COVID-19 Report

Robert Forrest

April 11th 2024

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

Intro

In this report, I'll be taking at look at US COVID-19 data from John Hopkins University. The data can be found hosted on their github here: https://github.com/CSSEGISandData/COVID-19. In it contains the number of cases and deaths for each date for all states in the US. I will perform some tidying of the data, data visualizations with analysis, create a model and go through some conclusions. I am interested in exploring how the cases and deaths trended over time for the US as well as how deadly the virus was.

Import Data

Read in 2 csv files from John Hopkin's github repo.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov
file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")

urls <- str_c(url_in, file_names)

# read data in
us_cases <- read_csv(urls[1], show_col_types = FALSE)
us_deaths <- read_csv(urls[2], show_col_types = FALSE)
```

Tidy and Transform

To clean the data, I will transform the data from a wide format into a long format. This will allow for better analysis later on. I will also drop some unnecessary columns and create some new columns.

```
us_cases_by_state <- us_cases %>%
    pivot_longer(cols = -c(UID, iso2, iso3, code3, FIPS, Admin2, Province_State,
                           Country_Region, Lat, Long_, Combined_Key),
                            names_to = "date",
                            values_to = "cases") %>%
  mutate(date = mdy(date)) %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Country_Region, Lat, Long_,
            Combined_Key)) %>%
  group_by(Province_State, date) %>%
  summarize(cases = sum(cases), .groups = 'keep')


us_deaths_by_state <- us_deaths %>%
    pivot_longer(cols = -c(UID, iso2, iso3, code3, FIPS, Admin2, Province_State,
                           Country_Region, Lat, Long_, Combined_Key, Population),
                            names_to = "date",
                            values_to = "deaths") %>%
  mutate(date = mdy(date)) %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Country_Region, Lat, Long_,
            Combined_Key, Population)) %>%
  group_by(Province_State, date) %>%
  summarize(deaths = sum(deaths), .groups = 'keep')
```

Here I will combine the data for cases and deaths into a single table. I will also add a deaths_per_cases column.

```
us_all_data <- us_cases_by_state %>%
  full_join(us_deaths_by_state, by=c("Province_State", "date")) %>%
  mutate(deaths_per_cases = deaths / cases)
```

Its important to get rid of any potentially bad data. Here I will make sure that the deaths_per_cases do not exceed 1.

```
us_all_data <- us_all_data %>%
  filter(deaths_per_cases < 1) %>%
  arrange(Province_State, date)
```

Visualization and Analysis

Lets take a look at the top 5 states with the most cases and deaths.

```
cases_and_deaths_by_state <- us_all_data %>%
  group_by(Province_State) %>%
  summarise(total_cases = sum(cases), total_deaths = sum(deaths)) %>%
  arrange(desc(total_cases))

