**Determining Similarity Between Cities and Recommending locations for Tourist/Business Owners**

**1.Introduction**

**1.1 Background:**

To compare the similarity between two biggest financial cities of the world Toronto and New York. This similarity well be compare based on the neighborhoods and venues nearby like hospital, entertainment, school, cafes, outdoor activities etc. This comparison will help people in migrating to either countries. People always have questions before migrating to any place like good schools for kids, hospitals, Jobs, entertainment, safety, public transport etc. This model will help give them all the answers and help them in determining whether to move or not. Target audience could be financial analysts or people from financial business who constantly move from one place to another. Second audience could be general public who is thinking to move to different country Canada especially in Toronto.

**2nd problem statement** is for business owners/entrepreneurs who wants to start a business in either cities. Model will help in determining which location is more appropriate for their business. This is very critical for business owners if opened in wrong place can led to a disaster results and can loss million of dollars. This model will not save their millions of dollars but provide them all the strategic information in quick time which help them in determining appropriate location. This model can also be used to recommend top places to tourists like Yelp. Target audience is Tourists, business owners, industrialist or entrepreneurs, small business.

**1.2 Objectives:**

**1.2.1 Similarity between cities**

This project aims to predict similarity between two financial cities include schools, entertainment, universities, hospitals, restaurants etc. which helps people to make strategic decision whether to migrate or not.

**1.2.2 Determining Location**

Second aim of this project to provide appropriate location to start their business which helps business owners, industrialist etc. to take strategic decision.

**2. Data Sources**

For this project we will be using below data sources:

* Four square data location data
* Extracting Toronto data from Wikipedia.
* Geospatial Data

**2.1 To get the New York neighborhoods and their longitude and latitude** [**https://cocl.us/new\_york\_dataset**](https://cocl.us/new_york_dataset)

Data from this link will be in JSON format will be converted into pandas data frame. This data has 4 attributes namely Borough, neighborhoods, latitude and longitude and 306 rows. 306 unique neighborhoods which will be classified in one of the boroughs.

**Borough Neighborhood Latitude Longitude**

Manhattan Marble Hill 40.876551 -73.9106601

Manhattan Chinatown 40.715618 -73.9942792

Manhattan Washington Heights40.851903 -73.9369003

Manhattan Inwood 40.867684 -73.9212104

Manhattan Hamilton Heights40.823604 -73.949688

* 1. **Four square New York location Data**

Four square location API will help in getting different venues like hospitals, education, restaurant data etc. to explore each individual borough and neighborhoods.

E.g. API Format is

url = 'https://api.foursquare.com/v2/venues/explore?&client\_id=**{}**&client\_secret=**{}**&v=**{}**&ll=**{}**,**{}**&radius=**{}**&limit=**{}**'.format(

CLIENT\_ID,

CLIENT\_SECRET,

VERSION,

lat,

lng,

radius,

LIMIT)

This API will response with venue details in that neighborhood like Coffee, Entertainment, restaurants, museums etc. that will help tourists to recommend places or comparing between different cities.

* 1. **Toronto data extracted from Wikipedia**

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>'

|  |  |  |  |
| --- | --- | --- | --- |
|  | Post Code | Borough | Neighbourhood |
| 0 | M3A | North York | Parkwoods |
| 1 | M4A | North York | Victoria Village |
| 2 | M5A | Downtown Toronto | Harbour Front |
| 3 | M5A | Downtown Toronto | Regent Park |
| 4 | M6A | North york | Lawrence Park |

* 1. **Geospatial\_Coordinates CSV file** provided by this course itself.

There are three attributes Postal code, Latitude, Longitude.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Post Code | Latitiude | Longitude |
| 0 | M3A | 43.806686 | -79.194353 |
| 1 | M4A | 42.806686 | -79.209876 |
| 2 | M5A | 43.780987 | -79.194345 |
| 3 | M5A | 43.809686 | -79.198452 |
| 4 | M6A | 43.807645 | -79.197612 |

* 1. **Four square Toronto location Data**

Four square location API will help in getting different venues like hospitals, education, restaurant data etc. to explore each individual borough and neighborhoods.

