Clustering of Similar Neighborhoods in Different Cities

New York City and Toronto

March 2020

Introduction

- Purpose: Explore and investigate the distribution of different venues in New York City and Toronto.

- Analysis to help new visitors to stay in better place.
- New York welcomes 65.2 million Tourists
- Toronto welcomes 27.5 million Tourists
- This will help understand the 2 cities better.

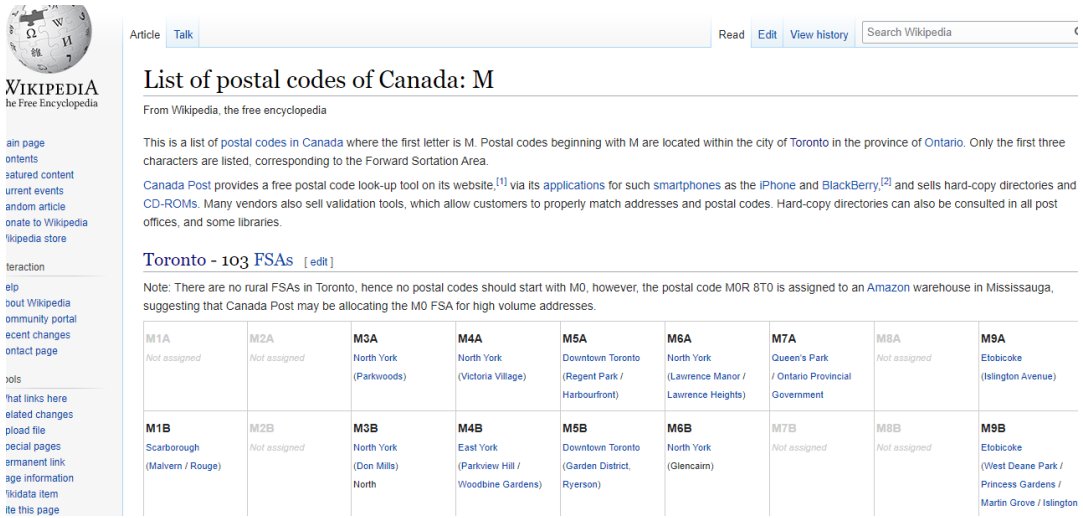
Data Preprocessing

FILES(GEOLOCATION DATA)

HTTPS://COCL.US/GEOSPATIAL_DATA
(.CSV) -- USED IN TORONTO.IPYNB

HTTPS://COCL.US/NEW_YORK_DATASET
(.JSON) -- USED IN NEWYORKCITY.IPYNB

Data Preprocessing



WIKIPEDIA
The Free Encyclopedia

Article Talk | Read Edit View history | Search Wikipedia

List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0, however, the postal code M0R 8T0 is assigned to an Amazon warehouse in Mississauga, suggesting that Canada Post may be allocating the M0 FSA for high volume addresses.

M1A *Not assigned*	M2A *Not assigned*	M3A North York (Parkwoods)	M4A North York (Victoria Village)	M5A Downtown Toronto (Regent Park / Harbourfront)	M6A North York (Lawrence Manor / Lawrence Heights)	M7A Queen's Park / Ontario Provincial Government	M8A *Not assigned*	M9A Etobicoke (Islington Avenue)
M1B Scarborough (Malvern / Rouge)	M2B *Not assigned*	M3B North York (Don Mills) North	M4B East York (Parkview Hill / Woodbine Gardens)	M5B Downtown Toronto (Garden District, Ryerson)	M6B North York (Glencairn)	M7B *Not assigned*	M8B *Not assigned*	M9B Etobicoke (West Deane Park / Princess Gardens / Martin Grove / Islington