# This one is sorted by cases
head(cases_and_deaths_by_state, 5)
```

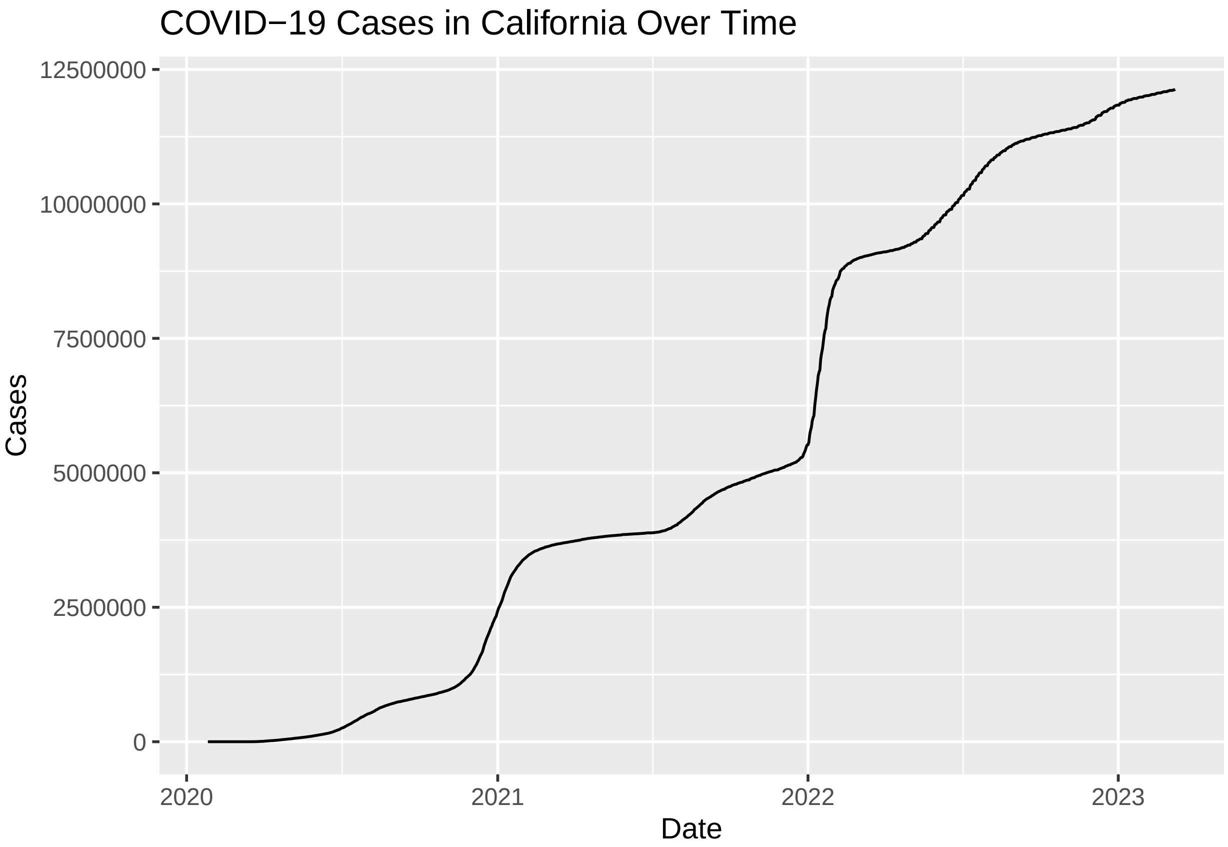
```
## # A tibble: 5 x 3
##   Province_State total_cases total_deaths
##   <chr>                <dbl>        <dbl>
## 1 California      6166190335     65490302
## 2 Texas           4566537657     61302166
## 3 Florida         3978357707     51475342
## 4 New York        3392006819     58121236
## 5 Illinois        2122240785     28240376
```

```
# This one is sorted by deaths
head(cases_and_deaths_by_state %>% arrange(desc(total_deaths)), 5)
```

```
## # A tibble: 5 x 3
##   Province_State total_cases total_deaths
##   <chr>                <dbl>        <dbl>
## 1 California      6166190335     65490302
## 2 Texas           4566537657     61302166
## 3 New York        3392006819     58121236
## 4 Florida         3978357707     51475342
## 5 Pennsylvania    1836846159     31912144
```

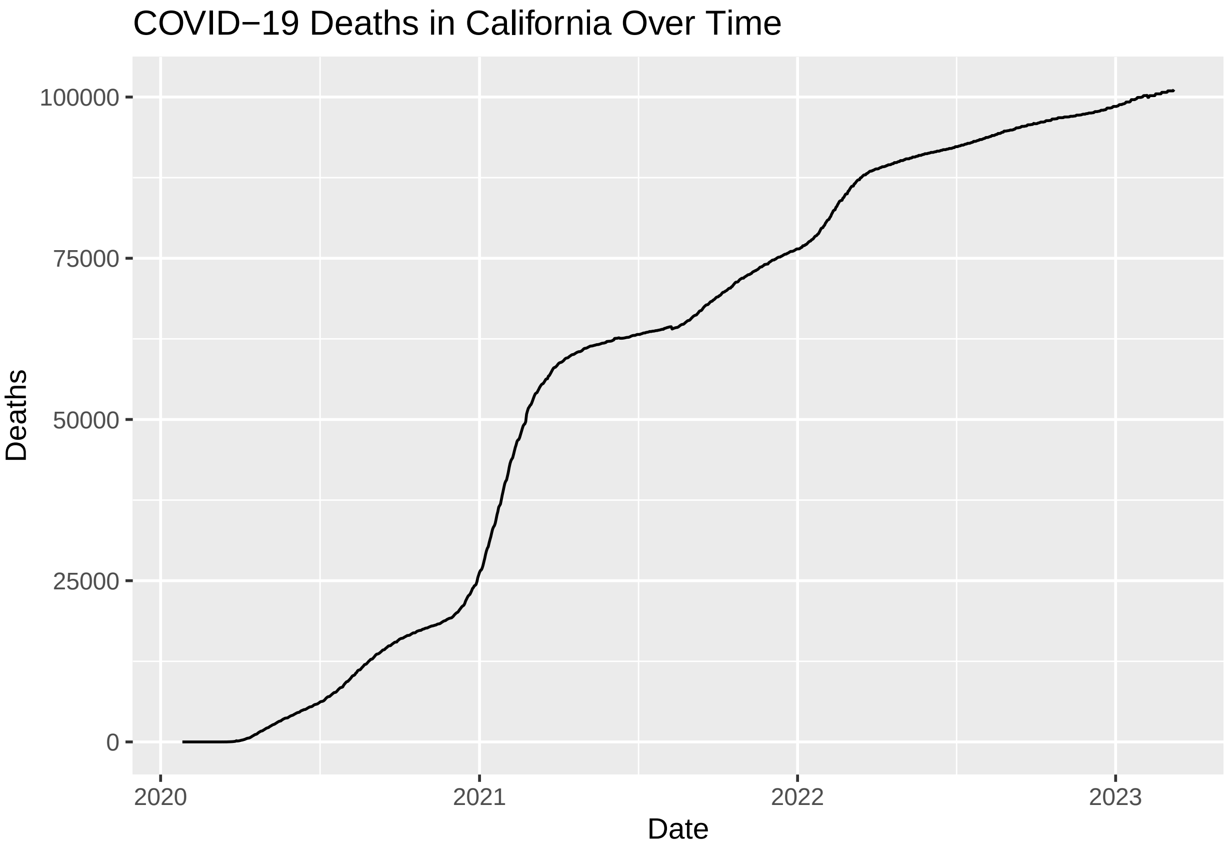
From the results, California had the most cases and deaths. Lets take a look at how California's cases and deaths trended over time.

