Exploring the potential relationship between Covid-19 Cases and other Variables

Ankhee Paul

22/04/2021

Abstract

The advent of Covid-19 which resulted in a global pandemic caused numerous deaths and economic disruptions. This study aims to infer a relationship between the total number of Covid-19 cases in a country and socio-demographic variables of interest. The results of multiple models yielded a positive relationship between population and the number of Covid-19 cases tested every 1 million people with the response variable, the total number Covid-19 cases. The results are significant and can be generalized to the population level.

1. Introduction

December 2019 saw the advent of Covid-19, a disease caused by severe acute respiratory syndrome coronavirus 2(SARS-CoV-2), which led to an ongoing global pandemic, as declared in March 2020. The virus has infected over 100 million people and has caused more than 2 million deaths. The pandemic caused by the virus has led to nationwide lockdowns, travel restrictions, economic disruptions and a rush to find a cure in the form of a vaccine. Even as newly produced vaccines are being tested in several countries as of 2021, and countries are slowly starting to open up, there are several new strains of the virus emerging, especially in the U.K and Africa, that are sending nations back into lockdowns.

The disease, itself, can show no to severe symptoms and can even be fatal. Health experts have encouraged the wearing of face masks, social distancing and maintaining hygiene as some of the preventive measures. The virus is especially fatal for people with underlying diseases as well as for elderly people.

This paper aims to understand the relationship between the total number of covid-19 cases and socio-demographic variables such as population as well as other variables of interest that might have a potential correlation. The independent variables chosen initially are the population of the country and the number of covid-19 tests conducted in each country for every 1 million people of the population. The dependent or response variable in this study is the total number of covid-19 cases in every country. In addition to the previous variables, a third independent variable, the health risk scores of countries, was added. In order to further explore the correlation of the total number of covid-19 cases with other variables, we added the most recent average annual temperatures of countries as well as the percentage of the population using basic sanitation in 2017, as potential independent variables. This is to explore whether climate and hygiene and sanitation levels affect the occurrence of the virus in countries in any way.

Since the aim is to infer a relationship between variables, multiple linear regressions are used. Regressions are a very common tool used in macroeconomic forecasting. A paper[1] by Smeekes and Wijler uses lasso-type penalized regression techniques to macroeconomic forecasting with high-dimensional datasets for forecasting. Similarly, another paper[2] by Pietro reinstates the idea that macroeconomic variables impact health conditions in a country. Many research papers use different forms of regressions to either predict, infer or establish a causal relationship between macroeconomic variables and diseases. For example, a paper[3] explores the effects of socioeconomic and environmental determinants on chronic obstructive pulmonary disease (COPD) mortality using geographically and temporally weighted regression model across Xi'an during 2014–2016. The paper also uses ordinary least square (OLS) and the geographically weighted regression models for cross-comparison.

Despite Covid-19 being a very recent occurrence, individual researches are being conducted all over the world with the goals of determining its causes or exploring the severity of the virus and the pandemic. A recent paper[4] from India uses multiple linear regression to predict the number of new active cases.

This paper simply uses multiple linear regression to determine the relationship between socio-demographic variables, such as population, climate, etc, and the total number of covid-19 cases in a country. The data section narrates each dataset used and its source. The summary statistics section provide key statistical estimates of the variables of interest. The visualization section uses scatterplots to visualize an initial relationship between the variables while maps describe the magnitude of each variable at a global scale. The Regression section of the paper describes the models used and the results derived from them which are discussed in the Discussion section.

2. Data

2.1. John Hopkins Covid-19 Data

The data has been obtained from the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) which is supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). It has been collected from several global entities such as the World Health Organization, Government health authorities in different countries, and data sources that have been tracking the logistics of the covid-19 disease and the global pandemic. The merged version of the datasets that have been used in this particular study, has been obtained from Kaggle.

[1] Link: https://www.sciencedirect.com/science/article/pii/S0169207018300074

[2]Link:https://www.researchgate.net/publication/320925730_Revisiting_the_impact_of_macroeconomic_conditions_on_health_behaviours

[3] Link: https://www.sciencedirect.com/science/article/pii/S0048969720374003

[4] Link: https://www.sciencedirect.com/science/article/pii/S1871402120302939

The John Hopkins data repository contains multiple datasets based on Covid-19 cases in the United States as well as time-series data of the case numbers on a daily basis. The data covers the case numbers at a global level as well as a state-level in the United States. The dataset used in our study is at a global level defining socio-demographic variables such as population as well as statistics pertaining to the pandemic such as the total number of deaths that occurred, etc. It has been obtained from Worldometer.

The data set consists of 16 columns including the index. Most of the columns simply display discrete values enumerating the number of covid-19 cases in each country, country's population, the total number of deaths and recoveries caused by covid-19, new as well as active cases in each country, and the number of covid-19 tests conducted. Other than the columns displaying countries, World Health Organization's regional divisions and continents- which are categorical values. There are a number of missing values. However, we do not remove them as the entire column would be deleted resulting in the loss of a particular country's covid-19 related information.

2.2. World Health Organization Data

Since we aim to include as many relevant variables as to establish inference, a number of datasets have been merged with the original data. Two such variables are the global health risk scores in the year 2019 and the level of basic sanitization usage in a country in 2017.

The data sets from which the health risk scores and the basic sanitation levels have been obtained are from the World Health Organization's "National and Global Health Risks"- assigns risk scores out of 100 to countries based on a comprehensive assessment of leading risks to global health- and " Basic and Safely Managed Sanitation Services" datasets, respectively. In this project, the most recent score in the year 2019 and the most recent year recorded for sanitation, 2017, are used and merged with the covid-19 world dataset in order to make the necessary comparisons. The health risk scores used have been collected using a self-assessment tool for annual reporting called the State Party Self-Assessment Annual Reporting Tool or SPAR. The SPAR (State Party Self-Assessment Annual Reporting) tool consists of 24 indicators for the 13 International Health Regulation capacities needed to detect, assess, notify, report and respond to public health risk and acute events of domestic and international concern.

Hence the data collected is observational data that is updated with time. It is stored in the Global Health Observatory (GHO) which is a data repository of the World Health Organization, containing numerous accessible data sets pertaining to health, medicines and financial protection, all at a global level.

The Global risk scores in 2019 is a comprehensive assessment of leading risks to global health. It provides detailed global and regional estimates of premature mortality, disability and loss of health attributable to 24 global risk factors. The scores have been indexed a given out of a total score of 100. The higher the score, the more susceptible a country is to a potential pandemic or virus/disease.

The basic sanitation level is tracked by the World Health Organisation as the usage of basic sanitation services by a country at an urban and rural level. The year 2017 has been taken as it is the most recent year available. The scores have been indexed out of 100. The higher the level, the more that country practices the use of safe and basic sanitation services.

2.3. Wikipedia Data

Since the aim of the project is to explore potential variables that affect the total number of covid-19 cases in a country, there is a possibility of exploring dynamic data, including the most recent information available. There can be numerous sources of data that can be easily downloaded but it is tedious to constantly go back to check for the most updated version of the data. It is much easier to acquire the data through web-scraping, especially if the data source is huge and complicated to search through.

The most prominent sources of data that provide potential variables related to our topic is the World Health Organization's Global Health Observatory which is constantly acquiring and updating data regarding Covid-19 and other global socio-demographic and health factors. The GHO has two APIs, namely, GHO OData API and Athena API which allows us to scrape their datasets. It would be ideal to scrape the website monthly if one was looking for dynamic data to evaluate time series. However, for our purpose scraping the data annually would be sufficient.

Another source of data is Kaggle.com which contains multiple projects conducted by individuals. Most of these projects use datasets that are readily available and pertinent to the topic in this project. Instead of searching manually through the entire website, trying to find a relevant dataset, it would be much efficient to scrape the data using Kaggle's own API. This would reduce the time and effort spent in searching for a relevant dataset. For this project, looking for variables that potentially affect the total number of Covid-19 cases in a country, we would ideally want the most recent information on political measures, country responses to the pandemic, medical accessibility in countries, etc. With Kaggle being a hub for researchers, there is are projects and datasets containing the relevant information which can be obtained through web scraping.

However, due to limited knowledge on web scraping especially converting API scraped data into text format, we are limiting ourselves to HTML based web scraping of the global temperatures.

In order to add relevant variables to our study, data has been obtained through HTML based web scraping of Wikipedia. The data comprises the average annual as well as the monthly temperature of a country over a year. The temperature is given in Celsius and Fahrenheit. The web scraping has been done to create a data frame that contains the countries and the corresponding temperature in Celsius and Fahrenheit. We will be working with temperature in Celsius only so we remove the Fahrenheit figures in parenthesis and convert our temperature values to a numeric type in order to continue our analysis. Since there are multiple entries for a single country due to different cities having different temperatures, we keep only the average of all temperatures for each country. Hence, our dataset contains the countries and the most recent average annual temperature over their cities in Celsius.

3. Summary Statistics

The summary statistics of each variable of interest has been computed along with boxplots to describe it. After removing a number of significant outliers in each variable using the

interquartile range method, observations between 25% and 75% of the numerical data has been subsetted in order to better visualize the distribution of the data. Each plot has been divided into subplots categorized by the World Health Organisation's regional divisions, namely, Americas, Africa, South- East Asia, Eastern Mediterranean and Western Pacific regions.

