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].

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

Methodology

Results

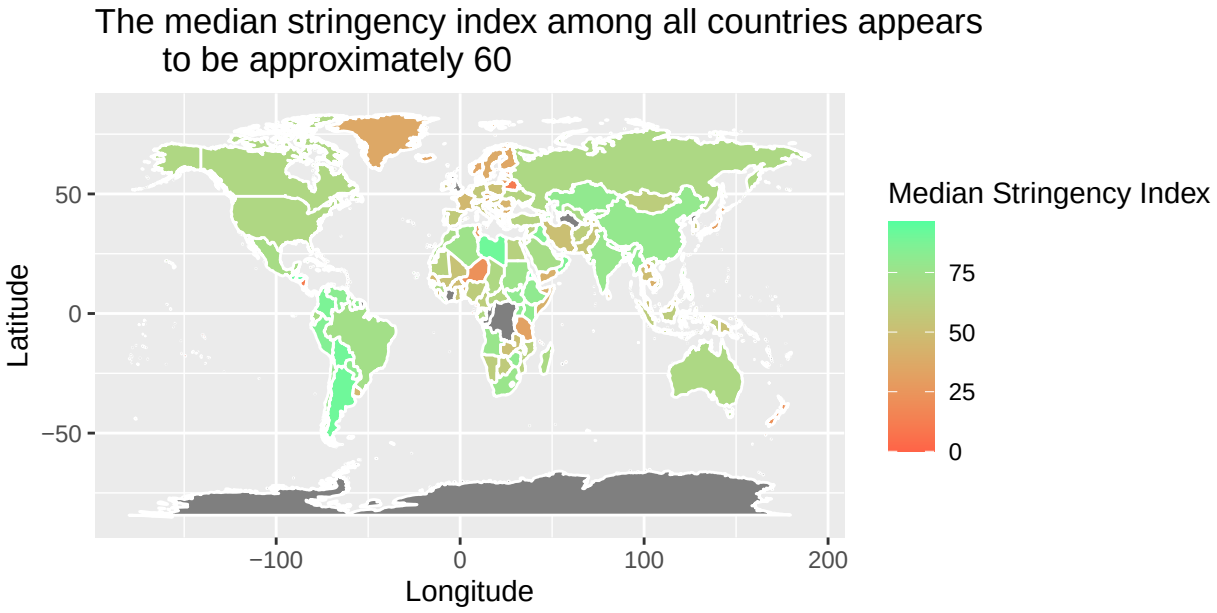
Question 1: Effectiveness of COVID-19 Related Policies

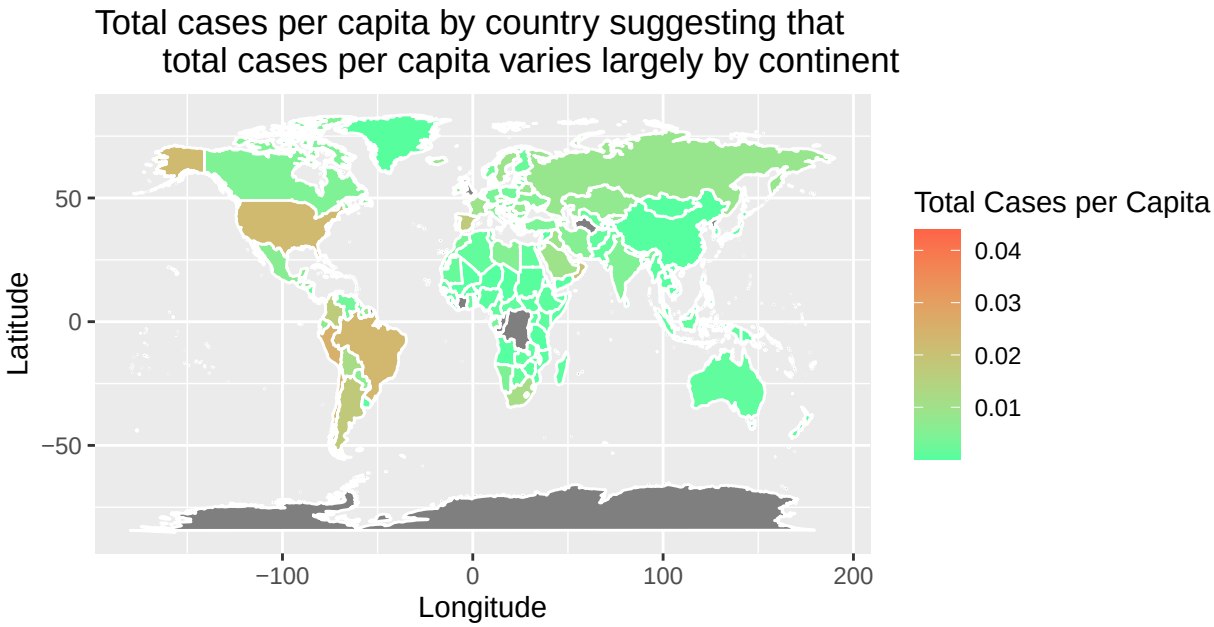
Politicians worldwide are pushing for governments to enact strict policies to mitigate the spread of COVID-19. They argue that social distancing, stay-at-home orders, mask wearing, and business closures are all crucial to flatten the curve and keep Covid-related deaths low.

A Lancet study from earlier during the pandemic (using Wuhan as a case study) found that restrictions to social activities helps delay the epidemic peak, and that lifting governmental restrictions can bring about a second peak. Thus, stricter masking policies seemingly have merit in reducing viral transmission [3]. We are therefore interested in analyzing the relationship between stringency index and both total cases per capita and the growth rate of COVID-19 cases.

