Coursera IBM Professional Data Science Certificate Course

Capstone Project – The Battle of Neighborhoods

Table of Content

A. Introduction/Business Problem

In module 3 of this Capstone Course, New York City in the United States and the city of Toronto in Canada were segmented and their neighborhoods were clustered. The two cities feature great diversity and are booming financial centers of their respective countries.

The idea here in this research is to compare the neighborhoods that are within the 16 boroughs (11 boroughs with hundreds of neighborhoods in Toronto, 5 boroughs with hundreds of neighborhoods in New York) of the two cities and find out, based on statistical techniques how similar or dissimilar they are. Are the boroughs and their associated neighborhoods make New York City like Toronto in terms of the venues, and categories of venues they harbor?

This idea will be refined here, through the use of statistical techniques that can pinpoint similarities and differences between New York and Toronto neighborhoods. This research will put emphasis on the use of data mining and machine learning techniques – Data Visualization, Analysis Of Variance (ANOVA), Correlation Analysis, Correspondence/Discriminant Analysis, and Canonical Correlation Analysis to gain insights.

In most part of the world, nowadays and with the advent of the Internet and smartphones, people/entities will to some extent, do their due diligence and research neighborhoods areas when planning to move from City A to City B. These stakeholders comprising of tourists, city migrant workers entities looking for new location to start office, and many others, will benefit from a user friendly information system that highlights to them similarities and differences between the two cities in terms of neighborhood venue profiles. Hence they will be able to make an informed choice beforehand and avoid costly moves.

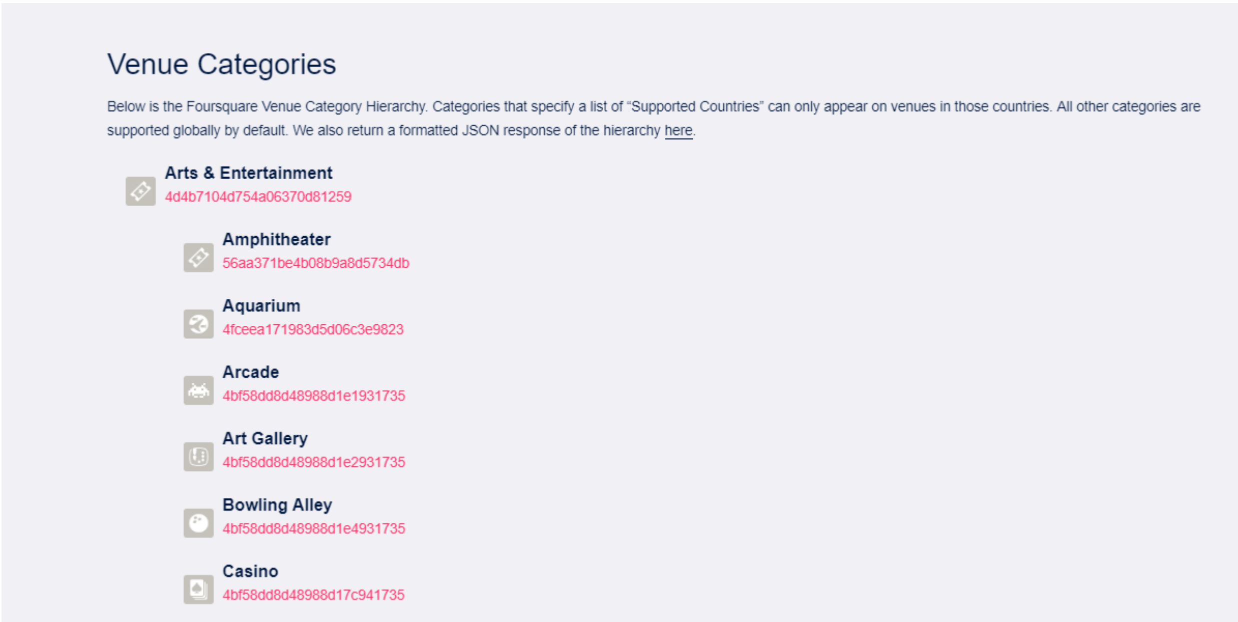
B. Data

In order to solve this problem – of clearly describing the similarities and differences between the two cities in terms of neighborhood venues – we will mainly use the Foursquare API location data.

For each City and borough/neighborhood, we will need a data set of the their venues and venue categories – borough/neighborhood name, location (latitude, longitude), venue name, venue category – shop/services, nightlife, arts, college/universities, etc.

The Venue Categories are presented in a hierarchy as suggested by the following picture from the Foursquare website.

Foursquare Venue Category Hierarchy



There are 10 top level categories with their ID:

- Arts & Entertainment (4d4b7104d754a06370d81259),
- College & University (4d4b7105d754a06372d81259),
- Food (4d4b7105d754a06374d81259),
- Professional & Other Places (4d4b7105d754a06375d81259),
- Nightlife Spot (4d4b7105d754a06376d81259),
- Outdoors & Recreation (4d4b7105d754a06377d81259),
- Shop & Service (4d4b7105d754a06378d81259),
- Travel & Transport (4d4b7105d754a06379d81259),
- Residence (4e67e38e036454776db1fb3a),
- Event (4d4b7105d754a06373d81259).

We did not make use of the Event category.

Hence, this research project would use Foursquare API as its core data gathering source as it has a database of more than 100 million places; the Foursquare API provides the ability to perform location search, venue searches by various venue categories through HTTP requests. Because of http requests limitations, the number of places per neighborhood will be limited to 100 and the radius parameter to 500.

Using our API credentials and the top level category Ids, we retrieved the basic data tables from the database.

The following two types of tables were used to create the all tables used to plot on maps or as inputs to the machine learning algorithms.

