Capstone Project - The world cuisine of New York

Qi Luo

March 12, 2020

1 INTRODUCTION

1.1 BACKGROUND

New York is one of the most diverse cities in the world, which is reflected in its cuisine. Food is an important part of people's lives, the accessibility of cuisines from all over the world is a big attraction for a city, both for residents and tourists.

1.2 PROBLEM AND SCOPE

In this project, I will collect and analyze the data of the world cuisine in New York, to understand the eating preference of New Yorkers. Thus, it will offer insight and recommendation when someone wants to find a neighborhood in New York to live in, based on his/her cuisine preference; or when someone wants to open a restaurant in any location in New York.

2 DATA ACUISITION AND CLEANING

2.1 DATA SOURCE

The New York city neighborhoods dataset is taken from the course material (link). It contains the name of the neighborhoods, the borough it is located, and the coordinates of it.

Type and location of restaurants in every neighborhood will be obtained using Foursquare API. The endpoint used was 'Explore' (https://developer.foursquare.com/docs/api/venues/explore). The coordinates of neighborhood used was taken from the neighborhood dataset, the exploration radius was set to 500 meters, the 'Food' category ID was use to filter the venues (https://developer.foursquare.com/docs/resources/categories).

2.2 DATA PREPROCESSING

2.2.1 Neighborhood dataset

The New York city neighborhoods dataset is a well-organized **json** dataset. We only need to convert it from **json** to **Pandas DataFrame.**

2.2.2 Restaurant dataset

After being obtained from Foursquare API, the data have a few problems. First, some restaurant categories are generic, for example, 'Restaurant', 'Food', 'Deli / Bodega','Café','Diner','Food Stand'. These restaurant categories will not give much information in the data analysis, therefore, we will get rid of them.

Second, some venue categories have very low counts. Therefore, I omitted the venue categories that has a total count lower than 5.

Third, some neighborhoods have low count of venues. Which will not offer enough data to analyze. I omitted those neighborhoods that has a venue count lower than 25.

The restaurant dataset has no missing values.

3 EXPLORATORY DATA ANALYSIS AND RESULT

We will look at restaurants in the whole New York City. First, I counted the restaurants by category, the resulting top 10 categories is shown in the bar chart:

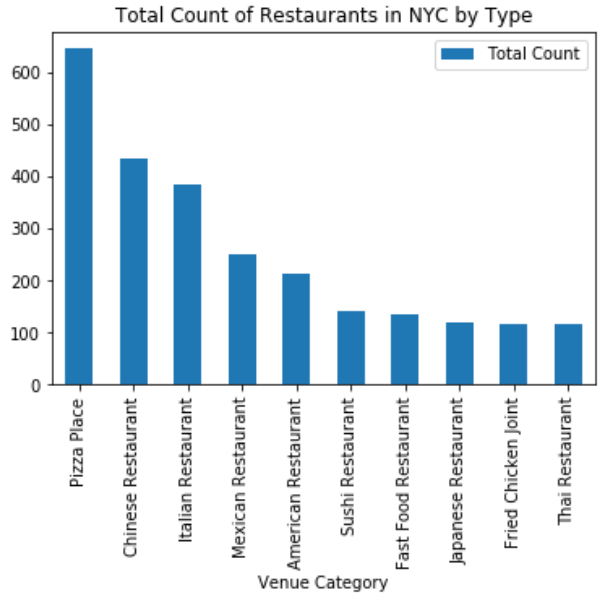

Figure 1 Total count of restaurants by type in NYC

From the bar chart, we can see that pizza place is leading, being the American favorite without a doubt. The Chinese, Italian and Mexican restaurants took the second to fourth place, followed by American, sushi, fast food, Japanese, fried chicken and Thai restaurant. This shows that New York is indeed an diverse city, with cuisines from all over the world.

By putting the top 6 cuisines on the map, with clustered markers, we can have an overview of the distribution of those cuisines (Figure 2-7).

