Choosing the Neighborhood for a Micro-Lending Company in Vancouver

Applied Data Science Capstone Project – IBM Data Science Professional Certificate

Author: Carlos Paiva González

August 5th, 2021

1. Business problem

1.1. Background

The rise of startups and micro/small companies in recent years has created a new group of customers that require financing for starting their new ventures, yet in several cases traditional financing is not a viable option for them due to the lack of good credit and a solid financial history. This scenario represents an opportunity for financial institutions that specialize in providing micro-loans to ventures that do not comply with the requirements of traditional banks.

Canada is a good example of this situation. Not only the country fosters the creation of startups by Canadians, but also encourages immigrants to set their new ventures within the country protected by specific migration policies. This makes Canada an attractive destiny for both startups and companies that provide support services to entrepreneurs.

1.2. Problem description

The main beneficiary of this report will be 'Argus' (fictitious entity), a small micro-lending Canadian company. Argus was formed 4 years ago, and provides financial assistance to entrepreneurs, immigrants, and veterans in the form of loans (up to US$ 50k) and consulting services (mentoring in business management and financial education with the objective of improving their financial credit history to access larger capital through traditional banking).

Argus started its business with a 100% online model, providing remote services to its customers and loans via wire-transfer to bank accounts already established by its customers. After its initial success, the company kept working the same operating model through the pandemic, although recently new customers have starting to request cash-only loans as well as consultancies via in-person meetings. Initially, Argus owners were reluctant to change their business model, but considering the increasing demand, they decided to set up their first physical office in Vancouver (where most of their customers are located).

Therefore, Argus is considering which neighborhood would be the best suited for their office. Besides price (which is beyond the scope of this project), they have two priorities: security (being

established in a safe neighborhood, as clients will go in and out the premises holding cash), and commercial (being closer to their main public segment – mostly young entrepreneurs with several different commercial businesses, from coffee places and restaurants to stores where they sell specific products such as sporting goods, hardware and tools, and clothes amongst others).

1.3. Interest

In real life, this project might interest any small/medium company that would like to select the right location for setting up their office in Vancouver, having commercial and security matters as their top priorities.

2. Data acquisition and cleaning

2.1. Data sources

All data used for the project comes from 3 main data sources.

The crime data comes from a Kaggle dataset (link here). This data comes in the form of a csv file containing 530,652 observations describing crimes in Vancouver from 2003 till 2017 (each observation corresponds to 1 crime). The attributes for each crime are: Type of crime, Year, Month, Day, Hour, Minute, Block, Neighborhood, Latitude and Longitude. Since this is a large dataset, it was cleaned and all the non-relevant data for the analysis was removed as described in the next section.

The coordinates of each neighborhood were initially obtained using the Geolocator package for Python3 using the name of each neighborhood taken from the crime dataset. However, the use of this data was disregarded in the end, as explained in the next section.

All data related to the commercial venues within each neighborhood was obtained via a Foursquare API call with defined parameters (radius=1,000m; limit of answers per neighborhood=100). The answer was a dataset of 1,330 observations (1 venue per observation). The attributes for each venue are: Neighborhood, Venue Latitude, Venue Longitude, Venue Name, and Venue Category.

2.2. Data cleaning and feature selection

Crime Data

The original dataset contained 530,652 observations and 12 features, yet it was reduced in size so it would not occupy too much space in the Github repository (where it will be called from with the Python3 code in the Jupyter Notebook). It was renamed as Vancouver_Crime_Filtered.csv, containing only data for years 2014, 2015 and 2016 (last 3 full years in the dataset) and the relevant

features: Type of Crime, Year, Month, Neighborhood, Latitude, and Longitude. Please refer to Table 1 for more detail on the feature selection for this first dataset.

Table 1: Feature selection for Crime Data

Feature	Decision	Reason for keep/discard
Type of Crime	Keep	Core data, not all types of crimes are relevant for commercial venues
Year	Keep	Will be used for assessing how level crimes move across time to simplify data
Month	Keep	Will be used for assessing how level crimes move across time to simplify data
Day	Discard	Too detailed for analysis, year and month is enough
Minute	Discard	Too detailed for analysis, year and month is enough
Block	Discard	Too detailed for analysis, neighborhood is enough
Neighborhood	Keep	Core data, will be used for classifying crimes per neighborhood
Latitude	Keep	Core data, will be used to know locations per crime
Longitude	Keep	Core data, will be used to know locations per crime

Coordinates Data

The first approach for obtaining the coordinates data for each neighborhood in Vancouver was to use the Geolocator package, which provides the latitude and longitude data for each neighborhood from its complete name, which must be in the following format: 'Name of Neighborhood, Vancouver, British Columbia'. So, a list of all unique neighborhood names from the Crime Dataset was obtained for this (24 neighborhoods in total). Table 2 presents an example of the information obtained from using the Geolocator package on each neighborhood from the list:

Table 2: Example of data obtained from the use of Geolocator package

Neighborhood	Latitude	Longitude
Arbutus Ridge	49.246305	-123.159636
Central Business District	49.336120	-123.078021
...	...	...

