[REPORT] COVID-19 DATA TABLE ANALYSIS IN R

Core Bioinformatics (Statistics) | MSc in Bioinformatics | UAB
by Víctor Fernández Oliveras / Professor: Antonio Barbadilla

Data source: https://www.worldometers.info/coronavirus

Code available: https://github.com/victor-fdz

INTRODUCTION

I selected the COVID-19 dataset (Table 1) because it had a type of information I had never analized before: epidemiological data. The dataset contains information of population, COVID-19 cases/deaths/tests and variable rates for 230 regions, either countries, overseas regions and others. Moreover, I introduced a new column with the Case Fatality Rate (deaths/cases*100). This dataset aims to be a census, not a sample. However, since 2024 it stopped updating data, so the analyisis is performed with the April 13th 2024 situation and some fields for certain countries are missing (NA values). I decided to present the results in this poster-like format because of its originality and fancy design for the reader.

Table 1: 10 first entries of the used COVID-19 dataset.

	order	Country_Other	Total_Cases	Total_Deaths	Total_Recovered	Active_Cases	Tot_Cases.1M_pop	Deaths.1M_pop	Total_Tests	Tests.1M_pop	Population
1	1	USA	98572011	1087976	95773910	1710125	294139	3247	1122147003	3348488	335120523
2	2	India	44616394	528822	44060198	27374	31625	375	896987772	635804	1410792039
3	3	France	35875626	155535	34817884	902207	546875	2371	271490188	4138498	65601133
4	4	Brazil	34766204	686928	33926118	153158	160964	318	63776166	295277	215987681
5	5	Germany	34121168	15072	32701100	1269348	404319	1786	122332384	1449579	84391659
6	6	S._Korea	24995246	28708	24536323	430215	48658	559	15804065	307656	51369259
7	7	UK	23735273	190888	23394134	150251	345517	2779	522526476	7606465	68695046
8	8	Italy	22830825	17757	22149828	503427	378871	2947	248648092	4126240	60260214
9	9	Japan	21564995	45538	20433996	1085461	171698	363	76201407	606706	125598560
10	10	Russia	21232963	388404	20470552	374007	145355	2659	273400000	1871624	146076385

Showing 1 to 10 of 230 entries Previous 1 2 3 4 5 … 23 Next

USE OF AI DECLARATION

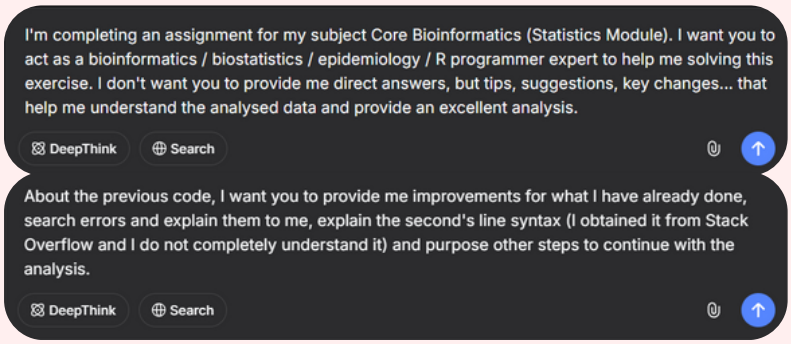


Fig. 1: Examples of prompts sent to DeepSeek.

I delcare I used generative-AI resource DeepSeek to assist me during the data analysis. Two clear examples of its usage are provided in Fig. 1. Using this tool allowed me faster and further analysis while learning new insights in R syntax, packages and general options.

HYPOTHESIS & QUESTIONS

With the analysis of this data it is aimed to answer to these questions:

- Do any of the rates (Cases/M, Deaths/M, Tests/M or Death rate) follow a normal distribution?
- Which countries or zones had higher rates for testing, cases and deaths?

And test these hypotheses:

- There is a correlation between Cases/M and Deaths/M.
- There is a correlation between Cases/M and Tests/M.
- There is a correlation betwen Population size and Tests/M.
- There is a correlation between Tests/M and Deaths/M.

