

## B. Data Acquisition and Cleaning

### B.1 Data Sources

To tackle our problem, I used the following data sources:

- Foursquare API was utilized to pull venue data for all Manhattan neighborhoods.
- NYC Health Coronavirus Data was utilized to sort COVID-19 data by ZCTA (Zip Code Tabulation Areas) Neighborhoods.

Using these two data sources, we can perform many tests to solve our problem. First, we can cross reference the two data sources by using neighborhood to check for correlations of high COVID-19 spread to certain venues. Once this is established, we can further this analysis by using K-mean clusters to first group similar neighborhoods together in order to check if the COVID-19 spread is uniform across members of the same cluster.