Final Project: SARS-CoV-2 Dataset

Due Friday, November 20, 11:59 PM

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

Introduction

Throughout the year 2020, the SARS-CoV-2 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 SARS-CoV-2 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 SARS-CoV-2. 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 SARS-CoV-2 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 SARS-CoV-2 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 SARS-CoV-2 [9]. We also know that patients with pre-existing conditions face a higher SARS-CoV-2 mortality rate [10].

Data Description

We selected a data set from "Our World in Data." Each observation in the data set shows relevant SARS-CoV-2 data for a particular country on a given date. The SARS-CoV-2 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 SARS-CoV-2 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).

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,...
```

Methodology

We will attempt to answer all of our research questions and hypotheses with similar approaches. We will create visualizations to graphically show any relationships we wish to analyze. We then plan to use linear

models to analyze which variables in particular have relationships with total cases per capita and deaths per case.

Our graphs will consist of both line plots and world maps. The world maps will serve to show different variables with respect to geography. This will provide clear visualizations for us to identify countries and regions with high numbers of cases per capita and deaths per case. We can use maps comparatively with stringency index and cases per capita to see if median stringency index (chose median because resistant to outliers) and cases per capita are related, helping us analyze the effect of governmental policy on the transmission of SARS-CoV-2.

Our line plots will be used to visualize relationships between two continuous variables. We would like to visualize our response variable over time, looking at deaths per case as the pandemic progresses. We can visualize the change in deaths per case over time when color is based on a categorical variable, allowing us to see not only the relationship between the variable and the response (deaths per case), but also how the relationship changes over time.

Our regressions will be simple linear models with predictor variables that we hypothesized. We do not plan on including interaction variables in our models. We plan on adjusting our models to try and improve R^2 to a level we are satisfied with while also keeping them simple. This is due to the fact that simpler models are more interpretable, less overfit, and can reduce collinearity among the variables. This will help us model the relationship between various country characteristics including age, population-related statistics (density, total), pre-existing conditions (cardiovascular death rate, diabetes prevalence), and wealth/quality of public health system (GDP per capita, life expectancy, hospital beds per thousand, human development index), and see how these relate with total cases per capita and deaths per cases.

In making our regression for deaths per cases (aka case fatality rate), we choose to use deaths per cases using all deaths and cases accrued up until our cutoff date (October 5), as this provides a more stable variable to predict that represents a country's response overall. Deaths per cases were essentially 0 for many countries for so long due to a lack of cases (and lack of reporting for many countries), such that to model deaths per cases and determine relevant variables, we must model it later in the pandemic.

