# **BATTLE**

# **OF**

# **NEIGHBORHOODS**

1. Introduction

**1.1 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 **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 **short listing a neighborhood**, 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 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.

1. Data

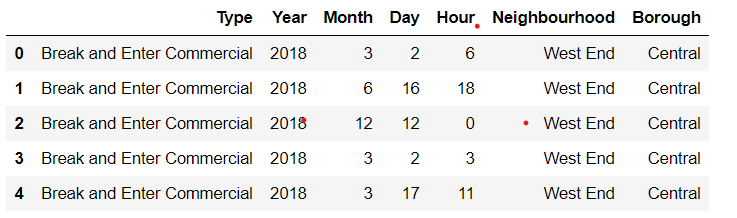
Based on definition of our problem, factors that will influence our decision are:

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

We will be using the geographical coordinates of Vancouver to plot neighborhoods in a borough that is safe and in the city's vicinity, and finally cluster our neighborhoods and present our findings.

Following are different dimensions used in the dataset for analysis.

* Type
* Year
* Month
* Hour
* Minute
* Hundred\_Block
* Neighborhood
* X-longitude
* Y-Latitude

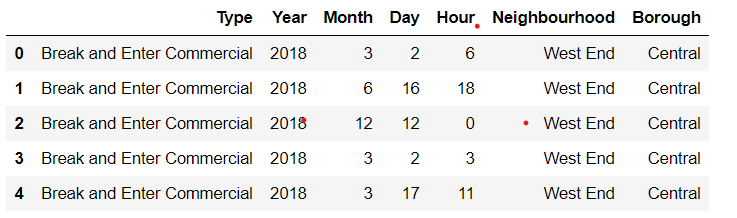


Below are the list of Columns:

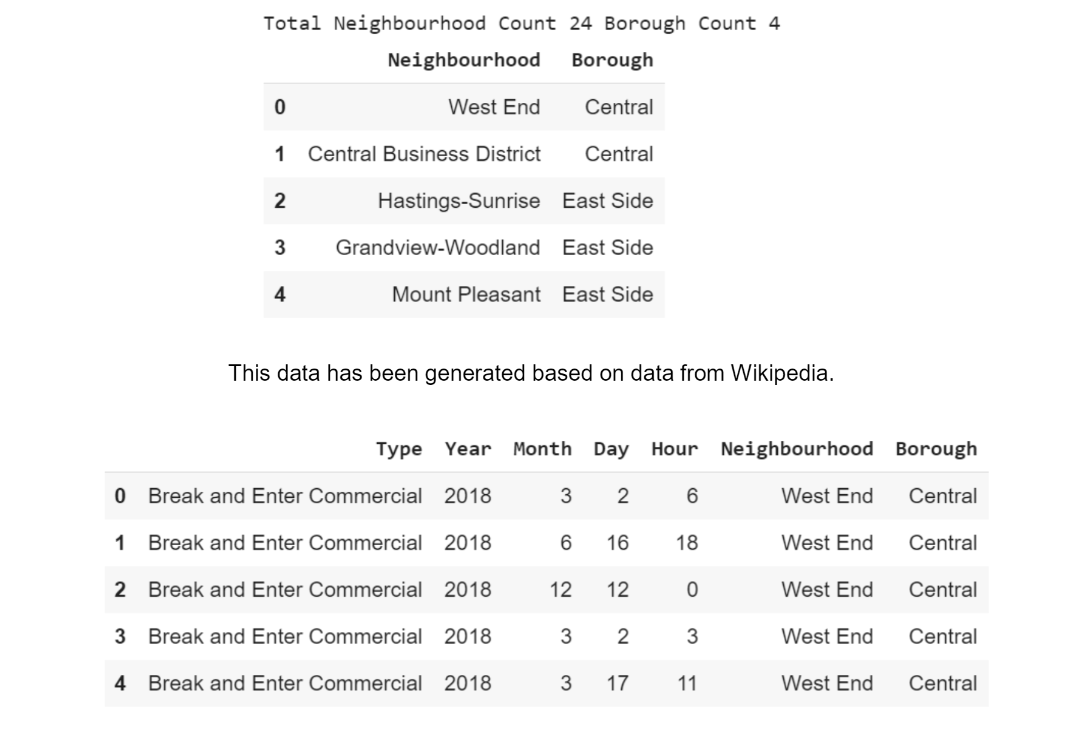
1. Neighborhood
2. Borough
3. Latitude
4. Longitude

