SUBMITED BY: Harsh Kakkad

MSC DATA SCIENCE

COVID-19 Death analysis by local authority area for England.

Date of Submission :- 4 January 2024

1 Contents

2 Table of figures

3 ABSTRACT

This research project is focused on analysing the total death counts in England that happened due to Covid-19, within the period of March 2020 to April 2021, and its possible causes. As Covid-19 is a global pandemic almost everything can be considered as a factor, hence there has been extensive research on age and population health data of England Districts, so this research is solely conducted considering only household themes such as number of cars owned per house, Unemployed household, Deprivation according to 4 dimensions etc.

Some of these variables represent the number of people in a house, how deprived are the people and others even compare the districts based on how dependent are people on others for daily needs (Unemployed people, dependent children in house and long-term health or disability in the house), as they might not have enough to survive through the quarantine.

The research is to find the different factors that have affected to what has happened, as we cannot change the results but we can definitely get a better understanding of these factors and their impact on Covid-19 deaths, which will help if something similar happened in future than we can prepare accordingly to reduce the effects of that event.

The culmination of our comprehensive process yields the result that we, indeed, fail to reject the null hypothesis. This signifies a clear association between our dependent variable and the variables scrutinized in this research. However, to provide a more nuanced perspective, it's important to note that the extent of the variance relation with our dependent variable (deaths per 1000 people) and other variables is not substantially large, approximately around 18 percent. This nuanced understanding sheds light on the intricacies of the relationships between our variables and emphasizes the need for further exploration and analysis to fully grasp the dynamics at play in our dataset.

4 INTRODUCTION

COVID-19 or SARS-CoV-2 is a virus that started spreading very quickly making it a disease to becoming a Pandemic that spread throughout the globe. It became like a test of people's immunity power, better the immunity lower the chances of dying. When someone with COVID-19 breathes, speaks, coughs or sneezes, they release small droplets containing the virus. You can catch it by breathing in these droplets or touching surfaces covered in them (NHS, 2023). Hence, it became highly contagious which resulted in so many deaths within a short period of time. So, to reduce its effects governments all across the globe started Quarantines and many other precautions, which resulted into many changes in daily lives of people.

Here, in this research I have analysed the death toll due to COVID-19 with some of the household themes for all the districts in England. All the variables can be generalized into 4 categories: Transportation availability for household, Family situation of the house, Financially or Physically Dependent people in household and Deprivation of household according to 4 dimensions. By using these factors, we are trying to find those that reflect the causation of COVID-19 deaths.

5 VARIABLE ANALYSIS

COVID-19 took the lives of millions all across globe and even just in England itself death tolls reach in tens of thousands in just the 14-month period of our data, which are divided among all the districts. These values have been analysed using Household themes, as COVID-19 spreads through a medium and the symptoms take time to show up, so if one person in household caught the virus, then there might be higher chance of others getting caught, which then might increase the Death toll.

5.1 TRANSPORTATION AVAILABILITY

The importance of transportation in our lives is so large that we are totally dependent on vehicles and public transport almost everyday to just get all our basic day to day requirements, this came into attention when Quarantine started as all these services were to be stopped and going out of house was allowed for a good reason, this created a chaos for people not having any vehicle.

To address this problem, we have taken the most common means of transport (cars) as our central factor and divided into 3 variables: Household with No cars, Household with 1 car and Household having 2 or more cars. This is very important as this variable touch both the transport availability and wealth factors together.

5.2 FAMILY SITUATION OF THE HOUSE

The family situation indicates the number and type of people living in the house, as let's say a young person lives with family catches COVID and affects their family and if they have someone old in house who got affected than their chance of dying is higher.

This theme consists of One-person household, one-person house where their age is over 65, one-person household where their age is lower than 65, One family household, one family household where all are over 65 and Other than these household types. This process of getting infected can be singled out by using these variables perfectly, as we are also considering age (not as primary factor but as secondary factor), which is very important factor considering the disease that we are discussing about infects our immunity.

5.3 FINANCIALLY OR PHYSICALLY DEPENDENT PEOPLE IN HOUSE

In both the terms either financially or physically dependent people had problems as they were incapable to get survivable resources hence they are highly likely to break the quarantine rules which would make them more likely to contribute in the death toll factor.

Here for our data we have used Unemployed people in household, dependent children in a household and having a long-term health problem or disability in household as our factors. These are not directly dependent to our disease but dependent to the factors that lead to getting infected.

5.4 DEPREVATION ACCORDING TO 4 DIMENSIONS

The dimensions of deprivation used to classify households are indicators based on four selected household characteristics.

Education

A household is classified as deprived in the education dimension if no one has at least level 2 education and no one aged 16 to 18 years is a full-time student.

Employment

A household is classified as deprived in the employment dimension if any member, not a full-time student, is either unemployed or economically inactive due to long-term sickness or disability.

Health

A household is classified as deprived in the health dimension if any person in the household has general health that is bad or very bad or is identified as disabled.

Housing

A household is classified as deprived in the housing dimension if the household's accommodation is either overcrowded, in a shared dwelling, or has no central heating.(ONS, UK).

