Country political regime and data transparency, and effects on forecasting COVID-19 cases per million

Word Count: 3185

# **Introduction**

The global pandemic caused by COVID-19 has had profound effects on daily life worldwide, with over 660 million total reported cases worldwide and 6.7 million confirmed deaths (as of January 2023) (WorldOMeter, 2022).

Previous cases have illustrated a correlation of democratic governance and increased health outcomes, such as increased life expectancy, reduced infant mortality, etc (Preston, 1975) (Kudamatsu, 2012) (Cassan & Van Steenvoort, 2021) (Bollyky et al., 2019). Whilst this is the case, COVID-19 has tended towards a correlation in the other direction, highlighting a correlation between democratic governance and COVID-19 metrics such as an increased case fatality rate (Sorci et al., 2020), illustrating that democratic governance has a negative effect on health outcomes. From this, and with such a vast amount of country level COVID-19 data available to analyse, the impacts of political regime on COVID-19 data have formed two differing bodies of evidence into the effects of political regime on national health outcomes, forming two different views: the ‘efficient autocracy’ view and the ‘biasing autocracy’ view (Cassan & Van Steenvoort, 2021).

On one hand, the ‘efficient autocracy’ view suggests that this correlation could be explained by autocratic governments are advantaged in the COVID-19 pandemic, as they are more efficient in implementing policies that could retain the spread of cases (Cassan & Van Steenvoort, 2021), owing to not being limited by the respect for individual rights and freedom that ideal democratic governance has (Norheim et al., 2020).

On the other hand, the ‘biasing autocracy’ view suggests that the increased COVID-19 cases and deaths in democratic governance could be explained by the idea that autocracies manipulate COVID-19 statistics, and hence under-report covid-19 case and death data (Cassan & Van Steenvoort, 2021).

One suggested explanation for this is detailed by (Kapoor et al., 2020), noting that as economic policies and political outcomes are affected by COVID-19 statistics, governments have the incentive to manipulate their data to favour better economic and political outcomes. (Kilani, 2021) analysed country level COVID-19 data of authoritarian countries, by moving average, and found a strong, significant correlation that data has been manipulated. Multiple other studies have also corroborated these findings (Wigley, 2022).

Furthermore, owing to the global effects of the COVID-19 pandemic, forecasting future cases using existing data has become an invaluable tool in allowing for better country level decision making. Forecasting COVID cases allows for the better implementation of prevention strategies, such as policies (lockdowns, mask mandates, social distancing rules, etc.), better planning and preparation, and better management of resources and medical facilities (Wang et al., 2022) (Pan American Health Organization, 2020). Forecasting COVID-19 cases is inherently tied to the accurate reporting of COVID-19 cases, wherein any under-reporting of cases could both decrease the accuracy of the model (if under-reporting is inconsistent) and lead to negative impacts on the benefits of forecasting in the first place (such as the model predicting lower levels of cases than what would otherwise be predicted as a result of the under-reporting, leading to worse planning, preparation, and changes to policy). An example of this is (Albani et al., 2021), wherein the paper notes how the under-reporting of COVID-19 data has had a negative impact on vaccination strategies.

# **Aims**

This essay aims to build on the work of papers such as (Annaka, 2021), analysing and comparing the ‘efficient autocracy’ vs ‘biasing autocracy’ views, first looking into the relationship of political regime and government transparency with a different COVID-19 statistic, cumulative cases per million, determining if there is any significant relationships (p < 0.05) with them, then analysing if there are any other COVID-19 variables that government transparency significantly correlates with.

Furthermore, as papers such as (Annaka, 2021) have shown a significant correlation of government transparency with COVID-19 metrics, it would also be interesting to analyse if government transparency has a relationship with the prediction accuracy of modelling COVID-19 cases. Therefore, this essay aims to forecast COVID-19 cumulative cases per million, and use the accuracy of the models of each country to determine if there is a significant correlation between government transparency and the accuracy in the prediction of future COVID-19 cases.

The aim of the essay is to answer the following questions, in order:

1. How does the political regime of a country correlate with COVID-19 cases per million?
2. Does government transparency correlate with COVID-19 cases per million?
3. Are there any other COVID-19 variables that government transparency significantly correlates with?
4. Does data transparency have a significant effect on the level of accuracy in predicting future COVID-19 cases?

# **Methodology**

## Datasets and Variables

Four datasets were used to conduct the aims of this paper:

1. Our World in Data: COVID-19 dataset

* location, date, total\_cases, total\_cases\_per\_million, gdp\_per\_capita, life\_expectancy, population\_density, stringency\_index, total\_deaths\_per\_million, aged\_70\_and\_older

<https://ourworldindata.org/coronavirus>

1. Our World in Data: democracy dataset

* electdem\_vdem\_owid, regime\_fh

<https://ourworldindata.org/democracy>

1. V-Dem Project dataset Version 12

* v2x\_polyarchy

<https://www.v-dem.net/data/the-v-dem-dataset/>

1. HRV Transparency Project dataset (2010 newest)

* gov\_transparency

<https://hrvtransparency.org/>

v2x\_polyarchy (also known as the Multiplicative Polyarchy Index (MPI)) is a measure of level of democracy, calculated through 5 other variables components of electoral democracy (freedom of association, clean elections, freedom of expression, elected officials, and suffrage). v2x\_polyarchy is coded on a scale from 0 to 1 (low to high).

𝑣2𝑥\_𝑝𝑜𝑙𝑦𝑎𝑟𝑐ℎ𝑦= .5(𝑣2𝑥\_𝑒𝑙𝑒𝑐𝑜𝑓𝑓 ∗ 𝑣2𝑥𝑒𝑙\_𝑓𝑟𝑒𝑓𝑎𝑖𝑟 ∗𝑣2𝑥\_𝑓𝑟𝑎𝑠𝑠𝑜𝑐\_𝑡ℎ𝑖𝑐𝑘 ∗𝑣2𝑥\_𝑠𝑢𝑓𝑓𝑟 ∗ 𝑣2𝑥\_𝑓𝑟𝑒𝑒\_𝑎𝑙𝑡𝑖𝑛𝑓)

Likewise, gov\_transparency is a measure of the level of a government’s willingness to disclose internal affairs, coded on a scale of -10 to 10 (low to high government transparency).

## Pre-Processing and Exploration of Dataset

Pre-processing of the data was conducted using the Tidyverse package in R Studio. Each dataset was first cleaned, filtering datasets by variables of interest, removing redundant/duplicate data by filtering data independent datasets by date, and renaming variables with unclear names (such as v2x\_polyarchy to level\_of\_democracy). Furthermore, each dataset was then the joined together (to the covid dataset) to form a complete/master dataset, with non-country locations being filtered out from the dataset as they are not relevant for this essay. The lubridate package was also used to convert dates from character data to date data, to allow for time-series analysis.

Subset datasets were created during each stage of the analysis to only include relevant variables and data.

Initial exploration into the data was conducted, with statistical analyses performed on each numerical variable to find the minimum, median, maximum, mean, standard deviation, percentage of missing data, and coefficient of skewness. Results were formulated into a clean data frame, with the results shown in appendix table 1.

Percentage of missing data was used to analyse whether a variable was able to be used in analysis, with a significant % of missing data suggesting that the variable should not be used. Initially, the plan was to conduct an investigation into how country the percentage of total vaccinations of a country affected the cumulative excess mortality (per million), however when looking at the percentages of missing values for cumulative excess mortality (per million), it was shown to be 96.41% (as detailed in appendix table 2), highlighting that the majority of data for the variable was absent. Therefore, analysis of the variable would either result in a low statistic power, or have a high chance of being incorrect from biased imputation. Therefore, other variables with lower percentages of missing values were explored, to retain a stronger statistical power. Total cases per million resulted a percentage of missing data of 4.24%, showing that the majority of data was present to be accurately analysed.

## Handling Missing Data

Missing data handling was performed on variables: population, total\_cases, total\_cases\_per\_million, level\_of\_democracy, and gov\_transparency. Percentage of missing data and coefficient of skewness were used alongside missing data visualisations (using the VIM and naniar packages) to analyse how the missing data should be handled. A coefficient of skewness < 0.5 would allow for imputation using the mean, a coefficient of skewness > 0.5 would suggest regression imputation would result more accurate values.

Missing data visualisations were used to: plot percentages of missing data and location of missing data using a histogram and pattern plot respectively (appendix figure 1), a matrix plot looking in detail at locations of missing data for all variables of interest (appendix figure 2), a missing values histogram visualising overlapping levels of missing data between variables (appendix figure 3), and finally a margin plot showing the distribution of missing and present data for level\_of\_democracy vs gov\_transparency to assess the missing at randomness of the variables (appendix figure 4).