As mentioned before, now the coordinates are plotted to check if the neighborhoods have been located correctly (Figure 1).

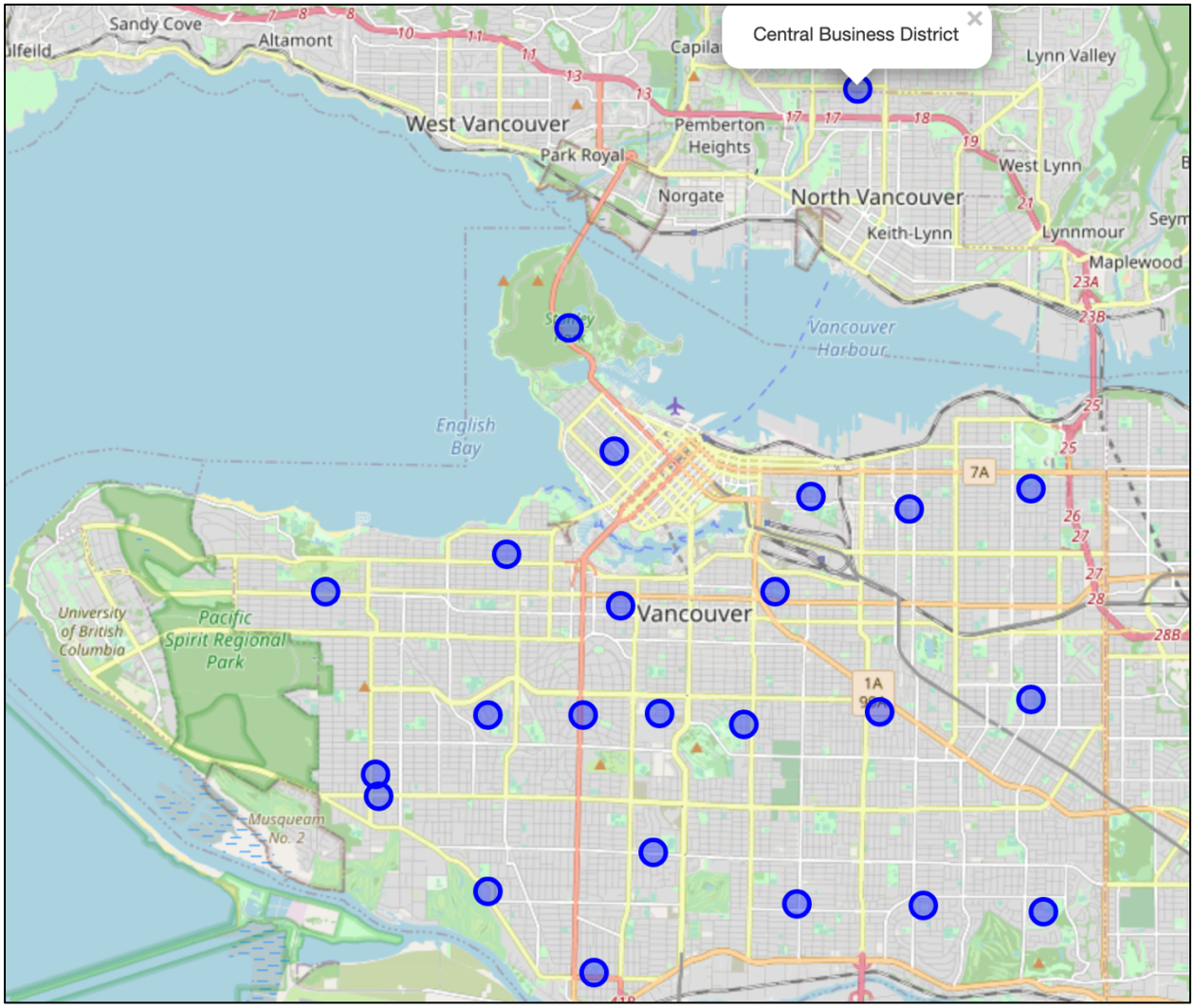

Figure 1: Map of Vancouver showing all 24 neighborhoods as per Geolocator Package

Comparing with a Vancouver city map from Google Maps (manual inspection), it can quickly be seen that for some neighborhoods the coordinates found via the Geolocator Package have not been accurate. Examples are Musqueam and Central Business District (which does not even show in Vancouver, but far north, which should not be the case).

Instead, another approach is to use the average of all the latitude and longitude data for all the crimes in the dataset grouped by neighborhood and plot this information in the map to see if a better geographical representation of each neighborhood can be found with this method. A table like Table 2 was obtained, and this new information can be used to plot a new Vancouver map showing the new neighborhood locations (Figure 2).

