Advanced Statistics for Genomics

Project (Final Report)

Jawad Chowdhury, 801135477

Topic: Statistical analysis of the influence of social, economic, and environmental factors on COVID-19 across different countries.

1 Introduction

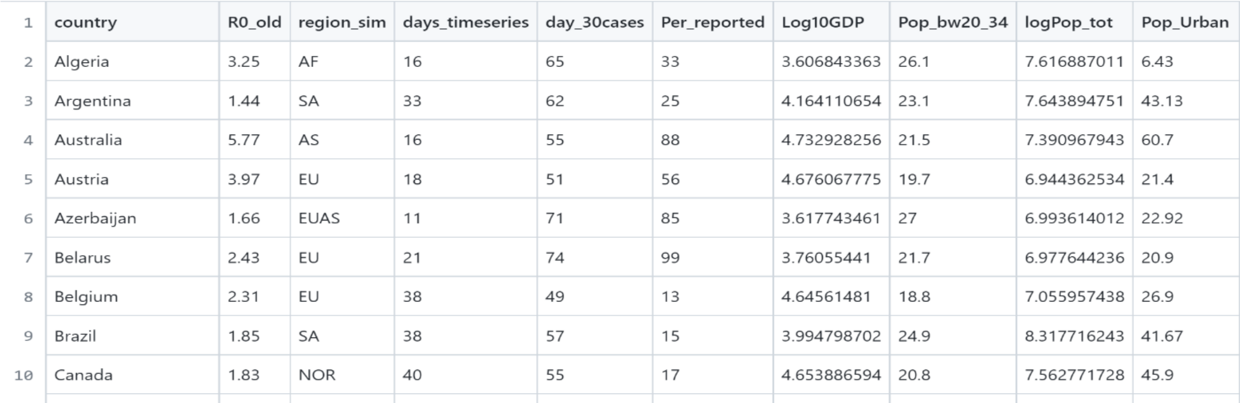
In recent times, there has been numerous research ongoing based on the COVID-19 pandemic. Existing research is mostly devoted to anticipate the nature of the pandemic and explaining its different characteristics. These researches consider different social, economic, and environmental factors and try to figure out the relation between the dependent variables such as the number of infected people or the basic reproduction number with those influencing factors. In this project, I aim to analyze how these influencing factors are related among themselves, how they impact the dependent variable, and how these trends differ between different countries or regions from a statistical perspective.

1.1 Baseline Work/Paper: For the statistical analysis, I am considering the following work or paper as the baseline work: “Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries” [1]. The main objectives of the study in Ref. [1] are-

i) to assess whether the basic reproduction number (R_0) of COVID-19 is different across countries/regions,

ii) and to analyze the demographic, social, and environmental factors other than interventions that characterize the vulnerability to the spread of the pandemic.

Since different countries or regions have its own characteristics that expose them to the vulnerability of COVID-19 in different extent, any study that aims to measure the effectiveness of interventions across different locals should consider these statistical differences.



1	country	R0_old	region_sim	days_timeseries	day_30cases	Per_reported	Log10GDP	Pop_bw20_34	logPop_tot	Pop_Urban
2	Algeria	3.25	AF	16	65	33	3.606843363	26.1	7.616887011	6.43
3	Argentina	1.44	SA	33	62	25	4.164110654	23.1	7.643894751	43.13
4	Australia	5.77	AS	16	55	88	4.732928256	21.5	7.390967943	60.7
5	Austria	3.97	EU	18	51	56	4.676067775	19.7	6.944362534	21.4
6	Azerbaijan	1.66	EUAS	11	71	85	3.617743461	27	6.993614012	22.92
7	Belarus	2.43	EU	21	74	99	3.76055441	21.7	6.977644236	20.9
8	Belgium	2.31	EU	38	49	13	4.64561481	18.8	7.055957438	26.9
9	Brazil	1.85	SA	38	57	15	3.994798702	24.9	8.317716243	41.67
10	Canada	1.83	NOR	40	55	17	4.653886594	20.8	7.562771728	45.9

Fig 1: Partial snapshot of considered dataset [3]

1.2 Dataset: The baseline work presents their data and implementation in Ref. [2]. In their study, they have considered 58 different samples (for this study it's 58 different countries). On the other hand, they have analyzed the data of these 58 countries on 7 different categories such as demographics, disease, economics, environmental, habitat, health, and social. The dataset has 31 different factors from which the study has chosen 16 covariates based on least amount of correlations. A glimpse of the considered dataset [3] is given in **Fig 1**.

2 Methodologies and Experimental Results

2.1 Un-correlated Covariate Selection (by Correlation Matrix & PCA): The initial step I have considered here in my study is to figure out what are the covariates that form the feature set with least amount of correlations. The dataset from the study [1] has 31 features where 29 of them are numeric and 2 are categorical. The correlation matrix with these 29 features are given in **Fig 2.**

