**Text

Description automatically generated**

**MSc. In Data Science**

**QUANTITATIVE DATA ANALYSIS(QDA)**

**DS7006**

COURSE WORK

BY

DAMILOLA OLUWASENI OMISORE

2272153

ON

SOCIAL AND ECONOMIC VARIABLES ON COVID-19 DEATHS



# **ABSTRACT**

The average death from COVID-19 and influencing factors, which would help inform optimal control strategies, remain unclear. Moreover, studies regarding this issue are limited in England, and no region-wise studies were conducted. Hence, this study aimed to investigate the deaths in England from COVID-19, and its predictors among patients. The large number of COVID-19 deaths in England brings about the hypothesis that there might be other factors that have a connection with the COVID-19 mortality of individuals. The objective of this work is to inspect and analyse the connections between social and economic variables such as age, sex, ethnicity, health, heating, social grade, travel to work and COVID-19 deaths. Each variable was analysed firstly by checking for normality through visualisation and statistical test. The data was then normalised per thousand population which was then ready for analysis. A correlation test was then carried out on all variables to test the relationship each variable has with the other. Before building the regression model, I needed to ascertain if factor analysis would have been needed which the Kaiser-Meyer-Olkin statistics test was used. With the factor analysis not being needed, choosing independent variables to build model was based on the degree of correlation the variables have with each other. The first model was built using all the variables to identify the significant variables in the set. Afterwards, the variables with no significance to first model were taken out and the focus remained the significant variables. At the end of the analysis, it was observed the bad health, age group 20 to 29, no heating in homes, Asians, and middle-class citizens all had a significant connection with COVID-19 deaths. For other models, old age and public transport were also significant factors but due to collinearity between variables, all these variables could not be in the same model and give close to an accurate result.

Table of Contents

[**ABSTRACT** 2](#_Toc121759811)

[**INTRODUCTION** 4](#_Toc121759812)

[**METHODOLOGY** 4](#_Toc121759813)

[**DATA ACQUISITION** 4](#_Toc121759814)

[**DATA CLEANING** 5](#_Toc121759815)

[**RESULT OF ANALYSIS** 5](#_Toc121759816)

[**DATA EXPLORATION** 5](#_Toc121759817)

[**REGRESSION MODEL** 10](#_Toc121759818)

[**CONCLUSION** 18](#_Toc121759819)

[**REFERENCE** 19](#_Toc121759820)

[**APPENDICES** 19](#_Toc121759821)

List of Figures

[Figure 1: A 2-D Cloumn graph of monthly COVID-19 deaths by Local authorities. 6](#_Toc123668356)

[Figure 2: Map of England showing COVID-19 deaths. 6](#_Toc123668357)

[Figure 3: Missingness map for all independent variables. 7](#_Toc123668358)

[Figure 4: Box plot, Histogram, and QQ plot of total COVID-19 deaths 7](#_Toc123668359)

[Figure 5: Box plot, Histogram, and QQ plot of total COVID-19 deaths per thousand 8](#_Toc123668360)

[Figure 6: Multivariate scatter plot of pDeaths and pAge 9](#_Toc123668361)

[Figure 7: Multivariate scatter plot of pDeaths and pHealth 9](#_Toc123668362)

[Figure 8: A correlation plot of all variables 10](#_Toc123668363)

List Of Tables

[Table 1: All variables used for the analysis 4](#_Toc121754117)

List of Boxes

[Box 1: Summary of Total COVID-19 deaths 7](#_Toc121758885)

[Box 2: Summary of Total COVID-19 deaths per thousand 7](#_Toc121758886)

[Box 3: KS test on pDeath 8](#_Toc121758887)

[Box 4: Kaiser-Meyer-Olkins statistical test 10](#_Toc121758888)

[Box 5: Multiple Regression Model using all variables (Model 1) 10](#_Toc121758889)

[Box 6: Multiple Regression Model using 13 variables (Model 2) 12](#_Toc121758890)

[Box 7: Multiple Regression Model using variables correlating with pDeaths (Model 3) 13](#_Toc121758891)

[Box 8: Multiple Regression Model using significant variables with pDeaths (Model 4) 14](#_Toc121758892)

[Box 9: Multiple Regression Model using significant variables with pDeaths (Model 5) 16](#_Toc121758893)

[Box 10: Multiple Regression Model using stepwise approach on Model 1 (Model 6) 17](#_Toc121758894)

# **INTRODUCTION**

Coronavirus disease 2019 (COVID-19) is caused by the virus associated with severe acute respiratory syndrome coronavirus which first appeared in late 2019 in Wuhan, China(WHO, 2020). Nearly 30 million cases were reported as of 16 September 2020, with close to one million deaths, including 41,664 in the United Kingdom (UK), signifying an unprecedented number of critically ill patients and a high demand for critical care services globally(Arabi, Murthy and Webb, 2020; Aziz *et al.*, 2020).

The aim of this project is to discover relationships between logical and common variables/themes on the COVID-19 deaths which is my dependent variable. COVID-19 is mainly contacted by inhaling contaminated air that contains the virus in the form of droplets, aerosols, and small airborne particles. These particles are exhaled by infected individuals while they breathe, speak, cough, sneeze, or sing. The closer people are to one another, the higher the likelihood of transmission. However, infection can spread farther, especially indoors. The themes considered for this analysis are:

1. Age
2. Sex
3. Ethnicity
4. Health status
5. Availability of heating
6. Social class
7. Means of travel to work

The Age theme was considered because biologically, ageing is as a result of the build-up of different types of molecular and cellular damage over time. As a result, physical and mental abilities gradually deteriorate, disease risk increases, and eventually, death occurs. People at old ages are assumed to have very high risks of COVID-19 death(Ioannidis, Axfors and Contopoulos-Ioannidis, 2020). The Sex theme is considered because men tend to have lesser life expectancy than women and global data show that men experience more COVID-19 case fatalities than women(Dehingia and Raj, 2021). The Ethnicity theme is taken into consideration because when it comes to being able to repel Covid, genetics is crucial. In the US, African Americans and other minority groups make up a larger number of COVID-19 deaths(Lamarque, 2020). The most common and predicted factor is the Health condition as the number of COVID deaths in England for people with pre-existing condition from March to June 2020 is 43,640 while for people with no pre-existing conditions from March to June is 4,169(gov.uk, no date). It is common knowledge that coronaviruses can be eradicated by heating up thereby leading to the Heating theme. In fact, as temperature and humidity rise, coronavirus inactivation on surfaces accelerates(Yap *et al.*, 2020). The reason for going with the Social class and Means of travel because, it contributes to social distancing which was declared as one of the major measures to taken during the fight of COVID-19. The selected themes were used to test the hypothesis.

# **METHODOLOGY**

## **DATA ACQUISITION**

The web page that served as the source for the information is “**Nomis**” (<https://www.nomisweb.co.uk/>) which is the UK’s largest independent producer of official statistics. Nomis is a service provided by Office for National Statistics (ONS). Nomis post statistics about the population, society, and labour market at the national, regional, and municipal levels. These comprise information from the most recent and prior censuses. The data retrieved were all from the 2011 census data which is the last census update on the web page. The themes/variables (Age, Sex, Ethnicity, Health status, Availability of heating, Social class, and Means of travel to work) were all downloaded by the Local authority/District administrative geography for England.

