Capstone Project - The Battle of Neighborhoods (Week 2)

To explore and analyze
The venues of neighborhoods of New York City and City of Toronto
Using Four Square Venue data

Table of Contents

1. Introduction

2. Data Description

3. Methodology and Exploratory Data Analysis

4. Results

5. Discussion

6. Conclusion

1.Introduction

To explore and analyze, the venues of the neighborhoods of New York City and City of Toronto using Four Square Venue data to ensure the ease of decision making.

1.1 Background

At times when we are visiting or moving to a new city, it becomes very difficult to discover or to choose restaurants, stores and other local business venues at the neighborhoods from user perspective and to start up a venue from a business point of view.

Decision making will be a huge task with lot of criteria in mind to select the venue based on distance, price tier, ratings and sometimes we often stuck at certain stages in deciding. In which the user contributions will also been an important element such as user likes of a venue, upload of venue photos by the user and user tips of a venue, which will be counted to the comparisons and the decision that we finally make out.

About the two cities

New York City is made up of five major areas or “boroughs” sitting where the Hudson River meets the Atlantic Ocean. some separated by rivers and connected via ferry or bridge. The five boroughs of New York are **Manhattan, Brooklyn, Queens, Staten Island** and **the Bronx.**

Toronto is a city of neighborhoods. each with its own style, vibe and scene. We might find ourselves in a shopping mecca in the morning, a historic market around lunchtime, and surrounded by popular bars at night. One thing Toronto doesn’t have a shortage of is shopping, whether it be outlet shopping, thrift or on trend pieces, there’s a neighborhood for it. The neighborhoods are **East York, Etobicoke, North York, Old City of Toronto, Scarborough** and **the York**

1.2 Business Problem

To explore four-square venue data, to get new insights to recommend venue owners to start a new venue. To compare and conclude the better city on venues. To cluster & segment the venues to enhance user experience on selecting the venues basis the below analysis.

- To compare and distinguish venue similarities of the two cities using frequency distribution all columns with percentile and by various methods, parameters.
- To analyze if there is any correlation exist between the user likes, rating, price tier, tips, photos and distance.
- To perform clustering and segmentation for the venues based on four square venue details.

1.3 Target Audience

Business personnel who wants to startup a new venue, this analysis will be a providing a detailed insight for them on the opportunity available to launch a venue on certain venue categories to target the users in fulfilling business requirements.

Stakeholders to add new venues to the site in the respective city or neighborhoods having less than two or NIL venues in the categorical segment.

The analysis will give an overview to the existing venue owners to have an insight of the user behavior, patterns and trends, which will be useful to enhance their business or to revisit the existing business model.

The analysis will also be useful for the venue owners to view as how the user likes, tips and photos of the venue will bring effectiveness to the Price Tier and Ratings of the venue.

The stakeholders both internal and external will be benefited on the venue clustering to choose their venues in time and to classify their venues better.

2. Data Description

2.1 Cities to be Analyzed

- New York City
- City of Toronto

2.2 Data Source

2.2.1 Wikipedia: To download cities neighborhood list. Links given below,

Link 1: New York City
https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

Link 2: City of Toronto
https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

2.2.2 Geopy Library: To download location coordinates, latitude & longitude.

```
from geopy.geocoders import Nominatim
geolocator = Nominatim()
geolocator = Nominatim(user_agent="LN-Capstone-test")
adrs='New York'
location = geolocator.geocode(adrs)
tlatitude = location.latitude
tlongitude = location.longitude
```

2.2.3 FourSquare: To download places data & Augment basic venues details

a: Link 3 using Lat & Long: **https://api.foursquare.com/v2/venues/explore?**
b: Response details: Venue ID, name, category, id, address, distance, lat and lng.
c: Limit: 100

a: Link 4 using Venue ID: https://api.foursquare.com/v2/venues/{}?
b: Response details: likes.count, photos.count, rating, ratingSignals, reasons.count, tips.count, verified and price.tier
c: Limit: 50

2.3 Wikipedia Source Data Frame for two cities

2.3.1 City of Toronto

```
dftn.head()
```

	Neighborhood	Borough	Population	Average Income	lanper	lannam
0	Agincourt	Scarborough	44,577	25,750	19.3	Cantonese
1	Alderwood	Etobicoke	11,656	35,239	6.2	Polish
2	Alexandra Park	Old City of Toronto	4,355	19,687	17.9	Cantonese
3	Allenby	Old City of Toronto	2,513	245,592	1.4	Russian
4	Amesbury	North York	17,318	27,546	6.1	Spanish

2.3.2 New York City

```
dfny.head()
```

	Borough	code	Neighborhood
0	Bronx	CB 1	Melrose
1	Bronx	CB 1	Mott Haven
2	Bronx	CB 1	Port Morris
3	Bronx	CB 2	Hunts Point
4	Bronx	CB 2	Longwood

2.4 GEOSPY to fetch location Coordinates – Latitude and Longitude

2.4.1 City of Toronto – Coordinates data frame

```
dftnbrgll.head(10)
```

	Borough	Latitude	Longitude
0	Scarborough,Toronto	43.773077	-79.257774
1	Etobicoke,Toronto	43.671459	-79.552492
2	Old City of Toronto,Toronto	43.653963	-79.387207
3	North York,Toronto	43.770817	-79.413300
4	East York,Toronto	43.691339	-79.327821
5	York,Toronto	43.689619	-79.479188

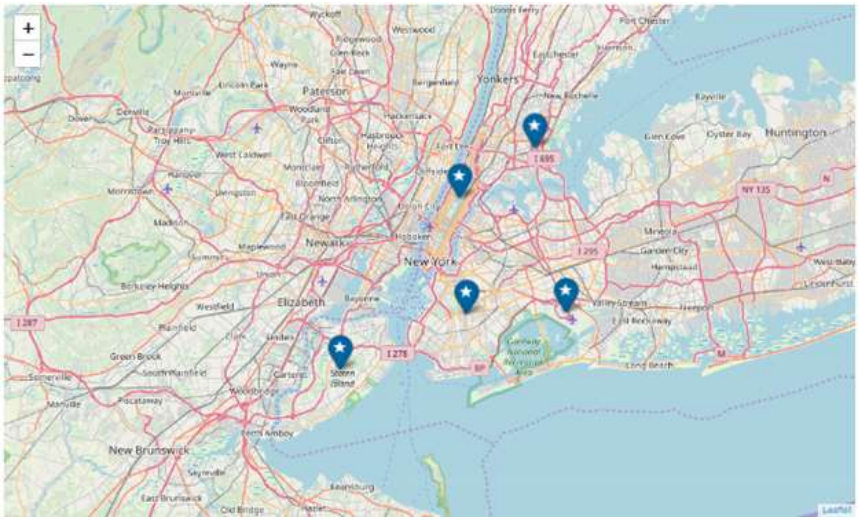
Visualization of Toronto City Coordinates

