Analysis of Top Restaurants and Auto Thefts in Toronto, Canada

Jeff Slee

February 21 2021

1.0 Introduction

When travelling to new cities it is helpful to know where the good restaurants are located. In addition, it can be useful to have an understanding of the crime in the area, so to avoid parking in those neighbourhoods.

The goal of this assignment is to find a top rated restaurant in a Toronto neighbourhood with a low number of auto thefts. Neighbourhoods will be divided into three clusters based on common auto theft characteristics. Then each restaurant will be scanned within a half-kilometer radius for auto thefts in the area. This will ensure to consider the surrounding neighbourhood in case there is a need to park slightly further away.

This problem would be of interest to tourists who are new in the city and would like to find a good place to eat while ensuring to avoid high crime areas. Tourist sites may also be interested as they can generate recommendations to indicate crime levels. Finally, potential restaurant owners may also want to find low crime neighbourhoods when researching where to open a new restaurant.

2.0 Data

A list of Toronto boroughs, neighbourhoods and postal codes will be web scrapped from Wikipedia. These postal codes will then be converted into geographical information (Latitude and Longitude). This process is called Geocoding, which is the computational process

of transforming a physical address description to a location on the Earth's surface. Geocoder will be the Python library used to do this. It is a simple and consistent geocoding library that can deal with multiple different geocoding providers. The provider will be ArcGIS World Geocoding Service. It finds addresses and places in all supported countries from a single endpoint. The geographical information returned from the Wikipedia postal codes will then be compared with the centroids and neighbourhood boundaries made available by the City of Toronto[1].

Foursquare City Guide, commonly known as Foursquare, will be used to determine the locations of all the top rated restaurants in Toronto. Foursquare is a technology company that has built a large dataset of location data that is currently the most comprehensive available. Many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, including over 100,000 developers, use it for its accuracy.

Two datasets made available by the Toronto Police Services containing auto theft locations and information on neighbourhood crime rates will be examined. The first dataset is called the Auto Theft dataset[2], which is a subset of the Major Crime Indicators (MCI) dataset. The dataset contains the closest intersection to where each auto theft occurred, between 2014 and 2019, as well as the neighbourhood, the time and date of the theft occurrence and many other attributes. The second dataset that will be used is the Neighbourhood Crime Rates Boundary File[3]. This dataset contains various crime statistics for each neighbourhood in Toronto.

3.0 Methodology

The following section will outline the steps taken in processing the data. The first part discusses how the neighbourhood locations are determined, then it reviews the auto theft datasets, clustering the

[1] https://open.toronto.ca/dataset/neighbourhoods/
[2] https://data.torontopolice.on.ca/datasets/auto-theft-2014-to-2019
[3] https://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-

neighbourhoods, then it filters for all restaurants in the venue dataset. The last part explains how the top restaurants in low risk neighbourhoods are selected.

3.1 Neighbourhood Locations

A list of Toronto postal codes have been web scrapped from Wikipedia, parsing the HTML using BeautifulSoup. This returned 180 unique rows of postal codes, however Wikipedia had several boroughs and neighbourhoods that were unassigned.

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Dropping the unassigned neighbourhoods and removing the Borough column left 103 rows of unique postal codes. The postal codes were then passed through ArcGIS using Geocoder and then merged back into the postal code dataframe.

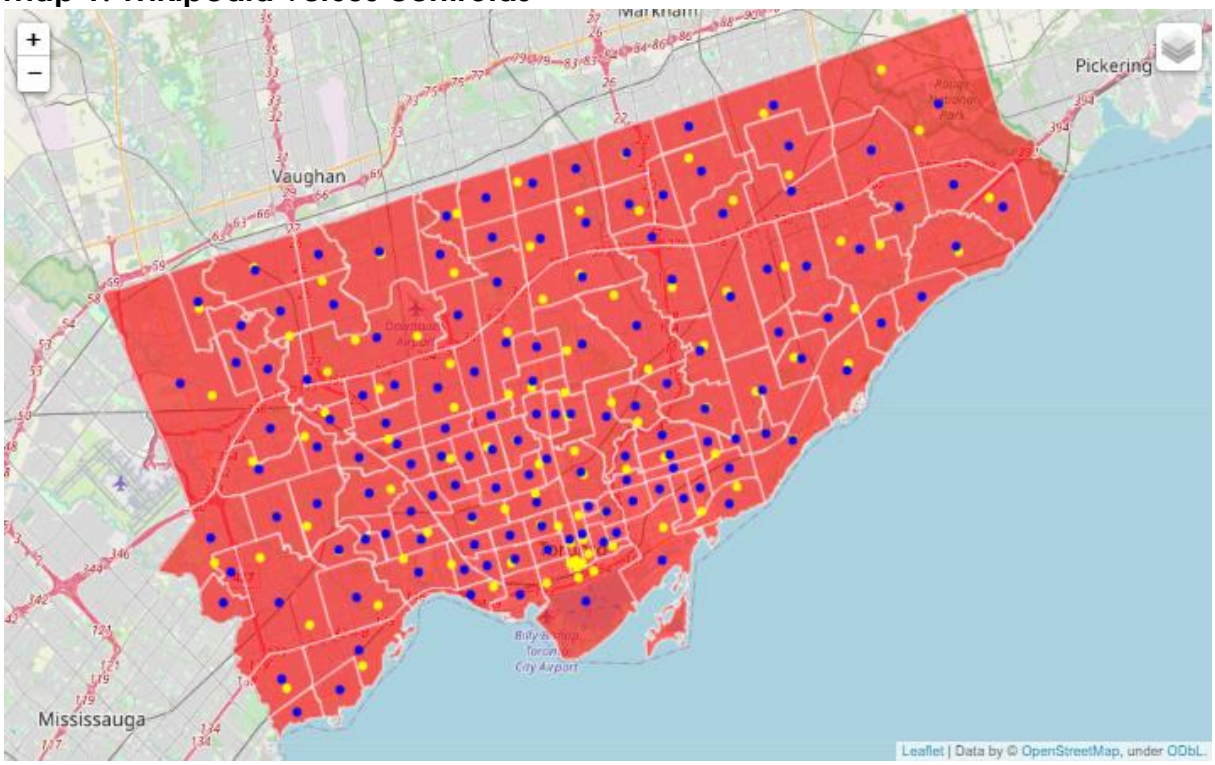
	Postal_Code	Neighbourhood	Latitude	Longitude
0	M3A	Parkwoods	43.75245	-79.32991
1	M4A	Victoria Village	43.73057	-79.31306
2	M5A	Regent Park, Harbourfront	43.65512	-79.36264
3	M6A	Lawrence Manor, Lawrence Heights	43.72327	-79.45042
4	M7A	Queen's Park, Ontario Provincial Government	43.66253	-79.39188

