Problem Set 4

Edgar Aguilar

March 22

Report on COVID-19, mask use, and vaccination

Summary

This report examines the relationship between COVID-19 mortality, mask usage, and vaccination rates across U.S. counties in 2022. Using publicly available data from the New York Times, the Centers for Disease Control and Prevention (CDC), and survey data collected by Dynata, I analyze both behavioral and structural factors that may help explain why some counties experienced higher death rates than others. The analysis focuses on three key variables: the total number of COVID-19 deaths per 100,000 residents, the percentage of residents who reported "always" wearing a mask in public, and the percentage of the population that completed a full COVID-19 vaccination series.

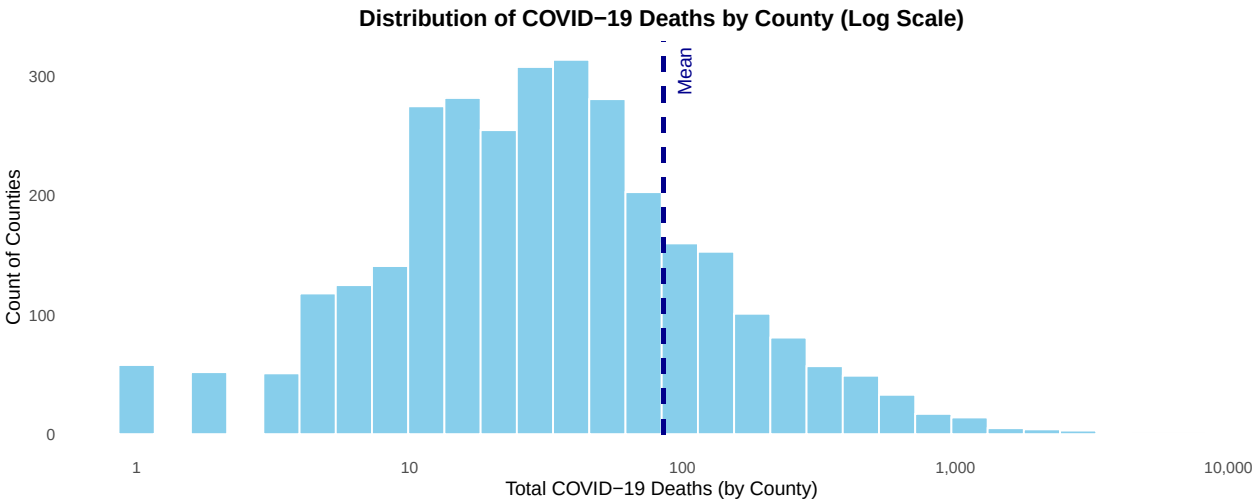
The report combines descriptive statistics, visualizations, and regression analysis to explore how these factors relate to COVID-19 outcomes. Special attention is given to how vaccination rates vary by social vulnerability and how county-level variation reflects broader disparities in public health. Findings from this report contribute to a deeper understanding of how individual behaviors and community conditions influenced the course of the pandemic.

