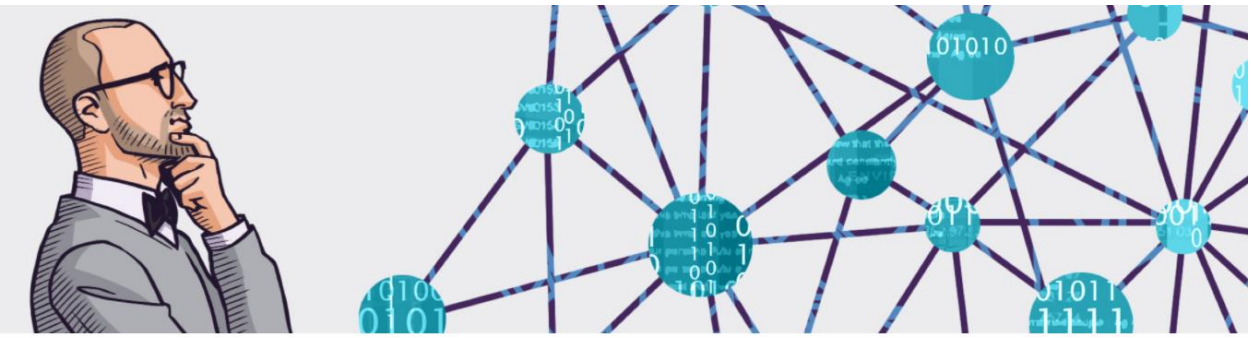

Study Covid-19 spread over Toronto city and identifying venues which are statistically significant in spreading using Foursquare API

Background

The World Health Organization (WHO) has declared the coronavirus disease 2019 (COVID-19) a pandemic. The virus is contagious and requires social distancing to minimize the spread. This even implies on travelling restrictions among different countries results in temperedly restrictions on travelling and sealing borders. Such strict measures bring tremendous stress to global economy. Many nations have also inforce lockdown situation within country in multiple cities. Despite strict measures, many nations still struggling to stop virus to spread. Hence it becomes extremely important to identify pattern of spreading viruses and identify if there is any key places or venues which could be a possible spreading hotspot. In this project we are focusing on identify such places which are statically significant in spreading the virus. The chosen city is Toronto to study Covid-19 spared rate among different neighborhood and find the key venues which are highly significant in spreading covid-19 virus.

Introduction: Problem Statement

Generally, within the proximity of a neighborhood there could be many venues like restaurants, food trucks, library, fitness center etc., but it is not certain among all these venues, what are the key places which is involved in spreading the virus and should be avoided to visit. This is the problem statement of our study to identify key places which are more significant in spreading the virus and should be avoided. This study is done using Toronto city data.

Data: Description

Data is collected from Toronto.ca website link https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto . Data is categorized based on different Neighborhood along with virus spread rate per 100K people and case count. Below figure is illustrating the data. Total numbers of rows are 140 and have 4 columns.

Table 1 Illustration of Covid-19 downloaded data

Neighborhood ID	Neighborhood Name	Rate per 100,000 people	Case Count
138.0	Eglinton East	430.277485	98
47.0	Don Valley Village	255.073750	69
38.0	Lansing-Westgate	222.717149	36
9.0	Edenbridge-Humber Valley	656.581912	102
44.0	Flemingdon Park	606.392194	133
...	...	...	...
113.0	Weston	1823.032459	328
95.0	Annex	281.727052	86
94.0	Wychwood	557.530142	80
37.0	Willowdale West	307.038262	52
76.0	Bay Street Corridor	220.955925	57

141 rows × 4 columns

In this study I have taken spread rate per 100K people to study. The spread rate varies within a wide rage among different neighborhood varies from 78 to 1823 with a standard deviation of 429.

