Capstone Project
The Battle of Neighborhoods

Christodoulos Christodoulou

May 17, 2020

Contents

1   Introduction

In this capstone project the concept of opening a new Italian restaurant in New York City is explored. restaurant entrepreneur is looking for a recommended area that would be more suitable for this project. Italian food is very popular worldwide and the idea is to open the restaurant at a neighborhood where not many other similar options are available. So, it may be a very good opportunity to attract new market segments from different cultures too. Choosing a proper location for a new restaurant is very important since it is considered to be a key for a successful and beneficial business.

The business Problem that this project is trying to solve, is to find the most suitable location for the opening of a new Italian restaurant in New York City. The objective of the proper selection is to enhance the profitable ability of the restaurant by maximizing the possible attracted customers. This project targets but not limits to aspire restaurant owners that looking to open new places in the New York City and makes a proposal of how a selection of the desired opening location can be done.

2   Data

This problem would be solved using Data Science and machine learning clustering method.

The required data in order to achieve the objectives are firstly the total neighborhoods of New York City, among with their corresponding lateral and longitudinal coordinates. In addition, venue data of existed Italian restaurants in the city. So, based on the venue data the neighborhoods of the city will be classified in clusters using the correspodning method. The data of the New York City are got form the NYU Spatial Data Repository, while the corresponding coordinates are extracted from the Geocoder imported package. Finally, the venue data for each neighborhood of the city are got through the Foursquare API.

3   Methodology

Firstly, the required data including names and coordinates of all the neighborhoods of New York city were collected. Then using folium package a map of the city presenting the locations of all neighborhoods was plotted as can be seen in the figure below.
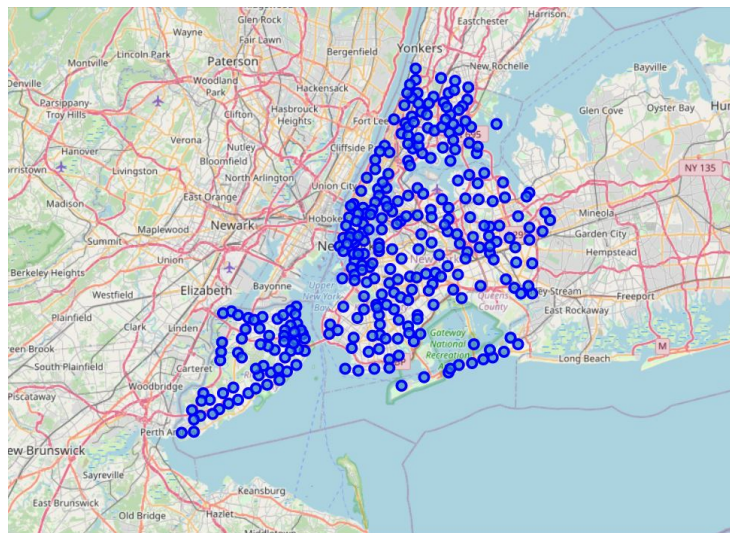

Figure 1: Representation of all neighborhoods of New York

Then using the Foursquare API an exploration was executed to every neighborhood in order to extract information regarding the local venues. Each neighborhood was analyzed and one-hot data frame was created showing the all the venues to the corresponding neighborhood. Then a selection was done to create a data frame that included information related to Italian restaurants only. Finally, a grouped data frame with respect to the neighborhoods and the number of Italian restaurants was created as can be seen in the figure below.

	Neighborhood	Italian Restaurant
0	Allerton	30
1	Annadale	15
2	Arden Heights	5
3	Arlington	6
4	Arrochar	20

Figure 2: Number of Italian restaurants per neighborhood

Since, the the preparing of the data was completed, next step was to decide which machine learning algorithm to use in order to classify the neighborhoods based on the density of Italian restaurants. It was decided to select the K-means algorithm because it is simple and able to provide a desirable result for the clustering problem.

4   Result

In this section the results that arise from the clustering of the neighborhoods of New York based on the density of Italian restaurants are presented. In the figure below it can be seen a map (created using folium) with the neighborhoods marked to the corresponding cluster.
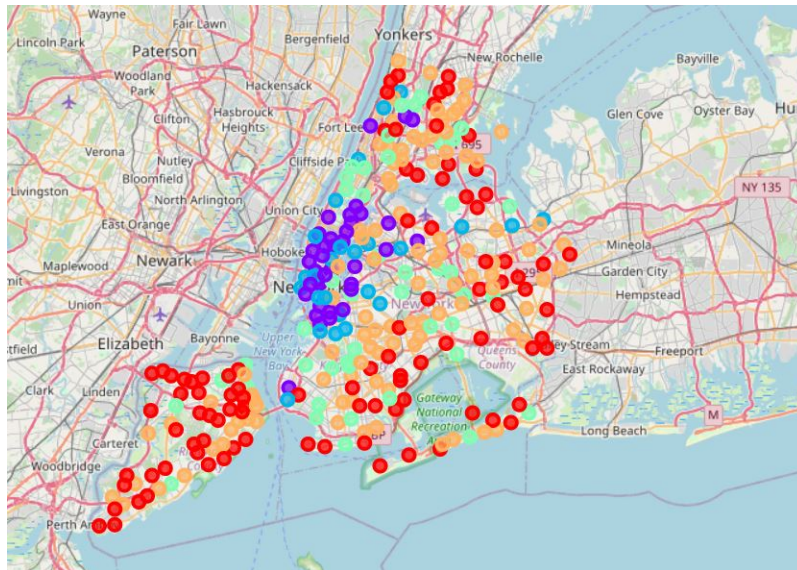

Figure 3: Clusters

It can be seen that in total 5 clusters were created. The details of each one are shown in the table below.

	Cluster	Italian Restaurants
0	Red	0-14
1	Orange	15-30
2	Light Blue	31-52
3	Light Green	53-80
4	Purple	80-130

Figure 4: Clusters Details

5 Discussion

It can be seen that the selected machine learning algorithm indeed classified the neighborhoods of New York based on the density of the Italian restaurants. As already expected the neighborhoods that are in a greater distance from the crowded center of the city (Manhattan) have less options for such restaurants.

However, choosing a neighborhood to open a new restaurant where less than 10 similar options are available seems to be a good choice but one should take into account if the chosen neighborhood is closed to other venues too. From this prospective the most recommended neighborhood to open a new restaurant is Astoria Heights, which is very close to the center and the other available options are just 13.

6 Conclusion

In this report, it was presented the procedure of trying to find a recommended neighborhood to open a new Italian restaurant in New York City using a machine learning algorithm. The recommendation was made based on the division of the several neighborhoods to clusters with respect to the the presence of other similar options in the area. It was highlighted that another important aspect was to check other venues in the area too. Finally, based on these a recommended neighborhood was provided.

It is important to mention that there are possible future developments around this research. For example average rating of the Italian restaurants of each neighborhood could be included in the data that will be use for classification using machine learning. In such case, it is possible to provide a recommendation in a more crowded neighborhood in terms of Italian restaurants but where the ratings are not very satisfying.