OPENING A RESTAURANT TARGETING TOURISTS IN LONDON

Qingan Wang,

September 2020

Table of Contents

Introduction

In this section, a description of the business problem, background, and target audience that will benefit from this project is provided in order to answer what the report is trying to solve.

Background

London is considered a touristic destination that people from all around the world visit annually. One of the major interests that tourists have when they visit a city for the first time is its people's eating habits, in other words, where one can find good restaurants, and what food is most popular. And since tourists will be looking for popular food places, then it is an interest for investors as well to meet this demand.

Problem

In order to find out which food type is popular in London, and where it is most needed, – assuming that the investor is targeting tourists – then data is needed to locate hotels in London, the level of competition among restaurants around these hotels, and the trending food type. This report will tell where it is most profitable to open a restaurant that tourists can visit, and what food type it should focus on.

Interest

This can be very helpful for investors looking to start a business in London and trying to follow the trend there, as the report provides insights on the hospitality and food industries around the city center. Visitors willing to travel to London would will also find it helpful.

Data

Data source and quality is the heart of any data science problem as it has a major effect on the final answers. Here the data used in this study is explained.

Data Source

Our problem requires location data that describes hotels and restaurants around London city center, specifically, location of hotels and restaurants, number of restaurants around hotels, foot traffic in restaurants and their category.

To retrieve this data, the Foursquare API was utilized, as it provides valuable location data that powers some of the world famous applications. Therefore, two Major datasets were retrieved from Foursquare, Hotels and Restaurants.

Data Cleaning

For the hotels dataset, a "**Search"** call was made to the Foursquare API with the search query "**Hotel"**. The aim was to get the hotels around London city center. However, after the call was made, results needed to be organized into a Pandas data frame.

Further cleaning was made to keep only data attributes related to the name and location of each venue as well as extracting the venues' categories to clearly verify the results.

After results were organized and viewed in a Pandas data frame, it turned out that some results had categories other than "Hotel". Therefore, these results were dropped from the data frame as our focus is on hotels only.

Second, the restaurants dataset was needed but this time the trending venues are the ones of interest, so am **“Explore”** call was made to the API to get a glimpse on which venues were most trending.

As in the case of hotel, results were organized in a data frame and cleaned from any unnecessary features other than venues’ names and location related data. Then, venues’ categories were extracted and only restaurants were kept and other venues were dropped from the data frame.

Since “Explore” call generates results that are trending at the moment when the call was made, two calls were made in different times of the day, thus different meal times, and the results were very close.

Feature Selection

For the intended analysis to be made, that is 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 the other is for restaurants, each holding only the latitude and longitude coordinates of the original datasets as these are the features that some Machine Learning algorithms will focus on in our analysis as will be explained in the next section.

Methodology

To clearly describe the methods of data analysis used in this project, a detailed explanation of our analysis will be presented here, showing the Machine Learning algorithms used, and how they contributed to the results as well as our statistical analysis of the opportunity of starting a restaurant in London city center.

Finding Touristic Areas

As a first step to finding the best spot for a restaurant targeting tourists near the city center of London, it is required to know which areas have high density in hotels and therefore are considered to be touristic areas 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 a touristic area where a 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 point 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, as mentioned earlier, and will serve 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 London using Folium library and tried to inspect the data visually first.

Although other techniques are available to find the optimum number of clusters, such as, Elbow Method, but due to the relatively small number of data point visual inspection [Figure 1] was enough where it was proposed that Three cluster would be suitable to group the data points so that number was used in the clustering algorithm.

