Capstone Project
The Battle of Neighborhoods

Marcio Henrique dos Santos Silveira

Introduction

With little research, one can figure out Toronto is not on the list of the most dangerous cities in the world. However, it does not mean Torotonians are exempt from crime occurrence as we can analyze using data from Toronto Police Service, which shows an increase of 7% in the 2019 cases of auto theft in comparison with 2018.

This kind of data is precious to insurance companies because they must take into account as it affects its market. Thus, this report aims to analyse auto theft occurrence in Toronto with respect to the urban and communitarian infrastructure and facilities, such as parks, restaurants and other venues. Which could help insurance companies on pricing tasks.

Thus, the main goal of this analysis is to describe the characteristics of neighbourhoods in which crime happens.

Data

Auto theft data obtained from Toronto Police is going to be used in this analysis. The data sets contains auto theft cases in Toronto from 2014 to 2019.

Foursquare API is going to be used to retrieve all the venues in each neighbourhood, including its coordinates, categories and sub-categories.

Methodology

The auto theft data will be shown on a map to give an overview of the auto theft locations as well as the most dangerous neighbourhoods.

For the data analysis and machine learning task, the neighbourhoods will be clustered using its venues and the results will be compared with the occurrences in an attempt to identify which venues are more prone to 'attract' the thefts.

For that I will follow these steps:

- Account the total auto theft occurrences in each neighbourhood
    - Show on the map
    - Identify the sazonality of the occurrences

- Machine Learning
    - Cluster the neighbourhoods
        - Standardize data
        - PCA analysis/feature selection
    - Identify the five most common venues for each type of neighbourhood

Results and Discussion

The map below (figure 1) shows every occurrence of auto theft between 2014 and 2019 (white dots) and the colours shows the neighbourhoods in regard to the sum of the occurrences.

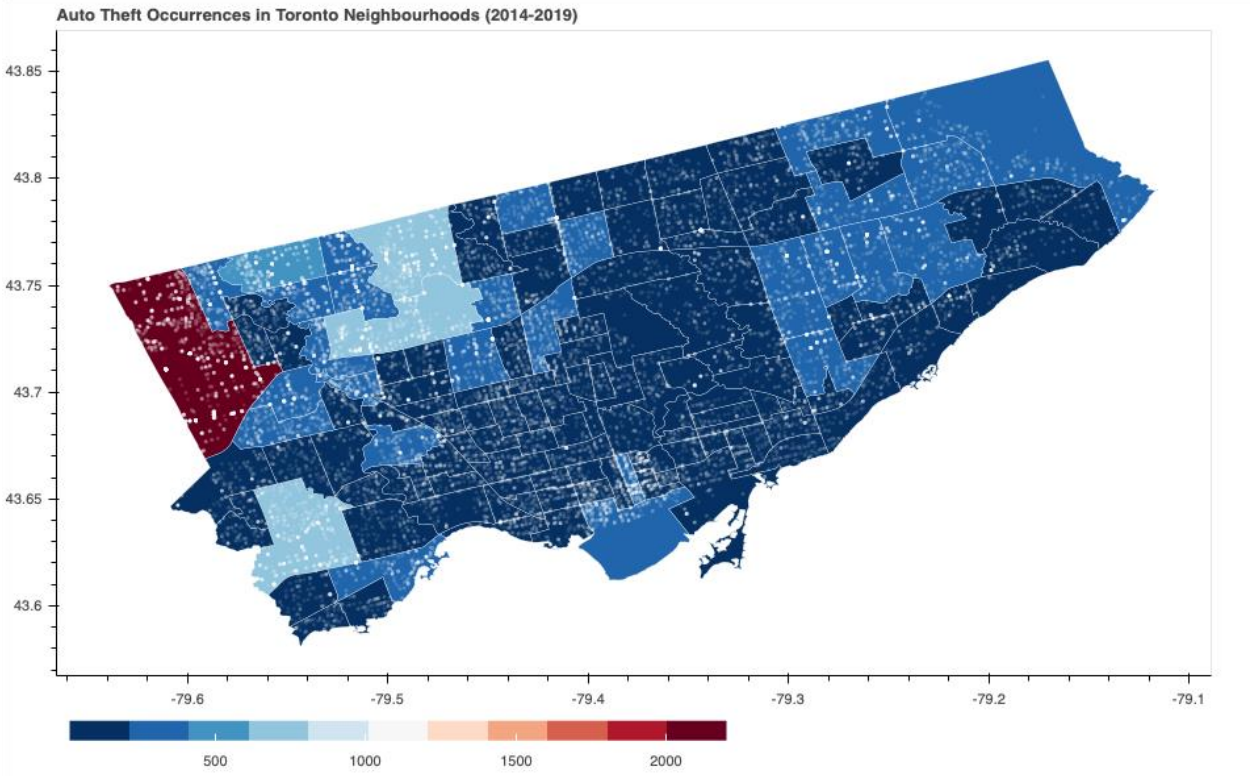

Figure 1: Auto theft in toronto Neighbourhoods

By far, West Humber-Clairville neighbourhood is the most dangerous, with 2200 records of auto theft, followed by Downsview-Roding-CFB, York University Heights and Islington-City Centre West with values around 1000 occurrences.

