COVID-19 - Time Series Analysis and Correlations:

Prediction of cases, deaths and finding correlations between cases and four human development indexes

Vinicius Rech Verza
School of Computing, National College of Ireland
Dublin, Ireland
Email: x19176503@student.ncirl.ie

Abstract — **The COVID-19 virus firstly appeared in 2019. With no vaccine nor effective proven method of cure, it still strikes the population on a global level. This project is aimed at forecasting cases and deaths numbers of the virus by using Time Series Analysis (Single Exponential Smoothing) as well as analysing if there are significant correlations between cases and deaths with percentage of expenditure of GDP on healthcare, human development, education and inequality indexes (Pearson's Correlation Coefficient). The results of forecasting were satisfactorily accurate given the changing environment conditions. On the correlation coefficient tests, some were little to none correlated, some variables had low correlation and Inequality Index with Infection Ratio were moderately correlated (r = 0.531). There were no significant strongly correlated variables.**

Keywords — COVID-19, Coronavirus, Time Series Analysis, Machine Learning, Prediction, Correlation Analysis, Statistics.

I. Introduction

A. *Brief Background on COVID-19*

The SARS-CoV-2 virus, also known as COVID-19 or Coronavirus was firstly reported in Wuhan City, China in the very end of 2019 [1]. From the first case until now, the virus infected more than 13 million people and killed almost 580 thousand individuals. The majority of the countries are taking strong precautions such as social distancing, government-controlled quarantine and other measures while other countries are adopting softer approaches and a few not adopting precautions at all [2]. It is difficult to predict the eradication time of this disease as there are multiple factors that involve this outcome, but with the progress of the situation and collection of data happening worldwide, it is possible to try predicting if the virus spread is on the rise or decline.

B. *Motivation*

While there are several analyses available worldwide about the virus itself, its infection rate and mortality, most of them do not take into consideration some socio-economic factors, such as inequality of income, education index, human development index and percentage of GDP expended on health. It is aimed that with all these data collected it will be possible to find patterns between infection, death numbers and the indexes, bringing into light behavioral relations of the virus with those indexes.

C. *Aim of Research*

This project has the goal of analyzing if there are correlations on infection and death rates with these indexes as well as predicting if the virus spread is on the rise or decline based on the data collected to date. The insights/questions are as follow:

- Are infection and death ratios of COVID-19 somehow correlated with the Education Index of a country?
- Are infection and death ratios of COVID-19 somehow correlated with the Human Development Index of a country?
- Are infection and death ratios of COVID-19 somehow correlated with Income Inequality Index of a country?
- Are infection and death ratios of COVID-19 somehow correlated with the Percentage of GDP expended on Health of a country?
- What is the forecast of the virus for the next eight weeks? Are the infections on the rise or decline? Are there fewer people dying?

D. *Definition of Central Terms*

- Time Series Analysis: A statistical technique used to analyze patterns of data points taken in a period of time in order to forecast the future [3].
- Correlation Analysis: A statistical technique that can determine whether and how strongly related are pairs of variables [4].

II. RELATED WORK

A. Introduction

It is in this topic that the literature review is found, which goes into details about the virus, its appearance, behavior, fatalities and related, relevant studies that have been made about forecasting casualties and infections as well as correlations of indexes with the virus.

B. Background on COVID-19

SARS-CoV-2, commonly known as COVID-19, firstly appeared in a case traced back to November 17th, 2019 [5]. This virus originated in bats and it made the jump to humans in the city of Wuhan, China, in possibly one of the open-air wet markets located in the city. It is common to see banned animal species being sold in those wet markets but it is difficult to say that Wuhan markets were selling bats at the time of the outbreak. It is important to highlight that some viruses can change within species, especially in crowded conditions. They can even jump and infect human beings [6]. From this day until now, the disease continues to spread throughout the globe, infecting more than 13 million people and killing almost 580 thousand. According to the World Health Organization Website [7], the symptoms are fever, tiredness and cough, amongst some other less common like sore throat, conjunctivitis, headache, aches and pains, diarrhea, skin rashes, loss of taste and smell, discoloration of toes and fingers, difficulty breathing or shortness of breath, loss of movement or speech and chest pressure or pain. The virus actually kills people by causing the development of ARDS (Acute Respiratory Distress Syndrome) and making it difficult to breath, requiring supplemental oxygen, mechanical ventilation and ICU to keep the patient alive [8]. People with pre-existing conditions such as diabetes, high blood pressure and others as well as the elderly are more susceptible to complications caused by the COVID-19.

C. Background on COVID-19 Analyses and Forecasting

There are plenty of analyses online, from dashboards and graphical approaches to proper scientific papers about the virus, developed big tech companies such as Google, for example, to governments, NGOs, private individuals and health sector entities, like The World Health Organization and much more. Google [9] developed a special page that provides a lot of real-time information about the virus on a global scale. It gathers different types of data, like newsworthy, government statements, numbers of infections and deaths per country and an interactive map. On this website you can select the country of your choice and manipulate the data as you like, displaying total number of cases, deaths and even filtering by a given period. The World Health Organization [10] provides an online dashboard with details about the virus, divided by regions and countries about total cases and deaths, also with an interactive map. It is also possible to pick a country or a region and tune the data as you like, splitting as daily or weekly cases, for example.

There are also websites and scientific papers that approach attempts to prediction of cases and/or casualties. While there is a variety of methods and approaches, such as Linear Regression, ARIMA, Wavelet-based Forecasting, BeCoDiS and much more, the most commonly adopted one is the Exponential Smoothing (Time Series) methodology. The paper from Chakraborty and Ghosh [11] is about forecasting the virus behavior in five countries (Canada, France, India, South Korea and UK) using a combination (hybrid) of two different forecasting techniques, ARIMA and Wavelet. While the results were fairly precise, this method is overly complicated for such results, compared to what you can achieve with Exponential Smoothing.

This is the case of the paper written by Petropoulos and Makridakis [12] where they approached forecasting COVID-19 cases using Exponential Smoothing models to predict how the infection numbers were going to behave and the obtained results were considerably accurate comparing with real numbers from data after the analysis.

D. Background on COVID-19 Correlation Analyses

Lastly, a research was done on analyses of the virus in correlation with some socio-economical indexes. The United Nations published a book [13] that approaches the effects of the COVID-19 pandemic on human development. One of the analysis is the measurement of the impact of some nonpharmaceutical interventions on countries based on the human development index as well as how the pandemic will affect unemployment, the education systems and even how much more vulnerable are groups of people to Coronavirus based on their income. While there are interesting in-depth analyses on this book, no actual correlation coefficient was calculated for any of these scenarios. The paper from Barua [14] brings into light the possible economic implications of the pandemic. He considers how the short and long terms effects would be, international trade and capital flows, micro and macroeconomic effects, employment and income, effects of exchange rates and much more. This paper is very thorough on the grounds of economic impact of the pandemic but again, it does not have correlations calculated between some of the mentioned indexes with COVID-19 case numbers. Furthermore, there is another paper [15] that statistically and mathematically analyses how quarantine can suppress COVID-19 transmission in developing and poor countries. The authors developed a mathematical model named ‘Social Distancing SEIQR Model’ that can predict the transmissibility of the virus in the country of Bangladesh. They have also calculated some correlation coefficients but only to determine relationships between five cases on lockdown situation, but those are not related to any human development indexes.

In conclusion, it is safe to say that there are very good, detailed analyses and forecasting methodologies around but when it comes to compare correlations between human development indexes to cases of COVID-19, there is not much out there. Some good papers and books were found on the matter of social-economical impact of the pandemic but they all lack the proper statistical correlation coefficient analysis.

III. METHODOLOGY

A. *Intro and Data Sources*

In order to perform all the proposed analyses, the necessary data was collected from reliable sources. A total of five data-sets were downloaded.

The first data-set was taken from The European Centre for Disease Prevention and Control [16] and contained data about the total number of cases and deaths of COVID-19, on a daily basis, separated by country and continent, from December 31st to May 31st. It consisted of 20503 rows and 11 variables, which 4 were nominal, 3 numerical and 4 were ordinal.

The second data-set was downloaded from The World Health Organization website [17] and contained data about the Health Expenditure as Percentage of GDP, separated by country, from 2000 to 2016. The data had 213 rows and 11 variables, of which 1 was nominal, 1 was ordinal and 9 were numerical.

The other three data-sets were taken from The United Nations website [18] and contained information of Human Development, Education and Inequality Indexes. As all three data-sets came from the same source, they contained the same structure: country name and indexes, from 1990 to 2018, which 1 variable was nominal and all others were numerical. The data-sets had between 208 and 214 rows and 31 variables.

B. *Data Preparation*

The software used for project were Microsoft Excel, R Studio, Tableau and JASP. The proposed solution was to clean the first data-set (COVID-19 cases) and keep only relevant data and to combine the remaining four data-sets into one, also discarding irrelevant data for this project. For the sake of understanding, the data-set with COVID-19 cases will be called 'Corona' and the other data-sets that will be merged into one will be called 'Indexes'.

The process of tidying the Corona data-set was performed as follows:

1 – All instances of December 31st, 2019 were removed, so the analysis takes only into consideration from January 1st to May 31st;

2 – The variables 'day', 'month', year', 'geold' and 'countryterritoryCode' were removed as they are irrelevant for this project;

3 – Renamed variables: 'data'Rep' to 'Date', 'countriesAndTerritories' to 'Country', 'popData2018' to 'Population' and 'continentExp' to 'Continent';

4 – Reordered variables as: Continent, Country, Population, Data, Cases, Deaths;

5 – Deleted all instances for 'Cases_on_international_conveyance_Japan';

6 – Deleted countries with black population values: Anguilla, Bonaire, Eritrea, Falklands, Western Sahara;

7 – Searched for negative numbers (wrong values) on cases to fix. Assumed the values that had negative values were mistaken and switched to positive. The countries were Benin once, Ecuador 3 times, Lithuania once, Portugal once, San Marino once, Spain twice, Uganda once and UK once. Spain had a negative instance for death cases. After an internet search, no correct, reliable information was found on that instance, so it was decided to use the same number as the day before. It was replaced from -1918 to 74;

8 – By using a Pivot Table, it was configured from Country as rows and Cases and Deaths as values in order to get total sum per country to be merged into the second Indexes data-set;

9 – Creation of death and infection ratio from population and total cases to be merged also into the second Indexes data-set.

Continent	Country	Population	Date	Cases	Deaths
Asia	Afghanistan	37172386	31/05/2020	866	3
Asia	Afghanistan	37172386	30/05/2020	623	11
Asia	Afghanistan	37172386	29/05/2020	580	8
Asia	Afghanistan	37172386	28/05/2020	625	7
Asia	Afghanistan	37172386	27/05/2020	658	1
Asia	Afghanistan	37172386	26/05/2020	591	1
Asia	Afghanistan	37172386	25/05/2020	584	2
Asia	Afghanistan	37172386	24/05/2020	782	11
Asia	Afghanistan	37172386	23/05/2020	540	12
Asia	Afghanistan	37172386	22/05/2020	531	6
Asia	Afghanistan	37172386	21/05/2020	492	9
Asia	Afghanistan	37172386	20/05/2020	581	5

Figure A – Corona Data-set Structure

Higher resolution graphic can be found on Appendix A – Corona Data-set Structure.

The image above represents the final structure of the first data-set. It contained 6 variables and 20082 instances.

