Data Management final project
Maréchal-Volaris

1. Presentation of the Research Question

Cleaning and merging the datasets allows us to obtain a clearer view of the question we are aiming to answer to: to what extent socio-economical factors have influenced the impact of the Covid-19 pandemic throughout the United States? Although Covid has been a global issue, it is fair to state that the United States were among the most impacted countries: according to the Centers for Disease Control and Prevention, almost 1.2 millions Americans have died of Covid-related complications, the most of any country worldwide. Once we take into account the overall populations, indicators are still concerning and should prompt us to try to understand better what can explain them. For example, the mortality rate (the ratio between total death and total cases reported) is equal to 1.17%, whereas it is only 0.43% in France. Such difference is explained thanks to various socio-economical factors, as well as political decisions. For example, the American healthcare system, relying on private companies rather than a mandatory public option, has been criticized for its negative impact on overall public health. Studying how uninsured Americans have been affected by the pandemic would therefore helps us understand if the weaknesses of the healthcare system played a role in an increased mortality rate. In addition to this first approach, our datasets allow us to study in a deeper way how economical and social differences may have played a role in how the mortality rate has evolved in various regions. The `counties` dataset, for example, provides us with a full racial make-up of each county-equivalent in the country, as well as basic economic indicators, such as the median income or the unemployment rate. To answer to our research question, we have chosen to base our study on the county-level, in order to get a more precise view of the country, which would not have been possible on the state-level.

2. Understanding Data

1. Dataset Descritption

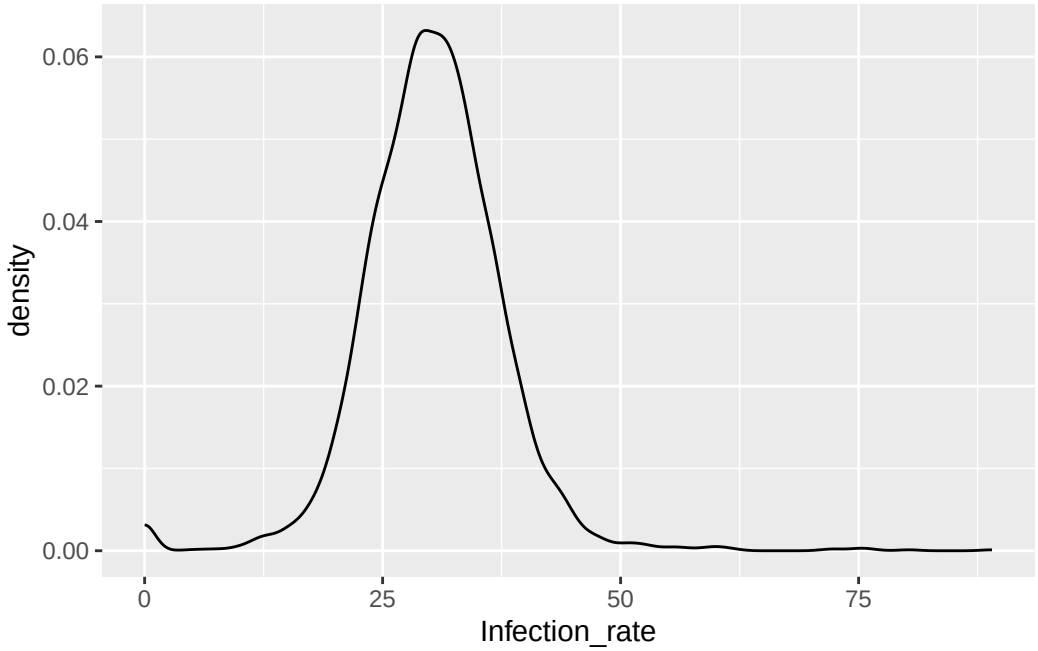
A - Dataset

To pursue our study, we need to establish per capita variables. Indeed, counties have widely varying populations.

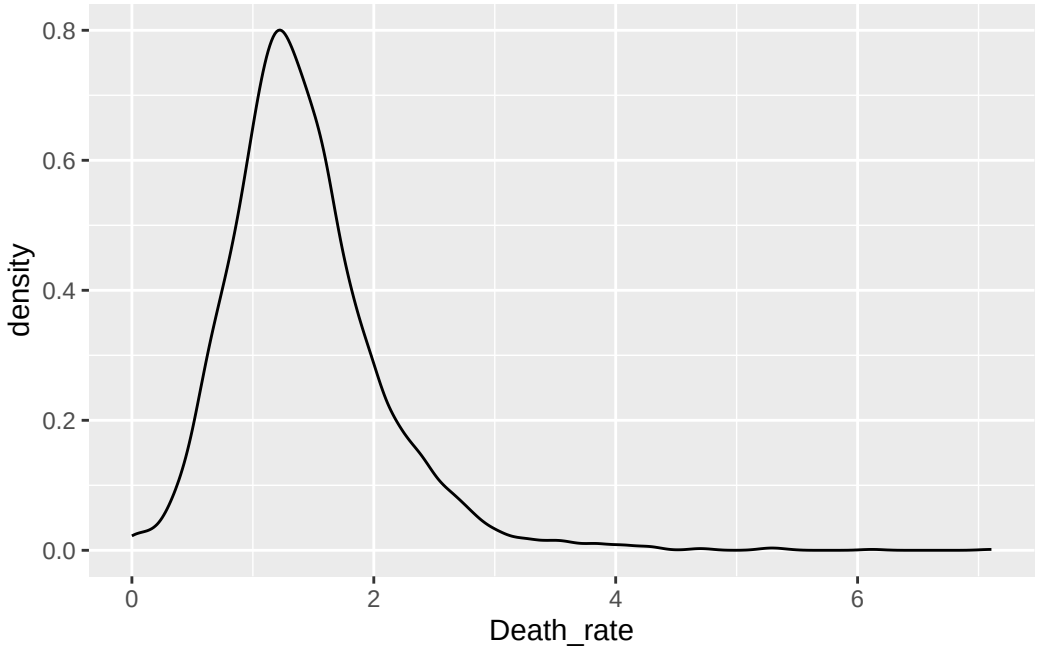
As we can see, in 2019, a county population ranges from 66 (Kalawao County, Hawaii) to more than 10 millions habitants (Los Angeles County, California). Aware of this disparity, we establish two per capita variables: the infection rate and the mortality rate. They are defined as: Infection rate=Total of confirmed cases/Total county population Mortality rate=Total death from Covid 19/Total of confirmed cases For a better understanding, we express these rates as percentages by multiplying them by 100.

We notice that two county-equivalents have infection rate higher than 100%. This abnormality can be explained by various reasons, such as an error in the database, a decline in population, a center diagnosing cases from neighboring counties, … For the sake of our study, we can drop these outliers.

In order to understand these two variables better, we represent them graphically. As continuous variables, a density function seems to be the best way to do so.



Covid has a heterogeneous impact on the United States, with 80% of counties having an infection rate ranging from 22.45% to 38.46%.



We can notice that both rates are distributed following a roughly normal law, centered around 30% for `Infection_rate` and a 1.2% for `Death-rate`. The disparities observed are significant, some counties having a death rate superior to 4%. These disparities are clearly shown by the quantiles of the `Death_rate` variable.

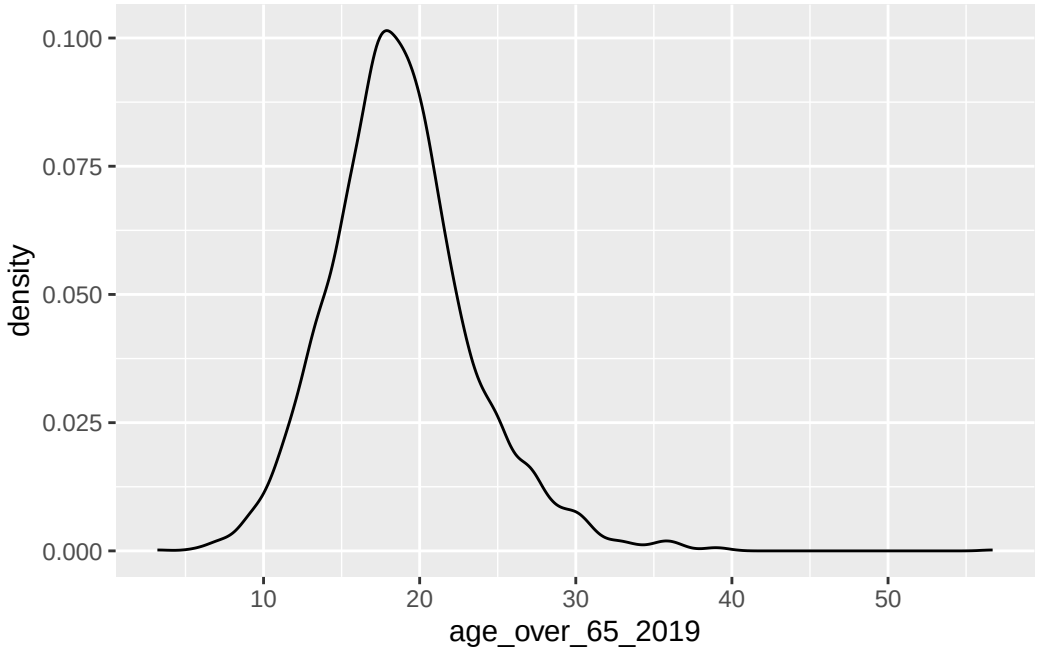
Here, we can notice a clear difference between quantiles: the two extremes are 0.73% and 2.2%. Therefore, we can ask ourselves what can explain such differences between counties.

