In Depth Analysis of COVID 19 Data for 2 Small European Nations

Steven Lee

2023-10-15

Library Used

```
library(tidyverse)
library(lubridate)
```

Import Data

Import COVID 19 Data by the given URL as a .csv file.

```
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_series_covid19_recovered_global.csv')
urls <- str_c(url_in, file_names)
global_cases <- read_csv(urls[1])
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): 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: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): 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.
```

```r
global_recovered <- read_csv(urls[3])
```

```
## Rows: 274 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): 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.
```

Data Clean Up

Cleaning the raw data so that it's more usable. This involves omitting some column values especially those that are marked NA and name editing.

```r
#For Global cases
global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
               names_to = 'date',
               values_to = 'cases') %>%
  select(-c(Lat, Long)) %>%
  rename(province_state = 'Province/State', country_region = 'Country/Region')

#For Global deaths
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
               names_to = 'date',
               values_to = 'deaths') %>%
  select(-c(Lat, Long)) %>%
  rename(province_state = 'Province/State', country_region = 'Country/Region')

#For Combining global cases and global deaths
global <- global_cases %>%
  full_join(global_deaths) %>%
  mutate(date = mdy(date)) %>%
  filter(cases > 0)
```

```
## Joining with `by = join_by(province_state, country_region, date)`
```

```r
global <- global %>% filter(cases > 0)

summary(global)
```

```
##  province_state     country_region          date                cases          
##  Length:306827      Length:306827      Min.   :2020-01-22   Min.   :        1  
##  Class :character   Class :character   1st Qu.:2020-12-12   1st Qu.:     1316  
##  Mode  :character   Mode  :character   Median :2021-09-16   Median :    20365  
##                                        Mean   :2021-09-11   Mean   :  1032863  
##                                        3rd Qu.:2022-06-15   3rd Qu.:   271281  
##                                        Max.   :2023-03-09   Max.   :103802702  
```

```
##      deaths
##  Min.   :      0
##  1st Qu.:      7
##  Median :    214
##  Mean   :  14405
##  3rd Qu.:   3665
##  Max.   :1123836
```

```r
#Global by country
global_by_country <- global %>%
  group_by(country_region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths))
```

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

```r
#Retrieving Iceland data
Iceland <- global_by_country %>%
  filter(country_region == 'Iceland') %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths)) %>%
  select(everything()) %>% drop_na()

summary(Iceland)
```

```
##  country_region          date                 cases            deaths
##  Length:1105        Min.   :2020-02-29   Min.   :     1   Min.   :  0.00
##  Class :character   1st Qu.:2020-12-01   1st Qu.:  5413   1st Qu.: 27.00
##  Mode  :character   Median :2021-09-03   Median : 10956   Median : 33.00
##                     Mean   :2021-09-03   Mean   : 74533   Mean   : 77.27
##                     3rd Qu.:2022-06-06   3rd Qu.:188924   3rd Qu.:153.00
##                     Max.   :2023-03-09   Max.   :209137   Max.   :263.00
##    new_cases         new_deaths
##  Min.   :-273.0   Min.   :-39.000
##  1st Qu.:   0.0   1st Qu.:  0.000
##  Median :   2.0   Median :  0.000
##  Mean   : 189.3   Mean   :  0.238
##  3rd Qu.:  47.0   3rd Qu.:  0.000
##  Max.   :7408.0   Max.   : 52.000
```

```r
#Retrieving Malta data
Malta <- global_by_country %>%
  filter(country_region == 'Malta') %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths)) %>%
  select(everything()) %>% drop_na()

summary(Malta)
```

```
##  country_region          date                 cases            deaths
##  Length:1097        Min.   :2020-03-08   Min.   :     3   Min.   :  0.0
##  Class :character   1st Qu.:2020-12-07   1st Qu.: 10659   1st Qu.:155.0
```

```
##  Mode  :character   Median :2021-09-07   Median : 36553   Median :445.0
##                     Mean   :2021-09-07   Mean   : 51021   Mean   :435.1
##                     3rd Qu.:2022-06-08   3rd Qu.: 95467   3rd Qu.:723.0
##                     Max.   :2023-03-09   Max.   :117610   Max.   :828.0
##    new_cases        new_deaths
##  Min.   : -42.0   Min.   :-2.0000
##  1st Qu.:  13.0   1st Qu.: 0.0000
##  Median :  43.0   Median : 0.0000
##  Mean   : 107.2   Mean   : 0.7548
##  3rd Qu.: 123.0   3rd Qu.: 1.0000
##  Max.   :1677.0   Max.   : 7.0000
```

