Finding the Safest Neighborhood in Vancouver

Ankur Roy

Background

- The Canadian city of Vancouver is located in the Lower Mainland region of British Columbia. The Greater Vancouver area has a population of about two and a half million, making it the third-largest metropolitan area in Canada. Criminal activity, like breaking and entering or theft, is prevalent throughout this area and can greatly impact business owners. It is therefore important for a new business owner to take crime statistics into account when selecting which neighbourhood, they would like to open their business. By analysing crime data in Vancouver, we aim to determine the safest neighbourhood that is also suitable for opening a small business like a grocery store.

Problem

- will involve first analysing crime data to shortlist safe neighbourhoods where grocery stores are not too common. Using various data science tools, we will pick the safest borough, and then look at its neighbourhoods, before looking at the most common businesses in each neighbourhood in order to select the neighbourhood that has both low crime and a low number of grocery stores.

Data Acquisition

- To fetch the crime details of Vancouver I used real world data set published on Kaggle. Though this dataset included 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. Further the coordinates of the data has been fetched using the OpenCage Geocoder API. Foursquare API is used to fetch venues for the listed neighbourhoods.
- The second source of data is based on data from a Wikipedia, which was not didn't require any scraping as it was direct categorizations. The page contains additional information about the neighbourhood and boroughs. The third data source is generated from OpenCage API
- The third data source is generated from OpenCage API

Data Cleaning

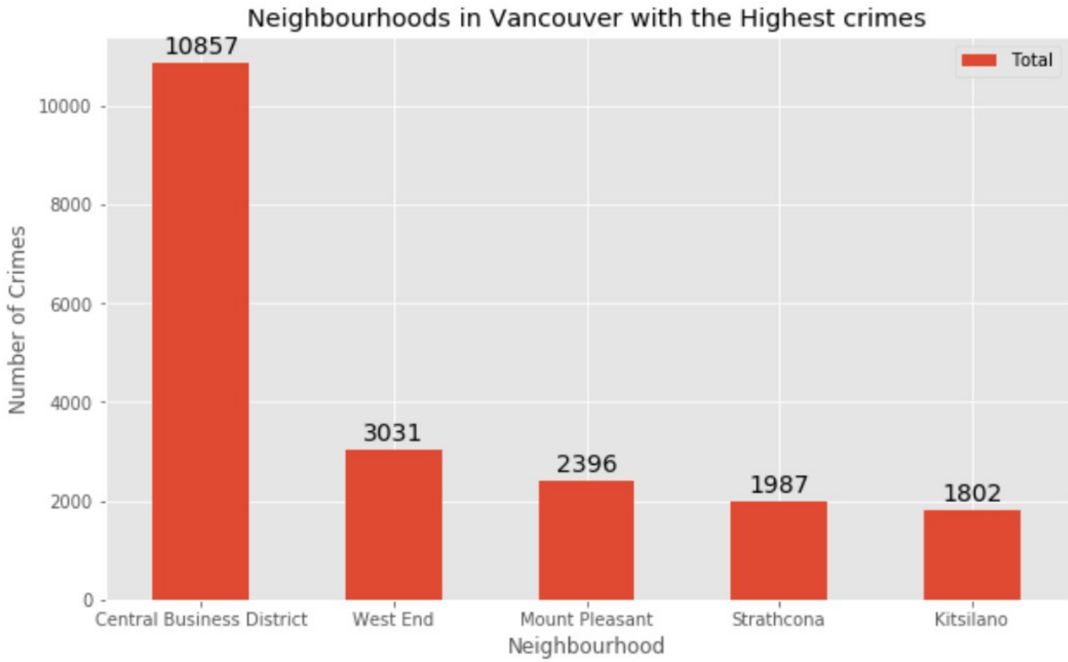
- Data from the kaggle data source was heavy file which Git could not accommodate. The Vancouver Crime report had close to ~600,000+ rows of information. Because of the sheer size of the dataset, we choose to take into consideration recent most crimes of the year 2018 which would greatly reduce the number of row 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.
- Due to improper encoding of the co-ordinates of the crime record, the exact same coordinates from the data couldn’t 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.