Beautiful Soup package

	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights
4	M6A	North York	Lawrence Manor
5	M7A	Downtown Toronto	Queen's Park
6	M9A	Etobicoke	Islington Avenue
7	M1B	Scarborough	Rouge
8	M1B	Scarborough	Malvern
9	M3B	North York	Don Mills North
10	M4B	East York	Woodbine Gardens
11	M4B	East York	Parkview Hill
12	M5B	Downtown Toronto	Ryerson
13	M5B	Downtown Toronto	Garden District
14	M6B	North York	Glencairn
15	M9B	Etobicoke	Cloverdale
16	M9B	Etobicoke	Islington
17	M9B	Etobicoke	Martin Grove
18	M9B	Etobicoke	Princess Gardens
19	M9B	Etobicoke	West Deane Park
20	M1C	Scarborough	Highland Creek
21	M1C	Scarborough	Rouge Hill
22	M1C	Scarborough	Port Union

Data sets

Geolocation Data

- NY_Neighborhood_Locations.csv
- Toronto_Neighborhood_Locations.csv

One hot coded with Neighbourhood and nearby venues

- NY_grouped.csv
- Toronto_grouped.csv

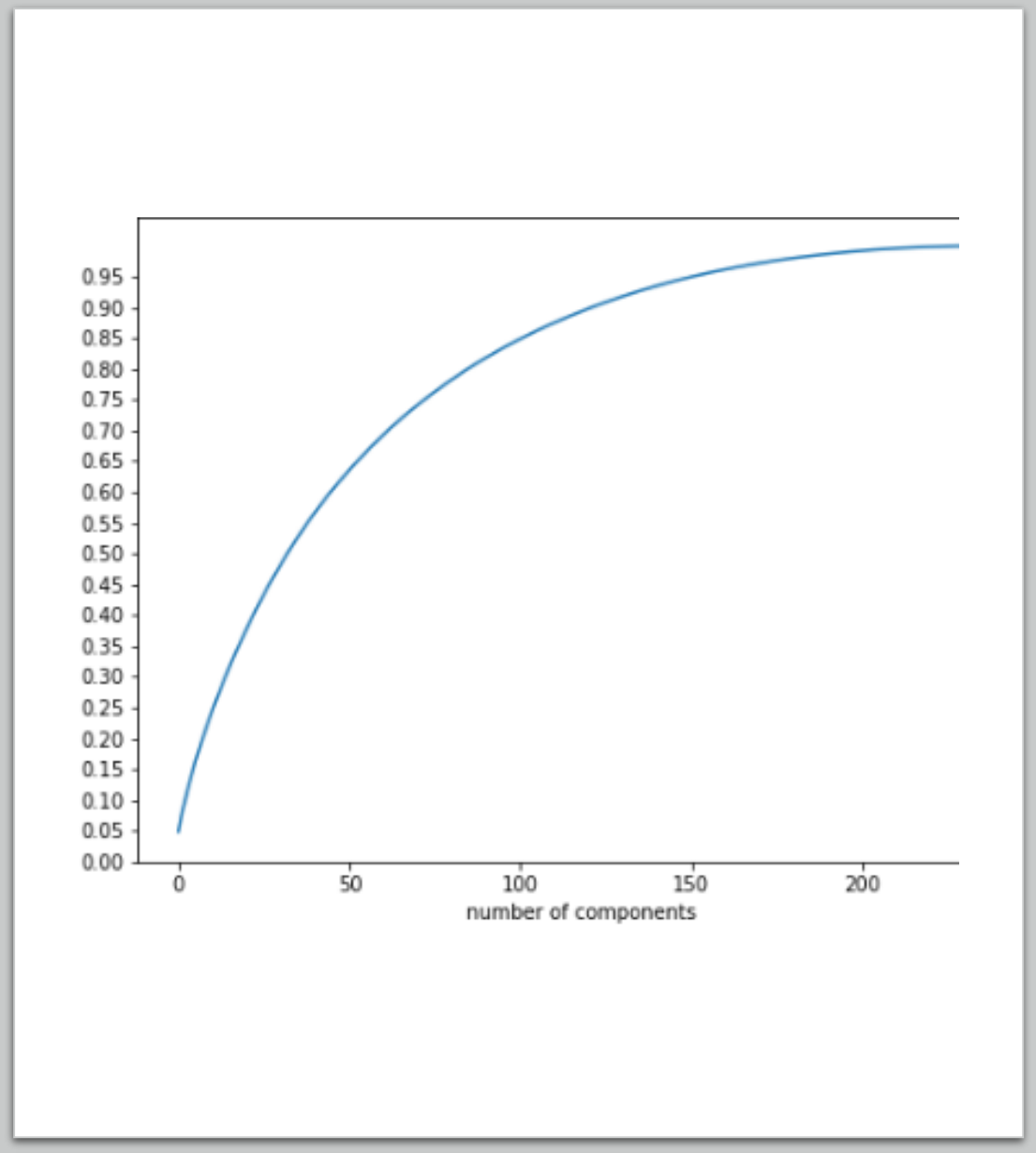
Packages used for Analysis

In [1]:

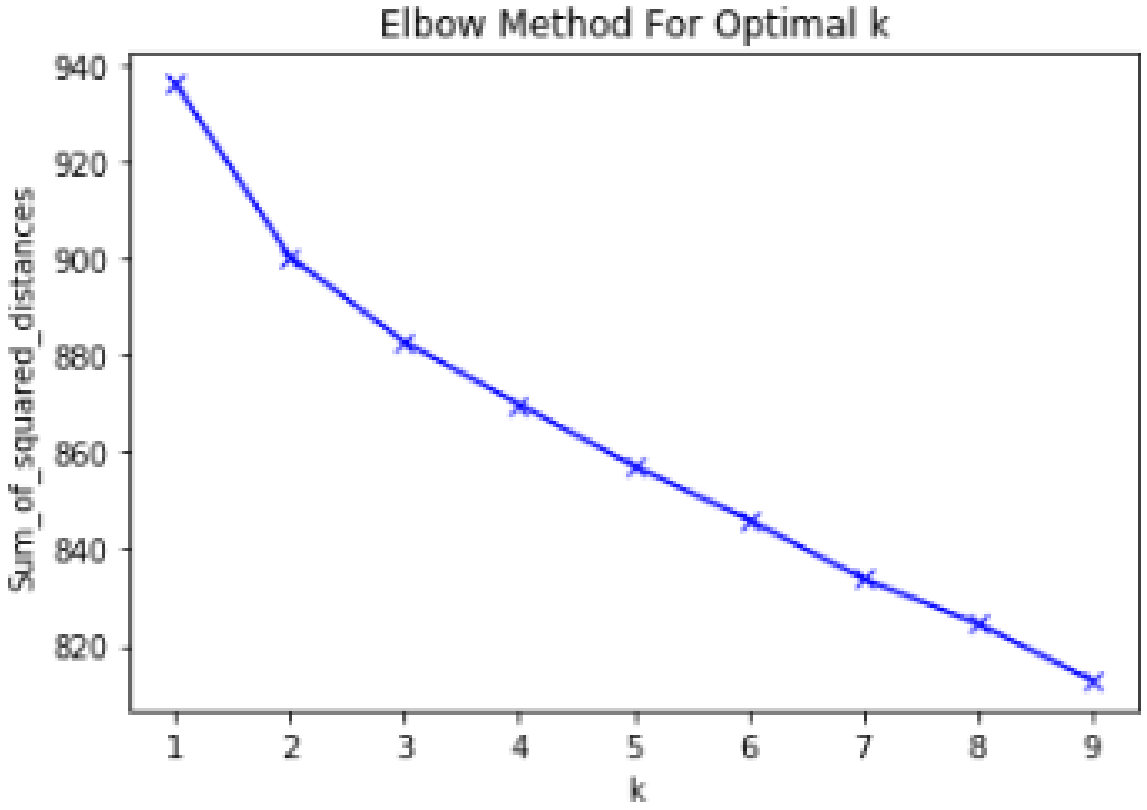
```
import pandas as pd
import numpy as np

import sklearn
from geopy.geocoders import Nominatim
import folium
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import silhouette_samples
from sklearn.metrics import silhouette_score
```

