Battle of Neighbourhoods

Contents

I. Introduction/Business Problem 2

II. Problem Statement 2

III. Scope/Limitations: 2

IV. Major Data Description: 3

V. Methodology: 4

VI. Results and Discussions: 7

A. City Maps 7

B. Cluster Maps 8

C. Similarities/Differences 9

D. Toronto Neighbourhoods with Top Number of Venues (as per Venue Category=50), with its Boroughs and Equivalent Clusters 10

E. New York Neighbourhood with Top Number of Venues (as per Venue Category=98), with its Boroughs and Equivalent Clusters 10

F. List of Top Five (5) Common Venues (as per Venue Category) for each Neighbourhood with Top Venues (as per Venue Category) 11

G. K-Means Results 11

VII. Observations and Recommendations: 12

VIII. ..... Conclusion 12

Annex A – Inventory of Dataframes 14

I. Introduction/Business Problem

This business problem performs a comparison of the two cities, i.e. Toronto and New York to determine how similar or dissimilar they are based on available amenities or services within the areas. This type of case or problem study was selected because it serves a lot of purposes other than selecting a place where to live or migrate. Major target clients or stakeholders of this case or problem study are (1) individuals or families who are planning or deciding to live or migrate to Toronto or New York City; (2) individuals who are planning or deciding to work or get employed in Toronto or New York City; (3) a company who plans or looking for a location to expand; and (4) Data Scientists or data analyst who wants to do analysis on using technologies like machine learning techniques or other data science tools. The case study facilitates the review or analysis in coming up with a decision to select a particular place based on comparison.

The major benefits of this business problem and its solution to clients or stakeholders are (1) improved decision making not only in migration issue, i.e. work movement, business expansion; and (2) Promote ability to use technologies like machine learning to build improved or better services.

II. Problem Statement

The problem statement is: *Is Toronto more like New York City before a family decides to migrate to Toronto?*

There are several factors to assess or compare places or cities like population, crime rate, cost of living, and others. For this case study, it will *leverage or focus on the Foursquare location* and available data to explore or compare the Toronto and New York cities and/or their neighbourhoods.

III. Scope/Limitations:

1. It will leverage or focus on the F*oursquare location and available data* to explore or compare the Toronto and New York cities and/or their neighbourhoods.

2. Statistical data depends on the source and date/time of last updates. See Section III – Major Data Description for the source references.

3. There is a limit on using Foursquare API. Since there are two (2) cities to be compared, the LIMIT was set to 50 and radius of 500 in running the programs for each city.

IV. Major Data Description:

The following are the data needed with its description, purpose or how it will solve the problems, possible source, and examples:

1. Data on postal codes of Toronto City, Canada;
   Purpose: The postal codes will be used to get the equivalent latitude and longitude coordinates from existing/available geospatial data;
   Source = https://en.wikipedia.org/wiki/List of postal codes of Canada: M ;
   Example: A matrix or table to generate the Toronto data on Borough, Neighbourhood, Latitude, Longitude, and/or other needed data

2. Data on geospatial data;
   Purpose = serves as source data for getting the latitude and longitude coordinates;
   Source = http://cocl.us/Geospatial data ;
   Example: A csv file with postal code, latitude, and longitude

3. Data on latitude and longitude coordinates of New York City, USA;
   Purpose: This is needed to create the dataframes with data on Borough, Neighbourhood, Latitude, Longitude, and/or other needed data.;
   Source = https://cocl.us/new york dataset ;
   Example: A matrix or table to generate the New York data on Borough, Neighbourhood, Latitude, Longitude, and/or other needed data

4. Foursquare application credentials' Data
   Purpose: The application credentials will link an application to Foursquare API and are needed in order for Foursquare users to access the website. The credential data are the Client ID and Client Secret. The version is also needed.
   Source = https://www.foursquare.com
   Examples:
   CLIENT_ID = '<Foursquare generated client_id>'
   CLIENT_SECRET = '<Foursquare generated client_secret>'
   VERSION = '<Foursquare generated version number>

5. Foursquare location data;
   Purpose: a site where one can explore/get locations, venues, webpages, users, and other relevant data/information;
   Source: https://api.foursquare.com ;
   Examples:

```
https://api.foursquare.com/v2/venues/explore?&client_id={}&
client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        lng,
        radius,
        LIMIT)
```

May return foursquare data on latitude and longitude coordinates, venues, users, among others depending on the application being executed.

