**Comparing Neighborhoods of New York City and Toronto**

Anuj Chaudhary

May 24, 2020

1. **Introduction**
   1. **Background**

New York City (NYC), is the most populous city in the United States. With an estimated 2018 population of 8,398,748. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined statistical area, it is one of the world's most populous megacities. New York City has been described as the [cultural](https://en.wikipedia.org/wiki/Culture_of_New_York_City), financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 as of 2016. Current to 2016, the Toronto census metropolitan area (CMA) held a population of 5,928,040, making it Canada's most populous CMA. Toronto is also an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada.

Therefore, it is advantageous to compare the neighborhoods in two big cities and understand how similar or diverse are the two cities. This information will be used by cross-cultural researchers and economists to understand how diversity could play a big role in a city’s success.

* 1. **Problem Statement**

We explored New York city and the city of Toronto and segmented and clustered their neighborhoods. Both cities are very diverse and are the financial capitals of their respective countries. The idea is to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. The comparison will be based of the venue categories associated with various venues within each city.

* 1. **Interest**

It will be interesting for the world to see how the different neighborhoods to world’s two biggest cities could be similar or dissimilar.

1. **Data acquisition and cleaning**
   1. **Data Sources**

We will start by looking at the different neighborhoods in both the cities. The data can be found for New York City and Toronto city [here](https://foursquare.com/explore?mode=url&near=New%20York%2C%20NY%2C%20United%20States&nearGeoId=72057594043056517) and [here](https://foursquare.com/explore?mode=url&near=Toronto%2C%20ON&nearGeoId=72057594044095801) respectively. Within each neighborhood, we will explore different venues to form our core dataframe for further analysis. For example, Marble Hill neighborhood in New York City has various venues like American restaurants, Donut shops, Bagel places, Yoga centers, etc.

* 1. **Data Cleaning**

Data downloaded from the sources were extracted into two tables. I used the Foursquare developer’s tool to fetch all the venues of all neighborhoods for each city. Next, to figure out the top ten venues under each neighborhood, count of the number of times a venue appears under a category was taken. This way we could use a mean to figure out the top ten neighborhoods. Now for each neighborhood, the top ten venues were captured in separate dataframes. Now we used the clustering algorithms to figure out the various clusters within each city and compared them against each other to know how similar or dissimilar they are against each other.

1. **Exploratory Data Analysis**

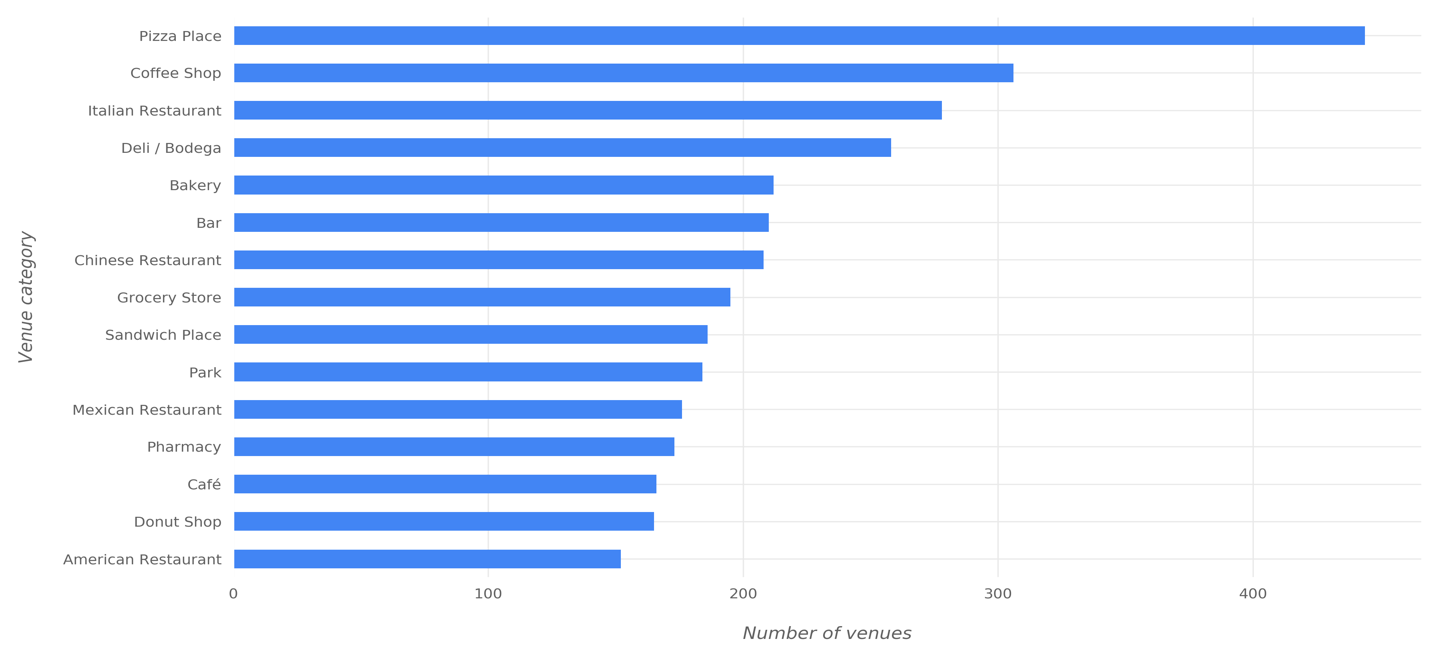
The datasets produced in previous section will be explored via effective visualizations to understand the two neighborhoods.

