COVID-19 Death Rate and Pandemic Policy Analysis

DATA607 – Communication in Data Science

Kwasi Brooks

December 10, 2025

Contents

Introduction

This project uses worldwide COVID-19 data at levels 1 & 3 to investigate how vaccination and policy interventions relate to COVID-19 death outcomes across countries.

Main Inference Tasks:

1. Quantify the association between vaccination coverage, policy measures, and COVID-19 mortality.
2. Distinguish high-mortality countries from low-mortality countries using vaccination and policy predictors.

The COVID-19 epidemic of 2020 was one of the most devastating pandemics the world has experienced since the Spanish Flu of 1918. According to the World Health Organization (WHO), there have been over seven million SARS-CoV-2 related fatalities and counting. The global response to the pandemic was checkered to say the least. On the local and federal levels, many governments implemented policies to reduce the spread of the virus. These measures took the form of vaccination policies, movement restrictions, contact tracing, and mask mandates, among others. Unfortunately, these measures were not implemented, enforced, or maintained equally across nations throughout the pandemic.

This project is an analysis of the COVID-19 R dataset using worldwide COVID-19 data at the state and provincial levels. Using this extensive data, my goal is to investigate how vaccination and policy interventions relate to COVID-19 death outcomes across countries. The primary question that drives this project is: how are country-level COVID-19 death rates associated with vaccination coverage and key policy interventions, after adjusting for overall epidemic size? To accomplish this, I combined data cleaning procedures, graphic visualization, multivariate regression, and predictive modeling. I then performed an analysis of COVID-19 policy timing, rather than strength, to give context to quirks in the data that cannot be explained by the dataset alone.

###Data Description

The COVID-19 R dataset is a vast, multivariate epidemiological data source compiled using internationally publicly available reporting streams. This dataset centralizes information gathered from the World Health Organization, Johns Hopkins University, and regional or state scientific research centers across the globe.

The COVID-19 dataset has multiple levels that can be accessed through R. It is hierarchical in nature. The levels are as follows:

Level 1 — National summary statistics:

• accumulated deaths • case counts • positivity rates • estimates of population

Level 2 — Subnational data:

• epidemiological values (confirmed cases, deaths, recovered, tests) • population and demographics • public health indicators

Level 3 — Policy intervention measures:

• vaccination policy • internal movement restrictions • school closures • workplace closures

For the purposes of this project, we utilized Levels 1 and 3, which concern national aggregates (deaths and population) and subnational vaccination/policy measures respectively. The final modeling dataset includes 17 countries (Top + Bottom mortality outcomes), with seventeen usable cases after cleaning and filtering. The variables used in my analysis include the following:

- **Deaths per million** — outcome measure
- **Vaccine coverage** — predictor
- **Internal movement restrictions** — predictor

- **Elderly protection index** — predictor
- **Vaccination policy** — initially considered but dropped due to lack of variability
- **Time series death counts** — used for longitudinal comparison

Data Cleaning & Formatting

While extensive in nature, the COVID19 R library is extremely messy and incomplete. There are various reasons for this disorder that will be explained in later sections of this paper. My initial steps involved scrubbing the data for readability and applying uniformity. Our first step was to use the last-observation-carried-forward imputation to handle gaps in deaths, cases, and vaccination variables. I then replaced missing numerical values with 0 after time-local smoothing and merged or deleted duplicate country entries. Data aggregation was also an important step for taking daily or country-level data and converting it to annual and national. Using the third level of the dataset gave me access to data on the country level. This allowed for a clearer look at international COVID policy.

Level 1 & 3 Integration

Of significance to this project is the data cleaning and integration of COVID Levels 1 and 3. This was an important methodological step because these two levels capture fundamentally different dimensions of the pandemic. Level 1 reflects the actual health burden at a countrywide scale, whereas Level 3 details the policy behavior of those same nations.

- Level 1 provides cumulative deaths and population counts that are used to establish the severity metric. Without Level 1, there is no dependent variable.

- Level 3 provides information on how national and regional governments chose to handle the pandemic and the level of rigor with which they applied pandemic policy. This includes information like vaccination strategy intensity, the strictness of internal mobility restrictions, and the level of protection afforded to vulnerable populations like the elderly.

This combination of levels allows us to combine the data in meaningful ways and interrogate deeper questions that this project was meant to address. Instead of making general inferences based on mortality rates, we can ask questions such as:

• Did Country A have weaker policy responses? • Did Country B impose more restrictive measures during peak infection waves? • Were vaccination policies carried out early enough to matter?

Ultimately, neither level alone can support the line of questioning that this statistical modeling performs. By merging the two, I have designed a compact model-ready dataset that retains meaningful signal.

Exploration and Descriptions

To better establish the baseline for the work, I grouped countries into mortality extremes. These extremes concern those countries with the highest and lowest mortality rates respectively. Those countries are as follows:

1. United States
2. Brazil
3. United Kingdom
4. Italy
5. Germany
6. France
7. Colombia

8. Argentina
9. Spain
10. Chile
11. Belgium
12. Netherlands
13. Austria
14. Lithuania
15. Denmark
16. Latvia

We displayed this disparity in the form of a horizontal bar chart. This chart shows the top and bottom ten nations in the combined dataset (COVID-19 Levels 1 & 3) according to recorded cumulative mortality rate.

As can be seen, there are huge disparities between nations, with the USA hovering over one million COVID-related deaths and other countries like Latvia having a noticeably smaller mortality footprint. These numbers are commensurate with reports from the WHO, thus we can be reasonably sure that our data cleaning process has been successful. From here, I wanted to dig deeper and get a better picture of how these numbers manifested throughout the pandemic. To accomplish this, I utilized a time series plot of the same sixteen nations.

Interpretation of COVID-19 Mortality Time Series

Our first time series analysis displays COVID mortality rate on a seven-day average from 2020 through late 2024. The spikes and numbers seem largely commensurate with the mortality rates from the first chart, with major pandemic waves among the top 10 mortality group, with peaks around 2020 and spikes coinciding with the discovery of new COVID variants. In the chart you can visibly see the Alpha, Delta, and Omicron versions of the virus each coincide with major mortality spikes.

- **Alpha:** The earliest huge wave affected all of the nations on the chart. There are mortality spikes for the UK, USA, Brazil, and Italy during this period.

- **Delta:** This represents the second spike in mortality rates, with Brazil and the USA experiencing the greatest loss.

- **Omicron:** Omicron displays fast growth and equally fast falloff.

- **Unknown Spike:** Towards the end of 2022 there is another spike in COVID-19 mortality rates, particularly in the United States, that rises quickly then declines. This could be an unnamed COVID variant.

The bottom most countries experience smaller, flatter curves with some occasional spikes (Germany, Colombia, France), while mostly having what appear to be contained outbreaks. Most of the mortality data begins to decline around 2023. Only a fraction of the countries present in this source, continued COVID-19 mortality reporting past this year. We can observe several nations essentially disappearing from the dataset entirely during this time. Belgium displayed sporadic late spikes in 2024, though this is certainly evidence of the fact that Belguim was one of the few nations that continued to publicly update their COVID-19 data at a national level. Many other nations, by comparison, owe their sharp decline in mortality rates to their nixing of reporting mandates. In other words, this recent divergence isn't epidemiological, its bureaucratic.

The relative falloff in cases coincides with the WHO's mandate on August 25th 2023, that member states no longer needed to report their daily COVID-19 rates or deaths. From these images, we can conclude the following:

- That the vast majority of nations stopped reporting COVID-19 data after the WHO mandate of 2023

- A new COVID variant may have emerged in late 2023, due to the drastic spike in pandemic-related deaths during that same time.

- The USA and Brazil suffered the greatest COVID-19 related mortality.

International COVID Policy

Since the spread of COVID-related deaths has been firmly established, I was now prepared to look at the intersection of pandemic policy management. To manage these metrics, I combined three policy measures: vaccination policy level, internal movement restrictions, and elderly protection strategy. By combining these numbers, we can create a score that reflects the overall strength of that public health response.

