COVID-19 Project Analysis

2024-10-13

Introduction

In this report, statistics about COVID-19 worldwide will be imported, tidied, and analyzed. Data was provided from the Johns Hopkin's Github website and includes a wide array of data points, all of which can be seen in a summary below. The two main questions that will be answered are whether the countries of residence have a large impact on overall cases as well as mortality rates.

Data Manipulation

We first will add the necessary libraries.

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(scales)
```

Next, we will input the data from Johns Hopkins

```
global_deaths <- read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/
```

```
## 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_cases <- read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/cs
```

```
## 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.
```

After that, we will make the data fit our needs. First, we will format and join the data sets of deaths and cases as well as discard regional data as we'll only be focusing on data from entire countries. Finally, we'll remove everything but the most recent date as we are only interested in totals our analysis.

```
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 %>%
  pivot_longer(cols =
-c(`Province/State`,

`Country/Region`, Lat, Long),
  names_to = "date",
  values_to = "cases") %>%
select(-c(Lat,Long))

global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country = `Country/Region`) %>%
  mutate(death_percentage = (deaths / cases * 100)) %>%
  filter( is.na(`Province/State`) ) %>%
  select(-c(`Province/State`)) %>%
  filter( date == "3/9/23") %>%
  select(-c(`date`)) %>%
  filter(Country != 'MS Zaandam' & Country != 'Winter Olympics 2022'
& Country != 'Holy See' & Country != 'Diamond Princess'
& Country != 'Summer Olympics 2020' )
```

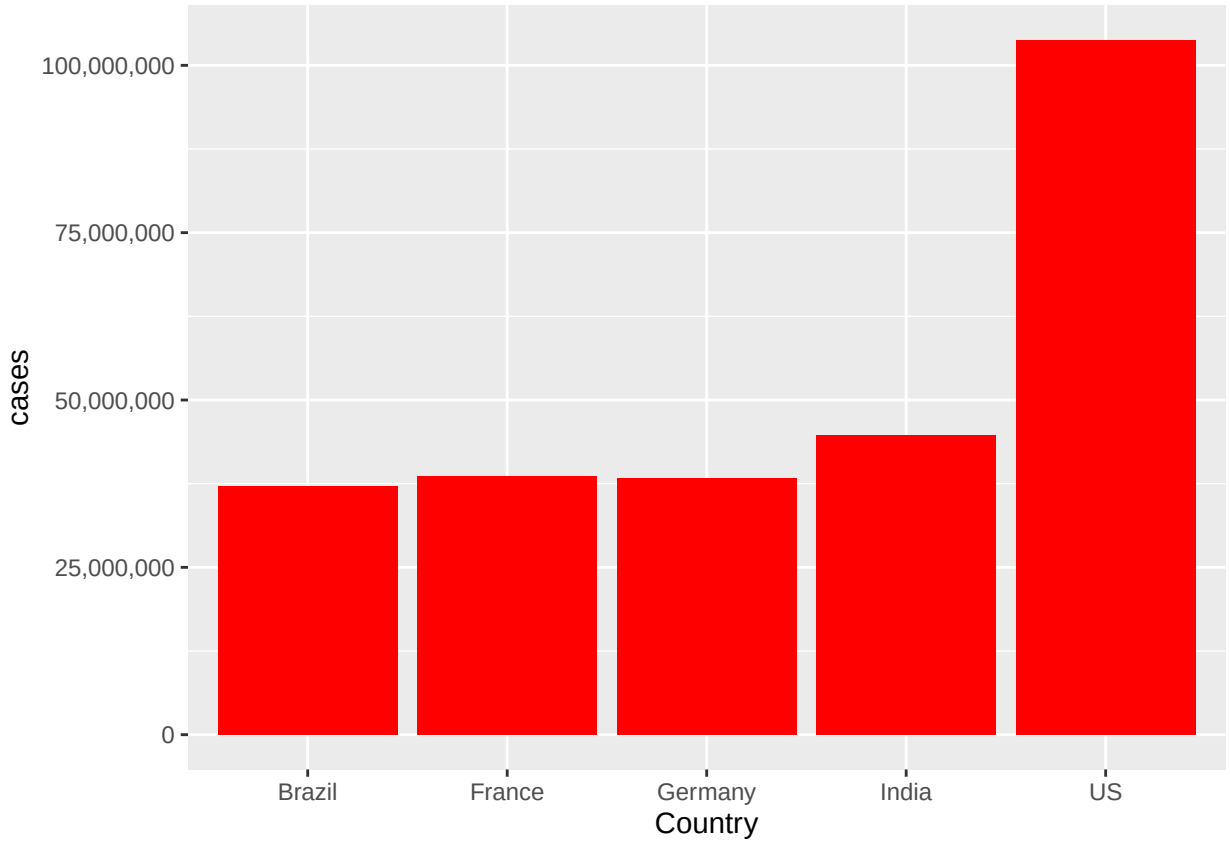
```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

