COVID-19 Mortality Analysis Report

A Statistical Investigation using R

Dane Pearson Dhriti Avala

2025-12-09

Introduction

The COVID-19 pandemic has had dramatic and lasting effect on the United States, specifically in public health, society, and policy. Between the years of 2020 and 2023, over one million Americans died due to COVID-19, with mortality patterns varying based on demographic groups, regions, and seasonality. Being able to understand these patterns and how specific characterisitcs affected mortality rates is important in characterizing the impact of the pandemic and evaluating public health implementations to combat the virus. The purpose of this study is to analyze these mortality patterns utilizing provisional death counts provided by the Center of Disease Control (CDC) in addition to state level population estimations.

With the help of prior research of the CDC and National Center of Health Statistics (NCHS), we know that the older generation seemed to have experienced significantly higher death counts, frequency increasing with age. Additionally, studied have shown that there is a higher mortality rate in men than women, potentially hinting at differences in comorbidity, immune response, and behavioral factors. Geographic variation has also been reported, where the South and Midwest seemed to have experienced mortality rates, potentially due to differences in vaccination rates, population density, healthcare policies, and the timeline of policy implementation.

The primary dataset used in this study is "Provisional COVID-19 Deaths by Sex and Age", (CDC, National Center for Health Statistics, 2020). This file contains key variables such as monthly and yearly death count for each U.S state from 2020 to 2023, broken down further with supporting variables such as sex and age group. The dataset contains over 138,000 observations and CDC-supressed values (NA values) in low population subgroups in efforts to preserve confidentiality and avoid systematic bias. In order to estimate comparalbe mortality rates across the states and demographics, the data was left-joined with 2020-2023 state population estimates from the U.S Census (US Census Beureau, 2024) on the "State" variable. By joining the datasets, we were able to calculate death rates of each state in different months and years. In addition to this, states were also grouped into four regions: West, MidWest, South, and Northeast to help us perform by-region analysis.

Our study aimed to focus on four main objectives. First, we aimed to analyze the patters in U.S COVID-19 using detailed analysis and graphical visualizations of death rates across all months from 2020 to 2023. Second, we aimed to examine demographic patterns by analyzing deaths across age groups and sex. Third, we used visualizations such as heatmaps to depict state-level variation in COVID-19 death rates. Lastly, we employed a variety of modeling approaches such as Multiple Linear Regression, Poisson Regression, Negatve Binomial Regression, and Two-Way ANOVA to quantify the effects of year, region, age group, and sex on COVID-19 mortality. Ultimately, these analyses helped to provide a comprehensive understanding of factors associated with COVID-19 death rates in the United States and revealed how mortality has changed over the course of the pandemic.

Exploratory Data Analysis

Table 1. Summary of Provisional COVID-19 Deaths

Variable	Categories
Group	By Month, By Year, By Total
Sex	All Sexes, Female, Male
Age Group	0–17 years, 1–4 years, 15–24 years, 18–29 years, 25–34 years, 30–39 years, (Other)

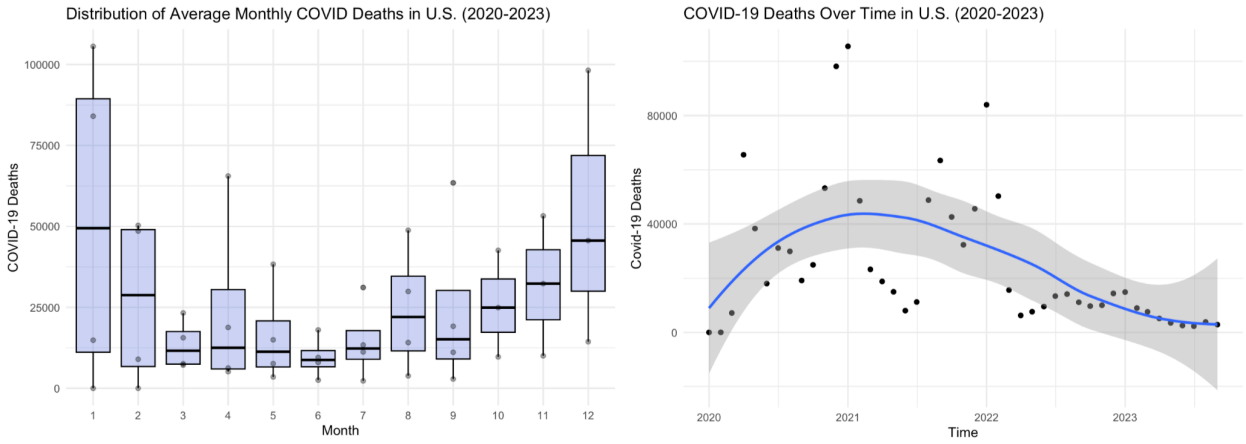
Table 2. Summary of State Populations Data

