Data Mining on Geographical Data of New York and Toronto

Yinuo Zeng

December 2019

1 Introduction

Cities are getting larger and larger and are playing an important role in modern people's lives. It is worthwhile to investigate the similarity of big cities and to find the common characteristics of them. By this way, we can make use of such similarity to solve various business problem.

In this project, I will study two representative cities, New York City and City of Toronto. As big cities, both of them are multicultural and are financial capitals of their respective countries. My goal is to study how similar their neighborhoods are and try to find latent factors behind the similarity.

This project provides clients who plan to invest big cities with an reliable and accurate reference. For example, audiences who are going to open a new Italian restaurant in City of Toronto can refer the results of project. Based on neighborhoods in New York City where Italian restaurants are popular with, and the similarity between these neighborhoods and neighborhoods in City of Toronto, they can locate an appropriate neighborhood in City of Toronto to open the Italian restaurant.

2 Data

2.1 Data Sources

The data for this project are from multiple sources.

Data of New York City is from website `https://geo.nyu.edu/catalog/nyu_2451_34572`, which includes four features *Borough*, *Neighborhood*, *Latitude*, and *Longitude*.

Data of City of Toronto is from City of Toronto website `https://www.toronto.ca/`. There are sixteen features, which are *ID*, *Area ID*, *Area attr ID*, *Parent Area ID*, *Area short code*,*Area long code*, *Area name*, *Area desc*, *X*, *Y*, *Longitude*, *Latitude*, *Object ID*, *Shape area*, *Shape length*, and *Geometry*.

The other features like different venue categories of New York City and City of Toronto can be extracted from *Foursquare* API. Foursquare is a location

data company that provides various location related data to mobile apps. In this project, I decide to scrape venues within 500 meters of the location of neighborhood. Also, I set the maximum number of venues for each neighborhood to be 100.

2.2 Data Cleaning

The New York City data is quite reliable. There is no missing value or duplicate. Also, all the features in the data sources are useful in this project. So, the dataset of New York City contains 306 observations (neighborhoods) and 4 features (*Borough*, *Neighborhood*, *Latitude*, and *Longitude*).

However, the data of City of Toronto is very messy since it is a location data that including geometric shapes. Also, unlike New York City data, there is no borough information in City of Toronto. So, there are 140 observations (neighborhoods) and 3 features (*Neighborhood*, *Latitude*, and *Longitude*) in the City of Toronto dataset.

Data of Venues and venue categories of each neighborhood in New York City is scraped from Foursquare by Foursquare API. This dataset includes 10229 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, and *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*). However, it only has 305 unique neighborhoods, which does not match the number of unique neighborhoods in New York City data. The observation with borough *Staten Island*, neighborhood *Howland Hook* is missing. After calling foursquare API based on longitude and latitude of this observation, it shows a warning message that *"There aren't a lot of results near you. Try something more general, reset your filters, or expand the search area"*. It seems the coordinate of this neighborhood is not accurate. Instead of deleting this neighborhood directly or expanding the search area, I use *GeoPy* library to extract more precise latitude and longitude of this neighborhood. Then, I call foursquare API based on improved location. The final dataset of venues and venues categories in New York City contains 10233 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*).

Similarly, venues and venue categories of neighborhoods in City of Toronto are obtained through Foursquare API. This dataset has 2092 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, and *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*). However, only 137 unique neighborhoods in this data, which does not match the number of unique neighborhoods in the dataset of City of Toronto. Three missing neighborhoods are *Newtonbrook West*, *St.Andrew-Windfields*, and *Willowridge-Martingrove-Richview*. By calling Foursquare API for these three neighborhoods, it returns warning message that *"There aren't a lot of results near you. Try something more general, reset your filters, or expand the search area."*, which is very similar to the situation of neighborhood *Howland Hook* in New York City data. So, *GeoPy* library is used to provide more accurate location of these three neighborhoods. However, only coordinate of neighborhood *Newton-*

brook West is available. Valid locations of neighborhoods *St.Andrew-Windfields* and *Willowridge-Martingrove-Richview* seems impossible to be obtained and I decide to delete them. Also, Foursquare API based on modified latitude and longitude works, so there are 2130 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*) of venues and venues categories in City of Toronto dataset.

3 Methodology

3.1 Exploratory Data Analysis

4 datasets are used in this project. New York City data has 306 observations (unique neighborhoods) and 3 features (*Neighborhood*, *Latitude*, and *Longitude*) while New York City venues and venues categories data has 10233 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*). There are 138 observations (unique neighborhoods) and 3 features (*Neighborhood*, *Latitude*, and *Longitude*) in City of Toronto data. For City of Toronto venues and venues categories data, there are 2130 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*).

3.1.1 New York City data

For New York City venues and venues categories dataset, there are 8033 unique venues, which indicates some of them either share the same name or are chain stores. The first ten most common venues in New York City are *Dunkin'*, *Chase Bank*, *Subway*, *Rite Aid*, *Starbucks*, *T-Mobile*, *Baskin-Robbins*, *CVS pharmacy*, *Popeyes Louisiana Kitchen*, and *Walgreens*. Their corresponding counts are listed in the following table:

Table of the number of Venues in New York City

Venues	Counts
Dunkin'	151
Chase Bank	91
Subway	68
Rite Aid	55
Starbucks	42
T-Mobile	37
Baskin-Robbins	35
CVS pharmacy	32
Popeyes Louisiana Kitchen	31
Walgreens	30

From the table, I find five of them are restaurants, three of them are drugstores, one of them is financial service, and the other one is telecommunication.

