PREDICTING
THE PERFECT NEIGHBORHOOD

A DATA SCIENCE PROJECT

BY ANURAG GUPTA

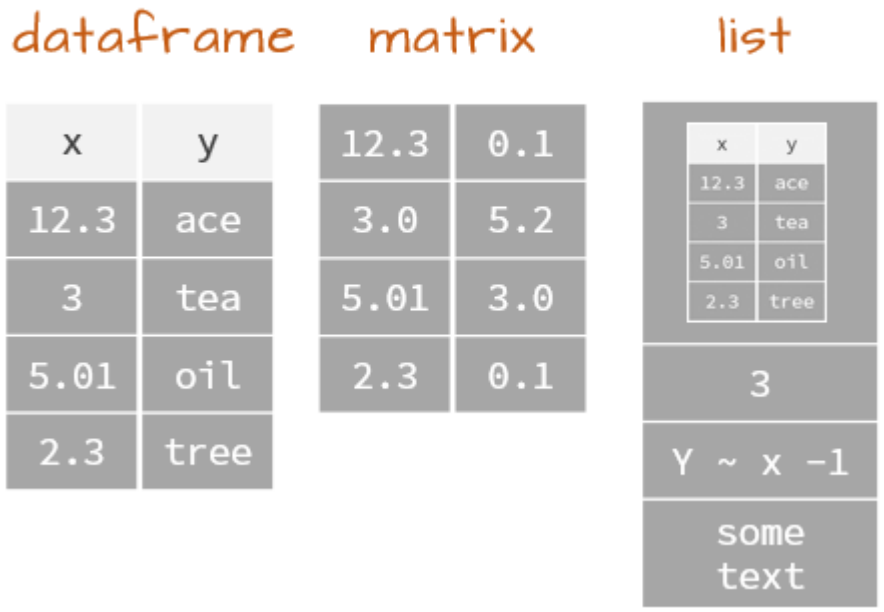
APRIL 10, 2020

INTRODUCTION

- Often times, people come across situations which forces them to move to a different city or even a different country. But in any case it is important that the neighborhood we choose to live in should be similar to our old one as much as possible. It allows us to adapt more easily as well as the places we like will be in close proximity.
- If we try to find the perfect neighborhood manually it will require us to check each and every neighborhood, find whether hundreds of venues such as restaurants, exist in close proximity or not. What is the distance from the schools, hospitals and many more factors.
- Doing this will require a lot of man-hours, and it still might not be effective enough. Thus, using a machine learning algorithm is a much better option.

MANAGING DATA

- The neighborhood names, boroughs and postal codes for Toronto were scraped from the Wikipedia page.
- The latitude and longitude coordinates of the neighborhoods was acquired using geopy library,.
- The venue details, categories and latitude and longitude were acquired using foursquare API. The neighborhoods were scanned in a 1000 meter radius for venues. The API only returned top 100 venues for each neighborhood.
- Wikipedia page: link
- After cleaning, getting location coordinates, the entire dataset had 7979 entries, which included all venues in all neighborhoods. The entire model generation is based on this dataset