Figures 1-6[5]in the Figure appendix section under 'Summary Statistics 'describes the summary statistics observed for each of the chosen variables. The variables are as follows:

Total cases:

The total number of covid-19 cases in each country is a discrete, dependent variable. From the summary statistics of "TotalCases", it is observed that the average total number of cases is highest in America followed by South-East Asia. The Western Pacific region has the smallest average with only 15 countries to account for. Consequently, the standard deviation is highest in America followed by South-East Asia. The highest median, however, lies in the Eastern Mediterranean region which accounts for 22 countries.

From Figure 1[6], a boxplot plotting the total number of cases categorized by WHO region, the medians lie towards the lower ends of the distributions, except for South-East Asia, where the median lies in the upper end of the distribution. There are some significant outliers to be observed, especially in America, Africa and the Western Pacific, at around 80,000 cases in America, and 40,000 cases in Africa. South-East Asia has the smallest interquartile range while the Eastern Mediterranean region has the largest in total cases. The maximum value for the total number of cases is highest in Europe followed by the Eastern Mediterranean region. The smallest maximum value is in South-East Asia which can be considered quite counter-intuitive given that the statistical summary table states the highest average number of cases to be in this region. This can be attested to the fact that this region had very significant outliers.

Population:

The population of each country is a discrete, independent variable that has been divided by the WHO's regional categorization for each country. It is seen in the summary statistics of "Population" that the highest average population of almost 199 million people is in South-east Asia, despite that region containing the least number of countries, that is, 10. This can probably be attested to the fact South-East Asia contains India, which alone carries a population of almost 1.3 billion, as observed in the "max" column. Europe, with the highest number of countries at 55, has the smallest average population of almost 16 million. Consequently, the highest standard deviation is projected in South- East Asia with the huge population difference between India and other countries, whereas Europe shows the smallest standard deviation.

Figure 2[7], displaying boxplots of the population of each country categorized by WHO regions, also shows the medians lying approximately towards the lower end of the distribution for each region. South-East Asia has the highest maximum value followed closely by Africa. The highest upper quartile is present in Eastern Mediterranean while the Inter-quartile range is

[5] See "**3. Summary Statistics**" under the Figures Appendix
[6] See "Figure 1" on Page 19, **3. Summary Statistics**" under the Figures Appendix
[7] See "Figure 2" on Page 19, **3. Summary Statistics**" under the Figures Appendix

highest for Eastern Mediterranean and South East Asia. America and Europe have the largest outlier in terms of magnitude although Europe has the smallest interquartile range.

Tests/1M pop:

This variable denotes the number of covid-19 tests conducted in each country, every one million people of its population. It is an independent discrete variable that essentially signifies the relative magnitude of testing in a country. From the summary statistics "Tests/1M pop", the highest average is from the European region at almost 139287.28 tests conducted every 1 million people. In contrast, Africa has the smallest average of approximately 17039.05 tests conducted every 1 million people with 35 countries to account for. Similarly, the standard deviations and medians, aligning with the mean, is highest in Europe, with 187473.10 and 82799, respectively, and smallest in Africa with 32609.39 and 5600, respectively.

Figure 3[8], displays boxplots of covid-19 tests conducted every 1 million people of the population, categorized by WHO regions. The medians lie towards the lower end of the distribution for each region, except the Eastern Mediterranean region. Europe has the highest median, upper-quartile, maximum number of tests and inter-quartile range with Africa being the least in all these factors. Each region displays several outliers, except Europe.

Global Health Risk Score 2019:

This variable denotes the health risk score each country received out of 100, by the World Health Organization. It is an independent discrete variable displaying the scores in the most recent year of 2019. The summary statistics show the average global health risk score to be highest in Europe, at almost 71, implying that the region is most likely to be at a health risk in case of a major disease outbreak. However, this could be due to the fact that the European region comprises the largest number of countries, 49. The lowest average risk score of around 45 is given to Africa. Both South-East Asia and the Western Pacific comprise only 10 countries yet have quite high average risk scores at 60 and 57, respectively. The median score is the highest in America as the country with the maximum score of 99 lies in that region. All regions except Africa and South-East Asia have countries with a minimum value of 0. The highest standard deviation of 38.24 is in the Western Pacific region implying variability in the health risk scores of the countries in that region.

Figure 4[9] plots the Global health risk scores 2019 of countries, distinguished by World Health Organization's regions in a boxplot. Eastern Mediterranean and South-East Asia have their median towards the middle of its distribution while Europe, America and Africa have their medians lie towards the upper portion of its distribution indicating the fact that most countries in those regions tend to have a higher risk score. The Western Pacific region has its median towards the lower end. Europe has an outlier in the lower end of its distribution indicating a country with an extremely small health risk score, almost 0. Western Pacific has the largest interquartile range and the lowest lower- quartile, while Africa and South-East Asia have the smallest interquartile range. The highest upper - quartile belongs to Europe and the Western Pacific. The maximum value, close to 99, lies in America while every region contains at least one country with the minimum score of 0, except South-East Asia.

[8] See "Figure 3" on Page 20, **3. Summary Statistics**" under the Figures Appendix
[9] See "Figure 4" on Page 20, **3. Summary Statistics**" under the Figures Appendix

Average Annual Temperature:

This variable denotes the most recent average annual temperature of each country in degree Celsius. It is an independent continuous variable displaying the temperatures in the most recent year. The summary statistics show the average annual temperature to be highest in South-East Asia followed by Africa, at almost 26 degrees Celsius and 25 degrees Celsius, implying that the regions are hottest due to them lying in the tropical zone, near the equator. The lowest average annual temperature of around 11 degrees Celsius is in Europe as Europe lies in the Northern Hemisphere. The highest standard deviation of 5 is in the American region implying variability in the temperatures across the region. Africa has the highest average annual temperature of almost 30 degrees. The inter-quartile range is highest in the Eastern Mediterranean region, and lowest in South-East Asia suggesting the fact that in countries South-East have mostly the same high temperature all year round.

Figure 5[10] plots the Average Annual temperature of countries, distinguished by World Health Organization's regions in histograms. Africa and the Eastern Mediterranean region have a more or less even distribution of temperature ranging from 18-30 degrees Celsius, except for an outlier in the Eastern Mediterranean region at 12.5 degrees. The medians lie somewhere in the lower 20 degrees for both these regions. Europe and America have a multimodal distribution with Europe being positively skewed and America being negatively skewed. America shows an outlier at an extremely low temperature of around 6-7 degrees which can be Greenland since it lies in the extreme northern hemisphere. The median is somewhere around 20 degrees while Europe has a much lower median between 10-12 degrees. Both South-East Asia and the Western Pacific are negatively skewed with outliers. The distribution is sparse since both these regions have fewer countries. South-East Asia has a high median temperature in the higher 20 degrees while the Western Pacific has its somewhere between 22-24 degrees. South-East Asia has a unimodal distribution while the Western Pacific has a bimodal one.

Percentage of population using basic sanitation in 2017:

This variable denotes the percentage of the population using basic sanitation, in each country, grouped by the World Health Organization region. It is an independent discrete variable displaying the basic sanitation level in the most recent year of 2017. The summary statistics show that the highest average is in Europe at almost 97% which means that almost all the population in European countries have good access and usage of basic sanitation facilities and maintain proper hygiene- a fact also emphasized by the lowest standard deviation of 3. The lowest average lies in Africa denoting the lack of practice of basic sanitation by a majority of the population. Indicating that this region might have a higher number of covid-19 cases. There are countries in Africa and the American region where 0% of the population use basic sanitation. The standard deviation is highest in the Western Pacific and followed by the American region signifying that the regions have countries with varying level of usage of basic sanitation.

Figure 6[11] plots the percentage of the population practising basic sanitation in each country in 2017, distinguished by World Health Organization's regions, in a boxplot. The European region has the highest median followed by Western Pacific. The lowest median, lower- quartile and upper-quantile belong to Africa. The Eastern Mediterranean has the highest inter-quartile range

[10] See "Figure 5" on Page 21, **3. Summary Statistics**" under the Figures Appendix
[11] See "Figure 6" on Page 21, **3. Summary Statistics**" under the Figures Appendix

while Europe has the smallest. America and Europe have the largest number of outliers towards a lower level of the population using basic sanitation.

4.Visualizations

4.1 Graphs

The visualizations for this study included several graphs that described the data, the chosen variables as well as the relationship between the two variables. After removing a number of significant outliers in each variable using the interquartile range method, observations between 25% and 75% of the numerical data has been subsetted into a new data frame and used to plot histograms boxplots and scatterplots to better visualize the distribution of the data.

Each plot has been divided into subplots categorized by the World Health Organisation's regional divisions, namely, Americas, Africa, South- East Asia, Eastern Mediterranean and Western Pacific regions.

Total Number of Cases and Population:

Figure 1[12] plots the relationship between the total number of covid-19 cases in a country and the population of that country in a scatterplot. Each country is categorized into their respective regions assigned by the World Health Organization. The figure shows each region in a different plot.

