Parks in the Neighbourhood

Mahaveer

October 13, 2019

1. Introduction

1.1 Background

There are approx 10,000 parks in UN and the vists of people to this parks are in millions per year. For refreshment,exercise,playing and yoga every age of person is coming to the park. So, building a park will increase the overall health of the population. So, we need good location for building a park. For a person not familiar with the city it takes a lot of research to select an appropriate location for building a park. Therefore, through this project the tedious task for selecting a location made easy.

1.2 Problem

The problem in establishing a successful business is to select a proper location to set up a new park in Toronto. For example, the person may select a location tht doesn't suit to set up a new park then we have to search and find the good place for that. Considering such challenges there is a requirement for a model which can provide with the right options for park location.

1.3 Interest

The problem is very common and therefore, shares interests with a government. The model would interest government looking to open a park, to increase overall health of the population in a city and need right choice for the target location.

2. Data

The data used in the project is of different format and origins. The types of data worked with in the project are as follows:

1. Web scraped data from the Wikipedia website. The data encompasses mainly the table of Postal code, Borough and Neighbourhoods associated with each. In the project as an example the data of city of Toronto is web scraped from the Wikipedia site. The data is in tabular form but not normalized.

2. The user input given . The first input is the selective neighbourhood in the city. The second input is the type of park the user is willing to make. This is mainly the search query which will be used in the search process of venues using the FourSquare API which provides accurate location data.

3. Along with the locations in the city the coordinates of those locations are obtained via the geopy library.

4. The list of venues , their category , their latitude and longitude with respect to search query are provided via the FourSquare API.

3. Exploratory Data Analysis

In this part the proper analysis of the type and origin of data is performed. The relationship between different data is explored. The relationship between the attributes neighbourhood and venues is examined via the frequency of the number of venues present at a particular neighbourhood of the same post code. The same type of neighbourhoods are grouped together in the data set. The neighbourhoods containing high number of search query venues are least likely to be included in the result.

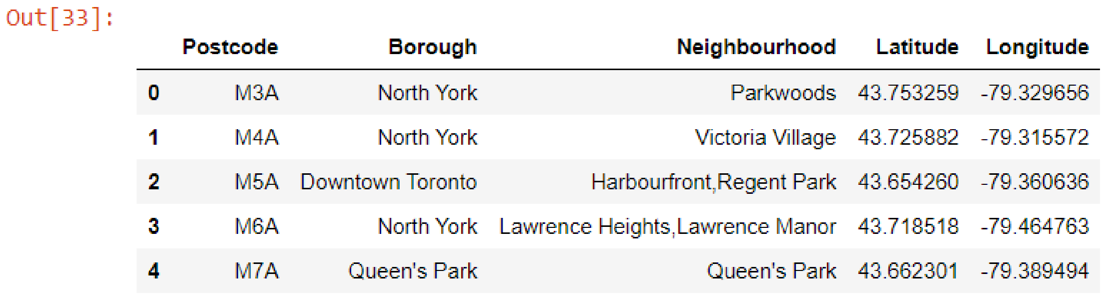
4. Analytic Approach

It is important because it helps identify what type of patterns will be needed to address the question most effectively. In this project, the problem is to search for appropriate locations in the neighbourhood to built a park. As the problem deals with exploring relationships between different factors therefore, a descriptive approach where clusters of similar data based on the number of venues in the neighbourhood and preferences are examined, would be the right analytic approach.

5. Data Understanding

In this part the proper understanding of the type and origin of data is performed. The data from different resources and different types is cleaned and transformed and then combined into dataframe. For example, the web scraping is done via the Wikipedia library in python. The data is then transformed and converted into a Pandas Data frame.

The location coordinates retrieved for each postal code are then added to the original data frame. The final data frame would look like this.



