Location, Location & Recommendation?

By Nicholas Tsangari

1. Introduction

1.1. Background

London has the largest number of migrants among all regions of the UK, 3,317,000 – or 35% of the UK's total foreign-born population. In 2019, people born outside of the UK made up an estimated 14% of the UK's population, or 9.5 million people. That makes London one of the most diverse cities in the world.

Being a diverse, cosmopolitan city, London is home to a number of restaurants serving a variety of cuisines, and dishes which capture the imagination of thousands with customers being able to enjoy delicacies from all around world, within 1 city. Quite rightfully, London has earned it's title as the 'most diverse and exciting food capital of the world'.

A portion of that diverse cuisine that you can find in London is made up of Greek dining, which has been long established within the heart of London since the late 50s. Bourne out of a desire to recreate food once served in the motherland, Greek restaurants have branched out throughout London, spilling over into suburbs where demand for the cuisine remains hot.

So let's set the scene, you're a budding entrepreneur, eager to open up a new Greek restaurant as close to London as possible but are worried about fierce competition. The first question on your entrepreneurial mind is 'Which area can I open a restaurant in?' That is where our story begins and we are going to attempt that with Data Science...

In this project we will direct our efforts on detecting areas of London that have low restaurant density for the Greek restaurants using a machine learning algorithm known as K-Means Clustering.

1.2. Problem Description

If someone wants to open a new restaurant wouldn't it be handy for them to have some information on where the nearest Greek Cypriot restaurant would be, right? Or what about knowing the local competitors within a particular borough?

That, in short, is our problem to solve. We want to help a particular audience of entrepreneurs who want to open a restaurant in London where the best Greek and ascertain the volume of those in a particular neighbourhood and present our audience with an opportunity to take the plunge in opening new businesses.

I will be using UK based restaurant data from FourSquare API as part of this project, as well London borough information taken directly from Wikipedia.

By utilizing the London restaurants data from the FourSquare API I will be looking to assess the risk category of restaurants, identify the closest restaurants to the center of London and preform an unsupervised machine learning algorithm to cluster those restaurants by neighbourhood, all whilst keep our audience at the forefront of our mind. Our final aim is to answer the question and offer recommendations to: In the City of London, where is the best place to open a Greek Restaurant based on location information.

1.3. Target Audience

- Ambitious entrepreneurs looking to open a Greek restaurant.

This project is timely coming off the back of COVID, with restaurants beginning to be relax restrictions and open their doors to the general public, there will be a desire for consumers to try out new restaurants. That increase in desire to experiment with new cuisines/restaurants could spill out to neighbouring towns around London, thus making this the perfect opportunity for entrepreneurs looking to take the 'plunge'.

We know the question that we need to answer and we also know that Data Science can help us to reach a sensible recommendation but what does that journey look like and what are the steps in between? First we need to identify the data we will be using...

2. Data to be used

To answer our questions and offer our recommendations, we will use the following data :

- List of London boroughs. This will define the scope of this project which is focussed around London. We will attempt to obtain a list of these boroughs directly from Wikipedia. (https://en.wikipedia.org/wiki/London_boroughs) contains a list of 32 boroughs in London. We will proceed with web scraping techniques to extract the data from the link, using the BeautifulSoup library.

- Latitude & Longitude Coordinates: Given this project heavily involves location data, it is imperative we obtain our latitude & longitude coordinates form a base of our code. We will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the boroughs.
-
- Data source : Foursquare API has one of the largest database of 105+ million places and is used by over 125,000 developers. The API allows us to connect to real-time information on the location of restaurants as well as their rating so we will be able to ascertain the closest location, best price and best rating all from one data source. Foursquare API will provide many categories of the venue data, we are particularly interested in the 'Greek Restaurant' category in order to help us to solve the question put forward.

This is project offers up the chance to work various data sources, and encourages the use of leaning on a variety of difference data science skills that would've been worked on over the past 4 months of the IBM course.

We will utilise our skills in web scraping by pulling in data from Wikipedia, visualizations of being able to produce maps through Folium and connecting/querying the Foursquare API to then wrangle, clean and apply our machine learning methods on top.

