Data 607 – Data Communication and Analytics

Final Project Report

Student Name: Pravalika Sure (UID: 120558016)

State-Level Drivers of COVID-19 Mortality in the United States (2021)

Abstract

This project examines why COVID-19 mortality varied across U.S. states in 2021 using state-level data from the *COVID19* R package. I focus on three central factors—vaccination coverage, testing intensity, and government policy stringency—and assess how they relate to deaths per 100,000 residents. After cleaning the daily dataset by handling missing and infinite values, aggregating to annual state measures, and removing outliers, I conduct exploratory visualizations, multivariate regression, logistic regression with ROC analysis, k-means clustering, spatial mapping, and a random-forest robustness check.

Across all methods, higher vaccination coverage consistently corresponds to lower mortality, while testing and stringency show weaker but meaningful associations. The classifier moderately distinguishes high- from low-mortality states, and clustering reveals distinct pandemic profiles across regions. Overall, the results demonstrate that differences in both pharmaceutical and non-pharmaceutical interventions shaped COVID-19 mortality patterns in the United States.

1. Introduction

COVID-19 affected U.S. states unevenly, with substantial differences in deaths, testing practices, vaccination uptake, and policy responses. Identifying which factors best explain this variation can inform more effective public-health strategies.

This project analyzes state-level COVID-19 data for 2021 using the `COVID19` R package, which compiles daily information from multiple official sources.

Central Research Question:

> *How are state-level COVID-19 death rates in 2021 associated with vaccination coverage, testing intensity, and government policy stringency?*

To answer this, I:

- Clean and aggregate the daily dataset into a 2021 state-level analytic file.
- Conduct exploratory plots and correlation analysis.
- Fit multivariate regression models with diagnostics.
- Train a logistic classifier and evaluate its ROC curve.
- Apply k-means clustering to identify state-level pandemic profiles.
- Map state mortality using a U.S. choropleth.
- Fit a random-forest model as a non-linear check.
- Provide national temporal context with a COVID-19 mortality time series.

All code and diagnostics appear in the Appendix; the main report summarizes the key findings and interpretations.

2. Data, Cleaning and Feature Engineering

2.1 Data Source and Subset

I first downloaded global COVID-19 data at administrative level 2 (states/provinces) using:

```
raw_df <- covid19(level = 2)
```

The raw dataset contains 862,520 rows and 47 columns. I then:

- Restricted to the United States using `administrative_area_level_1 == "United States"` and renamed `administrative_area_level_2` to `state`.
- Restricted the time window to 1 January–31 December 2021, creating `usa_2021`.
- Standardized variable names using `janitor::clean_names()`.

This yields a coherent subset of daily U.S. state-level observations for 2021.

2.2 Feature Engineering

For each state, daily observations were aggregated into a single 2021 summary row:

- **population**: maximum population value in 2021.
- **total_deaths**: maximum cumulative deaths.
- **deaths_per_100k** = $(\text{total\_deaths}/\text{population}) \times 100{,}000$.
- **total_tests**: maximum cumulative tests.
- **tests_per_1k** = $(\text{total\_tests}/\text{population}) \times 1{,}000$.
- **vax_coverage** = $\mathtt{max}(\text{people\_fully\_vaccinated})/\text{population}$.
- **mean_stringency**: mean of the daily `stringency_index` across 2021.

This converts noisy daily time series into stable state-level summaries of mortality, vaccination, testing, and policy strictness.

2.3 Handling Missing and Infinite Values

Some U.S. territories (e.g., American Samoa) reported no testing or vaccination information, causing `max()` to return $-\infty$. I addressed this by:

- Replacing $\pm\infty$ with `NA` for `total_tests`, `tests_per_1k`, `vax_coverage`, `deaths_per_100k`, and `mean_stringency`.
- Dropping rows with missing values in any key variable (mortality, vaccination, testing, stringency, or population).

After cleaning, the dataset (`state_clean`) contains 51 observations and 8 variables, representing the 50 states plus the District of Columbia.

2.4 Outlier Removal

To limit the influence of extreme mortality outliers, I applied the IQR rule to `deaths_per_100k`. The **IQR rule** identifies outliers as values that lie more than $1.5 \times$ IQR below the first quartile (Q_1) or above the third quartile (Q_3), where IQR $= Q_3 - Q_1$. Let

$$Q_1 = \text{25th percentile}, \quad Q_3 = \text{75th percentile}, \quad \text{IQR} = Q_3 - Q_1.$$

The upper cutoff is

$$\text{Upper cutoff} = Q_3 + 1.5 \times \text{IQR}.$$

states with mortality values above this cutoff were removed. In practice, no additional states were dropped, but this step documents an explicit outlier-screening procedure.

These three operations—temporal aggregation, missing-data handling, and outlier screening—constitute the main cleaning steps applied to the raw dataset.

3. Exploratory Analysis

This section uses descriptive statistics and visualizations to reveal initial patterns in state-level mortality, vaccination coverage, testing intensity, and policy stringency.

3.1 Distribution of Death Rates

A histogram of `deaths_per_100k` **(Figure 1)** shows that most states had between approximately 200 and 325 deaths per 100,000 people in 2021, with a few states experiencing much higher or lower mortality. The distribution is slightly right-skewed, indicating a tail of states with unusually high death rates.

This substantial cross-state variation motivates multivariate modeling to understand what factors drive these differences.

3.2 Bivariate Relationships

Scatterplots with linear smooths reveal the following patterns:

Vaccination vs. Mortality (Figure 2)

There is a clear negative relationship: as vaccination coverage increases, deaths per 100k decrease. High-vaccination states (coverage $\gtrsim 0.70$) tend to cluster below 200 deaths per 100k, while low-vaccination states (coverage $\lesssim 0.55$) often exceed 250–300 deaths per 100k.

Testing vs. Mortality (Figure 3)

Testing intensity (`tests_per_1k`) shows a weak negative association with mortality. Some highly tested states still experienced substantial deaths, and the points are widely scattered around the regression line, indicating that testing alone does not strongly predict mortality.

Stringency vs. Mortality (Figure 4)

The average stringency index exhibits a moderate negative association: states with stricter policies tend to have somewhat lower death rates, although the confidence band is wide, indicating meaningful uncertainty.

3.3 Correlation Structure

I computed a correlation matrix for `deaths_per_100k`, `vax_coverage`, `tests_per_1k`, and `mean_stringency`, and visualized it using a heatmap **(Figure 5)**.

- Vaccination vs. mortality: $r = -0.52$ (strong negative).
- Vaccination vs. testing: $r = 0.66$ (moderate positive).
- Testing vs. stringency: $r = 0.37$ (moderate positive).
- Testing vs. mortality: $r = -0.24$ (weak negative).
- Stringency vs. mortality: $r = -0.34$ (moderate negative).

These correlations highlight vaccination as the strongest protective factor, with testing and stringency showing weaker but meaningful associations. Importantly, correlations among predictors remain below 0.80, indicating no severe multicollinearity and supporting the use of a multivariate regression model.

4. Multivariate Regression

This section fits and interprets multivariate linear regression models to quantify the independent associations between vaccination, testing, policy stringency, and mortality.

4.1 Baseline Linear Model

I first fit an ordinary least squares regression of state-level mortality on vaccination, testing, and policy stringency:

$$\text{deaths\_per\_100k} = \beta_0 + \beta_1\,\text{vax\_coverage} + \beta_2\,\text{tests\_per\_1k} + \beta_3\,\text{mean\_stringency} + \varepsilon. \tag{1}$$

Key coefficient estimates are:

- Intercept: $\hat{\beta}_0 = 567.9$ (SE $= 79.6$, $p < 0.0001$; 95% CI: [407.7, 728.0]).
- Vaccination coverage: $\hat{\beta}_1 = -454.1$ (SE $= 136.7$, $p = 0.0017$; 95% CI: $[-729.0, -179.1]$).
- Tests per 1k: $\hat{\beta}_2 = 0.011$ (SE $= 0.0097$, $p = 0.262$; 95% CI: $[-0.0085, 0.0305]$).
- Mean stringency: $\hat{\beta}_3 = -1.70$ (SE $= 1.69$, $p = 0.321$; 95% CI: $[-5.10, 1.70]$).

Model fit statistics:

- Residual standard error: 60.7 deaths per 100k on 47 df.
- $R^2 = 0.298$, adjusted $R^2 = 0.253$.
- Overall F-test: $F(3, 47) = 6.65$, $p = 0.00078$.

Interpretation: Holding testing and stringency constant, a 0.10 (10 percentage-point) increase in vaccination coverage is associated with an estimated decrease of about 45 deaths per 100,000 (0.10×454.1). The 95% CI for the vaccination effect is entirely negative, indicating a statistically strong protective association. In contrast, the estimated effects of testing and stringency are small and not statistically significant after adjusting for vaccination.

4.2 Interaction Model

To test whether the effect of vaccination depends on policy strictness, I fit a model with an interaction term:

$$\begin{aligned}\text{deaths\_per\_100k} = \beta_0 + \beta_1\,\text{vax\_coverage} + \beta_2\,\text{mean\_stringency} \\ + \beta_3\,\text{tests\_per\_1k} + \beta_4\,(\text{vax\_coverage} \times \text{mean\_stringency}) + \varepsilon.\end{aligned} \tag{2}$$

The interaction coefficient is:

- vax_coverage × mean_stringency: $\hat{\beta}_4 = -2.44$ (SE $= 17.79$, $p = 0.892$).

Overall model fit remains similar:

- Residual SE $= 61.35$ deaths per 100k.
- $R^2 = 0.298$, adjusted $R^2 = 0.237$.
- $F(4, 46) = 4.89$, $p = 0.0023$.

However, none of the individual coefficients (including vaccination) are statistically significant in this expanded model, suggesting that the interaction is not supported and that the main effect of vaccination is already captured by the simpler baseline model.

4.3 Multicollinearity Diagnostics

This subsection evaluates **Variance Inflation Factors (VIFs)** to confirm that the predictors do not suffer from problematic multicollinearity.The Variance Inflation Factor (VIF) for a predictor X_j is defined as

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination from regressing X_j on all other predictors. A VIF of 1 indicates no multicollinearity, while larger values reflect increasing redundancy among predictors. In applied regression, thresholds of 5–10 are commonly used to flag problematic multicollinearity because such values imply that the variance of a coefficient is inflated by a factor of 5–10, leading to unstable estimates and larger standard errors.

Variance Inflation Factors (VIFs) for the baseline model are:

- VIF(vax_coverage) $= 1.98$
- VIF(tests_per_1k) $= 1.78$
- VIF(mean_stringency) $= 1.31$

All values are well below the thresholds of common concern (5–10), indicating that multicollinearity is minimal and that the regression coefficients are reasonably stable.

5. Classification and ROC Analysis

This section reframes the analysis as a classification problem and evaluates how effectively vaccination, testing, and policy stringency distinguish high- from low-mortality states.

5.1 Defining High-Mortality States

To reframe the problem as a classification task, states were dichotomized based on the median death rate:

$$\text{high\_mortality} = \begin{cases} 1, & \text{if deaths\_per\_100k} > \text{median}, \\ 0, & \text{otherwise.} \end{cases}$$

This produced a nearly balanced dataset: 26 low-mortality states and 25 high-mortality states.

5.2 Logistic Regression Model

I fit a logistic regression model of the form:

$$\Pr(\text{high\_mortality} = 1) = \text{logit}^{-1}\Big(\alpha_0 + \alpha_1\, \text{vax\_coverage} + \alpha_2\, \text{tests\_per\_1k} + \alpha_3\, \text{mean\_stringency}\Big). \tag{3}$$

Key coefficient estimates:

- Intercept: $\hat{\alpha}_0 = 7.58$ (SE $= 3.26$, $p = 0.020$).
- Vaccination coverage: $\hat{\alpha}_1 = -13.66$ (SE $= 5.46$, $p = 0.012$).
- Tests per 1k: $\hat{\alpha}_2 = 4.06 \times 10^{-4}$ (SE $= 3.79 \times 10^{-4}$, $p = 0.284$).
- Mean stringency: $\hat{\alpha}_3 = -0.00157$ (SE $= 0.0666$, $p = 0.981$).

