Investigating John Hopkins University Covid-19 Data

DTSA 5301, University of Colorado

April 1st, 2022

OBJECTIVE

Analyze the worldwide rates of Covid-19 cases that resulted in deaths.

DATA IMPORT & TIDYING

Importing Data:

Data for the number of Covid-19 cases and deaths are collected from the John Hopkins University Git-Hub account and each transformed into their own respective data frames.

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

cases <- "csse_covid_19_time_series/time_series_covid19_confirmed_global.csv"
deaths <- "csse_covid_19_time_series/time_series_covid19_deaths_global.csv"

cases <- read_csv(paste(url_path, git_path, cases, sep = ""))
deaths <- read_csv(paste(url_path, git_path, deaths, sep = ""))
```

Tidying Data:

These two data frames are then joined together based on the common dates for cases and deaths, as well as the common countries between the data frames. The latitude and longitude data is dropped from the data frame as it's not required for the analysis of rates of death due to Covid-19.

```
cases <- cases %>%
    select(-c("Lat", "Long")) %>%
    pivot_longer(cols = -c("Province/State", "Country/Region"),
        names_to = "Date", values_to = "Cases") %>%
    mutate(Date = mdy(Date)) %>%
    group_by(`Country/Region`, Date) %>%
    summarise(Cases = sum(Cases)) %>%
    ungroup()

deaths <- deaths %>%
    select(-c("Lat", "Long")) %>%
    pivot_longer(cols = -c("Province/State", "Country/Region"),
```

```
        names_to = "Date", values_to = "Deaths") %>%
    mutate(Date = mdy(Date)) %>%
    group_by(`Country/Region`, Date) %>%
    summarise(Deaths = sum(Deaths)) %>%
    ungroup()

covid <- list(cases, deaths) %>%
    reduce(left_join, by = c("Country/Region", "Date"))
```

Data Summary:

The now tidied Covid data frame indicates that there four columns of features, and nearly 160,000 rows of data.

```
glimpse(covid)
```

```
## Rows: 158,598
## Columns: 4
## $ `Country/Region` <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghani~
## $ Date             <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020~
## $ Cases            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Deaths           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

> (**Note:** The John Hopkins University Covid-19 data set has new data added daily. There may be the possibility that there are greater than 160,000 rows of data depending on when the code in this report is run. Because the data set is dynamic it may have an effect on the data analysis results as stated in this report. All rates as indicated throughout this report are based on data up to April 1st, 2022)

DATA ANALYSIS

Global Death Rate:

To begin the analysis it would be interesting to determine what the total worldwide death rate is for all Covid-19 cases, where the death rate is defined as:

$$\text{Death Rate} = \frac{\text{Total Deaths}}{\text{Total Cases}}$$

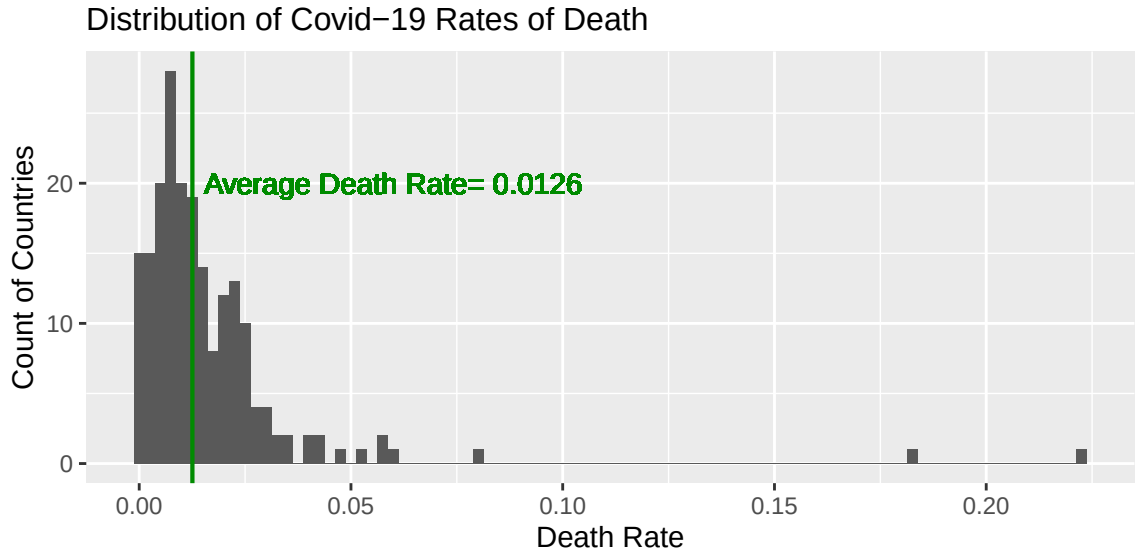
```
# Table detailing the average worldwide death rate
death_rate <- covid %>%
    group_by(`Country/Region`) %>%
    summarise(Cases = max(Cases), Deaths = max(Deaths)) %>%
    summarise(`Total Cases` = sum(Cases), `Total Deaths` = sum(Deaths)) %>%
    mutate(`Death Rate` = `Total Deaths`/`Total Cases`)

death_rate %>%
    knitr::kable(digits = 4, format.args = list(big.mark = ",",
        scientific = FALSE))
```

Total Cases	Total Deaths	Death Rate
489,611,309	6,148,767	0.0126

Distribution of Death Rates:

The above table shows that, on average, of all the Covid-19 cases contracted worldwide only 1.26% of those cases resulted in a death. Knowing the global death rate is 1.26%, it would be important to understand how this rate varies by country, throughout the world. To begin investigating this we will look at how the death rates are distributed:

```
# Histogram of death rates
covid %>%
    group_by(`Country/Region`) %>%
    summarise(Cases = max(Cases), Deaths = max(Deaths)) %>%
    mutate(`Death Rate` = Deaths/Cases) %>%
    ggplot(aes(x = `Death Rate`)) + geom_histogram(binwidth = 0.0025) +
    geom_vline(aes(xintercept = death_rate[[3]]), color = "green4",
        size = 0.75) + geom_text(aes(x = 0.06, y = 20, label = paste("Average Death Rate=",
    format(death_rate[[3]], digits = 3))), color = "green4") +
    ggtitle("Distribution of Covid-19 Rates of Death") + labs(y = "Count of Countries") +
    theme(plot.title = element_text(size = 12))
```



Countries with Well Above Average Death Rates:

The majority of death rates worldwide are below approximately 4.0%, as shown in the plot above. The next step is to remove the countries below the 4.0% threshold from the analysis, in order to study the countries which are above this threshold. A table of the countries with death rates above 4.0% is presented below:

```
# Data frame of outlier countries with death rates above 4%
high_death_rate <- covid %>%
    group_by(`Country/Region`) %>%
    summarise(Cases = sum(max(Cases)), Deaths = sum(max(Deaths))) %>%
```

```
    mutate(`Death Rate` = Deaths/Cases) %>%
    filter(`Death Rate` >= 0.04) %>%
    arrange(desc(`Death Rate`))

# Table of outlier death rates
high_death_rate %>%
    knitr::kable(digits = 3, format.args = list(big.mark = ",",
        scientific = FALSE))
```

Country/Region	Cases	Deaths	Death Rate
MS Zaandam	9	2	0.222
Yemen	11,806	2,143	0.182
Sudan	61,955	4,907	0.079
Peru	3,547,606	212,256	0.060
Mexico	5,662,073	323,127	0.057
Syria	55,701	3,142	0.056
Somalia	26,410	1,361	0.052
Egypt	505,264	24,417	0.048
Afghanistan	177,782	7,670	0.043
Bosnia and Herzegovina	375,554	15,718	0.042
Ecuador	859,890	35,421	0.041

