Photo by: Aditya Chinchure
(https://unsplash.com/s/photos/vancouver)

Choosing the Neighborhood for a Micro-Lending Company in Vancouver

Applied Data Science Capstone Project

Project Author: Carlos Paiva González

Table of Contents

1 Business problem

2 Data acquisition and cleaning

3 Exploratory Data Analysis for crime data

4 Neighborhood classification by crime threat

5 Exploratory Data Analysis for commercial venues data

6 K-Means Clustering for neighborhood classification by type of commercial venues

7 Conclusions and next steps

A micro-lending company wants to know which Vancouver neighborhood offers them both security and commercial prospects for its first office

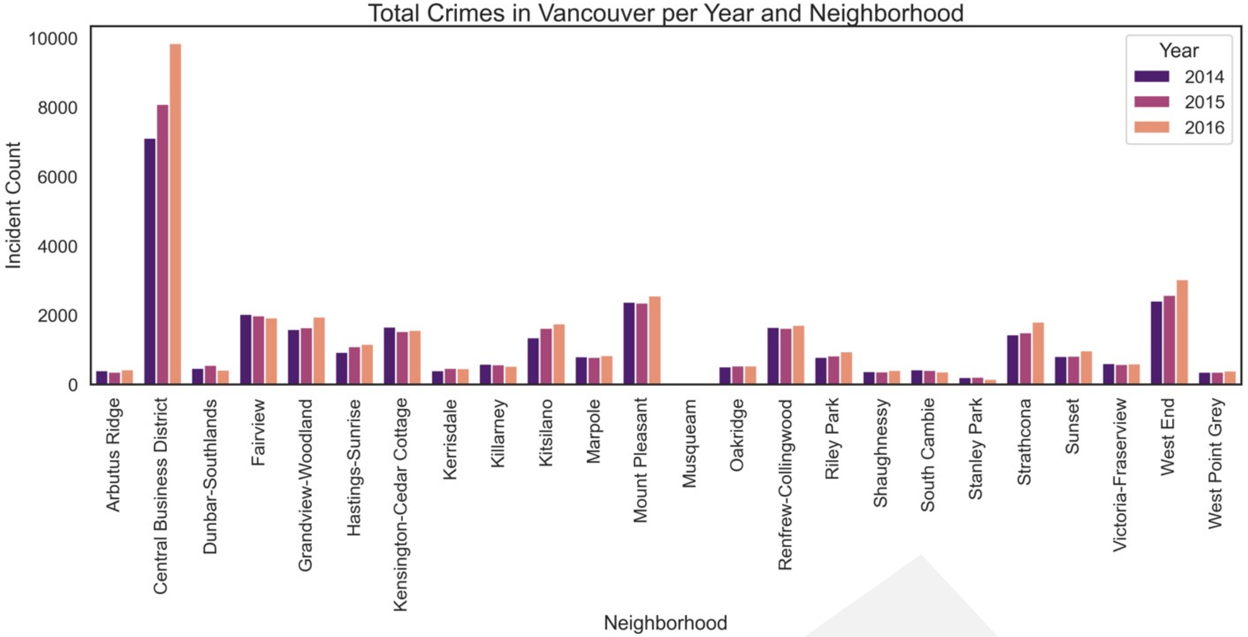
- Argus is a small micro-lending Canadian Company that provides financial assistance to entrepreneurs in the form of loans and other consulting services
- Initially, they had a 100% online model , but new clients are requesting cash-only loans and other in-person services, motivating Argus to establish its first physical office in Vancouver (where most of their clients are located)
- Considering which neighborhood would be the best suited for their office, Argus has two priorities:
    - **Security** (being established in a safe neighborhood, as clients will go in and out the premises holding considerable amounts of cash)
    - **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)

Note: Rental cost per sqm (another important factor for the decision) is out of scope for this project

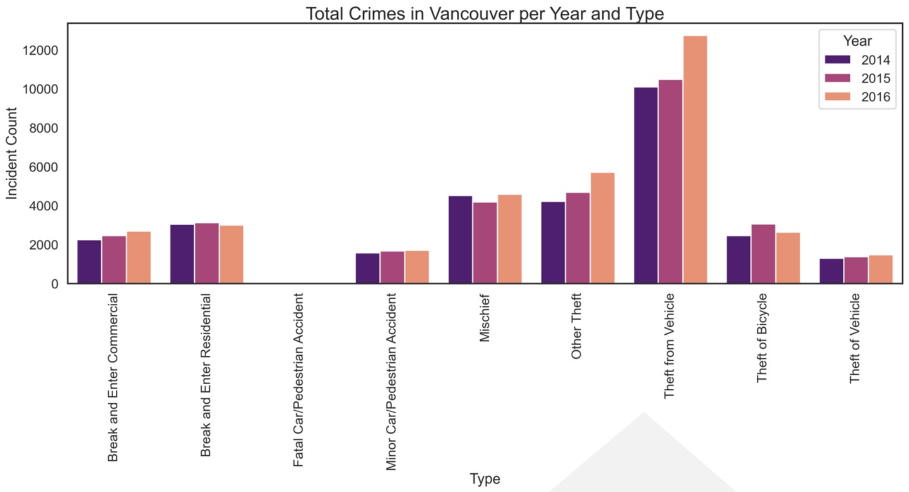
3 main data sources were used: Vancouver crime historic data, coordinates of all 24 neighborhoods, and venues data for all neighborhoods

