**The Battle of the Neighborhoods: Brooklyn Edition**

Kelly Gunneson

August 14th, 2020

**1. Introduction**

**1.1 Background**

Your friend is considering opening a coffee shop in the Brooklyn borough of New York City, NY. Brooklyn is the up and coming borough of New York City and they are requesting help selecting a neighborhood to open their shop in to be successful. They have asked for your assistance with your expertise in data analytics with Python to help them select a location for an increased likelihood of success.

**1.2 Problem**

The problem you need to solve is to select a neighborhood to recommend your friend open the coffee shop in. You should employ data analytics with Python to select the optimal neighborhood for the coffee shop based on neighborhood segmenting and clustering as well as analysis of the types of venues in the neighborhood.

**1.3 Interest**

The target audience of this problem is your friend who is opening the coffee shop as well as any investors or stakeholders involved in the opening of the coffee shop. This presentation will provide a recommendation for the neighborhood to open the coffee shop in as well as provide the documentation of the data analysis performed to inform the recommendation. Your friend, investors, and stakeholders will care about the recommendations and the supporting analysis because it can make them confident that they are making a data informed decision optimizing their success.

**2. Data Acquisition and Cleaning**

**2.1 Data Sources**

For the data to solve this business problem, we will use the dataset of New York City neighborhoods and boroughs at the following link of data collected and stored as a shapefile by NYU. <https://geo.nyu.edu/catalog/nyu_2451_34572>

**2.2 Data Cleaning**

The data is first extracted from the url and then looped to create a data frame from the features key filling in one row at a time ensuring that the data from all 5 boroughs and 306 neighborhoods is included. This transforms the data into a workable form for analysis in with Pandas in Python. Additionally, this problem requires the use of data only for the borough of Brooklyn. The data frame is sliced to create a new data frame only including Brooklyn’s data. Then, Foursquare API is utilized to access venue data regarding the neighborhoods in Brooklyn. The original dataset is utilized to develop a json Folium map in combination with the Foursquare data to cluster the venues and determine the most common venue in each cluster.

**2.3 Features Selection**

This dataset consists of the 306 neighborhoods in New York City including the neighborhood name, borough, latitude, longitude, geometry type, and annotation. We can extract each neighborhood including the neighborhood name, borough, latitude, and longitude into a Pandas data frame. This data will allow for the neighborhoods to be plotted on the map and grouped by borough. Then, we will filter the dataset and data frame to only include the Brooklyn borough. This resultant dataset and data frame can be utilized with the Foursquare data of venues to analyze each neighborhood and make a recommendation.

**3. Exploratory Data Analysis Methodology**

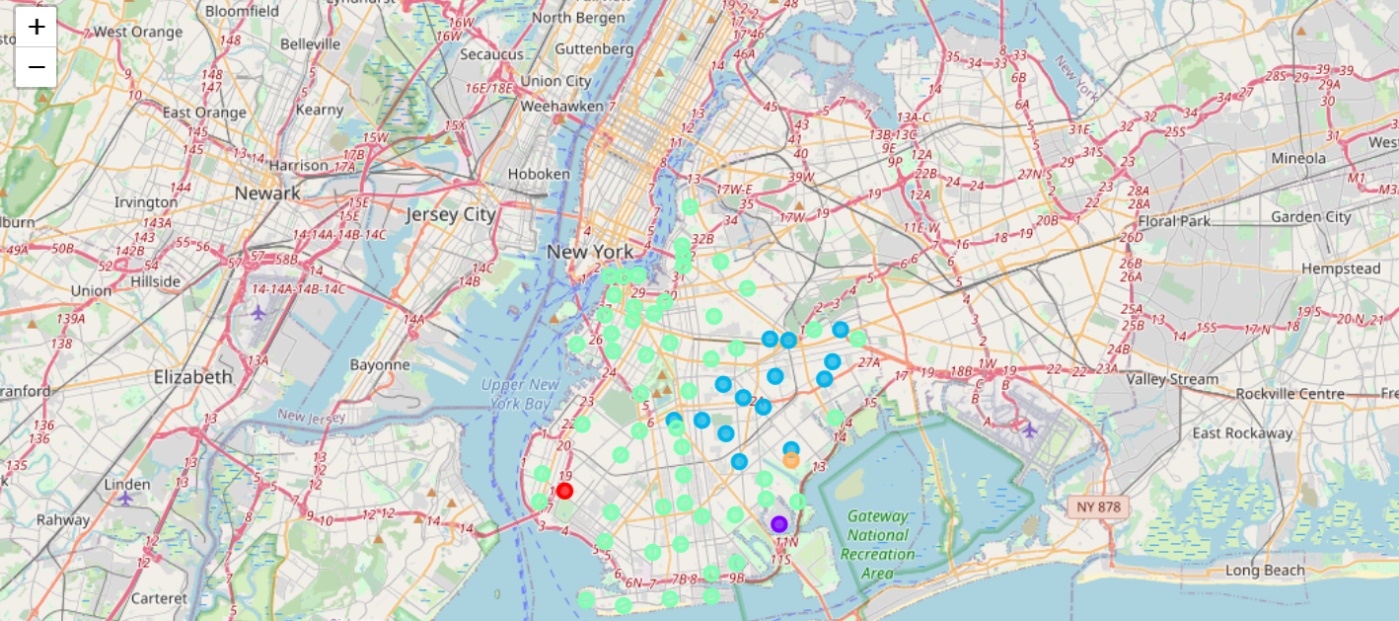
Here is the link to the Jupyter notebook on Github containing the full data analysis <https://github.com/kgunneson/Coursera_Capstone/blob/master/The%20Battle%20of%20the%20Neighborhoods%20(5).ipynb>

