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

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.

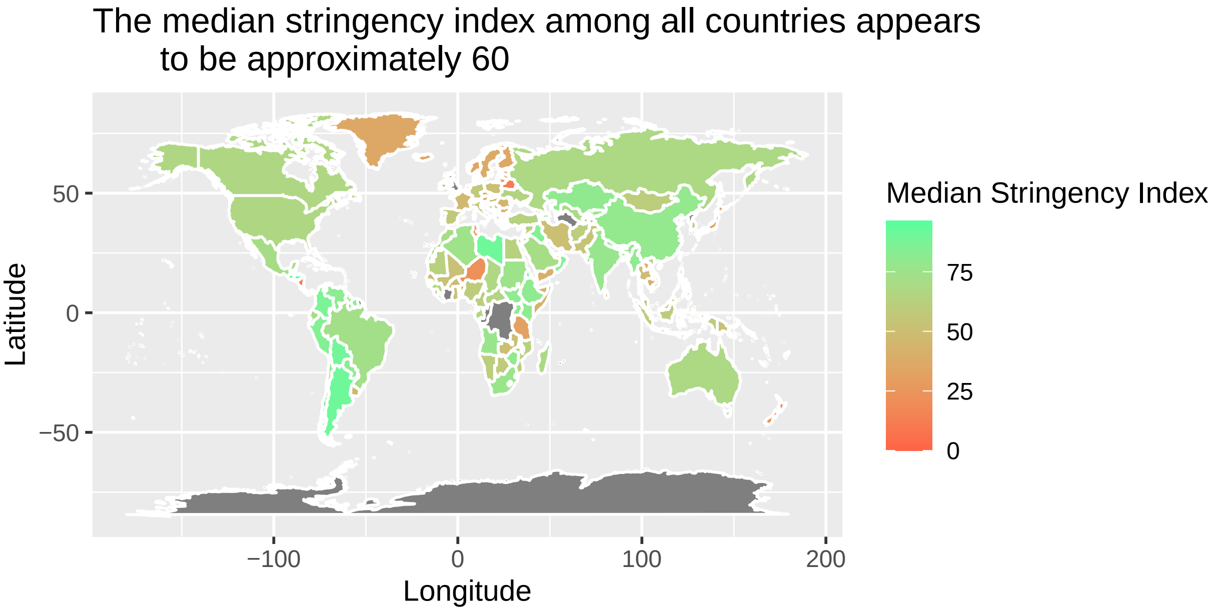
Results

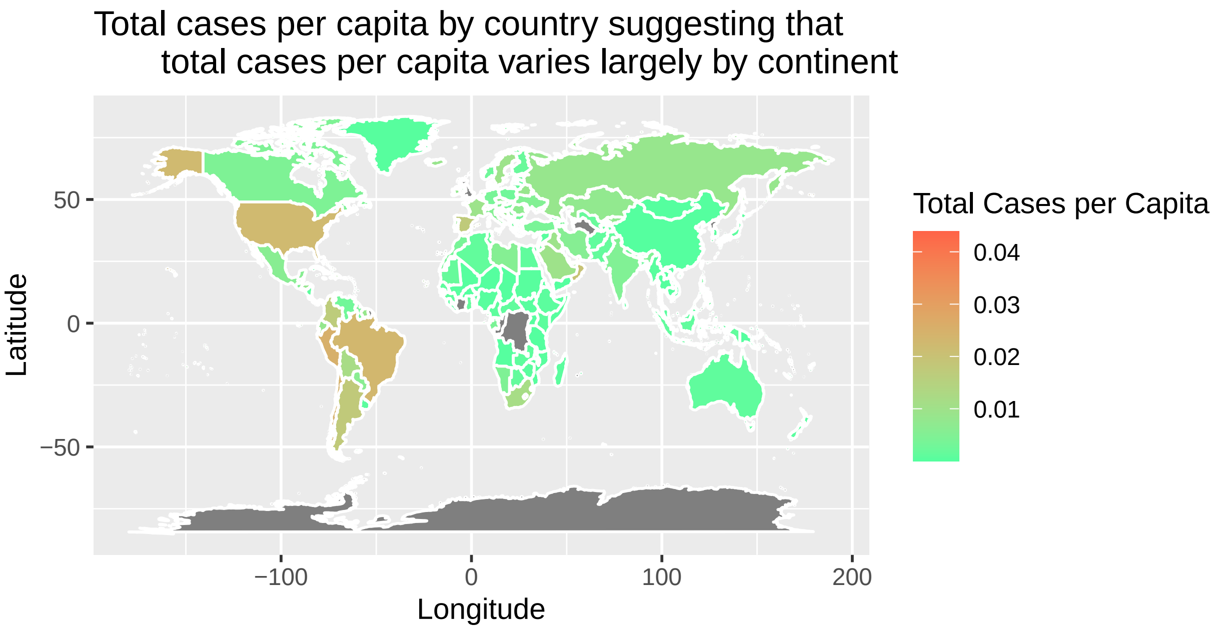
Question 1: Do Strict COVID-19 Related Policies Keep Total Cases per Capita Low?

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 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 COVID-19 cases.