There are 103 postal codes in 99 neighbourhoods available on Wikipedia. However, according to the City of Toronto there are 140 neighbourhoods[4]. Due to this discrepancy, a comparison was made using the geo-coordinates from Wikipedia and the centroids available from the actual boundary file provided by the City of Toronto. The below map indicates that the Wikipedia points in yellow do not appear to be accurate when compared with the centroids

[4] https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

marked in blue. The centroids will instead be used to mark the neighbourhoods going forward.

Map 1. Wikipedia versus centroids



3.2 Auto Thefts

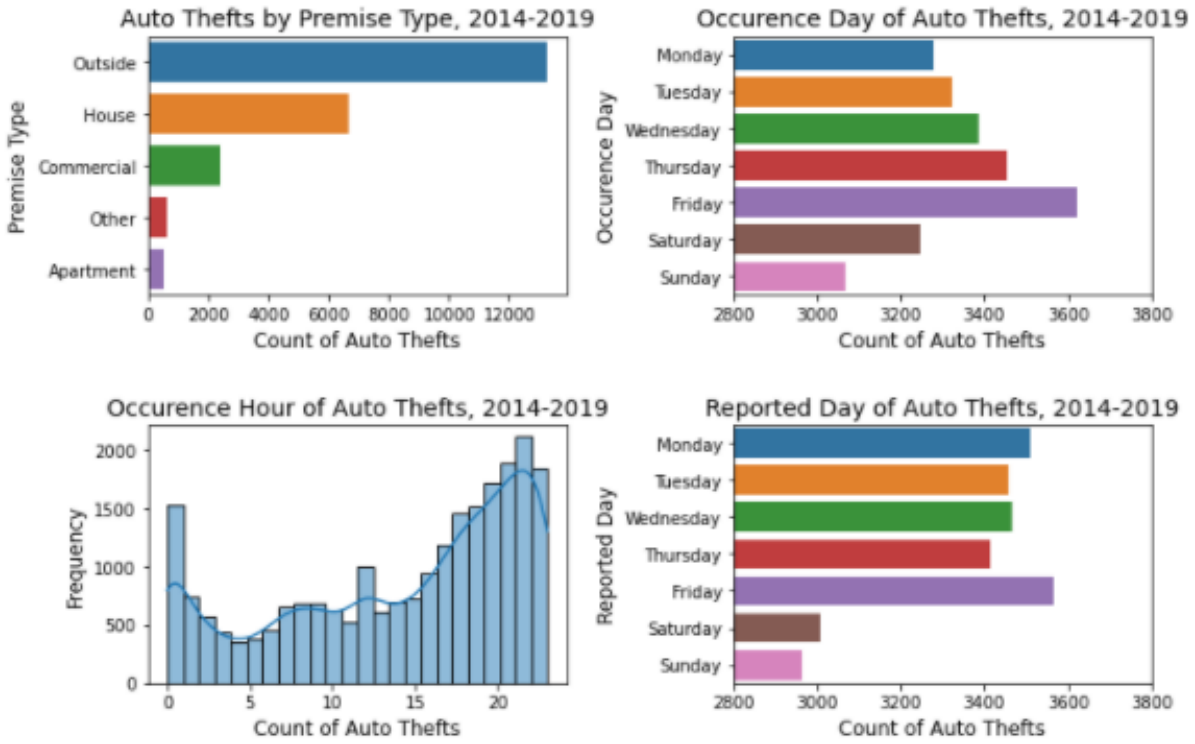
The next step is to load the auto theft data that is made publicly available by the Toronto Police. A total auto theft count per location is calculated based on the number of unique longitude and latitude codes. There are a total of 23,380 auto thefts that occurred during 2014 to 2019 across 9,016 unique locations.

On inspection, there are 315 auto theft locations that are recorded within different neighbourhoods. This could be possible as the auto theft locations reported may fall on the neighbourhood boundaries. For this analysis, the duplicates will be reclassified to be based on the highest frequency of the neighbourhoods reported for each location. Where there are duplicate neighbourhoods reported with only single cases, then the first neighbourhood in the list will be chosen. This process will ensure to assign a single neighbourhood to

each auto theft location. There are also three cases where the occurrence day is missing. The highest frequency day is used as an estimate.

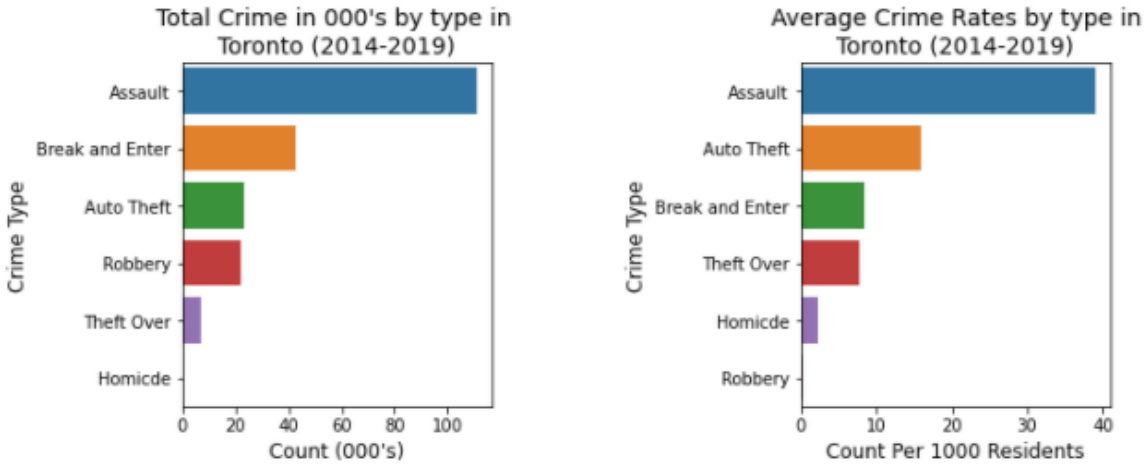
The majority of auto thefts in Toronto occur on Friday evenings, peaking at around 10:00 PM, and occur away from a person's place of residence. It is interesting to see a steady increase in auto thefts as the evening progresses and then starts to decline in the early mornings.

