Case 1: COVID-19

Instructor: Prashant Mittal

Maddie Santora, Maddie Semeraro, Jisoo Moon, Hannah Newbold, and Brianna Rash

This report is the result of a homework assignment for HMP 712: Introduction to Health Analytics– Due Date: October 3, 2020

I. Introduction

Using the data given to the group, analytical tools and JMP software, this report looks at the coronavirus prevalence throughout the world. In this report there are brief summaries and data tables describing the analytical analysis that were performed. Three different groups were created based on population, GDP and BMI to break down the data into smaller sections. The remainder of this report is organized as follows. Section II: Analysis and Results, Section III: Methods, Section IV: Limitations and Section V: Acknowledgements.

II. Analysis and Results

A. COVID-19 Prevalence in Continents

COVID-19 has affected different parts of the world in different ways. Recent studies have shown that some nations were affected more than others. There is also research of differences across continents. The continent with the highest prevalence of cases is North America, with South America following behind. The continent with the lowest prevalence is Oceania. Table I shows the results of the analysis for the case and death prevalence along with the total population of each continent.

B. COVID-19 Prevalence in Individual Countries (based on population size)

Nations were broken down into three groups, “high”, “low” and “medium” based on their population size. In tables II, III and IV it shows the case and death prevalence, the total population, the sum of cases and deaths. The country with the highest prevalence of cases in the low population nation table is Andorra, and the lowest prevalence of cases is Montenegro. The country with the highest prevalence of cases in the medium population nation table is Chile and the lowest is Zambia. The country with the highest prevalence of cases in the high population nation table is the US and the country with the lowest prevalence of cases is China.

C. COVID-19 and Risk Factors

There are often theories describing risks for COVID-19. These theories include higher risks for those with preexisting chronic conditions, respiratory conditions, racial diversity, and differences in socio-economic status. The risk factors that were examined were population, GDP and BMI. As shown below, the most affected and the least affected nations were ranked for each specific risk factor.

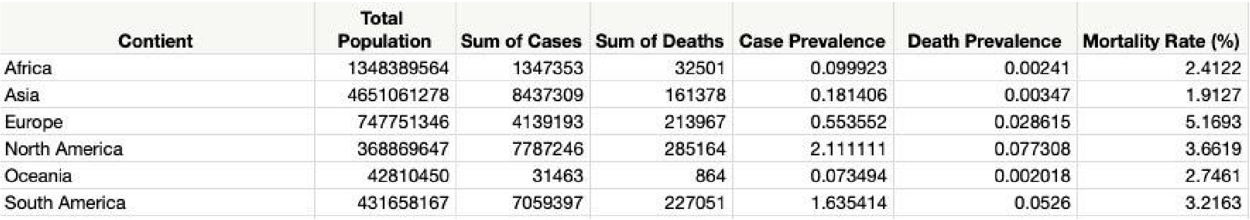
D. COVID-19 Testing Prevalence

There are often arguments around the prevalence of testing, one group against testing and one group for testing. An analysis was conducted, looking at the testing prevalence, checking if testing is responsible or associated with higher or lower case prevalence amongst nations. The countries that were selected to look at the case vs total tests were USA, Columbia, Russia and India. They were chosen due to the impact it had in the countries over the 98th percentile number of cases. This is expanded further on later in the report.

I. COVID-19 Prevalence in Continents

The table below is the data set created using the data given in JMP. Out of the six countries listed, Asia has the highest total population followed by Africa, Europe, South America, North America and Oceania. Asia also had the highest number of cases followed by North America and South America. In the sum of cases column it is shown that although Asia has a higher total population than North America and South America, there is a higher case prevalence in North America and South America. This data shows that higher population does not necessarily mean a higher case prevalence. That is also evident because despite North and South America having a lower population they still have a higher number of cases. So by looking at the case prevalence column, North America has the highest case prevalence and Oceania has the lowest case prevalence. The death prevalence column follows a pattern of a higher case prevalence resulting in a higher death prevalence. North America has the highest death prevalence of 0.077308 followed by South America, Europe, Asia, Africa, and Oceania. The increasing case prevalence matches the increasing death prevalence.

TABLE I

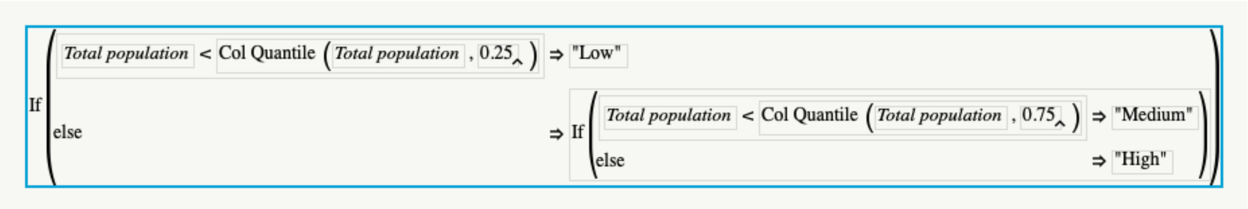


Contient	Total Population	Sum of Cases	Sum of Deaths	Case Prevalence	Death Prevalence	Mortality Rate (%)
Africa	1348389564	1347353	32501	0.099923	0.00241	2.4122
Asia	4651061278	8437309	161378	0.181406	0.00347	1.9127
Europe	747751346	4139193	213967	0.553552	0.028615	5.1693
North America	368869647	7787246	285164	2.111111	0.077308	3.6619
Oceania	42810450	31463	864	0.073494	0.002018	2.7461
South America	431658167	7059397	227051	1.635414	0.0526	3.2163

II. COVID-19 Prevalence in Individual Countries (based on population size)

All the countries in the world were affected by COVID-19, but some more drastically than others. Therefore it is important to analyze the prevalence of cases, deaths, and mortality rates. All the listed countries were identified and separated into three groups. The three groups were based on the sizes of their total populations, GDP per capita, and life expectancy; listed as low, medium, and high. Some countries from the given list that were not included because they were either considered an outlier or lacked the data that was being used for the analysis.

