

INTRODUCTION:

In this project we will explore data from the city of New York.

New York City is one of the most populous city in United States with a population of 8.39 million in 2020. It is a hub of diverse cultures combining all facets of the globe. NYC is a major industrial center and the financial capital of the world. There are 5 boroughs in the city, and with such a large geographical area there is a huge competition between companies. Thus there is big challenge in figuring out the most ideal spots to open up a new business and maximizing profits.

Business Problem:

The aim is to identify the best place to open a new Chinese food restaurant in the city, and which location would be the most appropriate, while taking into account the competitors and the inhabitants of the different neighborhoods of the city. We'll use data analytic methodologies that we learned from the IBM Coursera course to analyze and visualize relevant information to come to valid a conclusion.

Target Audience:

For our project, the target audience are developers and investors who might be interested in opening up a Chinese restaurant in the city of New York. In the city of NY, because it's such a ethnically diverse hub there are a lot of cuisine categories such as Japanese, Indian, Mexican etc. Chinese restaurants have become very popular through out the Untied States and we wanted to see how it stacked with other cuisines in the state.

Data:

Data 1:

Neighborhood has a total of 5 boroughs and 306 neighborhoods. To segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

This dataset exists for free on the web. Link to the dataset is : https://geo.nyu.edu/catalog/nyu_2451_34572

Data 2: For the below analysis we will get data from wikipedia as given below:

New York Population

New York City Demographics

Cuisine of New York city

https://en.wikipedia.org/wiki/New_York_City

https://en.wikipedia.org/wiki/Economy_of_New_York_City

https://en.wikipedia.org/wiki/Portal:New_York_City

https://en.wikipedia.org/wiki/Cuisine_of_New_York_City

https://en.wikipedia.org/wiki/List_of_Michelin_starred_restaurants_in_New_York_City

Data 3:

New York City geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each neighborhood. We will use the Foursquare API to explore neighborhoods in New York City. The below is image of the Foursquare API data.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896521	-73.844680	Pharmacy
2	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898283	-73.850478	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin Donuts	40.890631	-73.849027	Donut Shop

Methodology:

- 1) Data will be collected from https://cocl.us/new_york_dataset and cleaned and processed into a dataframe.
- 2) FourSquare be used to locate all venues and then filtered by Japanese restaurants. Ratings, tips, and likes by users will be counted and added to the dataframe.
- 3) Data will be sorted based on rankings.
- 4) Finally, the data be will be visually assessed using graphing from Python libraries.

Analysis

Data Identification, Capturing & Cleaning

Search & identify the relevant data source & capture it. Here we are using Wikipedia to get data about New York City. Then we remove all the redundant values (Data Cleaning). Then we combine neighborhoods similar to the Bronx. Now the data is cleaned & ready to use.

Combining Different Data Sources and Sorting Neighborhoods Based on Longitude & latitude

Now, we will combine the neighborhood datasets with the postal address' alongside the dataset with Latitude & Longitude and save them into separate data frames. The resulting data frame will contain details about Postal Codes, Brougths, Neighborhoods, and Latitude & Longitude. We finally then visualize it using the folium map.

Explore the NYC's Neighborhoods

Firstly, we explored all the neighborhoods in the city of New York using the Latitude & Longitude data, using the Foresquare API to get the restaurant venues available in NYC. Then, we explored the unique categories in the neighborhoods by filtering the venue details for all possible "Chinese Restaurants". Next, we found each neighborhood along with the top most common venues. Finally, we identified the top 10 venues for each neighborhood.

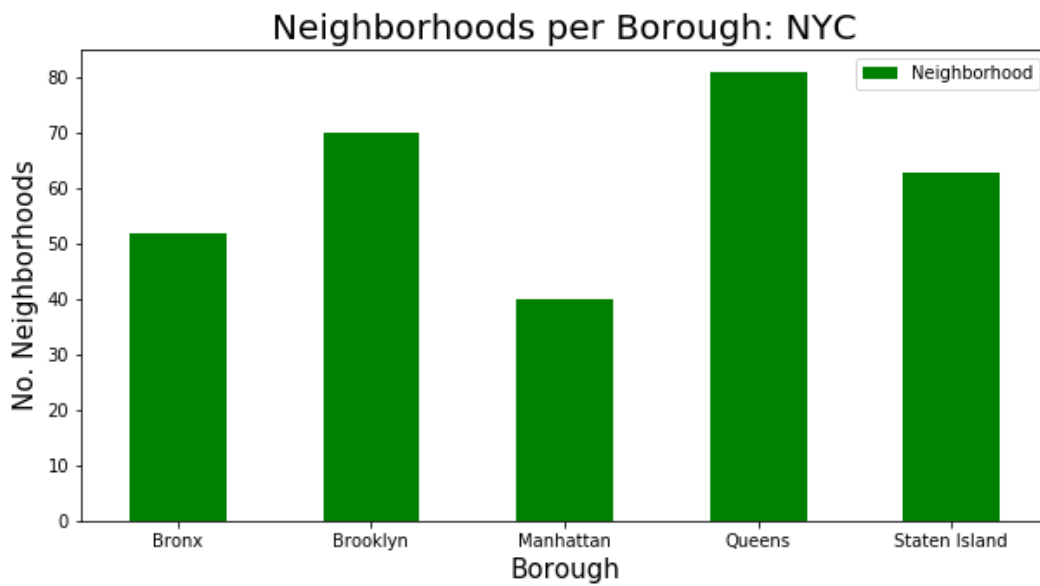
Clustering

With an assumption of five clusters, we use the K-Cluster algorithm to come up with five different clusters in NYC with a similar set of venues. We then explore each cluster & determine the discriminating venue categories that distinguishes each cluster. Identify the clusters & Boroughs/Neighborhoods with maximum number restaurants & their types.