2.4.2 New York City – Coordinates data frame

```
dfnybcsvl.head(10)
```

	Borough	Latitude	Longitude
1	Bronx ,New York	40.850485	-73.840404
2	Brooklyn ,New York	40.650104	-73.949582
3	Manhattan ,New York	40.789624	-73.959894
4	Queens ,New York	40.652493	-73.791421
5	Staten Island ,New York	40.583456	-74.149605

Visualization of New York City Coordinates



2.5 FourSquare (Part 1): To download Venue id based on Latitude and Longitude

2.5.1 City of Toronto: Data frame of venue id, name, category & distance

	name	categories	distance	lat	lng	id	Borough
0	Knuckle Sandwich	Sandwich Place	545	43.696194	-79.328749	56df62a4498e96e8608b5e94	East York
1	East York Farmers' Market	Farmers Market	110	43.690482	-79.328509	4e3816b18877541e90eba62f	East York
2	Mon K Patisserie	Pastry Shop	636	43.696922	-79.329520	51b0bcb6498e4f0309a58a65	East York
3	The Wren	American Restaurant	987	43.682467	-79.328079	5155ca12e4b0e05526806bdf	East York
4	Rendez-Vous Restaurant Bar & Cafe	Ethiopian Restaurant	976	43.682570	-79.327544	4ad9221cf964a520671821e3	East York

2.5.2 New York City: Data frame of venue id, name, category & distance

	name	categories	distance	lat	lng	id	Borough
0	LA Fitness	Gym / Fitness Center	154	40.849739	-73.841949	53d1b12d498ea039475dec73	Bronx
1	Residence Inn by Marriott New York The Bronx a...	Hotel	190	40.850020	-73.842579	54932887498ee0902b1ed511	Bronx
2	Starbucks	Coffee Shop	325	40.851371	-73.844087	57b25375498e76f51c083656	Bronx
3	Zeppieri & Sons Italian Bakery	Bakery	796	40.847119	-73.832057	4c9205941adc370460a134d1	Bronx
4	Empire Bagels	Bagel Shop	840	40.849392	-73.830527	4c249bffa852c9285f52e36c	Bronx

2.6 FourSquare (Part 2): To download user interaction details based on venue id

2.6.1 City of Toronto data frame of user likes, rating, price tier, photo count & tips

```
dftrvopt.head()
```

	Unnamed: 0	id	likes.count	name	photos.count	price.tier	rating	ratingSignals	reasons.count	tips.count	verified
0	0.0	56df62a4498e96e8608b5e94	12	Knuckle Sandwich	14	1.0	8.4	20.0	0	7	False
1	0.0	4e3816b18877541e90eba62f	8	East York Farmers' Market	37	NaN	7.8	12.0	0	5	False
2	0.0	51b0bcb6498e4f0309a58a65	19	Mon K Patisserie	38	NaN	8.1	29.0	0	14	False
3	0.0	5155ca12e4b0e05526806bdf	132	The Wren	175	2.0	8.8	183.0	1	48	False
4	0.0	4ad9221cf964a520671821e3	35	Rendez-Vous Restaurant Bar & Cafe	24	2.0	9.0	53.0	1	23	False

2.6.2 New York City data frame of user likes, rating, price tier, photo count & tips

```
dfnyvopt.head()
```

	Unnamed: 0	id	likes.count	name	photos.count	price.tier	rating	ratingSignals	reasons.count	tips.count	verified
0	0.0	53d1b12d498ea039475dec73	43	LA Fitness	1	NaN	8.6	52	1	4	True
1	0.0	54932887498ee0902b1ed511	25	Residence Inn by Marriott New York The Bronx a...	43	NaN	8.7	27	0	0	True
2	0.0	57b25375498e76f51c083656	26	Starbucks	11	1.0	8.5	31	0	3	True
3	0.0	4c9205941adc370460a134d1	31	Zeppieri & Sons Italian Bakery	27	1.0	9.4	41	1	10	True
4	0.0	4c249bffa852c9285f52e36c	31	Empire Bagels	19	1.0	8.6	39	1	8	False

2.7 Consolidation of final data frame from all 4 data sources

2.7.1 City of Toronto Final Data Frame

```
In [104]: FinVenueTor
```

	name	categories	distance	lat	lng	id	Borough	State	likes.count	photos.count	rating	ratingSignals	reasons.count
0	Knuckle Sandwich	Sandwich Place	545	43.696194	-79.328749	56df62a4498e96e8608b5e94	East York	Toronto	12	14	8.4	20.0	0
1	East York Farmers' Market	Farmers Market	110	43.690482	-79.328509	4e3816b18877541e90eba62f	East York	Toronto	8	37	7.8	12.0	0
2	Mon K Patisserie	Pastry Shop	636	43.696922	-79.329520	51b0bcb6498e4f0309a58a65	East York	Toronto	19	38	8.1	29.0	0
3	The Wren	American Restaurant	987	43.682467	-79.328079	5155ca12e4b0e05526806bdf	East York	Toronto	132	175	8.8	183.0	1
4	Rendez-Vous Restaurant Bar & Cafe	Ethiopian Restaurant	976	43.682570	-79.327544	4ad9221cf964a520671821e3	East York	Toronto	35	24	9.0	53.0	1
5	Red Rocket Coffee	Café	1003	43.682340	-79.328530	4f0f11fbe4b035449917e927	East York	Toronto	64	119	8.7	92.0	1
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	East	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