In order to identify the countries into these categories a formula was inputted into JMP in which separated the list of countries by using quantiles. If the country was in the bottom 25% of size it was considered low population, if the country was more than 25% but less than 75% it was medium, and if it was more than 75% it was high. Below is the formula inputted into JMP to manipulate the data into groupings.



```
If( Total population < Col Quantile( Total population , 0.25 ) ⇒ "Low"
    else ⇒ If( Total population < Col Quantile( Total population , 0.75 ) ⇒ "Medium"
               else ⇒ "High" ) )
```

IMAGE I

After separating all the listed countries into the sub categories, individual analyses were performed by calculating case prevalence and death prevalence. Table II shows the “high” group of nations. This group fell above the 75th percentile in regards to total population, GDP per capita, and life expectancy. Case prevalence in this group ranged from 4% to .06% and averaged around 1%. The death prevalence ranged from .088% to .001% and averaged around 0.025%. The highest case prevalence in this group was Bahrain and the highest death prevalence was Belgium.

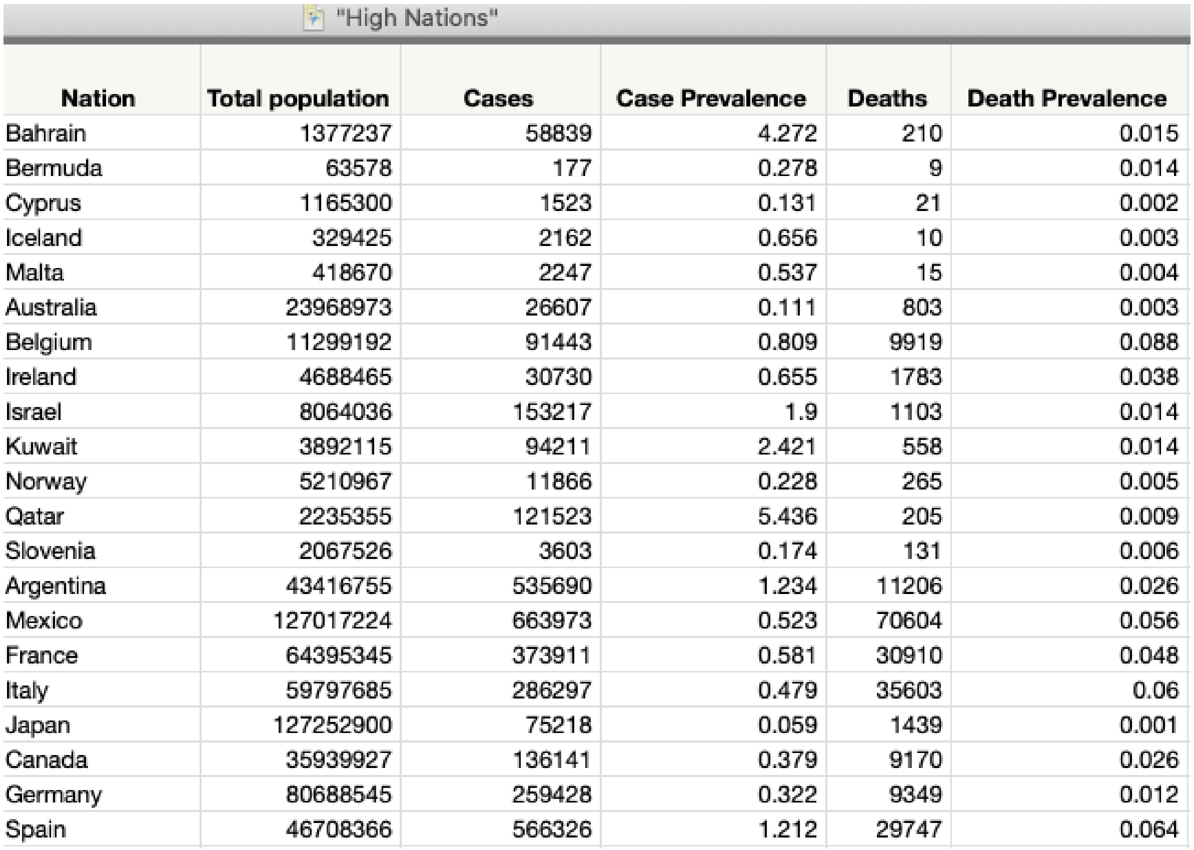
Table III shows the “medium” group. This group fell within the 50th percentile of total population, GDP per capita, and life expectancy. Case prevalence in this group ranged from

.001% to 2.572% and averaged around .49%. The death prevalence ranged from 0% to .0975% and averaged around 0.013%. The highest case prevalence in this group was Panama and the highest death prevalence was Peru.

Table III shows the “low” group. This group fell within the bottom 25th percentile in regards to total population, GDP per capita, and life expectancy. Majority of the nations have a case prevalence below 0.09%, ranging from 0.012% to .209% with an average of .05%. Death prevalence for this group is extremely low, ranging from 0 to .00437% with an average of .001%. This represents that a lower total population, GDP per capita and life expectancy does not increase your chances of contracting and dying from COVID-19.

After analyzing and performing calculations for these three groups of nations, it can be said that there isn’t a huge difference in case and death prevalence to relate it to the three factors we chose (Population, GDP, and life expectancy). Table I and II which represents nations with high and medium nations shows higher prevalence rates, but not high enough to make a correlation of COVID-19 prevalence and the three factors. In conclusion, there is no strong relationship between total population, GDP per capita, and life expectancy with the prevalence of COVID-19 cases and deaths.

