CAPSTONE PROJECT

Opening Italian Cuisine Business in Brooklyn

OCTOBER 26, 2020

AFRIONI ROMA RIO

1. Introduction

We have a request from a stakeholder to analyze a potential location in Brooklyn for an Italian cuisine business. Italian cuisines like pizza, pasta, and other cuisines are very famous worldwide and might have potential in a crowded place like Brooklyn, New York City (2.6 million residents). We try to get the result by using unsupervised machine learning k-means and obtain the best position.

2. Business Problem

This project aims to satisfy the stakeholder's demand, which is to find the best location for an Italian cuisine business in Brooklyn. This project will explain how to obtain the result by using data science methodology, such as unsupervised machine learning k-means.

3. Data Description

The raw data for analyzing is from this link:

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json.

By using a little bit of data preprocessing, we will obtain data set for New York location (borough and neighborhood). After that, we will select borough = Brooklyn for further analysis (see table 1).

Table 1. Example of data set for Brooklyn

Borough	Neighborhood	Latitude	Longitude
Brooklyn	Bay Ridge	40.625801	-74.030621
Brooklyn	Bensonhurst	40.611009	-73.99518
Brooklyn	Sunset Park	40.645103	-74.010316
Brooklyn	Greenpoint	40.730201	-73.954241
Brooklyn	Gravesend	40.59526	-73.973471

To obtain all the venues in Brooklyn, we use foursquare (table 2). Then, by using Jupyter Notebook we can obtain the best venues for each neighborhood in Brooklyn. Then, we cluster the cleaned data using a package from scikit-learn

Table 2. Example of venues obtained from foursquare

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.6242	-74.030931	Pizza Place
Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar

4. Methodology

In this project, we use *k*-means clustering to obtain the best location for Italian Cuisines. By definition *k*-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. *k*-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using *k*-medians and *k*-medoids [1].

First, we do data preprocessing. Then we search for the best venues in each neighborhood. After we transform the data, it will look like in the figure 1

```
In [23]: Brooklyn_grouped = Brooklyn_onehot.groupby('Neighborhood').mean().reset_index()
         Brooklyn_grouped
```

Out[23]:

	Neighborhood	Yoga Studio	Accessories Store	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auto Garage	
0	Bath Beach	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.021739	0.000000	0.00	0.(
1	Bay Ridge	0.012346	0.000000	0.037037	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.(
2	Bedford Stuyvesant	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.(
3	Bensonhurst	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.027778	0.000000	0.00	0.(
4	Bergen Beach	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.142857	0.00	0.(
5	Boerum Hill	0.022222	0.000000	0.011111	0.011111	0.00	0.000000	0.000000	0.033333	0.000000	0.000000	0.011111	0.00	0.(
6	Borough Park	0.000000	0.000000	0.045455	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.(
7	Brighton Beach	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.(
8	Broadway Junction	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.(

Figure 1. Cleaned Data for Modelling

To run the *k*-means clustering model we use a library from scikit-learn, and run the code like this.

```
# set number of clusters
kclusters = 3

Brooklyn_grouped_clustering = Brooklyn_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Brooklyn_grouped_clustering)
```

