COVID-19
Data Analysis Project

May 7, 2020

Team Blue
Nicole Gee, Nicholas Richmond,
Matthew Clark, Swati Kohli, Joseph Carozza

Table of Contents

Introduction

In 2020 a coronavirus called SARS-CoV-2, also known as COVID-19, infected many people worldwide. The pandemic made a mark on history and left many wondering, how did this happen? In our project, we will attempt to analyze data in order to learn what factors influence a country's positive COVID-19 cases and deaths. First, we will collect, combine, and clean different data sources into one. Followed by exploratory data analysis to describe the data and look for interesting correlation and trends. Finally, we will use regression to further explore the data and interpret any significant correlations found in the model. We decided to take two different approaches to the regression analysis. In Part A, we will use linear regression to explore specific questions about the variable's impact on positive COVID-19 cases, and in Part B, we will use supervised Machine Learning techniques with LASSO and Multiple Linear Regression to narrow down a large list of variables to identify more significant factors related to deaths. These steps can be visualized in *Figure 1*.

Figure 1: Data Analysis Project Steps

Data Preparation

Right off the bat, we realized how hard it was to find the data we wanted to include in our analysis. Our strategy to approach the task was to first brainstorm different variables that may have an influence on COVID-19 including multiple health, finance, and travel factors for each country. We ended up finding only a small selection of the variables from various data sources like John Hopkins, Data Blogs, and Wikipedia. We acknowledge some challenges like the validity of data from sources like Wikipedia and the fact these data sources were all published at different times. For our project, we decide to move forward using the data knowing any insights we discovered will need to be researched further with better quality data before making strong claims. We downloaded all of the data we wanted to include in the analysis as separate .csv or .xlsx files. We read the separate data files into Rstudio and, using the John Hopkins data as the primary data set, combined additional variables from the other data files using simple left-join commands. This was done after cleaning variations of country names and labels so that the new data being added would match the primary John Hopkins data. A table of the variables collected and their data sources can be found in Appendix 1. Also, we noticed a number of missing values or special characters like '..' in certain variables that would affect the analysis. Therefore, we decided to remove these countries if the variable was being used in the regression.

Descriptive Analysis

After cleaning the COVID-19 data and adding fields from various data sources to serve as independent variables, we utilized a Python script to calculate summary statistics of each field, create histograms of the COVID-19 related data, and perform a correlation analysis between each independent and COVID-19 outcome variable. Before running the final dataset through the Python script, we also calculated rates of key COVID-19 fields to control for population (Appendix 3). Understanding raw counts of every infection and death is important since each instance of data represents millions of human lives; however, since each country differs vastly in population, analyzing raw counts alone can become misleading in understanding how a country is doing in response to the virus. The descriptive table (Appendix 3) helped our team begin to focus on more meaningful independent variables and understand the distribution of each field. We then used the script to generate histograms of the key COVID-19 raw totals and calculated rate fields (Appendix 3). The histograms revealed that the COVID-19 data generally has a skewed distribution for every field, underscoring the importance of identifying potential outliers that could impact our future regression models. On average, we observed 34,053 cases and 2,457 deaths per country; however given the distributions, we see that a majority of countries are clustered on the lower end of each graph and there are a few countries with significant outliers (mainly the US) (Appendix 4). Of the countries analyzed, the average infection rate by population (# of cases / total population) was 0.09%, but infection rates globally ranged from 0.000177% to .60%. As demonstrated by Appendix 4, several countries with high infection rates were likely skewing the average. On average, the death rate by population (# of deaths / total population) was 0.005%. However, the average death rate among positively identified cases (# of deaths / # of cases) was 4.10% on average. After this initial exploratory analysis and visualization, we created heatmaps using Python to determine one-on-one correlations between each “predictor” (non-COVID-19 fields) and “outcome” (COVID-19 fields) variable. We grouped the predictor variables into “financial” and “demographic” categories to better digest each matrix.

On the financial matrix, we observed that the infection rate (# of cases / total population) is [positively] correlated to GDP per capita at .77 (Appendix 5). This correlation may exist because developed countries are more likely to have interactions with the global economy and more opportunity for physical interactions due to high density in populated areas. We also observed that the test rate (# of tests / total population) is also [positively] correlated to GDP per capita by .63, indicating that countries with more resources can likely acquire the testing infrastructure to effectively test more of their population (Appendix 5). From the demographic matrix (Appendix 6), we observed a moderate correlation between both median age and the urban percentage of the population with both testing and infection rates (.45). Our main takeaway from this initial analysis was the interpretation that some of these fields could be potential influencers of COVID-19 data. We then decided to further investigate these relationships more holistically through multiple regression analyses.

