Capstone Project: Battle of Neighbourhoods

1. Introduction

Modernization is taking its toll on the world today. The increase in the number of facilities, infrastructure etc. in a region has led in the increase of population in that area. People keep on moving to areas where these facilities are found in abundance. Also it depends on what the people like. Sometimes they prefer to be near a good school, some prefer to be near train stations, bus stops and so on. So the moving is based on things which make people happy and necessary.

And people do a lot of research on the places so they don’t waste time and money. Safety is a major factor in deciding this process.

Problem

The aim is to select the top ten places in each neighbourhood with low crime rates for a particular country. Here we use the kaggle data set for crime rates in London Boroughs from 2008-2016. We use K-means Clustering for this purpose.

2. Data Acquisition and Cleaning

The data acquired for this project is from three sources:

The first is a dataset of London crimes which contains the following columns:

* Isoa\_code: Area code in London
* Borough: Name of London borough
* Major category: High level crimes
* Minor category: Low level crimes
* Value: Monthly report of crimes
* Year: Year between 2008-16
* Month: Month between 1-12

The second is data scraped from a Wikipedia page that contains London Boroughs.

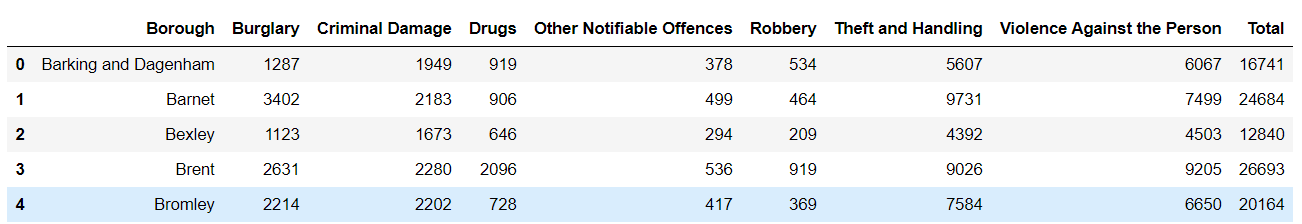
* Boroughs-Names of London Boroughs
* Inner-Checking if it is inner or outer London.
* Status-Checking if it is royal, city etc.
* Local Authority-Local Authority assigned to the Borough
* Political Control-Political party that controls the Borough
* Headquarters-Headquarters of the Borough
* Area-Area of the Borough in miles.
* Co-ordinates-Latitude and Longitude of the Boroughs

The third is also from a Wikipedia page which is created from scratch. It is the list of neighbourhoods in Royal Borough of Kingston.

