Battle of neighborhood : Toronto city Starting a new coffee shop

Business Problem Toronto is Canada's largest city with a population of more than 2,7 million and a density of 4,334.4 people per square kilometer. The city is renowned as one of the most multicultural cities globally due to its large population of immigrants from all over the globe.This also means that the market is highly competitive.Thus, any new business venture or expansion needs to be analysed carefully. The insights derived from analysis will give a good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

Problem Overview Now, imagine that you have a coffee franchise called Coffee Costa that has been doing business successfully in New York. you plan to expand the business and decide to look for a city that shares the same trait as New York, and one of the cities is Toronto.To ensure this project's success, the team requires insights into the demographics, neighboring businesses, and crime rates. For each neighborhood, we can ask: How many other cafes exist?What are the most popular hangouts? Can we get information about the other recreational spots? What is the neighborhoods' crime rate? Thus, the project goal is to figure out the best locations for opening up a new coffee shop in Toronto City. Other factors needed to be taken in consideration are -Toronto Population, Toronto City Demographics

Are there any other venues like, Entertainment zones, Parks etc nearby where floating population is high etc

Who are the competitors in that location?

Cuisine served / Menu of the competitors
Segmentation of the Borough
Untapped markets
Saturated markets etc

Target Audience Entrepreneurs who are passionate about opening a coffee shop in a metropolitan city would be very interested in this project. The project is also for business owners and stakeholders who want to invest in the franchise chain.

Data Description

Data Requirements and Collection We need historical data about crime incidents, busiest roads, and popular venues. Luckily, Toronto has an open data portal that makes it public. We can also leverage Foursquare Location data to compare neighborhoods in terms of service.Followings are data sources that we can use for this project-:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M The list of Toronto neighborhoods represented by postal codes and their boroughs

https://ckan0.cf.opendata.inter.prod-toronto.ca/en/dataset/traffic-signal-vehicle-and-pedestrian-volumes The most updated record of traffic signal vehicle and pedestrian volumes in Toronto City

https://tinyurl.com/toronto-mci The most updated record of crime incidents reported in Toronto City provided by Toronto Police Services. :

https://developer.foursquare.com/ The popular or most common venues of a given neighborhood in Toronto.

Data Cleaning and Feature Extraction The first is a Wikipedia page about Toronto postal code. We will scrape the page and create a data frame consisting of three columns; PostalCode, Borough, and Neighborhood. We remove any rows that do not have a borough assigned. Then, we will be using the Geocoder python package to retrieve the postal code's coordinates. It will return 103 rows and 5 columns.

The second data is in a CSV file. It contains 2280 rows and 11 columns. The data is typically collected between 7:30 a.m. and 6:00 p.m at intersections where there are traffic signals. Each intersection holds vehicle and pedestrian volumes data, along with its coordinates. We will focus on 5 columns; those are Main, 8 Peak Hr Pedestrian Volume, 8 Peak Hr Vehicle Volume, Latitude, and Longitude. We will use these features to diagnose each main road's characteristics and locate the busiest main roads in the city.

The third data is also in a CSV file. It contains 206,435 rows and 9 columns. The rows represent crime incidents that were reported from 2014 to 2019. It has 5 Major Crime Indicators (MCIs) scattered to 17 divisions and 140 listed neighborhoods. We will group the data based on division and get statistics about crime rates

The fourth data is stored inside Foursquare Location Data, and we will use Foursquare API to access it. We utilize the postal coordinates to retrieve popular venues around a specific radius. As a result, the same venue categories will be returned to different neighborhoods. We can use this idea to cluster the neighborhoods based on their venues representing services and amenities.

We will run the k-Means algorithm to perform this clustering with a different number of clusters (k). The features will be the mean of the frequency of occurrence of each venue category. Finally, we can visualize the cluster model using the Folium module.

