Text

Description automatically generated with medium confidence

**DS7006 - (M-LEVEL MODULE)**

**QUANTITATIVE DATA ANALYSIS- FINAL**

**NAME:** Kalai Selvi Gandhi **STUDENT ID:** U2216077

**MODULE NAME:** Quantitative Data Analysis

**MODULE CODE:** DS7006

**COURSE NAME:** MSC, Data Science (January 2022)

**TOPIC:** Analysis of COVID Death Rate in England

**MODULE LEADER:** Dr. Yang Li.

**ANALYSIS OF COVID DEATH RATE IN ENGLAND.**

**ABSTRACT:**

This coursework provides preliminary assessments COVID-19-related deaths in England's local authorities using population density, household, and health. COVID-19-related deaths that occurred between 2020 and 2022 are included. UK-wide By August 12th, the amount of COVID-19 reported cases and fatalities were 313,798 and 46,706, accordingly (70 deaths and 472 case per 100,000 population). Regional differences existed in England in comparison with other countries, with London, the North, and the West being disproportionately impacted (666 and 756 case per 100,000 population, accordingly). 13,618,470 analyzes had been completed in the UK as of August 11th. Male gender, mortality rate (65 years old), and selected groups were all linked to a higher chance of passing away. Since, the start of the epidemic there has been a decrease in the use of initial care, cancer referrals, outpatient/inpatient referrals. Government initiatives to lessen the impact on industry and the economy have been put in place, but the number of people applying for state aid has climbed. The management of a hospital is improved by the analysis of many patients using machine learning and data mining algorithms. Big data technologies and solutions are utilized since it is challenging to analyze, retrieve, manage, and store both structured and unstructured information due to the enormous volume of data that is increasing in all fields. Initially, SARS-CoV-2, the highly contagious source of Covid-19, was identified and screened using a combination of ultrasound scans, entire genotyping, and electron microscopy. There are very few Covid-19 test kits in health care facilities because of the growing number of patients diagnosed each day. As an outcome, a conscience structure must be employed as a quick substitute analysis to prevent Covid-19 from spreading throughout the world.

A prolonged closure of 42–56 days is preferred to 21–28 days of closure in order to significantly "flatten the curve." Our models only forecast the number of COVID-19 infections, providing policymakers with information about one part of this complex problem. In order to enable data-driven policymaking during a pandemic, finish by discussing the crucial roles of greater testing, trustworthy and transparent data, appropriate uncertainties assessment, accurate analysis of prediction model, reproducible data science methodologies, and tools. From a wide range of rich data sources, massive quantities of big data can be created and gathered. These huge data contain significant knowledge and practical information. For instance, healthcare and epidemiological information about people with viral disorders such the coronavirus disease in 2019 (COVID-19). Data scientists' knowledge gained from these epidemiologic studies helps researchers, epidemiologists, and policy makers better understand the disease, which may motivate them to devise strategies for detecting, containing, and battling it. With an emphasis on the geographic information analytics across various geographic locations, we provide a geographic information science system in this study for analyzing large COVID-19 epidemiological data.

***KEYWORDS:***

*Household, Population density, Health condition, Covid-19, local authorities in England, Area in hectares and Death Rates.*

**Table of Contents:**

[INTRODUCTION 4](#_Toc123575332)

[METHODOLOGIES USED 5](#_Toc123575333)

[DATA SET USED 5](#_Toc123575334)

[DATA SET DESCRIPTION 5](#_Toc123575335)

[TECHNIQUES USED 5](#_Toc123575336)

[DATA FRAME 5](#_Toc123575337)

[DATA EXPLORATION 6](#_Toc123575338)

[MISSING MAP 6](#_Toc123575339)

[BOXPLOT 7](#_Toc123575340)

[VISUALIZATION 7](#_Toc123575342)

[BOXPLOT OF OUTLIERS VARIABLES 8](#_Toc123575343)

[HISTOGRAM 8](#_Toc123575344)

[NORMALIZATION 8](#_Toc123575345)

[MIN-MAX SCALING 9](#_Toc123575346)

[Z-SCORE SCALING 9](#_Toc123575347)

[SOFTMAX SCALING 10](#_Toc123575348)

[KOLMOGOROV-SMIRNOV TEST OF NORMALITY FOR EACH VARIABLES 11](#_Toc123575349)

[CORRELATION 13](#_Toc123575354)

[SPEARMAN’S CORRELATION 13](#_Toc123575355)

[PEARSON’S CORRELATION 13](#_Toc123575356)

[INTERNAL CORRELATION USING PEARSON’S METHOD 16](#_Toc123575361)

[PARTIAL CORRELATION 18](#_Toc123575365)

[DEATH ANALYSIS BY AREA 19](#_Toc123575366)

[LOW DEATH RATE AREAS 19](#_Toc123575367)

[MEDIUM DEATH RATE AREAS 20](#_Toc123575368)

[HIGH DEATH RATE AREA 20](#_Toc123575369)

[GG-PLOT FOR EACH VARIABLES 21](#_Toc123575370)

[KAISER-MEYER-OLKIN STATISTICAL TEST 24](#_Toc123575379)

[SCATTERPLOT 24](#_Toc123575380)

[NORMALITY TEST 25](#_Toc123575381)

[I.QQ-Plot 25](#_Toc123575382)

[II. Regression Lined QQ-Plot 25](#_Toc123575383)

[LINEAR REGRESSION 25](#_Toc123575384)

[HISTOGRAM FOR RESIDUALS 26](#_Toc123575385)

[MULTIPLE LINEAR REGRESSION 27](#_Toc123575386)

[HISTOGRAM FOR RESIDUALS 28](#_Toc123575387)

[POISSON REGRESSION 29](#_Toc123575388)

[HISTOGRAM FOR RESIDUALS 30](#_Toc123575389)

[ASYMPTOTIC ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST 31](#_Toc123575390)

[LINEAR REGRESSION 31](#_Toc123575391)

[MULTIPLE LINEAR REGRESSION 32](#_Toc123575392)

[POISSON REGRESSION 32](#_Toc123575393)

[DATA ANALYSIS 32](#_Toc123575394)

[RESULT AND DISCUSSION 34](#_Toc123575395)

[CONCLUSION 34](#_Toc123575396)

[Appendix 39](#_Toc123575397)

# **INTRODUCTION:**

SARS-CoV-2 has been to take responsibility for the initial COVID-19 reported cases in the Great Britain in Jan 2020, and individual transmission had been acknowledged all through the country by mid - February. Regarding an exponential rise in COVID-19 cases, numerous measures were quickly implemented, which include social rejection, self-isolation, and lockdown tactics, to slow the spread of Covid-19 as well as begin minimizing the burden on the health-care system. To protect the UK economy, financial interventions were quickly implemented.

This Coursework describes the epidemiologic studies statistical data on COVID-19 reported in the UK populations, as well as the government feedback to global epidemic, which includes all the impacts on the health sector, business, and economic activity. A socioeconomic profile of the United Kingdom is depicted in, which includes data on **size of the population**, **density of population**, living standards, and **household composition**. The key health indicators in the United Kingdom are described through health behaviors, socioeconomic factors, and mental health. The dataset used in this course work is obtained from the Nomis website which contains various themes and the exploratory part using data mining is discussed in this coursework. There are now several models that have been employed to examine the COVID-19 case-count data at the time this coursework was being written. The methods for simulating the spread of a disease and then predicting the number of cases in the future could be broadly splited into two groups: apportioned epidemiological prototypes and exponential/Poisson type models.

This article explores two main forecasting goals using the Bayesian modification of the Model (eSIR) (a) Projecting case counts in the future, both in the long term and the short term, with various suppression methods in place (post closure) (a) Analyzing how a lockdown's length or duration affects our projections of cumulative COVID-19 infection. To evaluate the reliability of our forecasting models, we perform thorough sensitivity analysis. We wrap up by talking about the necessity of accurate case-count data, additional testing, quantification of anticipated case-count uncertainty, and open data science methodologies that might impact and educate policy during a pandemic.