6 OBJECTIVES

This research project is laser-focused on finding out the factors that affect the deaths through COVID-19 over all the districts in England. We are trying to also check our hypothesis of using these household factors we can get a significant relation with our death variable. Our research and analysis of all the variables will conclude that if household values play any significant role in increase or decrease of the death tolls due to COVID-19.

Using this as my Hypothesis,

Null Hypothesis: COVID-19 deaths are significantly related to the household variables used here.

Alternate Hypothesis: There is no significant relation of household variables used with COVID-19 deaths.

To work on this hypothesis, we have used 17 variables (columns of independent variables) with different themes in mind to get desired output.

7 DATA ACQUISITION

This research includes death due to COVID-19 data from the period of March 2020 to April 2021 of all the districts over England. With this as our dependent variable we are using the household themed variables. These variables are obtained from the UK government website called nomisweb where we have chosen the data from 2011 census which is properly refined to not create any confusions as the districts name and codes are not as constant from 2011 to 2021 so for simplicity we have used 2011 census districts to our data.

The other variables can be obtained from selection option in the website where we can directly select the name of theme we want and select the districts data of England and wales prior to 2015. After getting the excel sheet we remove the ones in wales by the code which starts with a W, we will then remain with 327 rows, where we will use SQL to join all the tables with each other and also join this data with our dependent variable table, where primary key is Geography code and LA code.

Now at the end we will remove duplicated columns and save that file as a csv file, which then can be further used for data exploration.

8 METHADOLOGY

8.1 DATA EXPLORATION

For the entire procedure, our initial step is to designate our working directory to the location where we've stored our data and files. This ensures that our libraries, once installed, are in the same location as other essential components. To confirm the directory, we employ getwd(). Subsequently, we install the required libraries for our project, namely:

1. Amelia: A package proficient in handling missing data, crucial for identifying null values in our original dataset.

2. Corrplot: An R package designed for visualizing correlation matrices, employing diverse plotting techniques to enhance our understanding of variable relationships.

3. Corrgram: A library instrumental in creating correlation diagrams, facilitating the identification of patterns and relationships among variables.

4. Ppcor: This package aids in obtaining partial correlation values, particularly useful when comparing the correlation of one variable with another while excluding a third variable.

5. Psych: A comprehensive package tailored for psychological and behavioral research. In our project, it serves the purpose of conducting factor analysis, enabling the analysis and interpretation of complex relations within our dataset.

6. Car: An essential library for applied regression analysis, pivotal for visualizing regression models and enhancing our understanding of them.

After the installation of these libraries, we proceed by reading our data, which is in CSV format and saved in the same folder as our working directory. Notably, we set the value of the function stringsAsFactors to false, ensuring that string values in geography_name and geography_code are treated as characters rather than factors, as evident in the output of the str function applied to our data.

Upon storing our dataset in a variable named mydata, we use the head function to verify the successful import of data into the variable, visible in the Environment variables section in the top-right corner of RStudio.

In the initial stages of data exploration, we check for any null values in our dataset using the apply function. This allows us to determine the sum of null values in each column, showcasing the overall count for each column. Additionally, we visualize these null values using a missmap plot for a more intuitive understanding.