5. Results

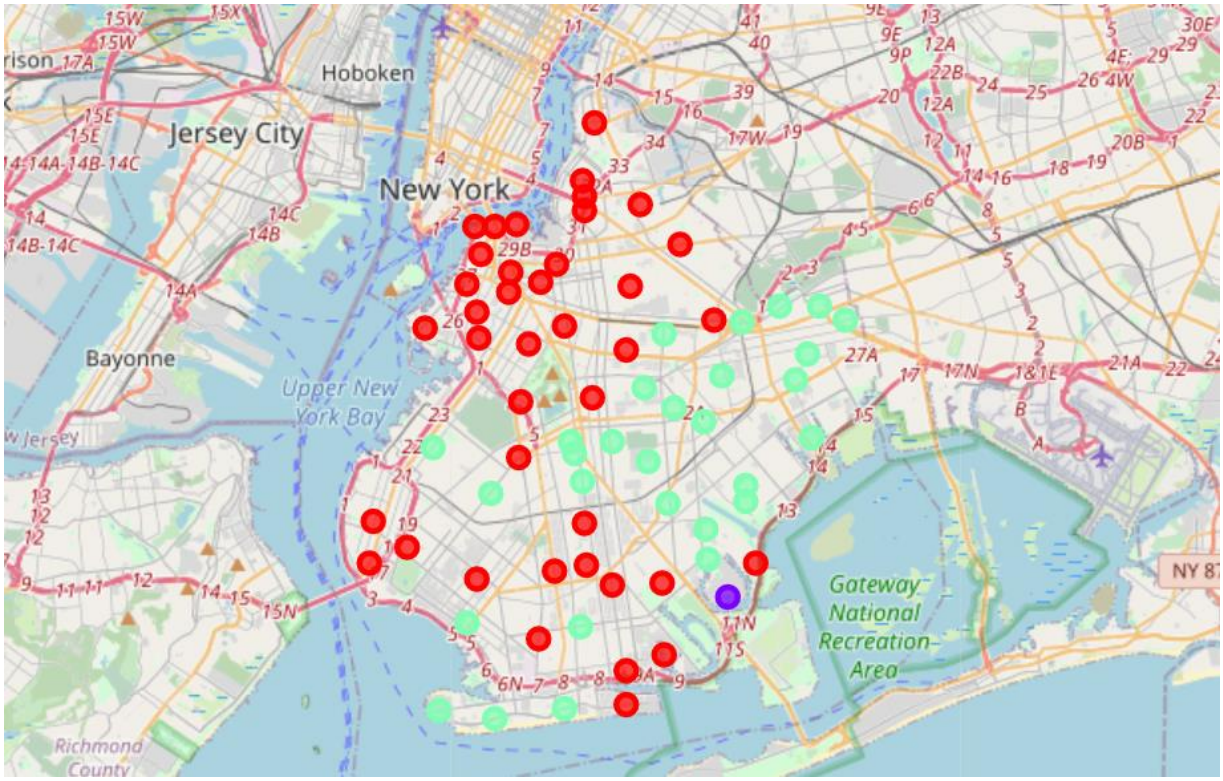

Figure 2. Brooklyn Cluster

After we run the model, we obtained the result like in figure 2. Where red, purple, and green circles represent cluster 1, cluster 2, and cluster 3 respectively.

6. Discussion

We try to examine the data for each cluster. In this case, we only take into consideration for 1st and 2nd most common venues.