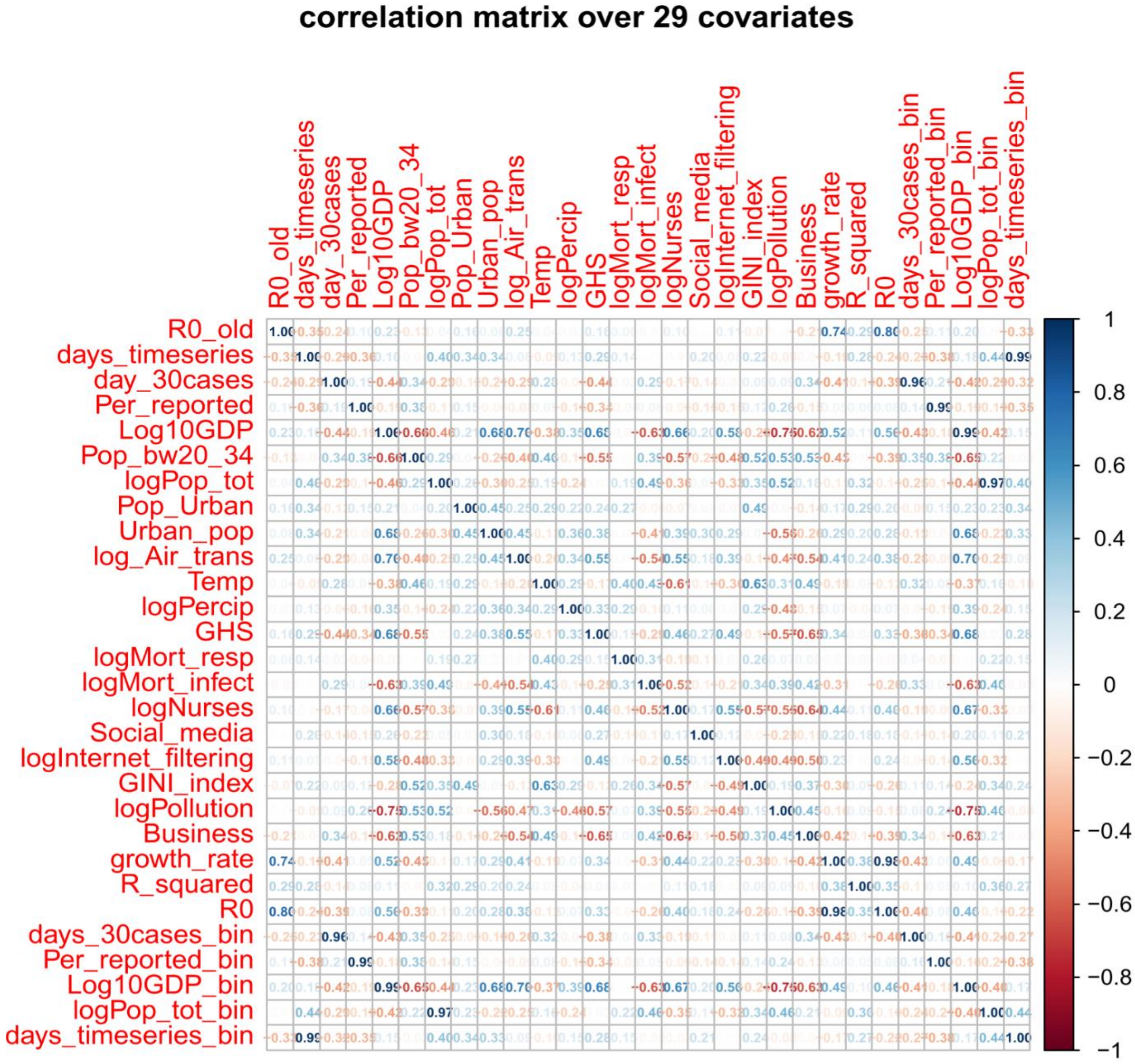
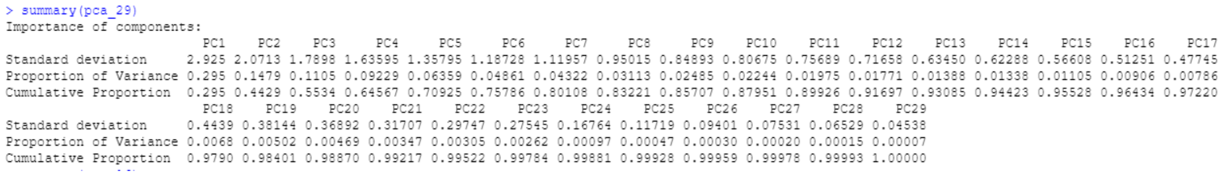


Fig 2: Correlation matrix over 29 covariates (before selection)

