ANALYSIS TO DETERMINE FACTORS INVOLVED IN COVID DEATHS

A QUANTITATIVE DATA ANALYSIS PROJECT TO DEDUCE THE FACTORS THAT CONTRIBUTED TO COVID DEATHS IN ENGLAND AND WALES BETWEEN MARCH 2020 AND APRIL 2021

AIYEGBENI GIFTY

A QUANTITATIVE DATA ANALYSIS PROJECT

Table of Contents

1. INTRODUCTION

Coronavirus disease (COVID) is a syndrome caused by an acute respiratory tract infection and patients that are critically ill with coronavirus have a high death rate (Dong *et al.*, 2021; Esposito *et al.*, 2021). The outbreak of COVID started in December 2019 in China, precisely Wuhan. This pandemic, with its rapid spread, has created an extraordinary challenge for the world in the 21st century (Jcma Rahmatizadeh et al., 2020).

COVID symptoms include cough, fever, respiratory symptoms, a low white blood cell (WBC) count, pneumonia, and shortness of breath. Other symptoms may include muscle pain, diarrhoea, abdominal pain, sore throat, sputum production, and loss of taste and smell (Salman *et al.*, 2020; Dong *et al.*, 2021).

COVID is commonly transmitted through close contact and respiratory droplets produced by sneezing or coughing. While breathing, respiratory droplets may form, but they are not well-planned airborne particles (Salman *et al.*, 2020).

The World Health Organization (WHO) announced the COVID-19 pandemic on March 11, 2020. As of 23rd December 2022, 651,918,402 COVID cases have been confirmed by WHO and 6,656,601 deaths have been reported. As of the 13th of December 2022, 13,008,560,983 Vaccination doses have been given (World Health Organization, 2022).

From recent articles, it has been discovered that there are significant associations between COVID incidence and mortality across social and economic domains. For religion, compared to Christians, people who identify with other religions died more from COVID (Gaughan *et al.*, 2021). (Karmakar, Lantz and Tipirneni, 2021) stated that social status and racial/ethnic minority status were associated with COVID incidence and mortality rates. (Roy and Ghosh, 2020) also stated that the COVID death rate seems to be driven by pre-existing health conditions while religions have little or no impact on the pandemic incidence or deaths.

Hence the objective of this project is to analyze COVID deaths with specific social and economic themes to deduce what variables are more associated with COVID deaths. The selected themes for this analysis are:

1) Religion
2) Ethnicity
3) General health condition
4) Social grade

Through this project, I will be showing the variables that can influence the possibility of a patient dying from COVID and to what extent. I will be using the R programming language for my research and my codes can be found here.

2 METHODOLOGY

2.1 Data Acquisition:

I was given the data of England by the Local Authority on COVID deaths in a CSV file. I then went on the ONS Neighbourhood Statistics portal at Nomis and chose themes that would support the objective of my analysis and downloaded by Local Authority administrative geography.

I edited the column names and loaded all the CSV files into SQLite. I joined all the tables using the same primary key (district) through some SQL queries (appendix A) then I copied them into an excel

document, edited and saved them as a CSV file ready to use in R. The final dataset has 322 observations of 37 variables. The selected variable and sub-variable of the dataset are given below:

Themes	**Variables**
Religion	Christian, Buddhist, Hindu, Muslim, Other religions, and no religion
Ethnicity	White, Mixed, Asian, Black, and Others
General health condition	Good health, fair health, and bad health
Social grade	Upper, upper middle, middle and lower

Table 1: Tables of all themes and their variables.

2.1.1 Definition of Terms Used in R Script

To have a better understanding of this study, the following few terms below are defined in the context of usage in this research;

Total_pop – number of people per district in England and Wales

Total_deaths – number of people that died from COVID per district

Other_religion – people that practice other religions not listed

Noreligion – people who have no religion or didn't state any

Other_Ethnicity – people from other races not listed

Good – people that are free from bodily or mental disease.

Fair – people that have neither good nor bad health conditions

Bad – people that have a bodily or mental disease

SocialGrade_Total – total number of people based on classification by occupation

Upper – people with higher & intermediate professional occupations

Upper_middle – people with supervisor and administrative professional occupations

Middle – people with skilled manual occupations

Lower – people that are semi-skilled, unskilled, unemployed, or with lowest-grade occupations.

2.2 Data Exploration:

I explored my data using various techniques by looking at the trend, central tendency of the data, spread, and checking for missing data and outliers. I also did hypothesis tests and correlation tests to explore the data.

2.3 Data Analysis:

For data analysis, I did the principal component analysis and factor analysis on the variables. I also did cluster analysis and modeling. I did the multiple regression modeling and the stepwise approach as well. Finally, I did the ANOVA test of variance to test for the best model.

3 DATA EXPLORATION

3.1 TREND ANALYSIS:

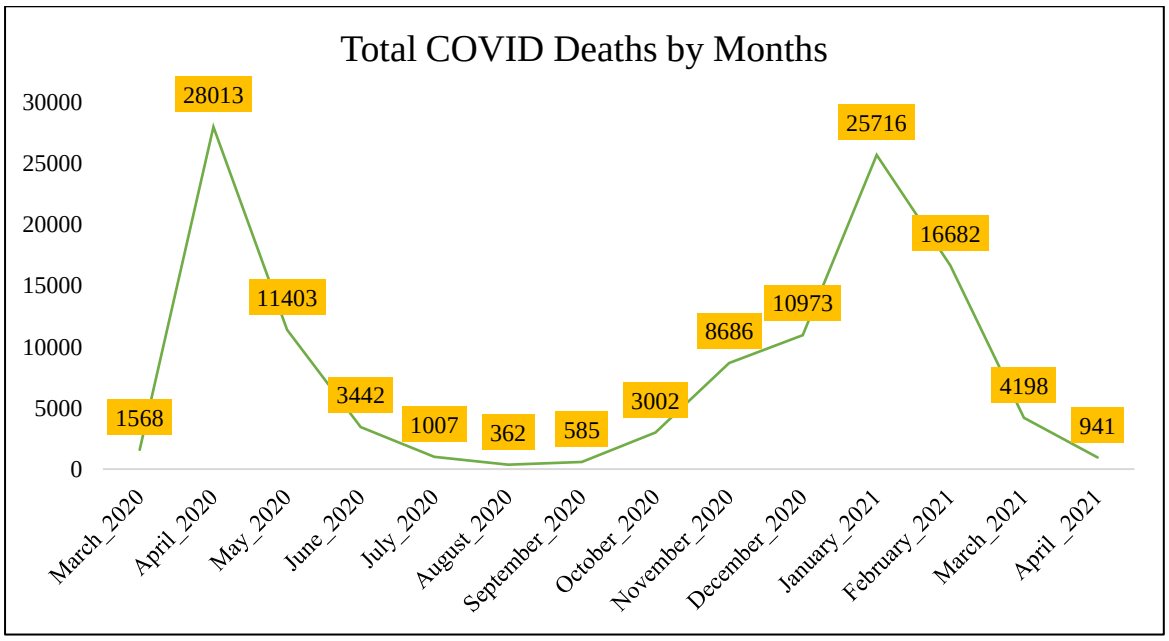

Figure 3.1.1: Plot of trend analysis for COVID deaths

I began by plotting a graph of the total COVID deaths by month in excel. This was to see the trend and pattern of the data. From the graph, April 2020 had the highest death this was the early period of the lockdown. August 2020 had the lowest death and this was when the quarantine measures were at their peak.

3.2 CENTRAL TENDENCY:

I checked the central tendency after loading the CSV file into R using the summary function to have an overview of the mean, median, and quartiles of all the variables. From observing the central tendency, I could deduce that most of the variables were not symmetric and this suggests that they are not normally distributed.

3.3 SPREAD

Looking at the minimum, maximum, and range values using the “describe” function, I saw that the spread of the data is high. This is due to the uneven population distribution. Places that are sparsely populated had lower values for each variable and vice versa.

3.4 MISSING DATA

I checked for the missing data by webbing the sum function and the 'is.na' function on the dataset. Then I visualized the missing data using the 'vis_miss’ and ‘vis_dat’ functions. Every plot returned a result stating that the dataset is complete and missing no entry.

Figure 3.4.1: Plot of missingness using ‘vis_miss’ function

3.5 CHECKING FOR OUTLIERS

I checked for outliers in the COVID deaths column by plotting a boxplot. The variable has a lot of outliers which I assumed was due to the population spread. I labeled the outliers and selected the highest three which were Birmingham, Leeds, and Durham to inspect them. Inspecting the outliers showed that there is no evidence of an error in the data as the population of these districts is similarly high.

