Aleksander Johansen 13.05.2019

Restaurant location optimization in London

1. Introduction

- In this report, I am going to investigate where an Indian restaurant owner should establish a new restaurant. The owner thinks that the area surrounding 'Kings Cross, London N1C 4AG, UK' (geo 51.533, -0.125) is a good starting point for the analysis and wishes to search the area around this location for potential competitors (i.e. other Indian restaurants).

- Ideally, it should be possible to group the restaurants (i.e. cluster) to see which areas near Kings Cross that have unused capacity with respect to Indian restaurants.

- This business problem is relevant since Kings Cross and especially Pancras Square is an attractive area for many businesses such as Google and media companies are located here.

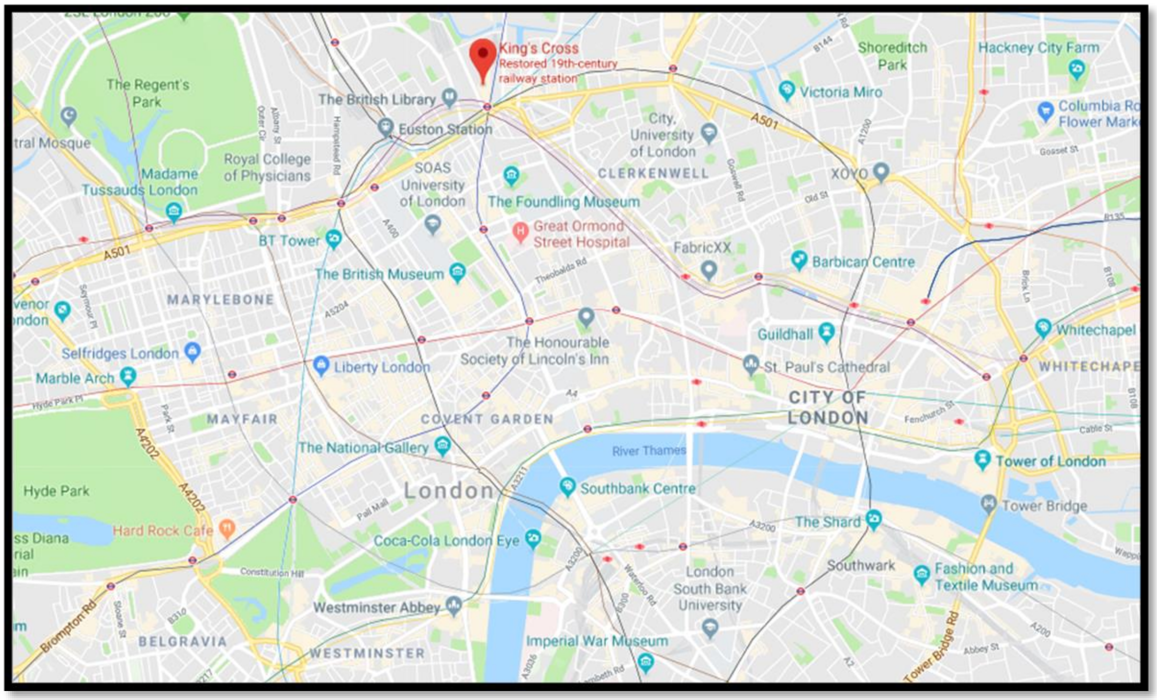


FIGURE 1 - OVERVIEW OF LONDON FROM GOOGLE MAPS

2. Data

- FYI: My starting point was a business problem similar to the one I described in section 1, only with focus on Oslo instead of London. However, the data on Foursquare for the Oslo area was very poor, so I had to change area due to the lack of quality data.

- The data that I utilize on in this report is Foursquare data for the London area. I have specifically extracted data for Pancras Square with the search query 'Indian Restaurants'. To limit the search, I have extracted all restaurants surrounding Pancras Square with a 3 km radius, given the search query criteria.

