THE SOCIOECONOMIC FACTOR THAT AFFECTS COVID-19 DEATHS IN ENGLAND

MODULE – DS7006

SUBMITTED BY – IYEOSE SIMON UHUMUAVBI

TABLE OF CONTENTS

ABSTRACT

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has had devastating effects worldwide–with England experiencing severe mortality. This study investigates the relationships between socioeconomic factors—age, gender, level of education, health status, socioeconomic deprivation, and living conditions—and COVID-19-related deaths in England in 2020.

Using data exploration and analysis, the study highlights key vulnerabilities among specific population groups. Findings therefore reveal that older adults, men, individuals with lower education levels, those with pre-existing health conditions, and residents of socioeconomically deprived areas or poor living conditions face heightened risks of COVID-19 mortality.

The final analysis shows that individuals living in two-bedroom flats, high levels of deprivation experienced higher mortality rates compared to those with no deprivation during the COVID-19 pandemic in England. These findings underscore the importance of implementing targeted public health strategies and initiatives tailored to address the specific risk factors associated with these conditions.

These results emphasizes the need for targeted public health interventions and policies to address and protect vulnerable populations. Also, an understanding of these demographic impacts can lead to more inclusive strategies for managing on-going health crises and mitigating the effects of emerging variants.

It is important to recognize the limitations of this study, notably the use of 2011 census data instead of more recent data from 2021. Despite these constraints, the findings offer valuable insights that can serve as a foundation for developing more inclusive and effective public health policies, taking into account the diverse demographic factors impacting COVID-19-related deaths.

Keywords: COVID-19, socioeconomic factors, mortality, pandemic, England

INTRODUCTION

In 2019, the worldwide outbreak of coronavirus disease (COVID-19) led to a significant threat to global public health (Shi et al., 2020). COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was first identified in December 2019 in patients linked to a seafood market in Wuhan City, Hubei Province, China (Zhu et al., 2020). Viruses naturally change and evolve as they spread among people over time. These changes that result in a virus that is significantly different from earlier versions, is classified as a "variant." Scientists identify these variants by sequencing the virus's genetic material and analyzing the differences to track any notable changes (WHO, 2021). The World Health Organisation also reported that since 2020, SARS-CoV-2, the virus responsible for COVID-19, has been mutating and spreading worldwide; and these changes have led to the emergence of numerous variants across various regions. Significant variants are categorized into three groups: Variants Under Monitoring (VUM), Variants of Interest (VOI), and Variants of Concern (VOC) (WHO, 2021).

These variants have led to mortalities over the years across the world. Statistics have reported that as aat May 2023, approximately 6.86 million deaths have been reported globally due to COVID-19, with Europe being about 2.16 million. However, by the end of December 2022, England had recorded over 177,000 deaths within 28 days of a positive COVID-19 test, corresponding to a crude mortality rate of 313 per 100,000 people (UK Government, 2022).

England, being one of the largest of the four countries in the United Kingdom–and located in the southern part of the island of Great Britain. It is also among the nations hardest hit by the impact of COVID-19. This study therefore examines the impact of demographic factors—age, gender, level of education, health status, socioeconomic deprivation and living conditions—on COVID-19-related mortality in England. Through data exploration and analysis, this study aims to uncover the relationships between these variables and COVID-19 deaths, highlighting the distinct vulnerabilities faced by different populations.

RESEARCH QUESTIONS

1. How does age influence the severity of COVID-19 outcomes or mortality?
2. How does gender impact the risk and outcomes of COVID-19-related mortality?
3. How does the level of education correlate with COVID-19-related outcomes?

4. In what ways does health status, particularly pre-existing conditions like cardiovascular diseases, diabetes, and respiratory disorders, intersect with COVID-19 severity?
5. How do living conditions and socioeconomic deprivation contribute to COVID-19 severity and mortality?
6. How can these findings guide the development of more inclusive and effective public health policies?

HYPOTHESIS

1. **Age and COVID-19 Mortality**

Older individuals are significantly more likely to experience mortality compared to younger age groups due to age-related health vulnerabilities.

2. **Gender Disparities**

Men are at a higher risk of COVID-19 mortality than women due to biological differences, occupational exposures, and health behaviors.

3. **Educational Level**

Individuals in England with lower levels of education are more likely to experience severe COVID-19 outcomes, reflecting disparities in health literacy and access to healthcare.

4. **Health Status**

Pre-existing health conditions such as cardiovascular disease, diabetes, and obesity significantly increase the likelihood of severe COVID-19 outcomes in England.

5. **Socioeconomic Deprivation**

Higher levels of socioeconomic deprivation in England are associated with greater COVID-19 mortality due to poor living conditions, reduced healthcare access, and occupational risks.

6. **Policy Recommendations**

Addressing demographic-specific vulnerabilities in England will lead to more inclusive and effective public health policies, reducing COVID-19 disparities across different population groups.

BACKGROUND / LITERATURE REVIEW

In 2020, when the first wave of the COVID-19 pandemic hit, it left a profound impact on public health (Borges do Nascimento et al., 2021) and particularly with England experiencing more waves of mortality (Flynn et al. 2020). Following the report, studies were conducted on how some socioeconomic factors could have affected the rate of mortality. These factors include; age, gender, level of education, health status, socioeconomic deprivation, and living conditions (Khanijahani, et al., 2021; Davies et al., 2020; ONS 2021; Tang et al., 2022). Understanding these factors is important for the development of specific public health strategies to curb the impact of other variants or future health crises.

Age

Age was, and is still one of the factors that led to the death of many who had coronavirus in 2020. In 2020, clinical evidence increasingly indicated that the risk of death from Coronavirus disease 2019 (COVID-19) was closely linked to age (e.g., Zhou et al., 2020). Early observations by clinicians revealed that COVID-19 fatality rates increased consistently with age. Unlike other respiratory illnesses, it did not follow the typical U-shaped risk curve, which shows heightened vulnerability among both infants and older adults (Raoult et al., 2020).

In Europe, which comprises four countries—England, Scotland, Wales, and Northern Ireland, studies showed that death was recorded in older ones compared to younger ones (Flynn et al. 2020, etc). As at 13th August 2020, approximately 91% of hospital deaths due to COVID-19 in England occurred among individuals aged 60 and older. In Scotland, as of 12th August 2020, around 76% of deaths were reported among those aged 75 and older. In Northern Ireland, by 31st July 2020, about 80% of deaths were recorded in individuals aged 75 and above. Similarly, in Wales, by the same date, nearly 90% of COVID-19-related deaths were among those aged 65 and older. The World Health Organization and the US Centers for Disease Control and Prevention also identified older adults—defined as individuals aged 60 and older by the WHO, and 65 and older by the CDC—as a particularly vulnerable population (Sasson, 2021; Davies et al., 2020).

Gender

Nguyen et al. (2021) found that adult males with COVID-19 had higher odds of mortality compared to females across all age groups, with the disparity being most significant in the 18–30 age range. The CDC in 2020, stated that men account for 54% of COVID-19 deaths (Nguyen et al., 2021). Additionally, a meta-analysis of 3,111,714 global cases revealed that male patients had higher odds of mortality compared to female patients (Peckham et al., 2020).