For our data analysis, we will look at stringency index, which is a composite variable based on nine other categories that determines how "strict" a country's SARS-CoV-2 prevention policies are, with 100 being the most strict and 0 meaning they have no SARS-CoV-2 related policies. These ordinal measurements of government responses include school closures, workplace closures, canceling of public events, restriction on gatherings, stay at home events, internal travel restrictions, international travel restrictions, and face covering policies [8]. Thus, it is able to provide us a comparative variable as to what the extent and severity a government's COVID response was. This will help us to analyze the relationship between country's government COVID response and the growth rate of SARS-CoV-2.

Additionally, we seek to find valuable predictors as to a country's total cases per capita and deaths per cases. Given the sheer number of variables, we can try hypothesis tests and bayesian paradigms on all of them, and thus we use a multiple linear regression to get an idea of the variables that have more statistically significant relationships with our response variable (deaths per case and cases per capita).

We can then extract these variables, and then test to confirm a relationship between these variables and the response.

With deaths per cases, we categorize our significant variables into categories before visualizing them on an aforementioned line plot, to observe how the relationship with the response variable by analyzing the difference between the lines for each category. Additionally, in plotting against time, we can see how that relationship between predictors and response changes over time.

Furthermore, we can test for a statistically significant relationship between a predictor and the response. This is done through a chi square test for independence, which determines if there is a statistically significant relationship or association between two categorical variables by calculating the distance between the observed response variable and expected response variable if the two variables are independent, and then squaring that difference (so its summation is positive regardless of if observed is greater than or less than expected),

scaling it (based on expected value), and summing it for all observations. It is similar to a hypothesis test in that you see how likely our data (or more extreme) is likely to occur given that the null hypothesis (of independence) is true, providing us with a P-value to reject or fail to reject our null hypothesis.

To do this, we categorize our predicted and response variables to meet the conditions for a chi square test for independence, creating bins that denote the extent of the magnitude of a country's predictor or respondent (cases per capita and deaths per case). In trying to make our statistical analysis unbiased, we try to make these choices through non-arbitrary measures, such as seeing how the agencies or who report the statistics or other reputable sources generally categorize them. In performing a chi square test for independence, this will allow us to assess which variables are statistically significant with deaths per case and cases per capita, helping us answer our research question.

Given that the data is time series and the dynamic nature of variables such as cases per capita or deaths per cases, we chose to perform chi square tests for independence using only the last week of data. This is because we categorized these variables based on percentile, however, the variables change so much that a high deaths per case in February is incredibly different from another month. In only using the last week, we know that our categories for our response variables are actually applicable to the data (something categorized as high will actually be high) for the duration used in the chi square test, reducing potential human bias.

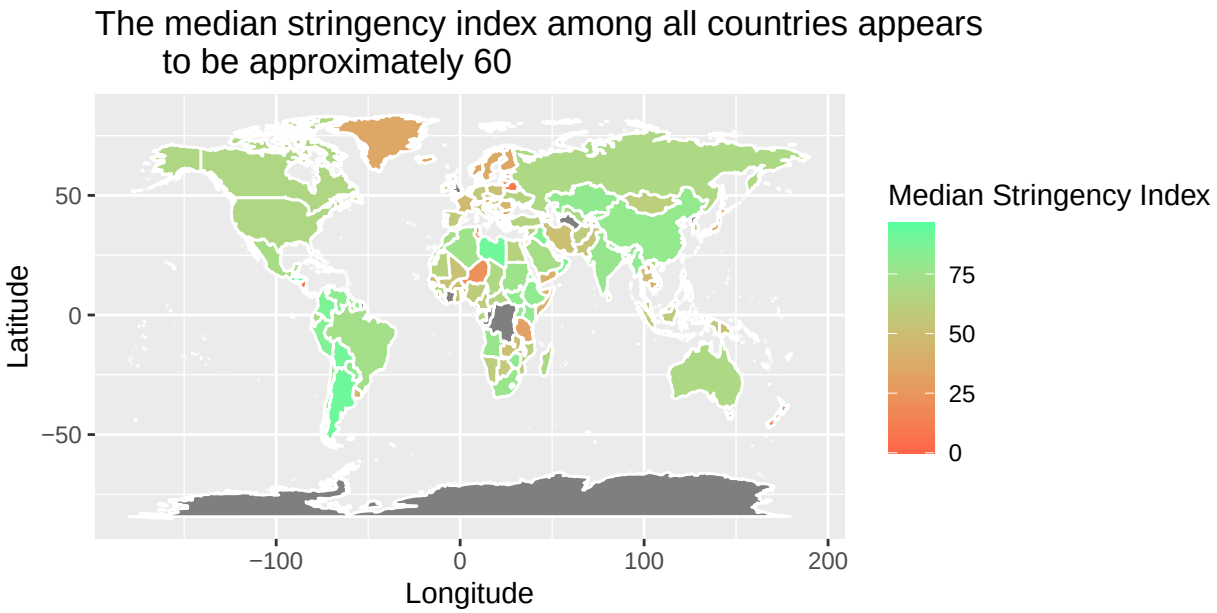
Results

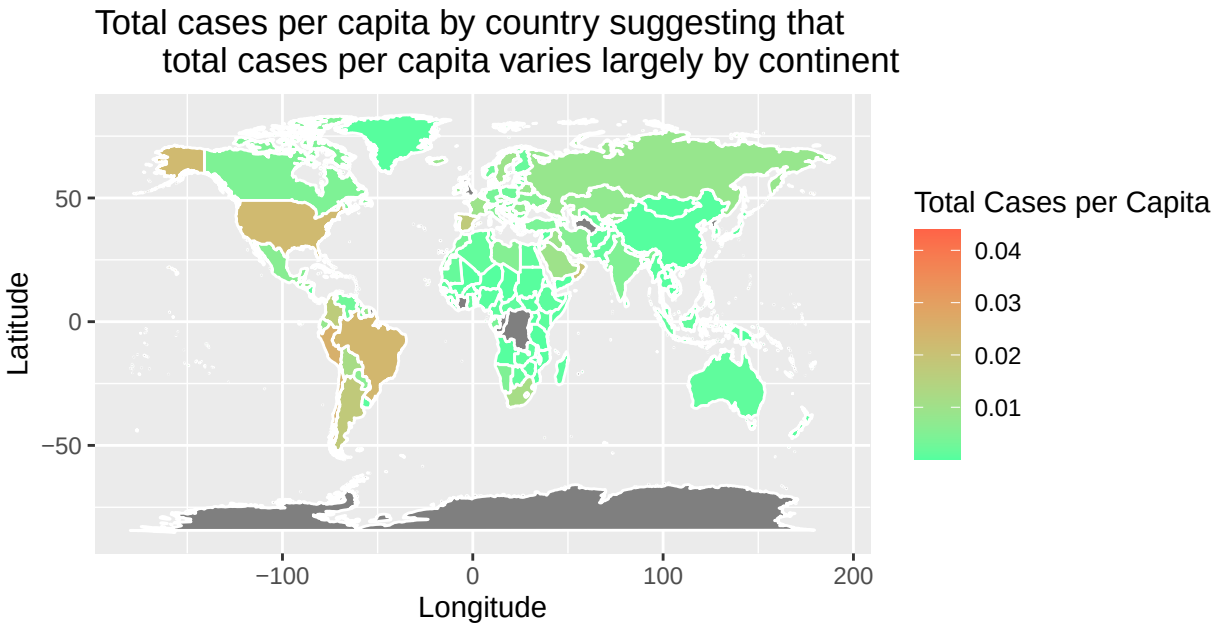
Hypothesis 1: Do Strict SARS-CoV-2 Related Policies Keep Total Cases per Capita Low?

