# Covid Cases & Venues Data Analysis of Toronto

# HAO CHEN

## Introduction

### A1 Description & Discussion of the Background

Covid-19, without any question, is one of the hottest topics in 2020. As an international student in Canada, I am interested in the trend of covid-19 in Toronto, the largest city of Canada. Toronto is the capital city of the [Canadian province](https://en.wikipedia.org/wiki/Provinces_and_territories_of_Canada) of [Ontario](https://en.wikipedia.org/wiki/Ontario). With a recorded population of 6,197,000, it is the [most populous city in Canada](https://en.wikipedia.org/wiki/List_of_the_100_largest_municipalities_in_Canada_by_population) and the [fourth most populous city in North America](https://en.wikipedia.org/wiki/List_of_North_American_cities_by_population). The city has a current area of 630.2 km2, and itis divided into 140 districts. As the largest and the most populous city in Canada, Toronto residents are undertaking the highest risk of being infected.

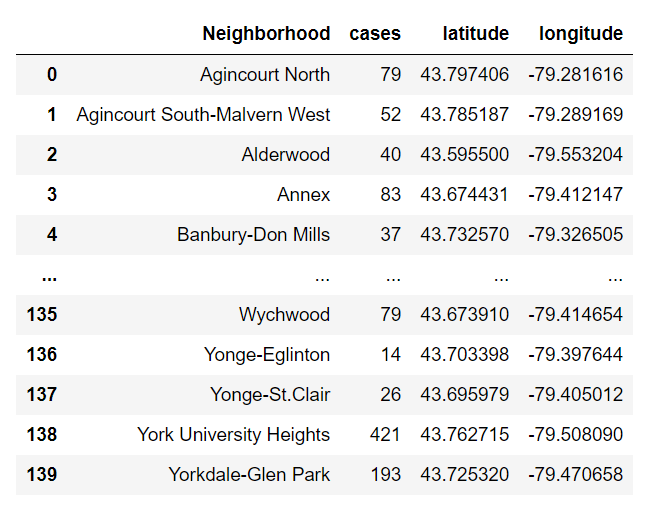
The purpose of this data analysis project is to figure out the relationship between the features of neighborhoods in Toronto and the level of covid-19.

### A2. Data Description

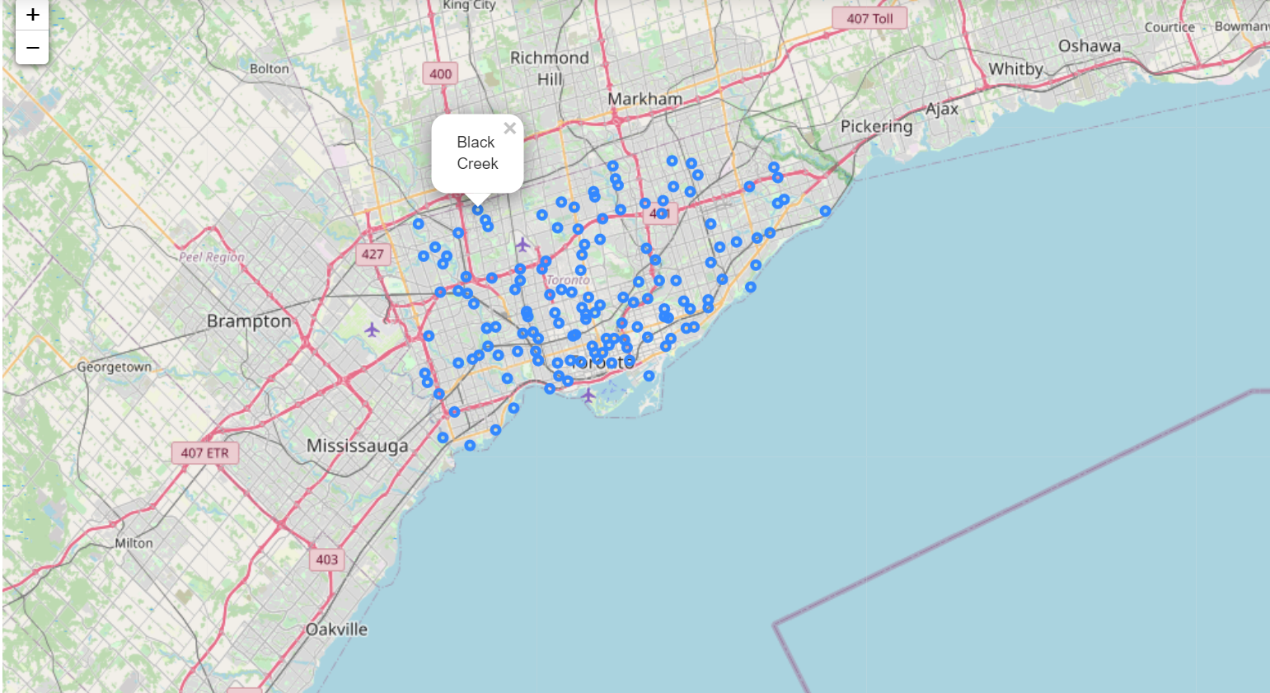
* I used Toronto neighborhood demographics and geographical and venues information from Kaggle.
* I used Toronto Covid-19 cases from Kaggle.
* I used Foursquare API to get the most common venues of neighborhoods of Toronto.