Figure 1. Total Auto Thefts in Toronto between 2014 to 2019



The second dataset provides neighbourhood crime statistics for each year between 2014 and 2019. The total number of crimes for each type are aggregated overall years. The most common crime are assaults. When looking at the average crime rates per 1000 residents, then auto thefts come up to second highest.

Figure 2. Total Crime and Crime Rates in Toronto between 2014 to 2019



The below figure provides the crime rates per 1000 residents in each neighbourhood. The total number of crimes by category are aggregated across all years. West Humber-Clairville neighbourhood has the highest rate of auto thefts. Bay Street Corridor has the highest assault and theft rates, University and Kensington-Chinatown are high on break and enter rates, Regent Park and Moss Park are top for homicide rates, and Moss Park has high robbery rates.

Figure 3. Total Crime Crime Rates by Neighbourhood between 2014 to 2019

3.3 Clustering

Using selected features of auto thefts for each neighbourhood, the neighbourhoods will be clustered into three groups using K-Means clustering, an unsupervised machine learning algorithm. K-Means is a type of partitioning clustering that divides the data into "k" non-overlapping subsets based on common characteristics. The best "k" is determined using the elbow method.

The first step is to build the dataframe in a way that is required for clustering. The following features are selected, based on each auto theft location and grouped by neighbourhood: the premise type, occurrence day, and occurrence hour. The occurrence hour is converted into four ranges: early morning, morning, afternoon and evening. All categorical features are converted into indicator variables and then the counts are aggregated by neighbourhood, giving the number of auto thefts by feature. The total number of auto thefts by neighbourhood is also added to the dataframe.

Neighbourhood	Apartment	Commercial	House	Other	Outside	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday	Afternoon
Agincourt North (129)	3.0	10.0	94.0	5.0	66.0	26.0	28.0	25.0	21.0	30.0	21.0	27.0	30.0
Agincourt South-Malvern West (128)	3.0	45.0	42.0	9.0	121.0	28.0	37.0	30.0	18.0	30.0	41.0	36.0	60.0
Alderwood (20)	0.0	14.0	22.0	3.0	58.0	13.0	12.0	21.0	15.0	13.0	13.0	10.0	21.0
Annex (95)	2.0	23.0	15.0	1.0	91.0	15.0	19.0	17.0	17.0	23.0	18.0	23.0	38.0
Banbury-Don Mills (42)	3.0	18.0	38.0	5.0	67.0	19.0	19.0	23.0	10.0	20.0	15.0	25.0	45.0

The features are then centred and scaled based on their mean and standard deviation. Standardization is a common requirement for many machine learning estimators.

```
# feature scaling
X = df_rest_crime_final.values[:,1:-2] # excludes neighbourhoods & locations
X = np.nan_to_num(X)
scaled_features = StandardScaler().fit_transform(X)
scaled_features
```

```
array([[-0.18727774, -0.18341701,  0.92849548, ...,  0.19157175,
        -0.11062765,  0.05249939],
       [-0.18727774,  0.73931743, -0.10950315, ..., -0.09666087,
         0.78204502,  0.25295162],
       [-0.97384425, -0.07796164, -0.5087334 , ..., -0.42812838,
        -0.28916218, -0.33408704],
```

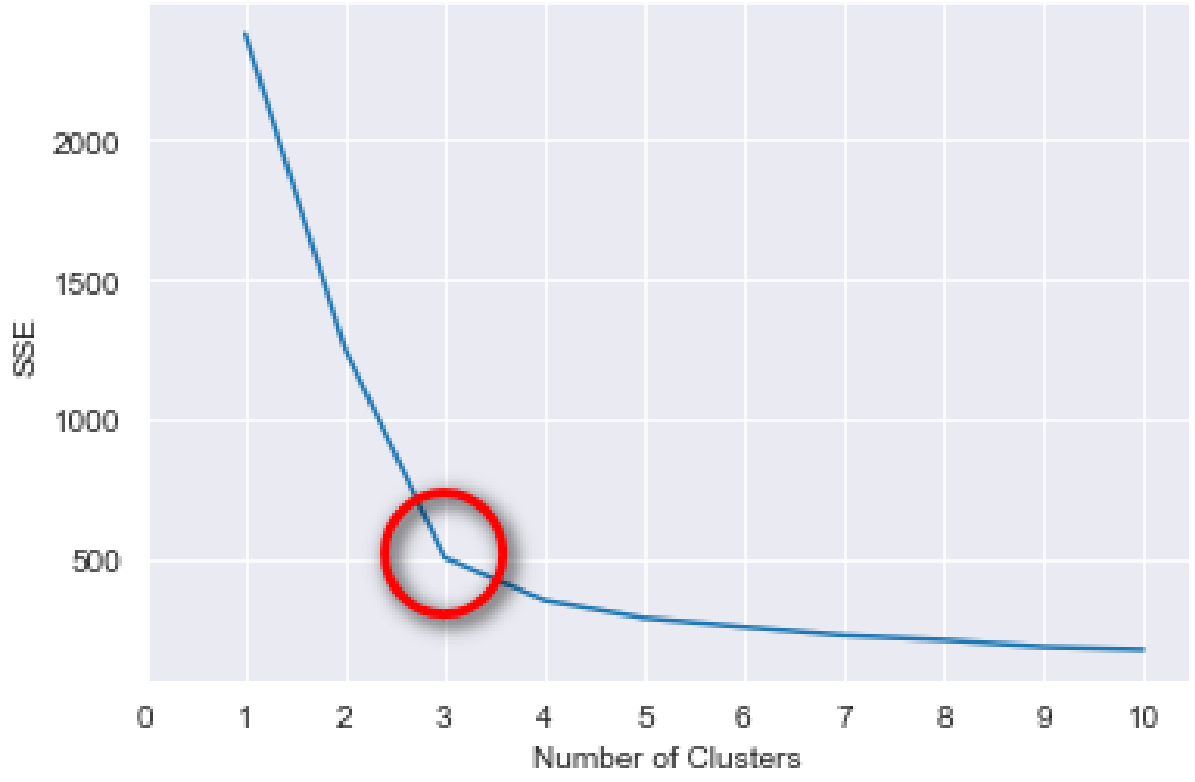
K-Means clustering is an iterative process. The elbow method is used to help determine the best "k". This is done by running a range of different "k" values and storing the intertia. The intertia is the sum of squared errors (SSE) of each data point to its closest cluster centre. If all data points are tightly congregated around their allocated centroid, then the SSE will be low — otherwise, it will be high.

```
kmeans_kwargs = {
                 "init": "random",
                 "n_init": 10,
                 "max_iter": 300,
                 "random_state": 42,
                   }

# A list holds the SSE values for each k
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, **kmeans_kwargs)
    kmeans.fit(scaled_features)
    sse.append(kmeans.inertia_)
```