## **DATA CLEANING**

Microsoft Excel was then used to carry out the first stage of data cleaning and simplification. For each theme, grouping of columns was necessary as raw data was ambiguous. After inspecting the CSV data downloaded from Nomis, it was observed that the two (2) major columns (LA\_name and LA\_code) that could be used as primary keys for the joining of data had some discrepancies which had to be treated before analysis. For the COVID Death CSV file, on the LA\_name column, “Shepway” which is the former name for “Folkestone and Hythe” had to be changed to “Folkestone and Hythe”(*Shepway - Wikipedia*, no date). While for the various themes/variables, “Bristol, City of” was changed to “Bristol”, “Herefordshire, County of” was changed to “Herefordshire”, “Kingston upon Hull, City of” was changed to “Kingston upon Hull”. SQL was used after to join all themes to make a single data set as shown in APPENDIX 1. DB Browser is the tool which was used to carry out this process. At the end of the cleaning process, the final dataset contained 323 rows with 48 variables and the following are the final variables showing what particular theme they represent:

Table : All variables used for the analysis

|  |  |
| --- | --- |
| **THEME** | **INDEPENDENT VARIABLES** |
| COVID Death | LA\_name, LA\_code, March\_2020 to April\_2021, Total\_Covid\_Deaths |
| Age | Age\_0\_to\_19, Age\_20\_to\_29, Age\_30\_to\_59, Age\_60\_and\_above, Total\_age |
| Sex | Males, Females, Total\_sex |
| Ethnicity | White, Mixed, Asian, Black, Other\_ethnic\_group, Total\_ethnicity |
| Heath status | Good\_Health, Fair\_Health, Bad\_Health, Total\_Health |
| Availability of heating | No\_heating, Heating, Total\_heating |
| Social class | Upper\_class, Middle\_class, Lower\_class, Total\_s\_grade |
| Means of travel to work | Public\_Travel, Private\_Transport, Other\_Transport, No\_Transport, Total\_transport |

A series of data exploration methods was then carried out on the data. Methods like test for normality, standardisation or normalisation of data, test for correlation between all variables, factor analysis requirement test, and variety of regression models built.

# **RESULTS OF ANALYSIS**

## **DATA EXPLORATION**

In total, England has 333 local councils and the total number of local authorities available in the retrieved data is 326 which is still quite significant(gov.uk, no date). Performed a trend check on the monthly data by the Local authorities using Excel to ascertain if there are any patterns or stories.

Figure : A 2-D Cloumn graph of monthly COVID-19 deaths by Local authorities.

The first COVID-19 case in England was registered in January 2020 and the first COVID death was registered in March 2020(gov.uk, no date). According to the above graph, the month of April 2020 appears to have the highest COVID-19 death for most of the local authorities in England with January 2021 following as the next highest. It may be as a result of the sudden discovery of the virus which had already spread to critical points before measures were put in place while for the January spike may be as a result of the second wave.

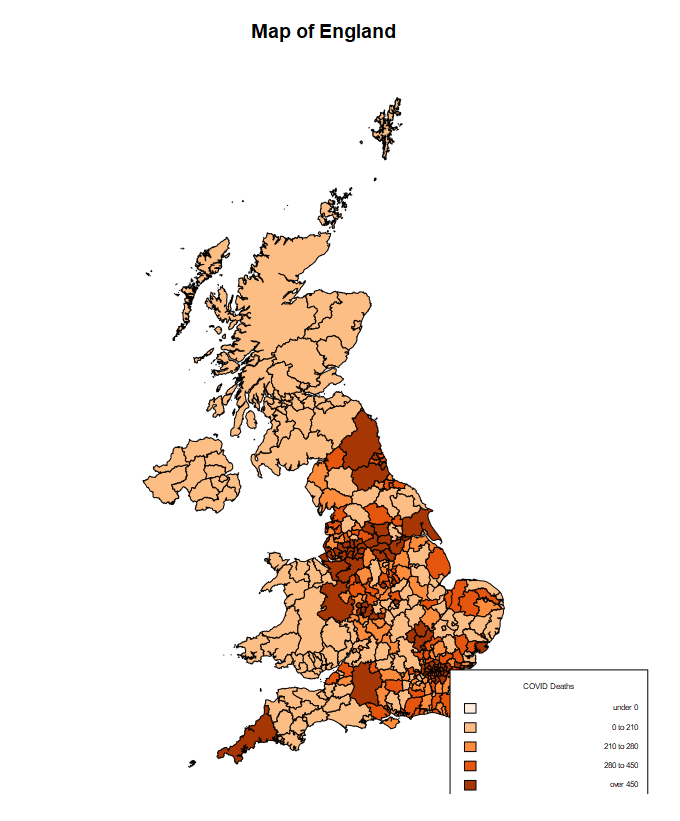


Figure : Map of England showing COVID-19 deaths.

The above is a choropleth map of England visualizing the level of variability of COVID-19 deaths within each region. Birmingham and neighbouring districts, along with London and neighbouring districts all appear to have a high level of COVID deaths.

A check for missing data was carried out to ascertain if all local authorities are accounted for. This is because if there are any missing data, the statistical methods used would result to error due to the fact that missing data are classified as N/A and numeric data types are required.

**Chart, treemap chart

Description automatically generated**

Figure : Missingness map for all independent variables.

As seen from the map above, there is no missing data meaning there was no need to perform any missing data handling techniques. Then I proceeded to exploring my variables with summary of the mean and median, box plot, histogram, QQ plot.

Box : Summary of Total COVID-19 deaths

Comparing the mean and the median, it is safe to say the dependent variable is not normally distributed as the difference gap is quite much.

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0 195.2 276.5 357.6 446.8 2716.0

To confirm this conclusion from the statistical summary, data visualization was used.

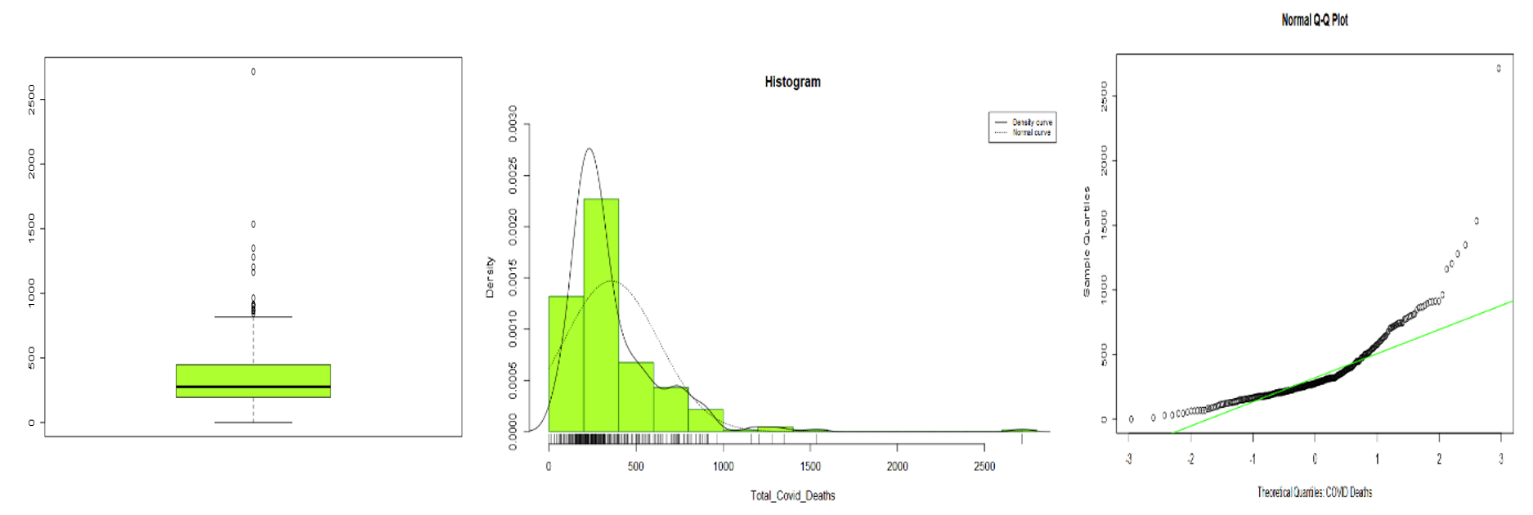


Figure : Box plot, Histogram, and QQ plot of total COVID-19 deaths

From the box plot, it is easily observed that the Total COVID-19 deaths data is not normally distributed and has quite a number of outliers. The histogram doesn’t conform to a bell shape as the normal curve shows in the plot. After a complete check of all the variables as shown in APPENDIX 3, it was then concluded that all variables were not normally distributed. The next process was to normalise all variables by per thousand (1000) as shown in APPENDIX 4 and then retest for normality .

