

|  |
| --- |
| Quantitative Data Analysis  **Analyzing Factors of COVID-19 Deaths Across England: A Quantitative and Statistical Analysis Using Social and Economic Variables**  SUBMITTED TO: DR. YANG LI  SUBMITED BY: ASHISH RAMPATI SHARMA  U2642080  MSc. DATA SCIENCE |

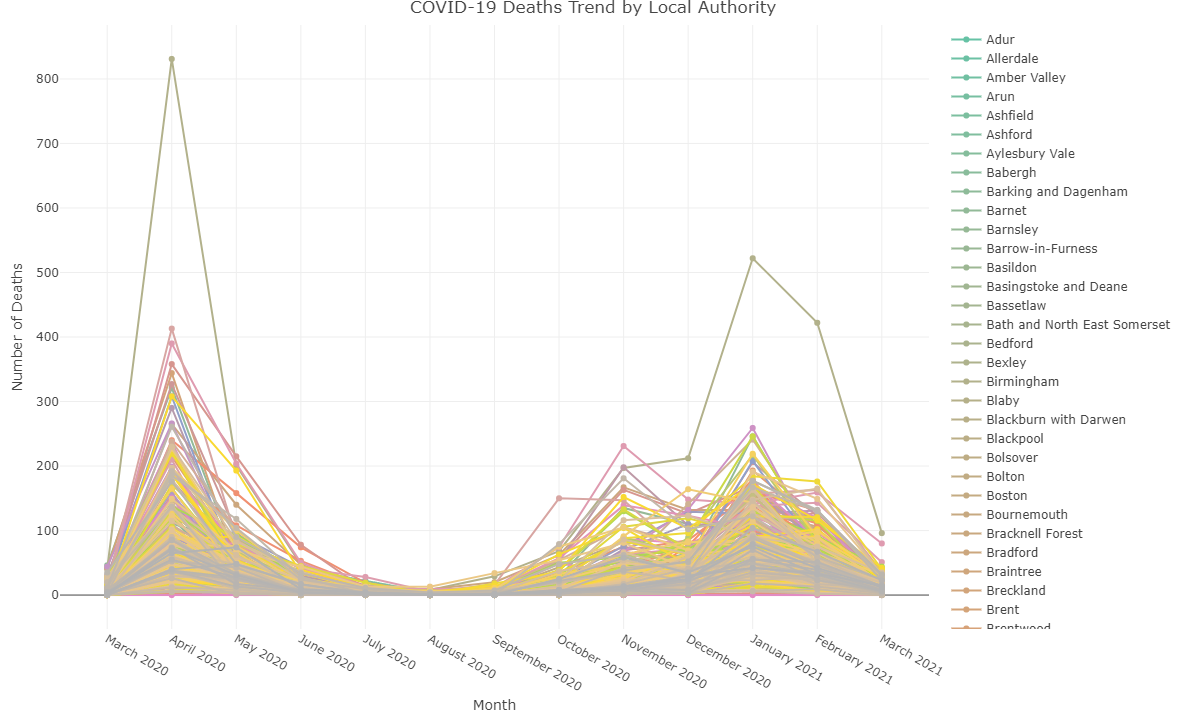


Table of Contents

[1. Abstract 3](#_Toc185360275)

[2. Introduction 4](#_Toc185360276)

[3. Literature Review 5](#_Toc185360277)

[4. Methodology 6](#_Toc185360278)

[4.1. About Dataset 6](#_Toc185360279)

[4.2. Data Collection and Processing 7](#_Toc185360280)

[4.3. Data Analysis Techniques 8](#_Toc185360281)

[5. Exploratory Data Analysis 9](#_Toc185360282)

[5.1. Descriptive statistics 11](#_Toc185360283)

[5.2. Data Visualization 12](#_Toc185360284)

[5.3. Hypothesis Testing 14](#_Toc185360285)

[5.4. Correlation Analysis 16](#_Toc185360286)

[5.5. Factor Analysis 19](#_Toc185360287)

[5.6. Regression modelling 21](#_Toc185360288)

[6. Results 26](#_Toc185360289)

[7. Future Work 27](#_Toc185360290)

[8. Conclusion 28](#_Toc185360291)

[9. References 29](#_Toc185360292)

[Appendix A 30](#_Toc185360293)

Table of Figures

[Figure 1: Missingness Map 11](#_Toc185397994)

[Figure 2: QQ plot for Total Deaths 13](#_Toc185397995)

[Figure 3: Boxplot of Total Deaths 14](#_Toc185397996)

[Figure 4: QQ plot for different variables 15](#_Toc185397997)

[Figure 4: QQ plot for different variables 15](#_Toc185397998)

[Figure 5: boxplots for different variables 15](#_Toc185397999)

[Figure 6: boxplots for different variables 16](#_Toc185398000)

[Figure 7: boxplots for different variables 16](#_Toc185398001)

[Figure 8: Scatter plot for different variables with Total deaths 17](#_Toc185398002)

[Figure 8: Scatter plot for different variables with Total deaths 17](#_Toc185398003)

[Figure 9: KS Test 18](#_Toc185398004)

[Figure 10: Anova Test 19](#_Toc185398005)

[Figure 11: Chi Squared Test 19](#_Toc185398006)

[Figure 12: Correlation for independent variables (Pearson) 20](#_Toc185398007)

[Figure 13: Correlation for independent variables (Spearman) 20](#_Toc185398008)

[Figure 14: Partial Correlation 21](#_Toc185398009)

[Figure 15: Correlation with total deaths 22](#_Toc185398010)

[Figure 16: Correlation with dependent variable 22](#_Toc185398011)

[Figure 17: Correlation with dependent variable(threshold>0.3) 23](#_Toc185398012)

[Figure 18: Scree plot and Cumulative Scree plot for eigen values 24](#_Toc185398013)

[Figure 19: Scree plot and Cumulative Scree plot for eigen values 24](#_Toc185398014)

[Figure 20: Correlation plot for rotated components 26](#_Toc185398015)

[Figure 21: Linear regression model with population density 27](#_Toc185398016)

[Figure 22: Histogram of residuals for linear regression model 28](#_Toc185398017)

# Abstract

This study investigates the factors of COVID-19 mortality, focusing on demographic, health, and socio-economic variables. Using a combination of data visualization techniques, hypothesis testing, correlation analysis, factor analysis, and regression modeling, the following study identifies key predictors of COVID-19 deaths. By the analysis, it is noticed that though population density does not contribute much to mortality rates, health-related factors such as Very\_bad\_health and Bad\_health, together with demographic variables like age (Young and Elderly) and socio-economic indicators, especially Household\_deprivation, contribute significantly to the variation in deaths. Regression modeling leads to a more refined model with these variables that explains a large amount of variation in mortality rates. It further emphasizes that consideration of interlinked factors in COVID-19 mortality modeling could be critical. However, some targeted interventions are needed concerning with health status, socio-economic disparity, and vulnerable populations in order to mitigate the impacts of this pandemic. Further research may include expanding the dataset, including temporal data, and considering various advanced machine learning techniques for more precise accuracy and prediction with high degree performance of models. The findings translate to actionable insights for policymakers desirous of addressing public health disparities and reducing deaths relating to COVID-19.

**Keywords:** COVID-19 Mortality, Statistical Analysis, Socioeconomic Variables, Demographic Factors, Health Determinants, England Pandemic Impact

# Introduction

The COVID-19 Pandemic had a huge impact on health systems, economics, and communities worldwide, which caused significant disruptions in the society. In England, the pattern of covid-19 mortality rate has revealed the regional disparities which are notable and are influenced by the social economic and to some extent demographic factors. Understanding these patterns to understand the underlying factors are essential for creating effective public health strategies, and it can also help in addressing the inequalities created due to the impact of pandemics. This report dives into these differences to identify such underlying causes of the observed variance.

