Covid-19, Short-term rent registrations & Venues Data Analysis of Toronto city

Souhail Berdaa

March 29, 2021

1. Introduction

1.1 Background

In Recent years, Covid-19 have changed the way we are looking at small businesses and had stroke hard the economy. During this period of uncertainty, it is getting more difficult to think about how much risky to start a small business. Scientists are trying to understand how the virus is behaving and what is impacting the Covid-19 cases numbers. Toronto is one of the most cities that have been impacted by the Virus in Canada. Analyzing Covid-19 data from different aspects using Neighborhoods, Venues with their latitude and longitudes data may help us cluster Toronto Neighborhoods to have clearer view of which type of venues are impacting the source of community infections for example. Combined with other insights we may propose a good neighborhood to start a small business in, depending on which type of business and the impact of that on the spreading of the virus.

1.2 Problem & Hypothesis

Small businesses are struggling to adapt with the frequent changes, every insight or story that we can get from data will help them find better solutions to adapt and stay in the market. Businesses may have to close if they have actives Covid-19 cases. That may lead them to lose lots of money. So, understanding the covid-19 situation around is very important to prevent that, the best way possible.

Hypothesis:

- Sources of community infections may be linked to higher number of some types of Venues (Businesses).
- Sources of Travel infections may be linked to higher numbers of short-term rent licenses.

2. Data acquisition and cleaning

2.1 Data sources/Webscraping

Covid-19 Dataset was collected from Toronto Opendata, postal codes and neighborhoods names was collected using Web-scraping Wikipedia page. Latitude, longitude and venues info around neighborhoods was collected using Foursquare API. Short term rental licenses dataset was collected from Toronto Opendata.

The data are extracted from the provincial Case & Contact Management System (CCM).

The Short-term rental regulation is based on Toronto Municipal Code Chapter 547, Licensing and Registration of Short-Term Rentals. A short-term rental is all or part of a dwelling unit rented out for less than 28 consecutive days in exchange for payment. This includes bed and breakfasts (B&Bs) but excludes hotels and motels. It also excludes other accommodations where there is no payment.

About COVID-19 Cases in Toronto:
Toronto Public Health is responding to an ongoing COVID-19 outbreak, in the context of an evolving global pandemic. This data set contains demographic, geographic, and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since the first case was reported in January 2020. This includes cases that are sporadic (occurring in the community) and outbreak-associated. The data are extracted from the provincial Case & Contact Management System (CCM).

Limitations:
The data in this spreadsheet are subject to change as public health investigations into reported cases and continuous quality improvement initiatives are ongoing, and additional cases continue to be reported. The data are extracted at 3 PM on the Monday of a given week, and posted by Wednesday. Please note that these numbers may differ from those posted elsewhere, as data are extracted at different times, and from different sources.

2.2 Data cleaning

Data downloaded or scraped from multiple data sources, and some rows had missing values or contained not assigned value. Missing values rows was removed, and not assigned value was replaced by neighborhoods names. Everything was combined in one data frame. Duplicated values were dropped.

2.3 feature selection

Features selected are: source of infections, covid cases numbers, source of infection_travel, source of infection community, venues, postal code, latitude, longitude, short-term rent

registrations. Some of them was selected based on the subject interests and others using correlation.

3. METHODOLOGY

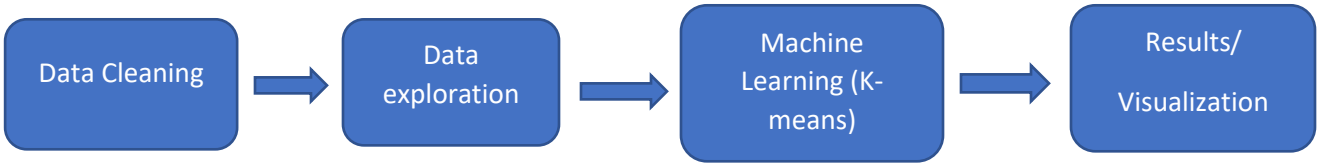


Figure 1: Project Workflow

3.1 Exploratory Data Analysis

3.1 Relationship between Source of infection and short-term rental registrations

We used the Pearson correlation as a measure of the extent of interdependence between variables. Correlation does not imply causation; causation may require independent experimentation.

Here below in figure 2, we can observe the distribution of short-term rentals registrations in Toronto neighborhoods. M5V have the highest number of short-term rentals registrations.

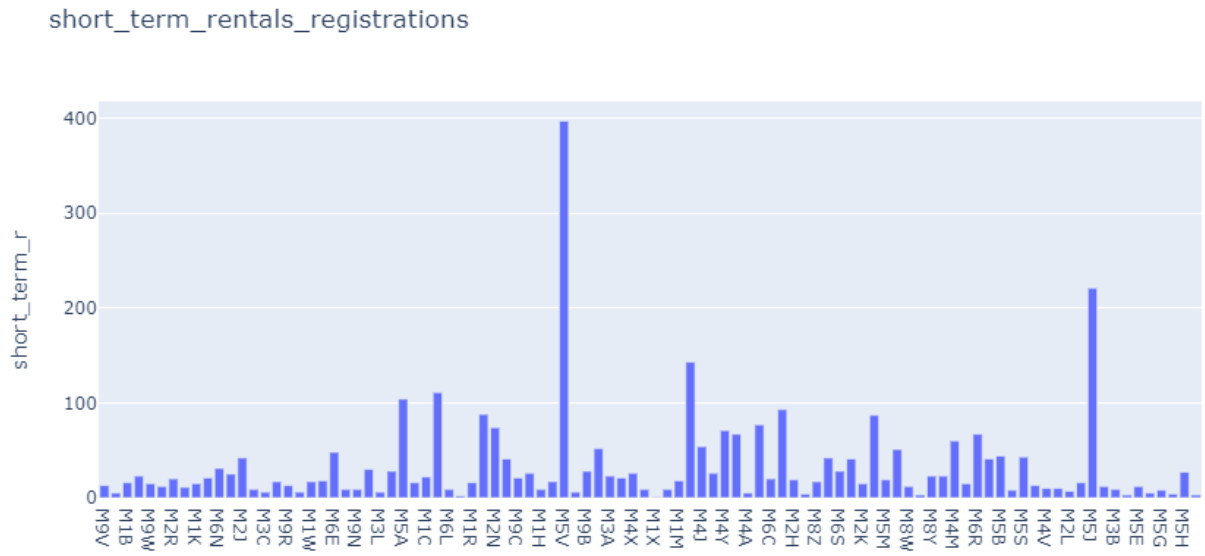


Figure 2: Short term rental registrations distribution in each Postal code

That's maybe explained by the fact that it's near the airport, this leads us to think that maybe there is a link between source of travel infections and short-term rentals registrations.

We can see in figure 3 that the bin with high number of short term rent registrations have travel related Covid-19 infections. That corresponds to M5V where we have high numbers of short term rent registrations, see figure 4 and figure 5.