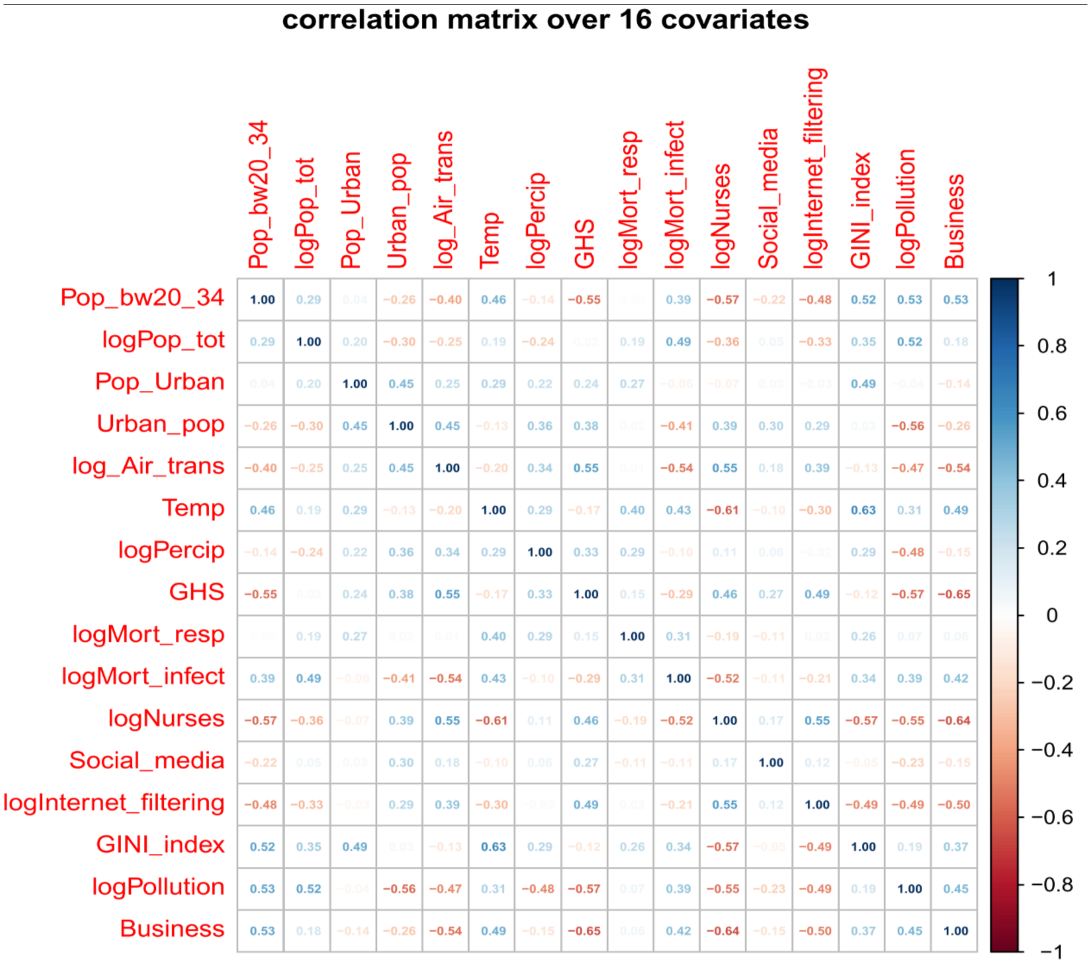
I have done a **Principle Component Analysis (PCA)** with these 29 covariates to see how much variance can be covered by these covariates individually and as cumulative.



```
> summary(pca_29)
Importance of components:
                         PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     2.925 2.0713 1.7898 1.63595 1.35795 1.18728 1.11957 0.95015 0.84893 0.80675 0.75689 0.71658 0.63450 0.62288 0.56608 0.51251 0.47745
Proportion of Variance 0.295 0.1479 0.1105 0.09229 0.06359 0.04861 0.04322 0.03113 0.02485 0.02244 0.01975 0.01771 0.01388 0.01338 0.01105 0.00906 0.00786
Cumulative Proportion  0.295 0.4429 0.5534 0.64567 0.70925 0.75786 0.80108 0.83221 0.85707 0.87951 0.89926 0.91697 0.93085 0.94423 0.95528 0.96434 0.97220
                         PC18    PC19    PC20    PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28    PC29
Standard deviation     0.4439 0.38144 0.36892 0.31707 0.29747 0.27545 0.16764 0.11719 0.09401 0.07531 0.06529 0.04538
Proportion of Variance 0.0068 0.00502 0.00469 0.00347 0.00305 0.00262 0.00097 0.00047 0.00030 0.00020 0.00015 0.00007
Cumulative Proportion  0.9790 0.98401 0.98870 0.99217 0.99522 0.99784 0.99881 0.99928 0.99959 0.99978 0.99993 1.00000
```

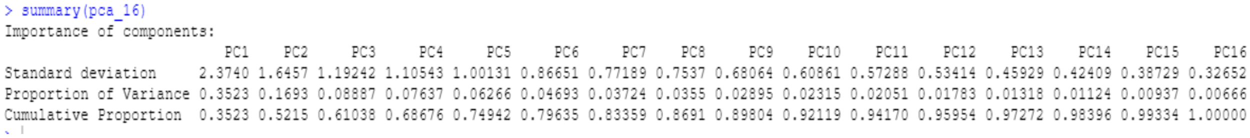
Fig 3: Principle Component Analysis with 29 covariates (before selection)

The baseline study [1] considered covariates that are diverse, specific, and do not covary much with each other. For example, in the correlation matrix with 29 covariates, I have found that Gross Domestic Product (Log10GDP) covaries with many others such as Pollution (logPollution), Nurses (logNurses), GHS (GHS), Air Transport (log_Air_trans), Urbanization (Urban_pop) and etc. The baseline study [1] excludes these covariates that are highly related with other covariates. The purpose of this filtering process is to reduce the collinearity in the final covariate set that lead us to 16 covariates with least amount of collinearity. The correlation matrix and the PCA with these final 16 covariates are given in **Fig 4** and **Fig 5.**



