Applied Data Science Capstone Course

Capstone Project – Final Report

Battle of the Neighborhoods: Manhattan vs. Queens

1. Introduction

1.1 Problem

Currently, the author resides in the Lenox Hill neighborhood of Manhattan. However, the author's lease is up soon, and the author and her roommate are faced with the question of whether they should try living somewhere new. They have settled on a couple of options: another neighborhood in Manhattan or a neighborhood in Queens.

1.2 Criteria to Consider

The author and her roommate really like Lenox Hill, so they would like to compare the other options to their current neighborhood to ensure they end up somewhere relatively similar. However, there are a few things that are important to have near them that will be focused on: Bagel shops, Italian and Ramen restaurants, Art Museums, and Wine Shops. Lenox Hill does not offer all of these things, and that is one of the reasons a move is being considered.

1.3 Target Audience

The target audience of this project is the author's roommate who is convinced all venues types in question could be found in a neighborhood in Queens and would like the author to prove or disprove that belief.

2. Data

To solve this problem, a dataset provided by New York University will be used. This data set provides information on Manhattan and Queens that outlines the state-recognized neighborhoods located in each borough. This data set can be found at "https://geo.nyu.edu/catalog/nyu_2451_34572". Foursquare location data will also be used in order to locate and sort through the myriad venues available in each of these neighborhoods. For example, if the NYU dataset for Queens includes Astoria, a built in Python package will be used to return the latitude and longitude coordinates for Astoria, Queens. Those coordinates will then be

utilized to get information on the venues in those neighborhoods from Foursquare. Once the venue data has been collected, further investigation can be done into which neighborhoods, if any, meet all five requirements (Bagel shops, Italian and Ramen restaurants, Art Museums, and Wine Shops), as well as which neighborhoods are similar to Lenox Hill. Unsupervised machine learning can then be carried out on the resulting data set.

3. Methodology

3.1 Preparation:

The first step in this analysis was to install the relevant libraries. These included NumPy, pandas, matplotlib, and several others for json manipulation and K-Means analysis. The next step was to clean the NYC json dataset. Column headers were added, as well as latitude and longitude coordinates for all of the neighborhoods in the dataset. Then Foursquare credentials were added as variables so they could be pulled more easily.

3.2 Queens – Prep the data and Visualize

The Queens data had to be extracted from of the NYC dataset because the dataset contained data for all five boroughs. Then it was necessary to "get" the venue data from Foursquare for all the neighborhoods. To do this, a URL using the Foursquare credential variables was created, and Foursquare was asked to return venues of each type up to a limit of 100 within a radius of 500. Then folium was used to create a map to visualize the neighborhoods in Queens. Then onehot encoding was used to transform the results to a count of each type of venues for each neighborhood and the result were grouped by neighborhood.
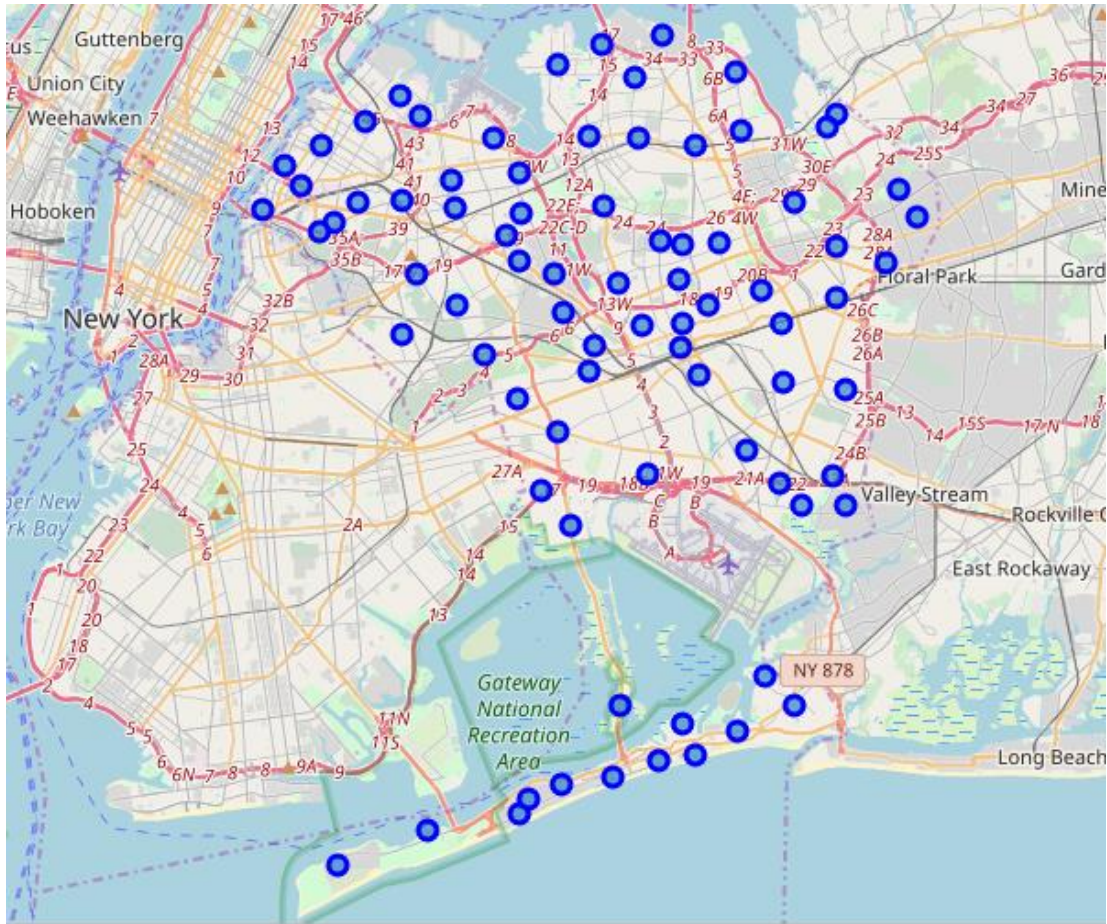
3.3 Queens – Analyze the Individual Neighborhoods

