Simon Foo
Comprehensive Assessment (ESP)

Analysing quantitative patterns in Covid-19 vaccinations through data processing and statistical evaluation

Submitted 26 July, 2021
for the Diploma of College Studies

Statistics and Probability,
201-HTH-05 Vanier College

Presented to:

Professor Chantal Linda Desrosiers

Abstract

This research paper aims to use the open-source data around us, analyse it and turn the data into useful information. Information like the number of daily vaccinations can be used estimate the number of people going in the following days using Poisson distribution. In a perfect world, that would have been enough and the number of people showing up everyday would equal the known average. Unfortunately, we will need to include a margin of safety also called the confidence interval. Furthermore, a method of approximation will be used for efficiency in computations.

The second part of the paper will introduce python-plotted graphs, in order to find patterns or correlation between data(All data manipulation and processing will be done by and on python). We are looking for grouped points hinting at a relationships between our variables, Covid test per thousands people and GDP per capita in USD. This relationship can than be represented by a line and we can use that same line to estimate or predict the amount of covid test taken given a country GDP. In addition to finding the best fit line, concepts concerning curve spread will be used, such as, Variance, Covariance and Mean.

Contents

1 Introduction

During these extreme times, we are surrounded by not only family but computers and electronics. Our exposure to news drastically increased due to the extra time we have on our hands. We are constantly faced with updates of Covid-19 vaccines efficiency, live outbreaks happening around the world and the increase or decrease in daily cases in each countries. Times like these makes us realize the abundance of collected data.

What can be done with the pool of data?

Perhaps this question can be broken down into two halves:

Question 1. How can we use the data?

Question 2. How can we apply it?

We are dealing with large files of data, so we have to ask: what must we do to select the right data? Then, how can we use this data selection and learn from it? These questions will be answered with 2 examples to show the practicality of data in the following research paper. All the data from this research paper was gathered through publicly available information by official sources and from many small surveys from manufacturers, builders

and retailers, as well as from trade and financial flows. These surveys are collected in order to estimate main factors of GDP, investments and production(net exports). The collected data is then entered into a CSV files for our use.

2 Daily Vaccinations: Assuming Consistency

2.1 Poisson Distribution

Noted as X ˜ P(λ)

The Poisson probability distribution is used to calculate the probability of occurrences in a scenario where a time interval and a known average is given. Notably, every occurrence should be independent of each other, meaning, the event or events that took place before X have no influence over the probability of X.

Therefore, the Poisson probability distribution can be applied on the data set of number of Covid-19 vaccinations. In order to do so, a daily average of number of doses will need to be calculated using,

$$\lambda = \frac{\sum_{i=1}^{N} x_i}{N}$$

where

λ = average amount of occurrences in a time interval

x_i = number of doses distributed in a day

N = number of days

2.1.1 Poisson Probability

In order to find the probability of a number of occurrences using Poisson distribution, the following equation is used,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(X \leq x) = \sum_{i=0}^{x} \frac{\lambda^i e^{-\lambda}}{i!}$$

$$P(X \geq x) = 1 - \sum_{i=0}^{x} \frac{\lambda^i e^{-\lambda}}{i!}$$

Notice that i (The number of doses distributed in a day) starts at 0 here, since no occurrences is a probability.

2.2 Poisson approximation using normal probability distribution

X ˜ P(λ) $\approx$ X ˜ N(μ, σ)

To be able to use the normal distribution to approximate a Poisson probability distribution, λ(the average) must be a large number($\lambda \geq 5$). Since we are dealing with large numbers, the expected value of the Poisson will be qualified to be equal the mean of a normal distribution ($\lambda = \mu$). In addition, for the second parameters of the normal distribution, it is simply $\sigma = \sqrt{\lambda}$.

The normal distribution curve is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

where the curve is centered on μ and the spread is determined by the standard deviation(σ). A property possessed by all normal distributed curve is the equal area under the curve of 1.

Hence, the curve is a probability density function, in which, the area equals to the probability of the given standardized bounds. For example, the probability of an event happening between A and B is,

$$P(A < X < B) = \int_A^B \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

To integrate the function above manually requires knowledge of the error function($erf(x)$). Luckily, the calculations are done by computers creating a table, called the z-score table.

