Sources of data

The primary data sources mainly include official statistics from the census 2021 data set which is accessible publicly through the data portal of NOMIS and COVID-19 records for the authority districts in England. A systematic approach was adopted which helped to align the district names from the older dataset of COVID-19 with the revised 2023 model to harmonize the data used in census 2021. The merging of the dataset of COVID-19 ensures a consistent basis for comparative analysis.

Dependent Variable

In this analysis, the dependent variable is Total Covid Deaths, which represents the number of deaths attributed to COVID-19 in each local authority.

Independent Variables

The independent variables are divided into three categories: age group distributions, household depravity and household compositions. The age group variables include population counts across different age brackets, ranging from Aged 4 years and under to Aged 85 years and over. Household depravity variables represent the number of households that are deprived ranging from Household is deprived in any dimension, Household is deprived in one dimension to Household is deprived in four dimensions while the household composition variables represent the number of households with varying sizes, from 1 person in a household to 8 or more people in a household.

Exploratory Data Analysis (EDA)

To get a clear understanding of the dataset, some data exploration was done. Checking the structure of the dataset revealed that the dataset contained 296 observations of 27 variables which include the dependent variable(Total Covid deaths). It was also revealed that the dataset contained missing values that needed to be handled. Records with missing values were removed from the dataset, the dependent and independent were then converted to numeric values to ensure data type consistency for numeric computations.

Standardization

The numeric columns in the dataset were standardized by transforming them to per thousand. In regression modeling, predictors with different scales can make coefficients difficult to compare. Standardization puts all predictors on a comparable scale, simplifying interpretation (James et al., 2013). Transforming to percentages or per thousand is ideal for making data more interpretable in real-world terms, which is important for policy and demographic analyses (Kaufman and Schervish, 2004).

Distribution Analysis

In data analysis, knowing a dataset's distribution is essential since it affects the selection of statistical methods and the reliability of the findings. Both visual and numerical techniques may be used to evaluate the distribution of a dataset. Histograms, box plots, and Q-Q (quantile-quantile) plots are examples of visual tools that offer intuitive insights on the existence, skewness, and form of outliers in datasets (Field, 2013). Statistical tests such as the Shapiro-Wilk test and the Kolmogorov-Smirnov test are frequently used to assess normalcy for numerical verification. According to Ghasemi and Zahediasl (2012), these tests provide p-values, and a significant result ($p<0.05$) usually means that the data does not follow a normal distribution.

The Shapiro-Wilk normality test evaluates whether a given dataset is normally distributed.

Null Hypothesis (H_0): The data is normally distributed.

Alternative Hypothesis (H_1): The data is not normally distributed.

The Shapiro-Wilk test on our dependent variable resulted in a W value of 0.78303, which indicates how closely the data follows a normal distribution. A value closer to 1 suggests normality; a value significantly less than 1 suggests deviation from normality (OpenAI, 2024).

The p-value was < 2.2e-16, which is extremely small (practically zero).

Since the p-value is much smaller than the typical significance level (0.05), we reject the null hypothesis (H_0).