The following is an explanation of the methodology used in the data analysis.

**3.1 Foursquare Venue Data**

After the data imported, cleaned, and the maps are created, then you are ready to begin data analysis. First, the Foursquare data of the venues in Brooklyn can be imported and analyzed to determine the number of unique venues in the borough. Then, the groupby method can be applied to determine the frequency of the occurrence of each type of venue in each neighborhood. This allows you to determine the top five and ten most frequently occurring venue types for each neighborhood. With this, you can determine the most frequent types of venues in each neighborhood and begin to gain perspective regarding the characteristics of the neighborhood.

**3.2** **K-Means Clustering Map**

Next, k-means clustering is utilized to cluster the neighborhoods in Brooklyn into five groups. Then, the clustered neighborhoods are color coded on the map below with the colors light green, red, blue, orange, and purple.

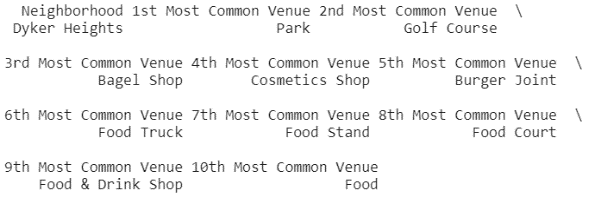
Cluster 1 is represented by the red dot. Cluster 2 is represented by the purple dot. Cluster 3 is represented by the blue dots. Cluster 4 is represented by the green dots. Cluster 5 is represented by the orange dot.

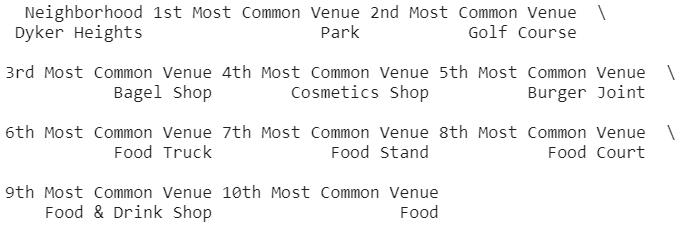
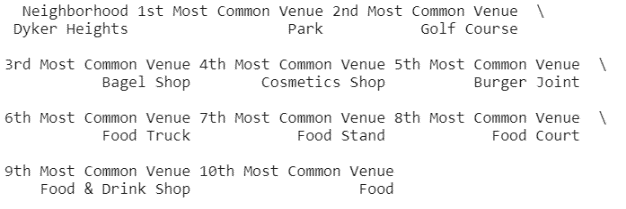
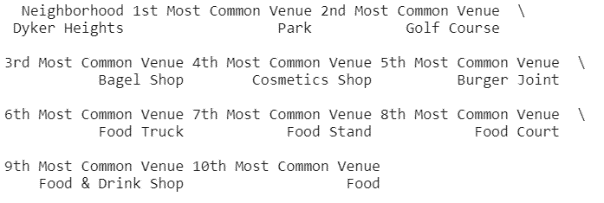
**3.3 Top 10 Venues for Each Cluster**

Then, data frames were constructed for each neighborhood cluster regarding the top ten most frequent venue types in each cluster. This will allow for the revelation of the trends in the neighborhood to determine where a coffee shop will likely do well.