2.1. Data Cleaning

We have ensured that the London Boroughs we extracted from Wikipedia have been cleaned and structure into a format that we can work with.

	Borough
0	Barking and Dagenham
1	Barnet
2	Bexley
3	Brent
4	Bromley
5	Camden
6	Croydon
7	Ealing

Working with Foursquare data was an interesting experience. There is such a vast quantity of information that you can obtain from the different endpoints and we it's about being streamlined in our approached rather than pulling in every data point unnecessarily.

The data is received in a JSON format, so there is a requirement for some level of data wrangling to ensure this is in a suitable format.

3. Methodology and Exploratory Data Analysis

In this project we will direct our efforts on detecting areas of London that have low restaurant density for the Greek cuisine.

The methodology behind our code will be focussed and bespoke to answering our question at hand. We want to work with the Foursquare API and pull off the top 100 venues that are within a radius of 2000km from the centre of London. We have used the center of London as our 'center spot' because this would be the most populous area for consumers and thus securing a restaurant as close to this spot would, in theory, suggest this would a lot busier. As a caveat, there has been no science behind proving restaurants within the center of London are busier.

By obtaining a list of top 100 venues this will enable us to indicate very quickly, which restaurants are the closest to our prime location and thus if you wanted to take this analysis in a different direction you could look at the average price/rating for those closer to the center, which I am sure would provide some interesting results.

With the list of boroughs, the aim is to work out the number of Greeks restaurants within a 350 metre radius of each borough and thus establish which boroughs are more densely populated. Using a 350 metre radius ensures there is full coverage and that we aren't missing any restaurants in the process.

With our dataset taking shape, the idea is them to perform an unsupervised machine learning algorithm such as 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.

We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for "Greek restaurants". The results will allow us to identify which neighbourhoods have higher concentration of restaurants while which neighbourhoods have fewer number of restaurants. We also want to maintain that the clusters with the fewest number of restaurants are still the closest they can be to the centre of the London.

Based on the occurrence of restaurants in different neighbourhoods, it will help us to answer the question as to which boroughs are most suitable to open a restaurant.

From a step by step perspective, we will undertake the following:

- Our first step is to obtain a list of boroughs in the City of London, which is where we lean on Wikipedia for this and utilize our web scraping skills in order to make that happen. As explained above, we will use packages such as BeautifulSoup and Requests in order to extract the lists.

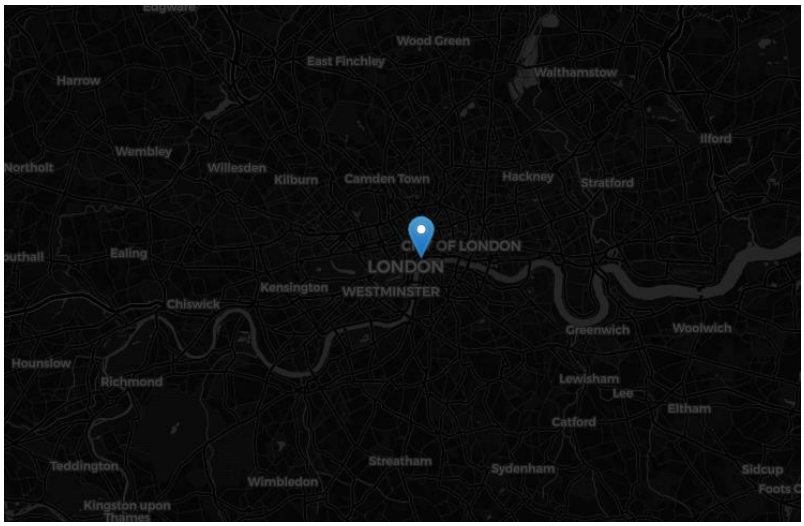
- Given that this is simply borough names, we still need to obtain a list of the matching coordinates. Our reasoning behind this is so that we can iterate through those coordinates using Foursquare API and find out the closest restaurants to those boroughs.

