# **Calgary’s Crime Impacted Businesses – Coursera Capstone Project**

## Introduction to Business Problem and Target Audience

### Background of the City

Calgary, Alberta, incorporated as a city in 1894, population 1,239,220 (2016 [census](https://www.thecanadianencyclopedia.ca/en/article/demographic-data-collection)) 1,096,833 (2011 census). The city of Calgary is situated on the Bow River in southern Alberta, about 220 km north of the American border at the meeting point of the Western prairies and mountain foothills. It is the financial centre of western Canada, based on its key role in the development of the region’s oil and gas industry.

In the recent past, Calgary has seen very polar economic conditions by virtue of Alberta’s dependence on Oil and Gas market. Like any other city, poor economic conditions lead to loss of jobs, which is often accompanied by an increase in crime rate. This increase in commercial crime rate leads to losses for businesses.

### Target Audience

Consequently, many passive investors find it tough to justify investing in businesses in Calgary. They are afraid of unforeseen losses incurred by business by virtue of commercial crimes in the city.

In order to obtain a decent return on investment, it is vital for both passive and active investors to avoid communities in the city where there is a historically high probability of commercial crimes by virtue of crime count.

## Data

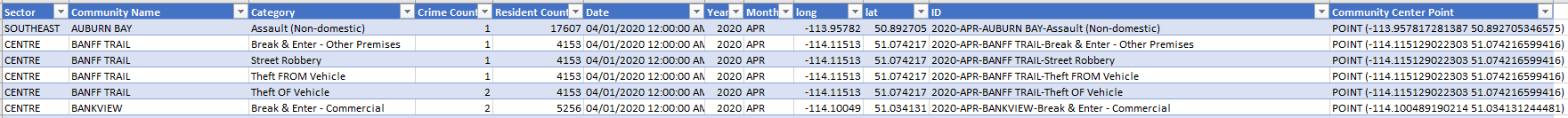
I will be using data from City of Calgary’s website to determine clusters of types of businesses in communities most impacted by commercial crimes. While Calgary is an economic hub in the province of Alberta, it is still a growing city with a moderate number of businesses, however not as many as you would expect to find in metro cities like New York, Toronto, or Vancouver.

### Crime Statistics (City of Calgary’s open data website)

We will be working with publicly available data repository managed by City of Calgary to find crime statistics in the various communities within Calgary.

<https://data.calgary.ca/Health-and-Safety/Community-Crime-Statistics/78gh-n26t>

I will scrape the data to only include types of crimes that have the potential to impact commercial businesses (e.g. commercial break-ins) and leave out data that has a low probability of affecting businesses (e.g. domestic violence).



### Nearby business venues (Foursquare API)

Subsequently, using geospatial data of these communities, we will use FourSquare API to find the businesses within five kilometer radius of these communities. We will be using the category of businesses to determine which type of businesses are most impacted by commercial crimes in Calgary. Finally, K-means clustering will be used to compartmentalize the findings and we shall attempt to identify the theme of each cluster.

## Methodology

### Data Cleaning

The data extracted from the City of Calgary’s website can be found at the following link:

<https://data.calgary.ca/Health-and-Safety/Community-Crime-Statistics/78gh-n26t>

We downloaded the CSV file and explored it briefly to understand the type and amount of data it contained. The original CSV file had 45570 rows vs 12 columns. While the data was formatted properly, we did find some challenges that had to be resolved before diving into data analysis and clustering.

Firstly, some columns had to be dropped as they were redundant to the target we had committed to. The following columns were dropped:

1. Sector: Shows which quadrant of the city community belongs to
2. Crime Count: Shows count of the specific type of crime in each community
3. Date: Shows date of the crime
4. Year: Shows year of the crime
5. Month: Shows month of the crime
6. ID: Contains short date concatenated with community name and category of crime
7. Community Center Point: shows the center point of the community

Next, since we decided to focus on commercial crimes, we filtered the Category column to truncate rows that didn’t mention the word “commercial” in its string. There were other types of crimes listed such as domestic abuse but we assumed that such crimes do not impact businesses within their vicinity.

Lastly, due to the multiple categories of crime, we had duplicates in the community name column. This would have caused us issues with Foursquare and sent more get requests for data to the API than were required. Therefore, we decided to drop duplicates from the “community name” column.

Once properly cleaned, we were left with 278 rows and 5 columns.

### Data Analysis

Once we had a refined list of communities in Calgary that had faced commercial crimes, we decided to pick out the first community to obtain venues for it using Foursquare API. We created a get URL request and passed the name of the community. We set the limit of results to 100 within a radius of 10 kms from the community’s center point using its latitude and longitude. The first community returned 100 results.

In order to extract venues for each community listed in our data frame, we used the getNearbyVenues to pass a list of community names to foursquare API and create a data frame that included venue name, its latitude, longitude, and its category against each neighborhood (community). 1466 venue results were returned by the FourSquare API.

