WorldWide Covid Final

2022-11-08

Data Science Covid Adventure

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

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

```
library(dplyr)
library(ggplot2)
```

1. Importing the Data

Creating the URLs

```
url_github <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19
_data/csse_covid_19_time_series/"

file_names <- c("time_series_covid19_confirmed_global.csv","time_series_covid19_deaths_globa
l.csv")

full_urls <- str_c(url_github, file_names)
```

Read Data into R

```
cases <- read_csv(full_urls[1])
```

```
## Rows: 289 Columns: 1026
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1024): 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.
```

```
deaths <- read_csv(full_urls[2])
```

```
## Rows: 289 Columns: 1026
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1024): 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.
```

2. Formating the Data

Changing the Column names.

```
cases <- cases %>% rename(Country = "Country/Region", Province = "Province/State")
deaths <- deaths %>% rename(Country = "Country/Region", Province = "Province/State")
```

Data is row based, we will Pivot it into the correct shape.

```
cases_pivot <- cases %>% pivot_longer(cols = -c("Province", "Country", Lat, Long), names_to =
"date", values_to = "cases") %>% select(-c(Lat,Long))
deaths_pivot <- deaths %>% pivot_longer(cols = -c("Province", "Country", Lat, Long), names_to
= "date", values_to = "deaths") %>% select(-c(Lat,Long))
```

Joining the Death and Cases Data into a single DataFrame.

```
global <- cases_pivot %>% full_join(deaths_pivot) %>% mutate(date = mdy(date))
```

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

Let’s look at a short summary of our data.

```
summary(global)
```

```
##    Province           Country              date                 cases
##  Length:295358      Length:295358      Min.   :2020-01-22   Min.   :       0
##  Class :character   Class :character   1st Qu.:2020-10-03   1st Qu.:     454
##  Mode  :character   Mode  :character   Median :2021-06-15   Median :   10917
##                                        Mean   :2021-06-15   Mean   :  802389
##                                        3rd Qu.:2022-02-26   3rd Qu.:  183064
##                                        Max.   :2022-11-08   Max.   :97797561
##      deaths
##  Min.   :      0
##  1st Qu.:      2
##  Median :    120
##  Mean   :  12201
##  3rd Qu.:   2563
##  Max.   :1072921
```

We can see 0 as a minimum which is plausible, let's checl if the maximums are feasible.

```
global %>% filter(cases > 95000000)
```

```
## # A tibble: 63 × 5
##    Province Country date          cases  deaths
##    <chr>    <chr>   <date>        <dbl>   <dbl>
##  1 <NA>     US      2022-09-07 95030572 1049261
##  2 <NA>     US      2022-09-08 95137693 1049949
##  3 <NA>     US      2022-09-09 95238504 1050496
##  4 <NA>     US      2022-09-10 95248621 1050534
##  5 <NA>     US      2022-09-11 95257606 1050554
##  6 <NA>     US      2022-09-12 95327128 1051015
##  7 <NA>     US      2022-09-13 95406852 1051534
##  8 <NA>     US      2022-09-14 95510510 1052505
##  9 <NA>     US      2022-09-15 95593873 1053176
## 10 <NA>     US      2022-09-16 95650114 1053608
## # … with 53 more rows
```

```
global %>% filter(deaths > 1000000)
```

```
## # A tibble: 184 × 5
##    Province Country date          cases  deaths
##    <chr>    <chr>   <date>        <dbl>   <dbl>
##  1 <NA>     US      2022-05-09 82138934 1000213
##  2 <NA>     US      2022-05-10 82209658 1000540
##  3 <NA>     US      2022-05-11 82370700 1001224
##  4 <NA>     US      2022-05-12 82476172 1001526
##  5 <NA>     US      2022-05-13 82563440 1001733
##  6 <NA>     US      2022-05-14 82581561 1001793
##  7 <NA>     US      2022-05-15 82614682 1001830
##  8 <NA>     US      2022-05-16 82791646 1002077
##  9 <NA>     US      2022-05-17 82877118 1002474
## 10 <NA>     US      2022-05-18 83081000 1003445
## # … with 174 more rows
```

Yes they seem to be okay.