Similarly, the study by Williamson et al. (2020) revealed that men showed higher mortality rates compared to women; and this was associated with the biological makeup of the man, his behaviours as well as occupations they engage in. This is similar to what Peckham et al. (2020) explained, that early findings suggest that declining estrogen levels in postmenopausal women may contribute to increased inflammatory cytokine production following SARS-CoV-2 infection, indicating a potential protective role of estradiol against hyperinflammatory immune responses linked to COVID-19 mortality (Averyanova et al., 2022). Conversely, testosterone, the male sex hormone, suppresses immune activity, with hypo-androgenism linked to elevated inflammatory markers and altered immune cell profiles. Sex-based differences in immune system aging also play a role. Males experience an age-related decline in B cells and faster immune aging, which, combined with the strong association between advanced age and COVID-19 severity, may further explain the observed sex bias in morbidity and mortality.

Level of Education

Zhuo and Harrigan (2023) reported in a study that education levels are strongly associated with higher COVID-19 mortality. They also stated that education should be recognized as a significant risk factor for COVID-19 mortality and addressed in efforts to mitigate the virus's impact on disadvantaged communities. This effect extends beyond the influence of vaccination rates, poor health, or economic disadvantages, as low-education communities appear to have collective cultural practices and literacy challenges that increase their vulnerability to COVID-19. Similarly, Marmot (2020) stated that the level of qualification of the individuals, especially those with higher levels make it easier for them to seek medical help compared to those with lower status who end up having and living with poorer health outcomes.

Health Status

In a study carried out by Choi (2021), the mortality rate for patients with underlying health conditions was 12%, which was four times higher than the rate for patients without such conditions. This clearly states that there is a strong connection between the mortality rate of individuals who contact COVID and those with an underlying health conditions. Neurological conditions, cardiovascular conditions and pneumopathies are examples of conditions that can increase the rate of mortality (Sousa et al., 2020).

Socioeconomic deprivation

Another factor that led to a higher rate of mortality during the COVID-19 pandemic was social deprivation due to barriers to healthcare, occupations, and the living conditions of the individuals (ONS, 2021).

Living conditions

Due to the mode of transmission of the virus–airborne, the living conditions of the several individuals led to an increased mortality rate. The number of individuals living in a room, the adherence to which social distancing was not properly upheld (Public Health England, 2020).

METHODOLOGY

This project adopts a thorough and multifaceted approach, utilizing advanced statistical techniques such as regression modeling to examine the impact of various factors on COVID-19 death rates in England. The objective is to generate meaningful insights that can inform decision-making by policymakers, healthcare practitioners, and public health experts. These insights will aid in the development of targeted strategies that address the specific needs of different population groups, fostering a more equitable and robust healthcare system capable of withstanding future health emergencies.

For the analytical process, the R programming language, accessed through the RStudio platform, has been selected due to its open-source nature and its extensive ecosystem of statistical libraries and packages. By leveraging R's capabilities, this project will implement a variety of analytical models, chosen based on their performance, precision, and alignment with the research parameters. The ultimate goal is to pinpoint the key variables that significantly influence the risk of COVID-19-related mortality in England, providing a foundation for evidence-based interventions and risk mitigation efforts.

IMPLEMENTATION

DATA PREPARATION

Dependent Variable; A list of the total deaths recorded from the covid-19 pandemic in all the 326 local authorities as at 2018 is provided in the module. This is going to be the dependent variable that will be used along side with other data set that would serve as the independent variable.

Independent Variable; The independent variables are downloaded from Nomis website. Nomis is a service provided by Office for National Statistics (ONS), the UK's largest independent producer of official statistics. ("Nomis - Official Census and Labour Market Statistics," n.d.).

The data set downloaded covers the population in the local authority districts for only England. These data sets include the 296 LA_name, LA_codes, population data for all the districts in different columns and saved in CSV format. Having a total of twenty-four (24) independent variable.

Social Factors:

SN	THEMES	VARIABLE	NOTE
01	Gender	-Male -Female	Gender identity by sex. Covers the total population aged 16 and over
02	Age band	-Children: 0-15 -Youth: 16-34 -Adult: 35-54 -Pre-elderly: 55-74 -Elderly: 75 and over	Covers the Total population form age 0 and over and provides more detail for the older age brackets.
03	General Health	-Good health -Very good health -Bad health -Very bad health	Covers the total population aged 16 and over

Economic Factors:

SN	THEMES	VARIABLE	NOTE
01	Deprivation	-No dimension deprivation -One-dimension deprivation -Two-dimension deprivation -Three-dimension deprivation -Four-dimension deprivation	Households by deprivation dimensions
02	Qualification level	-No qualifications -Low level qualification; Level 1, 2 and entry level qualifications, -High level qualification; Level 3 and above qualifications -Other qualification; Apprenticeship and others	All usual residents aged 16 years and over
03	Accommodation type	-One bedroom -Two bedrooms -Three bedrooms -Four bedrooms and more.	Accommodation type based on number of rooms

❖ UPDATING THE COVID DEATH LIST

The COVID-19 death data gotten from the Moodle uses district codes and names based on the 2018 version, whereas the Census 2021 data which will be used for this analysis reflects the 2023 version. This discrepancy arises from changes of English local authority districts over the years. To make the data set compactable with the downloaded data from nomis, the LA names and LA_code needs to be updated.

A table showing the changes in English local authority districts from 2018 to 2023, associated with their descriptions is provided in the Moodle. This will be used to update the Covid-19 data set by updating the names of the local authority districts and merging the district were necessary.

STEPS

1. The COVID-19 death data table is downloaded from the module and opened in an Excel

   work sheet. It includes the LA_name, LA_code, recorded deaths in the different districts for every month from March 2020 to April 2021, and the sum Total in different columns. Having 326 Rows excluding the header row.

2. For this analysis just the LA_name, LA_code, and Total will be needed. The rest are deleted

   and the fill is saved name; **COVID_19_deaths** as a Comma-Separated Value (CSV) file.

3. A new SQL data base is created in SQLite and the **COVID_19_deaths** table is

   imported to the data base.

4. With the list of the updated/merged Local authorities districts provided in the Moodle, the COVID_19_deaths data set is processed.

   To change the old LA_name and LA_code in the COVID_19_deaths table to the updated version, the LA_name and LA_code to be changed is selected and replaced using the queries in Appendix

Year	Changes From	Changes To
2023	E07000026 – Allerdale E07000028 – Carlisle E07000029 – Copeland	E06000063 – Cumberland
2023	E07000027 – Barrow-in-Furness E07000030 – Eden E07000031 – South Lakeland	E06000064 – Westmorland and Furness
2023	E07000163 – Craven	E06000065 – North Yorkshire

	E07000164 – Hambleton E07000165 – Harrogate E07000166 – Richmondshire E07000167 – Ryedale E07000168 – Scarborough E07000169 – Selby	
2023	E07000187 – Mendip E07000188 – Sedgemoor E07000189 – South Somerset E07000246 – Somerset West and Taunton	E06000066 – Somerset
2021	E07000150 – Corby E07000152 – East Northamptonshire E07000153 – Kettering E07000156 – Wellingborough	E06000061 – North Northamptonshire
2021	E07000151 – Daventry E07000154 – Northampton E07000155 – South Northamptonshire	E06000062 – West Northamptonshire
2020	E07000004 – Aylesbury Vale E07000005 – Chiltern E07000006 – South Bucks E07000007 – Wycombe	E06000060 – Buckinghamshire
2019	E06000028 – Bournemouth E07000048 – Christchurch E06000029 – Poole	E06000058 – Bournemouth, Christchurch and Poole
2019	E07000049 – East Dorset E07000050 – North Dorset E07000053 – Weymouth and Portland E07000051 – Purbeck E07000052 – West Dorset	E06000059 – Dorset
2019	E07000205 – Suffolk Coastal E07000206 – Waveney	E07000244 – East Suffolk
2019	E07000201 – Forest Heath E07000204 – St Edmundsbury	E07000245 – West Suffolk
2019	E07000190 – Taunton Deane E07000191 – West Somerset	E07000246 – Somerset West and Taunton
2018	E07000112 – Shepway	E07000112 – Folkestone and Hythe

The changes are being made starting from 2018 to 2023. And some LA_code and LA names were corrected using data gotten from GOV.UK

5. Changing the LA_name and LA_code creates duplicate rows which are summed up together in SQL.

After the whole merging process the total rows of the COVID_19_deaths are reduced to 296.