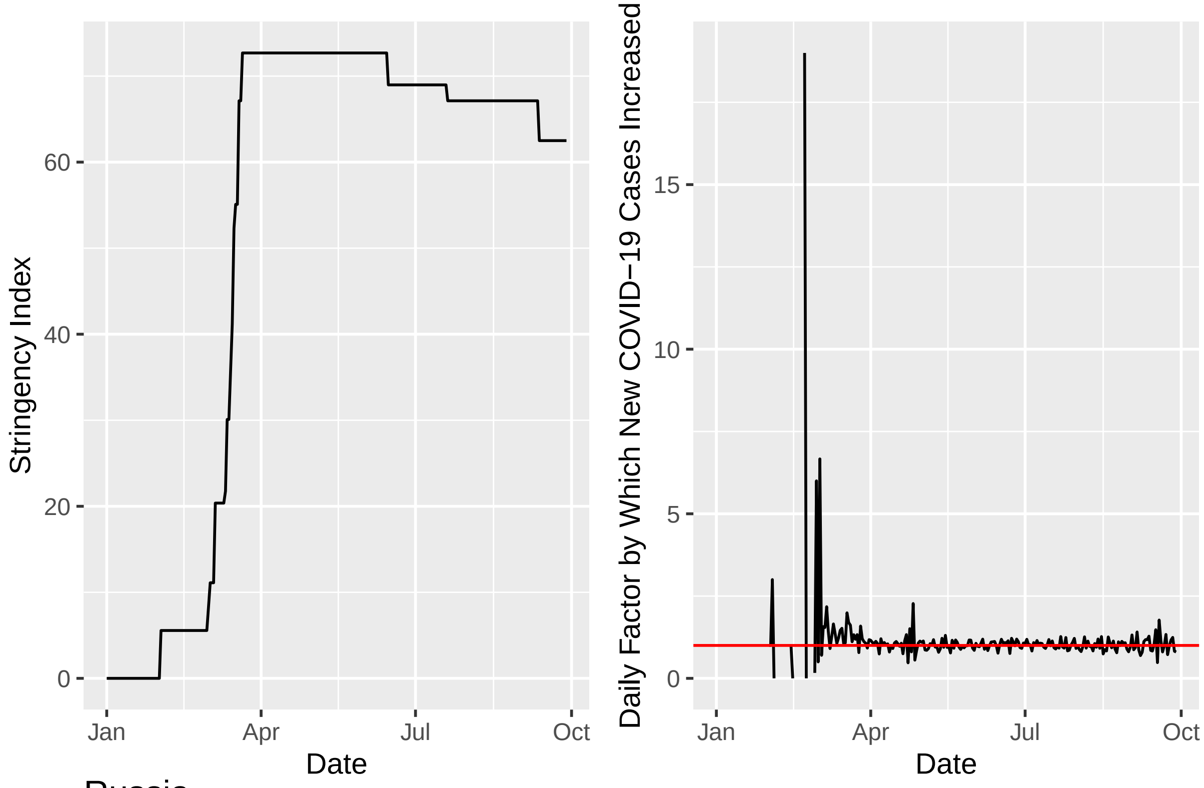
In order to visualize the relationship between a stringency index and the total cases per capita we created two world map plots. The first shows the median stringency index of a country during the entire pandemic while the second shows the total cases per capita on October 5th, 2020.


The median stringency index among all countries appears to be approximately 60
Latitude
50
0
−50
−100
0
100
200
Longitude
Median Stringency Index
75
50
25
0

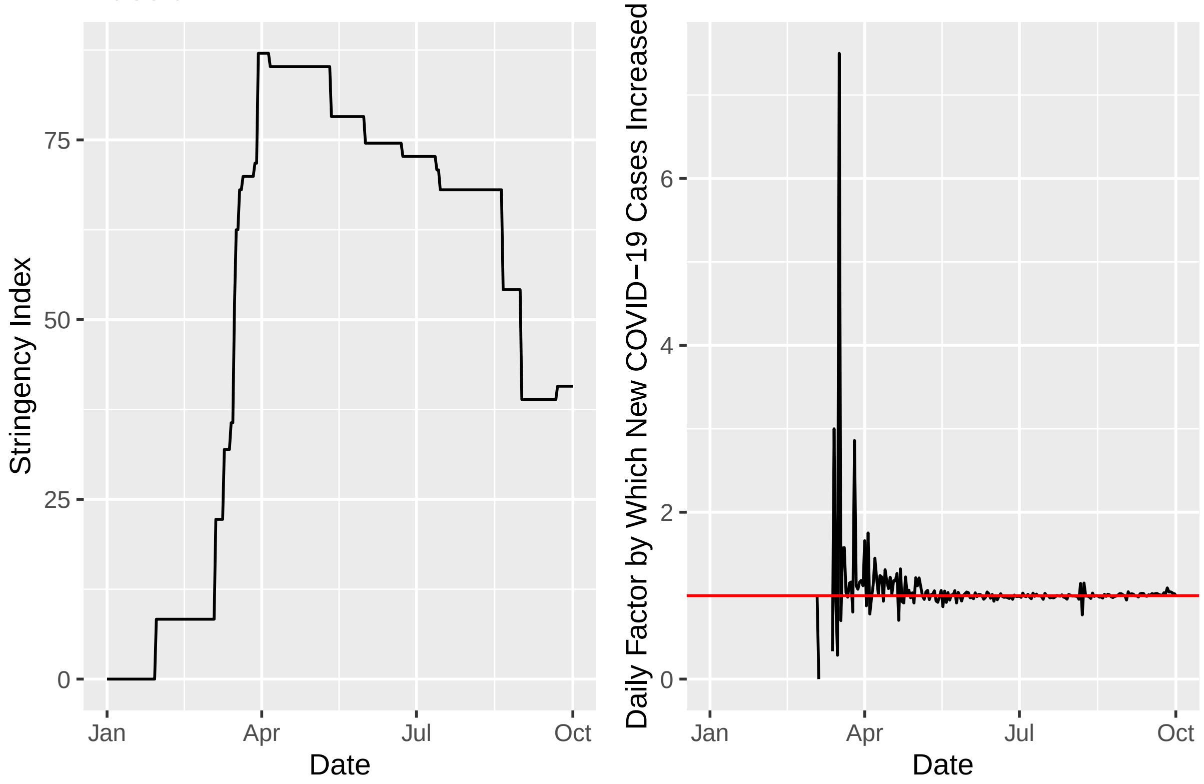

Total cases per capita by country suggesting that
total cases per capita varies largely by continent
Latitude
50
0
−50
−100
0
100
200
Longitude
Total Cases per Capita
0.04
0.03
0.02
0.01

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.