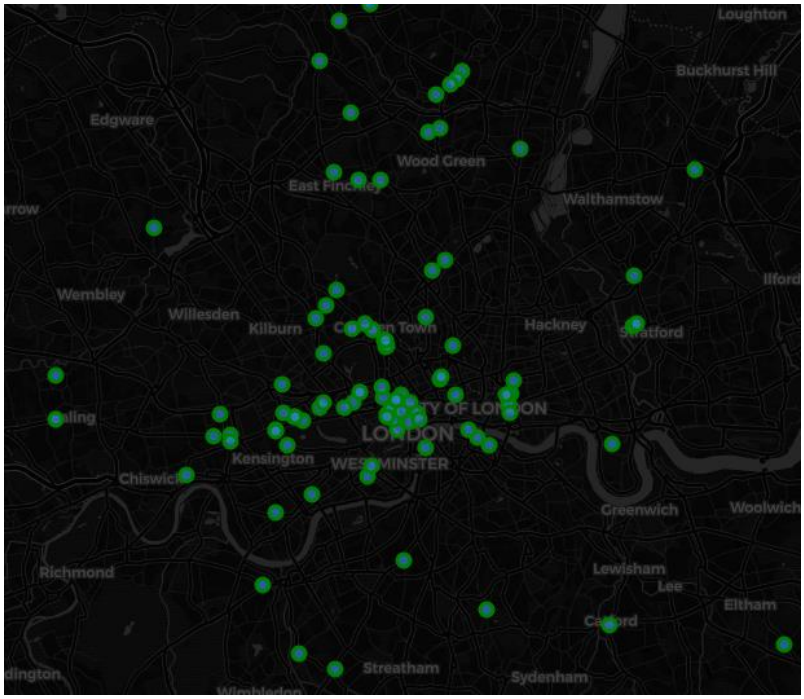
name	lat	long
Barking and Dagenham	51.554117	0.150504
Barnet	51.653090	-0.200226
Bexley	51.441679	0.150488
Brent	51.563996	-0.275906

- To find out the closest boroughs to the centre of London, we work out the distance between both coordinates using Haversine formula which determines the great-circle distance between two points on a sphere given their longitudes and latitudes.

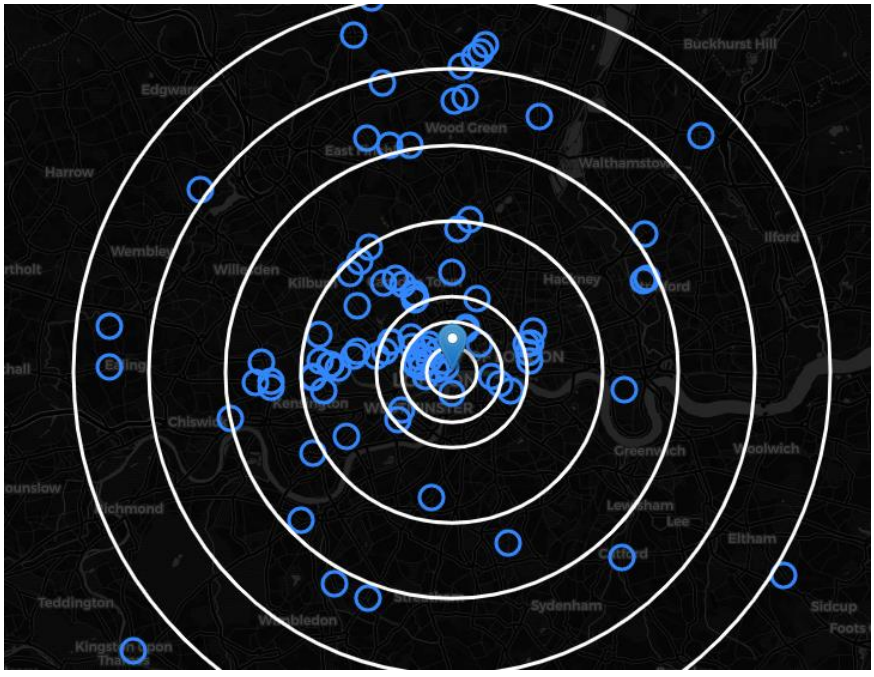
	name	lat	long	Ldn_lat	Ldn_long	distance_km
0	Barnet	51.653090	-0.200226	51.509865	-0.118092	16.921623
1	Bexley	51.441679	0.150488	51.509865	-0.118092	20.143313
2	Brent	51.563996	-0.275906	51.509865	-0.118092	12.496790
3	Bromley	51.366857	0.061709	51.509865	-0.118092	20.234981
4	Camden	51.542305	-0.139560	51.509865	-0.118092	3.904606
5	Croydon	51.371305	-0.101957	51.509865	-0.118092	15.456508
6	Ealing	51.512655	-0.305195	51.509865	-0.118092	12.993047
7	Enfield	51.652085	-0.081018	51.509865	-0.118092	16.030654

