Covid-19 Regression Analysis: Death Predictions

Database & Analytics Programming CA 2 Project

Alun, Price
School of Computing
National College Ireland
Dublin, Ireland
x21123659@student.ncirl.ie

Maria, Migrova
School of Computing
National College Ireland
Dublin, Ireland
x21146021@student.ncirl.ie

Polina, Prinii
School of Computing
National College Ireland
Dublin, Ireland
x21137757@student.ncirl.ie

***Abstract* — Covid 19 has been an overarching presence in the lives of almost every person on earth for the last 3 years. This global threat has brought about a global effort. In parallel to this, the global information system has never been so advanced, allowing for extraordinary levels of data collection and transmission. This report intended to leverage this by analyzing key factors related to the pandemic. These data for this were programmatically accessed and analyzed. The goal for this report was to investigate the prediction potential of global deaths using regression analysis. Analysis of the data determined that not only was prediction possible but using the chosen independent variables a prediction potential of 86% was achieved. Further transformation of the data increased this value to 95%. Showing that 95% of the variance of global death rates were explainable through the variance of the chosen predictor variables**

I. Introduction

January 9th, 2019 was the date the WHO announced an unknown pneumonia like disease located in Wuhan, China, later identified as the novel SARS-CoV-2 virus or, Covid-19, since then the disease has dominated the lives of almost every person on the planet. As 2020 ended the United States alone had reached over twenty million confirmed cases and over three hundred thousand deaths. [1]

Much of this now three-year pandemic has been focused on the idea of "surges" or "waves". These are rapid and intense escalations in confirmed infections and in some cases paralleled rises in deaths due to the virus. These surges have made responding to the virus exceptionally difficult for both governments, society and most critically, the health care systems of countries. [2]

While this is not the first global pandemic in history it is the first to happen in an age where technology and information systems are as widespread as they are. This has allowed unprecedented production of data surrounding all aspects of the virus, aspects of the pandemic and excellent tracking of infection and mortality rates.

Despite the tragedy of the pandemic and given the ease of access the public have to the data being generated, there has been an increase in independent research into the data surrounding the Covid19 pandemic and its inherent statistical potential.

A. Motivation

While the scope of potential for data analysis of this topic if extremely large, the primary motivation for this project is the idea of being able to leverage the exceptional wealth of data that has been produced during and surrounding the Covid-19 pandemic to investigate what actionable information can be extracted and ultimately to find if the value of deaths could be predicted using a regression model by utilizing existing, livestreamed data from multiple sources.

The high-level categories of data that were chosen are:

- Confirmed Cases and deaths globally.
- Vaccination program overview globally.
- Restriction and testing data.

B. Research Questions

The questions to be investigated in this report are as follows:

- Can the global death rate be predicted using the variables contained within the chosen datasets?
- How accurate can the prediction be?
- What factors have the largest influence over the dependant variable?
- To what degree has vaccination roll out been adopted?
- What countries had the highest mortality rate?
- How have restriction methods changed over the course of the pandemic?

II. Related Work

Since the emergence of the Corona Virus Disease aka Covid 19 the world have made multiple attempts to make effective predictions for the relevant departments to allow for the response strategies to be more effective.

At a time where there was still no effective treatment before the vaccine rollout worldwide. A study by the name of **"The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method"** has been undertaken by combining the Long Short-Term Memory network and the Markov model based on the data of confirmed cases for four countries.

Whilst the vaccine rollout commenced in late 2020 on the 29th of December in Ireland, the world was still actively and rigorously fighting with the novel corona virus. Please note, that the vaccine rollout quoted is since the current report in question is being undertaken within Ireland.

With the use of the defined statistical tools above the study predicted that the number of cases of infection will stabilize and that the pandemic will be brought under control in Britain by February 2021. [3]

The accuracy of that study is quite outstanding, to be more statistical the LSTM-Markov method showed a .94 R^2 value. Taking the example of Britain, by the end of February 2021 the country managed to put the disease under control with the extensive health and security measures as well as the roll out of the COVID-19 vaccine. By the end of the month Britain successfully had over 20 million people vaccinated with the Prime Minister proposing the re-opening of the country to the public. [4]

The report in question will attempt to undertake similar analysis to derive an answer whether the identified independent variables can predict further mortality number as the infection rates to this day keep fluctuating around the world, especially during colder months alongside the mutation of the virus itself.