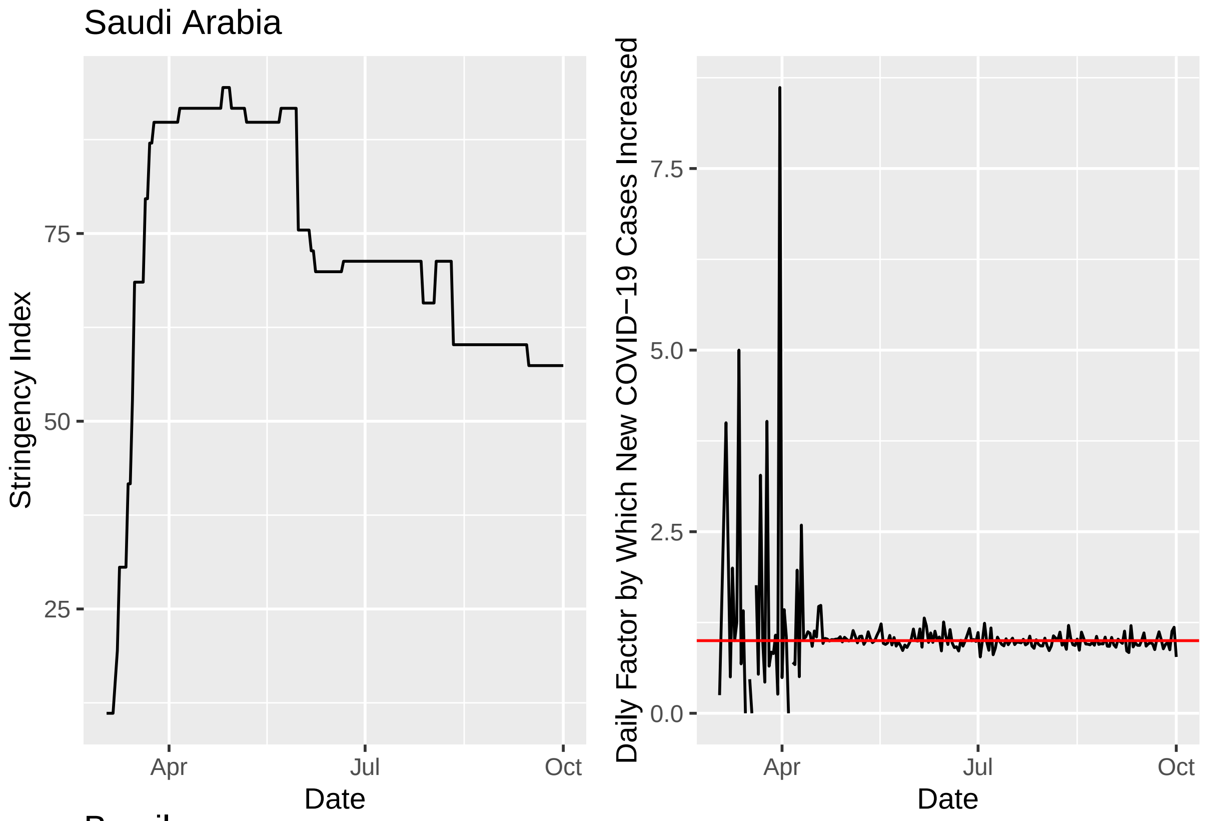
It also is useful to see how stringency index impacts the growth of cases over time. To visualize this, we will plot several countries that have both relatively high and low total cases per capita. The growth of cases will be represented by the factor for which the new cases changed from the previous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