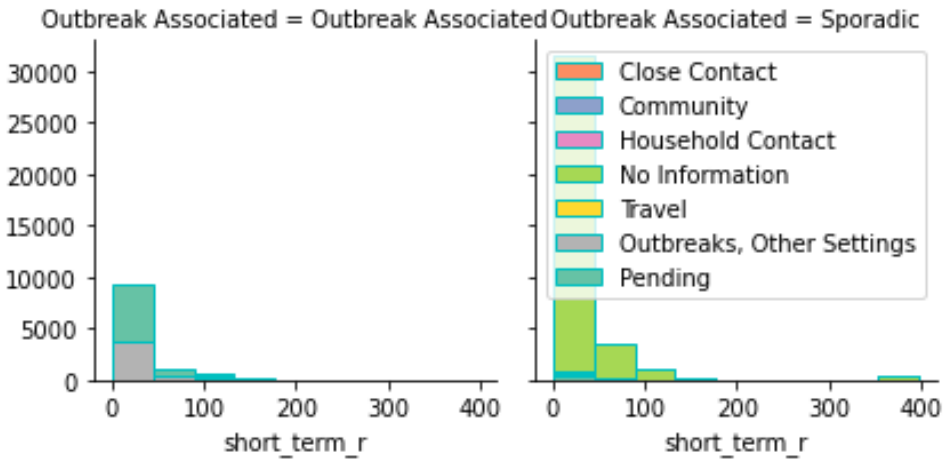


Figure 3: Outbreak associated for each short_term_r bin

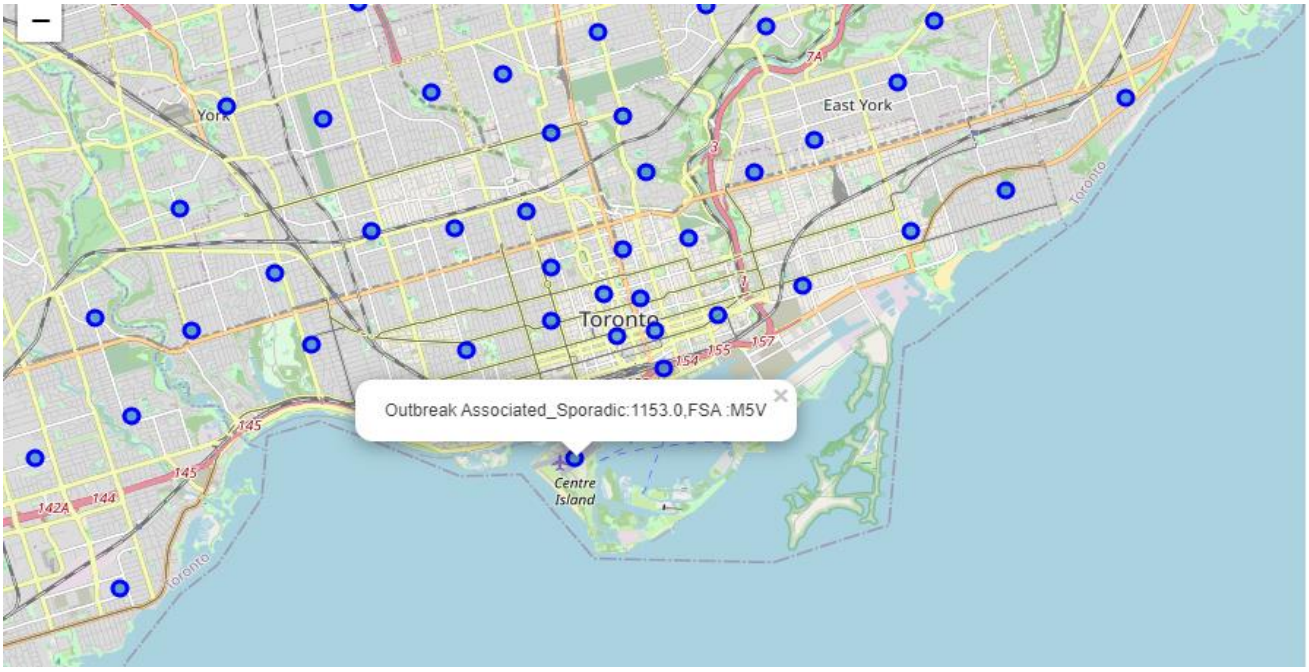


Figure 4: Map of Toronto city Neighborhoods

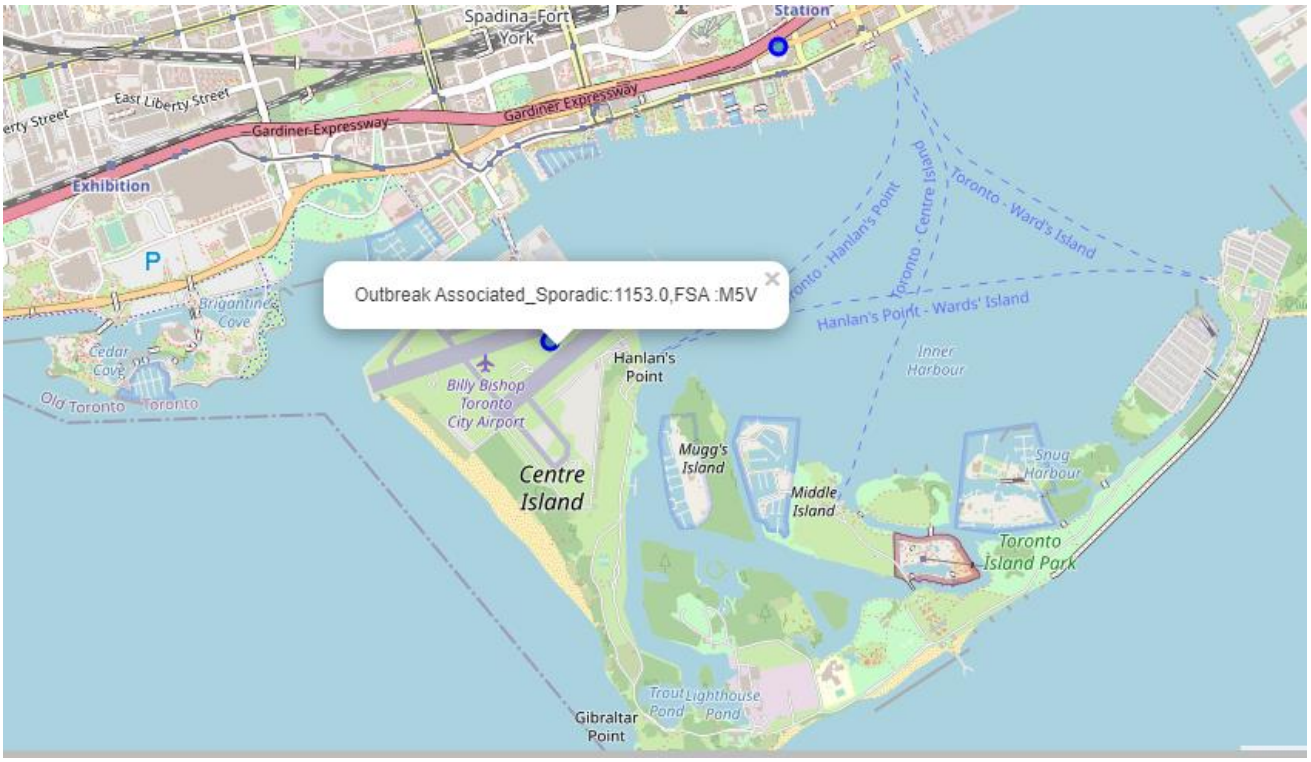


Figure 5: map of Toronto city showing M5V neighborhood

3.2 Relationship between source of travel infections and short-term rental registrations

The Pearson correlation coefficient is 0.43 between these two variables with a p- value less than 0.001.

Conclusion:

Since the p-value is <0.001, the correlation between source of infections_travel and short-term rental registrations is statistically significant, but the linear relationship is only moderate (~0.43). see figure 6.

