



# IBM Data Science Professional Certificate – Capstone Project: Battle of Neighborhoods

Exploring neighborhoods of Chicago and comparing them with  
neighborhoods of New York and Toronto.

---

Capstone Project Report

Chinmay Govilkar

June 2019

## Table of contents:

- Introduction/business problem
- Data gathering
- Data science methodology
- Analysis and results
- Discussions/observations
- Conclusion

Introduction: I am going to explore the city of Chicago on basis of tourist outlooks, food habits, recreation and sports and cultural outlook, so that either the residents or tourists have readily available access to city's neighbourhood exploration, and the observations would be substantiated with sufficient data. Lastly I will compare this exploration data with similar data of neighbourhoods of two other cities namely New York and Toronto, to cite the similarities as well as the differences between the people's culture of surrounding cities as well as countries.

Background: Chicago is a city in the state of Illinois in United States and is the third most populated city in the US with approximate population of 2.7 million. Chicago is widely known for its food, architecture and loyal sports fan, it also houses a few of the most reputed college and high school institutions in the world. So I am going to explore the neighbourhoods of the city on the basis of following pointers:

- Cuisines — Restaurants, bars, stalls and cafes
- Sports — Stadiums, arenas and clubs
- Tourist attractions — Architectural buildings and museums
- Miscellaneous — Pubs, Shops, Street food carts etc.

Data Gathering: The required data is gathered from the following sources:

- ▶ Chicago dataset : [https://en.wikipedia.org/wiki/List\\_of\\_neighborhoods\\_in\\_Chicago](https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago)
- ▶ Toronto dataset : [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- ▶ New York dataset : [https://geo.nyu.edu/catalog/nyu\\_2451\\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572)

Location data: Fetched from **Geocoder** package.

Venue data: Fetched from **Foursquare API** : <https://foursquare.com/developers/apps>

- The 3 links referencing to three datasets for New York, Toronto and Chicago consist of borough and neighbourhood information in the respective cities.
- The New York dataset also includes the list of latitudes and longitudes along with the borough and neighbourhood information.
- The data for Toronto and Chicago however needs to be scraped from the Wikipedia link references using the **Beautiful Soup** for Python package.
- We would fetch the geo-spatial data co-ordinates using the **Geo-coder** for Python package for Toronto and Chicago cities to determine the location and the proximity to each of the desired locations in a specified neighbourhood.
- Finally we would fetch the venue details for each neighbourhood using the **Foursquare API**, by specifically querying the API database regarding the information we need to fetch.
- With the help of foursquare API we will find the following **Venue** related information such as :
  - ◆ Search for a venue
  - ◆ Fetch details of a venue
  - ◆ Fetch category of a venue
  - ◆ Fetch opening and closing hours for a venue
  - ◆ Fetch venue recommendations and trending venues
- Each observation will be plotted as a map using the **Folium** for Python library.

#### Data science methodology – life cycle:

As already mentioned above, we have discussed the business problem and the analytic approach phases of the life cycle, we then also discussed the data requirements, data collection and the data understanding phases. Now we will dive deep into data preparation.

**Data cleaning:** The New York dataset was readily available, hence needed no further attention with respect to cleaning or dropping the ‘NA’s’, the information present was structured and no manipulation was required for adjusting a few values. The geo-spatial information too was provided and it was appended to our data frame from a geo-json file.

However Toronto dataset was not readily available, hence it needed to be scraped from the wiki-link mentioned in data gathering section. I decided to use the Beautiful Soup library for Python to scrape the data. I installed the required packages and instantiated an object of request class in order to store the data. The class will allow us to read the HTML page, I then used the HTML tags, to filter out the data required for my analysis. Fortunately the required data was grouped under <table> tag and in <tr> rows, hence made it easier to sort.

Once the data was sorted, the “borough”, “postal code” and “neighbourhood” information was fetched and later converted and stored in a pandas data frame. Unlike the NY dataset, this did have a few “Not Assigned” values which were treated by providing the borough name itself.

Apart from this, it needed no further manipulation or cleaning of data. Here too the geo-spatial information was provided in a csv file which we used and appended it to our data frame having all the required information for exploration and analysis.

