Correlation between Venues in Toronto Neighborhoods and Covid-19 Transmission

Aaron Ji

December 28, 2020

1 Introduction: Business Problem

2020 has been a very special year due to the Covid-19 pandemic. This contagious virus can cause mild to moderate respiratory illness and recover without requiring special treatment. However, older people, and those with underlying medical problems are more likely to develop serious illness. Till December 30, 2020 there are totally 572982 Covid-19 cases and costed 15472 lives in Canada. Ontario has the second highest total cases (178831) among all provinces and Toronto has 60000 cases since the beginning of pandemic. This report is trying to answer following question:

Is there any correlation between venues in Toronto neighborhoods and Covid-19 transmission?

Everyone live in Toronto area could be interested in the answer of this problem, especially decision makers in government.

2 Data

2.1 Data Sources

2.1.1 Statistics of Covid-19 cases in Toronto as of Dec 22

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

Datalink:https://drive.google.com/file/d/1jzH64LvFQ-UsDibXO0MOtvjbL2CvnV3N/view

Second tab named “All cases and Rates by Neighbou” is used as data source.

Features in data source:

Neighbourhood ID: Unique ID assigned to neighborhood

Neighbourhood Name: Name of neighborhood

Rate per 100,000 people: Number of Covid-19 cases for every 100,000 people in neighborhood

Case Count: Total number of cases in neighborhood

2.1.2 GEOJSON file for neighborhoods in Toronto

Website: https://open.toronto.ca/dataset/neighbourhoods/

GEOJSON file is used to:

(1) Obtain coordinates for each neighborhood. Coordinates are used to obtain venues of neighborhoods from Foursquare.

(2) Calculate area of each neighborhood. The area of each neighborhood will define the radius passed to Foursquare venues explore API.

(3) Plot choropleth map of Toronto.

2.2 Data Cleaning

For Covid cases since the dataset is in csv format, it needs to be read into pandas dataframe and assign preferable names to features. Also total population of each neighborhood is calculated by $\frac{Case\ Count}{Rate\ per\ 100{,}000\ people} \times 100000$.

For GEOJSON file it will also be read into pandas dataframe. To obtain the coordinates and area of each neighborhood we need to extract following values:

(1) feature['properties']['AREA_SHORT_CODE'] corresponding to Neighborhood ID in Covid cases dataset is extracted as ‘ID’.

(2) feature['properties']['AREA_NAME'] corresponding to Neighborhood Name in Covid cases dataset is extracted as 'AREA_NAME'.

(3) feature['geometry']['coordinates'][0] as the coordinates of Neighborhood boundary is extracted as 'geometry'.

Then we calculate the average of all coordinates of each neighborhood as the longitude and latitude of neighborhood. Area package is used to calculate the area of each neighborhood using the coordinates of Neighborhood boundary. This area is stored in column 'AREA_M2'

Now we combine Covid cases dataset and data we extracted from GEOJSON file into one pandas dataframe called 'toronto_data'. Since Foursquare does not have API that takes geometry feature in GEOJSON, a circle with equivalent area of neighborhood located at the neighborhood coordinates is used as an approximation to calculate radius for Foursquare API. Such radius is stored in column 'RADIUS_M'. Lastly the population density (population per km^2) of each neighborhood is calculated as $\frac{Population}{AREA\_M2} \times 1000000$ as stored in 'Population_density' column.

2.3 Explore Neighborhoods in Toronto

Now we use the coordinates and radii of neighborhoods to obtain venues from Foursquare. Returned venues are stored in dataframe 'toronto_venues'. After exploratory data analysis in 'toronto_venues' we found in toronto_venues dataframe there are 328 different venue categories and quite a few has quantity of 1 or 2 which can be noise in clustering. Therefore we further group them into bigger categories per Table 1. Since neighborhoods have various of sizes, it makes sense to normalize number of venues with the area of neighborhood.