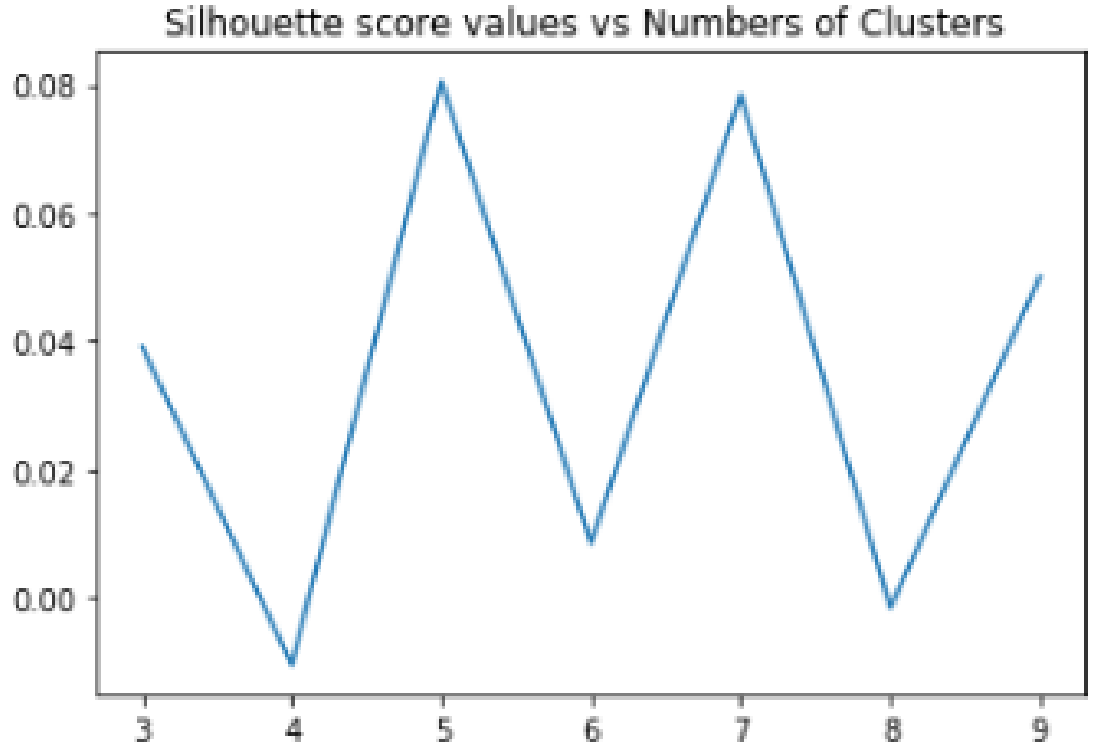
PCA

- Principal Component Analysis is a dimension reduction tool that can be used to reduce a large set of variables to a smaller set by grouping the similar variables that explains the whole dataset.
- With the use of MinMaxScalar functionality of Scikit learn preprocessing, it scales the dataset such as all features values are in the range {0,1}. This retains the variance of the dataset.
- With this it was safe to reduce the number of components to 175 retaining the variance.