6. The table is saved as ‘**Covid_death**’ in CSV format.

❖ MERGING THE DEPENDENT AND INDEPENDENT VARIABLES TOGETHER

1. A new SQL data base is created in SQLite
2. The CSV files; Covid_death, Age_band, Accommodation_type, Gender, Deprivation, Qualification_level, are imported into the SQL database.
3. The tables are to be joined together using the LA_code column as the related column between them. To do this the Full Join syntax is used to return all the matching records from all the tables selected.

❖ NORMALIZATION

The merged data set now includes all the 24 independent variable and 1 dependent variable the (updated Covid death). These variables are gotten from different sources and may have different scales. To adjust the values measured on different scales to a common scale, it is normalized to per 1000 using the total population gotten from the total age band that was downloaded from Nomis. This covers the population from age 0 – infinity. (variable * 1000 / total population).

The final dataset now consists of the 25 normalized variables which is saved as a CSV file

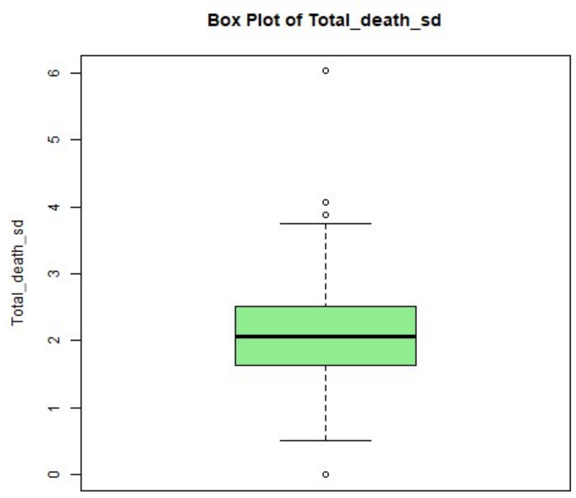
DATA EXPLORATION AND ANALYSIS

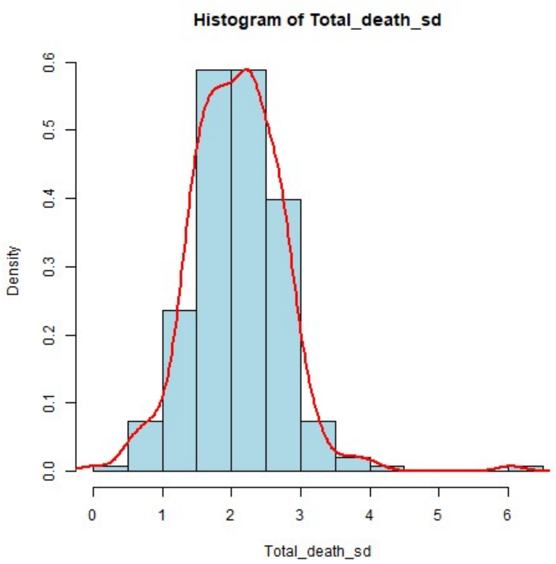
The Final dataset is imported to Rstudio. The first 6 rows structure was inspected.

❖ Distribution Check:

To check for the distribution of all the different variables, a box blot, quantile-quantile (Q-Q plot) and Histogram with density curve was made for each.

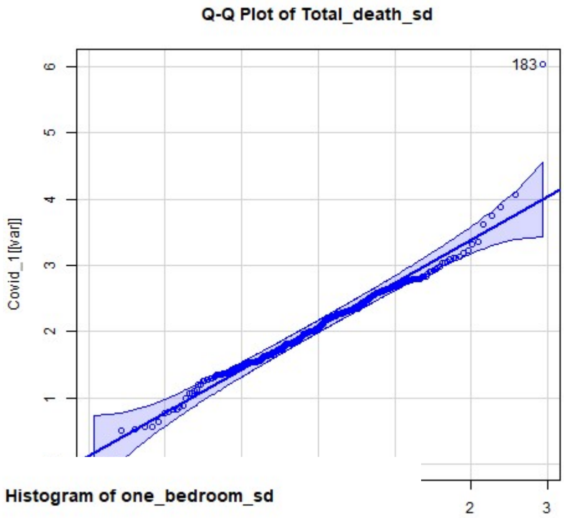
Total Covid-19 deaths; the graphs below show the distribution of the Covid-19 population





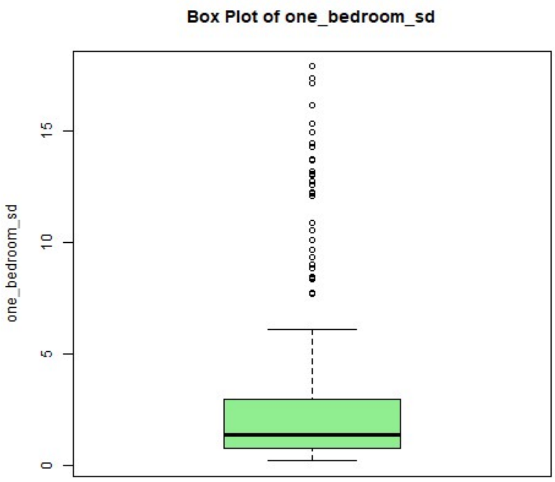
Interpretation;

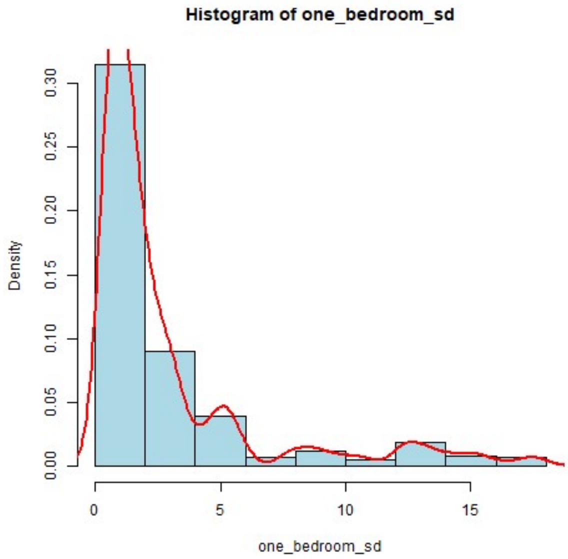
The Total Covid death variable generally follows a Normal Distribution, as evidenced by the Q-Q plot showing that most points lie along the diagonal line and Histogram with the shape closely resembling a bell curve.



However, the box plot highlights some outliers which is also seen on the Q-Q plot as some deviations at the tails (extreme values). This represents specific areas or populations experiencing significantly higher or lower or no deaths during COVID-19.

Accommodation type by number of rooms: the graph below shows the distribution of One-bedroom flat population.

