The Impact of Neighbourhood Composition on Resident COVID-19 Infection Rates in Toronto

Liam Browne

August 17, 2020

1. Introduction

1.1. Background

The 2020 COVID-19 pandemic has gravely impacted every nation on Earth. Within the span of a few months, borders have closed, the hospitality industry has vastly dissipated, and remote working has become the norm in much of the developed world. The impact of this on the human experience is perhaps most widely felt at the community level as it is local communities that constitute the fabric of modern society. Resultingly, in Canada, much of the decision-making authority regarding the public health response relating to the transmission of the SARS-CoV-2 virus (the virus responsible for the coronavirus disease) has been entrusted to the provincial governments. The province of Ontario has opted for a regional approach in applying its public health policies due to the idiosyncratic nature of its cities. Ontario's capital and its largest city, Toronto, can be further thought of as a tapestry of distinct communities largely individualized by their local venues. The conventional wisdom, it follows, is that to decrease the transmission of SARS-CoV-2 in Toronto, one must engage in an examination of the local neighbourhoods that comprise the city.

1.2. Objective

Given the idiosyncratic nature of Toronto's neighbourhoods and the previously discussed conventional wisdom, the objective of this report is to determine if a segmentation of Toronto's neighbourhoods via their most popular venues provides an insight into the COVID-19 infection rates of the neighbourhoods' residents.

1.3. Relevant Stakeholders

Those conducting research into the spread of COVID-19 might be particularly interested in the transmission of the disease as it relates to the composition of neighbourhoods – in this case, using Toronto as a case study. Further to that group, public health officials in the offices of the Chief Medical Officer of Health for Ontario and the Medical Officer of Health for Toronto as well as those in the Ontario Premier's office and city officials would likely be interested in this information as they begin to lift so-called "lockdown restrictions" within the city of Toronto and neighbouring regions.

2. Data

2.1 Data Usage and Sources

The data used in this report come entirely from three different sources. All data relating to the venues that comprise the different neighbourhoods in Toronto is taken from Foursquare's location data.[1] This data includes the various venues in Toronto and their geographical coordinates that are used to determine which types of venues are most popular in each of Toronto's neighbourhoods. The COVID-19 data is the cases per 100,000 people in each neighbourhood. It should be noted that the cases used to create these case rates are based on where the residents live and are not necessarily where the residents contracted the disease. Nonetheless, this data is useful as it suffices as a proxy for where the infection was contracted. As for sources, the COVID-19 neighbourhood cumulative case rates are published by the department of Public Health within the City of Toronto.[2] This is also where the names of all of Toronto's official neighbourhoods comes from. The geographical coordinates for each neighbourhood were taken from Google.[3]

The list of Toronto's venues and their geographical coordinates are used to determine which types of venues are most popular within a given neighbourhood. These venue types are then used as the features associated with the neighbourhoods for the purpose of neighbourhood segmentation via machine learning algorithms. The COVID-19 case rates, on the other hand, are used to create a violin plot with a swarm plot superimposed on it to examine if there are any noticeable differences in COVID-19 rates between neighbourhood clusters.

3. Methodology

3.1 Segmenting Neighbourhoods

The records from Toronto Public Health are fully comprehensive and do not contain any missing values. As such, there was no need to deal with any missing values. For each neighbourhood as defined by the City of Toronto, its geographical coordinates were found via Google. The Foursquare API was then used to find the first 100 venues within a 500 m radius of the geographical coordinates of each neighbourhood. The neighbourhoods then needed to be segmented. To do this, the k-means clustering machine learning algorithm was chosen. For this algorithm, dummy variables were used to code each venue type (for the top ten venue types in each neighbourhood) in order to cluster the neighbourhoods.

---

1. https://developer.foursquare.com
2. https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/
3. https://www.google.ca/

Of course, the first step in k-means clustering is to determine the optimal number of clusters, *K*. At first, the Gap Statistic method as defined by Tibshirani, Walther, and Hastie was chosen.[4] After changing the hyperparameter controlling the maximum number of clusters a few times, it became apparent that the algorithm was selecting the maximum number of clusters each time. From the below graph (Figure 1), it becomes apparent that the gap statistic increases with the number of clusters without a sufficiently large drop as to clearly identify an optimal *K*.

