Implementing K-means clustering to find an optimal location to open an exclusive gym in Manhattan

Table of Contents

Introduction

While exciting, starting a new business can also be very risky. There are many factors that need to be taken into consideration for the success of a new business, such as location, competition, safety. The purpose of this project is to find an optimal location for opening an exclusive gym at Manhattan, NY. To answer this question for a client who is seeking the best area to open a new exclusive gym, we first have to determine which neighborhoods are considered safe, find which neighborhoods have high income, and look into the nearby competition since we want an area that is not very competitive to increase the chances of a successful business.

Data

The following datasets were used in this project:

- NYC census to determine neighborhoods with high income
- NYPD arrests to determine the safest neighborhoods
- NYC borough and neighborhood data with Latitude and Longitude
- Venues data was obtained from Foursquare API

Methodology

The first step of is to find Manhattan neighborhoods with high income using the Census dataset. To accomplish this, we are going to merge the census track and census block datasets, and filter to only keep data that contains Manhattan income information. Since we need to extract neighborhood information from the latitude and longitude data, we are going to use reverse Geocoding to extract neighborhoods from the address and create two new columns in data frame: Neighborhood and Suburb. The reason we need Suburb information is because we notice that some of the rows were missing neighborhood information but had suburb data, and we will use that to filter data only to Manhattan. Figure 1. shows merged census data that contains 41 columns and over 18 thousand rows.

	Latitude	Longitude	BlockCode	County_x	State	Tract	County_y	Borough	TotalPop	Men	...	Walk	OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemp
15	40.480000	-74.232513	360859901000011	Richmond	NY	36085990100	Richmond	Staten Island	0	0	...	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	NaN	
16	40.480000	-74.229347	360859901000011	Richmond	NY	36085990100	Richmond	Staten Island	0	0	...	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	NaN	
17	40.480000	-74.226181	360859901000011	Richmond	NY	36085990100	Richmond	Staten Island	0	0	...	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	NaN	
18	40.480000	-74.223015	360859901000011	Richmond	NY	36085990100	Richmond	Staten Island	0	0	...	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	NaN	
19	40.480000	-74.219849	360859901000011	Richmond	NY	36085990100	Richmond	Staten Island	0	0	...	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
36715	40.911910	-73.903266	360050319001006	Bronx	NY	36005031900	Bronx	Bronx	645	163	...	57.4	0.9	6.5	20.3	216	88.4	11.6	0.0	0.0	
36911	40.914171	-73.915930	360050319000001	Bronx	NY	36005031900	Bronx	Bronx	645	163	...	57.4	0.9	6.5	20.3	216	88.4	11.6	0.0	0.0	
36912	40.914171	-73.912764	360050319000001	Bronx	NY	36005031900	Bronx	Bronx	645	163	...	57.4	0.9	6.5	20.3	216	88.4	11.6	0.0	0.0	
36913	40.914171	-73.909598	360050319001002	Bronx	NY	36005031900	Bronx	Bronx	645	163	...	57.4	0.9	6.5	20.3	216	88.4	11.6	0.0	0.0	
37111	40.916432	-73.915930	360050319000001	Bronx	NY	36005031900	Bronx	Bronx	645	163	...	57.4	0.9	6.5	20.3	216	88.4	11.6	0.0	0.0	

18052 rows × 41 columns

Figure 1. Merged data

Once we have all neighborhoods, next we focus on income. Figure 2. Represents distribution of income.

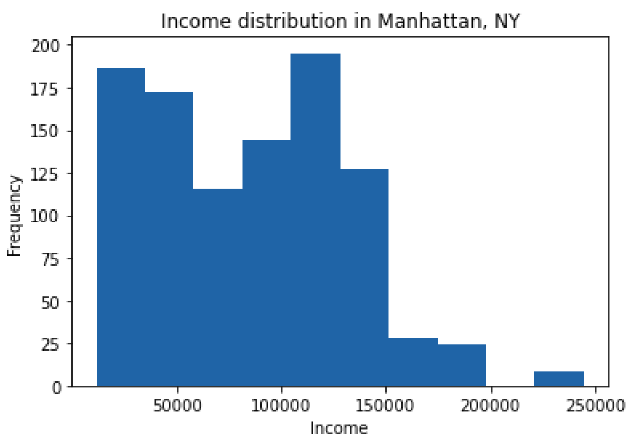

Figure 2. Income distribution

We are looking for neighborhoods with high income, so the data is filtered on neighborhoods with more than 10 occurrences of an annual income higher than $100k per neighborhood. As a result, we have a list of 16 neighborhoods (Figure 3.).

