# Coursera Capstone

IBM Applied Data Science Project

## Opening a new Cult Coffee Shop in London, United Kingdom



## Contents

**Owner of the document: Milind Dhanak**

**July 2020**

# Introduction

Most of the coffee shops that operate in prime locations within London are homogenized Transatlantic chains providing the same coffee taste, service and experience. To provide a new refreshing cult coffee experience which not only includes great coffee, coffee cocktails and all day brunch options but also a cult experience of enjoying the coffee experience, a new coffee chain is looking to identify locations within London Borough where it can attract young generation of office employees, students and other professionals.

Opening a coffee shop may sound like a simple straight forward process however, in order to compete with the established coffee chains and attract young generation of coffee drinkers with the cult coffee experience serious consideration is required towards identifying the right location that has the right amount of footfall and the right mix of crowd which would be attracted towards this concept i.e. location becomes the key for such cult coffee chain startup.

# Business Problem

The objective of this project is to identify locations within London Boroughs where this new cult coffee shop an open its store. Using Pythons powerful Data Science tools and Machine Learning techniques like clustering, this project aims to provide solution to the problem of identifying the right location with the Boroughs of London for a new cult coffee chain startup to setup is store. The aim of the this project is to provide multiple viable options for the startup chain to setup its store.

# Data

To come up with a list of viable options, following data will be required:

- List of Boroughs within London with Latitudes and Longitudes. This will define the scope of area for this project.
- Data cleansing will be required to restrict the search closer to the central London area.
- Venue data, particularly data related to existing coffee shops, train stations, educational establishments and shopping areas/centres

## Data Source

Data required for this project will be sourced from following location:

a) Wikipedia page containing list of London Boroughs and its coordinates - https://en.wikipedia.org/wiki/List_of_London_boroughs
b) Use Foursquare API using developer account to get venue data for the London Boroughs to be analysed. The venue data will include information on coffee shops, train stations, educational establishments and shopping centres/areas.

c) For Chloropeth maps for London including Borough names, I used json file of London from the following location: https://skgrange.github.io/www/data/london_boroughs.json

## Methodology

1. Scrape the London Boroughs data using pandas html reading functionality
2. Co-ordinates contained within the Wikipedia page for London Boroughs is converted into Latitutes and Longitudes described in the Process section below
3. Data cleansing will be performed on with London Boroughs data
4. Foursquare API will be used to source venue data using "Explore" endpoint
5. Folium library will be used to plot maps relating to London Boroughs, venues data sourced from Foursquare
6. Clustering will be performed on specific venue category of "Coffee Shop" using K-Means and scikit-learn algorithm to determine the number of optimal K
7. Finally, I used Folium to visualize the clusters on the Chloropeth map of London Boroughs with population density

## Process Steps

### Data Loading and Cleaning

Following steps are performed to load and clean the data:

1) Data is scraped from Wikipedia page

| | Borough | Inner | Status | Local authority | Political control | Headquarters | Area (sq mi) | Population (2013 est)[1] | Co-ordinates | Nr. in map |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Barking and Dagenham [note 1] | NaN | NaN | Barking and Dagenham London Borough Council | Labour | Town Hall, 1 Town Square | 13.93 | 194352 | 51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E | 25 |
| 1 | Barnet | NaN | NaN | Barnet London Borough Council | Conservative | Barnet House, 2 Bristol Avenue, Colindale | 33.49 | 369088 | 51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W | 31 |
| 2 | Bexley | NaN | NaN | Bexley London Borough Council | Conservative | Civic Offices, 2 Watling Street | 23.38 | 236687 | 51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E | 23 |
| 3 | Brent | NaN | NaN | Brent London Borough Council | Labour | Brent Civic Centre, Engineers Way | 16.70 | 317264 | 51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W | 12 |
| 4 | Bromley | NaN | NaN | Bromley London Borough Council | Conservative | Civic Centre, Stockwell Close | 57.97 | 317899 | 51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E | 20 |

2) Redundant columns are deleted, new columns are added to convert co-ordinates into latitude and longitude and additional strings like ºW, ºE and ºN are removed and ºW coordinates are converted into minus.
3) Latitude co-ordinates contained link strings which could not be removed and therefore I added a new column with manually inserting the latitudes and then replaced the existing latitude column.