III. Methodology

This section looks at the group approach to analyze the data and the methods applied to extract, transform, and load the data in preparation for the final analysis.

A. Approach:

With the world actively fighting to control the ongoing Covid-19 pandemic, this report aims to perform a multiple linear regression to facilitate the understanding of the criteria that may dictate the Covid-19 mortality rates and potentially allow for the prediction of the same.

With the outlined statistical approach, the report aims to take the dependent variable of 'Death' number and evaluate in the *Results* section if the independent variables hold any explanation for the outcome. Ultimately, answering the question of "**Can the independent variables be used to predict the dependent variable?".**

An additional approach of data visualization such as heatmaps, bar charts and waterfall charts are undertaken to compliment the data analysis.

B. Methods:

To prepare the raw data for final analysis the following programmatic approach was undertaken following the Knowledge Discovery in Databases (KDD) process (see Fig.1):

1) Selection:

The selected data was extracted from open-source databases which are hosted and maintained on GitHub.

The raw data was extracted in multiple formats such as CSV, JSON and or XML using both Python and R to process the extraction.

2) Pre-Processing:

The extracted data was then programmatically stored in MongoDB for a pre-processing, here a decisional approach was applied to extract and store useful information to the final regression analysis.

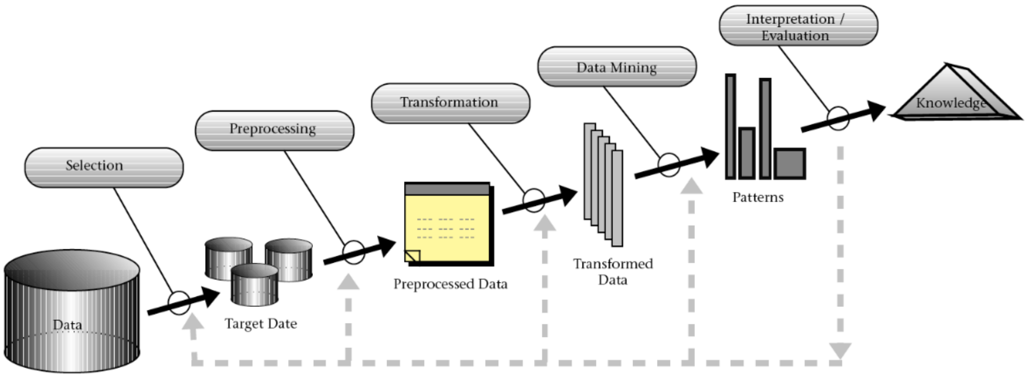


Figure 1: KDD Process

The programmatic approach was undertaken using Python and MQL through pymongo.

3) *Transformation:*

The pre-processed data is then programmatically stored in PostgreSQL with the approach of extracting the data into mini-CSV files for future final merge.

The programmatic approach was undertaken using Python and SQL.

4) *Data Mining:*

The final dataset was merged using Python based on two common columns being '*ISO Code'* and '*Country'* in preparation of the regression analysis which is undertaken in R.

The final merged file is then imported into R for the regression analysis where the results are derived.

Additionally, during this stage the data is selected and prepared for data visualization.

5) *Interpretation / Evaluation:*

Lastly, the results derived as part of the Data Mining step are interpreted and evaluated.

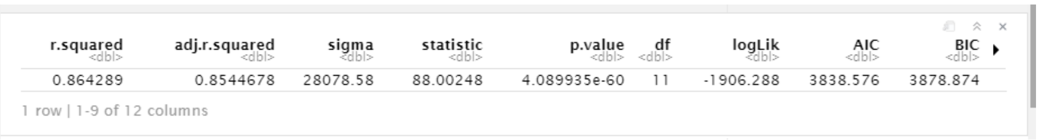
IV. Results & Evaluation

The absolute goal of the project was to perform a multiple linear regression analysis by building a model which aims to analyze a relationship between number of deaths and other variables.