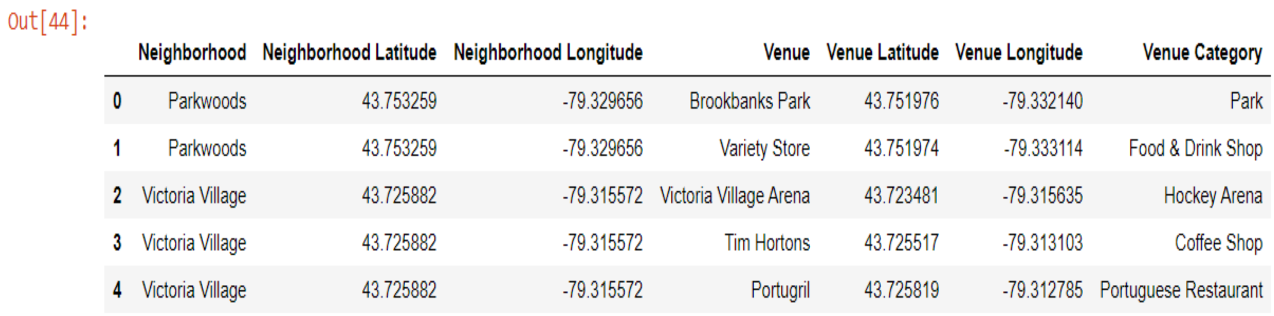
Out[33]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

6. Data Preparation

The FourSquare API is used in the retrieval of venues located in every neighbourhood. The search query is passed in the URL providing the result. The result which is a json file is converted to dataframe and then the resultant dataframe is the combination of the neighbourhood dataframe and the venues and venue categories associated to each.

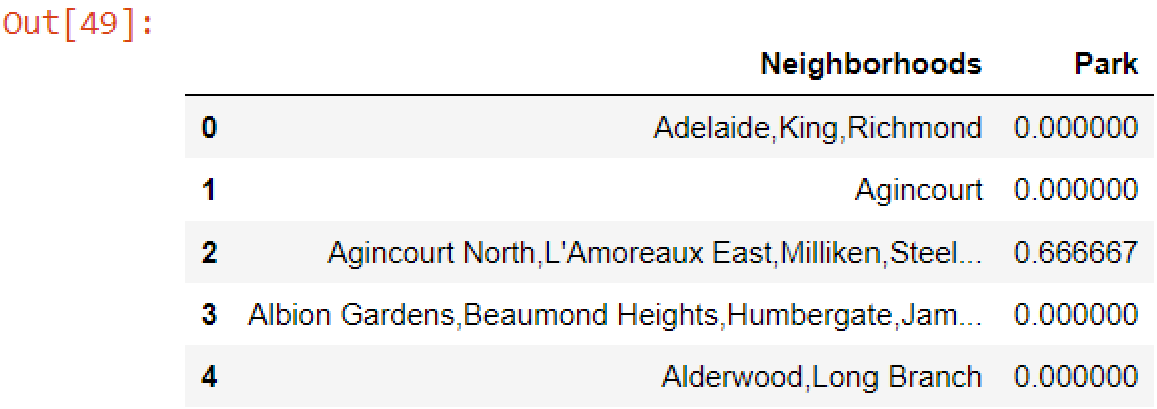
The dataframe might contain missing values which can be dropped . The resulting dataframe would be like this :



Out[44]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

The dataset contains the venues that belong to a same category and in the same neighbourhood. Therefore, group the data of the same neighbourhood and frequency count the number of such rows. The resulting data frame would be like this:



Out[49]:

	Neighborhoods	Park
0	Adelaide,King,Richmond	0.000000
1	Agincourt	0.000000
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	0.666667
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	0.000000
4	Alderwood,Long Branch	0.000000