- The results obtained by that search was a 49x19 based matrix with restaurants, grocery stores and cafes in the nearby area. Hence, the 'Indian Restaurants' search query did not restrict the search to only restaurants. I used some time to analyze why these irrelevant search results showed up. However, I did not find the main reason.

- Given the biased search results, I had to clean my data for grocery stores, cafes and museum results. This meant dropping certain rows and columns.

- Having normalized the data, dropped certain columns and dropped all rows not containing a category (column) value equal to 'Indian restaurants', my final data frame was a 14x9 matrix with only Indian restaurants. This means that I dropped (49-14) = 35 observations and (19-9) = 10 column values due to irrelevance.

- Plotting the data on a map using a Folium map, shows that the observations are spread quite widely throughout the City of London. The red marker is the center of the search query, whereas the blue points represents each observation of the data frame.

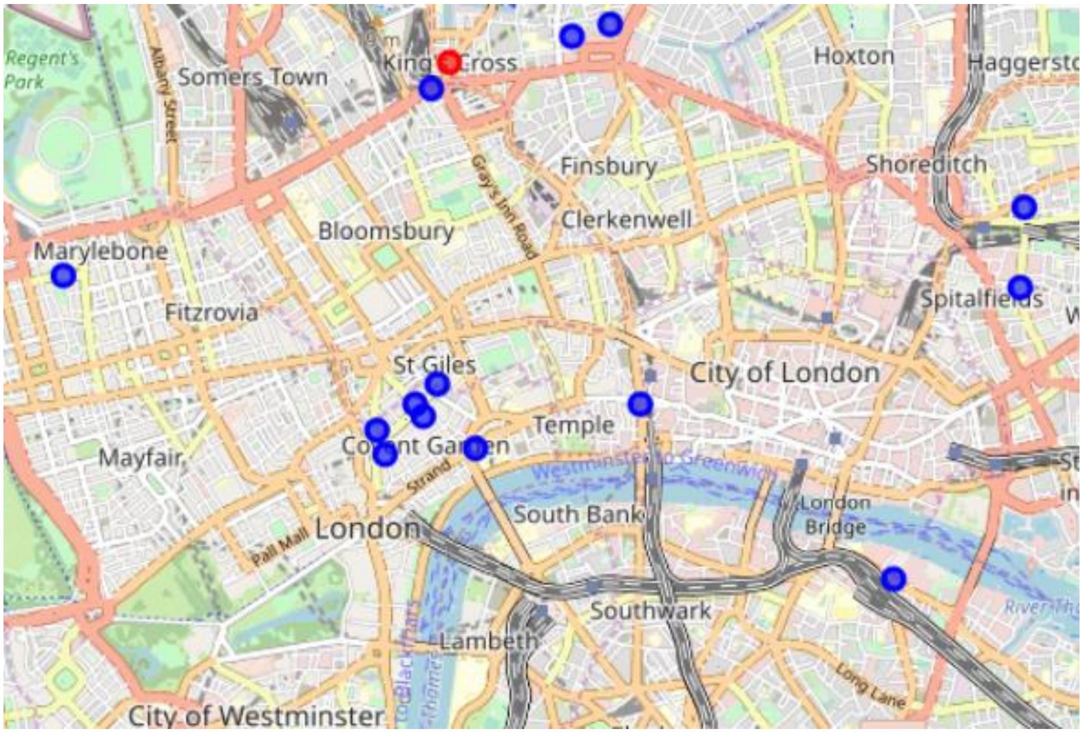


FIGURE 2 - MAP OF LONDON WITH OBSERVATIONS

- Immediately, it is quite easy to see that this data centers around two to three different areas in the South, North and East of the picture, whereas there is one outlier in the West. This is due to the large search grid of 3 km of the API call towards Foursquare.

3. Methodology

- I intend to use the machine learning technique named clustering ("KMeans") to group the observations based on the distance from random, optimized centers of each observation group. My assumption in this is that there are between 3 and 5 clusters, and I will have to apply different modules to find the optimized fit.

- Using the clusters, I will be able to define the center of each observation group to say something useful about the optimized location for a restaurant in the area. I assume that the ideal location is the one which satisfies the following criteria:
- a) It is in the center of the surrounding observations.
- b) It has an even distance to the surrounding observation.