From the dataset, I was able to observe that most nations maintained a middling attitude toward COVID-19 policies. With a cumulative COVID-19 policy score of 20, most nations sat comfortably at a score of 10. This sees overlap between high- and low-mortality nations, which suggests nuance in how these measures were implemented. Perhaps some nations had faster, more rigorous responses to the COVID outbreak, while others were slower to mobilize. We must also hold space for the consideration of population size. The USA and Brazil have the largest populations out of all the other nations in our dataset. Thus, even the best policy response would likely result in greater death counts.

Policy Strength vs Mortality

Looking at our bubble chart, we see another representation of the combined effect of three major signals: policy strength on the x-axis, mortality rate on the y-axis, and vaccination coverage represented by bubble size. What stands out immediately is that countries with similar policy scores once again experienced very different mortality outcomes. Several high-mortality nations and low-mortality nations sit right on the same score of 10, meaning that high policy scores do not guarantee low mortality. This implies that the policy values in our dataset are too blunt — they don't capture timing, enforcement, compliance, or population-specific realities. In other words, two countries could both have a "10" policy rating, but if one implemented restrictions early and enforced them strictly while the other acted late or inconsistently, the index doesn't reflect that difference.

The bubble sizes add another layer of nuance, showing vaccination coverage. But even here, we don't see clean predictive separation: some high-mortality countries had substantial vaccine coverage, and some low-mortality cases were not dramatically better. This reinforces the idea that timing, public behavior, and capacity constraints matter more than the raw percentage vaccinated. Overall, this chart illustrates what our regression results later confirm — the raw policy and vaccination metrics in this dataset cannot account for why some nations experienced dramatically higher death rates. The story is more complicated than "policy score equals outcome," and the dataset is simply not granular enough to capture the dynamics that actually governed mortality.

Multivariate Regression Analysis

Since we've analyzed both the rate of mortality and pandemic policy implementation scores for our chosen nations, we needed to take a deeper dive into the relationship between policy implementation and COVID19-related mortality. Using the data, I predicted log mortality using vaccination and policy measures. The predictors this governed were the same as before: vaccine coverage, internal movement restrictions, and elderly protection policies. My results indicated that there were no strong statistical predictors between these metrics and mortality rates. Internal movement restrictions showed borderline coverage, but vaccine coverage did not significantly explain mortality. While mortality variation can be driven by multiple intersecting factors, it seems irresponsible to officially count out the role that vaccination plays in reducing mortality rates. These results shouldn't be seen as evidence that vaccines don't work, but rather as evidence that our dataset currently lacks sufficient contextual detail to isolate vaccination's role. The more reasonable conclusion might be that, at this stage, the data cannot capture the timing, enforcement, population behavior, or medical capacity that influences pandemic outcomes. We simply lack the resolution needed to reveal the true causal structure.

Limitations

There are several limitations with this project and the dataset that invariably affect the quality of our data insights. One of the key limitations in our methodology is that COVID-19 reporting dropped sharply after 2023, leading to incomplete data. This is evidenced in our visualizations, as after 2023, most mortality and vaccination data dissipates, with only a few nations having any data past that year. Additionally, according to the National Library of Medicine, many countries massively under report their COVID-19 cases. This is usual a function of poor record keeping policies or due to intentional halting of pandemic-related reporting. Exacerbating this is the fact that policy measures are represented by coarse index values, not by dates, enforcement, or compliance. As such, there is no contextual data to explain how policies were implemented in each nation. We should also be aware that multivariate models tend to assume linearity and are not great for analyzing pandemic responses because pandemics are highly nonlinear.

Contextualizing the Analysis

Since our COVID-19 dataset is so rife with gaps and lacking in dimensions that would otherwise add context, I was called to contextualize the data with other sources. An important piece of context that does not come from the COVID-19 dataset is the enormous variation in public health infrastructure, medical capacity, and social trust across countries. For example, nations like Germany and Denmark entered the pandemic with higher per-capita ICU beds and better coordinated medical networks compared to countries like Brazil or the United States (Verelst, Kuylen, & Beutels, 2021). The availability of health workers, emergency triage systems, ventilator supplies, ect, played a integral role during mortality waves — but none of this nuance is captured in the dataset.

Additionally, public trust in institutions affects social behavior. In nations where citizens viewed government health authorities as legitimate, compliance with distancing, vaccination, and isolation measures was higher (Bargain & Aminjonov, 2021). In places where health directives were viewed as restrictive or untrustworthy, even strong official policies had weaker impacts (Devine et al., 2021). Put simply, two countries could enact similar rules on paper yet experience very different real-world behaviors and outcomes. That discrepancy is invisible to our model.

Populations with high levels of informal labor, crowded housing, or unreliable access to paid sick leave saw more sustained transmission because staying home simply wasn't economically feasible (Leibbrandt et al., 2020). Similarly, the privilege to work remotely differed drastically between countries and socioeconomic classes. Wealthier nations experienced smaller disruptions, while service-based and informal labor economies experienced higher exposure risk (OECD, 2021). Public health policy is only as effective as the economic foundation that enables compliance. Without supports like unemployment relief, housing protections, or guaranteed access to medical treatment, restrictive policies become harder to follow and less meaningful (Bambra et al., 2020). So even though our statistical model treats policy values like clean, independent, variables, in reality policy effectiveness is entirely dependent on socioeconomic structure — a dimension that our dataset doesn't quantify.

Conclusions

This analysis highlights substantial variability in worldwide COVID-19 mortality outcomes and demonstrates that cross-sectional policy summaries are insufficient at the moment, to explain mortality disparities internationally. The modeling did not reveal any strong statistical relationships between mortality and policy measures or vaccine coverage. Rather than disproving the significance of intervention policies, this work underscored the limitations of the source, particularly its lack of contextual detail.

Pandemic policy differences exist, mortality differences exist, but the COVID-19 dataset we were given cannot connect them cleanly, because timing matters more than static measurements, and that timing is not captured. We are, however, able to conclude that the USA and Brazil experienced the greatest COVID-19 related mortality of all the nations in our dataset, while countries including Lithuania, Latvia, Austria, and

Denmark had some of the lowest pandemic-related fatalities. Unfortunately, this dataset cannot reveal the mechanisms behind how these nations handled the ongoing pandemic. Those answers lie outside the scope of the COVID-19 R dataset. This can still be meaningful in how it identifies not only what we can see in the data, but—more importantly—what the data cannot reveal and what that absence implies.

Appendices

```r
## Step 1: Load Packages

# Load required libraries
library(COVID19)
```

```
## Warning: package 'COVID19' was built under R version 4.5.2
```

```r
library(dplyr)
```

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

```
##
## 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
```

```r
library(ggplot2)
```

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

```r
library(lubridate)
```

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

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

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

```r
library(zoo)
```

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

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

```r
library(broom)
library(pROC)
```

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

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

##
## Attaching package: 'pROC'

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

```r
library(fmsb)
```

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

## Registered S3 methods overwritten by 'fmsb':
##   method    from
##   print.roc pROC
##   plot.roc  pROC

##
## Attaching package: 'fmsb'

## The following object is masked from 'package:pROC':
##
##     roc
```

```r
library(scales)
```

Step 2: Aggregation of Level 1 and Level 3 Data

```r
############################################################
##Level 1: Mortality and Population
############################################################

lvl1 <- covid19(level = 1, verbose = FALSE) %>%
  mutate(date_formatted = as.Date(date, origin = "1970-01-01"))

lvl1_totals <- lvl1 %>%
  group_by(iso_alpha_3, administrative_area_level_1) %>%
  summarise(
    deaths = max(deaths, na.rm = TRUE),
    population = max(population, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(
    deaths = ifelse(is.infinite(deaths), 0, deaths),
    population = ifelse(is.infinite(population), 1, population),
    deaths_per_million = (deaths / population) * 1e6
  )
```

