Coursera Applied Data Science Capstone Project

Eoghan Chelmiah

eoghan.chelmiah.research@gmail.com

Introduction

New York City is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over 784 square kilometres, New York is also the most densely populated major city in the United States.

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 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.

New York City is composed of five boroughs, each of which is a county of the State of New York. The five boroughs; Brooklyn, Queens, Manhattan, the Bronx, and Staten Island were consolidated into a single city in 1898.

New York City is the world's undisputed sports champion, with professional teams playing in every type of game you could want to watch. It's the place where moments like the famous "Shot Heard 'Round the World" and Willis Reed's limping out onto the court in the NBA Finals were etched into history, and where the world's best athletes compete in marquee annual events like the US Open Tennis Championships and the New York City Marathon.

Business Problem

This objective of this project is to identify the best locations in New York City to open a gym of fitness studio facility.
Location data will be analysed using a combination of machine learning and data science techniques, the optimal location for a property developer to build or establish a new gym facility will be determined.

Target Audience

The target audience of this project are property developers and leisure centre managers who plan on opening a new gym or fitness studio in New York City.
Determining the optimal location for the gym will help in obtaining valuable investors for such a project.
As New York City has such a high reputation in the domain of sports and fitness, finding the optimal location is critical for being able to compete with the existing establishments in this field.

Data

To begin this project, data will need to be acquired in order to study the geographical location, scan the area for existing facilities, and segment the area into zones to analyse further. Once the area has been studied, locations can be ranked into an order which illustrates and highlights the optimal locations, based on the existing facilities in each of the neighbourhoods in New York City.

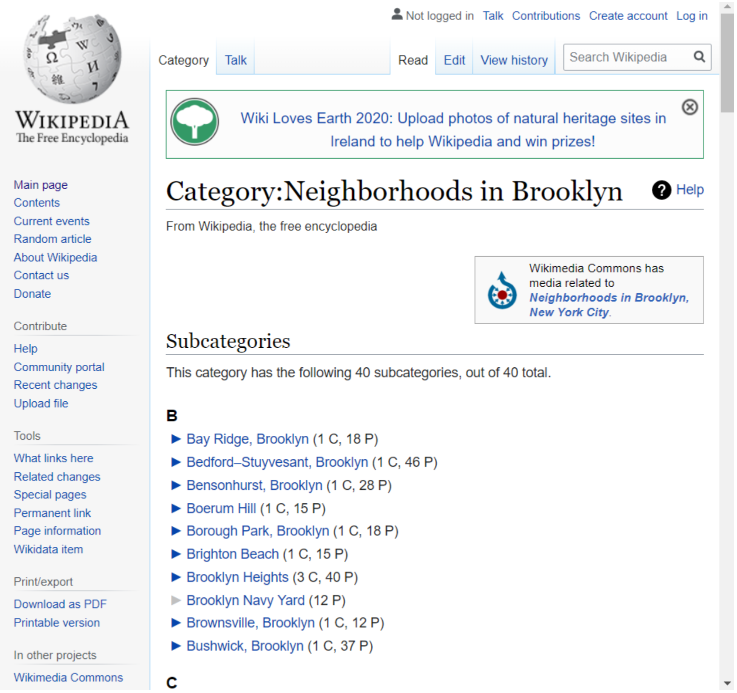
The data required includes the following:

Neighbourhoods

A list of neighborhoods in New York City can be found at;
https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Brooklyn

Brooklyn is one of the largest of the central New York districts and home to the majority of the state's population. The names of over 40 neighbourhoods are listed on the Wiki page.

Data mining techniques to extract this list of locations from the webpage are performed using the computer programming language Python. A support library called 'beautifulsoup' can be used to scrape the text and convert to a dataframe in the Python environment which can then be filtered, reordered and cleaned.



Wiki Loves Earth 2020: Upload photos of natural heritage sites in Ireland to help Wikipedia and win prizes!

Category:Neighborhoods in Brooklyn

Help

From Wikipedia, the free encyclopedia

Wikimedia Commons has media related to *Neighborhoods in Brooklyn, New York City*.

Subcategories

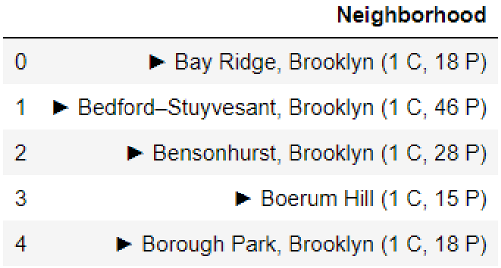
This category has the following 40 subcategories, out of 40 total.