2.2.1 Continuity Correction Calculation

Before being able to approximate a Poisson distribution, there is a slight adjustment that needs to be done. It is the continuity correction calculation, it is essentially a transformation that converts every integer into intervals. Otherwise, finding the probability of an integers in range of finite numbers would be equal to 0. For simplicity, these are all the equations to correct every integer into their proper interval.

$$P(X = A) = P(A - 0.5 < X < A + 0.5)$$

$$P(X < A) = P(A < x - 0.5)$$

$$P(X \leq A) = P(X < A + 0.5)$$

$$P(A < X \leq B) = P(A + 0.5 < X < x + 0.5)$$

$$P(A \leq X < B) = P(A - 0.5 < X < x - 0.5)$$

$$P(A \leq X \leq B) = P(A - 0.5 < X < B + 0.5)$$

Therefore, to use the probability density function to approximate a Poisson distribution, the bound needs to be changed. Thus, the equation becomes,

$$P(A < X < B) = \int_{A+0.5}^{B-0.5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

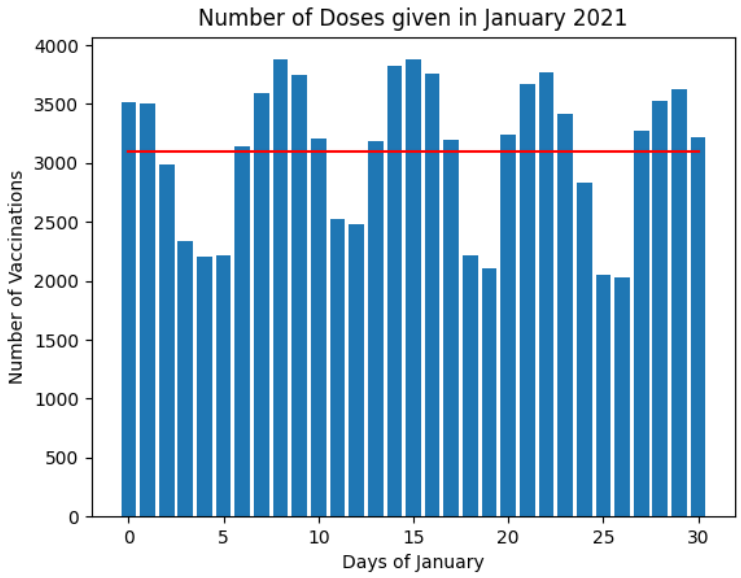
2.3 Application of Poisson

The likelihood of people getting vaccinated are influenced by many factors, such as, vaccine shortages, insufficient incentives by governments, lack of trust, the absence of urgency, and many more. Hence, the distribution can only be used in small time intervals, since daily averages could vary drastically from a season to another due to macro factors. For instance, vaccination centers could use the Poisson probability distribution to estimate the amount of patient coming through the door the next day using monthly-based data. Using the estimation, it allows the clinics to prepare sufficient doses for everyone that wants to get vaccinated. On the other hand, it would

be inaccurate for a center to calculate λ using all the data since countries has started administering Covid-19 shots.

2.4 Calculation of Poisson

Let's take Canada for the purpose of the calculations and set the time interval to be the first month of 2021. Using python, the data found in the csv file from OurWorldInData can be manipulated to yield the average of 3,100. Since the value of vaccine shots are per thousands, the actual average of doses per day during the month of January 2021 is 3100. Therefore the standard deviation would be $\sqrt{3100}$ or 55,68. Knowing the two parameters for a normal distribution and given an arbitrary confidence interval(CI), an interval of mean can be determined.