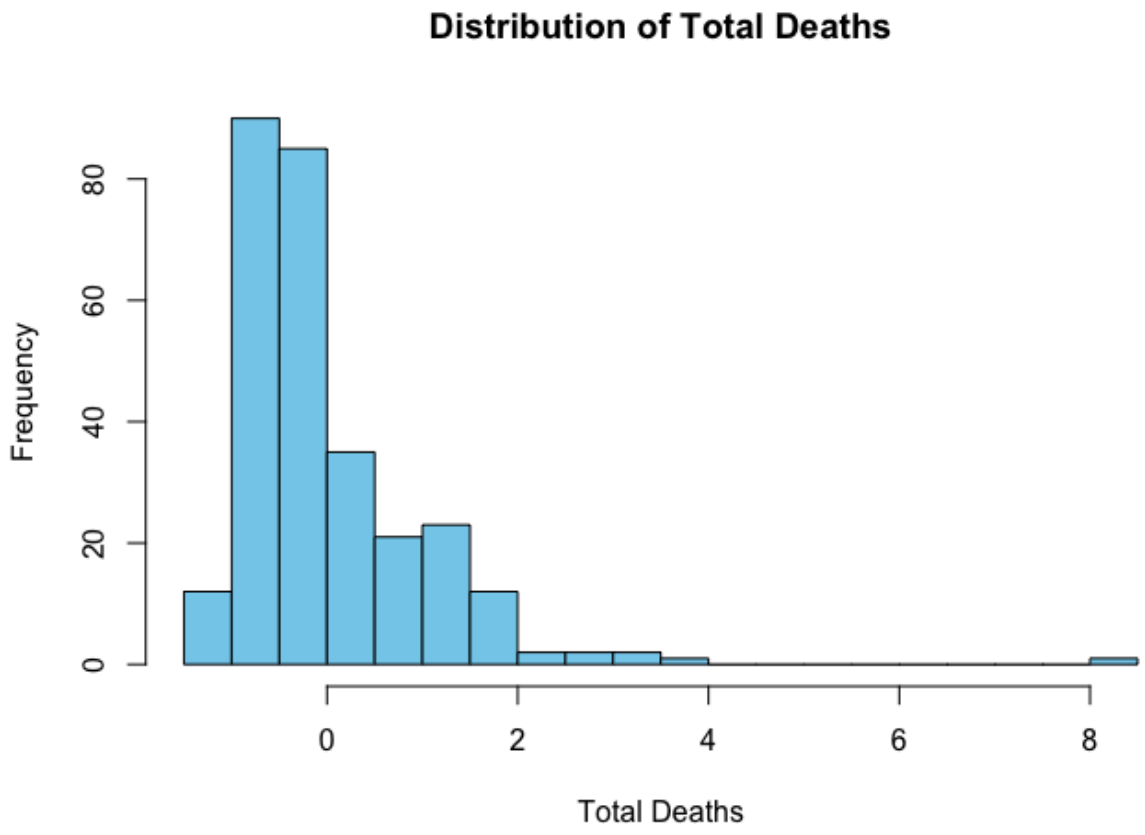

Fig.1: Histogram of the dependent variable

The distribution of the "Total Covid deaths" variable is shown Fig.1 above. The majority of observations are concentrated in the bottom end of the distribution, which is heavily skewed to the right. A small number of values deviate sharply to the right, indicating the existence of extreme or outlier values. As the overall number of deaths rises, the frequency gradually decreases. Because the data is not normally distributed, statistical modeling, including regression, may be impacted by the existence of extreme values.

