Introduction.

Background

A fortune 500 company wants to establish their headquarters in one of the cities New York or Toronto. The company needs insights on neighbourhoods and local business in these cities, which will provide maximum revenue as well as provide their employees a quality living. Despite the dissimilarities, the different venues in the city can be segmented based on different categories. Later on these can be grouped by the neighbourhood so that they will all have similar kind of neighbourhood.

Problem Statement

- Compare the neighbourhoods of the two cities (**New York City and Toronto**) and determine how similar or dissimilar they are.
- A company wants to start a new business and needs recommendations on the preferable city (**New York City or Toronto**) based on similar neighbourhoods.

Data Description

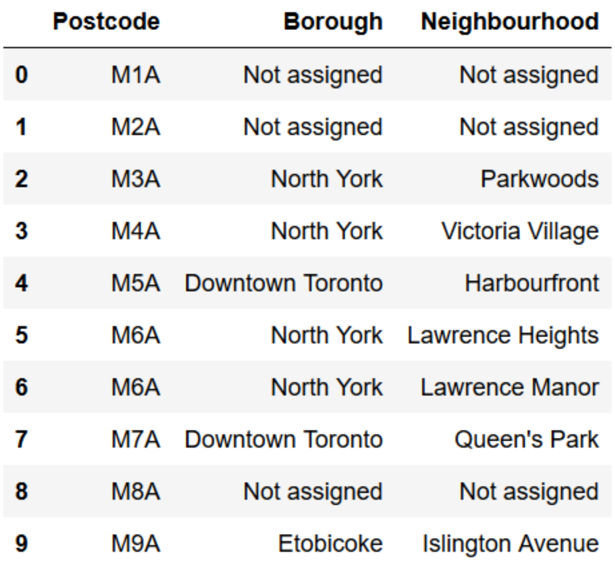
Data Source

The source of the data for both the cities is from the Wikipedia website.

Toronto City, Canada

The link for Toronto city Canada (**https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M**) has the postal code details along with Neighbourhood and Borough details.

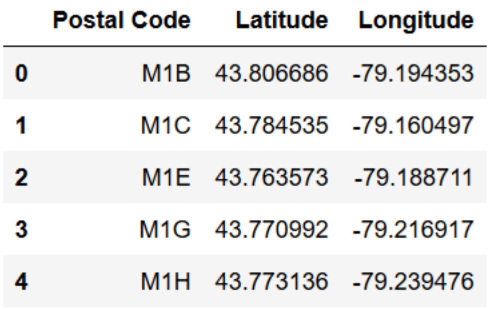
Toronto Postal Data



	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights
6	M6A	North York	Lawrence Manor
7	M7A	Downtown Toronto	Queen's Park
8	M8A	Not assigned	Not assigned
9	M9A	Etobicoke	Islington Avenue

While a second dataset has the geospatial data which will be used to extract the geospatial details of the postal code from the first file.

Toronto Geospatial Data

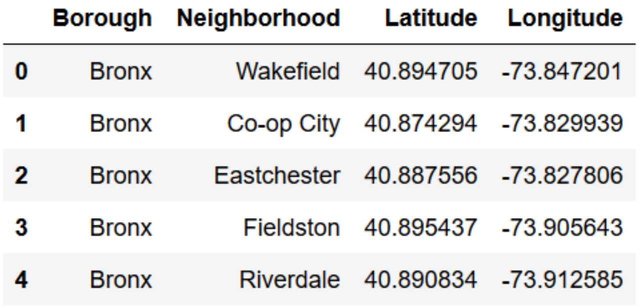


	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

New York City

The New York data is within the Json file. The data from the Json file will be extracted with the information of the borough and geospatial details of the Neighbourhood in New York.

New York Neighbourhood data



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Data processing

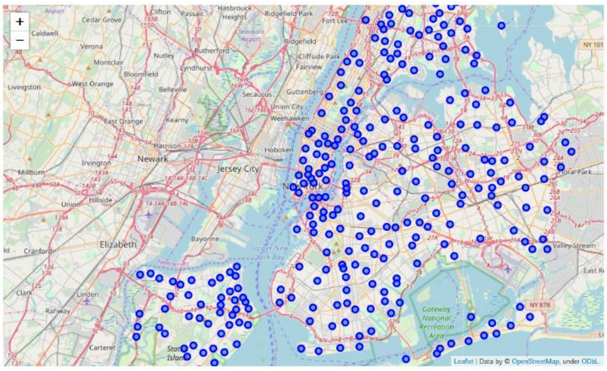
The Toronto data from the Wikipedia website has many rows with “Not assigned” values in borough and Neighbourhood columns, these rows will be considered as missing values and will not be considered for further analysis.In order to make the data complete for the Toronto city, the geospatial details will be extracted from the second file.