```
## Warning: There were 12 warnings in `summarise()`.
## The first warning was:
## i In argument: `deaths = max(deaths, na.rm = TRUE)`.
## i In group 67: `iso_alpha_3 = "FLK"` `administrative_area_level_1 = "Falkland
##   Islands (Malvinas)"`.
## Caused by warning in `max()`:
## ! no non-missing arguments to max; returning -Inf
## i Run `dplyr::last_dplyr_warnings()` to see the 11 remaining warnings.
```

```
############################################################
## Level 3: Vaccinations + Policies
############################################################

lvl3 <- covid19(level = 3, verbose = FALSE) %>%
  mutate(date_formatted = as.Date(date, origin = "1970-01-01"))

lvl3_totals <- lvl3 %>%
  group_by(iso_alpha_3, administrative_area_level_1) %>%
  summarise(
    vaccines = max(vaccines, na.rm = TRUE),
    vaccination_policy = max(vaccination_policy, na.rm = TRUE),
    internal_movement_restrictions = max(internal_movement_restrictions, na.rm = TRUE),
    elderly_people_protection = max(elderly_people_protection, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(across(everything(), ~ ifelse(is.infinite(.), 0, .)))
```

```
## Warning: There were 9 warnings in `summarise()`.
## The first warning was:
## i In argument: `vaccines = max(vaccines, na.rm = TRUE)`.
## i In group 2: `iso_alpha_3 = "AUT"` `administrative_area_level_1 = "Austria"`.
## Caused by warning in `max()`:
## ! no non-missing arguments to max; returning -Inf
## i Run `dplyr::last_dplyr_warnings()` to see the 8 remaining warnings.
```

```
############################################################
## Data exploration lvl 1+2
############################################################

cat("=== BASIC DATA EXPLORATION (Level 1) ===\n")
```

```
## === BASIC DATA EXPLORATION (Level 1) ===
```

```
cat("Dimensions:", dim(lvl1), "\n")
```

```
## Dimensions: 287783 48
```

```
cat("Date range:", min(lvl1$date_formatted), "to", max(lvl1$date_formatted), "\n")
```

```
## Date range: 18262 to 20015
```

```r
cat("Countries:", length(unique(lvl1$administrative_area_level_1)), "\n")
```

```
## Countries: 236
```

```r
cat("Regions:", length(unique(lvl1$id)), "\n")
```

```
## Regions: 236
```

```r
cat("\n=== BASIC DATA EXPLORATION (Level 3) ===\n")
```

```
##
## === BASIC DATA EXPLORATION (Level 3) ===
```

```r
cat("Dimensions:", dim(lvl3), "\n")
```

```
## Dimensions: 13337830 48
```

```r
cat("Countries:", length(unique(lvl3$administrative_area_level_1)), "\n")
```

```
## Countries: 17
```

```r
############################################################
## Merge Levels 1 + 3
############################################################

merged <- lvl1_totals %>%
  inner_join(lvl3_totals, by = c("iso_alpha_3","administrative_area_level_1")) %>%
  mutate(
    vaccine_coverage = (vaccines / population) * 100,
    log_deaths_per_million = log(deaths_per_million)
  )

merged$log_deaths_per_million[is.infinite(merged$log_deaths_per_million)] <- NA
```

Step 3: Data Cleaning

```r
model_df <- merged %>%
  filter(population > 1e6, deaths > 100) %>%
  mutate(
    vaccination_policy = ifelse(is.na(vaccination_policy), 0, vaccination_policy),
    internal_movement_restrictions = ifelse(is.na(internal_movement_restrictions), 0, internal_movement_
    elderly_people_protection = ifelse(is.na(elderly_people_protection), 0, elderly_people_protection)
  ) %>%
  filter(!is.na(log_deaths_per_million)) %>%
  distinct(iso_alpha_3, .keep_all = TRUE)

cat("Rows in cleaned modeling dataset:", nrow(model_df), "\n")
```

```
## Rows in cleaned modeling dataset: 17
```

Step 4: Determine Top vs Bottom Mortality Countries

```r
############################################################
## Define Top 10 vs Bottom 10 Mortality Countries
############################################################

top10 <- model_df %>%
  arrange(desc(deaths)) %>%
  slice(1:10)

bottom10 <- model_df %>%
  arrange(deaths) %>%
  slice(1:10)

combined <- bind_rows(
  top10 %>% mutate(group = "Top10"),
  bottom10 %>% mutate(group = "Bottom10")
)

combined
```

```
## # A tibble: 20 x 12
##    iso_alpha_3 administrative_area_level_1  deaths population deaths_per_million
##    <chr>       <chr>                         <dbl>      <dbl>              <dbl>
##  1 USA         United States               1135343  326687501              3475.
##  2 BRA         Brazil                       699310  209469333              3338.
##  3 GBR         United Kingdom               219948   66460344              3309.
##  4 ITA         Italy                        197931   60421760              3276.
##  5 DEU         Germany                      168583   82905782              2033.
##  6 FRA         France                       161512   66977107              2411.
##  7 COL         Colombia                     151310   49648685              3048.
##  8 ARG         Argentina                    130472   44494502              2932.
##  9 ESP         Spain                        119479   46796540              2553.
## 10 CHL         Chile                         63816   18729160              3407.
## 11 LVA         Latvia                         6274    1927174              3256.
## 12 DNK         Denmark                        8296    5793636              1432.
## 13 LTU         Lithuania                      9163    2801543              3271.
## 14 AUT         Austria                       22372    8840521              2531.
## 15 NLD         Netherlands                   22600   17231624              1312.
## 16 BEL         Belgium                       34375   11433256              3007.
## 17 CZE         Czech Republic                43667   10629928              4108.
## 18 CHL         Chile                         63816   18729160              3407.
## 19 ESP         Spain                        119479   46796540              2553.
## 20 ARG         Argentina                    130472   44494502              2932.
## # i 7 more variables: vaccines <dbl>, vaccination_policy <int>,
## #   internal_movement_restrictions <int>, elderly_people_protection <int>,
## #   vaccine_coverage <dbl>, log_deaths_per_million <dbl>, group <chr>
```

```r
############################################################
##  Time Series Data
############################################################