V. Methodology:

This section discusses and describes the exploratory data analysis, inferential statistical testing, and/or machine learning and reason(s) used in the case or problem study.

1. Business/Problem Understanding

This defined the intention of the project or understanding of the case or problem. The purpose/reason was to establish a clear and concise statement of the problem; and complete perspective of its objectives.

2. Data Understanding

Data were collected from appropriate sources. The purpose/reason was to determine what data needs to be collected and by what methods based on the understanding of the business/problem understanding.

- See Section IV on Major Data Description with its source.

3. Data Preparation

Collected data were transformed into usable form or format. The purpose/reason was to check for questionable, missing, or ambiguous cases; and prepared the data for further analysis. This included the **Pre-processing of Data**, i.e. Data Cleaning or Data Wrangling. This included the following:

- Renaming of columns
- Dropping of unnecessary or irrelevant columns
- Removing duplicates by merging
- Masking of data
- Getting the coordinates of data
- Transformation of the file(s) or data into a pandas dataframe library

- Loading of the data on latitude and longitude
- Defining information of interest and filtering dataframes like keeping only columns that included venue name and anything that was associated with locations

4. Data Normalization

This analyzed the data if there is a need to bring them into similar range(s). The purpose/reason was to aid in making comparisons like centering/scaling, *as applicable*.

5. Exploratory Data Analysis

This provided the context in summarizing the main characteristics of the data, gaining better understanding of the data set, determining relationships, and/or setting or extracting important variables. Important results are saved in a *dataframe* (stat1_df) to facilitate the analysis of data. The following were the exploratory data analyses used:

a. Descriptive Statistics – described the features of the data sets. The purpose/reason was to obtain a brief summary and/or measure of the data.

    - Generating basic statistics …
    - Loaded data

b. Group By – grouped the data. The purpose/reason was to help in transforming the data sets.

    - Grouped the data by neighbourhood to get the total venue for each neighbourhood.

c. Modelling - analyzed the cities and/or neighbourhoods using location data. ***K-Means,*** *a form of unsupervised machine learning* was used for this case. The purpose/reason was to explore neighbourhoods, segment them, and group them into clusters to find similar neighbourhoods in Toronto and New York cities.

    As part of modelling, important dataframes were created or generated to facilitate the analysis. See Annex A for the inventory of dataframes.

    - All Toronto and New York neighbourhoods were explored using a limit of 50 to access the Foursquare API for each city. This resulted to a dataframe consisting of all neighbourhoods and its corresponding venues.
    - A dataframe with number of unique categories of venues for each neighbourhood was created or generated.

- Unique categories were spread out in terms of column presentation by each neighbourhood. This became the important source for generating the dataframe of top ten (10) common venues for each neighbourhood.

- K-means was used to cluster the neighbourhoods using five (5) clusters. The resulting cluster labels or numbers were included in the dataframe of top ten (10) common venues.

- Each of the clusters was generated to have a better understanding of the data, i.e. 1 to 5 clusters.

d. Visualization - placed the data in a visual display or context, i.e. maps and charts to give a clear idea of what it means. The purpose/reason was to make the data more understandable for the human mind to easily comprehend and analyze like identifying patterns, trends, similarities, differences, among others.

    - Created maps like map of Toronto and New York.