The Notebooks are (NewYorkVenueCat.ipynb and TorontoVenueCat.ipynb)

New York City Neighborhoods and Venues by College/University Venue Category

Launcher | NewYorkVenueCat.ipynb | NewYorkVenueColU.csv | TorontoVenueCat.ipynb | StripNYDataset.ipynb | NT9VenueCat408Neighb.ipy

Delimiter: ,

		Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	0	Wakefield	40.89470517661	-73.84720052054902	mt cernon chamber of commerce	40.8942	-73.8454	College Rec Center
2	1	Wakefield	40.89470517661	-73.84720052054902	KNIGHTmare Gym	40.891589869109815	-73.8440744271743	College Gym
3	2	Co-op City	40.87429419303012	-73.82993910812398	G.B. Wise	40.87808	-73.832621	Student Center
4	3	Eastchester	40.887555677350775	-73.82780644716412	Nativity of Our Lady School	40.888728179458404	-73.83202473417445	General College & University
5	4	Eastchester	40.887555677350775	-73.82780644716412	Cornerstone Academy PS/IS 189	40.88270358897625	-73.83109426973273	Student Center
6	5	Fieldston	40.89543742690383	-73.90564259591682	Quad	40.89042309417936	-73.9062203725553	College Quad
7	6	Fieldston	40.89543742690383	-73.90564259591682	Kleinman	40.88991812990233	-73.90614334179558	Fraternity House
8	7	Fieldston	40.89543742690383	-73.90564259591682	J3	40.890485	-73.902294	Sorority House
9	8	Fieldston	40.89543742690383	-73.90564259591682	Jasper Hall	40.890477143573975	-73.90245592896831	College Residence Hall
10	9	Fieldston	40.89543742690383	-73.90564259591682	J4	40.890482530256946	-73.90250622474919	College Residence Hall
11	10	Riverdale	40.890834493891305	-73.9125854610857	Yeshivat Chovevei Torah Rabbinical School	40.888189312034626	-73.9107731087732	General College & University
12	11	Riverdale	40.890834493891305	-73.9125854610857	The David A. Stein Riverdale/Kingsbridge Academy	40.888363348134824	-73.91468065496096	General College & University
13	12	Riverdale	40.890834493891305	-73.9125854610857	Meyers Physics	40.889622	-73.906686	College Lab
14	13	Riverdale	40.890834493891305	-73.9125854610857	Saint Gabriel's school	40.885544035530174	-73.91136796927037	College Academic Building
15	14	Riverdale	40.890834493891305	-73.9125854610857	Smith Hall	40.88956174815586	-73.90885327595306	College Auditorium
16	15	Kingsbridge	40.88168737120521	-73.90281798724604	Manhattan College bookstore	40.88214691666667	-73.90271751666667	College Bookstore
17	16	Kingsbridge	40.88168737120521	-73.90281798724604	MC Dept. ChemE Conference Room	40.88456273674023	-73.90053388806048	College Engineering Building
18	17	Kingsbridge	40.88168737120521	-73.90281798724604	Columbia University Medical	40.876950291469946	-73.90160182106129	Medical School
19	18	Kingsbridge	40.88168737120521	-73.90281798724604	U.S.K. Tae Kwon Do	40.87925374096617	-73.90461841311556	College Gym
20	19	Kingsbridge	40.88168737120521	-73.90281798724604	Kingsbridge School P.S. 7	40.88109742034722	-73.9054746333788	College Classroom
21	20	Kingsbridge	40.88168737120521	-73.90281798724604	Spuyten Duyvil Preschool	40.879244392033605	-73.90720545790641	University

Toronto Neighborhoods and Venues by College/University Venue Category

Launcher | NewYorkVenueCat.ipynb | NewYorkVenueColU.csv | TorontoVenueColU.csv | TorontoVenueCat.ipynb | StripNYDataset.ipynb | NT9VenueCat408

Delimiter: ,

		Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	0	Parkwoods	43.7532586	-79.3296565	Ranchdale	43.75231914332775	-79.32344787202561	College Classroom
2	1	Victoria Village	43.7258823	-79.31557159999998	Toronto Fire Services Operations Training Centre	43.723972237085476	-79.31712598976469	Trade School
3	2	Harbourfront,Regent Park	43.6542599	-79.3606359	The George	43.653245935894304	-79.35722189473297	College Residence Hall
4	3	Harbourfront,Regent Park	43.6542599	-79.3606359	Charles MacPherson Associates	43.654681262392735	-79.35920876327621	Trade School
5	4	Harbourfront,Regent Park	43.6542599	-79.3606359	George Brown College - School of ESL	43.65187195086978	-79.36557973642425	College Academic Building
6	5	Harbourfront,Regent Park	43.6542599	-79.3606359	George Brown School of Design	43.651871139300646	-79.36579671873575	College Technology Building
7	6	Harbourfront,Regent Park	43.6542599	-79.3606359	George Brown College - SJG Building	43.65188773426242	-79.36557437967805	College Academic Building
8	7	Harbourfront,Regent Park	43.6542599	-79.3606359	TAIE International Institute	43.659135	-79.365487	College Academic Building
9	8	Harbourfront,Regent Park	43.6542599	-79.3606359	George Brown School Of Design	43.65189490212039	-79.36560093952392	College Technology Building
10	9	Harbourfront,Regent Park	43.6542599	-79.3606359	George Brown College Theatre School	43.65077615191524	-79.35758081681655	College Theater
11	10	Lawrence Heights,Lawrence Manor	43.718518	-79.46476329999999	Kumon	43.71784327668243	-79.46250823846198	Student Center
12	11	Lawrence Heights,Lawrence Manor	43.718518	-79.46476329999999	Bluenotes' Head Office	43.718053735782526	-79.46320043192486	Fraternity House
13	12	Lawrence Heights,Lawrence Manor	43.718518	-79.46476329999999	Yorkdale Adult Learning Centre	43.71678648144151	-79.45860768569246	Student Center
14	13	Queen's Park	43.6623015	-79.3894938	University of Toronto	43.6624934706167	-79.39521976633822	University
15	14	Queen's Park	43.6623015	-79.3894938	Banting Institute	43.660300526777704	-79.38809455925548	College Science Building
16	15	Queen's Park	43.6623015	-79.3894938	Best Institute	43.660409308489015	-79.38941231620524	College Science Building
17	16	Queen's Park	43.6623015	-79.3894938	Regis College Library	43.66378714912873	-79.39072809094115	College Library
18	17	Queen's Park	43.6623015	-79.3894938	St Josephs College Secondary School	43.6642585579444	-79.38868089155856	High School
19	18	Queen's Park	43.6623015	-79.3894938	Student Kitchen	43.665841	-79.390135	College Cafeteria
20	19	Queen's Park	43.6623015	-79.3894938	Teefy Hall	43.665299046033205	-79.39113842105318	College Arts Building

