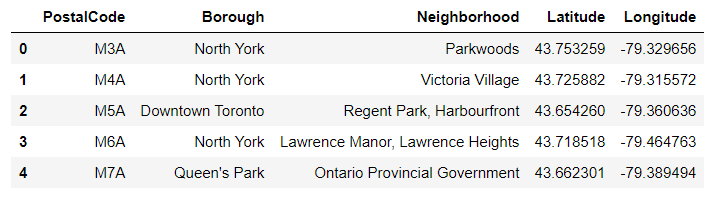
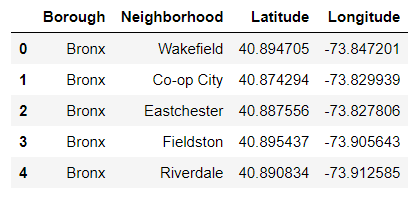
**The Battle of Neighborhoods – Toronto vs. New York**

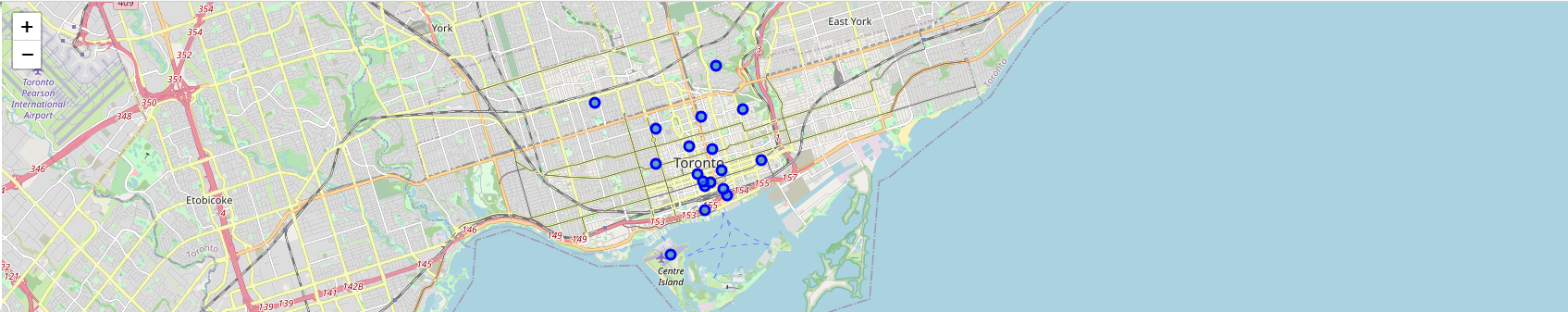
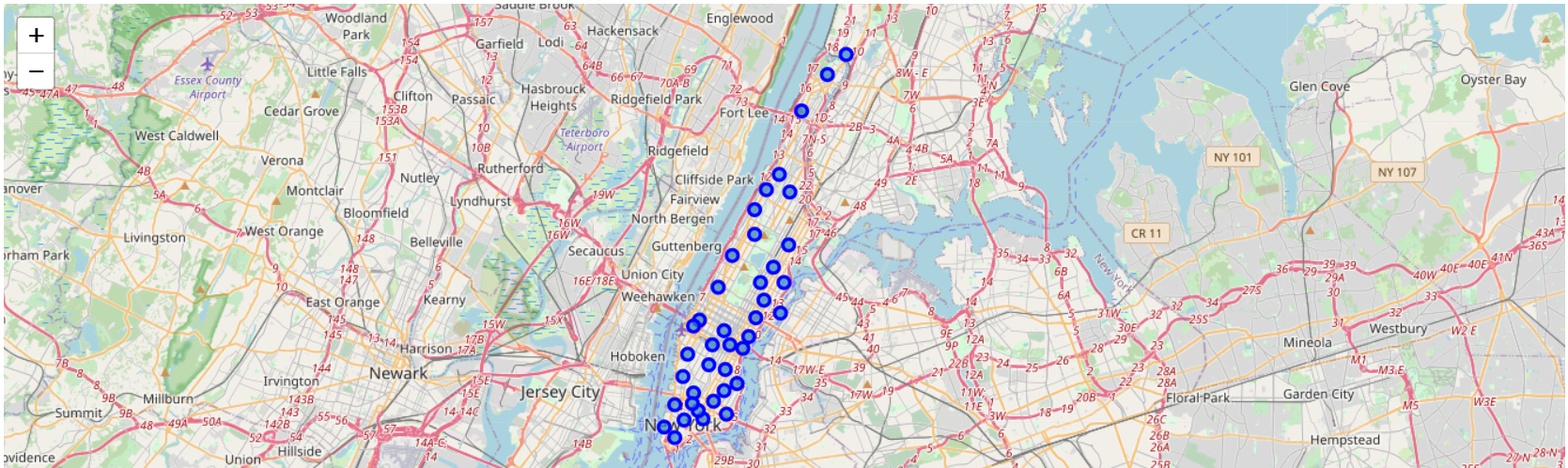
**Hareen Subramaniam**