As we can see below on figure 2, the majority of the occurrences are outside houses.

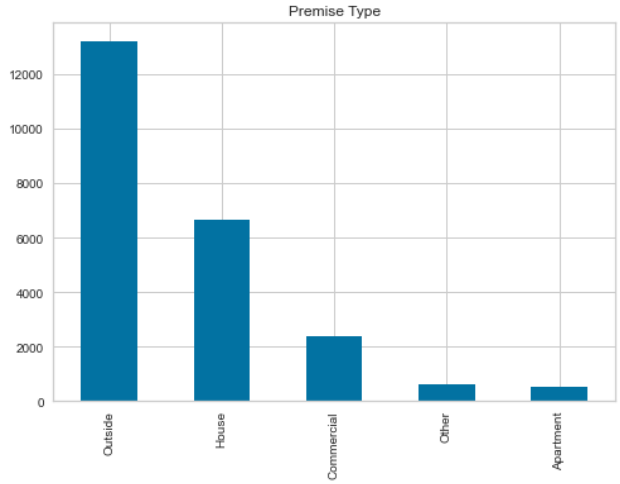

Figure 2: Auto theft occurrences premise types

The analysis of seasonality of the occurrences, can be seen on figures 3 and 4.

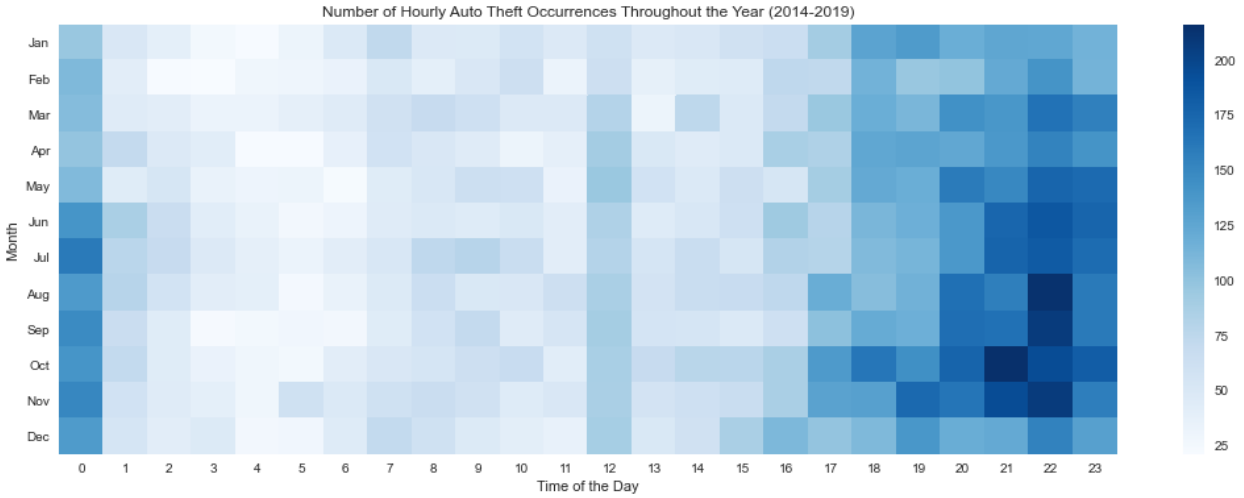

Figure 3: Hourly auto theft occurrences divided by month count.

The autumn months are more subjected to auto theft, with felonies happening predominantly at night between 8pm and midnight.