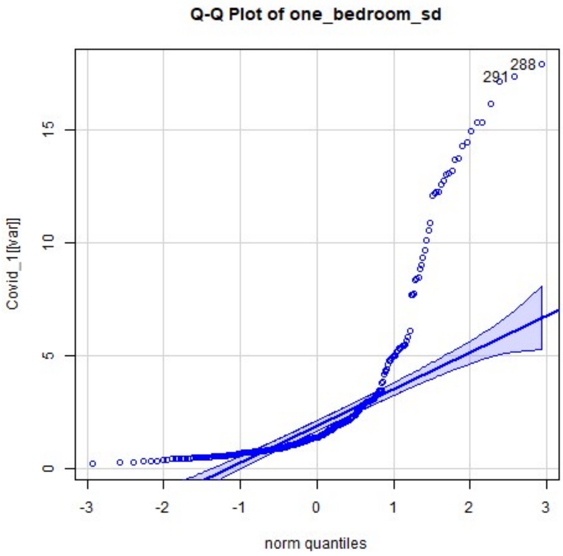




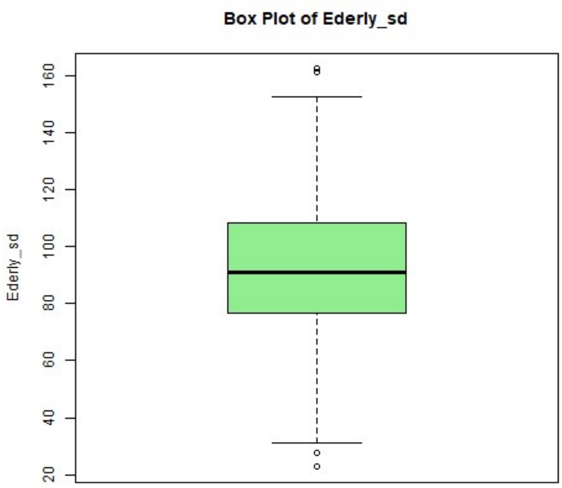
Interpretation:

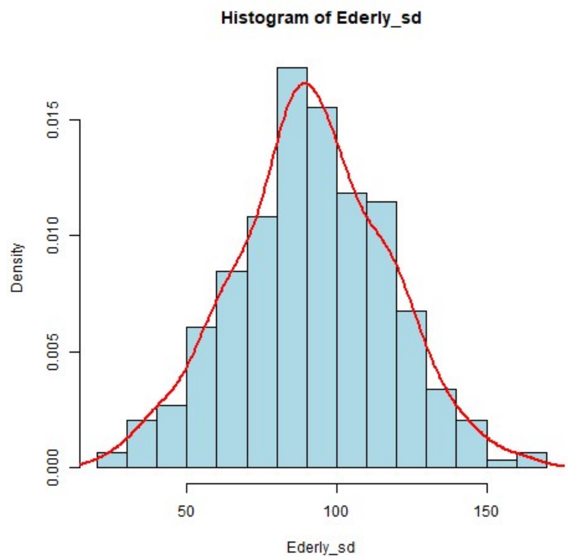
The data is not normally distributed, as evidenced by the Q-Q plot and histogram. There is significant skewness and the presence of outliers, with the majority of values clustered around lower ranges (below 5) as seen in the box plot.

These characteristics indicate a small subset of areas in London with significantly higher populations in one-bedroom flats, skewing the distribution.



Age Band: The graph below shows the distribution of population aged 75 and above classified as Elderly

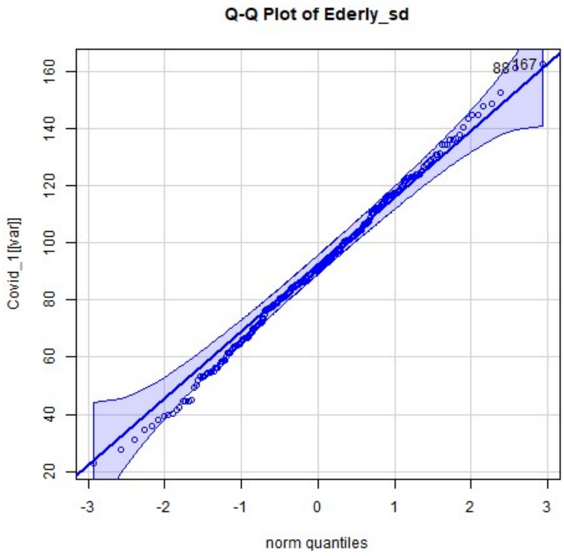




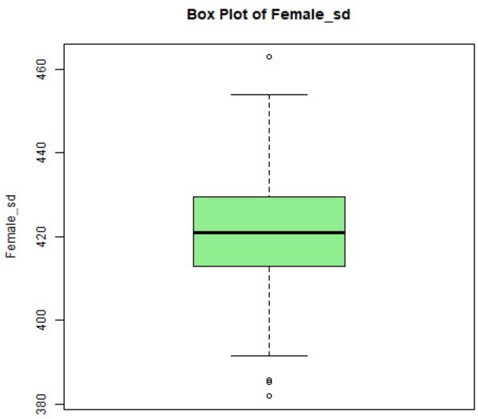
Interpretation:

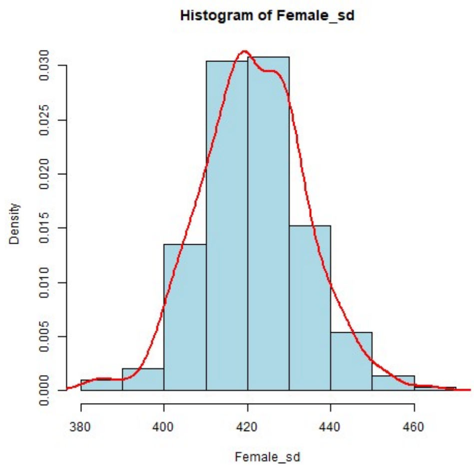
The Elderly age band variable is Normally Distributed, as evidenced by the Q-Q plot showing that most points lie along the diagonal line and Histogram with the shape closely resembling a bell curve.

However, the box plot highlights some outliers which is also seen on the Q-Q plot. This represents specific areas or populations with relatively high or low number of old people.



Gender: The graph below shows the distribution of Female population



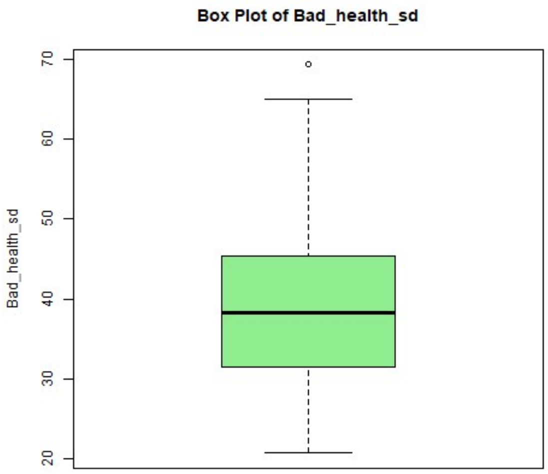


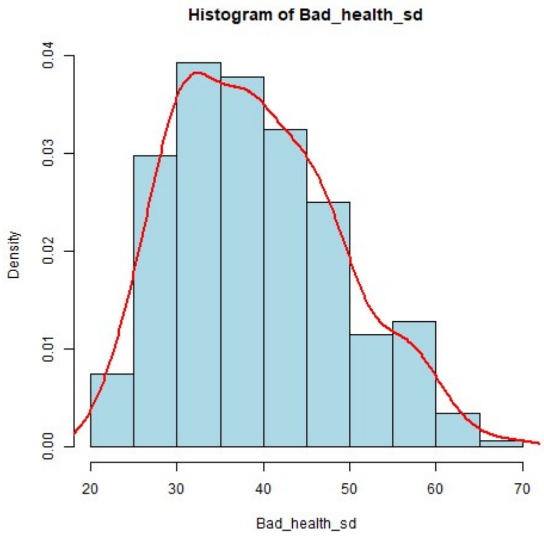
Interpretation:

The female population variable is Normally Distributed, as evidenced by the Histogram with the shape closely resembling a bell curve.