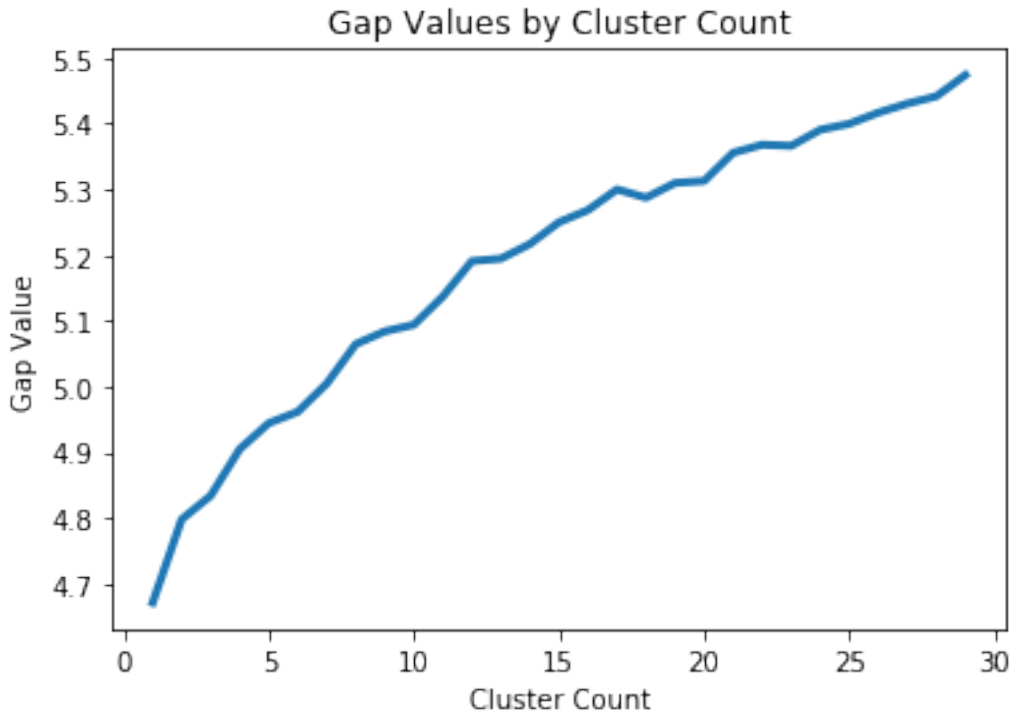

Figure 1: Gap values by cluster count for the types of venues of Toronto's neighbourhoods.

Resultingly, the gap statistic method was abandoned in favour of the Elbow method. Figure 2, below, is a graph of the distortion values by maximum number of clusters. From this graph, it is apparent that the distortion values are decreasing as the number of clusters increases without a sufficiently large drop in the marginal decrease to identify a so-called "elbow." Hence, the Elbow Method was not useful in trying to identify the optimal *K*. Instead, the optimal *K* was chosen to five as this is a workably small number.