* 1. **Most Common Venue Categories**

We will start by looking at different venues in each neighborhood and understand the various categories associated with venues. The best way to look at these details is to create a bar plot of the number of occurrences of each venue category.

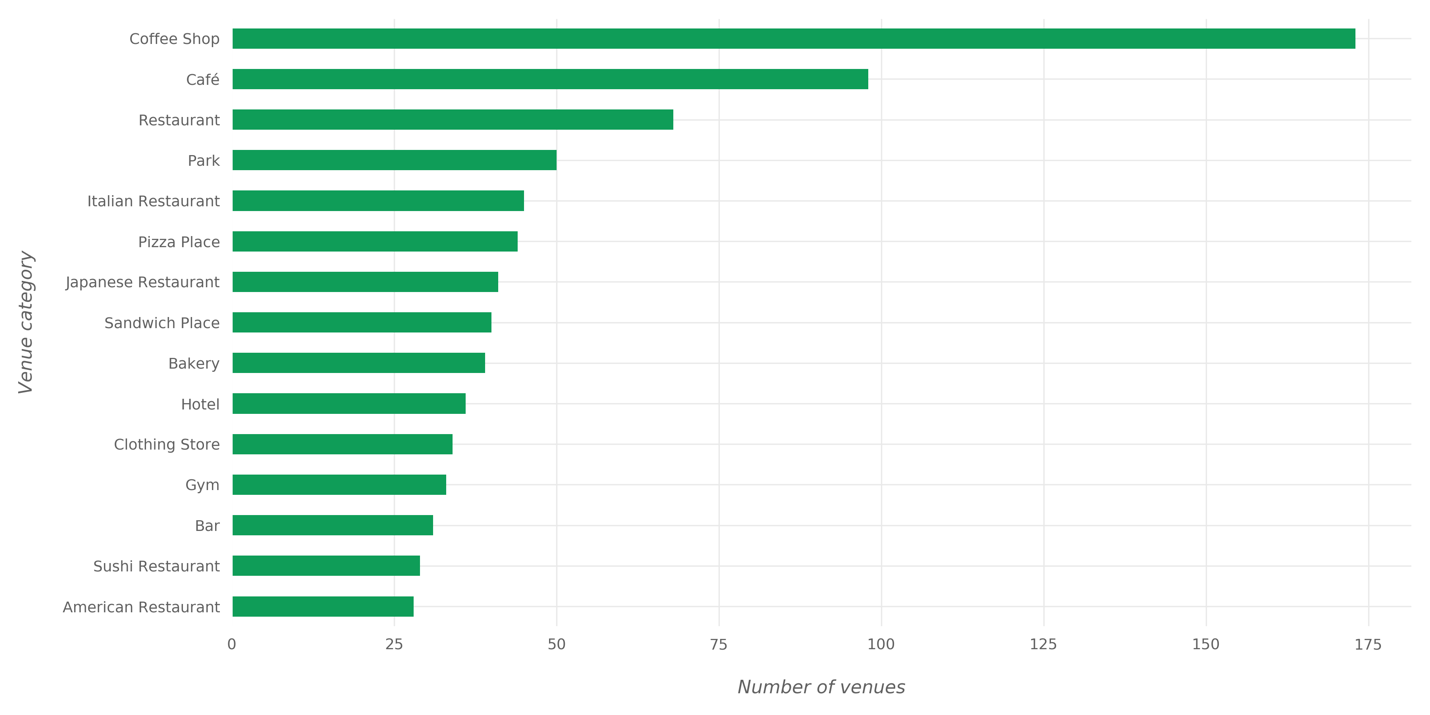
* + 1. **New York City**

Below figure shows a bar plot of the most common venues in NYC.