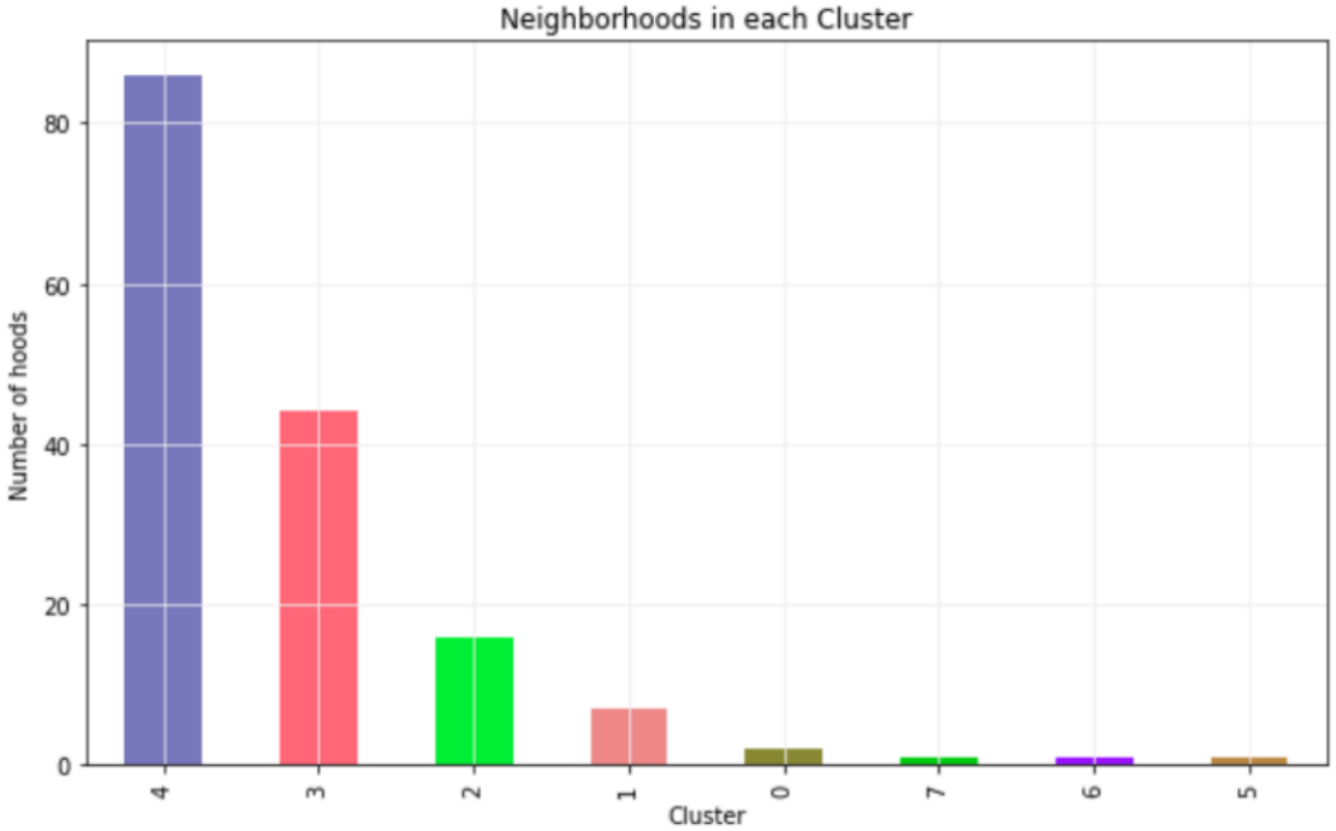


BASE DATAFRAME

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Venue	VLatitude	VLongitude	Category
0	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Richmond Station	43.651569	-79.379266	American Restaurant
1	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Dineen Coffee	43.650497	-79.378765	Café
2	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	GoodLife Fitness Toronto 137 Yonge Street	43.651242	-79.378068	Gym
3	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Beerbistro	43.649419	-79.377237	Gastropub
4	M5H	Downtown Toronto	Adelaide	43.650486	-79.379498	Pilot Coffee Roasters	43.648835	-79.380936	Coffee Shop

```
(7979, 9)
Number of neighborhoods : 158
Number of venue categories : 355
```

