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 9th, 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 up 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 with 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=1,000 m; 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

2.2.1. 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. In Table 2, an example of the first rows of the Crime Data can be seen after removing the unneeded features.

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

Table 2: Example of the first rows of the Crime Dataset

	Type	Year	Month	Neighborhood	Latitude	Longitude
0	Other Theft	2014	12	Hastings-Sunrise	49.269348	-123.046810
1	Mischief	2014	7	Victoria-Fraserview	49.232509	-123.077009
2	Other Theft	2014	11	Hastings-Sunrise	49.269348	-123.046810
...	...	...	...	...	...	...

2.2.2. 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 3 presents an example of the information obtained from using the Geolocator package on each neighborhood from the list:

Table 3: 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
...	...	...

Now the coordinates are plotted to check if the neighborhoods have been located correctly (Figure 1). The blue circles represent the exact coordinates of each of the 24 Vancouver neighborhoods.

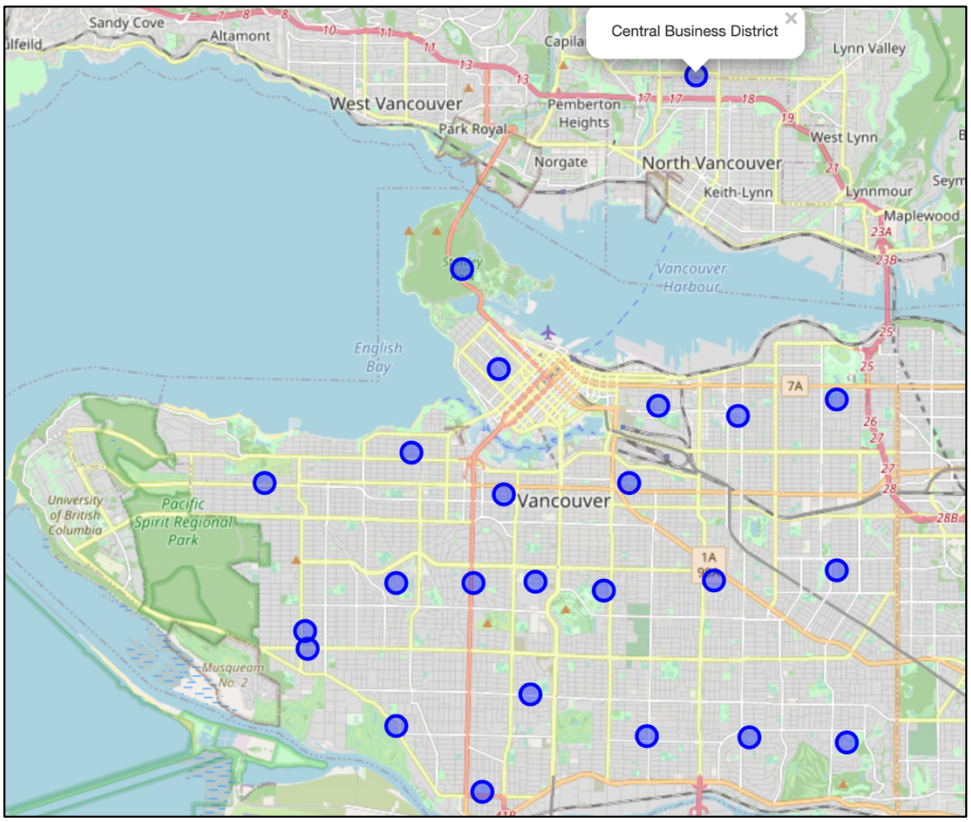

Figure 1: Map of Vancouver showing all 24 neighborhoods as per Geolocator Package

Comparing with a Vancouver city map from Google Maps (process done via manual inspection), the coordinates found for some neighborhoods with 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).

Therefore, another approach must be used. The project will use the latitude and longitude data for all the crimes in the Crime dataset grouped by neighborhood to plot this information in the map to see if a better geographical representation of each neighborhood can be found with this method. Figure 2 shows the plotting of the location of each neighborhood using this new method to visually check if the results were better than the ones obtained with the Geolocator Package.

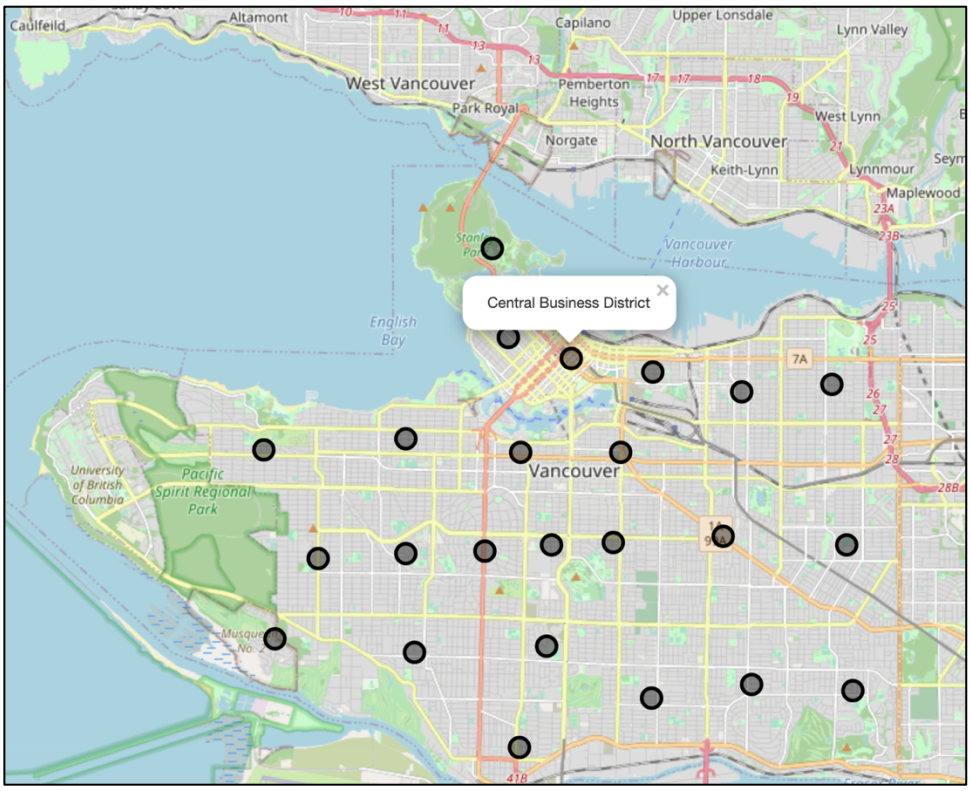

Figure 2: Map of Vancouver showing all 24 neighborhoods as per the Crime Dataset average crime locations

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.

