Global Covid Death Trends

Lendel Deguia,
Caitlyn Nguyen, Hamza Javed

November 2022

Contents

1 Introduction

This project was created with the intent to analyze trends and correlations between various external factors and deaths due to COVID. R, alongside its many packages, were used to clean the data and analyze these trends. Our primary data set is titled "Our World in Data, Covid-19" acquired through OWID's Github COVID data repository [1]. The complete dataset contains observations from January 22, 2020 through September 27, 2022 from various countries throughout the world amounting to 219,637 observations with each observation described by 67 variables. The variables primarily focused on in this report include: location, demographics, confirmed cases, confirmed deaths, vaccinations, GDP per capita, and the date of the observations.

From the earlier half of 2020 with a lingering effect on present day, the COVID-19 pandemic has negatively affected the world on a global level and disrupted the lives of many individuals. The death of millions of people has brought about significant change to society on a health, economic, and daily livelihood scale leading us to find interest in the statistics of such a large-scale disease outbreak. With an interest on the impact of the recent pandemic on the world, our report aims to investigate the effects of geographical region, socioeconomic status, and vaccination status on mortality due to COVID; our target variables in this regard are cumulative deaths with respect to time and/or region, the proportion of total deaths to total cases within a certain region and a given time frame, and the average daily deaths in a certain region over a given time frame that at least amounts to a month.

In order to investigate the relationship between a region's socioeconomic status and COVID mortality, we will focus on a country's GDP per capita as this is a suitable metric indicative of a country's standard of living. OWID's dataset provides precisely what we need in this case: it contains the feature variable `gdp_per_capita`. Although OWID provides regional data, the dataset only provides data on a national scale. One of our objectives is to investigate the relationship between COVID deaths and geographical region; we are also interested in data for each state in the continental United States. We turn to JHU [2] for this information; JHU provides regional data on both a national and state/provincial scale regarding total deaths and total cases which will also be useful for evaluating the proportion of deaths to cases.

Moreover, OWID does contain vaccine data; however, we are interested in the effects of vaccination status on COVID related deaths. In order to accomplish this, we would need data that provides information on how many deaths are associated with a particular vaccination category: unvaccinated, regular two doses, or regular doses plus one or more boosters. OWID does not provide such information, so we turn to the CDC [3] for this kind of data. Unfortunately, we decided to limit our scope of analysis to the United States; there appears to be no cohesive vaccination data of the like on a global scale. Various countries provided their own individual reports, but reporting methodologies varied widely.

2 Data Cleaning

2.1 General Data Cleaning Approaches

As can be seen repeatedly in Appendix A, our method of filtering NA values for all datasets is something along the lines of the following process:

```
some_dataset <- read.csv('some_dataset.csv')

filtered_dataset <- some_dataset[
  -which(is.na(some_dataset$target_column_to_filter)), ]
```

This method was also used for further derived datasets from the initial filtered dataset. Another notable method was how we subsetted data when multiple instances of a categorical variable had different numerical values for the same category value; for instance, United States has different reported new daily deaths for each date in the OWID dataset. We would also extract other variables of interest with values that remained constant with respect to each value of the categorical variable. All of these values would be appended to an initially empty vector for each unique category and then used to construct a new dataframe. In this case, we used the following generic code block:

```
# initiate empty vectors to store desired quantities in
avg_num_var <- c()

constant_var_1 <- c()
          .
          .
          .
constant_var_n <- c()

# initiate counter so categorical variable can be iterated over directly
counter <- 1
for (category in unique(some_dataset$target_categorical_variable)){
  # append the mean of numerical_variable for a given category to avg_num_var
  avg_num_var[counter] <- mean(some_dataset[
    some_dataset$target_categorical_variable == category,
    "numerical_variable"])

  # append other desired variables with a
  # constant value with respect to category
  constant_var_1[counter] <- some_dataset[
    some_dataset$target_categorical_variable == category,
    "constant_variable_1"]
                    .
                    .
                    .
  constant_var_n[counter] <- some_dataset[
    some_dataset$target_categorical_variable == category,
    "constant_variable_1"]
  # increment the counter
  counter <- counter + 1
}

# generate new dataframe using newly filled vectors
# (column names may be changed here)
new_dataframe <- data.frame(avg_num_var = avg_num_var,
                            constant_var_1 = constant_var_1,
                                          .
                                          .
                                          .
                            constant_var_n = constant_var_n)
```

Other forms of data cleaning done primarily consisted of multiple instances of subsetting data, appending new columns, and filtering out NA values either using the above described methods or something similar (perhaps comparatively trivial). More details can be found in Appendix A.

2.2 The JHU Dataset

Perhaps a considerable amount of effort towards data cleaning went towards dealing with the JHU dataset. Our objective with the JHU dataset was to utilize their regional data for total deaths and total cases in each country.

In order to map data from the JHU dataset onto a world map, we used the "world" map dataset using the map_data() command from ggplot2. Unfortunately, several country names in the "world" dataset were listed differently than the country names in the JHU dataset; for instance, JHU listed “Burma” whereas "world" listed “Myanmar”. Considering it would be tremendously tedious to manually match each country name, we tried to match as many names as possible with the intersect() command. Fortunately 178 countries had matching names between the two datasets. Manual work led us to match 18 more countries bringing us to a total of 196.

Each of the unmatched country names had to be altered manually. Since "world" contained all the latitude and longitude data to generate a map, we decided to do the alterations on the JHU dataset. The JHU dataset had 23 unmatched country names, so we primarily focused on those. It turns out that four of the unmatched names were not even countries; two of them were the cruises “Diamond Princess” and “MS Zaandam” while the other two were “Summer Olympics 2020” and “Winter Olympics 2022”. Additionally, there did not appear to be any corresponding names in the "world" dataset for the country “Tuvalu” found in the JHU dataset. In the end, we decided to leave Tuvalu out of the analysis along with other unmatched countries.

After matching as many country names as possible, the total_cases and total_cases data from JHU were appended to the "world" dataset for each matched country. These values were also used to derive the values in the added cf_ratios column. Unmatched countries were given values of 0 for the total_cases, total_cases, and cf_ratios columns. Filtering the country names was the most work that had to be done in regards to data preparation for this project.

Furthermore, a similar approach had to be taken to match US state names from JHU to the state names in the "state" map dataset from ggplot2; fortunately, the only difference was that states names from the JHU dataset had the first letter capitalized whereas state names from the "state" dataset did not.