Variable	Min	Median	Max
Pop 2020	577664	6025563	331526933
Pop 2021	579548	6025186	332048977
Pop 2022	581629	6027262	333271411
Pop 2023	584057	6045604	334914895

Our original, unaltered "Provisional COVID-19 Deaths by Sex and Age" (CDC, National Center for Health Statistics, 2020) dataset from the CDC has 138,000 rows and 16 columns. The key columns we focused on for this project are "Group" (whether data was measured by Month, by Year, or Total), "Year", "Month", "State", "Sex", "Age Group", and "COVID.19.Deaths". Before conducting the main exploratory analysis, we examined the number of missing values (NA) in each of the variables of interest. We did not count the number of missing values of the "Year" and "Month" columns, as they have missing values depending on whether the respective row was measured by year or by month. Each key column had no missing values, aside from "COVID.19.Deaths", which had 39,430 rows with no value.

Following this, we explored how the frequency of deaths changed over time in the data, plotting the average total deaths across in the entire US for each month, from 2020-2023. We found that deaths from COVID-19 spiked drastically in late 2020 and early 2021, with the death rate lowering in mid 2021 and leveling off by early 2023.

The decline COVID-19 deaths in mid-2021 coincided with the release and administration of the vaccine, as more than 70% of the U.S. population had already received at least one vaccine dose, and approximately 60% were fully vaccinated by this time. These developments may have contributed to the observed reduction in mortality, as vaccines have been shown to substantially decrease the risk of severe illness, hospitalization, and transmission (CDC, 2025). Although this was insightful, we wanted to dive deeper by examining the distribution of each month's COVID-19 deaths in the US, averaged across 2020-2023. By creating a boxplot for each month, a clear seasonal pattern was revealed: mortality was highest in December and January and lowest during the summer months, particularly June.

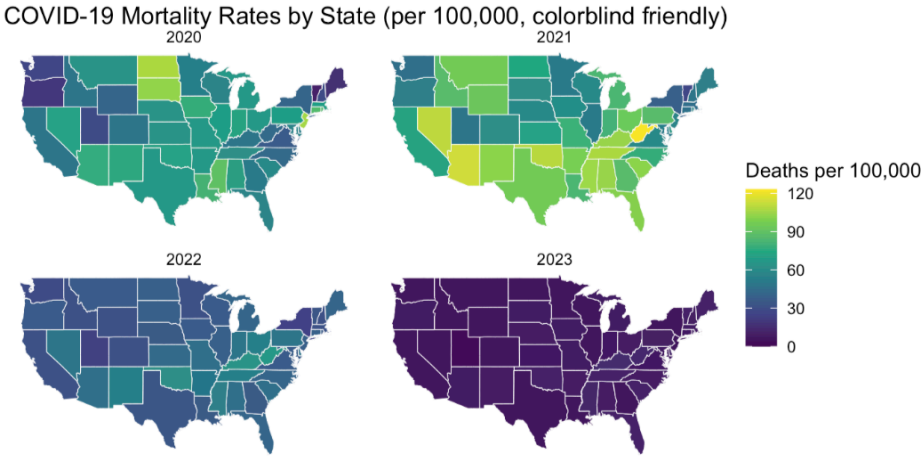


This pattern indicates that higher COVID-19 mortality in the United States was associated with colder periods of the year, whereas lower mortality aligned with warmer periods. Seasonal factors may help explain

this, as colder and drier conditions can weaken immune defenses and aid viruses in surviving longer. Increased time inside during colder months may have also promoted greater transmission of the COVID-19 virus.

