Data Analysis Individual Project

Part 1: Dataset Selection & Description

Dataset chosen: The complete Our World in Data COVID-19 dataset as of 9/23/2020

Hyperlinks: https://ourworldindata.org/coronavirus-source-data
https://github.com/owid/covid-19-data/tree/master/public/data/
https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv

Dataset description

Domain of this dataset - Health

Size of the dataset, i.e. required storage:

- as a csv file, it is 9.55 MB
- as an excel workbook file, it is 7.16 MB

Metadata

Column name & data type	*Brief field description*	*Source*
iso_code - nominal	Three-letter country codes (in all caps)	International Organization for Standardization
continent - nominal	The continent of the geographical location	Our World in Data
location - nominal	Geographical location	Our World in Data
date - interval	The date of the observation in each record	Our World in Data
total_cases - ratio	Total confirmed cases of COVID-19	European Centre for Disease Prevention and Control
new_cases - ratio	New confirmed cases of COVID-19	European Centre for Disease Prevention and Control
new_cases_smoothed - ratio	New confirmed cases of COVID-19 (7-day smoothed)	European Centre for Disease Prevention and Control
total_deaths - ratio	Total deaths attributed to COVID-19	European Centre for Disease Prevention and Control
new_deaths - ratio	New deaths attributed to COVID-19	European Centre for Disease Prevention and Control
new_deaths_smoothed - ratio	New deaths attributed to COVID-19 (7-day smoothed)	European Centre for Disease Prevention and Control
total_cases_per_million - ratio	Total confirmed cases of COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control
new_cases_per_million - ratio	New confirmed cases of COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control
new_cases_smoothed_per_million - ratio	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people	European Centre for Disease Prevention and Control
total_deaths_per_million - ratio	Total deaths attributed to COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control
new_deaths_per_million - ratio	New deaths attributed to COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control

new_deaths_smoothed_per_million - ratio	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people	European Centre for Disease Prevention and Control
total_tests - ratio	Total tests for COVID-19	National government reports
new_tests - ratio	New tests for COVID-19	National government reports
new_tests_smoothed - ratio	New tests for COVID-19 (7-day smoothed)	National government reports
total_tests_per_thousand - ratio	Total tests for COVID-19 per 1,000 people	National government reports
new_tests_per_thousand - ratio	New tests for COVID-19 per 1,000 people	National government reports
new_tests_smoothed_per_thousand - ratio	New tests for COVID-19 (7-day smoothed) per 1,000 people	National government reports
tests_per_case - ratio	A rolling 7-day average of tests conducted per new confirmed case of COVID-19	National government reports
positive_rate - ratio	A rolling 7-day average of the share of COVID-19 tests that are positive	National government reports
tests_units - nominal	Units used by the location to report its testing data	National government reports
stringency_index - ratio	Government Response Stringency Index from 0 to 100 (100 = strictest response)	Oxford COVID-19 Government Response Tracker
population - ratio	Population in 2020	The U.N.
population_density - ratio	Number of people divided by land area, measured in square kilometers	World Bank
median_age - ratio	Median age of the population	The U.N.
aged_65_older - ratio	Share of population that's 65 years & older	World Bank
aged_70_older - ratio	Share of the population that is 70 years and older in 2015	The U.N.
gdp_per_capita - ratio	Gross domestic product at purchasing power parity (constant 2011 international dollars)	World Bank
extreme_poverty - ratio	Share of the population living in extreme poverty	World Bank
cardiovasc_death_rate - ratio	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)	Global Burden of Disease Study 2017 Results, Global Burden of Disease Collaborative Network
diabetes_prevalence - ratio	Diabetes prevalence (% of population aged 20 to 79) in 2017	World Bank
female_smokers - ratio	Share of women who smoke	World Bank
male_smokers - ratio	Share of men who smoke	World Bank
handwashing_facilities - ratio	Share of the population with basic handwashing facilities	The U.N.
hospital_beds_per_thousand - ratio	Hospital beds per 1,000 people	OECD, Eurostat, World Bank, national government records and other sources
life_expectancy - ratio	Life expectancy at birth in 2019	The U.N.
human_development_index - ordinal	Summary measure of average achievement in key dimensions of human development	The U.N.

1) a. All of this data on the Coronavirus pandemic was aggregated by Our World in Data which is a project and website put together by a collaborative effort between researchers at Oxford University in England and the non-profit organization Global Change Data Lab.

b. Their purpose, broadly speaking, is to publish data and tools for analyzing that data which are publicly available in order to help make the world a better place. To lift a quote from their website's about page, "To work towards a better future, we also need to understand how and why the world is changing."