However, the box plot highlights some outliers. This represents specific areas or populations with relatively high or low number female population.

General Health: The graph below shows the distribution people with very bad health condition





Interpretation:

The population is Normally Distributed, as evidenced by the Histogram with the shape closely resembling a bell curve. However, the box plot highlights an outlier. This represents a specific area with relatively high population of people with very bad health condition.

- ❖ TESTS

Kolmogorov-Smirnov Test and Shapiro-Wilk Test was carried out to further confirm the variables that are normally distributed.

The Null Hypothesis (Ho) = The population data sample follows a normal distribution.

Below are the test results

	Variable	KS_p_value	KS_Normal	SW_p_value	SW_Normal
1	Total_death_sd	5.541283e-01	Yes	3.776642e-07	No
2	Adults_sd	1.878038e-01	Yes	2.221774e-03	No
3	Children_sd	8.636270e-01	Yes	2.742374e-05	No
4	Ederly_sd	8.673163e-01	Yes	9.306823e-01	Yes
5	Pre_ederly_sd	5.189564e-01	Yes	3.760732e-03	No
6	Youth_sd	1.216428e-05	No	1.765374e-15	No
7	four_plus_bedrooms_sd	1.383827e-01	Yes	1.768104e-03	No
8	one_bedroom_sd	1.926215e-16	No	3.798782e-24	No
9	three_bedrooms_sd	5.043828e-01	Yes	1.333147e-02	No
10	two_bedrooms_sd	3.142490e-05	No	1.201483e-13	No
11	Female_sd	9.823169e-01	Yes	7.526679e-01	Yes

12	Male_sd	2.963140e-02	No	8.713391e-17	No
13	Bad_health_sd	9.220106e-02	Yes	6.548486e-05	No
14	Good_health_sd	2.041992e-01	Yes	1.635762e-04	No
15	Verybad_health_sd	3.091910e-01	Yes	8.719685e-06	No
16	Verygood_health_sd	3.044056e-01	Yes	3.736841e-02	No
17	Four_dimension_deprivation_sd	5.964123e-03	No	8.436821e-13	No
18	No_dimension_deprivation_sd	2.404292e-01	Yes	3.853253e-06	No
19	One_dimension_deprivation_sd	9.300951e-01	Yes	7.008525e-01	Yes
20	Three_dimension_deprivation_sd	2.163702e-02	No	2.930783e-07	No
21	Two_dimension_deprivation_sd	8.731720e-01	Yes	6.428630e-02	Yes
22	High_level_qualifications_sd	1.172386e-01	Yes	1.537912e-08	No
23	Low_level_qualifications_sd	2.277651e-04	No	3.025575e-11	No
24	No_qualifications_sd	8.007925e-01	Yes	1.439287e-01	Yes
25	Other_qualifications_sd	2.656498e-01	Yes	6.652164e-04	No

Judging with the P-value, if the P-value is less than 0.05 (KS_Normal = No) we reject the Null Hypothesis. But if the P-value is grater then 0.005 (KS_Normal = Yes), we fail to reject the Null Hypothesis.

For Every Variable that the KS_Normal = Yes, we accept the Null Hypothesis.

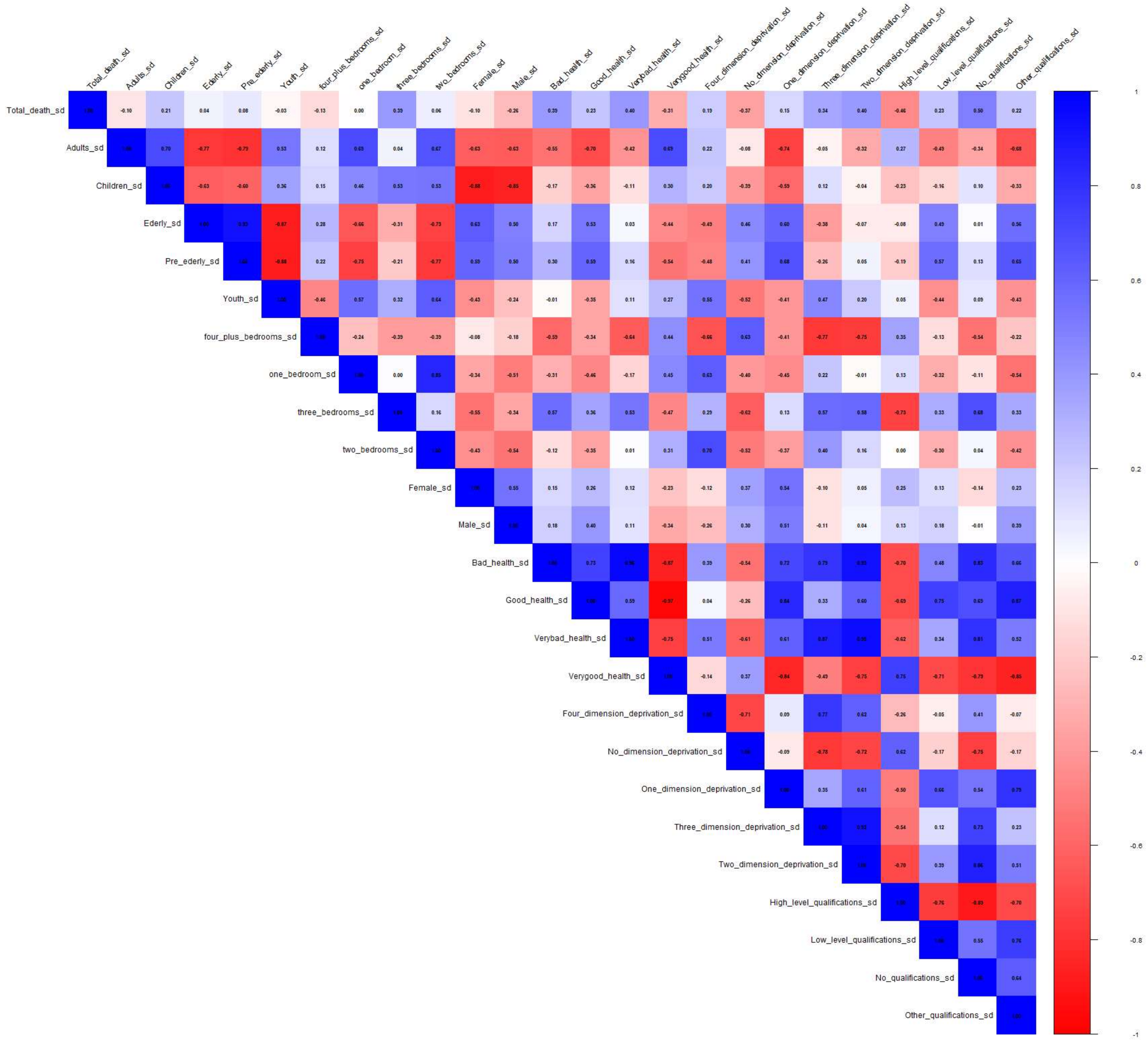
For Every Variable that the KS_Normal = No, we fail to accept the Null Hypothesis.

