New York City Neighborhood Venue Analysis

Manthan Shah

August 3rd, 2019

1. Introduction

1.1 Background

New York city was called many names, Big Apple, Financial Capital of the World, the Garment Capitol, Wall Street, and many more. Jay Z even wrote a song about it where he refers to the city as “a concrete jungle where dreams are made of”. New York is called many things, but never has it been called slow, boring, or uneventful. One can find interesting places to visit, eat, drink, enjoy, in one of the five boroughs of the city, namely Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The city, knowing to be one of the top tourist destinations of the world, attracts a large number of travelers everyday who throng the city streets morning to night. This provides a good opportunity for a business-person to open a business in areas that are frequently visited by large groups of people, tourists. This helps the business-person personally and the nation economically.

1.2 Problem and Target Audience

Due to the vastness of the area and the ample locations available to visit, it can sometimes become quite intimidating for a new timer to know one’s way around the city. Many a times it can become quite confusing given with so many options to visit. It would be quite helpful to know which and where are interesting places to visit around the city. Having knowledge about different nearby locations might help one plan their day around the city.

Similarly, having knowledge about different types of locations around a neighborhood can help a business-person make an informed decision about the location where a new shop should be set up to incur profit.

This paper attempts to give a general idea of different types of venues situated across 300 neighborhoods spanning 5 boroughs of the city. The target audience is classified under two cohorts of population: firstly, people with little to no knowledge about the locations of different venues around the 5 boroughs of the city, and secondly, prospective business-people looking for locations to set up new or expand old businesses.

2. Data acquisition and cleaning

2.1 Data Sources

Firstly, to know the neighborhoods of data, spatial information (latitude and longitude coordinates) was needed. JSON (JavaScript Object Notation) was obtained containing spatial coordinates of 306 neighborhoods of New York City.

Secondly, and most important, source of data was FOURSQUARE. Its API (Application Programming Interface) was used extensively to obtain venue data using latitude and longitude coordinates obtained from the JSON file.

2.2 Data Collection Methodology

Firstly, the JSON file was downloaded and parsed. From it, 306 neighborhood's name, spatial coordinates, and the borough in which its situated was collected. Using this spatial data, and python's Folium mapping library, the neighborhoods of the city were mapped.

Then, using the spatial coordinates of the neighborhood and the developer account of FOURSQUARE, requests were made to FOURSQUARE's API by providing the neighborhood's spatial coordinates and user developer account credentials to obtain 100 venues situated within 500 meters of the spatial coordinates provided. These requests were answered with a JSON file containing information about 100 different venues (if present) situated in the spatial coordinates provided for a neighborhood and metadata associated with each venue, like **v**enue id, venue name, venue category, venue spatial coordinates, and more. Using the venue id, requests were further made to the API to obtain detailed metadata on each individual venue, which included, but not limited to, venue rating, venue tips, venue distance, venue menu, venue photos, venue likes, venue comments by users, user photos, etc.

A collective table of 6201 venues and 13 features was generated where each venue had information on its 13 features which provided information on venue's borough, venue's neighborhood, neighborhood latitude, neighborhood longitude, venue id, venue name, venue latitude, venue longitude, venue category, venue shortname, venue costliness, venue likes, and venue rating, respectively.

3. Exploratory Data Analysis

3.1 Relationship Between Borough and Neighborhood

Firstly, from the large collection of venues obtained from the 306 neighborhoods, it was time to see how the venues were spread out across the boroughs. Finding which boroughs contain more venues and different types of venues was important for further analysis. Exploring the data, it was found that Brooklyn and Queens were the neighborhoods which contained the maximum number of neighborhoods and thus, venues. It was evident due the size size of the boroughs that they contain the maximum number of venues.

