Final Project: COVID-19 Dataset

Due Friday, November 20, 11:59 PM

The Lads: Frankie Willard, Manny Mokel, Alex Katopodis, Parker Dingman

Introduction

Throughout the year 2020, the COVID-19 pandemic took the world by storm, deeply impacting every country on the planet, albeit with differing degrees of severity. As cases continued to rise, families suffered from the loss of family members, jobs, social interactions, disposable income, and more.

This public health crisis became severe enough such that many countries took decisive action, shutting down their economies to prioritize the lives of citizens. Meanwhile, other countries were less strict in their policies, attempting to preserve their economy at the potential expense of their citizen's lives. The difference in each country's characteristics, demographics, public health capacities, and the strictness of COVID-19 policies led to vastly different effects of the pandemic on different countries. Given our personal connections to the effects of the pandemic through our lives, our friends, and our families, we wanted to determine what led to the pandemic affecting some places worse than others.

We are interested in investigating how a country's demographics impact the domestic severity of COVID-19. More specifically, we would like to see which demographics lead to higher cases per capita and deaths per case. We are also interested in analyzing how effective lockdowns and COVID-19 related policies have been in mitigating the spread of the virus.

We hypothesize that stringency-index, GDP per capita, population density, and human development index will have a strong impact on cases per capita. We also hypothesize that deaths per case will be largely determined by GDP per capita, the number of citizens aged 65+, hospital beds per thousand, and prevalence of pre-existing conditions (ex. diabetes prevalence, cardiovascular death rate, etc.). Finally, we expect that strict COVID-19 policy has effectively slowed the transmission of the virus.

These hypotheses are based on prior experiences and research. There is evidence that a high stringency index, a composite score based on how strict a country's restrictions are, slows the spread of COVID-19 [1]. We also know that patients with pre-existing conditions face a higher COVID-19 mortality rate [2, NEED A SOURCE!!!!!].

Data Description

We selected a data set from "Our World in Data." Each observation in the data set shows relevant COVID-19 data for a particular country on a given date. The COVID-19 data in the data set includes total deaths, total cases, new deaths, new cases, total cases per million, total deaths per million, total tests, new tests, total tests per thousand, positive rate, as well as telling country numbers such as stringency index (composite measure of government strictness policy) and hospital beds per thousand. Additionally, the data set includes country characteristics including population density, median age, GDP per capita, diabetes prevalence, life expectancy, and extreme poverty rate. While the previous variables are quantitative, the data set also includes categorical variables when it comes to geography such as the country and continent.

"Our World In Data" uses data from the European Center for Disease Prevention and Control (ECDC), a world leader for COVID-19 data. The ECDC has a team of epidemiologists that works every day to screen up to 500 sources to get the latest figures. These sources include ministries of health (43%), websites of public health institutes (9%), websites of public health institutes (6%), World Health Organization (WHO)

websites, WHO situation reports (2%), and official dashboards and interactive maps from national and international institutions (10%). The EDEC also utilizes social media accounts maintained by national authorities, ministries of health, and official media outlets (30%). These social media sources are screened and validated by the other sources mentioned previously. The data is recorded daily, and we will be using the data set updated as of October 9, 2020 (10:30, London time).

We used latitude data from the website "Kaggle" to supplement the COVID-19 data set. This data was collected from a Google data set and was merged with the "Our World in Data" COVID-19 data set to create one larger data set.

Here is a glimpse of our data set:

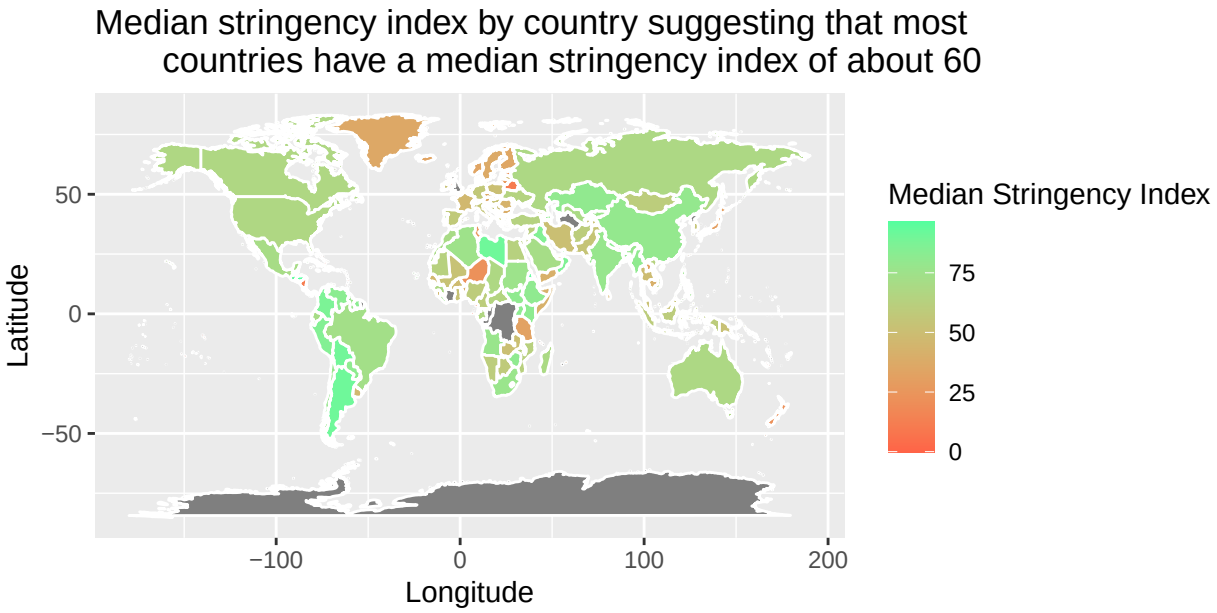
```
## Rows: 49,016
## Columns: 41
## $ iso_code                          <chr> "ABW", "ABW", "ABW", "ABW", "ABW", ...
## $ continent                         <chr> "North America", "North America", "...
## $ location                          <chr> "Aruba", "Aruba", "Aruba", "Aruba",...
## $ date                              <date> 2020-03-13, 2020-03-19, 2020-03-20...
## $ total_cases                       <dbl> 2, NA, 4, NA, NA, NA, 12, 17, 19, 2...
## $ new_cases                         <dbl> 2, NA, 2, NA, NA, NA, 8, 5, 2, 9, 0...
## $ new_cases_smoothed                <dbl> NA, 0.286, 0.286, 0.286, 0.286, 0.2...
## $ total_deaths                      <dbl> 0, NA, 0, NA, NA, NA, 0, 0, 0, 0, 0...
## $ new_deaths                        <dbl> 0, NA, 0, NA, NA, NA, 0, 0, 0, 0, 0...
## $ new_deaths_smoothed               <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ total_cases_per_million           <dbl> 18.733, NA, 37.465, NA, NA, NA, 112...
## $ new_cases_per_million             <dbl> 18.733, NA, 18.733, NA, NA, NA, 74....
## $ new_cases_smoothed_per_million    <dbl> NA, 2.676, 2.676, 2.676, 2.676, 2.6...
## $ total_deaths_per_million          <dbl> 0, NA, 0, NA, NA, NA, 0, 0, 0, 0, 0...