This analysis was performed in R Studio using the lm () function. For the reports final regression model, it was decided to use all the variables, as outlined below:

- Deaths – total deaths by each country
- Confirmed – total confirmed cases by each country
- Tests – total positive tests by each country
- MaxStringecyIndex – Maximal stringency index by each country
- MeanStringencyIndex – Average stringency index by each country
- MaxFacialCoverings – Maximal facial covering restriction value by each country
- MeanFacialCoverings – Average facial covering restriction value by each country
- MaxStayHome – Maximal stay home restriction value by each country
- MeanStayHome – Average stay home restriction value by each country
- Totalvaccinations – Number of total vaccinations by each country
- Totalbooster – Number of total boosters by each country
- Fullyvacnumber – Number of fully vaccinated people by each country

Results were evaluated from the model using the glance () function. The resulting R-squared value is 0.86, which is very good. R-squared value measures and demonstrates the strength of the relationship between our model and the dependent variables. In this case the strength of the relationship is 86%.



r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>
0.864289	0.8544678	28078.58	88.00248	4.089935e-60	11	-1906.288	3838.576	3878.874

1 row | 1-9 of 12 columns

Figure 2 Values from glance ()

First Figure 3 shows the predicted deaths using linear regression model. Second Figure 3 represents plotted residuals. Residuals are placed around 0, which indicates that the model is positively accurate.

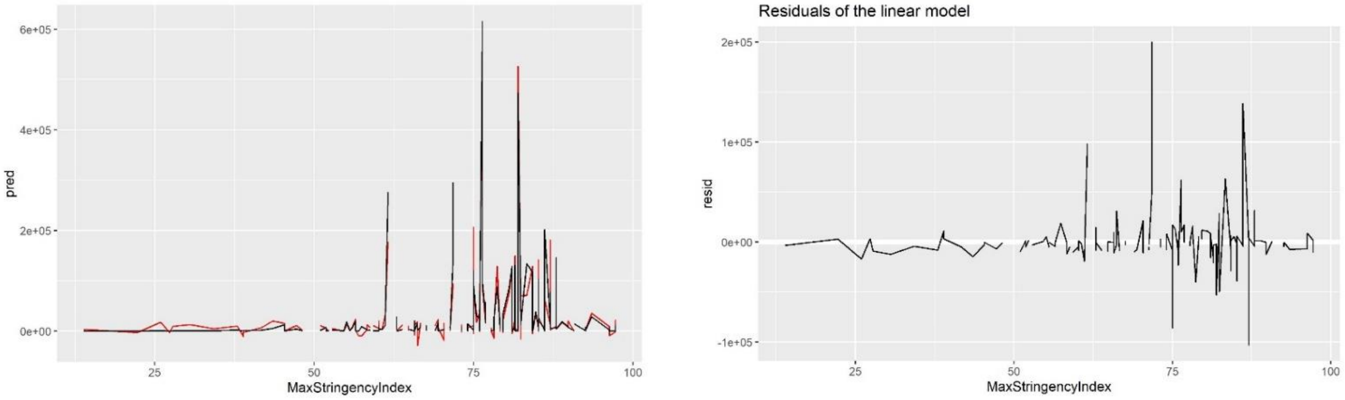


Figure 3: Linear Regression Model & Residuals vs Fitted

The next part of the analysis was to check if all the assumptions were justified for our model.

