

# BATTLE OF THE NEIGHBORHOODS

**NEW YORK CITY  
FY 2020**

This report discusses the challenges of finding a good neighborhood to move to and start a business considering living expenses, availability of venues that are preferred, etc., and offers a solution by using modern Data Science Tools.



# TABLE OF CONTENTS

## Contents

Introduction	1
Data	2
Methodology	4
Results	7
Discussion	9
Conclusion	10

# BATTLE OF THE NEIGHBORHOODS

## Introduction

### BUSINESS PROBLEM

I want to start living in New York City (NYC), but I don't know what neighborhood to live in. I love to eat Chinese food so I want to live in a borough that has a high number of Chinese food restaurants. But I want my neighborhood to be small, but have a few number of Chinese food restaurants. Eventually, if the neighborhood is well-suited, I may consider opening my own restaurant – I want to open mine in a community of similar minded eaters, (If you want good Chinese food – go to China Town!) but not many options right by where I live.

### TARGET AUDIENCE / STAKEHOLDERS

This analysis is a good project for data scientists seeking to explore available open-source tools. It is also a useful tool for someone who wants to move to NYC! The simple problem statement above is actually quite complex to resolve, but easily done with the correct tools and code

### BACKGROUND

NYC is exciting and interesting and I like tall buildings. I enjoy the cultural diversity. I like to eat a lot of food, so I want there to be lots options to eat near-by. Specifically, I like to eat Chinese food.

Unfortunately (or fortunately – depending on your perspective) NYC is a gigantic city with many neighborhoods and a significant number of venues to choose from. This project would be nearly impossible to do by flipping through a phone book, or trying to search manually though google.

This project seeks to utilize the advantages of modern-day data science tools to solve the problem discussed above.

# BATTLE OF THE NEIGHBORHOODS

## Data

### DATA SETS

The analysis begins with an examination of various publically available data sets. It is critical that the data science solution be able to utilize up-to-date sources of data, so the web and services such as Four Square are heavily leveraged.

1. Neighborhoods: This data, including neighborhood name, borough, and location, is extracted from websites by using the Python extension called *Beautiful Soup*. Its parses html data and makes it available to insert into a pandas dataframe. For this project I used Wikipedia

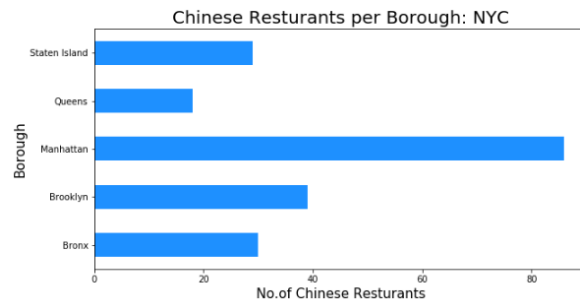


2. Neighborhoods: The data can also be obtained from a publically available csv file at this location - [https://cocl.us/new\\_york\\_dataset](https://cocl.us/new_york_dataset)

	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

3. Specific Venue Data: This is obtained through use of the Four Square API. Four Square provides continuously updated, user-supplied information regarding venues. All data is geo-tagged so that it can be compared to other locations with relative ease.

# BATTLE OF THE NEIGHBORHOODS



4. Location Data: Location data about almost any area can be obtained using the Geopy python extension. In this project it is used to find the latitude and longitude of New York City and some neighborhoods within it, but can be used to find any similar location desired.

```
: # Use geopy library to get the Latitude and Longitude values of New York City.  
address = 'New York City, NY'  
  
geolocator = Nominatim(user_agent="ny_explorer")  
location = geolocator.geocode(address)  
latitude = location.latitude  
longitude = location.longitude  
print('The geograppical coordinate of New York City are {}, {}'.format(latitude, longitude))
```

The geograppical coordinate of New York City are 40.7127281, -74.0060152.

