2020

Safest Neighborhood in Toronto for opening a commercial establishment

CAPSTONE PROJECT

FER VÁZQUEZ

1. INTRODUCTION:

1.1 BACKGROUND

Toronto is a great place to live, the shopping is great, thousands of restaurants and cafes to get a fantastic meal, and there are lots of things you can do at any hour from strolling through parks, catching a movie or concert, or watching some live sports. But opening a business in Toronto isn't always so good, especially if you ask about crime. Fortunately, if you want to open your business in Toronto, with this project we will be looking to understand the crime, and which will be the best neighborhood to open your own business.

1.2. BUSINESS PROBLEM:

The purpose of this project is to understand which neighborhood will be the best to open a commercial business in Toronto and which type of commercial business. The first task will be to understand which neighborhood is the safest by analyzing the crime data and the second task will be to analyze the 10 most common venue in these neighborhoods. We will use our knowledge of Data Science to do this analysis.

1.3 INTEREST

For all entrepreneur this project will be useful to know which neighborhood is the safest to opening a commercial establishment.

2. DATA ACQUISITION AND CLEANING

2.1 DATA SOURCE AND DATA CLEANING

The data of crimes I will use the real data that it is published in Kaggle dataset for this page:
https://www.kaggle.com/kapastor/toronto-police-data-crime-rates-by-neighbourhood

In the next table I describe the columns and the transformation that I will apply for each column:

Column	Description	Transformation
X	Latitude	Remove
Y	Longitude	Remove
Index_	Unique ID	I will use as unique id
event_unique_id	Event ID	Remove
occurrencedate	Date of crime occurred	Remove
reporteddate	Date of crime reported	Remove
premisetype	Location of crime occurred (commercial, house, apartment, outside, other)	I will use to filter the premise only with commercial and outside types.
ucr_code	Code	Remove
ucr_ext	Ext	Remove
offence	Crime description	Remove
reportedyear	Year of the report	Remove
reportedmonth	Month of the report	Remove
reportedday	Day of the report	Remove
reporteddayofyear	Year day of the report	Remove
reporteddayofweek	Week day of the report	Remove
reportedhour	Hour of the report	Remove
occurrenceyear	Year of the crime occurred	Remove
occurrencemonth	Month of crime occurred	I will use to known which month has more crimes

occurrenceday	Day of crime occurred	Remove
occurrencedayofyear	Year Day of crime occurred	Remove
occurrencedayofweek	Day of week of crime occurred	I will use to known which day of week has more crimes
occurrencehour	Hour of crime occurred	Remove
MCI	Type of crime	I will use to know the type of crime
Division	Division	Remove
Hood_ID	Neighborhood Id	Remove
Neighbourhood	Neighborhood	I will use to know the name of the Neighborhood
Long	Longitude	I will use to create the map
Lat	Latitude	I will use to create the map
ObjectId	Object ID	Remove

For data of Toronto Neighborhoods, I will use the WikIpedia source:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This list, I will clean it to have the next dataframe:

- PostCode
- Borough

- Neighborhood

Then, I will use another dataset to get the Latitude and Longitude of each neighborhoods, the final dataframe will be:

- PostCode
- Borough
- Neighborhood
- Latitude
- Longitude

And Finally, I will use the Foursquare location data to know the 10 most common venue in the safest neighborhood.

3. EXPLOROTORY DATA ANALYSIS

STATISTICAL SUMMARY OF CRIMES

The describe function in python is used to get statistics of the crime data, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime.