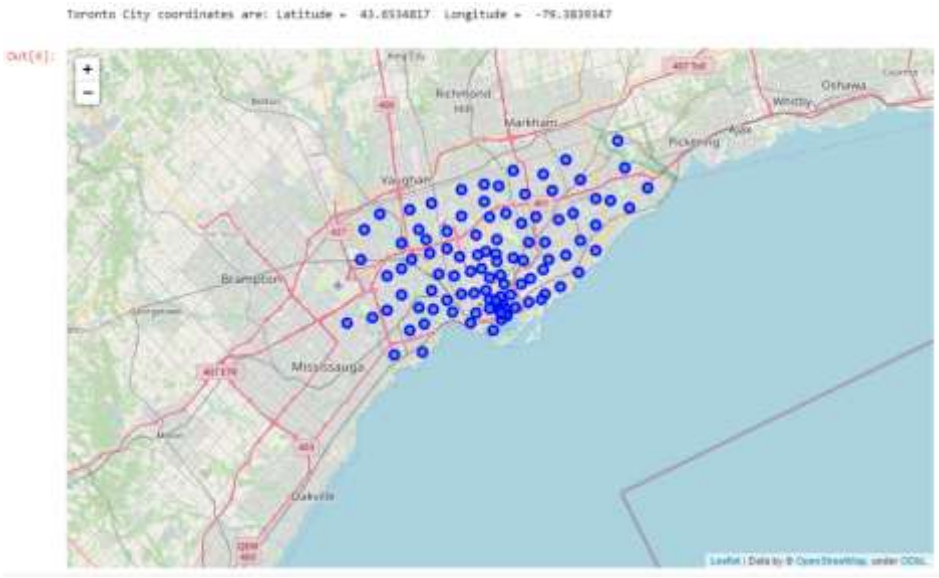
    - Resulting cluster labels or numbers were used in the creation or generation of Cluster Maps to better view the different clusters.

    - Graphed the statistics on similarity/differences as per # of boroughs, neighbourhoods, venues, unique categories, maximum number of venue categories, and number of neighbourhoods with the maximum number of venue categories.

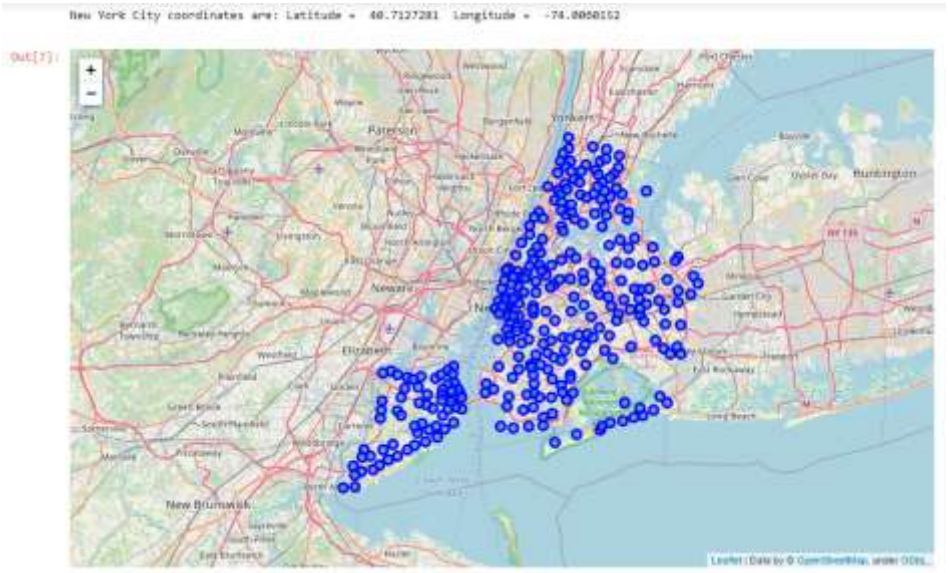
VI. Results and Discussions:

A. City Maps

Toronto Map



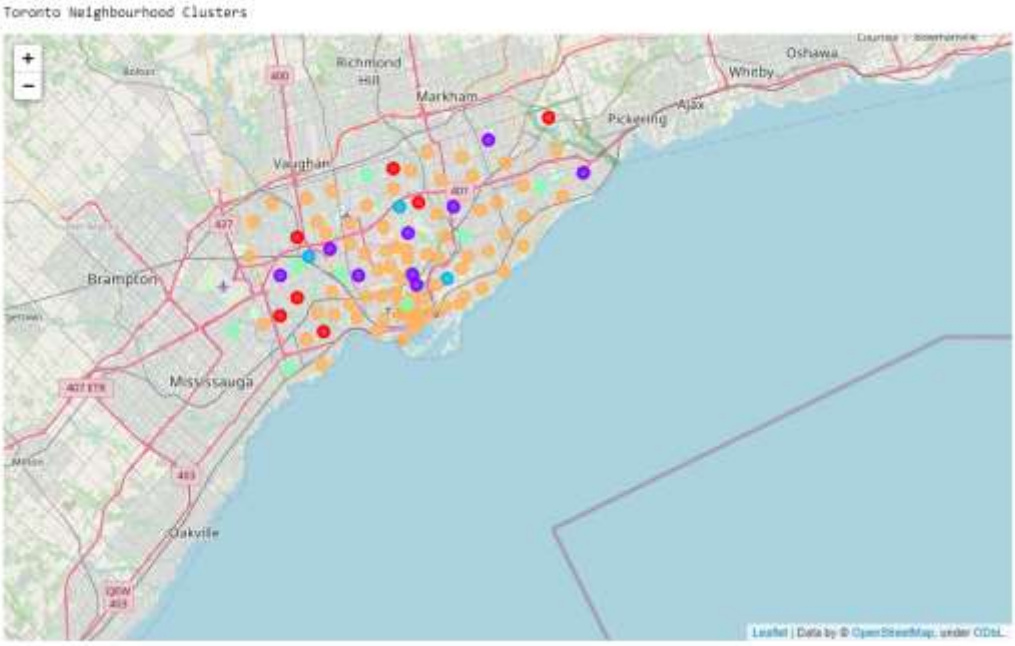
New York Map



B.Cluster Maps

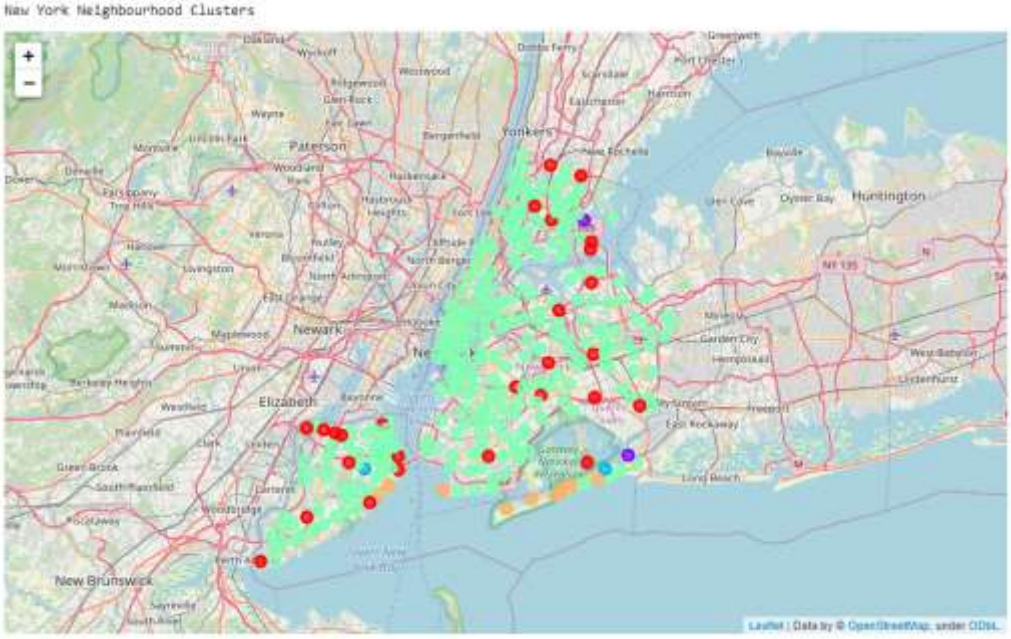
Cluster Map of Toronto Neighbourhoods

Out[28]:



Cluster Map of New York Neighbourhoods

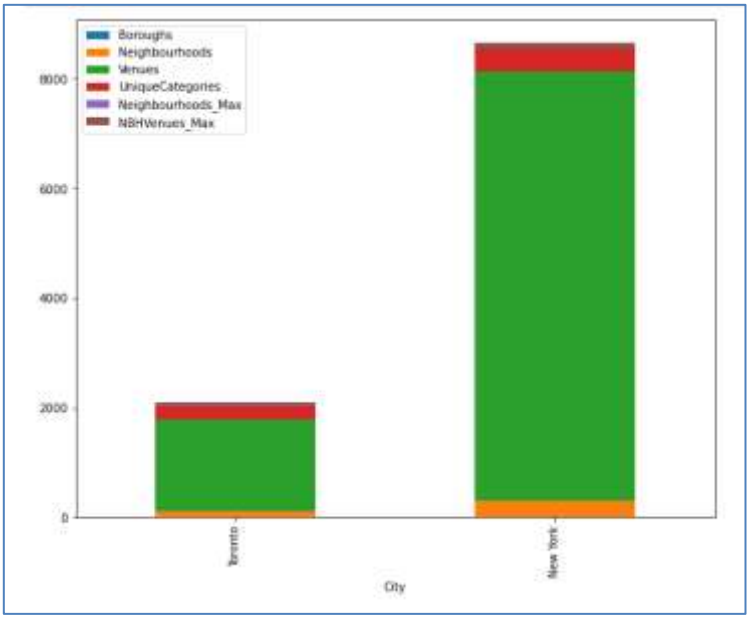
Out[29]:



C. Similarities/Differences

Out[89]:

	City	Boroughs	Neighbourhoods	Venues	UniqueCategories	Neighbourhoods_Max	NBHVenues_Max
0	Toronto	10	99	1671	251	13	50
1	New York	5	302	7824	412	1	98



1. **Boroughs & Neighbourhoods**: Toronto has 10 boroughs and 99 neighbourhoods; whereas New York has 5 boroughs and 302 neighbourhoods. This showed that Toronto has more boroughs than New York; but New York has more neighbourhoods that Toronto.

2. **Venues/Amenities**: New York has less boroughs than Toronto but with the highest number of neigbourhoods and number of venues/amenities (7,824). Toronto has 1,671 venues/amenities across its 99 neighbourhoods.

3. **Venue Categories**:

   a. With the 1,671 venues/amenities across the 99 neighbourhoods of Toronto, there are 251 unique venue categories. The highest number as per venue category among the neighbourhoods is 50. There are 13 neighbourhoods who tied up with the highest number as per venue category.

   b. With the 7,824 venues/amenities across the 302 neighbourhoods of New York, there are 412 unique venue categories. The highest number as per venue category among

the neighbourhoods is 98. There is 1 neighbourhood who has the highest number as per venue category.

D. Toronto Neighbourhoods with Top Number of Venues (as per Venue Category=50), with its Boroughs and Equivalent Clusters

Toronto

Out[90]:

	Neighbourhood	Venue	Maximum	Borough	ClusterNo
0	Berczy Park	50	Yes	Downtown Toronto	4.0
1	Central Bay Street	50	Yes	Downtown Toronto	4.0
2	Church and Wellesley	50	Yes	Downtown Toronto	4.0
3	Commerce Court, Victoria Hotel	50	Yes	Downtown Toronto	4.0
4	Fairview, Henry Farm, Oriole	50	Yes	North York	4.0
5	First Canadian Place, Underground city	50	Yes	Downtown Toronto	4.0
6	Garden District, Ryerson	50	Yes	Downtown Toronto	4.0
7	Harbourfront East, Union Station, Toronto Islands	50	Yes	Downtown Toronto	4.0
8	Kensington Market, Chinatown, Grange Park	50	Yes	Downtown Toronto	4.0
9	Richmond, Adelaide, King	50	Yes	Downtown Toronto	4.0
10	St. James Town	50	Yes	Downtown Toronto	4.0
11	Stn A PO Boxes	50	Yes	Downtown Toronto	4.0
12	Toronto Dominion Centre, Design Exchange	50	Yes	Downtown Toronto	4.0