Data Visualizations

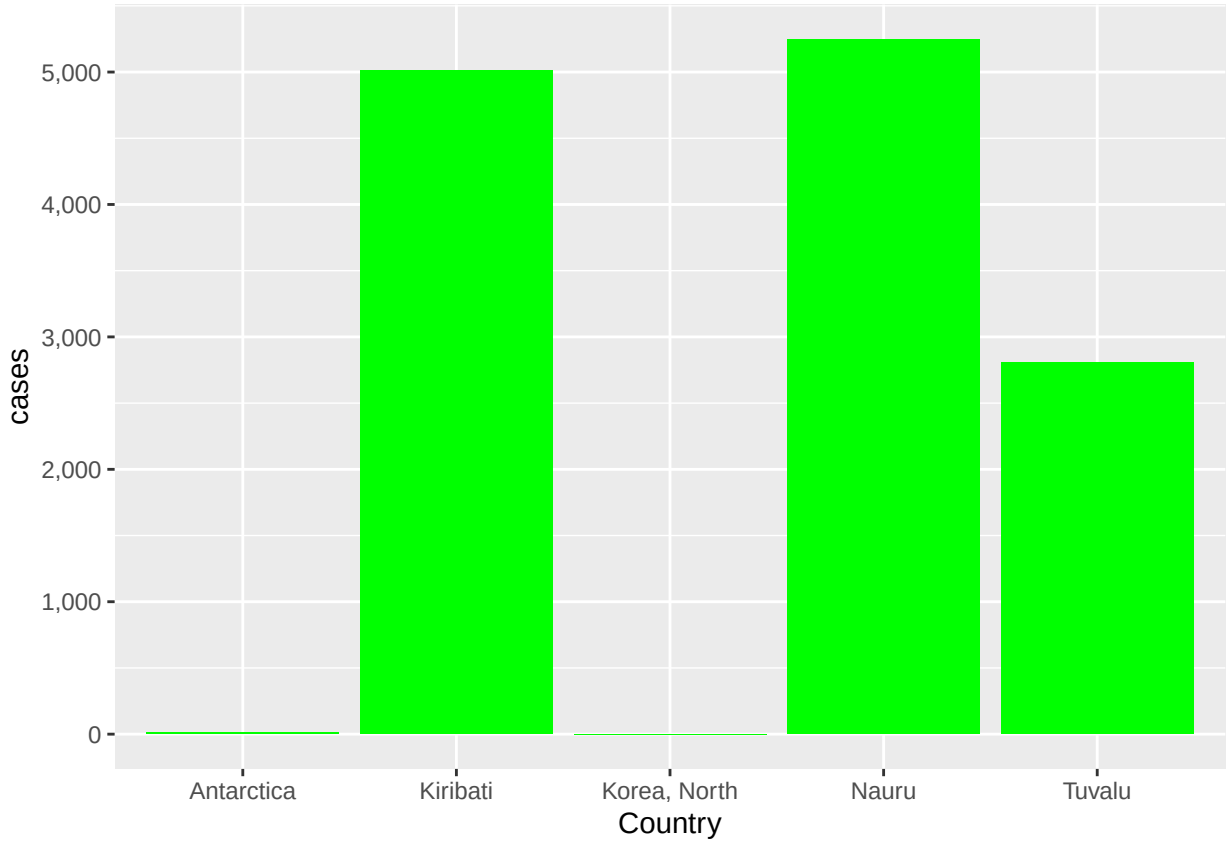
After cleaning our data, we can manipulate it further to extract data sets that contain the countries with the highest and lowest amount of both deaths and cases. In the below charts, we can see that the United States not only has the most cases, but also the most deaths reported. But does this truly mean that it was the deadliest country as far as COVID-19 statistics go? Conversely, we can see that North Korea was in the bottom 5 for both deaths and cases. This shows that the sourcing of this data has some inconsistencies, and that while more developed countries statistics may be more accurate, others might need to be taken with a grain of salt.

```
lowest_cases <- global[order(global$cases)[1:5],]

highest_cases <- global[rev(order(global$cases))[1:5],]

lowest_deaths <- global[order(global$deaths)[1:5],]
```