**3. Exploratory Data Analysis**

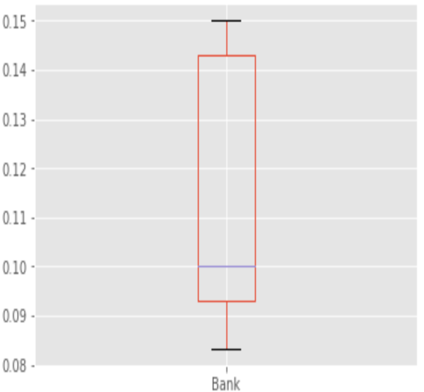
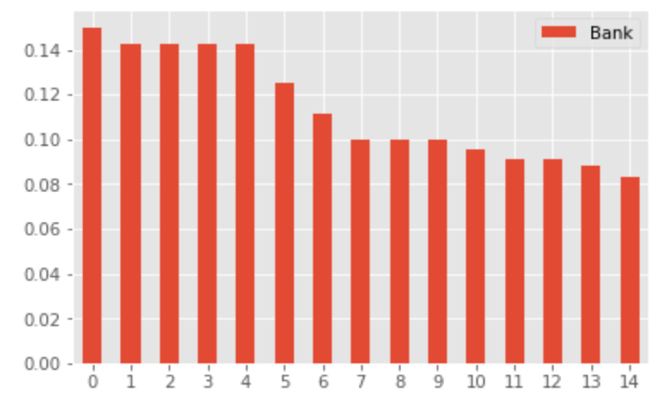
In this section we will be understanding the relationship between different neighborhoods and some of their important venues such as bank, train station and supermarket.

**3.1 Relationship between Neighborhoods and Banks**

Here, we are doing the analysis on New York Neighborhoods and Banks. We will be understanding it through bar chart and box plot.

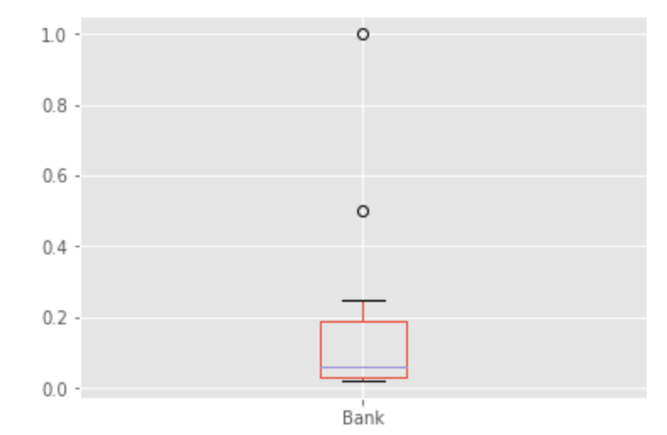
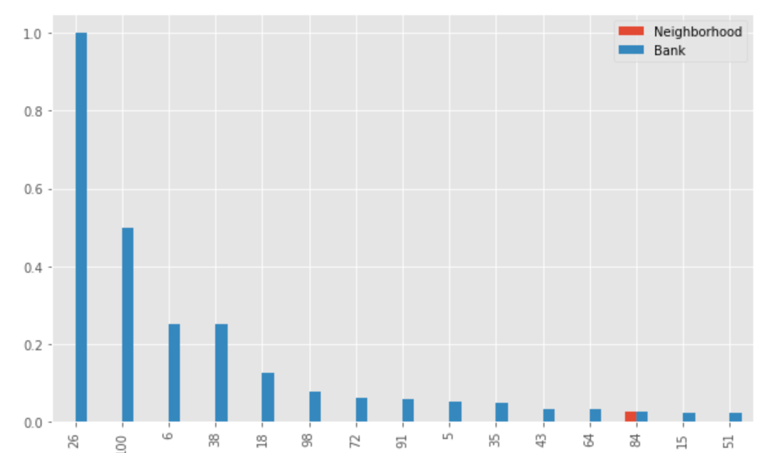
***3.1.1 New York***

