The Right Conditions for COVID-19

Executive Summary

Finding a cure or vaccine for the COVID-19 virus, which has infected more than 13 million people worldwide with over 1.5 million deaths, is arguably the most important medical event in our lifetime. The virus seems to prey on those less fortunate in society – elderly, people in certain blue-collar jobs, and people of color; however, in the United States, one of the wealthiest countries in the world, we have the highest number of cases and deaths. Our paper focuses on sixteen different universal conditions that could contribute to a person being infected with COVID-19. Those conditions are the stringency index, population, population density, portion of the population over age 65, GDP per capita, extreme poverty, cardiovascular death rate, diabetes prevalence, smoker or not, number of hospitals, life expectancy and human development index. Our findings indicate that deaths from the virus are mainly dependent on the number of cases. The number of new cases is based on the size of the population and the stringency level of the country. In many situations, the higher the stringency index, then the lower number of cases. With the variations in the dependent variables, the results from the response variable are questionable. This is noted when reviewing the results from clustering and the best model for predicting the number of new cases. The results from clustering were limited. The program simply clustered the groups according to the index. This information was not informative. One of the best methods visually was the growth factor. From it, we are seeing a decline in the number of cases and deaths worldwide. We are seeing individual countries responding by either increasing or decreasing the stringency index. We forecasted by November 26 in the United States, there were 248,738 deaths and 10,386,105 cases. Due to the ongoing surge, we underestimated the number of cases by 2.5 million. Lastly, we determined the best model for predicting the test data was Neutral Net (with Tour 20) which was able to predict approximately 70% of the test data.

1 Introduction

We are completing this course to demonstrate our ability to become an independent business analyst. The report on the Covid19 virus has allowed us to use our descriptive, prescriptive, and predictive skills to analyze the dataset. It allowed us to build our confidence and determine areas we may need to focus on

to become proficient. In addition, this program allowed us to build our confidence and determine areas we may need to focus on to become proficient. During the process, we used Microsoft Excel, R programming language, JPM Pro15, and Tableau. In the future, we plan to learn the Python programming language, increase proficiency in Tableau, and consider getting a master's degree in Statistics or taking more senior level courses.

There are many events that have shaped the Year 2020. North Carolina experienced a 7.0 magnitude earthquake which originated in Virginia, mega fires are consuming one-million acres in California and Colorado, President Trump faced the possibility of impeachment, and the list goes on and on. The one event that will shape the world more than anything else is the global pandemic caused by a new coronavirus called SARS-CoV-2, more commonly known as COVID-19. It is spread through droplets that are created when a person coughs, talks, sneezes, sings, or breathes, which makes COVID-19 easily transmittable. This virus made headlines on December 31, 2019, when health officials in Wuhan, China reported groups of pneumonia cases with an unknown cause. Ten months later, according to John Hopkins' university dashboard, as of October 23, over 41.9 million people have been infected worldwide with over 1.1 million deaths. As of December 6, over 66 million people have been infected worldwide with more than 1.5 million deaths. In the United States, there are over 14.6 million people infected with 281,347 deaths.

1.1 Problem Motivation and Context

Since the beginning of the pandemic, the question has been which people are more likely to be infected by the virus. In the United States, it has been reported that Asian people, people of color, people with underlying conditions like diabetes and heart disease, people over 65, children, and now young people under 35. Presently, anyone can be infected by the virus. We plan to gain insight into which conditions will increase the chances one could be infected with the COVID-19 virus and build a model to predicts which conditions will cause a person to get COVID-19. We will be analyzing whether the conditions in each country is likely to increase the chances that someone would be infected with the virus.

1.2 Research Questions and Contributions

After studying the COVID-19 dataset, we created a problem statement which addressed why a person is infected with Covid-19. To address this question, we used scatter charts to visualize the relationship between new death (new cases) to our conditions. We used the R programming environment to further determine if there is a correlated relationship, determine which variables are significant, complete one-sample hypothesis testing and ANOVA testing. In addition, we used JMP Pro15 to create models to determine which model best predicts the test data. Lastly, we created a model to forecast the number of cases and deaths that will happen in the future.

2 The Dataset

The data set contains COVID-19 worldwide data as of October 18, 2020. Using the pivot table feature in Excel to explore the data, it contains 50,350 observations with 41 variables. The data collection begins on December 31, 2019 and ends on October 18, 2020. It has data on 210 countries and the city of Hong Kong which are appropriately placed in six continents – Africa, Asia, Europe, North America, Oceania, and South America. Antarctica currently does not have any infections from COVID-19. There are 39,772,036 cases worldwide with 1,110,863 deaths from the virus. The Asian region leads in the number of cases with 12,472,669; however, the North American continent leads in the number of deaths with 329,830.

Variable	Description	Type of Data	Type of Scale
iso_code	ISO 3166-1 alpha-3 – three-letter country codes	Character	Categorical
continent	Continent of the geographical location	Character	Categorical
location	Geographical location	Character	Categorical
date	Date of observation	Character (Date)	Interval
total_cases	Total confirmed cases of COVID-19	Double	Ratio

Table 1 Variables in the Dataset

Table 1 - A portion of the 41 variables, including their descriptions and original data types, which were provided as part of the project. Note under “Type of Data” we added the new conversion which is noted in parenthesis. data type. We have added the type of scale for the dataset. To view the entire dataset, please see Appendix A

2.1 Visualization Through Exploring Data: *Entire Dataset – All 210 countries*

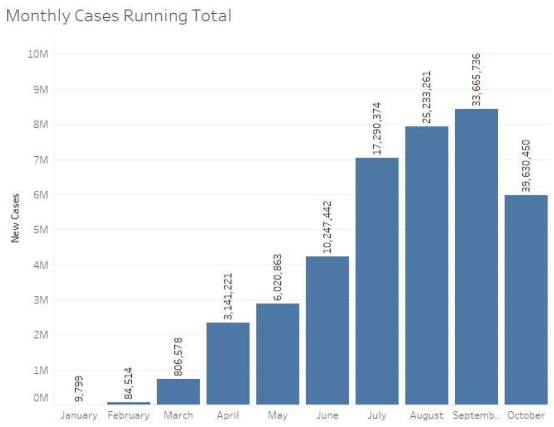


Figure 1 Monthly Cases Running Total

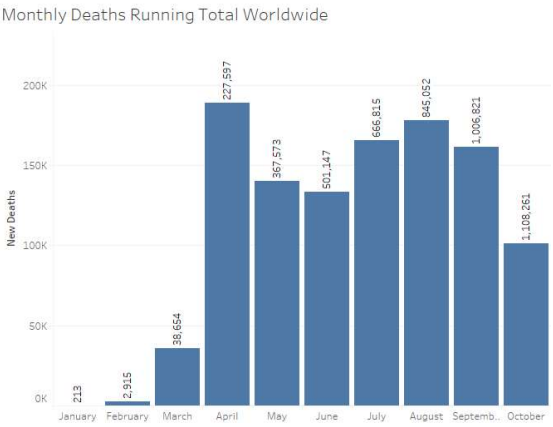


Figure 2 Monthly Deaths Running Total Worldwide

Figure 1 – Monthly running totals for new cases. As of October 18, 2020, there were 39,630,450 cases.

Figure 2 – Monthly running totals for new deaths. As of October 18, 2020, there were 1,108,261 deaths.

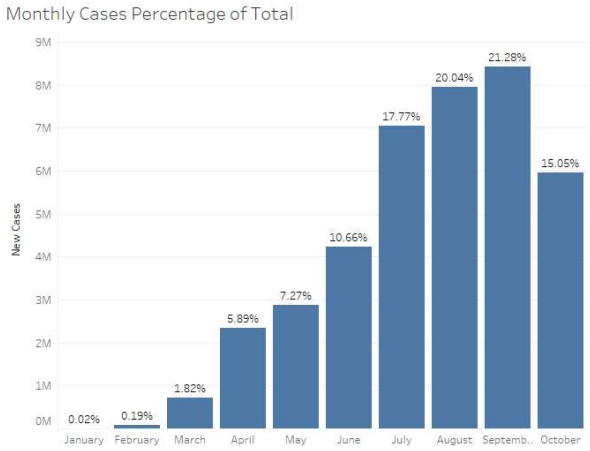


Figure 3 Monthly Cases Percentage of Total

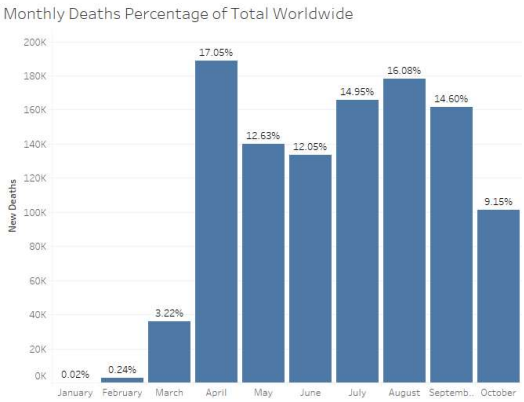


Figure 4 Monthly Deaths Percentage of Total Worldwide

Figure 3 – Monthly percentage of totals for new cases. The largest percentage of cases occurred in September with 21.28%.

Figure 4 – Monthly percentage of totals for new deaths. The largest percentage of cases occurred in April with 17.05%.

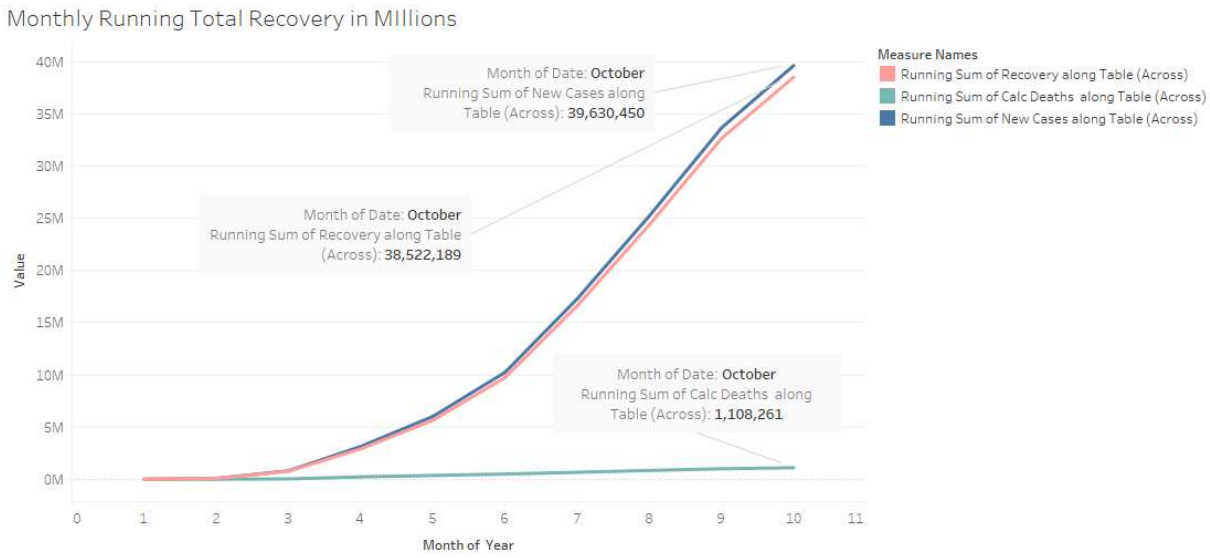


Figure 5 Monthly Running Total Recovery

Figure 5 – Chart provides the running totals for cases, deaths, and recovery. There were 39,630,450 cases as of October 18 with 38,522,189 people who had recovered from the virus.

As of October 18, 2020

Continent	Population 2020	Avg. Stringency Index 2020	New Cases 2020	New Deaths 2020
Africa	319,342,472,937	61	1,628,908	39,407
Asia	1,307,581,932,947	57	12,398,464	222,014
Europe	208,400,135,770	47	6,805,360	237,285
North America	166,167,027,744	62	9,809,929	329,665
Oceania	11,024,771,980	53	33,146	1,004
South America	113,683,577,930	71	8,954,643	278,886

Table 2 Overall view of dataset

Table 2 – Overall view of dataset by Region. Provides immediately information on population, stringency index, and cases. A higher average stringency does not indicate a lower number of deaths or new cases.

The table above demonstrates the role the stringency index plays in managing the number of new cases and deaths. The continent of Africa has average index of 61 with 39,407 deaths. South America has an index of 71 with even lower cases. At the same time North American has an average index of 62 with the 329,665 deaths. Based on this information, it is difficult to determine if business and schools being closed contributes to a reduction on new cases on a continent level. We switched from looking at continents to looking at individual countries.

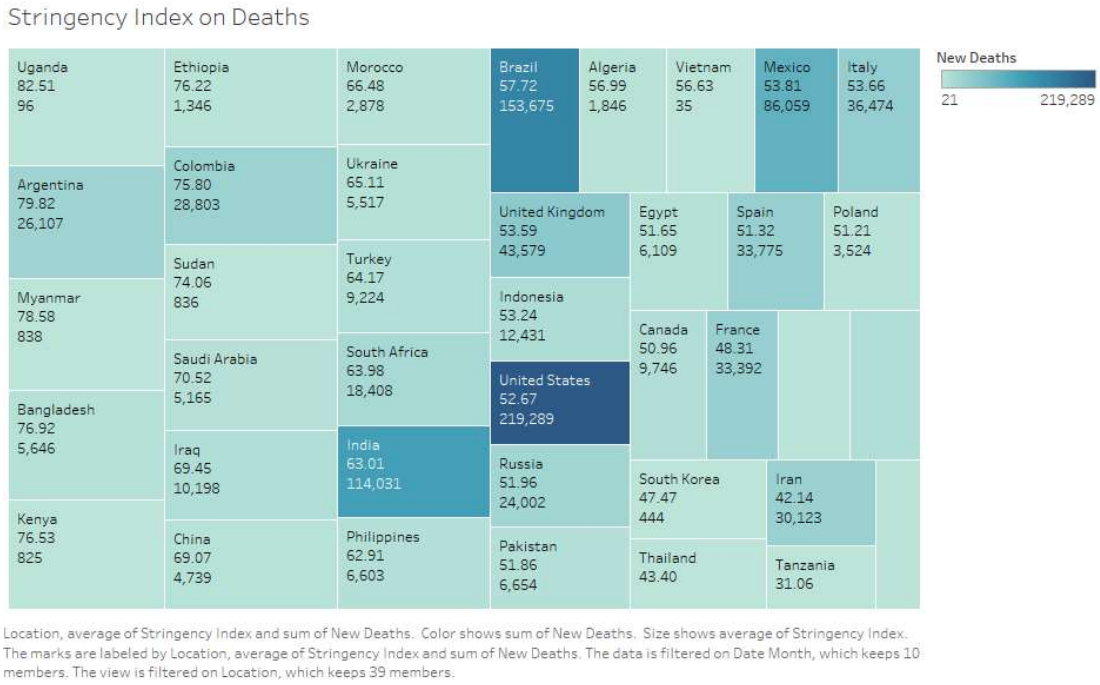


Figure 6 Strringency index on deaths

Figure 6 – Heatmap displays number of deaths and the stringeny index for the country. The darker the shade of blue to higher number of deaths due to COVID-19. Note, The United States which is part of North America has an average index of 52.67 and accounts for 62% of deaths and 83% of new cases.

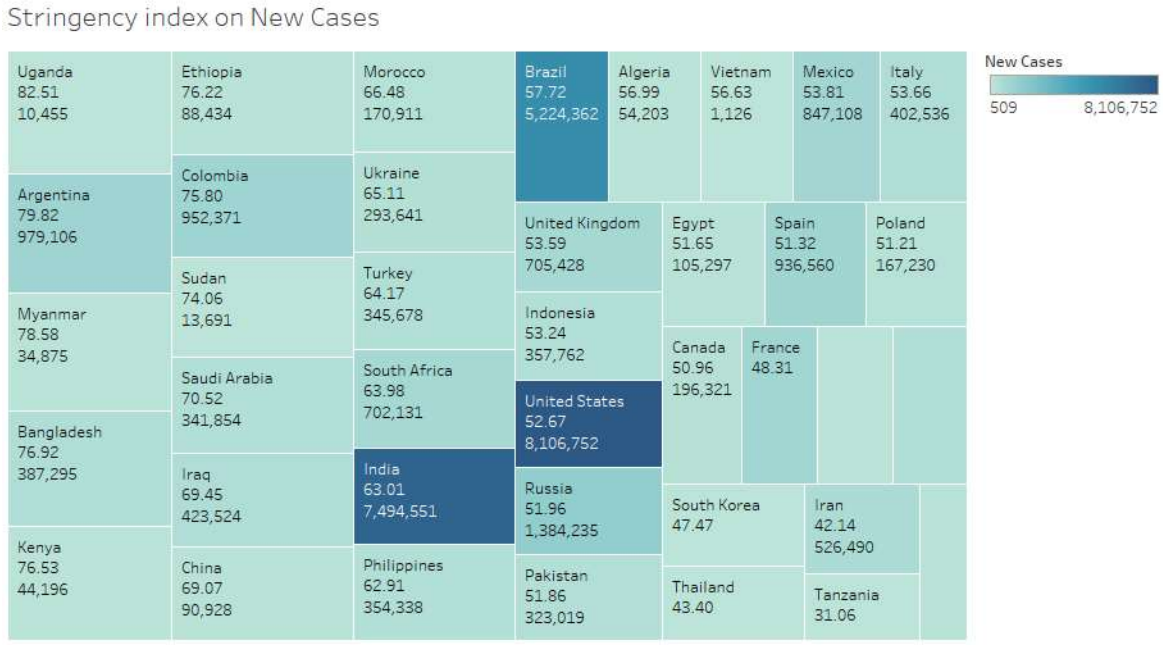


Figure 7 Strringency index on New Cases

Figure 7 – Heatmap displays number of deaths and the stringeny index for the country. The darker the shade of blue to higher number of deaths due to COVID-19.

2.3 – Cleaning the Data

The COVID-19 dataset contains a larger number of NAs or missing data. 37 variables out of the 41 variables were missing some type of data.

aged_65_older	median_age	new_tests_smoothed_per_thousand
aged_70_older	new_cases	population_density
cardiovasc_death_rate	new_cases_smoothed	positive_rate
diabetes_prevalence	new_cases_per_million	stringency_index
extreme_poverty	new_cases_smoothed_per_million	tests_per_case
female_smokers	new_deaths	tests_units
gdp_per_capita	new_deaths_per_million	total_cases
handwashing_facilities	new_deaths_smoothed	total_deaths
Hospital_beds_per_thousands	new_deaths_smoothed_per_million	total_cases-per_million
human_development_index	new_tests	total_deaths_per_million
life_expectancy	new_tests_per_thousand	total_tests
male_smokers	new_tests_smoothed	total_tests_per_thousand
		New_deaths_smoothed_per_million

Table 3 Variables Missing Data

Table 3 contains the variables identified as having NA or missing data by using the R programming environment.

We decided the best way to clean up the data was to impute the dataset of missing data. The options available are substituting missing data with zero, average value of the column, duplicating a value above or below, or removing the data.

The *Date* field is currently set up as a Character, which was converted to a Date format. It was converted using R programming environment as the dataset was imported in.

The following variables were in the wrong data type: *new_tests, new_tests_per_thousand, new_tests_smoothed, new_tests_smoothed_per_thousand, positive_rate, tests_per_case, total_tests, and total_tests_per_thousand.* These variables were listed as logical; however, a numerical value is stored rather than True or False. These variables were converted to Double using R programming environment as the data was brought in. The field also has missing data which does not exist. We cannot assume the missing data is 0, because it may not be accurate. These fields were not be used in analyzing the dataset for any country that has missing data in this field.

