**IBM Data Science Professional Certificate**

**IBM Applied Data Science Capstone**

**Capstone Project – The Battle of Neighborhoods**

**EVALUATING JAPANESE RESTAURANTS IN NEW YORK CITY USING K‑MEANS MACHINE LEARNING CLUSTERING METHOD**

**MARCELO ABBEHUSEN MAGALHÃES**

**March, 2021**

1. INTRODUCTION

By the year of 2011, there were about 20 thousand japanese immigrants living in New York City. According to the 2017 United States Census Bureau, there were 1,466,514 Americans with japanese ancestry living in USA.

The Asian American population is greatly urbanized, with nearly three-quarters of them living in metropolitan areas with population greater than 2.5 million. New York City is one of the three areas with the highest Asian American Populations (Greater Los Angeles Area, New York Metropolitan Area and San Francisco Bay Area).

According to the USA 2010 Census, New York is home to more that one million Asian Americans.

This project will attempt to answer the questions “Where should an investor open a Japanese Restaurant in NYC?” and “Where should I go If I want a great and highly rated Japanese food?”.

1. DATA

In order to answer the above questions, data on New York City neighborhoods and boroughs, including boundaries, latitude, longitude, restaurants and restaurants ratings are required.

New York City data containing the neighborhoods and boroughs, latitudes, and longitudes will be obtained from the data source: https://cocl.us/new\_york\_dataset

All data related to locations and quality of Japanese restaurants will be obtained via the FourSquare API, using the Requests library in Python.

1. METHODS

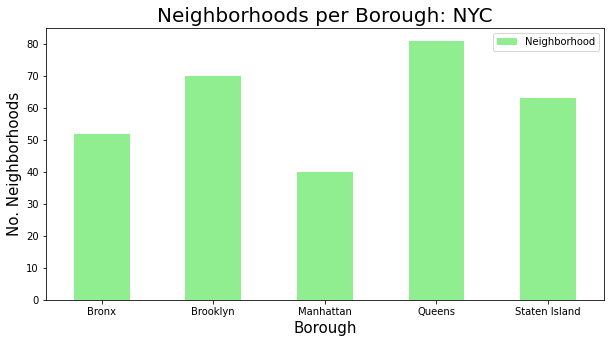
Data will be collected from https://cocl.us/new\_york\_dataset, cleaned and processed into a dataframe. FourSquare will be used to locate all venues and then filtered by Japanese restaurants. Users ratings will be counted and added to the dataframe.

After the data is preprocessed, It will be sorted based on ratings. Finally, It will be visualized using graphics from Python libraries and also divided into different clusters, using machine learning K-Means algorithm, from scikit-learn library.

1. RESULTS

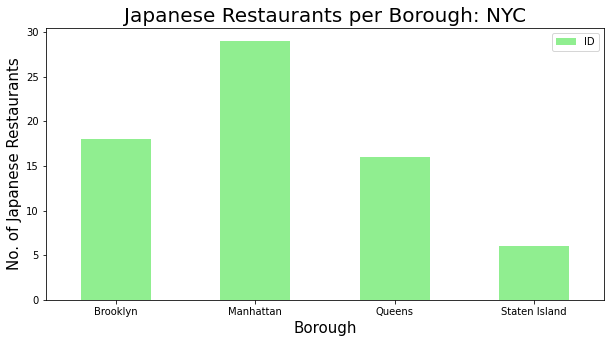
After using the Requests library to retrieve New York city data from the URL, the data was visualized in a bar graphic and in a pandas dataframe.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Borough** | **Neighborhood** | **Latitude** | **Longitude** |
| **0** | Bronx | Wakefield | 40,89471 | -73,8472 |
| **1** | Bronx | Co-op City | 40,87429 | -73,8299 |
| **2** | Bronx | Eastchester | 40,88756 | -73,8278 |
| **3** | Bronx | Fieldston | 40,89544 | -73,9056 |
| **4** | Bronx | Riverdale | 40,89083 | -73,9126 |
| **5** | Bronx | Kingsbridge | 40,88169 | -73,9028 |



Using the Foursquare API, It was able to retrieve data from 69 Japanese Restaurants in New York City, as shown in the table and figures below.