Table 1 Venue Types to be Combined into Larger Categories

Category	Venue Types in Category
To be dropped	'Intersection', 'Bus Line', 'Business Service', 'Neighborhood', 'Bridge', 'Moving Target', 'General Entertainment'
Personal_Service	'Spa', 'Salon / Barbershop', 'Nail Salon', 'Massage Studio', 'Health & Beauty Service'
Indoor_Public	'Movie Theater', 'Theater', 'Art Gallery', 'Event Space', 'Music Venue', 'History Museum', 'Concert Hall', 'Museum', 'Indie Movie Theater', 'Poke Place', 'Performing Arts Venue', 'Art Museum', 'Science Museum'
Store	'Store', 'Shopping Mall', 'Sporting Goods Shop', 'Cosmetics Shop', 'Gourmet Shop', 'Gift Shop', 'Flower Shop', 'Farmers Market', 'Hobby Shop', 'Fish Market', 'Boutique', 'Butcher', 'Flea Market', 'Antique Shop', 'Market', 'Print Shop', 'Organic Grocery', 'Record Shop', 'Bike Shop', 'Comic Shop'
Outdoor	'Park', 'Trail', 'Golf Course', 'Dog Run', 'Playground', 'Harbor / Marina', 'Beach', 'Scenic Lookout', 'Zoo Exhibit', 'Garden', 'Plaza', 'Racetrack', 'Historic Site', 'Garden Center', 'Botanical Garden', 'Lake', 'Other Great Outdoors'
Coffee_Shop	'Café', 'Gaming Cafe', 'Coffee Shop'
Bar	'Pub', 'Bar', 'Gastropub', 'Brewery', 'Cocktail Bar', 'Beer Bar', 'Wine Bar', 'Sports Bar', 'Jazz Club'
Food_Service	'Restaurant', 'Pizza Place', 'Bakery', 'Burger Joint', 'Breakfast Spot', 'Ice Cream Shop', 'Fried Chicken Joint', 'Dessert Shop', 'Diner', 'Juice Bar', 'Deli / Bodega', 'Food & Drink Shop', 'Bubble Tea Shop', 'BBQ Joint', 'Steakhouse', 'Food Court', 'Bagel Shop', 'Burrito Place', 'Fish & Chips Shop', 'Wings Joint', 'Tea Room', 'Frozen Yogurt Shop', 'Chocolate Shop', 'Salad Place', 'Lounge', 'Noodle House', 'Smoothie Shop', 'Food Truck', 'Donut Shop', 'Cupcake Shop', 'Dive Bar', 'Cheese Shop', 'Taco Place', 'Snack Place', 'Sandwich Place'
Sports_Place	'Gym', 'Skating Rink', 'Athletics & Sports', 'Yoga Studio', 'Gym / Fitness Center', 'Pool', 'Baseball Field', 'Dance Studio', 'Hockey Arena', 'Curling Ice', 'Soccer Field', 'Pool Hall', 'Soccer Stadium', 'Tennis Court', 'Bowling Alley', 'Martial Arts School'
Transportation	'Train Station', 'Metro Station', 'Bus Station', 'Light Rail Station', 'Bus Stop'

After these cleaning steps there are 19 unique venue categories and we save dataframe as toronto_venues_cleaned. This concludes all data cleaning and prepration.

3 Methodology

In this project we are trying to find out correlation between venues in Toronto neighborhoods and Covid-19 transmission, which is corresponding to 'Venue Category' in toronto_venues_cleaned dataframe and 'rate_per_100K' in toronto_data dafaframe.

First step is to cluster neighborhoods by venues using _k_-means clustering algorithm with cluster number k=5.

In second step we use the Folium library to visualize the neighborhoods in Toronto and their emerging clusters as well as transmission rate in each neighborhood.

In final step is to compare transmission rates between different clusters. Find correlation between venues in cluster and cluster transmission rate.

4 Results and Discussion

From population density map (Figure 1) the population density is evenly distributed except a few neighborhoods. As we can see from the transmission rate map (Figure 2) green and purple clusters happen to have higher transmission rate. They are located outside the core area of Toronto. The box plot in Figure 3 shows transmission rate range in each cluster.