4. Results

- To run KMeans clustering on the data, it is important to ask what data should be applied. In the geographical plot above, the data is used is latitude and longitude of each restaurant. This means that the data can be plotted as a scatter plot with y representing latitude and x representing longitude:

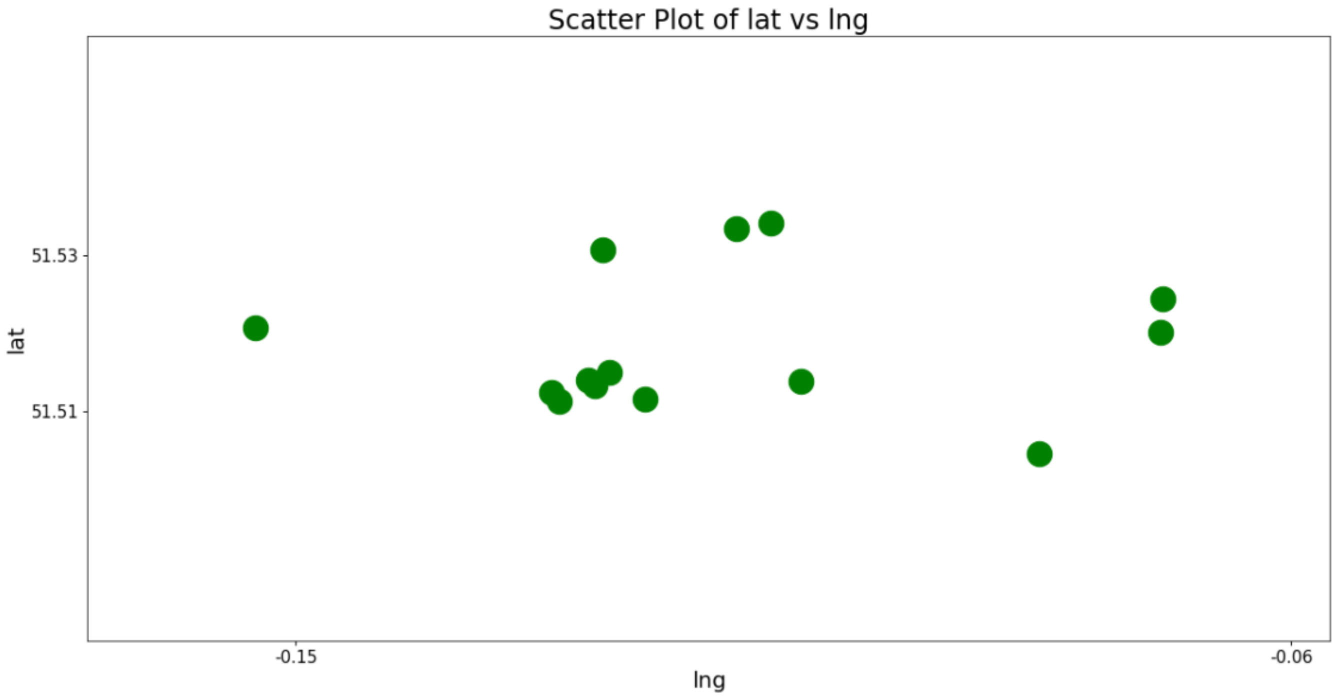
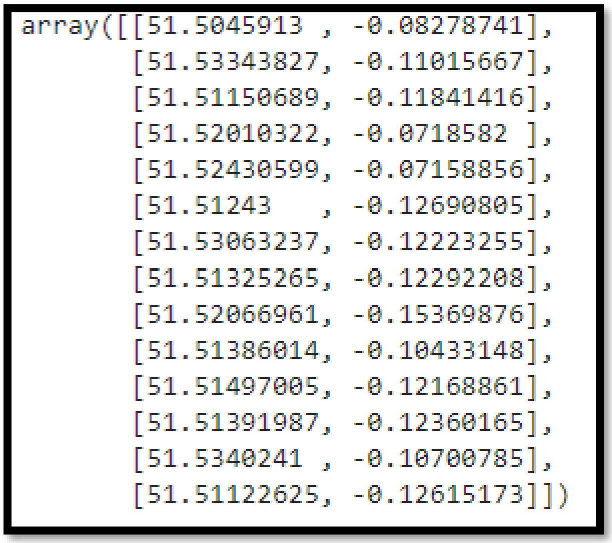


