Table of Contents

[1. Introduction 2](#_Toc50126004)

[1.1. Background 2](#_Toc50126005)

[1.2. Problem 2](#_Toc50126006)

[1.3. Target Audience 2](#_Toc50126007)

[2. Data acquisition 2](#_Toc50126008)

[2.1. Data Sources 2](#_Toc50126009)

[2.1.1. Source data for Toronto City 2](#_Toc50126010)

[2.1.2. Source data for New York City 4](#_Toc50126011)

[3. Methodology 5](#_Toc50126012)

[3.1. Data Cleaning 5](#_Toc50126013)

[3.2. Data Exploration 6](#_Toc50126014)

[3.3. Machine Learning 7](#_Toc50126015)

[3.4. Data Analysis 9](#_Toc50126016)

[3.4.1. Comparison of the top ten venues in Toronto and New York 9](#_Toc50126017)

[3.4.2. Comparison of the number of venues per Borough in Toronto and New York 11](#_Toc50126018)

[3.5. Cluster Analysis 13](#_Toc50126019)

[3.5.1. Comparison of number of venues per cluster in Toronto and New York 14](#_Toc50126020)

[3.5.2. Comparison of Top ten of the first most common venues from the 5 clusters in Toronto and New York 16](#_Toc50126021)

[4. Discussion 21](#_Toc50126022)

[5. Conclusion 21](#_Toc50126023)

**Comparison of Toronto and New York Cities**

# Introduction

## Background

Toronto and Network cities are the most vibrant cities in Canada and USA, respectively. Both cities in turn are divided into smaller areas known as Boroughs which are the administrative regions for these smaller areas. The Boroughs are further divided in to neighbourhoods. One basic measure of that determines the vibrancy of a city is the available facilities and services. Exploring the number of and kinds of venues in a city is advantageous to examine the progress of the city and ease of access for people in demand. Basically, it would help to discover the nearest venues within a given radius for explorers, and knowing the distinguishing the venues helps for comparison with other cities.

## Problem

New York and Toronto cities are very diverse and are the financial capitals of their respective countries. This project aim is to compare the neighbourhoods of the two cities and determine how similar or dissimilar they are. We examine whether New York City is more like or completely different from Toronto.

## Target Audience

Firstly, the city administrators would be very interested in knowing the number and types of venues available in their cities, for estimating the progress of the cities and identifying the lacking services. Others who are interested in the results of the comparison of the two cities include such as tourists, investors, residents, and immigrants who want to work and live in a best city.

# Data acquisition

## Data Sources

The dataset mainly consists of Boroughs, neighborhoods, and geographical coordinates of the neighborhoods (latitude and longitude) for both cities.

### Source data for Toronto City

For Toronto city, the dataset is not freely available on internet, so we fetch it from two sources.

**Source 1:**

The postal codes address dataset is available in Wikipedia paged. Hence, we scrape <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> to get the list of Postal codes, Boroughs, and neighborhoods.



Figure 1: Wikipedia page that contains the postal codes in Canada

The dataset has many problems, so we scape the data, wrangle it, clean it, and then read it into a pandas dataframe so that it is in a structured required format.

First, we create a Beautiful Soup package which is used to parse the html, that is, take the raw html text and break it into Python objects. Since the html contains a text and a table, we used a specific script to get the list of tables in the website. There are two tables in the site; we filter out the table of our interest which is the first table in this case.

Second, each retrieved row which contains Postalcode, Borough, and neighborhood in the dataframe that are separated by newline feed (\n). Since a Neighborhood may contain multiple cities separated by comma, we use newline break to split each row in to three columns instead of comma. The column names are exctracted using a BeautifulSoup object and get\_text() function, then convert to a dataframe and merge with the previous data. By removing unnecessary marks and renaming the column names, we modify the dataframe as per the required structure.

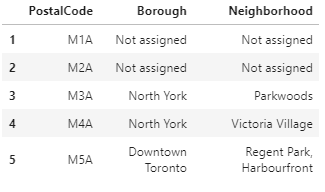


Figure 2: Postal address in Canada scraped as pandas data frame

**Source 2:**

The geographical coordinates for the neighborhoods are available as csv file on another site, <http://cocl.us/Geospatial_data>, and we read the csv file to obtain latitude and longitude values.

|  |  |
| --- | --- |
| Figure 3: Geographical data of Neighborhoods in Toronto | Figure 4: Conversion of csv file into pandas data frame |

After reading the csv file which contains the geographical coordinates of each postal code from the provided url, we modify the column names to match with scraped dataset.