One of the discrete epidemiologic models included is the especially vulnerable (SIR) scheme, which would be influenced by a differential equation set linking the percentage of sensitive persons, the quantity of infected individuals (instances), and the number of individuals who have been removed. Because of the COVID-19 epidemic, numerous studies have focused on various aspects of the COVID-19 illness. For instance, research on conflict resolution for the COVID-19 epidemic has been done from a social science perspective. There have been publications concentrating on medical and treatment information, as well as clinical research and vaccine development, from the perspective of medicine and health science. Researchers have investigated ai technologies (AI)-driven informatics, sensing, and scanning for monitoring, testing, diagnosis, therapy, and prognosis. One example is the imaging-based diagnostic test of COVID-19 utilizing lung computed tomography (CT) scans. Additionally, researchers have developed mathematical models to predict the COVID-19 outbreak.

# **METHODOLOGIES USED:**

The data set used in the data exploration with using data mining techniques is obtained from the data census of 2011 from the official Nomis Website which consists of 323 entries and 21 total columns.

## **DATA SET USED:**



This data set is used in this course work to obtain the Covid death rate in England.

## **DATA SET DESCRIPTION:**

The above data set is used to analyze the Covid Death rate in the Local authorities of England from 2020 to 2022. There are several attributes used in this dataset are Area, Lives in a household, Population density, General Health, Death in 2020, Death in 2022 and Total Death rate. The main themes used in this course work are Household, Population density and Health. One of the discrete epidemiologic models included is the especially vulnerable (SIR) scheme, which would be influenced by a differential equation set linking the percentage of sensitive persons, the quantity of infected individuals (instances), and the number of individuals who have been removed. Because of the COVID-19 pandemic, numerous studies have focused on various aspects of the COVID-19 illness. Here, a dependent variable is Covid Death and the independent variables such as Rental Household, Own Household, Bad Health, Very Bad Health, White, Asian Ethnic groups, Resident\_Male and Resident\_Female are utilized in the Data Exploration.

## **TECHNIQUES USED:**

A statistical study of the data used to forecast the Covid Death Rates from the Local Authority Areas in England dataset on the Nomis website. The linear regression is used to predict the cumulative shortfall between the actual procedures executed and the projected procedures, based on previous pre-pandemic trends, in order to estimate the possible backlog. Following that, three additional predictions of the modification to the backlog are made: recovery to a specific capacity level, varying between 90% and 130%; and further activity disruption (such as a second pandemic wave). Additionally, the Poisson Regression and Multiple Regression are estimated and contrasted appropriately.

## **DATA FRAME:**

A data frame is the most common data framework for analysis of data in R, and it is also the most common method for storing information in R. A data set comprises a set of feature vector of equal length. The elements of the list can be thought of as rows, and the length of each element relates to the column count.

Graphical user interface, application

Description automatically generated

# **DATA EXPLORATION:**

* Load the data set into R studio and visualize the missing values. Thus, the missingness map can be obtained.
* Then, create a data frame containing the integer data type variables and name the columns and check the new variables created.
* Create another data frame which contains the themes alone. Hence, the boxplot is acquired.
* The Normalization can be done by using three techniques such as Minmax scaling, Z-score and SoftMax scaling, and the respective boxplots can be observed.
* Then, the KS-test of normality is done to check whether the variables is normally distributed or not.
* Finally, the three models of Regression such as Linear Regression, Multiple Linear Regression and Poisson Regression is evaluated and compared.

## **MISSING MAP:**

A missingness map depicts the destination of missing data in each variable. It shows who dropped out of the analysis earlier or whether numerous factors were absent at the same time. If this is the case, the analyst can conclude that there is a link between these variables. The missingness map for the selected data set is as follows:

A picture containing table

Description automatically generated

## **BOXPLOT:**

A box plot graphically displays statistical evidence predicated on the minimal level, first percentiles, average, third quartile, and maximum. The phrase "box plot" arose from the chart's presentation as a rectangular prism with line frequency as from top and bottom.

### **BOXPLOT OF ALL VARIABLES PROPORTION BY 1000:**

Chart, box and whisker chart

Description automatically generated

## **VISUALIZATION:**

An approach for graphically representing data is data visualization. It will make the data simpler to comprehend by incorporating components such as graphical representations, statistics, graphs, histograms, maps, and so on. Visualization is a technique that will make data easier to analyze

## **BOXPLOT OF OUTLIERS VARIABLES:**

Chart, box and whisker chart

Description automatically generated

## **HISTOGRAM:**

A histogram depicts the value systems of a variable's frequency components as bucketed ranges. Although bar graphs and a histogram are similar, a histogram organizes the attributes into continuous ranges. The percentage of value systems that fall within that range is represented by the altitude of every bar in a histogram. Histograms are created in R using the hist () function.

Diagram

Description automatically generated

## **NORMALIZATION:**

The process of organizing statistics in a database is known as normalization. This consists of creating charts and integrating them as according to criteria’s designated to protect the information and increase the flexibility of the database by eliminating recurrence and inconsistent reliance. This includes three methods such as Min-Max scaling, Z-score technique and SoftMax scaling.

## **MIN-MAX SCALING:**

Min-max normalization is a popular technique for normalizing data. The maximum and minimum value of each parameter are converted to 0 and 1, including both, whereas all other attributes are transformed to a fraction between 0 and 1.

Chart, box and whisker chart

Description automatically generated

## **Z-SCORE SCALING:**

The Z-Score number is used to evaluate how far a data set is from the mean. It calculates the mean and standard deviation that are either above or below the mean. Its range is from -3 mean difference to +3 standard deviation.

The Z-score scaling can be estimated by assuming the value of Standard deviation. Here, the value of Sd=1.

Chart, box and whisker chart

Description automatically generated

The Z-score scaling can be estimated by assuming the value of Standard deviation. Here, the value of Sd=2.

Chart, box and whisker chart

Description automatically generated

## **SOFTMAX SCALING:**

A coordinates of K (Lambda) actual figures is transformed into a probability density function with K (Lambda) different alternatives by the SoftMax function. It is also referred to as the normalized exponential function or the soft argmax. It is used in multinomial regression and is a multidimensional generalization of the logistic function.

**Soft Max lambda=1 Soft Max lambda=2**

Chart, bar chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**Soft Max** **lambda=3 Soft Max lambda=4**

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

**Soft Max lambda =10**

Chart, box and whisker chart

Description automatically generated

Here, the **Dependent variable** is **Death** and the population density, lives in household and Health are the attributes which is considered as an independent variable.

## **KOLMOGOROV-SMIRNOV TEST OF NORMALITY FOR EACH VARIABLES:**

The Kolmogorov-Smirnov Test, as well identified as a one K-S Test, is a non-parametric trial of an equality of non - continuous and continuous probability distributions in a 1D probability density function that can be used to compare the sample to the reference probability trial (which is known as two-sample K-S test). To check the given variables are normally distributed through KS-test: if the value of norm>0.05 then it is normally distributed. Therefore, the normality test for each variable is illustrated below:

### **I.KS-Test for Rental Household and Own Household:**

**Graphical user interface, text, application

Description automatically generated**

The One-sample Kolmogorov-Smirnov test for Rental Household and Own Household is performed and the obtained p-value 0.1089 for both variables and the hypothesis is two-sided.

### **II.KS-Test for Bad Health and Very Bad Health:**

**Graphical user interface, text, application

Description automatically generated**

The One-sample Kolmogorov-Smirnov test for Bad Health and Very Bad Health is performed and the obtained p-values are 0.09492 and 0.2544 respectively and the alternative hypothesis is two-sided.

### **III.KS-Test for White and Asian British Ethnic groups:**

**Graphical user interface, text, application

Description automatically generated**

The One-sample Kolmogorov-Smirnov test for White and Asian British ethnic group is performed and the p-values obtained is same for both the variables as 1.088e-07 and the hypothesis is two-sided.

### **IV.KS-Test for Resident\_Male and Resident\_Female:**

**Graphical user interface, text, application

Description automatically generated**

The One-sample Kolmogorov-Smirnov test for Resident\_Male and Resident\_Female is performed, and the p-values obtained is 0.07809 for both the variables and the alternative hypothesis is two-sided.

Therefore, the given Variables are Normally Distributed as the pnorm value is greater than 0.05 as evaluated above.

# **CORRELATION:**

The sample correlation coefficient (r), as shown above for cumulative savings through time, is a metric that indicates how closely the attributes in a scatter graph are affiliated with a line of linear regression based on those points. The correlations performed between dependent variables with other independent variables are given as follows.

# **SPEARMAN’S CORRELATION:**

The coefficient of correlation is a metric that measures the strength as well as direction of a relationship. (r). It is depicted as a negative or positive count between -1 and 1. The integer value represents the strength of the correlation: r = 0 indicates an absence of correlation.

