Business understanding

In this project I will analyze the cities of New York (USA) and Toronto (Canada) and cluster/ compare their neighborhoods regarding their businesses. I will investigate which businesses are common or not common in both cities and which neighborhoods are suitable for each business.

The aim is to find out similarities and differences in their business infrastructure between the two cities. The results from the data analysis should help people to decide in which neighborhood they should open a business and what are the specific business which thrive in this neighborhood.

Data understanding

1. Location data: Required is data about all the neighborhoods (Postal codes, Borough, Neighborhood) in those cities and the latitude/ longitude information from these neighborhoods.

1.1 New York: The location data from the NY neighborhoods is provided by the IBM data science course as a JSON file. This file contains all the required information (neighborhoods, latitude, longitude) and will be store in a pandas data frame.

1.2 Toronto: Unfortunately, there is no data which contains all the required information for Toronto like in the case whit the New York data. The information about all neighborhoods in Toronto can be retrieved from a Wikipedia page ('List_of_postal_codes_of_canada:_M'). The latitude and longitude data can be retrieved as a file (.csv) from: http://cocl.us/Geospatial_data Both information will be stored in a pandas data frame and the join on the Postal code attribute.

2. Venue data: For each city, data that describes the venues of its neighborhoods and the categories of these venues is needed. Venues data will be retrieved from Foursquare which is a popular source of location and venue data. Foursquare API service will be utilized to access and download venues data. The API call will return a data frame with venue information for a given neighborhood with neighborhood name, latitude and longitude data. This venue data can then be analyzed to cluster the neighborhoods in both cities and compare the clusters to find similarities / differences. The data for the venues in each city will be retrieved through the Foursquare API with my personal credentials.