The best "k" is selected at the "elbow" point, after which the inertia starts decreasing in a linear fashion. The below graph provides a visual of the generated SSE and the cluster number. The best "k" is determined to be based on three clusters.

Figure 4. The Elbow Method using Inertia



This can also be verified using the KneeLocator function in Python.

```
In [170]:
# Auto detect best number of clusters
kl = KneeLocator(
 range(1, 11), sse, curve="convex", direction="decreasing"
)

kclusters = kl.elbow
kclusters
```

```
Out[170]: 3
```

The cluster labels are then inserted back into the original dataframe and the average number of auto thefts are generated by feature.

Table 1. Statistics of crimes by cluster and feature

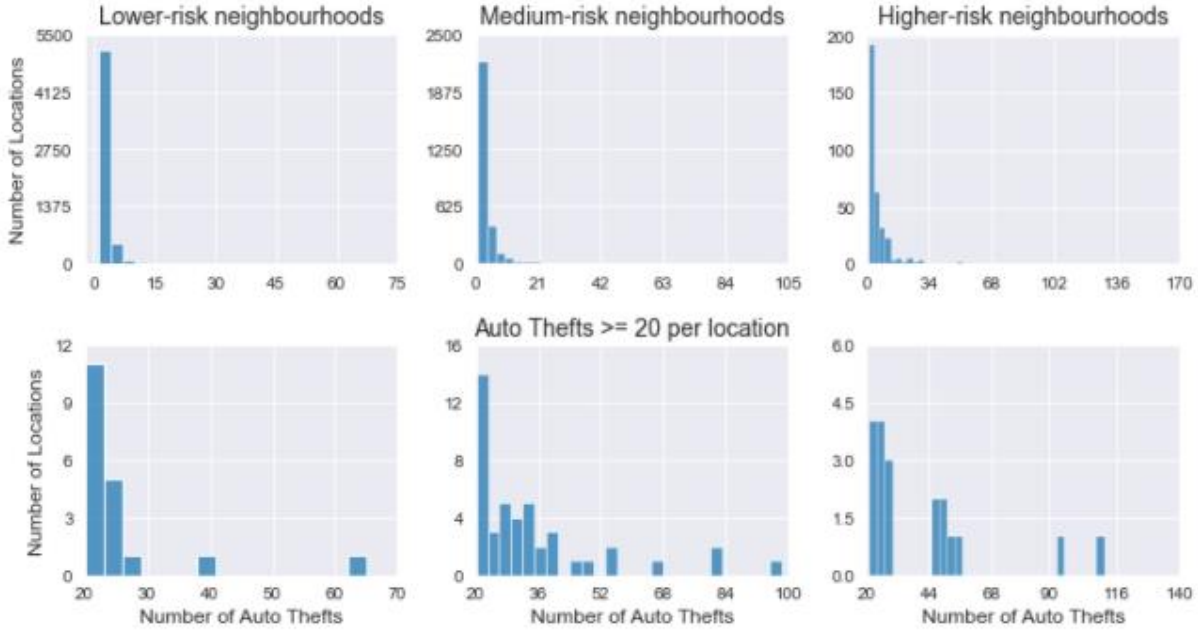
	Low-Risk	Medium-Risk	High-Risk
Apartment	3.0	8.0	12.0
Commercial	7.0	45.0	362.0
House	34.0	92.0	405.0
Other	3.0	9.0	50.0
Outside	59.0	197.0	1415.0
Friday	16.0	55.0	375.0
Monday	15.0	48.0	300.0
Saturday	14.0	50.0	352.0
Sunday	14.0	46.0	265.0
Thursday	16.0	50.0	305.0
Tuesday	15.0	51.0	309.0
Wednesday	15.0	51.0	338.0
Afternoon	25.0	87.0	551.0
Early Morning	20.0	67.0	503.0
Evening	44.0	130.0	725.0
Morning	18.0	67.0	465.0
Auto Thefts	106.0	351.0	2244.0
Neighbourhoods (#)	113.0	26.0	1.0
Total Thefts (#)	12010.0	9126.0	2244.0
Unique Locations (#)	5815.0	2863.0	338.0
Auto-theft Density	2.1	3.2	6.6

It is clear from the above table that the first cluster, containing 113 neighbourhoods, has a lower number of auto thefts on average across the various features than the other clusters. The clusters are appropriately renamed as Low, Medium and High Risk clusters accordingly. The high-risk cluster is based on a single neighbourhood and will be looked at closer on its own later.

There are a total of 12,010 auto thefts in low-risk neighbourhoods, across 5,815 unique auto theft locations, with an auto theft density of 2.1 thefts per location. The medium-risk neighbourhoods have a total of 9,126 auto thefts across 2,863 unique locations, with an auto theft density of 3.2 thefts per location. Finally, the high-risk cluster has 2,244 auto thefts in 338 locations, with an auto theft density of 6.6 thefts per location.

