Starting an Indian restaurant in Scotland

Murali Rajamanickam

13 May 2020

Table of Contents

Introduction ***3***

Clients request **3**

The suggested study **3**

Data Acquisition and Cleaning ***4***

Data Acquisition **4**

Data Cleaning **4**

Exploratory Data analysis ***6***

Distribution of each restaurant by post code **6**

Postcodes with no Indian restaurant in top 5 ..... **7**

Postcodes with Indian restaurant in top 5 ..... **7**

Revising the initial data set **8**

Predictive Modelling ***9***

Preparing the data for Clustering **9**

K-Means Clustering **9**

Map cluster 3 9

Map Cluster 4 10

Map Cluster 5 11

Results and Discussion ***12***

Results **12**

Cluster 0 12

Cluster 1 12

Cluster 2 13

Cluster 3 13

Cluster 4 13

Discussion **13**

Conclusion ***13***

Introduction

Clients request

A client living in Scotland would like to invest some money opening an Indian restaurant, but he doesn't know where to start from. He has looked into several bit cities, such as Glasgow, Edinburgh, Stirling, Dundee, Inverness, and Aberdeen. The reason for him picking these cities are because they are major historical major cities and attract a lot of tourists every year. He would like us to guide him in picking the most suitable city and the area of choice.

The suggested study

Scotland is a very outgoing country with a huge emphasis on tourism. According to the Office of National Statistics (ONS), in Q2 2017 – Q1 2018 3.4 million overseas tourists visited Scotland, spending £2.4 billion which was an increase of29% compared to the year before. Furthermore, the food and drink industry are worth around £14 billion each year and accounts for one in five manufacturing jobs in Scotland. On the other hand, according to the U.S. Bureau of Labour Statistics (BLS), approximately 20% of new businesses fail during the first two years of being open, 45% during the first five years, and 65% during the first 10 years. Data from the ONS from 2018 also suggest that the food and beverage service is very competitive, as 18,820 new restaurants and mobile food services were born, and nearly as much 15,930 were observed dead. With an overall industry of 105,730 restaurant and mobile services, it's a nearly 15% of the market that is lost, and 18% of the market born.

To secure a sustainable investment several factors has to be taken into account, this includes:

1. The city where the restaurant is located (the ratio of restaurant facilities to amount of population)

2. The area where the restaurant is located (the amount of restaurants located in a certain area)

3. The type of food served (the cuisine served should be unique or least represented in the area of choice)

With all of these criteria in mind, a set of data will be gathered and studied to demonstrate the most viable place to start a new business.

The client has decided to choose between 6 cities in Scotland (Glasgow, Edinburgh, Stirling, Dundee, Inverness, and Aberdeen), therefore data on restaurants from these cities will be collected with the location of these places on a map. Then the several criteria's will be studied, such as the number of restaurants to the city's population. Are there too many restaurants in a small city? or which city presents the least amount of restaurant per head? The proportion of Indian restaurants will also be looked at and compared to the amount of people present in

the city. Finally, with the most appropriate city in hand, the restaurants will be clustered and the most ideal area will to open the restaurant will be defined.

Data Acquisition and Cleaning

Data Acquisition

The initial data was extracted using the Foursquare API. The coordinates of each city was collected from Google Maps, and used to collect the required data. The Initial research was to get the data on all restaurants around the city (with a radius around the centre point defined for each individual city), but the API search could only allow a maximum of 50 results (in reality I never observed over 30 results), therefore it was decided to only get query from existing Indian restaurants.

Data Cleaning

All initial results were gathered on a table for better appreciation of the results. As shown in the results below (Table 1) all query gave below 30 results; suggesting that they were all captured, and none were missed. Another few values were also added, such as the number of population in each city (according to the latest figures) and the ratio of population/restaurant. This ratio allows us to define the possible frequency of usage per restaurant, i.e. the higher it is, the more customers the restaurant can attract. When the values are distributed in descending order by ratio, Stirling comes as the city with the most potential for opening an Indian restaurant.