<u>Regression Analysis</u>

Part A

For our first regression, we approached the problem with the driving question: Why are richer countries seemingly more impacted? To evaluate a country's "richness" we used the following variables to evaluate: GDP per capita, household income, income per capita, and healthcare spending per capita. We took additional data cleaning steps to prepare for this analysis like correcting the size of cases from raw count to per million residents and removed countries if they had missing GDP, population, or testing information with left 106 countries. By quickly ranking of the countries with most COVID-19 cases and the highest GDP we can extrapolate the following insights, of the top 10 countries with most cases:

- 4 are within the top 10% of GDP/ capita
- 8 are within the top 20% of GDP/ capita
- All are within the top 30% of GDP/ capita

Of the 10 countries with the least cases:

- 6 are within the lowest 10% of GDP/ capita
- 8 are within the lowest 20% of GFP/ capita
- All are within the lowest 30% of GDP/ capita

This leaves the glaring question, are richer countries more infectious? First, we approached the question by using linear regression in excel with the dependent variable, cases per million, and the independent variable, GPD per capita, to see if there is a strong correlation between the two. Our model produced a 0.58 R^2 value (Appendix 8) which measures the variation in the dependent variable that can be attributed to the independent variable. While 0.58 is not a terrible score we wondered if this could be explained because richer countries had more testing so we ran another linear regression with the dependent variable, tests per million, and independent variable, GPD per capita, and found a 0.43 R^2 value (Appendix 9). While it's not as strong as the previous model there is some correlation. This led us to our next question, is testing the confounding variable? In our third linear regression, we assigned the dependent variable as case per test, and the independent variable, GDP per capita, we got an R^2 value of 0.002 (Appendix 10). From our analysis, we can assume richer countries aren't more infectious but that the original correlation we found may be due to the increased rate of testing in richer countries. Although we cannot determine the causation of these correlations, we assume that richer countries simply have the means to detect more cases.

Part B

Through the data it turns out that the average deaths out of total positive cases has been 7.2! Therefore, we dig deeper into this aspect. In our next regression, we take a different approach with a driving question: Which predictors in financial, demographic and mobility categories all put together potentially explain the deaths? The deaths have been considered as response variable taken with population in millions and Covid infections (# positive cases). For statistical analysis of deaths/million population, due to skewed response variable (to the right), (see histogram in Appendix 10), log transformation is applied to make it more normal or symmetric such that it moves the big countries closer together and space out the smaller ones. This also helps meet the assumption of constant variance in the context of linear modeling.

The approach for this analysis is Supervised Machine Learning with a method called LASSO regression in Python to narrow down all of our variables of interest to learn which factors could bear some influence on Covid deaths. LASSO regression is a type of linear regression that shrinks the coefficients of less impactful variables which will be useful for our goal. Further, Multiple Linear Regression with hypothesis testing is employed to determine which predictors seem significant. Due to relatively smaller data for analysis and considering the volatility in the situation, we chose to increase the range of probability with level of significance at 10% (p-value as 0.01).

In order for the regression to work, we took some more data cleaning steps. We removed all countries where there were null values and standardized(and scaled) the data, being in different units for the ease of comparison after which we get 92 countries for analysis. Additionally, we make some assumptions for 5 countries which have significant cases and deaths to keep them in the analysis. They are:

1. China	• Total Tests=1000,000 (random assumption)
2. South Korea	• Female proportion= 50(assumed equal proportion of males and females)
3. Switzerland	• Median Per Capita Income 37466/3.2 = 11708 (applied formula: median per capita Household income/ avg household size *assuming 3.2*)
4. Czechia	• Female proportion 50 (equal proportion) • smoking prevalence (% of adults) 21.6 (average across the world)
5. Ireland	• Median Per Capita Income 28234/3.2 = 8823 (applied formula: median per capita Household income/ avg household size *assuming 3.2*)

After implementing the LASSO algorithm we observed some variables were removed from the model which we can assume have no correlation to deaths. Also, coefficients in the log of response gives the percent increase (or decrease) in the response for every one-unit increase in the independent variable. Even though R square values are 0.82, 0.36 respectively(Appendix 11,12), indicating over and underfitting, the analysis can be used to understand the relations. Predictors of interest with relation direction through this technique on deaths are as follows:

Y = Log(Deaths/ Million)	Y = Deaths/ Positive Cases
• Cases/million (+) • Pop. Proportion Female (+) • Median Age (+) • Import/Export (% of GDP)(-)	• Per Capita Spending on Health (+) • Intrnl Inbound Tourists(+)

From the analysis (and Appendix 13) , it can be seen that population age structure has some relation with the vulnerability towards the death toll due to Covid-19. While 65 years or older are not the only ones at maximum risk, the data analysis backs the current scenario as evidence that the disease is fatal as people testing positive and relatively older (together with underlying health conditions) have a significant relation to Covid-19 fatality rate. Since December, novel coronavirus has spread to numerous countries from the Chinese city of Wuhan. Logically, due to contagious nature of the disease, interactions play a role and mobility, i.e. transmission through active virus carriers through humans or surfaces imply increased effect on death rate due to risk in contracting the fatal infection (also evident through analysis result- International inbound tourists). Though, an inverse relation of Import/Export (% of GDP) with increasing deaths is an interesting insight because of low correlation (0.19) which can be explored further to understand the importance. Further, even though, as per the latest news, males appear to be dying at higher rates than women, the population proportion of females in analysis (predominantly ranging between 46-55%) might not be such an important variable but worth exploring with disaggregated data by sex to improve real-time targeted forecasting. Finally, nations spending more on healthcare does not mean more deaths due to Coronavirus but implies that through more testing more cases are detected some of which are not able to recover resulting in deaths.

In conclusion, our analysis pointed to some interesting insights, that with more time, could expand to more in-depth projects.