The Spearman's rank coefficient of correlation is the non-parametric technique which is utilized for evaluating rank correlation (statistical relevance of ranking between two variables). It is commonly represented as the Greek letter "rho" and is primarily used in data analysis. Charles Spearman inspired the name.

Company name

Description automatically generated

# **PEARSON’S CORRELATION:**

The Pearson coefficient of correlation is the most widely used method for identifying a linear connection (r). The direction and strength of the relationship between two variables are expressed as numbers ranging from -1 to 1.

Diagram

Description automatically generated with medium confidence

Correlation testing refers to the statistical test that evaluate the statistical correlation, or correlation, between continuous variables. Hence, the test correlation between dependent variable with other independent variables is illustrated below:

## **I. Test Correlation between Total Population vs Rental household and Total Population vs Own household:**

Graphical user interface, text, application

Description automatically generated

The test correlation between Total Population and Rental Household is negatively correlated whereas the test correlation between Total population and Own Household yields Positive Correlation.

## **II. Test Correlation between Total Population vs Bad health and Total Population vs Very Bad health:**

Graphical user interface, text, application

Description automatically generated

Both the correlation test of Total Population vs Bad Health and Total population vs Very Bad Health shows negative correlation.

## **III. Test Correlation between Total Population vs White and Total Population vs Asian:**

Graphical user interface, text, application

Description automatically generated

The test correlation between Total Population and White ethnic groups is negatively correlated whether the correlation test between Total Population and Asian is Positively correlated.

## **IV. Test Correlation between Total Population vs Resident Male and Total Population vs Resident Female:**

Graphical user interface, text, application

Description automatically generated

The correlation test of Total Population vs Resident\_Male is negatively correlated whereas, the correlation test between Total Population and Residen\_Female illustrates Positive correlation.

# **INTERNAL CORRELATION USING PEARSON’S METHOD:**

A measure of dependency in a group of variables known as the internal correlation is examined and generalized. This coefficient serves as the product's upper bound. A set of variables can define instant correlations, multiple correlations, and canonical correlations. The internal correlation between Four variables is given as follows:

## **I. Internal Correlation between Rental household and Bad health:**

Graphical user interface, text, application

Description automatically generated

The p-value of Internal correlation between Rental Household and Bad Health is 0.0068 where the true correlation is not equal to 0 and the sample estimates a negatively correlated with the value of -0.1502.

## **II. Internal Correlation between white and Bad health:**

Graphical user interface, text, application

Description automatically generated

The Internal Correlation between White ethnic group and Bad Health yields a Negative correlation with the value of -0.04276 whereas the p-value is obtained as 0.4437 and the true correlation is said to be not equal to 0.

## **III. Internal Correlation between Rental household and White:**

Graphical user interface, text, application

Description automatically generated

The Internal correlation between Rental Household and White is said to be effectively correlated, because it shows Positive correlation between each variable (i.e., the correlated value is 0.034) and the obtained p-value is 0.5416 where the true correlation is not equal to 0.

A correlation matrix is an array of correlations for a given group of variables that can be used to determine whether there is relationship between those variables. The coefficient shows both the intensity and the direction of the association (positive vs. negative correlations). The overall correlation matrix is shown below:

**Correlation Matrix**

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated

The overall correlation Matrix data frame and the round values of correlation matrix is illustrated above.

# **PARTIAL CORRELATION:**

The partial correlation is the correlation of two variables when at least one additional variable is taken into consideration. The partial coefficient of correlation is a standard measure of an intensity of the relationship that exists between the two variables after all other factors are considered.

Pearson’s partial correlation between the Total Population, Bad health and Rental household is observed in the figure given below.

A screenshot of a computer

Description automatically generated

The library used to obtain Pearson’s Partial Correlation is ppcor. The Partial correlation between Total Population, Bad Health and Rental Household shows a negative correlation with p-value of 0.005596, whereas the Pearson’s Partial Correlation between Total Population, Rental Household and Bad Health is also yielding negative correlation with p-value 0.01534.

# **DEATH ANALYSIS BY AREA:**

## **LOW DEATH RATE AREA:**

The Low Death Rate by local authority areas in England is illustrated below by using a function gg-plot. The graphical representation can be obtained by analyzing the selected data set and divided the death rate by areas such as Low Death Areas, Medium Death Areas and High Death Areas. Then, by creating a data frame and insert the analyzed data by areas into the data frame to obtain the gg-plot of respective areas.

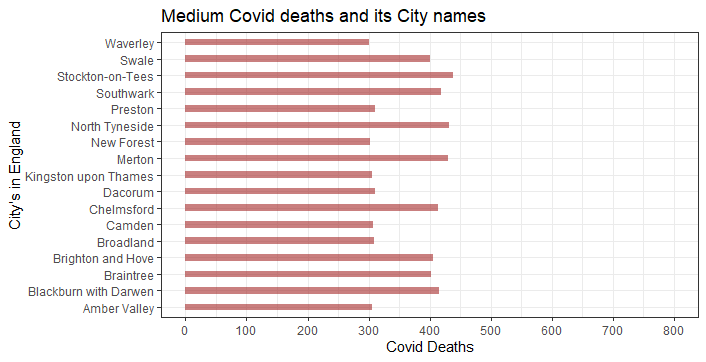
Chart

Description automatically generated

These are the local authorities in England which has lowest death rates.

## **MEDIUM DEATH RATE AREA:**

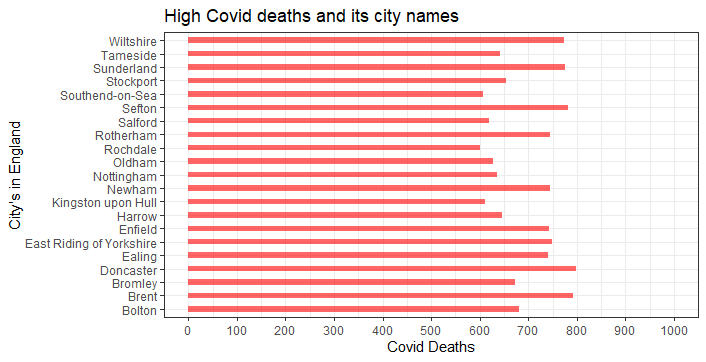
The Medium Death Rate by local authority areas in England is illustrated below by following the same interpretation above and the graphical representation is obtained as follows.



These are the areas in the local authorities in England which has medium death rates.

## **HIGH DEATH RATE AREA:**

The Highest Death Rate Areas can be obtained by analyzing the selected dataset and split them according to the number of deaths in each local authority areas. Thus, the obtained graphical representation for High Death Rate Area is illustrated below.



These are the areas in the Local authorities in England which is referred to be a Highest Death Rate Areas.

# **GG-PLOT FOR EACH VARIABLES:**

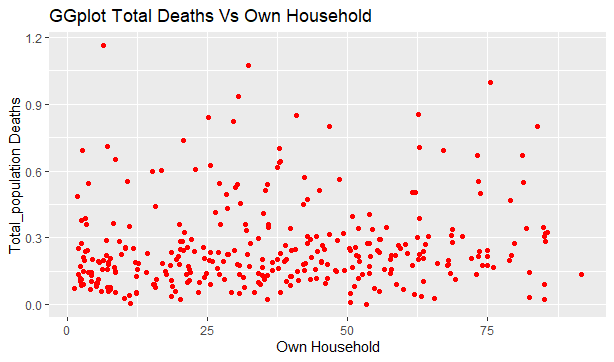
The plotting software ggplot2 provides useful functions for creating complex graphical representations from underlying data within a dataset. It provides much more efficient inputs for determining which variables to plot, how to display them, and other graphical properties. The gg-plot (Graphical Representation) for each independent variables with dependent variable is evaluated and illustrated below.