To sum up, we will use the 2nd and 3rd data to analyze the pedestrian/vehicle volume and crime rates. Then, we load the 1st data to obtain the exact coordinates for each neighborhood based on the postal code, allowing us to explore and map the city. Finally, we will use the coordinates and Foursquare credentials to access the 4th data source through its API and retrieve the popular venues along with their details, especially for coffee shops. The venue frequency in each neighborhood will be the features of the clustering model.

Methodology

statistics needed to answer questions concerning crime incidents, and vehicle and foot traffic records are gathered.Then, we approach the problem using the clustering technique, namely k-Means. This approach enables the audience to see how similar neighborhoods are about their demographics. We can then examine each cluster and determine the discriminating venue categories that distinguish each cluster.k-Means is one of the common machine learning algorithms used to cluster data points based on similar characteristics. The algorithm is fast and efficient for a medium and large-sized database and is useful to discover insights from unlabeled data quickly.

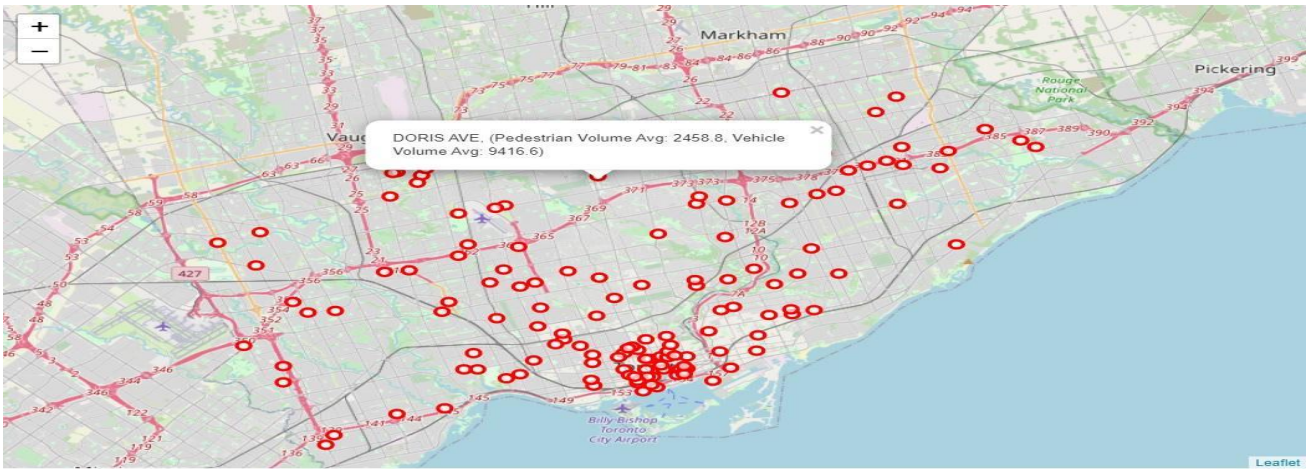
Exploratory Data Analysis Vehicle and Foot Traffic

We begin by analyzing the data about the pedestrian and vehicle volumes. The column Main contains the main street name that appears several times indicating it contains intersections. We can group by the street name and aggregate this either by summing those values up or averaging it. We will choose to average it for simplicity. This returns 248 main roads.

	8 Peak Hr Pedestrian Volume	8 Peak Hr Vehicle Volume	Latitude	Longitude
count	248.000000	248.000000	248.000000	248.000000
mean	1855.100736	11274.239194	43.710040	-79.395862
std	3190.819880	5193.129205	0.056145	0.102267
min	0.000000	1081.000000	43.603757	-79.622225
25%	343.500000	7403.750000	43.660016	-79.472370
50%	675.000000	10466.433824	43.703423	-79.390913
75%	1653.678571	13938.031250	43.761615	-79.330196
max	23335.000000	29797.428571	43.825259	-79.140419

Then we filter out the roads. In this example, we only show the roads with an average of pedestrian volume above 1,200or vehicle volume above 12,000during peak hour (above ~70%). This gives us 208 main roads.

We plot the map of those 208 roads busiest roads.