Data

1. COVID-19 Deaths

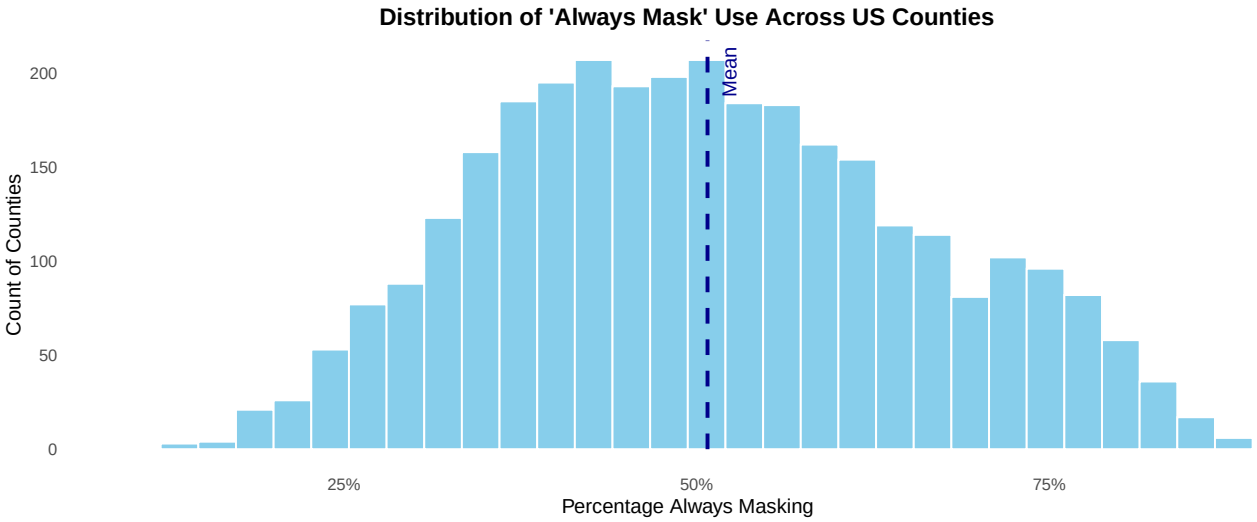
The outcome variable—COVID-19 deaths in 2022—was constructed using cumulative death counts reported by the New York Times. Since the dataset provides cumulative totals over time, I calculated the number of deaths that occurred specifically during 2022 by taking the difference between the maximum and minimum cumulative death counts recorded for each county that year. This method isolates new deaths within the calendar year without requiring daily-level data and avoids double-counting, making it a straightforward and reliable way to extract annual totals. To allow meaningful comparisons across counties of different sizes, I scaled this figure to reflect deaths per 100,000 residents using 2019 U.S. Census population estimates.

The distribution of deaths is highly skewed: while a small number of counties experienced extremely high mortality, most saw relatively modest totals. Summary statistics show that COVID-19 deaths in 2022 ranged from 0 to 7034, with a mean of 84.1, median of 29, and first and third quartiles at 12 and 70, respectively. These figures highlight both the wide variation in pandemic impact across counties and the importance of using log transformations or per capita adjustments in the analysis.



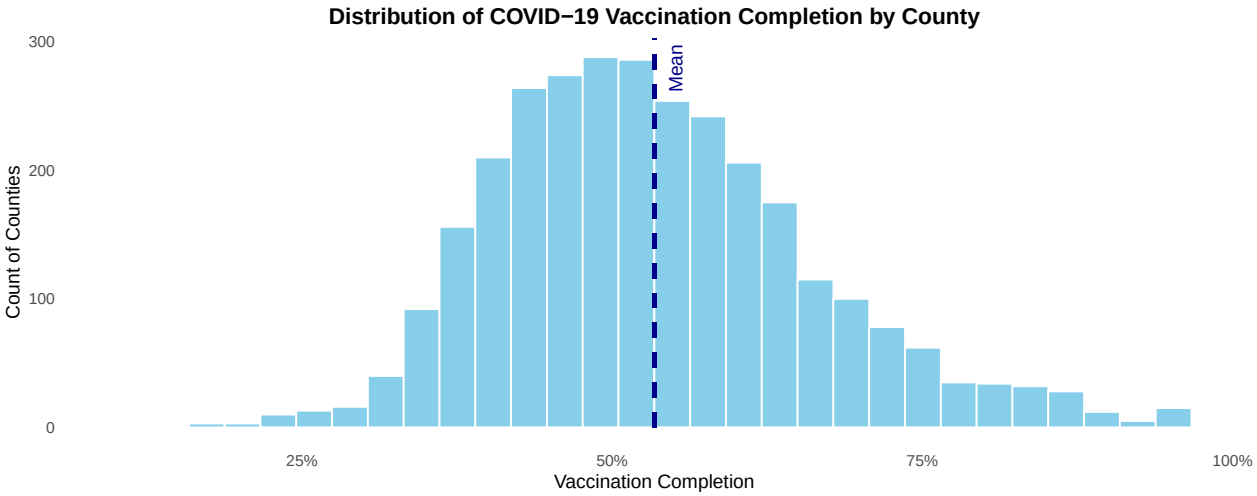
2. Mask Usage

The primary behavioral variable—mask usage—was measured using county-level survey data collected in July 2020 by Dynata in partnership with The New York Times. The dataset captures the percentage of county residents who reported "always" wearing a mask in public, serving as a proxy for local adherence to public health guidelines during the early stages of the pandemic. The distribution of mask usage rates varies widely across counties. Summary statistics show that reported "always masking" ranged from 11.5% to 88.9%, with a mean of 50.8%, median of 49.7%, and interquartile range from 39.3% to 61.3%. These figures reflect both regional variation and differing levels of compliance with non-pharmaceutical interventions across the United States.