B

- ▶ Bay Ridge, Brooklyn (1 C, 18 P)
- ▶ Bedford–Stuyvesant, Brooklyn (1 C, 46 P)
- ▶ Bensonhurst, Brooklyn (1 C, 28 P)
- ▶ Boerum Hill (1 C, 15 P)
- ▶ Borough Park, Brooklyn (1 C, 18 P)
- ▶ Brighton Beach (1 C, 15 P)
- ▶ Brooklyn Heights (3 C, 40 P)
- ▷ Brooklyn Navy Yard (12 P)
- ▶ Brownsville, Brooklyn (1 C, 12 P)
- ▶ Bushwick, Brooklyn (1 C, 37 P)

C

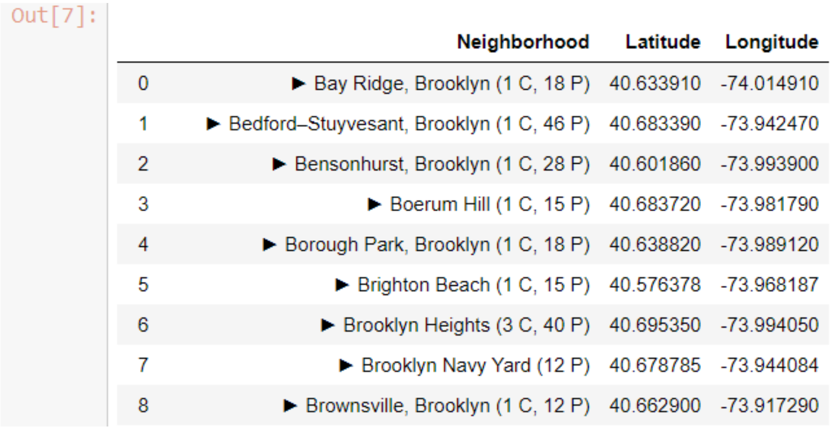
Out[2]:



	Neighborhood
0	▶ Bay Ridge, Brooklyn (1 C, 18 P)
1	▶ Bedford–Stuyvesant, Brooklyn (1 C, 46 P)
2	▶ Bensonhurst, Brooklyn (1 C, 28 P)
3	▶ Boerum Hill (1 C, 15 P)
4	▶ Borough Park, Brooklyn (1 C, 18 P)

Geocoding

The coordinated details of each neighbourhood consisting of the longitude and latitude values was acquired using the ‘geocoder’ Python package.
These coordinates were then attached to their corresponding locations which were scraped from the Wiki page in the previous step.

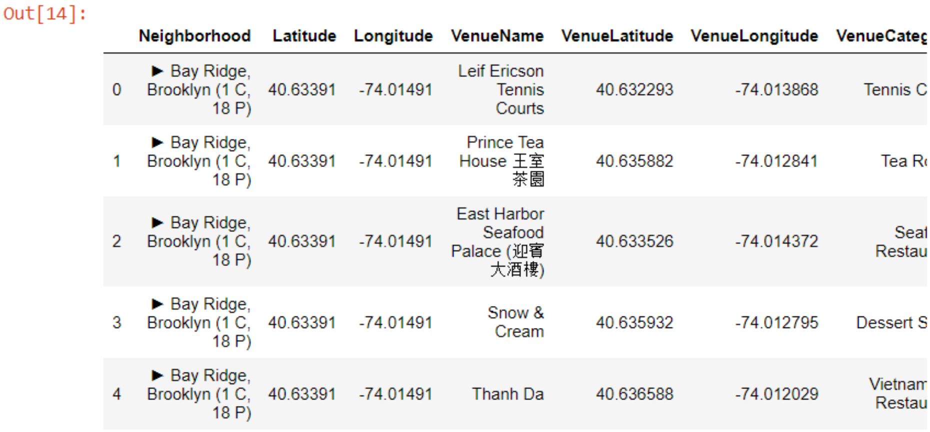


Out[7]:

	Neighborhood	Latitude	Longitude
0	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.633910	-74.014910
1	▶ Bedford–Stuyvesant, Brooklyn (1 C, 46 P)	40.683390	-73.942470
2	▶ Bensonhurst, Brooklyn (1 C, 28 P)	40.601860	-73.993900
3	▶ Boerum Hill (1 C, 15 P)	40.683720	-73.981790
4	▶ Borough Park, Brooklyn (1 C, 18 P)	40.638820	-73.989120
5	▶ Brighton Beach (1 C, 15 P)	40.576378	-73.968187
6	▶ Brooklyn Heights (3 C, 40 P)	40.695350	-73.994050
7	▶ Brooklyn Navy Yard (12 P)	40.678785	-73.944084
8	▶ Brownsville, Brooklyn (1 C, 12 P)	40.662900	-73.917290