Politicians worldwide are pushing for governments to enact strict policies to mitigate the spread of SARS-CoV-2. 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 early in 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 SARS-CoV-2 cases.

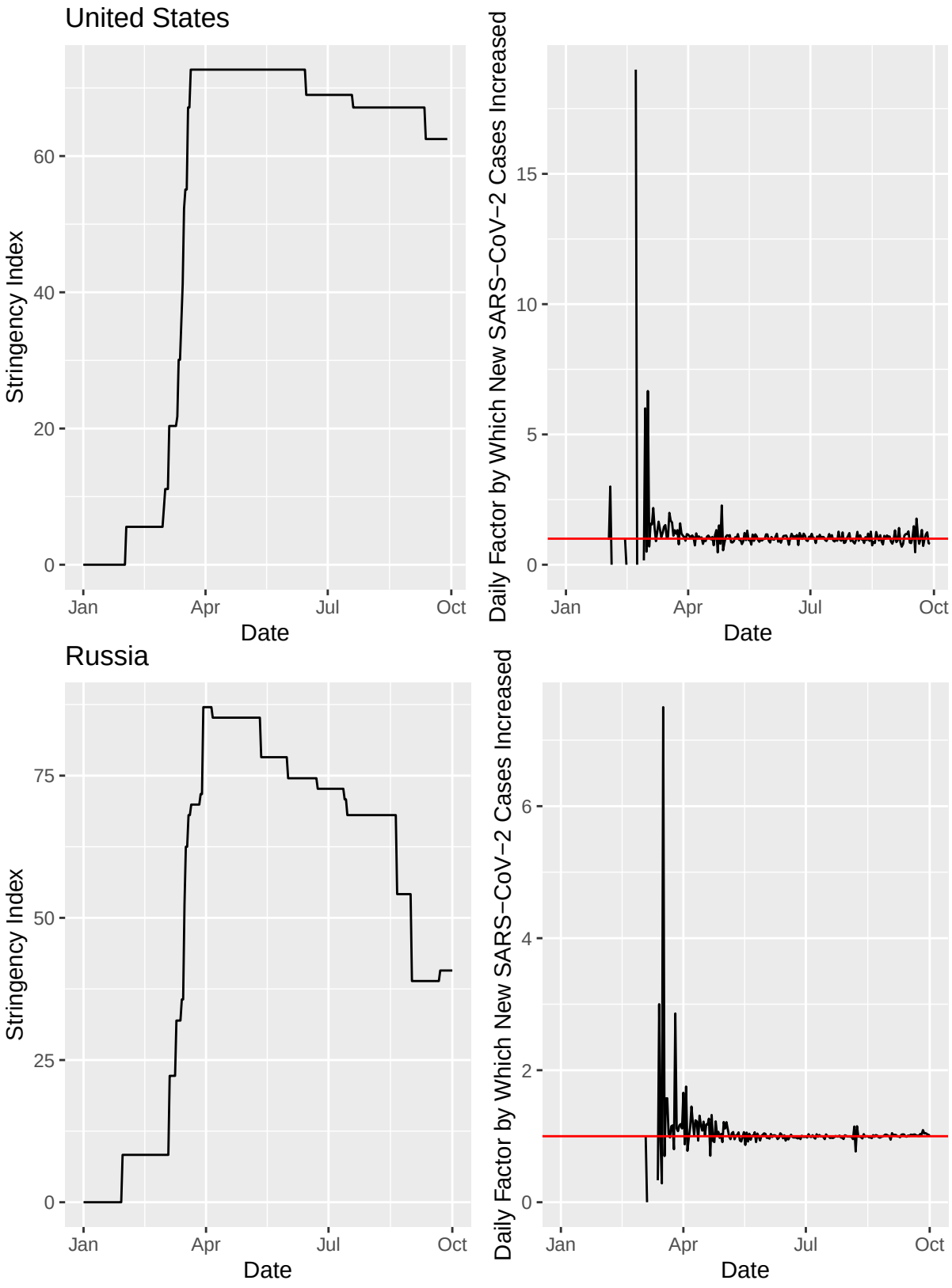
In order to visualize the relationship between a stringency index and the total cases per capita we created two world map plots. The first shows the median stringency index of a country during the entire pandemic while the second shows the total cases per capita on October 5th, 2020.

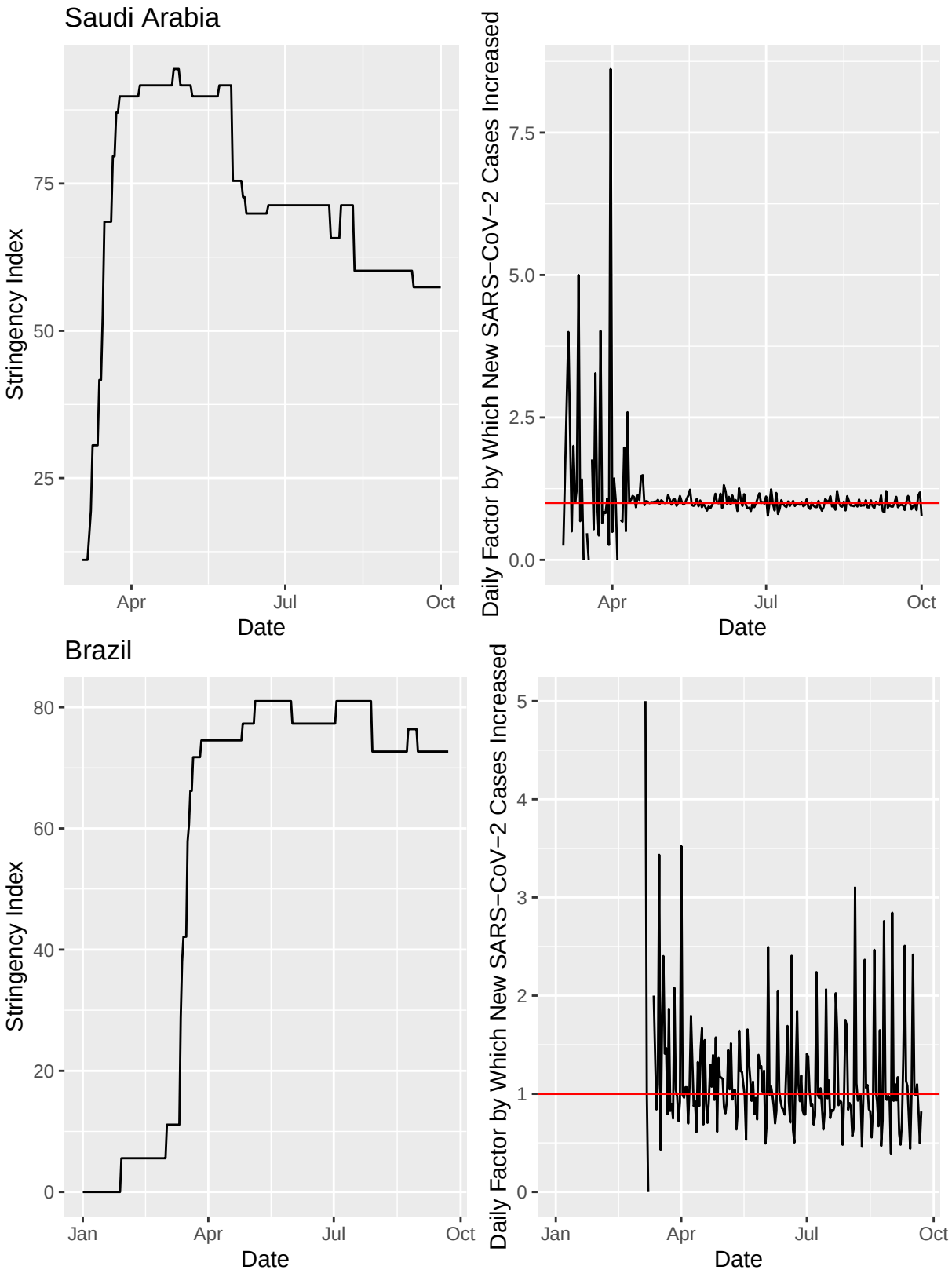


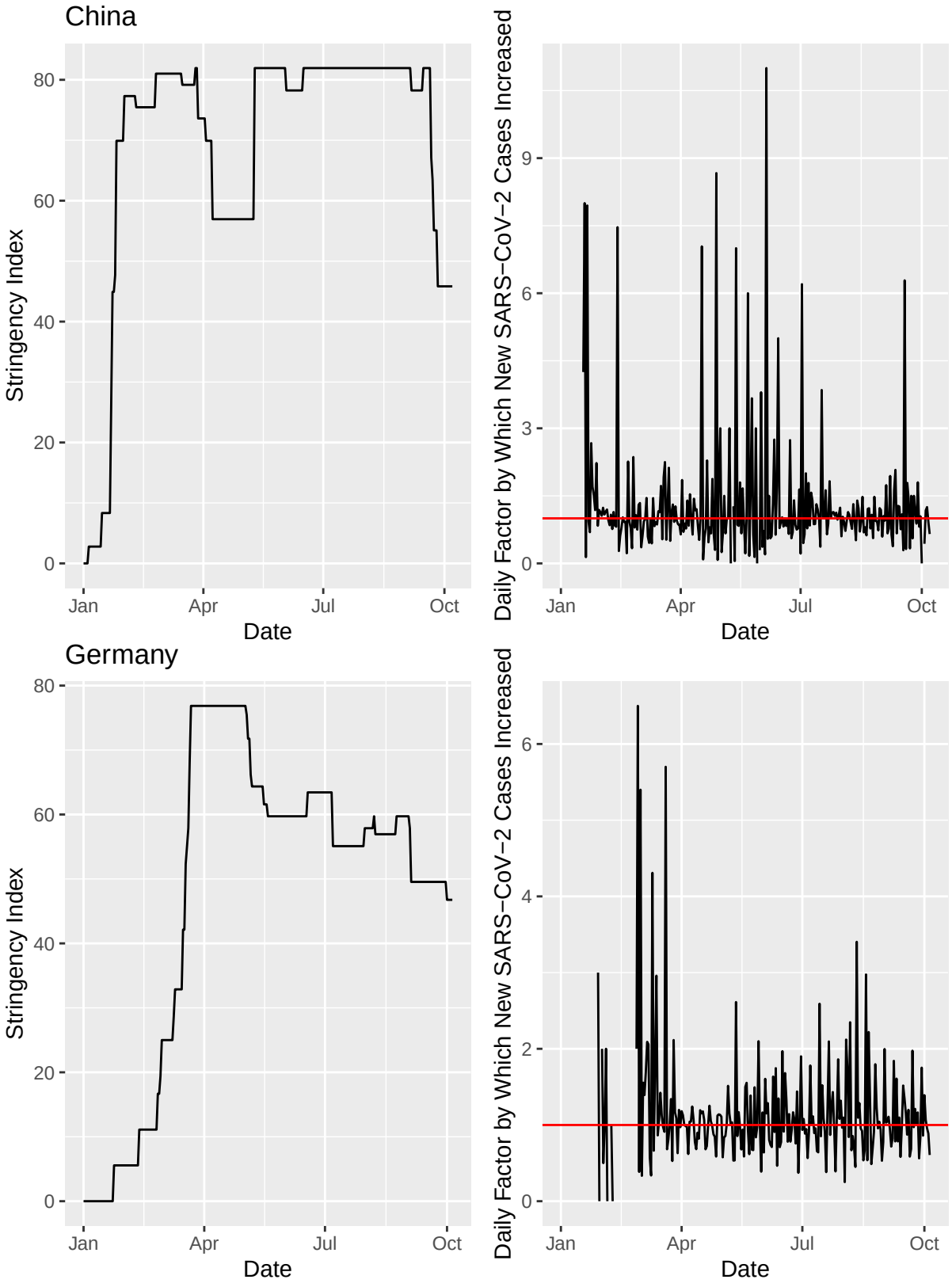


From these world maps alone, it is difficult to determine if a high stringency index is associated with a low total cases per capita. There does not seem to be any pattern of high stringency index countries ending up with low total cases per capita. In fact, it appears from the visual alone that total cases per capita is mostly associated with continent. South America looks to have the highest stringency indices, but also the highest total cases per capita. From these two visualizations, it's hard to asses any relationship between stringency index and total cases per capita.

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
Daily Factor by Which New SARS−CoV−2 Cases Increased
0
5
10
15
Jan
Apr
Jul
Oct
Date
Russia
Stringency Index
0
25
50
75
Jan
Apr
Jul
Oct
Date
Daily Factor by Which New SARS−CoV−2 Cases Increased
0
2
4
6
Jan
Apr
Jul
Oct
Date


Saudi Arabia
Stringency Index
75
50
25
Apr
Jul
Oct
Date
Daily Factor by Which New SARS−CoV−2 Cases Increased
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
Daily Factor by Which New SARS−CoV−2 Cases Increased
5
4
3
2
1
0
Jan
Apr
Jul
Oct
Date



From the graphs of individual countries, it appears that a high stringency index is associated with a lower volatility in the transmission rate of SARS-CoV-2. 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. 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 even after the mid-March spike.

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 SARS-CoV-2 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 SARS-CoV-2" 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 SARS-CoV-2 cases was not being consistently reported each day before then. Thus, there is no information to plot.

Hypothesis 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, we will create a linear model to predict total cases per capita. This model will determine if certain variables are statistically significant in determining total cases per capita.

We will use stringency index, GDP per capita, human development index, and population density as predictors since they were all in our hypothesis. In addition, we added continent as a predictor after observing its potential relevance in the world maps above. We will test our variables at the $\alpha = 0.05$ significance level.

Stringency Index:

H_0: There is no relationship between stringency index and total cases per capita. $\beta_{si} = 0$

H_1: There is a relationship between stringency index and total cases per capita. $\beta_{si} \neq 0$

GDP per Capita:

H_0: There is no relationship between GDP per Capita and total cases per capita. $\beta_{gdp} = 0$

H_1: There is a relationship between GDP per Capita and total cases per capita. $\beta_{gdp} \neq 0$

Human Development Index:

H_0: There is no relationship between Human Development Index and total cases per capita. $\beta_{hdi} = 0$

H_1: There is a relationship between Human Development Index and total cases per capita. $\beta_{hdi} \neq 0$

Population Density:

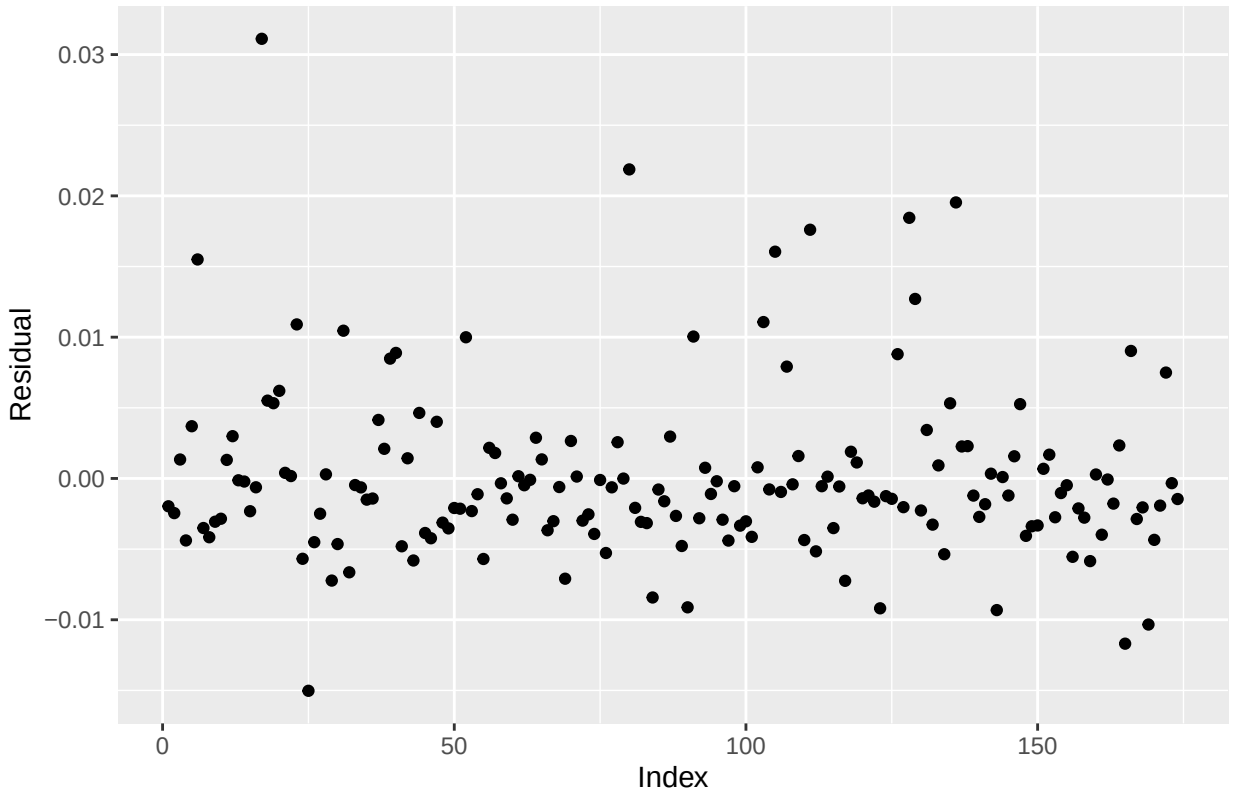
H_0: There is no relationship between Population Density and total cases per capita. $\beta_{pd} = 0$

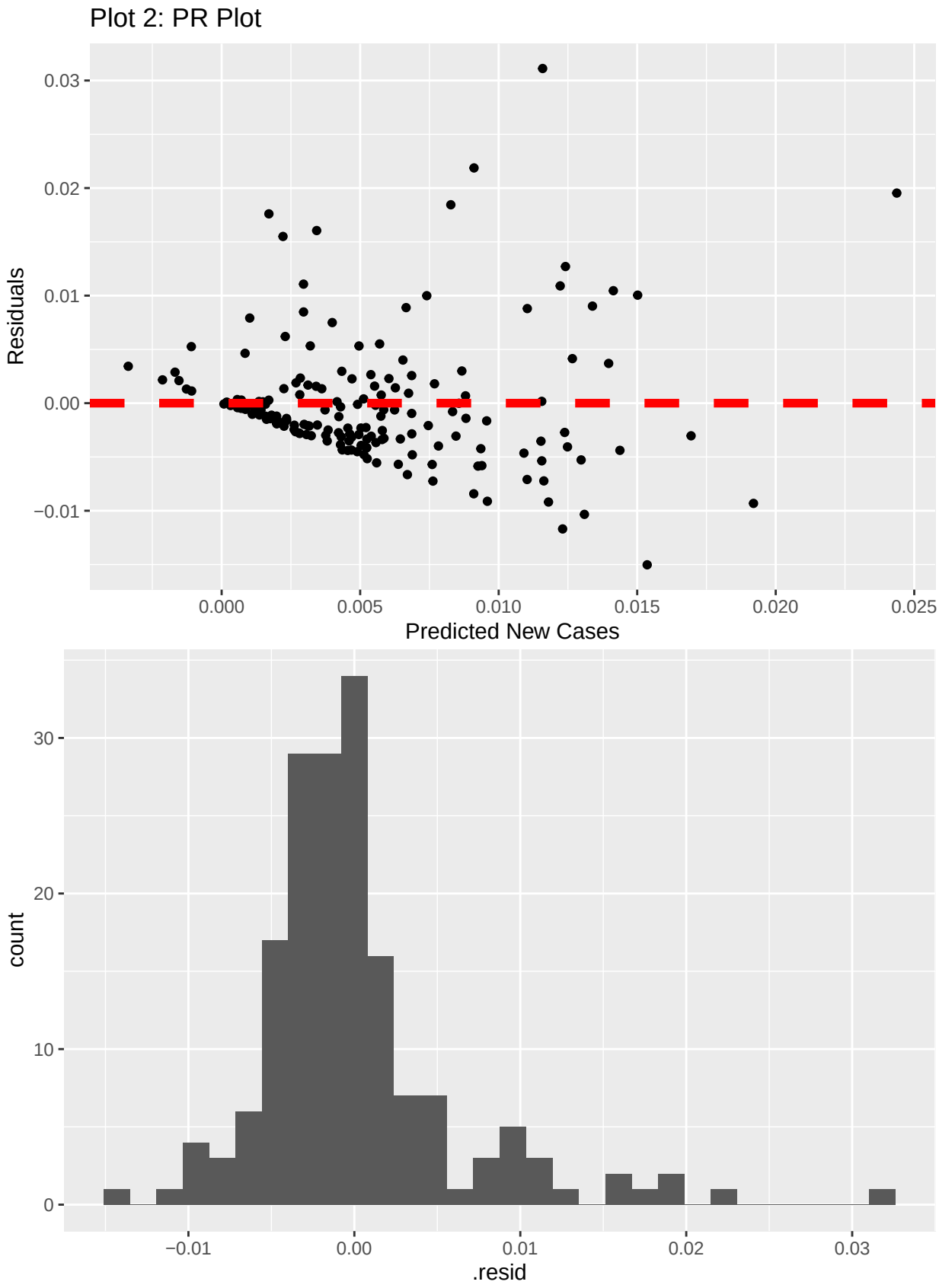
H_1: There is a relationship between Population Densityand total cases per capita. $\beta_{pd} \neq 0$