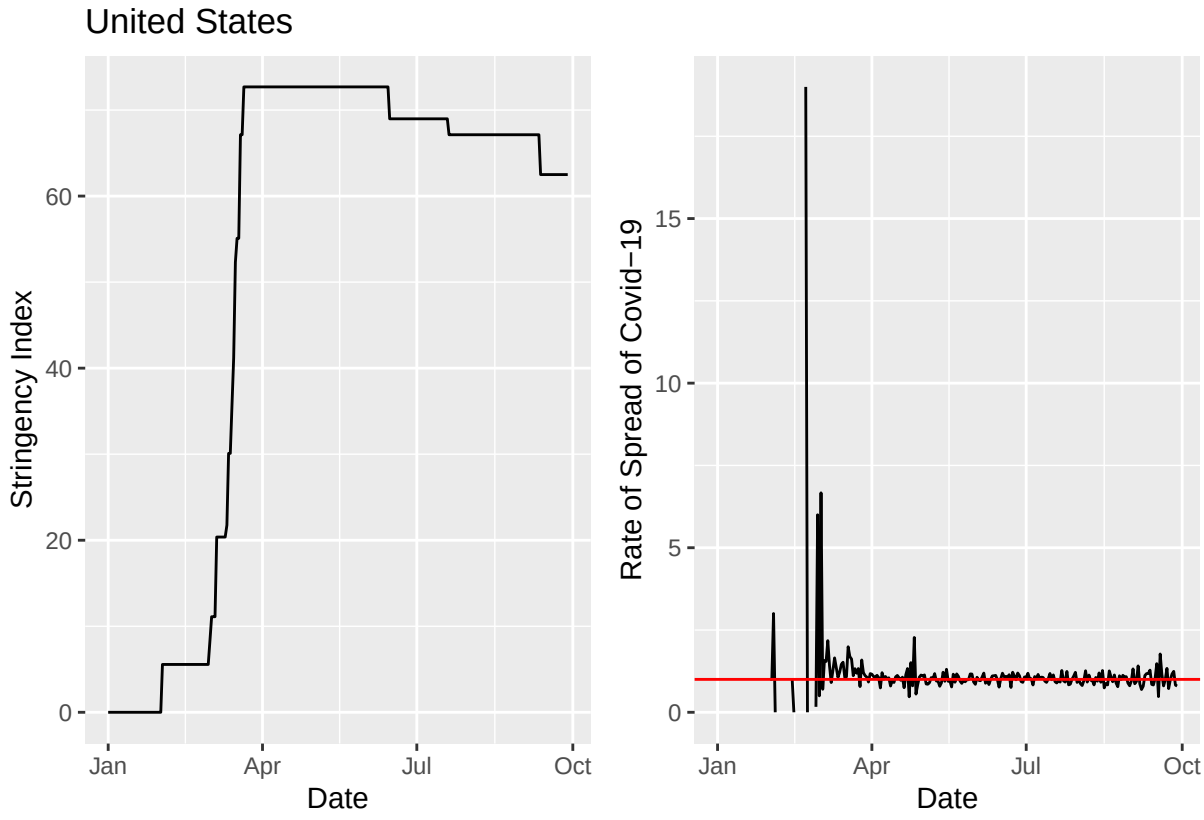
Stringency index is a composite variable based on nine other categories that determines how "strict" a country's COVID-19 prevention policies are, with 100 being the most strict and 0 meaning they have no COVID-19 related policies.

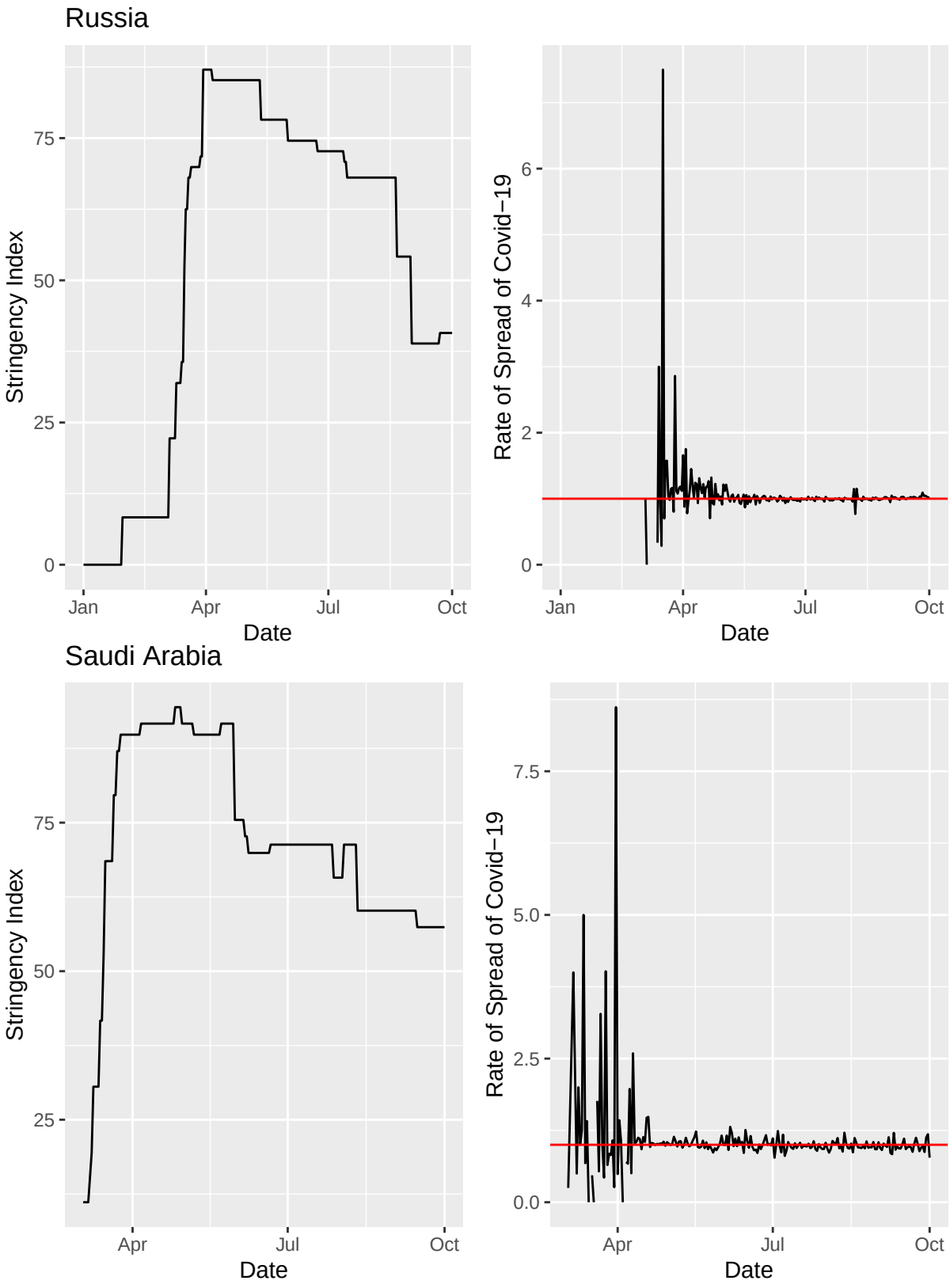
In order to visualize the relationship between a stringency index and the total cases per capita we created two world map plots. The first shows the median stringency index of a country during the entire pandemic while the second shows the total cases per capita on October 5th, 2020.