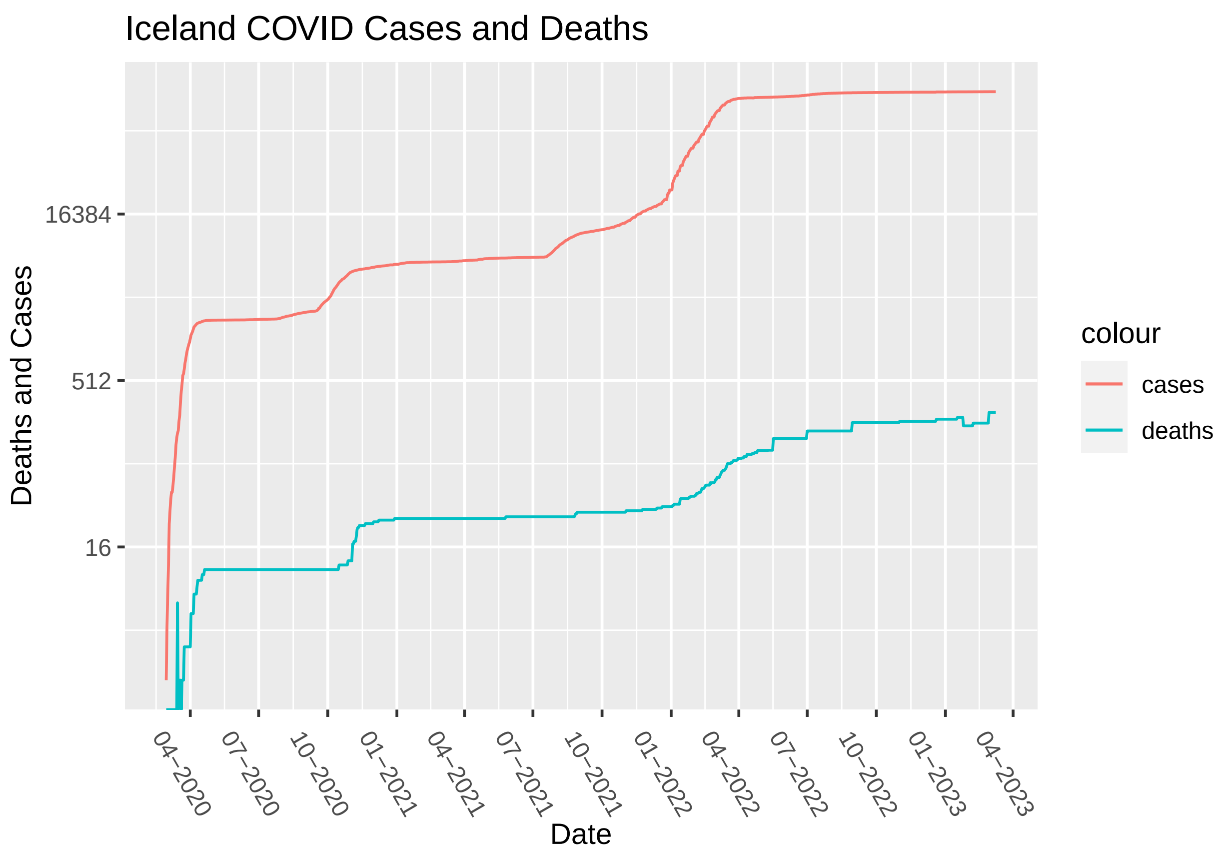
Visualizations

Now that all the data is organized, we can make some visualizations to answer some questions especially how changes in Canadian cases and deaths relate to the world as whole.

1. How many cases and deaths were there to date in Iceland?

```r
Iceland %>%
  ggplot() +
  geom_line(aes(x = date, y = deaths, color = 'deaths')) +
  geom_line(aes(x = date, y = cases, color = 'cases')) +
  scale_y_continuous(trans = 'log2') +
  labs(title = 'Iceland COVID Cases and Deaths',
       x = 'Date', y = 'Deaths and Cases') +
  scale_x_date(date_labels = '%m-%Y', date_breaks = '3 month') +
  theme(axis.text.x = element_text(angle = 300))
```