```
## # 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





Total Cases Per Capita = -3.001862e-03 + 1.719410e-07 * (GDP per Capita) + 4.255473e-05 * (Median

Stringency Index) + 1.503030e-07 * (Population Density) + 2.293517e-03 * (Human Development Index) - 4.502342e-03 * (Continent == Oceania) + 2.082339e-03 * (Continent == North America) + 1.298823e-05 * (Continent == Europe) + 7.959857e-03 * (Continent == South America) + 1.958916e-03 * (Continent == Asia)

After creating the linear model, we see that stringency index and GDP per capita have corresponding p values less than our significance level of 0.05. We reject the null hypotheses. There is sufficient evidence to suggest that there is a relationship between stringency index and total cases per capita as well as GDP per capita and total cases per capita.

As for the coefficients of population density and human development index, both have a p value greater than our significance level of 0.05. Therefore, we fail to reject the null hypothesis. There is not enough evidence to suggest that population density has any relationship to total cases per capita nor that human development index has any relationship to total cases per capita.

The R^2 is 0.3448232 and adjusted R^2 is 0.3088684. Roughly 34.8% of the variability in total cases per capita can be explained by the model which includes the stringency index, GDP per capita, population density, human development index, and continents.

Our model also shows that the continent of South America is statistically significant in determining total cases per capita, also hinting at some association between continent and total cases per capita.

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.

