IBM Data Science – Capstone Project

Finding an Optimal Restaurant Location and Cuisine Preference for Visitors in Birmingham, UK

K. Patel

February 2021

Table of Contents

Introduction

Background

Birmingham is a diverse, multicultural city located in the UK with many visitors every year for leisure or business purposes. The city is well known for the number of different types of restaurants and cuisines available and is famous for being the home of Balti-houses, a type of Indian restaurant.

Visitors to Birmingham often look for hotels located near amenities and restaurants, which local businesses often target.

Business Problem

The relationship between hotel locations and restaurants, and restaurant types needs to be investigated. Data is needed to locate hotels in Birmingham, the density of restaurants around these hotels, and the most popular or represented cuisines. This report will provide an insight into the distribution of restaurants vs hotel locations and where there could be opportunities to open a restaurant as well as the ideal restaurant type.

Interest

This analysis can be a useful tool for an investor looking to start a restaurant business in Birmingham, as the report provides useful information on the hospitality services around the city center.

Data

Data Source

For the purpose of this study, location data is required that describes hotels and restaurants around Birmingham city center. Specifically, location data, number of restaurants around hotels, foot traffic in restaurants and their category. To retrieve this data, the Foursquare API was utilized in two stages, one for hotel information, and one for restaurant information.

Data Cleaning

For the hotels dataset, a “Search” call was made to the Foursquare API with the search query “Hotel”, with the aim was to identify hotels around Birmingham city center. After the call is made, results need to be organized into a Pandas data frame.

Data wrangling is performed to keep only the required data attributes related to the name and location of each venue as well as extracting the venues’ categories to verify the results. Results are then organized and viewed in a Pandas data frame, displaying some categories other than “Hotel”. These results were dropped from the data frame in order to focus on hotels only.

For restaurant data, the “Explore” call was made to the API to retrieve data on which venues were trending. Likewise, results were organized in a data frame and cleaned from any unnecessary features other than venue name, cuisine type and location data.

Feature Selection

To find which areas have groups of hotels close to one another, and the number of trending restaurants in each area, two new data frames were created one for hotels and for restaurants, each holding only the latitude and longitude coordinates of the original datasets to aid the Machine Learning algorithms explained in the next section.

Methodology

Finding Key Areas

As a first step to finding the best spot for a restaurant targeting visitors near the city center of Birmingham, it is required to know which areas have high density of hotels while meeting the criteria of being close to the city center. Since the criteria referred to earlier is already met when collecting the data from Foursquare API by selecting a suitable radius for our search, the critical step here is to combine these hotels in groups (clusters) where each cluster will represent an area where a certain number of hotels are located.

K-Means Clustering

To cluster the retrieved hotels into clusters, we need a Machine Learning clustering algorithm that can find similarities among data points and group them accordingly. Thus, K-Means Clustering algorithm was found to be most suitable due to its simplicity and effectiveness.

A Pandas data frame containing hotels location coordinates was created, and serves as our input to the clustering algorithm, but the challenge is finding the appropriate number of clusters to group the data into. To overcome this issue, we visualized the hotels on a map centered around Birmingham using Folium library and tried to inspect the data visually first. Upon visual inspection [Figure 1], it is confirmed that 3 cluster would be suitable to group the data points and was therefore used in the clustering algorithm.