DESCRIPTIVE STATISTICS

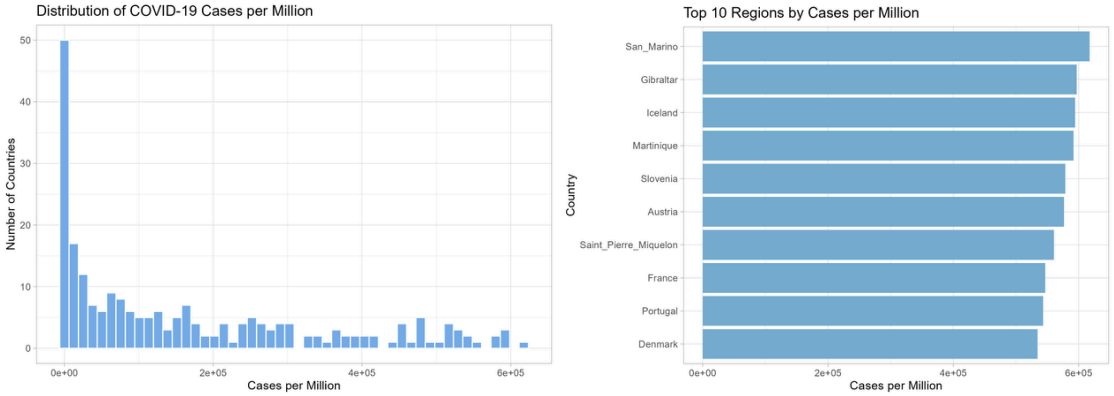
Initial descriptive analysis (Fig. 2) show an extreme disparity among countries, with high differences between medians and means and extremely large minmax ranges.

```
Tot_Cases.1M_pop Deaths.1M_pop    Tests.1M_pop      Death_Rate
Min.   :    16   Min.   :   2.0   Min.   :   2499   Min.   : 0.00155
1st Qu.: 10898   1st Qu.: 133.5   1st Qu.:  142138   1st Qu.: 0.35091
Median : 92506   Median : 615.0   Median :  850715   Median : 0.87055
Mean   :162096   Mean   :1108.5   Mean   : 2034940   Mean   : 2.09527
3rd Qu.:269809   3rd Qu.:1763.5   3rd Qu.: 2331568   3rd Qu.: 2.05695
Max.   :617517   Max.   :6373.0   Max.   :22006216   Max.   :25.42819
NA's   :2        NA's   :7        NA's   :16         NA's   :5
```

Fig. 2: Trimmed screenshot of R's output for summary(dataset).

DISTRIBUTIONS

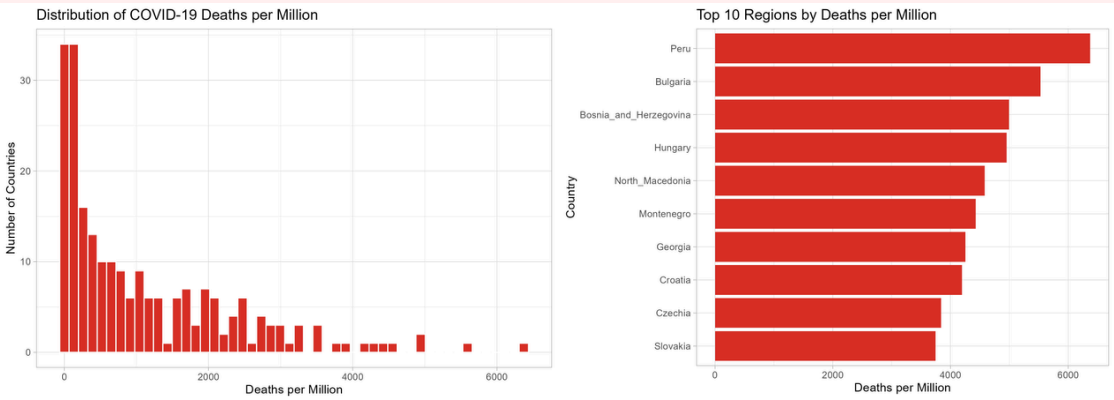
The distributions of Cases, Tests and Deaths per 1 Million population, as well as of Case Fatality Rate (CFR), were plotted (Fig. 2-5) to visually inspect them. The top 10 countries' values for each variable were also plotted (Fig. 2-5) to visually inspect differences or tendencies. To extract meaningful conclusions of these type of analysis, they should include other socioeconomic and geographical data. For these reason, the comments on this part are merely descriptive, with no epidemiological conclusions.



Not normally distributed, with low values being the more frequent.

Top 10 regions are mainly European countries and there is a high diversity of population size.