E. New York Neighbourhood with Top Number of Venues (as per Venue Category=98), with its Boroughs and Equivalent Clusters

New York

Out[91]:

	Neighbourhood	Venue	Maximum	Borough	ClusterNo
0	Murray Hill	98	Yes	Manhattan	3.0

F. List of Top Five (5) Common Venues (as per Venue Category) for each Neighbourhood with Top Venues (as per Venue Category)

Toronto

	Neighbourhood	Venue	Maximum	Borough	ClusterNo	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Berczy Park	50	Yes	Downtown Toronto	4.0	Coffee Shop	Cocktail Bar	Beer Bar	Seafood Restaurant	Restaurant
1	Central Bay Street	50	Yes	Downtown Toronto	4.0	Coffee Shop	Café	Sandwich Place	Bubble Tea Shop	Burger Joint
2	Church and Wellesley	50	Yes	Downtown Toronto	4.0	Sushi Restaurant	Coffee Shop	Yoga Studio	Gay Bar	Japanese Restaurant
3	Commerce Court, Victoria Hotel	50	Yes	Downtown Toronto	4.0	Hotel	Coffee Shop	Café	Restaurant	Gym
4	Fairview, Henry Farm, Oriole	50	Yes	North York	4.0	Clothing Store	Coffee Shop	Fast Food Restaurant	Restaurant	Juice Bar
5	First Canadian Place, Underground city	50	Yes	Downtown Toronto	4.0	Café	Coffee Shop	Restaurant	Concert Hall	Pizza Place
6	Garden District, Ryerson	50	Yes	Downtown Toronto	4.0	Coffee Shop	Café	Clothing Store	Fast Food Restaurant	Cosmetics Shop
7	Harbourfront East, Union Station, Toronto Islands	50	Yes	Downtown Toronto	4.0	Coffee Shop	Aquarium	Brewery	Hotel	Park
8	Kensington Market, Chinatown, Grange Park	50	Yes	Downtown Toronto	4.0	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Mexican Restaurant	Gaming Cafe
9	Richmond, Adelaide, King	50	Yes	Downtown Toronto	4.0	Coffee Shop	Café	Steakhouse	Concert Hall	American Restaurant
10	St. James Town	50	Yes	Downtown Toronto	4.0	Café	Gastropub	Coffee Shop	Seafood Restaurant	Farmers Market
11	Stn A PO Boxes	50	Yes	Downtown Toronto	4.0	Café	Coffee Shop	Cheese Shop	Hotel	Farmers Market
12	Toronto Dominion Centre, Design Exchange	50	Yes	Downtown Toronto	4.0	Café	Coffee Shop	Seafood Restaurant	Hotel	Restaurant

New York

Out[117]:

	Neighbourhood	Venue	Maximum	Borough	ClusterNo	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Murray Hill	98	Yes	Manhattan	3.0	Korean Restaurant	Coffee Shop	Bar	Sandwich Place	Japanese Restaurant

The first most common venue in New York is a Korean Restaurant; while Toronto has coffee shops, Sushi Restaurant, Hotel, Café,

G. K-Means Results

The use of K-means to cluster the neighbourhoods into 5 clusters confirmed the presence of the 13 Toronto neighbourhoods with the top venues (as per venue category). Similarly, it also confirmed the presence of 1 New York neighbourhood with the top venues (as per venue category).

VII. Observations and Recommendations:

Observations:

1. There are different sources where one can get the applicable or appropriate data to be used or analyzed. One has to analyze its contents, structure, and other factors that will be needed for better use or appreciation.

2. Dataframes are powerful but one has to gain deeper understanding on how to do further manipulations and familiarity with the syntaxes.

3. There is a need to strategize beforehand on how the results can be presented for better understanding of the results towards decision making or conclusion.

4. As one executes a program code, there is a need to analyze further and comprehensively the results and do tests to confirm the accuracy of the data.

Recommendations based on the results:

K-means is a good model to use. In this case or problem study, the k-means clustering was executed beforehand. Further analysis using dataframes and analysis of data have confirmed its appropriateness and importance. Thus, recommendation is: one has to really focus on the data and do tests to confirm its accuracy.