For the second data-set (Indexes), the process of tidying happened as follows:

1 – From the education index original data-set, it was taken only the 'Country' and '2018' variables. The variable 2018 was renamed as 'Education';

2 – From the inequality index original data-set, it was taken only the '2018' variable and renamed as 'Inequality';

3 – From the health expend gdp index, it was taken only the ‘2017’ value (2018 is still not available to this date) and renamed as ‘GDP-Health’;

4 – From the hdi index, it was taken the ‘2018’ and ‘rank 2018’ variables and renamed as ‘HDI-Index’ and ‘HDI-Rank’;

5 – From the first data-set, it was taken the population data, total cases, total deaths, death-ratio and infection ratio. Death ratio is number of deaths divided by number of infected and Infection Ratio is number of infected divided by total population;

6 – Some countries instances were deleted as they didn’t have some of the indexes available. The deleted countries were: Nauru, Tuvalu, Cook Islands, Monaco, Niue, San Marino, Hong Kong, Libya, Liechtenstein, Montenegro, Palestine, Syria, Yemen, Brunei, Qatar, Eritrea, Singapore, Kiribati, Marshall Islands, Micronesia, Palau, Samoa, Solomon Islands, Tonga, Turkmenistan, Vanuatu.

Country	Population	Education	Inequality	GDP-Health	HDI-Index	HDI-Rank	Total-Cases	Total-Deaths	Death-Ratio	Infection-Ratio
Afghanistan	37172386	0.413	0.432	11.8	0.496	170	14525	249	0.017	0.0004
Albania	2866376	0.758	0.727	6.2	0.791	69	1122	33	0.029	0.0004
Algeria	42228429	0.675	0.743	6.4	0.759	82	9267	646	0.070	0.0002
Andorra	77006	0.708	0.935	10.3	0.857	36	764	51	0.067	0.0099
Angola	30809762	0.498	0.607	2.8	0.574	149	84	4	0.048	0.0000
Antigua and I	96286	0.655	0.816	4.5	0.776	74	25	3	0.120	0.0003
Argentina	44494502	0.842	0.781	9.1	0.83	48	16201	528	0.033	0.0004
Armenia	2951776	0.759	0.684	10.4	0.76	81	8927	127	0.014	0.0030
Australia	24992369	0.923	0.92	9.2	0.938	6	7185	103	0.014	0.0003
Austria	8847037	0.871	0.927	10.4	0.914	20	16638	668	0.040	0.0019
Azerbaijan	9942334	0.694	0.759	6.7	0.754	87	5246	61	0.012	0.0005
Bahamas	385640	0.741	0.853	5.8	0.805	60	102	11	0.108	0.0003
Bahrain	1569439	0.738	0.907	4.7	0.838	45	10793	17	0.002	0.0069
Bangladesh	161356039	0.513	0.559	2.3	0.614	135	44608	610	0.014	0.0003
Barbados	286641	0.773	0.766	6.8	0.813	56	92	7	0.076	0.0003
Belarus	9485386	0.837	0.776	5.9	0.817	50	41658	229	0.005	0.0044
Belgium	11422068	0.893	0.919	10.3	0.919	17	58186	9453	0.162	0.0051
Belize	383071	0.691	0.645	5.6	0.72	103	18	2	0.111	0.0000
Benin	11485048	0.476	0.462	3.7	0.52	163	650	3	0.005	0.0001
Bhutan	754394	0.441	0.673	3.2	0.617	134	43	0	0.000	0.0001

Figure B – Indexes Data-set Structure

Higher resolution graphic can be found on Appendix B – Indexes Data-set Structure.

Above, the image represents how the structure of the Indexes data-set was configured.

C. Time Series Analysis – Single Exponential Smoothing

After some exploratory analysis of the two data-sets, the first statistical method that will be reported is the Time Series Analysis (Simple Exponential Smoothing) on the first data-set ‘Corona’.

The Simple Exponential Smoothing method is basically a way to forecast univariate data that has no clear trend behavior. It is required to have a single parameter (alpha), also called smoothing coefficient, to control the rate at which the influence of observations at initial steps exponentially decay [19] [20].

The formula is as follows:

$$St + 1 = \alpha yt + (1 - \alpha)\, St,\ 0 < \alpha \leq 1, t > 0.$$

That is the basic equation, where ϵt is the forecast error for period t [22].

D. Correlation – Pearson’s Correlation Coefficient

Following that, a Correlation Coefficient was calculated on the second data-set ‘Indexes’ in order to analyse if there are correlations between case numbers and deaths with various indexes.

The approach that was used on this project was Pearson’s Correlation Coefficient, which measures statistical relationships between two variables. The main result is called ‘correlation coefficient’, which ranges from -1 to +1 and it tells how closely variables are related. The closer to zero, the lower the relationship and the further from 0, the higher the relationship (positive or inverse correlation) [4][22].

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

The equation above represents the correlation coefficient formula chosen for this project (Pearson’s approach) [23].

IV. Results

A. Exploratory Analysis

Firstly, an exploratory analysis was performed to grasp initial ideas about the data-sets. Please keep in mind all the figures can be found in higher resolution on the Appendices section at the end of this report.

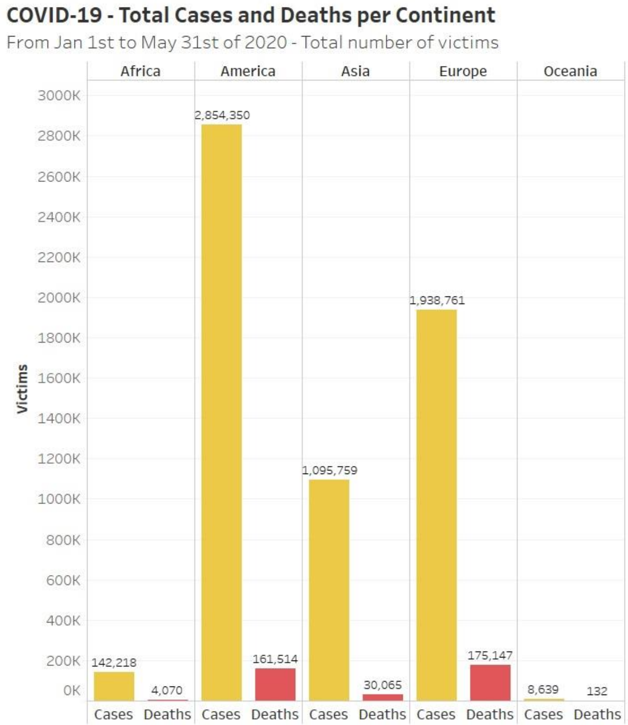


Figure C – Total Cases and Deaths per Continent

This first figure indicates the total number of cases and deaths of COVID-19, per continent, from January 1st to May 31st. As observed, America, Europe and Asia hold the vast majority of cases and this is to be expected as a behavior to follow on the Time Series Analysis.

Figure D – COVID-19 Infection Rate Map

This map displays the infection rate of COVID-19. This was calculated by dividing the total number of cases by the number of each country's population.

COVID-19 Infection Rate - Top 15 Countries

Percentage of the population who became infected with COVID-19

* Data from January 1st to May 31st of 2020

Andorra	0.99%
Bahrain	0.69%
Luxembourg	0.66%
Kuwait	0.63%
United States	0.54%
Spain	0.52%
Ireland	0.51%
Iceland	0.51%
Belgium	0.51%
Chile	0.51%
Peru	0.49%
Belarus	0.44%
United Kingdom	0.41%
Italy	0.39%
Sweden	0.36%

Figure E – COVID-19 Infection Rate – Top 15 Countries

The table above brings into light which are the 15 countries with highest rates of infection.

Figure F – COVID-19 Mortality Rate Map

The map above represents the mortality rate. This was calculated by dividing the number of deaths by the number of cases instead of the population. This was done to have an idea of the actual probability of death once an individual contracts COVID-19.

COVID-19 Mortality Rate - Top 15 Countries

Percentage of people who died after infection from COVID-19

* Data from January 1st to May 31st of 2020

France	18.99%
Belgium	16.25%
Italy	14.33%
United Kingdom	14.01%
Hungary	13.55%
Netherlands	12.87%
Spain	12.06%
Antigua and Barbuda	12.00%
Sweden	11.84%
Mexico	11.17%
Belize	11.11%
Bahamas	10.78%
Liberia	9.64%
Chad	8.56%
Guyana	7.89%

Figure G – COVID-19 Mortality Rate – Top 15 Countries

The table above displays the top 15 countries in regards of mortality rate. As observed, the vast majority of countries are from Europe.

Human Development Index Ranking - Top 20 Countries

Top 20 countries by HDI with anexed Education Index, Inequality Index and Percentage Expenditure of Total GDP on Healthcare

* Data from January 1st to May 31st of 2020

Country	HDI Index	Education Index	Inequality Index	% of GDP on Health
Norway	0.954	0.919	0.985	10.4%
Switzerland	0.946	0.896	0.965	12.3%
Ireland	0.942	0.918	0.955	7.2%
Germany	0.939	0.946	0.929	11.2%
Iceland	0.938	0.918	0.931	8.3%
Australia	0.938	0.923	0.920	9.2%
Sweden	0.937	0.914	0.932	11.0%
Netherlands	0.934	0.906	0.939	10.1%
Denmark	0.930	0.920	0.935	10.1%
Finland	0.925	0.915	0.912	9.2%
Canada	0.922	0.891	0.918	10.6%
New Zealand	0.921	0.923	0.885	9.2%
United States	0.920	0.899	0.956	17.1%
United Kingdom	0.920	0.916	0.903	9.6%
Belgium	0.919	0.893	0.919	10.3%
Japan	0.915	0.850	0.908	10.9%
Austria	0.914	0.871	0.927	10.4%
Luxembourg	0.909	0.802	0.980	5.5%
South Korea	0.906	0.862	0.892	7.6%
Israel	0.906	0.876	0.879	7.4%

Figure H – Indexes Table – Top 20 Countries

The table above displays the top 20 countries with highest HDI Index. It also displays Education Index and Inequality Index as well as the percentage expenditure of total GDP on healthcare per country.

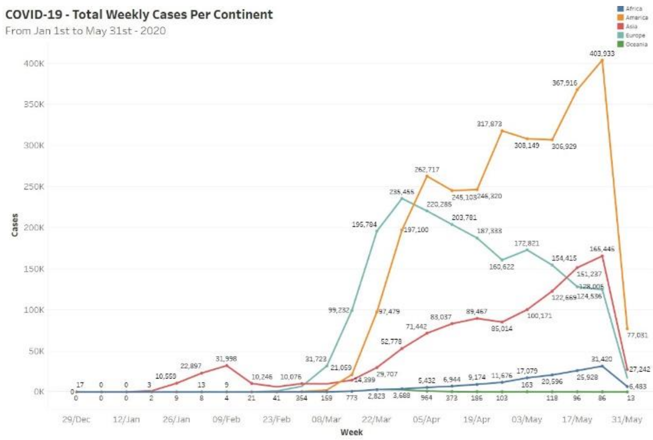


Figure I – Total Weekly Cases per Continent

The line graph above represents the total weekly cases per continent. It is important to be able to see all the values per week and this graph makes it easy to visually interpret the rise and decline of the numbers.