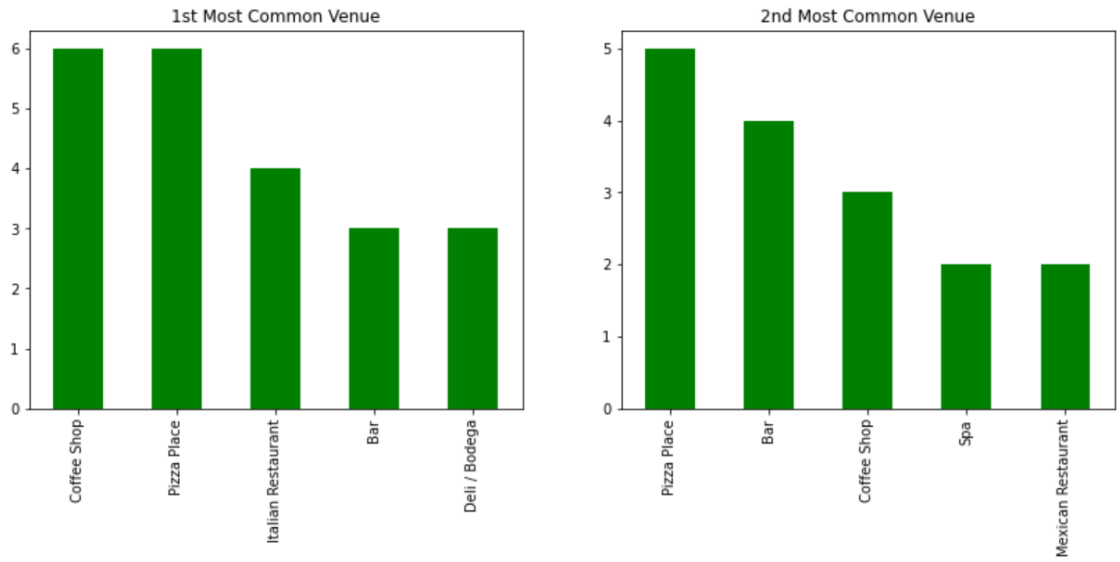

Figure 3. Cluster 1 Most Common Venues

In cluster 1, the frequency of Italian cuisines business is a lot (figure 3). In the 1st most common venue and 2nd most common venue, pizza place or Italian Restaurant are the dominant venues. Which make this cluster is the best location for opening a business in Italian cuisine

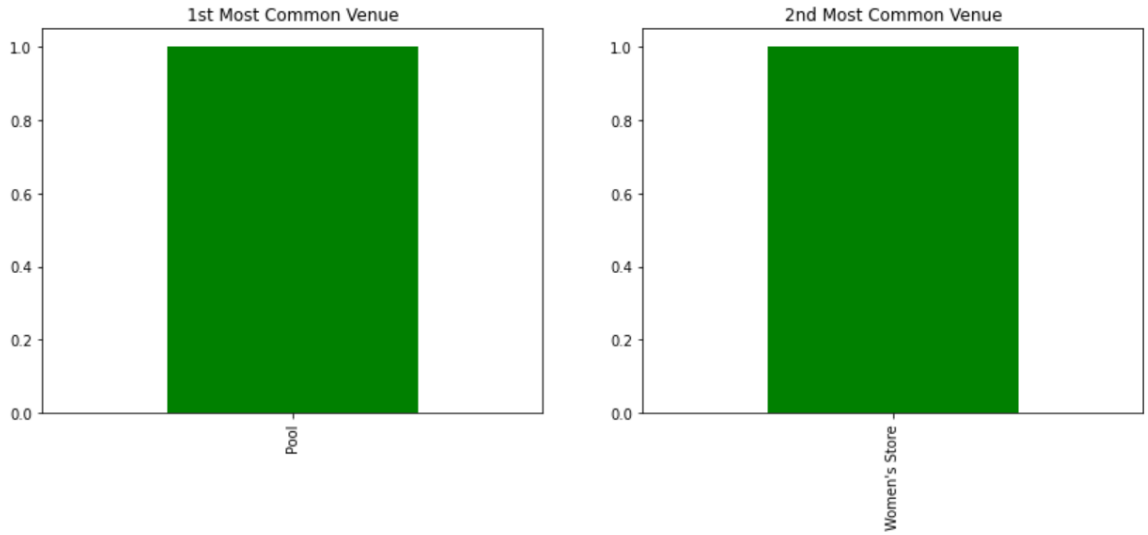

Figure 4. Cluster 2 Most Common Venues

Cluster 2 is likely an anomaly of data, where the most common venues is different from the other cluster and, only has 1 location.