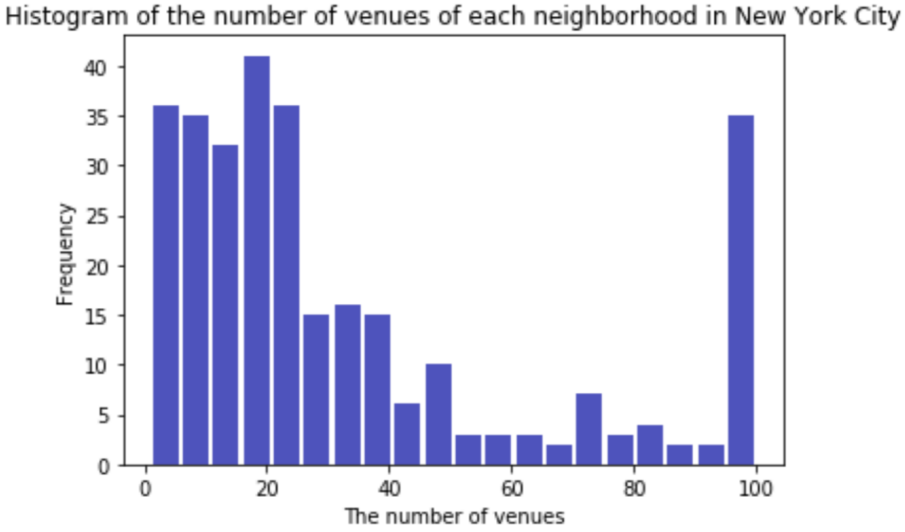
In addition, there are 434 unique venues categories in this dataset. The first ten most common venues categories are shown in the following table:

Table of the number of Venues Categories in New York City

Venues Categroy	Counts
Pizza Place	434
Italian Restaurant	321
Coffee Shop	281
Deli / Bodega	262
Bakery	229
Bar	218
Chinese Restaurant	214
Sandwich Place	185
Mexican Restaurant	175
American Restaurant	175

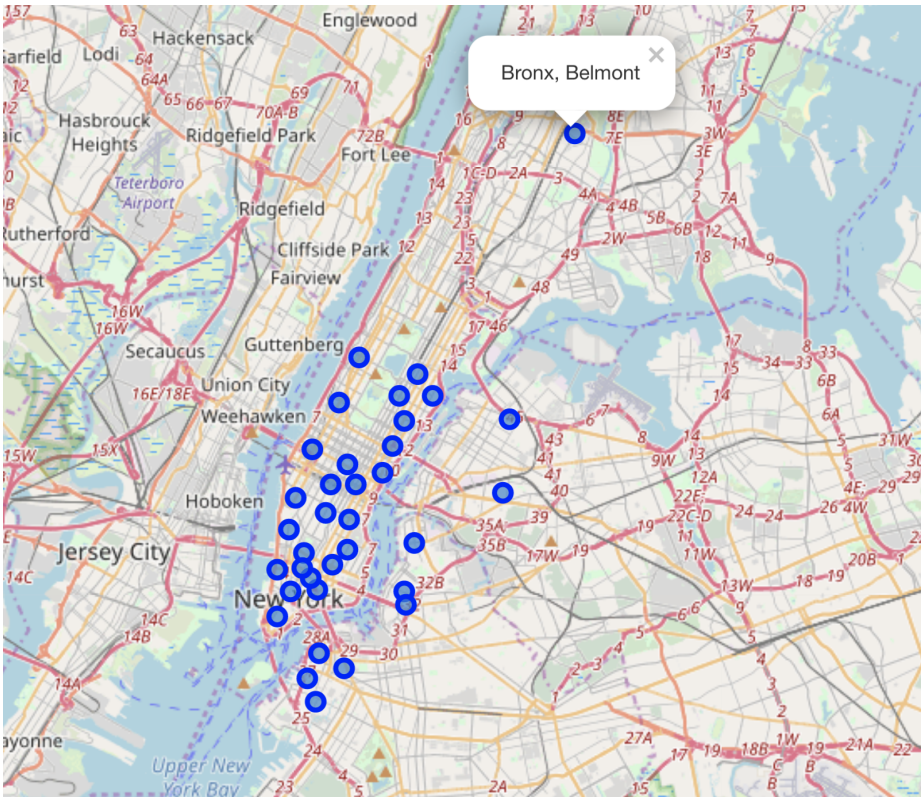
In this table, the first ten most common venue categories in New York City are *Pizza Place*, *Italian Restaurant*, *Coffee Shop*, *Deli / Bodega*, *Bakery*, *Bar*, *Chinese Restaurant*, *Sandwich Place*, *Mexican Restaurant*, and *American Restaurant*. All of them are related to food.

Furthermore, the number of venues for each neighborhood in New York City are shown in the following histogram:



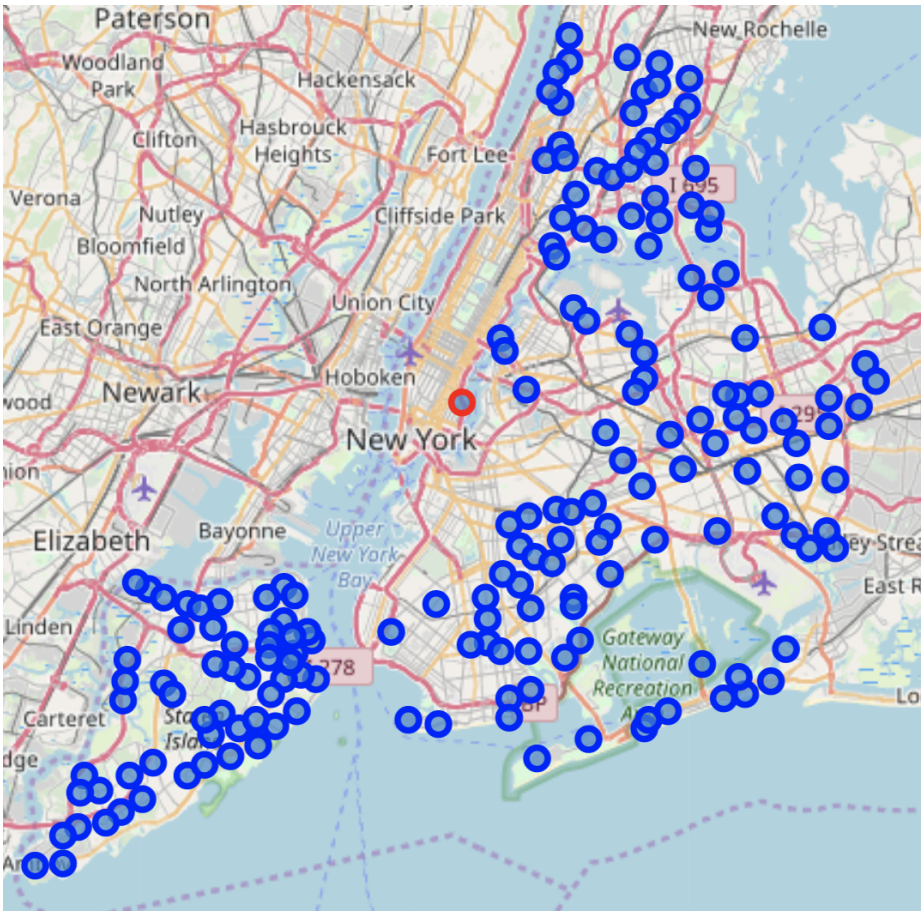
From histogram, I find neighborhoods with 95-100 venues or with 1-25 venues have higher frequency. There are 35 neighborhoods in New York City with 95-100 venues while 180 neighborhoods in New York City with 1-25 venues.

The following map shows the locations of 35 neighborhoods that have 95-100 venues:



In the map, most of neighborhoods with 95-100 venues are in borough *Manhattan* while the others are in boroughs like *Brooklyn* and *Queens* that are very close to *Manhattan*. The only special point is neighborhood *Belmont* in borough *Bronx*, which is far away from *Manhattan*.

The following map shows the locations of 180 neighborhoods that have 1-25 venues:



In this map, most of neighborhoods with 1-25 venues are spread around the New York City except borough *Manhattan*. The only neighborhood with 1-25 venues in borough *Manhattan* is *Stuyvesant Town*.

3.1.2 City of Toronto data

For City of Toronto venues and venues categories dataset, there are 1654 unique venues, which indicates some of them either share the same name or are chain stores. The first ten most common venues in City of Toronto are *Tim Hortons*, *Subway*, *Starbucks*, *Pizza Pizza*, *TD Canada Trust*, *Shoppers Drug Mart*, *Dollarama*, *The Beer Store*, *McDonald's*, and *LCBO*. Their corresponding counts are listed in the following table:

Table of the number of Venues in City of Toronto

Venues	Counts
Tim Hortons	62
Subway	43
Starbucks	35
Pizza Pizza	27
TD Canada Trust	25
Shoppers Drug Mart	22
Dollarama	14
The Beer Store	13
McDonald's	13
LCBO	12

From the table, I find seven of them are related to food, one of them is financial service, one of them is drugstore, and the other one of them is market.

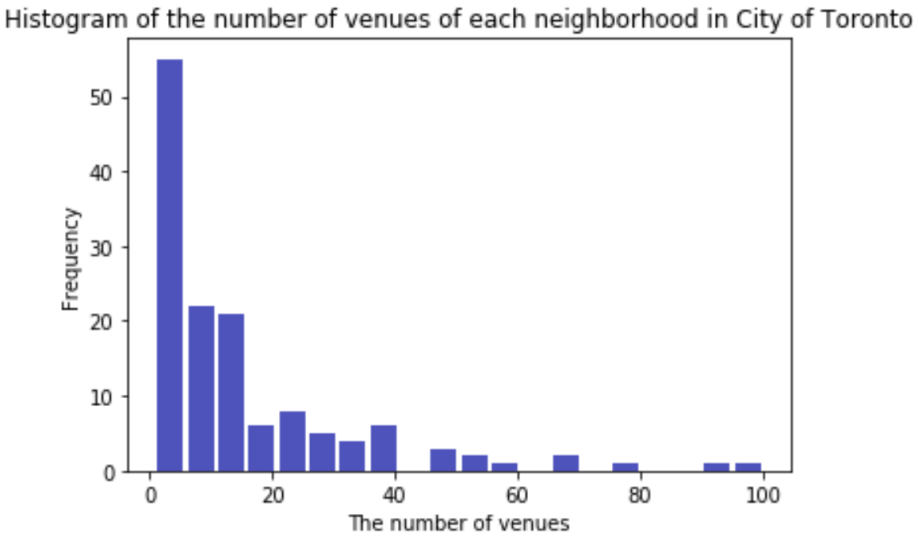
In addition, there are 282 unique venues categories in this dataset. The first ten most common venues categories are shown in the following table:

Table of the number of Venues Categories in City of Toronto

Venues Category	Counts
Coffee Shop	151
Park	89
Cafe	80
Pizza Place	76
Sandwich Place	63
Italian Restaurant	58
Fast Food Restaurant	45
Grocery Store	42
Bakery	41
Bar	41

