

Coursera Applied Data Science Capstone Project

Week 4 Assignment

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 or 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.



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	Seal 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