Since n is greater than 30, the z-test will be used and,

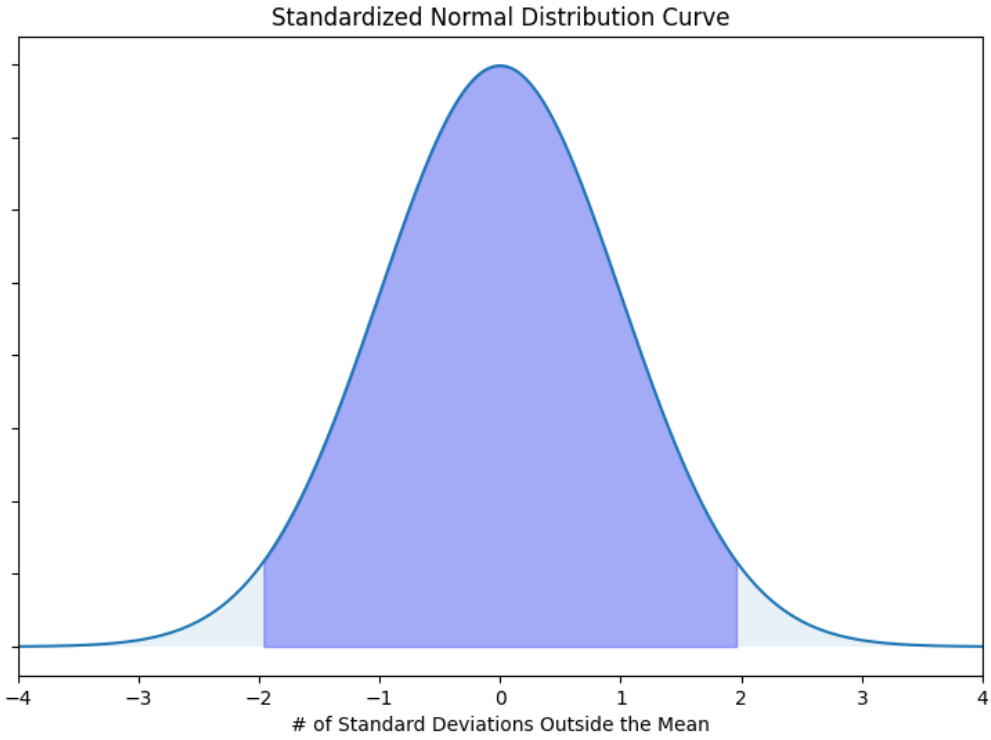
$$z = \frac{x - \mu}{\sigma}\sqrt{n}$$

thus,

$$x = \frac{z\sigma}{\sqrt{n}} + \mu$$

Let CI = 0.95

The Z-score for $\alpha_{/2}$ of 0.95 is ±1.95996.



The confidence interval is found as such,

$$\mu_{min} < \mu < \mu_{max}$$

$$\frac{-z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} + \mu < \mu < \frac{+z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} + \mu$$

$$\frac{-z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} + \mu < \mu < \frac{+z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} + \mu$$

$$-19,60 + 3100 < \mu < 19,60 + 3100$$

$$3080,40 < \mu < 3119,60$$

Using this method, Canadian clinics can, with a confidence interval of 0.95, expect between 2437.85 and 2462.73 the next day. Hence, they can prepare a little over 2463 vaccines to make sure that every patient will get a dose. In fact, on February 1st 2021, there was 2704 vaccine distributed in Canada, which means that confidence interval failed and overshot it.

The probability that there is 2704 or less shot administered on day can be found using the Poisson approximation instead of summing every probability from 0 to 2704 using the Poisson probability distribution.

X˜N(μ, σ)

$$P(X < \mu_{min}) = P(Z < \frac{\mu_{min} - \mu}{\sigma}\sqrt{n})$$

$$P(Z < -39.60) = 0$$

Thereby, it was highly unlikely that there will be 2704 or less. Although it was an overshot, it also means that there was enough vaccine for everyone and some surplus as backups.

