* [Code GitHub](https://github.com/shreyasagrawal/Covid-County-Project.git)
* [Datasets Dropbox](https://www.dropbox.com/scl/fo/6i2dcck84cdhwylavss69/h?dl=0&rlkey=upj9bty4q30cq3q6eb7ho7wdv)

**Question:** What are the key variables that predict the COVID-19 mortality rate **by FIPS code**?

**Methods:** Multiple linear regression and random forest

**Response variable:** Case Fatality Ratio (= #Deaths / #Cases) by FIPS code

**Explanatory variables:** As listed in the following sections

**Project Plan by Steps:**

1. Data cleaning regarding the selected variables for each dataset. We will integrate all datasets together by week and FIPS code. N/A datapoints (only a few for each dataset) and cases without sufficient number of observations will be deleted to reduce the uncertainty. Average or sum calculations will be considered as per the class for each variable.
2. Appropriate plots will be produced to observe the general trend of variables. Outliers and some outstanding cases may be identified and to which we will pay extra attention on an ongoing basis.
3. The following models will be generated to identify the best variables for the model. All models will be tested for assumptions (i.e., multicollinearity, homoscedasticity, and equal variance) through the consistency of R-sq’s, ANOVA tests, and multiple plots. A reduced model will be generated after performing the following methods.
   1. Multiple regression models: stepwise regression & best subsets regression
   2. Random forest method
4. (IF time’s allowed) Creating a county-specific comprehensive metric based on the county’s demographic, sociodemographic, and healthcare datasets to help administration departments better allocate healthcare resources for areas in need to reduce CFR.
   1. Based on real life observations, healthcare usually takes a much higher patient load after holidays or weekends due to the extensive social activities and increasing unhealthy behaviors such as drinking and using substances. We wish to take special dates as mentioned into account for the ultimate prediction model.

**Data sources**

* [*Centers for Disease Control and Prevention*](https://www.bing.com/ck/a?!&&p=2e3ab699b8fcb82aJmltdHM9MTY4MTY4OTYwMCZpZ3VpZD0xYzJkMzlmYy0zODZjLTZlZGMtMWFiZC0yYjMwMzk5NTZmMDcmaW5zaWQ9NTcwMQ&ptn=3&hsh=3&fclid=1c2d39fc-386c-6edc-1abd-2b3039956f07&psq=CDC&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvQ2VudGVyc19mb3JfRGlzZWFzZV9Db250cm9sX2FuZF9QcmV2ZW50aW9u&ntb=1) *(CDC)*
  + [PLACES: County Data (GIS Friendly Format), 2021 release](https://chronicdata.cdc.gov/500-Cities-Places/PLACES-County-Data-GIS-Friendly-Format-2021-releas/kmvs-jkvx) ([*Data Dictionary*](https://www.cdc.gov/places/measure-definitions/index.html))
  + [COVID-19 Vaccination by county](https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh)
  + [COVID Patient Impacts and Hospital Occupancy](https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u)
  + [COVID Cases and Deaths Across Counties](https://www.nytimes.com/interactive/2021/us/covid-cases.html) (NY Times cited CDC)
* *U.S. Census Bureau*
  + [DP05 American Community Survey Demographics and Housing Estimates](https://data.census.gov/table?q=county+DP05&tid=ACSDP1Y2021.DP05&moe=false)
  + [DP03 Selected Economic Characteristics](https://data.census.gov/table?tid=ACSDP1Y2021.DP03)

**Data Collection**

**PLACES: County Data**

PLACES is a joint effort between the CDC and the Robert Wood Johnson Foundation to provide county-level health data for a better understanding of the geographic distribution of health measures.

PLACES uses data from Behavioral Risk Factor Surveillance System (BRFSS) data, Census Bureau data, and American Community Survey data to provide both crude and age-adjusted rates for various health measures.

For our project, we will be using 7 age-adjusted variables from data collected in 2020, 2021, and 2022 as explanatory variables. A list of useful variables for this project can be found in Appendix I.

**COVID-19 Reported Patient Impact and Hospital Capacity by Facility Data**

This dataset provides information on the total number of inpatient, outpatient, and ICU beds in each sample hospital, as well as the number of COVID-19 patients occupying these beds. The statistics are reported to the U.S. Department of Health and Human Services on a weekly basis by hospital facilities.

This dataset can be used to track the impact of COVID-19 on the healthcare system in the United States. It can also help identify areas with high numbers of COVID-19 hospitalizations and monitor changes in hospital capacity over time, which could be a great factor dominating future healthcare resource allocations.

A list of useful variables for our project can be found in Appendix I.

**Demographic and Housing Estimates**

The United States Census Bureau produced this dataset to study each county’s demographic and housing statuses. This data provides a 5-year estimate for variables such as age, race, ethnicity, household numbers, and the total population of each county.

This information can help us understand the extent of the pandemic’s impact on senior citizens and people of color particularly.