K-Means(Optimal K clusters)





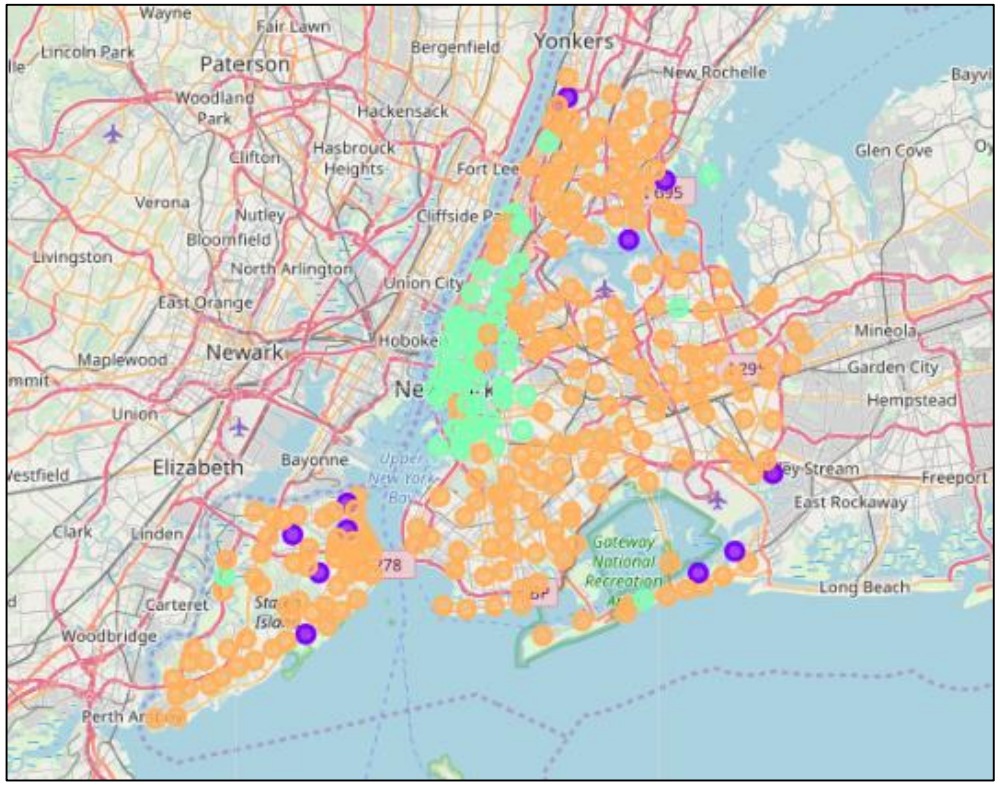

```
Optimal number of components is:
5
```