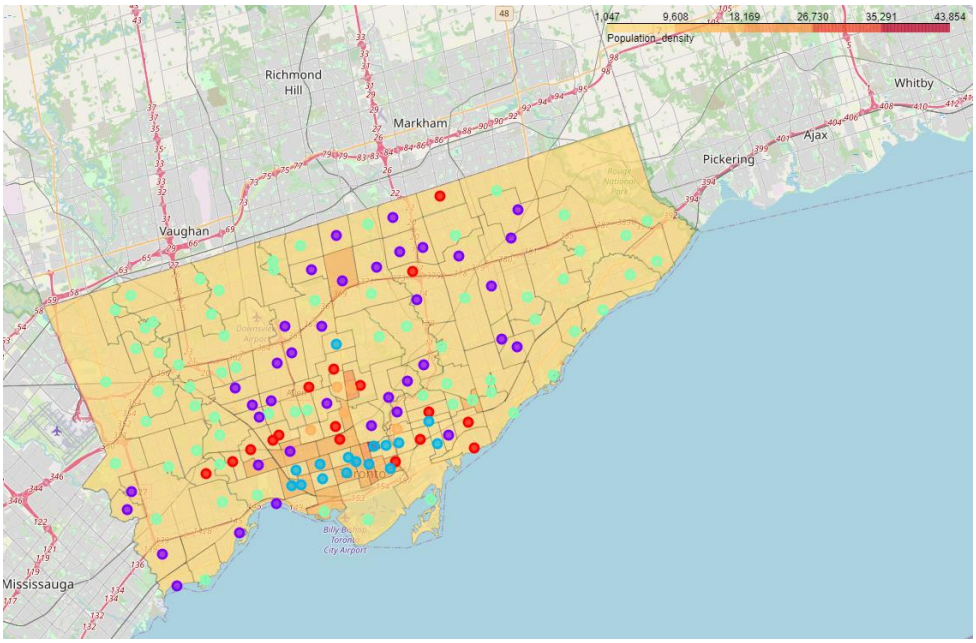

Figure 1 Population Density in Toronto

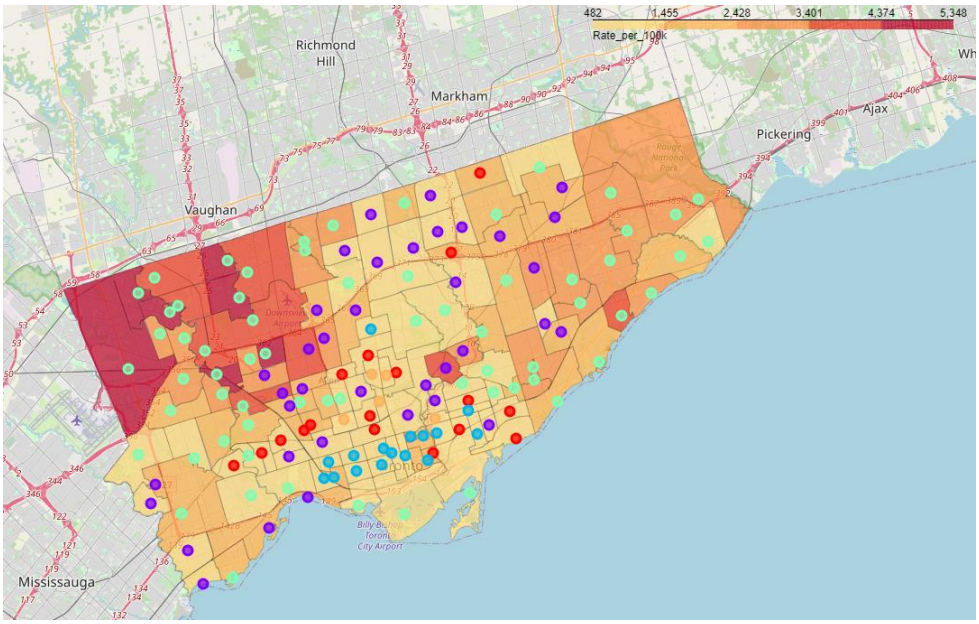

Figure 2 Transmission Rate in Toronto

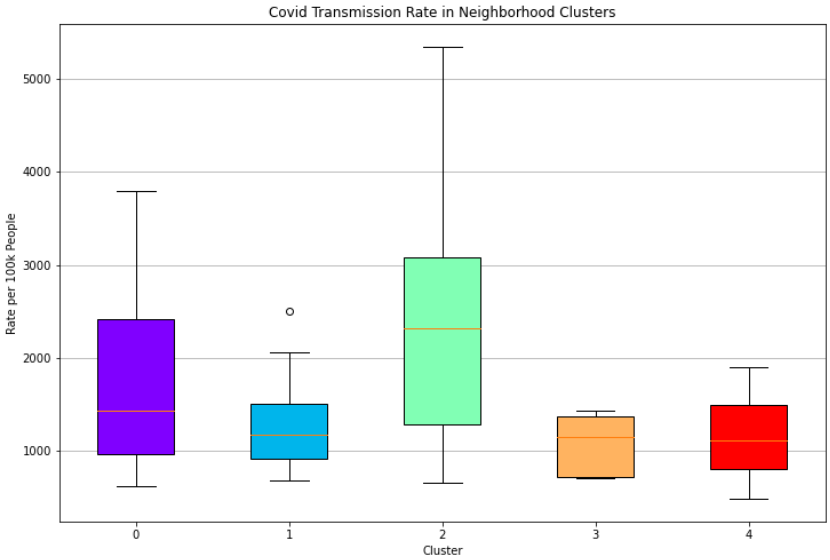

Figure 3 Box Plot of Transmission Rate in Each Cluster

Then we plot each cluster's venue density for all venue categories using bar plot (Figure 4 and Figure 5). Interestingly cluster green or purple does not have any venue that has significant higher density than other three clusters. In fact, cluster blue and yellow have very high venue density in food service, stores, coffee shops and bars because they are at the core of city. There seems to be little correlation between venue in Toronto neighborhoods and Covid-19 transmission rate based on the analysis we performed so far.