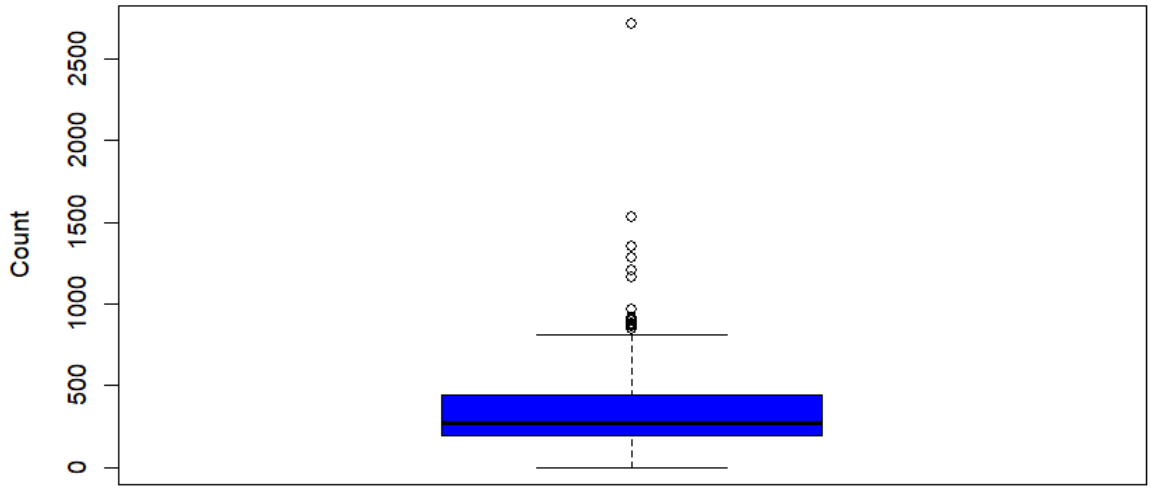

Figure 3.5.1: Boxplot of the total COVID deaths

3.6 CHECKING FOR NORMALITY

I did the normality test to compare the distribution of the COVID death (target variable) data to a normal distribution. This was done because, from a statistical point of view, most standard tests assume normal distributions. I made use of 3 normality tests;

3.6.1 Kolmogorov-Smirnov test (KS Test):

This KS test was done to compare the target variable to a normal distribution using a significance level of 5%. Hence, the null hypothesis was "there is no significant difference between the COVID death data and a normal distribution".

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Total_deaths

D = 0.1841, p-value = 6.634e-10

alternative hypothesis: two-sided
```

Figure 3.6.1.1: KS test of the COVID deaths variable

From the figure above, we can safely reject the null hypothesis as there is a significant difference between the data and a normal distribution. This is because the P-value is significantly below the 0.05 benchmark.

3.6.2 Shapiro-Wilk's test (SW Test):

To get further insights into the normality of the data, I did the SW test which is an alternative test of normality in statistics.

```
        Shapiro-Wilk normality test

data:  Total_deaths

W = 0.76286, p-value < 2.2e-16
```

Figure 3.6.2.2: Shapiro-Wilk's test of the COVID deaths variable

The P-value of the test was less than 0.05 and this means that the data is not normally distributed.

3.6.3 Quantile-Quantile (Q-Q) plots:

Finally, I did a normal probability plot using the Q-Q plot to compare the data distribution to the expected normal distribution. We expect observations from normally distributed data to lie approximately in a straight line.

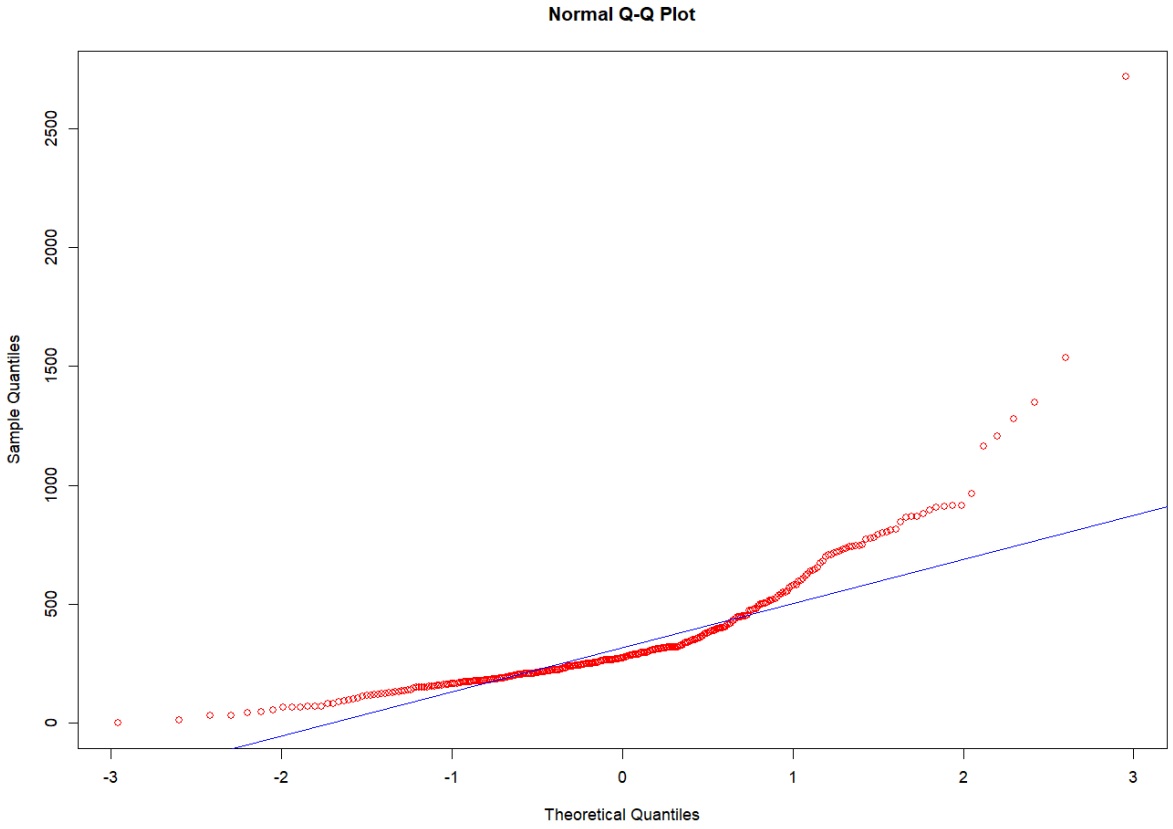

Figure 3.6.2.3: QQ plot of total COVID death data

This shows that the data does not lie on a straight line and the qq-line is not positioned at an approximate angle of 45º. This confirms that the data is not normally distributed.

3.7 STANDARDIZING THE DATA

Since the spread of the data is large and there are several outliers, I standardized all variables to ensure uniformity and improve the accuracy of the data.

To standardize the data, I made all variables to be in a similar scale without changing the difference between them. I indexed all the variables and calculated the percentage using the appropriate base

population. After that, I visualized all the independent and dependent variables on boxplots and qq-plots (appendix b) to check for normality.

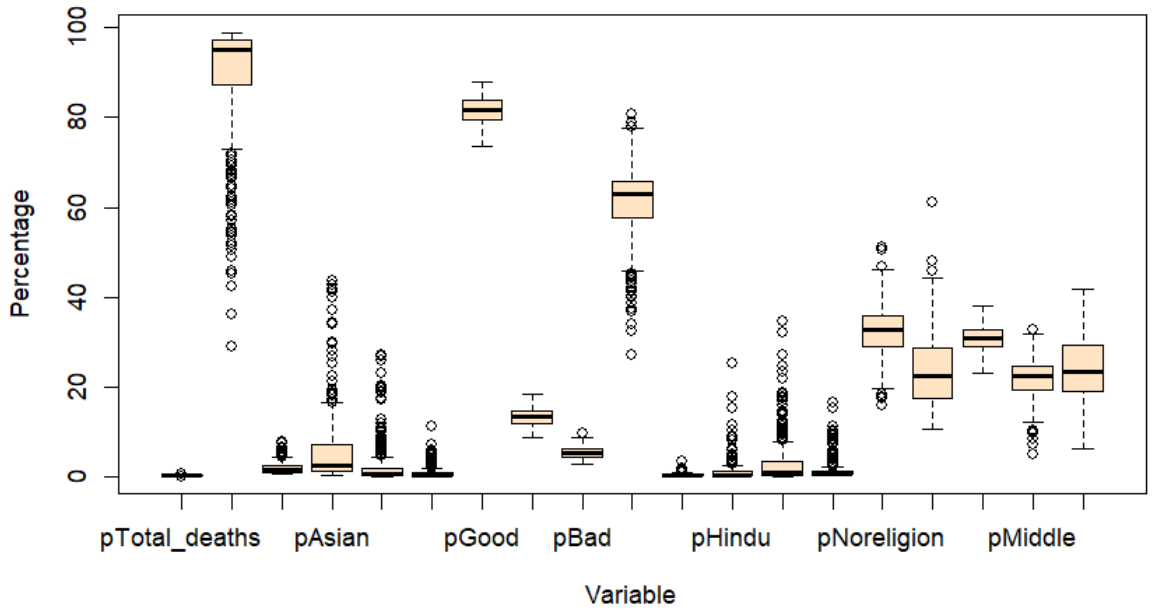

Figure 3.7.1: Boxplots of the percentage independent variables

These are the boxplots of the percentage variables and we can see that some variables are normally distributed. Although there are still a great number of outliers in the data.

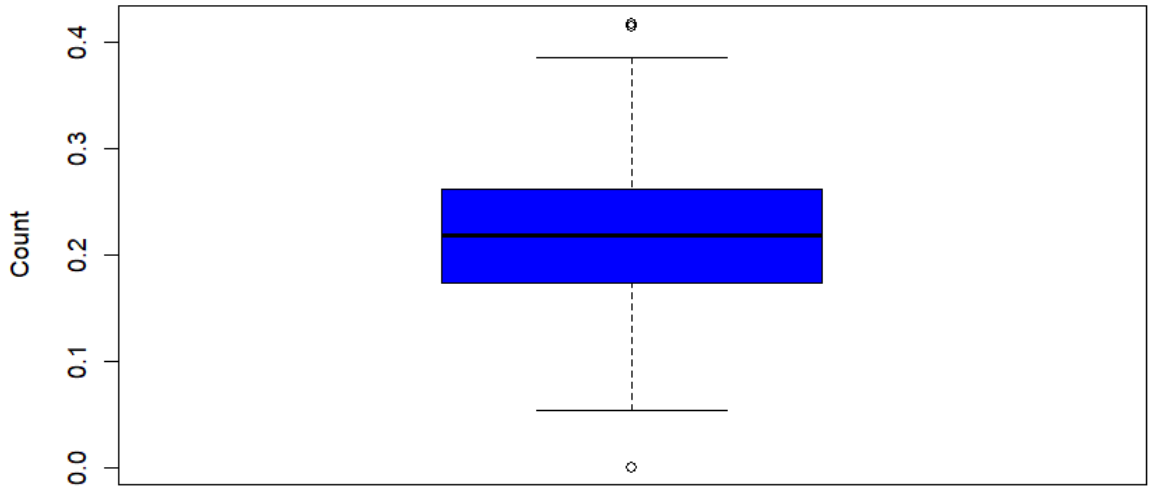

Figure 3.7.2: Boxplot of the percentage of COVID deaths

The plot infers that the percentage of COVID death data is normally distributed. To confirm this, I did the Shapiro-Wilk's test, QQ plot, and Kolmogorov-Smirnov test. The null hypothesis for these tests was "there is no significant difference between the percentage of COVID death data and a normal distribution".