2.2.3. Commercial Venues Data

As mentioned before, a Foursquare API call was used to obtain all the information for each commercial venue in each neighborhood. The outcome of this call was a file containing raw data in JSON format, which had to be arranged into the form of a Pandas data frame using a custom made Python function. Table 4 shows an example of this data frame (1 observation or row represents 1 venue). The total dataset has 1,330 observations (one per venue) and 5 features.

Table 4: 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 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 by combining all the assessment of intensity for 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, one 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 will be created to facilitate the crime data EDA, by classifying crime count per neighborhood and years (2014, 2015 and 2016) as indexes, and type of crime as columns. An example of the first rows of this data frame is provided in the table below (Table 5).

Table 5: 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	...
...	...	...	...	...	...	...	...

EDA will be done on the Crime Data to see if crimes have been occurring consistently across the 3 years per neighborhood, type of crime and month. One of the purposes of this exercise is to simplify the data as possible. Some visualizations might be helpful for this (Figures 3, 4 and 5).

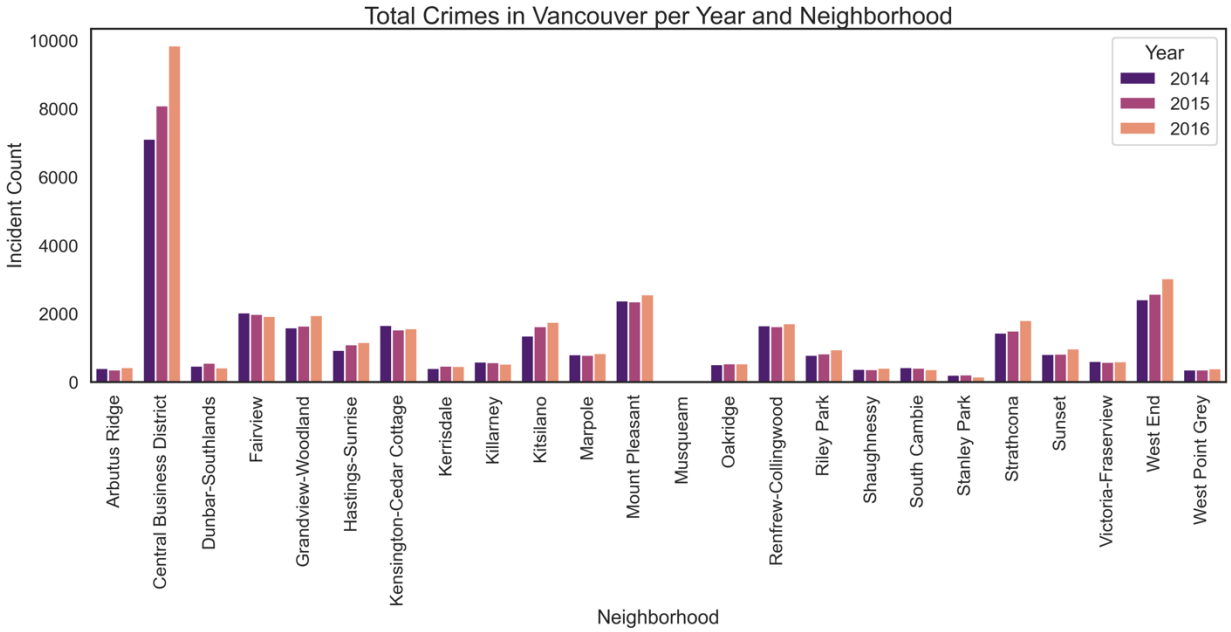

Figure 3: Evolution of crime count in Vancouver per Neighborhood across the last 3 years available in data

The total number of crimes per neighborhood appears to be quite consistent across the 3 years (only Central B.D. shows an increase in crimes every year). Now, looking at crimes per type:

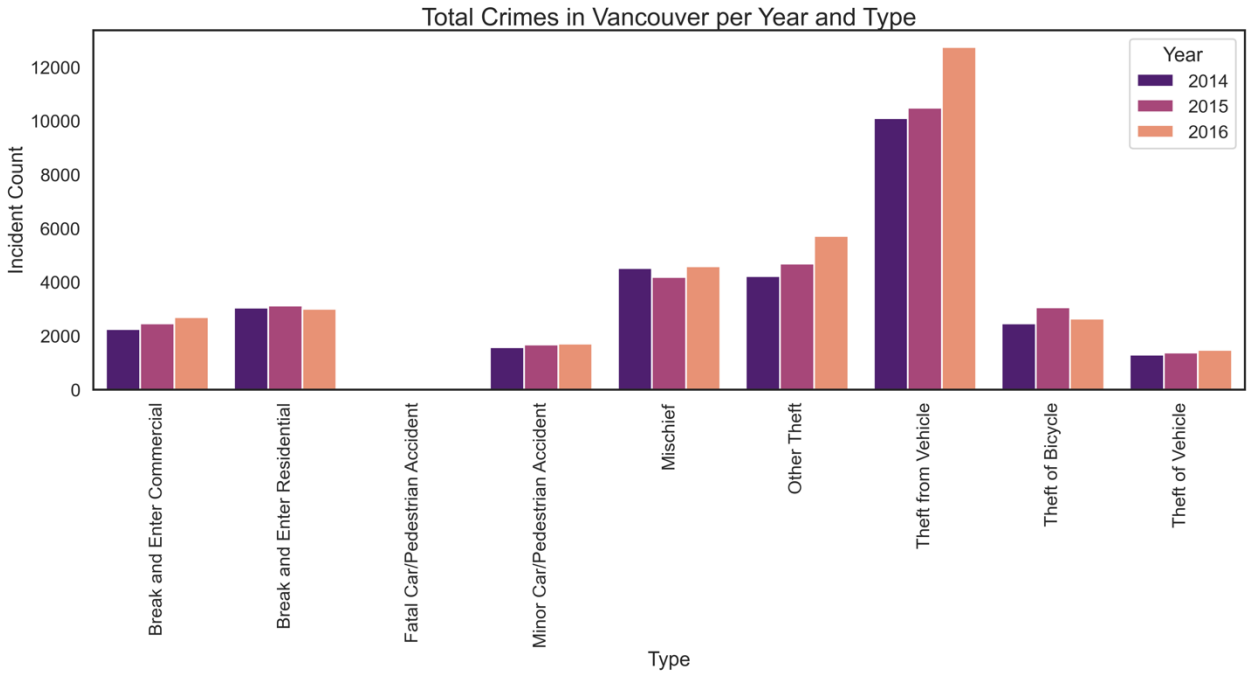

Figure 4: Evolution of crime count in Vancouver per Type across the last 3 years available in data

The number of crimes per type appears to be quite consistent, with a slight increase per year for all types except for 'Theft from Vehicle', which shows a bigger increase in 2016.