### Source data for New York City

For New York City the dataset is freely available on <https://geo.nyu.edu/catalog/nyu_2451_34572> , but it is downloaded and resides on the skills Networks lab which has 5 boroughs and 302 neighbourhoods. The dataset contains 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood. Hence, we simply run *wget* command and access the data. Since the file is json type, we transform this data of nested Python dictionaries into a pandas dataframe.

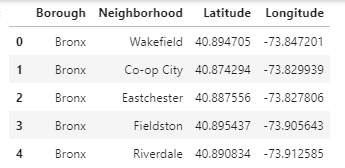


Figure 5: Dataset for New York City

# Methodology

## Data Cleaning

**Toronto dataset:**

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made:

1. Only the cells that have an assigned borough will be processed. Borough’s that were not assigned get ignored.
2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighbourhoods separated with a comma.
3. If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

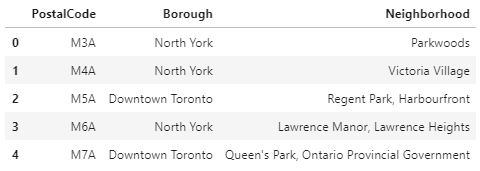


Figure 6: scraped and cleaned dataset for Toronto

Then, we merge the first dataframes which contains postalcode, borough, and neighborhood with the dataframe of geographical coordinates based on the postalcode.

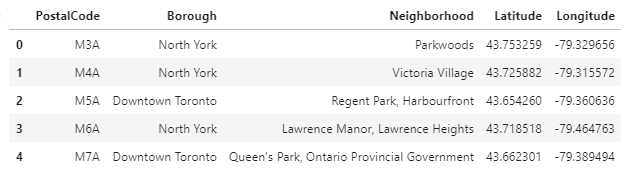


Figure 7: Postal code and geographical coordinates for Toronto

## Data Exploration

Having cleaned the dataset, we have 10 Boroughs and 103 PostalCode(neighborhoods) for Toronto, and 5 Boroughs and 302 neighborhoods for New York City. The dataframe for Toronto city contains both PostalCode and neighborhood fields. Since the PostalCode is provided with latitude and longitude values, we ignore the neighborhood for the time being and use Postalcode values inplace of neighborhood and use this to explore the nearest venues.

Furthermore, we convert Toronto and New York into their equivalent latitude and longitude values. We apply the Foursquare API to explore neighborhoods in both Cities.

|  |  |
| --- | --- |
| Figure 8: Toronto city with neighborhoods superimposed on it | Figure 9: New York city with neighborhoods super imposed on it |

We use the explore function to get the venues within 500 meter radious of each neighbourhood. After exploring the neighborhoods using the Foursqare API, we get 2146 venues in Toronto and 10128 venues in New York which is almost five times that of Toronto.

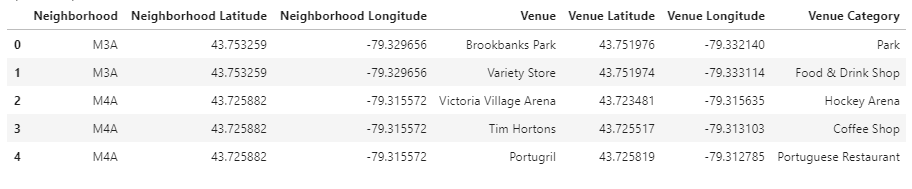


Fig 10: Venues within 500 meter radius in Toronto neighborhoods



Figure 11: Venues within 500 meter radius in New York neighborhoods

## Machine Learning

Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called **One hot encoding**. For each of the neighbourhoods in Toronto and New York City, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.