2.7.2 New York City Final Data Frame

lat	lng	id	Borough	State	likes.count	photos.count	rating	ratingSignals	reasons.count	tips.count	verified	price.tier
3.696194	-79.328749	56df62a4498e96e8608b5e94	East York	Toronto	12	14	8.4	20.0	0	7	False	1.0
3.690482	-79.328509	4e3816b18877541e90eba62f	East York	Toronto	8	37	7.8	12.0	0	5	False	NaN
3.696922	-79.329520	51b0bcb6498e4f0309a58a65	East York	Toronto	19	38	8.1	29.0	0	14	False	NaN
3.682467	-79.328079	5155ca12e4b0e05526806bdf	East York	Toronto	132	175	8.8	183.0	1	48	False	2.0
3.682570	-79.327544	4ad9221cf964a520671821e3	East York	Toronto	35	24	9.0	53.0	1	23	False	2.0

2.7.3 Final Structure of the Consolidated Data Frame from 4 Data Sources

City of Toronto data frame info

```
dfmsttn.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 568 entries, 0 to 567
Data columns (total 16 columns):
name             568 non-null object
categories       568 non-null object
distance         568 non-null int64
lat              568 non-null float64
lng              568 non-null float64
id               568 non-null object
Borough          568 non-null object
State            568 non-null object
likes.count      568 non-null int64
photos.count     568 non-null int64
rating           553 non-null float64
ratingSignals    553 non-null float64
reasons.count    568 non-null int64
tips.count       568 non-null int64
verified         568 non-null bool
price.tier       355 non-null float64
dtypes: bool(1), float64(5), int64(5), object(5)
memory usage: 67.2+ KB
```

New York City data frame info

```
dfmstny.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 16 columns):
name             500 non-null object
categories       500 non-null object
distance         500 non-null int64
lat              500 non-null float64
lng              500 non-null float64
id               500 non-null object
Borough          500 non-null object
State            500 non-null object
likes.count      500 non-null int64
photos.count     500 non-null int64
rating           498 non-null float64
ratingSignals    498 non-null float64
reasons.count    500 non-null int64
tips.count       500 non-null int64
verified         500 non-null bool
price.tier       256 non-null float64
dtypes: bool(1), float64(5), int64(5), object(5)
memory usage: 59.2+ KB
```

2.8. How data will be used to solve the problem?

The above final data frame will be used to solve the problem statement

- Frequency distribution table with Venue category and Neighborhoods will be calculated. Top 15 venues will be filtered to identify opportunity of the missing slots or neighborhood having less than 2 venues will be recommended to business owner to start up new venues or opportunity to add more venue in the respective neighborhood to the site.

- Regression analysis will be performed to explore if there is any correlation exist between rating, price tier, user likes, tips, photo count and distance.

- Overall frequency tables will be calculated with percentile to distinguish the similarities between the two cities on all the columns and various visualization technique will be performed to observe the difference between two cities under various methods and parameters.

- K-Mean algorithm will be performed to cluster and segment the venues based on distance, price tier, user likes, photo count, tips and rating to enhance user experience in selecting the venues.

3. Methodology and Exploratory Data Analysis

3A. Methodology

In line with the problem statement, the methods mentioned below are used to bring the insights as a resultant to our decision making.

- To start with the initial process, the data source of the two cities will be acquired from Wikipedia for details such as Borough and Neighborhoods.
- Python Geopy, Nominatim library is used to fetch the geo coordinates of the two cities with Borough and Neighborhood location details.
- Based on the above two points, we will obtain the main source from Foursquare on the Venue details in two parts.
- The final dataset will be created using the above said details in a single file and cleaning process will be completed to start the exploratory data analysis.
- In the exploratory data analysis, we will be using the frequency table for each city to analyze the neighborhood percentile, venue category location wise spread to analyze new venue opportunities and enhancements.
- Data visualization methods applied such as seaborn bar plot, strip plot and regression plots to identify the strength and correlation between key entities.
- K-Means cluster algorithm applied for venue segmentation, Elbow method used to identify number of clusters, k-means prediction algorithm used to generate the cluster labels for each record in the dataset. Scatter plot used for graphical visualization and python folium library utilized to generate visualization on maps using their respective venue location coordinates.

The following exploratory data analysis will provide a detailed insight on various parameters.

3B. Exploratory Data Analysis

3.1 The Neighborhood Comparison of the two cities.

3.1.1 City of Toronto

	Neighborhood	%-age
Old City of Toronto	64	36.78%
North York	40	22.99%
Scarborough	29	16.67%
Etobicoke	25	14.37%
York	10	5.75%
East York	6	3.45%

3.1.2 New York City

	Neighborhood	%-age
Queens	86	26.22%
Brooklyn	79	24.09%
Bronx	60	18.29%
Staten Island	56	17.07%
Manhattan	47	14.33%

Observations	Toronto	New York
No of boroughs	6	5
No of neighborhood	174	328
Average neighborhood/borough	29	67
Top 3 neighborhood count and %	133(76%)	225(67%)

***New York City neighborhood count is 47% higher when compared to Toronto City**

3.2 Venue comparison of the two cities

3.2.1 New York City

Borough	Venue-s	User Likes	like %	Avg-Rating	Avg-Price	Total Photos	Total Tips
Bronx	100	3066	6.27%	7.78	1.55	2637	1131
Brooklyn	100	3245	6.64%	7.80	1.47	2565	947
Manhattan	100	35499	72.62%	8.15	1.81	57630	4807
Queens	100	4232	8.66%	6.99	1.47	6960	1360
Staten Island	100	2839	5.81%	7.76	1.67	2330	1059

3.2.2 City of Toronto

	Venue-s	User Likes	likes %	Avg-Rating	Avg-Price	Total Photos	Total Tips
East York	100	2928	11.24%	7.86	1.61	4041	1334
Etobicoke	101	1901	7.3%	7.40	1.76	2126	719
North York	66	2047	7.86%	7.64	1.64	1975	600
Old City of Toronto	100	15282	58.67%	8.20	1.86	26420	3897
Scarborough	100	1873	7.19%	7.26	1.52	2265	804
York	101	2015	7.74%	7.55	1.53	2853	864

Observation:

*New York City Venue user likes are higher by 47% compared to Toronto City

*New York City Venue user photo count are higher by 45% and user tips are higher by 12% compared to Toronto.

*Average rating of a venue is similar for both the cities, marginal difference found by 0.05.

*Average price tier for both the cities are similar with marginal difference of 0.06.

