5301_project_2

HB

2024-03-03

Introduction

This document reviews **COVID19 Data Report** based on a dataset retrieved from Johns Hopkins University data repository.The purpose of this document is to review the Covid19 data and produce educational visuals and models. The general trend of change in the number of cases and deaths are analysed throughout 2020-2023 followed by a deep dive into the states of Washington and California. Finally, the relationship between the population is analyzed to create a predictionb model.

Requirements

The following libraries are used in this module: tidyverse, zoo, dplyr, ggplot2,

Importing the Data

To keep this analysis reproducible, data is directly imported into the environment from the source repository. Since this dataset covers a lot of details. It needs to go through filtering and conditioning before analysis. In the next few steps the dataset is prepared in a more plot friendly format. Here is a sample of unfiltered data of US cases:

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi

file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")
urls <- str_c(url_in, file_names)
US_cases <- read_csv(urls[1])
US_deaths <- read_csv(urls[2])
head(US_cases,5)
```

```
## # A tibble: 5 x 1,154
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## # i 1,145 more variables: Long_ <dbl>, Combined_Key <chr>, ‘1/22/20‘ <dbl>,
## #   ‘1/23/20‘ <dbl>, ‘1/24/20‘ <dbl>, ‘1/25/20‘ <dbl>, ‘1/26/20‘ <dbl>,
## #   ‘1/27/20‘ <dbl>, ‘1/28/20‘ <dbl>, ‘1/29/20‘ <dbl>, ‘1/30/20‘ <dbl>,
## #   ‘1/31/20‘ <dbl>, ‘2/1/20‘ <dbl>, ‘2/2/20‘ <dbl>, ‘2/3/20‘ <dbl>,
## #   ‘2/4/20‘ <dbl>, ‘2/5/20‘ <dbl>, ‘2/6/20‘ <dbl>, ‘2/7/20‘ <dbl>,
## #   ‘2/8/20‘ <dbl>, ‘2/9/20‘ <dbl>, ‘2/10/20‘ <dbl>, ‘2/11/20‘ <dbl>,
## #   ‘2/12/20‘ <dbl>, ‘2/13/20‘ <dbl>, ‘2/14/20‘ <dbl>, ‘2/15/20‘ <dbl>, ...
```

As shown above, the raw data is not much plot friendly, thus to get a cleaner data the following steps are done: - Removing unnecessary columns “Lat” and “Long” - Pivoting the date columns - Merging overall cases along with the number of deaths - Removing data for dates with zero cases to keep the focus on the more valuable data Here is a snapshot of the cleaned US cases and death data:

```
## # A tibble: 5 x 6
##   Admin2  Province_State date       cases Population deaths
##   <chr>   <chr>          <date>     <dbl>      <dbl>  <dbl>
## 1 Autauga Alabama        2020-03-24     1      55869      0
## 2 Autauga Alabama        2020-03-25     5      55869      0
## 3 Autauga Alabama        2020-03-26     6      55869      0
## 4 Autauga Alabama        2020-03-27     6      55869      0
## 5 Autauga Alabama        2020-03-28     6      55869      0
```

```
##     Admin2          Province_State          date                cases
##  Length:3468325     Length:3468325     Min.   :2020-01-22   Min.   :      1
##  Class :character   Class :character   1st Qu.:2020-12-27   1st Qu.:    690
##  Mode  :character   Mode  :character   Median :2021-09-20   Median :   2852
##                                        Mean   :2021-09-19   Mean   :  15502
##                                        3rd Qu.:2022-06-15   3rd Qu.:   9347
##                                        Max.   :2023-03-09   Max.   :3710586
##    Population           deaths
##  Min.   :       0   Min.   :    0.0
##  1st Qu.:   10953   1st Qu.:   10.0
##  Median :   26234   Median :   47.0
##  Mean   :  104571   Mean   :  205.4
##  3rd Qu.:   67997   3rd Qu.:  137.0
##  Max.   :10039107   Max.   :35545.0
```

Check Data Validity

Looking at the summary for the US_all_clean, it is surprisingly shown that the minimum number of deaths are negative on the US data set. Therefore a filer is applied to remove the invalid chunk of data. Checking the summary as shown below:

```
##     Admin2          Province_State          date                cases
##  Length:3420617     Length:3420617     Min.   :2020-01-22   Min.   :      1
##  Class :character   Class :character   1st Qu.:2020-12-27   1st Qu.:    699
##  Mode  :character   Mode  :character   Median :2021-09-20   Median :   2863
##                                        Mean   :2021-09-19   Mean   :  15563
##                                        3rd Qu.:2022-06-15   3rd Qu.:   9347
##                                        Max.   :2023-03-09   Max.   :3710586
##    Population           deaths
##  Min.   :      86   Min.   :    0.0
##  1st Qu.:   11663   1st Qu.:   10.0
##  Median :   26794   Median :   47.0
##  Mean   :  106029   Mean   :  204.3
##  3rd Qu.:   69761   3rd Qu.:  138.0
##  Max.   :10039107   Max.   :35545.0
```

When looking at the number of cases and deaths an important reference is the population on which this data has been retrieved. Thus comparing data without knowing the population is not valid. Fortunately the population data was already included in US deaths.

At this point the data is imported, cleaned, and validated. Therefore it is ready for analysis and modeling.

Analyzing the US cases and Death

As mentioned before, to have a valid comparison among the number of cases or number of death between different areas, those numbers need to be unified based on the population. This can be done by looking at cases per million population as shown below.

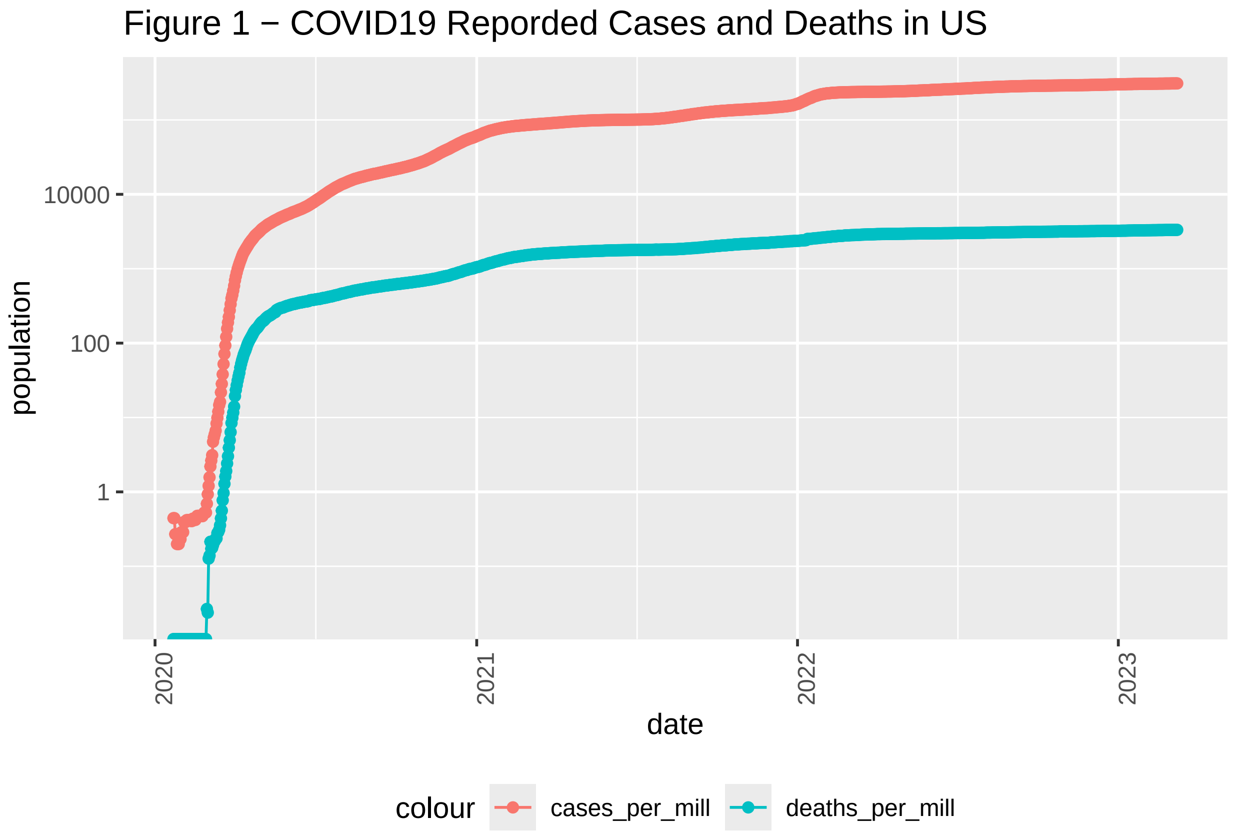
```
US_by_state <- US_all_clean %>%
    group_by(Province_State, date) %>%
    summarize(cases= sum(cases), deaths= sum(deaths), Population= sum(Population)) %>%
    mutate(deaths_per_mill= deaths* 1000000 / Population) %>%
    mutate(cases_per_mill= cases* 1000000 / Population) %>%
    ungroup()
head(US_by_state,5)
```