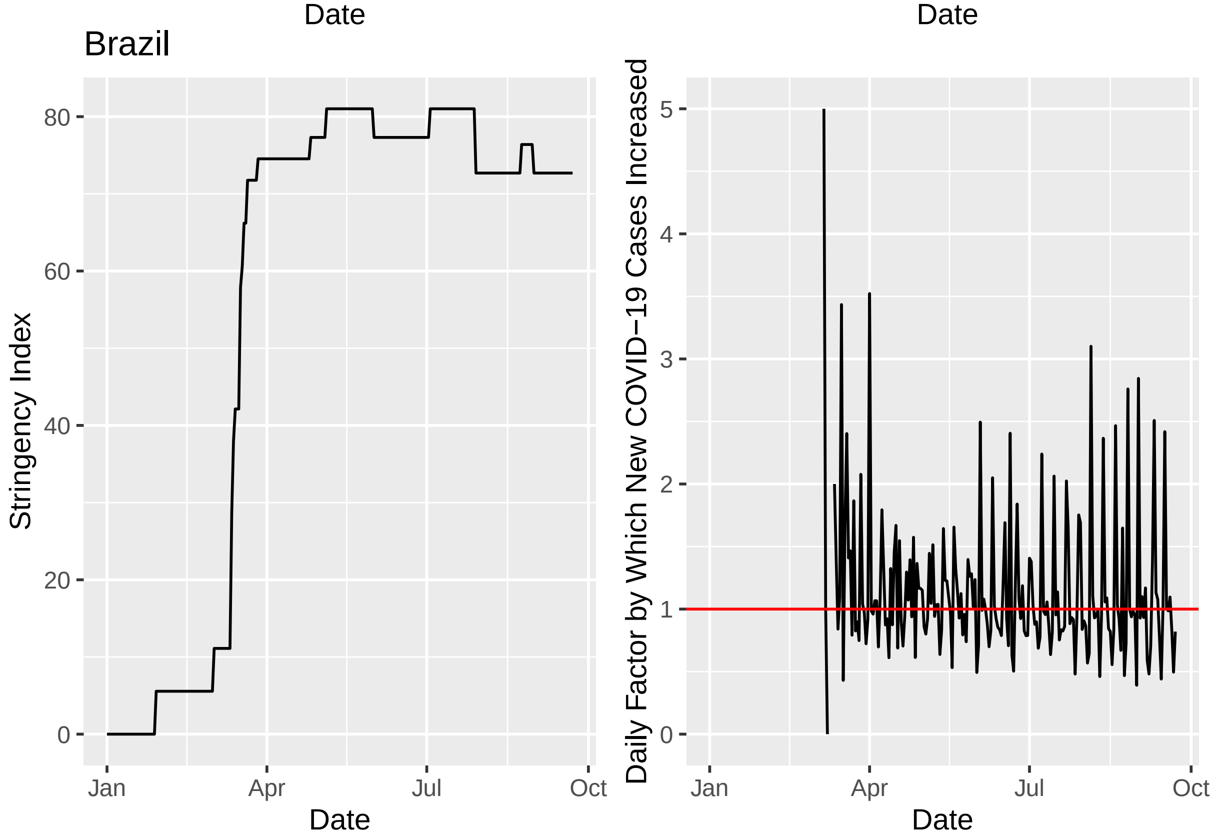
United States



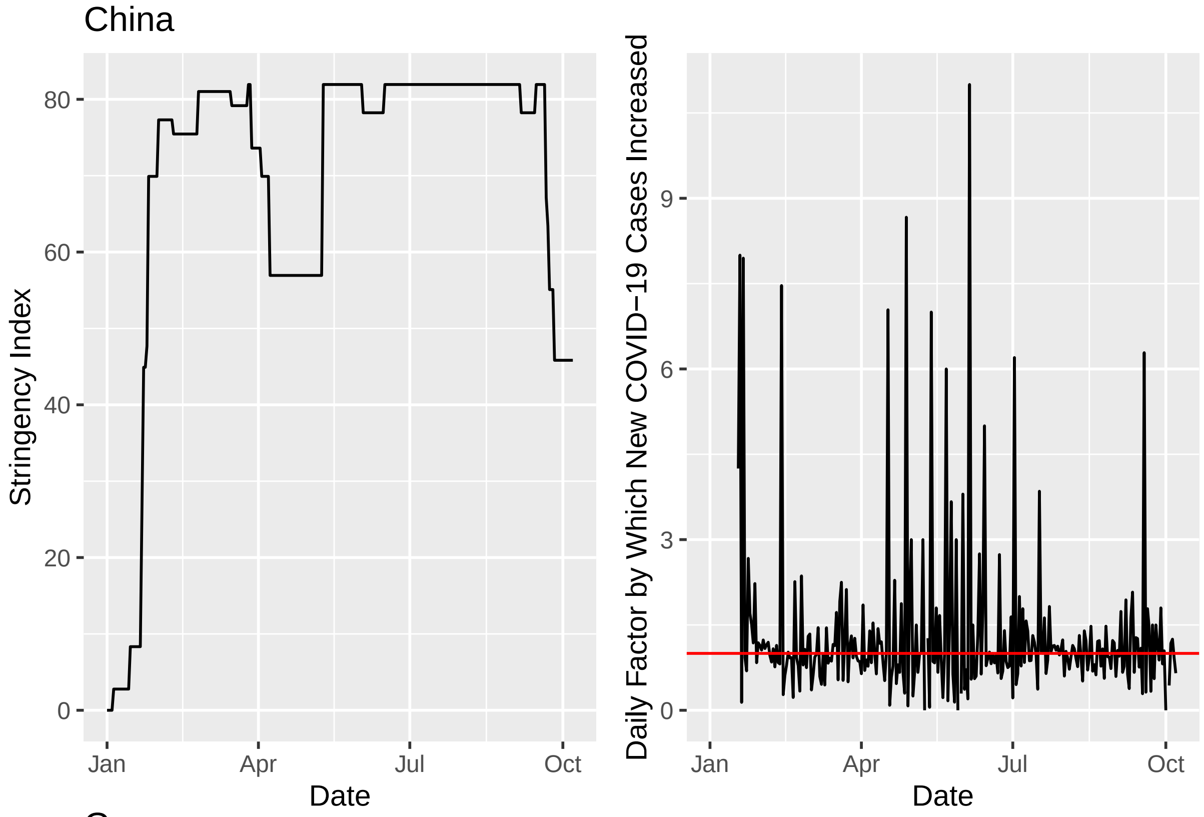
Russia

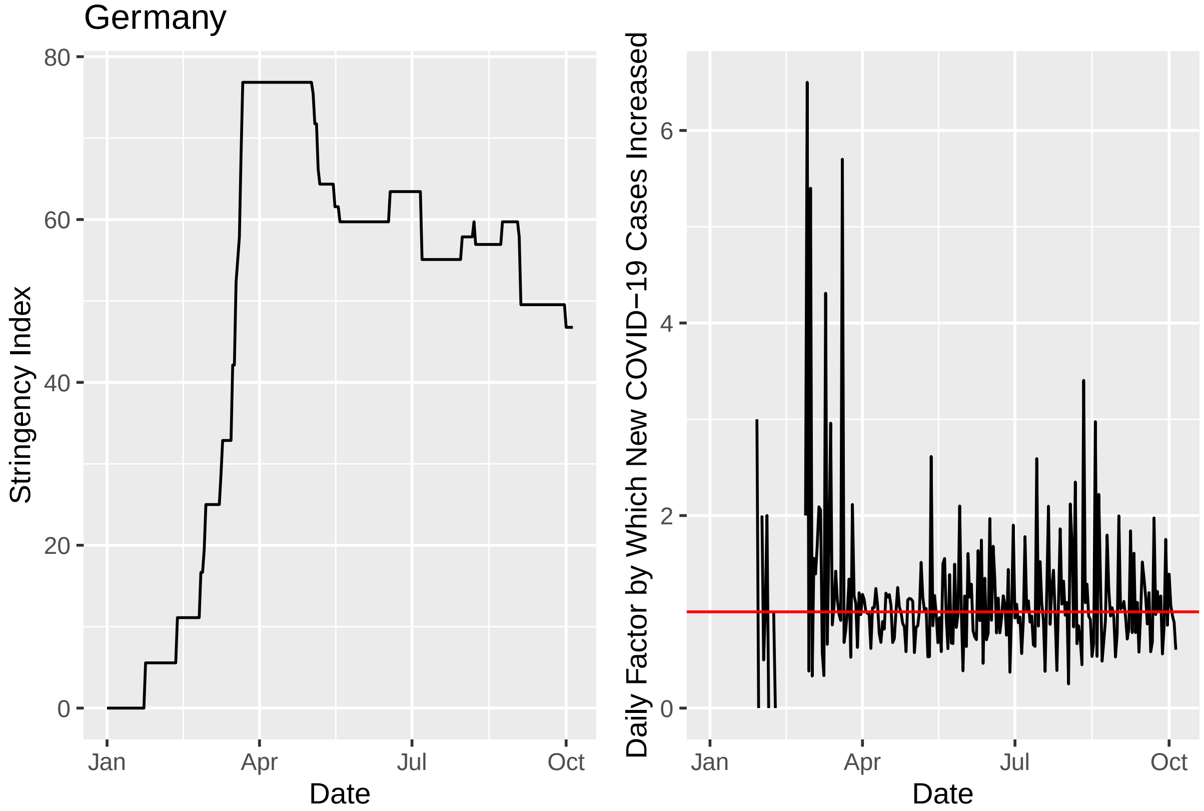


Saudi Arabia



Brazil





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. 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 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, 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. $\beta_{si} \neq 0$

