Executive Summary*:*

This paper uses a cross-sectional regression methodology to analyse the demographic characteristics of 179 countries and examine their significance in explaining the number of confirmed COVID-19 cases and deaths per capita.

The dataset used consists of 59 country characteristics spanning areas including economics, the environment and societal indicators for the 13th April 2020. The dataset was constructed using information collated from three online sources.

Three unique ordinary least squares (OLS) regression approaches were used to determine which predictor variables have a statistically significant relationship with the dependent variables of interest. I find that 'Total Greenhouse Gas Emissions' is a significant factor in explaining both the total number of confirmed cases and deaths per capita in two of the three approaches used. Urban Population as a % of total population and the proportion of people using safely managed drinking water services are also good predictors of total cases per capita.

These results nicely reflect the conclusions of other recent studies by (Martelletti and Martelletti, 2020) and (Wu, C Nethery, Braun and Dominici, 2020) which indicate that air pollution has a significant impact on the COVID-19 death rate. This outlines that further analysis of this relationship is required.

1. Introduction

The outbreak of the novel Coronavirus, COVID-19, in Wuhan, China has caused damage, destruction and death on a scale not witnessed from a global disease since the Spanish Flu epidemic back in 1918. As of 19/05/2020, over 315,000 people have died as a direct result of the COVID-19 virus with over 4.7 million confirmed cases distributed across 216 countries and territories. (Roser, Ritchie, Ortiz-Ospina and Hasell, 2020).

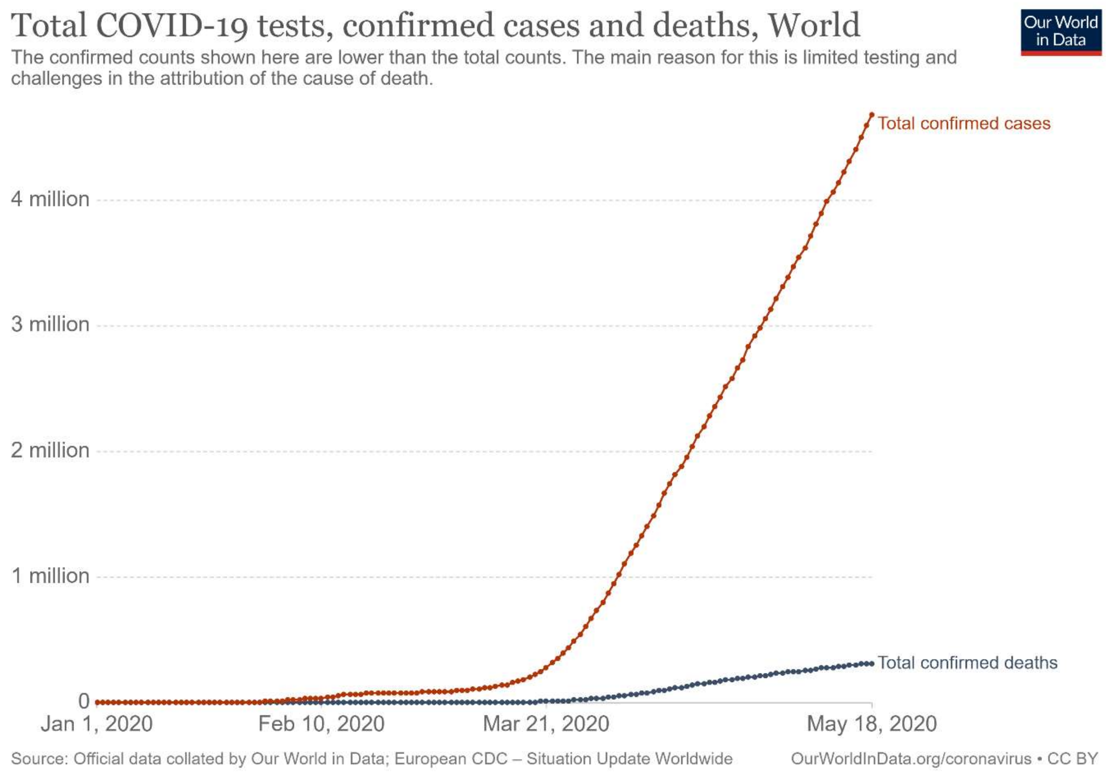

Figure 1: Total Confirmed COVID-19 Cases and Deaths. Source: https://ourworldindata.org/