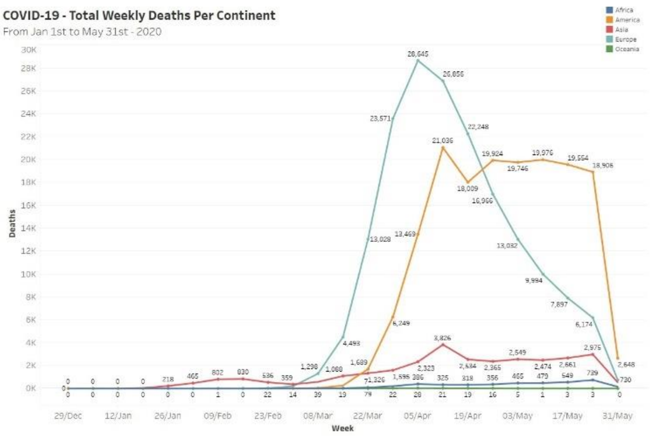


Figure J – Total Weekly Deaths per Continent

This time, this line graph represents the total weekly deaths per continent, following the same standard of representation of the previous one.

With all these exploratory images generated, it was possible to have an initial idea of what generally to expect on the Correlation and Time Series Analyses. It also brought to light how to actual situation of COVID-19 is on the planet, understanding the rates of infection and mortality as well as the sum of all cases, divided by continent.

B. Definition of Hypotheses

Following the first analyses on the data, it was possible to define the hypotheses for the Time Series Analysis and Correlation Analysis. The hypotheses are as follow:

- Time Series Analysis

$$H_0: \mu_{\text{Cases-Before}} = \mu_{\text{Cases-After}}$$
$$H_1: \mu_{\text{Cases-Before}} \neq \mu_{\text{Cases-After}}$$

The null hypothesis states that the number of cases will not change in the next 8 weeks. The alternative hypothesis states that the number of cases will change in the next 8 weeks.

$$H_0: \mu_{\text{Deaths-Before}} = \mu_{\text{Deaths-After}}$$
$$H_1: \mu_{\text{Deaths-Before}} \neq \mu_{\text{Deaths-After}}$$

The null hypothesis states that the number of deaths will not change in the next 8 weeks. The alternative hypothesis states that the number of deaths will change in the next 8 weeks.

Remembering that for both cases, the dependent variables will be ‘cases’/’deaths’ and the independent variables will be ‘weeks’ (time).

- Correlation Analysis

$$H_0: \rho_{\text{Indexes-Ratios}} = 0$$
$$H_1: \rho_{\text{Indexes-Ratios}} \neq 0$$

The null hypotheses state that the correlation coefficients between the indexes and infection/death ratios are 0 (there are no correlations). The alternative hypotheses state that the correlation coefficients between the indexes and infection/death ratios are higher or lower than 0 (there are positive or negative correlations). The level of significance for all tests is set as $\alpha = 0.05$ (95%). It is important to highlight that the importance of this test is actually interpreting how different the correlations are instead of just testing if they are statistically different than zero.

C. Time Series Analysis

The Time Series Analysis method was used twice on this project. The first time was to forecast the number of cases for the next 8 weeks and the second time was to forecast the number of deaths, again for the next 8 weeks. By the time of the writing of this report (July 19th of 2020), we already have actual numbers of the dates the forecasting was performed, so a comparison of the values of real vs forecasted was also done.

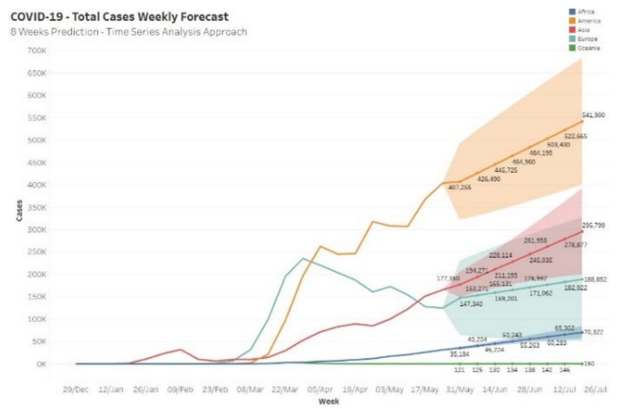


Figure K – Total Cases Weekly Forecast

The image above contains the numbers of prediction of the total cases for the next 8 weeks, starting 31st of May until 26th of July. As we can interpret, all continents are on the rise, however Oceania has very low values compared to the other continents.

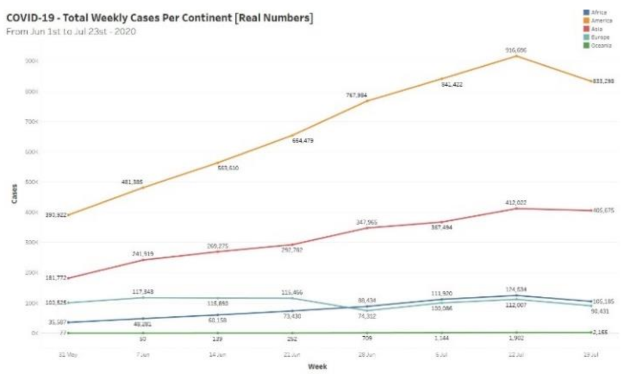


Figure L – Total Weekly Cases after May – Real Numbers

The figure above contains the actual numbers for almost all the forecasted weeks, ending on 19th of July. As we can observe, the values of Africa, America and Asia started very accurately but they ended up higher than the forecasted. The case of Europe was actually the opposite, where the prediction was actually higher than the real numbers. The forecasting of Oceania had lower values but the continent still has very low cases.

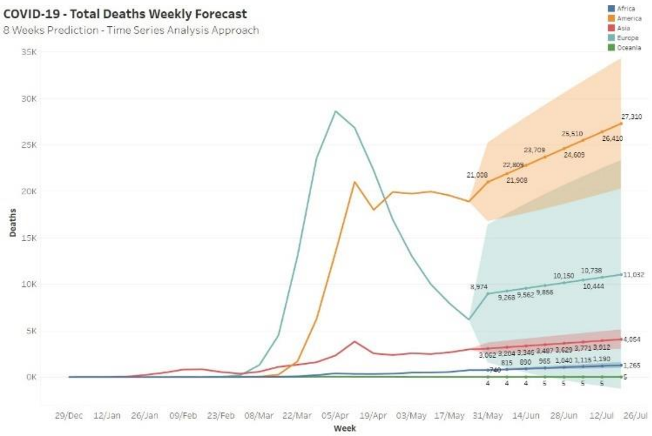


Figure M – Total Deaths Weekly Forecast

The figure above contains the prediction of total deaths for the next 8 weeks, also starting on 31st of May until 26th of July. All the predictions had rising values, although not accentuated.

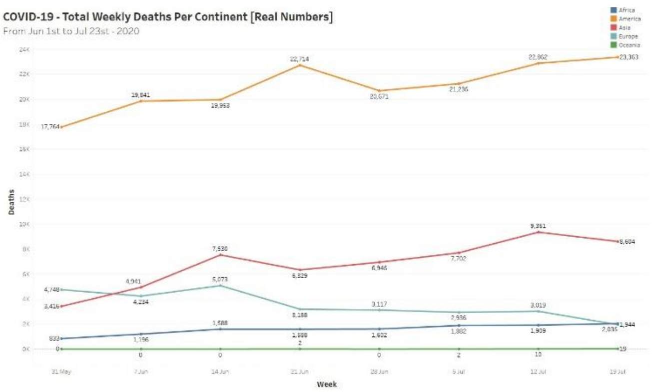


Figure N – Total Weekly Deaths after May – Real Numbers

The figure N contains the actual numbers of deaths, ending on July 19th. As observed, the values for America were considerably accurate, given the conditions of the virus. The predictions for Africa and Oceania started accurately but then the real numbers were slightly higher than predicted. Forecasting for Asia were lower than real numbers and Europe's forecast were significantly higher than real numbers. Concluding, we can reject the null hypothesis in favour of the alternative one. It appears that the number of cases and deaths change significantly after 8 weeks.

REGION	WEEK ENDING							
AFRICA	31-May	07-Jun	14-Jun	21-Jun	28-Jun	05-Jul	12-Jul	26-Jul*
Predicted Cases	35,184	40,204	45,224	50,243	55,263	60,283	65,302	70,322
Real Cases	35,587	48,281	60,158	73,430	88,434	111,920	134,534	105,185
Predicted Deaths	740	815	890	965	1,040	1,115	1,190	1,265
Real Deaths	833	1,196	1,588	1,588	1,602	1,882	1,909	2,036
AMERICA	31-May	07-Jun	14-Jun	21-Jun	28-Jun	05-Jul	12-Jul	26-Jul
Predicted Cases	407,255	426,490	445,725	464,960	484,195	503,430	522,655	541,900
Real Cases	390,922	481,385	563,610	654,479	767,984	841,422	916,696	833,298
Predicted Deaths	21,008	21,908	22,809	23,709	24,609	25,510	26,410	27,310
Real Deaths	17,764	19,841	19,953	22,714	20,671	21,236	22,862	23,363
ASIA	31-May	07-Jun	14-Jun	21-Jun	28-Jun	05-Jul	12-Jul	26-Jul
Predicted Cases	177,350	194,271	211,193	228,114	245,035	261,956	278,877	295,799
Real Cases	181,772	241,919	269,275	292,782	347,965	367,494	412,022	405,675
Predicted Deaths	3,062	3,204	3,346	3,487	3,629	3,771	3,912	4,054
Real Deaths	3,416	4,941	7,530	6,329	6,946	7,702	9,351	8,604
EUROPE	31-May	07-Jun	14-Jun	21-Jun	28-Jun	05-Jul	12-Jul	26-Jul
Predicted Cases	147,340	153,271	159,201	165,201	171,062	176,992	182,922	188,852
Real Cases	100,525	117,343	115,890	115,466	74,312	100,086	112,007	90,431
Predicted Deaths	8,974	9,268	9,562	9,856	10,150	10,444	10,738	11,032
Real Deaths	4,748	4,234	5,073	3,188	3,117	2,936	3,019	1,944
OCEANIA	31-May	07-Jun	14-Jun	21-Jun	28-Jun	05-Jul	12-Jul	26-Jul
Predicted Cases	121	125	130	134	138	142	146	150
Real Cases	77	50	139	252	709	1,444	1,902	2,155
Predicted Deaths	4	4	4	5	5	5	5	5
Real Deaths	0	0	0	2	0	2	10	19

* Forecasting until July 26th. Real numbers until July 19th.

Figure O – Time Series Analysis Chart – Predicted VS Real

The chart above brings all the information together, allowing an easier comparison between the predicted data VS the real collected data.

D. Correlation Analysis

Lastly, the correlation test (Pearson's) was performed to analyse if there are significant correlations between cases/deaths and human development, education, inequality and percentage of expenditure of GDP on healthcare.

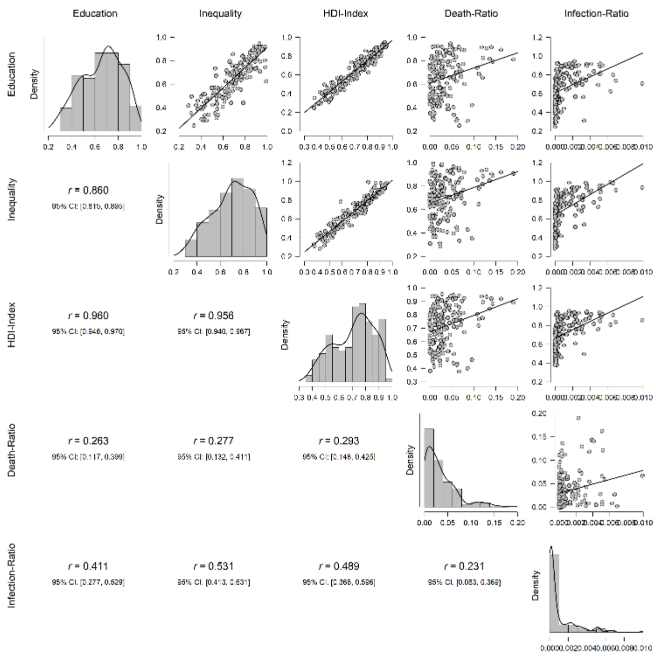


Figure P – Correlation Coefficients

The figure above shows us the correlation coefficients between the dependent and the independent variables that we are analysing, together with scatter and bar plots. As we can observe on the scatter plots and also the coefficients, all values are positively correlated.

