# Restaurant Location Analysis

## IBM Applied Data Science Capstone

**By: Luanluan Xu**

**July, 2019**

## Table of contents



## 1. Introduction

As a Chinese proverb goes, food is the first necessity of the people, food is the most basic conditions for survival. This shows the importance of eating to human beings. Three meals a day are indispensable. However, with the speeding pace of life and high intensity work under pressure in the modern society, spending a lot of time and energy to cook at home has become a luxury. More and more people choose to eat out or increase the frequency of eating out. Especially in modern metropolises like New York City, restaurants are everywhere, opening and closing every day. But why some restaurants after opening can be filled with customers and make a lot of money, while others have slow bussiness. Location is one of the important factors.

### 1.1 Bussiness Problem

So how to decide a location-selection? That is the question to be answered in this project. Based on the data of 2,698 restaurants in five Boroughs of New York City, It aims to study the influence of household income level, population density, existing competitors and other factors on the location-selection，and find out the patterns.

### 1.2 Targeted Audiences

This report is intended for owners or investors who want to open a restaurant. A good location is helpful to bring more customers and more sales for a restaurant. The report will be used as evaluation criteria and scientific suggestions for location-selection.

# 2. Data Section

## 2.1 Data Requirement and Collection

For the project, those data are needed:

1) Neighborhoods in New York City From Module 3 in this course, we can get 5 boroughs in NYC and 306 neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood, I have saved it to dy_newyork.csv.

2) Restaurants in New York City Use the Foursquare API to explore the neighborhoods and get all the venues data. We only need the restaurants data, so I just filter the data that the venue category contains 'Restaurant'.

3) Population density by Neighborhoods a Wikipedia page ("https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City (https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City) ") has those information. I scraped the page and get the data.

4) Median Household Income by Neighborhoods Those data can be found in a web page (https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability (https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability)), I downloaded them and had saved to 'Household_income_By_Neighborhoods_NYC.csv'.

## 2.1 Data Cleaning and Preparation

I clean the data and drop some features that I think they useless or irrelevant. For example, I drop two features, area and population by Community Board(CB) in the table- ‘Population density by Neighborhoods’, because in order to maintain data consistency, we only need the data by neighborhoods.

There are several problems in the cleaning process.

1) Because the original population density is calculated by community board, some neighborhoods in the same board are in the same cell. I split them and make sure they can be combined with Foursquare location data.

2) There are 2155 population density data and 861 median household income data missed in a total of 2698 restaurant samples. Those data is too much to be ignored, so I give them the data by borough which is the mean of data by neighborhoods instead.

3) Some restaurants have opened another store in the same or different neighborhood, this is an independent sample, but it may result in an error. So I combine the restaurant name with 'index' to give every restaurant a unique name.

4) Competitors around the location can divert customers, but I can't find such data online. So I build a new feature named competitors which is the number of restaurants in the same neighborhood.

Now, data from all sources are combined into one table. There are 2698 restaurant samples and 11 features.

(2698, 11)

| | Restaurant name | Neighborhood | Borough | Median Household Income | Neighborhood Latitude | Neighborhood Longitude | Pop./km2 | Venue Category | Venue Latitude | Venue Longitude | Competitors |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | White Castle_0 | Allerton | Bronx | 37816.894737 | 40.865788 | -73.859319 | 12149.0 | Fast Food Restaurant | 40.866065 | -73.862307 | 3 |
| 1 | Chef King_1 | Allerton | Bronx | 37816.894737 | 40.865788 | -73.859319 | 12149.0 | Chinese Restaurant | 40.865561 | -73.856752 | 3 |
| 2 | Internacional Restaurant & Deli_2 | Allerton | Bronx | 37816.894737 | 40.865788 | -73.859319 | 12149.0 | Spanish Restaurant | 40.863809 | -73.856640 | 3 |
| 3 | Il Sogno_758 | Annadale | Staten Island | 66764.200000 | 40.538114 | -74.178549 | 2593.0 | Restaurant | 40.541286 | -74.178489 | 3 |
| 4 | Diesel Bagels_759 | Annadale | Staten Island | 66764.200000 | 40.538114 | -74.178549 | 2593.0 | American Restaurant | 40.540373 | -74.177374 | 3 |

## 3.Methodology

In this project, I first try to classify restaurants using k-means algorithm. Then, I perform statistical analysis. Last, I make a conclusion based on my analysis.