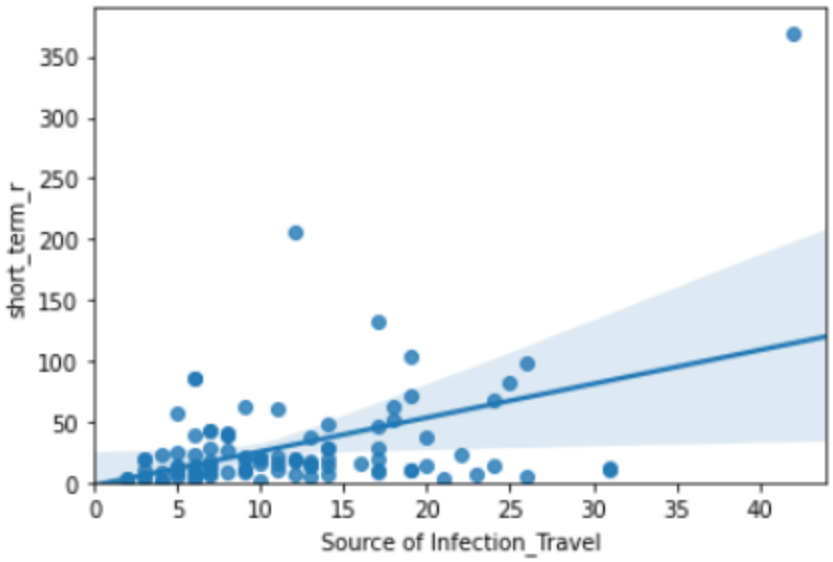


Figure 6 : Regplot showing linear correlation between short_term_r and source of infection_travel

That could be explained by the fact that travelers are staying next to the airport for a short period of less than 28 days.

3.3 one-hot encoding

We used indicator variables or dummy variables to be able to use our categorical data to explore relationship between data such as the correlation between features.

	FSA	Source of Infection_Close Contact	Source of Infection_Community	Source of Infection_Household Contact	Source of Infection_No Information	Source of Infection_Outbreaks, Congregate Settings	Source of Infection_Outbreaks, Healthcare Institutions	Source of Infection_Outbreaks, Other Settings
0	M9V	0	0	0	0	0	1	0
1	M9V	1	0	0	0	0	0	0
2	M9V	0	1	0	0	0	0	0
3	M9V	0	1	0	0	0	0	0
4	M9V	0	1	0	0	0	0	0

	FSA	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athl & Sp
0	M1B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	M1C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	M1E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	M1G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	M1H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

3.4 Relationship between source of community infections and most Venues

In Figure 7, we can observe the distribution of covid-19 cases in Toronto city neighborhood, we can see that M9V have the highest number of covid-19 cases

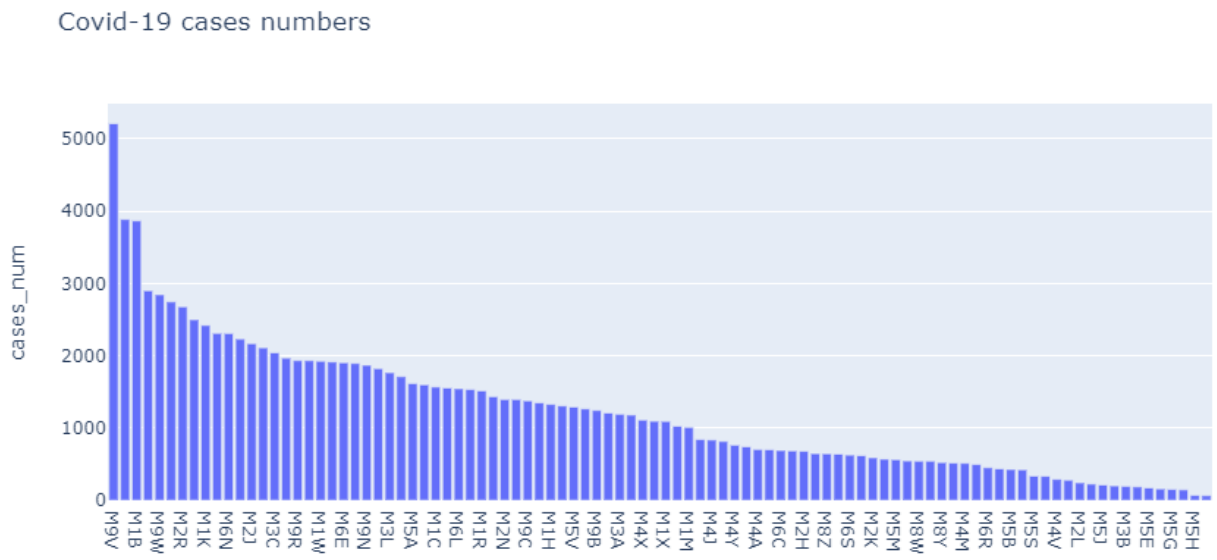


Figure 7: distribution of covid-19 cases in each neighborhood

In figure 8 below, we can observe that we have the third wave kicking in. The number of cases is going up in Toronto, and it's getting higher than the second wave. It seems like there is more propagation of the virus and that may be related to the new Covid-19 variants of concern.

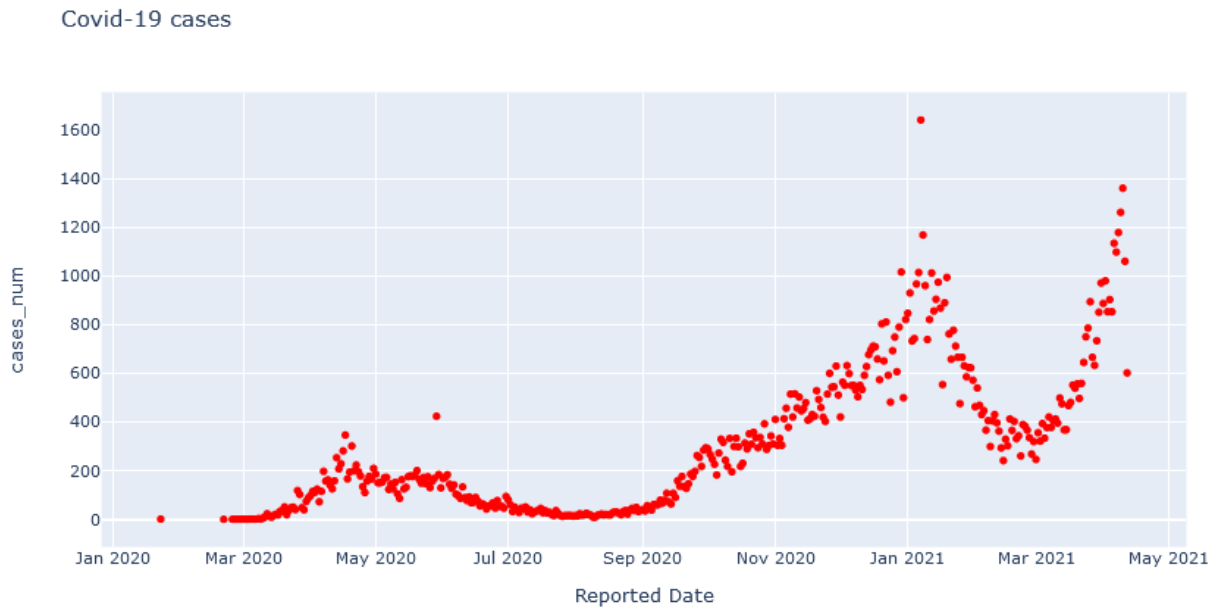


Figure 8: Scatter plot showing the evolution of Covid-19 cases number in time

VOC: variant of concern

We can observe here in figure 9, that numbers of VOC confirmed or screened positive are increasing as of Covid-19 non-VOC are increasing too, and it looks like it represents a big portion of the total daily cases. That raises lot of concern about the spreading of the variants.

