Introduction

We have been approached by a company interested in investing in a new restaurant opening within London. They have asked us to determine where would be the best location and what sort of restaurant would be most appropriate for that area. The business problem and purpose of this project therefore is to select a borough in London to open a new restaurant and select its cuisine. In order to optimise the location and restaurant type we will characterise the boroughs of London using k-means clustering. The people interested in our results will be those looking to invest in restaurant within London.

Data

For the borough data we are scraping the webpage:
https://en.wikipedia.org/wiki/List_of_London_boroughs

This data will be supplemented with data on up to the top 100 venues within a 1km area the London borough. This additional data will be acquired from Foursquare and is explained in detail below.

Method

1. **Acquire Borough Data**
   The first part of the project will involve finding data relating to the borough of London and their location.
2. **Data Wrangling**
   Using data wrangling techniques this data will be processed and we will keep and collate relevant parameters associated with each borough.
3. **Acquire Venue Data (Foursquare API)**
   We will then use the Foursquare API to get information about the venues within a 1km radius of the borough coordinates. A 1km radius is chosen as it is assumed that this is a suitable walking distance.
4. **Clustering by Restaurant Population**
   Using k-means clustering we will cluster the boroughs based on the number of restaurants per population density into low, medium and high. We opted to use population density rather than population as it was assumed that each borough location would only serve a subpopulation with the borough rather than the entire population. Since we are only searching within a 1km vicinity of a location in each borough this helps account for variations in the area of each borough. For the analysis we assume that the boroughs with a relatively high number of restaurants (relative to the population density) have features in common. Relevant features could include a disposable income, business vs residential areas etc. We further assume that these predictive features correlate with features such as the frequency of other types of venue categories (i.e venue categories excluding restaurants).
5. **Clustering by Venue Categories**
   Using k-means clustering we will then cluster the boroughs based on the frequency of types of venues excluding restaurants. We will use the Silhouette Score to select the most appropriate number of clusters. The generated clusters will be referred to hereafter as "feature clusters". If we find a feature cluster which contains many of the relatively high restaurant boroughs, then we assume that those features are predictive of the type of boroughs which should have a relatively high number of successful restaurants.
6. **Borough Selection**
   We will then see what other boroughs are contained within the same cluster as many of the high restaurant boroughs and then select the borough with the relatively lowest number of restaurants. The assumption here is that a borough with a relative low number of restaurants but features which are shared with the high restaurant borough is a good candidate for a new restaurant. The best-case scenario would be if there are boroughs

which appear simultaneously in the low number of restaurant cluster and in the feature cluster with a high number of high restaurant boroughs.

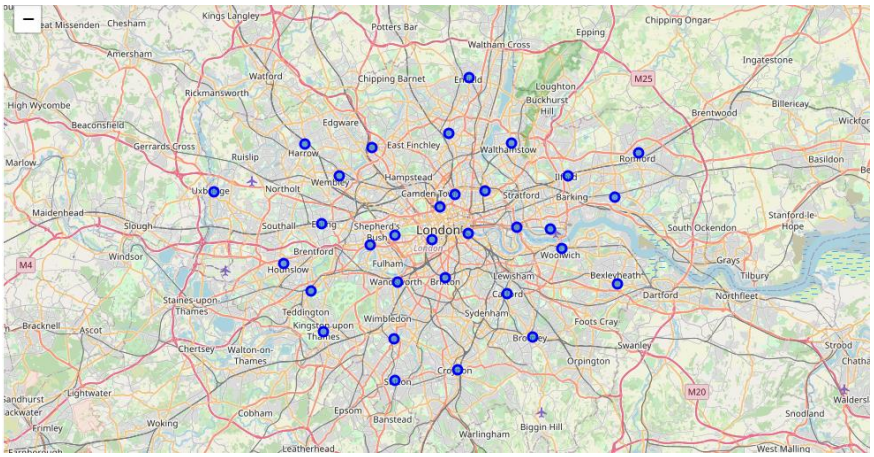
7. **Restaurant Selection**
   After selecting the best borough for a restaurant we will then analysis the best type of restaurant to open in that area. For this part of the project we estimate the number of different restaurant categories you would expect in the selected borough based on the proportions of different restaurant categories in the high restaurant cluster. Using this comparison, we will decide which type of restaurant will have the highest demand in the chosen borough.
8. **Conclusions**

Results

To begin with we process the scraped wiki data to get a dataframe which looks like the following;