| | Borough | Local authority | Area (sq mi) | Population (2013 est)[1] | Nr. in map | Longitude | Latitude |
|---|---|---|---|---|---|---|---|
| 0 | City of London | Corporation of London | 1.12 | 7000 | 1.0 | -0.0922 | 51.5155 |
| 1 | Westminster | Westminster City Council | 8.29 | 226841 | 2.0 | -0.1372 | 51.4973 |
| 2 | Kensington and Chelsea | Kensington and Chelsea London Borough Council | 4.68 | 155594 | 3.0 | -0.1947 | 51.5020 |
| 3 | Hammersmith and Fulham [note 4] | Hammersmith and Fulham London Borough Council | 6.33 | 178685 | 4.0 | -0.2339 | 51.4927 |
| 4 | Wandsworth | Wandsworth London Borough Council | 13.23 | 310516 | 5.0 | -0.1910 | 51.4567 |
| 5 | Lambeth | Lambeth London Borough Council | 10.36 | 314242 | 6.0 | -0.1163 | 51.4607 |
| 6 | Southwark | Southwark London Borough Council | 11.14 | 298464 | 7.0 | -0.0804 | 51.5035 |
| 7 | Tower Hamlets | Tower Hamlets London Borough Council | 7.63 | 272890 | 8.0 | -0.0059 | 51.5099 |
| 8 | Hackney | Hackney London Borough Council | 7.36 | 257379 | 9.0 | -0.0553 | 51.5450 |
| 9 | Islington | Islington London Borough Council | 5.74 | 215667 | 10.0 | -0.1022 | 51.5416 |
| 10 | Camden | Camden London Borough Council | 8.4 | 229719 | 11.0 | -0.1255 | 51.5290 |

```
Borough                      object
Local authority              object
Area (sq mi)                 float64
Population (2013 est)[1]      int64
Nr. in map                   float64
Longitude                    float64
Latitude                     float64
dtype: object
```

4) After cleaning the data, I used folium library to plot the visualize the Geographic map of London with the Boroughs superimposed on top based on the latitudes and longitudes of the Boroughs.
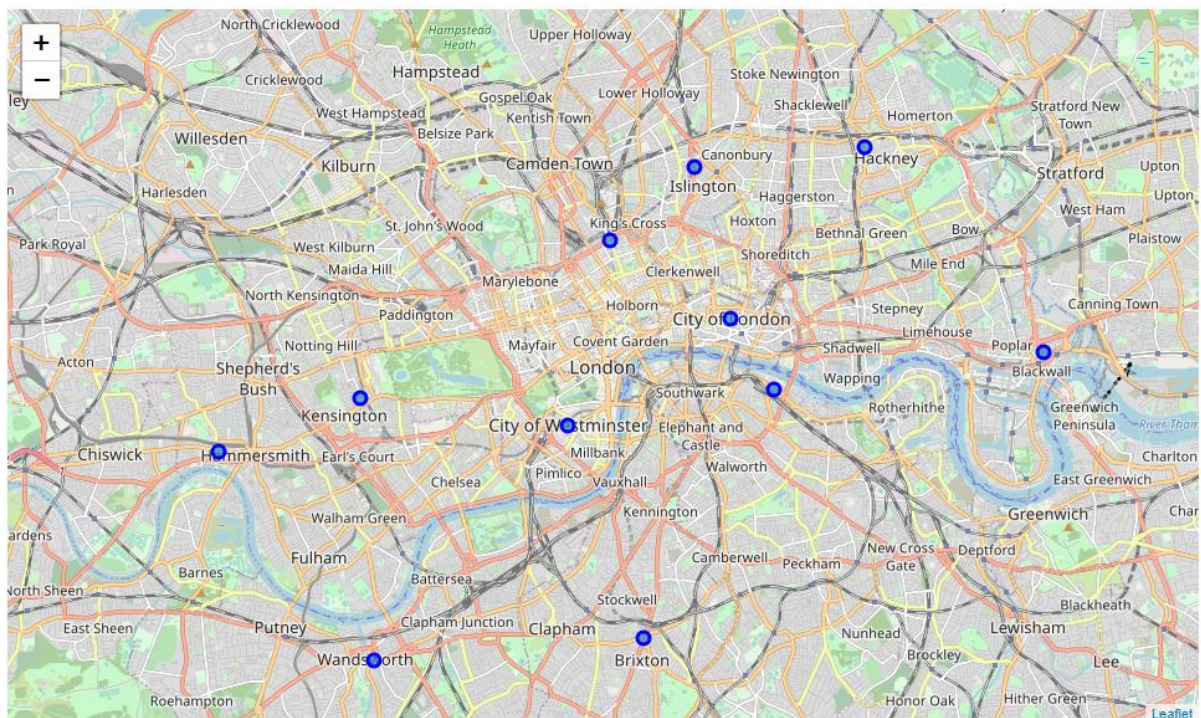