3. Data Analysis and Interpretation

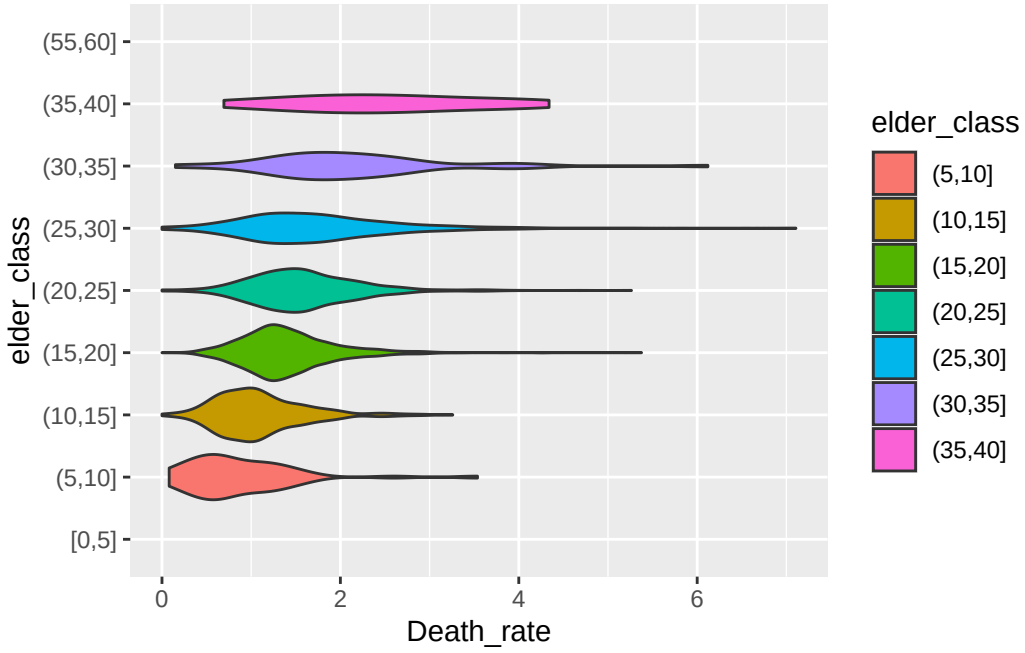
I. Analysis of Demographic Data

Age

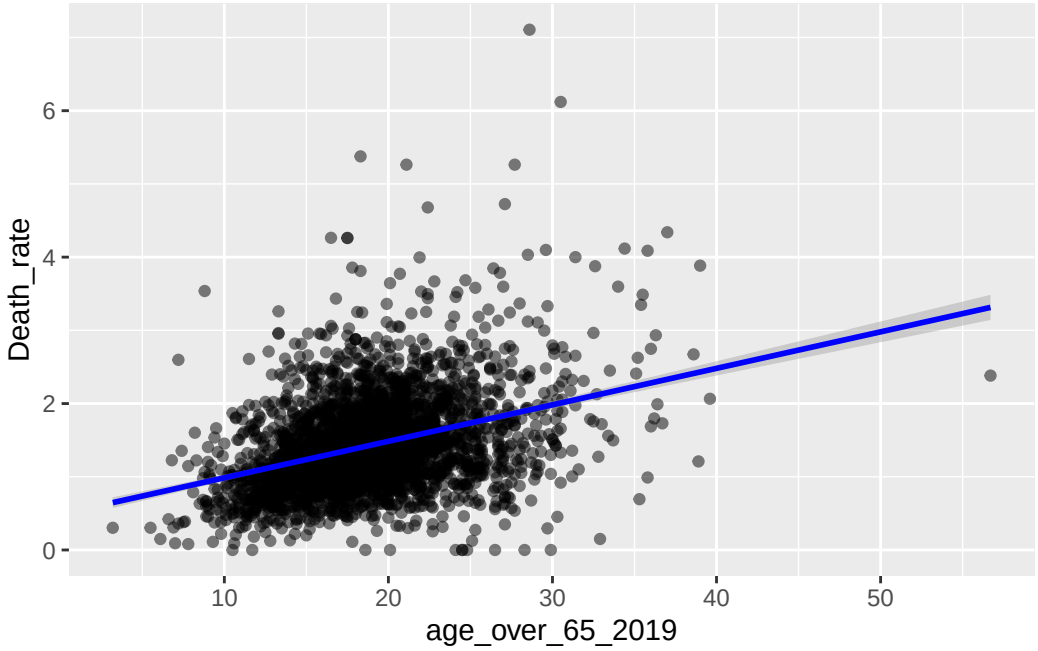
It is safe to state that the age is among the most important factors explaining Covid-19 mortality. Older people represent a majority of fatalities from the disease and are more sensible to its symptoms. Our dataset provides us with various variables regarding age. We can first test how the percentage of people older than 65 years old impact the death rate within a county.



Most counties have an elder population representing between 10 and 30% of its citizens. As a numeric continuous variable, we can establish classes in order to obtain an overview of the variable and its correlation with the death rate from Covid.

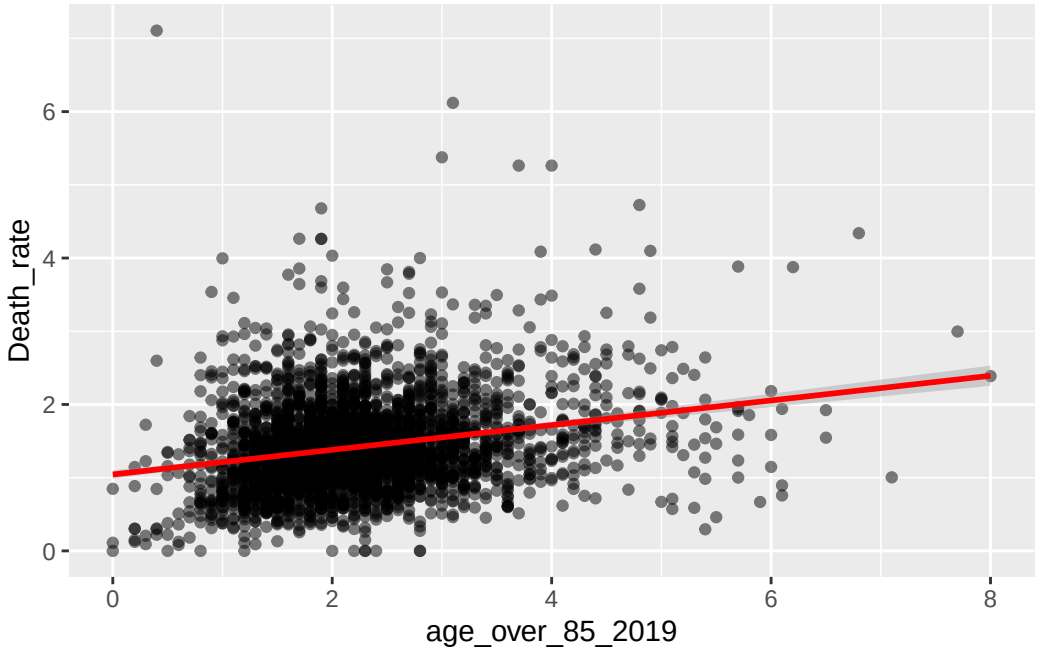


The graph seems to indicate a positive correlation between the two variables: the higher the population over 65 years old, the higher the mortality from Covid. This hypothesis prompts us to test a linear regression model involving the two variables.

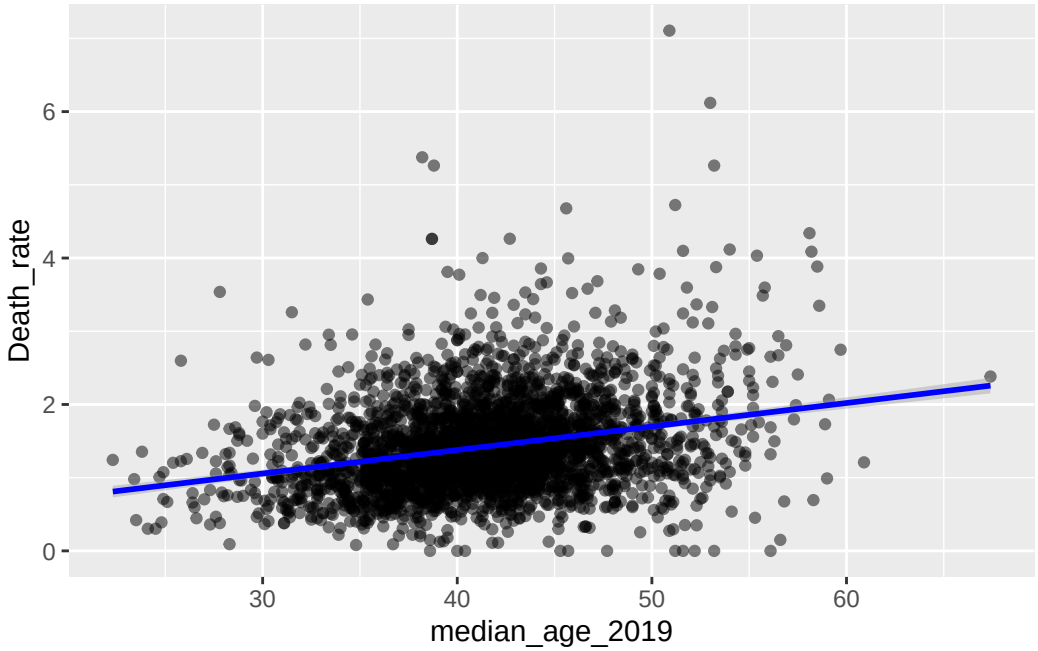


We can observe a clear correlation, although we need to evaluate its precision.

