**Identifying Restaurant Locations & Venues Data Analysis for New York City**

1. Introduction

A.1 Description & Discussion of the Background

New York has always been a lucrative and dream destination for most of the people across the world. It has some of the world’s finest education institutions and from time to time it succeeded in securing, one of the positions in – ‘Word’s Most Livable City’. This fascinated many and some immigrated to New York City over the due course of time.

Last month while having a discussion with my friend, he mentioned that one of his relatives wants to open a restaurant in New York and interested in settling down there. However, my friend’s relative is not sure which neighborhood of New York city would be ideal for him to choose from, for this business venture.

I show this as a nice opportunity to apply my knowledge and agreed to use my Data Scientists skill to help my friend’s relative. This is how the problem got passed down to me and I was entrusted to advise some neighborhoods in New York city for opening a restaurant with some facts and figures. However, this analysis can be used by anyone who has a similar notion for such business ventures.

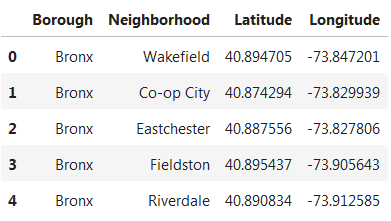
A.2 Data Description

The quest for New York’s city borough and neighborhood data took me to a web page [1] where I was able to find the required data in JSON format.

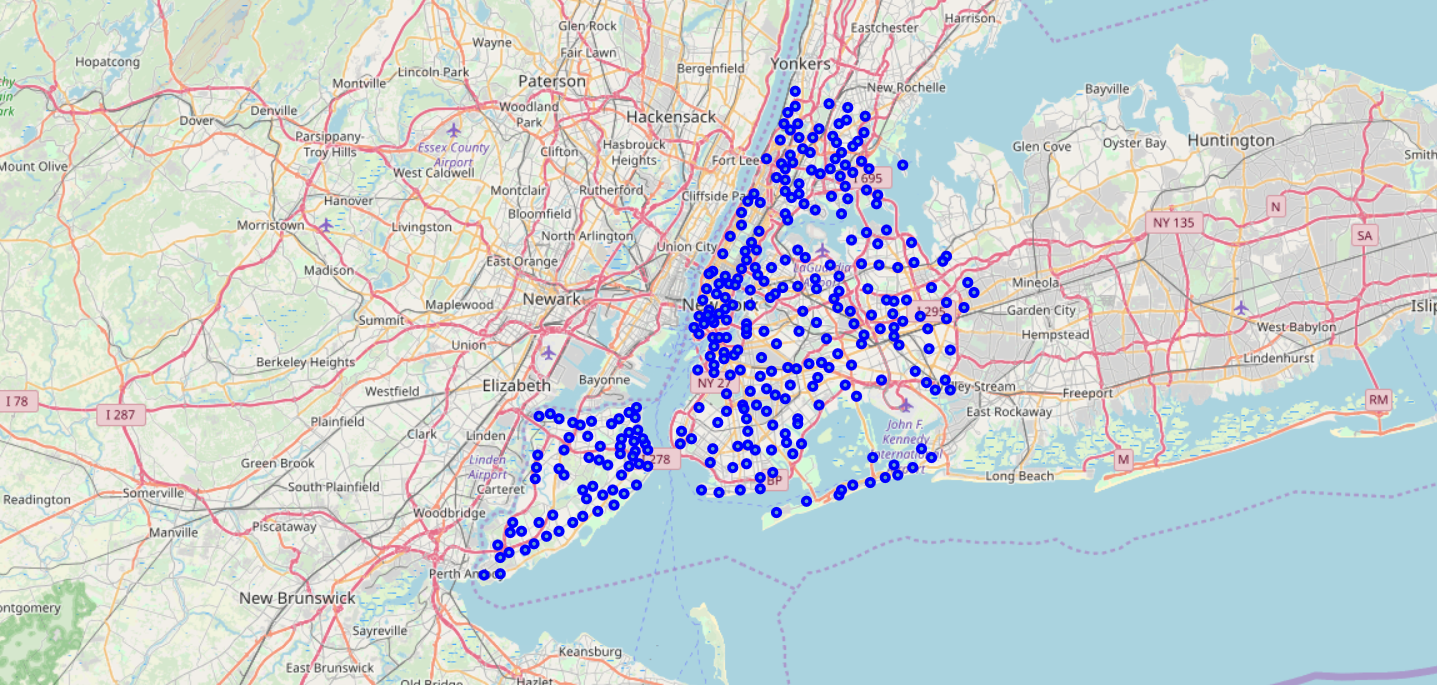
The extracted data has the required details to proceed with the analysis. It has got 5 boroughs with 306 neighborhoods’ latitude and longitude details. We will use Foursquare API [2]to search for the restaurants having keyword ‘Restaurant’ in the venue name. We will then use the explore function of Foursquare API to get 10 most common venue categories for each neighborhood. On this k-means clustering algorithm will be applied to group the neighborhoods into clusters and then information will be deduced from each of the derived clusters.