ASSIGNING CLUSTER TO NEIGHBORHOOD

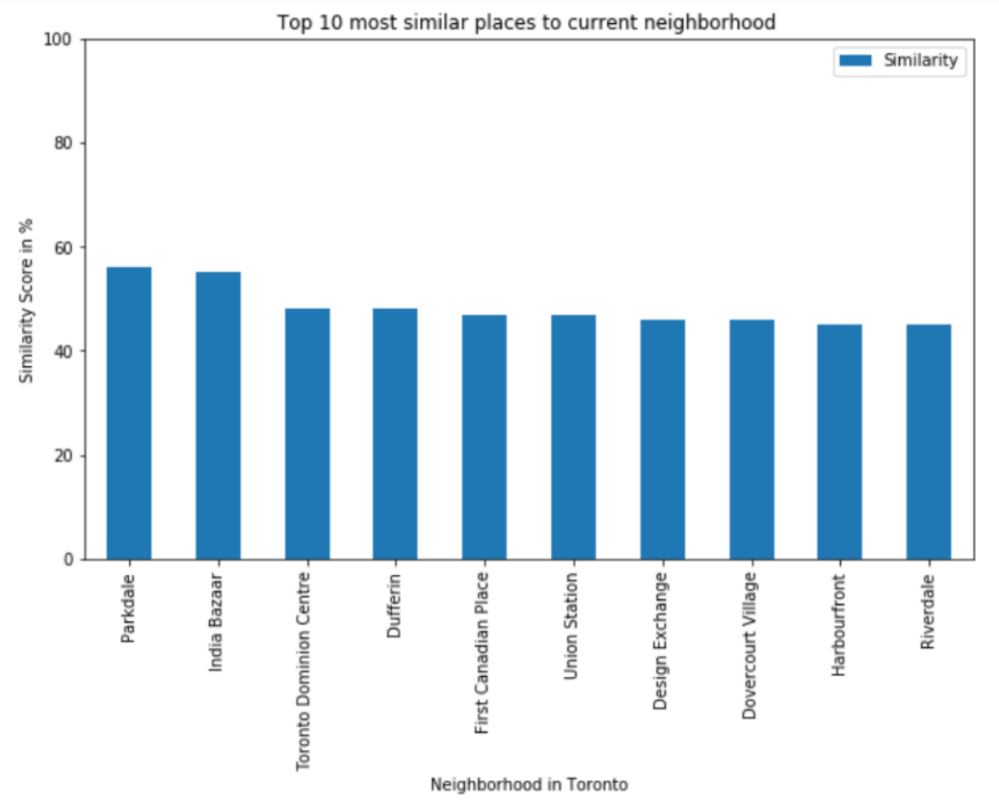


- The cluster to which each neighborhood belonged to was determined using K – Means Clustering.
- The neighborhoods were divided into 8 clusters.
- The mean number of neighborhood per cluster is 19.7 and the standard deviation was 30.5 indication the distribution of neighborhoods into cluster is not uniform.
- With the top 3 clusters having 86, 44, 16 neighborhoods out of 158.