*Similarity: There is only 1 location for both the cities contributing more than 55% on user likes, photo count and user tips.

*Old City of Toronto with 59% user likes and Manhattan with 73% user likes are the highlights of these two cities.

3.3 Total Number of Venues. Category and Location wise status.

3.3.1 Toronto City - Top 15 Categories Venue count

categories	East York	Etobicoke	North York	Old City of Toronto	Scarborough	York	Total
Coffee Shop	5	11	4	8	9	7	44
Café	10	1	2	3	0	5	21
Bakery	3	3	1	2	3	6	18
Chinese Restaurant	0	2	0	1	10	1	14
Grocery Store	0	3	5	0	1	4	13
Sandwich Place	1	3	1	2	3	2	12
Italian Restaurant	1	2	0	1	0	8	12
Park	2	2	4	1	1	1	11
Indian Restaurant	3	0	0	1	5	1	10
Burger Joint	1	3	1	1	1	3	10
Pizza Place	5	1	2	1	0	1	10
Breakfast Spot	2	1	0	2	3	2	10
Sushi Restaurant	1	2	2	2	1	2	10
Restaurant	0	4	0	1	2	2	9
Bar	3	0	0	3	0	3	9

3.3.1.1 Observation

- Top 5 venue categories are coffee shop, bakery, Chinese restaurant and grocery stores.
- Coffee shops, bakery, sandwich, park, sushi and burger joints are found across all locations.
- Scarborough has the highest number of Chinese Restaurant and Coffee shops.
- Grocery venue found only at North York and Etobicoke. New venues can be opened.
- York location has the maximum no of Italian Restaurants and all venue categories are present.

3.3.1.2 Business Opportunities

We have observed zero venues at certain location and categories, business owners can review the same to start up new venues based on these categories

- Grocery store venue can be opened at East York and Old City of Toronto, which has highest 64 neighborhood, contributes to 37%.
- Chinese, Italian and Indian restaurants can be opened at location having zero venues
- Old City of Toronto with highest user likes at 59%, huge opportunity to increase venues like bakery, grocery stores, burger joints and restaurants.

3.3.2 New York City - Top 15 Categories Venue count

categories	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Total
Italian Restaurant	8	1	3	0	9	21
Pizza Place	8	5	2	1	4	20
Coffee Shop	4	0	3	7	3	17
Caribbean Restaurant	0	14	0	1	0	15
Bakery	5	5	1	0	4	15
Park	1	1	8	0	2	12
Deli / Bodega	3	3	3	1	2	12
Airport Lounge	0	0	0	12	0	12
Donut Shop	4	2	0	3	2	11
Mexican Restaurant	2	2	2	2	2	10
Grocery Store	1	3	3	0	3	10
Cosmetics Shop	2	0	1	5	2	10
Café	0	3	3	2	1	9
Playground	0	1	8	0	0	9
Rental Car Location	1	0	0	8	0	9

3.3.2.1 Observation

- Top 5 venue are under pizza place, coffee shop, Italian & Caribbean restaurant.
- Queens location doesn't have bakery, grocery, park & Italian restaurant venues.

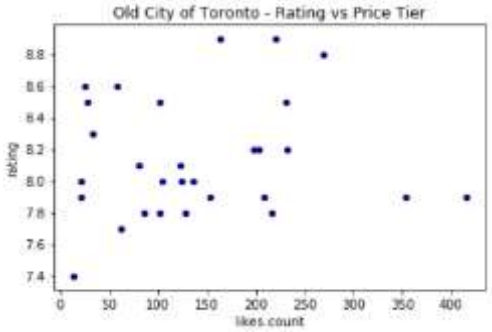
3.3.2.1 Business Opportunities

We have observed zero venues at certain location and categories, business owners can review the same to start up new venues based on these categories.

- Queens with the highest (86, 26%) neighborhoods. Bakery and Grocery store can be started.
- Coffee shop at Brooklyn location can be started.
- Manhattan location having highest number of users likes (73%), business owners can startup new venues on Airport Lounge, Donut shop, Rental Car Location, Caribbean restaurant and bakeries also can be increased since the users are high.

3.4 To analyze price tier 2: user likes vs rating on venue categories of Manhattan and Old City of Toronto.

3.4.1 Old City of Toronto



categories	7.4	7.7	7.8	7.9	8.0	8.1	8.2	8.3	8.5	8.6	8.8	8.9	Total
Bar	0	0	0	1	0	1	0	1	0	0	0	0	3
Vegetarian / Vegan Restaurant	1	0	0	0	0	0	0	0	0	2	0	0	3
Steakhouse	0	1	0	0	0	0	0	0	0	0	1	0	2
Tea Room	0	0	0	0	0	1	0	0	1	0	0	0	2
American Restaurant	0	0	1	1	0	0	0	0	0	0	0	0	2

3.4.1.1 Observation

- In the old city of Toronto. Bar, vegetarian restaurant, steakhouse tops the list based on total venues.
- The two American restaurants having less than 8.0 ratings when compared to the vegetarian restaurant having rating at 8.6.
- Scatter plot indicates higher the likes higher the ratings, barring the two outliers at 350 and 400.

3.4.2 Manhattan, New York City

categories	7.2	7.4	7.5	7.8	8.0	8.3	8.4	8.7	8.9	Total
American Restaurant	0	0	1	0	0	0	0	0	1	2
Café	0	0	0	0	0	1	0	1	0	2
Bar	0	0	1	0	0	0	0	0	0	1
Burger Joint	0	0	0	0	1	0	0	0	0	1
Cocktail Bar	0	0	0	0	0	0	1	0	0	1

3.4.2.1 Observation

- At Manhattan location, American restaurant has the highest rating topping the list with two venues.
- Burger Joint and Bars has lesser rating which is less than and equal to 8.0.
- The scatter plot indicates as the user likes increases the ratings also increasing.