The main objective of this study is to analyse the factors affecting COVID-19 deaths across the different local authorities of England by using many socio-economic and demographic factors. By integrating and examining the data from the census report collected from NOMIS data portal, this analysis aims to provide the insights into how the variables such as population density, deprivation levels, age distribution, modes of travel, worker type and health conditions can correlate with the COVID-19 mortality rates. These findings will help to further our understanding of the pandemic's varied effect and influence policy decisions that promote fairness and resilience in public health interventions.

This study uses a systematic approach that integrates the data using excel and later the integrated data is used in the R for analysis purpose. The dataset is evaluated for dependability, investigated using descriptive statistics and visualizations, and subjected to hypothesis testing and regression modelling. These methods enable us to perform a thorough analysis into the relation between the COVID-19 deaths and a range of independent variables. Special emphasis is placed on factors that reveal socioeconomic inequities, providing a thorough examination of the pandemic's complex dynamics.

The research is guided by the following key questions:

* Which social and economic variables have the most significant correlation with COVID-19 deaths across local authorities in England?
* How do demographic factors, such as age distribution and population density, influence mortality patterns?
* What roles do health-related variables, including self-reported health status and deprivation levels, play in explaining variations in COVID-19 mortality?
* How robust are the statistical models in predicting COVID-19 mortality based on the selected variables?

In addition, Addressing the research questions, this report also highlights the broader societal implications of the findings by emphasizing on the need of targeted interventions in public health, and it also stresses on emphasizing the importance of resource allocation. The integration of statistical techniques, including regression modelling and normality tests, ensures that the conclusions drawn are both scientifically valid and practically relevant.

This work adds to the increasing corpus of COVID-19 research by giving a detailed, data-driven view of mortality trends, resulting in meaningful findings for policymakers, public health professionals, and academics.

# Literature Review

COVID-19 took the world by surprise by highlighting the reasonable disparities in the health outcomes, which are influenced by the socio-economic factors. Many studies had emphasized on how the social determinants for example, ethnicity, income, housing conditions and access to healthcare; can play the key role when it comes to understanding the infection and mortality rates caused by pandemics. In order to make efficient policies, understanding these factors is very crucial.

In the year 2020, the research by Filipa Sá examines the COVID-19 mortality across the England and Wales. This study refers that there is some correlation between the occurrence of deaths caused by the COVID-19 virus and the socioeconomic characteristics of the diseased. Areas with older populations, larger ethnic minority groups, and poorer self-reported health had higher mortality rates. Sá's results emphasized the necessity for improved housing and public health policies to reduce these inequities. (Sá, 2020)

Similarly, , a rapid review published in PLOS ONE synthesizes global studies, noting that race, income, housing status, and employment conditions significantly influence COVID-19 outcomes. The study highlights that those in low-income and overcrowded homes had higher risks, which were aggravated by inadequate access to healthcare and social assistance. (Upshaw *et al.*, 2021)

Another study from BMJ Open (2021) focuses on institutional disparities and their effects on health outcomes during the outbreak of the epidemic. It reveals that death rates were disproportionately higher among those from ethnic minority groups and deprived neighbourhoods. The authors highlight that underlying health conditions, occupational exposure, and limited healthcare access compounded the risks for these populations

An analysis by Bambra in 2020, published in The Lancet Public Health, which explores the "syndemic" nature of COVID-19, where socioeconomic disadvantage intersects with chronic health conditions, and how together they can intensify the pandemic's impact. (Bambra *et al.*, 2020). Evidence shows that COVID-19 infection rates are much greater in disadvantaged areas than in more prosperous places. The authors propose systemic adjustments to address these overlapping vulnerabilities through fair healthcare policy and social protection measures.

In the year 2021, The findings of (Mena *et al.*, 2021) in Nature Human Behaviour reveal that COVID-19-related deaths were strongly correlated with socioeconomic deprivation factors and pre-existing health inequalities. This study also talks about the long-term consequences of these gaps and recommends for a more focused public health policies to identify and address the underlying causes.

These highlights the significance of social and economic factors influencing COVID-19 deaths. They present compelling proof that efforts focusing on housing, healthcare access, and socioeconomic inequities are crucial to minimizing the pandemic's impact. Moving ahead, authorities must use these findings to promote equal health outcomes during future public health crises.

# Methodology

## About Dataset

Data for this analysis comes mainly from publicly available, with a key focus on the Census 2021 published by the Office for National Statistics (from NOMIS portal). It contains detailed information on several socio-economic factors in England that are of relevance to understanding the broader impacts of COVID-19.

The variables selected for this analysis were guided by their relevance to both the spread and severity of COVID-19. These include:

* **COVID-19 Deaths:** This is a categorical variable of the mortality due to COVID-19 across various regions in England, and it is a record from Public Health England.
* **Employment Activity:** The information relating to employed, actively seeking work, or being economically inactive gives an idea that this could be one of those factors relating to COVID-19 deaths. The essential workers in contact professions are more liable to the virus.
* **Hours Worked:** Part-time versus full-time was used as a control for the number of hours worked per week. Full-time, frontline workers were more likely to be exposed to the virus, hence making this variable critical.
* **Travel Distance to Work:** Commuting patterns during the pandemic highlighted how travel behaviour might contribute to exposure to the virus. The more people travelled in congested public transportation systems, the greater their risk of infection.
* **Working from Home:** This helped us to understand the level of remote work that was taking place within the pandemic and therefore contrast those people with others who remained commuters working in close proximity to other people.
* **Demographics:** These were analysed because of the prior knowledge that age and gender affect the severity of outcomes in COVID-19. More aged populations and populations of certain genders faced higher risks than others, hence the interest in analyzing them.

These variables are chosen to provide a meaningful insight into how different socio-economic and demographic factors could influence the spread and severity of COVID-19 deaths. In addition, these variables reflect the key risk factors and behaviours identified in prior research.

For the dataset and meta-data please refer to Appendix A 1 & 2 respectively.

## Data Collection and Processing

The Census 2021 data provided a rich source of information on the social, economic, and demographic factors in England, with data on economic activity, commuting, and household characteristics. This was combined with data from Public Health England on the number of COVID-19-related deaths by region. For the count of total deaths occurred by covid-19, we had the data distributed by months and total, but for this study total count has been used. However, you can check the distribution of covid 19 deaths distributed by months in Appendix-A 5. The data is merged using excel v-lookup and saved as CSV file. As the Local district boundaries have changed over the years, the covid date variable is adjusted as per the updated boundaries using the excel functions.

Afterwards, data preprocessing and cleaning were done accordingly by analyzing the missing and duplicate values, ensuring categorical variables consistency, and transforming a few of them into analysis ready formats. For instance, employment status was categorized to have clear labels such as actively employed without full time students, actively employed as full-time students and economically inactive for better clarity on categorizing into respective categories and making them comparable easily.

## Data Analysis Techniques

. The dependent variable in this study was the total number of COVID-19 deaths, whereas the independent variables included socio-economic factors such as employment status, commuting distance, working hours, age, and gender, among others, including Population\_density, Household\_deprivation, Household\_not\_deprived, Young, Adult, Elderly, On\_foot, Public\_transport, personal\_vehivle, Very\_good\_health, Good\_health, Fair\_health, Bad\_health, Very\_bad\_health, Lives\_in\_a\_household, Lives\_in\_Communal\_Establishment, Travel\_for\_work, Works\_mainly\_from\_home, Works\_offshore, EA\_Ex\_FTS, EA\_FTS, Economically\_inactive, Part\_time, and Full\_time. Various statistical and analytical methods were employed in the analysis: starting with descriptive statistics, just to understand the general distribution of the overall data and the trends within them; followed by the development of histograms, boxplots, and scatter plots to expose relationships and distributions. Hypothesis testing was then used to confirm the relationships observed in previous steps and to establish the statistical significance of those relationships. The correlation analysis was done to check the strength and direction of the relationship between socio-economic factors and COVID-19 death rates, which helped in identifying the key associations. Further, factor analysis and multivariate regression analysis were done to assess the interaction and contribution of multiple independent variables toward COVID-19 death rates, controlling for other factors. This regression model helped predict the possible impact that changes in these variables may have on COVID-19 deaths by showing the most influential factors to explain mortality

# Exploratory Data Analysis

The R-script used in the analysis can be found in Appendix A 3.