A list of useful variables for our project can be found in Appendix I.

**COVID-19 Vaccinations in the United States (County) Data**

This dataset is published by the CDC with information on administrated COVID-19 vaccinations in each U.S. county. Variables in this dataset were broken down by age, demographic group, and vaccine dosage and type.

By analyzing this data, we may confirm and measure the extent of the positive impact of vaccinations on reducing the communal CFR. The data is reported in terms of 7-day sums.

A list of useful variables for our project can be found in Appendix I.

**Selected Economics Characteristics**

This dataset provides financial predictors regarding the relationship between economic status and COVID-19 CFR. In counties with low employment and income, we expect to find a higher CFR due to the lack of access to healthcare facilities.

Additionally, the use of public transportation could be a factor distinguishing the different impacts COVID-19 had on metropolitan areas than the rural, low-income areas.

A list of useful variables for our project can be found in Appendix I.

**New York Times Coronavirus in the U.S. Data**

This dataset reports the accumulative numbers of COVID-19 cases and deaths in each county. Observations were reported once in a few days regularly throughout 2020, 2021, and 2022.

A list of useful variables for our project can be found in Appendix I.

**Missing Data**

* The 2020 PLACES data did not include information on depression (DEPRESSION) and high cholesterol (HIGHCHOL).
* The 2021 PLACES data did not include data for New Jersey counties.
* The 2022 PLACES data did not include data for several variables for New Jersey counties (HIGHCHOL, CHOLSCREEN, BPMED, and BPHIGH).
* Missing variables would not be a problem, as those are not our variables of interest, yet we would have to remove NJ counties from our analysis due to the missing variables.

**Confounding Variables**

Variables such as CHD, COPD, and CSMOKING show some correlation.

**New Variables Calculated from above Sources**

* Case fatality ratio (CFR) = # of deaths / # of COVID-19 cases
* Percent of total/inpatient/ICU beds occupied by patients confirmed COVID-19 = # beds occupied by COVID-19 patients / # total beds
* Household income in relation to national average = Average household income / national income
* Weighted vaccination effectiveness = Effectiveness of vaccination \* Percentage of vaccinated population

**Appendix I: Data Dictionary**

* NY Times Coronavirus in the U.S. Data
* Economics characteristics Data
* Demographic and Housing Data
* PLACES Dataset
* Covid-19 reported patient impact and hospital capacity by facility
* COVID-19 Vaccinations in the United States (County) Data

