Najdi Aboubakr

The battle of Neighbours – Week 2

Applied Data Science Capston

# Introduction

## 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 a prevalent distress to the people in Metropolitan cities and Vancouver is no exception. Criminal activity is an ongoing practice by offenders causing disruption of public peace, people who own commercial establishments especially bear the brunt of these acts. Crimes like break into commercial property to for theft are on rise and people thinking to enter into similar business should bear in mind criminal activity of the neighborhood 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 neighborhood with in the borough which best suits the requirements of our business problem.

## 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.

## 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.

# Data Acquisition and Cleaning

## Data Acquisition

To fetch the crime details of Vancouver I used real world data set published on Kaggle datasets from [here](https://www.kaggle.com/agilesifaka/vancouver-crime-report/version/2). Though this dataset included type of crime, recorded time and coordinates of the criminal activity along with neighborhood, the neighborhoods were not properly categorized into boroughs which I fetched from Wikipedia from [here](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver). Further the coordinates of the data has been 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 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 neighborhood and its boroughs.

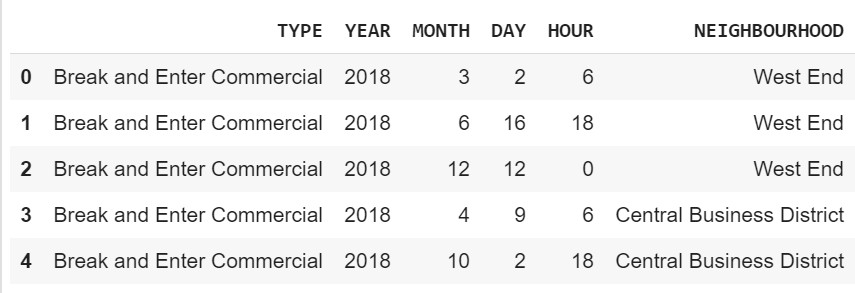
The third data source is generated from OpenCage API. The data is generated as follows below are the list of columns:

* **Neighborhood:** Name of the neighborhood in the Borough.
* **Borough:** Name of the Borough.
* **Latitude:** Latitude of the Borough.
* **Longitude:** Longitude of the Borough.

## 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.



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

Due to improper encoding of the co-ordinates of the crime record, the exact same co- ordinates 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.

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.

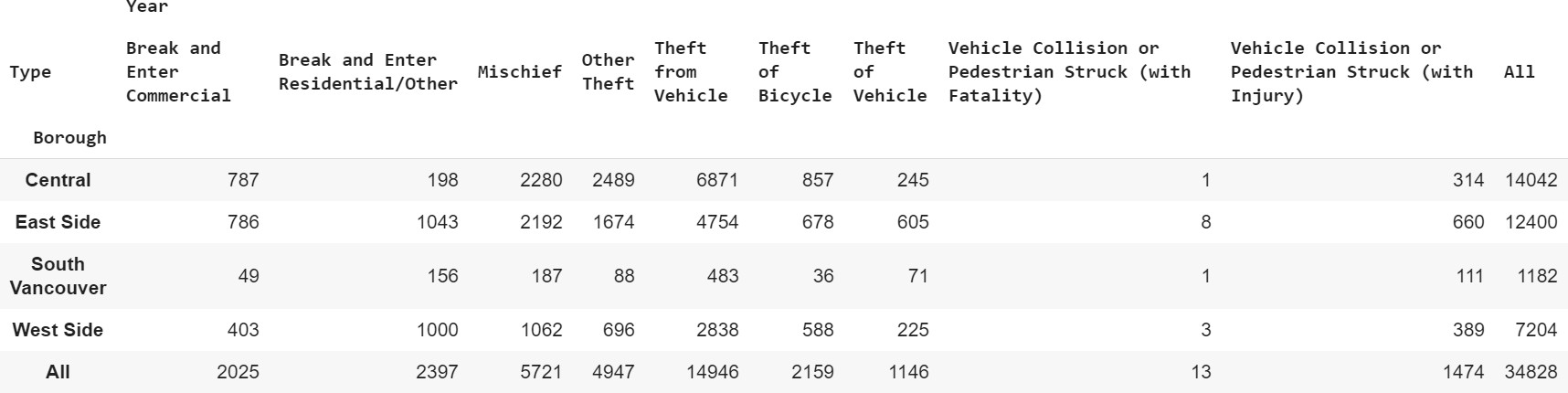


This data has been generated based on data from Wikipedia.



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.