As total\_cases is cumulative, using the mean to impute could result in earlier dates having higher total\_cases than later dates, meaning that mean imputation should not be used. Using the zoo package, imputation by linear interpolation was performed on total\_cases, predicting missing values off present data, and replacing total\_cases missing data with the approximations. Any countries lacking any data to allow for approximations to be made were omitted from the dataset. These countries included: Guam, Guernsey, Jersey, Nauru, Niue, Northern Mariana Islands, Pitcairn, Puerto Rico, Sint Maarten (Dutch part), Tokelau, Turkmenistan, Tuvalu, United States Virgin Islands.

As rows with missing data for population (0.17% of the data) were omitted from analysis, missing data between total\_cases and total\_cases\_per\_million were found to be the same. total\_cases\_per\_million missing values were imputed using (population / 1,000,000 \* total\_cases), converting the imputed total\_cases missing data approximations into total\_cases\_per\_million.

Both electoral democracy and political freedom variables from ‘Our World in Data’ were used to impute missing values of level\_of\_democracy. electoral democracy and level\_of\_democracy plotted 1 to 1 data (figure 1) even though some electoral democracy data was present where level\_of\_democracy data was missing.

Chart, line chart

Description automatically generated

*Figure 1: A scatter graph plotting electoral democracy vs level of democracy.*

political freedom and level\_of\_democracy plotted a strong correlation by way of distinct separation of level\_of\_democracy by level of political freedom (Not Free, Partly Free, Free) (figure 2).

Chart, scatter chart

Description automatically generated

*Figure 2: A scatter graph with x-axis jitter plotting the category of political freedom (not free, partly free, free) against level of democracy.*

Where data for both electoral democracy and political freedom were missing, the mean of level\_of\_democracy was imputed. Therefore, both of level\_of\_democracy missing data was filled using the following criteria:

1. impute electoral democracy
2. impute mean values by political freedom category
3. impute mean value of level\_of\_democracy

Finally, countries lacking any government transparency data were omitted from the dataset. This was performed last to retain statistical power whilst performing previous variable imputations.

## Analysis

All data analysis was performed using R Studio. Significance was calculated using a 95% confidence interval for the project (p < 0.05).

#### How does the political regime of a country correlate with COVID-19 cases per million? Does government transparency correlate with COVID-19 cases per million?

Two scatter graphs were used (using the ggplot2 package in Tidyverse) to plot political regime and government transparency versus COVID-19 cases per million. The scatter graphs included a linear line of best fit, correlation coefficient, and significance value (R and p values were calculated using Pearson correlation).

#### Are there any other COVID-19 variables that government transparency significantly correlates with?

A correlation matrix was plotted using the corrplot package, to analyse correlations between government transparency and other numerical COVID-19 variables. The correlation plot used pairwise deletion when comparing each variable with another, with the correlation plot only displaying significant correlations between variables on a scale of -1 to 1 with -1 being a perfect negative correlation and 1 being a perfect positive correlation. Variables used in this correlation analysis were population, stringency\_index, population\_density, gdp\_per\_capita, life\_expectancy, total\_cases, total\_cases\_per\_million, level\_of\_democracy, electoral\_democracy, total\_deaths\_per\_million, and gov\_transparency

Does data transparency have a significant effect on the level of accuracy in predicting future COVID-19 cases?

First, analysis into the nature of total\_cases\_per\_million (if total\_cases\_per\_million over time was linear/logarithmic) was conducted using a line graph with a line of best fit (using the ggplot2 package in Tidyverse), plotting the time series data for an example country, Bulgaria, and calculating the Pearson correlation coefficient, and significance value (shown in figure 3).

Chart, line chart

Description automatically generated

*Figure 3: A line graph plotting the time-series of total cases per million between 2020-03-06 and 2022-07-03.*

The p < 0.001 highlights a significant linear relationship in the time-series data and therefore, a linear forecasting model could be used in further analysis. Furthermore, the creation of training and testing datasets (80/20% split) were pre-processed for forecasting. Time-Series forecasting was carried out using the forecast package (along with the tsibble, fable, and feasts packages to allow for forecasting and analysis of data grouped by country) using the autoregressive integrated moving average (ARIMA) model, predicting 192 future cases (the length of the testing dataset) using the training dataset. ARIMA models and predicted forecasts were calculated for each country. An example of time-series forecasting is plotted using a line graph of the train, test, and forecasting data for Bulgaria (shown in figure 4).

Chart, line chart

Description automatically generated

*Figure 4: A line graph of the forecast of total cases per million, with training data shown in blue, actual (test) data shown in green and predicted (forecast) data shown in red.*

Using the predicted forecasts for each country, the Root Mean Squared Error (RMSE) was calculated using the Metrics package. RMSE is a performance indicator used to analyse the measure of the average difference between predicted and actual values, with a lower value indicating a better model fit. From this, a normalised Root Mean Squared Error (nRMSE) was calculated for each of the countries using the formula:

Normalisation of RMSE was performed to allow for the analysis of RMSE values from different datasets (countries), as RMSE values from different datasets are not comparable. RMSE and nRMSE results are shown in appendix table 2. Finally, nRMSE values were plotted on a scatter graph to look for any significant correlation between government transparency and nRMSE.

## R Studio Packages

|  |  |
| --- | --- |
| **Package Used** | **Use** |
| tidyverse | Involved in data exploration, manipulation, and plotting |
| lubridate | Dealing with date data |
| VIM | Plotting missing data visualisations |
| naniar | Plotting missing data visualisations |
| zoo | Linear interpolation of linear time series data (in NA handling for this essay) |
| ggpubr | Calculation of correlation coefficient and statistical significance for ggplot graphs |
| corrplot | Plotting correlation plots |
| forecast | Time series forecasting |
| tsibble | Allows for data frames using temporal structure to analyse time series by group |
| fable | Time series forecasting |
| feasts | Analysing tidy time series data |
| Metrics | Calculating RMSE values for predicted vs actual data |

*Table 1: A table of the packages used for the project, including their use case*

# **Results and Discussion**

## How does the political regime of a country correlate with COVID-19 cases per million?

Figure 5 shows the relationship between level of democracy (MPI) and total cases per million between January 1st 2020 and July 3rd 2022. The scatter graph results a Pearson correlation coefficient (R) value of 0.67 and a p value < 0.001, highlighting a moderate, positive correlation. These findings highlight that an increase in Multiplicative Polyarchy Index (an increased MPI meaning more democratic) is significantly correlated (compared to p < 0.05) with an increase in total cases per million.

Chart, scatter chart

Description automatically generated

*Figure 5: A scatter graph of the level of democracy against total cases per million, with a line of best shown in orange.*

## Does country data transparency correlate with COVID-19 cases per million?

Figure 6 shows the relationship between government transparency and total cases per million between January 1st 2020 and July 3rd 2022. The scatter graph results a Pearson correlation coefficient (R) value of 0.58 and a p value < 0.001, highlighting, similarly to that of the previous section, a moderate, positive correlation that is significant. Therefore, countries with higher government transparency, also tend to report higher total cases per million.

These results support the ‘biasing autocracy’ view that autocratic favourability in health outcomes in COVID-19 is as a result of under-reporting data, owing to the significant correlation between variables implying that countries with higher tendencies to disclose reliable information also report higher total cases per million.

Chart, scatter chart

Description automatically generated

*Figure 6: A scatter graph of the government transparency against total cases per million, with a line of best shown in orange.*

## Are there any other COVID-19 variables that government transparency significantly correlates with?

Figure 7 shows a correlation plot noting any significant correlations between all variables noted. When looking at the correlation between government transparency (gov\_transparency) and other related COVID-19 variables of interest, 3 significant correlations can be observed. These are life expectancy (with a correlation coefficient of 0.65), total deaths per million (a correlation coefficient of 0.63), and population aged 70 and older (a correlation coefficient of 0.73). All 3 variables show a moderate positive correlation with that of government transparency.

Chart

Description automatically generated

*Figure 7: A correlation plot analysing significant correlations of 11 variables, with significant positive associations shown in blue and significant negative correlations shown in red.*

These findings support existing literature (Cassan & Van Steenvoort, 2021) reporting a correlation between government transparency and total deaths per million. The correlation between government transparency and total deaths per million could be argued in the same way as total cases per million, supporting the ‘biasing autocracy’ view. As there is a positive correlation between government transparency and total deaths per million, this implies that countries with higher tendencies to disclose reliable information also tend to report higher cumulative deaths per million, similar to that of the findings of this paper and total cases per million.

