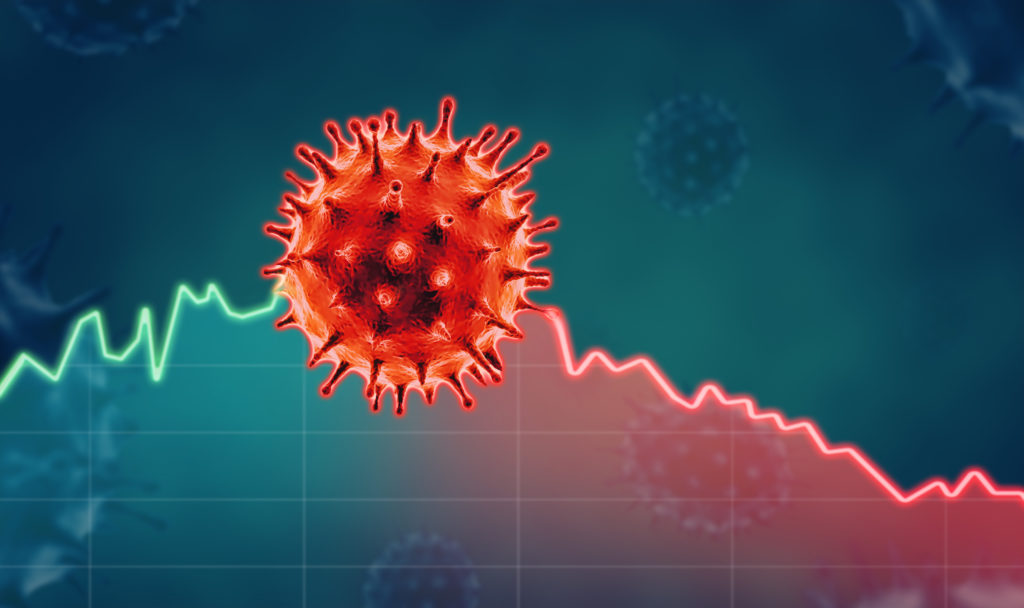
IAC- 622 Project Stage – 1

Data and Project Understanding





Bala Mallampati

Neethu Sreerangam

Samip Thapa

**Introduction**

As COVID-19 pandemic continues to spread, many people around the world have been affected in many different ways. We can understand patterns of COVID-19 effect and the spread by designing and developing an analytical framework with real life data. The data used for this report is COVID-19 data and relevant impact data, which reflects aspects of demographics, socioeconomics, employment, politics, and hospital beds as enrichment datasets. These datasets will provide additional insight into the American municipalities. This project provides us with opportunities to not only demonstrate our technical skills and knowledge in data science, but also to turn uncovered scientific outcomes into insightful knowledge and information that will be highly beneficial to public health and quality of living.

**Primary COVID-19 Data**

**Variable Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Variable | Description | Data type |
| Covid Confirmed Cases and  Deaths Cases | countyFIPS | County FIPS code, which is a five-digit that stands for Federal Information Processing Standards | integer |
| (both datasets have same variables) |  | (FIPS uniquely identifies [counties](https://en.wikipedia.org/wiki/County_(United_States)) and county equivalents in the [United States](https://en.wikipedia.org/wiki/United_States), certain  U.S. possessions, and certain freely associated states.) |  |
|  | County Name | County name | string |
|  | State | State name | string |
|  | StateFIPS | Two-letter alphabetic [codes](https://en.wikipedia.org/wiki/Codes) to identify U.S. states and certain other associated areas | integer |
|  | Specific dates (1/22/20 ~ 2/8/21) | Count of confirmed cases on the specific date | integer |

|  |  |  |  |
| --- | --- | --- | --- |
| Population | countyFIPS | County FIPS code, which is a five-digit that stands for Federal Information Processing Standards  (FIPS uniquely identifies [counties](https://en.wikipedia.org/wiki/County_(United_States)) and county equivalents in the [United States](https://en.wikipedia.org/wiki/United_States), certain  U.S. possessions, and certain freely associated states.) | integer |
| County Name | County name | string |
| State | State name | string |
| population | Population each county | integer |

**Merging process of dataframes:**

In merging data for this project, the merging process consists of two steps. First, confirmed and death cases data frames are reshaped from wide forms into vertically long forms. The wide form of the dataframe summarizes confirmed or death cases each date like a pivot table with extended data with summary. So, its form does not allow us to merge the data frames in the way we want to combine them. We create two variables.

One represents dates and the other indicates confirmed or death counts for each date correspondingly.

The *County Population Metrics Dataset* had only one unique column that needed to be included in the merged dataset. The *Confirmed Cases Dataset* included multiple dated entries which needed to be addressed similarly in the *Death Cases Dataset.* These two datasets also contained similar data and structures. In looking at the datasets, the most logical approach was to merge the data frames that have similar structures, *County Population Metrics Dataset* and *Confirmed Cases Dataset*. So after changing the structure of these datasets it is important to note the need to pay attention to the different dimensions between the newly merged dataset and the County *Population Metrics Dataset* in combining these datasets. The three datasets shared the variables *countyFIPS* and *State*.

**Preliminary Observations (intuition from the data)**

Both confirmed and death cases each county seems to exponentially grow over time given the time period. They are highly likely to be correlated to each other In looking at the data, the data presented is from January 22, 2020 to February 8, 2022. It shows the increase in the number of confirmed cases and the number of deaths in the various counties across America. The population ranges from 0 to over 10,000,000 people.

Underlying data sets have been leveraged to facilitate analysis of the independent variables that have an influence on the progress of the COVID-19 pandemic. In tandem with the variance of population density and socioeconomic status throughout the country, newly discovered trends, such as the political majority by state gives a picture of how the human element is steering the direction of the COVID-19 outbreak.