The following variables were missing data which was not recorded*: new_cases, new_cases_smoothed, new_cases_per_million, new_cases_smoothed_per_million, new_deaths, new_deaths_per_million, new_deaths_smoothed, new_deaths_smoothed_per_million, stringency_index, total_cases, total_deaths, total_cases-per_million, total_deaths_per_million, and New_deaths_smoothed_per_ million*. For example, prior to the first incident of COVID-19 in a country or region, the number was not recorded. For those variables, the NAs was replaced with 0. We used the R function *replace_na* to replace these variables.

Several variables were missing data which was replaced with the average of the data (or the data provided before the missing data). The missing data will be the same as the data in the country's field. Those variables are *age 65 older, age 70 older, cardio death data, diabetic prevalence, extreme poverty, gdp per capita, life expectancy, median age, and population density*. Also, when we reviewed the data fields, some data collected was from the past. So, it is expected that it would not change. Any countries or regions with missing data were excluded from the project.

Several variables were missing data which does not exist. Those variables are *female_smokers, handwashing_facilities, Hospital_beds_per_thousands, and male_smokers.* Any country missing these variables were not be used in the analysis.

The variable *human_development_index* was missing data which exists but was not recorded. Any country missing this data were not used in the analysis.

The variable *test_unit* which is a categorical variable will not be used in the project.

Next, we removed the rows of data containing NAs. After the initial descriptive statistical results, the following variables had negative minimum values: *new_cases, new_cases_smoothed, new_deaths, new_deaths_smoothed, new_cases_per_million, new_cases_smoothed_per_million, new_cases_smoothed_per_million, and new_deaths_smoothed _per_million.* Originally, we considered this was an error and spent time considering how to address the problem. We decided to either update the value to 0, change the value from negative to positive or to remove the value. This was a difficult decision. In R programming environment, we planned to use an if_else statement to update any negative value in those columns. After additional research on the John Hopkins COVID-19 dashboard, we determined that the negative values were likely corrections. The countries with negative *new_cases, and new_deaths* are Benin, Ecuador, Italy, Lithuania, and Luxembourg. In many cases, if we updated the negative value to a positive value, the number of cases in the past would be greater than the current number of cases.

Continent	Location	Date	gdp_per_capita	
total_cases	new_cases	new_cases_smoothed	diabetes_prevalence	cardiovasc_death_rate
total_deaths	new_deaths	new_deaths_smoothed	hospital_beds_per_thousand	aged_70_older
total_cases_per_million	new_cases_per_million	new_cases_smoothed_per_million	extreme_poverty	aged_65_older
total_deaths_per_million	new_deaths_per_million	new_deaths_smoothed_per_million	female_smokers male_smokers	median_age
stringency_index	population	population_density	life_expectancy	human_development_index

Table 4

Table 4 contains the 30 variables used in the analysis of the COVID-19 dataset.

In addition, the following countries (Locations) used in the analysis:

Albania	*Colombia*	*Greece*	*Liberia*	*Pakistan*	*Timor*
Algeria	*Costa Rica*	*Haiti*	*Lithuania*	*Panama*	*Togo*
Argentina	*Croatia*	*Hungary*	*Luxembourg*	*Paraguay*	*Tunisia*
Australia	*Denmark*	*Iceland*	*Malawi*	*Portugal*	*Turkey*
Austria	*Djibouti*	*India*	*Malaysia*	*Romania*	*Uganda*
Bangladesh	*Dominican Republic*	*Indonesia*	*Mauritius*	*Russia*	*Ukraine*
Belgium	*Ecuador*	*Iran*	*Mexico*	*Seychelles*	*United Kingdom*
Benin	*Egypt*	*Ireland*	*Moldova*	*Slovakia*	*United States*
Bosnia and Herzegovina	*El Salvador*	*Israel*	*Mongolia*	*South Africa*	*Uruguay*
Brazil	*Estonia*	*Italy*	*Morocco*	*South Korea*	*Vietnam*
Bulgaria	*Ethiopia*	*Kazakhstan*	*Mozambique*	*Spain*	*Yemen*
Burkina Faso	*Fiji*	*Kenya*	*Myanmar*	*Sri Lanka*	*Zambia*
Canada	*Gambia*	*Kyrgyzstan*	*Nepal*	*Sweden*	*Zimbabwe*
Chile	*Georgia*	*Laos*	*Niger*	*Tanzania*	
China	*Ghana*	*Latvia*	*Norway*	*Thailand*	

Table 5

Table 5 contains the 88 random countries out of 210 used in the analysis of the COVID-19 dataset.

3 – Descriptive / Exploratory Analysis

We used Excel to create a quick descriptive analysis on the final dataset used for the COVID-19 project.

Average new Cases	1,511 cases per day
Average new deaths	44 deaths per day
Average Stringency index	55.43
Average Life Expectancy	74
Percentage Male Smokers	32.5%
Percentage over 65	10.45%

3.1 – Visualization Through Exploring Data: Dataset – Filtered 88 countries

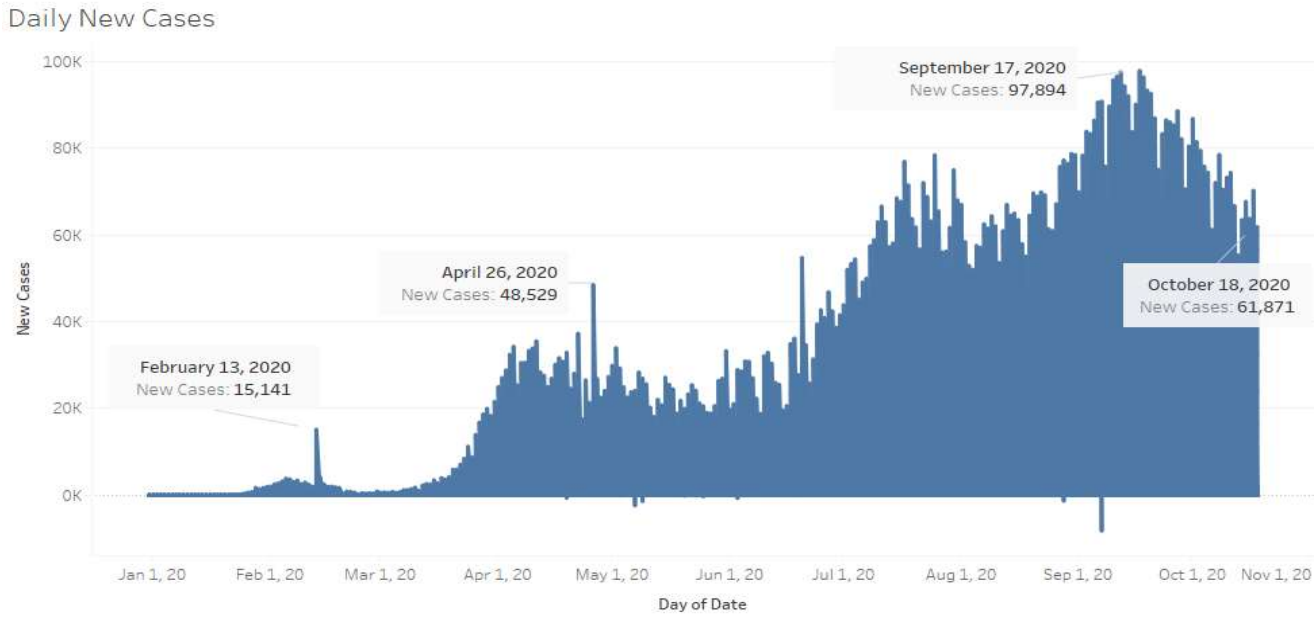


Figure 8 Daily New Cases

Figure 8 is a chart of the daily new cases for the sample 88 countries. The number of cases peaked on Sepember 17. As of October 18, the number of cases have begun to decline slightly.

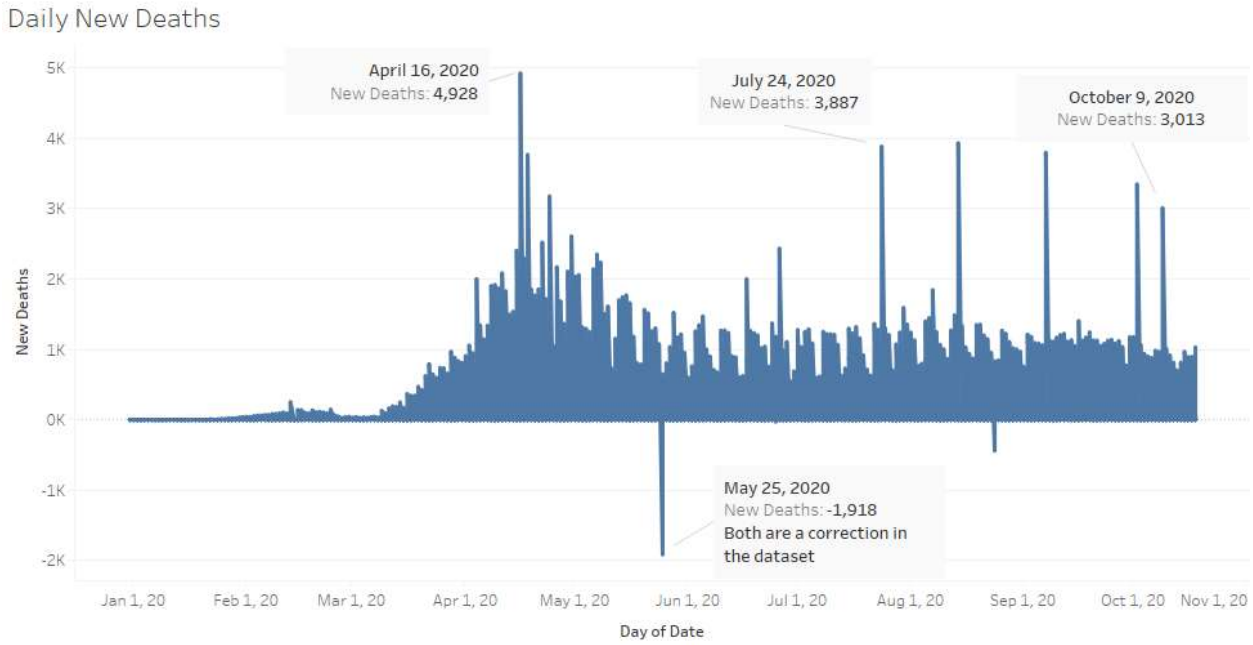


Figure 9 Daily New Deaths

Figure 9 is a chart of the daily new deaths for the sample 88 countries. The number of cases peaked on April 16 . As of October 18, the number of cases have continued to decline slight. However, the number of deaths continue to have random spikes.

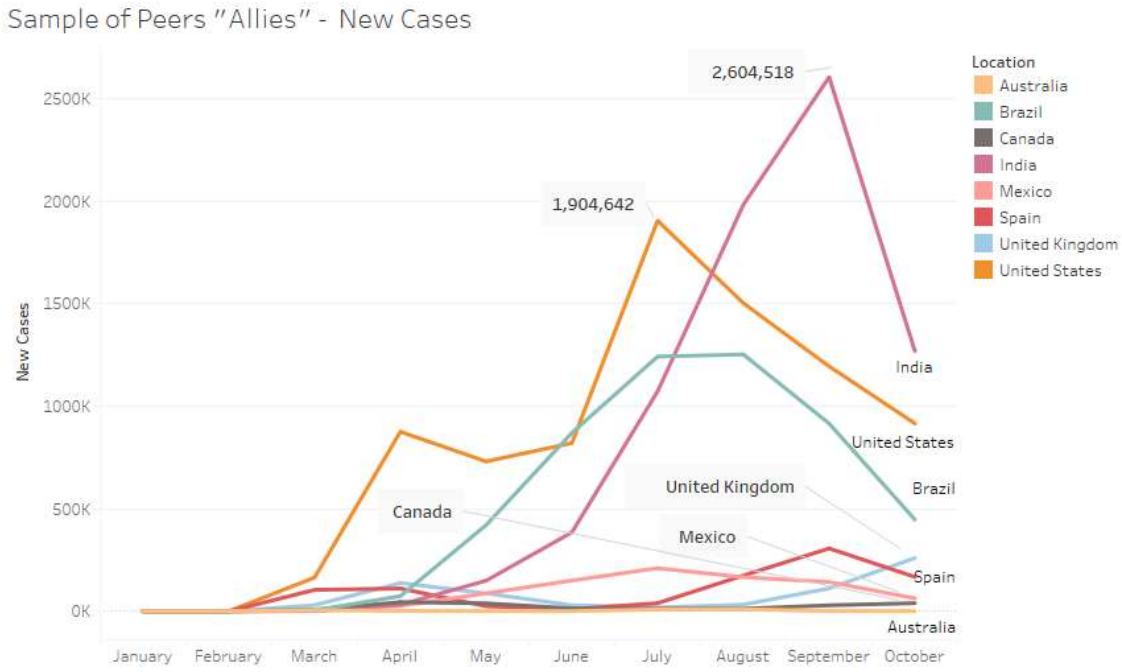


Figure 10 Peer Handling – New Cases

Figure 10 is a chart of how our peers (allies) have dealt with the number of new cases. In September, India had a large spike in cases, but the number of cases has been declining. As of October 18, Spain Mexico, Australia, and Canada had less than 170,000 new cases.

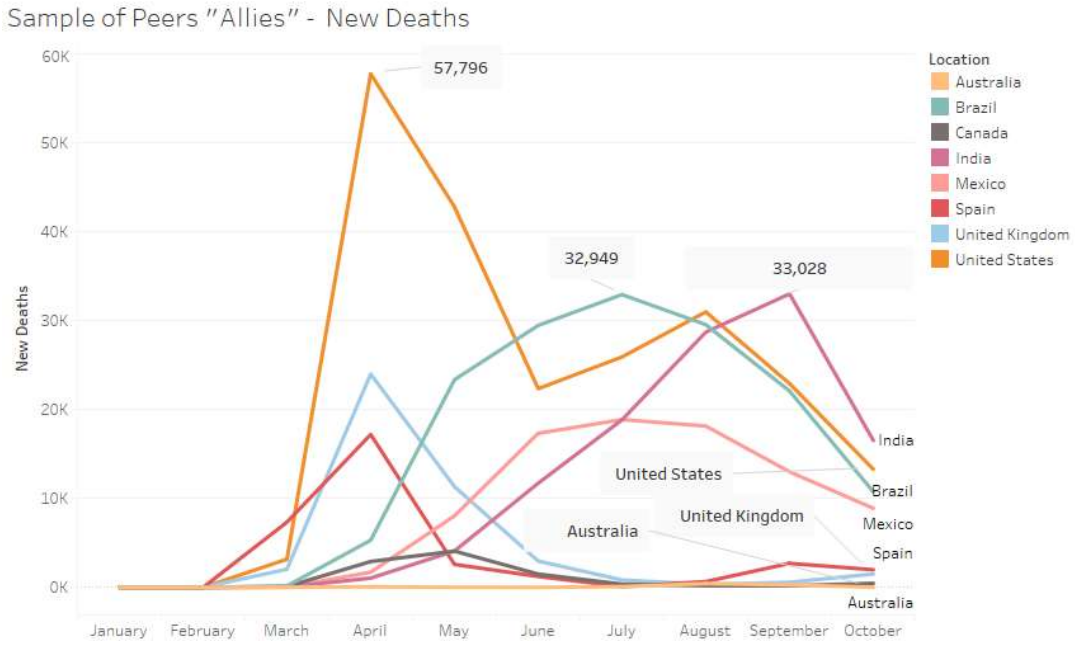


Figure 11 Peer Handling – New Deaths

Figure 11 is a chart of how our peers (allies) have dealt with the number of new deaths. All countries are seeing a decline in number of deaths. The country with the least number of deaths at October 18,2020 was Australia with 22 deaths.

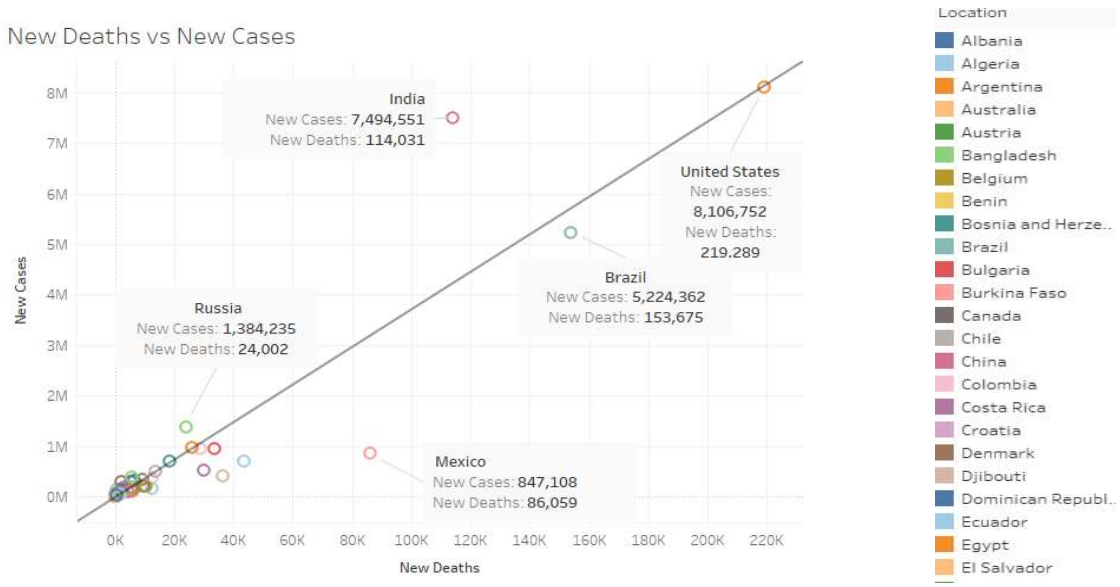


Figure 12 New Deaths vs New Cases

Figure 12 Charts new deaths and new cases. Most of the countries are gathered around 10,000 to 15,000 deaths verses the number of cases. We have highlighted outliers.

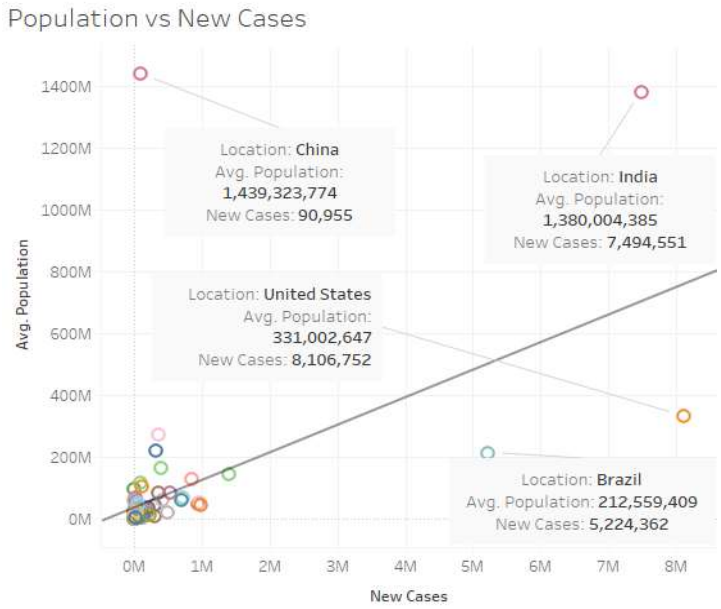


Figure 13 Population vs cases

Figure 13 Charts new case verses population. Most countries have less than half million cases. Note the outliers: China has a populaton with 1.4 billion, but has less than 100,000 cases. Also the United State, it has a populaton of 331 million, but more than 8 million cases of the virus.

