Ideal Neighborhoods for Parks in York, Canada

by

Rithwik

1. Introduction /Business Problem

➢ Objective

- Our aim is to find the neighborhoods that do not have parks in York, Canada. We also aim to find the neighborhoods that already have parks in their neighborhood, which can be enjoyed by the residents.

➢ Purpose

- This analysis will be useful to those who are planning to construct a park based on whether the neighborhoods in York, Canada already contain an existing park or not. It will also be beneficial to those choosing a residential or office space depending on which neighborhoods have a Park or not.
- This will help the people who seek or are trying to avoid parks in the vicinity of their neighborhood, be it for professional reasons or for private reasons.
- Constructing parks in areas where parks don't already exist would make more sense as it will bring in more residents or people which could include the youth, families, pet owners, fitness enthusiasts or just regular folks who are trying to get away from the digital world for a bit.
- It also allows people in York looking to avoid parks in their vicinity which may be due to health reasons such as allergies or the increased rent due to the availability of such amenities or maybe just a personal preference altogether.

➢ Location

- York is a former city within the current city of Toronto, Ontario, Canada. It is located northwest of Old Toronto, southwest of North York and east of Etobicoke, where it is bounded by the Humber River. As a separate city, it was one of six municipalities that amalgamated in 1998 to form the current city of Toronto.

- In 1998 it amalgamated with the cities of Toronto, Etobicoke, Scarborough, and North York and the borough of East York to form the City of Toronto. York was established as a borough in 1967, through the amalgamation of the township of York and the town of Weston (incorporated 1881). It was incorporated as a city in 1983. The original York Township was formed in 1793, and it was once called Dublin.

2. Data Collection

➢ Postal Codes of Canada

- The data which consists of the Postal codes of Canada starting with the letter "m" is t aken from the Wikipedia page: "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
- The data is scraped from a Wikipedia page and then loaded into a Pandas dataframe. This is done using the "read_html" method.

➢ Geospatial data

- The Geospatial data is available in the form of a CSV file in "https://cocl.us/Geospatial_data".

  We extract the data from CSV file to panda dataframe. It contains the Latitude and lo ngitude data against the postal codes which are required to identify the location on the map s.

➢ Foursquare API:

- The Foursquare Places API provides location-based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag.
- Additionally, foursquare allows developers to build audience segments for analysis an d measurement. JSON is the preferred response format.

- The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there's really nothing to download onto your server.
- There are both GET and POST methods you can use, which means you aren't just limited to reading from feeds, you can also do useful stuff like checking in and creating locations using the API.
- As for rate limits, your application is limited to 200 requests per hour per method, so you probably want to implement some kind of result caching to "play nice."

3. Machine Learning Algorithm Used

➢ One Hot Encoding

- For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.
- In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).
- In this case, a onehot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.
- Here we use onehot encoding of the data using the get_dummies () method. Then we group the new onehot encoded dataset by neighborhood. This is done to find out the common venues in each neighborhood by their means. This helps to find out which neighborhoods have pre-built parks (mean > 0) and which neighborhoods do not.

➢ K- Means Clustering

- Kmeans clustering is one of the simplest and popular unsupervised machine learning algorithms.

- The objective of Kmeans is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset."
- A cluster refers to a collection of data points aggregated together because of certain similarities.
- You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.
- Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.
- In other words, the Kmeans algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
- The 'means' in the Kmeans refers to averaging of the data; that is,finding the centroid
- Here, we use k -means to cluster the neighnorhoods with and without parks.