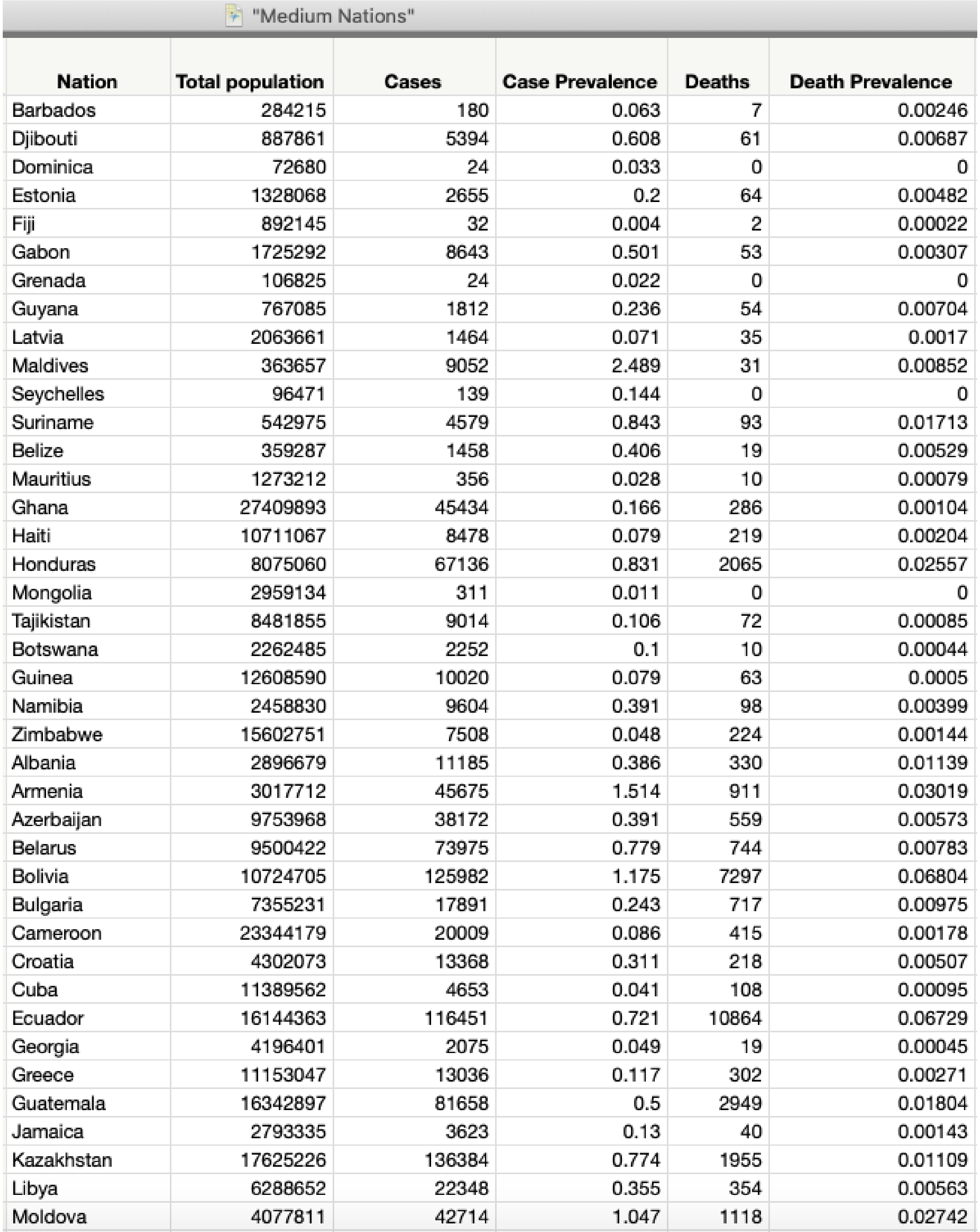
TABLE II



"High Nations"

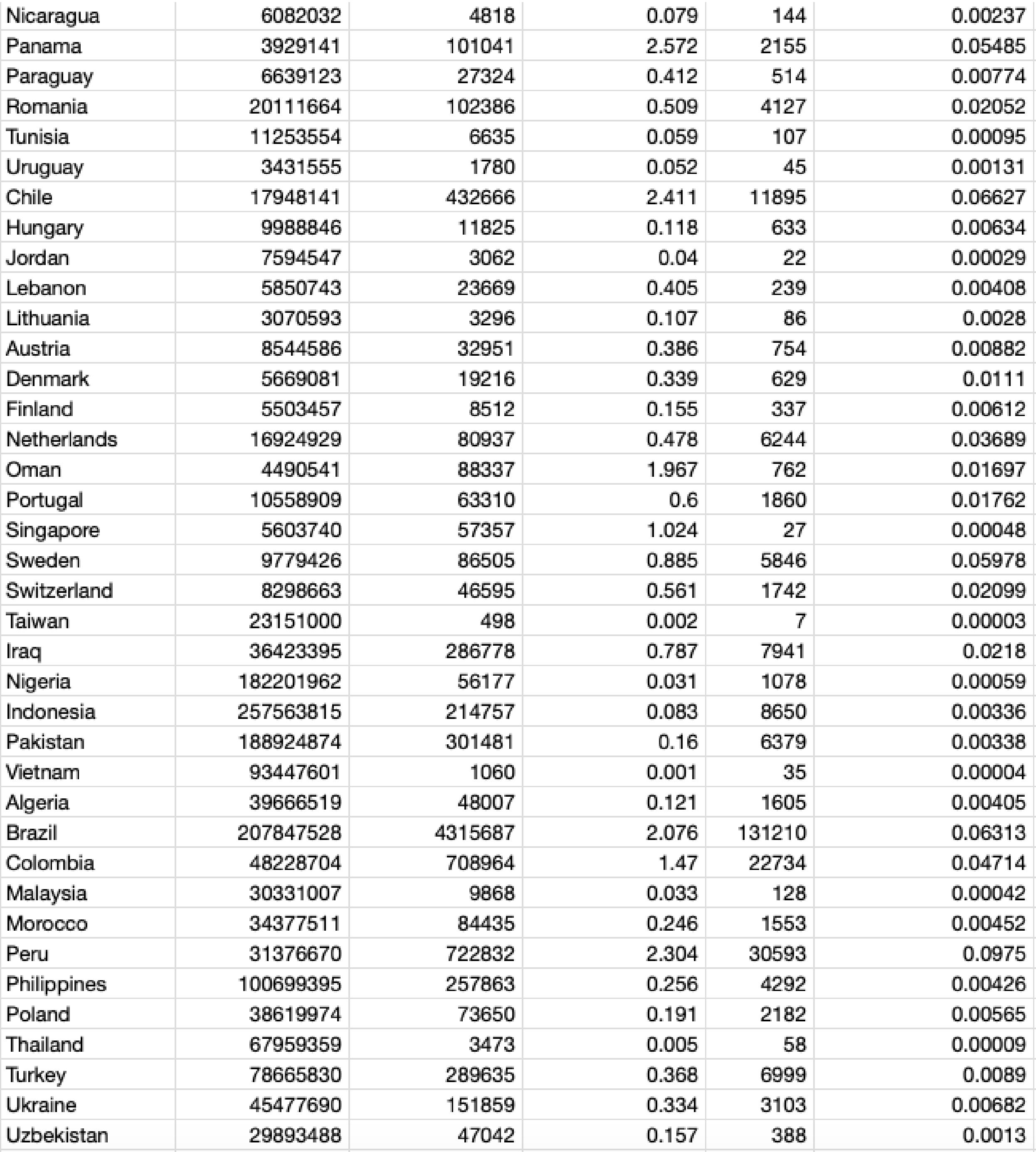
Nation	Total population	Cases	Case Prevalence	Deaths	Death Prevalence
Bahrain	1377237	58839	4.272	210	0.015
Bermuda	63578	177	0.278	9	0.014
Cyprus	1165300	1523	0.131	21	0.002
Iceland	329425	2162	0.656	10	0.003
Malta	418670	2247	0.537	15	0.004
Australia	23968973	26607	0.111	803	0.003
Belgium	11299192	91443	0.809	9919	0.088
Ireland	4688465	30730	0.655	1783	0.038
Israel	8064036	153217	1.9	1103	0.014
Kuwait	3892115	94211	2.421	558	0.014
Norway	5210967	11866	0.228	265	0.005
Qatar	2235355	121523	5.436	205	0.009
Slovenia	2067526	3603	0.174	131	0.006
Argentina	43416755	535690	1.234	11206	0.026
Mexico	127017224	663973	0.523	70604	0.056
France	64395345	373911	0.581	30910	0.048
Italy	59797685	286297	0.479	35603	0.06
Japan	127252900	75218	0.059	1439	0.001
Canada	35939927	136141	0.379	9170	0.026
Germany	80688545	259428	0.322	9349	0.012
Spain	46708366	566326	1.212	29747	0.064