3 Questions and Findings

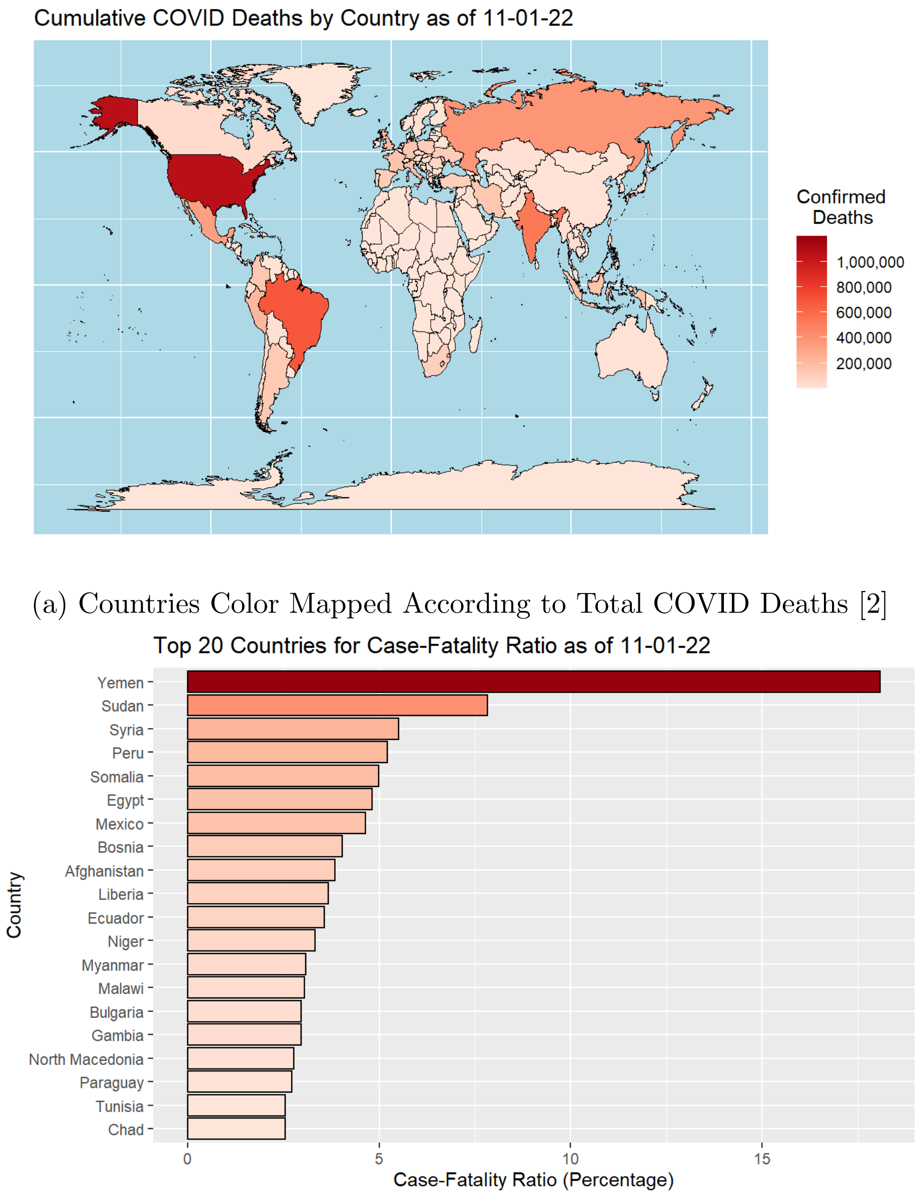
3.1 Total COVID Deaths and Deaths to Cases Ratio

Figure 1a depicts the total COVID deaths around the world as of November 01, 2022. According to the acquired data from JHU, the United States with a population of approximately 330 million has the most cumulative deaths around the world with 1,070,948 total deaths due to COVID; second highest is Brazil with 688,219 deaths, third highest is India with 530,452 deaths, and fourth highest is Russia with 382,372 deaths. This is a fairly suspect result considering, for example, China has a population of approximately 1.4 billion, a population of a little over 4 times that of the United States, yet has reported a total of 15,642 total deaths due to COVID (almost 1.5 percent the total deaths of the United States). This observation can be the result of either under-reporting or very effective efforts from China to combat the virus.

However, the total deaths is not all there is to consider. In figure 1b, countries are ranked according to their case-fatility ratio displayed as a percentage (we refer to it as “Deaths to Cases Ratio” as that describes the quantity better). The Deaths to Cases ratio is the total number of deaths divided by the total number of cases for a particular region; i.e. it is the proportion of COVID cases that result in death for a region of interest. Apparently, the United States does not even rank in the top 20 countries according to deaths to cases ratio — neither does Brazil, India, or Russia.

It turns out we see almost an entirely different worldwide trend with deaths to cases ratio compared to the trend for total deaths. Here, the top 20 countries according to deaths to cases ratio are primarily located in the Middle East and Africa. Yemen in particular has been reported to be facing a humanitarian crisis which might explain its high deaths to cases ratio since this could be due to a higher lack of healthcare resources compared to other countries. An additional point to make is that North Korea was not considered here since it yielded a nonsensical deaths to cases ratio of 600 percent; this would mean that there were somehow more deaths due to COVID than there were cases. An easy conclusion to make from this is that COVID numbers are severely under-reported in North Korea; any other interpretation (whatever that would be) is left as an exercise to the reader.

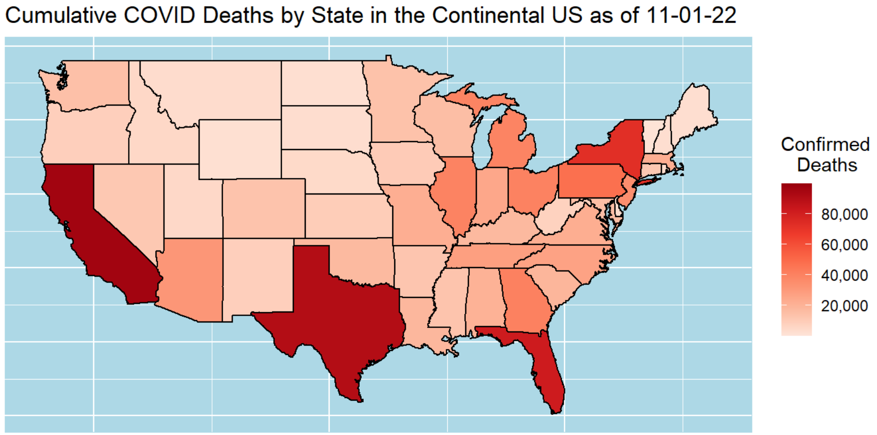
We made a similar observation in regards to different trends between total deaths and deaths to cases ratios when only considering the continental United States. The highest number of total deaths due to COVID among the US states considered comes from California which yields 96,896 total deaths; the second highest is Texas with 91,178 total deaths due to COVID, the third highest is Florida with 82,176 total deaths due to COVID, and the fourth highest is New York state with 73,186 total deaths due to COVID. Again, just as on a global scale, the top four states according to total COVID deaths do not appear in the top 20 states according to deaths to cases ratio.



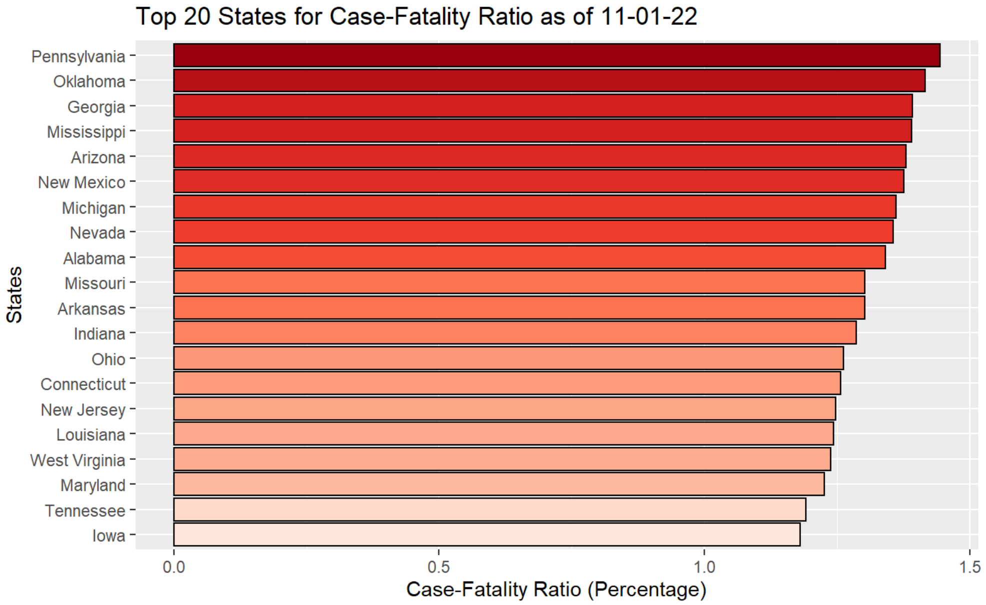
(a) Countries Color Mapped According to Total COVID Deaths [2]

