New Restaurant in NYC

Contents

Introductory Notes

Someone looking to open a restaurant in New York City, where would you recommend that they open it?

In this project will try to find an optimal location for a new Spanish restaurant in NYC, we need to find locations (Neighborhood) that that isn't full of Spanish restaurants.

On the other hand, We will associate the success of each restaurant to the number of 'likes' and their 'rating we have to find the neighborhood with the number of rating and their food security

For food security we will use the New York City Restaurant Inspection Data. This dataset provides restaurant inspections, violations, grades and adjudication information.

Using the information obtained through the Foursquare API we will collect the rating and the number of likes.

Data Acquisition and Preparation

We will collect our data from two sources:

1. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
2. Foursquare API

The first one contains the Restaurant fields:

- Names
- Location (latitude and longitude)
- Grade of Inspection
- Cuisine

The second one contains the rating and number of likes for each Restaurants.

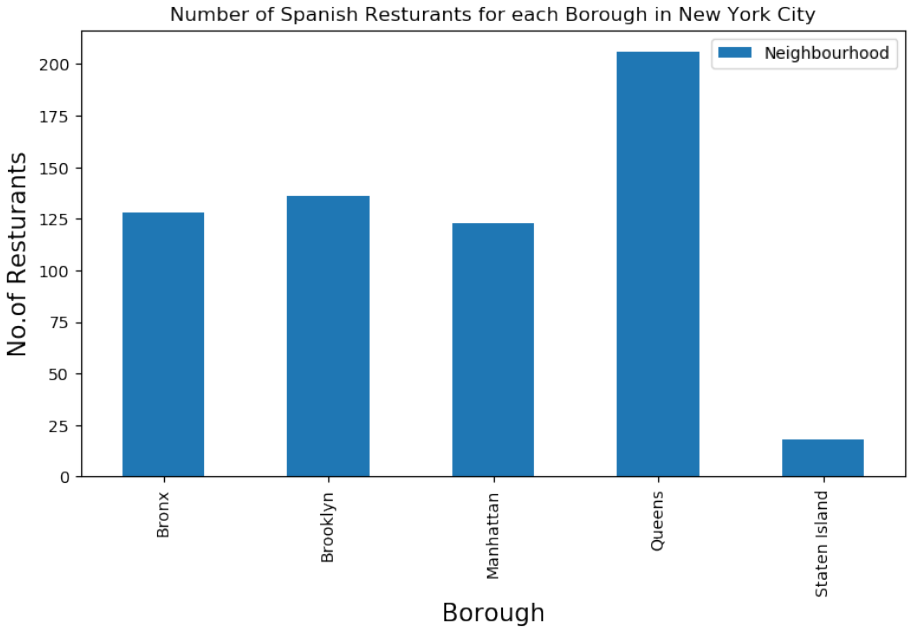
We will eliminate duplicate restaurants and obtain the number and filter the dataset per Spanish cuisine.

