COVID19_Analysis

Engin Deniz Dogu

2024-10-05

Introduction

With this project, we will analyze the COVID-19 dataset from the Johns Hopkins Github site. There are four csv files that we are going to use:

- time_series_covid19_confirmed_global.csv
- time_series_covid19_deaths_global.csv
- time_series_covid19_confirmed_US.csv
- time_series_covid19_deaths_US.csv

csv files contain case/death numbers, country, region and date information. This project mainly uses the “global” data. At the end of this markdown, you will find a simple linear model to pretict number of deaths given the number of cases.

Link to the github page: https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/ (https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/)

Total Cases and Total Death

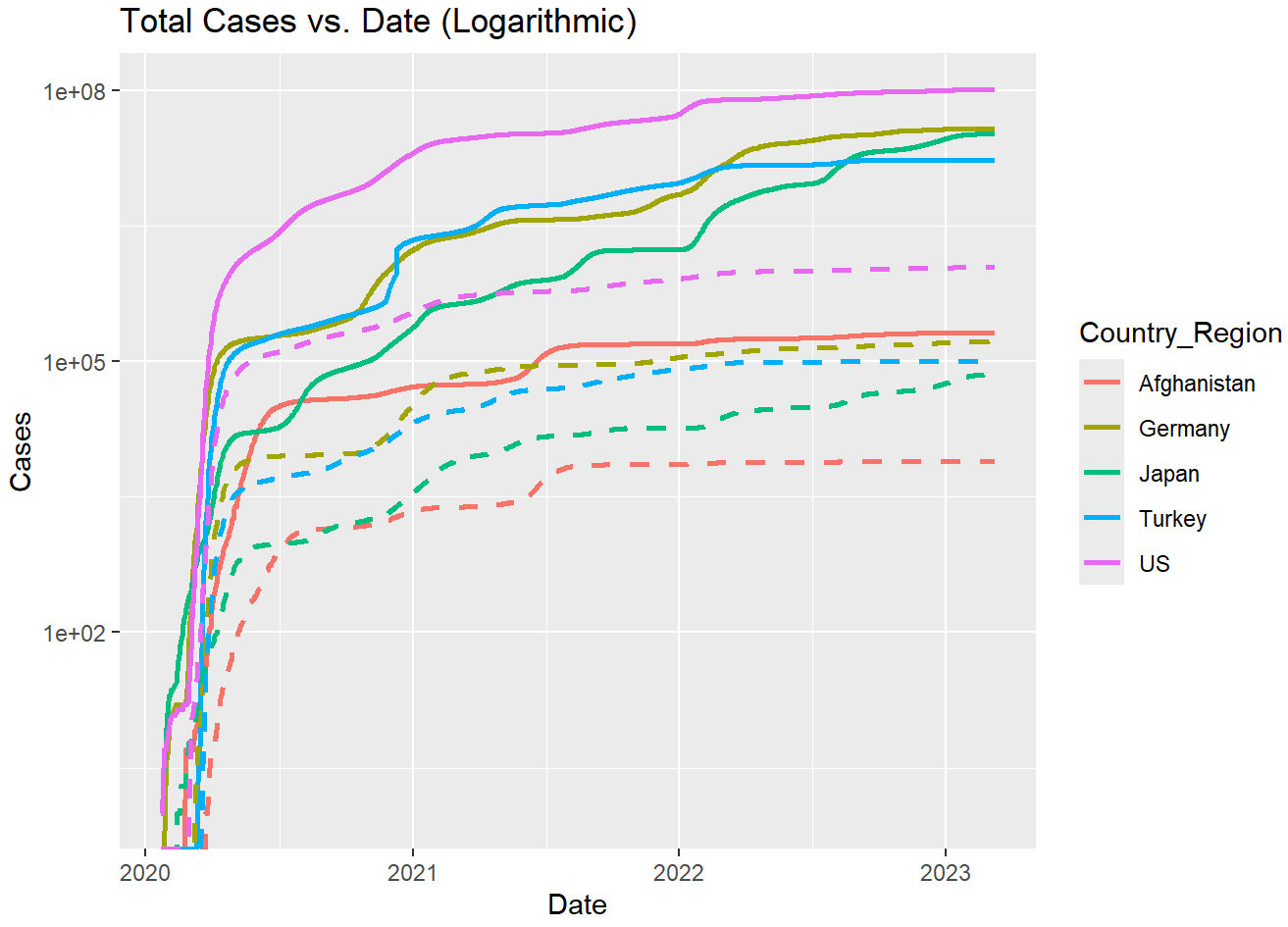
Throughout this analysis, we will mainly focus on five countries for simplicity:

- US
- Germany
- Turkey
- Afghanistan
- Japan

First, let’s look at the total number of cases (straight line) between 2020 and 2023. The dashed line shows the total number of deaths within the same period. y-axis scale is logarithmic.

```
# Filter five countries
daily_cases_of_five <- global |> filter(Country_Region %in% c("Afghanistan","Turkey","US","Germa
ny","Japan")) |> select(-c(Province_State, Combined_Key, Population))

# Plot cases and deaths in logarithmic scale
daily_cases_of_five |>
  ggplot(aes(x=date,y=cases,color=Country_Region)) +
  geom_line(linetype=1,lwd=1) +
  geom_line(aes(y=deaths,color=Country_Region),linetype=2,lwd=1) +
  scale_y_log10()+
  xlab("Date") +
  ylab("Cases") +
  ggtitle("Total Cases vs. Date (Logarithmic)")
```

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
## log-10 transformation introduced infinite values.
```



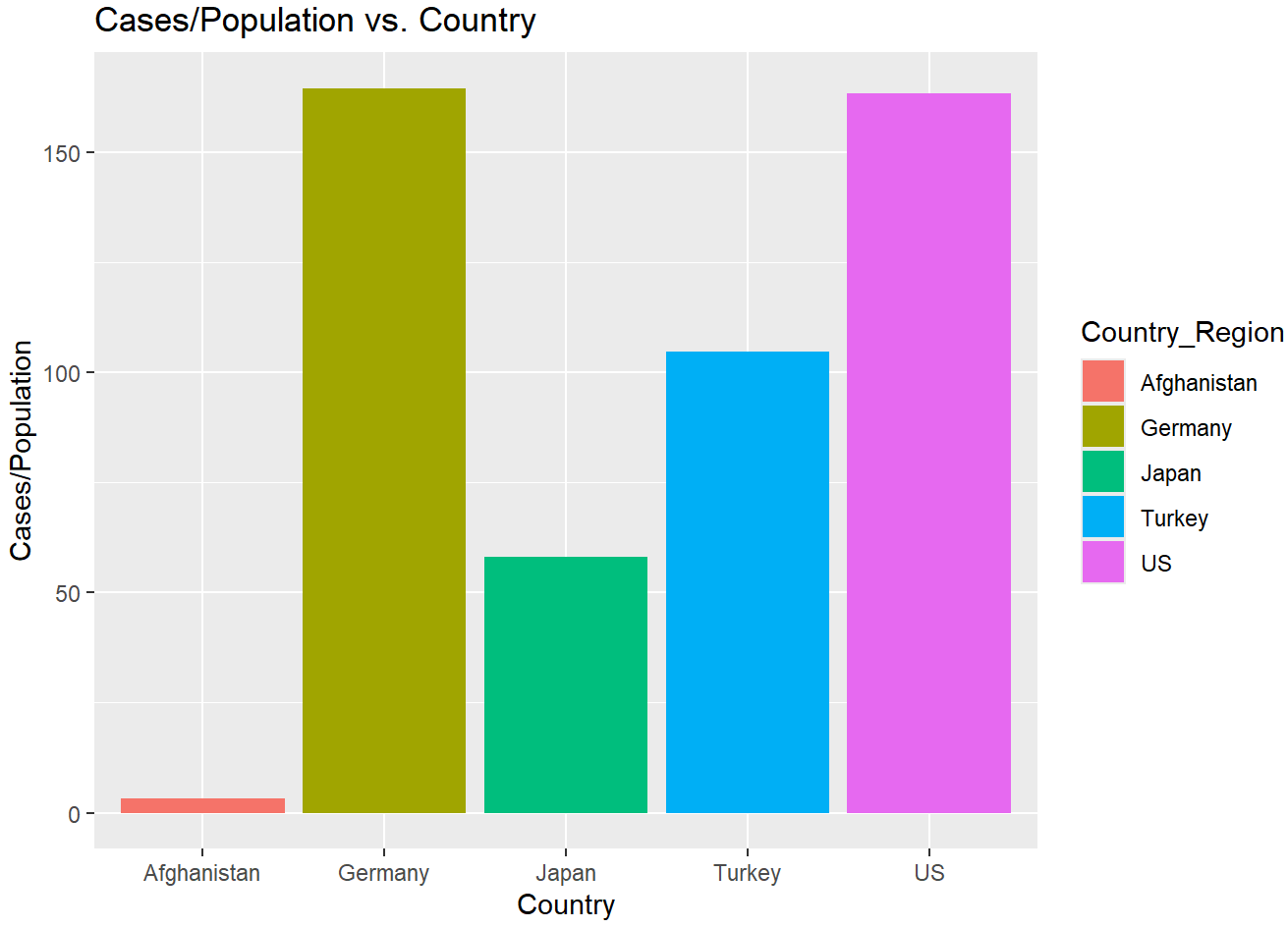
Case/Population vs. Country

Next, we look at the number of total cases. Each value is divided by the population of the country to get a better understanding of how much of the population got diagnosed. Notice some countries have a value higher than 100%. This might be due to people getting sick more than once, or the number reflect people on top of residents (tourist etc.).