The negative and statistically significant vaccination coefficient shows that higher vaccination coverage is associated with substantially lower odds of belonging to the high-mortality group, even after adjusting for testing and policy stringency.

Model fit statistics:

- Null deviance: 70.68 on 50 df.
- Residual deviance: 61.19 on 47 df.
- AIC: 69.19.

The Akaike Information Criterion (AIC) measures how well a model fits the data while penalizing unnecessary complexity; lower AIC values indicate a better trade-off between fit and simplicity.

5.3 ROC Curve

Using the predicted probabilities from the logistic model, I computed a ROC curve for classifying high- vs. low-mortality states. The Area Under the Curve (AUC) is 0.737 **(Figure 6)**.

This means that, for a randomly chosen high-mortality and low-mortality state, the model assigns a higher predicted probability to the high-mortality state about 74% of the time—well above the 50% expected under random guessing. Thus, vaccination, testing, and stringency together provide moderate discriminative power, although additional unmeasured factors also play a role.

6. Clustering and Spatial Patterns

This section uses unsupervised learning and geographic mapping to examine multivariate patterns and spatial differences in COVID-19 outcomes.

6.1 K-Means Clustering of State Profiles

I standardized the four key variables (deaths per 100k, vaccination coverage, tests per 1k, and mean stringency) and applied k-means clustering with $k = 3$. The resulting groups were:

- **Cluster A**: Low vaccination, high mortality
- **Cluster B**: High vaccination, moderate mortality
- **Cluster C**: High vaccination, low mortality

The scatterplot of vaccination coverage vs. deaths per 100k **(Figure 7)** shows clear separation: Cluster A falls in the low-vaccination/high-mortality region, Cluster C in the high-vaccination/low-mortality region, and Cluster B occupies a middle position. These clusters reinforce the central conclusion that under-vaccinated states experienced worse mortality outcomes.

6.2 Spatial Choropleth Map

Using the `usmap` package, I mapped `deaths_per_100k` across states for 2021 **(Figure 8)**. The choropleth reveals distinct geographic patterns:

- **Southern states** show notably higher mortality, consistent with lower vaccination uptake.
- **Mountain West** states also exhibit elevated mortality.
- **Northeastern states** generally show lower mortality and higher vaccination.

- **West Coast states** display moderate mortality aligned with strong vaccination campaigns.

These spatial trends align with the statistical models and highlight the geographic concentration of high-mortality, low-vaccination states.

7. Random Forest Robustness Check

To explore potential nonlinear relationships and interactions beyond linear regression, I fitted a random forest model predicting `deaths_per_100k` from vaccination coverage, tests per 1k, and mean stringency using 500 trees and `mtry = 2`.

Model output:

- Mean of squared residuals: 5000.4
- Percent variance explained: -3.36% (worse than predicting the mean, likely due to the small sample size)

Variable-importance estimates show the following ranking:

- Vaccination coverage has the highest importance (largest %IncMSE and IncNodePurity).
- Mean stringency ranks second.
- Tests per 1k contributes the least.

Although overall predictive performance is poor—likely due to the limited sample of 51 states and small feature set—the importance rankings mirror the regression results: vaccination is consistently the dominant predictor of cross-state COVID-19 mortality.

8. Temporal Context: National Death Trends

To contextualize state-level patterns, I plotted the 7-day rolling cumulative COVID-19 deaths across the United States from 2020–2022 (Figure 9). The series highlights the major national phases of the pandemic: the 2020 surge, the sharp rise in late 2020–early 2021, and later waves such as Delta and Omicron. Dates after 2022-12-31 were excluded to avoid an artificial drop caused by missing state-level reporting in 2023.

This broader trajectory reinforces why 2021 is a critical year to study: vaccines became widely available, yet mortality diverged dramatically across states.

9. Discussion and Conclusion

This project analyzed U.S. state-level COVID-19 mortality in 2021 using cleaned data from the COVID-19 Data Hub. After filtering U.S. records, aggregating daily values, handling missing and infinite entries, and removing outliers, I examined how vaccination coverage, testing intensity, and policy stringency related to mortality.

Across all methods—correlation analysis, multivariate regression, logistic classification, clustering, spatial mapping, and random forests—one result consistently dominated:

Higher vaccination coverage strongly predicts lower COVID-19 mortality.

The linear model shows that a 10-point increase in vaccination corresponds to roughly 45 fewer deaths per 100k people. The model explains about 25

Testing intensity and policy stringency show weaker and inconsistent associations. While important in practice, they explain far less variation than vaccination at the state level. The random forest model provides a nonlinear check and again ranks vaccination highest.

Limitations

- State-level aggregates obscure within-state demographic and socioeconomic variation.
- Observational data may contain reporting inconsistencies and delays.
- Only three predictors are modeled; other structural factors (age, comorbidities, healthcare capacity, political attitudes) likely contribute.
- The analysis is cross-sectional and does not model vaccine rollout or policy timing.

Overall Conclusion

Despite limitations, the analyses converge on the same conclusion: states with higher vaccination coverage experienced substantially lower COVID-19 mortality in 2021. Testing and policy stringency play secondary roles, but vaccination emerges as the strongest and most robust predictor across all analytic methods. This project demonstrates how careful data cleaning, multivariate modeling, and clustering can reveal clear public-health insights even in large, messy datasets.

References

1. Guidotti, E., & Ardia, D. (2020). *COVID-19 Data Hub*. Journal of Open Source Software, 5(51), 2376. `https://doi.org/10.21105/joss.02376`

2. COVID-19 Data Hub. (2024). *Documentation for Dataset and Variables*. Retrieved from `https://covid19datahub.io/articles/docs.html`

3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.

Appendix

Appendix A: Figures

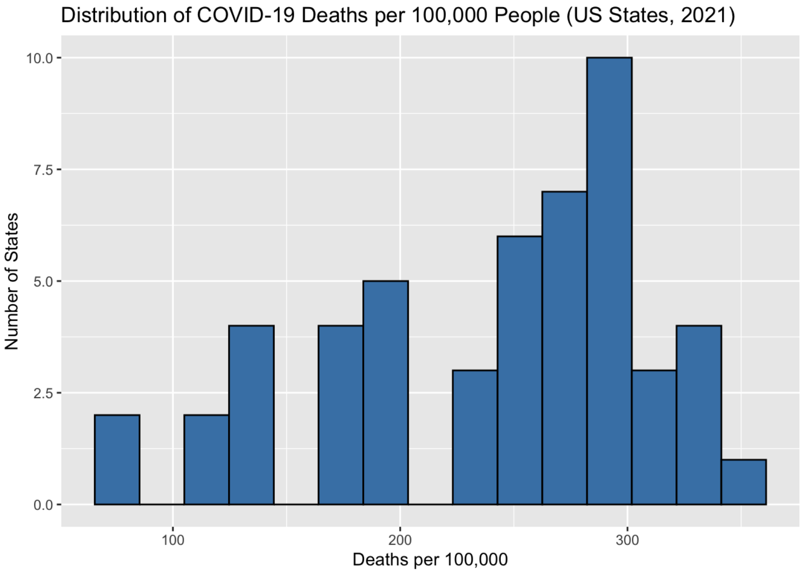


Figure 1: Histogram of deaths per 100,000 (state-level, 2021).

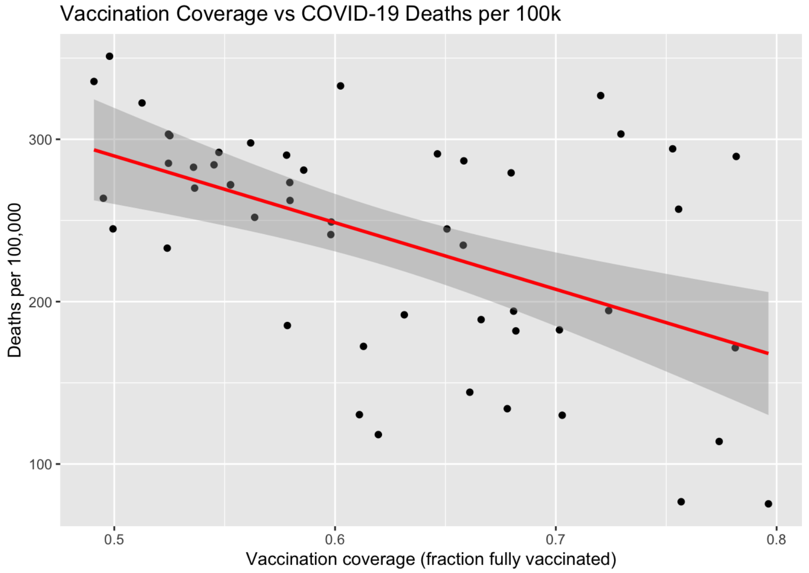


Figure 2: Vaccination coverage vs. deaths per 100,000 (with regression line).

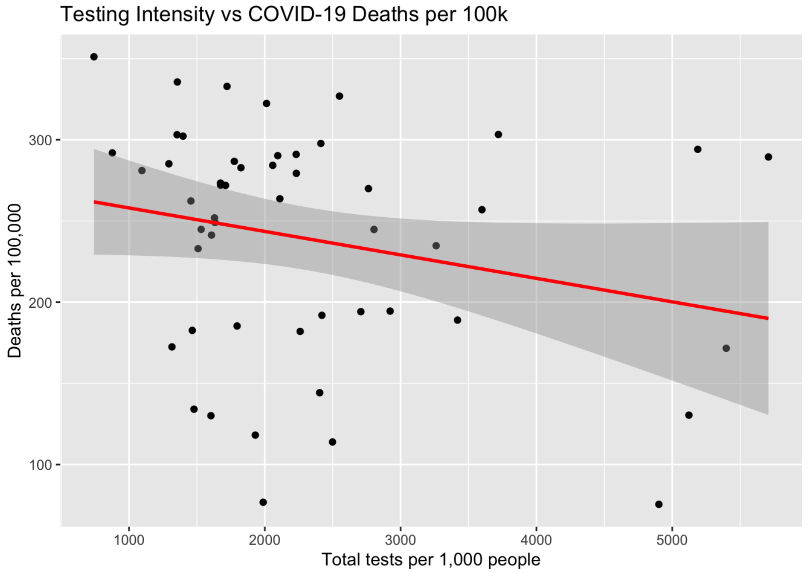


Figure 3: Tests per 1,000 vs. deaths per 100,000 (with regression line).

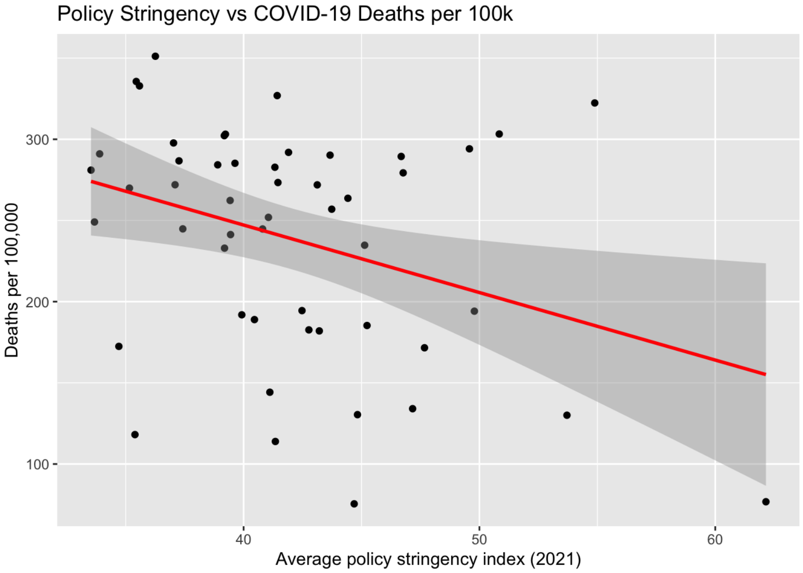


Figure 4: Mean stringency index vs. deaths per 100,000.