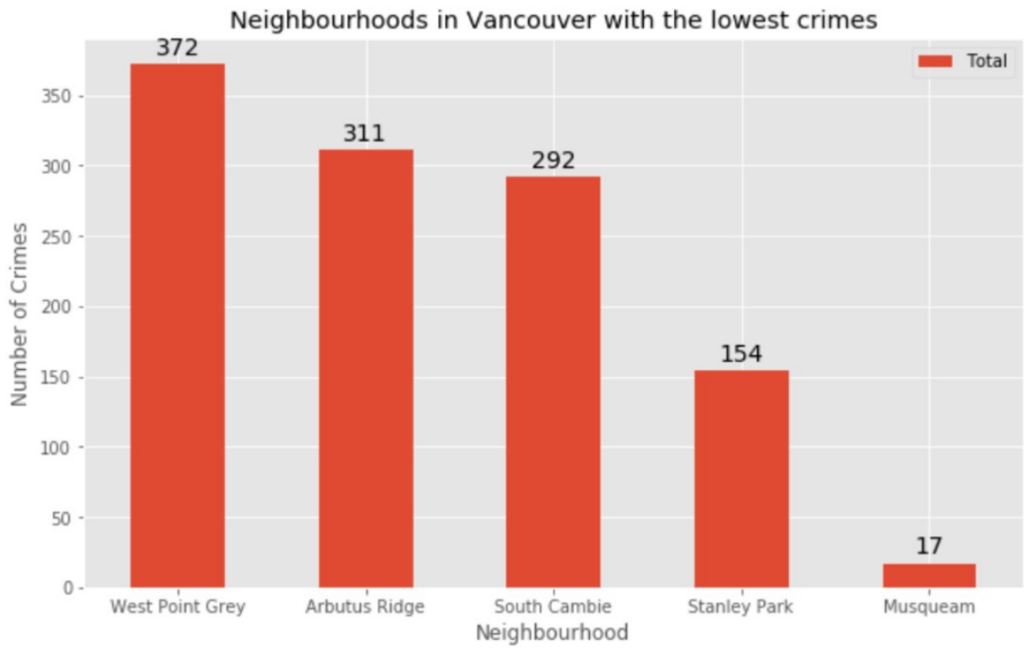
Methodology

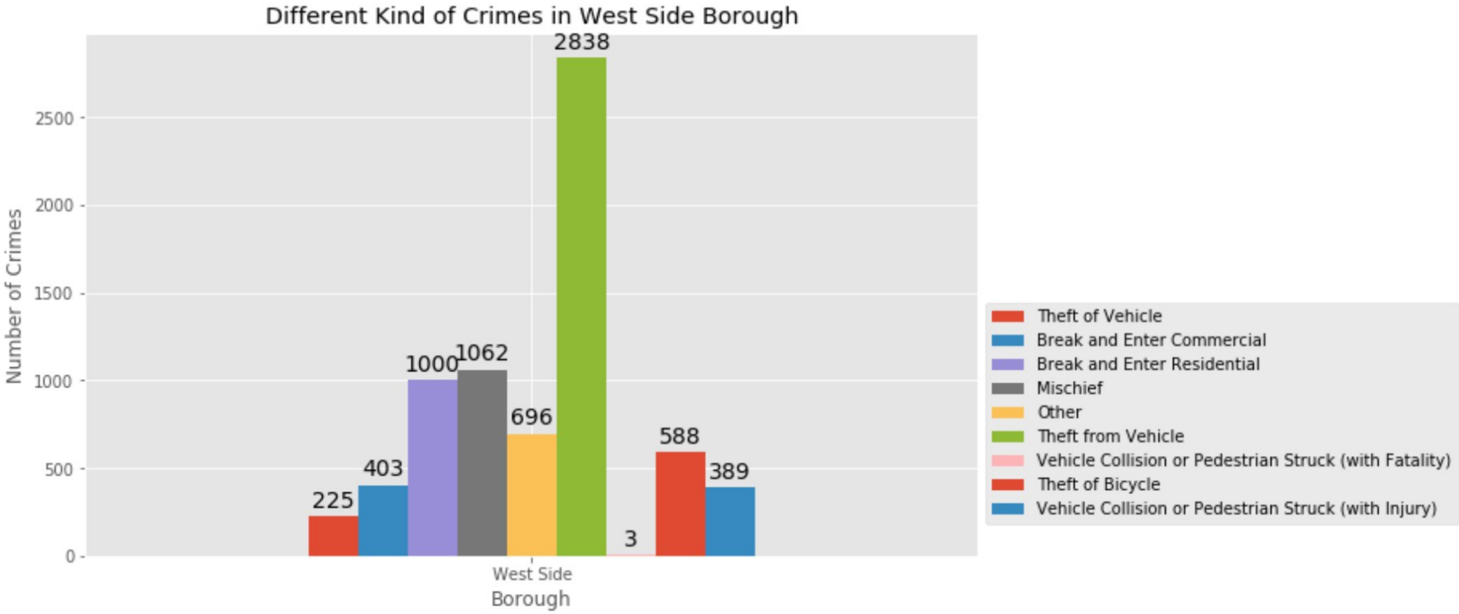
	YearBreak and Enter Commercial	YearBreak and Enter Residential/Other	YearMischief	YearOther Theft	YearTheft from Vehicle	YearTheft of Bicycle	YearTheft of Vehicle	YearVehicle Collision or Pedestrian Struck (with Fatality)	YearVehicle Collision or Pedestrian Struck (with Injury)
count	4.000000	4.000000	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

