Aakash Arora

Coursera Capstone Report

April 2, 2019

Capstone Project - Best stay in London

## Introduction

* 1. **Problem Background**

European countries are a hub for tourism. People around the world are attracted towards its variants, beauty and marvellous structures. Although planning a visit and selecting the countries to travel to is not a problem with European countries. Its the stay in the country that requires a lot of time and research. There are ample amount of queries that arise like food, availability, locality, etc.

London, the capital city of England, attracts a lot of tourists round the clock. It has been termed as one of the most desirable destination and is the most visited city in the world. With all this, it is also one of the costliest city in the world.

There are a lot of neighbourhoods and boroughs in London city. To plan a visit to this city a tourist needs to include a lot of locations in the bucket list. Thames river, Trafalgar square, London Bridge, Canary Wharf, London Eye, Tower Bridge and what not.

It can create nuisance for anyone who plans for a visit to this great city, that what could be the best location to live in while in London.

* 1. **Problem Description**

The best options that we get when we go to an unknown place is views of family, friends or anyone who has visited the place. Based on the pros and cons, recommendations of everyone we decide the best place to live in. But in today’s era of technology we cannot be limited by just the opinions of our friends and family and known associates. If we dive into the world we can get all the information and details over the internet. Best place where we can look into is different online portals and search them for reviews by people who have visited the country and based on their views and experience, anyone can make their journey memorable. But the information can be biased or be a one sided story of any accountant and may be even unfurnished or a lie. It is very difficult to take decisions on the basis of these information.

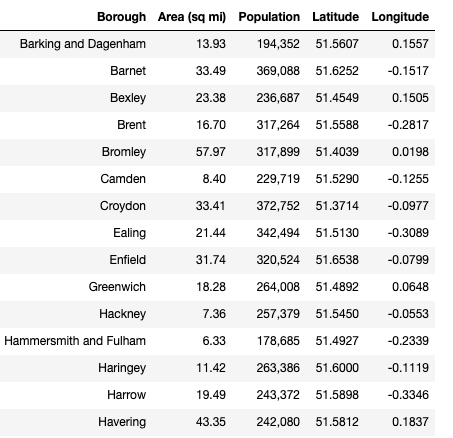
## Data Description

To consider the problem we can list the data as below.

* 1. First part of research was to find all the boroughs in London city. Which I found the Wikipedia page at following URL:

<https://en.wikipedia.org/wiki/List_of_London_boroughs>

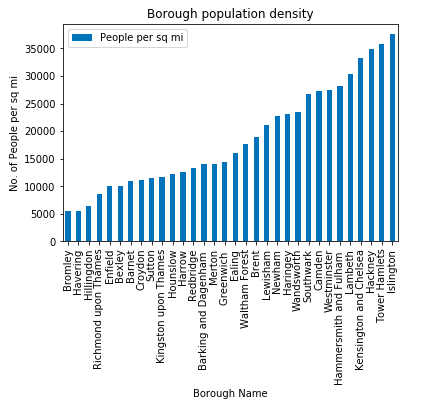
* 1. The above link provides the boroughs with their latitude and longitude and also area wise population of each borough with the area of each borough.
  2. The ratings for venues in the borough are assumed and may be inaccurate but since this is a demonstrating project, the main idea is to get the working model.

**Snapshot of the data**

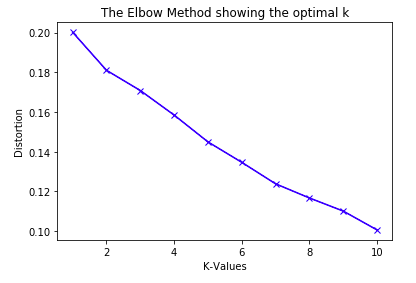
## Methodology:

**Exploratory analysis:**

Scrapping the data from different sources and then combining it to form a single-ton dataset is a difficult task. To do so, we need to explore the current state of dataset and then list up all the features needed to be fetched.