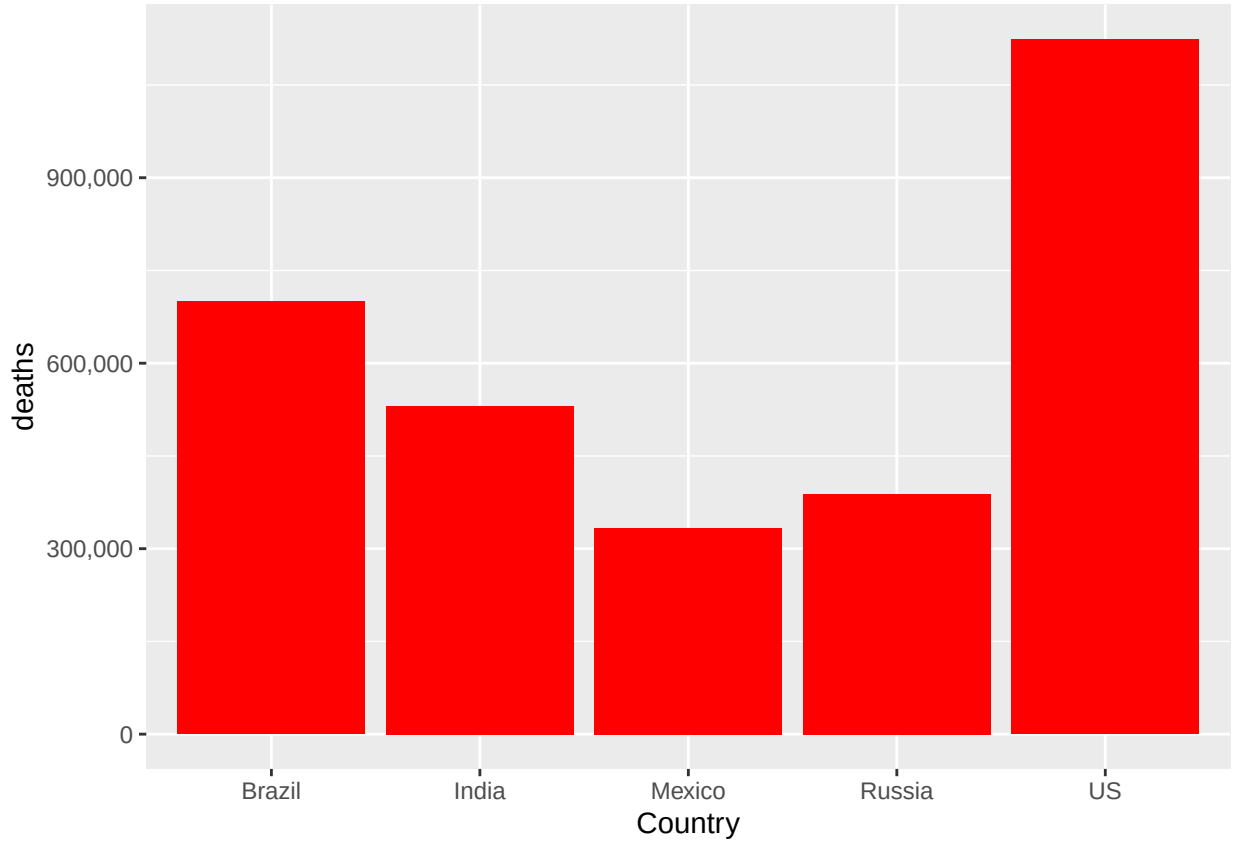
```
highest_deaths <- global[rev(order(global$deaths))[1:5],]

ggplot(data=highest_cases, aes(x=Country, y=cases)) + geom_bar(stat="identity", fill="red") + scale_y_c
```