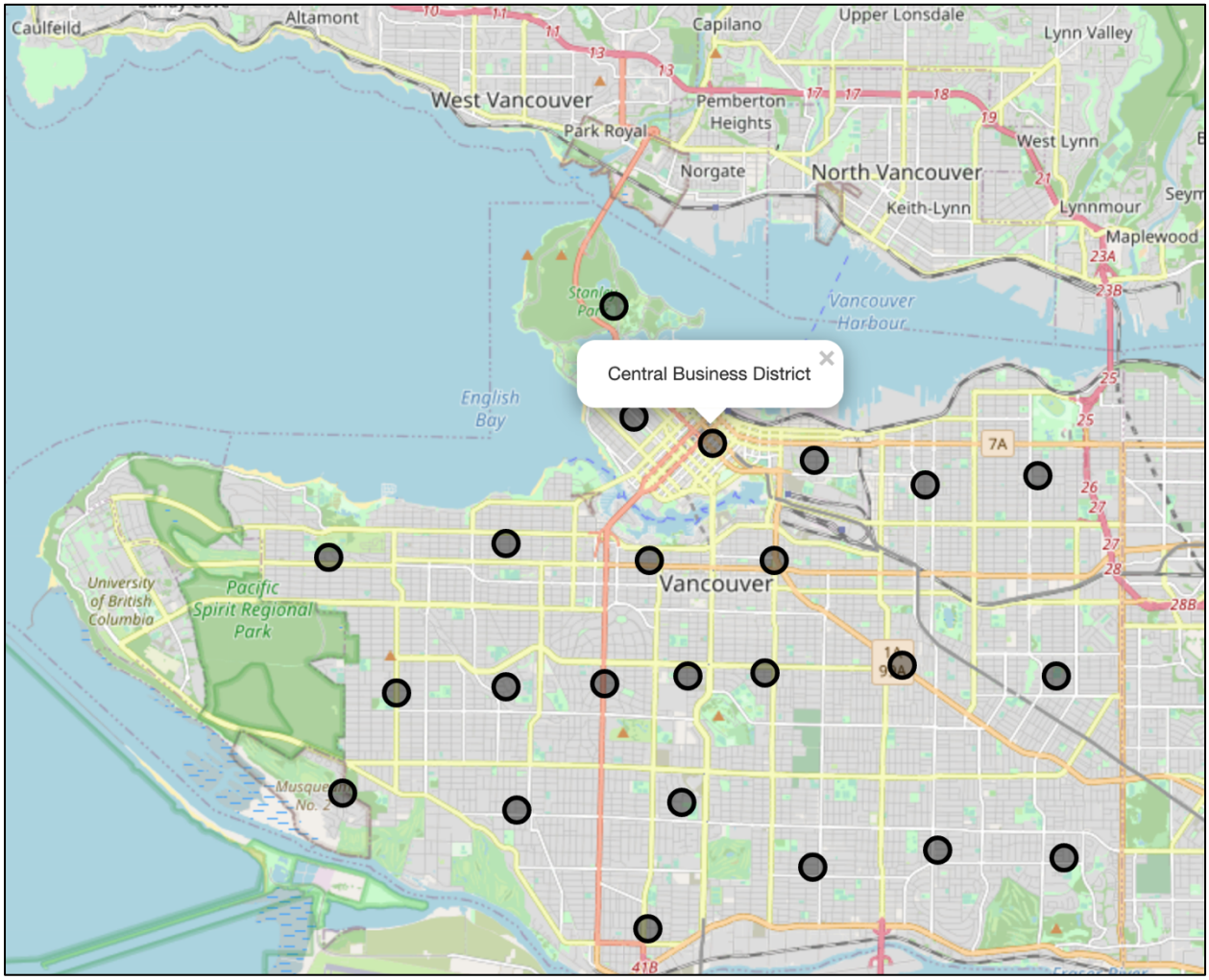

Figure 2: Map of Vancouver showing all 24 neighborhoods as per the Crime Dataset average crime locations

This time, all neighborhoods are correctly mapped in the visualization, as it can be manually checked for each one (refer for instance to Musqueam o Central Business District compared to the previous visualization). Therefore, the locations provided by this second approach will be the official coordinates reference for the rest of the project.

Commercial Venues Data

As mentioned before, a Foursquare API call was used to obtain all the information for each commercial venue in each neighborhood. Table 3 shows an example of the provided data, including venue names, categories, and locations for each venue within each neighborhood (1 observation or row represents 1 venue). The total dataset has 1,330 observations (one per venue) and 5 features as shown below.

Table 3: Foursquare API Vancouver Venue Data

Neighborhood	Venue Name	Venue Latitude	Venue Longitude	Category
Arbutus Ridge	The Patty Shop	49.250680	-123.167916	Caribbean restaurant
Arbutus Ridge	Butter Baked Goods	49.242209	-123.170381	Bakery
Arbutus Ridge	The Arbutus Club	49.248507	-123.152152	Event Space
...	...	...	...	...

