# The Battle of the Neighborhoods

## Introduction

A restaurant investment group with several venues along the east coast is interested in opening a new restaurant in New York City. In over 20 years of operations, the group has found that its most successful investments have been restaurants that specialize in cuisines that have regional popularity but are underrepresented in the vicinity of the restaurant location. The group leadership has asked its lead data scientist to identify the most popular restaurant types/cuisines in New York City and to identify neighborhoods with the least competition for these restaurant types/cuisines. The results of this report will help group leadership select an appropriate restaurant type/cuisine and narrow down the search for potential restaurant locations.

## Data

The data used to conduct this analysis consists of the following:

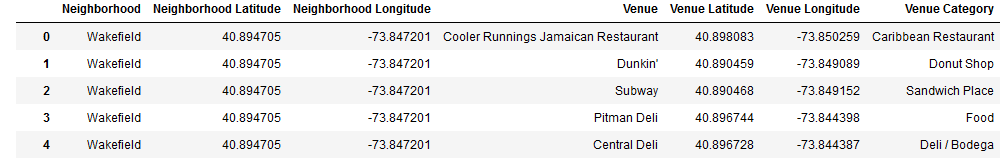
* A listing of New York City neighborhoods with associated boroughs and central latitudinal and longitudinal coordinates.
* Food-related venue data for each neighborhood, including name of venue, venue latitudinal and longitudinal coordinates, and venue category.

The New York City neighborhood data is obtained from the following link: <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json>

The food-related venue data is obtained via the Foursquare API, using a default radius of 500 meters from the center coordinates of each neighborhood. It should be noted that the food-related venue data used in this analysis is limited to the first 100 results for each neighborhood.

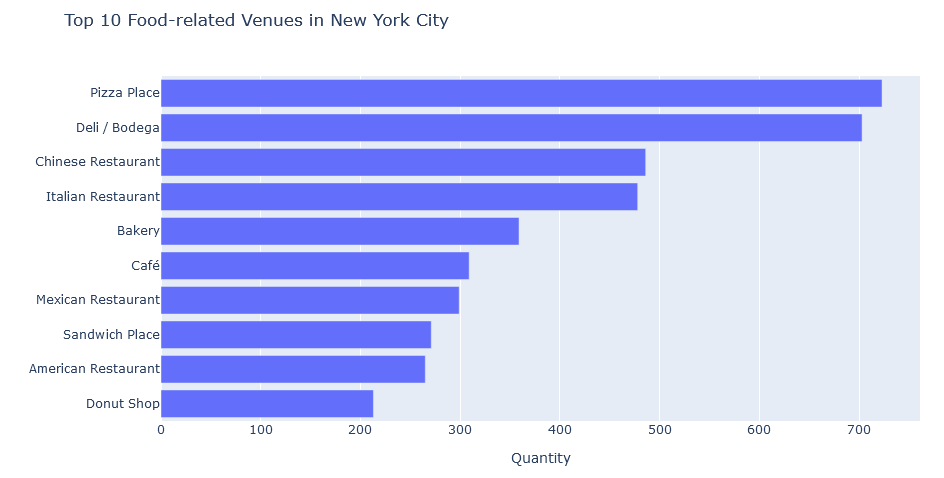
## Methodology

The first step in the analysis was to acquire the New York City neighborhood data described in the previous section. The neighborhood central coordinates from this data set were then used to make Foursquare API venue explore queries to obtain the venue data. To restrict the results of these queries to food-related venues, the categoryID for the parent category ‘Food’ provided in Foursquare API documentation was used in the queries. The query results were then compiled into a new Pandas dataframe that combined the neighborhood and venue data for each venue. A snapshot of the first 5 rows of this dataframe is provided in the figure below. Note, the borough data was excluded because they represent geographical areas that are too large to be useful in the analysis.