After obtaining class labels from the k-means classification, I visualize all restaurants in NYC in a map using Folium. Each restaurant is represented by a filled circle. The color of the circle is assigned according to its class. I notice that there are more competitors in some area while less in others.

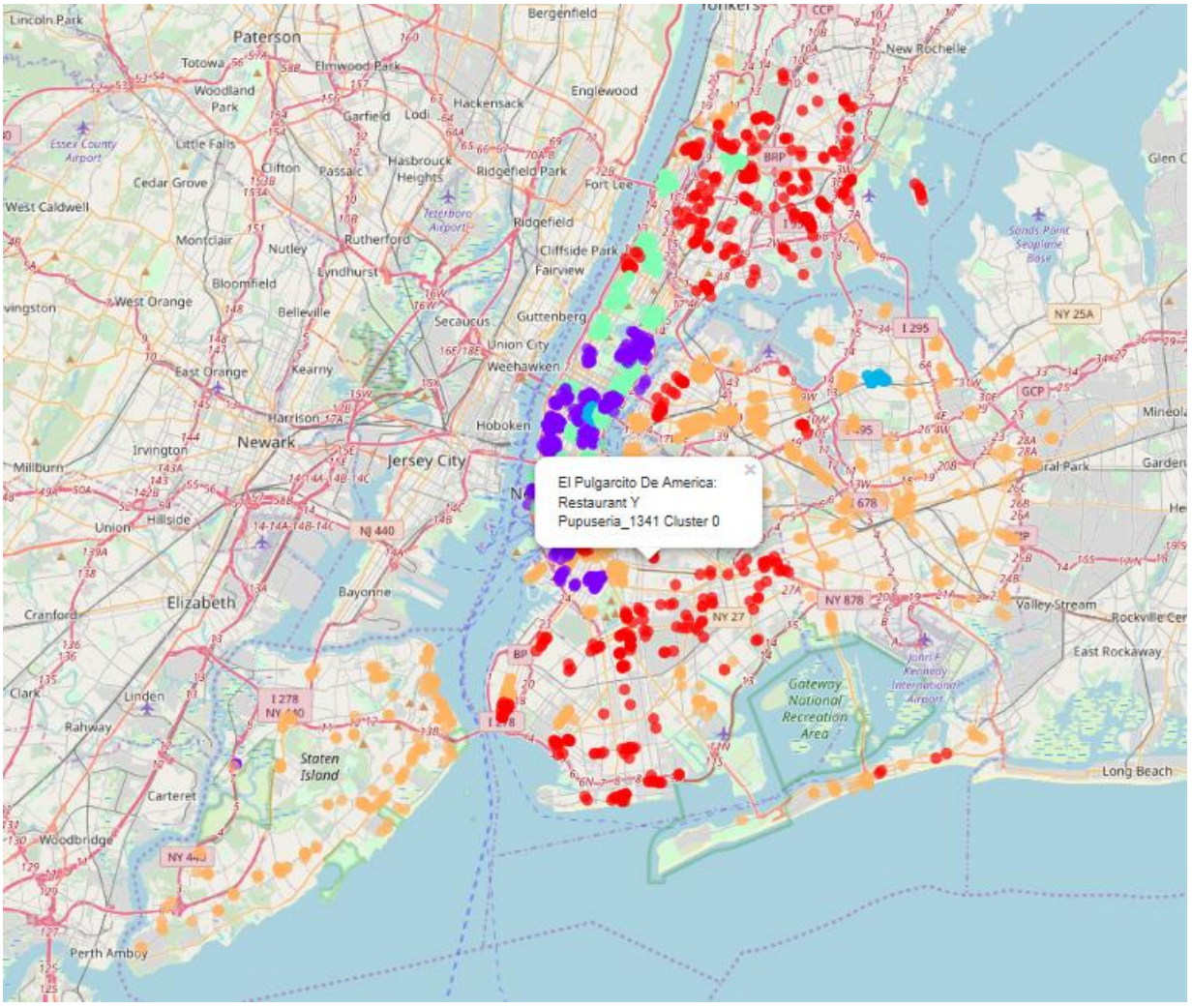
Then, I summarize individual feature's statistic patterns and also calculate the Pearson correlations between income and number of competitors and between population density and number of competitors. Both correlations are statistically significant: number of competitors are positively correlated with median household income and population density. I filter the neighborhoods with 'Median Household Income' and 'Population density' higher than 75% and 'Competitors' less than mean, 8 out of 306 neighborhoods are chosen.