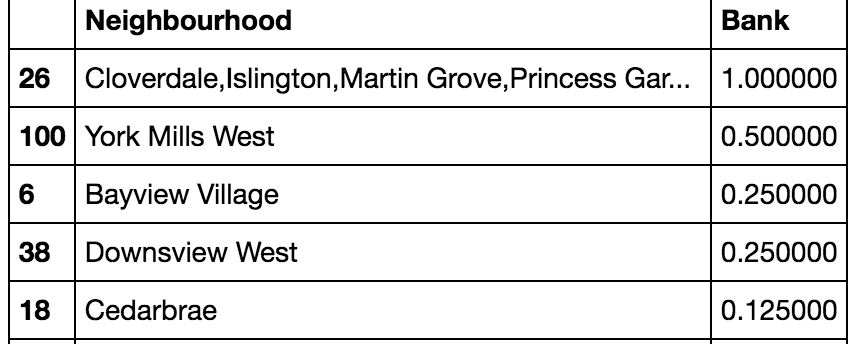
Some of the top banks are in Borough Park, Richmond Valley, Castle Hill, Prince's Bay, Huguenot.

****

***3.1.2 Toronto***

Top banks in Toronto are in Cloverdale, Islington, Martin Grove, Princess Gar, York Mills West and every neigbhorhood has a bank.



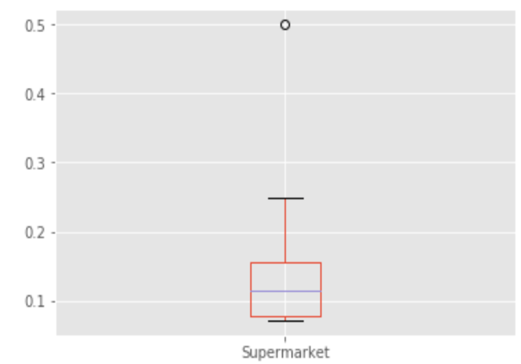
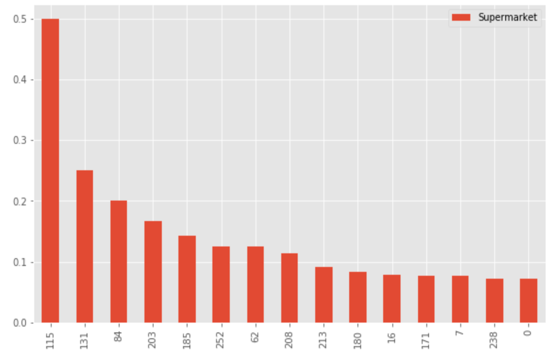
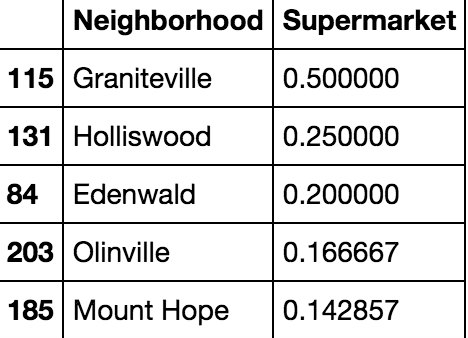


**3.2 Relationship between Neighborhoods and Supermarket.**

Here, we are doing the analysis on New York Neighborhoods and Supermarket. We will be understanding it through bar chart and box plot.