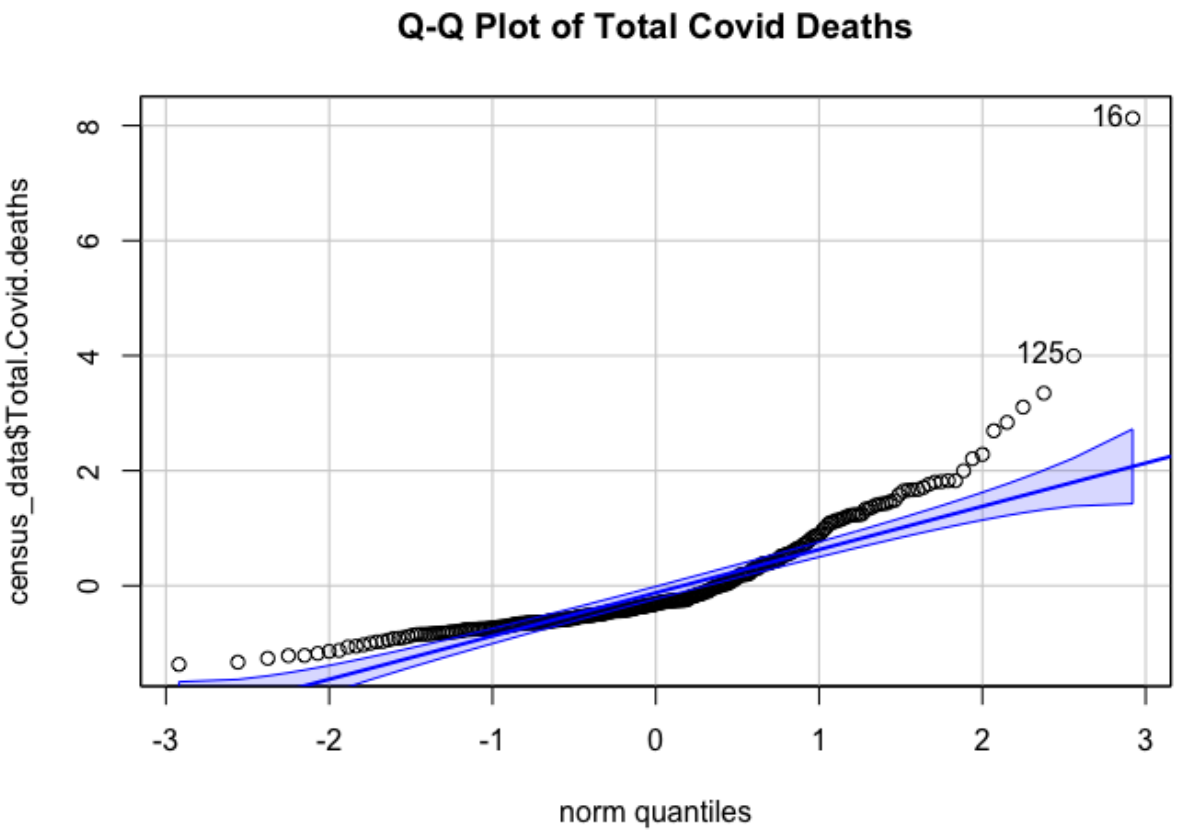

Fig.2: Q-Q plot of the dependent variable

Figure 2 above further proves the result of the histogram by comparing the quantiles of the raw "Total Covid Deaths" data against the quantiles of a standard normal distribution. The points deviate significantly from the diagonal line, especially at the tails. This confirms that the raw data is not normally distributed.

Through the Shapiro test and the Histogram and Q-Q plots, it was established that the dependent variable was not normally distributed. This indicated that the data is unsuitable for parametric statistical analysis and would need to be transformed. The dependent variable underwent two transformations: square root followed by logarithmic.

The Shapiro-Wilk test on the transformed variable Total.Covid.deaths resulted in a Test Statistic (W) value of 0.98479 and a p-value of 0.3135. The W value of 0.98479 is close to 1, which suggests that the data is now much closer to following a normal distribution compared to earlier tests. The p-value of 0.3135 is greater than the conventional significance threshold (0.05). This means we fail to reject the null hypothesis of normality, indicating that the data can be considered normally distributed.

The transformations applied (square root followed by log) have successfully improved the normality of the Total.Covid.deaths variable.

The variable now passes the Shapiro-Wilk test for normality, so it can be treated as approximately normal for subsequent parametric statistical analyses.

The results were then confirmed visually using a histogram and Q-Q plot to ensure the data aligns with a normal distribution

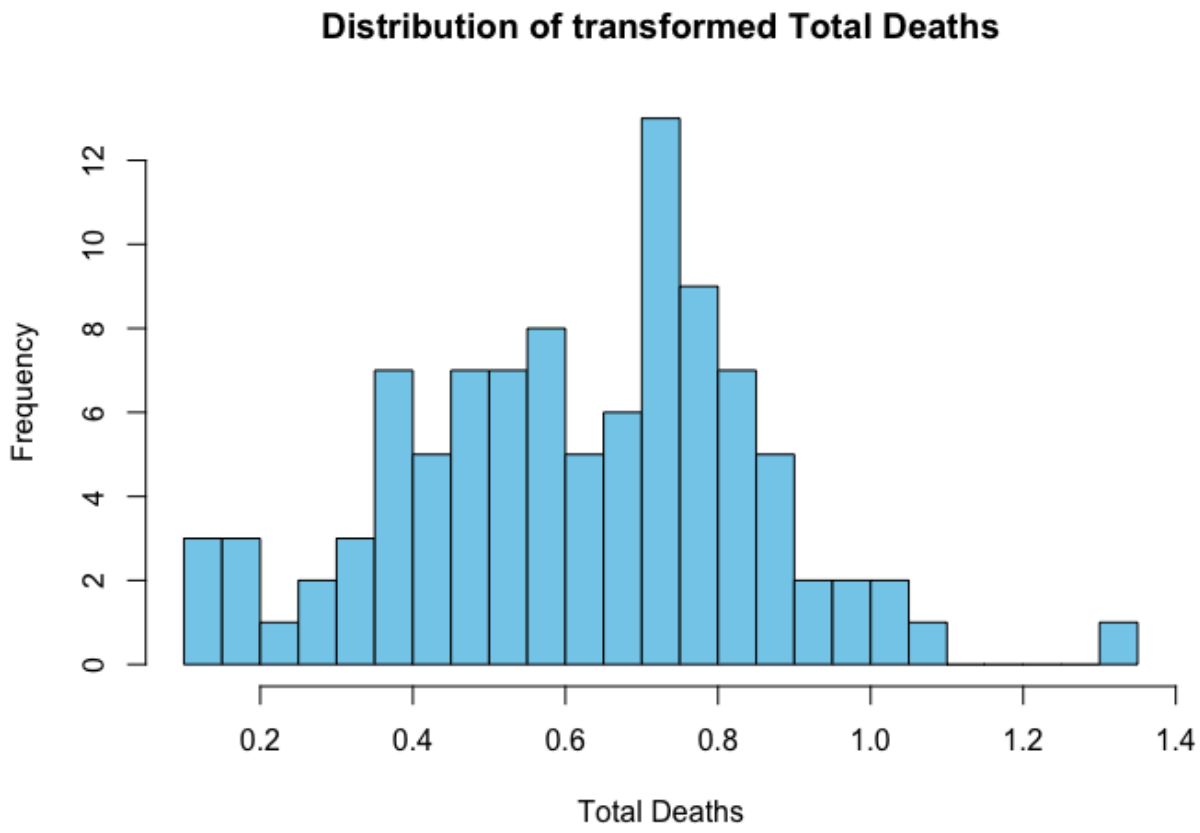

Fig.3: Histogram of the dependent variable after transformation

Figure 3 represents the distribution of "Total Deaths" after applying a transformation. The transformation has made the distribution more symmetric, approaching a normal distribution.