It can be observed that South-East Asia has the smallest number of observations while most of the observations are clustered towards the lower end of the graph America and Europe show a weak positive correlation indicating that the higher the population of the country, the higher is the total number of cases. However, the rest of the regions show a neutral relationship between the two variables which means that there is no relationship between the total number of covid-19 cases in a country and its population. America, Europe and the Eastern Mediterranean have the largest outliers in terms of magnitude. However, America shows a large number of covid-19 cases given a comparatively less population (slightly greater than 30 million) which might potentially cause the correlation.

Total Number of Cases and Tests Conducted Every 1 Million People:

Figure 2[13] plots the relationship between the total number of covid-19 cases in a country and the number of covid-19 tests conducted every 1 million people in a scatterplot. Similar to the previous figure each country is categorized into their respective regions assigned by the World Health Organization. The figure shows each region in a different plot.

It is observed that the number of observations is smallest in South-East Asia and largest in Europe and Africa. Most of the observations are clustered towards the bottom left of the graph. Europe has a random scattering of the data points indicating no correlation. Likewise, for other

[12] See "Figure 1" on Page 22, "**4.1 Graphs**" under the Figures Appendix
[13] See "Figure 2" on Page 22, "**4.1 Graphs**" under the Figures Appendix

regions except for Africa which is denoting a weak, positive correlation between the variables. America has some significant outliers where the number of covid-19 cases is quite high (around 80,000) given to a lower number of tests conducted (less than 10,000) compared to the outliers in the Eastern Mediterranean where the outliers show around 80,000 covid-19 cases for around 10,000 tests conducted every 1 million people.

Total Number of Cases and Global health risk scores:

Figure 3[14] plots the relationship between the total number of covid-19 cases in a country and the health risk score assigned to each country out of 100 by the World Health Organization, in a scatterplot. Similar to the previous two figures each country is categorized into their respective regions assigned by the World Health Organization. The figure shows each region in a different plot.

The largest number of observations is in Europe while the smallest numbers are in South East Asia and the Western Pacific. Most of the observations are clustered towards the bottom of the plot. Like the previous two figures, there appears to be no relationship between the two variables in the regions as there is a random scattering of data points. Every region except South-East Asia has some significant outliers. A country in America shows more than 80,000 cases for a smaller health risk score compared to the Western Pacific where the outliers show around 60,000 cases with a health risk score greater than 50%.

Total Number of Cases and Average Annual Temperature of the country:

Figure 4[15] plots the relationship between the total number of covid-19 cases in a country and the average annual temperature of the country, grouped by the World Health Organization regions, in scatterplots.

The largest number of observations is in Europe and Africa while the smallest number is in South East Asia. Most of the observations are clustered towards the bottom of the plot towards a higher temperature. Like the previous three figures, there appears to be no relationship between the two variables in most of the regions as there is a random scattering of data points. This may be due to the lack of observations. Every region except South-East Asia has some significant outliers. A country in America shows more than 80,000 cases for a higher temperature compared to Europe where the outliers show around 60,000 cases at lower temperatures.

Total Number of Cases and Sanitization level:

Figure 5[16] plots the relationship between the total number of covid-19 cases in a country and the percentage of the population using basic sanitation. Similar to the previous figures each country is categorized into their respective regions assigned by the World Health Organization. The figure shows each region in a different plot.

[14] See "Figure 3" on Page 22, "**4.1 Graphs**" under the Figures Appendix
[15] See "Figure 4" on Page 23, "**4.1 Graphs**" under the Figures Appendix
[16] See "Figure 5" on Page 23, "**4.1 Graphs**" under the Figures Appendix

The largest number of observations is in Europe and Africa while the smallest number is in South East Asia. Most of the observations are clustered towards the bottom of the plot. Like the previous figures, there appears to be no relationship between the two variables in most of the regions as there is a random scattering of data points. This may be due to the lack of observations. Every region except South-East Asia has some significant outliers. Europe, America and the Western Pacific have a percentage of its population using basic sanitation, yet face a large number of cases. Africa has some of its countries showing a smaller percentage population using basic sanitation which faces a lower number of cases.

4.2. Maps

Map 1[17] shows the total number of covid-19 cases on a global level. The highest number of cases is in the United States followed by Brazil and then India with roughly 5 million, 4 million and 3 million cases respectively. Russia, South-American countries like Peru and Columbia as well as South Africa are at around a million cases. Most of Africa, Greenland and Australia has the least number of cases, lying below a million.

Map 2[18] shows the total population in each country. India has the highest population with more than 1.2 billion people. It is easily a great distance away from the population of the rest of the countries. The United States has the second-highest population at around 0.4 billion. Brazil, Pakistan and Nigeria seem to have a similar level of population at around 200 million. Australia, Greenland, Kazakhstan along with parts of northern Europe have a relatively small population.

Map 3[19] displays the number of tests conducted every one million people at a global level in each country. Russia has the highest number of tests at almost 200,000 followed closely by the United States and Australia in the higher 100,000 range. Canada follows with approximately 114,000 tests per million people while India and parts of Africa have the smallest number.

Map 4[20] displays the global health risk scores out of 100 assigned to each country in 2019 by the World Health Organization. The highest health risk score is given to Canada in the 90s followed by Australia, Brazil and North-European countries like Sweden in the higher 80s and lower 90s. India, Kazakhstan, parts of Northern Africa and Mexico lie in the lower 80s and high 70s. The smallest score, below 50, have been given to Malaysia, Botswana, Somalia and Belarus. It is observed from the map that there are a lot of missing values.

Map 5[21] displays the average annual temperature of each country in degree Celsius. The deeper the colour red, the higher is the temperature of the particular country. As can be observed Africa has some of the hottest countries given its proximity to the equator. Most of Africa, India, the Middle-Eastern countries, South American countries like Brazil and Columbia, as well as the islands in the Indian Ocean like Indonesia boast some of the higher temperatures at around 20-25 degrees, given the fact that they lie in the tropical zone. Australia lies at around 20 degrees.

[17] See "Map 1" on Page 24, "**4.2 Maps**" under the Figures Appendix
[18] See "Map 2" on Page 24, "**4.2 Maps**" under the Figures Appendix
[19] See "Map 3" on Page 25, "**4.2 Maps**" under the Figures Appendix
[20] See "Map 4" on Page 25, "**4.2 Maps**" under the Figures Appendix
[21] See "Map 5" on Page 26, "**4.2 Maps**" under the Figures Appendix

As we move further north, the temperature decreases drastically as seen in Ukraine and the European countries. Canada, Russia and Sweden have some of the lowest average annual temperatures at around 4-6 degrees.

Map 6[22] displays the percentage of the population using basic sanitation in a country in 2017. Canada, European countries, Chile, Kazakhstan and Australia boast some of the highest percentages close to 98%. Mexico, Brazil, Peru, Algeria and Saudi Arabia also have higher percentages, at around 90% of their population using basic sanitation in 2017. Most of the South-African countries like Nigeria and Sudan show less than 50% of their population using basic sanitation. There are also a number of countries that have missing data.

5. Regression

The paper runs various regression models on the dependent variable which is the total number of Covid-19 cases in a country. The chosen independent variables are the population of the country, the number of covid-19 tests conducted every 1 million people in the country and the annual average temperature of the country. Additional standardized independent variables such as the global health risk scores in 2019 and the level of basic sanitation usage in a country in 2017- scores are given by the World Health Organization- have also been considered for the model.

The independent variables in our regression models have been carefully chosen. The population of a country and the number of tests conducted every one million people in the population helps determine the number of covid-19 cases that occurs in the country. The susceptibility of a country to the virus in terms of how much risk the population of that country is at if there is a potential health risk is measured by the global health risk scores. The global health risk score of the most recent year of 2019, given by the World Health Organization might help determine the magnitude of cases in a country.

Since the onset of Covid-19, global health experts have been advising people to wear masks, use sanitisers and maintain proper hygiene as a preventive measure. Hence, the level of basic sanitization used in a country, as determined by the World Health Organization, may potentially be related to the number of cases occurring in a country. Moreover, in the initial stages of the pandemic, there were numerous reports and speculation about the virus being correlated to the weather. In an article by the Huffington Post, it was hypothesized that the virus spreads in colder weather[23]. Although this hypothesis is still being explored as numerous new variants arise, it might be a potential predictor and hence, has been chosen in this project as an independent variable.

We aim to infer what may cause the total number of covid-19 cases in a country. Five regression models are run with the first one being a simple linear regression and the subsequent ones adding our chosen independent variables at a time. For the variables with extremely large values, we convert them to the logarithmic form to define a change in those variables. This

[22] See "Map 6" on Page 26, "**4.2 Maps**" under the Figures Appendix

[23] Link: https://www.huffingtonpost.ca/entry/cold-weather-spread-covid-explainer_ca_5ff78146c5b612d958ea6d29

comprises of population, the total number of covid-19 cases, and the tests per one million people and the average annual temperature being converted to logarithmic values.

5.1. Model

Given that most of the variables chosen (except temperature) are discrete variables that either define a socio-demographic characteristic or is a standardized score, it can be assumed that the relationship between the dependent and independent variable is linear. Moreover, a non-linear relationship between the dependent and the independent variables emerge only if the dependent variable is a non-linear function of the independent variable. Among the variables chosen, the total number of covid-19 cases in the country can only be a function of the population and the tests conducted every one million people of the population. In both these cases, the dependent variable is a proportion of the independent variable. Hence the relationship between the dependent and the independent variables is linear.