3. Analysis

Totals

Let's get some insights on the total amount of cases and deaths per Country.

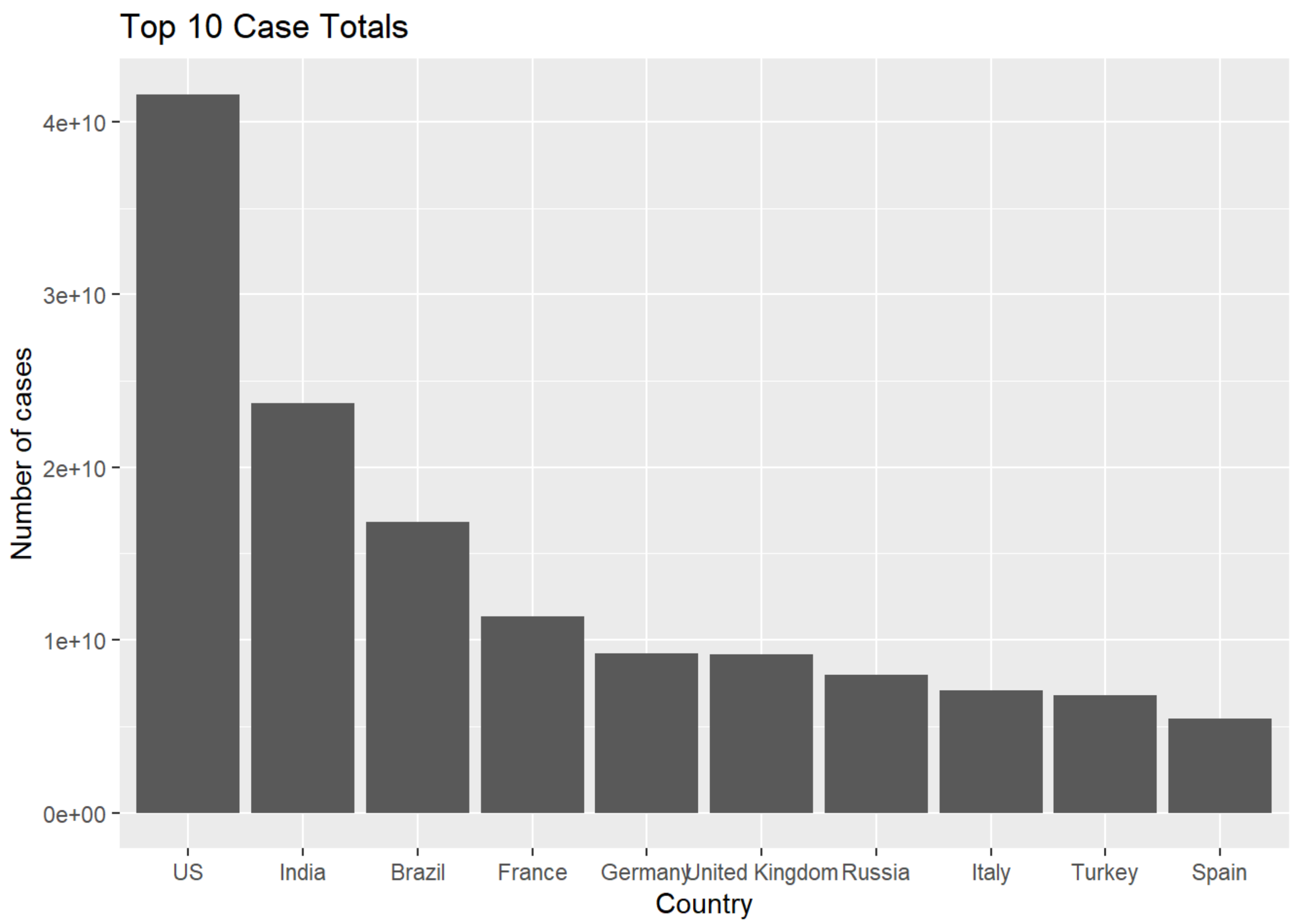
```
global_totals <- global %>% group_by(Country) %>% summarise(total_cases = sum(cases), total_d
eaths = sum(deaths))

cases_totals <-  global_totals %>% arrange(desc(total_cases))
death_totals <-  global_totals %>% arrange(desc(total_deaths))
```