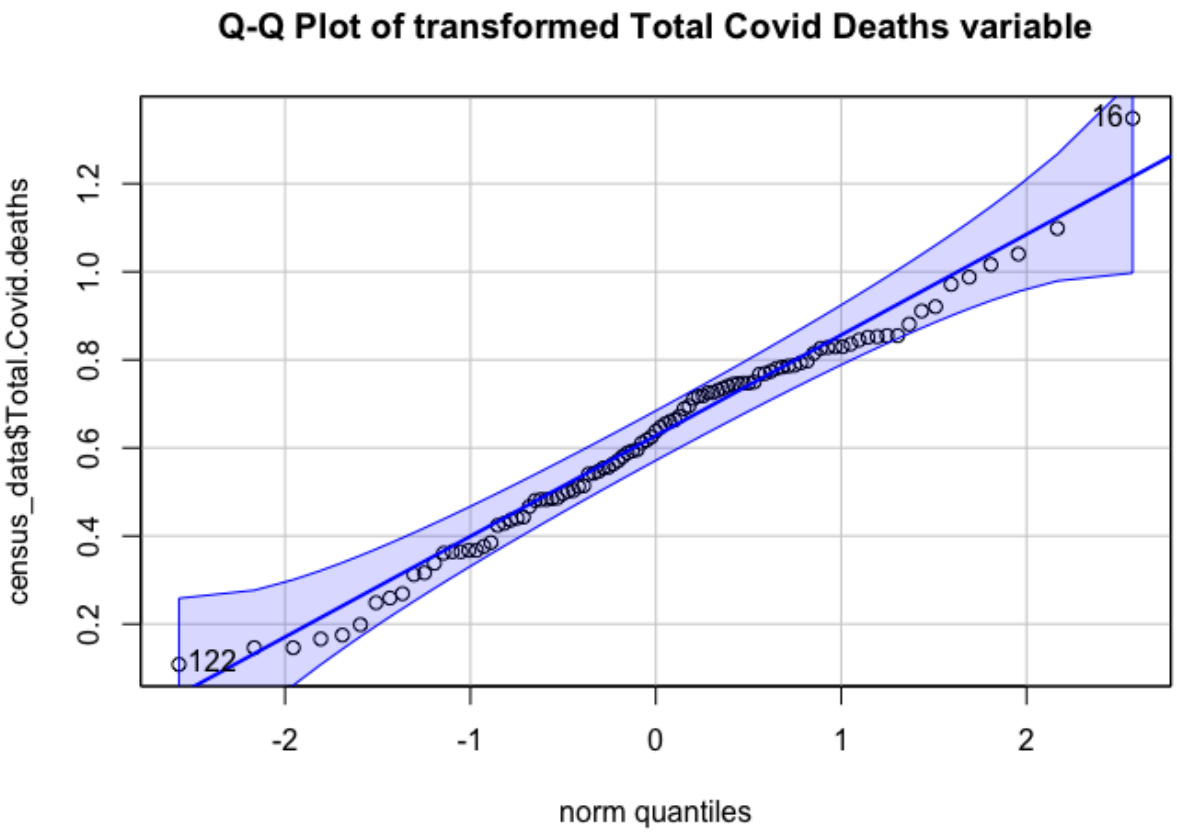

Fig.4: Q-Q plot of the dependent variable after transformation

Figure 4 represents a Q-Q plot of the transformed "Total Covid Deaths" data against a normal distribution. After the transformation, the points align much better with the diagonal line, particularly in the center of the distribution. There is some deviation at the extremes, but overall, the transformation improves normality.

Correlation Analysis

Using R's cor() function, the correlation between the dependent variable (Total.Covid.deaths) and all independent variables (IDVs) in the dataset showed that the following variables are strongly positively correlated with Total.Covid.deaths:

- Household.is.deprived.in.two.dimensions (0.747)
- Household.is.deprived.in.one.dimension (0.738)
- Age groups: Aged.5.to.9.years, Aged.10.to.15.years, Aged.50.to.64.years, and others.

These variables are positively associated with higher COVID-19 deaths. For example, areas with more households deprived in certain dimensions or specific age groups tend to have higher COVID-19 deaths.