	Cities	Population	Lattitude	Longitude	N_restuaurants	ratio
2	Stirling	94330	56.11903	-3.93682	1	94330
5	Aberdeen	227560	57.14369	-2.09814	4	56890
3	Dundee	148750	56.46913	-2.97489	3	49583
0	Glasgow	626410	55.86515	-4.25763	13	48185
1	Edinburgh	518500	55.95206	-3.19648	25	20740
4	Inverness	47380	57.47908	-4.22398	5	9476

Table 1: Number of Indian restaurants in major cities in Scotland and their ratios per population

On a critical note, one can observe that this data is a bit incoherent. The number of Indian restaurants in Stirling seems too low, furthermore, Glasgow with the highest population in Scotland have lower numbers of Indian restaurants than Edinburgh. So to have a better

appreciation, a query was launched on TripAdvisor website for Indian restaurants in each city and the results are reported.
As observed below (Table 2) the results with TripAdvisor are much higher and more realistic than those observed with foursquare. The results indicate that the Glasgow has the biggest amount to Indian restaurants, and Stirling has 11 not 1. When calculating the ratio with the new results and sorting them in ascending order Stirling and Dundee come up as the two places with the highest potential.

	Cities	Population	Lattitude	Longitude	N_restuaurants	ratio	tripadvisor_data	ratio_tripadvisor
0	Glasgow	626410	55.86515	-4.25763	13	48185	192	3262
4	Inverness	47380	57.47908	-4.22398	5	9476	14	3384
[illegible]	Aberdeen	[illegible]	[illegible]	-2.09814	4	56890	59	3856
1	Edinburgh	518500	55.95206	-3.19648	25	20740	131	3958
3	Dundee	148750	56.46913	-2.97489	3	49583	29	5129
2	Stirling	94330	56.11903	-3.93682	1	94330	11	8575

Table 2: Restaurant query using Foursquare and TripAdvisor, and their respective ratio population/restaurants

Unfortunately, the restrictions applied on the Coursera project using Foursquare API data, restricts us from using these cities. It appears that Foursquare is not as readily used in Scotland and therefore doesn't provide an accurate picture of reality, which in consequence would not allow us to provide any conclusive data. Furthermore, the search on Foursquare doesn't provide enough data point on small cities such as Stirling and Dundee to do any work with data science tools.

With all of the above in mind, it was decided to randomly pick Glasgow as the city to focus our attention on. Being the most populated city, and having lived there for over three years, I could make better judgements of the obtained results.

To have a better search/results pattern Glasgow was divided manually in different points. This has been done as the postcode system in the UK is too precise and only covers very small distances. On the other hand, the larger postcode are, covers a wider and non-symmetrical area which would have the consequence of not covering all the areas.

Once all areas of Glasgow covered, the query was launched on all existing restaurants within those coordinates. From the obtained JSON file was the relevant data was extracted, such as the name of the restaurant, its unique ID, the corresponding coordinates and also the category.
While the method used to collect data was in such a way that the entirety of the data was collected within Glasgow, a lot of the restaurants were obtained in duplicates. Once the duplicates removed (by restaurant id, and therefore none were missed) the shape of the dataframe showed the presence of 72 restaurants within the vicinity of Glasgow.
It should be noted this value is way below those observed when the query of Indian restaurants in Glasgow was carried out on Tripadvisor, suggesting a lack of data on Foursquare.

Once all the categories section filled with data (two of them didn't have their categories identified) the number of Indian restaurants in Glasgow was of 6 over a total of 33 different types of restaurants. A shown in the map below (Figure 1) all the distribution of restaurants seems to cover most of the area of Glasgow.