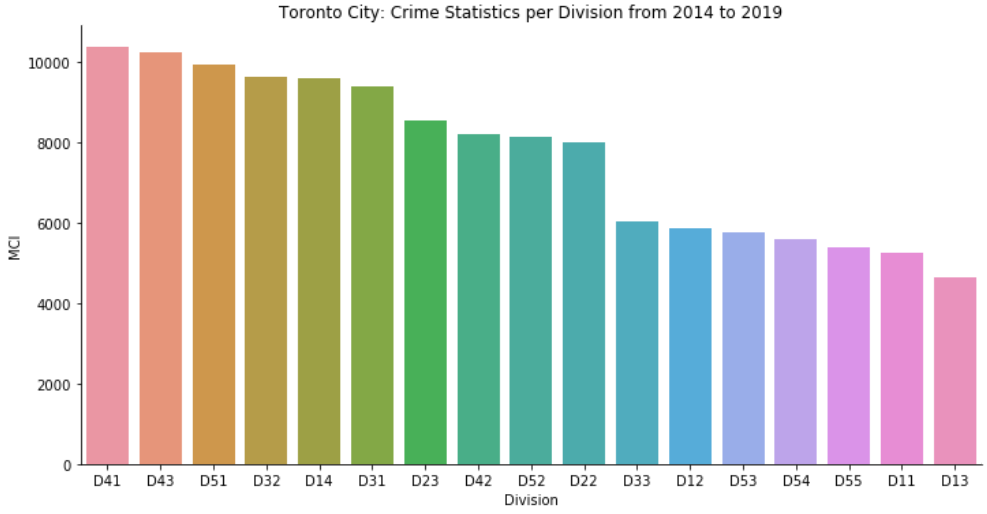


Crime Statistics

Next, we analyze the crime statistics from 2014 to 2019. It gives us 130374 crime incidents segmented by police divisional boundaries, neighborhoods, and Major Crime Indicators (MCI). Toronto Police Service divides the major crimes into 5 categories scattered to 17 divisions and 141 neighborhood IDs.

	premisetype	offence	MCI	Division	Hood_ID	Neighbourhood	Lat	Long	re
0	Other	Assault With Weapon	Assault	D32	36	Newtonbrook West (36)	43.781639	-79.416	
1	Other	Assault With Weapon	Assault	D32	36	Newtonbrook West (36)	43.781639	-79.416	
2	Other	Assault With Weapon	Assault	D32	36	Newtonbrook West (36)	43.781639	-79.416	
3	Other	Assault With Weapon	Assault	D32	36	Newtonbrook West (36)	43.781639	-79.416	
4	Commercial	B&E	Break and Enter	D14	79	University (79)	43.665390	-79.410	
...	...	...	...	...	...	...	...	...	...

We will group the data based on division (**Division**), not neighborhood **(Hood_ID**). This will give us insight into the safest boroughs and their neighborhoods.



Among the 5 MCIs, Assault incidents have the most occurred for 6 consecutive years. In the same period, several divisions are consistent about their crime rates. We can segment them into three groups:

- **High Crime Rates** (D51, D43, D41, D32, D31, D14)
- **Middle Crime Rates** (D52, D42, D23, D22)
- **Low Crime Rates** (D55, D54, D53, D33, D13, D12, D11)

Since we expect our candidate neighborhoods to be:

- **safe** — having low crime rates
- **lively** — crowded by people, vehicles, and easy to access
- **close to downtown**,

Therefore, the divisions qualified are **D55**, **D54**, **D53**, and **D13**.

Referring to Toronto Police Service Wikipedia, these divisions cover:

- **Central Toronto** (D53)
- **East York** (D53, D54, D 55)
- **York** (D13)

Then ,we will explore the neighborhoods inside Central Toronto, East York, and York as the selected boroughs.

