**Opening a Coffee Shop in New York City**

Philip Ross

December 20, 2020

**1. Introduction**

**1.1 Background**

Starting a new business is a huge risk for any person. In fact, according to the U. S. Bureau of Labor Statistics, 20% fail within the first two years, and 45% during the first 5. As such, any new business owner should use all the knowledge and data available to them to give their business the best shot at thriving. To do otherwise is just wasting vast amounts of easily available data.

**1.2 Problem**

In this project, we will be looking at a new business owner that is looking to open up a new coffee shop in New York City. Given the size of New York City, our business owner is looking to find a neighborhood in which their coffee shop can thrive. That means that they need to locate a neighborhood where coffee shops are something the residents want, and that the neighborhood is not already over-saturated with coffee shops. We can use existing data to determine the layout of coffee shops in New York City and see where this new one might fit.

Also, many coffee shops sell other goods in addition to coffee (pastries, sandwiches, books). Our new business owner wants to analyze what non-coffee businesses exist in the neighborhoods and determine what additional goods they should look into selling. In fact, they might be able to partner with other businesses in the neighborhoods and sell their goods as well. Coffee shops also tend to act as a hub for information about local events, so knowledge of the neighborhood would also help integrate the shop in the community.

Finally, any new business owner wants to make sure their investment will be safe and protected. As such, our owner would like to look into crime rates of the various neighborhoods to help make an informed decision when ultimately determining which neighborhood is the best fit.

**1.3 Interest**

Parties primarily interested in this project will be those looking to open new businesses. While the results discussed in this report will be specific to New York City, the methodology will be able to be applied to other locations as well. Similarly, just because this specific usecase involves a coffee shop does not mean that it cannot be applied to other businesses as well. Also, this information will be valuable to both first time owners, as well as those who already own one or more businesses.

**2. Data**

**2.1 Location Data**

In this capstone project, we will be using Foursquare location data in conjunction with outside data sources in order to achieve our goal.

The first bit of data that we will need is a list of all the neighborhoods in New York City. To get this information, we are able to use either the new\_york.json file that was part of week 3 in this capstone course. Alternatively, we are able to use:

<https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm>

Either way, this will give us the Neighborhood, Borough, and Postal Codes that we need.

We are also going to need the Latitude and Longitude values for the various neighborhoods. This information comes with the new\_york.json file. We are also able to obtain this information using the Geocode API with the postal codes we got from above.

**2.2 Neighborhood Data**

With this data, we can begin using the Foursquare APIs. We are going to use the previously achieved location data to get additional data for each neighborhood. First, we want to search for coffee shops in the various neighborhoods. We can do so using the 'search' endpoint. This will give us data on existing coffee shops in various neighborhoods, and help us determine which neighborhoods can support coffee shops while not being over crowded.

We can also use the 'explore' endpoint to determine the top 5 types of venues in each neighborhood. This will give us information on what each neighborhood is like, what is popular with it. It will help us determine what additional items the coffee shop should stock if it picks a specific neighborhood. 'details' and 'tips' will also give us data about existing businesses and how a coffee shop would integrate itself.

The 'events' endpoint will give us data about various events taking place in the neighborhood. As mentioned in the Business Problem section, coffee shops often act as a central hub for neighborhoods. Therefore, this information is vital for determining what events the coffee shop would advertise and what the current residents would like to find information about.

Overall, we are looking to find neighborhoods that are most similar to our needs outlined above.

**2.3 Crime Data**

Finally, our business owner wants to make sure they are opening their business in a secure and safe location. Historical crime data can be found here:

<https://www1.nyc.gov/site/nypd/stats/crime-statistics/historical.page>

We will match the precincts with the various neighborhoods in order to get an overall view of the safety of the locations. We will want to match for neighborhoods that share a similar high safety level.

**3. Methodology**

**3.1 Data Wrangling**

The first step to the project was to acquire data pertaining to New York City. Using the data source mentioned in the Data section, we obtained neighborhood, borough, and zip code data. As this project was neighborhood focused, each row in our dataframe represented a neighborhood with its corresponding borough and zip codes. It should be noted that a single neighborhood often had multiple zip codes.

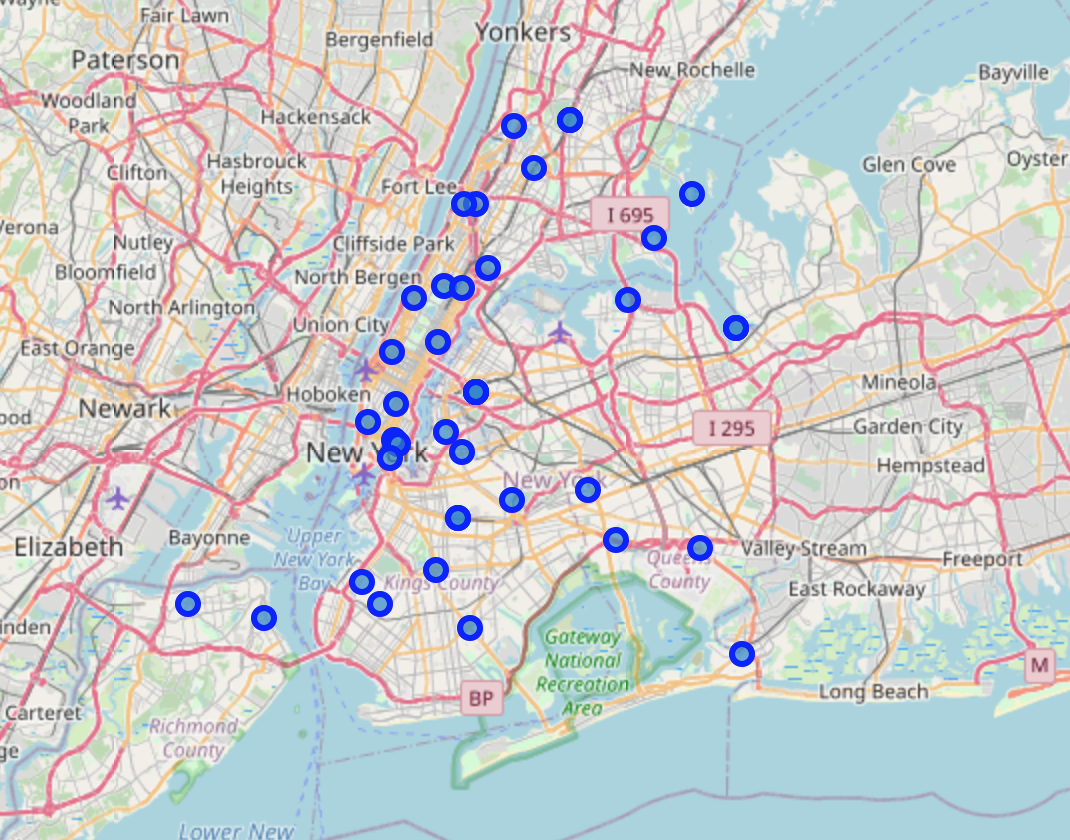
Next we set about acquiring location data. Using Geocoder, we were able to search for the longitude and latitude of each neighborhood in New York City. We were then able to append this data onto our dataframe.