According to these statistics, the correlation between the percentage of people older than 65 years old and the mortality rate within a country is robust. A one-percent increase of the share of elders leads to a 0.05% increase of the mortality rate. Since the latter ranges mostly from 0 to 3, such correlation is far from being uninteresting. However, our model only explains about 13% of the mortality rate. This weak R-squared should warn us about other factors influencing the mortality rate on the county-level. First, we can try to explore a correlation with a more precise age range. Our dataset provides us with the percentage of people older than 85 years old in each county. We can run this linear model.



We can see that the R-squared is even smaller for such regression, at around 5%: to generalize, the percentage of habitants older than 85 years old within a county fails to explain 95% of mortality from Covid. As a final attempt to modelize the relationship between age and Covid mortality, we use the median age of each county. Such analysis is advantageous to the extent that it offers a more precise overview of the demographic makeup of each county.



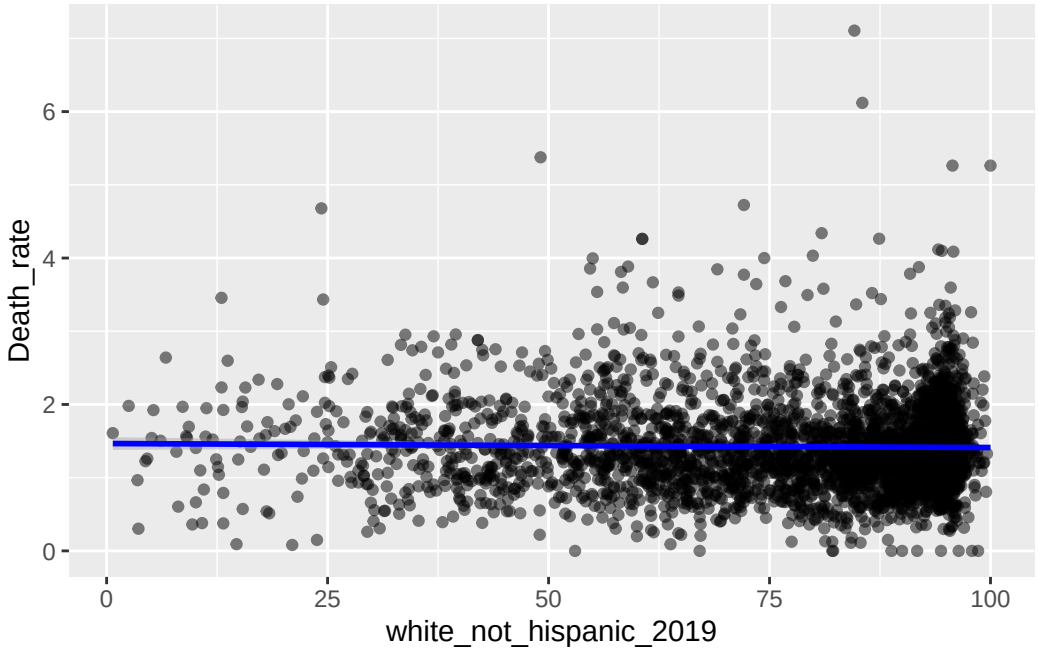
This new regression, once again, fails to explain much of the mortality rate. Furthermore, we have to reject its intercept on the 95% confidence level. Therefore, these three regressions seem to indicate age, while significant, cannot be the only factor explaining why Covid has had different levels of mortality throughout the country. This conclusion prompts us to study other factors not related to age.

Race

It has been established by scholars that ancestry does not play a significant role regarding risks of Covid for a patient. However, race continues to play a major role in American society, characterized by a long-lasting system of racial segregation. Although legally disbanded since the 1960s, the negative effects of it are still visible today. Therefore, it is safe to suppose a greater vulnerability for some ethnic background, such as African-Americans or Latinos. By exploring the racial makeup of the counties, we can test this hypothesis.

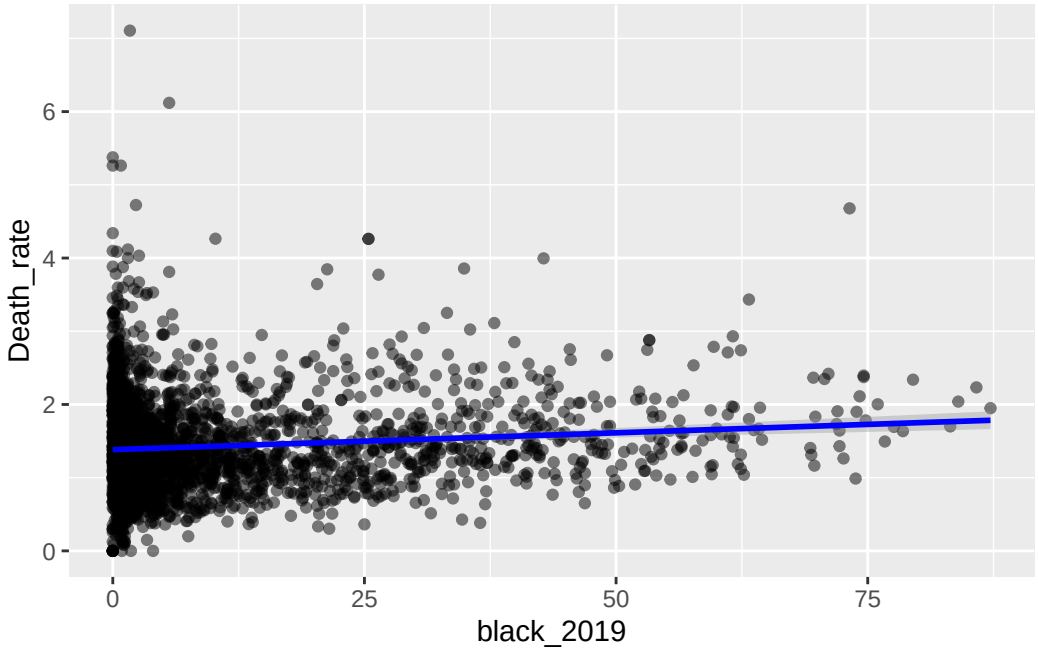
Non-Hispanic Whites

Following the guidelines of the United States census Bureau, the Latinos are not considered as a separate race, but rather as an ethnicity. Therefore, most Latinos declare their race as White, and precise a Hispanic ethnicity. In the case of our study, Hispanics are considered as a distinct group. Therefore, rather than Whites, we study here the correlation between Covid mortality and the percentage of non-Hispanic Whites.

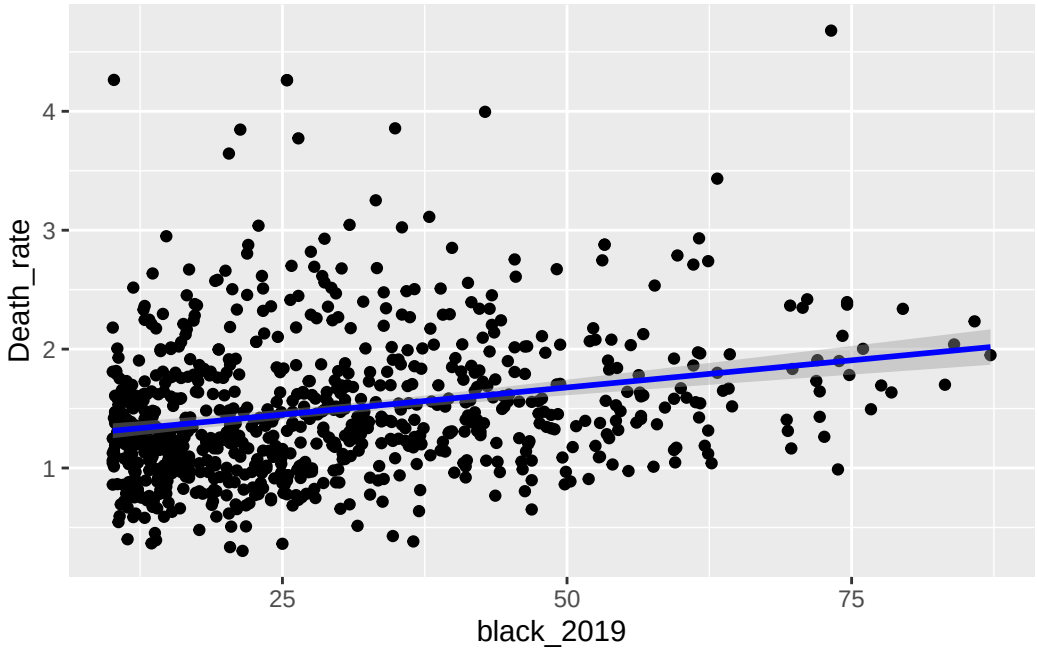


The percentage of non-hispanic Whites within a county does not have any significant effect on the mortality rate. This absence of correlation is not surprising: Non-Hispanic Whites represent a majority of the American population: while by average richer than other ethnic groups, they do not form a homogeneous group. Studying other factors, such as socio-economic ones, would therefore be more relevant in this case.

