Capstone Project - The Battle of the Neighbourhoods (Part2/Week2)

Applied Data Science Capstone by IBM/Coursera

Table of contents

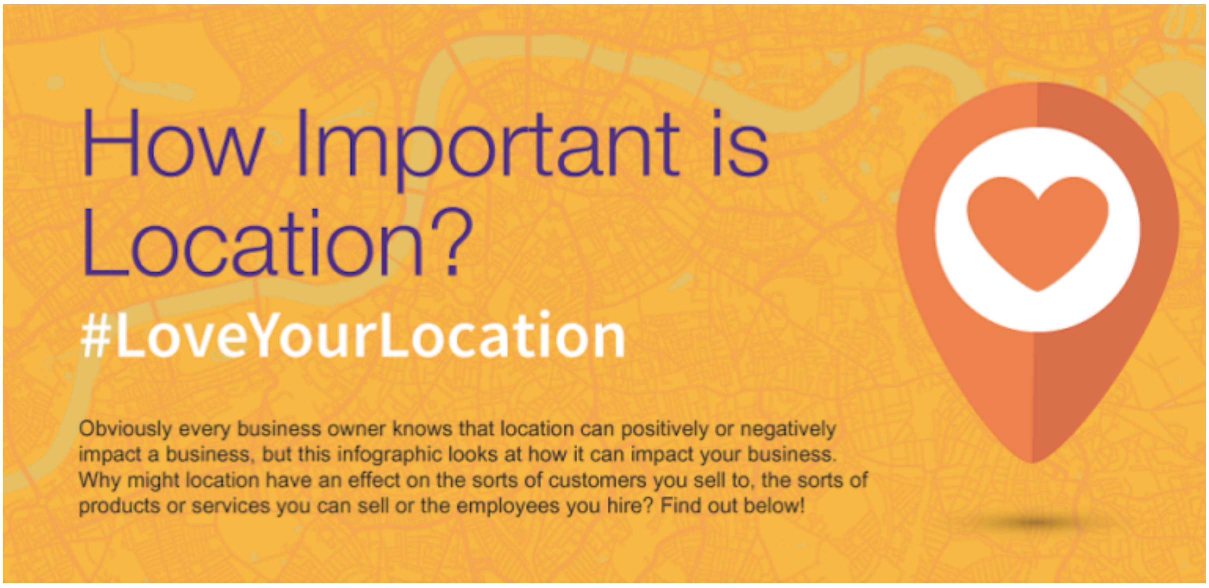
1. Introduction: Business Problem

In this project, we will try to find similarity within the city's neighbourhoods, if a user or agent wants to move or shift from his current city locality to another city. The target audience or the stakeholders can be either an individual looking to relocate or a business contractor who wants to start or expand his business to new city. This project will focus on 3 cities **London, Sydney and Singapore**. It's all about **understanding the locality**

Moving or establishing to a new city is always exciting. However, with the excitement comes difficulties. Sometimes, essentials such as food, transportation, water, accommodation, competitors, supplies etc. need more focus and planning. Also comes the **Cultural Shocks**.

We will use our data science powers to generate a few most promising neighbourhoods based on the stakeholders criteria, **recommend them a cluster of Neighbourhoods with desired venues based on their choice of city**.

Importance of House



Obviously every business owner knows that location can positively or negatively impact a business, but this infographic looks at how it can impact your business. Why might location have an effect on the sorts of customers you sell to, the sorts of products or services you can sell or the employees you hire? Find out below!

Importance of Business Location

2. Data

Business problem defines the factors that will influence our decisions and recommendations which are:

- Cities with business facilities
- Location within the city closer to the centre of the city
- Location co-ordinates matching the cities co-ordinates, as many areas or Borough exist in multiple cities.

Dataset used for the projects are collected to generate the required information are:

- **Geoname Dataset** for countries of **GB(Great Britain), AU(Australia), SG(Singapore)**, from the website www.geonames.org.
- Various venues and their type and location in every neighbourhood within each city will be obtained using **Foursquare API**

The above datasets will be used to create city based maps and find the neighborhood locations within the city using latitude and longitude co-ordinates. Foursquare API will give the types of venues for each neighborhoods so that they can be clustered among themselves to find related neighborhoods and recommend neighborhoods.

3. Methodology

We use python libraries like, **numpy, pandas** to perform data exploration and numeric operations on the data which we extracted from geoname website for the countries where the cities belong to. **Matplotlib** to plots graphical viz., **folium** to visualize geographic details such as maps of the cities, **geopy.geocoders** to get the geographic co-ordinates. **sklearn** to bring in the unsupervised machine learning technique of Clustering using the method of KMeans, to cluster the neighborhoods based on the common venues identified by **FourSquare API**.