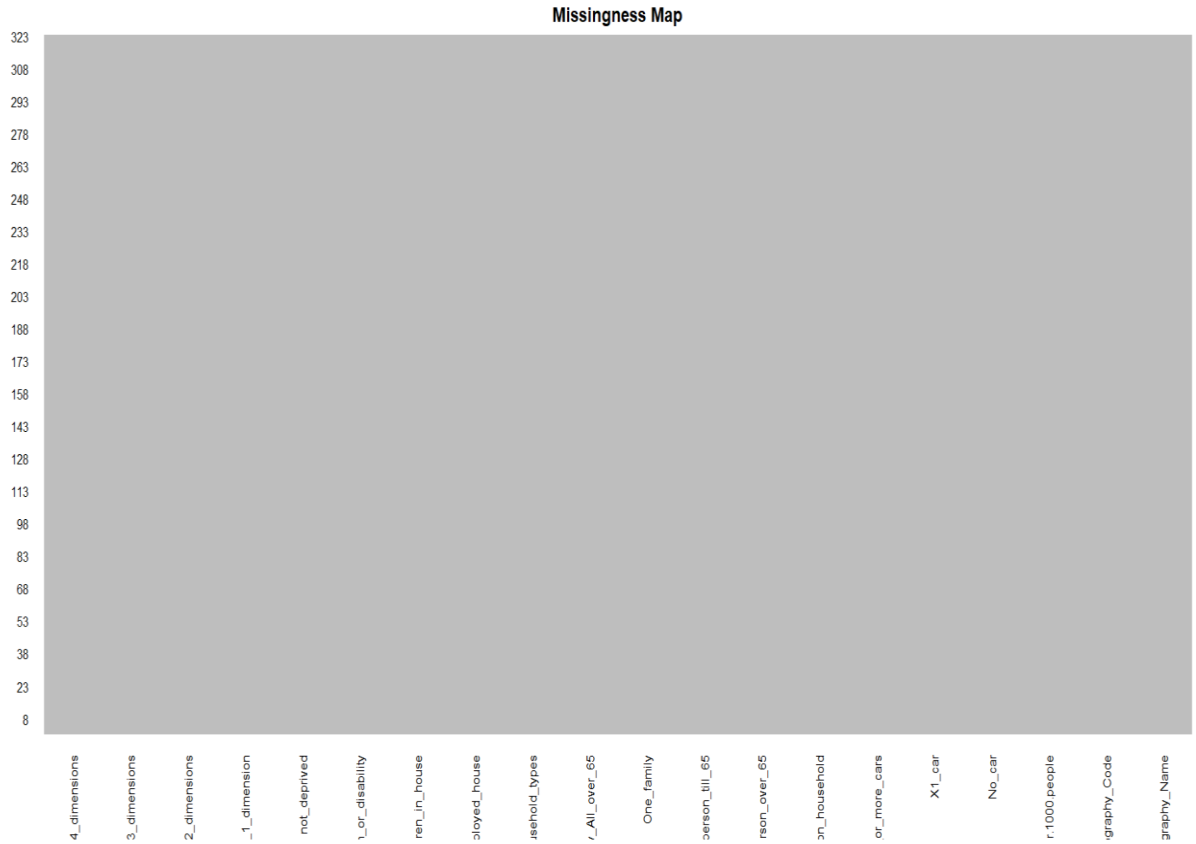

Figure 1-Missingness Map

As we can see there is not a single null value in our dataset from figure-1, hence we can move forward with our next processes and there is no need for applying techniques such as mean or median values to remove null values as there are none.

We will now remove the unnecessary columns as there is no need for Geography code and Geography name in any of our processes, which are done on numeric values. After finishing all these processes, we can add these again if required, but in that case, we cannot remove a single row out of our dataset to get consistent results. Hence, we have used subset function on mydata variable – the column names that we don't need. After this we checked the result if the function worked or no by str function. Just for reference, every value should be numeric for further processing.

Now, as we got our table, we will attach this table to work within the table only which will directly help us as there will be no need to refer this table again and again for every function.

As we move forward we have 2 types of variables:

- Dependent variable: Deaths per 1000 people, which is our target variable, that is we will use all other variables to get a significant relation with this variable.
- Independent variables: Every variable except Deaths per 1000 people that we are using to get some significant relation with dependent variable.

To check the relation of our dependent variable we need to check if the dependent variable is normally distributed or not, as both cases have different routes for data processing in later stages.

Here, in our case, first we have taken the summary statistic which gave us all the quartile values and mean value. Here to check significance we can observe that how different are mean and median values. In our case, the difference between mean and median in our dependent variable is only 0.002 (|2.170 - 2.172|). Hence, we can say that our dependent variable is normally distributed using summary statistic. But to be sure of this result we need other tests to confirm by using other tests.

The first test that we will use is by making boxplot of dependent variable:

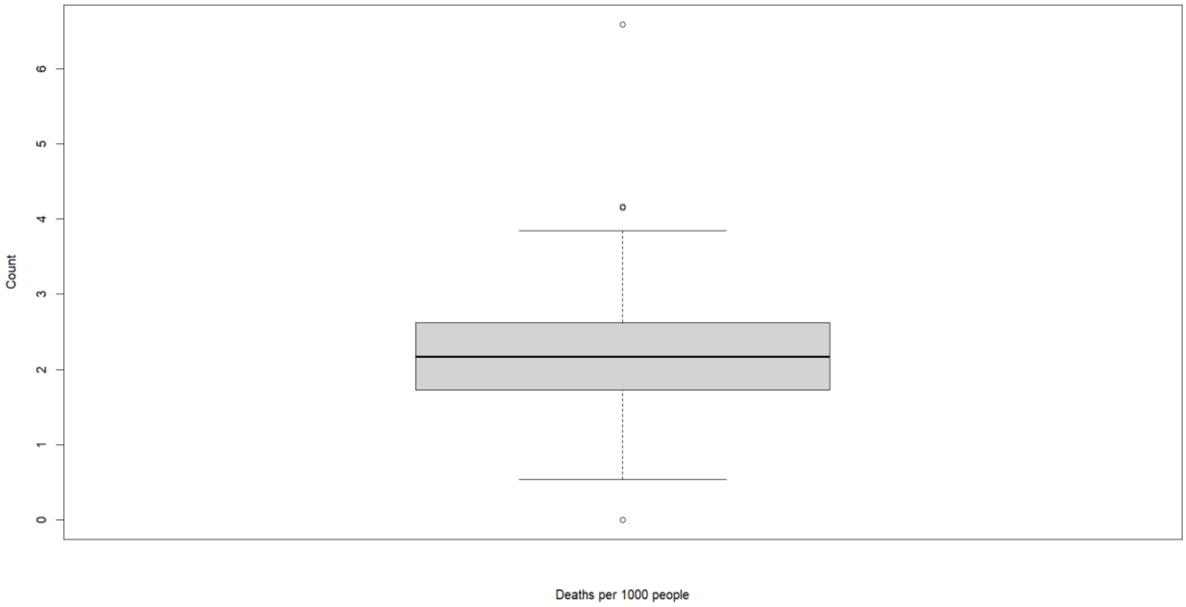

Figure 2- Boxplot of Deaths per 1000 people

As we can see in this boxplot, the middle line is median and other lines are quartile values, the values above median are similar to values below median and there are a smaller number of outliers to make this not normal. Hence, by boxplot test we can say that Deaths per 1000 people is normally distributed.