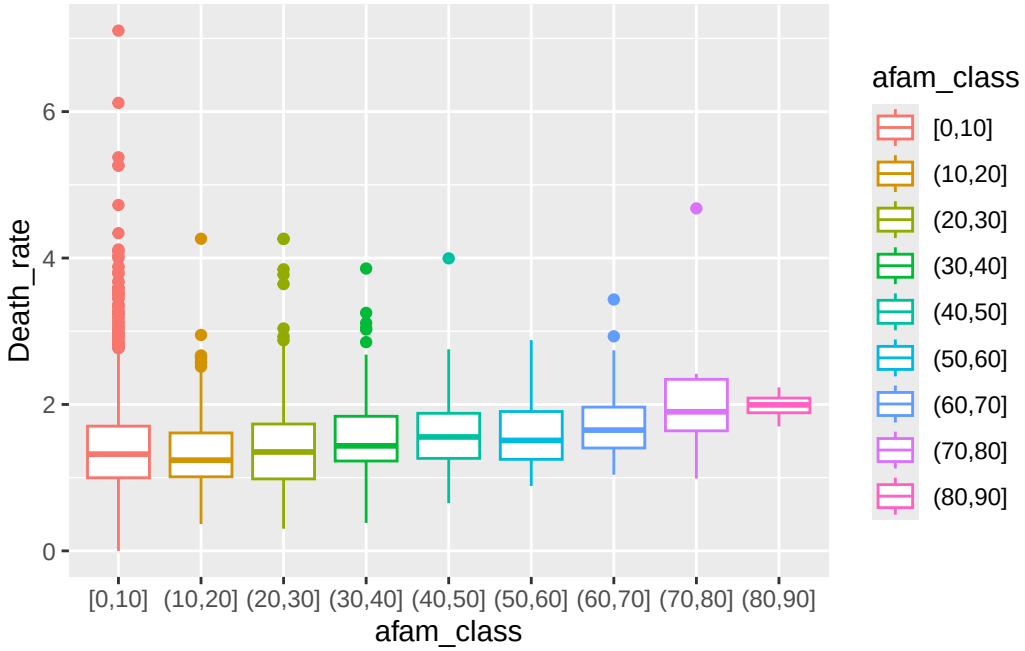
African-Americans



The percentage of African-Americans within a county have a weak, yet significant effect on the risks regarding Covid. A one percent increase in the Black share of the population is expected to increase the mortality rate by 0.004%. In order to obtain a clearer view of the correlation, we can drop the counties less than 10% African-American in their population.

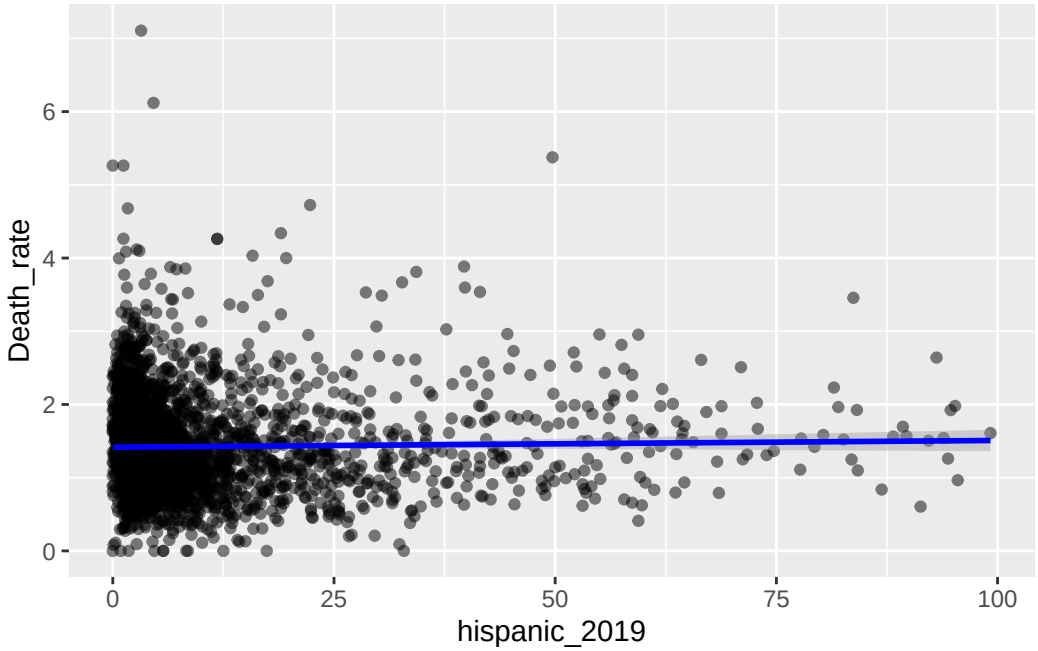


This adjusted model confirms a positive correlation between the two variables: the more African-Americans live in a county, the more sensible to Covid this county is. We can explore further this weak correlation by creating a new variable, `afam_class`. `afam_class` divides the counties in ten classes based on the share of African-American in their population.



This new variable furthers the study of the link between a large African-American population and an important mortality from Covid. As we can see, counties with a black population representing between 70 and 80% of its total population has a median of 1.9% of death among confirmed cases. This median is only 1.32% for counties where the black share of the population does not exceed 10%. Although the small number of counties with a large black population should warn us about the precision of these results, they still show a clear pattern of vulnerability from Covid for Black Americans.

Hispanics



The percentage of Latinos in a county does not have any significant effect on our studied variable.

Conclusion

Following this part of our study, we can conclude by putting the role of race as an explaining variable of the mortality into perspective. Intra-group differences seem to reduce the relevance of such correlation. Only the share of African-Americans play a significant role. This can be explained by the specific challenges faced by the Black community. Aware of the limited impact of race, we can now look for other explaining variables. A study of income and other economic factors would be a better fit to strengthen our model.

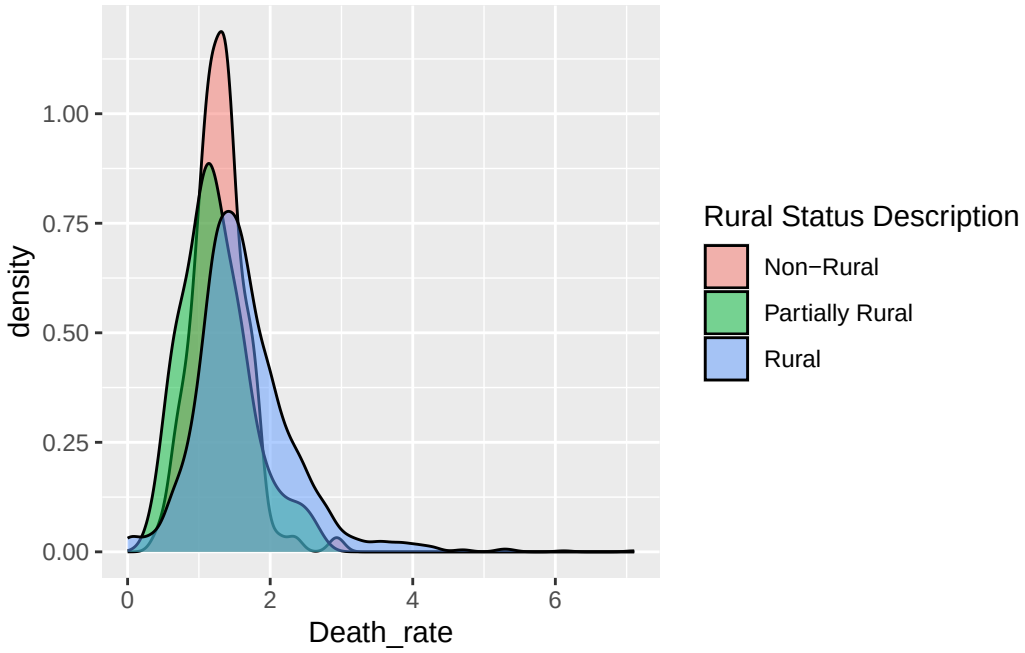
Housing and location

The United States are the the third largest country in the world and include a wide variety of natural regions and communities. As of 2021, 83% of Americans were living in cities and other urban communities. One can expect a difference of mortality and infection between different types of communities, a hypothesis we can test now based on our dataset.

Urban and rural counties

Our dataset only provides partial data regarding the urbanization of counties. Indeed, only Medically Underserved Areas, a classification we will explain more deeply later in our study, have a specified degree of urbanization. Three of them are specified: “Rural”, “Partially Rural” and “Non-Rural”.

Among the 1,729 counties with data provided, a large majority are considered as rural ones by the Department of Health and Human Services. Since only Medically Underserved counties are taken into account here, this fact in itself indicates an overexposure of rural counties to public health issues.