(b) Top 20 Countries Ranked According to Case-Fatality Ratio [2]

Figure 1: Global COVID Death Trends


Cumulative COVID Deaths by State in the Continental US as of 11-01-22
Confirmed Deaths
80,000
60,000
40,000
20,000

(a) US States Color Mapped According to Total COVID Deaths [2]



(b) Top 20 Continental US States Ranked According to Case-Fatality Ratio [2]

Figure 2: Continental US COVID Death Trends

3.2 GDP Per Capita vs. Daily COVID Deaths

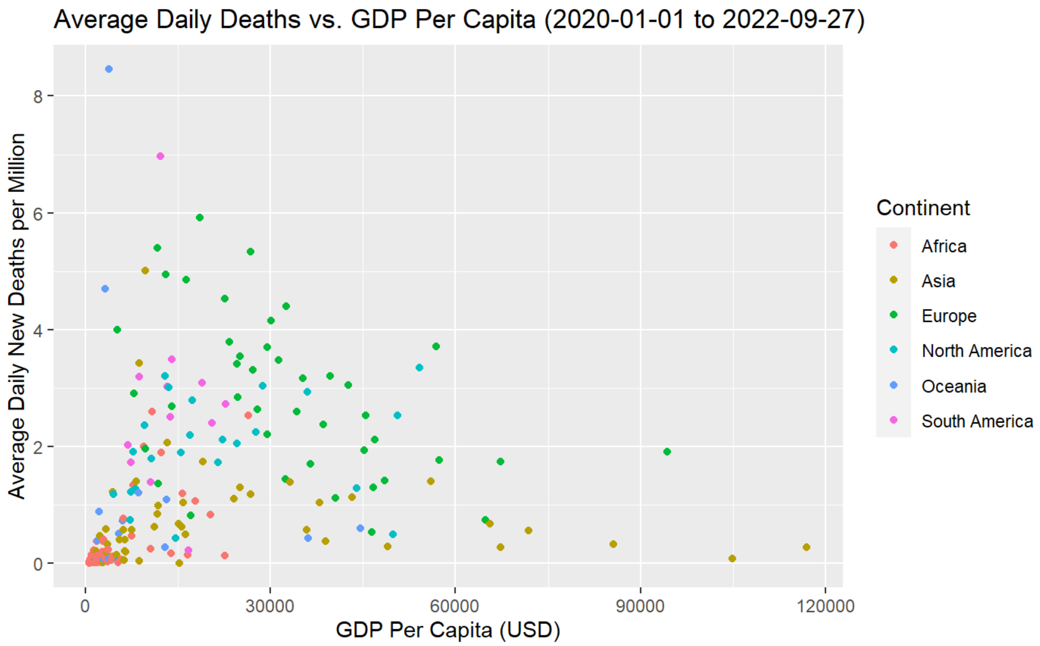


Figure 3: Average Daily Deaths vs. GDP Per Capita Colored by Continent

Our next primary question considered how a country's wealth affected its COVID daily death rate. Here, we took the average daily deaths for each country and plotted it against its GDP per capita via the method described in Section 2 and detailed in Appendix A. Unfortunately, according to Figure 3, we see that deducing a relationship is hard to infer from the acquired plot. However, notice that in Figure 3, there is a cluster of data points packed tightly in the bottom-left corner of the figure. This indicates that our GDP per capita data and average daily deaths data are skewed.

Indeed, when applying the following equation for the method of moments estimator for skewness

$$\gamma = \sqrt{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

we find that for GDP per capita, $\gamma \approx 1.887$ and for the average daily COVID deaths, $\gamma \approx 1.442$; here, for each variable, the x_i are the data points, $\bar{x}$ is the sample mean, and n is the number of data points. This indicates that GDP per capita and average daily deaths are highly skewed. Thus, a log transform of both variables will reveal more information about their relationship.

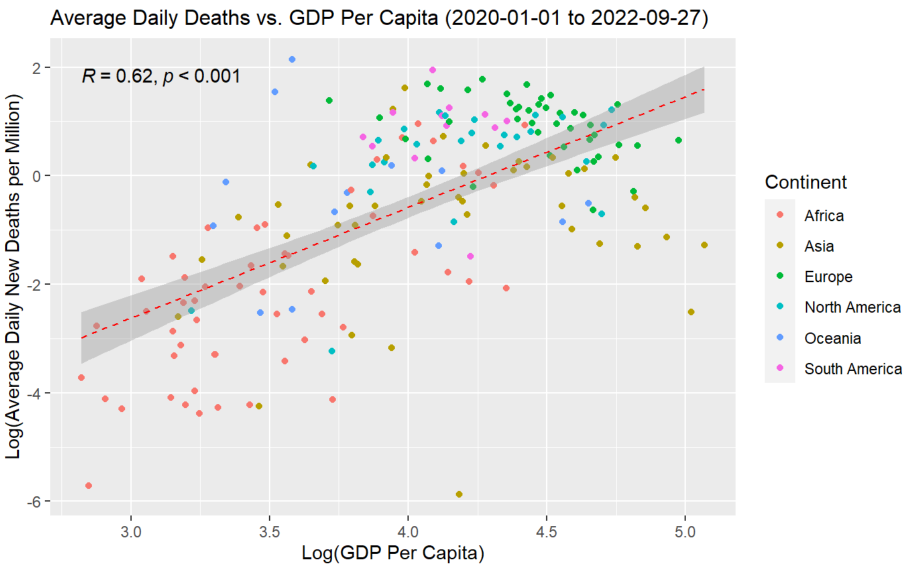


Figure 4: Log Transformed Average Daily Deaths vs. GDP Per Capita Colored by Continent

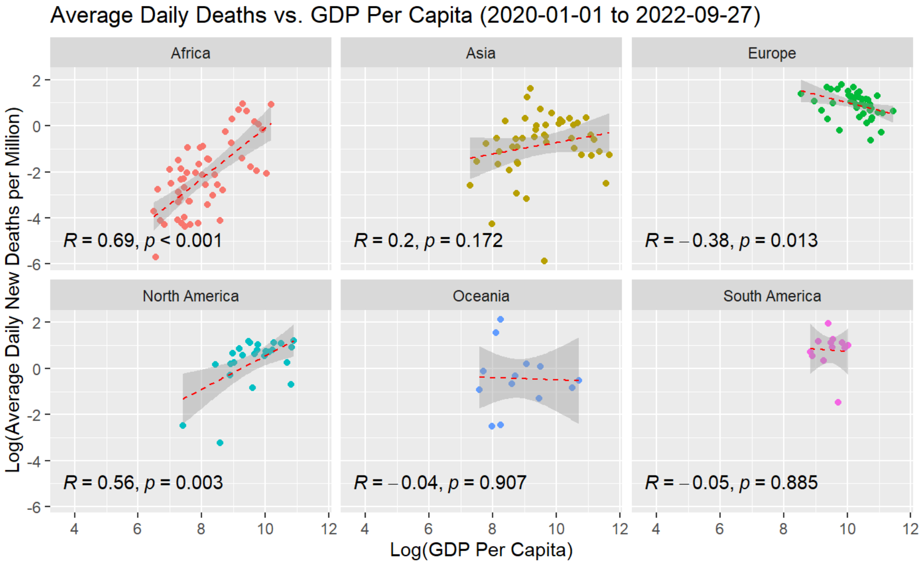


Figure 5: Figure 4 Facet Wrapped by Continent

It turns out that on a global scale, under a log transform, the average daily deaths and GDP per capita yield a positive correlation ($R = 0.62$) with extremely high statistical significance ($p < 0.001$). Globally, we see a positive trend between average daily deaths and GDP per capita which means that richer countries yield higher average daily deaths. One might think that the trend should be the opposite scenario where poorer countries yield higher average daily COVID deaths considering our observations according to deaths to cases ratios, but our observation here is likely due to the fact that richer countries have higher populations. Additionally, poorer countries may not have the proper resources for adequate reporting, and the corresponding populace might have stronger immune systems considering they may be subjected to less sanitary environments.