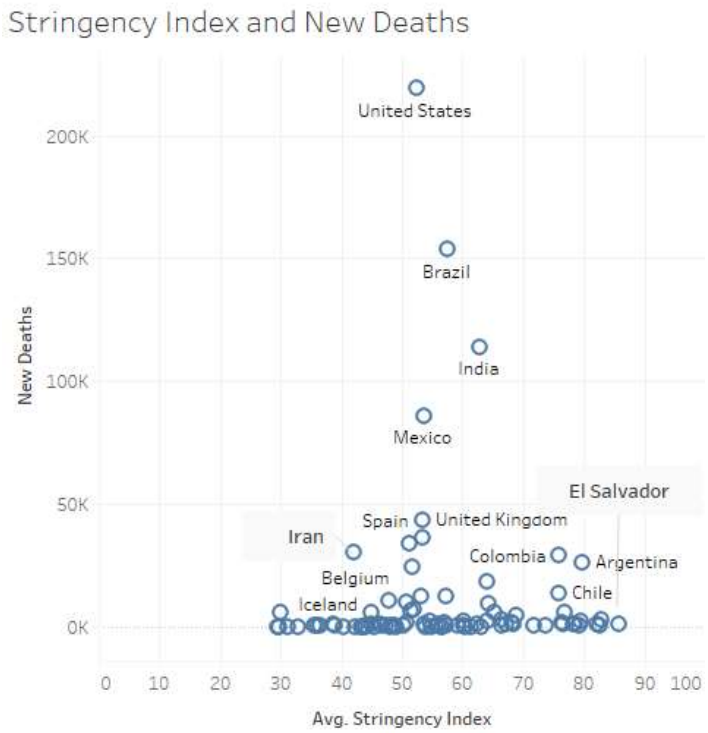


Figure 14 Stringency Index vs New Deaths

Figure 14 charts New deaths verses Stringency index. There are outliers – US, India, and Mexico, but a majority of the countries have less than 5,000 deaths.

3.2 – Statistical Inference

For this project, **new_deaths** and **new_cases** are the response variables. The following variables are the predictor variables - ***stringency_index, population, population_density, median_age, aged_65_older, aged_70_older, gdp_per_capita, extreme_poverty, cardiovasc_death_rate, diabetes_prevalence, female_smokers, male_smokers, hospital_beds_per_thousand, life_expectancy, and human_development_index.***

Correlation Matrix is a method used to determine if a relationship exists between two or more variables. The relationship can be either negative or positive. A correlation of 0 indicates that the variables have no relationship. Typically, we are interested in correlation values above 0.07 which means there is a strong relationship. We performed correlation testing on new_deaths and new_cases.

The strongest correlation is between new_deaths and new_cases. It has a positive relationship, which means as the number of cases increases the number of deaths from the virus increases. The correlation value is 0.7533873. It is the only strong relationship.

Correlation on New_deaths

Other correlations to new_deaths which are not as strong are the Stringency index at 16.05%, Population at 25.59%, Cardiovascular death rate at (11.51%), Diabetes prevalence at 13.62%, Males Smokers at (11.69%), Human Dev Index at 10.36% and GDP per capita at 9.09%. One interesting item is the correlation between males smokers and new deaths. This has a negative relationship. As deaths increase, we see a decrease in the number of male smokers by 11.69%.

Correlation on New_cases

Other correlations to new_cases which are not as strong are the Stringency index at 13.45%, Population at 35.84%, Diabetes prevalence at 12.79%, and Males Smokers at (9.61%). The strongest relationship is Population. This indicates as the population increases, the number of cases will increase. Since we know this virus transmits quickly from person to person, we should not be surprised by this result.

3.3 – Regression Analysis

Regression Analysis is a method used to determine if the correlated variables are significant. We are typically looking for the value less than .05.

Significant variables to New_deaths

We are considering if the following variables -Stringency index, Population, Cardiovascular death rate, Diabetes prevalence, Males Smokers, Human Development Index, and GDP per capita were correlated with new deaths.

```
Call:
lm(formula = Covid_Corr$new_deaths ~ Covid_Corr$new_cases + Covid_Corr$stringency_index
+
    Covid_Corr$population + Covid_Corr$cardiovasc_death_rate +
    Covid_Corr$diabetes_prevalence + Covid_Corr$male_smokers +
    Covid_Corr$human_development_index)

Residuals:
    Min      1Q  Median      3Q     Max
-1953.8   -27.2   -10.1     6.2  4309.6

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -4.202e+01  5.939e+00  -7.074 1.55e-12 ***
Covid_Corr$new_cases             1.889e-02  1.225e-04 154.184  < 2e-16 ***
Covid_Corr$stringency_index      4.749e-01  2.926e-02  16.228  < 2e-16 ***
Covid_Corr$population           -1.168e-08  3.733e-09  -3.128  0.00176 **
Covid_Corr$cardiovasc_death_rate -5.982e-02  9.464e-03  -6.321 2.65e-10 ***
Covid_Corr$diabetes_prevalence   2.464e+00  2.469e-01   9.981  < 2e-16 ***
Covid_Corr$male_smokers         -4.373e-01  6.964e-02  -6.279 3.48e-10 ***
Covid_Corr$human_development_index  5.786e+01  6.556e+00   8.826  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 115.4 on 22284 degrees of freedom
Multiple R-squared:  0.5809,    Adjusted R-squared:  0.5808
F-statistic:  4412 on 7 and 22284 DF,  p-value: < 2.2e-16
```

Table 6 Results from Correlation to Deaths

Table 6 - The most significant variables for new deaths are new cases, stringency index, diabetes prevalence and the human development index.

Significant variables to New_cases

```
Call:
lm(formula = Covid_Corr$new_cases ~ Covid_Corr$stringency_index +
    Covid_Corr$population + Covid_Corr$diabetes_prevalence +
    Covid_Corr$male_smokers)

Residuals:
   Min     1Q Median     3Q    Max
-17532  -1539   -649    296  80805

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.723e+01  1.553e+02   0.111    0.912
Covid_Corr$stringency_index    2.704e+01  1.570e+00  17.217   <2e-16 ***
Covid_Corr$population          1.041e-05  1.910e-07  54.494   <2e-16 ***
Covid_Corr$diabetes_prevalence 1.727e+02  1.310e+01  13.182   <2e-16 ***
Covid_Corr$male_smokers       -6.217e+01  3.014e+00 -20.624   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

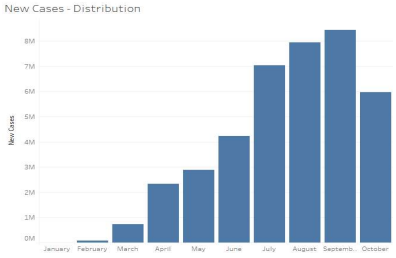
Residual standard error: 6348 on 22287 degrees of freedom
Multiple R-squared:  0.1601,    Adjusted R-squared:  0.1599
F-statistic:  1062 on 4 and 22287 DF,  p-value: < 2.2e-16
```

Table 7 Results from Correlation to Cases

Table 7 - The most signatificant variables for new cases are stringency index, population, diabetes prevalence and if the smoker was a male.

3.4 One – Sample Hypothesis Testing

We visually explored and used some statistical tools to learn more about the dataset. We now want to see if there is a hypothetical difference between the mean of new cases for all 210 countries and the 88 countries we randomly selected for the project based on meeting all of the conditions.



We made the assumption that new cases has a normal distribtuion in order to complete the one-sample hypothesis testing.

The distribution of new cases is actually skewed the right. See figure to the left.

We want to see if the mean of new cases is larger than the hypothetical value of 1,000. The hypothesis for one -sample hypothesis testing is the following:

$$H_o: \mu = \mu_o \quad (1)$$

$$H_1: \mu > \mu_o \quad (2)$$

Using the Test Statistic: $t_{obs} = \frac{\times - \mu_0}{\frac{s}{\sqrt{n}}}$

Hypothesis Testing: Is there a significant difference between the mean of new cases (or new deaths) for all countries compared to our sample of countries?

(1) **Null Hypothesis**: Our null hypothesis is there is no difference between the mean of new cases for all countries compared to our sample. Ho: population mean cases = 100,000 cases.
(2) **Alternate Hypothesis**: Our alternate hypothesis is there is a difference between mean of new cases for all countries compared to our sample. H1: population mean cases < 100,000 cases.

Using the R programming environment, we used the t.test function to complete the t-statistical test on the sample data. The t value was -6.03068 with a p-value of 1.441 e^{-10} with a .05 significant level or 95% confidence interval. Since the p-value was less than .05, we can reject the null and concluded that the population mean of the new cases was significantly less than 100,000 cases. We want to see if the mean of new deaths is larger than the hypotheized value of 1,000 cases.

The t value was -1252.1 with a p-value of 2.2 e^{-16} with a .05 significant level or 95% confidence interval. Since the p-value was less than .05, we rejected the null and concluded that the population mean of the new cases was significantly less than 1,000 cases.

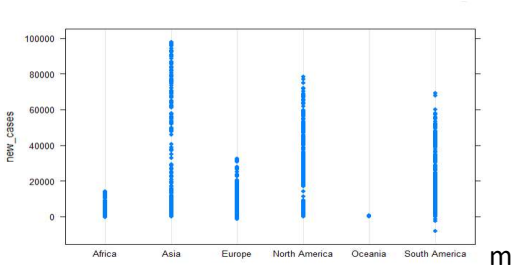
3.5 – AVONA

AVONA is a statistial method which will help us answer these questions - “Are the means of the new cases or new deaths different for the five continents”.

Prior to completing the ANOVA testing, we completed descriptive statistical analysis on the regions and created a chart to see if we can determine if the means are significantly different.

continent <chr>	mean <dbl>	sd <dbl>
Africa	184.09633	855.66689
Asia	1138.83972	6740.37675
Europe	680.24697	2034.19841
North America	2734.08729	9790.05317
Oceania	33.97653	96.07114
South America	3151.92890	8164.22362

6 rows



m

Table 8 Mean and Standard Deviation of Continents **Figure 15**

Table 8 provides the mean and standard deviation of new cases by continents. From these values, we can see that the means are different.

Figure 15 provides a plot of the data which also indicates a variance in the different continents. We also see outliers within the regions. Although, visually it looks as if the means are different, we must perform an ANOVA test to be certain.

We used two methods of ANOVA testing using the R programming environment. In both methods, the p-value is 208.4 e.16 which was less than .05. So, we rejected the hypothesis that all means are equal. We must conclude that at least one continent was different than the others regarding the new cases. The results from the ANOVA testing are displayed in Appendix F.

5 Data Mining and Predictive Analytics

As we continue our exploration of the COVID-19 dataset, we used data mining and predictive analytics to predict the future of the virus. We used 5 general methods to create models for the data. Those models are Growth Factor, Cluster Analysis, Linear Models, Polynomial (Multiple) Regression, and Machine Learning Algorithms. We continued to use the final dataset containing the 88 countries selected in the descriptive analysis process.

5.1 Growth Factor

The growth factor is the percentage increase in the value of a variable from the prior day. We used the following formula: *Growth Factor = Number of new cases at period t / Number of new cases a t-1*

A value greater than 1 indicates the variable is increasing. A value less than 1 indicates the variable is declining. While a value equal to 1 indicated the variable is steady.

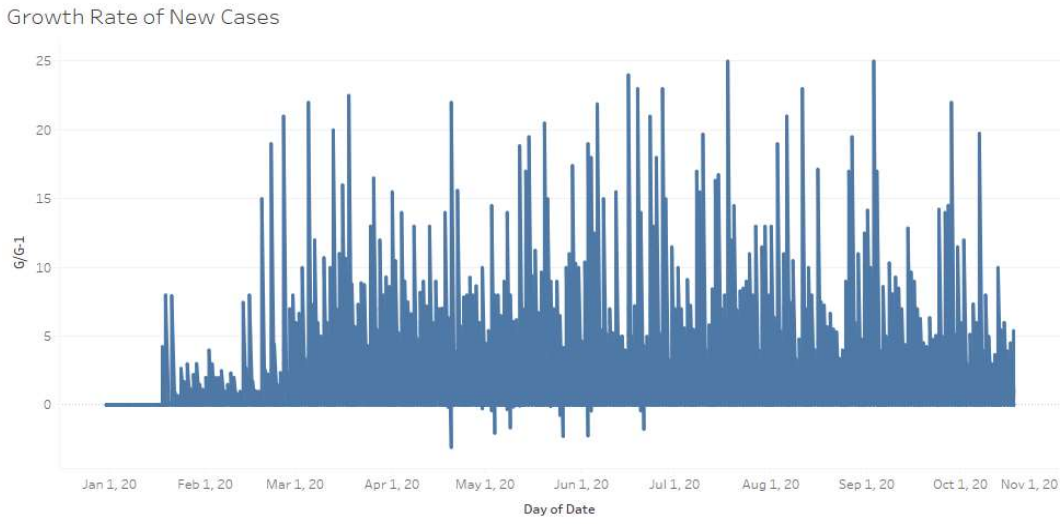


Figure 16 Growth Factor – New Cases

Figure 16 is the daily growth factor of new cases for all 88 countries in the final sample dataset. The growth factor for new cases continue to increase.

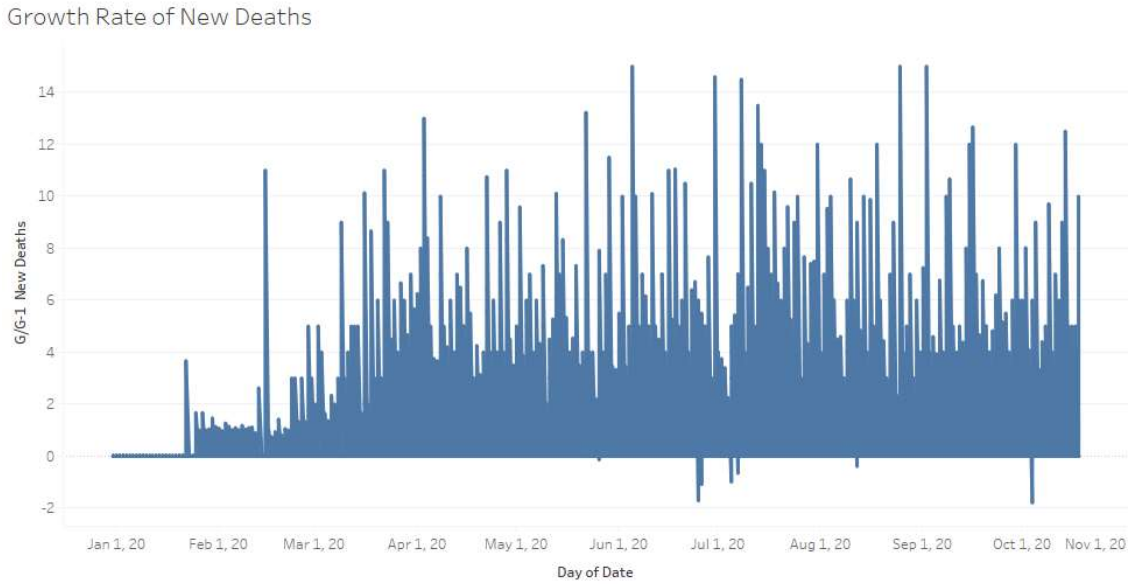


Figure 17 Growth Factor – New Deaths

Figure 17 is the daily growth factor of new deaths for all 88 countries in the final sample dataset. The growth factor for new deaths continue to increase.

Both charts for the growth factors using all 88 location were hard to read, so we created charts for random "allies" of the United States. Focusing on the individual countries will make the charts easier to read.

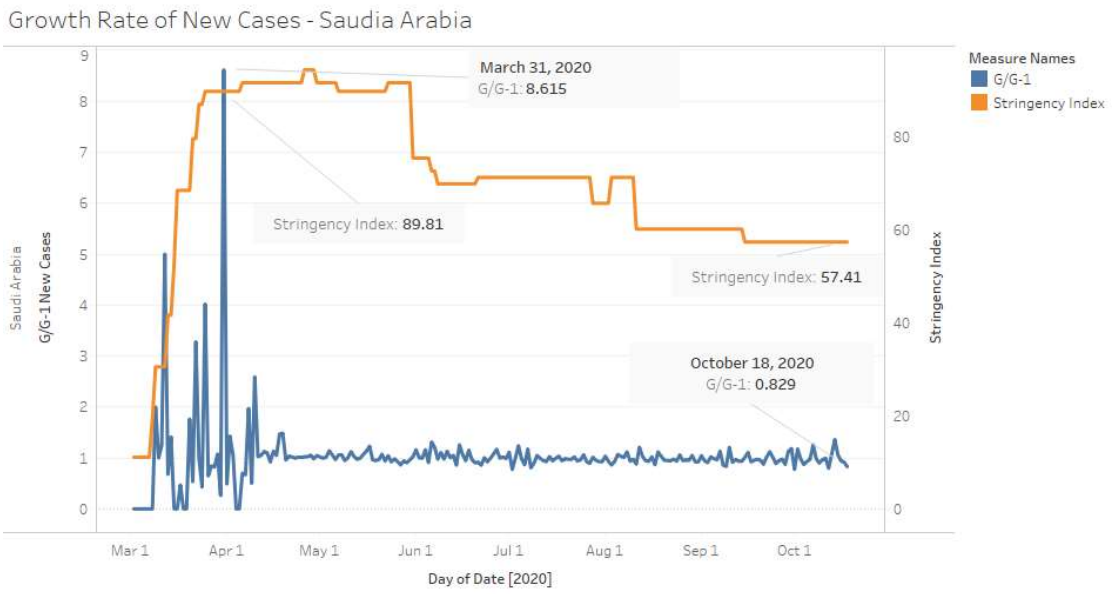


Figure 18 Growth Factor – New Cases Saudi Arabia with Stringency Index

Figure 18 is the daily growth factor of new cases in Saudi Arabia with stringency index. The growth factor increases until at reached a peak factor of 8.615. As the number of cases increased, the country increased the index. Currently, the country is seeing a decline in the growth factor for cases.

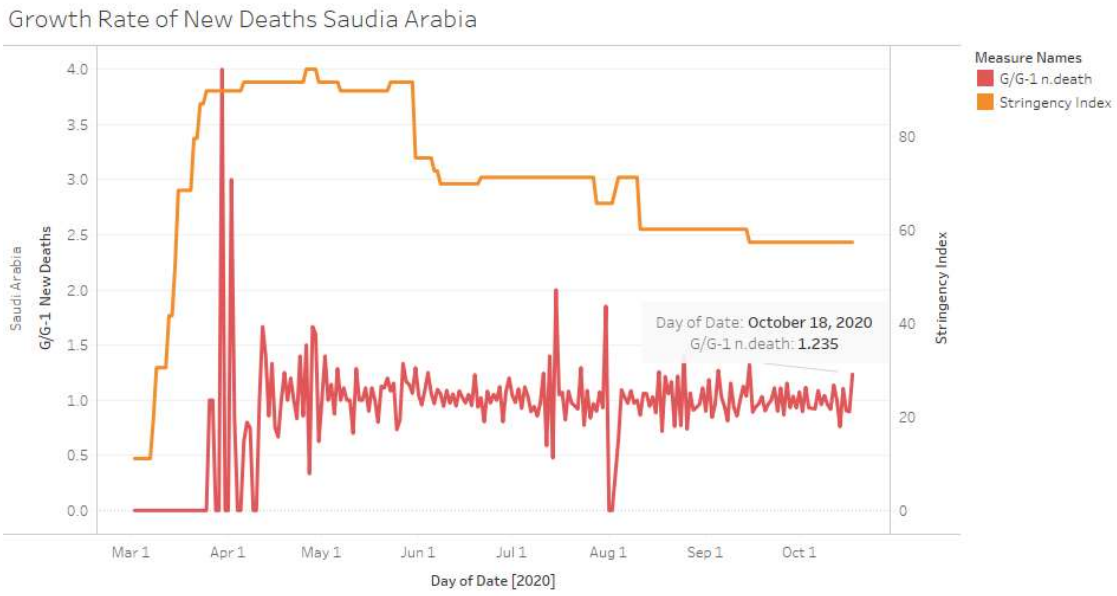


Figure 19 Growth Factor – New Deaths Saudi Arabia with Stringency Index

Figure 19 is the daily growth factor of new deaths in Saudi Arabia with stringency index. Currently, the country is seeing a slight increase in the growth factor for cases.