The World Health Organisation (WHO) declared a worldwide public health emergency on 30th January 2020 and recommended that countries initiate 'lockdown' procedures to ease the spread of the virus and reduce pressure on health systems around the globe. The result of these measures have been catastrophic to the global economy with indices such as the Dow Jones and FTSE experiencing their largest quarterly decreases since 1987 (Jones, Palumbo and Brown, 2020). The economic outlook is bleak, with a looming global recession almost inevitable (Fernandes, 2020). Considering the drastic impact on humanity the UN has declared the disease as a social, human and economic crisis (Sannigrahi, Pillaa, Basua and Sarkar, 2020).

2. Research Purpose

The purpose of this research paper is to examine the relationship between certain demographic variables and the total number of cases and deaths per capita across countries using a cross-sectional regression analysis. The goal is to identify country characteristics which are significant in explaining the number of confirmed COVID-19 cases and deaths. The identification of these variables can aid in understanding which countries are at the greatest risk based on their underlying characteristics. This will enable a more efficient allocation of resources thereby helping to alleviate the impact of the coronavirus globally by focussing solely on characteristics which have a causal relationship with cases and deaths per capita.

3. Research Design

Using several cross-sectional multiple-linear regression models, I analyse the effect of numerous country characteristics on two dependent variables namely, cases per capita and deaths per capita. The dataset I employ in this study comprises 59 explanatory variables for 179 unique countries/territories on the 13^{th} April 2020. The data was obtained from three online sources, https://www.worldometers.info/coronavirus/, https://ourworldindata.org/ and the World Bank Data API (application program interface). A table of variable descriptions and their respective sources is included in the appendix.

In this paper I employ three varying ordinary least squares (OLS) multiple linear regression techniques in order to analyse the relationship between the characteristics of the 179 countries and their respective deaths and cases per capita as at 13^{th} April 2020. Firstly, I perform two OLS regressions, one for each dependent variable, on the entire dataset in order to get an overview of the significance and magnitude of the explanatory variables in predicting our two key variables. The two regressions have Newey-West standard errors, that is they are heteroskedastic and autocorrelation robust standard errors.

Multicollinearity analysis is then performed in order to exclude highly correlated explanatory variables using the variance inflation factor (VIF) from the cross-sectional regression model. Using the threshold value of 10, as originally suggested by (Greene and Kennedy, 1993), any explanatory variable with a VIF above this threshold is removed from the dataset resulting in a reduced number of predictors. In this way the multicollinearity of the predictors is greatly reduced and thus the accuracy of the significance of model variables is improved.

The second OLS cross-sectional model is constructed from the reduced dataset (outlined above) and uses forward stepwise selection to choose explanatory variables. The Akaike and Bayesian information criterion, AIC and BIC respectively are used to determine the optimal model parameters as discussed in (Kadane and Lazar, 2004). Both criterion are used as the BIC usually results in a more parsimonious model than the AIC (Zhang, 2016).

The final cross-sectional model is a lasso regression model where the optimal lambda parameters are calculated using AIC,BIC and cross-validation methods. A lasso regression like an OLS regression minimizes the sum of squared residuals, however the lasso imposes the restriction that the sum of the absolute value of the coefficients have to be less than a constant. This produces interpretable and parsimonious models that are stable (Tibshirani, 1996).

Equation 1: Cost Function for Lasso Regression

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

4. Findings

Included in table 1 below are the results of the first of the cross-sectional regression models using the entire dataset.

