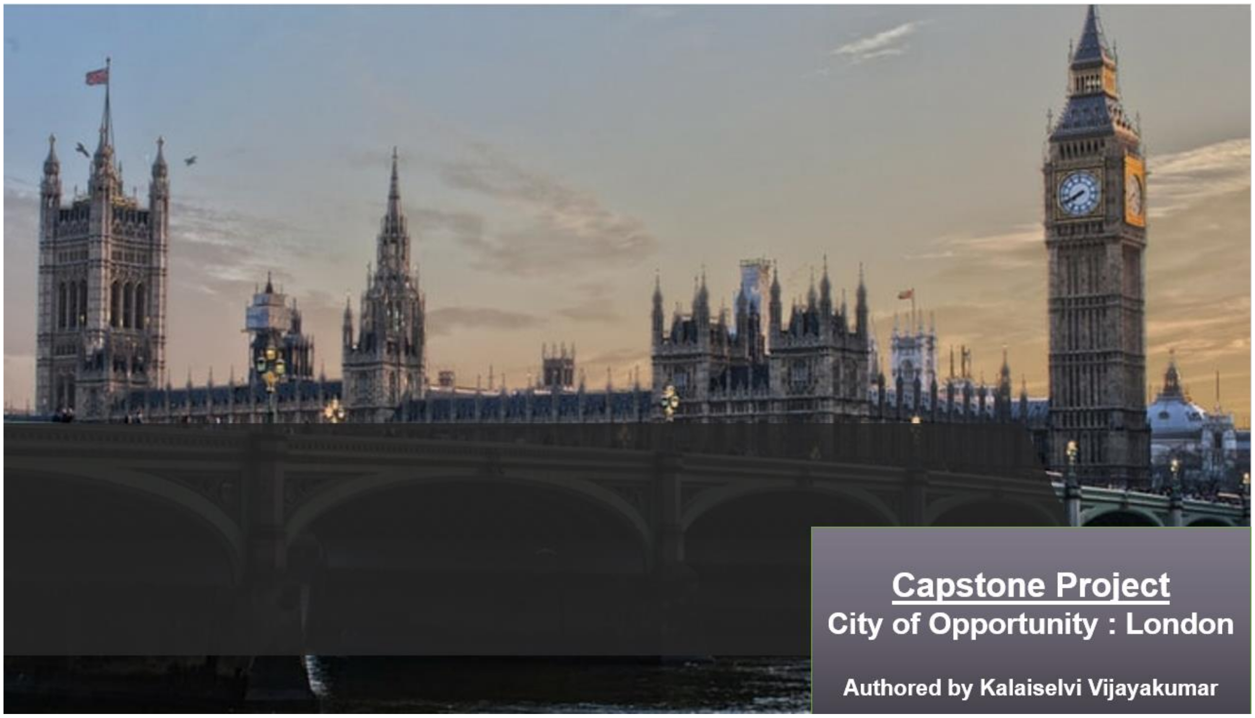

Capstone Project
City of Opportunity : London
Authored by Kalaiselvi Vijayakumar

IBM Applied Data Science Capstone

City of Opportunity: London

Introduction:

London is definitely one of the best places to start a business. Even though it has high competition it still offers plenty of opportunities if you have a lucrative idea. Being the capital of both England and the UK, it is the center for commerce, finance, fashion, healthcare, research, development, business, tourism, and plenty more. As such there are many opportunities for people in London that want to start a business. London is also a very expensive place to live.

Many businesses want to open a branch here and are ready to pay highest price for their accommodations. With a huge competition and high business cost, it will be very difficult for the newcomers to find a perfect place with less expensive to start their business. It is not possible to find a good place based on their need in the city limits of London. Let's talk about the business problems to choose the best possible option for newcomers to start their business in London boroughs.

Business Problem:

The questions that I aim to answer in this project are the following:

- Who want to start their own business and cannot really afford to open in the City yet?
- Where is it best to open a new place?
- Where will it be cheapest and will have enough people living around to be popular?
- Where the competition is not too overwhelming?