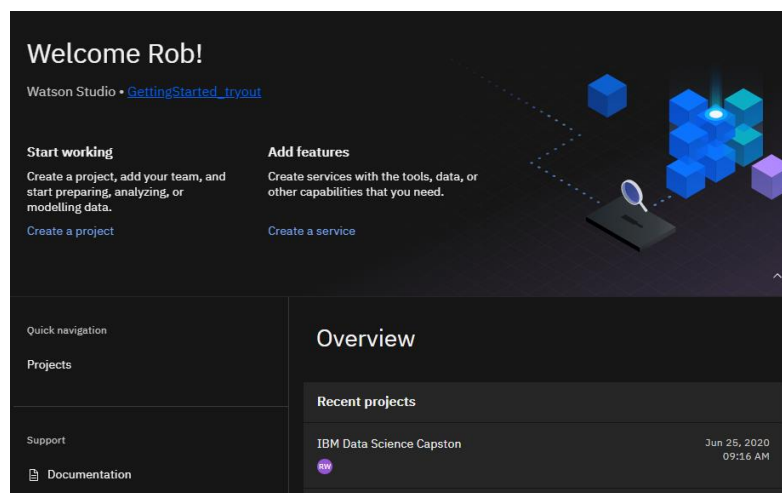
Latitude and longitude values of Marble Hill are 40.87655077879964, -73.91065965862981.

# BATTLE OF THE NEIGHBORHOODS

## Methodology

### SOFTWARE USED

- Anaconda – Simple tool that manages notebooks and maintains the python environment
- Python – Open-source programming language designed for data science and analysis
- Jupyter Notebook – Open-source web application allowing creation and sharing of documents that contain live code and markdown text
- GitHub – Free web-based repository of files and codes. Partial snippets of code from multiple contributors informed the final code.
- Microsoft Word – Create the report
- Microsoft Powerpoint – Create the presentation
- Adobe – Saves files as shareable and flattened pdf's
- IBM Watson – Free cloud-based services that includes Python and Notebooks



### PYTHON LIBRARIES USED.

- Pandas – Puts data into dataframes
- Numpy – For solving matrix based problems, including clustering and one-hot coding
- Geopy – Geographic location tool
- Json – Reads json formatted files that are provided by Four Square API
- BeautifulSoup – Scrapes html into dataframes
- Matplotlib – Plotting tool
- Sklearn – Clustering tool, introductory machine learning.

# BATTLE OF THE NEIGHBORHOODS

- Folium – Mapping tool

All libraries were either native to Anaconda install or Watson, or imported from the Conda Forge when necessary. Source code on this project can be found on github.

```
# Import Libraries already installed in environment

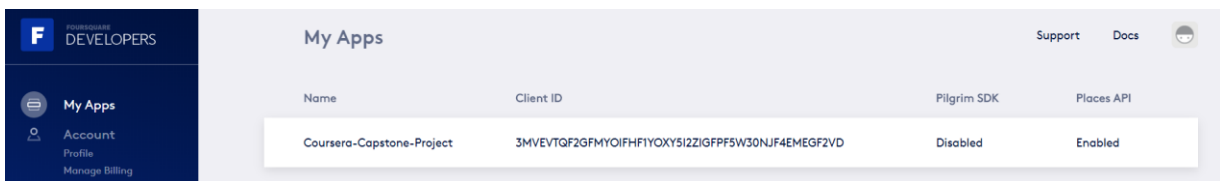
import json
import numpy as np # Library to handle data in a vectorized manner
import pandas as pd # Library for data analysis
import folium
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
import requests # Library to handle requests
from pandas.io.json import json_normalize # transform JSON file into a pandas dataframe
from sklearn.cluster import KMeans
import matplotlib.cm as cm
import matplotlib.colors as colors
```

## DATA COLLECTION

- For this project the New York neighborhood, borough, and other information was obtained from a IBM-provided dataset. It is located at [https://cocl.us/new\\_york\\_dataset](https://cocl.us/new_york_dataset), or [https://geo.nyu.edu/catalog/nyu\\_2451\\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572)
- Additional information can be gleaned from sources online. Wikipedia is a fantastic example. The website [https://en.wikipedia.org/wiki/Demographics\\_of\\_New\\_York\\_City](https://en.wikipedia.org/wiki/Demographics_of_New_York_City) has been used as a source during the execution of this project.
- This data is reported as a JSON. The next step is to transfer to a Pandas dataframe for ease of future analysis.

## APPROACH

- The data set was collected as described above.
- Foursquare API was utilized to determine venue information in the area surrounding the neighborhoods we examined as potential candidates for both living in and starting a business.
- Venue data is also returned as a JSON. It must be transferred into a Pandas dataframe for analysis.