Data Preparation

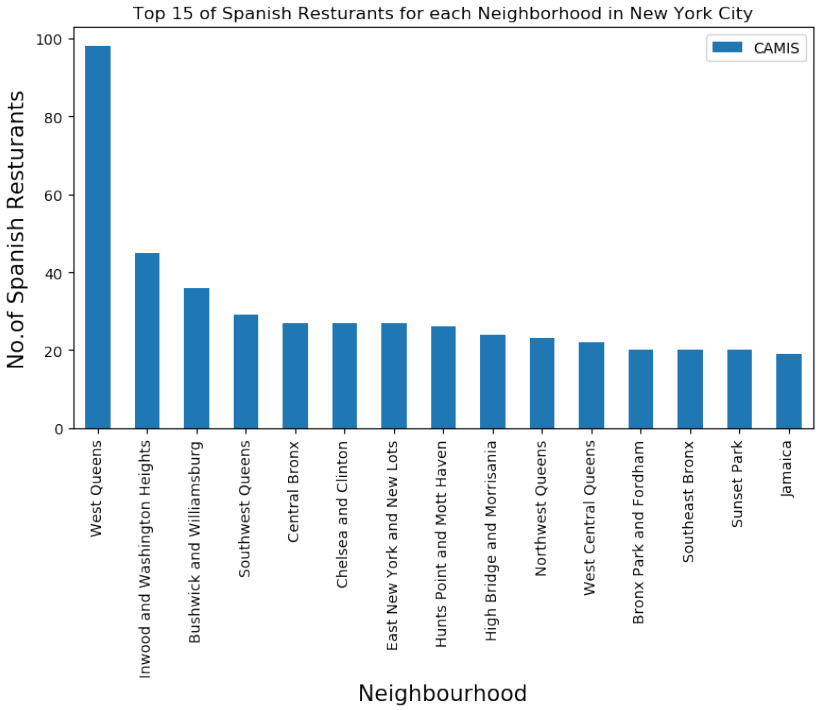
- Remove some unnecessary information
- checking missing data in data
- Create some new columns (SCORE%,Rating,Likes)
- Convert date columns
- Delete duplicate restaurants
- Filter restaurant by Spanish cuisine