**27th June, 2021**

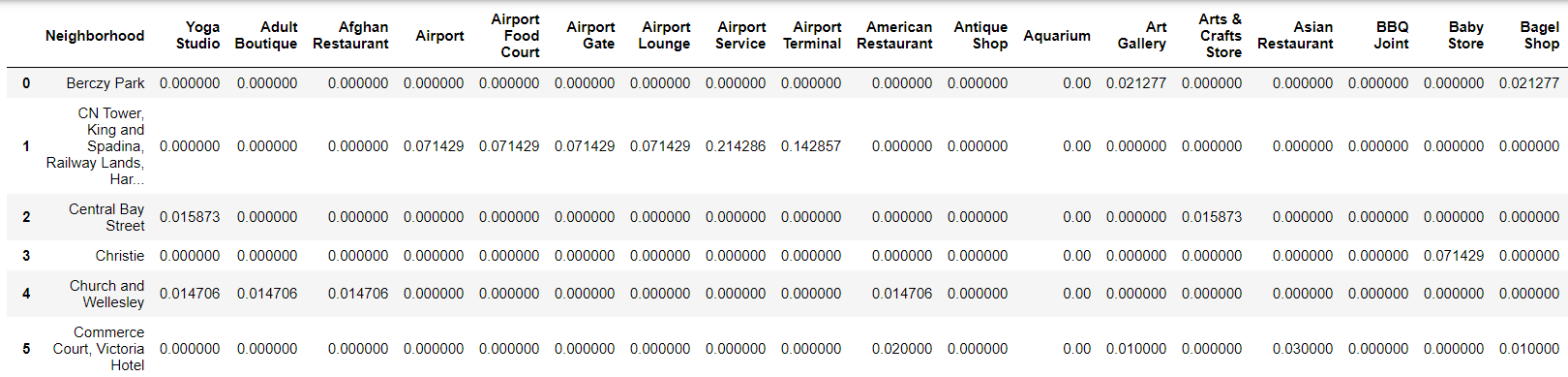
**1. Introduction  
  
1.1 Background**As data scientists, we solve real world problems by analyzing data, applying complex mathematical and problem-solving skills, and validating theories to address open-ended questions. New York City and Toronto are two large cities which are well-known and can be quite attractive to immigrants from many countries around the globe.   
  
**1.2 Problem**If one were thinking about migrating to either cities, which would be the better city to live in? New York City or Toronto in the past couple of decades are home to many immigrants around the world. Each of these two cities are unique to their own heritages.  
  
**1.3 Interest**  
Exploring and comparing these two cities can be quite easy to someone who would be migrating since most of the information such as ratings of restaurants, venue vicinities to a specified neighborhood, etc. are at our fingertips. With a wide range of data and tools to process and analyze them from our data science toolbox, we will be exploring how similar/dissimilar they are to one another with regards to common venues for food, tourist attractions, and more.   
  
**2. Data acquisition and cleaning**  
  
**2.1 Data Sources**  
In answering the problem statement at hand, we will be pulling data from our previous course labs in order to get the dataset for the neighborhoods of both New York city and Toronto. We will process these into usable data frames as well as intend to leverage the Foursquare API to further explore the datasets for these two cities. The datasets will include information about their surrounding neighborhoods i.e hotels, cafes, coffee shops, restaurants, museums, theaters, etc.   
  