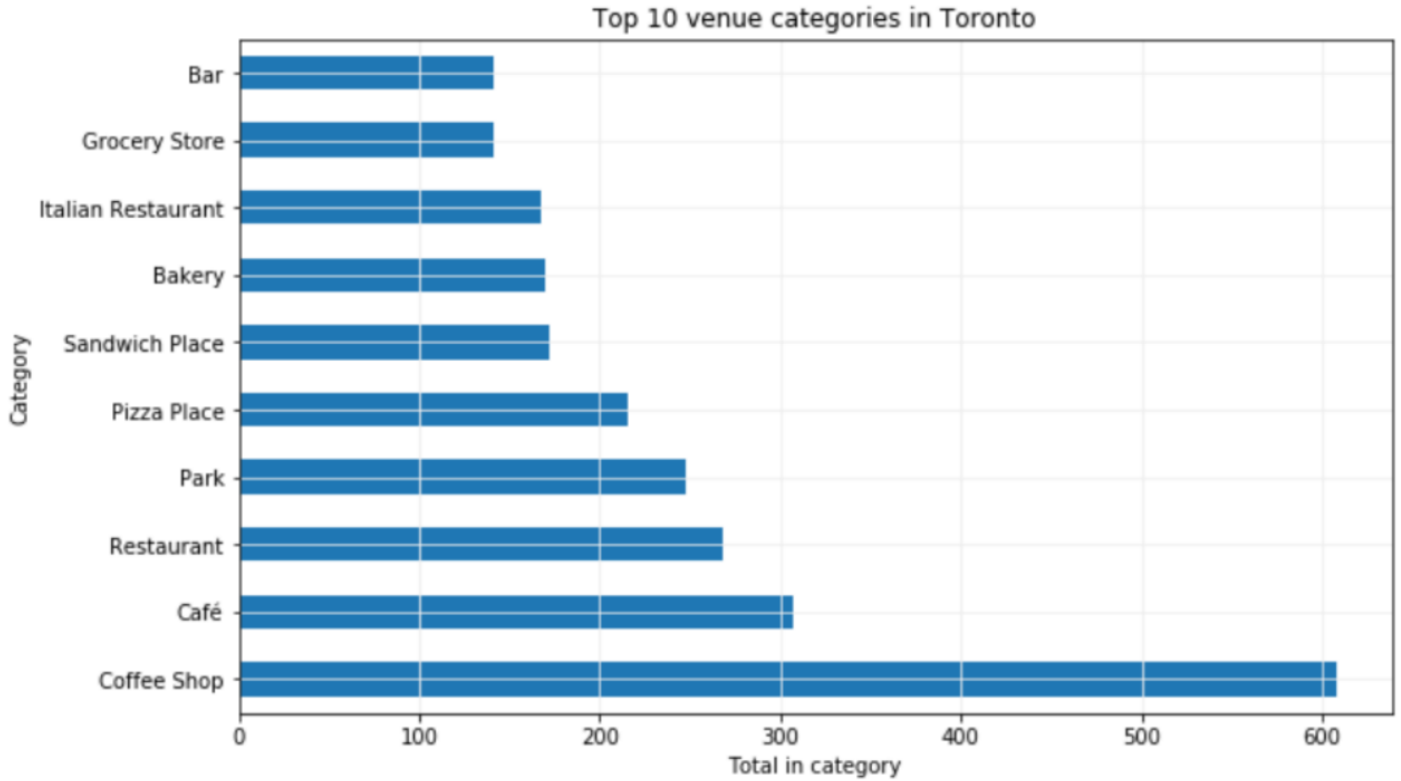
CLASSIFYING CURRENT NEIGHBORHOOD

	Neighborhood	Distance	Similarity
0	Parkdale	0.182004	56.0
1	India Bazaar	0.195937	55.0
2	Toronto Dominion Centre	0.198032	48.0
3	Dufferin	0.209851	48.0
4	First Canadian Place	0.202616	47.0
5	Union Station	0.203119	47.0
6	Design Exchange	0.205274	46.0
7	Dovercourt Village	0.207214	46.0
8	Harbourfront	0.207449	45.0
