QUANTITATIVE DATA ANALYSIS PROJECT

Covid-19 Deaths & Age, Type of Work, Distance Travel to Work, Ethnicity, Disability

Noor Mohammad Atapoor (2590537)

DECEMBER 10, 2023

UNIVERSITY OF EAST LONDON(UEL)

Table of Contents

Abstract: Covid-19 pandemic was dramatically spread over the world, took many life, and disrupted aspects of people social life and activities across countries. Countries around the world took different measures to prevent the pandemic outbreak, and save people life. This research aims to study the association between a number of demographic and socioeconomic factors with the Covid-19 infection and fatalities rates. The factors such as age, type of work, disability, distance travel to work, gender, ethnicity are examined in association with Covid-19 deaths number using the census data obtained from the England local authorities. Quantitative data analysis was performed on the Covid-19 deaths number as dependent variable, and census data as independent variables to explore, analyze and model the Covid-19 deaths with socioeconomic effective factors. The regression model created based on the data set, implies that restring movement is a significant measure for reducing the pandemic infection as it reveals that 'Part-Time', and 'Work from Home' are the most effective factors associated with Covid-19 deaths in the model and indicate a remarkable negative association with the dependent variable among all other factors were examined in the research.

Keywords: Covid-19, Census Data, Correlation, Collinearity, Regression, Clustering

1 Introduction

Research studies conducted on identifying effective factors and measures in avoiding or controlling the spread of the Covid-19 viruses suggest that there are a number of effective measures were taken by different countries in the pandemic period. For example lockdown, working from home, part-time job are of those controlling measures that have been effective in reducing the infection and mortality rates of the virus (Alfano & Ercolano, 2020), & (Vinceti, et al., 2022). A number of other research findings suggest that age and disabilities play role in the virus infection (Goujon, et al., 2020), & (Fadinger & Schymik, 2020). Due to disastrous impact the pandemic has had on different aspect of our life, especially in education, economy and health (Tarkar, 2020), & (Padhan & Prabheesh, 2021), it is worthwhile to do further research, and examine different measures and factors are felt to be having associations with the Covid-19 pandemic infections and fatalities.

This data analysis project is designed to quantitatively study the association of Covid-19 deaths with a group of socioeconomic variables obtained from England local authorities' census data. The research focuses on examining of the association between Covid-19 deaths, and demographic features or variables obtained from different themes such as the population age groups, disabilities, types of work, distance travel to work, gender, and ethnicity, and try to achieve objectives through answering research questions defined below:

1.1 Objectives

1. To analyze the relationships between the Covid-19 deaths and the demographic variables such as age, disabilities, working types, work traveling distance, and ethnicities based on the England Local Authorities population;
2. To identify the most effective socioeconomic variables of the population explaining the Covid-19 deaths rate;
3. To formulize the associations of identified effective socioeconomic variables with the covid-19 deaths;

1.2 Research Questions

1. Is there any correlation or association between the number of Covid-19 deaths and socioeconomic variables of the populations such as ages, disabilities, working types, work traveling distance, and ethnicities?
2. Which of the socioeconomic variables of the populations are more associated and explaining well the Covid-19 deaths?
3. How to mathematically define the association of Covid-19 deaths with the identified effective socioeconomic variables of the population?

1.3 Methodology

The quantitative data analysis approach was taken in analyzing the Covid-19 and census data obtained from (https://www.nomisweb.co.uk). The census data on the population Ages, Disabilities, Distance travel to Work, Work type, and Ethnicity of the England Local Authorities were analyzed considering the following major steps of quantitative data analysis:

- **Data Exploration:** the downloaded CSV files were uploaded into tables of the SQLite database, merged and exported into a single CSV file. Using a number of appropriate statistical tools in R, the data were cleaned, assessed for missing, outliers and normality, as well as were standardized, checked for collinearity, and finally get ready for analysis;
- **Data Analysis:** all the dependent variables, and a number independent variables undergone hypothesis tests for examining their representativeness of the population. Using the Principal Component Analysis (PCA), factor analysis process was performed on the independent variable in purpose of dimension reduction;
- **Data Modeling:** using the regression modelling technique, a number of different model were created out of different set of the data variables; the models were assessed and compared with each other based on a common number of statistical characteristics, and finally the best model was selected out of others and formulated mathematically.

2 Literature Review

Detected the first case in December 2019, China, the Covid-19 viruses rapidly spread worldwide and was declared as pandemic by the World Health Organization(WHO) on the March 11, 2020 (Ciotti, et al., 2020). The pandemic was disastrous in all aspect of human life. Till January 2020, It infected more than 99.7 million and took the lives of more than 2.14 million people around the world (Cifuentes-Faura, 2021). As of November 10, 2023, the total deaths of Covid-19 in United Kingdom have been 197270 people. The pandemic also drastically disrupted other areas of human life; Schools and other training institutions were physically closed (Tarkar, 2020); the education systems around the world witnessed a paradigm shift towards online education system (Pokhrel & Chhetri, 2021); and the global economy damaged unprecedentedly (Padhan & Prabheesh, 2021).

To contain the virus outbreak, a number of different measures such washing hands, wearing masks were put in action in all over the world (Pokhrel & Chhetri, 2021). The non-pharmacological measure is one of them. This measure focuses on the restriction of mobility and reducing the people interaction that are used widely in all countries (Vinceti, et al., 2022). This type of measure is implemented in the form of changing working style, social distancing, or any type of activities reducing the movement and mobility of the people such as lockdown practiced by many countries that has been very effective in declining the infection rate (Alfano & Ercolano, 2020).

The pandemic has more negative impact on the children with disabilities, not because of their poor health condition, also because of social circumstances in which they live (Shakespeare, et al., 2021). A study in Germany suggest that work from home is effective in reducing Covid-19 infections rates and fatalities, as it is high among employees working less from home (Fadinger & Schymik, 2020). There are studies suggesting that overall the rate of infection is higher among the people who are 35 to 65 years' old (Goujon, et al., 2020).

3 Data Exploration

3.1 Preparing Data

The downloaded census data as CSV files were uploaded into SQLite database tables, as the following screenshotted database scheme shows.

Tables (12)	
Age	CREATE TABLE "Age" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "Age_Total" INTEGER, "Age_0to9" IN
COVID19_Death	CREATE TABLE "COVID19_Death" ("Geography" TEXT, "Geo_Code" TEXT, "March_2020" INTEGER, "April_2020" INTE
Disability	CREATE TABLE "Disability" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "Disb_Total" INTEGER, "DDLimit
Distance_to_Work	CREATE TABLE "Distance_to_Work" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "Dist_Total" INTEGER,
Dweling_Type	CREATE TABLE "Dweling_Type" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Cetegory" INTEGER, "L
Ethnic_Group	CREATE TABLE "Ethnic_Group" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Ethnic" INTEGER, "While
Family_Type	CREATE TABLE "Family_Type" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Families_Households" IN
Hours_Worked	CREATE TABLE "Hours_Worked" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "HW_All_Usual_Residents1
Household	CREATE TABLE "Household" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Categories" INTEGER, "On
Living_Arrangement	CREATE TABLE "Living_Arrangement" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Categories" INTE
Occupation	CREATE TABLE "Occupation" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Occupations" INTEGER, "N
Qualification	CREATE TABLE "Qualification" ("Date" INTEGER, "Geography" TEXT, "Geo_Code" TEXT, "All_Categories" INTEGER, "N
Indices (0)	
Views (0)	
Triggers (0)	

The *Covid19 Deaths, Age, Disability, Distance to Work, Hours Worked, and Ethnic* Groups themes were selected to establish the data set. Executing the following SQL query, the data tables for each selected themes were joined, and the query result was exported as a single CSV file named 'DS7006E'.