FIGURE 3 - SCATTER PLOT OF LATITUDE (LAT) AND LONGITUDE (LNG)

- The main difference in this plot compared to the Folium map plot is that the x and y axis has specific values and that there is no map in the background. This basically illustrates that the Folium map is based on two arrays of data being plotted.

- To run the KMeans technique on the data, I would have to transform the latitude and longitude columns for each observation into a NumPy array.



```
array([[51.5045913 , -0.08278741],
       [51.53343827, -0.11015667],
       [51.51150689, -0.11841416],
       [51.52010322, -0.0718582 ],
       [51.52430599, -0.07158856],
       [51.51243   , -0.12690805],
       [51.53063237, -0.12223255],
       [51.51325265, -0.12292208],
       [51.52066961, -0.15369876],
       [51.51386014, -0.10433148],
       [51.51497005, -0.12168861],
       [51.51391987, -0.12360165],
       [51.5340241 , -0.10700785],
       [51.51122625, -0.12615173]])
```

FIGURE 4 - ARRAY OF LAT AND LNG

Then I can run the clustering technique with a set of different clusters. I start with the minimum, which is 3:

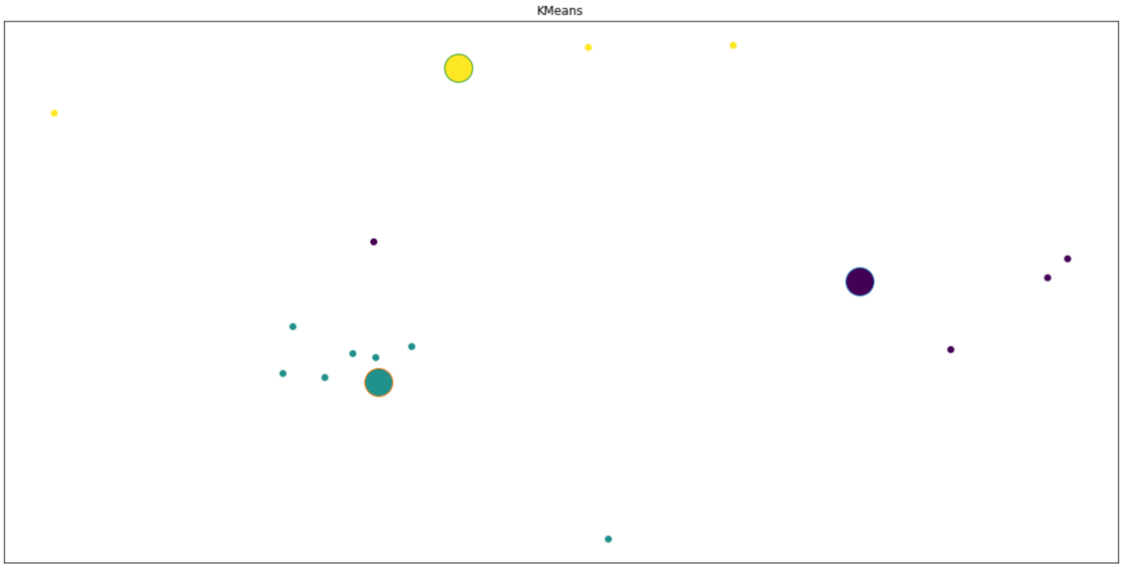


FIGURE 5 - KMEANS WITH NUMBER OF CLUSTERS = 3 AND INTERATIONS = 12

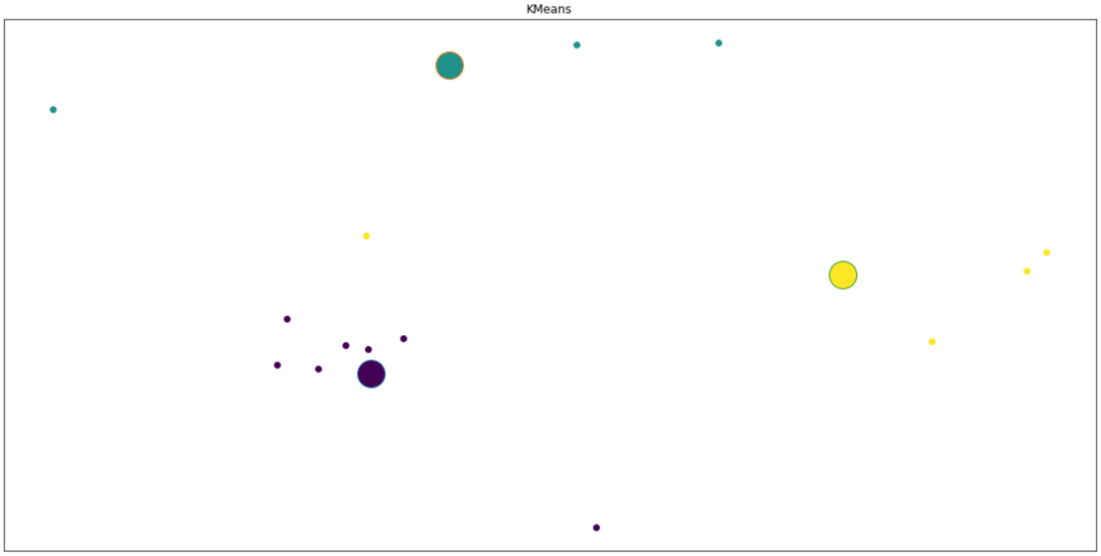


FIGURE 6 - KMEANS WITH NUMBER OF CLUSTERS = 3 AND ITERATIONS = 24

- As illustrated when comparing figure 5 and 6, there is little or no difference in increasing the number of iterations by 100% from 12 to 24.