Based on these data analysis, I found that the neighborhood with lower population density and low median household income have the least competitors. This suggests that there may be rooms for opening new restaurant in those neighborhoods. I also discuss some other factors that may be cofounds of the results with an example.

# 4. Exploratory Data Analysis

#### 4.1 Using *k*-means for Restaurants Segmentation

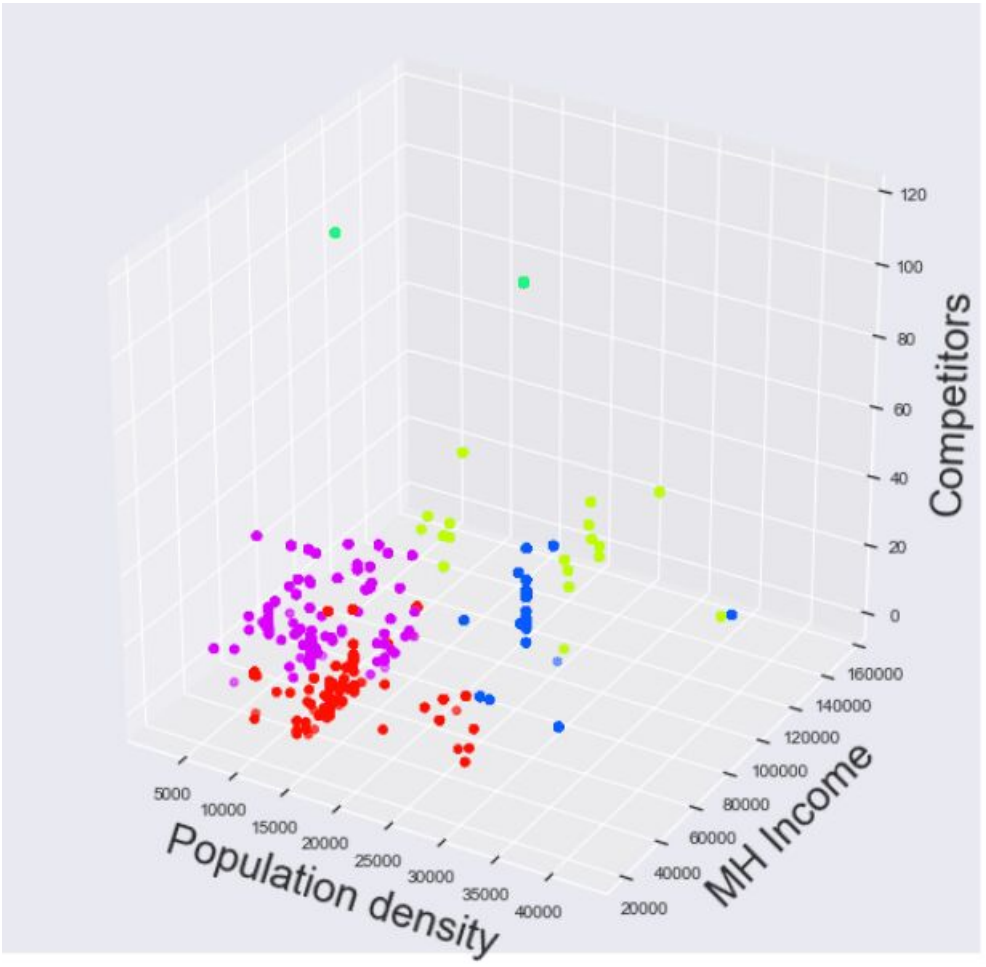
In the original database, there are 11 features. Here I create a new dataframe- 'df1' for clustering, which only contains 4 features: 'Restaurant name', 'Median Household Income', 'Pop./km2' (Population density) and 'Competitors'. I normalize the dataset, run k-means to cluster the restaurants into 5 clusters, and visualize all restaurants in NYC in a map using Folium.

