Advanced Statistics for Genomics

Project (Preliminary Version)

Jawad Chowdhury, 801135477

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

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.

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,
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.

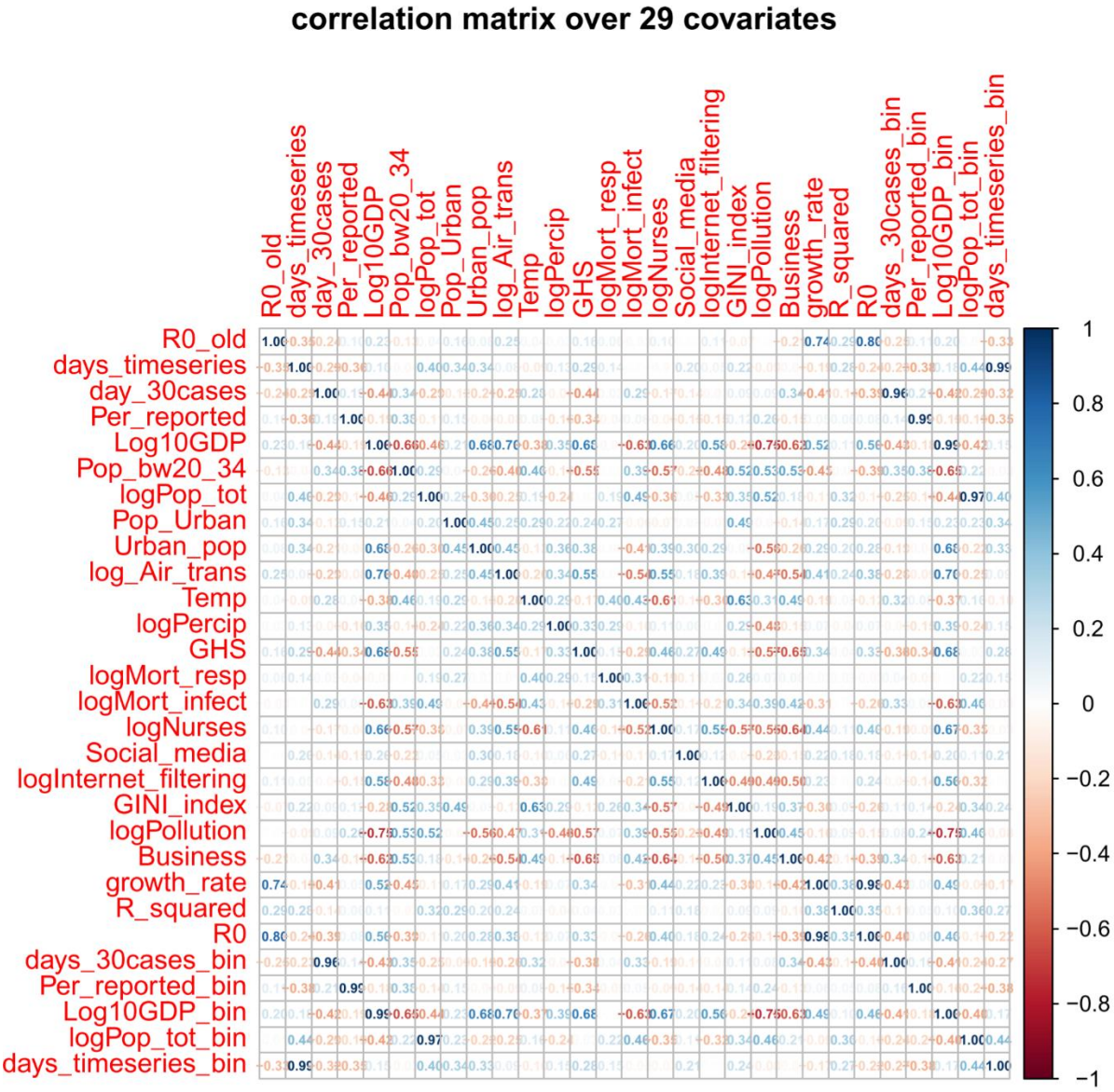
Dataset: The baseline work presents their data and implementation in Ref. [2]. In their study, they have considered 58 different samples (for this study it's 58 different countries). On the other hand, they have analyzed the data of these 58 countries on 7 different categories such as demographics, disease,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

