STAT461 Course Project

Daniel Fitzgerald, Peter Phillips, Haichen Wei, Ailin Zhang

May 5, 2022

Executive Summary

There are many different factors that can determine whether or not someone may survive an infection with COVID-19. In our study, we took CDC data on COVID-19 deaths and compared the death totals to the age group of the case, whether or not that case was vaccinated to protect against severe COVID-19, and when that death took place, as we have seen throughout the pandemic the coronavirus comes in waves, and during those waves people have a higher chance of infection and/or death than in between those waves.

As a result of our study, we found all three factors contributed to the number of deaths in a given week, with vaccination status contributing the most to the variation in number of deaths. Age group also had a large effect, and the week the death occurred had the least effect on number of deaths, though it was still statistically significant.

Introduction

In 2020, a new coronavirus, COVID-19, spread out around the world. COVID-19 affects different people in different ways. Some people may have mild symptoms like fever or cough, and some may have difficulty breathing or turn into more serious illnesses.

According to the CDC Data Tracker, the United States has had 991,439 people died since COVID-19 first arrived in this country. Though the World Health Organization (WHO) indicated that anyone can get sick with COVID-19 and become seriously ill or die at any age, they also point out older people are more likely to develop serious illnesses. Thus, we want to see whether age group is an important factor related to deaths caused by COVID-19.

Since there are 82.5% of people older than 5 years old who have at least received one vaccination dose in the U.S., we can also explore the effectiveness of vaccines. Do vaccines indeed help to reduce death cases?

To this end, we want to explore the effect age, immunization status, and time has on COVID-19 related deaths.

Literature Review

We use *Rates of COVID-19 Cases or Deaths by Age Group and Vaccination Status* dataset[1] which was collected by the CDC COVID-19 Response, Epidemiology Task Force. Based on our research question, we did some data cleaning and reorganized the format of our data. We filtered and selected all the death outcomes, kept age group and week, and created a new column named Vaccination Status, and assigned case numbers to the Death column.