The final portion of data wrangling involved acquiring crime data. There were some difficulties with this section that should be noted. While our project is ‘neighborhood’ focused, we were unable to find New York City crime data at that level. However, by using some data manipulation we were able to get crime data at the ‘borough’ level. Since neighborhoods share boroughs, this was not ideal but still seemed valuable. Still, there was worry over this possibly skewing neighborhood clustering. As such, we would ultimately have two dataframes, one with and one without crime data. This is discussed more in a later section. Also, for the sake of this report, the crime data used was 2019 crime statistics relating to the seven major crimes in New York City.

To acquire the crime data, we first got the various precincts and matched them to the boroughs to which they belonged. After some additional formatting we were able to merge the precinct information to the neighborhood dataframe. Next we pulled in 2019 crime data for the seven major crimes in New York City. The data was broken down by precinct so those values were added together to get the totals for each borough. Once the totals were acquired, they were appended to the crime dataframe.

**3.2 Neighborhood Mapping**

Using the location information for the neighborhoods, we were able to plot them so as to get a visual of potential locations for our coffee shop.



**3.3 Single Neighborhood Exploration**

In order to get familiar with the data we would be working with, we decided to first explore one New York City neighborhood at random. We ended up choosing Central Bronx. After acquiring the location data for it, we were able to utilize the Foursquare API so as to explore the neighborhood. Since we are focused on opening a coffee shop which is considered a venue, we then filtered the results to just get the venues for Central Bronx. We received back 6 venues from the API call that will be discussed in the Results section.

**3.4 Exploring All Neighborhoods**

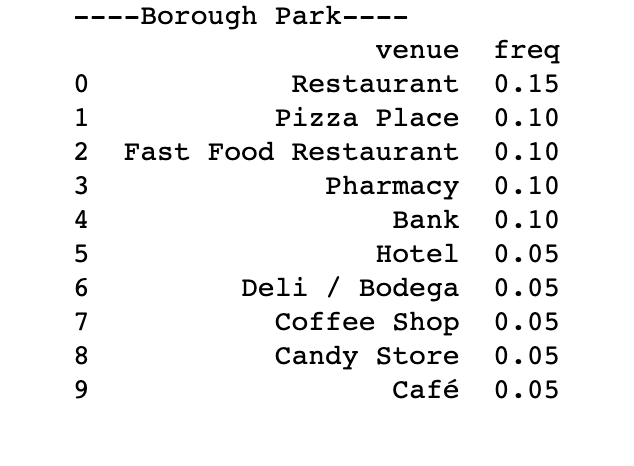
Our next goal was to explore all the neighborhoods in New York City. In order to do that, we utilized the Foursquare explore API for each neighborhood and then filtered the results for venues. Once again, we decided to filter for venues as that is what our coffee shop would be classified as.

Next, we grouped the venue results by neighborhood. We then performed one-hot encoding on the venue results so as to get the venue information in a format that would allow us to perform data analysis and clustering on it more easily. The mean values were taken for each venue and the data was combined with the grouped neighborhood dataframe. At this point we had a dataframe consisting of each neighborhood and the average occurrence of each venue type in it.

**3.5 Acquiring the Top 10 Most Common Venues for each Neighborhood**

At this point, we needed a way to limit our information as we were dealing with too much data noise to get an accurate view of each neighborhood. For instance, most venues did not even appear in any given neighborhood. To solve this problem, we decided that it would be best to limit our results to the top X most common venues in a given neighborhood. This would give us an accurate view of what types of venues make up a neighborhood while not overwhelming us with data. We needed a good value of X for this. Too small a value could potentially exclude valuable venue information, while too large a value could still include too much irrelevant data. Ultimately, we settled on the value 10 for X.

We took the dataframe and acquired a list of the top 10 most frequent venues for each neighborhood. An example:



Observe the rankings are in order with Restaurants at the top with a frequency of 0.15. You will also note that some venues are similar. We will discuss this more in later sections.

We then converted these results into a dataframe consisting of neighborhoods and their top 10 most common venues.

**3.6 Clustering Neighborhoods with Crime Data**

As mentioned above, we decided to perform clustering both with and without crime data. This section will discuss clustering without crime data.

When attempting to perform K Means clustering, one of the first things to do is to determine the number of clusters. How to determine this number is a very hard problem with many different approaches. For this project, we decided to calculate the sum of squared distances for each k value from 1 to 25 and take the ‘elbow’ value as our k. Observe below:



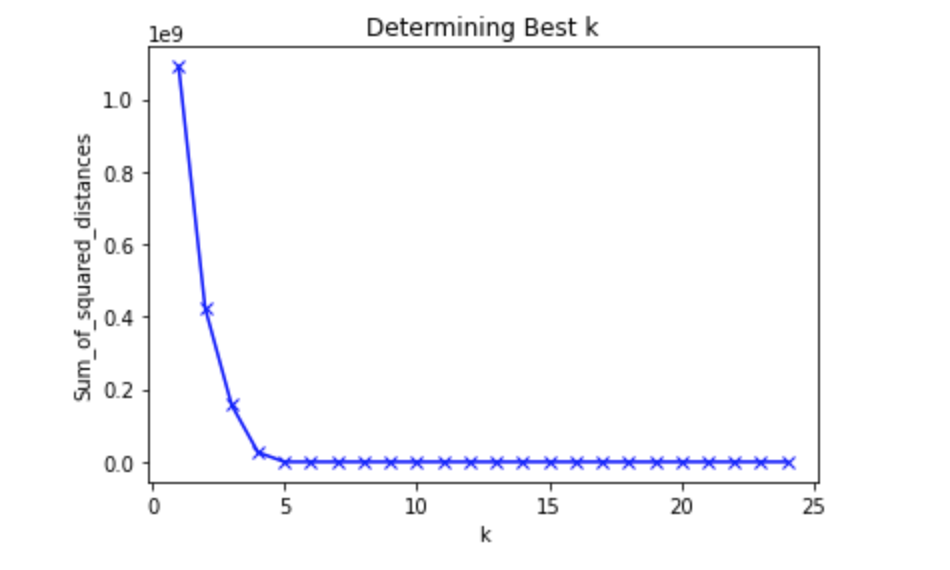
An issue with the data was that there did not appear to be a clear elbow. As such, clustering was performed on multiple k’s in order to observe the grouping for multiple values. It was ultimately determined to use k = 9 for the number of clusters.