Neighborhoods Analysis

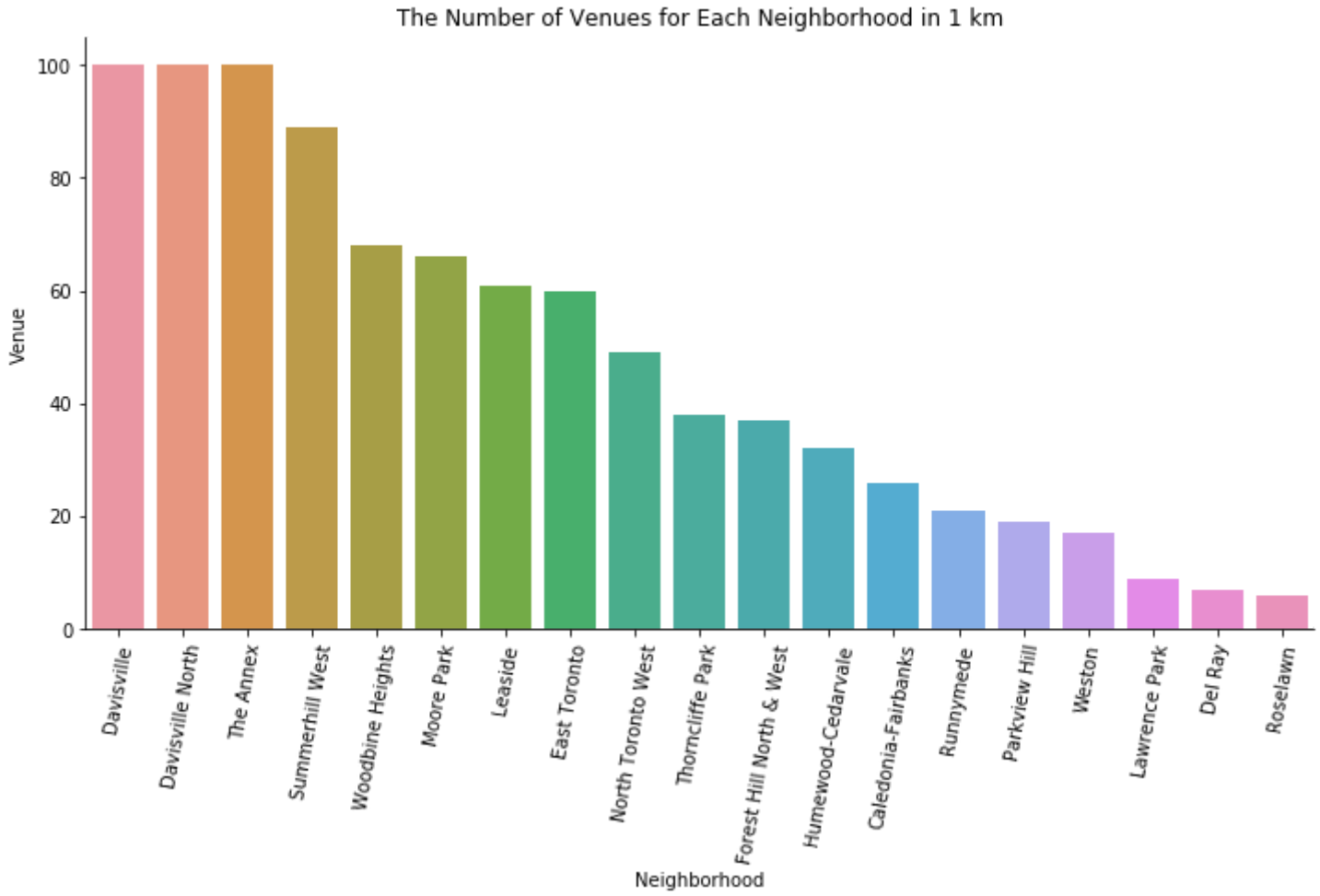
We have built a neighborhood data frame that contains 103 postal codes, 10 boroughs, neighborhood names inside each borough, and their coordinates. Since we are interested in neighborhoods inside Central Toronto, East York, and York only, we filter the data frame. This results in having 3 boroughs and 19 neighborhoods.