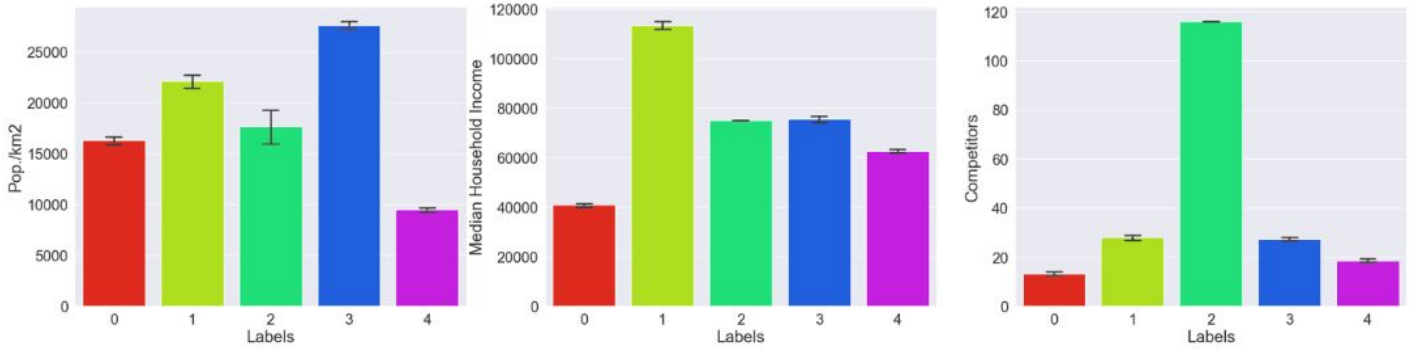



From the map, we can see there are more competitors in some area, while less in others. But what's the relationship between 'Competitors' and other variables? Let's find out.

#### 4.2 Analyzing Individual Feature Patterns using Visualization

In this section, a scater diagram and 3 histograms are ploted.




From the figures, cluser 5('Lables'==4) have the lowest population density, and cluser 1 ('Lables'==0) have the lowest Median Household Income, they have less Competiors too. what's the relationships between Median Household Income, Population density and Competitors? Let's continue analyzing.

### 4.3 Neighborhoods analysis

In order to find the patterns, I create a new dataframe called 'df_nei', which contains 4 features: 'Neighborhood', 'Median Household Income', 'Pop./km2', 'Competitors'(number of restaurants in the neighborhood).

Let's first take a look at the variables by utilizing a description method.

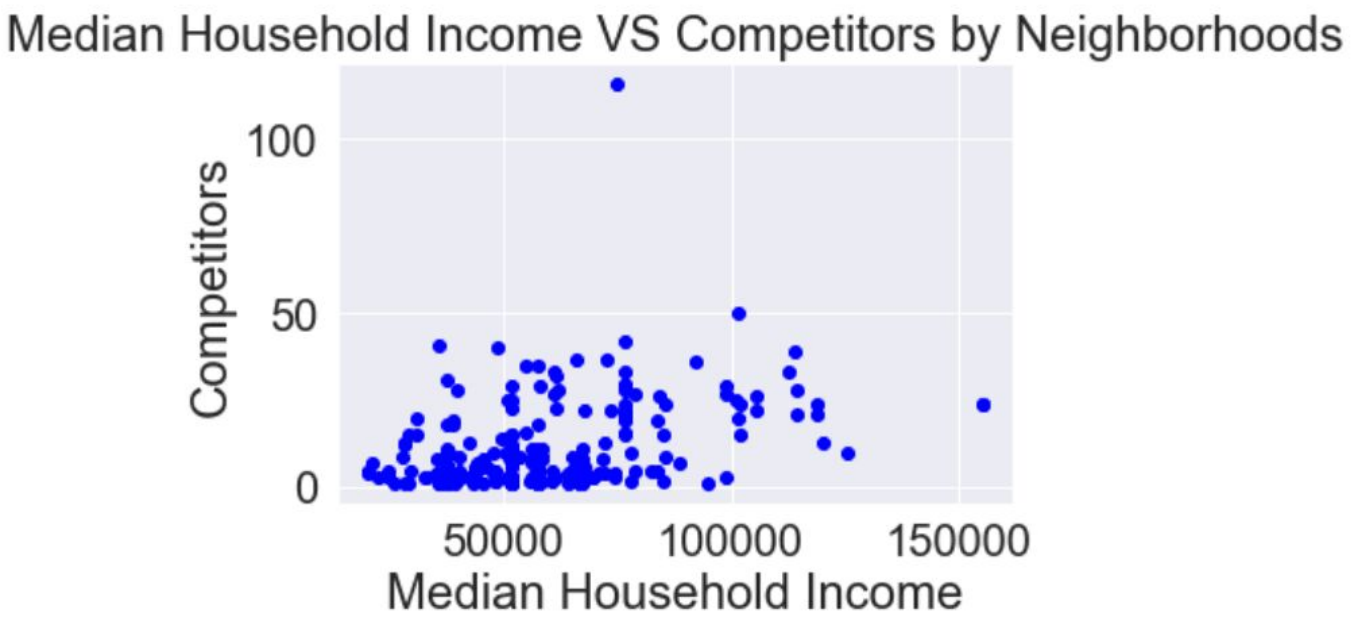
| | Median Household Income | Pop./km2 | Competitors |
|---|---|---|---|
| **count** | 265.000000 | 265.000000 | 265.000000 |
| **mean** | 59699.984978 | 13035.127676 | 10.181132 |
| **std** | 21844.841354 | 7505.374929 | 12.168219 |
| **min** | 20334.000000 | 2326.000000 | 1.000000 |
| **25%** | 44027.000000 | 8803.967071 | 3.000000 |
| **50%** | 57813.625000 | 14153.791939 | 5.000000 |
| **75%** | 66764.200000 | 14353.379074 | 14.000000 |
| **max** | 155213.000000 | 42312.000000 | 116.000000 |