	Crime Data	Coordinates Data	Commercial Venues Data
Source	▪ Raw data in CSV format from Kaggle: link here	▪ First approach was to use Geolocator Python package (disregarded) ▪ Final approach: use 'Longitude' and 'Latitude' features from Crime Data	▪ Foursquare API call to obtain information for all commercial venues in each neighborhood. Limit of results: 100 venues / neighborhood
Original data content	▪ Original dataset contained 530,652 observations and 12 features ▪ Historic data from 2003 till 2017 (2017 only from January till July)	▪ Average of latitude and longitude of all crimes in Crime Dataset grouped by neighborhood to map the coordinates of each neighborhood	▪ Result of API call: raw data in JSON format that requires to be transferred to a Pandas data frame with the use of a custom function
Data after cleaning	▪ Features kept (6): 'Type of Crime', 'Year', 'Month', 'Neighborhood', 'Latitude', and 'Longitude' ▪ Historical data kept: years 2014, 2015, and 2016	▪ Features in created data frame: 'Neighborhood', 'Latitude', and 'Longitude'	▪ Final dataset contains 1,330 observations (one per venue) and 5 features: 'Neighborhood', 'Venue Name', 'Venue Latitude', 'Venue Longitude', and 'Venue Category'

Overall number of crimes remained consistent across the 3 years, showing a small increase of 5% from 2014 to 2015, and 11% from 2015 to 2016



Total number of crimes is consistent across the 3 years for all neighborhoods, except for Central Business District (showing a small increase per year)



Total number of crimes per type is consistent across the 3 years for all types except for 'Theft from Vehicle' (which shows a bigger increase in 2016)

Total number of crimes	2014	2015	2016
All neighborhoods	29,430	31,020	34,515

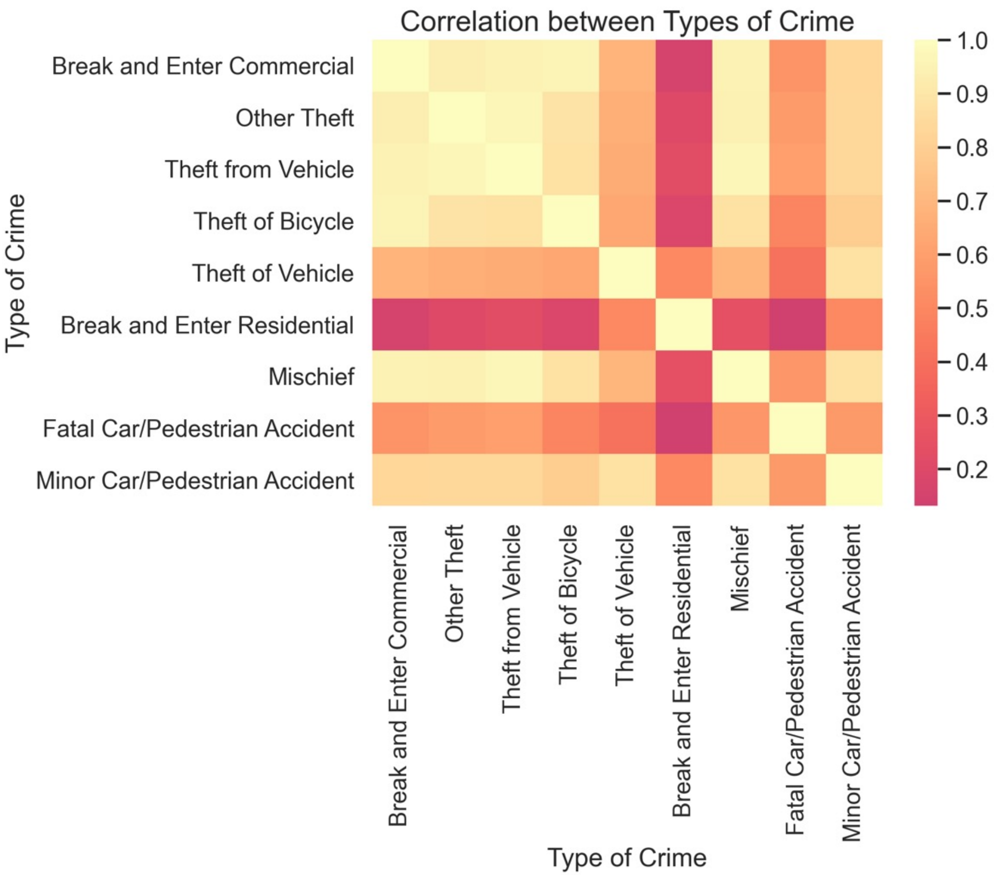
Since the project goal is to classify the neighborhoods by number of incidents to find the safest ones, and across all 3 years the number of incidents/neighborhood remains stable, it is possible to simplify the data by averaging the number of crimes per type and neighborhood