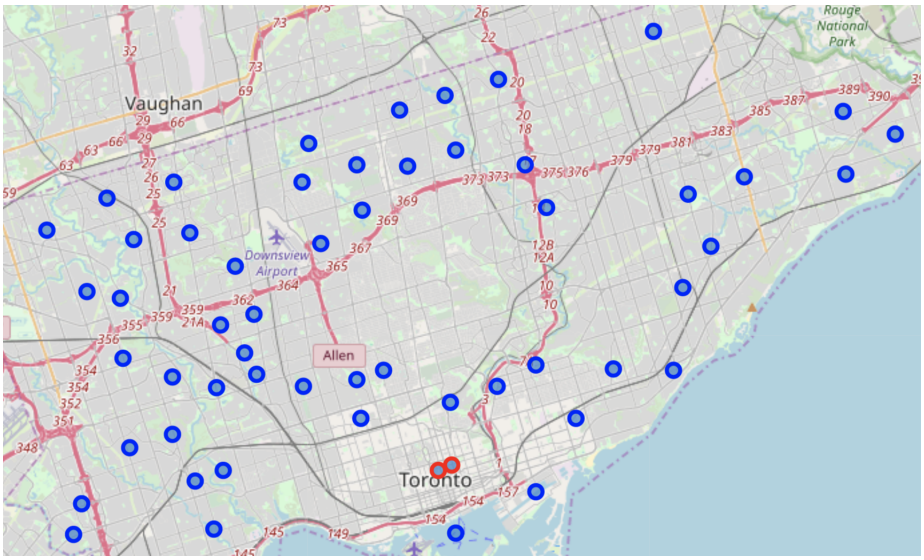
In this table, the first ten most common venue categories in New York City are *Coffee Shop*, *Park*, *Cafe*, *Pizza Place*, *Sandwich Place*, *Italian Restaurant*, *Fast Food Restaurant*, *Grocery Store*, *Bakery*, and *Bar*. Nine of them are related to food (include grocery store) while the other one is park.

Furthermore, the number of venues for each neighborhood in City of Toronto are shown in the following histogram:



From histogram, I find neighborhoods with 1-5 venues have higher frequency and the number of these neighborhoods is 55. Also, there are 2 neighborhoods with 90-100 venues.

The following map shows the locations of these neighborhoods (blue points represent neighborhoods with 1-5 venues while red points represent neighborhoods with 90-100 venues):



In this map, all the neighborhoods with 1-5 venues spread around the City of

Toronto. However, only neighborhoods with 90-100 venues are in the downtown Toronto area.

3.2 Clustering Analysis

3.2.1 Preparation

First, I combine New York City data with City of Toronto data. The combined City data contains 444 observations (unique neighborhoods of New York City and City of Toronto) and 3 features (*Neighborhood*, *Latitude*, and *Longitude*.) Also, by combining New York City venues and venues categories data with City of Toronto venues and venues categories data, there are 12363 observations and 7 features (*Neighborhood*, *Neighborhood Latitude*, *Neighborhood Longitude*, *Venue*, *Venue Latitude*, *Venue Longitude*, and *Venue Category*) in the combined City venues and venues categories dataset.

In order to conduct clustering analysis, I perform one-hot encoding to convert categorical features into multiple dummary numerical features. The transformed dataset has 12363 observations and 466 features.

Below is the sample of dataframe after one-hot encoding:

	neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	...
0	Bronx, Wakefield	0	0	0	0	0	0	0	0	0	...
1	Bronx, Wakefield	0	0	0	0	0	0	0	0	0	...
2	Bronx, Wakefield	0	0	0	0	0	0	0	0	0	...
3	Bronx, Wakefield	0	0	0	0	0	0	0	0	0	...
4	Bronx, Wakefield	0	0	0	0	0	0	0	0	0	...

From the dataframe, 466 features are neighborhood and 465 venues categories.

Next, the dataframe is grouped based on unique neighborhood and the average number of venues categories for each neighborhood is calculated. The new dataframe includes 444 observations (unique neighborhoods) and 466 features.

Below is the sample of dataframe after grouping:

	neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	...
0	Agincourt North	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...
1	Agincourt South-Malvern West	0.000000	0.0	0.0	0.0	0.045455	0.0	0.0	0.0	0.0	...
2	Alderwood	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...
3	Annex	0.000000	0.0	0.0	0.0	0.041667	0.0	0.0	0.0	0.0	...
4	Banbury-Don Mills	0.045455	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...

Different clustering methods in next few sections will be performed on this dataframe.

Moreover, the first ten most common venues categories for each neighborhood are calculated and are used as features in the new dataset. This dataset has 444 observations (unique neighborhoods) and 11 features (neghborhood and the first ten most common venues categories).

Below is the sample of this dataset:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt North	Chinese Restaurant	Park	Convenience Store	Sandwich Place	Pizza Place	Liquor Store	Bakery	Discount Store	Dim Sum Restaurant	Vietnamese Restaurant
1	Agincourt South-Malvern West	Chinese Restaurant	Shopping Mall	Malay Restaurant	Dim Sum Restaurant	Mediterranean Restaurant	Bank	Seafood Restaurant	Filipino Restaurant	Asian Restaurant	Cantonese Restaurant
2	Alderwood	Pizza Place	Athletics & Sports	Coffee Shop	Convenience Store	Pharmacy	Zoo Exhibit	Event Service	Event Space	Exhibit	Eye Doctor
3	Annex	Sandwich Place	Pub	Café	BBQ Joint	Donut Shop	Social Club	Pizza Place	Vegetarian / Vegan Restaurant	Liquor Store	Pharmacy
4	Banbury-Don Mills	Gourmet Shop	Shoe Store	Coffee Shop	Liquor Store	Shopping Mall	Sandwich Place	Pizza Place	Pharmacy	Movie Theater	Italian Restaurant

3.2.2 K-means clustering

K-means clustering is one of the famous unsupervised learning methods. It divides observations into K groups based on the centroid (mean) of each group. Specifically, for each observation, the distances between it and K centroids are calculated. Then this observation is assigned to the group, which the distance between the observation and group centroid is minimum. The detailed algorithm of this method is the following (P388, Introduction to Statistical Learning, James et al. (2013)):

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

3.2.3 Non-negative Matrix Factorization and K-means clustering

I notice that dataset grouped based on unique neighborhoods is sparse and high-dimensional, which may affect the performance of K-means clustering.