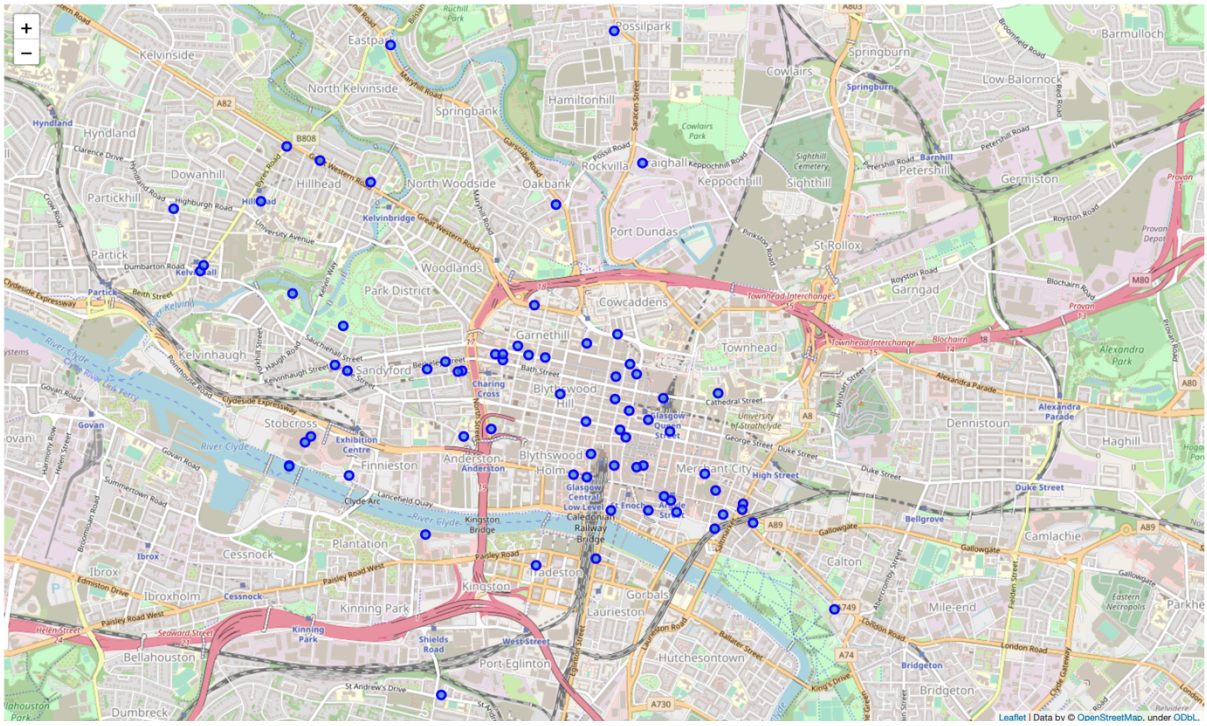

Figure 1: Map of Glasgow with all restaurants and their locations

Exploratory Data analysis

Distribution of each restaurant by post code

The data obtained so far has been string values corresponding to the name of the venue, therefore in order to change them to calculatable integers, each character type was changed to a column and given a value (number of existing restaurant type in each post code). Once this has been done the data was standardized.

Note: The total number of postcodes has decreased since removal of the duplicates. This was expected as more postcode positions were given to cover the full extent of data.

All the top 5 categories of each postcode were defined. At a first glance of the 12 postcodes used, 50% of then didn't have Indian in their top 5 categories, and only one (G2_1) had Indian restaurants as the major type of restaurant in its vicinity.

But a more detailed look at the situation is required. Let’s look at the 50% of values that doesn’t contain Indian restaurants in their top 5 categories.

Postcodes with no Indian restaurant in top 5

As shown in the graph below (Figure 2) when plotted on the ratio of the top 5 categories in each post code, some don't categories are not listed. When in G1_1 the frequency of Restaurants, bar, Mediterranean Turkish and Chinese have a standardised value associated with it, in G3_6 only a value for Chinese in present, and no other values are available.