	country	R0_old	region_sim	days_timeseries	day_30cases	Per_reported	Log10GDP	Pop_bw20_34	logPop_tot	Pop_Urban	Urban_pop	log_Air_trans	Temp	logPercip	GHS	lo
2	Algeria	3.25	AF	16	65	33	3.606843363	26.1	7.616887011	6.43	72.05	-0.82	14.94	0.954242509	23.6	1.3
3	Argentina	1.44	SA	33	62	25	4.164110654	23.1	7.643894751	43.13	91.75	-0.42	19.41	1.827304641	58.6	1.8
4	Australia	5.77	AS	16	55	88	4.732928256	21.5	7.390967943	60.7	85.9	0.48	26.79	1.862012051	75.5	1.2
5	Austria	3.97	EU	18	51	56	4.676067775	19.7	6.944362534	21.4	58.1	0.26	-1.36	1.84416641	58.5	1.0
6	Azerbaijan	1.66	EUAS	11	71	85	3.617743461	27	6.993614012	22.92	55.34	-0.62	1.86	1.489677292	34.2	1.2
7	Belarus	2.43	EU	21	74	99	3.76055441	21.7	6.977644236	20.9	78.13	-0.59	-4.33	1.512150537	35.3	0.8
8	Belgium	2.31	EU	38	49	13	4.64561481	18.8	7.055957438	26.9	98	0.08	3.26	1.84167225	61	1.7
9	Brazil	1.85	SA	38	57	15	3.994798702	24.9	8.317716243	41.67	86.31	-0.34	25.58	2.355566448	59.7	1.6
10	Canada	1.83	NOR	40	55	17	4.653886594	20.8	7.562771728	45.9	81.4	0.4	-22.2	1.410608543	75.3	1.2
11	Chile	1.93	SA	26	58	86	4.177172753	23.9	7.266477319	25.9	87.5	-0.03	11.93	1.526179533	59.2	1.1

Preliminary Work:

A. Covariate Selection (Un-correlated)-

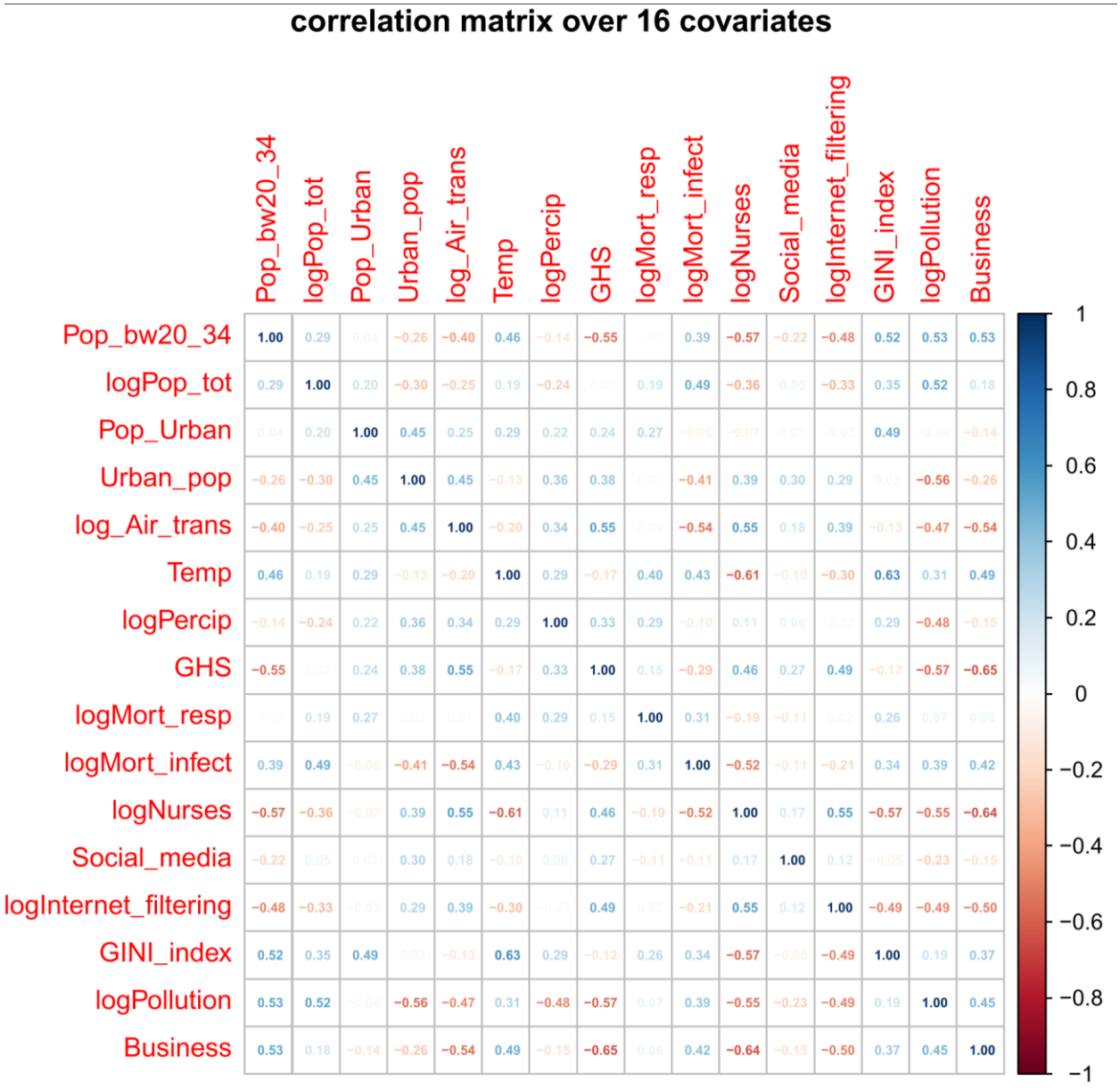
The initial step I have considered here is to figure out what are the covariates that form the feature set with least amount of correlations. The dataset from the study [1] has 31 features where 29 of them are numeric and 2 are categorical. The correlation matrix with these 29 features are given below:



I have also done a **Principle Component Analysis (PCA)** with these 29 features 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
```

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 below:




```
> 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
```

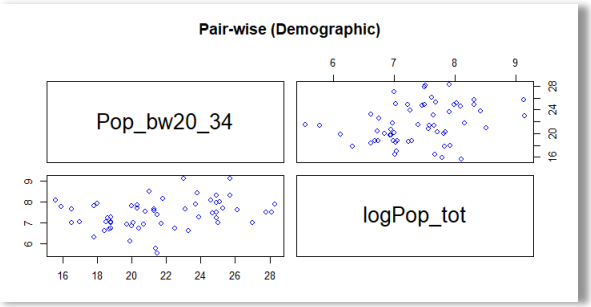
Here after filtering out the highly correlated covariates, the maximum correlation coefficient I have found is between Business and GHS with a value of -0.65.

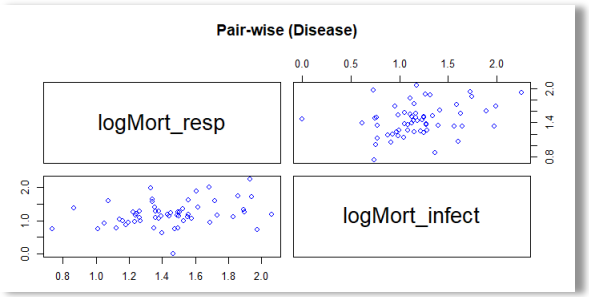
B. Pairwise Analyses of Selected Covariates (Per Category)-

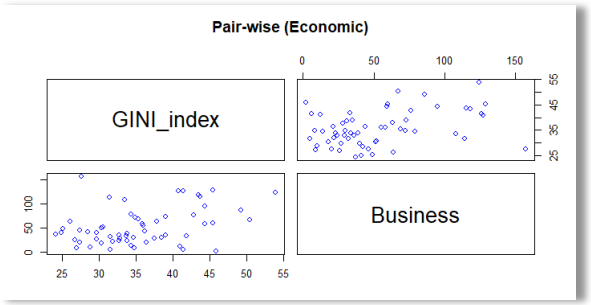
After the initial filtering process, we now have 7 categories with 16 covariates with least collinearity. They are as follows:

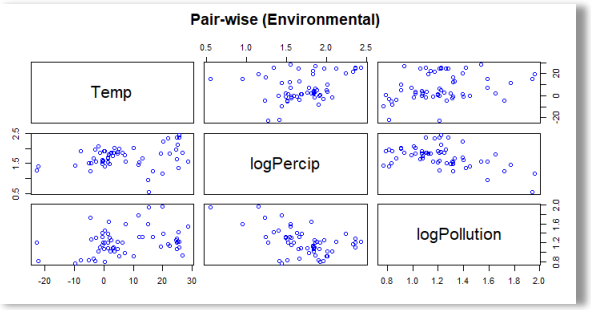
Categories	Selected Covariates
Demographics	Youth, Total Population
Disease	Mortality Respiratory, Mortality by
Economic	GINI index, Business
Environmental	Temperature, Precipitation, Pollution
Habitat	City, Urbanization
Health	GHS, Nurses
Social	Social Media, Internet Filter, Air Transport

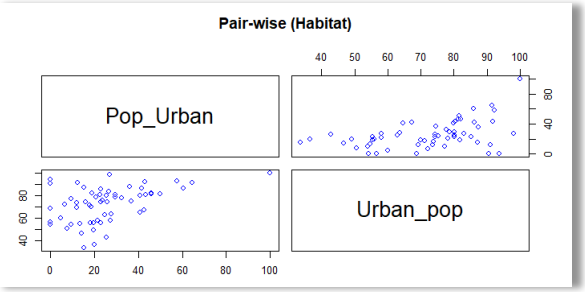
Below I have tried to analyze how the covariates are related to each other within the categories by a 2 dimensional plotting for each pairs:

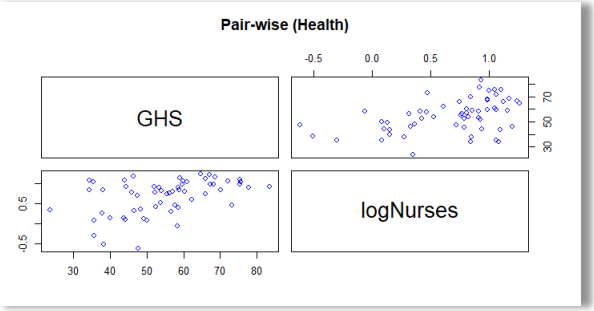


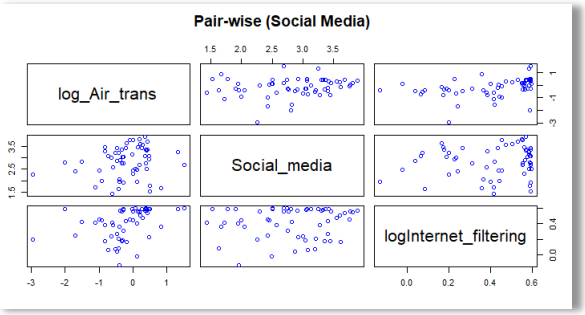












C. **Modeling (linear regression/ANOVA)-** To be done in the further steps of the project.

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