```
## # A tibble: 5 x 7
##   Province_State date       cases deaths Population deaths_per_mill
##   <chr>          <date>     <dbl>  <dbl>      <dbl>           <dbl>
## 1 Alabama        2020-03-11     3      0     534756               0
## 2 Alabama        2020-03-12     4      0     907665               0
## 3 Alabama        2020-03-13     8      0    1647447               0
## 4 Alabama        2020-03-14    15      0    1647447               0
## 5 Alabama        2020-03-15    28      0    2252925               0
## # i 1 more variable: cases_per_mill <dbl>
```

```
US_totals <- US_all_clean %>%
    group_by(date) %>%
    summarize(cases= sum(cases), deaths= sum(deaths), Population= sum(Population)) %>%
    mutate(deaths_per_mill= deaths * 1000000 / Population) %>%
    mutate(cases_per_mill= cases* 1000000 / Population) %>%
    select(date,cases,deaths, cases_per_mill, deaths_per_mill, Population) %>%
    ungroup()
head(US_totals,5)
```

```
## # A tibble: 5 x 6
##   date       cases deaths cases_per_mill deaths_per_mill Population
##   <date>     <dbl>  <dbl>          <dbl>           <dbl>      <dbl>
## 1 2020-01-22     1      0          0.444               0    2252782
## 2 2020-01-23     1      0          0.444               0    2252782
## 3 2020-01-24     2      0          0.270               0    7403015
## 4 2020-01-25     2      0          0.270               0    7403015
## 5 2020-01-26     5      0          0.199               0   25103228
```

```
US_totals %>%
    filter(cases_per_mill > 0) %>%
    ggplot(aes(x = date, y= cases_per_mill)) +
    geom_line(aes(color= "cases_per_mill")) +
    geom_point(aes(color= "cases_per_mill")) +
    geom_line(aes(y= deaths_per_mill, color= "deaths_per_mill")) +
    geom_point(aes(y= deaths_per_mill, color= "deaths_per_mill")) +
    scale_y_log10() +
    theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
    labs(title = "Figure 1 - COVID19 Reporded Cases and Deaths in US", y= "population")
```



As shown in figure 1, the total number of cases and deaths in US can simply be viewed over the tears. The main observation here is the big spike of the the growing number of cases during 2020 which plateaus by end of 2022. This could indicate the some measures that were implemented in US have been successful in controlling the increasing spread.

Looking more closely into the state of Washington and California:

```
wa_cases <- US_by_state %>%
    filter(Province_State == "Washington")
ca_cases <- US_by_state %>%
    filter(Province_State == "California")
wa_ca <- wa_cases %>%
  left_join(ca_cases,by=c("date"))%>%
  rename(WA_deaths_per_mill = `deaths_per_mill.x`, WA_cases_per_mill = `cases_per_mill.x`)%>%
  rename(CA_deaths_per_mill = `deaths_per_mill.y`, CA_cases_per_mill = `cases_per_mill.y`)
wa_ca
```

```
## # A tibble: 1,143 x 13
##    Province_State.x date       cases.x deaths.x Population.x WA_deaths_per_mill
##    <chr>            <date>       <dbl>    <dbl>        <dbl>              <dbl>
##  1 Washington       2020-01-22       1        0      2252782                  0
##  2 Washington       2020-01-23       1        0      2252782                  0
##  3 Washington       2020-01-24       1        0      2252782                  0
##  4 Washington       2020-01-25       1        0      2252782                  0
##  5 Washington       2020-01-26       1        0      2252782                  0
##  6 Washington       2020-01-27       1        0      2252782                  0
##  7 Washington       2020-01-28       1        0      2252782                  0
##  8 Washington       2020-01-29       1        0      2252782                  0
##  9 Washington       2020-01-30       1        0      2252782                  0
## 10 Washington       2020-01-31       1        0      2252782                  0
## # i 1,133 more rows
## # i 7 more variables: WA_cases_per_mill <dbl>, Province_State.y <chr>,
## #   cases.y <dbl>, deaths.y <dbl>, Population.y <dbl>,
## #   CA_deaths_per_mill <dbl>, CA_cases_per_mill <dbl>
```