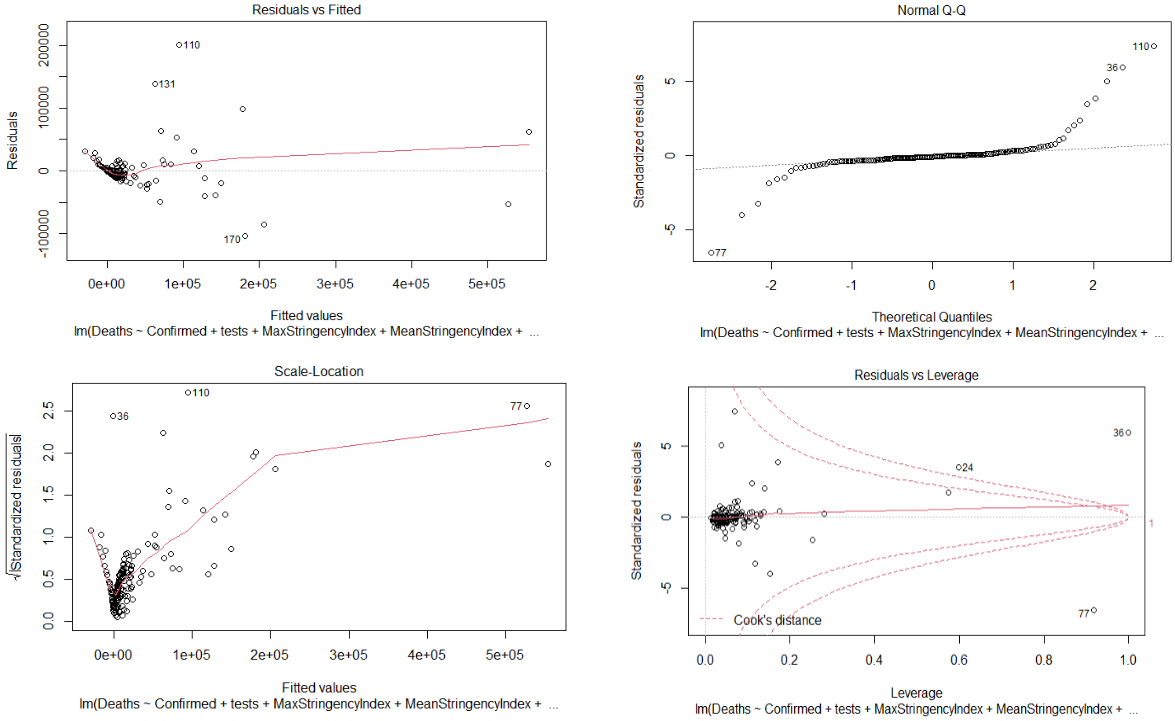


Figure 4: Regression Model pre-Transformation

For getting the assumptions graphed, the plot () function is used.
1. Residuals vs Fitted - this graph is showing heteroscedasticity.
2. At the second graph (Normal Q-Q) checks the normality assumption. The points follow a straight line with some outliers around. This assumption is justified. Outliers are on line 36,110 and 77.
3. At the third graph (Scale – Location) checks the constant variance assumption. The analysis can notice heteroskedasticity in the model as the points are not located around a horizontal line, but they follow a 'V' shape.
4. The third graph shows us the outliers. The outliers are on line 24,36,77.
As these assumptions are not justified, the use a log version of the linear regression model is required. The analyses started by plotting histograms of all the variables involved within the dataset to see which variables are showing skewness. Out of 10 variables, six were transformed using the log of nature function and re-evaluated. The 6 variables are: confirmed, deaths, tests, totalvaccines, totalboosters and fullyvaccinated.
After creating a new linear regression model using log () function the model showed an R-squared value equal to 0.95, which means that the strength of the relationship between the model and the dependent variables is 95%. The new model is 9% better than the old one.

