Capstone Project - The Battle of Neighborhoods

Manhattan Grocery Store Location Suggestion

Project Report

1. Business Problem:

New York City (NYC) is the most populous city and most densely populated major city in the United States. A global power city, New York City has been described uniquely as the financial capital of the world and exerts a significant impact upon commerce. Needless to say, the city and especially its borough Manhattan provides immense opportunities for entrepreneurs and businessmen.

NYC's food culture includes an array of international cuisines influenced by the city's immigrant history. This is evident from the fact that the city has a humongous number of restaurants, bars, cafes, joints et al. in its neighborhoods which creates a huge demand for raw materials to be supplied to these food outlets with quality and in a timely manner. In other words, a large scale Grocery store which would cater mainly in storing and supplying items required for restaurants' inventory. One of the challenge for any business to flourish is to carefully select the neighborhoods where it wants to target its customers.

Here finding the right Manhattan's neighborhoods for opening a Grocery store is the mission and we will try to achieve it by using dataset containing NYC Boroughs and Neighborhood features coupled with data science techniques, which would eventually help potential clients to make informed decision.

2. Data Needs

- **Data title:**
  Coordinates of each NYC Neighborhood

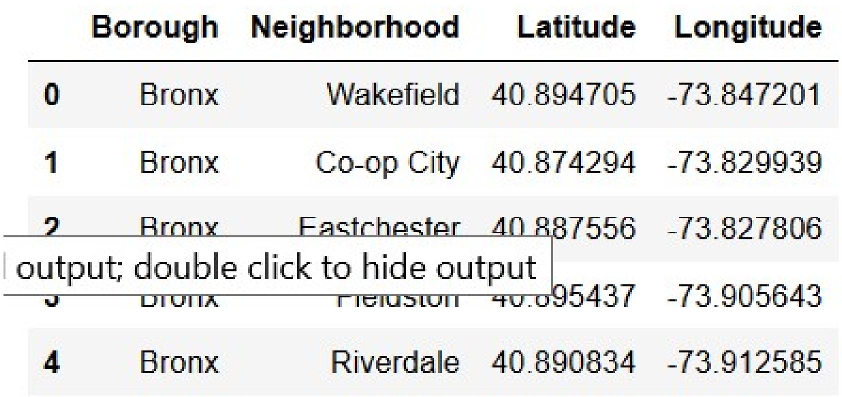
  Type of data:
  JSON

  Description of the data:
  A dataset that contains the New York City’s 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. This dataset exists for free on the web.

  Source:

https://geo.nyu.edu/catalog/nyu_2451_34572

Data Example: The latitude and longitude of all the Neighborhoods of Borough Bronx in New York City can be retried from this dataset:



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

output; double click to hide output

- **Data title:**
  Foursquare location data

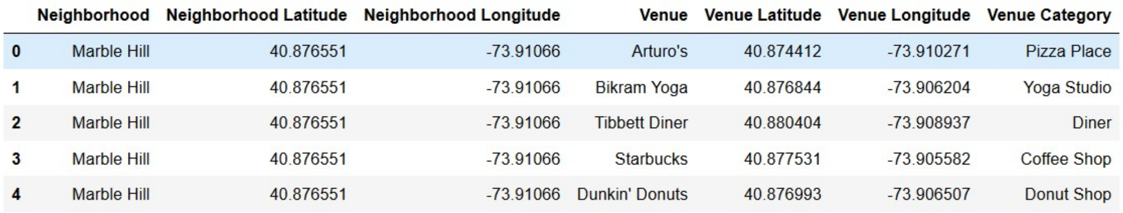
  Type of data:
  JSON

  Description of the data:
  Location coordinates obtained by Foursquare API calls.

  To determine the proximity of various amenities as per the client's requirement, Foursquare location data is used.