‘Break/Enter Commercial’ is the most relevant crime for Argus, ‘Other Theft’ and ‘Theft from Vehicle’ are the crimes most correlated with it

Heatmap Objective: to explore the relationship between the different types of crime that occur in Vancouver

Insights:

- Types of crime that are highly correlated amongst them: ‘Break and Enter Commercial’, ‘Other Theft’, ‘Theft from Vehicle’, and ‘Theft from Bicycle’
- ‘Break and Enter Residential’ shows the lowest correlation with the other types of crime
- Not all types of crimes are relevant. Some crimes like ‘Fatal Car/Pedestrian Accident’ and ‘Minor Car/Pedestrian Accident’ will be disregarded from the analysis since they do not pose a direct threat to Argus’ business

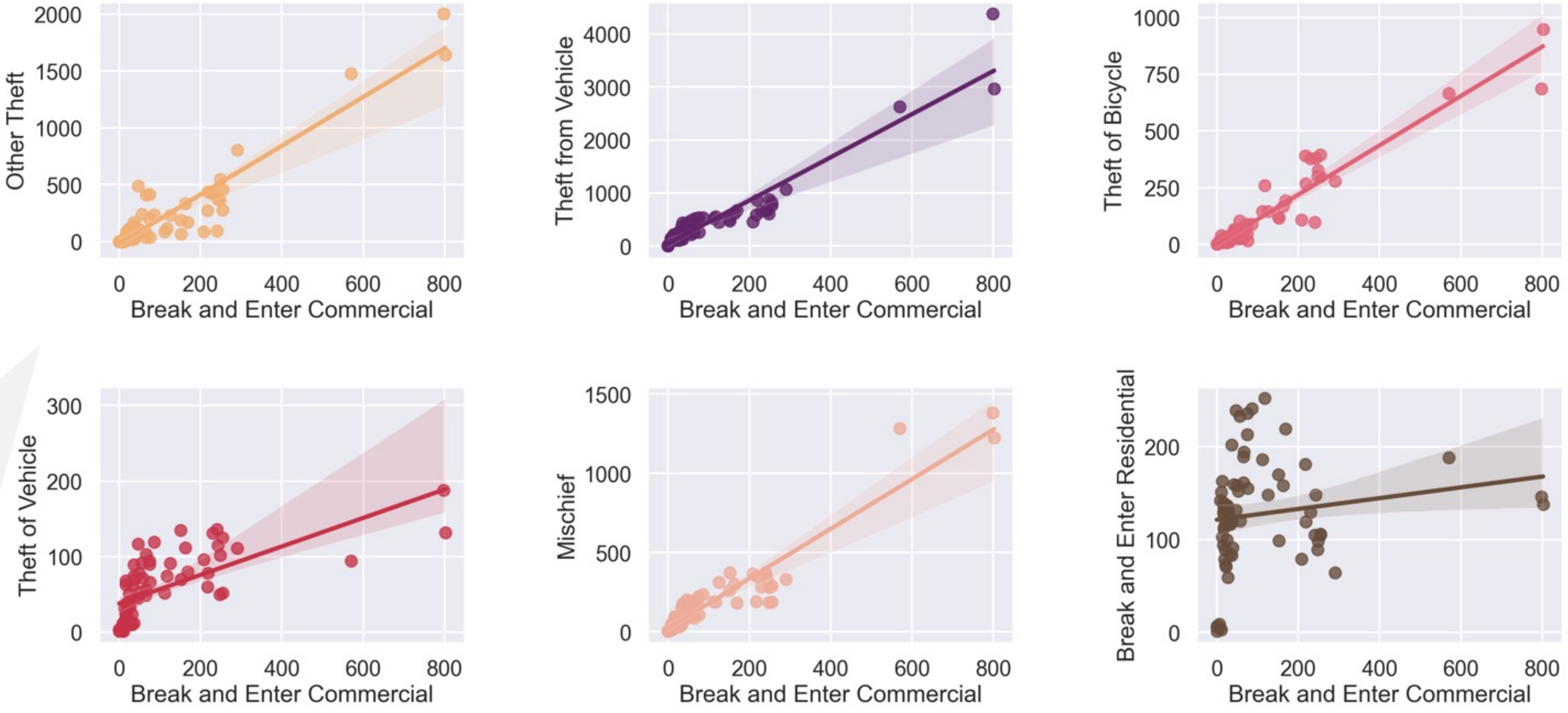


‘Break/Enter Commercial’, ‘Theft of Vehicle’, and ‘Theft from Vehicle’ are crimes selected to evaluate the criminal threat level of each neighborhood