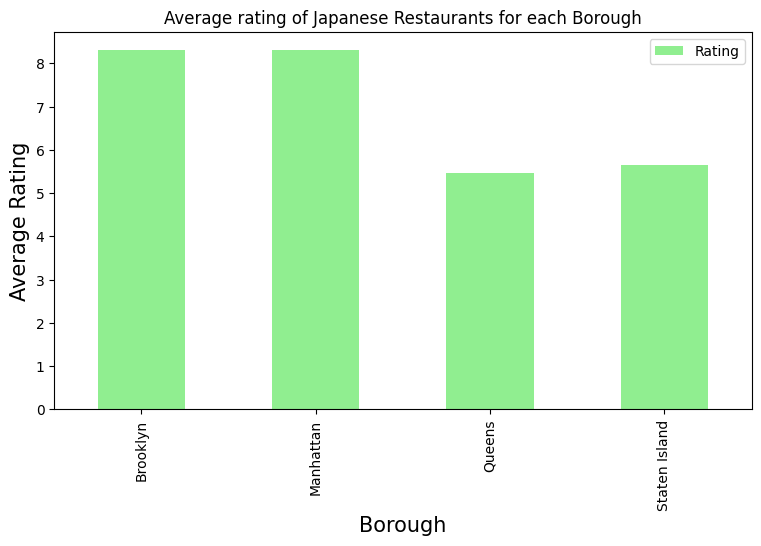
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Borough** | **Neighborhood** | **ID** | **Name** |
| 0 | Brooklyn | Kensington | 4d5c12a01e43236a87eb1583 | Sake Sushi |
| 1 | Brooklyn | Prospect Heights | 5cb5e5f9a35f4600255406c6 | Maison Yaki |
| 2 | Brooklyn | Williamsburg | 51f9b7b3498eefe896caeb23 | Shalom Japan |
| 3 | Brooklyn | Bedford Stuyvesant | 5b3bcb69bfc6d0002ca9bf17 | Warude |
| 4 | Brooklyn | Brooklyn Heights | 479ccb47f964a5206b4d1fe3 | Iron Chef House |
| 5 | Brooklyn | Cobble Hill | 48a41073f964a52091511fe3 | Hibino |



After retrieving data from japanese, data from likes and ratings were obtained also via Foursquare API, and the results were grouped by neighborhoods and borougs, as shown in the tables and chart below.

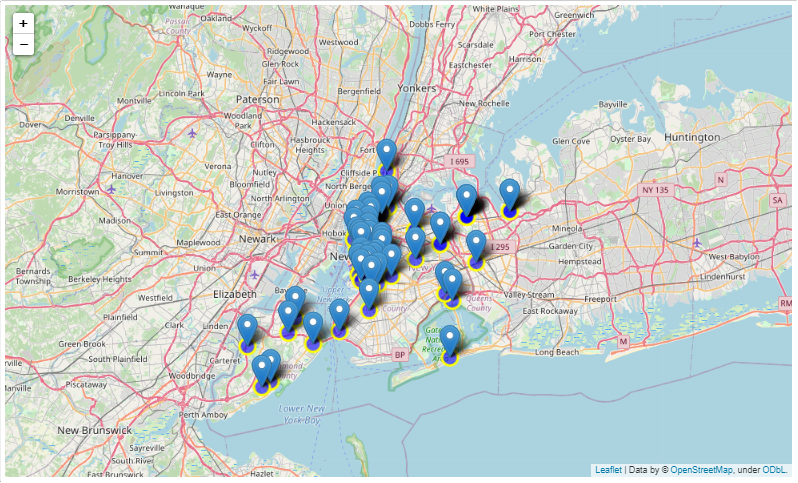
|  |  |
| --- | --- |
| **Neighborhood** | **Average Rating** |
| East Village | 9.25 |
| Cobble Hill | 9.10 |
| North Side | 9.00 |
| Park Slope | 8.90 |
| Downtown | 8.90 |
| Soho | 8.80 |
| Lindenwood | 8.80 |
| Chelsea | 8.80 |
| Boerum Hill | 8.70 |
| Fort Greene | 8.60 |

|  |  |
| --- | --- |
| **Borough** | **Average Rating** |
| Manhattan | 8.310.526 |
| Brooklyn | 8.306.667 |
| Staten Island | 5.660.000 |
| Queens | 5.472.727 |