7. Descriptive Modelling

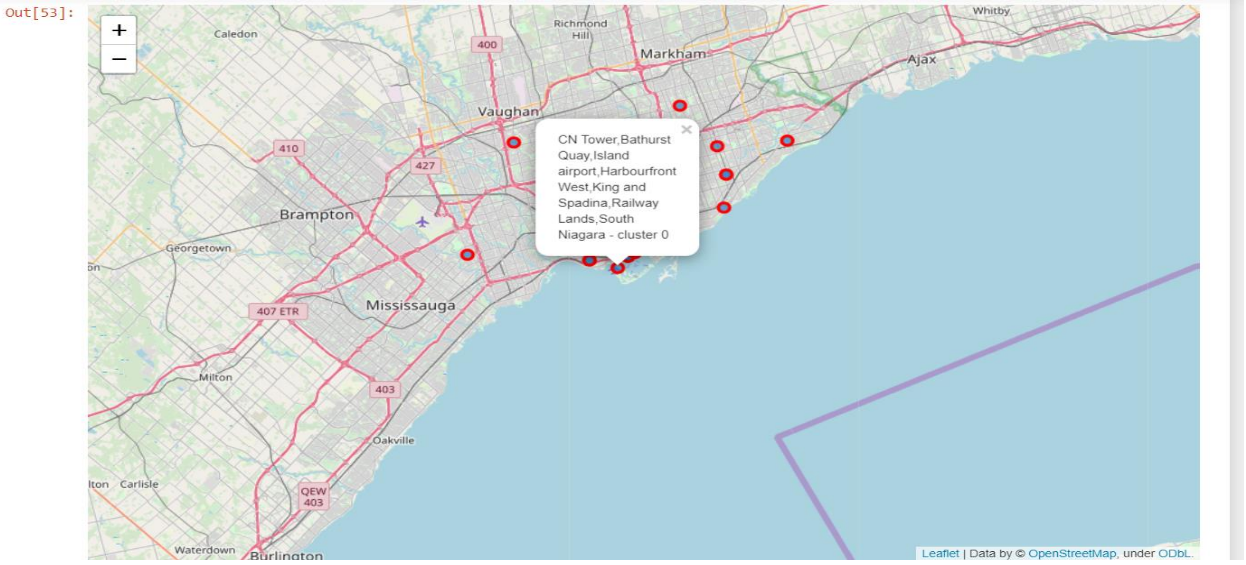
As discussed in the analytic approach section , the model used in the project would be descriptive model. The algorithm used in the procedure is the K-means clustering algorithm.

The number of clusters are 3 considering the suitable number of venues in the entire borough. The clusters are then plotted along the original map of the borough. In the evaluation phase, the clusters having the maximum number of venues and those having least number of venues are eliminated . This is because the objective of this project is to ensure the right options to built a park. And the neighborhoods containing large number of such venues would result in good natural environment and the neighborhoods having least number of venues considering (0,1,2) would be not feasible for building a park due to various different reasons. Therefore, the most right option is to select the neighborhoods containing an average number of such venues. And the result of right location for park depends on so many different factors including domain expertise. Therefore, this is the best option.

7. Result

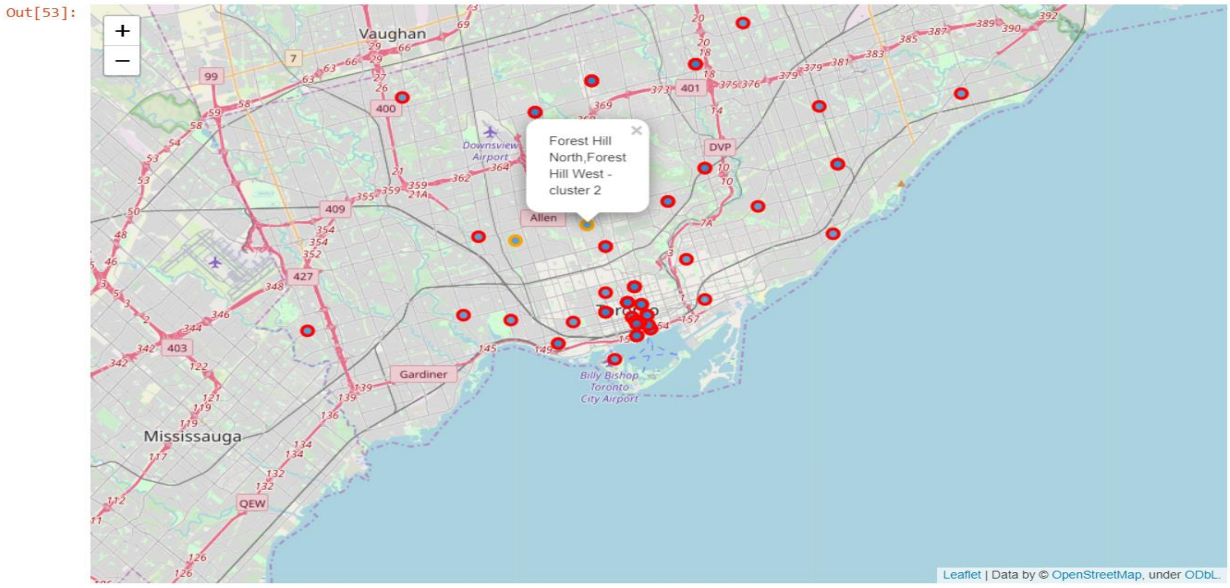
The final result would be a map of borough built using library folium which on which the clusters are superimposed. Apart from this the result would contain the dataframe with appropriate locations along the borough and their respective location coordinates.