Furthermore, when accounting for the same relationship among different continents (Figure 5), we observe something different. The positive trend still holds for Africa ($R = 0.69$) and extremely high statistical significance ($p < 0.001$) and for North America ($R = 0.56$) with high statistical significance ($p = 0.003$). However, Europe yields the opposite trend ($R = -0.38$) with moderate statistical significance (p = 0.013). Also, the trend is inconclusive for Oceania, South America, and Asia according to the high p values (0.907, 0.885, and 0.172 respectively); considering the confidence intervals surrounding the trend lines for each of those continents, the relatively low correlations may likely be due to the wider variation of countries in regards to wealth in those continents.

3.3 COVID Vaccine Trends for USA

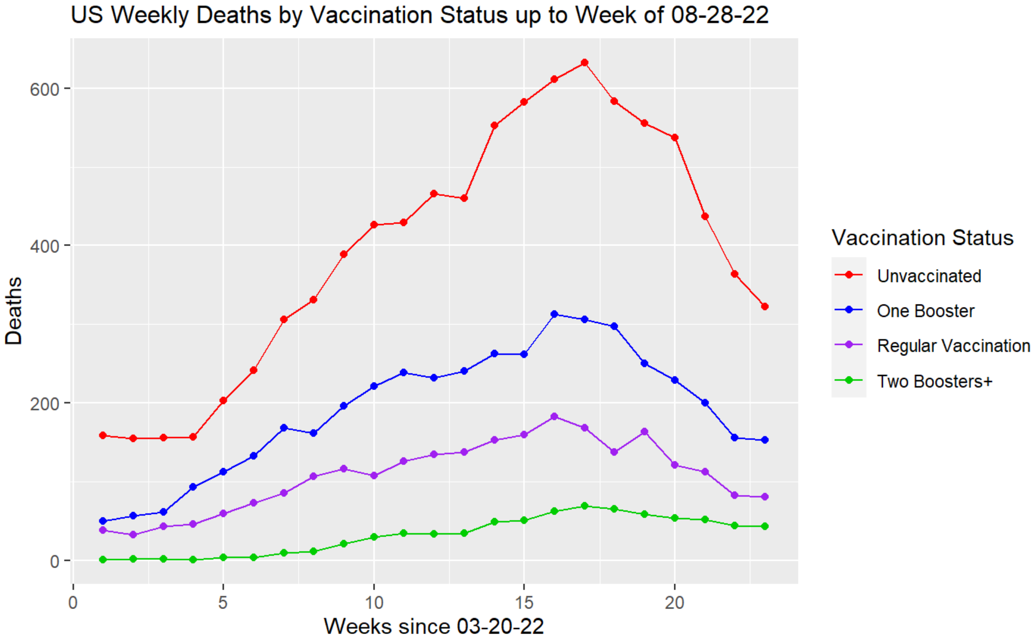


Figure 6: US Weekly Vaccine Trends Color Coded by Vaccination Status

Perhaps our biggest question of interest was regarding the effectiveness of vaccines and how vaccine status affects COVID related deaths. Figure 6 depicts weekly deaths due to COVID in the United States according to vaccination status since March 20, 2022 (e.g. the first point on each trendline indicates the deaths reported on March 27, 2022). Unfortunately, as previously mentioned, we were not able to find a cohesive dataset that relates deaths to vaccination status on a global scale and thus, limited our scope of analysis to the United States. We wanted data with deaths involving two or more booster doses in the US which we were only able to find for late March 2022 to late August 2022 from the CDC [3].

This vaccination data from the CDC involved vaccinations from Moderna, Pfizer, and Janssen (J&J). We only considered Moderna and Pfizer as these are the more prominent brands sought after; additionally according to the CDC, “In most situations, Pfizer-BioNTech, Moderna, or Novavax COVID-19 vaccines are recommended over the J&J/Janssen COVID-19 vaccine for primary and booster vaccination due to the risk of serious adverse events.” [4]. This dataset also categorized data according to different age groups; we decided to consider all age groups here.

As seen in the figure, trend lines are colored according to vaccination status. The red line indicates deaths involving people where no vaccination was detected; the purple line indicates deaths involving people that were known to have only completed the primary series (two vaccines); the blue line indicates deaths involving people that were known to have a primary series in addition to an additional dose; the green line indicates deaths involving people that were known to have taken the primary series in addition to two or more doses.

For the most part, Figure 6 falls within our expectations: “unvaccinated” deaths are the highest as there would be much less protection against the virus; “regular vaccination” deaths are higher than “two or more boosters” deaths considering the latter would yield more protection. Perhaps the most unexpected result we observed was that “one booster” deaths were higher than “regular vaccination” deaths.

However, upon further inspection, we have found this explanation from the CDC: “Because data on the immune status of cases and associated deaths are unavailable, an additional dose in an immunocompromised person cannot be distinguished from a booster dose. This is a relevant consideration because vaccines can be less effective in this group”[3]. In other words, there is ambiguity in the data in regards to booster doses since it is indeed the case that the data we acquired does not indicate the immune statuses of those that have died from COVID. It is unclear whether or not this observation applies to the case with two or more boosters; based on the data, it appears that it likely does not which makes sense as perhaps many immunocompromised individuals may have died before a second additional dose could be attempted.

4 Other Questions of Interest

4.1 Lifestyle Choices and COVID

It has been suspected that smoking increases susceptibility to COVID, but in other reports, there have even been claims that smokers are protected against the virus [5]. This idea is motivated from a medical point of view. Smoking products that contain nicotine and tobacco are known to hinder the immune system of the user. If smokers are known to have weaker immune systems, does this mean they are more likely to get COVID? Our initial assumption of our data was that we had enough information to explore lifestyle choices such as smoking and COVID. The OWID data contained two variables of interest: male smokers and female smokers; values of these variables indicated the percent of the population of a given region were known to be smokers. We wondered; do certain lifestyle choices like smoking correspond to higher deaths due to COVID? Unfortunately, we were unable to observe any clear relationship between smoking (either male or female) and COVID deaths based on the OWID data.

4.2 Demographic Trends

Other questions we wished to investigate was the effect of population densities regarding COVID deaths and how much more susceptible certain demographics are to COVID. For instance, are older people more susceptible to COVID-19? What about immunocompromised individuals?

Due to their old age, elderly people slowly lose their ability to fight off immune challenges. We hypothesized that due to the lack of a healthy immune system older people and immunocompromised individuals are more susceptible. We were able to find (via OWID's dataset) a slight correspondence between the median age of a country versus average daily deaths due to COVID and a clear correspondence between those countries that had a higher portion of the population aged 65 or higher and average daily deaths due to COVID. However, we were fairly well aware of the idea that older people were more susceptible and were not very interested in exploring the idea further. Also, we did not find any data regarding immunocompromised individuals and COVID. For future work, hopefully we will be able to find such data and also explore other demographic relationships such as ethnicity and sex.