Finally the Chicago dataset just like our Toronto dataset, needed to be scraped from wiki-link, which I did as mentioned in the above paragraph, just that this dataset did only include 2 columns, i.e. “neighbourhood” and the “community area”. We later stored the scraped information again in a pandas data frame.

Unlike the other two, no geo-spatial information was provided here and also with lack of borough and postal code information, it became unlikely to group the neighbourhoods. Hence to tackle this problem and add an extra element to this capstone project, I researched the best 15 neighbourhoods from various references and links based on various outlooks, including area, population, density, median income, median rent, males to females ratio, median age and many tourist point of views which would form a base of our analysis.

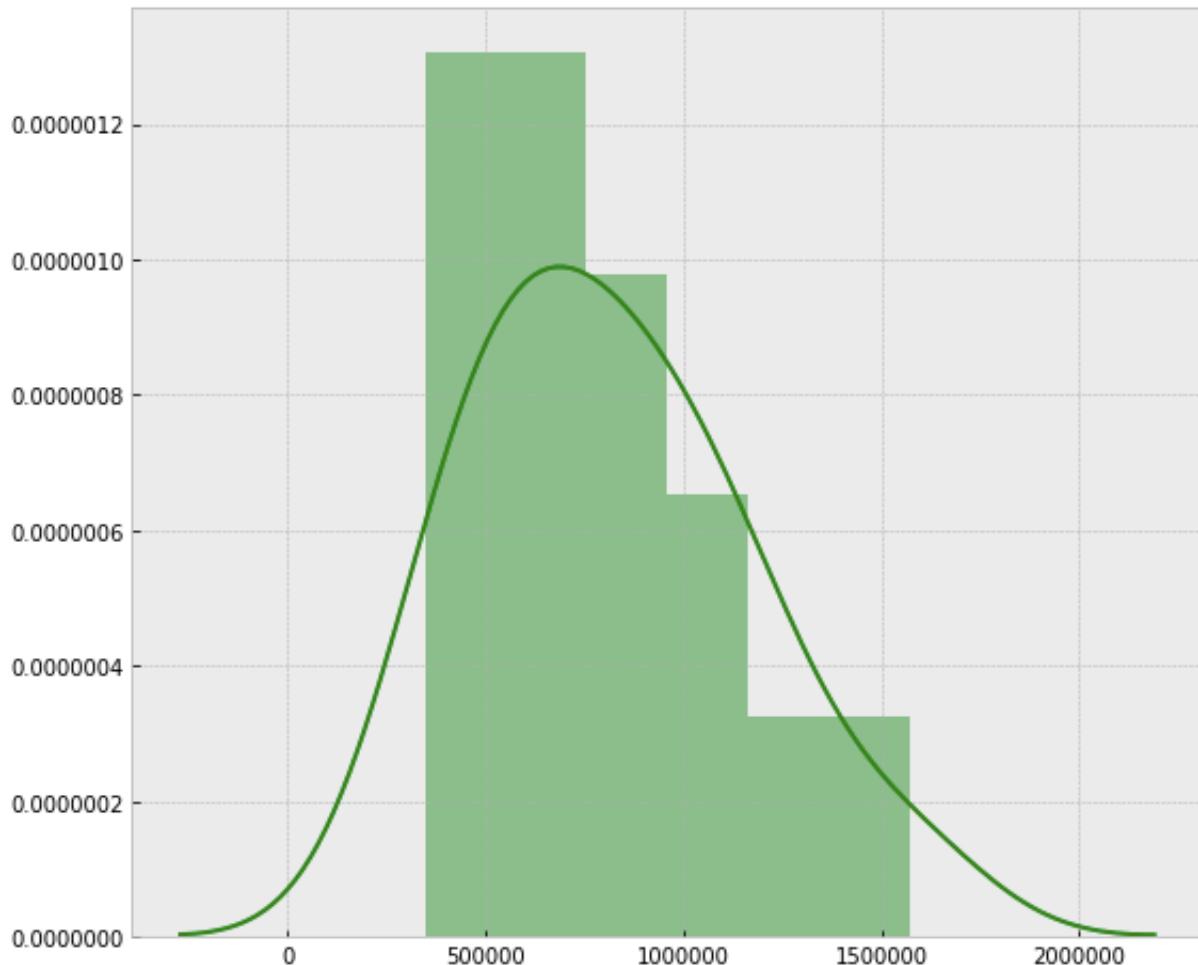
I then used the geo-encoder and Nominatim packages to fetch the latitudes and longitudes of these 15 neighbourhoods and append them to our data frame which was created along with the neighbourhoods and the geo-spatial information.