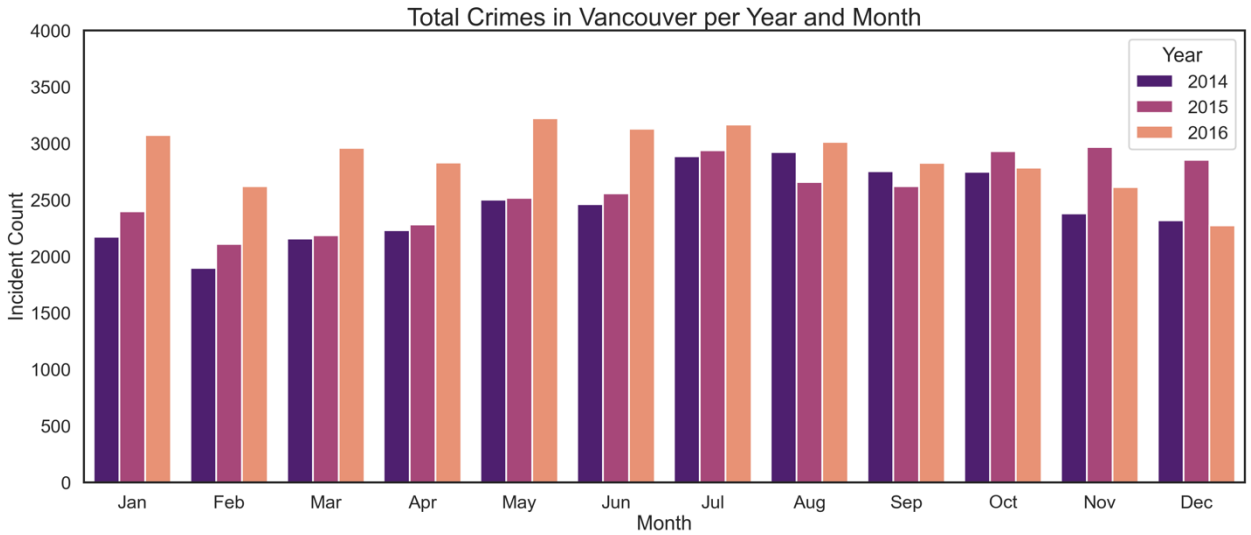

Figure 5: Evolution of crime count in Vancouver per Month across the last 3 years available in data

In this case, the number of crimes per month increases noticeably from January till July during 2016 compared to the previous 2 years, then it remains consistent during the next 3 months for all 3 years. For November and December, there is a crime peak during 2015.

From assessing all 3 breakdowns (Figures 3, 4, and 5), it can be concluded that overall criminal incident levels remained consistent across the 3 years, showing a small increase of 5% from 2014 to 2015, and of 11% from 2015 to 2016. Since this project focuses on classifying neighborhoods amongst them by the number of incidents to find the safest ones, and it has been found that across years the number of incidents per neighborhood remains stable, it is accurately possible to simplify the database to use by averaging the number of crimes per type and neighborhood. But before doing this, it would be useful to use all data points (no averages) to explore the relationships between each type of crime.

Considering all types of crimes in the pivot table (an example is shown in Table 5, there are 9 types of crime 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 6).

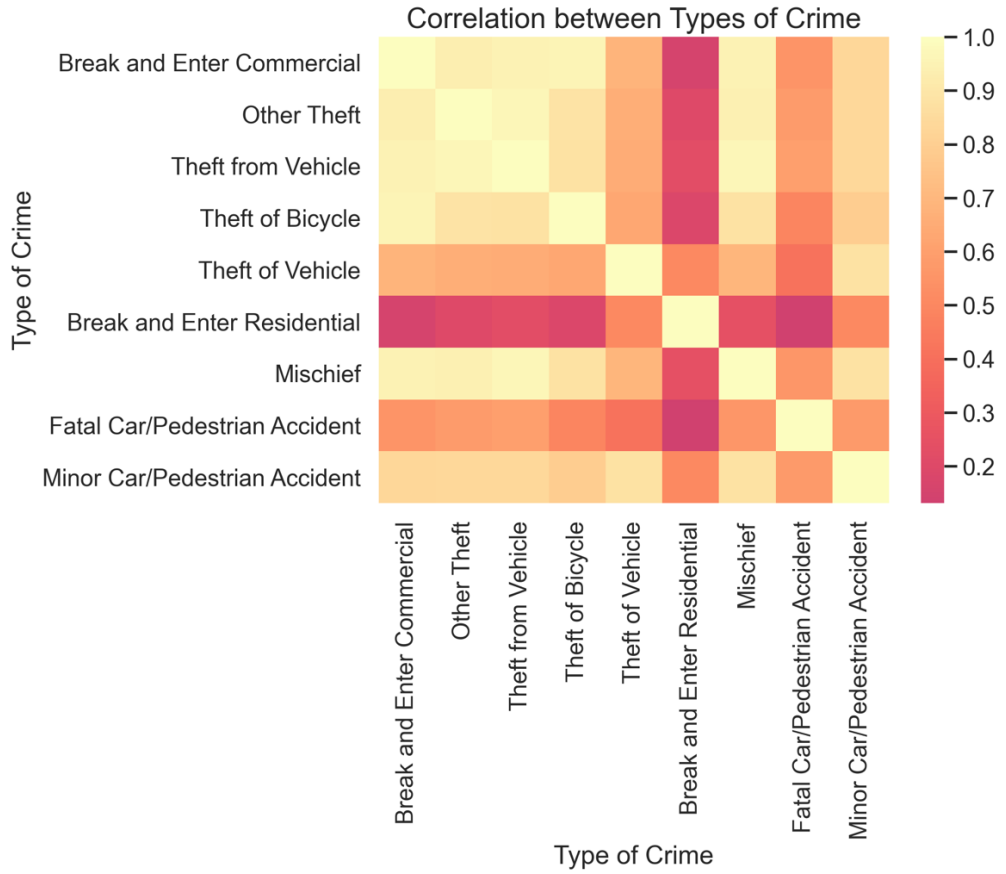

Figure 6: Correlation between the different types of crime

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, the project 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 it would be preferrable to capture in the 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 7.