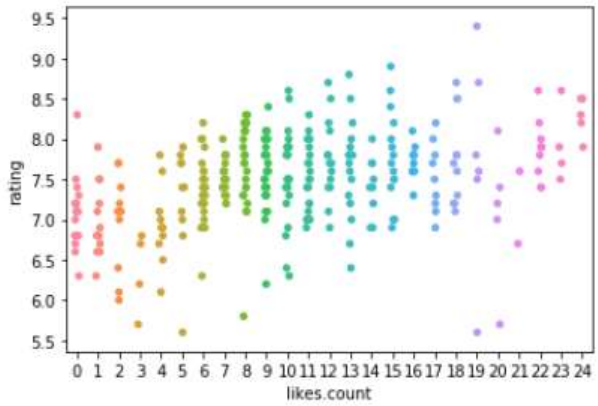
3.4.3 Conclusion about Manhattan and Old City of Toronto

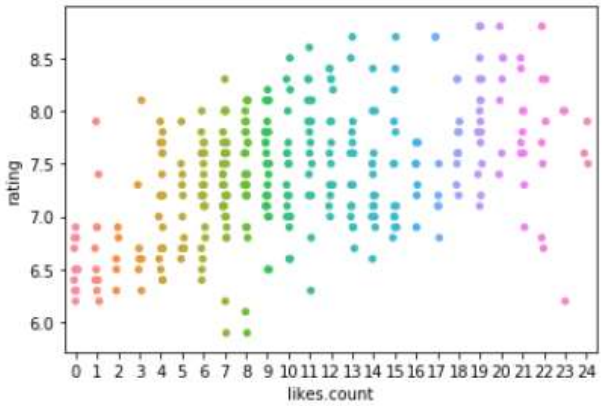
- Similarities: Both the locations have the same trend on user likes vs venue rating as observed in the scatterplot.
- Dissimilarities: American restaurant venues has higher rating in Manhattan and lower rating at the Old City of Toronto.

3.5 Correlation between Rating, User Like and Distance

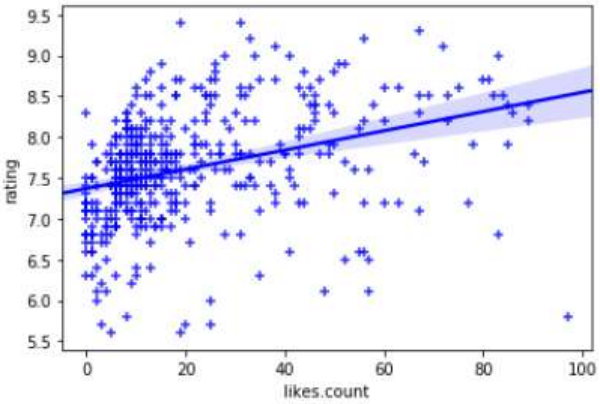
New York City | **Toronto City**

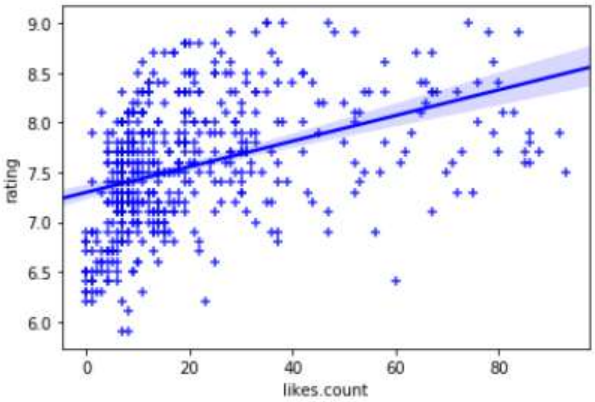
1. **Rating vs User Likes scale up to 25**



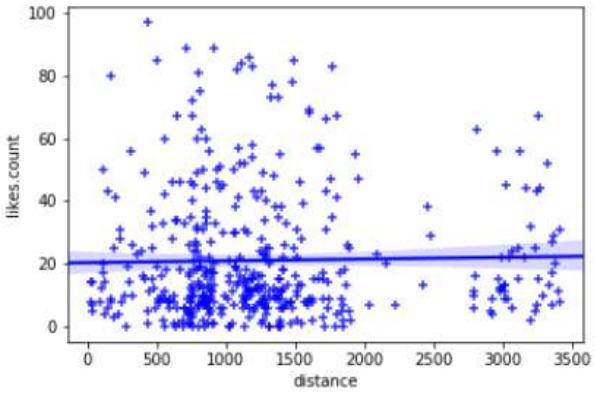


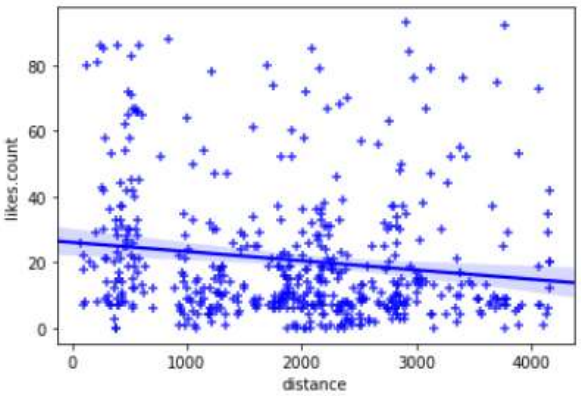
2. **Rating vs User likes scale up to 100**