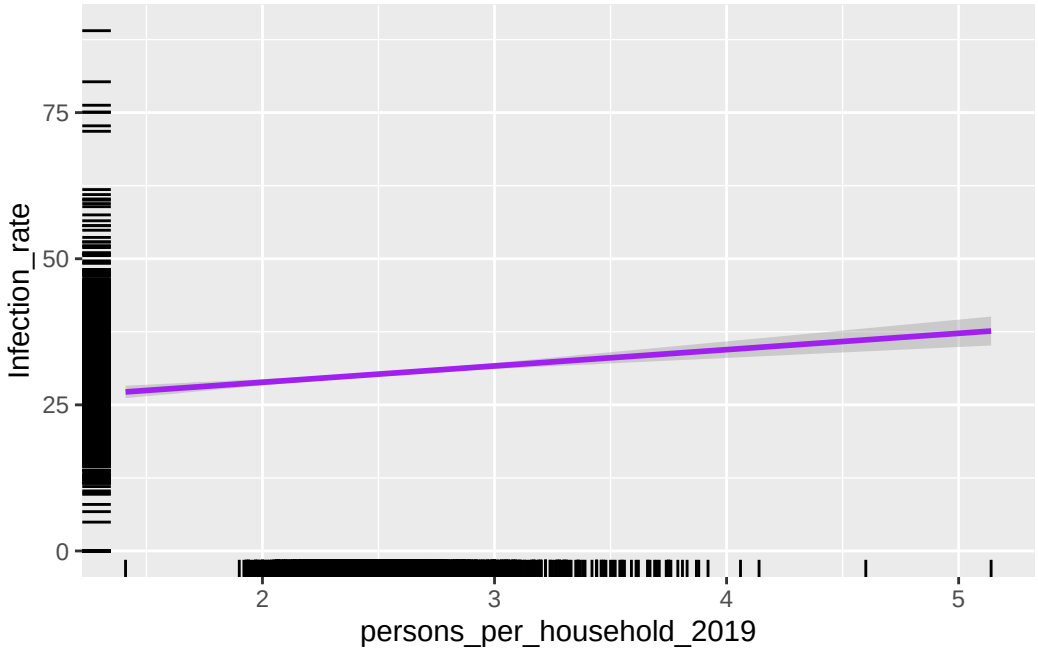
Comparing the density functions of the three different status, however, seems to indicate an absence of clear difference between them. To confirm it, we can compare them with basic statistics:

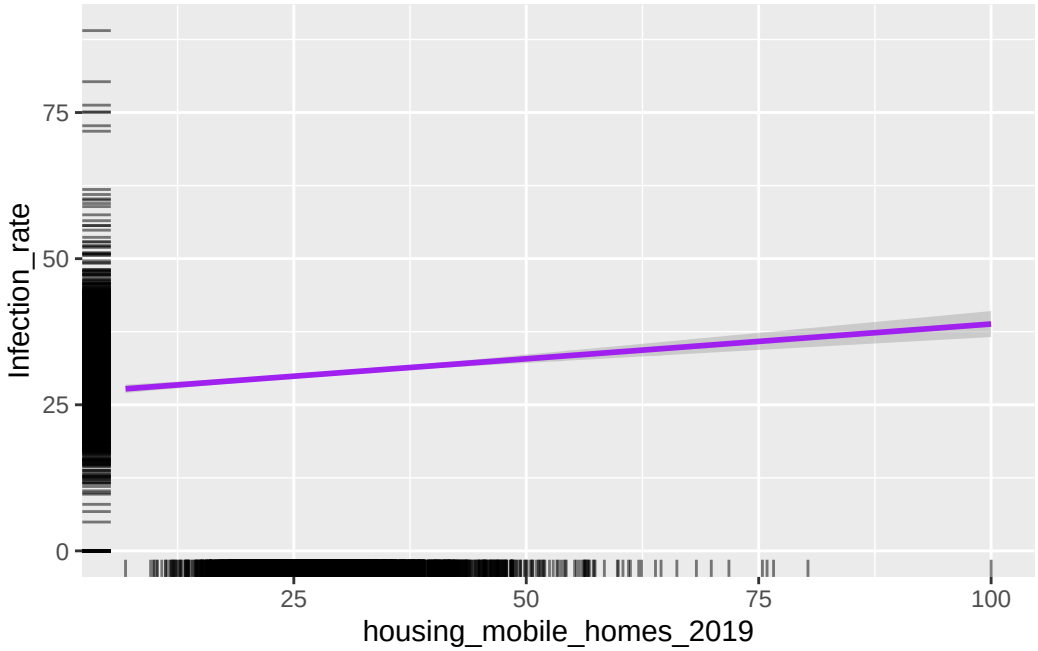
The table shown above proves the absence of clear difference between the three types of counties. Nevertheless, we can still notice a slightly higher median mortality rate in rural communities, but a smaller median infection rate. The latter can be explained by the lower rate of detection of the disease in areas less equipped with medical structures. Therefore, the absence of detection of mild symptoms of Covid might have led to a higher mortality rate among detected cases. Overall, the difference is not large enough to suppose a clear distinction. However, this analysis is only about underserved areas, already more exposed than the rest of the country. Since most rural counties are often among the more exposed, we cannot rule

out the impact of living far from a urban center on a high mortality rate from Covid, but the limits of our dataset prevent us to prove it.

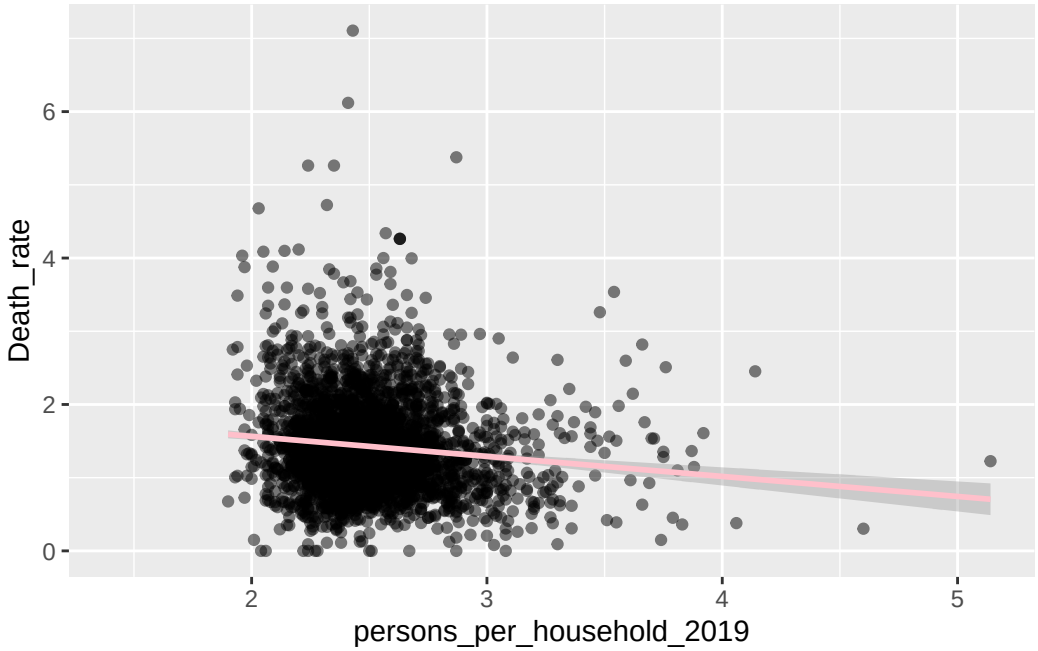
Housing

Covid being an infectious disease transmitted through air, the accommodation can play a major role in its transmission. Overcrowded place, indeed, are places with a high risk of infection. To verify this assumption, we use two variables included in our dataset hinting towards overcrowded accommodations. The first one ,`persons_per_household_2019`, specifies the average number of people living under the same roof. The second, `housing_mobile_homes_2019`, indicates the share of mobile homes in the total number of accommodations.





We can observe a positive correlation between each of these variables and the infection rate, in accordance with our hypothesis. However, we do not have any reason to assume that the mortality rate is positively impacted by the number of people in each household. We can test the correlation in order to obtain a clearer idea.



As expected, the number of persons per household is not positively correlated with the mortality rate. Inversely, rather than a positive correlation, a negative one is observed: ceteris paribus, a county with an average of 3 persons per household is expected to have a mortality rate 0.27% lower than one with an average of 2. This negative correlation can be explained by a higher incidence in counties with more populated accommodations, most notably with children, lowering the share of severe cases.

II. Analysis of Socio-Economic Data

A - Income

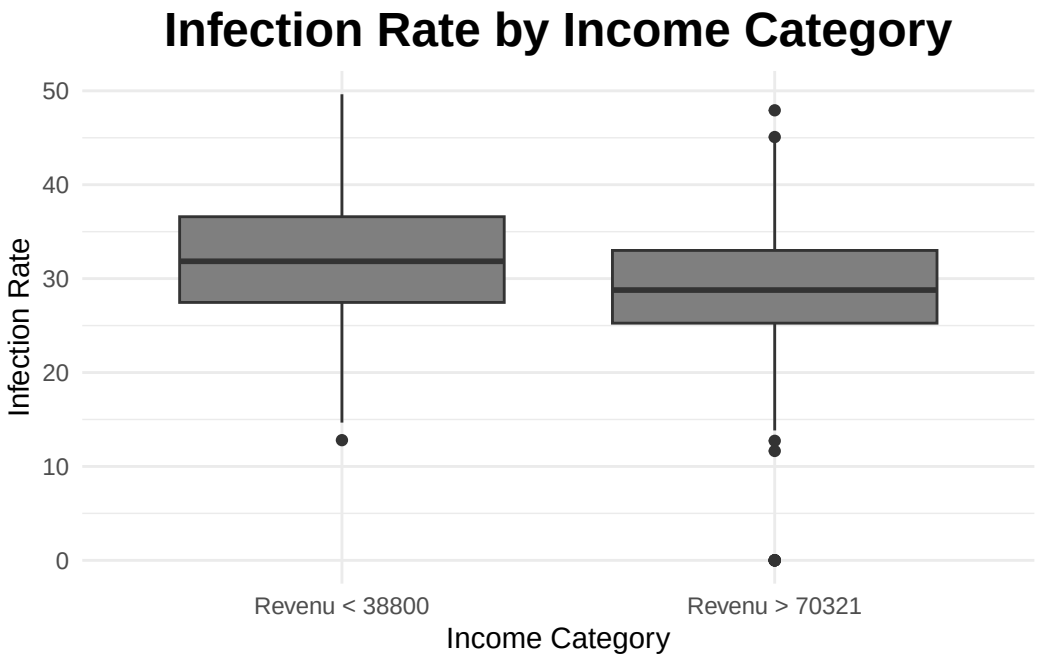
To explore the relationship between economic factors and the impact of COVID-19, this analysis focuses on the variable median_household_income_2019. Median household income serves as a key indicator of the financial well-being of populations. To examine disparities, we will compare COVID-19 death and infection rates between counties in the first and tenth income quantiles, representing the lowest and highest income brackets, respectively.