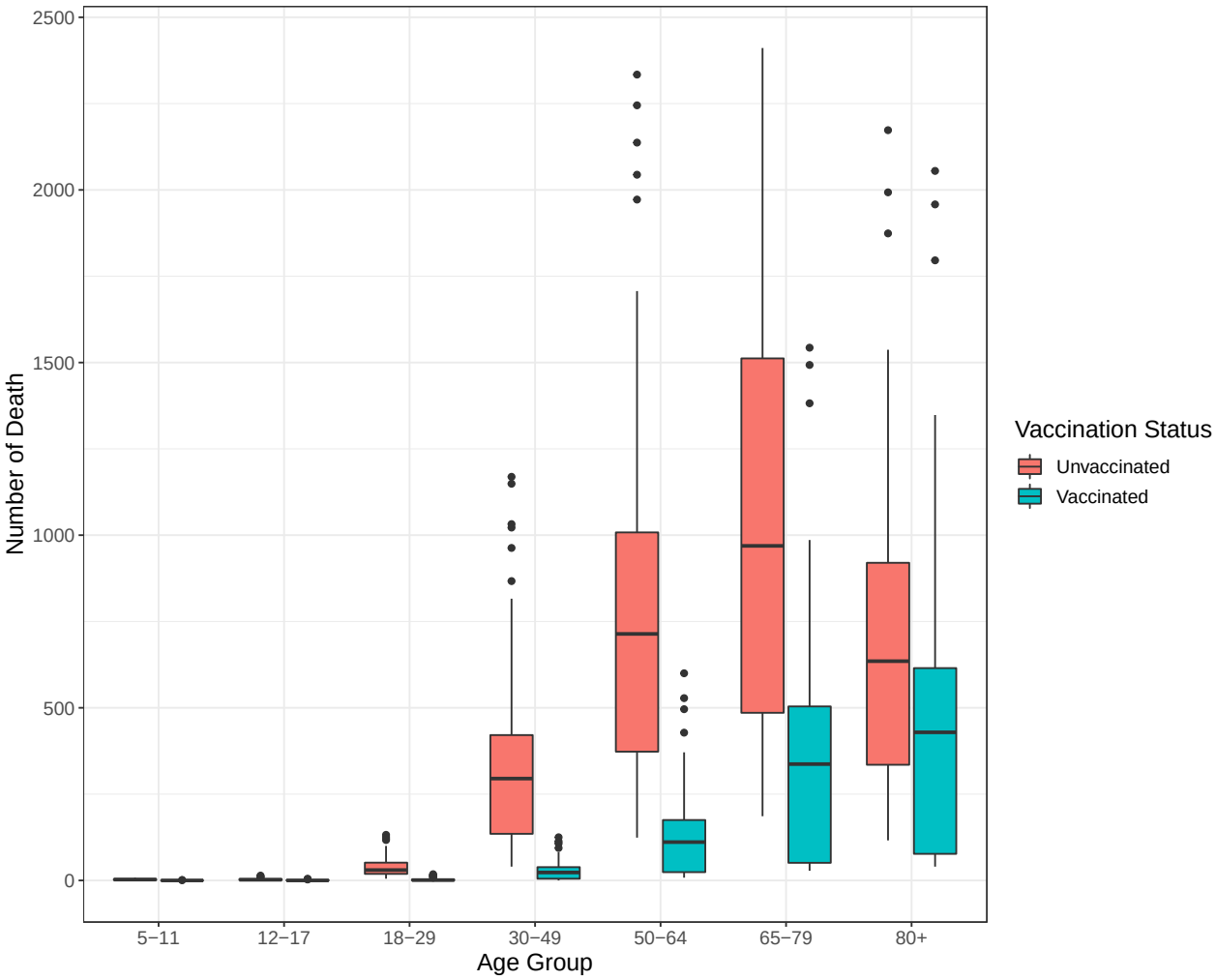


Figure 1: Box Plot With Age Group and Vaccination Status

Figure 1 provides the box plot with the age group and vaccination status. In examining the box plot, we can see there is rare death numbers for people under 30 regardless of vaccination status. For people above 30 years old, vaccination status starts to show its effect. Unvaccinated people have larger numbers of deaths than vaccinated in each age group. There are rare death numbers for vaccinated people in the 30-49 age group. Figure 1 obviously shows that the death cases increase as the age group increase.

[1]Data available at https://data.cdc.gov/Public-Health-Surveillance/Rates-of-COVID-19-Cases-or-Deaths-by-Age-Group-and/3rge-nu2a

Table 1: Summary Statistics for COVID-19 Deaths

Age Group	Vaccination Status	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
5-11	Unvaccinated	12	0	1.0	2.5	4.25	8	2.224	2.833	2.368	0.633	-0.584
5-11	Vaccinated	12	0	0.0	0.0	0.00	1	0.000	0.167	0.389	1.570	0.529
12-17	Unvaccinated	47	0	1.0	2.0	4.00	14	1.483	3.043	2.992	1.775	3.417
12-17	Vaccinated	47	0	0.0	0.0	1.00	5	0.000	0.489	0.997	2.668	7.783
18-29	Unvaccinated	47	5	19.0	30.0	51.50	132	23.722	43.340	35.057	1.143	0.186
18-29	Vaccinated	47	0	0.0	1.0	3.00	18	1.483	2.745	4.336	2.064	3.418
30-49	Unvaccinated	47	40	135.0	295.0	421.00	1,169	235.733	362.383	307.743	1.284	0.623
30-49	Vaccinated	47	0	5.0	23.0	38.50	125	23.722	29.638	30.312	1.558	2.037
50-64	Unvaccinated	47	124	373.0	714.0	1,008.00	2,334	456.641	821.553	593.727	1.043	0.255
50-64	Vaccinated	47	8	24.0	111.0	175.00	600	106.747	141.128	140.630	1.606	2.224
65-79	Unvaccinated	47	186	485.5	969.0	1,512.00	2,411	797.639	1,057.830	666.366	0.503	-0.946
65-79	Vaccinated	47	28	51.0	337.0	504.00	1,543	367.685	398.894	375.708	1.442	1.885
80+	Unvaccinated	47	116	335.0	635.0	920.00	2,173	465.536	708.468	499.943	1.083	0.776
80+	Vaccinated	47	40	77.0	429.0	615.00	2,055	406.232	498.191	488.344	1.637	2.483

Table 1 shows the value of various descriptive statistics broken out by age group and vaccination status. Visually, we can see that there are less than 20 deaths in each age group under 18. For the age group 18-29, deaths of vaccinated people are 18, but deaths of unvaccinated people are 132 which is even more than the deaths of vaccinated people in the 30-49 age group. We can also see this in the bar graph of Figure 2. For people above 80 years old, the vaccines seem not to have a significant effect as other age groups.