3

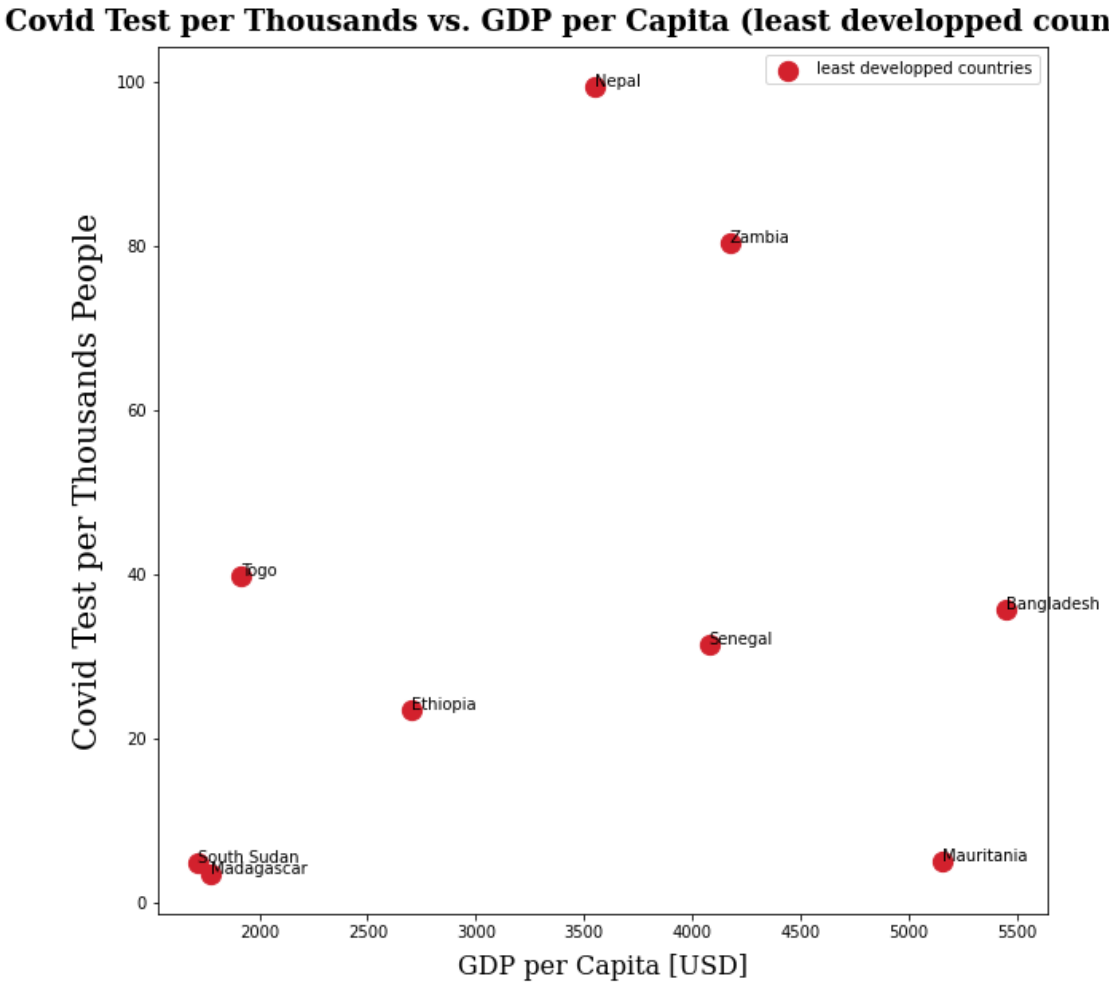
External Factor:

Country GDP per Capita

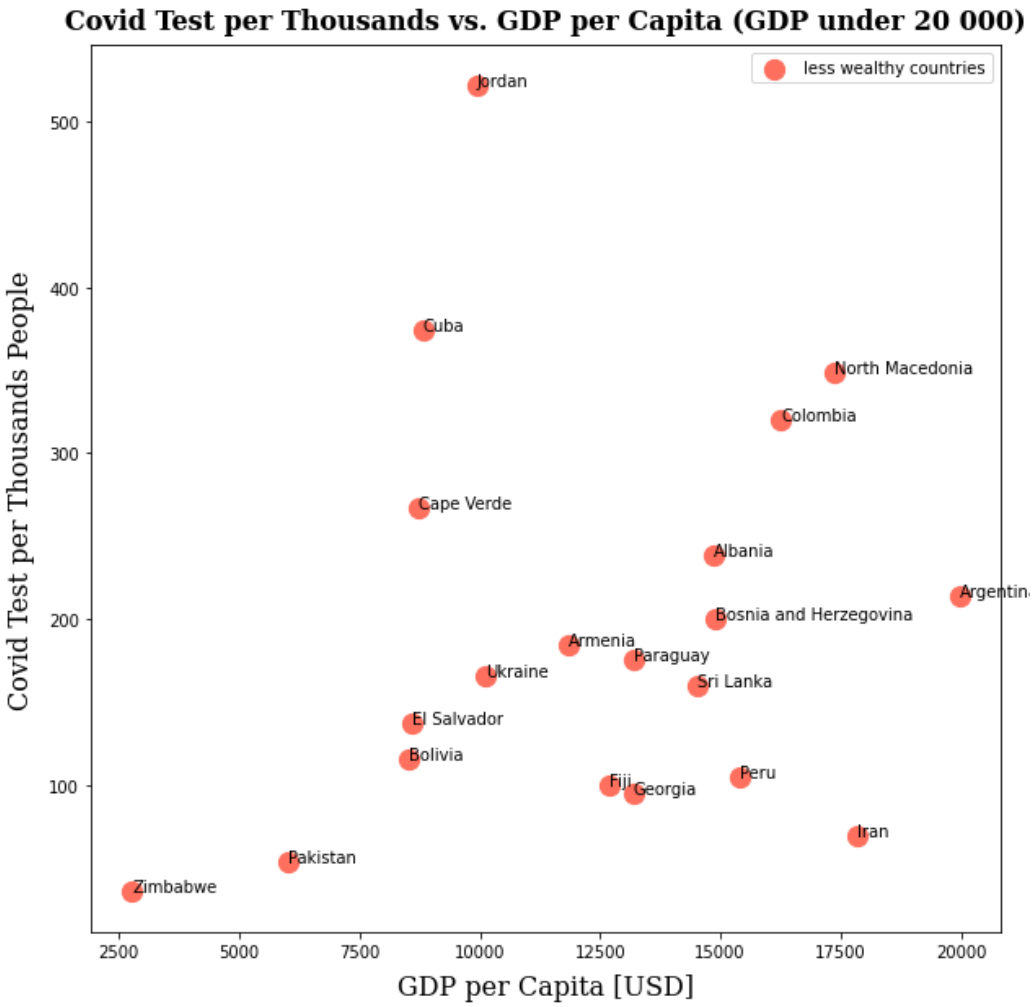
3.1 Graphing the data

First and foremost, the countries will be separated into three categories or samples. They will be color coded and divided depending on their wealth. For all python code regarding the scatter plots, you can find it on Colab: https://drive.google.com/file/d/1F4zaKC44bbDqETFjLFnsfqmS126E-0yn/view?usp=sharing.