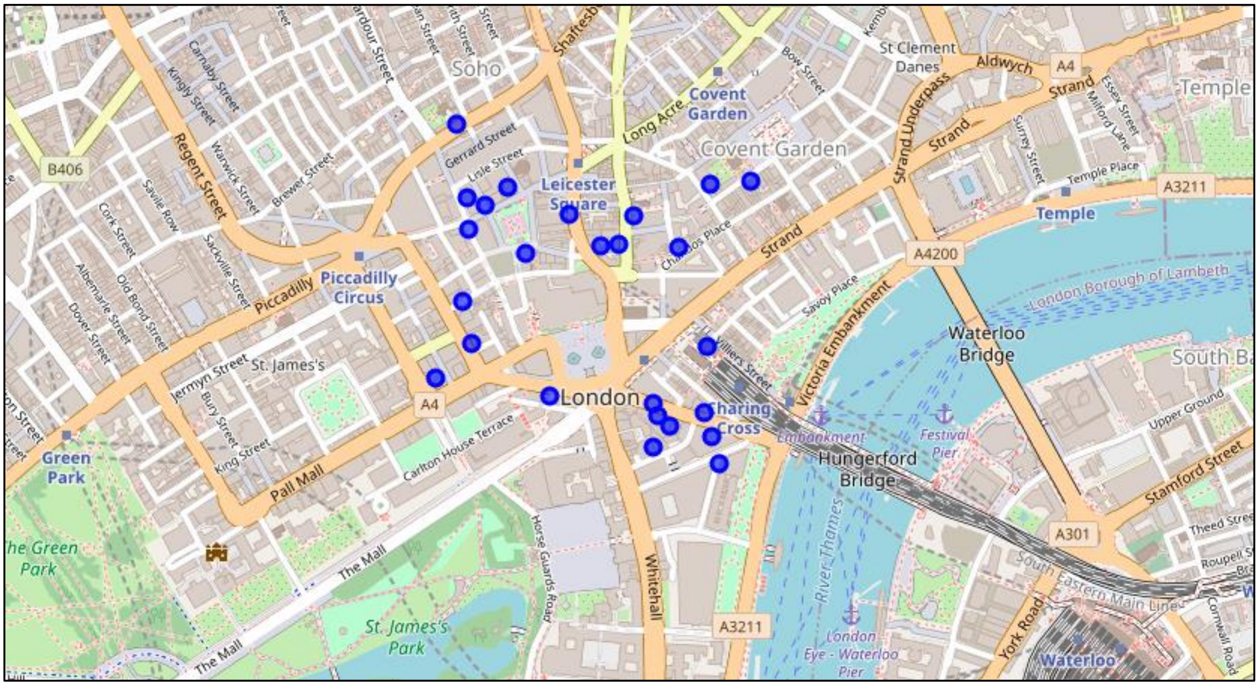

Figure 1: Hotels Around London City Center Map

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 two outputs were visualized on maps having the hotels in each cluster given a different color on the first map and the centers marked in larger circles on the other as will be seen in the Results section.

Locating The Best Spot

Once the touristic 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 data, but first a visualization of the restaurants locations is generated [Figure 2] to try to predict the KNN results, and it was observed by comparison with the hotels locations, that at least one hotels' cluster did not have any restaurants.

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.

