COVID 19 by County: An Analysis of the COVID 19 Infection and Death Rates in US Counties

# Introduction

Over the past year and a half our world has been profoundly changed by the outbreak of COVID 19. Millions have died and no part of humanity has been untouched by this global pandemic. What this project attempts to discover is what quantitative factors contribute to the infection and mortality rate of COVID 19 patients by US county by using publicly available socioeconomic data about all the 3142 counties in the United States of America.

# Literature Review

### **1.** Poverty, inequality and COVID-19: the forgotten vulnerable

The above study states that a number of factors increase people of low socio-economic status(SES) in getting COVID 19. First, they are more likely to live in overcrowded accommodation. Second, they are often employed in jobs were work from home is not an option. Third, people in low SES groups are more likely to have unstable work which can contribute to stress, which is known to lower a person’s immune system. Fourth, people of low SES tend to seek medical attention at a more advanced stage of illness. Also, access to health is determined by a person’s ability to access the healthcare system with ease, something people of low SES may feel that there are barriers to. They conclude that an increased exposure the virus, increase stress and lower access to healthcare makes people of low SES particularly vunerable to COVID 19.

### 2. INCOME AND POVERTY IN THE COVID-19 PANDEMIC

The purpose of the above study was to create real time information as to the level of poverty in the United States and assess the effectiveness of government stimulus. The study is using data from Basic Monthly Current Population Survey (Monthly CPS), which contains high frequency data for a large, representative sample of U.S. families and individuals. The study shows that poverty fell in the early months of the pandemic due to government assistance. However, since some of the government assistance was a one-time payment, the study concludes that the initial drop in poverty was unlikely to continue as the pandemic and recovery occurs.

### 3.COVID-19, school closures, and child poverty: a social crisis in the making

This paper is of the opinion that a long period of school closures would have negative social and health consequences for children living in poverty and could increase the current inequities in society. One key way that children living in poverty would be affected by school closures is an increase in food insecurity as schools are often a key source of meals for children living in poverty. This paper also cites research to suggest that children living in poverty are more affected by breaks and disruptions in school than their peers who are not living in poverty. The paper recommends that officials adapt their school closures to address the above points as well as prepare for considerable challenges that will arise once the pandemic is over.

### 4. Monthly Poverty Rates in the United States during the COVID-19 Pandemic

The above study attempts to produce monthly estimates of poverty in the US. This is done by creating real times estimates of family income. This study found that the monthly poverty rate in the United States increased from 15% to 16.7% from February to September 2020. This increase in poverty has been particularly bad for children, and Black and Hispanic individuals. This study also notes the increase in poverty rates is also due to the expiration of certain benefits that had begun in the early days of the pandemic but had expired in the summer of 2020.

### 5. Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States

This study identified 12 US states that reported the race/ethnicity of the people being hospitalized for COVID 19 between April 30 and June 24, 2020. The study found that the share of white people being hospitalized was much smaller than their overall share of the state population in general. On the flip side, the share of black people being hospitalized was larger than their overall share in the state’s representation. The same disparity found in Black patients was also found in Hispanic patients and Native American patients(where that data existed). This pattern was not seen in the Asian population where hospitalization was lower than state representation.

### 6.Racial, Economic, and Health Inequality and COVID-19 Infection in the United States

The purpose of this study was to report the connection of COVID 19 with respect to race, economic inequality and health in the United States. The study looked at demographic, socioeconomic, and mobility data from 369 US counties in 7 states. What it found is that the risk factors for infection and mortality were different. What it found was that more affluent counties were more prone to infection, but less affluent counties had a higher death rate. African Americans were the more vulnerable than other ethnicities to COIVD 19

### 7. Impacts of the COVID-19 pandemic on rural America

The impact of COVID 19 in rural areas in under-studied compared to urban areas. The study found that rural unemployment rates increased more than their urban counterparts during the pandemic. They found major decline in the perceptions of local economic health in rural areas, increased reliance on unemployment benefits and negative impacts to mental health. The study predicts a long road to recovery of rural America from the pandemic.

### 8. Obesity Is a Risk Factor for Greater COVID-19 Severity

This article examples the connection between obesity and COVID 19 severity among patients. The study found that the presences of obesity in a patient made that patient three times more likely to have a severe infection compared to a non-obese patient. The strong relationship between obesity and severity of COIVD 19 infection was not known at the time of the study, however the study notes several technical theories and states that obesity commonly aggravates the severity of respiratory diseases.

### 9. The effect of smoking on COVID-19 severity: A systematic review and meta-analysis

This study aims to definitively quantify the effects of smoking on COVID 19 severity by analysing all studies that were published on the connection between smoking and COVID 19. What they found is that the smokers and former smokers had a significant increase in the chance of having a severe reaction to COVID 19 compared to never smokers.

