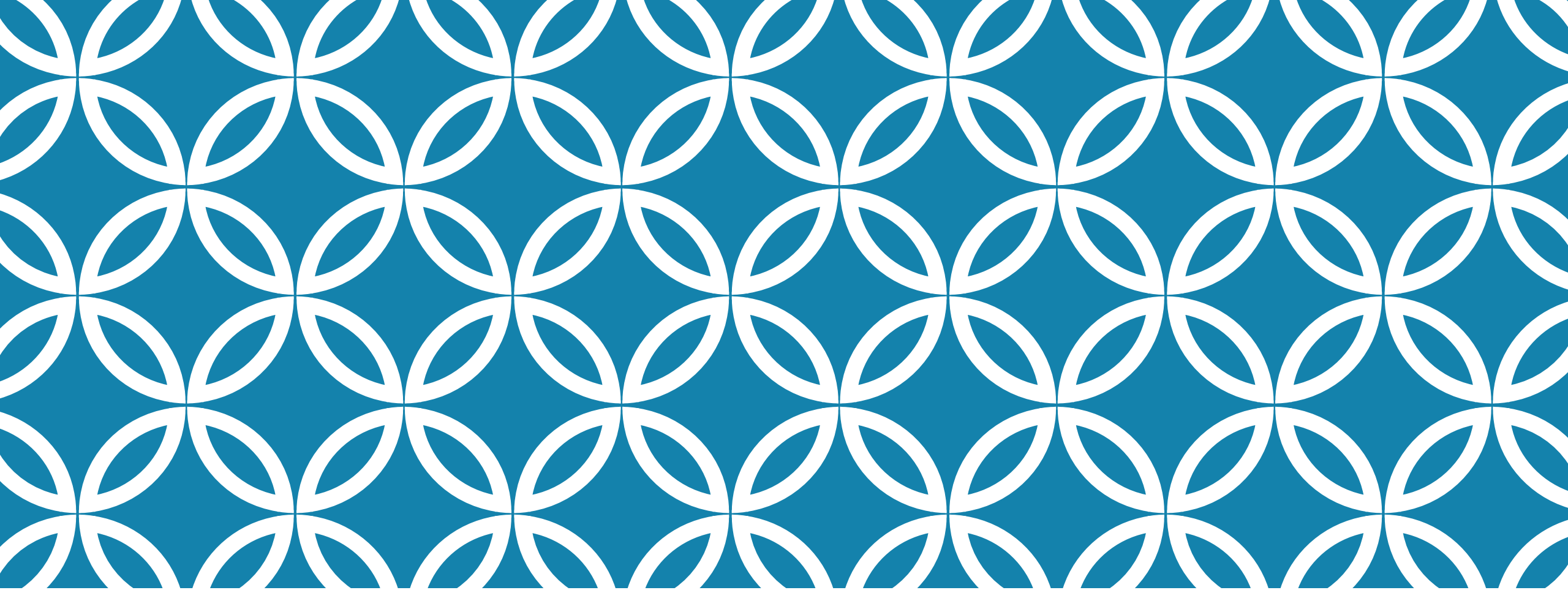


ASSIGNMENT – THE BATTLE OF NEIGHBORHOODS

Mohammad Sulaiman

December 8, 2020

BACKGROUND

Toronto is the capital city of the Canadian province of Ontario. It is Canada's largest city. Its large population of immigrants from all over the globe. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto is known for its many skyscrapers and high-rise buildings.

The global COVID-19 pandemic has affected several countries in the world with Canada being one of them too. On February 24, Toronto's third and Ontario's fourth case of COVID-19 was confirmed after a woman tested positive following travel to Wuhan. Toronto Mayor John Tory announced a local state of emergency on March 23, 2020.

PROBLEM

The big question is which neighborhood should one choose in Toronto amidst this pandemic?
In this project, I will discuss different parameters to consider and use foursquare location data and clustering of venue information to determine a less impacted neighborhood in Toronto.

TARGET AUDIENCE

Anyone who is planning to move to the city of Toronto.

Students in Toronto who are looking for off-campus housing.

Anyone who is planning to move to another location in the city of Toronto.

DATA

All data related to locations and will be obtained via the FourSquare API utilized via the Request library in Python.