Table 1: OLS Regression Results - Entire Dataset

	Cases per Capita		*Deaths per Capita*	
Variables	**Coefficient**	**Standard Error**	**Coefficient**	**Standard Error**
3.0.Gini	0.00185	(0.00245)	0.00043*	(0.00024)
6.0.Conspc	-0.00000	(0.00000)	-0.00000	(0.00000)
9.0.Unemp.All	-0.00005	(0.00006)	-0.00000	(0.00000)
Days Since First Case	0.00000	(0.00001)	0.00000	(0.00000)
EN.ATM.CO2E.PC	-0.00001	(0.00002)	-0.00000	(0.00000)
EN.ATM.GHGT.KT.CE	0.00005	(0.00019)	0.00002	(0.00001)
EN.ATM.PM25.MC.M3	-0.00000	(0.00001)	-0.00000	(0.00000)
EN.POP.SLUM.UR.ZS	0.00000	(0.00001)	0.00000	(0.00000)
EN.URB.MCTY.TL.ZS	-0.00002**	(0.00001)	-0.00000***	(0.00000)
IC.BUS.EASE.XQ	0.00000	(0.00000)	0.00000	(0.00000)
IS.AIR.PSGR	0.00001	(0.00000)	0.00000**	(0.00000)
IS.SHP.GCNW.XQ	0.00000	(0.00000)	0.00000	(0.00000)
NE.EXP.GNFS.ZS	0.00000	(0.00000)	0.00000	(0.00000)
NE.IMP.GNFS.ZS	-0.00000	(0.00000)	-0.00000	(0.00000)
NE.TRD.GNFS.ZS	0.00000	(0.00000)	0.00000	(0.00000)
SE.ADT.LITR.ZS	0.00000	(0.00000)	0.00000	(0.00000)
SE.PRM.CMPT.ZS	-0.00001	(0.00001)	0.00000	(0.00000)
SH.ALC.PCAP.LI	0.00002	(0.00003)	-0.00000	(0.00000)
SH.DTH.0514	-0.02474*	(0.01424)	-0.00316***	(0.00120)
SH.DTH.COMM.ZS	-0.00000	(0.00002)	0.00000	(0.00000)
SH.DYN.AIDS.ZS	0.00002	(0.00003)	0.00000	(0.00000)
SH.DYN.MORT	-0.00003	(0.00002)	-0.00000	(0.00000)
SH.H2O.BASW.ZS	-0.00001	(0.00001)	-0.00000**	(0.00000)
SH.H2O.SMDW.ZS	0.00001*	(0.00000)	0.00000	(0.00000)
SH.IMM.HEPB	-0.00001	(0.00001)	-0.00000	(0.00000)
SH.IMM.IDPT	0.00000	(0.00002)	0.00000	(0.00000)
SH.IMM.MEAS	0.00002	(0.00001)	-0.00000	(0.00000)
SH.MED.CMHW.P3	0.00015	(0.00090)	0.00003	(0.00004)
SH.MED.NUMW.P3	0.00003	(0.00004)	-0.00000	(0.00000)
SH.MED.PHYS.ZS	-0.00004	(0.00008)	0.00000	(0.00001)
SH.STA.AIRP.P5	0.00001***	(0.00000)	0.00000***	(0.00000)
SH.STA.BASS.ZS	-0.00001	(0.00001)	0.00000	(0.00000)
SH.STA.BFED.ZS	-0.00001	(0.00001)	-0.00000	(0.00000)
SH.STA.DIAB.ZS	0.00006	(0.00004)	0.00000	(0.00000)
SH.STA.ODFC.ZS	-0.00001*	(0.00001)	-0.00000	(0.00000)
SH.STA.WASH.P5	0.00002**	(0.00001)	0.00000	(0.00000)
SH.TBS.INCD	0.00000	(0.00000)	-0.00000	(0.00000)
SH.VAC.TTNS.ZS	-0.00004*	(0.00002)	-0.00000	(0.00000)
SI.DST.FRST.20	-0.00006	(0.00007)	0.00000	(0.00000)
SI.POV.DDAY	-0.00001*	(0.00001)	-0.00000***	(0.00000)
SM.POP.NETM	0.00000	(0.00000)	-0.00000	(0.00000)
SN.ITK.DEFC.ZS	-0.00001	(0.00001)	0.00000	(0.00000)
SP.DYN.AMRT.MA	0.00000	(0.00000)	0.00000	(0.00000)