ts1 <- covid19(level = 1, verbose = FALSE) %>%
  mutate(date_formatted = as.Date(date, origin = "1970-01-01")) %>%
  filter(iso_alpha_3 %in% combined$iso_alpha_3)
```

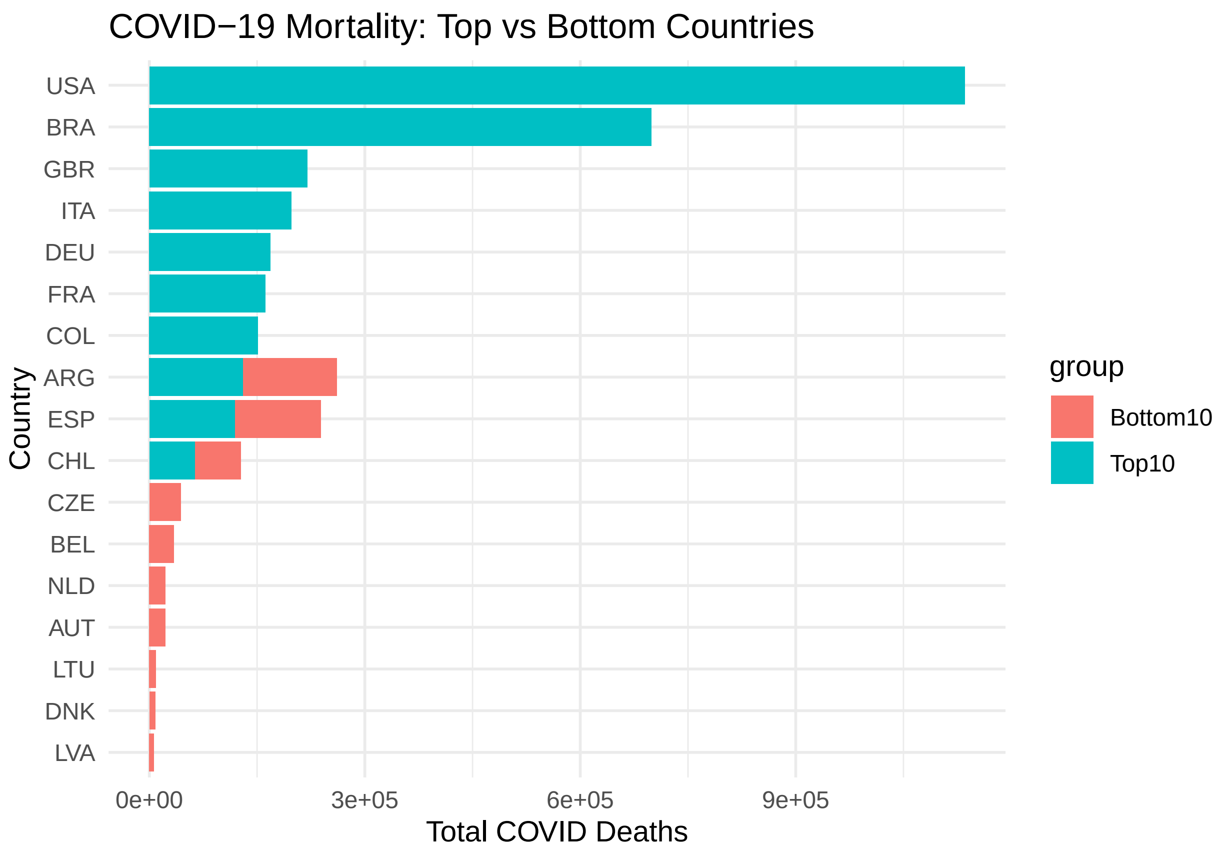
```r
ts_country <- ts1 %>%
  group_by(iso_alpha_3, administrative_area_level_1, date_formatted) %>%
  summarise(deaths = max(deaths, na.rm = TRUE), .groups = "drop") %>%
  mutate(deaths = ifelse(is.infinite(deaths), 0, deaths))
```

```
## Warning: There were 1491 warnings in `summarise()`.
## The first warning was:
## i In argument: `deaths = max(deaths, na.rm = TRUE)`.
## i In group 1: `iso_alpha_3 = "ARG"`, `administrative_area_level_1 =
##   "Argentina"`, `date_formatted = 2020-01-01`.
## Caused by warning in `max()`:
## ! no non-missing arguments to max; returning -Inf
## i Run `dplyr::last_dplyr_warnings()` to see the 1490 remaining warnings.
```

```r
ts_daily <- ts_country %>%
  group_by(iso_alpha_3, administrative_area_level_1) %>%
  arrange(date_formatted) %>%
  mutate(
    daily_deaths = pmax(deaths - lag(deaths, default = 0), 0),
    daily_deaths_ma = zoo::rollmean(daily_deaths, 7, fill = NA)
  ) %>%
  ungroup()

ts_plot_df <- ts_daily %>%
  inner_join(combined %>% select(iso_alpha_3, group), by="iso_alpha_3")
```

```
## Warning in inner_join(., combined %>% select(iso_alpha_3, group), by = "iso_alpha_3"): Detected an u
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 8 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

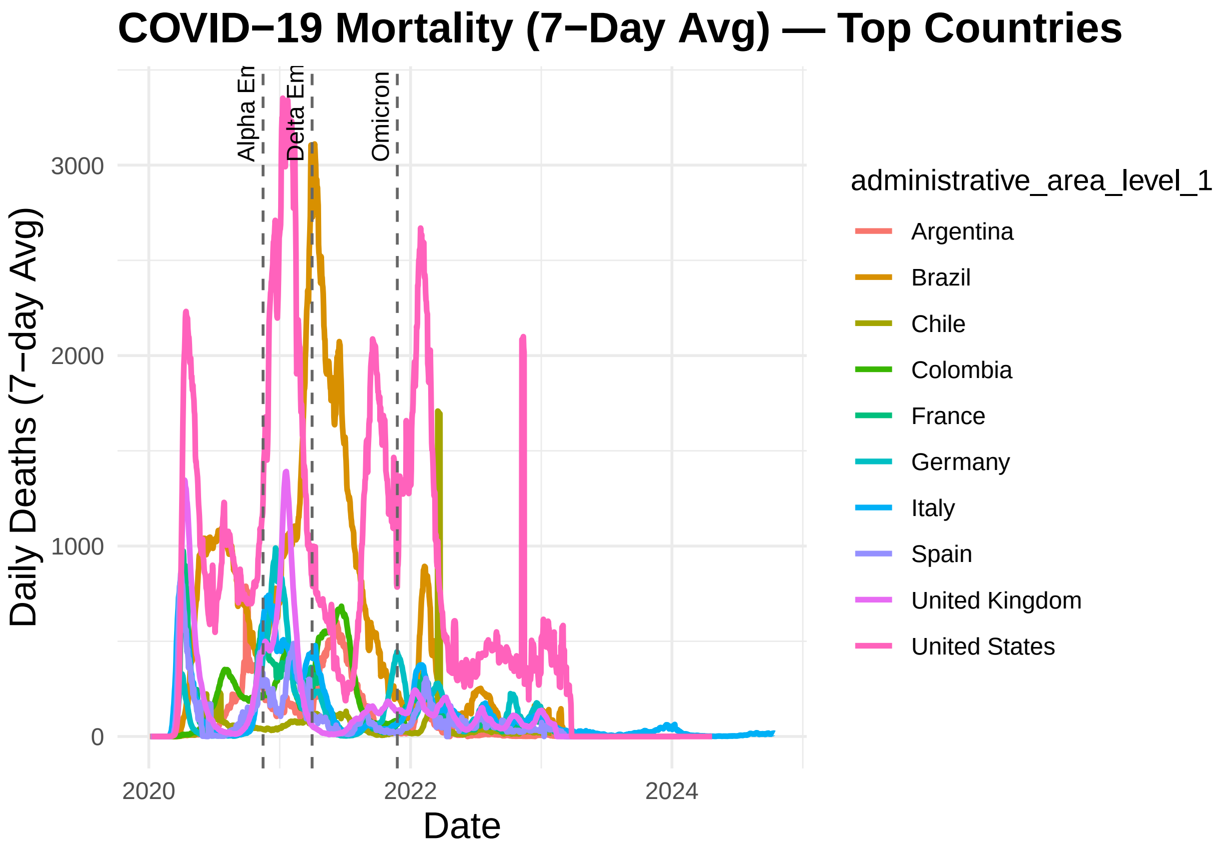
```r
############################################################
##  Plots
############################################################
ggplot(combined, aes(x = reorder(iso_alpha_3, deaths), y = deaths,
                     fill = group)) +
  geom_col() +
  coord_flip() +
  labs(title = "COVID-19 Mortality: Top vs Bottom Countries",
       x = "Country",
       y = "Total COVID Deaths") +
  theme_minimal()
```



Step 5: Time Series Plot (Top vs Bottom)