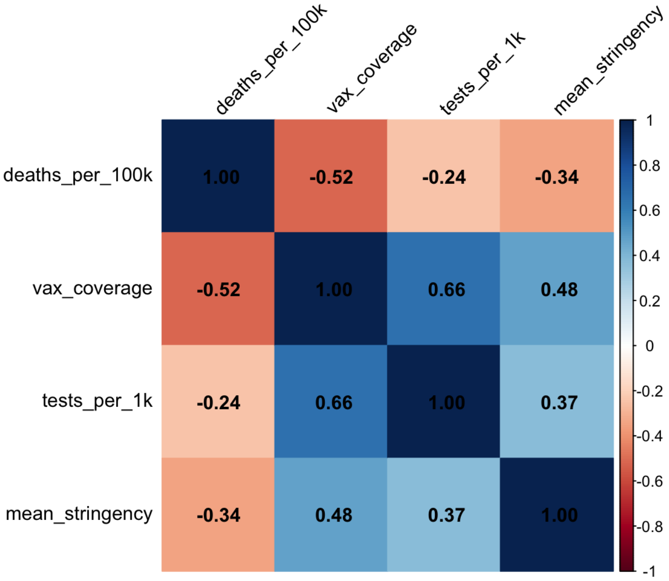


Figure 5: Correlation heatmap of key variables.

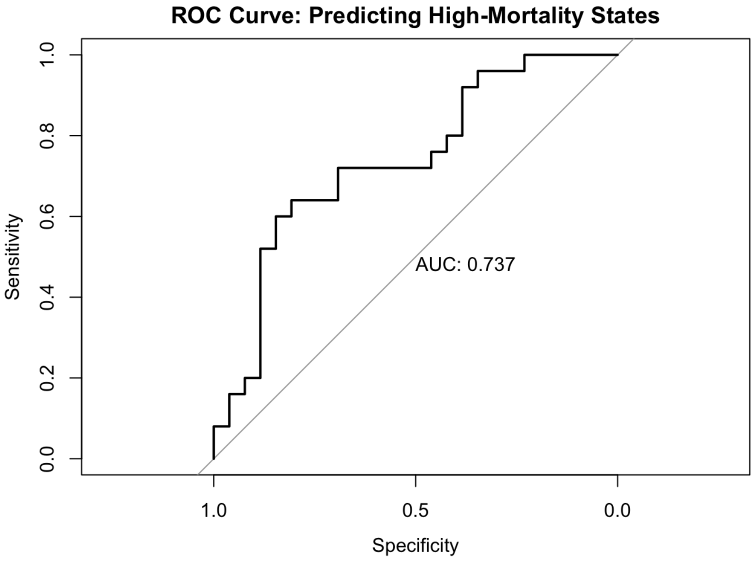


Figure 6: ROC curve for logistic regression classifier (AUC = 0.737).

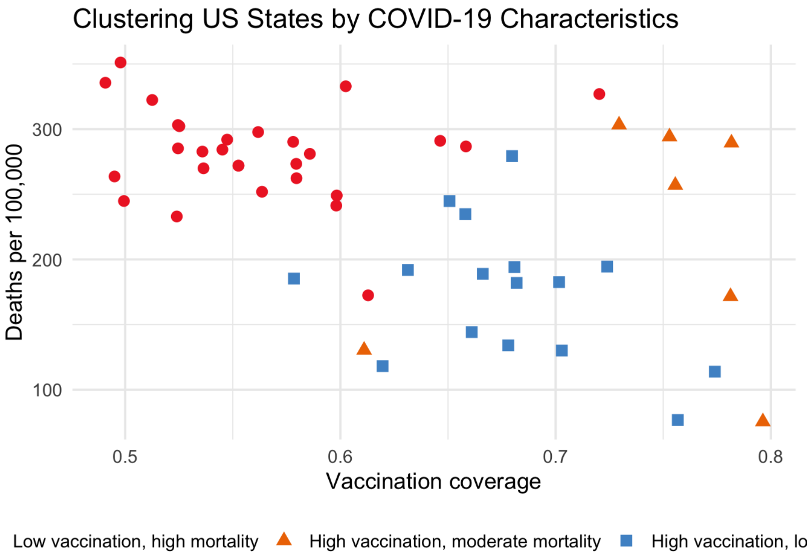


Figure 7: K-means clustering of states (3 clusters).

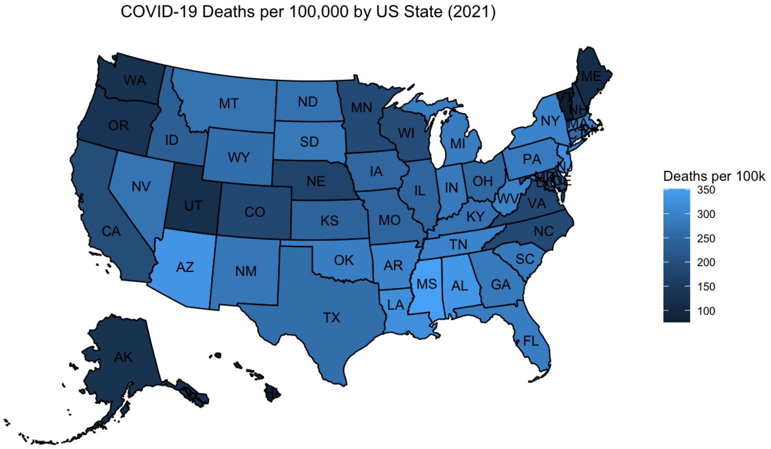


Figure 8: U.S. choropleth map of deaths per 100,000 (2021).

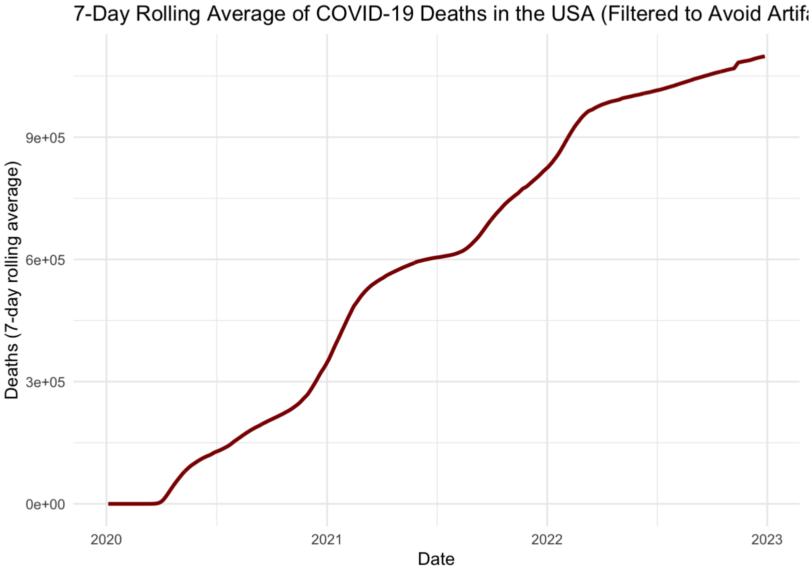


Figure 9: National cumulative deaths (7-day rolling average), 2020–2022.

Data 607 – Final Project

Student_Name: Pravalika Sure

UID: 120558016

State-Level Drivers of COVID-19 Mortality in the United States (2021)

1.1 Loading required packages

```
# Load required packages

library(COVID19)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(usmap)
library(randomForest)
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(car)       # for VIF
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

2.Data Loading and Initial Cleaning

Exploring entire dataset

```
# Load global COVID-19 data at level 2 (states/provinces)
raw_df <- covid19(level = 2)
```

```
## We have invested a lot of time and effort in creating COVID-19 Data
## Hub, please cite the following when using it:
##
##   Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open
##   Source Software 5(51):2376, doi: 10.21105/joss.02376
##
## The implementation details and the latest version of the data are
## described in:
##
##   Guidotti, E., (2022), "A worldwide epidemiological database for
##   COVID-19 at fine-grained spatial resolution", Sci Data 9(1):112, doi:
##   10.1038/s41597-022-01245-1
## To print citations in BibTeX format use:
##  > print(citation('COVID19'), bibtex=TRUE)
##
## To hide this message use 'verbose = FALSE'.
```

```
# Basic structure
glimpse(raw_df)
```

```
## Rows: 862,520
## Columns: 47
## $ id                                  <chr> "0042529a", "0042529a", "0042529a"…
## $ date                                <date> 2020-01-12, 2020-01-13, 2020-01-1…
## $ confirmed                           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ deaths                              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ recovered                           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ tests                               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ vaccines                            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ people_vaccinated                   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ people_fully_vaccinated             <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ hosp                                <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ icu                                 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ vent                                <int> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ school_closing                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ workplace_closing                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ cancel_events                       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ gatherings_restrictions             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ transport_closing                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ stay_home_restrictions              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ internal_movement_restrictions      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ international_movement_restrictions <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ information_campaigns               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ testing_policy                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ contact_tracing                     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ facial_coverings                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ vaccination_policy                  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ elderly_people_protection           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ government_response_index           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ stringency_index                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ containment_health_index            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ economic_support_index              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ administrative_area_level           <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2…
## $ administrative_area_level_1         <chr> "Thailand", "Thailand", "Thailand"…
## $ administrative_area_level_2         <chr> "Yasothon", "Yasothon", "Yasothon"…
## $ administrative_area_level_3         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ latitude                            <dbl> 15.81802, 15.81802, 15.81802, 15.8…
## $ longitude                           <dbl> 104.2755, 104.2755, 104.2755, 104.…
## $ population                          <int> 487976, 487976, 487976, 487976, 48…
## $ iso_alpha_3                         <chr> "THA", "THA", "THA", "THA", "THA",…
## $ iso_alpha_2                         <chr> "TH", "TH", "TH", "TH", "TH", "TH"…
## $ iso_numeric                         <int> 764, 764, 764, 764, 764, 764, 764,…
## $ iso_currency                        <chr> "THB", "THB", "THB", "THB", "THB",…
## $ key_local                           <chr> "35", "35", "35", "35", "35", "35"…
## $ key_google_mobility                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ key_apple_mobility                  <chr> "Yasothon Province", "Yasothon Pro…
## $ key_jhu_csse                        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ key_nuts                            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ key_gadm                            <chr> "THA.77_1", "THA.77_1", "THA.77_1"…
```

Filtering out USA Data

```
df <- covid19(level = 2) %>% clean_names()
```

```
## We have invested a lot of time and effort in creating COVID-19 Data
## Hub, please cite the following when using it:
##
##   Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open
##   Source Software 5(51):2376, doi: 10.21105/joss.02376
##
## The implementation details and the latest version of the data are
## described in:
##
##   Guidotti, E., (2022), "A worldwide epidemiological database for
##   COVID-19 at fine-grained spatial resolution", Sci Data 9(1):112, doi:
##   10.1038/s41597-022-01245-1
## To print citations in BibTeX format use:
##  > print(citation('COVID19'), bibtex=TRUE)
##
## To hide this message use 'verbose = FALSE'.
```

```
usa <- df %>%
  filter(administrative_area_level_1 == "United States") %>%
  rename(state = administrative_area_level_2)
```

Checking for the filtered data