Variables like Household.is.deprived.in.three.dimensions (0.673) and Aged.65.to.74.years (0.644) also show moderate positive correlations with COVID-19 deaths. These variables are still relevant but contribute less to the overall association than the strongest correlates.

Variables such as X8.or.more.people.in.household (-0.115) and X7.people.in.household (-0.075) have weak negative correlations. Other household size variables (X1.person.in.household, X2.people.in.household, etc.) show minimal or no significant correlation.

Household size appears to have a weak relationship with COVID-19 deaths.

Further exploration of the relationship between the dependent variable and certain independent variables was done using scatter plots.

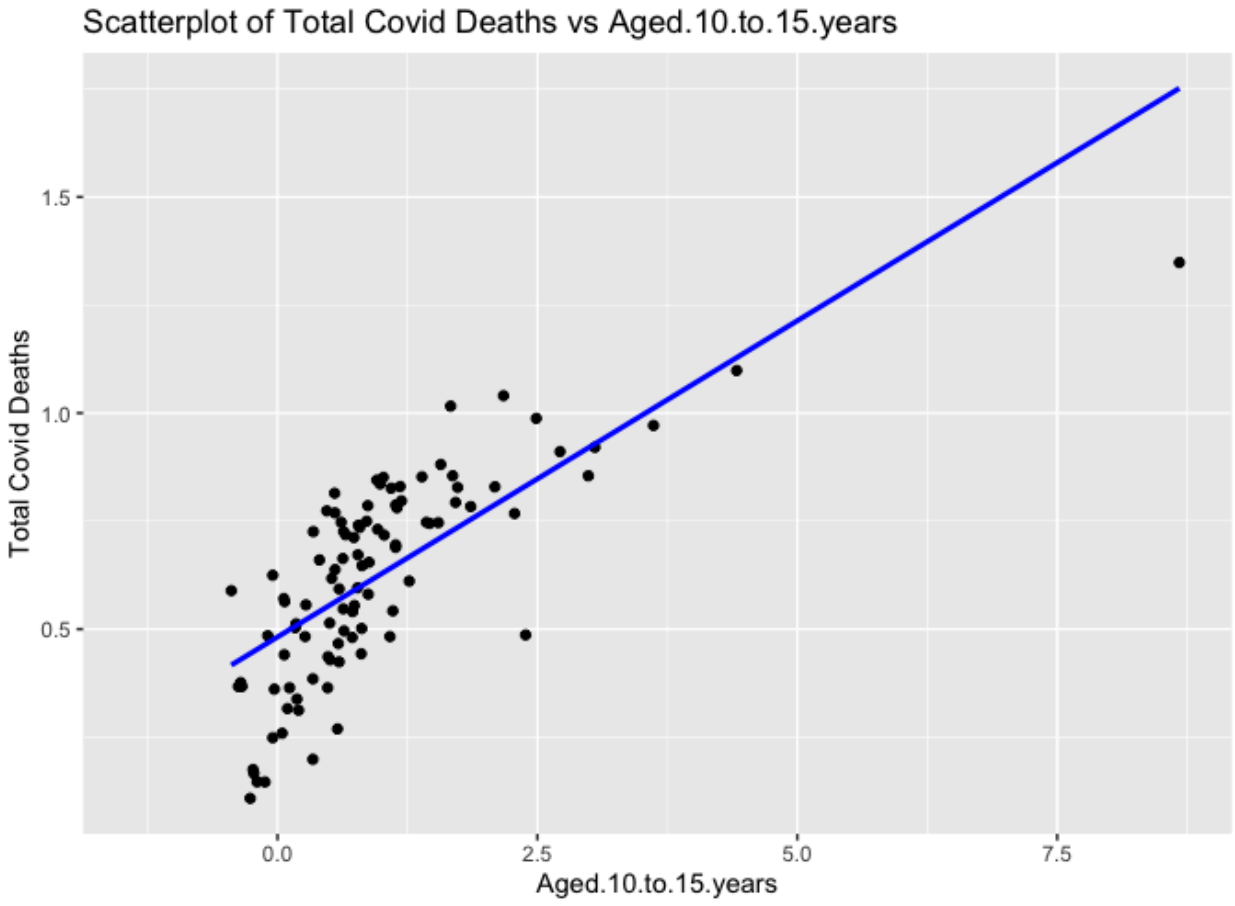

Fig.5: Scatter plot of Total Covid deaths vs Aged 10 to 15 years variable

Figure 5 shows a positive linear relationship between the proportion of children aged 10 to 15 years and total COVID deaths, as evidenced by the upward slope of the regression line. This suggests that as the proportion of children aged 10 to 15 years in a population increases, there is a tendency for total COVID deaths to increase as well.