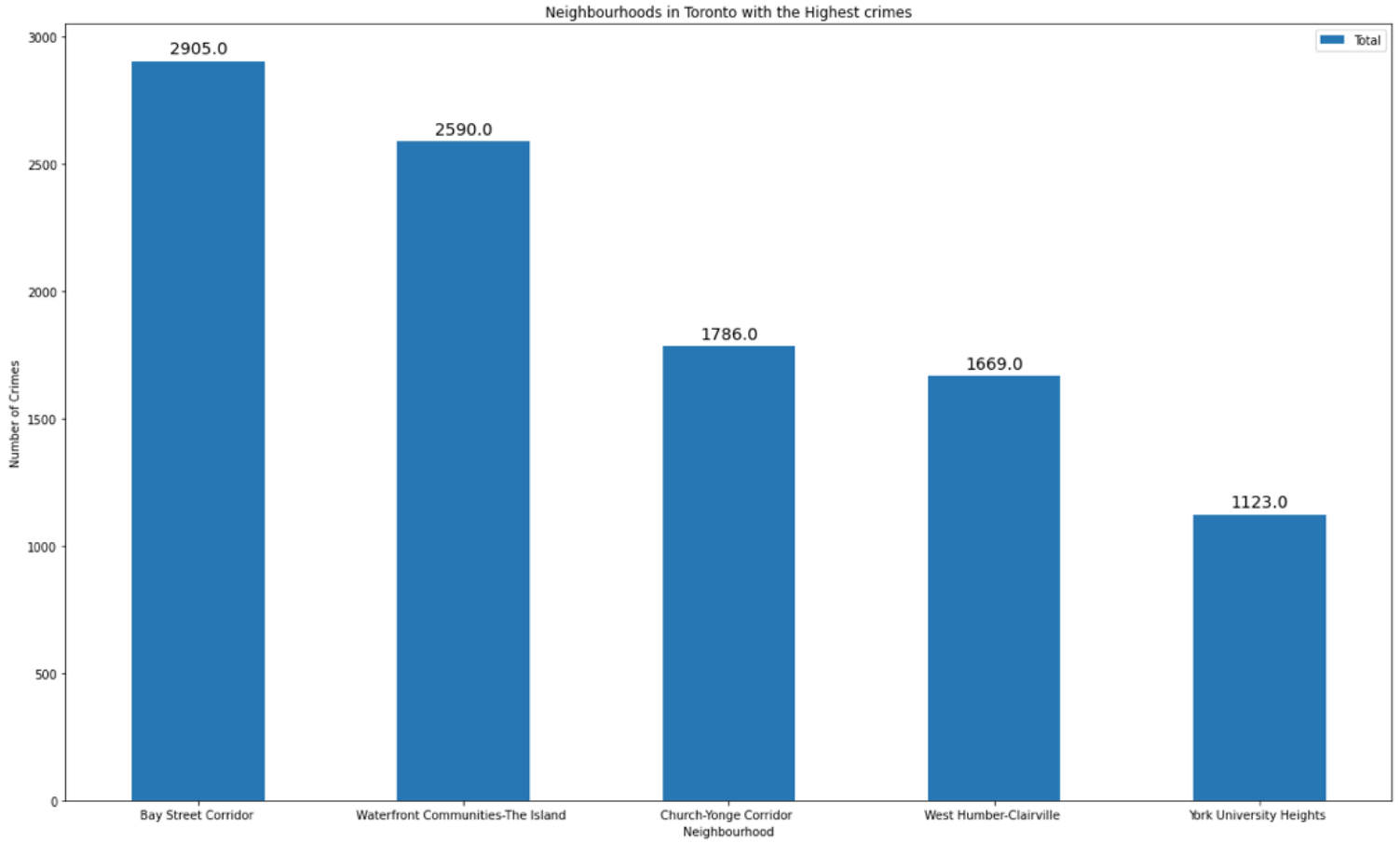
[13]:
```
df_crime_cat.describe()
```

[13]:

	occurrenceyearAssault	occurrenceyearAuto Theft	occurrenceyearBreak and Enter	occurrenceyearRobbery	occurrenceyearTheft Over	Total
count	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000
mean	249.191489	33.673759	194.063830	71.985816	33.773050	582.666667
std	1485.352055	202.090379	1150.696201	425.861463	201.227496	3459.193320
min	3.000000	0.000000	5.000000	0.000000	0.000000	20.000000
25%	32.000000	2.000000	29.000000	14.000000	4.000000	86.000000
50%	66.000000	5.000000	61.000000	25.000000	7.000000	171.000000
75%	144.000000	16.000000	107.000000	43.000000	17.000000	291.000000
max	17568.000000	2374.000000	13681.000000	5074.000000	2381.000000	41078.000000