* + 1. **Toronto City**

Below figure shows a bar plot of the most common venues in Toronto.

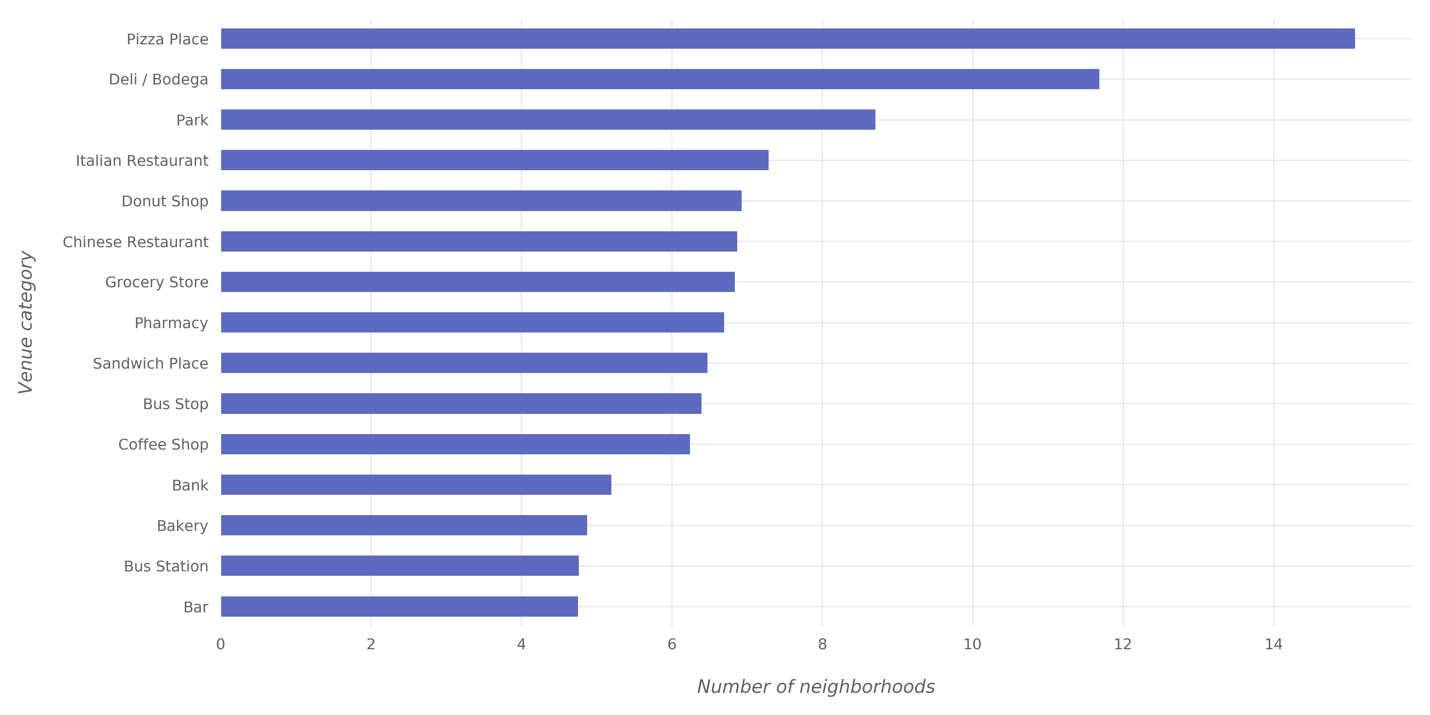


* 1. **Most Widespread Venue Categories**

We will also look at the most widespread venue category which is essentially telling us in how many neighborhoods does a venue category appear? We will also create a bar plot for each city

