Capstone Project

The Battle of Neighborhoods

Finding the Best Neighborhood in New York

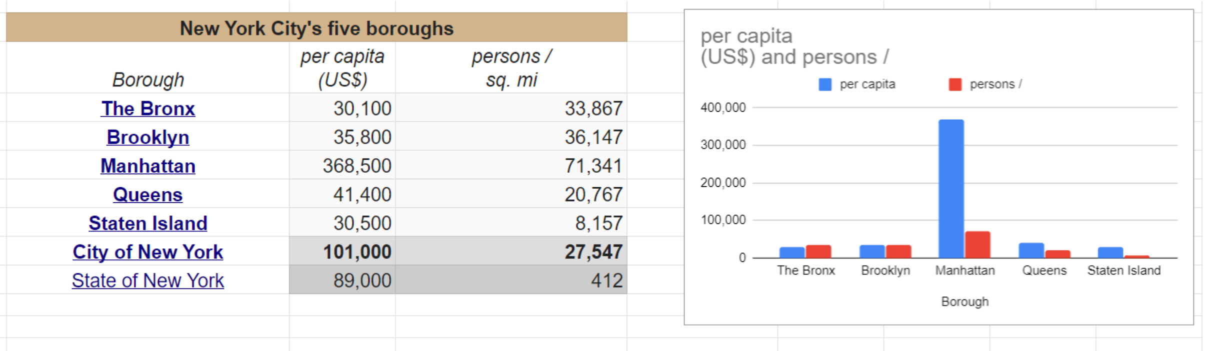
Introduction:

Recently, one of my client has approached me to find out the most suitable place in New York to live.

He wants to move to the most happening place in New York.

Business Problem:

According to Wikipidea – “New York City (NYC), often called New York (NY), is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined statistical area, it is one of the world's most populous megacities.



New York City's five boroughs

Borough	per capita (US$)	persons / sq. mi
The Bronx	30,100	33,867
Brooklyn	35,800	36,147
Manhattan	368,500	71,341
Queens	41,400	20,767
Staten Island	30,500	8,157
City of New York	**101,000**	**27,547**
State of New York	89,000	412

New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. Home to the headquarters of the United Nations, New York is an important center for international diplomacy.

Long Island Sound
Wave Hill
Van Cortland House Museum
Woodlawn Cemetery
New York Botanical Garden
City Island
NASSAU
The Cloisters
Bronx Zoo
BRONX
George Washington Bridge
Morris Jumel Mansion House
Yankee
Audubon Terrace
Hudson River
East River
Little Neck Bay
Shea Stadium
Flushing Meadows Corona Park
New York Hall of Science
American Museum of the Moving Image
QUEENS
MANHATTAN
New Jersey
Brooklyn Academy of Music
Brooklyn Children's Museum
Grand Army Plaza
Ellis Island
Brooklyn Museum
Botanic Garden
Statue of Liberty
Park Slope Historic District
Prospect Park
Jamaica Bay
Jamaica Bay Wildlife Refuge Center
Upper New York Bay
BROOKLYN
Newark Bay
Snug Harbour
Alice Austen House
STATEN ISLAND
Lower New York Bay
Coney Island
New York Aquarium
Jacques Marchais Center of Tibetan Art
Richmond Town
Atlantic Ocean
* of particular interest
** of outstanding interest
Sights outside city centre
N
3 mi
5 km
© Baedeker
www.PlanetWare.com

Given that there are many neighbourhoods in New York, I will aim at exploring the neighbouhoods and using the Foursquare API to analyse the most sought after, trendy, popular and commented venues. I will use web scrapping techniques to take out the location and neighbourhood data of New York and then use the various Machine Learning tool to find out the most suitable place for our client.

Data Collection:

Data is sourced from Wikipedia.

For this problem, we will get the services of Foursquare API to explore the data of two cities, in terms of their neighborhoods. The data also include the information about the places around each neighborhood like restaurants, hotels, coffee shops, parks, theaters, art galleries, museums and many more.

Data Preparation:

Suitable Web scraping tools are applied to do the same including Beautiful Soup.

Using Foursquare API, a large database of nearby venues for each neighborhood in JSON format is sourced for a radius of 5 km radius.

Methodology:

First of all, the Foursquare API was utilized because basic geographical location information and particular venues can be explored through Foursquare API. With a specific credentials client ID and secret, the required foursquare data can be accessed.

