IBM Data Science Capstone Project – The Battle of Neighborhoods

Opening an Italian restaurant in Brooklyn, New York

(Finding optimal neighborhood in Brooklyn for new restaurant)

By: Zdravko Radulovic

December 2019

Introduction

Italian restaurants are one of the most popular restaurants in USA. The Italian immigrants who traveled to America in late 19th, early 20th century bring their culinary culture onto US soil. During World War I Italian immigration paused, and resumed afterward resulting in the establishment of rustic Italian restaurants that catered mainly to immigrants[1]. According to the National Restaurant Association, Italian-American food and Mediterranean cuisine has been highly influential in United States, and it is one of the top tree cuisines.[2] Taking this in consideration, opening an Italian restaurant would be great idea. Narrowing this project with Brooklyn borough, and analyzing each neighborhood using Foursquare as a source of data we found that there are 28 Italian restaurants dispersed in neighborhoods of Brooklyn. Finding the right neighborhoods for potential location of a new Italian restaurant would be a key factor that will determine whether the restaurant will be successful investment opportunity or not.

Business problem

The objective of this project is to analyze neighborhoods in Brooklyn and propose possible location for opening new Italian restaurant. Using data science techniques, we will try to answer the business question: *What are the best neighborhoods for potential opening a new Italian restaurant?*

Target Audience

Choosing location is key to success for almost any business, especially restaurant business. So this project is useful for potential Italian restaurant chains or restaurant investors that are looking to open or invest in new restaurant in neighborhood that lacks this type of restaurants.

Data

For this project we need following data:

1. List of all neighborhoods in Brooklyn
2. Venue data, related to Italian restaurants located in Brooklyn
3. Geospatial data – latitude and longitude of each neighborhood

[1] Simone Cinotto, 2013 "*The Italian American Table*, *Soft Soil, Black Grapes*, and *Making Italian America*"

[2] Italian-American cuisine, Wikipedia.org

Sources, description of data and methods of extraction

List of all neighborhoods of Brooklyn define scope of this project. We will use data set “2014 New York City Neighborhood Names” (Department of City Planning) from following source: https://geo.nyu.edu/catalog/nyu_2451_34572 which contains list of neighborhoods and boroughs of the New York city.

We will filter this data set to get only neighborhoods of Brooklyn with coordinates of each neighborhood.

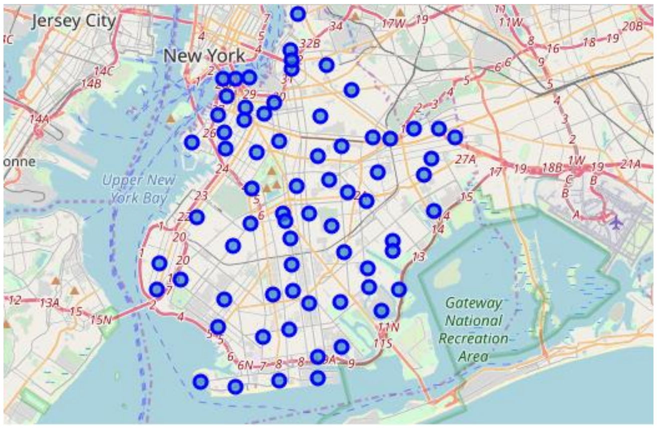
Venue data, related to Italian restaurants located in Brooklyn we can find using Foursquare API, as one of the largest data provider. We will search for venue data related to Italian restaurants in neighborhoods of Brooklyn.

In this project we will use different data science skills and tools: Pandas library for data extraction, cleaning, wrangling; Foursquare API for data source; machine learning algorithm for neighborhood clustering, K-means; visualization library Folium, for neighborhood visualization on interactive map.

In Methodology section we will describe all steps taken in this project, data analysis that we performed and machine learning technique that was used.

Methodology

First, we need to download list of neighborhoods in Brooklyn. We have available data set “2014 New York City Neighborhood Names” that contain Brooklyn neighborhoods from following source: https://geo.nyu.edu/catalog/nyu_2451_34572 which contains list of neighborhoods and boroughs of the New York city. This data set contains longitude and latitude for all neighborhoods. Using Pandas library, we will filter out only neighborhoods of Brooklyn. New data frame contains 70 neighborhoods. Using Folium library, we can visualize on map all neighborhoods in Brooklyn (picture 1).