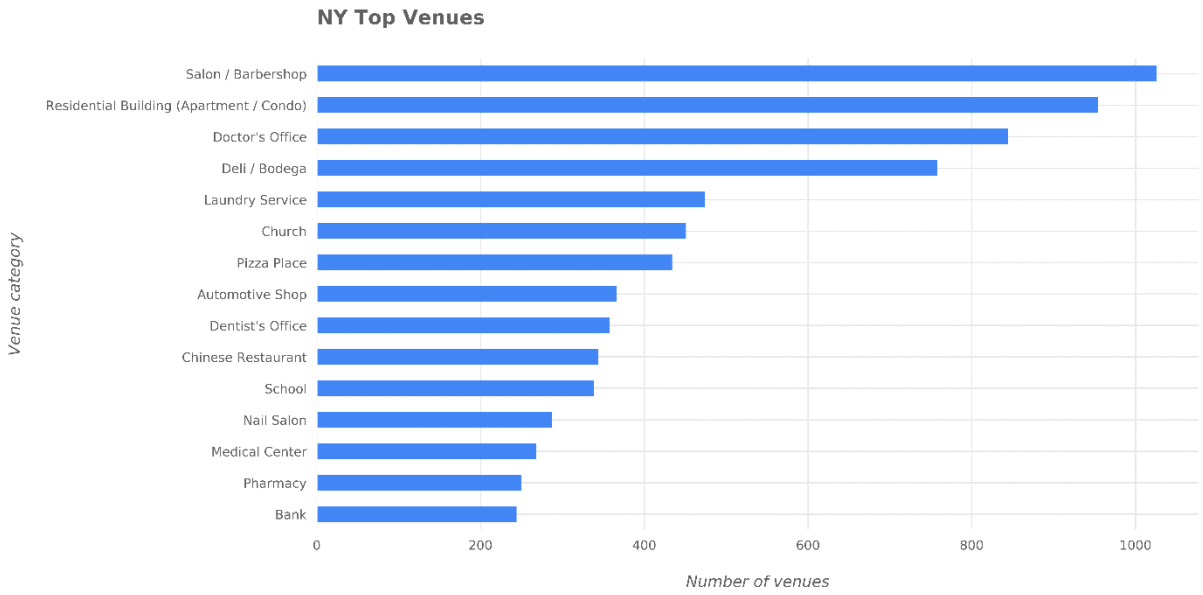
Methodology

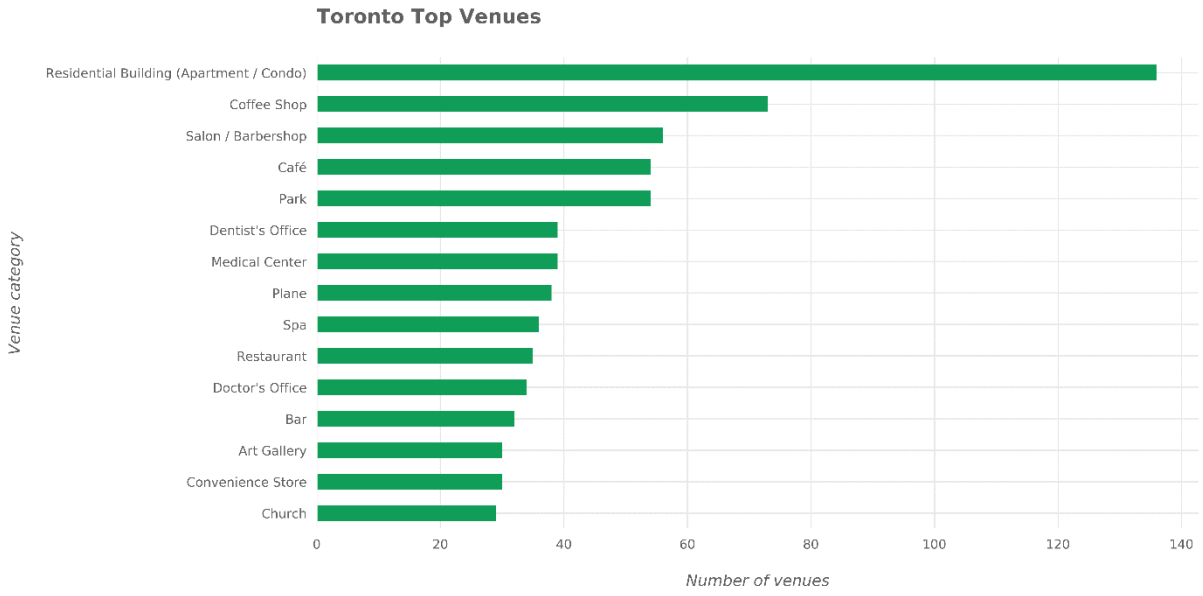
The approach for this data science projects will be in two steps. First, I will apply an exploratory analysis on the venue data for both cities, in order to get a better understanding about the neighborhood in venue structure in those cities.
Second, I will apply clustering on NYC and Toronto neighborhoods to find similar neighborhoods in the two cities. Clustering is the process of finding similar items in a dataset based on the characteristics of items in the dataset. In particular, K-means clustering algorithm of the Scikit-learn Python library will be used. To be able to perform clustering,

Results/ Discussion

1. Exploratory Analysis

For a better understanding of the venues in those cities, number of occurrences is counted for each venue category. The resulting data is displayed in the following bar plots.





Comparing the venues of both cities, it can be seen that there are multiple venues, with are identical among the most common venues (Residential building, Church, Medical Center). There are also some venues which are very specific for NY (Deli/ Bodega) and Toronto (Coffee Shop).

2. Clustering

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. The following figures show the most common 7 venue categories in each cluster; for each common category, the percentage of venues of that category in the neighborhoods of the cluster is shown also.