	PostalCode	Borough	Neighbourhood	latitude	longitude
0	M4B	East York	Parkview Hill	43.70718	-79.31192
1	M4C	East York	Woodbine Heights	43.68970	-79.30682
2	M6C	York	Humewood-Cedarvale	43.69211	-79.43036
3	M6E	York	Caledonia-Fairbanks	43.68784	-79.45046
4	M4G	East York	Leaside	43.70902	-79.36349

Given the coordinates information, we can use the Foursquare API to access the 2nd data source, explore the neighborhoods, and get the top 100 venues within a radius of 1 km for each. As a result, it returns 905 venues with 172 unique venue categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkview Hill	43.70718	-79.31192	Toronto Climbing Academy	43.709362	-79.315006	Rock Climbing Spot
1	Parkview Hill	43.70718	-79.31192	Jawny Bakers	43.705783	-79.312913	Gastropub
2	Parkview Hill	43.70718	-79.31192	Muddy York Brewing Co.	43.712362	-79.312019	Brewery
3	Parkview Hill	43.70718	-79.31192	East York Gymnastics	43.710654	-79.309279	Gym / Fitness Center
4	Parkview Hill	43.70718	-79.31192	Peek Freans Cookie Outlet	43.713260	-79.308063	Bakery
...	...	...	...	...	...	...	...

For each neighborhood, we can create the top 10 venues based on occurrences as follows.



The data frame below indicates that we have the same venue categories returned to different neighborhoods. We can use this idea to cluster the neighborhoods based on their venues representing services and amenities.

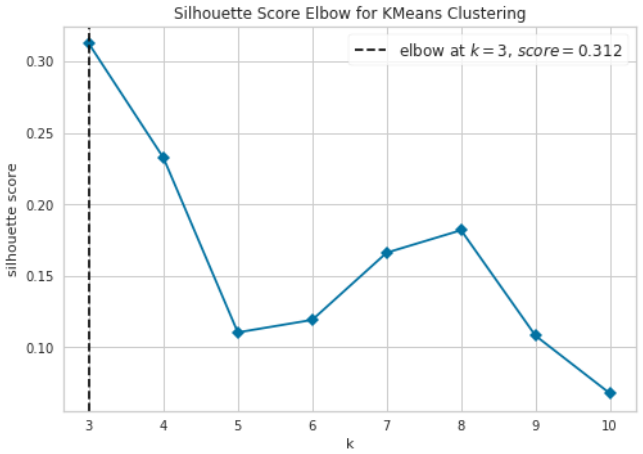
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Caledonia-Fairbanks	Pizza Place	Park	Coffee Shop	Portuguese Restaurant	Bus Line	Grocery Store	Women's Store	Japanese Restaurant	Food Truck	Mexican Restaurant
1	Davisville	Italian Restaurant	Coffee Shop	Sushi Restaurant	Indian Restaurant	Café	Restaurant	Pizza Place	Dessert Shop	Gym	Bakery
2	Davisville North	Coffee Shop	Italian Restaurant	Café	Pizza Place	Dessert Shop	Restaurant	Gym	Fast Food Restaurant	Park	Japanese Restaurant
3	Del Ray	Park	Convenience Store	Grocery Store	Coffee Shop	Sandwich Place	Fast Food Restaurant	Gas Station	Discount Store	Falafel Restaurant	Ethiopian Restaurant
4	East Toronto	Coffee Shop	Café	Sandwich Place	Ethiopian Restaurant	Pizza Place	Convenience Store	Park	Thai Restaurant	Beer Store	Beer Bar

Clustering the Neighborhoods

We will run the k-Means algorithm to build a clustering model with a different number of clusters (k). The features will be the mean of the frequency of occurrence of each venue

category. Using Silhouette Score Elbow, we can measure and plot the clustering performances.