1. Methodology

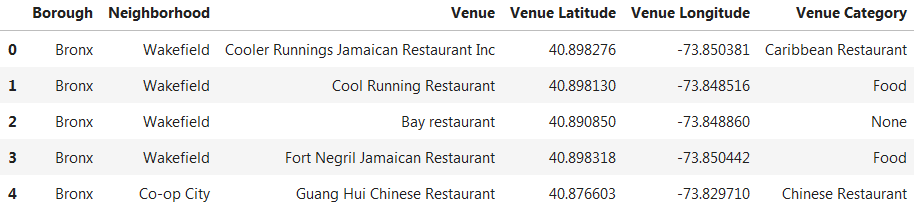
The master data extracted from JSON file has the following main components – Borough, Neighborhood, Latitude, and Longitude. It comprises of 5 Boroughs and 306 Neighborhoods. Head of the data looks like as follows:



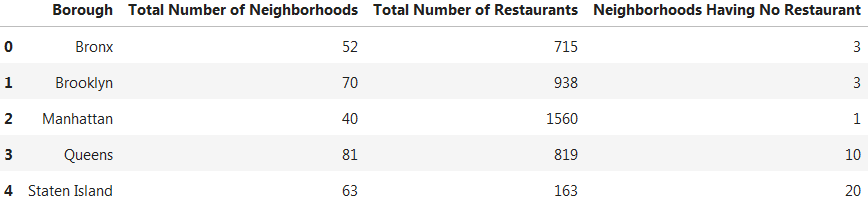
The folium library is used to create a map with neighborhoods superimposed on top.

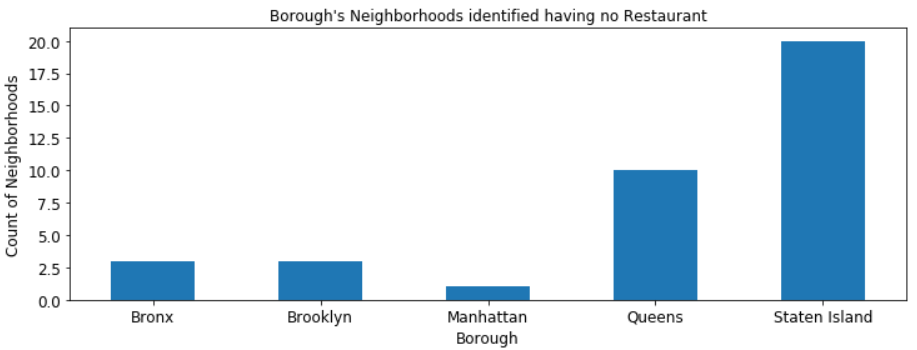


I then used the Foursquare API to search for the restaurants in the radius of 500 meters for each of the neighborhood. Below is the head of the list of venues having keyword ‘Restaurant’ or ‘restaurant’ in its name with coordinates as returned from the Foursquare API.



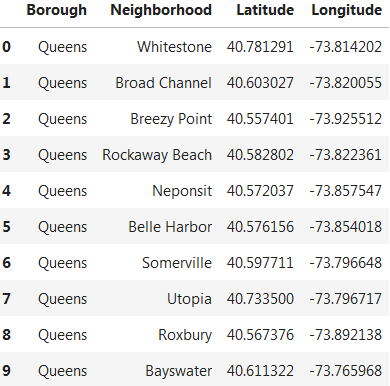
For some of the neighborhoods, no restaurant data has been returned. From this, we can derive the below analysis.





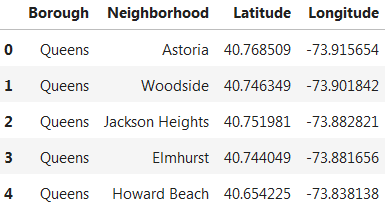
Boroughs ‘Staten Island’ and ‘Queens’ has many neighborhoods which have got no restaurants. However, we will use ‘Queens’ for our further analysis as it has got almost double the number of neighborhoods and half the number of neighborhoods as compared to ‘Manhattan’. We will skip ‘Staten Island’ as it is topographically and geologically a part of New Jersey [3] and can be studied later if needed.

The following neighborhoods of ‘Queens’ has been identified as having no restaurants.