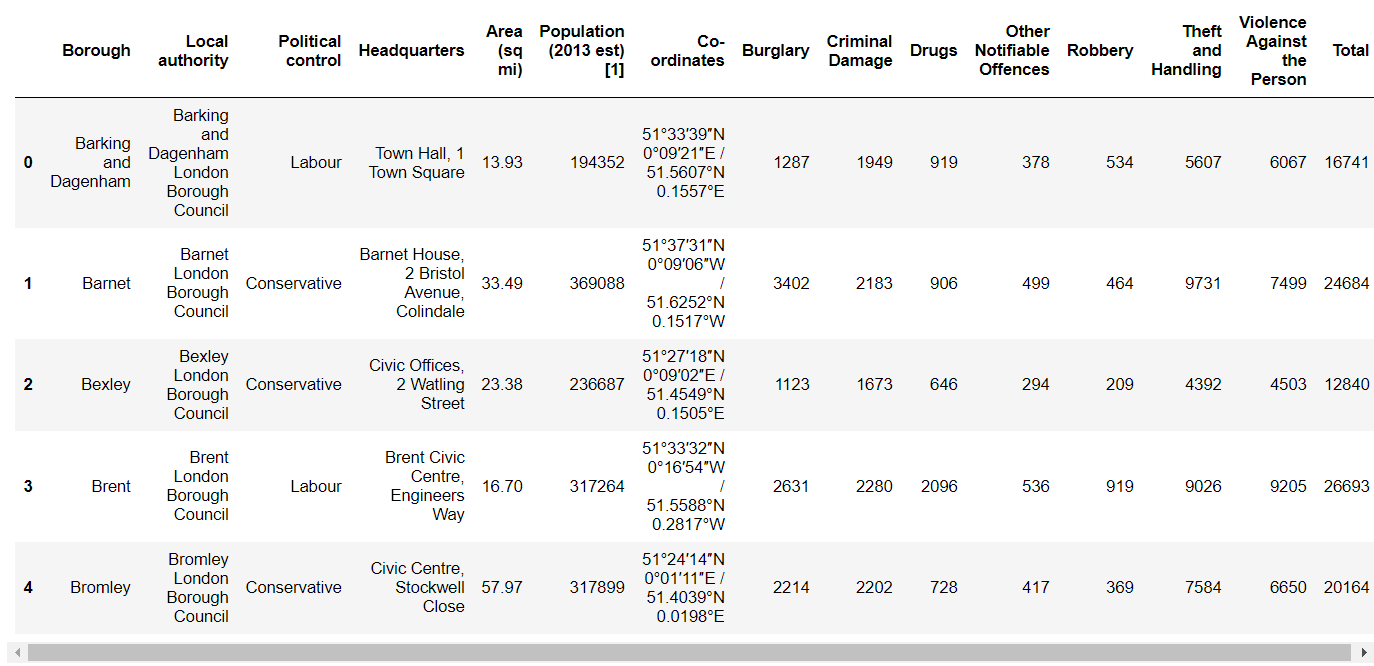
* Neighbourhood-Name of the Neighbourhood in the Borough
* Borough-Name of the Borough
* Latitude-Latitude of the Borough
* Longitude-Longitude of the Borough

Data Cleaning:

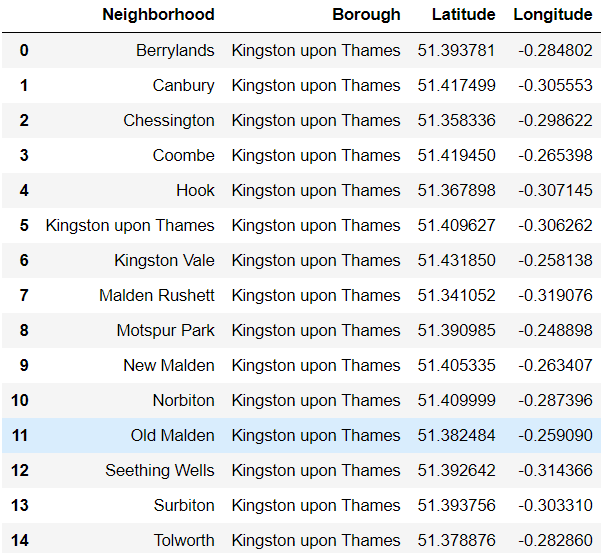
For the London Crimes Dataset the crimes during the most recent year (2016) are only selected.



For the second dataset the web scraping is done using Beautiful Soup library. The table is obtained and after string manipulation the following table is created.



The third source of data is acquired from the list of neighbourhoods in the safest borough on Wikipedia. This dataset is created from scratch; the panda’s data frame is created with the names of the neighbourhoods and the name of the borough with the latitude and longitude.