In order to handle both sparsity and high-dimension problem, I decide to conduct non-negative matrix factorization (NMF) first to reduce the dimension of data and then use K-means clustering for transformed data.

Similar to Principal component analysis (PCA), NMF is also a dimension reduction method. but unlike PCA, which creates new features by making linear transformations of all the original features, the advantage of NMF is that it keeps all the original features in the data after dimension reduction.

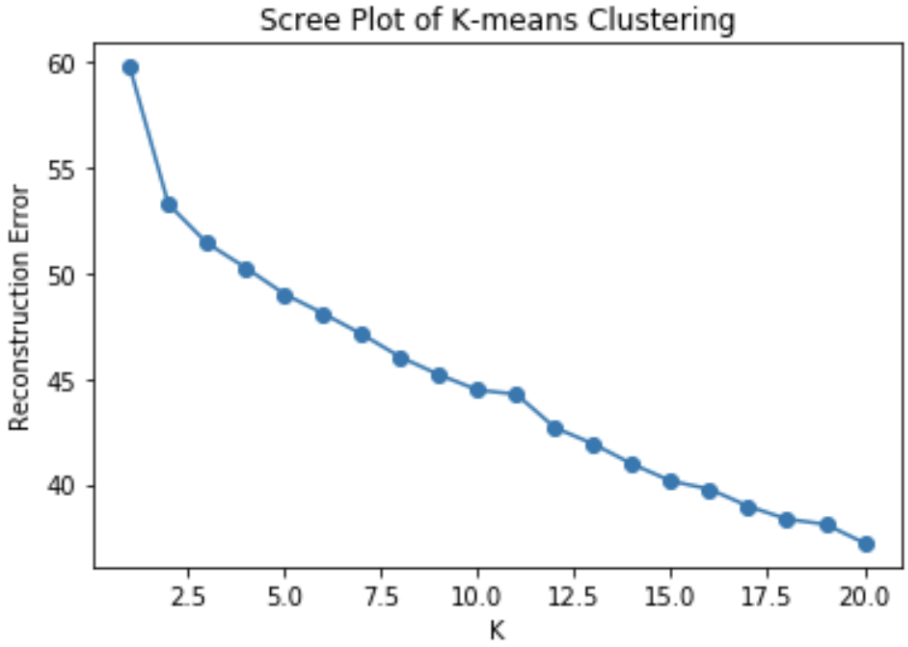
Technically, NMF decomposes a data matrix X, where all the entries are non-negative, into two matrix W and H. W is the transformed dataset after dimension reduction while H keeps all the information between original features and new features. The detail of this decomposition is complicated and is ignored intentionally. Audiences who are interested in this part can refer to materials on the Internet.

4 Results

4.1 Results of K-means Clustering

I perform K-means clustering on dataset that is grouped based on unique neighborhoods.

Since K-means clustering is an unsupervised learning method, I use reconstruction error as the evaluation criterion and the number of K is chosen based on the scree plot:

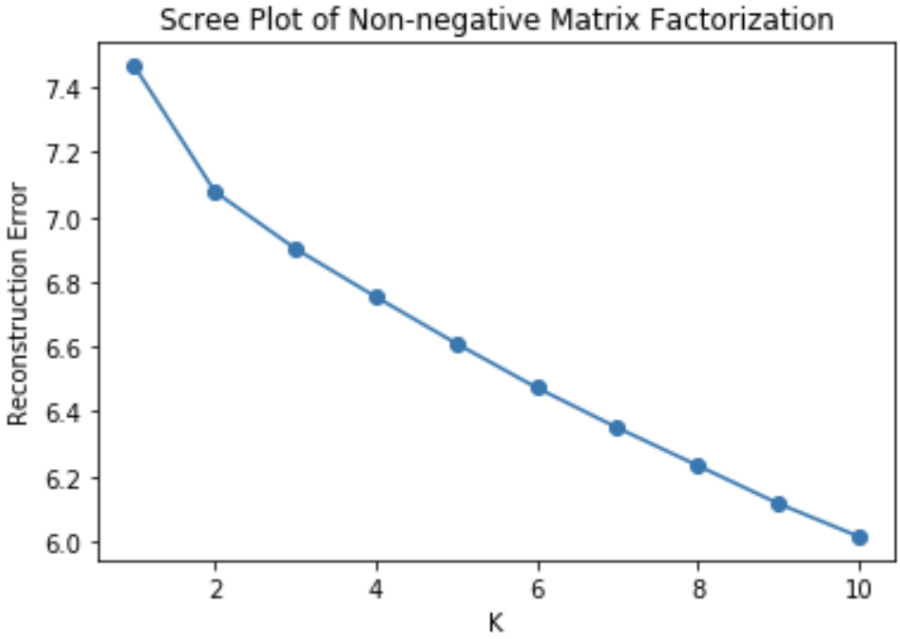


From the scree plot, the elbow point is at K=2. So, I select K=3 for conservative purposes.