NEIGHBORHOODS WITH THE HIGHEST CRIME RATES

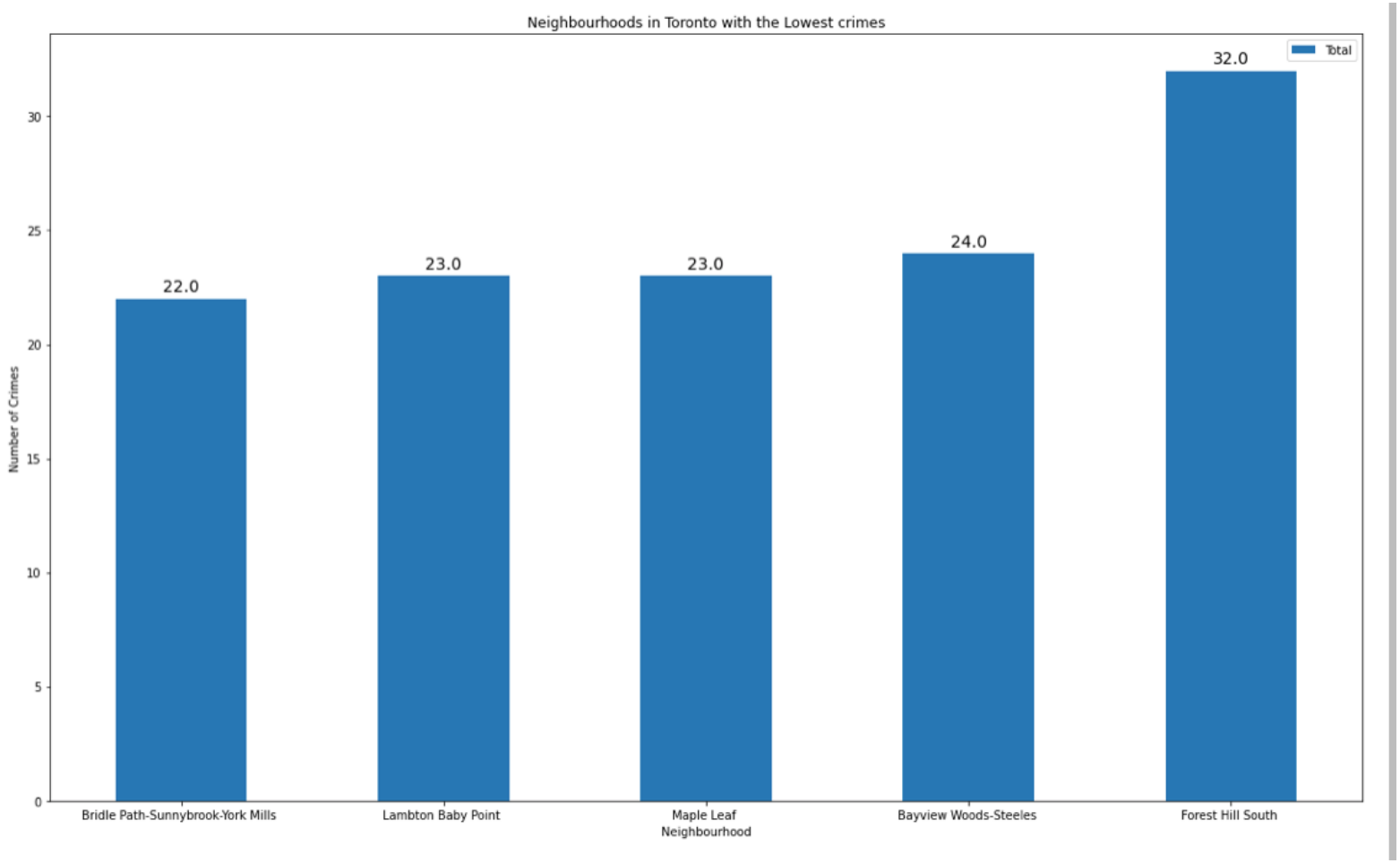
With the Gran Total count of crimes, we have the top 5 of neighborhoods with highest crime:



The number one is Bay Street Condor neighborhood and takes the major chunk of the crime records.