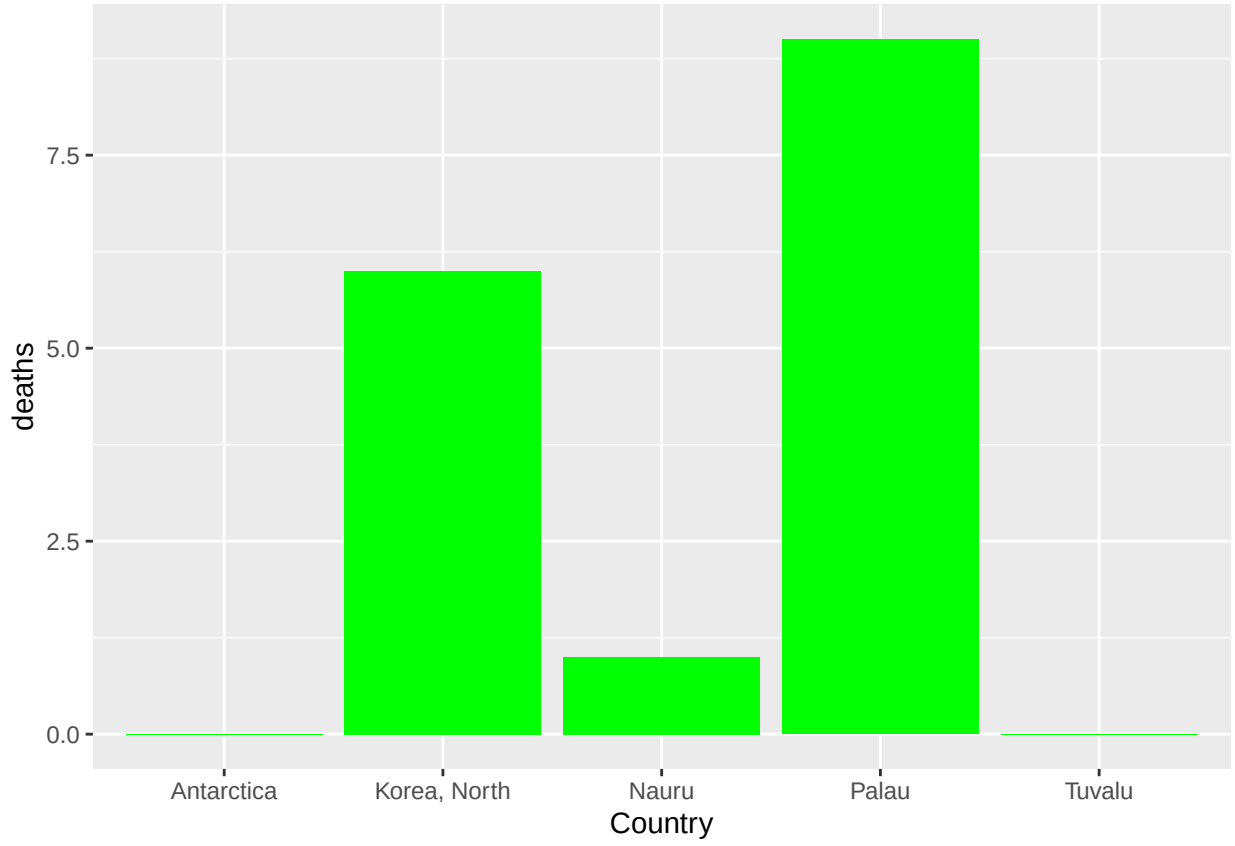

```
ggplot(data=lowest_cases, aes(x=Country, y=cases)) + geom_bar(stat="identity", fill="green") + scale_y_
```




```
ggplot(data=highest_deaths, aes(x=Country, y=deaths)) + geom_bar(stat="identity", fill="red") + scale_y
```




```
ggplot(data=lowest_deaths, aes(x=Country, y=deaths)) + geom_bar(stat="identity", fill="green") + scale_
```