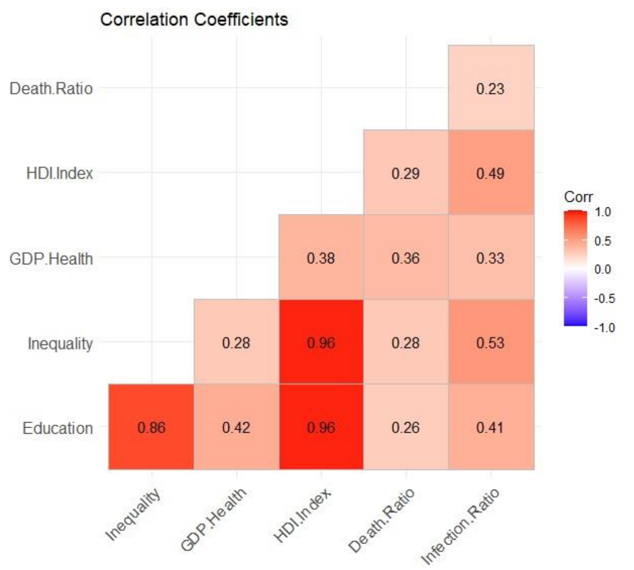


Figure Q – Correlation Matrix

The correlation matrix makes the interpretation easier as it colors the correlations. The darker the red, the stronger the correlation.

Pearson Correlations

			Pearson's r	p
Education	-	Inequality	0.860	< .001
Education	-	GDP-Health	0.420	< .001
Education	-	HDI-Index	0.960	< .001
Education	-	Death-Ratio	0.263	< .001
Education	-	Infection-Ratio	0.411	< .001
Inequality	-	GDP-Health	0.283	< .001
Inequality	-	HDI-Index	0.956	< .001
Inequality	-	Death-Ratio	0.277	< .001
Inequality	-	Infection-Ratio	0.531	< .001
GDP-Health	-	HDI-Index	0.380	< .001
GDP-Health	-	Death-Ratio	0.361	< .001
GDP-Health	-	Infection-Ratio	0.334	< .001
HDI-Index	-	Death-Ratio	0.293	< .001
HDI-Index	-	Infection-Ratio	0.489	< .001
Death-Ratio	-	Infection-Ratio	0.231	0.003

Figure R – Pairwise Correlation Matrix

The pairwise table also makes the correlation indexes easier to read and interpret. It also shows us the level of significance (P value). We can observe the p values are 0.001 for all the relevant correlations we are looking for.

The relevant correlation coefficients are:

Index	Ratio	Correlation Coefficient
Education	Death	0.263
Education	Infection	0.411
Inequality	Death	0.277
Inequality	Infection	0.531
% of GDP	Death	0.361
% of GDP	Infection	0.334
HDI	Death	0.293
HDI	Infection	0.489

As we can observe on the table above, Education Index, Inequality Index and HDI with Death Ratio have little to none correlation. The % of GDP with Death Ratio have low correlation. Education Index, HDI and % of GDP have also low correlation. Lastly, Inequality Index and Infection Rate are moderately correlated. Concluding, all the relevant obtained significance levels (p values) were < 0.001, therefore we can reject the null hypothesis in favour of the alternative one. It appears that there are, even mildly, correlations within the variables. Of course, what we were really interested in this test is not weather there are nor not correlations at all but how significant are they.

V. Conclusions

A. Conclusions

The results of the Time Series Analysis were satisfactory given the changing conditions surrounding the virus. It is important to highlight that forecasting the behavior of this virus (for cases and deaths) is relatively difficult as there are many changing variables that directly impact on these numbers, like how strict is the quarantine at the time of the prediction, number of the population being tested, weather conditions, and much more. It was possible to forecast fairly accurately in most of the regions. Furthermore, we are considering the numbers used on this analysis were accurate and the method of collection was always the same (pre and post analysis) as they came from the same source.

Regarding the correlations, most of them were either little to none of low correlations with exception of the correlation of Inequality Index with Infection Rates. Could this be happening due to the approach of quarantine of countries that have high rates of inequality? Could this be connected to the fact that poor people have less access to basic sanitation or even masks? This could be a job for a future analysis.

B. Future Work

With the obtained results of this project, it is hoped that a future work on the correlation between Inequality Index and Infection Ratio will bring a more detailed approach on factors and characteristics of this connection.

References

[1] "Coronavirus disease 2019 (COVID-19)", Who.int, 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf?sfvrsn=b8304bf0_2#:~:text=Retrospective%20investigations%20by%20Chinese%20authorities,%2C%20some%20did%20not. [Accessed: 19- Jul- 2020].

[2] "Google News", News.google.com, 2020. [Online]. Available: https://news.google.com/covid19/map?hl=en-IE&mid=/m/02j71&gl=IE&ceid=IE:en. [Accessed: 19- Jul- 2020].

[3] "Time Series Analysis for Better Decision Making in Business", Researchoptimus.com, 2020. [Online]. Available: https://www.researchoptimus.com/article/what-is-time-series-analysis.php. [Accessed: 19- Jul- 2020].

[4] "Correlation - Statistical Techniques, Rating Scales, Correlation Coefficients, and More - Creative Research Systems", Surveysystem.com, 2020. [Online]. Available: https://www.surveysystem.com/correlation.htm. [Accessed: 19- Jul- 2020].

[5] J. Ma, "China's first confirmed Covid-19 case traced back to November 17", South China Morning Post, 2020. [Online]. Available: https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back. [Accessed: 19- Jul- 2020].

[6] H. Bhargava, "Coronavirus History", WebMD, 2020. [Online]. Available: https://www.webmd.com/lung/coronavirus-history. [Accessed: 19- Jul- 2020].

[7] "Q&A on coronaviruses (COVID-19)", Who.int, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses#:~:text=symptoms. [Accessed: 19- Jul- 2020].

[8] M. Wadman, J. Couzin-Frankel, J. Kaiser and C. Matacic, "How does coronavirus kill? Clinicians trace a ferocious rampage through the body, from brain to toes", Science Mag, 2020. Available: 10.1126/science.abc3208 [Accessed 19 July 2020].

[9] "Coronavirus (COVID-19) - Google News", News.google.com, 2020. [Online]. Available: https://news.google.com/covid19/map?hl=en-IE&gl=IE&ceid=IE:en. [Accessed: 19- Jul- 2020].

[10] "WHO Coronavirus Disease (COVID-19) Dashboard", Covid19.who.int, 2020. [Online]. Available: https://covid19.who.int/. [Accessed: 19- Jul- 2020].

[11] T. Chakraborty and I. Ghoshb, "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis", Indian Statistical Institute, 2020.

[12] F. Petropoulos, S. Makridakis, "Forecasting the novel coronavirus COVID-19", Universidad Nacional de Mar del Plata, 2020.

[13] P. Conceição et al., COVID-19 and Human Development, 1st ed. New York: United Nations Development Programme, 2020.

[14] S. Barua, "Understanding Corononomics: The Economic Implications of the Coronavirus (COVID-19) Pandemic", University of Dhaka - Department of International Business, 2020.

[15] A. Chowdhury, K. Kabir, J. Tanimoto, "How quarantine and social distancing policy cansuppress the outbreak of novel coronavirus indeveloping or under poverty level countries : amathematical and statistical analysis", New MexicoState University, 2020.

[16] "Today's data on the geographic distribution of COVID-19 cases worldwide", European Centre for Disease Prevention and Control, 2020. [Online]. Available: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. [Accessed: 19- Jul- 2020].

[17] "World Health Organization - GHO - Global Health Observatory Data Repository", Apps.who.int, 2020. [Online]. Available: https://apps.who.int/gho/data/view.main.GHEDCHEGDPSHA2011v?lang=en. [Accessed: 19- Jul- 2020].

[18] "United Nations Development Programee - Human Development Reports", Hdr.undp.org, 2020. [Online]. Available: http://hdr.undp.org/en/data#. [Accessed: 19- Jul- 2020].

[19] R. Hyndman and G. Athanasopoulos, "Simple exponential smoothing", Forecasting: Principles and Practice, 2020. [Online]. Available: https://otexts.com/fpp2/ses.html. [Accessed: 19- Jul- 2020].

[20] J. Brownlee, "A Gentle Introduction to Exponential Smoothing for Time Series Forecasting in Python", Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/. [Accessed: 19- Jul- 2020].

[21] "Forecasting with Single Exponential Smoothing", NIST/SEMATECH e-Handbook of Statistical Methods, 2020. [Online]. Available: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc432.htm. [Accessed: 19- Jul- 2020].

[22] "Pearson's Correlation Coefficient", Statistics Solutions, 2020. [Online]. Available: https://www.statisticssolutions.com/pearsons-correlation-coefficient/. [Accessed: 19- Jul- 2020].

[23] "Pearson Correlation Coefficient Calculator", Socscistatistics.com, 2020. [Online]. Available: https://www.socscistatistics.com/tests/pearson/. [Accessed: 19- Jul- 2020]