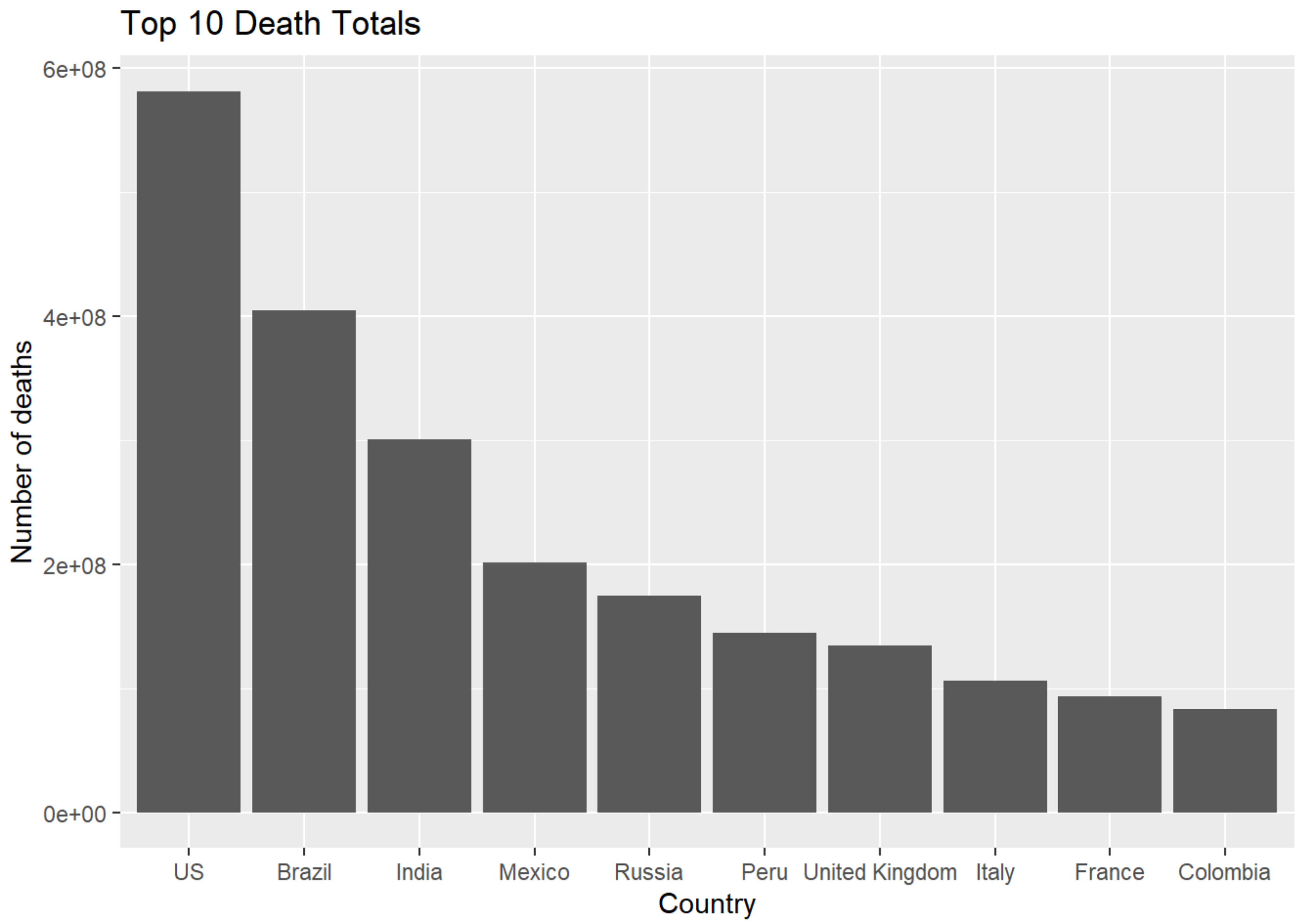
Now that we have that, why don't we check the Top 10 of each Category.

```
top_cases <- cases_totals %>% slice(1:10) %>% .$Country
top_deaths <- death_totals %>% slice(1:10) %>% .$Country

cases_totals %>% slice(1:10)  %>% ggplot(., aes(x=factor(Country, level = top_cases), y=total
_cases))+ geom_bar(stat='identity') + labs(title = "Top 10 Case Totals",x="Country", y="Numbe
r of cases")
```




```
death_totals %>% slice(1:10)  %>% ggplot(., aes(x=factor(Country, level = top_deaths), y=tota
l_deaths))+ geom_bar(stat='identity') + labs(title = "Top 10 Death Totals",x="Country", y="Nu
mber of deaths")
```