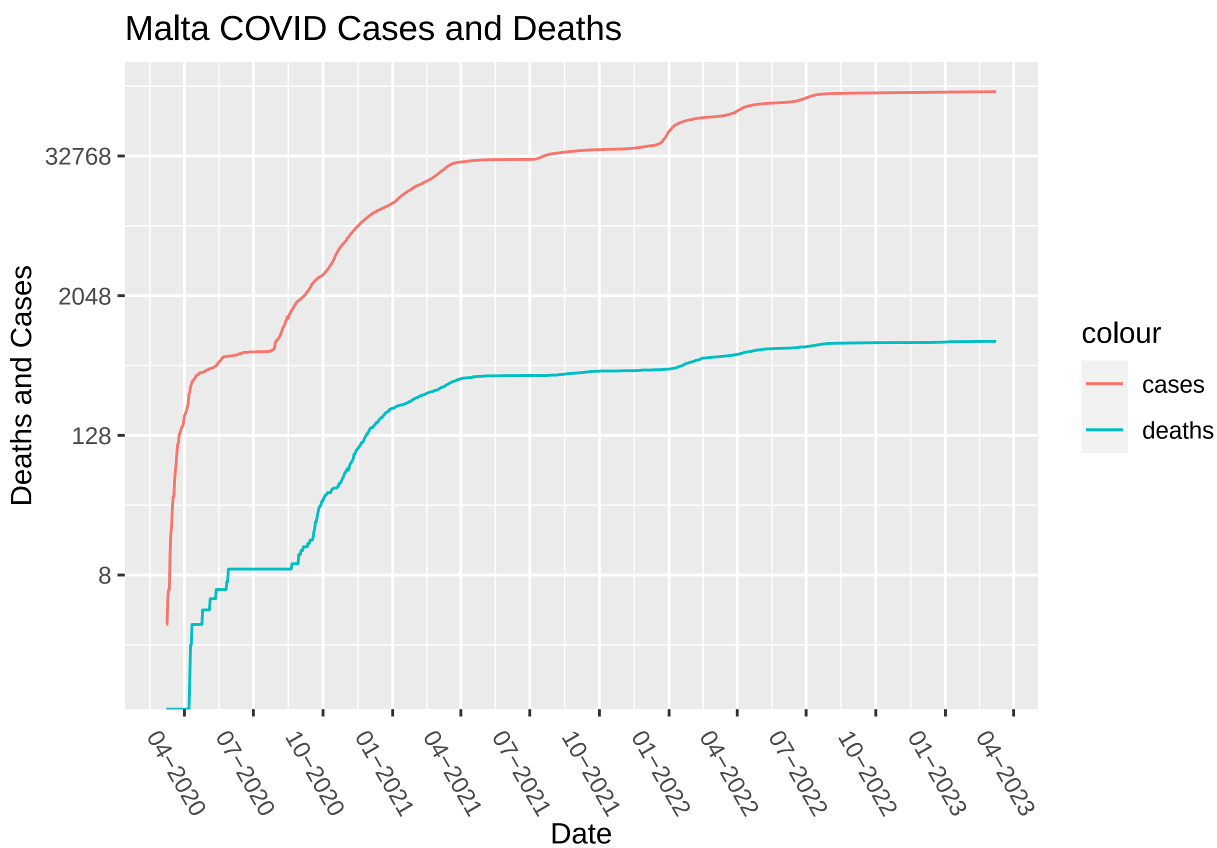
```
## Warning: Transformation introduced infinite values in continuous y-axis
```



2. How many cases and deaths were there to date in Malta?

```
Malta %>%
  ggplot() +
  geom_line(aes(x = date, y = deaths, color = 'deaths')) +
  geom_line(aes(x = date, y = cases, color = 'cases')) +
  scale_y_continuous(trans = 'log2') +
  labs(title = 'Malta COVID Cases and Deaths',
       x = 'Date', y = 'Deaths and Cases') +
  scale_x_date(date_labels = '%m-%Y', date_breaks = '3 month') +
  theme(axis.text.x = element_text(angle = 300))
```