Venue Data

Acquiring data about venues currently present in the surrounding area of New York City contributes to a further understanding of whether the neighbourhood could foster the development of a new gym or fitness studio.
Foursquare is used to analyse the venue data for all the listed neighbourhoods.
Once details regarding the venues existing are obtained, clustering can be done using the K-Means algorithm in order to determine the optimal location to place a new gym or fitness studio.



Out[14]:

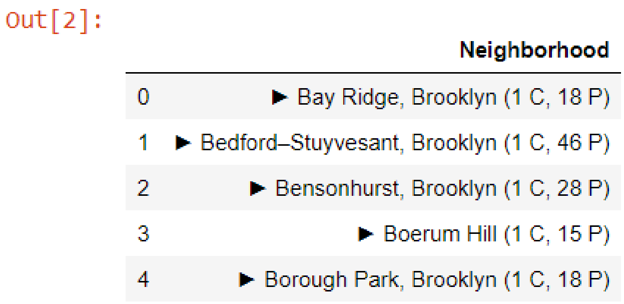
	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCateg
0	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Leif Ericson Tennis Courts	40.632293	-74.013868	Tennis C
1	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Prince Tea House 王室茶園	40.635882	-74.012841	Tea R
2	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	East Harbor Seafood Palace (迎賓大酒樓)	40.633526	-74.014372	Seaf Restau
3	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Snow & Cream	40.635932	-74.012795	Dessert S
4	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Thanh Da	40.636588	-74.012029	Vietnam Restau

Methodology

A list of neighbourhoods in New York City from the Wiki page;
https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Brooklyn
was used to gather location names.

Brooklyn is one of the largest of the central New York districts and home to the majority of the state's population. The names of over 40 neighbourhoods are listed on the Wiki page.

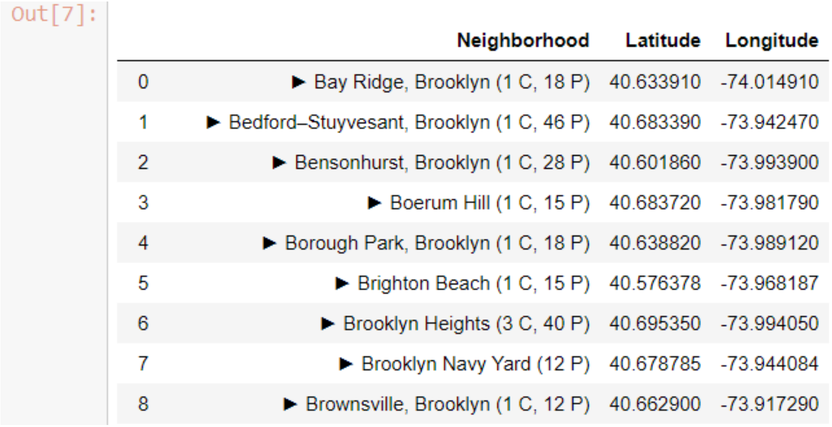
Data mining techniques to extract this list of locations from the webpage are performed using the computer programming language Python. A support library called 'beautifulsoup' can be used to scrape the text and convert to a data frame in the Python environment using the Pandas library. This data frame was then filtered, reordered and cleaned.



Out[2]:

	Neighborhood
0	▶ Bay Ridge, Brooklyn (1 C, 18 P)
1	▶ Bedford–Stuyvesant, Brooklyn (1 C, 46 P)
2	▶ Bensonhurst, Brooklyn (1 C, 28 P)
3	▶ Boerum Hill (1 C, 15 P)
4	▶ Borough Park, Brooklyn (1 C, 18 P)

The geographical coordinates of each neighbourhood were obtained using the 'Geocoder' Python package. The longitude and latitude values of each of the neighbourhoods in the data frame were collected and matched to their corresponding locations.