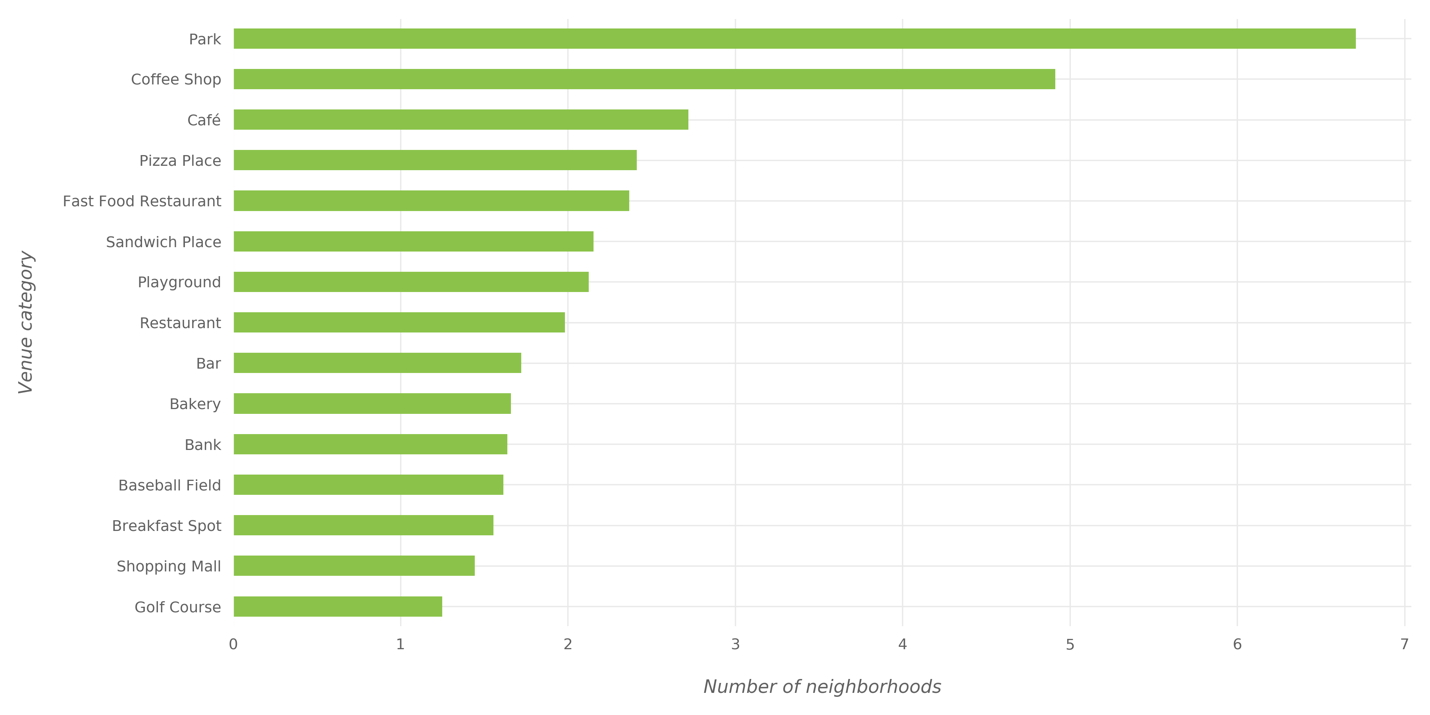
* + 1. **New York City**

Below image shows the most widespread venue categories in NYC.



* + 1. **Toronto City**

Below image shows the most widespread venue categories in NYC.