Next, we analyzed how COVID-19 mortality varied across the different age groups. We found that older ages were directly associated with increased deaths. Those age 85 and older accounted for 311,863 total COVID-19 deaths across all states from 2020-2023 in the United States, while individuals younger than 35 accounted for 16,735 COVID-19 deaths.

Finally, we examined how mortalities varied both temporally and spatially in the United States by creating choropleth heatmaps of the mortality rate of each state, for each year in the dataset. By joining our original dataset with data on 2020-2023 state population estimates from the U.S Census, we were able to calculate the COVID-19 deaths per 100,000 people in each state. Our heatmap revealed that in 2020 and 2021, mortality rates were higher and demonstrated substantial variation across states, with states in the Midwest, South, and Northeast exhibiting the highest rates. By 2022, overall mortality levels declined across the United States, though southern states still maintained higher mortality rates. In 2023, mortality rates were lower nationwide with little-to-no variation between states.



Preliminary patterns in the data indicated that year, region, age group, and sex were important sources of variability in COVID-19 mortality. The substantial variation across states and demographic groups drove us to use modeling approaches suited for count data and population-standardization.

Methods

In order to understand how COVID-19 mortality varies across time, demographics groups, and U.S regions, we used a combination of exploratory visualization and statistical modeling techniques. Each method was selected accordingly to adhere to the structure of the dataset, count based, non-normally distributed outcomes with strong demographic groupings. We conducted our analysis in R using the following packages: tidyverse (Wickham et al., 2019) for data manipulation and visualization, lubridate (Grolemund & Wichhamm, 2011) for date handling, ggplot (Wickham, 2016) for plotting, maps (Becker et al, 2023) for state analysis, MASS (Venables & Ripley, 2002) for Negative Binomial modeling, patchwork (Pedersen et al., 2024) to combine plots and knitr (Xie, Y., 2024) for embedding code. For formatting tables we used flextable (Gohel, et al., 2024), modelsummary (Arel-Bundock, V., 2022), and kableExtra (Zhu, H., 2024).

Multiple Linear Regression (MLR)

As a baseline model we fit a multiple linear regression predicting standardized mortality rates:

$$Y_i = \beta_0 + \beta_1 \text{Year}_i + \beta_2 \text{AgeGroup}_i + \beta_3 \text{Sex}_i + \beta_4 \text{Region}_i + \varepsilon_i$$

We included Year as a factor rather than a numeric variable because early exploratory analysis showed a non-linear decline in mortality rates over time. Mulitple Linear Regression provides estimates that quantify differences between demographics and geographic subgroups. Additionally, MLR assumes linearity between

the predictors and outcomes, error homoscedasticity, and normally distributed residuals. We developed some diagnostic plots which indicated violations of linearity and constant variance, suggesting that linear regression may provide a biased standard error, pushing for the use of a count-based generalized linear model.

Poisson Regression

Since COVID-19 death counts are non-negative integers, Poisson was an appropriate next step for us. We modeled observed COVID-19 deaths as:

$$Y_i \sim \text{Poisson}(\lambda_i), \qquad \log(\lambda_i) = \beta_0 + \beta_1 \text{Year}_i + \beta_2 \text{AgeGroup}_i + \beta_3 \text{Sex}_i + \beta_4 \text{Region}_i + \log(\text{Population}_i)$$

The log link ensures that the predicted count values are positive, and the offset term log(Populationi) adjusts for differences in population size so that expected deaths scale according to the state's population. Poisson regression is appropriate when the variance and mean of the count data are similar, however, COVID-19 mortality data often has a higher variance than mean, leading to bias estimations and our consideration for a model that is a bit more flexible.

Negative Binomial Regression

Due to overdispersion in mortality rate, we implemented a Negative Binomial regression model using the MASS package. This model assumes:

$$Y_i \sim \text{NegBin}(\mu_i, k), \qquad \log(\mu_i) = X_i\beta + \log(\text{Population}_i),$$

where (k) is the dispersion parameter. The Negative Binomial model incorporates an additional variance term:

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{k}$$

making it a more flexible model in comparison to the Poisson model. This model was able to provide a better fit and lower AIC values, proving to be more reliable and a better predictor.