	Pop_bw20_34	logPop_tot	Pop_Urban	Urban_pop	log_Air_trans	Temp	logPercip	GHS	logMort_resp	logMort_infect	logNurses	Social_media	logInternet_filtering	GINI_index	logPollution	Business
Pop_bw20_34	1.00	0.29	0.04	−0.26	−0.40	0.46	−0.14	−0.55	0.00	0.39	−0.57	−0.22	−0.48	0.52	0.53	0.53
logPop_tot	0.29	1.00	0.20	−0.30	−0.25	0.19	−0.24	0.02	0.19	0.49	−0.36	0.05	−0.33	0.35	0.52	0.18
Pop_Urban	0.04	0.20	1.00	0.45	0.25	0.29	0.22	0.24	0.27	−0.06	−0.07	0.03	−0.03	0.49	−0.04	−0.14
Urban_pop	−0.26	−0.30	0.45	1.00	0.45	−0.13	0.36	0.38	0.02	−0.41	0.39	0.30	0.29	0.03	−0.56	−0.26
log_Air_trans	−0.40	−0.25	0.25	0.45	1.00	−0.20	0.34	0.55	0.01	−0.54	0.55	0.18	0.39	−0.13	−0.47	−0.54
Temp	0.46	0.19	0.29	−0.13	−0.20	1.00	0.29	−0.17	0.40	0.43	−0.61	−0.10	−0.30	0.63	0.31	0.49
logPercip	−0.14	−0.24	0.22	0.36	0.34	0.29	1.00	0.33	0.29	−0.10	0.11	0.06	−0.02	0.29	−0.48	−0.15
GHS	−0.55	0.02	0.24	0.38	0.55	−0.17	0.33	1.00	0.15	−0.29	0.46	0.27	0.49	−0.12	−0.57	−0.65
logMort_resp	0.00	0.19	0.27	0.02	0.01	0.40	0.29	0.15	1.00	0.31	−0.19	−0.11	0.02	0.26	0.07	0.06
logMort_infect	0.39	0.49	−0.06	−0.41	−0.54	0.43	−0.10	−0.29	0.31	1.00	−0.52	−0.11	−0.21	0.34	0.39	0.42
logNurses	−0.57	−0.36	−0.07	0.39	0.55	−0.61	0.11	0.46	−0.19	−0.52	1.00	0.17	0.55	−0.57	−0.55	−0.64
Social_media	−0.22	0.05	0.03	0.30	0.18	−0.10	0.06	0.27	−0.11	−0.11	0.17	1.00	0.12	−0.05	−0.23	−0.15
logInternet_filtering	−0.48	−0.33	−0.03	0.29	0.39	−0.30	−0.02	0.49	0.02	−0.21	0.55	0.12	1.00	−0.49	−0.49	−0.50
GINI_index	0.52	0.35	0.49	0.03	−0.13	0.63	0.29	−0.12	0.26	0.34	−0.57	−0.05	−0.49	1.00	0.19	0.37
logPollution	0.53	0.52	−0.04	−0.56	−0.47	0.31	−0.48	−0.57	0.07	0.39	−0.55	−0.23	−0.49	0.19	1.00	0.45
Business	0.53	0.18	−0.14	−0.26	−0.54	0.49	−0.15	−0.65	0.06	0.42	−0.64	−0.15	−0.50	0.37	0.45	1.00

Fig 4: Correlation matrix over 16 covariates (after selection)



```
> summary(pca_16)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8     PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16
Standard deviation     2.3740 1.6457 1.19242 1.10543 1.00131 0.86651 0.77189 0.7537 0.68064 0.60861 0.57288 0.53414 0.45929 0.42409 0.38729 0.32652
Proportion of Variance 0.3523 0.1693 0.08887 0.07637 0.06266 0.04693 0.03724 0.0355 0.02895 0.02315 0.02051 0.01783 0.01318 0.01124 0.00937 0.00666
Cumulative Proportion  0.3523 0.5215 0.61038 0.68676 0.74942 0.79635 0.83359 0.8691 0.89804 0.92119 0.94170 0.95954 0.97272 0.98396 0.99334 1.00000
```

Fig 5: Principle Component Analysis with 16 covariates (after selection)

Here after filtering out the highly correlated covariates, the maximum correlation coefficient I have found is between Business and GHS with a value of -0.65.

2.2 Pairwise Analyses of Selected Covariates (Per Category)-

After the initial filtering process, we now have 7 categories with 16 covariates with least collinearity. They are as in **Table 1.**

Categories	**Selected Covariates**
Demographics	Youth, Total Population
Disease	Mortality Respiratory, Mortality by
Economic	GINI index, Business
Environmental	Temperature, Precipitation, Pollution
Habitat	City, Urbanization
Health	GHS, Nurses
Social	Social Media, Internet Filter, Air Transport

Table 1: Covariates (Selected) per Categories

In **Fig 6,** I have tried to analyze how the covariates are related to each other within these categories by a two-dimensional plotting for each pair.