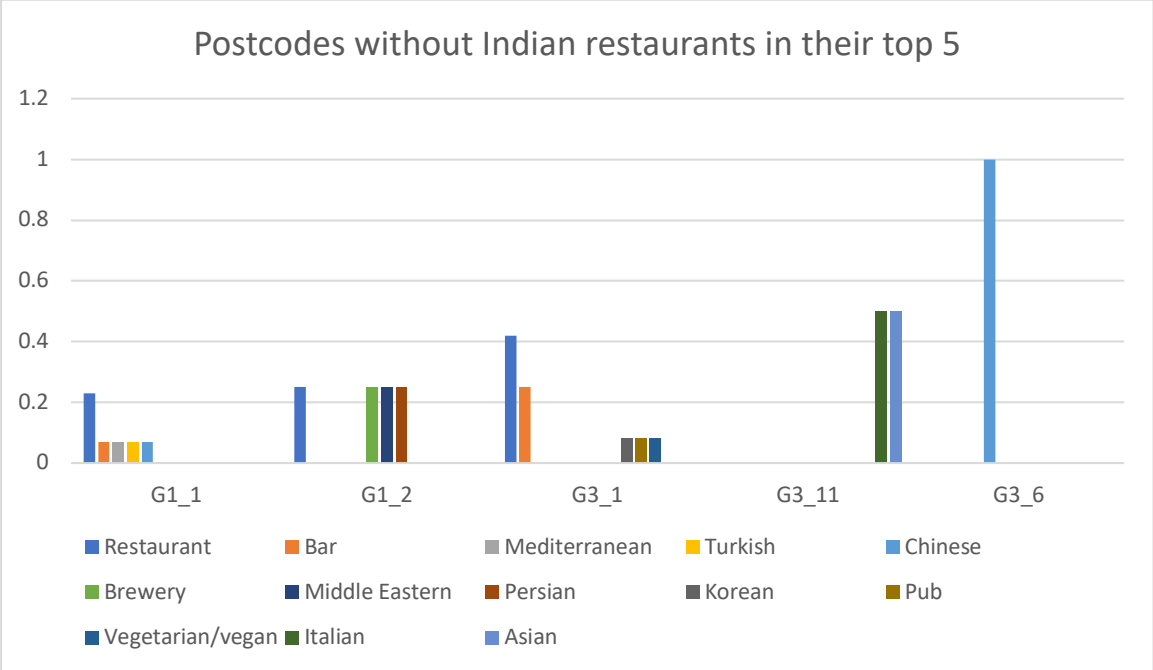

Figure 2: Distribution of restaurants in Glasgow without Indian restaurants in their top 5 category

This indicates that the lack of representation in restaurants and types in each postcode only gives a few categories.

Postcodes with Indian restaurant in top 5

The other set of results, showing the top 5 categories having Indian restaurants is represented below (Figure 3). As it can be observed below, only three of the 5 postcodes (G2_1, G3_2, G5_1) contain Indian restaurant values, the other have only few restaurants and the other values have been populated randomly.

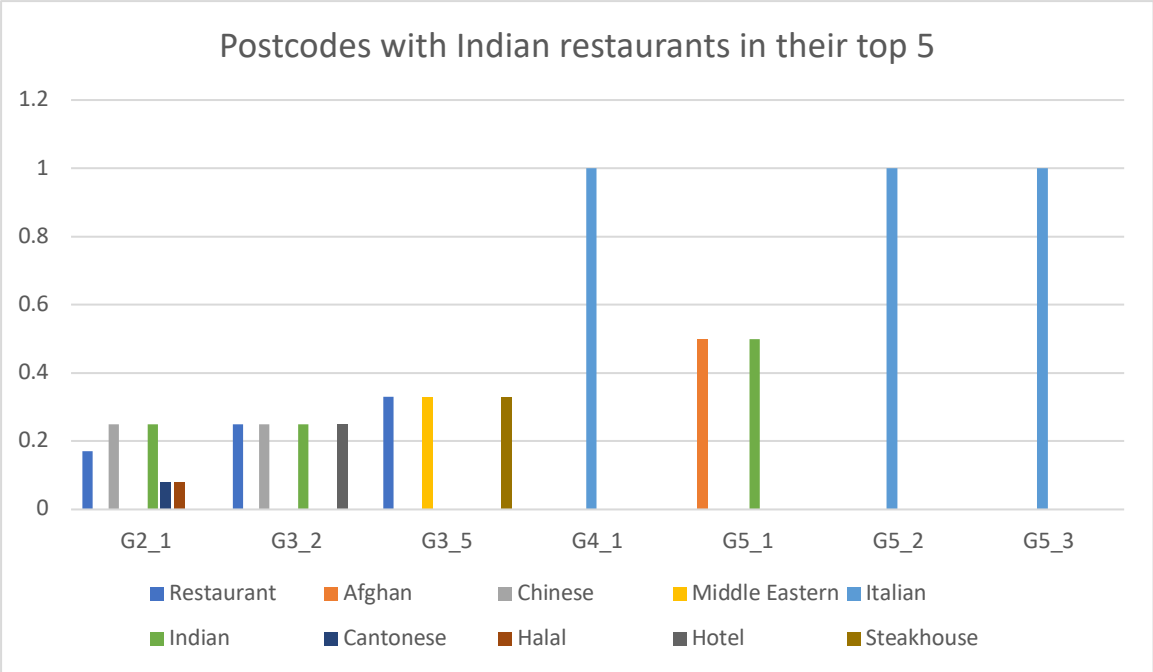

Figure 3: Distribution of restaurants in Glasgow with Indian restaurants in their top 5 category