The next test we will use here is Q-Q plot test, here if the expected normal line matches with our data line then we can say that data is normally distributed.

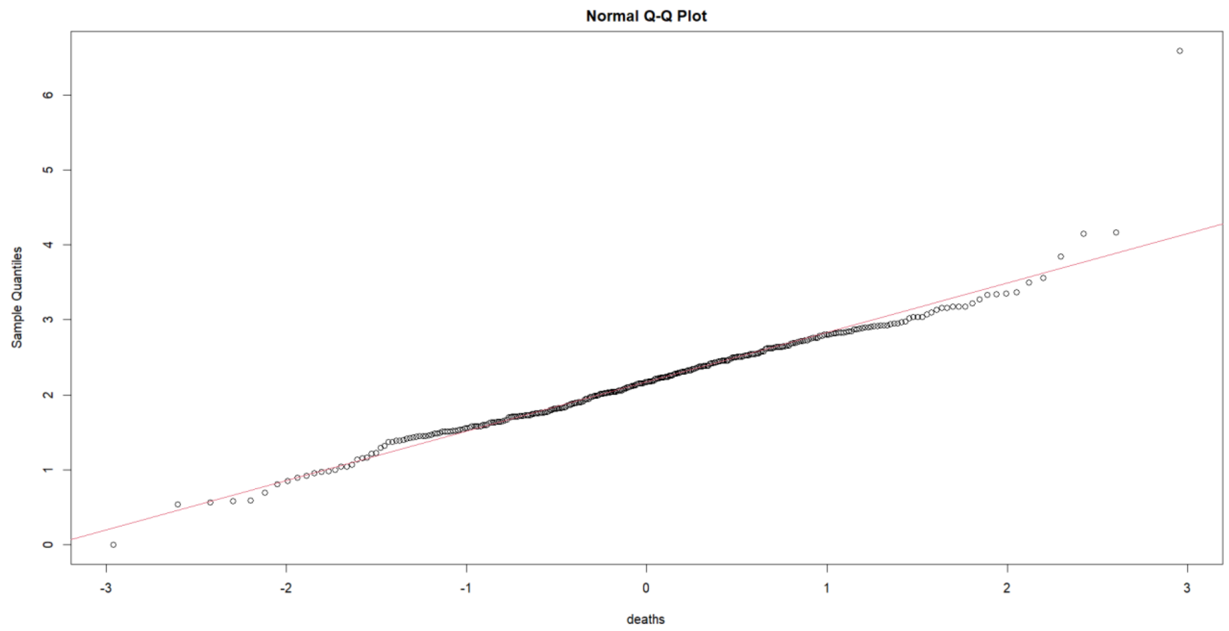

Figure 3- Q-Q plot of deaths per 1000 people

In this Q-Q plot, the red line closely aligns with our data trend, indicating that our dependent variable follows a normal distribution.

Our subsequent test involves a histogram analysis. We created a histogram of our data and incorporated a normal distribution line for comparison. If the line and our histogram exhibit similar trajectories, we can confidently assert that the data is distributed normally.

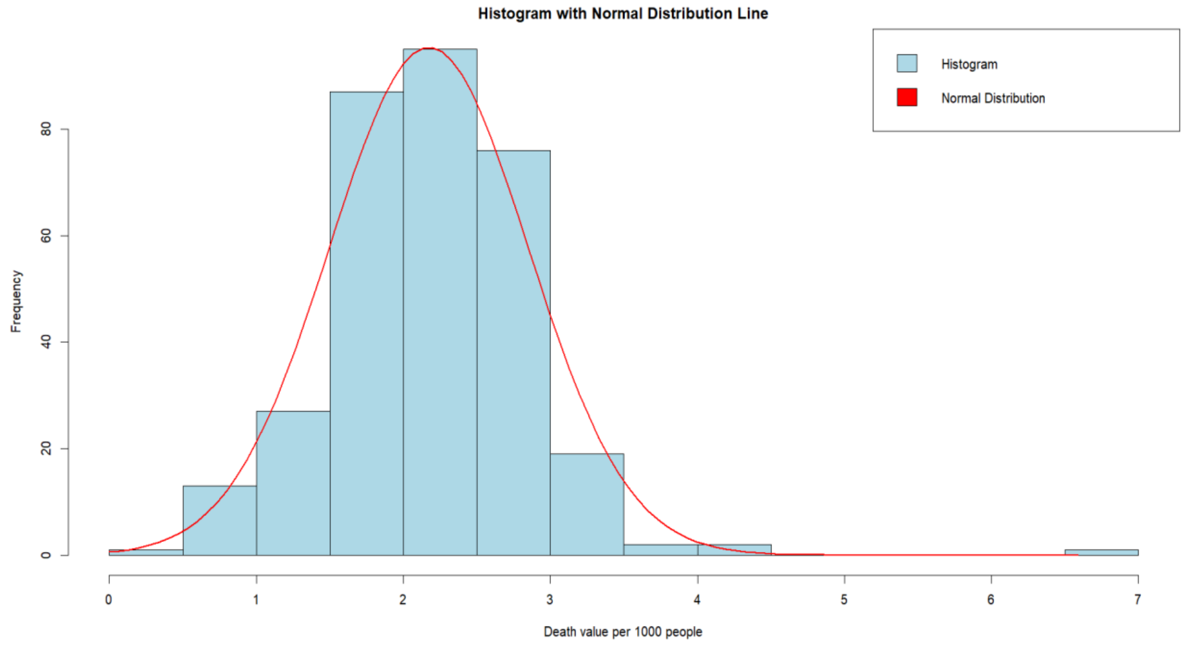

Figure 4- Histogram of Deaths per 1000 people

As evident from this graph, the normal distribution line and histogram follow similar trajectories. Therefore, based on the histogram test, we can assert that our dependent variable is normally distributed.