Two-Way ANOVA and Interaction Effects

To understand whether mortality differed based on age group, sex, and region, as well as understanding how these factors interacted, we conducted a Two-Way ANOVA.

$$Y_{ijk} = \mu + \alpha_i(\text{AgeGroup}) + \beta_j(\text{Sex}) + \gamma_k(\text{Region}) + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijk}$$

Although death count is discrete, our dependent variable was mean deaths conditional on demographic subgroups, making ANOVA valid for comparison of group level means. Our two-way ANOVA exhibited a significant interaction between Age Group x Sex, supporting the idea that mortality differences across age groups depends on sex, and vice versa.

Spatial and Panel Data Methods

We merged the CDC dataset with the U.S state populations from the maps package in order create a heatmap to visualize geographic variation in mortality rates. In order to do this, we had to standardize state names, convert populate estimates into long format, and compute yearly death rates per 100,000 residents.

We also developed a region-level panel dataset by aggregating monthly deaths and population counts across states within each region. This allowed us to create a visualization displaying month-to-month trends and regional differences. We also used ggplot2 to plot mortality trajectories, dedicating a line to each region, allowing for clear comparison and pattern analysis.

Why These Methods

Using a combination of linear regression, Poisson and Negative Binomial linear models, ANOVA, and panel-data visualization allowed us to conduct a thorough analysis of this multidimentional dataset. We chose Poisson and Negative Binomial models since COVID-19 mortality is count-based, high skewed, and overdispersed, something that classic linear models would struggle to handle. ANOVA and visualizations helped us to provide insightful subgroup comparisons. These methods aligned with our goals of quantifying differences across demographic groups, regions, and years, as well as indentifying which factors had strong influences on COVID-19 mortality between 2020 and 2023.

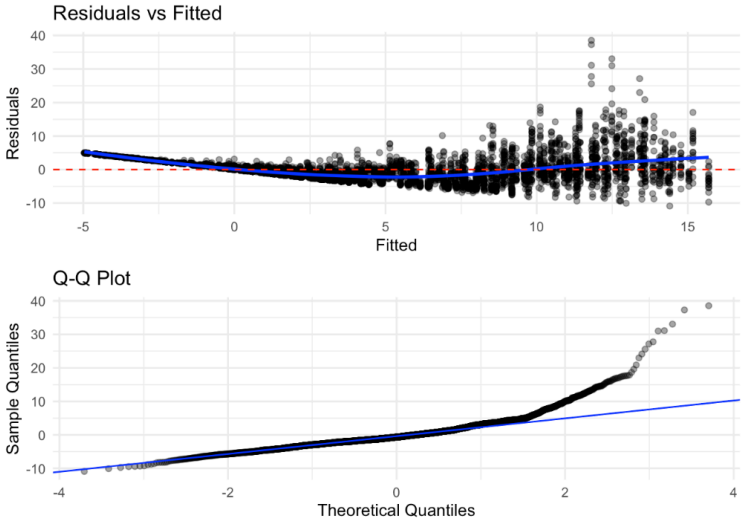
Results

Multiple linear Regression

With an adjusted R^2 of 0.636, our multiple linear regression model explained 63.6% of the variance in COVID-19 mortality rates. The results demonstrate that mortality did not follow a uniform national pattern but instead shifted over time and varied across demographics. Mortality rates rose above the 2020 baseline and reached their maximum in 2021, then declined quickly in 2022 and 2023. Age was the strongest predictor (both $p < 0.001$), as those aged 50-64 experienced rates 7.17 per 100,000 higher than baseline, while those 85+ showed rates 11.45 per 100,000 higher. States in the South had higher mortality than those in the Midwest, while the Northeast and West experienced significantly lower rates. This suggests potential regional disparities in exposure or health policies. Finally, males had consistently higher mortality than females.

Multiple Linear Regression Results

Coefficient	Estimate	P-value
(Intercept)	1.03412	0.002121
factor(Year)2021	1.62987	< 2e-16
factor(Year)2022	-1.70353	< 2e-16
factor(Year)2023	-4.88121	< 2e-16
Age.Group1-4 years	-0.05610	0.882791
Age.Group15-24 years	0.04009	0.921787
Age.Group18-29 years	-0.44203	0.263045
Age.Group25-34 years	-0.36135	0.349848
Age.Group30-39 years	-0.11993	0.753292
Age.Group35-44 years	0.37865	0.315075
Age.Group40-49 years	1.12080	0.002462
Age.Group45-54 years	2.19247	2.02e-09
Age.Group5-14 years	0.21367	0.592276
Age.Group50-64 years	7.17042	< 2e-16
Age.Group55-64 years	5.60433	< 2e-16
Age.Group65-74 years	9.33868	< 2e-16
Age.Group75-84 years	10.93079	< 2e-16
Age.Group85 years and over	11.45333	< 2e-16
Age.GroupUnder 1 year	0.19791	0.620597
regionnortheast	-0.67781	0.000117
regionsouth	0.65506	1.08e-05
regionwest	-0.63542	6.15e-05
SexMale	0.91958	< 2e-16



