Introduction/Business Problem

New York City is known for having a plethora of dining options all over the city. With a population of 8.4 million people and over 65 million visitors a year, the demand for restaurants will remain constant. While many people enjoy the diversity of cuisine and the opportunity to get "authentic" foreign dishes, there is still a demand for authentic American dishes: residents who did not grow up in New York occasionally yearn for comfort food from their childhood, and visitors from overseas want to try traditional American food. One of the quintessential type of foods in this categories is barbecue (BBQ). And not just any BBQ – real, authentic southern style pit BBQ.

Many BBQ restaurants use cooking techniques and ingredients that are focused on cost and volume, not taste. Because of that, the taste and quality is ok but not great. Being a native Southerner I know the difference quality ingredients and more importantly proper cooking techniques. This provides an opportunity to open several BBQ restaurants in NYC that will stand out from the competition. Rather than open a single restaurant, opening a restaurant in each of the 5 boroughs will provide multiple locations that are far enough apart to not overlap.

For this project, I will segment and cluster each of the 5 boroughs in NYC and identify the best neighborhood to open an authentic BBQ restaurant. To predict the ideal location, I'll identify locations with other restaurants nearby, which will show the area already supports food services. I'll also find areas without other BBQ joints nearby to minimize direct competition.

Data

Similar to some of the labs, Foursquare will be used to identify the various venues and locations in a given neighborhood. This data is ideal to sort by venue type and frequency of location/how many of a given type are within a given distance to a neighborhood's center. To provide a comprehensive list of all the neighborhoods in NYC, I'll use the free data available from NYU that was used in one of the skills labs. I did look for similar data in other cities (specifically my hometown of Atlanta) however that particular information is not available except for a fee. This data will provide the latitude & longitude for each neighborhood. That information will be cross referenced and linked with the Foursquare data.

Once the dataframes are built and segmented into the 5 boroughs, I'll then extract the top venues in each neighborhood. I'll then separate using one-hot encoding and group by neighborhood and frequency. After this, I'll search the results and segment any neighborhoods that already have a "BBQ Joint" listed (that is the Foursquare category listing). If the frequency of BBQ joints is high for a given neighborhood, I'll remove that neighborhood from the dataframe. Alternatively, if the number of neighborhoods with BBQ is relatively high, I'll keep them in the list as to not skew any subsequent filtering or sorting however I will not select them as the ideal location should the analyses indicate they are.

After all the neighborhoods are analyzed, I'll use k-means clustering to cluster them appropriately. After the clustering is complete, I'll look at the most frequent venues in the cluster and rank them from best to worst to open a BBQ restaurant, specifically naming each neighborhood in a given cluster. This entire process will be done for all 5 boroughs.

Methodology

The initial method used followed on of the Labs used during the course. After importing all the necessary libraries (i.e. Pandas) I imported the New York City JSON data and parsed it into the various neighborhoods and associated borough. A composite map off all the data was put into a NY map to help me visualize it was the correct data and area.

The remaining steps I repeated separately for all five Boroughs, beginning with Manhattan. I segmented and clustered the individual neighborhoods, then created a map to make sure I was only looking at Manhattan. I then created a Foursquare query to retrieve the venue category by type and applied it to the first neighborhood, again using the top 100 venues within 500m of the neighborhood latitude & longitude. The results were then cleaned up and put into a pandas dataframe. Once this was completed (and checked to look reasonable) a function was used to repeat the above process for the remaining neighborhoods in Manhattan.

The data were now ready to be analyzed. I used one-hot encoding techniques learned in the course lab to create a new dataframe to separate out the different venues for each neighborhood. That dataframe was then grouped by neighborhood and the frequencies of each venue type were displayed. As described earlier, at this stage I wanted to exclude any neighborhoods that had BBQ joints already. In Manhattan's case, I found 4 neighborhoods with BBQ joints so these were removed from consideration for our best place to open up a new BBQ restaurant (reduces direct competition). I grouped the top 5 venues in each neighborhood and displayed them, as a double check that I didn't miss any neighborhoods with a high BBQ presence. I will note that I did not delete the neighborhoods with BBQ joints from the dataframe as I originally planned; as I'll describe later I wanted to target neighborhoods with lots of other restaurants (an assumption that the area can support many restaurants) and removing neighborhoods with BBQ joints would skew the results (since those neighborhoods would contribute to overall restaurant count).

The next step involved creating a new dataframe and listed the top 10 venue types for every single neighborhood. To better manage this data, I used the machine learning method of k-means clustering to separate the data into smaller sets that are easier to work with. Using a k-value of 5, this created 5 clusters of neighborhoods with their top 10 venues shown. I examined each cluster and identified the neighborhoods with the greatest number of restaurants. I somewhat arbitrarily defined a restaurant for this project as a place where one can buy a full meal – so venues such as bars, donut shops, coffee shops, etc. were not counted as restaurants. Also not counted were grocery stores or farmer's markets, though food trucks and generic descriptions of "food" were included.

Results

First, the neighborhoods where we do NOT want to open a BBQ restaurant are shown in the table below. They are listed by Borough and the frequency of BBQ joints they have are listed (in case someone wants to open one with direct competition but not many BBQ joints.