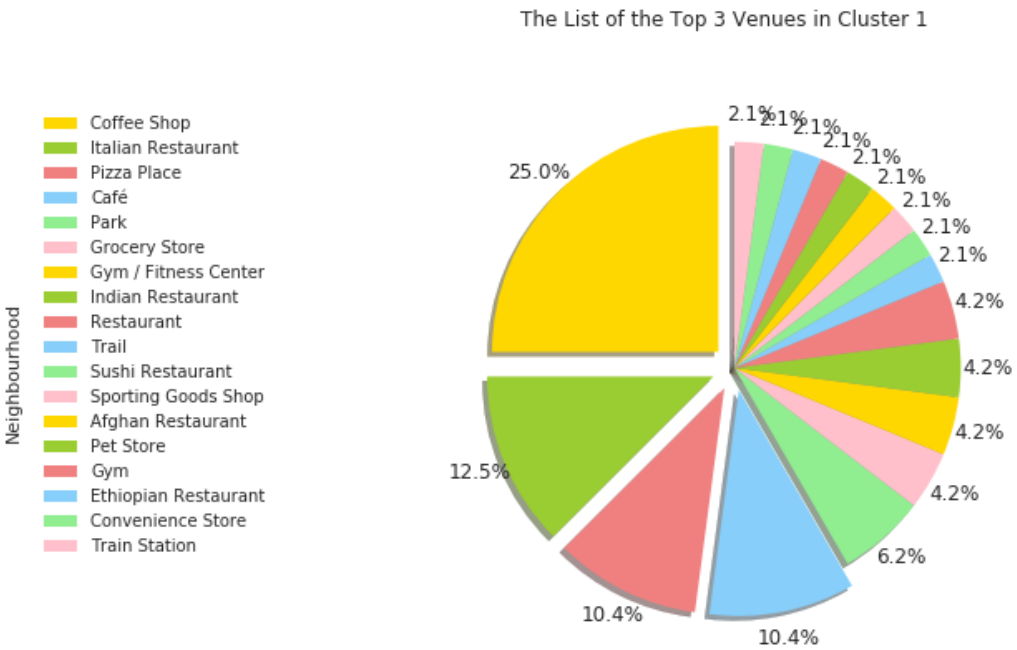
We can inspect that the best k value for this task is 4. Hence, we will have **4 cluster neighborhoods** at the end.



	Borough	Neighbourhood	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	9th Most Common Venue	10th Most Common Venue
0	East York	Parkview Hill	1	Gym / Fitness Center	Pizza Place	Intersection	Office	Rock Climbing Spot	Coffee Shop	Fast Food Restaurant	Soccer Stadium	Pet Store	Brewery
1	East York	Woodbine Heights	1	Pizza Place	Coffee Shop	Ice Cream Shop	Grocery Store	Park	Café	Bank	Bakery	Sushi Restaurant	Arts & Crafts Store
2	York	Humewood-Cedarvale	1	Pizza Place	Coffee Shop	Convenience Store	Beer Store	Grocery Store	Middle Eastern Restaurant	Rental Service	Restaurant	Sandwich Place	Seafood Restaurant
3	York	Caledonia-Fairbanks	1	Pizza Place	Park	Coffee Shop	Portuguese Restaurant	Bus Line	Grocery Store	Women's Store	Japanese Restaurant	Food Truck	Mexican Restaurant
4	East York	Leaside	1	Coffee Shop	Sporting Goods Shop	Furniture / Home Store	Grocery Store	Electronics Store	Department Store	Burger Joint	Shopping Mall	Sports Bar	Restaurant

We Visualise the clusters

We now examine venues listed inside each cluster and define the discriminating venue categories that distinguish them.



We have merged cluster 1, cluster 2, cluster 3, and cluster 4 to have get the most common venues

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Cluster Labels						
0	Parkview Hill	Gym / Fitness Center	Pizza Place	Pet Store	Bus Line	Gastropub
0	Woodbine Heights	Pizza Place	Coffee Shop	Park	Grocery Store	Ice Cream Shop
0	Humewood-Cedarvale	Pizza Place	Coffee Shop	Convenience Store	Hockey Arena	Field
0	Caledonia-Fairbanks	Pizza Place	Park	Coffee Shop	Mexican Restaurant	Beer Store
0	Leaside	Coffee Shop	Sporting Goods Shop	Grocery Store	Electronics Store	Furniture / Home Stor
0	Thorncliffe Park	Indian Restaurant	Afghan Restaurant	Restaurant	Grocery Store	Coffee Shop
0	East Toronto	Coffee Shop	Café	Ethiopian Restaurant	Sandwich Place	Pizza Place
0	Lawrence Park	Café	Trail	Gym / Fitness Center	Park	Coffee Shop

