The Battle of Neighborhoods

Manhattan vs Toronto

By Asavari Paluskar

1. Introduction and Background:

(A) Neighborhoods Analysed:

New York and Manhattan:
The city of New York is the most populous city in USA. It is home to a wide variety of businesses and is the financial capital of the US. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States.

A wide customer base and ever increasing population facilitates the growth of various types of businesses in New York. This also means that competition is high in New York, and aspiring entrepreneurs have to be careful while setting up a new business. Being the financial capital of the US, the cost of starting a new business is also one of the highest.

Manhattan is one of the largest neighborhoods in New York. The size of New York, as a whole, is much larger as compared to Toronto, Canada. Hence, we have narrowed down our study to the area of Manhattan in particular, in order to make it comparable.

Toronto:
Toronto is the most populous city in Canada. Toronto is an international centre 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.

The population of Toronto is ever increasing, which makes it an ideal location for starting a new business venture. The increasing number of immigrants also means that the competition in business environment is ever increasing.

(B) Problem Description:

Starting a new business venture in a city like New York or Toronto is very risky, considering the ever increasing competition and the high cost of establishing and operating a business. Hence, it is necessary for an aspiring entrepreneur to study the customer base thoroughly and select an area of business that is most favoured by the target audience and has the highest customer foot-fall.

Various types of businesses such as restaurants offering various cuisines, cafes, coffee shops, amusement parks, gyms, bars, bakeries etc are operating in Manhattan and Toronto. In order to be able to set-up a new business in these geographical areas, one needs to analyse the customer market and find the type of business that is most visited by customers in the Manhattan and Toronto.

Further, the demand structure in Manhattan and Toronto are very different, so deciding on establish the same type of business in both geographical regions would not be a profitable option. Hence, we need to analyse data of both geographical regions separately, visualize it and find the most profitable business venture in each geographical area, in terms of customer foot-fall.

(C) Target Audience:

ABC & Co. wishes to set up a new business in Manhattan and Toronto each. To recommend the correct type of business venture, ABC & Co. has appointed me as the Data Analyst. The objective is to analyse and explore the data of Manhattan and Toronto in terms of the most favoured type of activity for the customers and find out the venue which has the maximum number of visitors (customer foot-fall), in order to be able to select the best business venture in each of the geographical regions.

(D) Success Criteria:

The success of the analysis will depend on being able to find out the most favoured activities/areas of business by the customers and being able to recommend the correct business sector in Manhattan and Toronto to ABC&Co.

2. Data:

(A) Libraries and dependancies used:

Numpy, Pandas, json, geopy, matplotlib, kmeans clustering, folium and BeautifulSoup

(B) Geographic Regions Analysed:

New York:

We will be using the following dataset to analyse New York data: "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json"

We start by exploring the data and transforming it in a pandas dataframe. There are 5 boroughs and 306 neighborhoods in New York. After creating a dataframe and arranging it in boroughs and neighborhoods, we use 'geopy' to get the geographic co-ordinates (latitude and longitude data) for New York.

	Borough
Queens	81
Brooklyn	70
Staten Island	63
Bronx	52
Manhattan	40

Using 'Folium', we create a map of New York, using the latitude and longitude data extracted using geopy. Folium is a visualisation library.
We segment the data using foursquare.

Manhattan:
The New York dataset has around 10,040 neighborhoods, making it much larger than the Toronto dataset. Hence, we have scaled down and sliced the data, and only selected the data for Manhattan.

We follow the same procedure of finding the geographical co-ordinates for Manhattan using 'geopy' and create a map using 'folium. We create a dataset using the data and segment it using foursquare. The Manhattan dataset has 3,183 neighborhoods, which is comparable to the Toronto dataset. Hence, we proceed with segmenting and clustering the data.

We use onehot encoding to group the data based on the most common venues visited by customers in different neighborhoods. We use KMeans clustering to divide the data in five clusters and examin the clusters. We use Folium to visualise the map of Manhattan with each cluster being marked in different colors.

Toronto:
Unlike the New York data, the Toronto data is not readily available in the internet. We use the data from the following Wikipedia page: "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

We use BeautifulSoup to get the data from the Wikipedia page and transform it into a dataframe containing 3 columns. The next part of the analysis is the same as for Manhattan dataset. We get the latitude and longitude data for each location from Foursquare and extract the data relevant to Toronto.