3. **Distance vs User Likes**



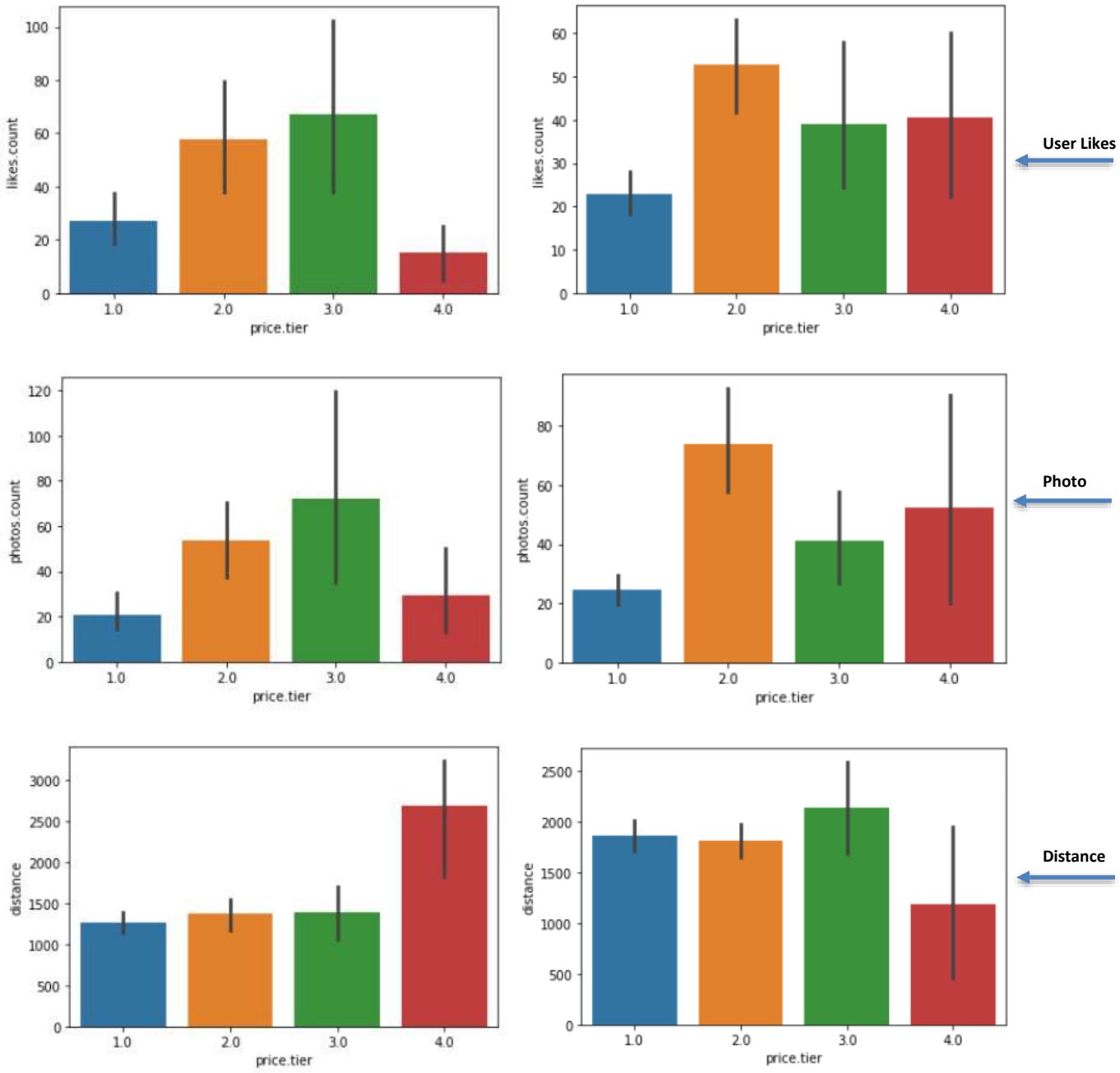


3.5.1 Observation:

The above regression plot indicates, there is a Positive relation between User Likes and Rating for both the cities and there is Negative relation between Distance and User likes.

3.6 Bar Plot analysis on Price Tier vs User Like, Photos and Distance

New York City | **Toronto City**



3.6.1 Observation on Price Tier vs User likes, Photo and Distance.

- NYC user likes are higher for price tier 3.0. Toronto user likes are higher for price tier 3.0
- NYC photo count are higher for price tier 3.0. Toronto user likes are higher for price tier 3.0
- NYC distance are higher for price tier 4.0. Toronto distance are higher for 3.0

3.7 K-Means : Cluster and Segmentation based on venue distance, user likes, rating, price tier, user tips and photo count.

3.7.1 Data Points Statistics

New York City

	distance	price.tier	rating	likes.count	photos.count	tips.count
count	254.000	254.000	254.000	254.000	254.000	254.000
mean	1330.476	1.575	7.692	41.559	37.323	16.035
std	830.020	0.717	0.693	77.825	72.400	24.402
min	103.000	1.000	5.600	0.000	0.000	0.000
25%	796.750	1.000	7.300	9.000	6.000	4.000
50%	1133.000	1.000	7.700	16.000	14.000	7.000
75%	1550.500	2.000	8.200	40.750	35.000	19.000
max	3411.000	4.000	9.400	590.000	568.000	170.000

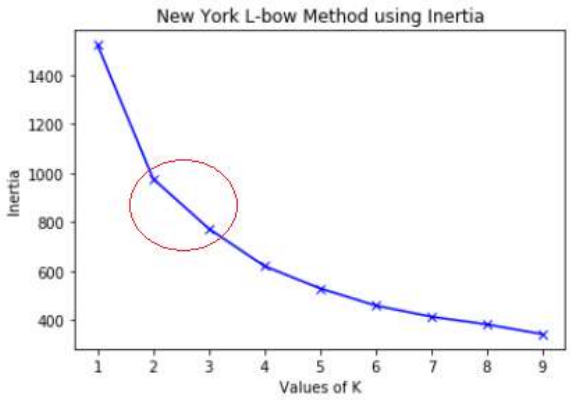
Toronto

	distance	price.tier	rating	likes.count	photos.count	tips.count
count	351.000	351.000	351.000	351.000	351.000	351.000
mean	1851.330	1.652	7.654	37.610	47.698	16.174
std	1062.199	0.716	0.668	54.085	83.337	22.773
min	106.000	1.000	5.900	0.000	0.000	0.000
25%	963.500	1.000	7.200	9.000	7.000	4.000
50%	1932.000	2.000	7.700	18.000	18.000	8.000
75%	2638.000	2.000	8.100	37.000	51.500	18.000
max	4164.000	4.000	9.200	415.000	705.000	179.000

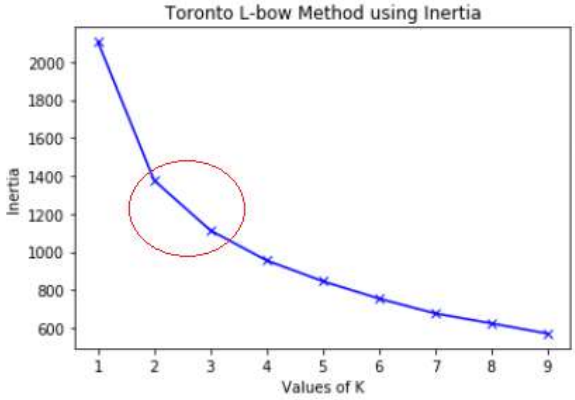
3.7.2 ELBOW method to decide number of clusters

To find the exact number of clusters using Elbow method

New York



Toronto



Above observation indicates Clusters between 2 and 3 for both the cities, we are going to set the total number of Cluster as 3 for both the cities

3.7.3 K-Means Fitting : Label count for each Clusters

New York

```
2    113
0    105
1     36
Name: Labels, dtype: int64
```

Cluster: 1-41%, 2-44% & 3-14%

Toronto

```
2    136
1    128
0     87
Name: Labels, dtype: int64
```

Cluster: 1-25%, 2-36% & 3-39%

3.7.4 Centroid Points and Cluster Visualization

K-means fitted the venues into three clusters and the venues in each cluster are similar to each other in terms of the features included in the dataset.