Five models are chosen to run. Each model adds an explanatory variable. The models are:

Model 1: $\log(\text{TotalCases}) = \beta_0 + \beta_1 \log(\text{Population}) + \text{epsilon}$

Model 2: $\log(\text{TotalCases}) = \beta_0 + \beta_1 \log(\text{Population}) + \beta_2 \log\left(\frac{\text{Tests}}{\text{1Mpop}}\right) + \text{epsilon}$

Model 3: $\log(\text{TotalCases}) = \beta_0 + \beta_1 \log(\text{Population}) + \beta_2 \log\left(\frac{\text{Tests}}{\text{1Mpop}}\right) + \beta_3 \log(\text{Year\_temp}) + \text{epsilon}$

Model 4: $\log(\text{TotalCases}) = \beta_0 + \beta_1 \log(\text{Population}) + \beta_2 \log\left(\frac{\text{Tests}}{\text{1Mpop}}\right) + \beta_3 \log(\text{Year\_temp}) + \beta_4 \text{Risk\_Scores\_19} + \text{epsilon}$

Model 5: $\log(\text{TotalCases}) = \beta_0 + \beta_1 \log(\text{Population}) + \beta_2 \log\left(\frac{\text{Tests}}{\text{1Mpop}}\right) + \beta_3 \log(\text{Year\_temp}) + \beta_4 \text{Risk\_Scores\_19} + \beta_5 \text{basic\_sanitation} + \text{epsilon}$

5.2. Results

Table 1: Regression results

	Model 1	Model 2	Model 3	Model 4	Model 5
const	-5.19***	-14.10***	-12.86***	-13.23***	-13.14***
	(0.81)	(1.07)	(1.86)	(1.91)	(1.93)
log_pop	0.88***	1.06***	1.03***	1.03***	1.02***
	(0.05)	(0.05)	(0.06)	(0.07)	(0.07)
log_tests		0.63***	0.61***	0.54***	0.53***
		(0.06)	(0.08)	(0.09)	(0.09)
log_temp			-0.18	0.02	0.04
			(0.28)	(0.29)	(0.30)
basic_sanitation(%)				0.01	0.01
				(0.01)	(0.01)
Risk_Scores_19					0.00
					(0.01)
R-squared	0.58	0.75	0.68	0.68	0.68
R-squared Adj.	0.58	0.74	0.68	0.67	0.66
No. observations	208	191	144	127	127

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Table 1 above shows the coefficients obtained for each predictor in the different models. Each unit of observation pertains to a particular country and the number of observations in each model varies. The data is taken on a global scale.

The coefficients for all the predictor variables, except temperature, is positive. An increase in the population growth, as well as a positive change in the tests per one million people, leads to an increase in the total number of cases, on average. Moreover, the risk scores and basic sanitation level show almost no correlation with the total number of cases in a country. A positive change in the average annual temperature, that is, an increase in the average annual temperature of a country is correlated with a fall in the number of covid-19 cases by 0.18 %, on average, in Model 3. This estimate is not significant and cannot be generalized to the population level. The coefficients for population and tests per one million people are significant at a 1% significance level which implies that the results can be generalized to the population level.

The mean squared error or the average squared difference is 2.17. The mean squared error is calculated by minimizing the average squared distance between the predicted values of the total number of cases using the explanatory variables and the actual value of the former. It gives us a form of the accuracy of our prediction. However, our objective through this study is inference and not a prediction. Hence the mean squared error simply becomes a measure of the strength of our model.

6. Discussion

Our research question which aimed to determine whether there is a relationship between the total number of Covid-19 cases in a country and the chosen variables is explored through the regression models. Specifically, the regressions are a way to explain whether the variation

caused in the explanatory variable is correlated with the variation caused in the dependent variable.

Since the models, that is, Model 4 and Model 5 with basic sanitization level and global health risk scores have insignificant coefficients close to zero and low adjusted R square, we can conclude that there is no relationship between these variables and the total number of cases in the country. This means that the variation in health risk scores and sanitization level of a country do not explain the variation in the change in the total number of covid-19 cases in a country, on average. Thus, model 4 and model 5 can be ignored as we search for the optimal model. Model 4 has the smallest Akaike's information criterion of 438.8. The smaller the AIC of the model, the better it is. However, given that the additional variable in model 4, global health risk scores, has an extremely small coefficient close to zero which is insignificant and lower adjusted R square, we choose to ignore it despite it having the smallest AIC.

The population change and the change in the tests conducted every 1 million people in the population have a positive relationship with the dependent variable implying they are relevant explanatory variables. Temperature change has a negative relationship with the dependent variable as shown in Model 3. However, when we add temperature as a predictor variable, the adjusted R square decreases implying that the variation in the change in temperature does not explain much of the variation in the dependent variable. In other words, adding the change in temperature as an additional explanatory variable reduces the goodness of fit in the model since the variable does provide much information about the dependent variable and simply increases the noise in the data.

The most appropriate model is Model 2 comprising of two explanatory variables, namely, the population growth and the change in the tests conducted per one million people. This is because the model has the highest R square adjusted. Adding further explanatory variables decrease the R square adjusted, or rather the goodness of the fit, implying that the explanatory variables do not contribute relevant information regarding the dependent variable. When we add the change in the tests conducted per one million people as an explanatory variable, the R square adjusted increases suggesting that the variation in the variable explains some of the variation caused in the change in the total number of covid-19 cases. Upon adding further explanatory variables simply decrease the R-square adjusted and hence, we choose model 2 as the optimal model.

Under model 2, β_1 indicates that a percentage change in population leads to a 1.06% change in the total number of covid-19 cases in the country, on average. Similarly, β_2 indicates that a percentage change in tests conducted every 1 million people leads to a 0.63 % change in the total number of covid-19 cases in the country, on average.

It is to be noted that correlation among variables does not imply causation. We aim to infer a relationship between the explanatory variables and the dependent variable and not establish a causal relationship between them. In this case, it would be particularly hard to do so since there are a lot of observed and unobserved confounders that lead to the problem of endogeneity. This means that the error term is correlated with the explanatory variables which violates the conditional independence assumption of the error term. Hence a causal relationship under such circumstances cannot be established.

7. Weaknesses and Future Work

In an attempt to explore the potential relationship between the total number of covid-19 cases and other socio-demographic variables, we began by choosing population and the number of tests conducted every 1 million people in a country as independent variables. We, then, proceeded to consider adding global health risk scores in 2019 and the level of basic sanitization usage in a country, assigned by the World Health Organization as an independent variable in addition to socio-demographic factors. We added the climate of the countries as well. However, we noticed no relevant relationship between each of the variables and the total number of Covid-19 cases in a country, through visualizing them by graphs and maps.

To explore our research question in-depth, five regression models were examined and the R square adjusted was used to choose the most appropriate model. This particular model contained the change in population and the change in the number of tests conducted every 1 million people in a country as independent variables and showed a positive relationship between each of these with the change in the total number of covid-19 cases.

However, our model has several weaknesses. To begin with, there are a lot of missing values in our data. Moreover, the most appropriate model has a higher AIC compared to other models. There may be multicollinearity between the explanatory variables implying that the explanatory variables may be correlated among themselves. This would lead to the problem of confounders and hence, the variation in the change in the total number of cases may not be just due to the variation of a particular independent variable. There may also be a lot of outliers in the data. Finally, correlation among the variables does not imply causation. Hence we cannot say that the change in population or the tests conducted caused a change in the total number of cases.

In future, it would be beneficial to explore the outliers to check for leverage points. The potential confounders and the Gauss-Markov conditions for OLS can be examined by checking for multicollinearity. Moreover, if we want to establish causation then controls are to be considered and a randomized test, through instrumental variables, could be considered.

8. Conclusion

As the covid-19 pandemic continues to affect millions of people and nations worldwide, there are already a number of variants of the virus as well as vaccines. Studies all over the world are being conducted to find potential cures and causes. In this particular paper, we try to determine the effect of socio-demographic and various other variables on the total number of covid-19 cases in every country. Choosing population, the number of covid-19 tests conducted every 1 million people and global health risk score in 2019 as our initial independent variables, and the total number of covid-19 cases as our dependent variable, we calculate the summary statistics for them by categorizing each country into World Health Organization’s regional divisions.

We also used HTML based web scraping to obtain climate information from Wikipedia as well as the World Health Organization's Global Health Observatory to get an idea about countries' basic sanitation usage. These factors were used as independent variables and explored to determine their relevance to covid-19 cases in countries.

Moreover, we plot the relationship between the response variable with each of the independent variables through scatterplots and maps and discover no correlation for most of the variables. Therefore, to explore our research question in-depth, multiple regression models were examined and the R square adjusted was used to choose the most appropriate model. This particular model contained the change in population and the change in the number of tests conducted every 1 million people in a country as independent variables and showed a positive relationship between each of these with the change in the total number of covid-19 cases.

The optimal model obtained in this paper has the change in the total number of covid-19 cases as the dependent variable and the change in population and test conducted every one million people, as the independent variables. The coefficients obtained show a positive linear relationship between the independent and dependent variables. The coefficients are significant at a 1% significance level.