The data set has 296 rows and 27 columns. Out of 27 columns, 2 columns have nominal data and remaining 25 columns have numeric. After having the first look at values from the numeric columns, it is seen that the precision formatting is not same for every variable. Hence, to overcome this the rounding of variable has been done. The formula used for the manipulation is as follows.



The second observation was that the column names for 4 columns i.e. Works\_mainly\_at\_an\_offshore\_installation.\_in\_no\_fixed\_place.\_or\_outside\_the\_UK, Economically\_active\_.excluding\_full\_time\_students., Lives\_in\_a\_communal\_establishment and Economically\_active\_and\_a\_full\_time\_student are very long, which is not easy for interpretations. Hence, the name for these 4 variables has been changed and renamed. After the change Works\_mainly\_at\_an\_offshore\_installation.\_in\_no\_fixed\_place.\_or\_outside\_the\_UK is renamed as Works\_offshore, Economically\_active\_.excluding\_full\_time\_students. as EA\_Ex\_FTS, Economically\_active\_and\_a\_full\_time\_student as EA\_FTS and Lives\_in\_a\_communal\_establishment as Lives\_in\_CE.

A white paper with black text

Description automatically generated

The 2 nominal columns have the character data, and they are converted to factors for ease of working with the dataset. There are no missing values in any of the variables, and there are no duplicate rows. The missingness map below confirms the same.

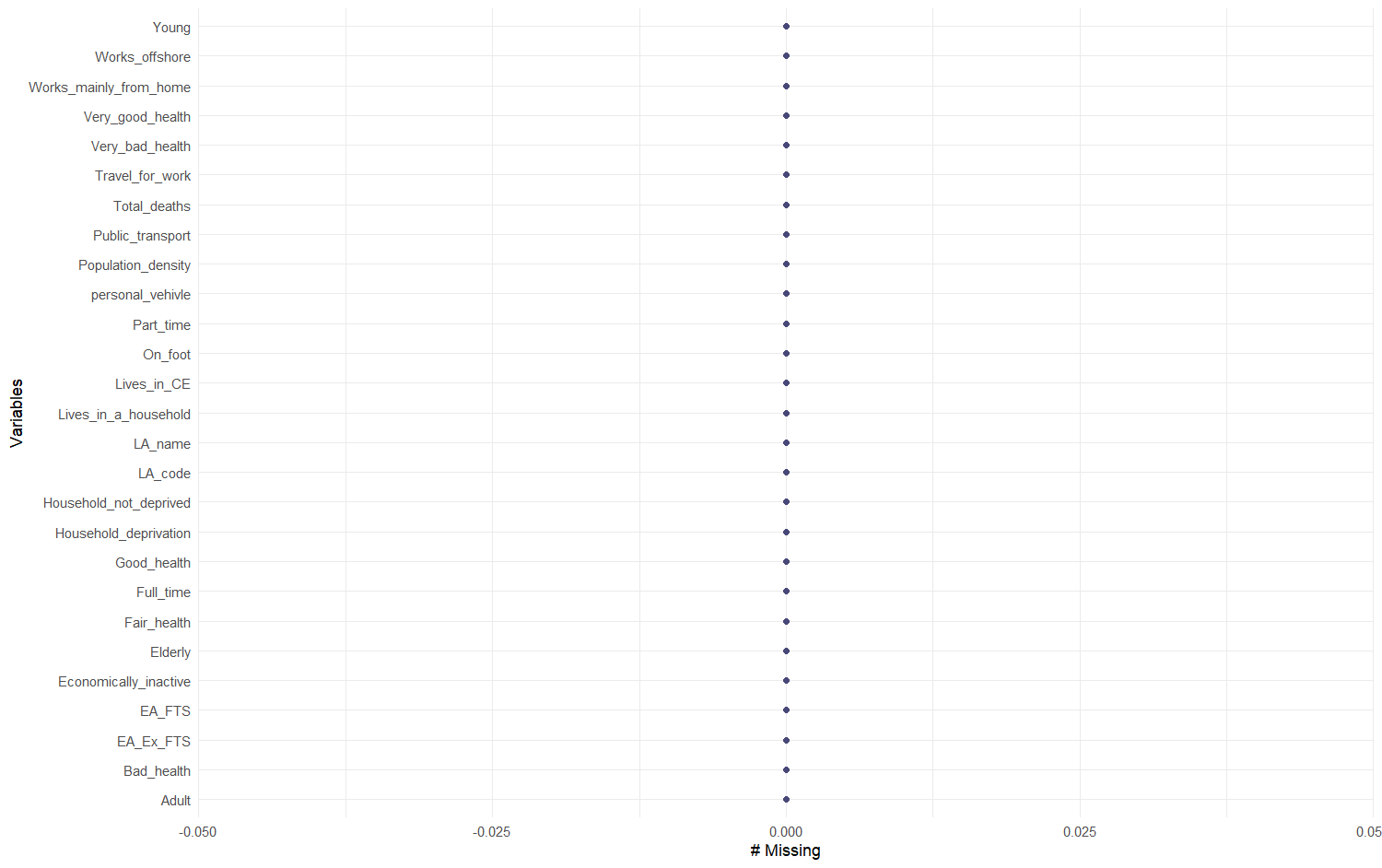
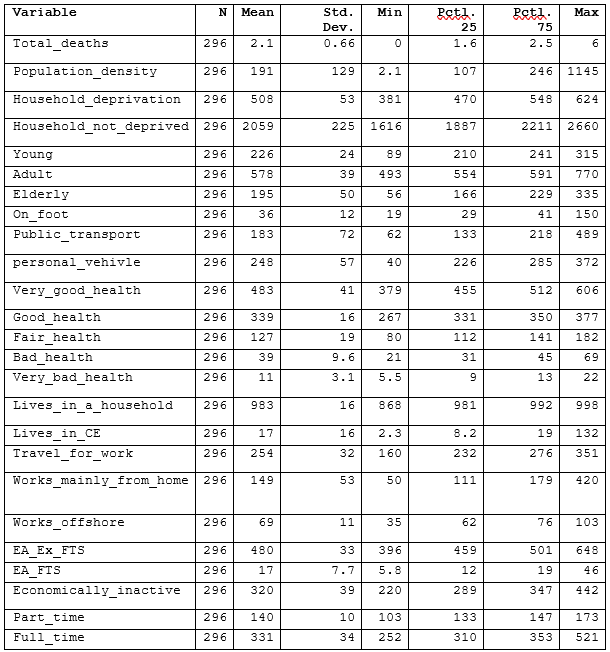


Figure : Missingness Map

## Descriptive statistics

The summary statistics provided include a range of demographic, health, and employment-related variables. The table below shows the summary statistics for all variables.



Based on the summary statistics Total\_deaths, Young, Adult, Elderly, Very\_Good\_Health, Good\_Health, Bad\_Health and Economically\_inactive are the variables that may be normally distributed. However, in order to confirm this QQ plots and Kolmogorov-Smirnov test for normality should be performed.

## Data Visualization

In order to test the normality of dependent variable, the QQ plot and histogram can be used.