By judiciously combining these types of tables for all the high levels venue categories, the following tables were created for New York and Toronto:

New York City boroughs, neighborhoods and venue top level venue categories

	Borough	Neighborhood	Latitude	Longitude	Arts	CollegeUniversity	Food	Professional	Nightlife	Outdoor	ShopServices	TravelTransport	Residence
0	Bronx	Wakefield	40.894705	-73.847201	4.0	2.0	19.0	21.0	2.0	4.0	32.0	5.0	1.0
1	Bronx	Co-op City	40.874294	-73.829939	4.0	1.0	25.0	42.0	3.0	14.0	42.0	21.0	13.0
2	Bronx	Eastchester	40.887556	-73.827806	1.0	2.0	24.0	30.0	3.0	5.0	42.0	21.0	2.0
3	Bronx	Fieldston	40.895437	-73.905643	4.0	5.0	3.0	25.0	2.0	12.0	9.0	1.0	4.0
4	Bronx	Riverdale	40.890834	-73.912585	4.0	5.0	35.0	41.0	3.0	32.0	39.0	5.0	25.0
5	Bronx	Kingsbridge	40.881687	-73.902818	8.0	10.0	50.0	49.0	27.0	40.0	50.0	44.0	26.0
6	Manhattan	Marble Hill	40.876551	-73.910660	8.0	7.0	48.0	47.0	9.0	36.0	50.0	44.0	23.0
7	Bronx	Woodlawn	40.898273	-73.867315	8.0	2.0	43.0	41.0	30.0	17.0	46.0	19.0	10.0
8	Bronx	Norwood	40.877224	-73.879391	6.0	4.0	49.0	48.0	4.0	35.0	48.0	34.0	25.0
9	Bronx	Williamsbridge	40.881039	-73.857446	6.0	7.0	37.0	38.0	12.0	9.0	34.0	8.0	5.0
10	Bronx	Baychester	40.866858	-73.835798	5.0	0.0	24.0	22.0	4.0	12.0	43.0	20.0	2.0
11	Bronx	Pelham Parkway	40.857413	-73.854756	4.0	1.0	37.0	44.0	2.0	12.0	30.0	7.0	6.0
12	Bronx	City Island	40.847247	-73.786488	4.0	0.0	19.0	13.0	7.0	25.0	27.0	12.0	3.0
13	Bronx	Bedford Park	40.870185	-73.885512	6.0	7.0	45.0	48.0	7.0	23.0	47.0	42.0	36.0
14	Bronx	University Heights	40.855727	-73.910416	4.0	42.0	42.0	44.0	11.0	8.0	45.0	18.0	19.0
15	Bronx	Morris Heights	40.847898	-73.919672	5.0	6.0	42.0	44.0	8.0	24.0	40.0	17.0	25.0
16	Bronx	Fordham	40.860997	-73.896427	13.0	40.0	50.0	49.0	8.0	32.0	50.0	46.0	32.0
17	Bronx	East Tremont	40.842696	-73.887356	4.0	5.0	41.0	49.0	7.0	17.0	47.0	16.0	26.0
18	Bronx	West Farms	40.839475	-73.877745	10.0	4.0	42.0	43.0	8.0	23.0	42.0	33.0	19.0

The Notebook is (NewYorkVenueCat.ipynb)

Toronto boroughs, neighborhoods and venue top level venue categories

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Arts	CollegeUniversity	Food	Professional	Nightlife	Outdoor	ShopServices	TravelTransport	Residence
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.0	1.0	1.0	9.0	0.0	3.0	4.0	4.0	4.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0	1.0	5.0	19.0	0.0	5.0	15.0	4.0	7.0
2	M5A	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636	50.0	8.0	50.0	49.0	27.0	47.0	50.0	38.0	35.0
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763	4.0	3.0	12.0	28.0	2.0	6.0	39.0	1.0	0.0
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494	45.0	48.0	50.0	50.0	41.0	45.0	50.0	28.0	49.0
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242	0.0	0.0	2.0	2.0	0.0	3.0	1.0	3.0	1.0
6	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	0.0	0.0	2.0	20.0	0.0	3.0	12.0	0.0	0.0
7	M3B	North York	Don Mills North	43.745906	-79.352188	3.0	5.0	8.0	43.0	0.0	13.0	8.0	1.0	0.0
8	M4B	East York	Woodbine Gardens,Parkview Hill	43.706397	-79.309937	2.0	0.0	22.0	18.0	4.0	7.0	26.0	7.0	2.0
9	M5B	Downtown Toronto	Ryerson,Garden District	43.657162	-79.378937	47.0	49.0	50.0	50.0	47.0	46.0	50.0	44.0	49.0
10	M6B	North York	Glencairn	43.709577	-79.445073	3.0	0.0	22.0	9.0	4.0	6.0	25.0	3.0	2.0
11	M9B	Etobicoke	Cloverdale,Islington,Martin Grove,Princess Gar...	43.650943	-79.554724	0.0	0.0	7.0	6.0	0.0	2.0	5.0	0.0	1.0
12	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	3.0	1.0	3.0	2.0	1.0	0.0	5.0	2.0	1.0
13	M3C	North York	Flemingdon Park,Don Mills South	43.725900	-79.340923	3.0	6.0	31.0	45.0	1.0	10.0	29.0	3.0	1.0
14	M4C	East York	Woodbine Heights	43.695344	-79.318389	2.0	1.0	8.0	20.0	4.0	10.0	21.0	11.0	0.0
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	47.0	49.0	50.0	50.0	50.0	43.0	50.0	40.0	37.0
16	M6C	York	Humewood-Cedarvale	43.693781	-79.428191	2.0	2.0	1.0	13.0	2.0	15.0	1.0	2.0	7.0
17	M9C	Etobicoke	Bloordale Gardens,Eringate,Markland Wood,Old B...	43.643515	-79.577201	3.0	0.0	4.0	11.0	4.0	4.0	23.0	4.0	0.0
18	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	0.0	0.0	10.0	17.0	0.0	7.0	23.0	6.0	4.0