VIII. Conclusion

Is Toronto more like New York City before a family decides to migrate to Toronto? Further review of the "more like" means something is better or more agreeable or satisfying. Based on Section VI Results, Toronto is more like New York City based on the following:

- Toronto has more neighbourhoods that top the maximum number of venues category; and where a family deciding to migrate has several choices. New York has only one neighbourhood inspite of the fact that it has more venue categories.

- 13 neighbourhoods of Toronto have the same number of different venues' categories which topped the list. A family has more choices when it comes to travelling within the area because the neighbourhoods belong to same cluster (Cluster 4) as well.

Further, since there are more neighbourhoods in Toronto to choose with wide variety of venues' categories, a family deciding to migrate to Toronto may use the data with the list of neighbourhoods and its top 5 most common venues.

Annex A – Inventory of Dataframes

#	Dataframe Name	Description	Contents
1	toronto_df	Toronto data	Index number Borough Neighbourhood Latitude Longitude
2	newyork_df	New York data	Index number Borough Neighbourhood Latitude Longitude
3	stat1_df	Consolidated Statistics of Toronto and New York	City Boroughs Neighbourhoods Venues UniqueCategories Neighbourhoods with maximum # of venues Maximum # of venues
4	toronto_venues	Toronto venues from Foursquare	Index number Neighbourhood Borough Neighbourhood Latitude Neighbourhood Longitude Venue Venue Latitude Venue Longitude Venue Category
5	newyork_venues	New York venues from Foursquare	Index number Neighbourhood Borough Neighbourhood Latitude Neighbourhood Longitude Venue Venue Latitude Venue Longitude Venue Category
6	toronto_groupdf	Toronto grouped by neighbourhood	Important Data Neighbourhood # of Venues (Other Data)
7	newyork_groupdf	New York grouped by neighbourhood	Important Data Neighbourhood # of Venues (Other Data)
8	toronto_onehot	Toronto one hot encoding (neighbourhood) Dataframe used : toronto_venues: venue_category	Index number Neighbourhood Columns [Accessories Store, Airport, Airport Food Court, Airport Lounge, Airport Service …]
9	newyork_onehot	New York one hot	Index number

		encoding (neighbourhood) Dataframe used : toronto_venues: venue_category	Neighbourhood Columns [Accessories Store, Adult Boutique, Afghan Restaurant, African Restaurant …]
10	toronto_grouped	Toronto grouped Dataframe used: toronto_onehot	Index number Neighbourhood *Columns [Accessories Store, Adult Boutique, Afghan Restaurant, African Restaurant …] *mean data
11	newyork_grouped	New York grouped Dataframe used: toronto_onehot	Index number Neighbourhood Columns [Accessories Store, Adult Boutique, Afghan Restaurant, African Restaurant …]
12	toronto_neighborhoods_venues_sorted	Toronto	Index number Neighbourhood 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue … 10th Most Common Value
13	newyork_neighborhoods_venues_sorted	New York	Index number Neighbourhood 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue … 10th Most Common Value
14	toronto_merged	Toronto clusters	Index number Neighbourhood Latitude Longitude Cluster Labels 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue … 10th Most Common Value
15	newyork_merged	New York clusters	Index number Neighbourhood Latitude Longitude Cluster Labels 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue … 10th Most Common Value
16	rslt_torontodf	Toronto Dataframe of Neighbourhoods with most venues	

17	rslt_newyorkdf	New York Dataframe of Neighbourhoods with most venues	
18	xxxfinaltor	Toronto Dataframe of Neighbourhoods with most venues + Boroughs	Index Neighbourhood Venue Maximum Borough
19	xxxfinalny	Toronto Dataframe of Neighbourhoods with most venues + Boroughs	Index Neighbourhood Venue Maximum Borough
20	stat1_df	Statistics data	Index number City No. of Boroughs No. of Neighbourhoods No. of Venues (Limit=50, Radius of 500) No. of Unique Categories