```
ggplot(data = subset(us_all_data, Province_State == "California"),
       aes(x = date, y = cases)) +
  geom_line() +
  labs(x = "Date", y = "Cases", title = "COVID-19 Cases in California Over Time")
```



Its interesting to see that were appears to be 2 major spikes in the number of cases near the end of 2020 and the beginning of 2022. I believe this could be due to the heavy amount of holiday traveling going on around those times.

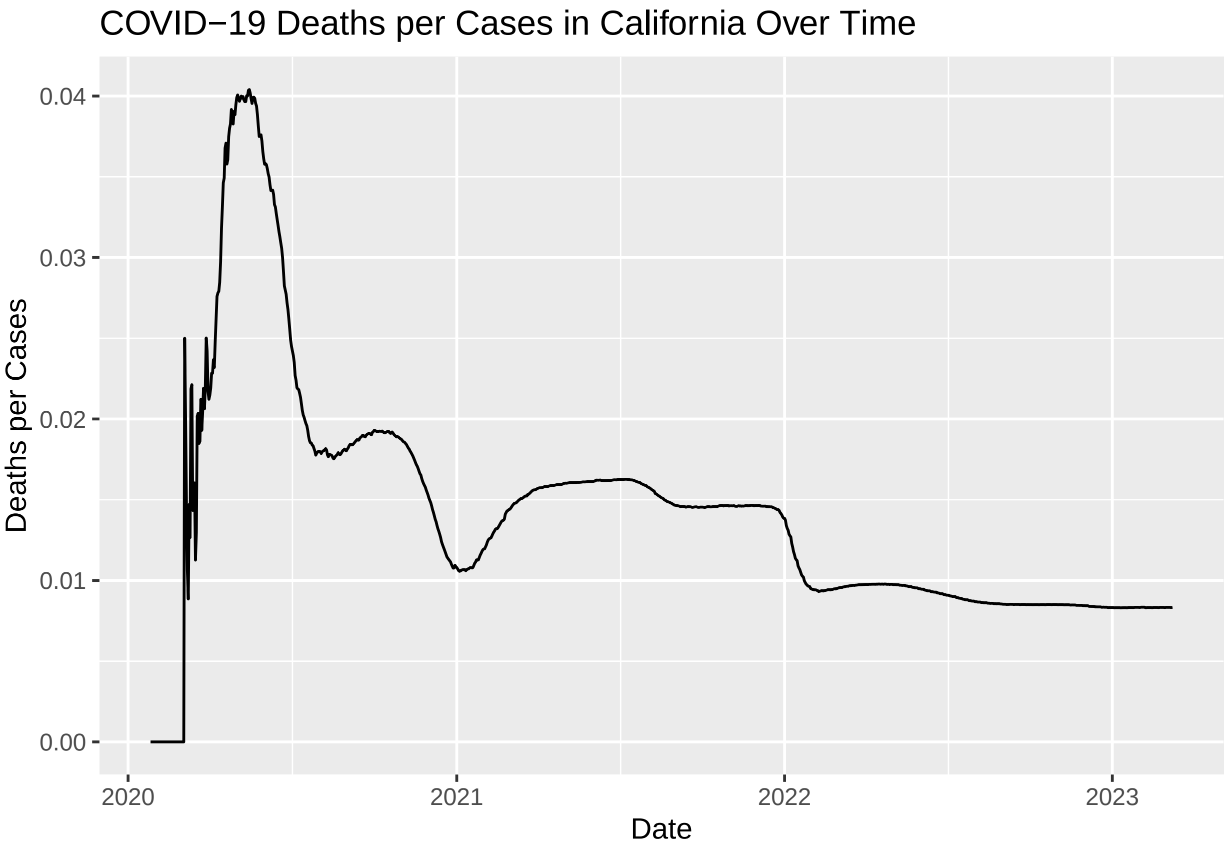
```
ggplot(data = subset(us_all_data, Province_State == "California"),
       aes(x = date, y = deaths)) +
  geom_line() +
  labs(x = "Date", y = "Deaths", title = "COVID-19 Deaths in California Over Time")