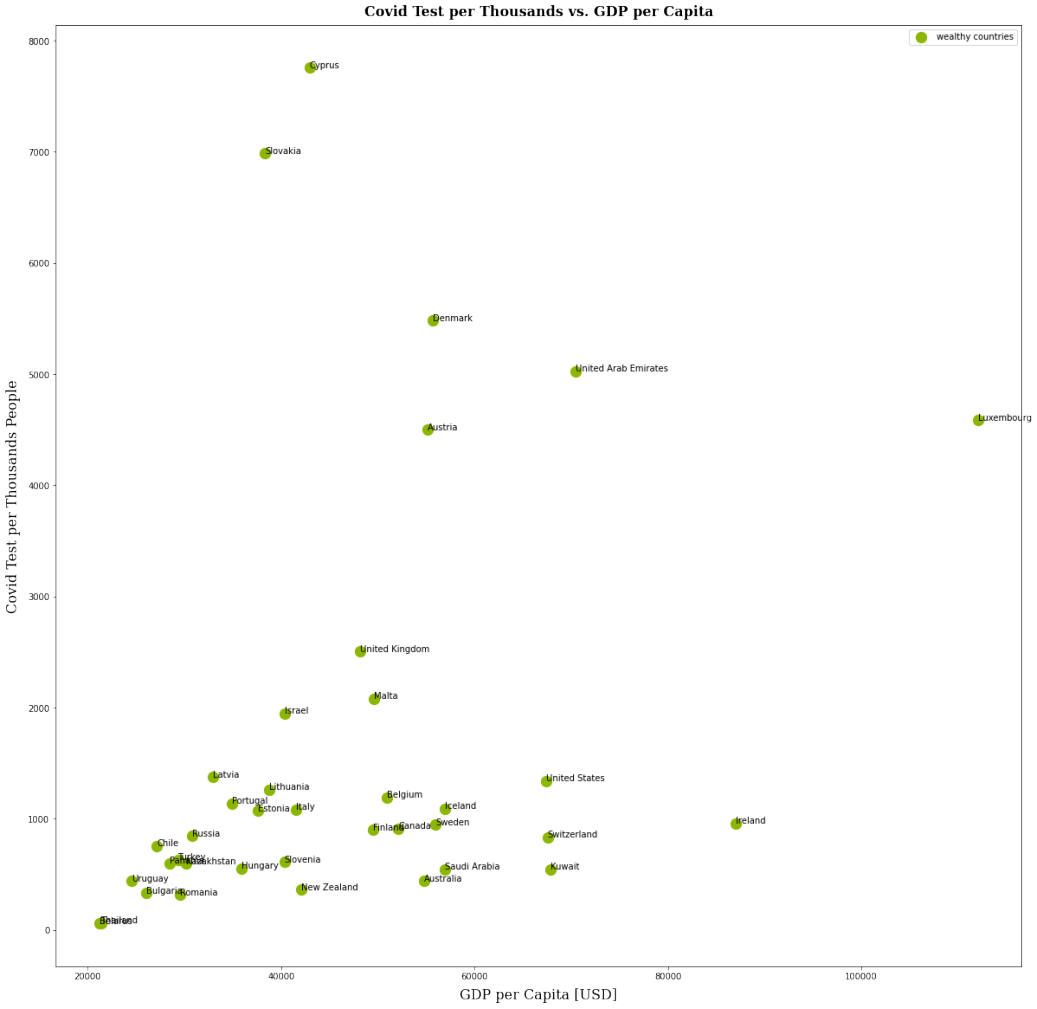
Plotting the first set of data yields:



Looking at the scatter-plot, there isn't much to say about it. The GDP per Capita seems to have no effect on the number of Covid test taken in countries.

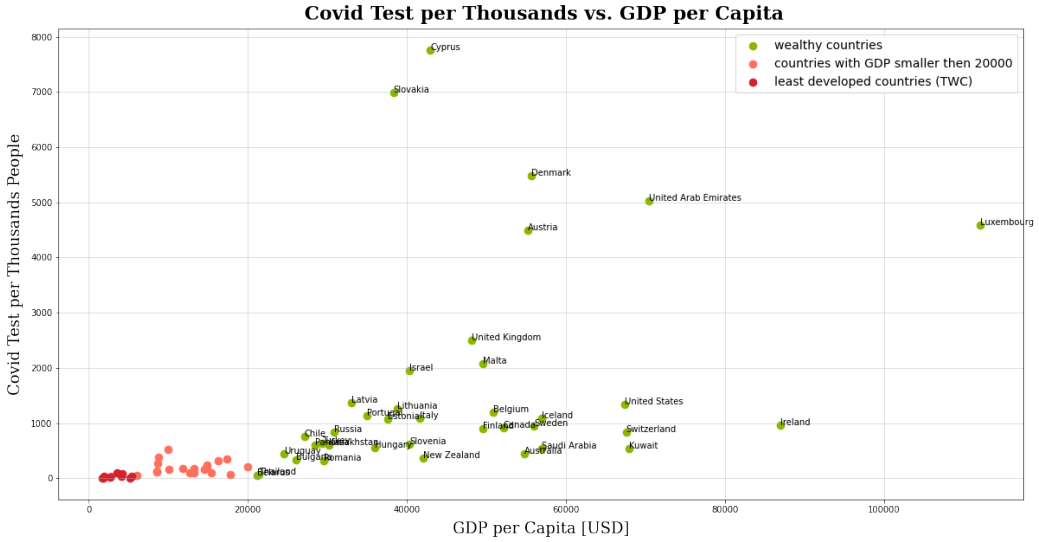


Apart from a few countries like Jordan and Cuba, the points are starting to clutter up near the (15000, 200) point. There seems to also have a sort of linear correlation from Zimbabwe to North Macedonia.