This will require the creation of a categorical variable to describe total cases per capita. This new categorical variable will be based on the following key:

Percentile Categorization 90th - 100th: Very High 70th - 90th: High 30th - 70th: Average 10th - 30th: low 0th - 10th: Very low

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.

$\alpha = 0.05$

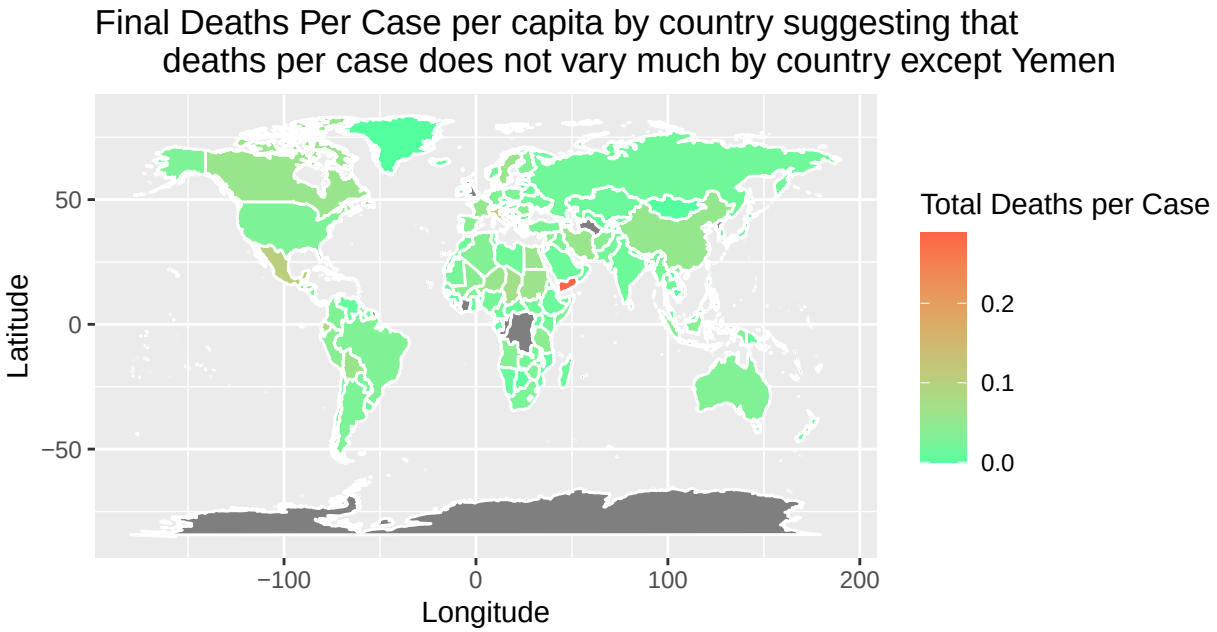
```
##
##  Pearson's Chi-squared test
##
## data:  table(covid_tcpc$continent, covid_tcpc$tcpc_cat)
## X-squared = 84.511, df = 20, p-value = 6.645e-10
```

After running a χ^2 test with a test statisitc of 84.511 and 20 degrees of freedom, we calculated a p value of 6.645e-10, which is less than our $\alpha = 0.05$ significance level. Thus we reject the null hypothesis. There is sufficient evidence to suggest that continents are not independent from total cases per capita.

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

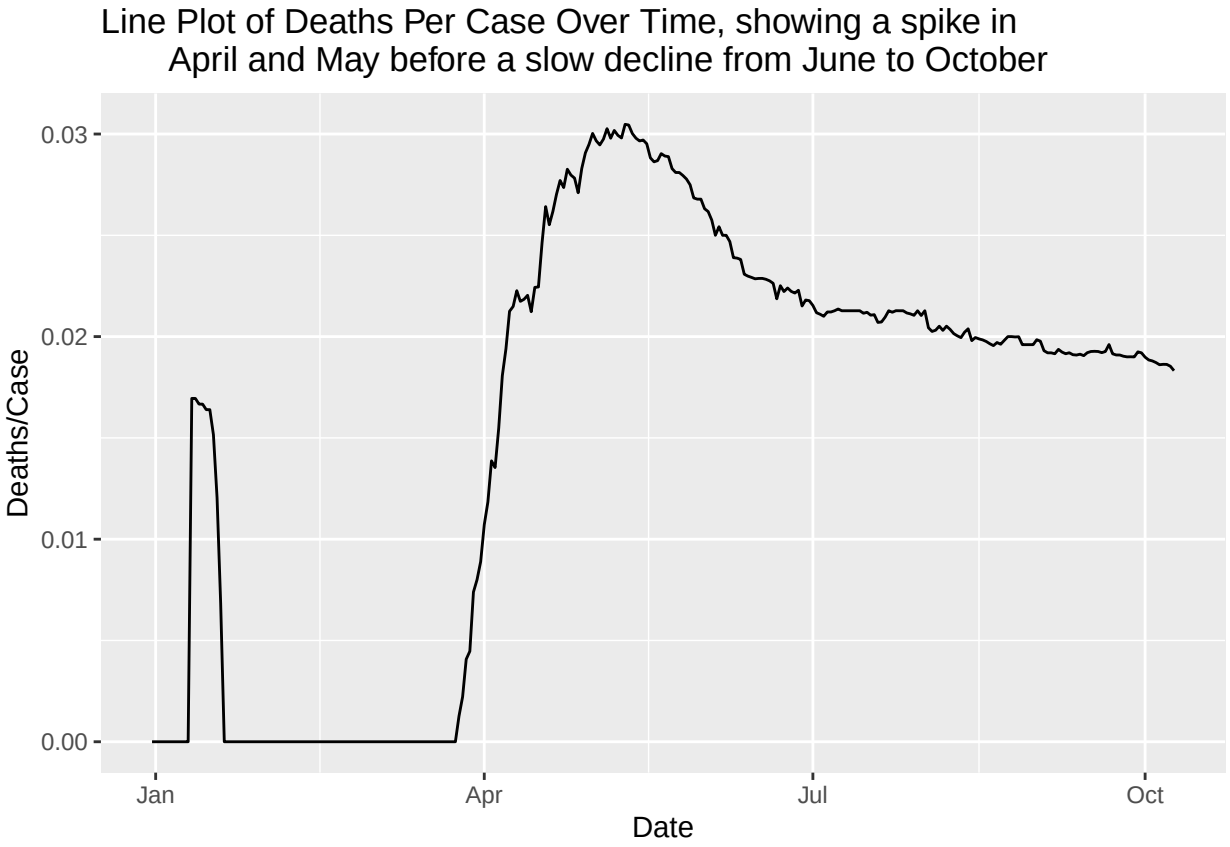
In looking at this hypothesis, we first seek to perform exploratory data analysis.

First, we are creating a map to visualize the final deaths per case by region/country, and see if there are any evident shared characteristics in countries with higher deaths per cases.



In looking at the plot, it is clear that most countries have a deaths per case in the 0 to 0.1 range, with Yemen being a clear outlier in the map.

In addition to exploring the regional deaths per case, we created a line plot to visualize how the median deaths per case changes in relation to time. We chose to visualize the median to prevent outlier countries from skewing the line in any direction.



