COVID-19 GLOBAL RATES

Portfolio Project Option 1

Taylor Jackson

CSU Global

COVID-19 Global Rates Analysis

COVID-19 is a very prominent issues across the globe and is the abbreviated term for Coronavirus Disease 2019. Not only is it affecting daily life of everyone, but it is also causing massive amount of deaths and changing quickly and dynamically. To help understand COVID-19 more, this report shows how deaths and recoveries have compared to confirmed cases globally and overtime. It is expected that there are more COVID-19 recoveries than deaths worldwide. The dataset used is provided on datahub by Datopian. The dataset has normalized a dataset created by [the Johns Hopkins University Center for Systems Science and Engineering](https://systems.jhu.edu/) (CSSE) team which has continuously logged the date, country, confirmed cases, recoveries and deaths COVID related (Novel Corona, para. 3). This dataset will be stored in an R directory, which enables one to recover stored data quickly in the R software. This is particularly useful because R will be used to analyze and interpret the data.

**Null Hypothesis:** The amount of globally reported COVID related deaths is greater than the number of reported recoveries.

**Alternative Hypothesis:** The amount of globally reported recoveries related to COVID

is greater than the number of COVID related deaths.

# Processing Data

There are some steps that must be taken to test the hypothesis and interpret our data for use. This includes data exploration and management. In other words, the data must be set up correctly and understood in order to move forward. Next, summary statistics and 2 sample t-testing shall be performed. From these steps, one can determine the accepted hypotheses and observe the trends in data (Spector, 2014, para. 1).

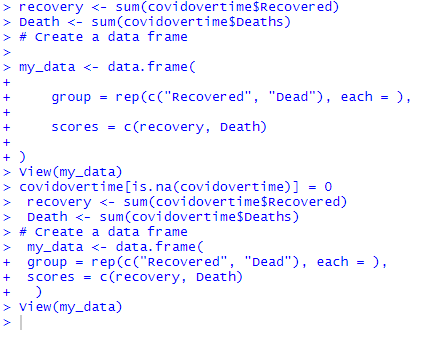
## Summary Statistics

The data is first processed so that it can be manipulated correctly. The data is difficult to process when NA values are present. The following code changes all NA values to 0:

|  |
| --- |
| covidovertime[is.na(covidovertime)] = 0 |

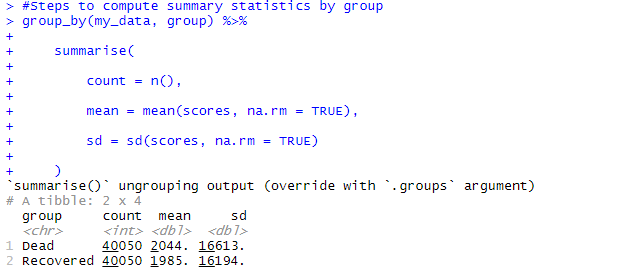
Now the data is ready for hypothesis testing. The next step in preparing the dataset is to create a data frame that represents the total reported recovery and death rates. Figure 1 shows the code used for this process:

Figure



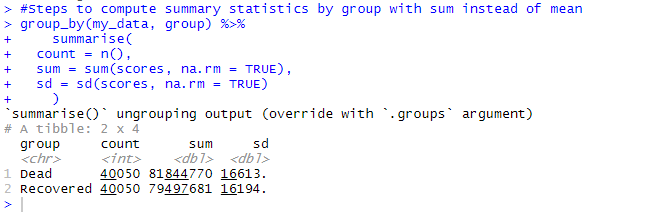
Now the data is ready to be tested. To test the hypothesis, the summary statistics is calculated first. Figure 2 shows both the code used and the results of the summary statistics.

Figure



In this case, the mean of the data is not as useful as using the sum of cases. Figure 3 shows the same summary statistics but with “mean” replaced with the “sum” function. This figure is only used to demonstrate the sum and enable one to predict our hypothesis result.

