Data Science Capstone Project: The battle of neighborhoods in New York City

APRIL 21, 2019

Contents

1. Introduction

The city of New York is described as the most populous city in the United States (US) with an estimated population of 8.3 million people as at 2017 (New York City, n.d.). Located in the state of New York, New York City (NYC) is the center of the New York Metropolitan Area (NYMA), the largest urban population area in the world by landmass (New York City, n.d.). Using location data to explore the geographical location of NYC, the neighborhoods of its five (5) boroughs will be segmented and clustered to determine the best borough and recommended neighborhoods in that borough, to open a new Caribbean cuisine restaurant.

1.1 Problem description

NYC consists of five boroughs: Brooklyn, Queens, Manhattan, The Bronx and Staten Island and is home to more than 3.2 million people born outside of the US (New York City, n.d.). Given NYC's multi-cultural make-up and equally diverse culinary scene, it is the aim of this exercise to obtain and explore data regarding the neighborhoods in NYC in order to determine the optimum location to establish the first of a franchise chain of dedicated Caribbean cuisine restaurants. The success of the first establishment will determine the opening of other restaurant locations within neighboring communities to build the franchise chain.

Today, thousands of restaurants exist in NYC; with varied menu offerings. The opening of a restaurant is a significant undertaking with a potentially huge profit margin. The success of a restaurant like any other business is dependent on choice of location, accessibility and visibility, demand, population and competition. There is a host of information available which may overwhelm an entrepreneur or investor. The cost of hiring a team of consultants to assist with the decision-making process may also be burdensome. As part of formulating a business plan, data science methodologies and Foursquare API will be applied to determine the most favorable NYC borough and neighborhood to open the first restaurant of the Caribbean Cuisine franchise, with the following criteria in mind:

Primary:

- ✓ Borough selection based on not more than 10 Caribbean restaurant categories currently in existence
- ✓ Borough selection based on not less than 7 farmers’ market categories currently in existence

Secondary:

- ✓ Borough selection based on the presence of bus stop venue categories
- ✓ Recommended neighborhoods based on the most common venue categories being non-competing restaurant menus (i.e. first most common and/or second most common venue is a non-Caribbean Restaurant)
- ✓ Presence of at least one landmark or monument

1.2 Target audience

This report is suitable for use by investors, banks and other commercial lenders approached to supply funding for this business venture. The report is also suitable for use by any entrepreneur interested in the restaurant business, to assist with drafting their business plan to determine the first site for his/her Caribbean cuisine restaurant within the NYC market. The completed report will accompany the entrepreneur's business plan and presented to investors or lenders in assessing the feasibility of the proposal.

2. Data Sources

In order to appropriately identify, sort, examine and present a solution to the problem of the most ideal location for a new Caribbean restaurant, the following approach will be applied:

1. Download and explore dataset for the city of New York
2. Load and explore neighborhoods of New York City using *Foursquare API*
3. Analyze each neighborhood using *one hot encoding*, *geopy library* and *json*
4. Cluster the neighborhoods into three (3) clusters to determine similarities using *k-means*
5. Examine each cluster based on the criteria outlined in the Problem Description section

Exclusions: Demographic and population data per borough/neighborhood will be excluded from this report.

2.1 Data sources and description

Having downloaded all the dependencies (libraries) required, the following datasets will be accessed in order to upon the information required to be examined for decision making:
The five (5) boroughs and their accompanying neighborhoods will be downloaded using a *wget command*, with the dataset available from:
https://geo.nyu.edu/catalog/nyu_2451_34572.

This dataset will allow for:

1. The identification of the five (5) unique boroughs that comprise NYC
2. The partitioning of the neighborhoods in their respective boroughs
3. The provisioning of the unique latitude and longitude coordinates of each neighborhood in its respective borough
4. The provisioning of a map of NYC neighborhoods
5. The conversion of unique neighborhood addresses to their equivalent latitude and longitude values
6. A unique data frame for each borough and its respective neighborhoods complete with specific venues and venue categories for analyses, borough selection and subsequent clustering of neighborhoods in the selected borough to determine the most suitable neighborhood to open a Caribbean cuisine restaurant.

3. Methodology

The list of NYC neighborhoods and boroughs will first be retrieved, and their corresponding latitude and longitude found using *GEOPY library*. These geographical coordinates will then be placed into *Foursquare* in order to obtain common venue categories, cuisine types and other variables in each neighborhood. The neighborhoods of the selected borough will then be grouped into clusters using *k-means*.

The following libraries will be used to assist with data manipulation from the data source:

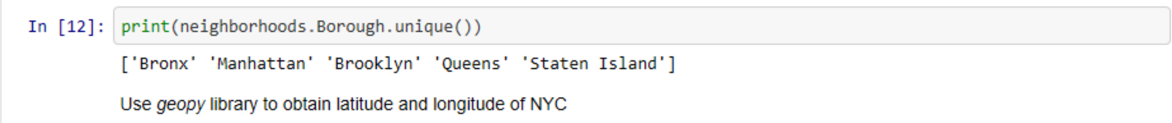
1. Numpy library for data structuring
2. Pandas library to read data into a pandas data frame for observation
3. Json library and json_data
4. Requests library
5. Sklear.cluster for KMeans to segment neighborhoods
6. Geopy to convert addresses to their equivalent latitude and longitude
7. Folium for map creation

The research will therefore include the following steps for analyses, after importing the appropriate libraries:

1. Retrieve list of neighborhoods and boroughs in NYC and create a data frame
2. Use geopy library to retrieve the coordinates of NYC
3. Use folium library to generate map of NYC
4. Input coordinates into Foursquare to explore NYC neighborhoods
5. Search for specific venue categories using Foursquare
6. Use the GET request to examine results
7. Group venues by category and total all venues per neighborhood per borough
8. Extract any restaurants labelled *Caribbean Restaurant* and *Farmers' Market* in the *venue* category
9. Select optimal borough
10. Define additional information of interest (Bus Stop and Landmark/Monument) and filter data frame
11. Cluster neighborhoods belong to the selected borough to determine ideal location for a new Caribbean cuisine restaurant

4. Results

NYC is comprised of five (5) unique boroughs containing multiple neighborhoods.