5) I used Foursquare API to get venue information for the Boroughs in my Dataframe. I had one quick observation with the "Search" endpoint functionality in Foursquare. I was searching for "Café" around the Boroughs but it returned errors as possibly it did not have sufficient number of "Cafés" in its database to return for each Borough. I therefore had to use "Explore"

endpoint in Foursquare and retrieve 100 venues for each London Borough and then performing some cleaning and restriction of the Dataframe.
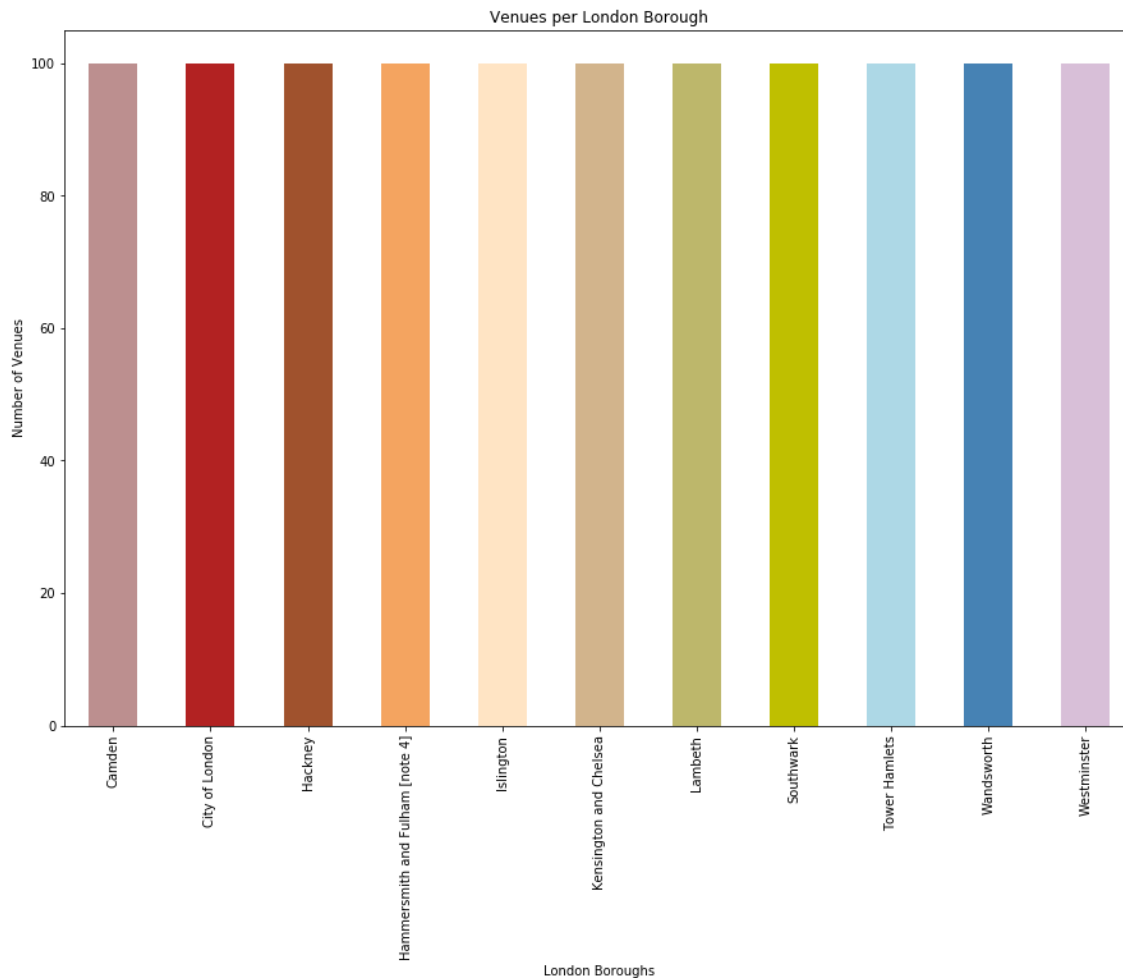