  Source:
  https://foursquare.com/

  Data Example: Using Foursquare API calls we extract features of the nearby venues, viz. venue names, venue categories and venue coordinates (latitude, longitude):



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin' Donuts	40.876993	-73.906507	Donut Shop

3. Methodology

The automated script developed as a part of Manhattan Grocery Store Location Suggestion project does the following:

1. Download all the dependencies that will be needed mainly pandas, json, geocoders, requests, sklearn and folium.

2. Download the dataset containing New York Boroughs and Neighborhoods.

3. Process the downloaded data and transform the New York features data of nested Python dictionaries into a Pandas DataFrame containing all the Venues in the Borough Manhattan.

4. Utilize Foursquare API to explore the Manhattan Neighborhoods. This will return only the relevant information for each venue viz. neighborhood, name, latitude, longitude and venue category.

5. Analyze each Manhattan Neighborhood using one-hot encoding. One-hot encoding is applied to the Manhattan venues, so that the categorical variables, in this case venue categories, are converted into a form that could be provided to Machine Learning technique to further explore.

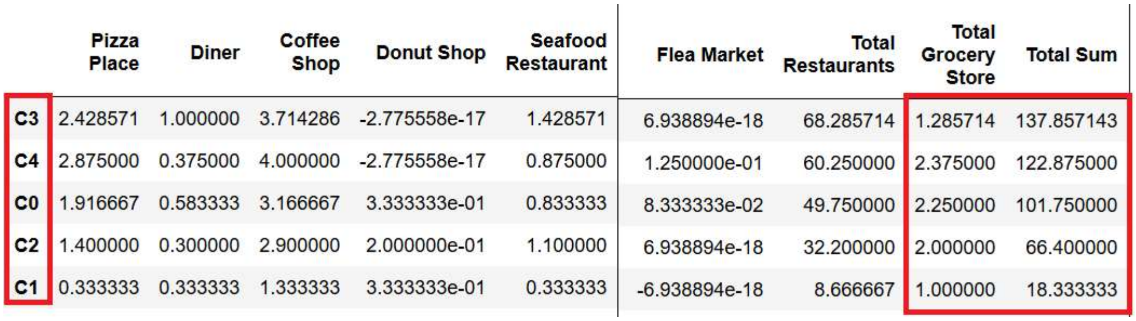
6. Create the restaurant category, grocery store category and needed features list for further integration, so that we can find the total restaurants and grocery stores. Also, to concentrate only on the required venue categories.

7. Apply one of the Machine Learning techniques (K-Means Clustering). As a result, Manhattan neighborhoods will be categorized into 5 Clusters. Now using the sum of all the features within each cluster, we will sort the clusters which will tell us the demand for a new grocery store considering the population of already existing grocery stores.

4. Results

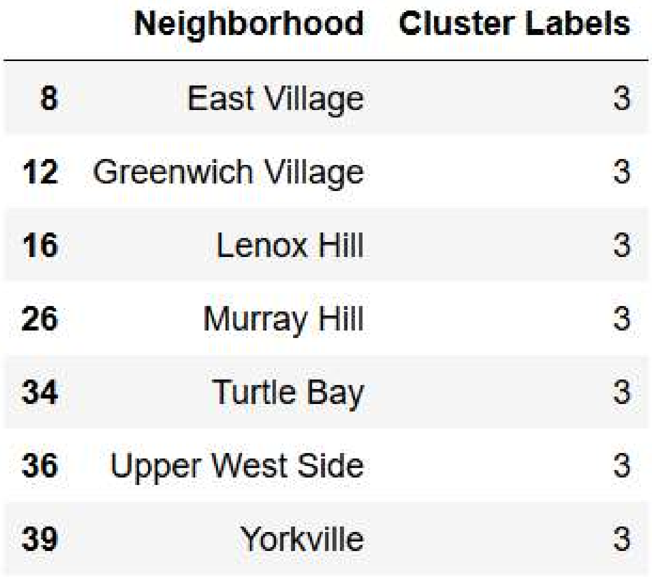
Focusing on centers of the clusters and comparing them for "Total Restaurants" against "Total Grocery Store" features, the result shows the Cluster whose center has the highest "Total Sum" and least “Total Grocery Store” will be suggested Cluster of neighborhoods, where a new Grocery Store will have the highest chances to flourish.