However, our model has several weaknesses including a higher AIC and the potential of having multicollinearity between the explanatory variables. Hence, in future, it would be beneficial to check for outliers and leverage points. Moreover, other methods such as instrumental variables can be used to establish a causal relationship between the explanatory variables and the dependent variable.

Citations

Date sources:

- "Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1".

- kp, D., 2020. COVID-19 Dataset: Number of Confirmed, Death and Recovered cases every day across the globe. Available at: https://www.kaggle.com/imdevskp/corona-virus-report/discussionhttps://www.kaggle.com/imdevskp/corona-virus-report/discussion

- World Health Organization. (2010–2019). National and Global Health Risk Scores [Dataset].Global Health Observatory. Retrieved from https://apps.who.int/gho/data/node.main.SDG3D?lang=en

- World Health Organization. (n.d.). Basic and safely managed sanitation services[Dataset]. Global Health Observatory. Retrieved March 21, 2021, from https://apps.who.int/gho/data/node.main.WSHSANITATION?lang=en

- Wikipedia contributors. (n.d.). List of cities by average temperature. Wikipedia. Retrieved March 21, 2021, from https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature

Other:

- Wikipedia contributors. (n.d.). COVID-19 pandemic. Wikipedia. https://en.wikipedia.org/wiki/COVID-19_pandemic

- Woods, M. (2021). Does Cold Weather Make COVID-19 Spread More Easily?. The Huffington Post. Retrieved 17 April 2021, from https://www.huffingtonpost.ca/entry/cold-weather-spread-covid-explainer_ca_5ff78146c5b612d958ea6d29

- Smeekes, S., & Wijler, E. (2018). *Macroeconomic forecasting using penalized regression methods*. International Journal of Forecasting. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207018300074

- Guo, B., Wang, Y., Pei, L., Yu, Y., Liu, F., & Zhang, D. et al. (2021). *Determining the effects of socioeconomic and environmental determinants on chronic obstructive pulmonary disease (COPD) mortality using geographically and temporally weighted regression model across Xi'an during 2014–2016*. Science of The Total Environment. Retrieved from https://www.sciencedirect.com/science/article/pii/S0048969720374003

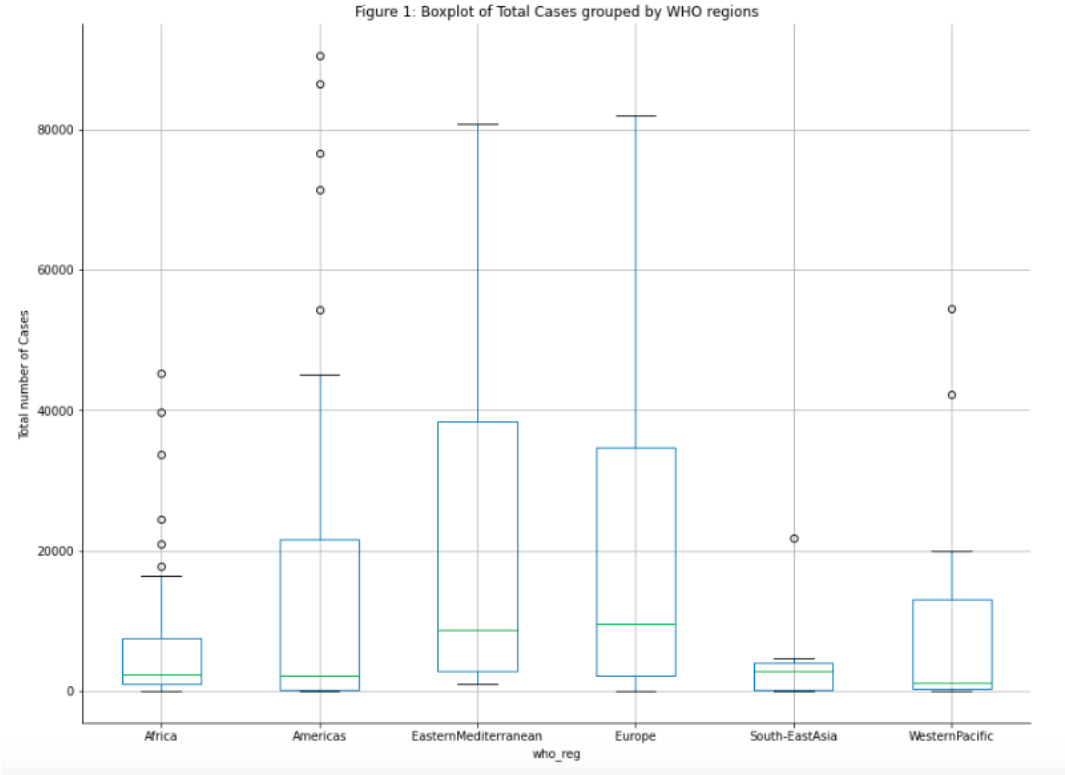
- Di Pietro, G. (2017). *Revisiting the impact of macroeconomic conditions on health behaviours*. Economics & Human Biology. Retrieved from

https://www.researchgate.net/publication/320925730_Revisiting_the_impact_of_macroeconomic_conditions_on_health_behaviours

- Rath, S., Tripathy, A., & Tripathy, A. (2020). *Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model*. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. Retrieved from https://www.sciencedirect.com/science/article/pii/S1871402120302939