Interesting. We can see the top 10 are mostly by similar Countries. US is first in deaths and cases for example, but some don't have matching spots. Let's look into it further.

Difference

```
death_totals$death_pos <- seq.int(nrow(death_totals))
cases_totals$cases_pos <- seq.int(nrow(cases_totals))
total_pos <- merge(cases_totals,death_totals)
total_pos$dif <- total_pos$cases_pos - total_pos$death_pos
total_pos <-  total_pos %>% arrange(desc(dif))
total_pos$fac <-  scales::percent(total_pos$total_deaths / total_pos$total_cases)
```

```
total_stats <- total_pos
total_stats$total_cases <- NULL
total_stats$total_deaths <- NULL
```

Top 10 by difference

```
total_stats <- total_stats %>% arrange(fac)
total_stats %>% slice(1:10)
```

```
##                 Country cases_pos death_pos dif       fac
## 1            Antarctica       199       197   2 0.000000%
## 2                Tuvalu       200       200   0 0.000000%
## 3              Holy See       197       198  -1 0.000000%
## 4   Summer Olympics 2020       195       199  -4 0.000000%
## 5   Winter Olympics 2022       196       201  -5 0.000000%
## 6                 Nauru       194       196  -2 0.021290%
## 7                Bhutan       153       186 -33 0.038432%
## 8                 Tonga       186       191  -5 0.088213%
## 9           New Zealand        76       142 -66 0.099101%
## 10              Iceland       123       174 -51 0.102185%
```

```
total_pos <- total_pos %>% arrange(fac)
# remove Winter Olympics/non countries
total_pos <- total_pos %>% slice(-c(2, 3, 4, 5))


total_stats <-total_stats %>% arrange(desc(fac))
total_pos <-total_pos %>% arrange(desc(fac))
total_stats %>% slice(1:10)
```

```
##                   Country cases_pos death_pos dif         fac
## 1                    Peru        24         6  18   7.338928%
## 2                   Sudan       132        89  43   7.312650%
## 3             Korea, North       201       195   6 600.000000%
## 4                  Mexico        16         4  12   6.560337%
## 5                   Syria       139       100  39   5.944483%
## 6                   Egypt        85        37  48   5.234181%
## 7                 Somalia       154       122  32   5.034458%
## 8                 Ecuador        64        26  38   4.815329%
## 9             Afghanistan       109        71  38   4.262304%
## 10 Bosnia and Herzegovina        94        51  43   4.211137%
```

```
#remove NK for having more deaths than cases
total_pos <- total_pos %>% slice(-c(3))
```

So let’s look at the top 10 for each side. Countries where they have much higher Deaths to Cases ratio and the other way.