Revising the initial data set

Instead of taking all obtained data in Glasgow as initially planned it was decided to clean some of the data dataset to avoid adding non existing values in the calculation. When looking at restaurants by postcodes (Figure 4), it was observed that one postcode englobed 30 restaurants (G1_1), whereas postcodes G2_1 and G3_1 had 12 each. Finally, G1_2, G3_2 and G3_5 had 4, 4 and 3 respectively. Only these values were kept, as the others had 2 or 1 restaurants in the allocated post code.

]:

Gla postcode	Gla Latitude	Gla Longitude	name	rest Latitude	rest Longitude	rest id	rest Category
G1_1	30	30	30	30	30	30	30
G1_2	4	4	4	4	4	4	4
G2_1	12	12	12	12	12	12	12
G3_1	12	12	12	12	12	12	12
G3_11	2	2	2	2	2	2	2
G3_2	4	4	4	4	4	4	4
G3_5	3	3	3	3	3	3	3
G3_6	1	1	1	1	1	1	1
G4_1	1	1	1	1	1	1	1
G5_1	2	2	2	2	2	2	2
G5_2	1	1	1	1	1	1	1
G5_3	1	1	1	1	1	1	1

Figure 4: Restaurants distribution in Glasgow per postcode

The new dataset thus created contains 65 restaurants, with 31 unique categories. This new cleaning therefore had small impact on the overall dataset.

Predictive Modelling

Given the range of different restaurants present in the city of Glasgow, each restaurant type was defined as a value. The idea behind this is to define the areas where the presence of Indian restaurants is low and therefore implementing a new Indian restaurant would increase its success. In fact the low presence of one type of restaurant gives a wide range of clients readily accessible, and also reduced the distance of Indian restaurants between two relevant zones.

Preparing the data for Clustering

First it was decided to define the best 3 best types of restaurants in each area. This was decided as some areas only had a 3 different restaurants types. To do this each restaurant type was given a number corresponding to the number of restaurants present in each post code. Then the values were standardised to give the relative proportion of each restaurant present. Finally the value of each ranked from high to low. As can be seen below (Figure 5), a quick representation of the top three restaurants for each postcodes show interesting results. For example G2_1 shows Indian restaurant as the most common type in that postcode, and therefore an area to avoid.

	Gla postcode	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant
0	G1_1	Restaurant	Turkish	Bar
1	G1_2	Persian	Brewery	Restaurant
2	G2_1	Indian	Chinese	Restaurant
3	G3_1	Restaurant	Bar	Vegetarian / Vegan
4	G3_2	Hotel	Restaurant	Chinese

Figure 5: Top 3 resturant types in each postcode

K-Means Clustering

K-Means clustering seems to the most efficient way to cluster the different types of values within the city of Glasgow. The range of different number of clusters were also explored, form 3 (having the top three values) to 5 clusters. The clustering was stopped at 5 to avoid too many isolated points and over fit the results.

Each of the clustered data points were placed on the map with a different colour for each cluster. Lets examine each map to better pick the most suitable cluster.

Map cluster 3

As it can be observed in the map below(Figure 6), all three cluster are represented. The cluster 0 represented in red, is present in the east and west part of Glasgow. This represents the one of the most active areas of Glasgow. The purple spots represent cluster 1, which is located in the middle of these cluster 0. And finally, the green spots represents cluster 2, which seems to be located in the outskirts of the city of Glasgow.