2.2 Data Cleaning

The original data fetched from Kaggle is very huge and Git cannot accommodate such huge data. Hence I processes the data set and applied a filter to return rows /crime data happened in the year of 2018.



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 later stage will be merged with the Crime data table.



Below is the dataset after fetching the latitude and longitude from Open cage API.



1. Methodology

Categorized the methodology section into two parts:

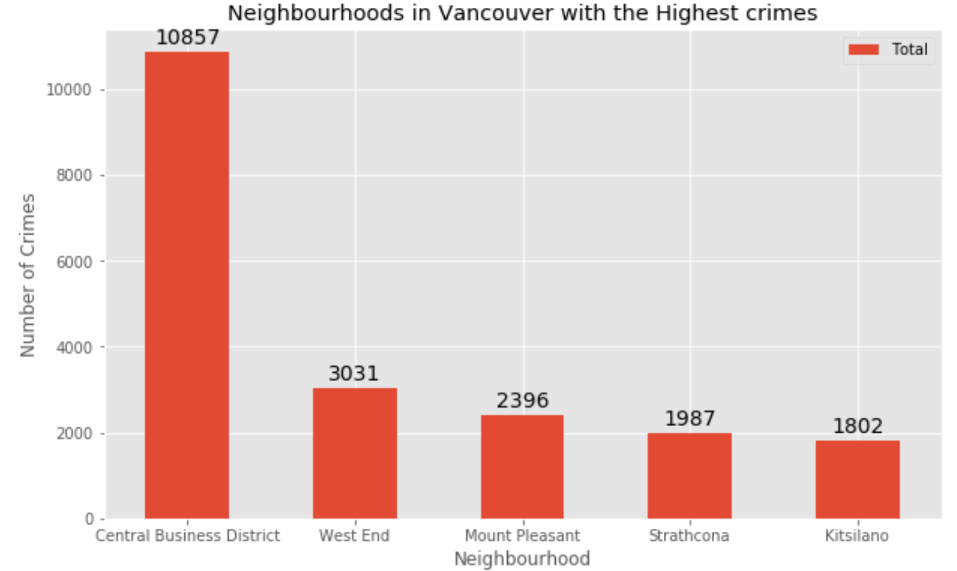
[**Exploratory Data Analysis**:](https://render.githubusercontent.com/view/ipynb?commit=434e438955d556ac9873b747f4d8aca0a4dc79db&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f52616d616e756a6153564c2f436f7572736572615f43617073746f6e652f343334653433383935356435353661633938373362373437663464386163613061346463373964622f43617073746f6e655f5361666573745f4e65696768626f72686f6f645f696e5f56616e636f757665722e6970796e62&nwo=RamanujaSVL%2FCoursera_Capstone&path=Capstone_Safest_Neighborhood_in_Vancouver.ipynb&repository_id=223603116&repository_type=Repository#eda) Visualize the crime reports in different Vancouver boroughs to identity 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.

[**Modelling**:](https://render.githubusercontent.com/view/ipynb?commit=434e438955d556ac9873b747f4d8aca0a4dc79db&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f52616d616e756a6153564c2f436f7572736572615f43617073746f6e652f343334653433383935356435353661633938373362373437663464386163613061346463373964622f43617073746f6e655f5361666573745f4e65696768626f72686f6f645f696e5f56616e636f757665722e6970796e62&nwo=RamanujaSVL%2FCoursera_Capstone&path=Capstone_Safest_Neighborhood_in_Vancouver.ipynb&repository_id=223603116&repository_type=Repository#mdl) 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.