In trying to determine what predictors have a relationship with deaths per case and what variables factor into

a country's deaths per case, we decided to run a multiple linear regression including variables that encompass different country characteristics including age, pre-existing conditions (diabetes, smoking, cardiovascular death rate), as well as variables to predict how well a country's healthcare system may be including human development index, life expectancy, and GDP per capita (how much they may be able to spend on pandemic).

```
## # A tibble: 10 x 5
##    term                       estimate  std.error statistic p.value
##    <chr>                         <dbl>      <dbl>     <dbl>   <dbl>
##  1 (Intercept)             -0.0593     0.171         -0.347  0.735
##  2 population_density      -0.0000126  0.0000257     -0.491  0.632
##  3 handwashing_facilities   0.000396   0.000445       0.890  0.391
##  4 diabetes_prevalence      0.00312    0.00337        0.927  0.372
##  5 life_expectancy          0.00555    0.00366        1.52   0.155
##  6 cardiovasc_death_rate   -0.000146   0.0000912     -1.60   0.135
##  7 female_smokers          -0.00395    0.00226       -1.75   0.105
##  8 gdp_per_capita           0.00000617 0.00000334     1.85   0.0896
##  9 human_development_index -0.408      0.214         -1.91   0.0810
## 10 aged_70_older           -0.0238     0.0118        -2.01   0.0672
```

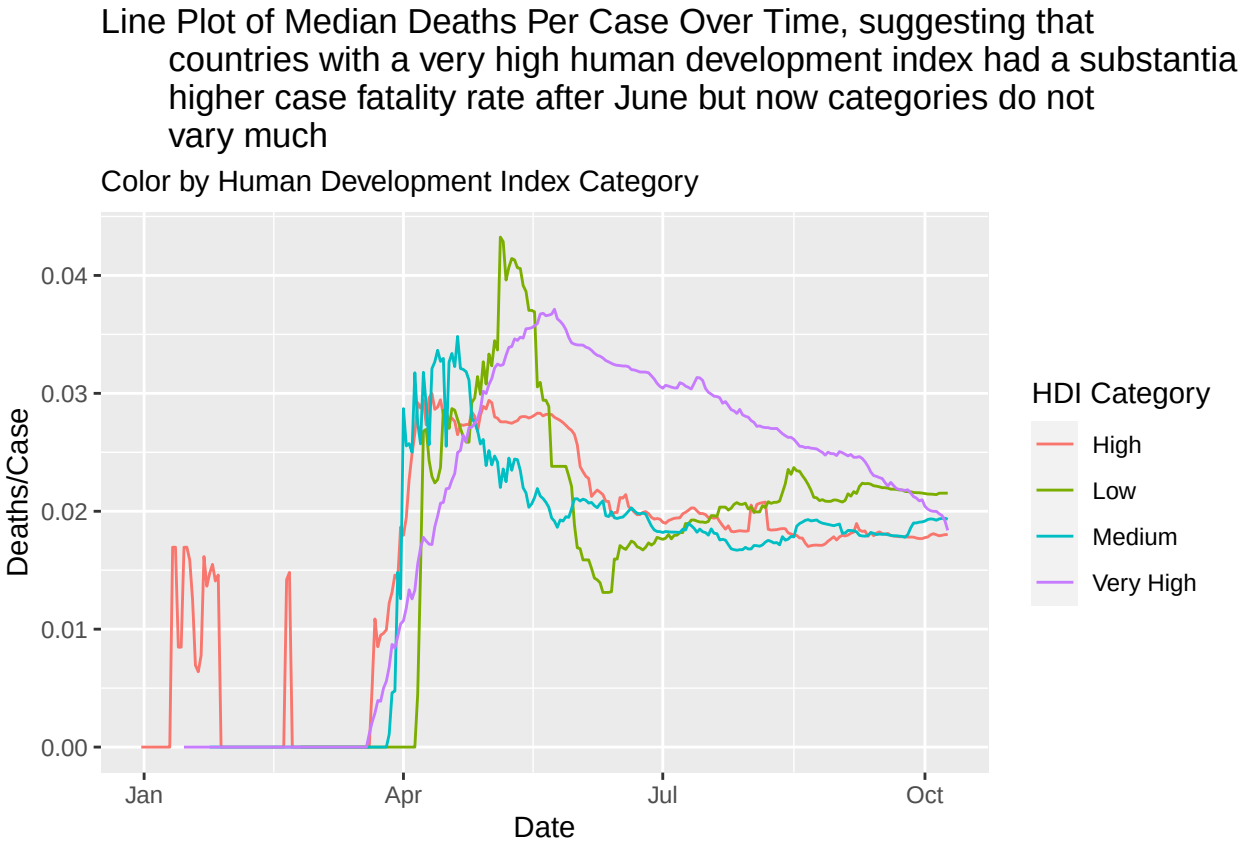
```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.508         0.139
```

Deaths per case = -5.933385e-02 + -1.263897e-05 * (population_density) + 3.961726e-04 * (handwashing_facilities) + 3.120720e-03 * (diabetes_prevalence) + 5.551136e-03 * (life expectancy) + -1.461799e-04 * (cardiovasc_death_rate) -3.953100e-03 * (female_smokers) + 6.167923e-06 * (gdp_per_capita) -4.079514e-01 * (human_development_index) -2.375371e-02 * aged_70_older

The R^2 is 0.5080334 and adjusted R^2 is 0.1390585 Roughly 50.8% of the variability in deaths per case of a country can be explained by population density, handwashing facilities, diabetes prevalence, life expectancy, cardiovascular death rate, female smokers, GDP per capita, human development index and the percentage of the population that is aged 70 or older.

In creating our regression, we found that the percentage of a country's population that is aged above 70, as well as human development index, to be the variables with the highest coefficients and most statistically significant P-values. Thus, we choose to explore them further in a line visualization, to see how they change over time.

First we will look at human development index. We will visualize deaths per case over time as we did before, however, we will have different lines denoting different human development index categories. We have categorized them into bins of "Very high", "High", "Medium", "Low", and "Very Low", based on the various thresholds used by the UN [7].



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.

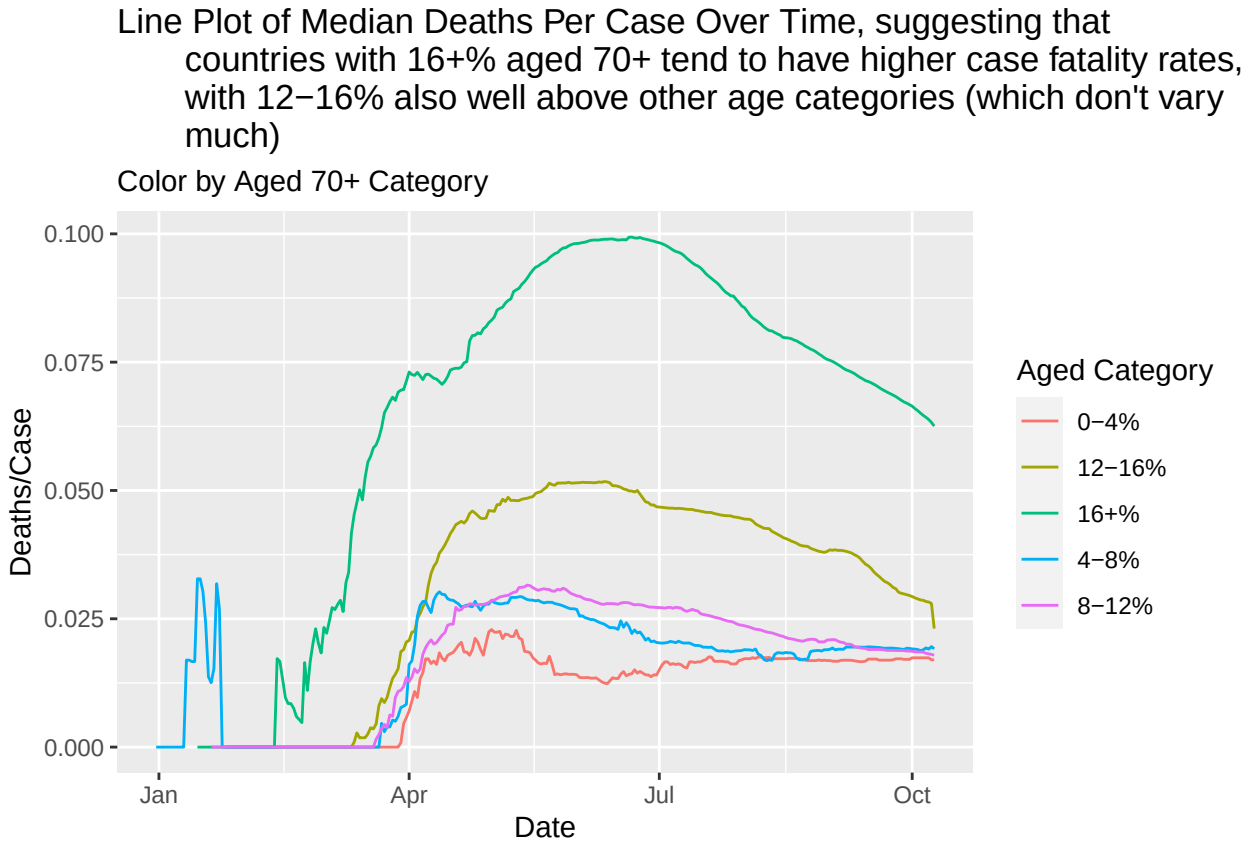
H_0: Human development index category and deaths per case are NOT independent.

```
##
##  Pearson's Chi-squared test
##
## data:  table(covid_hdic$hdi_columns, covid_hdic$dpc_cat)
## X-squared = 110.58, df = 12, p-value < 2.2e-16
```

```
## [1] 4.588415e-18
```

Our test statistic was 110.58, which has a chi-square distribution with 12 degree of freedom under H_0. This correlates with a P-value = 4.588415e-18 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 human development index category and deaths per case are not independent.