Figure

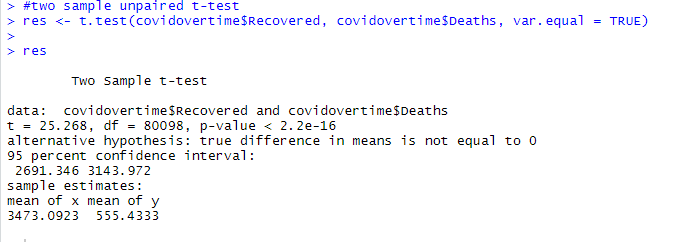


## T-testing and Hypothesis Testing

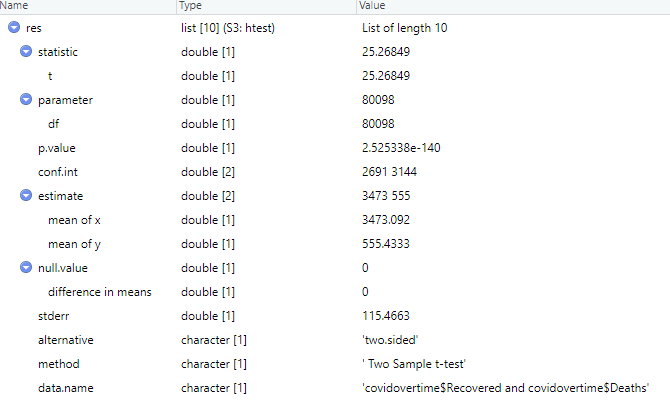
With summary statistics calculated, the next step is an unpaired two sample t-test using the steps out lined in ‘T-tests” (Kabacoff. 2017, para. 3). The following code was used to produce the results of this test. Figure 4 and 5 show the results.

|  |
| --- |
| #two sample unpaired t-test  > res <- t.test(covidovertime$Recovered, covidovertime$Deaths, var.equal = TRUE) |

Figure



Figure



# Hypothesis Conclusion

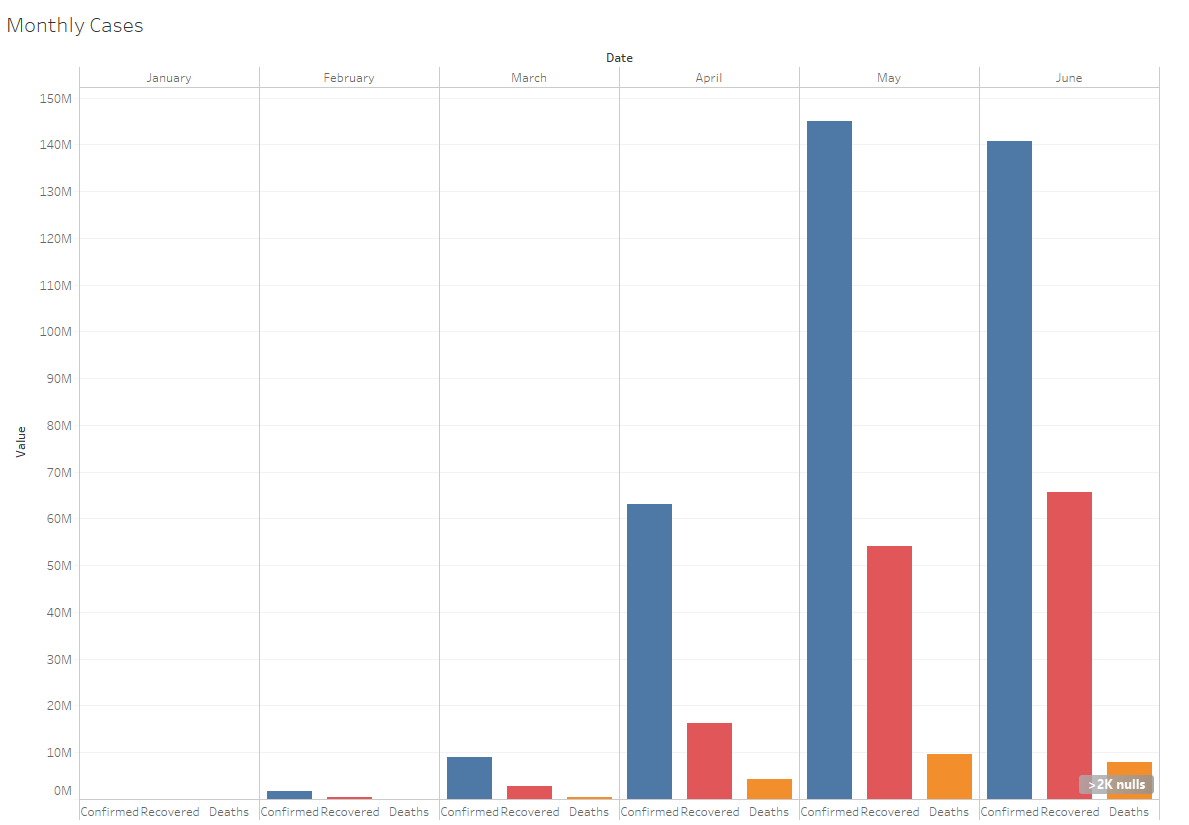
By observing the mean calculated in our hypothesis test in figure 4, the null hypothesis can be rejected or accepted. The results presented are mean of x (total deaths) = 3473.1 and the mean of y (total recoveries) = 555.4 which means the number of recoveries is greater than that of deaths in relation to COVID-19 global cases. Therefore, the Null hypothesis can be rejected, and the alternative hypothesis accepted.

