Stage 1 Task 2 Report

Group 1: Derek Fox, Sarah Barber, David Mack, Trevor Church

COVID-19 Dataset:

Data Dictionary -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Definition | Data Type | Example Values | Required? |
| countyFIPS | County ID Number | Integer | 1000, 1001, ... | yes |
| County Name | Name of County | String | Bibb County, Coffee County, ... | yes |
| State | State Abbreviation for County | String | AL, NC, ... | yes |
| State FIPS | State ID Number | Integer | 1, 2, ... | yes |
| Cases – by Date | Number of Cases on given Date | Integer | 0, 100, ... | yes |
| Deaths – by Date | Number of Deaths on given Date | Integer | 0, 100, ... | yes |
| Population | County Population | Integer | 1000, 5000, ... | yes |

2020 Presidential Election Enrichment Dataset – Derek Fox:

Data Dictionary -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Definition | Data Type | Example Values | Required? |
| state | State Name | String | Florida, Alabama, ... | yes |
| county | County Name | String | Orange County, Washington County, ... | yes |
| candidate | Candidate Name | String | Joe Biden, Donald Trump, ... | yes |
| party | Abbreviation of Party Name of Candidate | String | DEM, REP, ... | yes |
| total\_votes | Number of Votes for Candidate | Integer | 1000, 5000, ... | yes |
| won | Did Candidate Win? | Boolean | True, False | yes |

How to merge – This dataset also contains county information, so we can merge with the COVID-19 dataset by a join on the county columns. It may first be necessary to create a column for the County FIPS to uniquely identify each county (and avoid cases where multiple counties share the same name).

Hypotheses – This dataset might provide insight into the population’s attitudes of each county regarding the COVID-19 pandemic. For instance, counties which voted Republican may contain more individuals who believed that the pandemic was a hoax. This may lead to a larger number of cases in counties who voted Republican.

2020 Employment Statistics – David Mack

Data Dictionary-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Definition** | **Data Type** | **Example Values** | **Required?** |
| Year | Year of data | Integer | 2020 | Yes |
| Quarter | Quarter of the year in question | Character | ‘A’, ‘1’, ‘2’, ‘3’, ‘4’ | Yes |
| Area Type | The kind of area in question | String | ‘County’ | Yes |
| State Name | Name of the State | String | ‘Alabama’, ‘New York’ | Yes |
| Area | Name of the county and State | String | ‘Baldwin County, Alabama’ | Yes |
| Ownership | Who owns and manages the Industry | String | ‘Total Covered’, ‘Private’, ‘Federal Government’ | Yes |
| Industry | Classification of Industry | String | 101 Goods-producing | Yes |
| Annual Average Status Code | Status code, or disclosure code ('N' for not disclosed) | Character | ‘N’, | No |
| Annual Average Establishment Count | The average establishment counts | Integer | 600,000 , 1,000,000 | Yes |
| Annual Average Employment | The average amount of employed people | Integer | 600,000 , 1,000,000 | Yes |
| Annual Total Wages | Average weekly wage based on the 12-monthly employment levels and total annual wage levels | Integer | 8,905,606,771 | Yes |
| Annual average weekly wages | Average annual pay based on employment and wage levels for a given year | Integer | 1,231 | Yes |
| Annual Average Pay | The average amount of annual income | Integer | 60,000 | Yes |
| Employment location quotient relative to the US | Self explanatory | Integer | 1.00, 0.68, 0.24 | Yes |
| Total Wage Quotient relative to the US | Self explanatory | Integer | 1.00, 0.68, 0.24 | Yes |
| Hidden Data | This includes numerical representations of all previous data. This includes things like FIPS and zip codes. | Integer | 1104, 28038, 600 | Yes |

How to Merge – This dataset can be merged easily via the FIPSs contained in the Hidden Data. An inner merge will preserve all relevant data. To make things easier I can remove any duplicate data from the BLS database before the merge. This is a good idea because I cannot be sure data of type string is formatted exactly the same way across csv files.

Hypothesis – Individuals of lower economic status will have higher rates of infection and death than those of comparatively higher economic status. This can be measured by looking average employment quality for a given county and evaluating if there is a correlation between this data and infection/death rates.

