Opening a Business in Vancouver using basic Data Analysis

1. Introduction

1.1 Background

Vancouver is a coastal seaport city in western Canada, located in the Lower Mainland region of British Columbia. The Greater Vancouver area had a population of 2,463,431 as in 2016, making it the third-largest metropolitan area in Canada. Crime in different forms is prevalent in Metropolitan cities and Vancouver is no exception. Criminal activity is an ongoing practice by offenders disrupting public peace, people who own commercial establishments especially bear the brunt of these acts. Crimes like break-in a commercial property are on the rise and people thinking to enter similar business should bear in mind criminal activity of the neighbourhood before finalizing a location. We look to address this issue by analyzing the crime data of Vancouver City and finding the safest borough and a neighbourhood within the borough which best suits the requirements of our business problem.

1.2 Problem

This project aims to find a safe and secure location for the opening of commercial establishments in Vancouver, Canada. Specifically, this report will be targeted to stakeholders interested in opening any business place like Grocery Store in Vancouver City, Canada. The first task would be to choose the safest borough by analyzing crime data for opening a grocery store and shortlisting a neighbourhood, where grocery stores are not amongst the most common venues, and yet as close to the city as possible.

We will make use of our data science tools to analyse data and focus on the safest borough and explore its neighbourhoods and the 10 most common venues in each neighbourhood so that the best neighbourhoods where a grocery store is not amongst the most common venue can be selected.

1.3 Interest

Vancouver is one of the most ethnically and linguistically diverse cities in Canada according to that census; 52% of its residents have a first language other than English. Such an ethnically diverse city finding a safe neighbourhood requires a great deal of effort.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

To fetch the crime details of Vancouver I used real-world data set published on Kaggle datasets from here. Though this dataset included a type of crime, recorded time and coordinates of the criminal activity along with neighbourhood, the neighbourhoods were not properly categorized into boroughs which I fetched from Wikipedia from here. Further, the coordinates of the data have been fetched using the OpenCage Geocoder API. Foursquare API is used to fetch venues for the listed neighbourhoods.

Following are the properties of the dataset:

- TYPE - Crime type
- YEAR - Recorded year
- MONTH - Recorded month
- HOUR - Recorded hour
- MINUTE - Recorded minute
- HUNDRED_BLOCK - Recorded block
- NEIGHBOURHOOD - Recorded neighbourhood
- X - GPS longitude
- Y - GPS latitude

The second source of data is based on data from a Wikipedia, which did not require any scraping as it was direct categorizations. The page contains additional information about the neighbourhood and its boroughs.

The third data source is generated from OpenCage API. The data is generated as follows below are the list of columns:

- **Neighbourhood:** Name of the neighbourhood in the Borough.
- **Borough:** Name of the Borough.
- **Latitude:** Latitude of the Borough.
- **Longitude:** Longitude of the Borough.

2.2 Data Cleaning

Data from the Kaggle data source was a heavy file which Git could not accommodate. The Vancouver Crime report had close to ~600,000+ rows of information. Because of the sheer size of the dataset, we choose to take into consideration recent most crimes of the year 2018 which would greatly reduce the number of rows in the dataset.

Since the original data source couldn’t be uploaded to git I processed the dataset in the runtime to filter the records of crimes that took place in the year 2018, created a new CSV out of it using pandas and uploaded it to git hub repository.

	TYPE	YEAR	MONTH	DAY	HOUR	NEIGHBOURHOOD
0	Break and Enter Commercial	2018	3	2	6	West End
1	Break and Enter Commercial	2018	6	16	18	West End
2	Break and Enter Commercial	2018	12	12	0	West End
3	Break and Enter Commercial	2018	4	9	6	Central Business District
4	Break and Enter Commercial	2018	10	2	18	Central Business District

The dataset looks like the above image, after reading it into the data frame.

Due to improper encoding of the co-ordinates of the criminal record, the same coordinates from the data could not be used for plotting because the co-ordinates seemed to be corrupted. Along with X, Y columns in the dataset which represented the GPS co-ordinates of the criminal activity, other fields such as month and hour in which the crime took place has been dropped because they were not in the scope of the problem.