Looking at the graph of wealthy countries, there is a significant number of countries showing correlation. It also shows the type of correlation between GDP per capita and the number of covid test, a proportional relationships.

Lastly, putting all points together in one graph, the following is obtained:



At first, in the third world countries, it seemed like country GDP was not a factors playing a role in influencing the numbers of Covid-19 shots. Another perspectives, could be that the changes are negligible when looking at the big picture. Furthermore, now looking at the data of the entire world, there is clear correlation. However, notice that some countries are way off the clutter of points. These outliers can be explained by their small population size relative to the rest of the world. Countries like Cyprus and Slovakia has a population of 875.899 and 5.45 million respectively(Data collected from Eurostat). As the points lower and gets closer to the general population of points, the population size of the country increases. For example Denmark, which is a bit lower than Slovakia, has a population of 5.79 million. Then a bit lower there is Austria, population of 8.86 million, and United Arab Emirates, population of 9.77 million.

A reason of these abandoned points could be that smaller countries won't run into problems like vaccine shortages. A different explanation could be simply because of a smaller sample size, thus the data collected could be inac-

curate or less accurate compared to large countries. The law of large numbers states that as n(number of trials) goes to infinity, the average will approach the expected value. In this case, the expected value can be calculated with a regression.

3.2 Expected value

3.2.1 Variance

Denoted as Var(x) or σ^2

The Variance is a function that measures the spread of a set of data compared to the mean. Population variance is found with the following equation:

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

3.2.2 Covariance

Denoted as Cov(x, y)

The Covariance is a measure for a the spread of two variables to their respective means. However, a known property of the covariance is its ability to measure proportionality. Which indicates, if two sets of a data are linearly proportional, the covariance will show a positive number and if the two given variable are inversely proportional, the covariance will be negative.

The Covariance is found using,

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$$

3.2.3 Mean

Recall that the average was found using:

$$\lambda = \frac{\sum_{i=1}^{N} x_i}{N}$$

Slightly modifying the variables, gives the equation for the means of x and y:

$$\mu_x = \frac{\sum_{i=1}^{N_x} x_i}{N_x}$$

$$\mu_y = \frac{\sum_{i=1}^{N_y} y_i}{N_y}$$

3.2.4 Linear Slope

Best fit line: f(x) = mx + b

where:

$$m = \frac{Cov(x, y)}{Var(x)}$$

substitute Cov(x, y) and Var(x)

$$m = \frac{\frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}}{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