DS7006Query.sql

```sql
-- Join a number of tables for creating the data set;
SELECT * FROM COVID19_Death cvd, Age ag, Disability disa, Distance_to_Work distw, Hours_Worked hw, Ethnic_Group eth
        WHERE cvd.Geo_Code = ag.Geo_Code AND cvd.Geo_Code = disa.Geo_Code AND cvd.Geo_Code = distw.Geo_Code
          AND cvd.Geo_Code = hw.Geo_Code AND cvd.Geo_Code = eth.Geo_Code

```

	Geography	Geo_Code	March_2020	April_2020	May_2020	June_2020	July_2020	August_2020	September_2020	October_
1	Adur	E07000223	0	22	15	0	0	0	3	
2	Allerdale	E07000026	1	42	15	4	4	8	0	
3	Amber Valley	E07000032	4	54	25	11	2	1	1	
4	Arun	E07000224	1	31	25	7	0	0	1	
5	Ashfield	E07000170	1	71	51	10	5	0	1	
6	Ashford	E07000105	1	48	40	41	22	4	5	
7	Aylesbury Vale	E07000004	1	82	30	6	5	3	0	
8	Babergh	E07000200	2	33	17	6	0	0	0	
9	Barking and Dagenham	E09000002	15	119	20	4	0	0	4	
10	Barnet	E09000003	46	321	55	6	2	1	2	
11	Barnsley	E08000016	1	111	80	30	6	0	2	

```
Execution finished without errors.
Result: 323 rows returned in 374ms
At line 2:
SELECT * FROM COVID19_Death cvd, Age ag, Disability disa, Distance_to_Work distw, Hours_Worked hw, Ethnic_Group eth
WHERE cvd.Geo_Code = ag.Geo_Code AND cvd.Geo_Code = disa.Geo_Code AND cvd.Geo_Code = distw.Geo_Code AND cvd.Geo_Code = hw.Geo_Code
AND cvd.Geo_Code = eth.Geo_Code
```

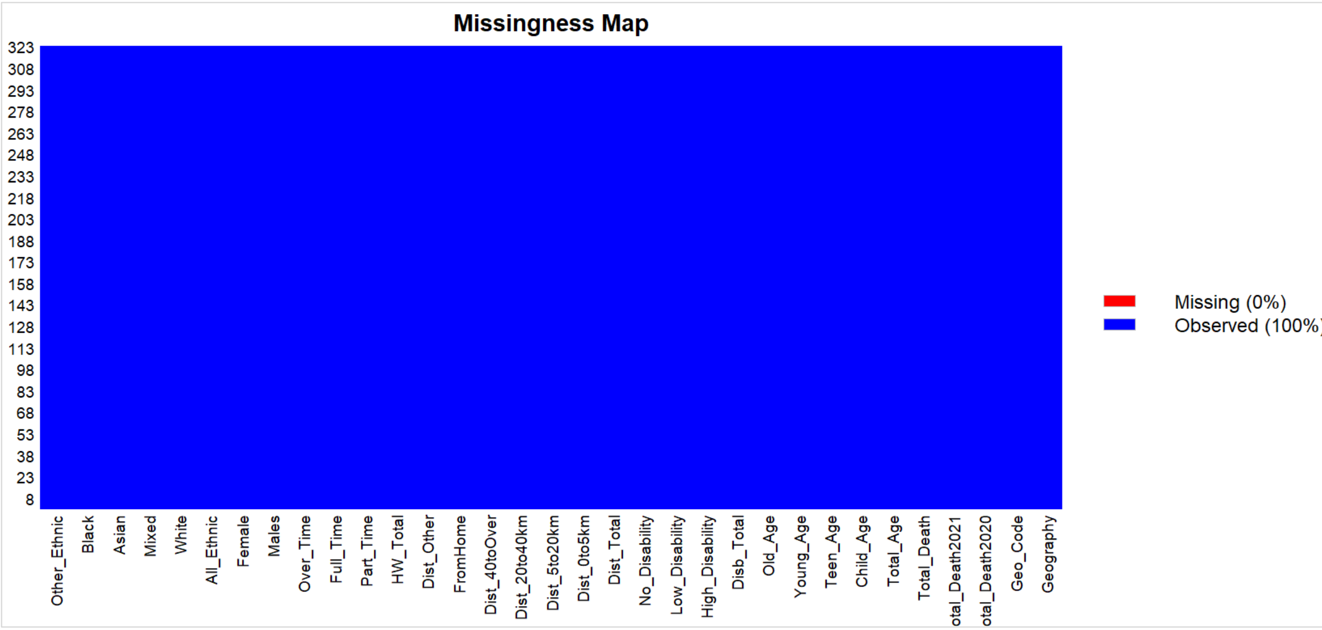
The final scheme of variables, after renaming and removing duplicates, looks as following:

Theme	No	Variable Name	Type	Description
Covid-19 Deaths	1	Total_Death	DV	Total deaths number in each district in England
	2	Total_Death2020	DV	Total deaths of each district from March to December 2020
	3	Total_Death2021	DV	Total deaths of each district from January to April 2021
Age	1	Total_Age	IV	Total population of each district in England
	2	Child_Age	IV	The population aged 0-9 years
	3	Teen_Age	IV	The population aged 10-20 years
	4	Young_Age	IV	The population aged 20-44 years
	5	Old_Age	IV	The population aged 45-Over
Disability	1	Total_Disability	IV	Total population of each district in England
	2	High_Disability	IV	The population with day to day limit in movement;
	3	Low_Disability	IV	The population with day to day little limit in movement;
	4	No_Disability	IV	The population with no day to day limit movement;
Distance to Work	1	Total_Dist	IV	
	2	Dist_0to5	IV	
	3	Dist_5t020	IV	
	4	Dist_20to40	IV	
	5	Dist_40orOver	IV	
	6	From_Home	IV	
	7	Dist_Other	IV	
Hours Worked	1	HW_Total	IV	
	2	Part_Time	IV	
	3	Full_Time	IV	
	4	Over_Time	IV	
	5	Male	IV	
	6	Female	IV	
Ethnic Groups	1	All_Ethnic	IV	
	2	White	IV	
	3	Mixed	IV	
	4	Asian	IV	
	5	Black	IV	
	6	Other_Ethnic	IV	

3.2 Checking for Missing Values

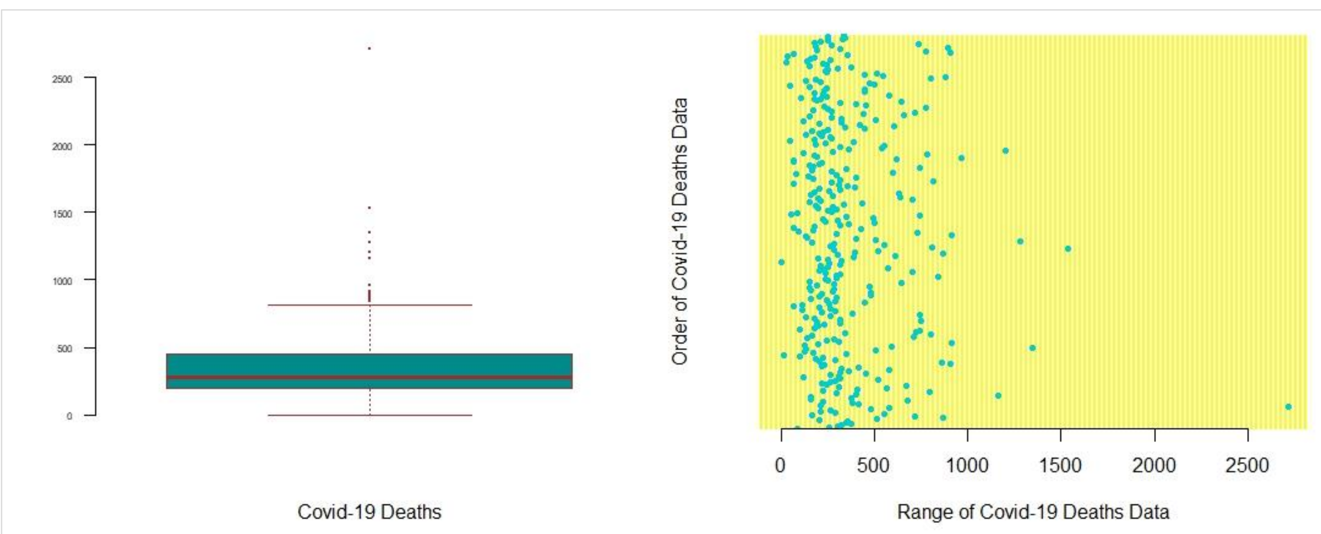
Using the following code, the data was checked for missing value; the check result ensure that there are no missing values in the data set.

```
missmap(DS7006E, col = c("red", "blue"), legend = TRUE) # Check missing data visually;
```



3.3 Checking for Outliers in the Dependent Variable

Box plot and Cleveland dotplot were used to check outliers in the dependent variable. The Box plot show many outliers, however, the Cleveland dotplot indicates except one of the outliers which is far out of range, the others are negligible.