Following on, population aged 70 and older and life expectancy are both age variables detailing an older age demographic. It has been numerously evidenced that democratic countries have an older demographic in comparison the autocratic countries, and therefore the correlation of older age demographic variables correlating with government transparency could argue against the ‘biasing autocracy’ view, with the significant correlation between level of democracy and total cases per million being explained by the older demographic of democracies. (Cassan & Van Steenvoort, 2021) reported similar results with a correlation between government transparency and the variable population ratio over age 65; the study modelled deaths per million when controlling for age and found a reduction in significance between MPI and total deaths per million. Whilst this is the case, the study also notes that the democratic/autocratic nature of a country is far more noticable than age demographic, providing evidence that government transparency remains significant, whilst age becomes insignificant.

Therefore, whilst this essay shows a correlation with variables other than government transparency (age demographic), it is likely that this correlation does not negate the relationship of transparency with effects of level of democracy on total cases per million.

## Does data transparency have a significant effect on the level of accuracy in predicting future COVID-19 cases?

Figure 8 shows a scattergraph of the relationship between government transparency and normalised RMSE. The scatter graph results a Pearson correlation coefficient (R) value of -0.058 and a p value < 0.52. These results conclude that there is no significant correlation between government transparency and the accuracy of a model in predicting COVID-19 cases per million.

Chart, scatter chart

Description automatically generated

*Figure 8: A scatter graph of the government transparency against nRMSE, with a line of best shown in orange.*

When linking to the previous analyses, the lack of significance could suggest that whilst there could be a relationship between accurate reporting of COVID-19 data and government regime, there is no significance of this found in the accuracy in forecasting future cases.

## Limitations

One limitation of the study was that HRV Transparency Project government transparency data was recorded in 2010. Therefore, analysis in this paper of government transparency does not account for changes in transparency since 2010.

Furthermore, another limitation of the study was in plotting the correlation between government transparency with other variables. This study only selected a handful of other variables of interest, as visualisations of results with many other COVID-19 variables of interest became unclear and, in many cases, produced errors rendering the correlation plot. Therefore, other significant correlations could have been missed due to the variables having not been included in the correlation analysis.

Additionally, another limitation of the study was that only 1 model (ARIMA) was used in the forecasting of total covid cases per million. As a result, no analysis was carried out into how good the ARIMA model was in predicting future total cases per million. An area for further research related to the time-series forecasting analysis conducted in this essay could be in comparing different models of predicting total cases per million and analysing if different models perform better than others, and if so, re-evaluating the correlation between government transparency and nRMSE with the best model.

# **Conclusion**

In conclusion, when linking back to the 4 aims of this essay, a significant positive correlation was found between both political regime and government transparency, and COVID-19 cases per million. There were also three other COVID-19 variables (life expectancy, total deaths per million, and population aged 70 and older) that had a significant correlation with that of government transparency. Finally, it was found that data transparency did not have a significant effect on the level of accuracy in predicting future COVID-19 cases.

# **References**

Albani, V., Loria, J., Massad, E., & Zubelli, J. (2021). COVID-19 underreporting and its impact on vaccination strategies. *BMC Infectious Diseases*, *21*(1). https://doi.org/10.1186/s12879-021-06780-7

Annaka, S. (2021). Political regime, data transparency, and COVID-19 death cases. *SSM - Population Health*, *15*, 100832. https://doi.org/10.1016/j.ssmph.2021.100832

Bollyky, T. J., Templin, T., Cohen, M., Schoder, D., Dieleman, J. L., & Wigley, S. (2019). The relationships between democratic experience, adult health, and cause-specific mortality in 170 countries between 1980 and 2016: an observational analysis. *The Lancet*, *393*(10181), 1628–1640. https://doi.org/10.1016/s0140-6736(19)30235-1

Cassan, G., & Van Steenvoort, M. (2021). Political regime and COVID 19 death rate: Efficient, biasing or simply different autocracies?An econometric analysis. *SSM - Population Health*, *16*, 100912. https://doi.org/10.1016/j.ssmph.2021.100912

Kapoor, M., Malani, A., Ravi, S., & Agrawal, A. (2020). *Authoritarian Governments Appear to Manipulate COVID Data*. https://doi.org/10.48550/arXiv.2007.09566

Kilani, A. (2021). Authoritarian regimes’ propensity to manipulate Covid-19 data: a statistical analysis using Benford’s Law. *Commonwealth & Comparative Politics*, 1–15. https://doi.org/10.1080/14662043.2021.1916207

Kudamatsu, M. (2012). Has Democratization Reduced Infant Mortality in Sub-Saharan Africa? Evidence from Micro Data. *Journal of the European Economic Association*, *10*(6), 1294–1317. https://doi.org/10.1111/j.1542-4774.2012.01092.x

Norheim, O. F., Abi-Rached, J. M., Bright, L. K., Bærøe, K., Ferraz, O. L. M., Gloppen, S., & Voorhoeve, A. (2020). Difficult trade-offs in response to COVID-19: the case for open and inclusive decision making. *Nature Medicine*, *27*(1), 10–13. https://doi.org/10.1038/s41591-020-01204-6

Pan American Health Organization. (2020). *WHY PREDICTIVE MODELING IS CRITICAL IN THE FIGHT AGAINST COVID-19?* https://iris.paho.org/bitstream/handle/10665.2/52276/PAHOEIHISCOVID-19200007\_eng.pdf?sequence=8

Preston, S. H. (1975). The Changing Relation between Mortality and level of Economic Development. *Population Studies*, *29*(2), 231–248. https://doi.org/10.1080/00324728.1975.10410201

Sorci, G., Faivre, B., & Morand, S. (2020). Explaining among-country variation in COVID-19 case fatality rate. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-75848-2

Wang, Y., Yan, Z., Wang, D., Yang, M., Li, Z., Gong, X., Wu, D., Zhai, L., Zhang, W., & Wang, Y. (2022). Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. *BMC Infectious Diseases*, *22*(1). https://doi.org/10.1186/s12879-022-07472-6

Wigley, S. (2022). Regime type and Data Manipulation: Evidence from the COVID-19 Pandemic. *COVID-19 SARS-CoV-2 Preprints from MedRxiv and BioRxiv*. https://doi.org/10.1101/2022.12.11.22283310

WorldOMeter. (2022). *Coronavirus toll update: Cases & deaths by country*. Worldometers. https://www.worldometers.info/coronavirus/