Challenges

Throughout the project, we noticed that obtaining the right data was the most challenging. We discovered the MICE (Multiple Imputation by Chained Equations) method that is effective in filing missing values. The technique involves running multiple regression models where each missing value is modeled conditionally depending on the observed (non-missing) values by taking the regression of the column. However, due to time constraints we could not apply this technique.

Modeling is not a perfectly determined prediction of the future, however, unlike weather conditions which we get accustomed to and incorporate into our day to day decisions, with pandemics we can actually influence the outcome through different techniques, measures and preventions.

Recommendations

In our analysis, we found some odd correlations - richer countries by GDP per capita have more cases per million and countries that spend more on healthcare per capita have higher death rates. Although we cannot find causation in this limited analysis, we assume the reason for these odd findings is due to these richer (by GDP and healthcare spending) countries simply having the means to test cases and detect deaths. Based on this assumption, we recommend international organizations provide funding and relief for countries unable to test at appropriate levels and for coalitions to support better healthcare. With better testing support and tracking, data can be better tracked worldwide and facilitate better learning about this disease’s behavior.

Appendix

Appendix 1:

Data & Source

Variable	Published Date	Data Source	URL
Country	2020	JHU	Link
# Positive Covid 19 Cases	Apr 29 2020	JHU	Link
# Deaths	Apr 29 2020	JHU	Link
Median Household Income	2006 - 2012	World Population Review	Link
Median Per Capita Income	2006 - 2012	World Population Review	Link
Population	2020	World Bank	Link
% Female population	2017	Our World in Data	Link
Per Capita Spending on Health	2015	Wikipedia	Link
Total Testing	2020	World Info Meter	Link
US $ GDP per Capita	2019	International Monetary Fund	Link
% Urban Population	2018	United Nations Development Programme	Link
People 65 and Over (Millions)	2018	United Nations Development Programme	Link
Median Age	2018	United Nations Development Programme	Link
International Inbond Tourists	2017	United Nations Development Programme	Link
Import/ Export % of GDP	2018	United Nations Development Programme	Link
International Student Mobilitty (% of total tertiary enrollement)	2010-2017	United Nations Development Programme	Link
R&D Expenditure (% of GDP)	2010-2017	United Nations Development Programme	Link
Cigarettes per person per year	2016	Wikipedia	Link
Smoking prevalence (% of auditis)	2016	Our World in Data	Link

Appendix 2:

Original Dataset: Screenshot

Country. Region	Cases	Deaths	medianHouseholdIncome	medianPerCapitaIncome	pop2020	Pop. Proportion Female	Per Capita Spending on Health	Total Tests	US$ GDP/Capita	% Urban Pop.	People 65 and Over (Millions)	Median Age	International Inbound Tourists(thousands)	Import/Export (% of GDP)	R&D Expenditure (% of GDP)	smoking prevalence (% of adults)
Luxembou	3769	89	52493	18418	625.978	49.74325	6236	42643	113196	91	8.57E-02	39.729	1046	415.4753	1.24366	23.5
Singapore	15641	14	32360	7345	5850.342	50.59025	2280	143919	63987	100	0.660004	42.226	13903	326.1947	2.15996	16.5
Malta	463	4	21141	6869	441.543	49.77706	2304	32989	30650	94.6	8.94E-02	42.596	2274	267.7761	0.59697	25.5
Ireland	20253	1190	28234	8823	4937.786	50.40446	4757	153954	77771	63.2	0.66815	38.246	10338	209.8109	1.17681	24.3
Slovakia	1391	22	17415	5455	5459.642	51.38578	1108	85922	19547	53.7	0.852265	41.249	2162	192.3459	0.78965	30.1
Belgium	47859	7501	26703	10189	11589.62	50.68928	4228	237963	45175	98	2.157357	41.928	8385	175.6481	2.48835	28.2
Hungary	2727	300	12445	4493	9660.351	52.43025	894	72951	17463	71.4	1.859736	43.336	5650	168.2775	1.20606	30.6
Lithuania	1375	45	12375	4719	2722.289	53.92806	923	125555	19266	67.7	0.55199	45.051	2523	161.9537	0.84724	28.8
Slovenia	1418	89	25969	8656	2078.938	50.34443	1772	52948	26170	54.5	0.407399	44.539	3586	160.9377	2.00202	22.5
Bahrain	2921	8	24693	4778	1701.575	37.27093	1190	129694	25273	89.3	3.81E-02	32.456	11370	159.1855	0.10116	26.4
Netherlanc	38998	4727	38584	14450	17134.87	50.24478	4746	219744	52367	91.5	3.274786	43.314	17924	155.3479	2.03247	25.8
Czechia	7579	227	22913	7821	10708.98	50	1322	242088	23078	73.8	2.071368	43.203	10160	151.4511	1.6783	21.6
Estonia	1666	50	12577	5031	1326.535	53.14547	1112	52741	23523	68.9	0.259641	42.424	3245	146.9634	1.28129	31.3
Belarus	13181	84	15085	5236	9449.323	53.46625	352	176625	6603	78.6	1.403255	40.335	11060.2	139.3435	0.58716	26.7
Malaysia	5945	100	11207	2267	32366	48.37675	386	160296	11136	76	2.103473	30.262	25948	132.2554	1.30069	21.5
Cyprus	843	15	18242	4932	1207.359	49.94015	1563	58109	27719	66.8	0.163156	37.25	3652	130.2879	0.50167	36.4
Bulgaria	1447	64	8487	2829	6948.445	51.3888	572	45208	9518	75	1.482383	44.596	8883	128.1416	0.78006	37
Cambodia	122	0	2308	451	16718.97	51.20651	70	11975	1620	23.4	0.742401	25.631	5602	124.8986	0.11823	17.2
Mongolia	38	0	5922	1440	3278.29	50.53249	152	7455	4132	68.4	0.129457	28.181	469	123.8266	0.13412	25.6
Thailand	2947	54	7029	1795	69799.98	51.2368	217	178083	7791	49.9	8.262606	40.102	35592	123.3069	0.78133	19.9
Georgia	517	6	2591	734	3989.167	52.25486	281	14718	4289	58.6	0.595057	38.268	6483	121.7427	0.30104	28.8
Switzerlan	29407	1716	37466	11708	8654.622	50.4543	9818	266200	83716	73.8	1.587743	43.053	9889	118.8316	3.3743	25.7
Latvia	849	15	10461	4000	1886.198	54.0734	784	57886	18171	68.1	0.386533	43.941	1949	118.3695	0.44292	37
Monteneg	322	7	11519	3123	628.066	50.66308	382	6864	8703	66.8	9.40E-02	38.8	1877	110.5984	0.374	45.9
Serbia	6630	125	8921	3020	8737.371	51.13594	491	85645	7397	56.1	1.614935	41.575	1497	110.1912	0.93022	38.9