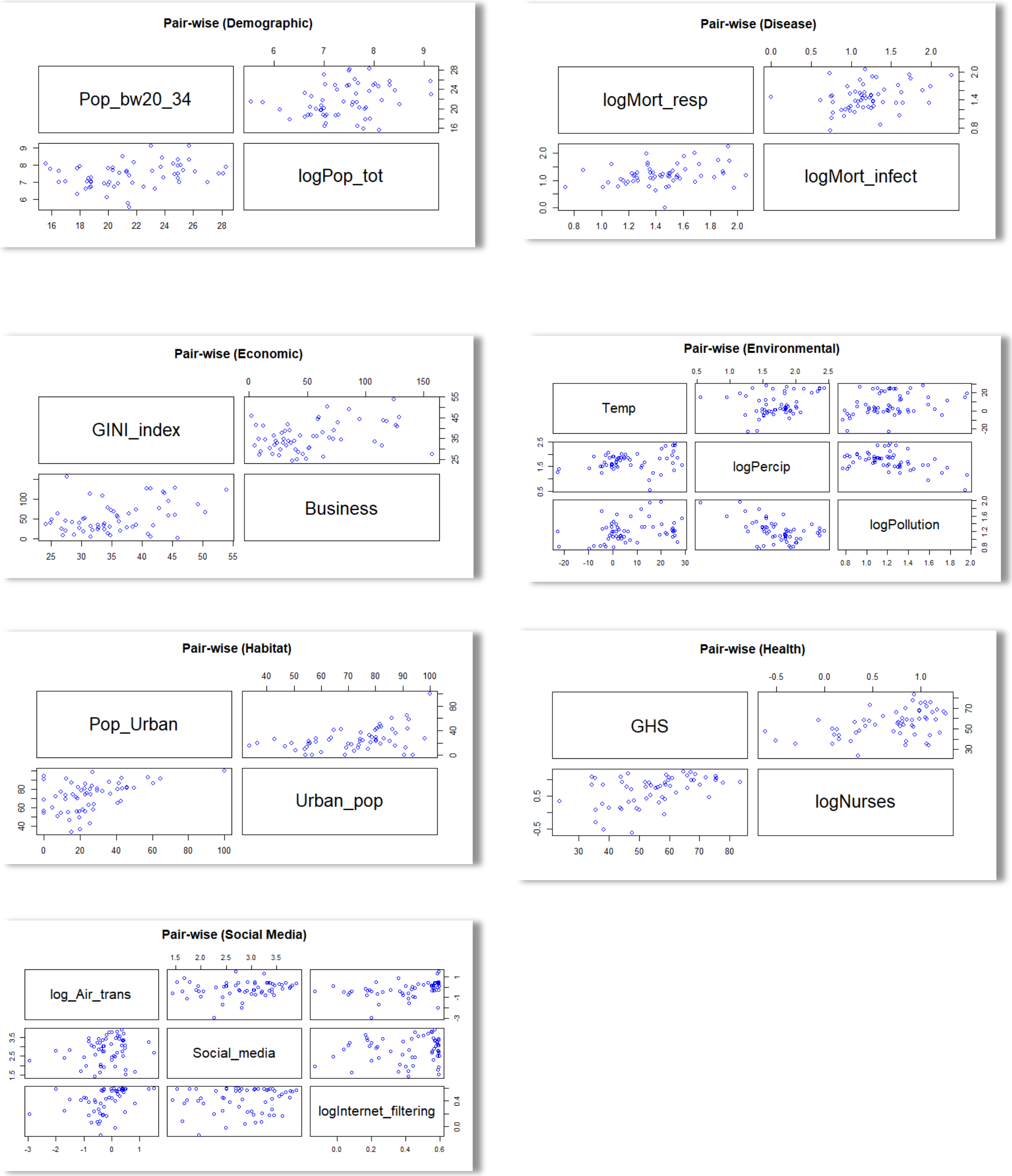


Fig 6: Pair-wise analyses of selected covariates within each category

2.3 Data Plotting: Principle Component (1) vs Principle Component (2)

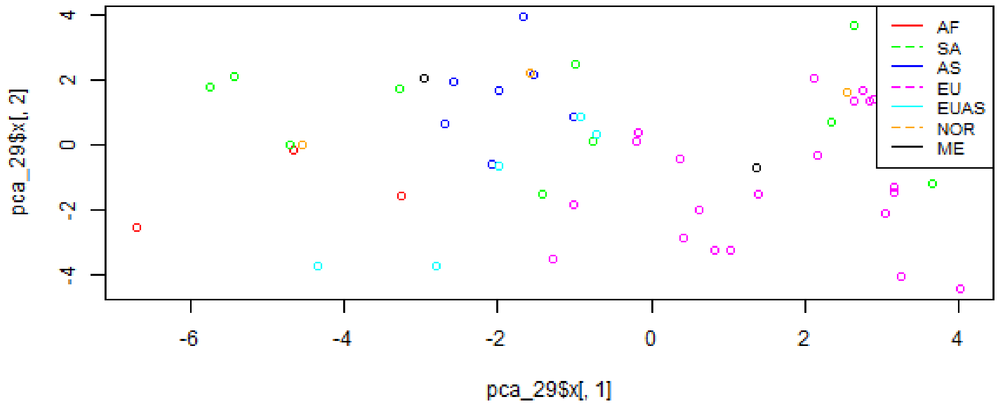


Fig 7: principle component 1 vs principle component 2 (using 29 covariates)

Next, I have tried to plot the 58 samples in a two-dimensional space where the x-axis refers to the first principle component and the y-axis to the second. From my prior analyses, I have two different sets of PCA components where in one case the components were being formed using all 29 covariates and in the other one, the PCA components are being formed using only the 16 (selected) covariates. The results from the plotting using first and second PCA components are given in **Fig 7** (using 29 covariates) and in **Fig 8** (using selected 16 covariates).