The details of New York city will be extracted from the Json file along with the geospatial details.

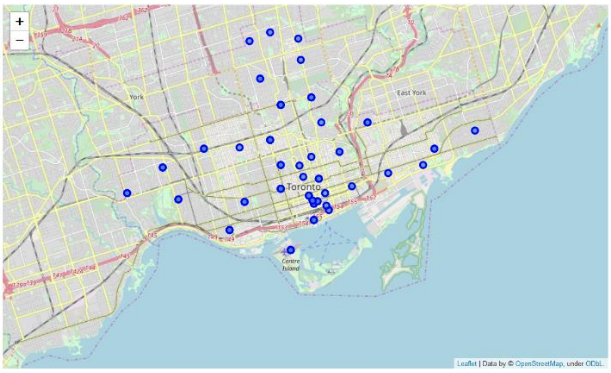
These geospatial data for each borough will be used to request the information of the venues using foursquare api. The foursquare api will require the client id, client secret, version , geospatial information (latitude and longitude) of borough, radius of area and venue information limit which will be fetched from foursquare api. For our analysis we will keep the radius of the area as 500 and limit as 100.

Based on the geospatial data the boroughs are superimposed on the world map as below

New York Neighbourhood data



Toronto Neighbourhood data



The four square api will provide the information of all the venues in the vicinity (based on the radius provided) of Neighbourhood.

This venue information contains the category information which will be One-Hot encoded. The clustered will be formed based on the grouped neighbourhood data.

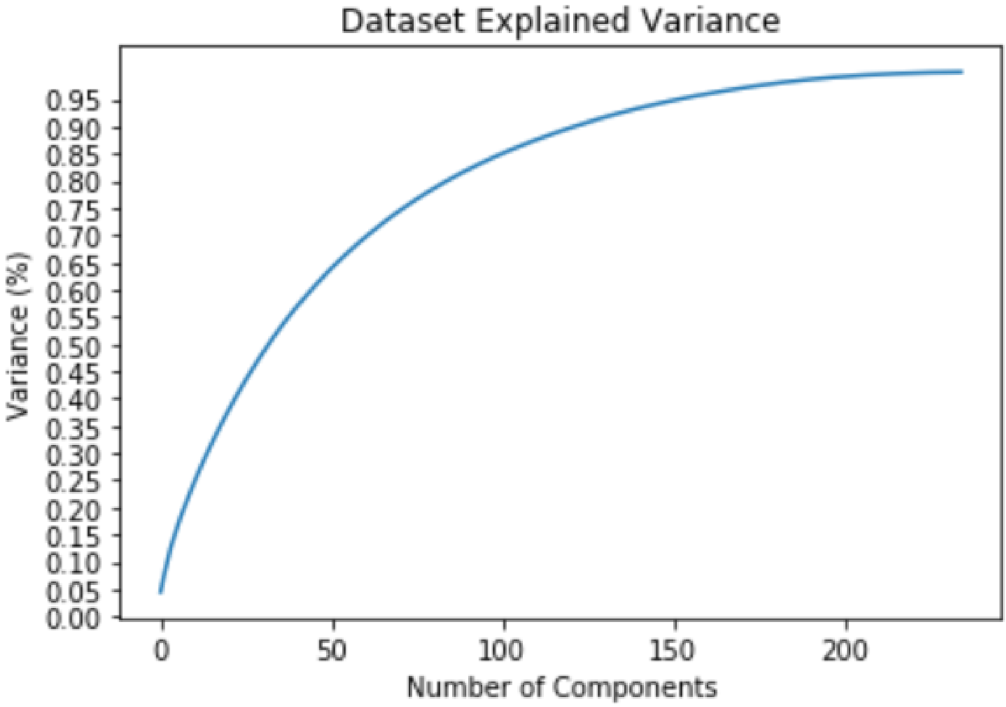
Dimentionality reduction

Before forming the clusters the columns can be reduced by eliminating the columns which have a high coorelation using the Principal Component Analysis (PCA). **Principal component analysis** (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**.

This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

So in order to perform a reasonable covariance analysis we need to normalize this data, putting all attributes in the same unit of measurement (the range between 0 and 1), it will improve the maximization of the variance for each component that our PCA needs to perform.

