IBM Data Science Capstone Project

# Title: Relationship between Covid-19 and neighborhood venues

# Objective:

The goal of this project is to explore the relationships, if any, between neighborhood venues and Covid-19 infection and death rates.

New York City is a diverse city with many distinct neighborhoods. Neighborhood venues are a reflection of the life-style within each neighborhood, and life-style may have a connection to how Covid-19 infected people through community spread. For example, several states that saw spiking Covid-19 infection rates have decided to shutdown bars and indoor dinning. Additionally, life-style choices have health implications on the community living in each neighborhood. Community health, in turn, has an impact on infection rate and death rate.

# Target Audience:

This analysis is intended to inform public health officials and hopefully trigger beneficial policy changes. For example, one category of venues used in this analysis is the “Take\_ out\_food” category (e.g., McDonald’s, cake shops) . These venues have little or no seating space, therefore customers generally order something and take it away to consume. So far, public health officials have not identified this type of venues as high risk. However, these venues generally have high volume of traffic, so more people can potentially come into contact with an infected employee, for example. Moreover, take out foods are often associated with bad health outcomes which can cause the neighborhood population to be more vulnerable to infection and death.

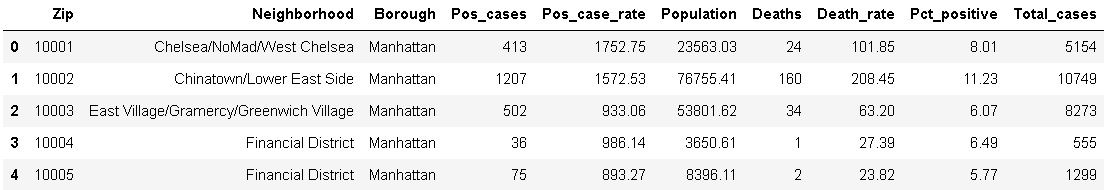
# Data

The New York City (NYC) Covid-19 data is organized by Modified Zip Code Tabulation Areas (MODZCTA). This is also the method used by the US Census Bureau. Therefore, the data required are:

* NYC Covid-19 infection and death data by MODZCTA ([ctl-click here](https://raw.githubusercontent.com/nychealth/coronavirus-data/master/data-by-modzcta.csv)).
* A mapping from MODZCTA to longitude and latitude (geospatial data) for extracting venue data from FourSquare. This data is available by state, [ctl-click here](https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?refine.state=NY&dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6InVzLXppcC1jb2RlLWxhdGl0dWRlLWFuZC1sb25naXR1ZGUiLCJvcHRpb25zIjp7InJlZmluZS5zdGF0ZSI6Ik5ZIn19LCJjaGFydHMiOlt7ImFsaWduTW9udGgiOnRydWUsInR5cGUiOiJjb2x1bW4iLCJmdW5jIjoiQVZHIiwieUF4aXMiOiJsYXRpdHVkZSIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlLCJjb2xvciI6IiNGRjUxNUEifV0sInhBeGlzIjoic3RhdGUiLCJtYXhwb2ludHMiOjUwLCJzb3J0IjoiIn1dLCJ0aW1lc2NhbGUiOiIiLCJkaXNwbGF5TGVnZW5kIjp0cnVlLCJhbGlnbk1vbnRoIjp0cnVlfQ%3D%3D&location=7,42.79,-75.84997&basemap=jawg.streets) for New York State data.
* Venues data from FourSquare extracted by MODZCTA.

Going forward, let’s use Zipcode instead of MODZCTA.

Here is a sample of the NYC Covid-19 data:



Here is a sample of the longitude and latitude data:



Here is a sample of venues data downloaded from FourSquare:



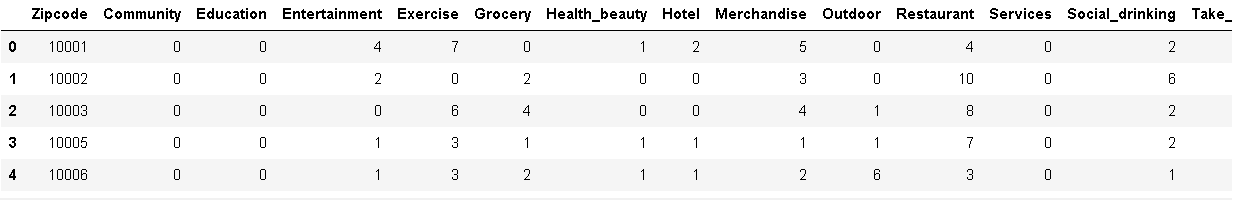
# Data Cleansing and Transformation

After merging the NYC Covid-19 data with the NY State geospatial data, it was discovered that four Zipcodes did not have longitude and latitude mappings. These were removed from the dataset. Additionally, one Zipcode’s (11237) geospatial data was rejected by FourSquare for unknown reasons. This is also removed from the dataset.

One transformation was needed to make the data more useable. FourSquare provided a total of 332 categories of venues across NYC. This is too granular. These 332 categories were condensed into the following 14 categories:

| **Category Name** | **Examples** |
| --- | --- |
| Community | Community center, church, synagogue |
| Education | School, university, professional training center |
| Entertainment | Movie theater, music hall, museum |
| Exercise | Sports center, gym, tennis facility |
| Grocery | Supermarket, grocery store, liquor store |
| Health\_beauty | Hospital, doctor’s office, beauty salon, barber shop |
| Hotel | Hotel, hostel |
| Merchandise | Arts and craft store, gift shop, department store |
| Outdoor | Parks, squares, attractions |
| Restaurant | Sit down restaurants, Italian, Japanese, African |
| Services | Bike rental, pet service, post office |
| Social\_drinking | Bars, lounges, beer hall |
| Take\_out\_food | McDonald’s, café, cake shop, take out burger |
| Transportation | Bus station, train station, airport |

After condensing the categories and using one-hot to create a separate column for each category, the data looks like:



Lastly, the venues data will be transformed into per-capita data through division by the population for each Zipcode. Population data is available from the NYC Covide-19 dataset. The resulting data is of the same format as above.