Spearman's Correlation; The Spearman correlation test was done to examine the correlation between the Covid deaths and the independent variables.

	Variable	Correlation
1	Adults_sd	-0.096862348
2	Children_sd	0.214688340
3	Ederly_sd	0.041971247
4	Pre_ederly_sd	0.081247658
5	Youth_sd	-0.028712688
6	four_plus_bedrooms_sd	-0.133950592
7	one_bedroom_sd	-0.003966833
8	three_bedrooms_sd	0.390090553
9	two_bedrooms_sd	0.061743408
10	Female_sd	-0.101445976
11	Male_sd	-0.264193007
12	Bad_health_sd	0.393022761
13	Good_health_sd	0.228799678
14	Verybad_health_sd	0.400893035
15	Verygood_health_sd	-0.314923398
16	Four_dimension_deprivation_sd	0.192810839
17	No_dimension_deprivation_sd	-0.367611525
18	One_dimension_deprivation_sd	0.153023131

```
19 Three_dimension_deprivation_sd  0.343995613
20   Two_dimension_deprivation_sd  0.399485001
21   High_level_qualifications_sd -0.458712179
22    Low_level_qualifications_sd  0.228240721
23           No_qualifications_sd  0.497270460
24        Other_qualifications_sd  0.222817733
```

Correlation Matrix plot with all the Variables



To further determine the independent variables with a real correlation positive or negative, we then filter variables based on correlation greater than 0.2

```
                         Variable  Correlation
2                     Children_sd    0.2146883
8               three_bedrooms_sd    0.3900906
11                        Male_sd   -0.2641930
12                  Bad_health_sd    0.3930228
13                 Good_health_sd    0.2287997
14              Verybad_health_sd    0.4008930
15             Verygood_health_sd   -0.3149234
17     No_dimension_deprivation_sd   -0.3676115
19  Three_dimension_deprivation_sd    0.3439956
20    Two_dimension_deprivation_sd    0.3994850
21    High_level_qualifications_sd   -0.4587122
22     Low_level_qualifications_sd    0.2282407
23           No_qualifications_sd    0.4972705
24        Other_qualifications_sd    0.2228177
```

It is seen that there are only 14 independent variables with a real correlation both positive and negative.

"Very bad health", "no qualification", "two-dimension deprivation" have higher positive correlation.

"High level qualification", "very good health", "no dimension of deprivation" have a higher negative correlation.

- Kaiser–Meyer–Olkin (KMO) test

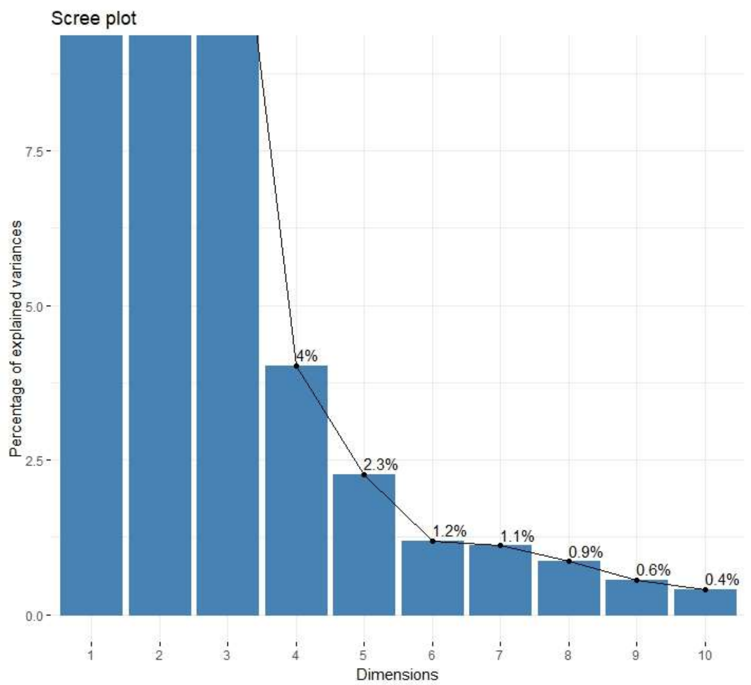
A KMO test was carried out to determine how suitable the data set is for factor analysis.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = pca_data)
Overall MSA =  0.47
MSA for each item =
                   Children_sd             three_bedrooms_sd
                          0.20                          0.88
               two_bedrooms_sd                       Male_sd
                          0.68                          0.63
                 Bad_health_sd                Good_health_sd
                          0.44                          0.40
             Verybad_health_sd            Verygood_health_sd
                          0.41                          0.44
   No_dimension_deprivation_sd Three_dimension_deprivation_sd
                          0.92                          0.88
  Two_dimension_deprivation_sd   High_level_qualifications_sd
                          0.89                          0.43
   Low_level_qualifications_sd          No_qualifications_sd
                          0.33                          0.45
       Other_qualifications_sd
                          0.38
```

The overall MSA **is 0.47**, which is below the recommended threshold of 0.6 for Factor analysis but it is close to 0.6 and it means that there are some inter-correlations present between the independent variable. So, I still carried out the Principal Component Analysis (PCA) regardless the low overall MSA.

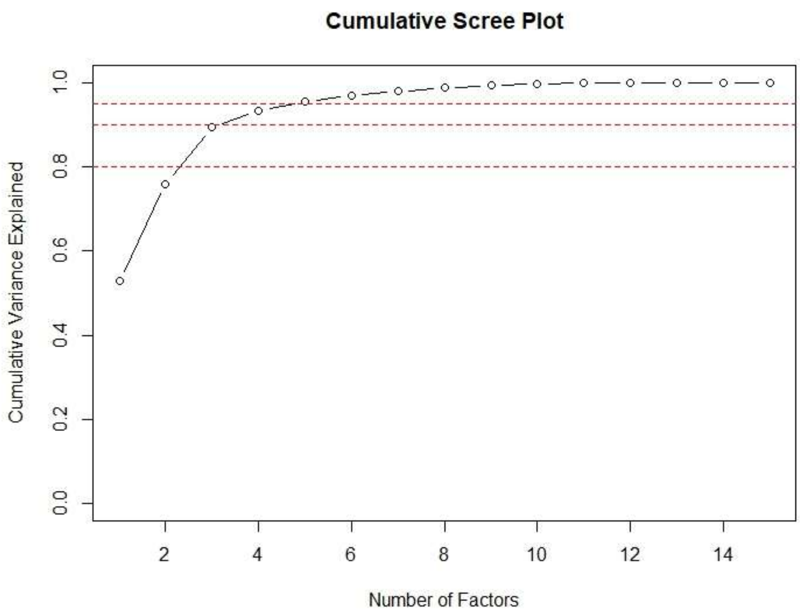
Eigenvalues

```
[1] 7.918955e+00  3.473567e+00 1.997746e+00 6.024744e-01
 [5] 3.401528e-01 1.787689e-01 1.670342e-01 1.287443e-01
 [9] 8.362230e-02 5.992274e-02 2.728950e-02 1.310454e-02
[13] 8.617022e-03 4.683329e-07 1.413981e-08
```



The "elbow" in the plot occurs around **4 dimensions**. This is the point where the explained variance begins to level off. This means the data set can effectively be reduced to 4 principal components as these capture most of the meaningful variance in the data set in order to simplify the data while retaining it core patterns and relationship.

Cumulative Proportion



There are two factors below the 80% line, three factors below 90% and four factors below 95%. The cumulative plot indicates that 4 components would explain approximately 95% of the variance in the dependent variables.

- PRINCIPAL COMPONENT ANALYSIS.