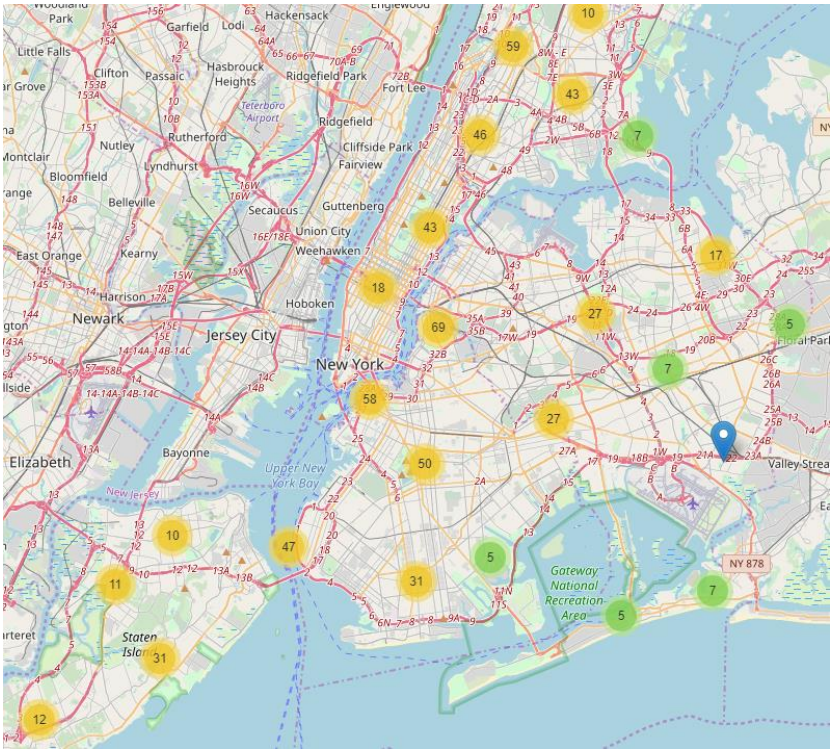

Figure 2 Pizza restaurants

Figure 3 Chinese restaurant

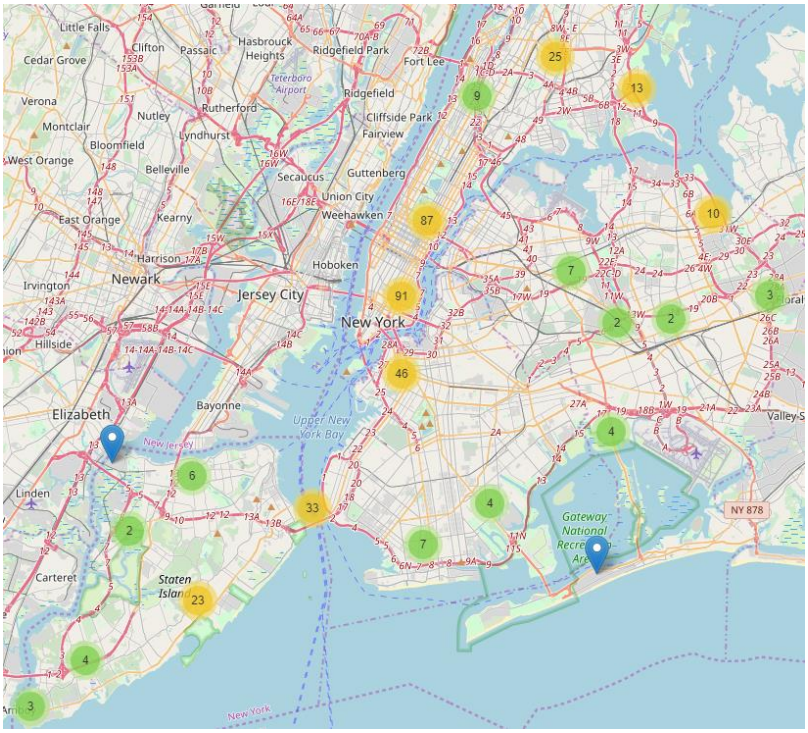

Figure 4 Italian restaurants

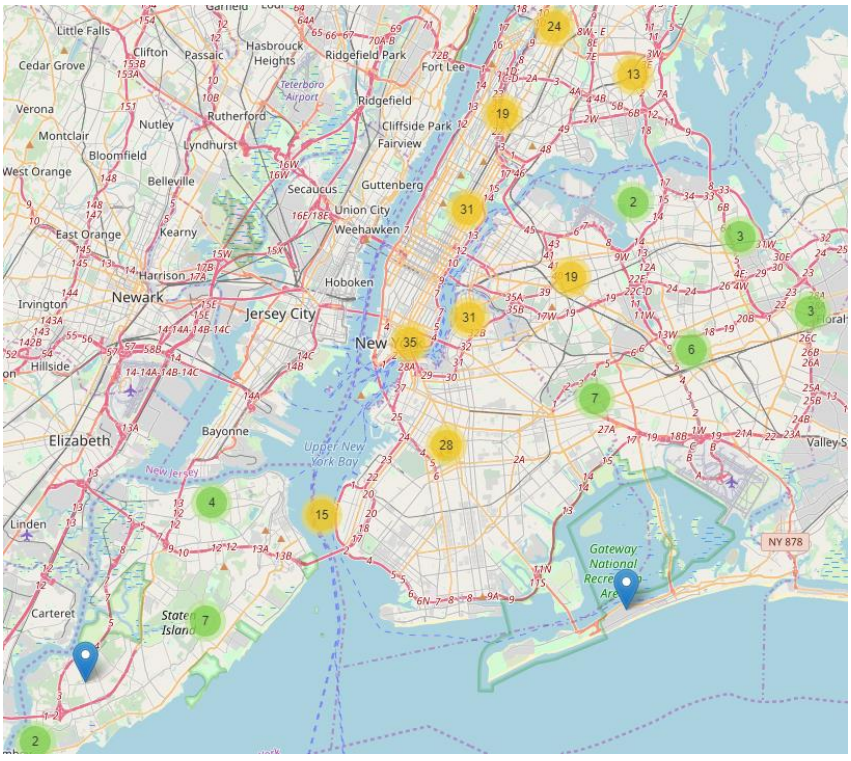

Figure 5 Mexican restaurants

Figure 6 American restaurants

