Coursera IBM Professional Data Science Certificate Course
Capstone Project – The Battle of Neighborhoods

Topics

INTRODUCTION/BUSINESS PROBLEM
DATA
METHODOLOGY
EDA AND INFERENTIAL STATISTICAL TESTS
MACHINE LEARNING
RESULTS
DISCUSSION
CONCLUSION

A. Introduction/Business Problem

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?

Use 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.

B. Data -- 0

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.

B. Data -- 1

Toronto

	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

B. Data -- 2

New York City

	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

C. Methodology -- 0

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

C. Methodology -- 1

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

D. Results – 0

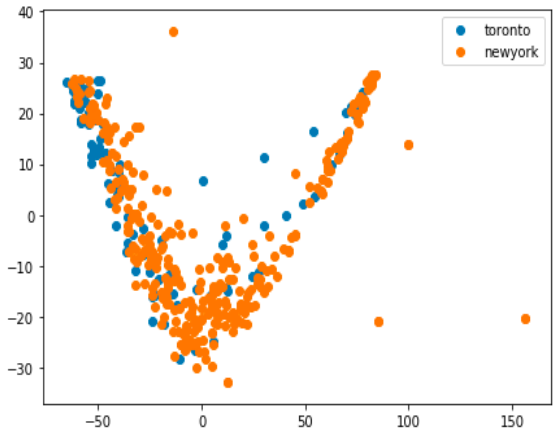
Principal Components Analysis