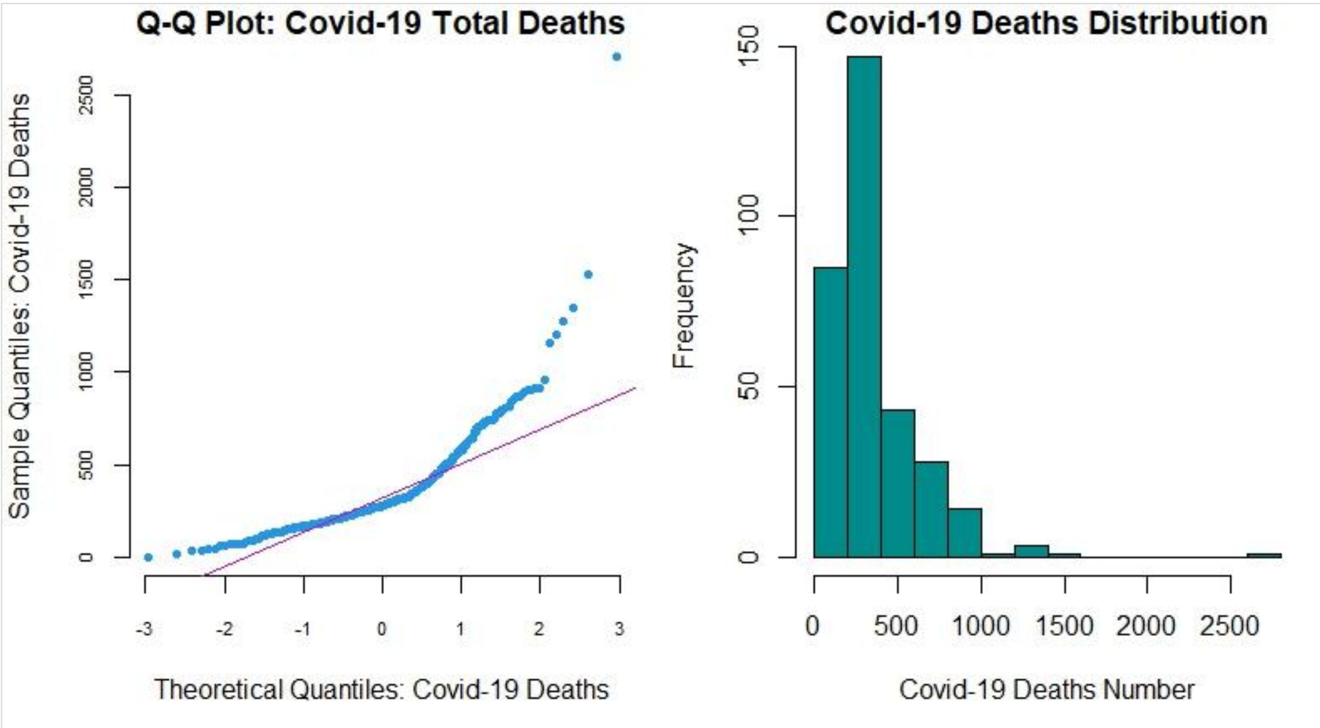
3.4 Checking for Normality: Dependent Variables

Three tools such as Q-Q plot, Histogram, and KS-Test for normality were used to check the normality of the dependent variables. As shown below, none of them confirm the normality of the dependent variable.

The hypothesis for the KS-Test is as following:

- $\boldsymbol{H_0}$: the sample drawn from the population which has normally been distributed;
- $\boldsymbol{H_a}$: the sample drawn from the population which has not normally been distributed;

```
R 4.3.1 · D:/UEL_Courses_2023/T01_Quantitative_Data_Analysis_2023/T01-Course_Works_2023/Data_Set/

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Total_Death
D = 0.18383, p-value = 6.602e-10
alternative hypothesis: two-sided
```



3.5 Normality Check for Independent Variables

The normality of all the independent variables also were checked using the KS-Test. The test results show that none of these variables are normally distributed. The summary of the test results has been written in the table below:

Theme	Variable	Test	Test Value	P-Value	Null Hypothesis	Normality
Age	Child_Age	KS-Test	0.17323	7.624e-09	Rejected	No normal distribution
	Teen_Age	KS-Test	0.17637	3.75e-09	Rejected	No normal distribution
	Young_Age	KS-Test	0.18483	5.214e-10	Rejected	No normal distribution
	Old_Age	KS-Test	0.16111	1.045e-07	Rejected	No normal distribution
Disability	High_Disability	KS-Test	0.18348	7.183e-10	Rejected	No normal distribution
	Low_Disability	KS-Test	0.17143	1.138e-08	Rejected	No normal distribution
	No_Disability	KS-Test	0.16745	2.716e-08	Rejected	No normal distribution
Distance to Work	Total_Dist	KS-Test	0.16428	5.369e-08	Rejected	No normal distribution
	Dist_0to5	KS-Test	0.15134	7.497e-07	Rejected	No normal distribution
	Dist_5t020	KS-Test	0.2099	8.712e-13	Rejected	No normal distribution
	Dist_20to40	KS-Test	0.12473	8.639e-05	Rejected	No normal distribution
	Dist_40orOver	KS-Test	0.16555	4.096e-08	Rejected	No normal distribution
	From_Home	KS-Test	0.14112	5.177e-06	Rejected	No normal distribution
Hours Worked	HW_Total	KS-Test	0.16428	5.369e-08	Rejected	No normal distribution
	Part_Time	KS-Test	0.17545	4.627e-09	Rejected	No normal distribution

	Full_Time	KS-Test	0.17391	6.547e-09	Rejected	No normal distribution
	Over_Time	KS-Test	0.14943	1.088e-06	Rejected	No normal distribution
	Male	KS-Test	0.16754	2.669e-08	Rejected	No normal distribution
	Female	KS-Test	0.16057	1.168e-07	Rejected	No normal distribution
Ethnic Groups	White	KS-Test	0.17688	3.338e-09	Rejected	No normal distribution
	Mixed	KS-Test	0.263	2.2e-16	Rejected	No normal distribution
	Asian	KS-Test	0.31684	2.2e-16	Rejected	No normal distribution
	Black	KS-Test	0.34859	2.2e-16	Rejected	No normal distribution
	Other_Ethnic	KS-Test	0.32325	2.2e-16	Rejected	No normal distribution