3.1. Data Exploration

We load 3 cities data into pandas Dataframe, one after another into a dataframe name by the city name itself. Each city is taken from the country csv files, from which the Boroughs are selected. Our dataset includes columns as country, postal code, neighborhoods, Country name, Boroughs/Areas, and co-ordinates.

Data Cleaning

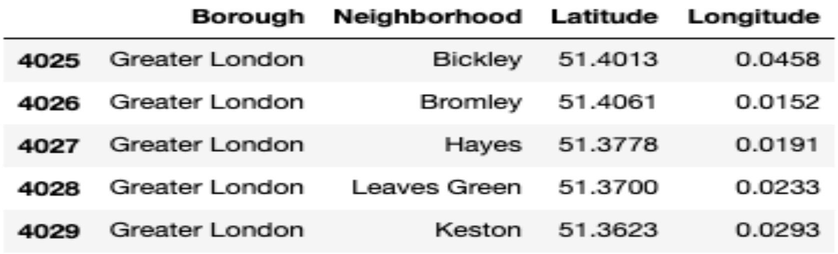
We drop the columns from the dataset we downloaded from geoname, Drop the duplicates neighborhoods after selecting the City as London, Sydney and Singapore to get uniques neighborhoods. We get the clean data in the form of **Borough, Neighborhood, Latitude, Longitude**.

Storage

We use GitHub repository as a database to store all our codes, data, images, reports.