Explained Variance plot

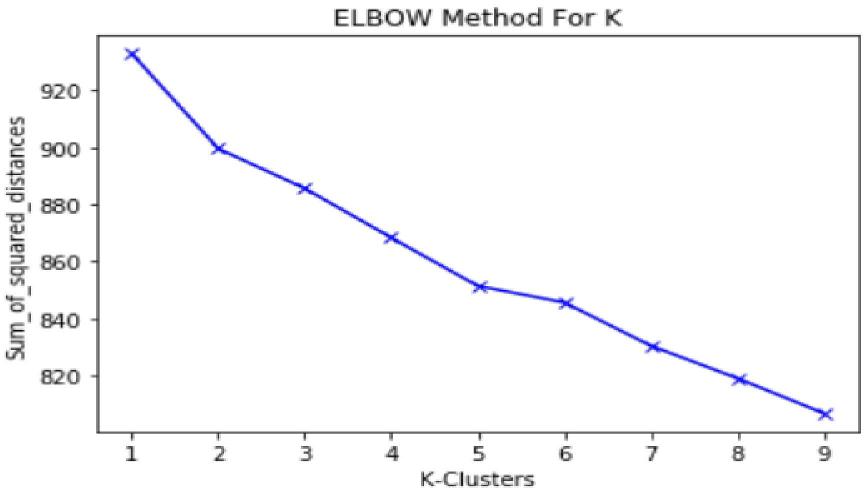


Finally, the number of features is reduced from 250 to 175.

The optimised features will be used to form clusters using the K-means clustering Algorithm. But the most important part is, how many number of clusters should be used.

This can be answered by plotting the variance for corresponding value of number of clusters used for clustering. This plot is known as elbow method to find the optimal K.