Toronto dataset has 1,602 neighborhoods, which is comparable to the Manhattan dataset.
We use onehot encoding to group the data based on the most common venues visited by customers in different neighborhoods, and use KMeans clustering to divide the data in five clusters

We get the co-ordinates for Toronto using 'geopy' and plot a map using 'Folium'. The map shows all the clusters in Toronto marked in different colors.

(C)Foursquare:
Location data from foursquare is used to get the geographic coordinates and data regarding the different venues visited by people in Toronto and Manhattan. This data is further used for clustering the database and plotting maps. It is also used to arrive at the final conclusion for the analysis using data visualisation.

3. Methodology:

(A) Understanding the objective:

Our main objective is to be able to find the most visited venues in Manhattan and Toronto in order to find the activities/areas of business that are most favoured by the customers and being able to recommend the correct business sector in Manhattan and Toronto to ABC&Co.

(B) Analytical Approach:

We start by analysing the data for New York. However, the dataset for New York has around 10,040 neighborhoods, which is much larger as compared to Toronto dataset. Hence, we have sliced the New York dataset and further only analysed the data for Manhattan, which contains aroun3,183 neighborhoods, and is comparable to the Toronto dataset, which has 1,602 neighborhoods.

(C) Exploratory Data Analysis:

New York:

- We load and explore the data from newyork_data.json file.
- The data is transformed into a pandas dataframe using python libraries. The data is looped to fill the dataframe.
- This dataframe contains geographical co-ordinates of neighborhoods in New York. This data will form the basis for getting the details of venues in New York using Foursquare.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

- Geopy and Folium libraries are used to get the geographic co-ordinates and create a map of New York with all the neighborhoods superimposed on top.

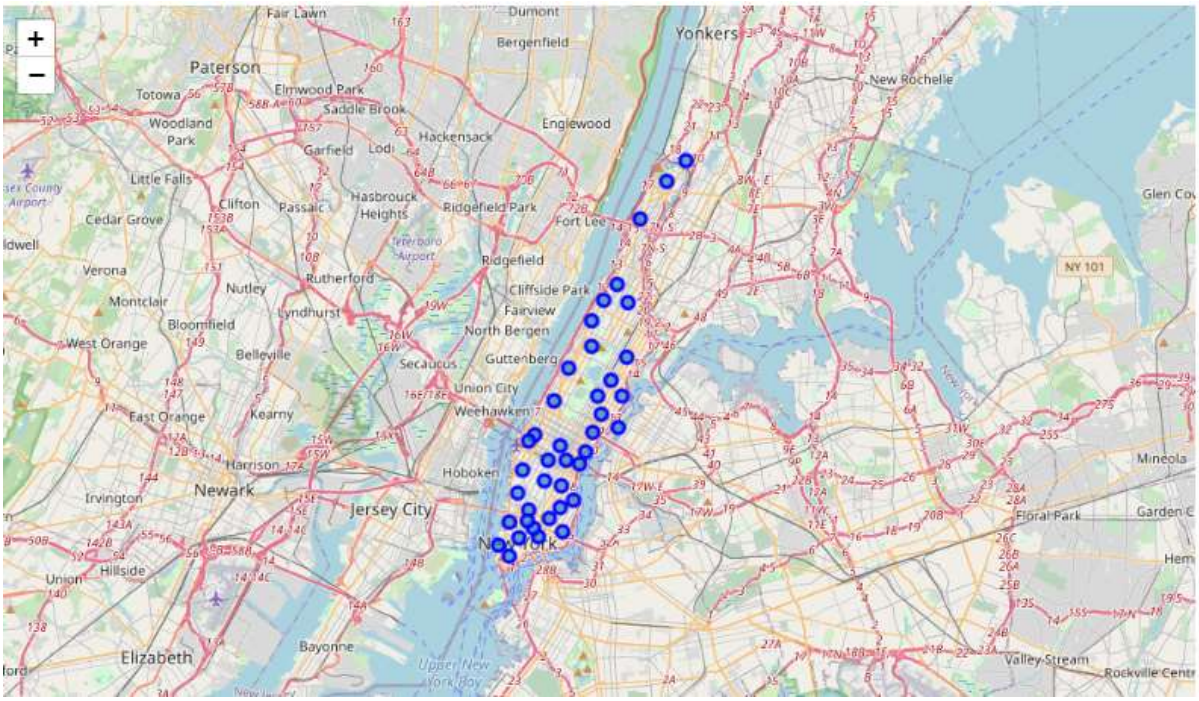
- Foursquare API is used to get details of venues in New York City. The New York dataset is much larger than the Toronto dataset, so we have decided to slice the data for New York and continue with the analysis of only the Manhattan dataset, which is comparable to Toronto dataset in terms of size.

