The Battle of the Neighborhoods

CC YAN

AUG/20/2020

1.Introduction

1.1 The business 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 will be targeted to stakeholders interested in opening any business place like Convenience Store in Vancouver, Canada.

1.2 Solution

The first task would be to choose the safest borough by analyzing crime data for opening a convenience store and short listing a neighborhood, where convenience stores are not amongst the most common venues, and yet as close to the city as possible, especially in the living areas.

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

2.Data Collection

2.1 Data Sources

Based on our problem, factors that will influence our decision are:

- finding the safest borough based on crime statistics
- finding the most common venues
- choosing the right neighborhood within the borough

So I will use the geographical coordinates of Vancouver to plot neighborhoods in a borough that is safe, and finally cluster our neighborhoods and present our findings.

Following data sources will be needed to extract/generate the required information:

Vancouver Crimes from 2003 to 2019

https://www.kaggle.com/agilesifaka/vancouver-crime-report/version/2

A dataset consisting of the crime statistics of each neighborhood in Vancouver along with type of crime, recorded year, month and hour.

Vancouver Crimes of 2018
https://raw.githubusercontent.com/RamanujaSVL/Coursera_Capstone/master/vancouver_crime_records_2018.csv

List of neighbourhoods in Vancouver
https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver

2.2 Data choosing and cleaning

After reading from the Dataset, we found that the "Vancouver Crimes from 2003 to 2019" mentioned before contains to many data (about 600,000 rows) to process, so we still choose the recent crime report of the 2018.

We count the number of crimes in different neighborhoods, plus the additional information about the Neighborhood from Wikipedia mentioned above, merge them together, drop the NULL data , make a new dataset and table showing the details of crimes in each area.

```
Central Business District    10857
West End                      3031
Mount Pleasant                2396
Strathcona                    1987
Kitsilano                     1802
Fairview                      1795
Renfrew-Collingwood           1762
Grandview-Woodland            1761
Kensington-Cedar Cottage      1391
Hastings-Sunrise              1270
Sunset                         967
Riley Park                     866
Marpole                        828
Victoria-Fraserview            600
Killarney                      565
Oakridge                       499
Dunbar-Southlands              474
Kerrisdale                     417
Shaughnessy                    414
West Point Grey                372
Arbutus Ridge                  311
South Cambie                   292
Stanley Park                   154
Musqueam                        17
Name: Neighbourhood, dtype: int64
```

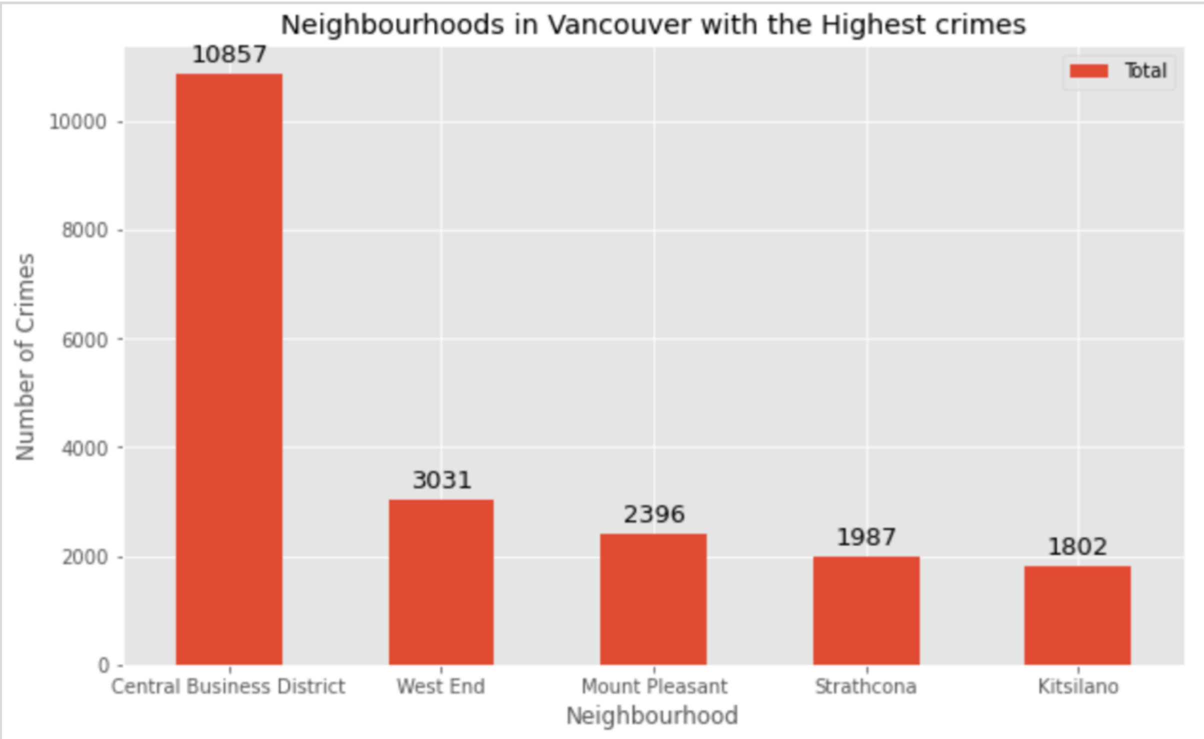
	Neighbourhood	YearBreak and Enter Commercial	YearBreak and Enter Residential/Other	YearMischief	YearOther Theft	YearTheft from Vehicle	YearTheft of Bicycle	YearTheft of Vehicle	YearVehicle Collision or Pedestrian Struck (with Fatality)	YearVehicle Collision or Pedestrian Struck (with Injury)	Total
0	Arbutus Ridge	12	78	49	18	111	12	12	1	18	311
1	Central Business District	551	124	1812	2034	5301	640	165	0	230	10857
2	Dunbar-Southlands	8	106	81	31	199	16	9	1	23	474
3	Fairview	138	73	233	297	692	245	55	0	62	1795
4	Grandview-Woodland	148	162	304	215	634	110	123	0	65	1761