Box : Summary of Total COVID-19 deaths per thousand

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.730 2.179 2.170 2.618 4.172

The Total COVID-19 deaths per thousand was then normalised as seen from the Box 2. The mean and median had a very insignificant difference.

To confirm this hypothesis, Kolmogorov-Smirnov test and visualisation methods such as box plot, histogram and QQ plot were used.

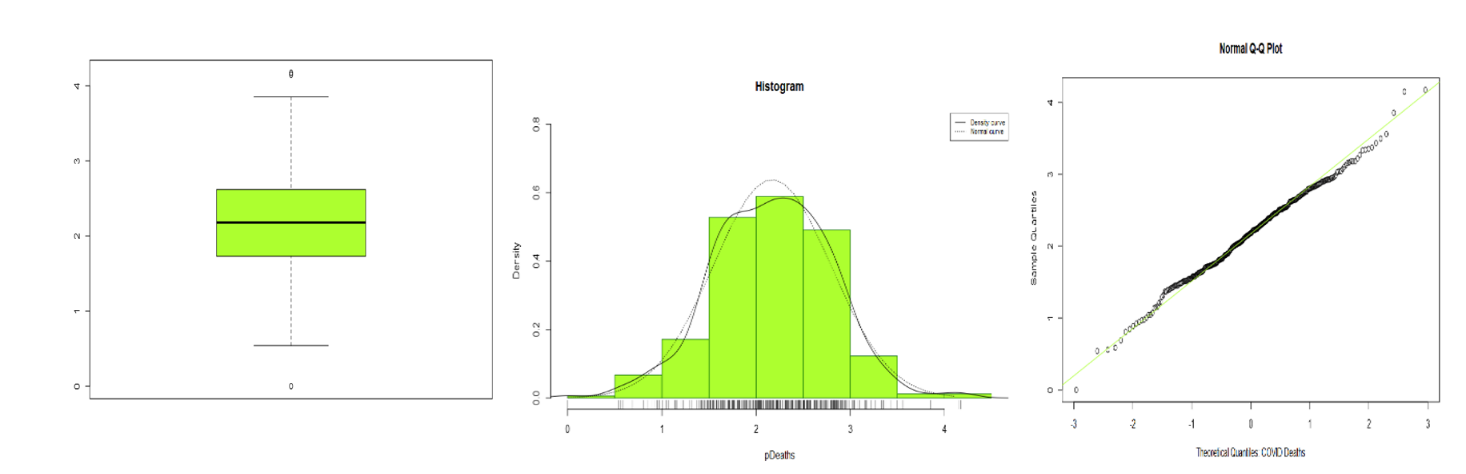


Figure : Box plot, Histogram, and QQ plot of total COVID-19 deaths per thousand

Box : KS test on pDeath

Asymptotic one-sample Kolmogorov-Smirnov test

data: pDeaths

D = 0.028772, p-value = 0.9501

alternative hypothesis: two-sided

As earlier stated, the newly standardized data was normalised as shown in the plots and in the KS test. As the p-value is greater than the confidence level of 0.05, the null hypothesis (H0) stating that there is no significant difference between the mean and standard deviation is then accepted. All the standardized independent variables were also normalised.

Correlation between the dependent (Total COVID deaths per thousand) and independent variables was carried out as shown in APPENDIX 5. Decided to create multivariate scatter plots for the dependent and each theme.

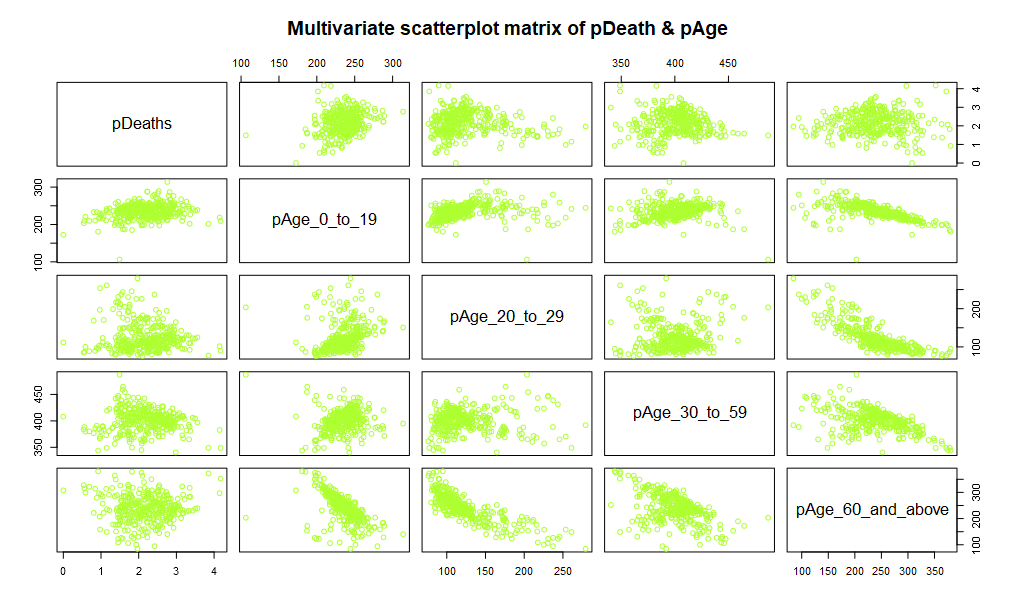


Figure : Multivariate scatter plot of pDeaths and pAge

After creating the correlation matrix and plotting the multivariate scatter plot for pDeath and pAge, it was observed the pAge\_60\_and\_above had internal correlations with other age groups. The strongest but negative correlation is with pAge\_20\_to\_29 while an average negative correlation with the remaining two groups. From this, it was hypothetically concluded that any model built with two or more age groups with one as pAge\_60\_and\_above would affect how much the variance of a regression coefficient is inflated due to multicollinearity.