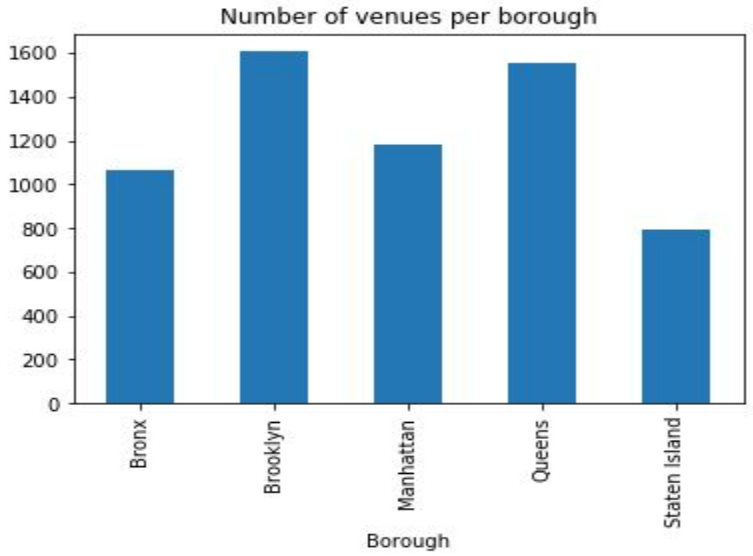

Figure 1. Bar-chart showing comparison of number of venues in each borough

3.2 Relationship Between Individual Borough and Venue Category

After finding the number of venues in each borough it was time to find what is the distribution of venues categories in each borough. Question like, is a specific borough having abundance of a certain type of venue, can be answered from this analysis. In order to perform this analysis, data was segmented on each borough and top 10 venue categories were extracted, grouped, and counted. Following below are bar charts representing venue categories in each borough in decreasing number of frequency. It was not uncommon to find that “Pizza Place” was the most abundant venue category in 4 out of 5 boroughs, knowing the fact that New York city is known for its Pizza places.

3.2.1 Relationship Between Bronx Borough and Venue Categories

Bronx borough is considered more a residential neighborhood. Hence, places like delis, pharmacies, restaurants, donut shops, supermarkets, etc. are abundant in neighborhoods of this borough. Businesses that tend to daily working class community’s needs are evident to flourish here.

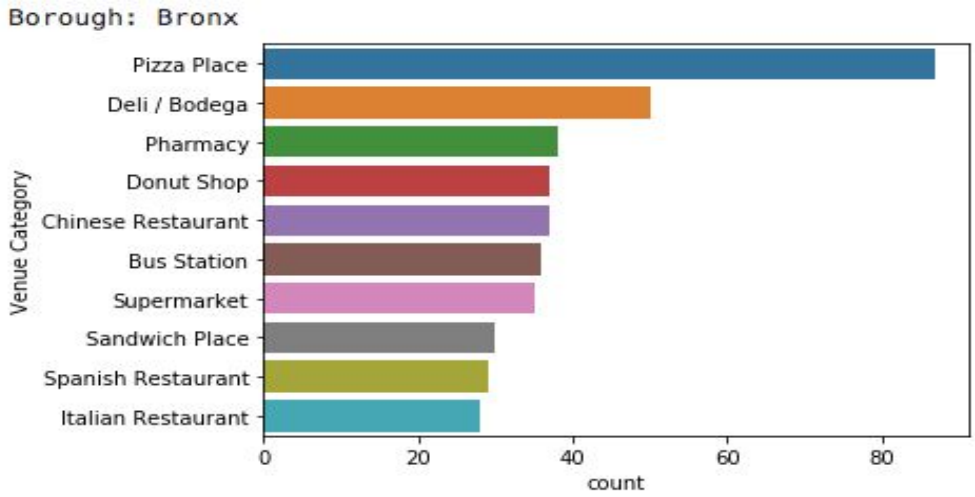

Figure 2. Venue Category bar chart for Bronx Borough

3.2.2 Relationship Between Brooklyn Borough and Venue Categories

Brooklyn is known for its hangout spots near the Brooklyn bridge, wood-burnt pizzeria's and its bars. This chart below shows that there is a uniform distribution of large number of hang-out spots available in and around Brooklyn brough.