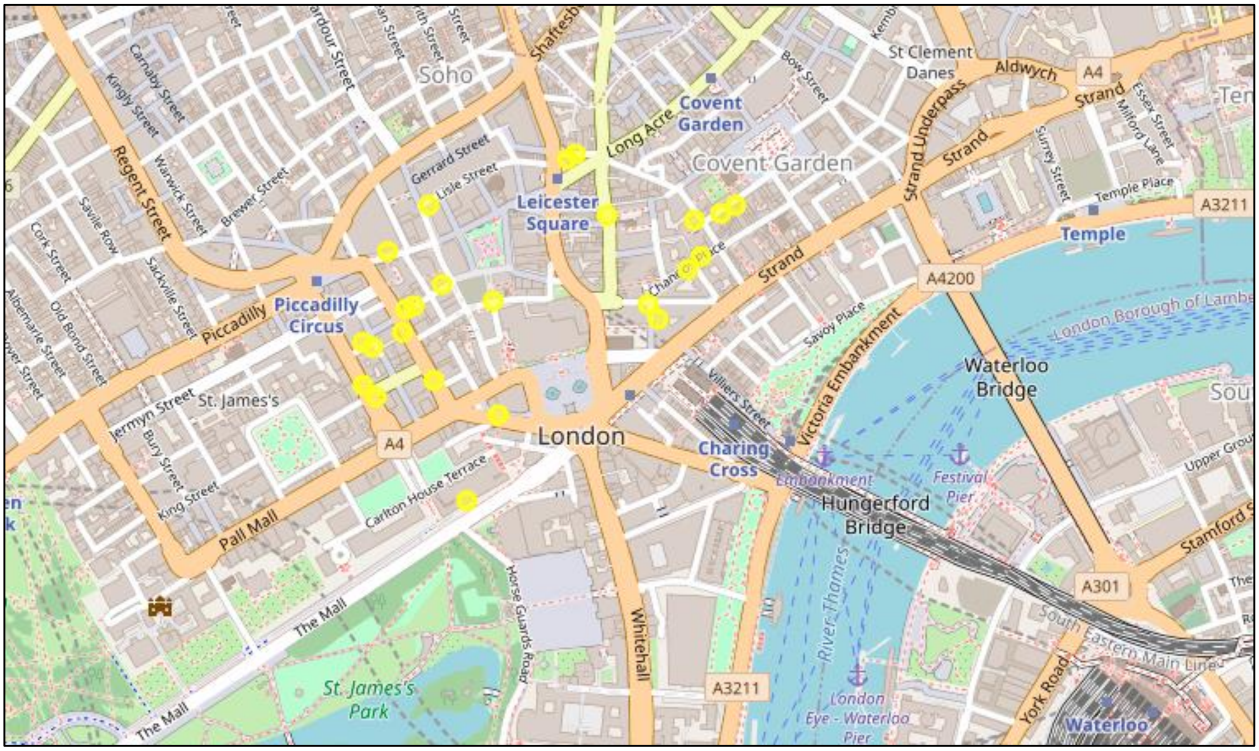

Figure 2: Restaurants Around London City Center Map

Therefore, we will use the clusters' centers' location data as a training set to tell the KNN model what location each class is center around, and we will 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 with 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; it is now that an evaluation of opportunities is possible.

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.

Results

Six main results of the previously explained methodology were obtained from our analysis and will be presented in this section in a simple and visualized manner.

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].