Cluster 1:

Category	% of venues
Doctor's Office	14.160225
Residential Building (Apartment / Condo)	4.708363
Dentist's Office	3.935348
Salon / Barbershop	2.881237
Medical Center	2.846100
Deli / Bodega	2.529866
Laundry Service	1.897400

Cluster 2:

Category	% of venues
Residential Building (Apartment / Condo)	3.047285
Salon / Barbershop	1.751313
Beach	1.611208
Bar	1.541156
Park	1.523643
Coffee Shop	1.488616
Art Gallery	1.436077

Cluster 3:

Category	% of venues
Salon / Barbershop	9.027222
Deli / Bodega	4.903923
Church	3.422738
Laundry Service	2.922338
Residential Building (Apartment / Condo)	2.922338
Chinese Restaurant	2.221777
Pizza Place	2.201761

Cluster 4:

Category	% of venues
Residential Building (Apartment / Condo)	15.447154
Salon / Barbershop	3.492924
Doctor's Office	3.432701
Deli / Bodega	2.770250
Laundry Service	2.619693
Church	1.866908
Dentist's Office	1.656128

Cluster 5:

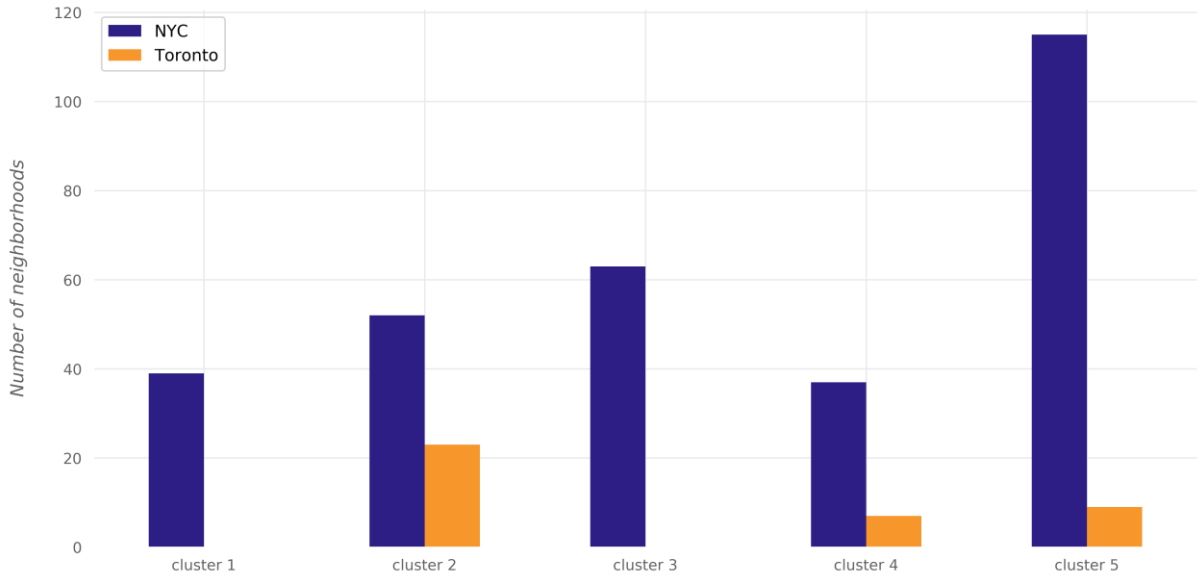
Category	% of venues
Salon / Barbershop	3.441505
Deli / Bodega	3.017776
Automotive Shop	2.470029
Doctor's Office	2.439024
Pizza Place	2.087640
Laundry Service	1.684580
Chinese Restaurant	1.539893

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 of Figures are:

- Residential buildings appear in the first four cluster with different percentages, ranging from 15% to 3%
- The first cluster has a large medical infrastructure with doctor's and dentist's office and medical center. In total 21% of medical venues.
- The second cluster is the only cluster with outdoor venues like beaches and parks. It's also the only cluster with bars and coffee shops.
- Salon/ Barbershop appears in every cluster

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



Conclusion

In this project, the neighborhoods of New York City and Toronto were clustered into multiple groups based on the categories 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.