Target Audience:

The purpose of this project is to help people in exploring better area around their neighborhood to start their business. It will help people making a smart and efficient decision on selecting great area out of numbers of other areas in London outer city limits. This analysis will be of interest to the following groups:

- First time entrepreneurs, who want to start their first business. Below dataset will give a comprehensive insight into where best to open a new venue, to maximize the value for money.
- People who already run a business and want to start a new branch. Given the extra information, it may provide some valuable information before decision making.

Data Description:

For this project we need the following data:

- List of London Boroughs with co-ordinates from Wikipedia

- The Most popular venues in the respective boroughs from Foursquare data

- Online based data on rent in London boroughs.

Using this data will allow exploration and examination to answer the questions. This is a project that will make use of many data science skills, from web scraping (wikipedia.com), working with API (Foursquare), data cleaning, data wrangling and map visualization (Folium) and to machine learning (K-means clustering).

Methodology:

Data Exploration -

Firstly, we need to get the list of boroughs in London. Fortunately, the list is available on the web page (https://en.wikipedia.org/wiki/List_of_London_boroughs). We need to do web scraping using Python requests to extract the list. Once the dataset of London boroughs has been downloaded, we must edit the dataset provided to only have information, necessary for our problem. Wikipedia provided information on political situation, headquarters of the borough council etc. that will not be required. After cleaning, we will only be left with Name, Area, Population, Coordinates and rent prices for each borough. Since the dataset will only include outer boroughs, all the inner ones will be omitted as well.

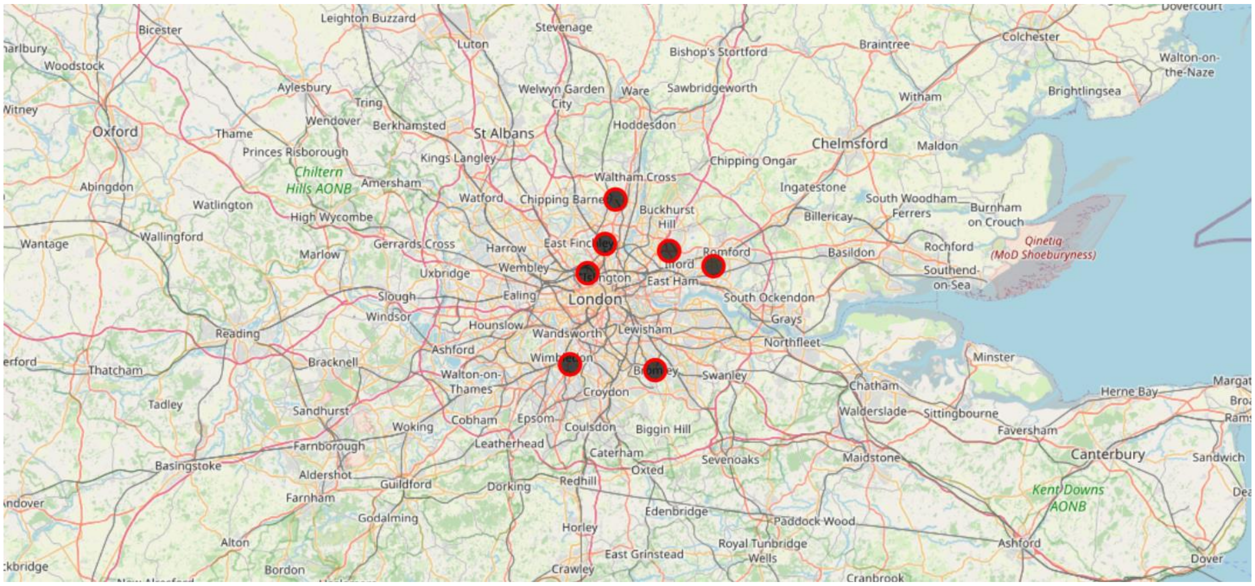
Foursquare provides a dataset of venues around the specific coordinates or venues, if we use the "Explore" function in the Developer tab. Once requested, we get a full breakdown of all recorder venues around the boroughs of interest.

Data Geocoding -

We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert the address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame.