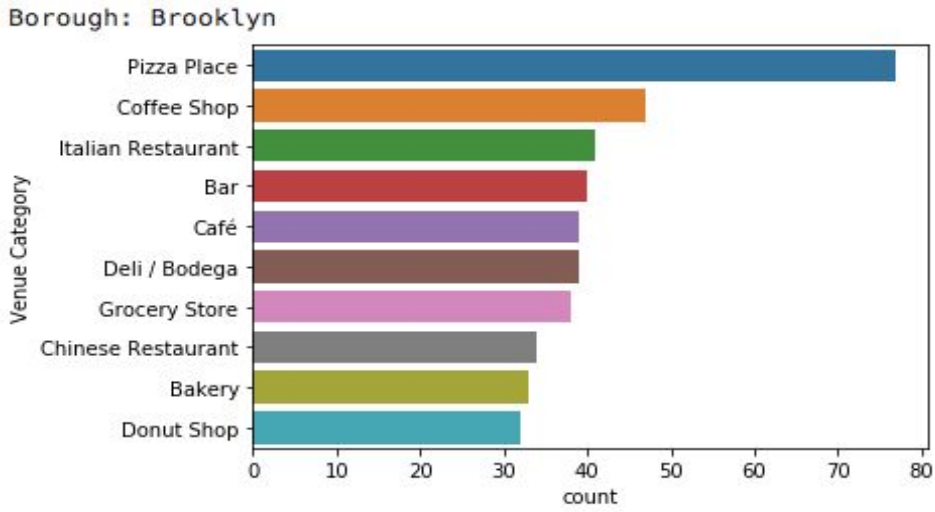

Figure 3. Venue Category bar chart for Brooklyn Borough

3.2.3 Relationship Between Manhattan Borough and Venue Categories

Manhattan is the heart of New York city. It contains a balance of please and business. It is frequented daily by working class people as well as tourists. Checking the top 5 abundant venue categories shows that they fall under dining places and park which are abundantly available across the borough. Parks are more tourist focused but the cafes, restaurants, coffee shops are frequented equally by both people who come to work in Manhattan as well the tourists. Health conscious and job-working people find it efficient and prefer to go to a gym near their workplaces than near their home as this helps destress and remove the possibility of not working-out after travelling back home. Thus we find gym, twice, in the top 10 categories.

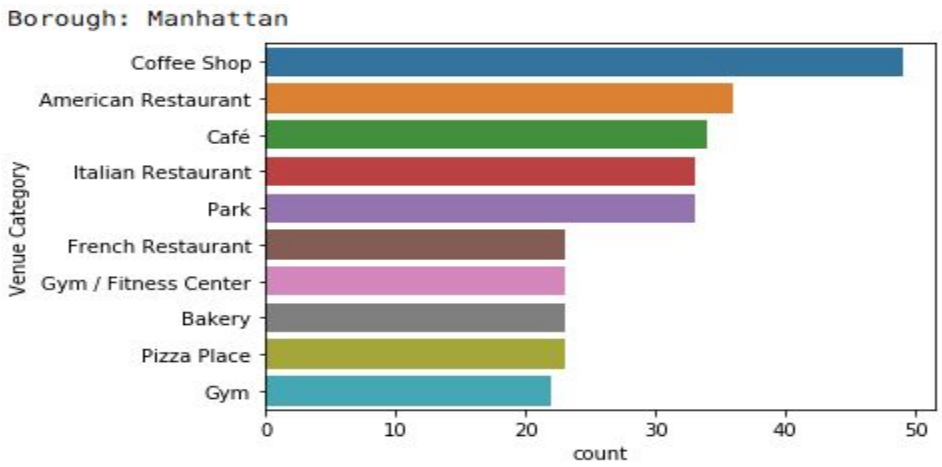

Figure 4. Venue Category bar chart for Manhattan Borough

3.2.4 Relationship Between Queens Borough and Venue Categories

Queens, again is a residential pro borough, just like the Bronx. Hence, it is evident to find restaurants, delis, pharmacies prevalent more in this borough. Contrary to Bronx, Queens contains almost equal amount of deli shops to pizza places. Seems, like residents of Queens like to eat pizza more than the residents of Bronx.