Next we will look at the relationship between percentage of population aged 70+ and deaths per case. We will visualize deaths per case over time as we did before, however, we will have different lines denoting different aged 70+ percentage categories. We have categorized them into bins of 4% intervals (e.g 0-4,4-8,8-12,12-16,16+), based on data from a comparative aging population study [11].



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.

```
##
##  Pearson's Chi-squared test
##
## data:  table(covid_dpc$aged_columns, covid_dpc$dpc_cat)
## X-squared = 264.18, df = 16, p-value < 2.2e-16
```

```
## [1] 6.328478e-47
```

Our test statistic was 264.18, which has a chi-square distribution with 16 degrees of freedom under H_0. This correlates with a P-value = 6.328478e-47 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 the total deaths per case in the final week are not independent.

Discussion

Throughout the project, we tried to answer the following research questions:

1) Is government policy effective in reducing the spread of SARS-CoV-2?

Here, we sought to determine if there is a relationship between stringency index and cases per capita, as well as the daily growth rate of cases within a country. Through the world map, we could not deduce a

relationship between stringency index and cases per capita. We do notice that Asia tends to have low cases per capita with high median stringency index, Africa tends to have low cases per capita with varying median stringency index, South America has high cases per capita and high median stringency index, and Europe has a lower median stringency index with higher cases per capita. Thus, Europe and Asia seem to support the conclusion that stringency index reduces cases per capita, however, there is uncertainty about the extent to the relationship given the other continents such as South America.

Thus, to further explore the impact of stringency index on the rate of transmission of SARS-CoV-2, we plotted the stringency index over time for various countries alongside the growth rate of cases over time. Given that cases per capita is increasing for virtually every country and is a reflection of the total cases as opposed to new cases, we found that growth rate is a more effective measure of transmission. Thus, we can see the immediate effect of governmental policy such as increasing stringency index when the growth rate changes afterward.

In analyzing the line plots for countries stringency indexes and rate of transmission, we found much more conclusive results suggesting that a higher stringency index is associated with lower case growth rate. In all countries, we observed that the stringency index spiked in mid-March in a response to a spike in cases. Before countries locked down for SARS-CoV-2, they all experienced surges in growth rate. However, these countries all then rapidly increased their stringency indexes and found subsequent drops in case growth rates. These effects were present in all countries, especially in the United States, Russia, and Saudi Arabia.

Ultimately, the comparative maps were limited in their storytelling ability due to the fact that we could not animate the visualization over time (due to PDF), but rather had to trust the median stringency index. Thus, it was harder to pinpoint the individual effect of a country's stringency index on transmission. However, the observed regional difference in cases per capita provided a great foundation for the rest of our research, leading us to further explore the relationship between continent and cases per capita in research question 2.

Meanwhile, the country plots illustrated the relationship between stringency index and daily growth rate of SARS-CoV-2, helping us determine its effectiveness in reducing its transmission. The visualizations corroborated a clear story among various countries that increasing a country's stringency index was associated with a reduction in case growth rate.

Given the implications of our results, we suggest further investigation into the matter of the effectiveness of increasing stringency index to reduce case growth rate through the use of proper hypothesis testing (via CLT or Bayesian).

2) What country characteristics are most significant in determining a country's cases per capita?

We were extremely interested to see which variables played a large role in determining cases per capita. We initially hypothesized that stringency index, GDP per capita, population density, and human development index would all be statistically significant in determining total cases per capita. We trained a linear model to test these hypotheses and concluded that GDP per capita and stringency index were both statistically significant in determining cases per capita, while population density and human development index were not.

We expected the relevance of GDP per capita and stringency index. Intuitively, stricter policy should lead to a slower spread of SARS-CoV-2 and countries with a high GDP per capita were likely able to afford programs to help slow the spread. Population density was surprising to us since we figured that people living in closer proximity would be more likely to transmit the virus. We also expected human development index to be relevant since countries with a low HDI would likely have less access to PPE and safe places to quarantine.

We also included continent as a predictor in our model even though it wasn't hypothesized. We felt compelled to after seeing an association between cases per capita and continent on the world maps mentioned in hypothesis 1. It turns out that South America in particular was statistically significant in determining cases per capita. Essentially, our model predicted that a country in South America would have a higher cases per capita than other continents regardless of the other predictors. This was shocking to us since South America's stringency indices were so high compared to the rest of the continents.

This led us to perform a chi squared test on continents versus cases per capita. We were able to conclude continents and cases per capita are NOT independent. This find was fascinating but led us to consider some

limitations. We recognized that the statistical significance of continent was likely a result of other variables that were continent specific but not captured in the dataset. This would explain why South America in particular was statistically significant in determing cases per capita while the rest of the continents were not.

All of this led us to realize that it's incredibly hard to accurately predict the spread of SARS-CoV-2. There are too many factors that impact the rate of spread for are models to capture in a reasonable manner. This was reflected in our relatively low adjusted R^2 of 0.309. If we wanted to further analyze which variables have a large impact on cases per capita, we might need to get more outside data and spend a lot more time thinking through which predictors to include. Still, we were satisfied to conclude that GDP per capita and stringency index were significant in determining some of the variability of cases per capita.

3) What country characteristics are most significant in determining a country's case fatality rate (deaths per case)?

In analyzing the relationship between country characteristics and case fatality rate, first we performed exploratory data analysis to determine which countries had high case fatality rates and how case fatality rate changed over time.

In creating a map with final case fatality rate, there were no clear trends in countries or country characteristics observed in the visualization. This is likely due to two limitations. We were not able to animate the visualization over time (due to PDF), such that we are only getting a limited piece of the data. Additionally, Yemen is an outlier with such a high case fatality rate, that other countries seem to all be similar in case fatality rate, as the disparity between them and Yemen is greater than the variance in the data among other countries. While the map was not incredibly beneficial, observing Yemen having a very high case fatality rate helped us consider the type of country that may have higher case fatality rates.

The visualization of median deaths per case over time starts relatively high due to lack of reporting (except for Asia where there was a spike), however, this decreases to approximately 0 from January to mid-March. In mid-March, we observe a spike in the deaths per cases through April, before a steady decline from May to October. This helps us understand the dynamic nature of case fatality rate, which spiked at a similar time to other case and death related variables, but is now experiencing a decline.

In trying to determine the most important predictors for deaths per case, we identified many different classes of variables for country characteristics that could affect the response, including pre-existing conditions, age, and healthcare/wealth indicators. Given the great number of potential variables, we could not perform an exhaustive data analysis and perform significance tests on all of them. Thus, we performed a multiple linear regression and exampled the statistical significance of the predictors. We then continued to remove variables that were shown as more insignificant, and kept them out of the model if the adjusted R^2 improved upon their removal (akin to backward stepwise regression but manual). Thus, we came to a regression that predicted approximately 50% of the variance in deaths per cases. The regression output suggested that human development index and the percentage of a country's population that is aged 70 or above are the most statistically significant predictors, however, neither variable had a P-value that demonstrated statistical significance at the $\alpha = 0.05$ level. Thus, we decided to further explore these variables through visualizations and chi square tests to determine if they individually showed a relationship with deaths per case (when not diluted by the model).

Based on a United Nations categorization, we categorized the human development index variable into categories ranging from very low to very high, before visualizing each category's median deaths per case over time. We found that countries with higher human development index had a substantially higher case fatality rate after June (with low peaking in June), but the categories varied less as time progressed. Thus, the relationship changed with time.

In seeing a relationship between human development index and deaths per case in the line plot, we decided to perform a chi square test to formally determine if there was a statistically significant relationship. The chi square test led to a rejection of the null hypothesis, suggesting that there is a statistically significant association. This was especially interesting given that the test was performed based on observations from the last week of data in the dataset, for which there was less variance in the dataset.