Appendix 3:

Descriptive Statistical Table (count, mean, std, min, .25, .50, .75, max of each field)
(Including new calculated rate columns: see table below)

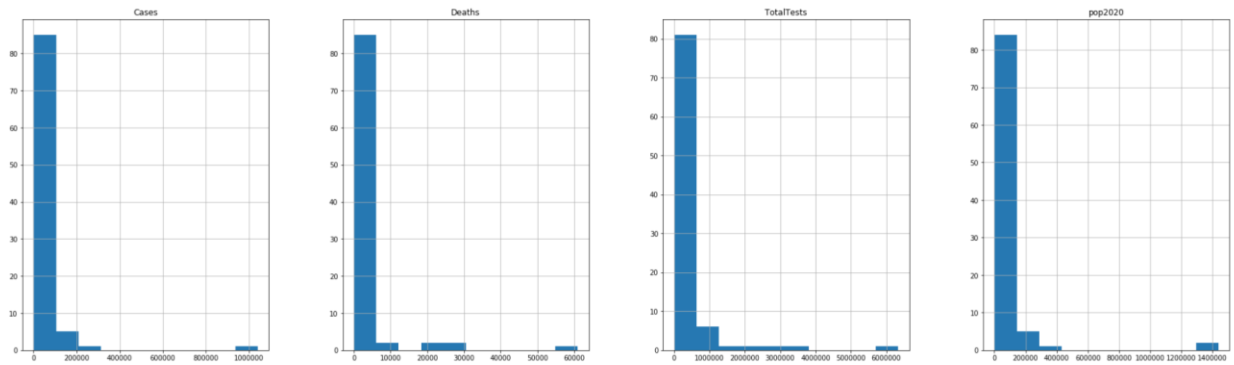
	Cases	Deaths	medianHouseholdIncome	medianPerCapitaIncome	pop2020	pop2020Total	PopProportionFemale	PerCapitaSpendingonHealth	TotalTests	USGDPCapita	UrbanPop&	People65andOverMillions	People65andOverTotal
count	92	92	92	92	92	92	92	92	92	92	92	92	92
mean	34,053	2,457	16,341.96	5,192.01	68,865.58	68,865,577	50.16	1,614.59	339,003	21,508.98	68.54	6.666701	6,666,701
std	115,563	8,142	13,954.04	5,177.52	209,961.79	209,961,787	3.43	2,105.23	822,455	23,023.89	18.53	19.285874	19,285,874
min	23	-	571.00	47.00	441.54	441,543	24.93	36.00	3,643	671.00	18.50	0.038100	38,100
0.25	1,387	18	5,745.75	1,273.25	5,339.46	5,339,460	50.00	277.25	32,201	4,257.50	57.05	0.505595	505,595
0.50	5,309	132	11,326.50	3,067.50	11,704.12	11,704,121	50.49	676.00	108,003	11,149.00	69.40	1.442819	1,442,819
0.75	16,325	716	25,227.25	7,380.50	47,786.81	47,786,806	51.18	2,092.50	238,994	30,845.00	81.68	5.081079	5,081,079
max	1,039,909	60,967	52,493.00	19,308.00	1,439,323.78	1,439,323,776	54.07	9,818.00	6,335,505	113,196.00	100.00	155.911750	155,911,750

	MedianAge	InternationalInboundTourists	ImportExport%ofGDP	R&DExpenditure%ofGDP	SmokingPrevalence%ofAdults	InfectionRateByPopulation	DeathRateByPopulation	DeathRateByPositiveCases	PositiveCasesPerTestsGiven	TestRatePerPopulation
count	92	92	92	92	92	92	92	92	92	92
mean	35.22	12,606	94.30	1.02	22.32	0.093646	0.005348	4.102905	8.271121	1.453742
std	8.00	17,633	61.02	0.99	9.37	0.125284	0.011787	3.648165	8.944926	1.680026
min	16.73	145	26.35	0.02	2.00	0.000177	-	-	0.276649	0.006641
0.25	29.64	2,151	57.35	0.28	15.60	0.008959	0.000275	1.527378	2.459079	0.192401
0.50	37.02	6,467	79.41	0.67	22.60	0.035443	0.000842	2.998159	5.773082	0.835623
0.75	42.24	14,088	112.54	1.28	28.33	0.139583	0.003669	5.589342	10.671508	2.293326
max	48.36	86,861	415.48	4.25	45.90	0.602098	0.064722	15.720886	59.200000	7.621997