In this way we finally had our 3 datasets complete, cleaned, manipulated and ready to explored and “neighbourhood”, “latitude” and “longitude” as our key features.

I will now start my analysis by first exploring the Chicago data alone with some descriptive and inferential statistics to begin with.

## Descriptive statistics:

First let us observe the range of the housing prices across the neighbourhoods from the distribution plot.



As we can see that the housing prices have been distributed from approx. 400,000 USD to roughly 15,50,000 USD.

We also infer the distribution of these prices i.e. maximum number of neighbourhoods have the prices relatively on the lower side, making them suitable for people who earn the wages as a lower middle class category.

Then we have moderate number of neighbourhoods have the houses intermediately priced which an average or an above average household income family could afford. We also notice the drop is gradual, meaning that these neighbourhoods do not show signs of extremity between the earnings.

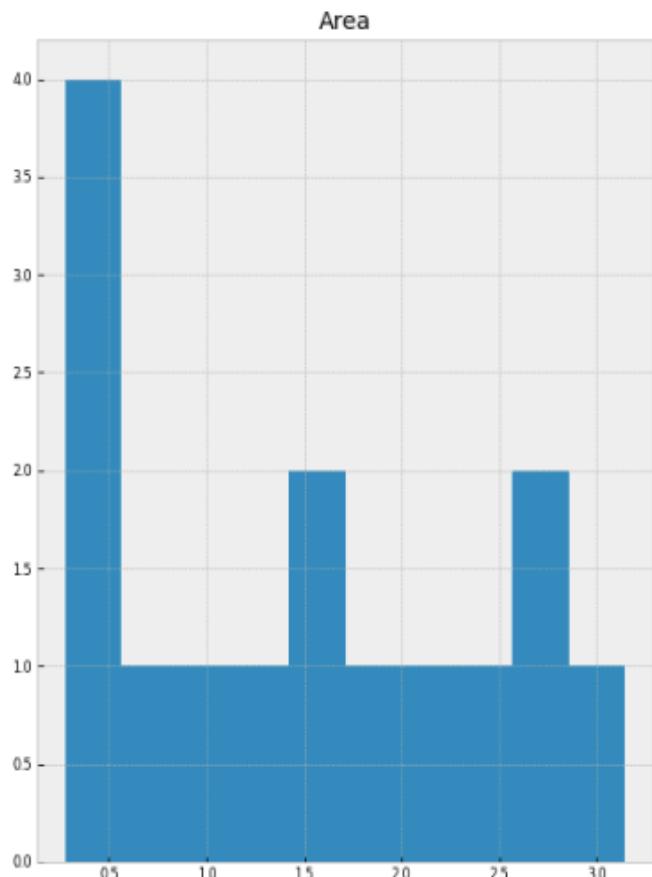
Finally we have the upper class category, where the number of neighbourhoods have significantly decreased compared to the other two categories. Meaning the number of neighbourhoods may be few, but the entirety of the neighbourhood would comprise of people with very high standard of living.

Now let us look at a scatter plot which will give us a pretty straight forward observation of relation between the income and the housing prices.



It is very evident that income and the housing prices are very closely correlated, i.e. the vertical movement on the Y axis leads to a drastic horizontal movement on the X axis.

Probably as observed from the distribution plot, the neighbourhoods with high standard of living, having high housing prices will also have high household income.



As we can infer from the histogram which shows the area of the neighborhoods. We can see that most of the neighborhoods are small, with area within a few sq. miles. Then few numbers of neighborhood are relatively small with area ranging from 1 to 2 sq. miles. Finally, the larger neighborhoods, which are not a lot but certainly depicts the picture of mix between the areas of the selected neighborhoods, which gives us the profile of each of smaller, medium in size and larger neighborhoods.