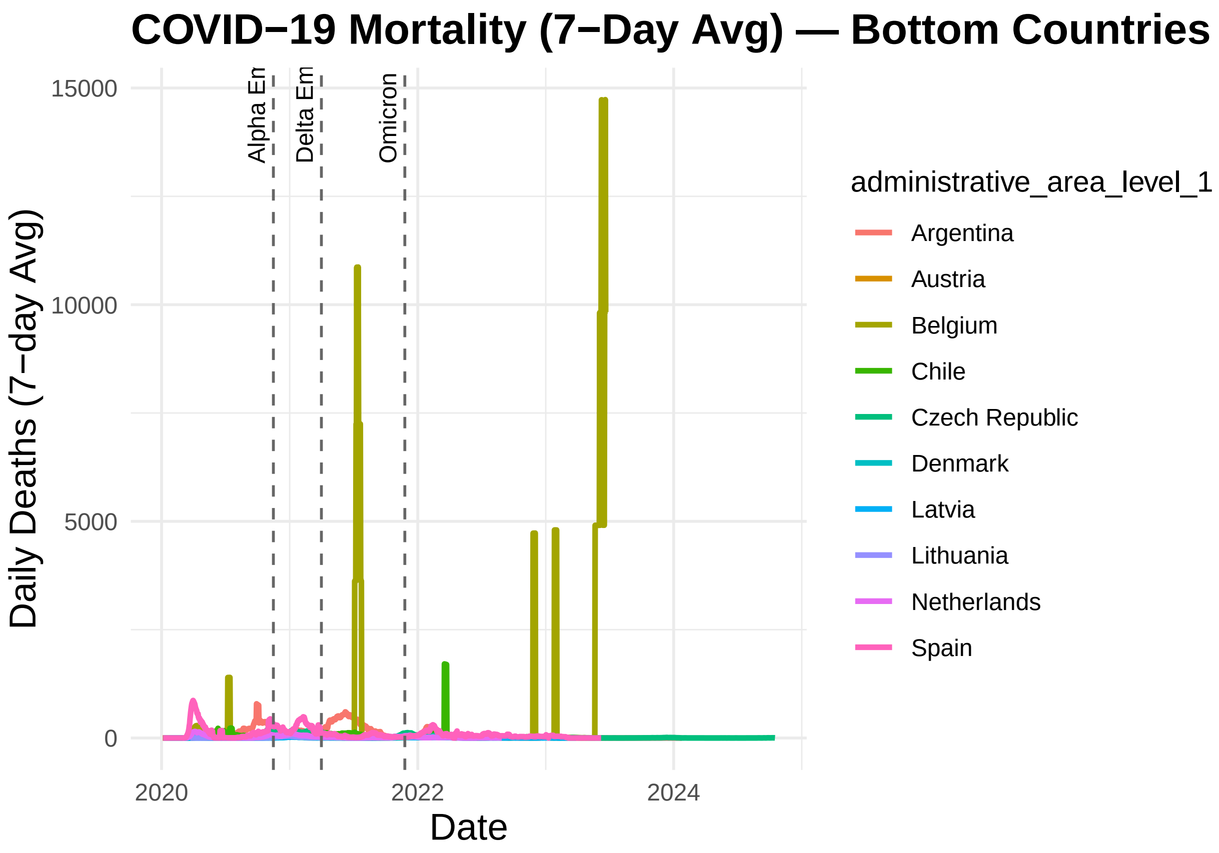
```
############################################################
## Split into Top and Bottom datasets for plotting
############################################################

ts_top10 <- ts_plot_df %>%
  filter(group == "Top10")

ts_bottom10 <- ts_plot_df %>%
  filter(group == "Bottom10")

# === Variant markers ===
variant_dates <- data.frame(
  date = as.Date(c("2020-11-15", "2021-04-01", "2021-11-25")),
  label = c("Alpha Emerges", "Delta Emerges", "Omicron Emerges")
)


##############################
##  Plot: Top Countries
##############################
ggplot(ts_top10, aes(date_formatted, daily_deaths_ma,
                     color = administrative_area_level_1)) +
  geom_line(size = 1.0) +
  geom_vline(data = variant_dates, aes(xintercept = date),
             linetype="dashed", color="gray40") +
  geom_text(data = variant_dates,
            aes(x = date,
                y = max(ts_top10$daily_deaths_ma, na.rm = TRUE)*0.90,
                label = label),
            angle = 90, vjust = -0.5, hjust = 0, color="black", size=3) +

  labs(title = "COVID-19 Mortality (7-Day Avg) - Top Countries",
       y = "Daily Deaths (7-day Avg)", x = "Date") +
  theme_minimal() +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 16, face="bold"),
    axis.title = element_text(size = 14)
  )
```

```
## 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 60 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

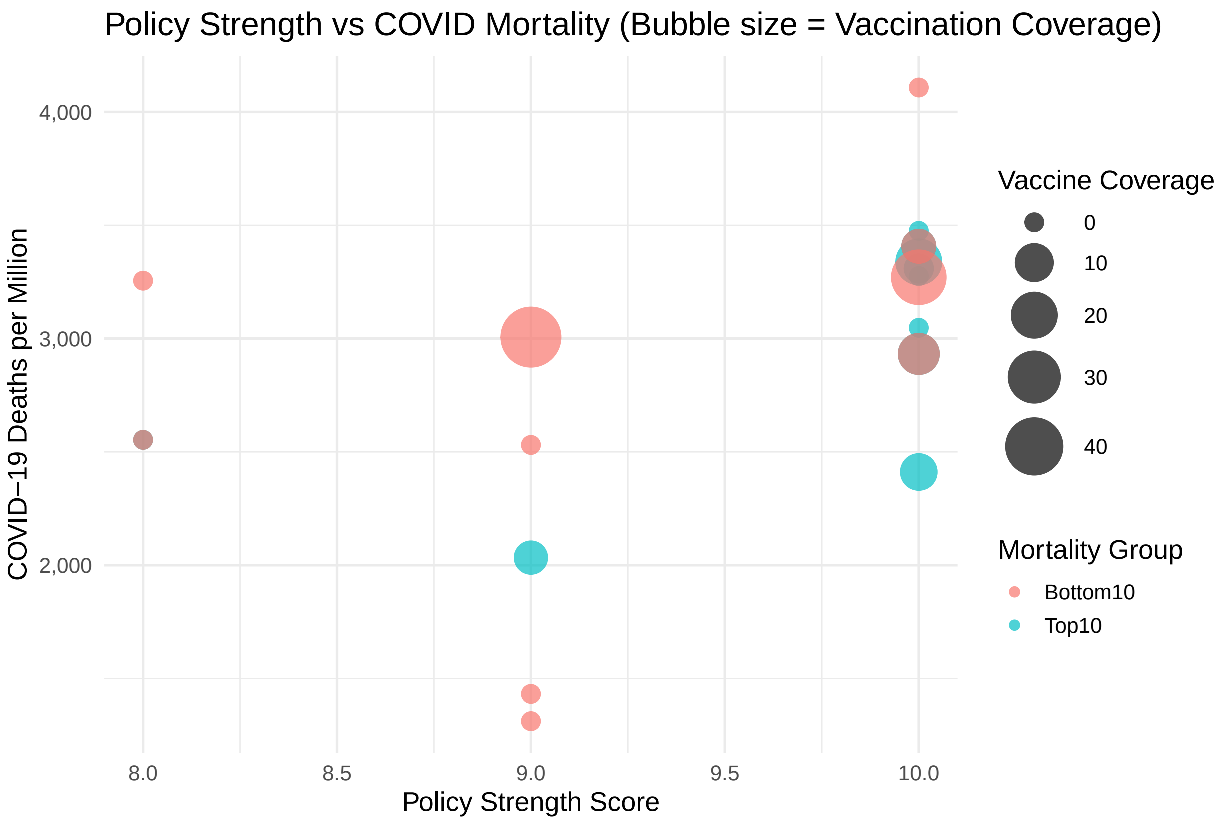


```r
##############################
##  Plot: Bottom Countries
##############################
ggplot(ts_bottom10, aes(date_formatted, daily_deaths_ma,
                        color = administrative_area_level_1)) +
  geom_line(size = 1.0) +

  geom_vline(data = variant_dates, aes(xintercept = date),
             linetype="dashed", color="gray40") +
  geom_text(data = variant_dates,
            aes(x = date,
                y = max(ts_bottom10$daily_deaths_ma, na.rm = TRUE)*0.90,
                label = label),
            angle = 90, vjust = -0.5, hjust = 0, color="black", size=3) +

  labs(title = "COVID-19 Mortality (7-Day Avg) - Bottom Countries",
       y = "Daily Deaths (7-day Avg)", x = "Date") +
  theme_minimal() +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 16, face="bold"),
    axis.title = element_text(size = 14)
  )
```