## Methodology

After data cleaning and refining of the Toronto neighborhood demographics and geographical and venues information dataset and the Toronto Covid-19 dataset by using python library Pandas, the main components of my dataset is Neighborhood name, number of cases, latitude of the neighborhood and longitude of the neighborhood. The view of the data is shown below.

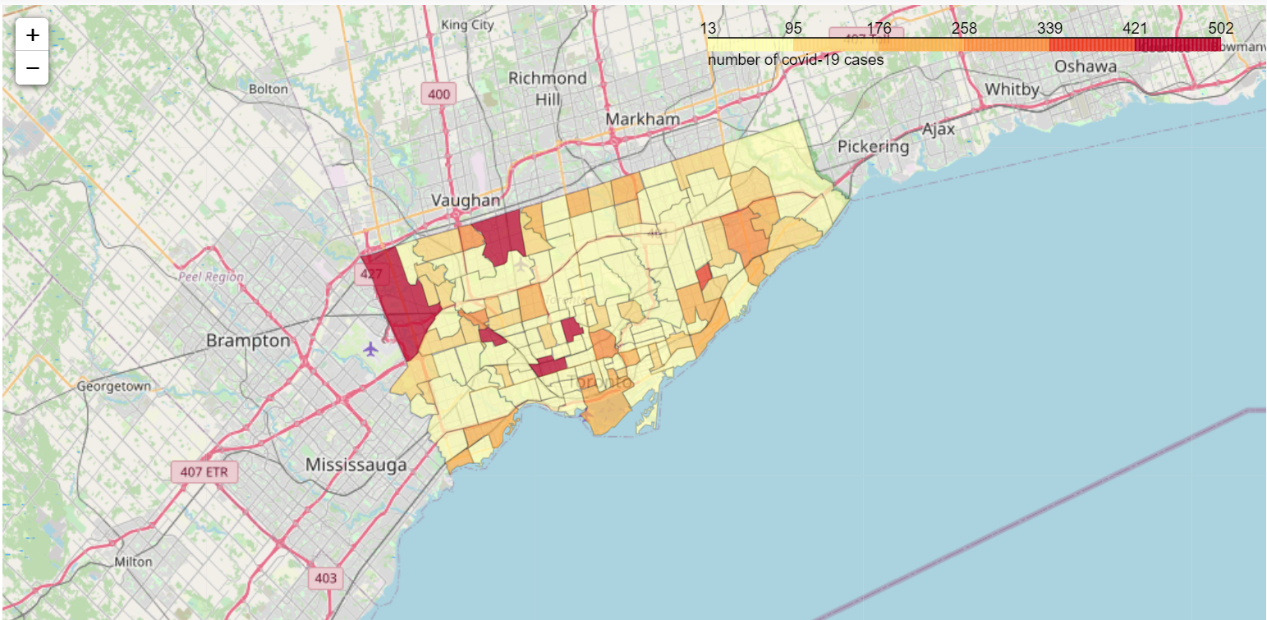


By using python map library Folium and the coordinates of Toronto neighborhoods, I created a map of Toronto on which all the neighborhood locations are marked. The figure is shown below.



Then, by using Geojson file from Toronto Open Data and python library folium, I drew a choropleth, which reflects the level of covid-19 of each neighborhoods in Toronto. The choropleth is shown below.