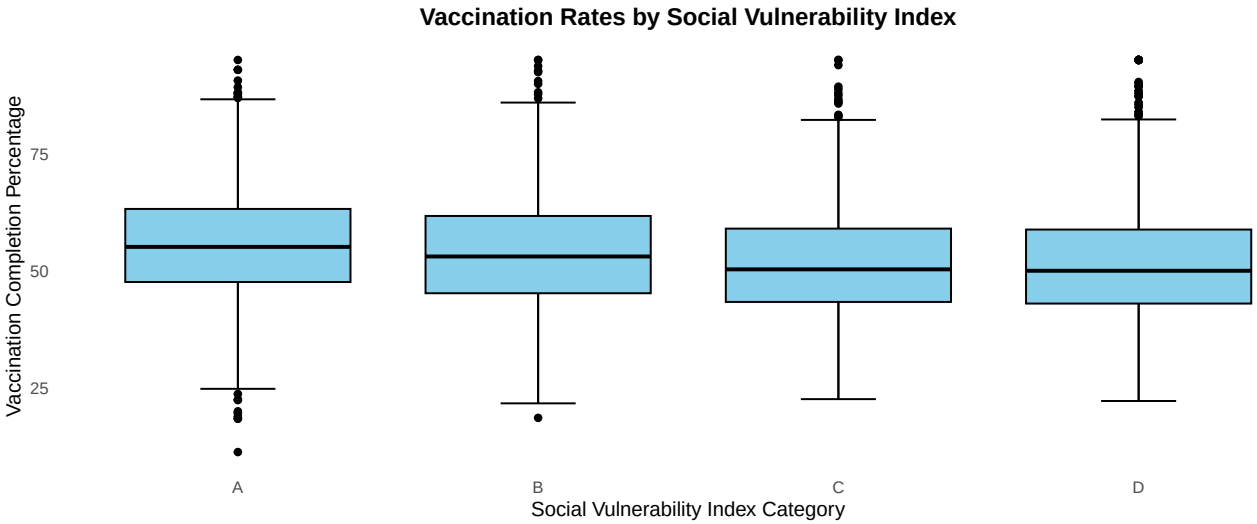


3. Vaccination Rates

The final key variable is COVID-19 vaccination completion, sourced from the CDC's March 2022 county-level dataset. This variable reflects the share of the county population that had completed a full primary vaccination series by that time. Completion rates vary significantly across counties, ranging from 11.3% to 95%, with a mean of 53.4%, median of 52.1%, and interquartile range from 44.3% to 61%. These statistics reveal a substantial gap between the most and least vaccinated areas, with some counties approaching near-universal coverage and others falling well below half the population.



The boxplot shows that COVID-19 vaccination completion rates tend to be lower in counties with higher social vulnerability. Counties in category A, the least vulnerable group, have the highest median vaccination rates, while those in category D, the most vulnerable, have the lowest. While there is some overlap between the groups, the general pattern suggests that vaccination coverage decreases as social vulnerability increases. This points to ongoing challenges in reaching more vulnerable communities with vaccination efforts.



The county-level vaccination data further illustrate the wide range of public health outcomes across the United States. The ten counties with the highest completion rates—each reporting 95%—include Apache County (Arizona), Imperial County (California), and McKinley County (New Mexico), reflecting the upper end of national vaccination coverage. In contrast, counties such as Slope County (North Dakota), Cameron Parish (Louisiana), and McCone County (Montana) reported some of the lowest vaccination rates, ranging from just 11.3% to 22.4%. These examples highlight the extremes of the distribution and reinforce the importance of local-level analysis to understand disparities in pandemic response and vaccine uptake across the country.