**2.2 Data Cleaning**  
To narrow down our search specifically which neighborhood would be appropriate to live in/migrate to, we will focus our analysis towards the island of Manhattan and Downtown Toronto to cluster and segmentize the neighborhoods. Based on the vicinity of common venues, we will compare each neighborhood in Manhattan and Downtown Toronto to help situate what would be the most ideal city to live in and ultimately, which neighborhood would be best in each city. Essentially, we will additionally be able to distinguish what the level of similarities there are between these two cities.   
  
**3. Methodology**Since this problem involves two cities, both exploration, analysis and visualization were conducted in a similar fashion, first for the city of Toronto and then New York.   
  
**3.1 Data Exploration**  
The neighborhood data was extracted from the Wikipedia page of the list of Toronto’s neighborhoods. The data was manipulated into a dataframe by using Python pandas. Data wrangling steps such as removing erroneous data, combining neighborhoods which have same neighborhood coordinates for each borough, and removing “Not Assigned” neighborhoods were applied to the dataframe. In the next step, coordinates which were provided in a .csv file were added to the dataframe for each neighborhood. Similarly, data wrangling was applied to the New York dataset as well. The datasets comprised of postal code, boroughs, neighborhoods, latitude and longitudes and could be visualized with the table below.

  
*Table indicating first five rows of the Toronto dataset.*  
  
  
  
*Table indicating first five rows of the New York dataset.*

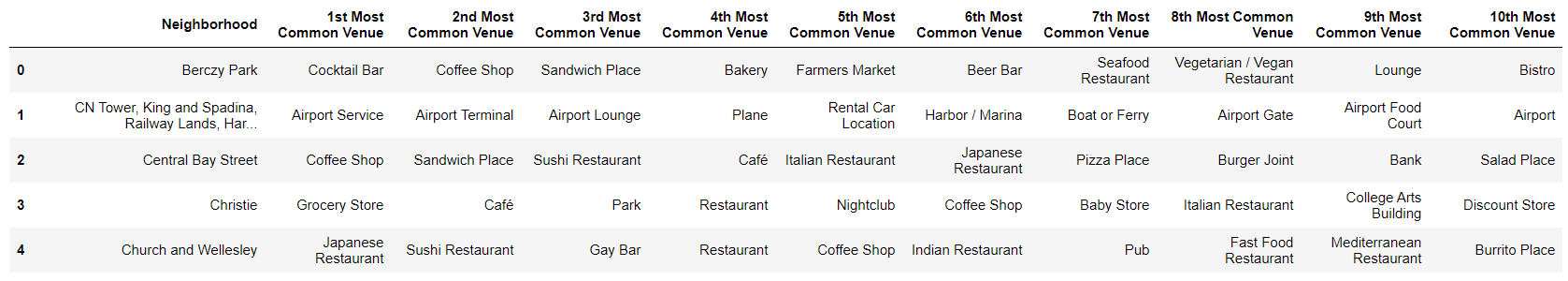
Preliminary calculations were made to ensure the accuracy of the steps to wrangle the data and were compared with the values achieved from the previous labs. For the Toronto dataset, using the .shape function from pandas, the dataframe comprised of 15 boroughs and 103 neighborhoods while for the New York dataset, the dataframe consisted of 5 boroughs and 306 neighborhoods.   
  
The next steps were to get the geographical coordinates of each city by utilizing the geopy library and hence, the geocode function. With this, the geographical coordinates of Toronto and New York were determined to be 43.6534817, -79.383934 and 40.7127281, -74.0060152 respectively. To narrow our search, only Downtown Toronto and the island of Manhattan were focused in this analysis. Hence, the Toronto and New York dataset were sliced to only focus on boroughs with the word “Downtown Toronto” and “Manhattan” in them. The coordinates for each neighborhood in both cities were plotted as follows:

  
*Map of Downtown Toronto indicating the location of each neighborhood.*  
  
  
*Map of Manhattan indicating the location of each neighborhood.*

**3.2 Analysis**The dataset were analyzed separately. Starting with the Downtown Toronto dataset, the first neighborhood was explored and looked at. From the data, it was found that the first neighborhood was Regent Park, Harbourfront. The latitude and longitude values of this neighborhood was 43.6542599, -79.360635.  
  