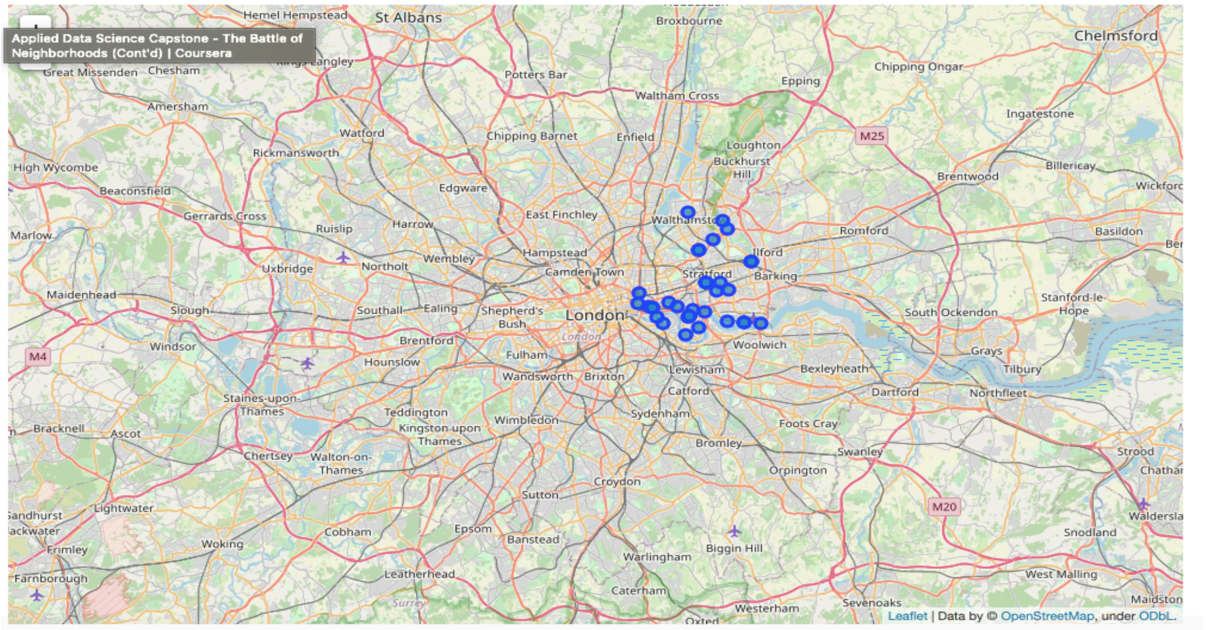
Click here to GitHub link

London Dataset after Cleaning

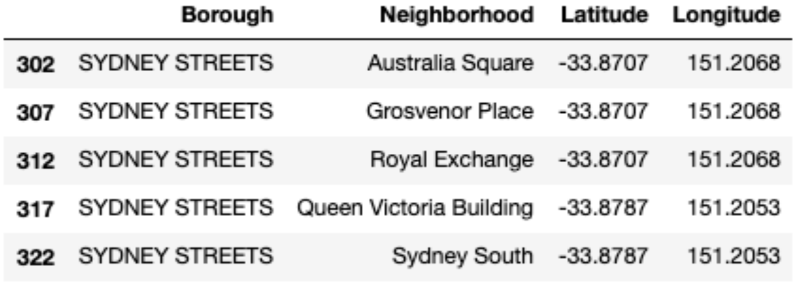


	Borough	Neighborhood	Latitude	Longitude
4025	Greater London	Bickley	51.4013	0.0458
4026	Greater London	Bromley	51.4061	0.0152
4027	Greater London	Hayes	51.3778	0.0191
4028	Greater London	Leaves Green	51.3700	0.0233
4029	Greater London	Keston	51.3623	0.0293

Using data with the Folium library to explore the neighborhoods in Greater London map.