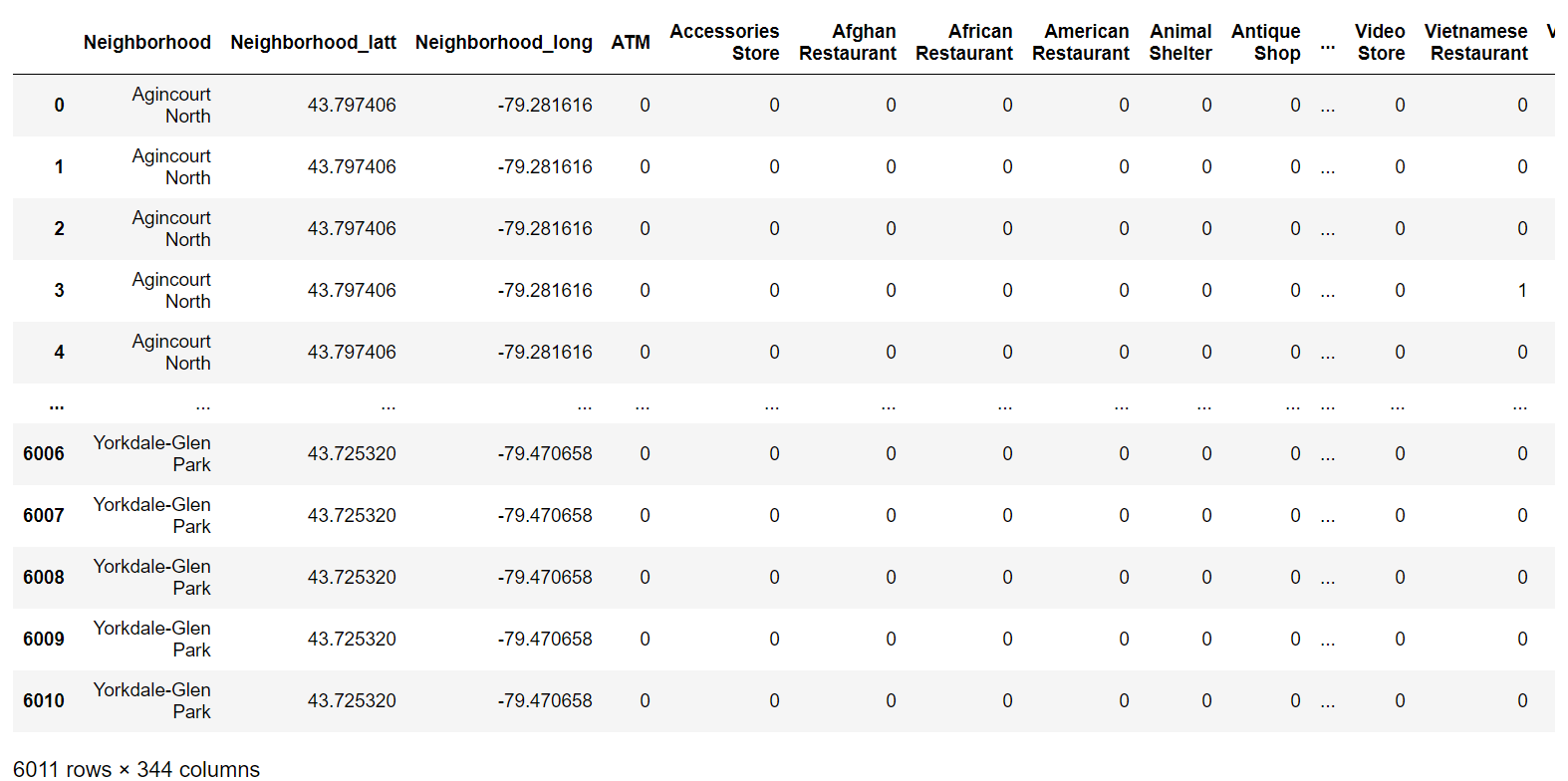
What we can observe from the choropleth is that the level of covid-19 is higher at the border of city.



After an overview of the covid-19 situation of Toronto, I used the Foursquare API to further explore venues of all the neighborhoods in Toronto. I set the radius parameter to 1100 and the limit parameter to 150 venues. A snippet of the sheet of the result is shown below.