The next step was to utilize the Foursquare API to construct a function which would return the categories of the venues and then another function to get the nearby venues in all the neighborhoods of the dataset. Before constructing these functions, the radius and limit of parameter of was set to 500 and 100 for the GET request URL. The function is then ran on each neighborhood and a new dataframe called toronto\_venues was created. By running these functions, 58 venues were returned by the Foursquare API and each venue was listed out.   
  
Using one hot encoding, each neighborhood was analyzed. Whichever venue had popped up in the neighborhood, a “1” would appear on the table. Afterwards, another dataset called toronto\_grouped was created by grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. The following indicates how the dataset looks:

  
*Screenshot of the table indicating the first few rows of toronto\_grouped.*

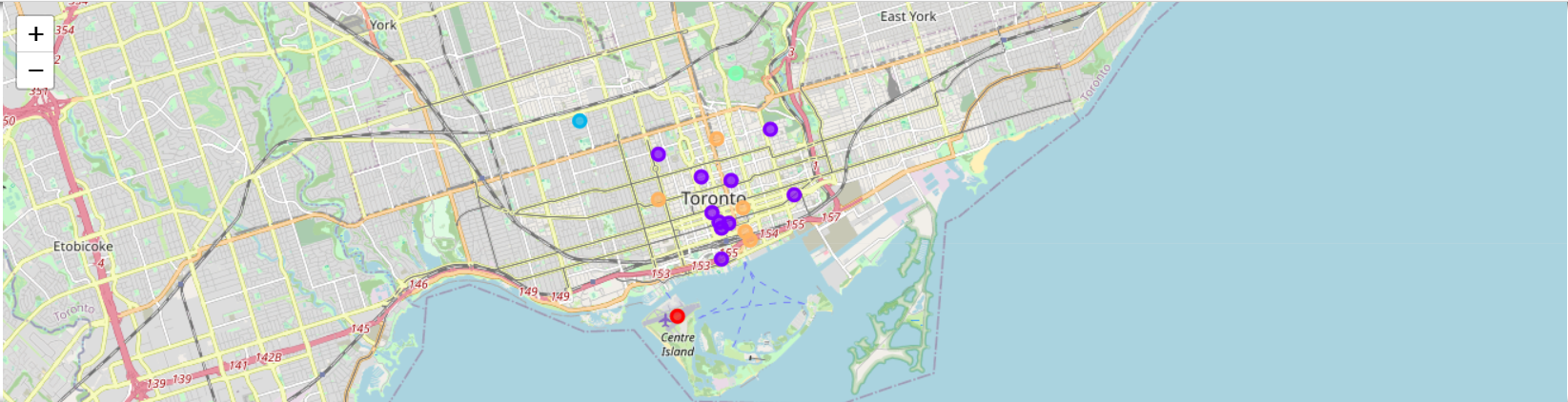
Each neighborhood was printed along with the top 5 most common venues. A function called return\_most\_common\_venues was created to sort the venues in descending order. A new dataframe called neighborhoods\_venues\_sorted was created by passing the function to return the top 10 venues for each neighborhood.

  
*Screenshot of the table indicating the first 5 rows of neighborhoods\_venues\_sorted for Downtown Toronto.*