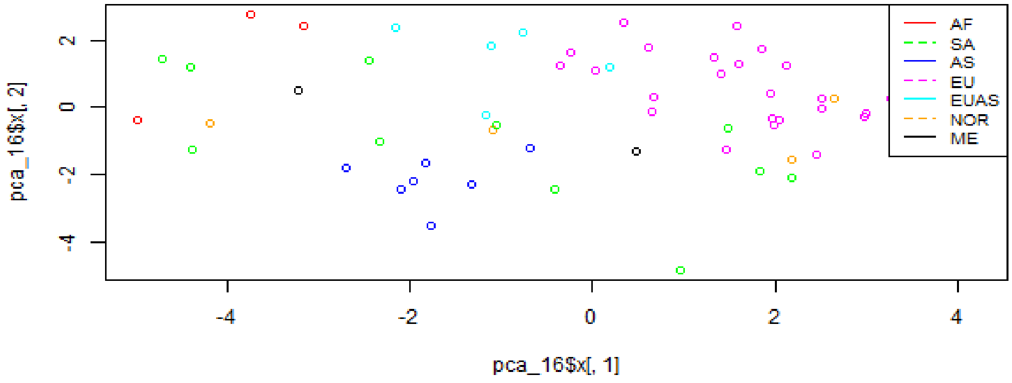
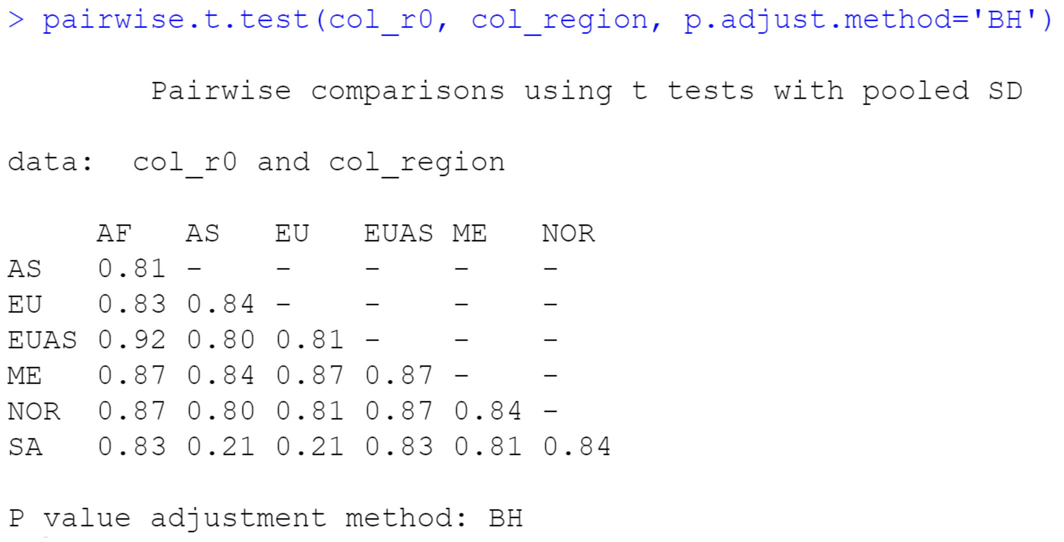


Fig 8: principle component 1 vs principle component 2 (using 16 covariates)

Here, the short-codes in the legends refers to each region such as- **AS**: Asia-Australia, **AF**: Africa, **EUAS**: Eurasia, **EU**: Europe, **ME**: Middle East, **NOR**: North America, and **SA**: South America.

2.4 Pairwise t-test (by Region)



```
> pairwise.t.test(col_r0, col_region, p.adjust.method='BH')

        Pairwise comparisons using t tests with pooled SD 

data:  col_r0 and col_region 

     AF   AS   EU   EUAS ME   NOR 
AS   0.81 -    -    -    -    -   
EU   0.83 0.84 -    -    -    -   
EUAS 0.92 0.80 0.81 -    -    -   
ME   0.87 0.84 0.87 0.87 -    -   
NOR  0.87 0.80 0.81 0.87 0.84 -   
SA   0.83 0.21 0.21 0.83 0.81 0.84

P value adjustment method: BH
```

Fig 9: Pairwise t-test by the Regions (**AS**: Asia-Australia, **AF**: Africa, **EUAS**: Eurasia, **EU**: Europe, **ME**: Middle East, **NOR**: North America, and **SA**: South America)

In this section, I tried to have some comparisons between different regions based on the reproduction number, R0 of the countries they include. **Fig 9** shows a chart with pairwise t-test values for the regions using ‘BH’ or ‘False Discovery Rate’ adjusted method.

2.5 Basic Reproduction Number (R0) vs each Covariates (16) (Simulating Analyses from Baseline Paper in Ref. [1] by Using Linear Model Instead of Nonlinear GAM Used in the Paper)

In this section, I tried to simulate one of the statistical analyses from the baseline paper [1] I considered for this project, even though there are some discrepancies in the experimental setup between the study in Ref. [1] and what I have used in my experiments.