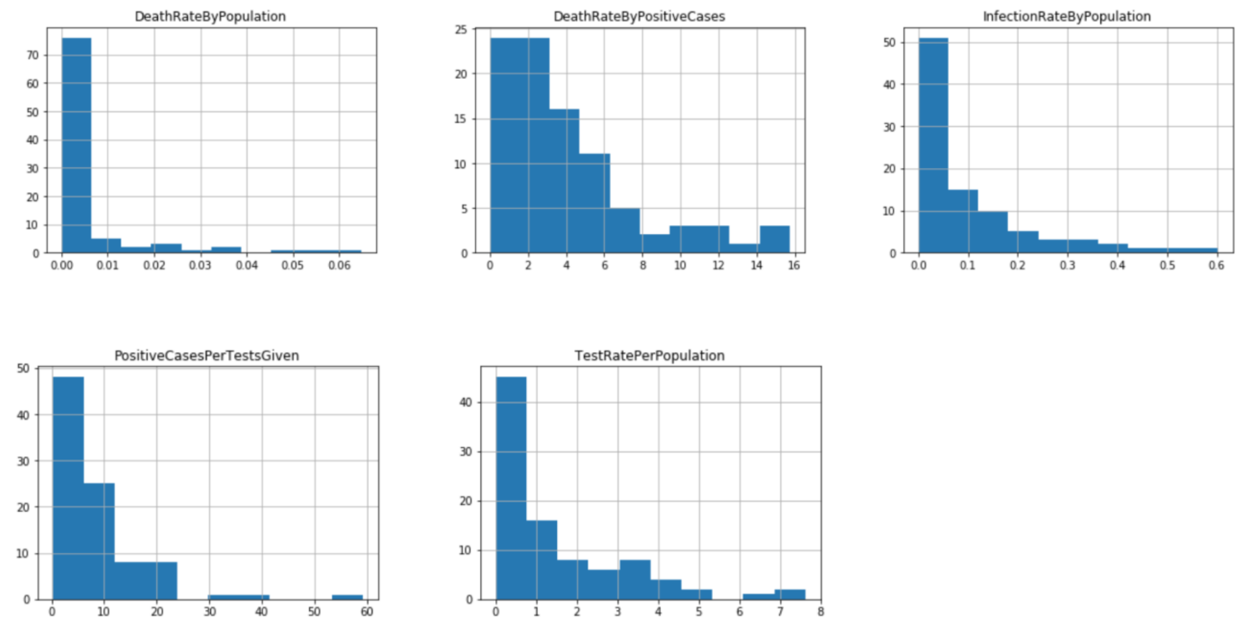
Rate Field	Calculation Method
InfectionRateByPopulation	Total Cases / Total Population
DeathRateByPopulation	Total Deaths / Total Population
DeathRateByPositiveCases	Total Deaths / Total Cases
PositiveCasesPerTestsGiven	Total Cases / Total Tests
TestRatePerPopulation	Total Tests / Total Population

Appendix 4:

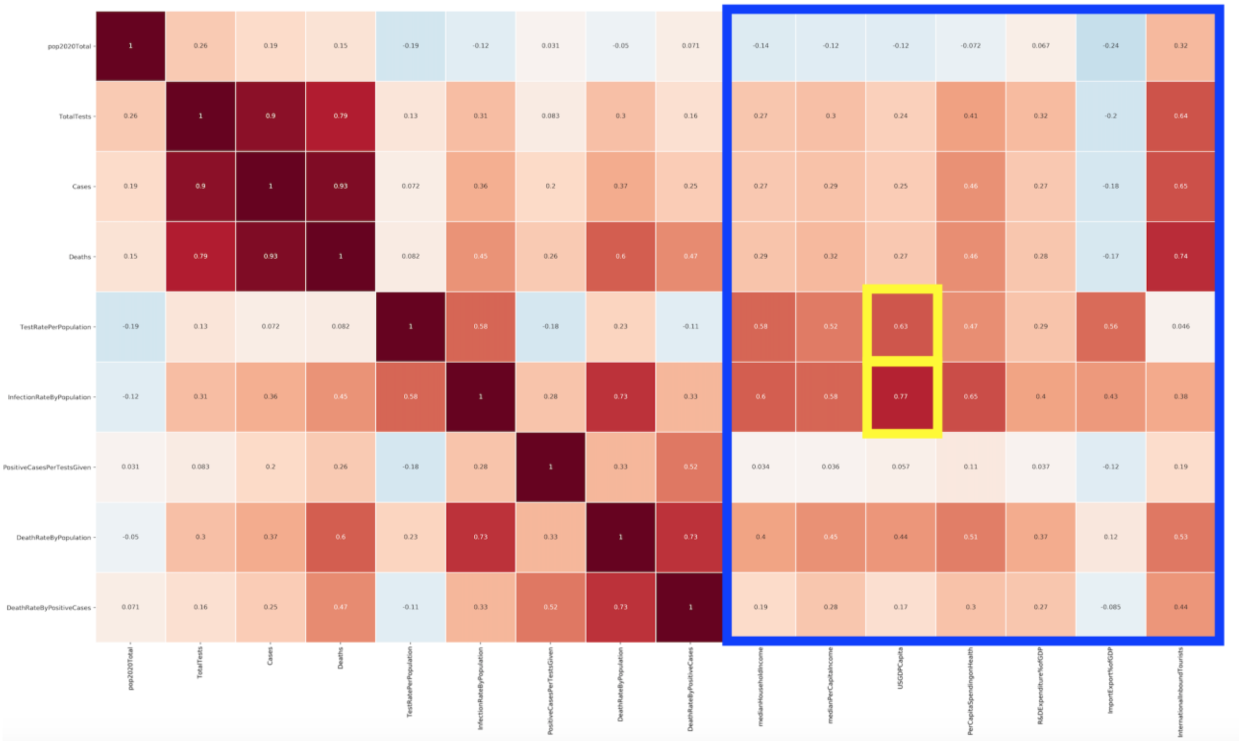
Histograms of raw totals (Cases, Deaths, Tests, Population)