TABLE III



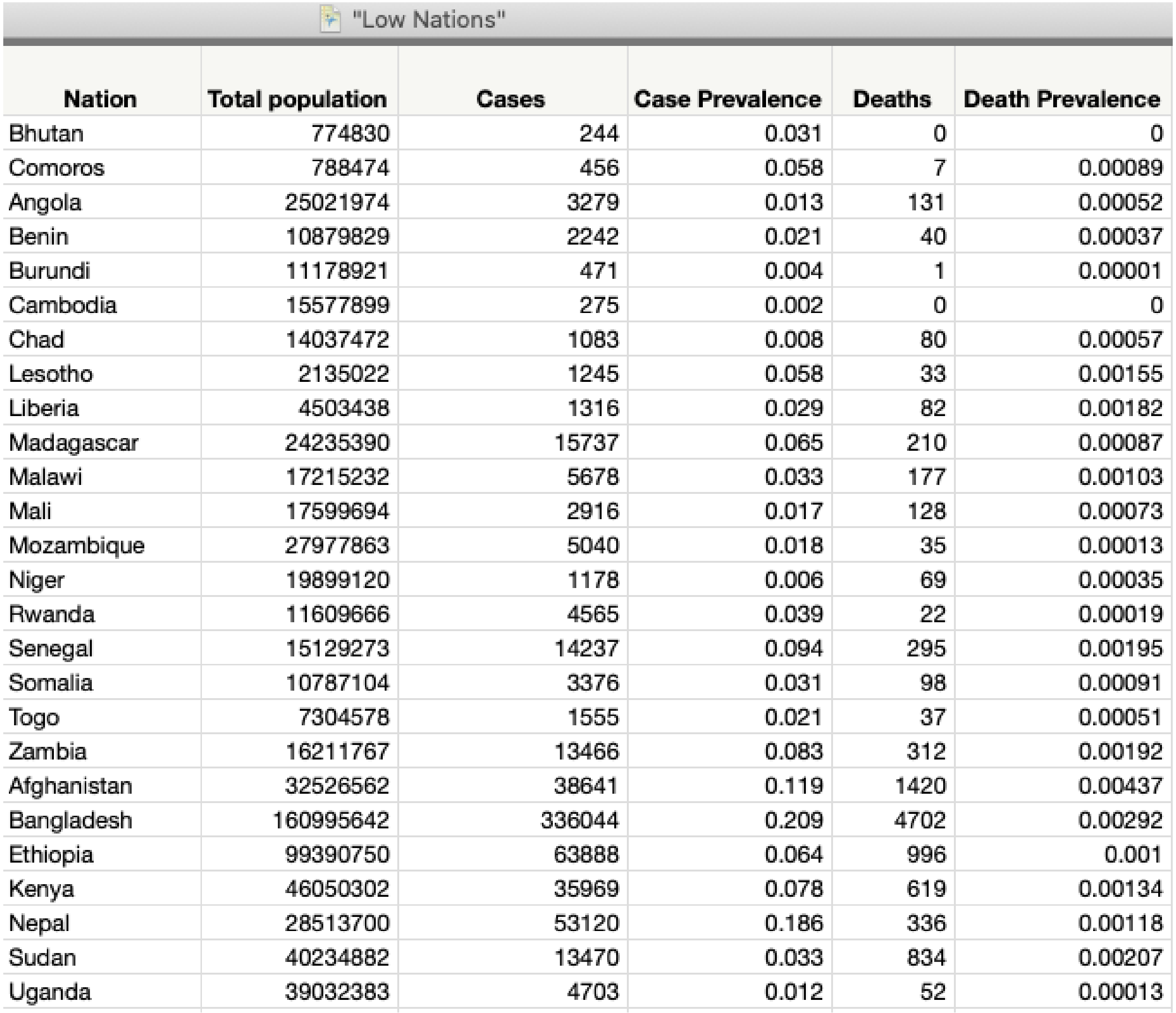
"Medium Nations"

Nation	Total population	Cases	Case Prevalence	Deaths	Death Prevalence
Barbados	284215	180	0.063	7	0.00246
Djibouti	887861	5394	0.608	61	0.00687
Dominica	72680	24	0.033	0	0
Estonia	1328068	2655	0.2	64	0.00482
Fiji	892145	32	0.004	2	0.00022
Gabon	1725292	8643	0.501	53	0.00307
Grenada	106825	24	0.022	0	0
Guyana	767085	1812	0.236	54	0.00704
Latvia	2063661	1464	0.071	35	0.0017
Maldives	363657	9052	2.489	31	0.00852
Seychelles	96471	139	0.144	0	0
Suriname	542975	4579	0.843	93	0.01713
Belize	359287	1458	0.406	19	0.00529
Mauritius	1273212	356	0.028	10	0.00079
Ghana	27409893	45434	0.166	286	0.00104
Haiti	10711067	8478	0.079	219	0.00204
Honduras	8075060	67136	0.831	2065	0.02557
Mongolia	2959134	311	0.011	0	0
Tajikistan	8481855	9014	0.106	72	0.00085
Botswana	2262485	2252	0.1	10	0.00044
Guinea	12608590	10020	0.079	63	0.0005
Namibia	2458830	9604	0.391	98	0.00399
Zimbabwe	15602751	7508	0.048	224	0.00144
Albania	2896679	11185	0.386	330	0.01139
Armenia	3017712	45675	1.514	911	0.03019
Azerbaijan	9753968	38172	0.391	559	0.00573
Belarus	9500422	73975	0.779	744	0.00783
Bolivia	10724705	125982	1.175	7297	0.06804
Bulgaria	7355231	17891	0.243	717	0.00975
Cameroon	23344179	20009	0.086	415	0.00178
Croatia	4302073	13368	0.311	218	0.00507
Cuba	11389562	4653	0.041	108	0.00095
Ecuador	16144363	116451	0.721	10864	0.06729
Georgia	4196401	2075	0.049	19	0.00045
Greece	11153047	13036	0.117	302	0.00271
Guatemala	16342897	81658	0.5	2949	0.01804
Jamaica	2793335	3623	0.13	40	0.00143
Kazakhstan	17625226	136384	0.774	1955	0.01109
Libya	6288652	22348	0.355	354	0.00563
Moldova	4077811	42714	1.047	1118	0.02742