Figures Appendix

3. Summary Statistics:



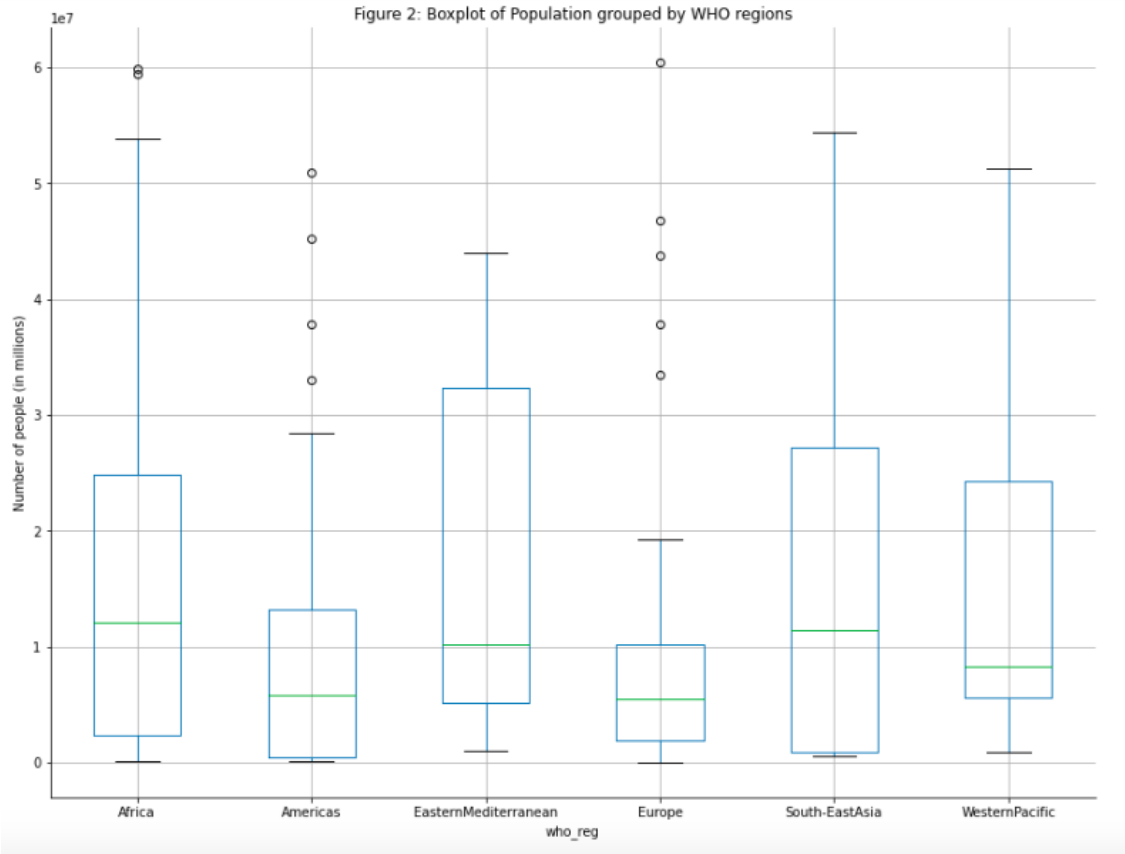


Figure 3: Boxplot of Tests/ 1 million people grouped by WHO regions

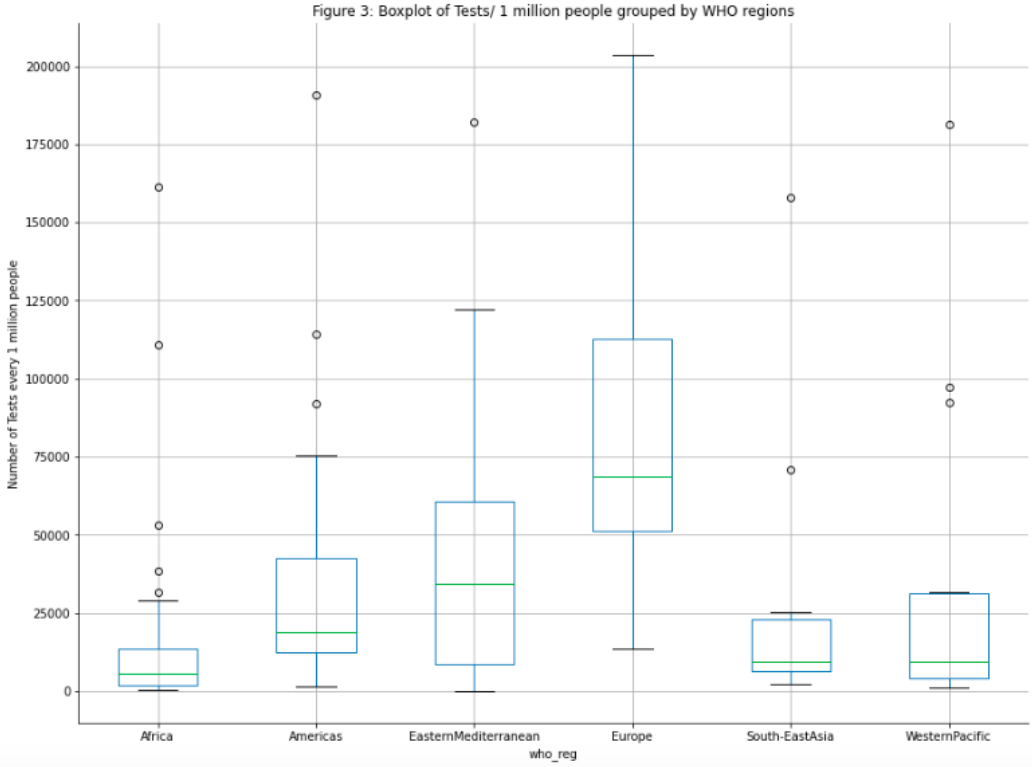
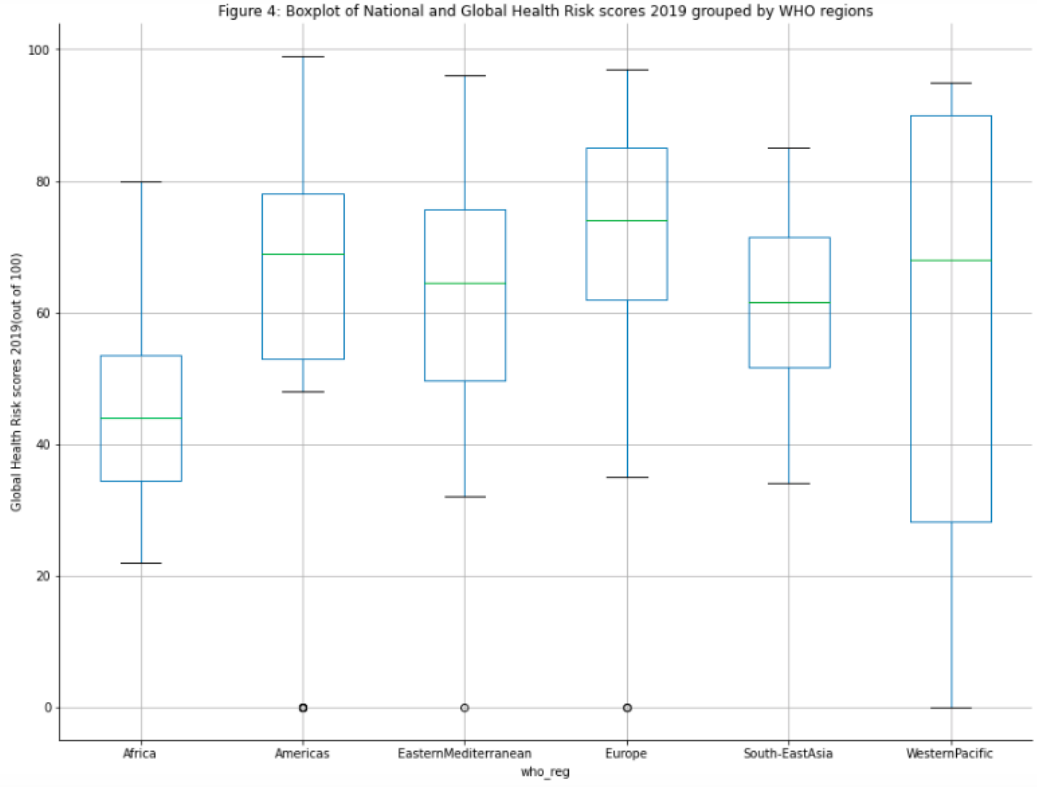


Figure 4: Boxplot of National and Global Health Risk scores 2019 grouped by WHO regions



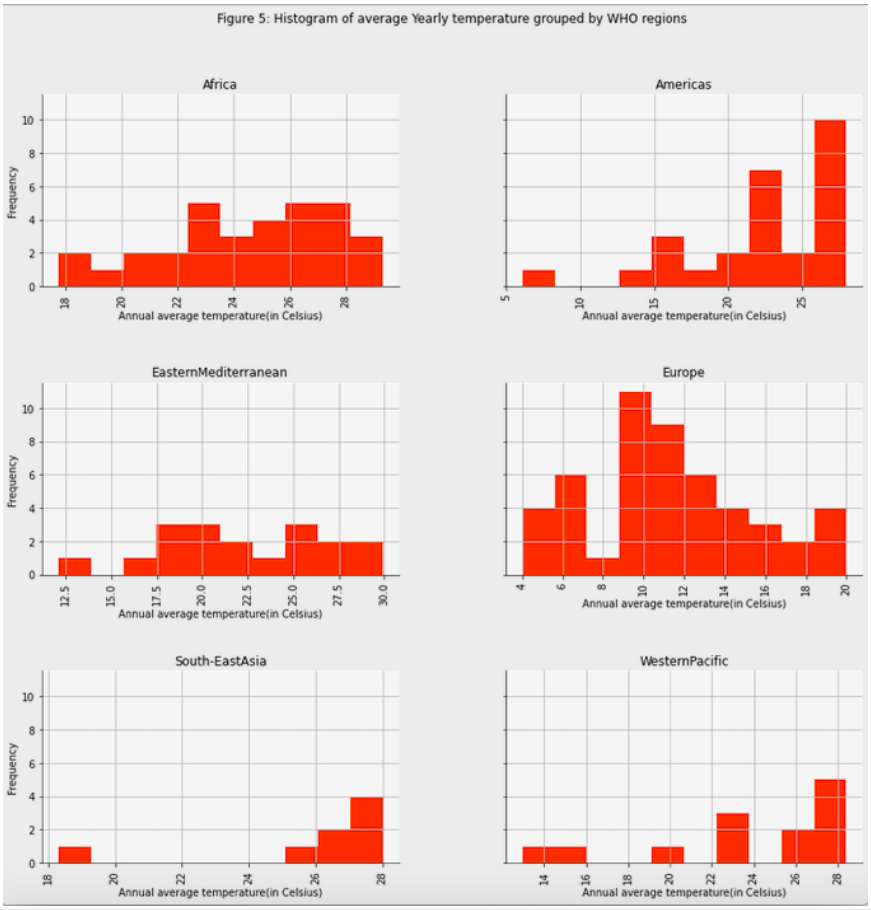


Figure 5: Histogram of average Yearly temperature grouped by WHO regions
Africa
Americas
EasternMediterranean
Europe
South-EastAsia
WesternPacific
Frequency
Annual average temperature(in Celsius)

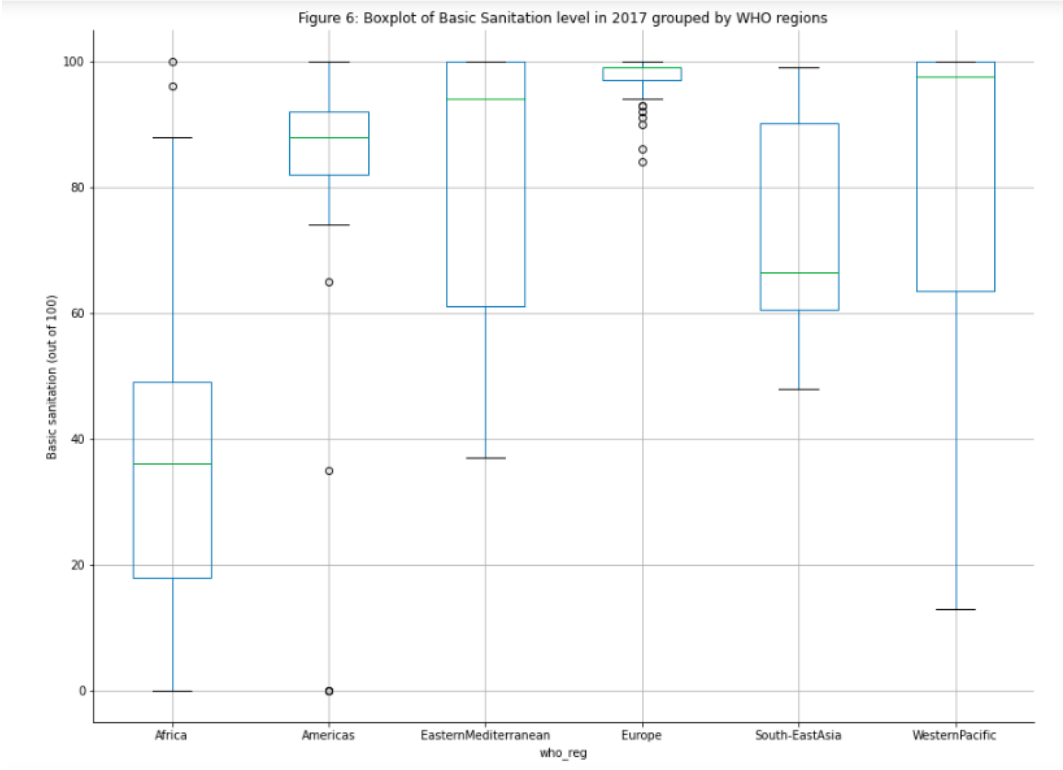


Figure 6: Boxplot of Basic Sanitation level in 2017 grouped by WHO regions
Basic sanitation (out of 100)
Africa
Americas
EasternMediterranean
Europe
South-EastAsia
WesternPacific
who_reg