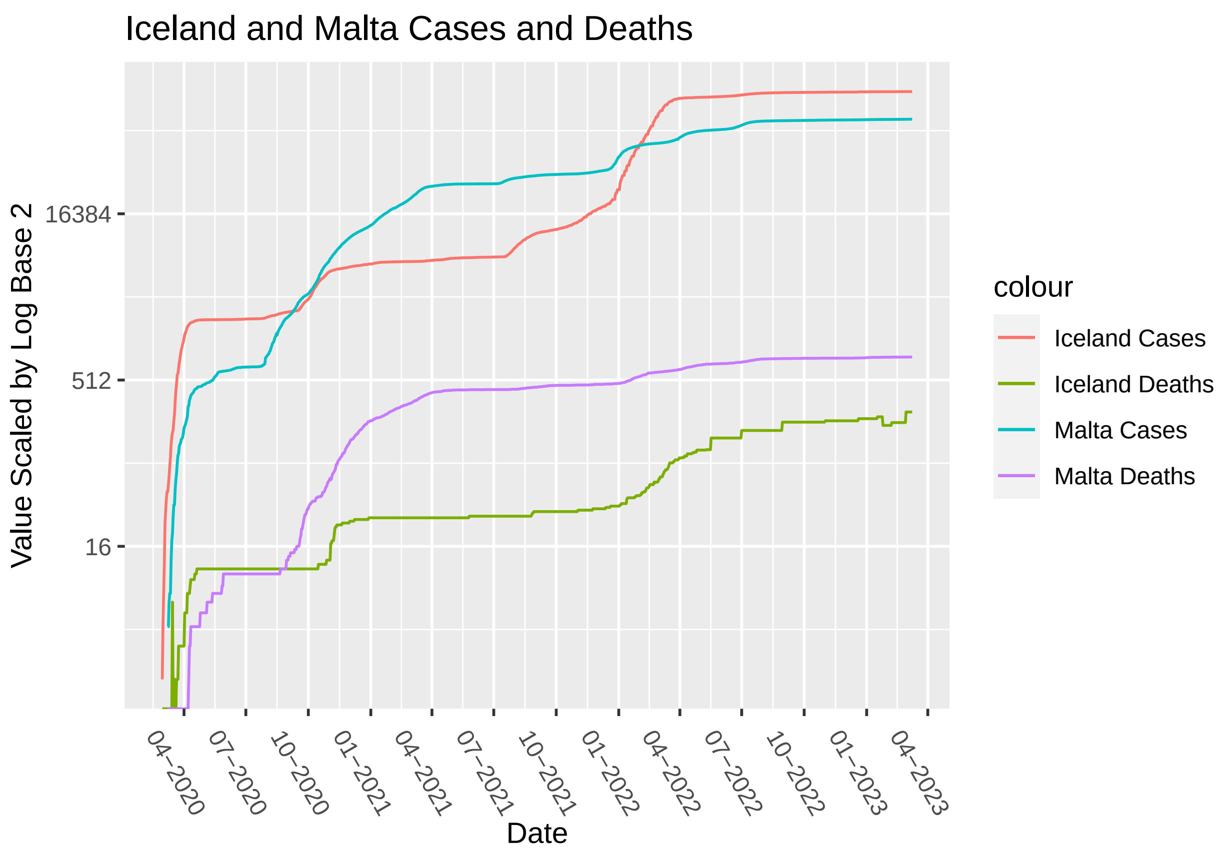
```
## Warning: Transformation introduced infinite values in continuous y-axis
```



