# Capstone Report - The Comparison of Neighborhoods

# Dharmesh Patel - July 09, 2020

## Introduction

### Background

Is New York City more like Toronto?

Whenever a person relocates to new city, they search for a suitable neighborhood. They are usually interested in knowing the different neighborhoods that the city has to offer, and find out the neighborhoods with specific venue or business of their interest, and compare the options. Combining the location and the information of venues near to neighborhood would help new people/business in a city to make better informed decisions and help them to choose neighborhood amongst the many neighborhood in the city.

**Objective*:*** Primary objective of this project is to set up the process to compare neighborhood across geographical location, and provide results that help us to analyze how the selected cities and its neighborhoods compares to each other, Cities in terms of the likelihood of their neighborhoods and neighborhoods by kind of venues they offer, such as:

* Compare the cities in terms of the number of restaurants serving their favorite cuisine
* Find neighborhoods which has venues of their choice like gym, park, restaurant, market etc.

### Interested audience

Specifically, this project will be targeted to people/business new to the selected city and interested in finding out the neighborhoods with specific venue or venues, and compare it with the other neighborhoods. It will also help analyze the businesses within city.

**Scope*:*** Anyone who wants to know more about Toronto, Canada and New York city of America, and needs to find out neighborhood having at least one top venue in following category: Gym, Park, Arts & Entertainment, Beach, Coffee Shop, Sports Bar, Market, Supermarket, Asian Restaurant, Indian Restaurant.

## Data

Both the selected cities are very diverse and are the financial capitals of their respective countries, in order to compare neighborhoods and Venues across both these cities, I started with compiling address and location (longitude, latitude) data for each neighborhood of both cities and then fetched venues data from Foursquare API.

### Neighborhood Data

New York has 5 boroughs and 300 neighborhoods. I have obtained this dataset which exists for free on the web, and contains the latitude & longitude(coordinates) of each neighborhood:

[New York neighborhood data source](https://geo.nyu.edu/catalog/nyu_2451_34572)

The Toronto city has 10 boroughs and 99 neighborhoods. I have compiled this data set by scraping the Toronto postcode from Wikipedia webpage for neighborhood address and then geocode address using Python Geocoder to obtain coordinates of neighborhoods: [GitHub](https://github.com/Dharmeshpatel25Sep/Coursera-Capstone/blob/master/Toronto_df.csv)

Clean neighborhood data of both the cities includes 15 boroughs and 401 neighborhoods:

*Top 5 Boroughs based on their neighborhood counts:*

|  |  |
| --- | --- |
| *New York:* |  |
| *Borough* | ***# Neighborhoods*** |
| *Queens* | *80* |
| *Brooklyn* | *70* |
| *Staten Island* | *60* |
| *Bronx* | *52* |
| *Manhattan* | *40* |

|  |  |
| --- | --- |
| *Toronto:* |  |
| *Borough* | ***# Neighborhoods*** |
| *North York* | *20* |
| *Downtown Toronto* | *19* |
| *Scarborough* | *17* |
| *Etobicoke* | *12* |
| *Central Toronto* | *9* |

*Few rows from neighbourhood Data:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *City* | *Borough* | *Neighborhood* | *Latitude* | *Longitude* |
| *New York* | *Bronx* | *Wakefield* | *40.894705* | *-73.847201* |
| *New York* | *Bronx* | *Eastchester* | *40.887556* | *-73.827806* |
| *Toronto* | *Etobicoke* | *Islington Ave., Humber Valley Village* | *43.667856* | *-79.532242* |
| *Toronto* | *Scarborough* | *Malvern, Rouge* | *43.806686* | *-79.194353* |

### Foursquare Venue Data

From Foursquare’s explore API (https://api.foursquare.com/v2/venues/explore?), I fetched all venues (up to 100) within 500 meter radius of neighborhood's geographical location and collected their names, categories and locations. I have used this data to analyse neighborhoods having similar kind of venues and explore similar venues with in city.

Also retrieved mapping data for mapping venue to general category such as spot for Food, Entertainment etc from categories API (https://api.foursquare.com/v2/venues/categories?)