Once k was chosen, clustering was performed, the data was fit, and cluster labels were generated for each neighborhood. The cluster labels were added to the dataframe which was then merged with the dataframe containing the most common venues information. The clusters were then plotted on a map to give a visual representation of the clustering. Finally, the clusters were examined individually. This will be discussed more in later sections.

**3.7 Clustering Neighborhoods with Crime Data**

At this point we decided to repeat the previous steps, but this time with the crime data included. We revisited the dataframe consisting of neighborhoods and the one-hot encoded venue frequencies and appended the crime statistics to them.

Here we also encountered the hard problem of needing to choose a value for the number of clusters. Once again, we calculated the sum of squared distances for k from 1 to 25 and plotted the results in order to locate the ‘elbow’:



The elbow was more apparent this time and we ended up choosing k = 4 for our number of clusters. With the number of clusters as 4, we applied K Means clustering, fit the data, and generated cluster labels for each row in the dataframe. This information was then merged with the dataframe containing the top 10 most common venues in each neighborhood. The clusters were then plotted on a map to give a visual representation of the clustering. Finally, the clusters were examined individually. This will be discussed more in later sections.

**4. Results**

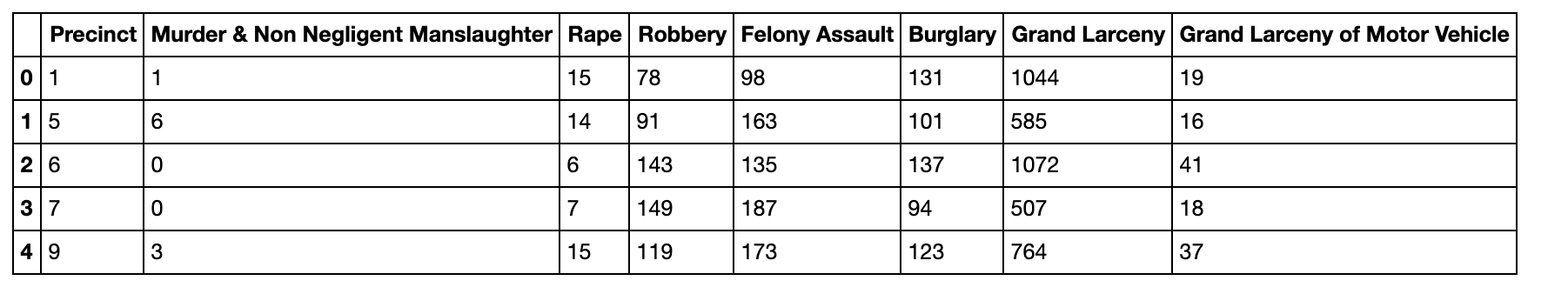
In this section we will be discussing the findings acquired from the Methodology section.

**4.1 Neighborhood Information**

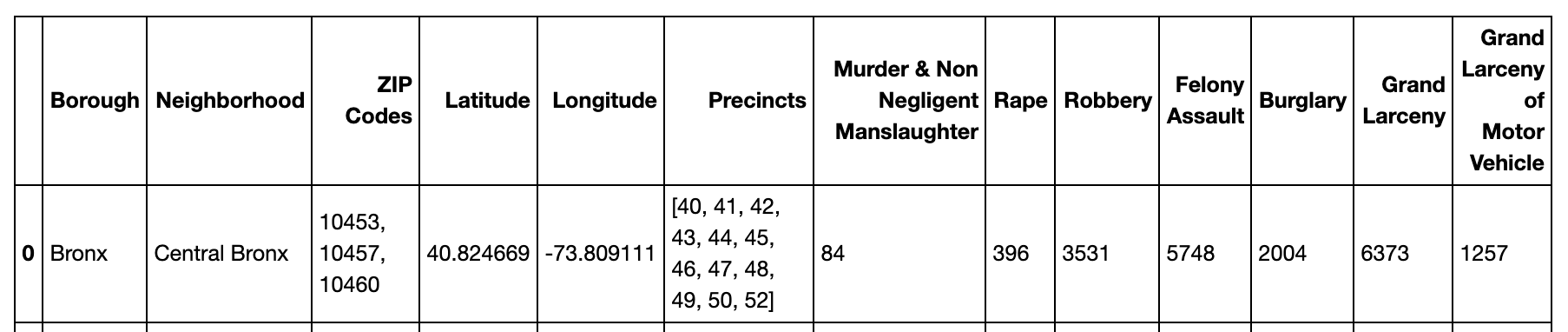
The following is the list of neighborhoods in New York City for reference.

**4.2 Crime Data per Precinct**

The following is an example of crime data for the 7 major crimes in New York City precincts for 2019.

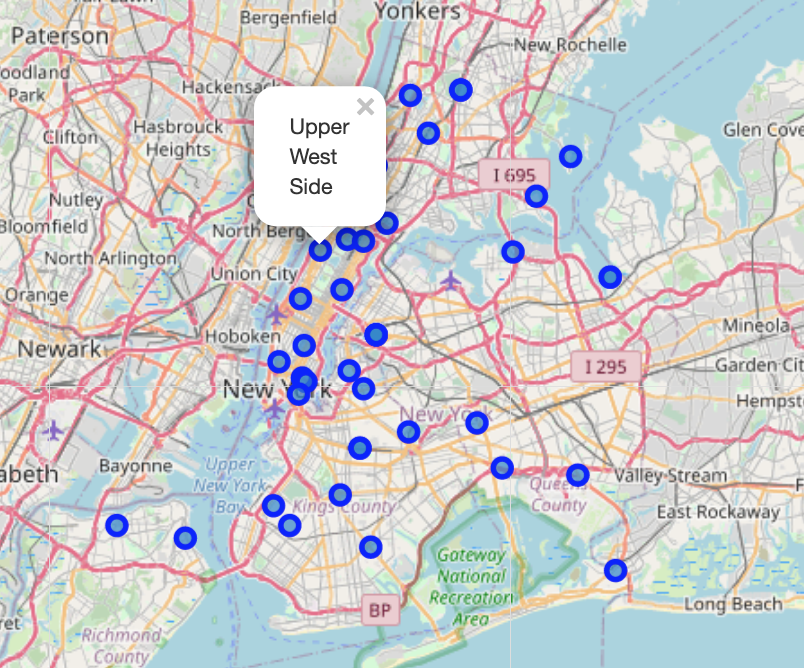


We were then able to apply this information to each neighborhood/borough.