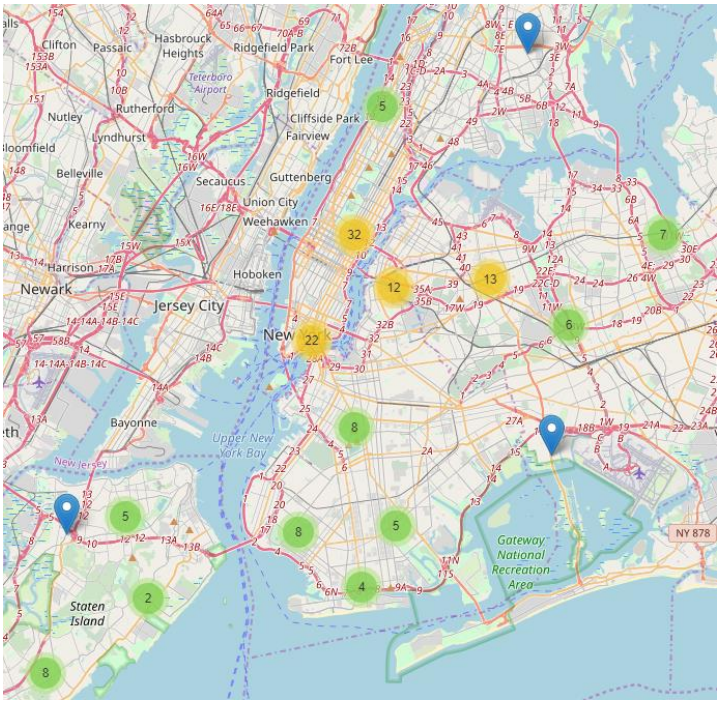

Figure 7 Sushi restaurants

We can see on the map that pizza places, being the most popular choice of New Yorkers, it is spread out across all neighborhoods and boroughs. While the other reginal cuisines, tend to be more concentrated on certain districts.