Second, Folium is a great visualization library, the map can directly show us the overall shape of the place we want to know. The detailed information on different venues such as restaurants can be seen in a direct way. The investors can have a better understanding of the city of New York to choose his properties. People feel free to zoom into the above map, and click on each circle mark to reveal the name of the neighborhood and its respective borough.

We will try to use the famed K Means Clustering to analyse clusters inside the NY city and then rate it so as to zero in on the best place to live in New York. All 4 clusters used with a K=4 is listed in the project.

According to Data Science Central:

With clustering methods, we get into the category of unsupervised ML because their goal is to group or cluster observations that have similar characteristics. Clustering methods don't use output information for training, but instead let the algorithm define the output. In clustering methods, we can only use visualizations to inspect the quality of the solution.

The most popular clustering method is K-Means, where "K" represents the number of clusters that the user chooses to create. (Note that there are various techniques for choosing the value of K, such as the elbow method.) {DataScience Central}

Roughly, what K-Means does with the data points:

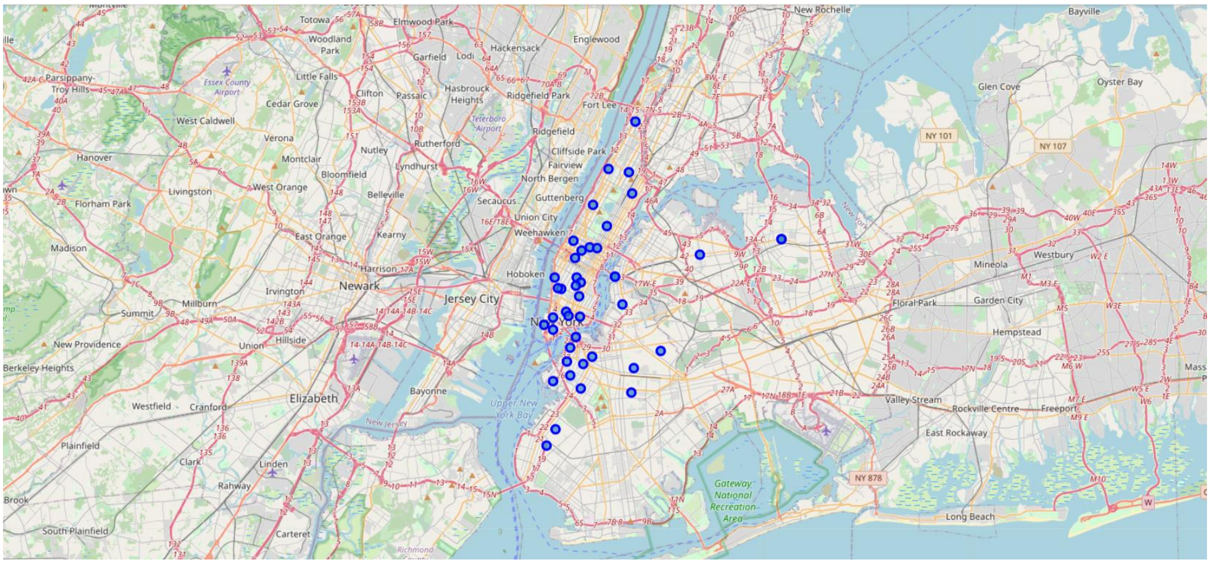
Randomly chooses K centers within the data.

Assigns each data point to the closest of the randomly created centers.

Re-computes the center of each cluster.

If centers don't change (or change very little), the process is finished. Otherwise, we return to step 2. (To prevent ending up in an infinite loop if the centers continue to change, set a maximum number of iterations in advance.)

The next plot applies K-Means to a data set of buildings. Each column in the plot indicates the efficiency for each building. The four measurements are related to air conditioning, plugged-in equipment (microwaves, refrigerators, etc…), domestic gas, and heating gas. We chose K=2 for clustering, which makes it easy to interpret one of the clusters as the group of efficient buildings and the other cluster as the group of inefficient buildings. To the left you see the location of the buildings and to right you see two of the four dimensions we used as inputs: plugged-in equipment and heating gas.



Last but not least, basic numpy and pandas libraries are fundamental methods because dataframes are the basic tool to do any calculations or simulations. Also, some basic python languages are better utilized.

Conclusion:

After doing the K means clustering and other analysis described in the python notebook we are of the Opinion that the person should reside in cluster 1. This is because the person wants to live in a happening place nearby. Cluster one has all the facilities the client is looking for.