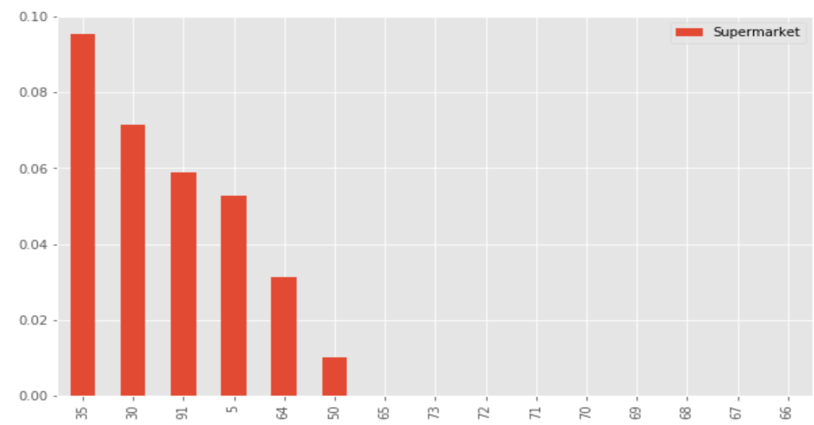
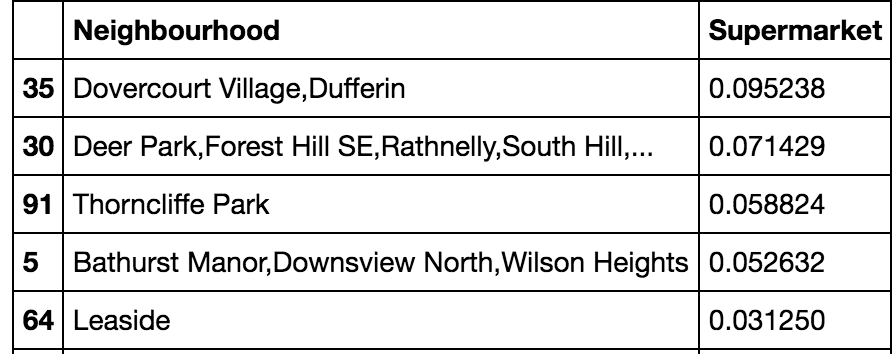
***3.2.1 New York***

Neighborhoods like Graniteville are like an outlier which has almost double stores as compared to any other neighborhood around it.

****

***3.2.2 Toronto***

There are very less supermarkets in the neighborhoods of Toronto and places like Little Portugal, Trinity, Parkswoods, parksdale has zero super markets.

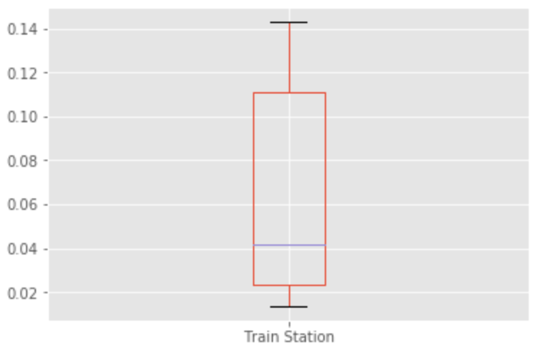
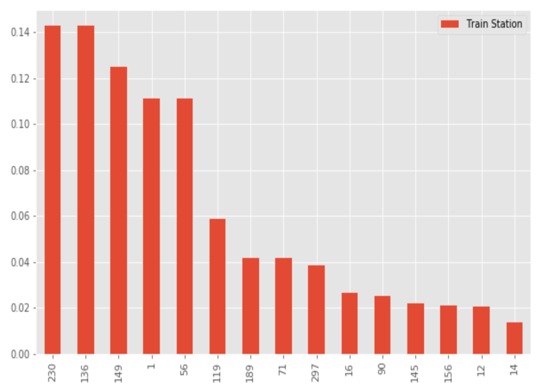
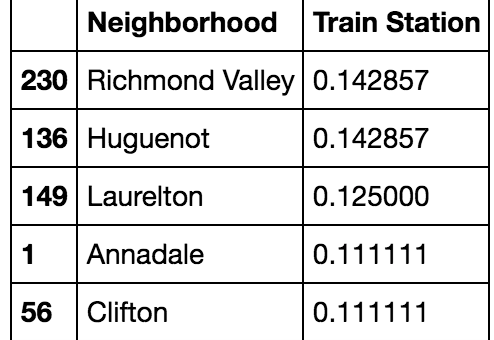
****

**3.3 Relationship between Neighborhoods and Train Market**

Here, we are doing the analysis on New York Neighborhoods and Train Market. We will be understanding it through bar chart and box plot.