```
Asymptotic one-sample Kolmogorov-Smirnov test

data:  pTotal_deaths

D = 0.030116, p-value = 0.9321

alternative hypothesis: two-sided


Shapiro-Wilk normality test

data:  pTotal_deaths

W = 0.99438, p-value = 0.2838
```

Figure 3.7.3: Results of the KS and SW test for the percentage of COVID deaths

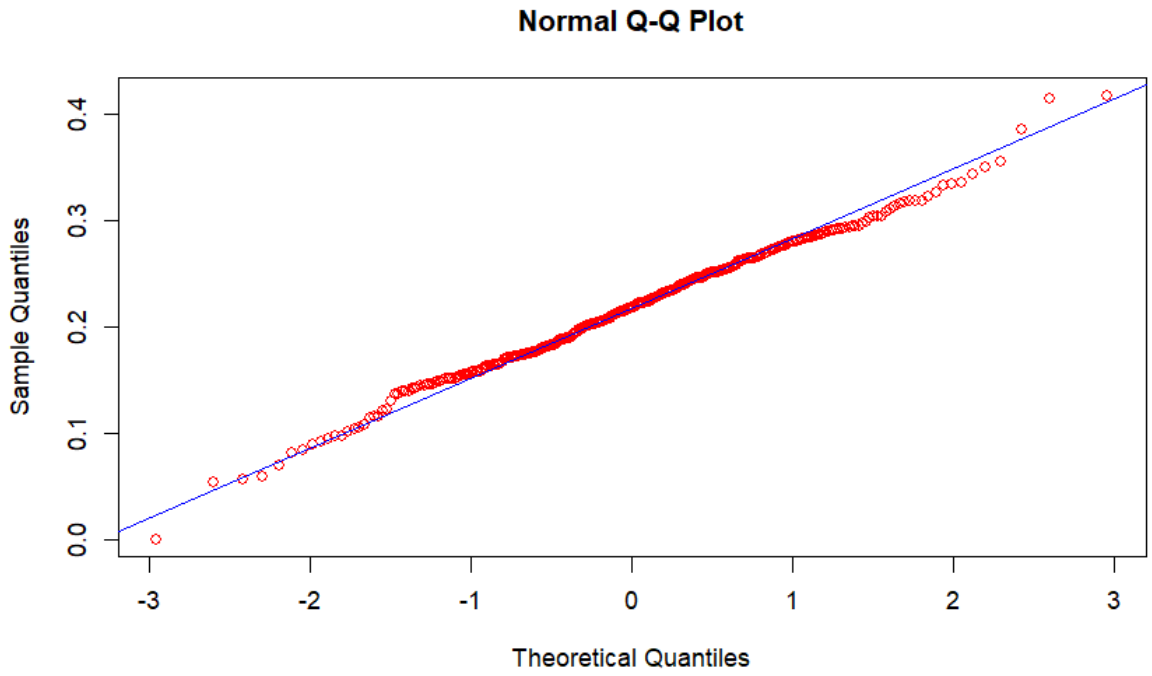

Figure 3.7.4: QQ plot of the percentage of COVID deaths

From the KS and SW tests results, the p-value is greater than 0.05 and this supports our initial assumption that the percentage COVID dataset is normally distributed. Hence it is safe to accept the null hypothesis. Additionally, I plotted a QQ plot for the data and as seen above, the observations lie approximately on a straight line. This confirms that the data is normally distributed.

3.8 MULTIVARIATE SCATTER PLOT

To understand how the multiple variables are related I plotted a multivariate scatter plot as a way to visualize the data to help classify and understand the relationships among the variables.

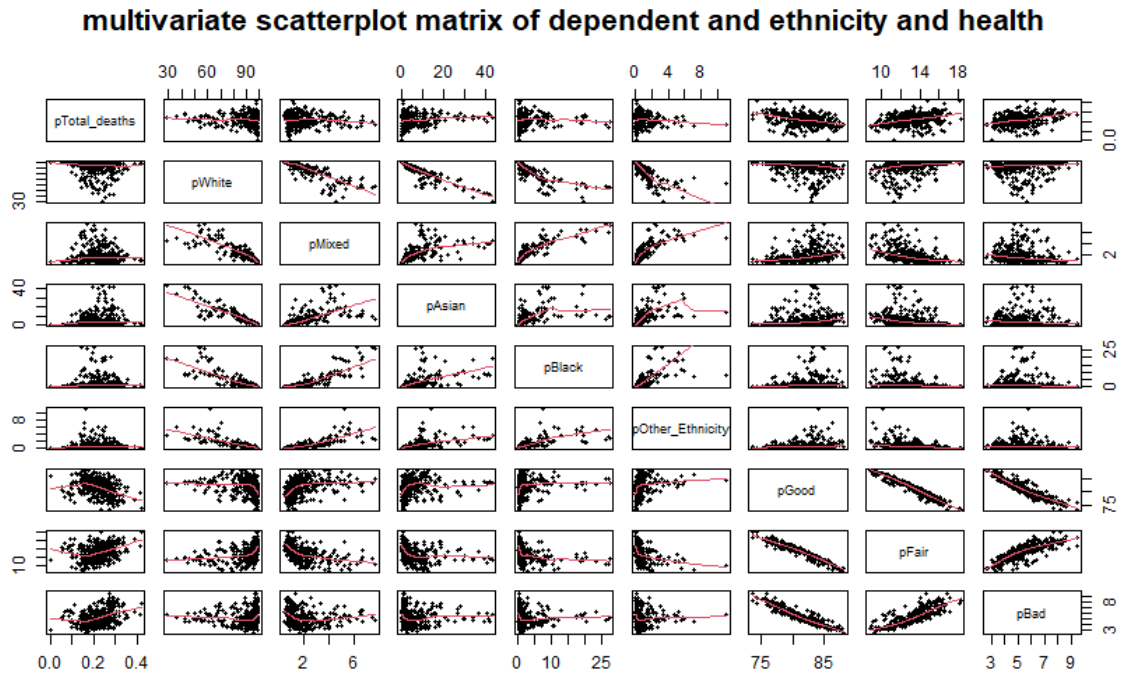

Figure 3.7.1: Multivariate Scatter plot of the percentage of COVID deaths and ethnicity and health

Looking at the scatterplot all the variables on ethnicity appear to have no correlation with COVID deaths but good health has a weak negative relationship while fair and bad health appear to have a weak positive relationship with COVID deaths.

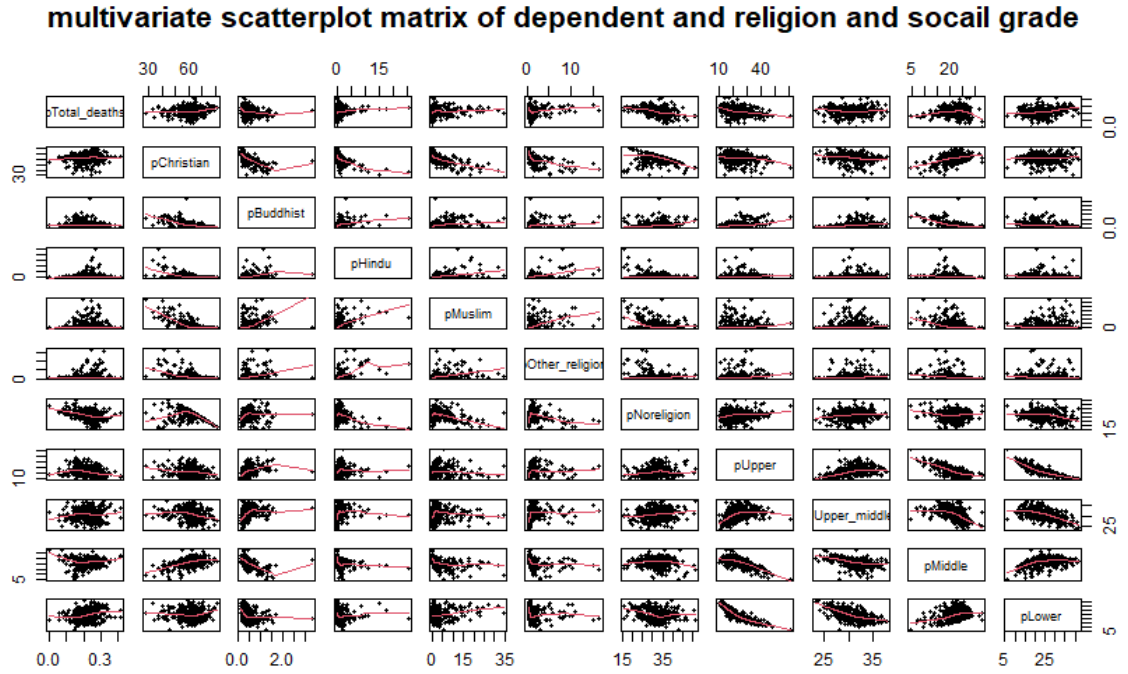

Figure 3.7.2: Multivariate Scatter plot of the percentage of COVID deaths and Religion and Social grade

Overall, most variables on religion seem to have no relationship with COVID deaths. But while no religion appears to have a negative relationship, Christians have a positive relationship.

The upper social grade appears to have a negative relationship with COVID deaths while upper middle, middle and lower appear to have a positive relationship.

3.9 CORRELATION MATRIX

To further analyze the strength of the relationship between the independent variables and COVID deaths, I did a correlation check between all independent variables and the target variable in a matrix format.

	pTotal_deaths	pWhite	pMixed	pAsian	pBlack	pOther_Ethnicity	pGood
pTotal_deaths	1.00	-0.11	0.02	0.13	0.08	0.05	-0.37

	pFair	pBad	pChristian	pBuddhist	pHindu	pMuslim
pTotal_deaths	0.33	0.37	0.15	-0.24	0.11	0.16

	pOther_religion	pNoreligion	pUpper	pUpper_middle	pMiddle	pLower
pTotal_deaths	-0.06	-0.42	-0.34	-0.04	0.10	0.31

Figure 3.8.1: Correlation matrix of the percentage of COVID deaths and the independent variables

Looking at the results, pMixed has almost no correlation with COVID deaths so I dropped it. All other variables showed some correlation so they moved to the next stage.

3.10 CORRELATION PLOT

I plotted a correlogram to visualize the correlation between the remaining independent variables and the dependent variable.

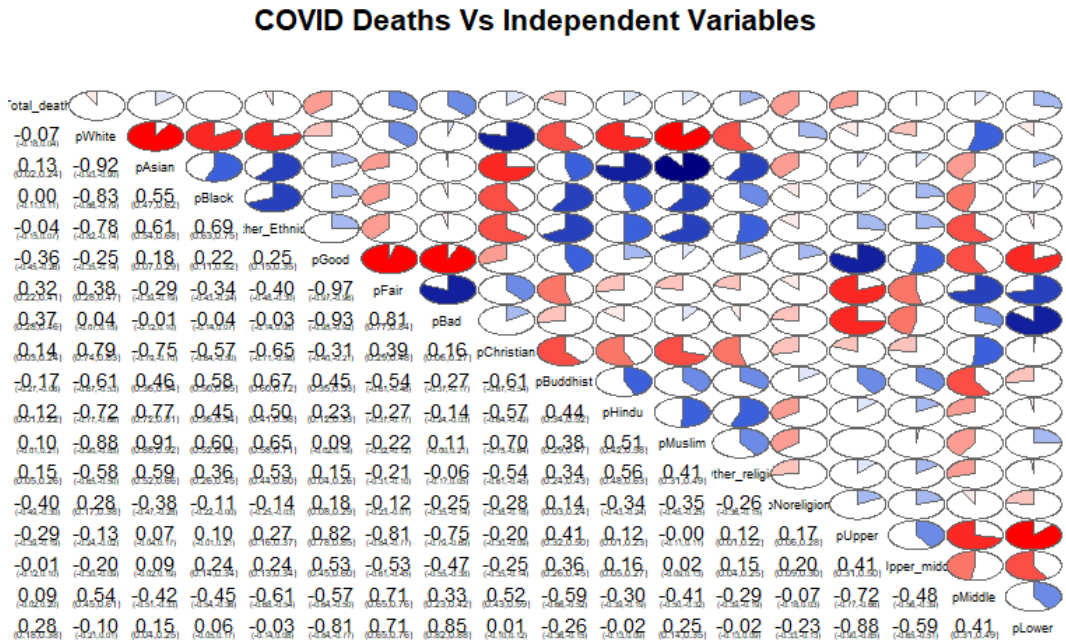

Figure 3.8.1: Correlation plot of the percentage of COVID deaths and the independent variables

The blue color represents positive correlations, while the red color represents negative correlations. All variables appeared to have some sort of correlation with the dependent variable so I proceeded.

3.11 CORRELATION TEST OF VARIABLES

I carried out correlation tests between the dependent and independent variables, checked for internal correlations between the independent variables and I also checked for partial correlations using both Pearson and Spearman methods. The internal correlation tests were done to see if there is any form of dependency between the variables or if there is an effect of one or more variables on other variables. While the Partial correlation test was done to check for a relationship between two variables while controlling for the effect of one other variable. There are some internal and partial correlations within the variables.

4 DATA ANALYSIS: RESULTS AND DISCUSSION

4.1 HYPOTHESIS TESTING:

4.1.1 Z-Test

I did a hypothesis testing between the first quarter of 2021 and the last quarter of 2020 COVID death rates. This was to test for differences in Covid death rate between quarters. I used the z-test as opposed to the t-test because the observations are more than 30. The null hypothesis was “there is no difference between the death rate in Q4 2020 and Q1 2021”