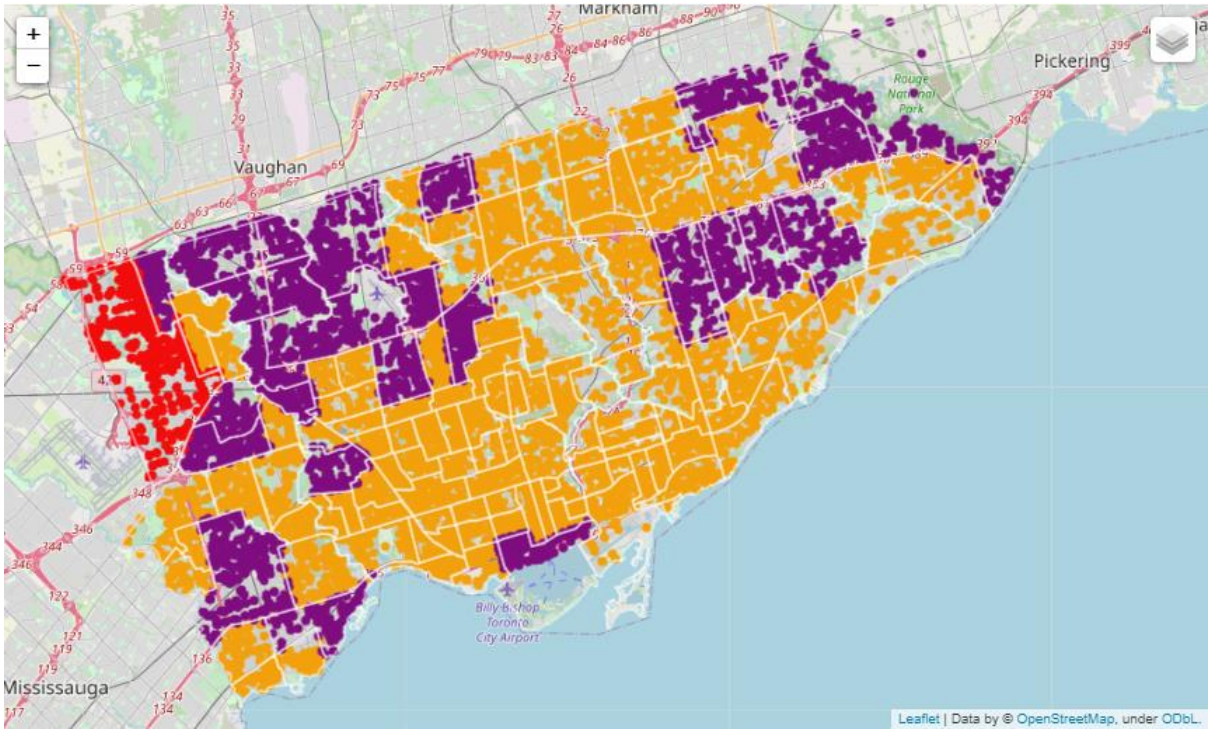
The majority of locations contained only a small number of auto thefts across the different risk categories (Figure 4). Filtering for the locations containing 20 or more auto thefts, there is evidence of auto theft "hot zones" within the various risk groups. A hot zone would be locations that are more frequently targeted. For example, within the medium-risk neighbourhoods there are around 15 different locations containing approximately 20 auto thefts and a lower frequency of many other locations containing well above 20 or more auto thefts.

Figure 5. Frequency of Auto Thefts by Neighbourhood Risk



The following is a map showing the auto theft locations based on the clustering results. Auto thefts in low-risk neighbourhoods are in yellow, medium-risk is in purple and high-risk is in red.

Map 2. Neighbourhood Clusters based on Auto Theft Characteristics



From the clustering above, only the low-risk neighbourhoods will be used to filter for restaurants in Toronto.

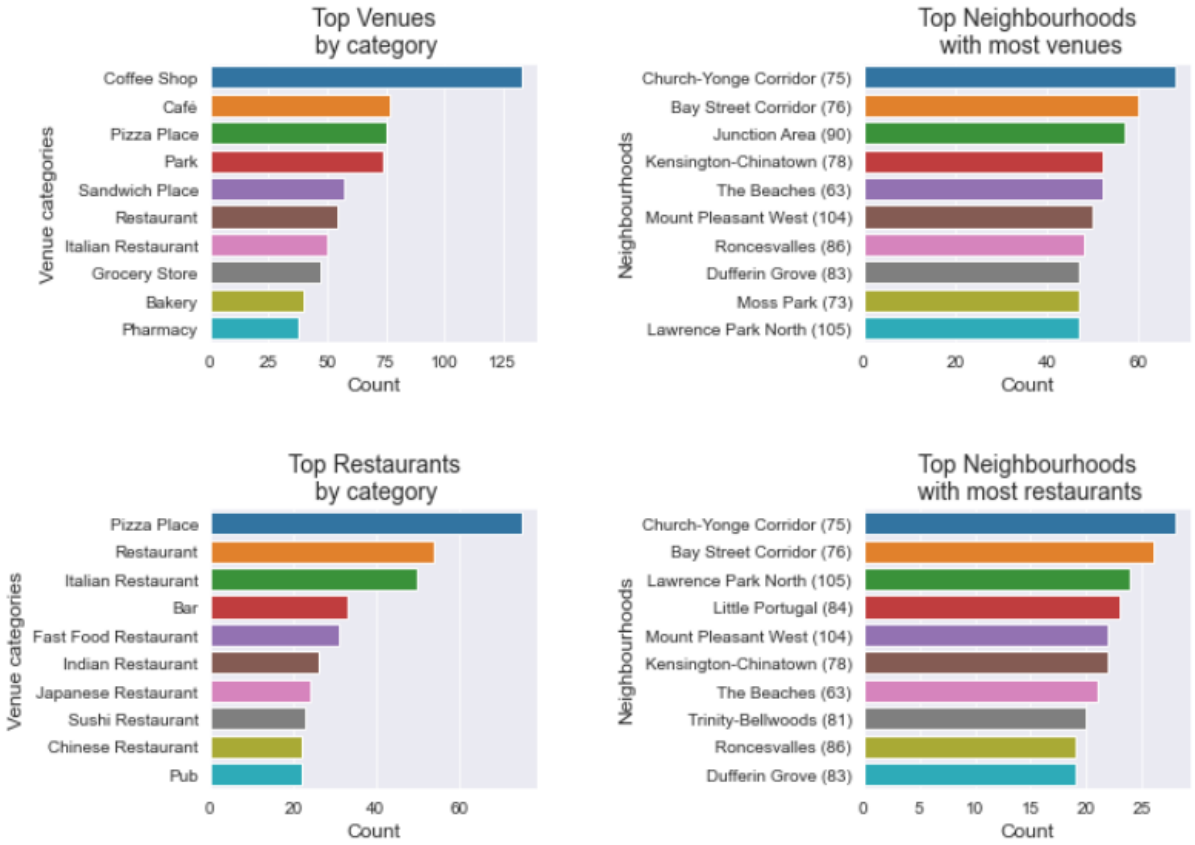
The next step is to determine the top rated venues and find the number of restaurants located in the lower risk neighbourhoods.