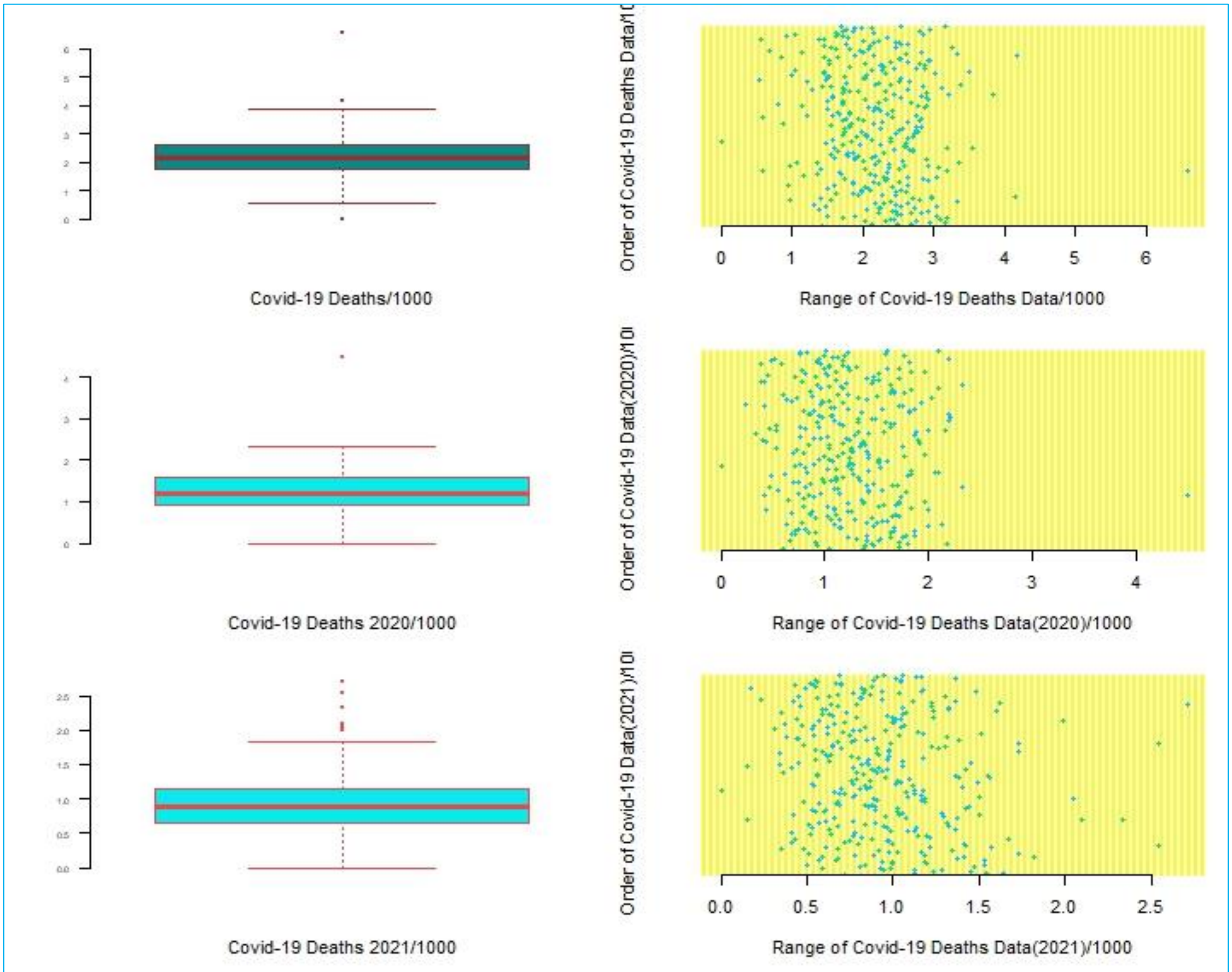
3.6 Data Standardization

The data, especially dependent variables, need to show normal distribution in order to fulfil minimal assumptions of the subsequent statistical process. There are a number of mechanism used for normalizing the data distribution; however, here normalization of distribution is sought through standardizing the data by calculating its percentage based on the total population (per 1000 total population in each local authority in England). To do so, the following code snip was executed:

```
DS7006E <- within(DS7006E, pTotal_Death <- (Total_Death/Total_Age * 1000))
DS7006E <- within(DS7006E, pTotal_Death2020 <- (Total_Death2020/Total_Age * 1000))
DS7006E <- within(DS7006E, pTotal_Death2021 <- (Total_Death2021/Total_Age * 1000))

DS7006E <- within(DS7006E, pChild_Age <- (Child_Age/Total_Age * 1000))
DS7006E <- within(DS7006E, pTeen_Age <- (Teen_Age/Total_Age * 1000))
DS7006E <- within(DS7006E, pYoung_Age <- (Young_Age/Total_Age * 1000))
DS7006E <- within(DS7006E, pOld_Age <- (Old_Age/Total_Age * 1000))

DS7006E <- within(DS7006E, pTotal_Disability <- (Total_Disability/Total_Age * 1000))
DS7006E <- within(DS7006E, pHigh_Disability <- (High_Disability/Total_Age * 1000))
DS7006E <- within(DS7006E, pLow_Disability <- (Low_Disability/Total_Age * 1000))
DS7006E <- within(DS7006E, pNo_Disability <- (No_Disability/Total_Age * 1000))

DS7006E <- within(DS7006E, pHW_Total <- (HW_Total/Total_Age * 1000))
DS7006E <- within(DS7006E, pPart_Time <- (Part_Time/Total_Age * 1000))
DS7006E <- within(DS7006E, pFull_Time <- (Full_Time/Total_Age * 1000))
DS7006E <- within(DS7006E, pOver_Time <- (Over_Time/Total_Age * 1000))

DS7006E <- within(DS7006E, pWhite <- (White/Total_Age * 1000))
DS7006E <- within(DS7006E, pMixed <- (Mixed/Total_Age * 1000))
```

3.6.1 Check for Outliers in Standardized DVs

The number of outlier in each DVs reduced after standardization. According to the *Cleveland dotplot*, only one outlier in each DV shows to has extreme value.



3.6.2 Display Rows in Data Set Containing Outliers

To know about the records having outliers in the data set, a function (*findOultier*), shown below, was developed to detect and display the records containing outliers. The more extreme value that both the box plot and Cleveland plot are agreed upon to be outlier, is 6.5856 and belong to the *East Riding of Yorkshire* district.

```
findOutlier <- function(boxP,dataVar,dataSet, empVec,colmNames){
  for (i in 1:length(boxP$group)) {
    empVec[i] <- which(dataVar == boxP$out[i])
  }
  outData <- dataSet[empVec,colmNames]
  return(outData)
}
# Prepare parameters for the function
colN1 <- c("Geo_Code","Geography","Total_Death", "pTotal_Death","Total_Age")
colN2 <- c("Geo_Code","Geography","Total_Death2020", "pTotal_Death2020","Total_Age")
colN3 <- c("Geo_Code","Geography","Total_Death2021", "pTotal_Death2021","Total_Age")
empV1 <- c()
empV2 <- c()
empV3 <- c()
# Call function for displaying outliers in each dependent variables
outDataT <- findOutlier(covBoxTotal,pTotal_Death,DS7006E,empV1,colN1)
outData21 <- findOutlier(covBox2020,pTotal_Death2020,DS7006E,empV2,colN2)
outData22 <- findOutlier(covBox2021,pTotal_Death2021,DS7006E,empV3,colN3)
```

	Geo_Code	Geography	Total_Death	pTotal_Death	Total_Age
48	E07000069	Castle Point	365	4.147209	88011
89	E07000193	East Riding of Yorkshire	748	6.585493	113583
137	E06000053	Isles of Scilly	0	0.000000	2203
274	E07000076	Tendring	576	4.172462	138048

	Geo_Code	Geography	Total_Death2020	pTotal_Death2020	Total_Age
89	E07000193	East Riding of Yorkshire	510	4.490109	113583

	Geo_Code	Geography	Total_Death2021	pTotal_Death2021	Total_Age
48	E07000069	Castle Point	224	2.545136	88011
89	E07000193	East Riding of Yorkshire	238	2.095384	113583
90	E07000061	Eastbourne	232	2.333722	99412
123	E07000062	Hastings	185	2.049771	90254
213	E07000064	Rother	230	2.538968	90588
248	E06000033	Southend-on-Sea	345	1.986663	173658
274	E07000076	Tendring	374	2.709203	138048

3.6.3 Check the Normality of Standardized Dependent Variables (DVs)

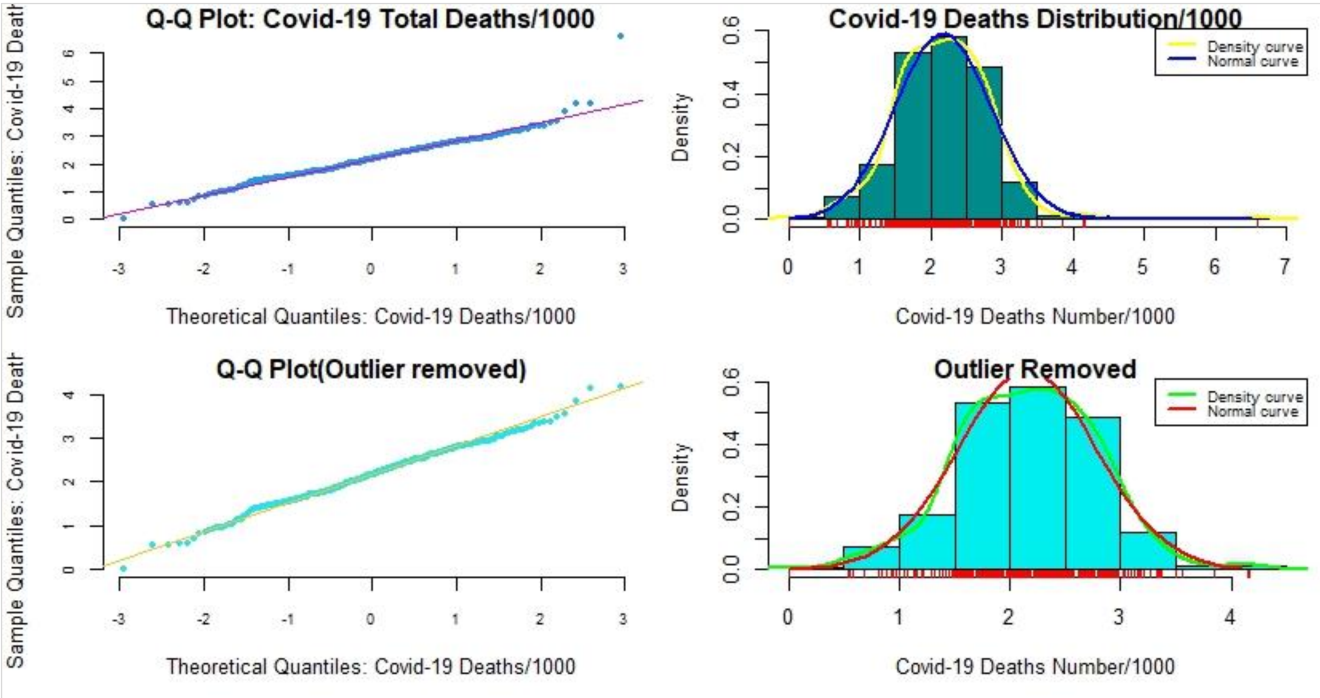
After scaling up the dependent variables, Kolmogorov-Smirnov (KS) test performed on DVs to check its normality. The test (D = 0.0435, P-Value (0.572) > 0.05) is no significant and the null hypothesis of the test was confirmed. Therefore, the variable *pTotal_Death* is normally distributed. After removing the extreme value (outlier) in row 89 of the dependent variable, the Shapiro-Wilk test (W = 0.9943, p-value (0.273) > 0.05) show no significant, thus, the null hypothesis of the test is maintained and the normality of the distribution is confirmed.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  pTotal_Death
D = 0.043551, p-value = 0.5725
alternative hypothesis: two-sided
```

```
#Test the normality of dependent variable having outlier;
shapiro.test(DS7006E$pTotal_Death)
#Remove outlier data points in dependent variable;
death.noOutlier <- pTotal_Death[-89]
#Test the normality of dependent variable with no outlier;
shapiro.test(death.noOutlier)
```