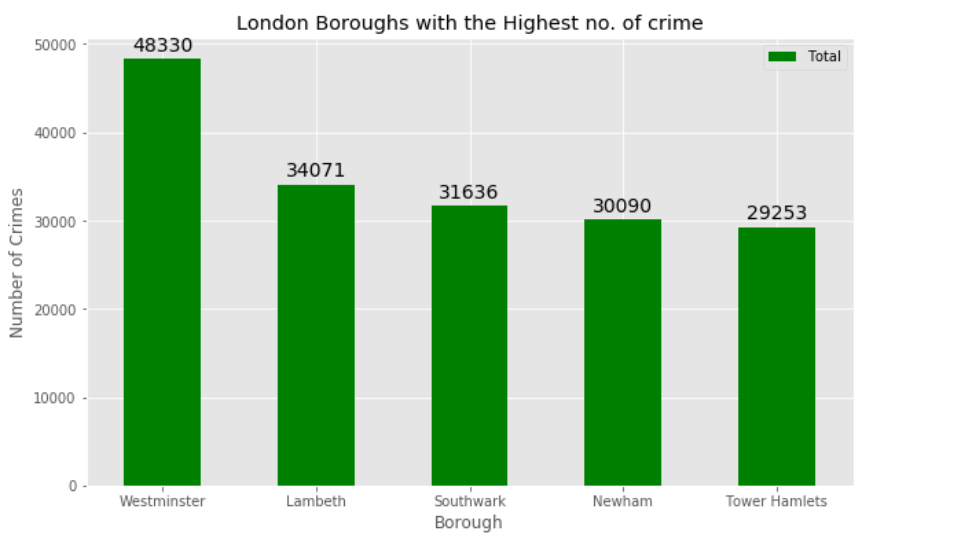


Methodology

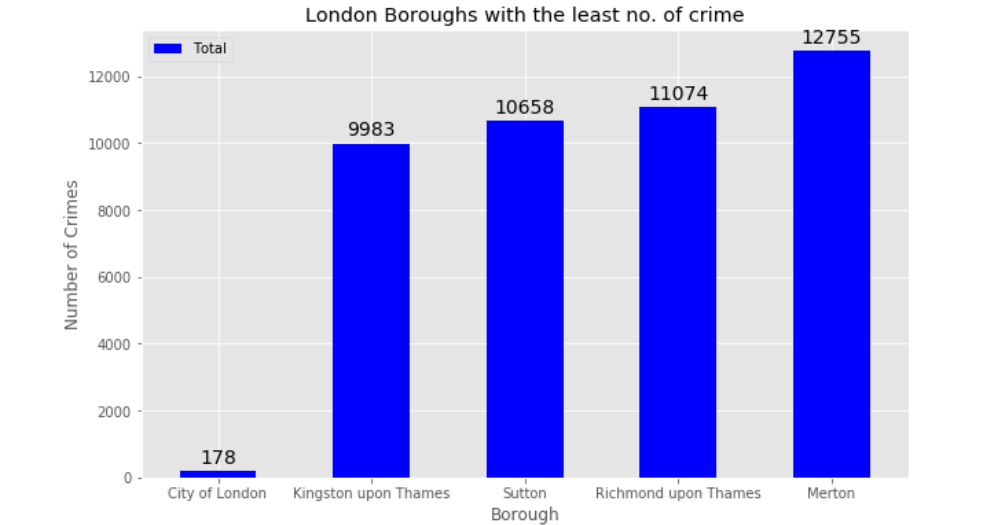
EDA

The describe function in python is used to get statistics of the London crime data, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime.