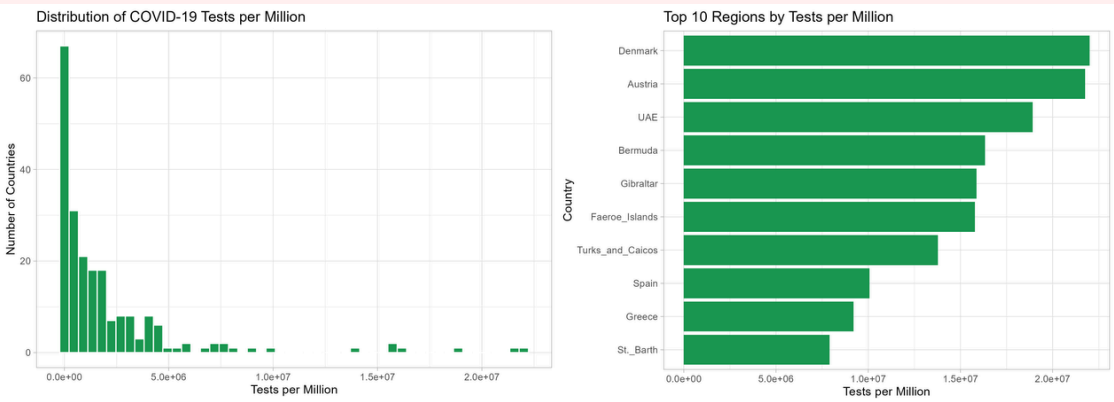
Fig. 3: Cases/M distribution and top 10 countries.



Not normally distributed, with low values being the more frequent.

Top 10 regions are all east-European countries, despite Peru. All of them are mid-sized countries.

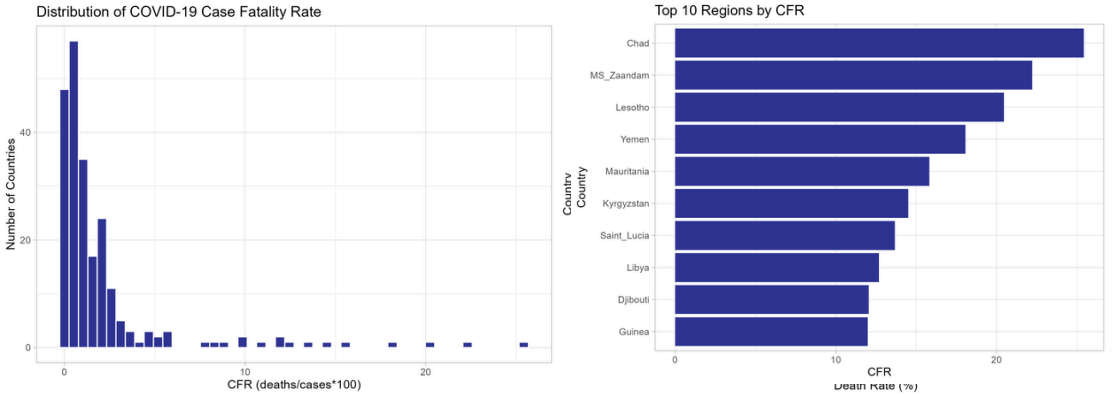
Fig. X: Deaths/M distribution and top 10 countries.



Not normally distributed, with low values being the more frequent.

Top 10 regions are all European countries or overseas territories. and there is a high diversity of population size.

Fig. 4: Tests/M distribution and top 10 countries.



Not normally distributed, with low values being the more frequent.

Top 10 regions are African and Asian countries + the famous case of the Zaandam cruise ship.

Fig. 5: CFR distribution and top 10 countries.

CORRELATIONS

The rate variables were tested with each other to search for interesting correlations using Spearman Correlation Test. NA values were discarded. The hypotheses set was:

- H_0: No correlation between variables.
- H_1: Correlation between variables (considering a two-sided test, as we are interested in both positive and negative correlations).

Cases/M VS Deaths/M (Fig. 6)

Significant moderate positive correlation.

- **rho:** 0.6493
- **p-value:** $4.3973*10^{-28}$

This was the expected trend, as more cases rate in a country is expected to generate a higher Deaths/M. This relation could fluctuate due to many factors, such as public healthcare, which makes the correlation be lower than if every country was equal.