## Citations:

1. Patel, J. A., Nielsen, F., Badiani, A. A., Assi, S., Unadkat, V. A., Patel, B., Ravindrane, R., & Wardle, H. (2020). Poverty, inequality and COVID-19: the forgotten vulnerable. *Public health*, *183*, 110–111. https://doi.org/10.1016/j.puhe.2020.05.006
2. Income and Poverty in the COVID-19 Pandemic Jeehoon Han, Bruce D. Meyer, and James X. Sullivan NBER Working Paper No. 27729 August 2020 JEL No. H53,I32,J65
3. Published Online April 7, 2020 https://doi.org/10.1016/ S2468-2667(20)30084-0
4. Parolin, Zachary, Megan A. Curran, Jordan Matsudaira, Jane Waldfogel and Christopher Wimer. 2020. “Monthly Poverty Rates in the United States during the COVID-19 Pandemic.” Poverty and Social Policy Discussion Paper. New York, NY: Center on Poverty and Social Policy.
5. Karaca-Mandic P, Georgiou A, Sen S. Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States. JAMA Intern Med. 2021;181(1):131–134. doi:10.1001/jamainternmed.2020.3857
6. Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., Li, J., & Zand, R. (2021). Racial, Economic, and Health Inequality and COVID-19 Infection in the United States. *Journal of racial and ethnic health disparities*, *8*(3), 732–742. https://doi.org/10.1007/s40615-020-00833-4
7. Impacts of the COVID-19 pandemic on rural America. J. Tom Mueller, Kathryn McConnell, Paul Berne Burow, Katie Pofahl, Alexis A. Merdjanoff, Justin Farrell. Proceedings of the National Academy of Sciences Jan 2021, 118 (1) 2019378118; DOI: 10.1073/pnas.2019378118
8. Obesity Is a Risk Factor for Greater COVID-19 Severity. Feng Gao, Kenneth I. Zheng, Xiao-Bo Wang, Qing-Feng Sun, Ke-Hua Pan, Ting-Yao Wang, Yong-Ping Chen, Giovanni Targher, Christopher D. Byrne, Jacob George, Ming-Hua Zheng. Diabetes Care Jul 2020, 43 (7) e72-e74; **DOI:** 10.2337/dc20-0682
9. Reddy, RK, Charles, WN, Sklavounos, A, Dutt, A, Seed, PT, Khajuria, A. The effect of smoking on COVID-19 severity: A systematic review and meta-analysis. *J Med Virol*. 2021; 93: 1045– 1056. <https://doi.org/10.1002/jmv.26389>