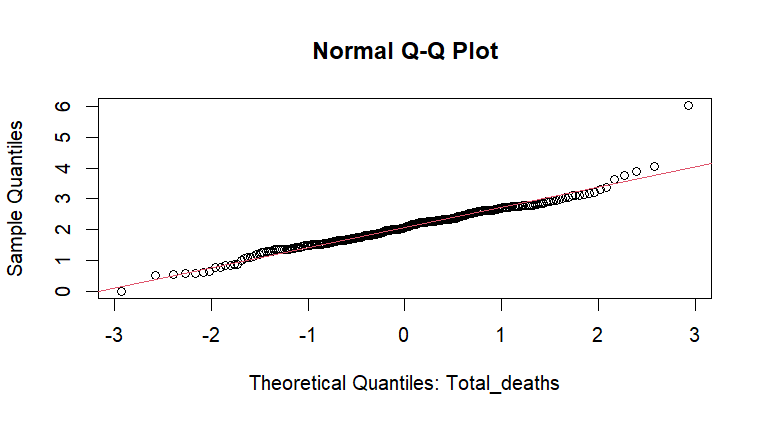


Figure : QQ plot for Total Deaths

The above plot represents that the variable is fairly normally distributed with some outlier at the extremes. To solidify this, the KS test is performed and the results from both tests are shown below. The null hypothesis for the test is that the distribution of Total deaths variable does not significantly differ from normal distribution.

A close-up of a computer code

Description automatically generated

Since the p-value of 0.5072 is grater that 0.05 threshold, we ca safely reject the Null hypothesis and state that the variable is fairly normally distributed. To understand this further the histogram and boxplots is used to check the distribution visually and check for outliers. The figures below confirms that the data is normally distributed. However, there are 4 outliers.

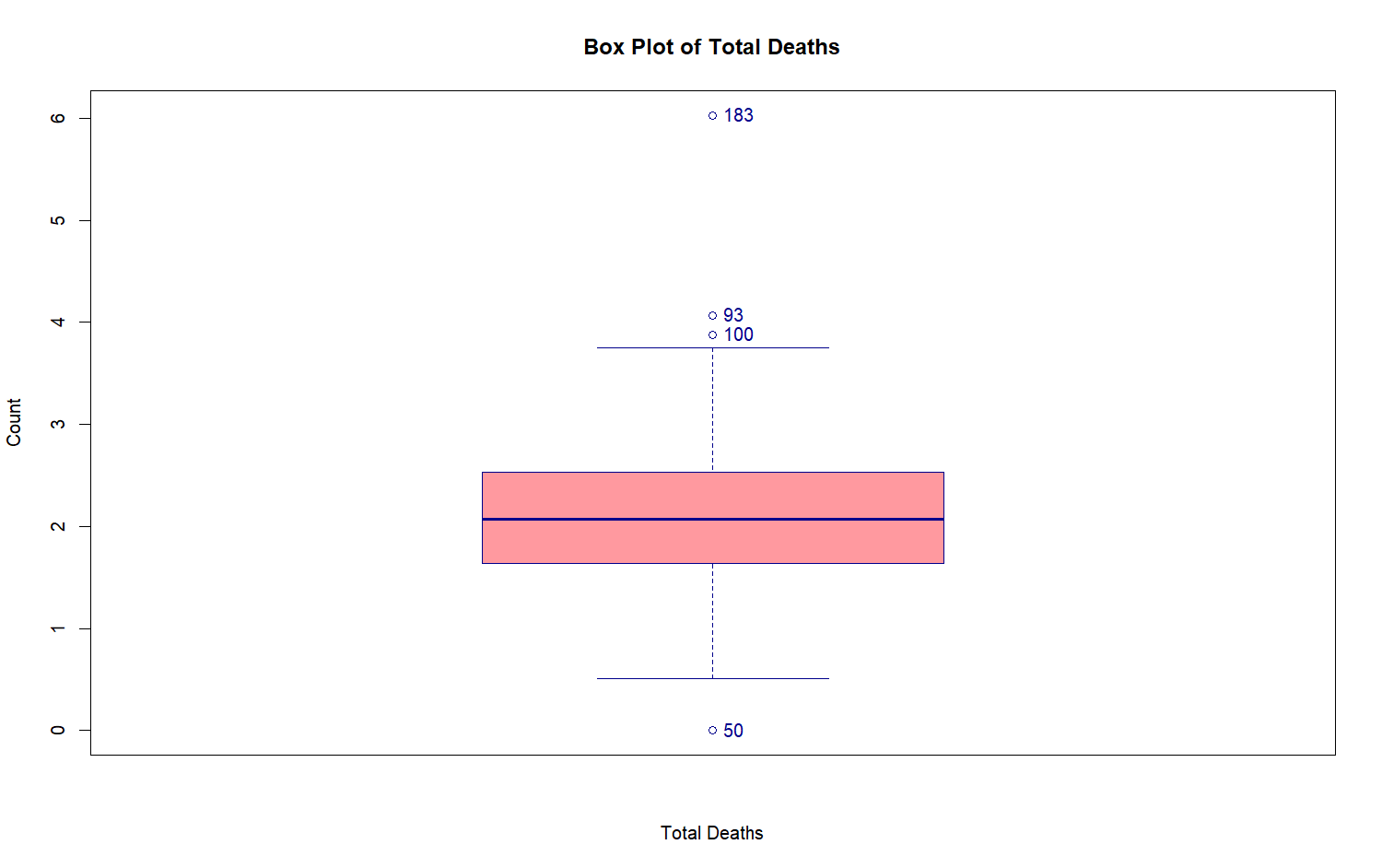


Figure : Boxplot of Total Deaths

A computer code with red lines and black text

Description automatically generated

There is an alternative approach that can be used to find the outlier using inter quartile range. Please refer to section 3 in r-script.

We cannot remove the outlier as the data is related to districts and it may discover the underlying impacts. In order to reduce the impact of outlier different normalization, techniques have been applied (Patel and Mehta, 2011).

The QQ plots for few independent variables, boxplots and Scatter plots with dependent variables to find if there is any visual correlation among the variables are shown below.Fromt these we can understand the distribution of each independent variable and their relationship with the dependent variable.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  | Figure 4: QQ plot for different variables |

Figure 4: QQ plot for different variables

|  |
| --- |
| Figure : boxplots for different variables |
|  |
| Figure : boxplots for different variables |
|  |
|  |
| Figure : boxplots for different variables |
|  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  | Figure 8: Scatter plot for different variables with Total deaths |

Figure 8: Scatter plot for different variables with Total deaths

## Hypothesis Testing

The KS test has been performed for each variable to assess which variables are normally distributed. The results are as follows:

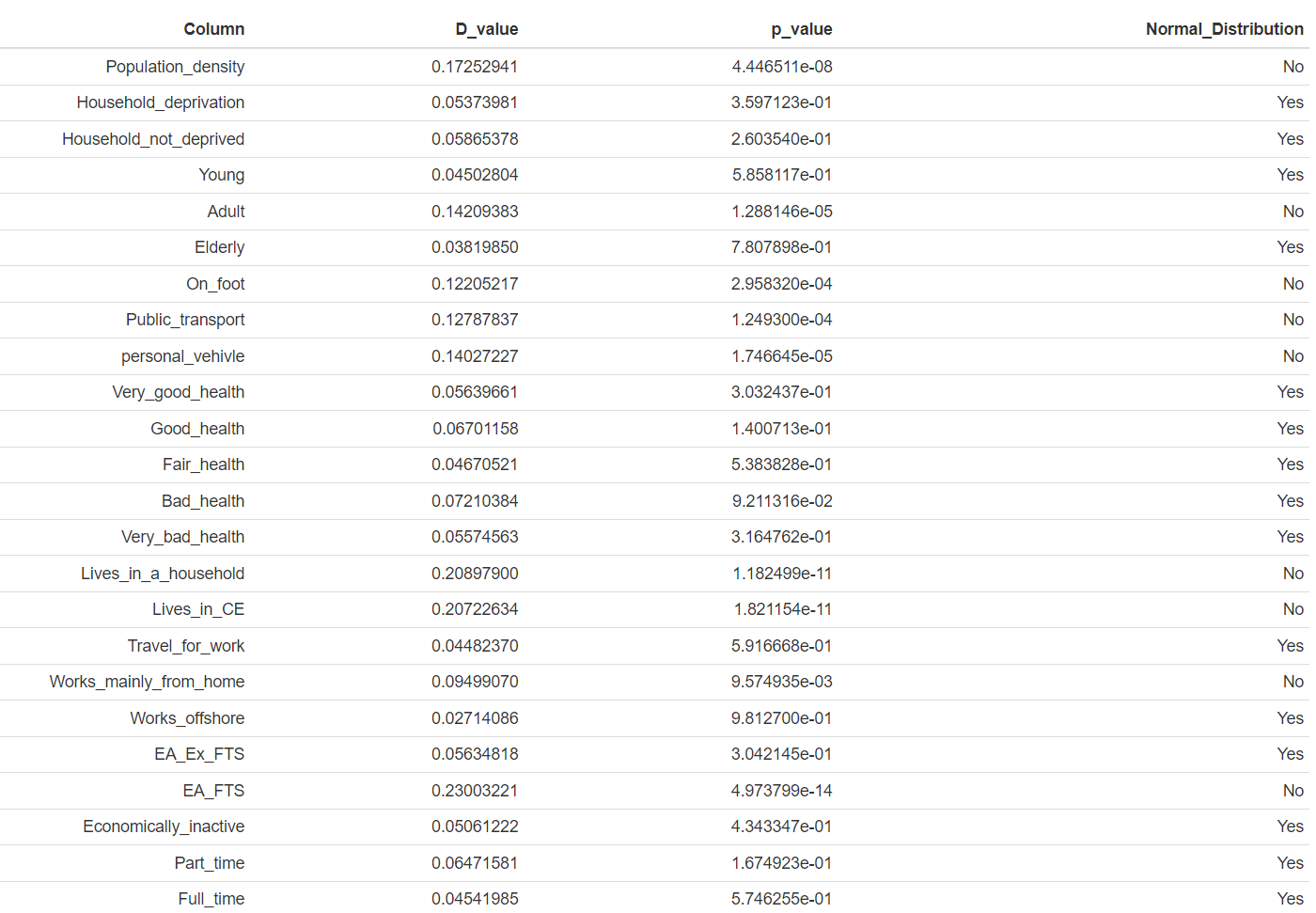
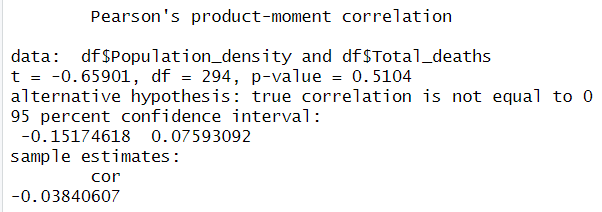


Figure : KS Test

Please refer to the Section 6 in the R-script.

The hypothesis that population density influences COVID-19 mortality is not substantiated, since the Pearson correlation test reveals no significant association between population density and total deaths (p = 0.5104; correlation = -0.038).



However, the hypothesis that the distribution of deaths differs significantly across health status, household deprivation, travel modes, work modes, and economic activity is accepted, with all ANOVA tests yielding significant p-values (p < 0.05).

A screenshot of a computer

Description automatically generated

Figure : Anova Test

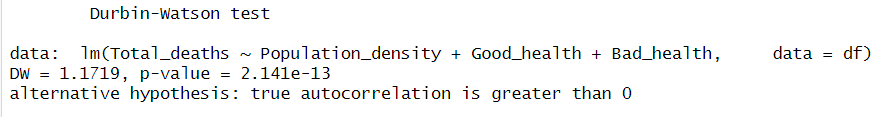
In contrast, chi-square tests show no significant relationships between household deprivation and health status, travel modes and economic activity, employment location and health status, age group and household type, or economic activity and health categories (p > 0.05).

A screenshot of a computer

Description automatically generated

Figure : Chi Squared Test

Lastly, the hypothesis of autocorrelation in total deaths is accepted, as the Durbin-Watson test indicates significant positive autocorrelation in residuals (p = 2.141×10−13 which is greater than 0.05 threshold).



## Correlation Analysis

In order to perform the correlation analysis, the new data frame has been created using only numerical columns. The pairwise correlation analysis has been performed for all numeric variable excluding the target variable using Pearson and spearman methods. Pearson method is usually preferred in case of numeric variables. However, there is no significant difference between the outputs of both methods.

|  |
| --- |
| Figure : Correlation for independent variables (Pearson) |
|  |
| Figure : Correlation for independent variables (Spearman) |
|  |

An alternative approach was made to perform the correlation analysis after scaling the data points in the data frame but the difference in results was not significant.

Based on the results Very\_good\_health, Public\_transport, Economically\_inactive, EA\_Ex\_FTS, personal\_vehicle, Household\_deprivation, Elderly, Works\_offshore, and Lives\_in\_CE can be removed as they are highly correlated with other variables.

However, to solidify this claim, partial correlation analysis has been conducted on the same dataset. The results for the partial correlation analysis shows that Household Deprivation, Bad Health, Very Bad Health, Adult, Elderly, Population Density and Personal Vehicle can be removed, and the explanation power of the model will not suffer much.

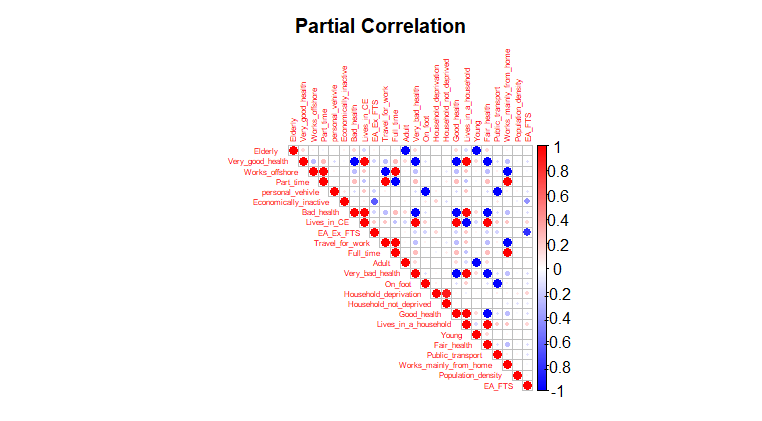


Figure : Partial Correlation

Later the correlation of these independent variables is conducted with the dependent variable to understand how many variables contribute to explain the total deaths.

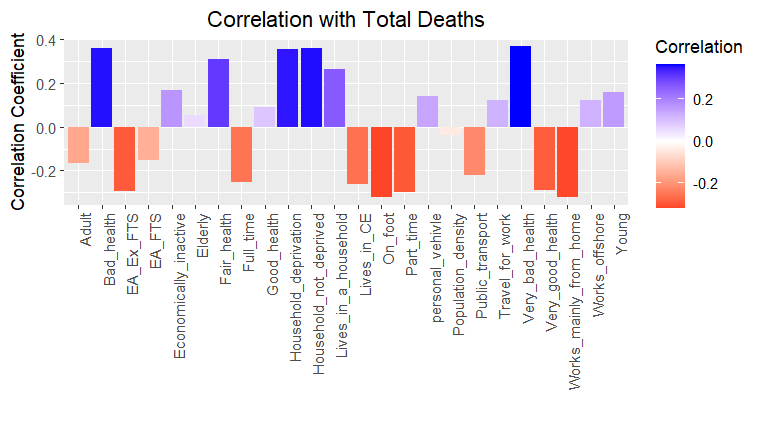


Figure : Correlation with total deaths

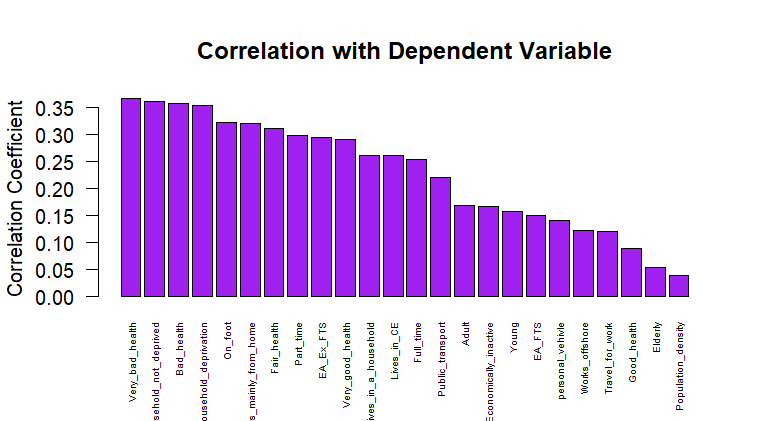


Figure : Correlation with dependent variable

The threshold of 0.3 was considered and only 7 variables can be selected are shown in the figure below.