Below its seen that Cluster 3 is the top one suggested:



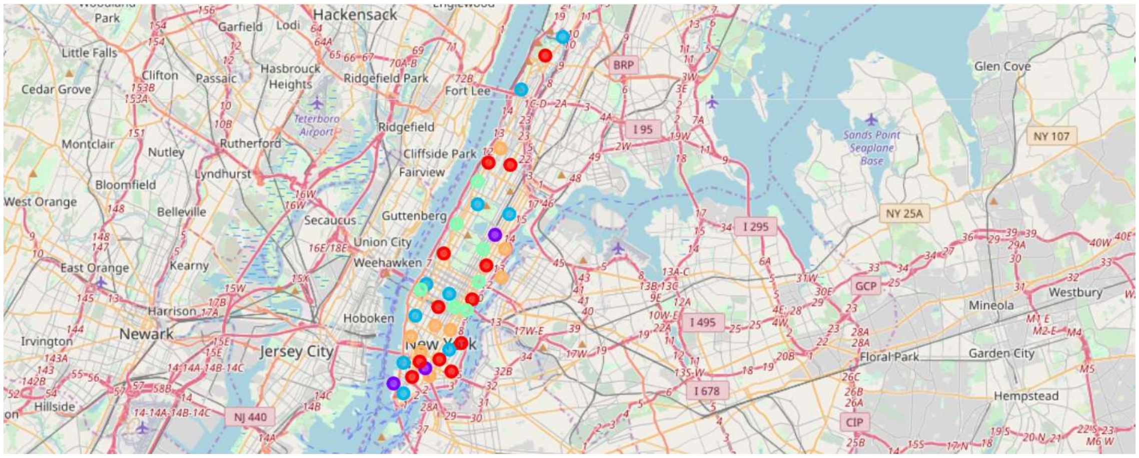
	Pizza Place	Diner	Coffee Shop	Donut Shop	Seafood Restaurant	Flea Market	Total Restaurants	Total Grocery Store	Total Sum
C3	2.428571	1.000000	3.714286	-2.775558e-17	1.428571	6.938894e-18	68.285714	1.285714	137.857143
C4	2.875000	0.375000	4.000000	-2.775558e-17	0.875000	1.250000e-01	60.250000	2.375000	122.875000
C0	1.916667	0.583333	3.166667	3.333333e-01	0.833333	8.333333e-02	49.750000	2.250000	101.750000
C2	1.400000	0.300000	2.900000	2.000000e-01	1.100000	6.938894e-18	32.200000	2.000000	66.400000
C1	0.333333	0.333333	1.333333	3.333333e-01	0.333333	-6.938894e-18	8.666667	1.000000	18.333333

Below is the list of Neighborhoods from Cluster 3:



	Neighborhood	Cluster Labels
8	East Village	3
12	Greenwich Village	3
16	Lenox Hill	3
26	Murray Hill	3
34	Turtle Bay	3
36	Upper West Side	3
39	Yorkville	3

Clusters on the Manhattan map. The markdowns colored in **Red**, **Purple** and **Blue** are the Top 3 neighborhood clusters where a new grocery store is suggested to be opened.



5. Discussion

Based on the findings in the result section, users can take an informed decision about choosing the neighborhood based upon their requirement.

The difference between the top 3 Clusters is not that significant, although Cluster 3 is a clear winner. Refining the venue category list to include only specific restaurants and omitting others like a café, bars, bakery, etc. would further help in choosing a Cluster.

6. Conclusion

Based on the findings in results and discussion section, the suggested neighborhoods for maximum probability of success of a new Grocery store are in Cluster 3. If the potential client has any specific preferences regarding the customer that needs targeted viz. restaurants serving specific cuisines or population, budget, demands of inventory, transportation preferences for supply, etc. then those must be considered and incorporated accordingly.