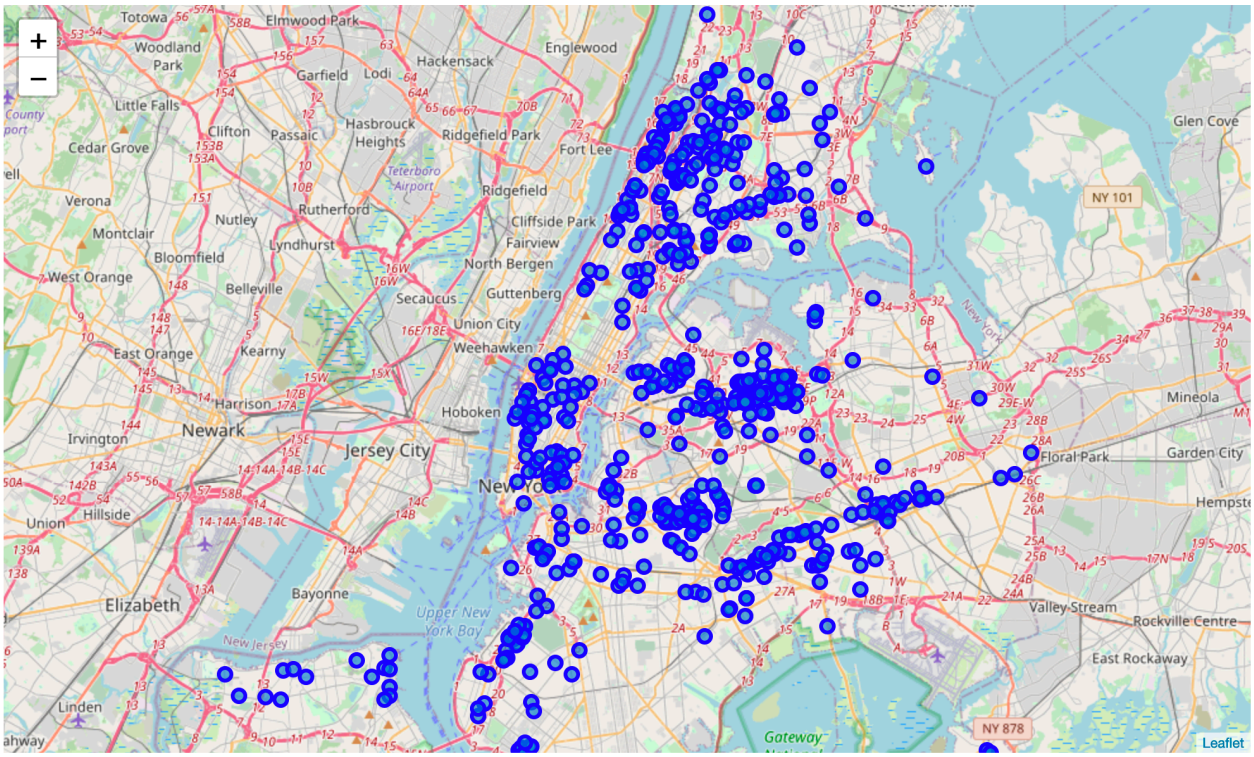
Data Analysis and Visualisation

First of all we want to visualise how the Spanish Restaurants are distributed around NYC for each Borough.



We see that most restaurants are located in the **West Queens** Neighborhood.




Totowa
Elmwood Park
Woodland Park
Garfield
Lodi
Hackensack
Englewood
Little Falls
Clifton
Passaic
Hasbrouck Heights
Ridgefield Park
Fort Lee
Cedar Grove
Teterboro Airport
Ridgefield
Glen Cove
Verona
Nutley
Rutherford
Cliffside Park
NY 101
Montclair
Lyndhurst
Fairview
Bloomfield
North Bergen
West Orange
Belleville
Secaucus
Guttenberg
Union City
East Orange
Kearny
Weehawken
Harrison
Hoboken
Mineola
Irvington
Newark
Jersey City
Floral Park
Garden City
Hillside
Union
Elizabeth
Bayonne
Upper New York Bay
Valley Stream
Rockville Centre
New Jersey
East Rockaway
Linden
Gateway
NY 878
Leaflet

Methodology

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.

Data points are clustered based on feature similarity.

The results of the K-means clustering algorithm are:
- The centroids of the K clusters, which can be used to label new data.
- Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

I will use trendy recommendation filtering approach in order to make recommendations.

Trendy Recommendation: we use the rating and the number of likes obtained through the Foursquare API and the grade of inspections to clustering restaurants into 10 groups

Based on the result above, the third, sixth and seventh are the clusters with the worst results, so we created a new dataset with the remaining data.

```
Cluster Labels =  0
SCORE Per_x    78.142857
Rating_x        7.614286
Likes_x        80.714286
dtype: float64
Likes_x    7
dtype: int64


Cluster Labels =  1
SCORE Per_x     88.0
Rating_x         9.0
Likes_x        835.0
dtype: float64
Likes_x    1
dtype: int64


Cluster Labels =  2
SCORE Per_x     78.333333
Rating_x         8.666667
Likes_x        379.000000
dtype: float64
Likes_x    3
dtype: int64


Cluster Labels =  3
SCORE Per_x    84.000000
Rating_x        7.186364
Likes_x        13.545455
dtype: float64
Likes_x    22
dtype: int64


Cluster Labels =  4
SCORE Per_x     70.333333
Rating_x         7.866667
Likes_x        202.333333
dtype: float64
Likes_x    3
dtype: int64
```