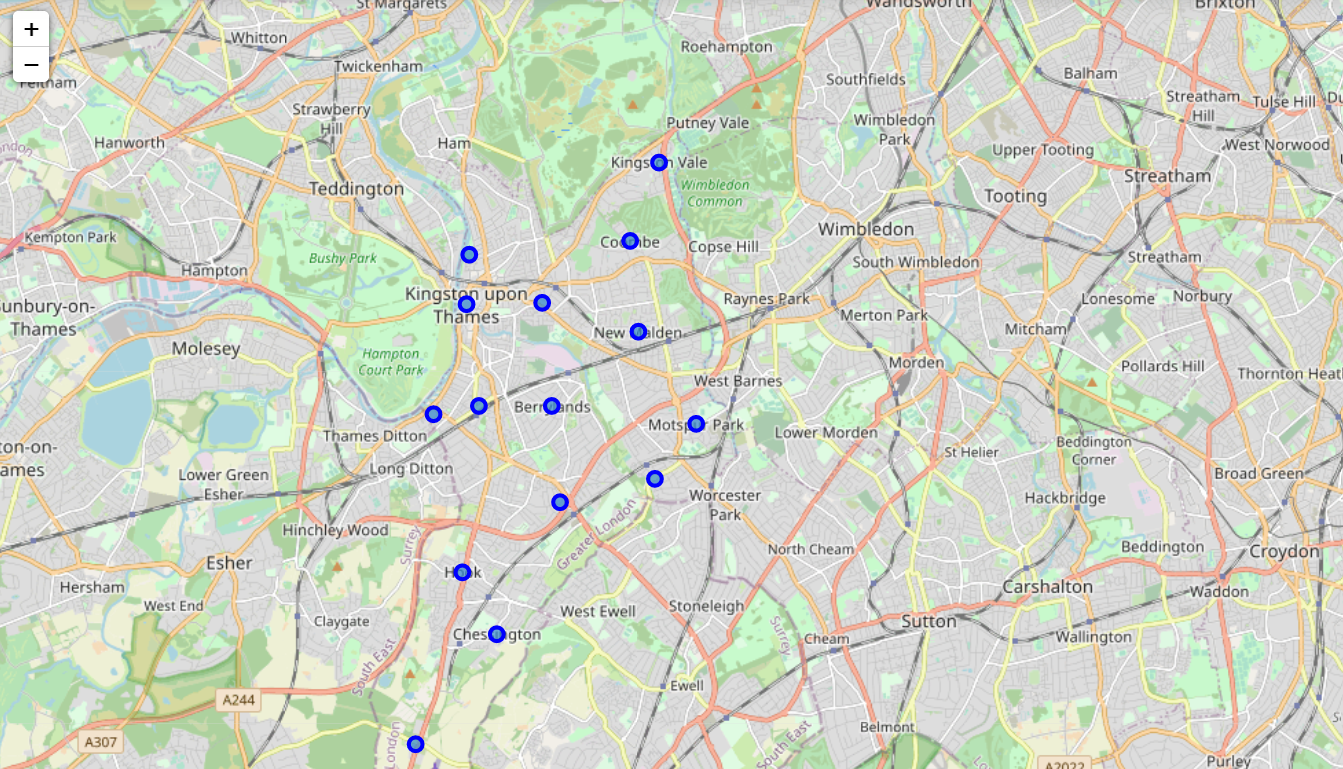
Comparing five boroughs with the highest crime rate during the year 2016 it is evident that Westminster has the highest crimes recorded followed by Lambeth, Southwark, Newham and Tower Hamlets. Westminster has a significantly higher crime rate than the other 4 boroughs.



Also, comparing five boroughs with the lowest crime rate during the year 2016, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton.

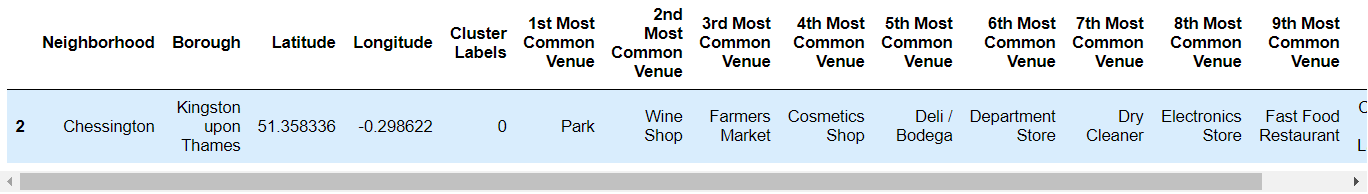


There are 15 neighbourhoods in the royal borough of Kingston upon Thames, they are visualised on a map using folium on python.



