Identifying variance of features between Manhattan and Toronto

Rahul S Nair

October 22, 2019

1. Introduction

Cities within countries can sometimes tend to show similarities if they are near to each other. Even then there can be factors like geography, community, etc. which make affect the overall functioning of a city. Manhattan in USA and Toronto located in Canada are very big cities but situated miles apparat. With this project, we are trying to establish the relations, similarities, dissimilarities between the venues located in both of these cities.

2. Data Acquisition and Cleaning

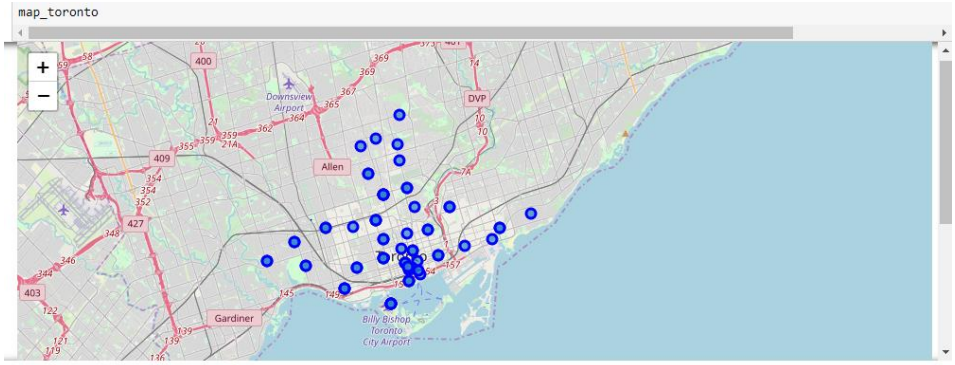
Commented [RSN1]:

Data were collected from Wikipedia and coursera provided links. When analyzed. All the postal codes available in a the respected city were taken. With the help of Foursquare api, the top 10 venues were selected for each of these postal codes.

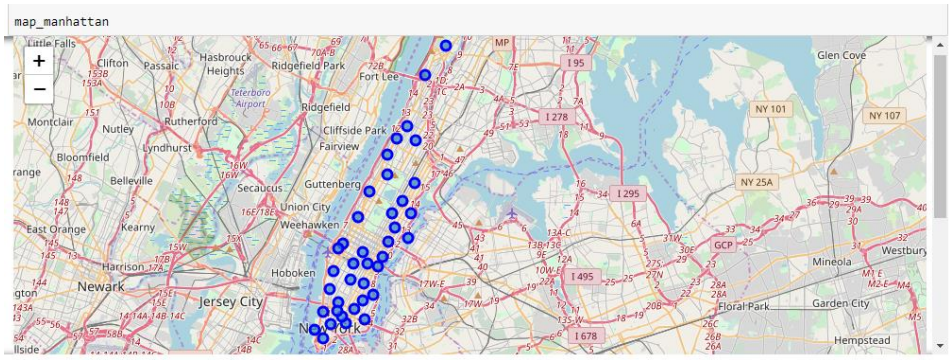
For categorical variables, one hot encoding was applied to extract the respected features. This resulted in a final training set of shape (73,117) in case of Toronto and (400,154) in case of Manhattan.

3. Exploratory Data Analysis

Plot of Neighborhoods in Toronto



Plot of Neighborhoods in Manhattan

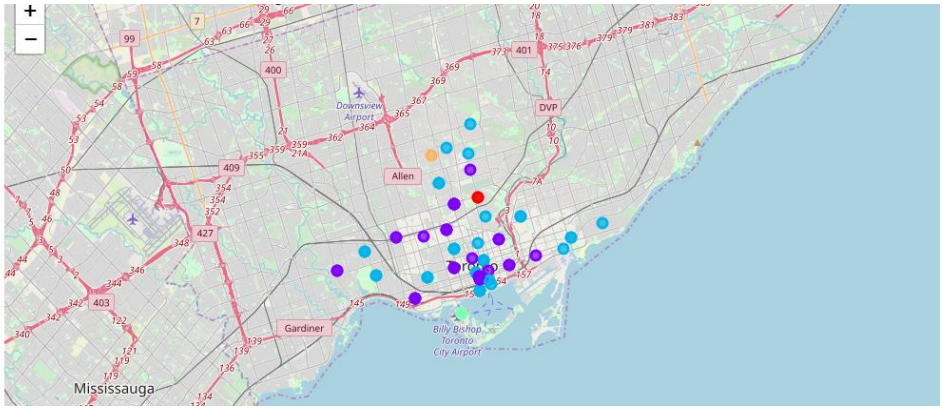


Both cities have different distribution of datapoint. For each of the datapoints to be clustered, rather than going with geographical distance clustering, top 10 venues of each neighborhood are considered to cluster the neighborhoods

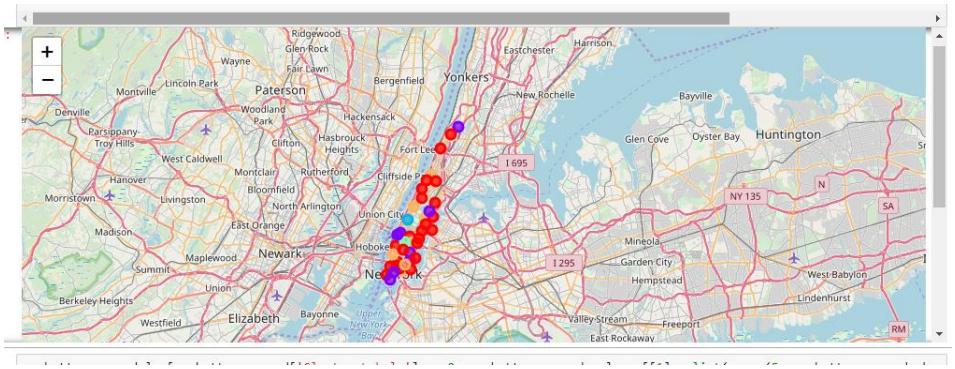
4. Clustering

K NN clustering was applied on both datasets and following data was generated.

Clustering done on Toronto



Clustering done on Manhattan



5. Conclusion

Toronto

More number of coffee shops are present. People tend to drink more coffee

People prefer entertainment venues such as comic book store, parks

Ice cream shops are more in some areas

Manhattan

A Cluster that shows domination on pubs are present.

People prefer entertainment venues such as pubs, parks

Higher number of gym/fitness centers are present