2) a. The reason they sought out and aggregated so much data on various aspects of the impacts of the Coronavirus pandemic from so many different sources is to allow curious members of the public (with or without formal training in statistics, inductive inference, epidemiology, or data analysis) to conduct their own investigations of much of the relevant data on it so they can come to their own conclusions and become more informed about it.

b. This is a big data problem because there are over 45,000 records in the dataset (volume), 42 fields (variety) on well over a hundred different countries all over the world (variety), and the figures are updated daily (velocity).

c. There are no significant privacy issues with any of this data because no individual names or other identifying information are included for any specific persons. As for data quality, there are potential issues with data consistency across different countries and regions because they may be gathered in different ways by different entities and institutions with different data collection protocols and even slightly different definitions or interpretations of the definitions of some of the variables. Another potential data quality issue is places with lest COVID-19 testing might be underestimating the true number of cases. Countries which are poor and underdeveloped without very much modern medical infrastructure might be undercounting the number of deaths from the virus because many people in these countries may get sick and die in their homes rather than in a hospital where they can be counted. All medical tests generate both false positives and false negatives, it is unclear without looking into it further which is likely to outweigh the other when it comes to COVID-19 tests. One further potential quality of data problem here is that some doctors or nurses may put COVID-19 as the only cause of death or a cause of death for persons who already had one or more underlying health condition to such an extent that COVID-19 might better be thought of as a catalyst for their death rather than the cause, but this is a tricky problem to settle epistemologically speaking.

3) What potential questions could be answered by studying this data?

a. What is the average fatality rate for COVID-19 worldwide cumulatively?

What is the individual fatality rate for each of a handful of different countries?

Did places with different public policy approaches to combating COVID-19 have different outcomes?

What is the relationship between population density and total covid cases per million?

Does having a higher GDP per capita lower the fatality rate for a country?

Does "flattening the curve" have any effect (so far) on the total number of deaths in the long run?

4) For hardware, I recently purchased an extra 8 GB of RAM for my laptop which only came with 8 GB, thus bringing the total up to 16 GB of memory so I can conduct the various parts of my analysis of this data more quickly. I also bought a 250 GB external hard drive early last month, so if I run into any problems with data storage capacity on my laptop, I can use that or my 120 GB flashdrive. As for software, I plan on definitely using MySQL, Python, and R in my analysis and possibly also Tableau Public or Microsoft Excel should the opportunity for their use present itself during the analysis.

Data Analysis Individual Project

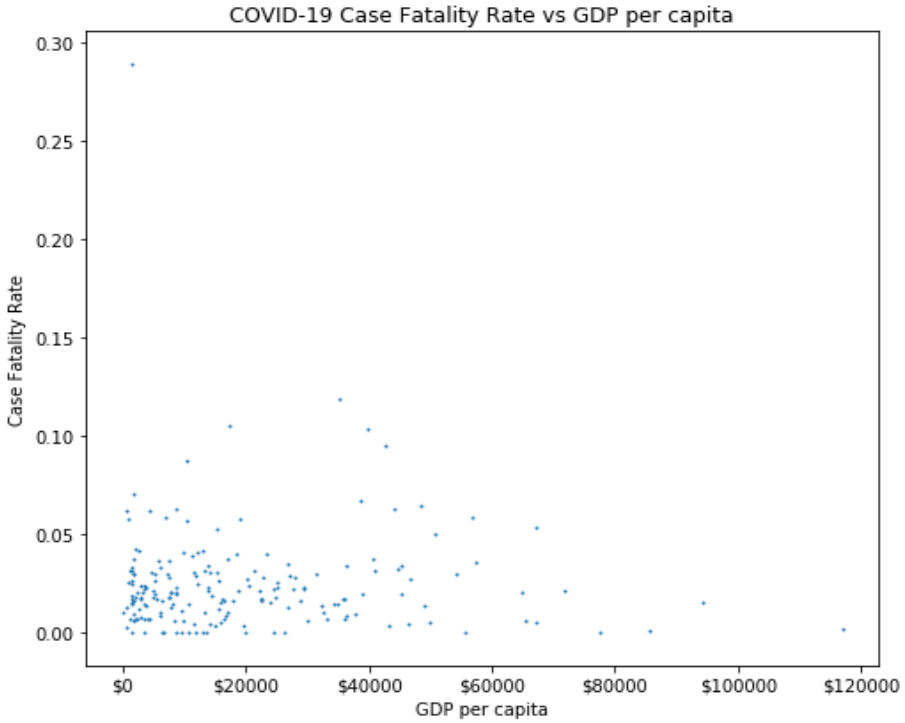
Deliverable 2: Dataset Analysis & Interpretation Report

1) a. Include at least one of each: scatterplot, boxplot, correlation analysis, regression analysis, & hypothesis test.

Scatterplot of GDP per capita vs the case fatality rate (total deaths ÷ total cases) of persons after contracting the SARS-CoV-2 virus created using Python with help from matplotlib.pyplot to produce the plot and pandas to read the data into Spyder, the IDE I wrote and ran the code in.