9	Riverdale	0.211510	45.0

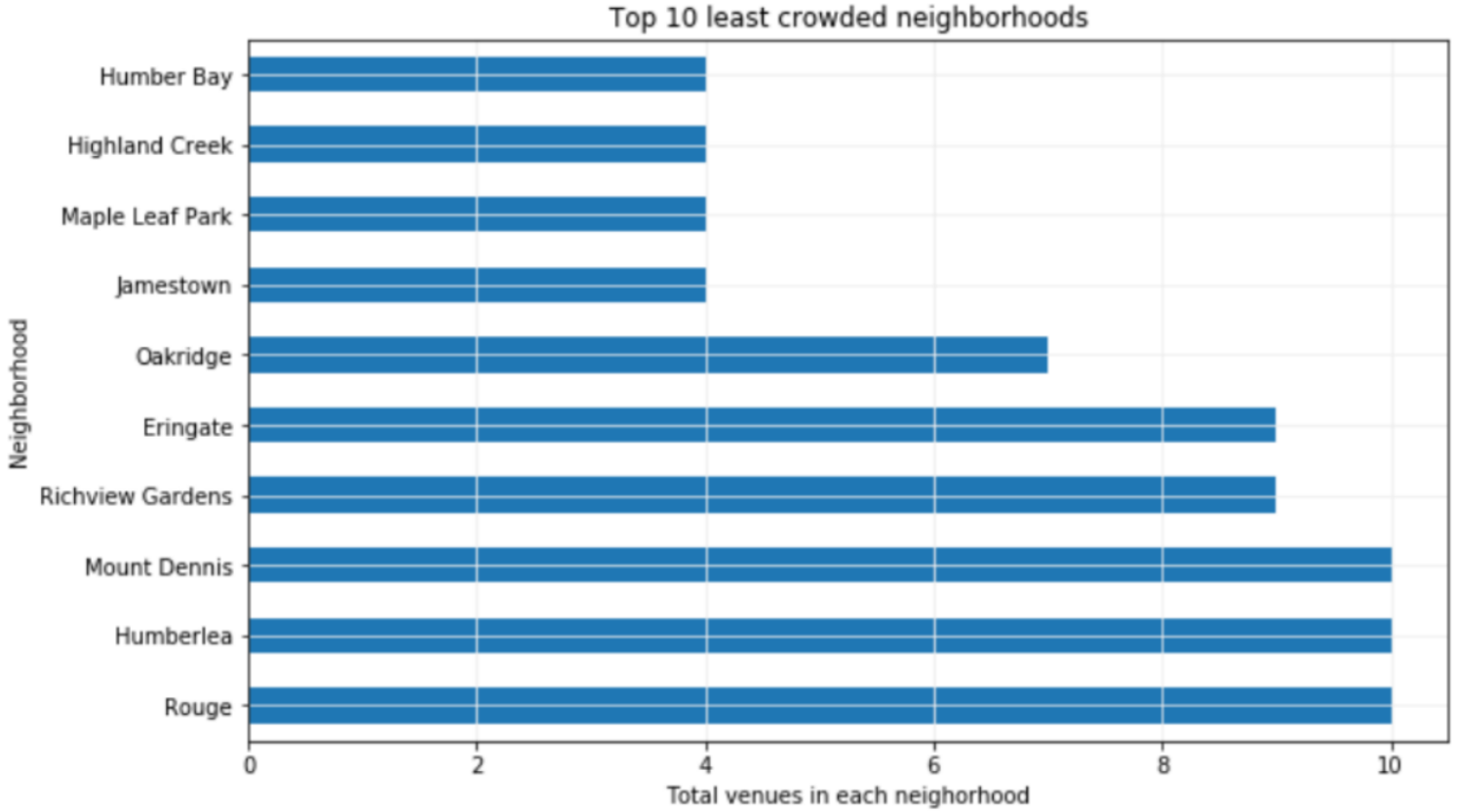
- Cluster 4 was assigned to the new neighborhood 'Connaught Place, New Delhi'.
- A similarity score of each neighborhood in the cluster with the current neighborhood was obtained, which is basically the correlation between them.
- The Euclidean distance between the neighborhoods was also obtained.