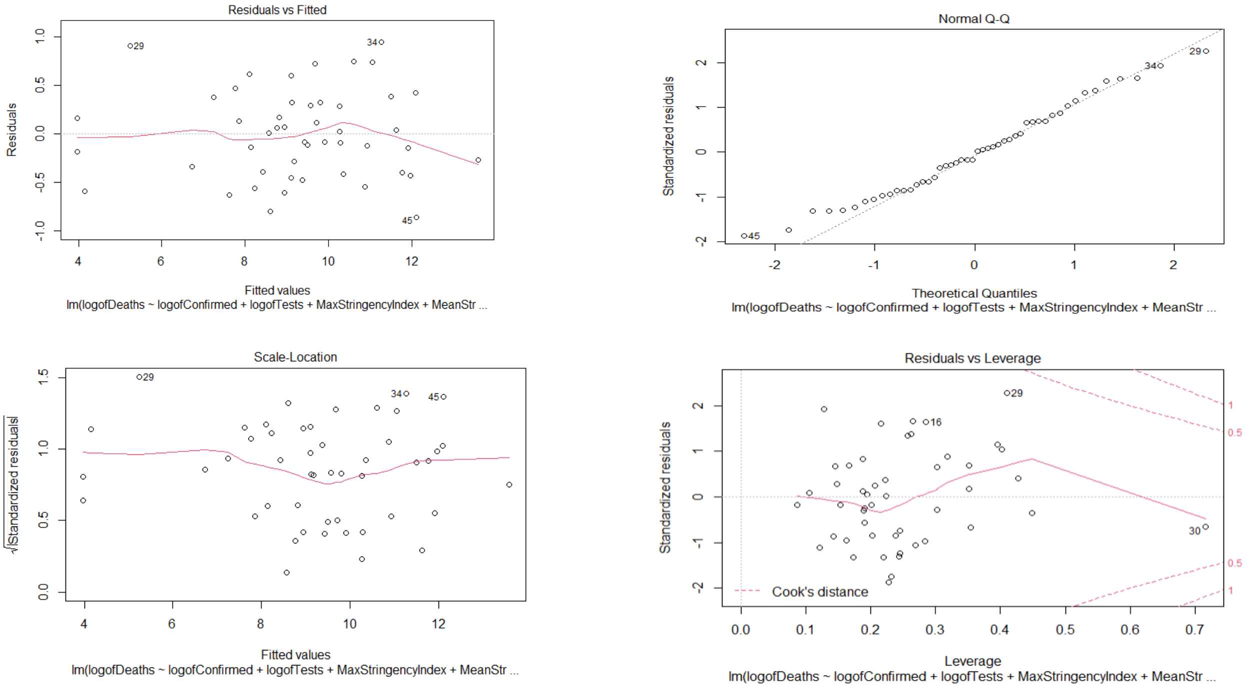


Figure 5: Regression Model after Transformation

It is seen that after changing the linear regression model into log form all the assumptions are justified.
The last part of this analysis was the evaluation of the model. For this part it was decided to use MAPE and MAE function.

The MAPE values is defined as actual to observed value minus the forecasted value. Value of 0.99 means that our model's predictions are, on average 1% off from the actual values.

The MAE value means that on average the forecasted distance from the true value is 50 403, which is OK as the values range from 35 to 615 636.

From these resutls it is seen that the linear regression model perfoms very well.

To support the above findings a number of intersting data visuals have been drafted. Figure 6 shows the overall Europe uptake of vaccinations. It is interesting to see that the numbers under review are showing a good healthy uptake, as when devided by 2 the final figure corresponds closely to the population of each country. This indicates a high and strong number of fully vaccinated people, however for the exeption for Russia with only 58 million people fully vaccinated being only 1/3 of the overall population.