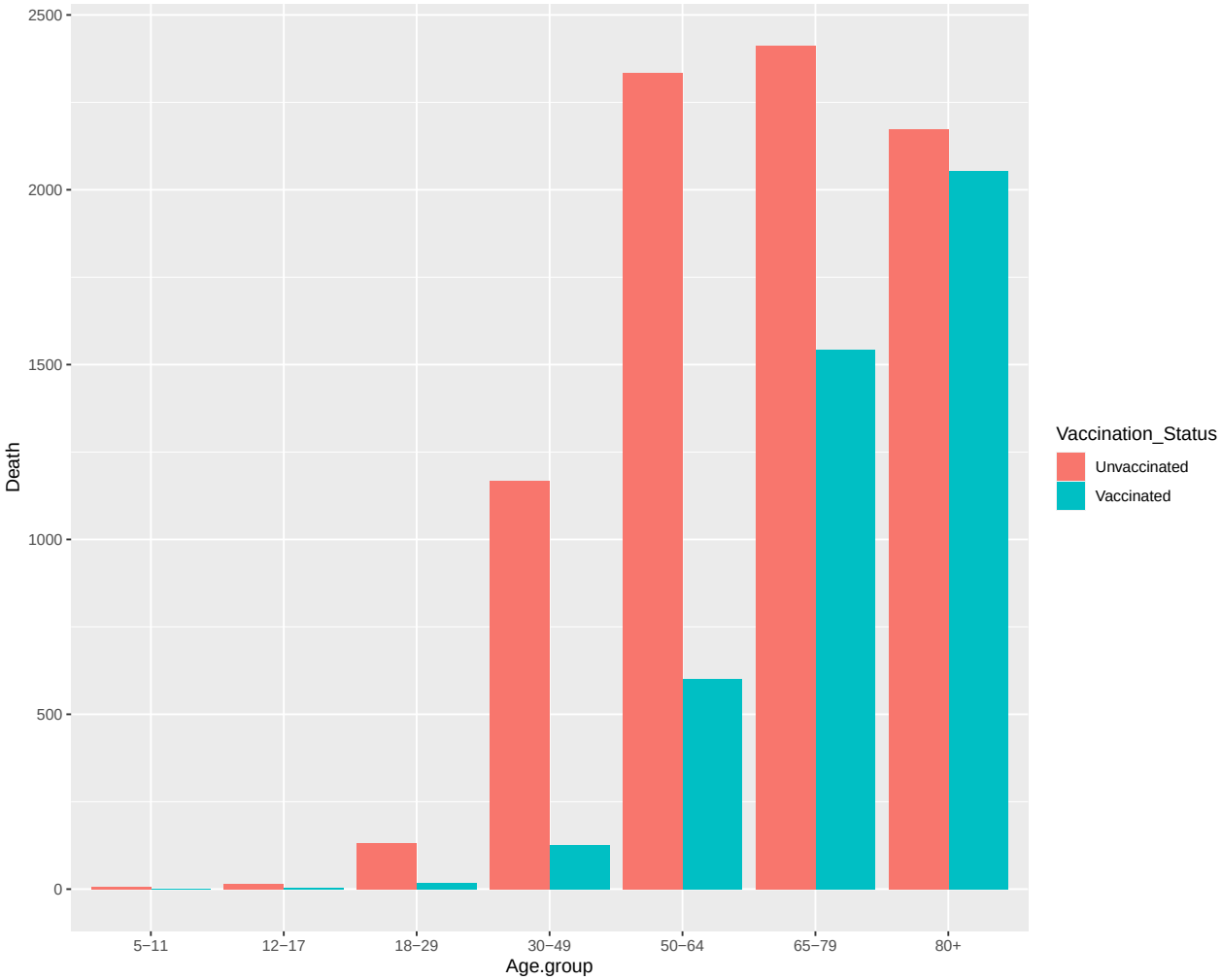


Figure 2: Bar Graph of Death Numbers in Various Age Groups

Methods

In order to assess our SRQ, we have determined which factors from the CDC data set will have an impact on the primary response, total number of deaths. We see in Figure 3 that Vaccination Status (vaccinated or not), age group, and point in time factor into how many deaths there are. Age and Vaccination Status also have an interaction with each other, and time point is independent from those two factors.