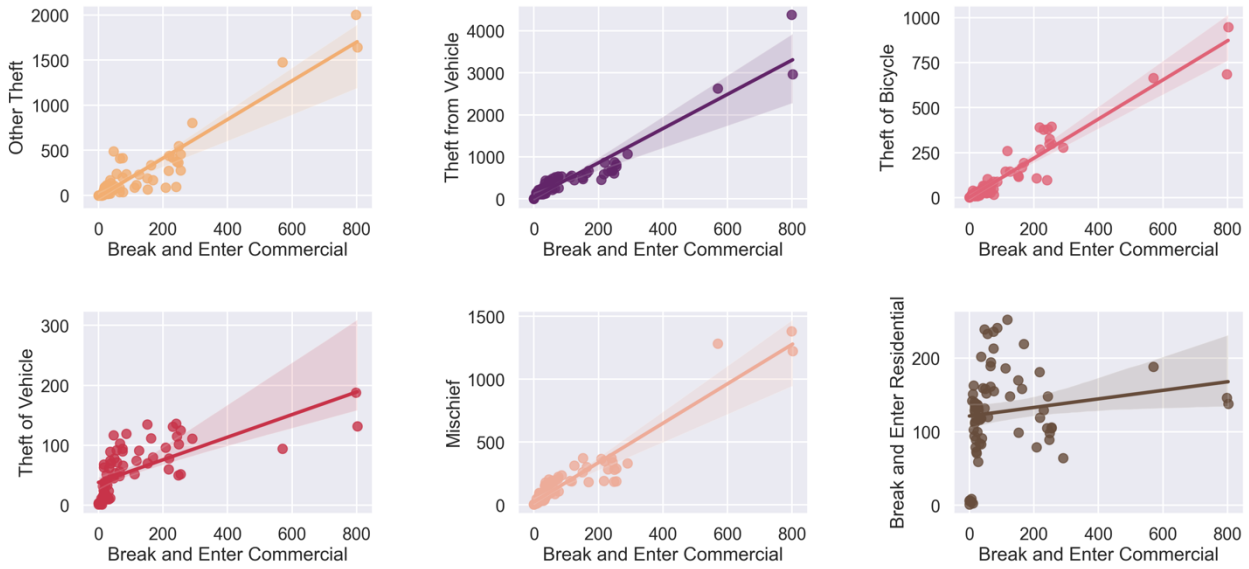

Figure 7: Regression Plot (Scatter + Linear Trend) between 'Break and Enter Commercial' and other types of crime

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 data frame is shown in Table 6 below.

Table 6: First 4 rows of the master pivot table showing the count of the selected crime types per neighborhood

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
...	...	...	...

For assessing the number of each type of crime for each neighborhood (e.g.: to make sense of how much 141 'thefts from vehicles' means, from Arbutus Ridge, above), 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 8).

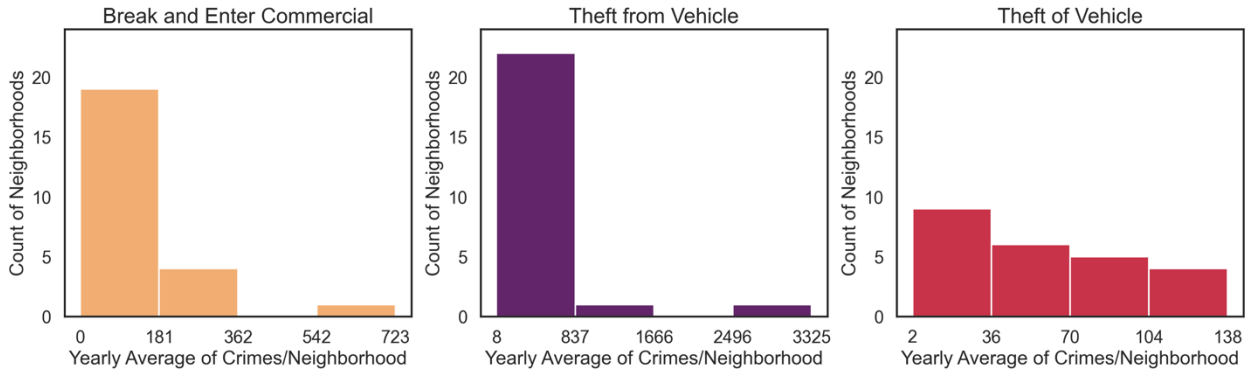

Figure 8: Distribution of yearly average crimes/neighborhood for the 3 selected types of crime

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 7).

Table 7: List of all 24 Vancouver neighborhoods classified by crime threat

	Neighborhood	Break/Enter Commercial Threat	Theft from Vehicle Threat	Theft of Vehicle Threat	Crime Threat Level
0	Arbutus Ridge	Low	Low	Low	Green
1	Central B.D.	High	High	High	Red
2	Dunbar-Southlands	Low	Low	Low	Green
3	Fairview	Mid-Low	Low	Mid-Low	Orange
4	Grandview-Woodland	Low	Low	High	Red
5	Hastings-Sunrise	Low	Low	Mid-High	Orange
6	Kensington-Cedar Cottage	Low	Low	Mid-High	Orange
7	Kerrisdale	Low	Low	Low	Green
8	Killarney	Low	Low	Mid-Low	Orange
9	Kitsilano	Low	Low	Mid-Low	Orange
10	Marpole	Low	Low	Mid-Low	Orange
11	Mount Pleasant	Mid-Low	Low	High	Red
12	Musqueam	Low	Low	Low	Green
13	Oakridge	Low	Low	Low	Green
14	Renfrew-Collingwood	Low	Low	High	Red
15	Riley Park	Low	Low	Mid-Low	Orange
16	Shaughnessy	Low	Low	Low	Green
17	South Cambie	Low	Low	Low	Green
18	Stanley Park	Low	Low	Low	Green
19	Strathcona	Mid-Low	Low	Mid-High	Orange
20	Sunset	Low	Low	Mid-High	Orange
21	Victoria-Fraserview	Low	Low	Mid-Low	Orange
22	West End	Mid-Low	Mid-Low	Mid-High	Orange
23	West Point Grey	Low	Low	Low	Green

Now, it is helpful to visualize a map (Figure 9) 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)