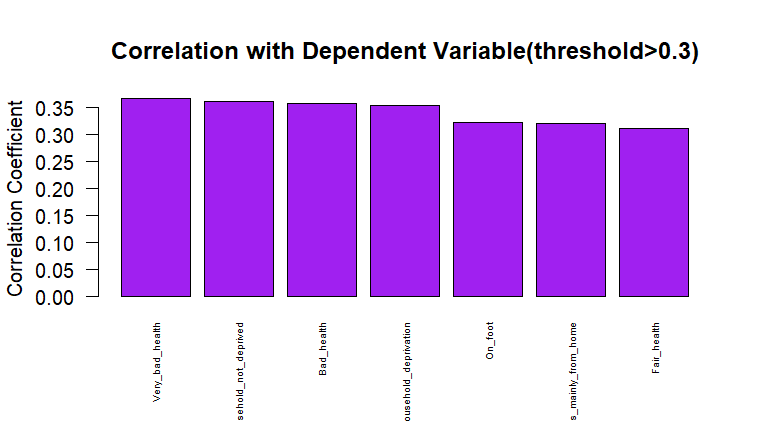
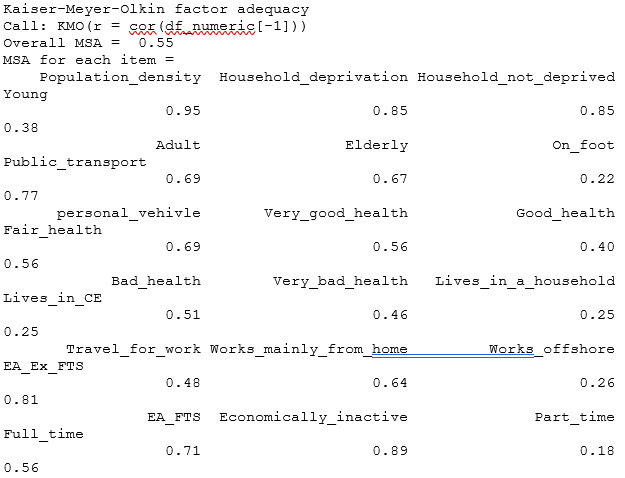


Figure : Correlation with dependent variable(threshold>0.3)

## Factor Analysis

In order to check if the Factor Analysis is required for this dataset, the Kaiser-Meyer-Olkin (KMO) test has been performed and the threshold was set to be 0.6, and overall MSA value exceeds 0.55, which mediocre in this case and hence, we can perform the Factor Analysis. Please refer to the Section 9 in R-script.



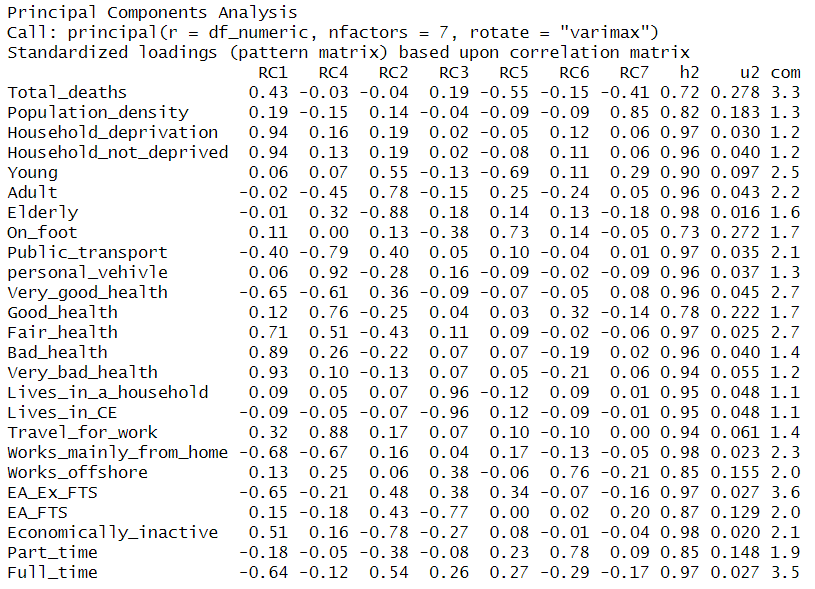
Based on MSA values, some of the variables have a low MSA score. Consider excluding these from further analysis since Young, On\_foot, Lives\_in\_a\_household, and Lives\_in\_CE are not suitable for factor analysis, therefore lowering the overall adequacy of your factor analysis model. This would enhance the robustness and validity of the factor extraction process.

|  |  |
| --- | --- |
|  | Figure 18: Scree plot and Cumulative Scree plot for eigen values |

Figure : Scree plot and Cumulative Scree plot for eigen values

Based on the scree plot generated after the eigen value analysis, it is can be concluded that the 7 variables can be used to make the new rotated factor components for Analysis.

The output for principal component analysis is as follows:



The factor analysis identified several key components representing different socio-economic factors. Component 1 (RC1) reflects health deprivation, with strong loadings on variables like Household\_deprivation and Bad\_health. Component 2 (RC2) captures age group proportions, with a negative correlation between the Elderly and younger age groups. Lives\_in\_a\_household and Lives\_in\_CE form RC3, contrasting household and communal living. Component 4 (RC4) describes commuting patterns, while RC5 highlights physical mobility, particularly walking. Other components include RC6 and RC7, which correspond to job types and urbanization, respectively. Variables such as Household\_not\_deprived, Very\_good\_health, EA\_Ex\_FTS, Full\_time, and Lives\_in\_a\_household are suggested for removal based on the criteria of multicollinearity, communalities, and interpretability. These variables are either redundant, have low contribution to variance, or are highly complex, which obscures clearer interpretations. In addition to this, a new data frame is created for rotated components after removing the suggested variable.

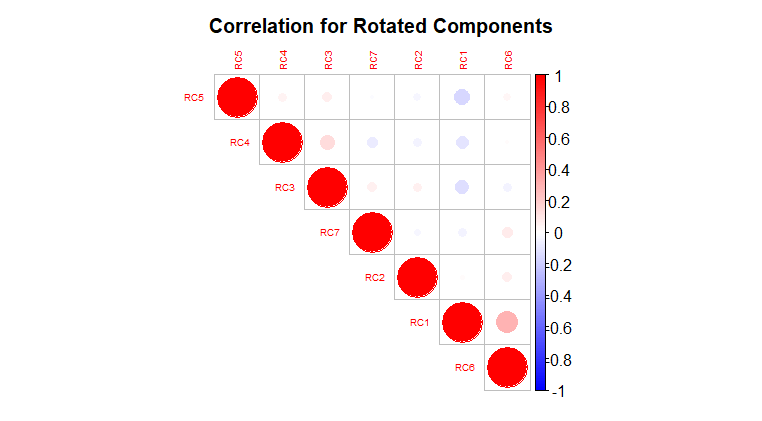


Figure : Correlation plot for rotated components

## Regression modelling

Please refer to Section 9 & 10 in Appendix A 3 R-Script.