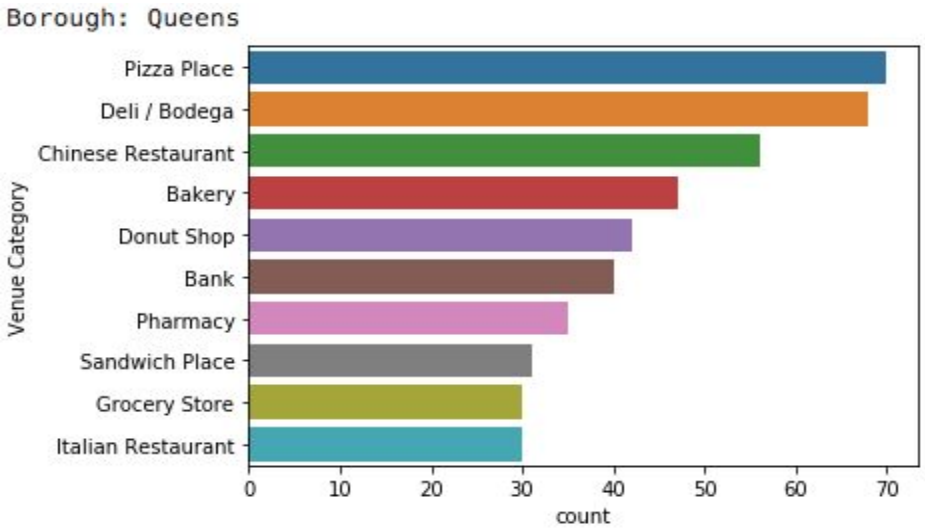

Figure 5. Venue Category bar chart for Queens Borough

3.2.5 Relationship Between Staten Island Borough and Venue Categories

Staten Island, similar to Bronx and Queens, is a residential pro borough. Finding similar venue categories like restaurants, delis, pharmacies, banks, to Queens and Bronx is not uncommon. As Staten Island, is literally an island, people have to depend majorly on buses for intra-borough public transportation as opposed to the availability of subways in all other boroughs. This is shown by the number of high number of bus stops in the neighborhoods of the borough.

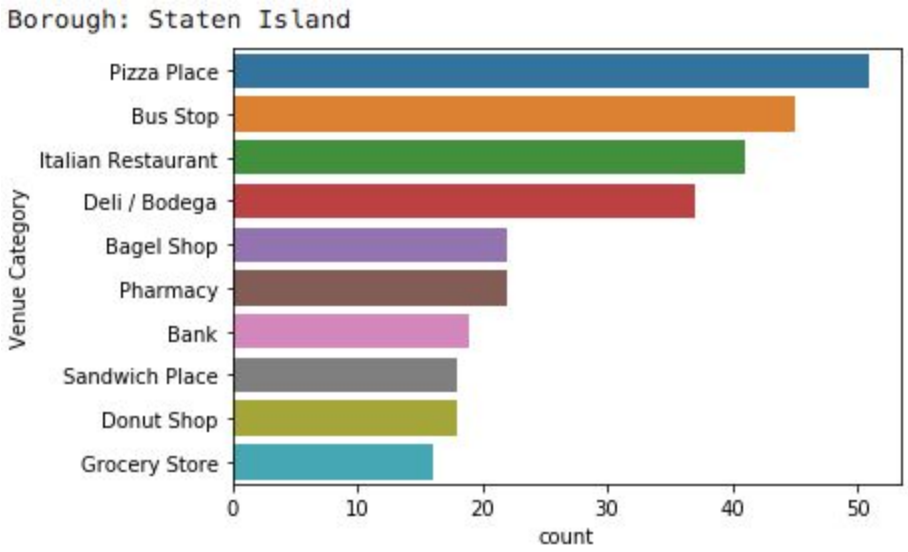

Figure 6. Venue Category bar chart for Staten Island Borough

3.3 Relationship Between Boroughs and Top 20 Venue Categories

After exploring individual boroughs, it was time to check how all borough compare across different venue categories. Data was formatted to display the top 20 abundant venue categories and plotted bar-chart showing how each borough compares to the group.