3. Project Methodology

The goal of the project is to find the neighborhood (or neighborhoods) in Vancouver that comply with the two main requirements from Argus: safety and having as many businesses founded by young entrepreneurs around (or that sell different goods consumed by this audience). For this, the project will come up with 2 ways of classifying the neighborhoods, one per security level, and another one per type of venues that each neighborhood holds for implying the kind of audience that visits each neighborhood the most. When combined, the two classifications should help us to recommend the best neighborhoods for hosting the new office.

Regarding crime, first Exploratory Data Analysis will be conducted to find the count of crimes per type and neighborhood. Since not all types of crime are relevant to commercial venues, only 3 types of crimes will be selected by using regression and correlation analysis. Once selected, with the help of histograms, the project can define levels of threat per crime type for all neighborhoods depending on the bucket that each neighborhood falls into as per the histograms. Finally, with the help of a defined rule, the overall level of threat per neighborhood is defined combining all types of relevant crimes. A map of Vancouver showing the neighborhoods' locations using circles with borders in a color scale (green for safe, red for dangerous and orange for safe with caution) will be generated to visually summarize the results.

Regarding the commercial venues' analysis, the project will classify the neighborhoods according to the types of venues around. The data will be re-arranged to count the number of venues per type that have been identified in each neighborhood, using each column for a different type of venue (~220 columns for each type of venue). This is done to use unsupervised machine learning to create clusters that group the neighborhoods by the number of similar venues that contain. For this, K-Means Clustering has been selected as the Machine learning tool. K-Means requires to find the ideal number of clusters prior to select the best classification, which will be done using the 'elbow' method. Finally, each cluster will be given a name according to the types of venues they have as majority, and this will be reflected in a final map plot of Vancouver that shows the neighborhoods classified by crime threat level (color of circle borders) and venue type cluster (color of circle filling). With all this information, the project will provide its final recommendation to Argus.

4. Exploratory Data Analysis (EDA)

4.1. EDA and arrangement of Crime Data

A data frame classifying crime count per neighborhood and years (2014, 2015 and 2016) as indexes, and type of crime as columns will be created to facilitate the crime data EDA. An example of the first rows of this data frame is provided in the table below (Table 4).

Table 4: Pivot table of count of crimes per type and neighborhood for 2014, 2015 and 2016 (not showing all columns)

Neighborhood	Year	Break/Enter Commercial	Break/Enter Residential	Mischief	Theft from Vehicle	Theft of Vehicle	…
Arbutus Ridge	2014	28	129	35	140	14	…
Arbutus Ridge	2015	13	103	40	128	15	…
Arbutus Ridge	2016	32	84	48	155	10	…
Central B.D.	2014	570	188	1,285	2,624	94	…
Central B.D.	2015	802	138	1,225	2,969	132	…
…	…	…	…	…	…	…	…

Considering all types of crimes in the pivot table (9 in total), it would be interesting to see if there is a relationship between the different types of crime occurrence. With the help of a heatmap, it is possible to graphically inspect the correlation between each type of crime, as below (Figure 3). The types of crime that are highly correlated amongst them are: Break and Enter Commercial, Other Theft, Theft from Vehicle, and Theft from Bicycle. On the other hand, Break and Enter Residential shows the lowest correlation with the other types of crime.

Since the purpose of the project is to find the safest location for a commercial location (Argus' office), not all types of crimes are relevant. Some crimes like 'Fatal Car/Pedestrian Accident' and 'Minor Car/Pedestrian Accident' can be disregarded from the analysis.

Clearly the most representative type of crime for the project purpose is 'Break and Enter Commercial'. Therefore, we will explore the relationship between this type of crime and the others, with the objective of selecting 2 features: one with a very high correlation with 'Break and Enter Commercial', and another with a smaller correlation with 'Break and Enter Commercial', since we would like to capture in our analysis another type of crime that is still relevant to the project. A regression plot between ‘'Break and Enter Commercial’ and the other 6 remaining types of crime is shown in Figure 4.