The final test we employed is the K-S test, which provides a precise result regarding the normality of the dependent variable. After conducting the K-S test, we obtained a p-value of

0.551, significantly greater than 0.05 (at a 95% confidence level). Consequently, we fail to reject the null hypothesis, indicating that the dependent variable is normally distributed.

Having conducted these tests, we can confidently state that our dependent variable follows a normal distribution.

Now, having established normality in the dependent variable, our focus shifts to understanding the influence and extent of impact that other variables have on the dependent variable. To achieve this, we will generate a correlation matrix of our variables. In the initial data analysis step, we can straightforwardly eliminate some variables that exhibit very low correlation with our dependent variable. Since these variables have such minimal correlation that removing them would not significantly affect our results, we will use the correlation between our dependent variable and all other variables. To gain a better understanding of their relationships, we have created a corrgram plot on mydata using the Pearson correlation method.

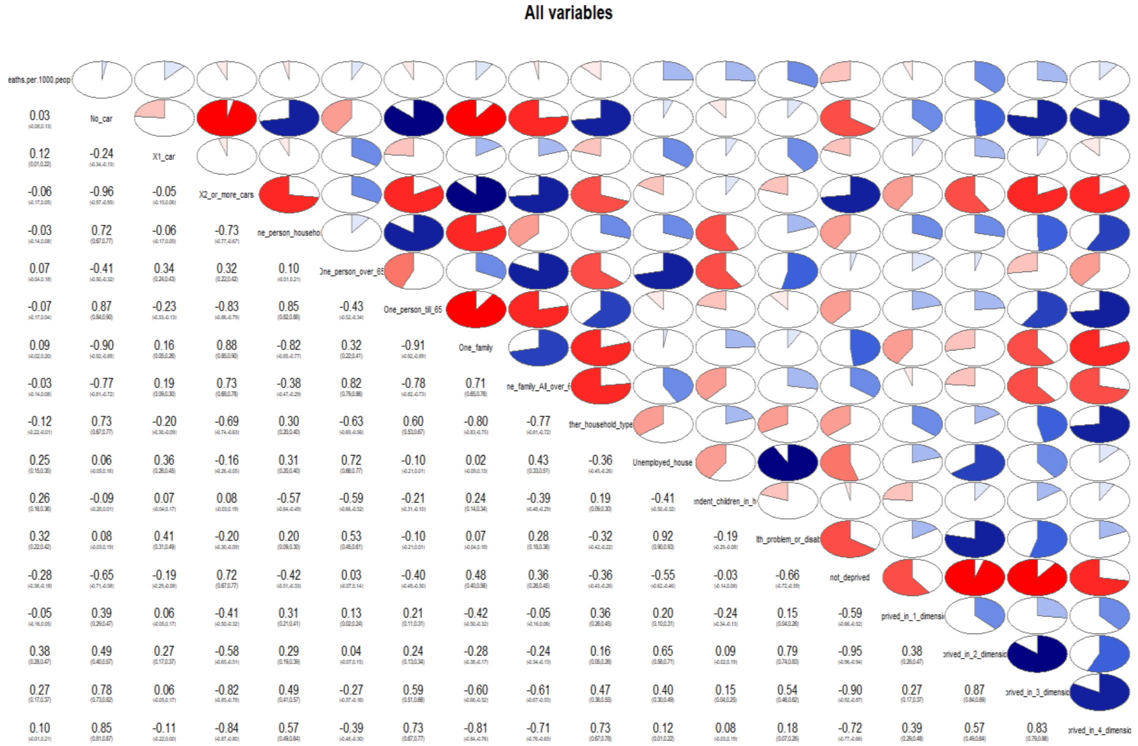

Figure 5 - Corrgram of all variables

There are numerous noticeable pie charts and numbers; however, for now, we will specifically focus on the correlation of the dependent variable with others. Based on our observation, we can identify the best variables by selecting those with a correlation of at least 0.08, and we will eliminate all variables with less than 0.08 correlation with the dependent variable. Consequently, we will remove the following variables:

- No_car
- X2_or_more_cars

- One_person_over_65
- One_person_till_65
- One_family_All_over_65
- Deprived_in_1_dimension
- One_person_household

Removing these variables will not significantly impact the end result, so we have excluded them. Additionally, we have temporarily removed the dependent variable as we proceed to the next step.

After this removal, we move on to the next stage in data exploration: Independent Variable Analysis. This involves comparing the correlation of two independent variables and attempting to eliminate one of them using partial correlation values.