# **Appendix**

## Appendix Table 1: A table of the statistical analyses of each numerical variable. The statistics calculated and outputted are min, median, max, mean, standard deviation, % blanks, and coefficient of skewness.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **min** | **median** | **max** | **mean** | **sd** | **percentblank** | **coef\_skewness** |
| aged\_65\_older\_ | 1.144 | 6.704 | 27.049 | 8.82899547805612 | 6.14976950003889 | 13.7177246513915 | 3.21686632873691 |
| aged\_70\_older\_ | 0.526 | 4.032 | 18.493 | 5.57006188120339 | 4.17691213111354 | 13.2495283296188 | 3.03530101798676 |
| cardiovasc\_death\_rate\_ | 79.37 | 243.964 | 724.417 | 261.013032116968 | 120.210631066277 | 12.9988615774368 | 4.48442114952123 |
| diabetes\_prevalence\_ | 0.99 | 7.2 | 30.53 | 8.35970542599351 | 4.69791072207456 | 8.7182858456128 | 3.80575905667387 |
| electoral\_democracy\_ | 0.016 | 0.522 | 0.908 | 0.513991471401935 | 0.255935933698279 | 20.5995692166263 | 3.98527240574294 |
| excess\_mortality\_ | -95.92 | 7.515 | 375 | 15.5439362336114 | 28.8782530993813 | 96.4126327491569 | 1.3545420689478 |
| excess\_mortality\_cumulative\_ | -28.45 | 6.78 | 111.01 | 9.908873659118 | 15.976705671302 | 96.4126327491569 | 1.4362548481175 |
| excess\_mortality\_cumulative\_absolute\_ | -37726.1 | 4599.3 | 1217716.1 | 42143.982479142 | 116604.189918283 | 96.4126327491569 | 1.04483936231457 |
| excess\_mortality\_cumulative\_per\_million\_ | -1826.595723 | 612.67140955 | 9725.192865 | 1161.87485479737 | 1606.46308246552 | 96.4126327491569 | 1.78837172556299 |
| extreme\_poverty\_ | 0.1 | 2.2 | 77.6 | 13.6464992129865 | 20.1121677006511 | 43.6336524123334 | 1.92617216679759 |
| female\_smokers\_ | 0.1 | 6.3 | 44 | 10.6701511386321 | 10.6094264113787 | 34.4017402365567 | 2.42335941821713 |
| gdp\_per\_capita\_ | 661.24 | 12951.839 | 116935.6 | 19645.4019130779 | 20627.1805135817 | 13.3751289410532 | 2.22930936726693 |
| gov\_transparency\_ | -6.67581885954938 | 0.923360329931058 | 5.63572688074547 | 0.96318918993017 | 1.98145371847007 | 42.7378795409966 | 0.992305407656762 |
| handwashing\_facilities\_ | 1.188 | 49.542 | 100 | 50.7968181738071 | 32.0156198233078 | 57.5453899230897 | 3.21244614625722 |
| hosp\_patients\_ | 0 | 756 | 154540 | 4162.56068208192 | 11189.0331353277 | 85.2999182259849 | 1.04849828437854 |
| hosp\_patients\_per\_million\_ | 0 | 89.586 | 1544.082 | 164.128529741129 | 201.143656909391 | 85.2999182259849 | 2.00254681361806 |
| hospital\_beds\_per\_thousand\_ | 0.1 | 2.4 | 13.8 | 3.08706853822073 | 2.55965877165951 | 22.8373979829076 | 2.68051573538993 |
| human\_development\_index\_ | 0.394 | 0.743 | 0.957 | 0.72503389359633 | 0.150138154638689 | 15.3516015414135 | 9.53855923056583 |
| icu\_patients\_ | 0 | 157 | 28891 | 859.409533592097 | 2544.95009797707 | 85.9509035226963 | 0.951385491880913 |
| icu\_patients\_per\_million\_ | 0 | 12.4875 | 177.282 | 22.8929733698547 | 27.085725225904 | 85.9509035226963 | 2.07457690871885 |
| level\_of\_democracy\_ | 0.016 | 0.526 | 0.908 | 0.51470736438209 | 0.256649347931601 | 24.2243494155563 | 3.96697712794348 |
| life\_expectancy\_ | 53.28 | 75.05 | 86.75 | 73.6499547242071 | 7.46176213277428 | 1.19400751465786 | 19.5530039120097 |
| male\_smokers\_ | 7.7 | 31.4 | 78.1 | 32.780863350958 | 13.5828401512343 | 35.3076680509457 | 4.9284677804878 |
| median\_age\_ | 15.1 | 30.6 | 48.2 | 30.6503321689036 | 9.08931476199895 | 12.7904180095243 | 6.74979336871576 |
| new\_cases\_ | 0 | 51 | 1383887 | 3064.35301616399 | 18403.7998186228 | 4.40831422600627 | 0.49674845078684 |
| new\_cases\_per\_million\_ | 0 | 8.21 | 208049.887 | 187.692256646532 | 964.899974736938 | 4.40831422600627 | 0.575051077279662 |
| new\_cases\_smoothed\_ | 0 | 78.571 | 807800.714 | 3066.17325969265 | 17357.965486814 | 5.00264562990043 | 0.525404246598251 |
| new\_cases\_smoothed\_per\_million\_ | 0 | 17.508 | 37617.428 | 187.629523463913 | 624.058873731526 | 5.00264562990043 | 0.873924870470421 |
| new\_deaths\_ | 0 | 1 | 4529 | 39.2319357049392 | 180.415179279464 | 14.212110036825 | 0.646818120187412 |
| new\_deaths\_per\_million\_ | 0 | 0.043 | 550.399 | 1.54643329387582 | 5.44870023706914 | 14.212110036825 | 0.843558955649176 |
| new\_deaths\_smoothed\_ | 0 | 1.429 | 4190 | 39.4286492594874 | 170.563591910609 | 14.796286497667 | 0.685122460599363 |
| new\_deaths\_smoothed\_per\_million\_ | 0 | 0.22 | 144.167 | 1.54685071855571 | 3.58613599312928 | 14.796286497667 | 1.23267833794828 |
| new\_people\_vaccinated\_smoothed\_ | 0 | 2105 | 6785334 | 45226.4323442447 | 265563.943300157 | 45.6213488971197 | 0.502983557830967 |
| new\_people\_vaccinated\_smoothed\_per\_hundred\_ | 0 | 0.048 | 11.75 | 0.122526906027933 | 0.230017626350796 | 45.6213488971197 | 1.38937490641006 |
| new\_tests\_ | 1 | 8783 | 35855632 | 67283.9098155117 | 247736.301840298 | 59.7003757328929 | 0.779331603855925 |
| new\_tests\_per\_thousand\_ | 0 | 0.969 | 534.013 | 3.23960218034244 | 8.96587095699105 | 59.7003757328929 | 0.975901458207442 |
| new\_tests\_smoothed\_ | 0 | 6570 | 14769984 | 142176.499417976 | 1138225.04934794 | 44.4348239720793 | 0.36895998598389 |
| new\_tests\_smoothed\_per\_thousand\_ | 0 | 0.858 | 147.603 | 2.80124436578399 | 7.26067528262056 | 44.4348239720793 | 1.03926051002635 |
| new\_vaccinations\_ | 0 | 23112 | 24741000 | 262564.747268839 | 1172193.69796725 | 79.5500825757211 | 0.652266125583519 |
| new\_vaccinations\_smoothed\_ | 0 | 5935 | 22424286 | 117522.294321636 | 709777.490986613 | 45.1258945703123 | 0.488366970447426 |
| new\_vaccinations\_smoothed\_per\_million\_ | 0 | 1556 | 117497 | 2857.4796142982 | 3831.52851912211 | 45.1258945703123 | 1.8312375355886 |
| people\_fully\_vaccinated\_ | 1 | 2212826 | 1260501000 | 18404305.4633049 | 72699754.4561602 | 76.9220902079625 | 0.729027089381369 |
| people\_fully\_vaccinated\_per\_hundred\_ | 0 | 39.77 | 122.94 | 39.3427291044258 | 29.1405305356488 | 76.9220902079625 | 2.68554435608304 |
| people\_vaccinated\_ | 0 | 2678773 | 1294045000 | 21965555.7148327 | 85653136.8882083 | 75.6874629210961 | 0.738068638712065 |
| people\_vaccinated\_per\_hundred\_ | 0 | 47.5 | 124.88 | 44.3089958011827 | 29.9598510139494 | 75.6874629210961 | 2.85138225032471 |
| political\_freedom\_ | 0 | 1 | 2 | 1.12868509826929 | 0.828341227174559 | 12.543492552151 | 2.88052220091303 |
| population\_ | 47 | 6769151 | 1444216102 | 36824863.016207 | 143129832.139771 | 0.16942720776479 | 0.724555017624482 |
| population\_density\_ | 0.137 | 88.125 | 20546.766 | 459.664269328865 | 2115.41520687496 | 5.88131543925473 | 0.610219593671899 |
| positive\_rate\_ | 0 | 0.055 | 1 | 0.0977101188490408 | 0.115152770921727 | 48.7335717072597 | 2.06795159717857 |
| reproduction\_rate\_ | -0.1 | 0.98 | 6.12 | 0.968483397255085 | 0.376117304837413 | 20.985457052608 | 5.1192810514198 |
| stringency\_index\_ | 0 | 52.31 | 100 | 52.1760436836503 | 21.1761946816388 | 17.6337913747121 | 4.92147586560087 |
| tests\_per\_case\_ | 1 | 17.5 | 1023631.9 | 2397.00302047636 | 33494.7485141674 | 49.7287561263703 | 0.214168172016424 |
| total\_boosters\_ | 1 | 1006780 | 790025000 | 8188608.31371083 | 27680407.9716103 | 88.7419094499762 | 0.851109021416707 |
| total\_boosters\_per\_hundred\_ | 0 | 13.55 | 127.44 | 22.6311260919104 | 23.2536606560614 | 88.7419094499762 | 2.33698165116922 |
| total\_cases\_ | 1 | 31115 | 87843561 | 893296.371500257 | 4172920.63334096 | 4.23568019411975 | 0.634753053613696 |
| total\_cases\_per\_million\_ | 0.001 | 7007.328 | 706541.904 | 45890.3435627429 | 85368.9003217078 | 4.23568019411975 | 1.53057732026335 |
| total\_deaths\_ | 1 | 796 | 1017848 | 17132.1081272702 | 67363.908532634 | 14.081164718521 | 0.75114887903658 |
| total\_deaths\_per\_million\_ | 0.001 | 168.475 | 6401.521 | 608.899906049579 | 903.420117206414 | 14.081164718521 | 1.83549678224607 |
| total\_tests\_ | 0 | 2067330 | 9.214e+09 | 21093607.5492437 | 84040153.5248413 | 57.5710445160635 | 0.728383874615807 |
| total\_tests\_per\_thousand\_ | 0 | 234.256 | 32925.9 | 916.217757712415 | 2168.24469330861 | 57.5710445160635 | 1.15964645544707 |
| total\_vaccinations\_ | 0 | 4530783 | 3403643000 | 59790248.1492167 | 274428098.849554 | 74.3063906660039 | 0.637106630773622 |
| total\_vaccinations\_per\_hundred\_ | 0 | 80.93 | 355.75 | 91.0834243754291 | 74.1369913619833 | 74.3063906660039 | 2.59412028453191 |
| weekly\_hosp\_admissions\_ | 0 | 1326 | 153995 | 5613.69290664228 | 13759.4684896208 | 93.1357929674347 | 1.12759288134061 |
| weekly\_hosp\_admissions\_per\_million\_ | 0 | 71.359 | 645.808 | 99.4983459472086 | 101.864997954217 | 93.1357929674347 | 2.22977511808041 |
| weekly\_icu\_admissions\_ | 0 | 205.5 | 4838 | 444.182805710737 | 603.213098853477 | 96.5558708932609 | 1.86840839377392 |
| weekly\_icu\_admissions\_per\_million\_ | 0 | 9.149 | 222.9 | 14.0671440099317 | 15.7738012157434 | 96.5558708932609 | 2.09540056817798 |