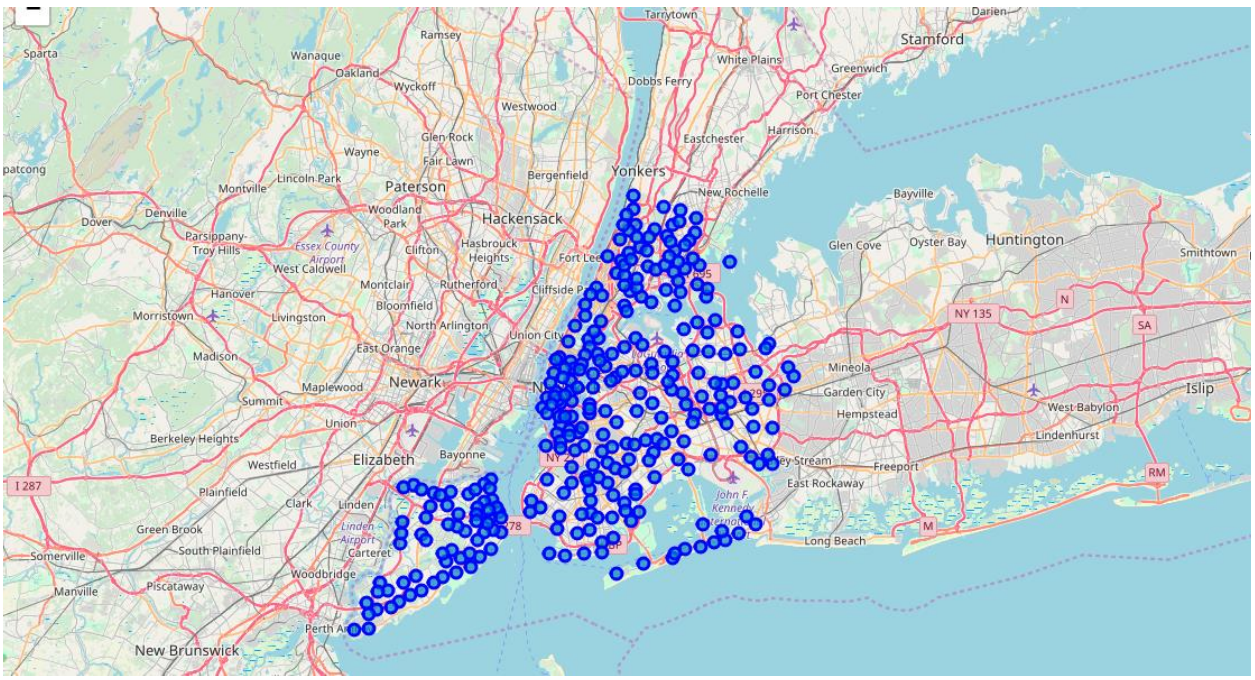
The Notebook is (TorontoVenueCat.ipynb)

By combining and merging these previous two tables, we obtained all the remaining tables needed.

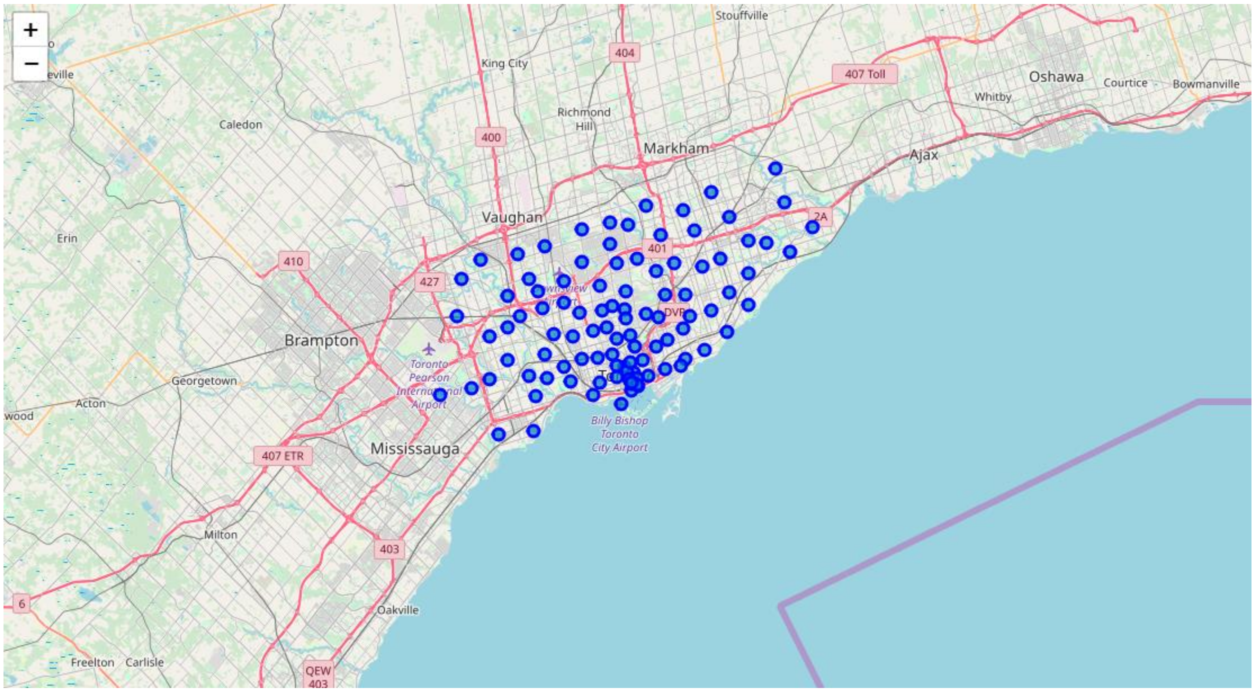
A number of Folium - Python visualization libraries, including libraries like Pandas, NumPy, Seaborn, Scikit-learn, Matplotlib, and others will be used for visualization and comparative analysis.