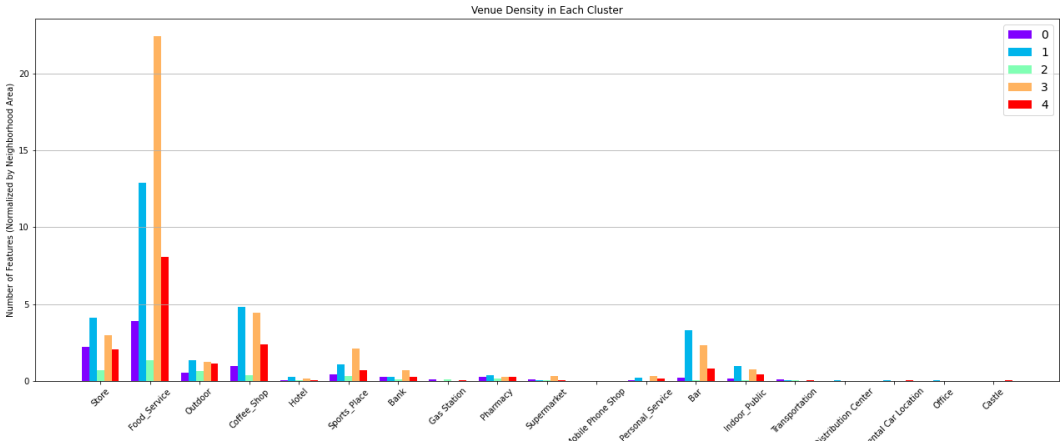

Figure 4 Bar Plot of Venues in Each Cluster

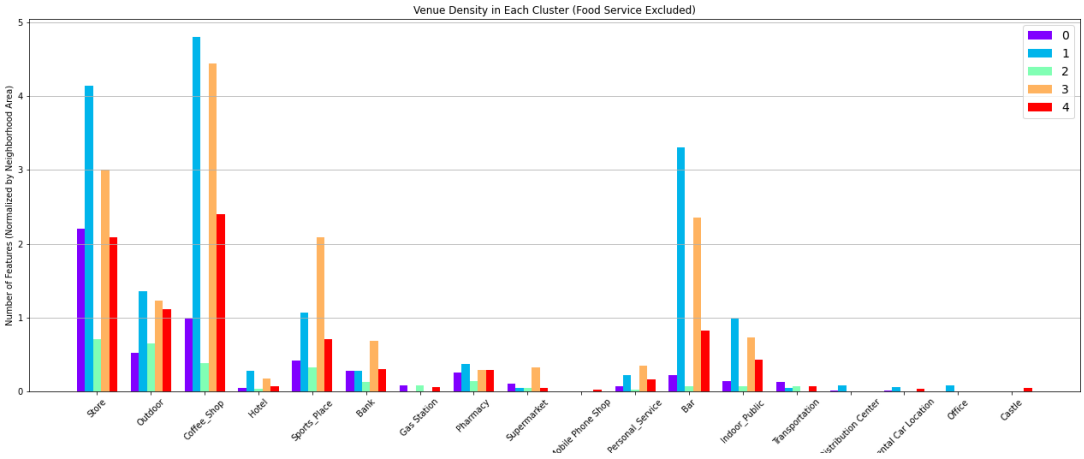

Figure 5 Bar Plot of Venues in Each Cluster (Food Service Excluded)

However we have to keep in mind that:

1. We used circle of equivalent area as boundary of neighborhood instead of the real boundary.
2. Foursquare limited number of venues per API call to 50. Therefore for each neighborhood we can only obtain 50 venues.

We also have to admit that there is chance that venues in neighborhood can be indeed irrelevant to transmission rate. However we need more data and probably improved analysis method to draw a solid answer.

5 Conclusion

In this report we explored the correlation between venues in Toronto neighborhood and Covid-19 transmission rate. Clustering algorithm and visualization methods such as choropleth map, box and bar plot were used to analysis data we have. We did not see strong correlation between venues in neighborhood and Covid-19 transmission rate. However this could be due to the approximation we took and the limitation of data we obtained.