```
        Shapiro-Wilk normality test

data:  DS7006E$pTotal_Death
W = 0.95449, p-value = 1.84e-08
```

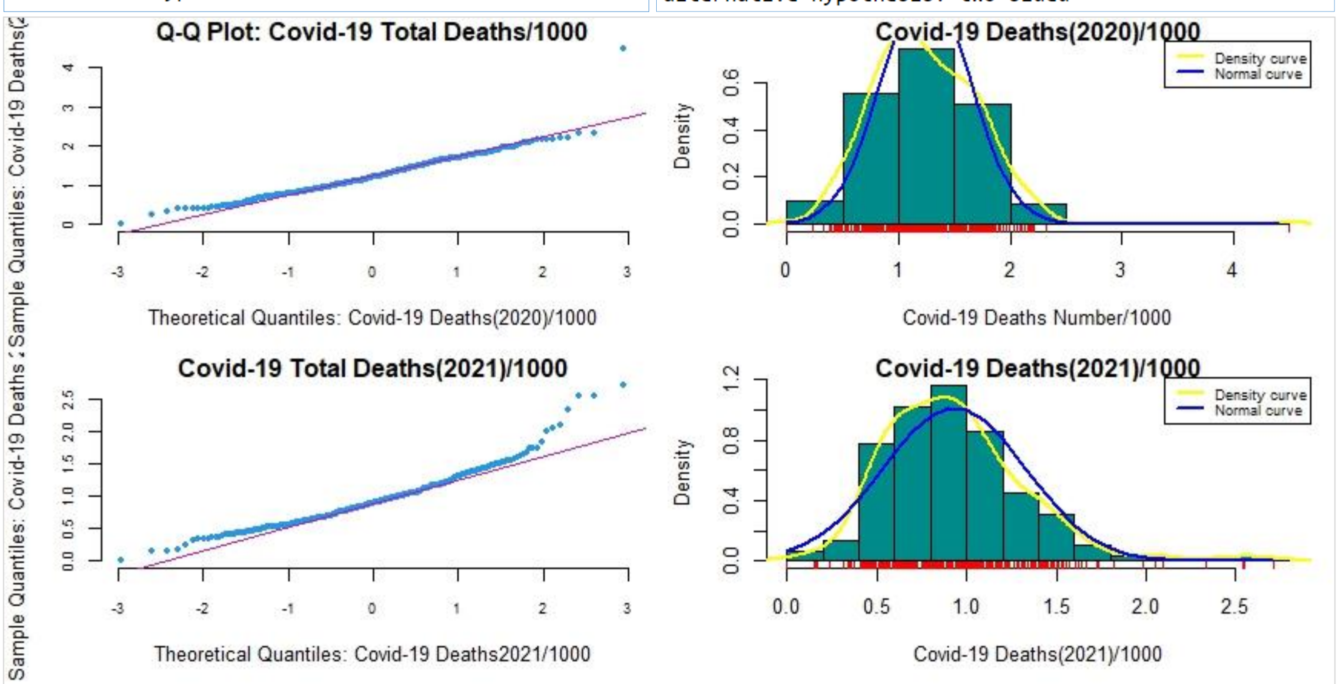
```
        Shapiro-Wilk normality test

data:  death.noOutlier
W = 0.9943, p-value = 0.273
```



The KS-Test on two other dependent variables, pTotal_Death2020 & pTotal_Death2021, confirm their normality.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  pTotal_Death2020
D = 0.046021, p-value = 0.5007
alternative hypothesis: two-sided
```

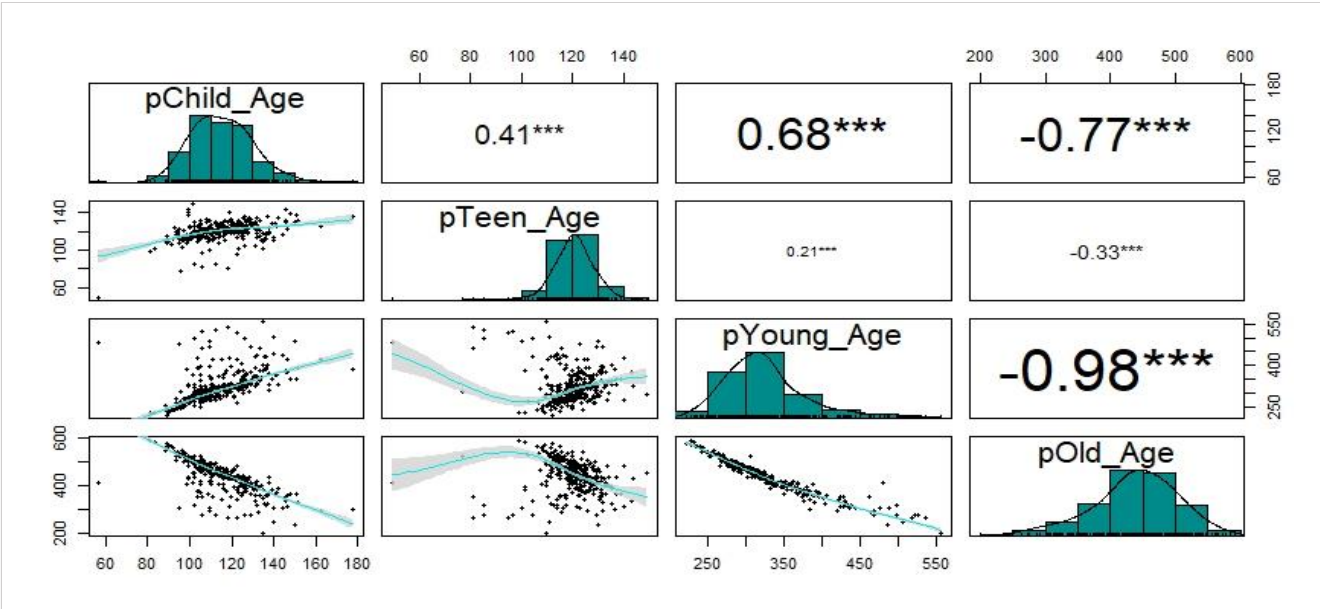
```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  pTotal_Death2021
D = 0.072115, p-value = 0.06949
alternative hypothesis: two-sided
```



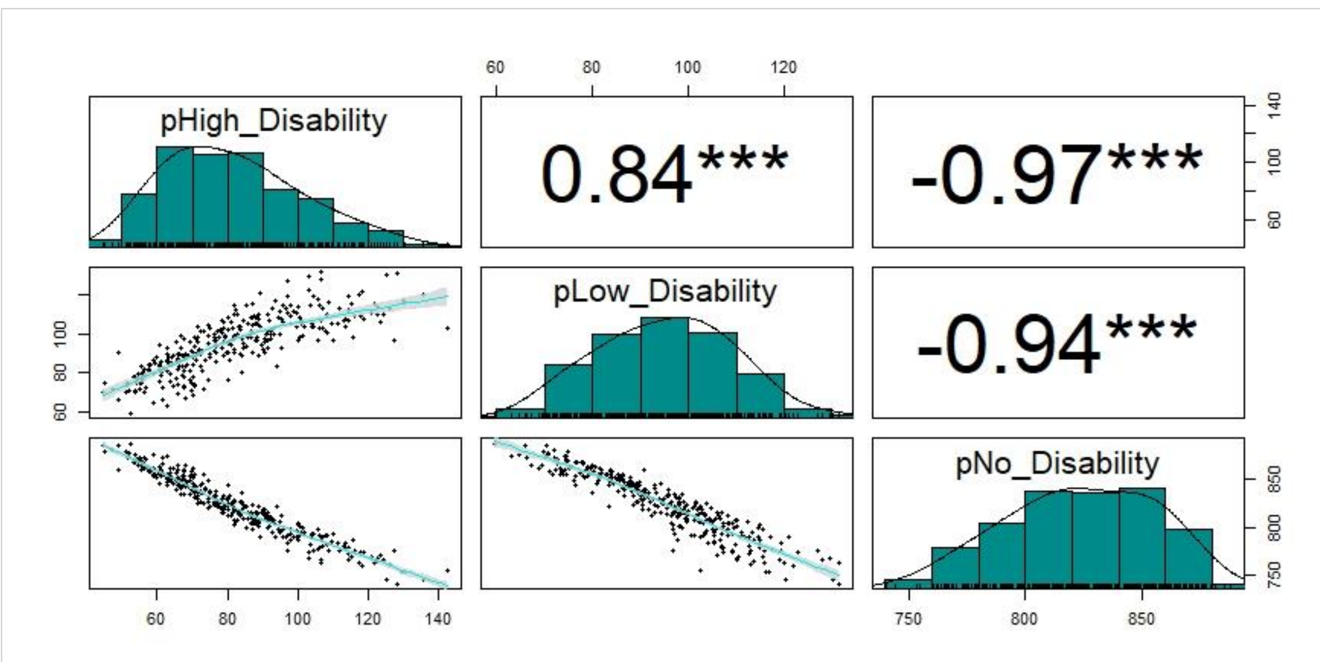
3.7 Collinearity Check for Independent Variables

To know about the collinearity between IVs, their correlations coefficients matrix of each variable need to be drawn; as illustrated bellow, it was done using the function ***pairs. panels ()*** in R. The results show that most of the variables in each theme are highly correlated; thus, there are collinearity between them, and need to be resolved by removing some of the variables.