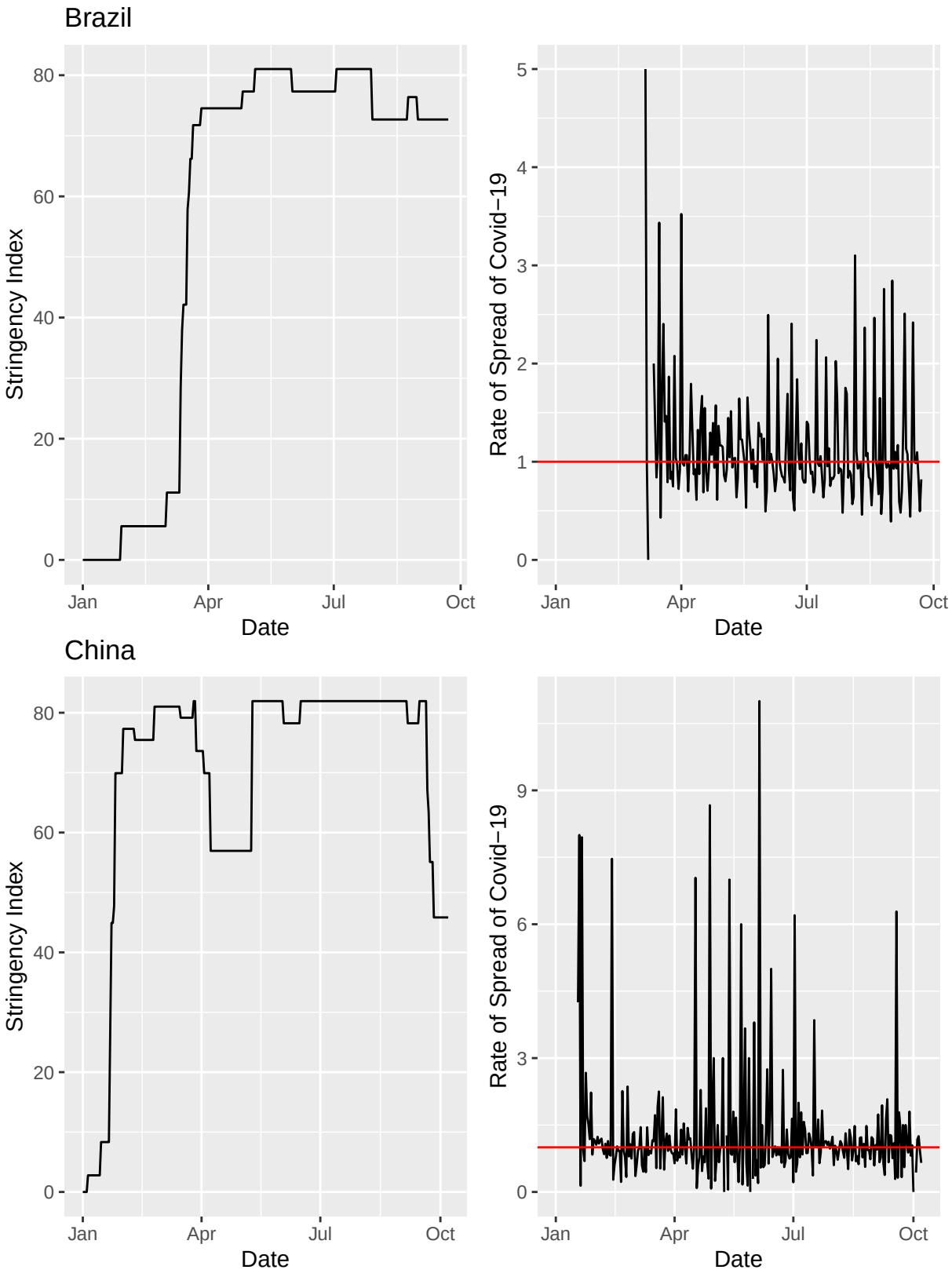


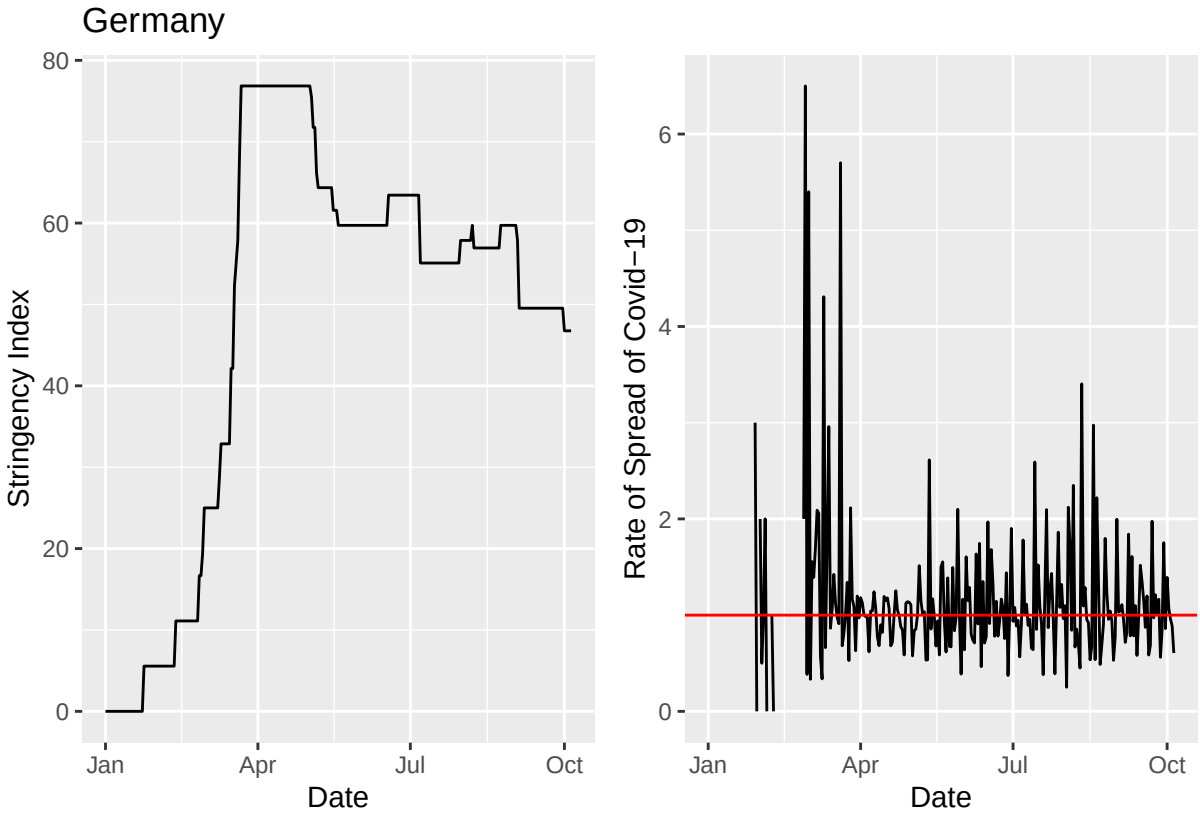


While these visuals can help capture the big picture, it also is useful to see how stringency index impacts the growth of cases over time. To visualize this, we will plot several countries that have both relatively high and low total cases per capita. The growth of cases will be represented by the factor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




Russia
Stringency Index
75
50
25
0
Jan
Apr
Jul
Oct
Date
Rate of Spread of Covid−19
6
4
2
0
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
China
Stringency Index
80
60
40
20
0
Jan
Apr
Jul
Oct
Date
Rate of Spread of Covid−19
9
6
3
0
Jan
Apr
Jul
Oct
Date



For all countries there is a massive spike in cases during mid-March. This can be explained by the fact that cases were exploding during those few weeks globally. It was also around the same exact time that the World Health Organization declared COVID-19 a global pandemic because they were "deeply concerned by the alarming levels of spread and severity of the outbreak" and "the alarming levels of inaction" [4].