4.1 Graphs:

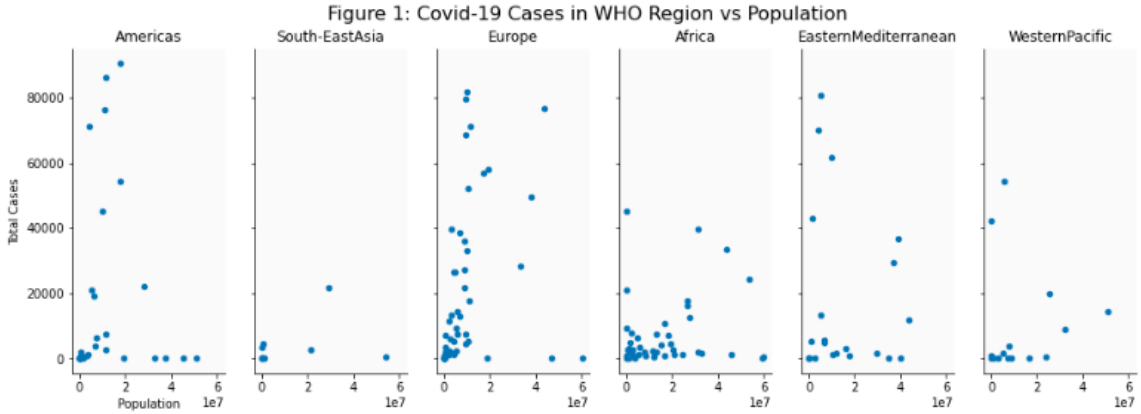


Figure 1: Covid-19 Cases in WHO Region vs Population
Americas
South-EastAsia
Europe
Africa
EasternMediterranean
WesternPacific
Total Cases
Population
1e7

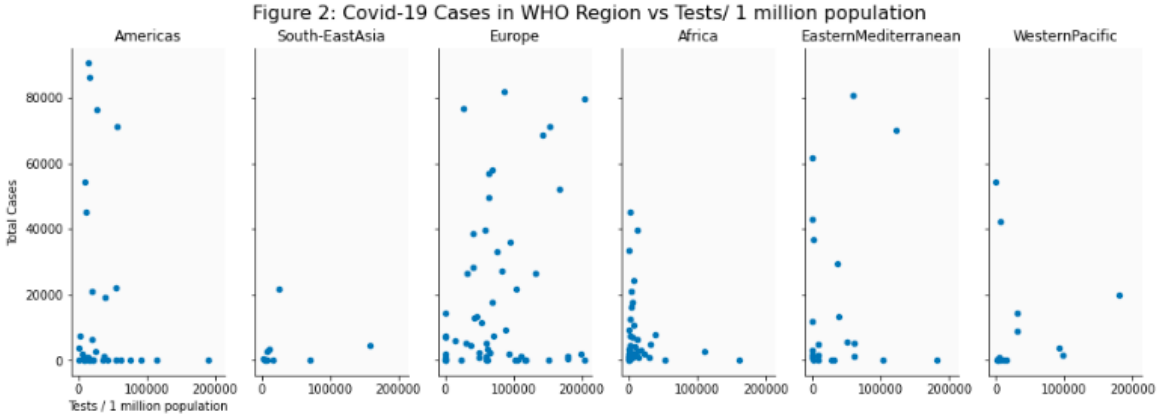


Figure 2: Covid-19 Cases in WHO Region vs Tests/ 1 million population
Americas
South-EastAsia
Europe
Africa
EasternMediterranean
WesternPacific
Total Cases
Tests / 1 million population

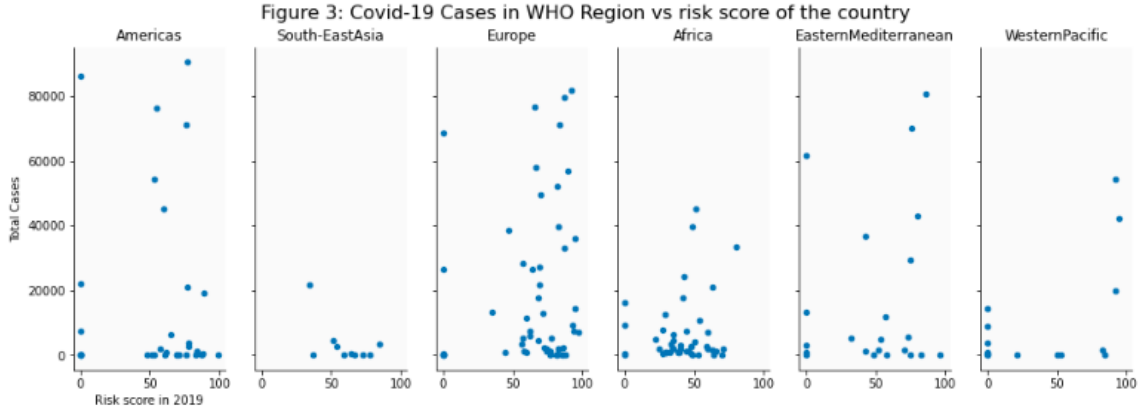


Figure 3: Covid-19 Cases in WHO Region vs risk score of the country
Americas
South-EastAsia
Europe
Africa
EasternMediterranean
WesternPacific
Total Cases
Risk score in 2019

Figure 4: Covid-19 Cases in WHO Region vs Average Yearly Temperature of the country

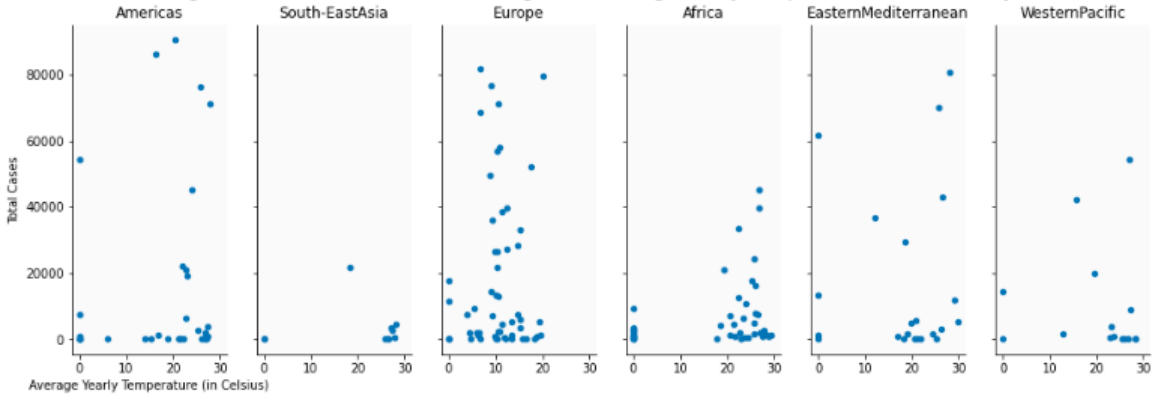
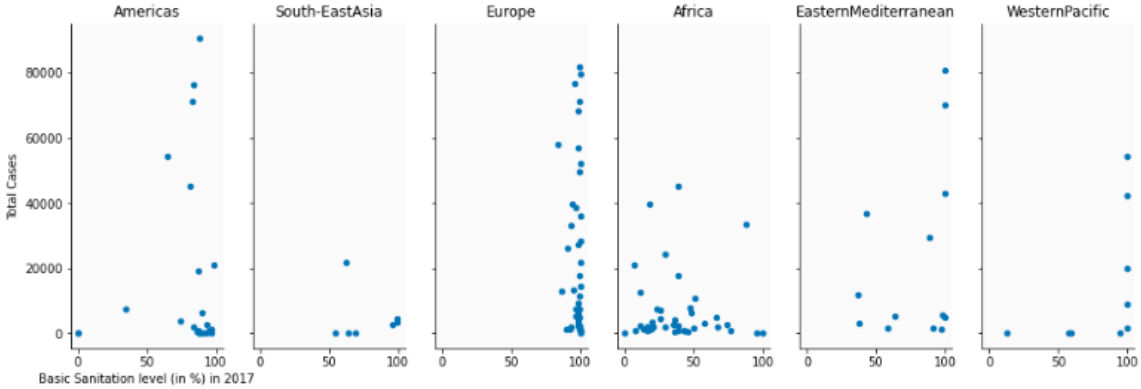
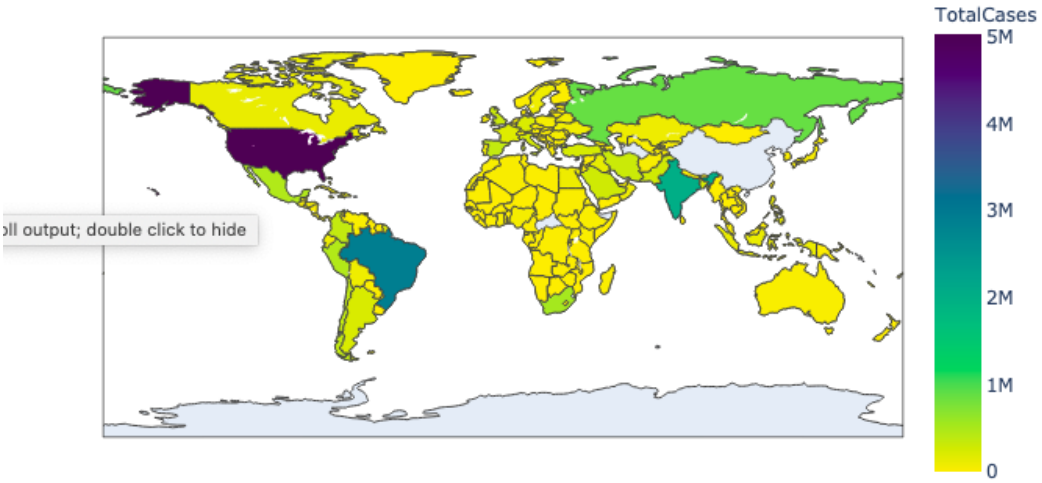