```
       Two-sample z-Test

data:  data$Q1_2021 and data$Q4_2020
z = 10.553, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 60.56990 88.20028
sample estimates:
mean of x mean of y
145.31677  70.93168
```

Figure 4.1.1.1: Z-test Results

The paired Z-test result gave a p-value less than 0.05 this shows that there is a difference between the death rate in the first quarter of 2021 and the fourth quarter of 2020. Hence, we can safely reject the null hypothesis.

4.1.2 Chi-Squared Test

I carried out the Chi-squared test to check for dependence between COVID deaths that fell below the average (High death rates) and those that fell above the average (low death rate).

I created a new categorical variable for the high and low death rates. Then ran a paired Chi-squared test on the categorical variable. The null hypothesis was "there is no statistically significant relationship between high and low death rates".

```
       Chi-squared test for given probabilities

data:  table
X-squared = 0.1118, df = 1, p-value = 0.7381

> output$observed

High  Low
 164  158
> output$expected
High  Low
 161  161
> output$stdres

      High        Low
 0.3343669 -0.3343669
```

Figure 4.1.1.2: Chi-squared test Results

The p-value of the test above is greater than 0.05 which means we can safely accept the null hypothesis. We can say that there is no statistically significant relationship between the high COVID death rate and low death rates i.e. both variables are independent.

4.2 FACTOR ANALYSIS*:*

Since I have a lot of independent variables in my dataset and I want to determine which of these variables influence the COVID deaths, I did a factor analysis to simplify the data and reduce the many variables into a fewer number of dimensions.

4.2.1 Kaiser Meyer Olkin (KMO) test:

I ran the KMO test on all the independent variables to determine how suited data is for factor analysis. Generally, if the measure of sampling adequacy (MSA) is 0.6 and above, factor analysis is done. Nevertheless, some authors keep this value at 0.5.

```
Kaiser-Meyer-Olkin factor adequacy

Call: KMO(r = cor(IV))

Overall MSA =  0.5

MSA for each item =

pWhite       pAsian     pBlack     pOther_Ethnicity       pGood        pFair
 0.5           0.5       0.5               0.5               0.5          0.5

 pBad       pChristian     pBuddhist       pHindu     pMuslim     pOther_religion
 0.5           0.5           0.5            0.5        0.5               0.5

 pNoreligion      pUpper     pUpper_middle      pMiddle           pLower
    0.5            0.5            0.5              0.5               0.5
```

Figure 4.2.1.1: KMO test Results

The overall MSA was 0.5 but using my discretion, I decided to proceed to factor analysis.

4.2.2 Eigen Values:

I calculated the eigenvalues and condensed the variance in the correlation matrix. Then I plotted a scree plot of the values. For the factor analysis, I selected factors whose eigenvalues were close to or greater than 1 and where the curve makes a relatively sharp drop (called "elbow") or where the slope levels off twice. Hence, I looked at the scree plot and first decided to go with 4 components that captured close to 90% variance in the dataset.