Figure 4. presents crime frequency, we can see that petit larceny, assault and burglary are top 3 crimes in Manhattan.

```
high_income.Neighborhood.value_counts()
```

```
Upper West Side        59
Chelsea                28
Financial District     21
Battery Park City      19
Lenox Hill             18
Hudson Yards           17
West Village           16
Midtown East           15
Hudson Square          15
Hell's Kitchen         14
Yorkville              13
Upper East Side        13
Morningside Heights    13
Greenwich Village      12
Tribeca                10
Turtle Bay             10
```

Figure 3. Neighborhoods with income over 100K

The next step is to find neighborhoods with low crime rates. For this, we will be using the NYPD arrest dataset. This dataset contains a lot of information on the date and time of the crime, the type of crime, the perpetrator’s sex, age, etc. For our purpose we narrowed down attributes to date, type of crime, arrest borough, and latitude and longitude. Figure 4. presents crime frequency, we can see that petit larceny, assault and burglary are top 3 crimes in Manhattan.

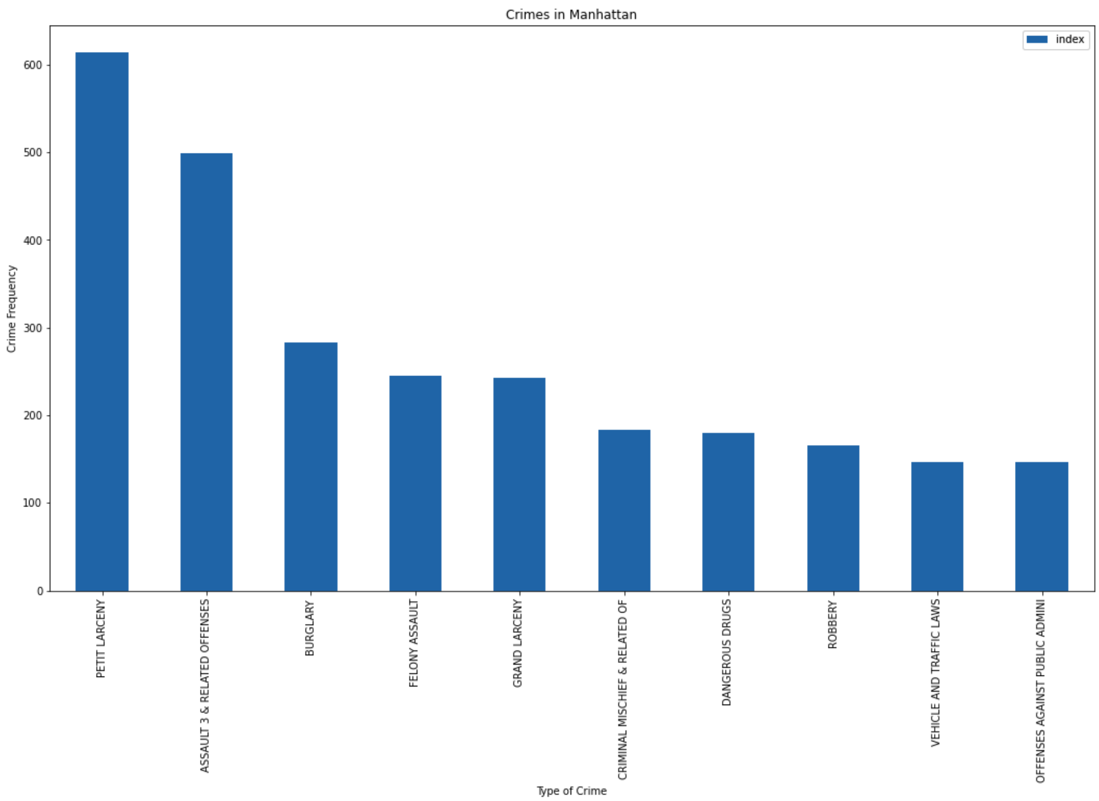

Figure 4. Frequency and type of crime

As with the previous dataset, we are going to use reverse Geocoder to extract information about the neighborhoods where the crimes occurred. Once we did that, we took a closer look into the frequency of crimes that occurred in Manhattan neighborhoods, setting the threshold to less than 300 occurrences. Our result was list of more than 30 neighborhoods with low crime. (Figure 5.)

```
low_crime.Neighborhood.value_counts(
```

```
Financial District        264
Yorkville                 245
Inwood                    243
Midtown East              242
Upper East Side           200
Alphabet City             197
Kips Bay                  193
Morningside Heights       189
Chinatown                 180
Flatiron District         163
Hudson Heights            146
Midtown South             132
Two Bridges               132
NoMad                     123
Carnegie Hill             110
Union Square               93
Hudson Square              81
Columbus Circle            70
Little Italy               69
Murray Hill                68
NoHo Historic District     58
NoHo                       53
Meatpacking District       41