The second source of data is fetched from the Wikipedia page as mentioned in the data section, a new data frame is created based on the data from Vancouver Neighborhood page which on a later stage will be merged with the Crime data table.

```
Total Neighbourhood Count 24 Borough Count 4
```

	Neighbourhood	Borough
0	West End	Central
1	Central Business District	Central
2	Hastings-Sunrise	East Side
3	Grandview-Woodland	East Side
4	Mount Pleasant	East Side

This data has been generated based on data from Wikipedia.

	Type	Year	Month	Day	Hour	Neighbourhood	Borough
0	Break and Enter Commercial	2018	3	2	6	West End	Central
1	Break and Enter Commercial	2018	6	16	18	West End	Central
2	Break and Enter Commercial	2018	12	12	0	West End	Central
3	Break and Enter Commercial	2018	3	2	3	West End	Central
4	Break and Enter Commercial	2018	3	17	11	West End	Central

This is how the data frame looks after merging both the crime and Neighborhood data

After merging the two table, the data frame is further cleaned by dropping records with inconsistent or invalid data like NaN values, to being with exploratory data analysis its essential that we first finish all sorts of data cleaning activities.

Year Type	Break and Enter Commercial	Break and Enter Residential/Other	Mischief	Other Theft	Theft from Vehicle	Theft of Bicycle	Theft of Vehicle	Vehicle Collision or Pedestrian Struck (with Fatality)	Vehicle Collision or Pedestrian Struck (with Injury)	All
Borough										
Central	787	198	2280	2489	6871	857	245	1	314	14042
East Side	786	1043	2192	1674	4754	678	605	8	660	12400
South Vancouver	49	156	187	88	483	36	71	1	111	1182
West Side	403	1000	1062	696	2838	588	225	3	389	7204
All	2025	2397	5721	4947	14946	2159	1146	13	1474	34828

Pivoting the table to represent the data in a format for better understanding.

Along with analyzing the crime data we also have to fetch the latitude and longitude data, to plot the neighbourhoods on the map for better visual depiction, to achieve this we create a data frame similar to the below one:

	Neighbourhood	Borough	Latitude	Longitude
0	Shaughnessy	West Side	49.251863	-123.138023
1	Fairview	West Side	49.264113	-123.126835
2	Oakridge	West Side	49.230829	-123.131134
3	Marpole	West Side	49.209223	-123.136150
4	Kitsilano	West Side	49.269410	-123.155267
5	Kerrisdale	West Side	49.234673	-123.155389
6	West Point Grey	West Side	49.264484	-123.185433
7	Arbutus Ridge	West Side	49.240968	-123.167001
8	South Cambie	West Side	49.246685	-123.120915
9	Dunbar-Southlands	West Side	49.253460	-123.185044

A glimpse of the dataset after fetching the latitude and longitude from OpenCage API.