Then, calculate the Pearson Correlation Coefficient and P-value between variables.

**a. Median Household Income VS Competitors**

The Pearson Correlation Coefficient is $r$ = 0.353 with a P-value of P = 3.116e-09. Since the p-value is << 0.001, the correlation between Median Household Income and Competitors is statistically significant, although the linear relationship isn't extremely strong (~0.35).

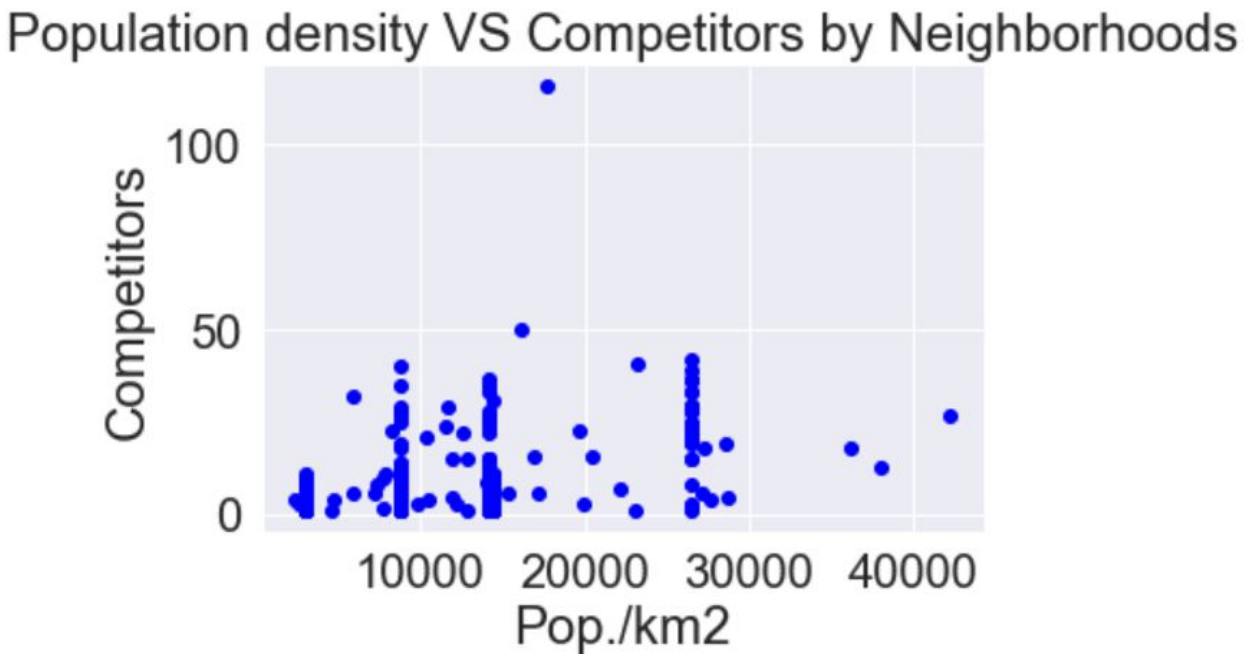
From the 2D Scater-plot, we can see that when the neighborhood tends to have more competitors with a higher Median Household Income.