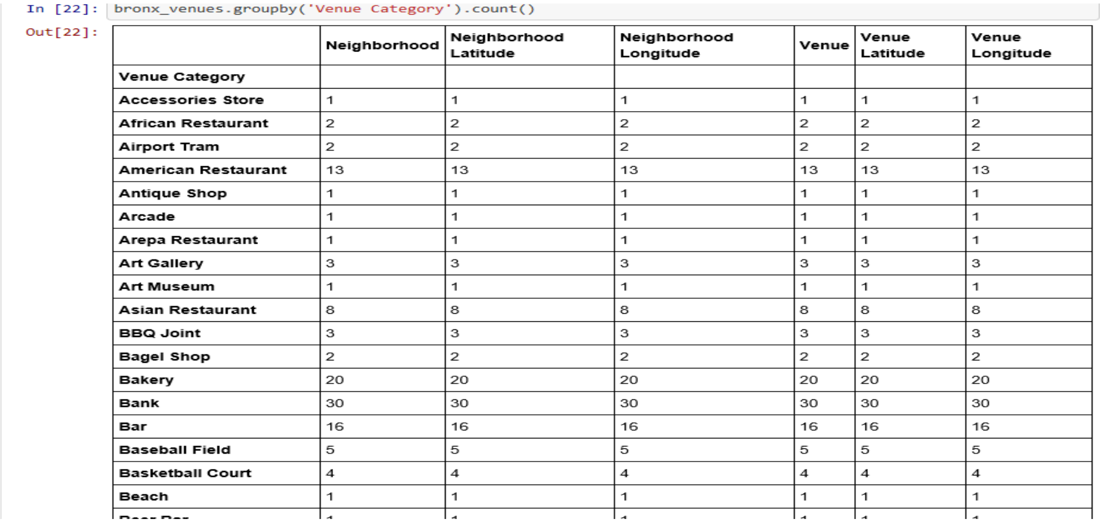
```
In [12]: print(neighborhoods.Borough.unique())

         ['Bronx' 'Manhattan' 'Brooklyn' 'Queens' 'Staten Island']
```

Use *geopy* library to obtain latitude and longitude of NYC

Close observation of the venues within each borough was significant in determining the borough for which the restaurant should be opened.

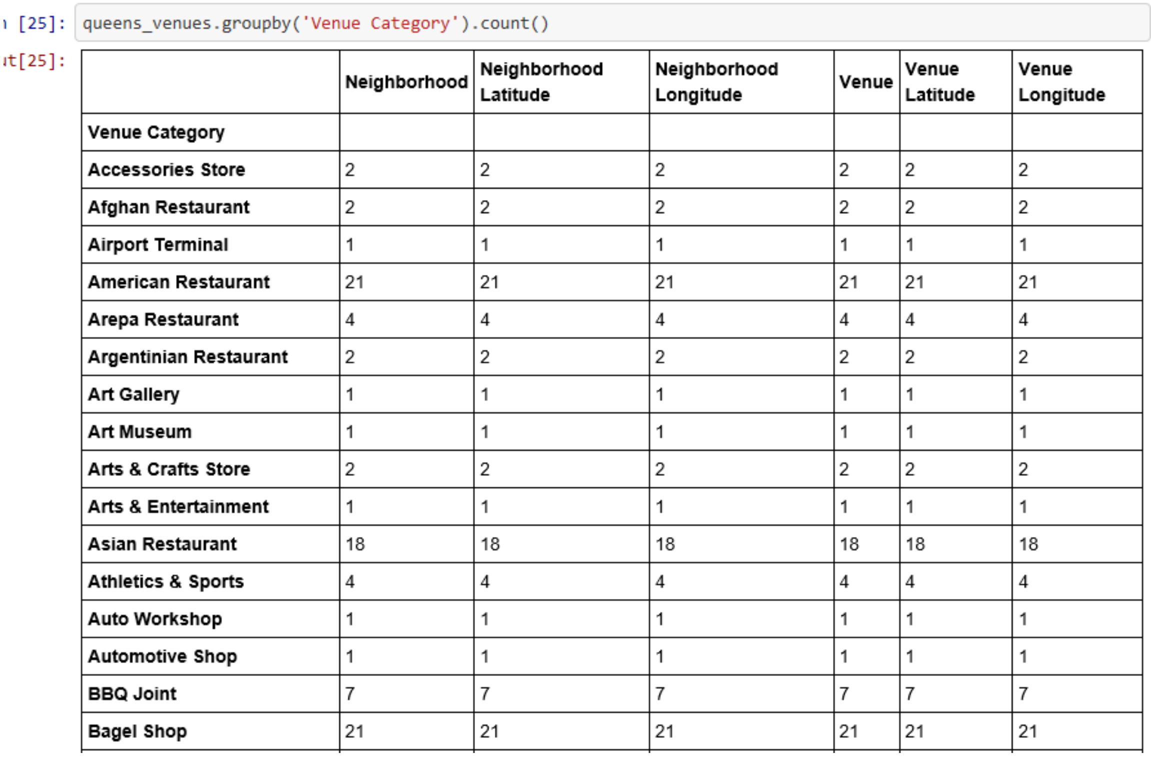
i. The Bronx Borough



```
In [22]: bronx_venues.groupby('Venue Category').count()
Out[22]:
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Accessories Store	1	1	1	1	1	1
African Restaurant	2	2	2	2	2	2
Airport Tram	2	2	2	2	2	2
American Restaurant	13	13	13	13	13	13
Antique Shop	1	1	1	1	1	1
Arcade	1	1	1	1	1	1
Arepa Restaurant	1	1	1	1	1	1
Art Gallery	3	3	3	3	3	3
Art Museum	1	1	1	1	1	1
Asian Restaurant	8	8	8	8	8	8
BBQ Joint	3	3	3	3	3	3
Bagel Shop	2	2	2	2	2	2
Bakery	20	20	20	20	20	20
Bank	30	30	30	30	30	30
Bar	16	16	16	16	16	16
Baseball Field	5	5	5	5	5	5
Basketball Court	4	4	4	4	4	4
Beach	1	1	1	1	1	1
Beer Bar	1	1	1	1	1	1

ii. Queens Borough

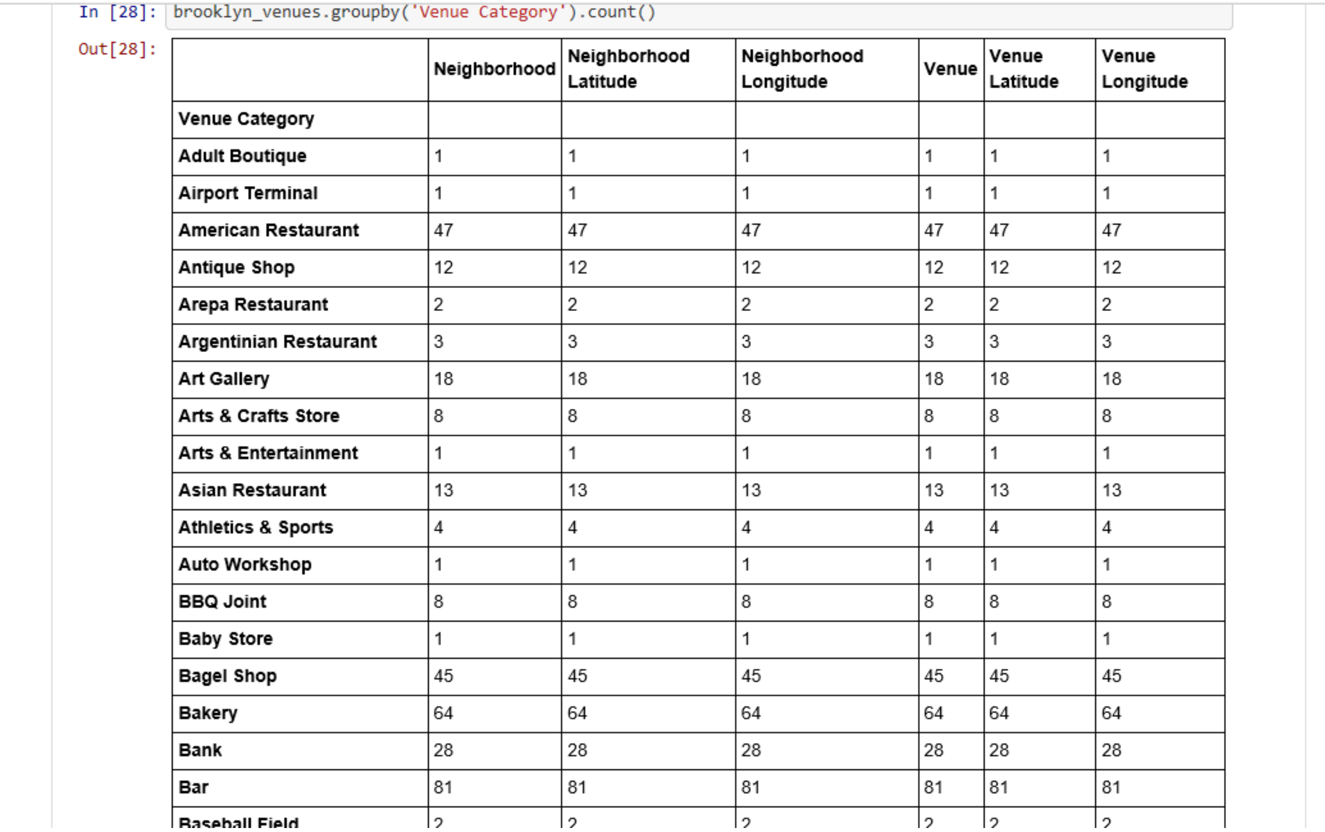