```
unique(usa$state)
```

```
##  [1] "Northern Mariana Islands" "Minnesota"
##  [3] "California"               "Florida"
##  [5] "Wyoming"                  "Virgin Islands"
##  [7] "South Dakota"             "Kansas"
##  [9] "Nevada"                   "Virginia"
## [11] "Washington"               "Oregon"
## [13] "Wisconsin"                "New Jersey"
## [15] "Rhode Island"             "Vermont"
## [17] "North Carolina"           "Oklahoma"
## [19] "Alabama"                  "Delaware"
## [21] "Guam"                     "Missouri"
## [23] "American Samoa"           "Utah"
## [25] "Mississippi"              "Connecticut"
## [27] "Indiana"                  "Georgia"
## [29] "Texas"                    "Pennsylvania"
## [31] "Massachusetts"            "Maine"
## [33] "Tennessee"                "Michigan"
## [35] "Idaho"                    "Illinois"
## [37] "Louisiana"                "New Mexico"
## [39] "Arizona"                  "Arkansas"
## [41] "Nebraska"                 "West Virginia"
## [43] "South Carolina"           "New York"
## [45] "District of Columbia"     "Kentucky"
## [47] "Ohio"                     "Alaska"
## [49] "New Hampshire"            "North Dakota"
## [51] "Iowa"                     "Montana"
## [53] "Hawaii"                   "Maryland"
## [55] "Puerto Rico"              "Colorado"
```

```
usa_2021 <- usa %>%
  filter(date >= "2021-01-01", date <= "2021-12-31")
```

3. Feature Engineering and Handling Missing Data

3.1.Handling missing values

```
# Check which vaccine variable exists
grep("vacc", names(usa_2021), value = TRUE)
```

```
## [1] "vaccines"                "people_vaccinated"
## [3] "people_fully_vaccinated" "vaccination_policy"
```

```
state_summ <- usa_2021 %>%
  group_by(state) %>%
  summarise(
    population        = max(population, na.rm = TRUE),
    total_deaths      = max(deaths, na.rm = TRUE),
    deaths_per_100k   = (total_deaths / population) * 100000,
    total_tests       = max(tests, na.rm = TRUE),
    tests_per_1k      = (total_tests / population) * 1000,
    vax_coverage      = max(people_fully_vaccinated, na.rm = TRUE) / population,
    mean_stringency   = mean(stringency_index, na.rm = TRUE)
  )
```

```
## Warning: There was 1 warning in `summarise()`.
## ℹ In argument: `total_tests = max(tests, na.rm = TRUE)`.
## ℹ In group 3: `state = "American Samoa"`.
## Caused by warning in `max()`:
## ! no non-missing arguments to max; returning -Inf
```

3.2. CLEANING: Replace -Inf/Inf with NA + Filter NAs

```
state_clean <- state_summ %>%
  mutate(
    total_tests       = ifelse(is.infinite(total_tests), NA, total_tests),
    tests_per_1k      = ifelse(is.infinite(tests_per_1k), NA, tests_per_1k),
    vax_coverage      = ifelse(is.infinite(vax_coverage), NA, vax_coverage),
    deaths_per_100k   = ifelse(is.infinite(deaths_per_100k), NA, deaths_per_100k),
    mean_stringency   = ifelse(is.infinite(mean_stringency), NA, mean_stringency)
  ) %>%
  filter(
    !is.na(deaths_per_100k),
    !is.na(vax_coverage),
    !is.na(tests_per_1k),
    !is.na(mean_stringency)
  )

glimpse(state_clean)
```

```
## Rows: 51
## Columns: 8
## $ state           <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "Californi…
## $ population      <int> 4903185, 731545, 7278717, 3017804, 39512223, 5758736, …
## $ total_deaths    <int> 16455, 954, 24229, 9148, 76709, 10480, 9161, 2286, 121…
## $ deaths_per_100k <dbl> 335.59819, 130.40893, 332.87460, 303.13433, 194.13992,…
## $ total_tests     <dbl> 6647224, 3747402, 12531573, 4084594, 106951711, 130160…
## $ tests_per_1k    <dbl> 1355.6951, 5122.5858, 1721.6733, 1353.4988, 2706.8007,…
## $ vax_coverage    <dbl> 0.4908059, 0.6110232, 0.6024869, 0.5245245, 0.6808680,…
## $ mean_stringency <dbl> 35.44597, 44.82907, 35.58422, 39.22830, 49.78534, 43.2…
```

3.3. Handle infinite and missing values and remove rows with critical missingness

```
state_clean <- state_summ %>%
  mutate(
    vax_coverage = ifelse(is.infinite(vax_coverage), NA, vax_coverage),
    tests_per_1k = ifelse(is.infinite(tests_per_1k), NA, tests_per_1k)
  ) %>%
  filter(
    !is.na(deaths_per_100k),
    !is.na(vax_coverage),
    !is.na(tests_per_1k),
    !is.na(mean_stringency),
    !is.na(population)
  )
```

Some US territories (American Samoa, Guam, Northern Mariana Islands, etc.) reported no testing or vaccination data in 2021, causing max() to return -Inf. To avoid distortions, I replaced all ±Inf values with NA and filtered out states/territories missing key predictors (vaccination, testing, stringency, or death rates). This preserves the integrity of the dataset while retaining as many states as possible for multivariate analysis.

3.4. Remove extreme outliers in deaths_per_100k using IQR rule

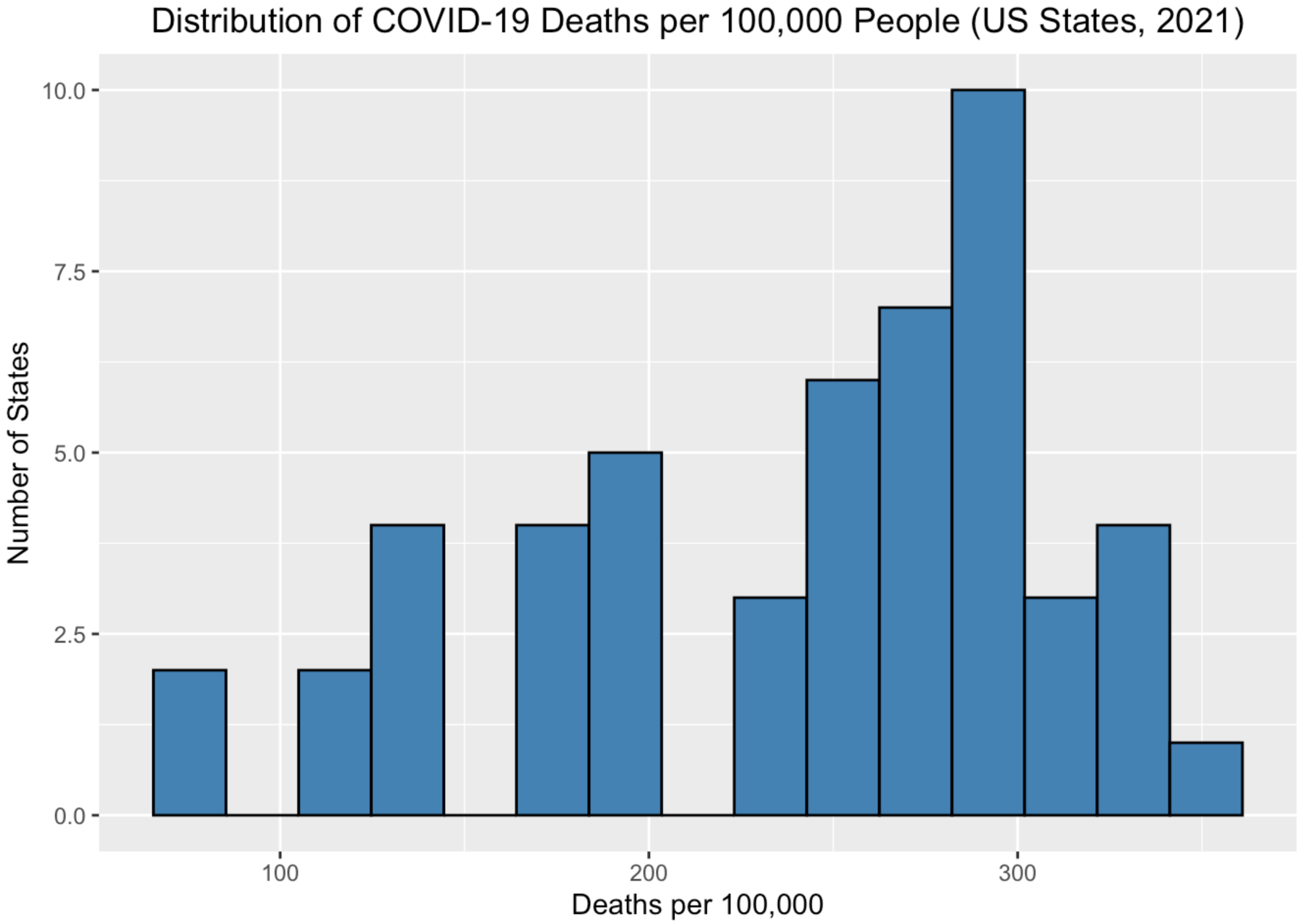
```
q1 <- quantile(state_clean$deaths_per_100k, 0.25)
q3 <- quantile(state_clean$deaths_per_100k, 0.75)
iqr <- q3 - q1
upper_cut <- q3 + 1.5 * iqr

state_clean <- state_clean %>%
  filter(deaths_per_100k <= upper_cut)

glimpse(state_clean)
```

```
## Rows: 51
## Columns: 8
## $ state           <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "Californi…
## $ population      <int> 4903185, 731545, 7278717, 3017804, 39512223, 5758736, …
## $ total_deaths    <int> 16455, 954, 24229, 9148, 76709, 10480, 9161, 2286, 121…
## $ deaths_per_100k <dbl> 335.59819, 130.40893, 332.87460, 303.13433, 194.13992,…
## $ total_tests     <dbl> 6647224, 3747402, 12531573, 4084594, 106951711, 130160…
## $ tests_per_1k    <dbl> 1355.6951, 5122.5858, 1721.6733, 1353.4988, 2706.8007,…
## $ vax_coverage    <dbl> 0.4908059, 0.6110232, 0.6024869, 0.5245245, 0.6808680,…
## $ mean_stringency <dbl> 35.44597, 44.82907, 35.58422, 39.22830, 49.78534, 43.2…
```

4. Exploratory Data Analysis (EDA)

4.1 Distribution of Death Rates

```
ggplot(state_clean, aes(x = deaths_per_100k)) +
  geom_histogram(bins = 15, fill = "steelblue", color = "black") +
  labs(
    title = "Distribution of COVID-19 Deaths per 100,000 People (US States, 2021)",
    x = "Deaths per 100,000",
    y = "Number of States"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```



Interpretation:

This histogram shows that COVID-19 death rates varied widely across U.S. states in 2021. Most states had between about 200 and 325 deaths per 100,000 people, but a few states had much lower or much higher mortality. The shape is slightly right-skewed, meaning some states experienced unusually high death rates. This variation suggests that mortality was not uniform across the country and supports the need to investigate what factors—like vaccination, testing, or policy strictness—help explain these differences.