4.3 Population Density

A big question of interest that we were unfortunately unable to answer was whether higher population densities corresponded to higher average COVID daily deaths. Our findings were that there was no relationship between population density and average COVID daily deaths; we even checked for average COVID daily cases and still found no relationship. This goes against our intuition and we are honestly not sure whether to believe it; surely, crowded places yield higher risk of the virus than less crowded regions? Then again, we only had access to an entire country's population data from the OWID dataset; instead, we would like to base future work on an accumulative analysis of smaller regions such as states, provinces or cities. We wanted to confirm our observation with other sources, but there appears to be no clear consensus. Overall, we decided not to explore the idea further and to consider it a dead-end for now.

4.4 Policy Effectiveness

As for government policy and COVID infection rate, the OWID dataset has a "stringency index" which measures the strictness of a government's COVID-19 related policies. Is there any evidence that strict governments combat the spread of COVID-19 better than less strict governments? The draconian lockdowns implemented in Shanghai were seen as a bit harsh and over the top to many countries. However, are these policies significantly more effective than comparatively more relaxed policies elsewhere; if so, are these policies worth the immense drawbacks that entail from such policies such as considerable damage to the economy and medical attention being driven away from serious non-COVID conditions (e.g. cancer)?

4.5 Impoverishment

In our analysis, we found that wealthier countries yielded higher average daily deaths which appears to be different than our observations with the trend regarding deaths to cases ratios (poorer countries comprise the top twenty). Perhaps for future work, we may consider average daily deaths to cases ratios and explore the relationship between deaths to cases ratios against average daily deaths. Furthermore, we would like to sample the highest poverty-stricken regions in the world (maybe top 20, 50, or 100) and compare them against each other regarding averge daily deaths due to COVID.

4.6 Climate

Our data provides the country for each observation. With this in mind, what is the effect of geographical climate on the transmissibility of the virus. Do warmer or colder climates increase the spread? Do humid or dry climates increase the spread? This question does have a few obstacles. We would need latitude and longitude variables per observation, for many countries like the US have multiple climate zones. Secondly, climates change based on the season. During the start of the pandemic as well as during its peak, one half of the weather experienced the opposite season as the other half. We believed that comparing different seasonal climates would lead us to false conclusions.

4.7 Air Travel

Air travel may have contributed to the spread of COVID-19. In this regard, did travel heavy regions across the globe have higher infection rates? This question may also help answer a previous question we posed. Did areas of high travel have lower infection rates if they had a stricter government? Especially in the US, many countries continued flights at the start of the pandemic. With this in mind we would have to look at data before countries entered lockdown and flight cancellations.

5 Conclusions

With mass amounts of data at our fingertips, our team analyzed multiple data sets and shifted through different data sources with the objective of determining what external factors influence COVID induced mortality. We have arrived at some interesting conclusions, but feel there is still much more to unveil.

We have found that under-reporting is well-within the realm of possibility. Not only China, but India (having a population over 1.3 billion: over 4 times more than the US) also falls behind the United States in regards to cumulative deaths due to COVID. This is hard to believe, but we have no basis to refute these observations. Understandably, it is inevitable for under-reporting to be circumstantial; therefore, finding evidence of under-reporting due to data tampering would be the objective of much more thorough future endeavors.

Next, we found that inspecting the proportion of COVID cases that resulted in deaths in a given region exhibits an almost entirely different trend when only considering total deaths. Our analysis on a global scale revealed that poorer countries seem to yield higher deaths to cases ratios which served as more motivation to investigate how the wealth of countries affects their average daily deaths due to COVID.

It turns out that wealthier countries yield higher average daily deaths due to COVID. Considering our results with the deaths to cases ratio, it appears there is much more work that needs to be done. Indeed, as we have discovered, when considering this relationship among different continents, we see differing trends regarding the relationship between GDP per capita and average daily COVID deaths. Therefore, we would like to explore this relationship on smaller scales; perhaps we will find something entirely different from the positive trend exhibited by North America when only considering the GDP per capita of each state (or of each city) in the USA.

Finally, according to our observations for the United States, vaccinations are effective against mitigating the frequency of deaths due to COVID. For the most part, it seems safe to conclude that boosters provide added layer of protection against COVID. However, in future efforts to analyze the effectiveness of vaccines, we would like to have access to data that not only indicates the vaccination status of a COVID related fatality, but also indicates other useful health related information such as immune status.

Also, since our scope of analysis was restricted to the continental United States, we hope to include such information on a global scale in future research. In addition to comparing COVID deaths according to vaccination status worldwide, we would like to explore the effectiveness of vaccine brands outside of the United States such as Sinovac and how they compare against brands in the United States like Pfizer and Moderna in regards to COVID related deaths.

References

[1] OWID. “COVID-19 Dataset by Our World in Data”. In: (Sept. 2022). URL: https://github.com/owid/covid-19-data.

[2] JHU CSSE. “JHU CSSE COVID-19 Dataset”. In: (Nov. 2022). URL: https://github.com/CSSEGISandData/COVID-19.

[3] CDC Public Health Surveillance. “Rates of COVID-19 Cases or Deaths by Age Group and Vaccination Status and Second Booster Dose”. In: (Oct. 2022). URL: https://data.cdc.gov/Public-Health-Surveillance/Rates-of-COVID-19-Cases-or-Deaths-by-Age-Group-and/ukww-au2k.

[4] CDC. “Overview of COVID-19 vaccines”. In: (Nov. 2022). URL: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/overview-COVID-19-vaccines.html.

[5] Naomi A van Westen-Lagerweij et al. “Are smokers protected against SARS-CoV-2 infection (COVID-19)? The origins of the myth”. In: *npj Primary Care Respiratory Medicine* 31.1 (Feb. 2021), p. 10.

A Source Code

```
# load relevant libraries
library(ggplot2)
library(xts)
library(dplyr)
library(stringr)
library(glue)
library(ggmap)
library(scales)
library(ggpubr)
library(gganimate)

# load global map data and jhu covid data
world <- map_data("world")
jhu_covid_raw <- read.csv('jhu_covid.csv')

# load US states/territories jhu covid data
us_states_data <- read.csv('us_states.csv')

# load owid covid data
raw_covid_data <- read.csv('owid-covid-data.csv')

# load cdc covid vaccine data
raw_booster_data <- read.csv('cdc_us_booster_data.csv')
```