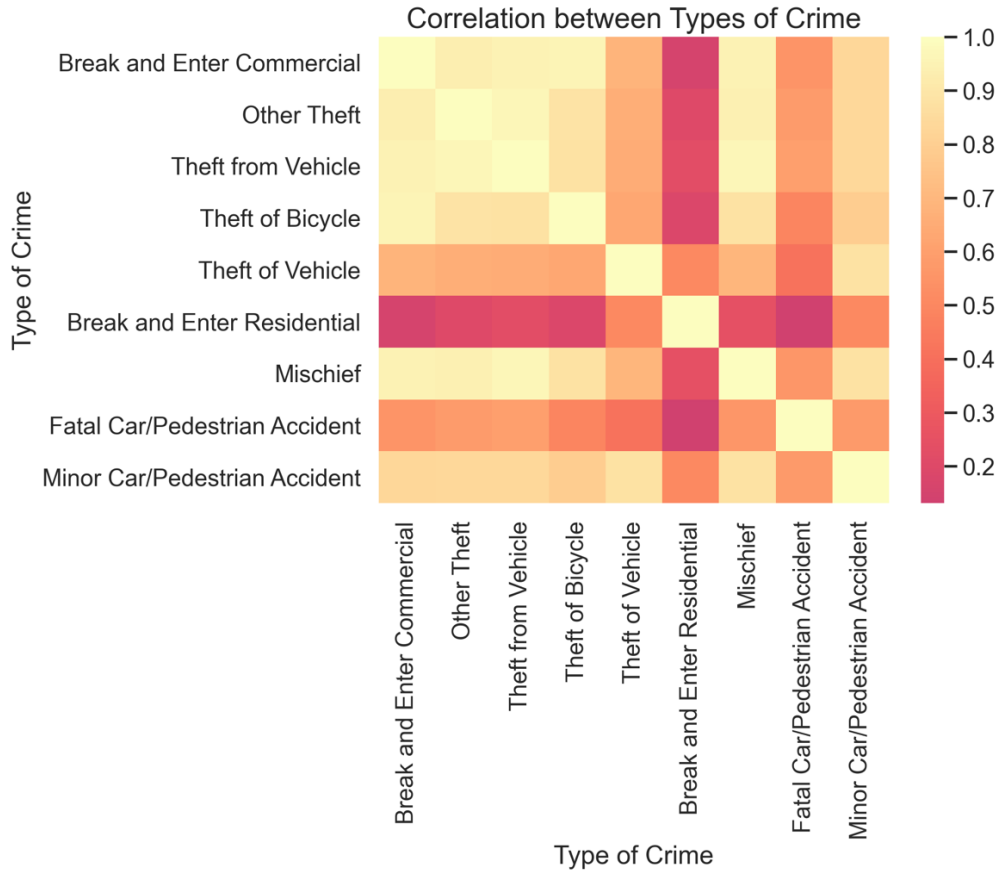

Figure 3: Correlation between the different types of crime

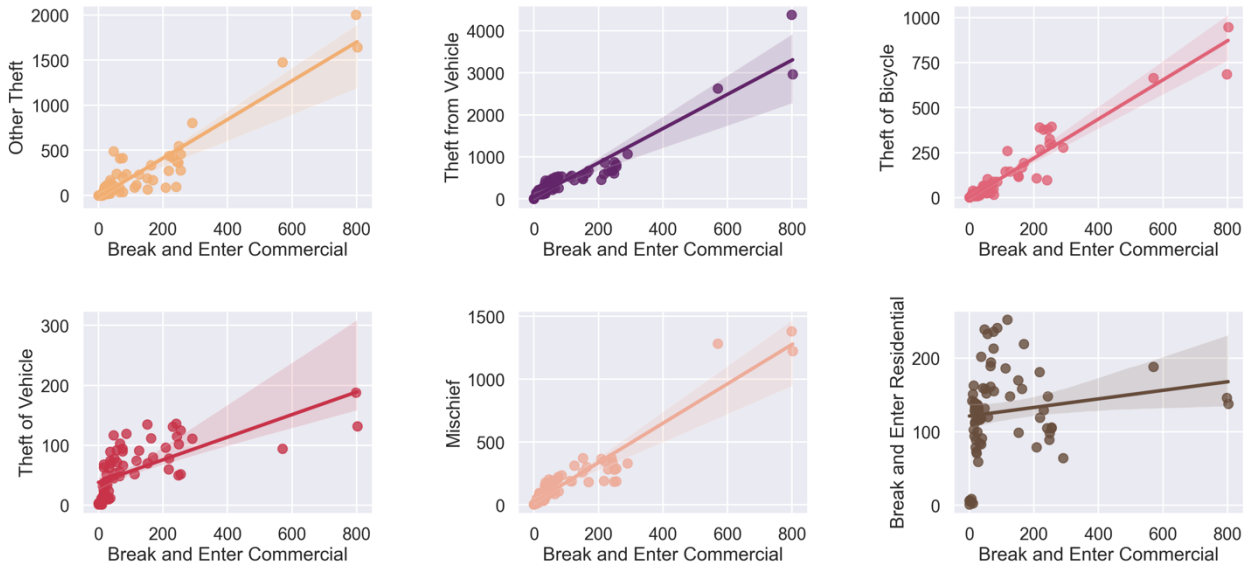

Figure 4: Regression Plot (Scatter + Linear Trend) between 'Break and Enter Commercial' and other types of crime

The 3 selected types of crime will be 'Break and Enter Commercial', as explained above; 'Theft from Vehicle', which shows the highest correlation with the first type of crime both in the heatmap and in the regression plot; and 'Theft of Vehicle', which is type of crime that does not show a very high correlation with 'Break and Enter Commercial' (around 0.6) but is a relevant type of crime for the project (the amount of vehicle robberies speaks about how dangerous a neighborhood can be for commercial purposes).