Manhattan:

- The New York dataset is sliced to select only the data that is pertaining to the borough Manhattan.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

- We get the geographical co-ordinates for Manhattan using geopy and create a map using Folium with all the neighborhoods superimposed on top.



- We get the details of venues in Manhattan using Foursquare API. This dataset contains details of all the venues in various neighborhoods and their geographic co-ordinates.

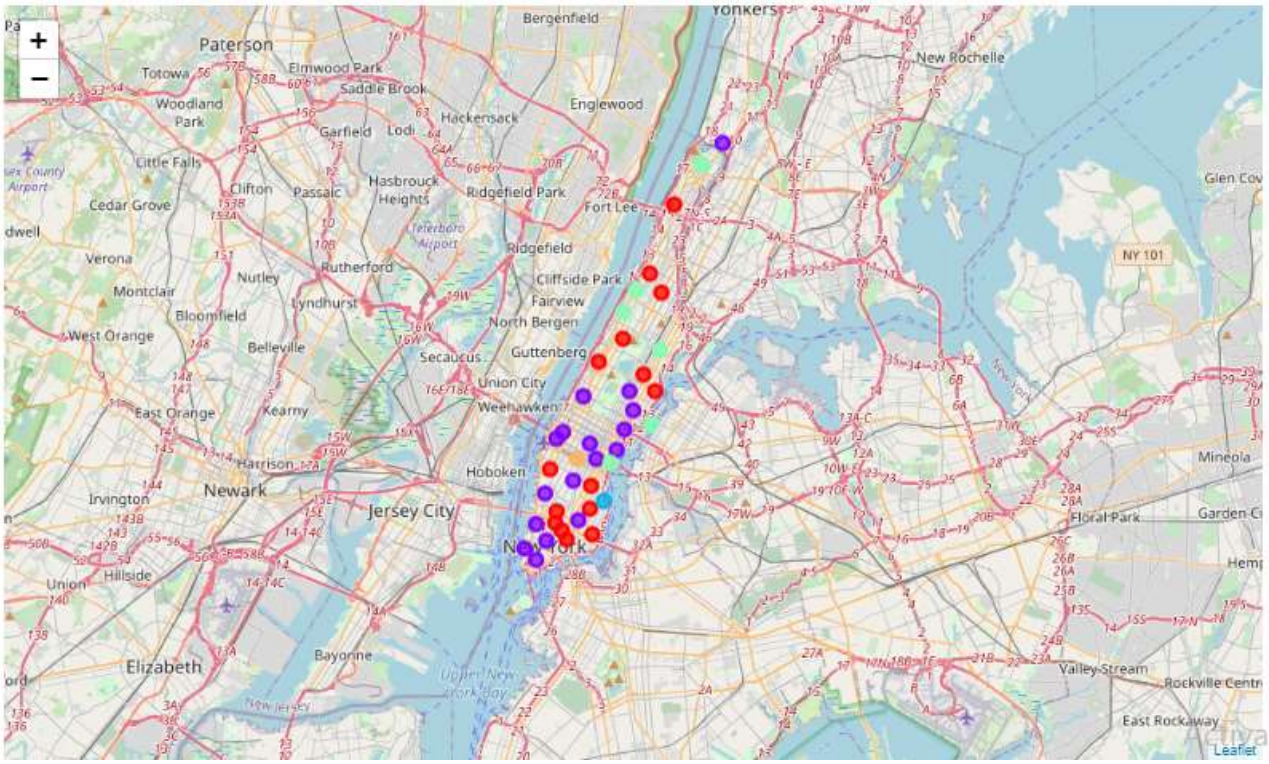
- We use onehot encoding to analyse each neighbourhood. The data is transformed into a dataset based on neighborhoods and venues.