Venue data for each neighborhood of both the cities includes 11,975 venues. Data obtained from API is imported to CSV files, which are available at [GitHub](https://github.com/Dharmeshpatel25Sep/Coursera-Capstone) repository.

*Few rows from consolidated venues Data:*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *City* | *Neighborhood* | *Venue*  *Name* | *Venue*  *Latitude* | *Venue*  *Longitude* | *Venue*  *Category* | | *Venue*  *Categoryid* |
| *New York* | *Wakefield* | *Lollipops Gelato* | *40.894123* | *-73.845892* | | *Dessert Shop* | *4bf58dd8d48988d1d…* |
| *Walgreens* | *40.896528* | *-73.844700* | | *Pharmacy* | *4bf58dd8d48988d10…* |
| *Carvel Ice Cream* | *40.890487* | *-73.848568* | | *Ice Cream Shop* | *4bf58dd8d48988d1c…* |
| *Dunkin'* | *40.890459* | *-73.849089* | | *Donut Shop* | *4bf58dd8d48988d14…* |

### Data wrangling and standardization

Now, we have comprehensive data set for neighborhoods of both the cities along with venues category. so, I decided to explore the data and check that we have required data for analysis and standardize it for further analysis.

#### Data Cleaning

Checked neighborhood data for duplicates, missing or miscellaneous values and cleaned it:

* Removed few neighborhoods for which no venues were found
* Replaced miscellaneous value found for venue category like “neighborhood”

#### Dataframe for unique venues with in city

In order to compare and see how the categories of venues are distributed in both the cities, I created new dataframe having distinct list of all venues within city, as the venues data for each neighborhood will have duplicates entries for venue that falls in range of 500 meter of nearby or adjacent neighborhoods.

I found 914 duplicates venues which is removed for City level analysis, the unique venue data from both of the city has total 11061 venues.

#### Dataframe to overview top 10 venue categories of neighborhoods

Next, I created new dataframe having one row for each neighborhood along with top 10 venue categories listed in columns. which I will use to classify and cluster neighborhoods., and then run exploratory data analysis of neighborhood to determine the likelihood of neighborhoods.

* Applied one hot encoding technique for venue category over venues dataframe
* Grouped it by neighborhood and took the mean frequency of the category occurrence

*Few rows from top 10 most common Venues Categories dataframe\*:*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *City* | *Neighborhood* | *1st Most* | *2nd Most* | *3rd Most* | *4th Most* | *5th Most* |
| *New York* | *Allerton* | *Pizza Place* | *Supermarket* | *Bus Station* | *Deli / Bodega* | *Breakfast Spot* |
| *Annadale* | *Pizza Place* | *Diner* | *Pub* | *Food* | *Cosmetics Shop* |
| *Arden Heights* | *Pizza Place* | *Deli / Bodega* | *Pharmacy* | *Coffee Shop* | *Dry Cleaner* |
| *Arrochar* | *Bus Stop* | *Italian Restaurant* | *Liquor Store* | *Deli / Bodega* | *Middle Eastern Restaurant* |
| *Toronto* | *Davisville North* | *Food & Drink Shop* | *Hotel* | *Gym / Fitness Center* | *Park* | *Sandwich Place* |
| *Christie* | *Grocery Store* | *Café* | *Park* | *Restaurant* | *Candy Store* |
| *Thorncliffe Park* | *Indian Restaurant* | *Sandwich Place* | *Yoga Studio* | *Supermarket* | *Coffee Shop* |
| *Glencairn* | *Metro Station* | *Pub* | *Japanese Restaurant* | *Asian Restaurant* | *Yoga Studio* |

\* Above shown table is for illustration only, actual dataframe includes top 10 venue categories.

## Methodology

#### Data Collection:

In first step, we have collected the required data for each neighborhood in both the cities: Address, location and type (category) of venues within 500 meter radius of Neighborhood location using Foursquare API.