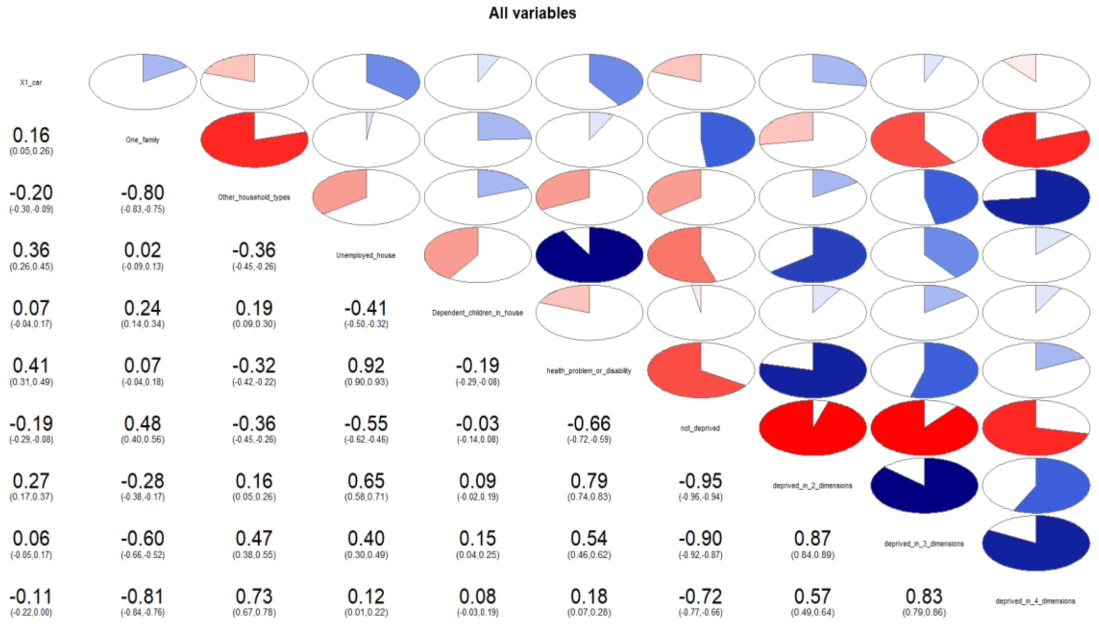

Figure 6-corrgram for independent variable analysis

In this phase, our objective is to pinpoint variables exhibiting higher correlation. Since highly correlated variables exert a comparable impact on our dependent variable, eliminating one should not drastically alter the outcomes.

Upon scrutinizing the corrgram, a notably high correlation (0.92) emerges between Unemployed_household and Health_problem_or_disability. Consequently, one of them can be excluded. To facilitate this decision, a partial correlation test is employed. We compare the correlation of Unemployed_household with Deaths per 1000 people, both with and without Health_problem_or_disability in the dataset. The results inform us about which variable to retain and which one to potentially remove.

In this instance, the first partial correlation estimate is -0.1084005 with a p-value of 0.05197468, while the second one is 0.2290368 with a p-value of 3.336086e-05. Clearly, the result indicates retaining Health_problem_or_disability and contemplating the removal of the Unemployed_household variable.

Subsequently, this process is iterated for other variables with high correlation values, employing partial correlation tests. Post-analysis, we confidently eliminate Unemployed_house, not_deprived, deprived_in_3_dimensions, One_family, and health_problem_or_disability. A new variable, mydata3, is introduced to store the resulting independent variable. To corroborate our findings, we compile a correlation matrix and corrgram for our resulting data.

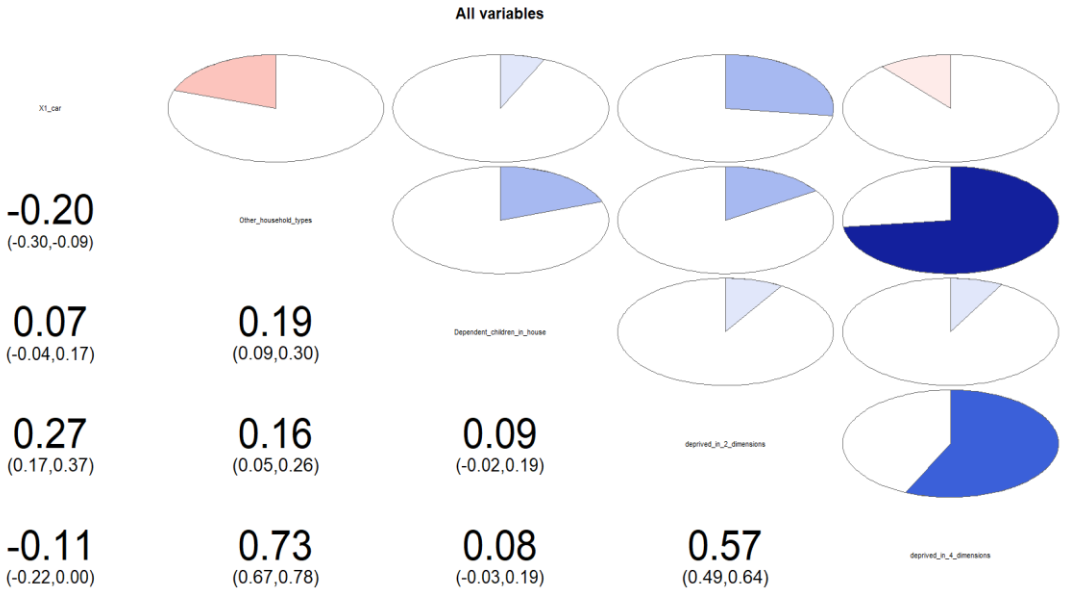

Figure 7- Corrgram after independent variable analysis

The subsequent phase in our analytical journey entails assessing the necessity of factor analysis. If deemed unnecessary, despite its potential benefits, we won't achieve significant variable elimination through this approach. To gauge this necessity, we executed the Kaiser-Meyer-Olkin (KMO) test, deriving the Measure of Sampling Adequacy (MSA). A value exceeding 0.6 indicates the need for factor analysis.