```
ı [25]: queens_venues.groupby('Venue Category').count()
```

ıt[25]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Accessories Store	2	2	2	2	2	2
Afghan Restaurant	2	2	2	2	2	2
Airport Terminal	1	1	1	1	1	1
American Restaurant	21	21	21	21	21	21
Arepa Restaurant	4	4	4	4	4	4
Argentinian Restaurant	2	2	2	2	2	2
Art Gallery	1	1	1	1	1	1
Art Museum	1	1	1	1	1	1
Arts & Crafts Store	2	2	2	2	2	2
Arts & Entertainment	1	1	1	1	1	1
Asian Restaurant	18	18	18	18	18	18
Athletics & Sports	4	4	4	4	4	4
Auto Workshop	1	1	1	1	1	1
Automotive Shop	1	1	1	1	1	1
BBQ Joint	7	7	7	7	7	7
Bagel Shop	21	21	21	21	21	21

iii. Brooklyn Borough

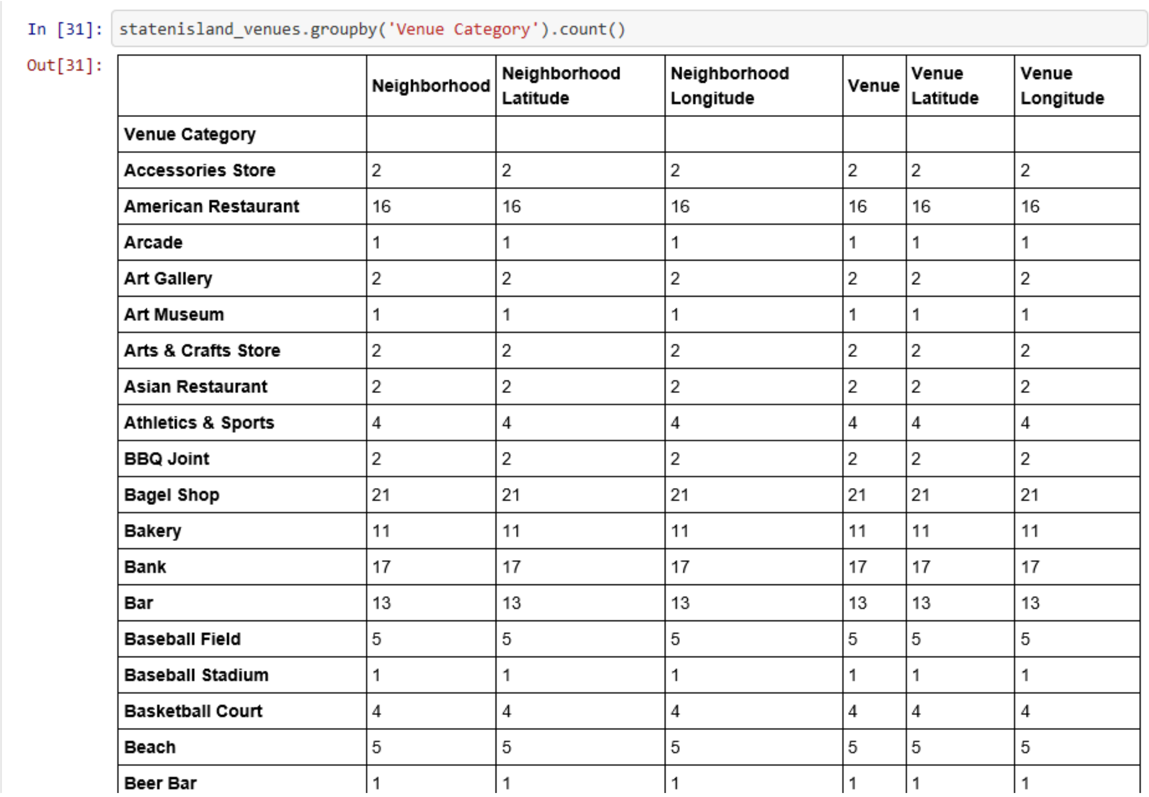