Borough	Neighborhood	BBQ Frequency
Manhattan	Battery City	0.027
	Central Harlem	0.022
	Lenox Hill	0.010
	Manhattanville	0.024

The Bronx	Hunts Point	0.071
	Pelham Bay	0.024
	Van Nest	0.053
Brooklyn	Bedford Stuyvesant	0.033
	Cobble Hill	0.010
	Downtown	0.010
	East Williamsberg	0.014
	Gowanus	0.016
	North Side	0.010
	Red Hook	0.020
	South Side	0.010
	Wingate	0.045
Queens	Astoria	0.010
	Flushing	0.015
	Jamaica Center	0.021
	Murray Hill	0.021
	Rockaway Beach	0.043
	Sunnyside	0.024
Staten Island	Grant City	0.045

Table 1: Neighborhoods that already have BBQ joints

Now that we know where we are <u>not</u> going, we will identify the best places where a BBQ restaurant should be successful. Again, that is defined as a neighborhood that supports lots of other restaurants but does not have any BBQ joints already in place. For Manhattan, the best neighborhood would be Washington Heights (Cluster 2), where 7 out of the top 10 venues are restaurants. Other good options in that cluster are Manhattan Valley, Morningside Heights or Tudor City, each with 6 restaurants in their top 10. Other options that are farther away are Hamilton Heights in Cluster 0, or Marble Hill and Flatiron in Cluster 1, each with 6 restaurants as well.

For an example, here is the map of Manhattan showing the 5 clusters:
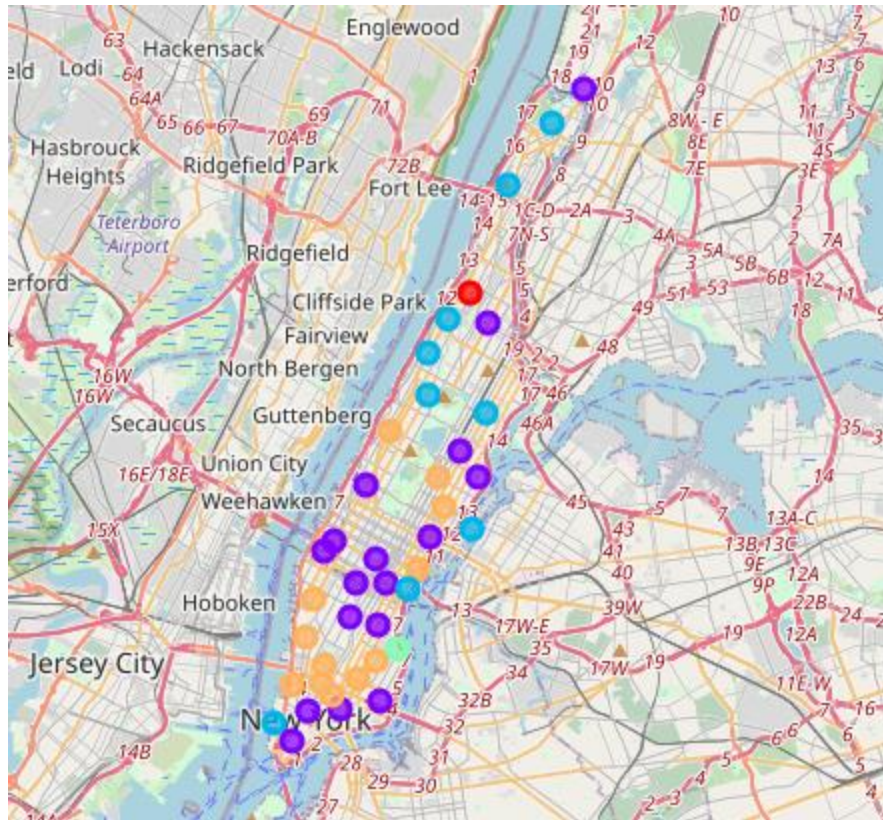
The same exact process as used for the other 4 Boroughs with the following results:

- For The Bronx, the best location would be in Bedford Park (Cluster 2) as well as Concourse Village or Kingsbridge Heights (7 restaurants in the top 10). This cluster had 3 other neighborhoods with 6 restaurants. For an alternative area, Concourse in Cluster 1 also had 6 restaurants.
- For Brooklyn, Clinton Hill (cluster 3) has 8 restaurants in the top 10. That cluster also has 2 neighborhoods with 7 restaurants (Prospect Lefferts Gardens and Ditmas Park), as well as 7 neighborhoods with 6 restaurants. A good alternative would be Dyker Heights (Cluster 0) with 6 restaurants.
- For Queens, Jackson Heights or Elmhurst (Cluster 1) each have 8 restaurants in their top 10. An alternative would be Bayswater (Cluster 4) with 5 restaurants.
- For Staten Island, Rosebank and Eltingville (Cluster 3) each have 7 restaurants in their top 10, as do Grymes Hill and Tompkinsville (Cluster 0) for alternatives.

Discussion

The results are pretty straightforward however they do hinge on a couple of key assumptions, namely that neighborhoods that already have many restaurants could support another one. The exclusion of neighborhoods that already have BBQ joints is another assumption – those areas show there is a demand for BBQ already and perhaps shouldn't be dismissed. An interesting observation is that none of the neighborhoods that were excluded show up on the best locations results.

Before picking a location, additional studies would be needed. Looking at metrics such as rent, proximity to transit, day traffic versus residents, or breaking the locations down into a primarily lunch or dinner crowd are potential studies that would provide better answers.

The cluster results were interesting, in that there was a clear cluster with more restaurants than the others for all 5 Boroughs (thus not a random event). If specific data were available to further separate the venues by cuisines, another clustering or regression model could show the best areas.

Conclusion

This project narrowed down the best areas to open a BBQ restaurant (by the assumptions/criteria listed) in each of New York's 5 Boroughs, from a total of 306 neighborhoods and 14,223 venues. This was a fun and enlightening exercise, and the results could help bring BBQ to the masses in NYC. It provided a good opportunity to apply our new skills to a practical problem, showing the importance of good design and being able to refine the approach as you go, as well as ID additional studies to provide better answers. Now let's eat!