In our specific case, the MSA value fell significantly below the 0.6 threshold at 0.4. Consequently, opting out of factor analysis becomes evident, directing us towards the direct construction of the Multiple Linear Regression Model. The rationale behind embracing this model lies in its ability to offer a comprehensive understanding of how our constructed model contributes to the variance in our dependent variable—deaths per 1000 people.

Our model integrates the dependent variable (deaths per 1000 people) and independent variables, including 1 car, various household types, dependent children in the house, and deprivation across four dimensions. Following model construction using the lm function, we presented a detailed summary of the outcomes.

This summary encompassed residuals presented as quartiles (minimum, 1Q, Median, 3Q, and Maximum), a vital dataset for subsequent analyses. Subsequently, a table outlined all variables with their Estimate value, Standard error, t value, and Pr value (analogous to p value). Significance was gauged by a Pr value smaller than 0.05, and in our case, one variable exceeded this threshold. To validate the necessity of removing that variable, we employed the vif function to calculate the Variance Inflation Factor (VIF) for each variable.

The results further featured R squared and adjusted R squared values, divulging the percentage of variance in the dependent variable elucidated by the linear regression model (approximately 18%). With a p-value of 8.88e-14, significantly less than 0.05, we refrained from rejecting the Null Hypothesis—COVID-19 deaths are significantly linked to the household variables.

To confirm the result validity, we scrutinized if the square root of the VIF value for each variable exceeded 2. In our case, all variables yielded a "False" result, affirming the accuracy of our process and negating the need for additional adjustments.

The final stage involved a meticulous examination of residual values. Employing the Kolmogorov-Smirnov (KS) test to scrutinize whether residuals exhibit normal distribution, we obtained a KS test result of 0.2664, surpassing 0.05, leading to the rejection of the hypothesis. To enhance this assessment, we crafted a histogram with a rug plot and a scatter plot of residuals, providing a visual inspection of normality in the residuals.

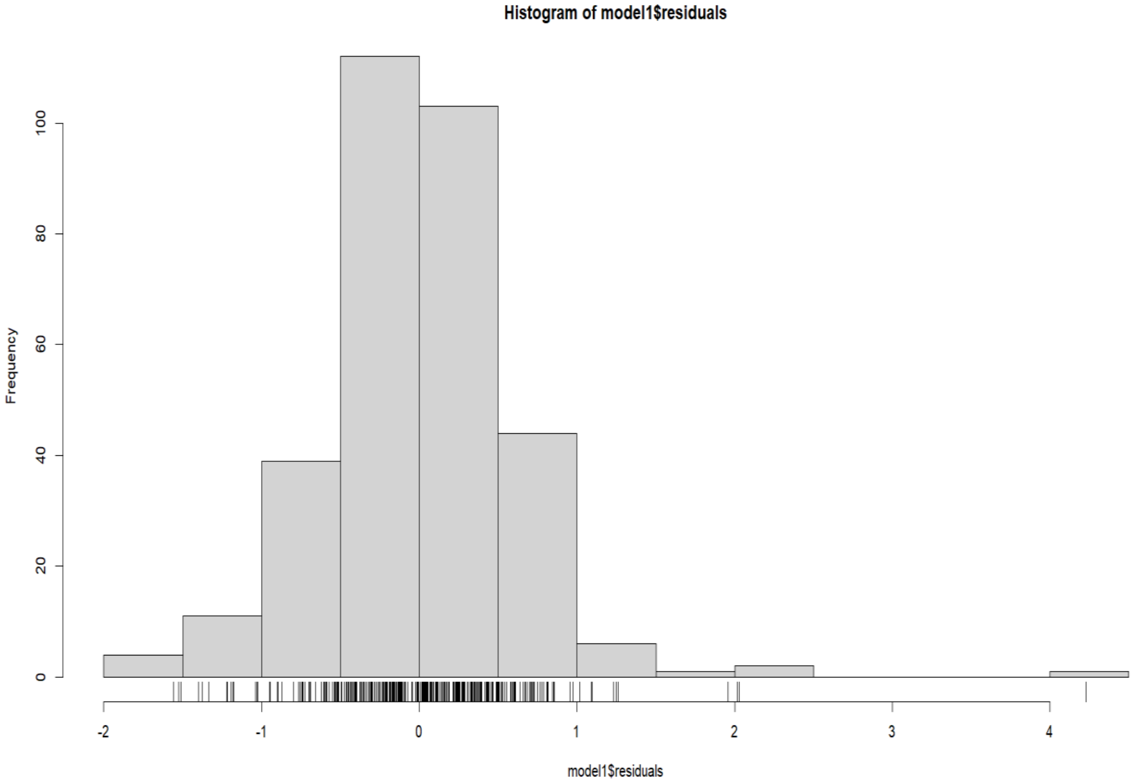

Figure 8- Histogram of residuals

This does not give a sure idea of results but it tends to normality, so we do another test to be sure if the residuals are normally distributed. Next, we did Scatterplot test where we observe that if the scatterplot has a common tendency or not.