**4.3 New York City Neighborhood Locations**

Visual representation of the neighborhoods located through New York City.

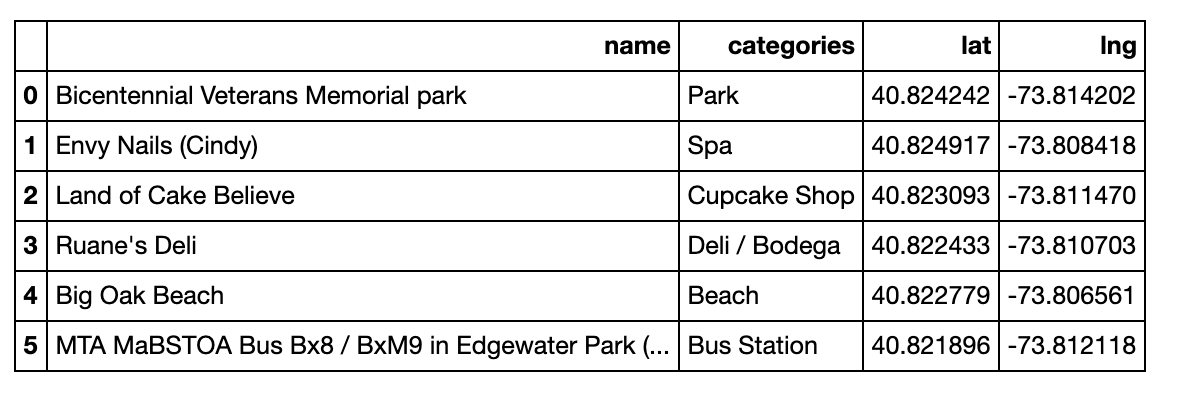


**4.4 Exploration of Central Bronx**

Sample of Foursquare exploration results:



Formatted venues returned for Central Bronx:

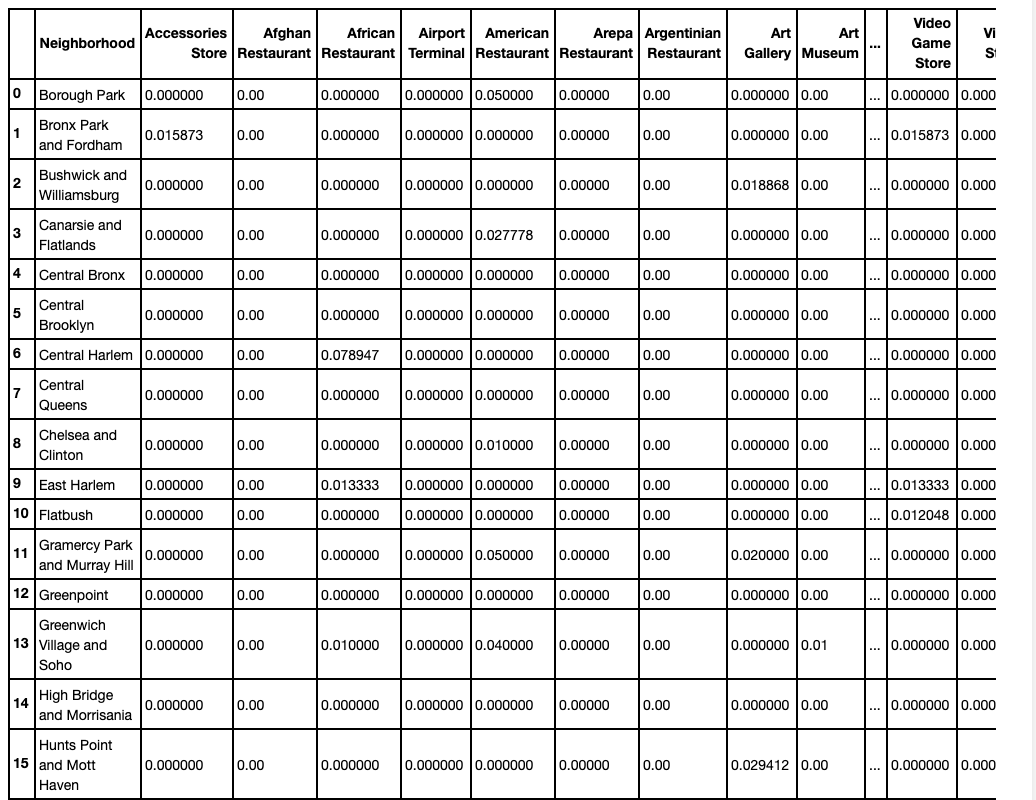


**4.5 Exploration of all New York City Neighborhoods**

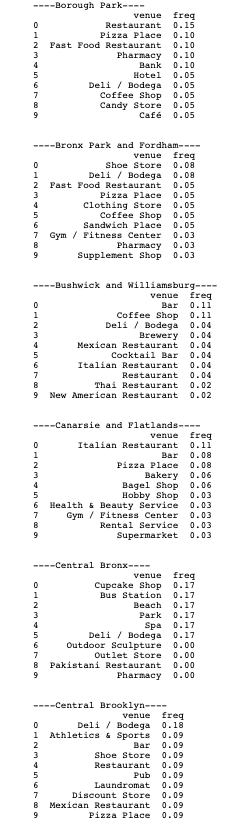
Venue count by neighborhood

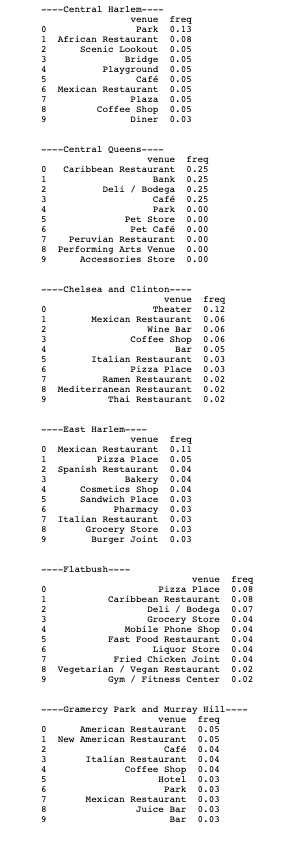


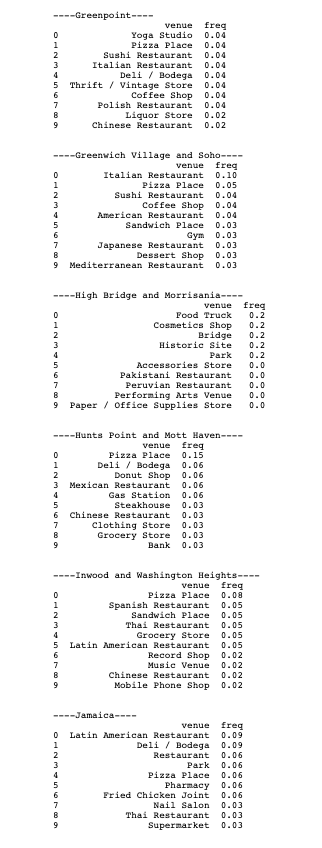
Sample of venue frequency broken down by neighborhood. Full results can be examined in the Notebook as they are too large to fit here.

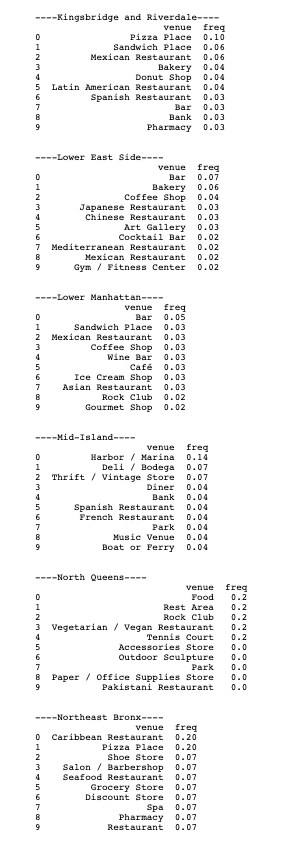


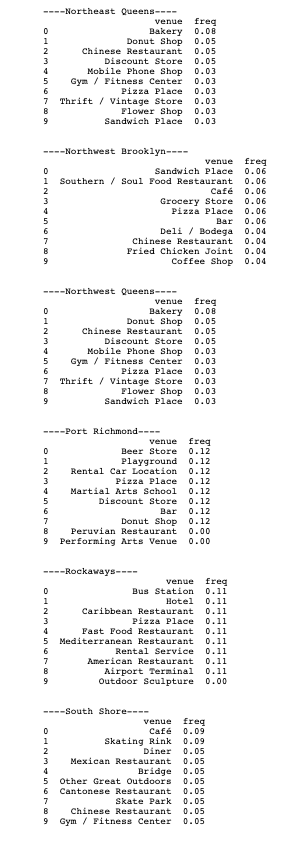
Top 10 most common venues by neighborhood.

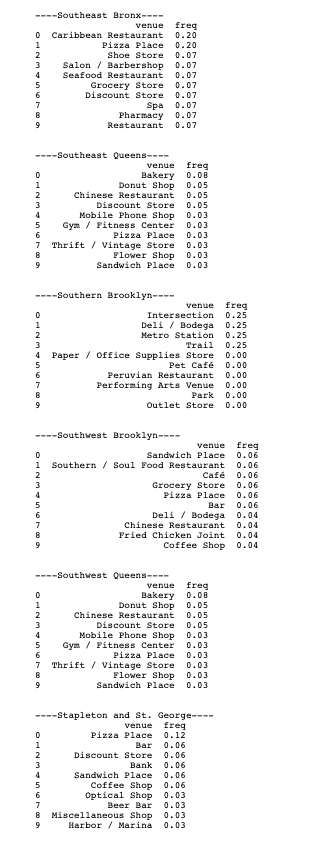


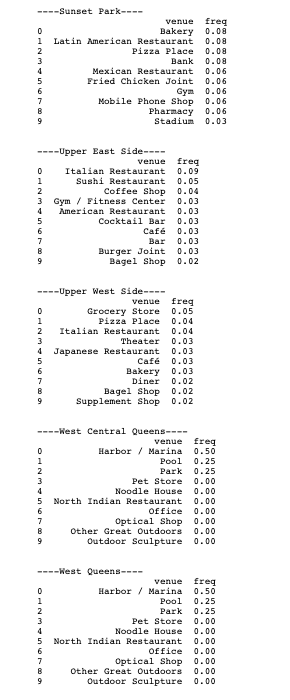










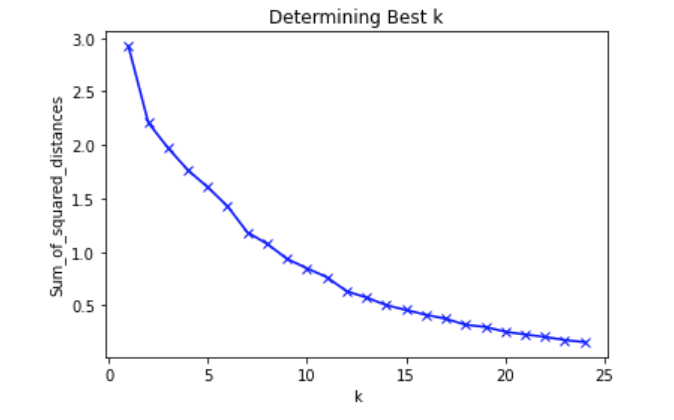


Sample of formatted venue information:



**4.6 Clustering without Crime**

Sum of squared distances for k

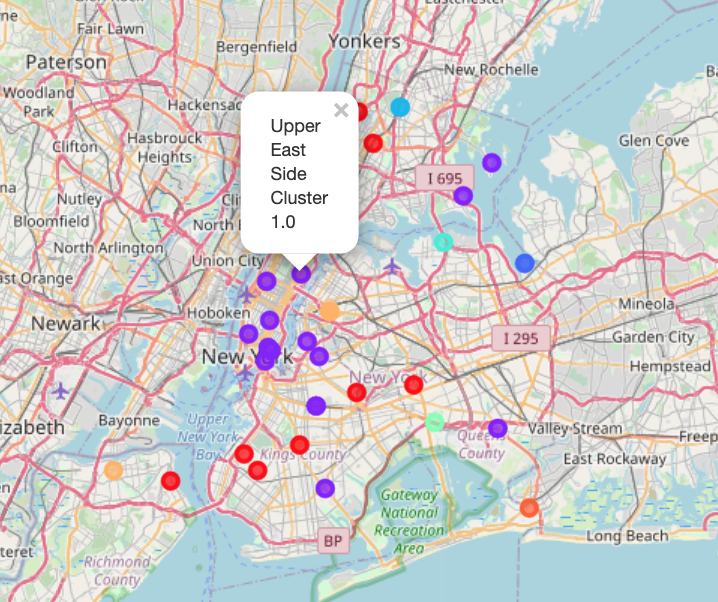


It was determined to use k = 9

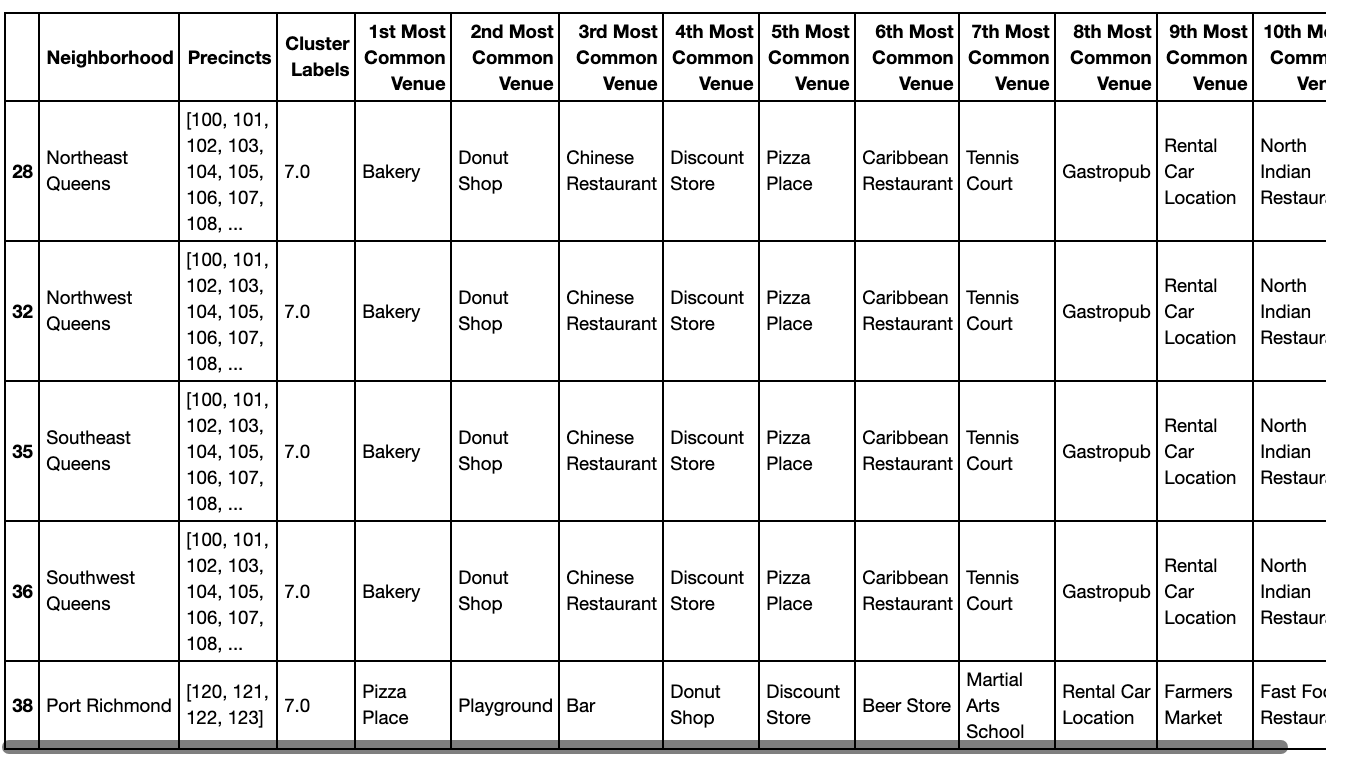
Sample of data with Cluster Label added:



Map with neighborhoods and cluster label



Example of cluster with full data (Cluster 7):



**4.7 Cluster Breakdown without Crime**

Cluster 0: Bronx Park and Fordham, Hunts Point and Mott Haven, Kingsbridge and Riverdale, Central Brooklyn, Borough Park, Flatbush, Sunset Park, Jamaica, Stapleton and St. George

Cluster 1: Central Bronx, Southwest Brooklyn, Canarsie and Flatlands, Northwest Brooklyn, Greenpoint, Bushwick and Williamsburg, Central Harlem, Chelsea and Clinton, East Harlem, Gramercy Park and Murray Hill, Greenwich Village and Soho, Lower Manhattan, Lower East Side, Upper East Side, Upper West Side, Inwood and Washington Heights, Rockaways, Southshore, Mid-Island

Cluster 2: West Central Queens, West Queens

Cluster 3: Northeast Bronx, Southeast Bronx

Cluster 4: North Queens

Cluster 5: Southern Brooklyn

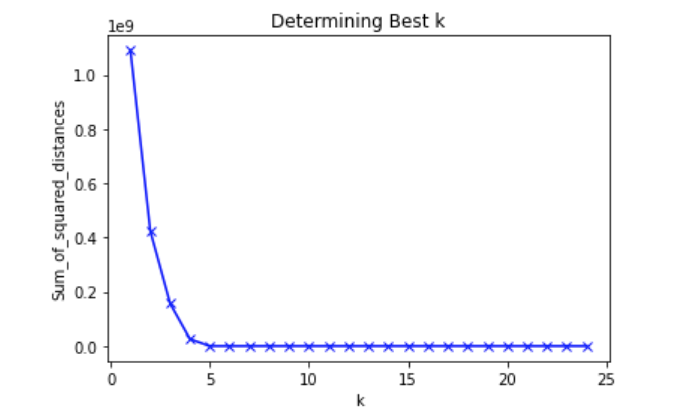
Cluster 6: High Bridge and Morrisania

Cluster 7: Northeast Queens, Northwest Queens, Southeast Queens, Southwest Queens, Port Richmond