3. <u>Methodology</u>

3.1 Exploratory Data Analysis

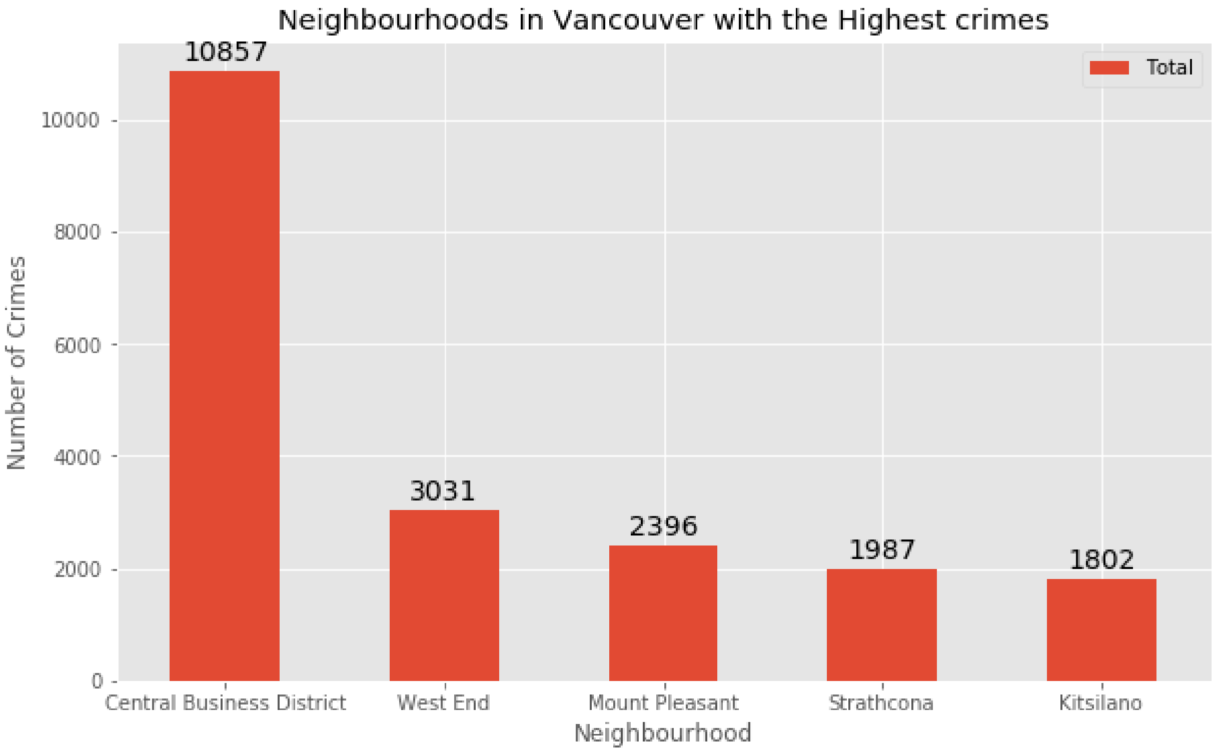
3.1.1 Statistical summary of crimes

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

	YearBreak and Enter Commercial	YearBreak and Enter Residential/Other	YearMischief	YearOther Theft	YearTheft from Vehicle	YearTheft of Bicycle	YearTheft of Vehicle	YearVehicle Collision or Pedestrian Struck (with Fatality)	YearVehicle Collision or Pedestrian Struck (with Injury)
count	4.000000	4.000000	4.00000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000
mean	506.250000	599.250000	1430.25000	1236.750000	3736.500000	539.750000	286.500000	3.250000	368.500000
std	354.409721	488.189427	997.26572	1060.087221	2723.536977	353.955153	226.117226	3.304038	227.060198
min	49.000000	156.000000	187.00000	88.000000	483.000000	36.000000	71.000000	1.000000	111.000000
25%	314.500000	187.500000	843.25000	544.000000	2249.250000	450.000000	186.500000	1.000000	263.250000
50%	594.500000	599.000000	1627.00000	1185.000000	3796.000000	633.000000	235.000000	2.000000	351.500000
75%	786.250000	1010.750000	2214.00000	1877.750000	5283.250000	722.750000	335.000000	4.250000	456.750000
max	787.000000	1043.000000	2280.00000	2489.000000	6871.000000	857.000000	605.000000	8.000000	660.000000

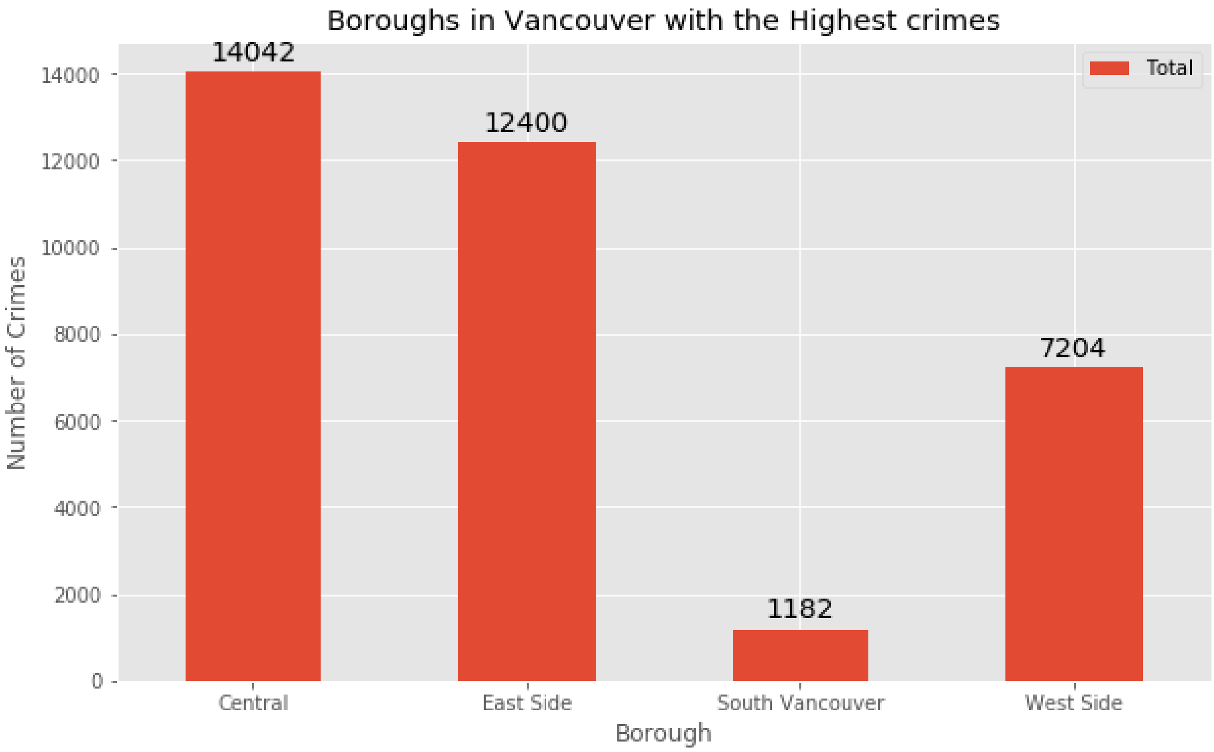
3.1.2 Neighbourhoods with the highest crime rates