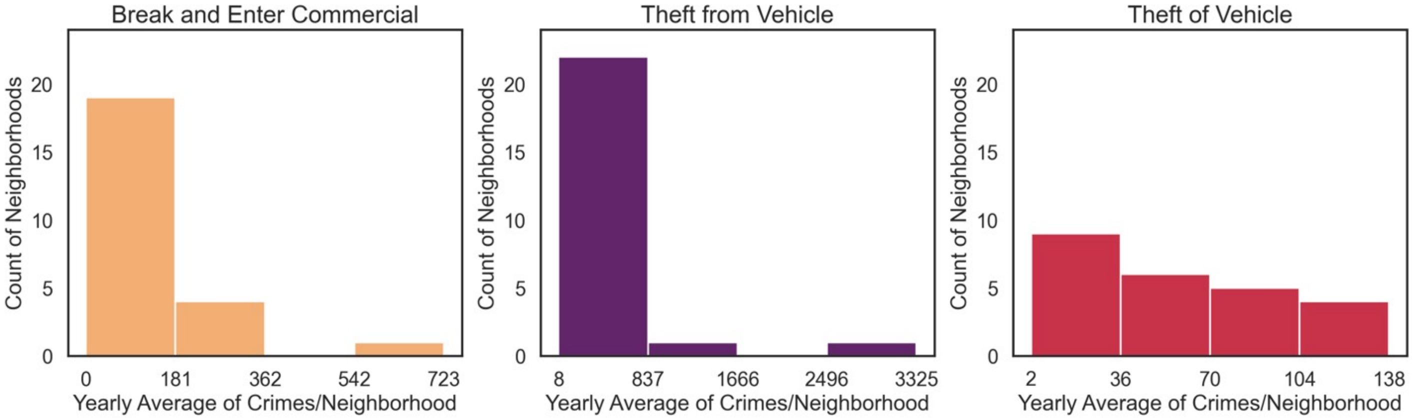
<u>Regression Plot Objective</u>: to explore the relationship between ‘Break/Enter Commercial’ and the relevant types of crime to select 2 types; one with a very high correlation with ‘Break/Enter Commercial’, and another with a smaller correlation but highly relevant to the project

<u>Insights</u>:

- ‘Theft from Vehicle’, ‘Theft of Bicycle’ and ‘Mischief’ show the highest correlation with ‘Break/Enter Commercial’
- ‘Theft of Bicycle’ will be disregarded since it is not very relevant to the project
- ‘Theft of Vehicle’ does not show a high correlation with ‘Break/Enter Commercial’, but is a relevant type of crime (count of vehicle thefts speaks about how dangerous an area can be for Argus’ businesses)
- Selected crime types: **‘Theft from Vehicle’, and ‘Theft of Vehicle’**

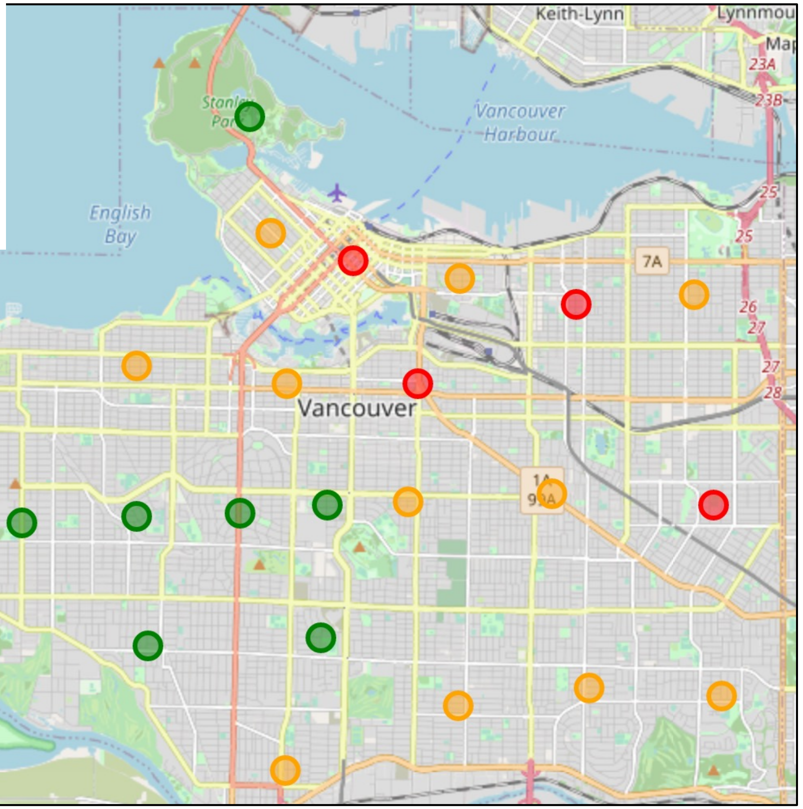


The 24 neighborhoods are classified (and color labeled) by crime threat level depending on the ‘bucket’ that they fall into for each crime type