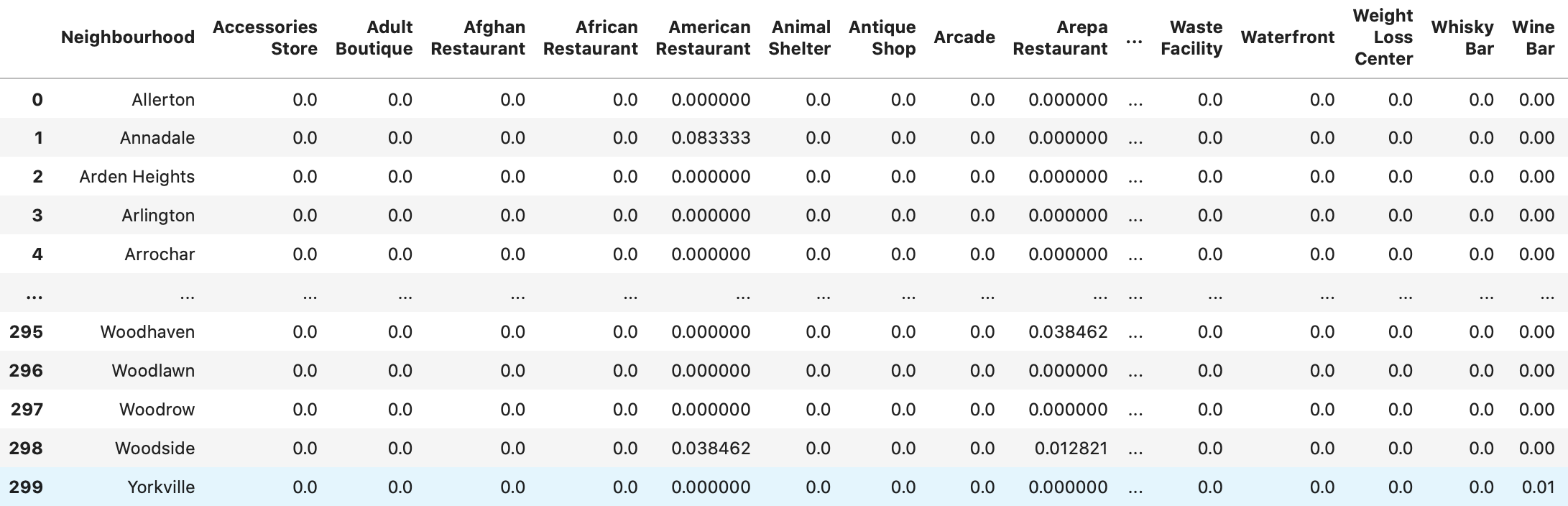


1. **Clustering of Neighborhoods**

We will apply K-means clustering algorithm of the Scikit-learn Python library to perform clustering of neighborhoods in each city.

* 1. **Feature Selection**

To perform clustering, we will start by performing one-hot encoding to the “Venue Category” column in New York City and Toronto city dataframes. The resulting dataframe looks like the one shown in figure below.



Result of one-hot encoding on NYC data



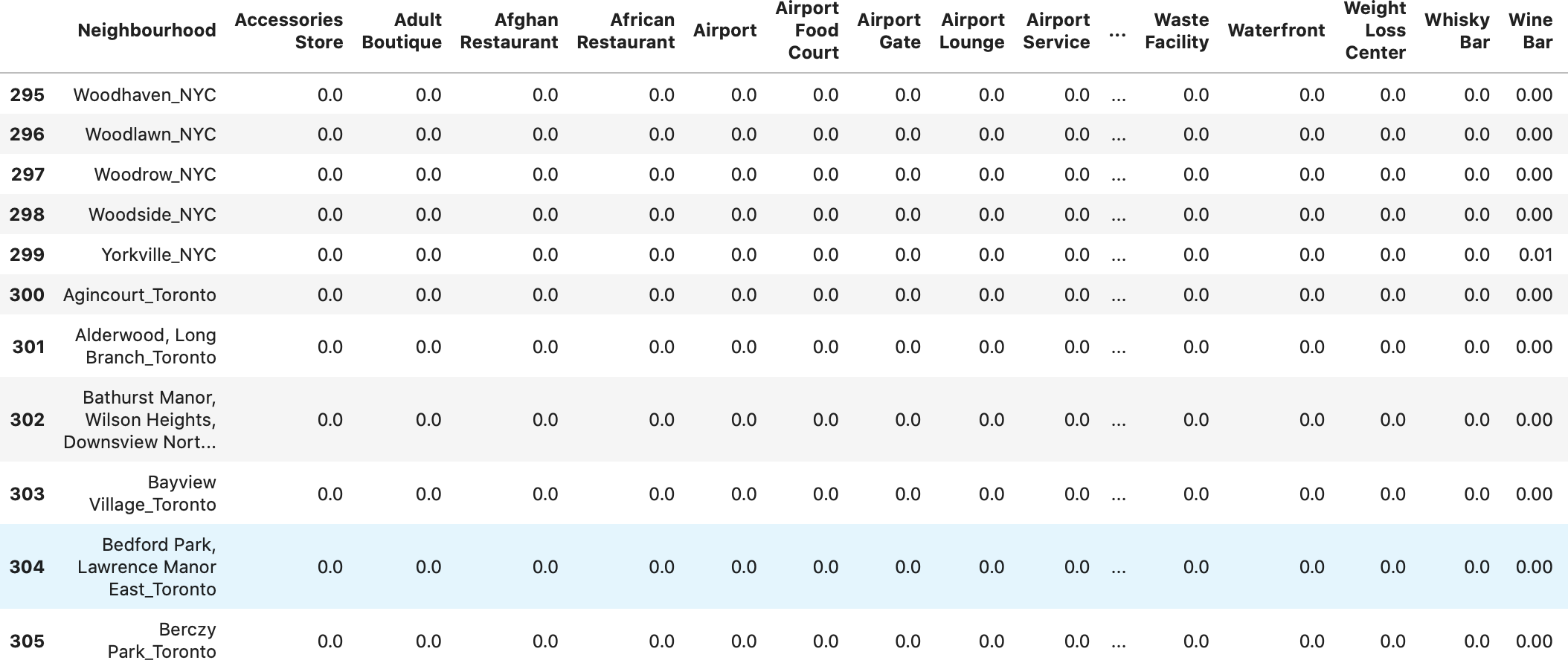
Result of one-hot encoding on Toronto data