A.1 Total COVID Deaths and Deaths to Cases Ratio

A.1.1 Global Scale

```
# in order to map global covid data onto a world map, we need uniformity in the
# country names; part of this process had to be done manually by first
# checking which country names matched between the "world" dataset and the
# jhu covid data set and then manually modifying the ones that did not match
jhu_covid_raw[jhu_covid_raw$Country_Region == "Taiwan*",
              "Country_Region"] <- "Taiwan"

n_us <- length(jhu_covid_raw[jhu_covid_raw$Country_Region == "US",
                             "Country_Region"])
jhu_covid_raw[jhu_covid_raw$Country_Region == "US", "Country_Region"] <-
  rep(c("USA"), times = n_us)

n_uk <- length(jhu_covid_raw[jhu_covid_raw$Country_Region == "United Kingdom",
                             "Country_Region"])
```

```
jhu_covid_raw[jhu_covid_raw$Country_Region == "United Kingdom" ,
              "Country_Region"] <- rep(c("UK"), times = n_uk)

# Koreas
jhu_covid_raw[jhu_covid_raw$Country_Region == "Korea, South",
              "Country_Region"] <- "South Korea"
jhu_covid_raw[jhu_covid_raw$Country_Region == "Korea, North",
              "Country_Region"] <- "North Korea"

# Congos
jhu_covid_raw[jhu_covid_raw$Country_Region == "Congo (Kinshasa)",
              "Country_Region"] <- "Democratic Republic of the Congo"
jhu_covid_raw[jhu_covid_raw$Country_Region == "Congo (Brazzaville)",
              "Country_Region"] <- "Republic of Congo"


# etc.
jhu_covid_raw[jhu_covid_raw$Country_Region == "Burma" ,
              "Country_Region"] <- "Myanmar"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Czechia",
              "Country_Region"] <- "Czech Republic"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Antigua and Barbuda",
              "Country_Region"] <- "Antigua"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Cabo Verde",
              "Country_Region"] <- "Cape Verde"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Cote d'Ivoire",
              "Country_Region"] <- "Ivory Coast"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Eswatini",
              "Country_Region"] <- "Swaziland"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Holy See",
              "Country_Region"] <- "Vatican"

jhu_covid_raw[jhu_covid_raw$Country_Region == "Saint Kitts and Nevis",
              "Country_Region"] <- "Saint Kitts"


jhu_covid_raw[jhu_covid_raw$Country_Region ==
                "Saint Vincent and the Grenadines", "Country_Region"] <-
  "Saint Vincent"
```

```
jhu_covid_raw[jhu_covid_raw$Country_Region == "Trinidad and Tobago",
              "Country_Region"] <- "Trinidad"



jhu_covid_raw[jhu_covid_raw$Country_Region == "West Bank and Gaza",
              "Country_Region"] <- "Palestine"
```

```
# check which country names match and which don't;
# store matched country names in "mutual_countries" vector
world_map_countries <- unique(world$region)
jhu_countries <- unique(jhu_covid_raw$Country_Region)

mutual_countries <- intersect(world_map_countries, jhu_countries)

# we will use the "world" dataset as our primary dataset;
# we will transfer values of interest from the jhu dataset
# over to the "world" dataset; here, we prepare
# new columns for the variables of interest;
# cf_ratio is case to fatalities ratio where we
# will have it simply be a derived quantity
# being total_deaths divided by total_cases
# for each country
world['total_cases'] <- NA
world['total_deaths'] <- NA
world['cf_ratio'] <- NA


# iterate through countries in mutual_countries vector; append jhu data
# to world dataset according to country; in the world dataset, there are
# multiple instances of each country for longitude and latitude columns
# so jhu data will need to be repeated for each country
for (country in mutual_countries){
  # check and store repeat instances of each country in world dataset
  rep_length <- length(world[world$region == country, ]$region)

  # append total cases for each country
  world[world$region == country, ]$total_cases <-
    rep(sum(jhu_covid_raw[jhu_covid_raw$Country_Region == country,
                          ]$Confirmed), times = rep_length)

  # append total deaths for each country
  world[world$region == country, ]$total_deaths <-
    rep(sum(jhu_covid_raw[jhu_covid_raw$Country_Region == country, ]$Deaths),
        times = rep_length)

  # evaluate and append total deaths to cases ratio for each country
  world[world$region == country, ]$cf_ratio <-
    world[world$region == country,
          ]$total_deaths / world[world$region == country, ]$total_cases
}
```

```
# make NA values zero so they appear on the map
world[is.na(world$total_cases), "total_cases"] <- 0
world[is.na(world$total_deaths), "total_deaths"] <- 0
world[is.na(world$cf_ratio), "cf_ratio"] <- 0

# plot world map with color mapping based on total deaths
plot_deaths <- ggplot() +
  geom_polygon(data = world,
               aes(long, lat, group = group, fill = total_deaths),
               color = "black", size = 0.1) +
  scale_fill_distiller(palette = "Reds", direction = 1,
                       breaks = c(200000, 400000, 600000, 800000, 1000000),
                       limits = c(0, 1200000), labels = comma) +
  labs(fill='Confirmed \n   Deaths') +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(), axis.ticks = element_blank()) +
  xlab("") + ylab("") + theme(
  panel.background = element_rect(fill = "lightblue",
                                  colour = "lightblue",
                                  size = 0.5, linetype = "solid"),
  panel.grid.major = element_line(size = 0.5, linetype = 'solid',
                                  colour = "white"),
  panel.grid.minor = element_line(size = 0.25, linetype = 'solid',
                                  colour = "white")
  ) + ggtitle("Cumulative COVID Deaths by Country as of 11-01-22")

plot_deaths
```

```
# extract deaths to cases ratio column from world dataset
# and store in a vector with names according to country
cf_ratios_raw <- world$cf_ratio
names(cf_ratios_raw) <- world$region

# filter out zero values when considering deaths to cases ratios
cf_ratios_1 <- cf_ratios_raw[cf_ratios_raw > 0]
```

```
# extract only unique values
cf_ratios <- cf_ratios_1[-which(duplicated(cf_ratios_1))]

# extract top 20 deaths to cases ratio values; filter out north korea
# by starting at index 2 up to 21; store in vector "top_cf_ratios"
n <- 20
top_cf_ratios <- sort(cf_ratios, decreasing = TRUE)[2:(n+1)]


# store values as percentages in a dataframe with a separate country column
top_cf_df <- data.frame(country = names(top_cf_ratios),
                        cf_ratio = top_cf_ratios*100)

# set up top_Cf_df so that deaths to cases ratios are sorted
# in descending order
top_cf_df$country <- factor(top_cf_df$country,
                            levels = top_cf_df$country[order(
                              top_cf_df$cf_ratio, decreasing = FALSE)])
```

```
# generate horizontal bar plot for deaths to cases ratio labeled by country
# in descending order
ggplot(data = top_cf_df, aes(country, cf_ratio)) +
  geom_col(aes(fill = cf_ratio), color = "black", show.legend = FALSE) +
  xlab("Country") + ylab("Case-Fatality Ratio (Percentage)") +
  ggtitle("Top 20 Countries for Case-Fatality Ratio as of 11-01-22") +
  coord_flip() + scale_fill_distiller(palette = "Reds", direction = 1)
```

```
# **CODE FOR ANIMATION**
# prepare dataframe for animation with 2 distinct frames "a" and "b" where
# frame "a" has all values at zero
top_cf_df$frame <- "b"

# create separate dataframe for the "begin" state with zero values
begin_state <- top_cf_df
begin_state$cf_ratio <- 0
begin_state$frame <- "a"



# set up begin_state so that countries are sorted
# in the same manner as top_cf_df
begin_state$country <- factor(begin_state$country,
                              levels = begin_state$country[
                                order(top_cf_df$cf_ratio, decreasing = FALSE)])
```

```r
# combine the begin_state and top_cf_df datasets
animation_df <- rbind(begin_state, top_cf_df)

# filter out redundant rownames
rownames(animation_df) <- NULL

# prepare plot to be animated
anim_cf <- ggplot(data = animation_df, aes(country, cf_ratio)) +
  geom_col(aes(fill = cf_ratio), color = "black", show.legend = FALSE) +
  xlab("Country") + ylab("Case-Fatality Ratio (Percentage)") +
  theme(plot.title = element_text(size=12), axis.text=element_text(size=12),
        axis.title=element_text(size=14)) +
  ggtitle("Top 20 Countries for Case-Fatality Ratio as of 11-01-22") +
  coord_flip() + scale_fill_distiller(palette = "Reds", direction = 1) +
  transition_states(
    frame,
    transition_length = 1,
    state_length = 1,
    wrap = FALSE
  ) + ease_aes('linear')


# store animated plot as an animation object called "anim"
anim <- animate(anim_cf, duration = 3, width = 1400, height = 865,
                res = 200, renderer = ffmpeg_renderer())

# export the animation object as an mp4 file
anim_save("cf_anim.mp4", animation = anim)
```