Regarding 'Theft of Bicycle', although highly correlated with 'Break and Enter Commercial', it will be disregarded since it is not very relevant for the project (the fact that many bicycles are stolen in a particular neighborhood does not mean that the neighborhood itself will be dangerous for Argus' business). Same goes for 'Other Theft'. Regarding 'Break and Enter Residential', it shows a very low correlation with 'Break and Enter Commercial', which makes sense since most likely there are not many houses to break into in commercial neighborhoods and vice-versa, and therefore it is difficult to assess if a neighborhood with crimes of this type might represent any danger to Argus or its customers.

Therefore, only 'Break and Enter Commercial', 'Theft from Vehicle', and 'Theft of Vehicle' will be included in the rest of the analysis. The next step will be to create a master data frame with 1 observation per neighborhood showing the total amount of crimes per type (average of all 3 years). An example of this initial data frame is shown in Table 5 below.

Table 5: First 3 rows of the master pivot table showing the

Neighborhood	Break/Enter Commercial	Theft from Vehicle	Theft of Vehicle
Arbutus Ridge	24	141	13
Central Business District	723	3325	138
Dunbar-Southlands	14	192	22
Fairview	240	656	54
Grandview-Woodland	146	527	113
...	...	...	...

To assess how much 141 thefts from vehicles (for example) means, it would be useful to have a measurement of the distribution of each type of crime to create buckets to classify each quantity relatively to the universe of all neighborhoods in Vancouver. Thus, histograms will help with the visualization of these buckets (Figure 5).

A column will be added in the master data frame for categorizing each neighborhood as 'Low Level Threat', 'Mid-Low Level Threat', 'Mid-High Level Threat', 'High Level Threat' according to the 4 bins in each histogram (Table 6).

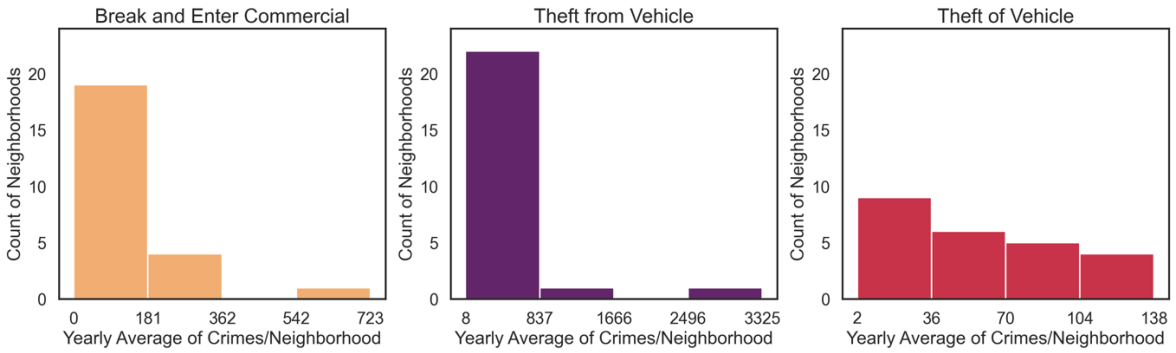

Figure 5: Distribution of yearly average crimes/neighborhood for the 3 selected types of crime

Table 6: List of Vancouver neighborhoods classified by crime threat

Neighborhood	Break/Enter Commercial Threat	Theft from Vehicle Threat	Theft of Vehicle Threat	Crime Threat Level
Arbutus Ridge	Low	Low	Low	Green
Central B.D.	High	High	High	Red
Dunbar-Southlands	Low	Low	Low	Green
Fairview	Mid-Low	Low	Mid-Low	Orange
Grandview-Woodland	Low	Low	High	Red
Hastings-Sunrise	Low	Low	Mid-High	Orange
Kensington-Cedar Cottage	Low	Low	Hid-High	Orange
Kerrisdale	Low	Low	Low	Green
Killarney	Low	Low	Mid-Low	Orange
Kitsilano	Low	Low	Mid-Low	Orange
Marpole	Low	Low	Mid-Low	Orange
Mount Pleasant	Mid-Low	Low	High	Red
Musqueam	Low	Low	Low	Green
Oakridge	Low	Low	Low	Green
Renfrew-Collingwood	Low	Low	High	Red
Riley Park	Low	Low	Mid-Low	Orange
Shaughnessy	Low	Low	Low	Green
South Cambie	Low	Low	Low	Green
Stanley Park	Low	Low	Low	Green
Strathcona	Mid-Low	Low	Mid-High	Orange
Sunset	Low	Low	Mid-High	Orange
Victoria-Fraserview	Low	Low	Mid-Low	Orange
West End	Mid-Low	Mid-Low	Mid-High	Orange
West Point Grey	Low	Low	Low	Green