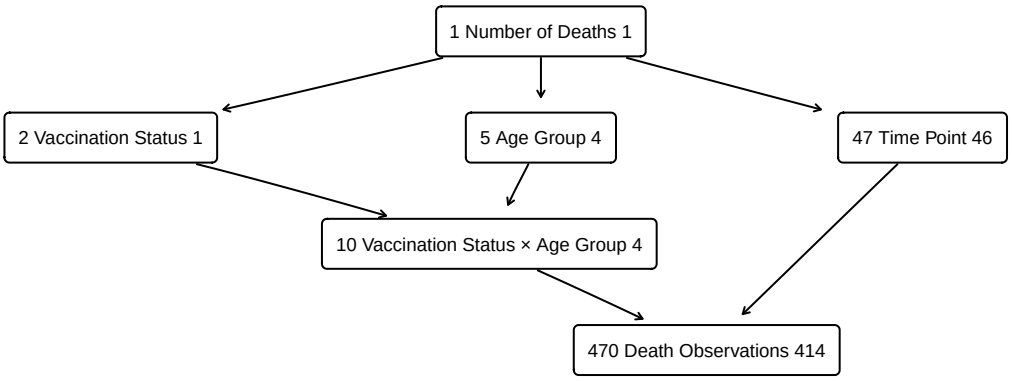


Figure 3: Hasse Diagram for COVID-19 Deaths Study

Based off the outline of our study, ANOVA models appear to be appropriate. We will conduct our study using an ANOVA F-test: Factorial Design Model, which allows us to compare each individual factor's impact on total deaths, and in this case, how the interaction between Vaccination Status and Age contribute to total deaths.

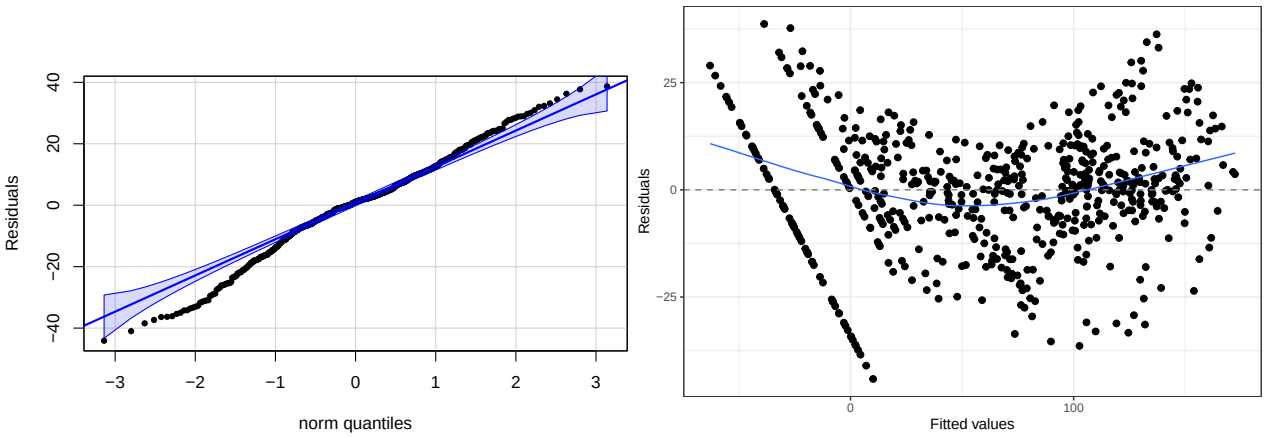
For our study, we will control our overall Type I risk at 15%, and use a personal unusualness threshold of 10 percent. Any factor (or interaction between two) within our unusualness threshold will be considered statistically significant as having an impact on number of deaths.