4. Libraries Used:

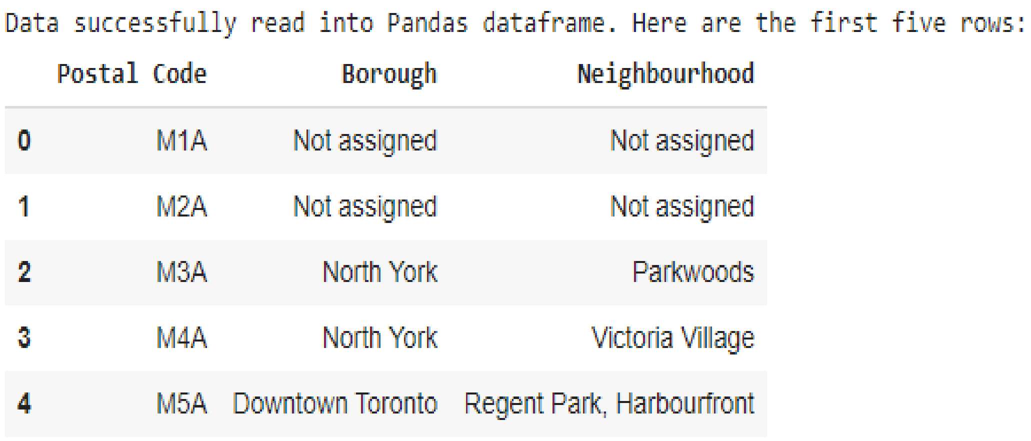
- **Pandas, NumPy:** for data manipulation and efficient handling
- **Matplotlib, Seaborn:** Data visualization and plotting libraries
- **Folium:** Python library used for visualizing geospatial data.
- **Geocoder:** Python library to convert an address into latitude and longitude values
- **JSON:** Library to handle and parse JSON files
- **Requests**: Library to handle requests

5. Procedure

a. Importing the necessary libraries such as pandas, NumPy matplotlib, seaborn, folium, geocoder, JSON, Requests

b. Introduce the data source which is the postal codes of Canada which is scrapped from the Wikipedia page "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

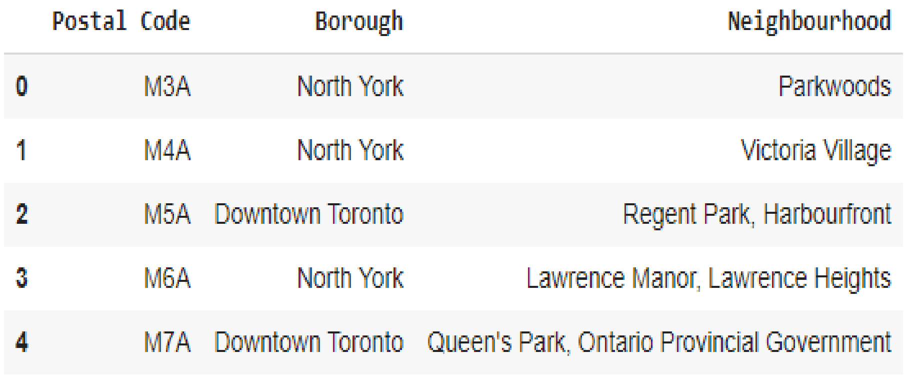
c. This new dataset is then put into pandas data frame. It looks as below



```
Data successfully read into Pandas dataframe. Here are the first five rows:
```

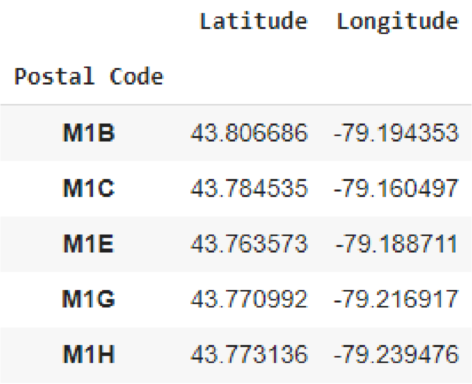
	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

d. Since the data contains multiple rows where the Borough values are “Not assigned “, we need to clean the data as it has no value to us. After cleaning the data, it looks like this



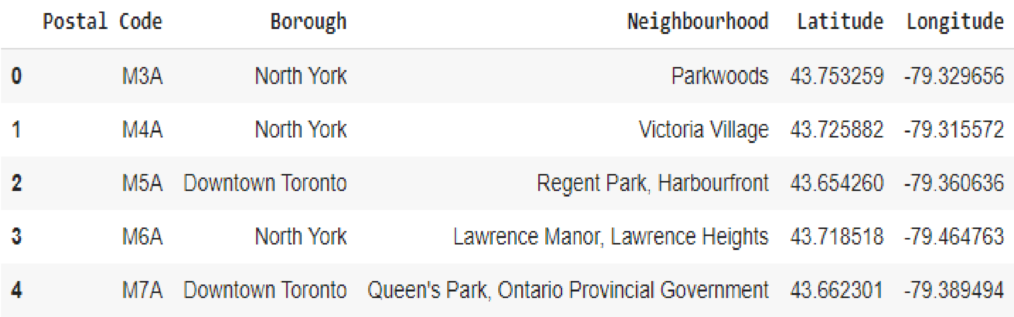
	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

e. We then read the Latitude and longitude values from CSV file containing geospatial data against their postal codes.