Python code I wrote to create the scatterplot:

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
# cd stands for current data, meaning all of the columns in the covid dataset for 9/23/20 only
cd = pd.read_csv("covid data for 9-23-20.csv")
CFR = cd['total_deaths']/cd['total_cases']
GDPpc = cd['gdp_per_capita']
fig, ax = plt.subplots()
plt.scatter(GDPpc, CFR, s=1)
plt.title("COVID-19 Case Fatality Rate vs GDP per capita")
plt.xlabel("GDP per capita")
plt.ylabel("Case Fatality Rate")
formatter = ticker.FormatStrFormatter('$%1.0f')
ax.xaxis.set_major_formatter(formatter)
plt.gcf().set_size_inches((8, 7));
```

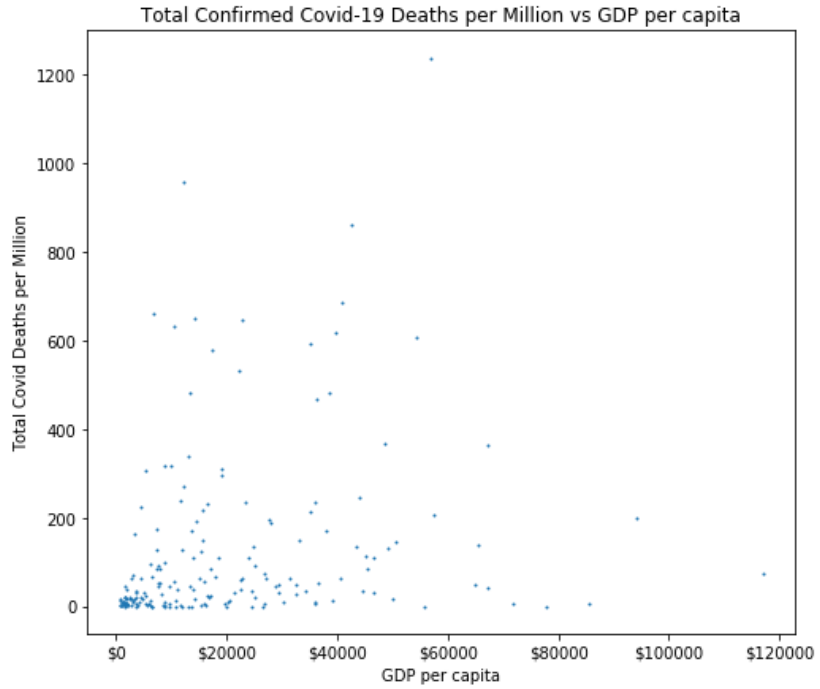


From a purely visual analysis of the scatterplot above, there does not appear to be any linear relationship between a country's GDP per capita and the fatality rate of its citizens from COVID-19.

I was not satisfied with the first scatterplot as I expected to see a correlation in the data, so I decided to create a second scatterplot attempting to visualize the same underlying comparison as before, but this time using total confirmed deaths from covid per million persons instead of case fatality rate as my proxy for the true death rate of covid in each country and compared that to GDP per capita for each country.

Python code I wrote to create the second scatterplot:

```
tdpm = cd['total_deaths_per_million']
fig, ax = plt.subplots()
plt.scatter(GDPpc, tdpm, s=1)
plt.title("Total Confirmed Covid-19 Deaths per Million vs GDP per capita")
plt.xlabel("GDP per capita")
plt.ylabel("Total Covid Deaths per Million")
formatter = ticker.FormatStrFormatter('$%1.0f')
ax.xaxis.set_major_formatter(formatter)
plt.gcf().set_size_inches((8, 7));
```



This second Graph tells roughly the same story as the first one, or rather the same lack of story, namely, there is no clearly visible linear relationship between the death rate in a country from contracting the COVID-19 illness after exposure to the SARS-CoV-2 virus and its income per capita. However, that could be due to how lousy of a proxy for how much liquid purchasing power the majority of a country's citizens posses at any given time to spend on quality medical care should they come down with a serious case of covid that gross domestic product per capita is more than it is due to there not being any real underlying relationship between one's ability to acquire medical treatment for covid and their probability of dying from it.

Boxplots

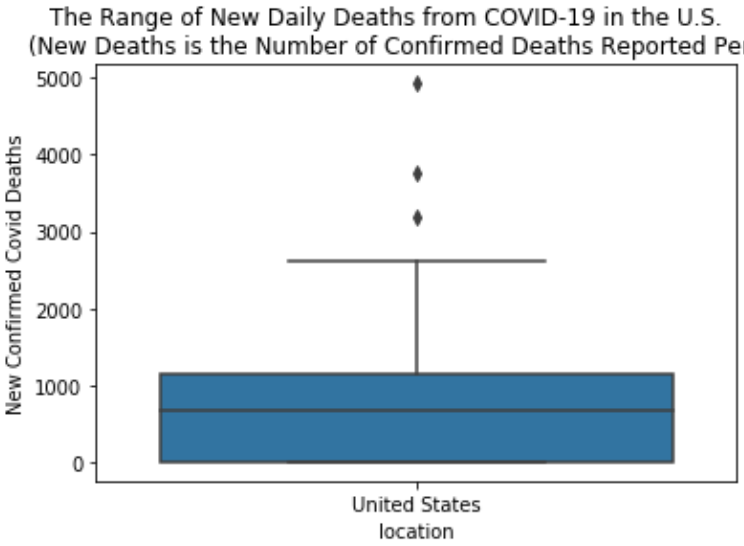
I created boxplots to visualize the range of new deaths from Covid figures for the United States and China in Python-Spyder with the assistance of the pandas library and the seaborn library. The new deaths column in Our World in Data's dataset I am using is new confirmed deaths per day tallied as occurring on the day reported.

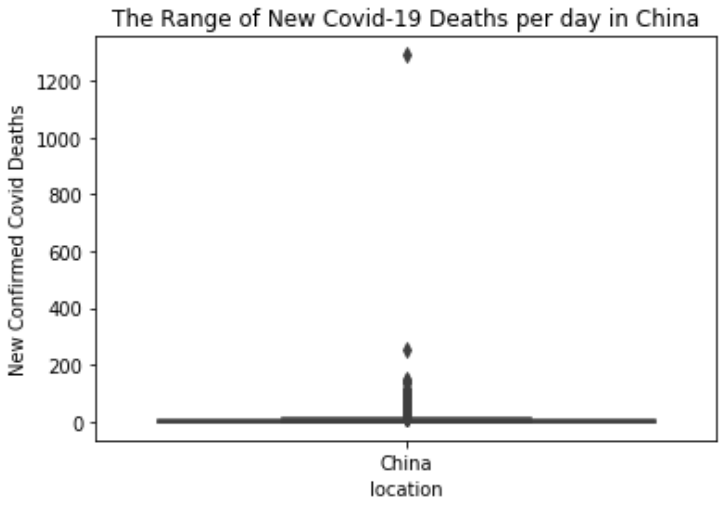
Python code used to create the boxplot showing the full range of new death numbers for the U.S.:

```
import pandas as pd
US = pd.read_csv("All USA COVID-19 data query results.csv")
import seaborn as sns
box = sns.boxplot(x='location', y='new_deaths', data=US)
box.set(ylabel="New Confirmed Covid Deaths")
box.set_title('The Range of New Daily Deaths from COVID-19 in the U.S.\
    \n (New Deaths is the Number of Confirmed Deaths Reported Per Day)');
```

Python code used to create the boxplot for China:

```
China = pd.read_csv("All China COVID-19 data query results.csv")
b2=sns.boxplot(x='location',y='new_deaths',data=China)
b2.set(ylabel = "New Confirmed Covid Deaths")
b2.set_title('The Range of New Covid-19 Deaths per day in China');
```





Analysis of the boxplots: Assuming the data coming out of the United States and China are accurate, which to be honest might be a pretty tenuous assumption for both of them, it is apparent after comparing these two box plots that the United States has consistently been having more new covid deaths, day after day and week after week than China for quite some time now and by large margins. This is not only unfortunate for us living in the U.S., but surprising as well considering this Coronavirus pandemic started in China, giving them less time to prepare for it than any other country, including the United States.

Something else noteworthy which is can be seen on these boxplots is both China and the US have extreme outliers in their new death records. There are three outliers in the American data and one in the Chinese data, but the single outlier in the Chinese covid data is much more extreme. These are more likely to be artifacts caused by bottlenecks and errors in the data production pipeline, i.e. the data gathering, transmission, or aggregation process than they are to accurately reflect such wild spikes in new deaths caused by covid.

Addendum: several hours after typing the analysis above, I discovered the following on OWID's website [1]: "For all global data sources on the pandemic, daily data does not necessarily refer to deaths on that day – but to the deaths reported on that day." Right next to a line graph with the title "Daily New confirmed COVID-19 deaths" and the subtitle "Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19."

Correlation Analysis

I used R via RStudio to investigate the correlation in OurWorldInData's COVID-19 dataset between the population density of a country and total COVID-19 cases per million in that country, here is the code:

```
tcpm <- data$total_cases_per_million
d <- data$population_density
r <- cor(d, tcpm, method = "pearson", use = "complete.obs")
r
```

According to RStudio's output console, there is only a correlation coefficient of $r = -0.0460$, i.e. 4.6% between the population density of a country and its total confirmed cases of covid per million of its citizens, so they are probably not really correlated.

In order to double check my interpretation of the weakness of the correlation implying that they are not correlated at all, I also conducted a *correlation hypothesis test* for the same two variables in RStudio. The null and alternative hypotheses for this test are: H_0: $r = 0$, H_1: $r \neq 0$.

The code I used to conduct the hypothesis test in RStudio:

```
test <- cor.test(population_density, total_cases_per_million, method = "pearson", use = "complete.obs")
test
```

The output in the console after running that code:

```
data:  population_density and total_cases_per_million
t = 9.5127, df = 42696, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.03651938  0.05544967
sample estimates:
      cor
0.04598865
```

So, it appears I was mistaken in the subjective judgment I made regarding the nonexistence of their correlation, this is a good illustration of why it is important to test your inferences and interpretations systematically whenever possible. While a correlation of around only 4 or 5% is small, having a p-value of less than $2.2 \times \frac{1}{10^{16}}$ means that it is extremely unlikely to have been nonzero due to chance.

Regression Analysis

In the regression analysis section of this analytics project, I will start off by running a simple linear regression (SLR), then I will construct and estimate a multiple regression model.

Simple linear regression

For my SLR, I am going to analyze how much of the variation in total confirmed covid deaths per million between different countries/locations can be explained or predicted by how stringent their governments' policy response to this pandemic has been. In other words, in this section, I am going to try to answer the question of whether places with different public policy approaches to combating COVID-19 had different outcomes. The only proxy for the severity of government responses to the pandemic in this dataset is to be found in the stringency_index column. This field tallies the score each country or region gets on the Government Response Stringency Index, a composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) which is estimated by researchers from Oxford's COVID-19 Government Response Tracker in the Blavatnik School of Government at Oxford. The stringency index is the only regressor I will need to include because simple linear regression means single variable regression.

This is the theoretical version of what my SLR is estimating:

$$TD_{PM} = \beta_0 + \beta_1 Stringency + \epsilon$$

RStudio code:

```
stringency <- stringency_index
SLR <- lm(tdpm ~ stringency)
summary(SLR)
```

The estimated SLR model for total confirmed covid deaths per million persons:

$$\widehat{TD}_{PM} = 37.3 + 0.309 Stringency$$

Analysis and interpretation of the output: With a p-value of $p < 2.2 \times 10^{-16}$, the stringency index is statistically significant. But the coefficient of determination for this simple linear regression model is only $R^2 = 0.00334$, so only 0.3% of the variation in total covid deaths per million can be explained by how strict the government response to it was which is a pitifully small amount.

There is another problem with the coefficient estimate for the stringency index, a much more fundamental which should be obvious if you step back for a second, its sign is backwards from what we would expect it to be. The whole justification for more stringent governmental responses to the COVID-19 pandemic rather than weaker responses is based on the hope and the belief that they will help reduce negative outcomes such as total deaths per million citizens, not increase them! The sign in front of $Stringency$ should be a minus sign, not a plus sign. But there is no need to panic yet because this was only a univariate analysis which is inevitably going to underwhelm us as an attempted explanation of such a complex phenomenon as a pandemic.

Multiple linear regression

For my multiple regression model, I am going to see if I can explain total deaths per million in a country by looking at a combination of its stringency index, the median age of its population, the percentage of its population aged 70 or older, its rate of extreme poverty, the prevalence of diabetes in its population, its score on the Human Development Index from the U.N., and maybe the percentage of its population who are smokers too. A more abstract expression of my multiple regression model would be:

$$TD_{PM} = \beta_0 + \beta_1 Stringency + \beta_2 Age + \beta_3 Old + \beta_4 Poverty + \beta_5 Diabetes + \beta_6 HDI + \epsilon$$

RStudio code:

```
# assign all of the regressors to new R variables with shorter names
age <- median_age
old <- aged_70_older
poverty <- extreme_poverty
diabetes <- diabetes_prevalence
HDI <- human_development_index
# create the multiple linear regression model and assign it to a variable
model <- lm(tdpm ~ stringency + age + old + poverty + diabetes + HDI)
summary(model)
```

The estimated multiple linear regression model for total confirmed deaths per million from covid:

$$\widehat{TDPM} = -144.3 + 0.995Stringency - 4.6Age + 12.7Old + 0.24Poverty - 2.1Diabetes + 319.4HDI$$

All 5 regressors and the intercept were individually statistically significant at the 1% level.
The overall model was highly statistically significant with a p-value of: $p < 2.2 \times 10^{-16}$
Adjusted R-squared: $R^2 = 0.1689$ which means that the variation in the regressors in this model specification only explain about 17% of the variation in the total confirmed deaths per million from covid.

Analysis and interpretation of the results of the model: In terms of trying to interpret what these coefficients mean practically speaking, they are a mess. The coefficient on $Stringency$ is not only positive again despite our hopes and expectations that it would be negative, but it is larger than in the simple linear regression which is movement in the wrong direction. Due to all of the reports in the news and from the CDC that the Coronavirus tends to be much more lethal to the elderly, you would expect the coefficient on the median age of a country's population to be positive, but it is negative which makes no sense. The regression coefficient on Old, which is the percentage of a country's population which is over the age of 70 is positive as you would expect it to be a priori. The estimate for the marginal effect of the proportion of a country's citizenry living in extreme poverty on the total death rate from covid is positive as one might predict, but its magnitude is so small that there is no practical significance to it despite it making it over the hurdle of statistical significance, it is always important to remember when interpreting statistical results that statistical significance and practical importance are not the same thing.

Insofar as the percentage of a country's population with diabetes serves as a good proxy for the proportion of their population which is obese (because type-II diabetes is often caused by obesity), you would expect the regression coefficient on $Diabetes$ to be large and positive, but it is neither which is very surprising to me. I have heard of the HDI index before, but I am not familiar enough with it be able to explain it well here, so I will quote the explanation of it on OurWorldInData's website:

"The Human Development Index (HDI) is an index that measures key dimensions of human development. The three key dimensions are:

– A long and healthy life – measured by life expectancy.

– Access to education – measured by expected years of schooling of children at school-entry age and mean years of schooling of the adult population.

– And a decent standard of living – measured by Gross National Income per capita adjusted for the price level of the country."

Given the fact that all three of the dimensions that go into calculating the HDI are positive things, one would expect to find a minus sign in front of its coefficient estimate, but the coefficient estimate for HDI after running my model has a plus sign in front of it and its magnitude is immense. It might be possible that this bizarre coefficient estimate for it has something to do with that the human development index is a composite of life expectancy which already has its own separate column in the dataset, mean years of schooling, and GNP per capita which is highly correlated with GDP per capita which also already has its own distinct column in the dataset.

Hypothesis Test

Since March of this year, many different people, in the news media and commentariat sphere, in the everyday acquaintance sphere on social media, and even some in the scientific community [2], have compared the lethality of this novel coronavirus pandemic the world is experiencing favorably with the seasonal flu. That is to say, many have made claims to the effect that this virus is no more dangerous than the seasonal influenza virus is every flu season. I have heard this claim made often enough that I thought I should try to find a way to test it during my analysis for this project, so I did.

The null hypothesis is that in the United States, the case fatality rate from COVID-19 is no worse than the case fatality rate from the seasonal flu. The alternative hypothesis is therefore that the case fatality rate from COVID-19 is worse than the CFR from the flu and because dying is bad, worse means higher.

$$H_0: \quad CFR_{covid} = CFR_{flu}$$
$$H_1: \quad CFR_{covid} > CFR_{flu}$$

The CFR from the flu in the U.S. for the 2019-2020 flu season I was able to calculate using information from the CDC is 0.01% ($\frac{55{,}00}{51{,}000{,}000} = 0.001078431$ to be exact).

Therefore, we have:

$$H_0: \quad CFR_{covid} = 0.00108$$
$$H_1: \quad CFR_{covid} > 0.00108$$

R code I wrote to conduct the hypothesis test:

```
US_CFR_data <- read.csv("US CFR data.csv")
t.test(x = US_CFR_data, mu = 0.001078431, alternative = "greater")
```

The output in the console after running the above lines of code in RStudio:

```
data:  US_CFR_data
t = 26.06, df = 246, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0.001078431
sample estimates:
 mean of x
0.03442915
```

As I expected, the null hypotheses that COVID-19 and the flu are equally deadly was rejected. And it was not rejected by a slight margin either, it had a test statistic of 26 which is cartoonishly large for a test statistic.

b. SQL Schema for the COVID-19 dataset & demonstrate several basic SQL-based queries of the dataset

```
-- create the table schema for the covid dataset
CREATE TABLE covid_data_table (iso_code CHAR(3),
        continent VARCHAR(20),
        location VARCHAR(30),
        date DATE,
        total_cases BIGINT,
        new_cases INT,
        new_cases_smoothed FLOAT,
        total_deaths BIGINT,
        new_deaths INT,
        new_deaths_smoothed FLOAT,
        total_cases_per_million FLOAT,
        new_cases_per_million FLOAT,
        new_cases_smoothed_per_million FLOAT,
        new_tests BIGINT,
        total_tests BIGINT,
        total_tests_per_thousand FLOAT,
        new_tests_per_thousand FLOAT,
        new_tests_smoothed FLOAT,
        new_tests_smoothed_per_thousand FLOAT,
        tests_per_case FLOAT,
        positive_rate FLOAT,
        test_units VARCHAR(25),
        stringency_index FLOAT,
        population BIGINT,
        population_density FLOAT,
        median_age FLOAT,
        aged_65_older FLOAT,
        aged_70_older FLOAT,
        gdp_per_capita FLOAT,
        extreme_poverty FLOAT,
        cardiovasc_death_rate FLOAT,
        diabetes_prevalence FLOAT,
        female_smokers FLOAT,
        male_smokers FLOAT,
        handwashing_facilities FLOAT,
        hospital_beds_per_thousand FLOAT,
        life_expectancy FLOAT,
        human_development_index FLOAT);
```

Some queries of the OWID covid dataset I ran in MySQL Workbench 8.0:

```
-- select all of the data for the USA
SELECT * FROM covid_data_table WHERE location = "United States";
# calculate the current CFR for the United States
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "United States";

SELECT (total_deaths/total_cases) AS cfr FROM covid_data_table
        WHERE location = "United States" AND (total_deaths/total_cases) IS NOT NULL;

 -- select all of the data for Canada
SELECT * FROM covid_data_table WHERE location = "Canada";

-- select all of the data for South Korea
SELECT * FROM covid_data_table WHERE location = "South Korea";
```

2) *Inferences, interpretations, and conclusions*

a. Questions I set out to answer in my analysis which were specified at the end of deliverable 1 of this project.

Q- What is the average fatality rate from COVID-19 worldwide since the pandemic started?

I calculated the cumulative mean global case fatality rate for COVID-19 (as of 9/23/20) in RStudio, it was 0.03145 or about 3.2%. Then I recalculated it again a slightly different way as a quick sanity check and got 0.03133 or roughly 3.1%, the different between the two was not substantial.

This means that on average around the world, about 3% of the people recorded as having caught COVID-19 have died so far, i.e. the ratio of all confirmed deaths from COVID-19 reported over total confirmed cases.

R code for global CFR calculations in RStudio:

```
data <- read.csv("owid-covid-data.csv")
# What is the average fatality rate for COVID-19 worldwide?
tdpm <- data$total_deaths_per_million
tcpm <- data$total_cases_per_million
global_death_rate <- mean(tdpm/tcpm, na.rm = TRUE)
global_death_rate
# calculate the same thing another way as a sanity check
td <- data$total_deaths
tc <- data$total_cases
global_death_rate2 <- mean(td/tc, na.rm = TRUE)
global_death_rate2
# calculate the current cumulative CFR worldwide
global_CFR <- max(td, na.rm = TRUE)/max(tc, na.rm = TRUE)
global_CFR
```

Q- What are the mean case fatality rates (CFR) and the current cumulative CFRs (as of 9/23/20) for a handful of different countries North America, Europe, and Asia.

Countries chosen: Canada, China, Italy, Japan, Mexico, Poland, South Korea, Spain, Sweden, UK, US, Vietnam

North American Countries

RStudio code to calculate the mean CFR for Mexico:

```
d.Mexico <- data[data$location == "Mexico", ]
td_Mex <- d.Mexico$total_deaths
tc_Mex <- d.Mexico$total_cases
CFR_Mex <- mean((td_Mex/tc_Mex), na.rm = TRUE)
CFR_Mex
label_percent()(CFR_Mex)
```

The mean CFR for Mexico is 0.09205 or 9.2% which is around 3 times higher than the worldwide CFR.

R code to calculate the current cumulative CFR for Mexico:

```
# load the Mexican data from the results of the SQL query I ran to isolate it
Mexico <- read.csv("All Mexico COVID-19 data query results.csv")
td_Mex <- Mexico$total_deaths
tc_Mex <- Mexico$total_cases
CFR_Mex <- max(td_Mex)/max(tc_Mex)
CFR_Mex
```

The current cumulative CFR for Mexico is 0.10542 or 10.5% which is extremely high unfortunately. As a Mexican American myself, the case fatality rate from Covid in Mexico being this high makes me sad.

R code to estimate the average CFR for the United States:

```
d.USA <- data[data$location == "United States", ]
td_USA <- d.USA$total_deaths
tc_USA <- d.USA$total_cases
CFR_USA <- mean((td_USA/tc_USA), na.rm = TRUE)
CFR_USA
label_percent()(CFR_USA)
```

The mean CFR for the U.S. is 0.03443 or 3.4% which is almost equal to the global mean CFR.

Code to calculate the cumulative CFR for the United States using R:

```
# load the American data from the results of the SQL query I ran to isolate it
USA <- read.csv("All USA data select query results.csv")
td_US <- USA$total_deaths
tc_US <- USA$total_cases
CFR_US <- max(td_US)/max(tc_US)
CFR_US
```

The current cumulative CFR for the U.S. is 0.02912 or 2.9% (as of 9/23/20).

RStudio code to calculate the mean CFR for Canada:

```
d.Can <- data[data$location == "Canada", ]
td_Can <- d.Can$total_deaths
tc_Can <- d.Can$total_cases
CFR_Can <- mean((td_Can/tc_Can), na.rm = TRUE)
CFR_Can
label_percent()(CFR_Can)
```

The average Canadian fatality rate from COVID-19 so far 0.05089 or roughly 5.1% which is almost two percentage points higher than the mean CFR for the U.S., its southern neighbor.

R code to calculate the cumulative CFR in Canada for the most recent date in my dataset:

```
td_Canada <- Canada$total_deaths
tc_Canada <- Canada$total_cases
CFR_Canada <- max(td_Canada)/max(tc_Canada)
CFR_Canada
```

The current Canadian CFR is 0.06296 or 6.3% which is over twice as high as the current American rate.

European Countries

MySQL query to get the cumulative CFR for the United Kingdom:

```
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "United Kingdom";
```

The case fatality rate in the U.K. is 0.1036 or roughly 10.4% (as of 9/23/20) which is surprisingly high.

MySQL query to get the cumulative CFR for Sweden:

```
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "Sweden";
```

The most current (9/23/20) CFR for in Sweden in my dataset is 0.0656 or about 6.6%.

MySQL code to calculate the current cumulative CFR in Italy:

```
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "Italy";
```

Current CFR in Italy is 0.1188 or roughly 11.9% which is the highest among the countries I picked.

MySQL code to calculate the current cumulative CFR for Poland:

```
SELECT total_deaths/total_cases AS cfr FROM covid_data_table
    WHERE location = "Poland" AND date = "2020-09-23";
```

The most current (9/23/20) CFR for in Poland in my dataset is 0.0287 or about 2.9%, the same as the U.S.

MySQL code to calculate the current cumulative CFR for Spain:

```
SELECT max(total_deaths)/max(total_cases) AS CFR FROM covid_data_table WHERE location = "Spain";
```

Current CFR in Spain is 0.0453 or roughly 4.5%.

Asian Countries

MySQL query to get the cumulative CFR for China:
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "China";
The current Chinese CFR is 0.0524 or about 5.2%.

Japan:
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "Japan";
Current Japanese CFR is 0.0190 or around 1.9% which is impressively low.

South Korea:
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "South Korea";
The CFR in South Korea (as of 9/23/20) is 0.0167 or about 1.7%, another impressively low rate indeed!

Vietnam:
SELECT max(total_deaths)/max(total_cases) FROM covid_data_table WHERE location = "Vietnam";
The current covid case fatality rate in Vietnam is 0.0328 or about 3.3%, which is just a tiny bit above the global rate.

Analysis and interpretation of the differences in case fatality rates between continents:
For the three North American countries selected, we have cumulative CFRs $(10.5, 2.9, 6.3)$. This means they have an average of $\overline{CFR}_{North\ America} = 6.6$.
We also have their individual mean CFRs $(9.2, 3.4, 5.1)$ which means the average of their individual mean case fatality rates is $\overline{\overline{CFR}}_{North\ America} = 5.9$.

For the five European countries selected, we have recent cumulative CFRs $(10.4, 6.6, 11.9, 2.9, 4.5)$, therefore, they have an average of $\overline{CFR}_{Europe} = 7.3$.

For the four Asian countries selected, we have recent cumulative CFRs $(5.2, 1.9, 1.7, 3.3)$, thus, their mean CFR is $\overline{CFR}_{Asia} = 3.0$.
As you can see, at least for the 12 countries I chose to analyze more closely on an individual basis, the Asian cohort has fared the best during this pandemic so far in terms of case fatality rate by a sizable margin. In fact, the Asian cohort was the only one to have a combined average case fatality rate anywhere close to the worldwide case fatality rate. The mean cumulative CFR for North America was well over 6% which is twice the worldwide rate of around 3.1% which is bad news for us given that we live here. The original three countries I chose to include in the European cohort, (The United Kingdom, Sweden, and Italy) all had such high case fatality rates that I ended up adding two more random European countries to my analysis (Poland & Spain) just to try to lower the European cohort's mean CFR which only brought it down to 7.3%. That is still an extremely high case fatality rate for any disease (e.g. as compared to 0.1% for the flu), it does appear thence that Europe has been hit pretty hard by the sudden invasion of the novel SARS-CoV-2 virus this year.

Analysis and interpretation of the differences in case fatality rates within continents:

As for the differences between the three North American countries selected, I think the data speaks for itself, the U.S. has done the best of the three so far while Mexico has had the worst outcomes, i.e. has suffered the most. I chose the United States, Canada, and Mexico because was born and raised in the United States, so if one goal of mine was to try to figure out how the United States has done, comparing these three would be a good comparison to make methodologically speaking because Canada and Mexico are its two closest neighbors and viruses don't care about national borders.

Q- Did regions with different public policy approaches to combating COVID-19 have different outcomes?

This question was already answered in the regression analysis section on pages 8, 9, & 10.

Q- What is the relationship between population density and the density or concentration of total covid cases?

This question was already answered in the correlation analysis section on page 7.

Q- Does having a higher GDP per capita lower the fatality rate or mortality rate?

This question was already answered by the scatterplots on pages 4 & 5, the answer it no, at least not in a linear fashion.

Q- Does "flattening the curve" have any effect (so far) on the total number of deaths in the medium term?

Unfortunately, I decided that this would be one of the questions I would try to answer before I had explored the data set thoroughly enough to realize that I could not answer it with this data set. That is why this is one of the questions I said I would answer in the data analysis section at the time I turned in the dataset selection deliverable.

b. *Value and limitations of this analysis and these findings*

I was able to provide more than one of everything required in part 2, the data analysis section of this project by the instructions for this project provided to us. And I am satisfied in that I was able to create all of the data visualizations I wanted to include.

However, what I was able to learn and the questions I was able to answer using this particular open source, freely available for download covid dataset turned out to be much more limited than I initially expected given the vast array of different types of charts, graphs, and visualizations of COVID-19 data you can make using the highly intuitive and interactive Data Explorer tool on Our World in Data's website which I have occasionally been using since May whenever I get curious about the current or cumulative covid figures. It took me until Thursday or Friday of this week to finally realize that many of the graphics on OWID's website must be using other data sources besides just the one they update daily as a downloadable csv, json, and xlsx file.

To take the most important example, there is no way you can estimate either excess mortality or the infection fatality rate (IFR) with their downloadable dataset, both of which are more important measures of lethality than case fatality rate. Excess mortality or excess death is a measure of the number of deaths in a country for a specific timespan due to all causes over and above the number of deaths in that same country due to all causes in previous years during each corresponding timespan, often done by the week. The case fatality rate of a disease is the total number of deaths divided by the total number of people that have the disease's clinical symptoms. In contrast, the infection fatality rate is the total number of deaths divided by the total number of people that carry the infection, including all of those who are asymptomatic. One final disappointing example is that the usual definition of mortality rate is deaths per 100,000 persons and this dataset had columns for total deaths, population, and total deaths per million, but no column for total for deaths per 100,000 which would have been easy to provide.