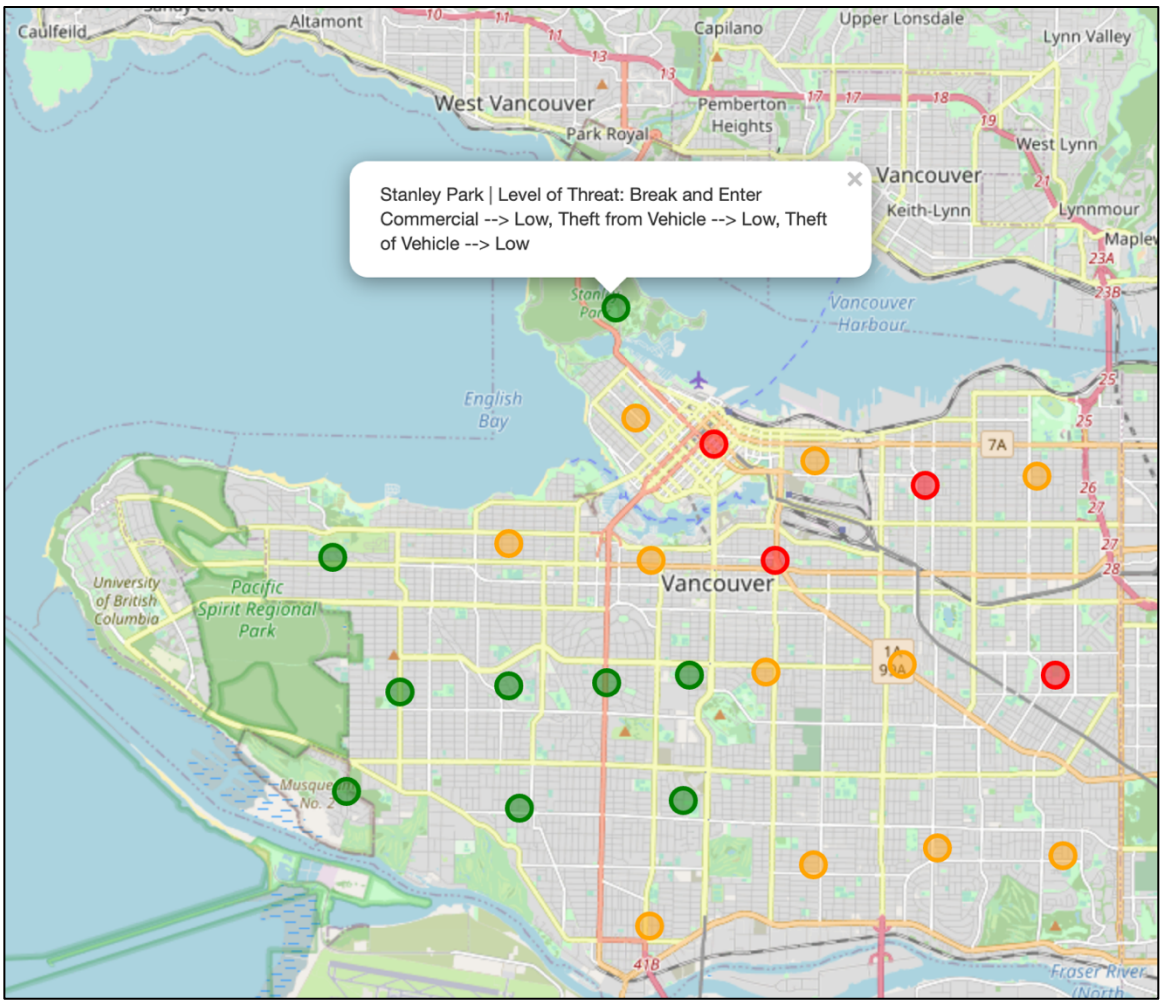

Figure 9: Map of Vancouver showing all 24 neighborhoods classified per crime threat

4.2. EDA and arrangement of Venues Data

With the data from commercial venue (refer to the previous Table 4), it would be interesting to visually explore the number of venues found per neighborhood (Figure 10).

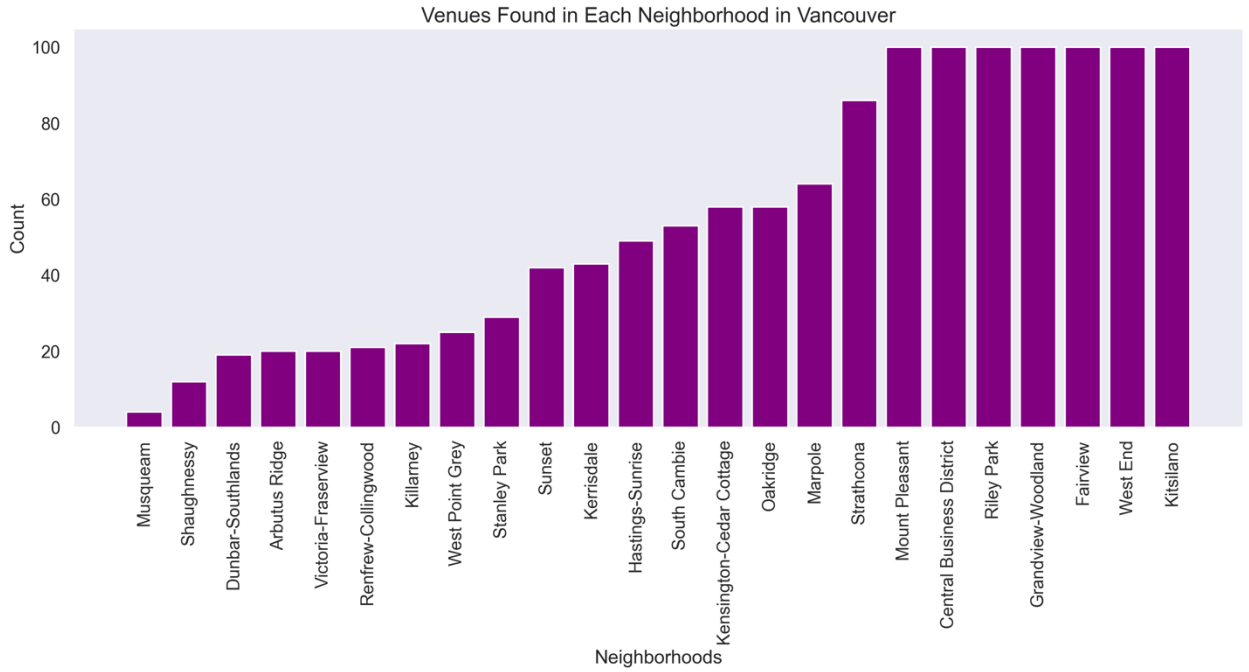

Figure 10: Count of number of venues found per neighborhood (limit 100 venues)

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 a one-hot encoding operation to create new columns with a binary value (0 or 1) for each type of venue for each neighborhood (one column per type of venue as per Table 4). 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 8 is an example of the first few rows of this data frame).

