FINDING THE BEST SPOT TO START A RESTAURANT IN NYC

Aushin Raj

June 2021

Contents

1 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.

1.1 Background

New York City is one of the most popular tourist destinations in the world. NYC is the mecca of business in the United States and as a melting pot of American culture, there is something for every style, taste and budget in New York City. One of the common 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. Since tourists will be looking for popular food places, then it is an interest for investors as well to meet this demand.

1.2 Problem

The objective of this capstone project is to analyse and select the best location and trending food type in the city of New York to open a new restaurant, targeting the tourists. Using the data science methodology and instruments such as Data Analysis, Visualization and Machine Learning, this project aims to provide solutions to answer the business question: Where in New York should investors open a restaurant, targeting the tourists and what food type should they serve?

1.3 Interest

This Project is particularly useful to investors looking to open or invest in a restaurant in the city of New York.

2 Data

Data is the heart of any data science problem as it has a major effect on the final answers

2.1 Data Source

Our problem requires location data that describes hotels and restaurants around New York City. Specifically, location of hotels and restaurants, number of restaurants around hotels, foot traffic in restaurants and their category. For retrieving this data Foursquare API was utilized. Foursquare API search feature would be used to get the location of hotels and restaurants. Due to Foursquare API call limitations, the number of hotels in search query is set to 50 and radius parameter to 500.

2.2 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 New York City. The results got from the call were first converted into pandas data frame. There were a lot of unwanted rows and columns in the data frame. The columns other than the name and location of hotels were dropped and the rows containing only hotel as the value for the column name is taken.

For the restaurant dataset, a "explore" call was made to the API to get a glimpse on which venues were most trending. The results got from the call were first converted to a pandas data frame. There were a lot of unwanted columns and rows in the data frame. The columns other than the venue name and venue location were dropped and the rows containing restaurant as the value for the column venue name is taken

2.3 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.

3 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 New York City.

3.1 Finding Touristic Areas

As a first step to finding the best spot for a restaurant targeting tourists near the city of New York, 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(500m).

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.

3.1.1 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 hotel's 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 New York 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.

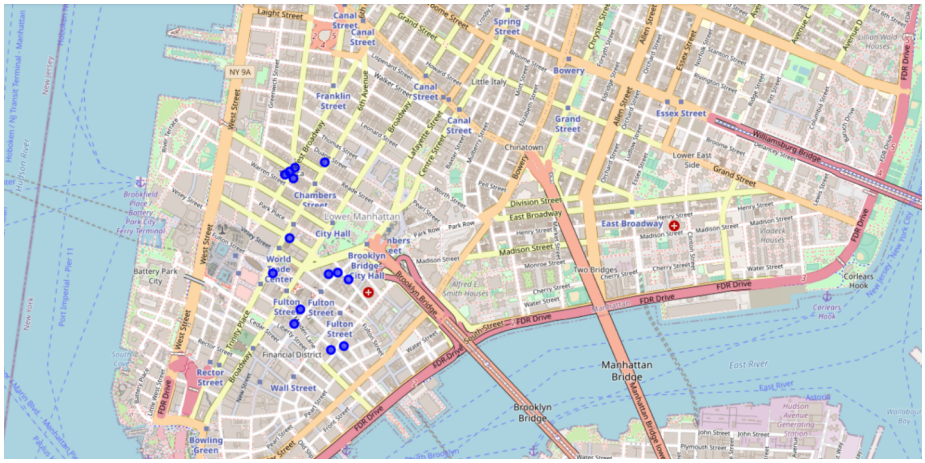

Figure 1: Hotels Around New York City Center Map

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 centre's location coordinates (latitude and longitude).

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

3.2 Locating The Best Spot

Once the touristic areas were defined, the next step is to analyse 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.

3.2.1 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.