```



When taking at look the deaths, it makes sense that given the large amount of cases near 2021, that there would also be a spike in the deaths. The vaccine was not widely available in time to prevent the deaths. When comparing the second spike in cases in 2022, we don't see such a drastic increase in deaths, which is most likely due to people either having the vaccine already or they may have already gotten sick prior.

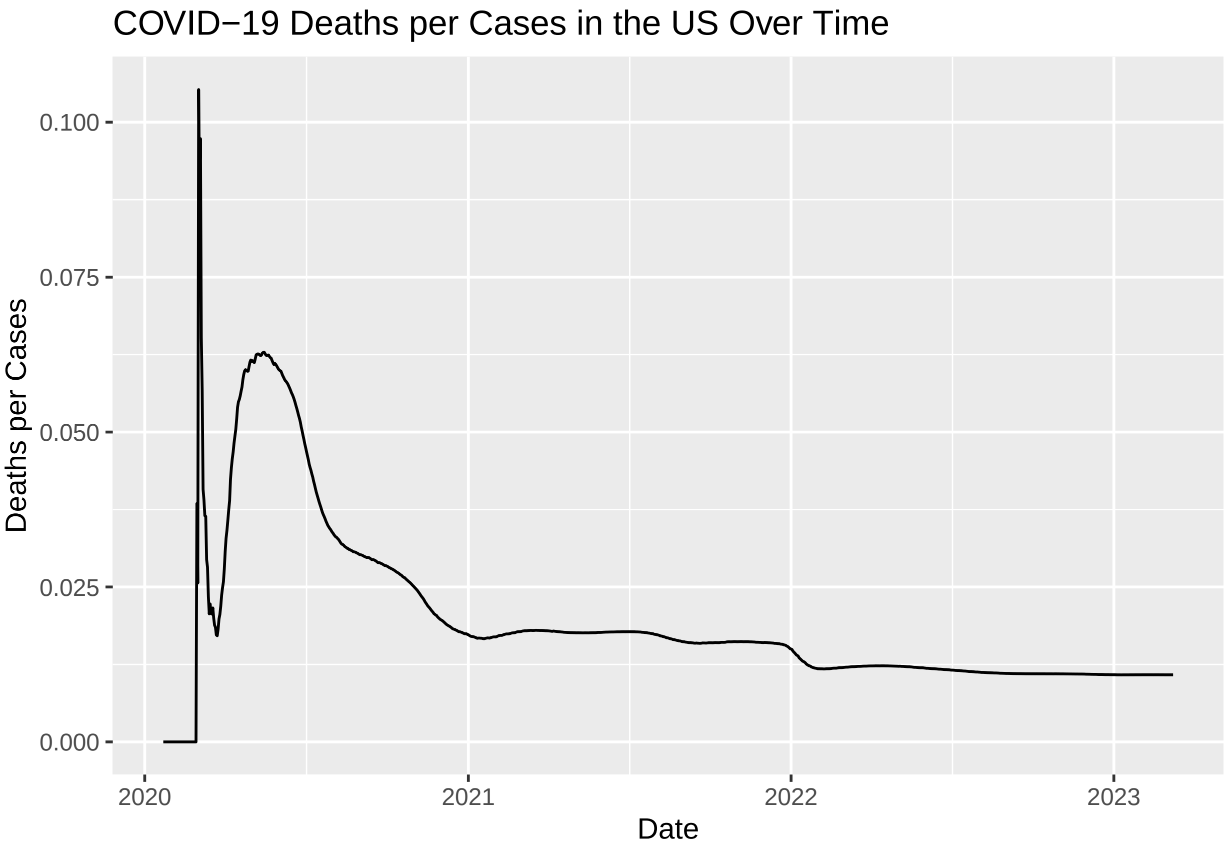
Here we will take a look at the number of deaths per cases over time. This will allow us to understand how deadly the virus was in California.