Table 8: 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 9 is an example of the first few rows of this data frame).

Table 9: 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

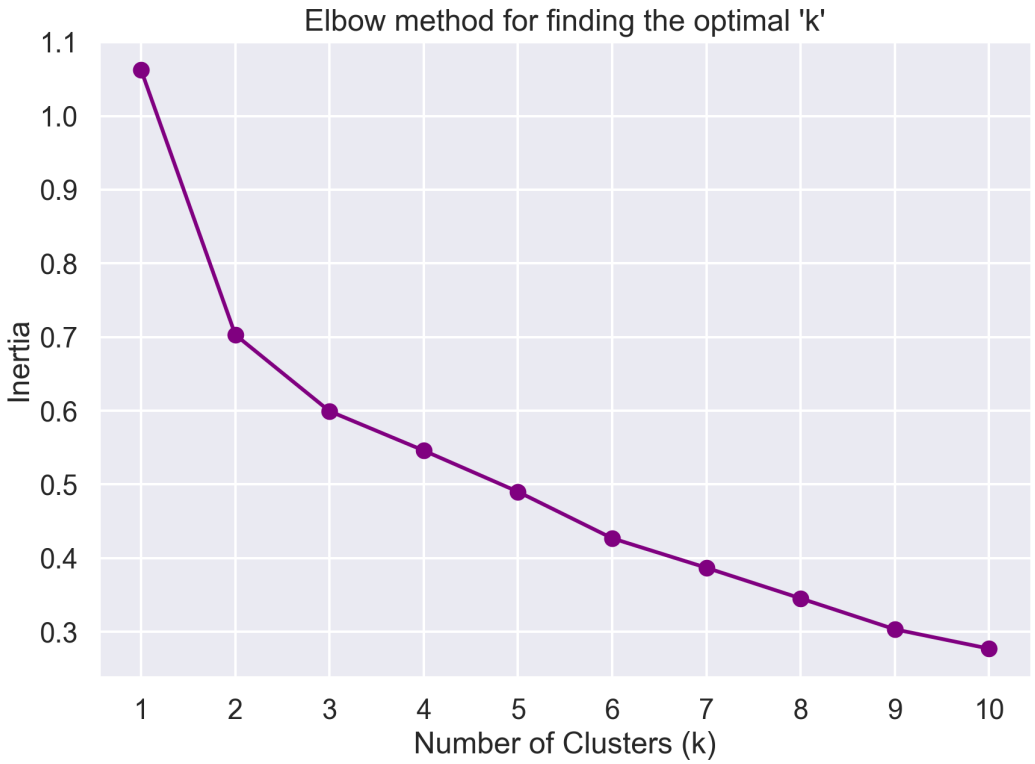

Figure 11: Inertia values for each 'k' value from applying K-Means Clustering to the Venue Data (example in Table 8)

Now, K-Means Clustering will be used for classifying the neighborhoods according to the frequency or appearance of the different venues in each of them. This machine learning tool was selected since it provides unsupervised learning (which is the required case) and specializes in classification. Yet, before using it, it is necessary to determine the optimal number of clusters or 'k' by exploring the inertia of the model. The figure above (Figure 11) shows the inertia values obtained from applying K-Means clustering to the prepared data (which follows the format shown in Table 8) for a list of 'k' values from 1 to 10, to find the 'k' that drives the elbow or drastic inertia reduction.

From the graph, there is a clear 'elbow' that can be identified for k=2, with another 'less severe' in k=3. This trend continues for the next k values, reducing inertia by ~0.05 with each k+1 increase. From exploring 'k' values from k=2 till k=6, the following results are obtained (Table 10):

Table 10: Number of clusters, inertia value, and number of neighborhoods for each 'k' value

K Value (Number of Clusters)	Number of Neighborhoods in Each Cluster	Inertia value
2	C 0: 1 N | C 1: 23 N	0.70
3	C 0: 21 N | C 1: 1 N | C 2: 2 N	0.60
4	C 0: 21 N | C 1: 1 N | C 2: 1 N | C3: 1 N	0.55
5	C 0: 10 N | C 1: 11 N | C 2: 1 N | C3: 1 N | C4: 1 N	0.49
6	C 0: 9 N | C 1: 11 N | C 2: 1 N | C3: 1 N | C4: 1 N | C5: 1 N	0.43

After analysing the results of each K-Means with different k values (from k=2 to k=6), some conclusions can be drawn:

- For k=2, only 1 neighborhood was separated from the rest. Although the inertia was reduced considerably, and an elbow is clearly seen in this instance, it is not very helpful having a cluster of 23 neighborhoods and another of only 1 neighborhood
- For k=3, 3 clusters are obtained. The biggest one still has 21 neighborhoods out of 24. Inertia keeps going down at a slower rate ~0.1
- For k=4, inertia keeps reducing, but the biggest one still has 21 neighborhoods out of 24. The only difference with k=3 is that the cluster with 2 neighborhoods was broken into 2 clusters, with 1 neighborhood each
- For k=5, there is a breakthrough, inertia keeps going down, but now there are 2 clearly defined clusters with 10 and 11 neighborhoods respectively
- For k=6. The inertia further reduces, but at a smaller rate, and creating one more cluster with only 1 neighborhood, which is not optimal. Therefore, the selected k value will be k=5.

The next step is to add the labels to each cluster in the master pivot table by giving names according to the types of venues that they group. A bar chart showing the number of venues per type in each cluster can be useful for determining which type of venue are in each cluster (for the graph, the project uses the 1st most common venue for each neighborhood). See Figure 12.