Exploring the dataset is important because it gives you initial insights and may help you to get partial idea of the answers that you are looking to find out from the data.

While exploring the dataset, I found out that Islington has most dense population of all of them.

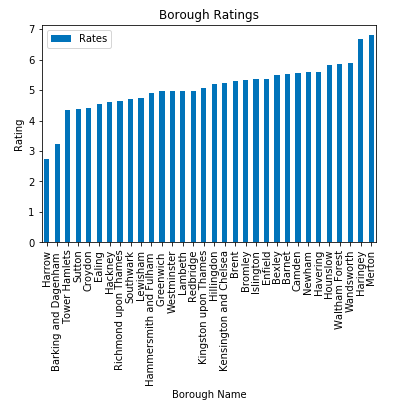
Also while producing graph for number of cluster, I produced a graph to explore all the values for n\_clusters and then finding the best by exploring the elbow graph.

**Inferential analysis:**

Most important factors while building the clustering system were venues, rating and location. They are the most important factor because they have a nonlinear relationship according to our dataset.

It needed to make some inferential analysis to understand this nonlinear relationship. As the amount of population increases, it does not necessarily mean that ratings of venues of a borough will also increase. It is true to most of the case but also many cases differ to follow this trend. Similarly, a borough with less number of people may not necessarily have less rating. It is possible to have less number of people and more income and vice versa. This can be inferred from the above graph

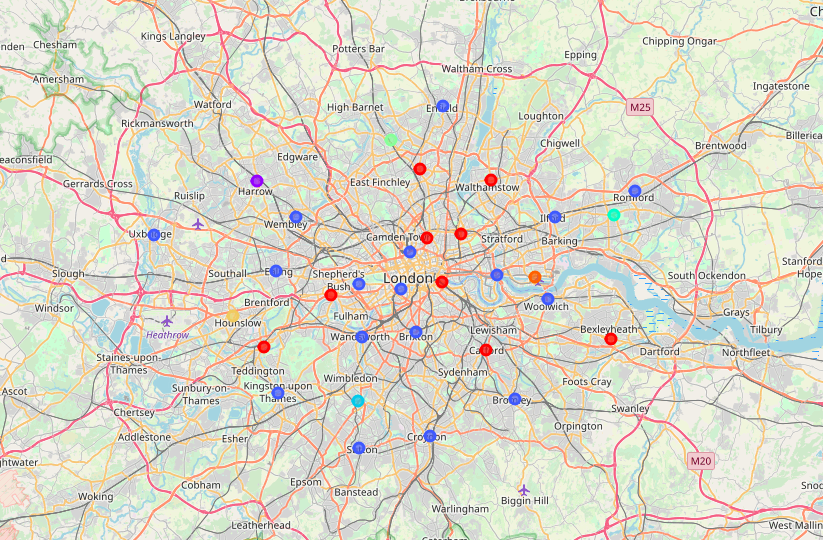
## Result :

The result of the clustering is that it produces a list of top venues, restaurants and recreational areas that the user can enjoy. Based on users preferences, users can put in their inputs of type of venue they would like to visit and they can get favourable outcome of their choice.

The following image shows the result:

## Discussion

Since there was a nonlinear relationship between rating and population, it can be concluded that we must always perform inferential approach to find relationship among different set of features. Also during clustering, similar boroughs must be dumped into the right cluster.

Another observation that we can make is that choosing number of clustering could produce very diverse results. Some may be over fitted or some may be under fitted. Hence analysis of number of clusters must be done. Ref elbow graph in the Methodology section.

The graph above shows the clusters.

## Conclusion

In the above model, K-Means clustering system considers factors such as population, ratings and makes use of Foursquare API to determine recommended boroughs. It is a powerful data driven model whose efficiency may decrease with more data but accuracy will increase. It will help tourists in finding the best spot to start their vacations.