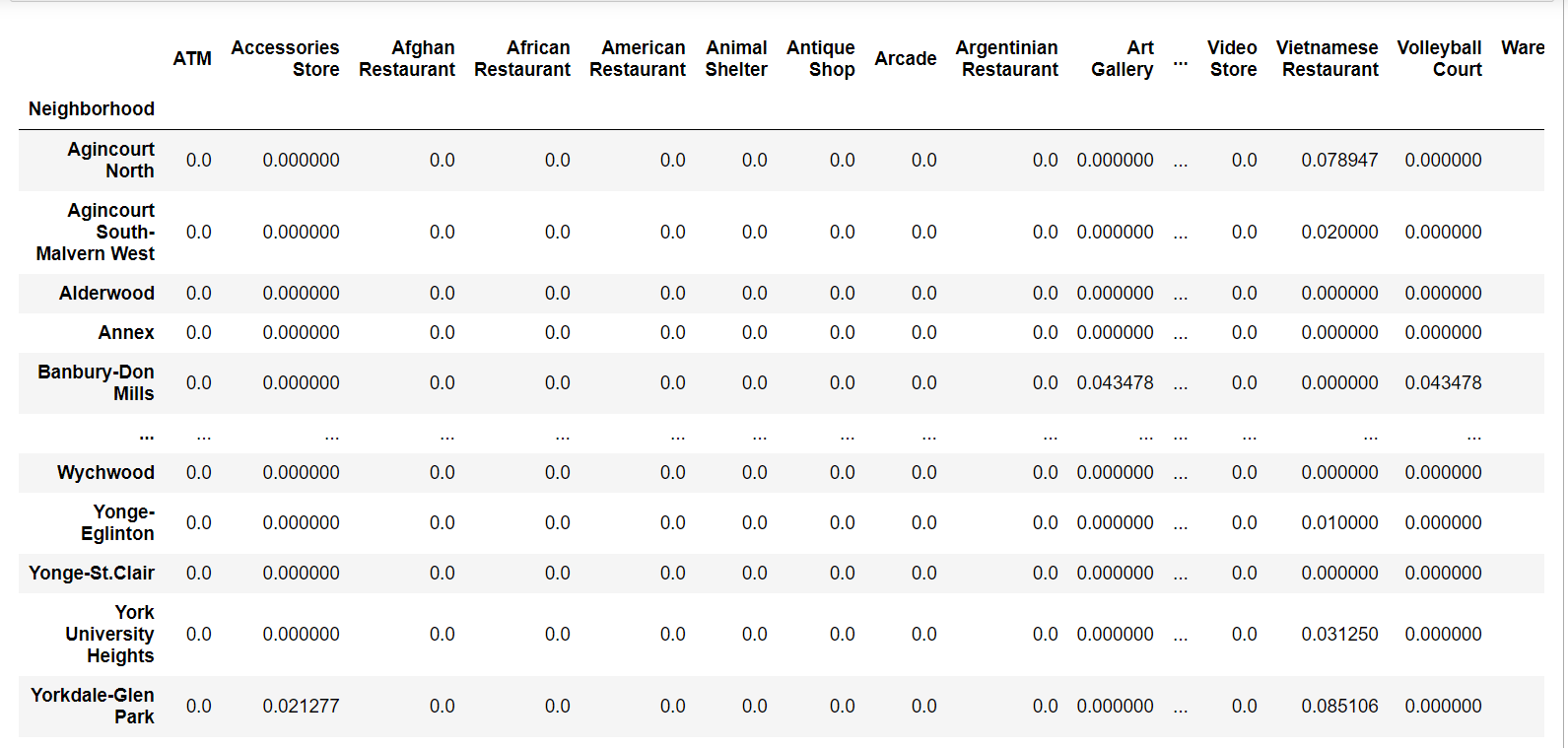
Then, I refine the data and create dummies for the venue category. The result of the refinement is shown below.



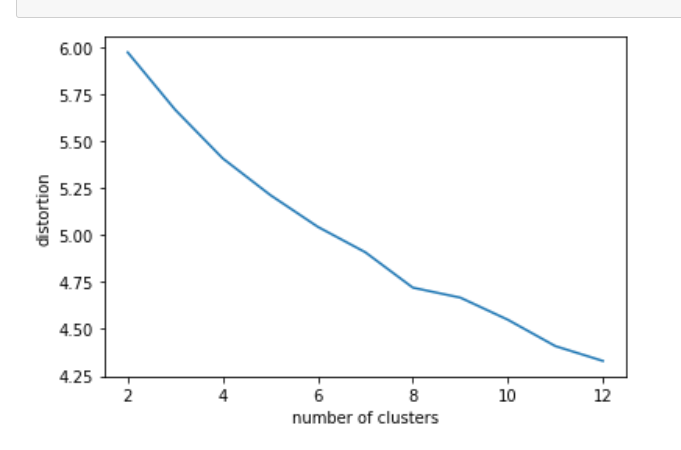
After created the dummy table, I grouped the data by neighborhoods and then selected top 5 venues of each neighborhood based on the ranking of average frequencies of each venue. The resulting data reflects what kind of venues appear the most in a certain neighborhood.