```
wa_ca %>%
    ggplot(aes(x = date, y= WA_cases_per_mill)) +
    geom_line(aes(color= "WA_cases_per_mill")) +
    geom_point(aes(color= "WA_cases_per_mill")) +
    geom_line(aes(y= WA_deaths_per_mill, color= "WA_deaths_per_mill")) +
    geom_point(aes(y= WA_deaths_per_mill, color= "WA_deaths_per_mill")) +
    geom_line(aes(y= CA_cases_per_mill, color= "CA_cases_per_mill")) +
    geom_point(aes(y= CA_cases_per_mill, color= "CA_cases_per_mill")) +
    geom_line(aes(y= CA_deaths_per_mill, color= "CA_deaths_per_mill")) +
    geom_point(aes(y= CA_deaths_per_mill, color= "CA_deaths_per_mill")) +
    scale_y_log10() +
    theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
    labs(title = "Figure 2 - COVID19 Reporded Cases and Deaths in WA and CA", y= "population")
```

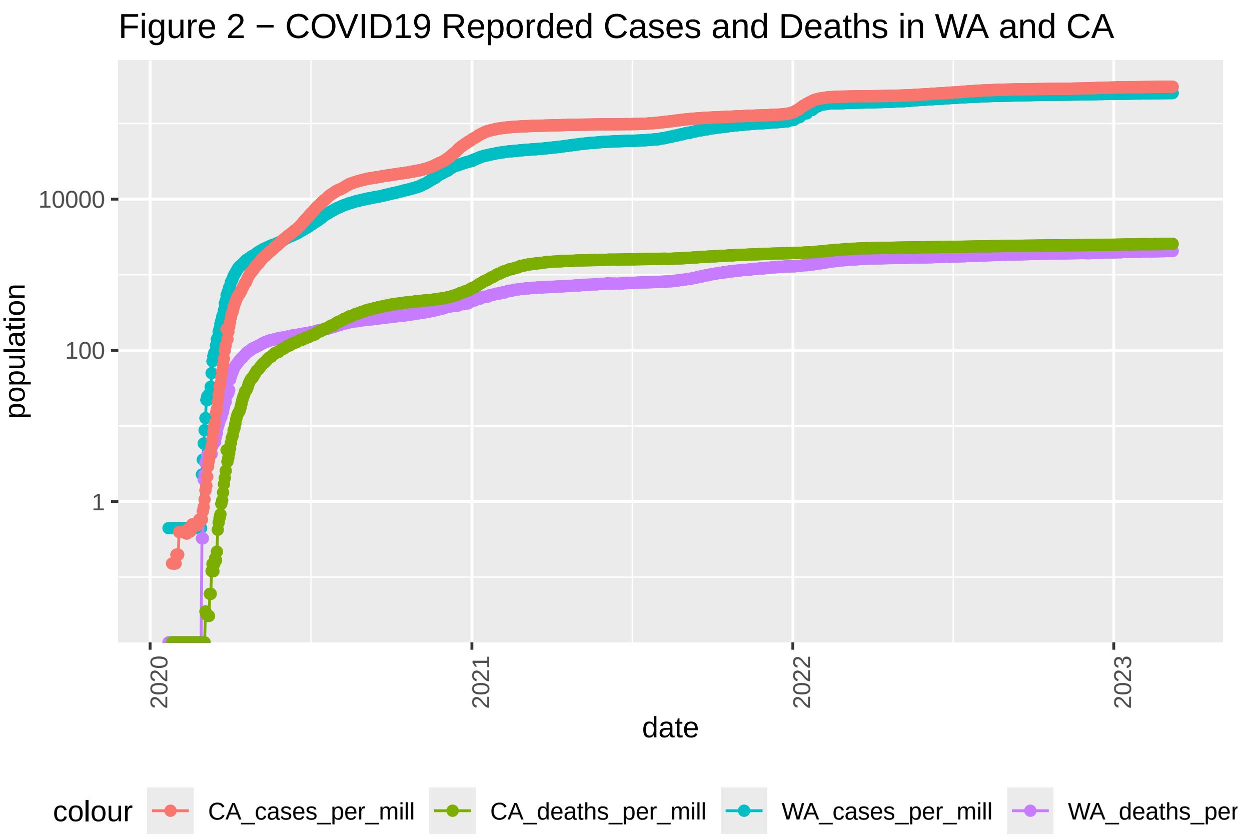
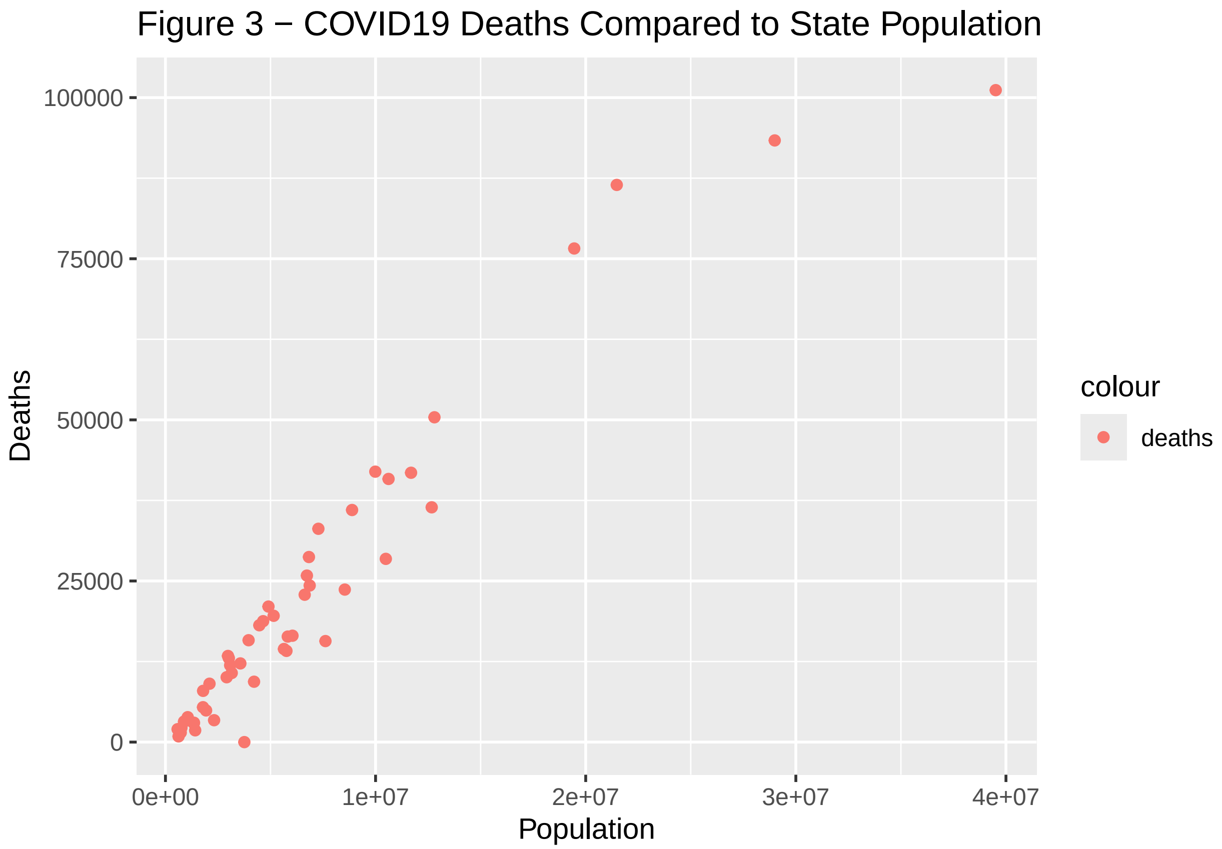


Figure 2 − COVID19 Reporded Cases and Deaths in WA and CA

Figure 2 compares the number of cases and deaths in states of California and Washington. It can be shown that both states are following a similar trend similar to the overall US cases. However, it can be pointed that Washington had a more gradual increase. This can be later studied based on the different measured deployed in each state.