```
# Filter five countries
total_cases_of_five <- global |> filter(Country_Region %in% c("Afghanistan","Turkey","US","Germa
ny","Japan")) |> select(-c(Province_State, Combined_Key)) |>
  group_by(Country_Region,Population) |>
  summarize(cases=sum(cases),deaths=sum(deaths))
```

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

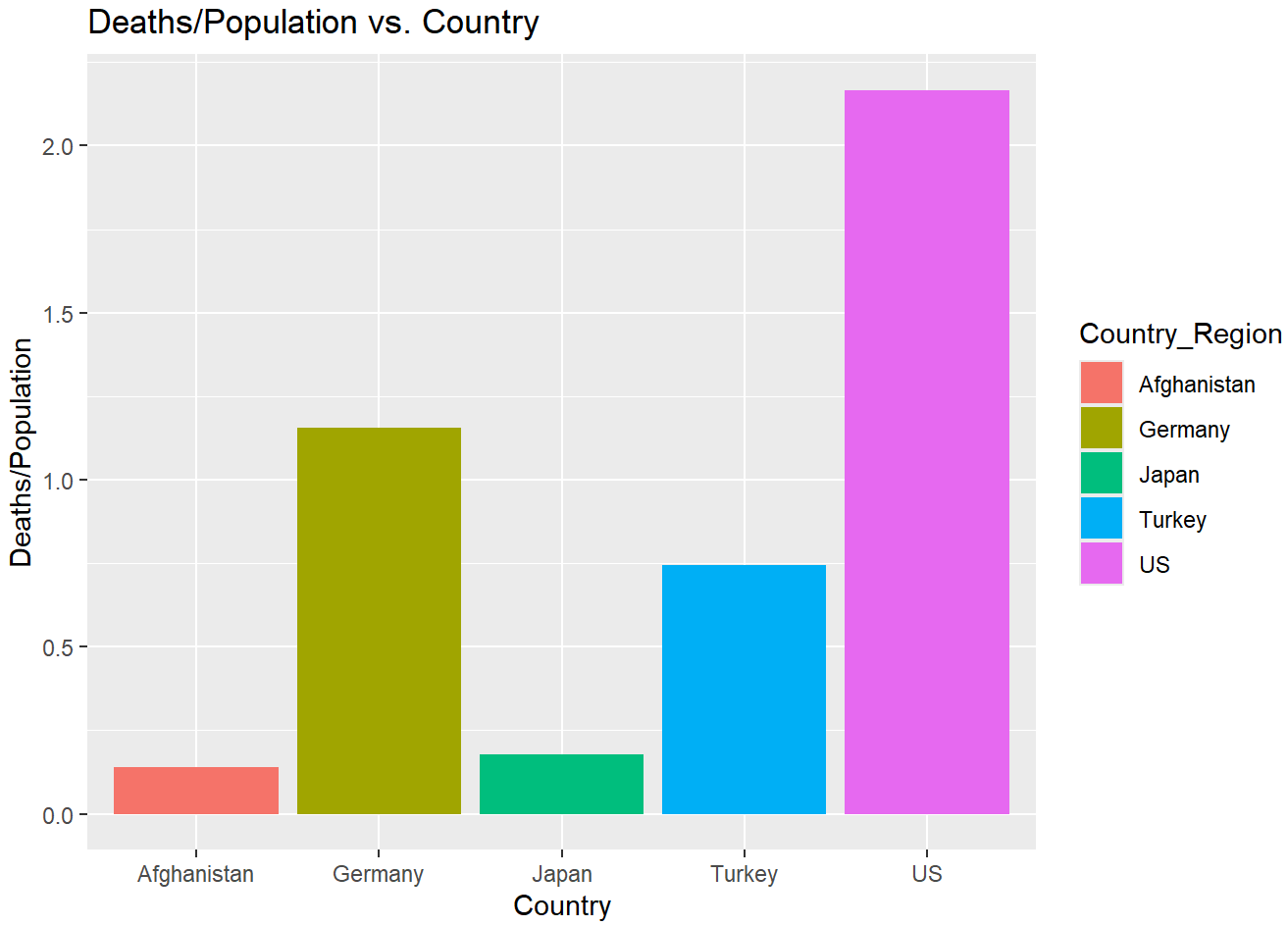
```
# Cases divided by population
total_cases_of_five <- total_cases_of_five |>
  mutate(cases_per_population = cases / Population,
         deaths_per_population = deaths / Population) |>
  select(-c(Population,cases,deaths))