In order to gain a complete comprehension of the data above, I decided to use unsupervised learning algorithm, K-Means algorithm to cluster the neighborhoods. In other words, I used this algorithm to categorize the neighborhoods in Toronto based on the venues. The [K-Means](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans) algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the *inertia* or within-cluster sum-of-squares . This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields. The table below is the dataset I utilized as sample in K-Means algorithm.

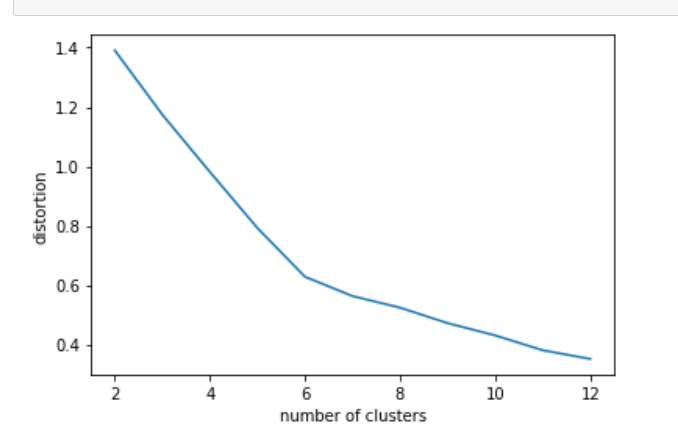


The first step to implement the K-Mean algorithm is to choose a proper number of clusters. I used the elbow method based on the inertia (Sum of squared distances of samples to their closest cluster center.) to determine the best number of clusters. By switching cluster numbers using loop, I finally got the elbow graph below.



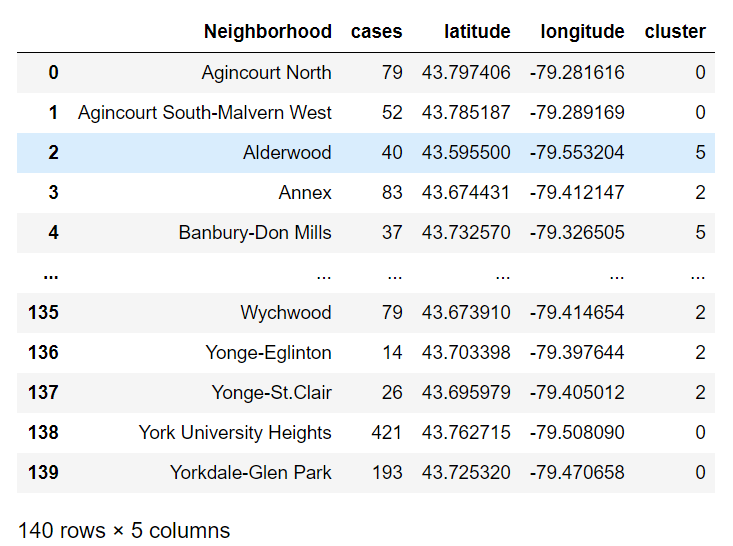
Based on the graph above, I observed that there is no elbow point. One possibility about why no elbow point is that clustering in very high dimensions is tricky for the Euclidean distance metric all distances tend to the same as the number of dimensions increases. Thus, I implemented Principal component analysis (PCA) to reduce dimensionality of the sample dataset. Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning.

After implemented PCA, the elbow graph looks like below.