Results

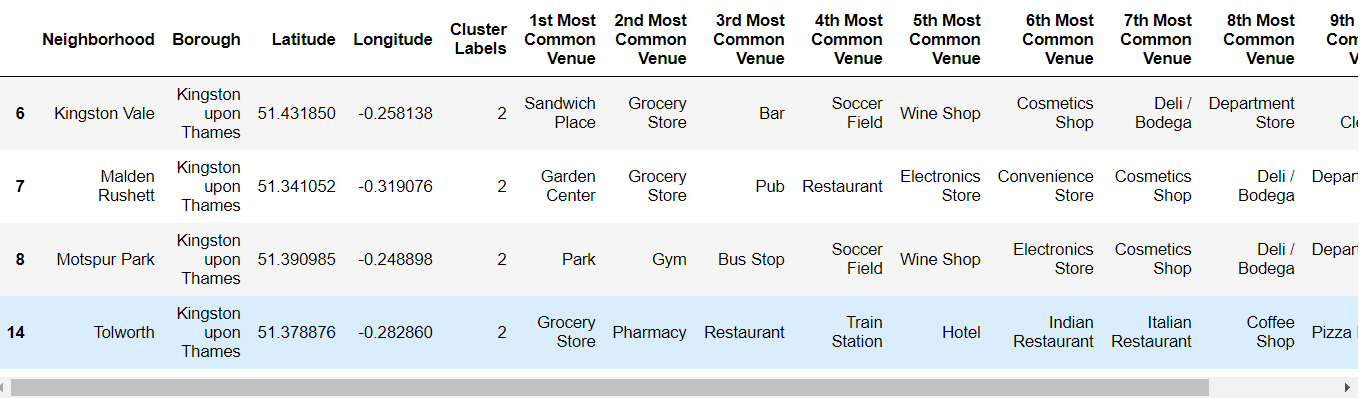
After running the K-means clustering we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters. Looking into the neighbourhoods in the first cluster :



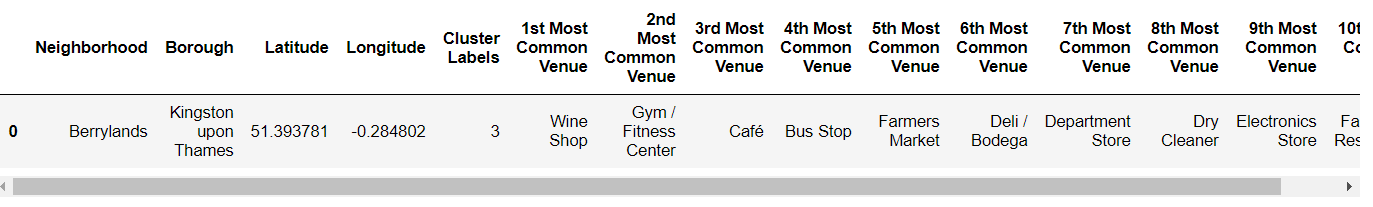
The first cluster consists of only one neighbourhood which consists of fast food restaurant, department stores etc.



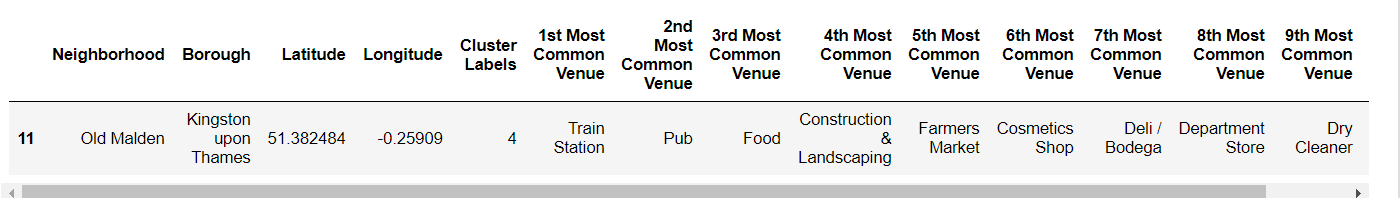
The second cluster consists of seven neighborhoods.These neighbourhoods consist of many places like cafes, pubs, electronics store etc.



The third cluster consists of four neighbourhoods consisting of gyms, hotels etc.



The fourth cluster consists of one neighbourhood.



The fifth cluster also consists of one neighbourhood.

Conclusion

So we can find out the most suitable place to live in based on these clusters. Each cluster consists of neighbourhoods, which consist of different venues.

So people can choose the locations they prefer. With the advancement of technology it would be useful for people to use it in situations which are important to them in terms of money and time.