# Histogram
ggplot(total_cases_of_five) +
  geom_bar(aes(x=Country_Region,y=cases_per_population,fill=Country_Region),stat="identity") +
  xlab("Country") +
  ylab("Cases/Population") +
  ggtitle("Cases/Population vs. Country")
```



Deaths/Population vs. Country

Similarly, let's look at the ratio of deaths. Compared to number of cases, it seems to be low.

```
# Histogram
ggplot(total_cases_of_five) +
  geom_bar(aes(x=Country_Region,y=deaths_per_population,fill=Country_Region),stat="identity") +
  xlab("Country") +
  ylab("Deaths/Population") +
  ggtitle("Deaths/Population vs. Country")
```



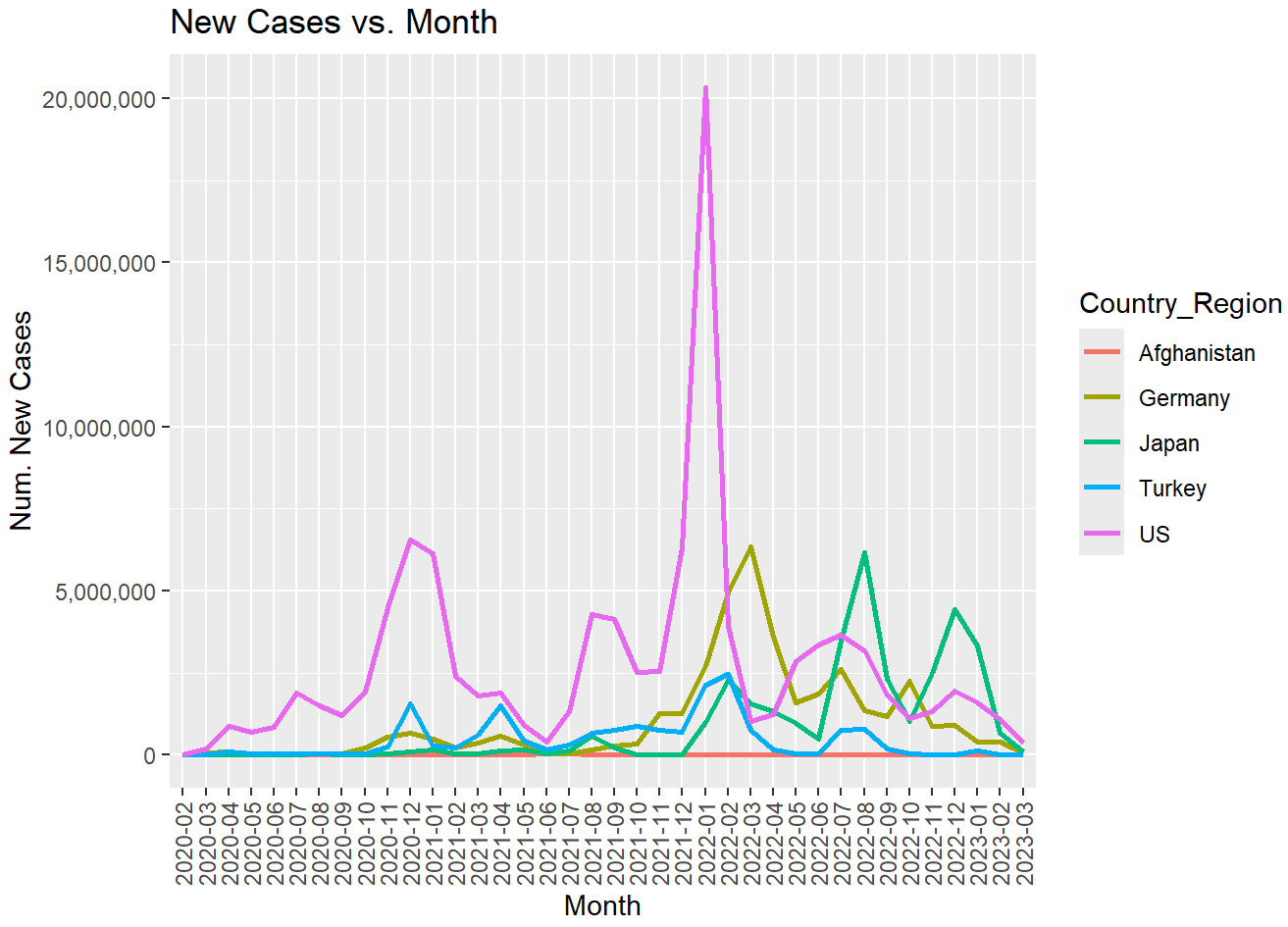
New Cases Through Months