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 crime categories.

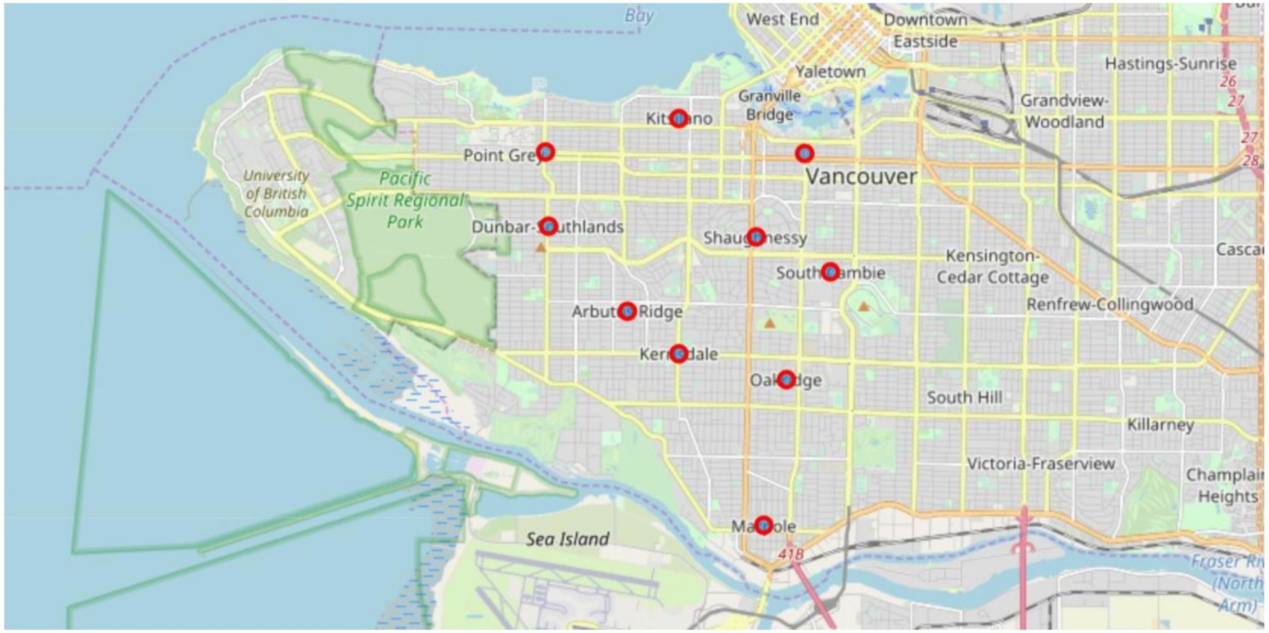
Data Visualisation




Neighbourhoods in Vancouver with the lowest crimes
Total
372
311
292
154
17
350
300
250
200
150
100
50
0
Number of Crimes
West Point Grey
Arbutus Ridge
South Cambie
Stanley Park
Musqueam
Neighbourhood