Hypothesis Development:

- Null Hypothesis (H_0): The proportion of younger children in a population has a statistically significant positive effect on total COVID deaths.
- Alternative Hypothesis (H_1): The proportion of younger children in a population has no statistically significant effect on total COVID deaths.

Fig.6: Scatter plot of Total Covid deaths vs X7 people in household variable

Figure 6 shows a weak negative relationship between the average number of people in a household and total COVID deaths, as indicated by the slight downward slope of the regression line. This suggests that populations with larger households tend to have slightly lower COVID deaths.

Hypothesis Development:

- Null Hypothesis (H_0): The average number of people in a household has no statistically significant effect on total COVID deaths.
- Alternative Hypothesis (H_1): The average number of people in a household has a statistically significant negative effect on total COVID deaths.

Principal Component Analysis

The Kaiser-Meyer-Olkin (KMO) statistic assesses the suitability of a dataset for factor analysis by measuring the proportion of variance among variables that could be attributed to shared variance (common factors) rather than unique variance. A higher KMO value (typically above 0.6) indicates that the data is appropriate for factor analysis (Kaiser, 1974).

Computation of the KMO of the independent variables in the dataset resulted in a value of 0.88 which means the data is highly suitable for factor analysis. MSA values for each variable range from 0.82 to 0.94. This suggests that all variables are adequate for inclusion in factor analysis. The lowest MSA values are for X1.person.in.household (0.82) and X4.people.in.household (0.83), which are still above the acceptable threshold of 0.50.

Since all MSA values are above 0.80 and the overall MSA is high, the dataset shows a high proportion of shared variance suitable for identifying underlying latent factors through techniques like Principal Component Analysis (PCA).

Variables with higher MSA values contribute more strongly to shared variance, while lower MSA values indicate variables that might have less common variance. The high individual MSA values mean no variables need to be excluded based on sampling adequacy.

PCA Table

Component	Eigenvalue	Variance_Explained	Cumulative_Variance
1	1.418082e+01	59.086735308	59.08674
2	6.655541e+00	27.731421731	86.81816
3	1.388006e+00	5.783356969	92.60151
4	9.499600e-01	3.958166831	96.55968
5	2.433210e-01	1.013837408	97.57352
6	1.573554e-01	0.655647672	98.22917
7	1.275932e-01	0.531638453	98.76080
8	1.047072e-01	0.436280165	99.19708
9	8.686013e-02	0.361917221	99.55900
10	2.798326e-02	0.116596903	99.67560
11	1.909005e-02	0.079541893	99.75514
12	1.739645e-02	0.072485199	99.82763
13	1.081651e-02	0.045068801	99.87269
14	7.077834e-03	0.029490976	99.90219

15	5.663646e-03	0.023598526	99.92578
16	4.286017e-03	0.017858403	99.94364
17	2.690462e-03	0.011210259	99.95485
18	2.519303e-03	0.010497097	99.96535
19	2.404938e-03	0.010020576	99.97537
20	2.065819e-03	0.008607578	99.98398
21	1.680846e-03	0.007003526	99.99098
22	1.286367e-03	0.005359864	99.99634
23	5.112929e-04	0.002130387	99.99847
24	3.667810e-04	0.001528254	100.00000

Table 1: PCA table

Eigenvalues represent the amount of variance explained by each principal component in a dataset. According to the Kaiser Criterion, components with eigenvalues greater than 1 are typically considered significant and retained for analysis (Kaiser, 1960).

The first two components (PC1 and PC2) in Table 1 have eigenvalues of 14.18 and 6.66, respectively, indicating they capture most of the variance. PC1 explains 59.1% of the variance, and PC2 explains an additional 27.7%. Together, the first two components explain 86.8% of the total variance, indicating that most of the information in the dataset can be captured by these two components. By PC4, 96.6% of the variance is explained, meaning subsequent components add little additional value.

The PCA analysis demonstrates that most of the dataset's variance can be captured using just a few components. This reduces complexity and makes further analysis and visualization easier.