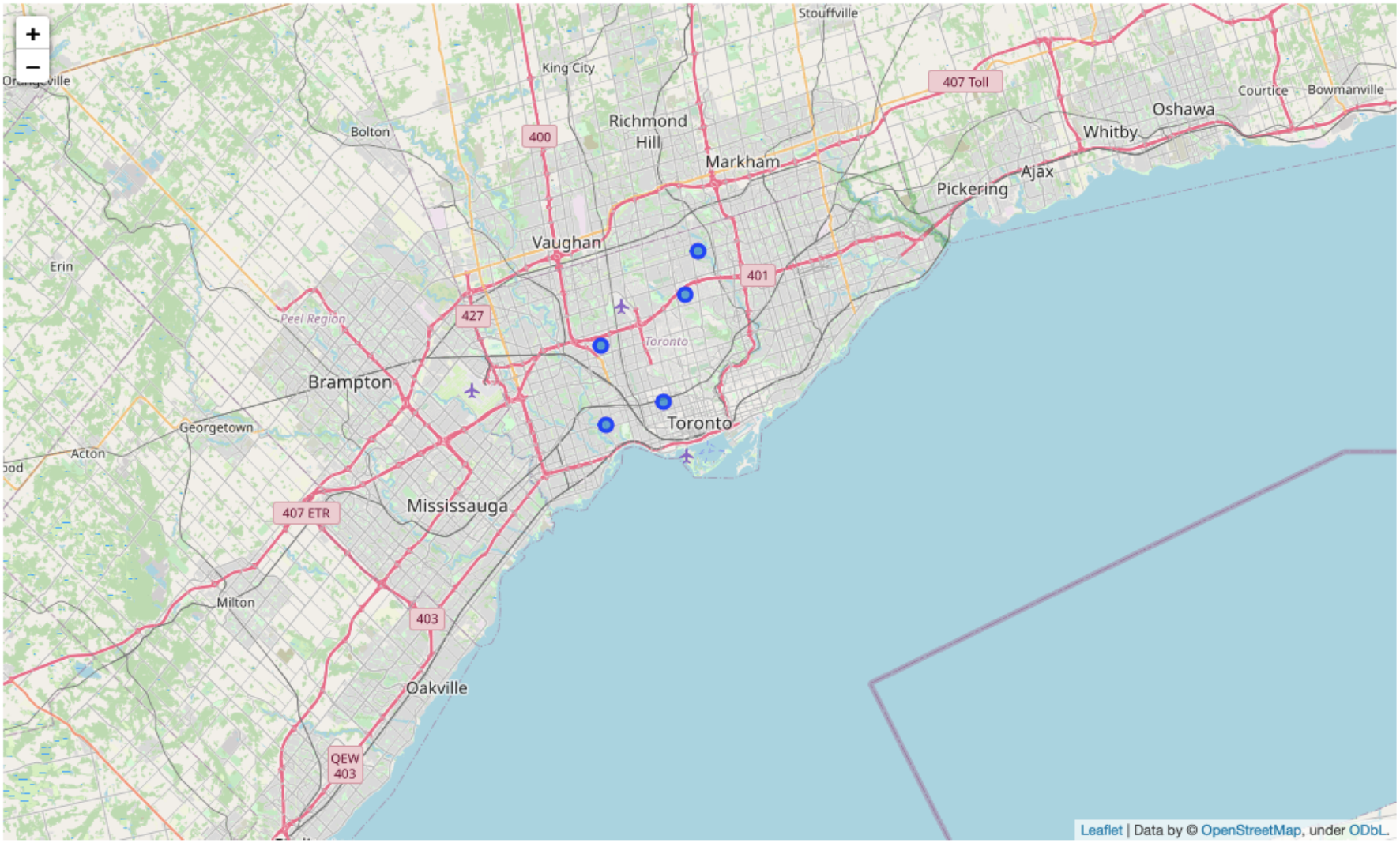
NEIGHBORHOODS WITH THE LOWEST CRIME RATES

With the Gran Total count of crimes, we also have the bottom 5 of Neighborhoods with lowest crimes:



The safest one is Bridle Path- Sunnybrook-York Miles.

