# Abstract on Data

## Data Sources

Our team investigated three data sets as candidates for modeling future effects of COVID-19 in areas resembling the state of New York. The first of the three was the Centers for Disease Control and Prevention (CDC, ref. https://data.cdc.gov). The CDC data while extensive did not have any attributes for location, which was vital for us to create a model for our main area of interest (New York) and then trying to contrast it with a similar location (e.g., Florida, Texas, California).

The next data source we vetted was the Johns Hopkins University Center for Systems Science and Engineering (ref. https://github.com/CSSEGISandData/COVID-19/tree/master/csse\_covid\_19\_data/csse\_covid\_19\_time\_series). The data set was a strong contender for modeling, but we found the volume of the data (424 features) did not come without the need for a lot of data manipulation. This included filtering out hundreds of unnecessary columns, a pivot execution in both directions for values and columns needing restructure, multiple data cleansing tasks including imputations, and finally we would still be left with a heavy data source for extracting and using for computations.

This finally left us with our selected data set, The COVID Tracking Project at the Atlantic (ref. https://covidtracking.com/data). The data was readily available for us to extract via a web API and since it is a compilation of data brought together by the Atlantic team of data scientists and analysts, the data for the most part was clean and manageable. Inclusive of this we included data to explore public transportation ridership from the Metropolitan Transportation Authority (ref. https://new.mta.info/coronavirus/ridership) in New York. To explore additional effects of behavior we also added an attribute depicting what dates in our time series fell on dates that were a national holiday (ref. https://www.opm.gov/policy-data-oversight/pay-leave/federal-holidays/) or dates of known seasonal gatherings such as spring break which garners large crowds of people, increasing the risk of spreading a virus.

## Methodology

Our platform for this project is on Python version 3.8. The main libraries used for the data wrangling portion is "pandas", "io", and "matplotlib" with "seaborn" for creating EDA extrapolations and visualizations. IDE choice varied per team member based on preference - for example PyCharm by Jet Brains and Spyder open-source software were used during development. Microsoft's GitHub was used as version control repository for project documents, resources, and executable code.

The COVID Tracking Project at the Atlantic supplies a public web API for data consumption. We were interested in region specific data sets and fortunately had the option to supply the query string of the API URL with any choice name of a US state. Unfortunately, the hosting of this resource had a schedule to be taken down on May 1, 2021. To make our code robust to this future take down, on every data pull we backed up the data to a physical file and included a "Try-Catch" block to use that physical file in the event the API raised an error or simply did not return any data.

The data itself had a manageable structure and clean data. There were many features excluded as a result of not being applicable such as "date modified" or having one value for every observation - for example, a variable with exclusively "NA" values. In conclusion, we considered half of the 56 original columns and joined two columns to the time series, a binary feature indicating time of holiday, and a numeric value to indicate public transportation ridership from the MTA of New York.

Our first secondary data of holiday data was corroborated online and by common knowledge into a hard coded list. We then looped through the time series and if a timestamp fell in that list of holiday times a binary value was inserted to indicate such. The next secondary data, New York's MTA ridership time series, was able to be joined simply by using the date index.

In conclusion after every run, we back up the final data set to a physical CSV file as a disaster recovery method in lieu of our various data sources being corrupted or absent.