Cluster 8: Central Queens

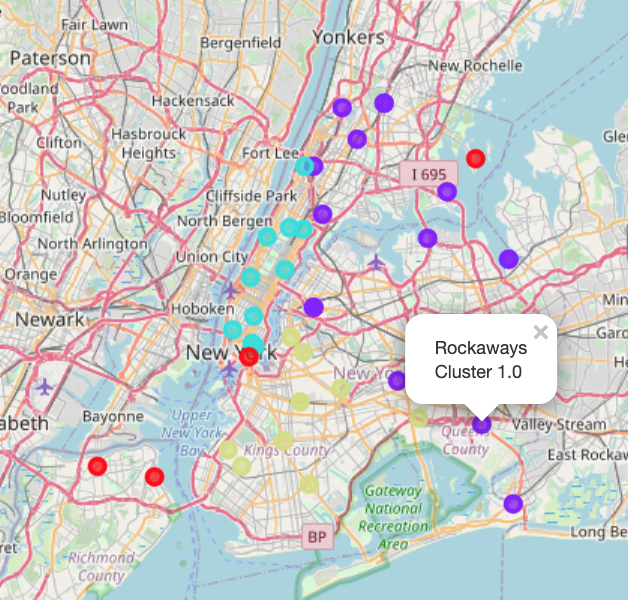
**4.8 Clustering with Crime**

Sum of squared distances for k

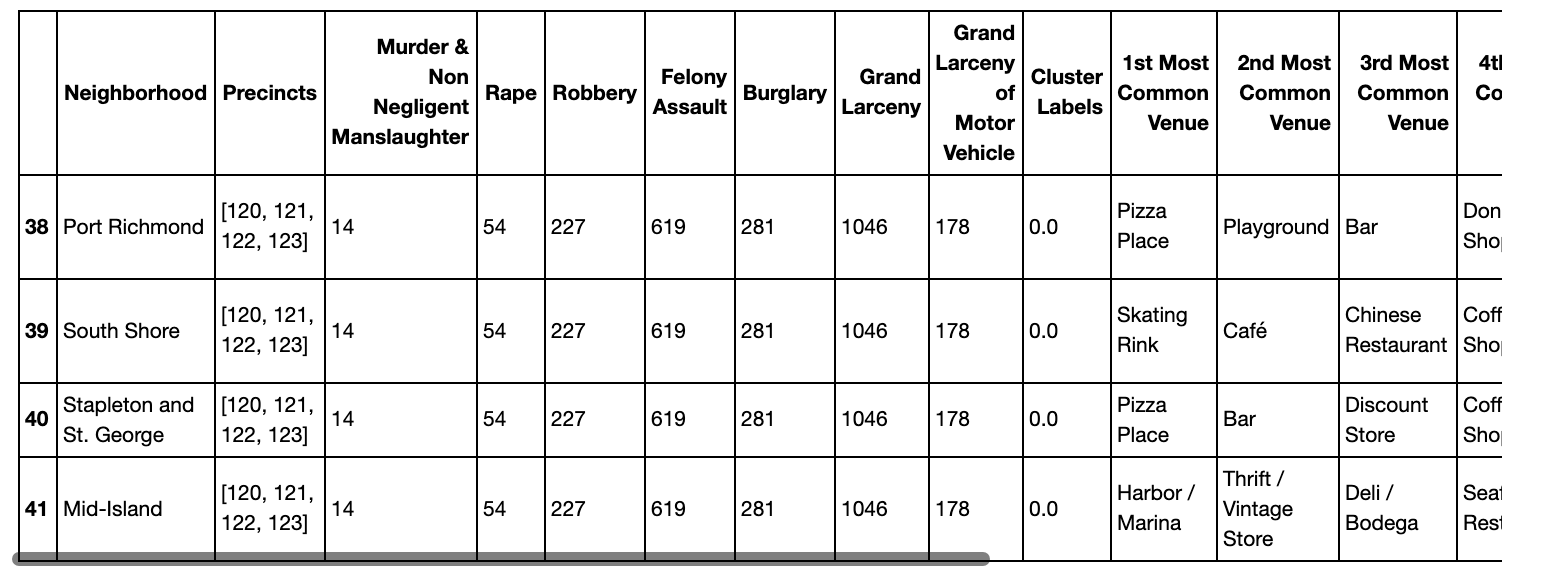


It was determined to use k = 4.

Map with neighborhoods and Cluster Label:



Example of cluster with full data (Cluster 0):



**4.9 Cluster Breakdown with Crime**

Cluster 0: Port Richmond, South Shore, Stapleton and St. George, Mid-Island

Cluster 1: Central Bronx, Bronx Park and Fordham, High Bridge and Morrisania, Hunts Point and Mott Haven, Kingsbridge and Riverdale, Northeast Bronx, Southeast Bronx, Northeast Queens, North Queens, Central Queens, Jamaica, Northwest Queens, West Central Queens, Rockaways, Southeast Queens, Southwest Queens, West Queens

Cluster 2: Central Harlem, Chelsea and Clinton, East Harlem, Gramercy Park and Murray Hill, Greenwich Village and Soho, Lower Manhattan, Lower East Side, Upper East Side, Upper West Side, Inwood and Washington Heights

Cluster 3: Central Brooklyn, Southwest Brooklyn, Borough Park, Canarsie and Flatlands, Southern Brooklyn, Northwest Brooklyn, Flatbush, Greenpoint, Sunset Park, Bushwick and Williamsburg

**5. Discussion**

The goal of this project was to determine which neighborhoods in New York City would be good locations to open a new coffee shop. In this section, we will discuss how the above results relate to that goal.

First, we should talk about the data. There were two main data concerns going into this project. The first was if there would be venue overlap between neighborhoods. Being unfamiliar with the size of New York City neighborhoods, I was curious whether or not a venue in one neighborhood might also appear in another. In a way, this is not really a problem since it does indicate that the two neighborhoods would be similar. However, it might suggest that there is a better way to group potential locations for the coffee shop (zip code, borough, etc.).