The geo coordinates for each neighborhood is extracted from ***geopy*** library. And appended to dataframe as shown below

Table 2 Data with latitude and longitude

Neighborhood ID	Neighborhood Name	Rate per 100,000 people	Case Count	latitude	longitude
138.0	Eglinton East	430.277485	98	43.739465	-79.232100
47.0	Don Valley Village	255.073750	69	43.792673	-79.354722
38.0	Lansing-Westgate	222.717149	36	NaN	NaN
9.0	Edenbridge-Humber Valley	656.581912	102	43.672223	-79.514685
44.0	Flemingdon Park	606.392194	133	43.718432	-79.333204
...	...	...	...	...	...
113.0	Weston	1823.032459	328	43.700161	-79.516247
95.0	Annex	281.727052	86	43.670338	-79.407117
94.0	Wychwood	557.530142	80	43.682122	-79.423839
37.0	Willowdale West	307.038262	52	43.761510	-79.410923
76.0	Bay Street Corridor	220.955925	57	43.667342	-79.388457

141 rows × 6 columns

As it is seen data could be having missing parameter or NaN value. Hence to clean the data all the rows with NaN are removed. The filtered data is of shape row =104, and column = 6. The data is used to explore and create map superimposing on map of Toronto to highlight neighborhood with virus spread rate. For this folium library is used. First coordinates of Toronto city is extracted using ***geopy*** library. The coordinates is used to create base map using folium lib. The neighborhood data from table 2 along with latitude and longitude information is used to create markers superimposing on Toronto map. Below figure illustrating code used to do this operation

3.1 Create Map of toronto and superimposing neibhorhood with Covid-19 Rates per 100k people

```
address = 'Toronto City, Canada'

geolocator = Nominatim(user_agent="to_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto City are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Toronto City are 43.6534817, -79.3839347.
```

```
# create map of Toronto using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, neighborhood,CovidRate in zip(df_clean['latitude'], df_clean['longitude'], df_clean['Neighbourhood Name'],df_clean['Rate per 100,000 people']):
    label = '{},{}'.format(neighborhood,CovidRate)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Figure 1 Code illustrating how to create MAP superimposing virus spread rate over each neighborhood on Toronto map

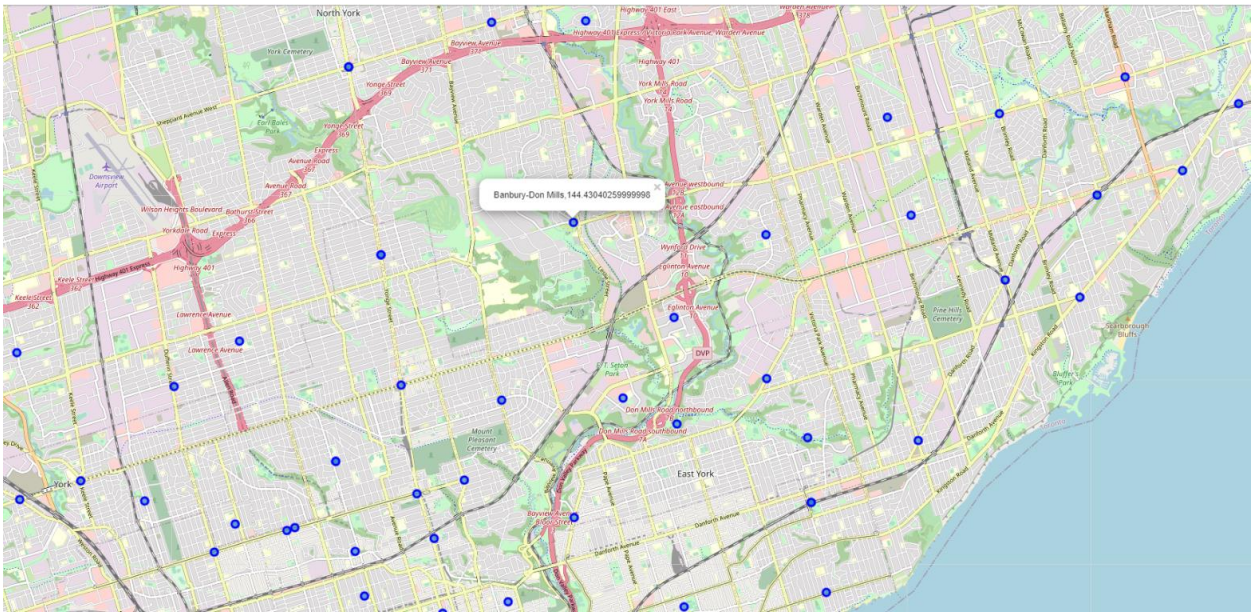

Figure 2 Generated MAP with Neighborhood information along with virus spread rate

Methodology: How Data is used to solve the problem

The primary goal of this project is to find venues which are statistically significant in spreading the virus. So the first requirement is to extract the venues located within 500 meter of each neighborhood, for this the data prepared in early steps (table 2) is used. The data has neighborhood along with latitude and longitude information. This information is used to extract the venues near that location using Foursquare API. The API provide list of venues located near given coordinates. For each venue present near a neighborhood a new row is appended to the data frame. Below table illustrating updated data frame with Venue name and Venue Category for each neighborhood. Shape of dataframe is row=2009, column=9.

Table 3 Data frame with Venue Category for each neighborhood