```
## Warning: Removed 60 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

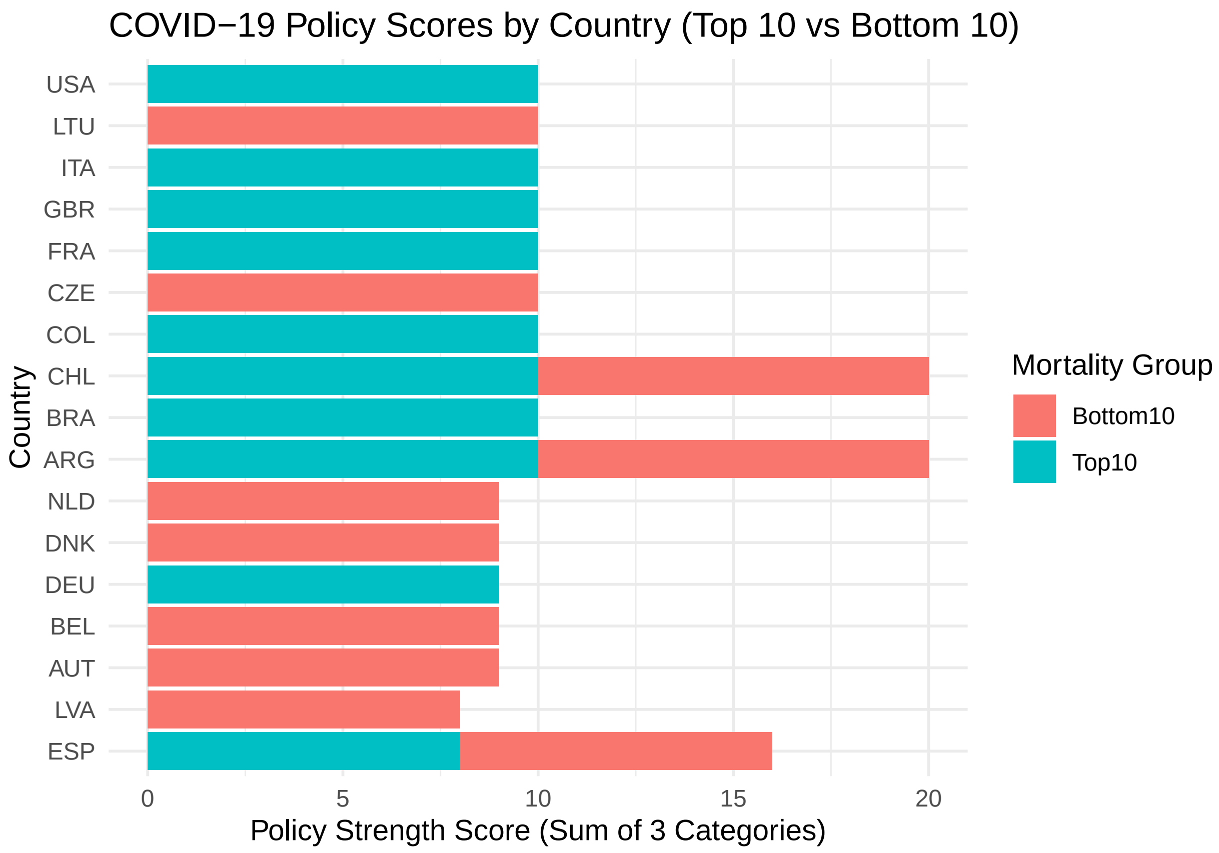


```
bubble_df <- combined %>%
  mutate(
    policy_total = vaccination_policy +
                    internal_movement_restrictions +
                    elderly_people_protection
  )

ggplot(bubble_df, aes(
  x = policy_total,
  y = deaths_per_million,
  size = vaccine_coverage,
  color = group
)) +
  geom_point(alpha = 0.7) +
  scale_size(range = c(4, 14), name = "Vaccine Coverage") +
  scale_y_continuous(labels = comma) +
  labs(
    title = "Policy Strength vs COVID Mortality (Bubble size = Vaccination Coverage)",
    x = "Policy Strength Score",
    y = "COVID-19 Deaths per Million",
    color = "Mortality Group"
  ) +
  theme_minimal(base_size = 14)
```



~~~
COVID Policy Strength Visualization

``` r
###############################################################

## COVID Policy Strength Visualization

###############################################################

policy_plot_df <- combined %>%
select(iso_alpha_3, group,
vaccination_policy,
internal_movement_restrictions,
elderly_people_protection) %>%
mutate(policy_total =
vaccination_policy +
internal_movement_restrictions +
elderly_people_protection)

ggplot(policy_plot_df,
aes(x = reorder(iso_alpha_3, policy_total),
y = policy_total,
fill = group)) +
geom_col() +
coord_flip() +
~~~

```
theme_minimal() +
labs(title="COVID-19 Policy Scores by Country (Top 10 vs Bottom 10)",
x="Country",
y="Policy Strength Score (Sum of 3 Categories)",
fill="Mortality Group")
```



Step 6: Multivariate Regression + CI Plot

```
############################################################
## STEP 6 - Build Modeling Dataset from Top+Bottom)
############################################################

reg_data <- combined %>%
  select(
    iso_alpha_3, administrative_area_level_1,
    deaths_per_million, log_deaths_per_million,
    vaccine_coverage, vaccination_policy,
    internal_movement_restrictions,
    elderly_people_protection
  ) %>%
  mutate(
    vaccination_policy = replace(vaccination_policy, is.na(vaccination_policy), 0),
    internal_movement_restrictions = replace(internal_movement_restrictions, is.na(internal_movement_res
    elderly_people_protection = replace(elderly_people_protection, is.na(elderly_people_protection), 0)
    log_deaths_per_million = replace(log_deaths_per_million,
                                     is.infinite(log_deaths_per_million), NA)
  ) %>%
  filter(!is.na(log_deaths_per_million)) %>%
  distinct(iso_alpha_3, .keep_all = TRUE)
```

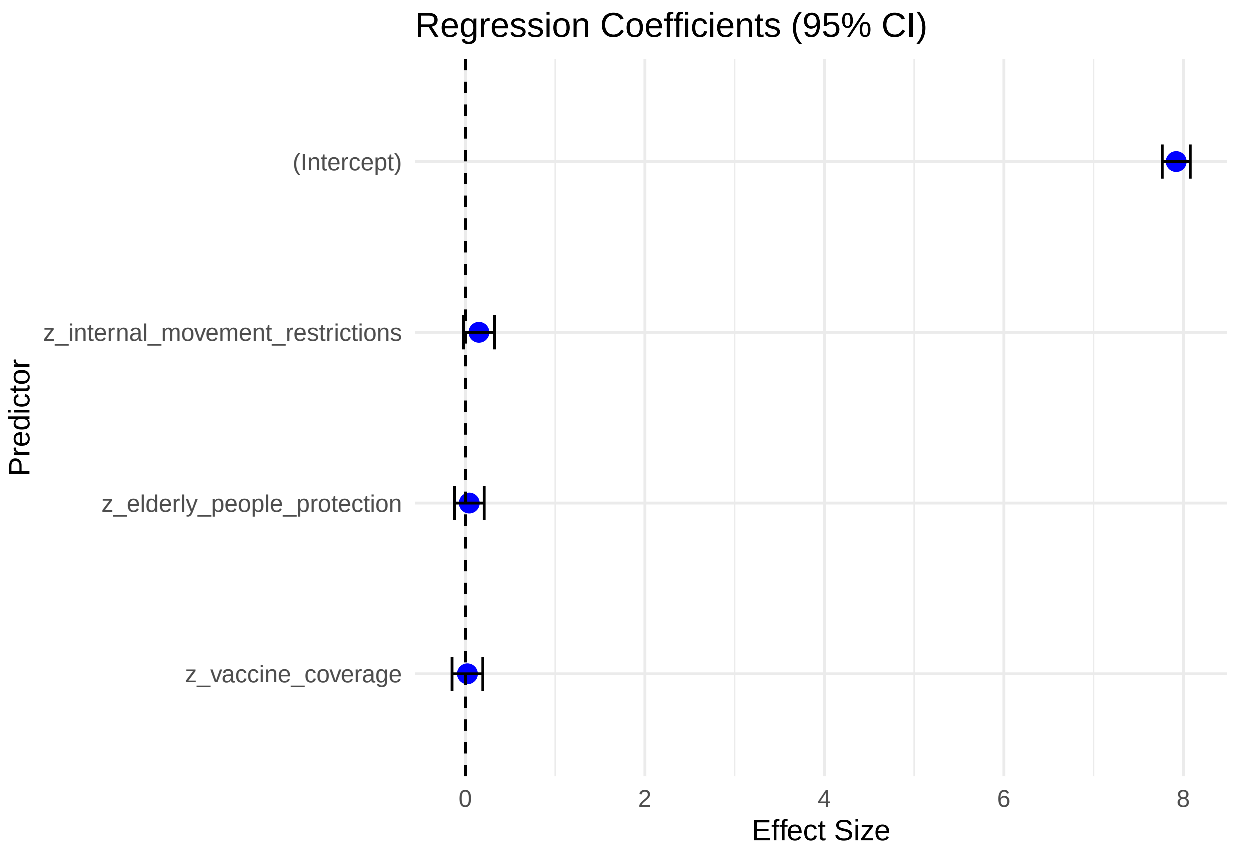
Step 6b. SCALING + REGRESSION