Figure 20 Growth Factor – New Cases Japan with Stringency Index

Figure 20 is the daily growth factor of new cases in Japan with stringency index. The growth factor increases until at reached a peak factor of 8.8. As the number of cases increased, the country increased the index. Currently, the country is seeing a decline in the growth factor for cases with a factor of .938. It is interesting that Japan has an overall lower stringency rate of 31.48 compared to other countries with lower rates.

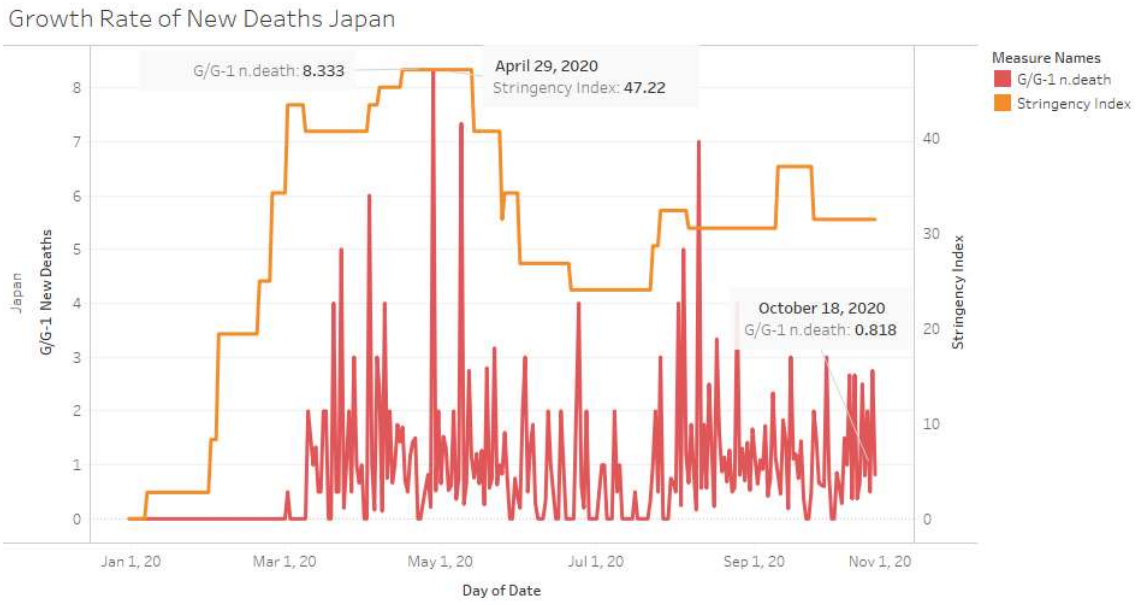


Figure 21 Growth Factor – New Deaths Japan with Stringency Index

Figure 21 is the daily growth factor of new deaths in Japan with stringency index. Currently, the country is seeing a decline in the growth factor for cases with a factor of .938. Currently, the country is on a decline with a factor of .818.

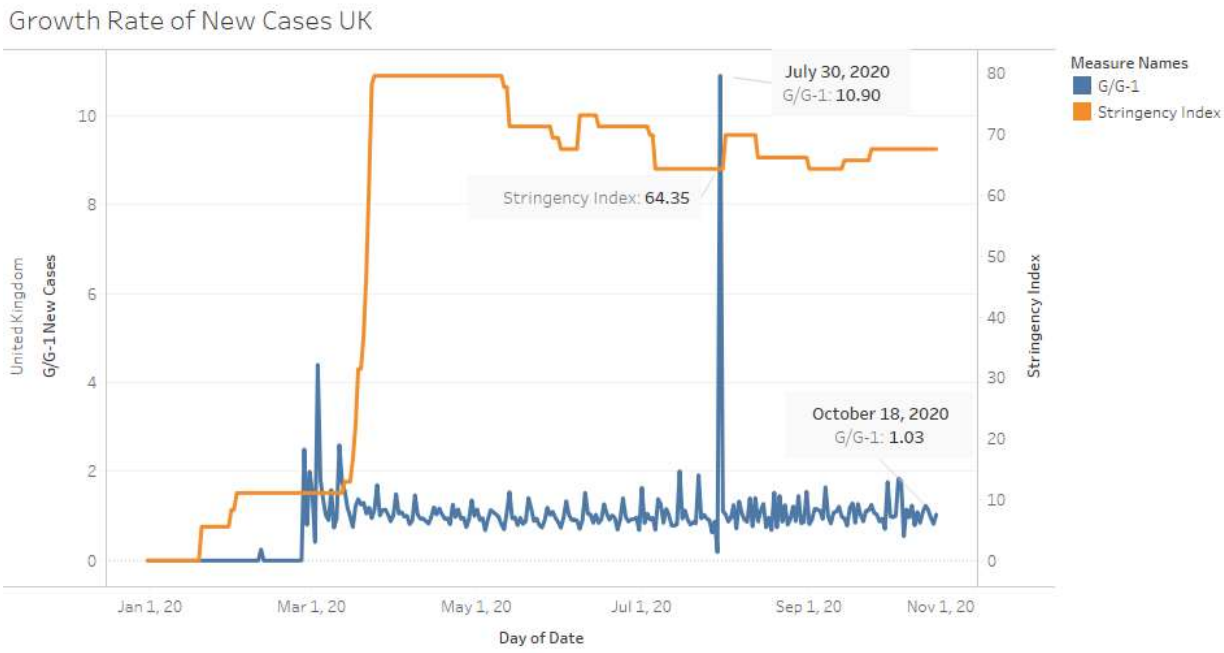


Figure 22 Growth Factor – New Cases Japan with Stringency Index

Figure 22 is the daily growth factor of new cases in United Kingdom with stringency index. Currently, the factor is 1.03 which indicated the growth rate is stable.

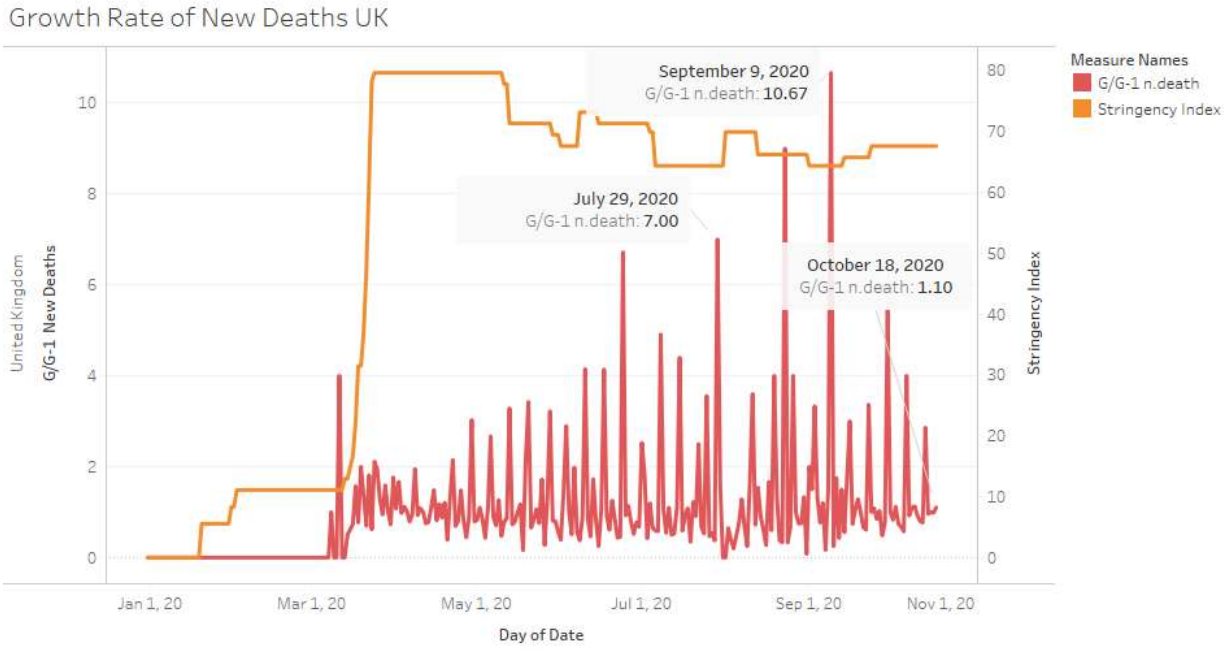


Figure 23 Growth Factor – New Deaths United Kingdom with Stringency Index

Figure 23 is the daily growth factor of new deaths in United Kingdom with stringency index. Currently, the country is seeing a slight increase in deaths with a factor of 1.10.

5.2 Cluster Analysis

Cluster Analysis is the process of dividing a group into subgroups and completing a census the data in those subgroups. For this analysis, we used Tableau.

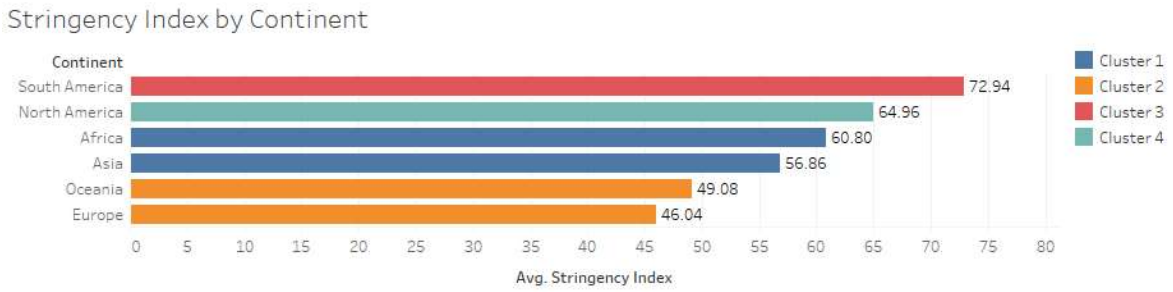


Figure 24 Index by Continent

Figure 24 is a chart of the stringency index for the six continents in the dataset. Tableau clustered the data into 4 subgroups. South America has the highest average index at 72.94%.

5.2.1 Cluster Analysis – New Deaths

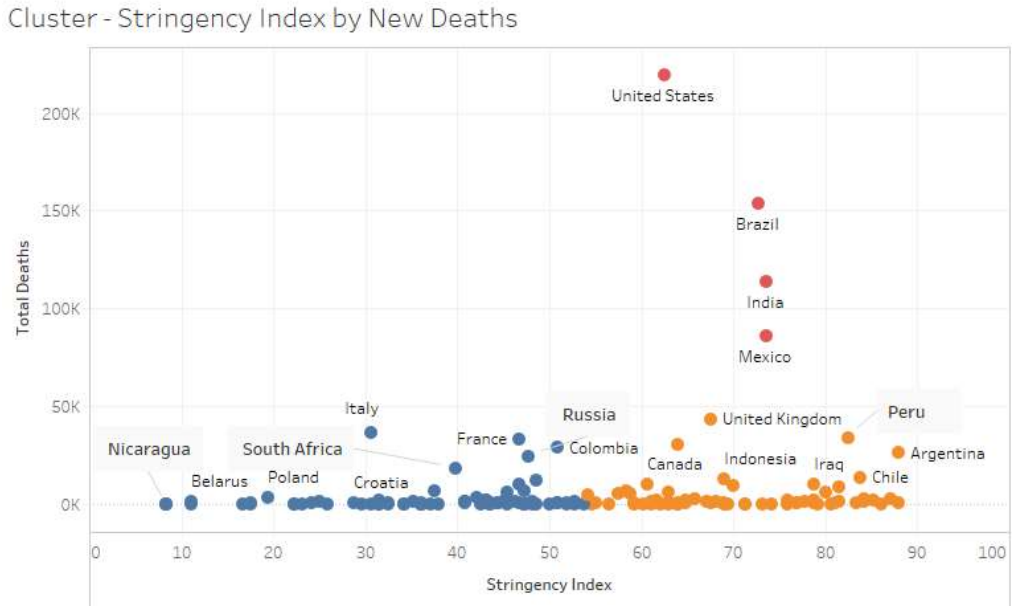


Figure 25 Clustering Stringency and Total Deaths

Figure 25 is a chart of the average stringency index and the total number of deaths on October 18, 2020.

We allowed Tableau to create the number of clusters for the 88 locations (countries). Tableau divided the group by the index and placed the three outliers in a cluster of their own. We exported the data from Tableau to get the names of the individual countries and computed the average and standard deviation.

Cluster 3 - Red

Clusters	Location	Avg. Stringency Index	Total Deaths
Cluster 3	Brazil	72.69	153,675
Cluster 3	India	73.61	114,031
Cluster 3	Mexico	73.61	86,059
Cluster 3	United States	62.5	219,289

Stringency Index	
Average Number of cases	70.6025
Standard Deviation of Sample	5.42

Total Deaths	
Average Number of cases	143,264
Standard Deviation of Sample	57,778.80

Table 8 Clustering Stringency and Total Deaths

Table 8 is a chart of the average stringency index and the total number of deaths on October 18, 2020. This information is for cluster 3.

Cluster 2 - Orange

Clusters	Location	Avg. Stringency Index	Total Deaths
Cluster 2	Algeria	75.93	1,846
Cluster 2	Argentina	87.96	26,107
Cluster 2	Australia	68.06	904
Cluster 2	Azerbaijan	68.98	623
Cluster 2	Bahrain	63.89	295
Cluster 2	Bangladesh	80.09	5,646
Cluster 2	Belize	69.44	43
Cluster 2	Bhutan	71.3	-
Cluster 2	Bolivia	81.48	8,463
Cluster 2	Botswana	63.89	20
Cluster 2	Canada	60.65	9,746
Cluster 2	Cape Verde	68.98	85
Cluster 2	Central African Republic	56.48	62
Cluster 2	Chile	83.8	13,588
Cluster 2	China	54.17	4,739
Cluster 2	Costa Rica	61.11	1,183
Cluster 2	Dominican Republic	64.81	2,195
Cluster 2	Egypt	62.96	6,109
Cluster 2	Eritrea	86.11	-
Cluster 2	Ethiopia	77.78	1,346
Cluster 2	Gabon	74.07	54
Cluster 2	Gambia	75.93	118
Cluster 2	Georgia	61.11	128
Cluster 2	Guyana	79.17	109
Cluster 2	Honduras	84.26	2,563
Cluster 2	Indonesia	68.98	12,431
Cluster 2	Iran	63.89	30,123
Cluster 2	Iraq	78.7	10,198
Cluster 2	Ireland	61.57	1,849
Cluster 2	Israel	85.19	2,190
Cluster 2	Jamaica	71.3	168
Cluster 2	Jordan	78.7	330
Cluster 2	Kazakhstan	78.7	2,149
Cluster 2	Kenya	67.59	825
Cluster 2	Kuwait	62.96	694
Cluster 2	Kyrgyzstan	67.13	1,108
Cluster 2	Lebanon	59.26	517
Cluster 2	Liberia	62.96	82
Cluster 2	Libya	83.33	699
Cluster 2	Malawi	54.63	181
Cluster 2	Malaysia	60.19	180
Cluster 2	Morocco	65.74	2,878
Cluster 2	Mozambique	62.04	74
Cluster 2	Myanmar	81.02	838
Cluster 2	Nepal	76.85	727
Cluster 2	Netherlands	58.33	6,728
Cluster 2	Oman	84.26	1,071
Cluster 2	Panama	87.04	2,557
Cluster 2	Paraguay	81.48	1,179
Cluster 2	Peru	82.41	33,702
Cluster 2	Qatar	64.81	223
Cluster 2	Saudi Arabia	57.41	5,165
Cluster 2	South Korea	55.09	444
Cluster 2	Suriname	63.89	109
Cluster 2	Swaziland	59.26	115
Cluster 2	Trinidad and Tobago	80.56	95
Cluster 2	Turkey	69.91	9,224
Cluster 2	Uganda	73.15	96
Cluster 2	Ukraine	58.8	5,517
Cluster 2	United Kingdom	67.59	43,579
Cluster 2	Uzbekistan	62.96	525
Cluster 2	Venezuela	87.96	725
Cluster 2	Zimbabwe	76.85	231

Stringency Index	
Average Number of cases	70.5539683
Standard Deviation of Sample	9.80855791
Total Deaths	
Average Number of cases	4,214
Standard Deviation of Sample	8471.22276

Table 9 Clustering Stringency and Total Deaths

Table 9 is a chart of the average stringency index and the total number of deaths on October 18, 2020. This information is for cluster 2.

Cluster 1 - Blue

Clusters	Location	Avg. Stringency Index	Total Deaths
Cluster 1	Afghanistan	25	1,488
Cluster 1	Albania	43.52	448
Cluster 1	Austria	44.91	907
Cluster 1	Barbados	31.48	7
Cluster 1	Belarus	11.11	921
Cluster 1	Benin	47.22	41
Cluster 1	Bosnia and Herzegovina	40.74	980
Cluster 1	Brunei	31.48	3
Cluster 1	Bulgaria	35.19	968
Cluster 1	Burkina Faso	36.11	65
Cluster 1	Burundi	11.11	1
Cluster 1	Cambodia	37.04	-
Cluster 1	Cameroon	36.11	423
Cluster 1	Colombia	50.93	28,803
Cluster 1	Croatia	28.7	355
Cluster 1	Cyprus	53.7	25
Cluster 1	Czech Republic	48.15	1,352
Cluster 1	Denmark	50.93	679
Cluster 1	Djibouti	37.96	61
Cluster 1	Ecuador	48.61	12,375
Cluster 1	El Salvador	52.78	922
Cluster 1	Estonia	22.22	68
Cluster 1	Fiji	45.37	2
Cluster 1	Finland	32.41	351
Cluster 1	France	46.76	33,392
Cluster 1	Germany	46.76	9,777
Cluster 1	Ghana	44.44	310
Cluster 1	Greece	46.76	500
Cluster 1	Guatemala	42.13	3,515
Cluster 1	Guinea	48.15	70
Cluster 1	Haiti	43.52	231
Cluster 1	Hungary	40.74	1,142
Cluster 1	Iceland	34.26	11
Cluster 1	Italy	30.56	36,474
Cluster 1	Japan	31.48	1,670
Cluster 1	Laos	32.41	-
Cluster 1	Latvia	30.56	43
Cluster 1	Lithuania	48.61	112
Cluster 1	Luxembourg	43.52	133
Cluster 1	Madagascar	40.74	238
Cluster 1	Mali	34.26	132
Cluster 1	Mauritius	16.67	10
Cluster 1	Moldova	46.3	1,569
Cluster 1	Mongolia	42.59	-
Cluster 1	New Zealand	22.22	25
Cluster 1	Nicaragua	8.33	154
Cluster 1	Niger	17.59	69
Cluster 1	Norway	17.59	278
Cluster 1	Pakistan	37.5	6,654
Cluster 1	Philippines	47.22	6,603
Cluster 1	Poland	19.44	3,524
Cluster 1	Portugal	43.06	2,162
Cluster 1	Romania	45.37	5,812
Cluster 1	Russia	47.69	24,002
Cluster 1	Serbia	50.93	774
Cluster 1	Seychelles	36.11	-
Cluster 1	Singapore	52.78	28
Cluster 1	Slovakia	50	82
Cluster 1	Slovenia	47.22	153
Cluster 1	Solomon Islands	8.33	-
Cluster 1	South Africa	39.81	18,408
Cluster 1	Sri Lanka	23.15	13
Cluster 1	Sudan	51.85	836
Cluster 1	Switzerland	43.06	1,822
Cluster 1	Tajikistan	36.11	80
Cluster 1	Tanzania	8.33	21
Cluster 1	Thailand	43.52	59
Cluster 1	Timor	29.63	-
Cluster 1	Togo	47.22	51
Cluster 1	Tunisia	24.07	626
Cluster 1	United Arab Emirates	52.78	459
Cluster 1	Uruguay	25.93	51
Cluster 1	Vietnam	51.85	35
Cluster 1	Yemen	45.37	597
Cluster 1	Zambia	44.44	346

Stringency Index	
Average Number of cases	37.5
Standard Deviation of Sample	12.2268526

Stringency Index	
Average Number of cases	2857.30667
Standard Deviation of Sample	7,352.092

Table 10 Clustering Stringency and Total Deaths

Table 10 is a chart of the average stringency index and the total number of deaths on October 18, 2020. This information is for cluster 1.