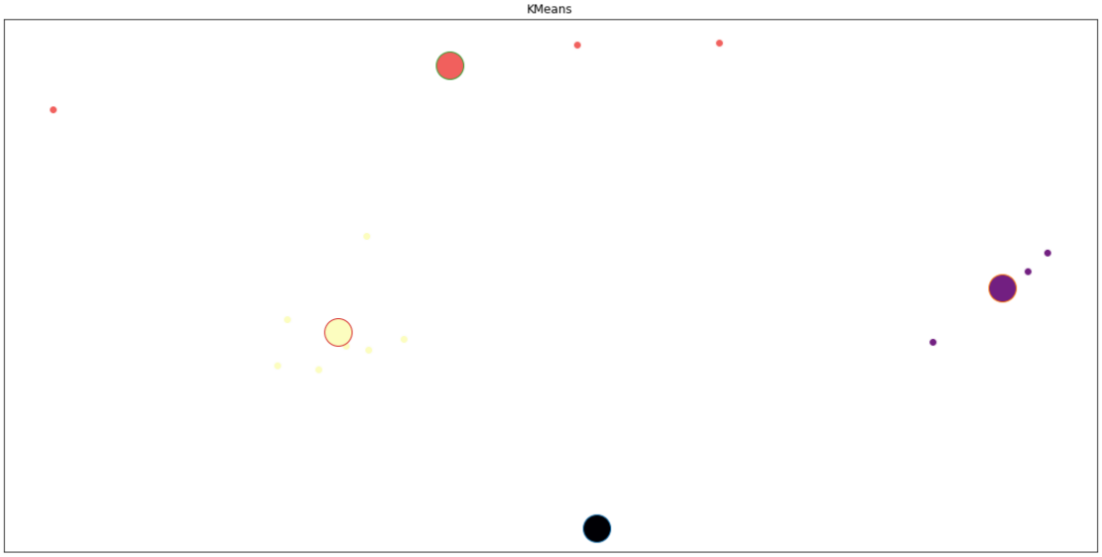


FIGURE 7 - KMEANS WITH NUMBER OF CLUSTERS = 4 AND ITERATIONS = 12

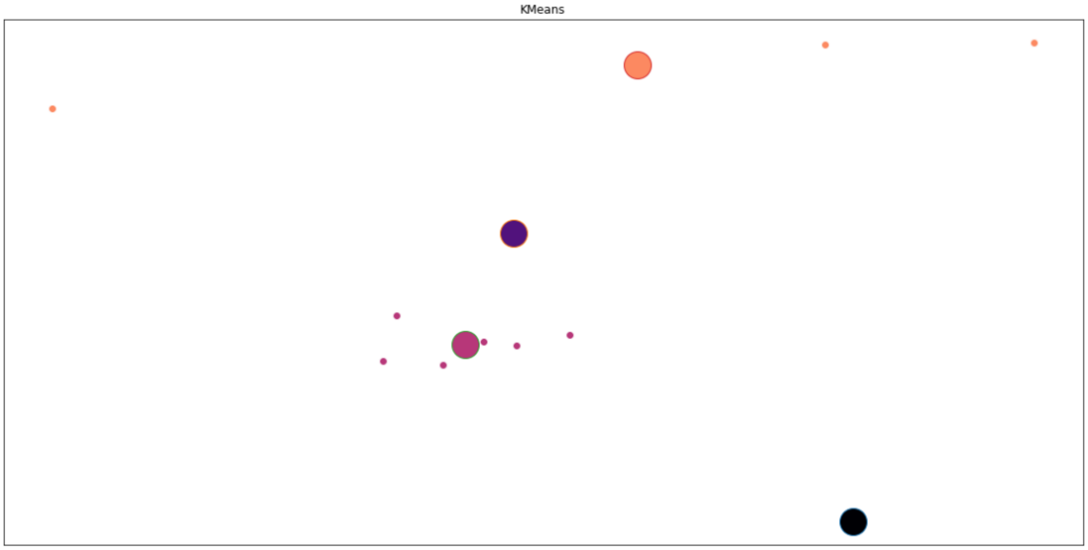


FIGURE 8 - KMEANS WITH NUMBER OF CLUSTERS = 5 AND ITERATIONS = 12

- As the number of clusters or iterations increase, the clustering does not seem to improve in performance. By looking at the data, it seems as if the total of 3 clusters and 12 iterations find the ideal spread of the clusters. However, we are most interested in the cluster furthest to the top (furthest to the North), seeing how this is where our starting point of the analysis was placed.

- From all clustering modules, this upper cluster (furthest to the top / north) is in the same area. That is, quite close to observation “number two” from the upper right corner.

- Given the assumptions of our analysis, the ideal location of our Indian restaurant should be this cluster. It satisfies the main criteria: It is centered around the surrounding observations and it has a distance weighted based on the means of all observations.