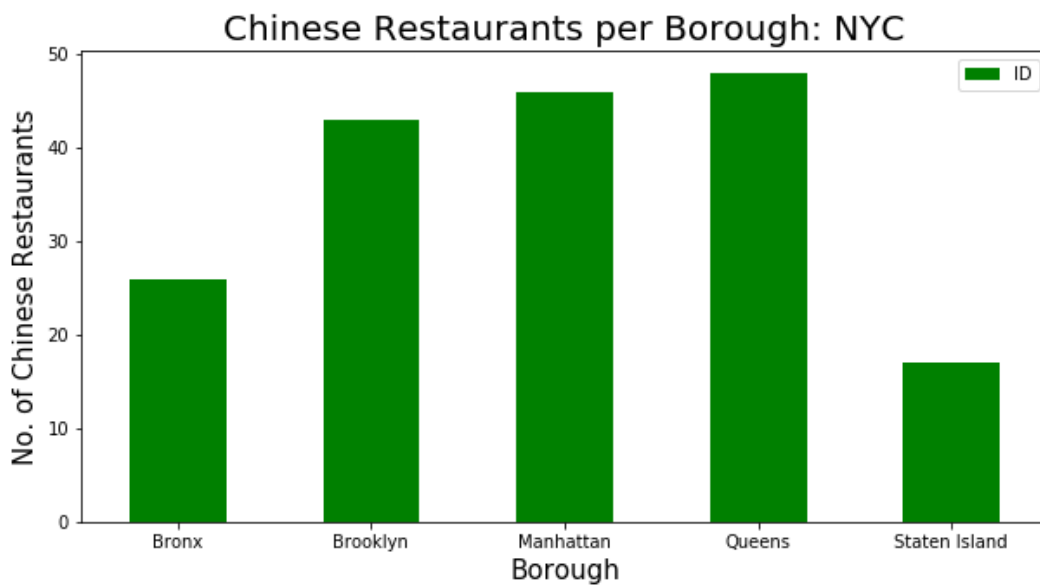
Results and Discussion

Results The results of our analysis showed below:

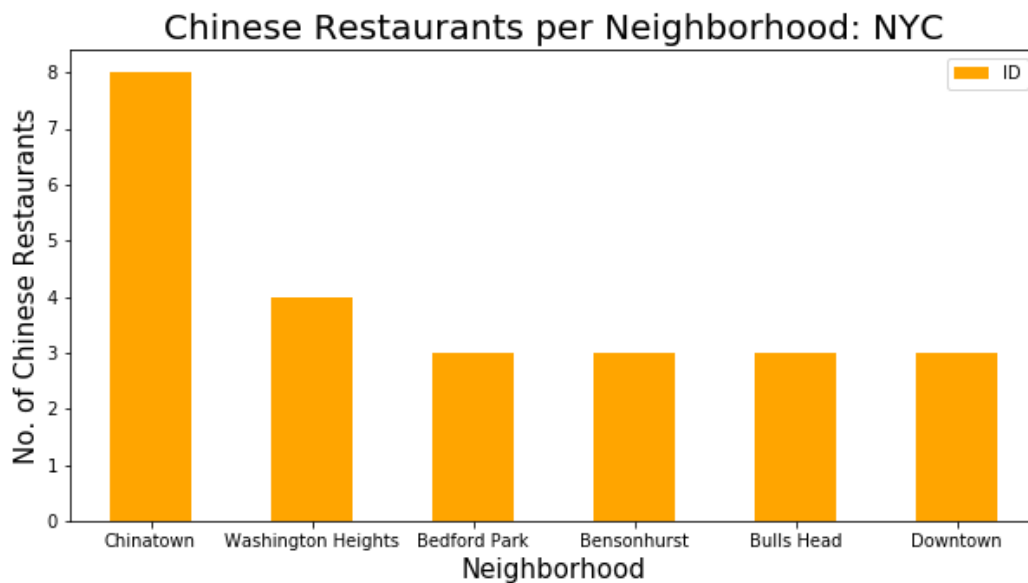
1. We see that Queens has the highest number of Neighborhoods.



2. Chinatown in Manhattan has the highest number of Chinese restaurant



3.) Chinatown in Manhattan has the highest number of Chinese Restaurants



Discussion Section

Based on the results of our analysis, I would state that Manhattan & Brooklyn are the best locations for Chinese cuisines in NYC. To have the best shot of success, I would open a Chinese restaurant in Brooklyn. Brooklyn has multiple neighborhoods with average ratings exceeding 8.0 on a scale of 1.0 to 10.0 and has less amount of Chinese restaurants than Manhattan, making competition easier. In addition, we should keep in mind that real estate prices in Brooklyn are much cheaper than in Manhattan. Particularly, I would recommend considering opening a Chinese Restaurant either in Cobble Hill or in North Side, because both of these neighborhoods have the highest rating for Chinese restaurants.

Limitations & Suggestions for Future Research

All of the above analysis depends on the accuracy of the Four Square data. During this project, we used a free Sandbox Tier Account of Foursquare API that goes with limitations as to the number of API calls & results returned. To get better results, future research work & more comprehensive analysis could consider using a paid account to bypass these limitations as well as incorporating data from other external databases.

Conclusions

In the project we have gone through the process of identifying the business problem, specifying the data required, extracting & preparing the data, performing data analysis, and lastly providing recommendations to the investors/developers. During the project, we applied different data science methods & instruments to get the answer to our main question: "Where in the City of New York, should the investor open a Chinese Restaurant?" The findings of this project will help the relevant investor better understand the advantages & disadvantages of different New York neighborhoods/boroughs in terms of opening a Chinese restaurant