## $ new_deaths_per_million            <dbl> 0, NA, 0, NA, NA, NA, 0, 0, 0, 0, 0...
## $ new_deaths_smoothed_per_million   <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ new_tests                         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ total_tests                       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ total_tests_per_thousand          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ new_tests_per_thousand            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ new_tests_smoothed                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ new_tests_smoothed_per_thousand   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ tests_per_case                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ positive_rate                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ tests_units                       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ stringency_index                  <dbl> 0.00, 33.33, 33.33, 44.44, 44.44, 4...
## $ population                        <dbl> 106766, 106766, 106766, 106766, 106...
## $ population_density                <dbl> 584.8, 584.8, 584.8, 584.8, 584.8, ...
## $ median_age                        <dbl> 41.2, 41.2, 41.2, 41.2, 41.2, 41.2,...
## $ aged_65_older                     <dbl> 13.085, 13.085, 13.085, 13.085, 13....
## $ aged_70_older                     <dbl> 7.452, 7.452, 7.452, 7.452, 7.452, ...
## $ gdp_per_capita                    <dbl> 35973.78, 35973.78, 35973.78, 35973...
## $ extreme_poverty                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ cardiovasc_death_rate             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ diabetes_prevalence               <dbl> 11.62, 11.62, 11.62, 11.62, 11.62, ...
## $ female_smokers                    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ male_smokers                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ handwashing_facilities            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ hospital_beds_per_thousand        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ life_expectancy                   <dbl> 76.29, 76.29, 76.29, 76.29, 76.29, ...
## $ human_development_index           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

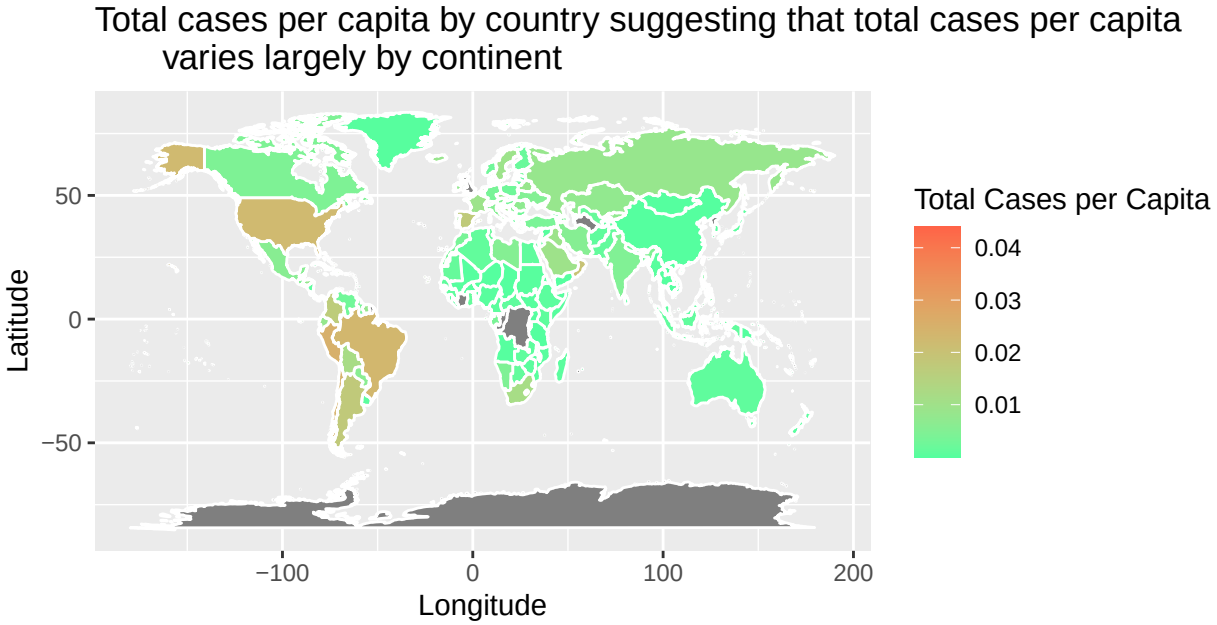
Sources: https://ourworldindata.org/coronavirus-source-data https://www.ecdc.europa.eu/en/covid-19/data-collection https://www.kaggle.com/paultimothymooney/latitude-and-longitude-for-every-country-and-state