3. How does Iceland and Malta compare to each other?

```
ggplot() +
  geom_line(data = Iceland, aes(x = date, y = deaths, color = 'Iceland Deaths')) +
  geom_line(data = Iceland, aes(x = date, y = cases, color = 'Iceland Cases')) +
  geom_line(data = Malta, aes(x = date, y = deaths, color = 'Malta Deaths')) +
  geom_line(data = Malta, aes(x = date, y = cases, color = 'Malta Cases')) +
  scale_y_continuous(trans = 'log2') +
  labs(title = 'Iceland and Malta Cases and Deaths',  x = 'Date', y = 'Value Scaled by Log Base 2') +
  scale_x_date(date_labels = '%m-%Y', date_breaks = '3 month') +
  theme(axis.text.x = element_text(angle = 300))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```



Modeling

Now that we have a basic picture of what cases look like over time in Iceland and Malta, let's dive into modeling for predicted death based on number of cases for both countries.

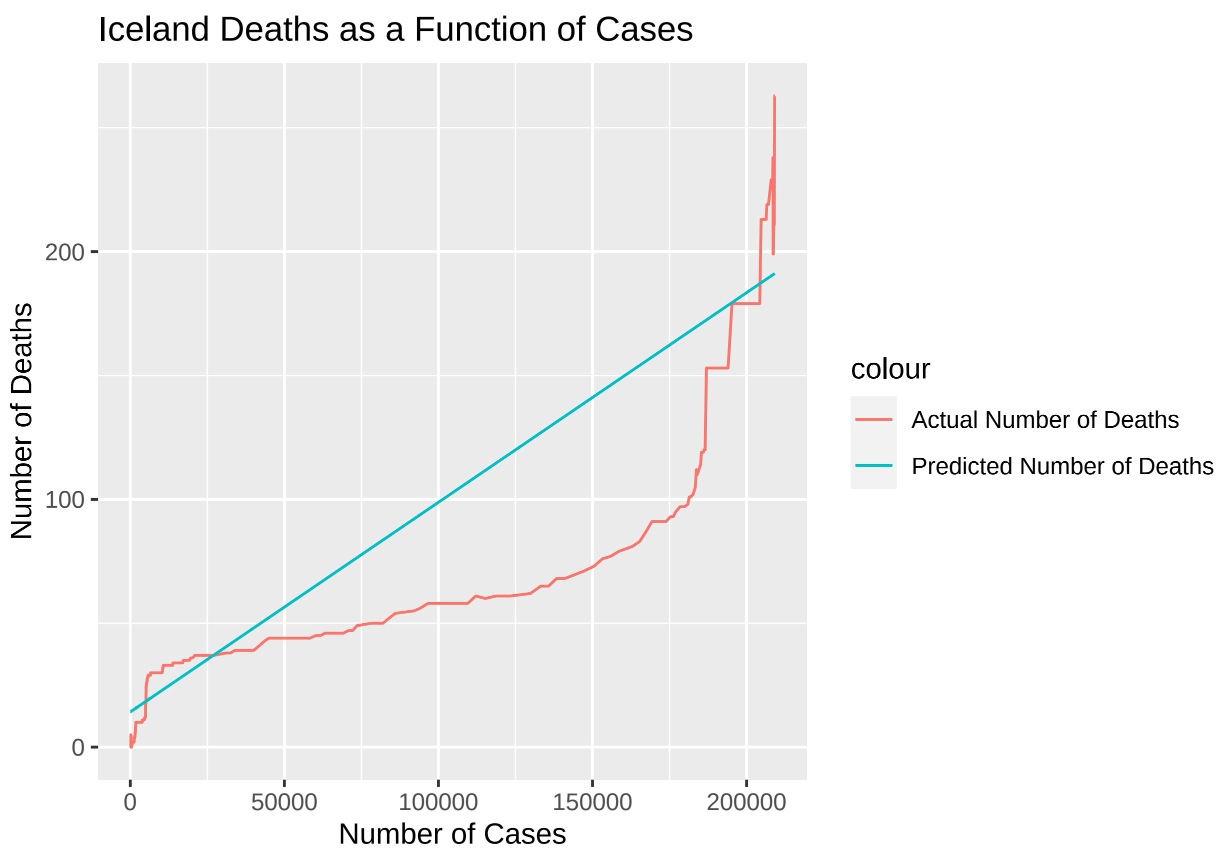
```
Iceland_death_model <- lm(deaths ~ cases, data = Iceland)

summary(Iceland_death_model)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = Iceland)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.067  -6.543   6.480   9.713  71.881
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.420e+01  9.288e-01   15.29   <2e-16 ***
## cases       8.461e-04  7.953e-06  106.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 23.77 on 1103 degrees of freedom
## Multiple R-squared:  0.9112, Adjusted R-squared:  0.9111
## F-statistic: 1.132e+04 on 1 and 1103 DF,  p-value: < 2.2e-16
```