```r
ggplot(data = subset(us_all_data, Province_State == "California"),
       aes(x = date, y = deaths_per_cases)) +
  geom_line() +
  labs(x = "Date", y = "Deaths per Cases", title = "COVID-19 Deaths per Cases in California Over Time")
```



Next, lets took at the US as a whole to understand how the deaths per cases is trending over time.

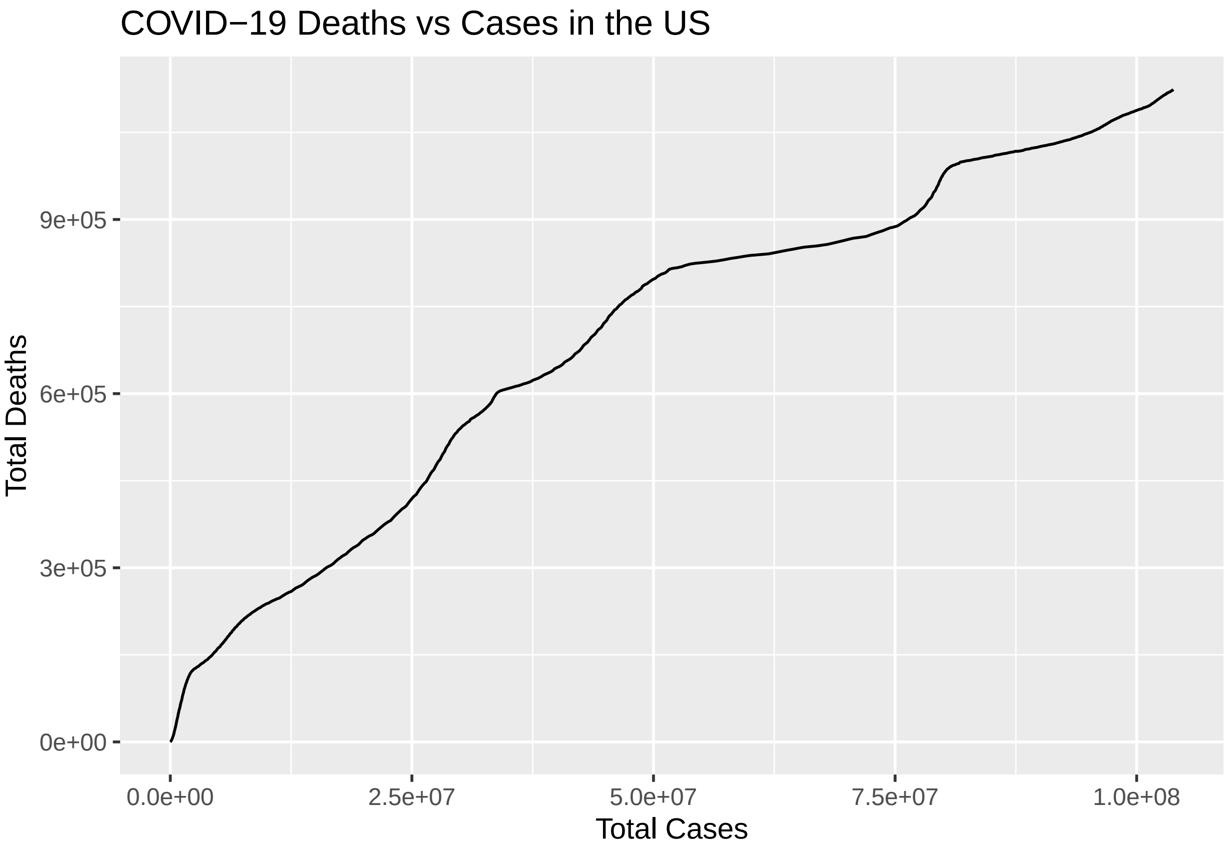
```
us_cases_and_deaths_by_date <- us_all_data %>%
  group_by(date) %>%
  summarise(total_cases = sum(cases), total_deaths = sum(deaths)) %>%
  mutate(deaths_per_cases = total_deaths / total_cases) %>%
  filter(deaths_per_cases < 1)