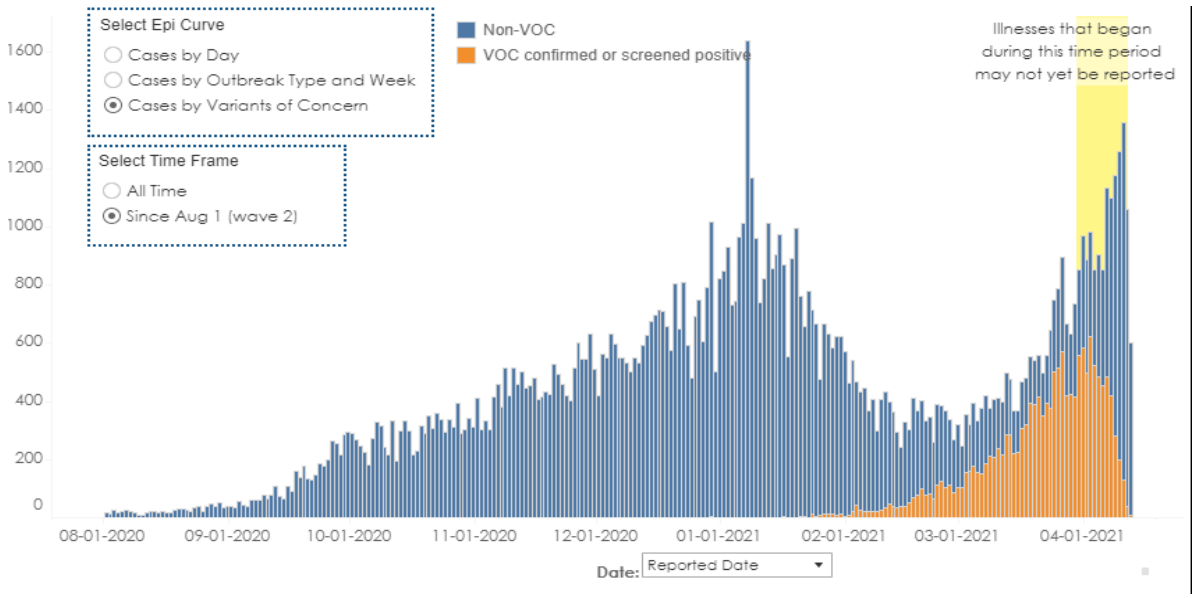


Figure 9: Evolution of variants of concern in time from Toronto open data visualization

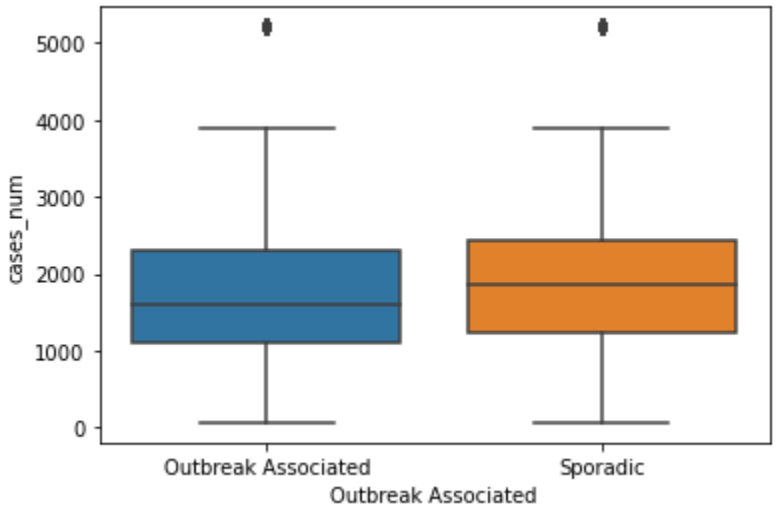
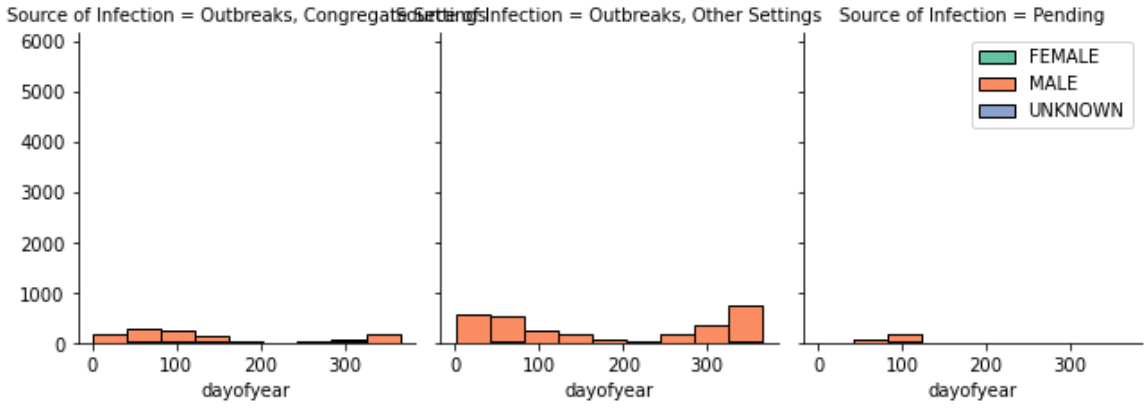


Figure 10: Boxplot of Outbreak associated Covid-19 cases numbers

Here we can observe in figure 11, the distribution of cases numbers in Toronto population genders, we can see that males represent the biggest portion of the population who have been confirmed positive to covid-19 during each period of time in every source of infection.



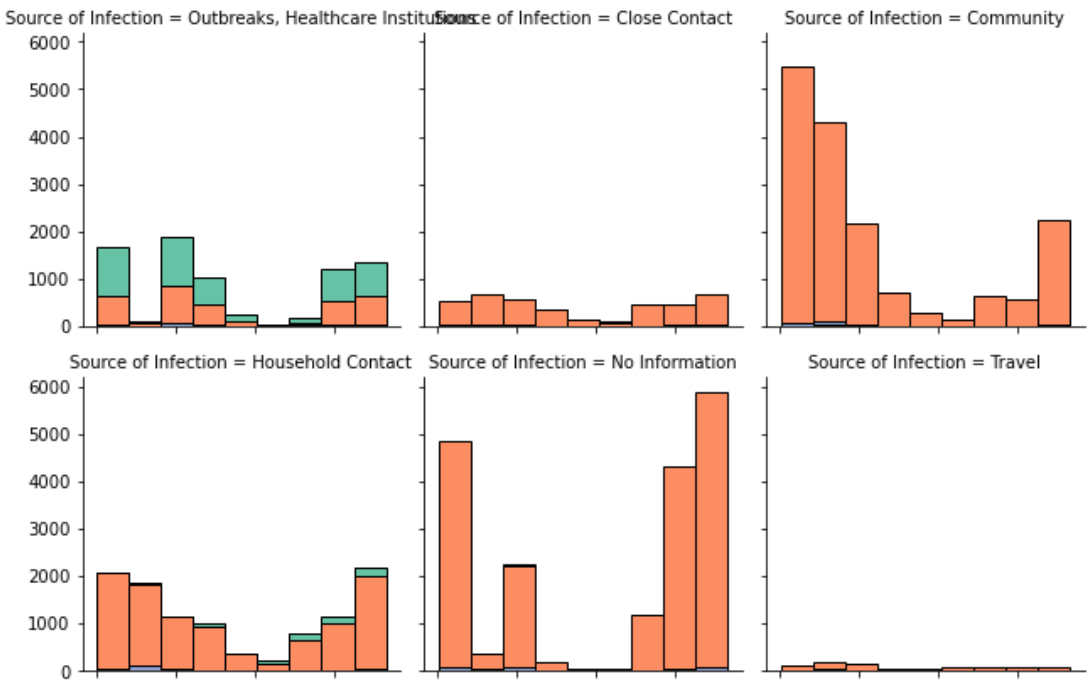


Figure 11: Client gender distribution of related source of infection Covid-19 cases in time

Here in figure 12, we observe that the biggest portion of affected people in sporadic outbreak associated is the young population 19 and younger and young adults 20 to 29 years. Which means this portion of population is more likely to spread to virus through community and this maybe explain the fact that we have a growing number of young adults hospitalizations.