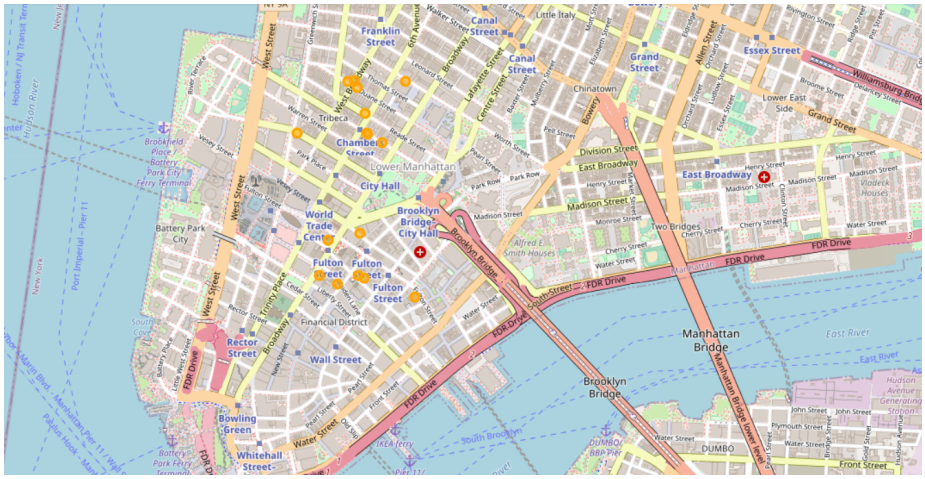

Figure 2: Restaurants Around New York City Center Map

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 colour to each class of restaurants similar to the hotels' output visualization (K-Means output).

3.2.2 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.

3.3 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.

4 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.

4.1 Hotels Clustering

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

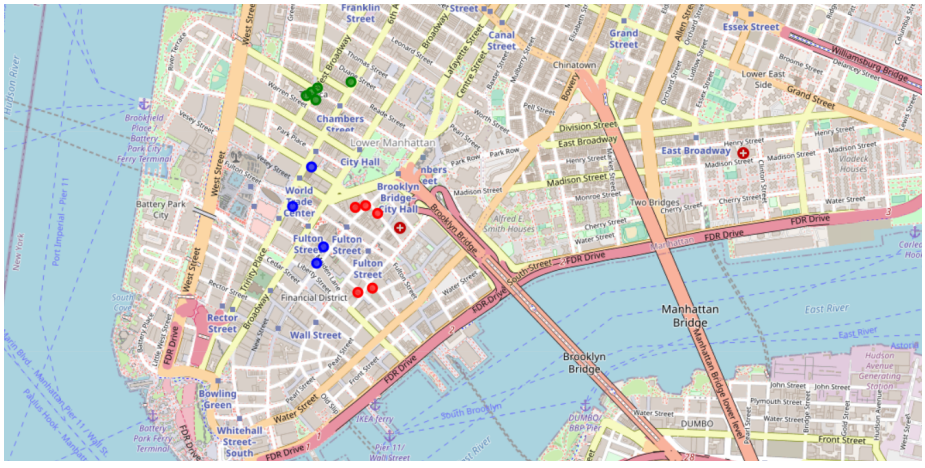

Figure 3: Hotels Clusters Map

Red-green-blue colouring scheme was used to distinguish hotels of different clusters.

4.2 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	40.70980691	-74.00669725
1/Green	40.71548148	-74.00895632
2/Blue	40.71057204	-74.00940412

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

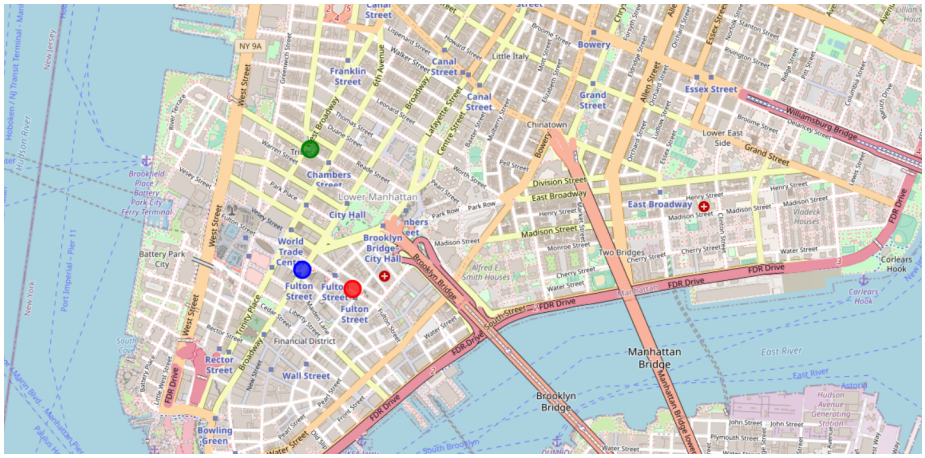

Figure 4: Centers of Clusters Map

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

4.3 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.