Modeling the relationship between the total death and the population of states

First visualizing the existing data by summing all the death for each state



As shown in figure 3, there is a linear relation between the population of each state and the total number of deaths. Thus a linear model is created to predict the number of deaths based on any population.

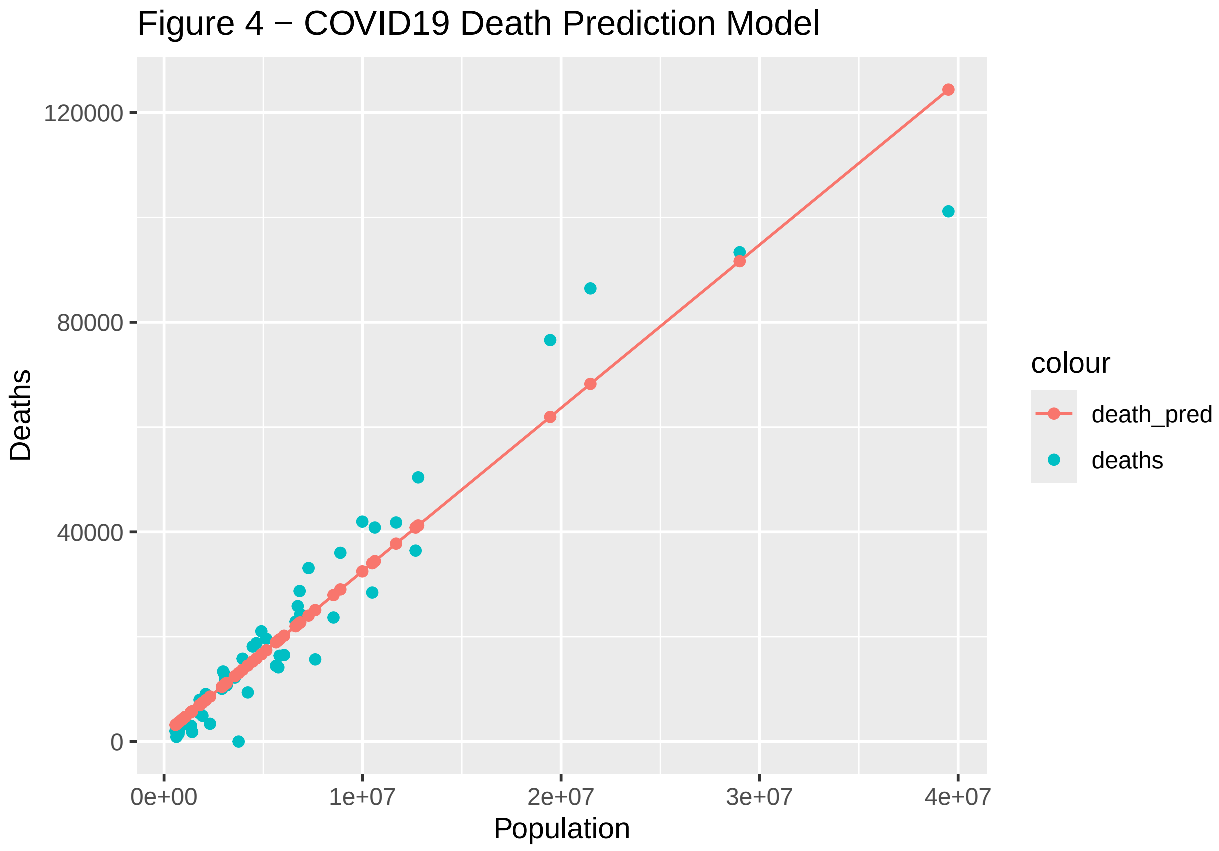
```
mod <- lm(deaths ~ Population, data = state_max)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths ~ Population, data = state_max)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23227.3  -2754.7   -638.4   2759.5  18213.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.374e+03  1.191e+03   1.153    0.254
## Population  3.113e-03  1.235e-04  25.203   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6445 on 50 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9256
## F-statistic: 635.2 on 1 and 50 DF,  p-value: < 2.2e-16
```