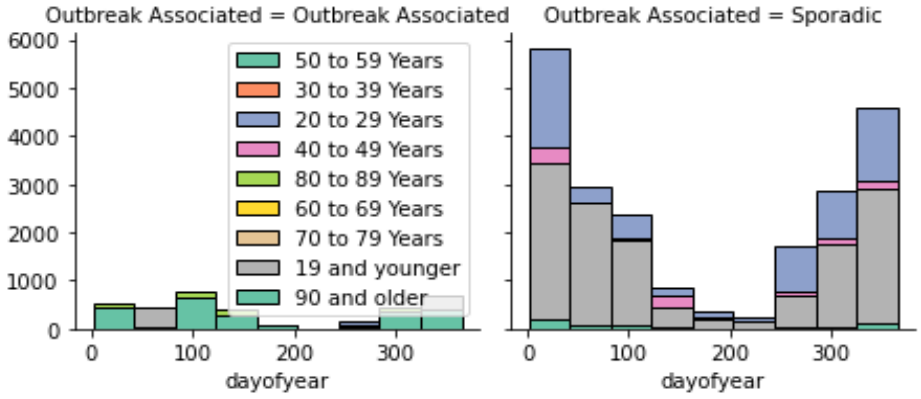


Figure 12: Age group distribution of related Outbreak associated Covid-19 cases in time

As shown below in figure 13, we see that only old people were hospitalized in the beginning of last year but we don't see any hospitalizations lately.

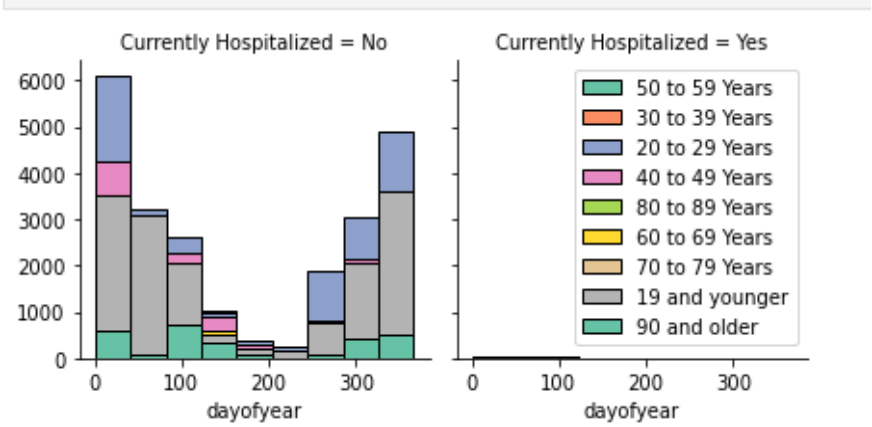


Figure 13: Age group distribution of related Currently Hospitalized Covid-19 cases in time

We can observe in figure 14, that close contact and household contact source of infections are mainly in the 19 and younger population, thus we see that source of infections related to healthcare institutions outbreak is mainly older population 50 to 59 years and 80 to 89 years.

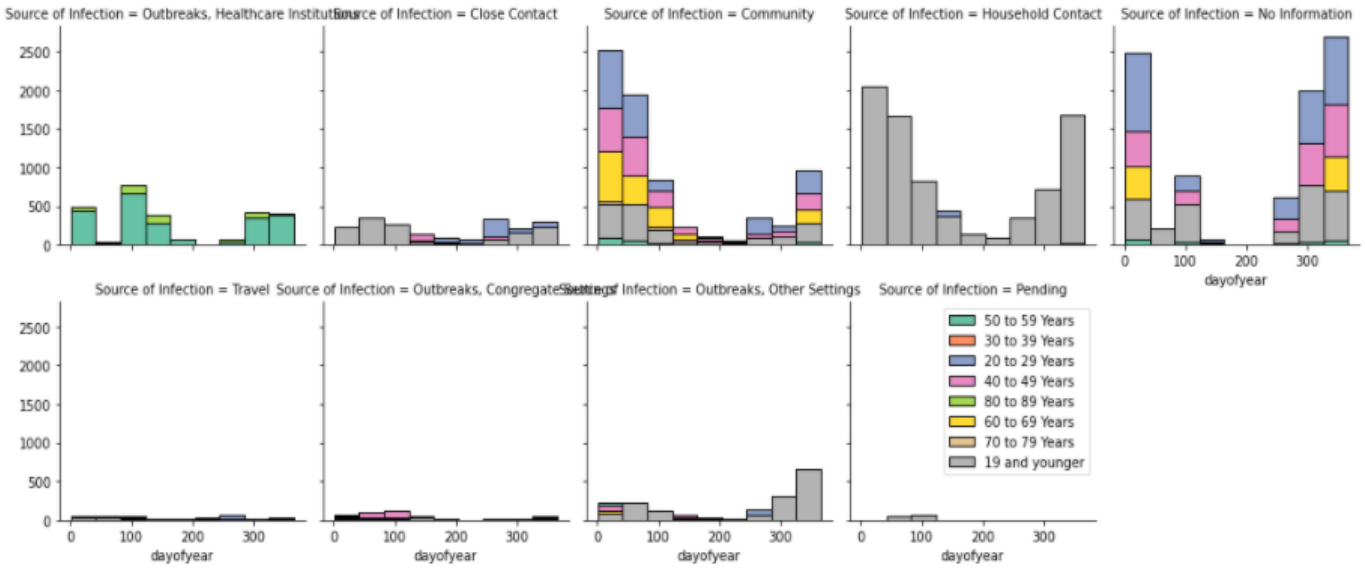


Figure 14: Age group distribution of related source of infection Covid-19 cases in time