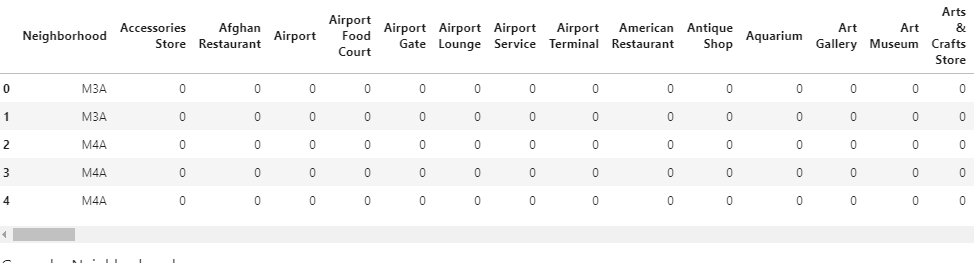


Figure 12: One hot encoding for Toronto venues

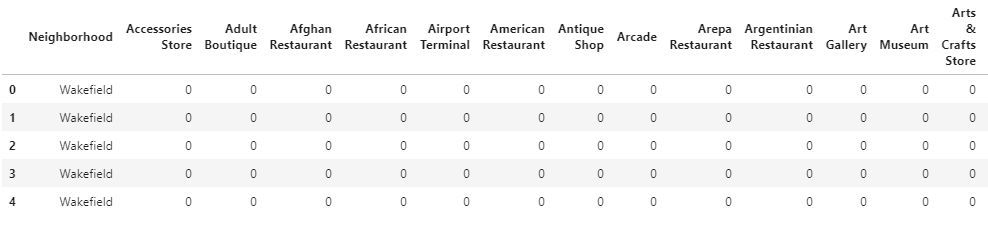


Figure 13: one hot encoding for venues in New York City

Then we grouped the venues based on similar neighborhoods in both cities and by taking the **average** of the frequency of occurrence of each Venue Category.

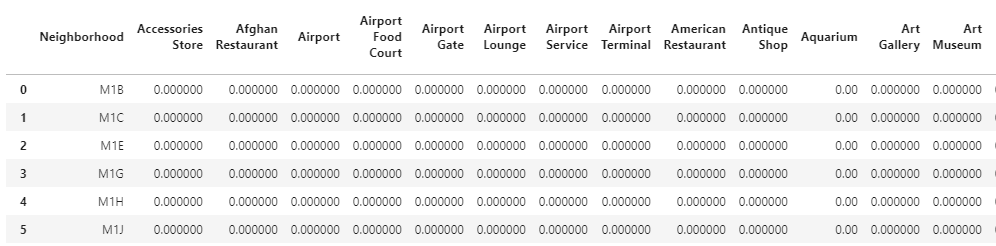


Figure xdf: Venues in Toronto grouped by neighborhood

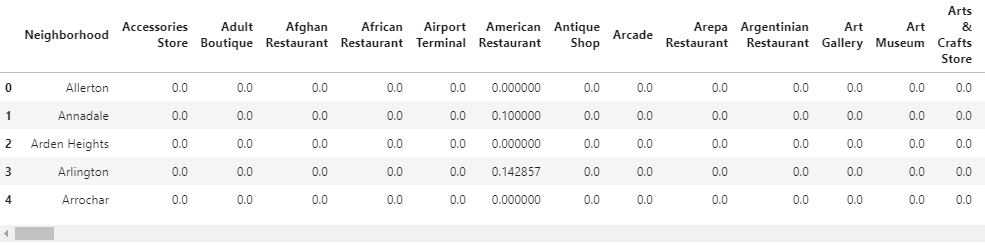


Figure xy: Venues in New York grouped by neighbourhood

**K-means Clustering**

To make the analysis more interesting, we wanted to cluster the neighbourhoods based on the neighbourhoods that had similar averages then use this feature to group the neighborhoods into clusters. The clustering is implemented using K-means clustering algorithm. For convenience we use, k to be 5 in both cities. Moreover, we use the Folium library to visualize the neighborhoods in New York City and their emerging clusters.

|  |  |
| --- | --- |
| **Toronto** | **New York** |
| Figure 14: Map of Toronto venues with 5 clusters | Figure 15: Map of New York venues with 5 clusters |

The Legends are described below

|  |  |
| --- | --- |
| **Color** | **Cluster** |
|  | 1 |
|  | 2 |
|  | 3 |
|  | 4 |
|  | 5 |

Table 1: Cluster legends

## Data Analysis

So far we have seen the neighborhoods, venues, and cluster results for Toronto and New York cities. In this section, we analyse our data and make comparisons between the two cities with different parameters.