```
In [28]: brooklyn_venues.groupby('Venue Category').count()
```

Out[28]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Adult Boutique	1	1	1	1	1	1
Airport Terminal	1	1	1	1	1	1
American Restaurant	47	47	47	47	47	47
Antique Shop	12	12	12	12	12	12
Arepa Restaurant	2	2	2	2	2	2
Argentinian Restaurant	3	3	3	3	3	3
Art Gallery	18	18	18	18	18	18
Arts & Crafts Store	8	8	8	8	8	8
Arts & Entertainment	1	1	1	1	1	1
Asian Restaurant	13	13	13	13	13	13
Athletics & Sports	4	4	4	4	4	4
Auto Workshop	1	1	1	1	1	1
BBQ Joint	8	8	8	8	8	8
Baby Store	1	1	1	1	1	1
Bagel Shop	45	45	45	45	45	45
Bakery	64	64	64	64	64	64
Bank	28	28	28	28	28	28
Bar	81	81	81	81	81	81
Baseball Field	2	2	2	2	2	2

iv. Staten Island Borough

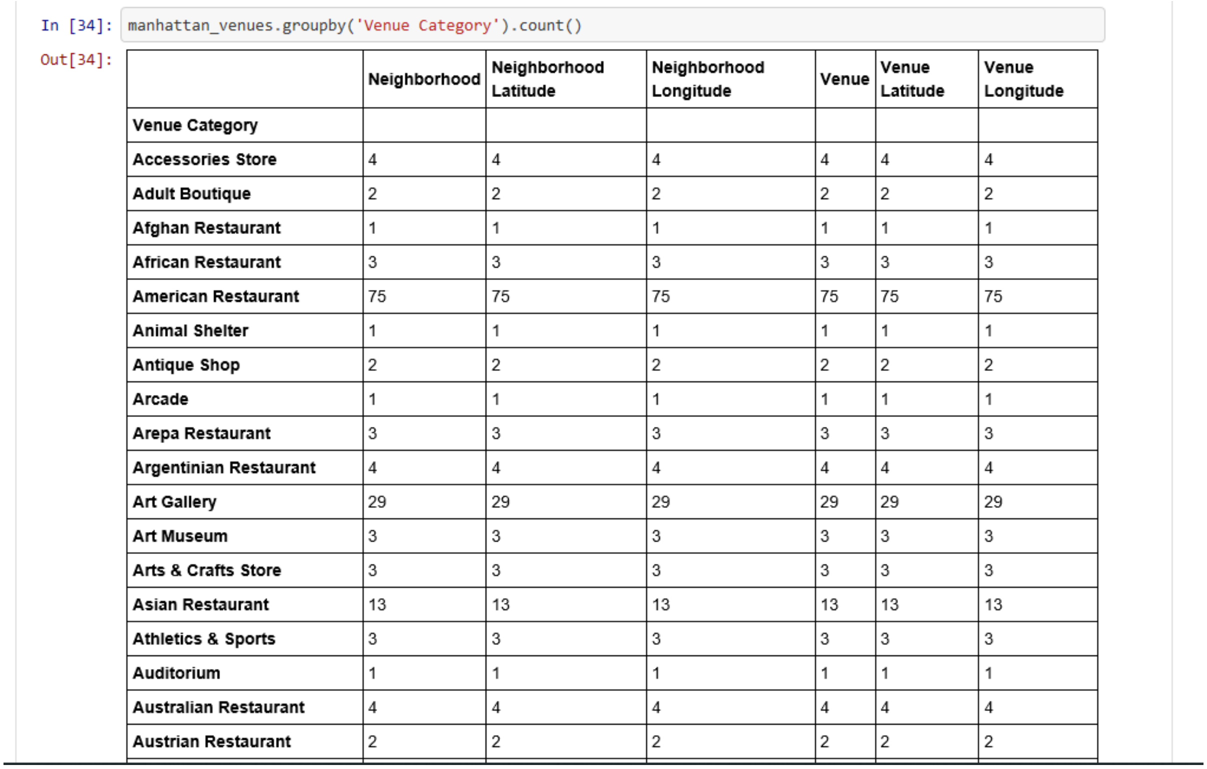