- We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the boroughs in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and perform adequate EDA on the results.





We will use marker maps to identify a few promising areas close to centre with low number of Greek restaurants focus our attention on those areas.

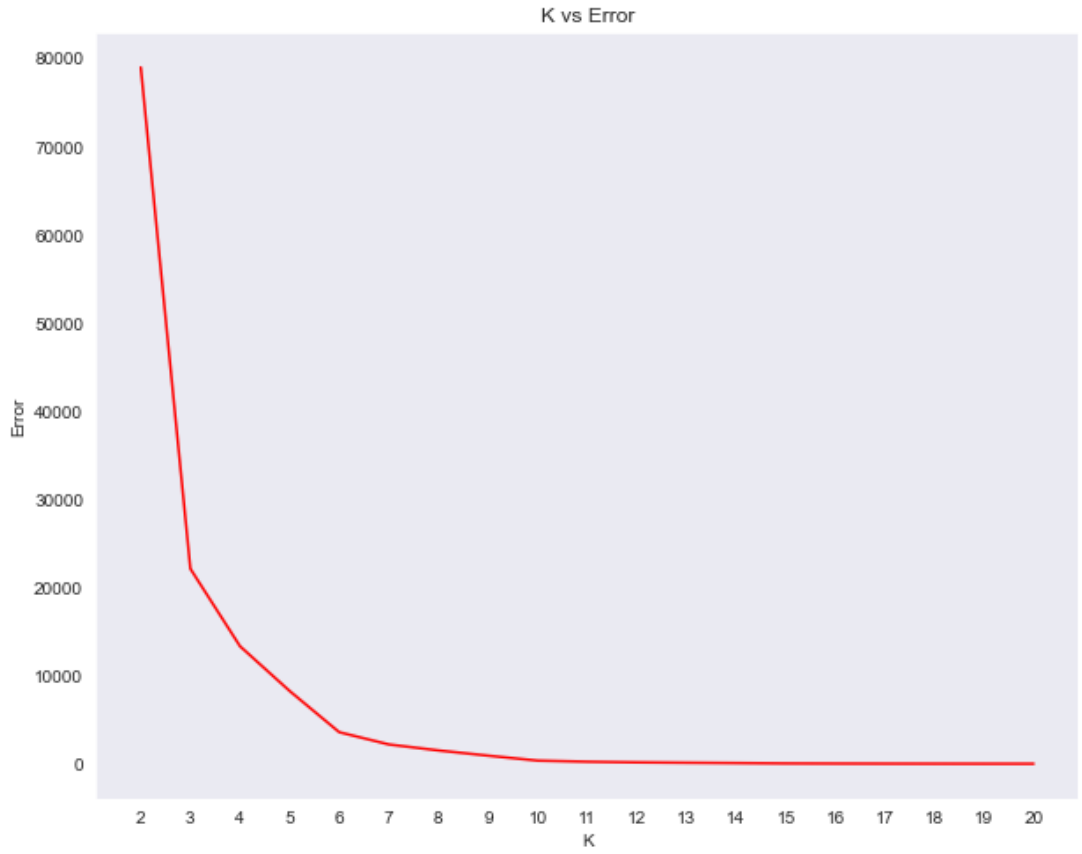


In our final step we will focus on those low density areas and create clusters of locations in order to answer our question.

Our results will be presented with a map of all such locations but also create clusters (using k-means clustering) of those locations to identify zones.

