Allocating Funds to London Boroughs Based on Crime Statistics

By: Talia Dagan
Dr. Shanker, Machine Learning
Kent State University
December 10th, 2021

Problem:

London crime rates have been on the rise for the last two decades throughout different boroughs in London. London has 32 boroughs, and each have dealt with a significant rise in crime rates through the years of 2008-2016. Her Majesty's Government, which is the United Kingdom's government, needs to make a decision on what boroughs to allocate more funds to in the next year. This is being done to combat the problem of rising crime rates. Allocating more funds to the boroughs will allow the Metropolitan Police Service in London to hire more police officers and acquire more protective equipment.

Data:

The data used to make a decision on what London borough needs more funding is found off of Kaggle. The data is real-world data; meaning it is real crime data from London. The data has 9 variables including ISO code, borough, major crime, minor crime, value, month, year, longitude, and latitude. Value refers to the monthly reported count of categorical crime in a given borough. 4 rows are qualitative, and the remaining 5 rows are quantitative data. There are 13 million rows, but since the dataset is so large, I randomly picked 10,556 rows to use to make a decision. Reference to data: Boysen, Jacob. "London Crime Data, 2008-2016." *Kaggle*, 3 Aug. 2017, https://www.kaggle.com/jboysen/london-crime.

Approach:

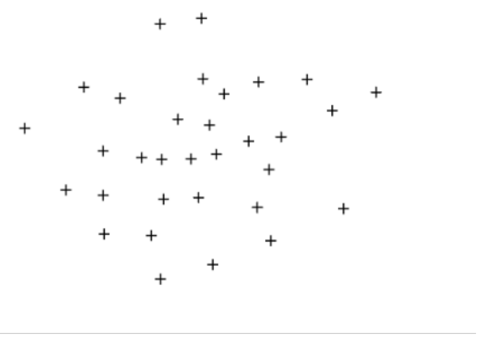
I approached this problem using unsupervised learning and clustering algorithms. It is necessary to use clustering algorithms because I want to classify the data into specific groups. The techniques used to cluster are K-means clustering as well as hierarchical clustering. Both clustering techniques were used to compare the output. Hierarchical clustering was used because it is more accurate when dealing with spatial data. k-means was used to compare the K value.

Since dealing with longitude and latitude, it is necessary to convert the coordinates into a spatial points data frame. A spatial points data frame is used for points that have 2 dimensions. Using spatial data also requires defining the distance threshold. In this case, 1000 M was used by averaging the distance in London's smallest borough and London's largest borough. This had to be done since dealing with distance, it is not possible to use Euclidean distance with coordinate points. This is because the world is globe shaped and the Euclidean distance computes the distance of a straight line.

Analysis:

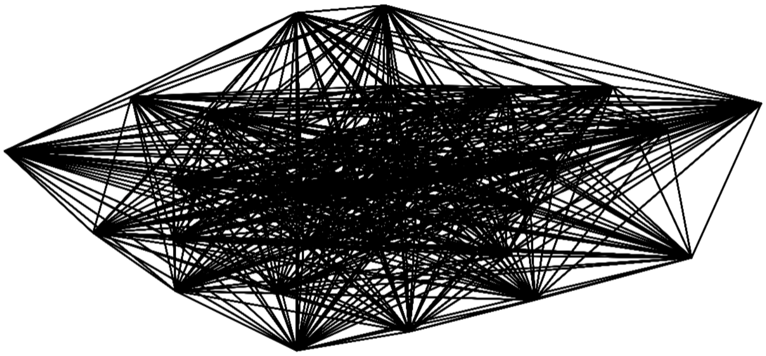
The first step was to visualize the spatial data. This is done using the data, coordinates, and plot functions in RStudio. The results are located in the image below (Image A). As you can see, there are 32 points to represent the 32 boroughs in London.