```
round(ev$values, 3)

 [1] 7.318 4.425 1.306 0.855 0.739 0.631 0.536 0.417 0.285 0.225 0.197 0.045
0.020 0.001 0.000 0.000 0.000
```

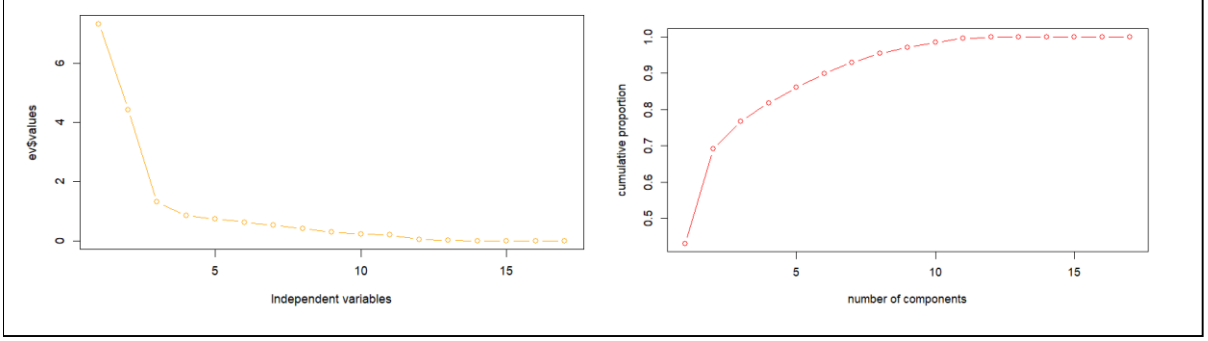

Figure 4.2.2.1: Eigenvalues and Scree plot with eigenvalues and eigen proportions

4.2.3 Principal Component Analysis (PCA):

I used PCA to display a multivariate data-table as summary indices (a smaller set of variables) to identify trends, clusters, jumps, and outliers.

After the first PCA, I weeded out the variables that had high communalities as they are non-contributing leaving only the dominant variables. I weeded out: Christians, Muslims, Buddhists, and white.

After removing the non-contributing variables, I again calculated the eigenvalues, plotted a scree plot, and calculated the PCA.

```
Principal Components Analysis
Call: principal(r = IV, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                    RC1   RC2   RC3   RC4   h2    u2 com
pAsian            -0.02  0.77  0.45 -0.18 0.82 0.179 1.7
pBlack             0.02  0.37  0.75  0.14 0.72 0.281 1.5
pOther_Ethnicity -0.08  0.41  0.81  0.02 0.83 0.174 1.5
pGood             -0.94  0.13  0.16  0.14 0.94 0.063 1.1
pFair              0.87 -0.15 -0.35 -0.09 0.91 0.087 1.4
pBad               0.92 -0.08  0.12 -0.19 0.90 0.099 1.1
pHindu            -0.11  0.86  0.20 -0.07 0.79 0.208 1.2
pOther_religion  -0.05  0.78  0.20  0.06 0.65 0.349 1.2
pNoreligion       -0.12 -0.48  0.05  0.70 0.74 0.261 1.8
pUpper            -0.91 -0.07  0.27 -0.03 0.91 0.087 1.2
pUpper_middle     -0.47  0.21  0.14  0.67 0.73 0.267 2.1
pMiddle            0.52 -0.09 -0.74 -0.04 0.84 0.161 1.8
pLower             0.92  0.05  0.09 -0.21 0.90 0.097 1.1

                       RC1  RC2  RC3  RC4
SS loadings           4.69 2.57 2.32 1.11
Proportion Var        0.36 0.20 0.18 0.09
Cumulative Var        0.36 0.56 0.74 0.82
Proportion Explained  0.44 0.24 0.22 0.10
Cumulative Proportion 0.44 0.68 0.90 1.00

Mean item complexity =  1.5
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  173.87  with prob <  2e-21

Fit based upon off diagonal values = 0.98
```

Figure 4.2.3.1: Second PCA of the independent variable

From the results, I selected Good, Hindu, other_ethnicity, and Noreligion as the variables for clustering based on strength of correlation with the target.

4.3 CLUSTERING

After factor analysis, I decided to cluster to see if the data might show something interesting and new if hierarchically classified. I clustered into 4 groups because, from the factor analysis, only 4 components are most significant. I performed two types of clustering:

4.3.1 Ward Hierarchical Clustering:

Using hierarchical clustering, I mapped the clusters into a hierarchy. This is to reflect inter-cluster similarities or similarities between features using the agglomerative or bottom-top clustering method.

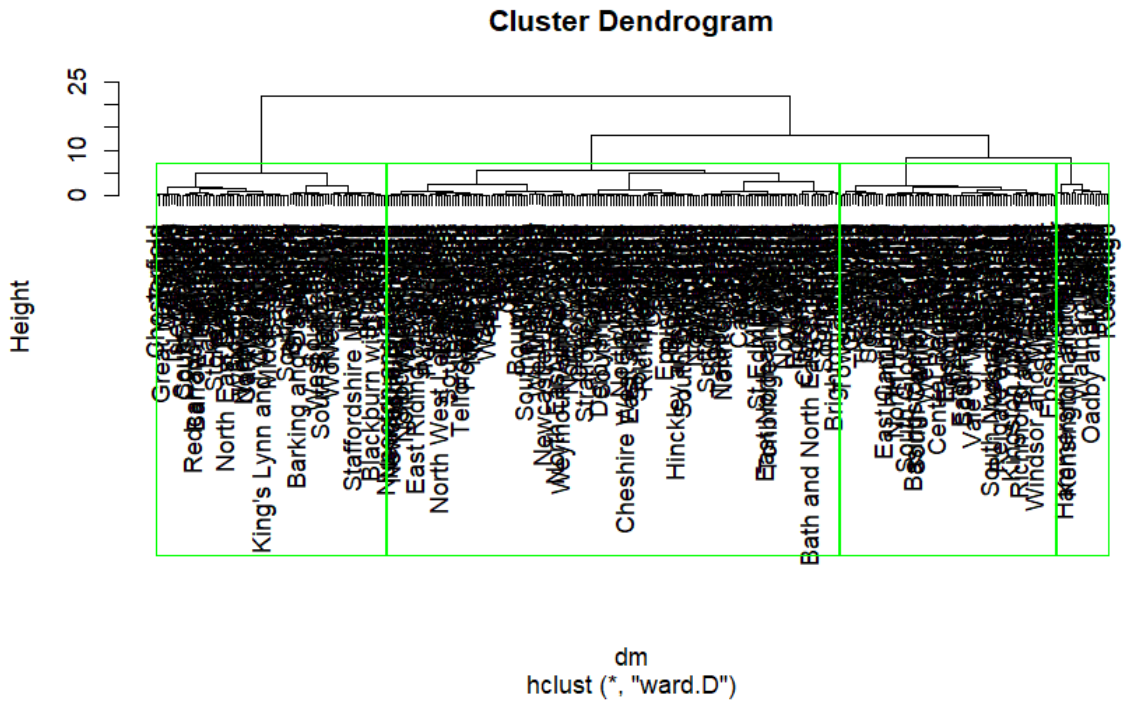

Figure 4.3.1.1: Cluster Dendrogram

As seen on the dendrogram above, a line was drawn from all districts (observation) then the closest clusters were combined, and so on until we were left with a single cluster.

4.3.2 K-Means Cluster Analysis:

Here I attempted to classify my data set into 4 clusters using the cluster library.

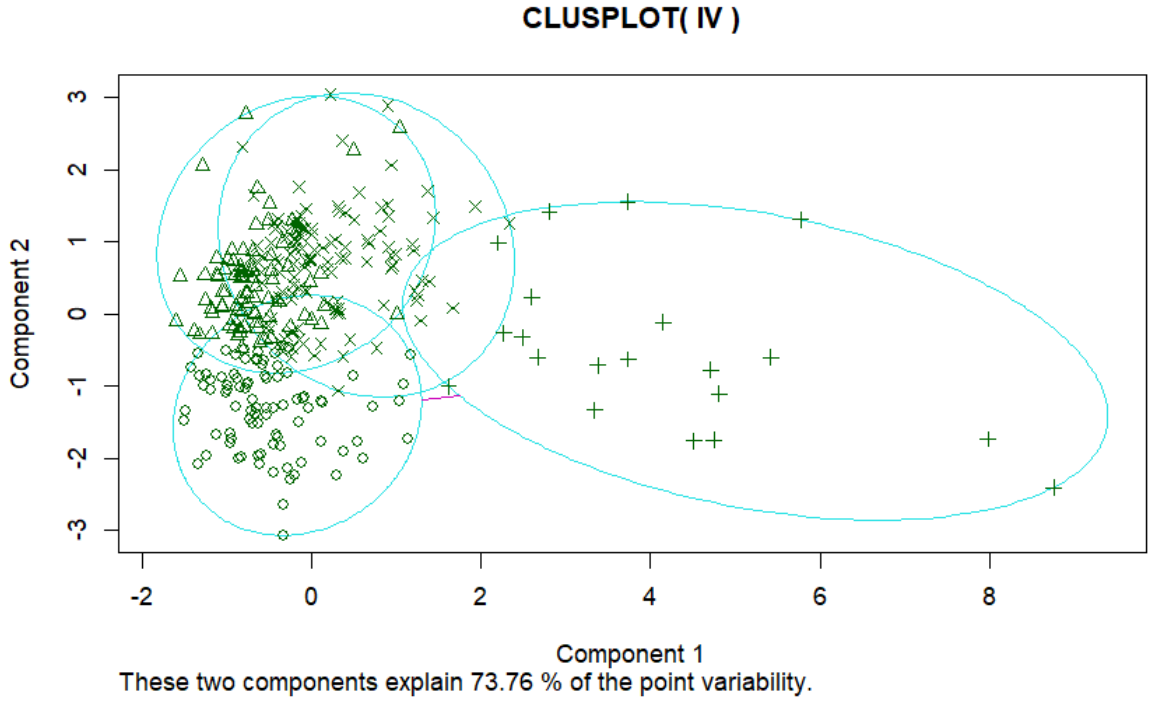

Figure 4.3.2.1: Cluster Plot using the Cluster Library

Above is the scatterplot of the first two principal components which captures 73.76% of the information about the data.

From the cluster analysis, we can see that what distinguishes each group from the others is distance. This means the clusters were formed based on similarities in terms of longitude and latitude.

4.4 REGRESSION

I then proceeded to build a multiple linear regression model to identify how the independent variables work together to explain the target variable.

4.4.1 MODEL 1

For model 1, I ran a linear regression on the target-dependent variable and all the independent variables. The summary of the model showed an error message that four of my independent variables have perfect multicollinearity. To fix this error, I took out the variables in the dataset that have a perfect correlation from the regression model and rerun the model.