[4] https://web.stanford.edu/~hastie/Papers/gap.pdf

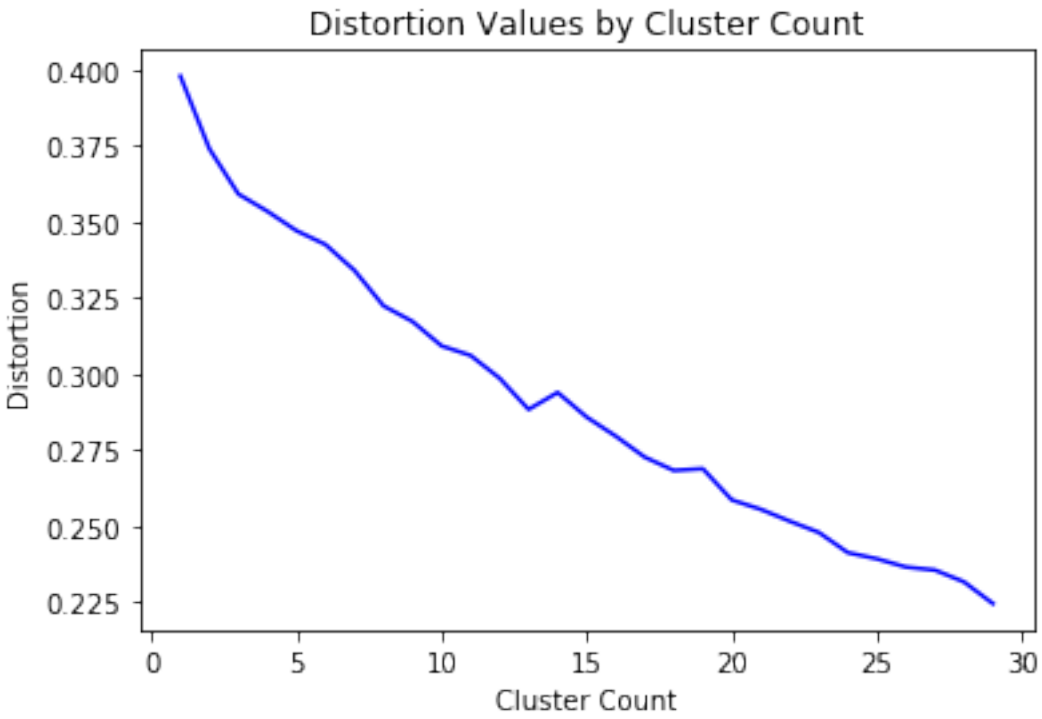

Figure 2: Distortion values by cluster count for the types of venues in Toronto's neighbourhoods.

The neighbourhoods were then clustered using Scikit-learn's k-means clustering. The results of this clustering of neighbourhoods by venue type are shown below via the Folium library.

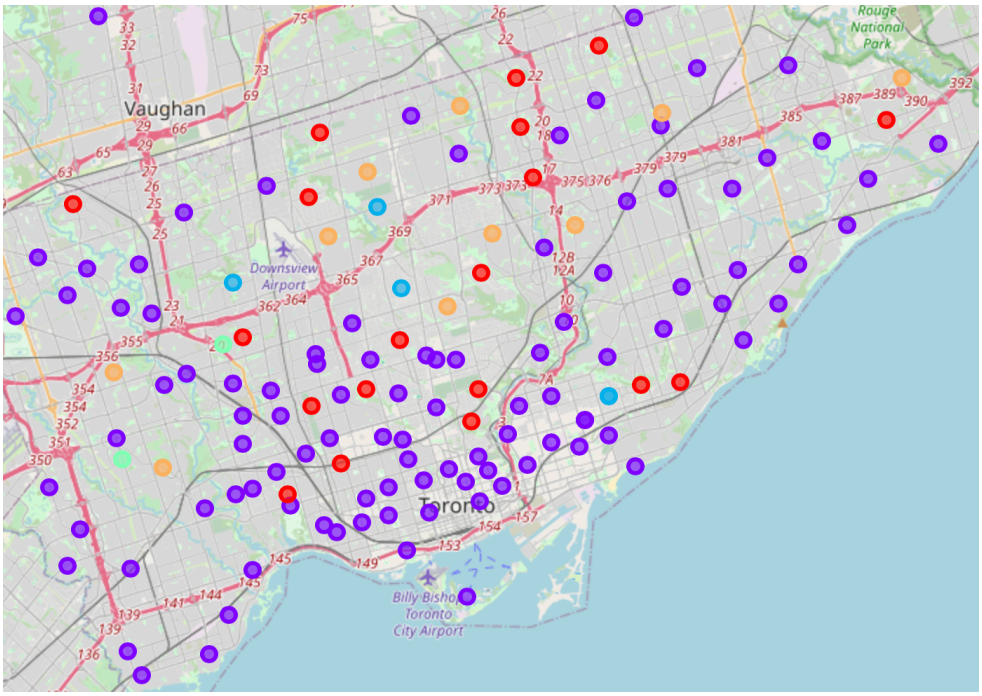

Figure 3: Map of Toronto's neighbourhoods segmented by the type of venues popular in each neighbourhood.

