

Business Problem

London is one of the largest metropolises in the world where over **9 million** people live and it has a population density of **1510** people per square kilometer. The city is divided into 32 boroughs in total.

As you can see from the figures, London is a city with a high population and population density. Being such a crowded city leads the owners of shops and social sharing places in the city where the population is dense. When we think of it by the investor, we expect from them to prefer the districts where there is a lower real estate cost and the type of business they want to install is less intense.

This project primarily deals with exploration and segmentation of neighborhoods in the city of London,UK.

I'm targeting investors and/or potential individual owners who are looking to invest in a restaurant business.

By means of this project, I'll be focusing on aspects such as property price, population density and cuisine to facilitate the decision making process of our customer.

Using various data sources from the internet and by exploiting the location data facility of the foursquare API, I'll begin by helping the customer narrow down his search for an appropriate location by considering the boroughs of London and discriminating on the basis of average property price and population density.

Once I have finalized the borough, I'll explore its neighborhoods using the foursquare API to provide insights into what type of cuisine the residents typically enjoy, using which , I'll cluster the neighborhoods into different categories. Our customer can then refer to these clusters which might help him decide what type of cuisine they would want their restaurant to serve and in which neighborhood they'd like to establish their business.

Data

Initially, the table of London Boroughs would be constructed by scraping the webpage https://en.wikipedia.org/wiki/List_of_London_boroughs

Their population densities would be scraped from their respective Wiki pages.

Further, I'll use <https://www.foxtons.co.uk/living-in/> to extract the approximate property prices in the respective boroughs.

Once I have this data in the form of a table, I'll choose the borough which best suits the budget requirement of the customer and has a reasonable density of population so as to expect a decent footfall.

Once the borough is finalized, I will construct a table which will contain the districts of the borough and their coordinates which would be fetched using the *geocoder* library of python.

Finally, I'll use the *search* endpoint of the Foursquare API to get a list of restaurants in each district and construct a table with the columns indicating the type of restaurant(Indian,Japanese,Café,etc) and the rows indicating the neighborhood. Finally I'll cluster these neighborhoods and display the most popular restaurant themes in each neighborhood within the clusters. This would help our customer narrow down their choice of neighborhood and the type of cuisine they wish to serve at their restaurant

Methodology

The initial focus of the project was to build a visualization which could give an overview of the Density and the average property prices in the various boroughs of London.

As discussed earlier, the first dataframe I intended to create was one that contained the names of boroughs, their average population densities and the average property prices within the boroughs.

For the borough names, I used BeautifulSoup to scrape the wiki page https://en.wikipedia.org/wiki/List_of_London_boroughs.

The names of all boroughs were contained in a table which I extracted into a bs4 object. Running a loop over all the rows, I filled the Borough column of my dataframe. Within the same loop, I filled the density column of my dataframe as well. In order to do this, I made an additional request to the Borough Name hyperlink in the wiki page being scraped, which opened the dedicated wiki page of that borough.