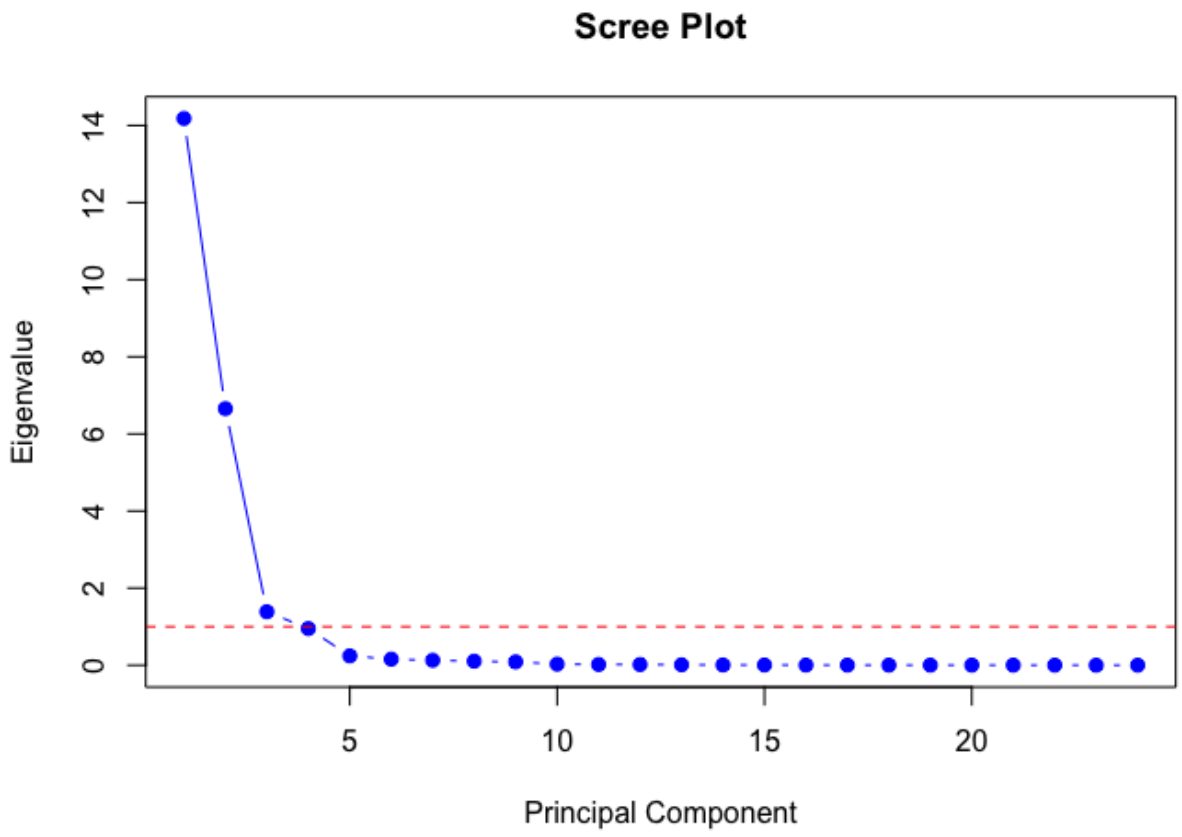

Fig.7: Scree plot of PCA analysis

Figure 7 shows the eigenvalues of each principal component in descending order. The steep drop-off in eigenvalues (elbow) occurs after the third or fourth principal component. This suggests that most of the variance in the dataset can be explained by the first 3–4 principal components. Components after the elbow have relatively small eigenvalues, meaning they contribute little to the overall variance. The Cumulative Scree plot also showed the same pattern.

Modelling

The regression model fits the dependent variable (DependentVar) using the first four principal components (PC1, PC2, PC3, and PC4) as predictors.

```
regression_model <- lm(DependentVar ~ PC1 + PC2 + PC3 + PC4, data = regression_data)

Residuals:

    Min       1Q   Median       3Q      Max

-0.46495 -0.10585  0.03775  0.11646  0.25096


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  0.478147   0.020566  23.249   <2e-16 ***

PC1          0.039327   0.003675  10.701   <2e-16 ***

PC2         -0.010601   0.004748  -2.233   0.0279 *

PC3          0.005911   0.009155   0.646   0.5201

PC4         -0.010340   0.013090  -0.790   0.4316

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.1572 on 94 degrees of freedom

  (187 observations deleted due to missingness)

Multiple R-squared:  0.5638,  Adjusted R-squared:  0.5453

F-statistic: 30.38 on 4 and 94 DF,  p-value: 3.18e-16
```

Residuals are the differences between the observed values and the predicted values of the dependent variable in a regression model, representing the portion of the data not explained by the model (Montgomery et al., 2012).

A median Residual close to zero (0.03775), indicated that the model does not exhibit a consistent bias. A residual range of -0.46495 to 0.25096 shows the distribution of prediction errors, with most falling within ±0.1. Extreme residuals indicate a few outliers where the model performs less well.

The coefficients in a regression model represent the contribution of each principal component to predicting the dependent variable, reflecting the strength and direction of their relationship (James et al., 2013).

An Intercept (0.4781) represents the baseline value of the dependent variable when all principal components are zero.

PC1 (Estimate = 0.03933, $p < 0.001$). A highly significant predictor (p-value < 2e-16), indicating that variation in PC1 is strongly associated with changes in the dependent variable.