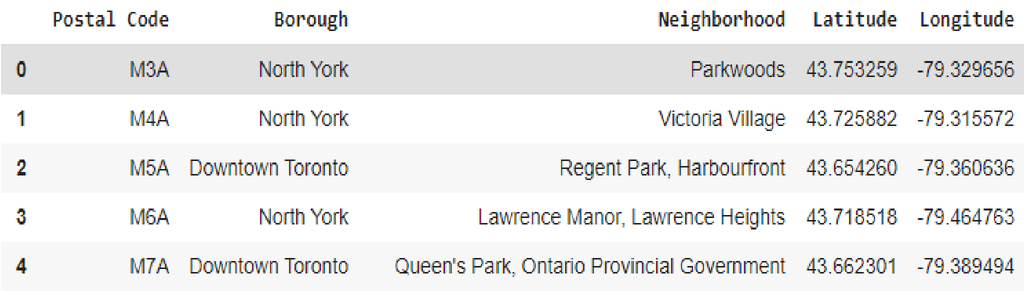
	Latitude	Longitude
Postal Code		
M1B	43.806686	-79.194353
M1C	43.784535	-79.160497
M1E	43.763573	-79.188711
M1G	43.770992	-79.216917
M1H	43.773136	-79.239476

f. Next, we merge the 2 data sets based on the postal codes to give the below table



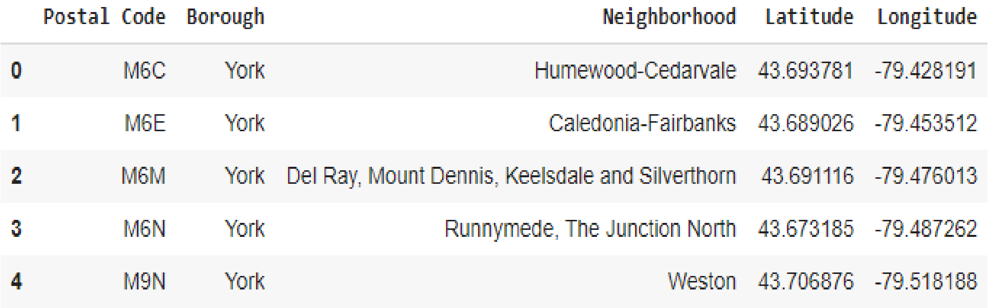
	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A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

g. Renaming neighbourhood to neighborhood (part of cleaning data)



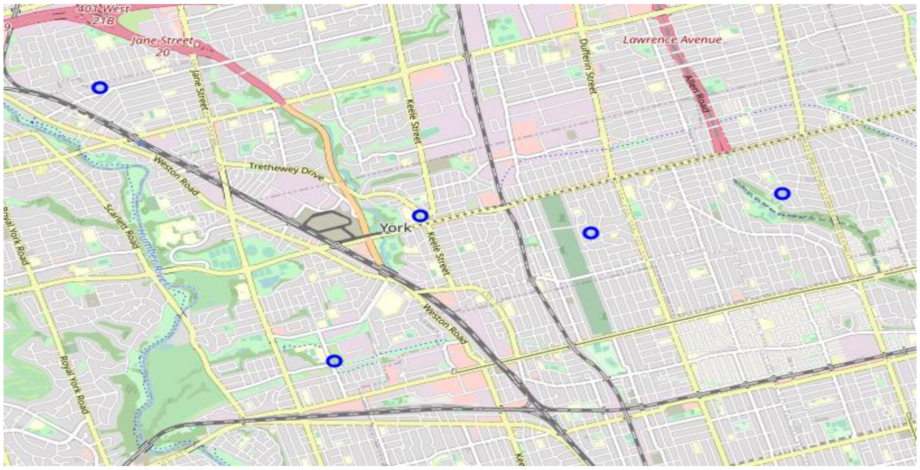
	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A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

h. Our next move is to take only data in regards to the location we chose: York, Canada. We see that we have only 5 neigborhoods in York.