An initial inspection of the resultant data showed that venue data were obtained for 8,139 venues representing a total of 143 distinct venue categories. Of the 306 neighborhoods included in the original neighborhood dataset, only 293 neighborhoods returned venue data via the Foursquare API queries. So, the 13 neighborhoods for which no venue data was available were excluded from the analysis.

To estimate the popularity of each restaurant type/cuisine, the assumption was made that popularity would be closely correlated to the number of venues since more demand would be required to sustain a greater number of venues. To begin exploring the venue data in more depth, one hot encoding was used to transform the dataframe so that unique columns for each of the 143 distinct venue categories were established. Each row of the new dataframe included the neighborhood and a value of 1 under the associated venue category column; a value of 0 was assigned to the remaining venue category columns. To determine the most popular restaurant types/cuisines in all of New York City, the one hot encoding dataframe was then grouped by neighborhood and summarized, with a city-wide total row added to capture the totals for each venue category. Sorting the results, the top ten most popular venue categories can be seen in the figure below.

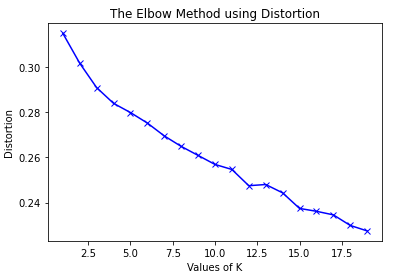


With a better understanding of the most popular restaurant types/cuisines in New York City, the analysis then focused on grouping the various neighborhoods into clusters based on their 10 most prevalent venue categories. To accomplish this task, the k-means clustering algorithm was chosen. K-means clustering is a popular unsupervised learning algorithm that can be used to identify patterns and likenesses in data which would otherwise be difficult to determine. The intent of applying this algorithm was to reduce the 293 neighborhoods into far fewer marketplaces where competition between the most prevalent venue categories would be similar, thus simplifying the analysis.

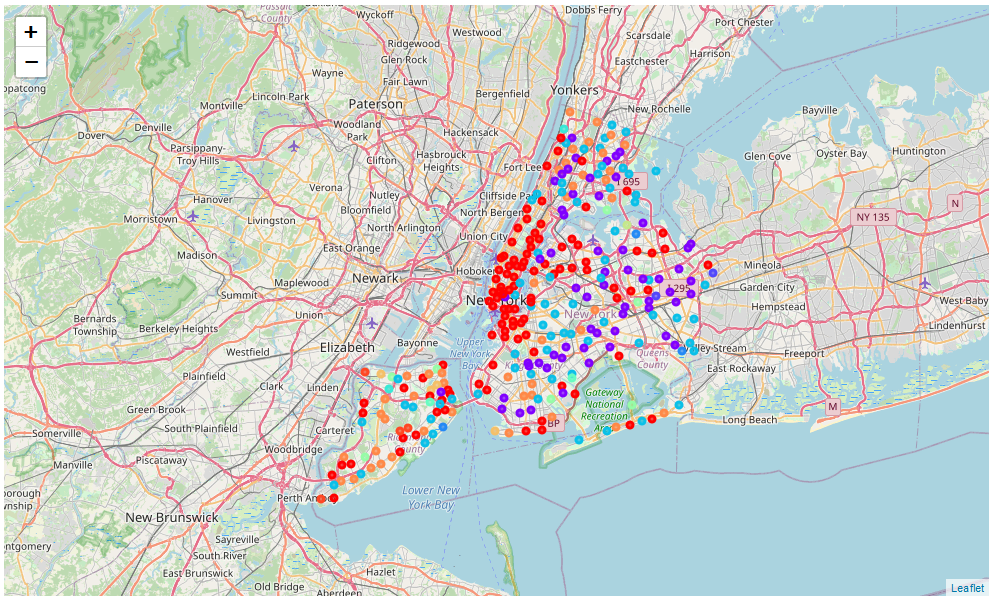
Before applying the k-means clustering algorithm, the one hot encoding dataframe was again grouped by neighborhood and normalized based on distribution frequency. This resulted in a value under each venue category column that represented the venue category as a percentage of total venues in the neighborhood. These percentages were then used to create a new dataframe that included the top ten venue categories for each neighborhood. A snapshot of the first 5 rows of this dataframe is provided in the figure below.



