covid19-final-project

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

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

COVID-19 Overview

This project uses the New York Times COVID-19 data and imports the global and U.S. COVID-19 data. Data exploration and analysis is performed on the U.S. COVID-19 data.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c("time_series_covid19_confirmed_global.csv","time_series_covid19_deaths_global.csv","time
urls <- str_c(url_in,file_names)
global_cases <- read_csv(urls[1])
```

```
## Rows: 285 Columns: 882
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (880): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_deaths <- read_csv(urls[2])
```

```
## Rows: 285 Columns: 882
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (880): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
US_cases <- read_csv(urls[3])
```

```
## Rows: 3342 Columns: 889
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (883): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
US_deaths <- read_csv(urls[4])
```

```
## Rows: 3342 Columns: 890
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (884): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24/...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

COVID-19 Tidying and Transformation

The following activities are performed to tidy and transform the data:

- Lat and Long columns are not required and are removed
- each date on a seperate row so that we have the cases per each date
- dates that have zero cases are removed
- cases and deaths are combined into one dataset

```
global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "cases") %>%
  select(-c(Lat,Long))

global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province/State','Country/Region', Lat, Long),
```

```
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(Lat,Long))




global_cases <- global_cases %>% filter(cases>0)

global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State') %>%
  mutate(date = mdy(date))
```

```
## Joining, by = c("Province/State", "Country/Region", "date")
```

```
US_cases <- US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

US_deaths <- US_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

US <- US_cases %>%
  full_join(US_deaths)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key",
## "date")
```

```
global <- global %>%
  unite("Combined_Key",
        c(Province_State, Country_Region),
        sep = ", ",
        na.rm = TRUE,
        remove = FALSE)

US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  select(Province_State, Country_Region, date,
```

```
        cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
```

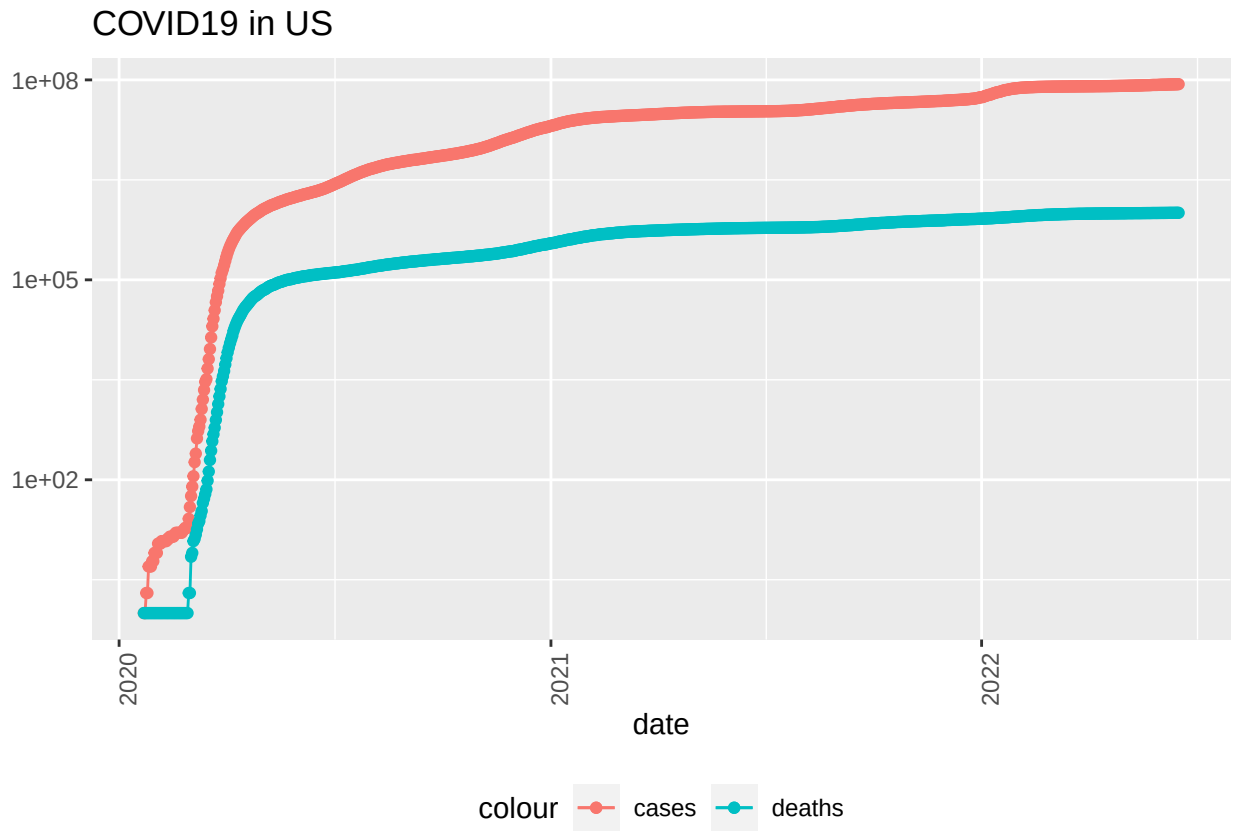
COVID-19 Data Exploration and Visualization