Few insights that stand out from this bar-chart are:

1. Compared to other boroughs Manhattan contains low number of cafe's, bank, pharmacies, donut shops, and supermarkets. Assuming banks and supermarkets require ample space s=to set up shop and the property rates of Manhattan being high, it can be inferred that there are less number of them present. For prospective business, it is viable to open bagel shops as they are in scarcity in Manhattan.
2. There are not many cafes seen across Queens, Bronx and Staten Island. One can profit greatly by setting shop in dense neighborhoods of these boroughs.
3. Bakery shops have prospects of flourishing in Staten Island
4. Bagel shops in Bronx can earn decent profit as seen from the chart below

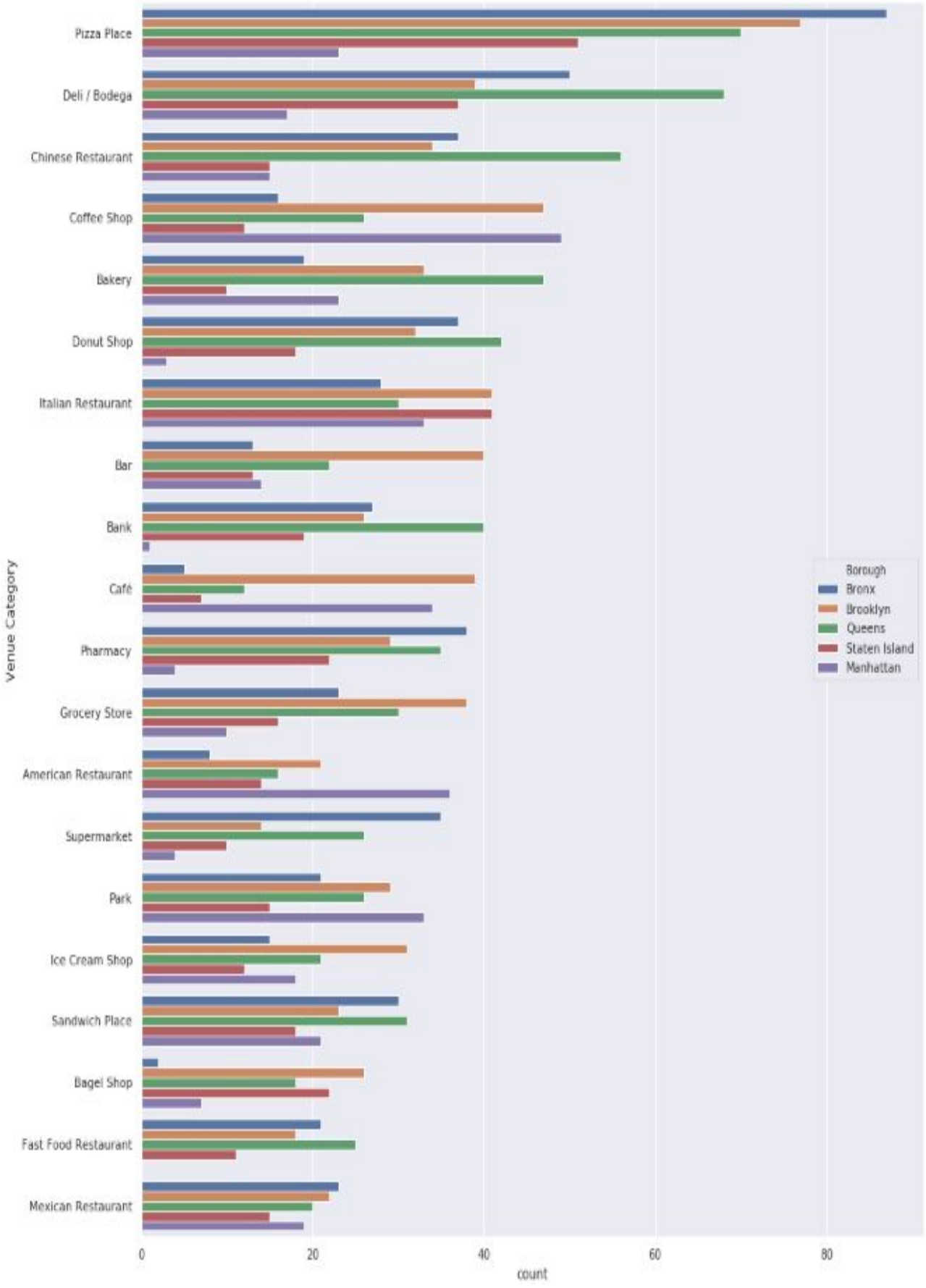

Figure 7. Top 20 Venue categories by borough comparison

4. Data Manipulation and Machine Learning

After exploring the data, then was time to perform machine learning and clustering venues across neighborhoods. We applied KMeans CLustering Algorithm to cluster venues across the city. Before, applying the algorithm the data had to be converted. Based on the data collected, the venue category had 373 different values. One-Hot Encoding was applied and KMeans clustering algorithm was run on it. Based on elbow method, number of clusters was set at 5 and algorithm was fit using the data. As a result, cluster labels were generated and assigned to the 300 neighborhoods.

The one-hot encoded data was further used to find the top 10 most abundant places associated with each neighborhood. A table was generated and used to denote top 10 frequent venue category in decreasing number of their counts for each neighborhood.

Due to previously available spatial coordinates of each neighborhood, and now the cluster labels, it was possible to visualize the clusters across the city.

5. Results

On plotting the cluster labels obtained from applying KMeans algorithm onto city's map it is clearly shown that the clusters are not separable from each other. This describes that the venue categories are almost mutually inclusive for majority of the categories. There are some categories which differentiate different clusters.