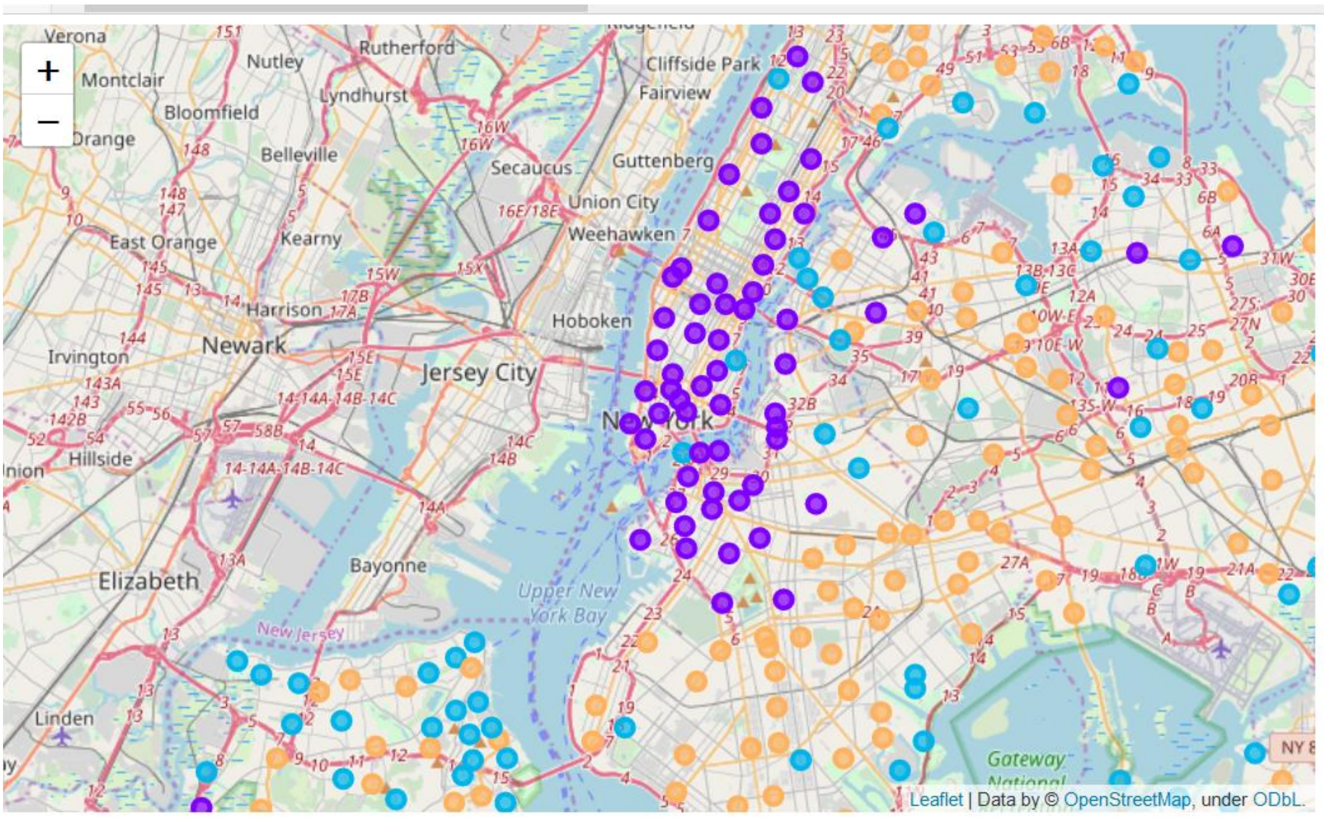
Elbow Method for K



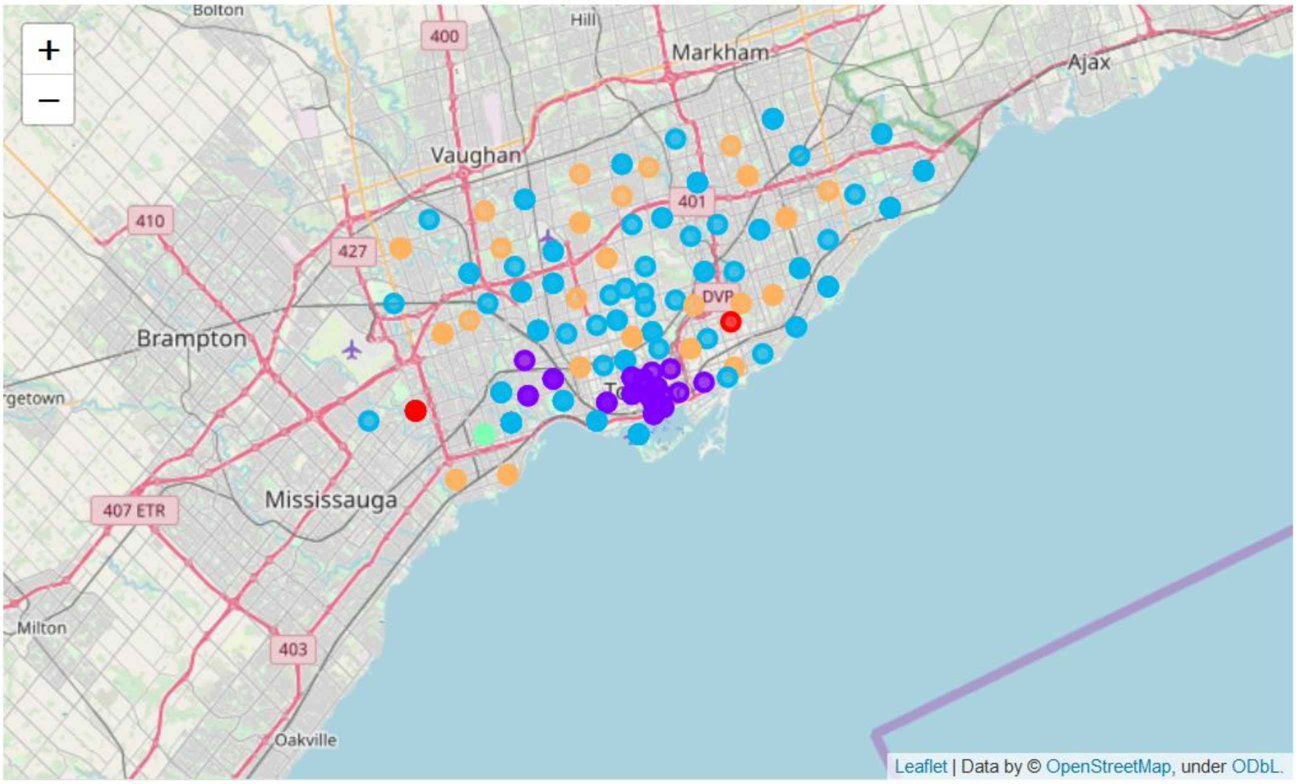
K-Means algorithm is an iterative algorithm that tries to partition the dataset into *K*-pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Finally, these cluster labels will be included in the original neighbourhood data of New York and Toronto to get the neighbourhoods which are similar in some respect.

New York Neighbourhood Clusters



Toronto Neighbourhood Clusters



Methodology

- Location-based services like Foursquare will be used to fetch the venues data based on the geospatial data of the neighbourhood.
- Foursquare API's 'explore' feature will be used to fetch the nearby venues of the neighbourhood of New York and Torornto.
- Folium python visualization library will be used to visualize the neighbourhood clusters distribution of New York and Toronto cities.
- After combining the Neighbourhood of both the cities, due to huge number of features, dimensionality reduction will be applied to the data and the features which are possibly more correlated will be transformed into a set of values of linearly uncorrelated variables.
- After determining the number of clusters required for the Neighbourhood data, the uncorrelated variables will be used to form the clusters using the unsupervised machine learning algorithm K-Means.
- K-means clustering will be applied to form the clusters of different categories of places in and around the neighbourhoods. The clusters from each of these neighbourhood cities will be analyzed collectively to understand the distribution of these venues.

Conclusion

From the cluster plotting using the folium library for both the cities (**New York City and Toronto**) we can see that there are more Business opportunities in **New York** than in Toronto.