3.4 Top-rated venues and restaurants

The neighbourhood locations are used to return the top 100 highly rated venues within a 500 radius around each neighbourhood in Toronto. This was done by using a Foursquare API connection, and returned a total of 1,536 unique venues across 1,923 different locations in Toronto.

There are over 288 unique venue categories that need to be further filtered in order to identify all restaurants in Toronto. The following key words have been used to filter each venue's category description: "Restaurant", "Pizza", "Burger", "Chicken", "BBQ", "Fish", "Steak", "Taco", "Burrito", "Bar", "Wings", "Buffet", and "Diner". This returned 600 unique restaurants.

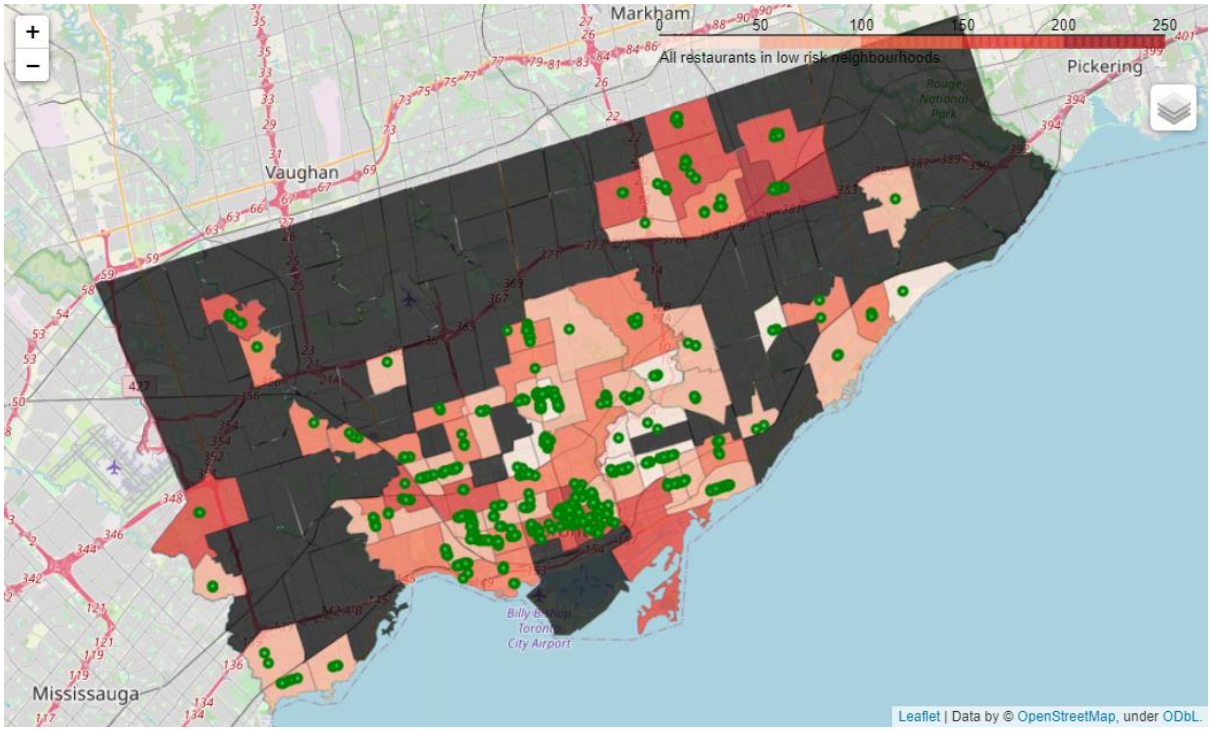
Graph 6. Top 10 Venues & Restaurants



The most common venues in Toronto are coffee shops and cafes, while the most common restaurants are pizza places. The Church-Young Corridor has the highest number of venues, including the highest number of restaurants.

The next step is to identify which restaurants are located in low risk neighbourhoods. This was done using geopandas to check if each venue's location falls within in each neighbourhood's geometric area. Based on this methodology there are 631 restaurants, shown in green, that fall into 80 low risk neighbourhoods.

Map 3. Restaurants located in low risk neighbourhoods



3.4 Calculating distance and determining low risk restaurants

Now that we have the low risk neighbourhoods and have identified all the restaurants then the next step is to determine how far each auto theft location is from each restaurant. The Haversine formula is used to calculate the distance as it is quite simple to set up and works well with geopandas. The Haversine formula calculates the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface. The formula is:

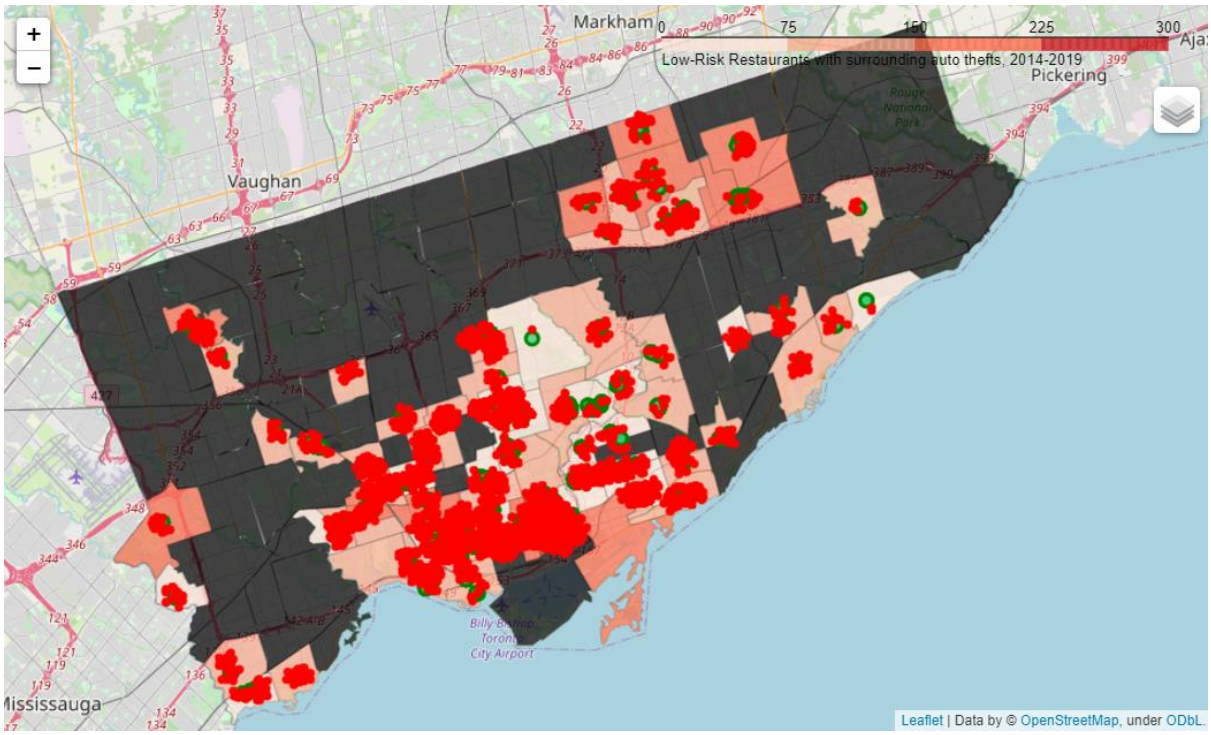
$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

After determining the distances between all restaurants that are located in low-risk neighbourhoods and all car theft locations, then only the distances that are a 1/2 KM or closer to the restaurant will be kept.

It is important to distinguish that the auto theft locations are not exclusive to being located in low-risk neighbourhoods, but account

for all auto thefts within the 1/2 KM surrounding area. This is to consider cases where a restaurant is located near low-risk boundaries. The following map provides the restaurants located in low-risk neighbourhoods and the surrounding auto thefts.

Map 4. Restaurants in low risk neighbourhoods and surrounding auto thefts

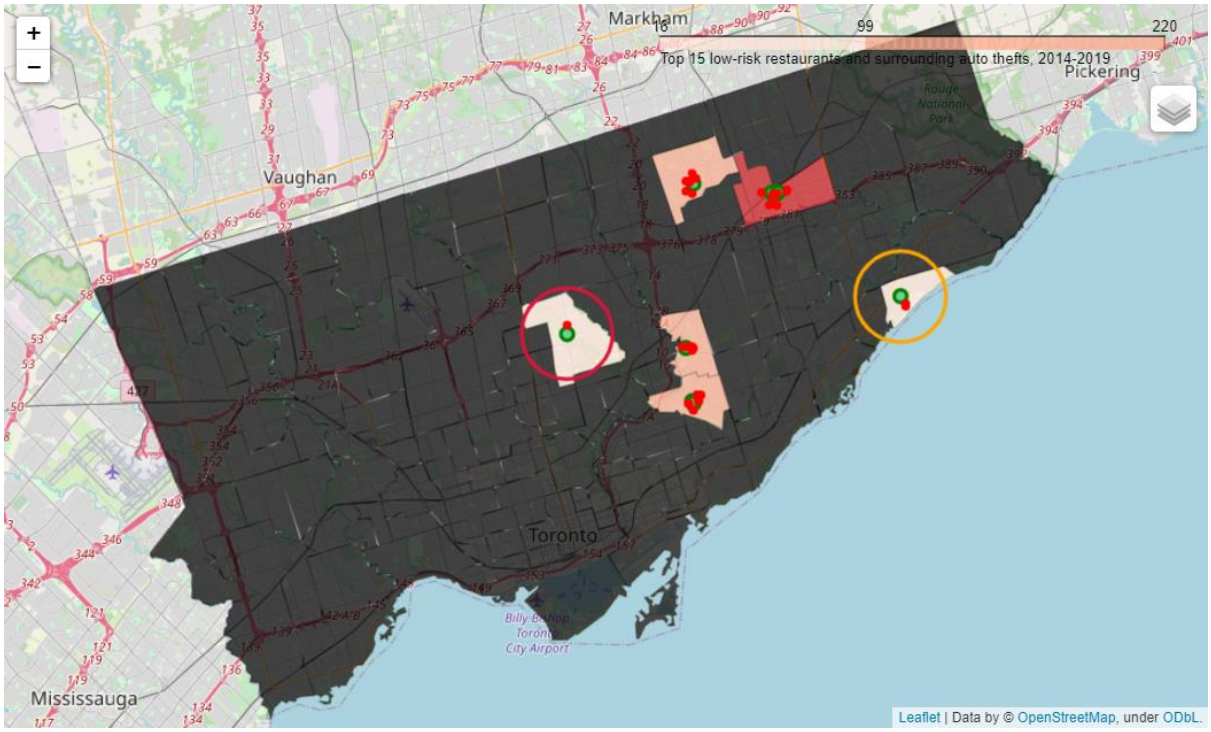


Now the top 15 restaurants with the lowest number of car thefts will be selected.

4.0 Results

The top 15 restaurants are mapped below. The Granite Club Dining Room, located in the Bridle Path-Sunnybrook-York Mills neighbourhood, is the top restaurant with only one auto theft within a half-kilometer radius (red circle). The Granite Club is a private social and athletic club, founded in 1875. The initial membership fee is $53,000 per couple!! It looks then like it will have to be pizza at the second top restaurant, Pizza Nova (yellow circle).