# Further Analysis

With the alternative hypothesis accepted, further data manipulation can be done to better understand the affect of COVID-19 cases around the world. Using the application Tableau, the data set from Datopian is processed to visualize the global rates of COVID-19.

Figure 6 shows how confirmed, deaths, and recovered cases have developed worldwide by months (2020). This visualization further demonstrates how recoveries have been greater than deaths, even as confirmed cases have grown.

Figure



In order to understand how each country has been affected, figure 7 shows the number of confirmed cases by number and the amount of recoveries by color density. As shown, the United States has the most confirmed cases and more recoveries, Brazil coming in second, and then Italy, Spain, and Russia close behind.

Figure

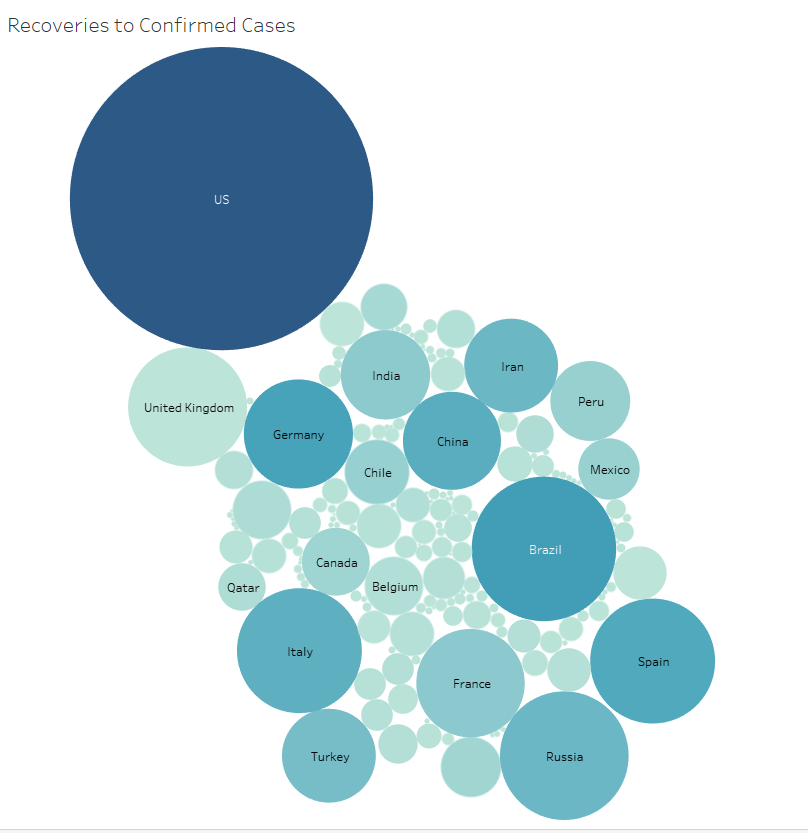
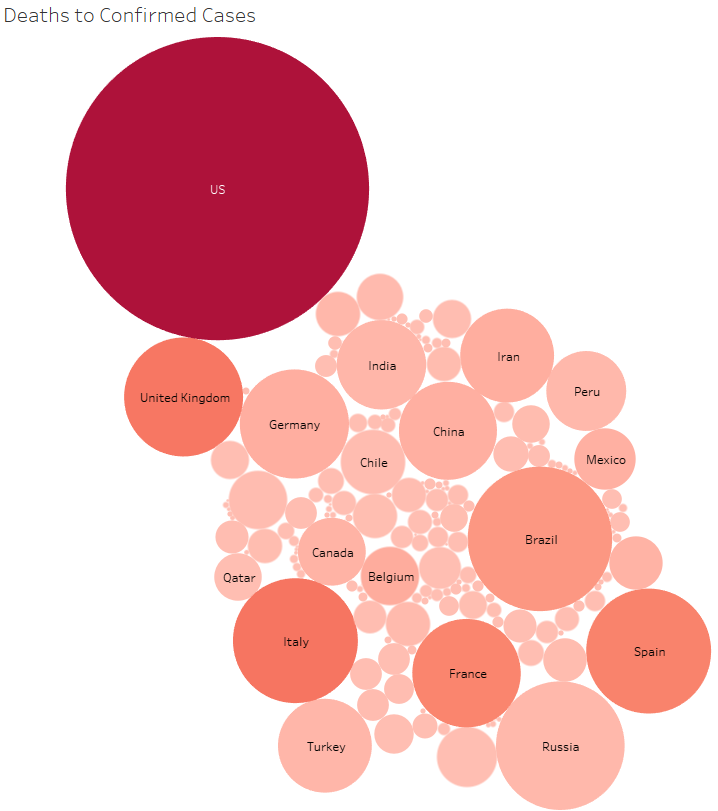