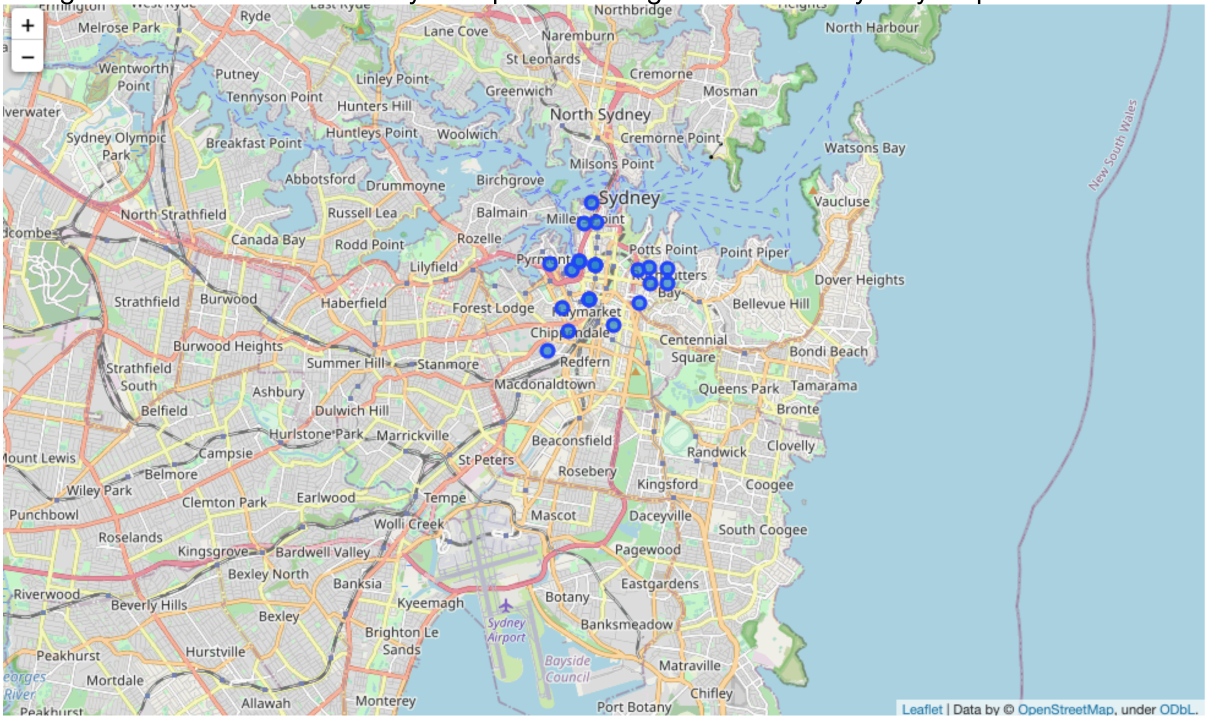


Sydney Dataset after Cleaning

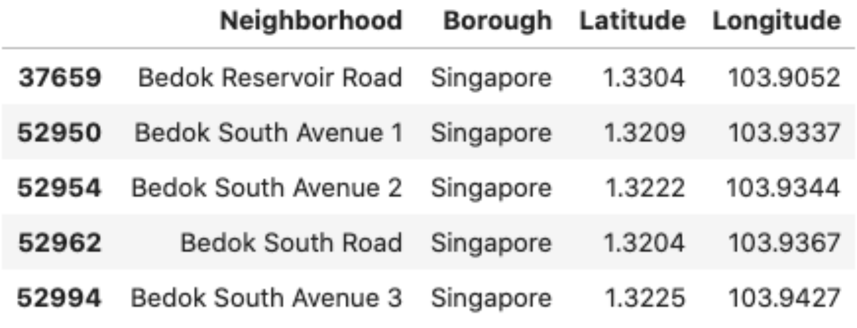


	Borough	Neighborhood	Latitude	Longitude
302	SYDNEY STREETS	Australia Square	-33.8707	151.2068
307	SYDNEY STREETS	Grosvenor Place	-33.8707	151.2068
312	SYDNEY STREETS	Royal Exchange	-33.8707	151.2068
317	SYDNEY STREETS	Queen Victoria Building	-33.8787	151.2053
322	SYDNEY STREETS	Sydney South	-33.8787	151.2053

Using data with the Folium library to explore the neighborhoods in Sydney map.

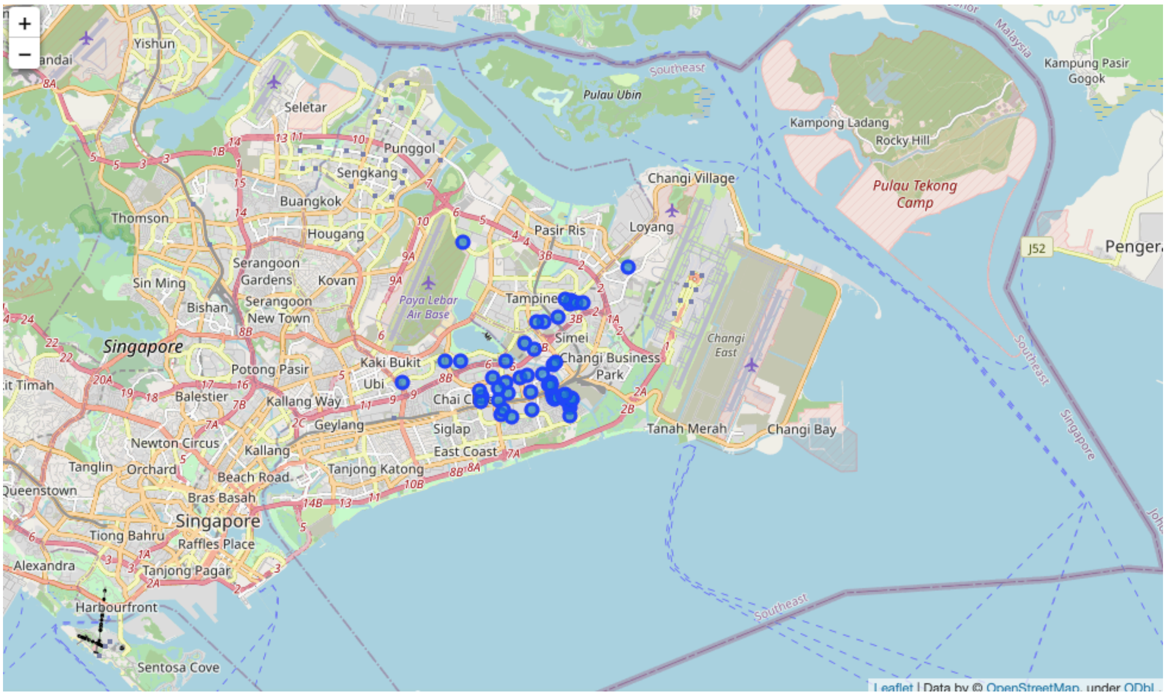


Singapore Dataset after Cleaning



	Neighborhood	Borough	Latitude	Longitude
37659	Bedok Reservoir Road	Singapore	1.3304	103.9052
52950	Bedok South Avenue 1	Singapore	1.3209	103.9337
52954	Bedok South Avenue 2	Singapore	1.3222	103.9344
52962	Bedok South Road	Singapore	1.3204	103.9367
52994	Bedok South Avenue 3	Singapore	1.3225	103.9427

Using data with the Folium library to explore the neighborhoods in Singapore city map.