To take a closer look at the five types of venues in question, the columns for Bagel shops, Italian and Ramen restaurants, art museums, and wine shops were extracted and placed into a new data frame. A total column and row were added to help with comparing the neighborhoods. Since there isn't a single neighborhood that contained all five types of venue, the analysis was expanded to three of the five. This narrowed the list to two neighborhoods – Ravenswood and Hunters Point.

Using the .loc command, more comprehensive data was pulled on those two neighborhoods. The more comprehensive data for each of the two neighborhoods was then examined to see what each neighborhood had to off in terms of variety of venues beyond the five types of venue on which the analysis was focused.

3.4 Queens – Clustering using K-Means (Machine Learning)

In order to conduct the second part of the analysis, determining which neighborhoods are similar to Lenox Hill, it was necessary to cluster the neighborhoods. This was done using K-Means Clustering, which is an unsupervised machine learning technique. Then folium was used to create a visual – a map of Queens showing the clusters of neighborhoods. The results of this analysis were then compared to the results of the venue analysis to determine which, if any, neighborhoods, met both sets of criteria.
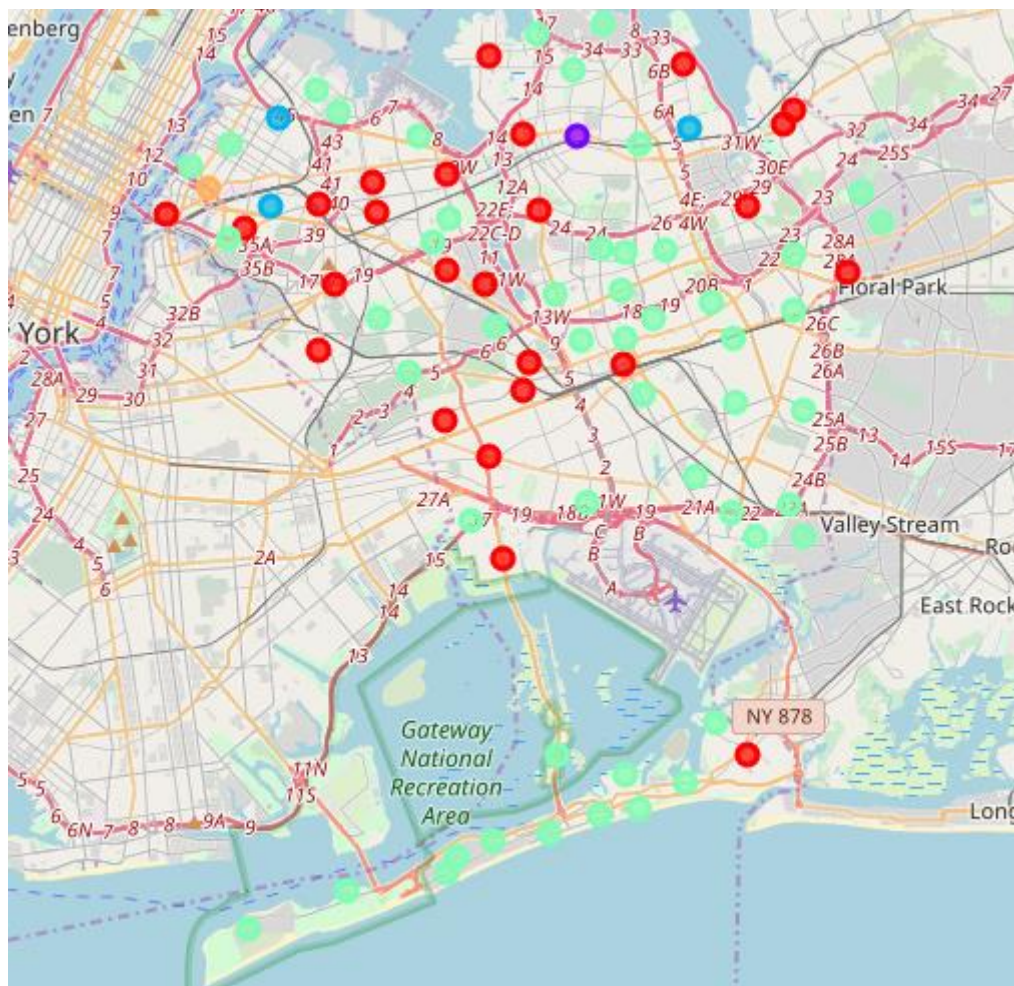

3.5 Manhattan – Prep the data and Visualize

The Manhattan data had to be extracted from of the NYC dataset as was done with Queens because the dataset contained data for all five boroughs. Then it was necessary to "get" the venue data from Foursquare for all the neighborhoods. To do this, a new URL using the Foursquare credential variables was created, and Foursquare was asked to return venues of each type up to a limit of 100

within a radius of 500. Then folium was used to create a map to visualize the neighborhoods in
Manhattan. Then onehot encoding was used to transform the results to a count of each type of
venues for each neighborhood and the result were grouped by neighborhood.
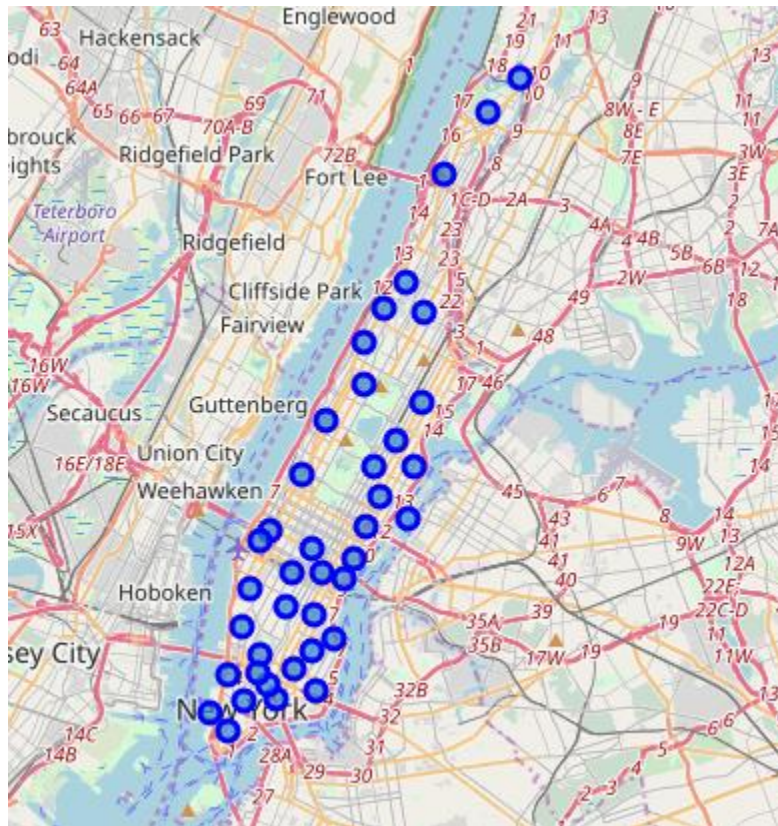