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 neighborhoods on map for better visual depiction, to achieve this we create a data frame similar to the below one:



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

# Methodology

## Exploratory Data Analysis

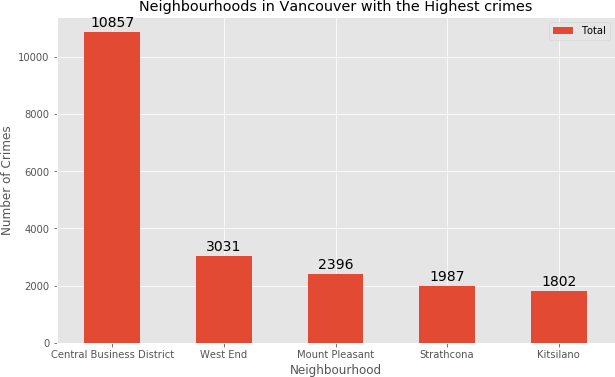
### Statistical summary of crimes

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



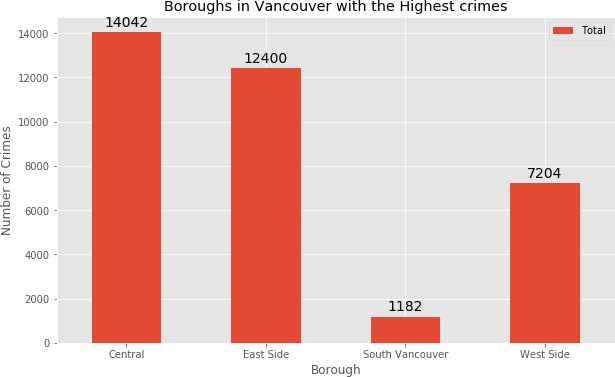
### Neighborhoods with the highest crime rates

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



### Boroughs crime Analysis

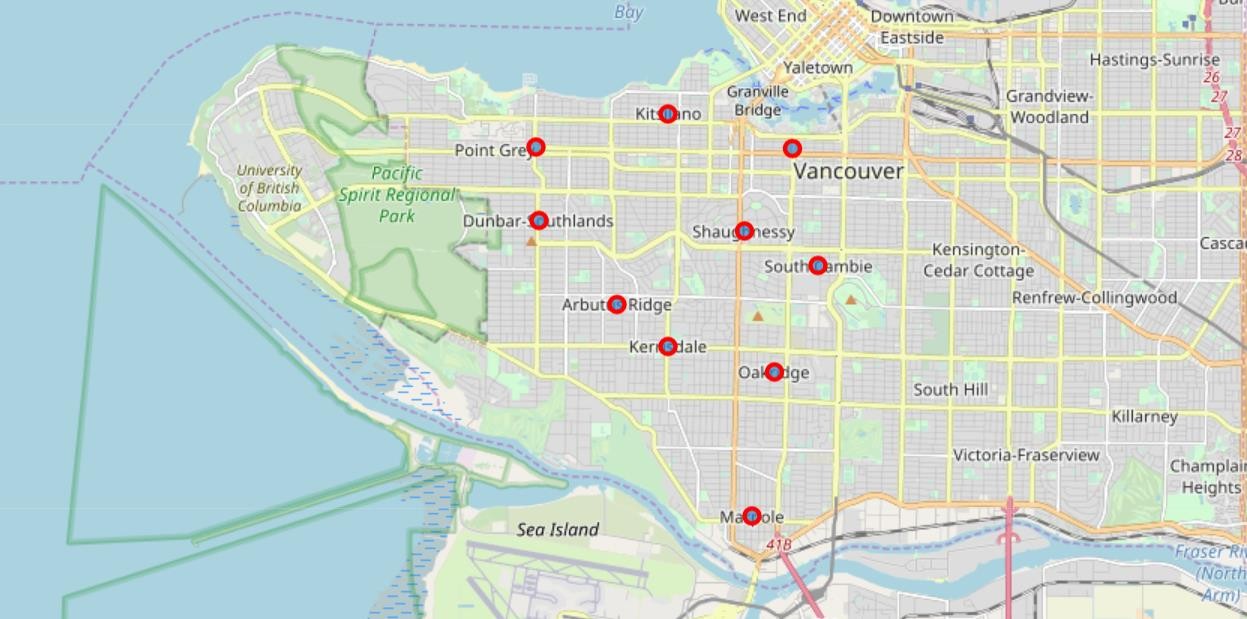
Comparing the crime report in the four boroughs of Vancouver during the year 2018, South Vancouver has the lowest crime rate probably because of its low neighborhood, followed by West Side which despite having up to 10 neighborhoods has less number of crimes compared with like of Central Vancouver.