To investigate the relationship between economic variables and COVID-19, we will analyze the variable "median_household_income_2019."Median household income provides insight into the financial well-being of different populations, to differentiate groups, we will compare the death and infection rates between the 1st and 10th income quantiles.

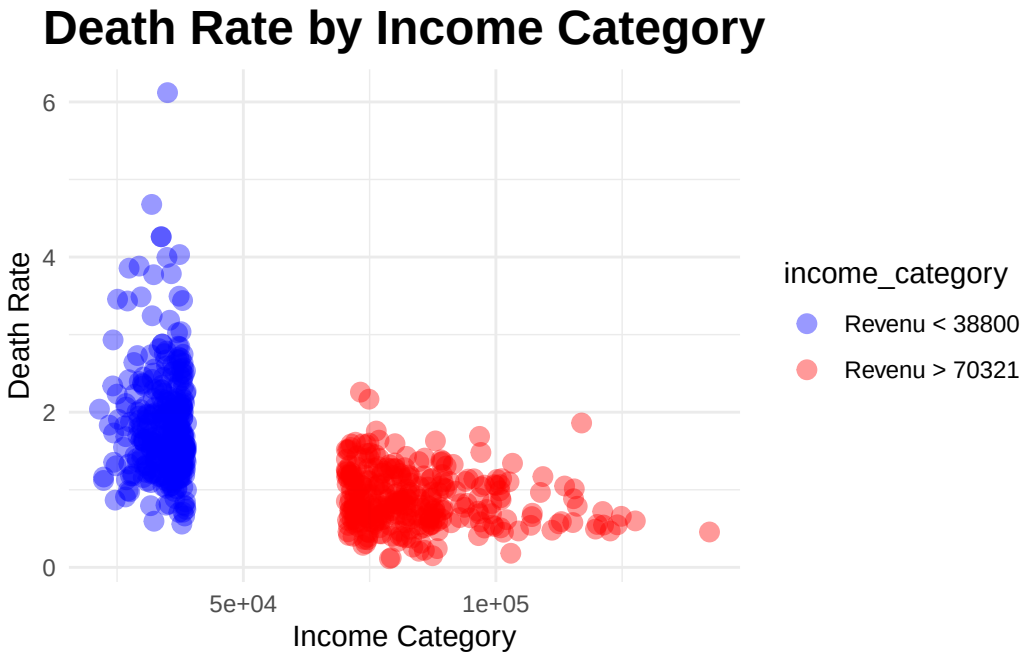
```
   Decile          Boundaries
1      D1     (21504 ; 38078)
2      D2   (38078 ; 42341.8)
3      D3   (42341.8 ; 45889)
4      D4   (45889 ; 49027.8)
5      D5   (49027.8 ; 51734)
6      D6   (51734 ; 54276.8)
7      D7 (54276.8 ; 57756.2)
8      D8   (57756.2 ; 62254)
9      D9   (62254 ; 70155.4)
10    D10  (70155.4 ; 142299)
```

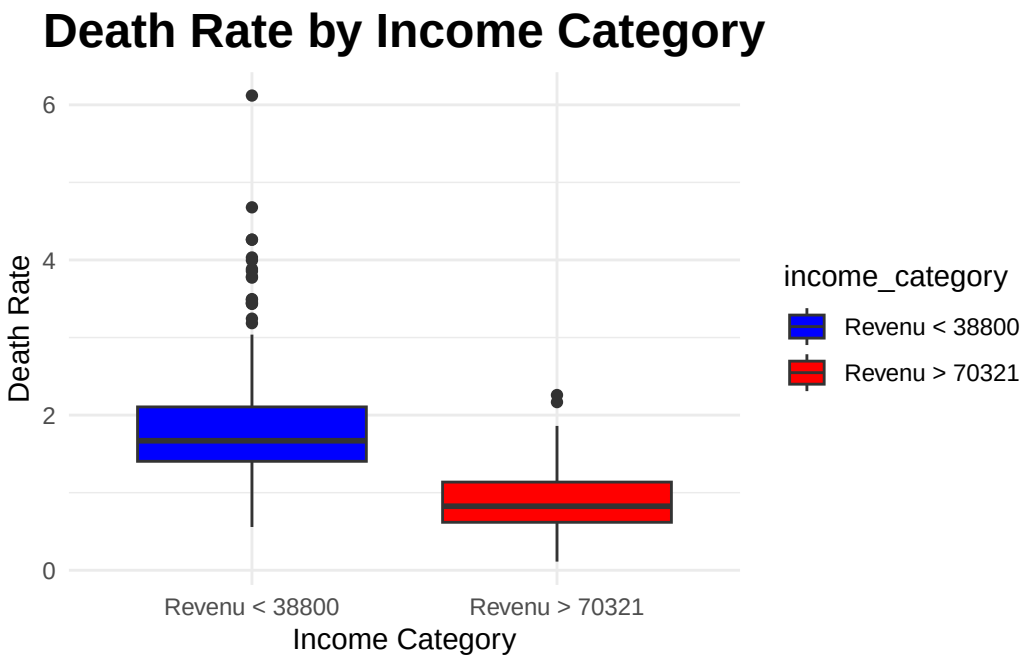
The first decile comprises counties with a median household income below $38,080, representing the lowest 10% of the dataset. Conversely, the tenth decile includes counties with a median household income exceeding $70,321, corresponding to the highest 10% of the data. The following charts illustrate disparities between counties in the first and tenth income deciles with respect to COVID-19 infection and death rates. Notably, the death rate reflects the proportion of COVID-19 cases that resulted in fatalities.

```
# A tibble: 1 x 3
  count_sup count_inf_or_equal total_count
      <int>              <int>       <int>
1       313                315        3147
```



Both Graphs illustrate Infection rates between the first and last decile. It shows no striking differences (mean 32% vs 28,7%)





The data reveal substantial disparities in infection and mortality rates across income categories. Households with low incomes (less than $38,880) experience a mean infection rate of 32.08%, which is notably higher than the 28.68% observed among high-income households (earning more than $70,321). The associated confidence intervals (31.35% to 32.82% for low-income households and 27.89% to 29.47% for high-income households) confirm the statistical reliability of these estimates, although the difference in infection rates is relatively moderate.

In contrast, the disparity in mortality rates is more pronounced. Low-income households report a mean mortality rate of 1.83%, nearly twice that of high-income households, which stands at 0.88%. This significant gap is reinforced by non-overlapping confidence intervals (1.75% to 1.90% for low-income households and 0.84% to 0.92% for high-income households), highlighting the critical impact of economic inequality on severe health outcomes.

These findings underscore the pivotal role of social determinants in shaping pandemic outcomes. While poverty itself may not exert a direct influence, its downstream effects—such as limited health insurance coverage, economic instability, and restricted access to essential resources—play a decisive role in exacerbating health disparities during crises. To deepen this analysis, we will investigate health insurance coverage, with a particular focus on uninsured populations, including vulnerable groups such as the elderly, to assess the impact of gaps in the U.S. healthcare system on the management of critical cases.

B - Insurance

One of the direct consequences of income inequality is the disparity in access to healthcare coverage, which is often closely tied to financial resources. Given that individuals aged 65 and older account for the majority of COVID-19-related deaths, this analysis will focus specifically on this demographic group.

```
# A tibble: 1 x 1
  mean_uninsured_over_65
                   <dbl>
1                  0.497
```

```
Call:
lm(formula = Death_rate ~ uninsured_65_and_older_2019, data = clean_merged)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4285 -0.4028 -0.0869  0.2964  5.6781

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.428533   0.012985  110.01   <2e-16 ***
uninsured_65_and_older_2019 -0.007082   0.012424   -0.57    0.569
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6378 on 3117 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.0001042, Adjusted R-squared:  -0.0002165
F-statistic: 0.325 on 1 and 3117 DF,  p-value: 0.5687
```

The insurance coverage rate among individuals aged 65 and older exceeds 99.5%, largely due to Medicare eligibility, which supplements or replaces private insurance. Regression analysis shows no significant relationship between the proportion of uninsured seniors and COVID-19 mortality rates (p-value: 0.5687; R-squared: ~0), suggesting that uninsurance does not directly impact mortality outcomes. While exploring differences by insurance type would be valuable, the dataset lacks the granularity needed for such an analysis.

C - Poverty

Having analyzed health insurance coverage among individuals aged 65 and older, we now turn our focus to examining poverty within the same demographic. The study of poverty in this age group is particularly relevant as it provides critical insights into economic vulnerability and its potential impact on health outcomes. Poverty among seniors can influence access to resources, quality of living conditions, and overall well-being, all of which play a significant role in shaping resilience to health crises such as the COVID-19 pandemic. Understanding these dynamics is essential for addressing disparities and improving outcomes in this vulnerable population.

