**Finding the Optimum Location to Open a Grocery Store in Vancouver**

Tom Corbin

**Introduction**

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

This project aims to find a safe neighbourhood to open a grocery store in Vancouver. This 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.

Interest

Vancouver is an incredibly diverse city, and as such, has incredibly diverse neighbourhoods. Selecting the right neighbourhood from such a wide array of options will be challenging.

**Data Acquisition and Cleaning**

Data Acquisition

We will get the crime data from a real-world dataset published on Kaggle. This dataset includes the type of crime, the time of the crime, and the coordinates of the crime, but it does not properly categorise the boroughs. We will therefore get this from Wikipedia. More coordinates will come from the OpenCage Geocoder API, and Foursquare API will be used to fetch the venues in the neighbourhoods.

Dataset properties are:

* Type
* Year
* Month
* Hour
* Minute
* Hundred Block
* Neighbourhood
* X – GPS longitude
* Y – GPS latitude

The Wikipedia dataset does not require scraping as it was properly categorised. The OpenCage dataset contains the following columns:

* Neighbourhood
* Borough
* Latitude
* Longitude

Data Cleaning

The data from Kaggle was a heavy file which Git could not accommodate. Due to the size of the data (~ 600,000 rows), we cannot process it all. Instead, we will consider a recent 2018 crime report which will greatly reduce the number of rows in the dataset. Since the original data source cannot be uploaded to Git we will process the dataset in the runtime to filter the records of crimes that took place in the year 2018, create a new csv out of it using pandas and upload it to Github repository.

Table

Description automatically generated

Due to the improper encoding of the co-ordinates, the exact coordinates from the data could not be used for plotting because the co-ordinates seemed to be corrupted. Along with the X, Y columns in the dataset which represented the GPS co-ordinates of 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 neighbourhood page which on a later stage will be merged with the Crime data table.

Table

Description automatically generated

Table

Description automatically generated

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.

Table

Description automatically generated

Along with analysing the crime data we also have to fetch the latitude and longitude data, to plot the neighbourhoods on map for better visual depiction, to achieve this we create a data frame similar to the below one:

Table

Description automatically generated

**Methodology**

Exploratory Analysis

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.

Table

Description automatically generated

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

Chart, bar chart

Description automatically generated

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 neighbourhood, followed by West Side which despite having up to 10 neighbourhoods has a smaller number of crimes compared with like of Central Vancouver.

Chart, bar chart

Description automatically generated

Since South Vancouver has very little number of neighbourhoods 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.

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

Map

Description automatically generated

Modelling

Based on the final dataset of neighbourhoods and boroughs along with latitude and longitude of neighbourhoods in West Side Vancouver, we can find all the venues within a 500-meter radius of each neighbourhoods 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:

Table

Description automatically generated

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 neighbourhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighbourhoods. To help people find similar neighbourhoods in the safest borough we will be clustering similar neighbourhoods 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 neighbourhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighbourhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighbourhood.

**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. Looking into the neighbourhoods in the first cluster

Table

Description automatically generated

The cluster one is the biggest cluster with 6 of the 10 neighbourhoods in the borough West Side. Upon closely examining these neighbourhoods we can see that the most common venues in these neighbourhoods are Restaurants, eateries, parks and food trucks, Grocery store is not among the most common venues which makes this cluster of neighbourhoods an ideal destination to set up a grocery store. Looking into the neighbourhoods in the second, third, fourth and fifth clusters, we can see these clusters have only one neighbourhood in each. This is because of the unique venues in each of the neighbourhoods, hence they couldn't be clustered into similar neighbourhoods.

Graphical user interface, application

Description automatically generated

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

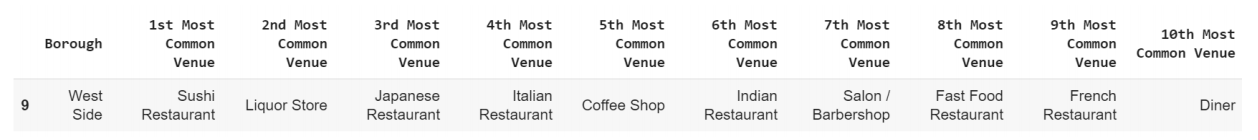


The third cluster has one neighbourhood which consists of Venues such as bus stop, park and other utility place like Yoga Studio along with restaurants and food trucks.

Graphical user interface, application

Description automatically generated

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



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

Map

Description automatically generated

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

Each cluster is colour coded for the ease of presentation; we can see that majority of the neighbourhood falls in the red cluster which is the first cluster. Remaining neighbourhood each is part of remaining four clusters and has been represented with different colours.

**Discussion**

The objective of the business problem was to help stakeholders identify one of the safest boroughs 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.