## Exploration analysis:

Now we have our 3 datasets ready with neighbourhood and geo-spatial co-ordinates information, we will now plot the maps using “folium”. We will need to provide the geo-spatial co-ordinates, once we plot the map, we will initiate an instance of the “Foursquare API” by providing your credentials, the instance will help us generate queries and query the database for fetching venue related information.

As observed from the map, the markers are spread out, hence I grouped them based on how each neighbourhood is near from other and formed 5 groups which I will now explore one by one, before we get to exploring, here are the 5 groups and the respective neighbourhoods:

- Cluster 1 - Gold Coast, Old Town, Near North Side, River North, West Loop
- Cluster 2 - Edgewater, Andersonville
- Cluster 3 - Lakeview, Lincoln Park
- Cluster 4 - Logan Square, Wicker Park
- Cluster 5 - Edison Park
- Cluster 6 - Bridgeport, Near South Side
- Cluster 7 - Hyde Park

Now let us explore each of the 7 clusters and fetch the venues and the information related to each venue.

Before we explore the neighbourhoods, please note that the exploration is carried out within a 1 or 2 kilometre radius with a maximum of 100 search results, therefore a few venue categories might not be listed due to the geographical limitations and their distances from our neighbourhoods.

Also for convenience the outputs obtained in python, were saved to an excel file and the venue categories were updated to a limited number of listings.

For cluster 1, we calculated the centre of the 5 neighbourhoods to be Near North Side, hence the exploration is carried out within a 2 kilometre radius.

For cluster 2, we again calculated the centre of the 2 neighbourhoods to be Andersonville, hence the exploration is carried out within a 1 kilometre radius.

Here are the results of both of the cluster groupings, each of the listing has the corresponding count, which denotes the occurrence of the venues.

Gold Coast, Old Town, Near North Side,  
River North and West Loop

Category Listing	Venue Count
Architecture	1
Clothing	1
Coffee and Snacks	20
Drinks	8
Entertainment	9
Fitness	7
Health	1
Hotel	10
Parks and Garden	2
Personal Care	5
Recreation	2
Restaurant	29
Shops and Supermarket	4
Stationary	1
<b>Total</b>	<b>100</b>

Edgewater and Andersonville

Category Listing	Venue Count
Antiques	5
Architecture	1
Bookstore	2
Coffee and Snacks	17
Drinks	15
Entertainment	3
Fitness	2
Health	2
Hotel	1
Parks and Garden	2
Personal Care	5
Restaurant	34
Shops and Supermarket	11
<b>Total</b>	<b>100</b>

### Observations:

As we can infer from the above venue information table, I have further classified the foursquare API venue categories into discrete sub-categories in order to publish as much information possible in limited entries.

Both neighbourhood groups show significant number of restaurants, coffee shops and light snacks or fast food joints compared to other categories. This is mainly because these neighbourhoods have higher number of food joints clustered in the areas which I have explored.

As previously mentioned, we also can see some shops categorized as “antiques and architecture” as Chicago is known for its unique and beautiful architecture rich structures. These neighbourhoods also comprise of number of bars and pubs depicting the rich night life culture, the tourists and residents can experience.

The neighbourhood also is in abundance of shops, supermarkets and hotels, providing the visitors with all the amenities within reach from respective destinations.

The neighbourhood also hosts parks, playgrounds and gardens along with fitness studios which residents and tourists can make use of to abide by their fitness regime.

For cluster 3, we calculated the centre of the 2 neighbourhoods to be Lincoln Park, hence the exploration is carried out within a 1 kilometre radius.

For cluster 4, we again calculated the centre of the 2 neighbourhoods to be Wicker Park, hence the exploration is carried out within a 1 kilometre radius.

Here are the results of both of the cluster groupings, each of the listing has the corresponding count, which denotes the occurrence of the venues:

Lakeview and Lincoln Park

Category Listing	Venue Count
Architecture	2
Clothing	3
Coffee and Snacks	30
Drinks	4
Entertainment	5
Fitness	6
Hotel	1
Parks and Garden	2
Personal Care	5
Restaurant	26
Shops and Supermarket	15
Stationery	1
<b>Total</b>	<b>100</b>

Logan Square and Wicker Park

Category Listing	Venue Count
Antiques	4
Clothing	4
Coffee and Snacks	10
Drinks	15
Education	2
Fitness	3
Health	1
Parks and Garden	2
Personal Care	4
Restaurant	35
Shops and Supermarket	20
<b>Total</b>	<b>100</b>

Both Lakeview and Lincoln Park are to be considered one of the most sophisticated neighbourhoods in the city, hence here the grouping is slightly tilted towards posh cafés and fine dine restaurants. It is also evident that, the number of night life activities were reduced, citing to the fact that median age is above 49, and also the fine dine restaurants, serve drinks rather than being just deli's or burger joints.