## Appendix Figure 1: A histogram and pattern plot of 8 variables detailing plot percentages of missing data and location of missing data using a histogram and pattern plot respectively (left and right). Missing data is shown in red, present data is shown in blue.

Chart, bar chart

Description automatically generated

## Appendix Figure 2: A matrix plot looking in detail at locations of missing data for all variables of interest. Missing data shown in red, values of data are shown by a gradient of white to black (largest number in variable = black).

Chart, surface chart

Description automatically generated

## Appendix Figure 3: A missing values histogram visualising overlapping levels of missing data between variables.

Chart, bar chart, histogram

Description automatically generated

## Appendix Figure 4: A margin plot showing the distribution of missing and present data for level\_of\_democracy vs gov\_transparency to assess the missing at randomness of the variables.

Chart, scatter chart

Description automatically generated

## Appendix Table 2: A table of the RMSE and nRMSE values of each country.

|  |  |  |
| --- | --- | --- |
| **location** | **rmse** | **nrmse** |
| Afghanistan | 395.32 | 0.63 |
| Albania | 12869.99 | 0.49 |
| Algeria | 370.09 | 0.33 |
| Angola | 2022.00 | 2.21 |
| Argentina | 41907.81 | 0.48 |
| Australia | 104412.28 | 0.34 |
| Austria | 257592.22 | 0.72 |
| Bangladesh | 1598.26 | 0.67 |
| Belgium | 89054.18 | 0.47 |
| Benin | 140.34 | 0.77 |
| Bolivia | 10440.18 | 0.35 |
| Botswana | 13288.76 | 0.28 |
| Brazil | 32971.44 | 0.69 |
| Bulgaria | 30440.15 | 0.48 |
| Burkina Faso | 162.06 | 0.87 |
| Burundi | 2820.19 | 1.82 |
| Cambodia | 711.12 | 0.76 |
| Cameroon | 120.48 | 0.28 |
| Canada | 337272.71 | 6.54 |
| Central African Republic | 158.07 | 0.29 |
| Chad | 43.22 | 0.42 |
| Chile | 68421.38 | 0.59 |
| China | 323.19 | 0.60 |
| Colombia | 12489.27 | 0.61 |
| Congo | 304.81 | 0.37 |
| Costa Rica | 47114.29 | 0.72 |
| Cote d'Ivoire | 237.93 | 0.32 |
| Cuba | 9238.21 | 0.74 |
| Cyprus | 141240.36 | 0.35 |
| Democratic Republic of Congo | 1316.74 | 5.70 |
| Denmark | 113071.72 | 0.26 |
| Dominican Republic | 11630.60 | 0.65 |
| Ecuador | 12906.78 | 0.63 |
| Egypt | 265.19 | 0.20 |
| El Salvador | 3369.12 | 0.46 |
| Eswatini | 30857.17 | 3.94 |
| Ethiopia | 3282.87 | 3.92 |
| Fiji | 10068.86 | 0.70 |
| Finland | 35434.68 | 0.22 |
| France | 143612.43 | 0.44 |
| Gabon | 1765.21 | 0.41 |
| Gambia | 296.90 | 0.39 |
| Ghana | 323.92 | 0.32 |
| Greece | 113696.37 | 0.44 |
| Guatemala | 9559.54 | 0.59 |
| Guinea | 173.86 | 0.38 |
| Guinea-Bissau | 381.84 | 0.40 |
| Guyana | 24051.66 | 0.67 |
| Haiti | 148.70 | 0.30 |
| Honduras | 4405.74 | 0.90 |
| Hungary | 37677.24 | 0.53 |
| India | 5207.00 | 0.83 |
| Indonesia | 5197.13 | 0.78 |
| Iran | 12245.33 | 0.98 |
| Iraq | 4945.70 | 0.77 |
| Ireland | 40048.04 | 0.22 |
| Israel | 234671.11 | 0.72 |
| Italy | 53120.24 | 0.24 |
| Jamaica | 9919.07 | 0.58 |
| Japan | 39256.42 | 0.65 |
| Jordan | 72052.60 | 1.14 |
| Kenya | 31544.79 | 30.62 |
| Kuwait | 39716.81 | 0.75 |
| Laos | 1609.49 | 0.11 |
| Lebanon | 26398.06 | 0.44 |
| Lesotho | 1303.31 | 0.44 |
| Liberia | 97.88 | 0.32 |
| Libya | 7310.95 | 0.43 |
| Madagascar | 179.83 | 0.28 |
| Malawi | 10259.65 | 11.97 |
| Malaysia | 33668.98 | 0.60 |
| Mali | 322.31 | 0.59 |
| Mauritania | 2252.47 | 0.54 |
| Mauritius | 79480.94 | 0.62 |
| Mexico | 10997.43 | 0.67 |
| Mongolia | 43273.05 | 0.60 |
| Morocco | 3540.34 | 0.49 |
| Mozambique | 12895.63 | 7.14 |
| Nepal | 3973.41 | 0.77 |
| Netherlands | 168575.63 | 0.56 |
| New Zealand | 147800.77 | 0.56 |
| Nicaragua | 18.22 | 0.11 |
| Niger | 23.77 | 0.34 |
| Nigeria | 346.48 | 3.19 |
| Norway | 123229.76 | 0.62 |
| Oman | 13045.51 | 0.80 |
| Pakistan | 768.94 | 0.71 |
| Panama | 44026.27 | 0.44 |
| Papua New Guinea | 537.34 | 0.57 |
| Paraguay | 18865.96 | 0.70 |
| Peru | 29235.02 | 0.71 |
| Philippines | 7267.18 | 0.93 |
| Poland | 21787.25 | 0.42 |
| Portugal | 72797.74 | 0.19 |
| Romania | 45536.82 | 0.78 |
| Russia | 31746.80 | 0.58 |
| Rwanda | 1121.12 | 0.54 |
| Saudi Arabia | 4558.22 | 0.66 |
| Senegal | 418.75 | 0.60 |
| Sierra Leone | 87.54 | 0.68 |
| Singapore | 145319.06 | 0.67 |
| Somalia | 181.82 | 0.91 |
| South Africa | 35999.33 | 3.59 |
| South Korea | 233346.69 | 0.67 |
| Spain | 116292.52 | 0.77 |
| Sri Lanka | 1428.23 | 0.38 |
| Sudan | 155.26 | 0.42 |
| Sweden | 61801.62 | 0.50 |
| Switzerland | 119178.94 | 0.41 |
| Syria | 82.78 | 0.26 |
| Tanzania | 118.95 | 0.79 |
| Thailand | 22009.32 | 0.66 |
| Togo | 2096.73 | 1.77 |
| Trinidad and Tobago | 12254.01 | 0.22 |
| Tunisia | 20371.49 | 0.74 |
| Turkey | 32913.10 | 0.48 |
| Uganda | 371.02 | 0.49 |
| United Arab Emirates | 3900.23 | 0.19 |
| United Kingdom | 158919.76 | 0.99 |
| United States | 39115.17 | 0.36 |
| Uruguay | 108143.66 | 0.68 |
| Venezuela | 1170.77 | 0.40 |
| Vietnam | 50291.32 | 0.54 |
| Zambia | 30740.80 | 5.98 |
| Zimbabwe | 14551.14 | 4.16 |