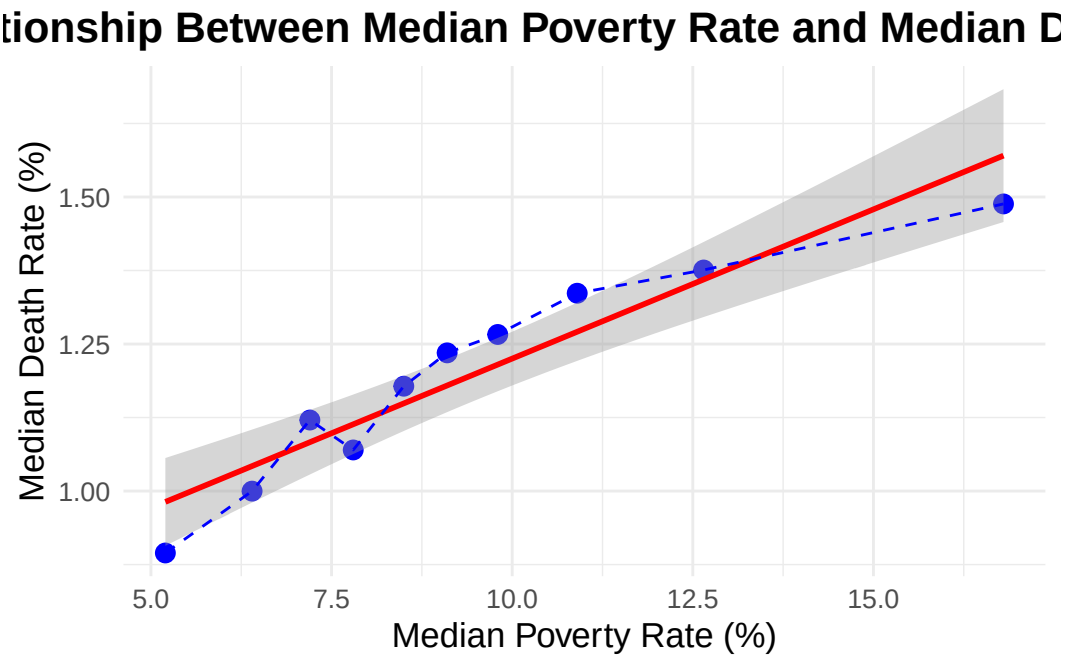
```
# A tibble: 1,248 x 77
    fips state   name         age_over_18_2019 age_over_65_2019 age_over_85_2019
   <dbl> <chr>   <chr>                   <dbl>            <dbl>            <dbl>
 1  1001 Alabama Autauga Cou~             76.2             15                1.6
 2  1003 Alabama Baldwin Cou~             78.3             20                1.9
 3  1005 Alabama Barbour Cou~             79.1             18.6              1.6
 4  1009 Alabama Blount Coun~             76.8             17.9              1.8
 5  1015 Alabama Calhoun Cou~             78.2             17.2              1.8
 6  1017 Alabama Chambers Co~             79               19.2              1.9
 7  1021 Alabama Chilton Cou~             75.9             16                1.4
 8  1031 Alabama Coffee Coun~             76.3             16.5              1.7
 9  1033 Alabama Colbert Cou~             78.9             19.4              2.2
10  1039 Alabama Covington C~             78.1             20.8              2.2
# i 1,238 more rows
# i 71 more variables: age_under_5_2019 <dbl>, asian_2019 <dbl>,
#   avg_family_size_2019 <dbl>, bachelors_2019 <dbl>, black_2019 <dbl>,
#   hispanic_2019 <dbl>, household_has_broadband_2019 <dbl>,
#   household_has_computer_2019 <dbl>, household_has_smartphone_2019 <dbl>,
#   households_2019 <dbl>, households_speak_asian_or_pac_isl_2019 <dbl>,
#   households_speak_limited_english_2019 <dbl>, ...
```

```
# A tibble: 10 x 5
   decile avg_death_rate median_death_rate median_poverty_rate avg_poverty_rate
   <fct>           <dbl>             <dbl>               <dbl>            <dbl>
 1 1               0.963             0.895                 5.2             5.07
 2 2               0.991             1.00                  6.4             6.34
 3 3               1.12              1.12                  7.2             7.13
 4 4               1.12              1.07                  7.8             7.82
 5 5               1.24              1.18                  8.5             8.48
 6 6               1.33              1.23                  9.1             9.11
 7 7               1.28              1.27                  9.8             9.85
 8 8               1.33              1.34                 10.9            10.9
```

```
 9 9                1.39                1.38                12.6                12.7
10 10               1.50                1.49                16.8                18.0
```

```
[1] ""
```

```
`geom_smooth()` using formula = 'y ~ x'
```



This graph represents the median death rate in function of the median poverty rate, each blue point represents a decile. The red lign illustrate the linear regression line, while the grey shadow is the confidence interval. It depicts a correlation, the higher the poverty rate, the higher the death rate.

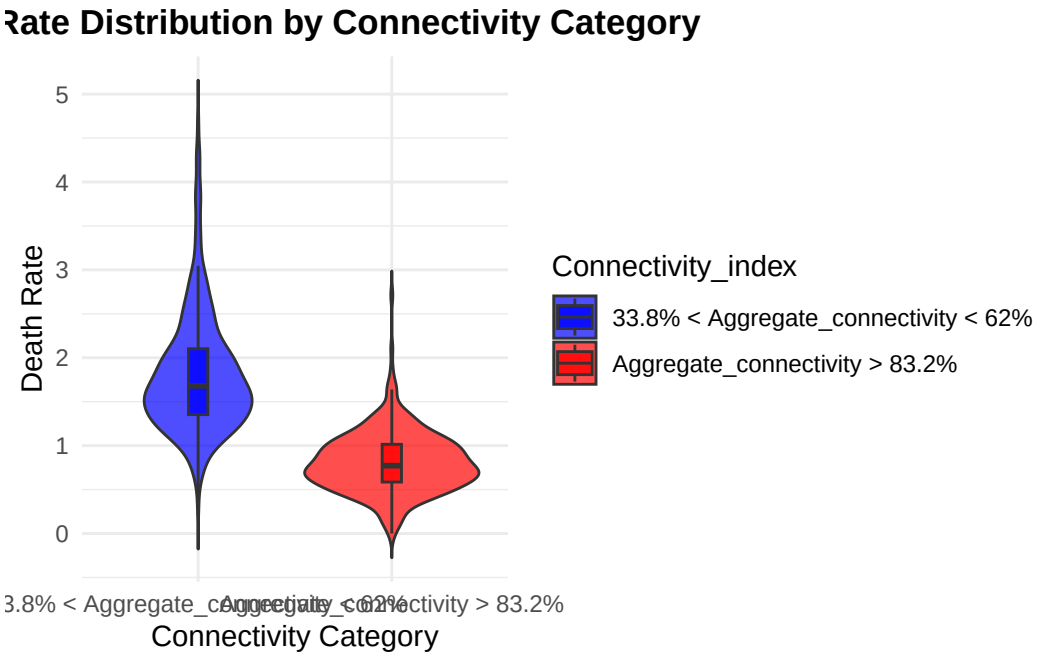
```
Call:
lm(formula = median_death_rate ~ median_poverty_rate, data = summary_by_decile)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08685 -0.04338  0.02269  0.04783  0.06548

Coefficients:
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.717618   0.060749  11.813 2.42e-06 ***
median_poverty_rate 0.050756   0.006097   8.325 3.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06177 on 8 degrees of freedom
Multiple R-squared:  0.8965,    Adjusted R-squared:  0.8836
F-statistic: 69.31 on 1 and 8 DF,  p-value: 3.275e-05
```

The analysis reveals a stark correlation between poverty and COVID-19 mortality. A 1% increase in the median poverty rate corresponds to a 0.05% rise in the death rate. With an R^2 of 89.65%, poverty explains most variability in mortality, underscoring the devastating impact of socioeconomic inequalities during health crises.

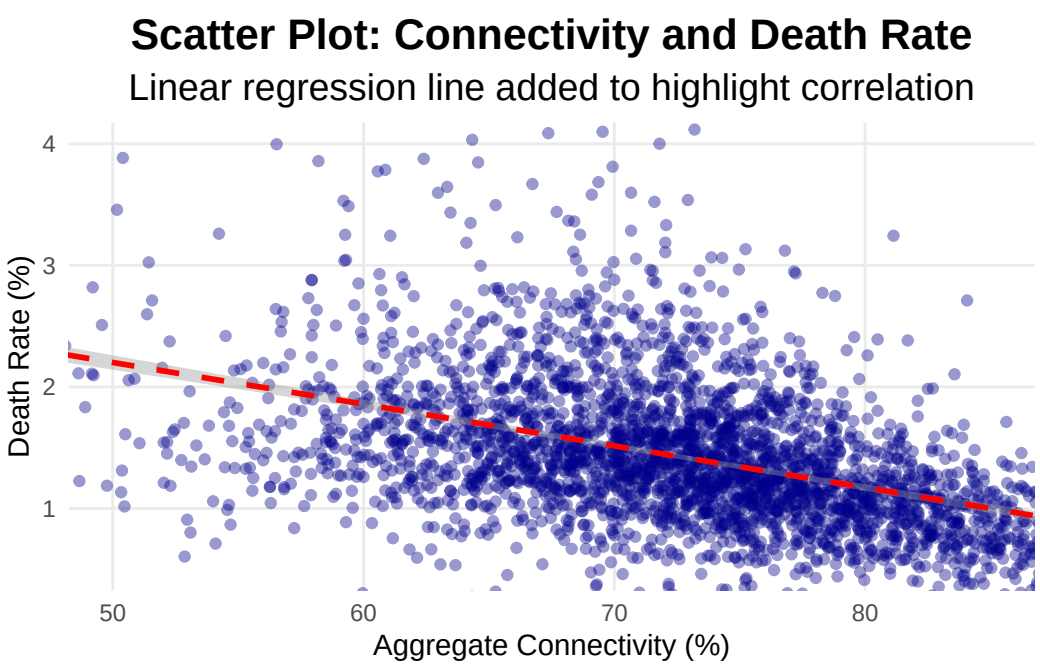
D - Information capacity

Aggregate_connectivity quantifies household digital access by averaging three key components: broadband, smartphone, and computer availability in 2019. The index ranges from 0.33 to 0.94 (indicating near-complete connectivity), excluding missing values. This measure provides a valuable tool for examining digital inclusion and its broader socioeconomic implications with precision and clarity.