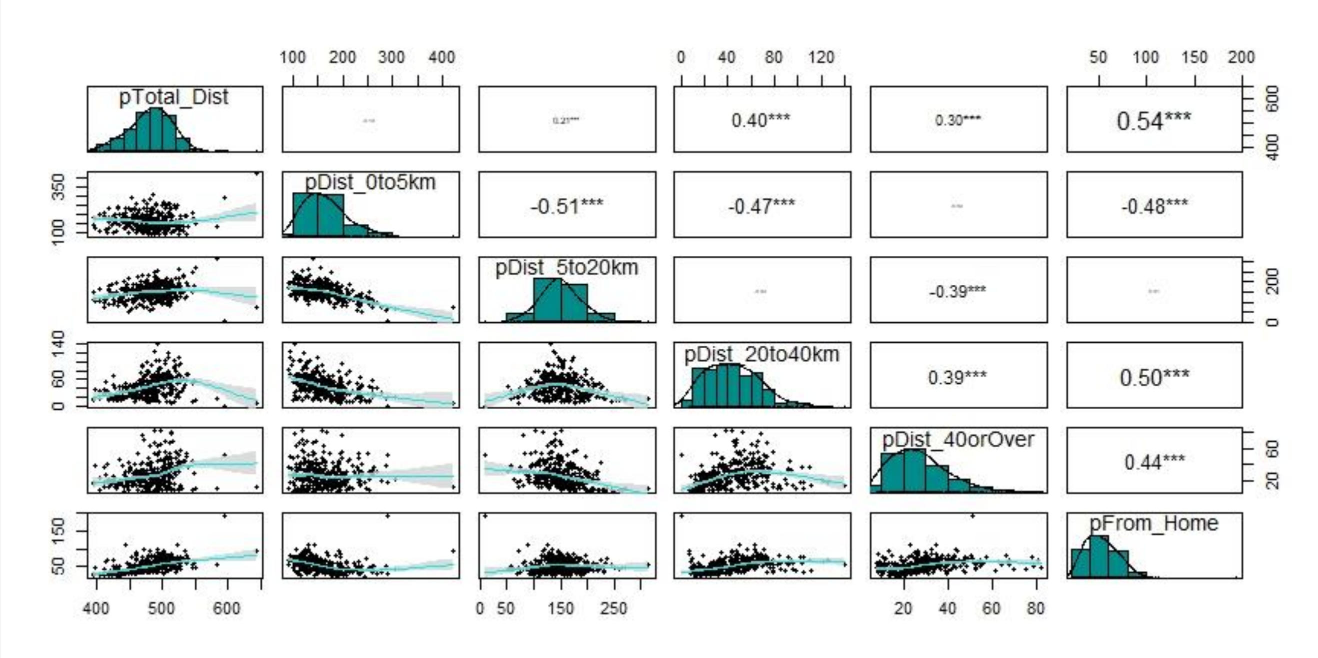
3.7.1 Age Groups Variables' Correlation Coefficients



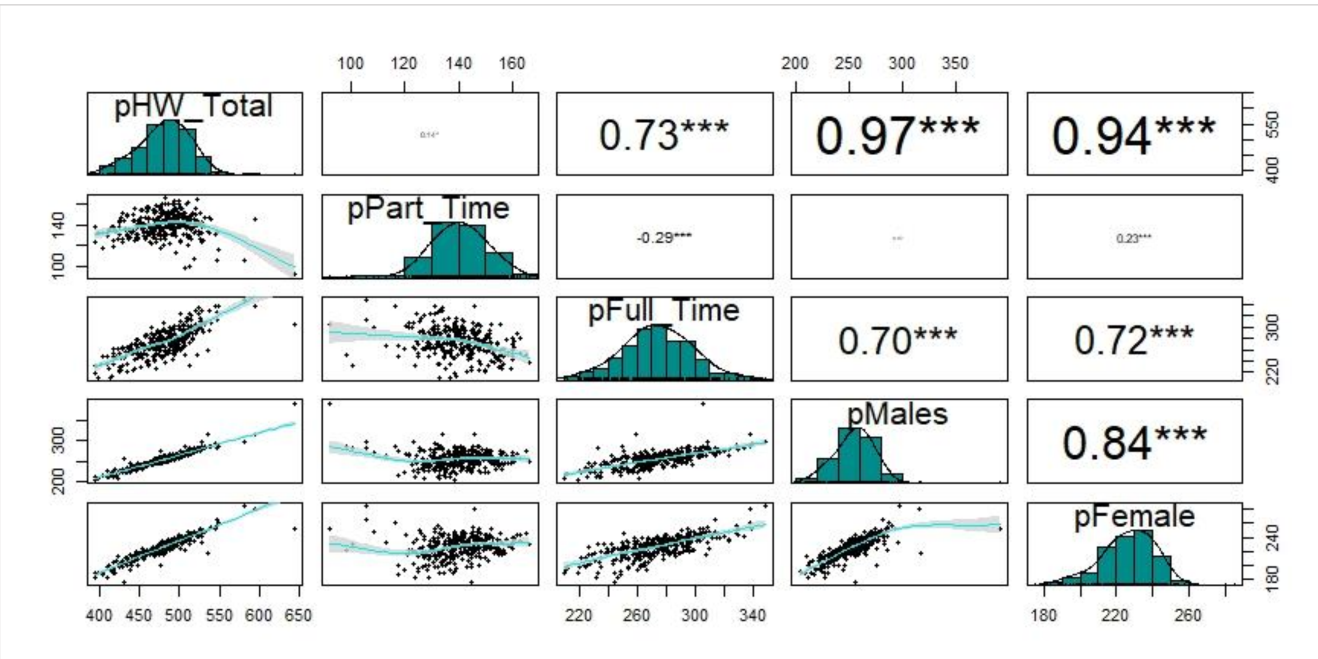
3.7.2 Disability Correlation Coefficients



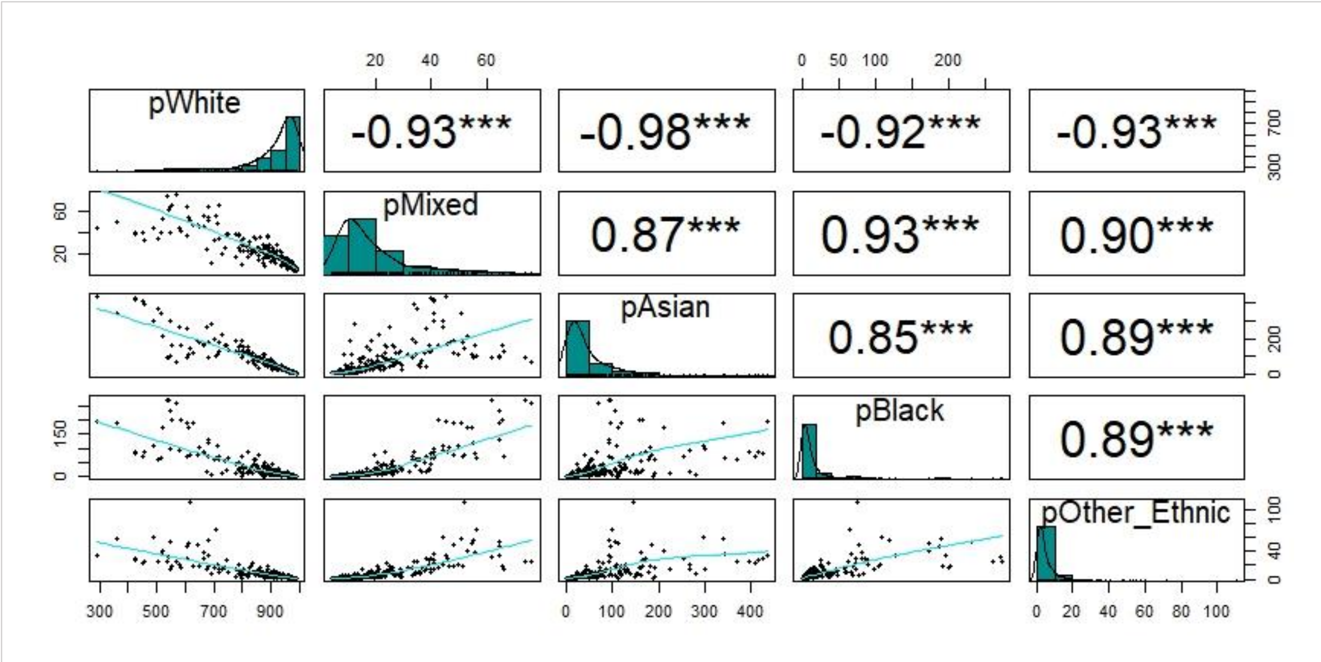
3.7.3 Distance Travel to Work



3.7.4 Hours Worked Correlations Coefficients



3.7.5 Ethnic Groups Correlation Coefficients



3.7.6 Resolving Collinearity within Themes

Collinearity should be resolved, otherwise, by ignoring it, most likely end up with a confusing statistical analysis in which nothing is significant (Zuur, et al., 2010). The threshold for considering a correlation between two independent variables are 0.5 – 0.7 (Dormann, et al., 2013). Here the 0.7 is considered as threshold. The following table show which variable in each them were selected to be removed:

Theme	Maintained	Removed due to collinearity	
Age	pChild_Age	pOld_Age	This variable has a strong correlation with pYoung_Age
	PTeen_Age		
	pYoung_Age		
Disability	pHigh_Disability	pLow_Disability	Strongly is correlated with pHigh_Disability
		pNo_Disability	Strongly are correlated with pHigh_Disability
Distance Travel to Work	pDist_0to5km		
	pDist_5to20km		
	pDist_20to40km		
	pDist_40orOver		
	pFromHome		
Hours Worked	pPart_Time	pHW_Total	Strongly is correlated with pMales and pFemales
	pFull_Time		
	pFemale	pMales	Strongly is correlated with pFemale
	pOver_Time		
Ethnicity Groups	pMixed	pWhite	Strongly are correlated with the pMixed
		pAsian	
		pBlack	
		pOther_Ethnic	

3.7.7 Collinearity Check Between All IVs

The collinearity is checked between all selected variables through creating their correlations coefficients matrix. The correlation matrix below shows that there some variables that highly correlated to each other. Therefore, collinearity exist between some of them.

	pChild_Age	pTeen_Age	pYoung_Age	pHigh_Disability	pDist_0to5km	pDist_5to20km	pDist_20to40km
pChild_Age	1.00000000	0.41078257	0.68327326	-0.29193116	0.13256144	0.16613251	-0.05849734
pTeen_Age	0.41078257	1.00000000	0.20897868	-0.02365355	0.04903919	-0.10000491	0.07295163
pYoung_Age	0.68327326	0.20897868	1.00000000	-0.24222249	0.48519744	0.04535315	-0.38834903
pHigh_Disability	-0.29193116	-0.02365355	-0.24222249	1.00000000	0.21121645	-0.15975040	-0.33074410
pDist_0to5km	0.13256144	0.04903919	0.48519744	0.21121645	1.00000000	-0.51201035	-0.46843843
pDist_5to20km	0.16613251	-0.10000491	0.04535315	-0.15975040	-0.51201035	1.00000000	-0.09386068
pDist_20to40km	-0.05849734	0.07295163	-0.38834903	-0.33074410	-0.46843843	-0.09386068	1.00000000
pDist_40orOver	-0.25837256	0.03717805	-0.37990650	-0.23841217	-0.08728414	-0.38677256	0.39451534
pFrom_Home	-0.44502632	-0.34236878	-0.55511719	-0.50245499	-0.48365657	0.01128425	0.50226055
pPart_Time	-0.58278974	-0.11162426	-0.60984694	-0.05314615	-0.17231983	-0.20574134	0.24857327
pFull_Time	0.44743430	-0.07267102	0.47018904	-0.51596311	0.16196822	0.32824032	0.13388864
pFemale	0.04167922	-0.22371216	0.04469258	-0.69410063	-0.08431066	0.24996955	0.36305883
pOver_Time	-0.17494005	-0.29961804	-0.25544075	-0.70685879	-0.44884163	0.12265745	0.52336795
pMixed	0.66811779	0.20790859	0.78005970	-0.47203473	0.20358548	0.10928251	-0.25793918

	pDist_40orOver	pFrom_Home	pPart_Time	pFull_Time	pFemale	pOver_Time	pMixed
pChild_Age	-0.25837256	-0.44502632	-0.58278974	0.44743430	0.04167922	-0.17494005	0.66811779
pTeen_Age	0.03717805	-0.34236878	-0.11162426	-0.07267102	-0.22371216	-0.29961804	0.20790859
pYoung_Age	-0.37990650	-0.55511719	-0.60984694	0.47018904	0.04469258	-0.25544075	0.78005970
pHigh_Disability	-0.23841217	-0.50245499	-0.05314615	-0.51596311	-0.69410063	-0.70685879	-0.47203473
pDist_0to5km	-0.08728414	-0.48365657	-0.17231983	0.16196822	-0.08431066	-0.44884163	0.20358548
pDist_5to20km	-0.38677256	0.01128425	-0.20574134	0.32824032	0.24996955	0.12265745	0.10928251
pDist_20to40km	0.39451534	0.50226055	0.24857327	0.13388864	0.36305883	0.52336795	-0.25793918
pDist_40orOver	1.00000000	0.43909290	0.44223481	0.01334645	0.25630145	0.36387965	-0.28392415
pFrom_Home	0.43909290	1.00000000	0.53702386	-0.05746926	0.50125349	0.84097962	-0.28384901
pPart_Time	0.44223481	0.53702386	1.00000000	-0.29225023	0.22787004	0.23580192	-0.49084490
pFull_Time	0.01334645	-0.05746926	-0.29225023	1.00000000	0.71870711	0.20035952	0.37769723
pFemale	0.25630145	0.50125349	0.22787004	0.71870711	1.00000000	0.63970205	0.11603141
pOver_Time	0.36387965	0.84097962	0.23580192	0.20035952	0.63970205	1.00000000	-0.03638999
pMixed	-0.28392415	-0.28384901	-0.49084490	0.37769723	0.11603141	-0.03638999	1.00000000

After removing variables such *pFull_Time*, *pOver_Time*, and *pYoung_Age*, the calculated correlation matrix of the survived independent variables, as depicted below, looks free of collinearity:

	pChild_Age	pTeen_Age	pHigh_Disability	pDist_0to5km	pDist_5to20km	pDist_20to40km	pDist_40orOver
pChild_Age	1.00000000	0.41078257	-0.29193116	0.13256144	0.16613251	-0.05849734	-0.25837256
pTeen_Age	0.41078257	1.00000000	-0.02365355	0.04903919	-0.10000491	0.07295163	0.03717805
pHigh_Disability	-0.29193116	-0.02365355	1.00000000	0.21121645	-0.15975040	-0.33074410	-0.23841217
pDist_0to5km	0.13256144	0.04903919	0.21121645	1.00000000	-0.51201035	-0.46843843	-0.08728414
pDist_5to20km	0.16613251	-0.10000491	-0.15975040	-0.51201035	1.00000000	-0.09386068	-0.38677256
pDist_20to40km	-0.05849734	0.07295163	-0.33074410	-0.46843843	-0.09386068	1.00000000	0.39451534
pDist_40orOver	-0.25837256	0.03717805	-0.23841217	-0.08728414	-0.38677256	0.39451534	1.00000000
pFrom_Home	-0.44502632	-0.34236878	-0.50245499	-0.48365657	0.01128425	0.50226055	0.43909290
pPart_Time	-0.58278974	-0.11162426	-0.05314615	-0.17231983	-0.20574134	0.24857327	0.44223481
pFemale	0.04167922	-0.22371216	-0.69410063	-0.08431066	0.24996955	0.36305883	0.25630145
pMixed	0.66811779	0.20790859	-0.47203473	0.20358548	0.10928251	-0.25793918	-0.28392415

	pFrom_Home	pPart_Time	pFemale	pMixed
pChild_Age	-0.44502632	-0.58278974	0.04167922	0.6681178
pTeen_Age	-0.34236878	-0.11162426	-0.22371216	0.2079086
pHigh_Disability	-0.50245499	-0.05314615	-0.69410063	-0.4720347
pDist_0to5km	-0.48365657	-0.17231983	-0.08431066	0.2035855
pDist_5to20km	0.01128425	-0.20574134	0.24996955	0.1092825
pDist_20to40km	0.50226055	0.24857327	0.36305883	-0.2579392
pDist_40orOver	0.43909290	0.44223481	0.25630145	-0.2839241
pFrom_Home	1.00000000	0.53702386	0.50125349	-0.2838490
pPart_Time	0.53702386	1.00000000	0.22787004	-0.4908449
pFemale	0.50125349	0.22787004	1.00000000	0.1160314
pMixed	-0.28384901	-0.49084490	0.11603141	1.0000000

3.8 Dependent and Independent Correlation Matrix

The correlations coefficients matrix of DV and IVs illustrate that there are weak correlations between DV and IVs.