Data Models

Here, we can make a model that predicts the the likely amount of deaths per cases for each country. When we first calculate the model, we can see that it seems to predict a mortality rate for each country that is much too high. To figure out the reason behind this, we can filter out some outliers that might be skewing our model. After doing this, it can be seen that North Korea has a whopping 600% mortality rate- a statistical impossibility.After removing the country, we can recalculate our model, which proves to be much more accurate.

```
mod <- lm(death_percentage ~ cases, data = global)
global <- global %>% mutate(pred = predict(mod))
global
```

```
## # A tibble: 193 x 5
##    Country                cases deaths death_percentage  pred
##    <chr>                  <dbl>  <dbl>            <dbl> <dbl>
##  1 Afghanistan           209451   7896            3.77   4.87
##  2 Albania               334457   3598            1.08   4.86
##  3 Algeria               271496   6881            2.53   4.87
##  4 Andorra                47890    165            0.345  4.89
##  5 Angola                105288   1933            1.84   4.89
##  6 Antarctica               11      0            0      4.90
##  7 Antigua and Barbuda    9106    146            1.60   4.90
##  8 Argentina           10044957 130472            1.30   3.77
```

```
##  9 Armenia               447308   8727            1.95   4.85
## 10 Austria              5961143  21970            0.369  4.23
## # i 183 more rows
```

```
global %>% filter(death_percentage > 5)
```

```
## # A tibble: 4 x 5
##   Country      cases deaths death_percentage  pred
##   <chr>        <dbl>  <dbl>            <dbl> <dbl>
## 1 Korea, North     1      6              600  4.90
## 2 Sudan        63829   5017             7.86  4.89
## 3 Syria        57467   3164             5.51  4.89
## 4 Yemen        11945   2159             18.1  4.90
```

```
global <- global %>% filter(Country != 'Korea, North')

mod <- lm(death_percentage ~ cases, data = global)
global <- global %>% mutate(pred = predict(mod))
global
```

```
## # A tibble: 192 x 5
##    Country                cases deaths death_percentage  pred
##    <chr>                  <dbl>  <dbl>            <dbl> <dbl>
##  1 Afghanistan           209451   7896            3.77   1.46
##  2 Albania               334457   3598            1.08   1.45
##  3 Algeria               271496   6881            2.53   1.45
##  4 Andorra                47890    165            0.345  1.46
##  5 Angola                105288   1933            1.84   1.46
##  6 Antarctica               11      0            0      1.46
##  7 Antigua and Barbuda    9106    146            1.60   1.46
##  8 Argentina           10044957 130472            1.30   1.33
##  9 Armenia               447308   8727            1.95   1.45
## 10 Austria              5961143  21970            0.369  1.38
## # i 182 more rows
```

```
global %>% filter( Country == "US" )
```

```
## # A tibble: 1 x 5
##   Country     cases  deaths death_percentage  pred
##   <chr>       <dbl>   <dbl>            <dbl> <dbl>
## 1 US      103802702 1123836             1.08 0.151
```

Bias and Conlusion

In this data, there are many possible sources of bias. To start, these statistics come from hundreds of different sources worldwide. It is certainly conceivable that some countries may under-report both case and death amounts to appear better from a public viewpoint. For example, we struck North Korea from our model because their reports were an extreme outlier that seemed incorrect. Another source of bias would be the data interpreter (me). As an American, I am part of the country with the highest reported deaths and cases. This could be seen as a bad look, so I would certainly be prone to blaming the reports of other countries as they all reported lower numbers than my own.

To conclude, It would appear that while the United States had both the highest amount of reported deaths and cases, it actually had a much lower death rate than many other countries. The model ended up still being very skewed for the US as well, projecting only a .151% mortality rate instead of the actual reported 1.08%. This leads me to believe that there are still large amounts of inaccurate data from other countries in the statstics/model. However, like I said, I am biased to believe that. More research will be needed to get to the truth.