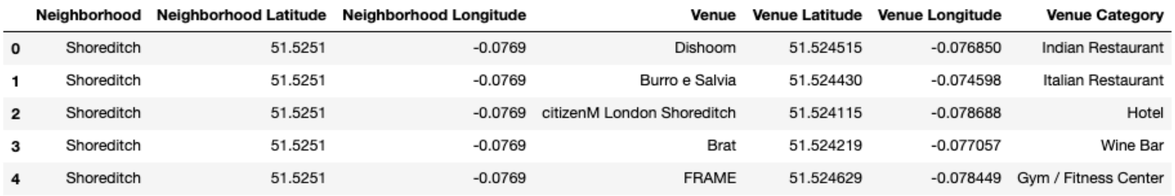


3.2. Defining FourSquare credentials:

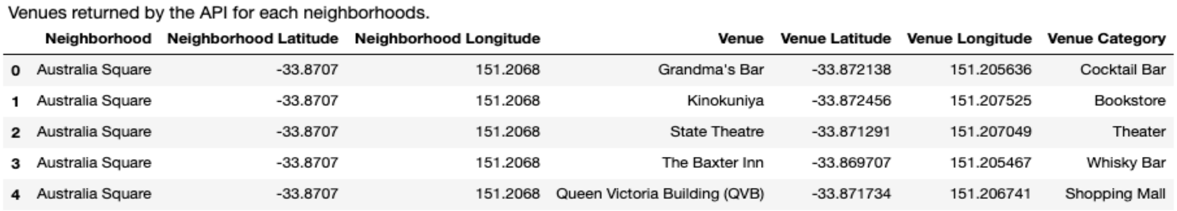
Foursquare API used below to explore the neighborhoods in the city. I have used the limit as 50 venue and the radius 500 meter for each neighborhoods to get the venues from their given latitude and longitude information.

url='https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, latitude, longitude, radius, LIMIT)

API response for London, Sydney and Singapore city:

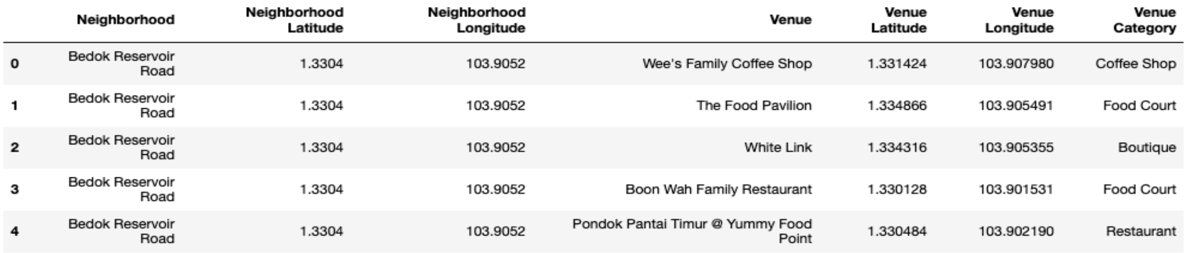


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Shoreditch	51.5251	-0.0769	Dishoom	51.524515	-0.076850	Indian Restaurant
1	Shoreditch	51.5251	-0.0769	Burro e Salvia	51.524430	-0.074598	Italian Restaurant
2	Shoreditch	51.5251	-0.0769	citizenM London Shoreditch	51.524115	-0.078688	Hotel
3	Shoreditch	51.5251	-0.0769	Brat	51.524219	-0.077057	Wine Bar
4	Shoreditch	51.5251	-0.0769	FRAME	51.524629	-0.078449	Gym / Fitness Center



Venues returned by the API for each neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Australia Square	-33.8707	151.2068	Grandma's Bar	-33.872138	151.205636	Cocktail Bar
1	Australia Square	-33.8707	151.2068	Kinokuniya	-33.872456	151.207525	Bookstore
2	Australia Square	-33.8707	151.2068	State Theatre	-33.871291	151.207049	Theater
3	Australia Square	-33.8707	151.2068	The Baxter Inn	-33.869707	151.205467	Whisky Bar
4	Australia Square	-33.8707	151.2068	Queen Victoria Building (QVB)	-33.871734	151.206741	Shopping Mall



