Capstone: The Battle of the Neighborhood

Analyzing Safest Neighborhood in Vancouver

JANUARY 30

Authored by: Ashish Rangari

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 [1]. Vancouver is no exception to crimes in different forms which are prevalent in other metropolitan cities. Criminal activity is an ongoing practice by offenders causing disruption of public peace, business owners. Therefore, it is important to consider the crime rate in the area before opening a business establishment. In this project, this very issue of finding a safe neighborhood is analyzed. For this purpose, the crime data of Vancouver City and finding the safest borough and a neighborhood within the borough is analyzed to resolve the business problem.

1.2. Problem

The aim of this project is to find a safe and secure location for opening of commercial establishments in Vancouver, Canada. Specifically, this report is catered towards individuals that are interested in opening any business place like liquor store in Vancouver City, Canada. The first step is to choose the safest borough by analyzing crime data for opening a liquor store and short listing a neighborhood, where liquor stores are not amongst the most common venues, and yet as close to the city as possible. Data science tools used to analyze data and focus on the safest borough and explore its neighborhoods and the 10 most common venues in each neighborhood. Then the best neighborhood where liquor stores are not amongst the most common venues 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 neighborhood requires great deal of effort.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

To make this project realistic and useful for the user, actual crime rate data set published on Kaggle datasets for Vancouver is used. This dataset included type of crime, recorded time and coordinates of the criminal activity along with neighborhoods. But the neighborhoods were not properly categorized into boroughs which were fetched from Wikipedia. Further the coordinates of the data were fetched using the OpenCage Geocoder API. Foursquare API is used to fetch venues for the listed neighborhoods.

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 neighborhood
- X - GPS longitude
- Y - GPS latitude

The second source of data was extracted from a Wikipedia. This data did not require any scraping, as it was direct categorizations. The page contains additional information about the neighborhood and its boroughs.

The third data source was generated from OpenCage API. The data was generated as follows below are the list of columns:
- Neighborhood: Name of the neighborhood in the Borough.
- Borough: Name of the Borough.
- Latitude: Latitude of the Borough.
- Longitude: Longitude of the Borough.

2.2 Data Cleaning

The data file extracted from Kaggle had close to 600,000 + data point. To simplify the project only 2018 crime data has been analyzed. The reference csv file is uploaded in the git repository.

Figure 1: Image of dataset after reading it into data frame

	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

It was observed that there was improper encoding of the co-ordinates of the crime record. Due to the erroneous nature of the information, these co-ordinates from the data couldn't be used for plotting. 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 was 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. This data frame at later stage was merged with the Crime data table.

Figure 2: Data generated from scrapping Wikipedia

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

Figure 3: Merging Crime and Neighborhood data

	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

After merging the two table, the data frame is further cleaned by dropping records with inconsistent or invalid data like NaN values. This step is important for exploratory data analysis.

Figure 4: Pivoting the table for better visualization

	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

In addition to analyzing the crime data, the latitude and longitude data is also fetched to plot the neighborhoods on the map for visualization by creating a data frame shown in figure 5.

Figure 5: Latitude and Longitude data fetched from OpenCage API

	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

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Statistical crime rate summary