5.2.2 Cluster Analysis – New Cases

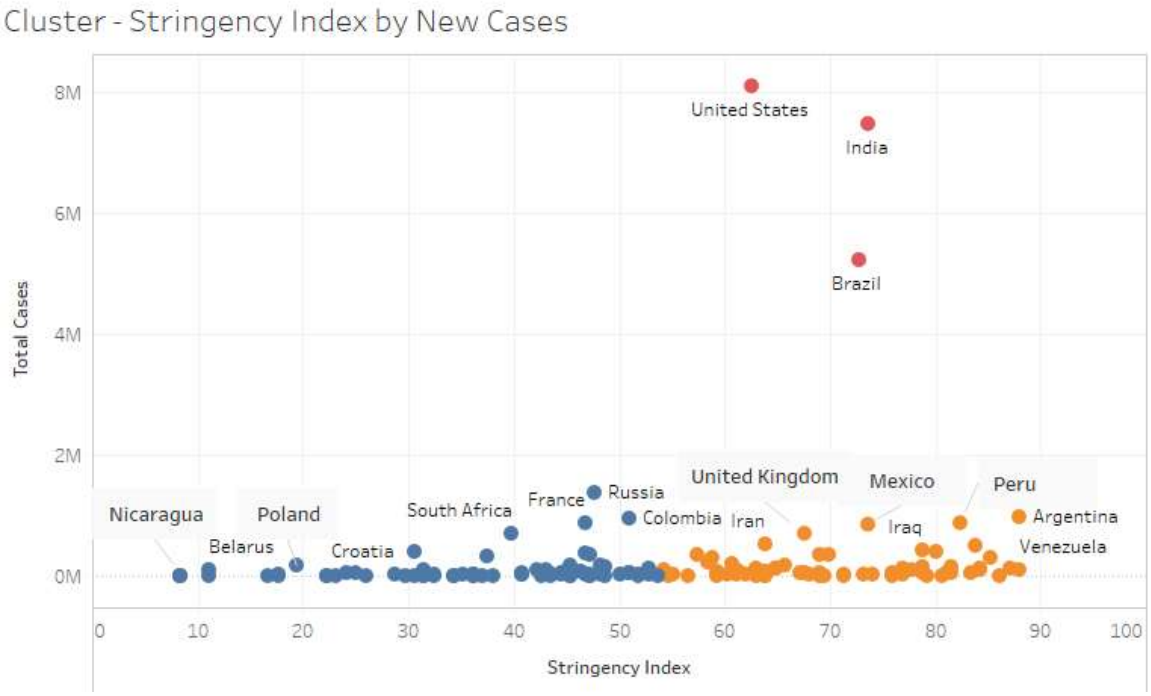


Figure 25 Clustering Stringency and Total Cases

Figure 25 is a chart of the average stringency index and the total number of cases on October 18, 2020

Cluster 3 - Red

Clusters	Location	Avg. Stringency Index	Total Cases
Cluster 3	Brazil	72.69	5,224,362
Cluster 3	India	73.61	7,494,551
Cluster 3	United States	62.5	8,106,752

Stringency Index	
Average Number of cases	69.6
Standard Deviation of Sample	6.17

Total Deaths	
Average Number of cases	6,941,888
Standard Deviation of Sample	1,518,591.47

Table 11 Clustering Stringency and Total Cases – Cluster 3

Table 11 is a chart of the average stringency index and the total number of cases on October 18, 2020. This information is for cluster 3.

Cluster 2 - Orange

Clusters	Location	Avg. Stringency Index	Total Cases
Cluster 2	Algeria	75.93	54,203
Cluster 2	Argentina	87.96	979,106
Cluster 2	Australia	68.06	27,383
Cluster 2	Azerbaijan	68.98	44,317
Cluster 2	Bahrain	63.89	77,571
Cluster 2	Bangladesh	80.09	387,295
Cluster 2	Belize	69.44	2,775
Cluster 2	Bhutan	71.3	325
Cluster 2	Bolivia	81.48	139,710
Cluster 2	Botswana	63.89	5,242
Cluster 2	Canada	60.65	196,321
Cluster 2	Cape Verde	68.98	7,638
Cluster 2	Central African Republic	56.48	4,855
Cluster 2	Chile	83.8	490,003
Cluster 2	China	54.17	90,955
Cluster 2	Costa Rica	61.11	95,514
Cluster 2	Dominican Republic	64.81	120,925
Cluster 2	Egypt	62.96	105,297
Cluster 2	Eritrea	86.11	422
Cluster 2	Ethiopia	77.78	88,434
Cluster 2	Gabon	74.07	8,881
Cluster 2	Gambia	75.93	3,649
Cluster 2	Georgia	61.11	16,285
Cluster 2	Guyana	79.17	3,710
Cluster 2	Honduras	84.26	87,594
Cluster 2	Indonesia	68.98	357,762
Cluster 2	Iran	63.89	526,490
Cluster 2	Iraq	78.7	423,524
Cluster 2	Ireland	61.57	48,678
Cluster 2	Israel	85.19	302,832
Cluster 2	Jamaica	71.3	8,195
Cluster 2	Jordan	78.7	36,053
Cluster 2	Kazakhstan	78.7	145,320
Cluster 2	Kenya	67.59	44,196
Cluster 2	Kuwait	62.96	115,483
Cluster 2	Kyrgyzstan	67.13	51,490
Cluster 2	Lebanon	59.26	61,284
Cluster 2	Liberia	62.96	1,377
Cluster 2	Libya	83.33	47,845
Cluster 2	Malawi	54.63	5,852
Cluster 2	Malaysia	60.19	19,627
Cluster 2	Mexico	73.61	847,108
Cluster 2	Morocco	65.74	170,911
Cluster 2	Mozambique	62.04	10,707
Cluster 2	Myanmar	81.02	34,875
Cluster 2	Nepal	76.85	129,304
Cluster 2	Netherlands	58.33	219,795
Cluster 2	Oman	84.26	108,296
Cluster 2	Panama	87.04	124,107
Cluster 2	Paraguay	81.48	54,015
Cluster 2	Peru	82.41	865,549
Cluster 2	Qatar	64.81	129,227
Cluster 2	Saudi Arabia	57.41	341,854
Cluster 2	South Korea	55.09	25,199
Cluster 2	Suriname	63.89	5,123
Cluster 2	Swaziland	59.26	5,765
Cluster 2	Trinidad and Tobago	80.56	5,281
Cluster 2	Turkey	69.91	345,678
Cluster 2	Uganda	73.15	10,455
Cluster 2	Ukraine	58.8	293,641
Cluster 2	United Kingdom	67.59	705,428
Cluster 2	Uzbekistan	62.96	63,124
Cluster 2	Venezuela	87.96	85,758
Cluster 2	Zimbabwe	76.85	8,110

Stringency Index	
Average Number of cases	70.60171875
Standard Deviation of Sample	9.737896444
Total Deaths	
Average Number of cases	153,496
Standard Deviation of Sample	223135.1693

Table 12 Clustering Stringency and Total Cases – Cluster 2

Table 12 is a chart of the average stringency index and the total number of deaths on October 18, 2020. This information is for cluster 2.

Cluster 1 - Blue

Clusters	Location	Avg. Stringency Index	Total Cases
Cluster 1	Afghanistan	25	40,141
Cluster 1	Albania	43.52	16,774
Cluster 1	Austria	44.91	64,495
Cluster 1	Barbados	31.48	219
Cluster 1	Belarus	11.11	86,392
Cluster 1	Benin	47.22	2,496
Cluster 1	Bosnia and Herzegovina	40.74	32,845
Cluster 1	Brunei	31.48	147
Cluster 1	Bulgaria	35.19	29,108
Cluster 1	Burkina Faso	36.11	2,343
Cluster 1	Burundi	11.11	536
Cluster 1	Cambodia	37.04	283
Cluster 1	Cameroon	36.11	21,441
Cluster 1	Colombia	50.93	952,371
Cluster 1	Croatia	28.7	24,761
Cluster 1	Cyprus	53.7	2,379
Cluster 1	Czech Republic	48.15	168,827
Cluster 1	Denmark	50.93	34,941
Cluster 1	Djibouti	37.96	5,452
Cluster 1	Ecuador	48.61	152,422
Cluster 1	El Salvador	52.78	31,456
Cluster 1	Estonia	22.22	4,052
Cluster 1	Fiji	45.37	32
Cluster 1	Finland	32.41	13,133
Cluster 1	France	46.76	867,197
Cluster 1	Germany	46.76	361,974
Cluster 1	Ghana	44.44	47,232
Cluster 1	Greece	46.76	24,932
Cluster 1	Guatemala	42.13	101,028
Cluster 1	Guinea	48.15	11,478
Cluster 1	Haiti	43.52	8,925
Cluster 1	Hungary	40.74	46,290
Cluster 1	Iceland	34.26	3,998
Cluster 1	Italy	30.56	402,536
Cluster 1	Japan	31.48	92,656
Cluster 1	Laos	32.41	23
Cluster 1	Latvia	30.56	3,392
Cluster 1	Lithuania	48.61	7,041
Cluster 1	Luxembourg	43.52	10,471
Cluster 1	Madagascar	40.74	16,810
Cluster 1	Mali	34.26	3,379
Cluster 1	Mauritius	16.67	407
Cluster 1	Moldova	46.3	66,652
Cluster 1	Mongolia	42.59	324
Cluster 1	New Zealand	22.22	1,530
Cluster 1	Nicaragua	8.33	5,353
Cluster 1	Niger	17.59	1,209
Cluster 1	Norway	17.59	16,136
Cluster 1	Pakistan	37.5	323,019
Cluster 1	Philippines	47.22	354,338
Cluster 1	Poland	19.44	167,230
Cluster 1	Portugal	43.06	98,055
Cluster 1	Romania	45.37	176,468
Cluster 1	Russia	47.69	1,384,235
Cluster 1	Serbia	50.93	35,946
Cluster 1	Seychelles	36.11	149
Cluster 1	Singapore	52.78	57,904
Cluster 1	Slovakia	50	28,268
Cluster 1	Slovenia	47.22	13,144
Cluster 1	Solomon Islands	8.33	3
Cluster 1	South Africa	39.81	702,131
Cluster 1	Sri Lanka	23.15	5,475
Cluster 1	Sudan	51.85	13,691
Cluster 1	Switzerland	43.06	74,227
Cluster 1	Tajikistan	36.11	10,455
Cluster 1	Tanzania	8.33	509
Cluster 1	Thailand	43.52	3,686
Cluster 1	Timor	29.63	29
Cluster 1	Togo	47.22	2,049
Cluster 1	Tunisia	24.07	40,542
Cluster 1	United Arab Emirates	52.78	114,387
Cluster 1	Uruguay	25.93	2,501
Cluster 1	Vietnam	51.85	1,126
Cluster 1	Yemen	45.37	2,059
Cluster 1	Zambia	44.44	15,789

Stringency Index	
Average Number of cases	37.5
Standard Deviation of Sample	12.22685259

Stringency Index	
Average Number of cases	98,872.453
Standard Deviation of Sample	234,030.601

Table 13 Clustering Stringency and Total Cases – Cluster 1

Table 13 is a chart of the average stringency index and the total number of deaths on October 18, 2020. This information is for cluster 1.

5.3 Observations

After the analysis of Growth Factor and Clustering, the overall growth factors for new cases and new deaths are continuing to decline. There continues to be random spikes in deaths. These spikes could be contributed to holidays, specifically in the United States. When we looked at the individual countries, we noticed if an increase occurred in either cases or death most countries increased the stringency index, which caused a steady decline in cases. The clustering for new deaths and new cases was interesting. Initially, Tableau only created two subgroups for new cases. The groups were divided by the index. Both had an exceptionally large standard deviation, which is caused by the larger variation of values. For example, the average number of cases was almost 6 million cases with a standard deviation of 1.5 million. This is difficult to observe when viewing the chart.

5.4 Models

We used JPM Pro15 to create models for testing the test data. We used Linear regression, Forward regression, Backward regression, Blackstrap (Random Forest), Neural Net (with and without Booster), and Boosted Tree. The procedure consists of creating a valuation column, which consists of dividing the data into training, validation, and test sets. To achieve this data set, we divided the data into 60/20/20 split. After running the model, we compared the models and selected the model with the highest R^2 and a lowest RASE.

Definitions:

R^2 is the percentage of data explained by the model. The higher the number the better the performance.

RASE is the root average squared error. The lower the number the best the performance.

AAE is the Average Absolute Error. The lower the number the best the performance.

AIC is the Akaike's information criteria which measures the fit. The lower the number, the better.

5.4.1 Linear Model

The standard linear regression is based on building a model to determine the relationship between dependent variable (response) and one or more independent variables (predictors). The relationship can be linear, non-linear or no relationship at all.

Using JMP Pro15, the linear model was able to predict 17.95% of the testing data.

5.5 Polynomial Regression

5.5.1 Forward Regression – New Cases

For the polynomial regression method, we used the forward and the backward regression method. The stepwise forward section begins with an empty model (no predictor variables). The predictor variables are added to the model one at a time. As a variable is added to the model, we are looking for the variables that provides the best improvement of R2.

The forward regression model was able to predict 17.95% of the testing data.

5.5.2 Backward Regression – New Cases

The stepwise backward selection begins with a full model (all predictor variables), then remove the least useful predictor, one at a time. For this method, continue removing predictor variables until no more can be removed without impacting the fit and maintaining the best R2 value.

The backward regression model was not succesful in predicting the testing data. The result was -3.39%. This method was the least effective.

5.6 Machine Learning Algorithms

For the Machine learning portion of the project, we will be using JMP15 and the following models – Random Forest, Neural Net, Boosted Tree, and Neural Net with Booster.

5.6.1 Random Forest

The Random Forest method forms successive independent decision trees using a bootstrap sample method. Random forest method is unlikely to have a problem with overfitting. Since each tree is constructed independently, the method is not affected by highly correlated variables.

Using Bootstrap (or Random Forest), the model was able to predict 67.74% of the testing data.

5.6.2 Neural Net

The Neutral Net models complex relationship between inputs and outputs. It uses an input layer, hidden layer(s), and an outer layer. The hidden layer contains one or more nodes. The Neutral Net used three types of transformation functions – TanH, Linear, and Gaussian.

Using the basic Neural Net model with Tour 20, it was able to predict 69.93% of the testing dataset. The addition of Tour is like using random 123. It completed the process 20 times and took the average result.

5.6.2.1 Neural Net with Booster

The basic Neural Net model with Booster 20, it was able to explain 67.93% of the test data.

5.6.3 Boosted Tree

Boosted tree is built by creating smaller decision trees, which predict the smaller residuals from the previous decision trees. The smaller trees are combined to form a larger tree.

The Booster tress model was able to explain 53.87% of the test data.

5.6.5 Model Comparison – New Cases

Model Comparison

Predictors

Target	Predictors		
new_cases	new_cases Prediction Formula - General Regr	Fit Generalized Standard Least Squares	Validation: Validation
	new_cases Prediction Formula - forward	Fit Generalized Forward Selection	Validation: Validation
	new_cases Prediction Formula - Backward	Fit Generalized Backward Elimination	Validation: Validation
	new_cases Predictor RF	Bootstrap Forest	Validation: Validation
	Predicted new_cases NN T20	Neural	Validation: Validation
	Predicted new_cases NN B20	Neural	Validation: Validation
	new_cases Predictor BT	Boosted Tree	Validation: Validation

Measures of Fit for new_cases

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	new_cases Prediction Formula - General Regr	Fit Generalized Standard Least Squares		0.1600	5064.0	1703.8	21632
Training	new_cases Prediction Formula - forward	Fit Generalized Forward Selection		0.1600	5064.0	1703.8	21632
Training	new_cases Prediction Formula - Backward	Fit Generalized Backward Elimination		-0.039	5631.9	1091.5	21632
Training	new_cases Predictor RF	Bootstrap Forest		0.6384	3322.6	801.76	21632
Training	Predicted new_cases NN T20	Neural		0.6416	3307.8	1115.7	21632
Training	Predicted new_cases NN B20	Neural		0.6417	3307.4	911.27	21632
Training	new_cases Predictor BT	Boosted Tree		0.5045	3889.3	1096.7	21632
Validation	new_cases Prediction Formula - General Regr	Fit Generalized Standard Least Squares		0.1792	5009.6	1690.0	7211
Validation	new_cases Prediction Formula - forward	Fit Generalized Forward Selection		0.1792	5009.6	1690.0	7211
Validation	new_cases Prediction Formula - Backward	Fit Generalized Backward Elimination		-0.039	5636.0	1091.4	7211
Validation	new_cases Predictor RF	Bootstrap Forest		0.6455	3292.1	792.41	7211
Validation	Predicted new_cases NN T20	Neural		0.6480	3280.5	1123.6	7211
Validation	Predicted new_cases NN B20	Neural		0.6458	3290.8	901.31	7211
Validation	new_cases Predictor BT	Boosted Tree		0.5254	3809.5	1067.1	7211
Test	new_cases Prediction Formula - General Regr	Fit Generalized Standard Least Squares		0.1795	5002.4	1675.1	7210
Test	new_cases Prediction Formula - forward	Fit Generalized Forward Selection		0.1795	5002.4	1675.1	7210
Test	new_cases Prediction Formula - Backward	Fit Generalized Backward Elimination		-0.039	5629.1	1091.0	7210
Test	new_cases Predictor RF	Bootstrap Forest		0.6774	3136.6	758.90	7210
Test	Predicted new_cases NN T20	Neural		0.6933	3058.2	1063.4	7210
Test	Predicted new_cases NN B20	Neural		0.6793	3127.4	863.11	7210
Test	new_cases Predictor BT	Boosted Tree		0.5387	3750.8	1057.9	7210

Table 14 Model Comparison for New Cases

Table 14 is the model comparison for the new cases. The best machine learning model to predict new case is the basic Neutral Net with the additional Tour 20. This model was able to predict 69.33% of the test model. The worst model for predicting the test data was the backward regression model which failed to predict the data.

As noted, the Neural Net with the additional Tour was able to predict 69.33% of the testing data set. For the selected model, we used Prediction tool. This tool does not forecast into the future but offers a prediction. Based on the following: Diabetes prevalence - 7.59, Hospital bed per thousand - 2.03, stringency index - 73.1, population - 430,000,000, cardiovascular death rate - 183, population over age 65 - 8.83%, life expectancy - age 75, median age - 30.9, and gdp per capita -13,200, the model determined that 16,487 new cases occurred. This number seems low, so we are not confident in this model even though this model had a largest R2.