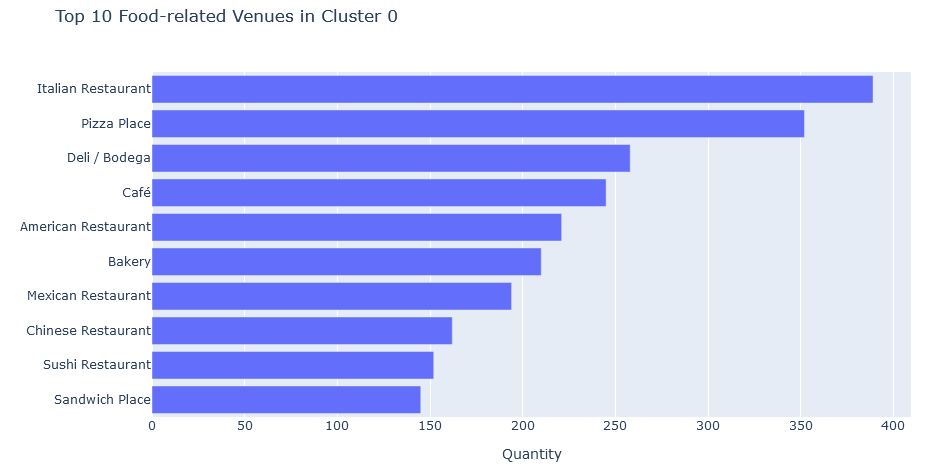
The Elbow Method was then used to determine the optimal number of clusters into which the data would be partitioned. Iterating through k values from 1 to 20, the resultant distortion values (or, the average of the squared distances from the cluster centers of the respective clusters) were plotted against the number of clusters, k. The results are depicted in the figure below.



Based on the figure, k=12 was chosen as the elbow point and the k-means clustering algorithm was applied to the dataframe to partition the data into 12 clusters. A ‘Cluster Labels’ column was inserted into the dataframe to capture the assigned cluster number for each neighborhood. This dataframe was then merged with the neighborhood dataframe to incorporate the neighborhood central coordinate data. The Folium library was then used to plot a map of New York City with each neighborhood represented by circles color-coded by cluster number, as shown in the figure below.



With the clusters defined, the top ten venue categories among the neighborhoods belonging to each cluster were than analyzed to gain a better understanding of how the k-means clustering algorithm grouped the neighborhoods. The results from Cluster 0 are presented in the figure below, as an example.



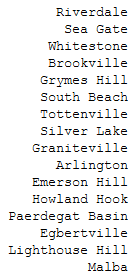
To get a better understanding of the cluster profiles, the venue category distribution frequency was calculated for each cluster. Again, this resulted in a value under each venue category column that represented the venue category as a percentage of total venues in the cluster. The resultant dataframe was then filtered to only include the top ten venue categories previously identified for all of New York City. A ‘Neighborhood Count’ column was also added to the dataframe to indicate the total number of neighborhoods comprising each cluster. The dataframe is depicted in the figure below.