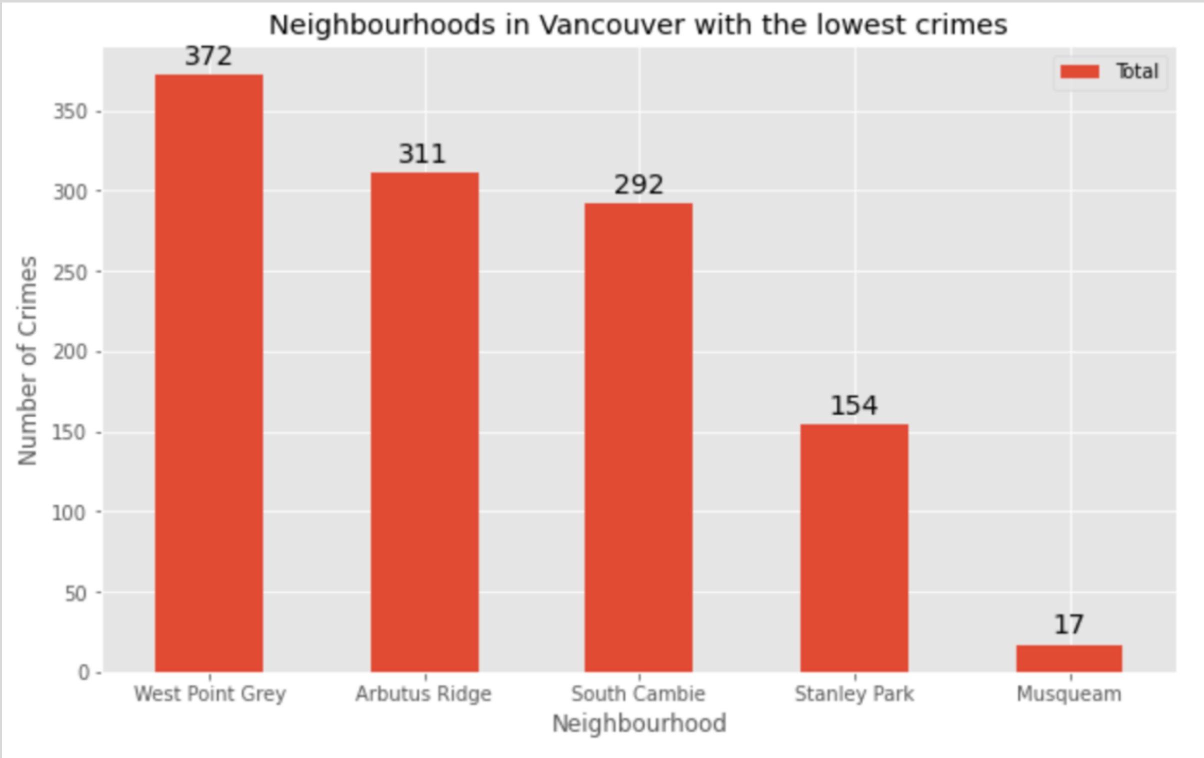
3.Methodology

3.1 Exploratory data analyzes

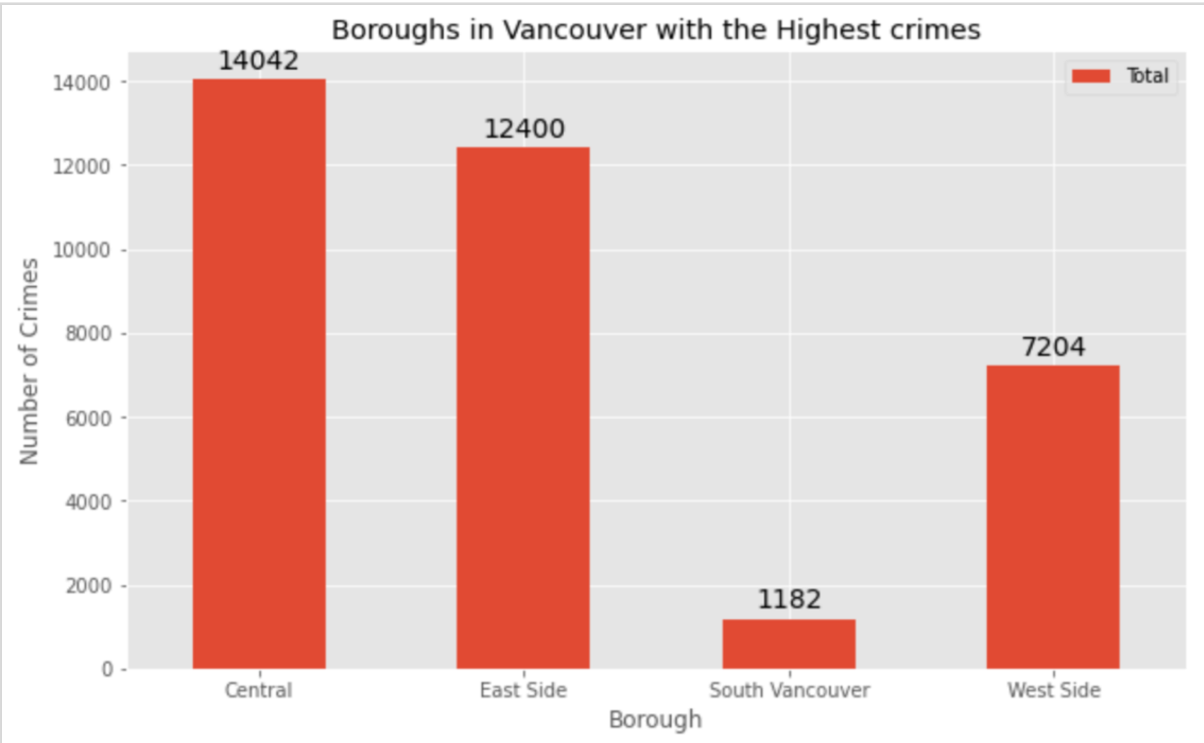
I will visualize the crime repots in different Vancouver boroughs to identify the safest borough and normalize the neighborhoods of that borough. We will use the resulting data and find 10 most common venues in each neighborhood.

First we use Pandas describe() to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values. Then we make a plot to show the neighborhoods in Vancouver with the highest crimes and lowest crimes.