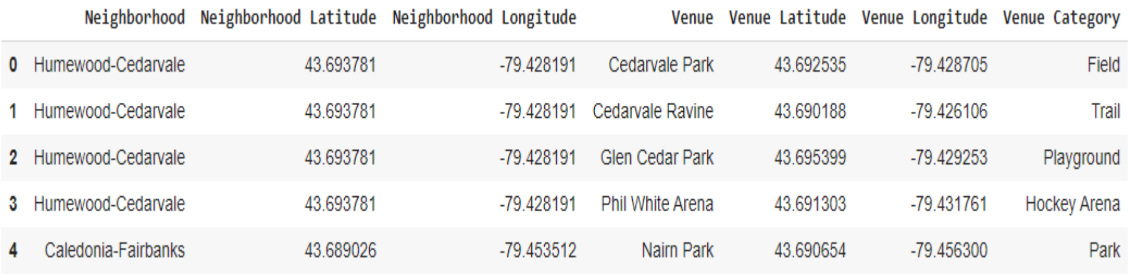


	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M6C	York	Humewood-Cedarvale	43.693781	-79.428191
1	M6E	York	Caledonia-Fairbanks	43.689026	-79.453512
2	M6M	York	Del Ray, Mount Dennis, Keelsdale and Silverthorn	43.691116	-79.476013
3	M6N	York	Runnymede, The Junction North	43.673185	-79.487262
4	M9N	York	Weston	43.706876	-79.518188

i. Our next step is to visualize this using the data we have so far and making use of the folium library and it looks like this.

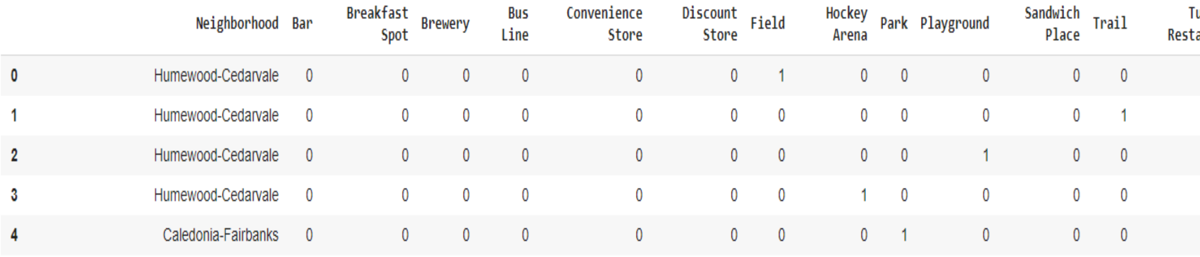


j. We then make use of the Foursquare API to get venue data. We use the credentials for foursquare and match the venues data with the neighborhood data. We also find that there are 14 unique categories



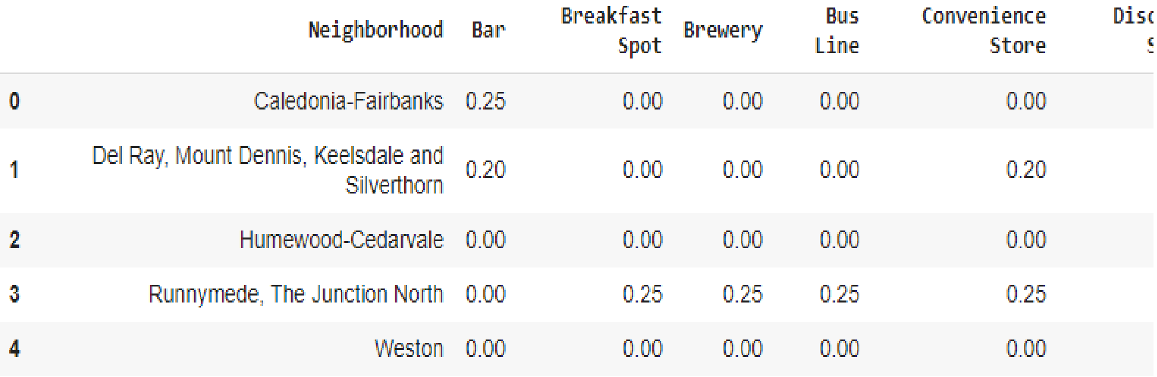
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Humewood-Cedarvale	43.693781	-79.428191	Cedarvale Park	43.692535	-79.428705	Field
1	Humewood-Cedarvale	43.693781	-79.428191	Cedarvale Ravine	43.690188	-79.426106	Trail
2	Humewood-Cedarvale	43.693781	-79.428191	Glen Cedar Park	43.695399	-79.429253	Playground
3	Humewood-Cedarvale	43.693781	-79.428191	Phil White Arena	43.691303	-79.431761	Hockey Arena
4	Caledonia-Fairbanks	43.689026	-79.453512	Nairn Park	43.690654	-79.456300	Park