It should be noted that the red line on the "Rate of Spread of COVID-19" graphs represents a growth factor of 1, meaning that at this line cases per day was unchanged from one day to the next. Also most graphs contain large gaps in data pre mid-March. This is because data on new COVID-19 cases was not being consistently reported each day before then. Thus, there is no information to plot.

Question 2: Which Demographics and Characteristics Impact Total Cases per Capita the Most?

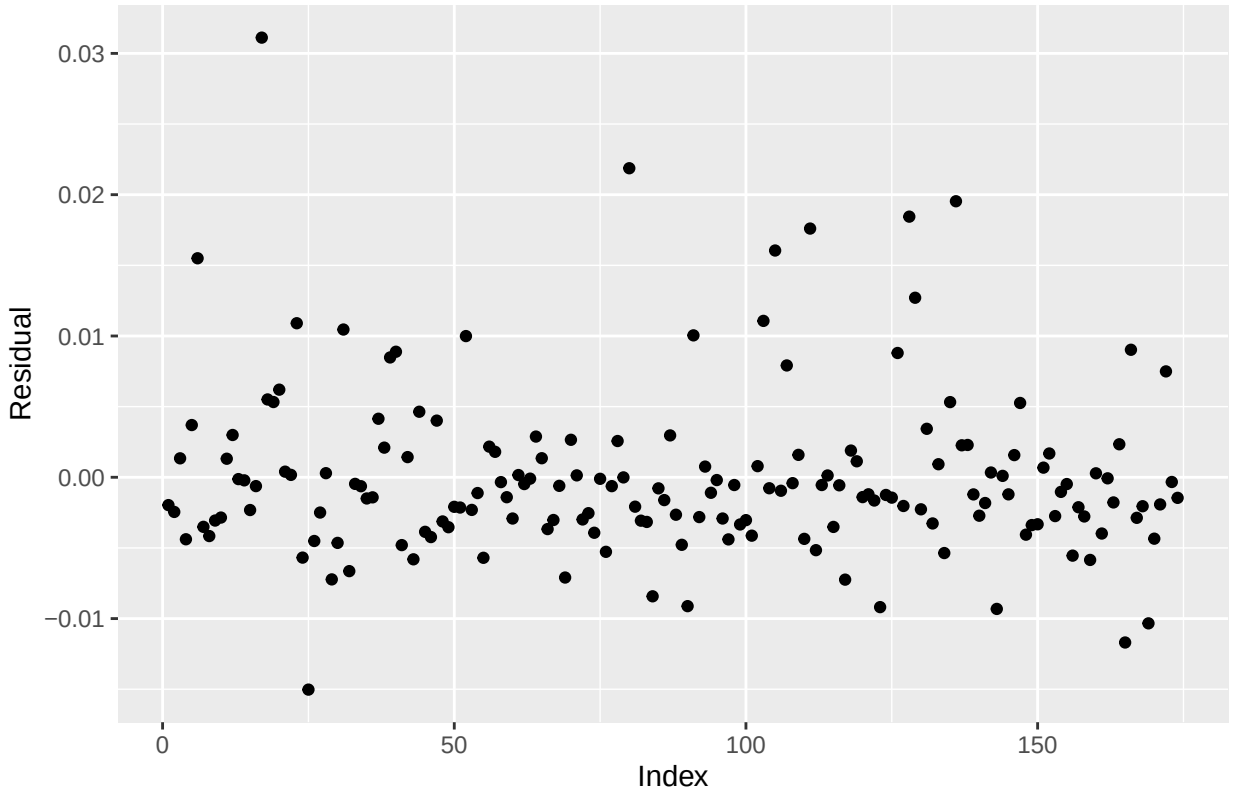
To assess which factors had the largest impact on determining a countries total cases per capita as well as test our hypothesis about total cases per capita, we will create a linear model to predict total cases per capita. We used stringency index, GDP per capita, human development index, and population density as predictors since they were all in our hypothesis.