Then, I conduct K-means clustering with K=3 on this dataset.

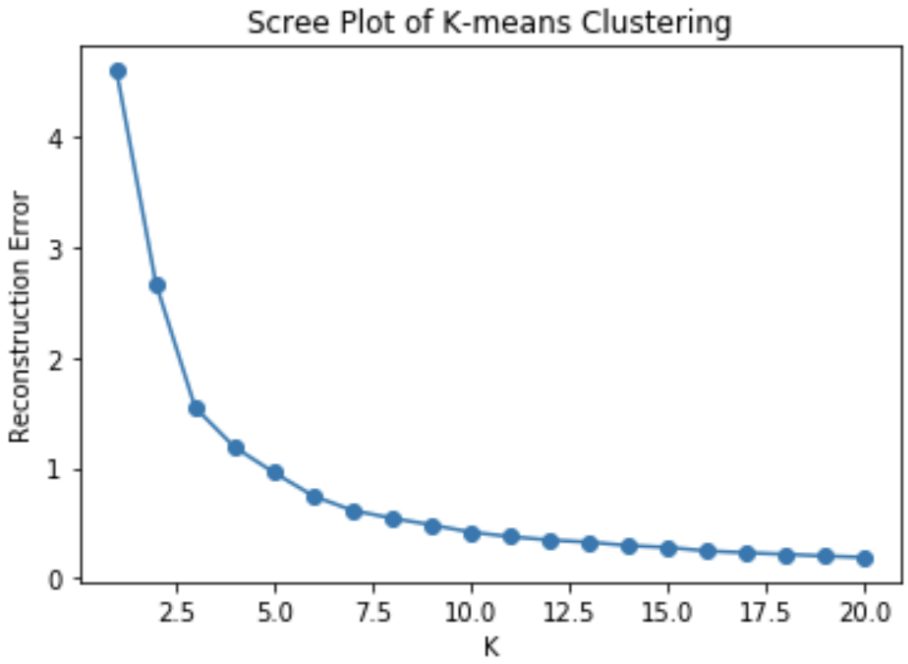
4.2 Results of NMF and K-means Clustering

The reduced dimension of data is selected by the scree plot based on reconstruction error:



From the scree plot, the elbow point is at K=2. So, I select K=3 for conservative purposes.

Then K-means clustering is performed on the transformed data, where K is chosen based on scree plot:



Scree plot indicates the elbow point is at K=3 and I select such K in order to make comparisons with the results of K-means Clustering.

5 Discussion

5.1 Discussion of K-means Clustering

After performing K-means clustering, neighborhoods of both cities are assigned to 3 groups separately.

Group 1 has 404 neighborhoods where 288 of them are from New York City. The first ten most common venues categories for this group are listed as below:

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Italian Restaurant	Italian Restaurant	Pizza Place	Pizza Place	Pizza Place	Pizza Place	Exhibit	Eye Doctor	Eye Doctor	Factory

The important characteristics of these neighborhoods are Italian restaurant and Pizza place. I conclude that people in these neighborhoods pay more attention on food especially Italian cuisine.

Group 2 contains 26 neighborhoods where 4 of them are form New York City. The first ten most common venues categories for this group are:

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Park	Park	Zoo Exhibit	Zoo Exhibit	Event Space	Exhibit	Eye Doctor	Factory	Falafel Restaurant	Farm

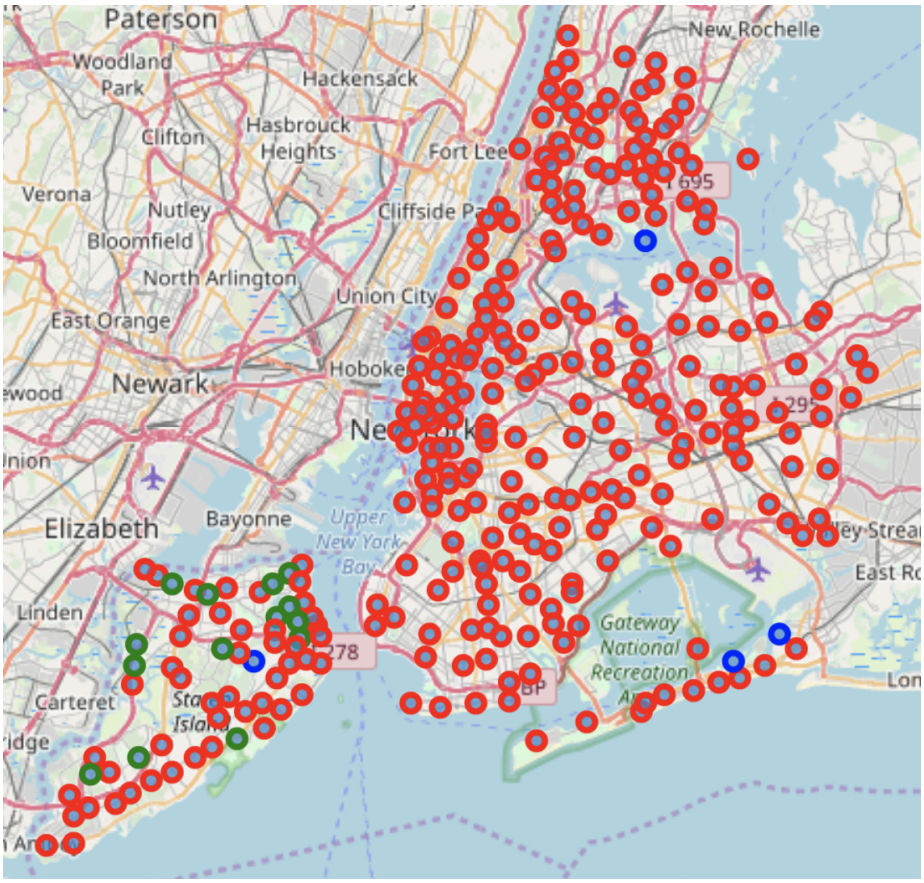
The important venue categories shared by these neighborhoods are Park and Zoo Exhibit. I conclude that neighborhoods in this group provide individuals with places that connect with nature.