This chart shows Death rates differences in function of connetivity. It shows highly connected counties are less impacted than the least ones.

```
`geom_smooth()` using formula = 'y ~ x'
```




```
Call:
lm(formula = Death_rate ~ Aggregate_connectivity, data = quantiles2_filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39268 -0.31305 -0.07433  0.21966  2.70029

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.667381   0.106739   34.36   <2e-16 ***
Aggregate_connectivity -0.032881   0.001462  -22.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5477 on 618 degrees of freedom
  (8 observations deleted due to missingness)
```

```
Multiple R-squared:   0.45, Adjusted R-squared:  0.4492
F-statistic: 505.7 on 1 and 618 DF,  p-value: < 2.2e-16
```

This analysis highlights the profound impact of connectivity on mortality during the pandemic. Low-connectivity areas (below 62%) experienced a mean death rate of 31.96%, significantly higher than the 28.51% observed in high-connectivity areas (above 83.2%). Each percentage point increase in connectivity reduces death rates by 0.0328%, explaining 45% of mortality variation. It can be explained by various factors : the is certainly a correlation between average age and connectivity as aged people own less digital products than younger ones. digital products are costly, so generally poor people have less electronical devices than rich ones, and poor people have generally higher obesity, chronical diseases which play an important role in COVID-19 mortality Moreover, strong connectivity also supports remote working, reducing exposure risks and enhancing resilience.

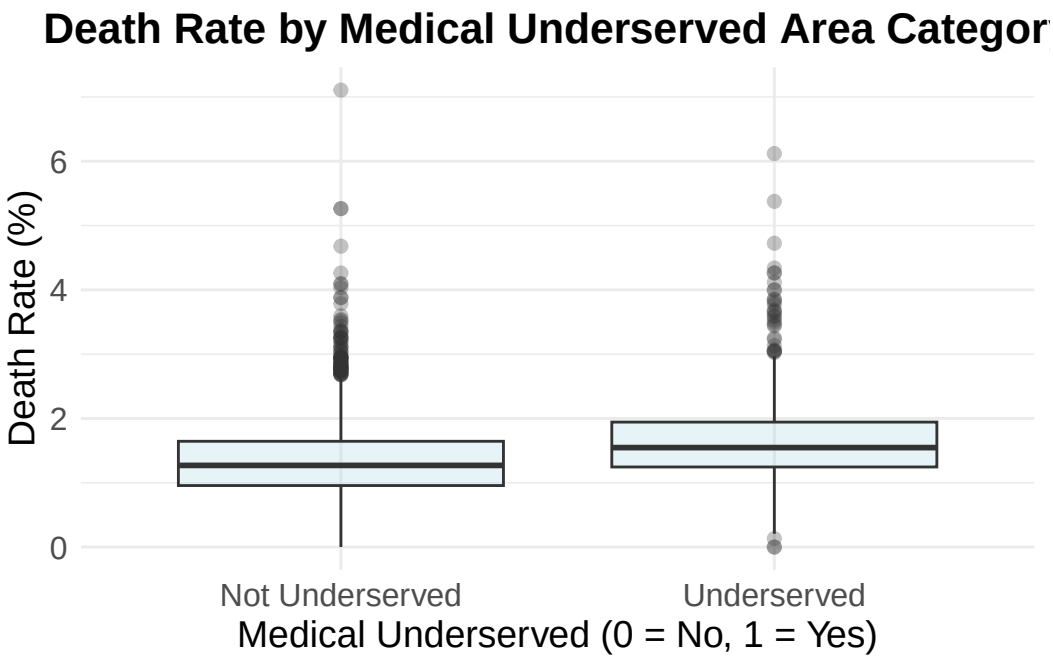
III. Health System

A - Healthcare Access Disparities: Medically Underserved Counties Compared to Others

Analyzing disparities between medically underserved and other counties highlights healthcare inequalities, revealing critical gaps in resources, access, and outcomes.

```
# A tibble: 2 x 11
  medically_underserved_d~1 mean_infection_rate2 sd_infection_rate2 n_infection2
                      <dbl>                <dbl>              <dbl>        <int>
1                         0                 30.3               6.62         2394
2                         1                 31.7               7.30          714
# i abbreviated name: 1: medically_underserved_dummy
# i 7 more variables: lower_ci_infection2 <dbl>, upper_ci_infection2 <dbl>,
#   mean_death_rate2 <dbl>, sd_death_rate2 <dbl>, n_death2 <int>,
#   lower_ci_death2 <dbl>, upper_ci_death2 <dbl>
```

The data reveal that medically underserved counties have higher mean infection rates (31.69%) and death rates (1.68%) compared to non-underserved counties (30.26% and 1.36%, respectively). These findings highlight an impact of the pandemic on underserved areas, emphasizing the need to address healthcare access disparities.



The difference in death rates between non-underserved counties (1.35%) and medically underserved counties (1.67%) is 0.32 percentage points, representing a 23.7% higher death rate in underserved areas. This emphasizes the significant impact of healthcare disparities on COVID-19 mortality outcomes.

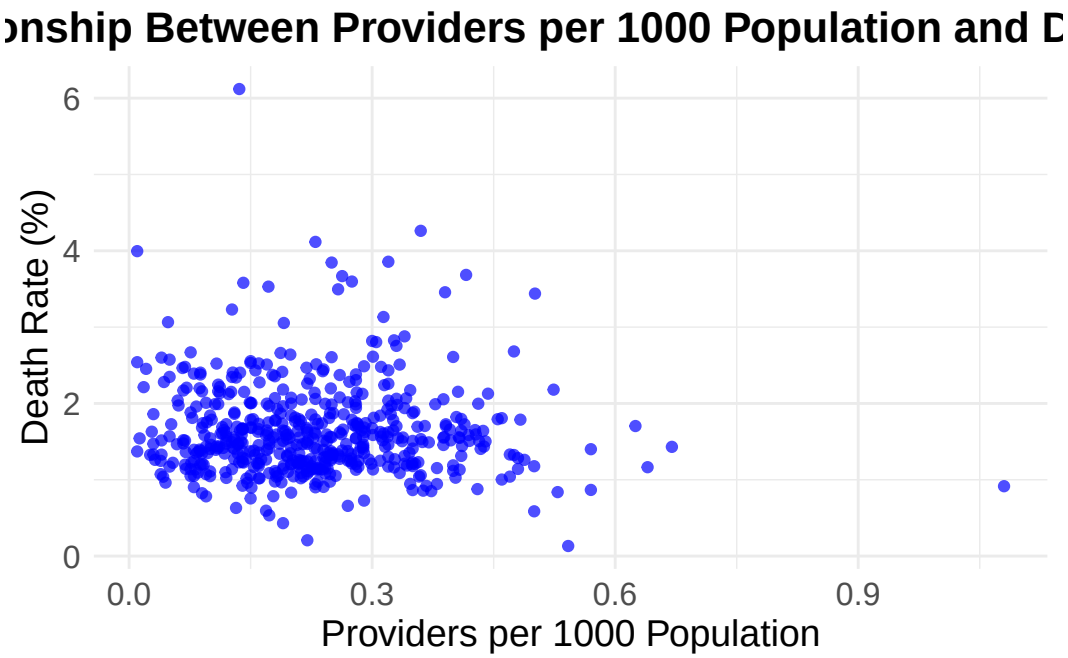
```
model <- lm(Death_rate ~ medically_underserved_dummy , data = clean_merged)
summary(model)
```

```
Call:
lm(formula = Death_rate ~ medically_underserved_dummy, data = clean_merged)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6727 -0.4065 -0.0946  0.2912  5.7554

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.35121    0.01272  106.27   <2e-16 ***
medically_underserved_dummy  0.32148    0.02654   12.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6233 on 3117 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.04496,   Adjusted R-squared:  0.04466
F-statistic: 146.8 on 1 and 3117 DF,  p-value: < 2.2e-16
```

B - Number of providers

”



The scatterplot demonstrates no clear correlation between the number of providers per 1000 population and the death rate. The data points are widely dispersed, and there is no discernible trend or pattern to indicate a strong or consistent relationship between these two variables.

The tests conducted in this study reveal no evidence of a correlation between medical coverage and COVID-19 mortality. However, the absence of evidence does not necessarily imply that no relationship exists. One critical factor to consider is the shortage of medical personnel, which likely played a significant role during the first wave of the pandemic, when healthcare systems were overwhelmed. It is important to note that the data used in this analysis extends only up to 2022, a period significantly removed from the initial wave. Consequently, it would be valuable to replicate this analysis using data specifically from the first wave to better understand the potential relationship between medical coverage and mortality during the most critical phase of the pandemic.