Brooklyn neighborhoods (Picture 1)

Next, we will use data from Foursquare API to get the top 100 venues in radius of 500 meters. Foursquare will return data in the JSON format, and we will extract the venue name, category, latitude and longitude. Than we will check how many unique categories can be recognized from all returned venues. We will analyze each neighborhood, groping them and taking the mean of the frequency of occurrence of each venue category. As we previously mention, we want to find potential neighborhood for new Italian restaurant, so we will filter only Italian restaurant venues, as potential "competition", and prepare data set for clustering. We will perform clustering using k-means clustering algorithm which identifies k number of centroids, and then allocates every data point to the nearest cluster. We decided to cluster neighborhoods into 3 clusters based on their frequency of occurrence for Italian restaurant, which will allow us to identify neighborhoods with low or no Italian restaurants, and vice versa. This will help us to find the answer on question: What are the best neighborhoods for potential opening a new Italian restaurant?

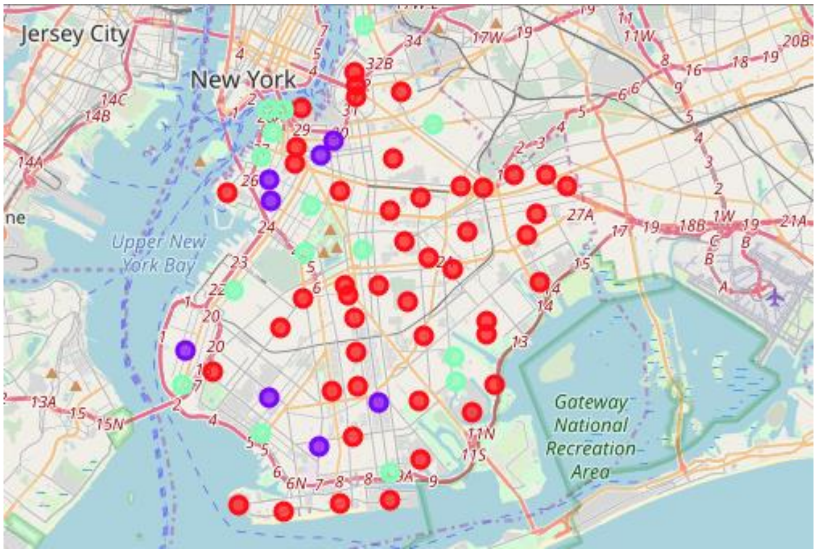
Results

The result from the k-means clustering based on frequency of occurrence for Italian restaurant shows following:

1. **Cluster 0:** neighborhoods with no or low number of Italian restaurants.
2. **Cluster 1:** neighborhoods with moderate number of Italian restaurants.
3. **Cluster 2:** neighborhoods with high concentration of Italian restaurants.

The results are visualized on map with different color (picture 2):

1. **Cluster 0: Red color**
2. **Cluster 1: Violet color**
3. **Cluster 2: Green color**



Clustered Brooklyn neighborhoods (Picture 2)

Discussion

Observing map and results of clustering, most of Italian restaurants are concentrated in Cluster 2. In Cluster 1 there is moderate dispersion of Italian restaurants, so opening an Italian restaurant in these two clusters of neighborhoods will be challenging because they are likely to have strong competition. Cluster 0 contains low concentration of Italian restaurants, and in most of neighborhoods in this cluster there is no this type of restaurants, so that makes this cluster perfect for opening new restaurant. However, further analysis must be taken, because in this project we only consider frequency of occurrence Italian restaurants in neighborhoods. Further research need to observe larger scope, not only frequency of occurrence, like population and income analysis, other venues etc.

Conclusion

In this project we start the process with identifying the business problem, asking a question: What are the best neighborhoods for potential opening a new Italian restaurant? To get answer on this question, we have gone through different phases of project: business understanding, data requirement, data collecting and preparation, performing machine learning algorithm for clustering the data into 3 clusters based on factor of occurrence. Using smaller scope for this project, and providing recommendations for further analysis, we can provide the answer: *The neighborhoods in Cluster 0 are the most preferred locations for new Italian restaurant.* These findings could provide insights for further research that will help potential investors to make relevant decision for opening new restaurant.

References

Department of City Planning, “2014 New York City Neighborhood Names”, https://geo.nyu.edu/catalog/nyu_2451_34572

Italian-American cuisine, http://www.wikipedia.org

Foursquare Developers Documentation, https://developer.foursquare.com/docs

Simone Cinotto, 2013 “*The Italian American Table*, *Soft Soil, Black Grapes*, and *Making Italian America*”