Introduction

- The City of New York, is the most populous city in the United States. It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business opportunities and business friendly environment. It has attracted many different players into the market

- It is a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States

- This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully

Business Problem

- The City of New York is famous for Wall Street. There are lots of financial elites, and bar is the most place for them to relax after work.

- In fact, if you open map, you will find bars are everywhere not only in major cities but also in smaller cities. Starting a bar can be a great business opportunity, but you need to distinguish yourself from others to enjoy long-term success

Data Collection

- The data of the neighborhoods in New York was scraped from https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json. The data is read into a pandas data frame using wget download method. The main reason for doing so is that the IMB Cloud provides a comprehensive and detailed table of the data which can easily be used

- The geographical coordinates for New York data has been obtained from the GeoPy library in python. This data is relevant for plotting the map of New York using the Folium library in python. The geocoder library in python has been used to obtain latitude and longitude data for various neighborhoods in New York. The coordinates of all neighborhoods in New York are used to check the accuracy of coordinates given on Wikipedia and replace them in our data frame if the absolute difference is more than 0.001. These coordinates are then further used for plotting using the Folium library in python.

- The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in New York and is used to study the popular venues of different neighborhoods.

Dataset for New York neighborhoods

	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

Dataset for top 5 most comment venues

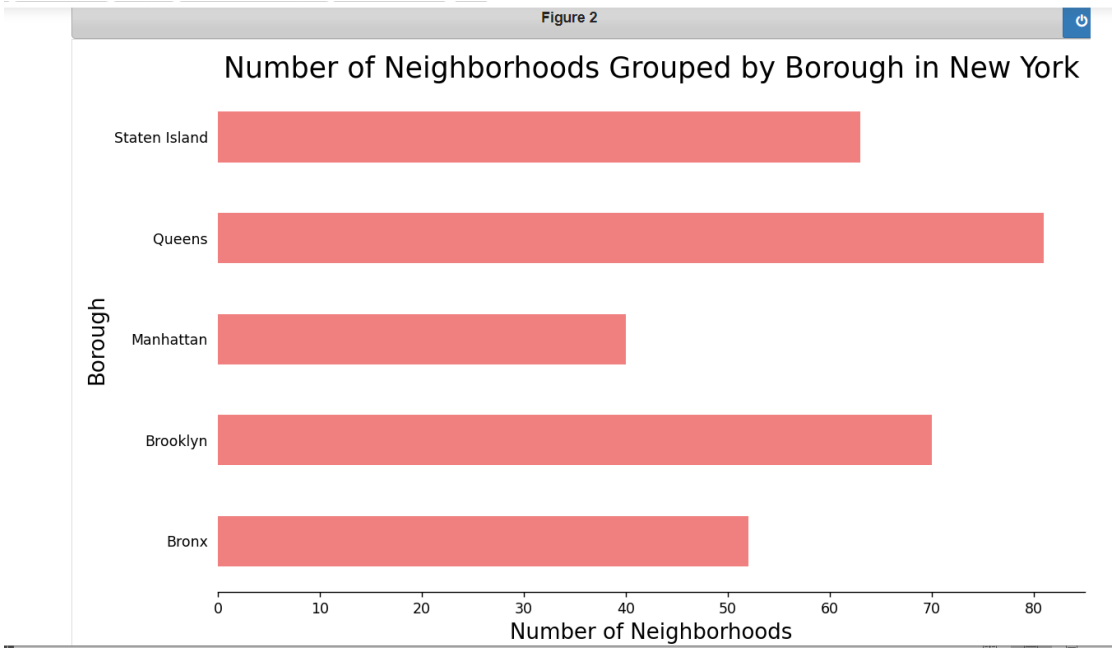
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Wakefield	Pharmacy	Supermarket	Caribbean Restaurant	Donut Shop	Fried Chicken Joint
1	Co-op City	Department Store	Pizza Place	Mobile Phone Shop	Shopping Mall	Pharmacy
2	Eastchester	Caribbean Restaurant	Diner	Pizza Place	Shopping Mall	Fast Food Restaurant
3	Fieldston	Bus Station	Pizza Place	Bar	Deli / Bodega	Mexican Restaurant
4	Riverdale	Bar	Park	Pizza Place	Bank	Bagel Shop
...	...	...	...	...	...	...
297	Hudson Yards	Gym / Fitness Center	Hotel	Theater	Coffee Shop	Lounge
298	Hammels	Beach	Surf Spot	Bar	Wine Bar	Pizza Place
299	Bayswater	Grocery Store	Playground	Park	Chinese Restaurant	Other Great Outdoors
300	Queensbridge	Hotel	Coffee Shop	Café	Sandwich Place	Bar
301	Fox Hills	Chinese Restaurant	Spanish Restaurant	Deli / Bodega	Pharmacy	Park