```
In [31]: statenisland_venues.groupby('Venue Category').count()
```

Out[31]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Accessories Store	2	2	2	2	2	2
American Restaurant	16	16	16	16	16	16
Arcade	1	1	1	1	1	1
Art Gallery	2	2	2	2	2	2
Art Museum	1	1	1	1	1	1
Arts & Crafts Store	2	2	2	2	2	2
Asian Restaurant	2	2	2	2	2	2
Athletics & Sports	4	4	4	4	4	4
BBQ Joint	2	2	2	2	2	2
Bagel Shop	21	21	21	21	21	21
Bakery	11	11	11	11	11	11
Bank	17	17	17	17	17	17
Bar	13	13	13	13	13	13
Baseball Field	5	5	5	5	5	5
Baseball Stadium	1	1	1	1	1	1
Basketball Court	4	4	4	4	4	4
Beach	5	5	5	5	5	5
Beer Bar	1	1	1	1	1	1

v. Manhattan Borough



In [34]:
```
manhattan_venues.groupby('Venue Category').count()
```

Out[34]:

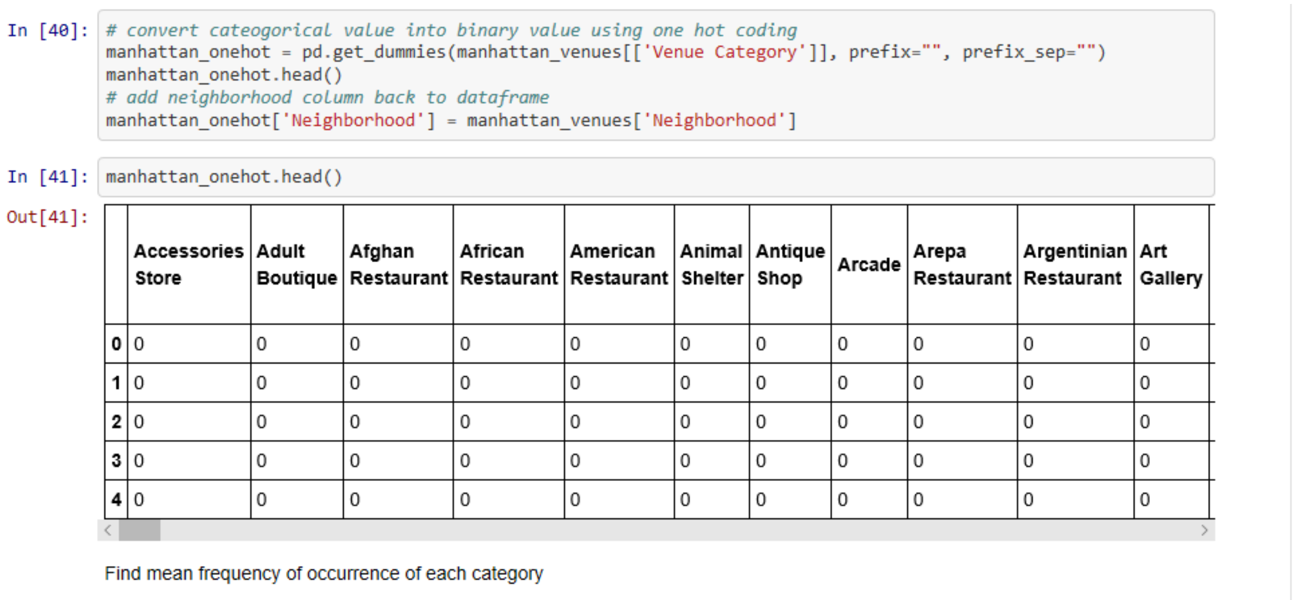
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Accessories Store	4	4	4	4	4	4
Adult Boutique	2	2	2	2	2	2
Afghan Restaurant	1	1	1	1	1	1
African Restaurant	3	3	3	3	3	3
American Restaurant	75	75	75	75	75	75
Animal Shelter	1	1	1	1	1	1
Antique Shop	2	2	2	2	2	2
Arcade	1	1	1	1	1	1
Arepa Restaurant	3	3	3	3	3	3
Argentinian Restaurant	4	4	4	4	4	4
Art Gallery	29	29	29	29	29	29
Art Museum	3	3	3	3	3	3
Arts & Crafts Store	3	3	3	3	3	3
Asian Restaurant	13	13	13	13	13	13
Athletics & Sports	3	3	3	3	3	3
Auditorium	1	1	1	1	1	1
Australian Restaurant	4	4	4	4	4	4
Austrian Restaurant	2	2	2	2	2	2

Based on the primary criteria-no more than 10 Caribbean Restaurants present and not less than 7 Farmers' Markets present in the preferred borough- the Manhattan borough was selected as the borough of choice having only nine (9) Caribbean restaurant and nine (9) Farmers' Market categories respectively. The remaining boroughs did not meet the primary criteria as they represented:

i. The Bronx: 15 Caribbean Restaurants and 1 Farmers' Markets
ii. Queens: 18 Caribbean Restaurants and 5 Farmer's Markets
iii. Brooklyn: 15 Caribbean Restaurants and 1 Farmers' Markets
iv. Staten Island: 2 Caribbean Restaurants and 0 Farmers' Markets

For the preferred borough of Manhattan, the *onehot encoding* function was then used to convert Manhattan's categorical data to binary data in order to find the mean frequency of occurrence of each venue category type. This would assist in determining which neighborhood in Manhattan met the secondary criteria of:

i. Presence of a landmark/monument
ii. Presence of bus stops
iii. Presence of non-competing venues as the first and most common venue



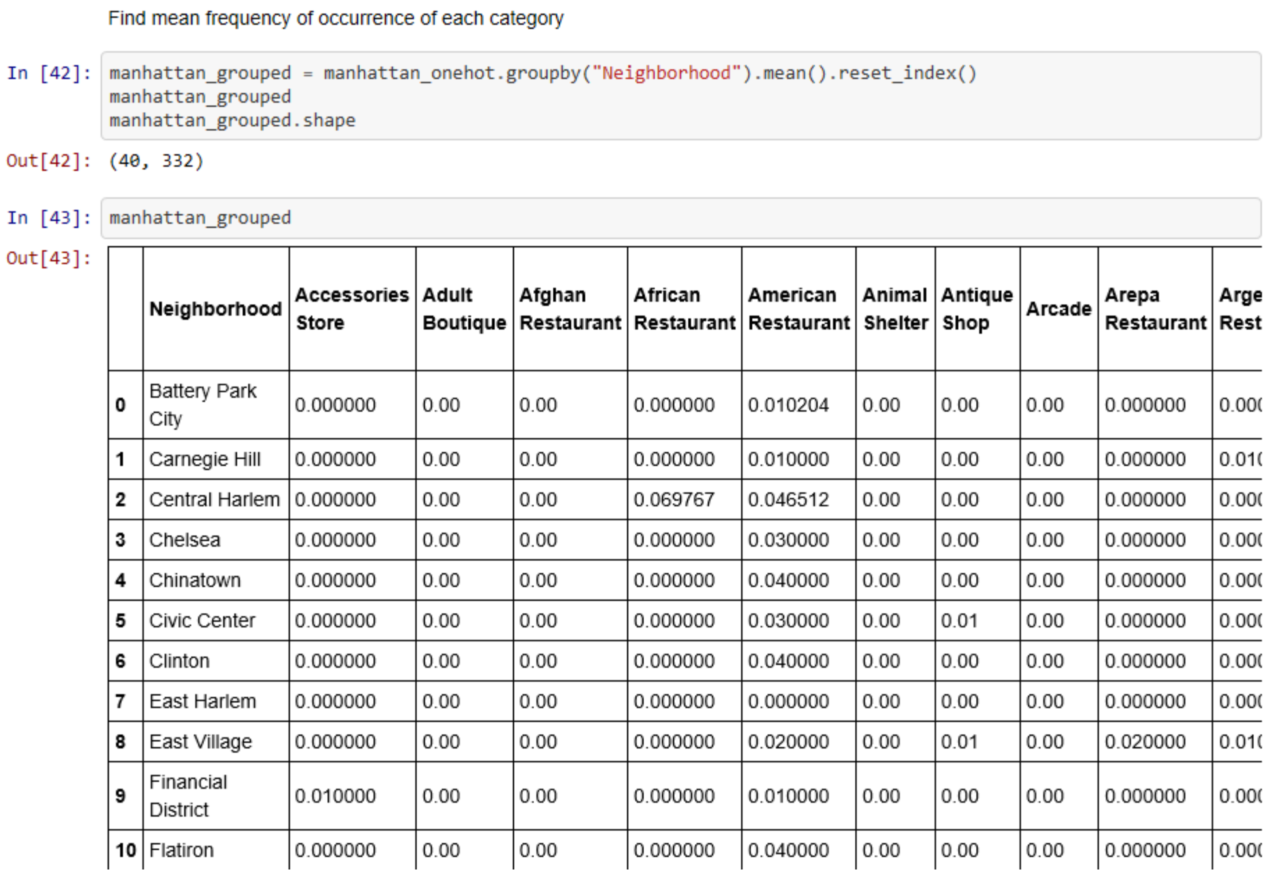
In [40]:
```
# convert cateogorical value into binary value using one hot coding
manhattan_onehot = pd.get_dummies(manhattan_venues[['Venue Category']], prefix="", prefix_sep="")
manhattan_onehot.head()
# add neighborhood column back to dataframe
manhattan_onehot['Neighborhood'] = manhattan_venues['Neighborhood']
```

In [41]:
```
manhattan_onehot.head()
```

Out[41]:

	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0

Find mean frequency of occurrence of each category



Find mean frequency of occurrence of each category

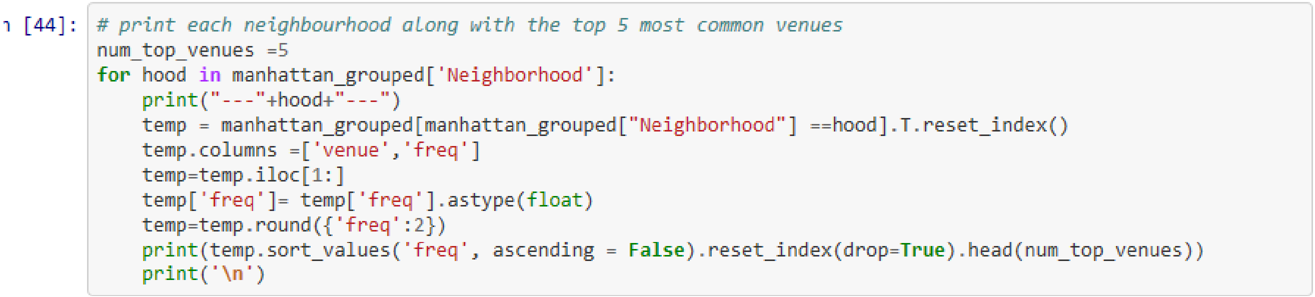
In [42]:
```
manhattan_grouped = manhattan_onehot.groupby("Neighborhood").mean().reset_index()
manhattan_grouped
manhattan_grouped.shape
```

Out[42]:
```
(40, 332)
```

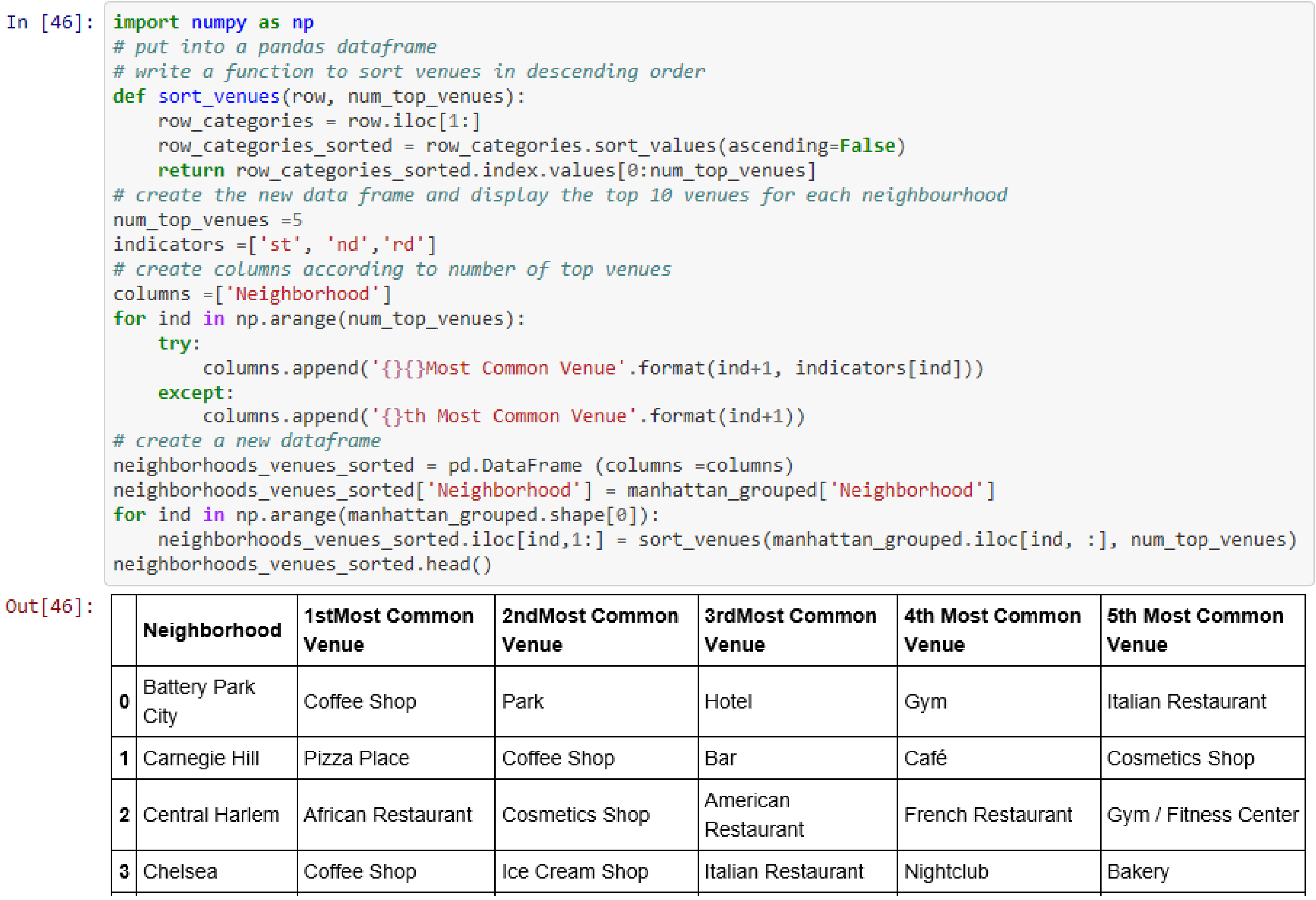
In [43]:
```
manhattan_grouped
```

Out[43]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Arge Rest
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010204	0.00	0.00	0.00	0.000000	0.000
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.000000	0.010
2	Central Harlem	0.000000	0.00	0.00	0.069767	0.046512	0.00	0.00	0.00	0.000000	0.000
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.00	0.000000	0.000
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000
5	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.01	0.00	0.000000	0.000
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.000
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.01	0.00	0.020000	0.010
9	Financial District	0.010000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.000000	0.000
10	Flatiron	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000



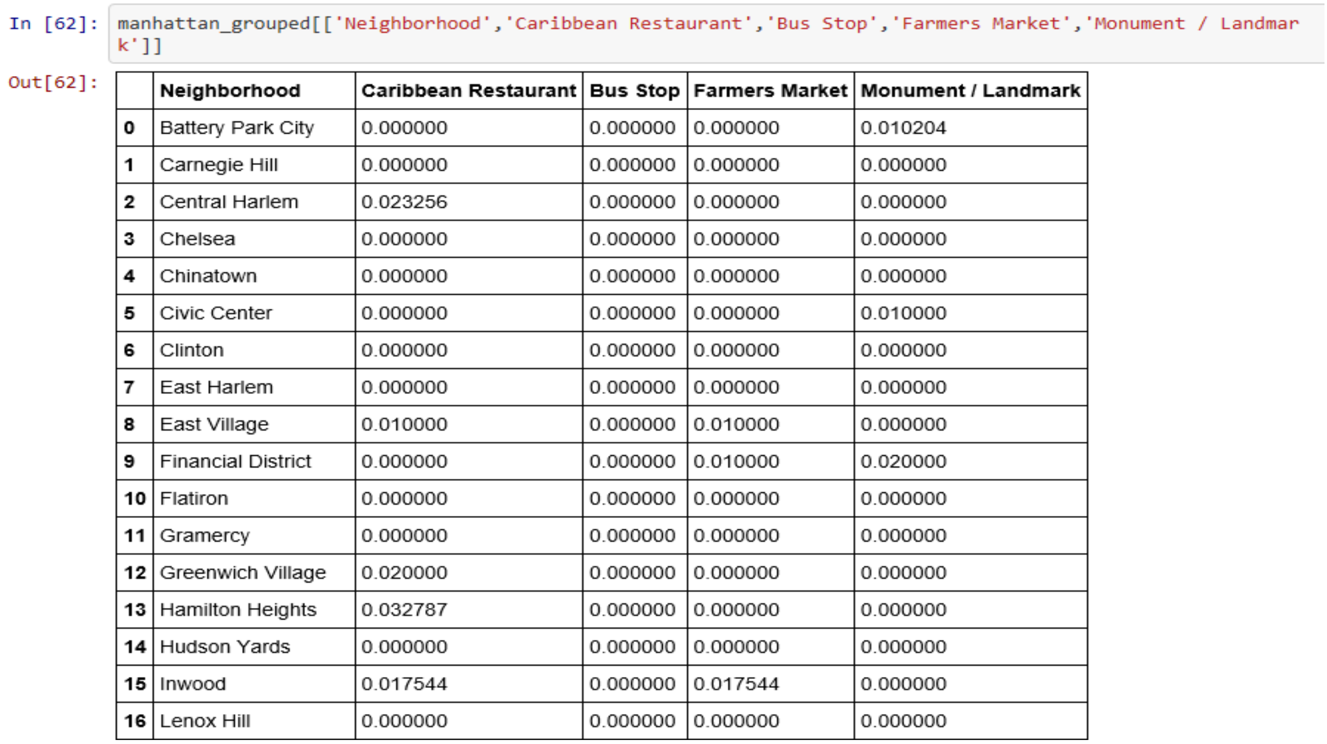
```
ı [44]: # print each neighbourhood along with the top 5 most common venues
        num_top_venues =5
        for hood in manhattan_grouped['Neighborhood']:
            print("---"+hood+"---")
            temp = manhattan_grouped[manhattan_grouped["Neighborhood"] ==hood].T.reset_index()
            temp.columns =['venue','freq']
            temp=temp.iloc[1:]
            temp['freq']= temp['freq'].astype(float)
            temp=temp.round({'freq':2})
            print(temp.sort_values('freq', ascending = False).reset_index(drop=True).head(num_top_venues))
            print('\n')
```