## **I.gg-plot for Total Death vs Rental Household:**

Chart, scatter chart

Description automatically generated

## **II.gg-plot for Total Death vs Own Household:**



## **III.gg-plot for Total Death vs Bad Health:**

Chart, scatter chart

Description automatically generated

## **IV.gg-plot for Total Death vs Very Bad Health:**

Chart, scatter chart

Description automatically generated

## **V.gg-plot for Total Death vs White ethnic group:**

Chart, scatter chart

Description automatically generated

## **VI.gg-plot for Total Death vs Asian British ethnic group:**

Chart, scatter chart

Description automatically generated

## **VII.gg-plot for Total Death vs Resident\_Male:**

Chart, scatter chart

Description automatically generated

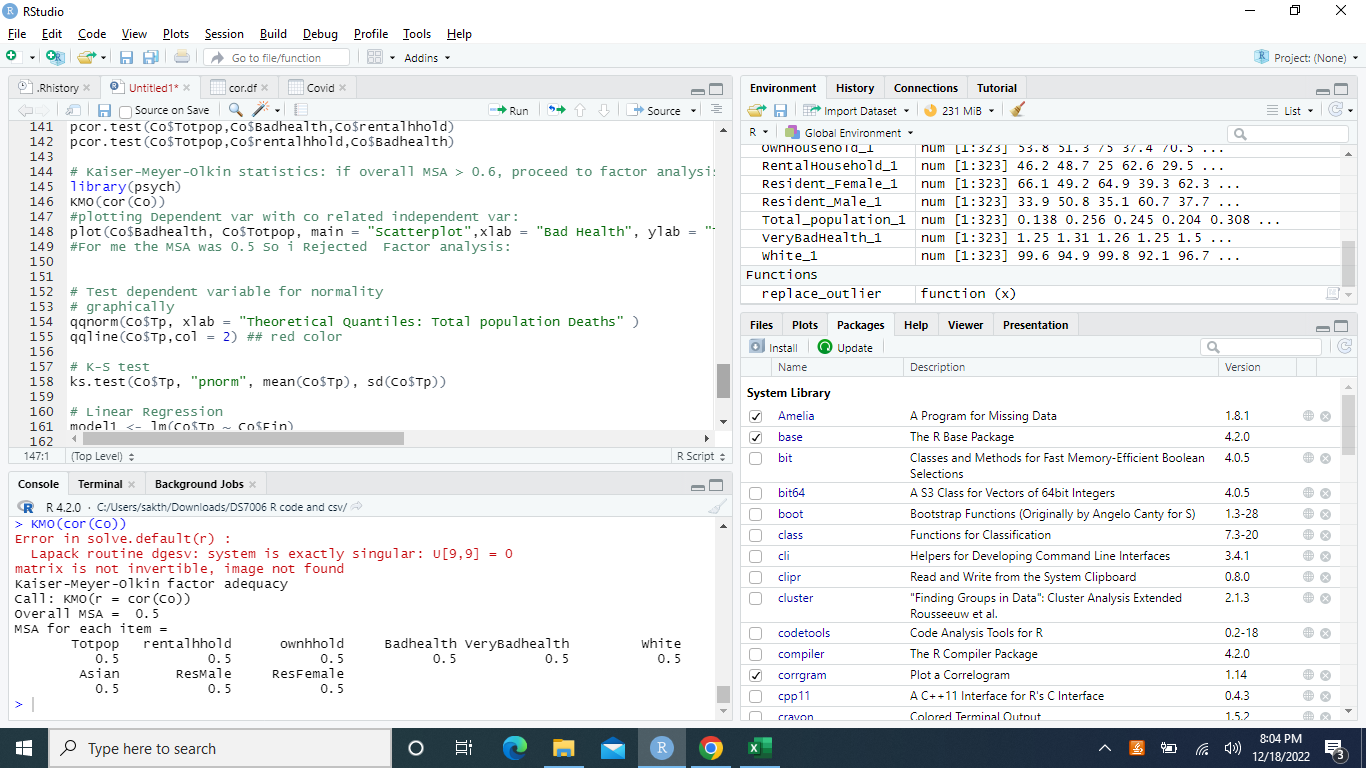
## **VIII.gg-plot for Total Death vs Resident\_Female:**

Chart, scatter chart

Description automatically generated

# **KAISER-MEYER-OLKIN STATISTICAL TEST:**

A Kaiser-Meyer-Olkin (KMO) check can be used to determine whether the information is appropriate for factor analysis. The check estimates regardless of whether sampling is sufficient for an entire model as well as for each individual variable. The statistic represents the percentage of variation among the variables is common variance. There are estimates for each item as well as the overall MSA. The Kaiser-Meyer-Olkin (KMO) index is the name of the index. The factor analysis can be done further is the MSA index is less than 0.5.



Here, the MSA index for each variable is equal to 0.5

# **SCATTERPLOT:**

One dot is plotted for each observation in a "scatter plot," a type of plot used to show the relationship between two numerical variables. It requires two vectors of the same length: one for the longitudinal x-axis (horizontal)  and another for the diagonal y-axis (vertical).

Chart, scatter chart

Description automatically generated

# **NORMALITY TEST:**

A type of hypothesis test called a normality test is used to draw conclusions about the population from which a sample of data was taken. For R, there are numerous normality tests available. Fundamentally, each of these tests evaluates the following hypothesis.

## **I.QQ-Plot:**

Chart, line chart, histogram

Description automatically generated

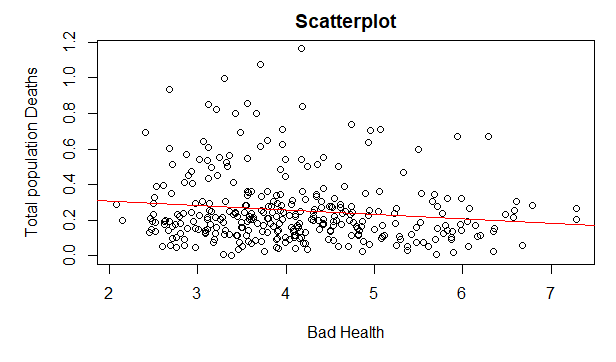
## **II. Regression Lined QQ-Plot:**

Chart, line chart

Description automatically generated

# **LINEAR REGRESSION:**

A form of correlations modelling technique called linear regression makes use of a strong correlation to show how two variables are related. By seeking out the values of the regression coefficient(s) that reduce the overall model error, it determines the optimum line for your data.



The summary of the residuals model is illustrated below:

Graphical user interface, application

Description automatically generated

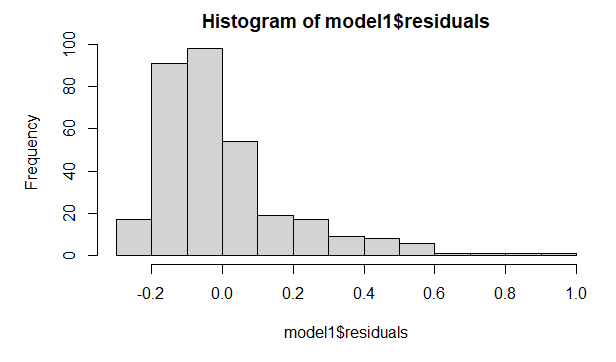
Hence, the outcomes obtained from the Linear Regression between Total Population and Bad Health produces Median as -0.05356, in which the p-value is obtained as 0.01554 and the R-Squared mean error value is 0.018.

## **HISTOGRAM FOR RESIDUALS:**

The residual histogram is used to see if the deviation is regularly distributed. If the bell-shaped histogram is geometric and uniformly distributed around zero, the assumption of normality is likely to be correct.

A rug plot is a data plot with marks along an axis for a single quantitative variable. It is used to display the data's dispersion. As a result, it is comparable to a one-dimensional scatter plot or a histogram with zero-width bins. The plot histogram is obtained below:

**Histogram Residuals. Rug Plot Histogram Residuals.**

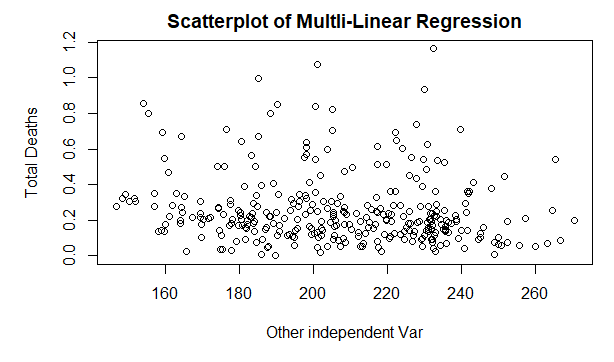
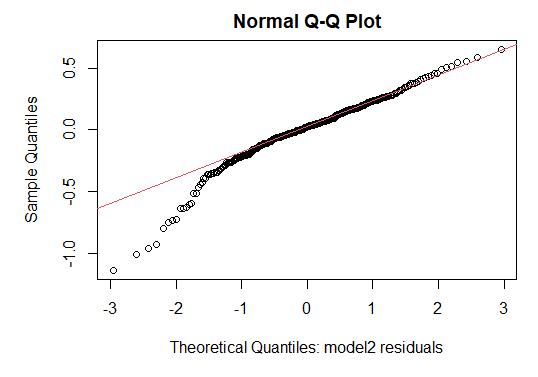
Chart, histogram

Description automatically generated

# **MULTIPLE LINEAR REGRESSION:**

Multiple regression analysis is an extended version of simple linear regression that includes various distinct predictor variables to determine an outcome variable (y) (x). The scatterplot for Multiple linear regression is given as follows.