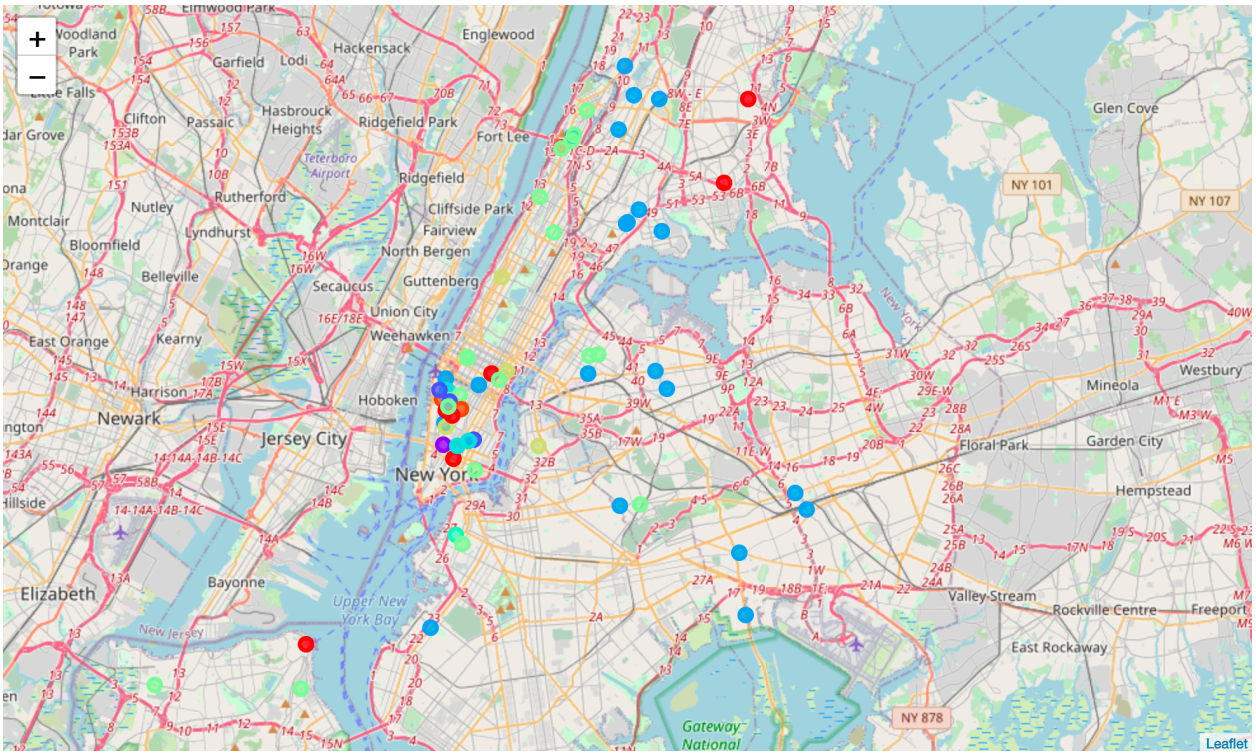
```
Cluster Labels =  5
SCORE Per_x     91.0
Rating_x         9.1
Likes_x        493.0
dtype: float64
Likes_x    1
dtype: int64


Cluster Labels =  6
SCORE Per_x    70.450
Rating_x        7.355
Likes_x        24.450
dtype: float64
Likes_x    20
dtype: int64


Cluster Labels =  7
SCORE Per_x     74.75
Rating_x         8.15
Likes_x        149.00
dtype: float64
Likes_x    4
dtype: int64


Cluster Labels =  8
SCORE Per_x     89.0
Rating_x         9.2
Likes_x        905.0
dtype: float64
Likes_x    1
dtype: int64


Cluster Labels =  9
SCORE Per_x     89.000000
Rating_x         8.166667
Likes_x        244.333333
dtype: float64
Likes_x    3
dtype: int64
```


Englewood
Hackensack
Garfield
Lodi
Clifton
Passaic
Hasbrouck Heights
Ridgefield Park
Fort Lee
Teterboro Airport
Ridgefield
Rutherford
Nutley
Lyndhurst
Cliffside Park
Fairview
North Bergen
Bloomfield
Belleville
Secaucus
Guttenberg
Union City
Weehawken
East Orange
Kearny
Harrison
Hoboken
Newark
Jersey City
New York
Elizabeth
Bayonne
Upper New York Bay
New Jersey
Glen Cove
NY 101
NY 107
Mineola
Westbury
Floral Park
Garden City
Hempstead
Valley Stream
Rockville Centre
Freeport
East Rockaway
NY 878
Gateway National
Leaflet

Conclusion

We recommend our business partner to open the new restaurant in a trendy Neighborhood with few Spanish restaurants open.

With that, we have concluded that the best recommendation will be neighborhood Chelsea and Clinton with a higher Trendy Recommendation, lower competition and easy replication for business expansion.