Histogram of rates (Death Rate of Population, Death Rate of Positive Cases, Infection Rate of Population, Positive Case Rate Per Tests Given, Test Rate of Population)

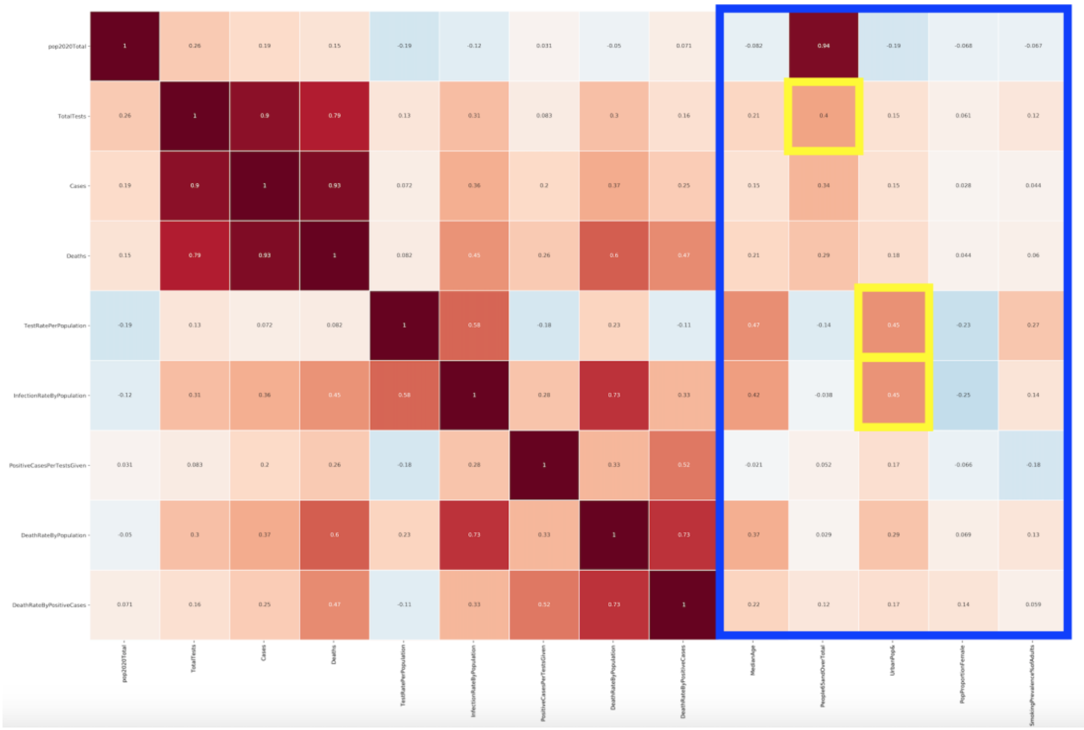


Appendix 5:
Heatmap of 1:1 Correlations (Financial Independent Variables: Median Household Income, Median Income Per Capita, GDP Per Capita, Healthcare Spending Per Capita, R&D Expenditure % Of GDP, Import/Export % Of GDP)



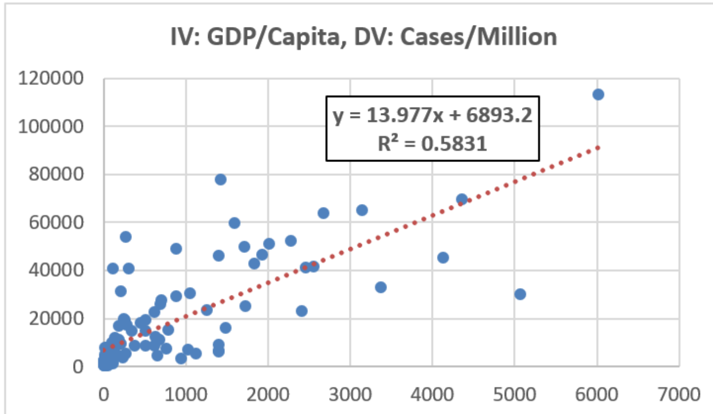
Appendix 6:

Heatmap of 1:1 Correlations (Demographic Independent Variables: Median Age, People Over 65, Urban Population %, Female Population %, Smoking Prevalence)



Appendix 7:

Independent Variable: GDP per Capita; Dependent: Cases/Million

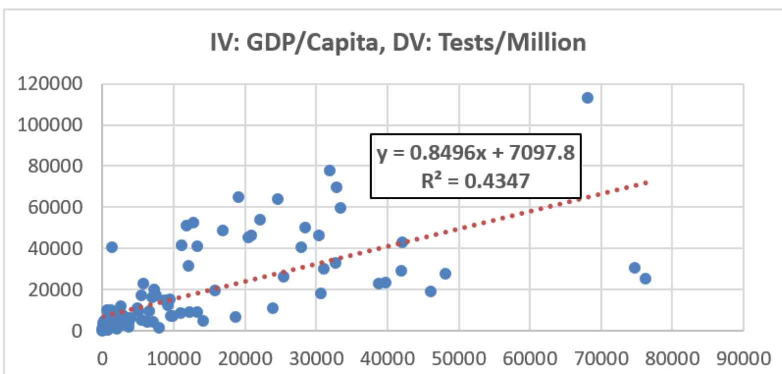


Regression Statistics	
Multiple R	0.763617357
R Square	0.583111468
Adjusted R Square	0.579102924
Standard Error	739.0203955
Observations	106

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	22.91235277	93.45969627	0.245158	0.806818	-162.4217322	208.2464377	-162.4217322	208.2464377
GDP/Capita	0.041719731	0.003459067	12.06098	0.00	0.034860271	0.048579191	0.034860271	0.048579191

Appendix 8:

Independent Variable: GDP per Capita; Dependent: Tests/Million

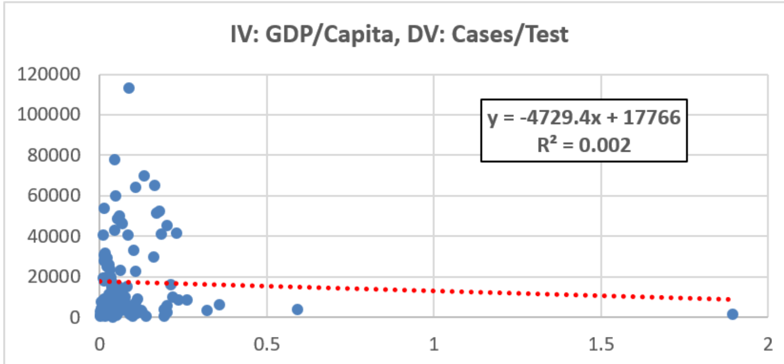


Regression Statistics	
Multiple R	0.659298909
R Square	0.434675052
Adjusted R Square	0.429239235
Standard Error	12223.30926
Observations	106

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3159.187403	1545.812237	2.043707073	0.043510105	93.78372796	6224.591078	93.78372796	6224.591078
GDP per Capita	0.511612728	0.057212561	8.942314733	1.54036E-14	0.398158073	0.625067384	0.398158073	0.625067384

Appendix 9:

Independent Variable: GDP per Capita; Dependent: Cases/Test

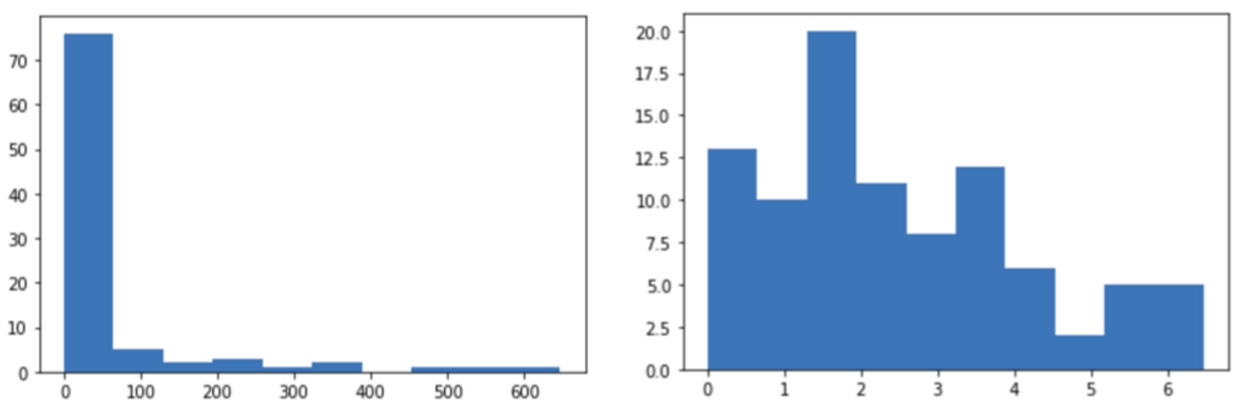


Regression Statistics	
Multiple R	0.044707911
R Square	0.001998797
Adjusted R Square	-0.007597368
Standard Error	0.197846308
Observations	106

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.105214202	0.025020495	4.20512078	5.54016E-05	0.055597622	0.154830781	0.055597622	0.154830781
GDP per Capita	-4.22636E-07	9.26042E-07	-0.456389366	0.649061704	-2.25901E-06	1.41374E-06	-2.25901E-06	1.41374E-06