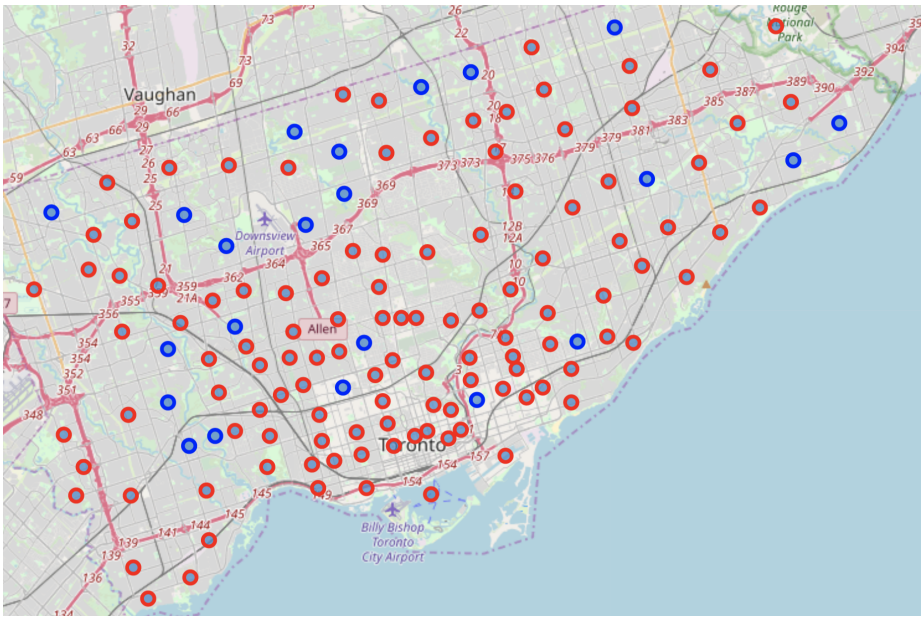
There are 14 neighborhoods in group 3 and all of them are from New York City. The first ten most common venues categories for this group are:

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bus Stop	Coffee Shop	Pizza Place	Zoo Exhibit	Financial or Legal Service	Event Space	Exhibit	Factory	Factory	Falafel Restaurant

The important features of these neighborhoods are bus stops and restaurants. Also, there are various exhibitions and financial/legal services. It seems these neighborhoods are integrated area, where people get off the bus here and eat some food, visit exhibitions, and use financial services.

The maps of grouped neighborhood in each city are listed below (group 1 is red, group 2 is blue, and group 3 is green).


Paterson
Woodland Park
Hackensack
Clifton
Hasbrouck Heights
Fort Lee
New Rochelle
Verona
Nutley
Bloomfield
Cliffside Park
695
North Arlington
Union City
East Orange
Hoboken
Newark
Elizabeth
Bayonne
Upper New York Bay
Linden
278
Gateway National Recreation
Carteret
Staten Island
East Rockaway



5.2 Discussion of NMF and K-means Clustering

After performing NMF and K-means clustering, neighborhoods of both cities are assigned to 3 groups separately.

Group 1 has 23 neighborhoods where 5 of them are from New York City. The first ten most common venues categories for this group are:

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Park	Park	Coffee Shop	Zoo Exhibit	Pizza Place	Exhibit	Eye Doctor	Factory	Falafel Restaurant	Farm

The important venues categories shared by these neighborhoods are Park, coffee shop and Zoo Exhibit. I conclude that these neighborhoods provide people with places that connect with nature. Also, coffee shop seems a complementary place for the rest when individuals get tired.

Group 2 contains 390 neighborhoods where 279 of them are form New York City. The first ten most common venues categories for this group are:

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Italian Restaurant	Italian Restaurant	Pizza Place	Pizza Place	Pizza Place	Exhibit	Exhibit	Eye Doctor	Factory	Falafel Restaurant