3.7.4.1 Centroid Points summary

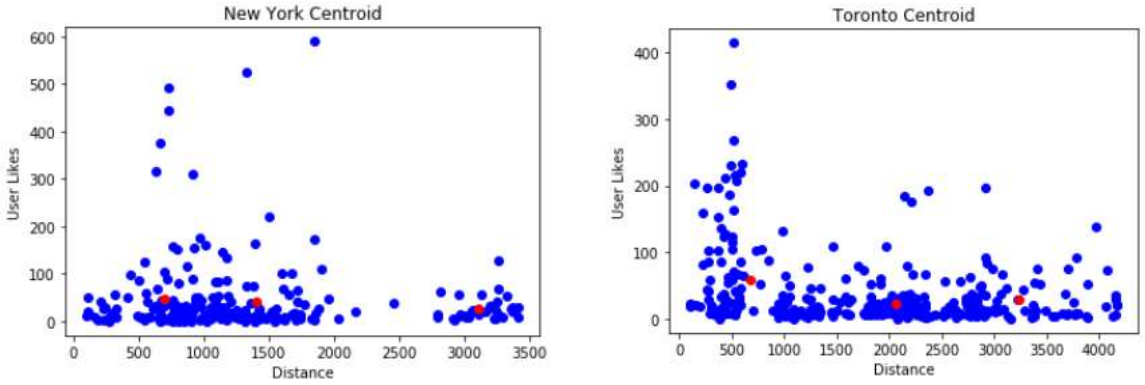
New York – Summary of the Centroid("Labels") based on mean values of the clusters

Labels	distance	price.tier	rating	likes.count	photos.count	tips.count
0	1403.009524	1.571429	7.720952	40.590476	33.114286	12.885714
1	3105.750000	1.750000	7.691667	26.305556	25.333333	15.527778
2	697.504425	1.522124	7.664602	47.318584	45.053097	19.123894

Toronto– Summary of the Centroid("Labels") based on mean values of the clusters

Labels	distance	price.tier	rating	likes.count	photos.count	tips.count
0	3236.586207	1.770115	7.616092	28.229885	36.758621	12.977011
1	679.898438	1.726562	7.808594	58.687500	70.218750	24.039062
2	2067.698529	1.507353	7.532353	23.772059	33.500000	10.816176

3.7.4.2 Visualization of Centroids for two Cities



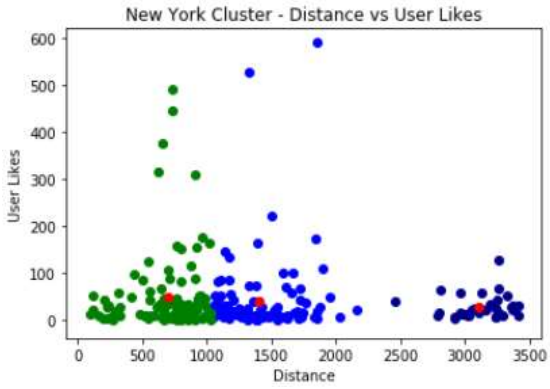
Observation:

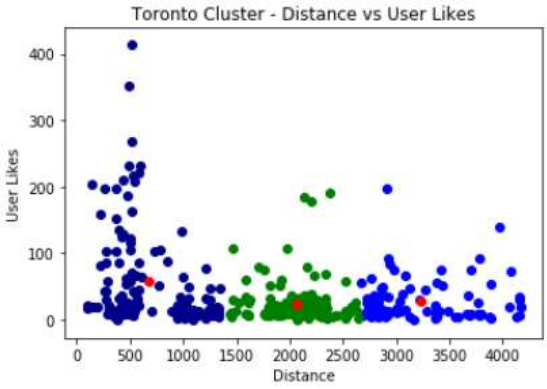
- The 1st centroid placed at the distance 1403 for NYC and 3237 for Toronto.
- The 2nd centroid placed at the distance 3106 for NYC and 680 for Toronto.
- The 3rd centroid placed at the distance for 697 NYC and 2068 for Toronto.

3.7.4.3 Visualization of Centroids and Clusters

Assigned each element to the nearest centroid

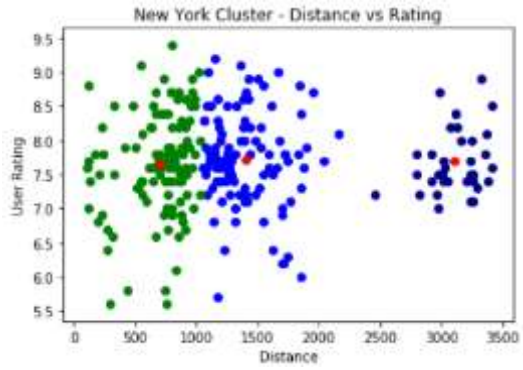
(a) Distance vs User Likes

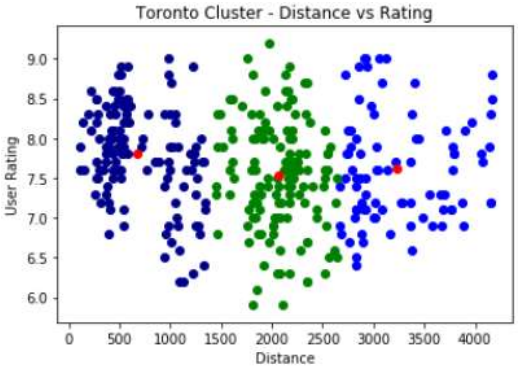




Observation:
The maximum user likes spread lies between the range 0 to 200 for both the cities.