**Scatterplot Regression lined qq-plot**

Here, the summary of the Regression Model with residuals is illustrated below:

Graphical user interface, application

Description automatically generated

The outcomes obtained from the Residuals produced by the Multiple Linear Regression shows the value of Median as -0.0557 and the p-value is 0.00429 with the Multiple R-Squared Error is obtained as 0.046.

The scatterplot for residuals is obtained as follows:

Chart, scatter chart

Description automatically generated

## **HISTOGRAM FOR RESIDUALS:**

For simple linear regression, the Multiple R-squared valuation is most used (one predictor). It expresses how much variation within the dependent variable can be related to an independent variable. In all other statements, it's an additional method for determining how well the model fits the data.

**Histogram Residuals with Rug plots.**

Chart, histogram

Description automatically generated

# **POISSON REGRESSION:**

A Poisson Regression method is a model of GLM that is utilized to predict numeric attributes and a contingency table. The outcome Y (count) is a Poisson distribution value. It is based on an exponential distribution of predicted value (mean), that can be patterned into a linear form with the help of some unknown parameters. The obtained scatterplot and regression lined qq-plot is illustrated below.

**Scatterplot Regression lined qq-plot**Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

The summary of the Regression model with Residuals is given below.

Graphical user interface, application

Description automatically generated

The Poisson Regression Model is said to effectively strong as it provides a value of Median as -0.116 and the Standard Error value of the overall intercepts is 1.6469. As a result, the Dispersion Attribute for the Poisson Family is set to 1.

Here, the scatterplot for residuals obtained from Poisson Regression is illustrated as follows.

Chart, scatter chart

Description automatically generated

## **HISTOGRAM FOR RESIDUALS:**

A residual histogram is used to see if the deviation is regularly distributed. Whereas if bell-shaped histogram is geometric and distributed uniformly to zero, and a chi-squared value of fit analysis is performed, the normality assumption is likely to be true. The goodfit () function found in the vcd module for count data can also be used. Because the procedure is a Poisson process, we cannot reject the h0 if p value is higher than 0.05. The Histogram for the Residuals obtained from Poisson Regression with Rug plots is illustrated below.

**Histogram for Residuals with rug Plots:**

**Chart, histogram

Description automatically generated**

# **ASYMPTOTIC ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST:**

The Kolmogorov-Smirnov Test, also defined as the one-sample K-S Test, is a non-parametric evaluation of an equality of discontinuous and continuous of a 1-dimensional density function of probabilities that can be used to evaluate samples using the reference's probability trial (which is known as a one-sample K-S test).

Command: ks.test (X,"pnorm")

## **LINEAR REGRESSION:**

The Kolmogorov-Smirnov Test, also defined as the one-sample K-S Test, is a non-parametric trial of an equality of continuous and discontinuous probability distributions in a 1D probability density function that can be used to compare the sample to the reference probability trial (known as two-sample K-S test).

Graphical user interface, application

Description automatically generated

The outcomes of Linear Regression with Asymptotic One-sample Kolmogorov-Smirnov Test provides a p-value of 4.714e-08 whereas the Alternative Hypothesis is said to two-sided.

## **MULTIPLE LINEAR REGRESSION:**

**Graphical user interface, text, application

Description automatically generated**

The result provided by the Multiple Linear Regression with Asymptotic One-sample Kolmogorov-Smirnov Test is relatively greater than the Linear Regression with the p-value 3.053e-07 and the Alternative Hypothesis is two-sided as Linear Regression.

## **POISSON REGRESSION:**

**Graphical user interface, text, application

Description automatically generated**

The outcomes provided from the Poisson Regression Model with Asymptotic One-sample Kolmogorov-Smirnov Test is said to be relatively greater than the Multiple Linear Regression with p-value of 7.105e-07, whereas the Alternative Hypothesis is obtained as two-sided Multiple Linear Regression. Hence, the Poisson Regression Analysis is said to be an efficient as it provides a higher p-value than the other two Regression Models.

# **DATA ANALYSIS:**

Data analysis is the process of analyzing, cleaning, adapting, and modelling data to identify relevant information, sustain inferences, and decision-making. Data analysis is used in a variety of areas, which include business, scientific method, and sociology, and it has a wide range of parameters and methodologies. It utilizes a variety of methods and goes by several names. Data mining is a subset of data analysis that concentrates on obtaining knowledge and data techniques to be predictive rather than descriptive. Business analytics is a subset of data analysis that focuses on business data and depends heavily on aggregation. In statistical applications, three types of analysis are used: analysis of data, data processing (EDA), and validation data analysis (CDA).EDA is involved with recognizing novel features in data, whereas CDA is considered with checking or disputing existing hypotheses.While sentimental analysis incorporates statistics, linguistics, and organizational methodologies to extract and categorize data from unstructured textual resources, predictive modelling emphasizes the utilization of predictive methods for prediction, forecasting, or categorization. These are all different types of data analysis. Data analysis is a step before data integration, and data integration and analysis are intertwined with data visualization and dissemination. The below figure describes a general process of Data Analysis.

**Data Analysis.**

Diagram

Description automatically generated

Hence, for the initial Exploratory analysis the selected data set is analyzed and compared by using data mining techniques as described above. To obtain the result the selected data set is loaded in the R-studio and the visualization is done to know the missing values. Then the boxplot for all variables after obtaining the variable (Total Population) which is proportion to 1000 is observed and the Normalization techniques were evaluated to check whether the variable is normally distributed. The correlation test has been done to know the variables are negatively correlated or positively correlated. Furthermore, the KS-test of normality is obtained to know the difference between the p-values of each variable. Finally, a Regression model is run to compare p-values and determine the most efficient model.

From the above, Exploration part it is evident that the Poisson Regression is highly efficient as it has a higher p-value (rho) than the other two models. (i.e., Linear Regression and Multiple Linear Regression Models). Therefore, the main outcome of this coursework is to estimate the deaths rates related to pandemic process and the changes in the p-values is analyzed and discussed as the variables were normally distributed with MSA value of 0.5. The Asymptotic one-sample Kolmogorov-Smirnov test is performed to obtain the p-value and data estimates and the correlated value to determine the Positive and Negative Correlation.

# **RESULT AND DISCUSSION:**

Temperature and humidity were taken into consideration because, in the past, coronaviruses caused respiratory illnesses, in order to determine whether they had any effect on the transmission of COVID-19. However, no clear correlation between COVID-19 transmission and climatic variables was found following numerous investigations that were carried out in various locations throughout the world and considering local authorities in England with the datasets obtained from Nomis.

The main results were an estimate of the backlog of pandemic-related processes and a change in the p-value of the linear regression analysis when compared to multiple linear regression and Poisson regression models. Estimating the impact of capacity expansion, capacity temporary decrease, and changes in the value of rho (p) with estimations on the backlog of procedures were secondary outcomes. The boxplot for each variable with proportion to 1000 and the inspection between the outliers is performed. The normalization is done by three methods namely, Min-Max scaling, Z-score scaling and Soft-Max scaling. As, the KS-test of normality is also performed for each variable which the MSA is equal to 0.5, therefore the factor analysis is rejected. The gg-plot (Grammar of Graphics) is obtained for all independent variables individually. When analyzing the actual procedure numbers, the overall significant difference between the three Regression Models is p=0.015.

Furthermore, there hasn't been a consistent assessment of the influence of socioeconomic and demographic factors including the proportion of elderly individuals who live in poverty, on overall mortality in relation to COVID-19 at a higher resolution. By identifying high-risk areas, local policies can be reinforced to effectively protect these extremely vulnerable subgroups. Eventually, as the Covid-19 epidemic evolved and changed, a few modelling methodologies were established. Many of these approaches are time-series and do not consciously account for spatial configuration while modelling. Although, these models' ability is to explain this phenomenon is limited by the importance of spatial point (i.e., tiny region) and the influence of nearby small areas. Additionally, the factor analysis has been rejected as the MSA value of each variable is equal to 0.5. In this coursework, the graphical representation of each variable and the estimated value of correlation is demonstrated and the covid death rate of local authorities of England is discussed by using various Data Mining Techniques.

