**Research Question:** Is race a significant variable in predicting weekly COVID case fatality ratio by county?

**Methods:** Multiple linear regression

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

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

**Analysis Plan**

**Step-by-step Analysis Plan**

1. Clean up each data set as planned below (filter out missing variables and variables that are not useful, create and calculate new variables)
2. Calculate the case fatality ratio (CFR) for each week\_number and each fips\_code (unique identifier for each county) for response variable
3. Merge CFR column to each dataset by fips\_code and week\_number
4. Conduct exploratory data analysis, look for outliers, correlated variables, etc.
5. Conduct best subsets or stepwise regression on each data set separately and select 5-10 variables (from each data set, 5-10 \* 5 explanatory variables in total) that have the best prediction of CFR with high R-square values
6. Make a combined data set by merging selected explanatory variables by FIPS code
7. Do a best subsets/stepwise regression to determine the most important variables for CFR, look into residual plots (perform transformations if necessary) to create a final reduced model
8. Pick one variable/category (e.g. race, income) and conduct an ESS test on the reduced model vs. reduced model + variable/category of interest (e.g. race, income) to test if the variable/category has a significant impact on CFR

**What is an experimental unit?**

The experimental unit is counties per week.

We are interested in each county’s case-fatality-ratio and whether race is a significant factor in predicting the weekly case-fatality-ratio of each county. All of our data will be joined together by FIPS codes (unique codes assigned to each county).

**Data Description**

**COVID-19 Cases and Deaths - Rosy**

How the data was collected (any reason the data may be biased?)

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

Missing data

* There are observations with county name “Unknown” and “Pending County Assignment,” and three counties (New York City, NY; Joplin, MO; Kansas City, MO) are missing FIPS codes. There is no missing data in cases and deaths.

The response variable

* The response variable would be weekly case fatality ratio, calculated by weekly deaths / weekly cases. This would be calculated after both deaths and cases are combined by weeks.

Confounding variables

* There seems to be no confounding variable in this data set.

Modifications before any analysis

* Rows with improper county names (“Unknown” and “Pending County Assignment”) would be removed, then missing FIPS codes would be added to the three aforementioned counties. Then, deaths and cases would be combined by weeks using the date column. Weekly case fatality ratio would be calculated for every county using weekly deaths and cases columns.

What is the n?

* After data cleaning, the merged 3-year COVID cases and deaths data will have 3138 \* 52 \* 3 = 489528 columns.

**PLACES Project - Rosy**

How the data was collected (any reason the data may be biased?)

* 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. This data might be biased as several health measures might be highly correlated with each other.

Missing Data

* The 2020 PLACES data does not include information on depression (DEPRESSION) and general health (GHLTH).
* The 2021 PLACES data does not include data for New Jersey counties.
* The 2022 PLACES data does not include data for several variables for New Jersey counties (HIGHCHOL, CHOLSCREEN, BPMED, and BPHIGH).
* The 2020 and 2021 PLACES data does not include data for counties Copper River and Chugach, AK, and the 2022 PLACES data does not include data for county Valdez-Cordova, AK.
* All rows for New Jersey counties and three Alaska counties (Copper River, Chugach, and Valdez-Cordova) would be removed, and missing variables (columns) would be removed as well.

Key explanatory variables

* Shown in Appendix I, color-coded by blue. We would be using age-adjusted estimates.

Confounding variables

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

Modifications before any analysis

* After missing data is removed, three data sets (2020, 2021, and 2022 PLACES) would be merged. Year would be added as a dummy variable.

What is the n?

* After data cleaning, the merged 3-year PLACES data will have 9360 columns.

**Patient Impact and Hospital Capacity - Cameron**

How the data was collected (any reason the data may be biased?)