4 NEIGHBORHOOD CLUSTERING AND RESULT

From the data exploration, we can see that New York has a rich variety of world cuisines, and there is a difference in the cuisine distribution across neighborhoods. Therefore, we will apply clustering to segment the neighborhoods.

The restaurants per category in each neighborhood was transformed into a matrix using one hot encoding, then K-means clustering was applied to the data. The k value was set to 5.

The resulting clusters are reported in the following map.

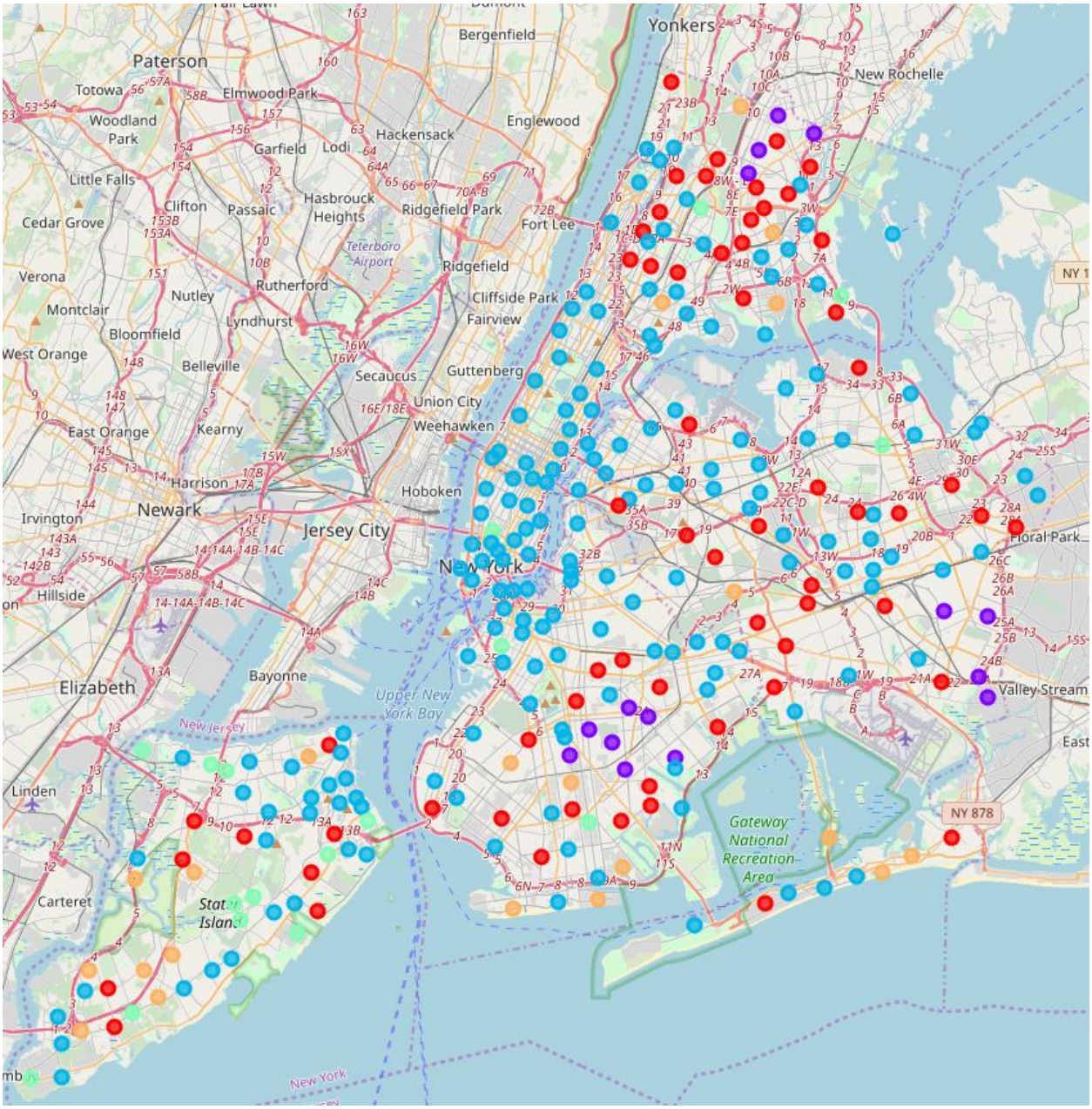

Figure 8 Clustering of neighborhoods. Red:0, purple: 1, blue: 2, cyan: 3, orange: 4