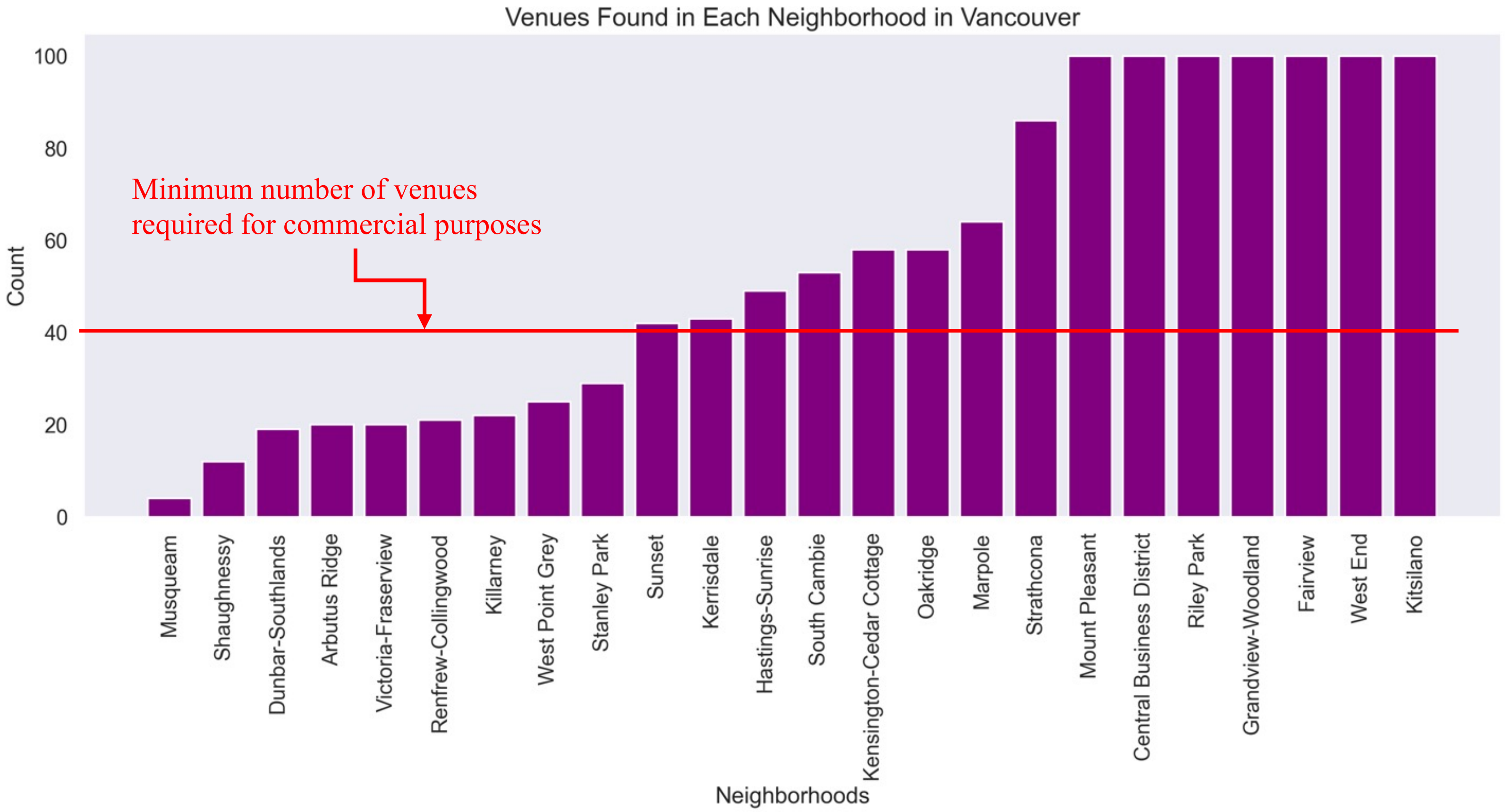
Vancouver Map

24 neighborhoods classified by crime threat



- Histograms are used to measure the distribution of each type of crime to create 4 buckets for classifying each neighborhood according to the crimes per type that they have
- The bucket categories are ‘Low Threat’, ‘Mid-Low Threat’, ‘Mid-High Threat’, and ‘High Threat’
- Then, neighborhoods will be given a danger color (plotted in map) depending on which category they fall for each type of crime:
  - Green or safe (if all 3 types of crime are in the ‘Low’ bucket)
  - Red or dangerous (if at least 1 type of crime is in the ‘High’ bucket)
  - Orange or relatively safe with caution (everything in between)