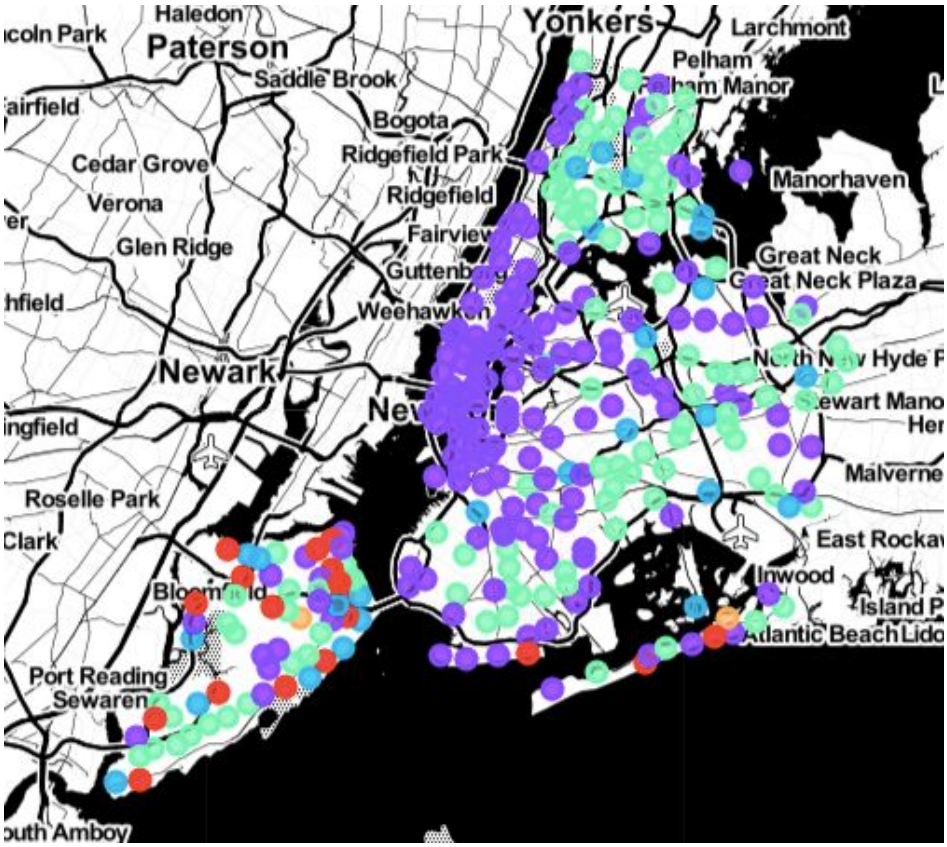

Figure 8. Cluster visualization

As this result was not much informative, generating information on the venue categories inside each cluster might help generate better segmented clusters. Generation of word clouds based on the top 10 most occuring venue category in each cluster was decided. The data was first subset to cluster label and the top 10 locations were extracted, manipulated and converted to a text file. The text fill was then input to wordcloud and visual representation of density of each venue category in a given cluster was depicted.

5.1 Cluster 1 analysis

Cluster one contained 18 neighborhoods,which were spread out across 5 boroughs in ratio: Staten Island (0.83%), Queens (0.11%), Brooklyn (0.056%). As seen here, major part of the neighborhoods were segregated in Staten Island area, hence we see bus-stops, restaurants, factory as abundant venue categories in the word cloud.

Figure 9. Cluster 1 word cloud

5.2 Cluster 2 analysis

Cluster two contains 148 neighborhoods,which are spread out across 5 Boroughs in ratio of: Brooklyn (0.32%), Manhattan (0.27%), Queens (0.25%), Staten Island (0.1%), Bronx (0.07%). As this cluster is majorly concentrated on Manhattan, Bronx, and Brooklyn boroughs, the word cloud contains majority of hang-out spots and dining places.

Figure 10. Cluster 2 word cloud

5.3 Cluster 3 analysis

Cluster three contains 26 neighborhoods,which are spread out across 5 Boroughs in ratio of: Staten Island (0.38%), Queens (0.31%), Bronx (0.23%), Brooklyn(0.8%). Majority of the neighborhoods situated in residential boroughs, this word cloud depicts majority of restaurants and delis in the neighborhoods.

Figure 11. Cluster 3 word cloud

5.4 Cluster 4 analysis

Cluster four contains 110 neighborhoods,which are spread out across 5 Boroughs in ratio of: Bronx(0.32%), Queens (0.30%), Staten Island (0.20%), Brooklyn(0.18%). This cluster is spread out across almost all boroughs except Manhattan.

Figure 12. Cluster 4 word cloud

5.5 Cluster 5 analysis

Cluster five is divided across two boroughs namely, Staten Island and Queens, in equal 50-50% ratio. Depicting restaurants and deli being the primary venue category.