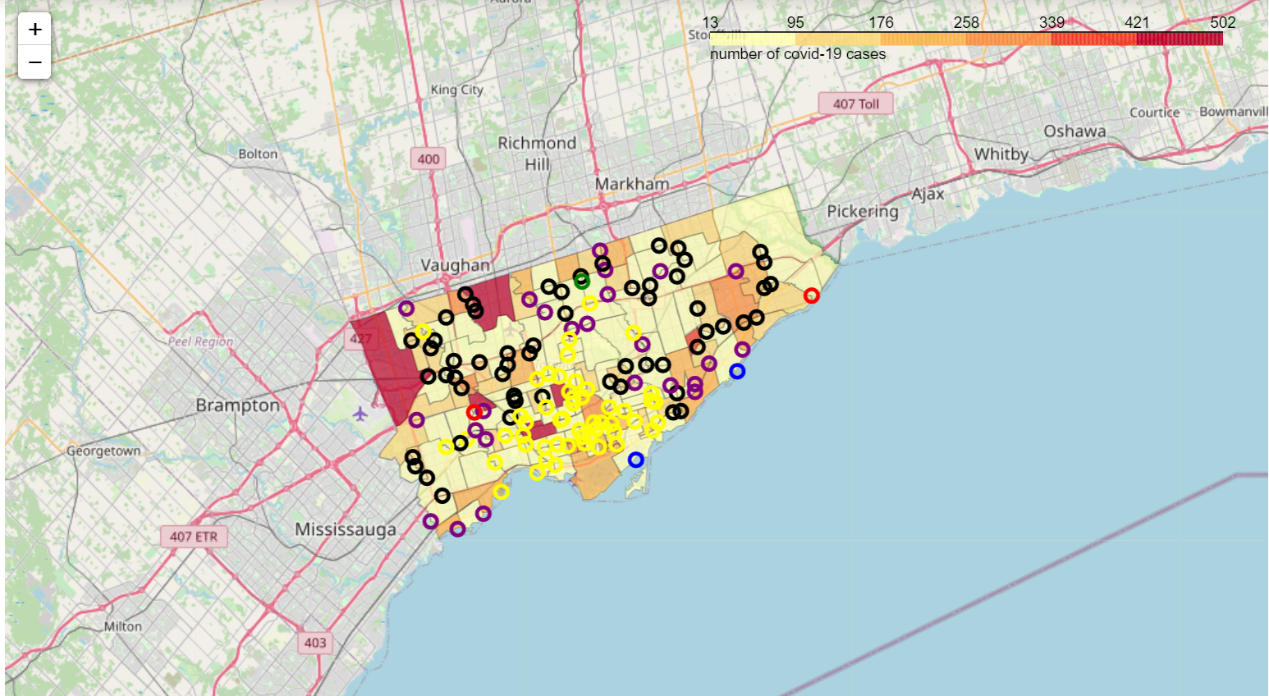


Hence, the best number of clusters should be 6.

Here is the merged table after clustering.

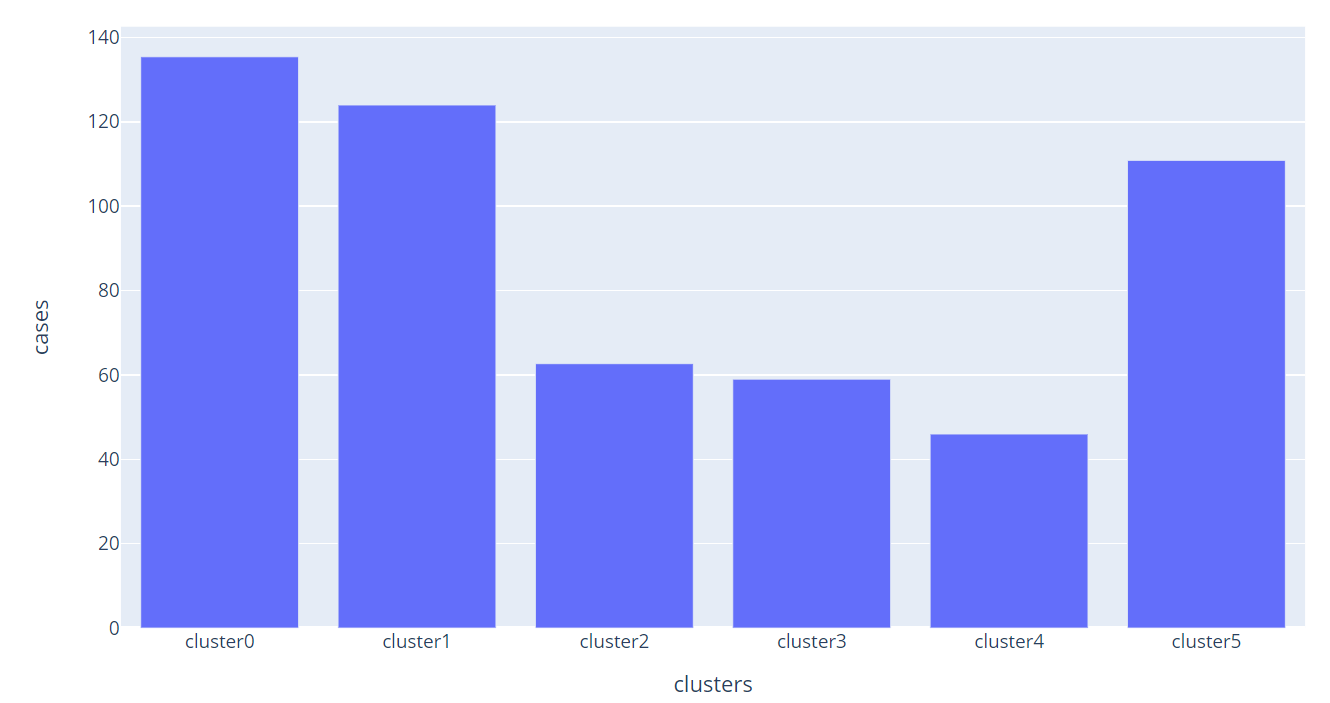


The cluster results on Toronto map combined with choropleth on cases:



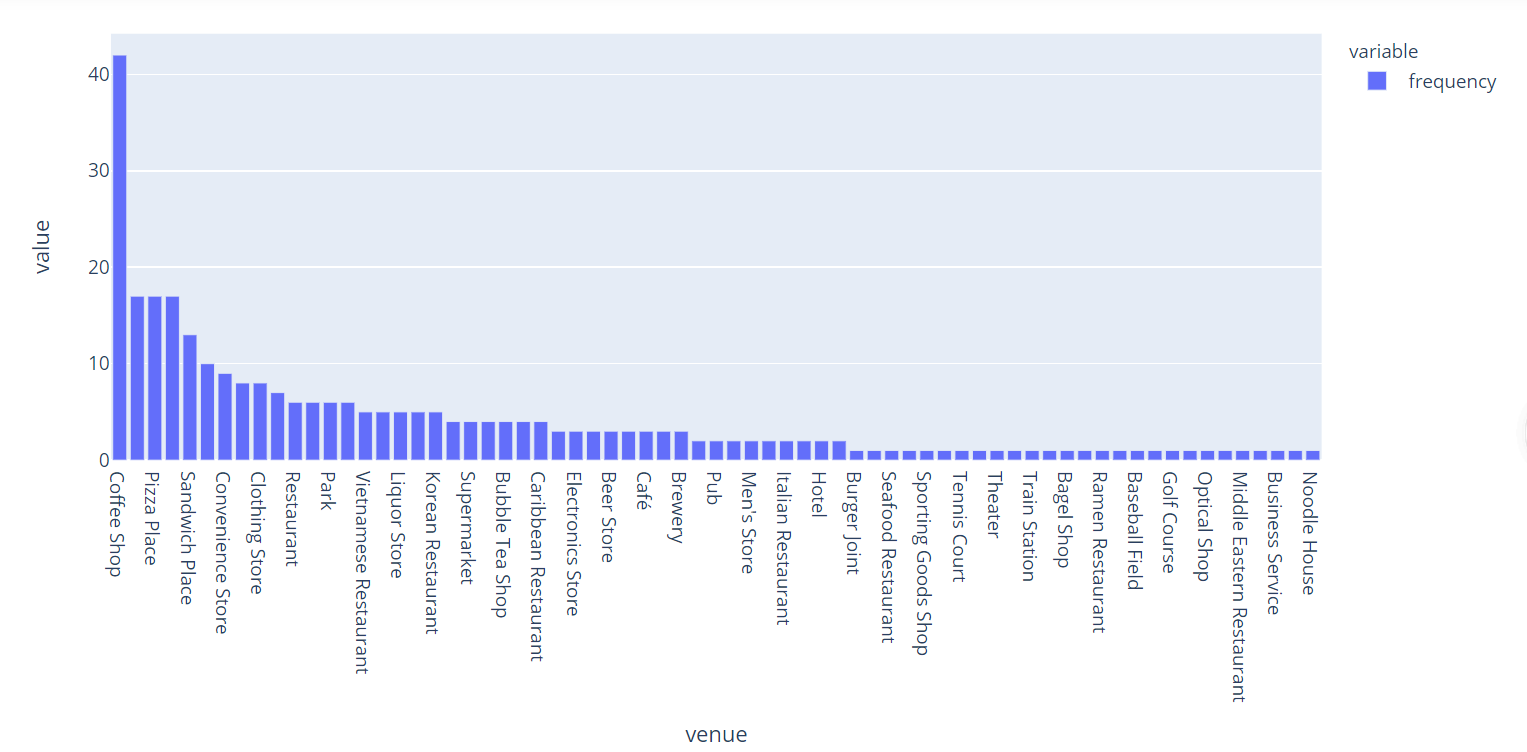
We can observe from the map that neighborhoods marked by black circles (cluster0) tend to have more cases.