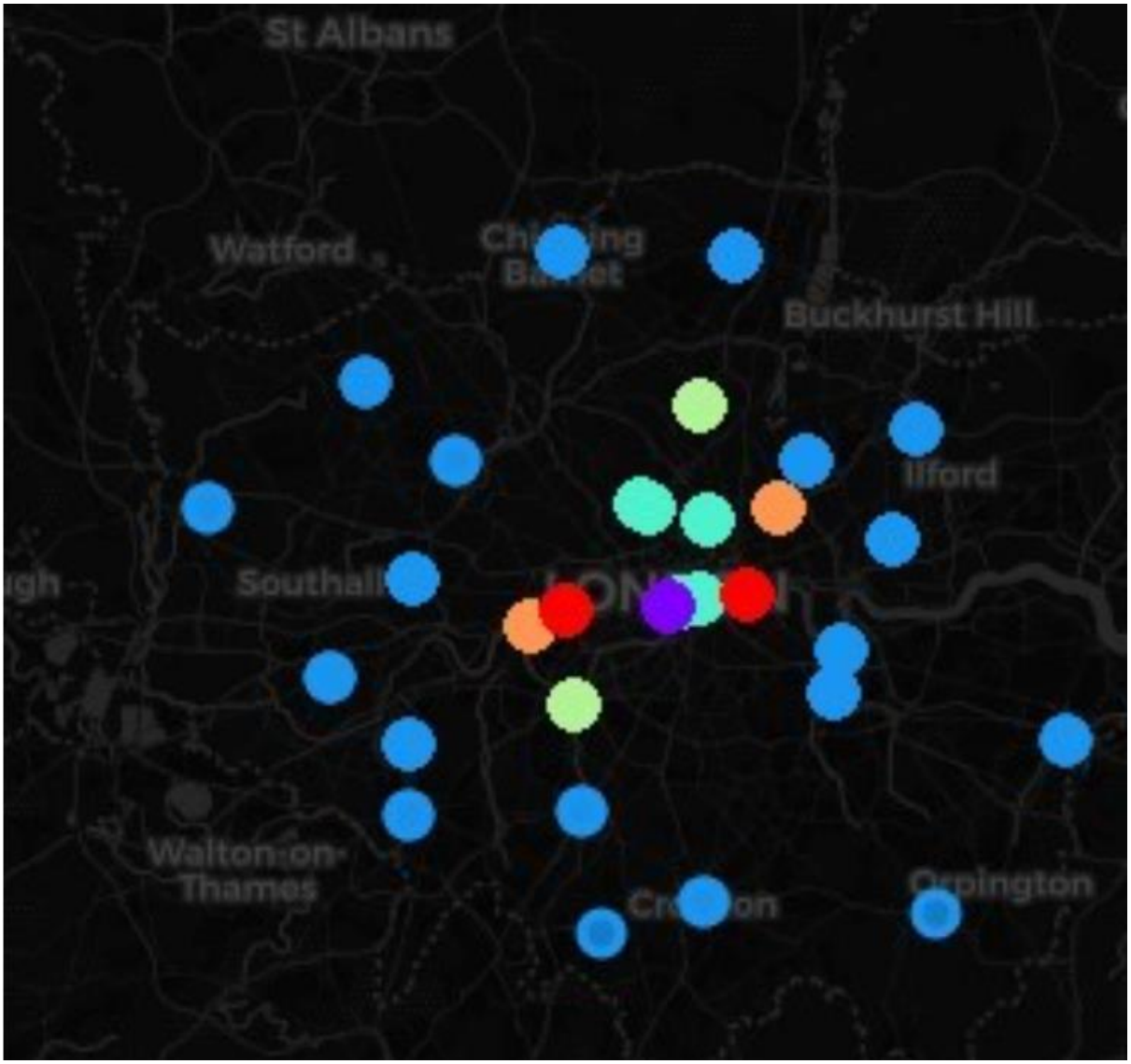
K-Means Clustering

As part of this project, we wanted to incorporate a machine learning algorithm to cluster the various boroughs accordingly and point us towards a recommendation. In order to create model, it needs to be fine tuned to find our optimum K value. We achieve this by fitting the model to our data and re-run the test several times using different values of K's. We then compare the results and see which value of K score the most accurate result. In layman's term it's all about 'finding the elbow point', which the graph below will show...our best value of K is 6:



4. Results

Our analysis reveals that although there are a high number of Greek restaurants towards the centre of London (as expected), we can clearly identify that there are pockets of low restaurant density around the suburbs of London, forming almost a circle to that effect.