**b. Population density VS Competitors**

The Pearson Correlation Coefficient is *r* = 0.406 with a P-value of P = 5.953e-12. Since the p-value is << 0.001, the correlation between Population density and Competitors is statistically significant, although the linear relationship isn't extremely strong (~0.41).

From the plot, less obviously with noise which is caused by the missing information, competitors tend to rise as the population density increases.




Let's find the neighborhoods which have Median Household Income > 66764(75%) and Population density > 14353(75%) but Competitor < 11(mean=10.18).

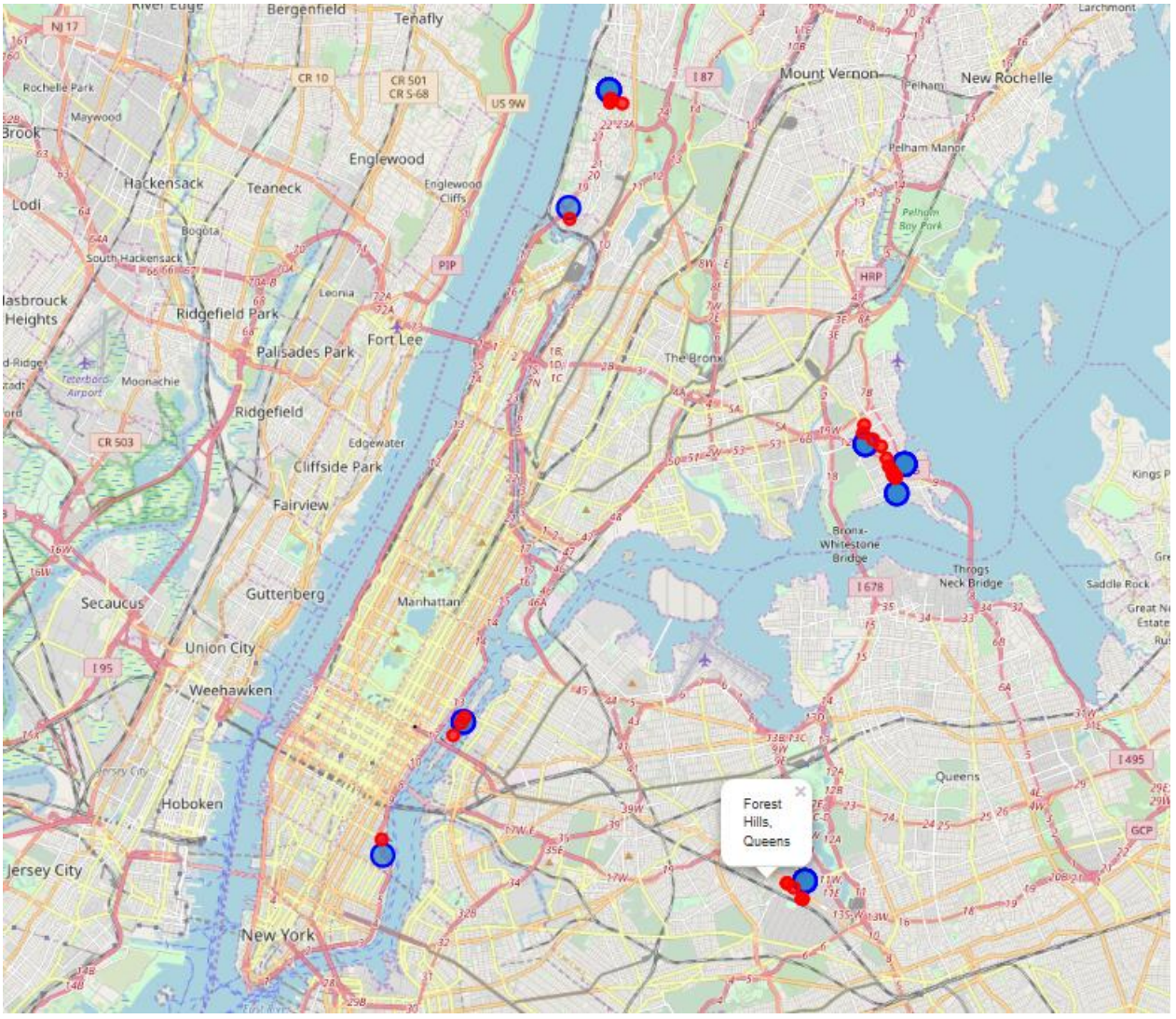
8 neighborhoods are chosen. They have higher Median Household Income and Population density, but number of restaurant is fewer than mean. That means their market consumption capacity is undervalued currently comparing to other neighborhoods.This suggests that there may be rooms for opening new restaurant in those neighborhoods.