Name	Client ID	Pilgrim SDK	Places API
Coursera-Capstone-Project	3MVEVTGQF2GFMYOIFHF1YQXY5I2ZIGFPF5W30NJF4EMEGF2VD	Disabled	Enabled

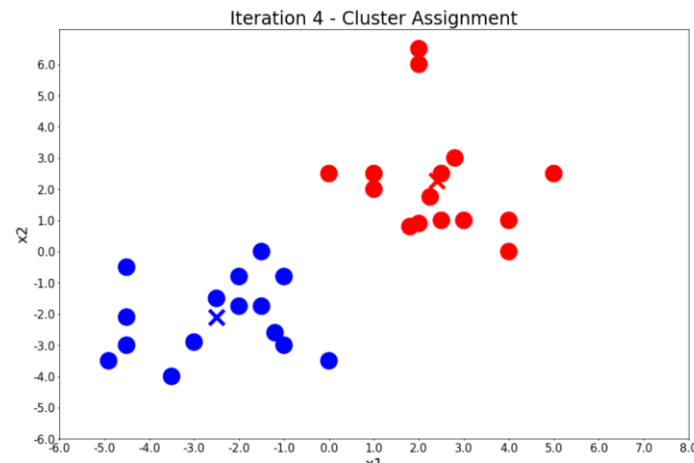
## DATA SORTED

- The data was sorted according to distance and other parameters to facilitate decision-making.

# BATTLE OF THE NEIGHBORHOODS

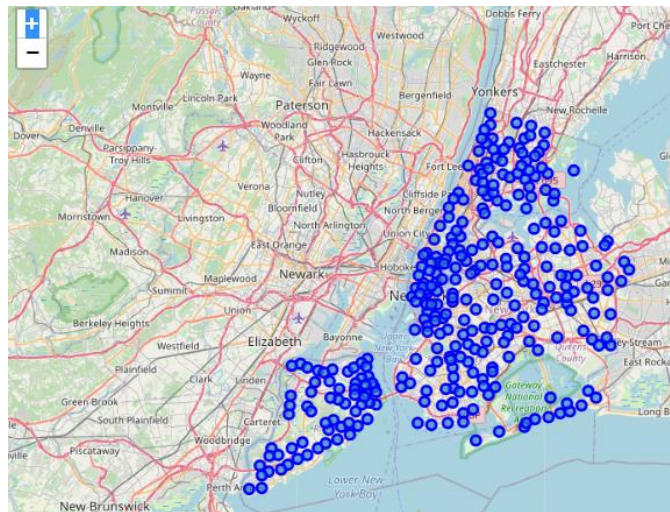
## OPTIONS FOR CLUSTERING

- Future iterations of this project could include options for clustering the data for deeper analysis.
- This would take advantage of Scikitlearn and the k-means algorithms.
- Below is an example of how clustering works.



## VISUALIZATIONS CREATED

- Matplotlib was used to create simple visualizations such as bar charts to help analyze the data. Other simple plots such as xy, line, pie, etc are also available, but not utilized in this project.
- Folium is a library that uses geodata to plot on a map. One of the benefits of Folium is that it has a strong user-interface, allowing the analyst to look deep in the data, wherever they desire. An example of folium is below.





# BATTLE OF THE NEIGHBORHOODS

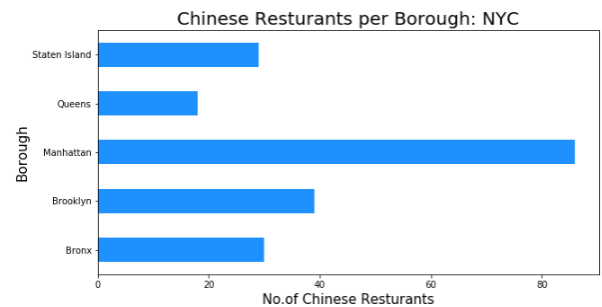
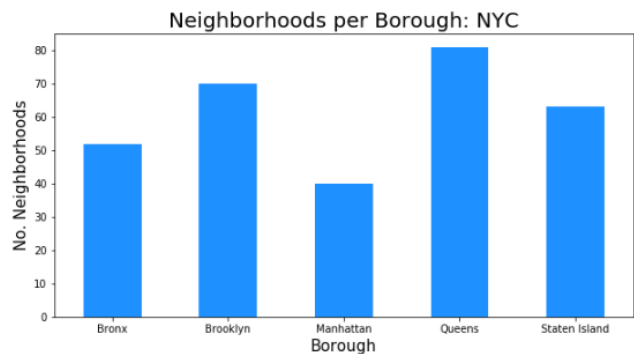
## Results

### WHAT IS THE BEST BOROUGH?

I chose to live in Queens, NY because it has a sufficient number of Chinese restaurants and has myriad other benefits. (<https://www.buzzfeed.com/perpetua/queens-is-great>). Queens also has the highest number of neighborhoods of all the boroughs, ensuring plenty of future options.