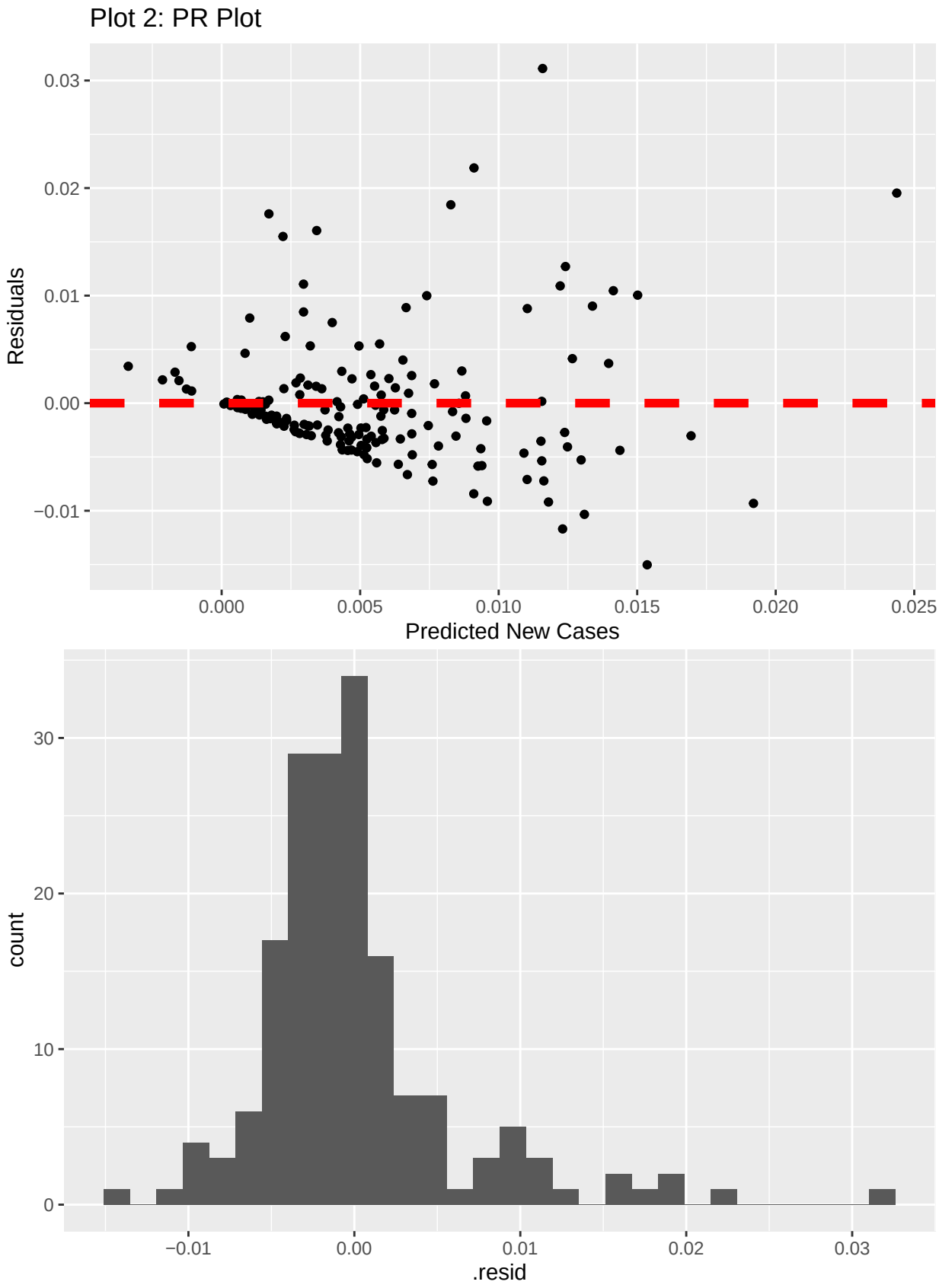
```
## # A tibble: 10 x 5
##    term                        estimate    std.error statistic   p.value
##    <chr>                          <dbl>        <dbl>     <dbl>     <dbl>
##  1 gdp_per_capita           0.000000172 0.0000000400   4.30    0.0000294
##  2 continentSouth America    0.00796     0.00235       3.39    0.000867
##  3 median_si                 0.0000426   0.0000205     2.07    0.0396
##  4 continentOceania         -0.00450     0.00344      -1.31    0.193
##  5 continentAsia             0.00196     0.00156       1.25    0.212
##  6 continentNorth America    0.00208     0.00189       1.10    0.273
##  7 (Intercept)              -0.00300     0.00379      -0.792   0.430
##  8 human_development_index   0.00229     0.00655       0.350   0.727
##  9 population_density        0.000000150 0.000000791   0.190   0.850
```

```
## 10 continentEurope          0.0000130   0.00205        0.00634 0.995
```

```
## # A tibble: 1 x 2
##   adj.r.squared r.squared
##           <dbl>     <dbl>
## 1         0.309     0.345
```