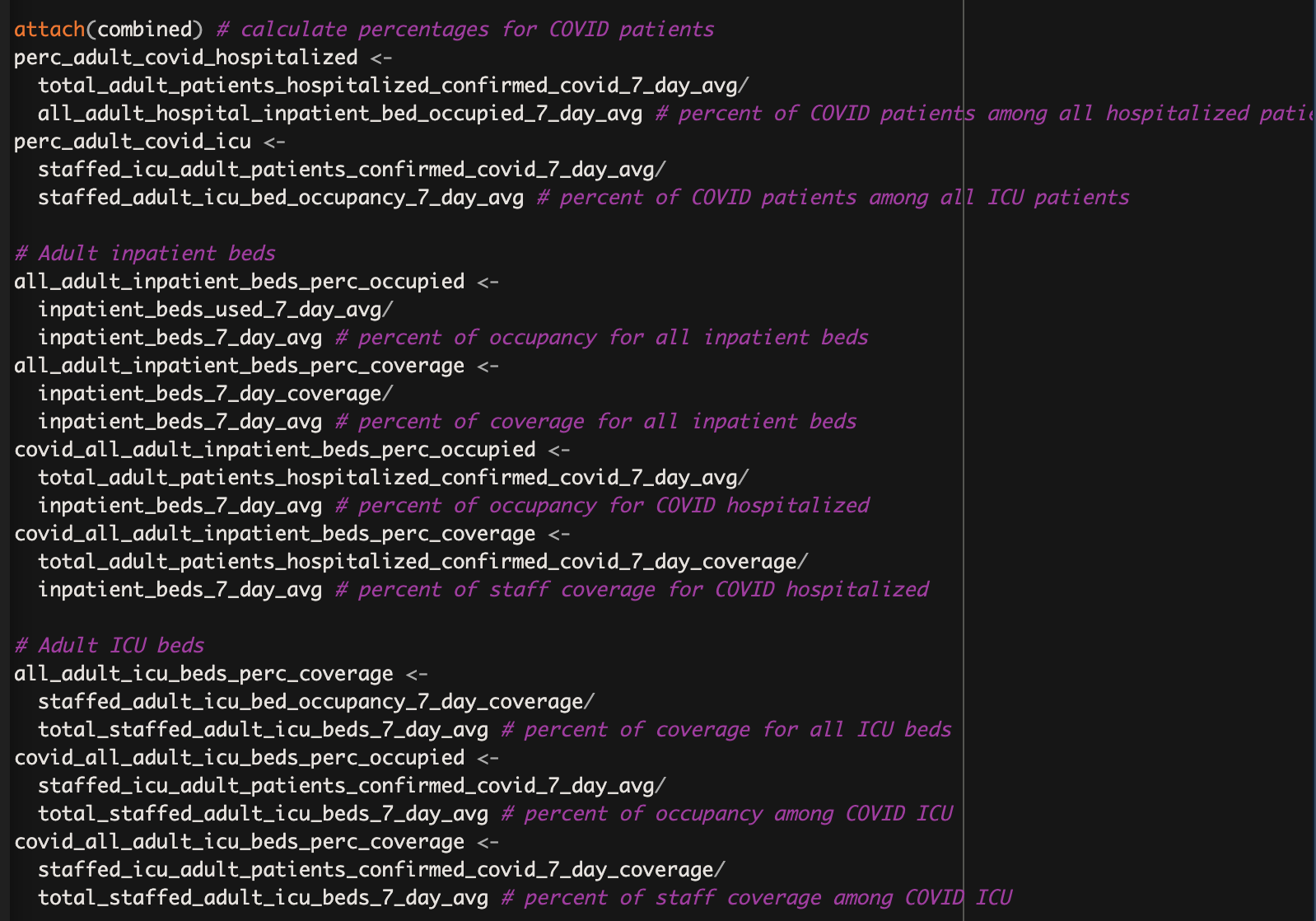
* The data was reported by selected hospitals weekly
* The data could be biased by outside factors (i.e., weather, season, policies, etc.)

Missing data (how many rows may need to be removed, how will you determine this?)