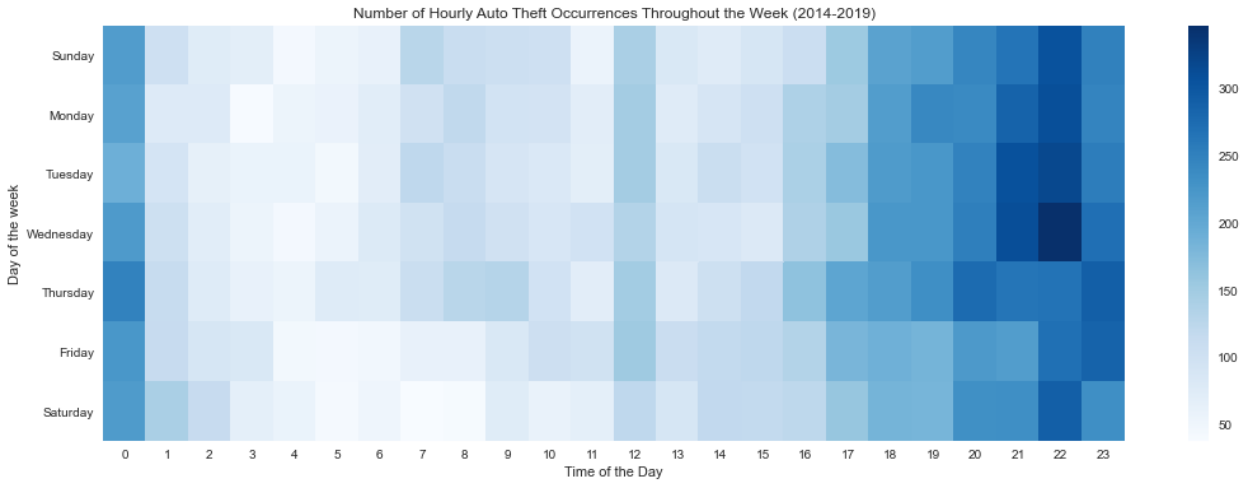

Figure 4: Hourly auto theft occurrences divided by week days count.

Occurrences are most common on Wednesday nights.

The venues used to cluster the neighbourhoods are plotted on the map at figure 5, where we can see the high concentration of commercial activity downtown, which does not precisely mean most crimes happened there.

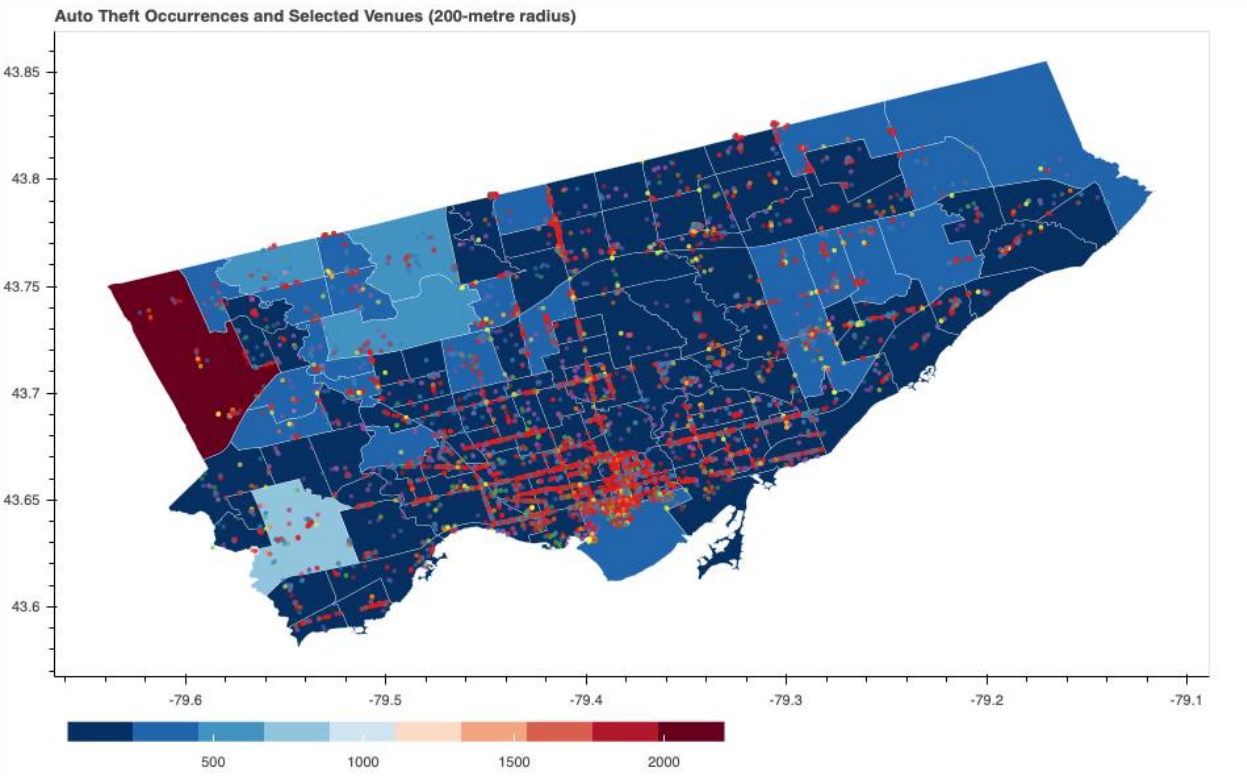

Figure 5: Venues analysed for characterizing the neighbourhoods.