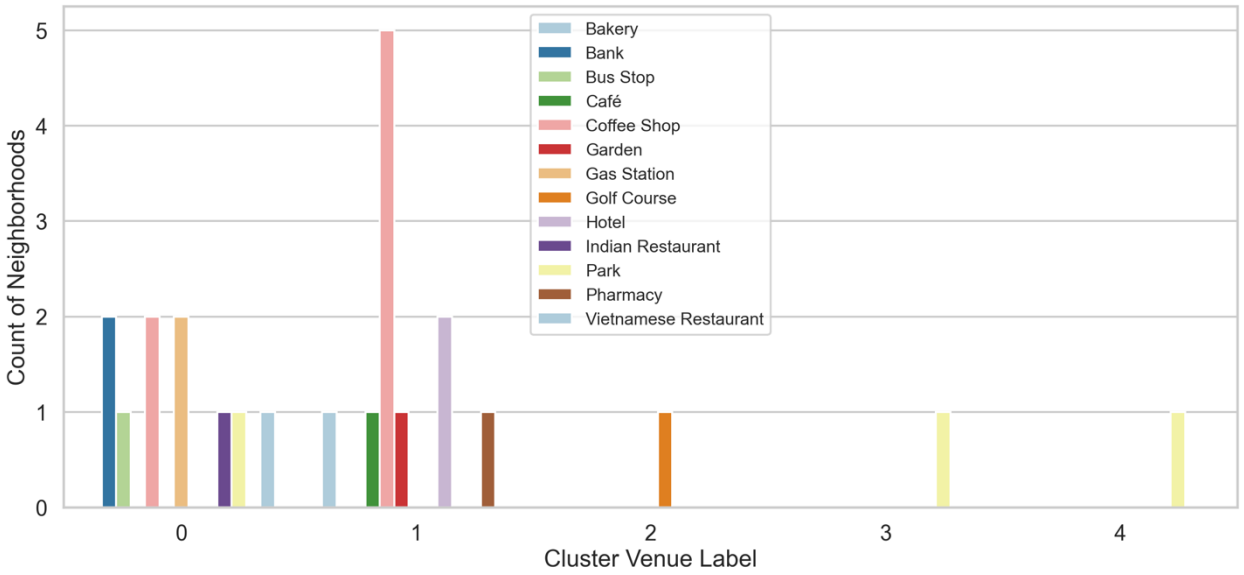

Figure 12: Count of venue types in each cluster as 1st Most Common Venue in each neighborhood

From the graph, the 4 groups can be identified as:

- Cluster 0: 'Restaurants, Gas Stations and Banks' (Banks, Coffee Shops, Gas Stations, Bus Stops, and ethnic restaurants)
- Cluster 1: 'Coffee Shops intensive and Hotels' (Coffee Shops, Cafés, Hotels, Bakeries)
- Cluster 2: 'Golf Course'
- Cluster 3: 'Outdoor Activities Locations 1' (small differences between Cluster 3 and 4 detected by K-Means)
- Cluster 4: 'Outdoor Activities Locations 2' (small differences between Cluster 3 and 4 detected by K-Means)

6. Results and Discussion

Before discussing the results, it will be useful to visualize the results in both table and map (Folium) format (Table 11 and Figure 13).

Table 11: Vancouver neighborhoods classified by threat of Crime and Type of venues in cluster

	Neighborhood	Latitude	Longitude	Crime Threat Level	1st Most Common Venue	Type of Venue in Cluster
0	Arbutus Ridge	49.245787	-123.160772	Green	Bakery	Coffee Shop int. & Hotels
1	Central Business District	49.281248	-123.114846	Red	Hotel	Coffee Shop int. & Hotels
2	Dunbar-Southlands	49.245016	-123.185095	Green	Pharmacy	Coffee Shop int. & Hotels
3	Fairview	49.264283	-123.128927	Orange	Coffee Shop	Coffee Shop int. & Hotels
4	Grandview-Woodland	49.275277	-123.067623	Red	Coffee Shop	Coffee Shop int. & Hotels
5	Hastings-Sunrise	49.276450	-123.042535	Orange	Park	Restaurants, Gas Stations and Banks
6	Kensington-Cedar Cottage	49.248876	-123.072525	Orange	Vietnamese Restaurant	Restaurants, Gas Stations and Banks
7	Kerrisdale	49.227909	-123.158447	Green	Coffee Shop	Restaurants, Gas Stations and Banks
8	Killarney	49.220965	-123.036435	Orange	Gas Station	Restaurants, Gas Stations and Banks
9	Kitsilano	49.266763	-123.160811	Orange	Coffee Shop	Coffee Shop int. & Hotels
10	Marpole	49.210556	-123.129410	Orange	Bank	Restaurants, Gas Stations and Banks
11	Mount Pleasant	49.264247	-123.101137	Red	Coffee Shop	Coffee Shop int. & Hotels
12	Musqueam	49.230380	-123.197089	Green	Golf Course	Golf Course
13	Oakridge	49.229036	-123.121757	Green	Bus Stop	Restaurants, Gas Stations and Banks
14	Renfrew-Collingwood	49.247458	-123.038422	Red	Park	Outdoor Act. Locations 2
15	Riley Park	49.247870	-123.103279	Orange	Coffee Shop	Coffee Shop int. & Hotels
16	Shaughnessy	49.246186	-123.138782	Green	Park	Outdoor Act. Locations 1
17	South Cambie	49.247459	-123.120235	Green	Coffee Shop	Restaurants, Gas Stations and Banks
18	Stanley Park	49.301157	-123.136932	Green	Garden	Coffee Shop int. & Hotels
19	Strathcona	49.278847	-123.092194	Orange	Café	Coffee Shop int. & Hotels
20	Sunset	49.219552	-123.092548	Orange	Indian Restaurant	Restaurants, Gas Stations and Banks
21	Victoria-Fraserview	49.222096	-123.064687	Orange	Gas Station	Restaurants, Gas Stations and Banks
22	West End	49.285099	-123.132311	Orange	Hotel	Coffee Shop int. & Hotels
23	West Point Grey	49.264637	-123.200172	Green	Bank	Restaurants, Gas Stations and Banks