302 rows × 6 columns

Methodology

- ## Data Visualization

New York neighborhoods data was plotted for providing a better understanding, The graph alongside depicts the number of neighborhoods in each location of New York. All neighborhoods on the outskirts of the city have been grouped as “New York”. Queens and Brooklyn contain the highest number of neighborhoods



The Folium library in python was used to visualize the spread of all neighborhoods across New York

OneHotEncoding

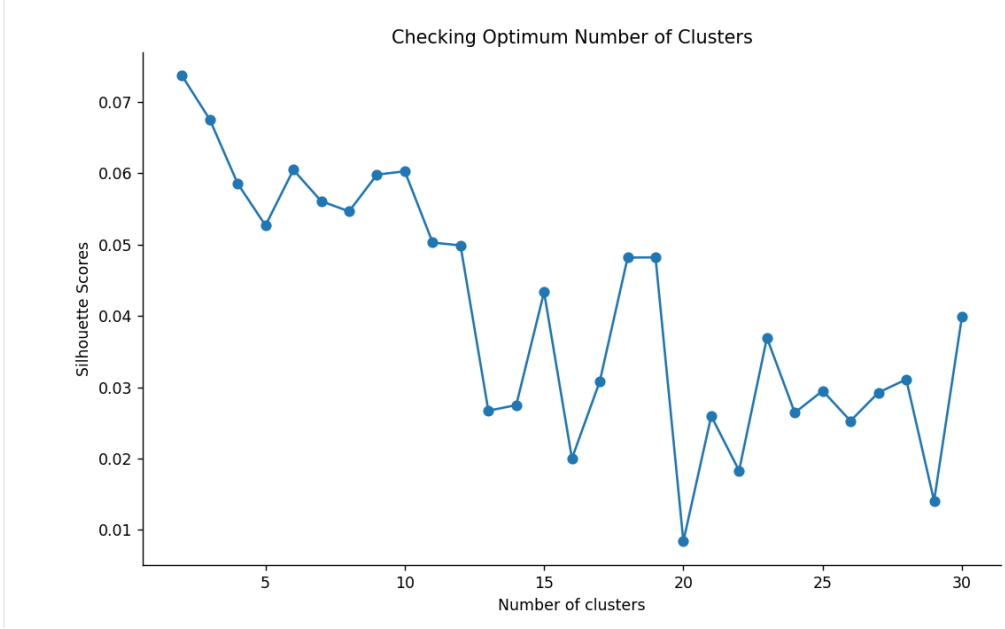
One-hot Encoding was used to encode venue categories to numeric values with 1 if a venue belongs to a category and 0 if a venue does not belong to a category for all neighborhoods. The average is then taken for all venue categories in a neighborhood to produce the data frame shown

	Neighborhood	ATM	Accessories Store	Adult Boutique	Adult Education Center	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	...	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Wome St
0	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
4	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 483 columns

Unsupervised Learning Model

KMeans clustering was used to cluster neighborhoods in New York based on venue categories. The plot shows a maximum Silhouette Score for 2 clusters and thus the n_clusters parameter in KMeans clustering was set to 2



