the 77 different community boundaries using the R libraries ggplot2 and maptools. The regions of this map were then coloured by the 4 cluster k-means and hierarchical clustering solutions found above. Figure 8 below shows the maps of Chicago coloured by the different clustering solutions. The first map shows the communities which were clustered together by the k-means algorithms, with the second representing those clustered by the hierarchical clustering algorithm. The final map represents the communities with the highest crime rates per population for the year 2005. From the first two maps it is clear that there are a large number of communities which were similarly clustered by the two clustering algorithms. It is more difficult to compare the total crime rates map against the other two maps, as the total crime rates is measured on a continuous rather than a discrete scale. However, there are some areas which appear to look similar between the total crime rates map and the maps created from the two clustering solutions. In particular the southern and central regions of Chicago appear to compare well with the k-means and hierarchical clustering maps.

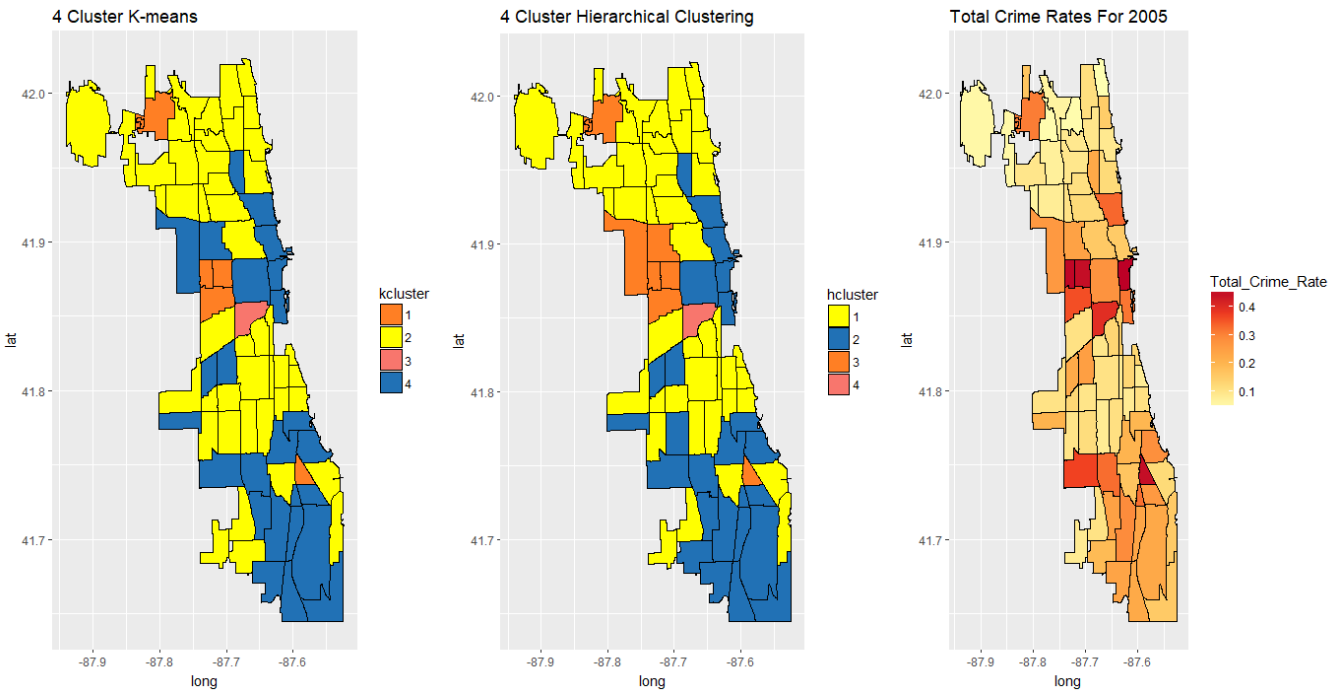

Figure 8: Maps of Chicago coloured by the different clustering solutions.

3.6 Further Discussion

One key observation after examining the crime dataset is that those who commit the crime might not necessarily live in the area where the crime was committed. Therefore, these communities might not fully represent areas which are very dangerous. This is particularly evident for the Loop community which has a very large level of theft but low levels of other crimes. From this investigation the Loop area might be classed as a dangerous area, however, this does not appear to be the case and the low levels of drug related crime and homicides would suggest that Loop is actually a safe neighbourhood in general.

It would be interesting to examine the number of schools present within a community or if the average education rate of the community was available. It would then be interesting to examine how the crime rates would vary with respect to education level.

4 Conclusion

It was found when using the k-means clustering algorithm that a 4 cluster model was the best fit to the crime data after examining the "elbow" plot. Similarly, for the hierarchical clustering algorithm employing an Euclidean distance measurement with the Ward's linkage method, it was found that when visually examining the plotted dendrogram, that the optimal number of clusters for the crime dataset was 4. These clustering solutions were examined in more detail and compared with each other using silhouette plots and the Rand and adjusted Rand indexes, respectively. The 3 cluster k-means solution obtained the highest average silhouette width of all the clustering solutions. However, it was decided that a 4 cluster solution was optimal, since there was such a large degree of similarity between the k-means and hierarchical clustering solutions. Therefore, it is possible to conclude that there are 4 groupings of communities present within the city of Chicago based on similar rates of thefts, drug related crimes and homicides.

Finally, we can see when fitting the clustering solutions to the map of Chicago that there is a large amount of similarity between the communities present within the k-means and hierarchical clustering solutions. There also appears to be some similarity with the total crime rate map which would suggest that our clustering solutions are a good representation of the different levels of crime across the 77 communities in the city of Chicago.

References

[1] chicagopolice.org. Crime Type Categories: Definition and Description. URL: http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html#N09.