Results

To answer our SRQ, we will seek to use the parametric shortcut known as the ANOVA F test. There are three assumptions that our data must satisfy to use this approach: residuals must follow a Gaussian distribution, homoscedasticity, and independence of observations.

Assumptions

When we ran our assumptions, we needed to perform a transformation to our response of adding .0001 and performing a "Box-Cox" transformation. Our data appeared to have a skew due to the presence of zeros for the 5-11 and 12-17 categories in terms of death. The Box-Cox transformation fixes the problem of having many zeroes in our response.



(a) Gaussian QQ Plot

(b) Tukey-Anscombe Plot

Figure 4: Assessing Assumptions for COVID-19 Deaths Study

Per Figure 4, after our transformation, it appears we have a reasonable Gaussian distribution, at worst "questionable". The Tukey-Anscombe plot also is quite questionable, but the blue line is relatively straight and not performing against expectations consistent with our data.

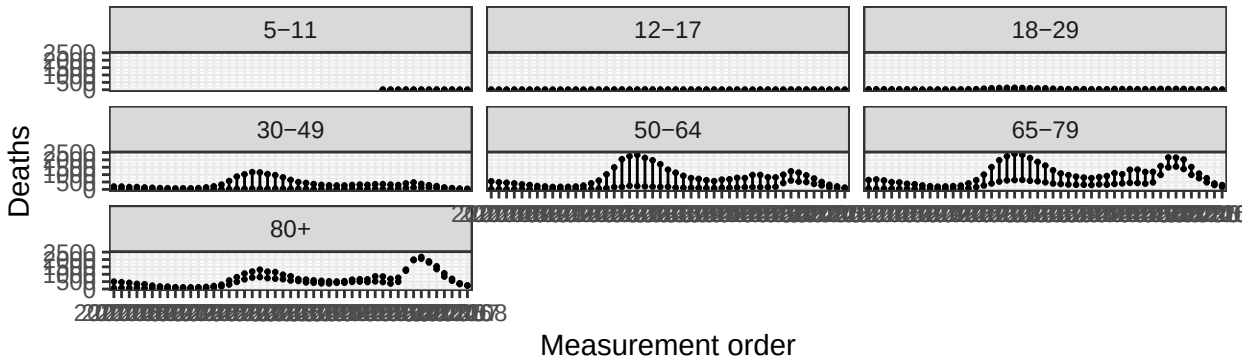


Figure 5: Independence of Observations for COVID-19 Deaths Study

Since we know the order of observations a la knowing the MMWR of the data, we can also form an "independence of observations chart", shown as Figure 5 which is split into separate age groups. As expected, the 5-11 and 12-17 groups seem to make our data lopsided, but given a larger sample size, it would be possible that those groups would conform with the rest. Thus, we can assume that the independence of observations is satisfied.

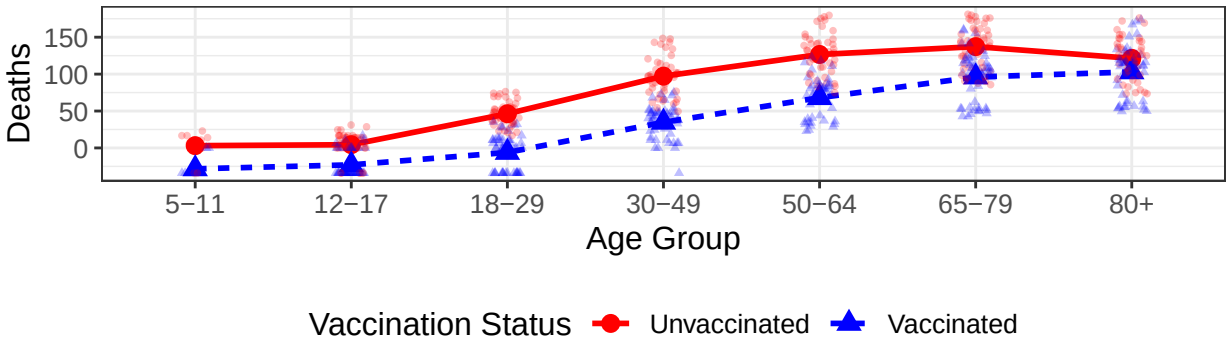


Figure 6: Independence of Observations for COVID-19 Deaths Study

Finally, when looking at our interactions plot in Figure 6, we can see that there is consistency between the two factors age group and vaccination status, meaning that there seems to be no confounding. While we also included time (MMWR) as a factor in our assessment, we are treating it as it does not interact with the other factors, thus it is not included in the interaction plot. This assumption is also satisfied.