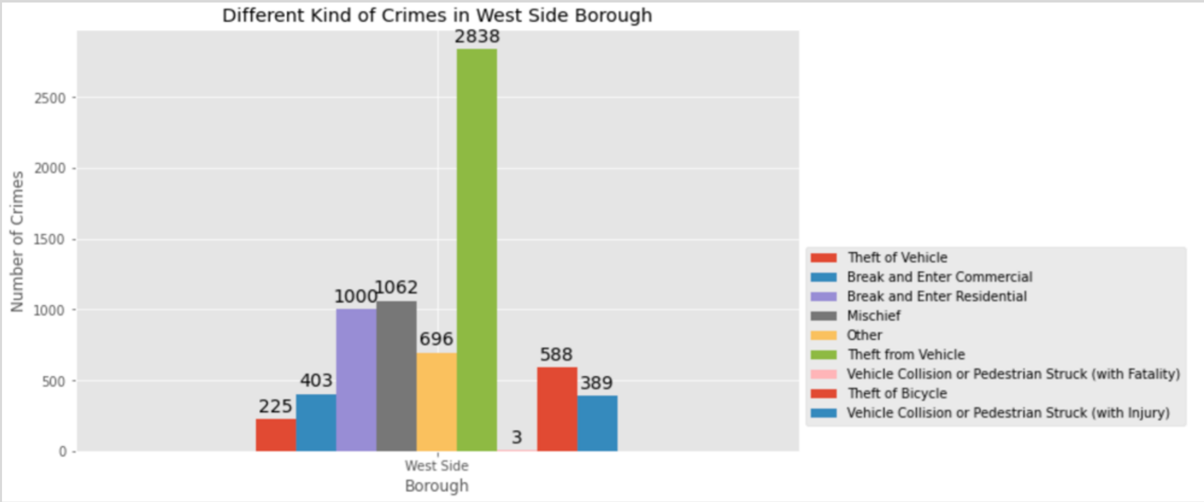


We also found the boroughs in Vancouver with the highest crimes.

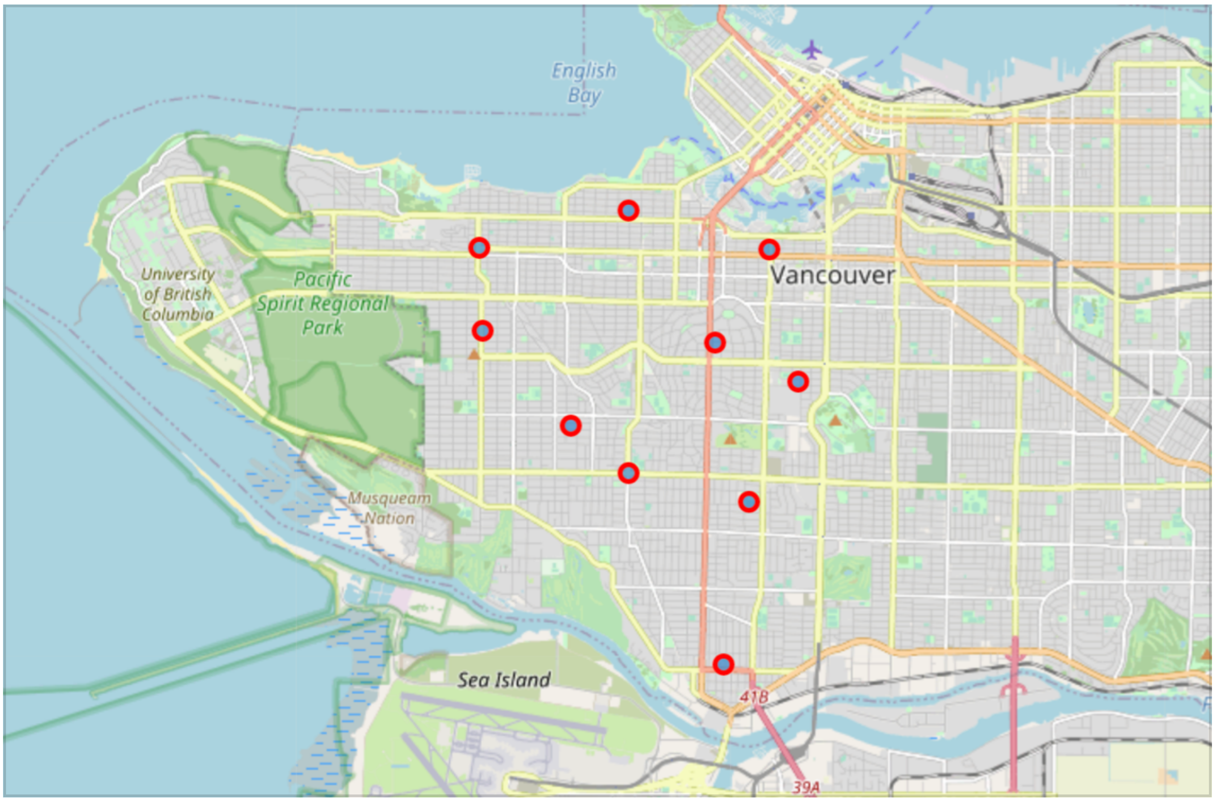


Based on data analysis , it is clear that South Vancouver has the lowest crimes. But South Vancouver has very little number of neighborhoods, so it would make no profit by a convenience store. So we choose the next borough with lowest crime which is West Side.