```
Iceland <- Iceland %>%
  mutate(pred_deaths = predict(Iceland_death_model))
Iceland %>%
  ggplot() +
  geom_line(aes(x = cases, y = deaths, color = 'Actual Number of Deaths')) +
  geom_line(aes(x = cases, y = pred_deaths, color = 'Predicted Number of Deaths')) +
  labs(title = 'Iceland Deaths as a Function of Cases', x = 'Number of Cases', y = 'Number of Deaths')
```



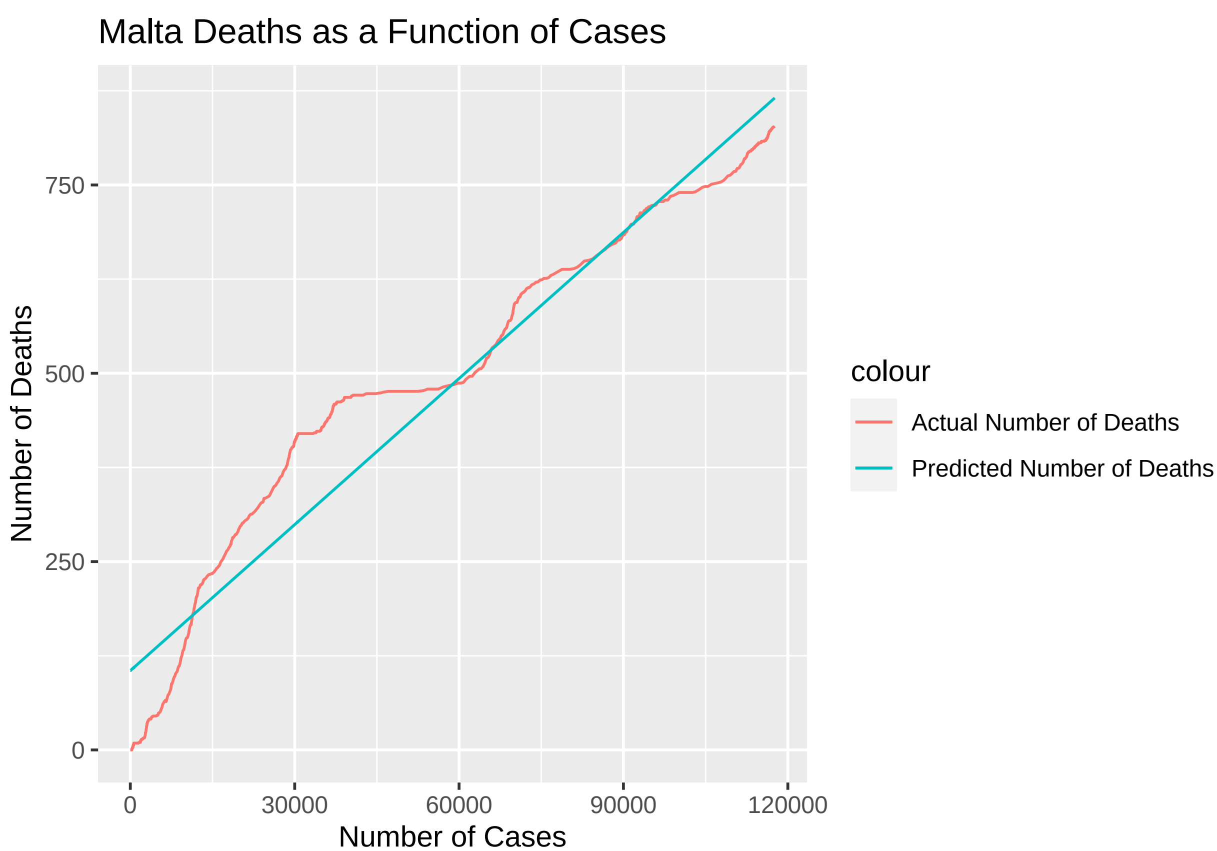

```
Malta_death_model <- lm(deaths ~ cases, data = Malta)

summary(Malta_death_model)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = Malta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.148  -48.159   -9.609   86.131  117.044
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.052e+02  3.610e+00   29.14   <2e-16 ***
## cases       6.465e-03  5.384e-05  120.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.58 on 1095 degrees of freedom
## Multiple R-squared:  0.9294, Adjusted R-squared:  0.9293
## F-statistic: 1.442e+04 on 1 and 1095 DF,  p-value: < 2.2e-16
```

```
Malta <- Malta %>%
  mutate(pred_deaths = predict(Malta_death_model))
Malta %>%
  ggplot() +
  geom_line(aes(x = cases, y = deaths, color = 'Actual Number of Deaths')) +
  geom_line(aes(x = cases, y = pred_deaths, color = 'Predicted Number of Deaths')) +
  labs(title = 'Malta Deaths as a Function of Cases', x = 'Number of Cases', y = 'Number of Deaths')
```



Based on the Linear Regression Models, we can see that Malta's trend follows very closely with the predicted outcomes based on the number of cases present. Iceland on the other hand managed to reduce deaths falling under the predicted outcome.

Bias

The possible biases present in this data include data collection, how cases were tested and deemed positive, how local climate might impact transmission rates, how the healthcare systems differ in both countries in how they manage patients with COVID 19. Responding to these biases may present a fuller picture as to how these two small European nations varied differently in how they handeld COVID 19 Deaths relative to the number of cases.