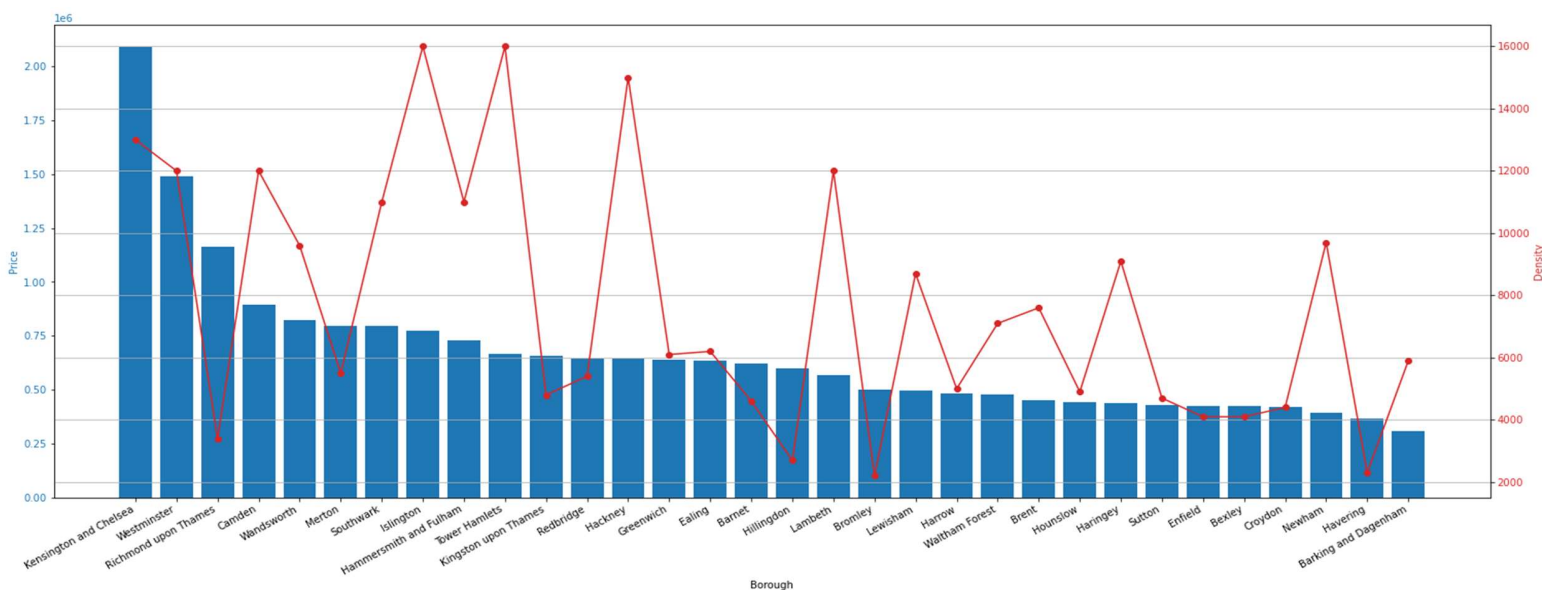
From their respective pages, I extracted the population density and appended it to the density column of the dataframe. Issues were encountered for two specific boroughs, for which I had to fill in the density values manually.

For the purpose of property prices, I took the help of

<https://www.foxtons.co.uk/living-in/>

A website dedicated to providing residents of all neighborhoods of London information including but not limited to property prices in their respective neighborhood. I was able to find out the information for most of the boroughs, but some boroughs were too large to be considered as consolidated units. So the website provided information not of those boroughs collectively, but of their sub districts individually. For the purpose of property price, I chose the most popular sub district within those boroughs to represent the collective average property price for that borough. This was supported by the evidence that the prices did not vary too much within a borough.

Once I had all the information in a dataframe, I visualized the results in the form of a dual axis plot, which clearly depicted the trend in terms of population density and property price.



GBP 600000 and GBP 900000 which is a reasonable budget for someone looking to invest in a restaurant.

Also, the population density should be moderate to high as a low density would mean low footfall and a high density would probably lead to difficulty in finding space to open a restaurant. Referring to the graph, the **London boroughs of Hammersmith and Fulham** and **Camden** seemed to be a reasonable choices.

The next step was to create a new dataframe containing the various districts of Hammersmith and Fulham and Camden along with their coordinates.

I scraped the names of districts from the wiki pages of these two districts which I further appended to the District column of my new Dataframe. Looping over the dataframe, I obtained the exact coordinates of the district using the geopy library, which I appended to the latitude and longitude columns.

The next step was to create a list of restaurants and their cuisines along with the district they're located in. For this purpose, I looped over the district dataframe and made a call to the Foursquare API using the explore endpoint. The parameters passed included the CLIENT_ID ,CLIENT_SECRET ,VERSION ,RADIUS and SECTION(which is used if we wish to search for a specific type of venue. I provided 'food' in order to get a list of restaurants).

The resultant dataframe consisted of some duplicate values as the boundaries of districts are not very well defined and the API considers some restaurants to be part of multiple districts. In order to fix this, I used the remove_duplicates command.

In order to build a dataframe suitable for applying ML models, I transformed the existing dataframe into a sparse matrix such that the columns represented all the unique types of cuisines and the each individual row had the name of the district the restaurant was situated in and had a value '1' in its respective cuisine column.

Next, I grouped the rows according to the district and summed up the column values. I thus obtained a dataframe which indicated the total number of restaurants of each category in every district.

I observed some themes were quite rare with only 1 or 2 restaurants appearing in all districts combined. I removed these themes from my analysis and included only those themes which had at least 3 restaurants present as per FourSquare.

Finally, I ran a K Means clustering algorithm on this data with $k=3$ which resulted in 3 clusters each with different districts. The cluster labels of each district were available using the method `kmeans.labels_`

The districts dataframe was modified by adding the cluster labels and including only the top 4 most common cuisines for every district.