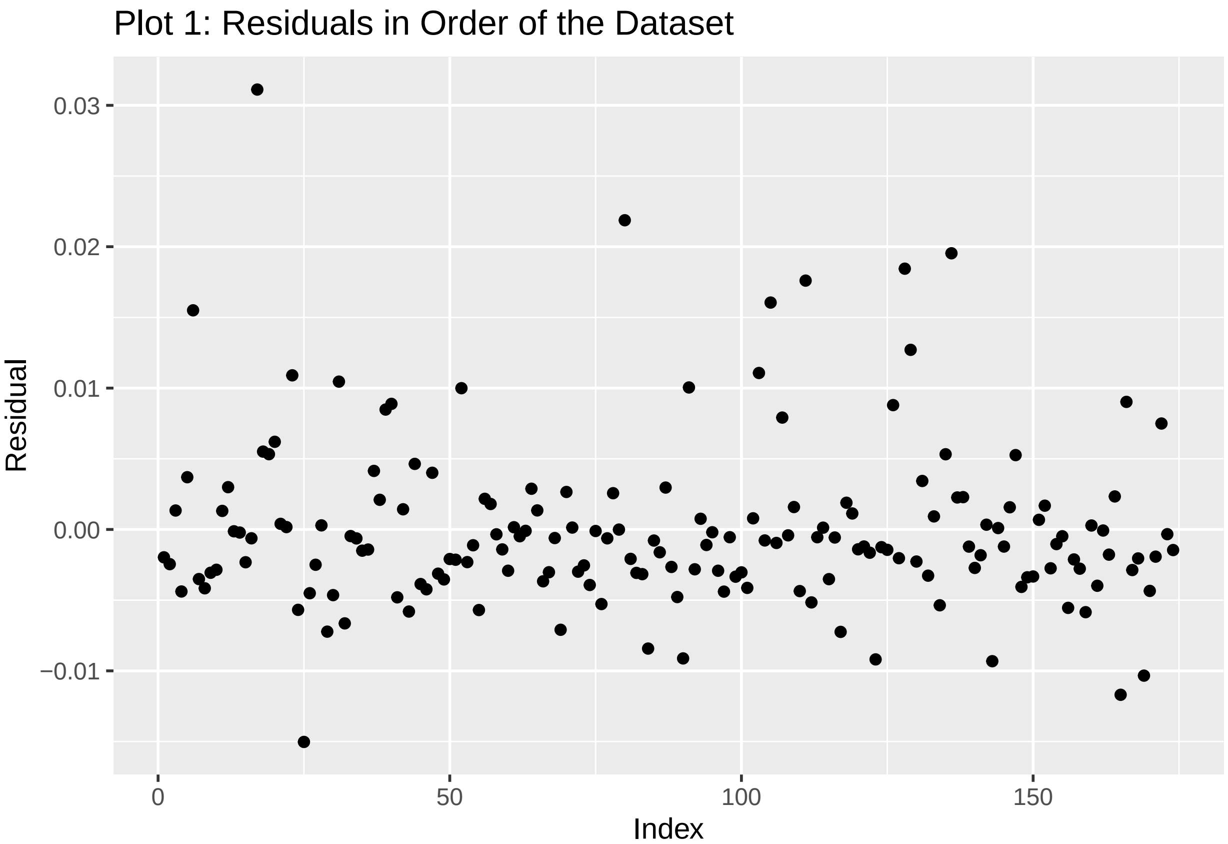
GDP per Capita: H_0: There is no relationship between GDP per Capita and total cases per capita. $\beta_{gdp} = 0$ H_1: There is a relationship between GDP per Capita and total cases per capita. $\beta_{gdp} \neq 0$

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. $\beta_{hdi} \neq 0$

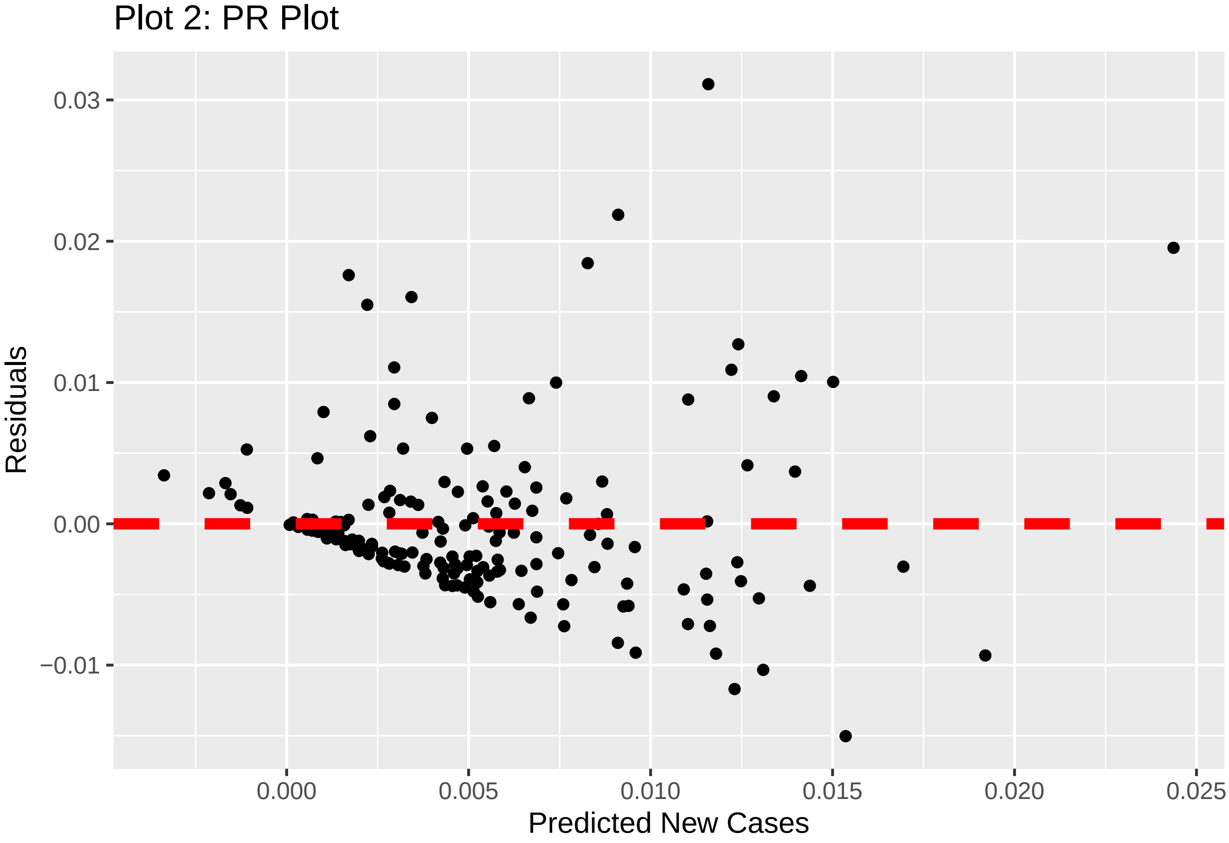
Population Density: H_0: There is no relationship between Population Density and total cases per capita. $\beta_{pd} = 0$ H_1: There is a relationship between Population Densityand total cases per capita. $\beta_{pd} \neq 0$

```
## # 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
```