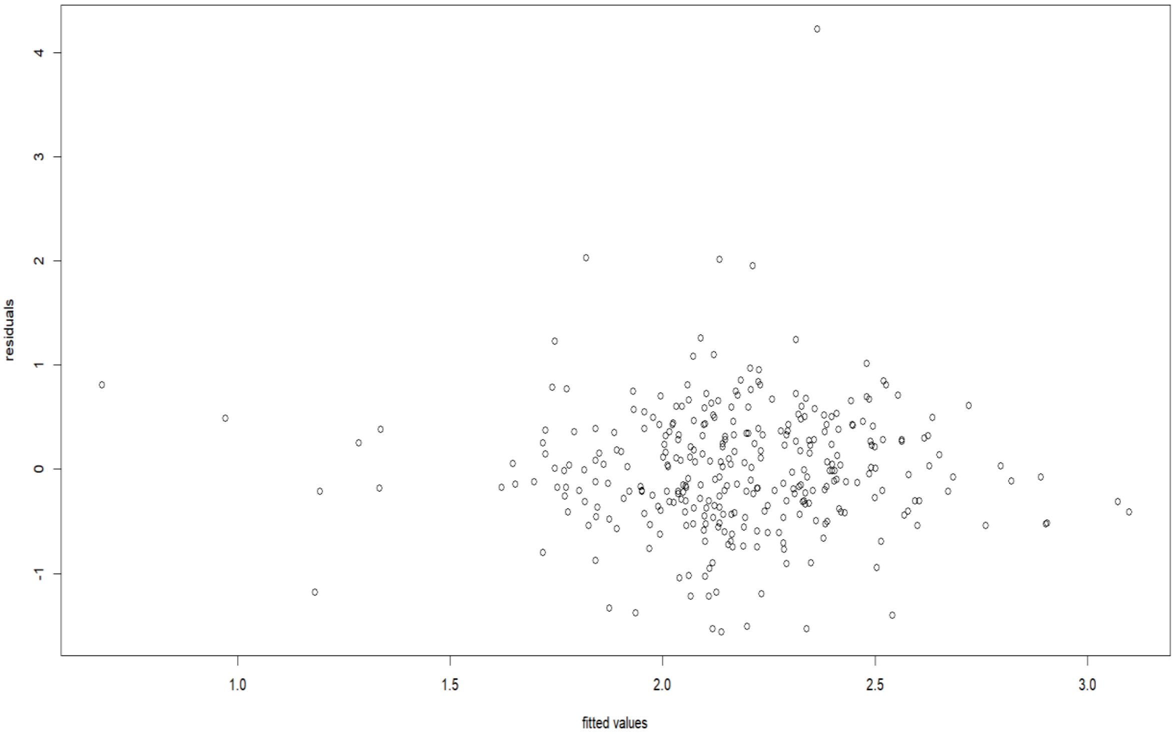

Figure 9 - Scatter plot of residuals

As we can see in this scatterplot, it is randomly distributed. This proves that the residuals are not normally distributed in any way, hence the processes that we did are correct and there is no need to go further.

At the very end, we used relimp function to calculate the importance of each variable to get the variance we got. Here, Dependent_children_in_house has a 0.41 lmg value, that is it contributes 41 percent of our resultant variance.

9 Conclusions

The profound impact of the COVID-19 pandemic on our shared consciousness emphasizes the pressing need to prevent and address potential future health crises. This study delves into the intricate relationship between COVID-19 deaths and various household variables. The null hypothesis, asserting a significant connection between COVID-19 deaths and the examined variables, remains unchallenged based on the obtained p-value. Consequently, it is reasonable to infer that the variables in question wield a discernible influence on the variance of deaths per 1000 people.

Yet, it is imperative to recognize that the impact observed, as denoted by the relatively moderate R squared value, suggests that these variables, in isolation, do not fully elucidate the variability in COVID-19 deaths in England. Unaccounted-for influential factors, omitted in this study, likely contribute significantly to the overall scenario. This underscores the compelling need for further research, incorporating additional variables and a more exhaustive analysis. Such an approach is essential to augment our comprehension of the intricate dynamics that shape pandemic outcomes, offering a more comprehensive and nuanced understanding of the factors at play.

10 References

1. NHS. (2023, March 21). *How to avoid catching and spreading COVID-19*. https://www.nhs.uk/conditions/covid-19/how-to-avoid-catching-and-spreading-covid-19/#:~:text=COVID%2D19%20spreads%20very%20easily,touching%20surfaces%20covered%20in%20them.
2. ONS, U. (n.d.). *Household deprivation variable: Census 2021*. Retrieved December 11, 2023, from https://www.ons.gov.uk/census/census2021dictionary/variablesbytopic/demographyvariablescensus2021/householddeprivation