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')
```

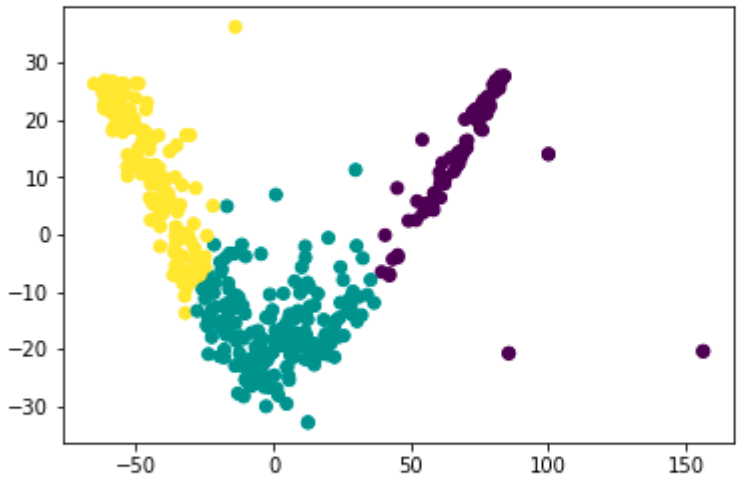
```
plt.figure(figsize=(7,5))
for lab in labels:
    plt.scatter(dfPCA.loc[dfPCA['City'] == lab, 'x1'],  dfPCA.loc[dfPCA['City'] == lab, 'x2'], label=lab)
    plt.legend()
```



D. Results – 1

kMeans Clustering

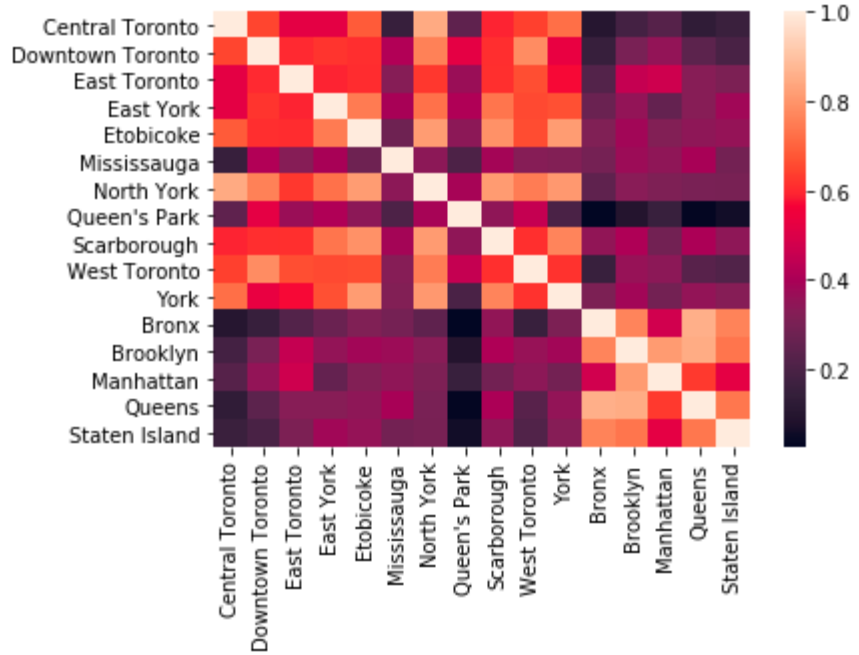
```
import pylab as pl
pl.figure('K-means with 3 clusters')
pl.scatter(X.iloc[:, 0], X.iloc[:, 1], c=model.labels_)
pl.show()
```




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

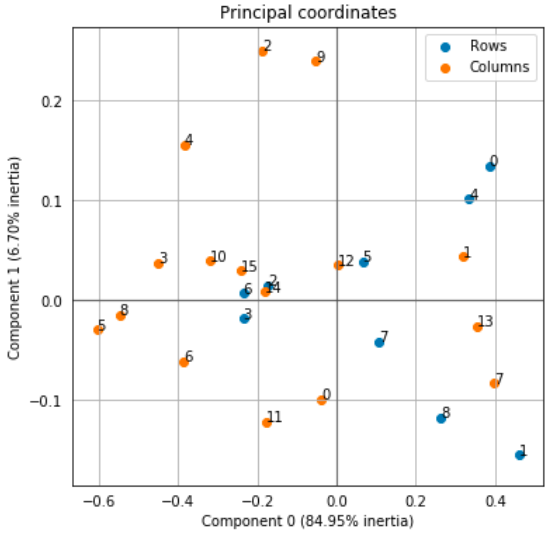
D. Results – 2

Heat map correlation



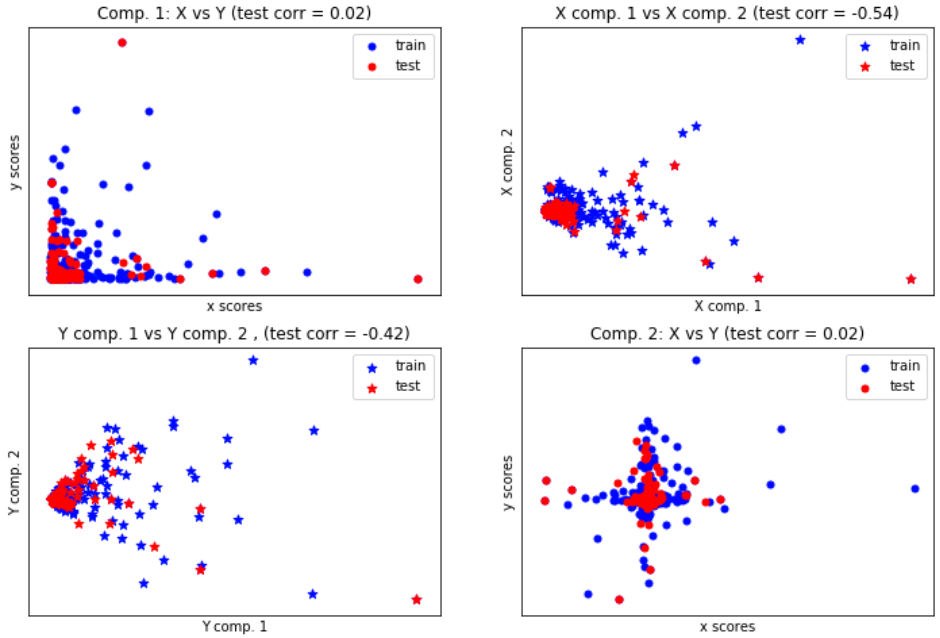
D. Results – 1
Correspondence Analysis

```
# 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')
```



D. Results – 1
Correspondence Analysis

D. Results – 1
CCA



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.