The next step is aggregating the values for each neighborhood so that each neighborhood is represented in a single row. The aggregation will be done by grouping rows by neighbor- hood and by taking the mean of the frequency of occurrence of each category.

* 1. **Combining New York City and Toronto Data**

The aggregated dataframes should be combined before applying the clustering algorithm. In order to distinguish New York City neighborhoods from Toronto neighborhoods in the new dataframe, add a suffix to each neighborhood name: for NYC, the string to be added is “\_NYC” and “\_Toronto” for Toronto.

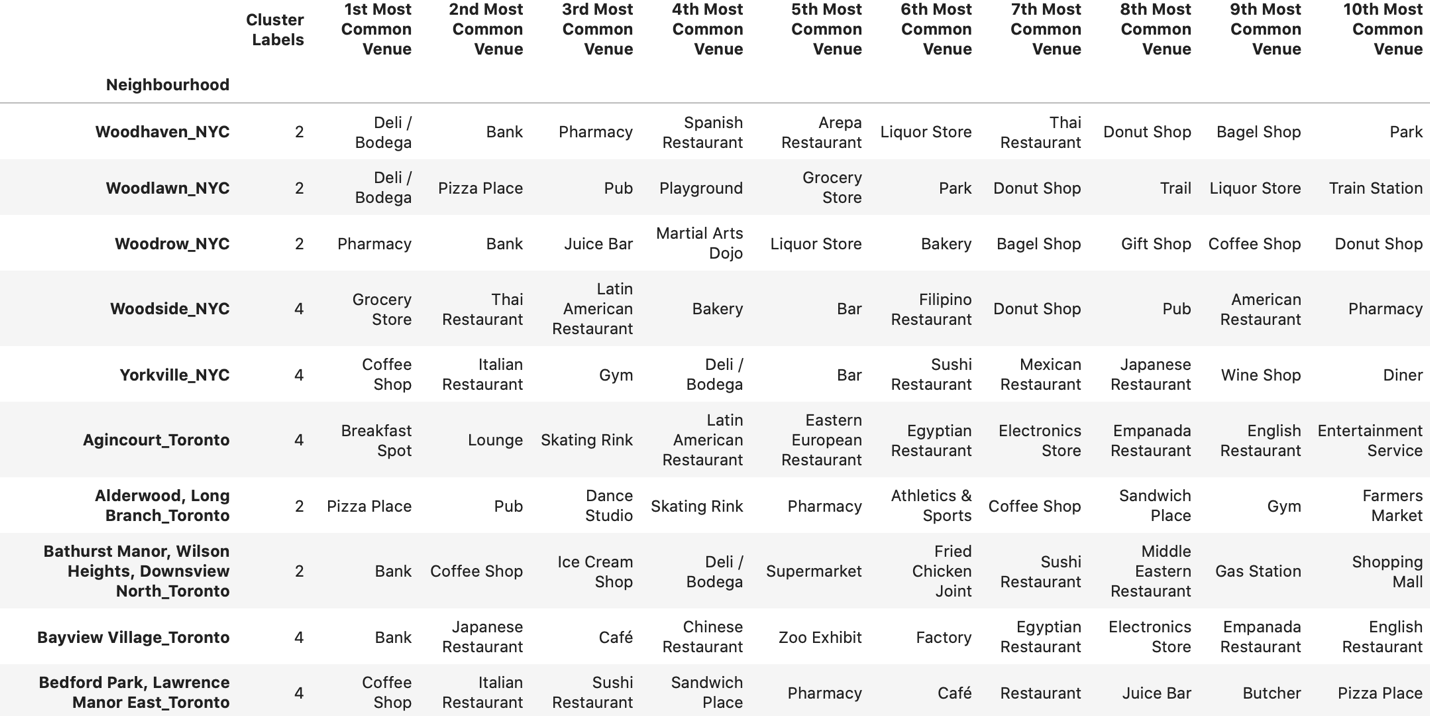
Also, NYC and Toronto don’t necessarily have the same venue categories. To deal with this issue before combining the dataframes, the columns of both dataframes are made the same by adding the columns that exist only in NYC dataframe to Toronto dataframe and vice versa; the newly-added columns have a value of 0 for all the rows.