15 of 24 neighborhoods have more than 40 venues, which speaks about its commercial importance. 220 unique venue types were found in Vancouver



Minimum number of venues required for commercial purposes

Number of unique venue types: 220

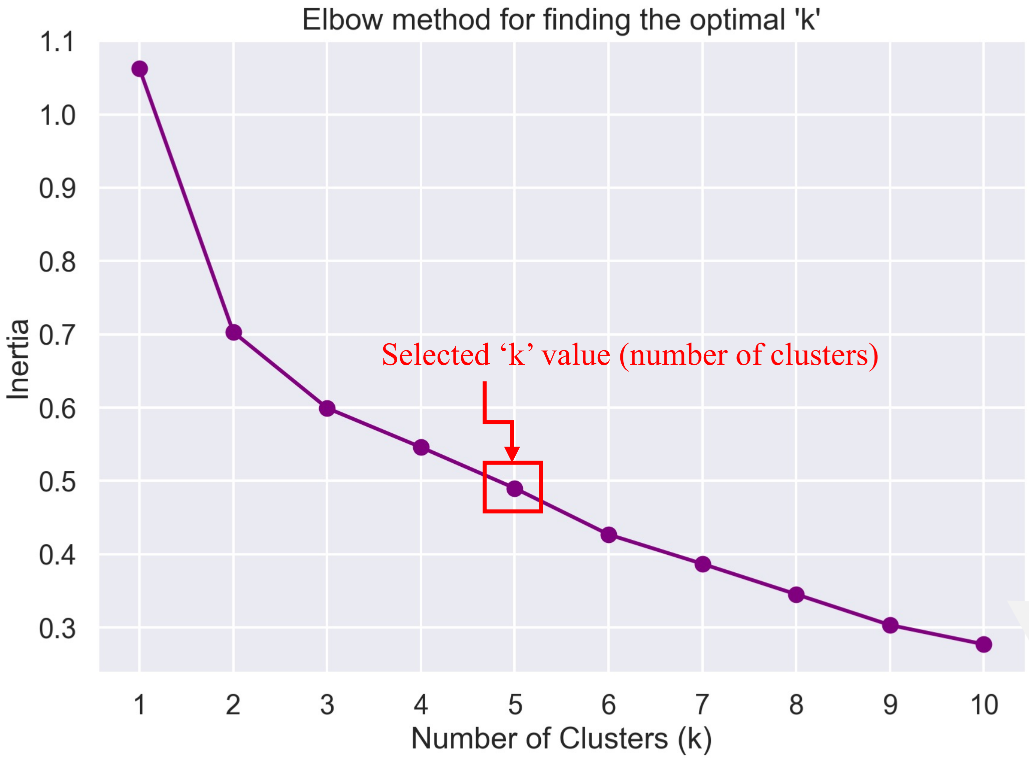
Total number of venues found: 1,326

Top 12 venue categories (# found):

- Coffee Shop (79)
- Park (52)
- Café (45)
- Sushi Restaurant (40)
- Chinese Restaurant (39)
- Vietnamese Restaurant (39)
- Bakery (34)
- Japanese Restaurant (33)
- Restaurant (27)
- Sandwich Place (26)
- Grocery Store (23)
- Bus Stop (23)

K-Means Clustering was used for classifying the 24 neighborhoods into 5 clusters according to the commercial characteristics of the venues within

K-Means Clustering will be used for classifying the neighborhoods according to the frequency or appearance of the different venues in each of them. Before using it, it is necessary to determine the optimal number of clusters or 'k' by exploring the inertia of the model ('elbow' method)