Top 10 by difference

```
total_pos %>% arrange(desc(fac)) %>% slice(1:10)
```

```
##                   Country total_cases total_deaths cases_pos death_pos dif
## 1                    Peru  1966603603    144327628        24         6  18
## 2                   Sudan    35228980      2576172       132        89  43
## 3                  Mexico  3063372182    200967538        16         4  12
## 4                   Syria    28260187      1679922       139       100  39
## 5                   Egypt   272206276     14247770        85        37  48
## 6                 Somalia    14560395       733037       154       122  32
## 7                 Ecuador   458768907     22091232        64        26  38
## 8             Afghanistan   104878576      4470244       109        71  38
## 9  Bosnia and Herzegovina   199047704      8382171        94        51  43
## 10                Liberia     4051402       168968       180       152  28
##          fac
## 1  7.338928%
## 2  7.312650%
## 3  6.560337%
## 4  5.944483%
## 5  5.234181%
## 6  5.034458%
## 7  4.815329%
## 8  4.262304%
## 9  4.211137%
## 10 4.170606%
```

```
total_pos %>% arrange(fac) %>% slice(1:10)
```

```
##             Country total_cases total_deaths cases_pos death_pos dif       fac
## 1        Antarctica        3630            0       199       197   2 0.000000%
## 2             Nauru      610604          130       194       196  -2 0.021290%
## 3            Bhutan    14761230         5673       153       186 -33 0.038432%
## 4             Tonga     2982552         2631       186       191  -5 0.088213%
## 5       New Zealand   335936914       332917        76       142 -66 0.099101%
## 6           Iceland    57185675        58435       123       174 -51 0.102185%
## 7         Singapore   457593893       483950        65       129 -64 0.105760%
## 8  Marshall Islands     1253663         1406       191       194  -3 0.112151%
## 9           Burundi    16834102        19033       150       183 -33 0.113062%
## 10            Palau     1359718         1627       189       193  -4 0.119657%
```