This time we are trying yo understand how each season affects number of cases. Case numbers seem to be increasing during winter except in Japan.

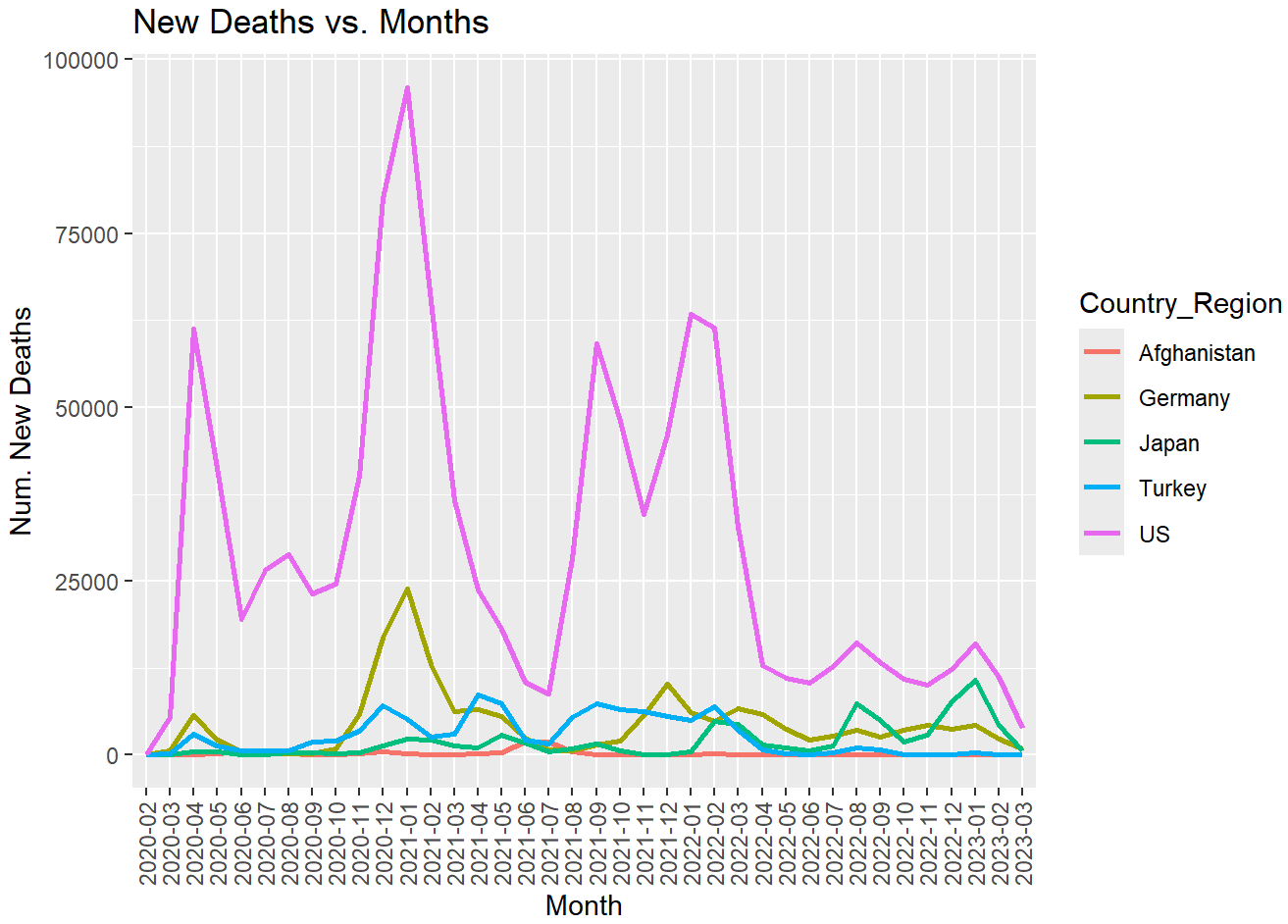
```
# Monthly changes
cases_seasonal <- global |>
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths),
         month_year = format(as.Date(global$date), "%Y-%m")) |>
  select(c(Country_Region,month_year,Population,new_cases,new_deaths)) |>
  group_by(Country_Region,month_year,Population) |>
  summarize(new_cases=sum(new_cases),new_deaths=sum(new_deaths)) |>
  filter(Country_Region %in% c("Afghanistan","Turkey","US","Germany","Japan")) |>
  filter(month_year != "2020-01")
```

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