Using Folium, for instance, we plotted neighborhoods and venues on maps of New York and Toronto as illustrated below (same Notebooks as above):

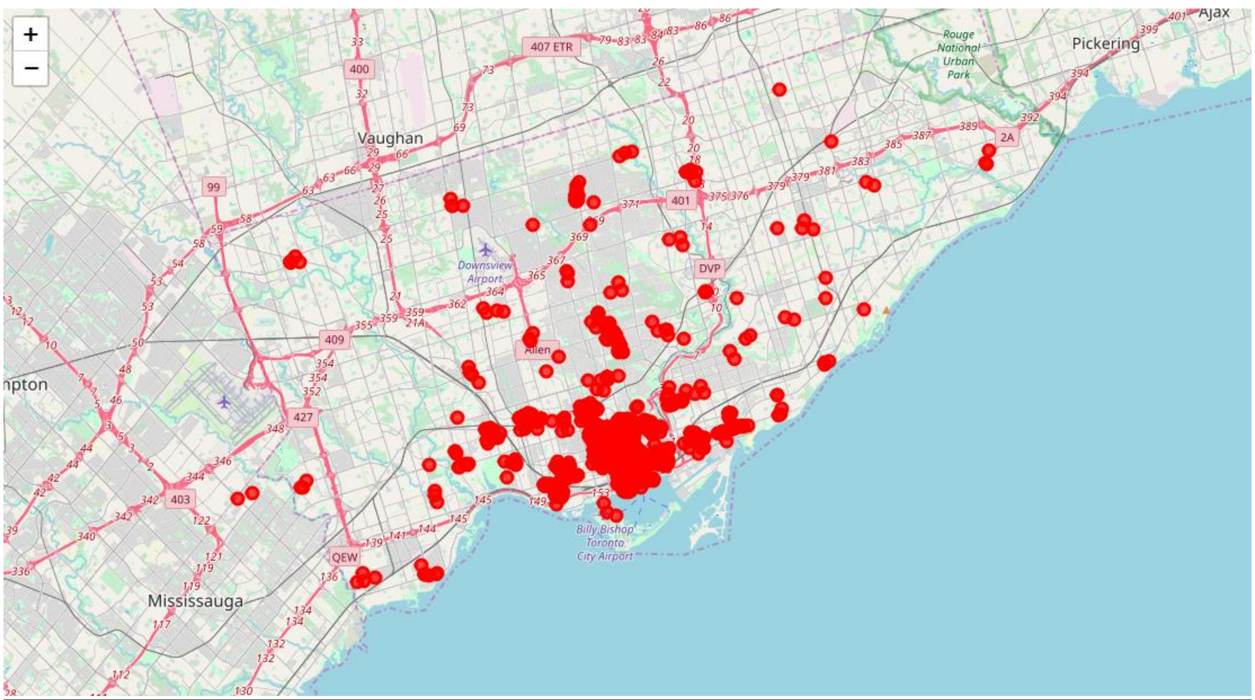
New York City neighborhoods (306)



Toronto neighborhoods (102)



Toronto Arts & Entertainment venue category



C. Methodology

In this section we discuss and describe exploratory data analysis (EDA) that was conducted, the inferential statistical testing performed, what machine learning were used and why.

C1. Eda and Inferential statistical tests

As already mentioned, statistical techniques have been used to look at similarities and differences between venues located within neighborhoods of boroughs of New York City and Toronto. These techniques require a number of data tables in the proper structure as inputs to their various algorithms.

We did Eda at a high level looking at averages, plotting boxplots and histograms, as well as performing Analysis of Variance (ANOVA) statistical tests. The table used had the following structure:

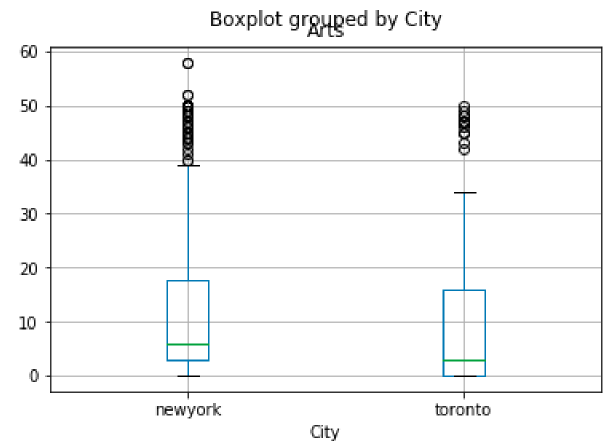
	Borough	City	Arts	CollegeUniversity	Food	Professional	Nightlife	Outdoor	ShopServices	TravelTransport	Residence
0	Central Toronto	toronto	66	59	200	227	60	145	210	57	145
1	Downtown Toronto	toronto	694	475	780	777	670	737	776	665	606
2	East Toronto	toronto	79	10	212	181	75	155	225	71	25
3	East York	toronto	13	7	123	132	20	48	141	36	16
4	Etobicoke	toronto	42	6	120	141	18	66	163	23	16

These are the numbers of venues in New York and Toronto by top level venue categories.

For instance this is the result of box plot with ANOVA statistical test results; it shoes that there is no statistical difference [PR (> F) is greater than 0.05] in average numbers of Arts & Entertainment venue categories for the venues in New York and Toronto.

```
# boxplot for Arts column
df2.boxplot('Arts', by='City')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f736415a278>
```

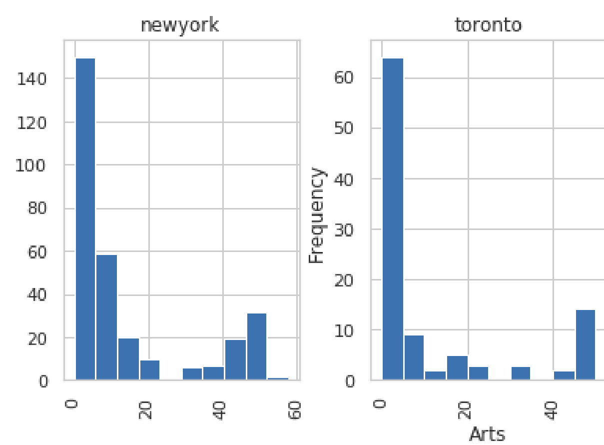



```
modl2 = ols('Arts ~ City', data=df2).fit()
anova_table = sm.stats.anova_lm(modl2,typ=2)
print(anova_table)
```

	sum_sq	df	F	PR(>F)
City	697.529412	1.0	2.394941	0.122507
Residual	118247.960784	406.0	NaN	NaN

The histogram plot is as follows:

```
# plotting histograms
df2['Arts'].hist(by=df2['City'])
# df2.Arts.hist()
# plt.title('Histogram of Arts')
plt.xlabel('Arts')
plt.ylabel('Frequency')
plt.savefig('hist_arts')
```



The details can be found in the following Notebook EDA_NYT.ipynt.