3.5 map covid cases and neighborhoods

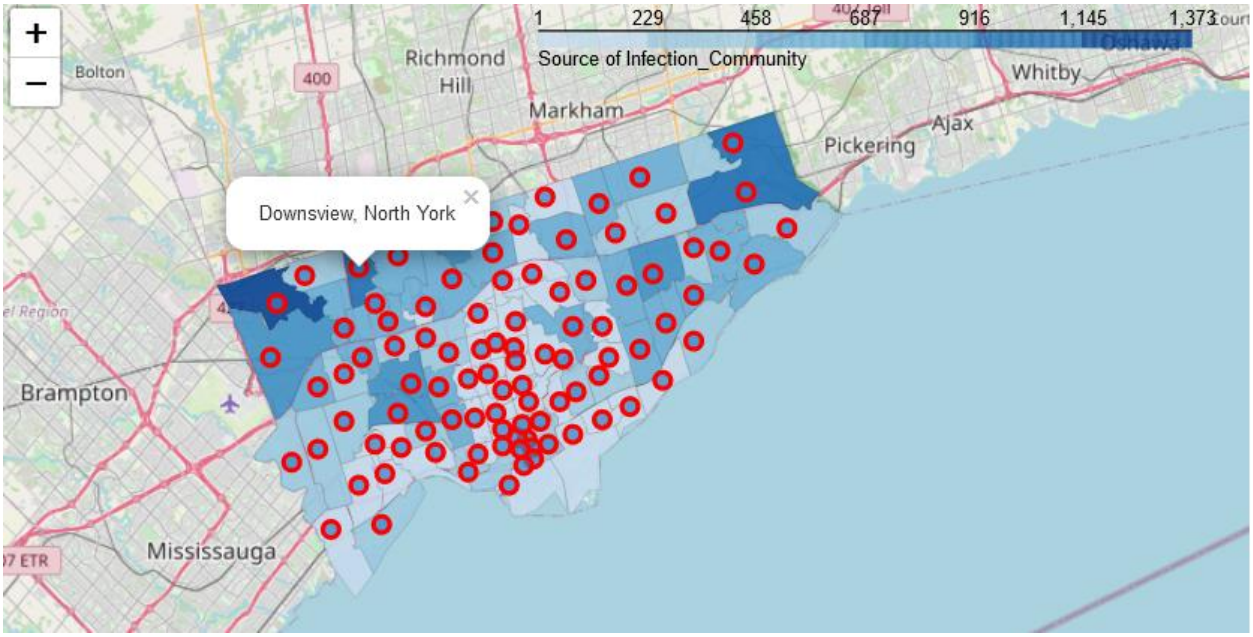


Figure 15: Map showing distribution and level of Covid-19 infections in Toronto city Neighborhoods

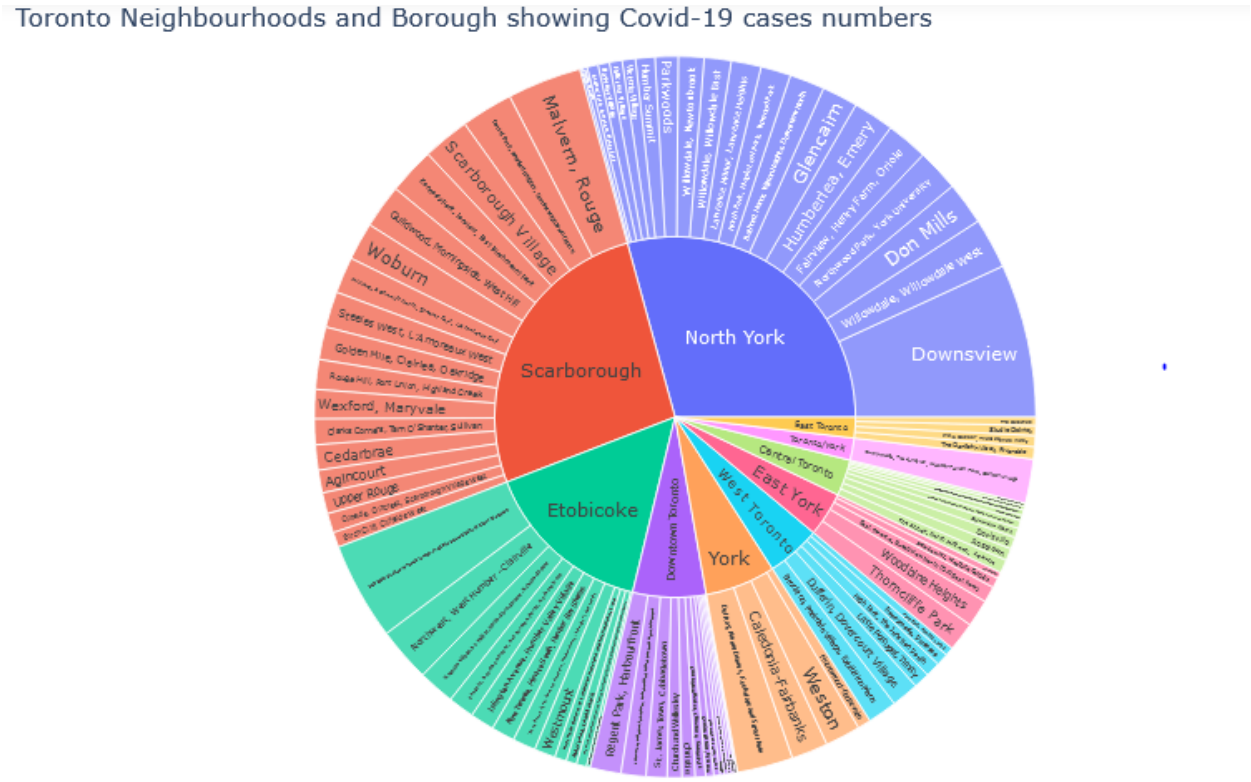


Figure 16: Toronto Neighborhoods and borough showing Covid-19 cases numbers

3.6 Machine learning algorithm

K-means clustering was used in this project to group neighborhoods based on the most venues types and the level of COVID-19 cases, to get the safest and the less competitive neighborhood in terms of number of venues to start a business in.

Optimal K

The optimal k was obtained using the Elbow Method. For this method, the dataset is fit with k-means model of a range of values (1-10). See figure 17.

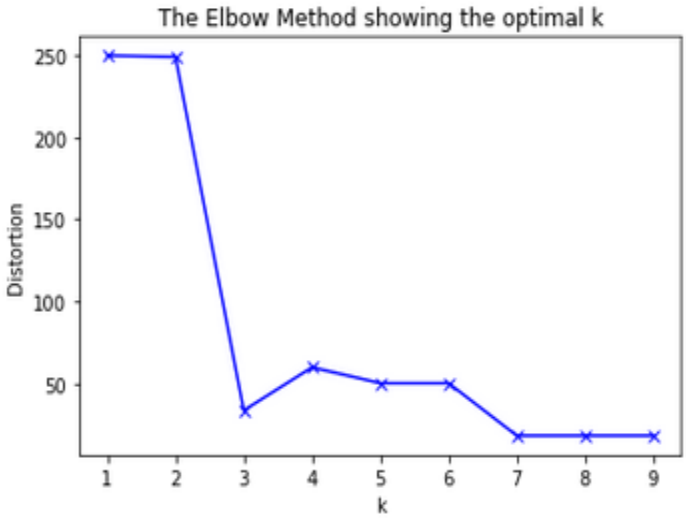


Figure 17: Elbow Method showing optimal k

4. Results and discussion

Using K-means clustering we were able to obtain 3 distinct clusters, mostly based on level of COVID-19 cases but also based on types of venues and most common venues. See table below.

Cluster Labels	Latitude	Longitude	Outbreak Associated_Outbreak Associated	Outbreak Associated_Sporadic	cases_num
0	43.728899	-79.380931	276.282051	1629.128205	1905.410256
1	43.687027	-79.399248	104.520833	454.125000	558.645833
2	43.741120	-79.474771	458.200000	3523.200000	3981.400000

- ***Cluster 0(Mid-Level):*** Neighborhoods with moderate covid-19 cases and diverse venues mostly coffee shop
- ***Cluster 1(Low Level):*** Neighborhoods with low covid-19 cases and diverse venues mostly Café
- ***Cluster 2(High Level):*** Neighborhoods with High covid-19 cases and the most venues are grocery stores, discount stores and fast-food restaurants.

4.1 map with clustered neighborhoods

We can see that density in neighborhoods such as Etobicoke is same or less than downtown Toronto, but the Covid-19 cases numbers are high there, which could means people aren't following the public health recommendations regarding social distancing, social gathering and masks correctly.