I calculated the average cases of each cluster, results shown below:

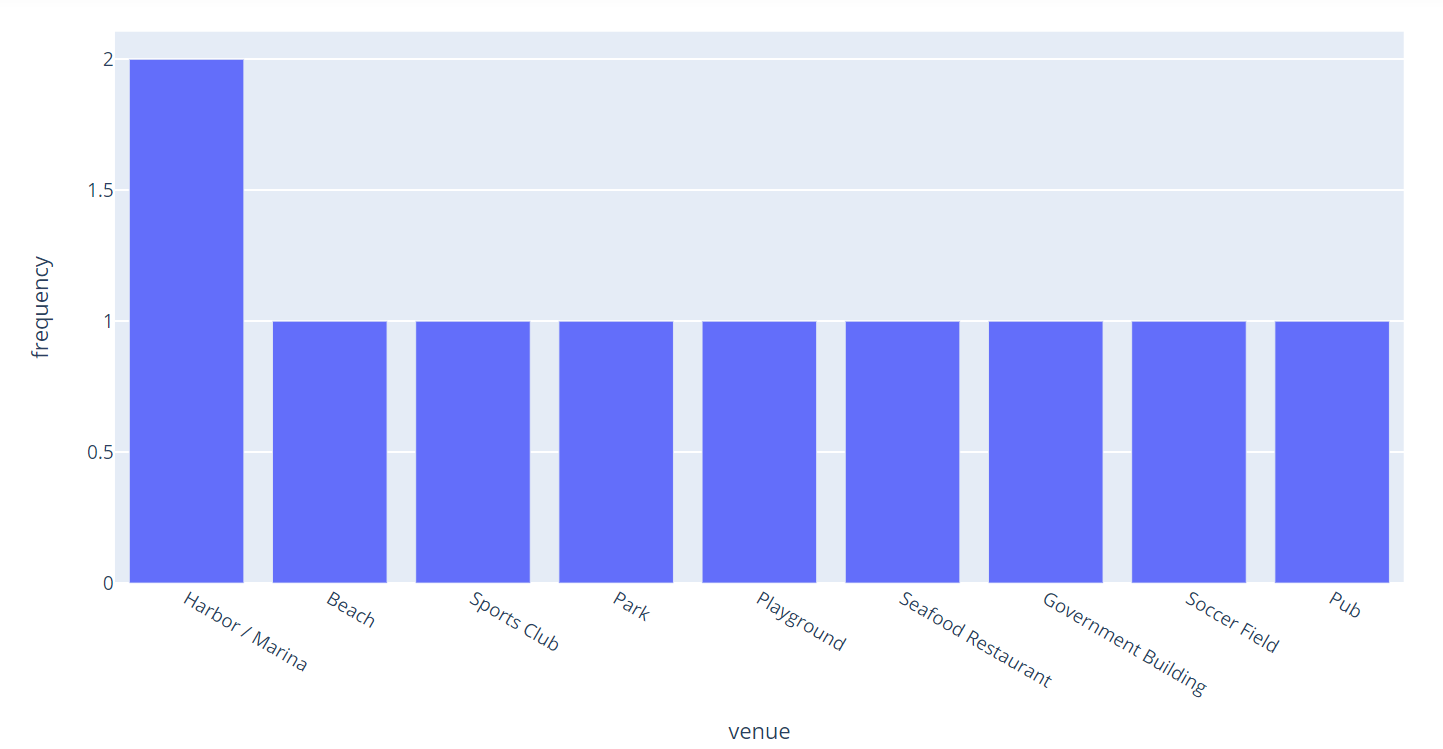


From the graph we can observe that cluster 0 has the most cases whereas cluster 4 has the least cases.

After gathered this information, I did some research to figure out venue features of cluster 0 and cluster 4.



The chart above displays the frequency of different categories of venues in neighborhoods of cluster0. We can observe that venues such as coffee shops and restaurants have a relatively high frequency. Thus, this trend can explain why neighborhood in cluster0 has the most cases since people are very likely to crowd in these venues which reduce the social distance and therefore increase the risk of getting infected.



The chart above shows the frequency of different categories of venues in neighborhoods of cluster4. It is obvious that venues in neighborhoods within cluster 4 are relatively large, so that social distances among residents are large enough to reduce the risk of being infected.

## Discussion & Conclusion

In conclusion, In Toronto, the more the venues that citizens are easily crowded in a certain neighborhood, the more the covid-19 cases. As the pandemic is still a big health issue around the world, I believe that this rule is applicable to every corner of the world. Personally, I would like to advise everyone to keep a safe social distance with other people to prevent from being infected.