Plot 1: Residuals in Order of the Dataset




Plot 2: PR Plot
Residuals
0.03
0.02
0.01
0.00
−0.01
0.000
0.005
0.010
0.015
0.020
0.025
Predicted New Cases
count
30
20
10
0
−0.01
0.00
0.01
0.02
0.03
.resid

Extra Analysis: Total Cases per Capita vs. Continent

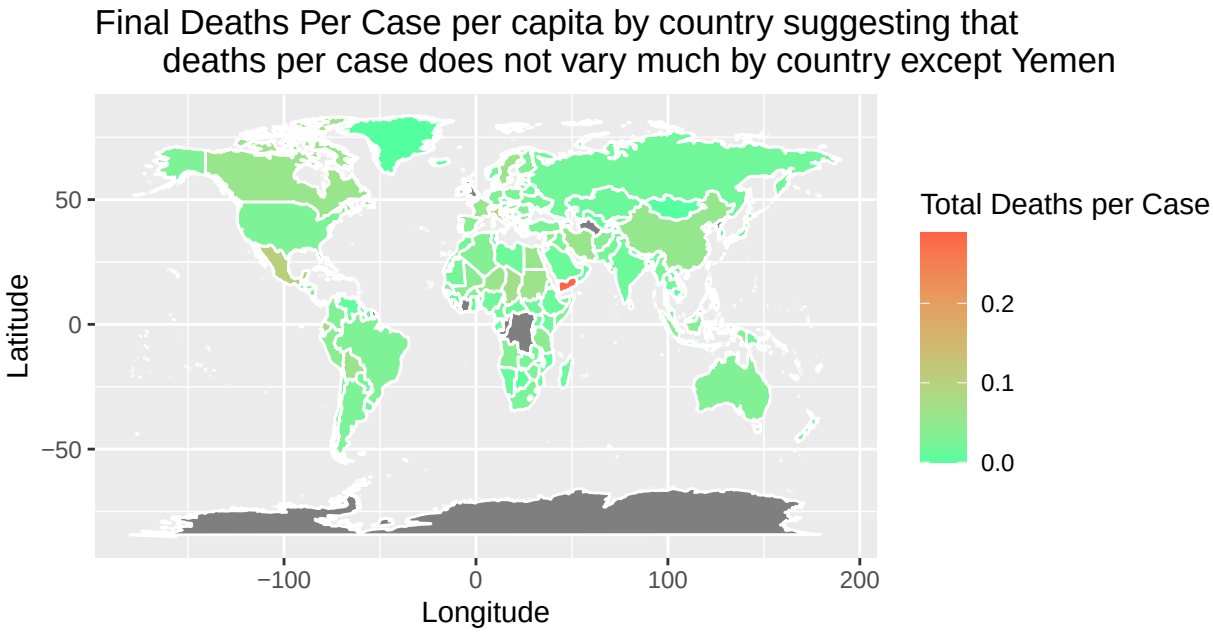
Our visualizations and linear models suggest there might be some sort of relationship between total cases per capita and continent. In the visualization of total cases per capita on the world map, it appeared that continents seemed to have similar total cases per capita. In the linear model predicting 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.

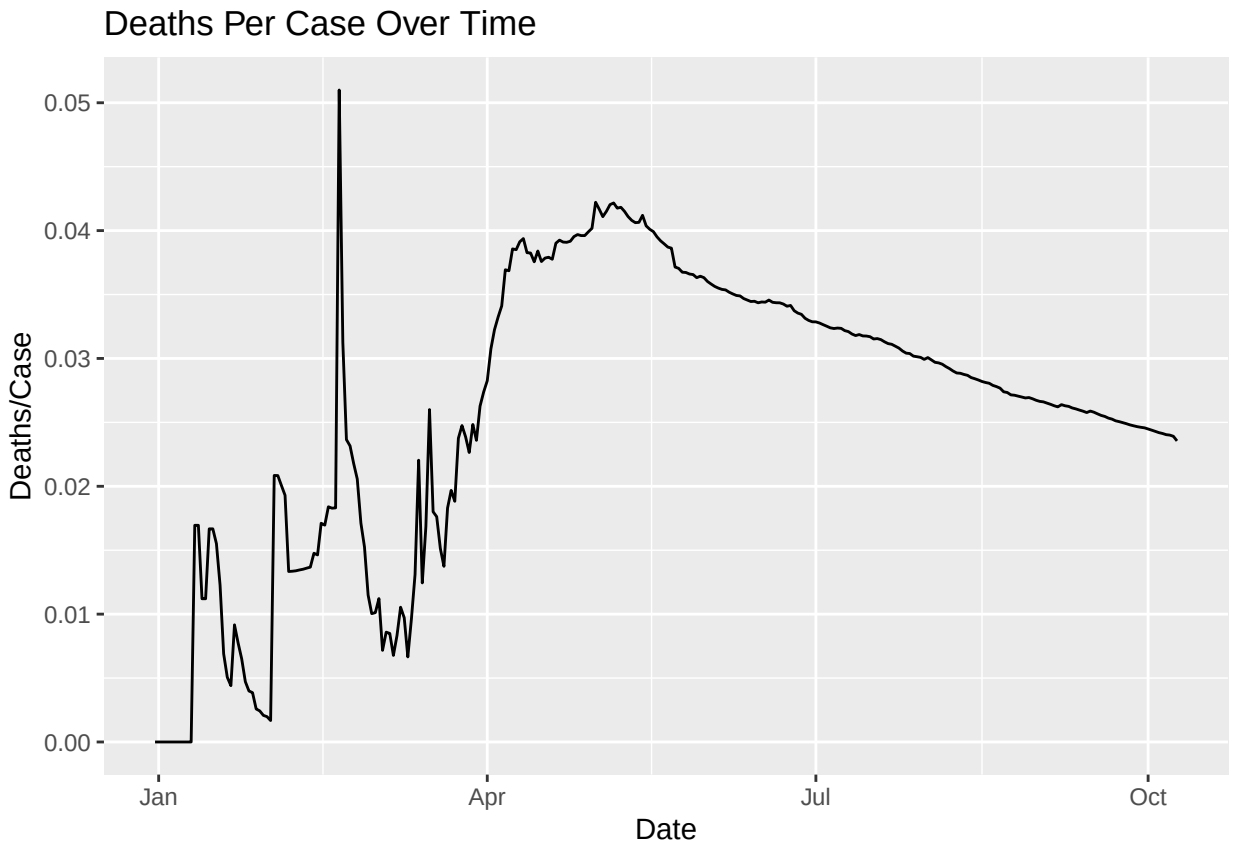
H_0: There is independence between continent and total cases per capita.