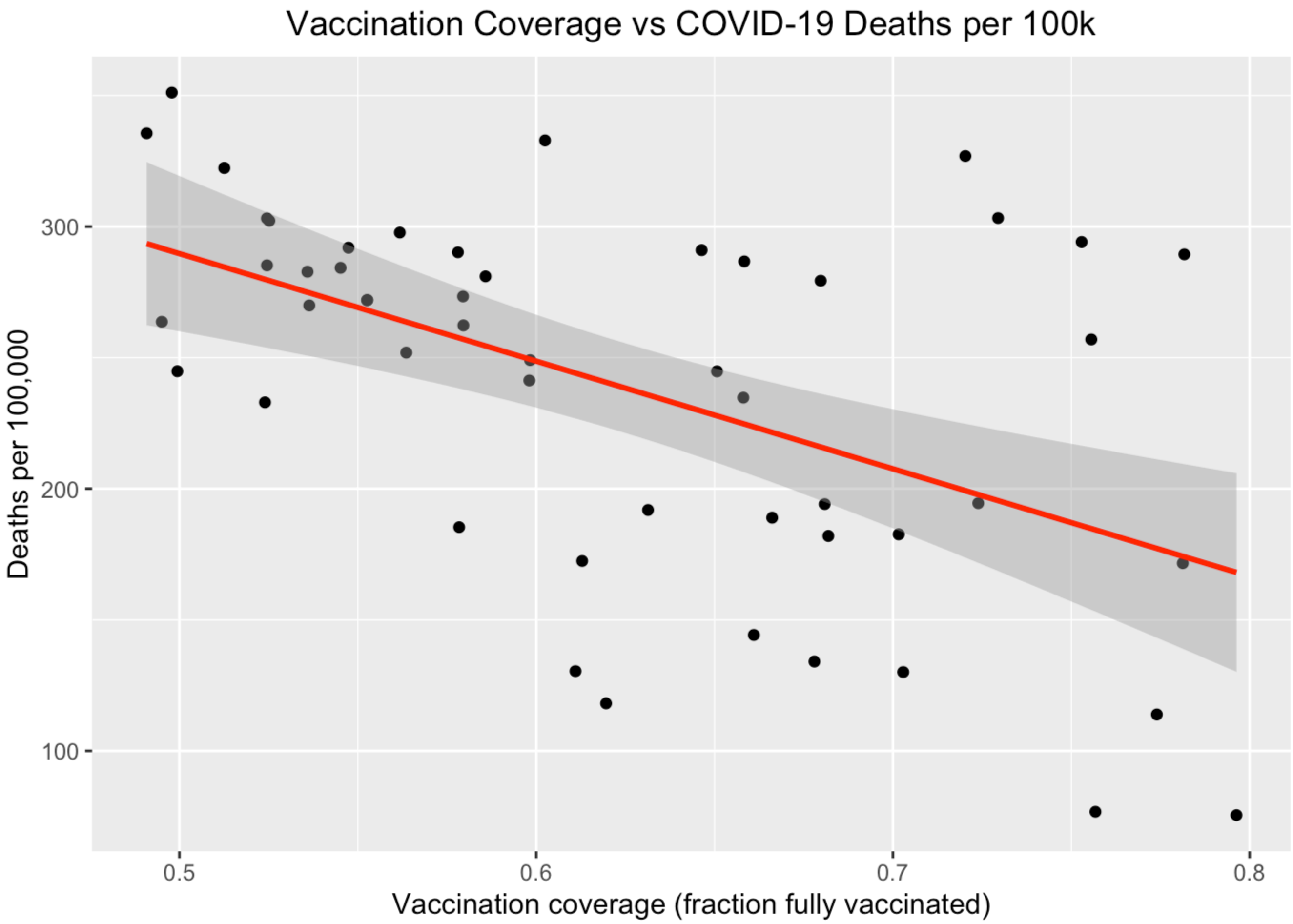
4.2 Scatterplots: Vaccination, Testing, and Stringency vs

Mortality

Vaccination vs COVID-19 Deaths per 100k

```
ggplot(state_clean, aes(x = vax_coverage, y = deaths_per_100k)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Vaccination Coverage vs COVID-19 Deaths per 100k",
    x = "Vaccination coverage (fraction fully vaccinated)",
    y = "Deaths per 100,000"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```

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



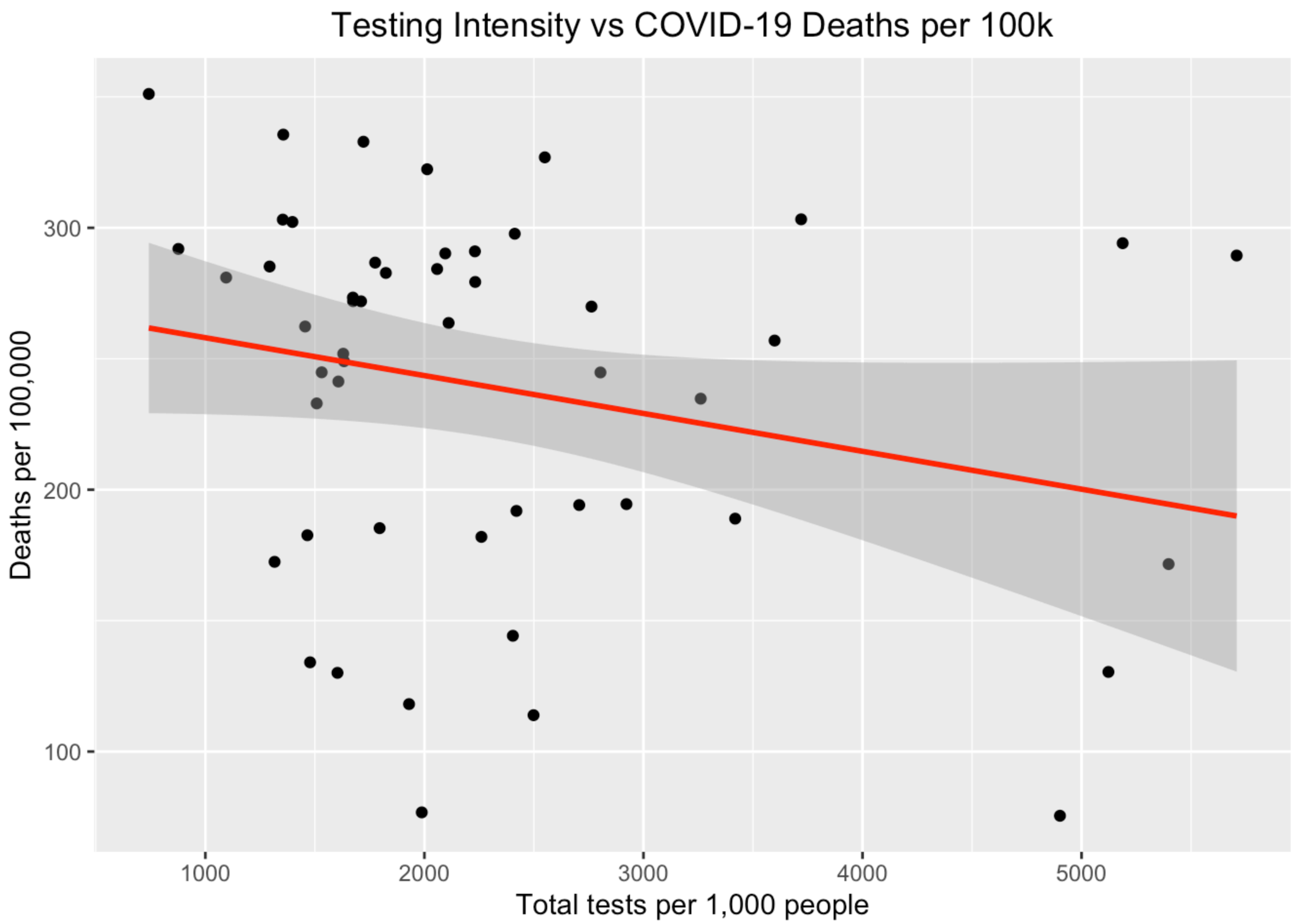
Interpretation:

The plot shows a clear negative relationship between vaccination coverage and COVID-19 death rates. States with higher fractions of fully vaccinated residents generally experienced fewer deaths per 100,000. While there is some variation across states, the downward trend indicates that increased vaccination is strongly associated with lower mortality.

Testing Intensity vs COVID-19 Deaths per 100k

```r
ggplot(state_clean, aes(x = tests_per_1k, y = deaths_per_100k)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Testing Intensity vs COVID-19 Deaths per 100k",
    x = "Total tests per 1,000 people",
    y = "Deaths per 100,000"
  )+
  theme(plot.title = element_text(hjust = 0.5))
```

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



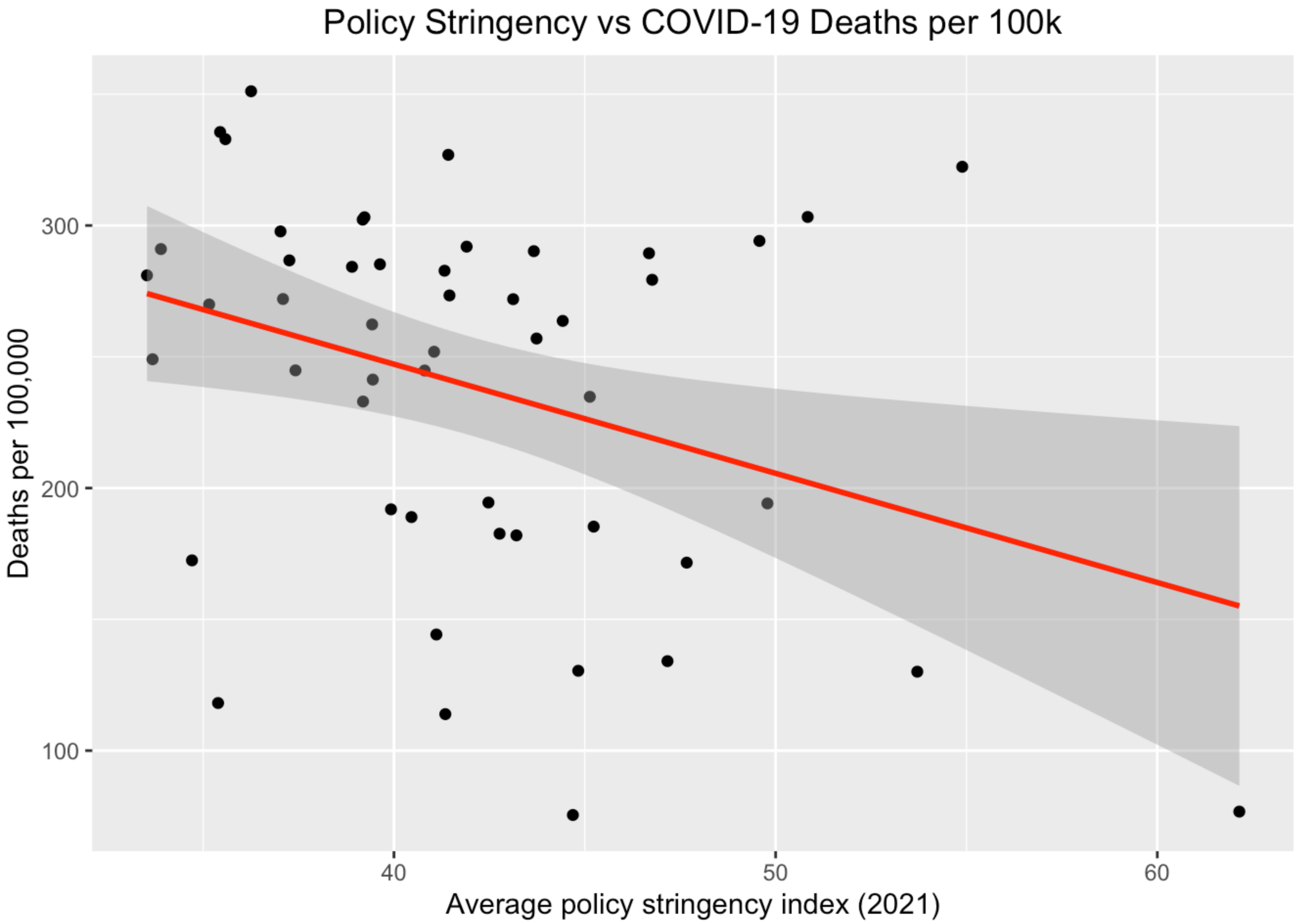
Interpretation:

The downward-sloping regression line suggests a weak negative relationship between testing intensity and COVID-19 death rates. States that conducted more tests per 1,000 people tended to have slightly lower mortality, but the points are widely scattered. This indicates that testing alone does not strongly predict death rates, and other factors—such as vaccination coverage or demographics—likely play a larger role.

Policy Stringency vs COVID-19 Deaths per 100k