#### Data Wrangling:

Data we have collected is comprehensive data set for neighborhoods of both the cities along with venues (up to 100). at this stage we explored the data & checked that we have data required for our analysis, and finally transformed it in standard format for further analysis.

Now, I will start with **Exploratory Data Analysis**.

#### City Level Analysis:

I will analyse both the city according the % of venues falling in different categories. I will use the mapping table ('Categoriesmap\_df') to map venue categories to general description such as Nightlife spot, Beach, Food etc. Further, I will plot the number of restaurants serving different cuisine and types of food for both the cities and compare them.

#### Neighborhood Level Analysis:

Once I finish my City Level analysis, I will start with segmentation and clustering of neighborhoods (Data Classification: based on top 10 venue categories) without filter of city., and then I will explore few neighborhoods within city and find similar neighborhoods within/across city. and finally, I will prepare map to showcase the similar neighborhoods based on specific cluster classification.

As a final step, I will analyse these plots and map and try to draw conclusions on how both the cities are similar or different. I will discuss my findings and any inferences I have drawn so far.

## Analysis

### City level analysis

Comparison of Cities based on Venue Characteristics

I begin my analysis by taking a look at the count and % of venues and various categories of venues that exist in both the city.

*Unique counts:*

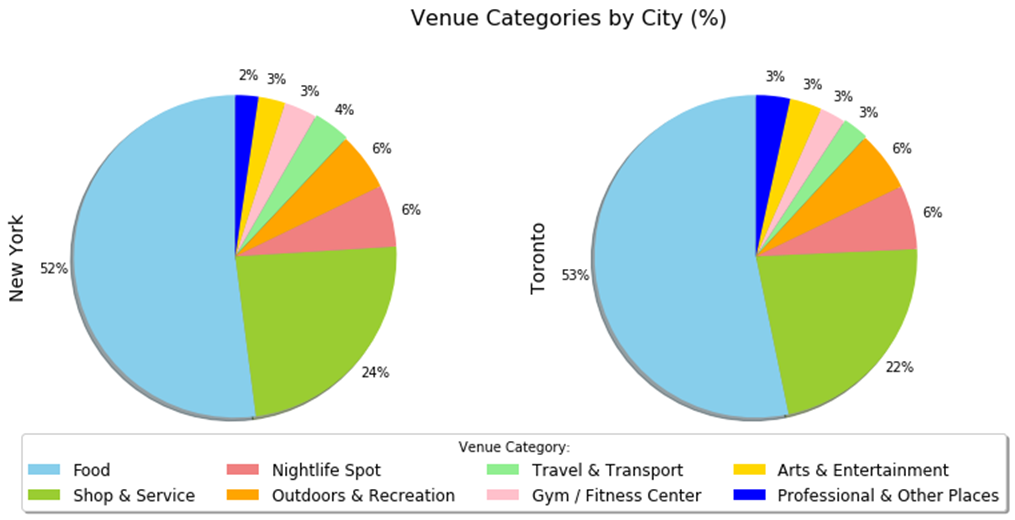
|  |  |  |  |
| --- | --- | --- | --- |
| *City* | *# Neighborhoods* | *# Venues* | *# Categories* |
| *Toronto* | *95* | *1,639* | *264* |
| *New York* | *301* | *9,422* | *428* |

New York has way too more venues compare to Toronto, also seems there are more choices available in New York. To see how the venues categories are distributed. I mapped venue categories to more general category so I can compare it across city. See example below:

|  |  |  |  |
| --- | --- | --- | --- |
| Venue | Venue Category | Venue Categoryid | Category |
| *Walgreens* | *Pharmacy* | *4bf58dd8d48988d10f951735* | *Shop & Service* |

Category wise venues distribution across cities:

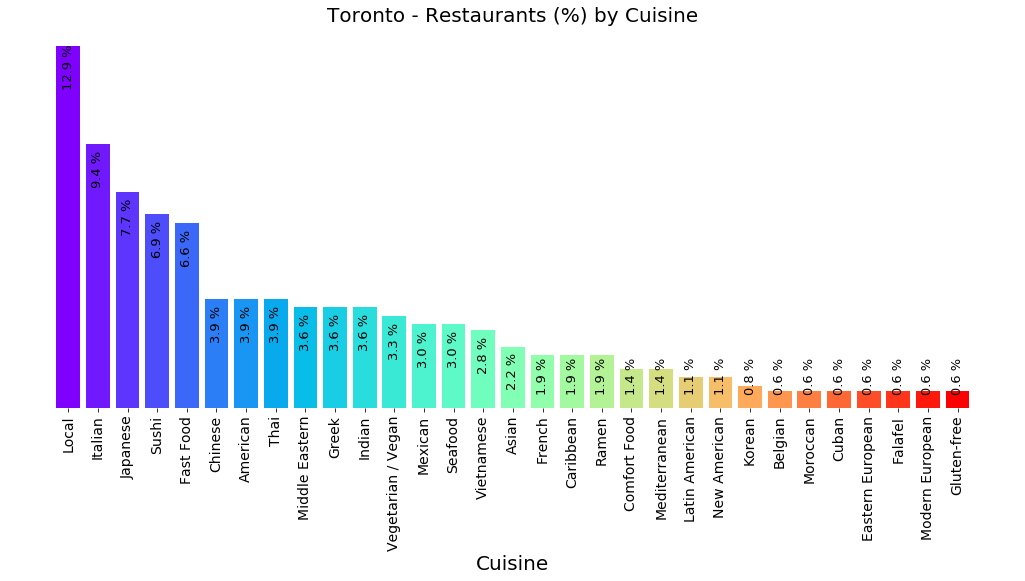
|  |  |  |
| --- | --- | --- |
| *Count of Venues by Category* | | |
| *Category* | ***New York*** | ***Toronto*** |
| *Food* | *4,895* | *870* |
| *Shop & Service* | *2,250* | *367* |
| *Nightlife Spot* | *587* | *105* |
| *Outdoors & Recreation* | *551* | *99* |
| *Travel & Transport* | *348* | *42* |
| *Gym / Fitness Center* | *308* | *42* |
| *Arts & Entertainment* | *255* | *53* |
| *Professional & Other Places* | *216* | *56* |

***Venues (%) by Category***

Pie charts of venue (%) by categories shows these cities have very identical contribution of venue categories. We can see difference of around 1-2 % for few venue categories. The top 2 venue categories contribute to **~75%** of venues, where ~52% venues fall in to **Food** and ~23% into **Shop & Services** category for both the cities.

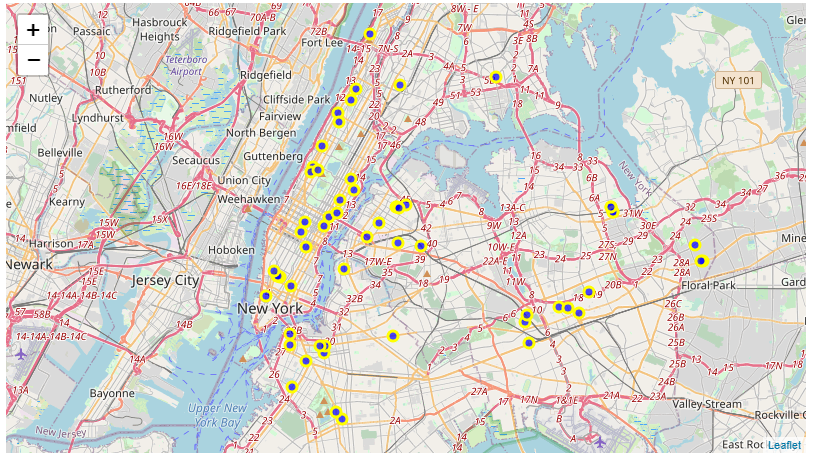
Further, plotted the % of restaurants serving different cuisine and types of food for both the cities and compare them.

