Final Project Proposal: COVID-19 Dataset

Due Friday, October 9, 11:59 PM

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

Section 1- 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 the citizens of their country. 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.

For our final project, we will be investigating country-level COVID-19 data to determine the relationship between their characteristics and demographics to virus transmission and deaths. By analyzing the effects of different country characteristics, we seek to determine specifically which variables are associated with stringency indexes, cases per capita, and deaths per case.

Research Question:

How do a country's characteristics, geography, and demographics impact the strictness of their COVID policy, as well as the total spread and effects of COVID-19?

Hypotheses:

We hypothesize that cases per capita will have a strong negative correlation to stringency index.

We hypothesize that deaths per case will be largely determined by the GDP per capita, the number of citizens aged 65+, hospital beds per thousand, the human development index, and columns concerning pre-existing conditions (i.e. diabetes prevalence, cardiovascular death rate, etc).

We hypothesize that the stringency index is going to vary the most by continent. We expect Europe, Asia, Oceania, and North America to have higher stringency indexes, while Africa and South America will have lower stringency indexes.

Section 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.

How The Data Was Originally Collected: "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 dataset updated as of October 9, 2020 (10:30, London time).

Sources: https://ourworldindata.org/coronavirus-source-data https://www.ecdc.europa.eu/en/covid-19/data-collection

Section 3- Glimpse of data

```r
library(tidyverse)

covid <- read_csv("data/covid-data.csv")
glimpse(covid)
```

```
## 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,...
```