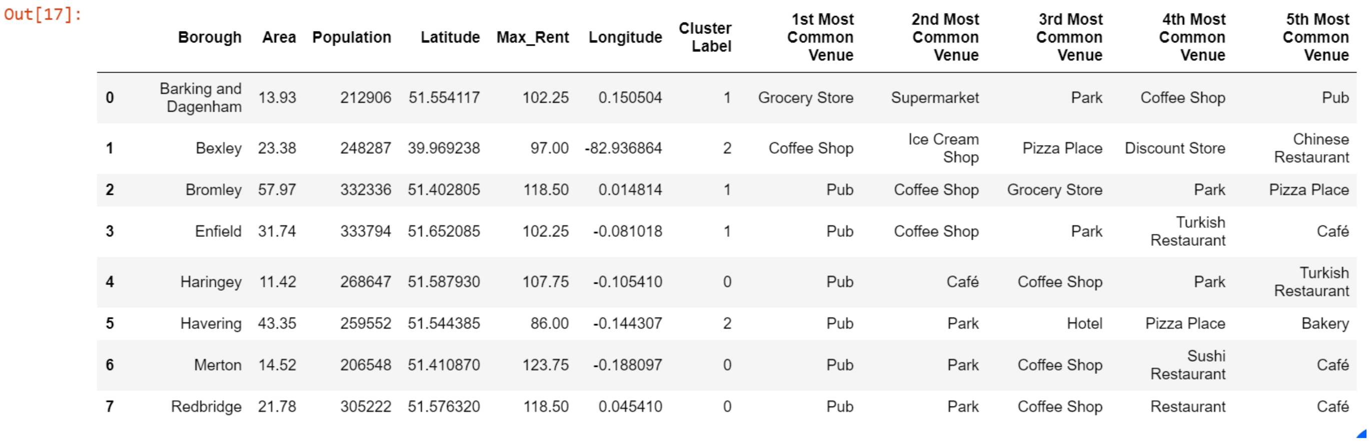
Data Visualization -

Visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of London.



Feature selection -

For convenience, we transform the dataset only to show top 5 places to work with. After that, we merge the data frames together for a comprehensive set of values, worth analyzing.



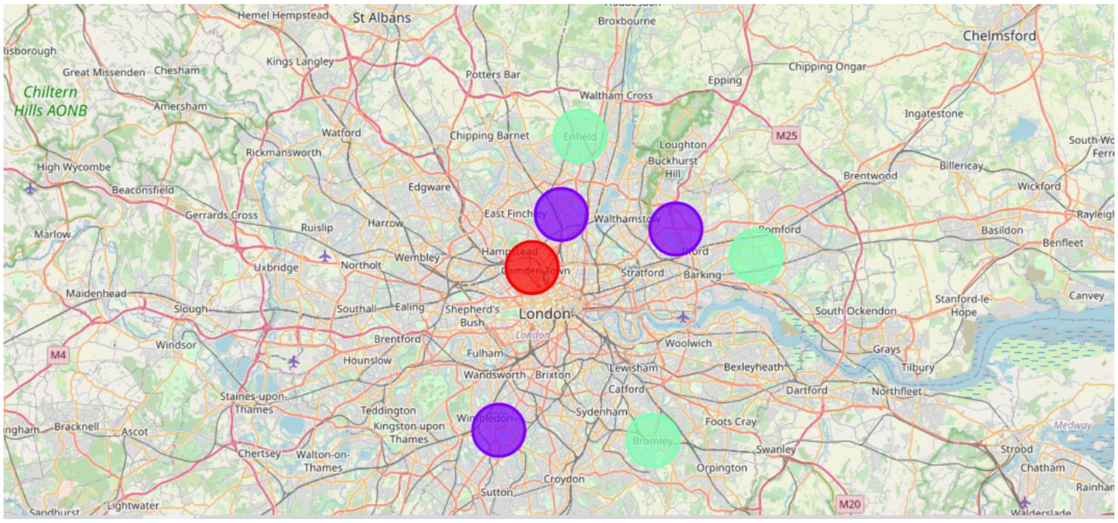
Out[17]:

	Borough	Area	Population	Latitude	Max_Rent	Longitude	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Barking and Dagenham	13.93	212906	51.554117	102.25	0.150504	1	Grocery Store	Supermarket	Park	Coffee Shop	Pub
1	Bexley	23.38	248287	39.969238	97.00	-82.936864	2	Coffee Shop	Ice Cream Shop	Pizza Place	Discount Store	Chinese Restaurant
2	Bromley	57.97	332336	51.402805	118.50	0.014814	1	Pub	Coffee Shop	Grocery Store	Park	Pizza Place
3	Enfield	31.74	333794	51.652085	102.25	-0.081018	1	Pub	Coffee Shop	Park	Turkish Restaurant	Café
4	Haringey	11.42	268647	51.587930	107.75	-0.105410	0	Pub	Café	Coffee Shop	Park	Turkish Restaurant
5	Havering	43.35	259552	51.544385	86.00	-0.144307	2	Pub	Park	Hotel	Pizza Place	Bakery
6	Merton	14.52	206548	51.410870	123.75	-0.188097	0	Pub	Park	Coffee Shop	Sushi Restaurant	Café
7	Redbridge	21.78	305222	51.576320	118.50	0.045410	0	Pub	Park	Coffee Shop	Restaurant	Café

Data Clustering -

Finally, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

Once the boroughs are selected, we will place them onto the map and cluster the boroughs to analyze the similar ones against each other and within the clusters themselves. Each of the clusters will be compared against the popularity of the venues, worth considering for a new business venture, as well as the hospitability of the venture climate. Depending on the area, population and rent, each of the clusters offer different advantages and disadvantages in terms of venue choices.



Results:

Reviewing each cluster as per method above, we learn:

- **Cluster 0**: Havering and Bexley are good fits to our target boroughs: both have entertainment venues as most common, both rather big and quite populated. In this situation perhaps Havering has a slight financial edge - cheaper rent a more incline to pubs and coffee shops in the area.
- **Cluster 1:** This cluster is heavy favoring pubs – predominantly as centers of socializing of the local communities. First cluster boroughs have majority of people living further from the central London, keeping to the old ways of pub socializing.
- **Cluster 2:** Most of the boroughs are not heavy drinkers, but neither are they very outgoing. Because I would not call visit to the grocery store as a social route! Enfield and Bromley are bigger outer London boroughs, and the data, that we get can differ massively within the same borough: those living closer to M25 in the suburban houses tend to stay indoors and spend more time with their families and cook at home. On the other hand, those living closer to the inner borough limits may be more outgoing, and because of this contribute the most to our dataset. Regardless, with not too high rent, big and a very diverse population, Enfield and Bromley are worth considering.
- Within top 5 places of interest in every borough is an ethnic restaurant. Because of the different ethnicities in the boroughs, some choices will be more favorable amongst the specific group in the area.
- Rent price is not so much a factor for going out - the demand is not affected by difference in costs. There is a spike in rent price going into London, but further away the cost is not too much of an issue.

Discussion:

(Based to constraint on API calls and search radius, the true result might vary.)

Looking at the data, Havering, Bromley, Bexley and Enfield are the best places outside of Central London where a new venue is worth opening. However, a lot of information is not considered, and cannot be obtained from Foursquare Developer:

- Higher ethnic presence in each borough can and will influence the popularity of a given cuisine.
- Closer proximity to Inner boroughs and better transport links allows people to travel to the neighboring borough and impact the measurements.
- Many small venues are not registered in Foursquare and are marketed via word-of-mouth and are not considered.

Regardless, the analysis provided an insight into what people like and opt for, when it comes to going out in their own neighborhoods.

Conclusion:

To conclude this project, I have gone through the process of identifying the business problems, specifying the data required, extracting and preparing the data, visualizing the results, performing machine learning by clustering the data into 3 clusters based on their frequency similarities, tackling and reaching to a definitive solution to business problems (mentioned in results).

Lastly, I have had a good trial run at solving a real-life problem, using available data to find a business solution - choosing to open a venue in London. I have made use of some frequently used python libraries to manipulate data, use Foursquare API to explore the information on the Boroughs I looked into and managed to make a map of results, that allowed me to illustrate my point graphicly and quite clearly to someone, not familiar with data manipulation and who only wants to know one thing - where will my venue be booming?