APPENDICES

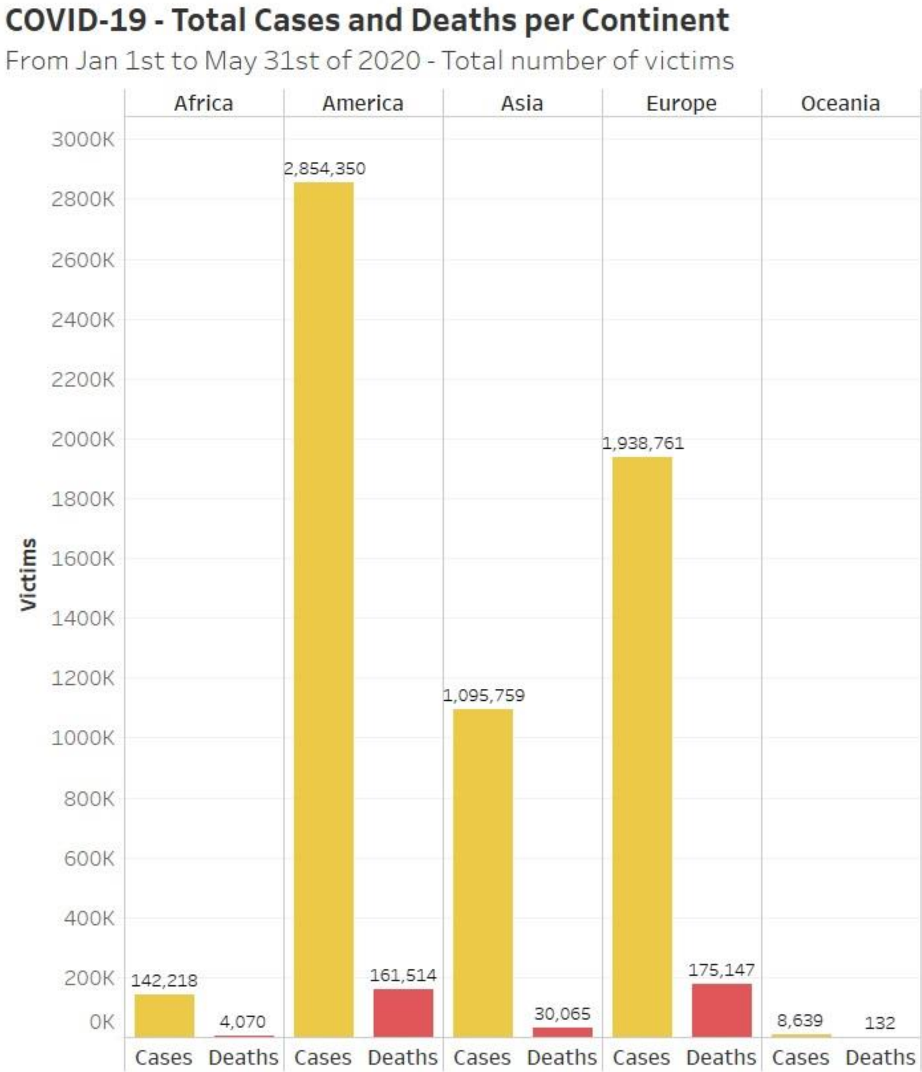
Appendix A – Corona Data-set Structure

Continent	Country	Population	Date	Cases	Deaths
Asia	Afghanistan	37172386	31/05/2020	866	3
Asia	Afghanistan	37172386	30/05/2020	623	11
Asia	Afghanistan	37172386	29/05/2020	580	8
Asia	Afghanistan	37172386	28/05/2020	625	7
Asia	Afghanistan	37172386	27/05/2020	658	1
Asia	Afghanistan	37172386	26/05/2020	591	1
Asia	Afghanistan	37172386	25/05/2020	584	2
Asia	Afghanistan	37172386	24/05/2020	782	11
Asia	Afghanistan	37172386	23/05/2020	540	12
Asia	Afghanistan	37172386	22/05/2020	531	6
Asia	Afghanistan	37172386	21/05/2020	492	9
Asia	Afghanistan	37172386	20/05/2020	581	5

Appendix B – Indexes Data-set Structure

Country	Population	Education	Inequality	GDP-Health	HDI-Index	HDI-Rank	Total-Cases	Total-Deaths	Death-Ratio	Infection-Ratio
Afghanistan	37172386	0.413	0.432	11.8	0.496	170	14525	249	0.017	0.0004
Albania	2866376	0.758	0.727	6.2	0.791	69	1122	33	0.029	0.0004
Algeria	42228429	0.675	0.743	6.4	0.759	82	9267	646	0.070	0.0002
Andorra	77006	0.708	0.935	10.3	0.857	36	764	51	0.067	0.0099
Angola	30809762	0.498	0.607	2.8	0.574	149	84	4	0.048	0.0000
Antigua and	96286	0.655	0.816	4.5	0.776	74	25	3	0.120	0.0003
Argentina	44494502	0.842	0.781	9.1	0.83	48	16201	528	0.033	0.0004
Armenia	2951776	0.759	0.684	10.4	0.76	81	8927	127	0.014	0.0030
Australia	24992369	0.923	0.92	9.2	0.938	6	7185	103	0.014	0.0003
Austria	8847037	0.871	0.927	10.4	0.914	20	16638	668	0.040	0.0019
Azerbaijan	9942334	0.694	0.759	6.7	0.754	87	5246	61	0.012	0.0005
Bahamas	385640	0.741	0.853	5.8	0.805	60	102	11	0.108	0.0003
Bahrain	1569439	0.738	0.907	4.7	0.838	45	10793	17	0.002	0.0069
Bangladesh	161356039	0.513	0.559	2.3	0.614	135	44608	610	0.014	0.0003
Barbados	286641	0.773	0.766	6.8	0.813	56	92	7	0.076	0.0003
Belarus	9485386	0.837	0.776	5.9	0.817	50	41658	229	0.005	0.0044
Belgium	11422068	0.893	0.919	10.3	0.919	17	58186	9453	0.162	0.0051
Belize	383071	0.691	0.645	5.6	0.72	103	18	2	0.111	0.0000
Benin	11485048	0.476	0.462	3.7	0.52	163	650	3	0.005	0.0001
Bhutan	754394	0.441	0.673	3.2	0.617	134	43	0	0.000	0.0001

Appendix C – Total Cases and Deaths per Continent



Appendix D – COVID-19 Infection Rate Map

COVID-19 Infection Rate

Percentage of the population who became infected with COVID-19

* Data from January 1st to May 31st of 2020

Appendix E – COVID-19 Infection Rate – Top 15 Countries

COVID-19 Infection Rate - Top 15 Countries

Percentage of the population who became infected with COVID-19

* Data from January 1st to May 31st of 2020

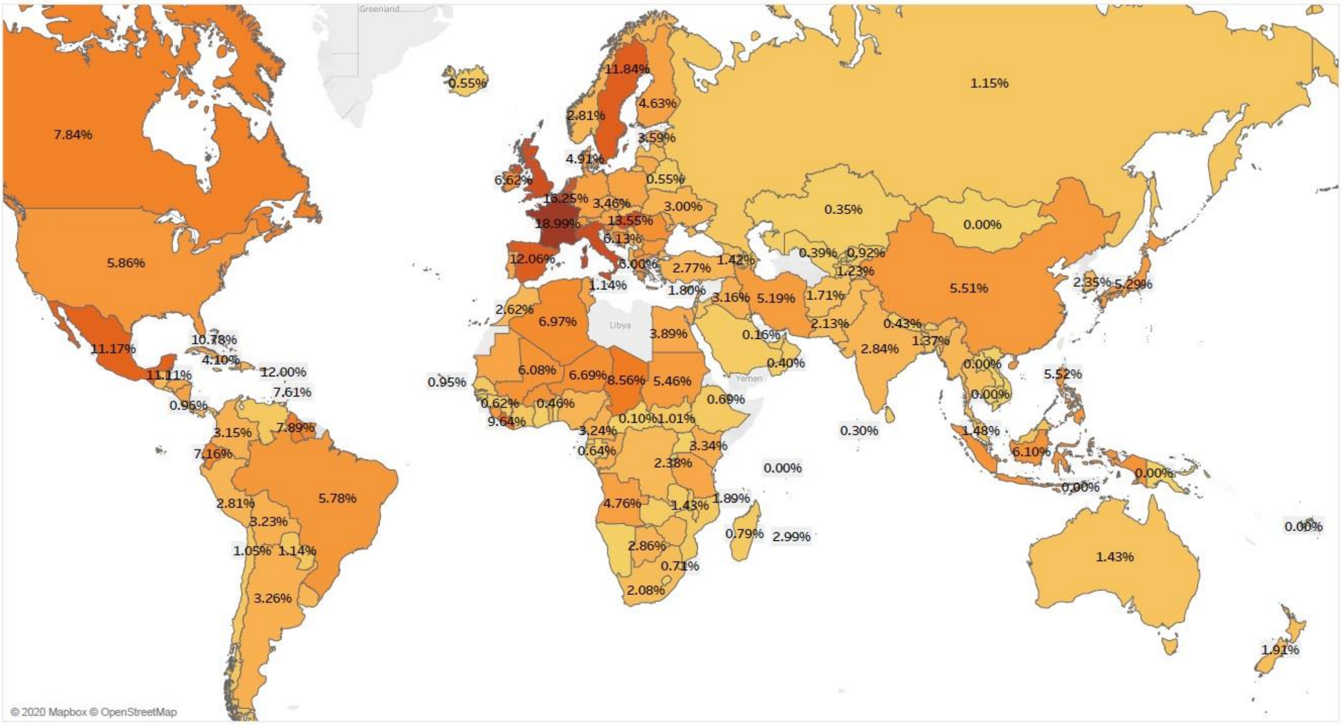
Andorra	0.99%
Bahrain	0.69%
Luxembourg	0.66%
Kuwait	0.63%
United States	0.54%
Spain	0.52%
Ireland	0.51%
Iceland	0.51%
Belgium	0.51%
Chile	0.51%
Peru	0.49%
Belarus	0.44%
United Kingdom	0.41%
Italy	0.39%
Sweden	0.36%

Appendix F – COVID-19 Mortality Rate Map

COVID-19 Mortality Rate

Percentage of people who died after infection from COVID-19

* Data from January 1st to May 31st of 2020



Appendix G – COVID-19 Mortality Rate – Top 15 Countries

COVID-19 Mortality Rate - Top 15 Countries

Percentage of people who died after infection from COVID-19

* Data from January 1st to May 31st of 2020

France	18.99%
Belgium	16.25%
Italy	14.33%
United Kingdom	14.01%
Hungary	13.55%
Netherlands	12.87%
Spain	12.06%
Antigua and Barbuda	12.00%
Sweden	11.84%
Mexico	11.17%
Belize	11.11%
Bahamas	10.78%
Liberia	9.64%
Chad	8.56%
Guyana	7.89%

Appendix H – Indexes Table – Top 20 Countries

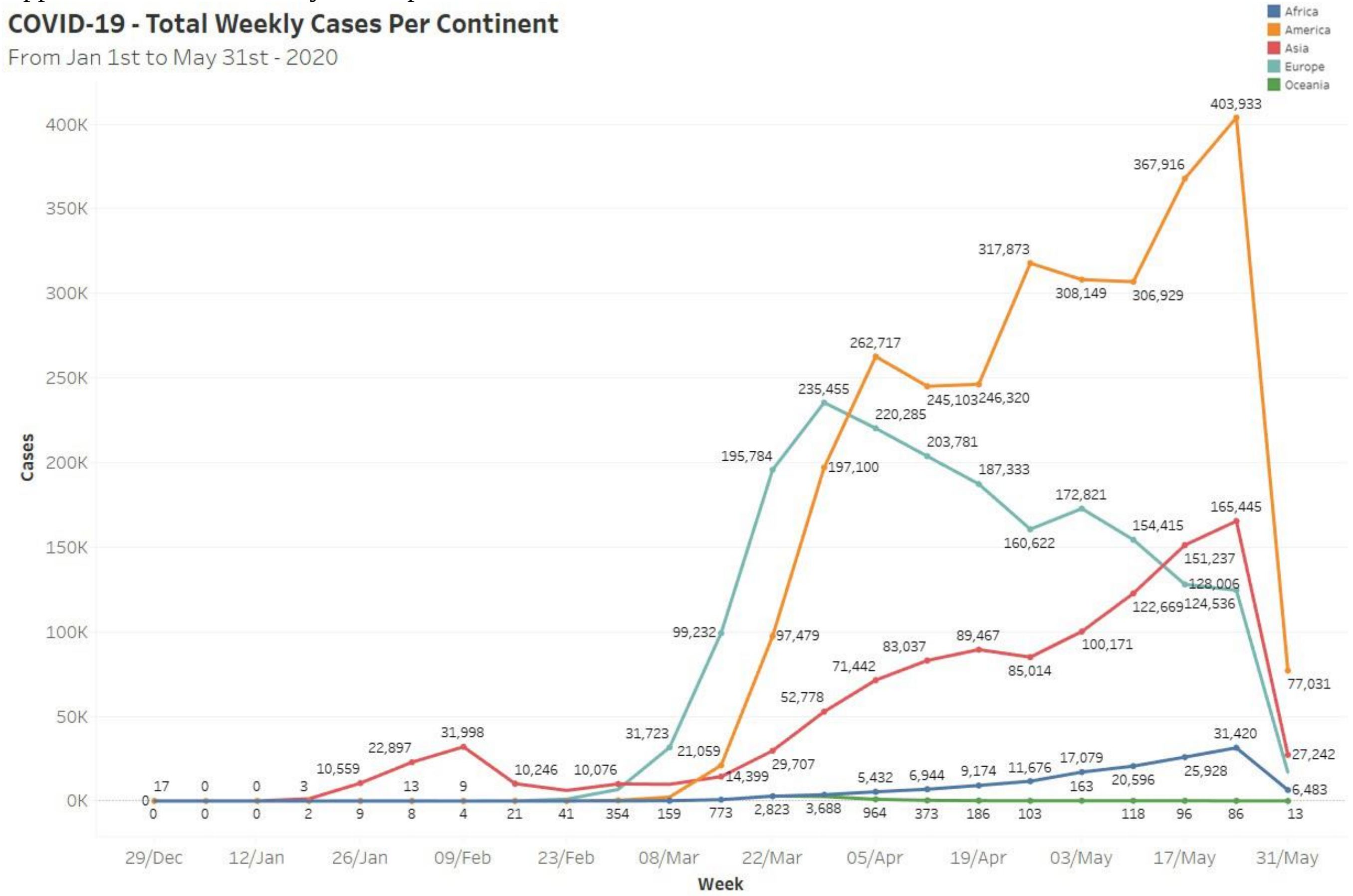
Human Development Index Ranking - Top 20 Countries

