**Identifying High-Risk Customers to Provide Better Health Coverage**

Juliana McCausland

The purpose of this assignment was to utilize MTA turnstile data to help identify individuals who may have a higher risk of exposure to infectious diseases so that their health insurance coverage may be adjusted accordingly. I used two additional datasets to give some insight into the locations of the MTA stations, as well as the locations of COVID-19 hot spots. I also created two visualizations using Tableau. Overall, there were no clear-cut results. If anything, this analysis provided insight into the number of factors this health insurance company must take into consideration to understand an individual's risk of contracting an infectious disease.

## Design

The backstory is that an imaginary health insurance company would like to better serve its customers. With the knowledge that tightly-packed places like a train are hotbeds for infectious diseases like COVID-19, they began collecting data about their customers' daily commutes to better understand who is more likely to be exposed in a time of crisis. They requested this analysis to contextualize the data they had collected on their customers. This analysis serves to recommend next steps, and narrow down the categorization of a high-risk individual.

## Data

MTA turnstile data for the year of 2019 and the year of 2020 was used. The data from the year of 2019 serves as a reference for typical life/commutes, and 2020 serves as a reference for a time of crisis. I also used data containing the geographic (latitude and longitude) information for NYC MTA stations and data with information about COVID rates by NYC zip code. I limited the geographic MTA data to only the top ten stations (in terms of traffic). I limited the COVID data to the top 15 locations (in terms of COVID cases).

## Algorithms

I primarily used Pandas and SQL. I cleaned the data, using the Pandas dropna function, and also dropping outliers that I found while exploring the data further. I created new columns for the MTA data, including a datetime object, and columns for the day of the week and the week of the year. I calculated the daily entries per turnstile, as well as the total number of entries per station across the span of each year (2019 and 2020). This allowed me to see the entry traffic at each station, and identify the stations with the most entries. I created categorical plots using seaborn and the daily entries for each of the most populated stations. I was also able to create a categorical plot of the traffic during each time frame for each of the most populated stations (using seaborn as well).

I attempted to create a map visualization using geopandas, and spent significant time on this, but it did not work out. I was able to plot the geographical locations of the train stations, but I could not overlay them onto a basemap using code, so I resorted to using Tableau.

I additionally manipulated data using SQL and SQLalchemy. This was especially useful for cleaning up the additional datasets I used in this project. I also used SQL to create databases and tables (within my terminal).

**Tools**

- Pandas, SQL, Tableau, Matplotlib, Seaborn