Nicaragua	6082032	4818	0.079	144	0.00237
Panama	3929141	101041	2.572	2155	0.05485
Paraguay	6639123	27324	0.412	514	0.00774
Romania	20111664	102386	0.509	4127	0.02052
Tunisia	11253554	6635	0.059	107	0.00095
Uruguay	3431555	1780	0.052	45	0.00131
Chile	17948141	432666	2.411	11895	0.06627
Hungary	9988846	11825	0.118	633	0.00634
Jordan	7594547	3062	0.04	22	0.00029
Lebanon	5850743	23669	0.405	239	0.00408
Lithuania	3070593	3296	0.107	86	0.0028
Austria	8544586	32951	0.386	754	0.00882
Denmark	5669081	19216	0.339	629	0.0111
Finland	5503457	8512	0.155	337	0.00612
Netherlands	16924929	80937	0.478	6244	0.03689
Oman	4490541	88337	1.967	762	0.01697
Portugal	10558909	63310	0.6	1860	0.01762
Singapore	5603740	57357	1.024	27	0.00048
Sweden	9779426	86505	0.885	5846	0.05978
Switzerland	8298663	46595	0.561	1742	0.02099
Taiwan	23151000	498	0.002	7	0.00003
Iraq	36423395	286778	0.787	7941	0.0218
Nigeria	182201962	56177	0.031	1078	0.00059
Indonesia	257563815	214757	0.083	8650	0.00336
Pakistan	188924874	301481	0.16	6379	0.00338
Vietnam	93447601	1060	0.001	35	0.00004
Algeria	39666519	48007	0.121	1605	0.00405
Brazil	207847528	4315687	2.076	131210	0.06313
Colombia	48228704	708964	1.47	22734	0.04714
Malaysia	30331007	9868	0.033	128	0.00042
Morocco	34377511	84435	0.246	1553	0.00452
Peru	31376670	722832	2.304	30593	0.0975
Philippines	100699395	257863	0.256	4292	0.00426
Poland	38619974	73650	0.191	2182	0.00565
Thailand	67959359	3473	0.005	58	0.00009
Turkey	78665830	289635	0.368	6999	0.0089
Ukraine	45477690	151859	0.334	3103	0.00682
Uzbekistan	29893488	47042	0.157	388	0.0013

TABLE IV


"Low Nations"

Nation	Total population	Cases	Case Prevalence	Deaths	Death Prevalence
Bhutan	774830	244	0.031	0	0
Comoros	788474	456	0.058	7	0.00089
Angola	25021974	3279	0.013	131	0.00052
Benin	10879829	2242	0.021	40	0.00037
Burundi	11178921	471	0.004	1	0.00001
Cambodia	15577899	275	0.002	0	0
Chad	14037472	1083	0.008	80	0.00057
Lesotho	2135022	1245	0.058	33	0.00155
Liberia	4503438	1316	0.029	82	0.00182
Madagascar	24235390	15737	0.065	210	0.00087
Malawi	17215232	5678	0.033	177	0.00103
Mali	17599694	2916	0.017	128	0.00073
Mozambique	27977863	5040	0.018	35	0.00013
Niger	19899120	1178	0.006	69	0.00035
Rwanda	11609666	4565	0.039	22	0.00019
Senegal	15129273	14237	0.094	295	0.00195
Somalia	10787104	3376	0.031	98	0.00091
Togo	7304578	1555	0.021	37	0.00051
Zambia	16211767	13466	0.083	312	0.00192
Afghanistan	32526562	38641	0.119	1420	0.00437
Bangladesh	160995642	336044	0.209	4702	0.00292
Ethiopia	99390750	63888	0.064	996	0.001
Kenya	46050302	35969	0.078	619	0.00134
Nepal	28513700	53120	0.186	336	0.00118
Sudan	40234882	13470	0.033	834	0.00207
Uganda	39032383	4703	0.012	52	0.00013

III. COVID-19 and Risk Factors

It is theorized that certain pre-existing conditions put individuals at higher risk when it comes to COVID-19. Conditions such as diabetes, asthma, COPD, as well as demographics such as race, and socioeconomic status are thought to impact case prevalence of a nation. The objective here is to look at certain health factors of a nation: Infectious TB detection rate, estimated HIV prevalence, and Lung Male Mortality to see if that has an impact on the COVID-19 prevalence of that nation. The analysis was performed three times on the three best and worst nations based on population size, GDP, and average BMI. The three best and three worst nations in terms of these three categories were included in the observation. Infectious TB detection rate, estimated HIV prevalence, and lung male mortality were selected because they impact the respiratory system, and it is proposed that diseases that impact the respiratory system increase COVID-19 risk. The data of these risk factors is from the given data set, and the

COVID-19 prevalence was calculated by dividing the total number of cases by the nation's population. The following tables represent the findings of the analysis:

TABLE V

Group A Based on Population					
Nation		**COVID-19 Prevalence**	**Infectious TB, detection rate (%)**	**Estimated HIV Prevalence % - (Ages 15-49)**	**Lung Male Mortality (per 100,000)**
Best #1	China	0.0066%	80	0.06	41.7
Best #2	India	0.3626%	68	0.34	7.9
Best #3	United States	2.0157%	N/A	0.6	48.49
Worst #1	Iceland	0.6563%	85	0.3	32.93
Worst #2	Belize	0.4058%	111	2.4	19.7
Worst #3	Maldives	2.4892%	94	0.06	N/A