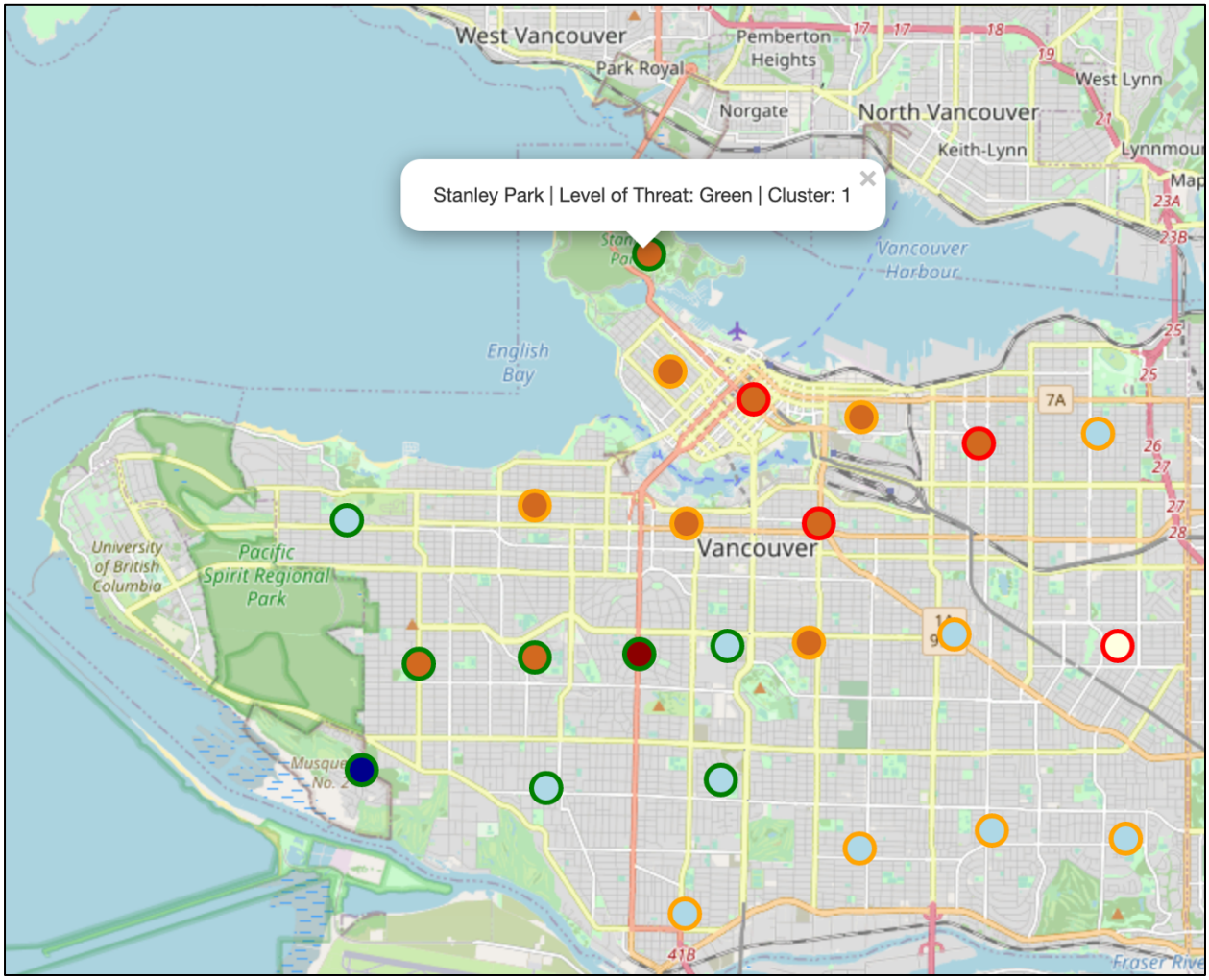

Figure 13: Map of Vancouver with neighborhoods classified by threat of Crime and Type of venues in cluster (Legend below)

Classification by level of threat (color of circle borders):

- 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)

Classification by type of venue cluster labels (color of circle fillings):

- Cluster 0 (light blue): 'Restaurants, Gas Stations and Banks'
- Cluster 1 (chocolate brown): 'Coffee Shops intensive and Hotels'
- Cluster 2 (dark blue): 'Golf Course'
- Cluster 3 (dark red): 'Outdoor Activities Locations 1'
- Cluster 4 (light yellow): 'Outdoor Activities Locations 2'

Regarding safety, the project classified the 24 neighborhoods of Vancouver in 3 clearly defined categories: 9 safe neighborhoods (green borders), 11 relatively safe with caution neighborhoods (orange borders), and 4 dangerous neighborhoods (red borders). This classification was made using the data from the Kaggle Vancouver Crime Dataset, which, although was quite detailed, did not include data for years after 2016 (2017 was incomplete). The analysis made showed little variation in the number of crimes per neighborhood and type across the years used in the project (2014, 2015 and 2016), which gave us enough confidence to use the data of these years to draw conclusions. Still, to further support the final recommendation, it would be better to have more recent crime information (to cover possible crime trends that might have changed due to COVID-19 pandemic).

Regarding the types of neighborhoods by venue, the project identified 5 main clusters of neighborhoods, which are: 'Coffee Shops intensive and Hotels' (11 neighborhoods), 'Restaurants, Gas Stations and Banks' (10 neighborhoods), 'Golf Course' which is Musqueam as expected, 'Outdoor Activities Locations 1' and 'Outdoor Activities Locations 2' (these 2 clusters have small differences detected by the machine learning algorithm). The method for finding these clusters was quite straightforward with the K-Means Clustering method, and the biggest challenge was to find the right 'k' value (finally established at k=5). The main concern here was to obtain the smallest number of clusters with the lowest inertia, while obtaining the most balanced clusters (similar amounts of neighborhoods amongst each). Other machine learning algorithms could have been used, like Hierarchal Clustering, but overall, the results obtained with the classification provided by K-Means are quite satisfying. It would have been interesting to compare the results of the two algorithms, which would be a task for another project.

7. Conclusion

To comply with Argus' priorities for finding a neighborhood that can host its first physical office, the project recommends:

Security-wise, the safest neighborhoods are the ones identified with circles with green borders: Arbutus Ridge, Dunbar-Southlands, Kerrisdale, Musqueam, Oakridge, Shaughnessy, South Cambie, Stanley Park, and West Point Grey. From the map, it can be noticed that most of these neighborhoods are on the west side of the city, mostly in more affluent areas.

Regarding type of commercial venues, the most convenient neighborhoods are the ones in Clusters 0 and 1 ('Coffee Shops and Other Commercial' and 'Restaurants, Gas Stations and Banks'). Argus mentioned that they want to be as close as possible to young entrepreneurs that own several types of business, mainly restaurants and coffee places which are split amongst both clusters. With more specific information about which type of business these entrepreneurs own (banks, bakeries, hotels, Indian restaurants, etc.), the project could make a call between Clusters 0 and 1, but for the

moment recommending both is the right choice. Combining this criterion with the previous one (safety) for finding the neighborhoods that comply with both, then the best neighborhoods for Argus first office are:

- Arbutus Ridge
- Dunbar-Southlands
- Kerrisdale
- Oakridge
- South Cambie
- West Point Grey

A third criterion is worth mentioning. The total amount of venues in each neighborhood is also relevant for Argus, as more venues means more business opportunities for the company. Considering this factor as well, and establishing a floor of having at least 40 commercial venues in the selected neighborhoods, the project can shorten the list to:

FINAL RECOMMENDATION

- **Kerrisdale**
- **Oakridge**
- **South Cambie**