# **CONCLUSION:**

This coursework emphasizes the value and applicability of data mining approaches to the characteristics of the Covid virus outbreaks at high resolution and the mortality rates in the English local authorities. It is important to predict how changes in mobility following the easing of a lockdown will affect the risk that an outbreak will recur locally and to recognize variations in daily/weekly network nodes. The risk of the COVID-19-related mortality rate is also much greater in geographical regions with high proportion of elderly people, especially those who are living in poverty, hence better monitoring and cases follow-up should be assured in these more critical communities. According to mortality trends, more people have been dying in local authorities than of England (normalized for population size).  To comprehend the impacts, as determined by the number of deaths per thousand people, differed so greatly among the various local authority areas in England.

The mean death age as in UK is around 80. Many of the fatalities were 80 years of age or older and may have been nearing the end of their lives due to coexisting conditions like cancer, heart disease, and respiratory problems. The 65+ age group accounts for 87percent of all deaths in the UK. Numerous reports indicate that the clinically obese are the ones who are most at risk in the under 65 age group, which accounts for 13% of all deaths. Diabetes, hypertension, and atherosclerosis are included as comorbid conditions in the descriptions of the deceased who were obese. Additionally, the removal of uncertainty, that has been a significant source of stress in the affected community, is a benefit of these explicit instructions. The science of outbreaks is well-established, and it requires that immediate action be taken to separate local clusters and protect the most vulnerable individuals. Thus, it appears to us that the unprecedented and undifferentiated global lockdown that has occurred in many regions has not been founded in sound scientific reasoning based on a calm analysis based on all prior knowledge but has instead fallen victim to hastily constructed models that stoked an environment of fear that was amplified by the media and social media as attention-seeking buyers.

Furthermore, by Comparing the p-values obtained from the various Regression model gives us the clear understanding of the data used. Normalization techniques or correlated factors, which is the statistical measurement of the correlation strength within two variables, have been estimated to explain the stark discrepancies in performance. Following the results of the epidemic in these local authorities and expanding the comparative study that will offer crucial insights to understand and utilize the successful practices to the greatest extent possible.

**REFERENCES:**

[1]. Alaa, A., Qian, Z., Rashbass, J., Benger, J. and van der Schaar, M. (2020). Retrospective cohort study of admission timing and mortality following COVID-19 infection in England. *BMJ Open*, 10(11), p.e042712. doi:[10.1136/bmjopen-2020-042712](file:///C:\Users\TEMP\Downloads\10.1136\bmjopen-2020-042712).

[2]. Atkins, J.L., Masoli, J.A.H., Delgado, J., Pilling, L.C., Kuo, C.-L., Kuchel, G.A. and Melzer, D. (2020). Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. *The Journals of Gerontology: Series A*, 75(11), pp.2224–2230. doi:[10.1093/gerona/glaa183](file:///C:\Users\TEMP\Downloads\10.1093\gerona\glaa183).

[3]. Barron, E., Bakhai, C., Kar, P., Weaver, A., Bradley, D., Ismail, H., Knighton, P., Holman, N., Khunti, K., Sattar, N., Wareham, N.J., Young, B. and Valabhji, J. (2020). Type 1 and Type 2 Diabetes and COVID-19 Related Mortality in England: A Whole Population Study. *SSRN Electronic Journal*, 8(4). doi:[10.2139/ssrn.3605225](file:///C:\Users\TEMP\Downloads\10.2139\ssrn.3605225).

[4]. Chakraborty, T. and Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, 8(4), p.109850. doi:[10.1016/j.chaos.2020.109850](file:///C:\Users\TEMP\Downloads\10.1016\j.chaos.2020.109850).

[5]. Dennis, J.M., McGovern, A.P., Vollmer, S.J. and Mateen, B.A. (2020). Improving Survival of Critical Care Patients With Coronavirus Disease 2019 in England. *Critical Care Medicine*, Publish Ahead of Print(67). doi:[10.1097/ccm.0000000000004747](file:///C:\Users\TEMP\Downloads\10.1097\ccm.0000000000004747).

[6]. Drake, T.M., Docherty, A.B., Weiser, T.G., Yule, S., Sheikh, A. and Harrison, E.M. (2020). The effects of physical distancing on population mobility during the COVID-19 pandemic in the UK. *The Lancet Digital Health*, [online] 2(8), pp.e385–e387. doi:[10.1016/S2589-7500(20)30134-5](file:///C:\Users\TEMP\Downloads\10.1016\S2589-7500(20)30134-5).

[7]. Ferrando-Vivas, P., Doidge, J., Thomas, K., Gould, D.W., Mouncey, P., Shankar-Hari, M., Young, J.D., Rowan, K.M. and Harrison, D.A. (2020). Prognostic Factors for 30-Day Mortality in Critically Ill Patients With Coronavirus Disease 2019: An Observational Cohort Study. *Critical Care Medicine*, 49(1), pp.102–111. doi:[10.1097/ccm.0000000000004740](file:///C:\Users\TEMP\Downloads\10.1097\ccm.0000000000004740).

[8]. Gray, W.K., Navaratnam, A.V., Day, J., Wendon, J. and Briggs, T.W.R. (2021). Changes in COVID-19 in-hospital mortality in hospitalised adults in England over the first seven months of the pandemic: An observational study using administrative data. *The Lancet Regional Health - Europe*, 5(48), p.100104. doi:[10.1016/j.lanepe.2021.100104](file:///C:\Users\TEMP\Downloads\10.1016\j.lanepe.2021.100104).

[9]. Hillson, K., Clemens, S.C., Madhi, S.A., Voysey, M., Pollard, A.J. and Minassian, A.M. (2021). Fertility rates and birth outcomes after ChAdOx1 nCoV-19 (AZD1222) vaccination. *The Lancet*, [online] 398(10312), pp.1683–1684. doi:[10.1016/S0140-6736(21)02282-0](file:///C:\Users\TEMP\Downloads\10.1016\S0140-6736(21)02282-0).

[10]. Hippisley-Cox, J., Coupland, C.A., Mehta, N., Keogh, R.H., Diaz-Ordaz, K., Khunti, K., Lyons, R.A., Kee, F., Sheikh, A., Rahman, S., Valabhji, J., Harrison, E.M., Sellen, P., Haq, N., Semple, M.G., Johnson, P.W.M., Hayward, A. and Nguyen-Van-Tam, J.S. (2021). Risk prediction of covid-19 related death and hospital admission in adults after covid-19 vaccination: national prospective cohort study. *BMJ*, 10(6), p.n2244. doi:[10.1136/bmj.n2244](file:///C:\Users\TEMP\Downloads\10.1136\bmj.n2244).

[11]. Law, J. (2009). Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology by LAWSON, A. B. *Biometrics*, 65(2), pp.661–662. doi:[10.1111/j.1541-0420.2009.01247\_2.x](file:///C:\Users\TEMP\Downloads\10.1111\j.1541-0420.2009.01247_2.x).

[12]. Maringe, C., Spicer, J., Morris, M., Purushotham, A., Nolte, E., Sullivan, R., Rachet, B. and Aggarwal, A. (2020). The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *The Lancet Oncology*, [online] 21(8), pp.1023–1034. doi:[10.1016/S1470-2045(20)30388-0](file:///C:\Users\TEMP\Downloads\10.1016\S1470-2045(20)30388-0).

[13]. Navaratnam, A.V., Gray, W.K., Day, J., Wendon, J. and Briggs, T.W.R. (2021). Patient factors and temporal trends associated with COVID-19 in-hospital mortality in England: an observational study using administrative data. *The Lancet Respiratory Medicine*, 9(4), pp.397–406. doi:[10.1016/s2213-2600(20)30579-8](file:///C:\Users\TEMP\Downloads\10.1016\s2213-2600(20)30579-8).

[14]. Rice, K., Wynne, B., Martin, V. and Ackland, G.J. (2020). Effect of school closures on mortality from coronavirus disease 2019: old and new predictions. *BMJ*, 6(8), p.m3588. doi:[10.1136/bmj.m3588](file:///C:\Users\TEMP\Downloads\10.1136\bmj.m3588).