Image from ivan\_ward instagram



### WHAT NEIGHBORHOOD IS BEST?

There are not very many Chinese restaurants in the neighborhood of Astoria, so while the borough is good for options, the neighborhood will be good in the future for business opportunities.

	name	categories	lat	lng
0	Favela Grill	Brazilian Restaurant	40.767348	-73.917897
1	Orange Blossom	Gourmet Shop	40.769856	-73.917012
2	Off The Hook	Seafood Restaurant	40.767200	-73.918104
3	CrossFit Queens	Gym	40.769404	-73.918977
4	Titan Foods Inc.	Gourmet Shop	40.769198	-73.919253

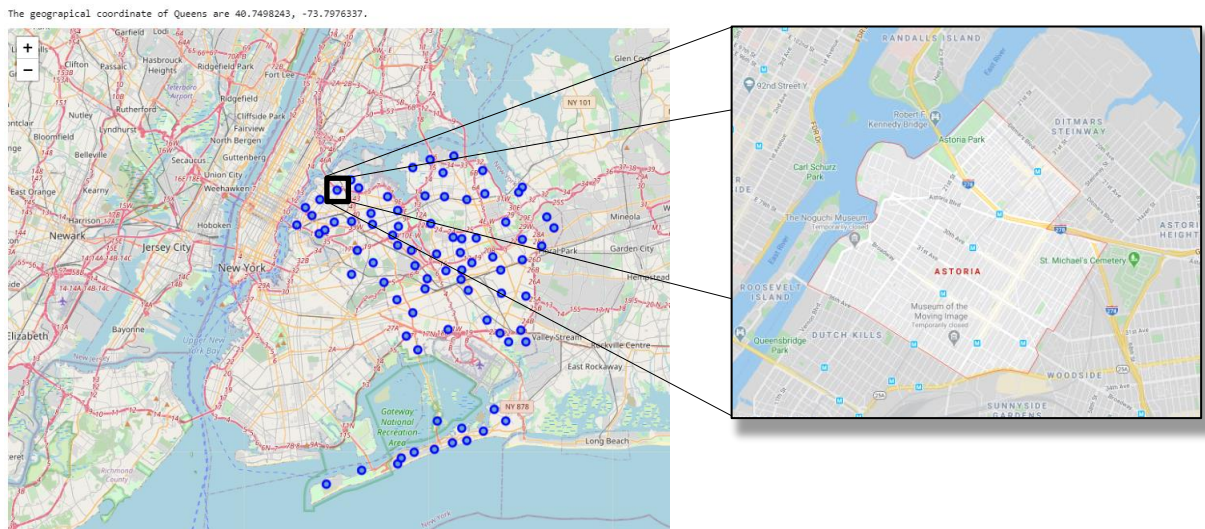


Image from NYCgo.com

# BATTLE OF THE NEIGHBORHOODS

## VISUALIZED USING FOLIUM

Below is a map of all the neighborhoods in Queens. The neighborhood of Astoria is highlighted in red.



## DECISION MADE

- After all of the data was collected, sorted, visualized and analyzed, the choice for best neighborhood to live in and ultimately start a business in, was easy.
- I decided to move to the neighborhood of Astoria.



By GK tramrunner229 - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=1394041>



# BATTLE OF THE NEIGHBORHOODS

## Discussion

Four Square API makes it easy to identify neighborhoods. It is relatively simple to go through many GB of data and sort them, compare them, and analyze it all to determine the solution to your problem.

Recommendations would be to expand the data set to other sources such as Kaggle.com. It would be good to go out to the city and check if the neighborhood just *felt* good too.



<https://www.newyorkfamily.com/top-family-friendly-neighborhoods-across-new-york-city/>

# BATTLE OF THE NEIGHBORHOODS

## Conclusion

I would move to Queens because it has sufficient Chinese restaurants. Paradoxically, I would move to the neighborhood of Astoria because it has the fewest restaurants. I would move to this neighborhood and eventually start up my own business.

This project has clearly demonstrated the utility of modern data science tools. Python and its associated library of open-source environments is a very powerful and easy to use tool. API's such as Four Square, Google Maps, and others, add an entirely different dimension to analysis, exponentially increasing efficiency.

Something as complex as this simple capstone project would have been impossible even a few years ago, without tremendous resources such as money and man-power.

I feel empowered by what I have learned in the IBM Data Science Professional Certificate course on Coursera, and I am looking forward to continuing my education in data science.