Another contrast is seen in the Logan Square and Wicker Park neighbourhoods, where shops and supermarkets are high and so are the fast food joints and bars. These neighbourhoods have a younger population hence the shops and restaurants depict a vastly different profile than observed in Lincoln Park neighbourhood.

They also have educational institutes, which justifies our initial observation of student population, a fact not yet observed in any of the neighbourhoods.

For clusters 5, 6 and 7 we calculated the centre of neighbourhoods as Edison Park, Hyde Park and Near South Side. Hence the exploration is carried out within a 2 kilometre radius.

#### Edison Park

Category Listing	Venue Count
Coffee and Snacks	3
Drinks	6
Entertainment	1
Parks and Garden	3
Restaurant	15
Shops and Supermarket	1
<b>Total</b>	<b>29</b>

#### Hyde Park

Category Listing	Venue Count
Coffee and Snacks	4
Drinks	2
Fitness	1
Health	1
Parks and Garden	1
Restaurant	11
Shops and Supermarket	8
Transport	4
<b>Total</b>	<b>32</b>

#### Bridgeport and Near South Side

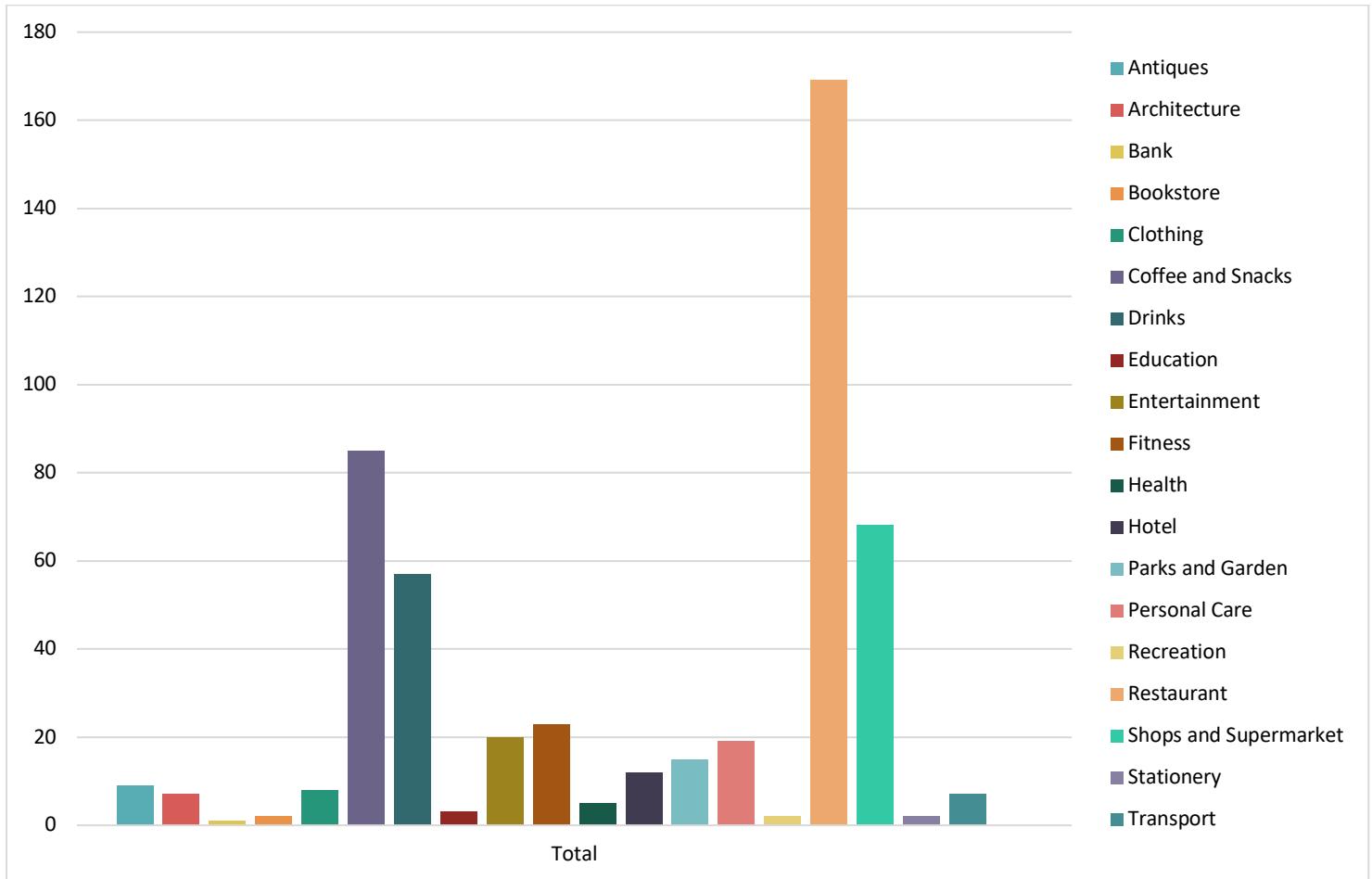
Category Listing	Venue Count
Architecture	3
Bank	1
Coffee and Snacks	1
Drinks	8
Education	1
Entertainment	2
Fitness	4
Parks and Garden	3
Restaurant	19
Shops and Supermarket	9
Transport	3
<b>Total</b>	<b>54</b>