4. Country Analysis and model

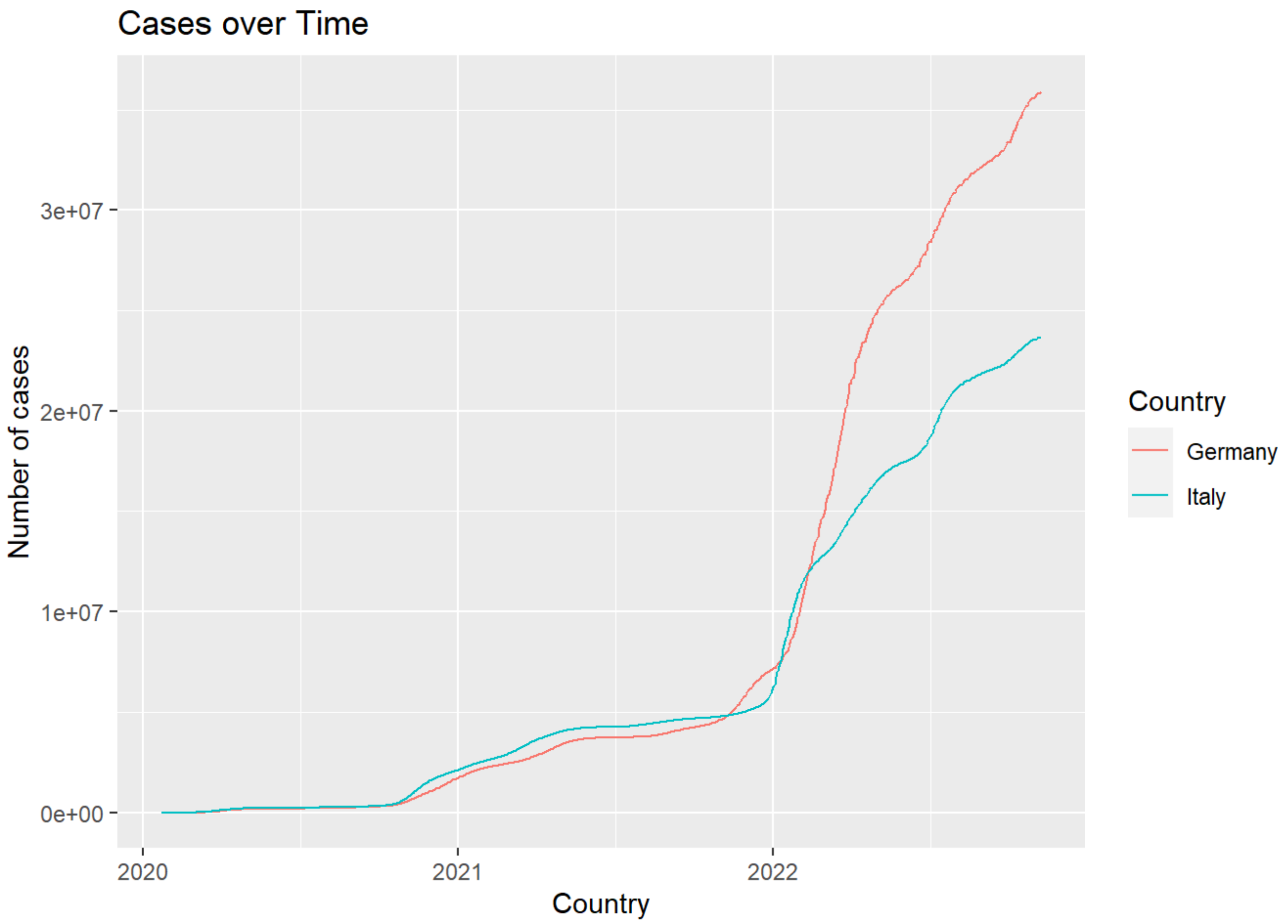
Analysis

Germany and Italy are countries with very different covid policies. So let’s compare the two.

```
geit <- filter(global, Country == "Germany" |  Country == "Italy")
italy <- filter(global,Country == "Italy")
germany <- filter(global,Country == "Germany")
```