Top 20 countries by HDI with anexed Education Index, Inequality Index and Percentage Expenditure of Total GDP on Healthcare

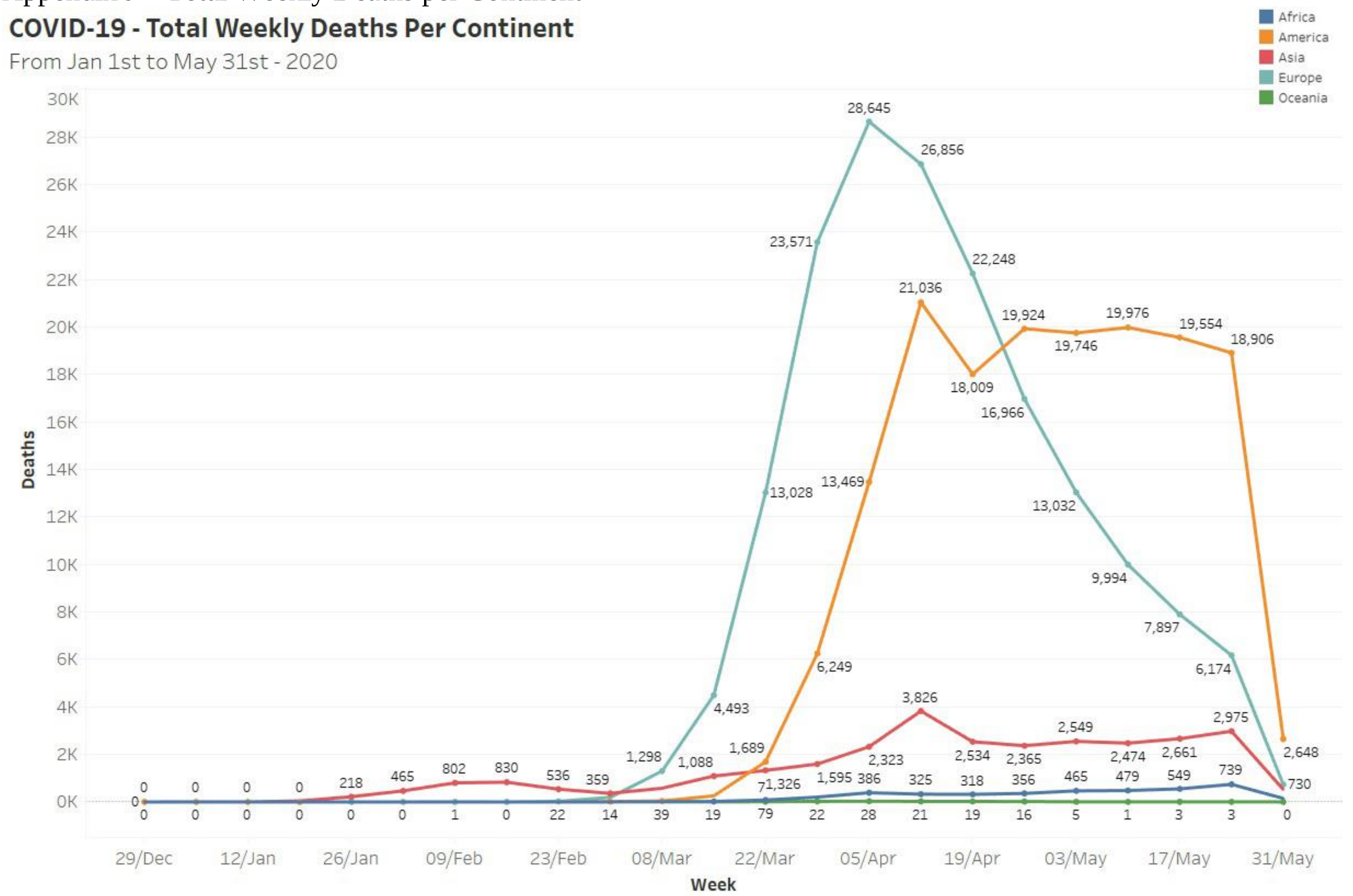
* Data from January 1st to May 31st of 2020

Country	HDI Index	Education Index	Inequality Index	% of GDP on Health
Norway	0.954	0.919	0.985	10.4%
Switzerland	0.946	0.896	0.965	12.3%
Ireland	0.942	0.918	0.955	7.2%
Germany	0.939	0.946	0.929	11.2%
Iceland	0.938	0.918	0.931	8.3%
Australia	0.938	0.923	0.920	9.2%
Sweden	0.937	0.914	0.932	11.0%
Netherlands	0.934	0.906	0.939	10.1%
Denmark	0.930	0.920	0.935	10.1%
Finland	0.925	0.915	0.912	9.2%
Canada	0.922	0.891	0.918	10.6%
New Zealand	0.921	0.923	0.885	9.2%
United States	0.920	0.899	0.956	17.1%
United Kingdom	0.920	0.916	0.903	9.6%
Belgium	0.919	0.893	0.919	10.3%
Japan	0.915	0.850	0.908	10.9%
Austria	0.914	0.871	0.927	10.4%
Luxembourg	0.909	0.802	0.980	5.5%
South Korea	0.906	0.862	0.892	7.6%
Israel	0.906	0.876	0.879	7.4%

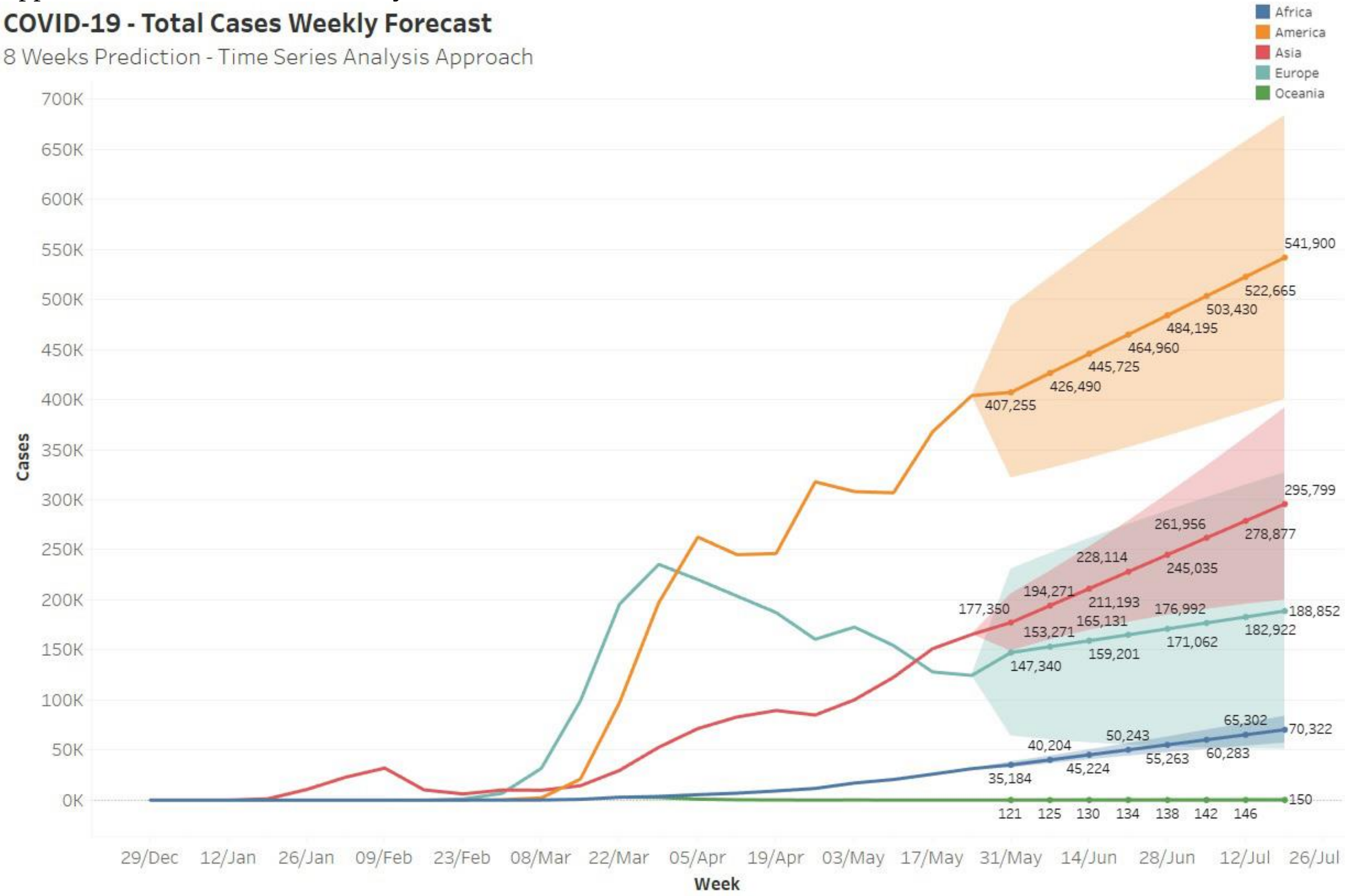
Appendix I – Total Weekly Cases per Continent