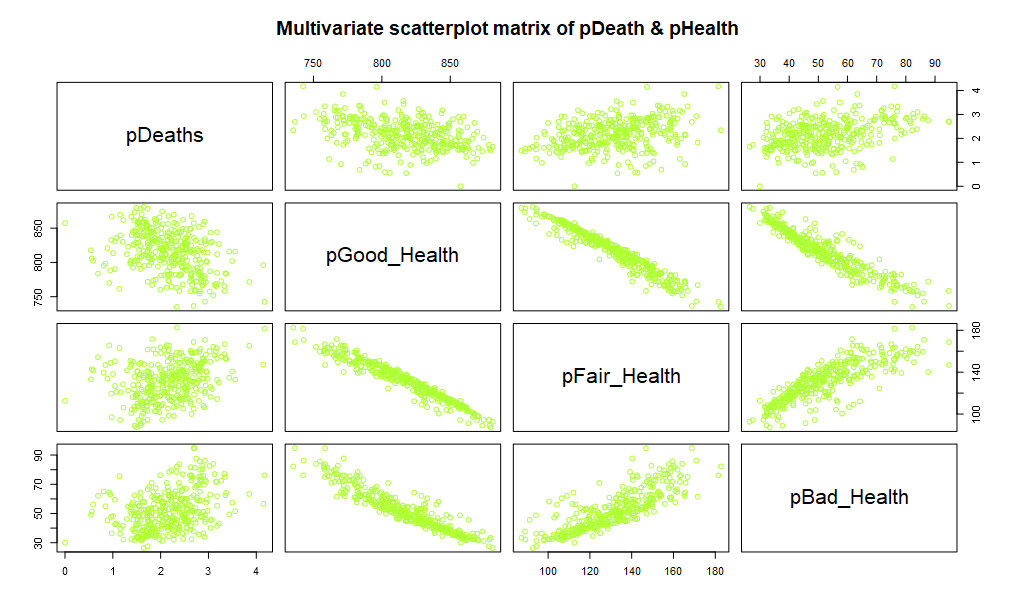


Figure : Multivariate scatter plot of pDeaths and pHealth

There was also internal correlation between all health classifications with the correlation being strong negative one. With this any model built with more than one health classification would inflate the variance of the inflation coefficient.

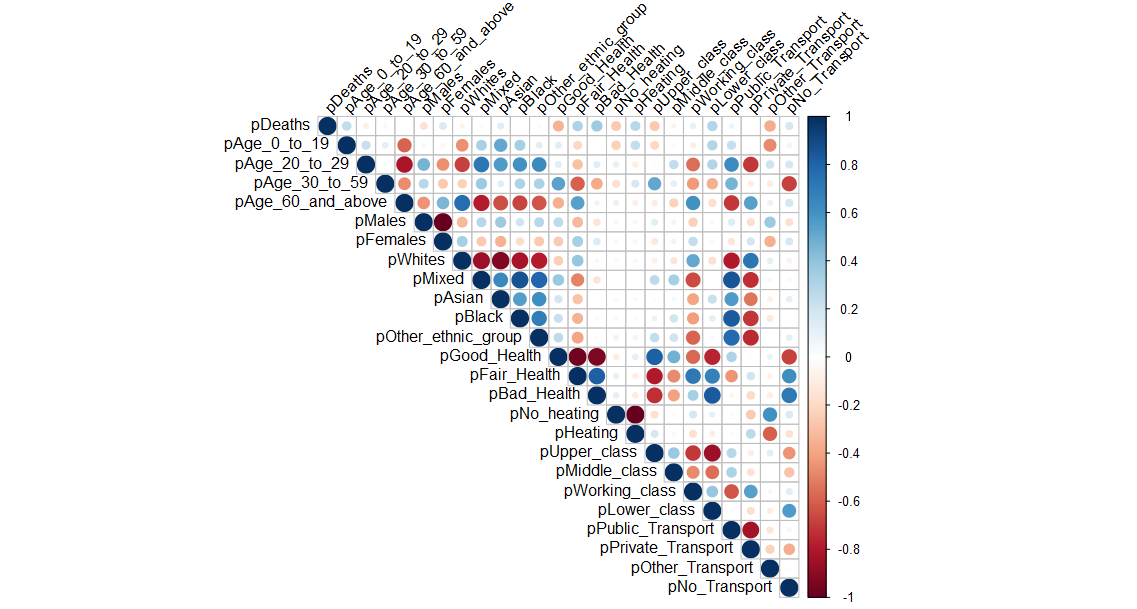


Figure : A correlation plot of all variables

The correlation plot visualizes the relationship between all the variables. From the plot, both sex have very strong negative correlation with each other and same goes for all ethnicities.

Box : Kaiser-Meyer-Olkins statistical test

Kaiser-Meyer-Olkin factor adequacy

Call: KMO(r = cor(cdeath\_data3))

Overall MSA = 0.5

MSA for each item =

pDeaths pAge\_0\_to\_19 pAge\_20\_to\_29 pAge\_30\_to\_59

0.5 0.5 0.5 0.5

pAge\_60\_and\_above pMales pFemales pWhites

0.5 0.5 0.5 0.5

pMixed pAsian pBlack pOther\_ethnic\_group

0.5 0.5 0.5 0.5

pGood\_Health pFair\_Health pBad\_Health pNo\_heating

0.5 0.5 0.5 0.5

pHeating pUpper\_class pMiddle\_class pWorking\_class

0.5 0.5 0.5 0.5

pLower\_class pPublic\_Transport pPrivate\_Transport pOther\_Transport

0.5 0.5 0.5 0.5

pNo\_Transport

0.5

The KMO test was carried out as shown in APPENDIX 6 to ascertain if the data is suited for factor analysis. The result shows that factor analysis is not needed as the overall MSA is less than 0.6. KMO value being less than 0.6 indicates that all the variance cannot be explained by factors.

## **REGRESSION MODEL**

The first model built comprised of all variables. This was to pick out the significant variables that contribute more to the model. Find code in APPENDIX 6.

Box : Multiple Regression Model using all variables (Model 1)

|  |
| --- |
| Call:  lm(formula = pDeaths ~ pAge\_0\_to\_19 + pAge\_20\_to\_29 + pAge\_30\_to\_59 +  pAge\_60\_and\_above + pMales + pFemales + pWhites + pMixed +  pAsian + pBlack + pOther\_ethnic\_group + pGood\_Health + pFair\_Health +  pBad\_Health + pNo\_heating + pHeating + pUpper\_class + pMiddle\_class +  pWorking\_class + pLower\_class + pPublic\_Transport + pPrivate\_Transport +  pOther\_Transport + pNo\_Transport)  Residuals:  Min 1Q Median 3Q Max  -1.29136 -0.29279 -0.02605 0.30235 1.64640  Coefficients: (7 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) 13.3444853 9.4416419 1.413 0.158558  pAge\_0\_to\_19 0.0064828 0.0029289 2.213 0.027605 \*  pAge\_20\_to\_29 -0.0081089 0.0023818 -3.404 0.000751 \*\*\*  pAge\_30\_to\_59 -0.0038849 0.0035189 -1.104 0.270455  pAge\_60\_and\_above NA NA NA NA  pMales 0.0056066 0.0056183 0.998 0.319101  pFemales NA NA NA NA  pWhites 0.0021342 0.0044999 0.474 0.635636  pMixed 0.0058251 0.0085103 0.684 0.494189  pAsian 0.0040044 0.0047100 0.850 0.395877  pBlack -0.0010568 0.0047133 -0.224 0.822733  pOther\_ethnic\_group NA NA NA NA  pGood\_Health -0.0209794 0.0076813 -2.731 0.006674 \*\*  pFair\_Health -0.0099324 0.0132624 -0.749 0.454480  pBad\_Health NA NA NA NA  pNo\_heating -0.0096422 0.0023335 -4.132 4.64e-05 \*\*\*  pHeating NA NA NA NA  pUpper\_class 0.0002491 0.0017101 0.146 0.884277  pMiddle\_class 0.0039908 0.0014549 2.743 0.006444 \*\*  pWorking\_class -0.0010325 0.0023786 -0.434 0.664555  pLower\_class NA NA NA NA  pPublic\_Transport 0.0068295 0.0022246 3.070 0.002331 \*\*  pPrivate\_Transport 0.0033152 0.0020047 1.654 0.099197 .  pOther\_Transport 0.0042646 0.0025621 1.665 0.097028 .  pNo\_Transport NA NA NA NA  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4779 on 308 degrees of freedom  Multiple R-squared: 0.4461, Adjusted R-squared: 0.4155  F-statistic: 14.59 on 17 and 308 DF, p-value: < 2.2e-16 |

For Model 1, out of 24 independent variables, only 6 were significant to the model. These variables are pAge\_0\_to\_19, pAge\_20\_to\_29, pGood\_Health, pNo\_heating, pMiddle\_class, and pPublic\_Transport. The model had a fitness of approximately 0.42 which can also be referred to as 42% accuracy of the model. Variance Inflation Factor (VIF) test could not be carried out on this model because of the NA values present.