## Machine Learning

▪ I restricted my K-means clustering to the Venue Category "Coffee Shop" from the Foursquare API as the purpose of the this project was to identify the concentration of "Coffee Shops" in London Boroughs and identify an optimal location based on the population density of the London Boroughs.

```
Boroughs_venues_coffee = London_onehot_grouped[['Borough', 'Coffee Shop']]
Boroughs_venues_coffee
```

|    | Borough | Coffee Shop |
|----|---------|-------------|
| 0 | Camden | 0.10 |
| 1 | City of London | 0.13 |
| 2 | Hackney | 0.09 |
| 3 | Hammersmith and Fulham [note 4] | 0.08 |
| 4 | Islington | 0.05 |
| 5 | Kensington and Chelsea | 0.01 |
| 6 | Lambeth | 0.08 |
| 7 | Southwark | 0.12 |
| 8 | Tower Hamlets | 0.08 |
| 9 | Wandsworth | 0.07 |
| 10 | Westminster | 0.01 |

- I used elbow method to identify the optimal K for K-Means clustering of "Coffee Shop" in London Boroughs

The Elbow Method showing the optimal k
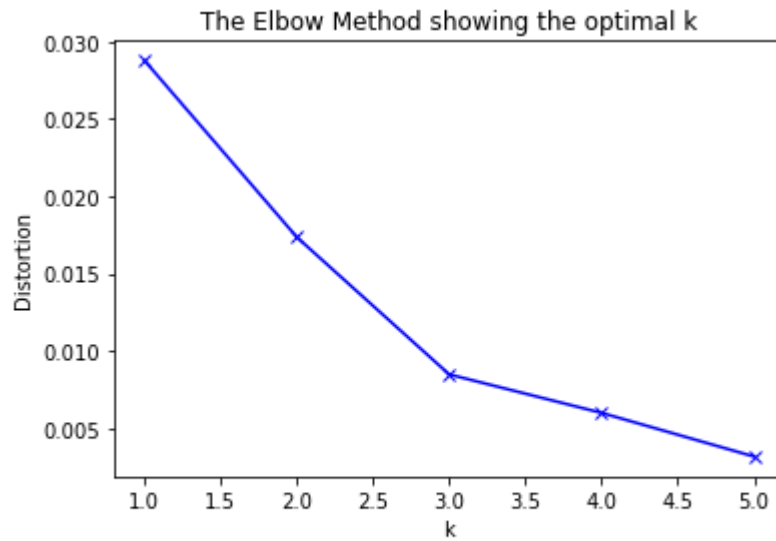


We can see from the above plot that the k elbow appears at 3

- The results from clustering were stored in a new dataframe.

| | Borough | Coffee Shop | Cluster Labels | Local authority | Area (sq mi) | Population (2013 est)[1] | Nr. in map | Longitude | Latitude |
|---|---|---|---|---|---|---|---|---|---|
| 5 | Kensington and Chelsea | 0.01 | 0 | Kensington and Chelsea London Borough Council | 4.68 | 155594 | 3.0 | -0.1947 | 51.5020 |
| 10 | Westminster | 0.01 | 0 | Westminster City Council | 8.29 | 226841 | 2.0 | -0.1372 | 51.4973 |
| 0 | Camden | 0.10 | 1 | Camden London Borough Council | 8.40 | 229719 | 11.0 | -0.1255 | 51.5290 |
| 1 | City of London | 0.13 | 1 | Corporation of London | 1.12 | 7000 | 1.0 | -0.0922 | 51.5155 |
| 7 | Southwark | 0.12 | 1 | Southwark London Borough Council | 11.14 | 298464 | 7.0 | -0.0804 | 51.5035 |
| 2 | Hackney | 0.09 | 2 | Hackney London Borough Council | 7.36 | 257379 | 9.0 | -0.0553 | 51.5450 |
| 3 | Hammersmith and Fulham [note 4] | 0.08 | 2 | Hammersmith and Fulham London Borough Council | 6.33 | 178685 | 4.0 | -0.2339 | 51.4927 |
| 4 | Islington | 0.05 | 2 | Islington London Borough Council | 5.74 | 215667 | 10.0 | -0.1022 | 51.5416 |
| 6 | Lambeth | 0.08 | 2 | Lambeth London Borough Council | 10.36 | 314242 | 6.0 | -0.1163 | 51.4607 |
| 8 | Tower Hamlets | 0.08 | 2 | Tower Hamlets London Borough Council | 7.63 | 272890 | 8.0 | -0.0059 | 51.5099 |
| 9 | Wandsworth | 0.07 | 2 | Wandsworth London Borough Council | 13.23 | 310516 | 5.0 | -0.1910 | 51.4567 |