```
Call:
lm(formula = pTotal_deaths ~ pWhite + pBlack + pOther_religion +
    Other_Ethnicity + pAsian + pGood + pFair + pChristian + pBuddhist +
    pHindu + pMuslim + pUpper + pMiddle + pUpper_middle)

Residuals:
      Min        1Q    Median        3Q       Max
-0.136812 -0.029582 -0.004332  0.032365  0.181575

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.740e-01  7.414e-01   0.505  0.61427
pWhite          -2.267e-03  4.321e-03  -0.525  0.60019
pBlack          -1.679e-03  5.194e-03  -0.323  0.74669
pOther_religion  6.968e-03  2.221e-03   3.137  0.00187 **
Other_Ethnicity -4.675e-06  1.890e-06  -2.474  0.01392 *
pAsian          -2.548e-04  4.775e-03  -0.053  0.95747
pGood           -4.984e-03  6.744e-03  -0.739  0.46049
pFair            9.183e-03  1.084e-02   0.847  0.39768
pChristian       4.127e-03  6.867e-04   6.010 5.26e-09 ***
pBuddhist       -9.213e-03  1.462e-02  -0.630  0.52897
pHindu           2.852e-03  3.071e-03   0.929  0.35383
pMuslim          2.034e-03  2.564e-03   0.793  0.42821
pUpper          -7.178e-04  1.074e-03  -0.668  0.50451
pMiddle         -3.971e-03  2.105e-03  -1.887  0.06014 .
pUpper_middle    5.787e-03  1.395e-03   4.148 4.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05092 on 307 degrees of freedom
Multiple R-squared:  0.3603,     Adjusted R-squared:  0.3311
F-statistic: 12.35 on 14 and 307 DF,  p-value: < 2.2e-16
```

Figure 4.4.1.1: Multiple regression model for all variables (Model1)

Since the p-value is less than 0.05, we reject the null hypothesis that the independent variables cannot explain the dependent variable (COVID deaths). Nevertheless, from the R-squared value, we can see that the independent variables can only estimate for 36% variation of the dependent variable. We have only 6 statistically significant variables meaning a lot of variables are unnecessary, hence this is a computationally expensive model. Therefore, it is not a good model.

4.4.1.1 Variance inflation factor (VIF):

Using the VIF we calculated the internal correlation between the independent variables. From the results, we can see there is high collinearity due to internal correlations within the independent variables. Hence, I reduced the number of variables to improve the model performance as high internal correlation is risky and can develop poor results.

```
> sqrt(vif(model1)) > 2  # if > 2 vif too high
pWhite     pBlack     pOther_religion     Other_Ethnicity      pAsian         pGood
 TRUE       TRUE           FALSE               TRUE               TRUE          TRUE
pFair     pChristian      pBuddhist         pHindu           pMuslim          pUpper
 TRUE       FALSE           FALSE             TRUE             TRUE            TRUE
pMiddle    pUpper_middle
 TRUE         FALSE
```

Figure 4.4.1.1.1: Variance inflation factor for independent variables

4.4.2 MODEL 2

Next, I used the four dominant variables representing each component after factor analysis in a multiple regression to see if the model will improve. I first plotted a correlation plot with the variables to see how they correlate.

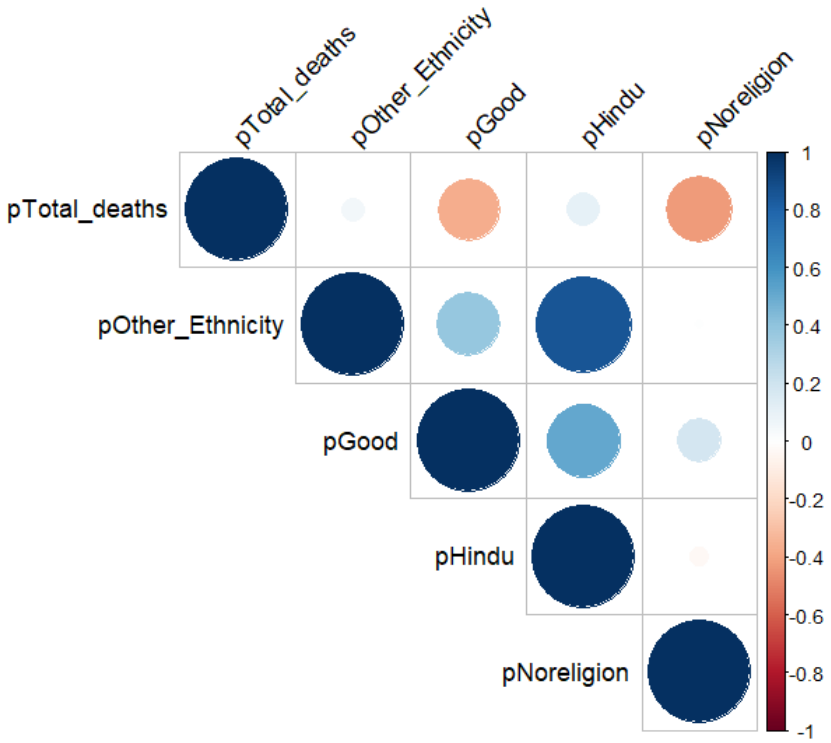

Figure 4.4.2.1: Correlation plot of variables from factor analysis

We can see that all the independent variables from the factor analysis correlate with COVID deaths with good health and noreligion being the strongest.

```
Call:
lm(formula = pTotal_deaths ~ pHindu + pGood + pOther_Ethnicity +
    pNoreligion)

Residuals:
      Min        1Q    Median        3Q       Max
-0.156366 -0.031493 -0.002215  0.032481  0.192965

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8404761  0.0827138  10.161  < 2e-16 ***
pHindu            0.0029340  0.0015488   1.894   0.0591 .
pGood            -0.0062766  0.0010616  -5.912 8.73e-09 ***
pOther_Ethnicity -0.0033760  0.0029114  -1.160   0.2471
pNoreligion      -0.0034725  0.0006002  -5.786 1.73e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05415 on 317 degrees of freedom
Multiple R-squared:  0.2529,     Adjusted R-squared:  0.2435
F-statistic: 26.83 on 4 and 317 DF,  p-value: < 2.2e-16

> sqrt(vif(model2)) > 2
          pHindu            pGood pOther_Ethnicity      pNoreligion

           FALSE            FALSE            FALSE            FALSE
```

Figure 4.4.2.2: Multiple regression and VIF of from the factor analysis (Model2)

From the results, R^2 indicates that the independent variables can only explain 25% of COVID deaths. We can also see that only two independent variables are significant and there are no internal correlations between the pairs from the VIF results.

I used ANOVA to test if model 1 and model 2 are statistically different using the F test. The null hypothesis was “there is no significant difference between the accuracy of model 1 and model 2”. Since the p-value was less than 0.05 (appendix D), model 1 (complex model) is significantly better than model 2 (simpler model). Hence, I rejected the null hypothesis and accept that model 2 is not a better model and that the accuracy of the model is poor.

4.4.2.1 Relative Importance of the Variables

To know which variable is the most important and to rank variables based on their contribution to the R-Squared value, I calculated the relative importance of the variables.

```
> calc.relimp(model2, type = c("lmg"), rela = TRUE)
Response variable: pTotal_deaths
Total response variance: 0.003876167
Analysis based on 322 observations

4 Regressors:
pHindu pGood pOther_Ethnicity pNoreligion
Proportion of variance explained by model: 25.29%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

                        lmg
pHindu           0.06049499
pGood            0.43462059
pOther_Ethnicity 0.01841689
pNoreligion      0.48646753
```

Figure 4.4.2.1.1: Relative importance of the independent variables

After calculating the relative importance of the 4 independent variables on COVID Deaths, results showed that No religion is most important followed by good health, Hindu religion, and then Other Ethnicity.

4.4.3 MODEL 3: STEPWISE APPROACH

I used the stepwise approach to search for the best model using model 1.

```
Call:
lm(formula = pTotal_deaths ~ pGood + pUpper_middle + pAsian +
    pChristian + pOther_religion)

Residuals:
      Min        1Q    Median        3Q       Max
-0.138515 -0.028312 -0.003864  0.031600  0.169209

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.5143185  0.0957501   5.371 1.52e-07 ***
pGood           -0.0095713  0.0011180  -8.561 4.97e-16 ***
pUpper_middle    0.0068127  0.0012275   5.550 6.05e-08 ***
pAsian           0.0038591  0.0006016   6.415 5.14e-10 ***
pChristian       0.0039876  0.0005644   7.065 1.03e-11 ***
pOther_religion  0.0052368  0.0017387   3.012  0.00281 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05128 on 316 degrees of freedom
Multiple R-squared:  0.3323,     Adjusted R-squared:  0.3217
F-statistic: 31.45 on 5 and 316 DF,  p-value: < 2.2e-16
```

Figure 4.4.3.1: Results of Model 3

The results of model 3 left us with just 5 variables. These five variables, good health, upper_middle social grade, and No religion, were significant and the R-squared is above 33%.

I did the ANOVA test on model 1 and model 3. The null hypothesis was “there is no significant difference between the accuracy of model 1 and model 3”. The p-value was greater than 0.05 (appendix

D) hence model 3 is significantly better than model 1. Hence, I rejected the null hypothesis and accepted that model 3 is the best model thus far.

I plotted a histogram of model 3 residuals and they appeared normally distributed. I plotted a scatter plot as well and the residuals looked random on the plot. Lastly, I carried out the KS test to test for the normality of the residuals with the null hypothesis that there is no significant difference between the residuals and a normal distribution. The p-value of the test was 0.4606 and this confirms that the residuals are normally distributed so we accept the null hypothesis.

I ran the VIF on the model and all variables have a result of less than 2 and which is great. I also checked for relative importance and good health was the highest followed by Christians, Asian ethnic groups, other_religion, and upper middle social grades sequentially.