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

### Neighborhoods in West Side, Vancouver, Canada

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



## Modelling

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

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 neighborhoods.

To help people find similar neighborhoods in the safest 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 a cluster size of 5 for this project that will cluster the 10 neighborhoods into 5 clusters.

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

# Results

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



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



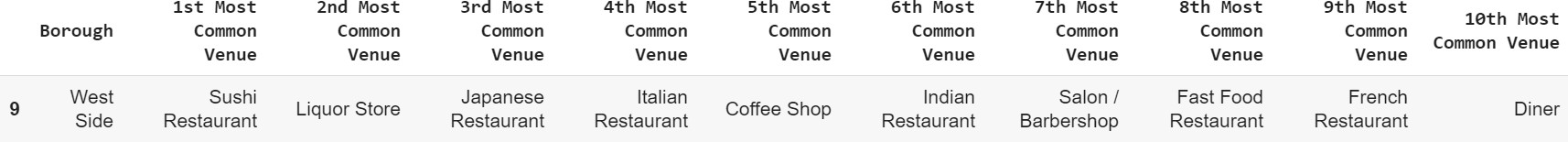
The second cluster has one neighborhood which consists of Venues mostly utility places like Spa, Yoga studio, pet studio, Grocery store and some Restaurants, Golf courses, and pubs.



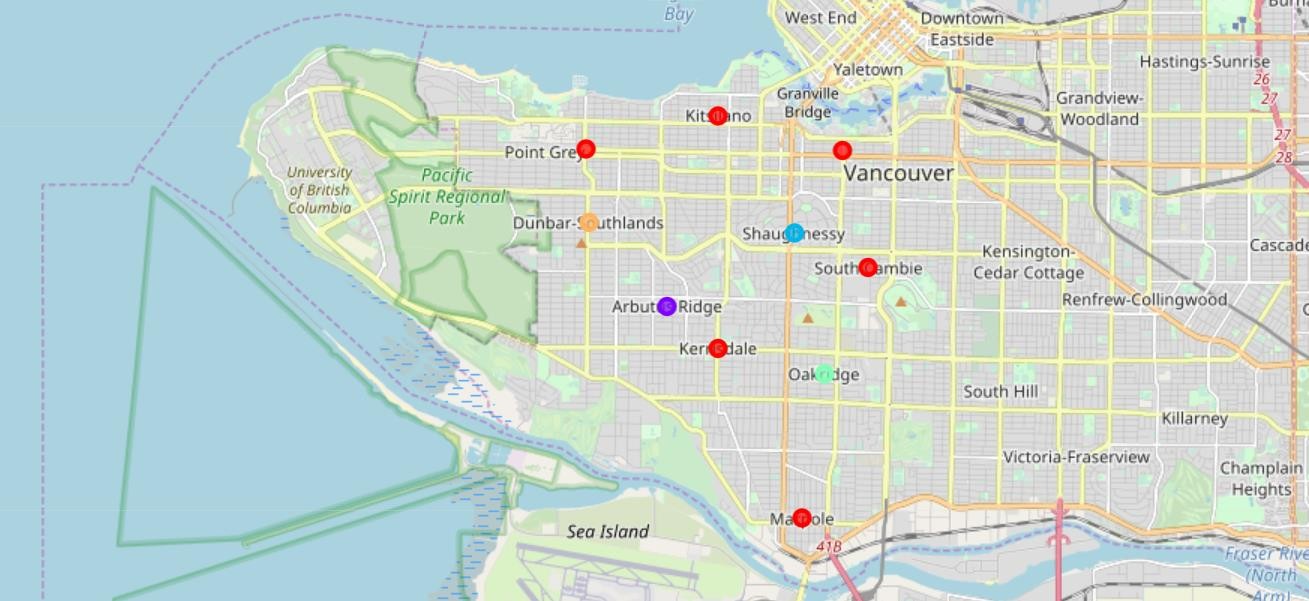
The third cluster has one neighborhood which consists of Venues such as bus stop, park and other utility place like Yoga Studio along with restaurants and food trucks.



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



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



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

Each cluster is color coded for the ease of presentation, we can see that majority of the neighborhood falls in the red cluster which is the first cluster. Remaining Neighborhood each is part of remaining four clusters and has been represented with different colors.

# Discussion

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 borough 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.

# 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.