```
In [46]: import numpy as np
         # put into a pandas dataframe
         # write a function to sort venues in descending order
         def sort_venues(row, num_top_venues):
             row_categories = row.iloc[1:]
             row_categories_sorted = row_categories.sort_values(ascending=False)
             return row_categories_sorted.index.values[0:num_top_venues]
         # create the new data frame and display the top 10 venues for each neighbourhood
         num_top_venues =5
         indicators =['st', 'nd','rd']
         # create columns according to number of top venues
         columns =['Neighborhood']
         for ind in np.arange(num_top_venues):
             try:
                 columns.append('{}{}Most Common Venue'.format(ind+1, indicators[ind]))
             except:
                 columns.append('{}th Most Common Venue'.format(ind+1))
         # create a new dataframe
         neighborhoods_venues_sorted = pd.DataFrame (columns =columns)
         neighborhoods_venues_sorted['Neighborhood'] = manhattan_grouped['Neighborhood']
         for ind in np.arange(manhattan_grouped.shape[0]):
             neighborhoods_venues_sorted.iloc[ind,1:] = sort_venues(manhattan_grouped.iloc[ind, :], num_top_venues)
         neighborhoods_venues_sorted.head()
```

Out[46]:

	Neighborhood	1stMost Common Venue	2ndMost Common Venue	3rdMost Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	Coffee Shop	Park	Hotel	Gym	Italian Restaurant
1	Carnegie Hill	Pizza Place	Coffee Shop	Bar	Café	Cosmetics Shop
2	Central Harlem	African Restaurant	Cosmetics Shop	American Restaurant	French Restaurant	Gym / Fitness Center
3	Chelsea	Coffee Shop	Ice Cream Shop	Italian Restaurant	Nightclub	Bakery

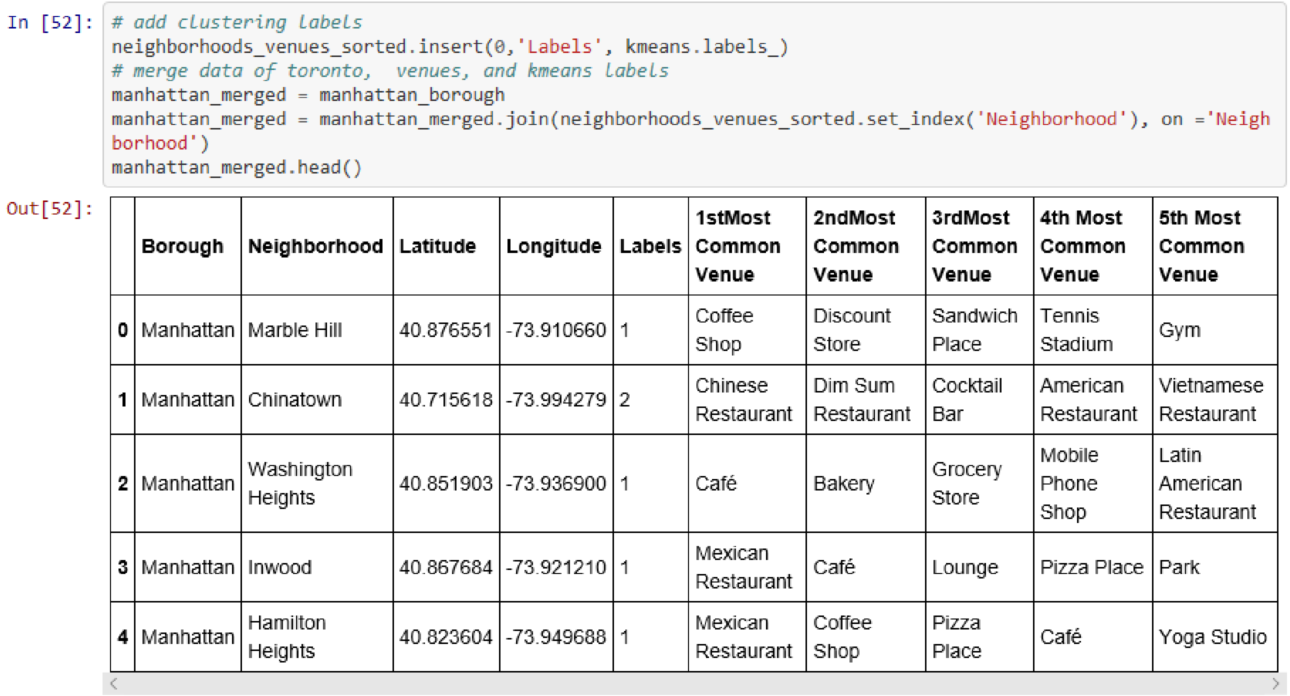