Koreatown                  38
Rose Hill                  35
Lincoln Square             32
Flower District            26
Stuy Town                  23
Battery Park City          21
Tudor City                 18
```

As a final result we got the intersection between the list of neighborhoods with low crime and neighborhoods with high income and we had total of 7 neighborhoods that will be used for further analyses: *Battery Park City, Financial District, Hudson Square, Midtown East, Morningside Heights*, *Upper East Side* and *Yorkville*. In order to segment the neighborhoods and explore them, we need a dataset that contains NYC's 5 boroughs and their neighborhoods, with latitude and longitude information for each of them. We created a dataset that contains common neighborhoods and their geographical locations. Next, we utilized the Foursquare API to obtain information about venues that are found in our neighborhoods and we sorted them by frequency to get only the top 10 most frequent ones.

All the work above was done to prepare our data for the next step which was applying the K-Means machine learning algorithm to cluster our data. The measure of centrality of data can be used to analyze the difference between the means of different groups of observations. We can utilize this difference to determine if observations belong to the same group. All data points within a single group should cluster around their central value. We used two methods to determine what would be optimal K value. First, we used The Elbow method which depends on a calculated value called inertia, which is the sum of the squared distances between each point and its closest K-means center. If K is 1, then the inertia will equal the sum of all squared distances to the dataset's mean.

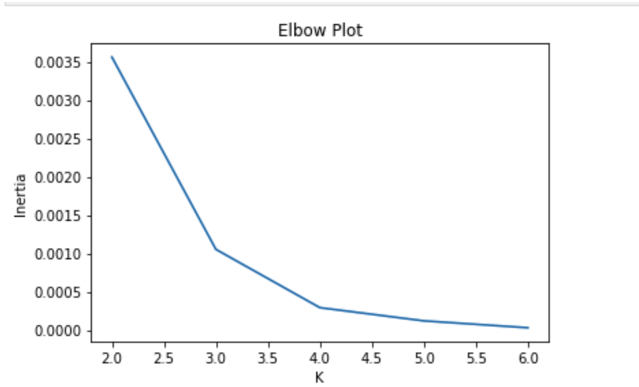

Figure 6. Elbow Plot

Second, we used the Silhouette score which captures the distance of each point to neighboring clusters. It can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of -1, 1. A higher Silhouette score would indicate optimal K value. We concluded that 3 is an optimal number for K value.

Results

After applying the K-Means clustering algorithm to our dataset, we had each neighborhood assigned to one of the 3 clusters. We used the Folium map to visualize clusters on map of Manhattan(Figure 7.)