The next model, Model 2, was built using the significant variables and also the variables that resulted in NA for Model 1.

Box : Multiple Regression Model using 13 variables (Model 2)

|  |
| --- |
| Call:  lm(formula = pDeaths ~ pAge\_0\_to\_19 + pAge\_20\_to\_29 + pAge\_60\_and\_above +  pFemales + pOther\_ethnic\_group + pGood\_Health + pBad\_Health +  pNo\_heating + pHeating + pMiddle\_class + pLower\_class + pPublic\_Transport +  pNo\_Transport)  Residuals:  Min 1Q Median 3Q Max  -1.32471 -0.30450 -0.00382 0.29316 1.70017  Coefficients: (1 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) 9.5499858 6.7537848 1.414 0.15835  pAge\_0\_to\_19 0.0124970 0.0038980 3.206 0.00148 \*\*  pAge\_20\_to\_29 -0.0026674 0.0031023 -0.860 0.39055  pAge\_60\_and\_above 0.0041331 0.0033863 1.221 0.22319  pFemales -0.0093942 0.0051909 -1.810 0.07129 .  pOther\_ethnic\_group 0.0011759 0.0041874 0.281 0.77903  pGood\_Health -0.0084453 0.0064100 -1.318 0.18863  pBad\_Health 0.0139304 0.0116839 1.192 0.23406  pNo\_heating -0.0090352 0.0016908 -5.344 1.76e-07 \*\*\*  pHeating NA NA NA NA  pMiddle\_class 0.0032936 0.0017358 1.897 0.05869 .  pLower\_class -0.0010090 0.0017248 -0.585 0.55896  pPublic\_Transport 0.0029272 0.0007108 4.118 4.90e-05 \*\*\*  pNo\_Transport -0.0024070 0.0018719 -1.286 0.19945  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4911 on 313 degrees of freedom  Multiple R-squared: 0.4056, Adjusted R-squared: 0.3828  F-statistic: 17.79 on 12 and 313 DF, p-value: < 2.2e-16 |

Model 2 had an adjusted R-squared of 0.38 and out of 13 variables used, just 3 were significant to the model. Variance Inflation Factor could also not be tested for this model because of the presence of the NA values. Then a Model 2b was then created with those three(3) significant variables. And the model had a fitness or adjusted R-squared of 0.12 approximately. VIF was tested for this model and there were no collinearities.

Model 3 was built using independent variables that had correlation with pDeaths and no internal correlation between each other.

Box : Multiple Regression Model using variables correlating with pDeaths (Model 3)

|  |
| --- |
| Call:  lm(formula = pDeaths ~ pAge\_0\_to\_19 + pAge\_60\_and\_above + pBad\_Health +  pNo\_heating + pMiddle\_class + pPublic\_Transport)  Residuals:  Min 1Q Median 3Q Max  -1.35025 -0.33246 -0.02621 0.28676 1.68684  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -4.0256344 0.7677299 -5.244 2.87e-07 \*\*\*  pAge\_0\_to\_19 0.0123621 0.0018429 6.708 9.00e-11 \*\*\*  pAge\_60\_and\_above 0.0059150 0.0009829 6.018 4.85e-09 \*\*\*  pBad\_Health 0.0202647 0.0022342 9.070 < 2e-16 \*\*\*  pNo\_heating -0.0100905 0.0016415 -6.147 2.35e-09 \*\*\*  pMiddle\_class 0.0025580 0.0011018 2.322 0.0209 \*  pPublic\_Transport 0.0025250 0.0005345 4.724 3.47e-06 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.499 on 319 degrees of freedom  Multiple R-squared: 0.3744, Adjusted R-squared: 0.3626  F-statistic: 31.81 on 6 and 319 DF, p-value: < 2.2e-16  > # calculate variance inflation factor  > vif(model3)  pAge\_0\_to\_19 pAge\_60\_and\_above pBad\_Health pNo\_heating pMiddle\_class pPublic\_Transport  1.902420 3.482030 1.227302 1.079541 1.332793 2.419397  > sqrt(vif(model3)) > 2 # if > 2 vif too high  pAge\_0\_to\_19 pAge\_60\_and\_above pBad\_Health pNo\_heating pMiddle\_class pPublic\_Transport  FALSE FALSE FALSE FALSE FALSE FALSE |

Model 3 with 6 variables has an adjusted R-squared of 0.36 and no collinearity after the VIF test. All variables were also very significant in the model. Created a fourth model with same variables as Model 3 but took out one of the variables with lesser significance (pMiddle\_Class).

Box : Multiple Regression Model using significant variables with pDeaths (Model 4)

|  |
| --- |
| Call:  lm(formula = pDeaths ~ pAge\_0\_to\_19 + pAge\_60\_and\_above + pBad\_Health +  pNo\_heating + pPublic\_Transport)  Residuals:  Min 1Q Median 3Q Max  -1.39329 -0.33684 0.00079 0.29998 1.60705  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -3.1052845 0.6619678 -4.691 4.03e-06 \*\*\*  pAge\_0\_to\_19 0.0121398 0.0018530 6.551 2.28e-10 \*\*\*  pAge\_60\_and\_above 0.0058572 0.0009893 5.920 8.27e-09 \*\*\*  pBad\_Health 0.0181340 0.0020509 8.842 < 2e-16 \*\*\*  pNo\_heating -0.0101258 0.0016527 -6.127 2.63e-09 \*\*\*  pPublic\_Transport 0.0027675 0.0005278 5.243 2.87e-07 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.5024 on 320 degrees of freedom  Multiple R-squared: 0.3638, Adjusted R-squared: 0.3538  F-statistic: 36.6 on 5 and 320 DF, p-value: < 2.2e-16  > # calculate variance inflation factor  > vif(model4)  pAge\_0\_to\_19 pAge\_60\_and\_above pBad\_Health pNo\_heating pPublic\_Transport  1.897280 3.479792 1.020220 1.079449 2.327033  > sqrt(vif(model4)) > 2 # if > 2 vif too high  pAge\_0\_to\_19 pAge\_60\_and\_above pBad\_Health pNo\_heating pPublic\_Transport  FALSE FALSE FALSE FALSE FALSE |

Model 4 with 5 variables has a fitness of 0.35 with no collinearity. Compared with the previous model (Model 3), there is just a difference of about 0.01 in the fitness. To test if the difference was actually a significant one, ANOVA test was then used to test for this.