For example :



Out[53]:

CN Tower,Bathurst Quay,Island airport,Harbourfront West,King and Spadina,Railway Lands,South Niagara - cluster 0

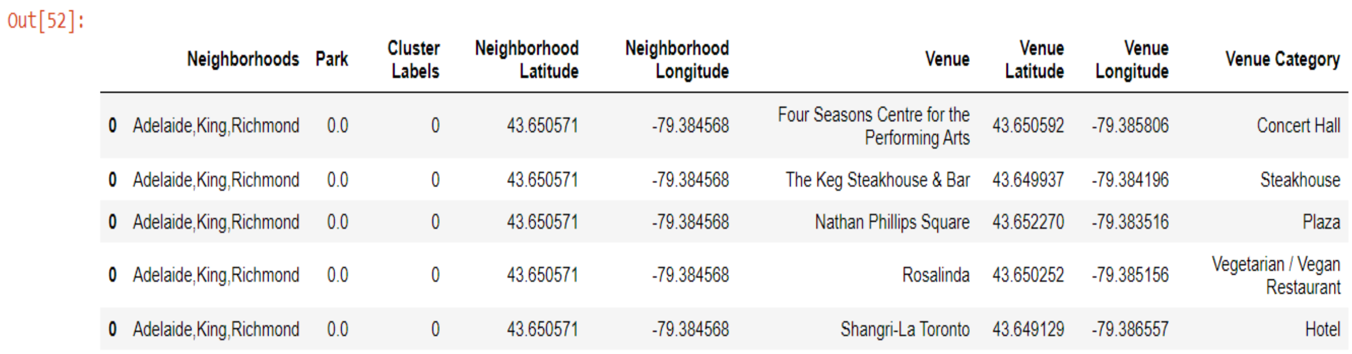


Out[53]:

Forest Hill North,Forest Hill West - cluster 2

+
−
Vaughan
Downsview Airport
Allen
DVP
401
409
427
400
403
Gardiner
QEW
Mississauga
Billy Bishop Toronto City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.

Along with the mapped data the data in the tabular format is also output of the model.



Out[52]:

	Neighborhoods	Park	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adelaide,King,Richmond	0.0	0	43.650571	-79.384568	Four Seasons Centre for the Performing Arts	43.650592	-79.385806	Concert Hall
0	Adelaide,King,Richmond	0.0	0	43.650571	-79.384568	The Keg Steakhouse & Bar	43.649937	-79.384196	Steakhouse
0	Adelaide,King,Richmond	0.0	0	43.650571	-79.384568	Nathan Phillips Square	43.652270	-79.383516	Plaza
0	Adelaide,King,Richmond	0.0	0	43.650571	-79.384568	Rosalinda	43.650252	-79.385156	Vegetarian / Vegan Restaurant
0	Adelaide,King,Richmond	0.0	0	43.650571	-79.384568	Shangri-La Toronto	43.649129	-79.386557	Hotel

8. Conclusion

The project is based on the problem which a government would face when searching for a proper location for building a park. The model build is a descriptive model defining relationships between data i.e. neighbourhoods containing different number of venues. Where the same type of neighbourhoods form one cluster . The objective is to provide the safest possible options to the government for the location. The modelling is done via the K-means clustering algorithm of ML. The user finally gets the valuable output in the form of map pointing to different clusters of neighbourhoods along with the tabular data of the list of neighbourhoods safe for building park.