The Residuals vs Fitted plot demonstrates that our MLR model clearly violates the asumption of linearity. The wavy LOESS curve demonstrates an uneven spread around the residuals = 0 line, indicating that the model is potentially overpredicting in some ranges and underpredicting in others. The assumption of constant variance (homoscedasticity) is also violated, exhibited by the widening funnel shape in the plot. This means that the variance in error grows as the predicted mortality increases. The Q-Q plot demonstrates that our MLR model also violates the asumption of normality of the residuals. The data is light-tailed, suggesting the data may be right-skewed and the model potentially underestimating high death-rate observations.

Poisson Regression Results

Coefficient	Estimate	IRR	P-value
(Intercept)	8.593	4.61e+372	<2e-16
SexMale	0.2448	1.277	<2e-16
regionnortheast	-0.1501	0.861	<2e-16
regionsouth	0.09375	1.098	<2e-16
regionwest	-0.1707	0.843	<2e-16
Year	-0.4322	0.649	<2e-16
Age.Group1-4 years	-3.727	0.024	<2e-16
Age.Group15-24 years	0.6648	1.944	<2e-16
Age.Group18-29 years	1.426	4.162	<2e-16
Age.Group25-34 years	1.936	6.933	<2e-16
Age.Group30-39 years	2.371	10.708	<2e-16
Age.Group35-44 years	2.761	15.819	<2e-16
Age.Group40-49 years	3.163	23.646	<2e-16
Age.Group45-54 years	3.577	35.778	<2e-16
Age.Group5-14 years	-2.404	0.090	<2e-16
Age.Group50-64 years	4.602	99.762	<2e-16
Age.Group55-64 years	4.363	78.490	<2e-16
Age.Group65-74 years	4.833	125.479	<2e-16
Age.Group75-84 years	4.990	146.940	<2e-16
Age.Group85 years and over	5.028	152.581	<2e-16
Age.GroupUnder 1 year	-2.733	0.065	<2e-16

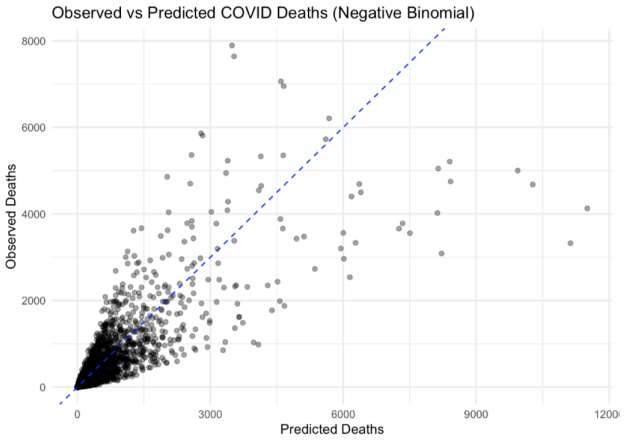
Negative Binomial Regression Results

Coefficient	Estimate	IRR	P-value
(Intercept)	1165.30209	5.08e+506	<2e-16
SexMale	0.31441	1.370	<2e-16
regionnortheast	-0.27157	0.762	6.82e-13
regionsouth	0.30610	1.358	<2e-16
regionwest	0.03863	1.039	0.252
Year	-0.58368	0.558	<2e-16
Age.Group1-4 years	-3.76598	0.023	<2e-16
Age.Group15-24 years	0.63030	1.878	2.32e-09
Age.Group18-29 years	1.50495	4.504	<2e-16
Age.Group25-34 years	2.05210	7.786	<2e-16
Age.Group30-39 years	2.54915	12.793	<2e-16
Age.Group35-44 years	2.90840	18.326	<2e-16
Age.Group40-49 years	3.26220	26.103	<2e-16
Age.Group45-54 years	3.64565	38.302	<2e-16
Age.Group5-14 years	-2.64737	0.071	<2e-16
Age.Group50-64 years	4.65791	105.309	<2e-16
Age.Group55-64 years	4.43404	84.254	<2e-16
Age.Group65-74 years	4.93283	138.593	<2e-16
Age.Group75-84 years	5.13196	168.952	<2e-16
Age.Group85 years and over	5.24508	189.502	<2e-16
Age.GroupUnder 1 year	-2.84356	0.058	<2e-16