Box : Multiple Regression Model using significant variables with pDeaths (Model 5)

|  |
| --- |
| Call:  lm(formula = pDeaths ~ pAge\_0\_to\_19 + pAge\_60\_and\_above + pBad\_Health +  pNo\_heating + pMiddle\_class + pPublic\_Transport + pAsian)  Residuals:  Min 1Q Median 3Q Max  -1.33324 -0.31335 -0.03699 0.32217 1.69834  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -3.9350517 0.7627246 -5.159 4.37e-07 \*\*\*  pAge\_0\_to\_19 0.0108572 0.0019297 5.626 4.04e-08 \*\*\*  pAge\_60\_and\_above 0.0063268 0.0009898 6.392 5.83e-10 \*\*\*  pBad\_Health 0.0206230 0.0022218 9.282 < 2e-16 \*\*\*  pNo\_heating -0.0108536 0.0016586 -6.544 2.40e-10 \*\*\*  pMiddle\_class 0.0030591 0.0011124 2.750 0.006300 \*\*  pPublic\_Transport 0.0020582 0.0005638 3.651 0.000306 \*\*\*  pAsian 0.0012598 0.0005158 2.443 0.015125 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4952 on 318 degrees of freedom  Multiple R-squared: 0.3859, Adjusted R-squared: 0.3724  F-statistic: 28.55 on 7 and 318 DF, p-value: < 2.2e-16  > # calculate variance inflation factor  > vif(model5)  pAge\_0\_to\_19 pAge\_60\_and\_above pBad\_Health pNo\_heating pMiddle\_class pPublic\_Transport  2.118366 3.586034 1.232673 1.119256 1.379712 2.733431  pAsian  2.108641  > sqrt(vif(model5)) > 2 # if > 2 vif too high  pAge\_0\_to\_19 pAge\_60\_and\_above pBad\_Health pNo\_heating pMiddle\_class pPublic\_Transport  FALSE FALSE FALSE FALSE FALSE FALSE  pAsian  FALSE |

Model 5 consists of 7 variables and has an adjusted R-squared as 0.37 but has lesser F-statistic than Model 3 and Model 4.

Box : Multiple Regression Model using stepwise approach on Model 1 (Model 6)

|  |
| --- |
| Call:  lm(formula = pDeaths ~ pOther\_Transport + pBad\_Health + pNo\_heating +  pAsian + pAge\_20\_to\_29 + pMiddle\_class)  Residuals:  Min 1Q Median 3Q Max  -1.35207 -0.29066 -0.04207 0.29245 1.70294  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0.2681937 0.4005268 0.670 0.504  pOther\_Transport -0.0011731 0.0012276 -0.956 0.340  pBad\_Health 0.0254804 0.0024390 10.447 < 2e-16 \*\*\*  pNo\_heating -0.0102112 0.0020396 -5.007 9.18e-07 \*\*\*  pAsian 0.0029454 0.0004718 6.243 1.37e-09 \*\*\*  pAge\_20\_to\_29 -0.0067537 0.0010711 -6.305 9.57e-10 \*\*\*  pMiddle\_class 0.0053084 0.0011324 4.688 4.10e-06 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4896 on 319 degrees of freedom  Multiple R-squared: 0.3978, Adjusted R-squared: 0.3864  F-statistic: 35.11 on 6 and 319 DF, p-value: < 2.2e-16  > sqrt(vif(model6)) > 2 # if > 2 vif too high  pOther\_Transport pBad\_Health pNo\_heating pAsian pAge\_20\_to\_29 pMiddle\_class  FALSE FALSE FALSE FALSE FALSE FALSE |

Model 6 was created using the stepwise approach on Model 1 and it is by far the best model. Model 6 has an adjusted R-squared of approximately 0.39 and 35.11 F-statistic. Compares to the next best out of all the models which is Model 5, it has lesser variables and more prediction accuracy. The model points out that bad health, no heating in homes, Asians, age groups 20 to 29, and middle-class individuals are all significant predictor variables for COVID-19 death.

# **CONCLUSION**

The analysis was focused on analysing the relationships between the social and economic variables on the COVID deaths in England. The aim was to analyse a set of variables which was done by testing the data for normality using visualisation and statistical test and then normalizing them. The correlation between those variables was then produced and it was discovered that majority of the variables from the same theme (Theme: Age, Sex, Ethnicity, Health,…) had a correlation with each other like Male and Female. From that, one variable but no more than two could have been used to build a regression model. The best model was Model 6 which contains one variable each from the Age, Ethnicity, Health, Heating availability, Travel to work, and Social grade themes.

Not all variables had a significant relationship with COVID deaths which is the dependent. It was observed that Bad health appeared to be one of the variables with a significant connection with COVID deaths. Meaning the difference in the COVID deaths with a pre-existing condition and no pre-existing condition is not just by chance but bad health does actually have a connection with COVID deaths. Another significant variable is the unavailability of heating which is considered logical as cold and dry temperatures weaken host defences and make humans more vulnerable(Wang *et al.*, 2021). Age\_20\_to\_29 was also a significant variable possible because it is the youthful and rebellious age where a lot of factors can come into a place like lifestyle, hygiene, social distancing law abiding and so on. All hypotheses can still further be tested but a significant and reliable amount of data would be needed.

# **REFERENCE**

Arabi, Y.M., Murthy, S. and Webb, S. (2020) ‘COVID-19: a novel coronavirus and a novel challenge for critical care’, *Intensive Care Medicine*, 46(5), pp. 833–836. Available at: https://doi.org/10.1007/s00134-020-05955-1.

Aziz, S. *et al.* (2020) ‘Managing ICU surge during the COVID-19 crisis: rapid guidelines’, *Intensive Care Medicine*, 46(7), pp. 1303–1325. Available at: https://doi.org/10.1007/s00134-020-06092-5.

Dehingia, N. and Raj, A. (2021) ‘Sex differences in COVID-19 case fatality: do we know enough?’, *The Lancet Global Health*. Elsevier Ltd, pp. e14–e15. Available at: https://doi.org/10.1016/S2214-109X(20)30464-2.

gov.uk (no date) *GOV.UK*. Available at: https://www.gov.uk/ (Accessed: 7 December 2022).

Ioannidis, J.P.A., Axfors, C. and Contopoulos-Ioannidis, D.G. (2020) ‘Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters’, *Environmental Research*, 188. Available at: https://doi.org/10.1016/j.envres.2020.109890.

Lamarque, K. (2020) *HEALTH-CORONAVIRUS/USA*. Available at: www.thelancet.com.

*Shepway - Wikipedia* (no date). Available at: https://en.wikipedia.org/wiki/Shepway (Accessed: 8 December 2022).

Wang, J. *et al.* (2021) ‘Impact of temperature and relative humidity on the transmission of COVID-19: A modelling study in China and the United States’, *BMJ Open*, 11(2). Available at: https://doi.org/10.1136/bmjopen-2020-043863.

World Health Organization (2020) ‘20200121-sitrep-1-2019-ncov’, *SITUATION REPORT - 1* [Preprint]. Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10\_4 (Accessed: 7 December 2022).

Yap, T.F. *et al.* (2020) ‘A predictive model of the temperature-dependent inactivation of coronaviruses’, *Applied Physics Letters*, 117(6). Available at: https://doi.org/10.1063/5.0020782.