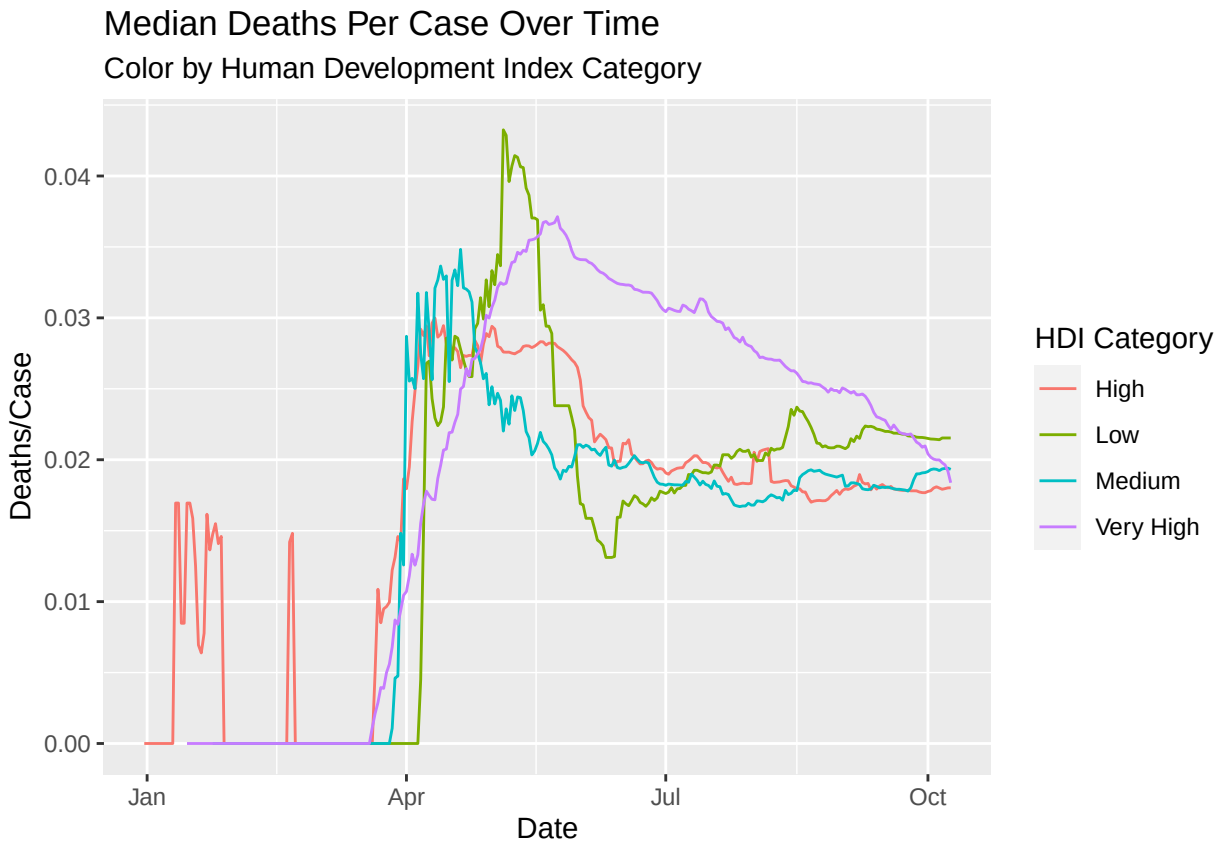
H_1: There is NOT independence between continent and total cases per capita.

```
##
##  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
```

Question 3: Which Demographics and Characteristics Impact Deaths per Case the Most?