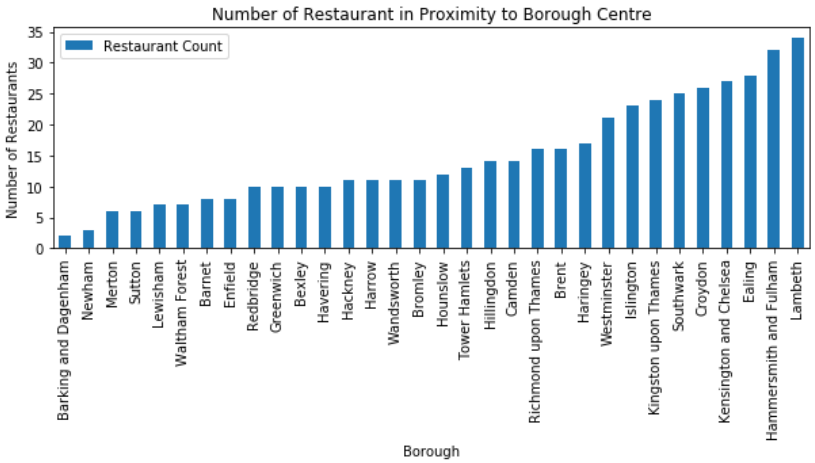
	Borough	Population Density (sq mi)	Latitude	Longitude
0	Barking and Dagenham	13952.0	51.5607	0.1557
1	Barnet	11021.0	51.6252	-0.1517

The dataframe contains information on all the London boroughs. The borough locations can then be plotted on a map.

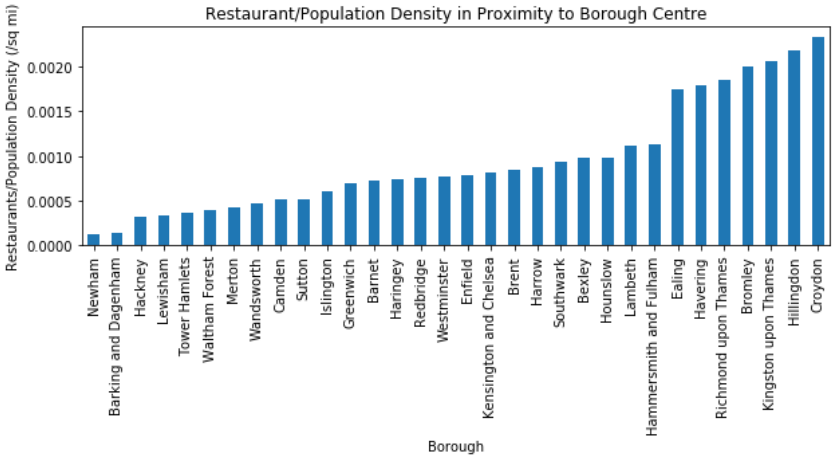


We then request data on up to the top hundred venues located within 1km of each borough point in the above map.

The figure below shows a graph of the total number of restaurants for each borough given by the Foursquare request:

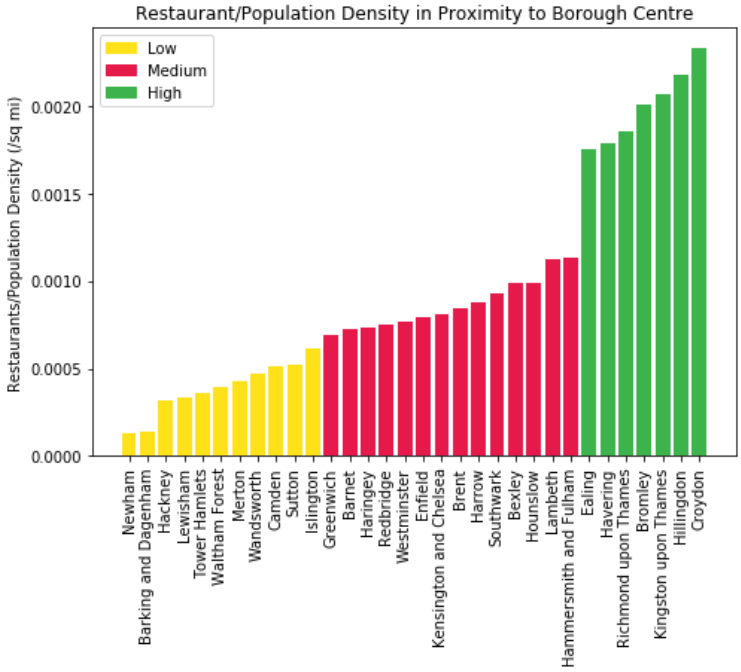


In the following graph we divide the number of restaurants by the population density:

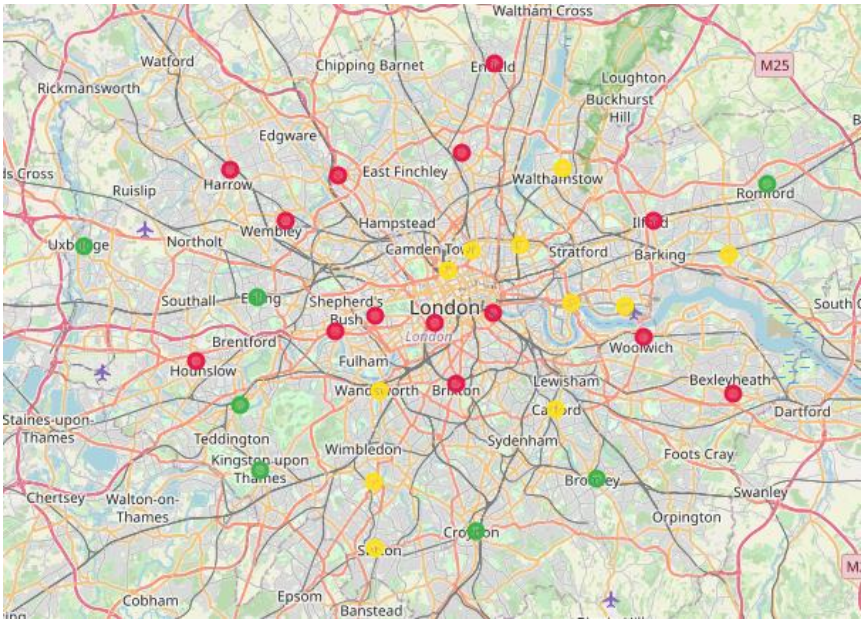


We opted to use population density rather than population as it was assumed that each borough location would only serve a subpopulation with the borough rather than the entire population. Since we are only searching within a 1km vicinity of a location in each borough this helps account for variations in the area of each borough. We can see some reordering of the boroughs compared to the previous graph.

Using this data, we then use a k-clustering to create clusters with a low, medium and high number restaurants per population density. That is to say with relatively low, medium or high number of restaurants.

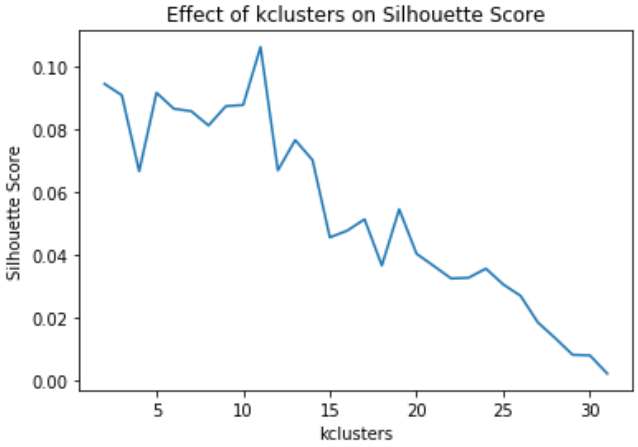


We can also plot these cluster on the map of London:



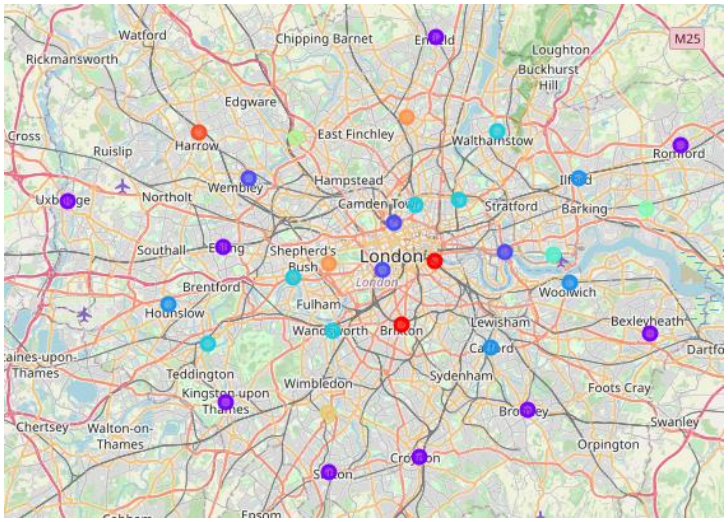
From the above map it looks like the most prolific restaurant industry relative to the population density (green dots) are around the edge of the city. These areas could correspond to the commuter belt and might indicate more disposable income to support restaurants.