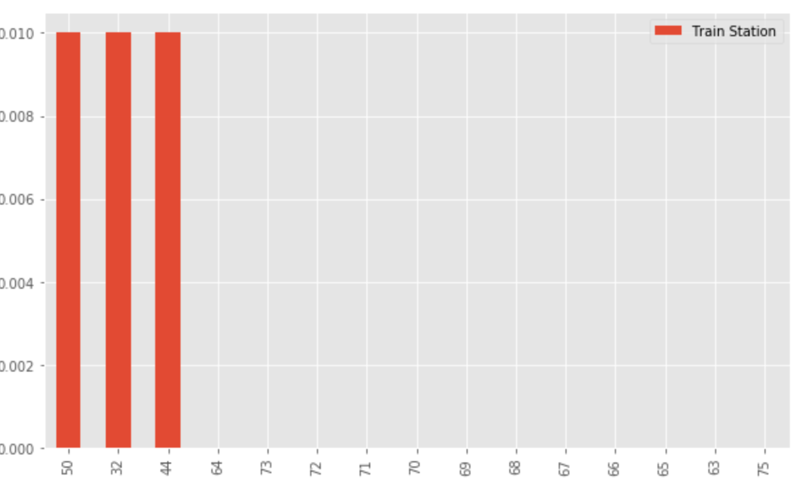
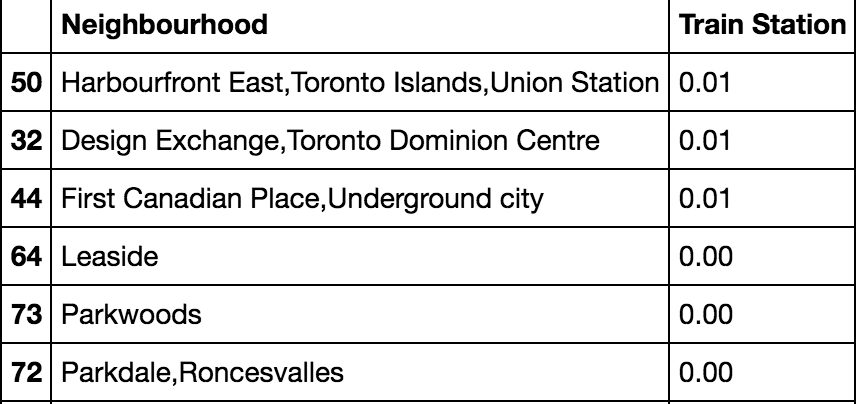
***3.3.1 New York***

There is good amount of train stations in Richmond valley, Huguernot and Laurelton but very less in other places.



***3.3.2 Toronto***

Neighborhoods of Toronto has very less train stations and those who loves to travel in metro then Toronto might not be a good option.

****

**4. Data Pre-processing and Feature Engineering**

In this section we will discuss some of the pre-processing and feature engineering which have been performed in this use case.

**4.1 Data Pre-processing**

Pre-processing is one of the most important steps. Here we will clean our data such as removing noise, replacing missing values, normalization, aggregation etc.

**4*.1.1 New York Data***

Data <https://cocl.us/new_york_dataset> extracted from this source will be in JSON Format and transformed into Pandas data frame. Next step is defining the columns for transformed data frame. The column names are Borough, Neighborhood, Latitude, Longitude. Using Four square link API we will request and explore venues for New York within the radius of 500m and pulling the 100 closest venues around each neighbor.

***4.1.2 Toronto Data***

Toronto data is pulled from Wikipedia '<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text> and transformed into pandas data frame. We have filtered the values where value of Borough is "Not assigned". We have combined the neighborhood which has same Postcode and Borough and dropped the duplicates. The next step we performed is checking the values of Neighborhood. If value of Neighbourhood is “Not assigned” then value of borough and neighborhood are same. We have used the Geospatial\_Coordinates.csv data to get the latitudes and longitudes of all the boroughs and neighborhoods and combined this data with Wikipedia data. Using Four square link API we will request and explore venues for Toronto within the radius of 500m and pulling the 100 closest venues around each neighbor.

4.2 **Feature Engineering**

Feature Engineering is used to convert your raw data into feature vector which model use to understand the patterns, trends in the data and help in classification, regression, clustering etc. activities.