Appendix 10:

Histogram plot of response variable: Death/million population and Log(Death/million population)



Appendix 11:

Y =Log (Deaths per Million)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      LogDeaths per mil   R-squared:                       0.820
Model:                            OLS   Adj. R-squared:                  0.807
Method:                 Least Squares   F-statistic:                     61.10
Date:                Wed, 06 May 2020   Prob (F-statistic):           1.27e-23
Time:                        16:22:21   Log-Likelihood:                -78.732
No. Observations:                  73   AIC:                             169.5
Df Residuals:                      67   BIC:                             183.2
Df Model:                           5
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       2.4331      0.087     27.992      0.000       2.260       2.607
medianPerCapitaIncome       0.1477      0.123      1.201      0.234      -0.098       0.393
Pop. Proportion Female      0.4810      0.096      5.026      0.000       0.290       0.672
Median Age                  0.4681      0.130      3.606      0.001       0.209       0.727
Import/Export (% of GDP)   -0.3756      0.096     -3.916      0.000      -0.567      -0.184
cases per million           1.2358      0.111     11.179      0.000       1.015       1.456
==============================================================================
Omnibus:                        0.165   Durbin-Watson:                   1.564
Prob(Omnibus):                  0.921   Jarque-Bera (JB):                0.256
Skew:                          -0.107   Prob(JB):                        0.880
Kurtosis:                       2.804   Cond. No.                         2.80
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
MSE on test set 1.0288700552388288
```

Appendix 12:

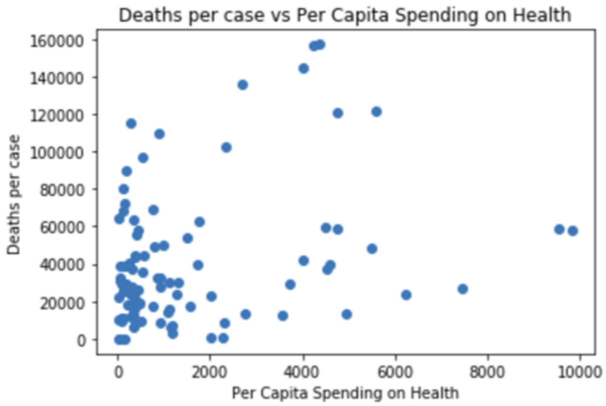
Y = Deaths per million Cases

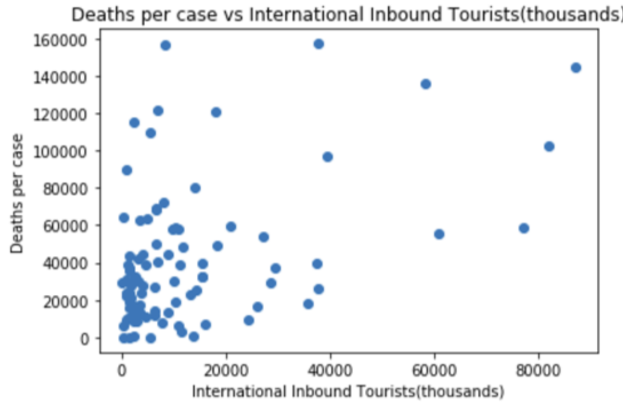
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    dpc   R-squared:                       0.360
Model:                            OLS   Adj. R-squared:                  0.269
Method:                 Least Squares   F-statistic:                     3.944
Date:                Wed, 06 May 2020   Prob (F-statistic):           0.000505
Time:                        17:01:54   Log-Likelihood:                -849.77
No. Observations:                  73   AIC:                             1720.
Df Residuals:                      63   BIC:                             1742.
Df Model:                           9
Covariance Type:            nonrobust
===============================================================================================================
                                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------
const                                    3.941e+04   3463.957     11.377      0.000    3.25e+04    4.63e+04
medianHouseholdIncome                   -1.238e+04   8316.371     -1.489      0.142    -2.9e+04    4238.903
Pop. Proportion Female                    795.4424   4095.873      0.194      0.847   -7389.506    8980.391
Per Capita Spending on Health            1.364e+04   6924.102      1.969      0.053    -201.168    2.75e+04
% Urban Pop.                             -383.0642   4771.824     -0.080      0.936   -9918.793    9152.665
People 65 and Over (Millions)           -5697.6957   4200.034     -1.357      0.180   -1.41e+04    2695.403
Median Age                               4103.3472   5845.859      0.702      0.485   -7578.669    1.58e+04
International Inbound Tourists(thousands)  1.823e+04   4611.038      3.954      0.000    9017.166    2.74e+04
Import/Export (% of GDP)                -1898.9889   4436.769     -0.428      0.670   -1.08e+04    6967.186
tests per million                       -3214.9875   5248.827     -0.613      0.542   -1.37e+04    7273.957
==============================================================================
Omnibus:                       32.455   Durbin-Watson:                   2.129
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               72.366
Skew:                           1.505   Prob(JB):                     1.93e-16
Kurtosis:                       6.837   Cond. No.                         5.71
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
MSE on test set 1837162570.3024642
```

Appendix 13:

Individual Scatter Plots - Deaths per Million Case vs significant predictors





Individual Scatter Plots - Log(Deaths per Million) vs significant predictors