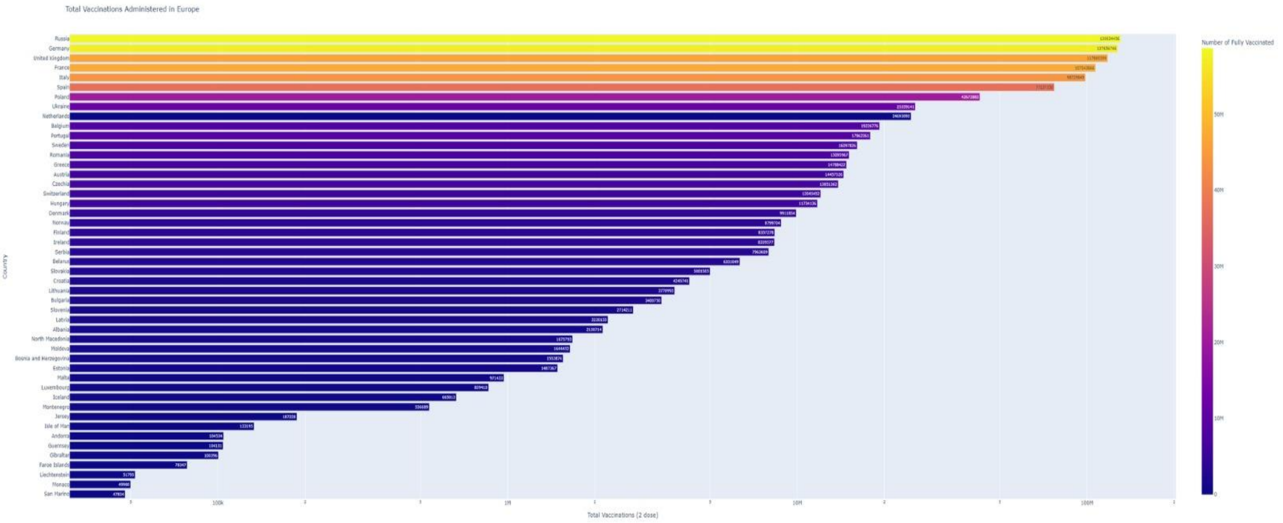


Figure 6: Total Vaccinations Administered in Europe

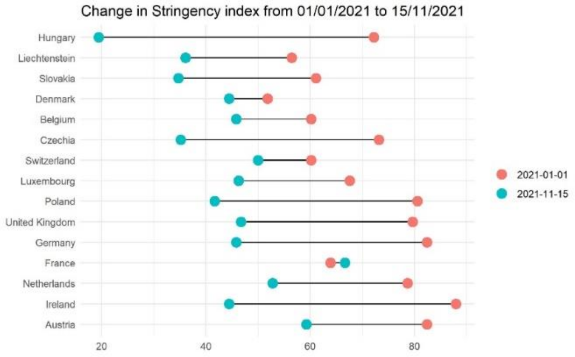


Figure 7: Comparison of Max Stringency Index

Figure 7 a list of 15 central Europe countries comparing a change in stringency index from 01/01/2021 to 15/11/2021 is shown.
It shows that the largest difference was for Hungary and the smallest difference was in France.
This tells that potentially the restriction put in place for protection may not always be bearing best results, as France to this day ranks as one of the highest infection rate countries in Europe.

Figure 8 demonstartes the top 20 countries with the largest number of deaths to date. Countries with the biggest numbers are United States, Brazil, India, Mexico and Russia.
These results can be explained very simply for two countries by the low uptake of vaccinated in both the United States of America and Russia. With both countries missing a very large percentage of fully vaccinated citizens.

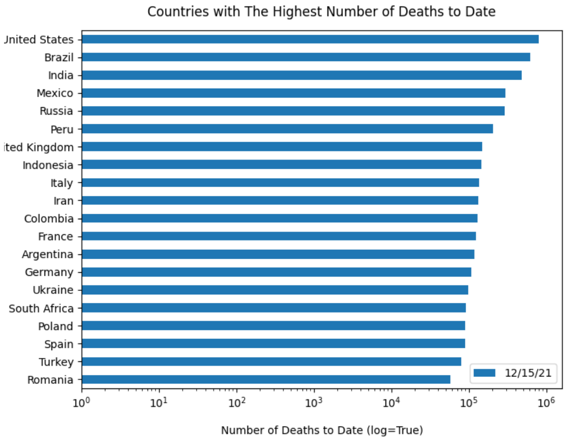


Figure 8: Top Country Death Rate

V. Conclusion & Future Work

The purpose of this project was to create a multiple linear regression model which is used to describe a relationship between worldwide death numbers and other independent variables like testing numbers, vaccination and government restrictions. Based on our analysis we can conclude that there is a strong linear relationship between these variables. Which means that we can predict total deaths using these variables.

Even though we got a really strong linear regression model, per future work could be oriented on using a polynomial regression instead. Polynomial regression provides the best approximation of the relationship between the dependent and independent variable.

We could also improve our linear regression model by adding more data or found another variables which would have bigger influence on the total deaths.

Acknowledgment

The team would like to express their appreciation to the National College of Ireland who have provided the possibility to complete this work and report. A special thank you goes to our lecturer Rashid Mijumbi who teaches Database and Analytics Programming.

Rashid has helped the team to coordinate their project, especially in the understanding of the objective and in writing this report.

References

[1] AJMC Staff, "A Timeline of COVID-19 Developments in 2020", 2021 Jan 1st, Accessed on: Dec 16th 2021, [Online], Available at: https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

[2] L.Maragakis, "Coronavirus Second Wave, Third Wave and Beyond: What Causes a COVID Surge", 2021 Oct 21st, Accessed on: Dec 16th 2021, [Online], Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus

[3] R. Ma, X, Zheng, P. Wang, H. Liu, C. Zhang, "The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method ", 2021 Aug 31st, Accessed on: Dec 16th 2021, [Online], Available at: https://www.nature.com/articles/s41598-021-97037-5#Abs1

[4] [4] Wikipedia, "Timeline of the COVID-19 pandemic in the United Kingdom (January–June 2021)", 2021, Accessed on Dec 17th 2021, [Online], Available at https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_Kingdom_(January%E2%80%93June_2021)