We take a look at the types of crimes in the West Side Borough, and find the Break and enter Commercial is low amongst other crimes types, which means it may be an ideal place.



Next step is combining these data above to make a new dataset that can show the safest borough on a map with Folium. We create a new dataset of the Neighborhoods, along with their boroughs, crime data and the respective Neighborhood's co-ordinates. And we create a new Data frame with Lat, Lng being fetched from OpenCage geocoder. By fetching the Geographical coordinates of Vancouver to plot on Map, We use Folium to plot Vancouver City's West Side Borough and it's Neighborhoods.



At last I will use Four Square API to explore the neighborhood venues and try machine learning algorithm (like K-Cluster) to make sure the safest and best location for opening a new convenience store.

First we create a new consolidated dataset of the Neighborhoods, boroughs, and the most common venues and the respective neighborhood along with coordinates.This data will be fetched using Four Square API to explore the neighborhood venues and to apply machine learning algorithm to cluster the neighborhoods and present the findings by plotting it on maps using Folium. Then I define a function to get top 10 venues around a given neighborhood, then create a data frame containing venues for each neighborhood in West Side, use one hot encoding to analyze each neighborhood.

	Neighbourhood	American Restaurant	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bank	Bar	Beach	Bookstore	...	Tea Room	Tennis Court	Thai Restaurant	Thrift / Vintage Store	Vegetarian / Vegan Restaurant	Video Store	Vietnamese Restaurant	Waterfall	Wi Sh
0	Arbutus Ridge	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
1	Dunbar-Southlands	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
2	Fairview	0.000000	0.000000	0.074074	0.037037	0.000000	0.037037	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.037037	0.037037	0.0000
3	Kerrisdale	0.000000	0.000000	0.026316	0.000000	0.026316	0.026316	0.000000	0.000000	0.000000	...	0.052632	0.000000	0.026316	0.000000	0.000000	0.000000	0.026316	0.000000	0.0000
4	Kitsilano	0.040816	0.000000	0.020408	0.000000	0.081633	0.000000	0.000000	0.020408	0.000000	...	0.020408	0.020408	0.040816	0.000000	0.020408	0.000000	0.000000	0.000000	0.0204
5	Marpole	0.000000	0.000000	0.000000	0.000000	0.000000	0.030303	0.030303	0.000000	0.000000	...	0.000000	0.000000	0.030303	0.000000	0.000000	0.030303	0.030303	0.000000	0.0000
6	Oakridge	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.000000	0.0000
7	Shaughnessy	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
8	South Cambie	0.000000	0.000000	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.058824	0.000000	0.0000
9	West Point Grey	0.000000	0.022222	0.022222	0.000000	0.022222	0.022222	0.022222	0.000000	0.044444	...	0.000000	0.000000	0.000000	0.022222	0.044444	0.000000	0.000000	0.000000	0.0222

10 rows × 91 columns

And find the 5 most common venues across neighborhoods.