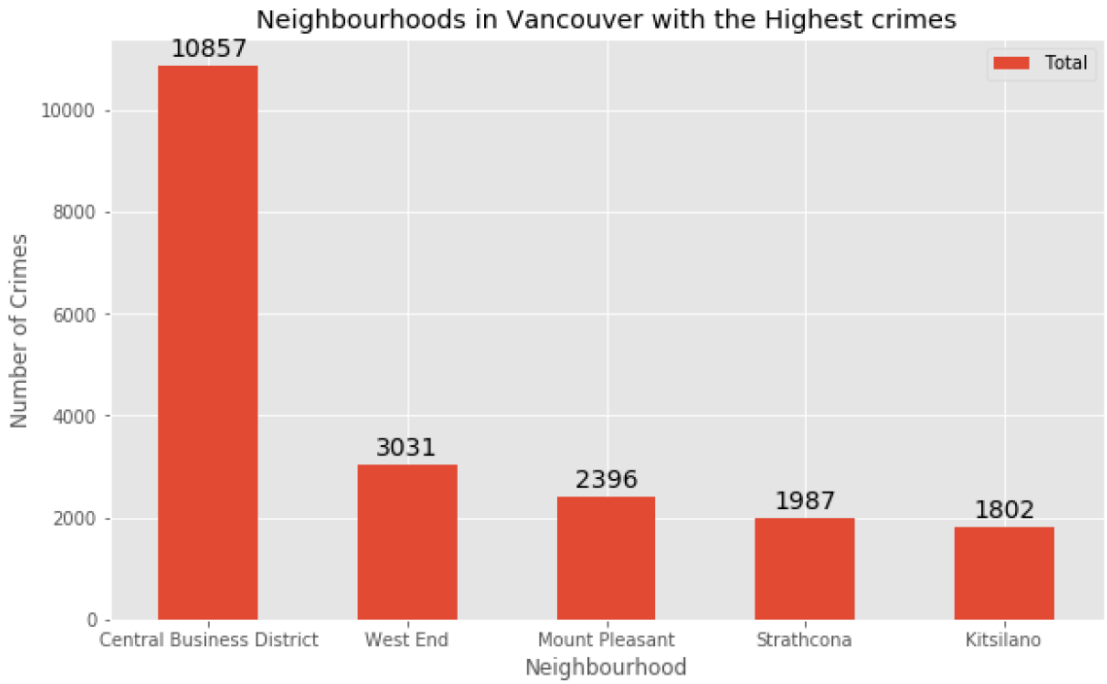
The describe function in python is used extract statistics of the crime data. This function returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crimes.

Figure 6: Describe function results

	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

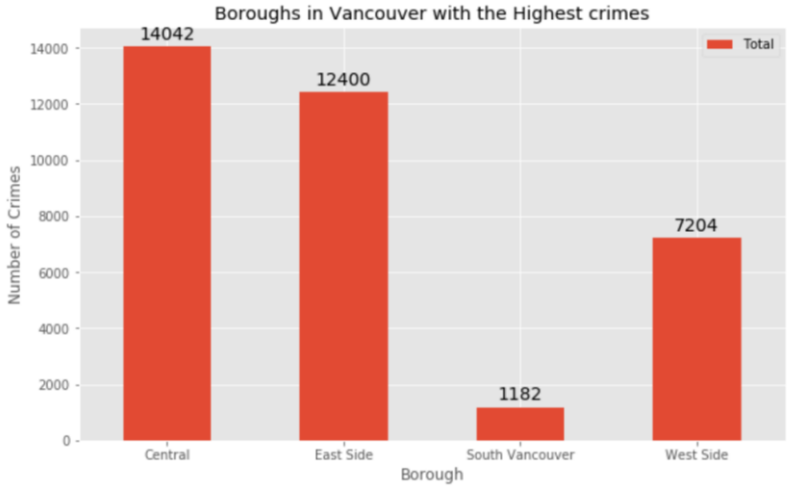
3.1.2 Highest crime rate neighborhood

Comparing crime rates among all the neighborhoods, it can be deduced that Central Business takes the major chunk of the crime records which explains why Central Vancouver borough has the greatest number of crimes which was explored in the next section. The only neighborhood from the west side borough among the lowest in the top five was Kitsilano.