```
> exp(coef(model3))
(Intercept)    pGood   pUpper_middle      pAsian    pChristian     pOther_religion
1.6724983   0.9904743   1.0068359     1.0038665   1.0039956          1.0052505
```

Figure 4.4.3.2: Odds ratios for model 3

After calculating the odds ratios for model 3, we can see that the upper middle class, Asian ethnicity, Christians, and other religions have an elevated risk by about 0.68%, 0.38%, 0.39%, and 0.52% respectively while people with Good health have a reduced risk of dying from COVID by about 1%. This means that the key factor associated with a reduced risk of COVID-related deaths is good health. Consequently, we can infer that having bad health would elevate the risk of dying from COVID.

5 CONCLUSION

I used multiple techniques to identify and rank the key factors associated with COVID death counts in England and Wales. From my analysis, good health conditions emerge as significant in reducing COVID deaths. This confirms the existing studies that the coronavirus was more severe in people who had pre-existing health conditions (Roy and Ghosh, 2020). People who identify with other religions had a higher risk of dying from COVID compared to Christians. This supports the report of (Gaughan *et al.*, 2021) that other religious affiliations died more than Christians. Furthermore, the results showed that being in the upper middle class slightly increases the risk of COVID death. According to the report by (ONS, 2020) jobs that involve proximity with people had higher risks and the upper-middle class involves supervisory, junior managerial, and administrative duties and this entails great interactions with colleagues and clients. Lastly, the Asian ethnicity have an increased risk of mortality rate. (Office of National Statistics, 2020) reported that most minority ethnicities had a higher mortality rate. Hence, the main finding from this project is that religion, ethnicity, social grade, and health conditions factors contribute to driving the incidence of COVID deaths.

The strengths of this work are that the target variable was normally distributed and that internally correlating independent variables were eliminated to improve accuracy. While the weakness of this project is the use of 2011 census data. This is an assumption that religious affiliations and other themes have not changed between 2011 and 2020 and this is very unlikely. Another assumption is that there is no link between any two observations. Ways to improve the model will be eliminating outliers from the data set and using a current dataset for population count. Additionally, adding more variables for each theme can increase the model's accuracy and precision. Further studies should include figuring out the reasons behind the unexplained COVID death risks as this will give insights into how to manage the death rates.

REFERENCE

Dong, D. *et al.* (2021) 'The Role of Imaging in the Detection and Management of COVID-19: A Review', *IEEE Reviews in Biomedical Engineering*, 14, pp. 16–29. Available at: https://doi.org/10.1109/RBME.2020.2990959.

Esposito, A. *et al.* (2021) 'Artificial intelligence in predicting clinical outcome in COVID-19 patients from clinical, biochemical and a qualitative chest X-ray scoring system', *Reports in Medical Imaging*, 14, pp. 27–39. Available at: https://doi.org/10.2147/RMI.S292314.

Gaughan, C.H. *et al.* (2021) 'Religious affiliation and COVID-19-related mortality: A retrospective cohort study of prelockdown and postlockdown risks in England and Wales', *Journal of Epidemiology and Community Health*, 75(6), pp. 509–514. Available at: https://doi.org/10.1136/jech-2020-215694.

Jcma) Rahmatizadeh, S., Valizadeh-Haghi, S. and Dabbagh, A. (2020) *The Role of Artificial Intelligence in Management of Critical COVID-19 Patients, JCMA) Journal of Cellular & Molecular Anesthesia*. Available at: https://nextstrain.org;

Karmakar, M., Lantz, P.M. and Tipirneni, R. (2021) 'Association of Social and Demographic Factors with COVID-19 Incidence and Death Rates in the US', *JAMA Network Open*, 4(1). Available at: https://doi.org/10.1001/jamanetworkopen.2020.36462.

Office of National Statistics (2020) *Why have Black and South Asian people been hit hardest by COVID-19?*

ONS (2020) *Coronavirus (COVID-19) related deaths by occupation, England and Wales: deaths registered between 9 March and 25 May 2020*.

Roy, S. and Ghosh, P. (2020) 'Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking', *PLoS ONE*, 15(10). Available at: https://doi.org/10.1371/journal.pone.0241165.

Salman, F.M. *et al.* (2020) *COVID-19 Detection using Artificial Intelligence*, *International Journal of Academic Engineering Research*. Available at: www.ijeais.org/ijaer.

World Health Organization (2022) *WHO Coronavirus (COVID-19) Dashboard*.

APPENDIX

APPENDIX A: SQL QUERY

```
SELECT *

FROM COVID_deaths, Ethnicity, Health_condition, Religion, Social_grade

WHERE COVID_deaths.LA_name = Ethnicity.District_E AND Ethnicity.District_E =
Health_condition.District_H