On careful examination of the clusters, some conclusive results could be derived.

Results

The clusters obtained were as follows.

District	Cluster	Most Common Theme	2nd Most Common Theme	3rd Most Common Theme	4th Most Common Theme
Belsize Park	0	Café	Indian	Italian	Bakery
Bloomsbury	0	Café	Greek	Bakery	French
Camden Town	0	Café	Italian	Caribbean	Vegetarian / Vegan
Chalk Farm	0	Café	Italian	Bakery	Pizza
Covent Garden	0	Bakery	Café	Steakhouse	French
Fitzrovia	0	Italian	Café	Bakery	Japanese
Fulham	0	Italian	Café	Pizza	Bakery
Hampstead	0	Café	Italian	Bakery	Japanese
Kentish Town	0	Café	Indian	Italian	French
West End of London	0	Café	Italian	Vietnamese	Chinese

District	Cluster	Most Common Theme	2nd Most Common Theme	3rd Most Common Theme	4th Most Common Theme
Brook Green	1	Café	Italian	Thai	Gastropub
Chelsea Harbour	1	Italian	Sandwiches	Café	Gastropub
College Park	1	Brazilian	Thai	Deli / Bodega	Spanish
Dartmouth Park	1	Italian	Café	Fast Food	Gastropub
East Acton	1	Chinese	Bakery	Fish & Chips	Falafel
Fortune Green	1	Café	Indian	Middle Eastern	Fish & Chips
Froggnal	1	Café	Japanese	Chinese	Persian
Gospel Oak	1	Café	Gastropub	Breakfast	Italian
Highgate	1	Café	Indian	Gastropub	Caribbean
Kings Cross	1	Café	Italian	Breakfast	Burgers
Old Oak Common	1	Café	Modern European	Gastropub	Breakfast
Parsons Green	1	Italian	Vegetarian / Vegan	Burgers	Spanish
Sands End	1	Café	Mediterranean	Thai	Chinese
Somers Town	1	Indian	Burgers	Pizza	Italian
South Hampstead	1	Café	Italian	Bakery	Chinese
St. Giles	1	French	Pizza	Italian	Peruvian
Swiss Cottage	1	Café	Italian	Spanish	Asian
Tufnell Park	1	Café	Italian	Thai	Turkish
Walham Green	1	Café	Gastropub	Fast Food	BBQ
West Hampstead	1	Café	Pizza	Thai	Indian
West Kensington	1	Italian	Café	Thai	Deli / Bodega
White City	1	Café	Italian	Peruvian	Modern European

District	Cluster	Most Common Theme	2nd Most Common Theme	3rd Most Common Theme	4th Most Common Theme
Hammersmith	2	Café	Pizza	Sandwiches	Burgers
Holborn	2	Café	Sandwiches	Food Truck	French
Shepherd's Bush	2	Café	Chinese	Fast Food	Indian
St Pancras	2	Café	Sandwiches	Indian	Italian

Following Conclusions can be drawn –

- The outer London boroughs have lower property prices but also have a low density of population which probably is not in favor of a footfall dependent business like a restaurant.
- Although the inner London boroughs have a high density of population and are popular among youth and tourists, they're prohibitively expensive.
- Within the London boroughs of Hammersmith and Fulham and Camden, which we concluded were reasonable choices to open a restaurant, Café is the most popular restaurant theme followed by Italian.
- For districts lying in the cluster 0, a Café with a predominantly Italian theme would be a good theme for a restaurant .
- Cluster 1 seems to have a somewhat varied distribution with café and Italian being the most common followed by Indian and Chinese. Different districts seem to have their own favourite theme. If our customer wants to experiment and try a different unconventional theme, cluster 1 would be a good spot as it seems the residents of these districts are open to experimentation.
- In cluster 2, apparently the residents prefer a quick snack or fast food such as sandwiches. These could be kept in mind while deciding on a theme. A corner café serving finger food could do well here

Conclusion

To conclude, we can say that this project , although not perfect, provides a reasonably accurate picture of the city of London to a potential customer in terms of what to expect when venturing out to start a restaurant business.

The results will definitely provide a good headstart and will definitely help our customer narrow down his location search.

The project, if built upon , could be made more general in nature and could even be developed into an app with a provision for the user to enter the specific type of business and the country/city they wish to invest in. The project could also be modified to include the provision to enter specific parameters which the user considers as priority while planning to make an investment, eg – distance from city centre, average rents, government policies, etc.