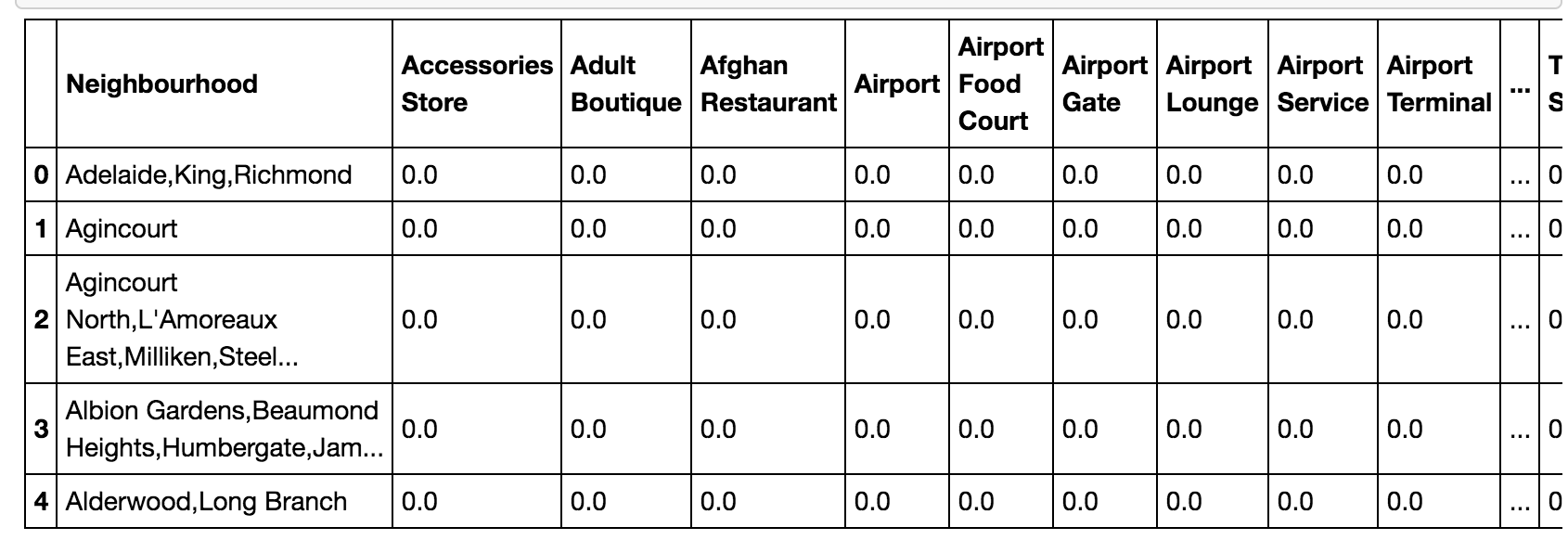
**4.2.1 Feature Engineering on New York Data**

One of the attributes in our dataset is venues which has categorical values such as airlines, coffee, gym, school etc. We have used one hot encoding technique to convert into dummy variables which has values 0 or 1.

Total attributes of the data frame before One-hot Encoding is 7. After one-hot encoding it increases to 280 columns.

****

*Figure 1: Before One-hot Encoding*

****

*Figure 2: After One-hot Encoding*

**4.2.2 Feature Engineering on Toronto Data**

One of the attributes in our dataset is venues which has categorical values such as airlines, coffee, gym, school etc. We have used one hot encoding technique to convert into dummy variables which has values 0 or 1 as shown in Figure 1 and Figure 2.

**5. Methodology**

The objective of this use case is to solve three problems:

1. Model recommends the cities which are similar to his residing city. This will help person to migrate to different city which is similar to his/her residing city.
2. It will recommend top places or venues in the city to the tourist.
3. Compare the neighborhoods of New York and Toronto and cluster them based on their similarity.

**5.1 K-Nearest Neighbor Algorithm**

KNN model was used to compare the cities between current residing and other cities and recommend K number of cities which are similar to each other. We calculate Euclidean distance between cities and lower value will give more similarity between the cities.

***5.1.1 Algorithm***

1. Initialize K to your chosen number of neighbors

3. If migrate to Toronto, For each record in the Toronto data

3.1 Calculate the Euclidean distance between the New York query data and Toronto Data

3.2 Add the distance and the index of the example to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances

5. Pick the first K entries from the sorted collection

***5.1.2 Pseudocode***

*toronto\_place.head()*

*toronto\_knn = toronto\_place.drop(['Neighbourhood'],axis=1)*

*from sklearn.neighbors import NearestNeighbors*

*neigh = NearestNeighbors(n\_neighbors=5)*

*neigh.fit(newyork\_knn)*

*print(neigh.kneighbors(toronto\_knn))*

**5.2 K-means Algorithm**

K-means algorithm is used to find the clusters between two cities. K-means will find the clusters or the group of cities which are close to each other.