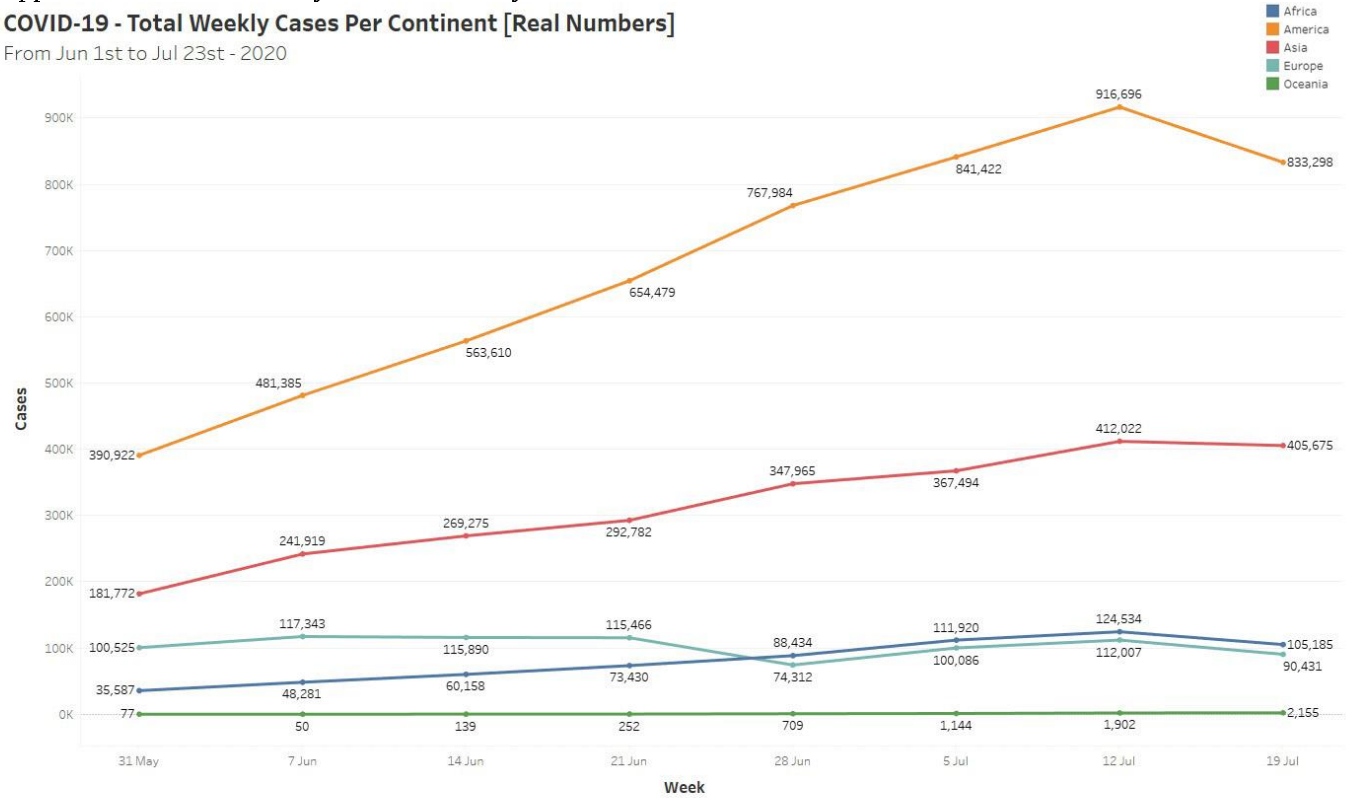
Appendix J – Total Weekly Deaths per Continent



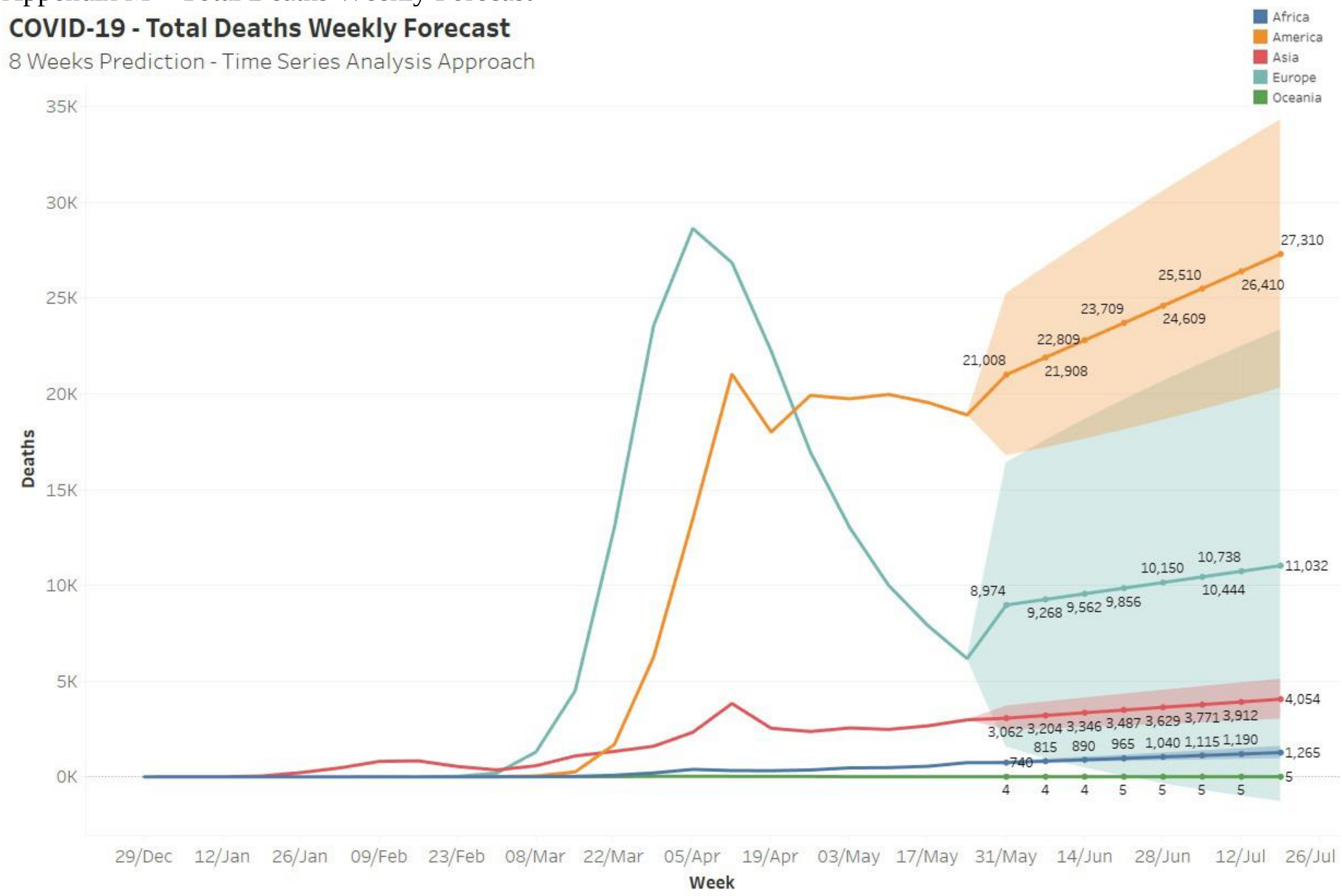
Appendix K – Total Cases Weekly Forecast



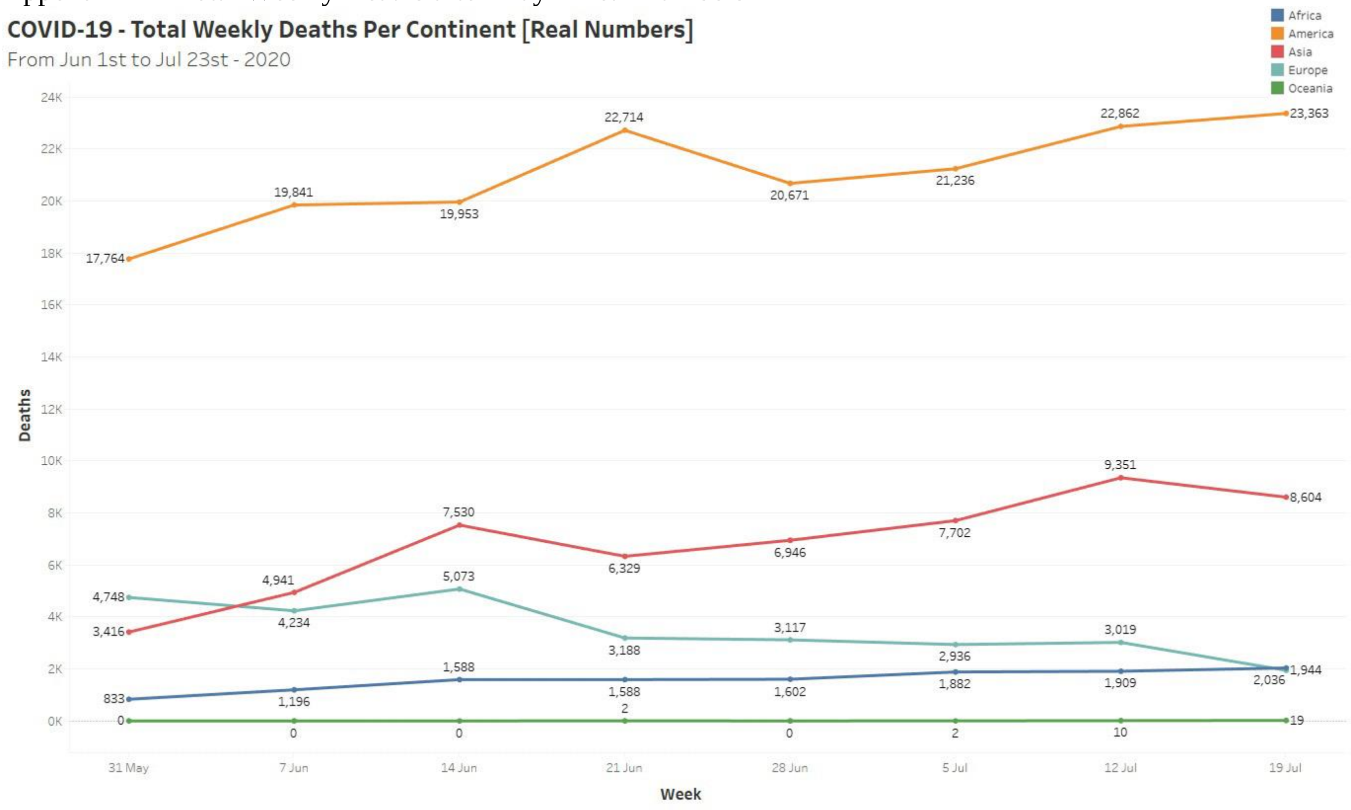
Appendix L – Total Weekly Cases after May – Real Numbers



