Capstone Project- The Battle of Neighbourhoods

Introduction:

This is capstone project for Applied Data Science in which we will consider the data set of London and its neighbours, as many people move to new cities but doesn't know where to stay as they don't know which neighboring city is safest for them.

Background:

As I've studies an average American moves about eleven times in their lifetime. This brings us to the question: Do people move until they find a place to settle down or do our needs change over time, we eventually leaving the town. What brings us satisfaction in the new area we are going to? Or do we often move to new area without knowing anything about that; leads to discomfort and unhappiness.

Problem:

The crime stats in this dataset of London found on Kaggle has crimes in each Boroughs of London from 2008-2016. The 2016 being the latest one, we will be considering the data of the year which is actually old information as of now. The crime rates in each borough may have changed over time.

This project is aimed to select the safest borough in London based on the total crimes, explore the neighborhoods of the borough to find the most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

People who are considering to relocate to London will be interested to identify the safest borough in London and explore its neighbourhood and common venues around each other.

Data Acquisition and Cleaning

Data Acquisition:

The data acquired for this project is a combination of data from three sources. The first data source of the project uses a London Crime data that shows the crime per borough in London. The dataset has 7 columns :

1. Isoa_code
2. Borough
3. Major_category
4. Minor_catergory
5. Value
6. Year
7. Month

The second source of data is carped from a Wikipedia page that contains the list of London boroughs. This page conatins additional information about the boroughs, the dataset contains 10 columns:

1. Borough
2. Inner
3. Status
4. Local Authority
5. Political Control
6. Headquarters
7. Area Population:
8. Co-ordinates
9. Nr. In map

Then we have third data source which is the list of Neighborhoods in the Royal Borough of Kingston upon Thames as well was their on Wikipedia page. The dataset contains 4 columns:

1. Neighborhood
2. Borough
3. Latitude
4. Longitude

Data Cleaning

Then I will be doing Data Cleaning for each of the three sources of the data is done separately. From the London crime data, only the crimes of 2016 will be selected as it's the recent.

The second data is gathered from a Wikipedia page using the Beautiful Soup library in Python. Using the library we can extract the data in tabular form.

Then we will merge the two datasets as both datasets contain borough so we will merge them using borough names. The basic purpose of merging is to see the visualized crime rates in each borough for identifying the lowest number of crimes in 2016.

After gathering or identifying the safest borough for we will obtain the co-ordinates using Google Maps API geocoding to get the final dataset. This new dataset is created using Foursquare API with most common venues (10) each; then using k-means clustering algorithm to make clusters.