The second, and more significant concern was with the crime data. I noted above that I was unable to find crime data on a per neighborhood basis. The best I could do was find crime statistics per precinct. Even then, this could not easily be translated to neighborhoods. Part of the issue is that neighborhoods are not always well defined. However, I was able to determine that specific precincts do belong to specific boroughs. I ultimately decided to sum up the precinct values for each borough and attach it to the neighborhood data. I figured that it might be likely that neighborhoods in the same borough would have similar crime statistics, so this might be acceptable. However, I still had my doubts and was worried this could skew the clustering so I decided to have two sets of data, one with crime statistics and one without. Also, I needed to determine what crime information to use. I ultimately ended up using data for the seven major crimes in New York City in 2019. In the future, additional types of crime could be used, along with other years so as to see crime trends in boroughs.

I next wanted to observe what the Foursquare data would look like for a neighborhood. For this purpose, I chose Central Bronx and filtered the results to ‘venues’. I did this as a coffee shop would be considered a venue. I also wanted to see what other venues would be nearby. This could give me information on competitors, collaborators, or neighborhood activities. All of this was information that could help the coffee shop prosper. Six venues were returned for Central Bronx. They were a park, a spa, a cupcake shop, a deli, a beach, and a bus station. With this information, let us consider Central Bronx as a potential spot for our coffee shop. On one hand, no coffee shop venues showed up. Why? Maybe this location simply cannot support a coffee shop. This should immediately make us a little wary. On the other hand this might mean that this is a prime location for a coffee shop, that we have found a need for this neighborhood. Let us take a look at the other venues and see how they might impact us. We could potentially partner with the cupcake shop and the deli to sell their items at our location. There is a park, and it is not uncommon for people to grab a coffee before a walk in the park. Also, there is a bus station, which indicates this could be a high traffic area. However, it might also be a “pass-through” area that people do not really stop at (maybe why there were only six venues). Central Bronx appears to have some potential, but there are also some red flags. As a result, this location would require closer inspection, possibly in-person.

We wanted to repeat the previous step for each neighborhood in New York City. However, since this would generate so many venues we would be completely overwhelmed by data. To solve this problem, we decided to limit the results to just the top ten most frequent venues for each neighborhood. The hope was that ten would give us enough information to get an accurate view of a neighborhood while not giving us too much noise. See the Results section for the top ten venues for each neighborhood.

The following are some of the more promising neighborhoods, along with reasons for why they might be good choices. One thing to note is that some venue categories are very similar to each other. For example, coffee shops and cafes are nearly the same thing. We need to be aware of this as otherwise we might not have an accurate view of what venues truly are most common.

Borough Park has both coffee shops and cafes, meaning it can support them. However we also need to be aware to not move to an oversaturated area. Hotels mean people are staying in that location. There are also delis to partner with.

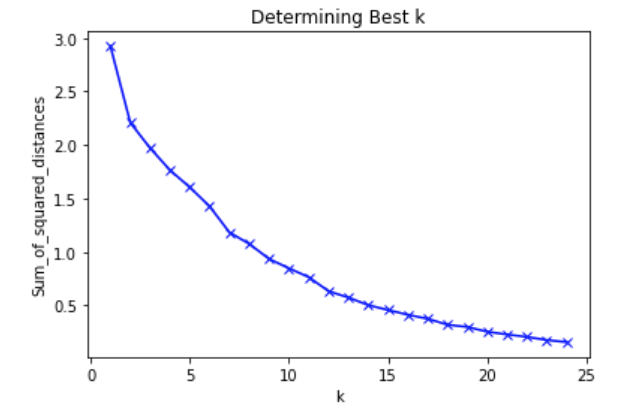
Central Harlem has parks and playgrounds that people might like to drink coffee at. There are already some existing shops. Also, given the high number of parks, we could look into events at the parks and advertise them in our shop.

Northeast Queens has bakeries and donut shops to partner with. Plus, coffee shops do not appear in their top ten. As a result, the location might be ripe for a coffee shop to move in.

Southwest Brooklyn supports both cafes and coffee shops. It has grocery stores in which to sell our coffee. There are also sandwich and pizza places to partner with.

This gives us some good information about the outlook of the neighborhood. Still, this is a lot of information to look through. We next decided to cluster the neighborhoods together so as to see which neighborhoods are most similar and focus our efforts there.

First, we did clustering without the crime data. To determine the number of clusters to use, we calculated the sum of squared distances from 1 to 25 and attempted to find the ‘elbow’ value. This value would ultimately be our cluster number. The results are below.



Unfortunately, no clear elbow value was produced. After some trial and error it was ultimately decided to use the value 9 for the number of clusters. We then applied the k-means algorithm to generate the clusters and mapped them for easy visualization. We observed that there were two main clusters, a medium one, with the rest being smaller clusters. The different Queens neighborhoods pretty much formed their own cluster, which was one of the areas we were interested in. However, Port Richmond also appeared in that cluster meaning it could be a potentially overlooked area for a coffee shop. We might never have seen this connection. Besides that, there was not a ton of overlap based strictly on geography which was something I had been worried about. As such, I was pleased to see such differentiation in top ten venues within clusters.

Next we moved onto the clustering with the crime data. We once again performed the elbow method to find the ideal number of clusters and ultimately settled on using 4. We applied the K Means algorithm and plotted the clusters so as to create an easily readable visualization. We then began to examine the individual clusters.

Unfortunately it became quickly apparent that the crime data seemed to overwhelm the venue information and the clusters seemed to just be determined by which precincts corresponded to the boroughs the neighborhoods belonged to. Three clusters were entirely their own precincts with the fourth being the two remaining precinct groupings. This was something I had been worried about and was not entirely unexpected given the data limitations.

**6. Conclusion**

Overall, the project was largely a success. We were able to take location data for the neighborhoods in New York City and utilize the Foursquare APIs with them to gain insight into the neighborhoods. From that we were able to begin to determine which neighborhoods could potentially be a good location for a new coffee shop. Some potential neighborhoods were Borough Park, Central Harlem, and Northeast Queens. We were then able to cluster the neighborhoods accordingly to discover additional prime locations for coffee shops more easily. For instance, we overlooked the fact that Port Richmond was another good location based on the clustering. We attempted to add crime data to the overall dataset, however this produced overall disappointing results as we were not able to get statistics specifically for each neighborhood. This could be an area to revisit in the future. Still, we have valuable feedback for our new business owner, and have additional ideas if they would like to continue exploring the data.