TABLE VI

Group B Based on GDP					
Nation		**COVID-19 Prevalence**	**Infectious TB, detection rate (%)**	**Estimated HIV Prevalence % - (Ages 15-49)**	**Lung Male Mortality (per 100,000)**
Best #1	Luxembourg	1.2624%	119	0.3	53.29
Best #2	Qatar	5.4364%	52	0.06	18.5
Best #3	United States	2.0157%	N/A	0.6	48.49
Worst #1	Liberia	0.0292%	69	1.8	1.6
Worst #2	Cambodia	0.0018%	68	0.6	20.9
Worst #3	Afghanistan	0.1188%	64	0.06	11.3

TABLE VII

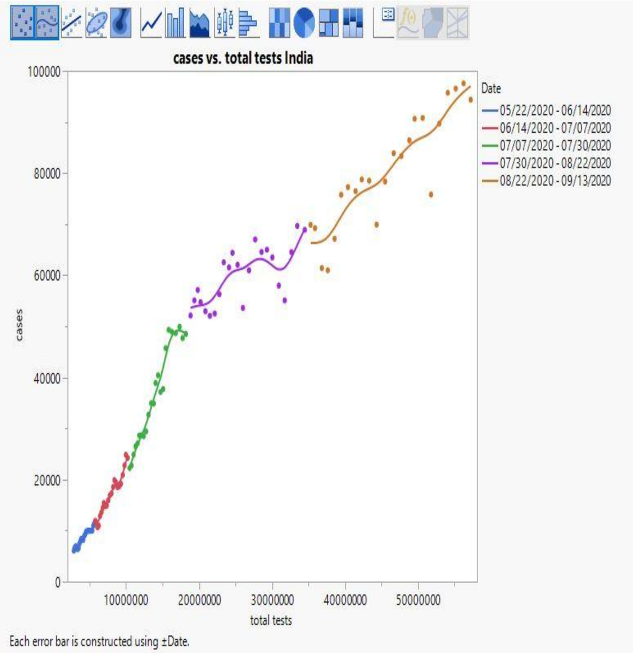
Group C Based on BMI					
Nation		**COVID-19 Prevalence**	**Infectious TB, detection rate (%)**	**Estimated HIV Prevalence % - (Ages 15-49)**	**Lung Male Mortality (per 100,000)**
Best #1	Ethiopia	0.0643%	31	1.7	2.6
Best #2	Bangladesh	0.2087%	66	0.06	20.8
Best #3	Afghanistan	0.1188%	64	0.06	11.3
Worst #1	New Zealand	0.0319%	65	0.1	33.42
Worst #2	Malta	0.5367%	74	0.1	41.35
Worst #3	Ireland	0.6554%	64	0.3	39.79

When looking at the risk factors in relation to prevalence based on population size, “best” is considered largest populations and “worst” is considered smallest populations. When looking at GDP, “best” are nations with the highest GDP and “worst” are the ones with lowest GDP. In the analysis for BMI, “best” are the nations with lowest BMIs and “worst” are the nations with highest BMIs.

It is difficult to detect a relationship between these risk factors and COVID-19 prevalence. It would be assumed that those nations with a low male lung mortality rate would have a lower rate of COVID-19, but based on these tables above, that is not true for all nations. For instance, Qatar has a relatively low male lung mortality rate compared to the other nations, being 18.5. However, it has the highest COVID-19 prevalence out of the nations in this analysis at 5.44%. There also does not appear to be a strong correlation between HIV prevalence of a nation and COVID-19 prevalence. Qatar only has an HIV prevalence of 0.06, yet it has the highest COVID-19 prevalence of the nation observed. Similarly, Ethiopia has the highest HIV prevalence of the nations analyzed at 1.7%, yet it has one of the lowest COVID-19 prevalence being only 0.06%. Following this trend, there is no clear relationship between TB detection rates and COVID-19 prevalence.

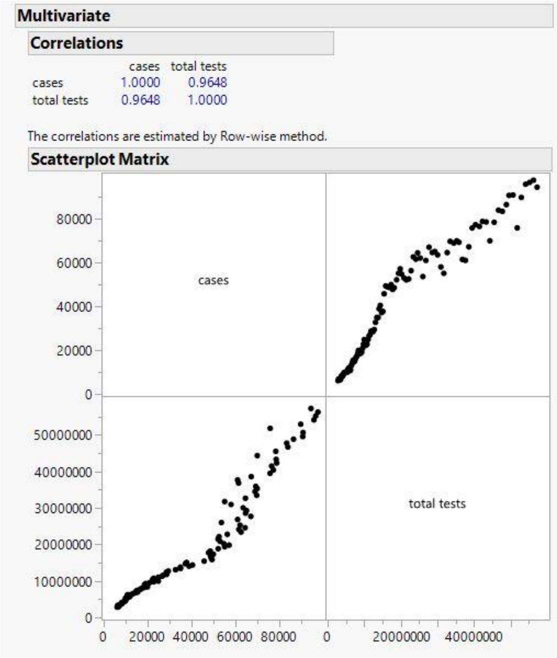
III. COVID-19 Testing Prevalence

TABLE VIII



Each error bar is constructed using ±Date.

TABLE IX



Multivariate

Correlations

	cases	total tests
cases	1.0000	0.9648
total tests	0.9648	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