- The number of Coffee Shop per Borough was sorted in descending order.

| Borough | Coffee Shop | Hotel | Pub | Café | Plaza | Total Top Venues |
|---|---|---|---|---|---|---|
| City of London | 13 | 9 | 1 | 1 | 0 | 24 |
| Southwark | 12 | 5 | 5 | 1 | 1 | 24 |
| Camden | 10 | 6 | 1 | 2 | 2 | 21 |
| Hackney | 9 | 1 | 12 | 5 | 0 | 27 |
| Hammersmith and Fulham [note 4] | 8 | 2 | 10 | 8 | 2 | 30 |
| Lambeth | 8 | 0 | 8 | 4 | 0 | 20 |
| Tower Hamlets | 8 | 7 | 4 | 1 | 3 | 23 |
| Wandsworth | 7 | 0 | 11 | 7 | 0 | 25 |
| Islington | 5 | 0 | 11 | 4 | 0 | 20 |
| Kensington and Chelsea | 1 | 2 | 6 | 5 | 0 | 14 |
| Westminster | 1 | 11 | 0 | 5 | 8 | 25 |

- Based on the above two tables, the cluster results can be interpreted as follows:
  - Cluster 0 – Lowest no. of Coffee Shops but medium population
  - Cluster 1 – Highest no. of Coffee Shops with low to medium population
  - Cluster 2 – Medium no. of Coffee Shops with medium to high average population

- I plotted the London Boroughs population along with Cluster lables on Chloropeth map
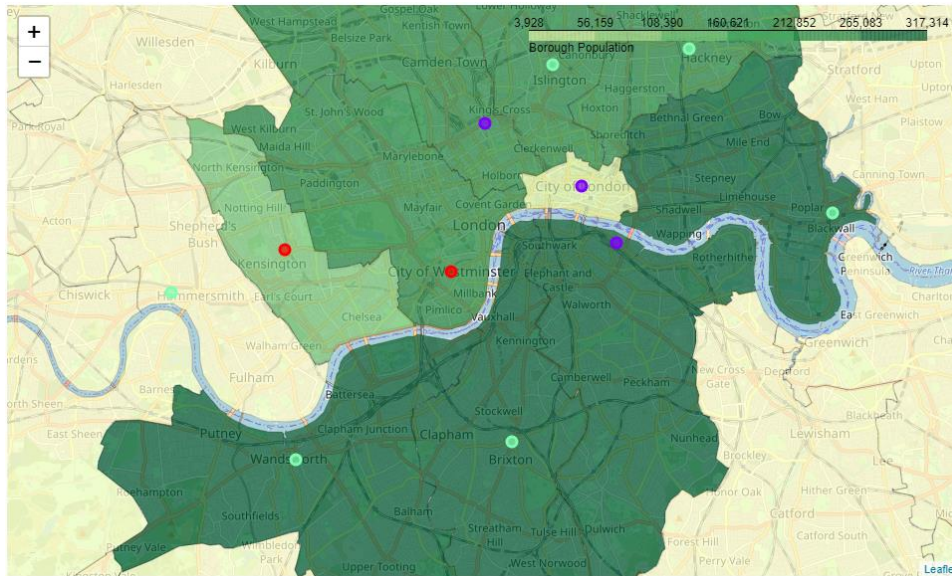


## Conclusion

From the Chloropeth above and the results from Clustering and sorted Coffee Shop dataframe above, the suggested location for opening a cult Coffee Shop would be most appropriate within one of these Boroughs which are highlighted in Dark Green and belong to Cluster 2 i.e. Medium no. of Coffee Shops with medium to high average population – Wandsworth, Lambeth & Tower Hamlets.

However, further detailed analysis should be performed within these identified Boroughs to pinpoint exact location for opening the cult coffee shop. Further analysis could include data points like No. of offices, average house prices, average population income, average population age etc. among other factors that can affect the decision of opening a successful cult coffee shop.