Plotting the total number of cases in the U.S. along side the total number of deaths would indicate that the percentages of cases that are deaths remains fairly constant.

```
US_totals <- US_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Country_Region'. You can override using the
## `.groups` argument.
```

```
US_totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y=deaths, color="deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID19 in US", y = NULL)
```




```
state <- "New York"
US_by_state <- US_by_state %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

US_totals <- US_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))
```

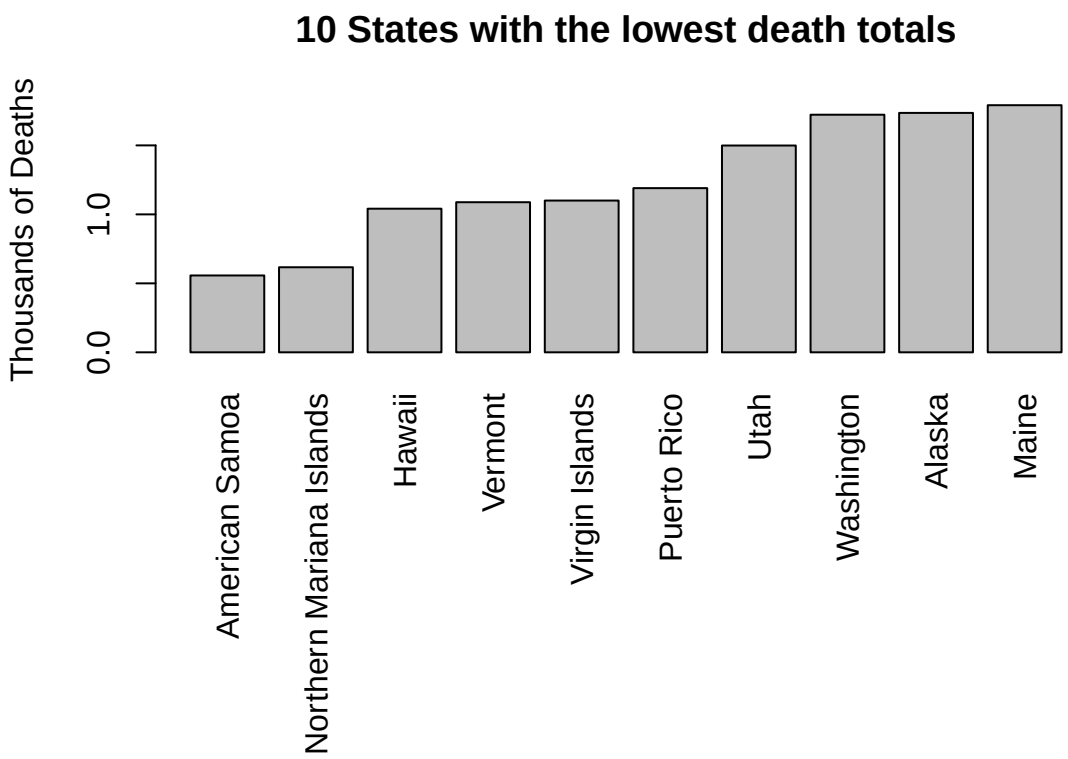
The three states with the highest number of deaths are:

- Mississippi
- Arizona
- Oklahoma

The three states with the lowest number of deaths are : * American Samoa * Northern Mariana Island * Hawaii

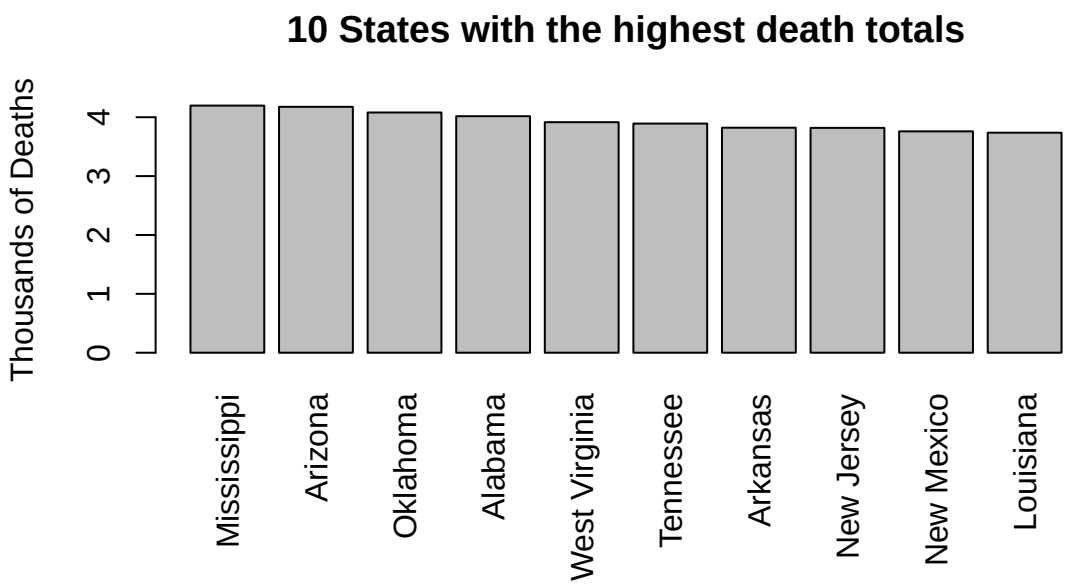
```
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thou = 1000 * cases / population,
            deaths_per_thou = 1000 * deaths / population) %>%
  filter(cases > 0, population > 0)
```

```r
par(mar=c(12,4,4,4))
# best 10 states
best_ten <- US_state_totals %>%
  slice_min(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
barplot(best_ten$deaths_per_thou, names.arg=best_ten$Province_State, ylab="Thousands of Deaths", main="
```




```r
# worst 10 states
worst_ten = US_state_totals %>%
  slice_max(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())


barplot(worst_ten$deaths_per_thou,names.arg=worst_ten$Province_State, ylab="Thousands of Deaths", main=
```



COVID-19 Data Modeling At the U.S. level, the first visualization graph plotting both cases and deaths would indicate that total number of cases could help predict total number of deaths, it is shown in this section that at the state level this is connect be shown (at not with the linear regression model that is used). The Adjusted R squared is only 0.278 which would indicate that at the state level, the number of cases is a poor predictor of deaths and the graphs plotted the actually number of deaths and the predicted deaths show that the linear lined plotted by the predictions does not fit the actual deaths very well. This would indicate that there are significant other reasons as to why the number of deaths vary significantly between states. This other factors could be number of hospitals, access to resperators, or government public health policies.

```
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2592 -0.5754  0.1064  0.7008  1.1840
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.199988   0.653028  -0.306    0.761
## cases_per_thou  0.011653   0.002494   4.673 2.01e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8217 on 54 degrees of freedom
## Multiple R-squared:  0.288,  Adjusted R-squared:  0.2748
## F-statistic: 21.84 on 1 and 54 DF,  p-value: 2.014e-05
```

```
x_grid <- seq(1,380)
new_df <- tibble(cases_per_thou = x_grid)
US_tot_w_pred <- US_state_totals %>% mutate(pred = predict(mod))

US_tot_w_pred %>% ggplot() +
  geom_point(aes(x= cases_per_thou, y = deaths_per_thou), color="blue") +
  geom_point(aes(x= cases_per_thou, y = pred), color="red")
```