3.1.3 Crime rate analysis in Boroughs

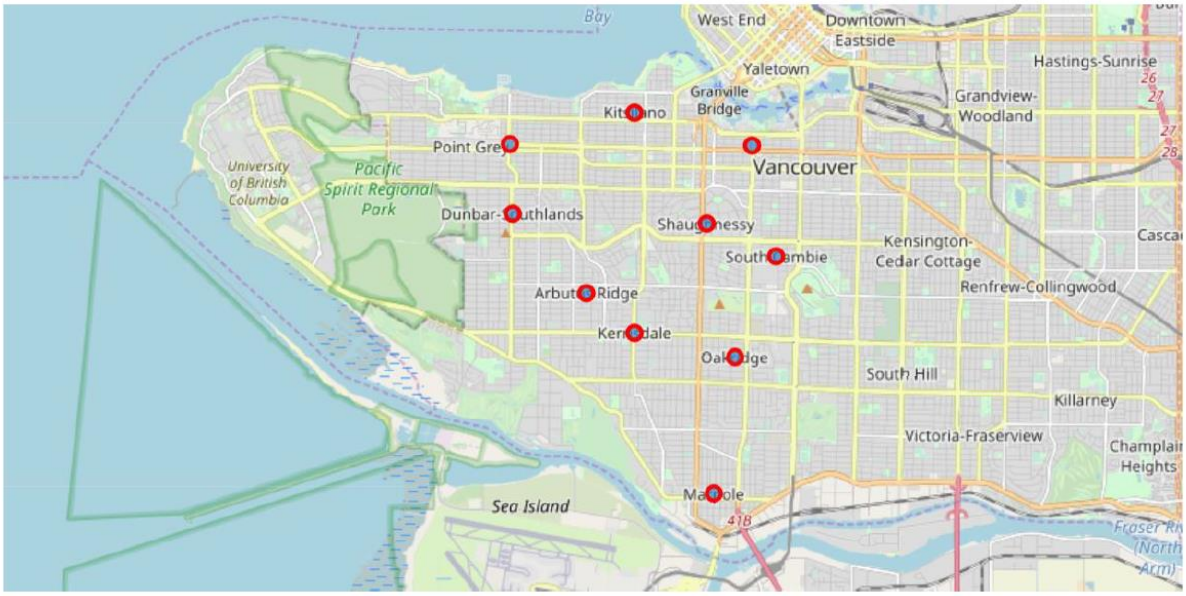
Comparison of the crime report in the four boroughs of Vancouver during the year 2018, it is noticed that South Vancouver has the lowest crime rate probably because of its low neighborhood, followed by West Side which despite having up to 10 neighborhoods has less number of crimes compared with like of Central Vancouver.



Analyzing South Vancouver further, it was noticed that South Vancouver has small number of neighborhoods and opening a commercial establishment would not be viable. Therefore, decided to select the next borough with lowest crime, in this scenario it was the West Side. The West side of Vancouver, was chosen because crime type of break and enter into commercial properties is low amongst other crimes types. This makes the West Side ideal location for opening of commercial establishments.

3.1.4 Analyzing neighborhood in the West Side of Vancouver

There are 10 neighborhoods in the West Side borough color coded in red circle filled with blue. Visualization is created using folium library.



3.2 Modeling

By connecting to the FourSquare API and using the final dataset of neighborhood and borough along with latitude and longitude of neighborhoods in West Side Vancouver, all the venues within a 500-meter radius of each neighborhood can be found. This returns a response in json format containing all the venues in each neighborhood which was converted to a pandas data frame. This data frame contains all the venues along with their coordinates and category.

Figure 7: Information extracted using Foursquare API

(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 was 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 find similar neighborhoods in the safest borough, clustering similar neighborhoods using K - means clustering was done. Which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. A cluster size of 5 was created for this project that will cluster the 10 neighborhoods into 5 clusters.

The idea behind conducting a K- means clustering is to cluster neighborhoods with similar venues together so that the area of interests based on the venues/amenities around each neighborhood can be shortlisted.

4. Result and Discussion

After running the K-means clustering, each cluster created to see which neighborhoods were assigned to each of the five clusters. The neighborhoods in the first cluster are shown in figure 8.

Figure 8: First Cluster results

	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 neighborhoods in the borough West Side. Examining these neighborhoods, it can be inferred that the most common venues in these neighborhoods are Restaurants, eateries, parks and food trucks, Liquor store is not among the most common venues which makes this cluster of neighborhoods an ideal destination to set up a liquor store.