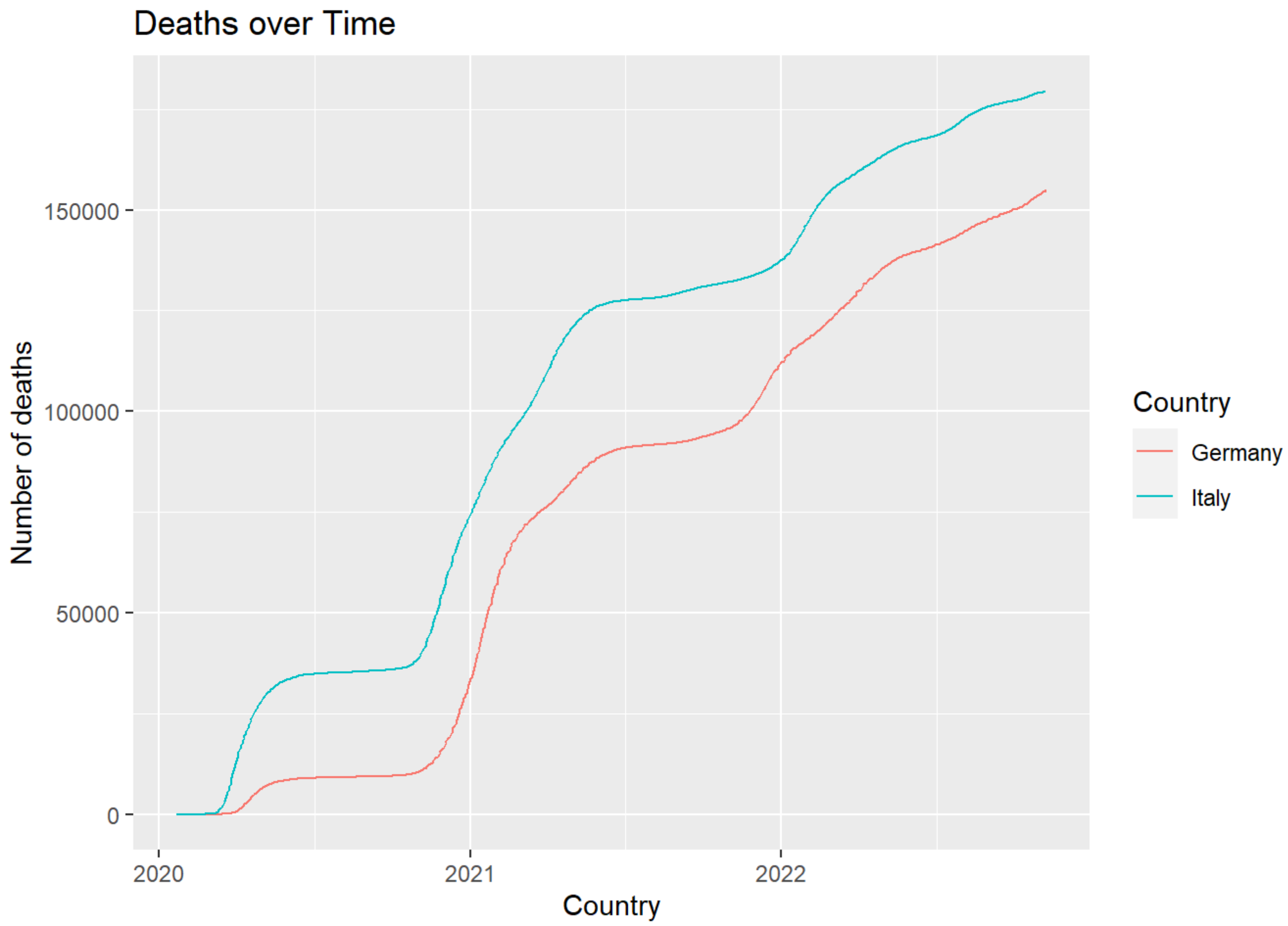
Cases over time

```
ggplot(geit, aes(x=date, y=cases, group=Country, color=Country)) + geom_line() + labs(title =
"Cases over Time",x="Country", y="Number of cases")
```



Deaths over time

```
ggplot(geit, aes(x=date, y=deaths, group=Country, color=Country)) +geom_line() + labs(title =
"Deaths over Time",x="Country", y="Number of deaths")
```