These neighbourhoods are located bit remote compared to the others and hence the concentration of the shops is very sparse and despite explored in a 2 km radius, we hardly have 110+ venues combined for 3 neighbourhoods.

One of the most defining observations here is the inclusion of transport services, which is mainly because of the locations of these neighbourhoods, which is a bit far from the city centre or downtown. Especially in Bridgeport and Near South Side, there are many architecture rich heritage sites which interest a lot of tourists.

The basic necessities do remain similar in comparison with other neighbourhoods, with most of them being fast food restaurants, few bars, few fitness studios and entertainment activities.

Here is the overview of all the venues and their categories which we just explored:



### Venue counts according to Neighbourhoods

```
chicago_venues.groupby('Neighbourhood').count()
```

	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
<b>Neighbourhood</b>						
<b>Andersonville</b>	100	100	100	100	100	100
<b>Bridgeport</b>	81	81	81	81	81	81
<b>Edgewater</b>	100	100	100	100	100	100
<b>Edison Park</b>	36	36	36	36	36	36
<b>Gold Coast</b>	100	100	100	100	100	100
<b>Hyde Park</b>	100	100	100	100	100	100
<b>Lakeview</b>	100	100	100	100	100	100
<b>Lincoln Park</b>	100	100	100	100	100	100
<b>Logan Square</b>	100	100	100	100	100	100
<b>Near North Side</b>	100	100	100	100	100	100
<b>Near South Side</b>	100	100	100	100	100	100
<b>Old Town</b>	100	100	100	100	100	100
<b>River North</b>	100	100	100	100	100	100
<b>West Loop</b>	100	100	100	100	100	100
<b>Wicker Park</b>	100	100	100	100	100	100

Finally here is the list of the 10 most common venue categories for all 15 neighbourhoods of Chicago. Later I will compare the same stat with those of Toronto and New York.