Map 5. Top 15 restaurants located in low risk neighbourhoods

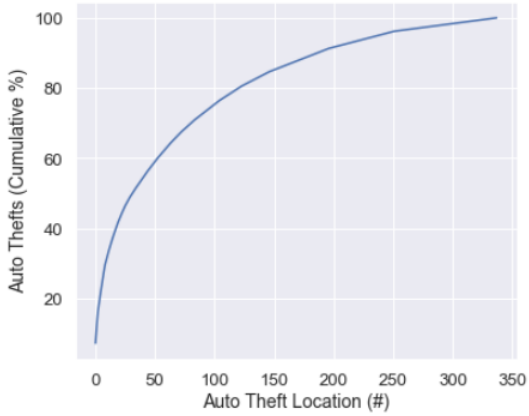


Name	Categories	Neighbourhood	Total Auto Thefts Around Restaurant	Number of Auto Theft Locations	Density	Average Distance	Minimum Distance
Granite Club Dining Room	Restaurant	Bridle Path-Sunnybrook-York Mills (41)	1	1	1.000000	0.380906	0.380906
Pizza Nova	Pizza Place	Guildwood (140)	2	2	1.000000	0.405166	0.362253
Yummy Cantonese Restaurant 老西関腸粉	Cantonese Restaurant	Agincourt South-Malvern West (128)	5	4	1.250000	0.088523	0.050876
Wonton Chai Noodle 雲吞仔	Noodle House	Agincourt South-Malvern West (128)	5	4	1.250000	0.098064	0.057525
Mike's BBQ 丰記燒臘	BBQ Joint	Agincourt South-Malvern West (128)	5	4	1.250000	0.110292	0.066128
Pizza Pizza	Pizza Place	O'Connor-Parkview (54)	6	5	1.200000	0.303235	0.231863
Venice Pizza	Pizza Place	O'Connor-Parkview (54)	6	5	1.200000	0.312742	0.196971
Jawny Bakers	Gastropub	O'Connor-Parkview (54)	7	6	1.166667	0.316291	0.235269
Congee Me 小米粥鋪	Chinese Restaurant	Agincourt South-Malvern West (128)	8	5	1.600000	0.169898	0.037581
Jesse Jr. (Filipino Foods & Restaurant)	Restaurant	Agincourt South-Malvern West (128)	8	5	1.600000	0.167416	0.024564
Happy Lamb Hot Pot	Hotpot Restaurant	L'Amoreaux (117)	9	7	1.285714	0.329652	0.147333
Shanghai Dim Sum	Chinese Restaurant	Agincourt South-Malvern West (128)	9	6	1.500000	0.203446	0.049918
Perfect Chinese Restaurant 雅瓊海鮮酒家	Chinese Restaurant	Agincourt South-Malvern West (128)	9	5	1.800000	0.193562	0.074730
The Frig	French Restaurant	Victoria Village (43)	10	6	1.666667	0.221276	0.065184
Asian Legend 味香村	Chinese Restaurant	Agincourt South-Malvern West (128)	11	6	1.833333	0.278344	0.112581

High-Risk Neighbourhoods

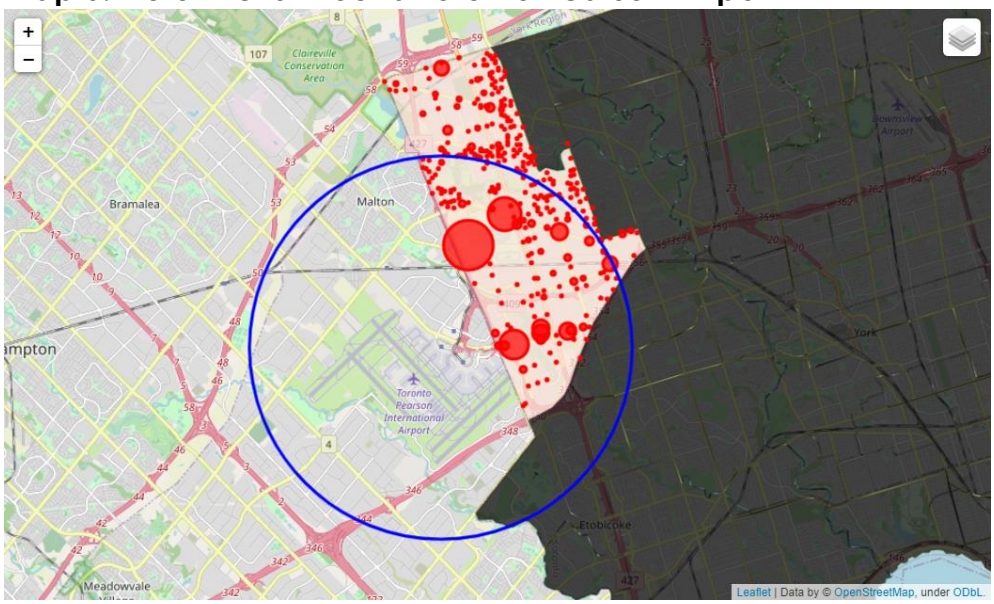
Returning briefly to the auto thefts located in the high-risk cluster. There are a total of 2,244 auto thefts located within a single neighbourhood, West Humber-Clairville. The auto thefts are spread across 338 locations and account for nearly 10% of the total auto thefts in Toronto between 2014 and 2019.

Figure 7. Cumulative Percent of High-Risk Auto Thefts



It is interesting to note that over 60% of the auto thefts in the high-risk neighbourhoods occur within 5KM from the departures gate of the Toronto Pearson International Airport. There is also evidence of "hot zones" near the airport, indicated by the larger circles. This area has a high concentration of parking lots, suggesting that travellers commuting from outside of Toronto may be unaware of the risks of parking near the airport, making an easy target for a career-minded criminal.

Map 6. Auto Thefts Around Toronto Pearson Airport



5.0 Conclusion

When travelling to Toronto it can be useful to have an idea of what areas have higher crime rates when looking to go out for a dinner at a nice restaurant.

Utilising auto theft data it is possible to be able to determine restaurants that are located in neighbourhoods with a low number of auto thefts in the surrounding area.

K-means clustering was used to group the locations of auto thefts into three clusters based on common characteristics. This generated 113 low-risk neighbourhoods, with over 600 restaurants to choose from.

Based on these results it can give visitors who are new to the city additional information on where they can go out for dinner and safely park their vehicle.