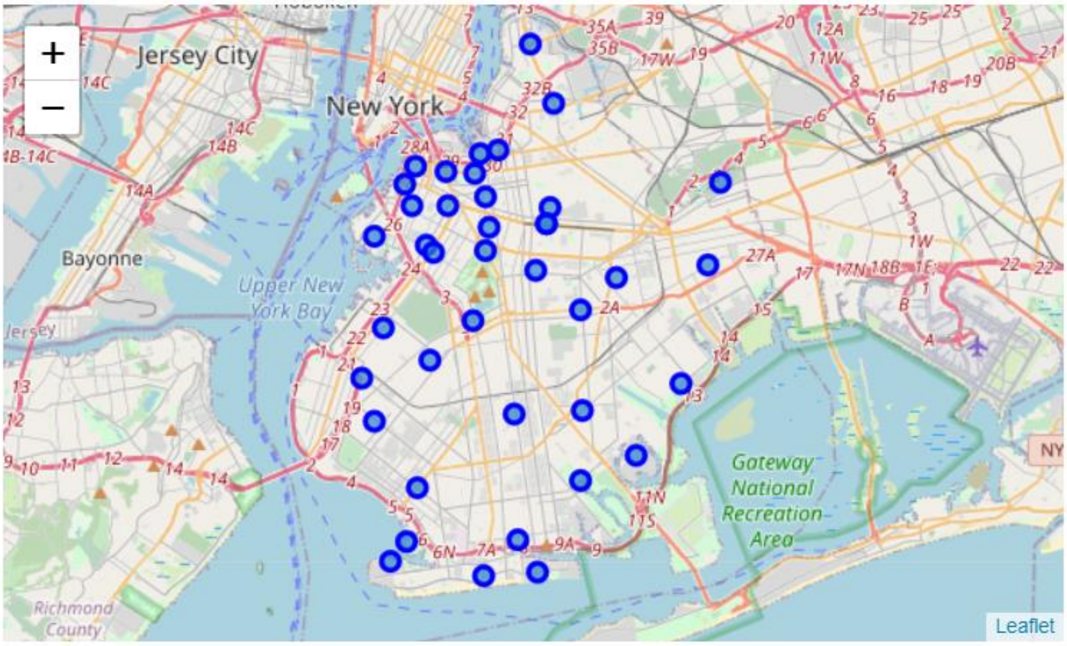


Out[7]:

	Neighborhood	Latitude	Longitude
0	▶ Bay Ridge, Brooklyn (1 C, 18 P)	40.633910	-74.014910
1	▶ Bedford–Stuyvesant, Brooklyn (1 C, 46 P)	40.683390	-73.942470
2	▶ Bensonhurst, Brooklyn (1 C, 28 P)	40.601860	-73.993900
3	▶ Boerum Hill (1 C, 15 P)	40.683720	-73.981790
4	▶ Borough Park, Brooklyn (1 C, 18 P)	40.638820	-73.989120
5	▶ Brighton Beach (1 C, 15 P)	40.576378	-73.968187
6	▶ Brooklyn Heights (3 C, 40 P)	40.695350	-73.994050
7	▶ Brooklyn Navy Yard (12 P)	40.678785	-73.944084
8	▶ Brownsville, Brooklyn (1 C, 12 P)	40.662900	-73.917290

The location names and coordinates were then used to create a map visualization of New York City, using the Folium Python package.

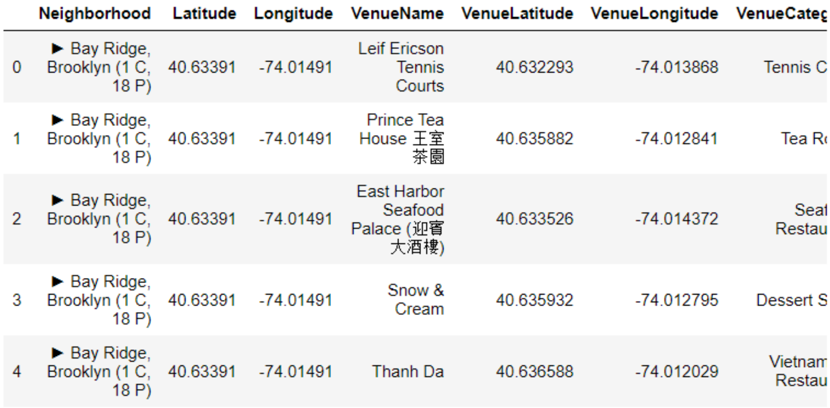
Out[10]:



Foursquare is used to analyse the venue data for all the listed neighbourhoods. The top 100 venues that are situated within a 2000 mile radius were determined. A Foursquare Developer account was setup and used to make API calls via a Client ID and Secret Key.

Looping through the coordinates in the data frame, Foursquare returned surrounding venue data in the form of a JSON file. This file was analysed to return venues isolated in specific locations from all of the New York City neighbourhoods.