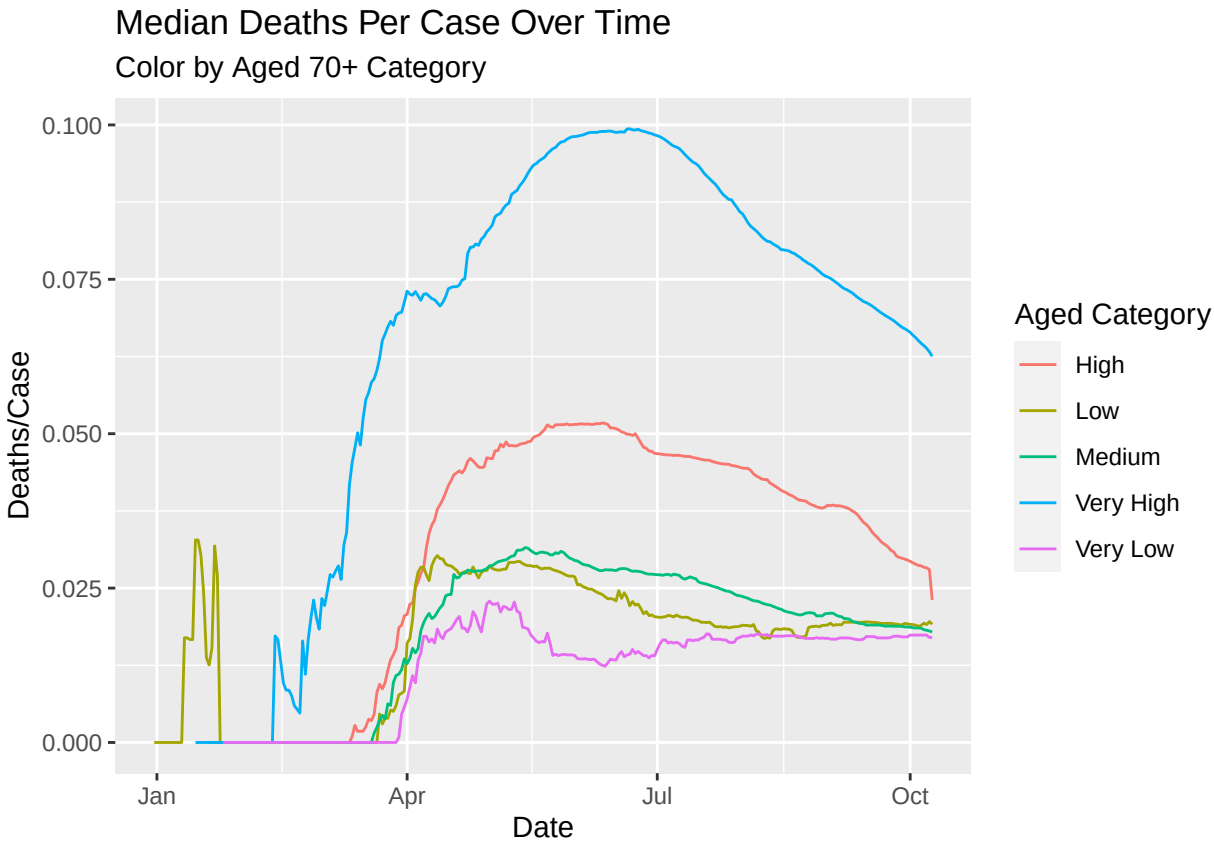
In visualizing the line plot of deaths per case over time colored by human development categories, there are

clear distinctions between the different categories, especially around May and June. Thus, we will test for independence between human development index category and deaths per case. We will do so using a chi-squared test at the $\alpha = 0.05$ significance level.

H_0: Human development index category and deaths per case are independent.

H_0: Human development index category and deaths per case are NOT independent.

```
## [1] 2.487605e-139
```



In visualizing the line plot of deaths per case over time colored by categories denoting the percentage of the country's population that is aged 70 or older, there are stark contrasts in the line throughout the entire duration of the pandemic, with higher percentages of populations aged 70 or older having higher median deaths per case. Thus, to test for statistical significance, we will perform a test for independence between aged category and deaths per case. We will do so using a chi-squared test at the $\alpha = 0.05$ significance level.

H_0: Human development index and deaths per case are independent.

H_0: Human development index and deaths per case are NOT independent.

```
## [1] 0
```

```
## # A tibble: 9 x 5
##   term                        estimate  std.error statistic p.value
##   <chr>                          <dbl>      <dbl>     <dbl>   <dbl>
## 1 (Intercept)              -0.0433     0.163         -0.266  0.795
## 2 diabetes_prevalence       0.00257    0.00308        0.834  0.419
## 3 handwashing_facilities    0.000485   0.000395       1.23   0.241
## 4 life_expectancy           0.00532    0.00352        1.51   0.155
## 5 cardiovasc_death_rate    -0.000147   0.0000885     -1.66   0.121
## 6 female_smokers           -0.00384    0.00218       -1.76   0.101
## 7 human_development_index -0.413      0.208         -1.99   0.0682
```

```
## 8 gdp_per_capita           0.00000678 0.00000301     2.25   0.0423
## 9 aged_70_older           -0.0253     0.0111        -2.29   0.0397
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.498         0.189
```

Discussion

Hypothesis 1

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. 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 2

Hypothesis 3

Map:

Figure 1:

Figure 2 (HDI):

Chi-Square Test (HDI):

Our test statistic was 83240, which has a chi-square distribution with 73212 degree of freedom under H_0. This correlates with a P-value approximately equal to 2.487605e-139 which is less than $\alpha = 0.05$, such that we reject the null hypothesis. There is sufficient evidence to suggest that human development index category and deaths per case are not independent.

Figure 3 (Aged):

Chi-Square Test (Aged):

Our test statistic was 126458, which has a chi-square distribution with 99284 degrees of freedom under H_0. This correlates with a P-value approximately equal to 0, which is less than $\alpha = 0.05$, such that we reject the null hypothesis. There is sufficient evidence to suggest that aged category and deaths per case are not independent.

###Limitations

All: * Lot of missing data for different variables, dates, can mess up the results * Time series data- have less understanding as to how to work with it * Dynamic and novel virus- hard to know what most impacts spread, death rate with so many variables changing * Different strains of coronavirus- can be a variable in terms of fatality rate * Cut off dataset in October, however, lots of new information has come in on coronavirus since then that may be helpful * Coronavirus data reporting- many countries do not have the comprehensive testing and thus our numbers for total cases/capita and deaths/case will be off

Regressions: * Doesn't pass diagnostic plots * Low R^2 or variance accounted for in regressions- not too many variables, missing important variables * Limited dataset * Still can not attribute causation- may be confounding variables

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418951/#:~:text=Our%20model%20implies%20that%20social,at%2021%
[2] https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.htm
l#:~:text=Adults%20of%20any%20age%20with%20the%20following%20conditions%20are%20at,COPD%20(chronic%20obstruc
[3] https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(20)30073-6/fulltext [4]
https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020 [5]