```
In [62]: manhattan_grouped[['Neighborhood','Caribbean Restaurant','Bus Stop','Farmers Market','Monument / Landmar
k']]
```

Out[62]:

	Neighborhood	Caribbean Restaurant	Bus Stop	Farmers Market	Monument / Landmark
0	Battery Park City	0.000000	0.000000	0.000000	0.010204
1	Carnegie Hill	0.000000	0.000000	0.000000	0.000000
2	Central Harlem	0.023256	0.000000	0.000000	0.000000
3	Chelsea	0.000000	0.000000	0.000000	0.000000
4	Chinatown	0.000000	0.000000	0.000000	0.000000
5	Civic Center	0.000000	0.000000	0.000000	0.010000
6	Clinton	0.000000	0.000000	0.000000	0.000000
7	East Harlem	0.000000	0.000000	0.000000	0.000000
8	East Village	0.010000	0.000000	0.010000	0.000000
9	Financial District	0.000000	0.000000	0.010000	0.020000
10	Flatiron	0.000000	0.000000	0.000000	0.000000
11	Gramercy	0.000000	0.000000	0.000000	0.000000
12	Greenwich Village	0.020000	0.000000	0.000000	0.000000
13	Hamilton Heights	0.032787	0.000000	0.000000	0.000000
14	Hudson Yards	0.000000	0.000000	0.000000	0.000000
15	Inwood	0.017544	0.000000	0.017544	0.000000
16	Lenox Hill	0.000000	0.000000	0.000000	0.000000

The neighborhoods of Manhattan were also clustered by venue category to assist with identifying the top five (5) most common venue type by category and the respective defined neighborhood cluster label. The below incudes the neighborhood of Inwood which has a non-competing restaurant venue and from the earlier frequency exercise also has a Farmers' Market.