* As reported in the missing data file
* Data for Joplin county, MO and Kansas City county, MO would be removed, since they are missing FIPS code
* There are no missing data for cases and deaths

The key explanatory variables (clearly labeled and described)

* As shown in the figure below: (more details in the R code)



Any confounding variables

* The variables could be confounded by outside factors such as hospital’s distance from population-concentrated areas and city/county policies
* Unproportional patient age populations could be a factor as elder populations suffered from more severe symptoms than younger individuals

Modifications before any analysis (missing data, creating new variables, merging datasets)

* As shown in files
* Missing data will be either omitted (if the missing% is high) or estimated by calculating the median regarding data points before/after (if only missing a few data points)
* Merging by week\_number, year\_number, and fips\_code
  + We will only investigate counties that were included in the hospital occupancy dataset

What is the n? How is the data collected? By week? By date? How many years?

* 776438 observations with 16 variables

**Vaccination - Shrey**

How the data was collected (any reason the data may be biased?)

* The data was collected by the Centers for Disease Control and Prevention (CDC) through reports submitted by state and local health departments. The data may be biased if there are variations in reporting standards or if some jurisdictions are less timely or accurate in their reporting.

Missing data (how many rows may need to be removed, how will you determine this?)

* It is possible that some counties or states may not have reported data for certain weeks or may have incomplete data for certain variables. These missing values will need to be identified and either removed or imputed before any analysis.

The response variable(s)- typically only one

* The response variable in this dataset is the number of COVID-19 vaccinations administered in each county. However, we are going to add Case Fatality Ratio variable to this dataset, and that will be the response variable

The key explanatory variables (clearly labeled and described)

* Given above in the data dictionary for this dataset

Any confounding variables

* There may be other variables that are not included in the dataset that could be confounding factors, such as demographic characteristics, socio-economic status, and access to healthcare.

Modifications before any analysis (missing data, creating new variables, merging datasets)

* Before any analysis, missing data will need to be imputed or removed. New variables may also be created, such as the percentage of people vaccinated in each county or state. The dataset may also be merged with other datasets, such as demographic data or COVID-19 case data.

What is the n? How is the data collected? By week? By date? How many years?

* 1048576 rows. Collected by week, from 9/9/21 - 3/15/23

**Demographic and Housing Dataset**

* How the data was collected (any reason the data may be biased?)
  + The data was collected by the United States Census Bureau across all counties in the nation.
* Missing data (how many rows may need to be removed, how will you determine this?)
  + There was no missing data in our variables of interest
* The response variable(s)- typically only one
  + There are no response variables here, only variables of explanatory interest.
* The key explanatory variables (clearly labeled and described)
  + Racial and Ethnic Makeup
  + Number of Households
  + Age
* Any confounding variables
  + There are no confounding variables
* Modifications before any analysis (missing data, creating new variables, merging datasets)
  + We may want to create new variables that combine the populations of all people of color in a county, such as hispanic and black, etc.
* What is the n? How is the data collected? By week? By date? How many years?
  + The data is collected by county, and separated into yearly datasets.

**Economics Characteristics Dataset - Pom**

* How the data was collected (any reason the data may be biased?) By week? By date? How many years?
* The data was collected by the U.S. Census Bureau through the American Community Survey using a combination of mail, internet-based questionnaires, and in-person interviews from a sample of households.
* The survey is conducted annually and the resulting data is subject to quality control checks and statistical weighting before being made publicly available through the Census Bureau's website.
* What is the n?
* n = 6444
* Missing data (how many rows may need to be removed, how will you determine this?)
* No missing data in chosen explanatory variables
* The response variable(s)- typically only one
* The explanatory variable is not in this dataset
* Case Fatality Ratio variable (deaths/cases) from the New York Master Covid Dataset
* The key explanatory variables (clearly labeled and described)
* Color-coded yellow in the Appendix I
* Any confounding variables
* Potential confounding variables are factors that could impact explanatory variables in this dataset and also be associated with the outcome variable including age, gender, education level, employment status, occupation, race/ethnicity, marital status, and geographic region. These confounding variables may exist in our other datasets
* Modifications before any analysis (missing data, creating new variables, merging datasets)
* No missing data to be removed
* The first row of the data should be renamed concisely
* The first row of the data should be rownames instead of the variable code (e.g. DP03\_0001E3)
* The econ data set from 2020 and 2021 will be merged together to make an econ.all data.
* The econ.all data will be merged with the remaining data from our group to form the final data set

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

**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 dabta 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 general health (GHLTH).
* 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 |