Comparing crime rates among all the neighbourhoods we can see that Central Business takes the major chunk of the criminal records which explains why Central Vancouver borough has the most number of crimes which we will explore in a while, the only neighbourhood from the west side borough is Kitsilano which is among the lowest in the top five.



3.1.3 Boroughs crime Analysis

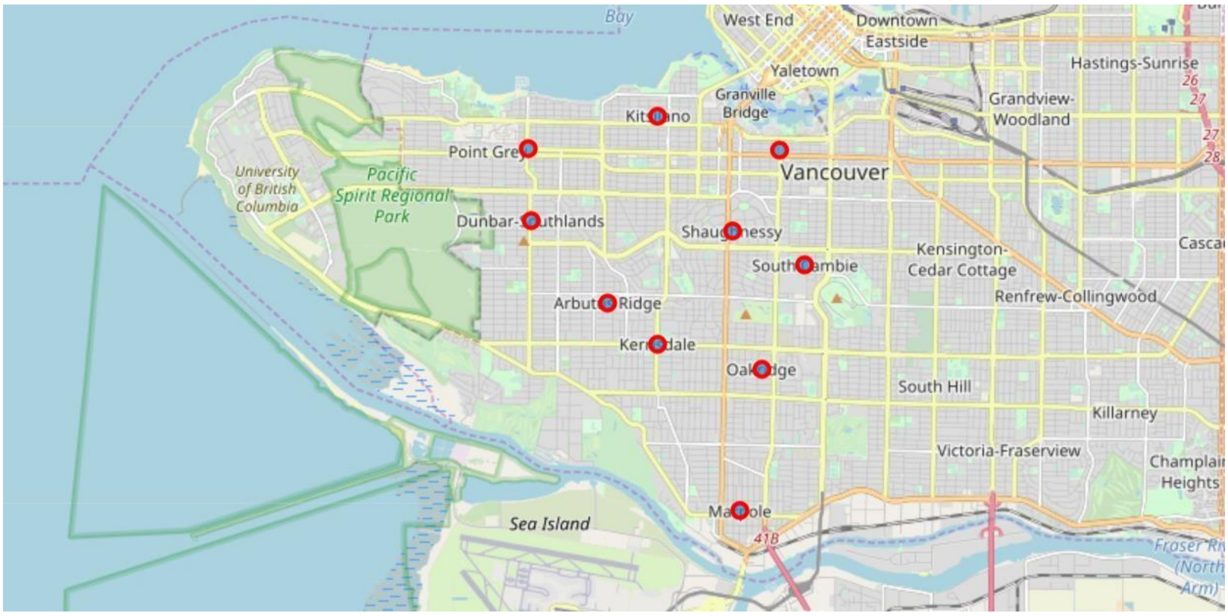
Comparing the crime report in the four boroughs of Vancouver during the year 2018, South Vancouver has the lowest crime rate probably because of its low neighbourhood, followed by West Side which despite having up to 10 neighbourhoods has less number of crimes compared with like of Central Vancouver.