A simple linear regression model has been created for total deaths with population density and the summary of model is shown below.

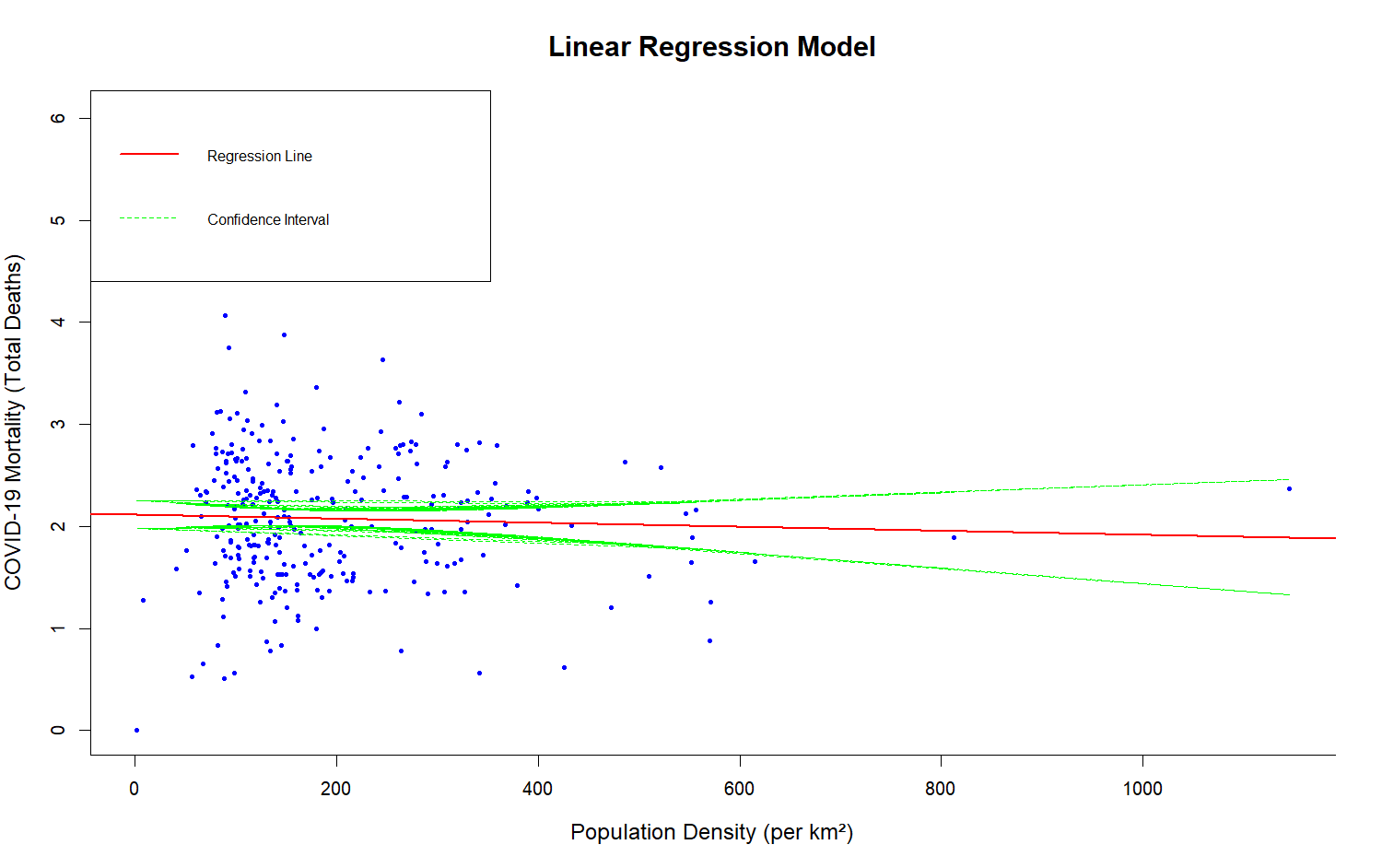


Figure : Linear regression model with population density

A screenshot of a computer

Description automatically generated

The model has a very poor R-squared value 0.001475, indicating only 0.15% of variability in Total\_deaths could be explained by Population\_density. The adjusted R-squared is negative, -0.001921, indicating that the overall model does not improve over just using the mean as a predictor of values. The F-statistic has a value of 0.4343, and the associated p-value of 0.5104 implies that this model is overall not significant. In other words, Population\_density does not seem to be a significant predictor of Total\_deaths in this model.

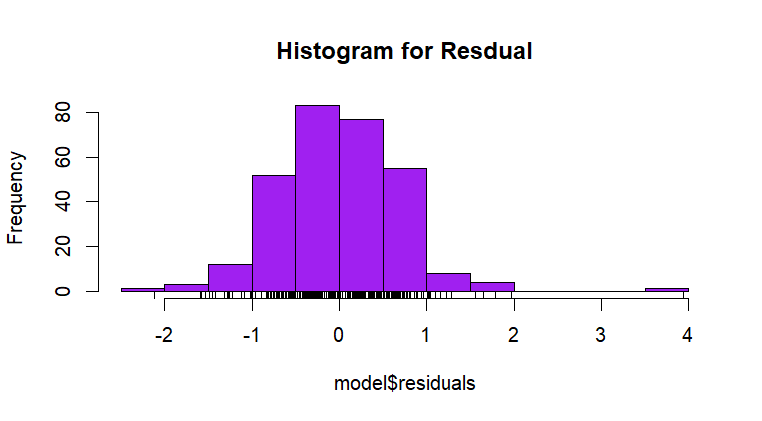


Figure : Histogram of residuals for linear regression model

In order to explain covid deaths in a better fashion. Multivariate linear regression with all 26 independent variables has been created, and the model summary is as follows.

**A screenshot of a computer

Description automatically generated**

The variance influence factor test suggests that there is multi collinearity amongst the factor used in the model.

A screenshot of a computer program

Description automatically generated

In similar fashion, the alternative model is also created with the highly correlated variables to understand which, model is better. The summary of the model is as follows.

A screenshot of a computer

Description automatically generated

The overall model explains a moderate proportion of the variation in death rates, suggesting room for improvement by adding other relevant predictors or addressing multicollinearity.

An attempt to improve the predictive power for model, the partial correlation was conducted on the variables used in model 2, and the result of partial correlation confirms that the variables such as Young, Elderly, Part\_time, and Travel\_for\_work have important relationships with Total\_deaths and should be considered in any further analyses.

A screenshot of a computer

Description automatically generated

The linear regression model explaining the relationship of Total\_deaths with predictors (Young, Elderly, Part\_time, and Travel\_for\_work) results in an R² value of 18.98%, showing that the mentioned predictors significantly influence. Both Young and Elderly are positively correlated with Total\_deaths, whereas Part\_time shows a negative relation. On the other hand, Travel\_for\_work is insignificant, with a p-value of 0.5756. Overall, the model is statistically strong (F = 17.04, p = 1.44e-12), with no multicollinearity concerns, as VIF < 2.

A stepwise approach is made to figure out the best model against the first model with all variables. The summary of model is as follows.

**A screenshot of a computer error

Description automatically generated**

Based on the model summary out of 7 variables only 6 variables have the significant importance. Very\_bad\_health doesn’t seem to have any significant impact. R2 value of 34.33% shows that predictors have significant impact on the total deaths. Overall, the model is statistically strong (F = 21.51, p = 2.2e-16), with no multicollinearity concerns, as VIF < 2.

Another model is made with the rotated components generated during the factor analysis step. 7 rotated components are used as independent variables. Only 4 rotated components seem to have significant impact on the total deaths. Overall, statistically model explains the variances in the total deaths quite well. The summary for the model is shown below.

A screenshot of a computer

Description automatically generated

Finally, the stepwise model is created from the rotated components models and below is the summary for the same. Only 3 rotated components seem to best explain the covid mortality patterns.