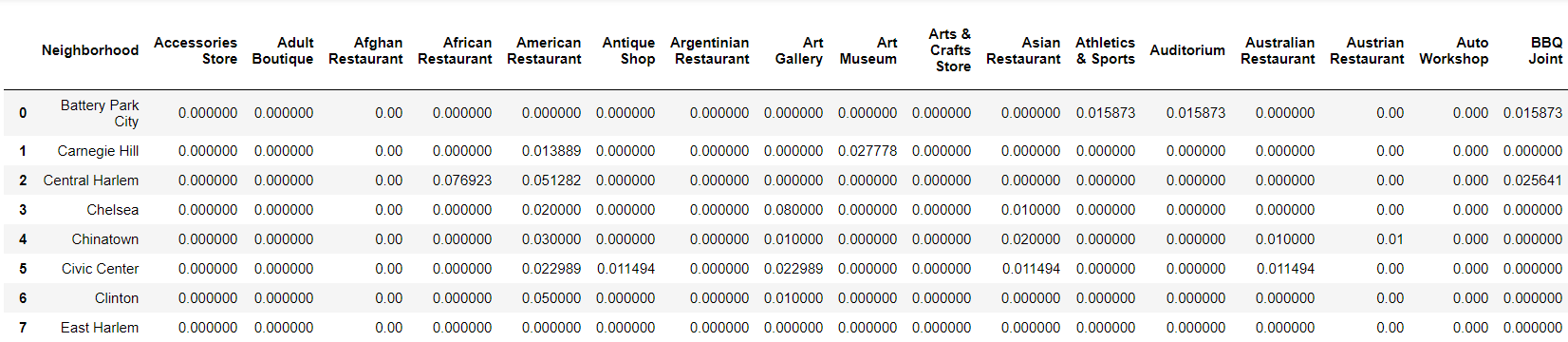
After this, the clustering and segmentation of the neighborhoods was carried out. The k-means was designated to 5 to cluster the neighborhood. The k-means clustering algorithm was ran and the cluster labels were generated and checked for each row in the dataframe. A new dataframe called toronto\_merged was created by merging, the downtown Toronto neighborhood dataset, cluster labels and the neighborhoods\_venue\_sorted dataset.

  
*Screenshot of the table indicating the first 5 rows of toronto\_merged.*

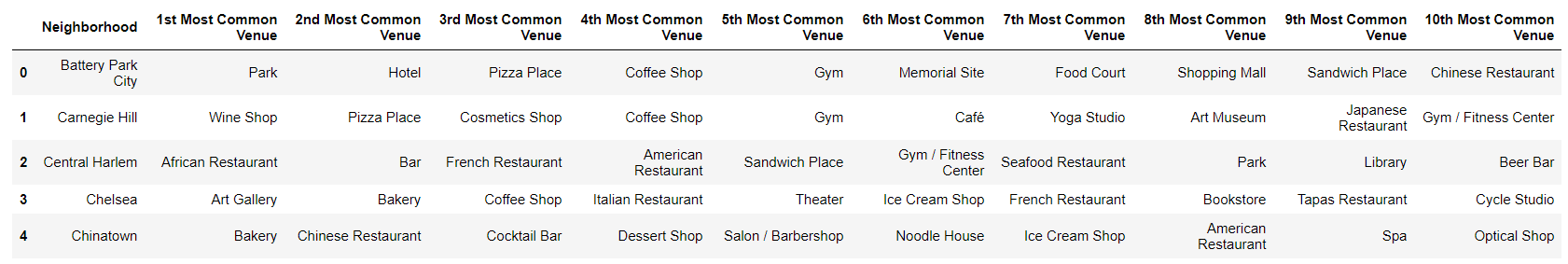
The resulting clusters were visualized on a map.

  
*Map of Downtown Toronto indicating the locations of each neighborhood cluster.*

The same, was repeated for the Manhattan dataset. The Foursquare API was utilized again to run the function that was created previously to return the categories of the venues and then the next function to get the nearby venues in all the neighborhoods of the Manhattan dataset. Before running these functions, the radius and limit of parameter of was set to 500 and 100 for the GET request URL. The function was then ran on each neighborhood and a new dataframe called manhattan\_venues was created. By running these functions, venues were returned by the Foursquare API and each venue was listed out.   
  
Using one hot encoding, each neighborhood was analyzed. Whichever venue had popped up in the neighborhood, a “1” would appear on the table. Afterwards, another dataset called manhattan\_grouped was created by grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. The following indicates how the dataset looks:

  
*Screenshot of the table indicating the first few rows of manhattan\_grouped.*

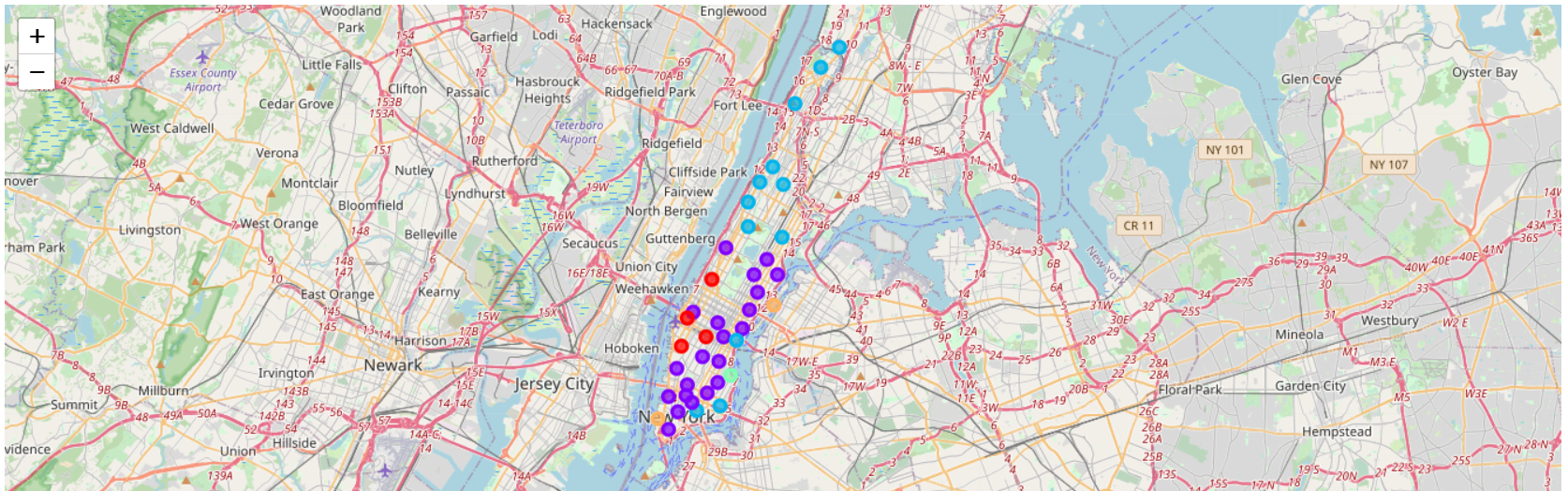
Each neighborhood was printed along with the top 5 most common venues. The function called return\_most\_common\_venues which was previously created, was used to sort the venues in descending order. A new dataframe called neighborhoods\_venues\_sorted was created by passing the function to return the top 10 venues for each neighborhood.

  
*Screenshot of the table indicating the first 5 rows of neighborhoods\_venue\_sorted for Manhattan.*

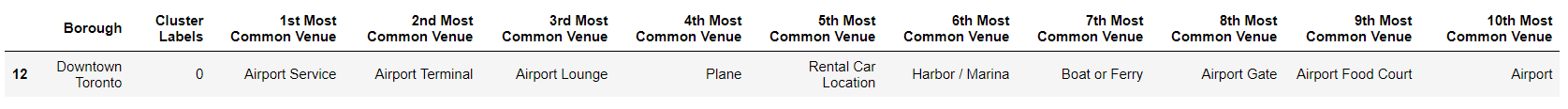
After this, the clustering and segmentation of the neighborhoods was carried out. The k-means was designated to 5 to cluster the neighborhood. The k-means clustering algorithm was ran and the cluster labels were generated and checked for each row in the dataframe. A new dataframe called manhattan\_merged was created by merging, the Manhattan neighborhood dataset, cluster labels and the neighborhoods\_venues\_sorted dataset.

  
*Screenshot of the table indicating the first 5 rows of manhattan\_merged.*

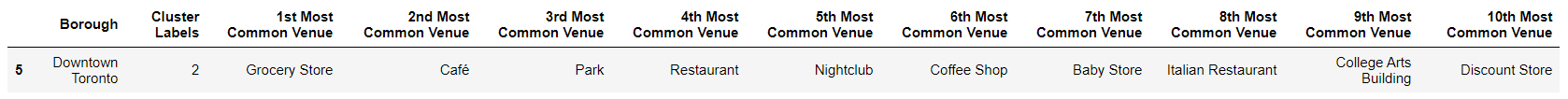
The resulting clusters were visualized on a map.