SP.DYN.CDRT.IN	-0.00009	(0.00013)	0.00000	(0.00001)
SP.DYN.LE00.IN	0.00008	(0.00009)	0.00001*	(0.00001)
SP.POP.BRTH.MF	-0.00186	(0.00427)	-0.00044	(0.00029)
SP.POP.TOTL.MA.ZS	-0.00004	(0.00003)	-0.00000	(0.00000)
SP.URB.TOTL.IN.ZS	0.00001	(0.00001)	0.00000***	(0.00000)
VC.PKP.TOTL.UN	0.00000*	(0.00000)	0.00000*	(0.00000)
aged_70_older	0.00001	(0.00005)	0.00000	(0.00000)
Intercept	0.00269	(0.00807)	-0.00042	(0.00051)
diabetes_prevalence	-0.00011***	(0.00004)	-0.00000**	(0.00000)
extreme_poverty	-0.00000	(0.00001)	0.00000	(0.00000)
gdp_per_capita	0.00000	(0.00000)	0.00000	(0.00000)
handwashing_facilities	0.00000	(0.00000)	0.00000	(0.00000)
hospital_beds_per_100k	-0.00008*	(0.00004)	-0.00000	(0.00000)
male_smokers	0.00000	(0.00001)	0.00000	(0.00000)
median_age	-0.00000	(0.00005)	-0.00000	(0.00000)
population_density	-0.00000	(0.00000)	-0.00000	(0.00000)
total_cases			0.00000	(0.00000)
total_tests	0.00037	(0.00027)		(0.00027)
N	179		179	
R^2	0.6547		0.6967	
AIC	-1946.7203		-2907.0342	
BIC	-1758.6646		-2718.9785	

As we can see in the above table many of the variables in both the cases per capita and deaths per capita regressions are insignificant and close to zero. We do observe however that there are some significant, non-zero coefficients such as SH.DTH.0514 (number of deaths ages 5-14 years) which is significant at the 10% level and 1% level for cases and deaths per capita respectively, and has the highest magnitude of any explanatory variable considered.

We see that the R-squared values for the two regressions are 0.65 and 0.70 respectively which indicates a relatively good fit considering the nature of the variability of the dependent variables and the cross-sectional nature of the model. The lower AIC and BIC values for the deaths per capita regression reflect the better fit compared to the cases model. However, the condition number for both regressions is extremely large indicating there is a strong multicollinearity problem between regressors. Therefore neither the significance nor the magnitude of the regression coefficients can be considered accurate or statistically robust. We must perform dimensionality reduction and parsimonious model selection in order to decrease the multicollinearity in predictor variables.

Table 2 displays the results of the OLS regressions using the AIC and BIC forward stepwise regressions for feature selection as previously mentioned. The stepwise regressions are performed on the reduced version of the dataset after discarding predictors with a VIF (variance inflation factor) greater than 10.

Table 2 OLS Regression results – Stepwise Regressions on reduced feature dataset

	Cases per Capita		Deaths per Capita	
Variables	**AIC**	**BIC**	**AIC**	**BIC**
Days Since First Case	0.00001***			
EN.ATM.GHGT.KT.CE	0.00048*	0.00034**	0.00005*	0.00005*
EN.ATM.PM25.MC.M3			-0.00000***	
EN.POP.SLUM.UR.ZS	0.00001***		0.00000***	
EN.URB.MCTY.TL.ZS	-0.00001***		-0.00000***	-0.00000***
IS.SHP.GCNW.XQ			0.00000**	
SH.ALC.PCAP.LI	0.00006**			
SH.DTH.0514	-0.00909		-0.00181	-0.00188
SH.H2O.SMDW.ZS	0.00001***			
SH.MED.NUMW.P3	0.00004	0.00008***		
SM.POP.NETM			-0.00000*	
SP.URB.TOTL.IN.ZS	0.00002***	0.00001***	0.00000***	0.00000***
Intercept	-0.00207***	-0.00089***	-0.00007**	-0.00004**
hospital_beds_per_100k	-0.00010***			
population_density	-0.00000		-0.00000	-0.00000
total_cases			0.00000	0.00000
N	179	179	179	179
R^2	0.4407	0.3509	0.4557	0.4125
AIC	-1954.3970	-1943.7626	-2898.3802	-2892.7162
BIC	-1916.1484	-1931.0131	-2863.3190	-2870.4045

We can see that after the VIF feature reduction and forward stepwise regressions the number of explanatory variables has significantly lessened. The R-squared values are sizably smaller than our overall regression suggesting a worse fit. However, the condition numbers have decreased enormously and we have obtained several significant predictor variables for both models.