the n cancels out and yields the final equation

$$m = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{n}(x_i - \mu)^2}$$

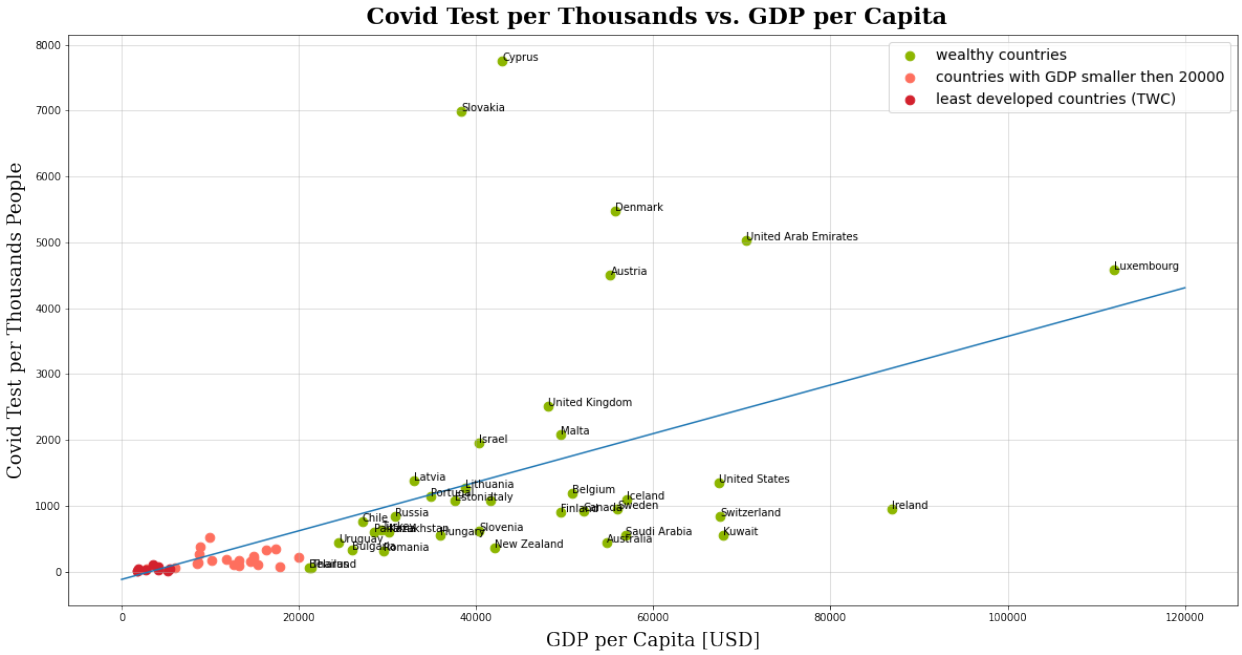
Subsequently, to find the b for the best fit line, a points on the plot needs to be given. Since, there is no way to know which point lays perfectly on the line, the mean of x and y will be used instead. Thus,

$$\mu_y = m(\mu_x) + b$$

$$b = \mu_y - m(\mu_x)$$

Solving for the best fit line by hand would be unpleasant, therefore, python will be used instead for efficiency.

Plotting the best-fit line on the graph above:



Apart from the few outliers, the line fits quite nicely on the points. In fact, the correlation coefficient r can determine the accuracy of the line.

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

Inserting the numbers into python, it outputs a r = 0,86. Thus, r^2 = 0,7396. This means that around 0.74 of the data can be explained by the linear regression. Therefore, the linear regression can also be a tool for predictions.

4 Conclusions

4.1 Conclusions

In this paper we used datasets for the GDP per capita of countries and the Covid tests performed. The purpose of our project was to determine if there was a correlation in the number of tests performed using data available to everyone. Testing to see if we would see a correlation or patterns and based on the information that was graphed we can see that indeed there was. In the daily vaccination, we saw that the daily dose performed in the beginning of the month is way below the calculated average. However, the bar chart showed that the month of January was consistent, demonstrating a monthly fluctuation. In GDP the graph, we can see that the least developed countries, in orange and in red are the ones that also performed the least amount of tests done. The results of these graphs show us something very important. Less wealthy countries have a bigger chance of spreading the virus because they may be unaware of their contamination due to the lack of tests performed on citizens. Thus, when analysing data sets such as the number of cases of Covid 19 per country, these numbers wouldn't be truthful as we now know that poorer countries tend to have more cases that pass unnoticed due to having less Covid 19 tests performed.

References

Hasell, J., Mathieu, E., Beltekian, D. et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020). https://doi.org/10.1038/s41597-020-00688-8

"Kaggle - https://www.kaggle.com/daniboy370/world-data-by-country-2020?select=GDP+per+capita.csv"

"MakingPrettyPictures - Solutions - 2021.ipynb"

"Scatter Plots with a Legend." Scatter Plots with a Legend - Matplotlib 3.4.2 Documentation. Web. 03 June 2021.

"Matplotlib Scatter Plot with Different Text at Each Data Point." Stack Overflow. 01 Sept. 1961. Web. 03 June 2021.

This Research Paper was written on

using

The graphs were rendered on

programmed in