C2. Machine learning

We have made use of those that can best allow the study of similarities and differences between entities on which a given set of characteristics have been measured. They are:

- Principal Component Analysis (PCA) followed by kMeans Clustering
- Pairwise Plotting, Correlations, Correspondence Analysis (CA), and Canonical Correlations (CCA)

PCA is a data reduction techniques that can be used to project the multidimensional data on a 2D space and hence will allow more relevant visualization of the variability observed within the cloud of data. Once the data is project ted in a 2D space it can then be clustered using kMeans to see what the different groups of observations are and how they are similar or different.

Correlations being indicators of distance –resemblance - will also show relationships between entities (venues, neighborhoods, boroughs). CA will show proximity between entities in a manner similar to PCA, but in a more comprehensive way since it simultaneously visualizes both entities and their features on a 2D plane. CCA is a generalization of the General Linear Model in the sense that it performs regression analysis of a group of dependent variables against a group of independent variables; for instance the 5 boroughs in New York City versus the 11 boroughs in Toronto.

Principal Component Analysis (PCA) followed by kMeans clustering are in the Notebook AnalysisPCA_kMeans.ipynb.

Pairwise Plotting, Correlations, Correspondence Analysis (CA), and Canonical Correlations (CCA) are in the Notebook Analysis_PP_Corr_CA_CCA.ipynb.

D. Results

The findings of the research will now be presented in this section.

<u>PCA & kMeans</u>

The data table used was:

```
df = pd.DataFrame(matX)
df.head()
```

	City	Arts	CollegeUniversity	Food	Professional	Nightlife	Outdoor	ShopServices	TravelTransport	Residence
0	toronto	0	1	1	9	0	3	4	4	4
1	toronto	1	1	5	19	0	5	15	4	7
2	toronto	50	8	50	49	27	47	50	38	35
3	toronto	4	3	12	28	2	6	39	1	0
4	toronto	45	48	50	50	41	45	50	28	49

```
labels = df['City'].unique().tolist()
```

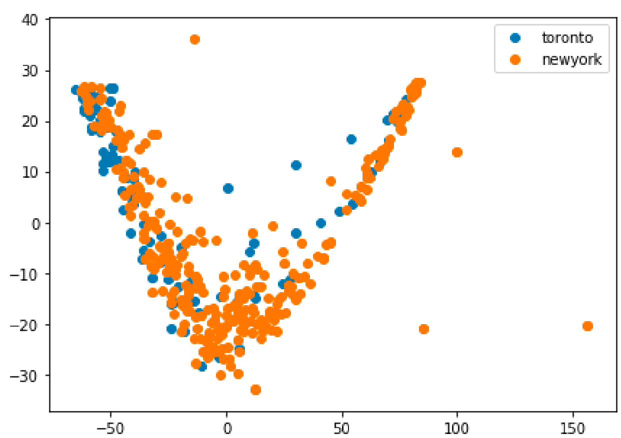
Column 'City' was used for labelling the observations (neighborhoods) and there were 9 features.

The PCA transformed the original 9 dimensions into two independent ones (2 principal components) for easy visualization. We obtained the following diagram:

```
pca = PCA(n_components=2)
pca.fit(X)
X_ = pca.transform(X)
```

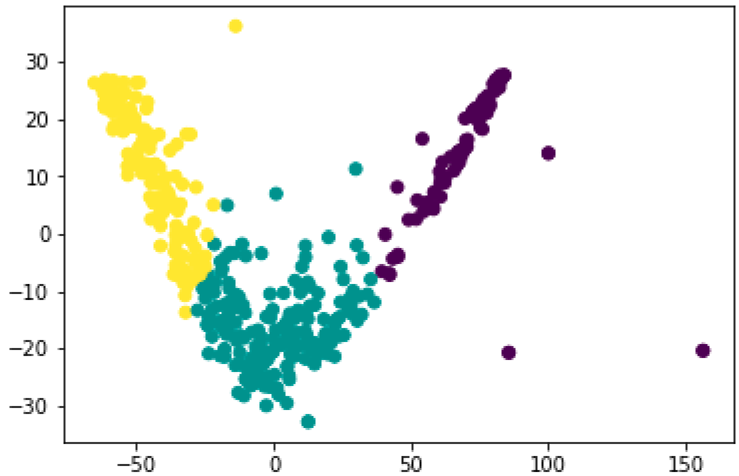
```
dfPCA = pd.DataFrame({'x1': X_[:,0], 'x2': X_[:,1]})
dfPCA['City'] = df['City']
dfPCA
dfPCA.to_csv('dfPCA.csv')
```

```
plt.figure(figsize=(7,5))
for lab in labels:
    plt.scatter(dfPCA.loc[dfPCA['City'] == lab, 'x1'],  dfPCA.loc[dfPCA['City'] == lab, 'x2'], label=lab)
    plt.legend()
```



The coordinates of the observations (venues) along the 2 principal components were then saved for kMeans clustering. Three clusters were obtained and plotted as follows:

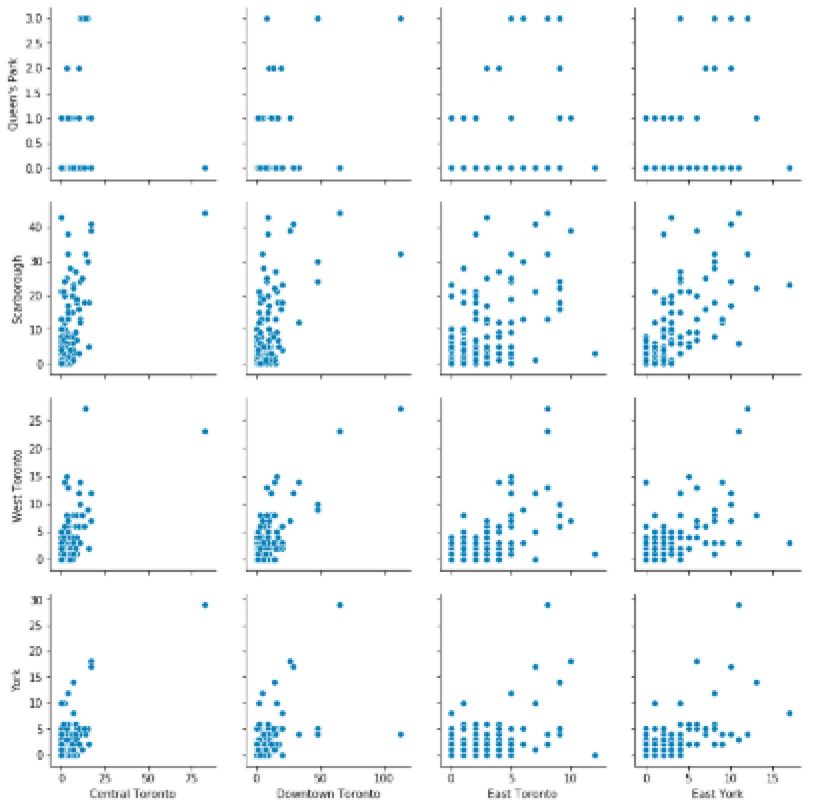
```python
import pylab as pl
pl.figure('K-means with 3 clusters')
pl.scatter(X.iloc[:, 0], X.iloc[:, 1], c=model.labels_)
pl.show()
```




```python
cluster_map = pd.DataFrame()
cluster_map['X_index'] = X.index.values
cluster_map['cluster'] = model.labels_
```

Pairwise Plotting, Correlations, Correspondence Analysis (CA), and Canonical Correlations (CCA)

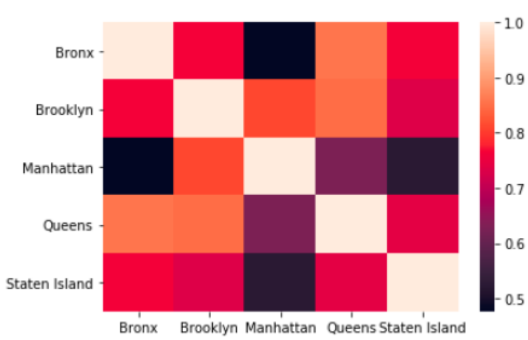
Sample pairwise plotting showing relationships between selected boroughs.



Correlations between New York boroughs and their heat map.

```
# Looking at New York borough correlations
corr = nds.corr()
print(corr)
sns.heatmap(corr)
```

```
                 Bronx  Brooklyn  Manhattan    Queens  Staten Island
Bronx         1.000000  0.765985   0.476835  0.858953       0.763439
Brooklyn      0.765985  1.000000   0.812482  0.850335       0.734550
Manhattan     0.476835  0.812482   1.000000  0.628317       0.523215
Queens        0.858953  0.850335   0.628317  1.000000       0.740938
Staten Island 0.763439  0.734550   0.523215  0.740938       1.000000
<matplotlib.axes._subplots.AxesSubplot at 0x7fa185e624e0>
```



Correlations between Toronto borough and their heat map

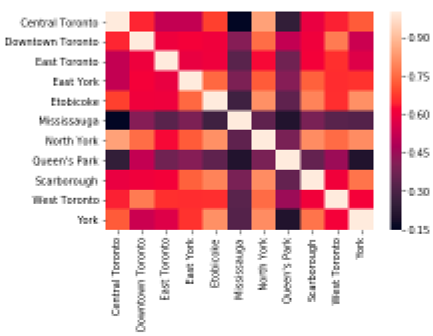
```
# Looking at Toronto borough correlations
corr = tds.corr()
print(corr)
sns.heatmap(corr)
```

	Central Toronto	Downtown Toronto	East Toronto	East York
Central Toronto	1.000000	0.646956	0.520860	0.520343
Downtown Toronto	0.646956	1.000000	0.603255	0.621222
East Toronto	0.520860	0.603255	1.000000	0.580980
East York	0.520343	0.621222	0.580980	1.000000
Etobicoke	0.686715	0.608928	0.605817	0.747816
Mississauga	0.149845	0.414821	0.329346	0.398502
North York	0.847785	0.757244	0.627380	0.727114
Queen's Park	0.245318	0.521314	0.372590	0.414068
Scarborough	0.592369	0.611005	0.615291	0.736413
West Toronto	0.638951	0.788251	0.663588	0.657248
York	0.725287	0.532110	0.578752	0.667314