```
Principal Components Analysis
Call: principal(r = pca_data, nfactors = 4, rotate = "varimax", scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
                                  RC1   RC3   RC2   RC4   h2    u2 com
Children_sd                     -0.04 -0.05  0.95  0.22 0.96 0.039 1.1
three_bedrooms_sd                0.49  0.40  0.67 -0.08 0.85 0.148 2.5
two_bedrooms_sd                  0.12 -0.39  0.31  0.83 0.95 0.053 1.8
Male_sd                          0.01  0.10 -0.89 -0.16 0.83 0.166 1.1
Bad_health_sd                    0.89  0.42 -0.03 -0.10 0.98 0.021 1.5
Good_health_sd                   0.30  0.91 -0.17 -0.04 0.94 0.055 1.3
Verybad_health_sd                0.95  0.26  0.01 -0.03 0.97 0.032 1.2
Verygood_health_sd              -0.51 -0.83  0.14  0.05 0.96 0.036 1.7
No_dimension_deprivation_sd     -0.54 -0.21 -0.46 -0.59 0.90 0.104 3.2
Three_dimension_deprivation_sd   0.94 -0.02  0.14  0.25 0.97 0.027 1.2
Two_dimension_deprivation_sd     0.91  0.34  0.02  0.18 0.98 0.023 1.4
High_level_qualifications_sd    -0.36 -0.82 -0.42 -0.06 0.98 0.021 1.9
Low_level_qualifications_sd      0.04  0.92  0.13 -0.18 0.90 0.097 1.1
No_qualifications_sd             0.69  0.60  0.19  0.20 0.91 0.085 2.3
Other_qualifications_sd          0.23  0.91 -0.09 -0.12 0.90 0.098 1.2
```

```
                         RC1  RC3  RC2  RC4
SS loadings             4.95 4.95 2.78 1.31
Proportion Var          0.33 0.33 0.19 0.09
Cumulative Var          0.33 0.66 0.85 0.93
Proportion Explained    0.35 0.35 0.20 0.09
Cumulative Proportion   0.35 0.71 0.91 1.00
```

```
Mean item complexity =  1.6
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.02
 with the empirical chi square  31.1  with prob <  0.99

Fit based upon off diagonal values = 1
```

❖ MULTIPLE LINEAR REGRESSION MODELLING

A multiple linear regression model was applied to include all independent variables, regardless of their correlation with the dependent variable. This is to find the best -fitting linear equation that explains the relationship between the dependent variable and independent variables.

After ensuring there are no missing values, the dependent variable is combined with the independent variables and fit the multiple linear regression model

```
Call:
lm(formula = dependent_var ~ ., data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6475 -0.2897 -0.0155  0.2643  3.9237

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.299e+03  1.062e+03  -2.165  0.03123 *
Children_sd                     7.834e-04  2.764e+00   0.000  0.99977
three_bedrooms_sd              -1.929e-04  4.520e-03  -0.043  0.96598
two_bedrooms_sd                 1.077e-02  6.361e-03   1.693  0.09165 .
Male_sd                        -6.923e-03  6.187e-03  -1.119  0.26416
Bad_health_sd                   2.338e+00  2.721e+00   0.859  0.39098
Good_health_sd                  2.295e+00  2.720e+00   0.844  0.39962
Verybad_health_sd               2.333e+00  2.722e+00   0.857  0.39209
Verygood_health_sd              2.301e+00  2.721e+00   0.846  0.39843
No_dimension_deprivation_sd     7.750e-03  3.151e-03   2.460  0.01451 *
Three_dimension_deprivation_sd -9.050e-02  3.437e-02  -2.633  0.00893 **
Two_dimension_deprivation_sd    3.293e-03  2.120e-02   0.155  0.87666
High_level_qualifications_sd    5.896e-04  2.765e+00   0.000  0.99983
Low_level_qualifications_sd     4.283e-03  2.765e+00   0.002  0.99877
No_qualifications_sd            1.956e-02  2.765e+00   0.007  0.99436
```

```
Other_qualifications_sd        -2.139e-02  2.765e+00  -0.008  0.99383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.556 on 280 degrees of freedom
Multiple R-squared:  0.3189,     Adjusted R-squared:  0.2824
F-statistic: 8.741 on 15 and 280 DF,  p-value: < 2.2e-16
```

Out of the initial 25 variables considered for the model, only three variable exhibited significant effects. They are; ‘two bed room’, ‘No dimension deprivation’ and ‘three-dimension deprivation’. The adjusted R-squared value for the model indicates it goodness of fit is 0.2824.

Multicollinearity Using VIF (Variance Inflation Factor).

```
                   Children_sd                three_bedrooms_sd                two_bedrooms_sd
                          TRUE                             TRUE                           TRUE
                       Male_sd                    Bad_health_sd                 Good_health_sd
                          TRUE                             TRUE                           TRUE
             Verybad_health_sd               Verygood_health_sd   No_dimension_deprivation_sd
                          TRUE                             TRUE                           TRUE
Three_dimension_deprivation_sd   Two_dimension_deprivation_sd  High_level_qualifications_sd
                          TRUE                             TRUE                           TRUE
   Low_level_qualifications_sd            No_qualifications_sd      Other_qualifications_sd
                          TRUE                             TRUE                           TRUE
```

The second model was constructed by utilizing only those variables that remained after the exclusion of variables exhibiting correlation coefficients during the analysis.

```
two_bedrooms_sd    No_dimension_deprivation_sd Three_dimension_deprivation_sd

          FALSE                          FALSE                          FALSE
```

the output of the “The variance inflation factor” (VIF) which detects multicollinearity between the variables to see if each independent variable in the linear regression model is greater than 2. These results been FALSE for all the variables suggests that multicollinearity is not a significant concern.

A multiple linear regression model was applied to include only those independent variables that exhibited a correlation coefficient of 0.4 or higher with the dependent variable.