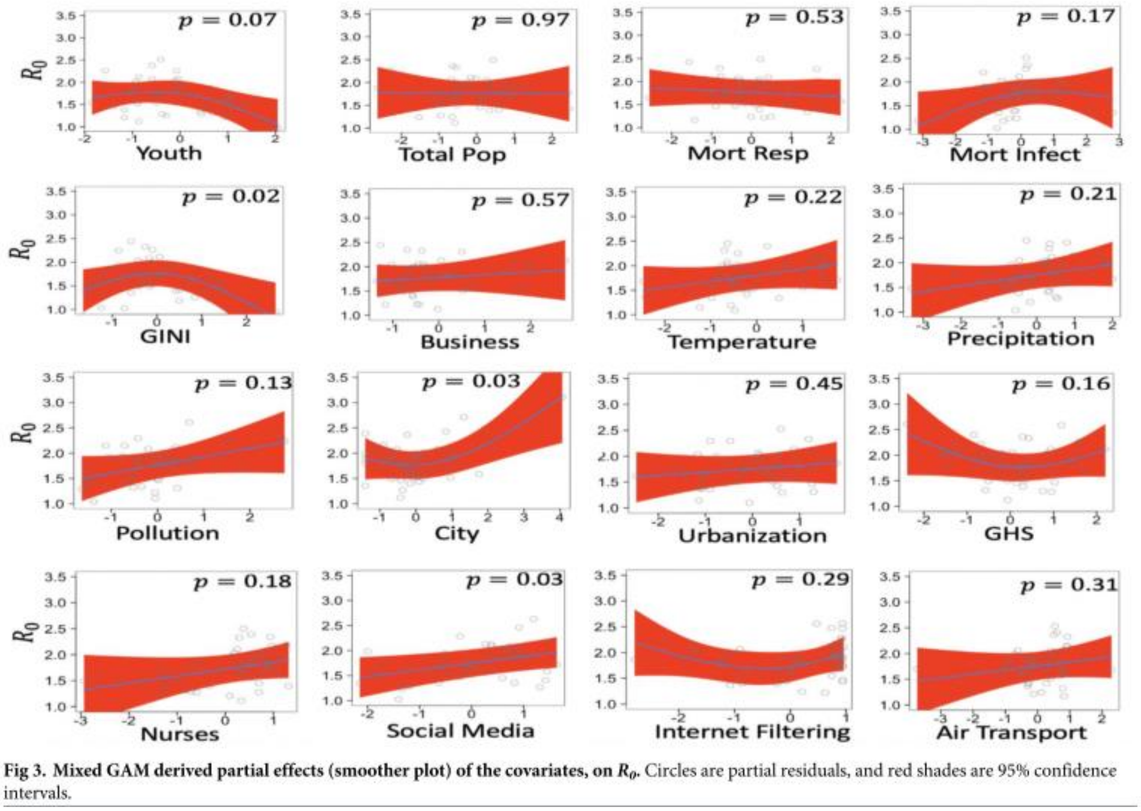


Fig 3. Mixed GAM derived partial effects (smoother plot) of the covariates, on R_0. Circles are partial residuals, and red shades are 95% confidence intervals.

Fig 10: Results from Baseline Paper [1] using Nonlinear GAM (Generalized Additive Model)- Basic Reproduction Number (R0) vs Covariates (16)

The analysis basically tries to regress the target variable which is in our case the reproduction number, R0 from each of the 16 selected covariates. Using the p-values this analysis tries to have inference on the model fit to the data. This to be noted here that the study in Ref. [1] used a Generalized Additive Model (GAM) to perform the regression. The general formula of the GAM is:

$$g(u_i) = \beta + \sum_{j=1}^{n} f_j(X_i) + \varepsilon_i$$

Where $g(u_i)$ is the monotonous link function relating the independent variable to the given covariate, β is the parametric component, the smoothing function $f_j(X_i)$ can be non-linear in nature.

However, in my experimental setup, I have used linear regression model to regress on the reproduction number, R0 from each covariate and the results are given in **Fig 11**.