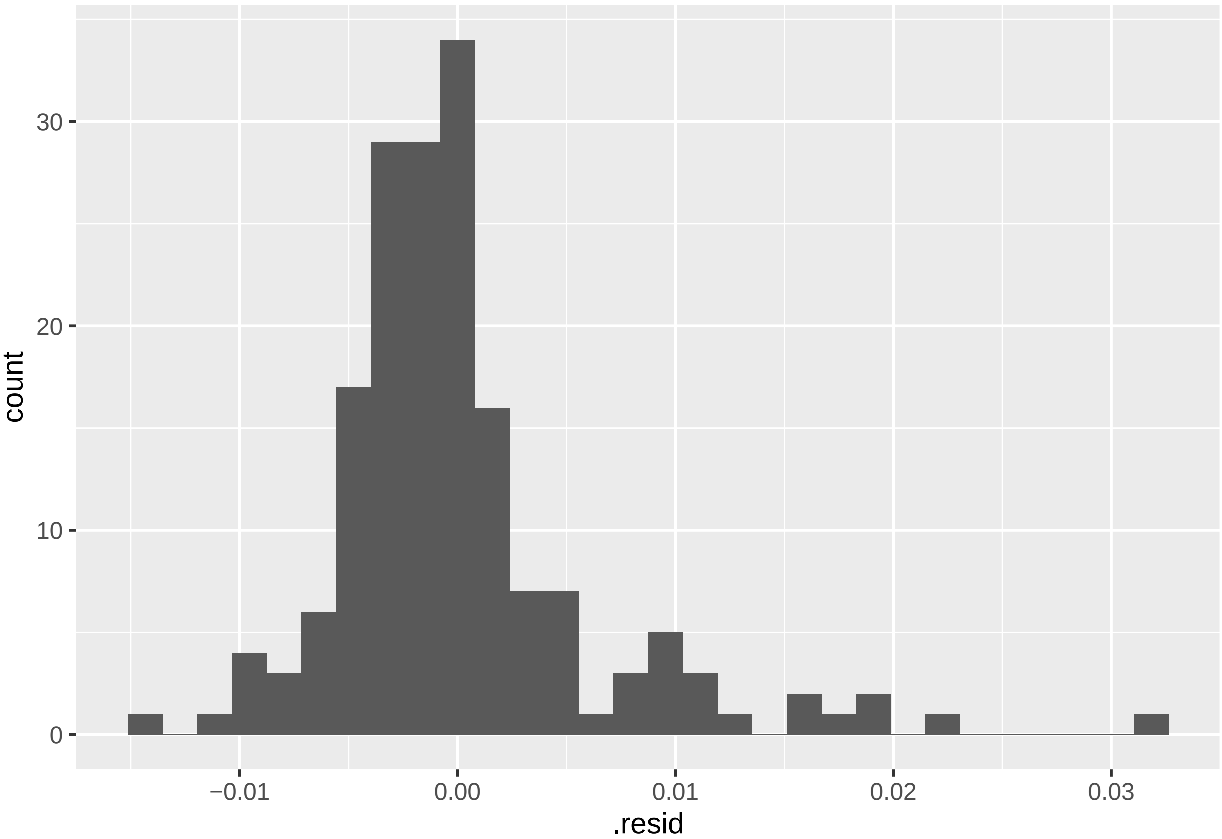
```
## # A tibble: 1 x 2
##   adj.r.squared r.squared
##           <dbl>     <dbl>
## 1         0.309     0.345
```



Plot 1: Residuals in Order of the Dataset



Plot 2: PR Plot



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.

Our model also has an adjusted R^2 of 0.3088684. In the context of our research, this means that 30.9% 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 statsitically 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. 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.

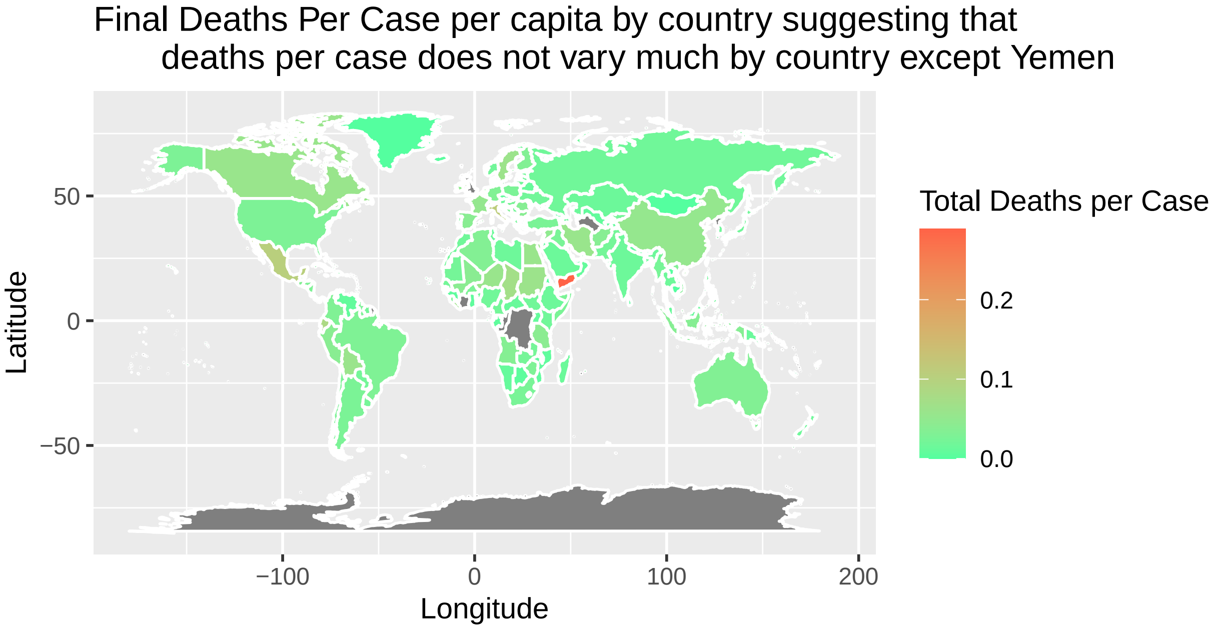
H_1: There is NOT independence between continent and total cases per capita.