[15]. Sartorius, B., Lawson, A.B. and Pullan, R.L. (2021). Modelling and predicting the spatio-temporal spread of COVID-19, associated deaths and impact of key risk factors in England. *Scientific Reports*, 11(1). doi:[10.1038/s41598-021-83780-2](file:///C:\Users\TEMP\Downloads\10.1038\s41598-021-83780-2).

[16]. Williamson, E.J., Walker, A.J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C.E., Curtis, H.J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H.I., MacKenna, B., Tomlinson, L., Douglas, I.J., Rentsch, C.T., Mathur, R., Wong, A.Y.S., Grieve, R. and Harrison, D. (2020). OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature*, [online] 584(548). doi:[10.1038/s41586-020-2521-4](file:///C:\Users\TEMP\Downloads\10.1038\s41586-020-2521-4).

[17]. Xiong, N. and Lin, Z. (2020). Feasibility of controlling COVID-19. *The Lancet Global Health*, 8(6), p.e774. doi:[10.1016/s2214-109x(20)30129-7](file:///C:\Users\TEMP\Downloads\10.1016\s2214-109x(20)30129-7).

[18]. Zakeri, R., Bendayan, R., Ashworth, M., Bean, D.M., Dodhia, H., Durbaba, S., O’Gallagher, K., Palmer, C., Curcin, V., Aitken, E., Bernal, W., Barker, R.D., Norton, S., Gulliford, M., Teo, J.T.H., Galloway, J., Dobson, R.J.B. and Shah, A.M. (2020). A case-control and cohort study to determine the relationship between ethnic background and severe COVID-19. *EClinicalMedicine*, 28(56), p.100574. doi:[10.1016/j.eclinm.2020.100574](file:///C:\Users\TEMP\Downloads\10.1016\j.eclinm.2020.100574).

**Appendix:**

The selected dataset is loaded to R-studio and the below code is used to execute the Data Exploration part.

**#To open the csv file:**

Setwd(dirname(file.choose()))

getwd()

covid <- read.csv ("Covid.data1.csv", stringsAsFactors = FALSE)

str(covid)

library (Amelia)

library(funModeling)

library(tidyverse)

library (Hmisc)

library(ggplot2)

library(forcats)

**#Visualize the missing values:**

apply (covid, MARGIN = 2, FUN = function(x) sum(is.na(x)))

library (Amelia)

missmap(covid, col = c("black", "Grey"), legend = FALSE)

str(covid)

**#Putting the Integer data types variables in one data frame and naming the column:**

Covid <- data.frame(covid$Area.Hectares, covid$Lives.in.Rental.Household, covid$Lives.in.own.household, covid$Total.Household, covid$Bad.health, covid$Very.bad.health, covid$Total.Bad.health, covid$White,covid$Asian.Asian.British, covid$Total.Ethnic.group, covid$Resident..Males, covid$Resident..Females, covid$Total.Residents, covid$Death.in.2020, covid$Death.in.2022, covid$Death.count\_Male, covid$Death.count\_Female, covid$Total.Death, covid$Total.Population)

colnames(Covid) <- c("AreaHectares", "Lives\_RentalHousehold", "Lives\_OwnHousehold", "TotalHousehold",

"BadHealth","VeryBadHealth", "TotalBadHealth", "White",

"AsianBritish", "TotalEthnics", "Resident\_Male", "Resident\_Female","TotalResidents",

"Death\_2020", "Death\_2022", "DeathCount\_Male", "DeathCount\_Female","TotalDeath", "TotalPopulation")

str (Covid)

View (Covid)

**#Converting the All the Attributes into propotion by 1000:**

**#Total population theme:**

Total\_population\_1<- (Covid$TotalDeath/Covid$TotalPopulation)\*1000

**#Household theme:**

RentalHousehold\_1<- (Covid$Lives\_RentalHousehold/Covid$TotalHousehold)\*1000

OwnHousehold\_1<- (Covid$Lives\_OwnHousehold/Covid$TotalHousehold)\*1000

**#Health theme:**

BadHealth\_1<- (Covid$BadHealth/Covid$TotalBadHealth)\*1000

VeryBadHealth\_1<- (Covid$VeryBadHealth/Covid$TotalBadHealth)\*1000

**#Ethnic theme:**

White\_1<- (Covid$White/Covid$TotalEthnics )\*1000

AsianBritish\_1<- (Covid$AsianBritish/Covid$TotalEthnics )\*1000

**#Resident Theme:**

Resident\_Male\_1<- (Covid$Resident\_Male/Covid$TotalResidents)\*1000

Resident\_Female\_1<- (Covid$Resident\_Female/Covid$TotalResidents)\*1000

**#Putting into Relevant data frame:**

Co<-data. frame(Total\_population\_1,RentalHousehold\_1,OwnHousehold\_1,BadHealth\_1,VeryBadHealth\_1 , White\_1, AsianBritish\_1, Resident\_Male\_1, Resident\_Female\_1)

colnames(Co) <-c("Totpop","rentalhhold","ownhhold","Badhealth","VeryBadhealth","White","Asian","ResMale","ResFemale")

str (Co)

boxplot (Co,main="Boxplot After Proportion by 100",xlab="Dependent and Independant Variables",ylab="count",col="deeppink")

**#Replacing the outlier functions:**

replace\_outlier <- function(x){

for (i in which(sapply(x, is.numeric))) { quantiles <- quantile( x[,i], c(.05, .95 ), na.rm =TRUE) x[,i] = ifelse(x[,i] < quantiles[1] , quantiles[1], x[,i])x[,i] = ifelse(x[,i] > quantiles[2] , quantiles[2], x[,i])} x}

**# Replacing the values with percentage**

London\_4 = replace\_outlier(Co)

boxplot(London\_4, xlab="numerical values", ylab="Count", col="deeppink", main="Outliers Replaced :")

**#Normalization(3 Methods Used):**

**#1st Method min-max scaling:**

London.mms<- apply(London\_4 , MARGIN = 2, FUN = function(x) (x - min(x))/diff(range(x)))

boxplot(London.mms, main= "Min Max Scaling ",xlab="numerical values",ylab="count")

**#2nd Method z-score scaling:**

London.z1 <- apply(London\_4, MARGIN = 2, FUN = function(x) (x - mean(x))/sd(x))

London.z2 <- apply(London\_4, MARGIN = 2, FUN = function(x) (x - mean(x))/(2\*sd(x)))

boxplot(London.z1,main= "Standard deviation 1",xlab="numerical values",ylab="count")

boxplot(London.z2,main= "Standard deviation 2",xlab="numerical values",ylab="count")

**##3rd Method soft Max Scaling:**

library(DMwR2)

help(SoftMax)

sts <- apply(London\_4, MARGIN = 2, FUN = function(x) (SoftMax(x,lambda = 1, mean(x), sd(x))))

boxplot (sts, main = "Soft Max, lambda = 1")

sts <- apply(London\_4, MARGIN = 2, FUN = function(x) (SoftMax(x,lambda = 2, mean(x), sd(x))))

boxplot (sts, main = "Soft Max, lambda = 2")

sts <- apply(London\_4, MARGIN = 2, FUN = function(x) (SoftMax(x,lambda = 3, mean(x), sd(x))))

boxplot (sts, main = "Soft Max, lambda = 3")

sts <- apply(London\_4, MARGIN = 2, FUN = function(x) (SoftMax(x,lambda = 4, mean(x), sd(x))))

boxplot (sts, main = "Soft Max, lambda = 4")

sts <- apply(London\_4, MARGIN = 2, FUN = function(x) (SoftMax(x,lambda = 5, mean(x), sd(x))))

boxplot (sts, main = "Soft Max, lambda = 10")

**#Low-Area deaths:**

d <- data.frame(Area\_names= c('Cambridge', 'Corby', 'Cotswold', 'Craven', 'Daventry', 'Eden', 'Exeter', 'Forest of Dean','Harborough','Hart','Maldon','Malvern Hills','North Warwickshire','Ribble Valley','Runnymede', 'Selby', 'South Northamptonshire', 'St Edmundsbury', 'Stevenage', 'Teignbridge', 'Torbay', 'Uttlesford', 'Wellingborough','West Dorset')

Death\_count= c(121,124,133,124,136,132,114,112,151,150,139,130,150,145,155,119,129,116,153,105,149,129,151,138))

ggplot(d,aes(x=Area\_names, y=Death\_count)) +

geom\_bar(stat="identity", fill="darkblue", alpha=.6, width=.4) +coord\_flip() + scale\_y\_continuous(breaks = seq(0,400,100),

limits=c(0,400)) +

ylab("Covid Deaths") +xlab("City's in England")+

ggtitle("Low Covid deaths and its City names") +

theme\_bw()