In our Poission model, males were predicted to have had about a 28% higher mortality rate than females, while regional effects showed that the Northeast and West had 14% and 16% lower mortality compared to the Midwest baseline. The South had about 10% higher mortality than the Midwest. Mortality declined over time, with each additional year predicting a 35% decrease in COVID-19 death rates, potentially being a result of the adinistrtion of the vaccine. Age was the strongest predictor, with risk increasing steadily across age groups and the 85+ population experiencing more than 150 times the mortality rate of children, which may be explained by older people having weaker immune systems.

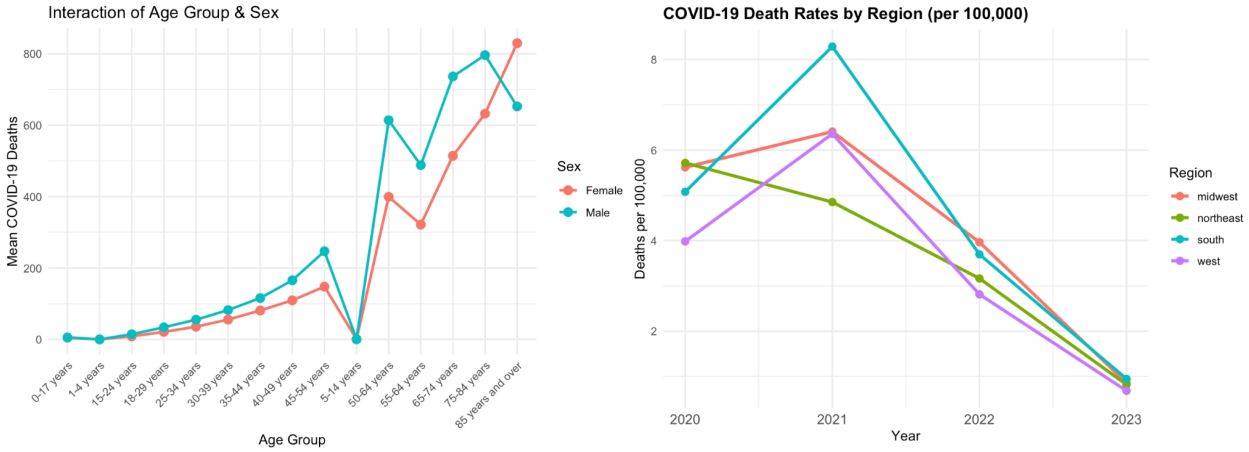
In the Negative Binomial Regression model, males were predicted to have had a 37% higher mortality rates than females. Southern states showed a 36% higher mortality rate, while the Northeast experienced about 24% lower mortality compared to the Midwest. Mortality declined substantially over time, with each additional year predicting a 44% decrease in COVID-19 death rates. This trend also might be explained by the release of the COVID-19 vaccine. Age maintained the most powerful predictor, with the oldest adults facing mortality rates more than 190 times higher than children, similar to the Poisson model.

The negative binomial regression model had a much lower AIC value of 44,747.3 compared to that of the Poisson Regression model (480,266.0). This demonstrates that the Negative Binomial approach exhibits a lower estimated prediction error and therefore better fits the mortality data. Despite this, the Negative Binomial model both under and overpredicted covid deaths. The fan-shaped pattern indicates that the model's prediction accuracy lessened as the number of deaths increased. Regardless, the model does a decent job of predicting the number of COVID deaths when the observed deaths are below about 1,500.



Finally, we plotted the interaction between Age Group and Sex across ages, as well as the mortality rates by region in a panel plot style. Our interaction demonstrated that after age 44, sex had a stronger effect on predicted COVID-19 deaths, given the average COVID-19 deaths for males is much higher than for females.