The combination of NYC and Toronto aggregated dataframes data

* 1. **Clustering results**

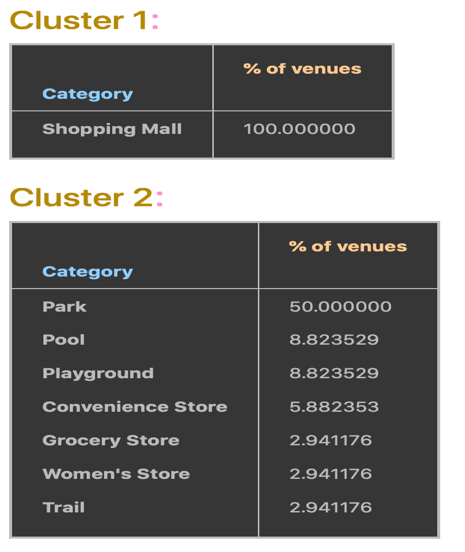
The clustering algorithm produced cluster-labels. These labels denote the cluster of each record (i.e. each neighborhood) in the data. Using these labels and the dataframe, a dataframe is constructed to show the neighborhoods of NYC and Toronto, the cluster to which each neighborhood belongs, and the most common venue categories in each neighborhood. This dataframe can be seen in below figure.

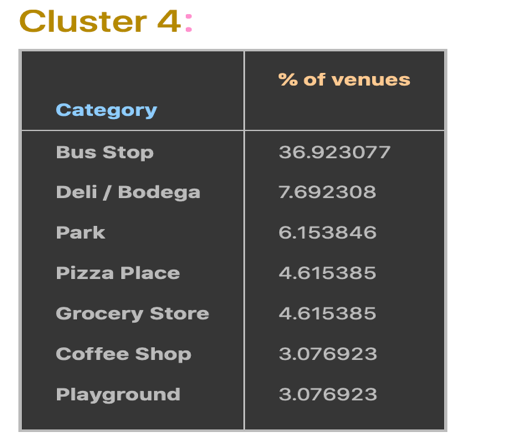


The output of the clustering operation is 5 clusters with cluster labels 0, 1, 2, 3, and 4. Each cluster is expected to contain a group of similar neighborhoods based on the categories of the venues in each neighborhood.

1. **Clustering Analysis**

The clustering algorithm grouped neighborhoods of NYC and Toronto in 5 clusters based on the similarity between their venues. Now, these clusters will be investigated to see the most common categories in each of them. Below figures show the most common 10 venue categories in each cluster.