To investigate features consistent with high restaurant boroughs we exclude restaurant venues from the data frame. To determine the optimal number of clusters for this analysis we use the Silhouette Score:



In this case we find that the best number of clusters = 11.

Performing k-cluster analysis we produce the following cluster map:



From the above map it looks like, with the expectation of Richmond upon Thames Cluster 1 captures the perimeter (possibly the commuter belt) of London. We can analysis the feature clusters in more detail below:

Cluster 1:

Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bexley	Pub	Clothing Store	Coffee Shop	Supermarket	Hotel	Pharmacy	Furniture / Home Store	Nightclub	Discount Store	Department Store
Bromley	Pub	Clothing Store	Coffee Shop	Gym / Fitness Center	Supermarket	Electronics Store	Pizza Place	Burger Joint	Café	Stationery Store
Croydon	Coffee Shop	Pub	Hotel	Clothing Store	Bookstore	Café	Sandwich Place	Platform	Gym / Fitness Center	Park
Ealing	Coffee Shop	Pub	Pizza Place	Bakery	Park	Burger Joint	Hotel	Café	Gym / Fitness Center	Supermarket
Enfield	Pub	Clothing Store	Coffee Shop	Grocery Store	Supermarket	Optical Shop	Shopping Mall	Gift Shop	Pharmacy	Café
Havering	Coffee Shop	Pub	Clothing Store	Shopping Mall	Park	Furniture / Home Store	Café	Supermarket	Grocery Store	Department Store
Hillingdon	Coffee Shop	Pub	Clothing Store	Pharmacy	Gym	Supermarket	Bar	Burger Joint	Bookstore	Park
Kingston upon Thames	Coffee Shop	Pub	Café	Burger Joint	Park	Clothing Store	Department Store	Hotel	Bookstore	Bar
Sutton	Coffee Shop	Clothing Store	Pub	Café	Hotel	Bar	Sandwich Place	Department Store	Pizza Place	Supermarket

This cluster has ~85% of the boroughs in the relatively high number of restaurant clusters, indicating that these shared features correlate with restaurant success.

Cluster 4:

Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Hackney	Pub	Coffee Shop	Bakery	Brewery	Café	Park	Supermarket	Flea Market	Beer Store	Wine Shop
Hammersmith and Fulham	Pub	Coffee Shop	Café	Gym / Fitness Center	Park	Sandwich Place	Gastropub	Plaza	Clothing Store	Bar
Islington	Pub	Coffee Shop	Park	Café	Gastropub	Boutique	Bakery	Trail	Cocktail Bar	Theater
Richmond upon Thames	Pub	Coffee Shop	Café	Grocery Store	Park	Bakery	Pharmacy	Pizza Place	Deli / Bodega	Boat or Ferry
Waltham Forest	Pub	Café	Art Gallery	Coffee Shop	Gym / Fitness Center	Pizza Place	Brewery	Burger Joint	Park	Multiplex
Wandsworth	Pub	Café	Gym / Fitness Center	Coffee Shop	Hotel	Clothing Store	Pizza Place	Grocery Store	Supermarket	Breakfast Spot

This cluster is similar to cluster 1 with high frequencies of pubs and coffee shops but not as high frequency of clothing stores. Since most of the high restaurant boroughs are in cluster 1 this

could suggest that the frequency of clothing shops is a good indicator of the success of restaurants, perhaps correlating with disposable income or the preference of a certain type of client.

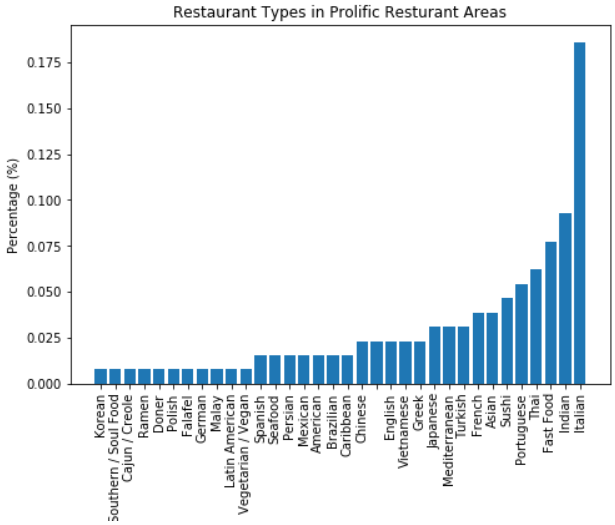
Boroughs in the high restaurant feature cluster: Bexley, Bromley, Croydon, Ealing, Enfield, Havering, Hillingdon, Kingston upon Thames, Sutton