k. Now, we analyze each neighborhood by using one hot encoding technique to get a data set that looks like this. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.



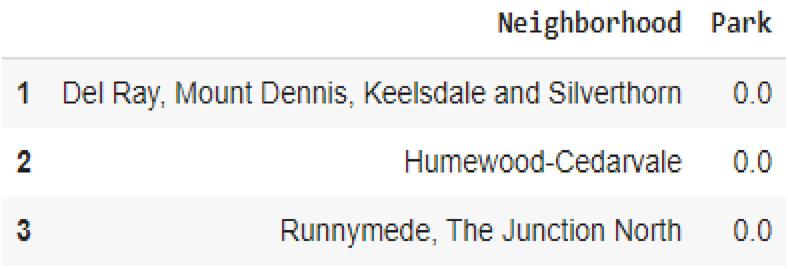
	Neighborhood	Bar	Breakfast Spot	Brewery	Bus Line	Convenience Store	Discount Store	Field	Hockey Arena	Park	Playground	Sandwich Place	Trail	Tu Resta
0	Humewood-Cedarvale	0	0	0	0	0	0	1	0	0	0	0	0	
1	Humewood-Cedarvale	0	0	0	0	0	0	0	0	0	0	0	1	
2	Humewood-Cedarvale	0	0	0	0	0	0	0	0	0	1	0	0	
3	Humewood-Cedarvale	0	0	0	0	0	0	0	1	0	0	0	0	
4	Caledonia-Fairbanks	0	0	0	0	0	0	0	0	1	0	0	0	

l. Furthermore, we group rows by neighborhood and by taking the mean of the frequency of occurrence of each category.



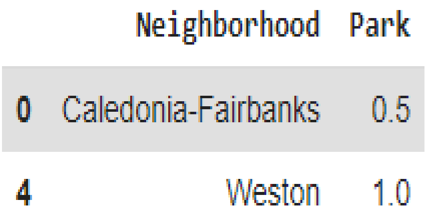
	Neighborhood	Bar	Breakfast Spot	Brewery	Bus Line	Convenience Store	Disc S
0	Caledonia-Fairbanks	0.25	0.00	0.00	0.00	0.00	
1	Del Ray, Mount Dennis, Keelsdale and Silverthorn	0.20	0.00	0.00	0.00	0.20	
2	Humewood-Cedarvale	0.00	0.00	0.00	0.00	0.00	
3	Runnymede, The Junction North	0.00	0.25	0.25	0.25	0.25	
4	Weston	0.00	0.00	0.00	0.00	0.00	

m. We can now search for all the neighborhoods without a Park as we only have to check the data where the mean value of park equal to 0