***New York - Restaurants (%) by Cuisine***

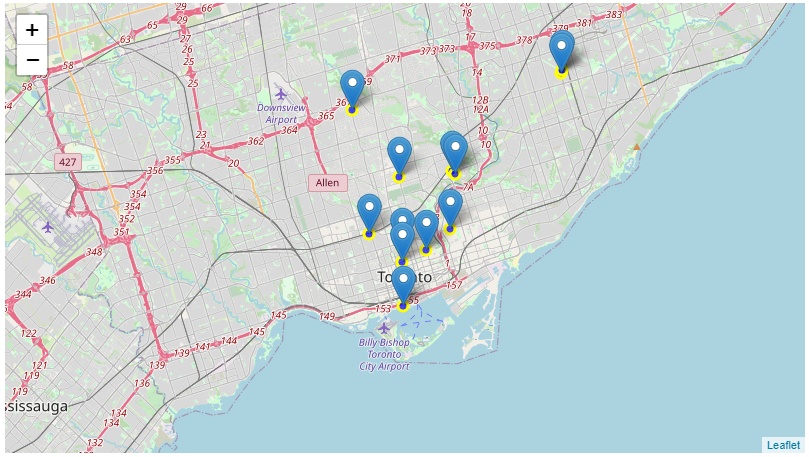
***Toronto - Restaurants (%) by Cuisine***

Looking at distribution of restaurants by cuisine, we can conclude that both cities have restaurants for cuisines famous across the globe. Closure look to the graph suggests that, cuisine in range of top 10-15 is available in both the city with different ranking by % contribution, and accounts to near about 70% of restaurants in that city. Next, I will explore Indian restaurants.

***Map of New York: Venues for Indian Restaurants***



***Map of Toronto: Venues for Indian Restaurants***

******

We can see that both the city has Indian restaurants which is spread across city and New York has much greater number of Indian restaurants than Toronto. which is expected due to different geographies and their difference in coverage area and population.

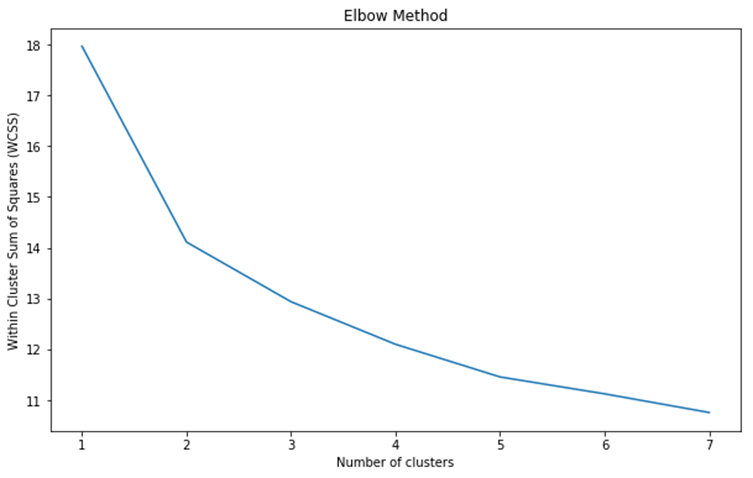
Now, we will move further to the next analysis of neighborhoods.

### Neighborhood level analysis

Comparison of Neighborhood based on top 10 Venue Categories

I took a subset of neighborhoods for clustering and further analysis as defined in scope. i.e. neighborhood which has some venues of our choice nearby such as Park, Gym, Market etc. input for clustering is grouped neighborhood data with venue categories occurrence values (refer to Data wrangling for further details).

I found out optimal number of clusters by elbow method, and cluster neighborhood data using *K-Means.*

**

Finally, I cluster all the neighborhood using K-means into 5 clusters. And reviewed, the counts of clustering number, to analyze how cluster have classified similar neighborhoods across city.