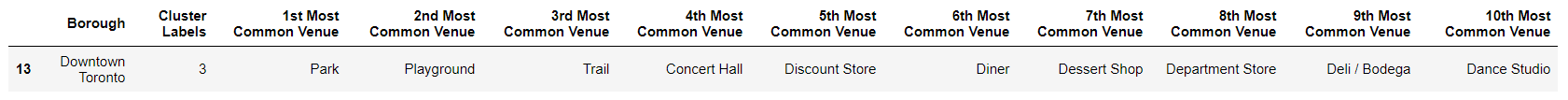
  
*Map of Manhattan indicating the locations of each neighborhood cluster.*

**4. Results**Both cities were segmented by 5 neighborhood clusters. Examining each cluster will allow for determination on distinctive venue categories that distinguish each cluster. Based on defining categories, each cluster is assigned a name. The clusters for Downtown Toronto are listed as follows:   
  
Cluster 1

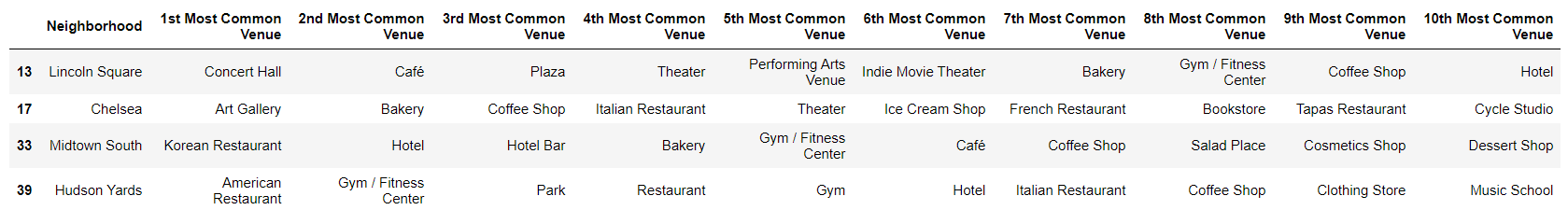
  
  
Cluster 2

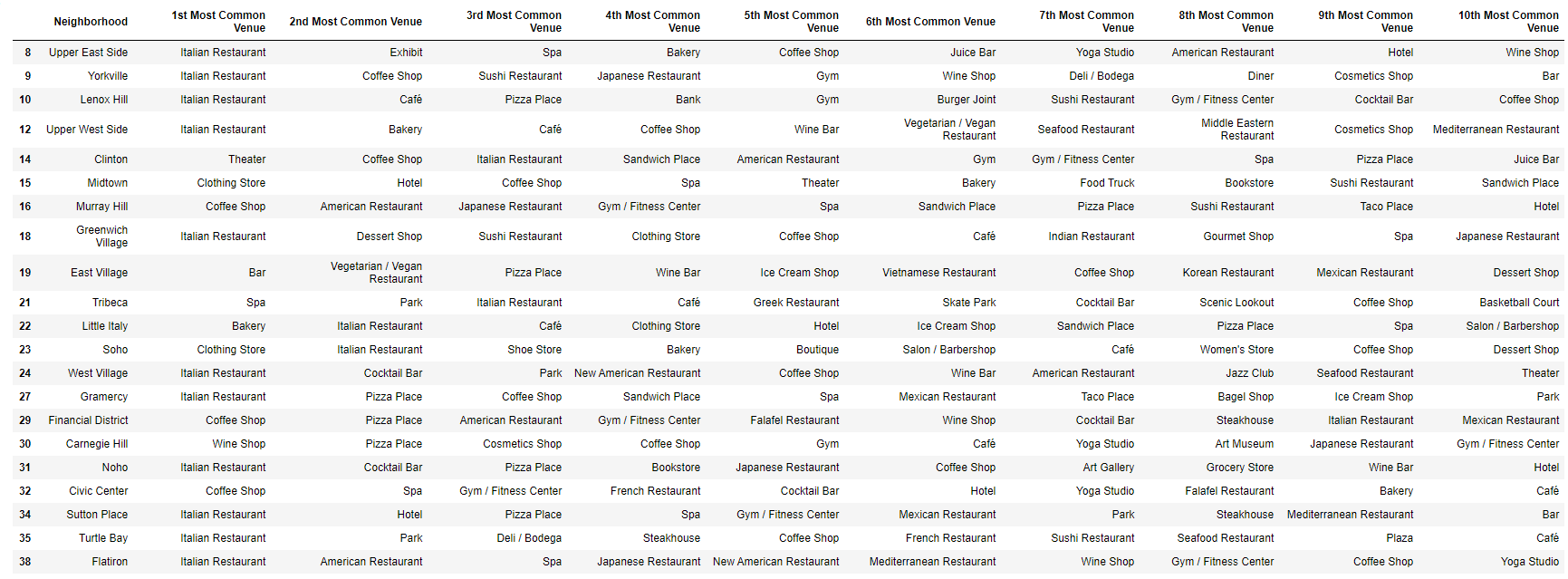
  
  
Cluster 3

  
  
Cluster 4

  
  
Cluster 5

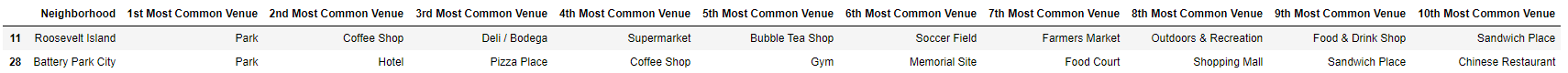
  
  
First, looking at cluster 1, it can be observed that the common venues were those of airport relation. This meant that this single borough was likely to be in proximity with the airport. In cluster 2, it seemed that the most common venue was coffee shops, parks, sandwich places or cafes for 10 boroughs. Clusters 3 and 4 have both single boroughs in them. In cluster 3, the first 3 common venues are those like cluster 2, however, in cluster 4, the top 3 common venues were park, playground and trail. Cluster 5 comprised of 5 boroughs in which, have a diverse set of venues ranging from coffee shops, to cocktail bars, to restaurants of diverse ethnic cuisines.   
  
The clusters of Manhattan are as follows:   
  
Cluster 1

  
  
Cluster 2

  
  
Cluster 3

  
  
Cluster 4

  
  
Cluster 5

  
  
Considering clusters 1, 2 & 3, they both have many neighborhoods in within their vicinities. A plethora of places such as concert halls, art galleries, diverse restaurants can be found. However, looking at clusters 4 and 5, these are clusters wherein there are only 1 or 2 neighborhoods. In cluster 4, the common venues are boat related which suggests that this neighborhood is near to the pier. In cluster 5, this neighborhood is close to the park as the 1st most common venue in both neighborhoods were denoted as parks.   
  