```
state_max <- state_max %>% mutate(death_pred = predict(mod))
head(state_max,5)
```

```
## # A tibble: 5 x 5
##   Province_State    cases deaths Population death_pred
##   <chr>             <dbl>  <dbl>      <dbl>      <dbl>
## 1 Alabama         1644533  21032    4903185     16639.
## 2 Alaska           307649   1486     728809      3643.
## 3 Arizona         2443514  33102    7278717     24035.
## 4 Arkansas         973278  13020    3017804     10769.
## 5 California     12125315 101159   39512223    124386.
```

```
state_max %>%
    ggplot(aes(x = Population, y= deaths)) +
    geom_point(aes(color= "deaths")) +
    geom_point(aes(y= death_pred, color= "death_pred")) +
    geom_line(aes(y= death_pred, color= "death_pred")) +
    labs(title = "Figure 4 - COVID19 Death Prediction Model", y= "Deaths")
```



Conclusion

By looking through the overall cases across US, figure 1 illustrated the general trend of number of cases followed by number of deaths. Further, by looking into the states of California and Washington a similar behavior was observed but in addition to the main trend, each state seems to have handled the Covid-19 experience differently at the start. Since Washington had a more gradual increase in both number of cases and deaths, the root cause can be studied as a potential successful measure for preventing such a disease in future. Finally, figure 4 has shown a prediction model for number of deaths based on the population. This model could be used to estimate any other population based on the prediction line.