Since South Vancouver has a very little number of neighbourhoods and opening a commercial establishment would not be viable, we can choose the next borough with the lowest crime which is West Side. Westside was chosen because crime type Break Commercial is also low amongst other crimes types which makes West Side ideal destination for the opening of commercial establishments.

3.1.4 Neighbourhoods in West Side, Vancouver, Canada

There are 10 neighbourhoods in the West Side borough colour-coded in a red circle filled with blue, they are visualized on a map using folium library.



3.2 Modelling

Based on the final dataset of neighbourhood and borough along with latitude and longitude of neighbourhoods in West Side Vancouver, we can find all the venues within a 500-meter radius of each neighbourhood by connecting to the Foursquare API. This returns a response in JSON format containing all the venues in each neighbourhood 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:

```
(229, 5)
```

	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category
0	Shaughnessy	49.251863	-123.138023	Bus Stop 50209 (10)	Bus Stop
1	Shaughnessy	49.251863	-123.138023	Angus Park	Park
2	Shaughnessy	49.251863	-123.138023	Crepe & Cafe	French Restaurant
3	Fairview	49.264113	-123.126835	Gyu-Kaku Japanese BBQ	BBQ Joint
4	Fairview	49.264113	-123.126835	CRESCENT nail and spa	Nail Salon

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 neighbourhoods.

To help people find similar neighbourhoods in the safest borough we will be clustering similar neighbourhoods 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 5 for this project that will cluster the 10 neighbourhoods into 5 clusters.

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

4. <u>Results</u>

After running the K-means clustering we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters. Looking into the neighbourhoods in the first cluster

	Borough	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
1	West Side	Coffee Shop	Asian Restaurant	Park	Chinese Restaurant	Sandwich Place	Indian Restaurant	Korean Restaurant	Malay Restaurant	Nail Salon	Fast Food Restaurant
3	West Side	Pizza Place	Chinese Restaurant	Sushi Restaurant	Japanese Restaurant	Lingerie Store	Noodle House	Dim Sum Restaurant	Falafel Restaurant	Plaza	Café
4	West Side	Bakery	Coffee Shop	Sushi Restaurant	American Restaurant	Thai Restaurant	Japanese Restaurant	Tea Room	Food Truck	French Restaurant	Ice Cream Shop
5	West Side	Coffee Shop	Chinese Restaurant	Pharmacy	Tea Room	Sushi Restaurant	Sandwich Place	Fast Food Restaurant	Noodle House	Dessert Shop	Pet Store
6	West Side	Japanese Restaurant	Coffee Shop	Café	Vegetarian / Vegan Restaurant	Bakery	Pub	Sushi Restaurant	Dessert Shop	Pizza Place	Pharmacy
8	West Side	Coffee Shop	Bus Stop	Malay Restaurant	Juice Bar	Cantonese Restaurant	Grocery Store	Sushi Restaurant	Park	Café	Bank

The cluster one is the biggest cluster with 6 of the 10 neighbourhoods in the borough West Side. Upon closely examining these neighbourhoods we can see that the most common venues in these neighbourhoods are Restaurants, eateries, parks and food trucks, the Grocery store is not among the most common venues which make this cluster of neighbourhoods an ideal destination to set up a grocery store.

Looking into the neighbourhoods in the second, third, fourth and fifth clusters, we can see these clusters have only one neighbourhood in each. This is because of the unique venues in each of the neighbourhoods, hence they could not be clustered into similar neighbourhoods.

	Borough	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
7	West Side	Spa	Bakery	Pet Store	Grocery Store	Nightlife Spot	Yoga Studio	Diner	Falafel Restaurant	Fast Food Restaurant	Food

The second cluster has one neighbourhood which consists of Venues mostly utility places like Spa, Yoga studio, pet studio, Grocery store and some Restaurants, Golf courses, and pubs.