# **APPENDICES**

APPENDIX 1

|  |
| --- |
| SELECT \*  FROM COVID\_Deaths  LEFT JOIN Age  ON COVID\_Deaths.LA\_name = Age.geography  LEFT JOIN Sex  ON COVID\_Deaths.LA\_name = Sex.geography  LEFT JOIN Ethnicity  ON COVID\_Deaths.LA\_name = Ethnicity.geography  LEFT JOIN Health  ON COVID\_Deaths.LA\_name = Health.geography  LEFT JOIN Heating  ON COVID\_Deaths.LA\_name = Heating.geography  LEFT JOIN "Social grade"  ON COVID\_Deaths.LA\_name = "Social grade".geography  LEFT JOIN "Travel to work"  ON COVID\_Deaths.LA\_name = "Travel to work".geography |

APPENDIX 2

|  |
| --- |
| #-----Reading Data-------------------------------------------  setwd(dirname(file.choose()))  getwd()  cdeath\_data <- read.csv("freshdata2.csv", stringsAsFactors = FALSE)  head(cdeath\_data) # Inspect top rows of the data  str(cdeath\_data) |

APPENDIX 3

|  |
| --- |
| #-----EDA-------------------------------------------  attach(cdeath\_data)  # check for missing data  apply(cdeath\_data, MARGIN = 2, FUN = function(x) sum(is.na(x)))  library(Amelia)  missmap(cdeath\_data, y.labels =District, col = c("black", "pink"), legend = TRUE)  #----England Map--------------  library(rgdal)  library(GISTools)  library(RColorBrewer)  # Read in the shapefile of london\_polygon  ld.polygon <- readOGR(".", "LAD\_DEC\_2021\_UK\_BGC")  write.csv(ld.polygon, file="ld.polygon.csv")  ld.polygon2 <- read.csv("ld.polygon.csv", stringsAsFactors = FALSE)  ld.polygon@data <- within(ld.polygon@data, death\_number <- (ld.polygon2$death\_number))  plot(ld.polygon, border = "black", col = "lightgrey")  names(ld.polygon@data)  # Set colour and number of classes  shades <- auto.shading(ld.polygon$death\_number, n = 5, cols = brewer.pal(5, "Oranges"))  # Draw the map polygons  choropleth(ld.polygon, ld.polygon$death\_number, shades)  title("Map of England")  choro.legend(557000, 182000, shades, fmt = "%g", title = "COVID Deaths", cex=0.5)  box(which = "outer")  #-----Normality check-------------------------------------------  summary(Total\_COVID\_Deaths)  #Histogram with density curve  hist(Total\_Covid\_Deaths, col = "greenyellow", border = "dark green", freq = F, ylim = c(0,0.003),  xlab = "Total\_Covid\_Deaths", main = "Histogram")  rug (Total\_Covid\_Deaths)  lines (density(sort(Total\_Covid\_Deaths)))  xfit <- seq(from = min(Total\_Covid\_Deaths), to = max(Total\_Covid\_Deaths), by = 0.1)  yfit = dnorm(xfit, mean(Total\_Covid\_Deaths), sd(Total\_Covid\_Deaths))  lines(xfit, yfit, lty = "dotted")  rm(xfit, yfit)  legend("topright", legend = c("Density curve", "Normal curve"),  lty = c("solid", "dotted"), cex = 0.7)  boxplot(Total\_Covid\_Deaths, col="greenyellow")  #boxplot(Total\_Covid\_Deaths, ylim= c(0,1000))  qqnorm(Total\_Covid\_Deaths, xlab = "Theoretical Quantiles: COVID Deaths" )  qqline(Total\_Covid\_Deaths, col="green") ## red color  ks.test(Total\_Covid\_Deaths,"pnorm", mean(Total\_Covid\_Deaths), sd(Total\_Covid\_Deaths))  #-----Check the variables for outliers using boxplot-------------------------------------------  boxplot(Age\_0\_to\_19, Age\_20\_to\_29, Age\_30\_to\_59, Age\_60\_and\_above,  names = c("Age0\_to\_19","Age\_20\_to\_29","Age\_30\_to\_59","Age\_60\_and\_above"),  xlab = "Age groups", ylab ="frequency", col = "greenyellow")  boxplot(Males, Females,  names = c("Males","Females"),  xlab = "Gender groups", ylab ="frequency", col = "greenyellow")  boxplot(White, Mixed, Asian, Black, Other\_ethnic\_group,  names = c("White","Mixed", "Asian", "Black", "Other\_ethnic\_groups"),  xlab = "Ethnic groups", ylab ="frequency", col = "greenyellow")  boxplot(Good\_Health, Fair\_Health, Bad\_Health,  names = c("Good\_Health","Fair\_Health", "Bad\_Health"),  xlab = "Health groups", ylab ="frequency", col = "greenyellow")  boxplot(No\_heating, Heating,  names = c("No\_heating","Heating"),  xlab = "Heating groups", ylab ="frequency", col = "greenyellow")  boxplot(Upper\_class, Middle\_class, Working\_class, Lower\_class,  names = c("Upper\_class","Middle\_class", "Working\_class", "Lower\_class"),  xlab = "Social class", ylab ="frequency", col = "greenyellow")  boxplot(Public\_Transport, Private\_Transport, Other\_Transport, No\_Transport,  names = c("Public\_Transport","Private\_Transport", "Other\_Transport", "No\_Transport"),  xlab = "Travel to work", ylab ="frequency", col = "greenyellow") |

APPENDIX 4

|  |
| --- |
| #-----Normalizing data in per 1000-------------------------------------------  summary(Total\_Covid\_Deaths)  cdeath\_data2 <- within(cdeath\_data2, pDeaths <- (Total\_Covid\_Deaths/Total\_age) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pAge\_0\_to\_19 <- (Age\_0\_to\_19/Total\_age) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pAge\_20\_to\_29 <- (Age\_20\_to\_29/Total\_age) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pAge\_30\_to\_59 <- (Age\_30\_to\_59/Total\_age) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pAge\_60\_and\_above <- (Age\_60\_and\_above/Total\_age) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pMales <- (Males/Total\_sex) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pFemales <- (Females/Total\_sex) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pWhites <- (White/Total\_ethnicity) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pMixed <- (Mixed/Total\_ethnicity) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pAsian <- (Asian/Total\_ethnicity) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pBlack <- (Black/Total\_ethnicity) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pOther\_ethnic\_group <- (Other\_ethnic\_group/Total\_ethnicity) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pGood\_Health <- (Good\_Health/Total\_Health) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pFair\_Health <- (Fair\_Health/Total\_Health) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pBad\_Health <- (Bad\_Health/Total\_Health) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pNo\_heating <- (No\_heating/Total\_heating) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pHeating <- (Heating/Total\_heating) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pUpper\_class <- (Upper\_class/Total\_s\_grade) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pMiddle\_class <- (Middle\_class/Total\_s\_grade) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pWorking\_class <- (Working\_class/Total\_s\_grade) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pLower\_class <- (Lower\_class/Total\_s\_grade) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pPublic\_Transport <- (Public\_Transport/Total\_transport) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pPrivate\_Transport <- (Private\_Transport/Total\_transport) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pOther\_Transport <- (Other\_Transport/Total\_transport) \* 1000)  cdeath\_data2 <- within(cdeath\_data2, pNo\_Transport <- (No\_Transport/Total\_transport) \* 1000)  #-----Normality check for standardized data-------------------------------------------  detach(cdeath\_data)  attach(cdeath\_data2)  summary(pDeaths)  hist(pDeaths)  hist(pDeaths, col = "greenyellow", border = "dark green", freq = F, ylim = c(0,0.8),  xlab = "pDeaths", main = "Histogram")  rug (pDeaths)  lines (density(sort(pDeaths)))  xfit <- seq(from = min(pDeaths), to = max(pDeaths), by = 0.1)  yfit = dnorm(xfit, mean(pDeaths), sd(pDeaths))  lines(xfit, yfit, lty = "dotted")  rm(xfit, yfit)  legend("topright", legend = c("Density curve", "Normal curve"),  lty = c("solid", "dotted"), cex = 0.7)  boxplot(pDeaths, col="greenyellow")  qqnorm(pDeaths, xlab = "Theoretical Quantiles: COVID Deaths" )  qqline(pDeaths, col="greenyellow") ## red color  ks.test(pDeaths,"pnorm", mean(pDeaths), sd(pDeaths))  boxplot(pAge\_0\_to\_19, pAge\_20\_to\_29, pAge\_30\_to\_59, pAge\_60\_and\_above,  names = c("pAge0\_to\_19","pAge\_20\_to\_29","pAge\_30\_to\_59","pAge\_60\_and\_above"),  xlab = "Age groups per 1000", ylab ="frequency", col = "greenyellow")  boxplot(pMales, pFemales,  names = c("pMales","pFemales"),  xlab = "Gender groups per 1000", ylab ="frequency", col = "greenyellow")  boxplot(pWhite, pMixed, pAsian, puBlack, pOther\_ethnic\_group,  names = c("pWhite","pMixed", "pAsian", "pBlack", "pOther\_ethnic\_groups"),  xlab = "Ethnic groups", ylab ="frequency", col = "greenyellow")  boxplot(Good\_Health, Fair\_Health, Bad\_Health,  names = c("Good\_Health","Fair\_Health", "Bad\_Health"),  xlab = "Health groups", ylab ="frequency", col = "greenyellow")  boxplot(No\_heating, Heating,  names = c("No\_heating","Heating"),  xlab = "Heating groups", ylab ="frequency", col = "greenyellow")  boxplot(Upper\_class, Middle\_class, Working\_class, Lower\_class,  names = c("Upper\_class","Middle\_class", "Working\_class", "Lower\_class"),  xlab = "Social class", ylab ="frequency", col = "greenyellow")  boxplot(Public\_Transport, Private\_Transport, Other\_Transport, No\_Transport,  names = c("Public\_Transport","Private\_Transport", "Other\_Transport", "No\_Transport"),  xlab = "Travel to work", ylab ="frequency", col = "greenyellow") |