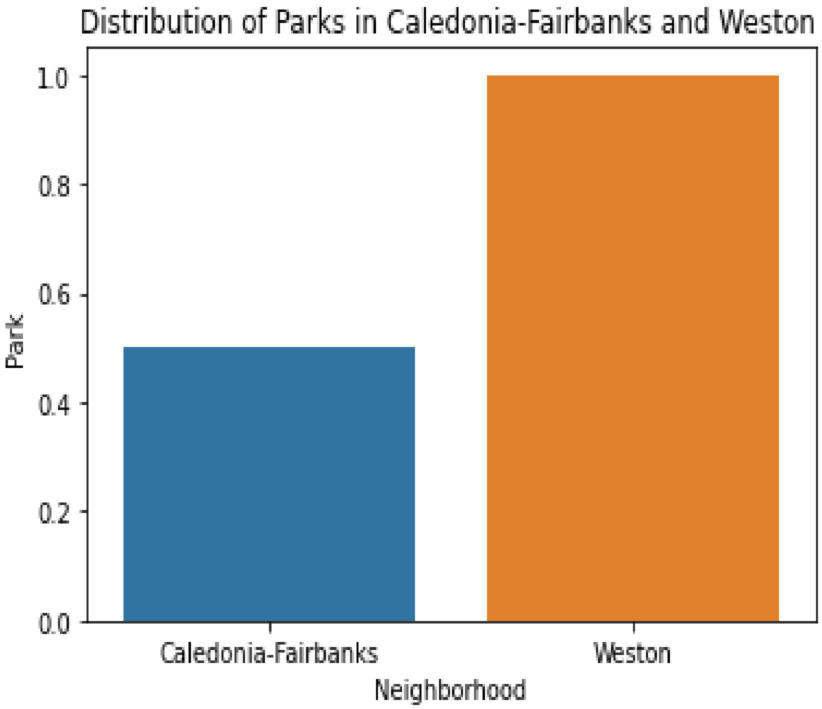
	Neighborhood	Park
1	Del Ray, Mount Dennis, Keelsdale and Silverthorn	0.0
2	Humewood-Cedarvale	0.0
3	Runnymede, The Junction North	0.0

n. Similarly, the neighborhoods where the Parks already exist have mean value of park more than or equal to 0.01



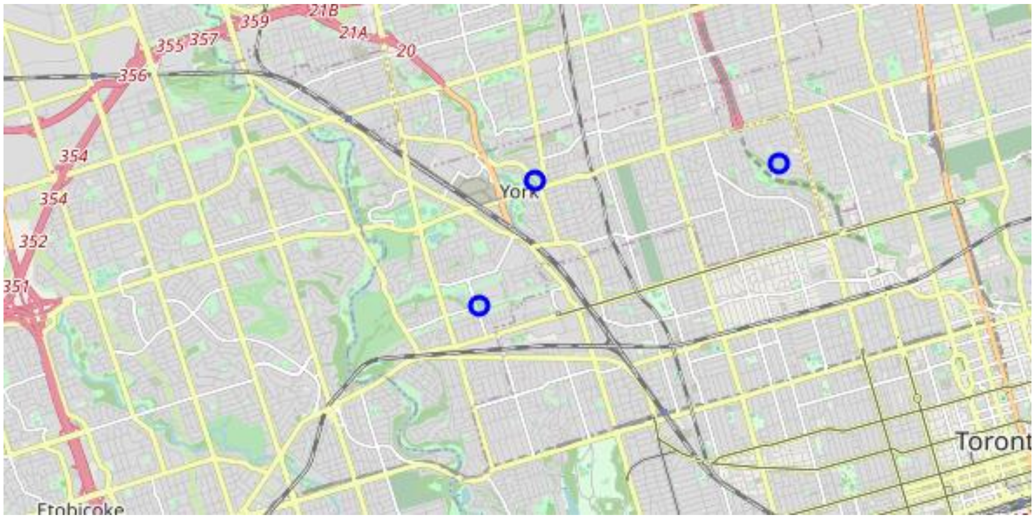
	Neighborhood	Park
0	Caledonia-Fairbanks	0.5
4	Weston	1.0

o. We see that only 2 neighborhoods have pre-built parks. We then visualize it to see the distribution of Parks in these neighborhoods. We use a bar graph for the same.