# **R Code**

#Sets the working directory to the correct folder

setwd("/Users/elysia/Sheffield Uni Work/Intro to Data Science")

#Packages:

library(tidyverse) #loads tidyverse package involved in data exploration, manipulation, and plotting

library(lubridate) #loads lubridate package involved in dealing with date data

#Loads 2 different packages that are used in plotting missing data visualisations, both for different purposes and analyses

library(VIM) #^

library(naniar) #^

library(zoo) #loads zoo package, used linear interpolation of linear time series data (used in NA handling)

library(ggpubr) #loads ggpubr package used for calculation of correlation coefficient and statistical significance for ggplot graphs

library(corrplot) #loads corrplot package used in plotting correlation plots

library(forecast) #loads forecast package used in time series forecasting

library(tsibble) #loads tsibble package used to analyse time series analysis of groups (allows for data frames using temporal structure)

library(fable) #loads fable package used in time series forecasting

library(feasts) #loads feasts package used in analysing tidy time series data

library(Metrics) #loads the Metrics package used in calculating RMSE values for predicted vs actual data

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 1: Setting up the data

##

#Imports the covid csv file from the data folder into a variable called covid\_dataset

covid\_dataset <- read.csv("data/owid-covid-data.csv")

#Imports 5 other csv files from the data file that will be analysed alongside the covid dataset

political\_regime\_dataset <- read.csv("data/V-Dem-v12.csv") #v2x\_polyarchy

transparency\_dataset <- read.csv("data/government-transparency.csv")

democracy\_dataset <- read.csv("data/democracy.csv")

freedom\_dataset <- read.csv("data/democracy-freedom.csv")

humanrights\_dataset <- read.csv("data/human-rights-vdem.csv")

#Views each of the initial datasets

View(covid\_dataset)

View(political\_regime\_dataset)

View(transparency\_dataset)

View(democracy\_dataset)

View(freedom\_dataset)

View(humanrights\_dataset)

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 2: Cleaning the initial datasets and combining them

##

#Cleans up the political regime dataset - ,

political\_regime\_cleaned <- political\_regime\_dataset %>%

select(country\_name, year, v2x\_polyarchy) %>% #selecting only relevant variables

filter(year == "2021") %>% #filters year by 2021 to remove duplicate instances, as level\_of\_democracy is not date dependent

mutate(year = NULL) %>%

rename("level\_of\_democracy" = "v2x\_polyarchy") #renames the unclear variable v2x\_polyarchy

#Cleans up the goverment transparency dataset

transparency\_dataset\_cleaned <- transparency\_dataset %>%

select(Entity, Year, gov\_transparency) %>%

filter(Year == "2010") %>%

mutate(Year = NULL)

#Cleans up the democracy dataset

democracy\_dataset\_cleaned <- democracy\_dataset %>%

select(Entity, Year, electdem\_vdem\_owid) %>%

filter(Year == "2021") %>%

mutate(Year = NULL) %>%

rename("electoral\_democracy" = "electdem\_vdem\_owid")

#Cleans up the freedom dataset

freedom\_dataset\_cleaned <- freedom\_dataset %>%

select(Entity, Year, regime\_fh) %>%

filter(Year == "2021") %>%

mutate(Year = NULL) %>%

rename("political\_freedom" = "regime\_fh")

#Views the cleaned datasets

View(political\_regime\_cleaned)

View(transparency\_dataset\_cleaned)

View(democracy\_dataset\_cleaned)

View(freedom\_dataset\_cleaned)

#Creates a variable that contains all of the non-country location strings to be removed

non\_country\_locations <- c("Africa", "Asia", "Europe", "North America", "South America", "European Union", "High income",

"International", "Low income", "Lower middle income", "Oceania", "Upper middle income", "World")

#Joins the cleaned datasets to the covid dataset, named covid\_dataset\_complete

covid\_dataset\_complete <- covid\_dataset %>%

mutate(date = as.Date(date, "%Y-%m-%d")) %>% #converts the date column into a date format

mutate(year = year(date), month = month(date), day = day(date)) %>% #splits each section of the date (year, month, day) into their own columns

#Joins the cleaned datasets to the covid dataset by the country and date/year (depending on the dataset)

left\_join(political\_regime\_cleaned, by = c("location" = "country\_name")) %>%

left\_join(transparency\_dataset\_cleaned, by = c("location" = "Entity")) %>%

left\_join(democracy\_dataset\_cleaned, by = c("location" = "Entity")) %>%

left\_join(freedom\_dataset\_cleaned, by = c("location" = "Entity")) %>%

filter(!location %in% non\_country\_locations) #filters out locations that are not countries

View(covid\_dataset\_complete) #views the complete dataset

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 3: Initial analysis of dataset - looking at different statistics and levels of missing values

##

ncol(covid\_dataset\_complete) #counts the amount of columns in the complete dataset #output: 74

covid\_nrow <- nrow(covid\_dataset\_complete) #counts the amount of rows in the complete dataset #output in variable: 187101

#Calculates min, median, max, mean, sd, and sum of blanks for each variable and stores it in covid\_dataset\_complete.sum

covid\_dataset\_complete.sum <- covid\_dataset\_complete %>%

select(-day, -year, -month) %>% #removes day, year and month columns from the dataset

summarise(across(where(is.numeric),

list(

ZZmin = min,

ZZmedian = median,

ZZmax = max,

ZZmean = mean,

ZZsd = sd,

ZZblank = ~ sum(is.na(.))

), na.rm=TRUE))

#Tidies the 1 row .sum table into a cleaner table split by variable and statistics (as listed above)

covid\_dataset\_complete.stats.tidy <- covid\_dataset\_complete.sum %>%

gather(stat, val) %>%

separate(stat, into = c("var", "stat"), sep = "ZZ") %>% ##uses ZZ character as a separator as underscores cannot be used due to variable names

spread(stat, val) %>%

select(var, min, median, max, mean, sd, blank) %>%

mutate(percentblank = blank / covid\_nrow \* 100) %>% #calculates the % of missing data

##Creates variable called coef\_skew that measures the skewness of each variable. Values >0.5/<-0.5 are significantly skewed

mutate(coef\_skewness = (3\* mean - median) / sd)

View(covid\_dataset\_complete.stats.tidy) #views the stats table that shows the stats for each variable

#Writes covid\_dataset\_complete.stats.tidy to a text document for easy input to word doc table

write.table(covid\_dataset\_complete.stats.tidy, file = "olstab.txt", sep = ",", quote = FALSE, row.names = F)

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 4: Sorting out missing data for variables of interest - mean for skew <0.5, regression for skew >0.5

##

#Creates a subset dataset from the complete dataset that only contains the variables of interest

covid\_dataset\_interest <- covid\_dataset\_complete %>%

select(

location, date, year, population, total\_cases, total\_cases\_per\_million, level\_of\_democracy,

gov\_transparency, political\_freedom, electoral\_democracy, gdp\_per\_capita, life\_expectancy,

population\_density, stringency\_index, total\_deaths\_per\_million, aged\_70\_older) %>%

mutate(political\_freedom = as.factor(political\_freedom)) %>% #converts political\_freedom variable to a factor (0,1,2)

filter(!is.na(population)) #removes rows that have no population recorded - filters out 0.17% of the dataset

View(covid\_dataset\_interest) #views the dataset of containing only variables of interest

#Visualisations of missing data for analysis of missing data:

#Creates subset dataset including only relevant variables for missing data visualisations

covid\_dataset\_interest\_NA\_vis <- covid\_dataset\_interest %>%

select(location, date, year, population, total\_cases, total\_cases\_per\_million, level\_of\_democracy, gov\_transparency)

#Creates a histogram and pattern plot looking into missing data

aggr\_plot <- aggr(covid\_dataset\_interest\_NA\_vis, col=c('lightblue','red'),

sortVars=TRUE, labels=names(covid\_dataset\_interest),

cex.axis=.4, gap=3, ylab=c("Histogram of proportion of missing data","Combination"))

#Creates a matrix plot looking at the positions of missing data in the dataset against each variable

matrixplot(covid\_dataset\_interest\_NA\_vis, sortby = 2, ylab = "Observation", cex.axis = 0.4)