Both EN.ATM.GHGT.KT.CE (Total greenhouse gas emissions) and SP.URB.TOTL.IN.ZS (Urban Population % of total population) are statistically significant in each model variation. The coefficients for both predictors are positive indicating that increased emissions and a larger proportionate urban population result in greater confirmed cases and deaths per capita. Interestingly, this mirrors results obtained in a recent publication by (Martelletti and Martelletti, 2020) linking increased levels of air pollution with increased COVID-19 death rates.

Other significant explanatory variables for cases per capita include Days Since First Case (number of days since first confirmed case), SH.ALC.PCAP.LI (total alcohol consumption per capita) and EN.POP.SLUM.UR.ZS (Population living in slums % of urban population).

Finally, table 3 presents the results from the lasso regressions where the optimal lambda parameter is estimated using AIC, BIC and cross-validation methods.

Table 3 Lasso Regression Results – Lasso Regressions on reduced feature set

	Cases per Capita			Deaths per Capita		
Variables	**AIC**	**BIC**	**CV**	**AIC**	**BIC**	**CV**
Days Since First Case	0.00000	0.00000	0.00000	0.00000	0.00000	
EN.ATM.GHGT.KT.CE	0.00036***	0.00036***	0.00036***	0.00004***	0.00004***	0.00002***
EN.URB.MCTY.TL.ZS				-0.00000*	-0.00000*	
IS.SHP.GCNW.XQ				0.00000*	0.00000*	
SH.ALC.PCAP.LI	0.00002	0.00002	0.00002			
SH.DTH.0514				-0.00143***	-0.00143***	
SH.H2O.SMDW.ZS	-0.00001**	-0.00001**	-0.00001**	0.00000	0.00000	
SH.MED.NUMW.P3	0.00006**	0.00006**	0.00006**			
SH.MED.PHYS.ZS				0.00001**	0.00001**	
SP.URB.TOTL.IN.ZS	0.00001*	0.00001*	0.00001*	0.00000	0.00000	
Intercept	0.00000	0.00000	0.00000	-0.00004*	-0.00004*	0.00001**
population_density				0.00000	0.00000	
total_cases				0.00000	0.00000	
N	179	179	179	179	179	179
R^2	0.3122	0.3122	0.3122	0.3449	0.3449	0.1386
AIC	-1927.4034	-1927.4034	-1927.4034	-2871.2149	-2871.2149	-2832.2066
BIC	-1905.0917	-1905.0917	-1905.0917	-2845.7159	-2845.7159	-2822.6444

We immediately observe from table 3 that all of the three cases per capita models are identical. The optimal lambda value estimated from each of the methods is the same and thus the same lasso regression model is fitted. Also, the AIC and BIC models are identical for the deaths per capita lasso regression. The R-squared values are much lower than the previous regressions and we have only about 30% of the variability in either cases or deaths per capita is explained by the predictor variables.

Observing the cases per capita lasso regression we see that again EN.ATM.GHGT.KT.CE (Total greenhouse gas emissions) is significant at the 1% level and is the largest coefficient in terms of magnitude. The predictors SH.H2O.SMDW.ZS (People using safely managed drinking water services (% of population)), SH.MED.NUMW.P3 (Nurses and midwives (per 1,000 people)) and SP.URB.TOTL.IN.ZS (Urban population (% of total population)) are all significant at the 10% level. Viewing the death per capita lasso regressions results, the significant predictors are EN.ATM.GHGT.KT.CE (Total greenhouse gas emissions) and SH.DTH.0514 (Number of deaths ages 5-14 years).

5. Research Limitations

The first and most critical limitation of this study is due to the fact that I employ a cross-sectional regression methodology. This study only considers data for one time point (April 13th 2020) and therefore suffers from a difficulty in making causal inferences and can be skewed by noise in the data occurring in the chosen time point (i.e. April 13th). This means that the results obtained in this study may not be representative of the true demographic characteristics that affect the number of confirmed cases and deaths due to April 13th being an outlier (Levin, 2006).

Secondly, in some cases the explanatory data obtained is over 8 years old. These variables pertaining to each of the 179 countries studied may have changed significantly in this time period and therefore had more recent data been available, better and more accurate inferences could be made. There was also a large number of missing values in the predictor variable dataset. A large amount of missing values in a study can introduce inconsistent bias and lead to an underestimation of standard errors of coefficients in regression models. Therefore caution must be taken when drawing substantial conclusions from these regression model results (Kang, 2013).