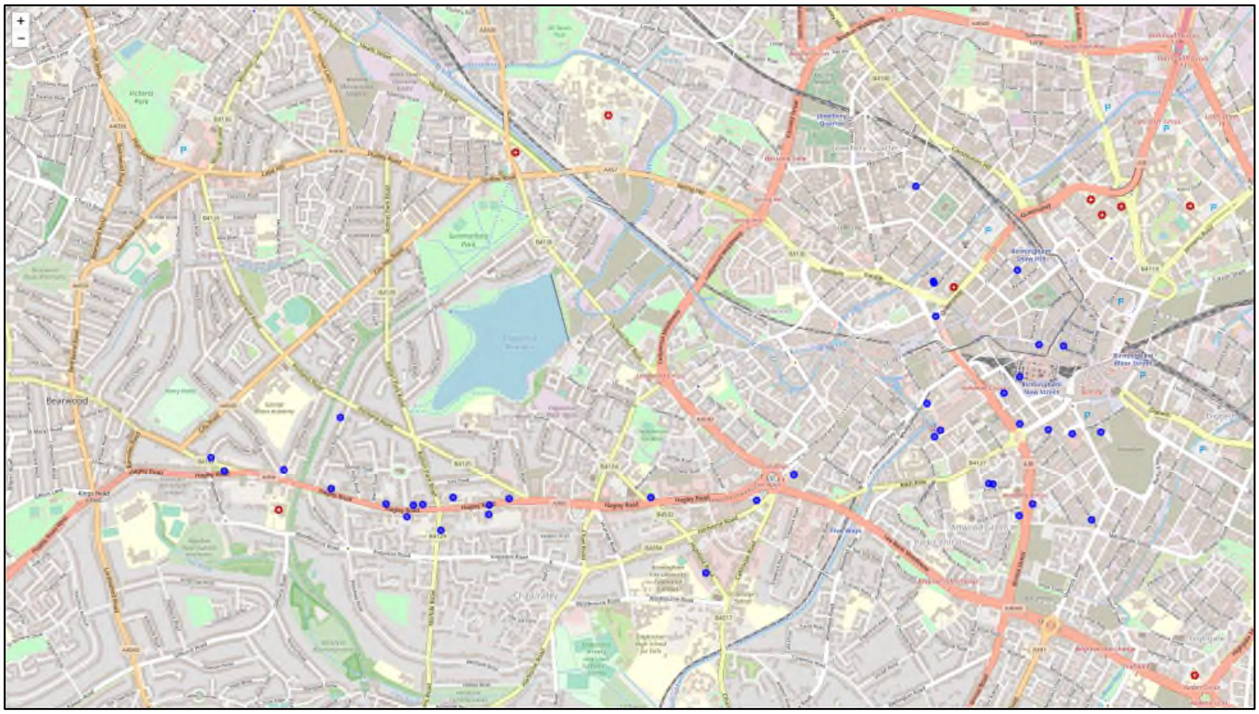

Figure 1. Hotel locations around Birmingham City Centre - extracted from Foursquare API.

The output from the clustering algorithm was added as a column to the locations data frame which was in turn merged back to the original hotels data frame, resulting in a complete table containing each hotel's name, location, cluster, etc. However, the most important output was the centers on each cluster, which was also saved in another data frame containing the label of each cluster and its center's location coordinates (latitude and longitude). These are summarized in the Results section.

Locating The Best Spot

Once the key areas were defined, the next step is to analyze competition in each area by finding the number of trending restaurants there. However, the first step in achieving this is to classify the trending restaurants retrieved from Foursquare API into the clusters obtained from the K-Means algorithm.

K Nearest Neighbors Classification (KNN)

For the completion of this task, KNN classification algorithm was used to assign a class label to each restaurant in a data frame created earlier that contains only the coordinates. First a visualization of the restaurants locations is generated [Figure 2] to try to predict the KNN results, with an early observation that with comparison to the hotels locations, all hotel clusters have some restaurants nearby.

KNN is a Supervised Machine Learning classification algorithm that assigns a label (class) to each data point according to the most frequent class among the nearest, user defined "K" number of neighbors. In other words, since the algorithm is supervised, we need to train the KNN model on a training dataset, and we need to specify the number of neighbors "K" that the classifier will use.