We have demonstrated that higher PC1 scores, which represent regions with a greater proportion of younger populations, are statistically significantly associated with an increase in the dependent variable, thereby failing to reject the null hypothesis that younger children in a population has a statistically significant positive effect on total COVID deaths.

PC2 (Estimate = -0.01060, $p = 0.0279$). A statistically significant predictor at the 5% level.

We have demonstrated that higher PC2 scores, which represent regions with larger household sizes, are statistically significantly associated with a slight decrease in the dependent variable, thereby supporting the alternative hypothesis and rejecting the null hypothesis that household size has no effect on total COVID deaths.

PC3 (Estimate = 0.00591, $p = 0.5201$). Not statistically significant ($p > 0.05$). This suggests that variation in PC3 does not meaningfully contribute to predicting the dependent variable.

PC4 (Estimate = -0.01034, p = 0.4316). Not statistically significant (p > 0.05). Variation in PC4 also does not significantly predict changes in the dependent variable.

In terms of the model fit, the Residual Standard Error(RSE) which is the average distance between observed and predicted values was 0.1572. A lower RSE indicates better model performance. The Multiple R-Squared(R^2) of 0.5638 indicates that 56.38% of the variation in the dependent variable is explained by the model. While not perfect, this R^2 suggests a reasonably good fit given the complexity of social data. An Adjusted R-Squared of 0.5453 which is a small difference from the R^2 indicates that including PC3 and PC4 (which are not significant) does not substantially harm the model.

F-Statistic (30.38, p < 0.001) tests whether the model as a whole is statistically significant. A high F-statistic and very low p-value (3.18e-16) confirm that at least one predictor (likely PC1) significantly explains the variance in the dependent variable.

The refined regression model fits the dependent variable (DependentVar) using only the first two principal components (PC1 and PC2) as predictors.

```
regression_model_refined <- lm(DependentVar ~ PC1 + PC2, data = regression_data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.44455 -0.10040  0.04647  0.11560  0.25042 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.477339   0.020427  23.368   <2e-16 ***
PC1          0.039461   0.003654  10.799   <2e-16 ***
```

PC2 -0.010061 0.004698 -2.142 0.0348 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1566 on 96 degrees of freedom

(187 observations deleted due to missingness)

Multiple R-squared: 0.5581, Adjusted R-squared: 0.5488

F-statistic: 60.61 on 2 and 96 DF, p-value: < 2.2e-16

Min, 1Q, Median, 3Q, Max values show that residuals are reasonably small and centered around zero, indicating a good model fit. Residual Standard Error (RSE) Value of 0.1566 which indicates better model fit.

About 55.81% of the variance in the dependent variable is explained by the predictors (PC1 and PC2). This indicates a moderately strong relationship.

An Adjusted R-squared Value of 0.5488 which is a small drop from the multiple R-squared value shows that the model remains robust with only two predictors.

F-statistic Value 60.61 on 2 and 96 degrees of freedom and p-value <2.2e-16 confirms that the overall model is statistically significant, meaning the predictors (PC1 and PC2) together explain a significant portion of the variance in the dependent variable.

The original model included four predictors (PC1, PC2, PC3, PC4). However, PC3 and PC4 were not statistically significant and did not contribute much to the model. Removing them simplified the model and slightly improved the Adjusted R-squared (from 0.5453 to 0.5488). The improved model uses fewer predictors and is more efficient, explaining almost the same amount of variance. The refined model is simpler, statistically robust, and explains the variance efficiently with fewer predictors, making it a better choice than the original model.

Conclusion

The findings showed that there was a strong correlation between higher COVID-19 mortality rates and places with higher PC1 scores, which indicate areas with a larger percentage of younger people. The null hypothesis, according to which there is a statistically significant positive correlation between COVID-19 mortality and the number of younger children in a population, is supported by this finding. Also, there was a small but statistically significant negative correlation between COVID-19 fatalities and locations with higher PC2 scores, which indicate bigger family sizes. This implies that bigger households could be protective possibly because of stronger caring skills.

Strong positive correlations were found between COVID-19 fatalities and certain age groups and variables such home deprivation in one or two categories. On the other hand, factors related to household size, such as "8 or more people in a household," showed slight negative correlations, suggesting that they had little bearing on pandemic outcomes. These results demonstrate the complex and multifaceted influence of demographic and socioeconomic variables on death rates.

This study emphasizes how crucial it is to combine demographic and socioeconomic data in order to comprehend the factors that influence health outcomes during a pandemic. These findings can be used by policymakers to pinpoint high-risk locations and better distribute resources, especially in places with high rates of poor or certain age distributions. More variables and their connections should be investigated in future research.