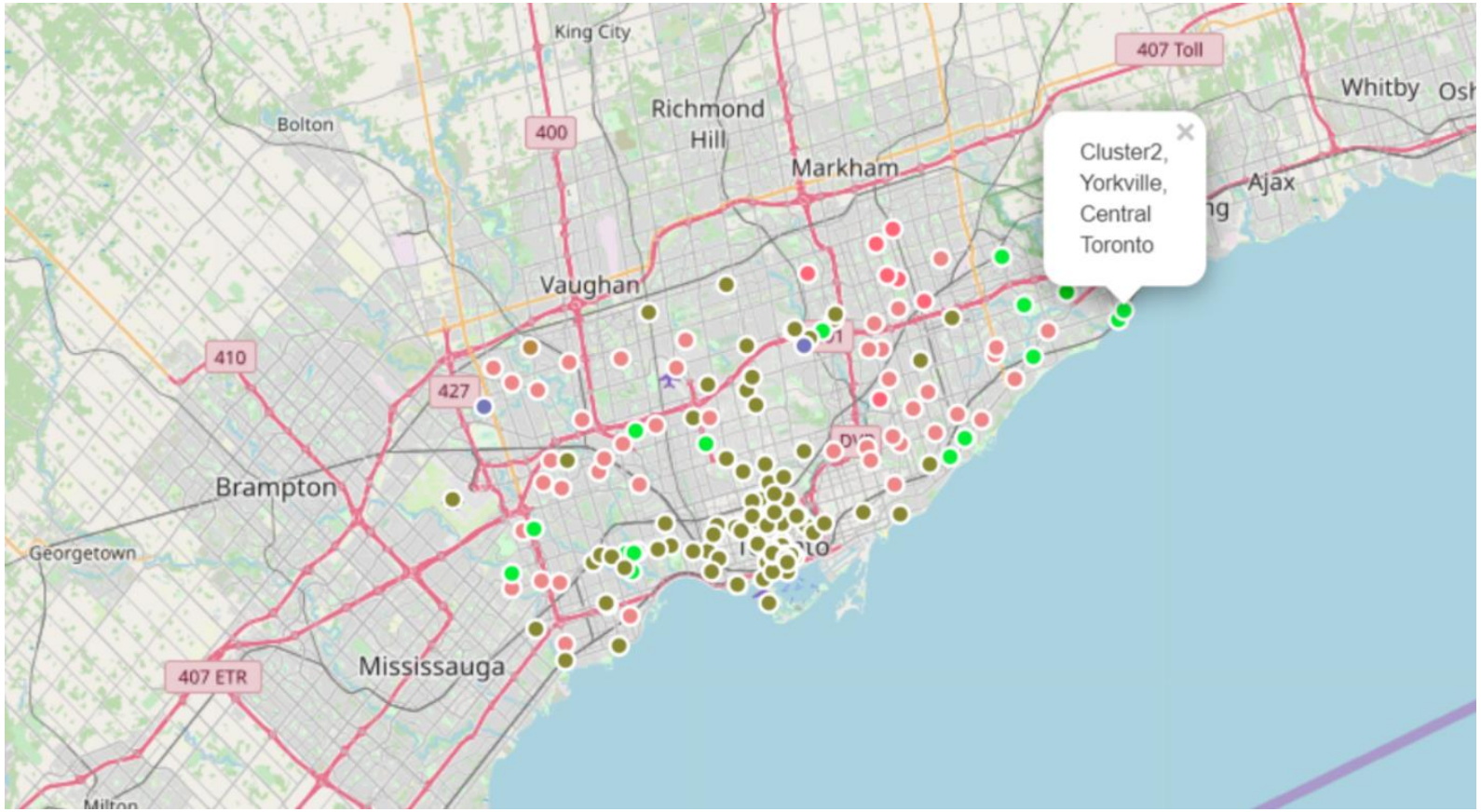
TOP VENUE CATEGORIES



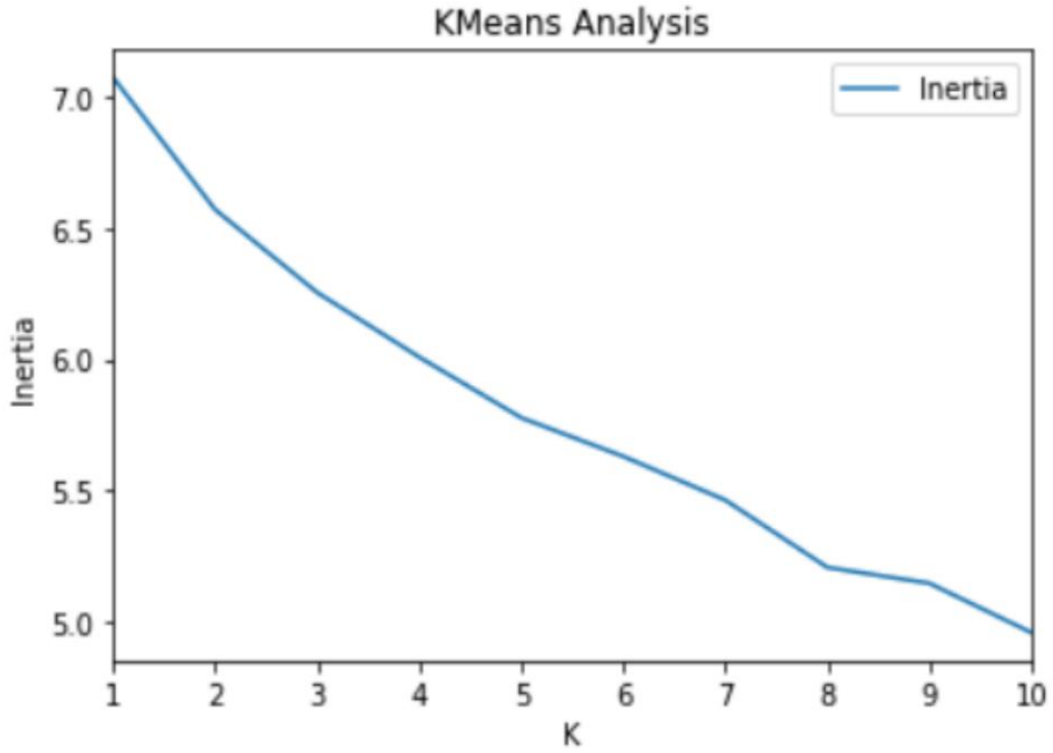
THE LEAST CROWDED NEIGHBORHOODS



MAP OF ALL 8 CLUSTERS IN TORONTO



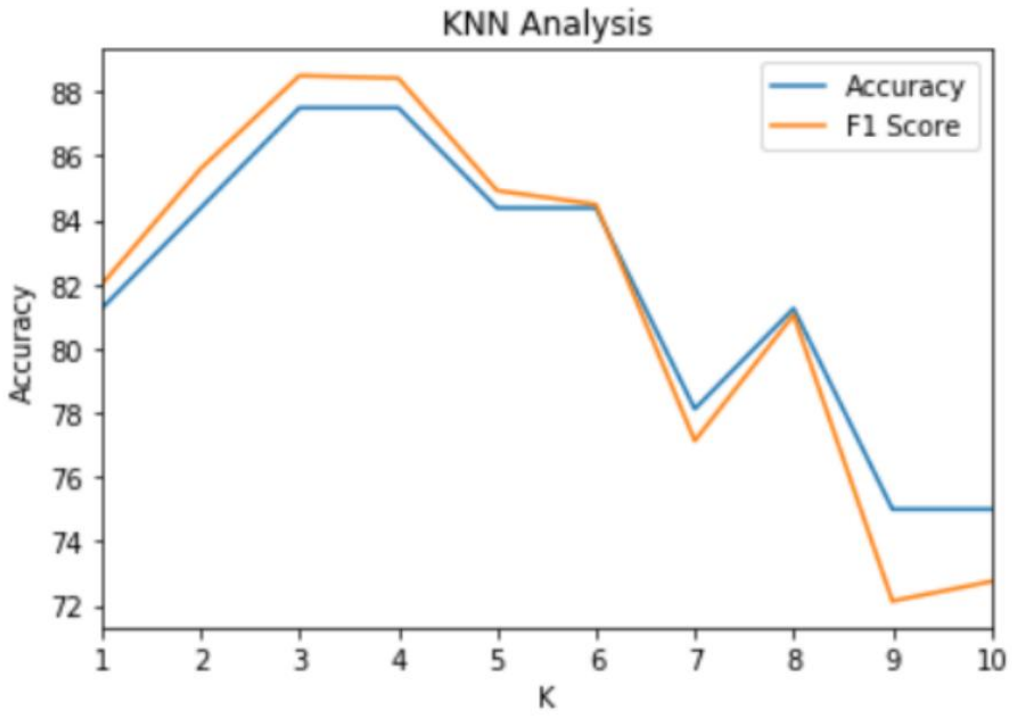
K-MEANS CLUSTERING



- K-Means clustering was used to group the similar neighborhoods in Toronto together.
- The neighborhoods were divided into 8 clusters.
- The best K was found using elbow technique.
- The clustering model was made using 352 features.

KNN CLASSIFICATION

	K	Accuracy	F1 Score
0	1	81.25	82.00
1	2	84.38	85.60
2	3	87.50	88.50
3	4	87.50	88.42
4	5	84.38	84.92
5	6	84.38	84.49
6	7	78.12	77.13
7	8	81.25	81.05
8	9	75.00	72.14
9	10	75.00	72.75



- KNN classification was used to classify the current neighborhood (given by user) to a cluster.
- The model was evaluated using different values of K and the best K was found to be 3.
- .The Jacquard similarity index for the model (K = 3) was 87.5 and the F1 score was 88.4.
- The cross validation score for the model was found to be 0.78

KNN CLASSIFICATION REPORT

Classification Report :

	precision	recall	f1-score	support
1	1.00	1.00	1.00	1
2	1.00	0.67	0.80	6
3	1.00	0.80	0.89	5
4	0.83	1.00	0.90	19
5	0.00	0.00	0.00	1
micro avg	0.88	0.88	0.88	32
macro avg	0.77	0.69	0.72	32
weighted avg	0.87	0.88	0.86	32

The cross validation score is : 0.78

CONCLUSION

- In this study I analyzed the neighborhoods present in the city of Toronto, Ontario, Canada. I found out the venues present in each neighborhood and clustered the similar neighborhoods together using K-Means Clustering.
- I found the most common venues in the city of Toronto, which neighborhoods are the most crowded, which are the least crowded, which venues are the most popular.
- I used KNN Classification to predict which cluster a new neighborhood will belong to considering the types of venues present in the neighborhood.
- The classification model had an accuracy of around 87.5 % using Jacquard similarity score, and 88.4% using F1 score, which is pretty good considering the small sample size. Thus this model can be used to predict a perfect new neighborhood similar to a given neighborhood.