Afterwards, the dataset was merged with the latitude/longitude data and finally plotted in a map, using the Folium library.

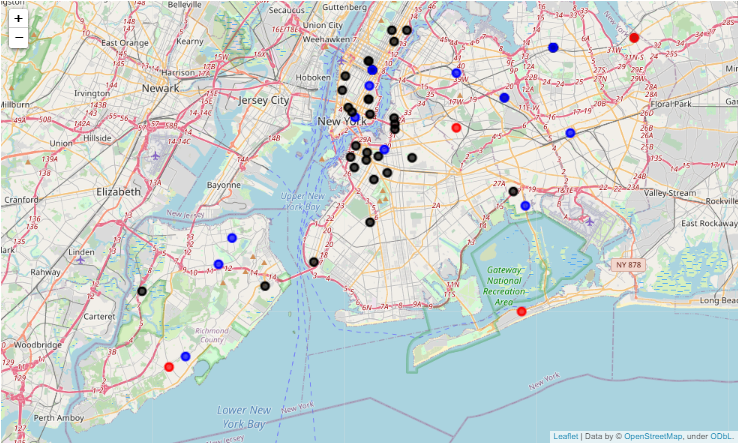
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Borough\_x** | **Neighborhood** | **Latitude** | **Longitude** | **Rating** |
| Brooklyn | Kensington | 40.642.382 | -73.980.421 | 7.9 |
| Brooklyn | Prospect Heights | 40.676.822 | -73.964.859 | 8.1 |
| Brooklyn | Williamsburg | 40.707.144 | -73.958.115 | 8.2 |
| Brooklyn | Bedford Stuyvesant | 40.687.232 | -73.941.785 | 8.2 |
| Brooklyn | Brooklyn Heights | 40.695.864 | -73.993.782 | 8.1 |
| Brooklyn | Cobble Hill | 40.687.920 | -73.998.561 | 9.1 |
| Brooklyn | Carroll Gardens | 40.680.540 | -73.994.654 | 7.8 |



Finally, the K-means algorithm was set to three clusters, that divided the data into three groups:

* Cluster 0: Mostly restaurants located in Brooklyn and Manhattan and that had the best ratings, from 7.9 to 9.3;
* Cluster 1: Restaurants with no rating, located basically in Queens;
* Cluster 2: Restaurants with ratings varying from 6.2 and 7.6, mostly located in Queens and Staten Island.

The following map shows the restaurants locations according to each cluster. The black spots are cluster 0, the red spots are cluster 1 and the blue are cluster 2.



1. CONCLUSIONS

In this project, It was possible to create a project that used most of the tools and methods that were studied during the course. I was able to identify a business problem, to specify the data required, to extract, prepare and visualize the data, to visualize the results and also to perform a machine learning clustering method, using K-Means and dividing the data into 3 clusters, based on their similarities, that was basically the rating each restaurant had.

The Manhattan and Brooklyn boroughs were clearly the regions with the higher rated japanese restaurants, labeled as Cluster 0 after KMeans algorithm were implemented. So, If one were to look for a good japanese restaurant, I'd recommend any neighborhood of these areas.

In other words, If a person is looking for a place to open a new Japanese restaurant, I'd recommend the boroughs of Queens or Staten Island, because besides having a small number of restaurants, the existing ones are low rated, which means these areas have potentially less competition. It's necessary to study more features, such as infrastructure, transports and wage rates of these regions, to be able to build a more robust and trustworthy model.