After consideration, we can put the assumptions into context and assume that overall, they are satisfied, which will allow us to perform the ANOVA F-shortcut: Factorial Designs Model.

Omnibus

Table 2: ANOVA Table for COVID-19 Deaths Study

Source	SS	df	MS	F	p-value	Partial Omega Sq.	Partial Eta Sq.	Partial Epsilon Sq.
Age.group	1409801.16	6	234966.8600	1061.5290	< 0.0001	0.9154	0.9234	0.9226
Vaccination_Status	270487.20	1	270487.1959	1222.0021	< 0.0001	0.6750	0.6983	0.6977
MMWR.week	258480.03	46	5619.1310	25.3860	< 0.0001	0.6561	0.6886	0.6615
Age.group:Vaccination_Status	37656.14	6	6276.0226	28.3537	< 0.0001	0.2182	0.2437	0.2351
Residuals	116871.51	528	221.3476					

Per Table 2, we would decide to reject the null hypothesis for each of the main factors and for the interaction term, as all are statistically significant for having an impact on the total number of COVID-19 related deaths. Of these, Vaccination Status appears to explain most of the variation in death rate.

Marginal Means

Table 3: Marginal Means-Tukey 85% Adjustment

Age group	Vaccine status	Marginal Mean	SE	DF	Lower Bound	Upper Bound
5-11	Unvaccinated	-10.7280	4.4261	528	-17.1088	-4.3473
12-17	Unvaccinated	4.3819	2.1701	528	1.2534	7.5105
18-29	Unvaccinated	46.3267	2.1701	528	43.1982	49.4552
30-49	Unvaccinated	97.2266	2.1701	528	94.0981	100.3552
50-64	Unvaccinated	126.5890	2.1701	528	123.4604	129.7175
65-79	Unvaccinated	137.2292	2.1701	528	134.1007	140.3578
80+	Unvaccinated	121.3499	2.1701	528	118.2214	124.4785
5-11	Vaccinated	-42.1400	4.4261	528	-48.5208	-35.7592
12-17	Vaccinated	-22.9027	2.1701	528	-26.0313	-19.7742
18-29	Vaccinated	-6.1185	2.1701	528	-9.2470	-2.9899
30-49	Vaccinated	34.7227	2.1701	528	31.5942	37.8513
50-64	Vaccinated	68.0553	2.1701	528	64.9267	71.1838
65-79	Vaccinated	96.0533	2.1701	528	92.9247	99.1818
80+	Vaccinated	102.9866	2.1701	528	99.8580	106.1151

From the Table 3, we can see that factor level of age groups effects estimates. At level age at 65-79 accumulated death at a rate of 1057 times. Age above 80 was 708 times related to death rate.

Discussion

Our research question was the effect of age and the vaccinated status on covid mortality. From the data of our study, we can conclude that the highest mortality rate was observed in the age group 65-79 years among those who were not vaccinated.

Our study has two limitations. First, the data do not include children under 5 years of age, which would make our data incomplete. We could break down the age and count the prevalence in children under 5 years of age. Second, the data do not include vaccine types. In a future improvement, we could compare mortality rates based on the vaccination population for different types of vaccines. This will give us a more accurate picture of the relationship between vaccine status and mortality.

References

COVID Data Tracker. *Trends in Number of COVID-19 Cases and Deaths in the US Reported to CDC, by State/Territory*. Centers for Disease Control and Prevention. Retrieved from https://covid.cdc.gov/covid-data-tracker/#trends_dailydeaths.

Coronavirus disease (COVID-19). World Health Organization. Retrieved from https://www.who.int/health-topics/coronavirus#tab=tab_3.