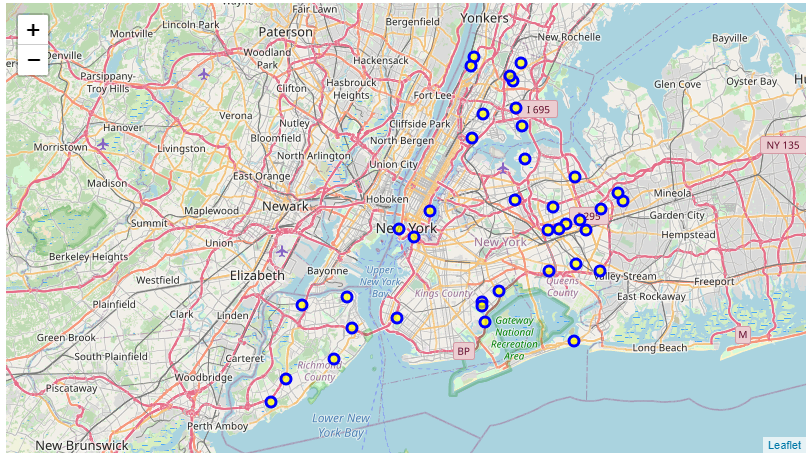
|  |  |  |
| --- | --- | --- |
| **Cluster Labels** | **City** | **# Neighborhood** |
| 0 | *New York* | *10* |
| *Toronto* | *1* |
| 1 | *New York* | *8* |
| *Toronto* | *19* |
| 2 | *New York* | *2* |
| *Toronto* | *3* |
| 3 | *New York* | *2* |
| *Toronto* | *8* |
| 4 | *New York* | *39* |
| *Toronto* | *13* |

We can see that each cluster has one or more neighborhoods from both the city, which suggest that in our dataset we have match for every neighborhood in both the city.

Now, we can examine cluster assigned to Neighborhood and determine the discriminating venue categories that distinguish it from another cluster and we can also find similar neighborhood by cluster label.

While examine the cluster (#4), I picked one of Toronto the neighborhood “Thorncliffe Park” which shown match with few of the neighborhood in New York, and tried to find out more about another match of this neighborhood with in Toronto and in New York.

I found 12 matches for “Thorncliffe Park” within Toronto. Next, I checked for neighborhood match in New York. Also, I found 38 matches from New York. Let's visualize the "Thorncliffe Park" neighborhood match for New York.



***Map: Matches of “Thorncliffe Park, Toronto” in New York***

Map above shows New York neighborhoods spread throughout city, which are like our selected neighborhood “Thorncliffe Park, Toronto” (from cluster # 4). likewise, we can explore, compare and find similar neighborhoods with in and across cities by cluster label. Finally, I plotted clusters on Map to see how the clusters are spread within city and to analyze.



***Map: Neighborhood clusters of New York City***

Above maps gives idea how the chosen cluster # 4 (marked with blue circle) differs from another cluster in our dataset. although every neighborhood has the few venues of our choice yet some are very distinct due to its proximity to specific venues nearby such as having beach (marked with green circle, cluster # 0). Next, I reviewed the cluster # 0 to validate the same.

***Neighborhood data for cluster # 0:***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *City* | *Neighborhood* | *1st Most* | *2nd Most* | *3rd Most* | *4th Most* | *5th Most* |
| *New York* | Belle Harbor | *Beach* | Pub | Deli / Bodega | Spa | Boutique |
| Breezy Point | *Beach* | Bus Stop | Supermarket | Monument / Landmark | Trail |
| Hammels | *Beach* | Diner | Fast Food Restaurant | Dog Run | Shoe Store |
| Neponsit | *Beach* | Beach Bar | Yoga Studio | Factory | Egyptian Restaurant |
| Rockaway Beach | *Beach* | Food Truck | Deli / Bodega | Bagel Shop | Seafood Restaurant |
| Rockaway Park | *Beach* | Donut Shop | Pizza Place | Ice Cream Shop | Pharmacy |
| South Beach | *Beach* | Deli / Bodega | Pier | Athletics & Sports | Yoga Studio |
| Midland Beach | Bus Stop | *Beach* | Russian Restaurant | Liquor Store | Bookstore |
| Sea Gate | Spa | *Beach* | Home Service | Bus Station | American Restaurant |
| Roxbury | Trail | Irish Pub | Deli / Bodega | Baseball Field | *Beach* |
| *Toronto* | The Beaches | Pub | Trail | Health Food Store | *Beach* | Yoga Studio |

As expected, neighborhoods in cluster # 0 are the one which have beach nearby and that specific venue sets them apart from neighborhoods of other clusters, also we can observe a match of similar neighborhood in Toronto which shares a same cluster classification (The Beaches, Toronto).