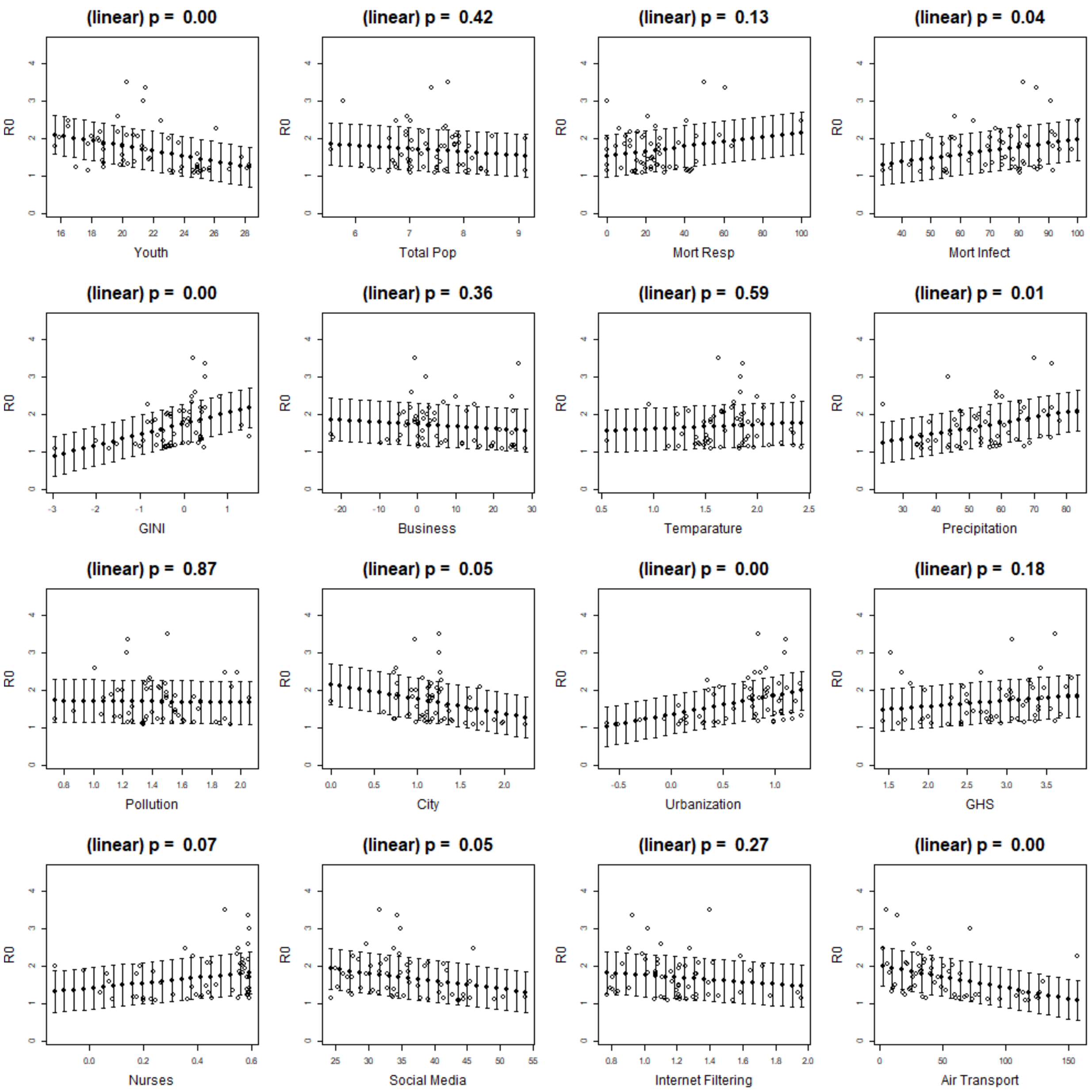


Fig 11: Results from this Study using Linear Regression Model- Basic Reproduction Number (R0) vs Covariates (16)

3 Discussion, Limitations & Concluding Remarks

Based on the observations from all these statistical analyses, it is clear that in the initial set, there were several features that highly covaried with each other. As a result, any statistical inference we would have drawn with all those 29 covariates would not make much sense or may lead to misinterpretation. The correlation matrix and PCA have given a good glimpse on which covariates should we be keeping and which we can discard.

Later on, in **Fig 6**, we have seen a pairwise plotting for the 16 covariates grouped by the regions. The trend in those pairwise plots look much flatter in most cases which is an indication that the remaining covariates has less collinearity with respect to each other.

In **Fig 7** and **Fig 8,** it's quite explicit that the principle component 1 and 2 do a pretty job on grouping the region-based samples. However, it seems like we see a better clustering of the samples in **Fig 8** than in **Fig 7.** In **Fig 8,** other than 'SA: South America' region, every other samples seem to be clustered based on their regions where as in **Fig 7,** we see mismatches in couple of cases (e.g. 3 different samples (green, red, and orange) are pretty close to each other than their own region-wise samples). The reason we see a better clustering with **Fig 8**, might be due to the fact that we use a selected set of 16 covariates in the corresponding PCA analysis and as a result, there is least amount of collinearity among these covariates, I believe the primary selection of the covariate set leads to a better component analysis and the distinction among the clusters are more explicit in **Fig 8**.

In **Fig 9**, we see the pairwise t-test based on the regions. Since we have used the reproduction number, R0 to regress on and the regions as the categories, from the values in the chart, it seems like most of these pairwise regions, the trend for the reproduction number, R0 are somewhat similar. As for 7 regions, we are considering 7C2 = 21 t-tests. Therefore, we have used an adjusted p-values based on the BH or False Discovery Rate. However, out of those 21 pairs, the most significant cases we have found are with pair SA-AS and SA-EU with a p-value of 0.21.

In **Fig 10**, we see the results from the baseline paper [1] that I have considered for this project whereas in **Fig 11**, the results are from my own study with a similar experimental setup. In these studies, we try to look for the trend of the model by setting reproduction number, R0 as the target for each of those 16 selected covariates. Considering the results from the baseline paper [1] in **Fig 10**, we can see that there are 4 covariates that are highly correlated with reproduction number, R0 where the p-values are below 0.10. On the other hand, based on my experimental results in **Fig 11,** there are 9 covariates that are highly correlated with the reproduction number, R0 where the p-values are below 0.10. This set of 9 covariates includes the 4 covariates that appeared to be significant in **Fig 10** along with 5 additional ones. It is to be noted here that the study in the paper uses nonlinear model (GAM) to fit the data whereas in my case I have considered basic linear regression model.

Considering all these observations on the COVID-19 dataset I have considered for this project, it seems like we can learn a lot about the influence of all these social, economic, environmental factors, about how they are related to each other, and how impactful they are on the target or the reproduction number, R0 just by using the linear models. However, we also noticed some limitations specifically when we compare **Fig 10** and **Fig 11**, it seems like we can learn more information or about the trend with nonlinear models. Although, using linear models gives us the benefit of having straightforward interpretability. Finally, based on the results from the paper [1] and the statistical analyses with this project, we can say that these influencing factors has their own characteristics based on different regions, and any study with similar goals or aims should consider these factors before drawing any inference or come to any conclusion.

References

[1] Kong, J. D., Tekwa, E. W., & Gignoux-Wolfsohn, S. A. (2021). Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries. *PloS one*, *16*(6), e0252373

[2] https://github.com/Jdkong/COVID-19

[3] https://github.com/Jdkong/COVID-19/blob/main/Data/variables.csv