Methodology

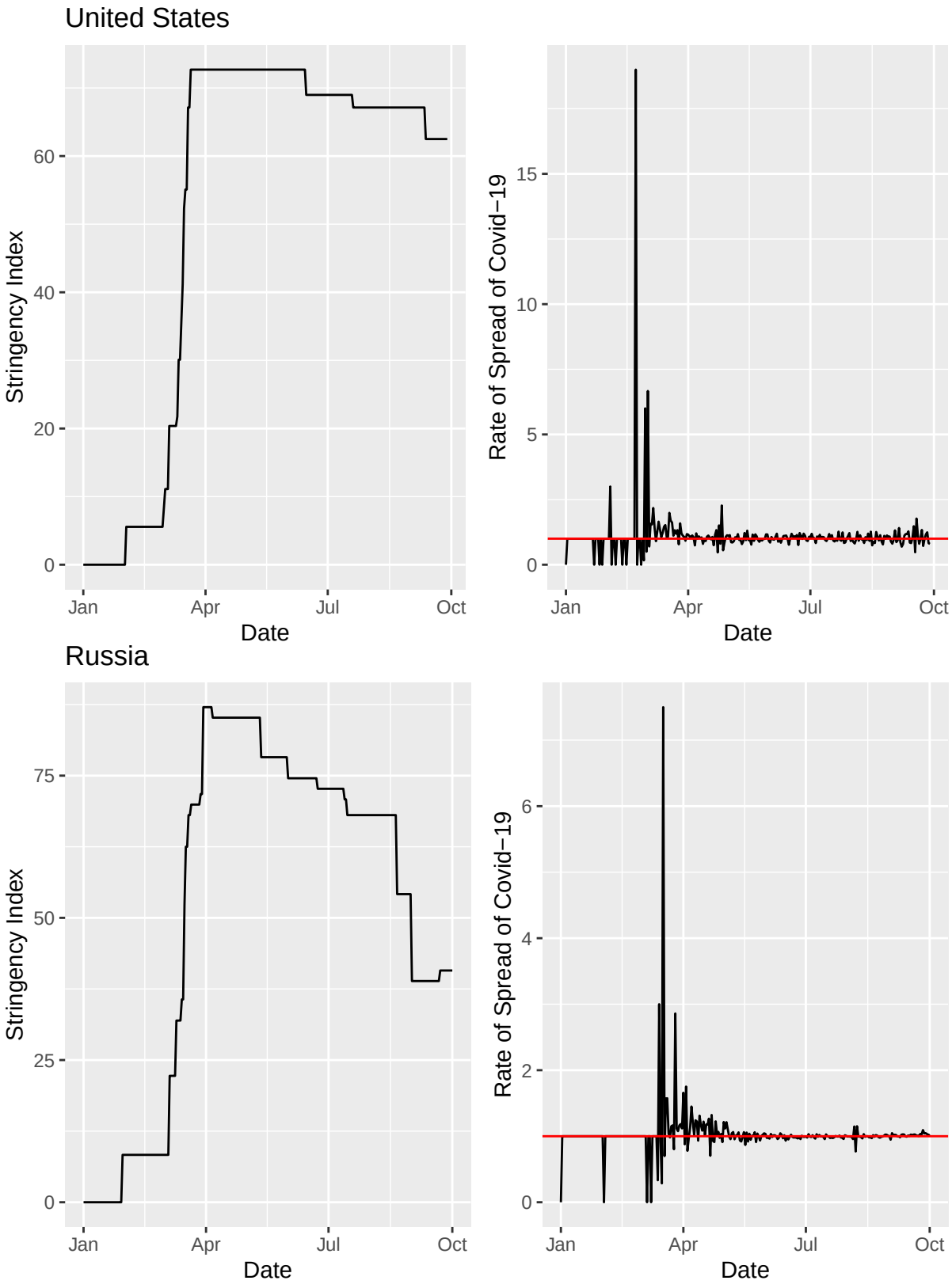
Visualizations

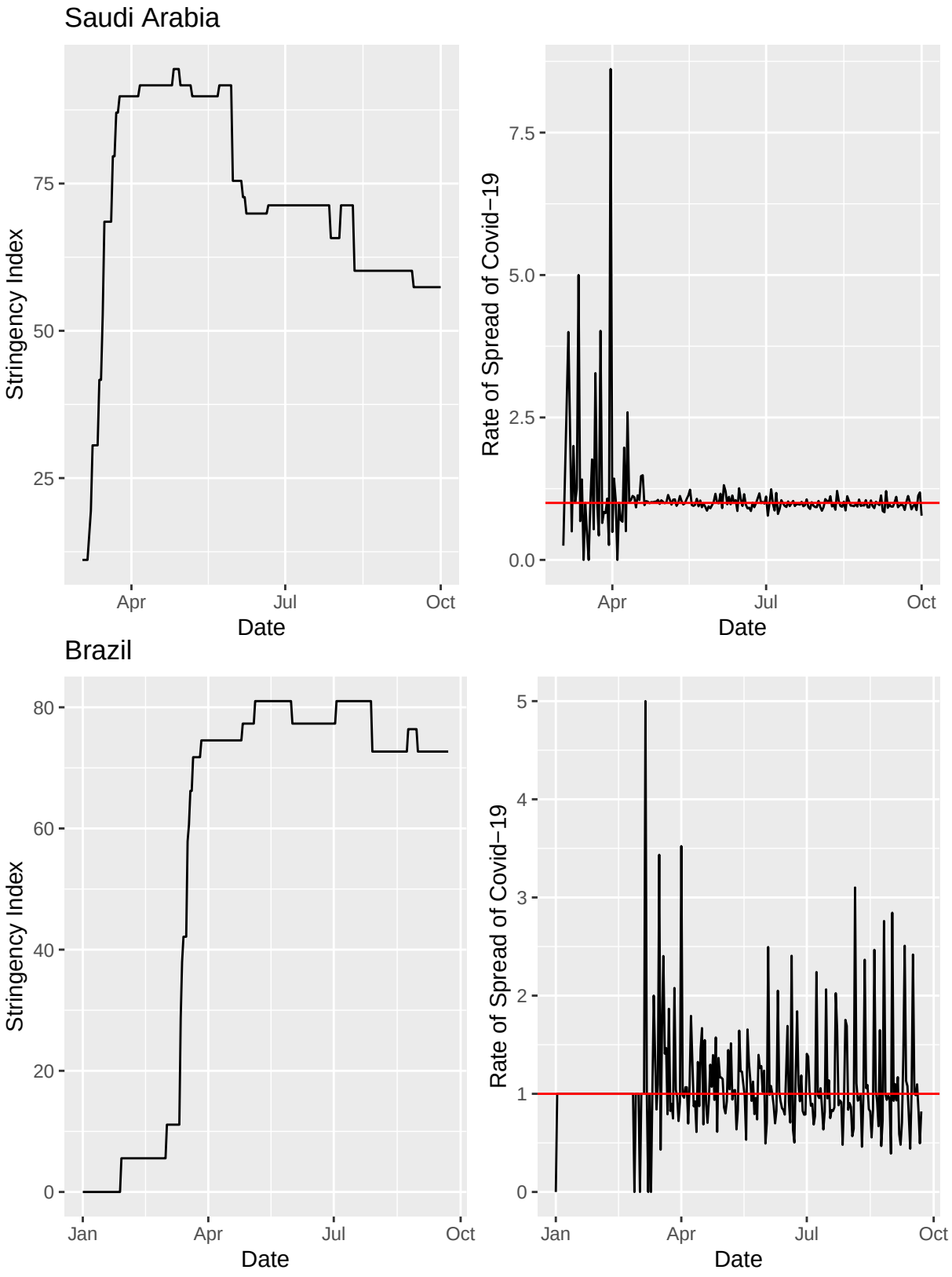
In order to visualize the relationship between stringency index and the total cases per capita we created two world map plots. The first shows the median stringency index of a country throughout the entire pandemic while the second shows the total cases per capita on October 5th, 2020.

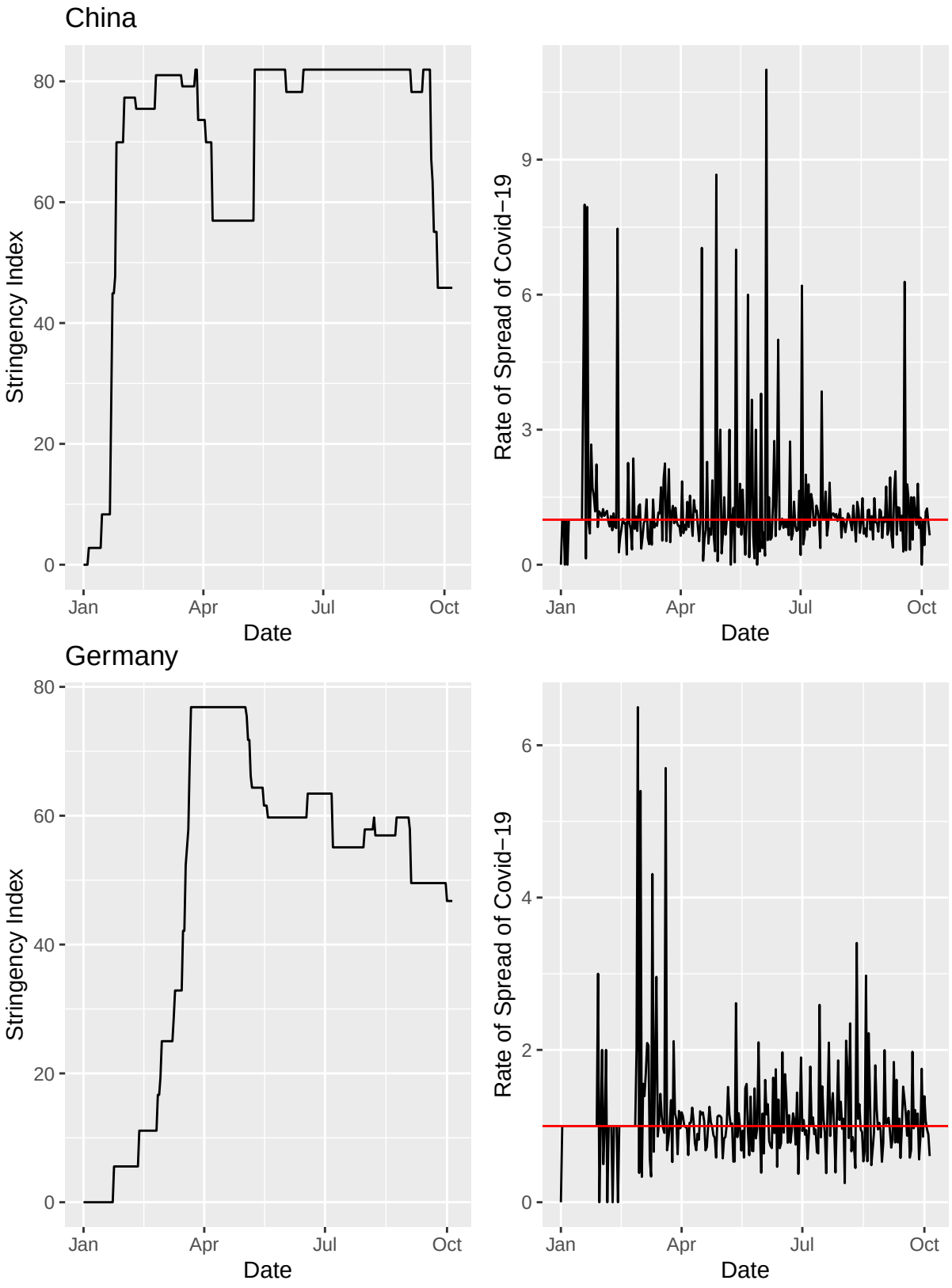




While these visuals can help capture the big picture, it also is useful to see how stringency index impacts the growth of cases over time. To visualize this, we will plot several countries that have both relatively high and low total cases per capita. The growth of cases will be represented by the factor for which the new cases changed from the previous day.

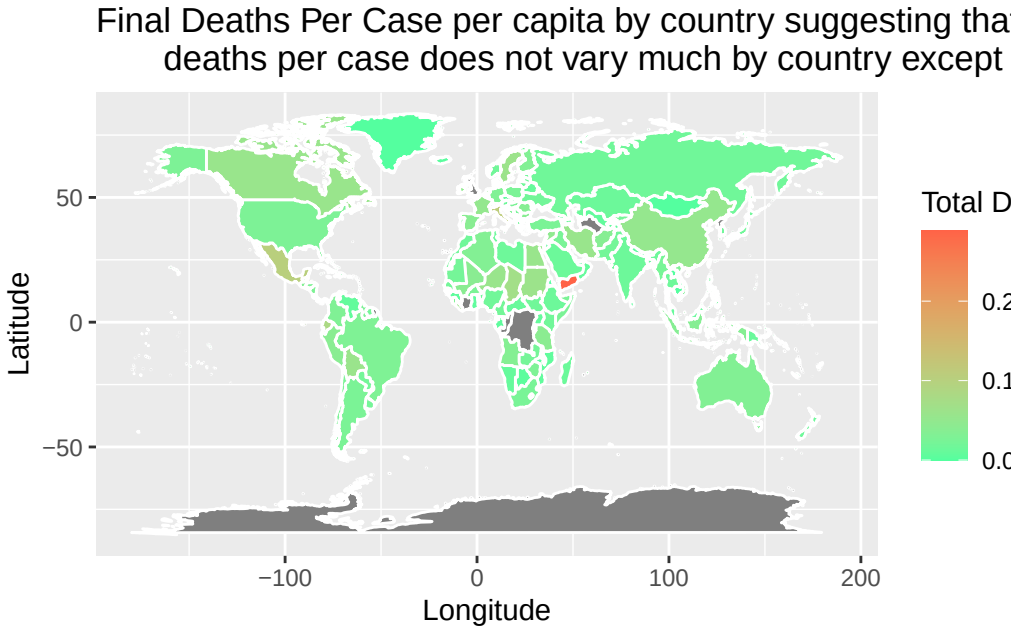

United States
Stringency Index
0
20
40
60
Jan
Apr
Jul
Oct
Date
Rate of Spread of Covid−19
0
5
10
15
Jan
Apr
Jul
Oct
Date
Russia
Stringency Index
0
25
50
75
Jan
Apr
Jul
Oct
Date
Rate of Spread of Covid−19
0
2
4
6
Jan
Apr
Jul
Oct
Date


Saudi Arabia
Stringency Index
75
50
25
Apr
Jul
Oct
Date
Rate of Spread of Covid−19
7.5
5.0
2.5
0.0
Apr
Jul
Oct
Date
Brazil
Stringency Index
80
60
40
20
0
Jan
Apr
Jul
Oct
Date