| | Neighborhood | MHI>66764 | Com<11 | Pop>14353 |
|---|---|---|---|---|
| **0** | Edgewater Park | 67549.0 | 8 | 14353.379074 |
| **1** | Forest Hills | 67881.0 | 6 | 15279.000000 |
| **2** | North Riverdale | 78895.0 | 5 | 14353.379074 |
| **3** | Roosevelt Island | 98797.0 | 3 | 26482.390996 |
| **4** | Schuylerville | 67549.0 | 4 | 14353.379074 |
| **5** | Spuyten Duyvil | 67534.0 | 1 | 14353.379074 |
| **6** | Stuyvesant Town | 95022.0 | 1 | 26482.390996 |
| **7** | Throgs Neck | 67549.0 | 3 | 14353.379074 |

There are 31 restaurants in those 8 neighborhoods. Let's visualize them by Folium.

```
(31, 11)
```

| | Restaurant name | Neighborhood | Borough | Median Household Income | Neighborhood Latitude | Neighborhood Longitude | Pop./km2 | Venue Category | Venue Latitude | Venue Longitude | Competitors |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Muscle Maker Grill_1403 | Edgewater Park | Bronx | 67549.0 | 40.821986 | -73.813885 | 14353.379074 | American Restaurant | 40.819391 | -73.817298 | 8 |
| 1 | Patricia's of Tremont_1404 | Edgewater Park | Bronx | 67549.0 | 40.821986 | -73.813885 | 14353.379074 | Italian Restaurant | 40.823119 | -73.819403 | 8 |
| 2 | Green Dragon_1405 | Edgewater Park | Bronx | 67549.0 | 40.821986 | -73.813885 | 14353.379074 | Asian Restaurant | 40.818878 | -73.816793 | 8 |
| 3 | Tosca Marquee_1406 | Edgewater Park | Bronx | 67549.0 | 40.821986 | -73.813885 | 14353.379074 | Italian Restaurant | 40.819222 | -73.817601 | 8 |
| 4 | Spoto's Italian Restaurant_1407 | Edgewater Park | Bronx | 67549.0 | 40.821986 | -73.813885 | 14353.379074 | Italian Restaurant | 40.820399 | -73.817702 | 8 |




## 5.Results and Discussion

According to data analysis, we have found 8 neighborhoods where there may be rooms for opening new restaurant. In fact, there are stil many other factors to be considered while deciding a location selection.

For example, if the owner wants to open a Italian restaurant. Edgewater Park is not a good choice. Because there have been 4 Italian Restaurants in this area. If he does so, the new restaurant may spend too much time on marketing because of the abundance of competitors around neighborhood. It's better to choose a location from the rest.

| | | Numbers of restaurants |
|---|---|---|
| **Neighborhood** | **Venue Category** | |
| **Edgewater Park** | **American Restaurant** | 1 |
| | **Asian Restaurant** | 1 |
| | **Chinese Restaurant** | 1 |
| | **Fast Food Restaurant** | 1 |
| | **Italian Restaurant** | 4 |

## 6.Conclusion

Overall, location selection is a very complicated process. There are many considerations that should be made in the selection. According to this report, 8 out of 306 neighborhoods are recommended to owners or investors who want to open a restaurant. Although it can't give the exact location, it helps to narrow it down!

**Thank you!**