For the clustering process, the data was scaled between zero and one. Then, a Principal Component Analysis were used to reduce the dimensionality of the data using Singular Value Decomposition of the data to project it to a lower dimensional space respecting 95% of explained variance. Finally, the number of clusters to be used in the KMeans algorithm was defined using the elbow technique, which found an optimal value of 3 clusters, as shown on figure 6, below.

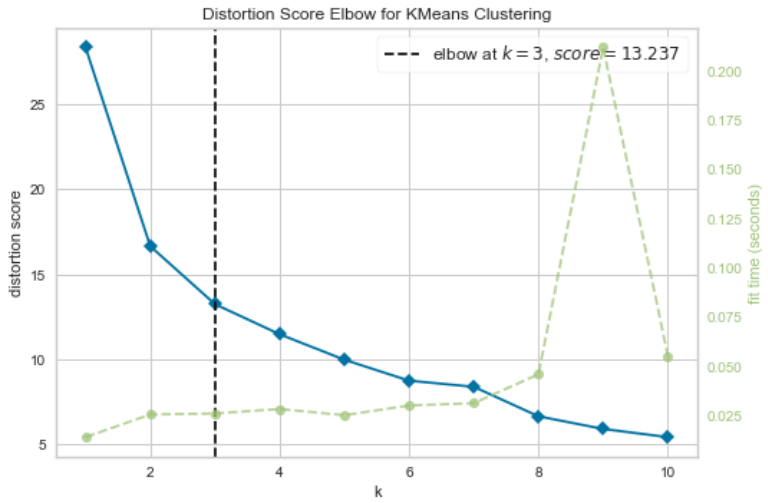

Figure 6: Elbow technique.

The three distinct neighbourhoods are described as folows. Cluster 0 is the one with most car theft occurrences (19665), followed by Cluster 2 (2433 occurrences) and Cluster 1, with 1282 occurrences. The figure 6, below, shows the most common venues associated with the clustered neighbourhoods.

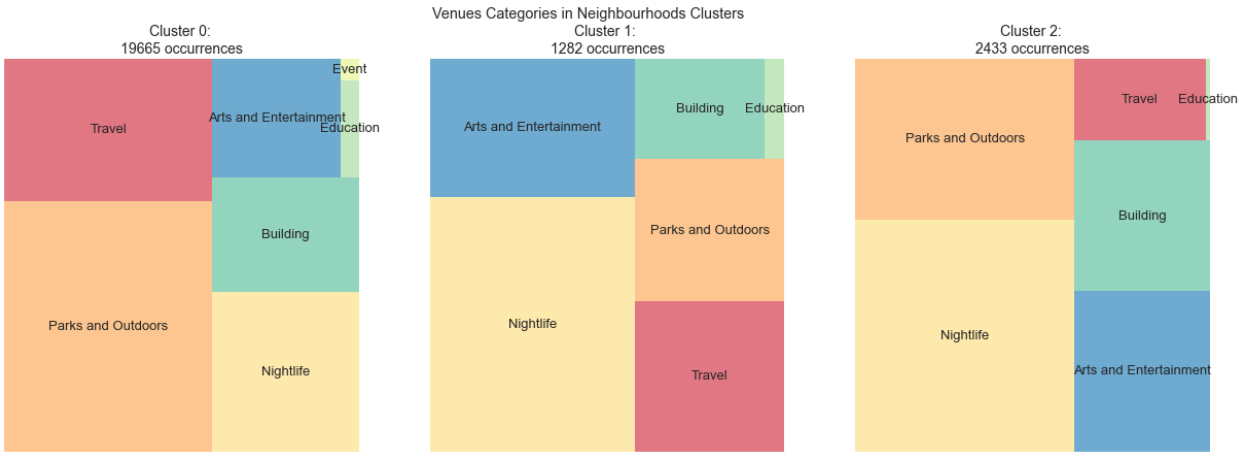

Figure 7: Most common categories of venues in the clustered Toronto Neighbourhoods.

It appears from the figure 6 that outdoor places like parks are prone to car robbery, as the premise type data on figure 8 indicated. On the other hand, nightlife types of venues tend to reduce the number of occurrences.

This page shows the venues subcategories, to give a better understanding.

It is worth noting that Food and Shops were purposely discarded from the analysis because they were common to every cluster, being the first and second most common venue category, respectively.

Conclusion

The analysis indicated that the most dangerous areas for car owners are those with open spaces such as parks, parking lots and hotels, while areas with active nightlife and arts/entertainment venues tend to have lower number of occurrences. Thus, this logic could be considered when defining the price of ca insurance.