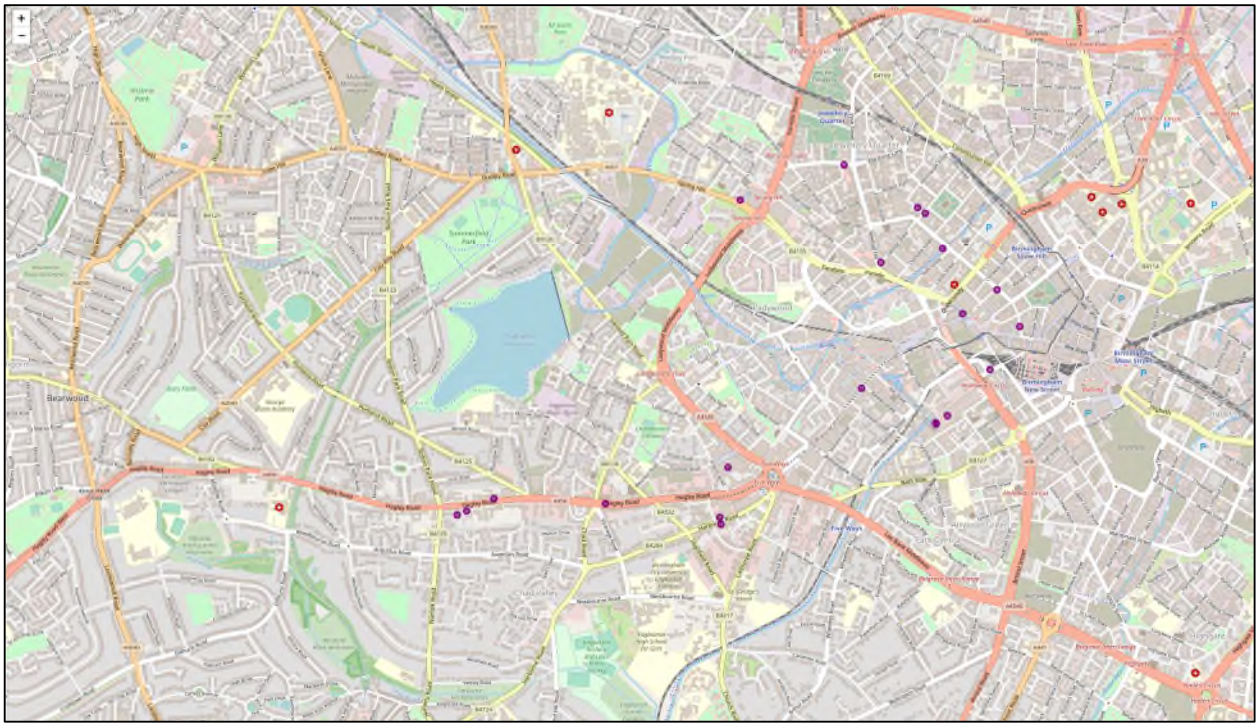

Figure 2. Restaurants (with specified cuisine types) locations around Birmingham City Centre - extracted from Foursquare API.

Therefore, we use the cluster center location data as a training set to tell the KNN model what location each class is center around, and we use a “K” value of ONE as we want to assign the label of the closest single center to the data point in the restaurants dataset. The resulting class labels were added as a column in the original restaurants data frame and visualized on the map by assigning a different color to each class of restaurants, similar to the hotels’ output visualization (K-Means output).

Opportunity Evaluation

After grouping hotels in clusters, finding the center coordinates of each cluster, and assigning a class to each restaurant according to the nearest cluster center; an evaluation of opportunities is carried forward. A Simple evaluation is to count the number of restaurants in each cluster to choose the one with the lowest level of competition, and then count the number of hotels in that cluster and compare it to the other clusters to evaluate opportunities. Simple value counting techniques were used in this evaluation.

Trending Restaurant Category

Once the location has been chosen, the question is “What should the new restaurant offer?” To answer, we used a simple frequency counting technique against the restaurants data frame and inspected the top five venues as these venues: first, have high foot-traffic because they were obtained through an “Explore” call to Foursquare API, second, are most occurring among the retrieved restaurants. The resulting ‘high trending’ restaurant type is compared against the opportunity evaluation described earlier to see if there are any potential opportunities in a certain area.

Results

Hotels Clustering

After grouping the hotels in three clusters by the K-Means Clustering algorithm, the resulting data frame was visualized by assigning different color to each cluster [Figure 3].