```
cases_seasonal |>
  ggplot(aes(x=month_year,y=new_cases,group=Country_Region,color=Country_Region)) +
  geom_line(linetype=1,lwd=1) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(name="Num. New Cases", labels = scales::comma) +
  xlab("Month") +
  ggtitle("New Cases vs. Month")
```



```
# Monthly changes
cases_seasonal |>
  ggplot(aes(x=month_year,y=new_deaths,group=Country_Region,color=Country_Region)) +
  geom_line(linetype=1,lwd=1) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  xlab("Month") +
  ylab("Num. New Deaths") +
  ggtitle("New Deaths vs. Months")
```



Modelling

At last, we will try to come up with a model that predicts the number of deaths given the total number of cases. Inputs to this model are the total number of cases and the total number of deaths in each country. Not surprisingly, the model tells us that if there are more cases, there will be more deaths. While this is an over-simplification, there are also points where the model overestimates or underestimates the number of deaths.

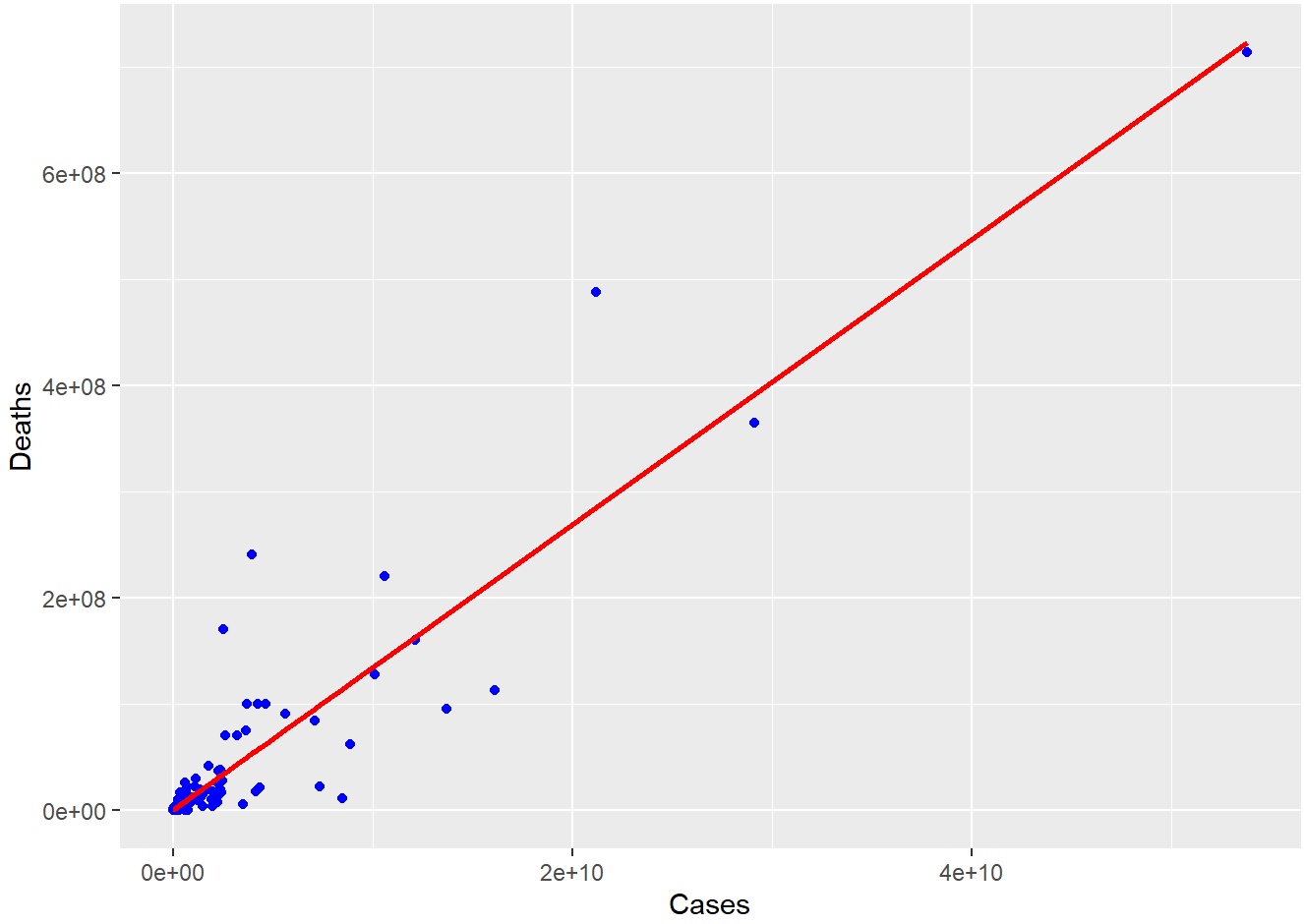
```
# Total cases per population of all countries
total_cases_deaths <- global |>
  select(-c(Province_State, Combined_Key)) |>
  group_by(Country_Region,Population) |>
  summarize(cases=sum(cases),deaths=sum(deaths)) |>
  group_by(Country_Region) |> #group by twice because some rows differentiate depending     on c
ountry/region, then we group by again
  summarize(population=sum(Population),cases=sum(cases),deaths=sum(deaths)) |>
  filter(!is.na(population))
```

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

```
model <- lm(deaths ~ cases, data=total_cases_deaths)