Results

- Each neighborhood received a cluster label based on clustering by the KMeans algorithm
- The Cluster Label column along with the Location, Latitude, and Longitude columns were added to the top 5 most common venues data frame to provide the final results
- This data frame is shown on the next slide

Cluster 1

Cluster 1

```
[341]: ny_merged.loc[ny_merged['Cluster Labels'] == 0, ny_merged.columns[[0] + [1] + list(range(5, ny_merged.shape[1]))]]
```

ıt[341]:

	Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bronx	Wakefield	Pharmacy	Supermarket	Caribbean Restaurant	Donut Shop	Fried Chicken Joint
1	Bronx	Co-op City	Department Store	Pizza Place	Mobile Phone Shop	Shopping Mall	Pharmacy
2	Bronx	Eastchester	Caribbean Restaurant	Diner	Pizza Place	Shopping Mall	Fast Food Restaurant
3	Bronx	Fieldston	Bus Station	Pizza Place	Bar	Deli / Bodega	Mexican Restaurant
5	Bronx	Kingsbridge	Pizza Place	Mexican Restaurant	Bar	Sandwich Place	Coffee Shop
...	...	...	...	...	...	...	...
298	Bronx	Allerton	Pizza Place	Deli / Bodega	Donut Shop	Fried Chicken Joint	Caribbean Restaurant
299	Bronx	Kingsbridge Heights	Gym	Diner	Coffee Shop	Latin American Restaurant	Pizza Place
300	Brooklyn	Erasmus	Caribbean Restaurant	Discount Store	Mobile Phone Shop	Pharmacy	Pizza Place
303	Queens	Bayswater	Grocery Store	Playground	Park	Chinese Restaurant	Other Great Outdoors
305	Staten Island	Fox Hills	Chinese Restaurant	Spanish Restaurant	Deli / Bodega	Pharmacy	Park