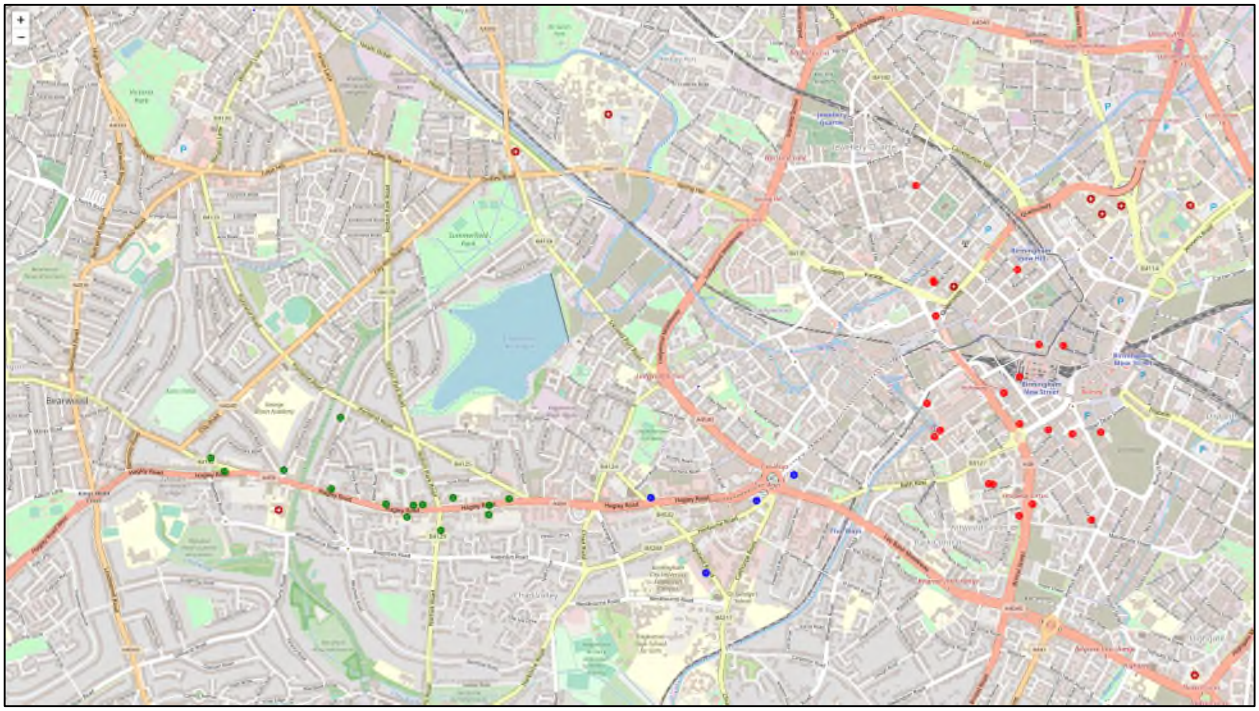

Figure 3. Red-green-blue coloring scheme used to distinguish hotels of different clusters.

Centers of Clusters

The second output of the K-Means clustering algorithm was the location of the center in each cluster. To visualize the centers, another map [Figure 4] was created with markers representing each center plotted on it.

