LIVING IN LONDON:
For the Everyday Person

Matthew Chi Siang Wong 12/07/2020

1 Table of Contents

2 Introduction

Living in a city is a dream for many. We grow up watching movies and series based on different cities that eventually form our perception of a "dream life" in a "dream city". For many, we start thinking "when I grow up, I want to live in (insert city name) and become an (insert dream job)" and, this is great. We start working towards our dreams, to live in London or New York or Melbourne or any other city. However, once we get the opportunity to live in this city, our main *question* becomes: **WHERE?**

Let us take London as an example and consider the three following questions/factors:

1. Which part of London do we stay in?
   Do we stay in Central London? What if it's too expensive, what else is there? How about Camden?
2. What do I need on a regular basis?

Is there a supermarket or train station nearby? **These will be our primary focus for this project**

3. What attractions are there?
   Is there a cafe in the area? Is the area lively at night?

Without even considering the price variations across the city, we already have this problem. We can't really know what is around each location (unless we manually look through Google Maps or TripAdvisor) and apart from asking other people or reading reviews online, not everyone knows a quick solution to this. Understanding what each area has or is known for can easily help the "everyday person" make a decision.

This introduction explains the "why" and the following parts will explain the "who", "what" and "how".

3 Target Audience

Our first assumption in this project is that people choose locations to stay in based on the presence of a supermarket or train station nearby. The justifications:

Location	Justification
Supermarket	Applicable to any type of demographic (i.e. families, working professionals, students, etc.)
Train Station	Most people commute around the city through trains, it's could be the London Underground, Overground, DLR, etc.

The obvious rebuttal question would be "how about schools or other amenities?" and that's a valid question. However, to make this applicable to most people, other criteria can't be included. Additionally, some people might not want to stay near restaurants or areas that are known for its night life.

Source of inspiration: https://www.investopedia.com/financial-edge/0410/the-5-factors-of-a-good-location.aspx

4 Objective

This capstone project aims to analyse the different locations in London, United Kingdom and focus on train stations and supermarkets. By segmenting and clustering the locations, we can develop a solution to the question:

If I move to London, where should I start looking at?

5 Data

To solve the problem, we will be using:

- List of areas in London: defines the scope of the project and the areas to consider.
- Latitude and longitude coordinates of these areas: needed to plot the map and get the venue data.
- Venue data: will be used to perform clustering.

5.1 Source of data and method of extraction

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London) contains the list of locations in London and the borough that it is located within. London has a total of 32 boroughs with each containing several locations. The table within the Wikipedia page provides the:

- Location
- Respective London Borough
- Post town
- Postcode District
- Dial Code
- OS grid ref

With the use of several libraries and packages, we will scrape the data from the Wikipedia page and get the coordinates of each location. Subsequently, we will use the Foursquare API to get the venue data for those neighbourhoods. As we are interested in train stations and supermarkets, our clustering of the locations will be based on a dataset that contains/combines both categories.

6 Methodology

Our project notebook is divided into 5 major sections, which consists of the downloading the needed libraries to visualising the final model. The section, respective libraries used and the work completed is detailed below:

6.1 Import libraries

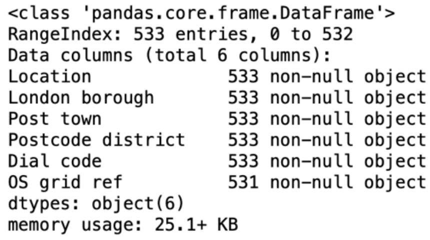
The required libraries were downloaded at the start of the notebook to ease our analyses later on. We included basic libraries such as numpy and pandas along with other specific. The libraries used for obtaining coordinates, map rendering and the k-means clustering were included as well.

6.2 Download and Explore Initials Dataset

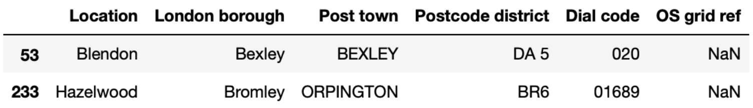
The dataset, as described in section 5.1, was downloaded from the Wikipedia page through pandas. The initial dataset contained 533 rows of locations in London.

Handling Null Values

During the initial check of the different columns, we see that the “OS grid ref” column has two nulls values. The rows with the null values were removed as the locations were in London Boroughs with many other locations and the “OS grid ref” could not be generated.