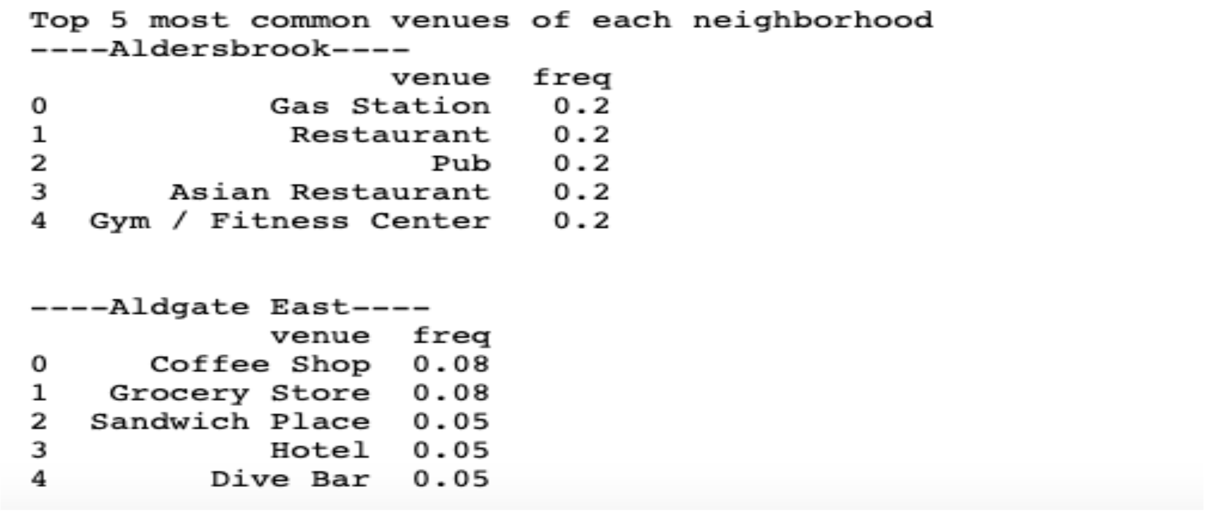
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bedok Reservoir Road	1.3304	103.9052	Wee's Family Coffee Shop	1.331424	103.907980	Coffee Shop
1	Bedok Reservoir Road	1.3304	103.9052	The Food Pavilion	1.334866	103.905491	Food Court
2	Bedok Reservoir Road	1.3304	103.9052	White Link	1.334316	103.905355	Boutique
3	Bedok Reservoir Road	1.3304	103.9052	Boon Wah Family Restaurant	1.330128	103.901531	Food Court
4	Bedok Reservoir Road	1.3304	103.9052	Pondok Pantai Timur @ Yummy Food Point	1.330484	103.902190	Restaurant

API resulted in unique venue categories for each city:
There are 164 unique categories for London city.
There are 134 unique categories for Sydney city.
There are 98 unique categories for Singapore city.

4. Analysis using ML

London city neighborhood analysis and Clustering of Neighborhoods using K-Means

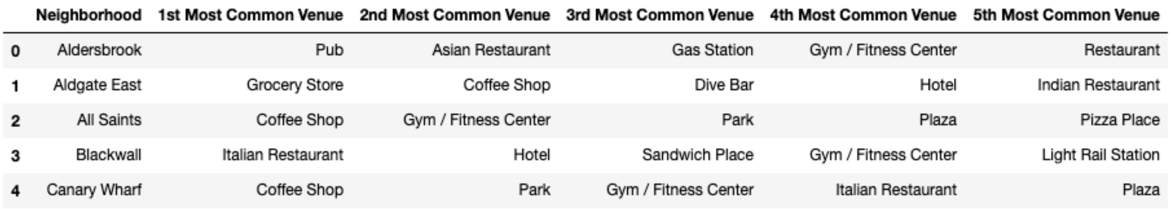
London City is analysed first by sorting every neighborhood with top 5 common venues within them.



```
Top 5 most common venues of each neighborhood
----Aldersbrook----
                 venue  freq
0          Gas Station   0.2
1           Restaurant   0.2
2                  Pub   0.2
3      Asian Restaurant   0.2
4  Gym / Fitness Center   0.2


----Aldgate East----
           venue  freq
0    Coffee Shop  0.08
1  Grocery Store  0.08
2 Sandwich Place  0.05
3          Hotel  0.05
4       Dive Bar  0.05
```

Venues are sorted as per most common venues

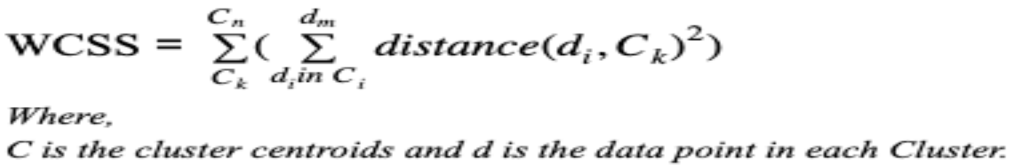


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Aldersbrook	Pub	Asian Restaurant	Gas Station	Gym / Fitness Center	Restaurant
1	Aldgate East	Grocery Store	Coffee Shop	Dive Bar	Hotel	Indian Restaurant
2	All Saints	Coffee Shop	Gym / Fitness Center	Park	Plaza	Pizza Place
3	Blackwall	Italian Restaurant	Hotel	Sandwich Place	Gym / Fitness Center	Light Rail Station
4	Canary Wharf	Coffee Shop	Park	Gym / Fitness Center	Italian Restaurant	Plaza

Finding the optimum K value for the London data using WCSS method:

An ideal way to figure out the right number of clusters would be to calculate the **Within-Cluster-Sum-of-Squares (WCSS)**.

WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids.



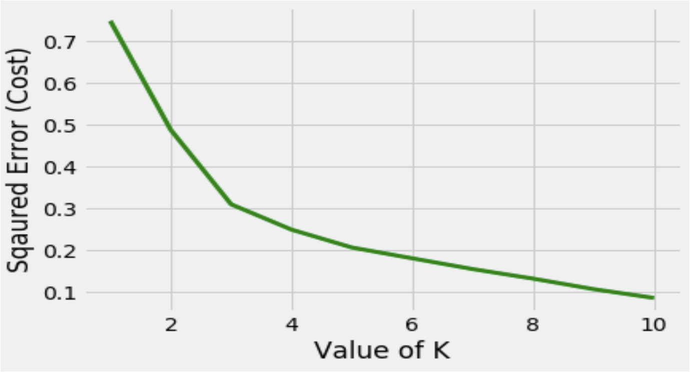
$$\text{WCSS} = \sum_{C_k}^{C_n}\left(\sum_{d_i\,in\,C_i}^{d_m} distance(d_i, C_k)^2\right)$$

Where,
C is the cluster centroids and d is the data point in each Cluster.

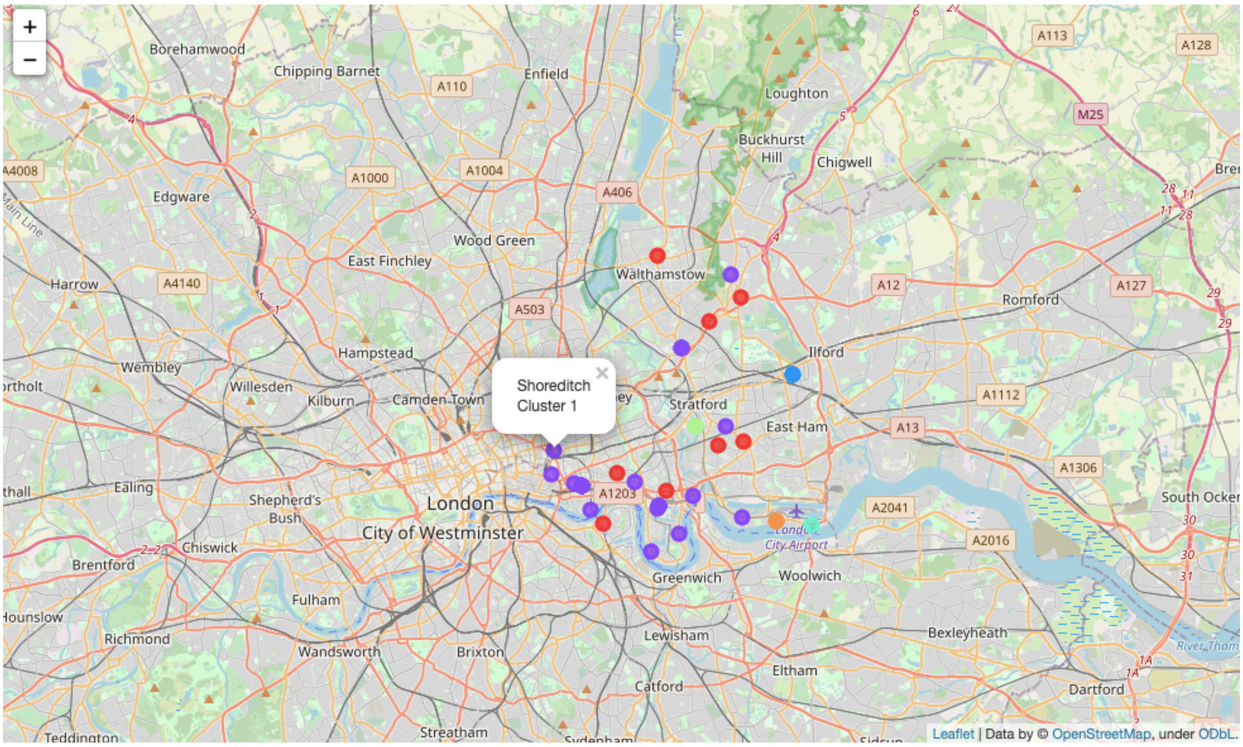
The idea is to minimise the sum. Suppose there are n observation in a given dataset and we specify n number of clusters (k = n) then WCSS will become zero since data points themselves will act as centroids and the distance will be zero and ideally this forms a perfect cluster, however this doesn't make any sense as we have as many clusters as the observations. Thus there exists a threshold value for K which we can find using the Elbow point graph.