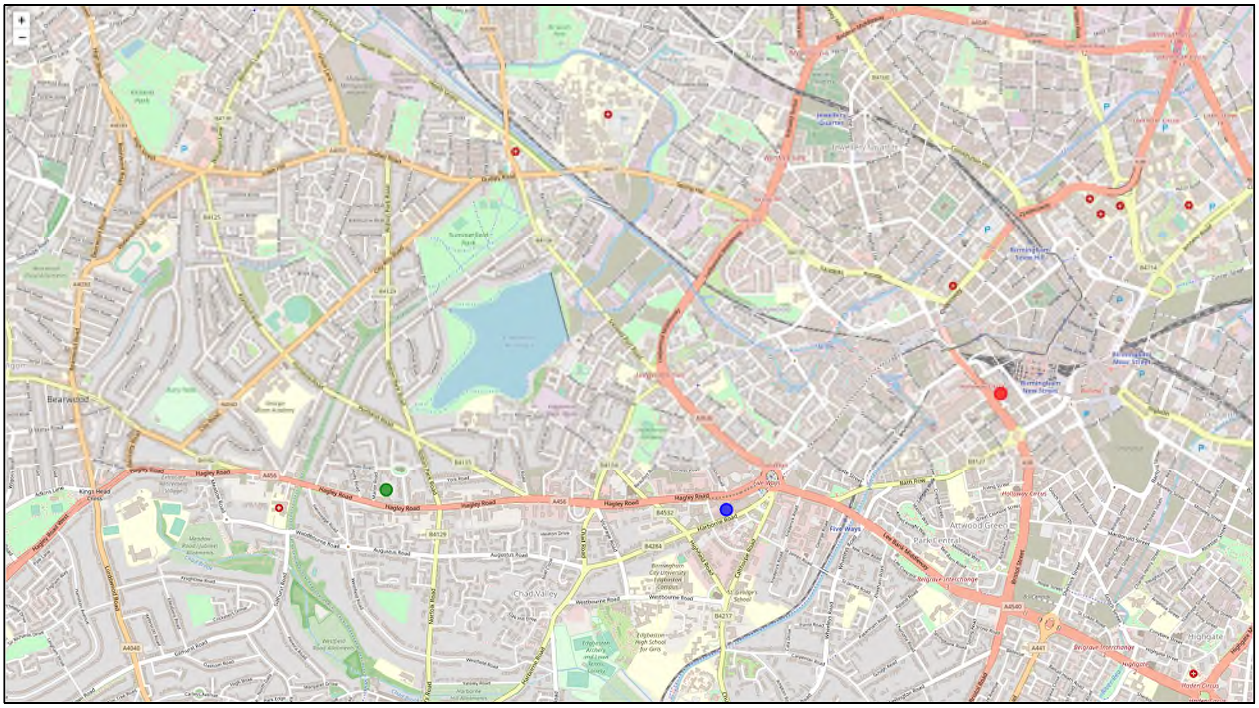

Figure 4. Centers of Hotel clusters map. (Red-green-blue coloring scheme consistent with Figure 3)

Restaurants Classification

The KNN classification algorithm output was also visualized on a map [Figure 5] to show each restaurant in a different color based on the class to which it was assigned.

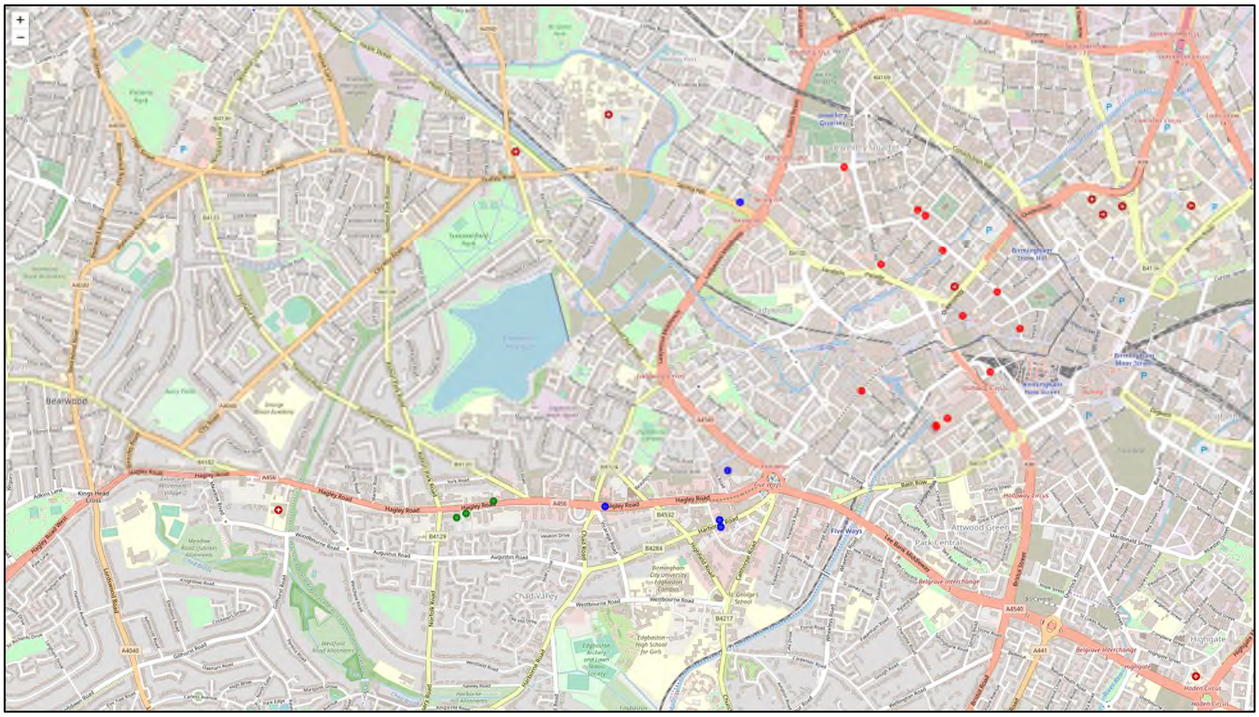

Figure 5. Restaurants classified by corresponding Hotel cluster map. (Red-green-blue coloring scheme consistent with Figure 3)