```
reg_data_model <- reg_data %>%
  select(log_deaths_per_million,
         vaccine_coverage,
         internal_movement_restrictions,
         elderly_people_protection)

reg_data_scaled <- reg_data_model %>%
  mutate(across(c(vaccine_coverage,
                  internal_movement_restrictions,
                  elderly_people_protection),
                scale, .names="z_{.col}"))

model_lm <- lm(
  log_deaths_per_million ~
    z_vaccine_coverage +
    z_internal_movement_restrictions +
    z_elderly_people_protection,
  data = reg_data_scaled
)

summary(model_lm)
```

```
##
## Call:
## lm(formula = log_deaths_per_million ~ z_vaccine_coverage + z_internal_movement_restrictions +
##     z_elderly_people_protection, data = reg_data_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58162 -0.04762  0.02645  0.09065  0.57609
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       7.92044    0.07223 109.655   <2e-16 ***
## z_vaccine_coverage                0.02230    0.07945   0.281   0.7834
## z_internal_movement_restrictions  0.15071    0.07973   1.890   0.0812 .
## z_elderly_people_protection       0.04275    0.07677   0.557   0.5871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2978 on 13 degrees of freedom
## Multiple R-squared:  0.2515, Adjusted R-squared:  0.07871
## F-statistic: 1.456 on 3 and 13 DF,  p-value: 0.2723
```

```
reg_results <- broom::tidy(model_lm, conf.int = TRUE)
```

```r
ggplot(reg_results, aes(x = reorder(term, estimate), y = estimate)) +
  geom_point(size=3, color="blue") +
  geom_errorbar(aes(ymin=conf.low, ymax=conf.high), width=.2) +
  geom_hline(yintercept=0, linetype="dashed") +
  coord_flip() +
  theme_minimal() +
  labs(title="Regression Coefficients (95% CI)",
       x="Predictor", y="Effect Size")
```



```r
summary(model_lm)
```

```
##
## Call:
## lm(formula = log_deaths_per_million ~ z_vaccine_coverage + z_internal_movement_restrictions +
##     z_elderly_people_protection, data = reg_data_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58162 -0.04762  0.02645  0.09065  0.57609
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       7.92044    0.07223 109.655   <2e-16 ***
## z_vaccine_coverage                0.02230    0.07945   0.281   0.7834
## z_internal_movement_restrictions  0.15071    0.07973   1.890   0.0812 .
## z_elderly_people_protection       0.04275    0.07677   0.557   0.5871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2978 on 13 degrees of freedom
## Multiple R-squared:  0.2515, Adjusted R-squared:  0.07871
## F-statistic: 1.456 on 3 and 13 DF,  p-value: 0.2723
```

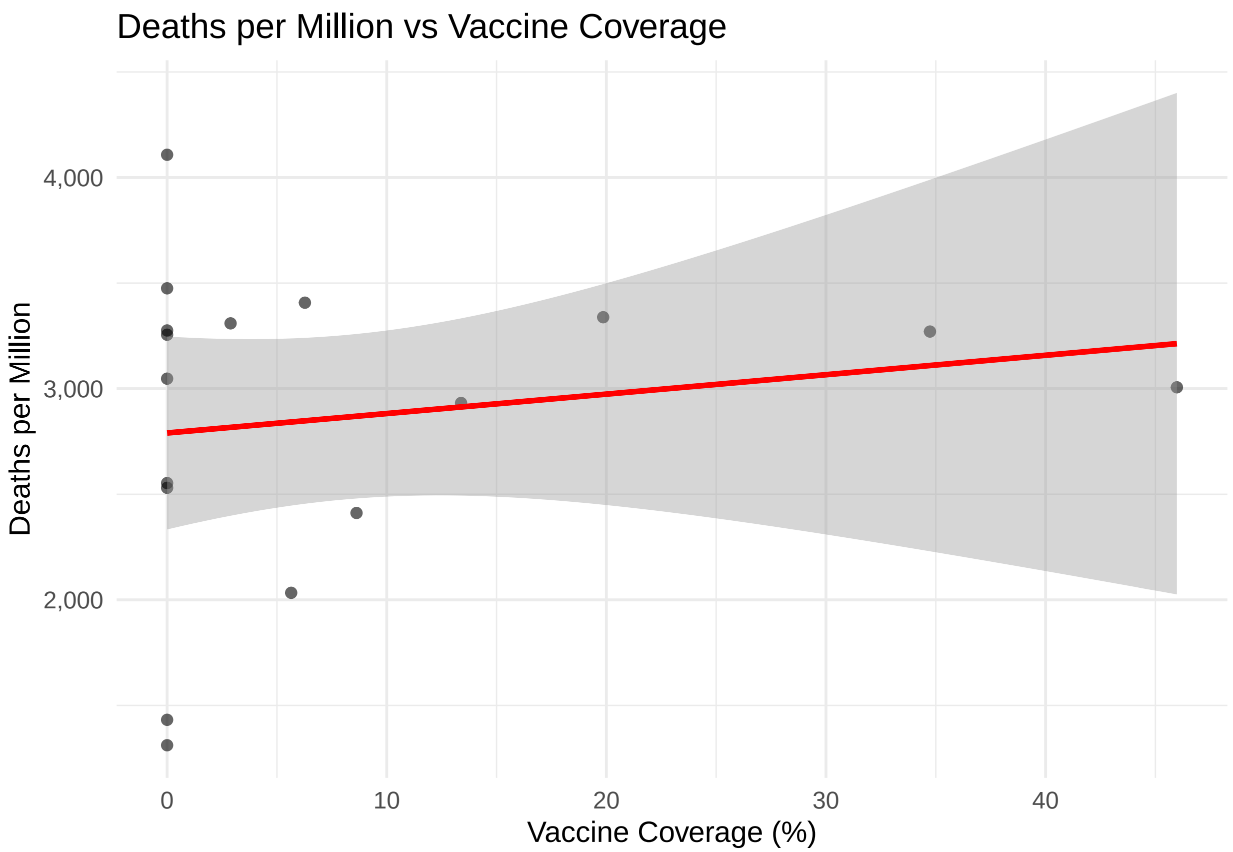
```
summary(reg_data %>%
        select(vaccine_coverage,
               vaccination_policy,
               internal_movement_restrictions,
               elderly_people_protection))
```