```
print(toronto_venues.shape)
toronto_venues.head()
```

(2009, 9)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Rate per 100,000 people	Case Count	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Eglinton East	43.739465	-79.2321	430.277485	98	Anjappar Authentic Chettinadu Restaurant	43.741592	-79.226799	Indian Restaurant
1	Eglinton East	43.739465	-79.2321	430.277485	98	Dairy Queen	43.739506	-79.236894	Ice Cream Shop
2	Eglinton East	43.739465	-79.2321	430.277485	98	Dairy Queen	43.739580	-79.236991	Ice Cream Shop
3	Eglinton East	43.739465	-79.2321	430.277485	98	Subway	43.738284	-79.236792	Sandwich Place
4	Eglinton East	43.739465	-79.2321	430.277485	98	Eglinton GO Station	43.739701	-79.232281	Train Station

The total number of unique venue category is 243. Data is further analyzed by sum up all the venues grouped by Neighborhood. This helps to understand neighborhood locality. Neighborhood densely placed with multiple venues might be more susceptible for spreading virus.

Table 4 Data Frame with total number of venue categories present within each neighbourhood

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Rate per 100,000 people	Case Count	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt North	27	27	27	27	27	27	27	27
Alderwood	6	6	6	6	6	6	6	6
Annex	42	42	42	42	42	42	42	42
Banbury-Don Mills	5	5	5	5	5	5	5	5
Bay Street Corridor	100	100	100	100	100	100	100	100
...	...	...	...	...	...	...	...	...
Wychwood	53	53	53	53	53	53	53	53
Yonge-Eglinton	72	72	72	72	72	72	72	72
Yonge-St.Clair	56	56	56	56	56	56	56	56
York University Heights	18	18	18	18	18	18	18	18
Yorkdale-Glen Park	12	12	12	12	12	12	12	12

To analyze each impact of each venue category, the data frame is transformed using one-hot coding for applied for each venue category. Below is the code involved and update data frame table.

```
# one hot encoding
toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
toronto_onehot = toronto_onehot[fixed_columns]

toronto_onehot.head()
```

	Yoga Studio	Afghan Restaurant	American Restaurant	Animal Shelter	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	...	Turkish Restaurant	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wings Joint	Women's Store
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 3 code illustrating code and updated data frame

Since the data is still have multiple rows for each neighborhood so next is to group the entire neighborhood and sum all the venue categories. These results update data frame as below figure 4.

3.3.1 Group rows by neighborhood and by adding each category

```
toronto_grouped = toronto_onehot.groupby('Neighborhood').sum().reset_index()
```

```
toronto_grouped
```

	Neighborhood	Yoga Studio	Afghan Restaurant	American Restaurant	Animal Shelter	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	...	Turkish Restaurant	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wings Joint	Women's Store
0	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	1	0
1	Alderwood	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Annex	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	1	0
3	Banbury-Don Mills	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Bay Street Corridor	1	1	0	0	1	1	0	0	1	...	0	1	0	1	0	0	0	0	0	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	Wychwood	0	0	1	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
90	Yonge-Eglinton	1	0	0	0	0	0	2	0	0	...	0	1	1	0	1	0	0	1	0	0
91	Yonge-St.Clair	1	0	1	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
92	York University Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
93	Yorkdale-Glen Park	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

94 rows × 243 columns

Figure 4 Data frame showing count of each venue categories located within each neighborhood

Next, this data frame is appended with Virus Infection rate as shown in figure 5

	Neighborhood	Yoga Studio	Afghan Restaurant	American Restaurant	Animal Shelter	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wings Joint	Women's Store	Infection Rate
0	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	1	0	291.966
1	Alderwood	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	365.024
2	Annex	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	1	0	281.727
3	Banbury-Don Mills	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	144.43
4	Bay Street Corridor	1	1	0	0	1	1	0	0	1	...	1	0	1	0	0	0	0	0	2	220.956
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	Wychwood	0	0	1	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	557.53
90	Yonge-Eglinton	1	0	0	0	0	0	2	0	0	...	1	1	0	1	0	0	1	0	0	135.398
91	Yonge-St.Clair	1	0	1	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	199.553
92	York University Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	1554.74
93	Yorkdale-Glen Park	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1391.52

Figure 5 Data Frame with venue categories as independent variable and Infection rate as dependent variable

Analysis: Statistical Analysis

To analyze data showing figure 5. An Ordinary Least Square Regression technique the model is used from ***Statsmodels*** library. The model is to predict infection rate using count of various venue categories. This makes the dependent variable ‘y’ as Infection rate and all the venues categories as independent variable represented by ‘X’.

4. Data Analysis

4.1 Model Ordinary Least Square Regression using Venues Categories as independent variable and Covid-19 Infection Rate as dependent variable using statsmodels lib