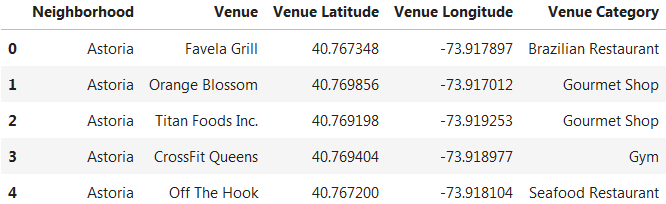


It may be possible that some of the above-identified neighborhoods have restaurants present in the radius of 500 meters. This is because we have considered checking for keyword ‘Restaurant’ or ‘restaurant’ in the ‘Venue’ name. There can be restaurants which do not suffice this above-stated requirement, however, there corresponding identified ‘Venue Category’ can fulfill this conditional requirement. For example, restaurant ‘Off The Hook’ is a restaurant, however, it won’t be picked as a restaurant as per the stated conditional requirement.

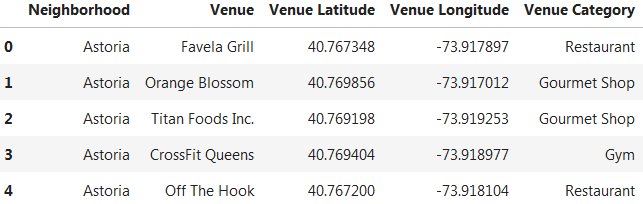
From the master data, ‘Queens’ data is filtered out. A head count of five of the lists of the data is as given below:



The explore function of the Foursquare API applied to the above data set returns the most common venue categories in each neighborhood. This is how the head of the data looks like:



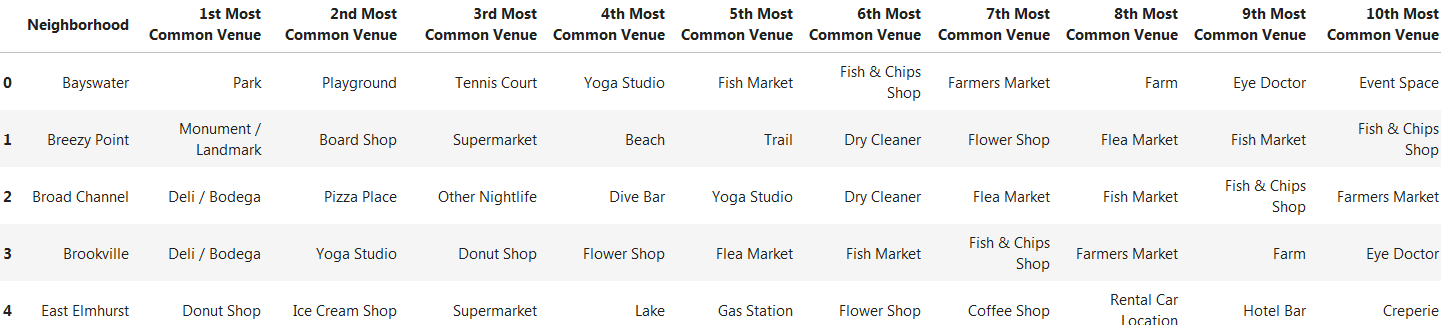
As we are targeting to find neighborhoods suitable for opening a restaurant, hence we will not focus on the type of restaurant for now. We will check if either ‘Venue’ or ‘Venue Category’ has keyword ‘Restaurant’ or ‘restaurant’ in its value. If this condition is met, then for the identified rows for the column ‘Venue Category’, its value will be replaced with the word ‘Restaurant’. The translated data will look like:



We get 206 unique categories on curating the ‘Venue Category’ data. From this, we extracted the top ten most common venues for each of the neighborhood and the resultant data looks something like as below:



We then get all the neighborhoods for which there is no restaurant listed in the ten most common venues. A head count of five rows is shown below:



The above-listed neighborhoods may have restaurants present but since restaurants are not that common as a venue category for such neighborhoods, we listed them as some of the probable locations for a new restaurant.