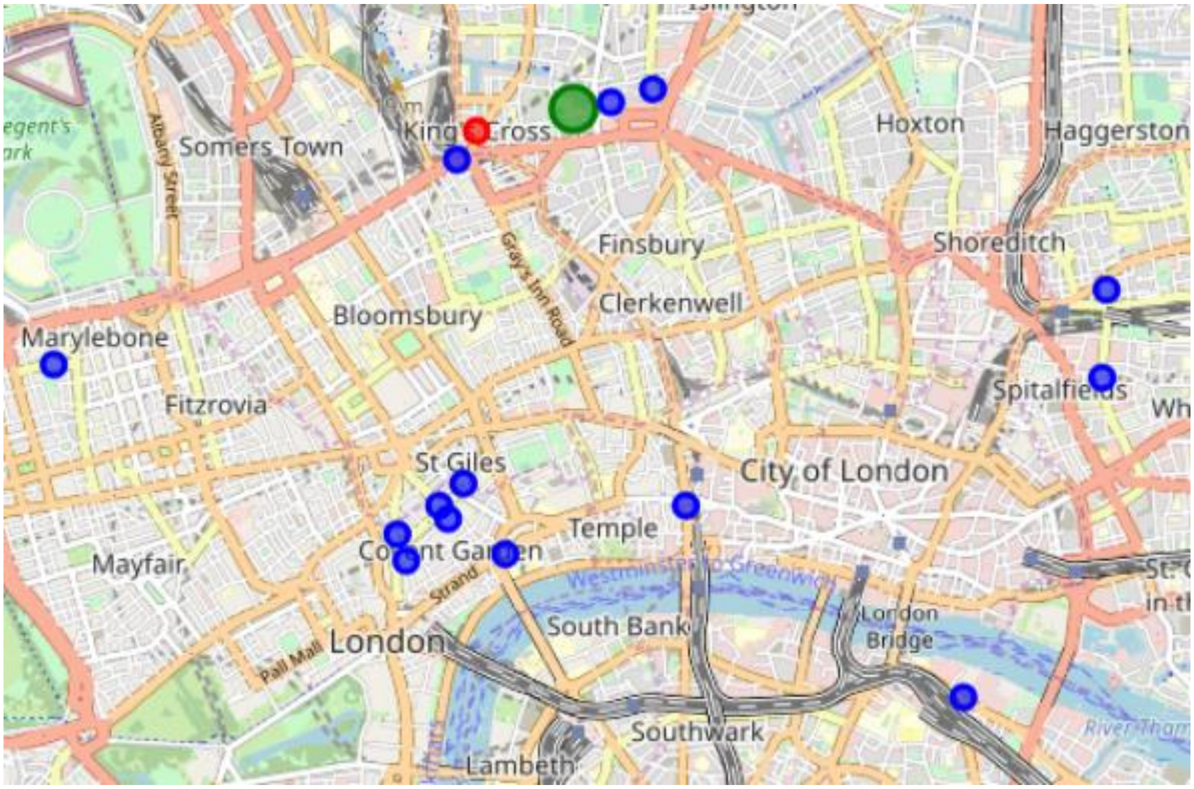


FIGURE 9 - FOLIUM MAP INCLUDING CLUSTER OF INTEREST

- To illustrate the main cluster of interest, I can plot the data in a Folium map, equal to what we did in Figure 2. This time, all observations are still marked as blue, and our starting point of the analysis at King's Cross is marked red. Additionally, notice the green marker in the upper/north part of the maps. This is our main cluster of interest, which was in the same area for each module. The plots for this location are 51.5330, -0.1131.

5. Discussion

- The data analysis in this report is limited by the availability of data on Foursquare. One might question whether Foursquare has a representative range of Indian restaurants in the northern parts of London. If, let's say, Indian restaurants tend not to register on Foursquare, then my assumptions for this analysis are not valid and hence my data quality is not good enough.