Together, these variables provide a comprehensive picture of the local burden of COVID-19 and the behavioral and structural factors that may explain differences in outcomes across the United States.

Analysis

The regression analysis confirms that both mask usage and vaccination rates are significantly associated with lower COVID-19 death rates across U.S. counties in 2022. In Model 1, a 1 percentage point increase in the share of people who always wore masks is associated with a decrease of approximately 113 deaths per 100,000 residents, holding other factors constant. Model 2 shows that higher vaccination rates also have a protective effect: each 1 percentage point increase in vaccination completion is associated with a reduction of about 1.12 deaths per 100,000. Model 3 includes both predictors and finds that the associations remain statistically significant, though the magnitude of the mask coefficient decreases slightly to –83.3, and the vaccination effect remains strong at –0.92. Across all models, counties with higher social vulnerability (categories B, C, and D) experienced significantly higher death rates compared to the least vulnerable counties. These results highlight the importance of both individual protective behaviors and structural conditions in shaping the pandemic's impact.

Regression Models: Predicting COVID-19 Death Rates

part	term	statistic	Model 1: Mask Use Only	Model 2: Vaccination Only	Model 3: Both Mask Use & Vaccination
estimates	always.mask	estimate	-112.887***		-83.296***
estimates	always.mask	std.error	(10.212)		(10.834)
estimates	population	estimate	-0.000***	-0.000***	
estimates	population	std.error	(0.000)	(0.000)	
estimates	svi.indexB	estimate	11.415***	9.968***	9.581***
estimates	svi.indexB	std.error	(2.884)	(2.878)	(2.859)
estimates	svi.indexC	estimate	15.142***	14.399***	13.128***
estimates	svi.indexC	std.error	(3.111)	(3.103)	(3.080)
estimates	svi.indexD	estimate	19.811***	18.353***	17.491***
estimates	svi.indexD	std.error	(3.401)	(3.393)	(3.367)
estimates	vax.complete	estimate		-1.123***	-0.923***
estimates	vax.complete	std.error		(0.093)	(0.099)
gof	Num.Obs.		3049	3049	3049

Together, these variables provide a comprehensive picture of how behavioral and structural factors influenced the local burden of COVID-19 across U.S. counties. The wide variation in deaths, mask usage, and vaccination rates—especially when viewed alongside social vulnerability—reveals the multifaceted nature of the pandemic's impact and the importance of place-based strategies in public health response.

Appendix: Replication code

```
# This "setup" chunk specifies global options
# for handling code, plots, etc in your doc.
  knitr::opts_chunk$set(
    eval = TRUE,
    echo = FALSE,
    warning = FALSE,
    message = FALSE,
    fig.align = 'center'
  )

# Load packages
library(tidyverse)
library(scales)
library(lfe)
library(modelsummary)
library(gt)
library(data.table)
library(knitr)
library(kableExtra)
```

```r
# Load the merged COVID dataset
covid <- read_csv("merged_covid_data.csv")
# Appendix: Replication code