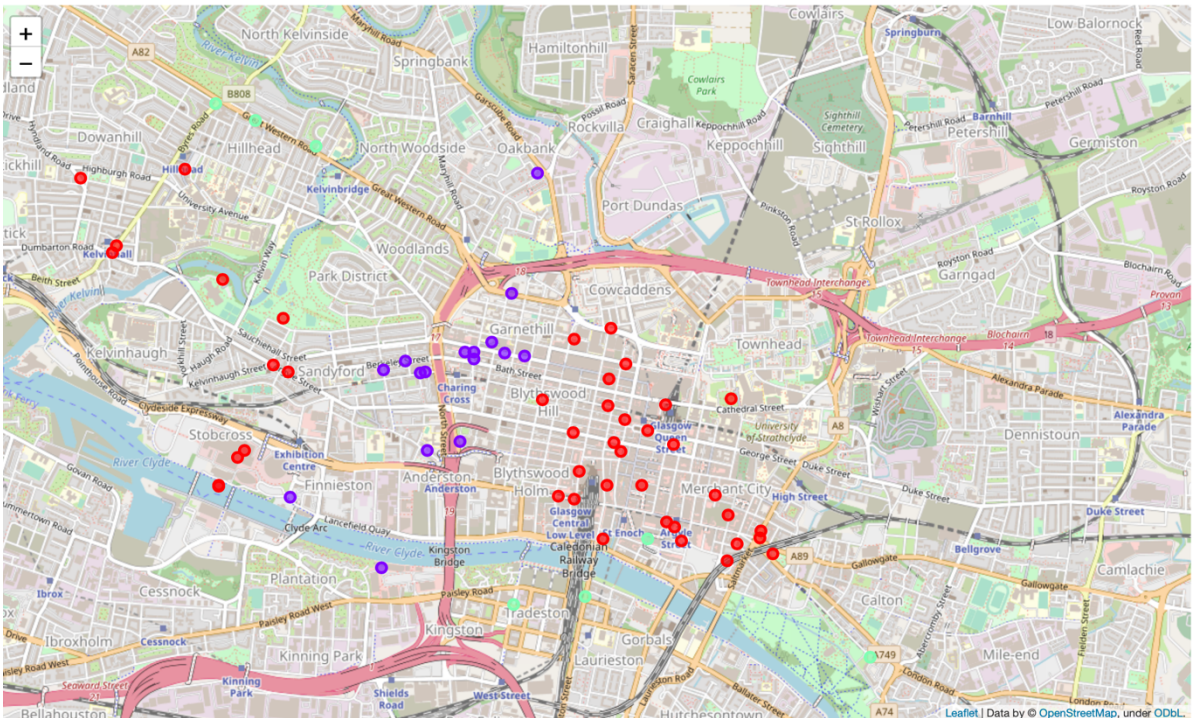

Figure 6: Clustering of restaurants in Glasgow with 3 clusters

The above representation seems too general as a wide area is occupied by cluster 0. Another cluster might be necessary to resolve this issue.

Map Cluster 4

As one can observe in the map below (Figure 7) the clustering is nearly similar to the one observed with 3 clusters. Dark green represents Cluster 3 occupying both east and west side of the map. Cluster 0 is in red, splitting cluster 3 and finally cluster 2 and 1 (blue and purple) are the two different clusters of the outskirts of Glasgow.