2020 Census Dataset – Sarah Barber

Data Dictionary -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Definition** | **Data Type** | **Example Values** | **Required?** |
| County Name | Name of each county in each state. | String | Randolph County, Clay County, ... | Yes |
| State | State abbreviation | String | NC, LA, OH, ... | Yes |
| Estimated Total Population | Estimated total population of the state/county. | Integer | 326569308, 4893186, ... | Yes |
| Estimated Total Population Under 5 years | Estimated total population of the state/county under 5 years old | Integer | 3235, 1332, 671, ... | Yes |
| *This continues for every age bracket until 85 and over* | *It’s just a lot and it would be excessive* | Integer | 2891, 633, 1174, ... | Yes |
| Estimated Total Male Population | Estimated total male population in a state/county | Integer | 6032, 105889, 7268, ... | Yes |
| Estimated Total Population Under 5 years | Estimated total male population that is under 5 years old in a state/county | Integer | 311, 728, ... | Yes |
| *This continues for every age bracket until 85 and over* | *It’s just a lot and it would be excessive* | Integer | 2891, 633, 1174, ... | Yes |
| Estimated Total Female Population | Estimated total female population in a state/county | Integer | 165750778, 4595, ... | Yes |
| Estimated Total Population Under 5 years | Estimated total female population that is under 5 years old in a state/county | Integer | 672, 627, 1508, ... | Yes |
| *This continues for every age bracket until 85 and over* | *It’s just a lot and it would be excessive* | Integer | 2891, 633, 1174, ... | Yes |

This dataset can be merged with the primary Covid-19 dataset because it contains information regarding sex, age, voting age population, and racial demographics that can merge with the Covid-19 dataset through the state and county name columns.

Hypothesis – This dataset can provide insight into the demographics who were more susceptible to Covid-19 by seeing which states had higher cases, and the demographics of said states. For example, if a state had a higher population of older people, they may have a higher number of Covid-19 cases and/or deaths.

2020 Census - Educational Attainment - Trevor Church

Data Dictionary -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Definition** | **Data Type** | **Example Values** | **Required?** |
| County Name | Name of each county | String | Guilford County, Autauga County | Yes |
| 18-24 - Total polled for educational attainment | Gives the total number of participants in the age group of each county, so that the rest of the data can be contextualized | Int | 4636, 15733, 2007 | Yes |
| 18-24 - Less than high school education | Number of citizens within the age range without a high school diploma | Int | 581, 2333, 463 | Yes |
| 18-24 - High school grad | Number of citizens within the age range who have graduated college | Int | 1723, 6471. 585 | Yes |
| 18-24 - Some college or Associate’s | Number of citizens within the age range with some college or an Associate’s degree | Int | 1720, 6471, 585 | Yes |
| 18-24 - Bachelor’s degree | Number of citizens within the age range with a Bachelor’s dgree | Int | 612, 1360, 84 | Yes |
| 25 and up - Total | Gives the total number of participants in the age group of each county, so that the rest of the data can be contextualized | Int | 37860, 155563, 17797 | Yes |
| 25 and up - Less than 9th grade | Number of citizens within the age range with no high school education | Int | 909, 3834, 1387 | Yes |
| 25 and up - Between 9th and 12th grade | Number of citizens within the age range with some high school education | Int | 3364, 10989, 3110 | Yes |
| 25 and up - High school grad | Number of citizens within the age range that finished high school | Int | 11880, 42272, 6361 | Yes |
| 25 and up - Some college, no degree | Number of citizends within the age range that did not finish college | Int | 7663, 34475, 3487 | Yes |
| 25 and up - Associate's degree | Number of citizens within the age range with an associate's degree | Int | 3323, 14357, 1385 | Yes |
| 25 and up - Bachelor’s degree | Number of citizens within the age range with a bachelor’s degree | Int | 6320, 31444, 1296 | Yes |

How to merge – The Covid and educational attainment sheets have county names listed in alphabetical order. With some slight manipulation of naming conventions in the educational sheet, the columns can be merged.

Hypothesis – This dataset can be useful in providing information on how COVID affected those with less education, who may be more susceptible to the rampant spread of misinformation alongside COVID. This can also be applicable to the election results datasheet, as the Republican candidate of the 2020 election has publicly stated his desire to appeal to the less educated.