CDC COVID-19 Response, Epidemiology Task Force (April 15, 2022). *Rates of COVID-19 Cases or Deaths by Age Group and Vaccination Status*. Centers for Disease Control and Prevention. Retrieved from https://data.cdc.gov/Public-Health-Surveillance/Rates-of-COVID-19-Cases-or-Deaths-by-Age-Group-and/3rge-nu2a.

Author Contributions

The authors of this report would like to acknowledge their individual contributions to the report.

- Peter contributed to the executive summary, methods, and general organization of the report.
- Haichen contributed to data wrangling, introduction and literature review.
- Daniel contributed to the assumptions and results of our data.
- Ailin contributed to results, marginal means, and discussion.

Code Appendix

```
rm(list = ls())
# Setting Document Options
knitr::opts_chunk$set(
  echo = FALSE,
  warning = FALSE,
  message = FALSE,
  fig.align = "center"
)

## Load packages
packages <- c("tidyverse", "knitr", "kableExtra",
              "parameters", "hasseDiagram", "car",
              "psych", "emmeans", "rstatix", "lme4", "nlme")
lapply(packages, library, character.only = TRUE)
options(knitr.kable.NA = "")
options(contrasts = c("contr.sum", "contr.poly"))
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")


data <- read.csv(
  file = "/Users/Zoey/Desktop/Rates_of_COVID-19_Cases_or_Deaths_by_Age_Group_and_Vaccination_Status.csv
  header = TRUE,
  sep = ","
)

#data cleaning
data <- filter(data, outcome == "death")
data <- data[, c('Age.group', 'MMWR.week', 'Vaccinated.with.outcome', 'Unvaccinated.with.outcome')]

data <- subset(data, Age.group != "all_ages_adj")

data <- data[rep(seq_len(nrow(data)), each = 2), ]
data$Vaccination_Status <- rep(c("Vaccinated", "Unvaccinated"), length.out=nrow(data))

data$Death <- ifelse(data$Vaccination_Status == "Vaccinated", data$Vaccinated.with.outcome, data$Unvacci
data <- data[, c('Age.group', 'MMWR.week', 'Vaccination_Status', 'Death')]

data$Age.group <- factor(data$Age.group , levels=c("5-11", "12-17", "18-29", "30-49", "50-64","65-79","8

data$MMWR.week <- as.factor(data$MMWR.week)

data$Vaccination_Status <- as.factor(data$Vaccination_Status)

data <- na.omit(data)

# Box Cox Transformation
data$tDeath <- data$Death + 0.0001
SGM <- psych::geometric.mean(data$tDeath)
data <- data %>%
  mutate(
    logDeath = log(tDeath),
```

```
    bLogDeath = SGM*log(tDeath),
    bcDeath = (tDeath^0.2222 - 1) /(0.2222 * SGM^(0.2222-1))
  )


ggplot(
  data = data,
  mapping = aes(
    x = Age.group,
    y = Death,
    fill = Vaccination_Status
) )+
  geom_boxplot() +
  theme_bw() +
  xlab("Age Group") +
  ylab("Number of Death") +
  labs(
fill = "Vaccination Status" )+
  theme(
    legend.position = "right",
    text = element_text(size = 14)
)

data %>%
  dplyr::group_by(Age.group, Vaccination_Status) %>%
  summarize(
    n = n(),
    min = min(Death),
    Q1 = quantile(Death, probs = c(0.25)),
    med = median(Death),
    Q3 = quantile(Death, probs = c(0.75)),
    max = max(Death),
    mad = mad(Death),
    sam = mean(Death),
    sd = sd(Death),
    skew = psych::skew(Death),
    kurtosis = psych::kurtosi(Death)
) %>%
knitr::kable(
  caption = "Summary Statistics for COVID-19 Deaths",
  digits = 3,
  format.args = list(big.mark = ","),
  align = rep('c', 13),
  col.names = c("Age Group", "Vaccination Status", "n", "Min", "Q1", "Median", "Q3", "Max", "MAD",
                "SAM", "SASD", "Sample Skew","Sample Ex. Kurtosis"),
  booktabs = TRUE
)  %>%
kableExtra::kable_styling(
  font_size = 12,
  latex_options = c("HOLD_position", "scale_down")