Looking into the neighborhoods in the second, third, fourth and fifth clusters, we can see these clusters have only one neighborhood in each. This is because of the unique venues in each of the neighborhoods, hence they couldn't be clustered into similar neighborhoods.
Following are the results of clusters 2, 3, 4 ,5.

Figure 9: Result from Cluster 2

	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	Bakery	Pet Store	Grocery Store	Spa	Nightlife Spot	Yoga Studio	Gas Station	Dim Sum Restaurant	Diner	Falafel Restaurant

Figure 10: Result from Cluster 3

	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	Park	Yoga Studio	Gastropub	Dim Sum Restaurant	Diner	Falafel Restaurant	Fast Food Restaurant	Food Truck	Gas Station

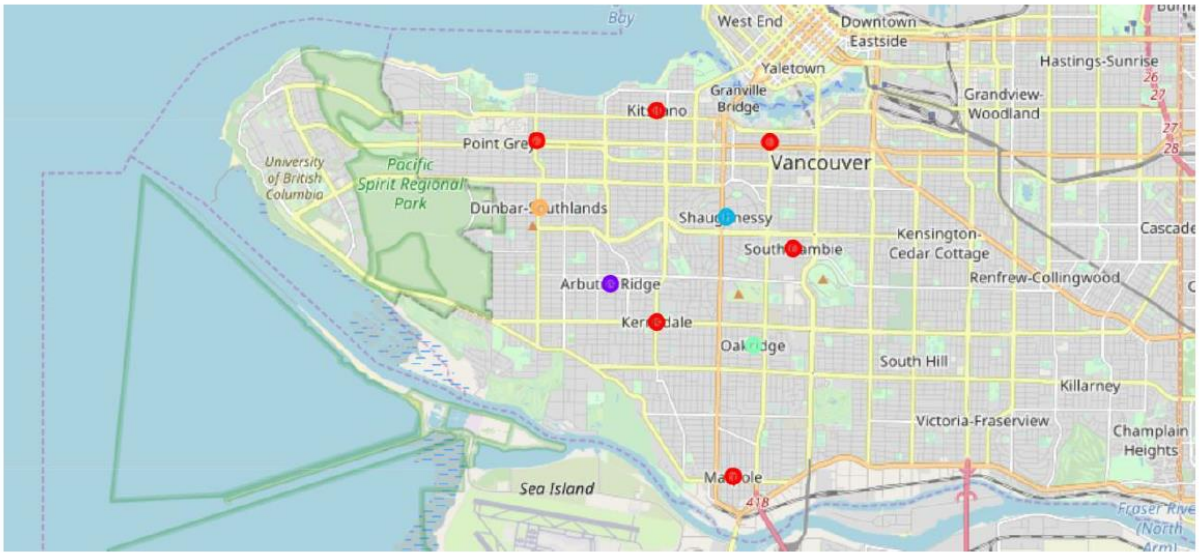
Figure 11: Result from Cluster 4

	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	Vietnamese Restaurant	Convenience Store	Pizza Place	Sandwich Place	Fast Food Restaurant	Sushi Restaurant	Bus Station	Yoga Studio	French Restaurant	Dim Sum Restaurant

Figure 12: Result from Cluster 5

	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	Coffee Shop	Italian Restaurant	Indian Restaurant	Ice Cream Shop	Gas Station	Dim Sum Restaurant	Diner	Falafel Restaurant

Figure 13: Visualization of clustered neighborhood



Each cluster in figure 13 is color coded for the ease of presentation, it can be noted that majority of the neighborhood falls in the red cluster which belongs to the first cluster. Remaining neighborhood are part of remaining four clusters and has been represented with different colors.

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

5. Conclusion

This project has explored the crime data to understand different types of crimes rate in all neighborhoods of Vancouver. Then categorized the data into different boroughs, this helped identify the safest borough. Once the borough was confirmed, the number of neighborhoods for consideration also were significantly reduced. Also, there was further shortlist the neighborhoods based on the common venues, and to choose a neighborhood which best resolves the business problem.