Neighbourhood	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 Andersonville	Coffee Shop	Italian Restaurant	Breakfast Spot	Asian Restaurant	Grocery Store	Bar	Antique Shop	Sandwich Place	Middle Eastern Restaurant	Mexican Restaurant
1 Bridgeport	Chinese Restaurant	Mexican Restaurant	Art Gallery	Pizza Place	Hot Dog Joint	Bar	Sporting Goods Shop	Grocery Store	Wings Joint	Pharmacy
2 Edgewater	Coffee Shop	Sandwich Place	Grocery Store	Asian Restaurant	Mexican Restaurant	Italian Restaurant	Sushi Restaurant	Lounge	Breakfast Spot	Gym
3 Edison Park	Park	Italian Restaurant	Bar	Pizza Place	Donut Shop	Bakery	Bus Station	Spa	Shipping Store	Seafood Restaurant
4 Gold Coast	Hotel	Italian Restaurant	American Restaurant	Gym	Boutique	Coffee Shop	Women's Store	Salon / Barbershop	Café	Steakhouse
5 Hyde Park	Science Museum	Bookstore	Coffee Shop	Café	Park	Sandwich Place	Pizza Place	Spa	Burger Joint	Caribbean Restaurant
6 Lakeview	Sandwich Place	Japanese Restaurant	Spa	Breakfast Spot	Pizza Place	Theater	Café	Sports Bar	Vegetarian / Vegan Restaurant	Sushi Restaurant
7 Lincoln Park	Pizza Place	Coffee Shop	Italian Restaurant	Sushi Restaurant	Hot Dog Joint	Theater	Indonesian Restaurant	Sandwich Place	Salon / Barbershop	New American Restaurant
8 Logan Square	Coffee Shop	Cocktail Bar	Bar	Latin American Restaurant	Ice Cream Shop	Dive Bar	Grocery Store	Café	Pizza Place	Park
9 Near North Side	Hotel	American Restaurant	Steakhouse	Coffee Shop	Gym	Pizza Place	Optical Shop	Yoga Studio	Women's Store	Gym / Fitness Center
10 Near South Side	Chinese Restaurant	Coffee Shop	Bar	Bakery	Korean Restaurant	Pizza Place	Park	Salon / Barbershop	Asian Restaurant	Bank
11 Old Town	Italian Restaurant	Café	Coffee Shop	Gym	Gym / Fitness Center	Comedy Club	Restaurant	Grocery Store	American Restaurant	Cycle Studio
12 River North	Steakhouse	Italian Restaurant	Hotel	Mexican Restaurant	Restaurant	Coffee Shop	Gym	Bar	American Restaurant	New American Restaurant
13 West Loop	New American Restaurant	Coffee Shop	Italian Restaurant	Cocktail Bar	Brewery	Sandwich Place	Pizza Place	Bar	Mexican Restaurant	Bakery
14 Wicker Park	Pizza Place	Bar	Coffee Shop	Boutique	Taco Place	Sushi Restaurant	Bookstore	Breakfast Spot	Café	Cajun / Creole Restaurant

We have the following information ready to be published:

- Coffee Shop – Is the most common venue for most of the neighbourhoods including likes of Andersonville, Edgewater and Logan Square. However if not the most common in other neighbourhoods, it is still making the 10 most common list for almost all.
- Multi-cuisine restaurants – make the 2<sup>nd</sup> most common venue with again topping the list for 3 neighbourhoods, including Bridgeport, Near South Side and West Loop. Again a very popular choice for both tourists and residents, making the top 10 list for all.
- Fast food joints – is our 3<sup>rd</sup> most common venue for 4 neighbourhoods including Lakeview, Lincoln Park, River North and Wicker Park. Well known for people on the run with not much time on hand, with quick delivery and service and affordable too.
- Hotels – tops the most common venue for 2 neighbourhoods with likes of Near North Side and Gold Coast. Both of these neighbourhoods offer an elegant view of the coastline which makes an ideal spot for business travellers and tourists to book a hotel offering an outstanding view of the coastline.
- Gyms and Fitness studios – Also plenty makes the 10 most common venues on almost all neighbourhoods, areas which do not offer parks, gardens where people can go for a run choose to hit these gyms and yoga studios for their workout regime.
- Parks, shops, stores are other venues which do make the top 10 list, but on few occasions and for selected neighbourhoods.

Concluding our exploration analysis is the comparison between the neighbourhoods of Chicago with those of Toronto and New York, which pretty much share the same cultural and professional outlook as Chicago.

Here is the list of 10 most common venues for the neighbourhoods of Toronto, specifically of Downtown Toronto.