#Create missing values plot detailing levels of missing data that overlap with different variables

gg\_miss\_upset(covid\_dataset\_interest\_NA\_vis, nsets = n\_var\_miss(covid\_dataset\_interest))

#Creates a margin plot plotting distribution of data and missing data of government transparency and level ofdemocracy

marginplot(covid\_dataset\_interest\_NA\_vis[ ,c("gov\_transparency","level\_of\_democracy")])

###Imputations:

###1. NA: total\_cases

#Groups dataset by country, and replaces missing values of total\_cases with a linear interpolation (in order of date)

covid\_dataset\_interest <- covid\_dataset\_interest %>%

group\_by(location) %>% #groups the data by country

arrange(date) %>% #orders data by date

#Creates a temporary variable and uses the zoo package na.approx to linearly approximate total case values per country

#with missing data removed, and always fill in the blanks (rule 2)

mutate(total\_cases2 = na.approx(total\_cases, na.rm=F, rule=2)) %>%

ungroup() %>% #ungroups the data for further manipulation

mutate(total\_cases = ifelse(is.na(total\_cases), total\_cases2, total\_cases)) %>% #replaces NA values with the approximate value

mutate(total\_cases2 = NULL) %>% #removes the temporary variable from the dataset

filter(!is.na(total\_cases)) #removes any countries wherein there are not enough data for total\_cases to make an approximation

sum(is.na(covid\_dataset\_interest$total\_cases)) #checks for the amount of missing values in the total\_cases column

#As total\_cases is cumulative, can't take an overall mean as this could mean previous values in time could be larger than later values

###2. NA: total\_cases\_per\_million

#Calculates total\_cases\_per\_million from the imputed total\_cases values and replaces NAs

covid\_dataset\_interest <- covid\_dataset\_interest %>%

mutate(total\_cases\_per\_million = ifelse(is.na(total\_cases\_per\_million), (population / 1000000 \* total\_cases), total\_cases\_per\_million))

sum(is.na(covid\_dataset\_interest$total\_cases\_per\_million)) #checks for the amount of missing values in the total\_cases\_per\_million column

###3. NA: level\_of\_democracy

#Creates a new subset dataset looking at electoral democracy and level of democracy correlation

covid\_dataset\_interest\_elect <- covid\_dataset\_interest %>%

filter(!is.na(electoral\_democracy)) %>% #filters out NA values

filter(!is.na(level\_of\_democracy)) %>% #filters out NA values

#Filters by specific date so that 1 point for each country is recorded (both variables are date independent so the date is not relevant)

filter(date == "2021-12-31")

#Plots a graph of electoral democracy vs level of freedom

ggplot(covid\_dataset\_interest\_elect, aes(x = electoral\_democracy, y = level\_of\_democracy)) +

geom\_point(size = 2, alpha = 0.4, colour = 'blue') +

labs(

x="Electoral Democracy",

y="Level of Democracy",

title='Electoral Democracy vs Level of Democracy',

caption='Source: Our World in Data and V-Dem Project',

tag='A')

#shows 1 to 1 correlation between variables

#no values of level\_of\_democracy for some (e.g. United States) but there is data for electoral\_democracy

#Creates a new subset dataset looking at political freedom and level of democracy correlation

covid\_dataset\_interest\_freedom <- covid\_dataset\_interest %>%

filter(!is.na(political\_freedom)) %>%

filter(!is.na(level\_of\_democracy)) %>%

filter(date == "2021-12-31") %>%

mutate(political\_freedom = as.factor(political\_freedom)) #converts political\_freedom variable to a factor (0,1,2)

#Plot of political freedom against level of democracy

ggplot(covid\_dataset\_interest\_freedom, aes(x = political\_freedom, y = level\_of\_democracy)) +

geom\_jitter(size = 2, alpha = 0.4, colour = 'blue', width = 0.1) + #adds jitter to points as plotting categorical data

scale\_x\_discrete(labels = c('Not Free','Partly Free','Free')) +

labs(

x="Political Freedom",

y="Level of Democracy",

title='Political Freedom vs Level of Democracy',

caption='Source: Our World in Data and V-Dem Project',

tag='B'

)

#Imputes level\_of\_democracy variable by (in order):

#1. electoral democracy values,

#2. mean values by level of freedom groups,

#3. the mean of the level of democracy

covid\_dataset\_interest <- covid\_dataset\_interest %>%

#Replaces any missing values in level\_of\_democracy with the electoral\_democracy value

mutate(level\_of\_democracy = ifelse(is.na(level\_of\_democracy), electoral\_democracy, level\_of\_democracy)) %>%

#Calculates the mean for level\_of\_democracy and creates a new variable to store it

mutate(level\_of\_democracy\_mean = mean(level\_of\_democracy, na.rm=TRUE)) %>%

#Calculates the mean for level\_of\_democracy for each category of political freedom and stores it in a new variable

group\_by(political\_freedom) %>%

mutate(level\_of\_democracy\_mean\_freedom = mean(level\_of\_democracy, na.rm=TRUE)) %>%

#If the political freedom row had no value, then replaces the mean\_freedom value with the overall mean

mutate(level\_of\_democracy\_mean\_freedom = ifelse(is.na(political\_freedom), level\_of\_democracy\_mean, level\_of\_democracy\_mean\_freedom)) %>%

#Replaces any missing values with the mean values calculated (either of groups of political freedom, or overall if political freedom NA)

mutate(level\_of\_democracy = ifelse(is.na(level\_of\_democracy), level\_of\_democracy\_mean\_freedom, level\_of\_democracy)) %>%

ungroup() %>% #ungroups the dataset

mutate(

level\_of\_democracy\_mean = NULL,

level\_of\_democracy\_mean\_freedom = NULL,

electoral\_democracy = NULL,

political\_freedom = NULL) #removes the temporary variables used to store means

sum(is.na(covid\_dataset\_interest$level\_of\_democracy)) #checks for the amount of missing values in the level\_of\_democracy column

###4. NA: government transparency

sum(is.na(covid\_dataset\_interest$gov\_transparency)) #checks for the amount of missing values in the gov\_transparency column

covid\_dataset\_interest <- covid\_dataset\_interest %>%

filter(!is.na(gov\_transparency)) #removes countries with no transparency data

###5.

summary(covid\_dataset\_interest) #shows statistics for each of the variables in the dataset

View(covid\_dataset\_interest)

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 5: Analysis of dataset: correlations

##

#Creates a new dataset that controls for date

covid\_dataset\_analysis\_dateconst <- covid\_dataset\_interest %>%

filter(date == "2022-07-03")

View(covid\_dataset\_analysis\_dateconst)

#Plots a scatterplot of level of democracy against total cases per million

ggplot(covid\_dataset\_analysis\_dateconst, aes(x = level\_of\_democracy, y = total\_cases\_per\_million)) +

geom\_point(size = 2, alpha = 0.4, colour = 'blue') +

labs(

x="Level of Democracy (MPI)",

y="Total Cases (per million)",

title='MPI vs Total Cases (per million)',

subtitle = 'between: 2020-01-01 and 2022-07-03',

caption='Source: Our World in Data and V-Dem Project',

tag='A') +

geom\_smooth(method = 'lm', se = FALSE, colour = 'darkorange') + #plots a linear line of best fit

stat\_cor(method = "pearson", label.x = 0.05, label.y = 5.1e+05, size = 3.5) #calculates R and p values and plots them

#Plots a scatterplot of government transparency against total cases per million

ggplot(covid\_dataset\_analysis\_dateconst, aes(x = gov\_transparency, y = total\_cases\_per\_million)) +

geom\_point(size = 2, alpha = 0.4, colour = 'blue') +

ylim(-2.1e+05, 6e+05) +

labs(

x="Government Transparency",

y="Total Cases (per million)",

title='Government Transparency vs Total Cases (per million)',

subtitle = '2022-07-03',

caption='Source: Our World in Data and HRV Transparency Project',

tag='B') +

geom\_smooth(method = 'lm', se = FALSE, colour = 'darkorange') + #plots a linear line of best fit

stat\_cor(method = "pearson", label.x = -6.4, label.y = 5.1e+05, size = 3.5) #calculates R and p values and plots them

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 6: Analysis of dataset: government transparency bar chart

##

#Creates a subset dataset including the 8 highest and 8 lowest transparency index countries

covid\_dataset\_country\_transparency <- covid\_dataset\_analysis\_dateconst %>%

filter(gov\_transparency >= 4.66 | gov\_transparency <= -0.95) %>%

arrange(desc(gov\_transparency)) #orders dataset by largest to smallest government transparency

View(covid\_dataset\_country\_transparency)