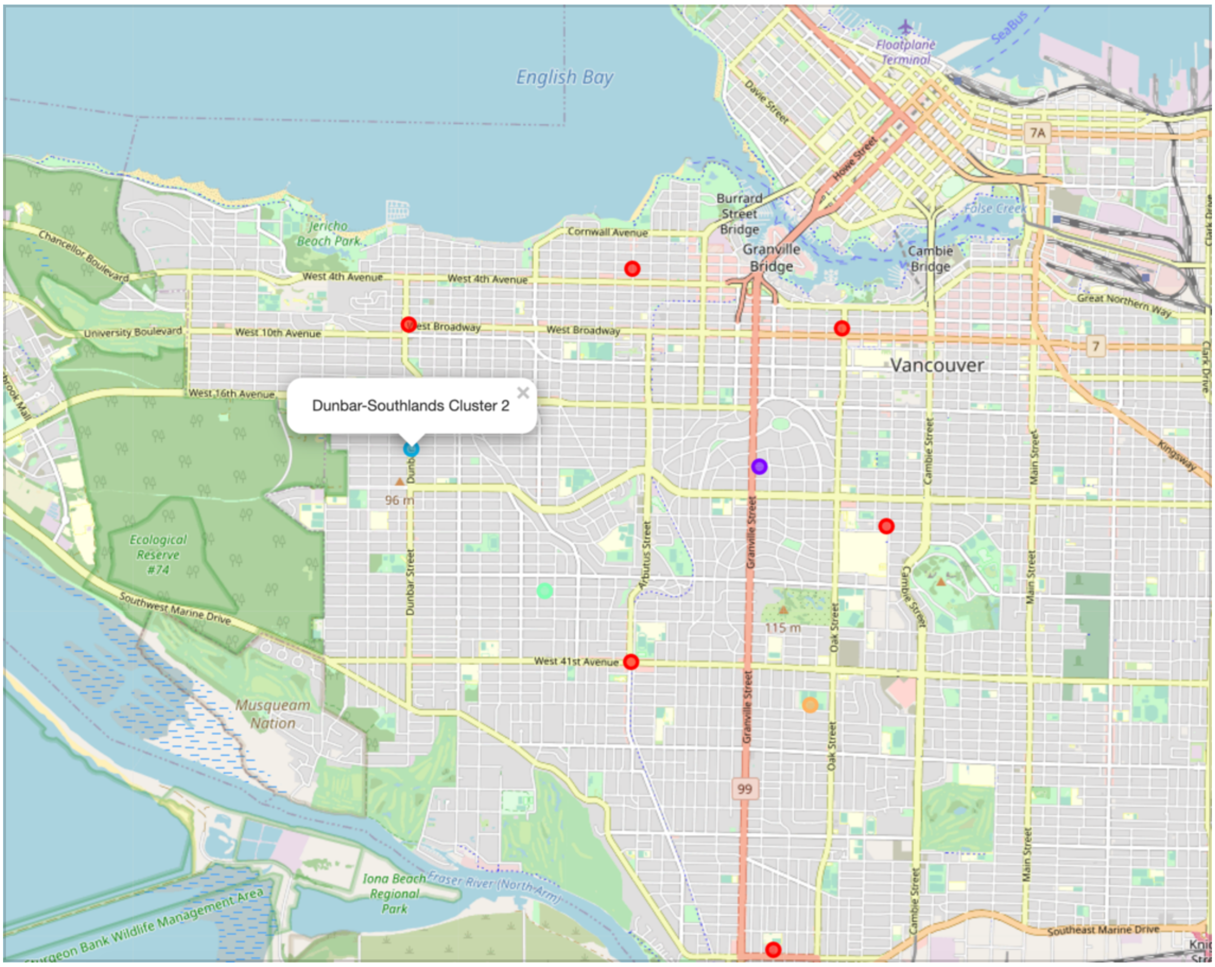
```
----Arbutus Ridge----
             venue  freq
0          Bakery   0.2
1  Nightlife Spot   0.2
2             Spa   0.2
3   Grocery Store   0.2
4       Pet Store   0.2


----Dunbar-Southlands----
                  venue  freq
0      Sushi Restaurant  0.33
1           Coffee Shop  0.17
2    Italian Restaurant  0.17
3        Ice Cream Shop  0.17
4  Sporting Goods Shop  0.17


----Fairview----
                  venue  freq
0           Coffee Shop  0.15
1      Asian Restaurant  0.07
2                  Park  0.07
3      Sushi Restaurant  0.04
4  Szechuan Restaurant  0.04


    Kerrisdale
```

```
----Marpole----
                venue  freq
0    Sushi Restaurant  0.09
1        Dessert Shop  0.06
2         Coffee Shop  0.06
3  Chinese Restaurant  0.06
4     Bubble Tea Shop  0.06


----Oakridge----
                  venue  freq
0     Convenience Store  0.17
1      Sushi Restaurant  0.17
2        Sandwich Place  0.17
3  Vietnamese Restaurant  0.17
4   Fast Food Restaurant  0.17


----Shaughnessy----
                  venue  freq
0                  Park   0.5
1     French Restaurant   0.5
2   American Restaurant   0.0
3              Pharmacy   0.0
4             Rest Area   0.0


----South Cambie----
                  venue  freq
0           Coffee Shop  0.29
1              Bus Stop  0.18
2             Juice Bar  0.06
3  Cantonese Restaurant  0.06
4         Grocery Store  0.06


----West Point Grey----
```

3.2 Modelling

To help stakeholders choose the right neighborhood within a borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use K-Means clustering to address this problem so as to group data based on existing venues which will help in the decision making process.
We create the new data frame and display the top 10 venues for each neighborhood. Then cluster neighborhoods.