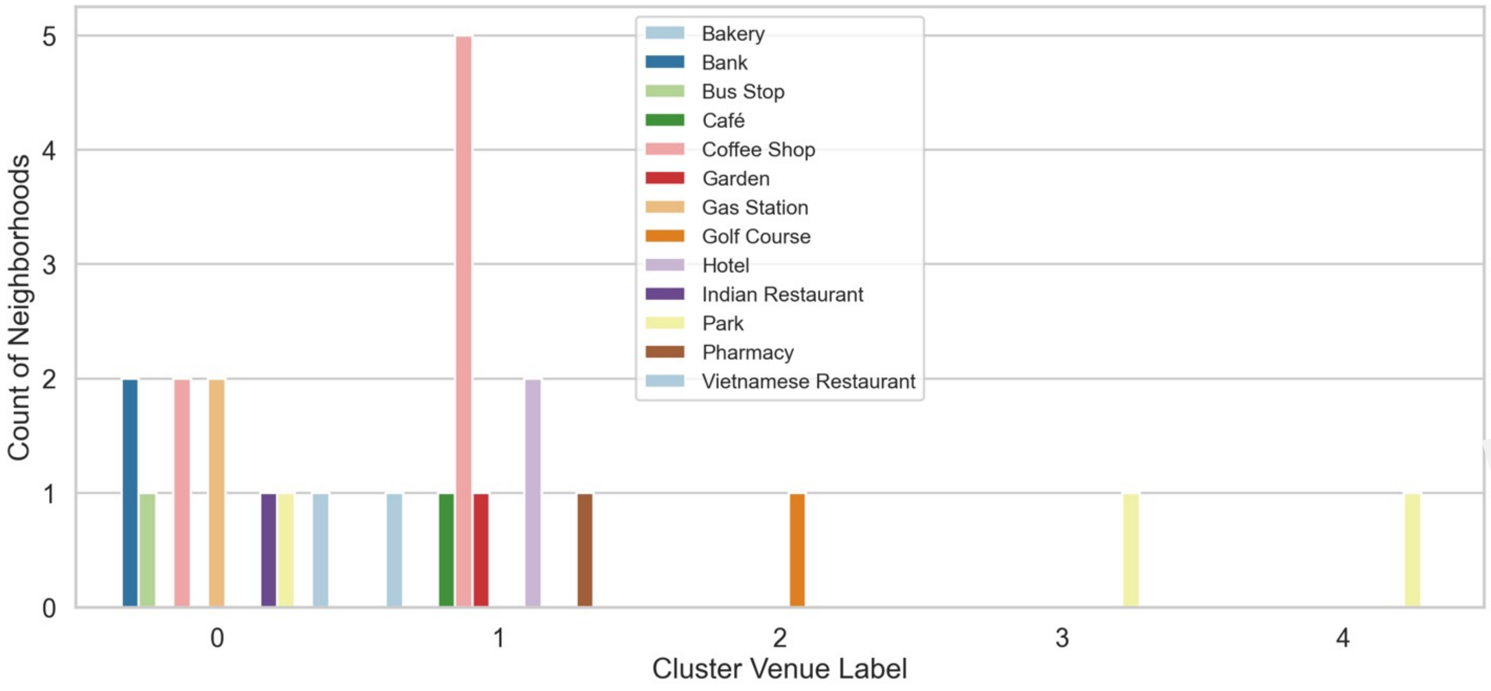


k value	Number of neighborhoods in each cluster	Inertia
2	C 0: 1 N | C 1: 23 N	0.70
3	C 0: 21 N | C 1: 1 N | C 2: 2 N	0.60
4	C 0: 21 N | C 1: 1 N | C 2: 1 N | C3: 1 N	0.55
5	C 0: 10 N | C 1: 11 N | C 2: 1 N | C3: 1 N | C4: 1 N	0.49
6	C 0: 9 N | C 1: 11 N | C 2: 1 N | C3: 1 N | C4: 1 N | C5: 1 N	0.42

- For k=2, inertia was reduced considerably, but only 1 neighborhood was separated from the rest (not useful). Refer to the table above
- For k=3, 3 clusters are obtained. Inertia keeps going down, but the biggest cluster still has 21 neighborhoods out of 24
- For k=4, inertia keeps reducing, yet results are still like k=3
- For k=5, there is a breakthrough, inertia keeps going down, but there are 2 defined clusters with 10 and 11 neighborhoods
- For k=6, inertia further reduces at a smaller rate, and creating one more cluster with only 1 neighborhood, which is not optimal

5 clusters have been identified, 1 with different commercial venues, 1 with coffee shops & similar, one golfing-specific, and 2 for outdoor activities