Elbow method We can find the optimum value for K using an Elbow point graph. We randomly initialise the K-Means algorithm for a range of K values and will plot it against the WCSS for each K value, with an increase in the number of clusters the WCSS value decreases. We select the value for K on the basis of the rate of decrease in WCSS.

E.g. of WCSS curve



KMeans clustering creates clusters and using Folium library we have created the map below with all the clusters of neighborhoods in the data for the city of London.



Similarly, the steps are repeated for the city of Sydney and Singapore in the analysis.

5. Results

Recommending user some of the desired neighborhoods based on the stakeholders preferences

Let’s take an example by taking user input, and say user passes a list of top 3 venues he thinks is nearby his locality before he/she moves into the location or set up a business. User input is taken as a form of a list of desired venues, and in return we can recommend the user to check the neighborhoods, also we can highlight the cluster number if the user wants to know more about neighboring venues and neighborhoods.

Purpose of this project was to only provide information on areas close to cities center and residential places. Recommended neighborhoods should be considered as a reference which can be further more analysed or grouped with other places to see if the desired location to shift or setup a business is more ideal.

6. Discussion & Further Recommendation

We can try with **DBSCAN clustering** to avoid outliers, as kmeans is very prone to be biased due to presence of outliers. There is further more analysis where we can consider more number of data points or locations with advance Foursqaure API plan which allows more API hits quota. Which then has multiple outliers some locations are far away from a cluster and KMeans is biased to such outliers and its then when we can use DBSCAN and also KMedians clustering.

Be it any city the neighborhoods are clustered either with coffee houses, cafes and hotels together, or with residentials, bus stations, bus lines, or with pubs parks, there are clusters specifically identifying vicinities closer to Airport and airport services. FourSquare API provides venues in detailed manner, when it comes to Restaurants where neighborhoods are clustered with cuisine based restaurants, such that the user can search for the locality with specific cuisine restaurants like Indian, Thai, Chinese etc. Clusters are also made up of sandwich plaza, pizza plaza.

Further improvement can be made on this project, by using more neighborhoods and boroughs without any limitations of using Sandbox version of FourSquare API, to get additional quotas. Cities considered here are mainly keeping in mind the business advantages this cities provide based on geographic locations like Europe(London), APAC(Singapore), ANZ(Sydney), more cities can be compared and considered.

7. Conclusion

The purpose of this project was to find similarity within the city's neighborhoods for stakeholders to decide which location suits their preferences,

If a user or agent wants to move or shift from his current city locality to another city, this is to serve the purpose of the stakeholder to decide whether to shift to which part of the city or which different city he/she wants to move. The target audience or the stakeholders can either be an individual looking to relocate or a business contractor who wants to start or expand his business to new city.

The project considers only on 3 cities London, Sydney and Singapore, but as mentioned before more cities can be considered with more number of areas/boroughs.

London: Good foundation, of course, is good for business. In London, it's easy to provide personal service to millions of potential customers within the city itself, across the U.K. and around the world. London's geographical positioning is uniquely advantageous. A large number of people move to London in their 20s, drawn from all corners of the country. This is because of the range and number of job opportunities that the capital offers.

Sydney: Sydney possesses a range of important economic advantages and attributes, including strong clusters of high value industries and closely-linked sectors, international connections and a highly skilled pool of workers attracted to Sydney's unique lifestyle. Sydney — a coastal metropolis whose five million residents make it the largest city in Australia — is famous for many things. The glittering harbour, complemented by landmarks like the Sydney Opera House and the Sydney Harbour Bridge. The dozens of sun-bathed beaches peppering the spectacular coastline

Singapore: Singapore has a well-developed free market economy that is based primarily on trade, finance and manufacturing. Services account for 75% of the country's GDP and employ 80% of its workforce. The country has managed to achieve very low unemployment rate while maintaining low inflation. With its high standard of living, business friendly climate and the ability to live and work in English, the city-state of Singapore is a very popular destination for both immigrants and expats.

Locations like **New York, San Francisco, Toronto, Quebec, Bangalore, Mumbai, Amsterdam, Barcelona, Shanghai, Tokyo** and many more can be considered into this analysis.

Thank You!!