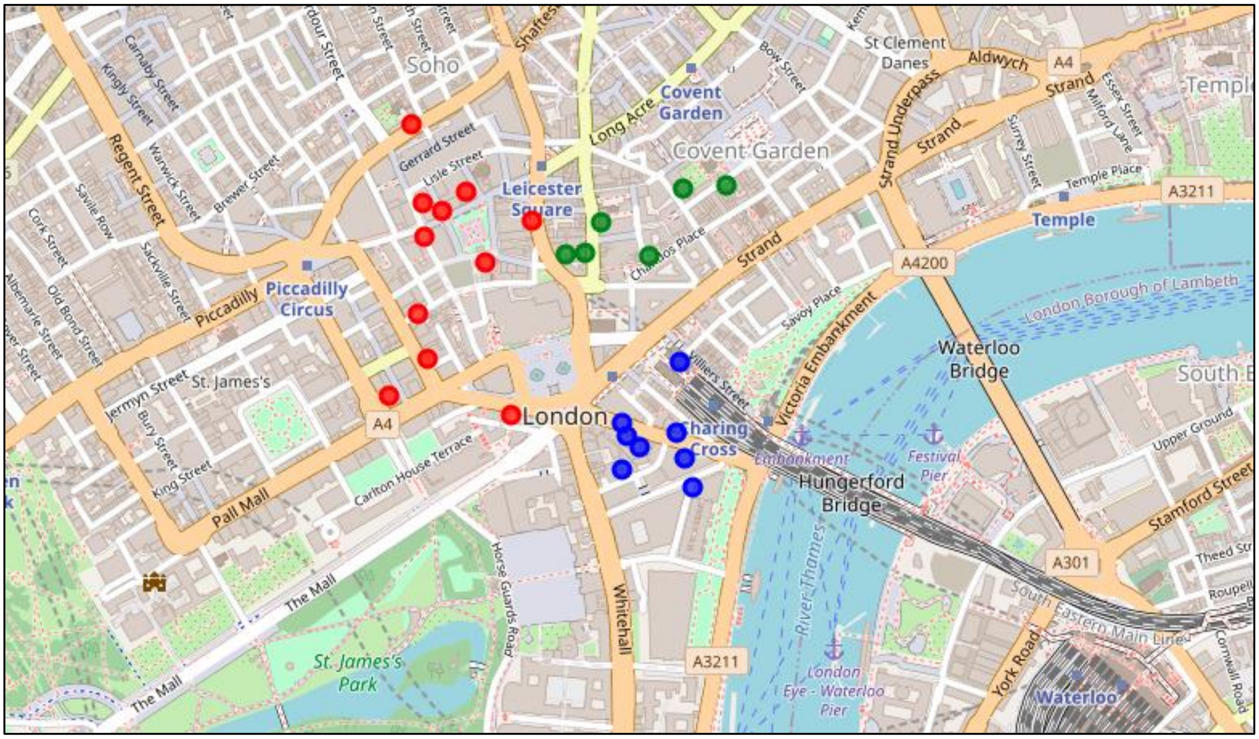

Figure 3: Hotels Clusters Map

Red-green-blue coloring scheme was 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. These centers' coordinates are presented in the following table [Table 1].

Table 1: Location Coordinates of Clusters' Centers

Cluster Label	Latitude	Longitude
0/Red	51.509792	-0.130766
1/Green	51.510464	-0.125776
2/Blue	51.506920	-0.125233

To visualize the above listed centers, another map [Figure 4] was created with markers representing each center plotted on it.

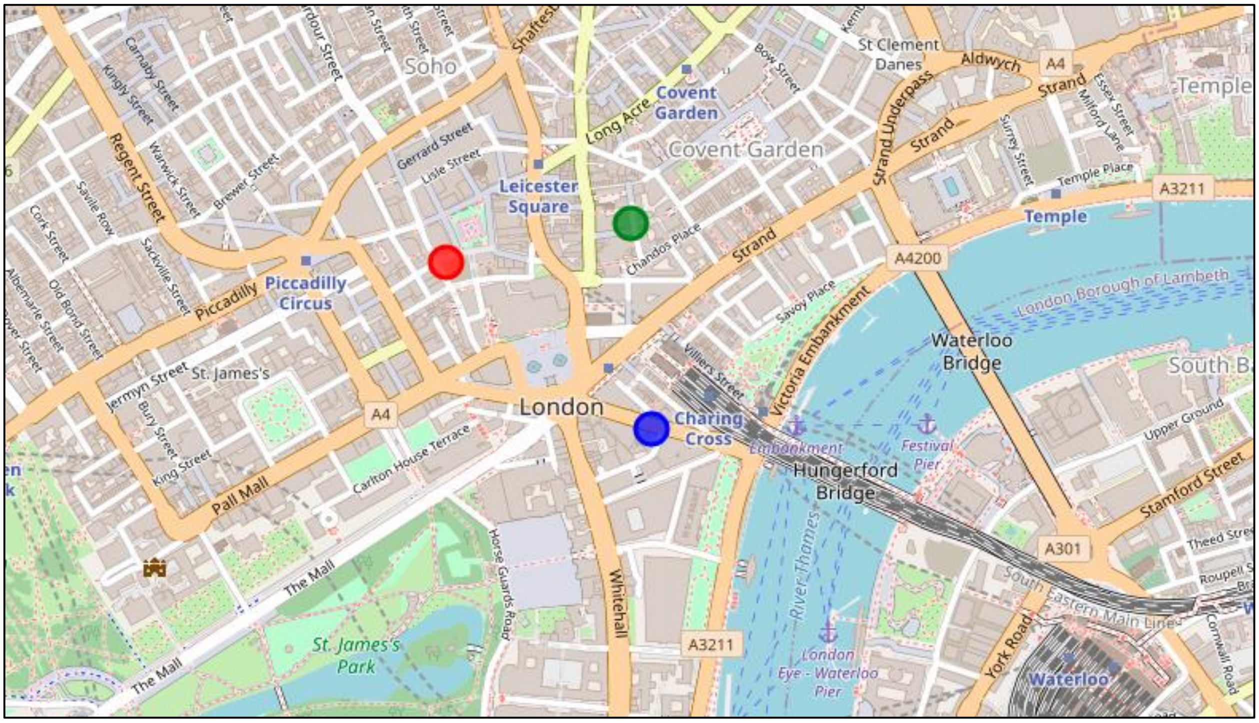

Figure 4: Centers of Clusters Map

The same Red-green-blue coloring scheme was used to differentiate the center of each cluster

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.