Rate of Spread of Covid−19
5
4
3
2
1
0
Jan
Apr
Jul
Oct
Date

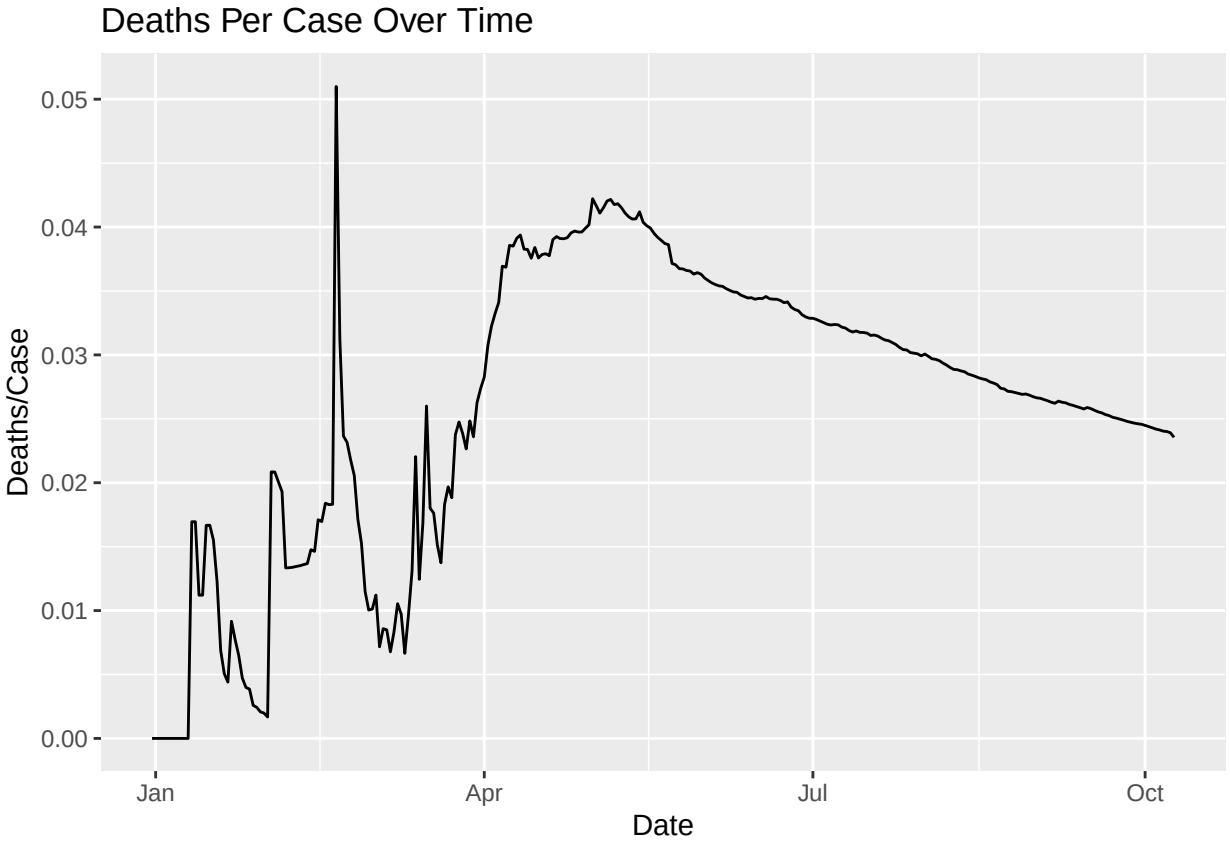


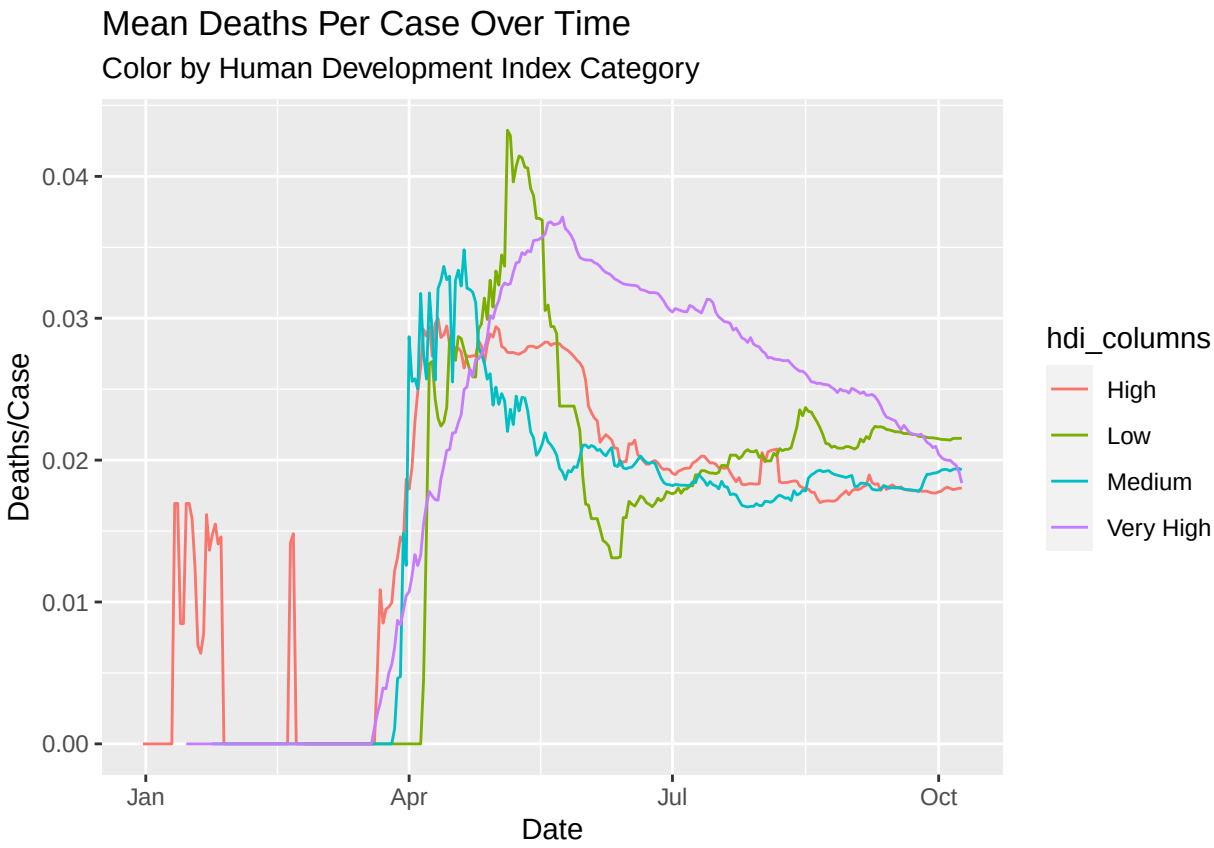
As is clear in all the graphs of individual countries, a high stringency index is associated with a lower volatility in the transmission rate of COVID-19. In the graphs of the US, Russia, and Saudi Arabia, this hold particularly true. There is a massive spike in cases during mid-March, which was roughly around the

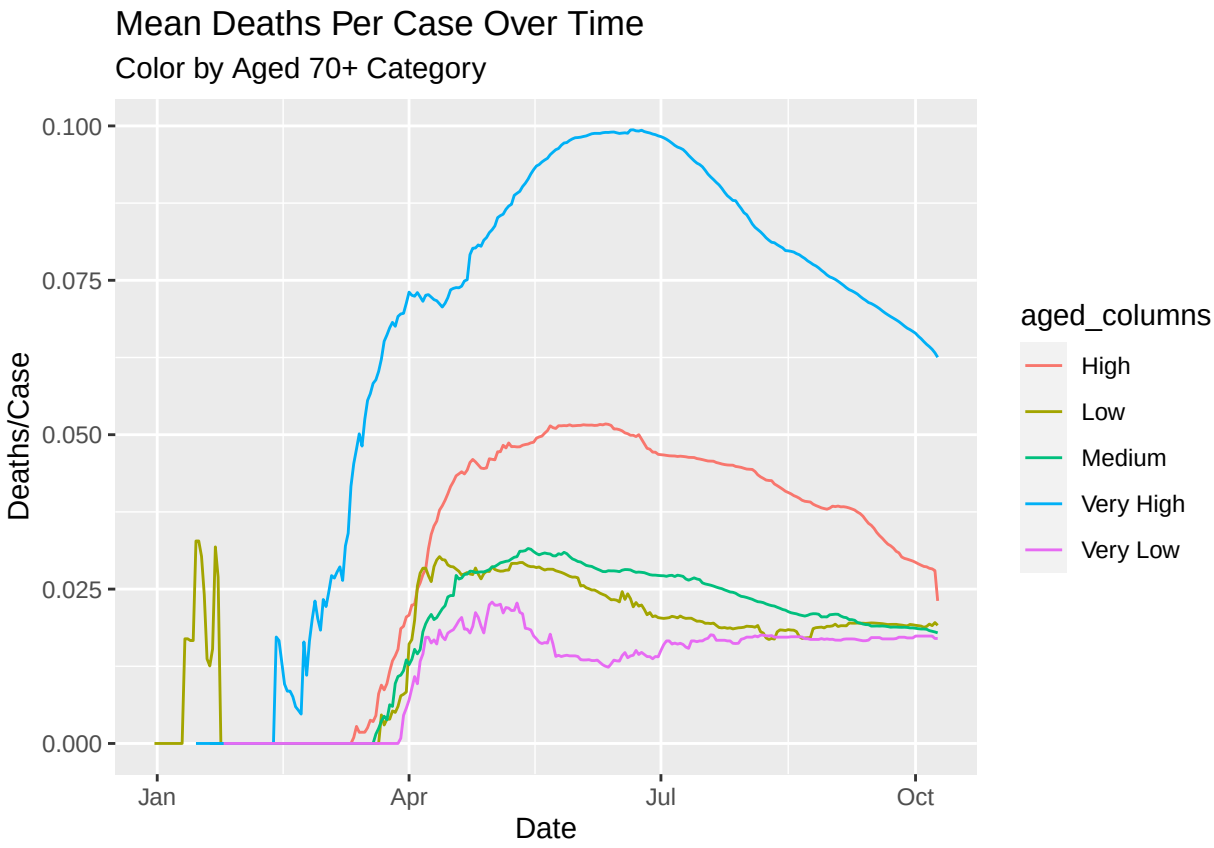
beginning of the first wave, and stringency index rises sharply. After stringency index rises, cases fall to a much lower growth rate. In the case of Brazil, China, and Germany, the graphs are a bit different. There are still major spikes around mid-March with stringency indexes going high almost immediately after. However, the transmission rates don't seem to flatten out nearly as much as they do in the US, Russia, and Saudi Arabia. There is still a good amount of volatility.



HYPOTHESIS: DEATHS PER CASE




Mean Deaths Per Case Over Time
Color by Human Development Index Category
Deaths/Case
0.04
0.03
0.02
0.01
0.00
Jan
Apr
Jul
Oct
Date
hdi_columns
High
Low
Medium
Very High

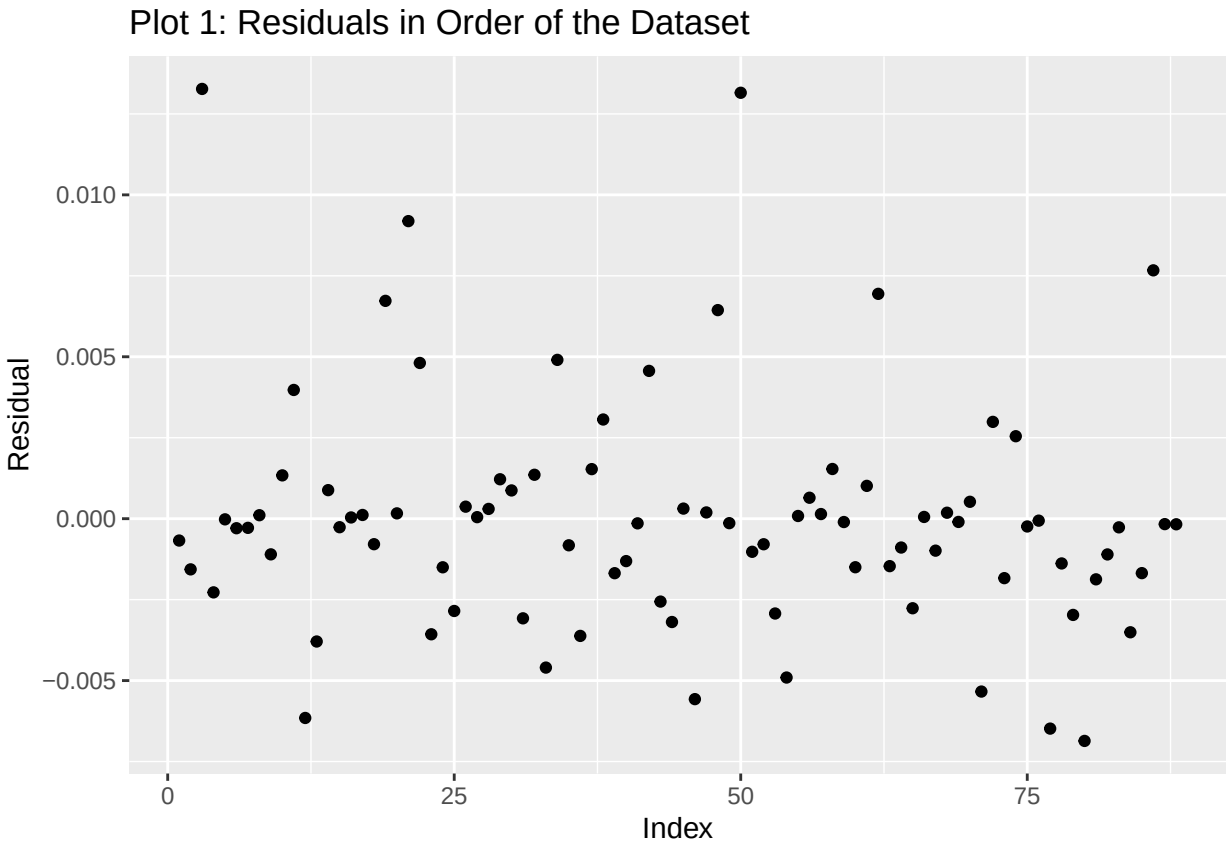

Mean Deaths Per Case Over Time
Color by Aged 70+ Category