After age 84, the effect of sex on reverses, indicating an interaction between the two predictive variables. One possible explanation for this change is that the aversge lifespan for females in the data may have been longer. Secondly, our panel plot confirmed that mortality rates were highest in all regions, with Southern states being the highest and Northeastern states being the lowest. This reflects possible differences in health policy and vaccine administration.



Discussion and Conclusion

This study examined how COVID-19 mortality varied across demographic groups, geographic regions, and time in the United States from 2020 to 2023. By combining the CDC's provisional death counts with state level population estimates, we were able to evaluate death rates per 100,000 people and use regression modeling in addition to visualizations to identify the factors strongly associated with COVID-19 mortality. Ultimately, these findings consistantly showed that age group, sex, region, and year played significant roles in shaping mortality patterns.

Mortality peaked nationally in late 2020 and early 2021, with a sharp decline in mid-2021. Age was the strongest predictor of mortality, as death rates grew as people got older. Additionally, males consistantly showed to have higher mortality in comparison to females. Regionally, the South showed elevated death rates, while regions such as the Northeast had very high early mortality followed by lower rates in later years. These trends are evident in the visualizations and further backed by our statistical models.

Our Poisson model was able to capture the count based characteristics of the data, but was also able to show its strong overdispersion. The Negative Binomial model was the best fit for the data, showing that mortality decreased significantly over time, males had higher death rates compared to females, and regional differences were statisically significant. The Two-Way ANOVA further revealed that the effect of age group on mortality changed based on males and females, emphasizing the significance of demographic interactions.

All of these findings have meaningful implications for public health, for example, the strong age related disparities emphasize the importance of ensuring that older adults are protected through the aid of vaccines, boosters, and early treatment access. Being able to understand these differences can help create interventions for potential future public health emergencies.

However, our project study does have its limitations. Firstly, our dataset being count data made regression hard more difficult and less insightful. Secondly, observations in our data are not statistically independent, so standard models such as linear regression may have uinderestimated uncertainty. In addition, the CDC intentionally suppresses values for smaller populations to maintain confidentiality, which left us with many missing values that we could not work with. Finally, identifying true statistical significance was almost impossible because our dataset has so many observations, giving each coefficient a small p-value by nature.

In the future, we would like to use additional predictors like vaccination coverage, variant spread, or county-level data to better understand patterns. We would also like to use more advanced modeling approaches, for example, time series models like a Generalized Additive Model to better capture seasonal patters.

In conclusion, our analysis revealed a clear and statistical understanding of how COVID-19 mortality varied across demographic and geographic dimensions in the United States, providing insights that may help implement preventative or early intervention health strategies in the future.

References

1. Arel-Bundock, V. (2022). modelsummary: Data and model summaries in R. *Journal of Statistical Software*, *103*(1), 1–23. https://doi.org/10.18637/jss.v103.i01

2. Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., & Deckmyn, A. (2023). maps: Draw geographical maps. R package version 3.4.2. https://CRAN.R-project.org/package=maps

3. Centers for Disease Control and Prevention. (2020). Provisional COVID-19 deaths by sex and age. National Center for Health Statistics. https://data.cdc.gov/

4. Centers for Disease Control and Prevention. (2025, June 11). Benefits of getting vaccinated. https://www.cdc.gov/covid/vaccines/benefits.html

5. Gohel, D., & Skintzos, P. (2024). *flextable: Functions for tabular reporting* (R package version 0.9.4). https://CRAN.R-project.org/package=flextable

6. Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3), 1–25. https://doi.org/10.18637/jss.v040.i03

7. Kassambara, A. (2023). *ggpubr: 'ggplot2' based publication ready plots* (R package version 0.6.0). https://CRAN.R-project.org/package=ggpubr

8. Pedersen, T. L. (2024). *patchwork: The composer of plots* (R package version 1.2.0). https://CRAN.R-project.org/package=patchwork

9. Robinson, D., Hayes, A., & Couch, S. (2024). broom: Convert statistical analysis objects into tidy tibbles. R package version 1.0.5. https://CRAN.R-project.org/package=broom

10. United States Census Bureau. (2023). State population totals: 2020–2023. https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html

11. Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

12. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis. Springer.*

13. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

14. Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data