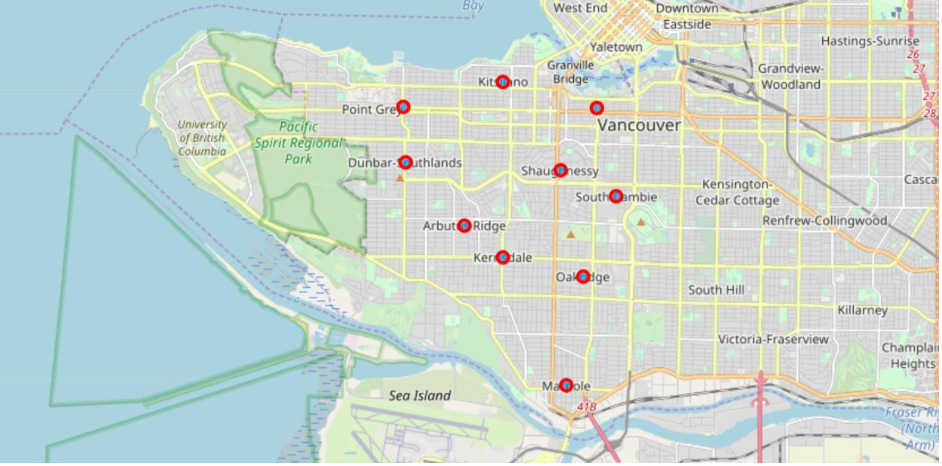
Below is the statistical summary of the crime data.



Below is the visualization of Neighborhood with highest crime data.



As per the observations, below are the neighborhoods in Westside, Vancouver plotted using folium library.

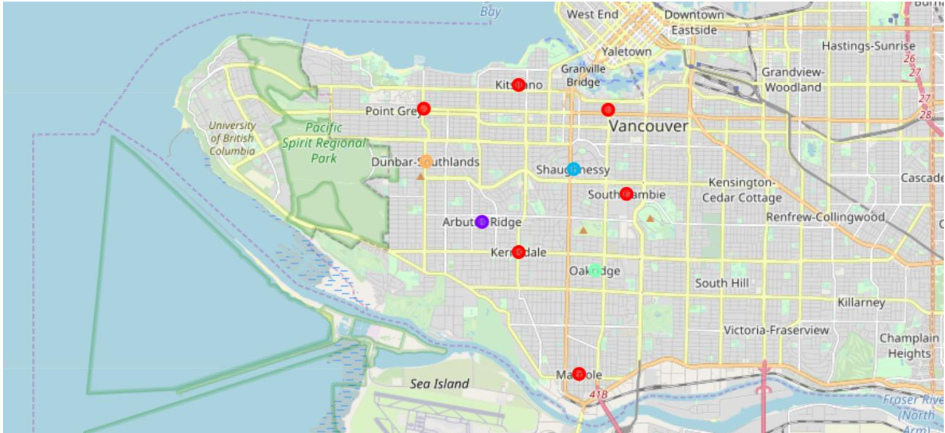


**Modelling:**

To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using k-means clustering. K-means is a form of unsupervised machine learning algorithm the clusters databased on predefined cluster size. The reason to cluster using k-means clustering is to cluster neighborhoods with similar venues together so that people can shortlist area based on their liking.

1. Results

The objective of the business problem was to help stakeholders identify one of the safest borough in Vancouver, and an appropriate neighborhood 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 boroughs with considerable number of neighborhood for any business to be viable. After selecting the borough it was imperative to choose the right neighborhood where grocery shops were not among venues in a close proximity to each other. We achieved this by grouping the neighborhoods into clusters to assist the stakeholders by providing them with relevant data about venues and safety of a given neighborhood.



1. 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.