Trending Restaurant Category

Restaurant categories were inspected by counting the frequency of each category to get an idea of the most popular cuisine type and thus, the restaurants receiving the highest demand, since the data frame used for this analysis already represents venues with the highest level of foot-traffic. After counting the frequency of categories, the top five are chosen to visualize and analyze as this will guide the recommendation for the type of restaurant to open. Figure 6 displays the top five categories plotted on a Bar Chart against the number of restaurants of that category.

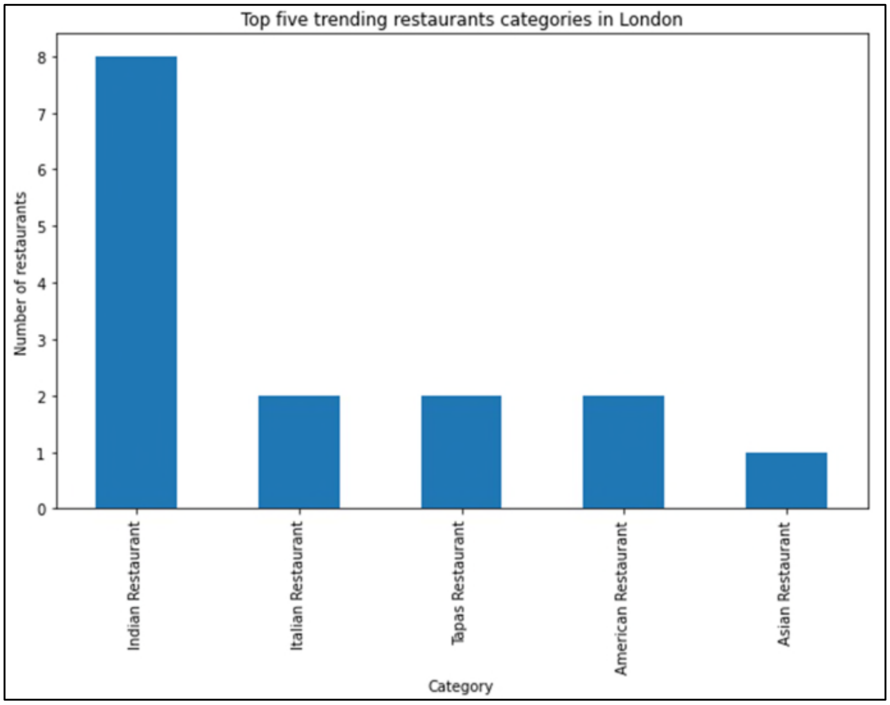

Figure 6. Top 5 trending restaurants in the Birmingham City Center area.

It is clearly evident from Figure 6 that Indian restaurants are the most popular category in Birmingham city center, with a total number of 8 restaurants vs 2 of the next popular category. The distribution of Indian restaurants are mapped in Figure 7 to aid the visualization of their scatter against the 3 hotel clusters identified.