Now, it is helpful to visualize the map of Vancouver presenting the different neighborhoods with tags detailing their names, number of crimes of each type; and assigning each marker a different color depending on whether a neighborhood is considered safe or not with the following criteria:

- Green or safe (if all 3 types of crime have the 'Low' category in the neighborhood)
- Red or dangerous (if at least 1 type of crime has the 'High' category)
- Orange or relatively safe with caution (everything in between)

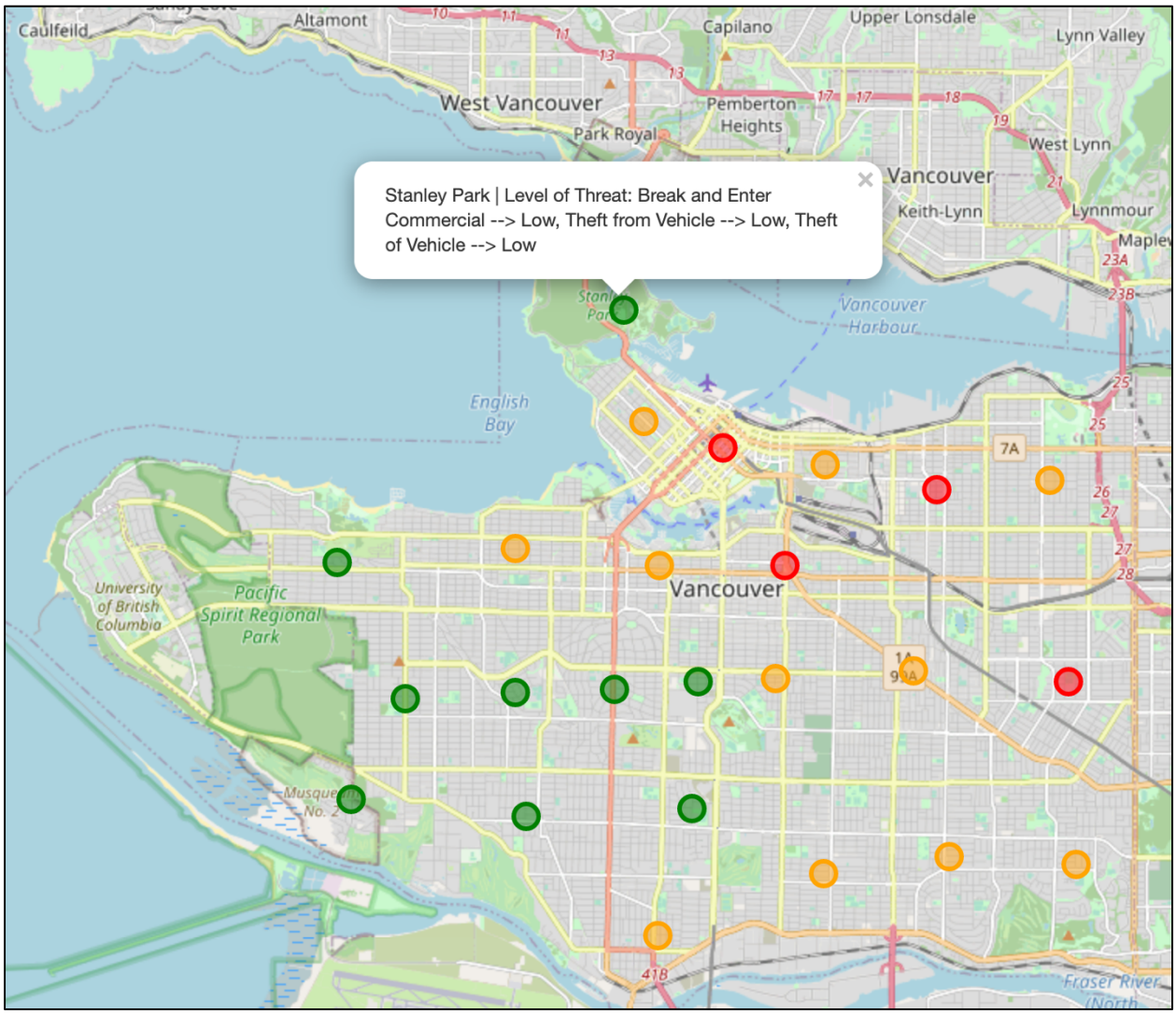

Figure 6: Map of Vancouver showing all 24 neighborhoods classified per crime threat

4.2. EDA and arrangement of Venues Data

With the data from commercial venue (refer to the previous Table 3), it would be interesting to visually explore the number of venues found per neighborhood (Figure 7).