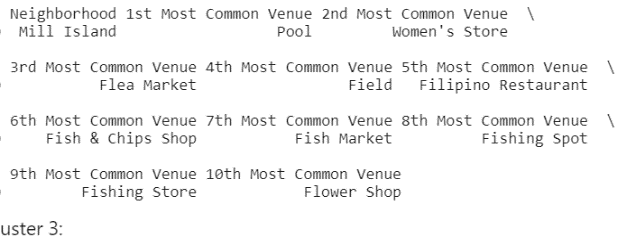
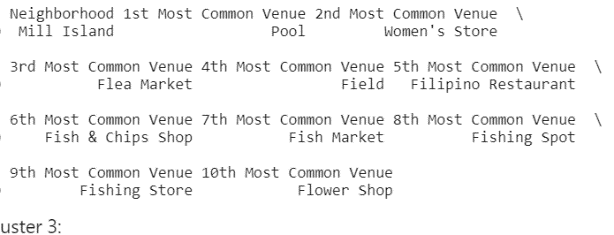
**4. Results**

The following page of results contains a summary of the neighborhoods and the top ten venue types in each neighborhood for each cluster. These resultant data frames can be utilized to observe commonalities in the clusters to determine the optimal neighborhood to open a coffee shop in Brooklyn.

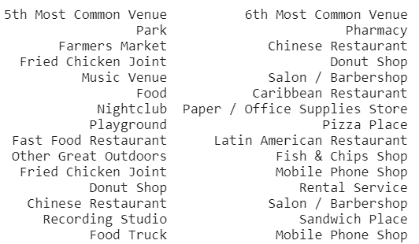
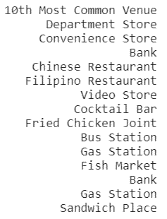
**Cluster 1:**

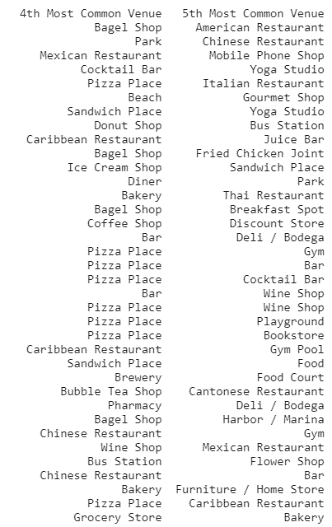
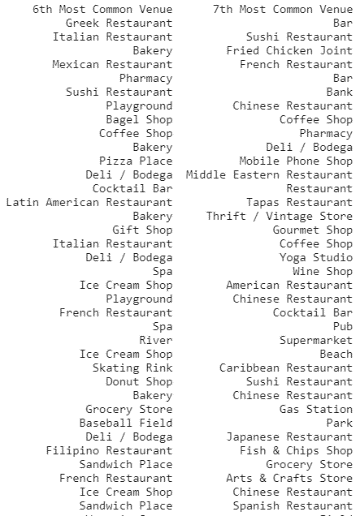
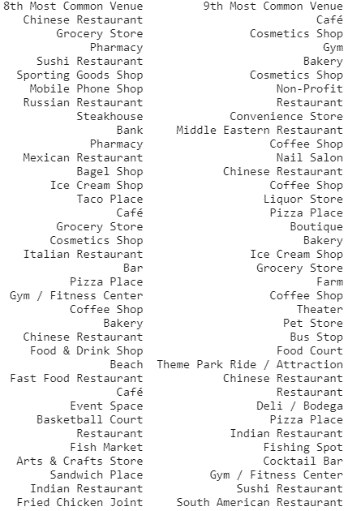


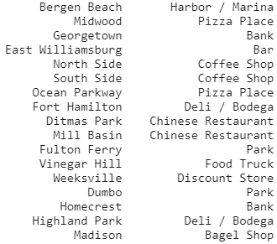
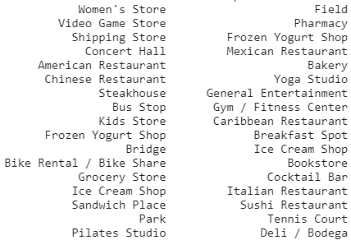
**Cluster 2:**



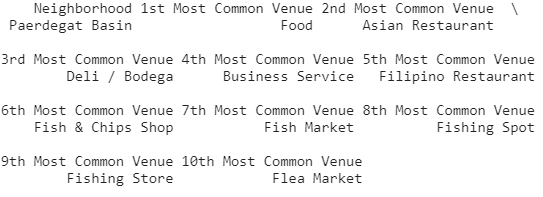
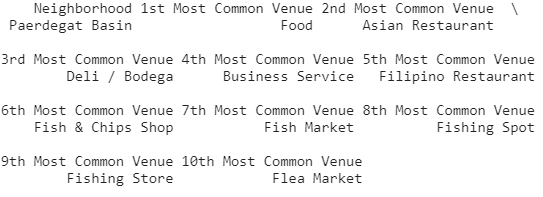
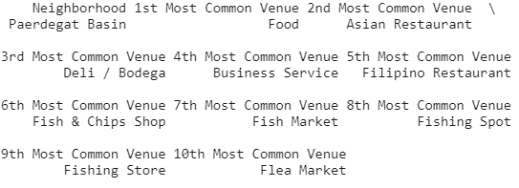
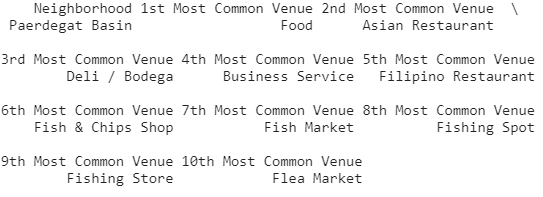
**Cluster 3:**