Deaths/Case
0.100
0.075
0.050
0.025
0.000
Jan
Apr
Jul
Oct
Date
aged_columns
High
Low
Medium
Very High
Very Low

Regressions

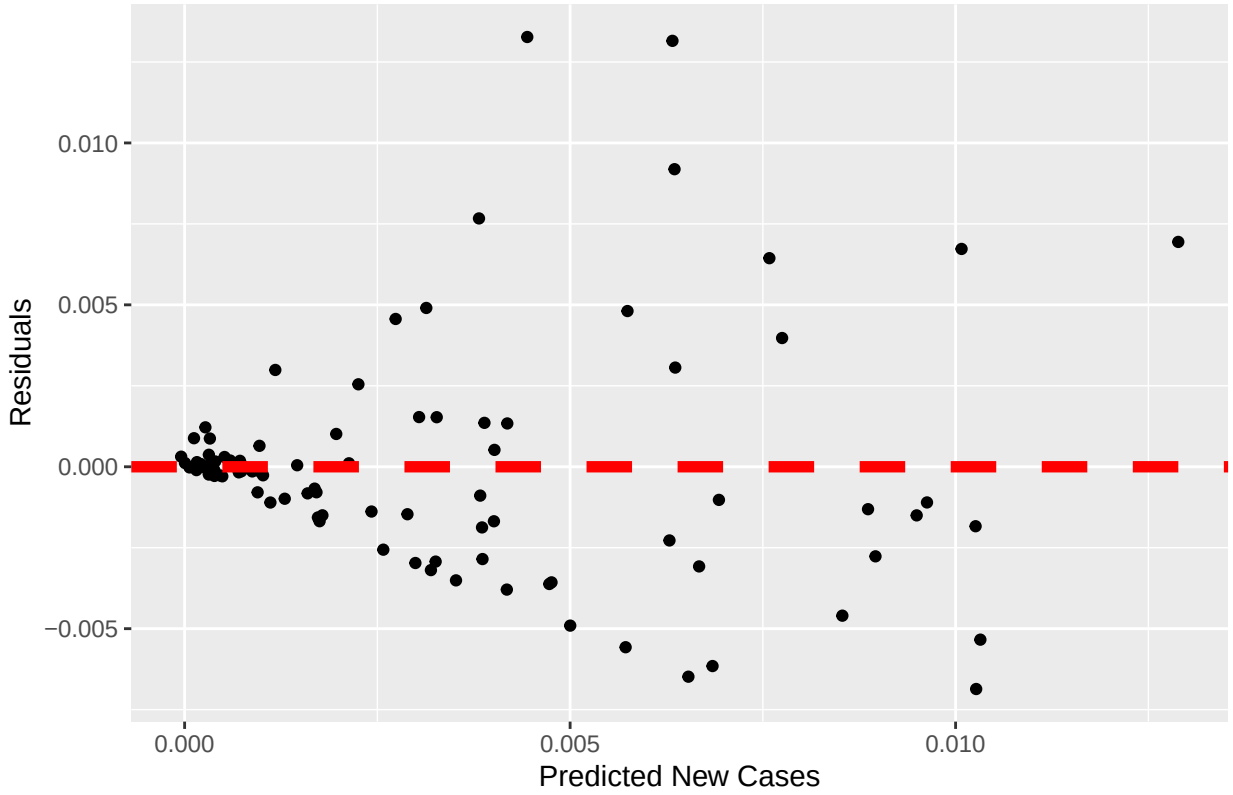
To assess which factors had the largest impact on determining a countries total cases per capita as well as test our hypothesis about total cases per capita, we will create a linear model to predict total cases per capita. We used the following linear model to try and predict total cases per capita based off of the demographics we hypothesized as well as a few additional variables.

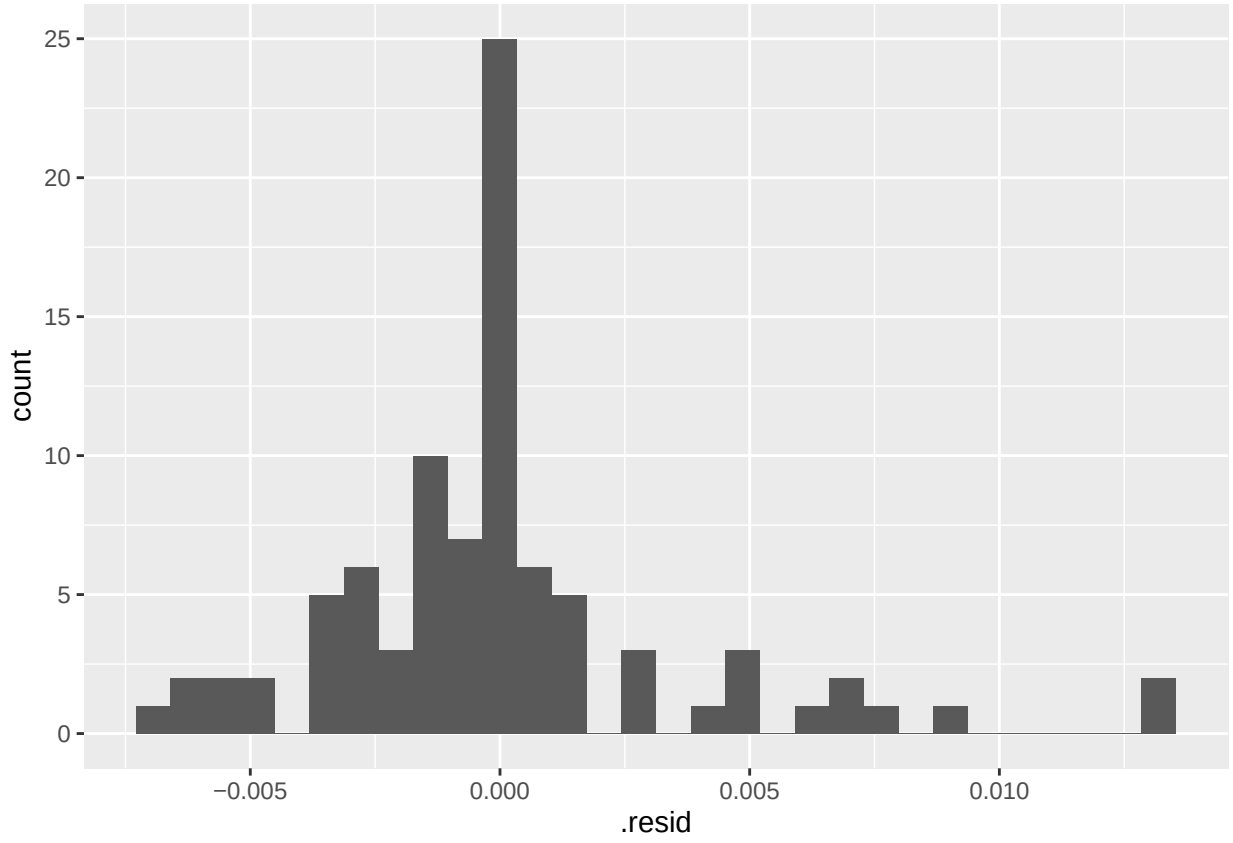
```
## # A tibble: 8 x 5
##   term                        estimate    std.error statistic  p.value
##   <chr>                          <dbl>        <dbl>     <dbl>    <dbl>
## 1 gdp_per_capita          0.000000291 0.0000000798    3.65   0.000466
## 2 continentSouth America  0.00574     0.00187         3.06   0.00298
## 3 continentEurope         0.00533     0.00263         2.03   0.0459
## 4 continentAsia           0.000978    0.00131         0.747  0.457
## 5 continentNorth America  0.00112     0.00161         0.699  0.487
## 6 handwashing_facilities  0.0000118   0.0000228       0.516  0.607
## 7 (Intercept)            -0.000198    0.00123        -0.161  0.872
## 8 median_si              -0.00000105  0.0000174      -0.0601 0.952
```

```
## # A tibble: 1 x 2
##   adj.r.squared r.squared
##           <dbl>     <dbl>
## 1         0.386     0.435
```



Plot 2: PR Plot





Chi-Squared

Our visualizations and linear models suggested there might be some sort of relationship between total cases per capita and continent. In the visualization of total cases per capita on the world map, it appeared that continents seemed to have similar total cases per capita. In the linear model predicting total cases per capita, we observed that certain continent predictor weights were statistically significant. Thus, we will test for independence between total cases per capita and continent. We will do so using a chi-squared test at the $\alpha = 0.05$ significance level.

H_0: There is independence between continent and total cases per capita.

H_0: There is NOT independence between continent and total cases per capita.

```
##
##  Pearson's Chi-squared test
##
## data:  table(covid_tcpc$continent, covid_tcpc$total_cases_per_cap)
## X-squared = 227767, df = 159400, p-value < 2.2e-16
```

Results

Discussion

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418951/#:~:text=Our%20model%20implies%20that%20social,at%2021%
[2] [3]