Figure 13. Cluster 5 word cloud

6. Conclusion

In this study, the focus of the analysis was to find out the distribution of different categories of venues in the 5 boroughs of New York City. Based on the data obtained from the FOURSQUARE api, and clustering using KMeans algorithm insights were discovered pertaining to individual boroughs, specific types of venues in different neighborhoods, and comparison of different venues across 5 boroughs. Some insights derived from the study suggested that one will have ample dining options if present in Bronx, Queens or Staten Island than anywhere else; opening a bakery shop in Staten Island, cafes in Bronx, Staten Island and Queens, and bagel shop in the Bronx can be a profitable business; one will likely find bar in Brooklyn easily when compared to other 4 boroughs; and finally irrespective of the borough one can find a pizza place much easily than any other eatery. Insights like this can help a tourist make informed decision while travelling through the city. Similarly, this can help business-persons on prospective business opportunities to expand or generate new businesses.

7. Discussion

On seeing the spread of the clusters and the variability of different venue categories, there is a possibility of increasing the cluster spread by using more features and data. For future steps following things can be implemented/improved:

- use FOURSQUARE's api to obtain additional information of the location like ratings, likes, number and distance of other locations nearby to gauge the popularity level of the location
- use housing data to know if property rates affected by the venues in the neighborhood
- use crime data to find which locations to avoid for a prospective business-person opening shop
- use other clustering algorithms like DBSCAN
- use comments given by users for individual venues to personalize location reporting to users as per user's preference

Applying, more functionalities and acquiring more data will improve model's understanding of the clusters, thereby providing better venue segmentation and reporting.