3.6 Manhattan – Analyze the Individual Neighborhoods

To take a closer look at the five types of venues in question, the columns for Bagel shops, Italian and
Ramen restaurants, art museums, and wine shops were extracted and placed into a new data frame.
A total column and row were added to help with comparing the neighborhoods. Since there was
only one neighborhood that contained all five types of venue, the analysis was expanded to four of
the five. This brought the list to three neighborhoods – Carnegie Hill, Grammercy, and Greenwich
Village.

Using the .loc command, more comprehensive data was pulled on those three neighborhoods. The
more comprehensive data for each of the three neighborhoods was then examined to see what each
neighborhood had to off in terms of variety of venues beyond the five types of venue on which the
analysis was focused.

3.7 Manhattan – Clustering using K-Means (Machine Learning)

In order to conduct the second part of the analysis, determining which neighborhoods are similar to Lenox Hill, it was necessary to cluster the neighborhoods. This was done using K-Means Clustering, which is an unsupervised machine learning technique. Then folium was used to create a visual – a map of Manhattan showing the clusters of neighborhoods. The results of this analysis were then compared to the results of the venue analysis to determine which, if any, neighborhoods, met both sets of criteria.
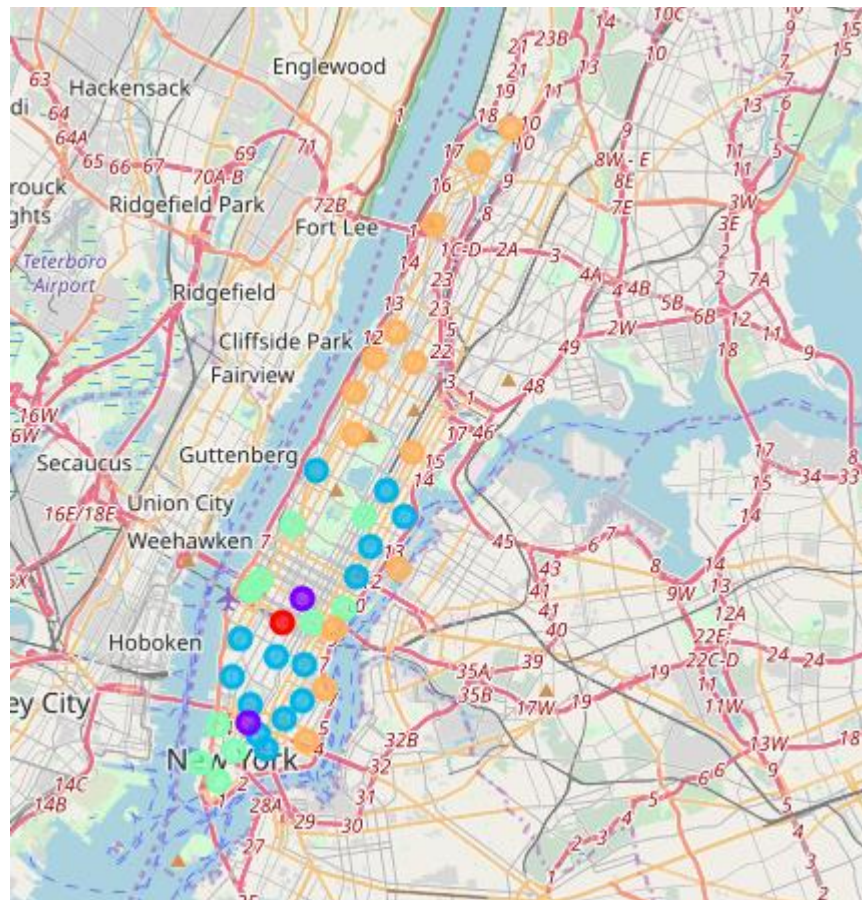
4. Results

4.1 Queens – Initial Analysis

The initial analysis of the neighborhoods in Queens revealed that there is not a single neighborhood in Queens that possesses all five required types of venue. There was also not a single neighborhood

that possessed four of the five. The analysis resulted in the selection of two possibilities on which to focus which each contained three of the 5 types of venue: Ravenswood and Hunters Point.