These safest neighborhoods are located in:



Based on the bottom 5 dataset of neighborhoods, we can merge with the Wikipedia Dataset to find the borough of this neighborhoods and the latitude and longitude. So, finally we have the finally dataset:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M2P	North York	York Mills West	43.752758	-79.400049
1	M6L	North York	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074
2	M6S	West Toronto	Runnymede, Swansea	43.651571	-79.484450
3	M2K	North York	Bayview Village	43.786947	-79.385975
4	M6G	Downtown Toronto	Christie	43.669542	-79.422564

4. MODELING

With the safest neighborhoods dataset, we were able to find the venues within a 500 meter radius of each neighborhood by connecting to the FourSquare API. This returns a response in json format containing all the venues in each neighborhood which we convert to a pandas data frame. This data frame contains all the venues along with their coordinates and category will look as follows:

[166]:

(58, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	York Mills West	43.752758	-79.400049	Kitchen Food Fair	43.751298	-79.401393	Convenience Store
1	York Mills West	43.752758	-79.400049	Tournament Park	43.751257	-79.399717	Park
2	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	Rustic Bakery	43.715414	-79.490300	Bakery
3	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	Maple leaf park	43.716188	-79.493531	Park
4	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	Mika's Trim	43.714068	-79.496113	Construction & Landscaping
5	Runnymede, Swansea	43.651571	-79.484450	Coffee Tree Roastery	43.649647	-79.483436	Café
6	Runnymede, Swansea	43.651571	-79.484450	Bryden's Pub	43.649259	-79.484651	Pub

One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

To help entrepreneurs understand similar venues in the safest neighborhood we will be clustering similar venues using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 3 for this project that will cluster the 5 safest neighborhoods into 3 clusters.

The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

5.RESULT

After running the K-means clustering the result is as follows:

[47]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	M2P	North York	York Mills West	43.752758	-79.400049	0	Convenience Store	Park	Yoga Studio	Candy Store	Diner	Dessert Shop	Construction & Landscaping	Coffee Shop
1	M6L	North York	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	2	Bakery	Construction & Landscaping	Park	Yoga Studio	Candy Store	Diner	Dessert Shop	Convenience Store
2	M6S	West Toronto	Runnymede, Swansea	43.651571	-79.484450	1	Café	Coffee Shop	Sushi Restaurant	Pub	Pizza Place	Italian Restaurant	Yoga Studio	Dessert Shop
3	M2K	North York	Bayview Village	43.786947	-79.385975	1	Chinese Restaurant	Bank	Café	Japanese Restaurant	Yoga Studio	Diner	Dessert Shop	Convenience Store
4	M6G	Downtown Toronto	Christie	43.669542	-79.422564	1	Grocery Store	Café	Park	Athletics & Sports	Candy Store	Baby Store	Restaurant	Coffee Shop

Looking into the neighborhoods in the first and third cluster, we can see these clusters have only one neighborhood in each. This is because of the unique venues in each of the neighborhoods, so they couldn't be clustered into similar neighborhoods.

The first cluster has one neighborhood which consists of Venues such as Yoga Studio, Candy Store, eateries (Diner, Dessert shop, coffee Shop) and the Convenience Store, so a Grocery Store is not good idea in this neighborhood.

The cluster second is the biggest cluster with 3 of the 5 neighborhoods in different borough. Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Restaurants, eateries, fitness amenities (yoga studios or park) and Candy/Dessert shops. For example, a Bookstore or Grocery store are not among the most common venues which makes this cluster of neighborhoods an ideal destination to set up a grocery store or a bookstore.

In the third cluster has one neighborhood which consists of Venues such as eateries (Bakery, Candy store, Diner, Dessert Shop) and the Convenience Store, so a Grocery Store is not good idea in this neighborhood too.

5. DISCUSSION

The objective of the business problem was to help entrepreneurs identify the safest neighborhoods in Toronto, and an appropriate venue to set up a commercial establishment such as Grocery Store, Book Store, Gym, Fitness Yoga, etc. This has been achieved by first making use of Toronto crime data to identify safest neighborhoods for any business to be viable. After selecting the neighborhoods, it was imperative to analyze the venues in the neighborhoods. We achieved this by grouping the neighborhoods into clusters to assist the entrepreneurs by providing them with relevant data about venues and safety of a given neighborhood.

6. CONCLUSION

We have explored the crime data to understand different types of crimes in all neighborhoods of Toronto, this helped us to know the safest neighborhoods first. Once we confirmed the safest neighborhoods, we analyzed neighborhoods based on the common venues, to choose a neighborhood and the type of commercial business which best suits the business problem.