| **Variable** | **Full variable name** | **Class** |
| --- | --- | --- |
| Date | Date when the data was updated | number |
| County | County | Text |
| State | State | Text |
| FIPS | FIPS code | Number |
| Cases | Number of COVID cases | Number |
| Deaths | Number of COVID deaths | Number |
| Estimate!!EMPLOYMENT STATUS!!Population 16 years and over!!In labor force!!Civilian labor force!!Unemployed | The number of civilian 16 years old or older in the labor force that are unemployed | Number |
| Estimate!!COMMUTING TO WORK!!Workers 16 years and over!!Public transportation (excluding taxicab) | The number of workers 16 years old or older commuting to work using public transportation, not including taxicab | Number |
| Estimate!!INCOME AND BENEFITS (IN 2019 INFLATION-ADJUSTED DOLLARS)!!Total households!!Median household income (dollars) | The median household income in dollars | Number |
| **DP05\_0024PE** | Percentage of population 65 years and older | Number (%) |
| **DP05\_0001E** | Total Population | Number |
| **DP05\_0086E** | Total housing units | Number |
| **DP05\_0037E** | Non-White Racial Makeup | Number |
| CASTHMA | Current asthma among adults aged >= 18 years | number |
| CHD | Coronary heart disease among adults aged >= 18 years | number |
| COPD | Chronic obstructive pulmonary disease among adults aged >= 18 years | number |
| OBESITY | Obesity among adults aged >= 18 years | number |
| BINGE | Binge drinking among adults aged >= 18 years | number |
| CSMOKING | Current smoking among adults aged >= 18 years | number |
| SLEEP | Sleeping less than 7 hours among adults aged >= 18 years | number |
| collection\_week | This date indicates the start of the period of reporting (the starting Friday). | Date & Time |
| fips\_code | The Federal Information Processing Standard (FIPS) code of the location of the hospital. | Plain Text |
| is\_metro\_micro | This is based on whether the facility serves a Metropolitan or Micropolitan area. True if yes, and false if no. | Plain Text |
| total\_beds\_7\_day\_sum | Sum of reports of total number of all staffed inpatient and outpatient beds in the hospital, including all overflow, observation, and active surge/expansion beds used for inpatients and for outpatients (including all ICU, ED, and observation) reported during the 7-day period. | Number |
| all\_adult\_hospital\_beds\_7\_day\_sum | Sum of reports of all staffed inpatient and outpatient adult beds in the hospital, including all overflow and active surge/expansion beds for inpatients and for outpatients (including all ICU, ED, and observation) reported during the 7-day period. | Number |
| all\_adult\_hospital\_inpatient\_beds\_7\_day\_sum | Sum of reports of all staffed inpatient and outpatient adult beds in the hospital, including all overflow and active surge/expansion beds for inpatients and for outpatients (including all ICU, ED, and observation) reported during the 7-day period. | Number |
| inpatient\_beds\_used\_7\_day\_sum | Sum of reports of total number of staffed inpatient beds that are occupied reported during the 7-day period. | Number |
| all\_adult\_hospital\_inpatient\_bed\_occupied\_7\_day\_sum | Sum of reports of total number of staffed inpatient adult beds that are occupied reported during the 7-day period. | Number |
| inpatient\_beds\_used\_covid\_7\_day\_sum | Sum of reported patients currently hospitalized in an inpatient bed who have suspected or confirmed COVID-19 reported during the 7-day period. | Number |
| total\_adult\_patients\_hospitalized\_confirmed\_covid\_7\_day\_sum | um of reports of patients currently hospitalized in an adult inpatient bed who have laboratory-confirmed COVID-19. Including those in observation beds. | Number |
| total\_pediatric\_patients\_hospitalized\_confirmed\_covid\_7\_day\_sum | Sum of reports of patients currently hospitalized in a pediatric inpatient bed, including NICU, PICU, newborn, and nursery, who have laboratory-confirmed COVID-19. Including those in observation beds. | Number |
| inpatient\_beds\_7\_day\_sum | Sum of reports of total number of staffed inpatient beds in your hospital including all overflow, observation, and active surge/expansion beds used for inpatients (including all ICU beds) reported in the 7-day period. | Number |
| total\_icu\_beds\_7\_day\_sum | Sum of reports of total number of staffed inpatient ICU beds reported in the 7-day period. | Number |
| total\_staffed\_adult\_icu\_beds\_7\_day\_sum | Sum of reports of total number of staffed inpatient adult ICU beds reported in the 7-day period. | Number |
| icu\_beds\_used\_7\_day\_sum | Sum of reports of total number of staffed inpatient ICU beds reported in the 7-day period. | Number |
| staffed\_adult\_icu\_bed\_occupancy\_7\_day\_sum | Sum of reports of total number of staffed inpatient adult ICU beds that are occupied reported in the 7-day period. | Number |
| staffed\_icu\_adult\_patients\_confirmed\_covid\_7\_day\_sum | Sum of reports of patients currently hospitalized in a designated adult ICU bed who have laboratory-confirmed COVID-19. Including patients who have both laboratory-confirmed COVID-19 and laboratory-confirmed influenza in this field reported in the 7-day period. | Number |
| Date | Date data are reported on CDC COVID Data Tracker | Date & Time |
| FIPS | Federal Information Processing Standard State Code | Plain Text |
| MMWR\_week | MMWR Week | Number |
| Recip\_County | County of residence | Plain Text |
| Completeness\_pct | Represents the proportion of people with a completed primary series whose Federal Information Processing Standards (FIPS) code is reported and matches a valid county FIPS code in the jurisdiction. | Number |
| Administered\_Dose1\_Recip | People with at least one Dose by State of Residence | Number |
| Administered\_Dose1\_Pop\_Pct | Percent of Total Pop with at least one Dose by State of Residence | Number |
| Administered\_Dose1\_Recip\_65Plus | People 65+ with at least one Dose by State of Residence | Number |
| Administered\_Dose1\_Recip\_65PlusPop\_Pct | Percent of 65+ Pop with at least one Dose by State of Residence | Number |
| Series\_Complete\_Yes | Total number of people who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where vaccine recipient lives | Number |
| Series\_Complete\_Pop\_Pct | Percent of people who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where vaccine recipient lives | Number |
| Series\_Complete\_65Plus | Total number of people ages 65+ who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction where vaccine recipient lives | Number |
| Series\_Complete\_65PlusPop\_Pct | Percent of people ages 65+ who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction where vaccine recipient lives | Number |
| Booster\_Doses | Total number of people who completed a primary series and have received a booster (or additional) dose. | Number |
| Booster\_Doses\_Vax\_Pct | Percent of people who completed a primary series and have received a booster (or additional) dose. | Number |
| Booster\_Doses\_65Plus | Total number of people ages 65+ who completed a primary series and have received a booster (or additional) dose. | Number |
| Booster\_Doses\_65Plus\_Vax\_Pct | Percent of people ages 65+ who completed a primary series and have received a booster (or additional) dose. | Number |
| Second\_Booster\_65Plus | Total number of people ages 65+ who have received a second booster dose. | Plain Text |
| Second\_Booster\_65Plus\_Vax\_Pct | Percentage of people ages 65+ with a first booster dose who received a second booster dose. | Plain Text |