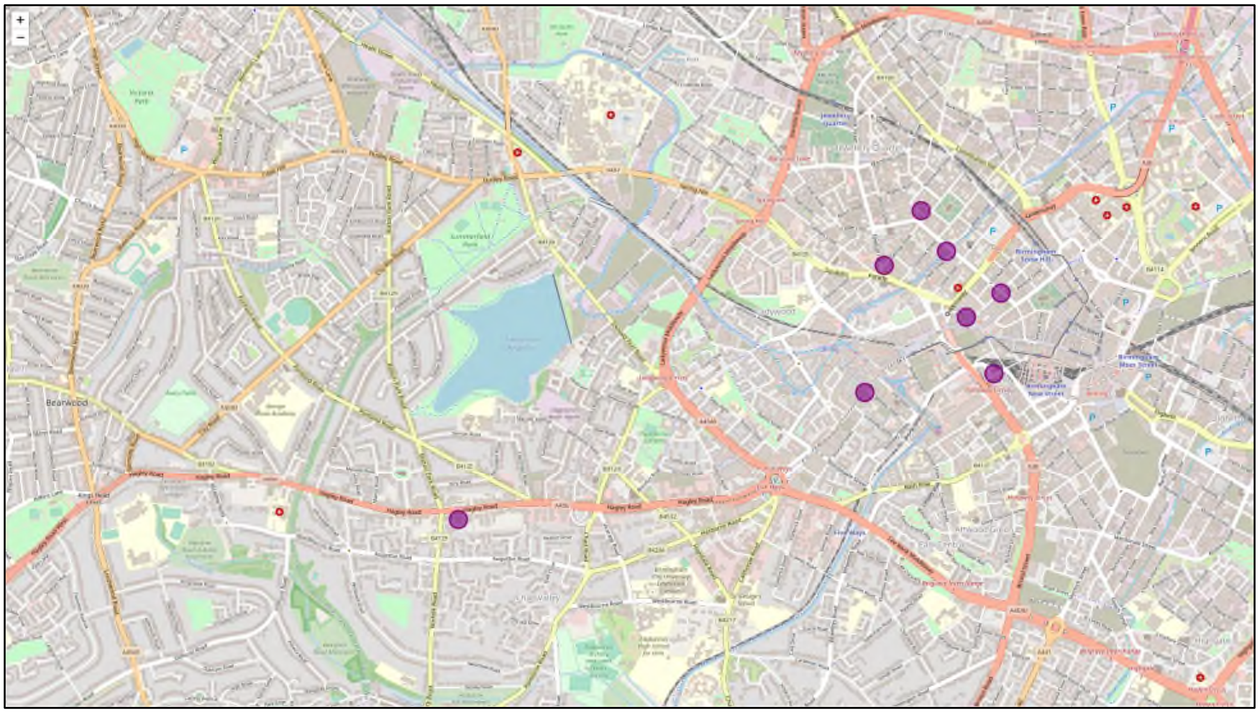

Figure 7. Distribution of Indian Restaurants in the Birmingham City Center area.

The results presented in this section aim to aid the decision-making process for the proposed business and is detailed in the next section.

Discussion

The analysis performed in this report intends to recommend which area in Birmingham City Center has the most potential to open a restaurant and what type of cuisine would be most popular with visitors.

After classifying the trending restaurants into the clusters generated by K-Means algorithm, and counting the numbers in each (Figures 4 & 5), Cluster 1 (green) has the fewest number of restaurants, although similar to the number of restaurants found in Cluster 2 (blue). This tells us that both areas have lower competition than Cluster 0 (Red).

When comparing the number of hotels between Cluster 1 and Cluster 2 (Figure 3), it is clear that more hotels are present in the area covered by Cluster 1. This observation couple with the fact that Cluster 1 has the fewest number of restaurants, raises the attractiveness of a potential opportunity for investors in this area.

It should be noted also that while Cluster 1 may be more attractive due to the higher hotel count and lower number of existing restaurants, among these there is an Indian restaurant already existing in the area. Cluster 2 has no Indian restaurants and therefore could be an alternative investment opportunity as the data suggests this is clearly the most popular cuisine type in Birmingham.

Conclusion

The aim of this study was to find the optimal location for a restaurant in Birmingham city center for visitors, and the favoured cuisine type in the city. To achieve this, data from Foursquare API about hotels and restaurants in Birmingham city center was acquired.

Hotel data was analyzed using K-Means Clustering algorithm to group the hotels in three clusters and find the center of each cluster. This was done in order to find the areas with a high density of hotels and thus considered to have a higher proportion of visitors to the city.

Restaurants' data was used as a target dataset for KNN Classification algorithm in order to see in which clusters the trending restaurants are located. This helped us find that Cluster 1 had the lowest level of competition as it had fewer restaurants compared to the other two areas.

By counting the number of hotels in Cluster 1, it was found to have the second highest number of hotels and so is considered as an area with an attractive opportunity to open a new restaurant.

Finally, by assessing the data for most popular cuisine type, Indian cuisine is found to be the most popular in restaurants. Interestingly there are no Indian restaurants in Cluster 2, which could be an alternative investment opportunity in the area.