### Comparison of the top ten venues in Toronto and New York

In order to know the most common types of venues in each city, we sum up similar venues based on their category and sorted them in descending order. So, we examine only the top ten venues from each city and the result is depicted in as follows:

|  |  |
| --- | --- |
| **Toronto** | **New York** |
|  |  |
|  | |

Figure 16: Top 10 venues in Toronto and New York

As we can see, the most common venue in Toronto is Coffee shop whilst Pizza place is in New York. Even though café takes the second place in Toronto, it is not available in the top ten lists of venues in New York. In general, coffee shops and cafes are the most common in Toronto followed by food zones, but it is the reverse in New York City.

### Comparison of the number of venues per Borough in Toronto and New York

We have 10 and 5 Boroughs in Toronto and New York, respectively. Then we computed the total number of venues per each Borough in each city, and we get the following results.

|  |  |
| --- | --- |
| **Toronto** | **New York** |
|  |  |
|  |  |

Figure 17: Total number of Boroughs and venues per Borough in Toronto and New York

From Toronto, Downtown Toronto has the largest number of venues above 1000 while Mississauga has the least number of venues which is just below 15. On the other hand, Manhattan comprises the largest venues nearly above 3000 whilst Staten Island has the least venues approximately 1000. Even though the number of Boroughs in Toronto is twice of the New York, the most vibrant city in Toronto has nearly equal number of venues with the Boroughs that are smallest in New York. All in all, the number of venues per Borough in Toronto is very few compared to New York City.

## Cluster Analysis

For analysing the clusters, we sorted the venues in each neighbourhood in descending order and take into consideration only the top ten most common venues.



Figure 18: Top 10 venues from each neighborhoods in Toronto

|  |
| --- |
|  |
|  |

Figure 19: Top 10 venues from each neighborhoods in New York

After Clustering, we examine each cluster separately based on their cluster labels. Here is sample of cluster 2 dataset from Toronto and New York.

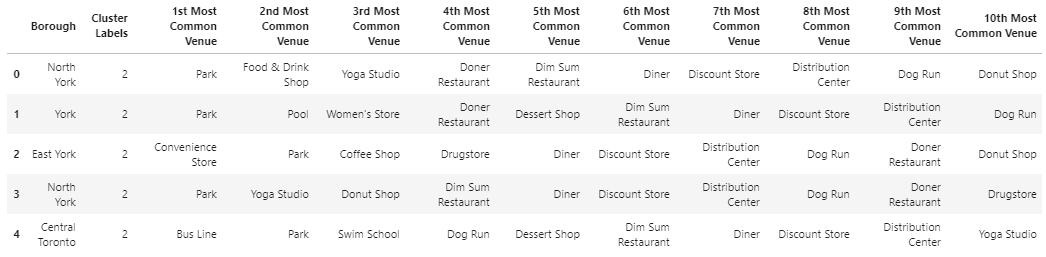


Figure 20: Cluster 3 datasets for Toronto

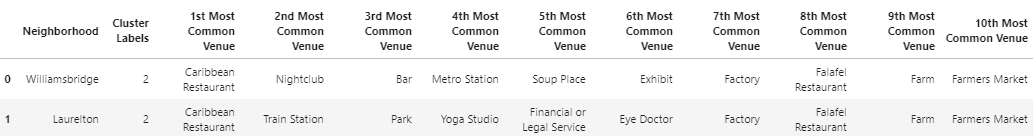


Figure 21: Cluster 3 datasets for New York

To compare and contrast the two cities, we have made basic analysis with the ten most common venues dataset from both cities and compared the results using different criteria as follows.