6. Research Implications

While there are several limitations affecting this study there are still many significant results and conclusions that can be drawn. A large number of countries were used in each of the above regressions adding to the robustness of the significant explanatory variables. It was found that total greenhouse gas emissions was a highly significant in every single regression aside from the overall regression using the entire dataset. This complements results obtained by (Martelletti and Martelletti, 2020) and (Wu, C Nethery, Braun and Dominici, 2020) which indicate that increased levels of air pollution significantly increase the death rate of COVID-19.

These results while only cross-sectional, warrant further investigation into the relationship between air pollution and the impact of the COVID-19 virus on deaths and cases per capita. These preliminary results can be used immediately however to more thoroughly prepare and equip countries with reliable information. This information can be used to aid governments in making correct policy decisions in terms of lockdown protocols and when to resume normal economic activity which has the power to save human lives and money.

Bibliography

1. Fernandes, N., 2020. Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy. *SSRN Electronic Journal*,.
2. Greene, W. and Kennedy, P., 1993. A Guide to Econometrics. *The Journal of Economic Education*, 24(4), p.371.
3. Jones, L., Palumbo, D. and Brown, D., 2020. *Coronavirus: A Visual Guide To The Economic Impact*. [online] BBC News. Available at: <https://www.bbc.com/news/business-51706225> [Accessed 19 May 2020].
4. Kadane, J. and Lazar, N., 2004. Methods and Criteria for Model Selection. *Journal of the American Statistical Association*, 99(465), pp.279-290.
5. Kang, H., 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), p.402.
6. Levin, K., 2006. Study design III: Cross-sectional studies. *Evidence-Based Dentistry*, 7(1), pp.24-25.
7. Martelletti, L. and Martelletti, P., 2020. Air Pollution and the Novel Covid-19 Disease: a Putative Disease Risk Factor. *SN Comprehensive Clinical Medicine*, 2(4), pp.383-387.
8. Roser, M., Ritchie, H., Ortiz-Ospina, E. and Hasell, J., 2020. *Coronavirus Pandemic (COVID-19)*. [online] Our World in Data. Available at: <https://ourworldindata.org/coronavirus> [Accessed 19 May 2020].
9. Sannigrahi, S., Pillaa, F., Basua, B. and Sarkar, A., 2020. The overall mortality caused by COVID-19 in the European region is highly associated with demographic composition: A spatial regression-based approach. [online] Available at: <https://arxiv.org/ftp/arxiv/papers/2005/2005.04029.pdf> [Accessed 19 May 2020].
10. Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp.267-288.
11. Wu, X., C Nethery, R., Braun, D. and Dominici, F., 2020. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study. [online] Available at: <https://www.medrxiv.org/content/medrxiv/early/2020/04/27/2020.04.05.20054502.full.pdf> [Accessed 19 May 2020].
12. Zhang, Z., 2016. Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, 4(7), pp.136-136.

Appendix

Table 4 Variable Descriptions and Sources

Variable Code	**Variable Description**	**Source**
3.0.Gini	Gini Coefficient	World Bank API
6.0.Conspc	Consumption per capita (2011 $)	World Bank API
9.0.Unemp.All	Unemployed (%)	World Bank API
aged_70_older	Share of the population that is 70 years and older in 2015	https://ourworldindata.org/coronavirus
Days Since First Case	Number of elapsed days since first confirmed case	https://www.worldometers.info/coronavirus/
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017	https://ourworldindata.org/coronavirus
EN.ATM.CO2E.PC	CO2 emissions (metric tons per capita)	World Bank API
EN.ATM.GHGT.KT.CE	Total greenhouse gas emissions (kt of CO2 equivalent)	World Bank API
EN.ATM.PM25.MC.M3	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)	World Bank API
EN.POP.SLUM.UR.ZS	Population living in slums (% of urban population)	World Bank API
EN.URB.MCTY.TL.ZS	Population in urban agglomerations of more than 1 million (% of total population)	World Bank API
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010	https://ourworldindata.org/coronavirus
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available	https://ourworldindata.org/coronavirus