	Neighbourhood	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	Arbutus Ridge	Spa	Bakery	Grocery Store	Pet Store	Nightlife Spot	Yoga Studio	Cosmetics Shop	Deli / Bodega	Dessert Shop	Dim Sum Restaurant
1	Dunbar-Southlands	Sushi Restaurant	Ice Cream Shop	Sporting Goods Shop	Coffee Shop	Italian Restaurant	Indian Restaurant	Fast Food Restaurant	Convenience Store	Cosmetics Shop	Deli / Bodega
2	Fairview	Coffee Shop	Park	Asian Restaurant	Korean Restaurant	Malay Restaurant	Chinese Restaurant	Camera Store	Restaurant	Nail Salon	Salon / Barbershop
3	Kerrisdale	Coffee Shop	Chinese Restaurant	Pharmacy	Sushi Restaurant	Tea Room	Sandwich Place	Noodle House	Café	Portuguese Restaurant	Pizza Place
4	Kitsilano	Bakery	Food Truck	American Restaurant	Coffee Shop	Japanese Restaurant	Ice Cream Shop	Sushi Restaurant	French Restaurant	Thai Restaurant	Clothing Store



Finally we need to examine the results.

Cluster 1

```
vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 0, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

	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	Park	Asian Restaurant	Korean Restaurant	Malay Restaurant	Chinese Restaurant	Camera Store	Restaurant	Nail Salon	Salon / Barbershop
3	West Side	Sushi Restaurant	Coffee Shop	Chinese Restaurant	Pizza Place	Dessert Shop	Bus Stop	Bubble Tea Shop	Liquor Store	Noodle House	Plaza
4	West Side	Bakery	Food Truck	American Restaurant	Coffee Shop	Japanese Restaurant	Ice Cream Shop	Sushi Restaurant	French Restaurant	Thai Restaurant	Clothing Store
5	West Side	Coffee Shop	Chinese Restaurant	Pharmacy	Sushi Restaurant	Tea Room	Sandwich Place	Noodle House	Café	Portuguese Restaurant	Pizza Place
6	West Side	Café	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Bus Station	Pizza Place	Bookstore	Pub	Vegetarian / Vegan Restaurant	Fast Food Restaurant
8	West Side	Coffee Shop	Bus Stop	Juice Bar	Park	Vietnamese Restaurant	Cantonese Restaurant	Bank	Malay Restaurant	Grocery Store	Sushi Restaurant

Cluster 2

```
vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 1, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

	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	Park	French Restaurant	Coffee Shop	Cosmetics Shop	Deli / Bodega	Dessert Shop	Dim Sum Restaurant	Diner	Falafel Restaurant	Fast Food Restaurant

Cluster 3

```
vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 2, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

	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	Ice Cream Shop	Sporting Goods Shop	Coffee Shop	Italian Restaurant	Indian Restaurant	Fast Food Restaurant	Convenience Store	Cosmetics Shop	Deli / Bodega

Cluster 4

```
vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 3, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

	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	Grocery Store	Pet Store	Nightlife Spot	Yoga Studio	Cosmetics Shop	Deli / Bodega	Dessert Shop	Dim Sum Restaurant

Cluster 5

```
vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 4, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

	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	Sushi Restaurant	Convenience Store	Vietnamese Restaurant	Café	Sandwich Place	Fast Food Restaurant	Food Truck	Cosmetics Shop	Deli / Bodega	Dessert Shop

4.Results and Discussion

After selecting the borough it was imperative to choose the right neighborhood where convenience store were not among venues in a close proximity to each other. We achieved this by grouping the neighborhoods into clusters to assist by providing them with relevant data about venues and safety of a given neighborhood.

5.Conclusion

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