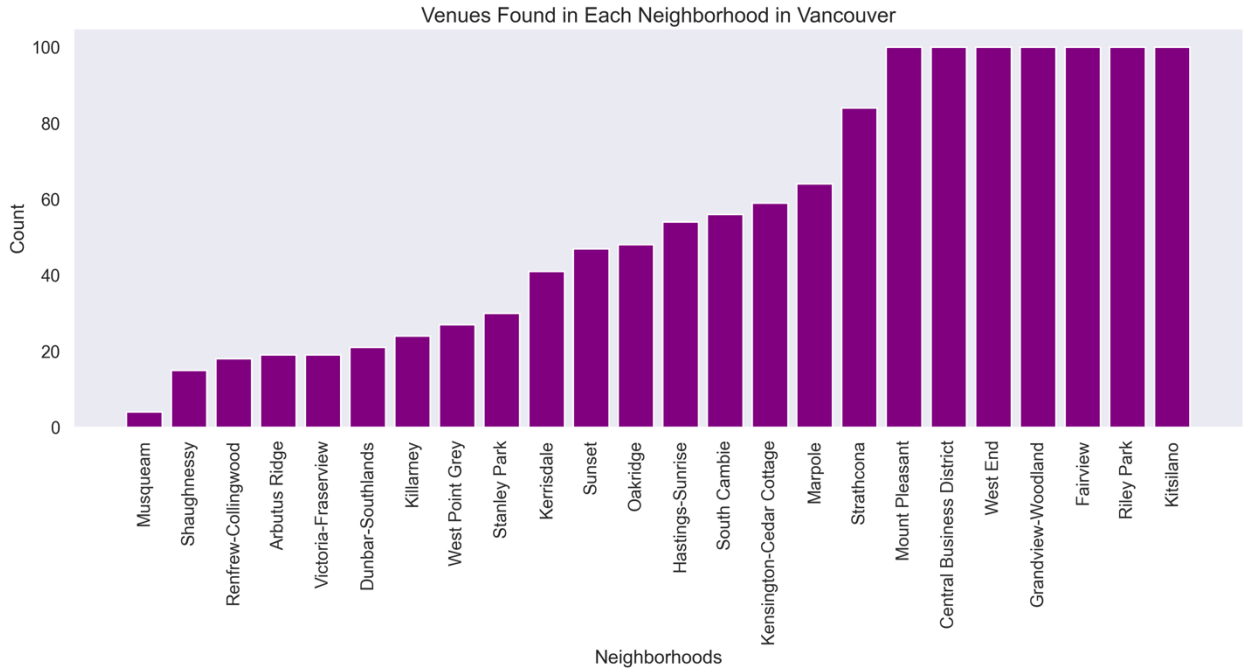

Figure 7: Count of number of venues found per neighborhood (limit 100 venues)

7 out of 24 neighborhoods already have at least 100 venues, while some others like Musqueam and Shaughnessy have less than 20 in the whole neighborhood. Later, the project will define a minimum of venues per neighborhood to consider for giving a final recommendation of which neighborhood would increase Argus' changes of increasing sales.

With this information, and to follow the path established in the methodology section (K-Means clustering for classifying the neighborhoods according to their commercial venue type), the next step is to perform one-hot encoding to create new columns with a binary choice (0 or 1) for each type of venue for each neighborhood (one column per type of venue as per Table 3). Then, the one-hot data frame will be grouped per neighborhood (using the mean of each type of venue) to find the frequency of each type of venue per neighborhood with a mean operation (Table 7 is an example of the first few rows of this data frame).

Table 7: Example of the first rows of the data used for the machine learning model

Neighborhood	American Restaurant	Amphitheater	Aquarium	Art Gallery	…
Arbutus Ridge	0.00	0.00	0.00	0.00	…
Central B.D.	0.01	0.00	0.00	0.01	…
Dunbar-Southlands	0.00	0.00	0.00	0.00	…
…	…	…	…	…	…

Finally, the venues can be arranged by type in columns to show the most frequent ones in a ranking for each neighborhood. This is another useful way for visualizing the data and will be particularly useful later for giving a name to the found clusters with the K-Means method by assessing which are the top venues in each neighborhood within each cluster (Table 8 is an example of the first few rows of this data frame).

Table 8: Top 5 most common venue types identified per neighborhood (example)

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Arbutus Ridge	Bakery	Burger Joint	Coffee Shop	Sushi Restaurant	Fast Food Restaurant
Central B.D.	Hotel	Coffee Shop	Restaurant	Desert Shop	Taco Place
Dunbar-Southlands	Sushi Restaurant	Pharmacy	Park	Bank	Bakery
…	…	…	…	…	…

5. Machine learning model

(To be developed on Week 5).

6. Conclusions

(To be developed on Week 5).

7. Next steps

(To be developed on Week 5).