handwashing_facilities	Share of the population with basic handwashing facilities on premises, most recent year available	https://ourworldindata.org/coronavirus
hospital_beds_per_100k	Hospital beds per 100,000 people, most recent year available since 2010	https://ourworldindata.org/coronavirus
IC.BUS.EASE.XQ	Ease of doing business index (1=most business-friendly regulations)	World Bank API
IS.AIR.PSGR	Air transport, passengers carried	World Bank API
IS.SHP.GCNW.XQ	Liner shipping connectivity index (maximum value in 2004 = 100)	World Bank API
male_smokers	Share of men who smoke, most recent year available	https://ourworldindata.org/coronavirus
median_age	Median age of the population, UN projection for 2020	https://ourworldindata.org/coronavirus
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)	World Bank API
NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)	World Bank API
NE.TRD.GNFS.ZS	Trade (% of GDP)	World Bank API
population_density	Number of people divided by land area, measured in square kilometers, most recent year available	https://ourworldindata.org/coronavirus
SE.ADT.LITR.ZS	Literacy rate, adult total (% of people ages 15 and above)	World Bank API
SE.PRM.CMPT.ZS	Primary completion rate, total (% of relevant age group)	World Bank API
SH.ALC.PCAP.LI	Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)	World Bank API
SH.DTH.0514	Number of deaths ages 5-14 years	World Bank API

SH.DTH.COMM.ZS	Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)	World Bank API
SH.DYN.AIDS.ZS	Prevalence of HIV, total (% of population ages 15-49)	World Bank API
SH.DYN.MORT	Mortality rate, under-5 (per 1,000 live births)	World Bank API
SH.H2O.BASW.ZS	People using at least basic drinking water services (% of population)	World Bank API
SH.H2O.SMDW.ZS	People using safely managed drinking water services (% of population)	World Bank API
SH.IMM.HEPB	Immunization, HepB3 (% of one-year-old children)	World Bank API
SH.IMM.IDPT	Immunization, DPT (% of children ages 12-23 months)	World Bank API
SH.IMM.MEAS	Immunization, measles (% of children ages 12-23 months)	World Bank API
SH.MED.CMHW.P3	Community health workers (per 1,000 people)	World Bank API
SH.MED.NUMW.P3	Nurses and midwives (per 1,000 people)	World Bank API
SH.MED.PHYS.ZS	Physicians (per 1,000 people)	World Bank API
SH.STA.AIRP.P5	Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 population)	World Bank API
SH.STA.BASS.ZS	People using at least basic sanitation services (% of population)	World Bank API

SH.STA.BFED.ZS	Exclusive breastfeeding (% of children under 6 months)	World Bank API
SH.STA.DIAB.ZS	Diabetes prevalence (% of population ages 20 to 79)	World Bank API
SH.STA.ODFC.ZS	People practicing open defecation (% of population)	World Bank API
SH.STA.WASH.P5	Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)	World Bank API
SH.TBS.INCD	Incidence of tuberculosis (per 100,000 people)	World Bank API
SH.VAC.TTNS.ZS	Newborns protected against tetanus (%)	World Bank API
SI.DST.FRST.20	Income share held by lowest 20%	World Bank API
SI.POV.DDAY	Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)	World Bank API
SM.POP.NETM	Net migration	World Bank API
SN.ITK.DEFC.ZS	Prevalence of undernourishment (% of population)	World Bank API
SP.DYN.AMRT.MA	Mortality rate, adult, male (per 1,000 male adults)	World Bank API
SP.DYN.CDRT.IN	Death rate, crude (per 1,000 people)	World Bank API
SP.DYN.LE00.IN	Life expectancy at birth, total (years)	World Bank API
SP.POP.BRTH.MF	Sex ratio at birth (male births per female births)	World Bank API
SP.POP.TOTL.MA.ZS	Population, male (% of total population)	World Bank API
SP.URB.TOTL.IN.ZS	Urban population (% of total population)	World Bank API
total_cases	Total confirmed cases of COVID-19	https://ourworldindata.org/coronavirus
total_deaths	Total deaths attributed to COVID-19	https://ourworldindata.org/coronavirus

total_tests	Total tests for COVID-19	https://ourworldindata.org/coronavirus
VC.PKP.TOTL.UN	Presence of peace keepers (number of troops, police, and military observers in mandate)	World Bank API