3.2 Classifying Neighbourhood Clusters

In order for the clusters to be useful, a meaningful label must be assigned to each of the five clusters of neighbourhoods. Table 1, below, is the first five rows of the neighbourhood data up to this point in the analysis. (NB: the table also contains the ninth and tenth most common venue type in each neighbourhood, although that is not displayed in Table 1.)

	Neighbourhood Name	Rate per 100,000 people	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Eglinton East	430.277485	43.7371	-79.2462	1	Pharmacy	Sandwich Place	Japanese Restaurant	Auto Garage	Beer Store	Liquor Store	Grocery Store	Bus Station
1	Don Valley Village	255.073750	43.7874	-79.3530	0	Bakery	Park	Fried Chicken Joint	Grocery Store	Caribbean Restaurant	Fruit & Vegetable Store	Sandwich Place	Beer Store
2	Lansing-Westgate	222.717149	43.7590	-79.4226	2	Spa	Women's Store	Dog Run	Fish Market	Fish & Chips Shop	Filipino Restaurant	Field	Fast Food Restaurant
3	Edenbridge-Humber Valley	656.581912	43.6671	-79.5280	4	Park	Baseball Field	Skating Rink	Women's Store	Event Space	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
4	Flemingdon Park	606.392194	43.7184	-79.3314	1	Gym	Fast Food Restaurant	Café	Coffee Shop	Grocery Store	Gym / Fitness Center	Women's Store	Dumpling Restaurant

Table 1: First five rows of the neighbourhood data immediately after clustering.

Since we have already determined the most common type of venue in each neighbourhood and the neighbourhoods have been grouped, we can determine the most common venue type in each of the five clusters. This is what was chosen as the labels for the clusters. The results are below in Table 2.

Cluster	Most Common Venue Type
0	Bakery
1	Coffee Shop
2	Convenience Store
3	Construction & Landscaping
4	Park

Table 2: Most common type of venue for each of the five neighbourhood clusters. These values were found by identifying the top ten most common venues in each neighbourhood, using them to cluster the neighbourhoods via k-mean clustering, then identifying the most common value in the "1st Most Common Venue" column for Table 1 for each cluster.

4. Results

4.1 COVID-19 Infection Rates by Neighbourhood Type

As defined in §2.1 of this report, the COVID-19 rates are the infection rates of the residents who live in each neighbourhood (not necessarily where they got exposed to the SARS-CoV-2 virus). The chief objective of this report, as set out in §1.2, "is to determine

if a segmentation of Toronto's neighbourhoods via their most popular venues provides an insight into the COVID-19 infection rates of the neighbourhoods' residents." In pursuit of this objective, a violin plot with a swarm plot superimposed on it is used to identify if there are any types of neighbourhoods with unusually high or low COVID-19 infection rates. The violin plot is as follows in Figure 4.