ggplot(us_cases_and_deaths_by_date,
       aes(x = date, y = deaths_per_cases)) +
  geom_line() +
  labs(x = "Date", y = "Deaths per Cases", title = "COVID-19 Deaths per Cases in the US Over Time")
```



It looks like when using at the entire United States, the trend is similar to what happened in California. The virus appears very deadly in the beginning but sometime after 2022, it begins to consistently stay lower.

Finally, lets see if the number of cases have any relationship with the number of deaths. A model will be created for it later.

```
ggplot(us_cases_and_deaths_by_date,
       aes(x = total_cases, y = total_deaths)) +
  geom_line() +
  labs(x = "Total Cases", y = "Total Deaths", title = "COVID-19 Deaths vs Cases in the US")
```



It looks like there is a relationship. So as the number of cases increases, the number of deaths do as well.

Model

Now a model will be created that will attempt to predict the total number of deaths, based on the total number of cases.

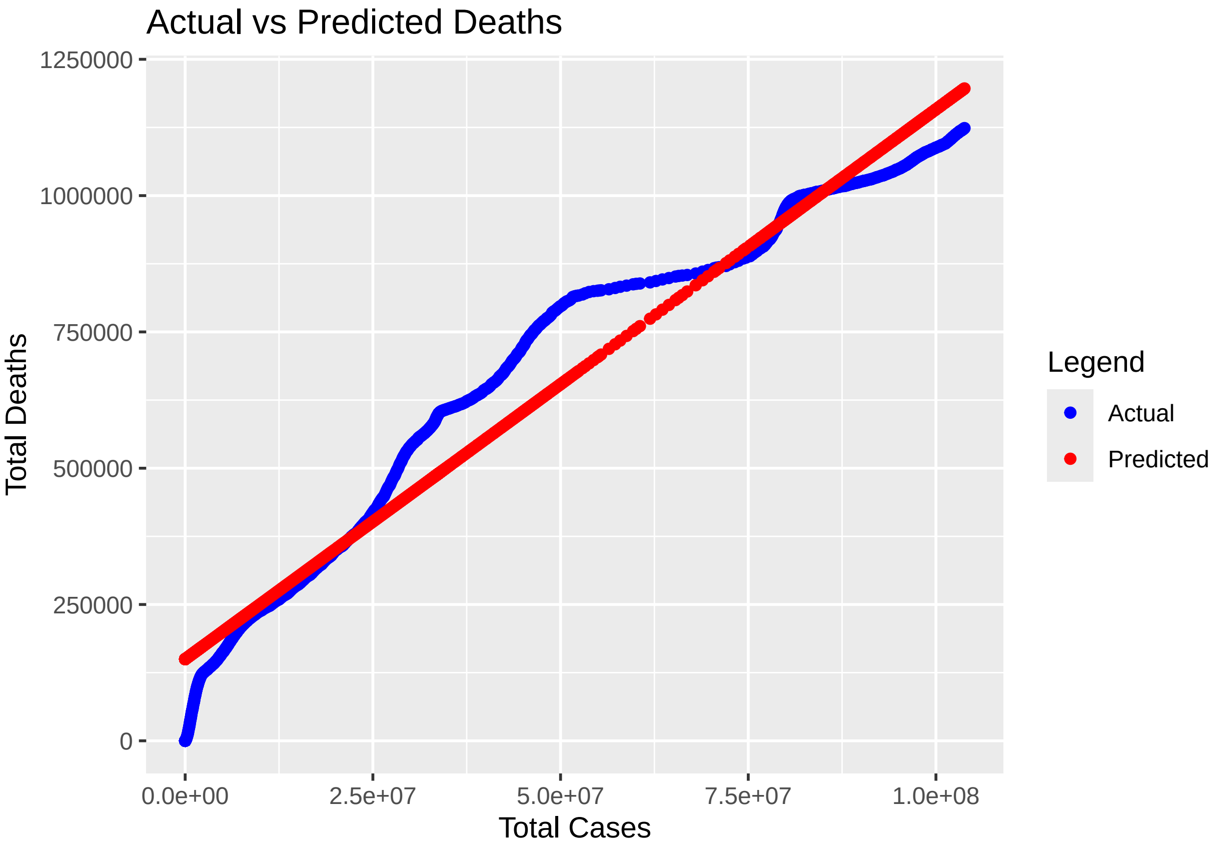
```r
mod <- lm(total_deaths ~ total_cases, data= us_cases_and_deaths_by_date)