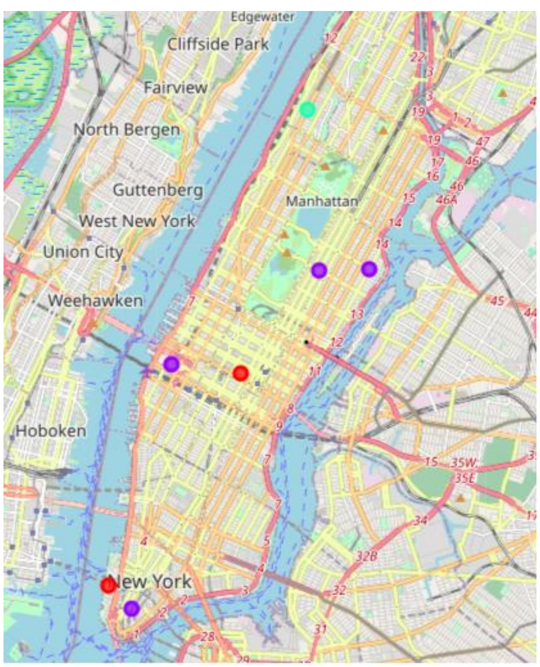

Figure 7. Visualizing clusters at map

Next, we carefully examined each cluster and its neighborhood venues.

Next, we will closely look into each cluster

First cluster

```python
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 0, manhattan_merged.columns[[2] + list(range(5, manhattan_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Midtown	0	Hotel	Clothing Store	Coffee Shop	Steakhouse	Theater	American Restaurant	Bakery	Bookstore	Gym	Sporting Goods Shop
4	Battery Park City	0	Park	Coffee Shop	Hotel	Clothing Store	Gym	Boat or Ferry	Memorial Site	Plaza	Food Court	Burger Joint

From the first cluster we can see that neighborhoods Midtown and Battery Park City are clustered together. Looking closer into venues we can see that both of them have a Gym listed in top 10 venues.

Second cluster

```python
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 1, manhattan_merged.columns[[2] + list(range(5, manhattan_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Upper East Side	1	Coffee Shop	Italian Restaurant	Exhibit	Bakery	Gym / Fitness Center	American Restaurant	Spa	French Restaurant	Hotel	Juice Bar
1	Yorkville	1	Italian Restaurant	Coffee Shop	Gym	Bar	Deli / Bodega	Sushi Restaurant	Japanese Restaurant	Diner	Wine Shop	Pharmacy
5	Financial District	1	Coffee Shop	Bar	American Restaurant	Café	Gym / Fitness Center	Italian Restaurant	Gym	Cocktail Bar	Pizza Place	Sandwich Place
6	Hudson Yards	1	Gym / Fitness Center	American Restaurant	Italian Restaurant	Hotel	Café	Gym	Dog Run	Park	Restaurant	Coffee Shop

Neighborhoods Upper East Side,Yorkville,Financial District and Hudson Yards were clustered together in second cluster. We also notice that all of them contains Gym in top 10 venues.

Third cluster

```python
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 2, manhattan_merged.columns[[2] + list(range(5, manhattan_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Morningside Heights	2	Coffee Shop	Park	Bookstore	American Restaurant	Café	Burger Joint	Arts & Crafts Store	New American Restaurant	Sandwich Place	Salad Place

Figure 8. Clustered Venues

- From the first cluster we could see that neighborhoods Midtown and Battery Park City are clustered together. Looking closer into the venues we can see that both of them have a Gym listed in top 10 venues.
- Neighborhoods *Upper East*, *Yorkville*, *Financial District* and *Hudson Yards* were clustered together in the second cluster. We also notice that all of them contain a Gym within the top 10 venues.
- In the third cluster we have only one Neighborhood: *Morningside Heights*. Taking a closer look into the venues we can see that this is the only cluster that does not have a Gym in the 10 most common venues.

Conclusion

In this project, we looked to find an optimal location to open an exclusive gym. To answer this question for our client, we had to take a few steps. Our approach was to evaluate Manhattan neighborhoods based on income, crime rate, and nearby competition. Once we had neighborhoods with high income and low crime rate, we used the Foursquare API to get most common venues in those areas. We applied K-Means to cluster neighborhoods and used the Folium map to visualize them on a map of Manhattan. Analyzing each cluster, we concluded that the third cluster, containing the neighborhood *Morningside Heights*, would be an optimal location to open a gym.