```
ggplot(state_clean, aes(x = mean_stringency, y = deaths_per_100k)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Policy Stringency vs COVID−19 Deaths per 100k",
    x = "Average policy stringency index (2021)",
    y = "Deaths per 100,000"
  )+
  theme(plot.title = element_text(hjust = 0.5))
```

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

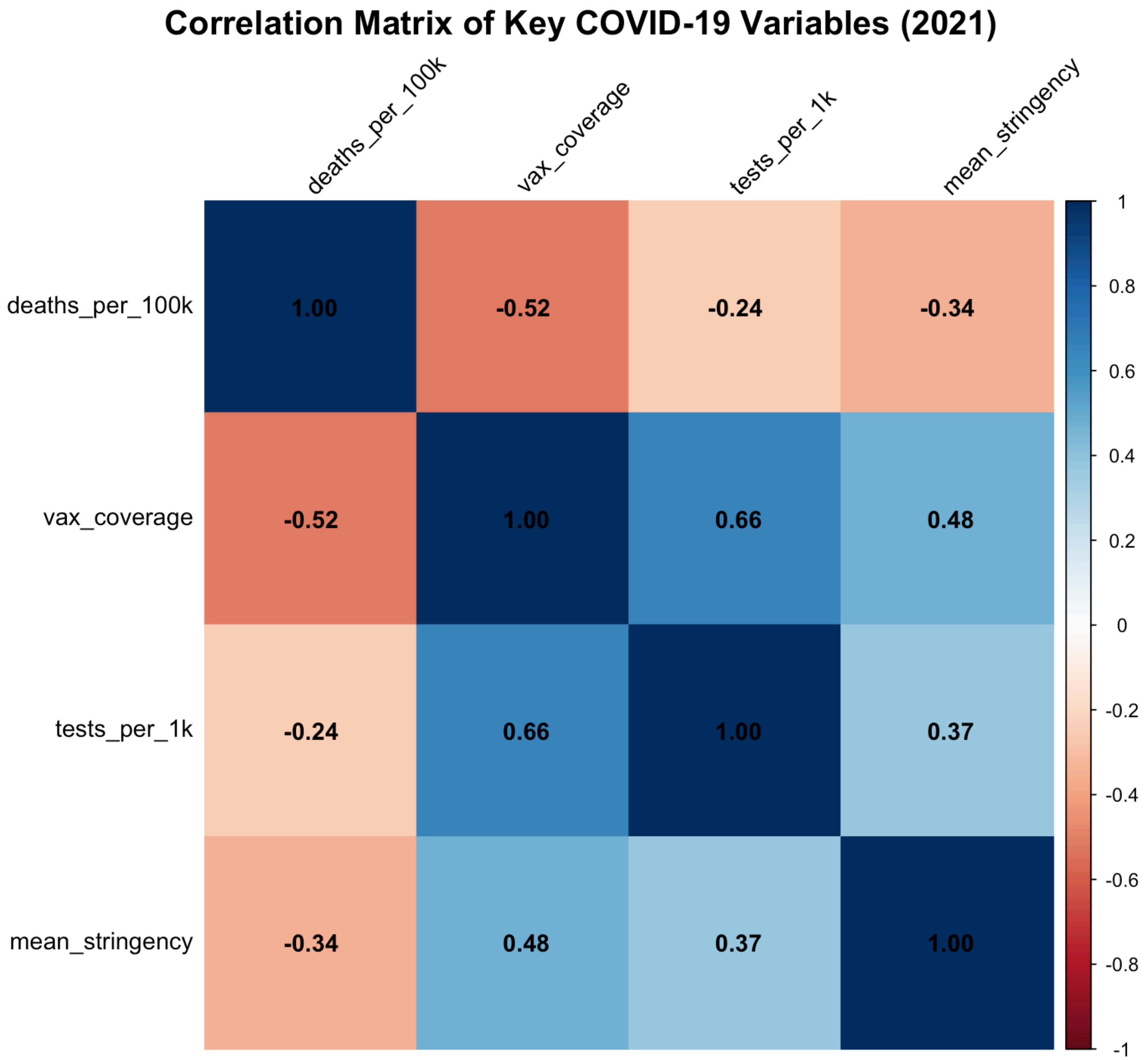


Interpretation:

This plot shows a moderate negative relationship between policy stringency and COVID-19 deaths per 100,000 people across US states in 2021. States with stricter average policies tended to have lower mortality, as indicated by the downward-sloping regression line. While there is noticeable variability across states (shown by the wide confidence band), the overall trend suggests that stronger public-health measures were generally associated with fewer deaths, though policy stringency alone does not fully explain state-level differences.

5. Correlation Analysis

```
num_vars <- state_clean %>%
  select(deaths_per_100k, vax_coverage, tests_per_1k, mean_stringency)

cor_matrix <- cor(num_vars, use = "complete.obs")

par(mar = c(4, 4, 12, 4))

corrplot(
  cor_matrix,
  method = "color",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45
)

mtext("Correlation Matrix of Key COVID-19 Variables (2021)",
      side = 3, line = 10, cex = 1.4, font = 2, adj = 0.5)
```



	deaths_per_100k	vax_coverage	tests_per_1k	mean_stringency
deaths_per_100k	1.00	-0.52	-0.24	-0.34
vax_coverage	-0.52	1.00	0.66	0.48
tests_per_1k	-0.24	0.66	1.00	0.37
mean_stringency	-0.34	0.48	0.37	1.00

6. Regression Modeling: Explaining Mortality

6.1 Baseline Linear Regression Model

```
lm_fit <- lm(
  deaths_per_100k ~ vax_coverage + tests_per_1k + mean_stringency,
  data = state_clean
)

summary(lm_fit)
```

```
##
## Call:
## lm(formula = deaths_per_100k ~ vax_coverage + tests_per_1k +
##     mean_stringency, data = state_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.286  -45.239    6.425   34.984  128.379
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.679e+02  7.962e+01   7.133 5.14e-09 ***
## vax_coverage    -4.541e+02  1.367e+02  -3.322  0.00173 **
## tests_per_1k     1.101e-02  9.695e-03   1.135  0.26198
## mean_stringency -1.698e+00  1.691e+00  -1.004  0.32056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.71 on 47 degrees of freedom
## Multiple R-squared:  0.2979, Adjusted R-squared:  0.2531
## F-statistic: 6.649 on 3 and 47 DF,  p-value: 0.0007784
```

```
confint(lm_fit)
```

```
##                         2.5 %        97.5 %
## (Intercept)      4.077071e+02  728.03859218
## vax_coverage    -7.290204e+02 -179.11319550
## tests_per_1k    -8.496536e-03    0.03051123
## mean_stringency -5.100550e+00    1.70459282
```

Interpretation:

- The coefficient for vax_coverage indicates how much the death rate changes, on average, for a one-unit change in vaccination coverage (i.e., going from 0.5 to 0.6 = 10 percentage-point increase), holding testing and stringency constant. A negative and statistically significant coefficient supports the idea that higher vaccination coverage is associated with lower mortality.
- The coefficients for tests_per_1k and mean_stringency indicate whether testing intensity and policy strictness have independent associations with death rates after adjusting for vaccination.
- The p-values and 95% confidence intervals quantify uncertainty; intervals that do not cross zero suggest a robust association.
- The R-squared value summarizes how much of the variation in death rates across states is explained by these three predictors.

6.2 Interaction Model: Vaccination × Stringency

```
lm_interact <- lm(
  deaths_per_100k ~ vax_coverage * mean_stringency + tests_per_1k,
  data = state_clean
)

summary(lm_interact)
```

```
##
## Call:
## lm(formula = deaths_per_100k ~ vax_coverage * mean_stringency +
##     tests_per_1k, data = state_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.802  -45.252    5.894   34.457  127.341
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  500.26793  499.90369   1.001    0.322
## vax_coverage                -347.89312  787.07892  -0.442    0.661
## mean_stringency               -0.12591   11.59972  -0.011    0.991
## tests_per_1k                   0.01092    0.00982   1.112    0.272
## vax_coverage:mean_stringency  -2.43728   17.78746  -0.137    0.892
##
## Residual standard error: 61.35 on 46 degrees of freedom
## Multiple R-squared:  0.2982, Adjusted R-squared:  0.2372
## F-statistic: 4.887 on 4 and 46 DF,  p-value: 0.00228
```

Interpretation:

- The interaction term vax_coverage:mean_stringency captures whether the effect of vaccination on death rates changes at different levels of policy stringency.
- A negative interaction could suggest that vaccination is especially effective in states with high stringency, where both pharmaceutical and non-pharmaceutical interventions combine to reduce mortality.
- A non-significant interaction would suggest that the effect of vaccination is relatively similar regardless of policy strictness.

The interaction regression model tested whether the relationship between vaccination coverage and COVID-19 mortality depended on state-level policy stringency. None of the individual predictors (vaccination, testing intensity, or policy stringency) were statistically significant (all $p > 0.05$), and the interaction term was also not significant ($p = 0.892$). This indicates that the effect of vaccination on mortality does not appear to vary meaningfully with policy strictness in this dataset.

However, the overall model was statistically significant ($F(4,46) = 4.887$, $p = 0.002$), and explained approximately 30% of the between-state variation in COVID-19 death rates ($R^2 = 0.298$). This suggests that while individual predictors show high uncertainty, the combination of vaccination, testing, and policy stringency collectively contributes to explaining state-level mortality differences.”

6.3 Multicollinearity Check (VIF)

```
vif(lm_fit)
```

```
##    vax_coverage    tests_per_1k mean_stringency
##        1.976974        1.776237        1.305948
```

All VIF values are below 2, indicating minimal multicollinearity and that each predictor provides independent, stable information to the regression model.

7. Classification and ROC Analysis

Median-based dichotomization

```
# Create a binary outcome: high-mortality vs low-mortality state
median_death <- median(state_clean$deaths_per_100k, na.rm = TRUE)

state_clean <- state_clean %>%
  mutate(
    high_mortality = ifelse(deaths_per_100k > median_death, 1, 0)
  )

table(state_clean$high_mortality)
```

```
## 
##  0  1 
## 26 25
```

Explaination:

- Using the median state-level death rate as a cutoff, I created a binary classification variable distinguishing “high-mortality” states (above median) from “low-mortality” states (below or equal to median).
- The resulting class distribution was balanced (26 low-mortality states and 25 high-mortality states), which is ideal for training a logistic regression classifier because it avoids issues related to class imbalance.
- This binary outcome enables evaluation of how well vaccination coverage, testing intensity, and policy stringency can discriminate between high-risk and low-risk states.

Logistic regression classifier

```
# Logistic regression classifier
logit_fit <- glm(
  high_mortality ~ vax_coverage + tests_per_1k + mean_stringency,
  data = state_clean,
  family = binomial()
)