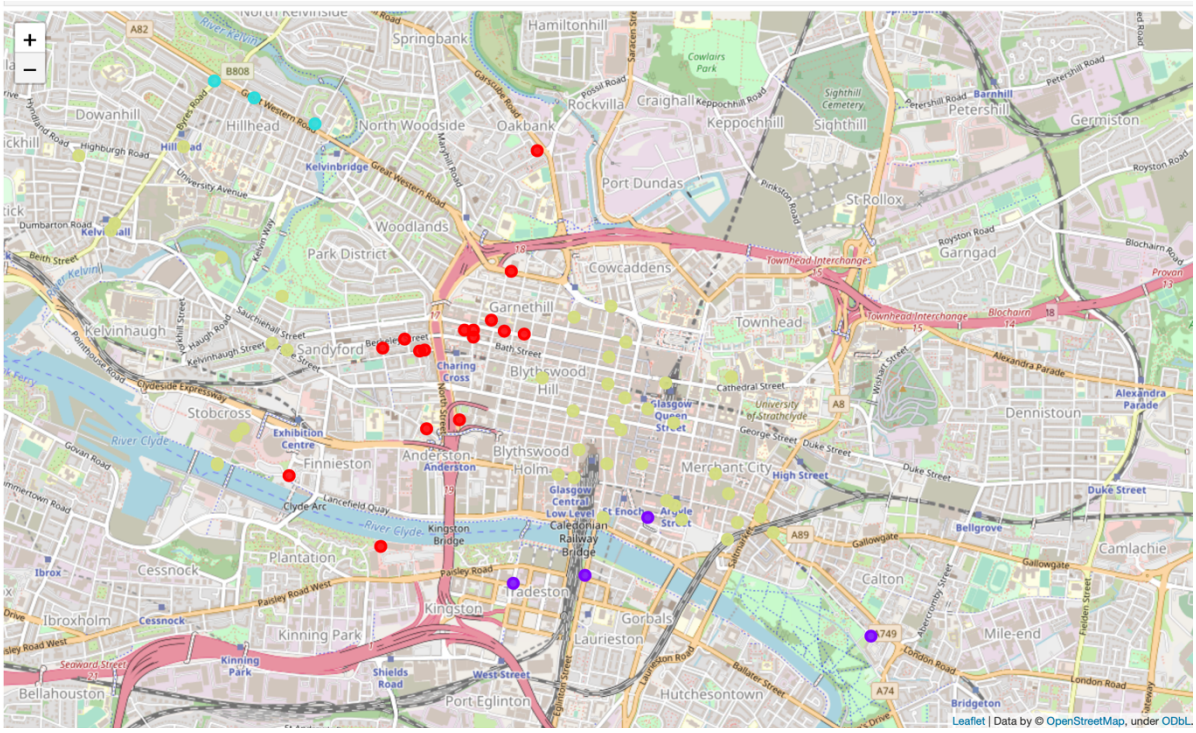

Figure 7:Clustering of restaurants in Glasgow with 4 clusters

At this point, although 4 clusters allow a better representation in space, this doesn't allow a separation of the two sections observed in cluster 0 making it too global to study.

Map Cluster 5

Here in the map (Figure 8) with 5 clusters calculated by K-mean, the results show a clear spatial representation of the areas. Red spots are cluster 0, purple cluster 1, blue cluster 3, turquoise cluster 2, and orange cluster 4. This represents a more homogeneous split of the area of Glasgow and allow us to look for the area which would be ideal for the new business.