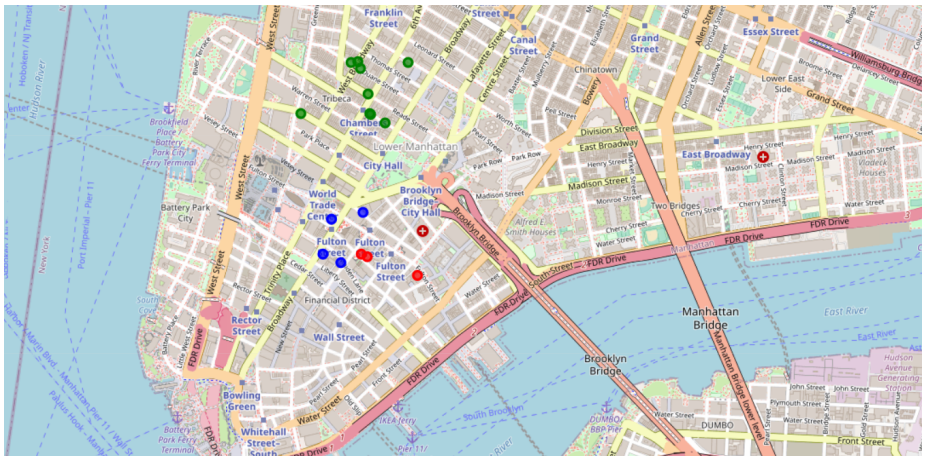

Figure 5: Restaurants Classes Map

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

4.4 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	3
1/green	9
2/blue	4

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

4.5 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	5
1/green	5
2/blue	4

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

4.6 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 analyse 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.

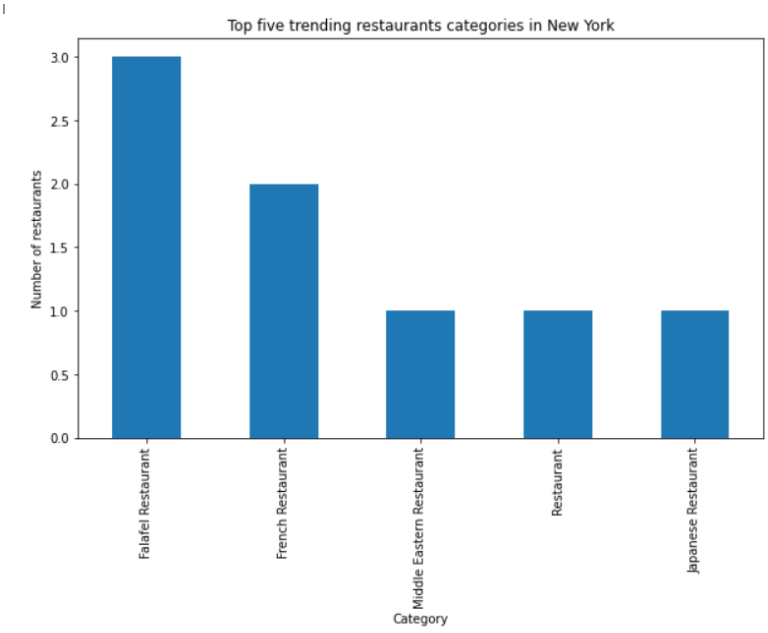

Figure 6: Top Five Restaurant Categories

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.

5 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 0 (shown in red) have minimum number of restaurants, which confirms our prediction when visualizing the restaurants New York'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 first in the number of hotel with a total of 5 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 0 located at (40.70980691 -74.00669725) 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 New York city by looking at the top five frequent categories, and we found that falafel was the food type in demand.

6 Conclusion

In this report our aim was to find the best location for a restaurant in New York city 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 analysed 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 0 had the lowest level of competition.

By counting the number of hotels in Cluster 0 , it was found to be the first cluster in the number of hotel (five 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 Falafel food was in demand and restaurants with Arabic food menus had the highest levels of foot-traffic.