#Plots a bar graph of the top 8 and bottom 8 countries for government transparency

ggplot(covid\_dataset\_country\_transparency, aes(x = gov\_transparency, y = reorder(location, gov\_transparency))) +

geom\_bar(stat="identity", width=0.7, fill="steelblue") +

geom\_text(

aes(label = sprintf("%.2f", round(gov\_transparency, 2))), size = 3,

hjust = ifelse(covid\_dataset\_country\_transparency$gov\_transparency >= 0, -0.2, 1.1)) + #includes values for each bar

theme\_minimal() + #changes the theme of the bar chart to minimal

xlim(-8, 6.8) + #forces x limit values onto the chart

labs(

x="Government Transparency",

y="Country",

title='Bar Chart of Government Transparency',

subtitle = 'top 8 & bottom 8 countries',

caption='Source: HRV Transparency Project',

tag='')

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 7: Analysis of dataset: correlations with other variables

##

#Creates a subset dataset called covid\_dataset\_corr

covid\_dataset\_corr <- covid\_dataset\_interest %>%

select(where(is.numeric) | date) %>% #selects numeric variables + date

select(-year) %>% #removes the year variable

filter(date == "2022-07-03") %>% #filters the data by date

mutate(date = NULL) #removes the date variable

View(covid\_dataset\_corr)

#Creates a variable correlations that stores the correlations values of covid\_dataset\_corr variables

correlations <- cor(covid\_dataset\_corr, use = "pairwise.complete.obs") #removes NAs by pairwise deletion

p.mat <- cor.mtest(correlations) #creates p values + confidence intervals for the correlations

#Creates a variable col that includes a list of hexidecimal colour list

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))

#Plots a correlation plot only including significant correlations between variables

corrplot(correlations, method="color", col=col(200),

type="upper", order="hclust",

addCoef.col = "black", #add correlation coefficient

tl.col="black", tl.srt=45, tl.cex = 0.5, #alters the text label colour, rotation, and position

p.mat = p.mat, sig.level = 0.01, insig = "blank", number.cex=0.75, #filters out squares with insignificant p values

diag=FALSE) #hides diagonal squares

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 8: Prediction of Time Series Data - total\_cases\_per\_million

##

#Example subset dataset of Bulgaria

covid\_dataset\_predict\_Bulgaria <- covid\_dataset\_interest %>%

filter(location == "Bulgaria")

#Example plot of Bulgaria showing the linear nature of date and total cases per million

ggplot(covid\_dataset\_predict\_Bulgaria, aes(x = date, y = total\_cases\_per\_million)) +

geom\_line(size = 0.8, alpha = 1, colour = 'blue') +

labs(

x="Date",

y="Total Cases (per million)",

title='Cumulative Total Cases (per million) of Bulgaria',

subtitle = 'between: 2020-03-08 and 2022-07-03',

caption='Source: Our World in Data',

tag='') +

geom\_smooth(method = 'lm', se = FALSE, colour = 'darkorange') +

stat\_cor(method = "pearson", label.y = 70000, size = 3.5)

#Creates a training dataset

train\_data <- covid\_dataset\_predict %>%

select(location, date, total\_cases\_per\_million) %>%

#Filters out 20% of the dates at the end of the data set #80% of 861 = 689

filter(date <= "2021-12-23") %>%

add\_column(forecast = "train") #adds a new variable to denote data as training data for plot differentiation

train\_data.cat <- train\_data %>%

as\_tsibble(key = location, index = date) %>% #converts training dataset to tsibble format by location

fill\_gaps() %>% #fills out missing dates

model(arima = ARIMA(total\_cases\_per\_million)) %>% #Creates a model of each country using a linear ARIMA model

forecast(h = 192) #forecasts/predicts 192 dates in the future (same length as test data)

train\_arima\_forecast <- as.data.frame(train\_data.cat) %>% #converts tsibble dataset back to a dataframe for analysis

select(location, date, .mean) %>% #selects the country, date and forecast columns

rename("total\_cases\_per\_million" = ".mean") %>% #renames total\_cases\_per\_million prediction data column

add\_column(forecast = "forecast") #denotes data as forecasting data

#Creates a test dataset

test\_data <- covid\_dataset\_predict %>%

select(location, date, total\_cases\_per\_million) %>%

#Contains 20% of the dates at the end of the data set, to be used to test against the forecast

filter(date > "2021-12-23") %>%

add\_column(forecast = "test") #denotes data as test data

#Views the training, forecasting, and test cleaned datasets

View(train\_data)

View(train\_arima\_forecast)

View(test\_data)

#Combines the 3 datasets above together by the 4 column variables

prediction\_dataset\_final <- train\_data %>%

full\_join(train\_arima\_forecast) %>%

full\_join(test\_data)

#Views the prediction dataset

View(prediction\_dataset\_final)

#Creates a subset of data from the prediction dataset for Bulgaria

example\_plot <- prediction\_dataset\_final %>%

filter(location == "Bulgaria")

#Plots a line graph of date vs total\_cases\_per\_million separating forecast, test, and training data by colour

ggplot(example\_plot, aes(x = date, y = total\_cases\_per\_million, group=forecast, color=forecast)) +

geom\_line() +

labs(

x="Date",

y="Total Cases (per million)",

title='Bulgaria Total Cases (per million) Forecasts',

tag='')

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 9: Root Mean Squared Error (RMSE) and Normalised Root Mean Squared Error (nRMSE) of each country

##

test\_data2 <- test\_data %>%

mutate(forecast = NULL) %>% #removes the forecast column

rename("actual" = "total\_cases\_per\_million") #renames test data to 'actual'

train\_arima\_forecast2 <- train\_arima\_forecast %>%

mutate(forecast = NULL) %>% #removes the forecast column

rename("predicted" = "total\_cases\_per\_million") #renames forecast data to 'predicted'

#Creates a subset dataset that joins the test and forecast data together

dataset\_rmse <- test\_data2 %>%

left\_join(train\_arima\_forecast2, by = c("location", "date")) #joins data by both location and date

View(test\_data2)

View(train\_arima\_forecast2)

View(dataset\_rmse)

dataset\_rmse\_corr <- dataset\_rmse %>%

group\_by(location) %>%

mutate(rmse = rmse(actual, predicted)) %>% #Creates a new variable that calculates the RMSE of actual vs predicted

mutate(minvalue = min(actual)) %>% #finds the minimum value of actual and temporarily stores it

mutate(maxvalue = max(actual)) %>% #finds the minimum value of actual and temporarily stores it

mutate(nrmse = rmse / (maxvalue - minvalue)) %>% #calculates a normalised RMSE based off RMSE/(max - min)

mutate(minvalue = NULL, maxvalue = NULL) #removes temporary min and max values

View(dataset\_rmse\_corr) #views the dataset containing both RMSE and nRMSE values

#---------------------------------------------------------------------------------------------------------------------------------------

##

###SECTION 10: nRMSE correlation analysis between accuracy and government transparency

##

#Creates a temporary datset that filters for a single date (to remove duplicates)

covid\_dataset\_interest\_temp <- covid\_dataset\_interest %>%

select(location, date, gov\_transparency) %>%

filter(date == "2022-07-03") %>%

mutate(date = NULL)

#Creates a correlation dataset that takes nRMSE values and joins the gov\_transparency variable back to the dataset

dataset\_rmse\_correlation <- dataset\_rmse\_corr %>%

left\_join(covid\_dataset\_interest\_temp, by = c("location")) %>%

filter(date == "2022-07-03") %>%

mutate(date = NULL, actual = NULL, predicted = NULL) #cleans up the dataset by removing non-used columns

#Plots government transparency vs nRMSE to test for correlation

ggplot(dataset\_rmse\_correlation, aes(x = gov\_transparency, y = nrmse)) +

geom\_point(size = 2, alpha = 0.4, colour = 'blue') +

labs(

x="Government Transparency",

y="Normalised RMSE (lower better)",

title='Government Transparency correlation against nRMSE',

tag='') +

geom\_smooth(method = 'lm', se = FALSE, colour = 'darkorange') + #plots a linear line of best fit

stat\_cor(method = "pearson", label.x = -6.4, label.y = 3, size = 3.5) #uses Pearson's correlation coefficient - plots R and p value

dataset\_rmse\_corr22 <- dataset\_rmse\_corr %>%

select(location, date, rmse, nrmse) %>%

filter(date == "2022-07-03") %>%

mutate(date = NULL)

write.table(dataset\_rmse\_corr22 , file = "olstab222.txt", sep = ",", quote = FALSE, row.names = F)