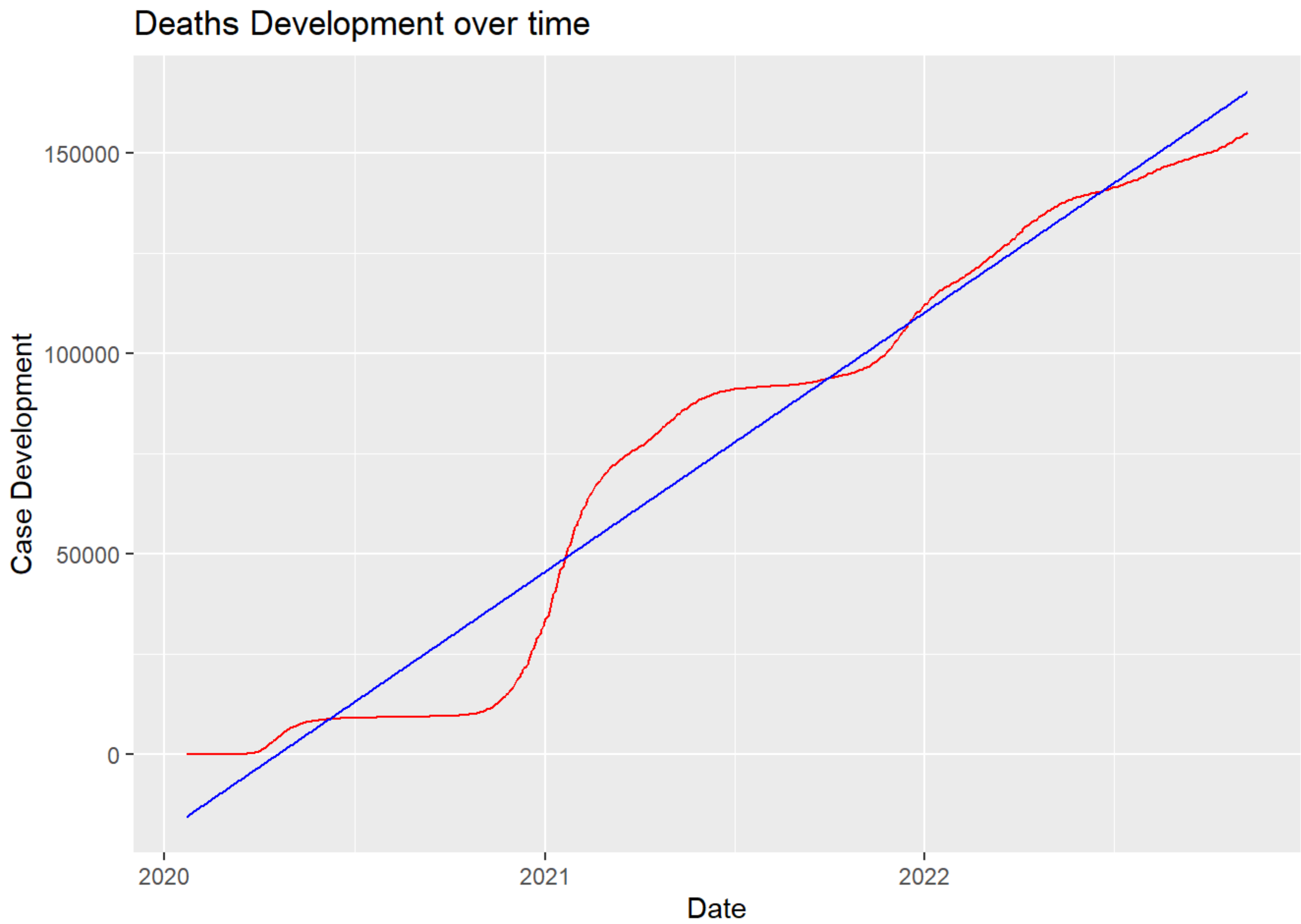
ML Model

```
mod <- lm(deaths ~ date, data=germany)
germany$pred <- predict(mod)


modi <- lm(deaths ~ date, data=italy)
germany$predi <- predict(modi)
```

Predictions are Red for Germany, Blue is actual Germany

```
ggplot(germany,aes(x=date)) + geom_line(aes(y = deaths), color="red") + geom_line(aes(y = pre
d), color="blue") + labs(title = "Deaths Development over time",x="Date", y="Case Developmen
t")
```



We can see that cases are still rising at a significant rate. The actual numbers are pretty close towards the trend line, which is pretty interesting. Before we were boomeranging around the trend line.

This does make a lot of sense as pandemic spread is very regularized and rises and drops very harmonically. We can see great curves and no sharp edges. We can see the actual number slowing down seeing we will most likely see easing of the curve. See a reduction in deaths.

5. Bias

There are some possible points of bias in the data and analysis, which I will briefly talk about here. This Data is not collected by one single entity, each country reports their own numbers. There are many different definitions cases and deaths can have, so comparing them one to one can only be done with a disclaimer.

Furthermore, the case numbers are very much dependent on how much testing is going on in these countries. If a country is poor it might not have access to as many Covid tests as a rich nation. There can only be as many confirmed cases as you test. One example of this is North Korea, while exploring the data we saw that they have more deaths than cases. Clearly showing that there are way more covid cases in North Korea than the data suggests.