	<b>Neighbourhood</b>	<b>1st Most Common Venue</b>	<b>2nd Most Common Venue</b>	<b>3rd Most Common Venue</b>	<b>4th Most Common Venue</b>	<b>5th Most Common Venue</b>	<b>6th Most Common Venue</b>	<b>7th Most Common Venue</b>	<b>8th Most Common Venue</b>	<b>9th Most Common Venue</b>
0	Adelaide, King, Richmond	Coffee Shop	Café	Bar	Thai Restaurant	American Restaurant	Steakhouse	Hotel	Cosmetics Shop	Bakery
1	Berczy Park	Coffee Shop	Cocktail Bar	Café	Cheese Shop	Bakery	Beer Bar	Seafood Restaurant	Steakhouse	Italian Restaurant
2	CN Tower, Bathurst Quay, Island airport, Harbo...	Airport Service	Airport Lounge	Airport Terminal	Boat or Ferry	Boutique	Airport	Airport Food Court	Airport Gate	Sculpture Garden
3	Cabbagetown, St. James Town	Coffee Shop	Bakery	Pizza Place	Pub	Café	Italian Restaurant	Restaurant	Bank	Pharmacy
4	Central Bay Street	Coffee Shop	Café	Italian Restaurant	Burger Joint	Sandwich Place	Japanese Restaurant	Spa	Sushi Restaurant	Middle Eastern Restaurant

As with neighbourhoods of Chicago, the most common venue for Downtown Toronto is also coffee shop. Downtown being the heart of the city and hub for all of the professionals, on the go cafes and coffee shops make them the most common choice for residents and tourists.

Multi-cuisine restaurants – just like we observed in Chicago, the 10 most common places include a lot of multi-cuisine restaurants, with likes of Thai, American, Italian, Japanese, and Middle Eastern.

The neighbourhood, also host an airport due to which all of the venues are nearby airport vicinity. It includes a lounge, terminal, food court, sculpture garden and a boutique.

Fast food joints, pizza places, bakeries and pubs are among other common venues which have high concentration in all of the neighbourhoods of Downtown Toronto.

Among the odd ones, the Bathurst Quay holds a Boat or a Ferry bay where tourists and residents can experience an excursion evening or an outing.

Now once Toronto is analysed, let's take a look at the New York neighbourhoods, more specifically ones in Manhattan.

	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	Park	Coffee Shop	Hotel	Memorial Site	Gym	Wine Shop	Clothing Store	Italian Restaurant	Women's Store	BBQ Joint
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Yoga Studio	Bookstore	Wine Shop	Cosmetics Shop	French Restaurant	Bar	Japanese Restaurant
2	Central Harlem	African Restaurant	Public Art	Art Gallery	Seafood Restaurant	Chinese Restaurant	Gym / Fitness Center	French Restaurant	American Restaurant	Cosmetics Shop	Liquor Store
3	Chelsea	Coffee Shop	Ice Cream Shop	Italian Restaurant	Bakery	Nightclub	Theater	Seafood Restaurant	American Restaurant	Hotel	Art Gallery
4	Chinatown	Chinese Restaurant	American Restaurant	Cocktail Bar	Spa	Dumpling Restaurant	Vietnamese Restaurant	Bubble Tea Shop	Optical Shop	Salon / Barbershop	Ice Cream Shop

Once again to no surprise, we have coffee shops and restaurants as the two most common venue categories. These neighbourhoods include multi-cuisine restaurants with likes of an African, Chinese, American, Italian, Vietnamese, French and Japanese. Compared to Toronto and Chicago, the cuisines are more and so is the popularity. The other restaurants and joints include pizza places, cafes, BBQ joints and seafood restaurant.

These neighbourhoods also host “Art galleries”, bookstores and apparel stores for both men and women to make some of the lifestyle market. Fitness studios, gyms and Yoga centres are among common venues in all of these neighbourhoods, allowing residents to keep a healthy lifestyle accommodating busy schedules. Finally to conclude the day, these neighbourhoods do host ice cream shops and bars where people can have a peaceful end to a hectic day.

### Conclusion:

The exploration of Chicago and then the comparison with neighbourhoods of Toronto and Manhattan, NY cite the fact that both US and Canada share a similar cultural and professional outlook with most of the venue categories matching in every department. However there were some anomalies which were unique and to the kind of the respective city, but which also adds as the city identification. This exploration is however small and restricted to a few neighbourhoods, but is an indication that it can be extended as we want and is scalable to all extent, in terms of results and amount of analysis which can be carried out.