AND Health_condition.District_H = Religion.District_R AND Religion.District_R =
Social_grade.District_So
```

Figure 1: SQL Queries

APPENDIX B: NORMALITY TEST (QQ-PLOTS)

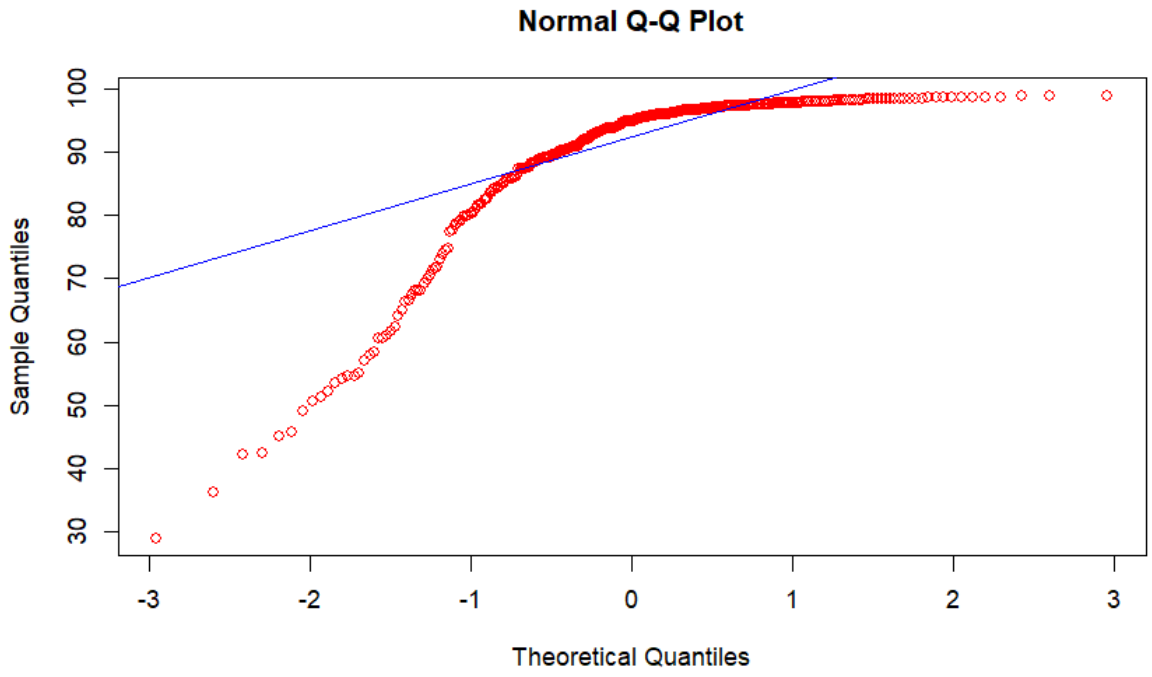

Figure 2: QQ plot for normality in White population data

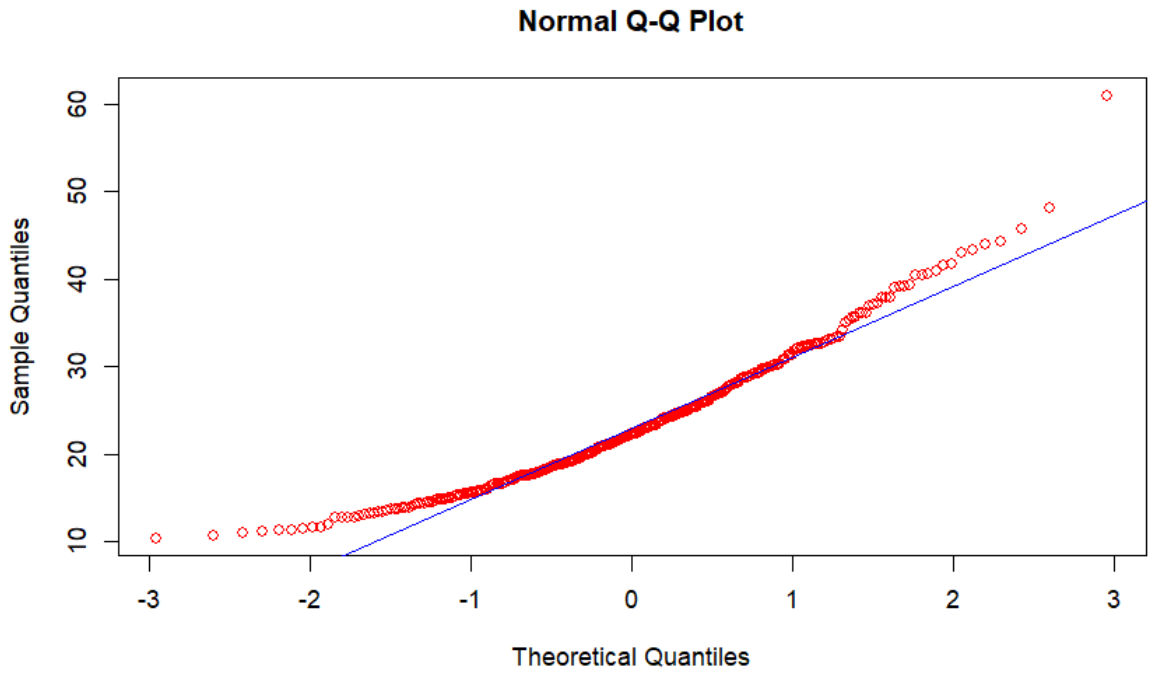

Figure 3: QQ plot for normality in Upper social status population data

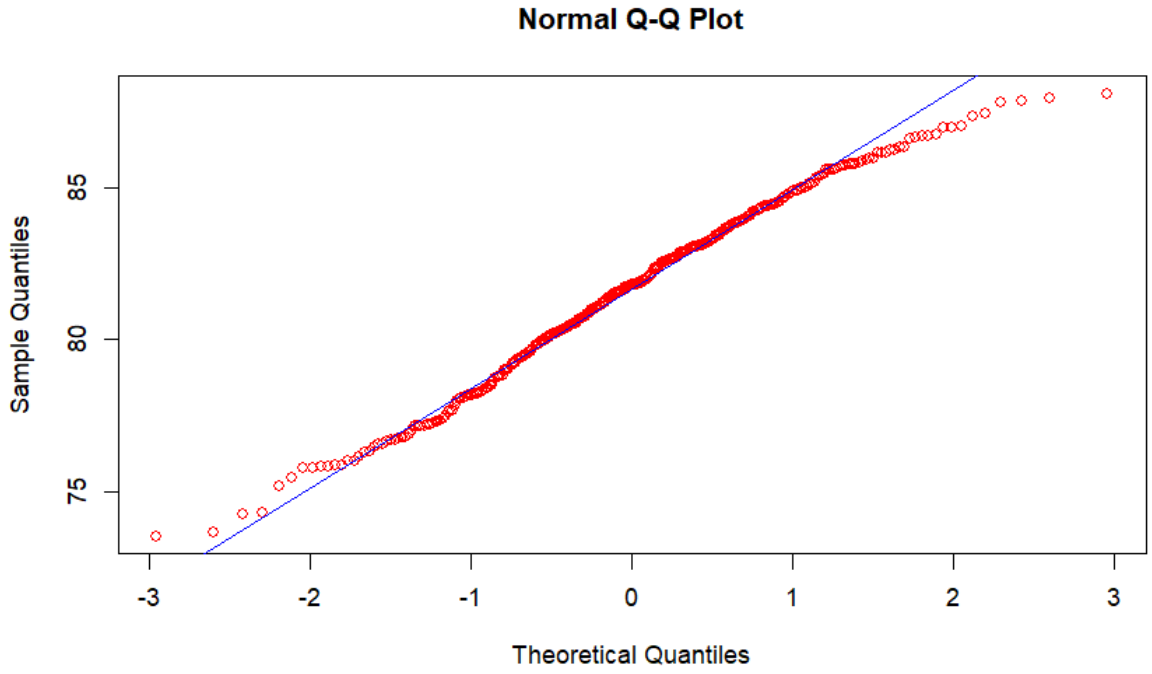

Figure 4: QQ plot for normality in Good health population data

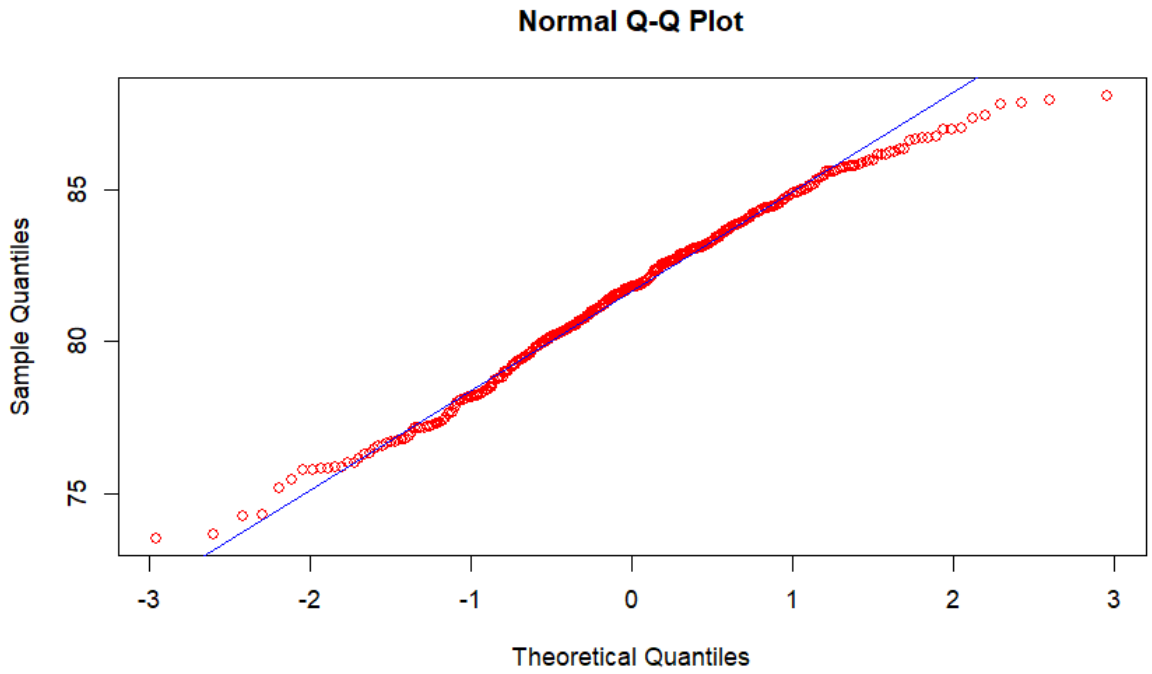

Figure 5: QQ plot for normality in Christian population data

APPENDIX C: HYPOTHESIS TESTING

```
#Hypothesis Testing

#Difference between COVID deaths in 2021 and 2020
colnames(data)

data <- within (data, Q1_2021 <- January_2021 + February_2021 + March_2021 +
April_2021)
data <- within (data, Q4_2020 <- September_2020 + October_2020 + November_2020 +
December_2020 )

#Checked for the Standard Deviation
sd(data$Q1_2021)
sd(data$Q4_2020)

#Z test to check if there is a difference between Q4 2020 and Q1 2021 deaths
z.test(data$Q1_2021,data$Q4_2020, mu=0, sigma.y=76.84573, sigma.x=100.464,
conf.level = 0.95)

#Creating a categorical variable

#Check for the mean of the total covid deaths
summary(df$pTotal_deaths)

#Create a categorical variable for above average mean (High) and Below average
mean (Low)

df$Average_death = ifelse(df$pTotal_deaths > 0.2170, "High", "Low")
table = table(df$Average_death)
table

#Chi squared test to check for relationship
chisq.test(table)

# simple chi-squared with additional outputs
output <- chisq.test(table)
output
output$observed
output$expected
output$stdres

# calculate effect size
CramerV(table, conf.level = 0.95)
```

Figure 6: Code snippets for hypothesis testing

APPENDIX D: ANOVA TESTS

```
Analysis of Variance Table

Model 1: pTotal_deaths ~ pWhite + pBlack + pOther_religion + Other_Ethnicity +
    pAsian + pGood + pFair + pChristian + pBuddhist + pHindu +
    pMuslim + pUpper + pMiddle + pUpper_middle
Model 2: pTotal_deaths ~ pHindu + pGood + pOther_Ethnicity + pNoreligion
  Res.Df     RSS  Df Sum of Sq      F   Pr(>F)
1    307 0.79593
2    317 0.92955 -10  -0.13362 5.1538 5.58e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: ANOVA test for significant difference between Model 1 and 2

```
> anova(model1, model3, test = "F")
Analysis of Variance Table

Model 1: pTotal_deaths ~ pWhite + pBlack + pOther_religion + Other_Ethnicity +
    pAsian + pGood + pFair + pChristian + pBuddhist + pHindu +
    pMuslim + pUpper + pMiddle + pUpper_middle
Model 2: pTotal_deaths ~ pGood + pUpper_middle + pAsian + pChristian +
    pOther_religion
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    307 0.79593
2    316 0.83083 -9 -0.034897 1.4956 0.1485
```

Figure 8: ANOVA test for significant difference between Model 1 and 3