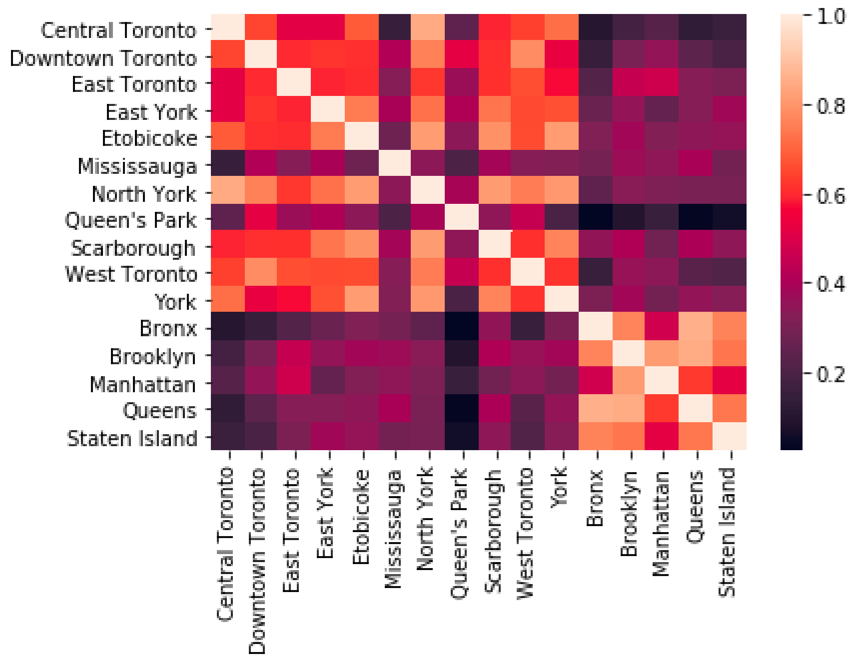
	Etobicoke	Mississauga	North York	Queen's Park
Central Toronto	0.686715	0.149845	0.847785	0.245318
Downtown Toronto	0.608928	0.414821	0.757244	0.521314
East Toronto	0.605817	0.329346	0.627380	0.372590
East York	0.747816	0.398502	0.727114	0.414068
Etobicoke	1.000000	0.278768	0.815715	0.339644
Mississauga	0.278768	1.000000	0.339720	0.205091
North York	0.815715	0.339720	1.000000	0.392011
Queen's Park	0.339644	0.205091	0.392011	1.000000
Scarborough	0.793112	0.390040	0.810292	0.340694
West Toronto	0.658620	0.330828	0.752871	0.454593
York	0.816790	0.321041	0.808690	0.201602

	Scarborough	West Toronto	York
Central Toronto	0.592369	0.638951	0.725287
Downtown Toronto	0.611005	0.788251	0.532110
East Toronto	0.615291	0.663588	0.578752
East York	0.736413	0.657248	0.667314
Etobicoke	0.793112	0.658620	0.816790
Mississauga	0.390040	0.330828	0.321041
North York	0.810292	0.752871	0.808690
Queen's Park	0.340694	0.454593	0.201602
Scarborough	1.000000	0.612667	0.767888
West Toronto	0.612667	1.000000	0.618352
York	0.767888	0.618352	1.000000

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa19812bb00>
```

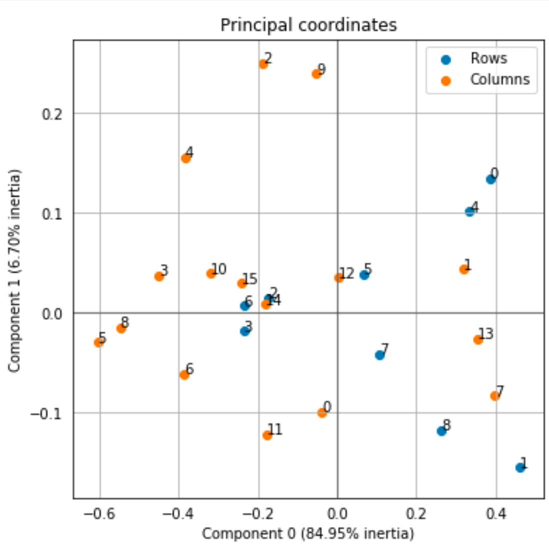


Correlation between New York and Toronto boroughs; only heat map is presented.

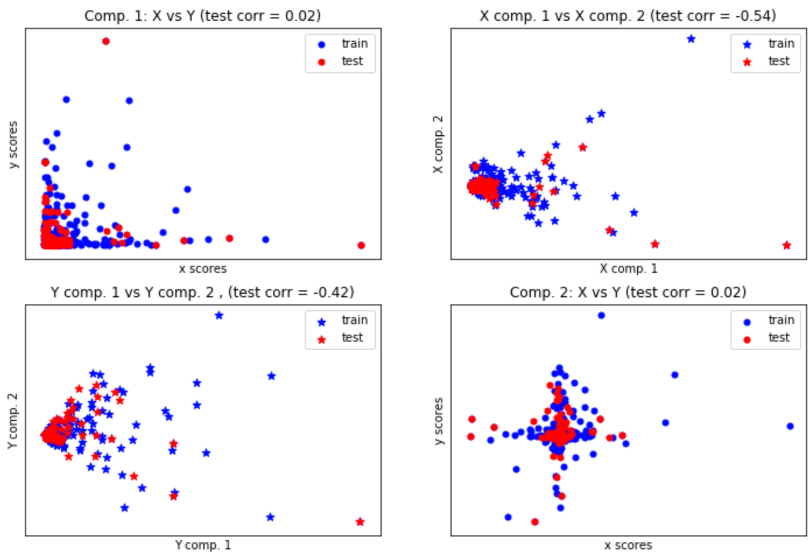


CA was performed for the 9 features (blue color) and the 16 boroughs (orange color) from New York and Toronto; the visualization on the 2 components show the proximity between boroughs and venue categories.

```
# plot both sets of principal coordinates with the plot_coordinates method.
ax = ca.plot_coordinates(X=X,ax=None,figsize=(6, 6),x_component=0,y_component=1,show_row_labels=True,show_col_labels=True)
ax.get_figure().savefig('cantg_coordinates.png')
```



CCA results are as shown.



This simply shows that there seems to be more resemblance between venues within the same city as compared with venues from outside the city.

E. Discussion

We look at observations and important points noted during the research and make some recommendations based on the results obtained.

In essence, looking at the results of the different statistical analysis techniques, one can see that neighborhoods within the same cities tend to be more similar than neighborhoods in the other city.

In Toronto, the venues in the boroughs of Mississauga and Queens' Park are quite different from the venues in the other boroughs.

In New York, venues in Manhattan and Brooklyn bear more resemblance; the same applies for Queens and Bronx together and they are all different from Staten Island.

There are more venues in East Toronto that bear resemblance with venues in Manhattan and Brooklyn.

F. Conclusion

This concludes the research findings and epilogues on lessons learned. Based on our experience with research project, we confirm that as data scientist, one spends more time – 80% -- on data preparation, data exploration than on model fitting. Also, if the statistical techniques are used properly, they tend to lead to the same conclusion no matter their different nature. This research can be extended further by trying other machine learning technique such Linear Discriminant Analysis and Multinomial Logistic Regression. In addition the CA part can be augmented by using supplementary feature and observation techniques