- The data is grouped based on the neighborhoods and by taking the mean of the frequency of occurance of each category.
- The data is group the data based on the most common venues visited by customers in different neighborhoods and transformed into a pandas dataframe. The dataframe contains details of 10 most common venues for each of the neighborhoods.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Coffee Shop	Clothing Store	Hotel	Park	Gym	Memorial Site	Gourmet Shop	Burger Joint	Beer Garden	Food Court
1	Carnegie Hill	Coffee Shop	Cosmetics Shop	Café	Yoga Studio	Bookstore	Gym	French Restaurant	Pizza Place	Wine Shop	Bakery
2	Central Harlem	African Restaurant	French Restaurant	Chinese Restaurant	American Restaurant	Cosmetics Shop	Seafood Restaurant	Bar	Southern / Soul Food Restaurant	Gym	Ethiopian Restaurant
3	Chelsea	Coffee Shop	Bakery	American Restaurant	Art Gallery	French Restaurant	Seafood Restaurant	Italian Restaurant	Ice Cream Shop	Hotel	Park
4	Chinatown	Bakery	Chinese Restaurant	Hotpot Restaurant	Cocktail Bar	Optical Shop	Spa	Dessert Shop	American Restaurant	Salon / Barbershop	Shanghai Restaurant

- Next, we use KMeans clustering to divide the data in five clusters. A dataframe is created containing details of all the neighborhoods, clusters and the top 10 venues for each neighbourhood.

- The cluster data is visualised using Folium. Each cluster is marked in a different colour on the map.



- Each of the cluster is examined. Here's an extract of the first cluster. Please note that only a part of the first cluster is presented here, as the entire dataset for cluster 1 is much bigger.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Chinatown	Bakery	Chinese Restaurant	Hotpot Restaurant	Cocktail Bar	Optical Shop	Spa	Dessert Shop	American Restaurant	Salon / Barbershop	Shanghai Restaurant
2	Washington Heights	Café	Bakery	Mobile Phone Shop	Bank	Park	New American Restaurant	Coffee Shop	Grocery Store	Gym	Deli / Bodega
4	Hamilton Heights	Pizza Place	Café	Coffee Shop	Mexican Restaurant	Deli / Bodega	Yoga Studio	Caribbean Restaurant	Chinese Restaurant	School	Sandwich Place
6	Central Harlem	African Restaurant	French Restaurant	Chinese Restaurant	American Restaurant	Cosmetics Shop	Seafood Restaurant	Bar	Southern / Soul Food Restaurant	Gym	Ethiopian Restaurant
9	Yorkville	Italian Restaurant	Gym	Coffee Shop	Bar	Sushi Restaurant	Deli / Bodega	Mexican Restaurant	Wine Shop	Diner	Japanese Restaurant

Toronto:

- Unlike the New York data, the Toronto data is not readily available in the internet. So, we use the data from the following Wikipedia page: "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

- We use BeautifulSoup to get the data from the Wikipedia page and transform it into a dataframe containing 3 columns. The data is transformed into a pandas dataframe using python libraries and the data is looped to fill the dataframe. The data is cleaned to remove any blanks.

- To get the co-ordinates for each neighbourhood, we use geocoder to get the latitudes and longitudes for each postal code. The two tables are then merged to create one dataset containing details of the neighbourhoods and their geographical co-ordinates.

- This dataset contains details for all boroughs, so we slice the dataset to extract data for Toronto only.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
42	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

- This data will form the basis for getting the details of venues in Toronto using Foursquare API.

- We get the details of venues in Toronto using Foursquare API. This dataset contains details of all the venues in various neighborhoods and their geographic co-ordinates.