5.6.5.1 Model Comparison – New Deaths

Model Comparison

Measures of Fit for new_deaths

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	new_deaths Prediction Formula - Linear	Fit Generalized Standard Least Squares		0.1038	105.97	37.022	21456
Training	new_deaths Predictor - RF	Bootstrap Forest		0.5769	72.808	15.831	21456
Training	Predicted new_deaths -NN	Neural		0.5168	77.811	20.700	21456
Training	Predicted new_deaths - NN L2 T20	Neural		0.6335	67.759	18.149	21456
Training	Predicted new_deaths - NN TLG T20	Neural		0.5831	72.276	22.051	21456
Training	Predicted new_deaths - NN L2	Neural		0.6042	70.421	19.005	21456
Training	new_deaths Predicted - Boosted	Boosted Tree		0.4179	85.401	24.420	21456
Training	Pred Formula new_deaths -Backward	Fit Least Squares		0.1038	105.97	37.026	21456
Training	Pred Formula new_deaths 2 - forward	Fit Least Squares		0.0000	111.93	37.026	21456
Validation	new_deaths Prediction Formula - Linear	Fit Generalized Standard Least Squares		0.0917	124.00	41.058	7152
Validation	new_deaths Predictor - RF	Bootstrap Forest		0.5244	89.727	18.246	7152
Validation	Predicted new_deaths -NN	Neural		0.4630	95.337	23.222	7152
Validation	Predicted new_deaths - NN L2 T20	Neural		0.5739	84.927	20.116	7152
Validation	Predicted new_deaths - NN TLG T20	Neural		0.5390	88.335	24.019	7152
Validation	Predicted new_deaths - NN L2	Neural		0.5421	88.041	21.261	7152
Validation	new_deaths Predicted - Boosted	Boosted Tree		0.3994	100.82	27.347	7152
Validation	Pred Formula new_deaths -Backward	Fit Least Squares		0.0918	123.99	41.064	7152
Validation	Pred Formula new_deaths 2 - forward	Fit Least Squares		-0.001	130.17	40.828	7152
Test	new_deaths Prediction Formula - Linear	Fit Generalized Standard Least Squares		0.1467	102.04	39.174	7152
Test	new_deaths Predictor - RF	Bootstrap Forest		0.6684	63.604	17.107	7152
Test	Predicted new_deaths -NN	Neural		0.5903	70.705	22.387	7152
Test	Predicted new_deaths - NN L2 T20	Neural		0.7247	57.958	19.572	7152
Test	Predicted new_deaths - NN TLG T20	Neural		0.6809	62.400	23.141	7152
Test	Predicted new_deaths - NN L2	Neural		0.6753	62.946	20.820	7152
Test	new_deaths Predicted - Boosted	Boosted Tree		0.5025	77.908	26.105	7152
Test	Pred Formula new_deaths -Backward	Fit Least Squares		0.1467	102.03	39.168	7152
Test	Pred Formula new_deaths 2 - forward	Fit Least Squares		-0.001	110.50	39.267	7152

Table 15 Model Comparison for New Deaths

Table 15 is the model comparison for the new death. The best machine learning model to predict new deaths is the basic Neural Net with two levels and Tour 20 which has the highest R^2 of 72.47%. This model was able to predict 72.47% of the test model. The worst model for predicting the test data was the Forward Regression model which was unable to create a prediction.

6.1 Forecasting / Prediction – New Cases

For the forecasting portion of the project, we used JMP Pro and their Time Series tool. Please see Appendix F for Excel table of forecast generated. The forecasting was completed on total new cases. We elected not to transform the data. The data was pretty much a straight line with little trend. We used the square root and there was a slight improvement, but we elected to forecast the data as is. For November 26, 2020, the program forecasted 10,386,105. On this day in the United States there were 12,877,783 cases. This is according to the John Hopkins Coronavirus tracker. Our forecasting was off more than 2.5 million cases.

6.2 Forecasting / Prediction – New Deaths

We also forecasted the total number of deaths using the same programming environment. The forecast was completed using total deaths. Please see Appendix G for Excel table of forecast generated. For November 26, 2020, the program forecasted 248,738. On this day in the United States there were 263,394 year to date deaths. This is according to the John Hopkins Coronavirus tracker. We forecasted 14,656 fewer deaths. The recent surge in the number of infections from the Covid-19 likely contributed to the lower number of cases forecasted.

7 Discussion and Conclusion

In this analysis, we used descriptive analysis to search for relationships in the Covid-19 dataset between the conditions and new cases or new deaths. Based on the conditions, the correlations were slightly significant. The most correlated relationship is between new cases and new death which was almost 80%. When we look at the correlations for new cases, stringency index has a positve correlation of 13.45%, and Diabetes prevlaence has a positive correlation of 12.79%. In both cases (new and death), population was also an item that has a positive correlation with a value of 35.54% and 25.59% respectively. In addition in both cases (new and death), if a male was a smoker, there was an inverse relationship with a value of (9.61%) and (11.69%) respectively. We can see that as the number of cases increases, we can expect an increase in the number of new deaths. We can conclude that only being infected with Covid-19 will contribute to the chances of dying from Covid-19. There are two items that I did not expect 1) effects due to stringency index and 2) effects of smoking if you are male. As mentioned, stringency index has a slightly positive effect on new cases. Recall stringency is how strict the government is on closing school, business,

etc. We would expect that as schools and business close, the new of cases decrease. The analysis indicates a positive relationship. As the stringency index increase, the number of cases increase. This could be caused by the United States being an outlier. It has the largest number of cases with an index of 57. The other unexpected item is males who smoke. Since Covid-19 is a virus that causes respiratory concerns like shortness of breath, being a smoker should have a positive relation. But the data tells us it has a negative relationship. From the one – sample hypothesis testing and ANOVA testing, we learned more about the population dataset. The ANOVA testing indicates the mean for the sample are different. The overall growth factors for new case and new deaths continue to decline. However, there continue to be spikes in death periodically. The clustering is subgroup based on the stringency index. Finally, using JPM, we have determined the basic Neural Net with two levels and Tour 20 was the best model with an R^2 of 72.47%. We also used JMP to forecast the number of cases and deaths. The program provided results which were less than actual number. It is likely due to the recent surge in cases.

As we conclude this project, we are hopeful that emergency use for a vaccine will be available as early as next month, December 2020. The FDA has authorized a monoclonal antibody for treatment and has scheduled an advisory committee meeting to discuss the Covid-19 vaccine candidates on December 10. We hope by mid-summer of 2021 a vast majority of the country will be vaccinated.

Appendix A – Description of Variables

Variable	Description	Type of Data	Type of Scale
iso_code	ISO 3166-1 alpha-3 – three-letter country codes	Character	Categorical
continent	Continent of the geographical location	Character	Categorical
location	Geographical location	Character	Categorical
date	Date of observation	Character (Date)	Interval
total_cases	Total confirmed cases of COVID-19	Double	Ratio
new_cases	New confirmed cases of COVID-19	Double	Ratio
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed)	Double	Ratio
total_deaths	Total deaths attributed to COVID-19	Double	Ratio
new_deaths	New deaths attributed to COVID-19	Double	Ratio
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed)	Double	Ratio
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people	Double	Ratio
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people	Double	Ratio
new_cases_smoothed_per_million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people	Double	Ratio
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people	Double	Ratio
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people	Double	Ratio
new_deaths_smoothed_per_million	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people	Double	Ratio
total_tests	Total tests for COVID-19	Logical (Double)	Categorical (Ratio)
new_tests	New tests for COVID-19	Logical (Double)	Categorical (Ratio)
new_tests_smoothed	New tests for COVID-19 (7-day smoothed).	Logical (Double)	Categorical (Ratio)
total_tests_per_thousand	Total tests for COVID-19 per 1,000 people	Logical (Double)	Categorical (Ratio)
new_tests_per_thousand	New tests for COVID-19 per 1,000 people	Logical (Double)	Categorical (Ratio)
new_tests_smoothed_per_thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people	Logical (Double)	Categorical (Ratio)
tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)	Logical (Double)	Categorical (Ratio)
positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)	Logical (Double)	Categorical (Ratio)
tests_units	Units used by the location to report its testing data	Logical	Categorical
stringency_index	Government Response Stringency Index: 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)	Double	Ratio
population	Population in 2020	Double	Ratio
population_density	Number of people divided by land area, measured in square kilometers, most recent year available	Double	Ratio
median_age	Median age of the population, UN projection for 2020	Double	Ratio
aged_65_older	Share of the population that is 65 years and older, most recent year available	Double	Ratio
aged_70_older	Share of the population that is 70 years and older in 2015	Double	Ratio
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available	Double	Ratio
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010	Double	Ratio
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)	Double	Ratio
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017	Double	Ratio
female_smokers	Share of women who smoke, most recent year available	Double	Ratio
male_smokers	Share of men who smoke, most recent year available	Double	Ratio
handwashing_facilities	Share of the population with basic handwashing facilities on premises, most recent year available	Double	Ratio
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010	Double	Ratio
life_expectancy	Life expectancy at birth in 2019	Double	Ratio
human_development_index	Summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living	Double	Ratio

Appendix B – Code Used for Cleaning Data

```
library(readxl)
covid_data_R <- read_excel("covid-data_R.xlsx",
    col_types = c("text", "text", "text",
        "date", "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "text", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric"))
```

Figure 9 R Code

Figure 9 is R code used to change the data type while importing the dataset.

```
covid_data_R1$total_cases <- covid_data_R1$total_cases %>% replace_na(0)

covid_data_R1$new_cases <- covid_data_R1$new_cases %>% replace_na(0)

covid_data_R1$new_cases_smoothed <- covid_data_R1$new_cases_smoothed %>% replace_na(0)

covid_data_R1$total_deaths <- covid_data_R1$total_deaths %>% replace_na(0)

covid_data_R1$new_deaths <- covid_data_R1$new_deaths %>% replace_na(0)

covid_data_R1$new_deaths_smoothed <- covid_data_R1$new_deaths_smoothed %>% replace_na(0)

covid_data_R1$total_cases_per_million <- covid_data_R1$total_cases_per_million %>% replace_na(0)

covid_data_R1$new_cases_per_million <- covid_data_R1$new_cases_per_million %>% replace_na(0)

covid_data_R1$new_cases_smoothed_per_million<- covid_data_R1$new_cases_smoothed_per_million %>% replace_na(0)

covid_data_R1$total_deaths_per_million <- covid_data_R1$total_deaths_per_million %>% replace_na(0)

covid_data_R1$new_deaths_per_million <- covid_data_R1$new_deaths_per_million %>% replace_na(0)

covid_data_R1$new_deaths_smoothed_per_million <- covid_data_R1$new_deaths_smoothed_per_million %>% replace_na(0)
```

Figure 10 R Code

Figure 10 is code used to replace missing data with 0.

```
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL1      %>% drop_na(stringency_index)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(median_age)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(aged_65_older)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(gdp_per_capita)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(cardiovasc_death_rate)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(diabetes_prevalence)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(hospital_beds_per_thousand)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2      %>% drop_na(life_expectancy)
covid_data_R_clean_FINAL2        <-covid_data_R_clean_FINAL2 %>% drop_na(human_development_index)
```

Figure 11 R Code

Figure 11 is R code used to remove any countries with missing data.

Appendix C – Descriptive Summary

Column1	***total_cases***	*Column2*	***new_cases***	*Column3*	***new_cases_smoothed***
Mean	~~117787.0698~~	Mean	~~1515.075588~~	Mean	~~1475.789661~~
Standard Error	3852.951488	Standard Error	46.39050198	Standard Error	45.04258656
Median	3039.5	Median	33	Median	37.0715
Mode	0	Mode	0	Mode	0
Standard Deviation	575265.1409	Standard Deviation	6926.336535	Standard Deviation	6725.085946
Sample Variance	3.3093E+11	Sample Variance	47974137.8	Sample Variance	45226780.99
Kurtosis	89.51519807	Kurtosis	73.38422752	Kurtosis	75.0305306
Skewness	8.937542524	Skewness	7.981451084	Skewness	8.037113201
Range	8106752	Range	106155	Range	93750.571
Minimum	0	Minimum	-8261	Minimum	-552
Maximum	8106752	Maximum	97894	Maximum	93198.571
Sum	2625709360	Sum	33774065	Sum	32898303.11
Count	22292	Count	22292	Count	22292

Column4	***total_deaths***	*Column5*	***new_deaths***	*Column6*	***new_deaths_smoothed***
Mean	~~4440.223982~~	Mean	~~43.14099228~~	Mean	~~42.46111327~~
Standard Error	119.0436091	Standard Error	1.193582962	Standard Error	1.10507144
Median	67	Median	0	Median	0.571
Mode	0	Mode	0	Mode	0
Standard Deviation	17773.81283	Standard Deviation	178.207972	Standard Deviation	164.9927542
Sample Variance	315908422.4	Sample Variance	31758.08128	Sample Variance	27222.60893
Kurtosis	59.80816718	Kurtosis	95.06999932	Kurtosis	53.47833381
Skewness	7.106176697	Skewness	7.729562337	Skewness	6.416739591
Range	219289	Range	6846	Range	2947.286
Minimum	0	Minimum	-1918	Minimum	-232.143
Maximum	219289	Maximum	4928	Maximum	2715.143
Sum	98981473	Sum	961699	Sum	946543.137
Count	22292	Count	22292	Count	22292

Column7	***total_cases_per_million***	*Column8*	***new_cases_per_million***	*Column9*	***new_cases_smoothed_per_million***
Mean	~~1911.55781~~	Mean	~~24.50299789~~	Mean	~~23.55026153~~
Standard Error	25.32485486	Standard Error	0.409481308	Standard Error	0.340045777
Median	268.131	Median	2.7025	Median	3.082
Mode	0	Mode	0	Mode	0
Standard Deviation	3781.128895	Standard Deviation	61.13762999	Standard Deviation	50.77055404
Sample Variance	14296935.72	Sample Variance	3737.8098	Sample Variance	2577.649157
Kurtosis	15.22304042	Kurtosis	205.0359393	Kurtosis	32.74523802
Skewness	3.482317821	Skewness	5.010509544	Skewness	4.404200777
Range	34987.068	Range	4105.127	Range	990.062
Minimum	0	Minimum	-2212.545	Minimum	-269.978
Maximum	34987.068	Maximum	1892.582	Maximum	720.084
Sum	42612446.69	Sum	546220.829	Sum	524982.43
Count	22292	Count	22292	Count	22292

Column10	***total_deaths_per_million***	*Column11*	***new_deaths_per_million***	*Column12*	***new_deaths_smoothed_per_million***
Mean	72.53560246	Mean	0.667015835	Mean	0.654564777
Standard Error	1.040827607	Standard Error	0.016962414	Standard Error	0.011928414
Median	5.4835	Median	0	Median	0.052
Mode	0	Mode	0	Mode	0
Standard Deviation	155.400825	Standard Deviation	2.532574162	Standard Deviation	1.78097248
Sample Variance	24149.41642	Sample Variance	6.413931888	Sample Variance	3.171862974
Kurtosis	7.853104085	Kurtosis	2567.107546	Kurtosis	72.89888161
Skewness	2.855843019	Skewness	33.76499566	Skewness	6.664262975
Range	893.904	Range	283.283	Range	43.378
Minimum	0	Minimum	-67.901	Minimum	-9.678
Maximum	893.904	Maximum	215.382	Maximum	33.7
Sum	1616963.65	Sum	14869.117	Sum	14591.558
Count	22292	Count	22292	Count	22292

Column13	***stringency_index***	*Column14*	***population***	*Column15*	***population_density***
Mean	55.42609905	Mean	77992229.12	Mean	132.719829
Standard Error	0.18188477	Standard Error	1524508.111	Standard Error	1.130121303
Median	60.19	Median	18776707	Median	81.721
Mode	0	Mode	43851043	Mode	17.348
Standard Deviation	27.15631586	Standard Deviation	227616770.1	Standard Deviation	168.7328254
Sample Variance	737.465491	Sample Variance	5.18094E+16	Sample Variance	28470.76636
Kurtosis	-0.679319273	Kurtosis	27.93558048	Kurtosis	19.51102176
Skewness	-0.565983971	Skewness	5.288195132	Skewness	3.710783062
Range	100	Range	1439225434	Range	1263.056
Minimum	0	Minimum	98340	Minimum	1.98
Maximum	100	Maximum	1439323774	Maximum	1265.036
Sum	1235558.6	Sum	1.7386E+12	Sum	2958590.428
Count	22292	Count	22292	Count	22292

Column16	***median_age***	*Column17*	***aged_65_older***	*Column18*	***aged_70_older***
Mean	32.71597883	Mean	10.45415409	Mean	6.74396649
Standard Error	0.059054082	Standard Error	0.041059505	Standard Error	0.028986336
Median	32.6	Median	8.552	Median	5.331
Mode	38.7	Mode	6.211	Mode	3.053
Standard Deviation	8.817073061	Standard Deviation	6.130391739	Standard Deviation	4.327806556
Sample Variance	77.74077737	Sample Variance	37.58170287	Sample Variance	18.72990959
Kurtosis	-1.066576383	Kurtosis	-1.265318178	Kurtosis	-1.129575885
Skewness	-0.247380464	Skewness	0.355654843	Skewness	0.484566026
Range	32.8	Range	20.853	Range	14.932
Minimum	15.1	Minimum	2.168	Minimum	1.308
Maximum	47.9	Maximum	23.021	Maximum	16.24
Sum	729304.6	Sum	233044.003	Sum	150336.501
Count	22292	Count	22292	Count	22292

Column19	***gdp_per_capita***	*Column20*	***extreme_poverty***	*Column21*	***cardiovasc_death_rate***
Mean	20076.84683	Mean	8.861501884	Mean	248.3219692
Standard Error	121.1065452	Standard Error	0.105916409	Standard Error	0.778453911
Median	14600.861	Median	1.3	Median	242.648
Mode	13913.839	Mode	0.2	Mode	278.364
Standard Deviation	18081.81962	Standard Deviation	15.81385543	Standard Deviation	116.2271054
Sample Variance	326952200.8	Sample Variance	250.0780235	Sample Variance	13508.74003
Kurtosis	2.650011185	Kurtosis	3.648835337	Kurtosis	-0.491822423
Skewness	1.462544749	Skewness	2.119493582	Skewness	0.609189467
Range	93525.177	Range	71.3	Range	453.851
Minimum	752.788	Minimum	0.1	Minimum	85.998
Maximum	94277.965	Maximum	71.4	Maximum	539.849
Sum	447553069.6	Sum	197540.6	Sum	5535593.337
Count	22292	Count	22292	Count	22292

Column22	diabetes_prevalence	Column23	female_smokers	Column24	male_smokers
Mean	7.008882559	Mean	10.75102727	Mean	32.54841199
Standard Error	0.022543262	Standard Error	0.066931512	Standard Error	0.09630279
Median	6.73	Median	6.9	Median	30.9
Mode	7.11	Mode	1.9	Mode	24.7
Standard Deviation	3.365822944	Standard Deviation	9.99321318	Standard Deviation	14.3784936
Sample Variance	11.32876409	Sample Variance	99.86430965	Sample Variance	206.7410782
Kurtosis	3.903446445	Kurtosis	-0.583794665	Kurtosis	0.451680484
Skewness	1.524247903	Skewness	0.790736832	Skewness	0.6983717
Range	21.03	Range	35.2	Range	70.4
Minimum	0.99	Minimum	0.1	Minimum	7.7
Maximum	22.02	Maximum	35.3	Maximum	78.1
Sum	156242.01	Sum	239661.9	Sum	725569.2
Count	22292	Count	22292	Count	22292