Toronto City data containing the neighborhoods and boroughs, latitudes, and longitudes will be obtained from the data source.

All data related to locations and will be obtained via the FourSquare API utilized via the Request library in Python.

DATA CLEANING

Code snippet

```
df_merged = pd.merge(df_comb, df_comb2, on='Borough')

df_merged.head()
```

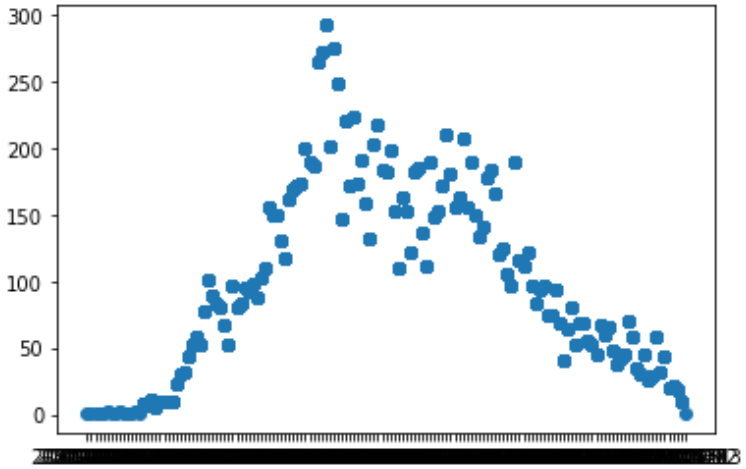
	Postal Code	Borough	Neighborhoods	Latitude	Longitude	CDN	Neighborhood	Case Count	Rate per 100,000 people
0	M3A	North York	Parkwoods	43.753259	-79.329656	42.0	Banbury-Don Mills	37	133.598122
1	M3A	North York	Parkwoods	43.753259	-79.329656	34.0	Bathurst Manor	126	793.800794
2	M3A	North York	Parkwoods	43.753259	-79.329656	52.0	Bayview Village	33	154.234436
3	M3A	North York	Parkwoods	43.753259	-79.329656	49.0	Bayview Woods-Steeles	116	881.861031
4	M3A	North York	Parkwoods	43.753259	-79.329656	39.0	Bedford Park-Nortown	81	348.597005

EXPLORATORY DATA ANALYSIS

The total number of cases per episode date was obtained using the following code:

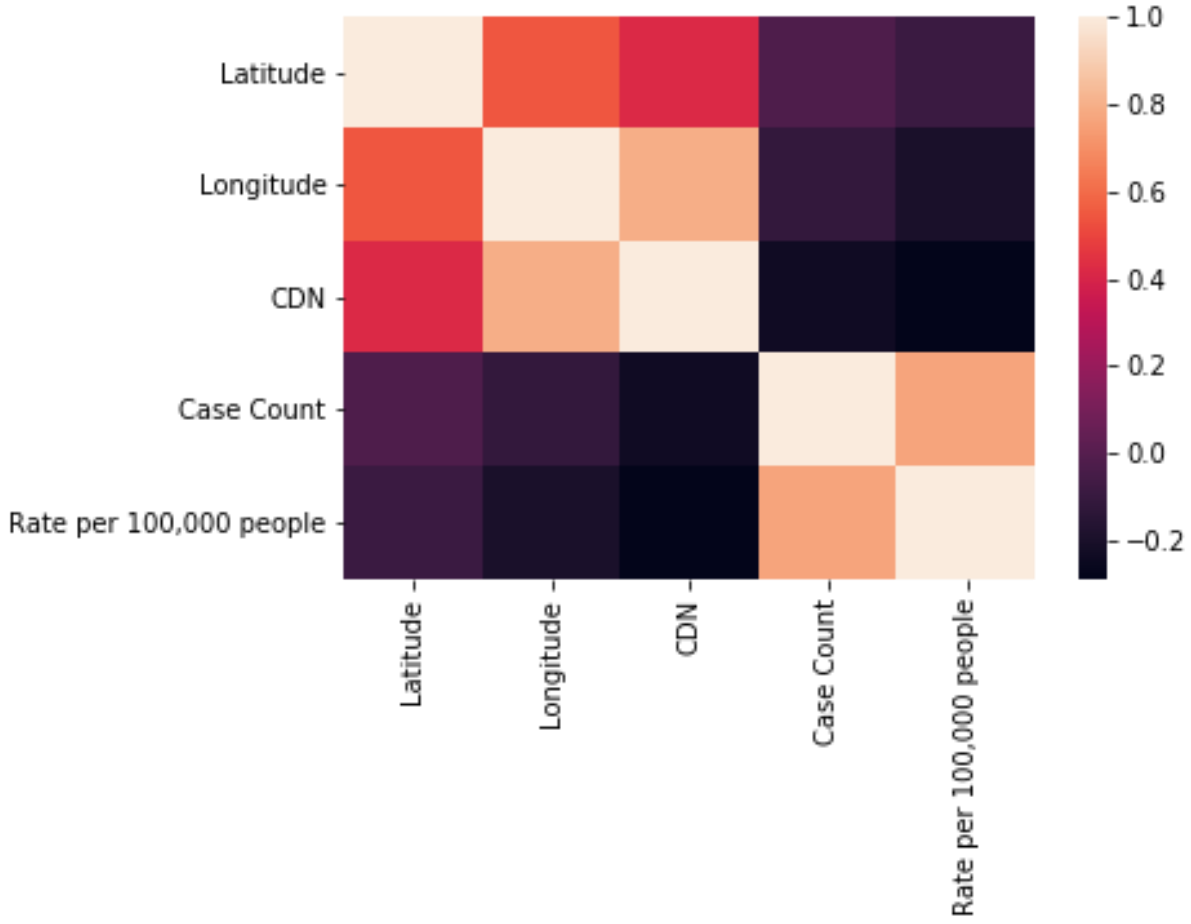
```
total_cases = df_cases['Episode Date'].value_counts()
total_cases = pd.DataFrame(total_cases)
total_cases.reset_index(level=0, inplace=True)
total_cases.columns = ['Episode Date','Total Cases']
total_cases.sort_values('Episode Date',inplace=True)
total_cases.head()
```

A scatter plot was obtained to analyze the relationship between total number of cases and the episode date:



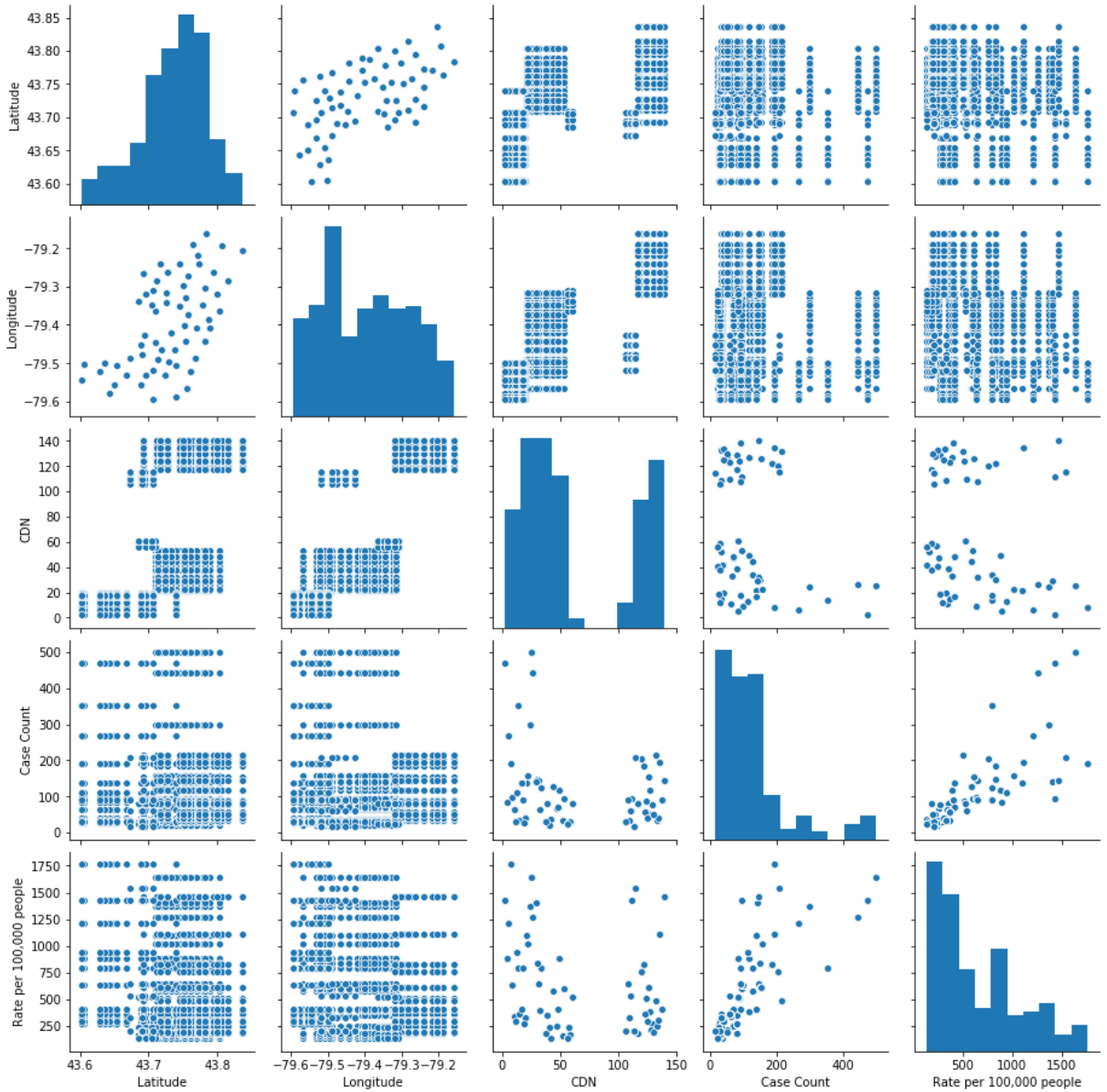
HEAT MAP

A heat map and a pairplot were made of all the factors to visually gauge average correlation of COVID-19 cases and factors in the dataset



PAIRPLOT

From the pairplot we can roughly tell that the neighborhoods (CDN numbers) were clustered based on the number of cases and rate per 100,000 people.