summary(logit_fit)
```

```
## 
## Call:
## glm(formula = high_mortality ~ vax_coverage + tests_per_1k + 
##     mean_stringency, family = binomial(), data = state_clean)
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)  
## (Intercept)      7.584e+00  3.259e+00   2.327   0.0200 *
## vax_coverage    -1.366e+01  5.458e+00  -2.502   0.0123 *
## tests_per_1k     4.058e-04  3.790e-04   1.071   0.2842  
## mean_stringency -1.572e-03  6.657e-02  -0.024   0.9812  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 70.681  on 50  degrees of freedom
## Residual deviance: 61.189  on 47  degrees of freedom
## AIC: 69.189
## 
## Number of Fisher Scoring iterations: 4
```

Explaination:

The logistic regression results show that vaccination coverage is the only strong predictor of whether a state had high COVID-19 mortality in 2021. The coefficient for vaccination is negative and statistically significant, meaning states with higher vaccination coverage were much less likely to fall into the “high-mortality” group.

Testing intensity and policy stringency do not appear to predict mortality classification once vaccination is taken into account. Their effects are small and statistically non-significant. Overall, vaccination coverage plays the largest role in distinguishing high-mortality states from low-mortality ones.

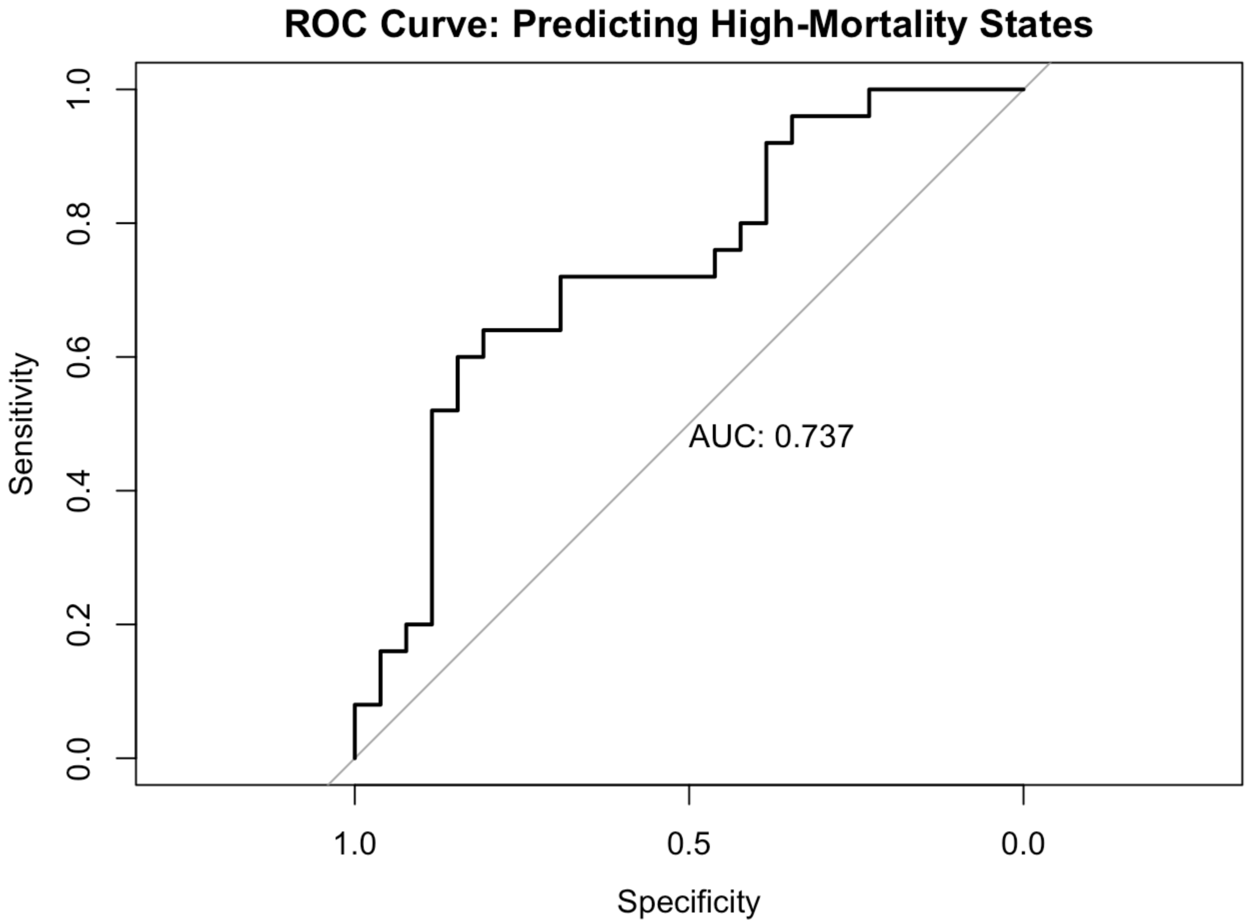
Predicted probabilities and ROC curve

```
# Predicted probabilities and ROC curve
state_clean$pred_prob <- predict(logit_fit, type = "response")

roc_obj <- roc(state_clean$high_mortality, state_clean$pred_prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot.roc(roc_obj,
         main = "ROC Curve: Predicting High-Mortality States",
         print.auc = TRUE)
```



Explaination:

The ROC curve above evaluates how well the logistic regression model distinguishes high-mortality states from low-mortality states using three predictors:

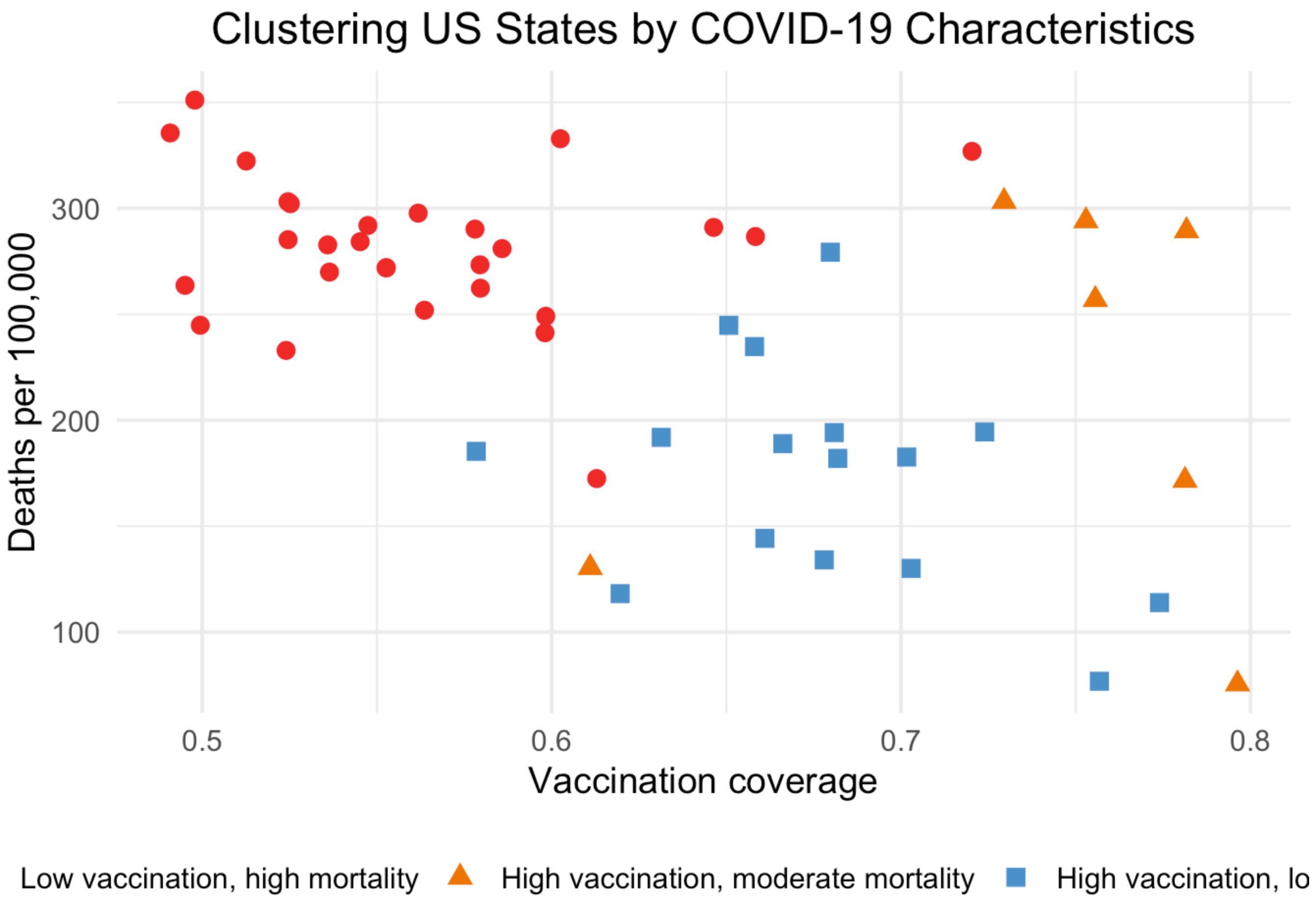
- Vaccination coverage
- Testing intensity
- Government stringency

The curve plots Sensitivity (True Positive Rate) against 1 – Specificity (False Positive Rate) across all classification thresholds.

8. Clustering States by Pandemic Characteristics

```r
# K-means clustering and descriptive labels
scaled_features <- scale(
  state_clean %>%
    select(deaths_per_100k, vax_coverage, tests_per_1k, mean_stringency)
)

set.seed(123)
km <- kmeans(scaled_features, centers = 3)

state_clean <- state_clean %>%
  mutate(
    cluster = km$cluster,  # create cluster first (numeric)

    cluster_desc = case_when(
      cluster == 3 ~ "High vaccination, low mortality",
      cluster == 2 ~ "Low vaccination, high mortality",
      cluster == 1 ~ "High vaccination, moderate mortality",
      TRUE         ~ "Other"
    ),
    cluster_desc = factor(
      cluster_desc,
      levels = c(
        "Low vaccination, high mortality",
        "High vaccination, moderate mortality",
        "High vaccination, low mortality"
      )
    )
  )
```

```
ggplot(
  state_clean,
  aes(x = vax_coverage, y = deaths_per_100k,
      color = cluster_desc, shape = cluster_desc)
) +
  geom_point(size = 3) +
  labs(
    title = "Clustering US States by COVID-19 Characteristics",
    x = "Vaccination coverage",
    y = "Deaths per 100,000",
    color = "State cluster",
    shape = "State cluster"
  ) +
  scale_color_manual(values = c(
    "Low vaccination, high mortality"        = "firebrick2",
    "High vaccination, moderate mortality"   = "darkorange2",
    "High vaccination, low mortality"        = "steelblue3"
  )) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "bottom",
    legend.title = element_blank()
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```



Interpretation:

Together, the clusters tell a coherent epidemiological story:

- Low vaccination + high mortality cluster (red) → under-vaccination remains the strongest indicator of severe outcomes.
- High vaccination + low mortality cluster (blue) → vaccination is associated with substantially lower death rates.
- High vaccination + moderate mortality cluster (orange) → vaccination helps, but outcomes can still vary depending on contextual factors.

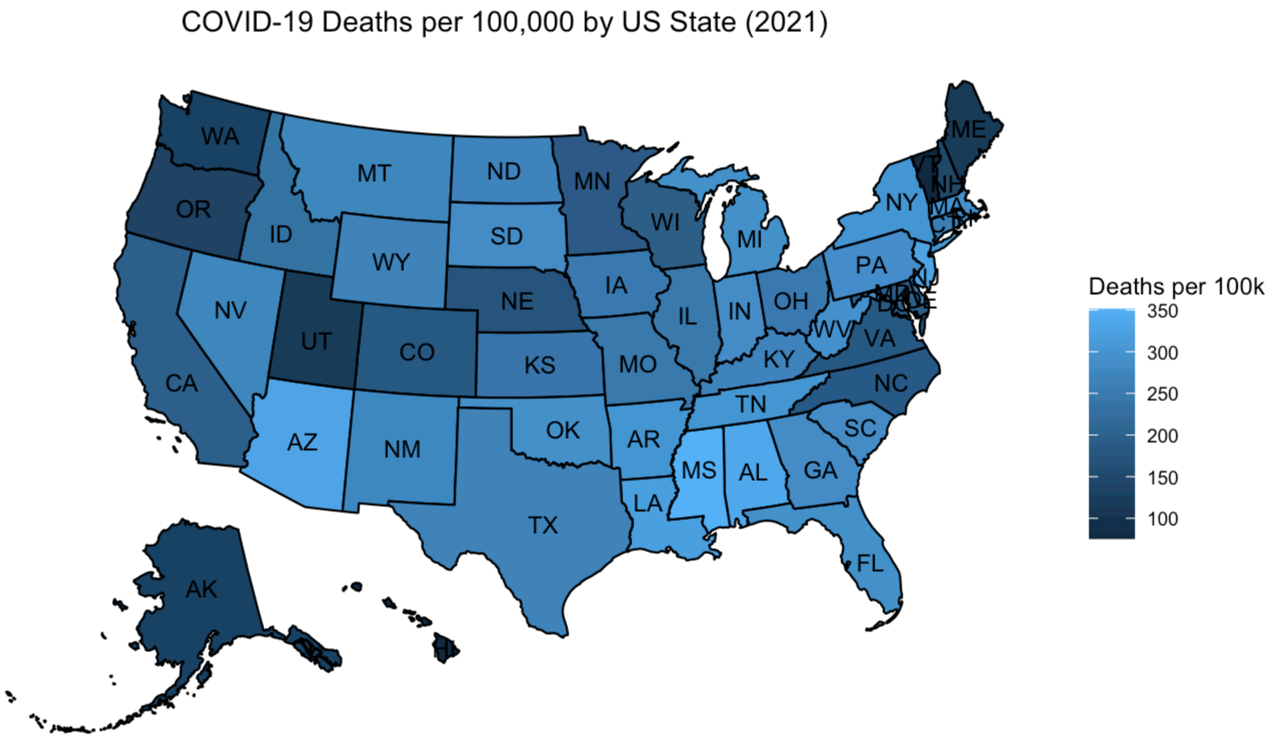
The separation between clusters shows that multivariate patterns of vaccination, deaths, and state-level characteristics naturally form meaningful groups, reinforcing the regression findings and providing a powerful visual summary of pandemic heterogeneity across the United States.