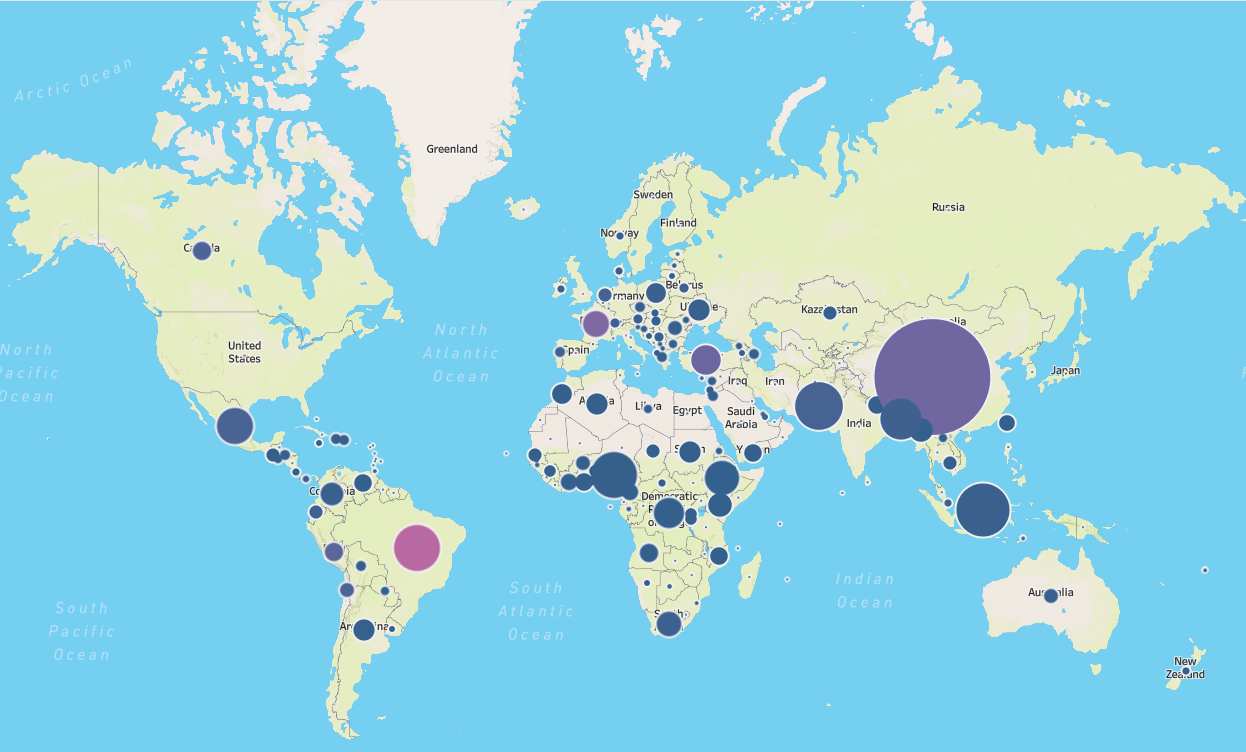
Figure 8 shows a similar trend and represents deaths to confirmed cases rather than recoveries. The United States is still the lead in number of cases and deaths. However, there are several countries that have more deaths to confirmed cases than recoveries to confirmed cases (see the UK in figure 7 & 8 as an example).

Figure



The last figure shows a very important consideration: population. It’s important to note that the US is missing values and is therefore not shown on the map (this is due to the way data is stored in this dataset). The population data is a different dataset (can be found in the dataset called “reference”) provided by the same organization as the COVID-19 data referenced for our hypothesis testing (Novel, 2019). Figure 9 shows the country and population (size of corresponding circle) and number of confirmed cases (color density) from 100 (blue) to 140,000,000 (red). Considering population is very important because a high number of confirmed cases in one country compared to another is relative to their population and can not be properly compared without comparing number of cases to total population.

Figure



# Conclusion

Using the datasets provided by Datopain, it has been determined that COVID-19 recoveries have been greater than deaths on a global scale. This observation and additional visualization can provide a deeper understanding of how COVID-19 has affected the world – month by month and country by country.

References

Kabacoff, Robert. "Quick R: T-tests." N.p., 2017. Web. 20 June 2020.

Novel Coronavirus 2019. (2020, April). Retrieved June 06, 2020, from <https://datahub.io/core/covid-19>

Spector, Phil. "Using T-tests in R." *Department of Statistics*. University of California, 2014. Web. 20 June 2020.