Different Kind of Crimes in West Side Borough
Number of Crimes
2500
2000
1500
1000
500
0
2838
1000
1062
696
588
403
389
225
3
West Side
Borough
Theft of Vehicle
Break and Enter Commercial
Break and Enter Residential
Mischief
Other
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Bicycle
Vehicle Collision or Pedestrian Struck (with Injury)

West Side Neighborhoods



Modelling

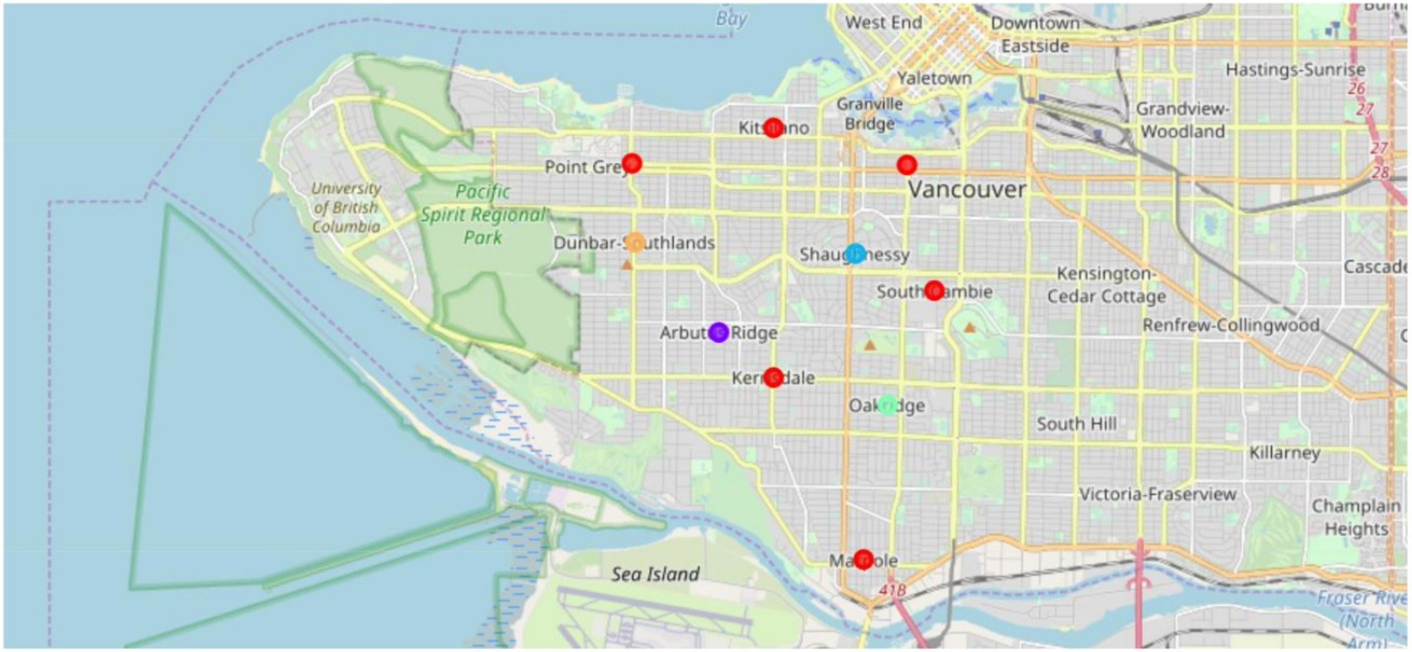
Based on the final dataset of neighbourhood and borough along with latitude and longitude of neighbourhoods in West Side Vancouver, we can find all the venues within a 500-meter radius of each neighbourhood by connecting to the FourSquare API. This returns a response in json format containing all the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

Results

After running the K-means clustering we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters. Here is how the map looks like:



The data of Cluster contains the following 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
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

Discussion

- The objective of the business problem was to help stakeholders identify one of the safest borough 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 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 a close 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.

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. The future scope of this project we can take into considerations population of the neighbourhood which is an additional factor that will have major impact on decision making.