A screenshot of a computer code

Description automatically generated

# Results

In order to decide the best explainable features from the dataset. The Anova test has been applied against each model to select the best model, as Anova test compares the goodness of fit, which can explain the covid deaths mortality in best way possible.

Please refer to Appendix A 3 R-script section 11 for the Anova tests.

The ANOVA tests below compare different linear regression models to see which ones are best to predict Total\_deaths. The full demographic, health, and socioeconomic model (Model 1) was compared to some simpler models. Model 4 is a reduced version of Model 1 and performed no worse, p = 0.4723, indicating that fewer variables can explain the variation in deaths. However, Model 2, which included additional predictors such as Household\_deprivation, was far superior to Model 3 (p < 0.001), indicating that socioeconomic factors are important. Model 1 and Model 2 were not significantly different from one another (p = 0.07355), but Model 2 was a better balance between predictive power and parsimony. A comparison of Model 4 to Model 2 demonstrated that Model 2 fitted significantly better, p < 0.001, which highlights the importance of including variables such as Household\_deprivation. Comparisons with models featuring reduced components, such as Model 5 and Model 6, showed these models performed similarly to Model 1 but without interpretability. Across all models, the most important predictors of Total\_deaths included health variables: Very\_bad\_health and Bad\_health; demographic factors: Young and Elderly; and socioeconomic variables: Household\_deprivation and Part\_time. These results suggest that targeted interventions in the domain of poor health, socioeconomic disparities, and vulnerable populations can effectively mitigate death rates. Overall, Model 2 was the most balanced model, offering strong predictive capability while retaining interpretability.

# Future Work

It is suggested that in future work, the drivers of total deaths are investigated further, focusing on refinement of model variables and exploring additional predictors not considered in the current analysis. The dataset can be expanded by more geographic regions or more detailed demographic data to get higher model generalizability and provide more insight into regional disparities in health outcomes. Interaction effects between variables, such as the relationship between health status, for example, Very\_bad\_health, and socioeconomic factors, for instance, Household\_deprivation, may further improve the performance and accuracy of the models. Exploring nonlinear models or machine learning algorithms may help in uncovering complex patterns and relationships not captured by linear regression. Incorporating temporal data might also enable the modeling of trends over time, offering predictive insights into future public health outcomes. This will, of course, involve collaboration with healthcare experts and policymakers to refine the models and ensure that the findings translate into actionable public health strategies.

# Conclusion

This study tries to determine the factors of COVID-19 mortality rates through statistical testing, correlation analysis, and regression modeling, considering some important demographic, health, and socioeconomic predictors. The research questions explored were:

1. Does population density influence COVID-19 mortality?
2. How do health status, household deprivation, travel, work modes, and economic activity impact COVID-19 mortality?
3. What are the most significant predictors of COVID-19 mortality, and how well can they explain variations in total deaths?

The analysis showed that population density did not contribute much to COVID-19 mortality, as evidenced by the Pearson correlation test (p = 0.5104) and the linear regression model (R² = 0.15%). In contrast, health, socio-economic, and behavioural factors contributed significantly to mortality. ANOVA tests showed that deaths varied significantly across health status, household deprivation, work modes, travel modes, and economic activity at p < 0.05. These included the variables Bad\_health, Very\_bad\_health, Young, Elderly, Household\_deprivation, and Part\_time as key contributors to mortality.

Regression, stepwise, and factor analysis revealed strong predictors for COVID-19 mortality. Among the significant drivers for mortality were poor health conditions such as Bad\_health and Very\_bad\_health, age factors comprising the Young and Elderly groups, and socio-economic status with respect to Household\_deprivation and employment status regarding Part\_time employees. The final stepwise regression model explained 34.33% of the variance, R² = 0.3433, without any multicollinearity issues. Model comparisons further confirmed that adding socioeconomic variables, such as Household\_deprivation, greatly enhanced the strength of the model and, therefore, these factors are very important to explain mortality patterns.

The findings have critical implications for public health strategies, underlining that health status and socioeconomic disparities are the prime drivers of COVID-19 mortality. Targeted interventions should be directed at improving access to healthcare for people in poor health, addressing socioeconomic vulnerabilities, and designing policies to protect vulnerable demographic groups, especially the elderly and those in precarious employment. These measures can enhance resilience during pandemics and reduce mortality risks.

Future studies should extend these findings by adding other predictors, such as healthcare accessibility and vaccination rates, and interaction effects of health and socioeconomic factors in order to capture more complex relationships. Advanced techniques, such as nonlinear models and machine learning, may improve predictive accuracy. The inclusion of temporal and geographic data would allow analysis of trends in mortality over time and regions.

This study, therefore, has illustrated how health and socioeconomic factors have led the ball in COVID-19 mortality. It brings the attention of policy decisions toward making evidence-based interventions over such mortality risks and further prepare better for future pandemics or public health hazards.

# References

1. Bambra, C. *et al.* (2020) ‘The COVID-19 pandemic and health inequalities’, *Journal of Epidemiology and Community Health*. BMJ Publishing Group, pp. 964–968. Available at: https://doi.org/10.1136/jech-2020-214401.
2. Mena, G.E. *et al.* (2021) ‘Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile’, *Science*, 372(6545). Available at: https://doi.org/10.1126/science.abg5298.
3. Patel, V.R. and Mehta, R.G. (2011) ‘Impact of Outlier Removal and Normalization Approach in’, *IJCSI International Journal of Computer Science Issues*, Vol. 8(Issue 5, No 2).
4. Sá, F. (2020) *Socioeconomic Determinants of COVID-19 Infections and Mortality: Evidence from England and Wales*. Available at: www.iza.org.
5. Upshaw, T.L. *et al.* (2021) ‘Social determinants of COVID-19 incidence and outcomes: A rapid review’, *PLoS ONE*, 16(3 March 2021). Available at: https://doi.org/10.1371/journal.pone.0248336.

# Appendix A

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Final Dataset | <https://uelac-my.sharepoint.com/:x:/g/personal/u2642080_uel_ac_uk/EfF1kRGP_BBBhbE-hvOXeooBVvSKah_ue9bK3WmGiw55iw?e=ssqyeV> |  |
| 2. | Meta-Data | <https://uelac-my.sharepoint.com/:x:/g/personal/u2642080_uel_ac_uk/ESUy0Bmc4GNOgaRnJP8HKXMBcKoyrH_ROQ5Lg3NAGY9AvQ?e=I3SmOM> |  |
| 3. | R-script for Analysis | <https://uelac-my.sharepoint.com/:u:/g/personal/u2642080_uel_ac_uk/EaG_oG3LaQZAtnifXgFLHSoB5qpCv4KD7ILo9zDNqM4ekw?e=q4OmN3> |  |
| 4. | Research Papers | <https://uelac-my.sharepoint.com/:f:/g/personal/u2642080_uel_ac_uk/EiWceiMWuNlAoJEZrYyxGPYBetdDidgd2RUAgw8joTu2jg?e=zgbZQ9> |  |
| 5. | Covid-19 death trends interactive html image | <https://uelac-my.sharepoint.com/:u:/g/personal/u2642080_uel_ac_uk/EVe1a8Um_slLqhe9eiV5dw0B2qKfKGWUZ7Q9jeZA9lBpiw?e=2EpAf5> |  |
|  | R-script for covid deaths trend | <https://uelac-my.sharepoint.com/:u:/g/personal/u2642080_uel_ac_uk/EaRmrNF_j2dKjx8M2ox0Q60Ba5dppL9miMrn3dQBZW5vdA?e=6gjRxJ> |  |
|  | Covid-19 death data by months | <https://uelac-my.sharepoint.com/:x:/g/personal/u2642080_uel_ac_uk/EVETsAXuUXNLr3yuoutwl2cB5bDLsFd2hWlLQRoAcAaYjw?e=u5VgyP> |  |