	Borough	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	West Side	French Restaurant	Bus Stop	Park	Yoga Studio	Dessert Shop	Diner	Falafel Restaurant	Fast Food Restaurant	Food	Food Truck

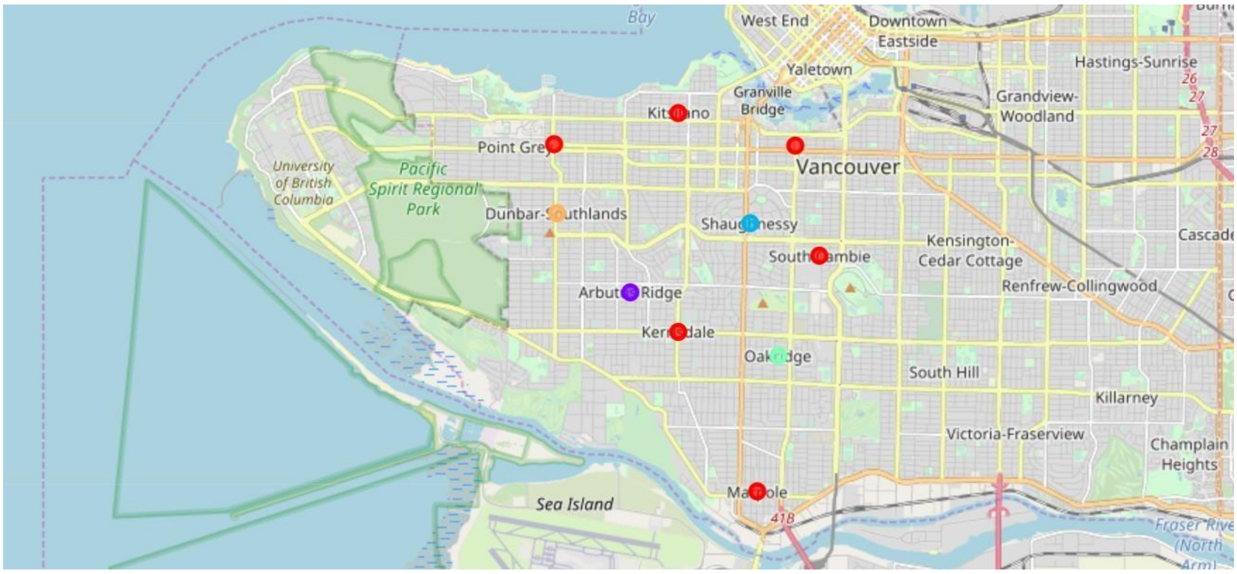
The third cluster has one neighbourhood which consists of Venues such as bus stop, park, and other utility places like Yoga Studio along with restaurants and food trucks.

	Borough	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
2	West Side	Convenience Store	Vietnamese Restaurant	Israeli Restaurant	Fast Food Restaurant	Sandwich Place	Food	Park	Sushi Restaurant	French Restaurant	Dim Sum Restaurant

The fourth cluster has one neighbourhood which consists of Venues such as parks, restaurants, food eateries and park.

	Borough	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
9	West Side	Sushi Restaurant	Liquor Store	Japanese Restaurant	Italian Restaurant	Coffee Shop	Indian Restaurant	Salon / Barbershop	Fast Food Restaurant	French Restaurant	Diner

The fifth cluster has one neighbourhood in it, these neighbourhoods have mostly venues Restaurants of different SEA countries along with a few European, coffee shop and saloon.



Visualizing the clustered neighbourhoods on a map using the folium library.

Each cluster is colour-coded for the ease of presentation; we can see that majority of the neighbourhood falls in the red cluster which is the first cluster. Remaining Neighborhood each is part of the remaining four clusters and has been represented with different colours.

5. Discussion

The objective of the business problem was to help stakeholders identify one of the safest boroughs in Vancouver, and an appropriate neighbourhood within the borough to set up a commercial establishment especially a Grocery store. This has been achieved by first making use of Vancouver crime data to identify a safe borough with a considerable number of neighbourhoods for any business to be viable. After selecting the borough, it was imperative to choose the right neighbourhood where grocery shops were not among venues in proximity to each other. We achieved this by grouping the neighbourhoods into clusters to assist the stakeholders by providing them with relevant data about venues and safety of a given neighbourhood.

6. Conclusion

We have explored the crime data to understand different types of crimes in all neighbourhoods of Vancouver and later categorized them into different boroughs, this helped us group the neighbourhoods into boroughs and choose the safest borough first. Once we confirmed the borough the number of neighbourhoods for consideration also comes down, we further shortlist the neighbourhoods based on the common venues, to choose a neighbourhood which best suits the business problem.