# Dataset

The dataset will be an amalgamation of several different datasets, however all the datapoints will be on a county level. The data is mainly from the US Government but also contains some data from NGOs and data from the New York Times. The data from the New York Times has been cited in article number seven in my literature review.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Hospitals per county | | | | |
| **Source(s)** | **Healthcare Cost Report Information System (HCRIS) data for hospitals** | | | | |
| **Description** | the number of hospitals that filed a cost report since the beginning of 2018, for each county. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 1.44 | | 1.00 | 2.56 | 74.00 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Poor or fair health | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Percentage of adults reporting fair or poor health (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 20.25 | | 20.00 | 5.13 | 42.00 | 9.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Poor physical health days | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Average number of physically unhealthy days reported in past 30 days (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.41 | | 4.41 | 0.78 | 8.30 | 2.37 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Poor mental health days | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Average number of mentally unhealthy days reported in past 30 days (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.69 | | 4.73 | 0.67 | 7.29 | 2.69 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Adult smoking | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Percentage of adults who are current smokers (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 21.36 | | 21.19 | 4.18 | 44.57 | 7.08 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Adult obesity | | | | |
| **Source(s)** | United States Diabetes Surveillance System : 2017 | | | | |
| **Description** | Percentage of the adult population (age 20 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m2. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 33.48 | | 33.80 | 6 | 58.9 | 11.80 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Physical inactivity | | | | |
| **Source(s)** | United States Diabetes Surveillance System : 2017 | | | | |
| **Description** | Percentage of adults age 20 and over reporting no leisure-time physical activity. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 26.8 | | 26.5 | 5.84 | 50.4 | 8.9 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Excessive drinking | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Percentage of adults reporting binge or heavy drinking (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 18.9 | | 18.83 | 3.25 | 28.34 | 6.45 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Uninsured | | | | |
| **Source(s)** | Small Area Health Insurance Estimates US Census: 2018 | | | | |
| **Description** | Percentage of population under age 65 without health insurance. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 18.08 | | 18.44 | 4.15 | 28.34 | 3.55 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | High school completion | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of adults ages 25 and over with a high school diploma or equivalent. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 86.79 | | 88.07 | 6.28 | 98.88 | 26.44 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Some college | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of adults ages 25-44 with some post-secondary education. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 57.94 | | 57.97 | 11.99 | 100 | .83 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Unemployment | | | | |
| **Source(s)** | Bureau of Labor Statistics : 2019 | | | | |
| **Description** | Percentage of population ages 16 and older unemployed but seeking work. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.01 | | 3.71 | 1.5 | 19.31 | .74 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Children in poverty | | | | |
| **Source(s)** | Small Area Income and Poverty Estimates US Census : 2019 | | | | |
| **Description** | Percentage of people under age 18 in poverty. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 20.15 | | 18.90 | 8.51 | 63.4 | 2.4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Income inequality | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Ratio of household income at the 80th percentile to income at the 20th percentile. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.52 | | 4.39 | .77 | 10.50 | 2.41 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Children in single-parent households | | | | |
| **Source(s)** | American Community Survey, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of children that live in a household headed by single parent. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 24.58 | | 23.08 | 10.02 | 75.1 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Social associations | | | | |
| **Source(s)** | County Business Patterns US Census: 2018 | | | | |
| **Description** | Number of membership associations per 10,000 population. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 11.57 | | 10.96 | 5.99 | 55.56 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Air pollution - particulate matter | | | | |
| **Source(s)** | Environmental Public Health Tracking Network : 2016 | | | | |
| **Description** | Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 7.68 | | 7.9 | 1.68 | 16 | 1.5 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Severe housing problems | | | | |
| **Source(s)** | Comprehensive Housing Affordability Strategy (CHAS) data : 2013-2017 | | | | |
| **Description** | Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen facilities, or lack of plumbing facilities. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.63 | | 13.06 | 4.56 | 69.14 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Driving alone to work | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of the workforce that drives alone to work. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 79.69 | | 81.22 | 7.7 | 97.10 | 5.7 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Frequent physical distress | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System : 2018 | | | | |
| **Description** | Percentage of adults reporting 14 or more days of poor physical health per month (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.68 | | 13.54 | 2.64 | 29.19 | 7.01 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Frequent mental distress | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System : 2018 | | | | |
| **Description** | Percentage of adults reporting 14 or more days of poor mental health per month (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 15.16 | | 15.16 | 2.37 | 24.67 | 8.84 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Diabetes prevalence | | | | |
| **Source(s)** | United States Diabetes Surveillance System : 2017 | | | | |
| **Description** | Percentage of adults aged 20 and above with diagnosed diabetes. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 12.38 | | 11.90 | 3.69 | 29.50 | 3.1 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Food insecurity | | | | |
| **Source(s)** | Map the Meal Gap : 2018 | | | | |
| **Description** | Percentage of population who lack adequate access to food. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.3 | | 13.1 | 3.75 | 30.40 | 3.6 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Limited access to healthy foods | | | | |
| **Source(s)** | USDA Food Environment Atlas : 2015 | | | | |
| **Description** | Percentage of population who are low-income and do not live close to a grocery store. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 8.74 | | 6.58 | 8.46 | 71.84 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Insufficient sleep | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System : 2018 | | | | |
| **Description** | Percentage of adults who report fewer than 7 hours of sleep on average (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 36.92 | | 36.94 | 3.96 | 49.06 | 25.62 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Uninsured adults | | | | |
| **Source(s)** | Small Area Health Insurance Estimates : 2018 | | | | |
| **Description** | Percentage of adults under age 65 without health insurance. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.75 | | 12.68 | 6.15 | 42.75 | 2.74 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Uninsured children | | | | |
| **Source(s)** | Small Area Health Insurance Estimates : 2018 | | | | |
| **Description** | Percentage of children under age 19 without health insurance. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 6.23 | | 5.35 | 3.34 | 27.49 | 0.85 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Median household income | | | | |
| **Source(s)** | Small Area Income and Poverty Estimates : 2019 | | | | |
| **Description** | The income where half of households in a county earn more and half of households earn less. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 55525.36 | | 53106.00 | 14562.33 | 151806 | 24732 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Children eligible for free or reduced price lunch | | | | |
| **Source(s)** | National Center for Education Statistics : 2018-2019 | | | | |
| **Description** | Percentage of children enrolled in public schools that are eligible for free or reduced price lunch. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 54.77 | | 52.96 | 18.63 | 100 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Homeownership | | | | |
| **Source(s)** | American Community Survey, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of occupied housing units that are owned. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 71.52 | | 72.78 | 8.39 | 93.06 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Population | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Resident population. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 105659.91 | | 25619 | 337990.39 | 10003107 | 86 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % below 18 years of age | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population below 18 years of age. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 21.95 | | 21.98 | 3.5 | 41.68 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % 65 and older | | | | |
| **Source(s)** | Percentage of population ages 65 and older. | | | | |
| **Description** | Census Population Estimates : 2019 | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 19.73 | | 19.39 | 4.82 | 58.17 | 4.86 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Non-Hispanic Black | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is non-Hispanic Black or African American. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 9.25 | | 2.43 | 14.43 | 85.87 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % American Indian & Alaska Native | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is American Indian or Alaska Native. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 2.37 | | 0.65 | 7.72 | 92.41 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Asian | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is Asian. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 1.59 | | 0.75 | 2.99 | 43.36 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Hispanic | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is Hispanic. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 9.92 | | 4.53 | 14.04 | 96.35 | 0.65 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Non-Hispanic White | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is non-Hispanic White. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 75.40 | | 82.63 | 20.33 | 97.83 | 2.69 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % not proficient in English | | | | |
| **Source(s)** | American Community Survey, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of population that is not proficient in English. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 1.72 | | 0.73 | 2.84 | 34.44 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Females | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is female. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 49.90 | | 50.31 | 2.29 | 57.01 | 26.51 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Rural | | | | |
| **Source(s)** | Census Population Estimates : 2010 | | | | |
| **Description** | Percentage of population living in a rural area. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 58.61 | | 59.54 | 31.56 | 100.00 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Provisional COVID Death Counts in the United States by County as of June 3, 2021 | | | | |
| **Source(s)** | Ny Times - https://github.com/nytimes/COVID-19-data | | | | |
| **Description** | Since late January 2020, The Times has tracked coronavirus cases and deaths as they are announced using data released by countries, states and local health officials. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 191.73 | | 50.00 | 900.70 | 33257.00 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Total ICU beds by county | | | | |
| **Source(s)** | **Healthcare Cost Report Information System (HCRIS) data for hospitals** | | | | |
| **Description** | the number of ICU beds reported in the most recent cost report for each hospital, including the categories “intensive care unit,” “coronary care unit,” “burn intensive care unit” and “surgical intensive care unit.” Aggregated by county. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 23.97 | | 0.00 | 85.14 | 2126.00 | 0.00 |