```
import statsmodels.api as sm
X = toronto_grouped.iloc[:, 1:-1].astpye(float)
y = toronto_grouped['Infection Rate'].astype(float)
```

```
tX = sm.add_constant(X)
est = sm.OLS(y, tX).fit()
est.summary()
```

Dep. Variable:	Infection Rate	**R-squared:**	0.972
Model:	OLS	**Adj. R-squared:**	0.562
Method:	Least Squares	**F-statistic:**	2.374
Date:	Fri, 14 Aug 2020	**Prob (F-statistic):**	0.137
Time:	09:10:24	**Log-Likelihood:**	-531.39
No. Observations:	94	**AIC:**	1239.
Df Residuals:	6	**BIC:**	1463.
Df Model:	87		
Covariance Type:	nonrobust		

OLS Regression Results

Result: Discussion

The computed RSquare value by model is ~0.972 and total of 15 categories are filtered with p-Value <0.05 which represent statistical significance in predicting the infection rate.

Below is the list of all the 15 categories in table 5.

Table 5 List of filtered categories having significance in predicting infection rate

index	P-value	
0	Bookstore	0.036599
1	Boutique	0.016180
2	Brewery	0.029301
3	Burger Joint	0.029436
4	Cuban Restaurant	0.029301
5	Department Store	0.040129
6	Farmers Market	0.040342
7	Fish Market	0.011843
8	Furniture / Home Store	0.004812
9	Gift Shop	0.040856
10	Karaoke Bar	0.008972
11	Modern European Restaurant	0.024861
12	Plaza	0.048322
13	Poke Place	0.020364
14	Vietnamese Restaurant	0.046480

The analysis and result is limited based on data and it could be possible that there are many more variables responsible of spread like population, which leaves opportunity to explore this idea to further extent.

Conclusion

In this study Covid-19 Data provided by Toronto.ca been analyzed. The data was having 140 Rows for different neighborhood with infection rate for each neighborhood. The Infection rate is provided per 100K people. The infection rate among different neighborhood varies in a wide range with min value of 78 and max value of 1823 with a standard deviation of 429. The data is further annotated by geo coordinates using geopy lib for each neighborhood lib. Using these coordinates a map is created and studied by superimposing different neighborhood with Infection rate over a map of Toronto. These geo coordinates further used to gather list of venues located within 500 meter for each neighborhood using foursquare lib. A total of 243 unique venue categories were identified in the data. To analyze these venue categories data is transformed using one-hot coding and added the similar venues categories. This data is modeled using Ordinary Least Square Regression Model from Statsmodels. All the venue categories count is used as independent variables and the Infection rate is used as depended variable. The model showed the R2squared value of ~0.97 and provided p-Values for each venue category. The p-Values represent the significance of independent variable in predicting dependent variable. If a p-Value of an independent variable is < 0.05 then it is highly significance in predicting the depended variable. With this exercise all the independent variables with $p<0.05$ are filtered and reported to full fill the goal of this project. There are total 15 venue categories which are found to be highly significant in spreading the virus as listed below:

Bookstore, Boutique, Brewery, Burger Joint, Cuban Restaurant, Department Store, Farmers Market, Fish Market, Furniture / Home Store, Gift Shop, Karaoke Bar, Modern European Restaurant, Plaza, Poke Place, Vietnamese Restaurant