In order to analyze our categorical data, we had to use one hot encoding method to convert our categorical variables to binary vectors. Subsequently, the results obtained from foursquare API were grouped by “Neighborhood” name and the mean for each neighborhood was found.

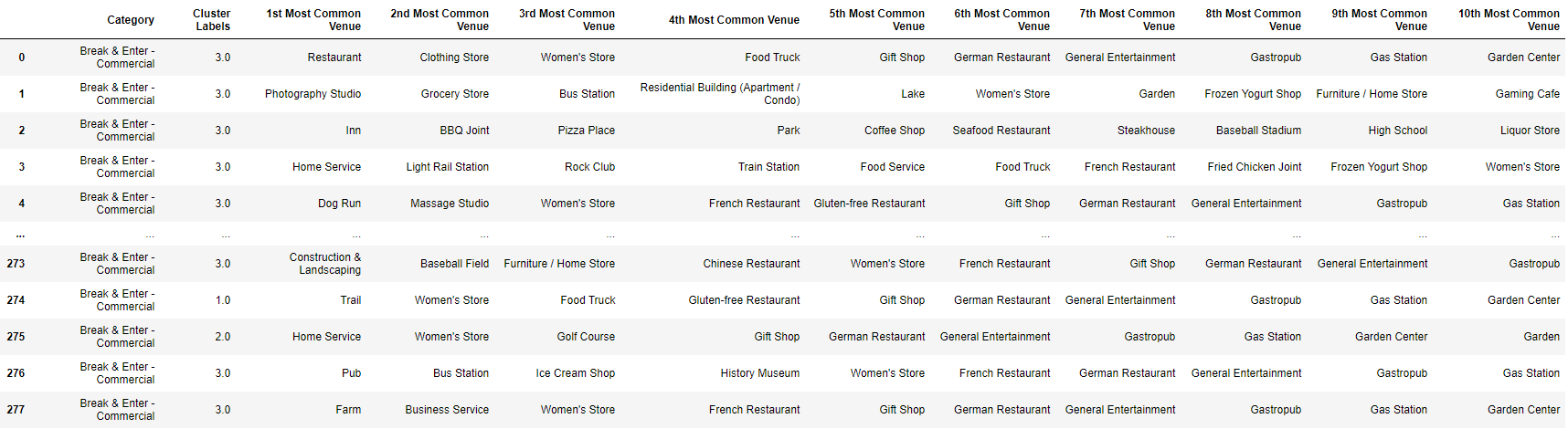
We then proceeded to find the top 5 venues in each neighborhood using frequency of venue category and then sort the venues in descending order.

#### Machine Learning

We used K-means clustering as our preferred Machine Learning method to create clusters of neighborhoods based on their categories. We assumed the correct number of clusters to be equal to 4.

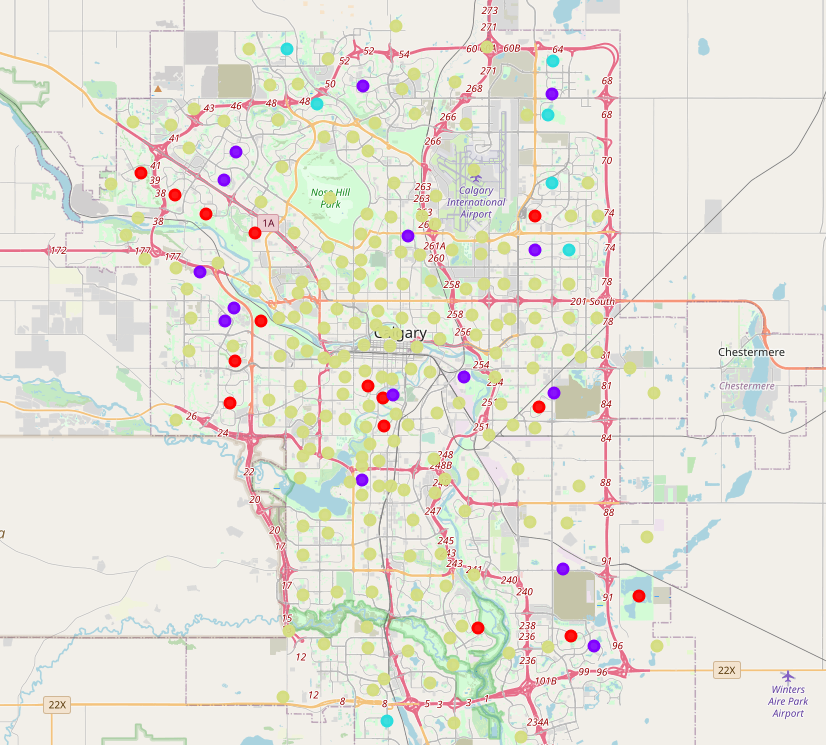
K-means clustering works by finding similarities in data and assigning them to centroids by method of minimizing the sum of squared errors between each data point and its associated cluster centroid. Since we had categorical data, K-means works by comparing the category of venues for each neighborhood.