Boroughs with a low restaurant industry relative to the population density are: Newham, Barking and Dagenham, Hackney, Lewisham, Tower Hamlets, Waltham Forest, Merton, Wandsworth, Camden, Sutton, Islington

We can then look for overlap between the two clusters and select the borough with the relatively lowest number of restaurants:

The boroughs with features consistent with a successful restaurant industry but a low restaurant presence are: **Sutton**

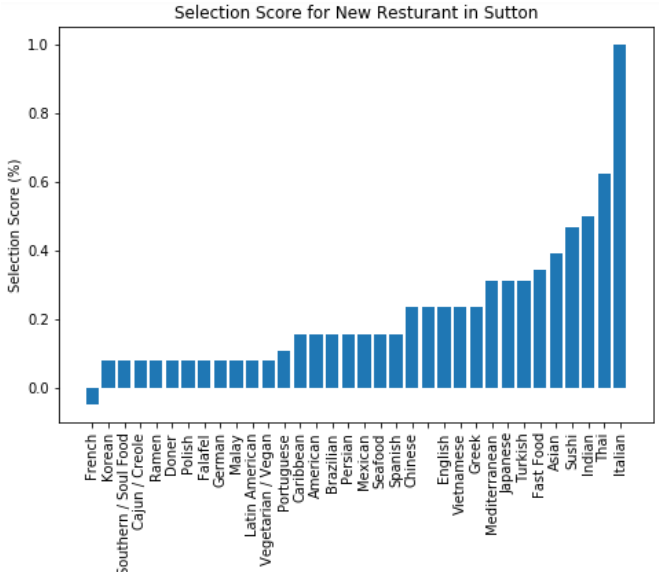
Having identified the best borough, we can now determine the best cuisine. The graph below plots the percentage of different cuisines within the high restaurant cluster.



If we assume that the high restaurant boroughs collectively indicate the viable restaurants categories and their frequency then we can normalise the total number of each restaurants with respect to the total population density of these boroughs. If we multiply the high restaurant count per population density by population density of our selected borough then we get an indication of the numbers of restaurant that the borough could accommodate:

	Expected number	True number	Difference
Sushi Restaurant	1.0	0.0	1.0
Portuguese Restaurant	1.0	1.0	0.0
Thai Restaurant	1.0	0.0	1.0
Fast Food Restaurant	1.0	1.0	0.0
Indian Restaurant	2.0	1.0	1.0
Italian Restaurant	4.0	2.0	2.0

Based on this analysis good restaurants to consider would be Italian, Indian, Thai or Sushi. We can order these more exactly by considering the difference in the expected number per population density and the true number per population density. The normalised result is displayed below:



Based on this analysis the best choice for a new restaurant would be either an Italian or Thai. Based on the high restaurant areas there is likely a high demand for another Italian, despite a presence already in Sutton. We should also seriously consider the possibility of a Thai restaurant which comes second in the above scoring metric and otherwise has no direct competitors in terms of cuisine. This exclusivity might make the choice of Thai promising compared to Italian than the above scoring system implies which consider a fairly saturated marketplace. The question of Italian or Thai could be answered more definitely with additional data which examines the benefit of exclusivity in the restaurant sector.

Discussion

Based on the analysis above we identified the most promising borough as Sutton. Sutton shares feature with many of the boroughs which have high population of restaurants relative to the population density, but Sutton has very few restaurants itself given its population density. Sutton was the only low restaurants borough in the feature cluster containing ~85% of the high restaurant boroughs. Having identified Sutton we developed a scoring system to determine the type of restaurants that would be the most successful. Using the data from the boroughs in the high restaurants cluster we estimated what sort of restaurants could be accommodated by Sutton given its population density. Based on this analysis the top two cuisines were Italian followed by Thai. It should be noted however that there are already two Italian restaurants in Sutton, but no Thai. We therefore recommend Thai as based on the available data from similar boroughs will be a high demand for Thai cuisine and there is at present no direct competitor.

Conclusions

Borough: Sutton

Restaurant: Thai

We might also want to consider whether:

1. Areas might have a low population but a high degree of footfall.
2. The location of boroughs may not be at major towns.
3. Whether a finer granularity at the scale of neighbourhoods would give more reliable results.