### Comparison of number of venues per cluster in Toronto and New York

In this section, we further analyse each cluster and compare similar clusters in both cities in terms of the number of venues. We have added all the venues in each cluster, and the output is depicted as follows:

|  |
| --- |
|  |
|  |

Figure 22: Total number of venues per cluster in Toronto and New York

From the clustering results, we can deduce that cluster 1 comprises the largest number of venues in both cities, but the total number of venues in New York is approximately three times as much as that of Toronto. On the contrary, Cluster 2 contains the least and equal number of venues in both cities. In general, the total number of venues for New York is greater than that of Toronto in each cluster but Cluster 5.

### Comparison of Top ten of the first most common venues from the 5 clusters in Toronto and New York

Out of the top ten most common clusters, we have taken only the 1st most common venues from each cluster. We have grouped them by their venue category, and below is shown cluster 3 results from both cities for illustration purpose.

|  |  |
| --- | --- |
| **Toronto** | **New York** |
|  |  |

Table 2: Top 10 venues out of the 1st most common venues in Cluster 3 of both Toronto and New York

From this we can see that in Toronto, there are only three venue categories while in New York we have only one category. So, we order the venues in descending order. As a result, we compare the top ten out of the first most common venues in the 5 clusters from both cities. The result is illustrated in the following:

|  |  |
| --- | --- |
| **Toronto** |  |
|  |
| **New York** |  |
|  |

Figure 23: Top ten venues out of the 1st most common venues in each cluster in Toronto and New York

When we combine the 5 clusters, we get ***NaN*** value for some venue category. The reason behind is that, the venue category is not within the top ten list of the 1st most common category in the specific cluster. As a result, it is automatically assigned null by the system. From the graphs we can deduce that Cluster 1 in both cities comprises almost all venue categories. However, in Toronto coffee shop, grocery store, and café are the top three venues whilst Italian restaurant, pizza place, and coffee shops are the top in New York. On the other hand, cluster 2 to cluster 5 in Toronto contains only one type of venue category each, i.e., Baseball field, fast food restaurant, pizza place, and park respectively. Similarly, Cluster 2 and Cluster 5 in New York contain park and playground venues, respectively. However, Cluster 3 and Cluster 4 comprise multiple venue categories.

# Discussion

After exploiting each neighborhood, we have obtained that New York has nearly 5 times as much venues as Toronto. This implies that the first city has more services to offer to residents and the target audience. Furthermore, the most common venues from each city are completely different. If we take, for example, the top three venues from each city we observe that food zones are in the list in New York whilst hot drink shops and cafes account for the top category in Toronto. To be specific, Pizza place, Italian restaurant, and Coffee shops are among the top in New York where as Coffee shop, café, and restaurant are in top three lists. As we go further, we noticed that parking is given greater attention in Toronto, but it’s not even in the top ten lists in New York. From this we can deduce that vehicle owners in Toronto would not face much difficulty when they are using other services. Even though both cities have Italian restaurants, Toronto has Japanese restaurant while New York has Mexican restaurants in their top ten lists. This indicates that there are more Japanese descendants in Toronto whereas New York has more Mexicans.

Moreover, there are 10 Boroughs in Toronto while there are only 5 Boroughs in New York. Each Borough in Toronto comprises fewer venue categories compared to the Boroughs in New York. The distribution of the venues in each borough implies that services in Toronto are dispersed and are scarce where as in New York there is plenty of access and choice within a Borough.

After clustering, we see that the total number of venues in the 5 clusters is much greater in New York compared to that of Toronto. Furthermore, most of the venue categories are type 1 clusters. This cluster contains almost all facilities in each respective city. However, the most common types of venues in this cluster are still in the order of the top list of venues in each city as described above.

All in all, New York is more vibrant compared to Toronto as it owns larger number and varieties of venues in each neighborhood within 500 meter radius. Hence, the target audience can easily get the services they require in nearest distances easily.

# Conclusion

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighbourhoods of Toronto and New York, and get a great measure of data from Wikipedia which we scraped with the Beautifulsoup Web scraping Library. We also visualized utilizing Matplotlib libraries. Similarly, we applied Folium to picture the clusters of both cities on a map.

Places that have room for improvement or certain drawbacks give us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize all the venues without filtering out only the top ten venues and get a better understanding the similarity or dissimilarity of the cities. Furthermore, we only visualize only the top ten of the 1st most common venues, but this can be extended to include the 2nd, 3rd, and so on for deeper understanding of the characteristics of the two cities.