## Results

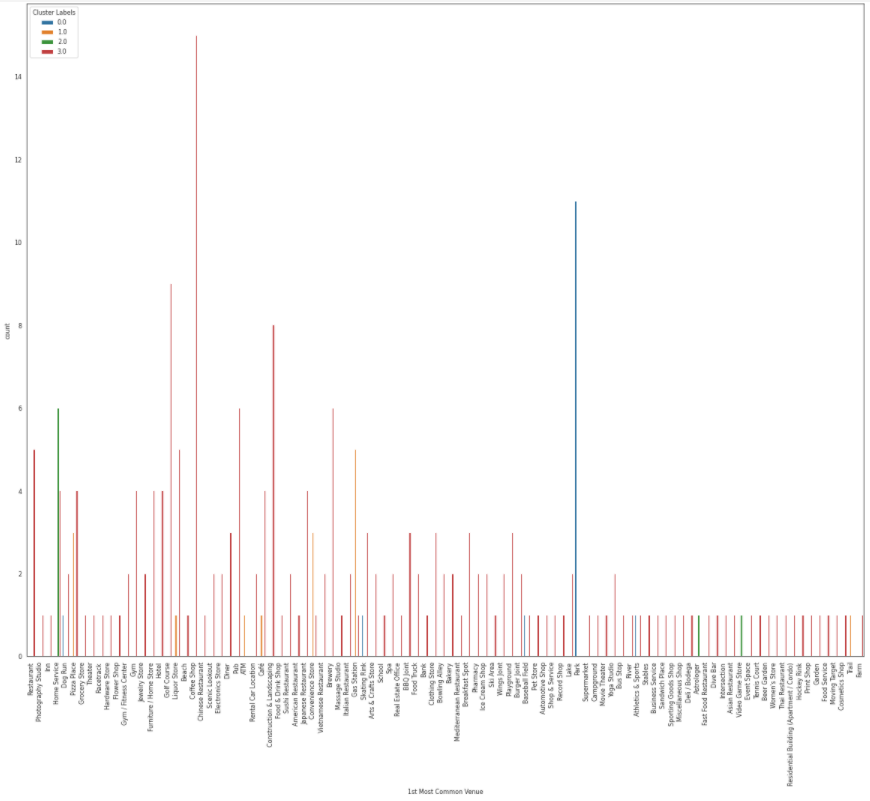


Using the results from our K-means clustering algorithm, we labelled each neighborhood with a cluster-label. These cluster results were plotted on a map of Calgary using folium.

The resulting clusters from our K-means clustering were plotted on a map. There was no geo-spatial pattern in our data as can be seen in the picture on the next page.



However, we decided to use seaborn to create a count plot of the 1st most common venues for each cluster. The resulting count plot is presented below.

Therefore, we can conclude the following from the our count plot with regards to the most common venues in each neighborhood impacted by commercial crimes:

1. Cluster 0: Parks, Skating Rinks, Dog Runs, Baseball Field
2. Cluster 1: Home Service, Gas station, Pizza Place, Convenience Store
3. Cluster 2: Home Service, video game store, astrologer
4. Cluster 3: Diner, Construction and Landscaping, Restaurant, Hotel.

As it can be seen, Foursquare doesn’t just return commercial venues but also locations of publicly owned places near each address. Consequently, we will be only using results whose categories are more often than not owned by public and private investors, and not held under government controlled non-profit assets.

## Discussion & Recommendations

Through observation of 1st most common venue by frequency for each cluster, we can make the following generalizations:

1. In cluster 0, businesses such as skating rinks and sporting complexes are most effected by commercial crimes.
2. In cluster 1, small size commercial shops such as gas stations, pizza shops, convenience stores are more effected by commercial crimes
3. In Cluster 2, businesses run out of home space are most effected by commercial crimes.
4. In Cluster 3, commercial lounging services are most effected by commercial crimes.

In conclusion, we can say that passive or active investors looking to open new businesses or invest in existing businesses, will have the highest possibility of suffering financial losses due to commercial crimes in the business categories defined above for communities listed in each cluster.

We will recommend that investors are extra careful and do their due diligence when opening up businesses in any of these neighborhood clusters.

## Conclusion

We can conclude from our study that we were indeed able to find a pattern in the data provided by foursquare API and make recommendations to our targeted audience based on the patterns found in each clusters.

While we could have benefited from choosing a bigger city for performing the same study in terms of results provided by Foursquare API, my choice for my hometown of City of Calgary was very crucial as it is a growing city where people are looking to open up new businesses.

Since we had categorical variables, we weren’t able to implement elbow method to predict the right number of k-means cluster but we can make this improvement next time.