The results from the K-means clustering show that we can group the restaurants into 6 groups based on frequency from highest to lowest:

• Cluster 0 (Red): Boroughs with the highest number of Greek restaurants in the vicinity

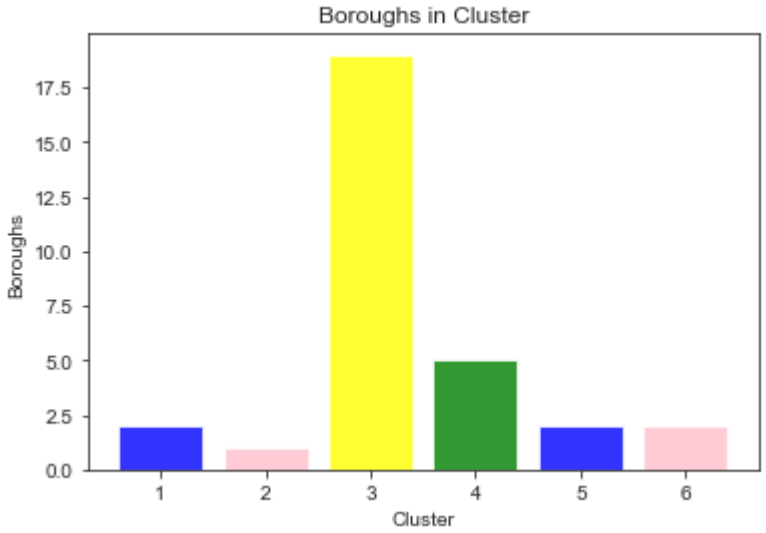
• Cluster 1 (Green)

• Cluster 2 (Purple)

• Cluster 3 (Red)

• Cluster 4 (Orange)

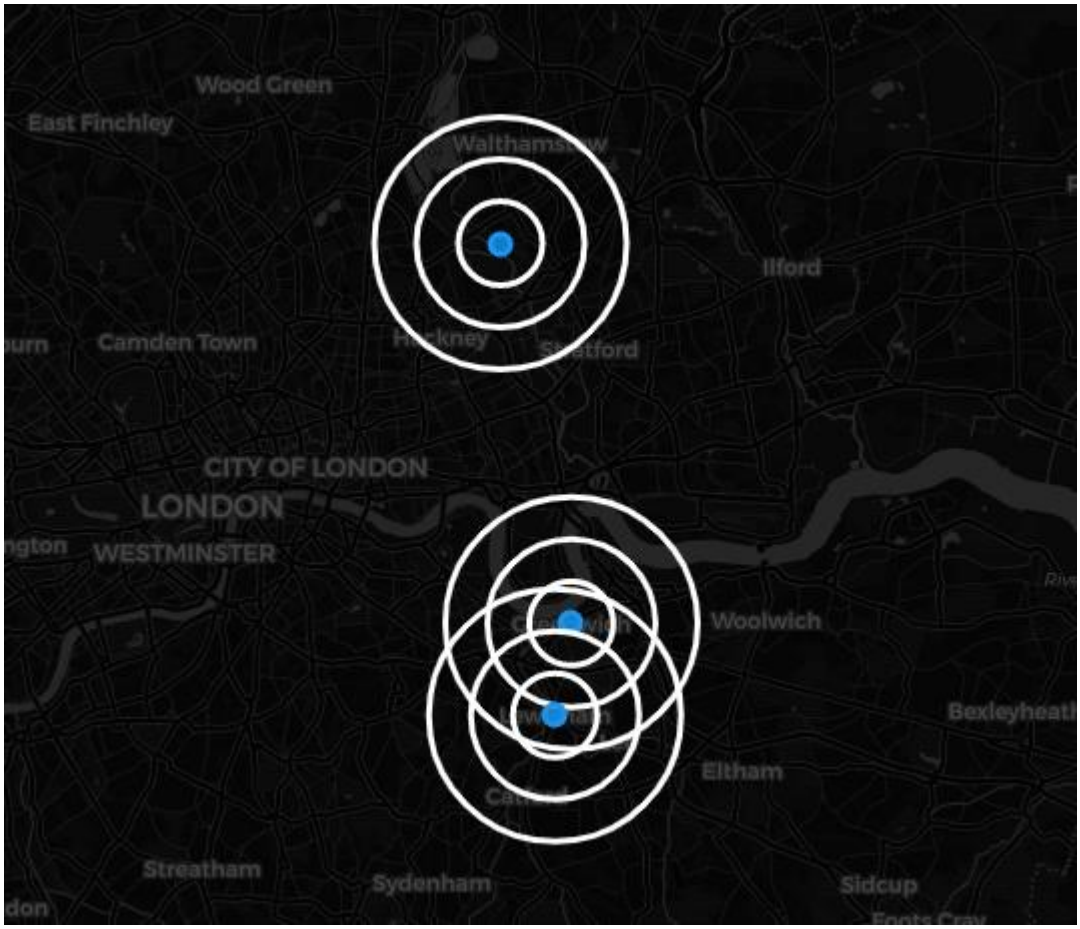
• Cluster 5 (Blue): Boroughs with the highest number of Greek restaurants in the vicinity