)
```

```r
ggplot(data, aes(x = Age.group, y = Death, fill = Vaccination_Status)) +
  geom_col(position = "dodge")

modelLabels <- c("1 Number of Deaths 1", "2 Vaccination Status 1", "5 Age Group 4", "47 Time Point 46",
modelMatrix <- matrix(
  data = c(FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALS
  nrow = 6,
  ncol = 6,
  byrow = FALSE
)
hasseDiagram::hasse(
 data = modelMatrix,
 labels = modelLabels
)
covidModelt <- aov(
  formula = bcDeath ~ Age.group*Vaccination_Status + MMWR.week,
  data = data
)
car::qqPlot(
  x = residuals(covidModelt),
  distribution = "norm",
  envelope = 0.85,
  id = FALSE,
  pch = 20,
  ylab = "Residuals"
)

ggplot(
  data = data.frame(
    residuals = residuals(covidModelt),
    fitted = fitted.values(covidModelt)
  ),
  mapping = aes(x = fitted, y = residuals)
) +
  geom_point(size = 2) +
  geom_hline(
    yintercept = 0,
    linetype = "dashed",
    color = "grey50"
  ) +
  geom_smooth(
    formula = y ~ x,
    method = stats::loess,
    method.args = list(degree = 1),
    se = FALSE,
    size = 0.5
  ) +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Residuals")
ggplot(
  data = data,
  mapping = aes(
```

```
    x = MMWR.week,
    y = Death
  )
) +
  geom_point(size = 0.5) +
  geom_line() +
  theme_bw() +
  xlab("Measurement order") +
  ylab("Deaths") +
  facet_wrap(
    . ~ Age.group
  )
ggplot(
  data = data,
  mapping = aes(
    x = Age.group,
    y = bcDeath,
    shape = Vaccination_Status,
    color = Vaccination_Status,
    linetype = Vaccination_Status,
    group = Vaccination_Status
  )
) +
  stat_summary(fun = "mean", geom = "point", size = 3) +
  stat_summary(fun = "mean", geom = "line", size = 1) +
  geom_jitter(width = 0.1, height = 0.1, alpha = 0.25, size = 1) +
  ggplot2::theme_bw() +
  xlab("Age Group") +
  ylab("Deaths") +
  labs(
    color = "Vaccination Status",
    shape = "Vaccination Status",
    linetype = "Vaccination Status"
  ) +
  scale_color_manual(values = c("red", "blue")) +
  theme(
    legend.position = "bottom",
    text = element_text(size = 12)
  )
parameters::model_parameters(
  model = covidModelt,
  omega_squared = "partial", # Notice the use of partial
  eta_squared = "partial",
  epsilon_squared = "partial",
  type = 1, # Use 1, 2, or 3 for the Type of SSQs you want
  drop = "(Intercept)", # Drop an unneeded row for ANOVA
  verbose = FALSE # Makes the function "quiet"
) %>%
  dplyr::mutate(
    p = ifelse(
      test = is.na(p),
      yes = NA,
      no = pvalRound(p)
```

```
    )
  ) %>%
  knitr::kable(
    digits = 4,
    col.names = c("Source", "SS", "df", "MS", "F", "p-value",
                  "Partial Omega Sq.", "Partial Eta Sq.", "Partial Epsilon Sq."),
    caption = "ANOVA Table for COVID-19 Deaths Study",
    align = c('l',rep('c',8)),
    booktab = TRUE
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
)

covidPHMeans <- emmeans::emmeans(
  object = covidModelt,
  # The order of factors does not really matter for this
  specs = pairwise ~ Age.group | Vaccination_Status,
  adjust = "tukey", # Where you specify your chosen method
  level = 0.85 # 1--Type I Risk
)
as.data.frame(covidPHMeans$emmeans) %>%
  knitr::kable(
    digits = 4,
    col.names = c("Age group", "Vaccine status", "Marginal Mean","SE", "DF",
                  "Lower Bound","Upper Bound"),
    caption = "Marginal Means-Tukey 85\\% Adjustment",
    format = "latex",
    align = rep("c", 7),
    booktabs = TRUE
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12,
    latex_options = c("HOLD_position")
)
```