Figure 5: Covid-19 Cases in WHO Region vs Percentage of Population using Basic Sanitation in 2017

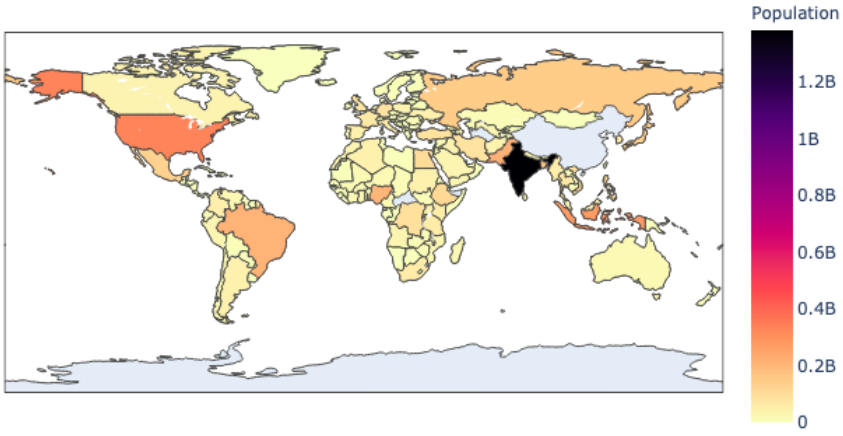


4.2 Maps:

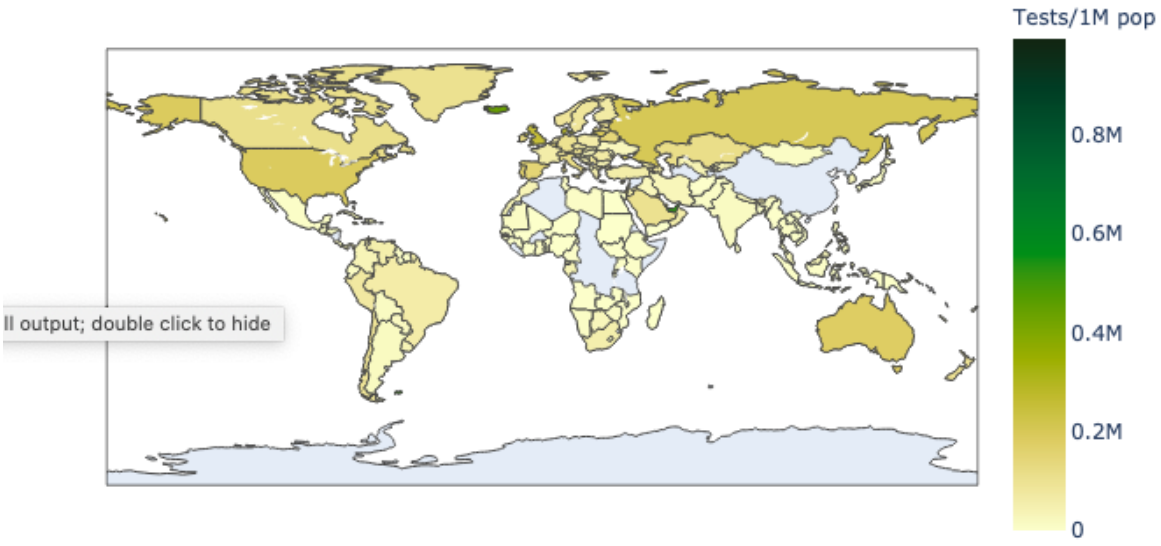
Map 1:Total Number of Covid-19 Cases



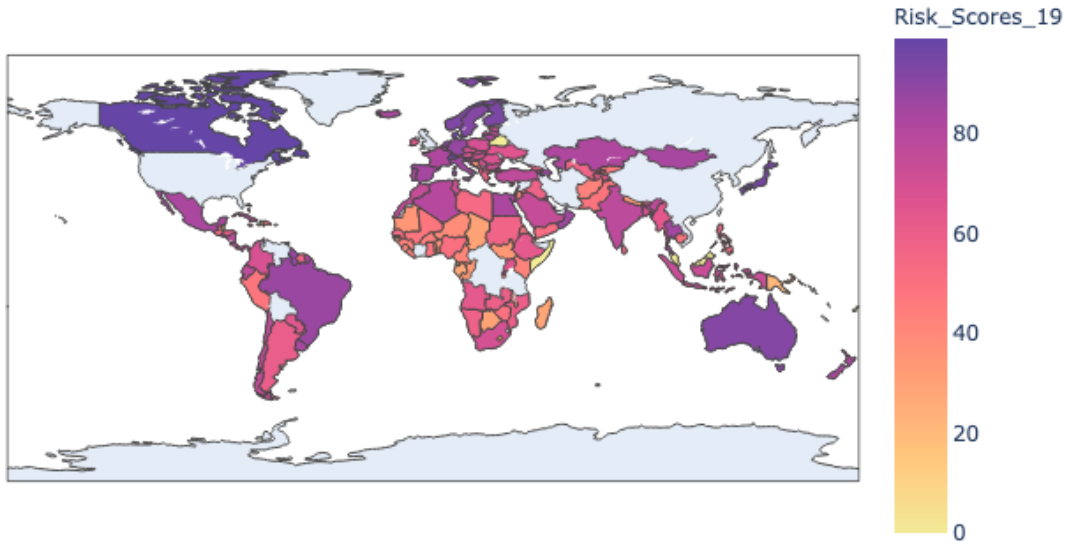
Map 2: Total Population



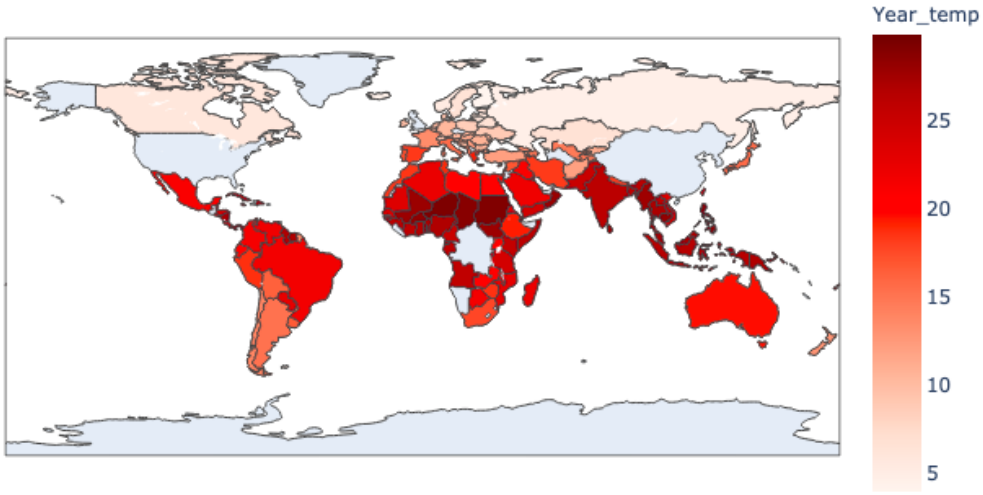
Map 3: Total Number of Covid-19 Tests conducted every 1 million people



Map 4: Level of Risk Scores in 2019



Map 5: Average Annual Temperature



Map 6: Level of basic sanitation in 2017