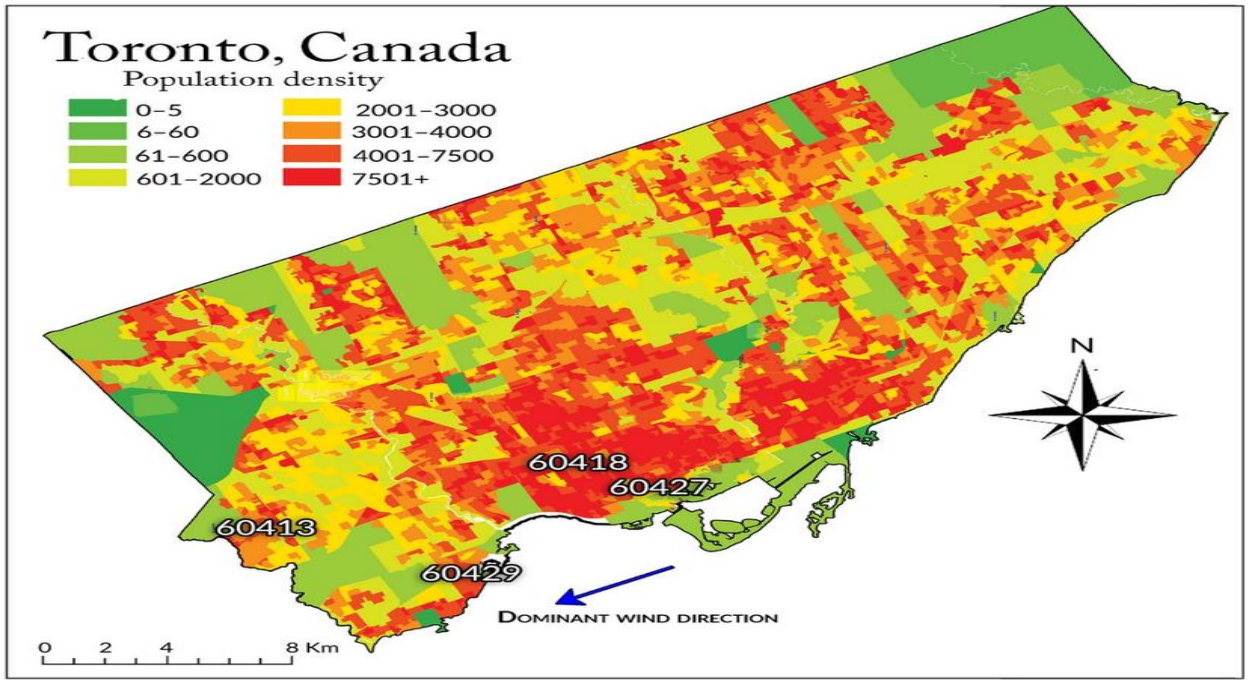


Figure 18: Population density map from researchgate

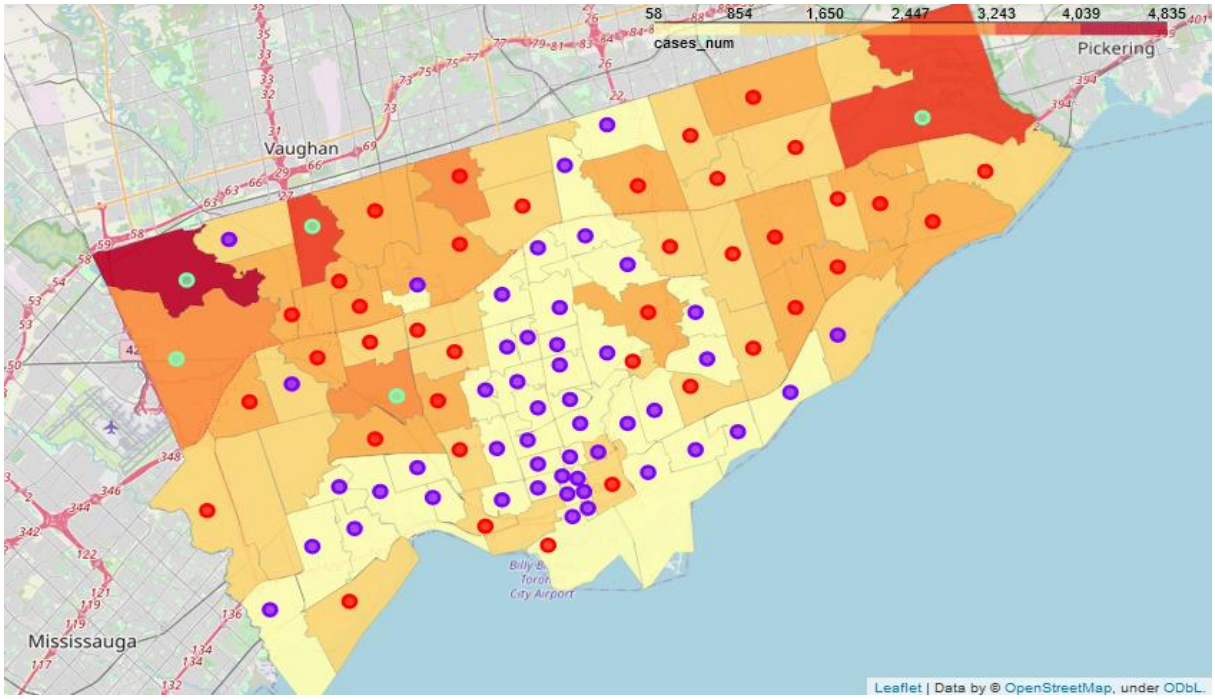


Figure 19: Choropleth map showing our neighborhoods clusters in Toronto city and Covid-19 distribution and level.

Here we have a sunburst figure showing postal codes in each cluster, details of cases in each neighborhoods for example can be found in the Jupyter notebook shared on my GitHub. See figure 20.