(b) Distance vs Rating





Observation:
The maximum rating spread lies between the range 7.0 to 8.5 closer to the centroid for both, the cities.

3.8 Visualization of Maps of the Two Cities using Cluster Label Latitude and Longitude

3.8.1 New York City: K-Means Before and After Clustering of Venues

Before

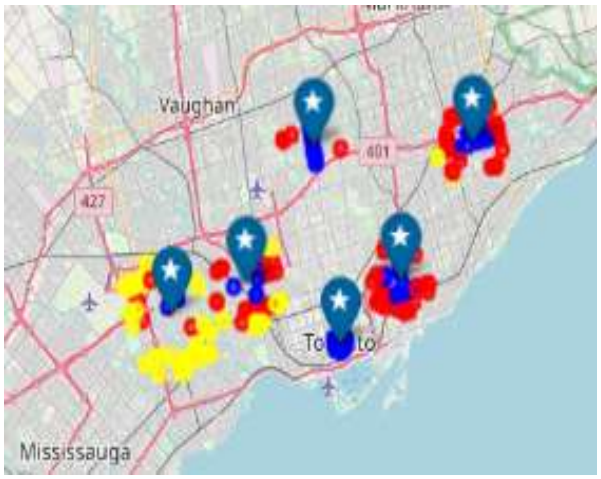
After

Clusters: 1-Red 2-Blue 3-Yellow

3.8.2 City of Toronto: K-Means Before and After Clustering of Venues

Before

After



Clusters: 1-Red 2-Blue 3-Yellow

4. Results

The summary of the exploratory analysis of the two cities are plotted below based on the data source and methods applied.

- New York City has high neighborhoods which is 47% higher compared to Toronto City.
- Queens borough contributes 26% with 86 neighborhoods in New York and old city of Toronto contributes 37% with 64 neighborhoods in Toronto city.
- New York Venue has 48.8K user likes higher by 47% compared to Toronto 26K user likes.
- The major contributors of user likes are Manhattan with 73% and old city of Toronto 59%.
- Compared to Toronto, New York Venue User Tips higher by 12% and Photo count by 45%.
- The average rating and price tier are Similar for both the cities with marginal difference.
- Top 5 Venue category: Pizza place, Coffee shop, Italian & Caribbean restaurant are the top 5 for New York. Coffee shop, Bakery, Chinese restaurant and Grocery stores are top 5 for Toronto.
- Venue Price Tier vs Rating analysis for Old City of Toronto and Manhattan
    - Bar, vegetarian restaurant, steakhouse top the list for Old city of Toronto
    - American restaurant, café, Burger Joint top the list of Manhattan
- Similarities for both the cities, where there is a positive correlation between venue user likes vs rating and negative correlation between distance vs user likes.
- New York user likes and photo count are higher for price tier 3.0 and for Toronto its 2.0
- Clustering of Venues created based on user likes, price tier, rating, distance, photo count and the number of clusters are 3 which is similar for both the cities using Elbow method.
- The percentage of venues grouped for each cluster are:
    - New York: Cluster 1 at 41%, 2 at 44% and 3 at 14%
    - Toronto: Cluster 1 at 25%, 2 at 36% and 3 at 39%
- The centroid mean value for price tier is similar for both cities with value 1.5 and 1.7 and the mean value for ratings are similar for both cities with the range 7.5 to 7.8
- The clustering w.r.t user likes and distance, the maximum user likes found between the range 0 to 200 for both cities
- Similarly, the clustering w.r.t rating and distance, the maximum rating found between the range 7.0 to 8.5 for both the cities.

5. Discussion

The exploratory analysis of the venues at the neighborhoods of New York City and City of Toronto and basis the results, we have observed the significance of venue user likes, photo count and user tips, which will be useful for venue owners to manage and increase their venue ratings and also segmentation of the venues into three clusters based on these significant parameters will improve user experience at the venues.

The descriptive analysis of the venue categories location wise shows, how the top categories of the venues distributed across locations and the findings provides a clear indicator to recommend business owners to startup new venues where there are no venues at certain locations and categories for both the cities.

In comparing the venues of the two cities, Manhattan, New York city and Old City of Toronto, Toronto are the major stakeholders, Manhattan with 73% user likes and Old City of Toronto with 59% user likes and both the location with an average rating at 8.1 and average price tier at 1.86 will be the highlight to their respective cities.

In the bar plot visualization, we have observed user likes and photo count are in higher side for the venues having price tier 3.0 at New York City and price tier 2.0 at City of Toronto. We strongly recommended the venue owners to revisit their business models on the price tiers to increase user likes, photo count thereby increasing the business. Venue owners at New York City to revisit price tier 1.0 and 4.0 and Toronto city venue owners to revisit price tier 1.0 (Low user likes).

And we have also observed similarities in our regression plots on user likes and ratings for both the cities. The visualization shows positive correlation, when the user likes increase’s the rating also is in increasing trend for both the cities and also visualized a negative correlation in terms of user likes and distance for both the cities.

In clustering of venues, we have observed the maximum user likes found between the range 0 to 200 for both the cities w.r.t user likes and distance and also maximum rating found between range 7.0 to 8.5 w.r.t rating and distance for both the cities, which indicates the centroids of the cluster may change whenever there is a change in the user likes at the venues.

6. Conclusion

In this exploration, we have analyzed the effectiveness of venue user likes, photo count, user tips, distance and rating. We have compared the venues at both the cities and analysis revealed strong similarities on the venue user likes and ratings for the both the cities and dissimilarities when compared with distance and user likes.

The descriptive analysis suggested the opportunities as how new business owner can startup new venues and the clustering of venues showed how users can choose venues of their preference to improve user experience at the venues.

The visualization of regression plots showed us a positive correlation with user likes and rating. The observation of various analysis showed the significance of user likes which is an important metric in rating the venues.

We have analyzed the venues of both the cities at various parameters, there are similarities and dissimilarities observed at both the cities. The venues of New York City when comparted to Toronto is higher at certain aspects of our analysis, basis the analysis we conclude New York City Venues are better than Toronto City.

The exploration and analysis have been performed based on Sample Venue data. Regardless of the above results and conclusions, we recommend the future analyst to explore and analyze the entire population of the Venue data to get more insights, more positive outcomes and to bring significant improvements to the venue owners and as well as to the users.