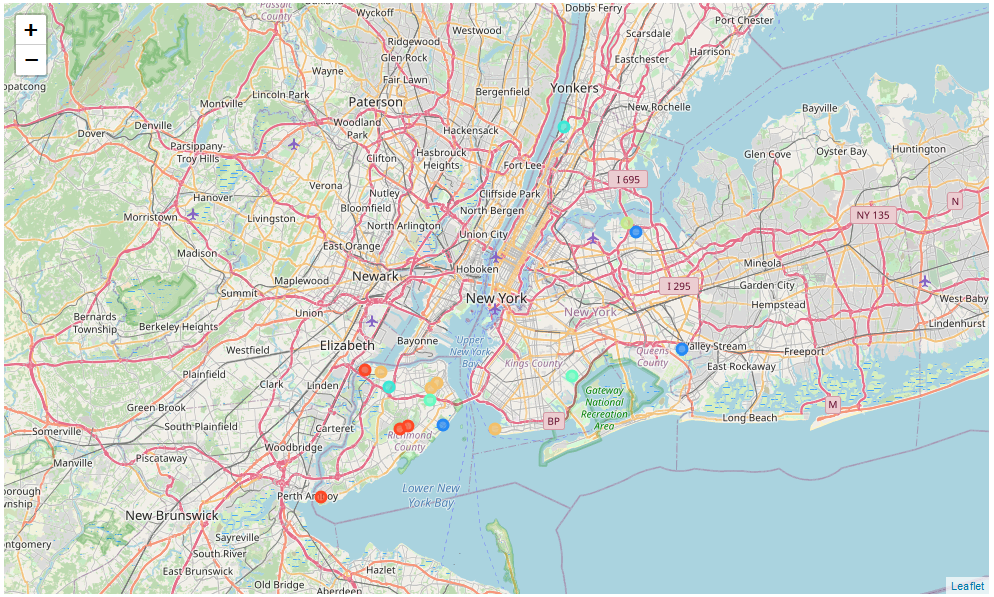
One of the immediate insights to be derived from the dataframe is that the partitioning of the neighborhoods into clusters is rather uneven. In other words, there are several clusters that consist of 7 or fewer neighborhoods while there are others with more than 40.

To identify clusters with the least competition for popular restaurant types/cuisines, the distribution frequencies for each cluster were subtracted from the distribution frequencies for the entire city to create a new dataframe capturing the resultant differences. The maximum values for each venue category in this new dataframe were then calculated and can be seen in the figure below.



Examining the results, it was determined that for each of the top ten most popular venue categories in New York City there were clusters that had no venues that fell under one or more of those categories. In other words, for each venue category there were one or more clusters where virtually no competition was found within the neighborhoods belonging to that cluster. Given that ‘Pizza Place’ was found to be the most popular restaurant type/cuisine in New York City, the analysis focused on identifying the clusters wherein no venues were found to fall under the ‘Pizza Place’ venue category. This included clusters 3, 5, 6, 8, 9, and 11. A listing of the neighborhoods belonging to these clusters is shown in the figure below.  


The Folium library was then used to plot a map of New York City with each of these neighborhoods represented by circles color-coded by cluster number, as shown in the figure below.



## Results

Based on the analysis, ‘Pizza Place’ is the most popular restaurant type/cuisine in all of New York City. The analysis identified 16 neighborhoods (see previous section) where there is currently no competition from other pizza restaurants.

## Discussion

Based on the analysis, it would appear that there are investment opportunities for each of the top ten most popular restaurant types/cuisines in New York City. Given that ‘Pizza Place’ is the most popular of them all, the recommendation from the data science team would be for the restaurant investment group to select a pizza restaurant as the new restaurant to open in New York City. The data science team also recommends that the search for a location to open the new pizza restaurant be narrowed to the 16 neighborhoods where there is currently no competition from other pizza restaurants.

There are also neighborhoods with no competition for each of the other nine most popular restaurant types/cuisines. If other factors would render the choice of a pizza restaurant less feasible, further analysis could be performed on one or more of these other nine most popular restaurant types/cuisines.

## Conclusion

The restaurant investment group data science team successfully determined the most popular restaurant types/cuisines in New York City and identified neighborhoods with the least competition for the most popular restaurant type/cuisine in the city. The recommendation from the data science team is to open a new pizza restaurant in one of the following neighborhoods: Riverdale, Sea Gate, Whitestone, Brookville, Grymes Hill, South Beach, Tottenville, Silver Lake, Graniteville, Arlington, Emerson Hill, Howland Hook, Paerdegat Basin, Egbertville, Lighthouse Hill, or Malba.