summary(mod)
```

```
##
## Call:
## lm(formula = total_deaths ~ total_cases, data = us_cases_and_deaths_by_date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149806  -62016  -13342   89500  143891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.498e+05  3.869e+03   38.72   <2e-16 ***
## total_cases 1.008e-02  6.498e-05  155.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80090 on 1141 degrees of freedom
## Multiple R-squared:  0.9548, Adjusted R-squared:  0.9547
## F-statistic: 2.408e+04 on 1 and 1141 DF,  p-value: < 2.2e-16
```

The results from the linear model look good. The p-value is < 0.05 and the R-squared is close to 1.

Now we can visualize the actual vs predicted deaths.

```
us_cases_and_deaths_by_date <- us_cases_and_deaths_by_date %>% mutate(pred_deaths=predict(mod))

us_cases_and_deaths_by_date %>%
  ggplot() +
  geom_point(aes(x = total_cases, y = total_deaths, color = "Actual")) +
  geom_point(aes(x = total_cases, y = pred_deaths, color = "Predicted")) +
  labs(title = "Actual vs Predicted Deaths",
       x = "Total Cases",
       y = "Total Deaths") +
  scale_color_manual(values = c("Actual" = "blue", "Predicted" = "red"),
                     labels = c("Actual", "Predicted")) +
  guides(color = guide_legend(title = "Legend"))
```



The model appears to fit well enough, but isn't very good at the extreme ends of the range or the middle. This is likely cause by external factors such as vaccines since the cases can increase drastically without leading to more deaths.

Conclusions and Bias Identification

In conclusion, COVID-19 started off a very deadly disease but as time went on and we learned more about its behavior, we were able to reduce its deadliness. Vaccines were also a likely contributor to the drastic decrease in deaths and cases. We did not experience an increase in deaths or cases near the beginning of 2023 like what was experienced for the past 2 years.

One source of bias in this data would be in how the data was gathered. The cases would be under represented by some amount since not everyone will report that they got the virus and would have likely just stayed at home. The number of deaths have a better chance of being more correct since any cause of death would eventually be reported.

In my analysis, I also only looked at the US since that is where I am based. The virus could have a wildly different behavior or outcome if it was in another country. John Hopkins does have more data on COVID-19 from other countries available. There could also be a misrepresentation of cases and deaths in this data since countries have different reporting practices.

Something else to keep in mind when thinking about bias, is that more populated areas would naturally have more cases and deaths. Data that is aggregated at the state level, may not be representative for someone in a less populated location of that same state.

Show Session Info

```r
sessionInfo()
```

```
## R version 4.3.3 (2024-02-29)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.0   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.4      highr_0.10        crayon_1.5.2
##  [5] compiler_4.3.3    tidyselect_1.2.1  parallel_4.3.3    scales_1.3.0
##  [9] yaml_2.3.8        fastmap_1.1.1     R6_2.5.1          labeling_0.4.3
## [13] generics_0.1.3    curl_5.2.1        knitr_1.45        munsell_0.5.0
## [17] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4
## [21] stringi_1.8.3     xfun_0.43         bit64_4.0.5       timechange_0.3.0
## [25] cli_3.6.2         withr_3.0.0       magrittr_2.0.3    digest_0.6.35
## [29] grid_4.3.3        vroom_1.6.5       rstudioapi_0.16.0 hms_1.1.3
## [33] lifecycle_1.0.4   vctrs_0.6.5       evaluate_0.23     glue_1.7.0
## [37] farver_2.1.1      fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.26
## [41] tools_4.3.3       pkgconfig_2.0.3   htmltools_0.5.8
```