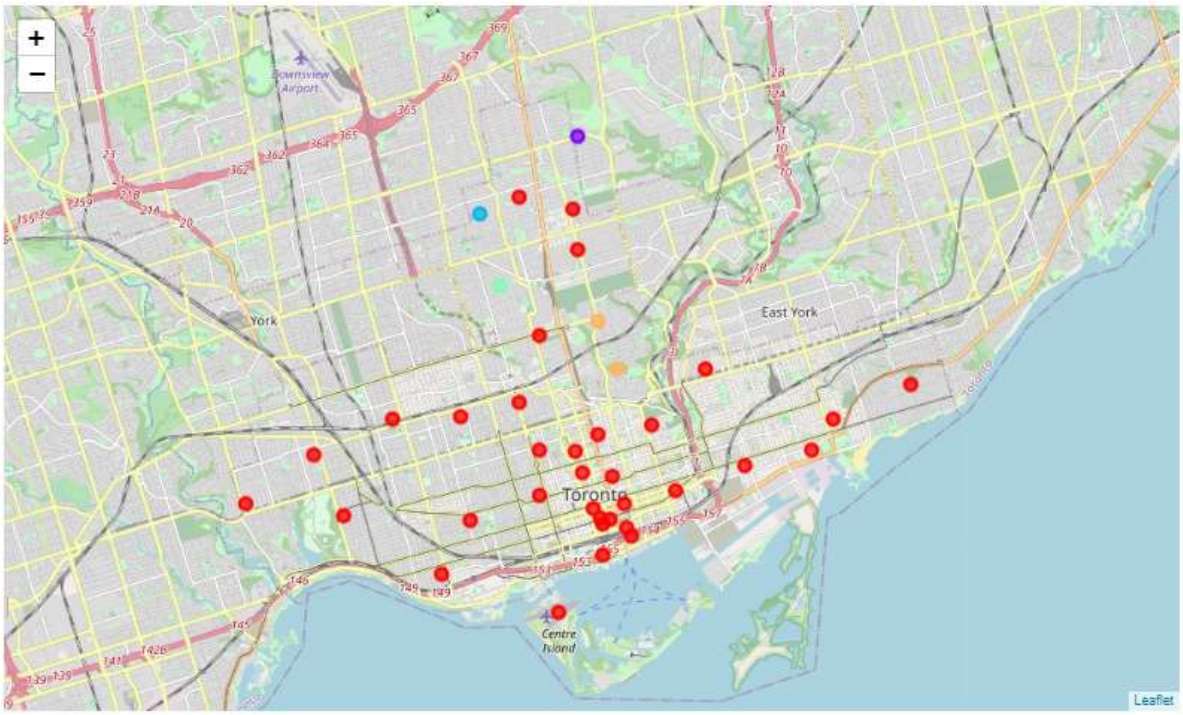
- We use onehot encoding to analyse each neighbourhood. The data is transformed into a dataset based on neighborhoods and venues.

- The data is grouped based on the neighborhoods and by taking the mean of the frequency of occurrence of each category.
- The data is group the data based on the most common venues visited by customers in different neighborhoods and transformed into a pandas dataframe. The dataframe contains details of 10 most common venues for each of the neighborhoods

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Bakery	Cocktail Bar	Farmers Market	Seafood Restaurant	Restaurant	Pharmacy	Cheese Shop	Beer Bar	Café
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Nightclub	Coffee Shop	Yoga Studio	Performing Arts Venue	Burrito Place	Restaurant	Climbing Gym	Convenience Store
2	Business reply mail Processing Centre, South C...	Yoga Studio	Smoke Shop	Auto Workshop	Brewery	Burrito Place	Butcher	Comic Shop	Farmers Market	Fast Food Restaurant	Garden
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Airport	Bar	Coffee Shop	Rental Car Location	Sculpture Garden	Boat or Ferry	Boutique
4	Central Bay Street	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Thai Restaurant	Japanese Restaurant	Burger Joint	Bubble Tea Shop	Salad Place	Portuguese Restaurant

- Next, we use KMeans clustering to divide the data in five clusters. A dataframe is created containing details of all the neighborhoods, clusters and the top 10 venues for each neighbourhood.

- The cluster data is visualised using Folium. Each cluster is marked in a different colour on the map.



- Each of the cluster is examined. Here's an extract of the first cluster. Please note that only a part of the first cluster is presented here, as the entire dataset for cluster 1 is much bigger.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
37	East Toronto	0	Asian Restaurant	Health Food Store	Trail	Pub	Yoga Studio	Dumpling Restaurant	Dog Run	Doner Restaurant	Donut Shop	Electronics Store
41	East Toronto	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Furniture / Home Store	Bookstore	Indian Restaurant	Spa	Pub	Japanese Restaurant
42	East Toronto	0	Sandwich Place	Fast Food Restaurant	Sushi Restaurant	Pub	Brewery	Board Shop	Restaurant	Italian Restaurant	Fish & Chips Shop	Steakhouse
43	East Toronto	0	Coffee Shop	American Restaurant	Bakery	Brewery	Café	Gastropub	Yoga Studio	Diner	Park	Middle Eastern Restaurant
45	Central Toronto	0	Gym / Fitness Center	Hotel	Pizza Place	Department Store	Sandwich Place	Breakfast Spot	Food & Drink Shop	Park	Gastropub	Gift Shop

(D) Analysing Manhattan and Toronto based on venues:

- Now that we have details of all the neighborhoods and the most common venues in Manhattan and Toronto, we can proceed with analysing both the geographical regions to find out the most favoured activity by the target audience in these regions.