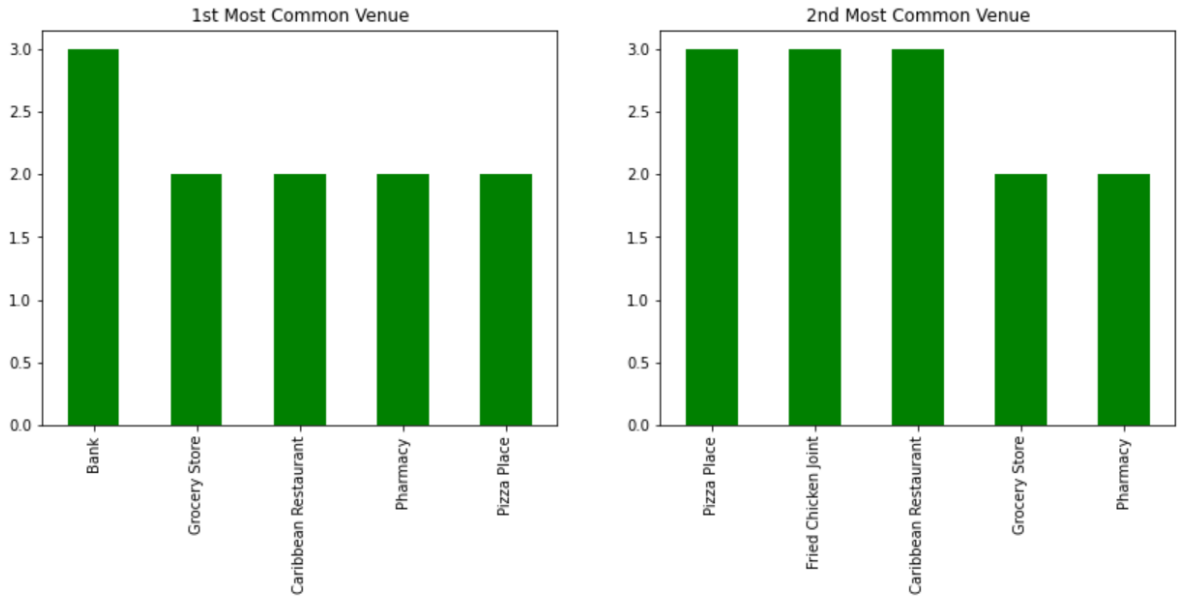

Figure 5. Cluster 3 Most Common Venues

In cluster 3, the most common venues are banks, grocery stores, Caribbean restaurants, pharmacy. For Italian cuisine compared to cluster 1 is still low (3 compared to 6)

7. Conclusion

In conclusion, the best location to start a business in Italian cuisine is on cluster 1. It will look like this on the map

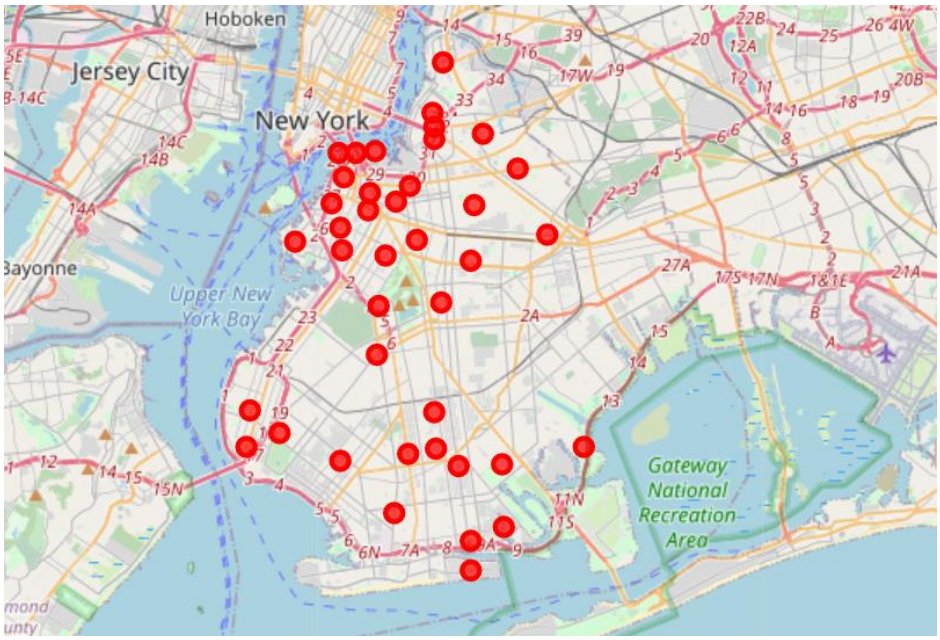

Figure 6. Best location for opening business in Italian cuisine

8. Reference

[1] https://en.wikipedia.org/wiki/K-means_clustering