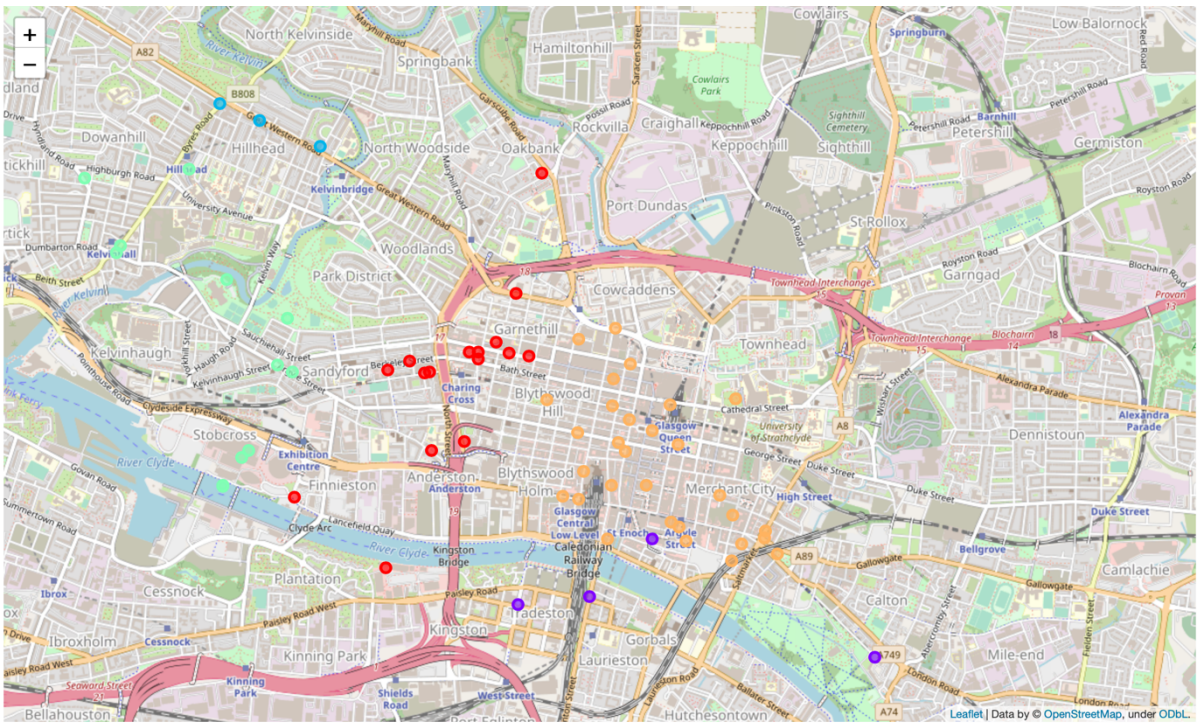

Figure 8:Clustering of restaurants in Glasgow with 5 clusters

The map with 5 clusters seems to show the best representation of what we want to achieve. It has split the area of Glasgow in several clusters geographically different, and allow us to study each cluster, and evaluate which cluster region is the most suitable for opening an Indian restaurant.

Results and Discussion

In here we will examine the results obtained with each cluster one by one and define the best outcome of this study.

Results

Cluster 0

When cluster 0 is studied carefully; it showed 16 restaurants, from which the most common ones observed were Indian, Chinese, Restaurant and Hotel. The high prevalence of Indian restaurants in this area shows that it's the least likely areas to invest in Glasgow while opening an Indian restaurant.

Cluster 1

Cluster 1 only contains 3 data points, a Brewery, a Persian and a Middle eastern restaurant. Although having no Indian restaurant in this cluster, the number of restaurants in this cluster seems to suggest that the area is not the most attractive. Nevertheless it should not be discarded.

Cluster 2

In Cluster 2, 12 restaurants are present, and the top 3 type are Restaurant (classic), Bar and Vegetarian/Vegan restaurants. This is an extremely interesting area containing a large number of classic restaurants and pubs, and no presence of Indian restaurants.

Cluster 3

In Cluster 3, there are again only 3 restaurants, a Steakhouse, a Restaurant (classic), and a Middle eastern. As with Custer 1, the small presence of restaurants is not attractive, but the non-existence of Indian restaurant should allow it to not be discarded.

Cluster 4

This Cluster contains 30 restaurants, representing the major portion of the Clusters, and the top three common restaurants are Restaurant (classic), Turkish and Bar. This area represents the busiest area with high potential for an alternative Indian restaurant.

Discussion

While looking through the clusters obtained above, clear evidence shows that cluster 0 should be discarded. Clusters 3 and 1 are interesting but containing only a small number of restaurants in those area shows a potential for lower customer presence. For that reason, those two clusters should also be discarded.

In light of Cluster 2 and 4, both seem to be sensible starting points for starting a new restaurant. Both have the advantage of not having an Indian restaurant in its top 3 favouring them and both have a high presence of restaurants signalling high customer presence.

My personal recommendation would be on cluster 2 rather than cluster 4. This is because of the high density of cluster 4, which is the reverse effect of clusters 3 and 1. In the cluster 4 there are too many restaurants in a small zone and therefore less head count, whereas the cluster 2 seems to a more even distribution and contains a lot of restaurants showing attractiveness and viability.

Conclusion

In this Study, a client request to find the most suitable place to start an Indian restaurant was carried out. Initially the study was interested in picking the city with the lowest ratio of Indian restaurant per population, but very quickly it was found that the data present on Foursquare was not enough to obtain a good representation of reality. Having used other media to understand the fact, it was randomly decided to pick Glasgow, the biggest city in population in Scotland to carry out the analysis. For this, all restaurants in Glasgow were collected and defined by their category as shown on Foursquare. Then the data was cleaned before clustering them according to the top three restaurant types in each cluster. K-means clustering was used and having 5 clusters was shown as the most appropriate separation. Finally, study of each cluster revealed that cluster 2 and 4 (See Figure 8) were suitable for starting an Indian restaurant as they contained a high amount of restaurant with no presence of Indian restaurants. The preference was given to cluster 2 (turquoise Figure 8) as it had the good ratio of density of restaurants with the absence of any Indian restaurants.