Column25	hospital_beds_per_thousand	Column26	life_expectancy	Column27	human_development_index
Mean	3.030614929	Mean	74.42537771	Mean	0.742227795
Standard Error	0.015541976	Standard Error	0.043798885	Standard Error	0.000993103
Median	2.5	Median	75.93	Median	0.77
Mode	0.7	Mode	76.68	Mode	0.752
Standard Deviation	2.320495555	Standard Deviation	6.539395012	Standard Deviation	0.148275231
Sample Variance	5.384699619	Sample Variance	42.76368712	Sample Variance	0.021985544
Kurtosis	2.082148275	Kurtosis	-0.801321726	Kurtosis	-0.493358727
Skewness	1.361172858	Skewness	-0.499122496	Skewness	-0.642619043
Range	11.97	Range	22.71	Range	0.599
Minimum	0.3	Minimum	60.85	Minimum	0.354
Maximum	12.27	Maximum	83.56	Maximum	0.953
Sum	67558.468	Sum	1659090.52	Sum	16545.742
Count	22292	Count	22292	Count	22292

Appendix C – Correlation Results – New Cases

	new_cases	stringency_index	population	population_density
new_cases	1.00000000	0.13453421	0.358454780	0.03903539
stringency_index	0.13453421	1.00000000	0.069631647	0.05242361
population	0.35845478	0.06963165	1.000000000	0.18802896
population_density	0.03903539	0.05242361	0.188028962	1.00000000
median_age	0.03038981	-0.18097923	-0.005968705	-0.01803927

	median_age
new_cases	0.030389814
stringency_index	-0.180979228
population	-0.005968705
population_density	-0.018039275
median_age	1.000000000

	new_cases	aged_65_older	aged_70_older	gdp_per_capita
new_cases	1.0000000000	0.009099573	-0.0008257297	0.0578026
aged_65_older	0.0090995732	1.000000000	0.9948036843	0.6859236
aged_70_older	-0.0008257297	0.994803684	1.0000000000	0.6708640
gdp_per_capita	0.0578025963	0.685923638	0.6708639978	1.0000000
extreme_poverty	-0.0274067179	-0.579444325	-0.5534792249	-0.4899173
cardiovasc_death_rate	-0.0640380442	-0.309284447	-0.2944734310	-0.5190485

	extreme_poverty	cardiovasc_death_rate
new_cases	-0.02740672	-0.06403804
aged_65_older	-0.57944433	-0.30928445
aged_70_older	-0.55347922	-0.29447343
gdp_per_capita	-0.48991725	-0.51904846
extreme_poverty	1.00000000	0.09641103
cardiovasc_death_rate	0.09641103	1.00000000

	new_cases	diabetes_prevalence	female_smokers
new_cases	1.00000000	0.12797393	0.02344171
diabetes_prevalence	0.12797393	1.00000000	-0.14748198
female_smokers	0.02344171	-0.14748198	1.00000000
male_smokers	-0.09618105	0.18965523	0.12948048
hospital_beds_per_thousand	-0.05913028	-0.02126573	0.52337069
life_expectancy	0.02922976	0.18817870	0.57863505
human_development_index	0.06136950	0.19392224	0.65132495

	male_smokers	hospital_beds_per_thousand	life_expectancy
new_cases	-0.09618105	-0.05913028	0.02922976
diabetes_prevalence	0.18965523	-0.02126573	0.18817870
female_smokers	0.12948048	0.52337069	0.57863505
male_smokers	1.00000000	0.38324660	0.04217673
hospital_beds_per_thousand	0.38324660	1.00000000	0.42517936
life_expectancy	0.04217673	0.42517936	1.00000000
human_development_index	0.07851300	0.54868224	0.92197962

	human_development_index
new_cases	0.0613695
diabetes_prevalence	0.1939222
female_smokers	0.6513249
male_smokers	0.0785130
hospital_beds_per_thousand	0.5486822
life_expectancy	0.9219796
human_development_index	1.0000000

Table 7 R results from correlation with new deaths

Figure 7 is the R output for the correlation based on new deaths as the response variable.

Appendix D – Correlation Results – New Deaths

Correlation Results for New Deaths

	new_deaths	new_cases	stringency_index	population
new_deaths	1.00000000	0.75338734	0.16058508	0.255918350
new_cases	0.75338734	1.00000000	0.13453421	0.358454780
stringency_index	0.16058508	0.13453421	1.00000000	0.069631647
population	0.25591835	0.35845478	0.06963165	1.000000000
population_density	-0.01272394	0.03903539	0.05242361	0.188028962
median_age	0.05834249	0.03038981	-0.18097923	-0.005968705

	population_density	median_age
new_deaths	-0.01272394	0.058342486
new_cases	0.03903539	0.030389814
stringency_index	0.05242361	-0.180979228
population	0.18802896	-0.005968705
population_density	1.00000000	-0.018039275
median_age	-0.01803927	1.000000000

---

	new_deaths	aged_65_older	aged_70_older	gdp_per_capita
new_deaths	1.00000000	0.03812646	0.02982515	0.09092178
aged_65_older	0.03812646	1.00000000	0.99480368	0.68592364
aged_70_older	0.02982515	0.99480368	1.00000000	0.67086400
gdp_per_capita	0.09092178	0.68592364	0.67086400	1.00000000
extreme_poverty	-0.06351457	-0.57944433	-0.55347922	-0.48991725
cardiovasc_death_rate	-0.11511647	-0.30928445	-0.29447343	-0.51904846

	extreme_poverty	cardiovasc_death_rate
new_deaths	-0.06351457	-0.11511647
aged_65_older	-0.57944433	-0.30928445
aged_70_older	-0.55347922	-0.29447343
gdp_per_capita	-0.48991725	-0.51904846
extreme_poverty	1.00000000	0.09641103
cardiovasc_death_rate	0.09641103	1.00000000

---

	new_deaths	diabetes_prevalence	female_smokers
new_deaths	1.00000000	0.13626327	0.04556542
diabetes_prevalence	0.13626327	1.00000000	-0.14748198
female_smokers	0.04556542	-0.14748198	1.00000000
male_smokers	-0.11698192	0.18965523	0.12948048
hospital_beds_per_thousand	-0.06205263	-0.02126573	0.52337069
life_expectancy	0.07403156	0.18817870	0.57863505
human_development_index	0.10364462	0.19392224	0.65132495

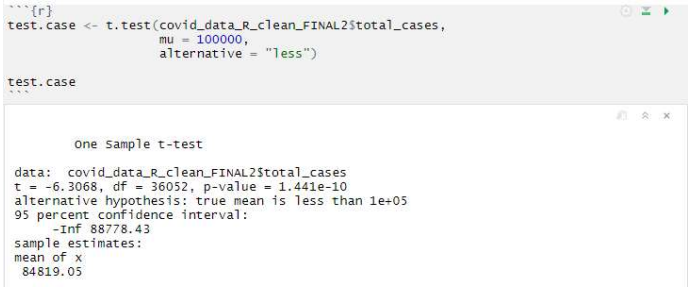
	male_smokers	hospital_beds_per_thousand	life_expectancy
new_deaths	-0.11698192	-0.06205263	0.07403156
diabetes_prevalence	0.18965523	-0.02126573	0.18817870
female_smokers	0.12948048	0.52337069	0.57863505
male_smokers	1.00000000	0.38324660	0.04217673
hospital_beds_per_thousand	0.38324660	1.00000000	0.42517936
life_expectancy	0.04217673	0.42517936	1.00000000
human_development_index	0.07851300	0.54868224	0.92197962

	human_development_index
new_deaths	0.1036446
diabetes_prevalence	0.1939222
female_smokers	0.6513249
male_smokers	0.0785130
hospital_beds_per_thousand	0.5486822
life_expectancy	0.9219796
human_development_index	1.0000000

Table 6 R results from correlation with new cases

Figure 6 is the R output for the correlation based on new cases as the response variable.

Appendix E - R Code and Output for One – Sample Hypothesis Testing



```
```{r}
test.case <- t.test(covid_data_R_clean_FINAL2$total_cases,
                    mu = 100000,
                    alternative = "less")

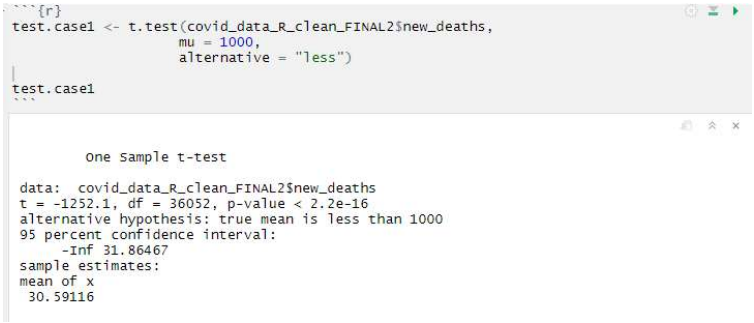
test.case
```

        One Sample t-test

data: covid_data_R_clean_FINAL2$total_cases
t = -6.3068, df = 36052, p-value = 1.441e-10
alternative hypothesis: true mean is less than 1e+05
95 percent confidence interval:
     -Inf 88778.43
sample estimates:
mean of x
 84819.05
```

Figure 20 One Sample Hypothesis Testing – New Cases

Figure 20 provides the results from our hypothesis testing for new cases using the R programming environment, we use the t.test function to complete the t-statistical test on the sample data. The t value is -6.03068 with a p-value of 1.441 e^{-10} with a .05 significant level or 95% confidence internal. Since the p-value is less than .05, we can reject the null and conclude that the population mean of the new cases if significantly less than 100,000 cases.



```
```{r}
test.case1 <- t.test(covid_data_R_clean_FINAL2$new_deaths,
                     mu = 1000,
                     alternative = "less")

test.case1
```

        One Sample t-test

data: covid_data_R_clean_FINAL2$new_deaths
t = -1252.1, df = 36052, p-value < 2.2e-16
alternative hypothesis: true mean is less than 1000
95 percent confidence interval:
     -Inf 31.86467
sample estimates:
mean of x
 30.59116
```

Figure 21 One Sample Hypothesis Testing – New Deaths

Figure 21 provides the results from our hypothesis testing for new deaths. The t value is -1252.1 with a p-value of 2.2 e^{-16} with a .05 significant level or 95% confidence internal. Since the p-value is less than .05, we can reject the null and conclude that the population mean of the new cases if significantly less than 1,000 cases.

Appendix F - R Code and Output for ANOVA testing

```
#ANOVA in R - new Cases

```{r}
oneway.test(new_cases ~continent,
            data = covid_data_R_clean_FINAL2,
            var.equal = TRUE  #assume equal variance)
            )
```

        One-way analysis of means

data: new_cases and continent
F = 208.36, num df = 5, denom df = 36047, p-value < 2.2e-16

```{r}
res_aov1 <- aov(new_cases ~continent,
                data = covid_data_R_clean_FINAL2
                )
summary(res_aov1)
```

                Df   Sum Sq   Mean Sq F value Pr(>F)
continent   5 3.092e+10 6.184e+09   208.4 <2e-16 ***
Residuals   36047 1.070e+12 2.968e+07

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 27 AVOVA Testing on New Cases

Figure 27 provides the ANOVA testing results of two methods using the R programming environment. In both methods, the p-value is 208.4 e^{-16} which is less than .05. So, we can reject the hypothesis that all means are equal. We must conclude that at least one continent is different than the others regarding the new cases.

Appendix G – Modeling using R

Forward Regression – New Cases using R

```
Step:  AIC=616164.4
new_cases ~ population

                           Df  Sum of Sq        RSS    AIC
+ stringency_index          1 9516872686 9.4374e+11 615805
+ human_development_index   1 6197398203 9.4706e+11 615931
+ cardiovasc_death_rate     1 5680427002 9.4758e+11 615951
+ gdp_per_capita            1 4453827079 9.4880e+11 615998
+ life_expectancy           1 2170838380 9.5109e+11 616084
+ aged_65_older             1 1738269795 9.5152e+11 616101
+ median_age                1 1266356934 9.5199e+11 616119
+ diabetes_prevalence       1  931097097 9.5233e+11 616131
+ hospital_beds_per_thousand 1  576070396 9.5268e+11 616145
<none>                                   9.5326e+11 616164
```

```
Call:
lm(formula = new_cases ~ ., data = test_grp)

Residuals:
   Min     1Q Median     3Q    Max
-17050  -1146   -507    379  81196

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.373e+03  6.101e+02  12.084  < 2e-16 ***
stringency_index           2.225e+01  1.007e+00  22.103  < 2e-16 ***
population                 1.140e-05  1.497e-07  76.137  < 2e-16 ***
median_age                -1.232e+02  1.195e+01 -10.308  < 2e-16 ***
aged_65_older              1.795e+02  1.284e+01  13.983  < 2e-16 ***
gdp_per_capita             3.912e-03  2.101e-03   1.862   0.0627 .
cardiovasc_death_rate     -1.784e+00  2.960e-01  -6.029 1.66e-09 ***
diabetes_prevalence        5.822e+01  9.082e+00   6.410 1.47e-10 ***
hospital_beds_per_thousand -1.942e+02  1.609e+01 -12.074  < 2e-16 ***
life_expectancy           -2.024e+02  1.177e+01 -17.205  < 2e-16 ***
human_development_index    1.291e+04  6.033e+02  21.404  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5040 on 36042 degrees of freedom
Multiple R-squared:  0.1682,    Adjusted R-squared:  0.1679
F-statistic: 728.7 on 10 and 36042 DF,  p-value: < 2.2e-16


Call:
lm(formula = new_cases ~ 1, data = test_grp)

Residuals:
   Min     1Q Median     3Q    Max
 -9351  -1090  -1059   -731  96804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1090.5       29.1   37.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5526 on 36052 degrees of freedom

Start:  AIC=621350
new_cases ~ 1

                           Df  Sum of Sq        RSS    AIC
+ population                1 1.4752e+11 9.5326e+11 616164
+ stringency_index          1 1.2613e+10 1.0882e+12 620936
+ cardiovasc_death_rate     1 4.3115e+09 1.0965e+12 621210
+ human_development_index   1 4.0831e+09 1.0967e+12 621218
+ diabetes_prevalence       1 3.8870e+09 1.0969e+12 621224
+ median_age                1 1.8519e+09 1.0989e+12 621291
+ life_expectancy           1 1.5566e+09 1.0992e+12 621301
+ hospital_beds_per_thousand 1 1.4293e+09 1.0993e+12 621305
+ gdp_per_capita            1 1.2407e+09 1.0995e+12 621311
```

```
Step:  AIC=616164.4
new_cases ~ population

                           Df  Sum of Sq        RSS    AIC
+ stringency_index          1 9516872686 9.4374e+11 615805
+ human_development_index   1 6197398203 9.4706e+11 615931
+ cardiovasc_death_rate     1 5680427002 9.4758e+11 615951
+ gdp_per_capita            1 4453827079 9.4880e+11 615998
+ life_expectancy           1 2170838380 9.5109e+11 616084
+ aged_65_older             1 1738269795 9.5152e+11 616101
+ median_age                1 1266356934 9.5199e+11 616119
+ diabetes_prevalence       1  931097097 9.5233e+11 616131
+ hospital_beds_per_thousand 1  576070396 9.5268e+11 616145
<none>                                   9.5326e+11 616164

Step:  AIC=615804.7
new_cases ~ population + stringency_index

                           Df  Sum of Sq        RSS    AIC
+ human_development_index   1 9649025491 9.3409e+11 615436
+ gdp_per_capita            1 7496774482 9.3625e+11 615519
+ cardiovasc_death_rate     1 6786001681 9.3696e+11 615547
+ aged_65_older             1 4691995373 9.3905e+11 615627
+ life_expectancy           1 4026675739 9.3972e+11 615653
+ median_age                1 3515922257 9.4023e+11 615672
+ diabetes_prevalence       1  608818106 9.4313e+11 615783
<none>                                   9.4374e+11 615805
+ hospital_beds_per_thousand 1   26774448 9.4372e+11 615806

Step:  AIC=615436.2
new_cases ~ population + stringency_index + human_development_index

                           Df  Sum of Sq        RSS    AIC
+ life_expectancy           1 4711096406 9.2938e+11 615256
+ hospital_beds_per_thousand 1 4451721829 9.2964e+11 615266
+ median_age                1 2982371964 9.3111e+11 615323
+ cardiovasc_death_rate     1 1814542003 9.3228e+11 615368
+ gdp_per_capita            1  420108829 9.3367e+11 615422
<none>                                   9.3409e+11 615436
+ aged_65_older             1   24884086 9.3407e+11 615437
+ diabetes_prevalence       1    9102574 9.3408e+11 615438

Step:  AIC=615255.9
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy

                           Df  Sum of Sq        RSS    AIC
+ hospital_beds_per_thousand 1 6303876459 9.2308e+11 615013
+ cardiovasc_death_rate     1 4928472765 9.2445e+11 615066
+ median_age                1 1706652110 9.2768e+11 615192
+ gdp_per_capita            1  363013124 9.2902e+11 615244
<none>                                   9.2938e+11 615256
+ aged_65_older             1   49235326 9.2933e+11 615256
+ diabetes_prevalence       1   14195992 9.2937e+11 615257


Step:  AIC=615012.5
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy + hospital_beds_per_thousand

                          Df  Sum of Sq        RSS    AIC
+ aged_65_older            1 2563513046 9.2051e+11 614914
+ cardiovasc_death_rate    1 2329285004 9.2075e+11 614923
+ diabetes_prevalence      1  268782324 9.2281e+11 615004
<none>                                  9.2308e+11 615013
+ median_age               1     646992 9.2308e+11 615014
+ gdp_per_capita           1     420067 9.2308e+11 615014
```

```
Step:  AIC=614914.2
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy + hospital_beds_per_thousand + aged_65_older

                         Df  Sum of Sq        RSS    AIC
+ median_age              1 2884934361 9.1763e+11 614803
+ cardiovasc_death_rate   1 1795878544 9.1872e+11 614846
+ gdp_per_capita          1   89851056 9.2042e+11 614913
+ diabetes_prevalence     1   72981582 9.2044e+11 614913
<none>                                 9.2051e+11 614914

Step:  AIC=614803.1
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy + hospital_beds_per_thousand + aged_65_older +
    median_age

                         Df Sum of Sq        RSS    AIC
+ cardiovasc_death_rate   1 851861316 9.1678e+11 614772
+ diabetes_prevalence     1 723376850 9.1691e+11 614777
+ gdp_per_capita          1 257231973 9.1737e+11 614795
<none>                                 9.1763e+11 614803

Step:  AIC=614771.6
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy + hospital_beds_per_thousand + aged_65_older +
    median_age + cardiovasc_death_rate

                       Df  Sum of Sq        RSS    AIC
+ diabetes_prevalence   1 1028627214 9.1575e+11 614733
+ gdp_per_capita        1   72725789 9.1670e+11 614771
<none>                               9.1678e+11 614772

Step:  AIC=614733.1
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy + hospital_beds_per_thousand + aged_65_older +
    median_age + cardiovasc_death_rate + diabetes_prevalence

                   Df Sum of Sq        RSS    AIC
+ gdp_per_capita    1  88036732 9.1566e+11 614732
<none>                          9.1575e+11 614733

Step:  AIC=614731.7
new_cases ~ population + stringency_index + human_development_index +
    life_expectancy + hospital_beds_per_thousand + aged_65_older +
    median_age + cardiovasc_death_rate + diabetes_prevalence +
    gdp_per_capita
```

Backward Regression – New Cases using R

```
Start:  AIC=614731.7
new_cases ~ stringency_index + population + median_age + aged_65_older +
    gdp_per_capita + cardiovasc_death_rate + diabetes_prevalence +
    hospital_beds_per_thousand + life_expectancy + human_development_index