1. Choose K, then select K random centroids
2. Assign records to the cluster to the closest centroid.
3. When records are moved to a new cluster, the centroid for the affected clusters needs to be recalculated.
4. This movement and recalculation are repeated until movement no longer results in an improvement.

***5.2.2 Pseudocode***

*Toronto\_cluster\_data = toronto\_grouped[['Bank', 'Bar','Church','Department Store', 'Kids Store', 'Pool','Supermarket','Train Station','Office','Museum','Pool']]*

*Newyork\_cluster\_data = newyork\_grouped[['Bank', 'Bar','Church','Department Store', 'Kids Store', 'Pool','Supermarket','Train Station','Office','Museum','Pool']]*

*frames = [Toronto\_cluster\_data, Newyork\_cluster\_data]*

*cluster\_data = pd.concat(frames)*

*kmeans = KMeans(n\_clusters=2, random\_state=0).fit(cluster\_data) # check cluster labels generated for each row in the dataframe*

*kmeans.labels\_*

**5.3 Recommend Top Places**

1. Using Four square API we explore all the venues of each neighborhood.
2. Values of the Venues are in categorical form such as Airport, Bar, Indian restaurant, School, Hospital.
3. We convert values of venues into 0 or 1 using one-hot encoding technique.
4. Take the mean of the rows.
5. Then transpose the data frame obtained in point 4 and sort it in descending order.
6. Then display Top 10 values or venues of each neighborhood.

***5.3.1 Pseudocode***

*sortt = newyork\_grouped[newyork\_grouped['Neighborhood']=='Battery Park City'].T.reset\_index()*

*sortt.head()*

*temp=sortt.iloc[1:]*

*temp.columns=['venue','freq']*

*temp.head()*

*top\_venues = temp.sort\_values(by=['freq'], ascending=****False****).head()*

*top\_venues.head(10)*

**6. Results**

In this section, we are presenting our findings and results.

**6.1 Model help person to migrate to different city which is similar to his/her residing city.**

*Example:- Person residing in Toronto and neighborhood is. This person personal preferences are Bank', 'Bar','Church','Department Store', 'Kids Store', 'Pool','Supermarket','Train Station' and wants to migrate to New York. The model explained in 4.1.1 will recommend below neighborhoods of new york.*

Neighborhood Chinatown

Neighborhood Financial District

Neighborhood Bensonhurst

Neighborhood Manhattan Valley

Neighborhood Vinegar Hill

**6.2 Recommend top places or venues in the city to the tourist.**

#### *Example. User is coming to New York and neighborhood is Battery Park City . Top Places or venues given by the algorithm explained in 4.3 are*

#### 

**6.3 Compare the neighborhoods of New York and Toronto and cluster them based on their similarity.**

*Using K-means clustering algorithm. We have found both cities are quite similar to each other if we consider feature vector: {'Bank', 'Bar','Church','Department Store', 'Kids Store', 'Pool','Supermarket','Train Station','Office','Museum','Pool'}*

**7. Conclusion**

In this study, I have created two different models. First one will help user to determine which cities they need to migrate based on their preferences using KNN model. Second one will recommend top places or venues of their travelling city. Third one is comparison between 2 different cities New York/Toronto and found to be quite similar to each other using K-means alogirthm see in section 5.2.

In this study, we also found Train stations are very less in Toronto and those who loves going in metro might not be a good option. In New York, there are good amount of train stations in Richmond valley, Huguernot and Laurelton but very less in other neighborhoods but still its not a matter of concern. Top banks in Toronto are in Cloverdale,Islington,Martin Grove,Princess Gar, York Mills West and every nieighborhood has a bank. Neighborhoods like Graniteville are like an outlier which has almost double stores as compared to any other neighborhood around it. There are very less supermarkets in the neighborhoods of Toronto and places like Little Portugal, Trinity, Parkswoods, parksdale has zero super markets.