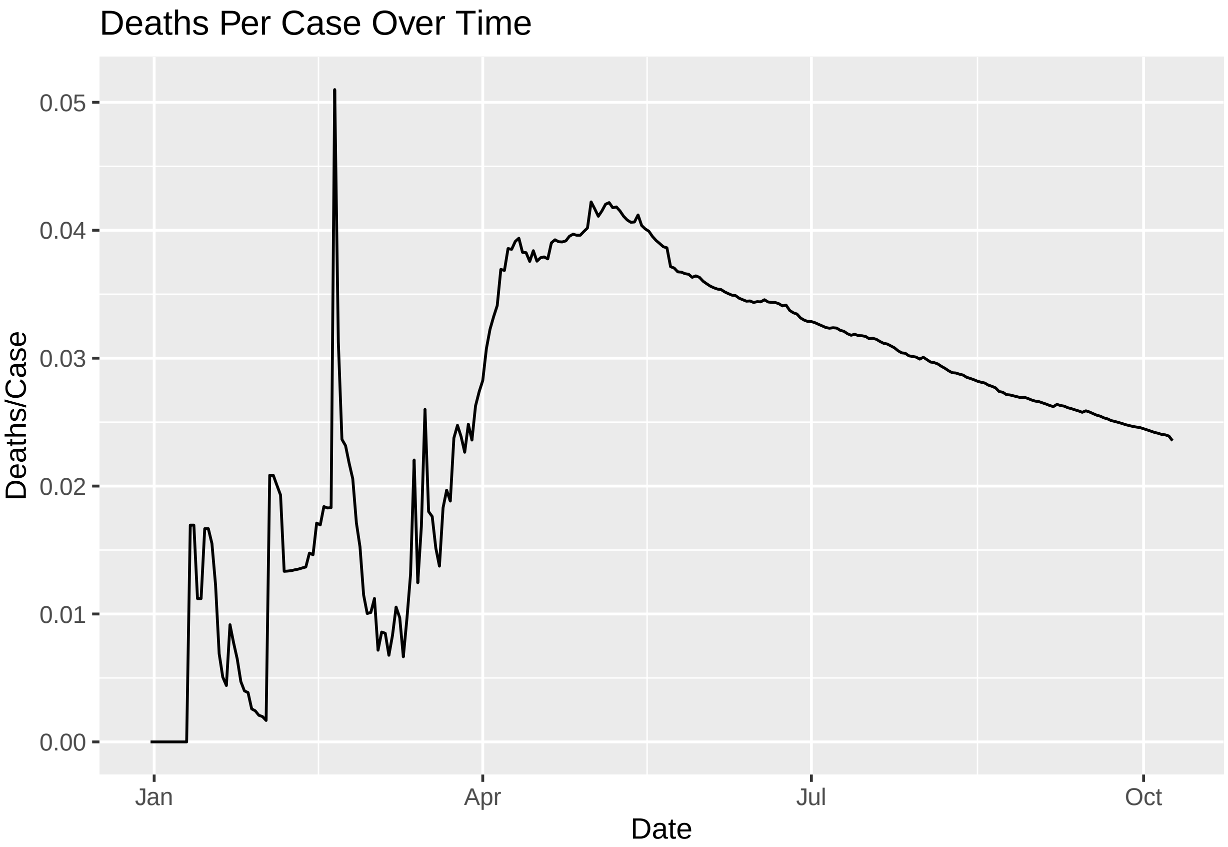
$\alpha = 0.05$

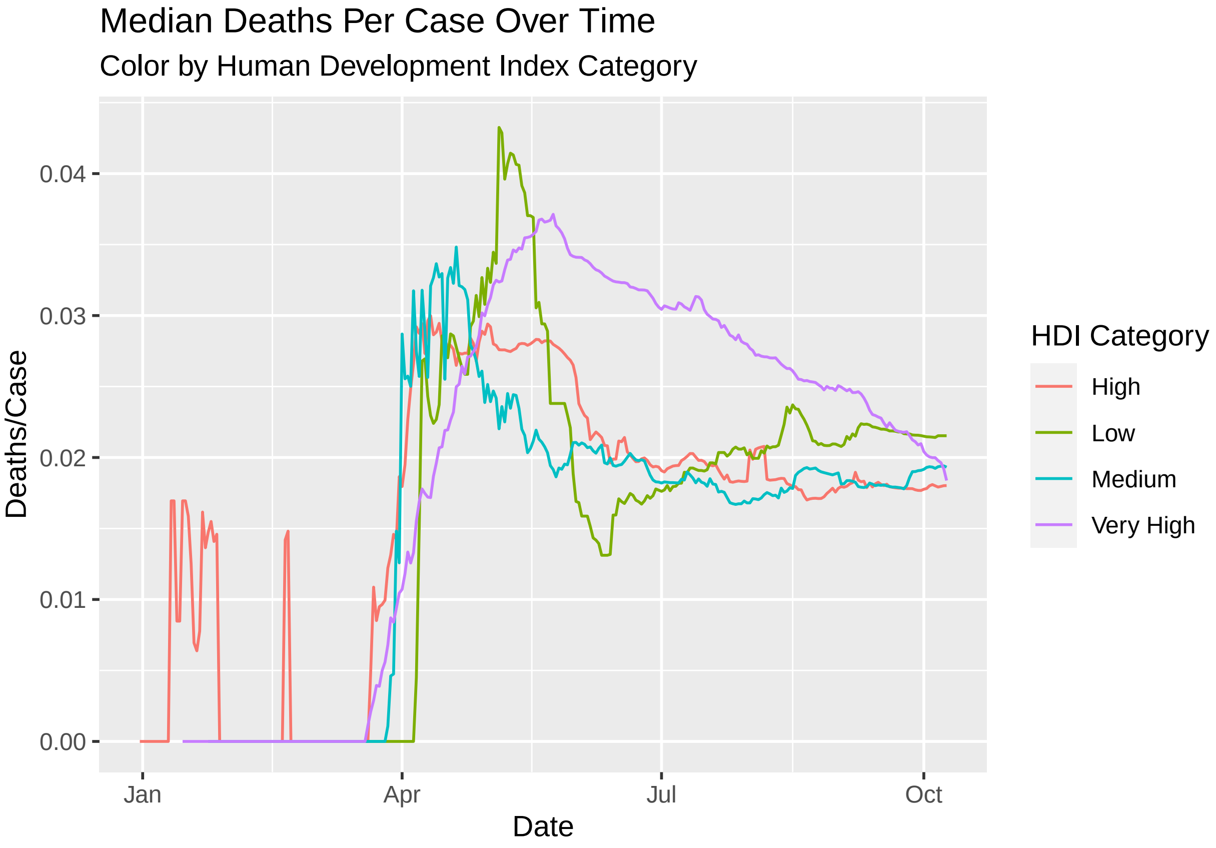
```
##
##  Pearson's Chi-squared test
##
## data:  table(covid_tcpc$continent, covid_tcpc$total_cases_per_cap)
## X-squared = 227767, df = 159400, p-value < 2.2e-16
```

After running a χ^2 test with a test statisitc of 227767 and 159400 degrees of freedom, we calculated a p value of 2.2e-16, 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.

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


Final Deaths Per Case per capita by country suggesting that
deaths per case does not vary much by country except Yemen
50
0
−50
Latitude
−100
0
100
200
Longitude
Total Deaths per Case
0.2
0.1
0.0



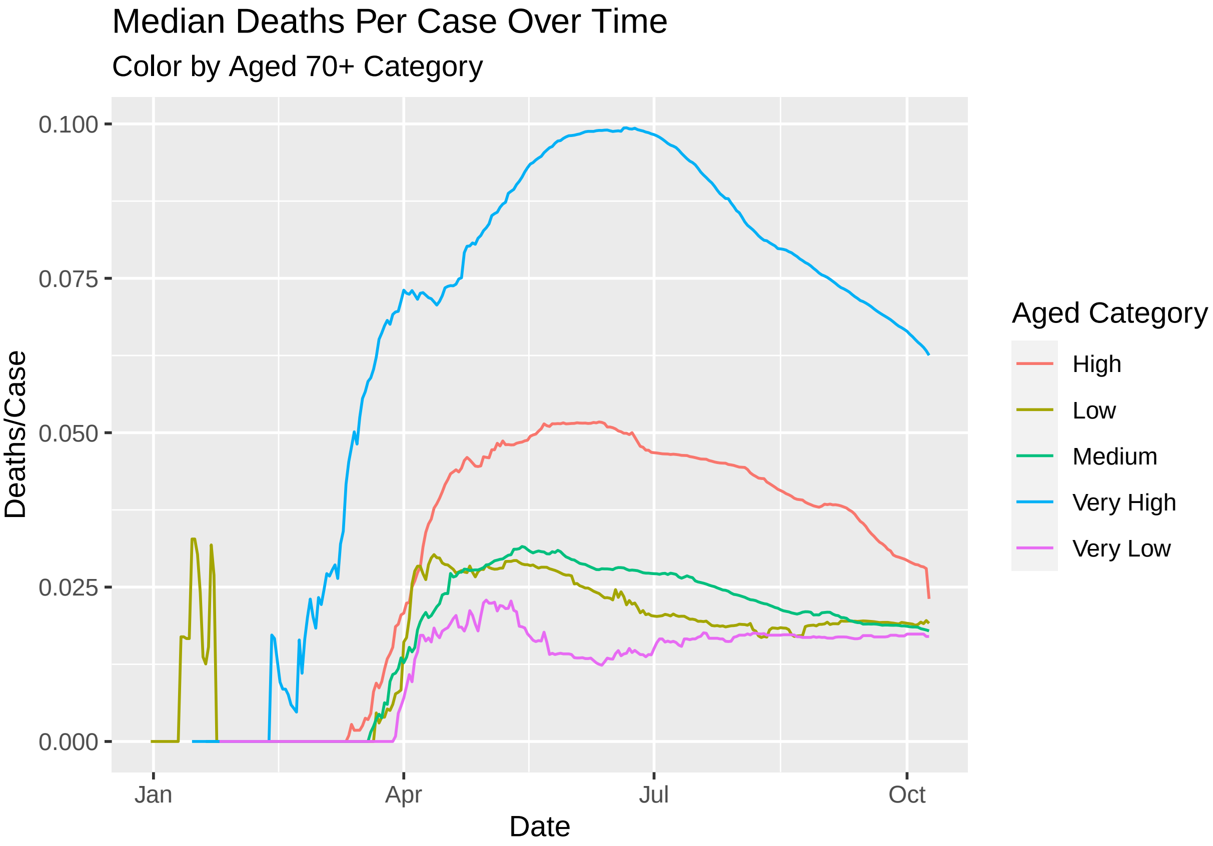


In visualizing the line plot of deaths per case over time colored by human development categories, there are clear distinctions between the different categories, especially around May and June. Thus, we will test for independence between human development index category and deaths per case. We will do 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

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

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, coronavirus data reporting and testing varies significantly by country, such that countries that lack comprehensive testing programs skew our anlaysis 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 coronavirus. 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 publishd 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.

Talk more about diagnostic plots: * Heteroscedastic? * Normality

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/