                             Df  Sum of Sq        RSS    AIC
<none>                                     9.1566e+11 614732
- gdp_per_capita              1 8.8037e+07 9.1575e+11 614733
- cardiovasc_death_rate       1 9.2357e+08 9.1658e+11 614766
- diabetes_prevalence         1 1.0439e+09 9.1670e+11 614771
- median_age                  1 2.6993e+09 9.1836e+11 614836
- hospital_beds_per_thousand  1 3.7036e+09 9.1936e+11 614875
- aged_65_older               1 4.9674e+09 9.2063e+11 614925
- life_expectancy             1 7.5201e+09 9.2318e+11 615025
- human_development_index     1 1.1639e+10 9.2730e+11 615185
- stringency_index            1 1.2412e+10 9.2807e+11 615215
- population                  1 1.4727e+11 1.0629e+12 620107

Call:
lm(formula = new_cases ~ stringency_index + population + median_age +
    aged_65_older + gdp_per_capita + cardiovasc_death_rate +
    diabetes_prevalence + hospital_beds_per_thousand + life_expectancy +
    human_development_index, data = test_grp)

Residuals:
   Min     1Q Median     3Q    Max
-17050  -1146   -507    379  81196

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.373e+03  6.101e+02  12.084  < 2e-16 ***
stringency_index            2.225e+01  1.007e+00  22.103  < 2e-16 ***
population                  1.140e-05  1.497e-07  76.137  < 2e-16 ***
median_age                 -1.232e+02  1.195e+01 -10.308  < 2e-16 ***
aged_65_older               1.795e+02  1.284e+01  13.983  < 2e-16 ***
gdp_per_capita              3.912e-03  2.101e-03   1.862   0.0627 .
cardiovasc_death_rate      -1.784e+00  2.960e-01  -6.029 1.66e-09 ***
diabetes_prevalence         5.822e+01  9.082e+00   6.410 1.47e-10 ***
hospital_beds_per_thousand -1.942e+02  1.609e+01 -12.074  < 2e-16 ***
life_expectancy            -2.024e+02  1.177e+01 -17.205  < 2e-16 ***
human_development_index     1.291e+04  6.033e+02  21.404  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5040 on 36042 degrees of freedom
Multiple R-squared:  0.1682,    Adjusted R-squared:  0.1679
F-statistic: 728.7 on 10 and 36042 DF,  p-value: < 2.2e-16
```

Forward Regression – New Deaths using R

```
Call:
lm(formula = new_deaths ~ ., data = test_grp)

Residuals:
    Min      1Q  Median      3Q     Max
-1945.8   -19.2    -8.3     3.7  4319.1

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.300e+01  1.203e+01   1.912   0.0559 .
new_cases                   1.904e-02  1.037e-04 183.647  < 2e-16 ***
stringency_index            3.366e-01  1.994e-02  16.882  < 2e-16 ***
population                 -1.986e-09  3.175e-09  -0.626   0.5316
median_age                 -5.013e-01  2.355e-01  -2.129   0.0333 *
aged_65_older               1.484e+00  2.533e-01   5.859 4.71e-09 ***
gdp_per_capita             -2.720e-04  4.135e-05  -6.578 4.83e-11 ***
cardiovasc_death_rate      -5.259e-02  5.827e-03  -9.026  < 2e-16 ***
diabetes_prevalence         3.882e-01  1.788e-01   2.171   0.0300 *
hospital_beds_per_thousand -2.993e+00  3.172e-01  -9.436  < 2e-16 ***
life_expectancy            -1.298e+00  2.325e-01  -5.584 2.37e-08 ***
human_development_index     1.231e+02  1.195e+01  10.304  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99.18 on 36041 degrees of freedom
Multiple R-squared:  0.5449,    Adjusted R-squared:  0.5448
F-statistic:  3923 on 11 and 36041 DF,  p-value: < 2.2e-16


Call:
lm(formula = new_deaths ~ 1, data = test_grp)

Residuals:
    Min      1Q  Median      3Q     Max
-1948.6   -30.6   -30.6   -24.6  4897.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.5912     0.7742   39.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 147 on 36052 degrees of freedom

Start:  AIC=359844.2
new_deaths ~ 1


                            Df  Sum of Sq       RSS    AIC
+ new_cases                  1 416714988 362398866 332251
+ population                 1  52948459 726165394 357309
+ stringency_index           1  13476909 765636944 359217
+ cardiovasc_death_rate      1   8347053 770766801 359458
+ human_development_index    1   6279380 772834474 359554
+ life_expectancy            1   3917218 775196636 359664
+ median_age                 1   2993406 776120448 359707
+ aged_65_older              1   2529408 776584446 359729
+ diabetes_prevalence        1   1987266 777126588 359754
+ gdp_per_capita             1   1637040 777476814 359770
+ hospital_beds_per_thousand 1   1051747 778062106 359797
<none>                                   779113854 359844

Step:  AIC=332250.8
new_deaths ~ new_cases

                            Df  Sum of Sq       RSS    AIC
+ cardiovasc_death_rate      1   2607345 359791521 331992
+ stringency_index           1   2233730 360165136 332030
+ human_development_index    1   1600095 360798771 332093
+ life_expectancy            1   1469957 360928909 332106
+ aged_65_older              1    939418 361459448 332159
+ median_age                 1    798534 361600331 332173
+ gdp_per_capita             1    353392 362045473 332218
+ hospital_beds_per_thousand 1     84184 362314682 332244
+ population                 1     44547 362354318 332248
+ diabetes_prevalence        1     38814 362360052 332249
<none>                                   362398866 332251
```

```
Step:  AIC=331992.5
new_deaths ~ new_cases + cardiovasc_death_rate

                           Df Sum of Sq       RSS    AIC
+ stringency_index          1   2642045 357149476 331729
+ human_development_index   1    355962 359435558 331959
+ aged_65_older             1    188604 359602916 331976
+ life_expectancy           1    167537 359623983 331978
+ median_age                1    138512 359653009 331981
+ hospital_beds_per_thousand 1     80554 359710967 331986
+ diabetes_prevalence       1     70362 359721158 331987
+ gdp_per_capita            1     40159 359751361 331990
<none>                                  359791521 331992
+ population                1     16274 359775246 331993

Step:  AIC=331728.8
new_deaths ~ new_cases + cardiovasc_death_rate + stringency_index

                           Df Sum of Sq       RSS    AIC
+ human_development_index   1    805550 356343926 331649
+ aged_65_older             1    742745 356406731 331656
+ median_age                1    557672 356591804 331674
+ life_expectancy           1    441103 356708373 331686
+ diabetes_prevalence       1     31435 357118041 331728
<none>                                  357149476 331729
+ population                1     14385 357135091 331729
+ gdp_per_capita            1      8769 357140707 331730
+ hospital_beds_per_thousand 1       967 357148509 331731

Step:  AIC=331649.3
new_deaths ~ new_cases + cardiovasc_death_rate + stringency_index +
    human_development_index

                           Df Sum of Sq       RSS    AIC
+ gdp_per_capita            1    470115 355873811 331604
+ hospital_beds_per_thousand 1    406038 355937888 331610
+ aged_65_older             1    113393 356230534 331640
+ life_expectancy           1    108607 356235319 331640
<none>                                  356343926 331649
+ population                1      5832 356338094 331651
+ diabetes_prevalence       1      3283 356340643 331651
+ median_age                1      2058 356341868 331651

Step:  AIC=331603.8
new_deaths ~ new_cases + cardiovasc_death_rate + stringency_index +
    human_development_index + gdp_per_capita

                           Df Sum of Sq       RSS    AIC
+ hospital_beds_per_thousand 1    592854 355280957 331546
+ life_expectancy           1    168848 355704963 331589
+ median_age                1     23959 355849852 331603
<none>                                  355873811 331604
+ aged_65_older             1     19396 355854415 331604
+ population                1     16620 355857191 331604
+ diabetes_prevalence       1      2894 355870917 331605


Step:  AIC=331545.6
new_deaths ~ new_cases + cardiovasc_death_rate + stringency_index +
    human_development_index + gdp_per_capita + hospital_beds_per_thousand

                       Df Sum of Sq       RSS    AIC
+ aged_65_older         1    283088 354997870 331519
+ life_expectancy       1    229544 355051413 331524
+ median_age            1     34374 355246583 331544
+ diabetes_prevalence   1     34151 355246806 331544
<none>                               355280957 331546
+ population            1     18596 355262361 331546

Step:  AIC=331518.9
new_deaths ~ new_cases + cardiovasc_death_rate + stringency_index +
    human_development_index + gdp_per_capita + hospital_beds_per_thousand +
    aged_65_older

                       Df Sum of Sq       RSS    AIC
+ life_expectancy       1    366443 354631426 331484
+ median_age            1    109621 354888249 331510
<none>                               354997870 331519
+ population            1     17549 354980321 331519
+ diabetes_prevalence   1      1813 354996056 331521

Step:  AIC=331483.7
new_deaths ~ new_cases + cardiovasc_death_rate + stringency_index +
    human_development_index + gdp_per_capita + hospital_beds_per_thousand +
    aged_65_older + life_expectancy

                       Df Sum of Sq       RSS    AIC
+ median_age            1   29839.6 354601587 331483
+ diabetes_prevalence   1   24167.5 354607259 331483
<none>                               354631426 331484
+ population            1    7690.4 354623736 331485
```

Backward Regression – New Deaths using R

```
Start:  AIC=331481.6
new_deaths ~ new_cases + stringency_index + population + median_age +
    aged_65_older + gdp_per_capita + cardiovasc_death_rate +
    diabetes_prevalence + hospital_beds_per_thousand + life_expectancy +
    human_development_index

                             Df Sum of Sq       RSS    AIC
- population                  1      3849 354556112 331480
<none>                                   354552263 331482
- median_age                  1     44588 354596850 331484
- diabetes_prevalence         1     46356 354598619 331484
- life_expectancy             1    306757 354859020 331511
- aged_65_older               1    337667 354889929 331514
- gdp_per_capita              1    425666 354977928 331523
- cardiovasc_death_rate       1    801389 355353651 331561
- hospital_beds_per_thousand  1    875841 355428104 331569
- human_development_index     1   1044417 355596679 331586
- stringency_index            1   2803540 357355803 331764
- new_cases                   1 331779504 686331767 355293

Step:  AIC=331480
new_deaths ~ new_cases + stringency_index + median_age + aged_65_older +
    gdp_per_capita + cardiovasc_death_rate + diabetes_prevalence +
    hospital_beds_per_thousand + life_expectancy + human_development_index

                             Df Sum of Sq       RSS    AIC
<none>                                   354556112 331480
- diabetes_prevalence         1     45475 354601587 331483
- median_age                  1     51147 354607259 331483
- life_expectancy             1    307799 354863911 331509
- aged_65_older               1    352455 354908566 331514
- gdp_per_capita              1    421857 354977969 331521
- cardiovasc_death_rate       1    798277 355354389 331559
- hospital_beds_per_thousand  1    873609 355429721 331567
- human_development_index     1   1067150 355623262 331586
- stringency_index            1   2812345 357368457 331763
- new_cases                   1 384165816 738721927 357943

Call:
lm(formula = new_deaths ~ new_cases + stringency_index + median_age +
    aged_65_older + gdp_per_capita + cardiovasc_death_rate +
    diabetes_prevalence + hospital_beds_per_thousand + life_expectancy +
    human_development_index, data = test_grp)

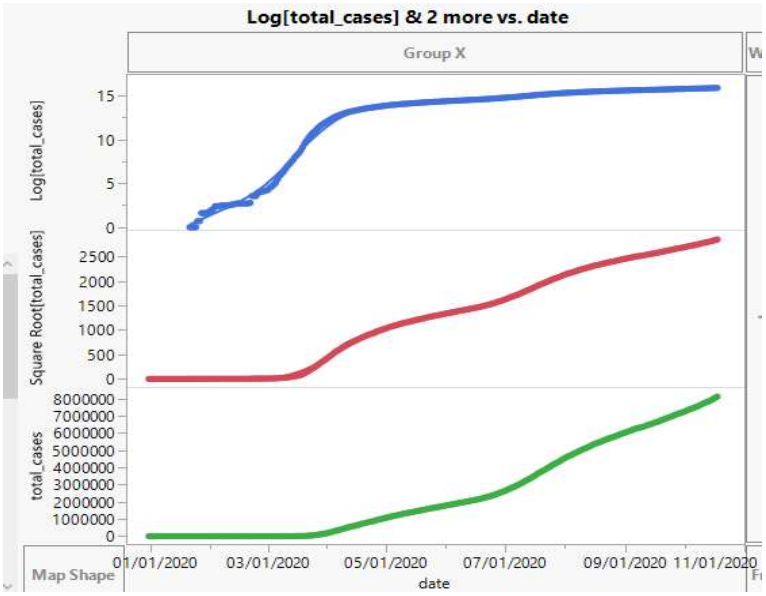
Residuals:
    Min      1Q  Median      3Q     Max
-1945.8   -19.2    -8.4     3.7  4319.1

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.309e+01  1.203e+01   1.919   0.0550 .
new_cases                    1.901e-02  9.620e-05 197.616  < 2e-16 ***
stringency_index             3.370e-01  1.993e-02  16.908  < 2e-16 ***
median_age                  -5.280e-01  2.315e-01  -2.280   0.0226 *
aged_65_older                1.504e+00  2.513e-01   5.986 2.18e-09 ***
gdp_per_capita              -2.693e-04  4.112e-05  -6.549 5.89e-11 ***
cardiovasc_death_rate       -5.245e-02  5.822e-03  -9.008  < 2e-16 ***


diabetes_prevalence          3.842e-01  1.787e-01   2.150   0.0316 *
hospital_beds_per_thousand -2.988e+00  3.171e-01  -9.424  < 2e-16 ***
life_expectancy            -1.300e+00  2.324e-01  -5.594 2.24e-08 ***
human_development_index     1.238e+02  1.189e+01  10.415  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.18 on 36042 degrees of freedom
Multiple R-squared:  0.5449,    Adjusted R-squared:  0.5448
F-statistic:  4316 on 10 and 36042 DF,  p-value: < 2.2e-16
```

Appendix H – Forecasting using JMP – New Cases



date	total_cases	total_cases (Lower 95%)	total_cases (Upper 95%)
10/19/2020	8,163,860.662	8,153,086.186	8,174,635.138
10/20/2020	8,222,340.769	8,202,064.403	8,242,617.134
10/21/2020	8,280,820.875	8,251,293.143	8,310,348.607
10/22/2020	8,339,300.982	8,300,120.135	8,378,481.829
10/23/2020	8,397,781.088	8,348,408.342	8,447,153.834
10/24/2020	8,456,261.194	8,396,125.877	8,516,396.512
10/25/2020	8,514,741.301	8,443,273.228	8,586,209.374
10/26/2020	8,573,221.407	8,489,862.324	8,656,580.491
10/27/2020	8,631,701.514	8,535,909.071	8,727,493.957
10/28/2020	8,690,181.620	8,581,430.378	8,798,932.863
10/29/2020	8,748,661.727	8,626,442.905	8,870,880.549
10/30/2020	8,807,141.833	8,670,962.551	8,943,321.115
10/31/2020	8,865,621.940	8,715,004.266	9,016,239.613
11/1/2020	8,924,102.046	8,758,582.011	9,089,622.082
11/2/2020	8,982,582.153	8,801,708.788	9,163,455.517
11/3/2020	9,041,062.259	8,844,396.703	9,237,727.816
11/4/2020	9,099,542.366	8,886,657.033	9,312,427.698
11/5/2020	9,158,022.472	8,928,500.302	9,387,544.643
11/6/2020	9,216,502.579	8,969,936.343	9,463,068.814
11/7/2020	9,274,982.685	9,010,974.367	9,538,991.003
11/8/2020	9,333,462.792	9,051,623.018	9,615,302.566
11/9/2020	9,391,942.898	9,091,890.421	9,691,995.375
11/10/2020	9,450,423.005	9,131,784.234	9,769,061.776
11/11/2020	9,508,903.111	9,171,311.684	9,846,494.538

11/12/2020	9,567,383.218	9,210,479.609	9,924,286.826
11/13/2020	9,625,863.324	9,249,294.484	10,002,432.164
11/14/2020	9,684,343.431	9,287,762.456	10,080,924.405
11/15/2020	9,742,823.537	9,325,889.367	10,159,757.707
11/16/2020	9,801,303.643	9,363,680.780	10,238,926.507
11/17/2020	9,859,783.750	9,401,141.999	10,318,425.501
11/18/2020	9,918,263.856	9,438,278.085	10,398,249.628
11/19/2020	9,976,743.963	9,475,093.880	10,478,394.046
11/20/2020	10,035,224.069	9,511,594.015	10,558,854.124
11/21/2020	10,093,704.176	9,547,782.931	10,639,625.420
11/22/2020	10,152,184.282	9,583,664.888	10,720,703.677
11/23/2020	10,210,664.389	9,619,243.975	10,802,084.803
11/24/2020	10,269,144.495	9,654,524.126	10,883,764.865
11/25/2020	10,327,624.602	9,689,509.126	10,965,740.078
11/26/2020	10,386,104.708	9,724,202.619	11,048,006.797

Appendix G – Forecasting using JMP – New Cases

date	total_deaths	total_deaths (Lower 95%)	total_deaths (Upper 95%)
44,123.000	220,028.183	219,254.965	220,801.401
44,124.000	220,783.682	219,446.320	222,121.044
44,125.000	221,539.181	219,673.049	223,405.313
44,126.000	222,294.680	219,892.997	224,696.362
44,127.000	223,050.179	220,094.862	226,005.496
44,128.000	223,805.678	220,274.618	227,336.737
44,129.000	224,561.177	220,430.801	228,691.552
44,130.000	225,316.676	220,563.021	230,070.330
44,131.000	226,072.175	220,671.383	231,472.966
44,132.000	226,827.673	220,756.230	232,899.117
44,133.000	227,583.172	220,818.016	234,348.329
44,134.000	228,338.671	220,857.244	235,820.099
44,135.000	229,094.170	220,874.429	237,313.912
44,136.000	229,849.669	220,870.081	238,829.257
44,137.000	230,605.168	220,844.698	240,365.638
44,138.000	231,360.667	220,798.757	241,922.578
44,139.000	232,116.166	220,732.713	243,499.620
44,140.000	232,871.665	220,647.001	245,096.329
44,141.000	233,627.164	220,542.034	246,712.295
44,142.000	234,382.663	220,418.202	248,347.124
44,143.000	235,138.162	220,275.879	250,000.445
44,144.000	235,893.661	220,115.417	251,671.905
44,145.000	236,649.160	219,937.151	253,361.169

44,146.000	237,404.659	219,741.401	255,067.917
44,147.000	238,160.158	219,528.470	256,791.846
44,148.000	238,915.657	219,298.647	258,532.666
44,149.000	239,671.156	219,052.210	260,290.102
44,150.000	240,426.655	218,789.421	262,063.888
44,151.000	241,182.154	218,510.533	263,853.774
44,152.000	241,937.653	218,215.787	265,659.519
44,153.000	242,693.152	217,905.413	267,480.890
44,154.000	243,448.651	217,579.634	269,317.667
44,155.000	244,204.150	217,238.662	271,169.637
44,156.000	244,959.649	216,882.702	273,036.595
44,157.000	245,715.148	216,511.950	274,918.345
44,158.000	246,470.647	216,126.595	276,814.698
44,159.000	247,226.145	215,726.820	278,725.471
44,160.000	247,981.644	215,312.800	280,650.489
44,161.000	248,737.143	214,884.705	282,589.581