The chart above displays some analysis on the number of boroughs that sit within various clusters.

Now we are able to find which cluster has the lowest density of restaurants (Cluster 5), I assessed which boroughs within those clusters were the closest to the Centre of London and found the following:

name	lat	long	distance_km	total_restaraunts_in_area	Clusters
Greenwich	51.482084	-0.004542	8.469772	6	2
Waltham Forest	51.563187	-0.028841	8.575778	9	2
Lewisham	51.462432	-0.010133	9.169732	8	2



5. Discussion

Given our results in the above section, there is a high concentration of restaurants in central London. Breaking this down further, we can see that the highest number in cluster 0 and moderate number in cluster 3. On the other hand, cluster 5 has very low number to no restaurants in those neighbourhoods.

This represents a great opportunity and high potential areas to open a new Greek restaurant as there is very little to no competition from existing restaurants. Meanwhile, restaurants in cluster 0 are likely suffering from intense competition due to oversupply and high concentration.

From another perspective, the results also show that the oversupply of restaurants mostly happened in the central area of the city, with the suburb area still have very few restaurants and thus it could be a case where there is still a high demand in those areas and perhaps opening a restaurant in those areas, ensuring the quality of food and prices remain competitive wouldn't necessarily become a huge problem, however the extent of this analysis doesn't go into depths on that.

This project recommends to entrepreneurs to capitalize on these findings to open new restaurants in Boroughs of Greenwich, Waltham Forest or Lewisham as the analysis reveals little to no competition. Entrepreneurs with unique selling propositions to stand out from the competition can also open a restaurant in cluster 4 with moderate competition. Lastly, entrepreneurs are advised to avoid boroughs in clusters 0-2 which already have high concentration of restaurants and suffering from intense competition.

6. Conclusion

Through this project we have undertaken the process of identifying the problem, specifying the data required, extracting and preparing the data, performing machine learning, and lastly providing recommendations to the relevant stakeholders.

To answer the question that was raised in the introduction section, the answer proposed by this project is: The top 3 boroughs in cluster 5 are the most preferred locations to open a Greek restaurant. We hope the findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decision.

7. Limitations

There are some limitations in the project which is useful to clarify. From the outset, we only considered one factor i.e. density of restaurants in a particular area. In fact there are other factors such as population and income of residents that could influence the location decision of a restaurant. Furthermore, this project utilized the ‘free account’ of Foursquare API that came with limitations as to the number of API calls and results returned. It would have been interesting to pull in the ratings and average price spent at restaurant to assist in helping us make our recommendations.