```
In [52]: # add clustering labels
         neighborhoods_venues_sorted.insert(0,'Labels', kmeans.labels_)
         # merge data of toronto,  venues, and kmeans labels
         manhattan_merged = manhattan_borough
         manhattan_merged = manhattan_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on ='Neigh
         borhood')
         manhattan_merged.head()
```

Out[52]:

	Borough	Neighborhood	Latitude	Longitude	Labels	1stMost Common Venue	2ndMost Common Venue	3rdMost Common Venue	4th Most Common Venue	5th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	1	Coffee Shop	Discount Store	Sandwich Place	Tennis Stadium	Gym
1	Manhattan	Chinatown	40.715618	-73.994279	2	Chinese Restaurant	Dim Sum Restaurant	Cocktail Bar	American Restaurant	Vietnamese Restaurant
2	Manhattan	Washington Heights	40.851903	-73.936900	1	Café	Bakery	Grocery Store	Mobile Phone Shop	Latin American Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	1	Mexican Restaurant	Café	Lounge	Pizza Place	Park
4	Manhattan	Hamilton Heights	40.823604	-73.949688	1	Mexican Restaurant	Coffee Shop	Pizza Place	Café	Yoga Studio

5. Recommendation and Conclusion

The city of NYC, the Manhattan borough in particular, is described as a multi-cultural community. With hundreds of restaurants offering over 100 cuisine types, the decision to open a restaurant in NYC requires substantial research in order to determine not only the best borough, but the best neighborhood in that borough to serve as the most profitable location.

Based on the analysis performed, the Manhattan borough was selected as it had the least amount of Caribbean restaurant venue categories and the most amount of Farmers' Markets in keeping with the primary requirements. Given the criteria outlined and the results from venue grouping and clustering, the Inwood neighborhood in cluster 1 would serve as a primary location to open a Caribbean cuisine restaurant. The neighborhood of Central Harlem in cluster 2 may also be considered as a suitable location for the open of a second location for the restaurant as it builds its franchise. The neighborhoods of Innwood and Central Harlem are suitable owing to:

1. A low mix of other restaurants exists which do not offer a Caribbean menu. It can be assumed that the frequency of the different restaurant types will also result in customer traffic being directed to the new Caribbean cuisine restaurant.
2. The existence of Farmers' Markets in order to source the fresh produce and spices required to support a Caribbean cuisine menu.
3. The presence of bus stops to support public transportation.

The entrepreneur seeking to open a Caribbean cuisine restaurant may therefore complete the business proposal with two appropriate locations to the team of investors or lending institution in order to obtain funding.

6. References

New York City (n.d.) Wikipedia. Retrieved April 12, 2019 from https://en.wikipedia.org/wiki/New_York_City