# Approach

## Step 1: Creating the dataset

I am pulling data from multiple sources to create my dataset. As a result, I will need to combine the different sources into a data frame. The data I am using has been exported from various websites as CSV and XLSX files. All the datasets share the same characteristic, namely that they show data on a county level, and the US contains 3142 counties (excluding Washington DC and US territories). Once completed the data frame will have a row for each county in the United States and columns for statistics that I am using for my analysis. Each county of the United States has a unique identifier named a Federal Information Process System Code (FIPS code) and the datasets that I am combining all contain the FIPS code. I will join the data using excel Index Match function against the FIPS code to create my final dataset.

I will be combining infections per county against the populations of the country to create a variable called ‘county infection rate’ which will be my first dependant variable for analysis. I will also be combining infections per county against deaths per county for a variable called ‘county infection death rate’ and this will be my second dependant variable.

## Step 2: Cleaning the data

The Datasets that I choose for the most part contain complete data for each 3142 counties. However, some attributes are missing data points for certain counties. For these null values, I will need to replace the null value with a not null value. I will be consulting with my supervisor on the best value to add, but some ideas I had was to substitute the null value with the mean of that attribute, the mean of the US State that the county is located in or another method.

There is no complete government dataset for COVID deaths by county in the United States, however the New York Times has created their own dataset that combines data from Federal, State and Local data. They have combined the data into a dataset that shows COVID 19 deaths n infections on a county level. The data does have some challenges that need to be dealt with. Firstly, New York City is located in 5 separate US counties, however due to how COVID deaths were reported by local health authorities the death data is combined into one geographic location called New York City. With the assistance of my supervisor, I will attempt to assign a fraction of that sum to each of the five counties that make up New York City. Two urban centers, Kansas City and Joplin are reported as their own geographic location and the deaths that occurred in those two locations are excluded from the respective counties that the cities are part of. I will consult with my supervisor on how to best address these anomalies.

## Step 3: Analysing the Data

I will be using Python to do regression analysis on the data. I will analyse my dependant variables, county infection rate and county infection death rate against independent variables that are in the dataset to identify how they are correlated. I will see if the dataset benefits from normalization, standardization or scaling. Once I have identified variables with a strong correlation, I will create a predictive model. This will be a cross sectional regression as all the independent variables are static.

## Step 4: Presenting the Data

I will present the data as outlined in the course shell.