- For this, we have used pie charts * as the visualisation tool.

- Using the base data for neighborhoods and venues that was used earlier, we create a pandas dataframe containing details of the most common venues in each neighbourhood in Manhattan and Toronto.

** Note: the preferred mode for visualising and presenting the results was wordcloud, but as there was some error while downloading the package, we had to use pie chart.*

Manhattan

	Neighborhood	Most Common Venue
0	Battery Park City	Coffee Shop
1	Carnegie Hill	Coffee Shop
2	Central Harlem	African Restaurant
3	Chelsea	Coffee Shop
4	Chinatown	Bakery

Toronto

	Neighborhood	Most Common Venue
0	Berczy Park	Coffee Shop
1	Brockton, Parkdale Village, Exhibition Place	Café
2	Business reply mail Processing Centre, South C...	Yoga Studio
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge
4	Central Bay Street	Coffee Shop

- Next, we proceed by taking a count of the neighborhoods for each of the ‘Most Common Venues’ for Manhattan and Toronto. This would help in creating pie charts.

Manhattan

Most Common Venue	Neighborhood
African Restaurant	1
American Restaurant	1
Bakery	2
Bar	2
Boat or Ferry	1

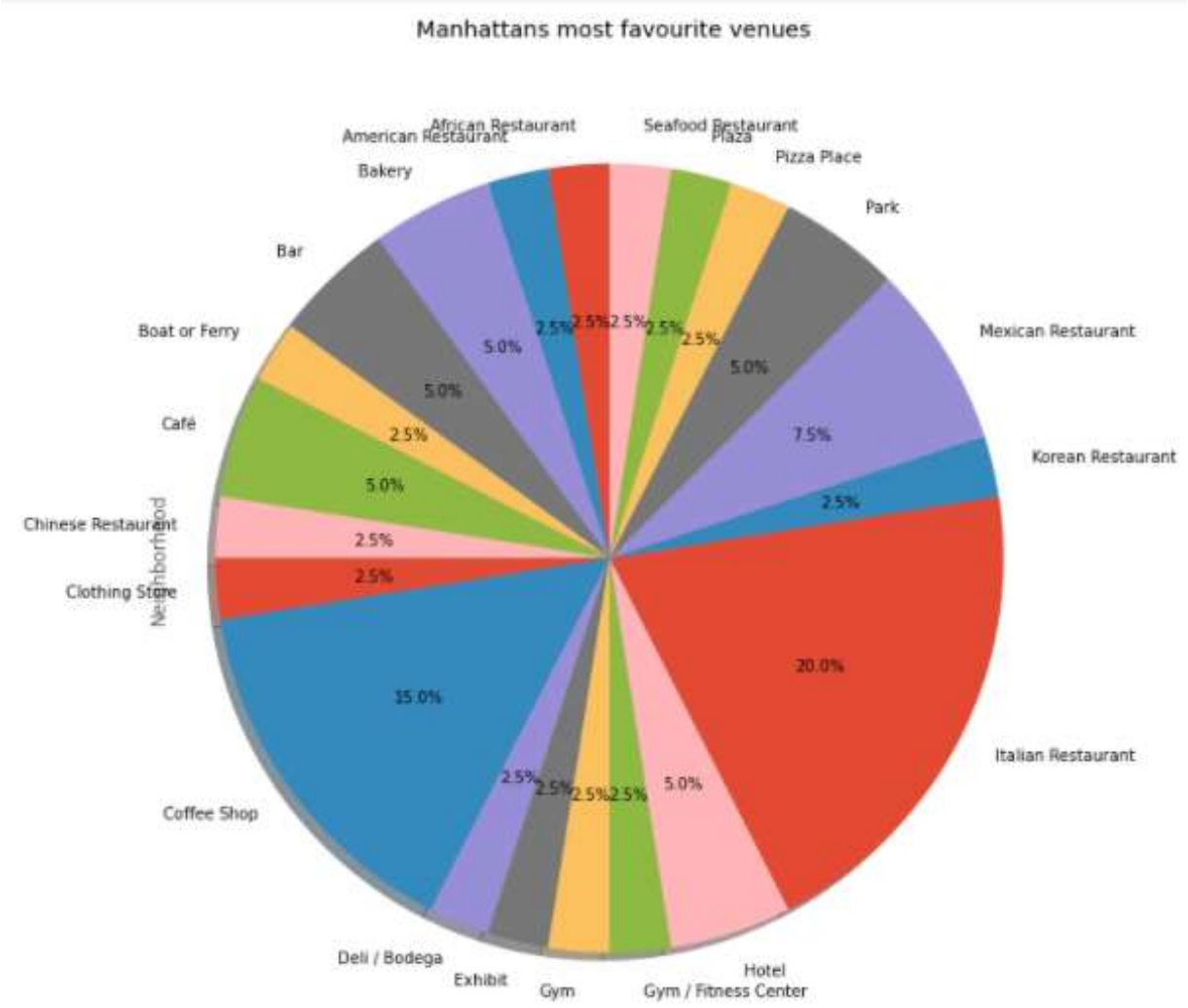
Toronto

Most Common Venue	Neighborhood
Airport Lounge	1
Asian Restaurant	1
Bakery	1
Bar	1
Breakfast Spot	1

- This data forms the basis for pie charts. We use matplotlib and ggplot for creating the pie charts.

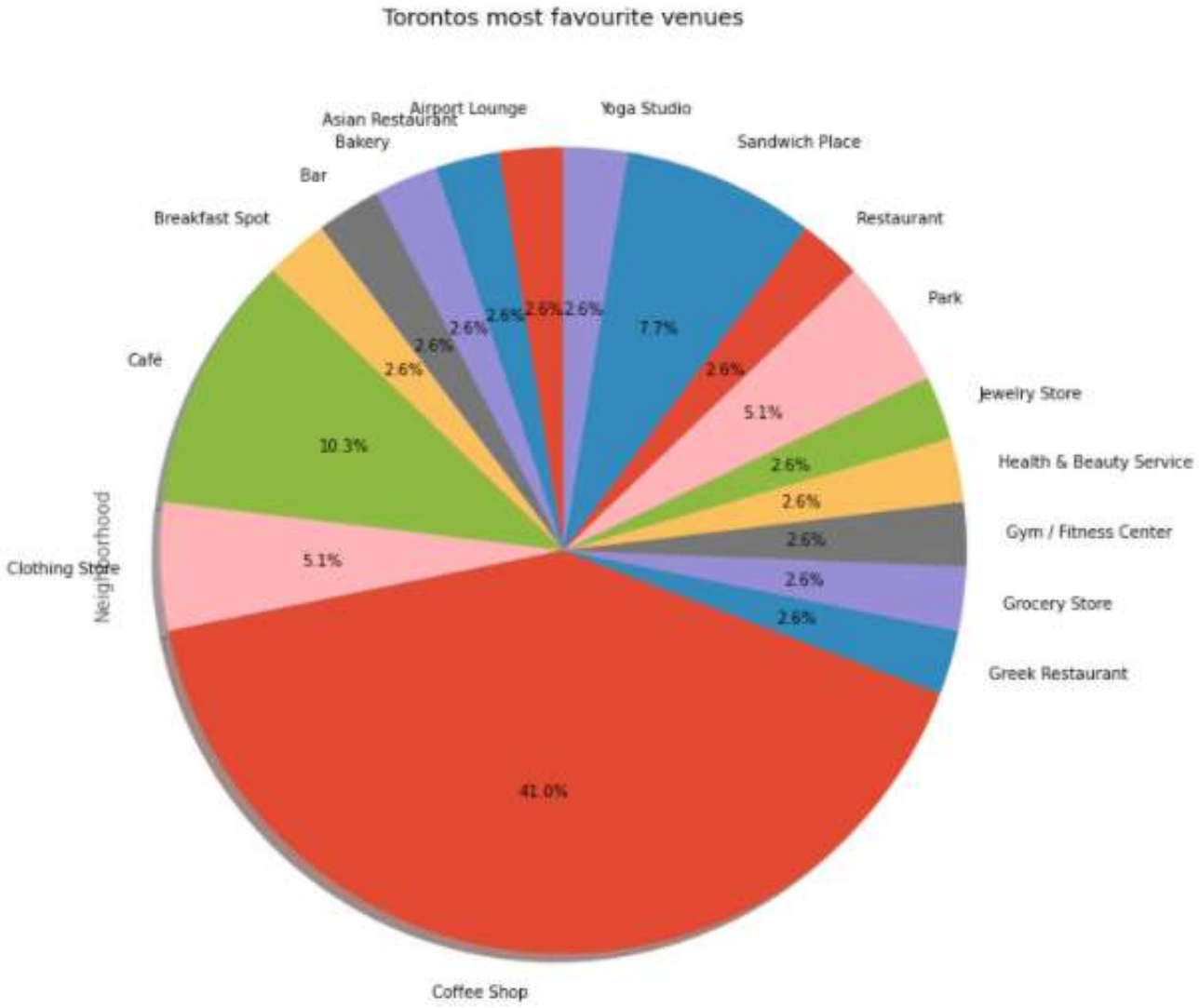
- The visualised data is as follows:

Manhattan



-

Toronto


Torontos most favourite venues
Airport Lounge
Yoga Studio
Asian Restaurant
Bakery
Sandwich Place
Bar
Breakfast Spot
Restaurant
Park
Café
Jewelry Store
Health & Beauty Service
Gym / Fitness Center
Clothing Store
Grocery Store
Greek Restaurant
Coffee Shop
Neighborhood
2.6%
2.6%
2.6%
2.6%
2.6%
2.6%
7.7%
2.6%
5.1%
2.6%
2.6%
2.6%
2.6%
2.6%
10.3%
5.1%
41.0%

4. Results:

Based on the analysis, we can see that both Manhattan and Toronto are similar in terms of being multi-cultural and cosmopolitan cities in the world. They are both major centre for various businesses and provide the businesses with a large customer base. This automatically attracts new businesses, increasing the competition in both the geographical areas.

Considering these factors, an aspiring entrepreneur needs to carefully analyse both the geographical regions before deciding on the type of business venture. This can be found out based on the information of the most commonly visited venues by the target customers in both the regions. The type of venue that has the most customer footfall could be considered as an ideal type of business for an aspiring entrepreneur, in order to ensure profit.

The analysis done by us in this project attempts to do the same. Based on the details of neighborhoods and venues in these neighborhoods, we have found out the details of the most and least commonly visited venues on Manhattan and Toronto.

Manhattan: Based on the pie-chart, we see that the most commonly visited venue in Manhattan is 'Italian Restaurants', which accounts for 20% of the total customer footfall. This is closely followed by 'Coffee Shops' with a 15% market share. 'Mexican Restaurants' are next with around 7.5% share. These three venues put together, account for about 42.5% of the total customer footfall. Following these three, are Parks, Bakeries, Bars, Cafes and Hotels, each with a share of 5%.

Toronto: Based on the pie-chart, we can see that the most commonly visited venue is 'Coffee Shops' with a clear majority of 41% of the total customer footfall. Following them, are 'Cafes' with a share of 10.3%, which is only about one-third of that for Coffee Shops. Cafes and Coffee Shops taken together account for about 51.3%, which is a little more than half of the total customer footfall. Following these are Sandwich shops (7.3%), Clothing stores and Parks (5% each).

5. Discussion:

The analysis is based on various sources of information and assumptions:

- The venue information extracted using Foursquare API, which may not account for all the customers in the geographical regions.
- Manhattan region from New York is used for the purpose of analysis and to make it comparable to Toronto data. However, the Manhattan data may not be an exact representation for the entire New York dataset. Customer preferences may differ depending on the geographical regions within New York.
- High customer footfall has been assumed to be the representative of high sales and high profit, however, it may not always be so.
- Information regarding the profit margins and initial capital expenditure for each of the businesses/venues analysed could further help in deciding the most profitable business.

6. Conclusion:

Based on the results and assumptions (as mentioned in discussion section), we can see that Manhattan and Toronto are very different in terms of activities that are most favoured by the customers. Hence, establishing the same type of business in both the regions may not yield the same level of returns.

In Manhattan, an aspiring entrepreneur could expect a favourable response from the target customers if they were to open an Italian restaurant or a Coffee Shop or a Mexican restaurant.

In Toronto, on the other hand, they could expect a favourable response from the target customers if they were to open a Coffee Shop or a Café.

It is important to note that the analysis is based on limited data and market segment information. Further information about customer preferences, profit margins and capital investments in various businesses in both the geographical regions could help refine the results.