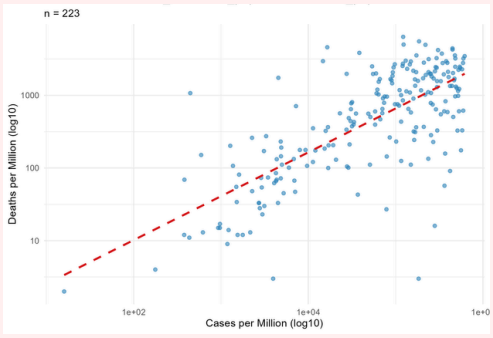


Fig. 6: Relation between Cases/M and Deaths/M.

Tests/M VS Cases/M (Fig. 7)

Significant high positive correlation.

- **rho:** 0.8139
- **p-value:** ~0

This high correlation is aligned with the obvious double fact that more testing implies detecting more cases and that countries with more cases tend to adopt testing politics to avoid further propagation of the virus.

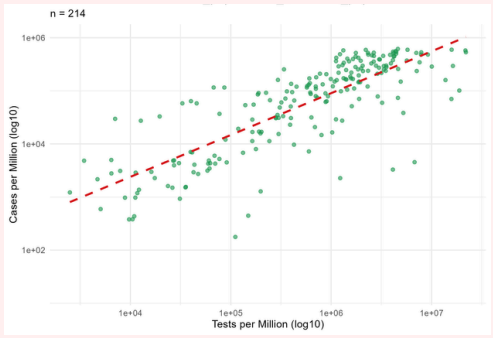


Fig. 7: Relation between Tests/M and Cases/M.

Population VS Cases/M (Fig. 8)

Significant moderate negative correlation.

- **rho:** -0.3909
- **p-value:** $1.3227*10^{-9}$.

The correlation is barely moderate because it is extremely multifactorial and there are contradictory arguments. For example, larger populations could be related to advanced society and better hygiene but also more interconnected mass groups.

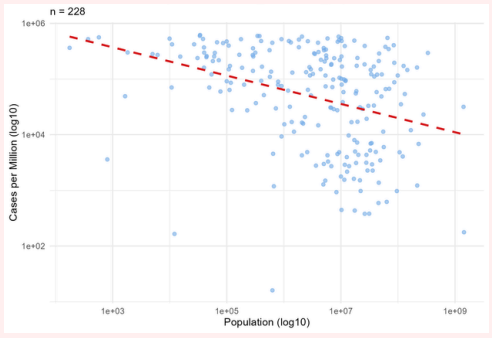


Fig. 8: Relation between Population size and Cases/M.

Did COVID-19 testing cause an increase of the deaths? Tests/M vs Deaths/M (Fig. 9)

Significant moderate positive correlation.

- **rho:** 0.5942
- **p-value:** $1.0226*10^{-21}$

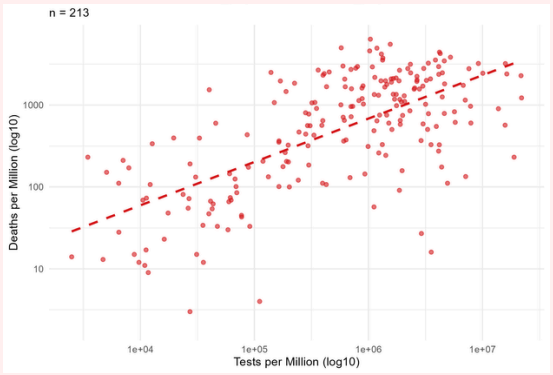


Fig. 9: Relation between Tests/M and Deaths/M.

However, this is a clear example that correlation does not imply causation. In these cases, there often is a hidden variable that explains this weird correlation. It is clear that the COVID-19 testing is not an invasive intervention that caused thousands of deaths, but an easy way to track the pandemic development and help improving population's security. Thanks to the previous correlations calculations, we can see that there are positive correlations between Tests/M and Cases/M, and Cases/M and Deaths/M. Then, we could have found the hidden third variable: Cases/M. More testing is correlated with more COVID-19 cases rate, which is correlated with a higher Deaths/M.

To sum up, although testing rate is positively correlated with Deaths/M, it is not a causal factor of it.

SUMMARY OF CONCLUSIONS

- The analysed variables of the dataset show irregular distributions with shallow trends.
- Cases/M was found to be correlated with Deaths/M and Tests/M. This is used as a hypotheses to explain the found correlation between Tests/M and Deaths/M.
- A shallow negative correlation was found between population size and Cases/M.