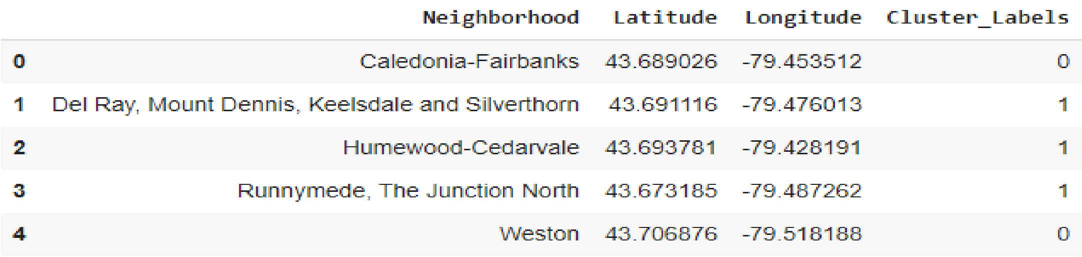


p. We observe from the graph that **Weston** has more Parks than **Caledonia-Fairbanks.**

q. If we visualize the, neighborhoods that lack parks on a map then it would appear like this

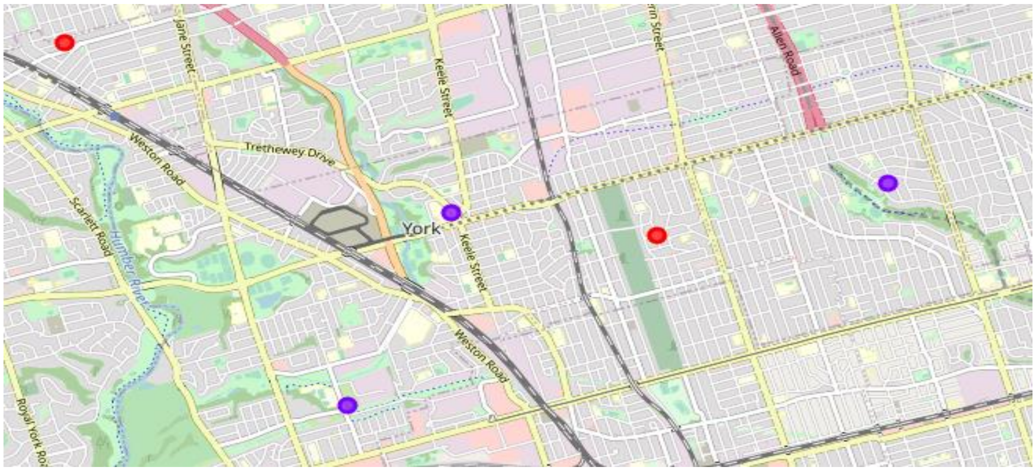


r. We then make use of the K – means clustering machine learning algorithm. We choose the k value as 2 here. We get the following result post the algorithm



	Neighborhood	Latitude	Longitude	Cluster_Labels
0	Caledonia-Fairbanks	43.689026	-79.453512	0
1	Del Ray, Mount Dennis, Keelsdale and Silverthorn	43.691116	-79.476013	1
2	Humewood-Cedarvale	43.693781	-79.428191	1
3	Runnymede, The Junction North	43.673185	-79.487262	1
4	Weston	43.706876	-79.518188	0

s. Visualizing the above data on a map gives us the following



6. Results and Observations

- We can infer from the above data and map that the 2 clusters denote whether a neighborhood has a park or not.
- The red clusters denote the neighborhoods where the parks are present while the blue clusters denote the neighborhoods where the parks are absent.
- Cluster 0 has Parks namely Caledonia -Fairbanks and Weston
- Cluster 1 has no Parks namely Del Ray, Mount Dennis, Keelsdale and Silverthorn, Humewood-Cedarvale and Runnymede, The Junction North

7. Conclusions

- If a person is looking for a new residential space or office space in York based on the availability of a park in their neighborhood then they can go for the Neighborhoods - Cluster 0 - **Caledonia-Fairbanks** and **Weston**. And if, anyone is planning to a construct a new park then it would be better to avoid this cluster or at the very least consider this cluster at a later point if the aim is to have a park in every neighborhood.
- If a person or group is looking for a neighborhood without parks in order to construct new ones then, Cluster 1 is the right group which contains **Del Ray, Mount Dennis, Keels dale and Silverthorn, Humewood-Cedarvale, Runnymede, The Junction North**. These are also apt for people looking for places without nearby parks which may be due to the rising rents, allergies and health conditions or even personal choices.