The important characteristics of these neighborhoods are Italian restaurant,

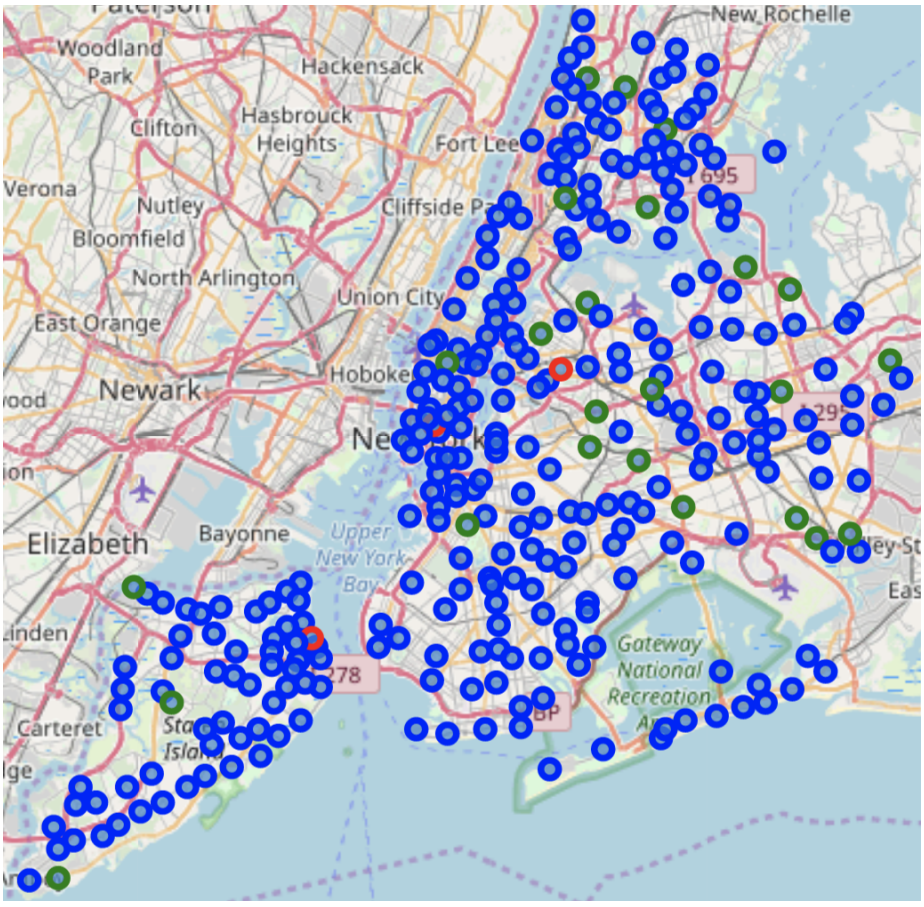
Pizza place, and Exhibit. I conclude that these neighborhoods provide people with food, especially Italian cuisine. Also, these neighborhoods have a lot of exhibitions

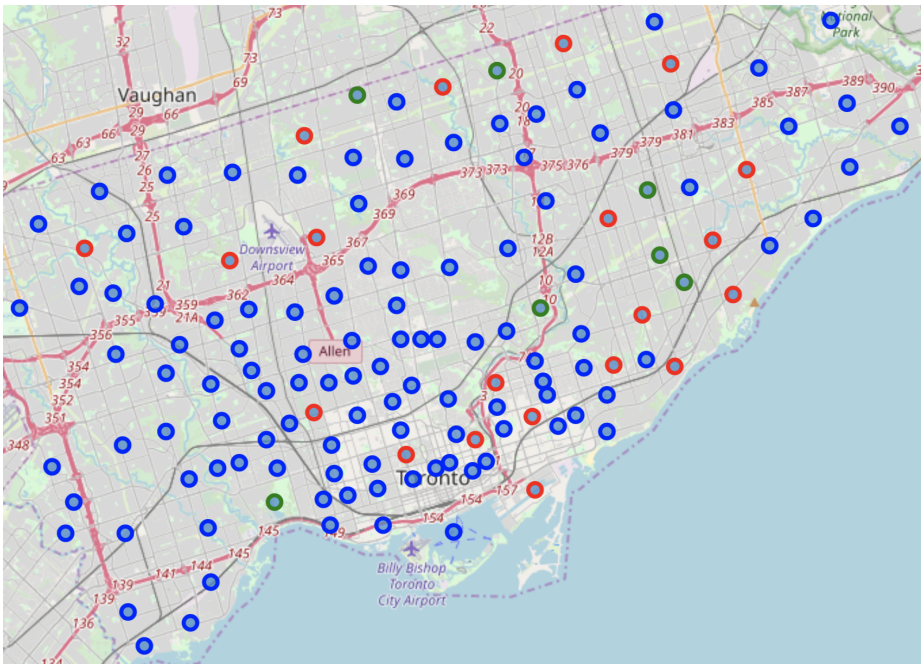
There are 31 neighborhoods in group 3 and 24 of them are from New York City. The first ten most common venues categories for this group are:

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Pizza Place	Pizza Place	Chinese Restaurant	Deli / Bodega	Pharmacy	Bank	Donut Shop	Fast Food Restaurant	Middle Eastern Restaurant	Falafel Restaurant

Most features are related to food. They can be interpreted as people in these neighborhoods pay more attention on various food.

The maps of grouped neighborhood in each city are listed below (group 1 is red, group 2 is blue, and group 3 is green).





5.3 Discussion of both methods

From discussion of K-means clustering and discussion of NMF and K-means clustering, I find these two methods share two groups, which are interpreted as neighborhoods with lots of Italian cuisine, and neighborhoods connecting with nature.

The only difference is that the last group in K-means clustering has neighborhoods with multiple functions while that in NMF and K-means clustering has neighborhoods with various kinds of food.

6 Conclusion

In this project, my goal is to study the similarity of neighborhoods in New York City and City of Toronto. Specifically, neighborhoods segmentation of two cities are performed based on K-means clustering and combination of Non-negative Matrix factorization and K-means clustering.

The results of this project can be used to answer business problems like which area in City of Toronto should be invested and what kind of business activities in this area are suggested based on the information of New York City. Also, the analysis of this project can be applied to the other cities

However, there are some drawbacks of this project. First, some neighborhoods in City of Toronto are missing. This would affect the range of application

of this project. Second, although Non-negative Matrix factorization can solve sparse and high-dimensional data problem, choosing the number of reduced dimension of transformed data may introduce extra noise. It is worthwhile to consider the other methods to handle sparse and high-dimensional data. Additionally, there is no mechanism of model selection for two clustering methods in this project. That is, I am not able to decide which one is better. All I can say is that two clustering methods provides different perspectives of neighborhoods segmentation, and people should balance the results of them when they make decisions. Nevertheless, I will delve into these topics and solve them in the future.