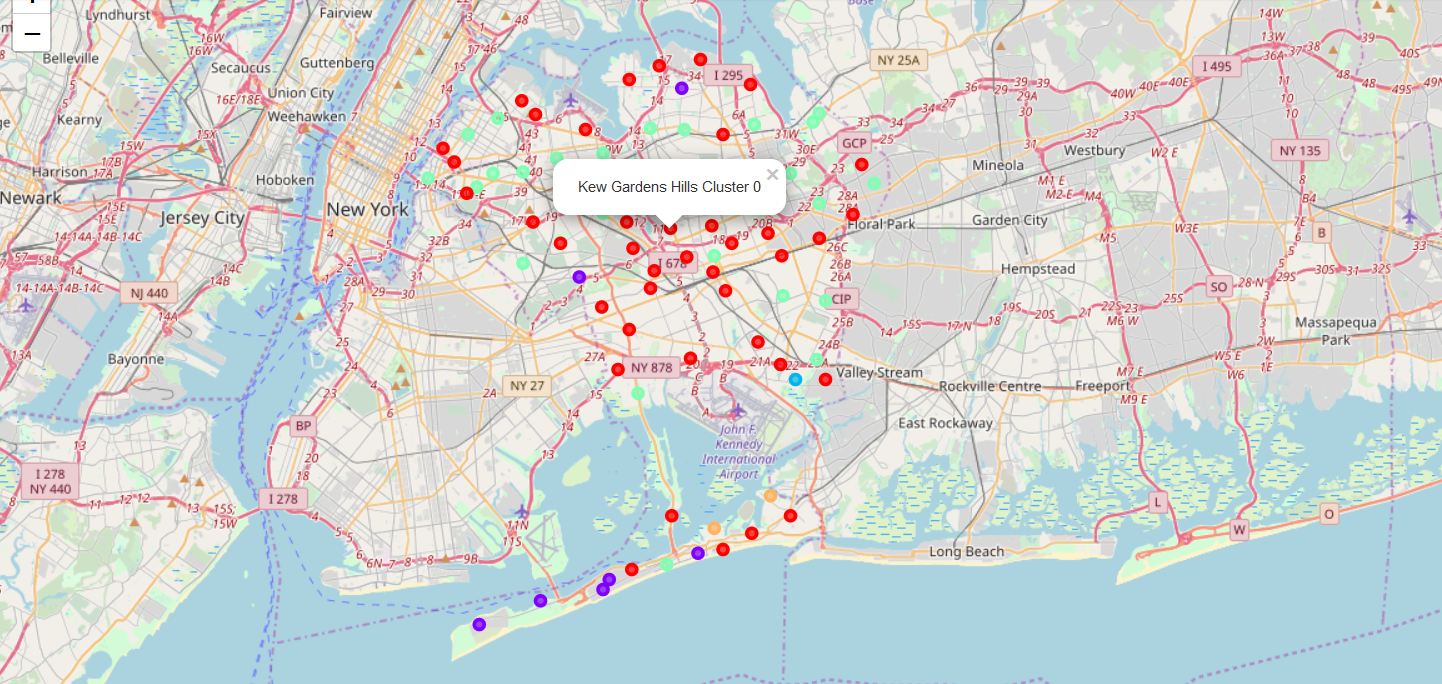
Finally, I used the k-means algorithm to group the top ten common values for all the neighborhoods for borough ‘Queens’. The reason behind using the k-means algorithm is to group the venue categories into clusters and derive some information relating to the restaurant from them.

1. Results

We got five clusters because of the application of the k-means algorithm. The resulting cluster column merged with top ten venues and the resulting data table from this is:



Also, a map view of the resulting cluster.



Most common venues for each of this five cluster is as below. Clusters 0 and 3 has got good number of restaurants.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |

1. Discussion

I took ‘Queens’ for my analysis. A similar approach can be applied to all the other remaining boroughs.

While searching for a restaurant using Foursquare API, it can be leveraged to find all such venue’s names too where keyword ‘Restaurant’ or ‘restaurant’ in the venue name doesn’t occur but can be identified as a restaurant with the help of that venue’s category.

Again, using Foursquare API, to explore the neighborhoods of ‘Queens’, only the top ten most common venues has been considered. An absence of restaurants in such neighborhoods doesn’t imply that the neighborhood is suitable for opening for a new restaurant. To consider such neighborhoods for opening a new restaurant, factors such as the size of the area and population also needs to be taken into consideration.

Lastly, I ended my analysis by analyzing the cluster derived from applying the k-means algorithm on the New York city map.

1. Conclusion

We are targeting to find neighborhoods where there is no restaurant for our minimum risk venture. However, the success of a restaurant depends on its type, nearby localities and taste preference of the people living there. Same neighborhood where there are already a good number of restaurants flourishing can be our location for a new restaurant in case we have some unique notions and great food to please the taste bud of the people and thus attract them.

I hope the above analysis can help someone to get some insight into what and where one should look for while opening a new restaurant and can leverage some benefits from the above-stated analysis.

1. References:

[1] [NYU Spatial Data Repository](https://geo.nyu.edu/catalog/nyu_2451_34572)

[2] [Foursquare API](https://developer.foursquare.com/)

[3] [Wikipedia – Staten Island](https://en.wikipedia.org/wiki/Staten_Island)