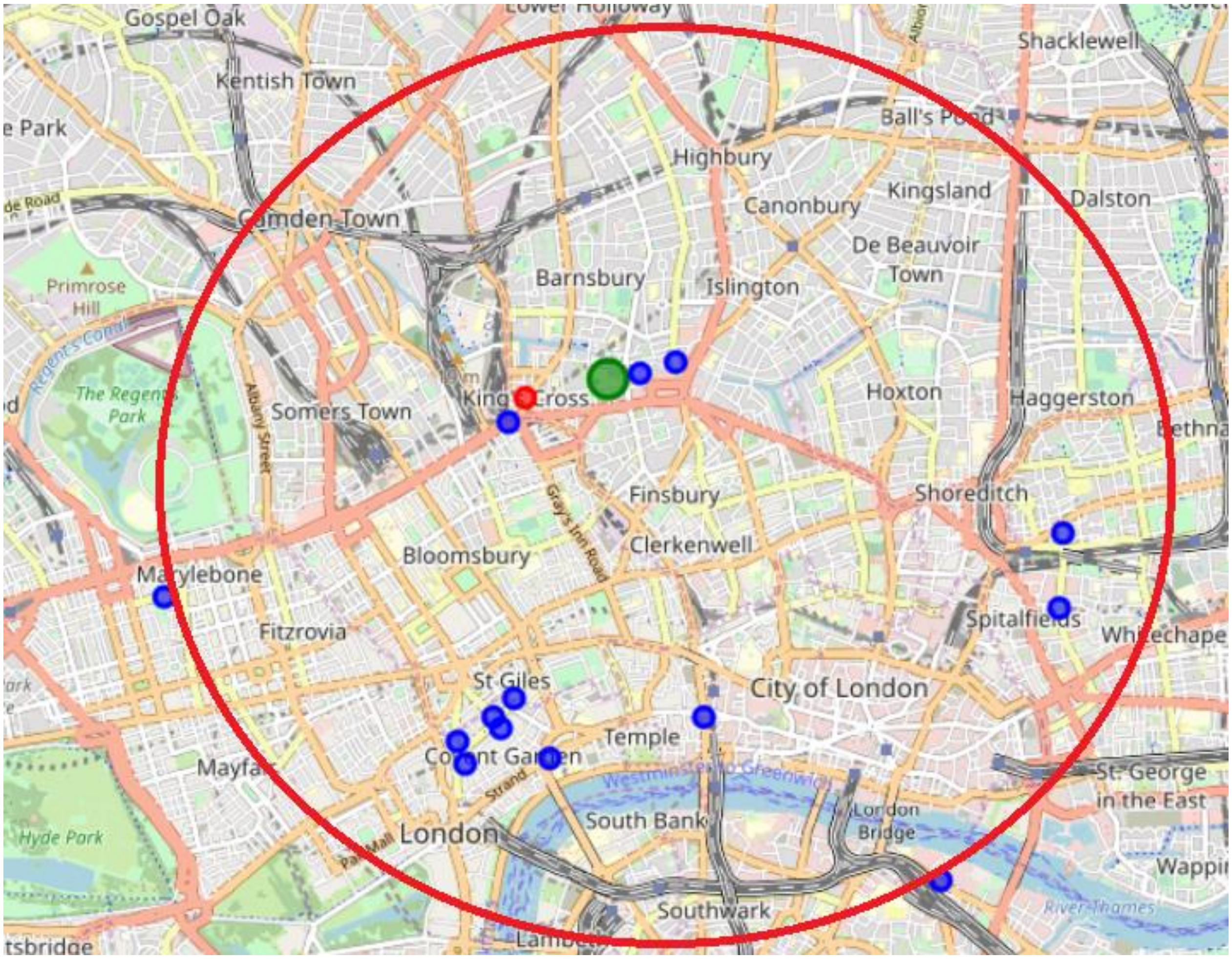


FIGURE 10 - GRID SEARCH FOR ANALYSIS

- My point is made clear when creating a circle / grid mark for the surrounding are of my analysis. If our search grid was 3 km surrounding the red marker, why is there no blue markers (i.e. observations of data) in the upper left, upper or upper right part of the circle? This is areas that are outside of the main city center in London, and one might question whether there is a tendency that only centralized restaurants are registered on Foursquare.

- Another point to question is the data quality when searching for ‘Indian restaurants’ on Foursquare. Can we assume that all Indian restaurants within the search grid are registered as an Indian restaurant for the category variable? This is not very likely, and hence, some observations of data are expected not to be included in the analysis simply because of this error.

- Improvements to be done could be to run cross-reference of other restaurant databases to see whether there are additional observations to be included.

6. Conclusion

- In this report, I have use KMeans clustering techniques to find the ideal location for an Indian restaurant owner in the London area. I optimized the location for the restaurant by creating clusters of the data which represents observations in groups.

- The ideal cluster was the one centering the starting point of the analysis. This cluster was found using multiple iterations and cluster settings, which were presented in the Results section of the report. The grid for the cluster was extracted to represent the ideal location for the restaurant, which is represented by a location in the middle of the surrounding observations.