Results

NY before clustering

Neighborhood Locations

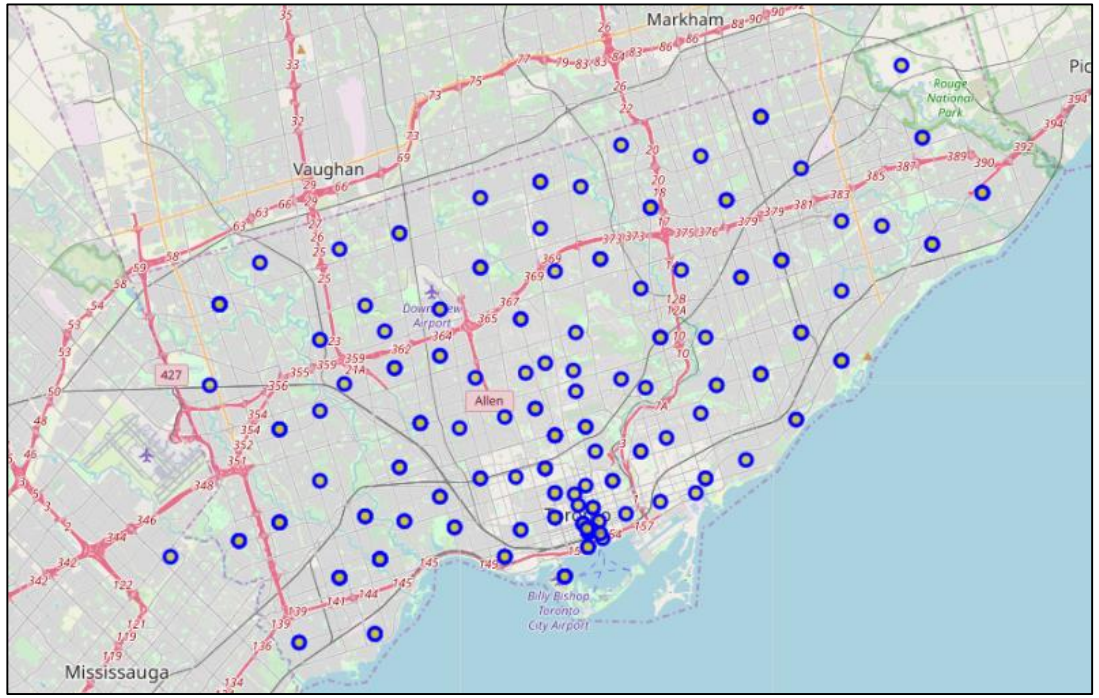
NY after clustering



Similar colors represent same clusters

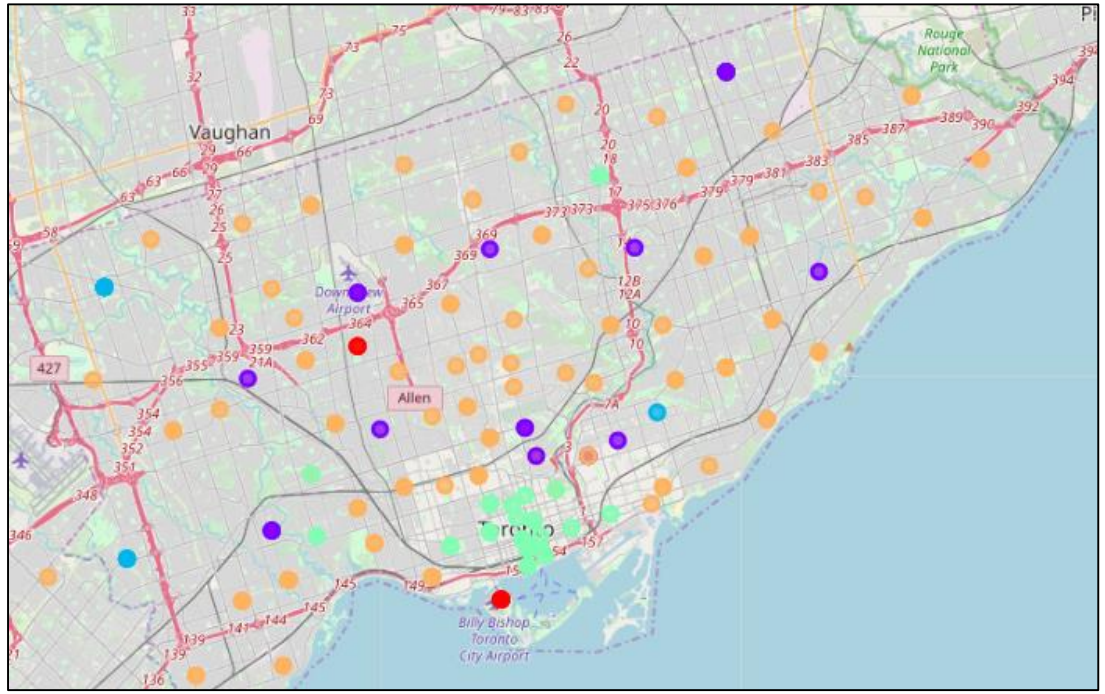
Results

Toronto before clustering



Neighborhood Locations

Toronto after clustering



Similar colors represent same clusters

Results

We can see how the same clusters in Toronto and New York City represent how these places are similar being in different cities

We can also see there are few clusters which are not available in New York city which can be unique to Toronto

Conclusion

- People travelling from one country to another can choose to stay in places that they are habitual
- It would benefit them with convenience and understanding of where to stay next.