**#Medium-Area deaths:**

t <- data.frame(Area\_names= c('Amber Valley', 'Blackburn with Darwen', 'Braintree', 'Brighton and Hove','Broadland', 'Camden', 'Chelmsford', 'Dacorum','Kingston upon Thames', 'Merton', 'New Forest', 'North Tyneside', 'Preston', 'Southwark', 'Stockton-on-Tees', 'Swale','Waverley'),

Death\_count= c(306,416,402,405,309,308,414,311,305,430,303,431,311,419,438,401,300))

ggplot(t,aes(x=Area\_names, y=Death\_count)) +

geom\_bar(stat="identity", fill="brown", alpha=.6, width=.4) +

coord\_flip() + scale\_y\_continuous(breaks = seq(0,800,100),

limits=c(0,800)) +

ylab("Covid Deaths") +xlab("City's in England")+

ggtitle("Medium Covid deaths and its City names") +

theme\_bw()

**#High-Area deaths:**

d <- data.frame(Area\_names= c('Bolton', 'Brent', 'Bromley', 'Doncaster','Ealing', 'East Riding of Yorkshire', 'Enfield', 'Harrow', 'Kingston upon Hull', 'Newham', 'Nottingham', 'Oldham', 'Rochdale','Rotherham', 'Salford', 'Sefton', 'Southend-on-Sea', 'Stockport', 'Sunderland', 'Tameside', 'Wiltshire'),

Death\_count c(680,792,673,799,741,748,743,645,611,745,636,628,600,745,619,781,607,654,775,641,773))

ggplot(d,aes(x=Area\_names, y=Death\_count)) +

geom\_bar(stat="identity", fill="red", alpha=.6, width=.4) +

coord\_flip() + scale\_y\_continuous(breaks = seq(0,1000,100),

limits=c(0,1000)) +

ylab("Covid Deaths") +xlab("City's in England")+

ggtitle("High Covid deaths and its city names") +

theme\_bw()

**#Co-relation Map:**

library(corrgram)

corrgram(sts, order=FALSE, cor.method = "spearman", lower.panel=panel.cor,upper.panel=panel.pie, text.panel=panel.txt, main="Deaths of Covid (Spearman correlation)")

**# Correlation between the binary values :**

cor.matrix <- cor(Co, use = "pairwise.complete.obs", method = "pearson")

round(cor.matrix, digits = 2)

cor.df <- as.data.frame(cor.matrix)

View(cor.df)

round(cor.df, 2)

**#Co-relation Map:**

library(corrgram)

corrgram(Co, order=FALSE, cor.method = "pearson", lower.panel=panel.cor,upper.panel=panel.pie, text.panel=panel.txt, main="Deaths of Covid (pearson correlation)")

**# correlation test of the dependent variable with all an independent variables:**

cor(Co)

cor.test(Co$Totpop, Co$rentalhhold, method ="pearson")

cor.test(Co$Totpop, Co$ownhhold,method ="pearson")

cor.test(Co$Totpop, Co$Badhealth, method ="pearson")

cor.test(Co$Totpop, Co$VeryBadhealth, method ="pearson")

cor.test(Co$Totpop, Co$White, method ="pearson")

cor.test(Co$Totpop, Co$Asian,method ="pearson")

cor.test(Co$Totpop, Co$ResMale, method ="pearson")

cor.test(Co$Totpop, Co$ResFemale, method ="pearson")

**#GGplot of the Dependent var with all independent Var:**

library(ggplot2)

ggplot() + geom\_point(aes(x = Co$rentalhhold, y = Co$Totpop),

colour = 'darkblue') +

ggtitle("GGplot Total Deaths Vs Rental Household") +

xlab('Rental Household') +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$ownhhold, y = Co$Totpop),

colour = 'red') +

ggtitle("GGplot Total Deaths Vs Own Household") +

xlab('Own Household) +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$Badhealth, y = Co$Totpop),

colour = 'yellow') +

ggtitle("GGplot Total Deaths Vs Bad Health") +

xlab('Bad Health') +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$Very Bad Health, y = Co$Totpop),

colour = 'brown') +

ggtitle("GGplot Total Deaths Vs Very Bad Health") +

xlab('Very Bad Health') +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$White, y = Co$Totpop),

colour = 'black') +

ggtitle("GGplot Total Deaths Vs White") +

xlab('White') +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$Asian, y = Co$Totpop),

colour = 'violet') +

ggtitle("GGplot Total Deaths Vs Asian") +

xlab('Asian') +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$ResMale, y = Co$Totpop),

colour = 'purple') +

ggtitle("GGplot Total Deaths Vs ResMale") +

xlab('ResMale') +

ylab('Total\_population Deaths')

ggplot() +

geom\_point(aes(x = Co$ResFemale, y = Co$Totpop),

colour = 'green') +

ggtitle("GGplot Total Deaths Vs ResFemale") +

xlab('ResFemale') +

ylab('Total\_population Deaths')

**## looking at internal correlations between Four variables:**

cor.test(Co$rentalhhold,Co$Badhealth, method = "pearson")

cor.test(Co$White,Co$Badhealth, method = "pearson")

cor.test(Co$rentalhhold,Co$White, method = "pearson")

**#partial correlation**

library(ppcor)

**#calculate partial correlation using Pearson**

pcor.test(Co$Totpop,Co$Badhealth,Co$rentalhhold)

pcor.test(Co$Totpop,Co$rentalhhold,Co$Badhealth)

**# Kaiser-Meyer-Olkin analysis: whether overall MSA < 0.5, factor analysis can be rejected**

library(psych)

KMO(cor(Co))

**#plotting Dependent var with co related independent var:**

plot(Co$Badhealth, Co$Totpop, main = "Scatterplot",xlab = "Bad Health", ylab = "Total\_population Deaths")

**#The MSA was 0.5 So the Factor analysis is rejected:**

**# Normality Test of dependent variable**

**# graphical representation**

qqnorm(Co$Totpop, xlab = "Theoretical Quantiles: Total population Deaths" )

qqline(Co$Totpop,col = 2) ## red color

**# K-S Normality test**

ks.test(Co$Totpop, "pnorm", mean(Co$Totpop), sd(Co$Totpop))

**# Linear Regression Analysis**

model1 <- lm(Co$Totpop ~ Co$Badhealth)

**# Add regression line to scatter plot**

plot(Co$Badhealth, Co$Totpop, main = "Scatterplot", xlab = "Bad Health", ylab = "Total population Deaths")

abline(model1, col = "red")

summary(model1)

hist(model1$residuals)

rug(model1$residuals)

**#Multi-Linear regression Model:**

model2<-lm(Co$Totpop~Co$Badhealth+Co$rentalhhold+Co$ResMale+Co$White)

**# Add regression line to scatter plot:**

Plot(Co$Badhealth+Co$rentalhhold+Co$ResMale+Co$White, Co$Totpop, main = "Scatterplot of Multi-Linear Regression", xlab = "Other independent Var", ylab = "Total Deaths")

abline(model2, col = "red")

summary(model2)

hist(model2$residuals)

rug(model2$residuals)

**# Normality of residuals:**

Plot (model2$residuals ~ model2$fitted.values, xlab = "fitted values", ylab = "residuals")

ks.test(model2$residuals, "pnorm", mean(model2$residuals), sd(model2$residuals))

**#Poisson regression:**

p1<-glm(formulaCo$Totpop ~ Co$Badhealth+Co$rentalhhold+Co$ResMale+Co$White, data = Co, family = poisson)

**# Add Poisson regression line to scatter plot:**

plot (Co$Badhealth+Co$rentalhhold+Co$ResMale+Co$White, Co$Totpop, main = "Scatterplot of poisson Regression", xlab = "Other independent Var", ylab = "Total Deaths")

abline(p1, col = "red")

summary(p1)

hist(p1$residuals)

rug(p1$residuals)

**# Normality of residuals:**

plot (p1$residuals ~ p1$fitted.values, xlab = "fitted values", ylab = "residuals")

ks.test(p1$residuals,"pnorm",mean(p1$residuals), sd(p1$residuals))

The overall global environment screen is attached below for reference.

A screenshot of a computer

Description automatically generated with medium confidence