A.1.2 Continental US Scale

```r
# load US state map data
map_data_us_states <- map_data("state")

# List of States, regions and other areas to be left out of analysis
del_states <- c("Alaska", "Hawaii", "American Samoa", "Diamond Princess",
                "Grand Princess", "Guam", "Northern Mariana Islands",
                "Puerto Rico","Virgin Islands")

# subset us_states_data based on regions not in del_states
continental_states_data <- subset(us_states_data,
                                  !Province_State %in% del_states)

# Store original region names
```

```r
original_state_list <- continental_states_data$Province_State

# uncapitalize first letter of region names; store as a vector
continental_states_data$Province_State <-
  unlist(lapply(continental_states_data$Province_State, tolower))

# prepare to iterate over lower case state names to find number of
# instance appearances in map_data_us_states for each name
state_list <- continental_states_data$Province_State

# prepare vector to store number of state name instances
state_list_lengths <- c()

# manual counter so iteration can be over state names directly
index <- 1
for (state in state_list){
  # find desired length and store in state_list_lengths vector
  state_list_lengths[index] <- length(map_data_us_states[
    map_data_us_states$region == state, "region"])
  index <- index + 1
}

confirmed_column <- rep(continental_states_data$Confirmed,
                        times = state_list_lengths)

deaths_column <- rep(continental_states_data$Deaths,
                     times = state_list_lengths)

cf_ratio_column <- rep(continental_states_data$Case_Fatality_Ratio,
                       times = state_list_lengths)


# prepare columns from map_data_us_states to extract for subset
new_columns <- c( "region", "long", "lat", "group", "order" )

# subset map_data_us_states according to the prepared columns
us_states_df <- map_data_us_states[, new_columns]

# append the other prepared columns
us_states_df["confirmed"] <- confirmed_column

us_states_df["deaths"] <- deaths_column

us_states_df["case_fatalities_ratio"] <- cf_ratio_column
```

```
# plot continental US map with color mapping based on total deaths
plot_deaths <- ggplot(data = us_states_df) +
  geom_polygon(aes(long, lat, group = group, fill = deaths), size = 0.5,
               color = "black", show.legend = TRUE) +
  coord_quickmap() + ggtitle(
    "Cumulative COVID Deaths by State in the Continental US as of 11-01-22") +
  labs(fill='Confirmed \n   Deaths') +
  scale_fill_distiller(palette = "Reds", direction = 1, limits = c(0, 100000),
                       breaks = c(20000, 40000, 60000, 80000), labels = comma) +
  theme(axis.text.x = element_blank(),  axis.text.y = element_blank(),
        axis.ticks = element_blank()) + xlab("") + ylab("") +
  theme(panel.background = element_rect(fill = "lightblue",
                                        colour = "lightblue",
                                        size = 0.5, linetype = "solid"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid',
                                        colour = "white"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid',
                                        colour = "white"))

plot_deaths
```

```
# reuse original state names (capital first letter)
cap_states <- original_state_list

# extract cf ratios from us_states_df
cf_ratios <- unique(us_states_df$case_fatalities_ratio)

# store cap_states and cf_ratios in a new dataframe called cf_df
cf_df <- data.frame("States" = cap_states,
                    "Case to Fatalities Ratio" = cf_ratios)

# rearrange cf_df in descending order according to case to fatalities ratio
reordered_cf_df <- cf_df %>% arrange(desc(Case.to.Fatalities.Ratio))

# extract top 20 states from reordered cf_df
top_cf_df <- reordered_cf_df[1:20, ]

# force ordering on top_cf_df
top_cf_df$States <- factor(top_cf_df$States,
                           levels = top_cf_df$States[
                             order(top_cf_df$Case.to.Fatalities.Ratio,
                                   decreasing = FALSE)])

# generate horizontal bar plot for deaths to cases ratio labeled by state
# in descending order
```

```r
ggplot(data = top_cf_df, aes(States, Case.to.Fatalities.Ratio)) +
  geom_col(aes(fill = Case.to.Fatalities.Ratio),
           color = "black", show.legend = FALSE) +
  ylab("Case-Fatality Ratio (Percentage)") +
  ggtitle("Top 20 States for Case-Fatality Ratio as of 11-01-22") +
  coord_flip() + scale_fill_distiller(palette = "Reds", direction = 1)
```

A.2 GDP Per Capita vs. Daily COVID Deaths

```r
# prepare columns desired for sub-dataframe;
# the locations column contains different countries;
# locations columns also contains other values
# not of interest such as "world"
columns <- c("location", "date", "gdp_per_capita",
             "new_cases_smoothed_per_million",
             "new_deaths_smoothed_per_million",
             "continent")

# extract desired columns from raw_covid_data
gdp_data <- raw_covid_data[, columns]

# filter out NA values of gdp_per_capita variable
gdp_data_nonull <- gdp_data[-which(is.na(gdp_data$gdp_per_capita)), ]

# create new column called "gdp_per_capita_lvl";
# it happens to be the case here that there
# is a unique gdp_per_capita_value for each
# country which means that the gdp per capita
# for each country not filtered out is assumed to be constant
# throughout the entire time-frame of our data
gdp_data_nonull[, "gdp_per_capita_lvl"] <-
  round(gdp_data_nonull$gdp_per_capita)
gdp_data_nonull[, "gdp_per_capita_lvl"] <-
  as.factor(gdp_data_nonull$gdp_per_capita)
```

```r
# Filter out NA values of new_cases_smoothed_per_million
semi_final_gdp_data <- gdp_data_nonull[
  -which(is.na(gdp_data_nonull$new_cases_smoothed_per_million)), ]

# Filter out NA values of new_deaths_smoothed_per_million
final_gdp_data <- semi_final_gdp_data[-which(
  is.na(semi_final_gdp_data$new_deaths_smoothed_per_million)), ]
```

```
# prepare empty vectors to store average daily deaths
# and average daily cases; also prepare empty vectors
# to store gdp_per_capita, continent, and location
avg_cases <- c()
avg_deaths <- c()
gdp_per_capita <- c()
continent <- c()
location <- c()


# iterate through gdp_per_capita factor levels and append elements
# to pertinent vectors

for (i in 1:length(levels(final_gdp_data$gdp_per_capita_lvl))){
  # append average daily cases to avg_cases vector
  avg_cases[i] <- mean(final_gdp_data[final_gdp_data$gdp_per_capita_lvl
                                      == levels(
                                        final_gdp_data$gdp_per_capita_lvl)[i],
                                      "new_cases_smoothed_per_million"])

  # append average daily deaths to avg_deaths vector
  avg_deaths[i] <- mean(final_gdp_data[
    final_gdp_data$gdp_per_capita_lvl == levels(
      final_gdp_data$gdp_per_capita_lvl)[i],"new_deaths_smoothed_per_million"])



  # append locations to location vector
  location[i] <- final_gdp_data[final_gdp_data$gdp_per_capita_lvl ==
                                  levels(final_gdp_data$gdp_per_capita_lvl)[i],
                                "location"][1]

  # append continents to continent vector
  continent[i] <- final_gdp_data[final_gdp_data$gdp_per_capita_lvl ==
                                   levels(final_gdp_data$gdp_per_capita_lvl
                                          )[i], "continent"][1]

  # append gdp_per_capita to gdp_per_capita vector
  gdp_per_capita[i] <- final_gdp_data[
    final_gdp_data$gdp_per_capita_lvl ==
      levels(final_gdp_data$gdp_per_capita_lvl)[i], "gdp_per_capita"][1]
}
```

```
# Generate new dataframe based on average daily deaths
# and average daily cases; also relabel the gdp_per_capita factor levels as
```

```r
# positive integers since each level is unique and ordered
new_gdp_frame_raw <- data.frame(gdp_lvl = 1:length(gdp_per_capita),
                                location = location, avg_cases = avg_cases,
                                avg_deaths = avg_deaths, Continent = continent,
                                gdp_per_capita = gdp_per_capita)

# filter out NA values in avg_cases; this also happens to be sufficient
# in filtering out NA values of other columns
new_gdp_frame <- new_gdp_frame_raw[
  -which(is.na(new_gdp_frame_raw$avg_cases)), ]

# prior analysis demonstrates we have skewed data for varaibles of interest;
# apply log transform to variables of interest
new_gdp_frame$log_gdp_per_capita <- log(new_gdp_frame$gdp_per_capita)
new_gdp_frame$log_avg_cases <- log(new_gdp_frame$avg_cases)
new_gdp_frame$log_avg_deaths <- log(new_gdp_frame$avg_deaths)

# filter out the location value "world"
new_gdp_frame <- new_gdp_frame[!new_gdp_frame$Continent == "", ]

# generate preliminary plot
ggplot(data = new_gdp_frame) +
  geom_point(aes(gdp_per_capita, avg_deaths, color = Continent)) +
  xlab("GDP Per Capita (USD)") + ylab("Average Daily New Deaths per Million") +
  ggtitle("Average Daily Deaths vs. GDP Per Capita (2020-01-01 to 2022-09-27)")
```

```r
# plot log transform of average daily deaths against log transform of
# gdp per capita; include regression line with confidence interval,
# correlation coefficient "R" and p-value "p"

gdp_plot_deaths <- ggplot(data = new_gdp_frame) +
  geom_point(aes(log_gdp_per_capita, log_avg_deaths, color = Continent)) +
  geom_smooth(formula = 'y~x', method = "lm",
              aes(log_gdp_per_capita, log_avg_deaths),
              color = "red", linetype = 2, size = 0.5) +
  xlab("Log(GDP Per Capita)") +
  ylab("Log(Average Daily New Deaths per Million)") +
  theme(plot.title = element_text(size=12))  +
  ggtitle("Average Daily Deaths vs. GDP Per Capita (2020-01-01 to 2022-09-27)")

gdp_plot_deaths + stat_cor(aes(log_gdp_per_capita, log_avg_deaths),
                           method="pearson", p.accuracy = 0.001,
                           r.accuracy = 0.01)
```

```r
# apply facet-wrap to the same plot based on different continents
gdp_facet <- ggplot(data = new_gdp_frame, aes(log_gdp_per_capita,
                                              log_avg_deaths)) +
  geom_point(aes(color = Continent), show.legend = FALSE) +
  geom_smooth(formula = 'y~x', method = "lm",
              aes(log_gdp_per_capita, log_avg_deaths),
              color = "red", linetype = 2, size = 0.5)  +
  stat_cor(aes(log_gdp_per_capita, log_avg_deaths),
           method="pearson", p.accuracy = 0.001, r.accuracy = 0.01,
           label.x = 3.63, label.y = -5) + xlab("Log(GDP Per Capita)") +
  ylab("Log(Average Daily New Deaths per Million)") +
  ggtitle("Average Daily Deaths vs. GDP Per Capita (2020-01-01 to 2022-09-27)")

gdp_facet + facet_wrap(vars(Continent))
```