```{r ref.label=knitr::all_labels(), echo=TRUE, eval=FALSE}
```

4a. COVID Deaths Nationally
# Calculate the mean of deaths
mean_deaths <- mean(covid$deaths, na.rm = TRUE)

covid %>%
  ggplot(aes(x = 1 + deaths)) +
  geom_histogram(fill = "skyblue", color = "white", bins = 30) +
  scale_x_log10(labels = scales::comma) +
  labs(
    title = "Distribution of COVID-19 Deaths by County (Log Scale)",
    x = "Total COVID-19 Deaths (by County)",
    y = "Count of Counties"
) +
Minimal theme + no grid lines + centered, bold, dark title
theme_minimal() +
theme(
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  plot.title = element_text(hjust = 0.5, face = "bold", color = "black")
) +
Dashed dark blue line from y=0 to y=Inf at the mean
geom_segment(
  aes(
    x = 1 + mean_deaths,
    xend = 1 + mean_deaths,
    y = 0,
    yend = Inf
  ),
  color = "darkblue",
  linetype = "dashed",
  size = 1
) +
Annotate the mean line
annotate(
  "text",
  x = 1 + mean_deaths,
  y = Inf,
  label = paste("Mean =", round(mean_deaths, 1)),
  color = "darkblue",
  angle = 90,
  vjust = 2
)

4b. Mask Usage
# Calculate mean of mask usage (as a fraction)
mean_mask <- mean(covid$always.mask, na.rm = TRUE)
```

```r
covid %>%
  ggplot(aes(x = always.mask)) +
  geom_histogram(fill = "skyblue", color = "white", bins = 30) +
# Convert x-axis to percentage format
scale_x_continuous(labels = scales::percent_format(scale = 100)) +
labs(
  title = "Distribution of 'Always Mask' Use Across US Counties",
  x = "Percentage Always Masking",
  y = "Count of Counties"
) +
theme_minimal() +
theme(
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  plot.title = element_text(hjust = 0.5, face = "bold", color = "black")
) +
# Add a dashed darkblue mean line from y=0 to the top of the plot
geom_segment(
  data = data.frame(x = mean_mask),
  aes(x = x, xend = x, y = 0, yend = Inf),
  color = "darkblue",
  linetype = "dashed",
  size = 1,
  inherit.aes = FALSE
) +
# Annotate the mean line with a label
annotate(
  "text",
  x = mean_mask,
  y = Inf,
  label = paste("Mean =", round(mean_mask * 100, 1)),
  color = "darkblue",
  angle = 90,
  vjust = 2
)

# Calculate the mean of vaccination completion (data is already in percentage form)
mean_vax <- mean(covid$vax.complete, na.rm = TRUE)

covid %>%
  ggplot(aes(x = vax.complete)) +
  geom_histogram(fill = "skyblue", color = "white", bins = 30) +
# Convert x-axis to display a percent sign (e.g., 75 becomes 75%)
scale_x_continuous(labels = scales::percent_format(scale = 1)) +
labs(
  title = "Distribution of COVID-19 Vaccination Completion by County",
  x = "Vaccination Completion",
  y = "Count of Counties"
) +
theme_minimal() +
theme(
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
```

```
  plot.title = element_text(hjust = 0.5, face = "bold", color = "black")
) +
# Dashed dark blue line from y=0 to y=Inf at the mean
geom_segment(
  aes(
    x = mean_vax,
    xend = mean_vax,
    y = 0,
    yend = Inf
  ),
  color = "darkblue",
  linetype = "dashed",
  size = 1
) +
# Annotate the mean line with a label
annotate(
  "text",
  x = mean_vax,
  y = Inf,
  label = paste("Mean =", round(mean_vax, 1)),
  color = "darkblue",
  angle = 90,
  vjust = 2
)

# Vaccination rates by Social Vulnerability Index (SVI)
covid %>%
  filter(!is.na(svi.index)) %>%  # remove counties with missing SVI
  ggplot(aes(x = factor(svi.index), y = vax.complete)) +
  # Draw whiskers (error bars) first so they appear behind the boxes
  stat_boxplot(geom = "errorbar", width = 0.25) +
  # Draw boxplots on top
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(
    title = "Vaccination Rates by Social Vulnerability Index",
    x = "Social Vulnerability Index Category",
    y = "Vaccination Completion Percentage"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5, face = "bold", color = "black")
  )
# Run regression models
mods <- list(
  "Model 1: Mask Use Only" = felm(deaths.scaled ~ always.mask + population + svi.index | state, data = 
  "Model 2: Vaccination Only" = felm(deaths.scaled ~ vax.complete + population + svi.index | state, data
  "Model 3: Both Mask Use & Vaccination" = felm(deaths.scaled ~ always.mask + vax.complete + svi.index
)

# Get model summary as a data frame to avoid LaTeX floats
reg_table <- modelsummary(mods, output = "data.frame", stars = TRUE, gof_map = "nobs")
```

```r
reg_table %>%
  kable(format = "latex", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = c("hold_position", "striped", "scale_down"), font_size = 10) %>%
  row_spec(0, bold = TRUE)
```