APPENDIX 5

|  |
| --- |
| #-----Correlation Matrix-------------------------------------------  cdeath\_data3 <- cdeath\_data2[33:57]  cor\_test2 <- cor(cdeath\_data3)  cor\_test2 <- round(cor\_test2, digits = 2)  library(corrplot)  corrplot(cor\_test2, type = "upper", tl.col = "black", tl.srt = 45)  detach(cdeath\_data2)  attach(cdeath\_data3)  pairs(~ pDeaths + pAge\_0\_to\_19 + pAge\_20\_to\_29 + pAge\_30\_to\_59 + pAge\_60\_and\_above, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pAge", col = "greenyellow")  pairs(~ pDeaths + pMales + pFemales, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pGender", col = "greenyellow")  pairs(~ pDeaths + pWhites + pMixed + pAsian + pBlack + pOther\_ethnic\_group, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pEthnicity", col = "greenyellow")  pairs(~ pDeaths + pGood\_Health + pFair\_Health + pBad\_Health, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pHealth", col = "greenyellow")  pairs(~ pDeaths + pNo\_heating + pHeating, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pHeating", col = "#8B7355")  pairs(~ pDeaths + pUpper\_class + pMiddle\_class + pLower\_class, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pSocial\_grade", col = "#8B2252")  pairs(~ pDeaths + pPublic\_Transport + pPrivate\_Transport + pOther\_Transport + pNo\_Transport, data = cdeath\_data3, main = "Multivariate scatterplot matrix of pDeath & pTransport", col = "#8B4513") |

APPENDIX 6

|  |
| --- |
| # Kaiser-Meyer-Olkin statistics: if overall MSA > 0.6, proceed to factor analysis  library(psych)  KMO(cor(cdeath\_data3))  #---------Multiple Regression Modelling---------------------  # model with all variables  model1 <- lm(pDeaths ~ pAge\_0\_to\_19 + pAge\_20\_to\_29 + pAge\_30\_to\_59 + pAge\_60\_and\_above + pMales + pFemales + pWhites  + pMixed + pAsian + pBlack + pOther\_ethnic\_group + pGood\_Health + pFair\_Health + pBad\_Health + pNo\_heating  + pHeating + pUpper\_class + pMiddle\_class +pWorking\_class + pLower\_class + pPublic\_Transport + pPrivate\_Transport + pOther\_Transport  + pNo\_Transport)  summary(model1)  # calculate variance inflation factor  library(car)  vif(model1)  sqrt(vif(model1)) > 2 # if > 2 vif too high  model2 <- lm(pDeaths ~ pAge\_0\_to\_19 + pAge\_20\_to\_29 + pAge\_60\_and\_above + pFemales + pOther\_ethnic\_group + pGood\_Health + pBad\_Health + pNo\_heating  + pHeating + pMiddle\_class + pLower\_class + pPublic\_Transport + pNo\_Transport)  summary(model2)  # calculate variance inflation factor  vif(model2)  sqrt(vif(model2)) > 2 # if > 2 vif too high  model2b <- lm(pDeaths ~ pAge\_0\_to\_19 + pNo\_heating + pPublic\_Transport)  summary(model2b)  # calculate variance inflation factor  vif(model2b)  sqrt(vif(model2b)) > 2 # if > 2 vif too high  model3 <- lm(pDeaths ~ pAge\_0\_to\_19 + pAge\_60\_and\_above + pBad\_Health + pNo\_heating  + pMiddle\_class + pPublic\_Transport)  summary(model3)  # calculate variance inflation factor  vif(model3)  sqrt(vif(model3)) > 2 # if > 2 vif too high  model4 <- lm(pDeaths ~ pAge\_0\_to\_19 + pAge\_60\_and\_above + pBad\_Health + pNo\_heating + pPublic\_Transport)  summary(model4)  hist(model4$residuals)  rug(model4$residuals)  plot(model4$residuals ~ model4$fitted.values, xlab = "fitted values", ylab = "residuals")  ks.test(model4$residuals, "pnorm", mean(model4$residuals), sd(model4$residuals))  # calculate variance inflation factor  vif(model4)  sqrt(vif(model4)) > 2 # if > 2 vif too high  # test whether model3 and model4 are significantly different using F test  anova(model3, model4, test = "F")  model5 <- lm(pDeaths ~ pAge\_0\_to\_19 + pAge\_60\_and\_above + pBad\_Health + pNo\_heating + pMiddle\_class+ pPublic\_Transport + pAsian)  summary(model5)  # calculate variance inflation factor  vif(model5)  sqrt(vif(model5)) > 2 # if > 2 vif too high  library(RcmdrMisc)  library(relaimpo)  model6 <- stepwise(model1, direction = "forward")  summary(model6)  vif(model6)  sqrt(vif(model6)) > 2 # if > 2 vif too high  hist(model6$residuals)  rug(model6$residuals)  plot(model5$residuals ~ model5$fitted.values, xlab = "fitted values", ylab = "residuals")  ks.test(model5$residuals, "pnorm", mean(model5$residuals), sd(model5$residuals))  sqrt(vif(model6)) > 2  calc.relimp(model6, type = c("lmg"), rela = TRUE)  # test whether model4 and model6 are significantly different using F test  anova(model4, model6, test = "F") |