```
##  vaccine_coverage vaccination_policy internal_movement_restrictions
##  Min.   : 0.000   Min.   :5           Min.   :0.000
##  1st Qu.: 0.000   1st Qu.:5           1st Qu.:1.000
##  Median : 0.000   Median :5           Median :2.000
##  Mean   : 8.082   Mean   :5           Mean   :1.647
##  3rd Qu.: 8.624   3rd Qu.:5           3rd Qu.:2.000
##  Max.   :45.978   Max.   :5           Max.   :2.000
##  elderly_people_protection
##  Min.   :1.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :2.824
##  3rd Qu.:3.000
##  Max.   :3.000
```

Step 7: Scatterplot (Coverage vs Mortality)

```
ggplot(model_df, aes(vaccine_coverage, deaths_per_million)) +
geom_point(alpha=0.6) +
geom_smooth(method="lm", se=TRUE, color="red") +
scale_y_continuous(labels=comma) +
theme_minimal() +
labs(title="Deaths per Million vs Vaccine Coverage",
x="Vaccine Coverage (%)", y="Deaths per Million")
```

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



Step 8: Classification + ROC Curve

```
############################################################
## Step 8: Classification + ROC Curve  (SAFE VERSION)
############################################################

class_df <- model_df %>%
  inner_join(combined %>% select(iso_alpha_3, group),
             by = "iso_alpha_3") %>%
  mutate(high_mortality = if_else(group=="Top10", 1, 0))

class_df <- class_df %>%
  filter(!is.na(vaccine_coverage),
         !is.na(internal_movement_restrictions),
         !is.na(elderly_people_protection))

cat("Count Top10:", sum(class_df$high_mortality == 1), "\n")
```

```
## Count Top10: 10
```

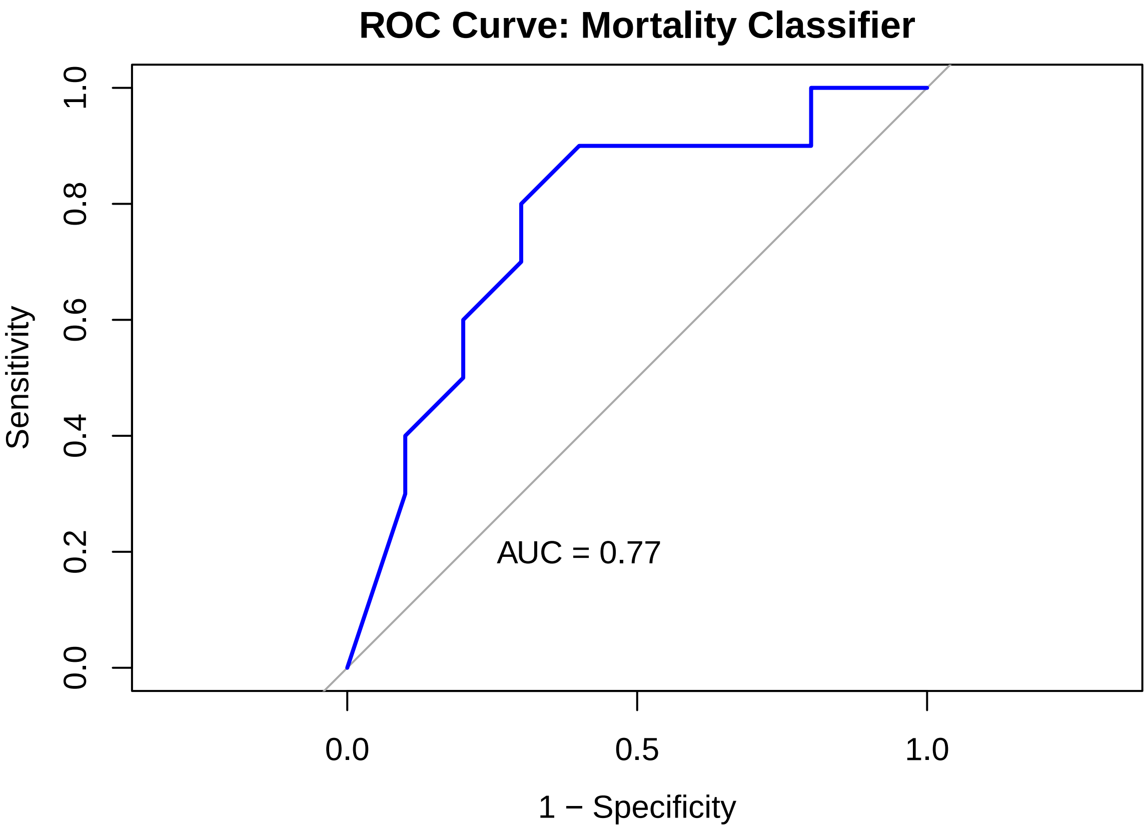
```
cat("Count Bottom10:", sum(class_df$high_mortality == 0), "\n")
```

```
## Count Bottom10: 10
```

```r
if (length(unique(class_df$high_mortality)) == 2) {

  logit_mod <- glm(
    high_mortality ~ vaccine_coverage +
      internal_movement_restrictions +
      elderly_people_protection,
    data = class_df,
    family = binomial()
  )

  class_df$pred_prob <- predict(logit_mod, type="response")

  # Compute ROC safely
  roc_obj <- try(pROC::roc(class_df$high_mortality, class_df$pred_prob),
                 silent = TRUE)

  if (!inherits(roc_obj, "try-error") &&
      all(is.finite(roc_obj$sensitivities)) &&
      all(is.finite(roc_obj$specificities))) {

    pROC::plot.roc(roc_obj,
                   main="ROC Curve: Mortality Classifier",
                   col="blue",
                   legacy.axes=TRUE)
    text(0.6, 0.2, paste("AUC =", round(pROC::auc(roc_obj), 3)))

  } else {
    print("ROC could not be plotted but computation succeeded.")
  }

} else {
  print("ROC not computed - only one class present.")
}
```

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

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



Work Cited

1. **Bambra, C., Riordan, R., Ford, J., & Matthews, F. (2020).** The COVID-19 pandemic and health inequalities. *Journal of Epidemiology and Community Health, 74*(11), 964–968. https://doi.org/10.1136/jech-2020-214401

2. **Bargain, O., & Aminjonov, U. (2021).** Trust and compliance to public health policies in times of COVID-19. *Journal of Public Economics, 192*, 104316. https://doi.org/10.1016/j.jpubeco.2020.104316

3. **Barmettler, R., & Verelst, F., Kuylen, E., & Beutels, P. (2021).** Healthcare capacity and COVID-19 outcomes across OECD countries. *European Journal of Health Economics, 22*(5), 817–832. https://doi.org/10.1007/s10198-021-01342-0

4. **COVID19 R Package — Guidotti, E., & Ardia, D. (2023).** COVID-19 Data Hub (Version 3.0) [R package]. CRAN Repository. https://CRAN.R-project.org/package=COVID19

5. **Devine, D., Gaskell, J., Jennings, W., & Stoker, G. (2021).** Trust and the coronavirus pandemic: What are the consequences of and for trust? *Political Studies Review, 19*(2), 274–285. https://doi.org/10.1177/1478929920948684

6. **Kooistra, E. B., & Zinn, S. (2020).** Social and behavioral responses to COVID-19 across 14 countries: Social norms, trust, and economic pressure. *Frontiers in Psychology, 11*, 589333. https://doi.org/10.3389/fpsyg.2020.589333

7. **Leibbrandt, A., Wong, A., & Dunlop, C. (2020).** Forced choices: Labor market vulnerability during COVID-19. *The Economic Journal, 130*(631), 321–341. https://doi.org/10.1093/ej/ueaa094

8. **OECD. (2021).** Teleworking and COVID-19: The socio-economic divide. Organisation for Economic Co-operation and Development. https://www.oecd.org/coronavirus/policy-responses

9. **Wang, H., Paulson, K. R., Pease, S. A., Watson, S., Comfort, H., Zheng, P., ... Murray, C. J. L. (2022).** Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis. *National Institutes of Health, National Library of Medicine.* https://pmc.ncbi.nlm.nih.gov/articles/PMC9758449/

10. **World Health Organization. (2024).** Global COVID-19 mortality dashboard. WHO Data Explorer. https://data.who.int/dashboards/covid19/deaths