TABLE X

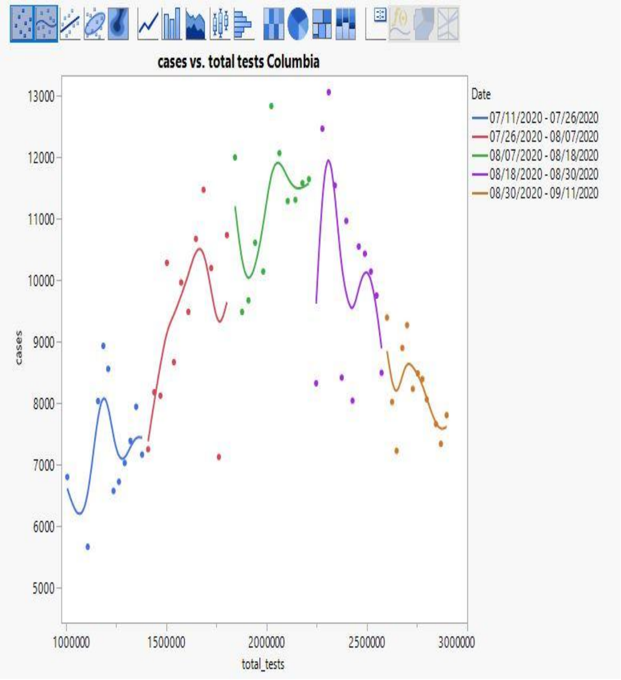
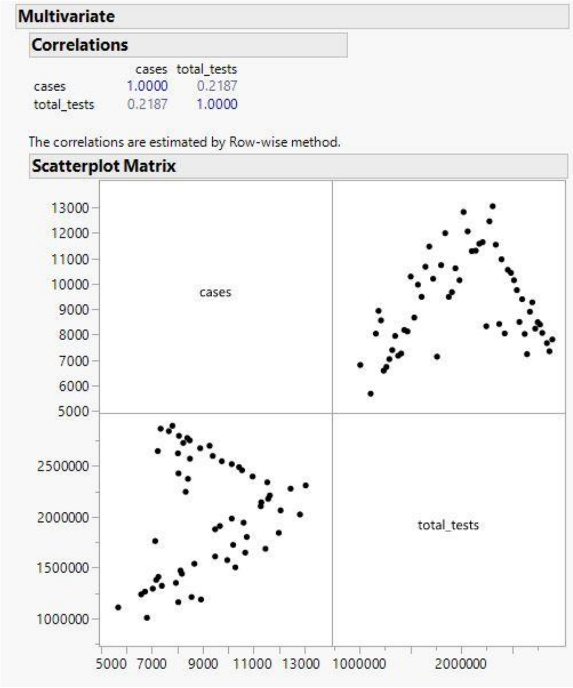


TABLE XI



Multivariate

Correlations

	cases	total_tests
cases	1.0000	0.2187
total_tests	0.2187	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

TABLE XII

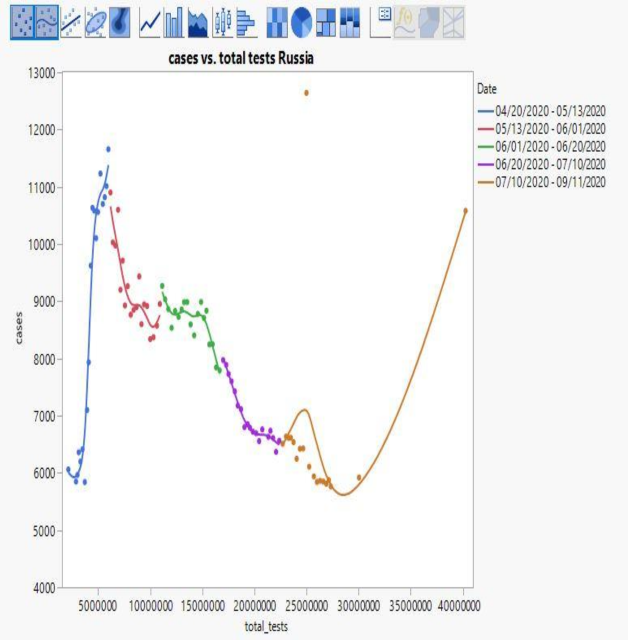
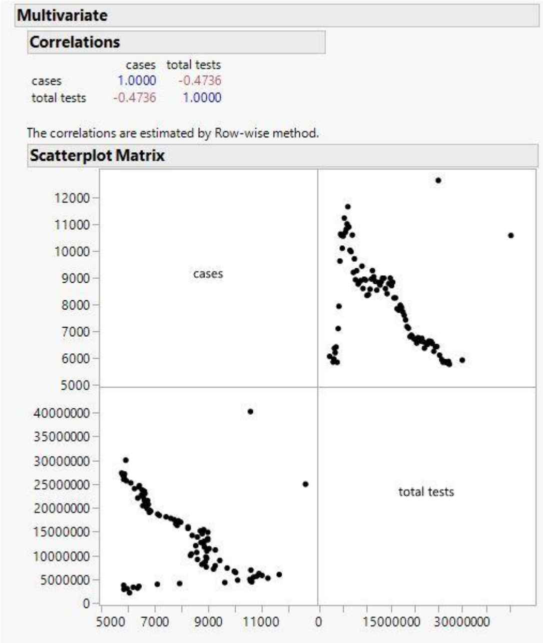


TABLE XIII



Multivariate

Correlations

	cases	total tests
cases	1.0000	-0.4736
total tests	-0.4736	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

TABLE XIV

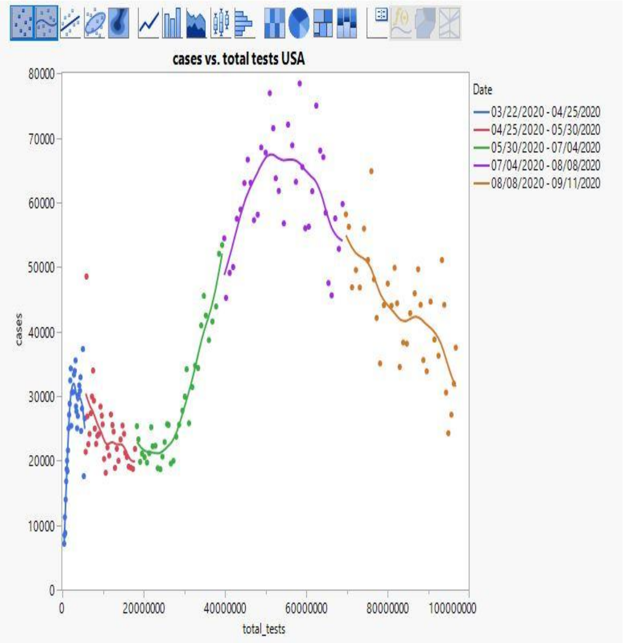
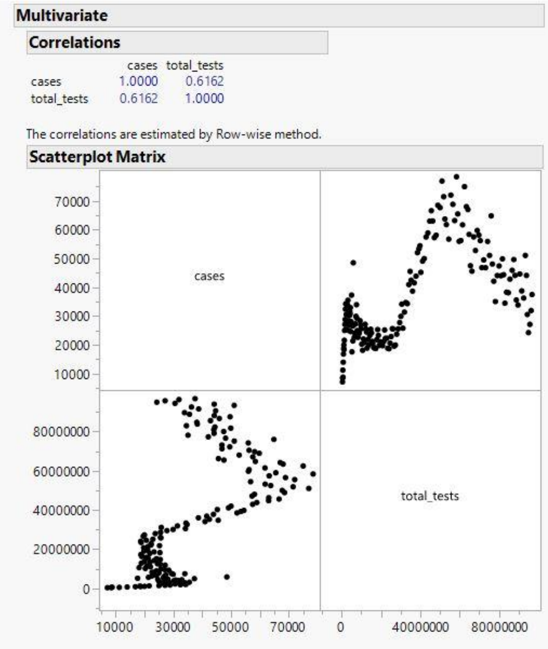


TABLE XV



Multivariate

Correlations

	cases	total_tests
cases	1.0000	0.6162
total_tests	0.6162	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

TABLE VIII is a scatter-plot of the relationship between cases and tests in India from 5/22-9/13. The graph presents a positive linear relationship between the cases and the test. A multivariate test (TABLE IX) was done to find a correlation of 96% between cases and total test. In this case, there is an association between cases and total tests for India.