Appendix M – Total Deaths Weekly Forecast



Appendix N – Total Weekly Deaths after May – Real Numbers

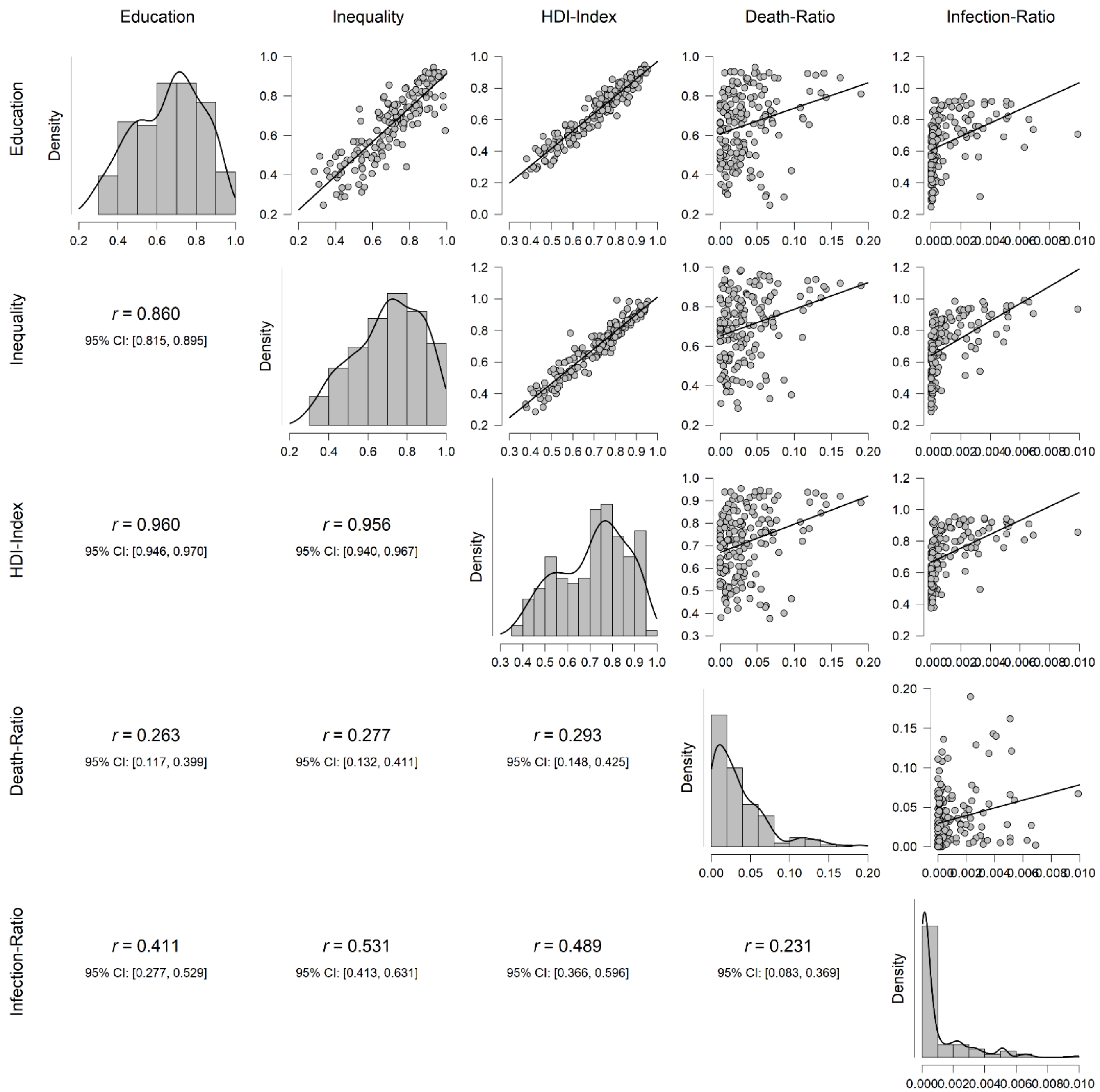


Appendix O – Time Series Analysis Chart – Predicted VS Real

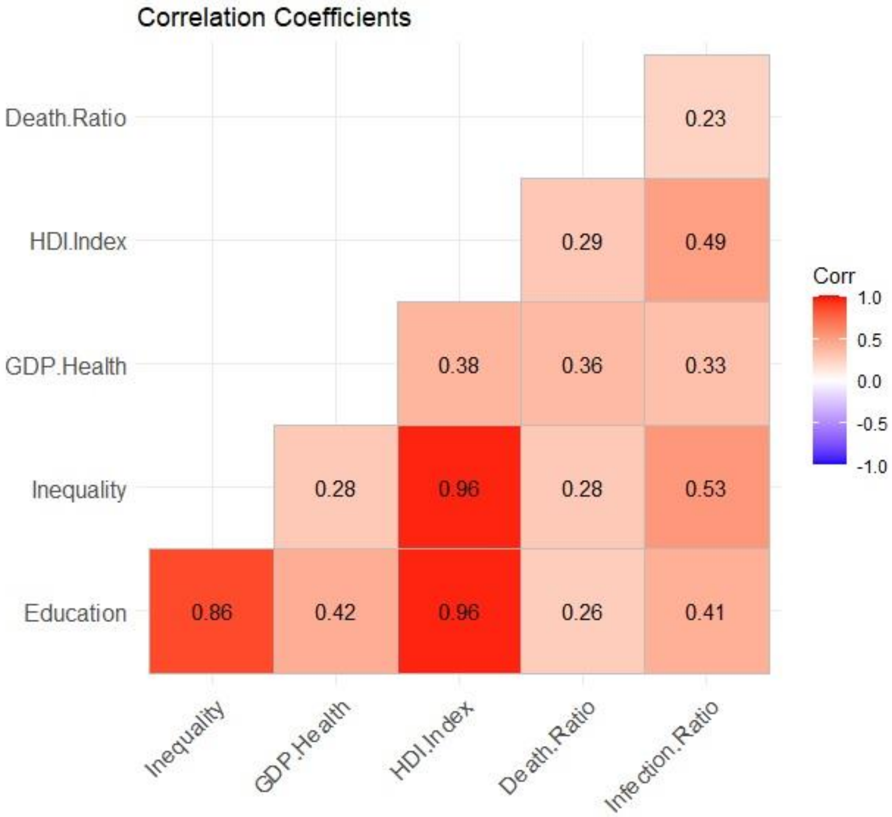
REGION	WEEK ENDING							
AFRICA	**31-May**	**07-Jun**	**14-Jun**	**21-Jun**	**28-Jun**	**05-Jul**	**12-Jul**	**26-Jul***
Predicted Cases	35,184	40,204	45,224	50,243	55,263	60,283	65,302	70,322
Real Cases	35,587	48,281	60,158	73,430	88,434	111,920	134,534	105,185
Predicted Deaths	740	815	890	965	1,040	1,115	1,190	1,265
Real Deaths	833	1,196	1,588	1,588	1,602	1,882	1,909	2,036
AMERICA	**31-May**	**07-Jun**	**14-Jun**	**21-Jun**	**28-Jun**	**05-Jul**	**12-Jul**	**26-Jul**
Predicted Cases	407,255	426,490	445,725	464,960	484,195	503,430	522,655	541,900
Real Cases	390,922	481,385	563,610	654,479	767,984	841,422	916,696	833,298
Predicted Deaths	21,008	21,908	22,809	23,709	24,609	25,510	26,410	27,310
Real Deaths	17,764	19,841	19,953	22,714	20,671	21,236	22,862	23,363
ASIA	**31-May**	**07-Jun**	**14-Jun**	**21-Jun**	**28-Jun**	**05-Jul**	**12-Jul**	**26-Jul**
Predicted Cases	177,350	194,271	211,193	228,114	245,035	261,956	278,877	295,799
Real Cases	181,772	241,919	269,275	292,782	347,965	367,494	412,022	405,675
Predicted Deaths	3,062	3,204	3,346	3,487	3,629	3,771	3,912	4,054
Real Deaths	3,416	4,941	7,530	6,329	6,946	7,702	9,351	8,604
EUROPE	**31-May**	**07-Jun**	**14-Jun**	**21-Jun**	**28-Jun**	**05-Jul**	**12-Jul**	**26-Jul**
Predicted Cases	147,340	153,271	159,201	165,201	171,062	176,992	182,922	188,852
Real Cases	100,525	117,343	115,890	115,466	74,312	100,086	112,007	90,431
Predicted Deaths	8,974	9,268	9,562	9,856	10,150	10,444	10,738	11,032
Real Deaths	4,748	4,234	5,073	3,188	3,117	2,936	3,019	1,944
OCEANIA	**31-May**	**07-Jun**	**14-Jun**	**21-Jun**	**28-Jun**	**05-Jul**	**12-Jul**	**26-Jul**
Predicted Cases	121	125	130	134	138	142	146	150
Real Cases	77	50	139	252	709	1,444	1,902	2,155
Predicted Deaths	4	4	4	5	5	5	5	5
Real Deaths	0	0	0	2	0	2	10	19

* Forecasting until July 26th. Real numbers until July 19th.

Appendix P – Correlation Coefficients



Appendix Q – Correlation Matrix



Appendix R – Pairwise Correlation Matrix

Pearson Correlations

			Pearson's r	p
Education	-	Inequality	0.860	< .001
Education	-	GDP-Health	0.420	< .001
Education	-	HDI-Index	0.960	< .001
Education	-	Death-Ratio	0.263	< .001
Education	-	Infection-Ratio	0.411	< .001
Inequality	-	GDP-Health	0.283	< .001
Inequality	-	HDI-Index	0.956	< .001
Inequality	-	Death-Ratio	0.277	< .001
Inequality	-	Infection-Ratio	0.531	< .001
GDP-Health	-	HDI-Index	0.380	< .001
GDP-Health	-	Death-Ratio	0.361	< .001
GDP-Health	-	Infection-Ratio	0.334	< .001
HDI-Index	-	Death-Ratio	0.293	< .001
HDI-Index	-	Infection-Ratio	0.489	< .001
Death-Ratio	-	Infection-Ratio	0.231	0.003

Appendix S – R Language Code for Correlation Coefficient

```
###### Installing Packages ######
install.packages('tidyverse')
install.packages('GGally')
install.packages('ggplot2')
install.packages("ggcorrplot")

###### Opening the databases ######
corona <- read.csv(file="Indexes.csv", head=TRUE, sep=",")

###### First Look ######
head(corona)
nrow(corona)
summary(corona)

###### TOTAL INFECTION RATIO CORRELATIONS ######

###### Correlation Education X Infection Ratio ######
educ_infec <- cor(corona$Education,corona$Infection.Ratio)
print(educ_infec)
###### Correlation Inequality X Death Ratio ######
inec_infec <- cor(corona$Inequality,corona$Infection.Ratio)
print(inec_infec)
###### Correlation % of GDP Expence on Health X Death Ratio ######
gdp_infec <- cor(corona$GDP.Health,corona$Infection.Ratio)
print(gdp_infec)
###### Correlation HDI X Death Ratio ######
hdi_infec <- cor(corona$HDI.Index,corona$Infection.Ratio)
print(hdi_infec)


###### TOTAL DEATH RATIO CORRELATIONS ######

###### Correlation Education X Death Ratio ######
educ_death <- cor(corona$Education,corona$Death.Ratio)
print(educ_death)
###### Correlation Inequality X Death Ratio ######
inec_death <- cor(corona$Inequality,corona$Death.Ratio)
print(inec_death)
###### Correlation % of GDP Expence on Health X Death Ratio ######
gdp_death <- cor(corona$GDP.Health,corona$Death.Ratio)
print(gdp_death)
###### Correlation HDI X Death Ratio ######
hdi_death <- cor(corona$HDI.Index,corona$Death.Ratio)
print(hdi_death)

###### Combining Results ######
Results <- rbind(educ_death,educ_infec,gdp_death,gdp_infec,hdi_death,hdi_infec,inec_death,inec_infec)
print(Results)

###### Saving Results into a File ######
write.csv(Results, file="Correlations.csv", row.names=TRUE)

###### Plot 1 ######
corona %>% ggpairs(columns = c('Education', 'Inequality', 'GDP.Health', 'HDI.Index','Death.Ratio', 'Infection.Ratio'),
             upper = list(continuous = wrap('cor', size = 6)))

###### Plot 2 ######
correlations_plot <- select (corona,-c(Country,Population,HDI.Rank,Total.Cases,Total.Deaths))
correlations_plot <- cor(correlations_plot)
ggcorrplot(correlations_plot,
         type = 'lower',
         method = 'square',
         lab = TRUE,
         title = 'Correlation Coefficients')
```