**Cluster 4:**



**Cluster 5:**



**5. Discussion**

**5.1 Observations**

Based on the analysis of the above results, you can observe trends in the types of venues that are successful in each cluster. The number of neighborhoods contained in each cluster as well as the most common venues are listed in the table below.

|  |  |  |
| --- | --- | --- |
| **Cluster #** | **Number of Neighborhoods Included** | **Most Common Venue Types** |
| Cluster 1 | 1 neighborhood | Park |
| Cluster 2 | 1 neighborhood | Pool |
| Cluster 3 | 14 neighborhoods | Various types of restaurants including Latin American, fast food, Chinese, Fried Chicken, and a Caribbean restaurant. |
| Cluster 4 | 53 neighborhoods | Various restaurants including Italian restaurants, pizza places, Chinese restaurants, cafes, delis/bodegas, and a few coffee shops |
| Cluster 5 | 1 neighborhood | Food |

Cluster 1, 2, and 5 include only one neighborhood and thus do not allow us to develop strong observations. Cluster 3 consists of 14 neighborhoods. The most common venue type in these neighborhoods are various types of restaurants including Latin American, fast food, Chinese, Fried Chicken, and a Caribbean restaurant. While restaurants have proven to be successful in these neighborhoods, the types of restaurants that are not similar to a coffee shop. Lastly, cluster 4 consists of 53 neighborhoods. The most common venue type in these neighborhoods are various types of restaurants including Italian restaurants, pizza places, Chinese restaurants, cafes, delis/bodegas, and a few coffee shops. This demonstrates that restaurants of varying levels of scale and price have been successful in these neighborhoods. Additionally, coffee shops are the most common venue in some clusters, but the majority of neighborhoods are not overrun with coffee shops illustrating room for development.

**5.2 Recommendations**

Based on these observations, I would recommend that the coffee shop be opened in one of the Brooklyn neighborhoods in cluster 4 where coffee shops are not one of the most common venues. These neighborhoods include Bay Ridge, Sunset Park, Gravesend, Brighton Beach, Sheephead Bay, Kensington, Winsor Terrace, Prospect Heights, Red Hook, Starrett City, Manhattan Beach, Coney Island, Bath Beach, Gerritsen Beach, Marine Park, Clinton Hill, Sea Gate, Prospect Lefferts Gardens, City Line, Bergen Beach, Midwood, Georgetown, Ocean Parkway, Fort Hamilton, Ditmas Park, Mill Basin, Weeksville, Dumbo, Homecrest, Highland Park, and Madison. These neighborhoods are included in the cluster where restaurants of various scales and price values have been successful and are not overly populated by coffee shops providing the opportunity for success.

**6. Conclusions**

In this study, I have analyzed data regarding the venues in each neighborhood in Brooklyn as well as their geographic location to determine the optimal neighborhood to open a coffee shop in. The neighborhoods were clustered based on their geographic location and the venue data was analyzed to reveal trends in the types of venues that are successful in each cluster of neighborhoods. Based on this analysis, I recommend that the coffee shop be opened in one of the neighborhoods in cluster 4 where coffee shops are not already one of the most common venues. These neighborhoods include Bay Ridge, Sunset Park, Gravesend, Brighton Beach, Sheephead Bay, Kensington, Winsor Terrace, Prospect Heights, Red Hook, Starrett City, Manhattan Beach, Coney Island, Bath Beach, Gerritsen Beach, Marine Park, Clinton Hill, Sea Gate, Prospect Lefferts Gardens, City Line, Bergen Beach, Midwood, Georgetown, Ocean Parkway, Fort Hamilton, Ditmas Park, Mill Basin, Weeksville, Dumbo, Homecrest, Highland Park, and Madison.

**7. Future Direction**

Future study and closer analysis can be performed on the neighborhoods within cluster 4 to determine the single optimal neighborhood to open the coffee shop in.