summary(model)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = total_cases_deaths)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -103711623   -2875946    -893056    -218369  202890293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.590e+05  2.166e+06   0.397    0.692
## cases       1.343e-02  4.057e-04  33.100   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28750000 on 192 degrees of freedom
## Multiple R-squared:  0.8509, Adjusted R-squared:  0.8501
## F-statistic:  1096 on 1 and 192 DF,  p-value: < 2.2e-16
```

```
total_cases_deaths <- total_cases_deaths |> mutate(pred=predict(model))
total_cases_deaths |> ggplot() +
  geom_point(aes(x=cases,y=deaths),color="blue") +
  geom_line(aes(x=cases,y=pred),color="red",linetype=1,lwd=1) +
  xlab("Cases") +
  ylab("Deaths")
```




```
  ggtitle("Prediction")
```

```
## $title
## [1] "Prediction"
##
## attr(,"class")
## [1] "labels"
```

Summary

We have completed our analysis and created a simple model that predicts number of deaths. Before we complete this markdown, there is one thing we have to talk about and that is bias. Like all datasets and analysis, this project is also prone to bias. For example looking at the “Cases/Population vs. Country” graph, we notice that the total number of cases for Afghanistan seems to be way lower than other countries. This might be true, but we might want to double-check our dataset and sources for missing data. Or similarly, do we have a certain prejudice for COVID-19? How dangerous do we perceive it? These should be the kind of questions we keep in mind when working on our analysis.