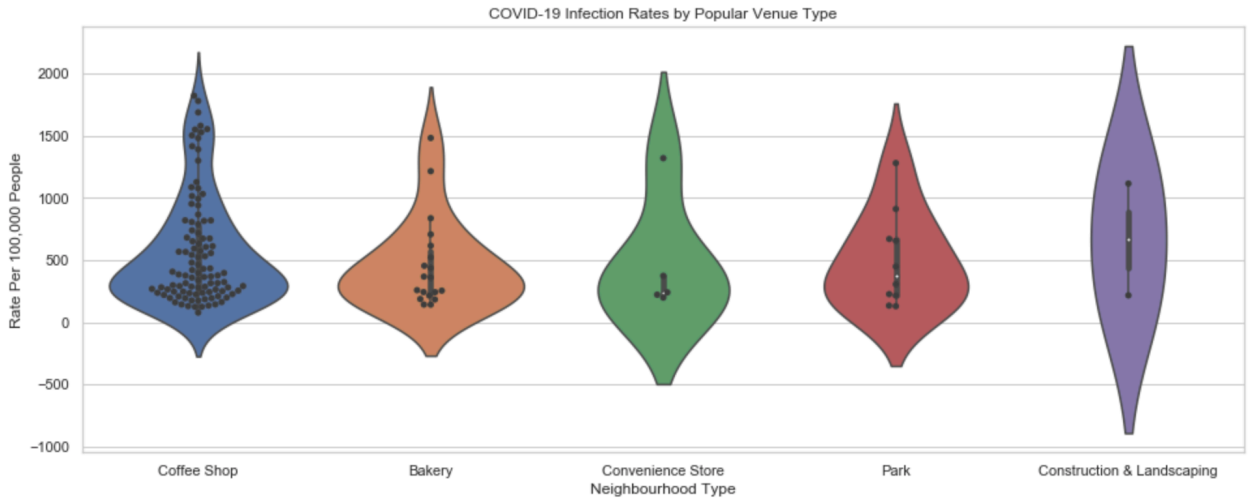

Figure 4: Violin plot of COVID-19 infection rates per 100,000 people separated by neighbourhood type.

5. Discussion

5.1 Impact of Neighbourhood Type on Infection Rates

Given Figure 4 above, we can now answer the main question laid out in §1.2: does the segmentation of Toronto's neighbourhoods by their most popular venue provide an insight into the COVID-19 infection rates of the neighbourhoods' residents? From Figure 4, we see that the interquartile range of the Coffee Shop, Bakery, and Park neighbourhoods are similar – as are their median infection rates. Moreover, the violin plots for the Convenience Store and Construction & Landscaping neighbourhoods are not significantly higher or lower than any of the other violin plots in terms of medians or overall spread. Albeit, the median of the Construction & Landscaping neighbourhood is somewhat higher than the others. However, given that this neighbourhood type contains only a few neighbourhoods, it is likely not very useful for this analysis. Also of note, the distribution of infection rates for all neighbourhood types – save for the Construction & Landscaping neighbourhoods – are all concentrated, with very degrees of concentration, at the lower end of their infection rate spread.

The chief conclusion to be drawn from Figure 4 is that the composition of neighbourhoods in Toronto do not appear to have a strong impact on the COVID-19 infection rates of their residents. This provides evidence contradictory to the conventional wisdom laid out in §1.1 of this report. Instead of focusing on the idiosyncrasies of Toronto's neighbourhoods, public health officials ought to be focusing

on the homogeneity of this large city. This gives rise to the following recommendation: although the neighbourhoods of Toronto might be unique in their culture and composition, the ease of mobility in the city (i.e., via public transit and properly maintained roadways) has led to a homogeneity of the use of the city's many venues that therefore favours a collective approach to COVID-19 reduction strategies as opposed to regional ones insofar as city governance is concerned.

It should also be noted that this recommendation only applies to strategies targeting neighbourhoods based on venue popularity. It does not hold any weight in a discussion of strategies concerning other factors such as the socioeconomic inequality between Toronto's neighbourhoods.

6. Conclusion

6.1 Summary

This report analyzes the impact of venue popularity by neighbourhood on the COVID-19 infection rates of the neighbourhoods' residents. In doing so, the neighbourhoods were clustered via k-means clustering into five clusters based on venue popularity. A violin plot was then created to analyze the spread of the resident infection rates by neighbourhood cluster. In doing so, this report concludes that the idiosyncratic culture of Toronto's neighbourhoods does not extend to venue popularity in a way that is sufficiently impactful to the COVID-19 infection rates of the neighbourhoods' residents as to warrant neighbourhood-specific COVID-19 reduction strategies. This gives support to the notion that the city's COVID-19 reduction plans should focus on collective protectionism and the homogeneity of Toronto insofar as venue popularity is concerned.

6.2 Possible Research Extensions

This report focuses exclusively on clustering Toronto's neighbourhoods by the popularity of types of venues. A natural extension would be to cluster the neighbourhoods by other factors such as socioeconomic standing, age, or other population demographic factors.