```
Call:
lm(formula = dependent_var ~ ., data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8146 -0.3381 -0.0359  0.3667  3.9250 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     3.683096   0.569950   6.462 4.32e-10 ***
two_bedrooms_sd                -0.008419   0.002837  -2.968 0.003249 ** 
No_dimension_deprivation_sd    -0.007660   0.002086  -3.672 0.000286 ***
Three_dimension_deprivation_sd  0.020383   0.009176   2.221 0.027100 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6122 on 292 degrees of freedom
Multiple R-squared:  0.139,	Adjusted R-squared:  0.1301 
F-statistic: 15.71 on 3 and 292 DF,  p-value: 1.69e-09
```

Key Takeaways:

The model suggests a statistically significant relationship between COVID-19 deaths and “three-dimension deprivation”, “two bed room”, “No dimension deprivation”. They all have a statistically significant impact on covid-19 mortality rate.

The F-statistic and it associated low P-value of 1.69e-09 indicates that the overall model is statistically significant.

Conclusion and Discussion

This study aimed to identify the key risk factors contributing to COVID-19 deaths. The dependent variable was the total number of COVID-19-related deaths, while independent variables, representing potential risk factors, were drawn from social and economic data obtained from the 2011 census provided by the Office for National Statistics.

The data tables were imported into SQLite and merged for analysis. R was chosen as the primary tool for data analysis. The initial phase involved checking for and visualizing any missing data. Factor analysis was then applied for dimensionality reduction, simplifying the set of independent variables into a smaller number of components.

Multiple regression analysis was performed using various models to explore the relationship between external variables (risk factors) and the rate of COVID-19 deaths. The results indicate that the effect of these external variables on COVID-19 death rates is multifaceted.

In conclusion, while other risk factors should not be disregarded, the study highlights that variable such as "two-bedroom housing," "No dimension deprivation," and "three-dimension deprivation" stand out as the most significant contributors to COVID-19 deaths in England.

Two- Bed Room Housing:

The relationship between two-bedroom homes and COVID-19 death rates is shaped by factors like housing density, socioeconomic status, and location. Overcrowded or multi-generational households in two-bedroom homes, common in urban areas, increase transmission risks and mortality among vulnerable members. Wealthier neighborhoods with two-bedroom homes benefit from better healthcare access and remote work opportunities, reducing exposure and death rates. In contrast, lower-income households often face higher transmission due to shared living spaces. Additionally, rural areas with lower population density and larger properties experience reduced risks, while urban two-bedroom homes in densely populated areas are linked to higher transmission and death rates.

No dimension deprivation:

The absence of deprivation, characterized by access to good housing, quality healthcare, economic stability, and education, was strongly linked to lower COVID-19 death rates. Improved living conditions facilitated social distancing and reduced transmission, while well-funded healthcare ensured timely interventions. Non-deprived populations benefited from better health, nutrition, and lower rates of chronic illnesses, enhancing resilience to the virus. Economic stability and remote work opportunities reduced exposure risks, while better vaccine access and reliable information promoted adherence to preventive measures. Additionally, lower stress levels in financially secure households supported better mental and physical health. Together, these factors significantly reduced mortality rates in non-deprived areas.

Three-Dimension Deprivation:

High levels of deprivation were positively correlated with increased COVID-19 deaths, driven by poor living conditions, lack of healthcare access, and economic instability. Overcrowded homes made social distancing and isolation difficult, leading to higher transmission rates, especially among vulnerable groups like the elderly. Poor housing quality, with inadequate ventilation and sanitation, further heightened the risk of virus spread. Limited healthcare access due to financial or logistical barriers delayed treatment, worsening outcomes and increasing preventable deaths, particularly in rural and low-income areas with under-resourced healthcare systems.

References

Averyanova, M., Vishnyakova, P., Yureneva, S., Yakushevskaya, O., Fatkhudinov, T., Elchaninov, A. and Sukhikh, G. (2022) ‘Sex hormones and immune system: Menopausal hormone therapy in the context of COVID-19 pandemic’, Frontiers in Immunology, 13. Available at: https://doi.org/10.3389/fimmu.2022.928171 (Accessed: 6 January, 2025).

Borges do Nascimento, I.J., O'Mathúna, D.P., von Groote, T.C., et al. (2021) 'Coronavirus disease (COVID-19) pandemic: an overview of systematic reviews', BMC Infectious Diseases, 21(525). Available at: https://doi.org/10.1186/s12879-021-06214-4 (Accessed: 6 January, 2025).

Choi, W.-Y. (2022) 'Mortality Rate of Patients With COVID-19 Based on Underlying Health Conditions', Disaster Medicine and Public Health Preparedness, 16(6), pp. 2480–2485. doi:10.1017/dmp.2021.139.

Flynn, D., Moloney, E., Bhattarai, N., Scott, J., Breckons, M., Avery, L. and Moy, N. (2020) 'COVID-19 pandemic in the United Kingdom', Health Policy and Technology, 9(4), pp. 673–691. Available at: https://doi.org/10.1016/j.hlpt.2020.08.003 (Accessed: 31 December, 2025).

Khanijahani, A., Iezadi, S., Gholipour, K., et al. (2021) 'A systematic review of racial/ethnic and socioeconomic disparities in COVID-19', International Journal for Equity in Health, 20(248). Available at: https://doi.org/10.1186/s12939-021-01582-4 (Accessed: 6 January, 2015).

Nguyen, N.T., Chinn, J., De Ferrante, M., Kirby, K.A., Hohmann, S.F., et al. (2021) 'Male gender is a predictor of higher mortality in hospitalized adults with COVID-19', PLOS ONE, 16(7), e0254066. Available at: https://doi.org/10.1371/journal.pone.0254066 (Accessed: 6 January, 2025).

Office for National Statistics (2020) Coronavirus (COVID-19) related deaths by occupation, England and Wales. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/datasets/coronaviruscovid19relateddeathsbyoccupationenglandandwales (Accessed: 31 December, 2024).

Peckham, H., de Gruijter, N.M., Raine, C., et al. (2020) 'Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission', Nature Communications, 11(6317). Available at: https://doi.org/10.1038/s41467-020-19741-6 (Accessed: 6 January, 2025)

Public Health England (2019) Disparities in the risk and outcomes of COVID-19. London: Public Health England. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/892085/disparities_review.pdf (Accessed: 3 January, 2025).

Raoult, D., Zumla, A., Locatelli, F., Ippolito, G., and Kroemer, G. (2020) 'Coronavirus infections: Epidemiological, clinical and immunological features and hypotheses', Cell Stress, 4(4), pp. 66–75. Available at: https://doi.org/10.15698/cst2020.04.216 (Accessed: 6 January, 2025).

Sasson, I. (2021) 'Age and COVID-19 mortality: A comparison of Gompertz doubling time across countries and causes of death', Demographic Research, 44(16), pp. 379–396. Available at: https://doi.org/10.4054/DemRes.2021.44.16 (Accessed: 6 January, 2025).

Sousa, G. Garces, T., Cestari, V., Florêncio, R., Moreira, T. and Pereira, T. (2020) 'Mortality and survival of COVID-19', Epidemiology and Infection, 148, p. e123. doi:10.1017/S0950268820001405.

Tang, I.W., Vieira, V.M. and Shearer, E. (2022) 'Effect of socioeconomic factors during the early COVID-19 pandemic: a spatial analysis', BMC Public Health, 22(1212). Available at: https://doi.org/10.1186/s12889-022-13618-7 (Accessed: 6 January, 2025).

UK Government (2022) COVID-19 confirmed deaths in England (to 31 December 2022): report. Available at https://www.gov.uk/government/publications/covid-19-reported-sars-cov-2-deaths-in-england/covid-19-confirmed-deaths-in-england-to-31-december-2022-report (Accessed: 6 January, 2025).

World Health Organization (2020) COVID-19: Vulnerable and high risk groups. Available at: https://www.who.int/westernpacific/emergencies/covid-19/information/high-risk-groups (Accessed: 31 December, 2024).

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. (2020) 'Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study', The Lancet, 395(10229), pp. 1054–1062. Available at: https://doi.org/10.1016/S0140-6736(20)30566-3 (Accessed: 5 January, 2025).

Zhuo, J. and Harrigan, N. (2023) ‘Low education predicts large increase in COVID-19 mortality: The role of collective culture and individual literacy’, Public Health, 221, pp. 201–207. Available at: https://doi.org/10.1016/j.puhe.2023.06.016 (Accessed: 5 January, 2025).