4.2 Queens – Further Exploration

When a closer look was taken at Ravenswood and Hunters Point, initial analysis revealed that while Ravenswood possessed three of the five required venues, it possessed little else and had little offer. Hunters Point was very different appeared to have a lot more options in comparison. Neither neighborhood appeared to be particularly similar to Lenox Hill.

4.3 Queens – Clustering Results

The K-Means Clustering Analysis clustered the neighborhoods of Queens into five clusters. However, all but a handful of neighborhoods fell into only two of those clusters. Ravenswood and Hunters Point were assigned to different clusters, supporting the conclusion of the initial analysis that they were very different neighborhoods. The majority of the neighborhoods not assigned to the two primary clusters fell closest to Manhattan. This suggests that the bulk of the neighborhoods of Queens are not similar to Manhattan, or by extension to Lenox Hill, as the Manhattan overflow neighborhoods make up their own clusters.

4.4 Manhattan – Initial Analysis

The initial analysis of the neighborhoods in Manhattan revealed that there is only one neighborhood which possesses all five required venues, and two neighborhoods that possess four of the five in abundance. These three neighborhoods (Carnegie Hill, Gramercy, and Greenwich Village) were selected for more in-depth analysis.

4.5 Manhattan – Further Exploration

Upon closer analysis, Carnegie Hill appears to have the greatest variety of venues in addition to being the only neighborhood offering all five required venue types. Gramercy and Greenwich appear to be very similar to each other and don't offer much beyond the four of the five required venues they each possess.

4.6 Manhattan – Clustering Results

The K-Means Clustering Analysis also clustered the neighborhoods of Manhattan into five clusters. Unlike the mix provided by Queens, the neighborhoods of Manhattan are much more evenly

dispersed among the five clusters. All three Manhattan neighborhoods in question, as well as Lenox Hill (the author's current neighborhood) were placed in the same cluster. This suggests that all of the neighborhoods are similar to Lenox Hill.

5. Discussion

5.1 Observations and Findings

There were two criteria to be considered in this analysis. The first was whether a neighborhood possessed five specific types of venue: Bagel shops, Italian and Ramen restaurants, Art Museums, and Wine Shops. The second was whether the neighborhood was similar to Lenox Hill, the neighborhood in which the author currently resides. The initial, individual analysis focused on whether each neighborhood possessed the five venues, and the K-Means Clustering determined which neighborhoods were similar to Lenox Hill.

The results of the individual analysis showed that only one neighborhood between both boroughs, Carnegie Hill, possessed all five venues. If the requirements are relaxed so more neighborhoods can be considered, there are two neighborhoods with four of the five venues and both are in Manhattan. The best possible options in Queens possess only three of the five venues. This suggests that remaining in Manhattan might be a better option than moving to Queens.

The results of the K-Means Analysis shows that all three of the neighborhoods in Manhattan meet the requirement of being similar to Lenox Hill. There were no neighborhoods in Queens that appear to meet this requirement.

5.2 Recommendation

The author and the author's roommate should remain in Manhattan. Queens doesn't meet either of the two requirements out forward in the problem. Of the three Manhattan neighborhoods to come into focus in the individual analysis, Carnegie Hill appears to be the best choice as it is the only option to offer all five required types of venue, it offers a variety of venues beyond those five, and is similar to Lenox Hill.

6. Conclusion

The author and the author's roommate are trying to decide whether they should move to a new neighborhood in Manhattan or move to Queens. The two criteria being considered are whether a

neighborhood possess five specific types of venues and whether a neighborhood is similar to Lenox Hill, the neighborhood where the author currently resides. A dataset containing all the neighborhoods of the five boroughs of New York City, provided by New York University, was used to star this analysis. The Foursquare Location dataset was then used to find venue data for the boroughs of Manhattan and Queens. An analysis was then done on the individual neighborhoods to find which neighborhoods, if any, in each borough possessed the five required venues. Further exploration was then done on each neighborhood selected to see what else it had to offer in terms of variety. Finally, a K-Means analysis was done on each borough to cluster similar neighborhoods and determine which neighborhoods, if any, were similar to Lenox Hill. The analysis determined that the author should remain in Manhattan and consider the Carnegie Hall neighborhood.