Out[14]:

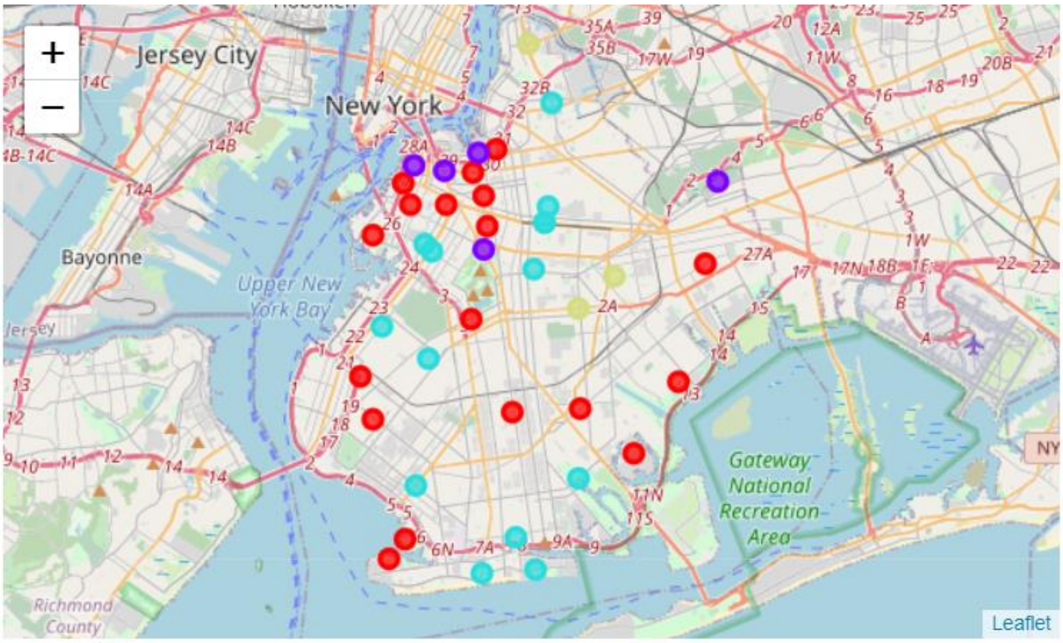


	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCateg
0	► Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Leif Ericson Tennis Courts	40.632293	-74.013868	Tennis C
1	► Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Prince Tea House 王室茶園	40.635882	-74.012841	Tea R
2	► Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	East Harbor Seafood Palace (迎賓大酒樓)	40.633526	-74.014372	Seaf Restau
3	► Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Snow & Cream	40.635932	-74.012795	Dessert S
4	► Bay Ridge, Brooklyn (1 C, 18 P)	40.63391	-74.01491	Thanh Da	40.636588	-74.012029	Vietnam Restau

The final steps were to filter the locations to display only the places of particular interest to this project. The locations of gyms and fitness suites were selected and extracted. These locations were filtered using the K-means algorithm, to identify the relative concentrations of gyms in each area.

4 clusters were identified which indicate the most suitable locations for establishing a new gym or fitness studio in New York City.

```
Out[27]:
```



Results

The results from the K-means clustering give a visual indication of the following patterns in the relative concentrations of gyms and fitness studios within the 4 clusters:

Clusters 0 and 1 had little or no gyms
Cluster 2 had a high concentration of gyms
Cluster 3 had a moderate concentration of gyms

On the map above, cluster 0 is marked by the red, cluster 1 by the purple, cluster 2 by the yellow and cluster 3 by the turquoise marker.

Discussion

As illustrated by the cluster locations in the visualization above, most of the gyms are concentrated in the central neighbourhoods with the highest cluster in 2.
A moderate concentration of gyms and fitness suites are depicted by the green marker on the coastal neighbourhoods.
There are very little gyms in the north western neighbourhoods, making these locations prime for the development of a new gym.
This is beneficial as the new gym will be facing relatively less competition with other gyms and fitness studios.

Conclusion

The purpose of this project was to analyse and identify the optimal locations to place a new gym or fitness studio in New York City.
Through data mining and machine learning methods in the Python computer programming language, we were able to segment Brooklyn into clusters of neighbourhoods and analyse these 4 clusters to determine which could foster the most optimal conditions for a new gym.

To definitively answer the original business question, the optimal location to build or establish a new gym or fitness suite would be the central neighbourhoods of Brooklyn where there are little to no competition with other fitness institutes.