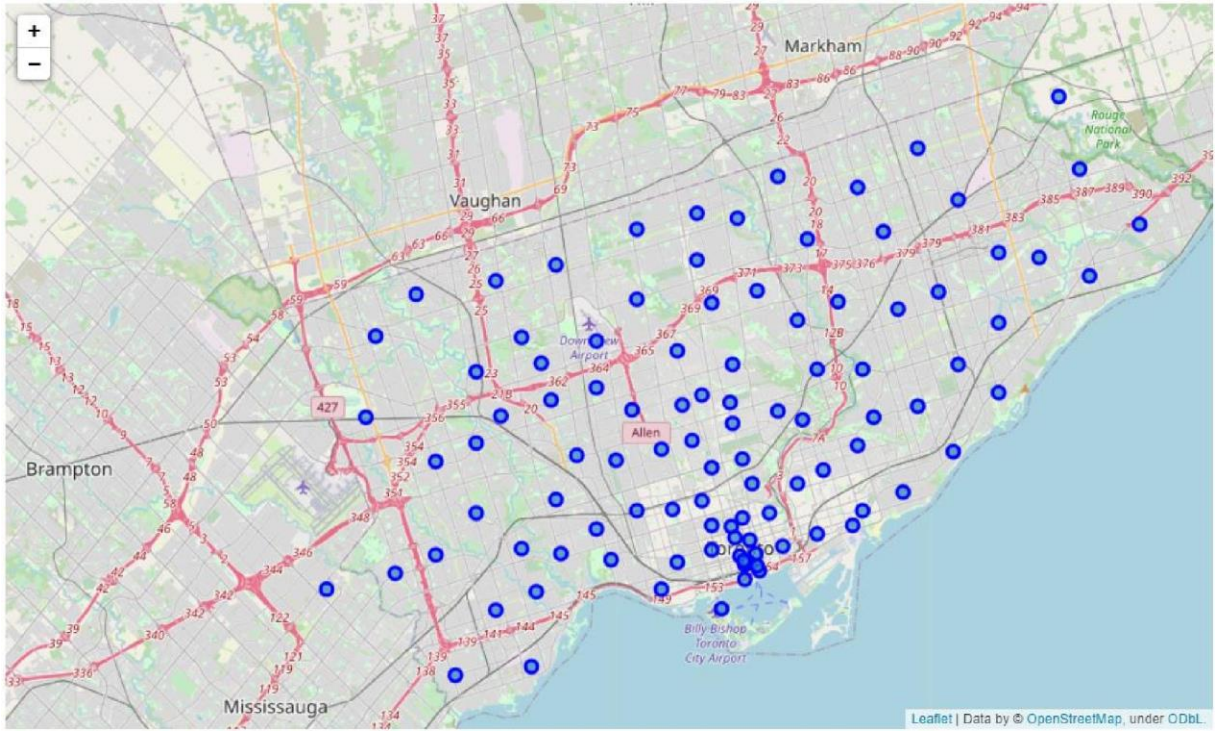
GEOLOCATION

Google maps geocoder API was used to obtain the latitude and longitude values of the city of Toronto.

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

FOLIUM

Folium was used to visualize the data in an interactive leaflet map. A map of the city of Toronto with the neighborhoods superimposed on it was created.



FOURSQUARE API

I utilized the Foursquare API to explore the boroughs and search for hospitals. I searched for hospitals around each borough in a 100 meter radius from their given latitude and longitude values.

```
def getNearbyVenues(names, latitudes, longitudes, radius=100):

    categoryId = '4bf58dd8d48988d196941735'
    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/search?&client_id={}&client_secret={}&v={}&ll={},{}
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT,
            categoryId)

        # make the GET request
        results = requests.get(url).json()["response"]

        # return only relevant information for each nearby venue
    venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])

    return(nearby_venues.head(5))
```

K MEANS CLUSTERING

After running K Means a number of times I observed that the optimal value for K was 4.

```
df_features = df_merged[['Rate per 100,000 people','Case Count']]
# set number of clusters
kclusters = 4

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(df_features)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]

df_merged['Cluster Labels'] = kmeans.labels_
```

RESULTS AND DISCUSSION

On the evaluation of the clusters I devised suitable labels for each cluster:

In Cluster labeled 1 the case count was below 150 for all neighborhoods and the rate per 100,000 approximately lied between 100-420. Therefore, its label name would be “Least Infected”.

In Cluster labeled 3 the case count was below 300 for all neighborhoods and the rate per 100,000 approximately lied between 450-700. Therefore, its label name would be “Considerably Infected”.

In Cluster labeled 0 the case count was below 400 for all neighborhoods and the rate per 100,000 approximately lied between 700-1200. Therefore, its label name would be “Highly Infected”.

In Cluster labeled 2 the case count was below 500 for all neighborhoods and the rate per 100,000 was above 1200 for all neighborhoods. Therefore, its label name would be “Extremely Infected”.

CONCLUSION

In this study after carefully analyzing COVID-19 in the city of Toronto from various datasets and using various visualizations and K Means Clustering I clustered neighborhoods based on case count and rate per 100,000.

Currently COVID-19 has a significant impact globally and therefore it is a major factor to take into account when deciding a neighborhood. Therefore, this model is useful to narrow down one's choice of neighborhood in the city of Toronto.

However, this model only clusters the neighborhoods based on the COVID-19 pandemic and other factors need to be considered when deciding a neighborhood for instance the price, distance from workplace/school, safety etc.

As I mentioned earlier the data is subject to change thus the results of this model are also subject to change.

RESULTS AND DISCUSSION

On the evaluation of the clusters I devised suitable labels for each cluster:

In Cluster labeled 1 the case count was below 150 for all neighborhoods and the rate per 100,000 approximately lied between 100-420. Therefore, its label name would be “Least Infected”.

In Cluster labeled 3 the case count was below 300 for all neighborhoods and the rate per 100,000 approximately lied between 450-700. Therefore, its label name would be “Considerably Infected”.

In Cluster labeled 0 the case count was below 400 for all neighborhoods and the rate per 100,000 approximately lied between 700-1200. Therefore, its label name would be “Highly Infected”.

In Cluster labeled 2 the case count was below 500 for all neighborhoods and the rate per 100,000 was above 1200 for all neighborhoods. Therefore, its label name would be “Extremely Infected”.

THANK YOU