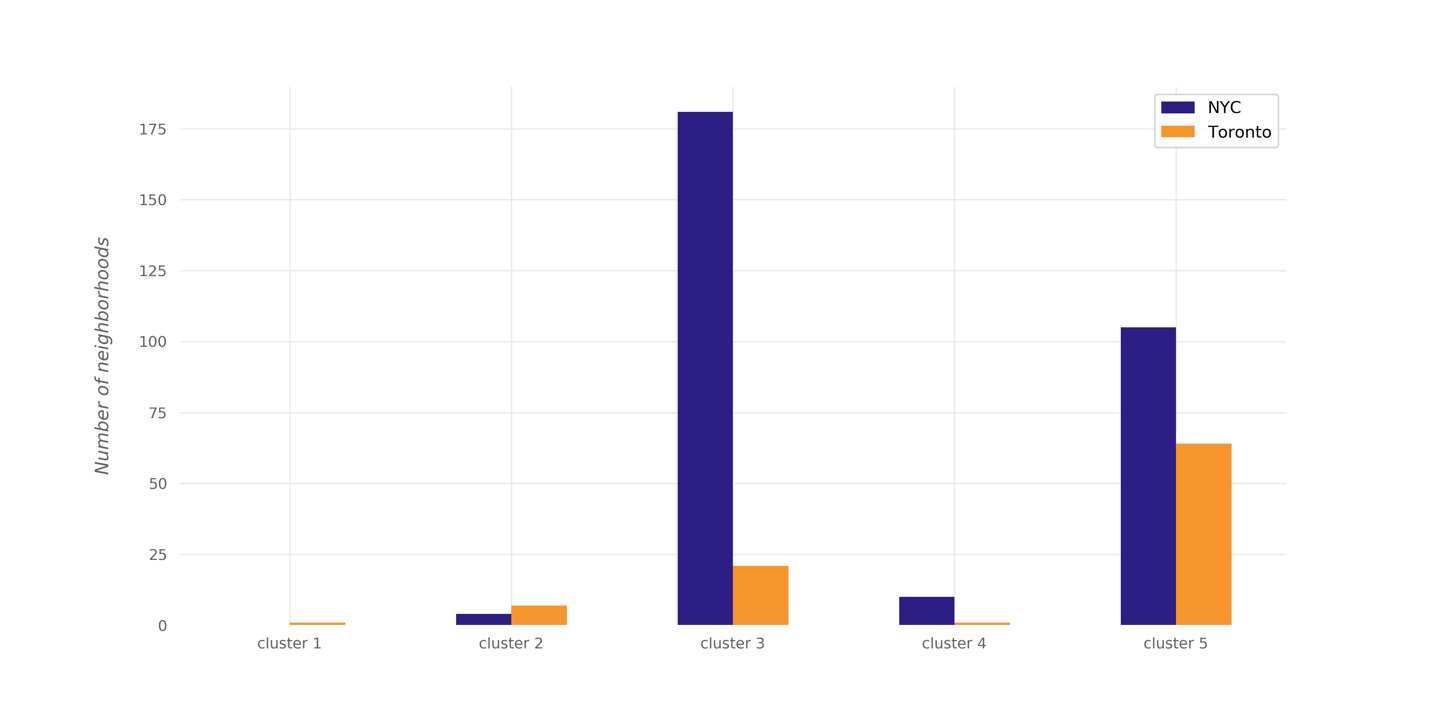
The differences between the clusters can be seen from the figure; each cluster distinguishably has different distribution of common venue categories than other clusters.

Some of the observations that can be made from the tables are:

1. While Pizza Place constitute ~6.65% of venues un the third cluster, they constitute ~4.6% of venues in the fourth cluster, and ~2.7% of the venues in the fifth cluster
2. Pizza places appear in the most common categories of the second, third, and fourth clusters only.
3. Park appear in the most common categories of the second cluster only.
4. Coffee shops appear in the most common categories of the fifth cluster only
5. Shopping mall is the only category in first cluster with 100% of venues.

Other differences can be observed as well.

Below figure shows the number of NYC neighborhoods and the number of Toronto neighborhoods in each cluster of the five resulting clusters.



1. **Conclusions**

In this project, the neighborhoods of New York City and Toronto were clustered into multiple groups based on the categories (types) of the venues in these neighborhoods. The results showed that there are venue categories that are more common in some cluster than the others; the most common venue categories differ from one cluster to the other. If a deeper analysis—taking more aspects into account—is performed, it might result in discovering different style in each cluster based on the most common categories in the cluster.