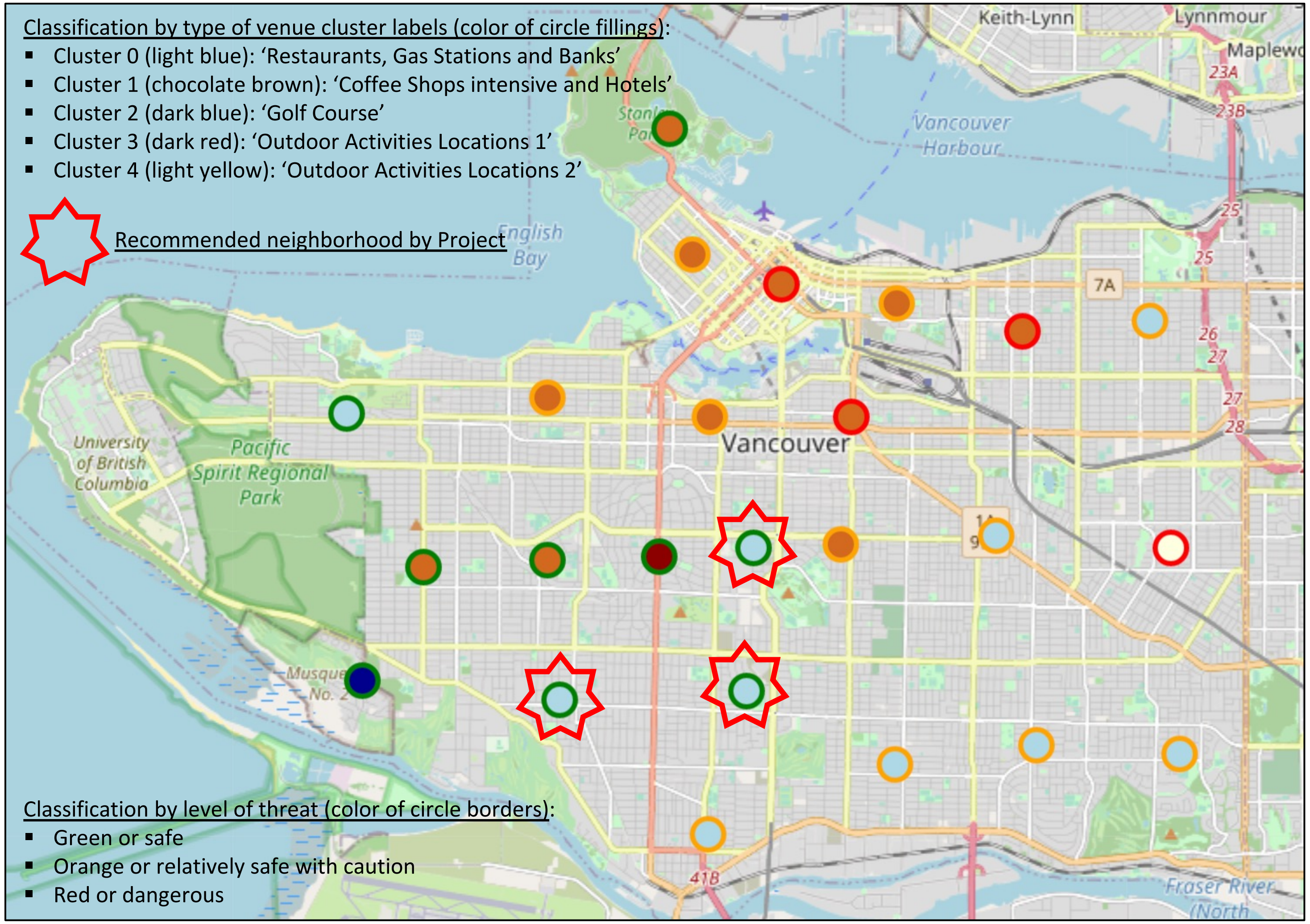
Objective: naming each cluster according to the types of venues that can be found within each of them. The bar chart below is helpful for determining which type of venue are within each cluster (the graph shows the 1st most common venue for each neighborhood)



The 5 clusters have been named as:

- **Cluster 0**: 'Restaurants, Gas Stations and Banks' (different commercial venues)
- **Cluster 1**: 'Coffee Shops intensive & Hotels' (Coffee Shops, Cafés, Hotels, Bakeries)
- **Cluster 2**: 'Golf Course'
- **Cluster 3**: 'Outdoor Activities Locations 1'
- **Cluster 4**: 'Outdoor Activities Locations 2' (like Cluster 3 with small differences between detected by K-Means)

Considering crime threats and commercial purposes, we recommend Argus to open its first office on Kerrisdale, Oakridge, or South Cambie


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'
Recommended neighborhood by Project
Classification by level of threat (color of circle borders):
Green or safe
Orange or relatively safe with caution
Red or dangerous
Keith-Lynn
Lynnmour
Maplewo
Vancouver Harbour
English Bay
Vancouver
University of British Columbia
Pacific Spirit Regional Park
Musque No. 2
Fraser River (North
7A
23A
23B
25
26
27
28
418

Conclusions:

The most recommended neighborhood are:

- Security-wise: green neighborhoods (total 9)
- Commercial-wise: Clusters 0 and 1 ('Coffee Shops & Other Commercial' and 'Restaurants, Gas Stations & Banks')
- Third criteria: minimum 40 commercial venues/neighborhood
- Final recommendation: **Kerrisdale, Oakridge, and South Cambie** (red stars on map)

Next steps for further improvement:

- Obtain more recent crime information (to cover crime trends that might have changed with COVID-19 pandemic)
- Comparing the results of K-Means Clustering for classifying neighborhoods by performing a similar exercise with Hierarchical clustering (another unsupervised ML algorithm)