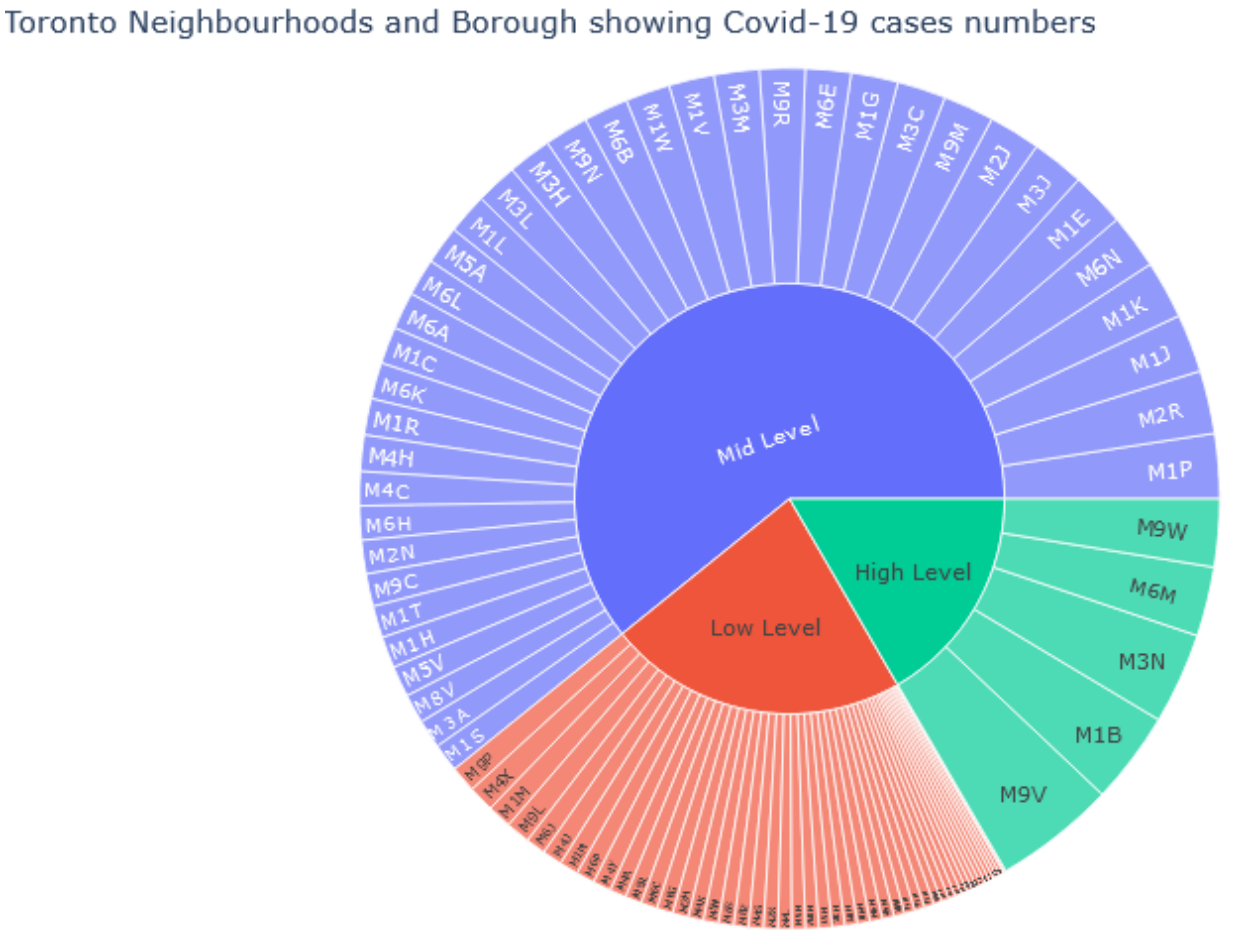


Figure 20: Postal codes in each cluster.

In Figure 21 and 22 below we can observe that around cafe, coffee shop, stores and fast-food restaurants we have high numbers of Covid-19 cases. And that could be linked to the impact of venues in the propagation of the virus.

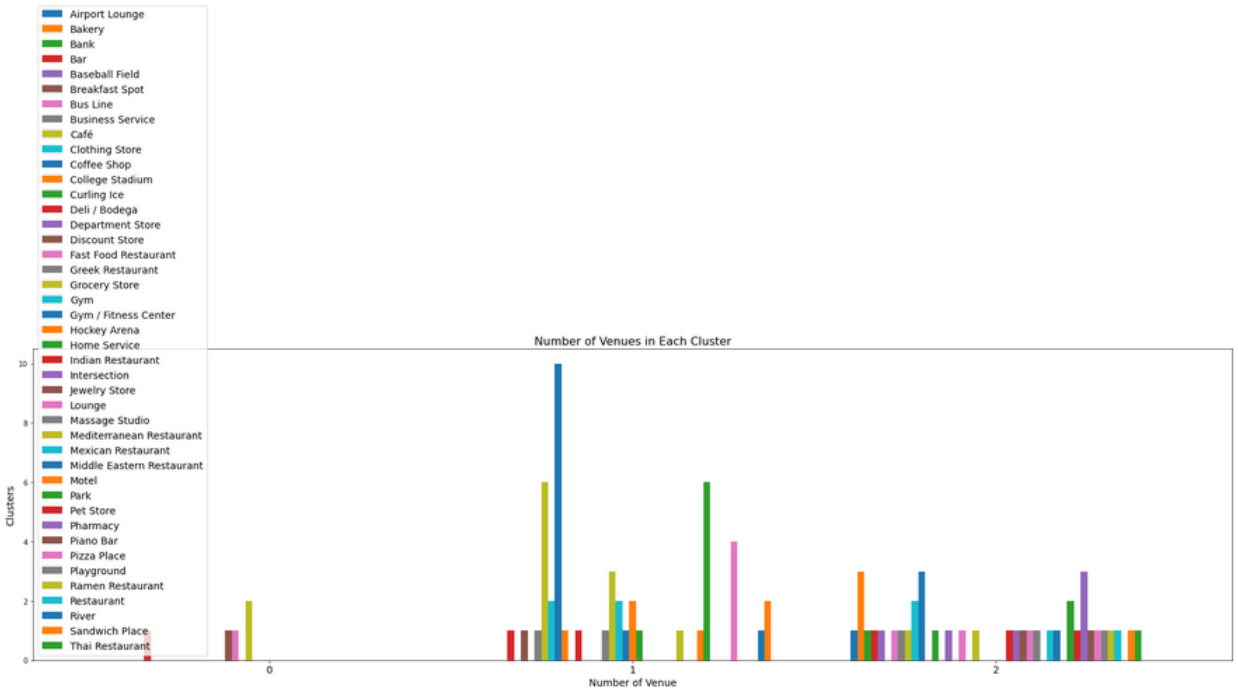


Figure 21: Distribution of Venues in each cluster

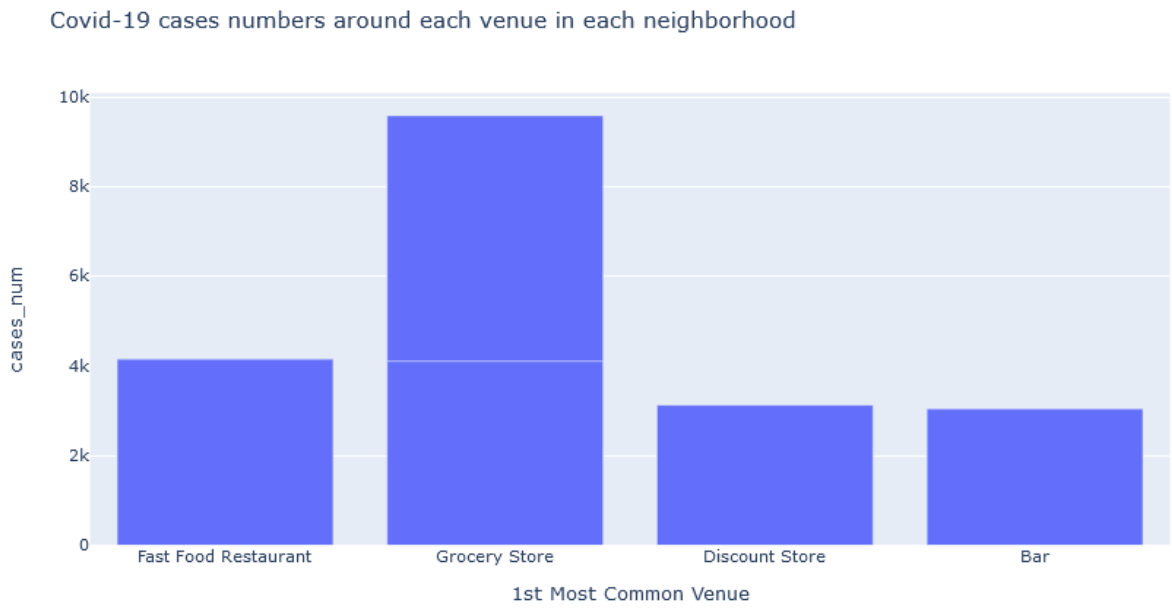


Figure 22: Cases numbers around each venue in each neighborhood

5. Conclusions

Recommandations:

- Population density is higher in low level cluster which means there is less COVID-19 community infections and more people to attract if you are willing to open a business
- Low Clusters and mid-level cluster have diverse venues so that could mean more competitive market for some businesses.
- Low Clusters and mid-level clusters have Less risk of having a Covid-19 outbreak related in your business.
- **To avoid:** high level cluster (Etobicoke and Scarborough, Malvern rouge) where high number of Covid-19 is observed knowing that this neighborhood has less population density than other city's neighborhoods.

References

- Wikipedia

- Foursquare API

- https://open.toronto.ca

- https://www.researchgate.net/figure/Toronto-Canada-population-density-and-locations-of-four-monitors-used-in-the-study_fig1_304186957

- https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/