193 rows × 7 columns

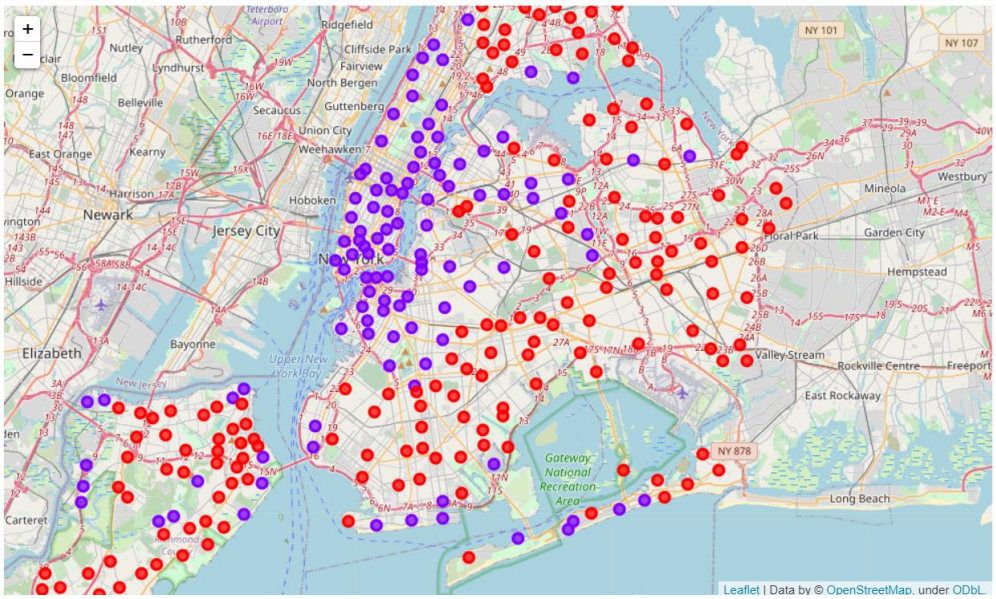
Cluster 2

	Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Bronx	Riverdale	Bar	Park	Pizza Place	Bank	Bagel Shop
12	Bronx	City Island	Harbor / Marina	Seafood Restaurant	Italian Restaurant	Bar	Boat or Ferry
19	Bronx	High Bridge	Baseball Stadium	Lounge	Plaza	Bar	Park
24	Bronx	Hunts Point	Coffee Shop	Park	Home Service	Bakery	Construction & Landscaping
27	Bronx	Clason Point	Park	Discount Store	Bus Stop	Pool	Gym / Fitness Center
...	...	...	...	...	...	...	...
288	Queens	Roxbury	Art Gallery	Baseball Field	Theater	Beach	National Park
292	Staten Island	Lighthouse Hill	Italian Restaurant	Trail	Cosmetics Shop	Bagel Shop	History Museum
301	Manhattan	Hudson Yards	Gym / Fitness Center	Hotel	Theater	Coffee Shop	Lounge
302	Queens	Hammels	Beach	Surf Spot	Bar	Wine Bar	Pizza Place
304	Queens	Queensbridge	Hotel	Coffee Shop	Café	Sandwich Place	Bar

113 rows × 7 columns

- Visualization of neighborhood clusters was done using Folium in python
- Different clusters correspond to different colors

Out[320]:

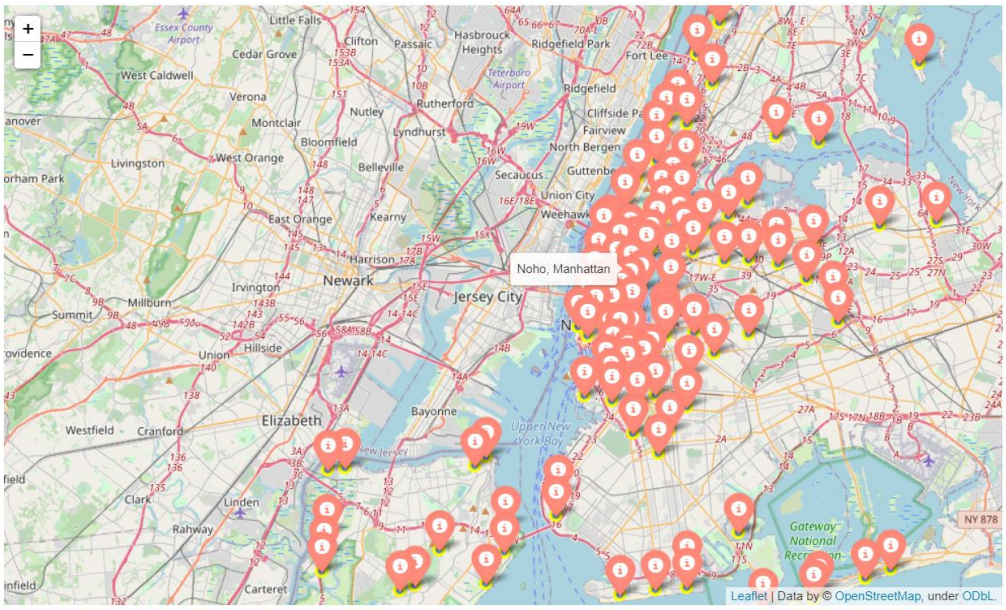


Discussion

By analyzing the 2 clusters obtained we can see, both of them are suit for Bar. cause the most common venues are Bars, Thus, the neighborhoods in these clusters would be well suited for opening a bar

Plotting final results using Folium library in python

Out[347]:



Conclusion

We have successfully analyzed the neighborhoods in New York for determining which would be the best neighborhoods for opening a new bar. Based on our analysis, neighborhoods in cluster 0 and 1 are recommended as locations for the new bar. This has also been plotted in the map above. The stakeholders and investors can further tune this by considering various other factors like transport, legal requirements, and costs associated. These were out of the scope for this project and thus were not considered.