0	Weston	Coffee Shop	Train Station	Pizza Place	Skating Rink	Gift Shop
0	Davisville North	Coffee Shop	Italian Restaurant	Café	Pizza Place	Dessert Shop
0	Forest Hill North & West	Park	Italian Restaurant	Café	Sushi Restaurant	Gym / Fitness Center
0	North Toronto West	Restaurant	Italian Restaurant	Coffee Shop	Sporting Goods Shop	Café
0	The Annex	Coffee Shop	Italian Restaurant	Gym	Café	Pub
0	Davisville	Italian Restaurant	Coffee Shop	Indian Restaurant	Sushi Restaurant	
0	Moore Park	Coffee Shop	Grocery Store	Italian Restaurant	Park	Thai Restaurant
0	Summerhill West	Coffee Shop	Sushi Restaurant	Café	Thai Restaurant	Bank
1	Roselawn	Pharmacy	Skating Rink	Trail	Café	Bank
2	Del Ray	Fast Food Restaurant	Convenience Store	Sandwich Place	Coffee Shop	Grocery Store
3	Runnymede	Brewery	Gas Station	Department Store	BBQ Joint	Coffee Shop

- ***Cluster 1:***

  The 1st cluster has only 18 neighborhoods,with the gym,pizza place,italian restaurant,park and Cafe venues appear to be the most common ones.

- ***Cluster 2: "Pharmacy"***

The 2nd cluster includes 1 neighborhood with pharmacy as the most occurrence venue category.

- ***Cluster 3: "Fast food restaurant"***

The 3rd cluster has 1 neighborhood with fast food restaurant as common venue

- ***Cluster 4: "Brewery"***

  The 4th cluster has 1 neighborhood with Brewery as common venue

Discussion

The project's main goal is to determine the best location for opening a coffee shop business in Toronto. Discussing what locations can be considered "the best" may vary, but we can equate it as the most conducive ones by considering the following criteria:

1. **Safety**

- The conducive locations are supposed to be safe; hence we analyze the crime statistics for all divisions of Toronto Police Service. We conclude that divisions D55, D54, D53, D33, D13, D12, D11 have the lowest crime rates. These cover Central Toronto, West Toronto, York, and East York.

2. **Demographics and Accessibility**

- **Vehicle and foot traffic** are important when we choose a location for the new coffee shop. We have shown the busiest main roads in the city where many are located around downtown. Then, we consider focusing on Central Toronto, York, and East York at first.Still, we need to understand the target market and discuss it further with the team.
- **Accessibility** is also another part to consider. Soon, if we have picked a few location candidates, knowing how and why your customers will get to your location are crucial, such as street visibility, parking slot, and location convenience. Thus, further discussion with the team is again needed.

3. **Neighboring businesses**

- Neighboring businesses can affect the profitability both positively and negatively.
- Cluster 1 has the most coffee shops and restaurants in their neighborhoods.Except these 7 neighborhoods rest 11 neighborhoods of **cluster 1 is recommended.**
- **Cluster 2, and 3 ,4are recommended neighborhoods** to inspect further. However, it is also wise to consider other businesses or amenities surrounding the area to complement your offerings. For example, if we target people who spend mostly for dining out, cluster 3 might be a good choice since it has "fast food restaurant" is the most common venue.

Conclusion

Finding the best location to start a business might not be 100 percent accurate relying on the statistical data. However, we can quickly gain meaningful insights into the city and its neighborhoods with data available today. This helps everyone, including entrepreneurs, business owners, and stakeholders, to make solid decisions based on facts.

By exploring viable coffee shop locations in Toronto, we hone our data exploring and data analysis abilities which will add value as our role in evolving data analysts.

Thank you