A.3 COVID Vaccine Trends for USA

```r
# filter booster data to only contain "all_ages" age group
raw_boost_data_f1 <- raw_booster_data[
  raw_booster_data$age_group == "all_ages" |
    raw_booster_data$age_group == "all_ages", ]

# filter booster data to only contain Moderna and Pfizer vaccine products
boost_data <- raw_boost_data_f1[
  raw_boost_data_f1$vaccine_product == "Moderna" |
    raw_boost_data_f1$vaccine_product == "Pfizer", ]

# filter out case and death outcomes and store them in seaparate
# dataframes; create new column "week" in each dataframe that
# counts the week since starting week
boost_cases <- boost_data[boost_data$outcome == "case", ]
boost_cases$week <- 1:length(boost_cases$mmwr_week)

boost_deaths <- boost_data[boost_data$outcome == "death", ]
boost_deaths$week <- 1:length(boost_deaths$mmwr_week)

# prepare columns that averages moderna and pfizer data together;
# week count is redundant since it is counting moderna and pfizer
# products as separate weeks so we will account for that too

boost_deaths$week <- NA
boost_deaths$avg_no_booster <- NA
boost_deaths$avg_one_booster <- NA
boost_deaths$avg_two_booster <- NA
```

```r
# manual counter so we can use iterator to call sections of
# the boost_deaths dataframe
i = 1

# iterate over unique values of mmwr_week
for (mmwr in unique(boost_deaths$mmwr_week)){
  # append count as week number
  # (should be numbered according to mmwr_week count)
  boost_deaths[boost_deaths$mmwr_week == mmwr, "week"] <- i

  # append average regular vaccine deaths
  boost_deaths[boost_deaths$mmwr_week == mmwr, "avg_no_booster"] <-
    mean(boost_deaths[boost_deaths$mmwr_week == mmwr,
                      ]$vaccinated_with_outcome)

  # append average regular vaccine plus one booster deaths
  boost_deaths[boost_deaths$mmwr_week == mmwr, "avg_one_booster"] <-
    mean(boost_deaths[boost_deaths$mmwr_week == mmwr,
                      ]$one_boosted_with_outcome)

  # append average regular vaccine plus two or more booster deaths
  boost_deaths[boost_deaths$mmwr_week == mmwr, "avg_two_booster"] <-
    mean(boost_deaths[boost_deaths$mmwr_week == mmwr,
                      ]$two_boosted_with_outcome)
  # increment counter
  i = i + 1
}
```

```r
# prepare color mappings according to vaccination status
cols <- c("Unvaccinated" = "red", "One Booster" = "blue",
          "Regular Vaccination" = "purple", "Two Boosters+" = "green3")

# plot vaccine trend lines according to vaccination status;
# only Moderna subset is considered since average deaths
# according to vaccine status is the same for both Moderna and Pfizer
ggplot(data = NULL, aes(x = week)) +
  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = unvaccinated_with_outcome, color = "Unvaccinated")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = unvaccinated_with_outcome, color = "Unvaccinated")) +

  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = avg_no_booster, color = "Regular Vaccination")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = avg_no_booster, color = "Regular Vaccination")) +
  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = avg_one_booster, color = "One Booster")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = avg_one_booster, color = "One Booster")) +
  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = avg_two_booster, color = "Two Boosters+")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = avg_two_booster, color = "Two Boosters+")) +
  ggtitle("US Weekly Deaths by Vaccination Status up to Week of 08-28-22") +
  theme(plot.title = element_text(size=12)) + xlab("Weeks since 03-20-22") +
  ylab("Deaths") + scale_color_manual(values=cols) +
  labs(color='Vaccination Status')
```

```
#  **ANIMATION CODE**
# store plot with animation effect in anim_vac
anim_vac <- ggplot(data = NULL, aes(x = week)) +
  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = unvaccinated_with_outcome, color = "Unvaccinated")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = unvaccinated_with_outcome, color = "Unvaccinated")) +

  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = avg_no_booster, color = "Regular Vaccination")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = avg_no_booster, color = "Regular Vaccination")) +

  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = avg_one_booster, color = "One Booster")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = avg_one_booster, color = "One Booster")) +

  geom_line(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
            aes(y = avg_two_booster, color = "Two Boosters+")) +
  geom_point(data = boost_deaths[boost_deaths$vaccine_product == "Moderna", ],
             aes(y = avg_two_booster, color = "Two Boosters+")) +
  ggtitle("US Weekly Deaths by Vaccination Status up to Week of 08-28-22") +
  theme(plot.title = element_text(size=12)) + xlab("Weeks since 03-20-22") +
  ylab("Deaths") + scale_color_manual(values=cols) +
  labs(color='Vaccination Status') + transition_reveal(week)

# generate animation object using anim_vac
anim <- animate(anim_vac, width = 1400, height = 865, res = 200,
                renderer = ffmpeg_renderer())
# save animation object as mp4 file
anim_save("vac_chart.mp4", animation = anim)
```