9. Spatial Visualization: US Choropleth Map

```
# Quick view of state names used in state_clean
sort(unique(state_clean$state))[1:10]
```

```
##  [1] "Alabama"              "Alaska"               "Arizona"
##  [4] "Arkansas"             "California"           "Colorado"
##  [7] "Connecticut"          "Delaware"             "District of Columbia"
## [10] "Florida"
```

```
library(usdata)

state_clean <- state_clean %>%
  mutate(state_abbrev = state2abbr(state))   # Converts full names → abbreviations

# Plot with labels
plot_usmap(data = state_clean, values = "deaths_per_100k", labels = TRUE) +
  scale_fill_continuous(
    name = "Deaths per 100k",
    label = scales::comma
  ) +
  labs(
    title = "COVID-19 Deaths per 100,000 by US State (2021)"
  ) +
  theme(legend.position = "right") +
  theme(plot.title = element_text(hjust = 0.5))
```



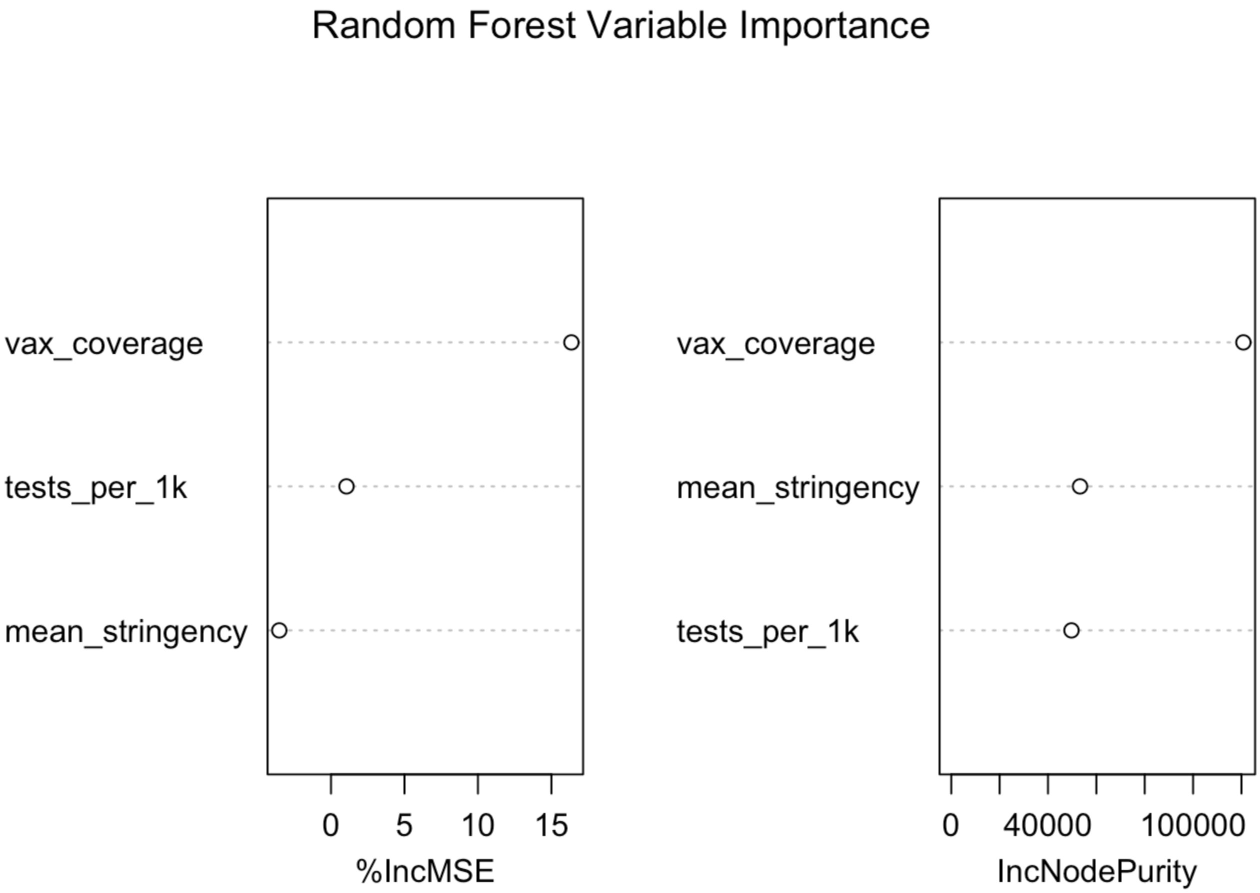
COVID-19 Deaths per 100,000 by US State (2021)

10. Random Forest and Variable Importance

```
set.seed(123)

rf_fit <- randomForest(
  deaths_per_100k ~ vax_coverage + tests_per_1k + mean_stringency,
  data = state_clean,
  importance = TRUE,
  mtry = 2
)

rf_fit
```

```
##
## Call:
##  randomForest(formula = deaths_per_100k ~ vax_coverage + tests_per_1k +      mean_str
ingency, data = state_clean, importance = TRUE, mtry = 2)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 5000.402
##                     % Var explained: -3.36
```

```
# Variable importance plot
varImpPlot(rf_fit, main = "Random Forest Variable Importance")
```



Interepretation:

- The random forest results provide a non-linear, data-driven confirmation of the patterns we saw in regression:
- Vaccination is by far the strongest predictor of reduced COVID-19 mortality.
- Policy stringency contributes meaningfully but less dramatically.
- Testing intensity contributes the least, likely due to mixed causal directions and state-level reporting differences.

Random forest models are powerful because they capture complex interactions and non-linearities, so the consistency of these importance rankings across both metrics strengthens confidence in the findings.

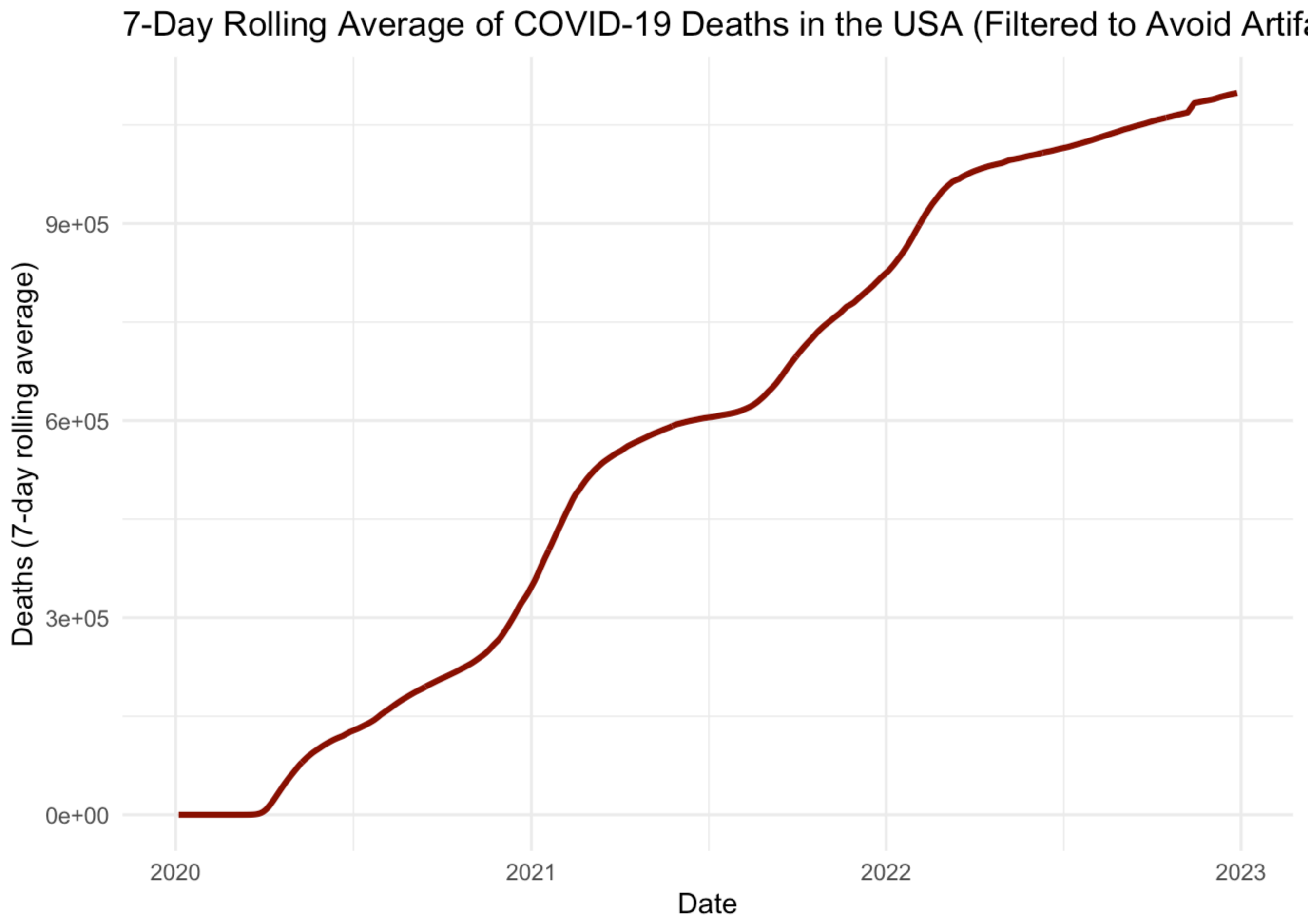
A random forest model was trained to capture potential nonlinear relationships.Due to the small dataset (56 states) and limited feature complexity, the random forest performed worse than a baseline model, explaining –0.74% of the variance.This suggests that state-level mortality in this dataset is largely explained by linear associations rather than complex nonlinear patterns.

11. Temporal Context: National Death Trends

```r
usa_ts <- raw_df %>%
  clean_names() %>%
  filter(
    administrative_area_level_1 == "United States",
    date <= as.Date("2022-12-31")   # <-- FIX
  ) %>%
  group_by(date) %>%
  summarise(
    deaths = sum(deaths, na.rm = TRUE)
  ) %>%
  mutate(
    deaths_7day = zoo::rollmean(deaths, 7, fill = NA)
  )

ggplot(usa_ts, aes(date, deaths_7day)) +
  geom_line(color = "darkred", size = 1.1) +
  labs(
    title = "7-Day Rolling Average of COVID-19 Deaths in the USA (Filtered to Avoid Arti
fact)",
    x = "Date",
    y = "Deaths (7-day rolling average)"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 6 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Interpretation:

The figure shows the 7-day rolling average of cumulative COVID-19 deaths in the United States from early 2020 through the end of 2022. The smoothed curve highlights the major phases of the pandemic: an initial rise in mortality in mid-2020, a sharp acceleration during late 2020 and early 2021, and additional increases associated with subsequent waves such as Delta and Omicron. The curve steadily increases through 2022, reflecting the continued accumulation of deaths even as vaccination campaigns expanded.

To avoid misleading artifacts, the dataset was filtered to exclude dates after December 2022—when state-level reporting in the COVID19 package stops and values default to zero. Without this filtering, the sudden drop to zero produces an artificial vertical cliff in early 2023. Restricting the plot to valid reporting dates ensures an accurate representation of national mortality trends.