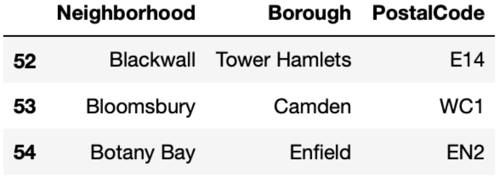
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 533 entries, 0 to 532
Data columns (total 6 columns):
Location              533 non-null object
London borough        533 non-null object
Post town             533 non-null object
Postcode district     533 non-null object
Dial code             533 non-null object
OS grid ref           531 non-null object
dtypes: object(6)
memory usage: 25.1+ KB
```



	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
53	Blendon	Bexley	BEXLEY	DA 5	020	NaN
233	Hazelwood	Bromley	ORPINGTON	BR6	01689	NaN

Cleaning Column Names and Values

From the original dataset, only the “Location”, “London borough” and “Postcode district” were kept. The “OS grid ref” and “Dial code” were removed. Within the three remaining columns, the values were cleaned to remove phrases or numbers within parenthesis or brackets. The columns were then renamed to “Neighborhood”, “Borough” and “PostalCode” respectively.



	Neighborhood	Borough	PostalCode
52	Blackwall	Tower Hamlets	E14
53	Bloomsbury	Camden	WC1
54	Botany Bay	Enfield	EN2

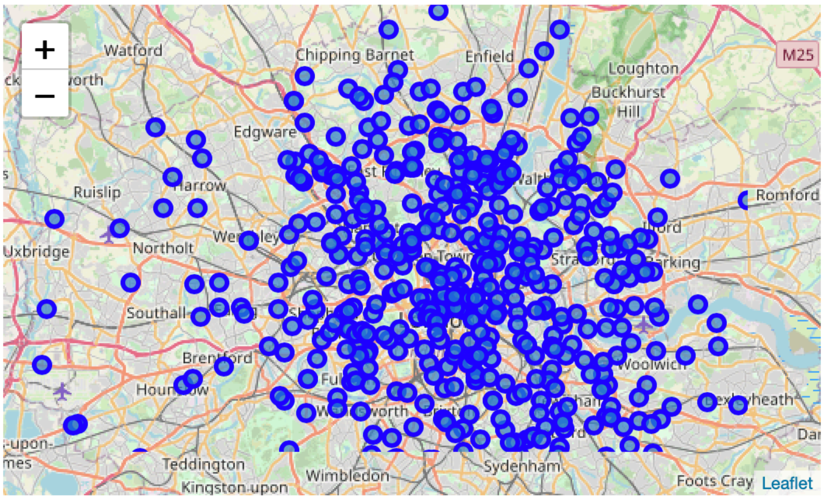
6.3 Get the Geographical coordinates

The coordinates for the respective locations were obtained through the geocoder library. Coordinates include the latitudinal and longitudinal positioning. These were merged with the dataset.

6.4 Explore and Cluster Neighborhoods

Explore locations in London through visualisations
We used the geopy library get the latitude and longitude coordinates of London, UK as a whole. A map was then created with folium and was added with markers of the coordinates from the original dataset. The blue markers show all the different coordinates that are considered part of London.

Observation: London is concentrated in the centre. Locations in the centre have a smaller coverage area than those further out of the city.



Explore the locations with Foursquare API
The application programming interface (API) in allows us to explore the neighbourhoods and coordinates that we have in the dataset. In this project, we are using the explore functionality to see *what’s in the area*, with a special focus on the venue’s:

- name,
- latitudinal location,
- longitudinal location, and
- category.

The only things that we need to input into the API as prerequisite is our client ID, client secret, version, radius from venue and limit of results. The results of this section the mean frequency information for each location, e.g at Abbey Wood, the mean of the frequency at each category is 0.143.

6.5 Analysis of Supermarkets and Train Station

For this process, we used the k-means clustering methods. The model divides the data into non-overlapping subsets (clusters) without any cluster-internal structure. It aims to minimize the "intra cluster" distances and maximize the "inter-cluster" distances (e.g. Euclidian, etc.). The steps are better visualised with numbering:

1. We combine our results "Supermarket" and "Train Station"
2. Select an initial 5 clusters.
3. Run k-means clustering on the combined results.
4. Combine the prediction from step 3 with the original dataset.

Additionally, we obtained the 10 most common venues in each location and added this to the final dataset. This would benefit users who wanted to see:

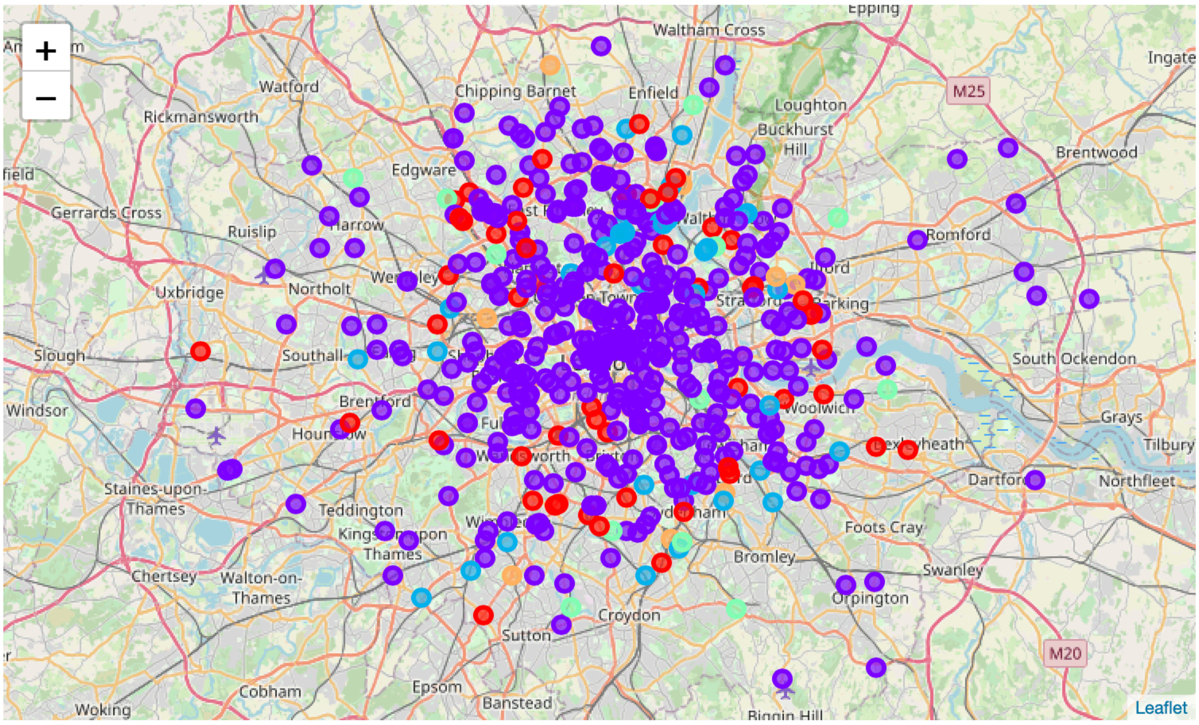
- The location with supermarkets and train stations, and
- What other locations are in the area.

6.6 Results

Our results are displayed through the map and seeing the separation of clusters.

7 Results

By visualising the map below, we can see that the purple cluster (cluster 1) is the biggest. Following up would be the red (cluster 0), blue (cluster 2), green (cluster 3) and orange (cluster 4).



However, this might not be very intuitive, as such we can refer to the tables produced. The compilation of results are:

Clusters	Size	Average Score within

		Supermarkets	Train Stations
0	57	0.09	0.00
1	411	0.01	0.00
2	38	0.02	0.09
3	15	0.27	0.03
4	8	0.00	0.25

As such we can see that cluster 3 is the best for supermarkets (and overall as well) and train stations are most present in cluster 4. Cluster 3 includes locations like Croydon, Edgware, etc.

8 Discussion and Limitations

8.1 Benefits

What's best: Based on the results, we can use this to make a more informed decision on where should we consider staying in. The chances of finding a place to stay with accessibility to supermarkets or train stations in cluster 3 are relatively high. However, if we wanted to stay in a place that has more accessibility to train stations, cluster 4 would be a better pick.

What's worst: This representation also shows us which locations we should avoid. Looking at cluster 1 (the purple dots), we can see that a few of the locations furthest away from the city have the little supermarkets or train stations.

What's in-between: We could also consider places in cluster 2, that have lesser chances of having a supermarket or train station nearby but is mostly in Central London. A person who stays and works directly in the city might have other considerations that are more important. This model can be adapted to individual users' preference.

8.2 Limitations and Future Possibilities

While the project has its benefits, its biggest limitation is the selection of information which is based on supermarkets and train stations only. This shows what's most prevalent in each location. However, it does not mean that central London does not have supermarkets or train stations, the concentration of other venues are just a lot higher in Central London.

As such, another factor to consider is rent and the persons spending ability. The question being "how much would a person pay based on the availability of other nearby venues?"

9 Conclusion

In this project, we have experienced the data science methodology. From understanding the business problem (or in this case, the individual benefits), specifying the data required, extracting and preparing the data, clustering the data through the k-Mean method and provided an evaluation of the model. The answer for the initial question of where to stay/consider when moving to London is to consider locations in cluster 3. Furthermore, we now know to avoid certain locations from cluster 1 and could possibly consider cluster 2. However, this recommendation is based on the assumption that supermarkets and train stations are the main priority. This methodology can be adapted to consider other venues like cafes, shopping centres, etc. It really depends on the user's demands.