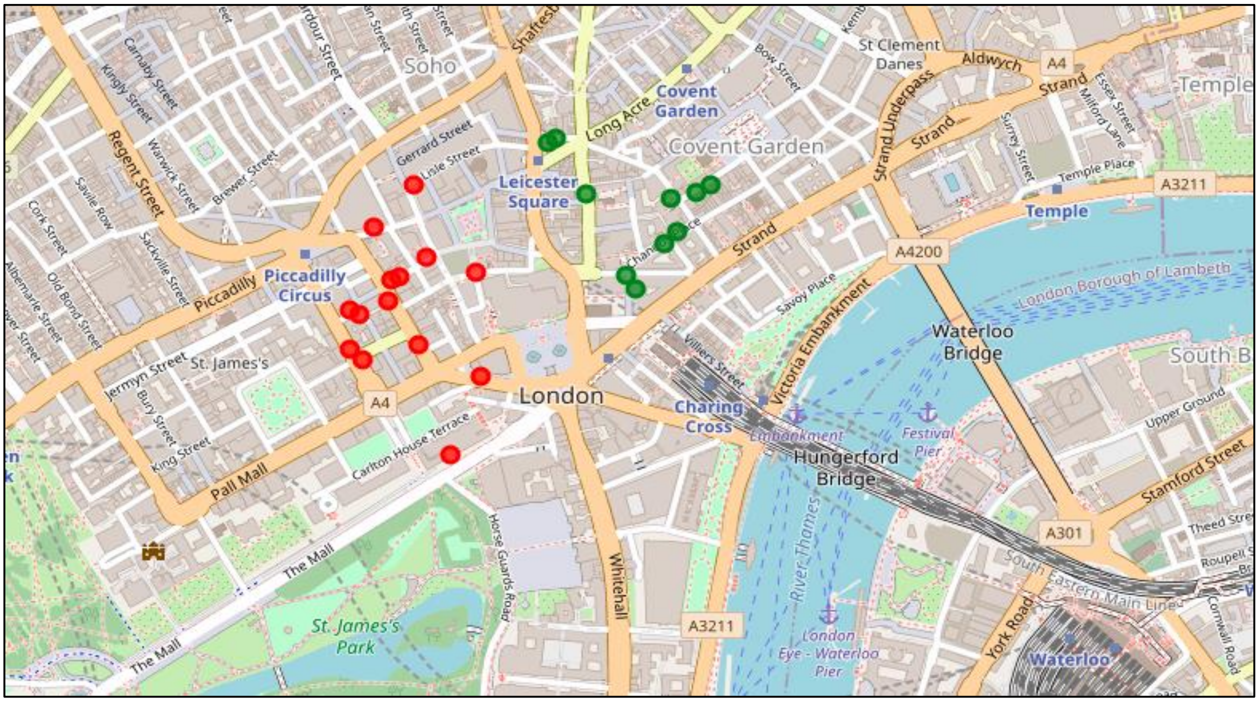

Figure 5: Restaurants Classes Map

The same Red-green-blue coloring scheme used for hotels clusters was used here to match classes/clusters easily.

Restaurants Counting

For the purpose of predicting completion levels, we counted the number of restaurants available in each map and the numbers are shown in the next table [Table 2].

Table 2: Restaurants in Each Class

Class Label	Number of Restaurants
0/red	14
1/green	10
2/blue	0

This evaluation will tell which cluster of hotels has the lowest number of restaurants and thus, a minimum level of completion.

Hotels Counting

Since the area with the minimum number of restaurants was found, it is convenient to verify that this area has a number of sufficient number of hotels to consider it a worthy opportunity. Therefore, we counted the number of hotels in each area [Table 3].

Table 3: Hotels in Each Cluster

Class Label	Number of Restaurants
0/red	11
1/green	6
2/blue	8

These numbers will tell whether or not the nominated area demonstrates a good opportunity.

Trending Restaurant Category

Restaurants' categories were inspected and we counted the frequency of each category to get an idea of the most occurring categories 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 as explained before.

After counting the categories' frequency, we chose the top five restaurants' categories to visualize and analyze as this will guide our selection of the type of food our restaurant will supply.

In the figure below [Figure 6] we visualize the top five categories in a Bar Chart against the number of restaurants of that category.