**5. Observations & Recommendations**  
  
When comparing both cities, they are both unique and diverse in their own ways. Toronto seems to be a quieter and calmer city compared to New York city. They both have pretty similar venues but it does seem that there are far more tourist attractions in the city of New York than in Toronto. For someone who loves the hustle and bustle lifestyle or would love to open a restaurant for the purposes of immigrating, the city of New York would seem more appropriate to be the recommendation to migrate to. Specifically, clusters 2 or 3 as they seem to be the best amongst the five clusters that were analyzed due to the proximities of many common venues to those neighborhoods.   
  
However, for someone who loves city lifestyle but rather not so “on-the-go”, Toronto would be more suitable. Cluster 1 would not be highly recommended as it is close by the airport which results in a lot of traffic noises which could be quite distracting and annoying. The best clusters would be cluster 2 as there are coffee shops, parks, sandwich places or cafes for over 10 boroughs.   
  
**6. Conclusion**  
  
In this study, the relationship between the clusters of each neighborhood and their performance was for each city was analyzed and reviewed. Both cities have their similarities as well as their dissimilarities. From one aspect, in terms of common venues, both cities have cafes, parks, coffee shops, gyms, shops, restaurants, etc. However, the difference lies were in historical landmarks and tourist attractions. From the maps, we can see that the island of Manhattan has a lot of venues for a smaller land area compared to Downtown Toronto.   
  
Both areas focused on this study are unique in their own ways. In terms of migration, if migrating to settle down in a quieter neighborhood, Toronto would be a more suitable city to do so. However, for the purposes of migrating in opening a restaurant, museum, etc. both cities are appropriate.