Number of neighborhoods in cluster 0: 63

Venue Category	Count per neighborhood
Chinese Restaurant	2.730159
Pizza Place	2.444444
Italian Restaurant	0.714286
Mexican Restaurant	0.460317
Fried Chicken Joint	0.412698
Spanish Restaurant	0.396825
Sushi Restaurant	0.333333
American Restaurant	0.317460
Japanese Restaurant	0.269841
Fast Food Restaurant	0.238095

Number of neighborhoods in cluster 1: 15

Venue Category	Count per neighborhood
Caribbean Restaurant	2.733333
Chinese Restaurant	1.133333
Fast Food Restaurant	0.666667
Pizza Place	0.533333
Fried Chicken Joint	0.533333
Seafood Restaurant	0.266667
Mexican Restaurant	0.200000
Southern / Soul Food Restaurant	0.133333
Asian Restaurant	0.133333
Breakfast Spot	0.133333

Number of neighborhoods in cluster 2: 171

Venue Category	Count per neighborhood
Pizza Place	2.350877
Italian Restaurant	1.538012
Chinese Restaurant	1.356725
Mexican Restaurant	1.251462
American Restaurant	1.023392
Sushi Restaurant	0.666667
Thai Restaurant	0.614035
Fast Food Restaurant	0.590643
Japanese Restaurant	0.573099
Latin American Restaurant	0.573099

Number of neighborhoods in cluster 3: 17

Venue Category	Count per neighborhood
Italian Restaurant	4.470588
Pizza Place	1.294118
American Restaurant	0.529412
Chinese Restaurant	0.352941
French Restaurant	0.352941
Sushi Restaurant	0.294118
Seafood Restaurant	0.235294
Fast Food Restaurant	0.235294
Greek Restaurant	0.235294
Thai Restaurant	0.176471

Number of neighborhoods in cluster 4: 21

Venue Category	Count per neighborhood
Pizza Place	2.857143
American Restaurant	0.380952
Chinese Restaurant	0.285714
Spanish Restaurant	0.190476
Fast Food Restaurant	0.190476
Seafood Restaurant	0.095238
Burger Joint	0.095238
Caribbean Restaurant	0.095238
Fried Chicken Joint	0.095238
Italian Restaurant	0.095238

The clustering is rather successful, the neighborhoods are separated into 5 distinctive groups. Sixty-three neighborhoods are in Cluster 0, which has an average of 2.73 Chinese restaurants per neighborhood, followed by 2.44 Italian restaurant. Only 15 neighborhoods are in Cluster 1, but it has a distinctive cuisine distribution: 2.73 Caribbean Resataurant and 1.13 Chinese restaurant per neighborhood. The majority is cluster 2, which counts 173 neighborhoods. This cluster has a well-balanced mix, where each neighborhood has at least one pizza, Italian, Chinese, Mexican and American restaurant on average. Cluster 3 and 4 are the neighborhoods that have persistent love on certain cuisines: cluster 3 have a whooping 4.47 Italian restaurants per neighborhood and less than one other restaurants per neighborhood, while cluster 4 have 2.85 pizza place per neighborhood and less than one of other restaurants per neighborhood.

5 CONCLUSIONS

In this project, I analyzed the restaurant categories in New York city, in the perspective of various cuisines from all over the world. The most prevalent restaurant is pizza place,which you can find in almost every neighborhood. It is followed by Chinese, Italian, and Mexican restaurants.

K-mean clustering was applied to identify the difference in the neighborhoods concerning the cuisine distribution. There are distinctive preference on cuisines in different clusters, there are pizza, Chinese, or Italian loving neighborhoods, etc. Carribean Restaurant is worth mentioning, it is not in the top ten popular restaurant across the New York City, but stands out in 15 neighborhoods which created a distinctive cluster.