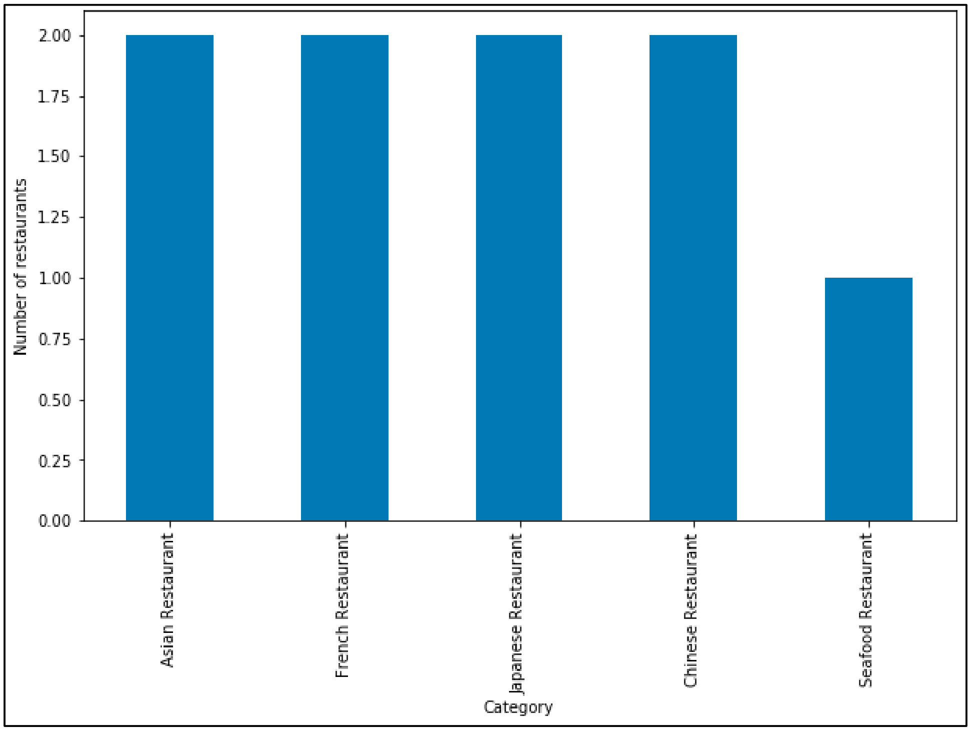

Figure 6: Top Five Restaurant Categories

The findings presented in this section act as the heart of our analysis and decision making for the proposed business and will be further discussed in the next section.

Discussion

The analysis performed in this report was intended to tell us where to open the proposed restaurant and what type of food it should supply.

After classifying the trending restaurants into the clusters generated by K-Means algorithm, and counting the numbers in each, we found that Cluster 2 (shown in blue) did not have any restaurants, which confirms our prediction when visualizing the restaurants London's map. This tells us that this area has the lowest level of competition.

Then we evaluated the opportunity in this area by counting the number of hotels there, as our restaurant is intended to target tourists, and we found that this cluster comes second in the number of hotel with a total of eight hotels, thus considered to be an attractive opportunity for investors.

In our analysis, we generated and visualized the center of each cluster as the point closest to all hotels in that cluster together. So in the case of our proposed restaurant, the center of Cluster 2 located at (51.506920, -0.125233) would be the optimum spot, or as close to it as possible.

Finally, we tried to learn what food category was in demand the most in London city center by looking at the top five frequent categories, and we found that four of them had an equal frequency of two restaurants. However, by looking at these categories we find that among them are "Asian Restaurant", "Japanese Restaurant", and "Chinese Restaurant". Since the latter two are also considered Asian food, we can safely say that Asian food is the category in demand the most.

Conclusion

In this report our aim was to find the best location for a restaurant in London city center targeting tourists, and the food type it should provide. To find that we retrieved data from Foursquare API about hotels and restaurants in the desired location.

Hotels' 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 as touristic areas.

Restaurants' data on the other hand 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 2 had the lowest level of competition as it had no restaurants at all.

By counting the number of hotels in Cluster 2, it was found to be the second cluster in the number of hotel (eight hotels) so it was considered as an attractive opportunity.

Finally, we looked at the food type in demand the most by counting the frequency each restaurant category in our data occurred, and we found that Asian food was in demand and restaurants with Asian food menus had the highest levels of foot-traffic.