Additionally, we categorized the percentage of those aged 70 or above by percentage and visualized their categories median's deaths per case over time. This line plot elucidated an incredibly clear relationship, in which deaths per cases were much higher for the duration of the pandemic for countries with 16 or more percent of their population aged 70+. This was backed up by countries with 12-16% of their population aged 70+ being less than the other, but still experiencing drastically higher case fatality rates than countries in the categories under 12%. We observed less variance in the lower categories, potentially suggesting that the relationship between aged 70+ and deaths per cases increases as the percentage of those aged 70+ increases (potentially non-linear relationship). Ultimately, the visualization suggested that countries with a higher percentage of those aged 70+ tend to experience a higher case fatality rate throughout the pandemic.

To determine whether the seemingly clear relationship between aged 70 and older and deaths per case in the line plot, we decided to perform a chi square test to formally determine if there was a statistically significant relationship. The chi square test led to a rejection of the null hypothesis, suggesting that there is a statistically significant association.

Ultimately, in finding these associations, we can not make affirmative statements on causation (due to the possibility of confounding), however, our statistical analysis has convincing evidence to suggest associations between percentage of population aged 70 and older and human development index on case fatality rate, where higher percentages of population aged 70 and older, as well as human development index, tend to have higher case fatality rates.

In observing these associations, we encourage further inquiry into the relationship between these variables and deaths per case. Ultimately, the multiple linear regression only accounted for 50% of the variance in case fatality rate, such that even more significant factors may not be in the dataset. Thus, we would like to not only encourage investigation of our variables, but potentially explore datasets with more potential predictors to help better explain and predict a country's case fatality rate.

Limitations

When it comes to our dataset and statistical analysis, there are many limitations to consider. In the dataset, there was missingness in the data, as many countries did not report testing, cases, positive rate, or deaths at different points especially early on, as well as data on country data such as percentage of male smokers and female smokers, cardiovascular death rate, extreme poverty rate, handwashing facilities, hospital beds per thousand, and human development index. This led to countries potentially being under or over-represented in the data based on whether or not their data was public and added to this dataset. These are key factors in the governmental policy, viral transmission, and death rate of countries that we seek to explore, such that this missingness reduces the reliability of our statistical analysis.

Additionally, SARS-CoV-2 data reporting and testing varies significantly by country, such that countries that lack comprehensive testing programs skew our analysis by reporting less cases than there are. Given that we are seeking to determine factors that affect cases per capita and deaths per case, knowledge of the true cases within each country is crucial to coming to this understanding. The disparity in the level of testing as well as reporting for cases and deaths among different countries adds bias to the model that we can not truly pinpoint and address without knowing the true numbers for these countries.

Given that the project started in October, our dataset only includes data up to October 5th. Over one month has passed since then, and we have seen a dramatic uptick in the new cases and total cases around the world, with the daily new cases increasing from ~270,000 when we downloaded the dataset to over 500,000 daily new cases globally now [5]. Many countries have experienced rapid growth/new spikes in cases and deaths, providing new and helpful information about the dissemination and deaths of the SARS-CoV-2. In not having this new data in our dataset, our statistical analysis will fail to interpret this new and potentially vital or enlightening information that could shed more light on the effect of different variables over time.

The previous limitation is ultimately rooted in the fact that SARS-CoV-2 is dynamic and novel, meaning that we are still learning about the virus and it is hard to know what variables have the most impact on spread and death rate. For example, experimentation on the virus published in the Journal of Translational Medicine found evidence for different geographical strains of SARS-CoV-2, as the virus is evolving and mutating. This

is an important variable that is not in our dataset, but could be a confounding variable if a strain is more infectious or more fatal, as we seek to determine what affects deaths per case and total cases per capita being. Our dataset likely does not include all relevant variables, which can be seen in our regressions as we are only able to account for 49% of the variance explained in deaths per case and 30.9% of the variance explained in total cases per capita. With so many variables changing over time and many relevant variables likely not in the dataset, it is difficult to determine the greatest factors in the spread and fatality rate of SARS-CoV-2.

Finally, our data is of a time series format, which we have less statistical experience dealing with. Time series data requires a different style of analysis as each observation is not independent of each other (e.g. cases yesterday is not independent of cases today). This was yet another problem in our linear regression model. Our linear regression model did not pass all of the diagnostic plots and thus must be interpreted as biased and virtually unusable. Our linear regression model is also limited in that it assumes a linear relationship between the variables, when in fact there may not be. For example, there could be a polynomial relationship between a variable and the response, the larger it gets, the more it affects the response variable. Additionally, the regression model can not attribute causality, but only an association. While a predictor and response variable may share a linear relationship, this could potentially be due to a confounding variable, such that the predictor does not actually cause the response. For example, people being aged 70 or older having a higher negative coefficient and negative relationship with deaths per case in the regression could be because of another variable. More people being aged 70 could lead to more retired people and less people in the workforce and medical system, causing deaths per case to be higher. This is thus a limit of using a linear regression.

With our current regression, another limitation is variables can be collinear, leading to coefficients that do not accurately reflect the relationship between the variable and a model. When running a linear regression, two multicollinear variables may have correlation coefficients that differ but rather additively have their true effect, as linear regressions struggle to differentiate the effects of two multicollinear variables. This is why machine learning specialists often perform feature selection (via ridge and lasso regression) and principal components analysis, to remove predictors and a model with more consistency in its correlation coefficients.

###Improvements

Given our many limitations, we would likely conduct our statistical analysis very differently were we to restart now. First, we would likely use an updated version of the dataset that includes more data after October, as it would provide new and relevant data with less missingness for us to evaluate and help analyze our research questions. Additionally, we would likely search for other datasets with other variables and country characteristics and join it with our own, to have more potential factors that account for variance in case fatality rate and cases per capita. These potential variables could include more direct quality of healthcare variables including ventilator and mask stock, whether there is universal healthcare, as well as more potential health indicators such as obesity. Ultimately, age and human development index seemed to be the most significant factors in our case fatality rate. Given that countries with a higher age tend to have higher life expectancy and likely healthcare, we believe this could have been a confounding variable for human development index. Given that higher human development index initially suggested a higher case fatality rate, however, human development index indicates a country's wealth, healthcare system, etc, it would originally seem as if these countries are best equipped to handle the pandemic. Thus, we could further investigate these variables by adding new variables, seeing what is most significant, and controlling for different variables.

Additionally, we could further explain research question one by looking at the data that led to the composite score for stringency index, and try to identify if any of the nine indicators have a particularly strong association with a lower case growth rate. This would allow us to make less broad statements about making government policy "stricter", and rather show which specific governmental policies (e.g. mask mandates, banning international travel, banning domestic travel) have stronger associations with reducing case growth rate. Perhaps certain policies may have a relationship, while others may have no association. If we were to restart now, we would certainly explore this possibility.

In adding these variables, our multiple linear regressions may end up diluted by the sheer number of potential variables. Thus, we would likely consider machine learning techniques including a principal components analysis to project similar (collinear) variables onto the same feature. Additionally, we can try running

regressions with interactions or potentially squared and cubed terms, and then compare these regressions with k-fold cross validation to compare the performance of linear and polynomial regressions. We can then choose the best regression model type for data, and use a lasso regression to perform feature selection and choose the variables it finds most significant. A lasso regression penalizes high coefficients and thus significantly decreases the variance in a model, preventing the regression from overfitting the abundance of data it would have.

Given that the data is time series, we would likely look into more techniques for analyzing time series data. This would lead to our regressions and techniques being more reliable and trustworthy, and allow us to further explore the dataset beyond our current abilities. It would also be worth exploring the field of Bayesian statistics and Bayesian paradigms to be able to make more affirmative statements as to the statistical significance of variables and likelihood of data as extreme as ours occurring, given a true null hypothesis. Thus, if we were to restart our project now, we would likely find more country data and explore new statistical analysis techniques to better ascertain the most statistically significant 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] https:
//www.worldometers.info/coronavirus/ [6] https://pubmed.ncbi.nlm.nih.gov/32321524/ [7]
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7507379/ [8] https://ourworldindata.org/gra
pher/covid-stringency-index [9] https://ourworldindata.org/coronavirus-source-data [10] https:
//www.ecdc.europa.eu/en/covid-19/data-collection [11] https://pubmed.ncbi.nlm.nih.gov/10812799/ [12]
https://www.economist.com/graphic-detail/2020/11/16/why-rich-countries-are-so-vulnerable-to-covid-19