## Results and Discussion

After collecting location data for all neighborhoods of both the city, we got a list of 15 boroughs and 401 neighborhoods. which consist 302 of New York and 99 neighborhoods of Toronto. as we know New York is bigger than Toronto in area and the count of neighborhood apparently suggest that it is much more dense, further search on web suggest that it’s because of huge difference in population of both the cities.

|  |  |  |
| --- | --- | --- |
| *City* | *Area* | *Population* |
| *Toronto* | 630.2 km² | 29.3 lakhs |
| *New York* | 783.8 km² | 84 lakhs |

However, as our objective is to analyze the kind of neighborhoods (based on venues categories) that both the city has to offer , we fetched the venues within 500 meter radius of each neighborhood from Foursquare API and cleaned them, we obtained total 11,975 venues, which includes duplicates of ~900 venues which are with in radius of more than neighborhood, we further explored the data and transformed it in standard format for analysis.

Unique counts after cleaning:

|  |  |  |  |
| --- | --- | --- | --- |
| *City* | *#Neighborhoods* | *#Venues* | *#Categories* |
| *Toronto* | 95 | 1,639 | 264 |
| *New York* | 301 | 9,422 | 428 |

#### Comparison of Cities based on Venue Categories:

Given that New York has way too more venues compare to Toronto, and has more venue categories. To see how the venue categories are distributed. We mapped venue categories to more general category and compared the contribution of general venue categories of both the cities.

* We found that these cities are similar and have very identical contribution of venue categories. The majority of venues falls in to top 2 categories which contributes to ~75% of venues, where ~52% venues falls in to Food and ~23% into Shop & Services category for both the cities.
* We then plotted distribution of restaurants by cuisine they serve, and concluded that both cities have restaurants serving cuisines famous across the globe. We also created a map of venues for Indian restaurants for both the cities which showed us that Indian restaurants spread across city and not concentrated around any single neighborhood.

#### Comparison of Neighborhood based on top 10 Venue Categories:

For Neighborhood comparison analysis we took a subset of neighborhood from both the cities, having at least one top venue in below mentioned categories, we found 105 neighborhoods, 61 of New York and 44 of Toronto, and classified them in 5 cluster using K-Means based on their top venue categories.

We filtered neighborhood having any of the following category: Gym, Park, Arts & Entertainment, Beach, Coffee Shop, Sports Bar, Market, Supermarket, Asian Restaurant, Indian Restaurant.

* We found that each cluster has mix of neighborhoods from both the city, and none of the cluster found city specific, which suggest that in our dataset we have match for every neighborhood in both the city.
* We took one neighborhood to find match within city as well as in other city and explored them with the help of map and concluded that with this framework we can find and compare neighborhoods.
* Finally, we plotted neighborhood cluster on map for New York, and examined how the neighborhoods cluster are spread within city and how they differ from another cluster. we took a closure look to one cluster and concluded that although every neighborhood has the few venues of our choice yet some are very distinct due to its proximity to specific venues nearby such as having beach and as such sets them apart from neighborhoods of other clusters.

## Conclusion

The purpose of this project was to compare Toronto, Canada and New York city of America, and check likelihood of cities and their neighborhoods by kind of venues they offer and workout the process of comparison of cities and neighborhoods across geographies.

*“Venues and their details for all the neighborhood have been identified, City and Neighborhood level Exploratory Data Analysis are complete. The results suggest that despite of being on two different geographies and of different size of cities. Both the cities, New York and Toronto shares similar contribution of venue types, and has many similar kinds of neighborhood which one can choose based on individual interest. we have our framework ready and cleaned/labeled data which we can use any time to query the neighborhood or venues of our choice for both the cities, we can also use this data to suggest neighborhoods to the new person/business to these cities.”*