TABLE X represents the association of cases and tests in Columbia from 7/11-9/11. It is shown in the visual that there is a little bit of correlation in the relationship but it is not a strong one. The scatter-plot and line does not match up as much as it did like India. A multivariate test (TABLE XI) presented to the group that the correlation between the cases and tests were barely where with 22%. For Columbia, there was not a strong case for any association between cases and total tests.

TABLE XII displays the relationship between cases and total tests from 4/20-9/11 in Russia. The graph does not seem to have any linear relationship between the cases and total tests. There are multiple outliers that affect the data set and it is shown in the multivariate table in TABLE XIII. The correlation coefficient is a -47%, representing no relationship between the cases and the tests. This means that as the number of tests increased, there was a decrease of cases.

Lastly, TABLE XIV represents the scatter-plot relationship between cases and total tests for the USA from 3/22-9/11. The graph does not represent a positive relationship between the number of cases and the number of tests. The visual displays that although the number of tests went up, the number of cases fluctuated and that was evident in the multivariate graph in TABLE XV. It only showed a 62% correlation coefficient which is not strong enough to suggest that there were any connection between the cases vs tests.

The analysis of the four countries represented above mostly represents a positive relationship between the cases and tests. However, it does not have any evidence that numbers of tests are the sole reason for the number of cases. This was the case for India, but it was not as strong in Columbia, Russia, and the United States. Three out of four graphs showed the fluctuation of cases as the tests increased.

III. Methods

For question one, using the “Nations Continents” and “Cases and deaths international” file, the nation’s names had to be recoded so that the names matched up when they were joined together. With the new file, the tabulate option was used to create a new data set that grouped the countries into their respective continents. The now calculated total sum of cases, sum of deaths, and total population columns grouped by continent were used to create the cases/pop ratio and death/pop ratio. The mortality rate was then created using the sum of cases divided by sum of deaths then multiplied by 100% and rounded to the ten thousandths place.

For question two, data was pulled from the “Combined Nations” file. Outliers were removed by using the “analyze” multivariate methods function to remove countries that were far away from the average line. From there the grouping was done using the formula from ***image 1***. From there, three separate files were made separating the low, medium, and high based off of total population, GDP per capita, and life expectancy. Having separated the groups, total populations from “Combined Nations” was imputed into the data set. From there, a summary of the cases and deaths from “Covid-19 geographic distribution” was calculated using JMP functions. Once put into three groups, using the population, cases, and death calculations for case and death prevalence were done. First case prevalence was calculated by dividing the number of cases by the size of the population. The death prevalence was calculated by dividing the number of deaths by the total population. Each number was rounded to the third and sixth decimal place. From there the numbers were analyzed and compared for trends in the categories.

For question three, data was pulled from the "Combined Nations" file. First, the data for population, GDP, and BMI was organized by descending order, and the top three and bottom three of each category that had the adequate information needed was pulled. They were then categorized as "Best #1, Best #2 … Worst #1, Worst #2" and so on. From there, the information that was going into the analysis was selected to be "infectious TB detection rate", Estimated HIV Prevalence", and "Lung Male Mortality". The data for those categories from the three best and worst nations was pulled and put into a table. Three separate tables were used depending on whether the nations were organized based on population, GDP, or BMI. The COVID-19 prevalence rate for each nation in the analysis was calculated by using JMP to find the sum of the total number of cases for each nation, and dividing that by the total population of the nation, and turning that into a percentage. From there, the numbers were analyzed and compared.

For question four, distribution of the cases were considered to pick out which countries were most affected by COVID-19. The group determined that the countries in the 98th percentile were the most affected and India, Columbia, Russia, and the United States was chosen. These four countries were categorized by "high" cases and were made subsets with multiple different dates on the dataset. The "high" cases were formulated by considering the 98% with the "if" statement function. The dates were sorted and coded to match up to the dates of the other files and later joined to create one file to make the graphs and the multivariate datas.

IV. Limitations

While performing the analyses with the provided data regarding covid-19 a lot of discrepancies were identified in which additionally research was conducted to accurately perform analyses. This limited the provided data that could be utilized, therefore the group had to research was done to fill in the missing data points in the provided jmp datasets. In certain data sets, the list of countries differed and were not the same in each data set. Therefore it limited the countries that could be individually analysed for problem two, three, and four. Prior to this lab, the researchers all had little to no experience with jmp. The confidence of utilizing the system was extremely low therefore it took a lot of trial and error when going through the problems. The JMP system is very technical when it comes to how data is put in. One of the original data sets, countries that had more than one name had underscores in them which threw off our numbers when trying to combine charts. The original data the group could pull from was limited because of the spelling and missing countries. Furthermore, the researchers had a hard time finding common times in their schedules to meet and were limited on the time when they could all collaborate together.

V. Acknowledgment

The Author of this paper wishes to acknowledge Maddie Santora, Maddie Semeraro, Jisoo Moon, Hannah Newbold, and Brianna Rash for their collaboration in this report. The division of labor had the following distribution:

Introduction: Brianna Rash

Analysis and Results A-D: Brianna Rash

Analysis and Results Section I: Jisoo Moon

JMP Data Analytical Analysis: Jisoo Moon and Maddie Santora

Analysis and Results Section II: Maddie Santora

JMP Data Analytical Analysis: Maddie Santora and Maddie Semeraro

Analysis and Results Section III: Maddie Semeraro

JMP Data Analytical Analysis: Maddie Semeraro

Analysis and Results Section IV: Jisoo Moon

JMP Data Analytical Analysis: Hannah Newbold, Maddie Santora, and Jisoo Moon

Methods: Jisoo Moon, Maddie Santora, Maddie Semeraro, and Hannah Newbold

Limitations: Maddie Santora