Image A:



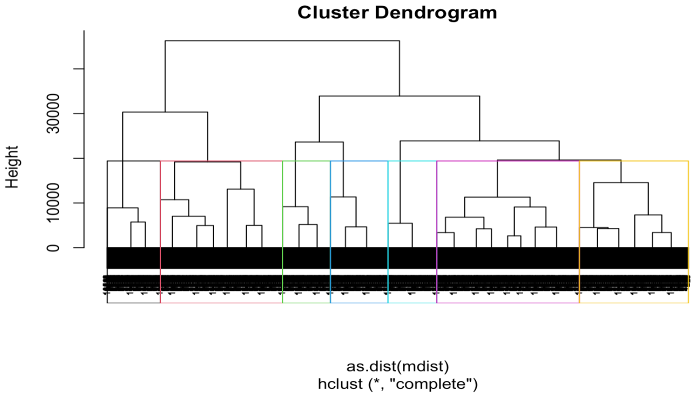
Next, to visualize the connection between the rows of data and the points using the spatial data points, it was necessary to use the spatial lines function. The image below (Image B) shows the connection between every coordinate point and the data. This is important to see the connection between the longitude and latitude of all the data.

Image B:



The first technique used to solve the problem of London crime rate was hierarchical clustering to cluster the spatial points. The dendrogram produced from hierarchical clustering is below in Image C. From the height of the dendrogram, it can be seen that 7 clusters are appropriate. This means that the 32 boroughs of London can be clustered together into 7 different clusters, composed of those boroughs' coordinates points. In the image below, it is not possible to read the coordinate points. This is because there are 10,556 points on the X-axis of the dendrogram.

Image C:



The next step was to use the K-means clustering algorithm to compare the clusters. To use K-means, it is necessary to determine the K value before plotting. This was done using both silhouette and elbow method. Using silhouette method (Image D), it was found that London Boroughs can be split up into 5 ideal clusters. Using the elbow method (Image E), 5 clusters are also deemed appropriate. In Image F, you will see the output for plotting k-means with 5 clusters. The longitude of the borough is located on the X-axis and the latitude is located on the Y-axis. From the cluster plot produced in Image F, it is seen that all London boroughs belong to a cluster.

Image D:

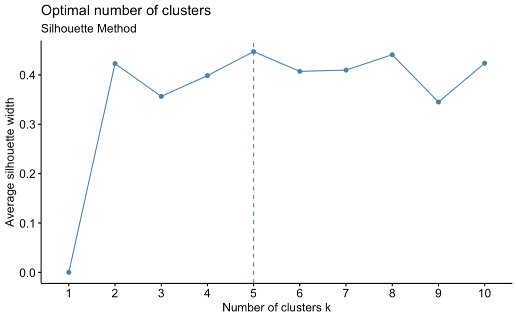


Image E:

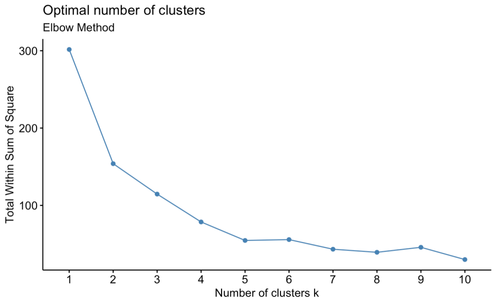
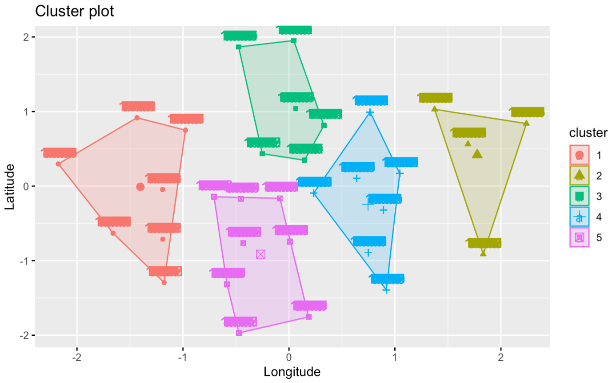


Image F:



Conclusion:

In conclusion, it is found that Her Majesty's Government should evenly distribute more funds to all 32 London boroughs. This is since all 32 London boroughs are experiencing a rise in crime rate. Looking at the clusters, 7 different clusters produced using hierarchical clustering and the 5 clusters produced using k-means, it is difficult to conclude which boroughs deserve more funding than the others. This being said, I can certainly conclude that Her Majesty's Government should equally distribute and allocate funds to all 32 London boroughs to tackle the rising crime rates in London.

Link to .rmd file: https://github.com/taliadagan/tdagan_64060.git