The highest death rate in this table is for the "MS Zaandam", which is a cruise ship, at 22.2%. Nine of the passengers caught Covid-19 while on board, with two of those passengers passing away from the virus. The ship was at sea while the passengers contracted the virus, so none of these cases or deaths can be assigned to any country in particular. Due to this fact, this data is removed from the data frame and the analysis continued focusing on these remaining countries:

```
# Filtering out the MS Zaandam data
high_death_rate <- high_death_rate %>%
    filter(`Country/Region` != "MS Zaandam")
# List of remaining outlier countries
high_death_rate[["Country/Region"]]
```

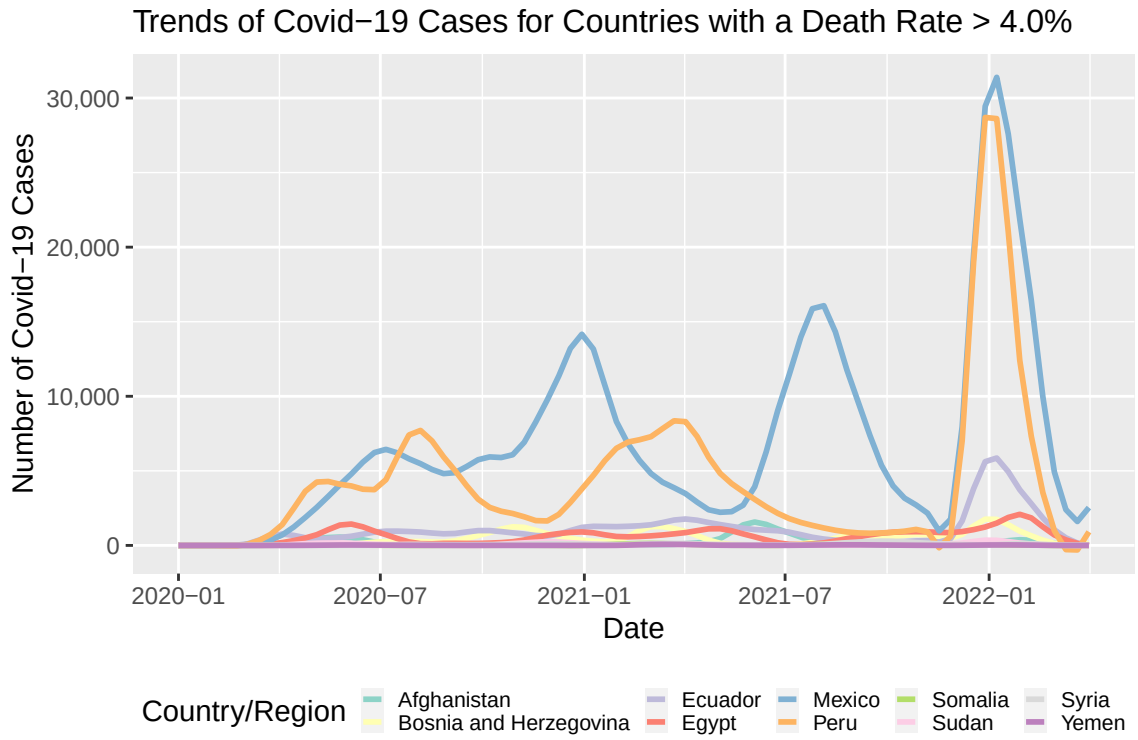
```
##  [1] "Yemen"                  "Sudan"                  "Peru"
##  [4] "Mexico"                 "Syria"                  "Somalia"
##  [7] "Egypt"                  "Afghanistan"            "Bosnia and Herzegovina"
## [10] "Ecuador"
```

Trends of Covid-19 Cases:

A chart of the trend of Covid-19 cases is shown below, analyzing only the countries with death rates above 4.0%. These trends of cases indicate that Mexico and Peru have been heavily impacted by the Covid-19 virus with each country experiencing multiple "waves" of cases.

```
covid %>%
    group_by(`Country/Region`) %>%
    filter(`Country/Region` %in% high_death_rate[["Country/Region"]]) %>%
    mutate(daily_cases = Cases - lag(Cases), daily_deaths = Deaths -
        lag(Deaths)) %>%
```

```
    drop_na() %>%
    mutate(Date = floor_date(ymd(Date), "month")) %>%
    group_by(`Country/Region`, Date) %>%
    summarise(Cases = mean(daily_cases), Deaths = mean(daily_deaths)) %>%
    arrange(`Country/Region`, Date) %>%
    ggplot(aes(x = Date, y = Cases, color = `Country/Region`)) +
    geom_smooth(span = 0.15, se = FALSE) + expand_limits(y = 0) +
    ggtitle("Trends of Covid-19 Cases for Countries with a Death Rate > 4.0%") +
    labs(y = "Number of Covid-19 Cases") + theme(plot.title = element_text(size = 12),
    legend.position = "bottom", legend.key.size = unit(3, "mm"),
    legend.text = element_text(size = 8)) + scale_y_continuous(labels = comma) +
    scale_color_brewer(palette = "Set3")
```

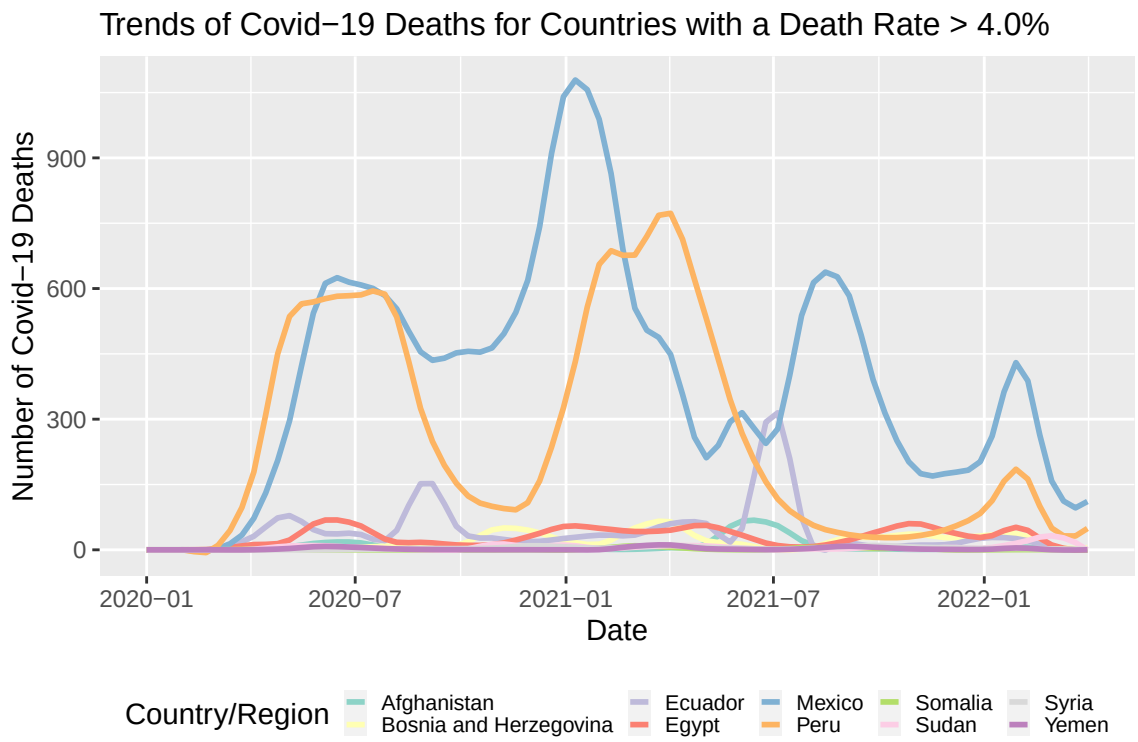


Trends of Covid-19 Deaths:

Corresponding to the amount of cases in Mexico and Peru, these countries also experienced a significant amount of deaths due to Covid-19 virus infections. It is encouraging that despite the large increase in the amount of cases over the last few months of data, the relative amount of deaths has decreased.

```
covid %>%
    group_by(`Country/Region`) %>%
    filter(`Country/Region` %in% high_death_rate[["Country/Region"]]) %>%
    mutate(daily_cases = Cases - lag(Cases), daily_deaths = Deaths -
        lag(Deaths)) %>%
    drop_na() %>%
    mutate(Date = floor_date(ymd(Date), "month")) %>%
    group_by(`Country/Region`, Date) %>%
```

```
    summarise(Cases = mean(daily_cases), Deaths = mean(daily_deaths)) %>%
    arrange(`Country/Region`, Date) %>%
    ggplot(aes(x = Date, y = Deaths, color = `Country/Region`)) +
    geom_smooth(span = 0.15, se = FALSE) + expand_limits(y = 0) +
    ggtitle("Trends of Covid-19 Deaths for Countries with a Death Rate > 4.0%") +
    labs(y = "Number of Covid-19 Deaths") + theme(plot.title = element_text(size = 12),
    legend.position = "bottom", legend.key.size = unit(3, "mm"),
    legend.text = element_text(size = 8)) + scale_color_brewer(palette = "Set3")
```



SOURCES OF BIAS:

The first possible bias would be to assume that only the poorest countries in the world have death rates above 4.0% because a lack of affordability of Covid-19 vaccines for those countries. Another bias would be to presume that the decrease in the rate of deaths over time in countries with death rates above 4.0%, despite an increase in cases, would be due to the implementation of vaccines in these countries.

Both of these biases cannot be proven unless relative financial indicators, such as GDP data, and vaccine distribution data sets are also included in this analysis.

CONCLUSION

It has been shown that the worldwide average death rate due to Covid-19 infections is currently equal to 1.26% of all cases reported. There are a number of outlier countries that are well above this rate, but further analysis is required to understand why Covid-19 has had a greater impact on these outlier countries when compared to countries with lower rates of deaths.

R SESSION INFORMATION:

```
##  package   * version date (UTC) lib source
##  dplyr     * 1.0.8   2022-02-08 [1] CRAN (R 4.1.2)
##  forcats   * 0.5.1   2021-01-27 [1] CRAN (R 4.1.1)
##  formatR   * 1.11    2021-06-01 [1] CRAN (R 4.1.3)
##  ggplot2   * 3.3.5   2021-06-25 [1] CRAN (R 4.1.0)
##  knitr     * 1.37    2021-12-16 [1] CRAN (R 4.1.2)
##  lubridate * 1.8.0   2021-10-07 [1] CRAN (R 4.1.2)
##  purrr     * 0.3.4   2020-04-17 [1] CRAN (R 4.1.1)
##  readr     * 2.1.2   2022-01-30 [1] CRAN (R 4.1.2)
##  scales    * 1.1.1   2020-05-11 [1] CRAN (R 4.1.1)
##  stringr   * 1.4.0   2019-02-10 [1] CRAN (R 4.1.1)
##  tibble    * 3.1.6   2021-11-07 [1] CRAN (R 4.1.2)
##  tidyr     * 1.2.0   2022-02-01 [1] CRAN (R 4.1.2)
##  tidyverse * 1.3.1   2021-04-15 [1] CRAN (R 4.1.1)
##
##  [1] C:/Users/ryand/Documents/R/win-library/4.1
##  [2] C:/Program Files/R/R-4.1.1/library
```