	pTotal_Death	pChild_Age	pTeen_Age	pHigh_Disability	pDist_0to5km	pDist_5to20km	pDist_20to40km	pDist_40orOver	pFrom_Home	pPart_Time	pFemale	pMixed
pTotal_Death	1.00000000	0.15881242	0.21882356	0.37850416	-0.07343950	0.07819528	-0.01029535	-0.21854722	-0.42338049	-0.32414512	-0.35346180	0.01778138
pChild_Age	0.15881242	1.00000000	0.41078257	-0.29193116	0.13256144	0.16613251	-0.05849734	-0.25837256	-0.44502632	-0.58278974	0.04167922	0.66811779
pTeen_Age	0.21882356	0.41078257	1.00000000	-0.02365355	0.04903919	-0.10000491	0.07295163	0.03717805	-0.34236878	-0.11162426	-0.22371216	0.20790859
pHigh_Disability	0.37850416	-0.29193116	-0.02365355	1.00000000	0.21121645	-0.15975040	-0.33074410	-0.23841217	-0.50245499	-0.05314615	-0.69410063	-0.47203473
pDist_0to5km	-0.07343950	0.13256144	0.04903919	0.21121645	1.00000000	-0.51201035	-0.46843843	-0.08728414	-0.48365657	-0.17231983	-0.08431066	0.20358548
pDist_5to20km	0.07819528	0.16613251	-0.10000491	-0.15975040	-0.51201035	1.00000000	-0.09386068	-0.38677256	0.01128425	-0.20574134	0.24996955	0.10928251
pDist_20to40km	-0.01029535	-0.05849734	0.07295163	-0.33074410	-0.46843843	-0.09386068	1.00000000	0.39451534	0.50226055	0.24857327	0.36305883	-0.25793918
pDist_40orOver	-0.21854722	-0.25837256	0.03717805	-0.23841217	-0.08728414	-0.38677256	0.39451534	1.00000000	0.43909290	0.44223481	0.25630145	-0.28392415
pFrom_Home	-0.42338049	-0.44502632	-0.34236878	-0.50245499	-0.48365657	0.01128425	0.50226055	0.43909290	1.00000000	0.53702386	0.50125349	-0.28384901
pPart_Time	-0.32414512	-0.58278974	-0.11162426	-0.05314615	-0.17231983	-0.20574134	0.24857327	0.44223481	0.53702386	1.00000000	0.22787004	-0.49084490
pFemale	-0.35346180	0.04167922	-0.22371216	-0.69410063	-0.08431066	0.24996955	0.36305883	0.25630145	0.50125349	0.22787004	1.00000000	0.11603141
pMixed	0.01778138	0.66811779	0.20790859	-0.47203473	0.20358548	0.10928251	-0.25793918	-0.28392415	-0.28384901	-0.49084490	0.11603141	1.00000000

3.8.1 DV and IVs Correlation Test

To know further about the significance of the correlation between IVs and DV, the correlation tests, using spearman method, were performed between each pair of DV and IVs, considering the following hypothesis:

- $\boldsymbol{H_0}$: the correlation coefficient between DV and IV is not significantly different from zero;
- $\boldsymbol{H_a}$: the correlation coefficient is significantly different from zero;

The following table illustrates the result of the correlation tests were performed:

Theme	IV	DV (pTotal_Death)					
		P-Value	ρ	Method	T- Result	$\boldsymbol{H_0}$	Correlation
Age	pChild_Age	0.0042	0.1588	Spearman	No Significant	Rejected	Positive weak
	pTeen_Age	7.612e-05	0.2188	Spearman	No Significant	Rejected	Positive moderate
Disability	pHigh_Disability	2.671e-12	0.378	Spearman	No Significant	Rejected	Positive moderate
Distance Travel to Work	pDist_0to5	0.1879	-0.073	Spearman	Significant	Maintained	No correlation
	pDist_5t020	0.1608	0.078	Spearman	Significant	Maintained	No correlation
	pDist_20to40	0.8537	-0.0102	Spearman	Significant	Maintained	No correlation
	pDist_40orOver	7.776e-05	-0.2185	Spearman	No significant	Rejected	Negative moderate
	pFrom_Home	2.2e-16	-0.4233	Spearman	No significant	Rejected	Negative moderate
Hours Worked	pPart_Time	3.031e-09	-0.3241	Spearman	No significant	Rejected	Negative moderate
	pFemale	8.256e-11	-0.3534	Spearman	No significant	Rejected	Negative moderate
Ethnicity	pMixed	0.7501	0.0177	Spearman	Significant	Maintained	No correlation

3.8.2 DV and IVs Partial Correlation

The spearman correlation coefficients test result reveals that some of the IVs have no correlations with the DV. However, sometime it happens that the correlation between IV with DV is influenced by the other IV in the group. To verify this, a partial correlation coefficient matrix of DV and IVs was calculated using the ***pcor()*** function. The matrix, shown below, confirms the existing of correlation of those IVs with the DVs which had previously been rejected by the correlation test. The partial correlations test p-values depicted in the following also reject the null hypothesis that denies the correlation between DV and the IVs.

	pTotal_Death	pChild_Age	pTeen_Age	pHigh_Disability	pDist_0to5km	pDist_5to20km	pDist_20to40km	pDist_40orOver	pFrom_Home	pPart_Time	pFemale	pMixed
pTotal_Death	1.000000000	-0.06200052	0.09995586	0.23948796	-0.19129993	-0.03448019	0.15129647	0.003093905	-0.194950474	-0.16969889	0.06061679	0.12258312
pChild_Age	-0.062000522	1.00000000	0.25006049	-0.18265704	-0.02008199	0.04575544	0.18624129	-0.019134368	-0.334285680	-0.29962585	0.04697055	0.27799558
pTeen_Age	0.099955862	0.25006049	1.00000000	-0.21869164	0.04321537	0.05705185	0.19085615	0.148254693	-0.272025158	0.28534848	-0.29194275	-0.02957914
pHigh_Disability	0.239487956	-0.18265704	-0.21869164	1.00000000	0.11353888	0.01970239	-0.04425020	-0.150849267	-0.436502219	0.06484760	-0.44343421	-0.58655640
pDist_0to5km	-0.191299929	-0.02008199	0.04321537	0.11353888	1.00000000	-0.78348480	-0.54368015	-0.184012818	-0.413985144	-0.17641653	0.61340532	-0.05152752
pDist_5to20km	-0.034480188	0.04575544	0.05705185	0.01970239	-0.78348480	1.00000000	-0.54176105	-0.408531289	-0.256850404	-0.22301989	0.61717725	-0.18817378
pDist_20to40km	0.151296468	0.18624129	0.19085615	-0.04425020	-0.54368015	-0.54176105	1.00000000	-0.077013690	0.039736703	-0.17228803	0.45680411	-0.34018490
pDist_40orOver	0.003093905	-0.01913437	0.14825469	-0.15084927	-0.18401282	-0.40853129	-0.07701369	1.000000000	-0.009636737	0.05491317	0.19866535	-0.23870779
pFrom_Home	-0.194950474	-0.33428568	-0.27202516	-0.43650222	-0.41398514	-0.25685040	0.03973670	-0.009636737	1.000000000	0.11709247	0.14119122	-0.18655244
pPart_Time	-0.169698888	-0.29962585	0.28534848	0.06484760	-0.17641653	-0.22301989	-0.17228803	0.054913173	0.117092474	1.00000000	0.26714162	-0.16087376
pFemale	0.060616788	0.04697055	-0.29194275	-0.44343421	0.61340532	0.61717725	0.45680411	0.198665350	0.141191217	0.26714162	1.00000000	-0.04042421
pMixed	0.122583119	0.27799558	-0.02957914	-0.58655640	-0.05152752	-0.18817378	-0.34018490	-0.238707793	-0.186552438	-0.16087376	-0.04042421	1.00000000

Partial Correlation Test P-Value:

	pTotal_Death	pChild_Age	pTeen_Age	pHigh_Disability	pDist_0to5km	pDist_5to20km	pDist_20to40km	pDist_40orOver	pFrom_Home	pPart_Time	pFemale	pMixed
pTotal_Death	0.000000e+00	2.741458e-01	7.743797e-02	1.848015e-05	6.680276e-04	5.433487e-01	7.330360e-03	9.565225e-01	5.232465e-04	2.594092e-03	2.850187e-01	3.014171e-02
pChild_Age	2.741458e-01	0.000000e+00	7.546605e-06	1.170580e-03	7.234132e-01	4.198508e-01	9.303806e-04	7.359677e-01	1.312641e-09	6.503264e-08	4.075990e-01	5.810371e-07
pTeen_Age	7.743797e-02	7.546605e-06	0.000000e+00	9.583905e-05	4.461465e-01	3.143522e-01	6.879567e-04	8.615873e-03	1.030084e-06	2.817077e-07	1.445693e-07	6.021377e-01
pHigh_Disability	1.848015e-05	1.170580e-03	9.583905e-05	0.000000e+00	4.473199e-02	7.284335e-01	4.353226e-01	7.507920e-03	5.424810e-16	2.526712e-01	1.647917e-16	2.557745e-30
pDist_0to5km	6.680276e-04	7.234132e-01	4.461465e-01	4.473199e-02	0.000000e+00	3.170660e-66	1.793371e-25	1.073709e-03	2.166318e-14	1.728821e-03	9.718344e-34	3.635764e-01
pDist_5to20km	5.433487e-01	4.198508e-01	3.143522e-01	7.284335e-01	3.170660e-66	0.000000e+00	2.848293e-25	5.080376e-14	4.155957e-06	6.889216e-05	3.023660e-34	8.205978e-04
pDist_20to40km	7.330360e-03	9.303806e-04	6.879567e-04	4.353226e-01	1.793371e-25	2.848293e-25	0.000000e+00	1.741215e-01	4.836288e-01	2.222287e-03	1.531017e-17	6.423091e-10
pDist_40orOver	9.565225e-01	7.359677e-01	8.615873e-03	7.507920e-03	1.073709e-03	5.080376e-14	1.741215e-01	0.000000e+00	8.651571e-01	3.328661e-01	4.062411e-04	1.971137e-05
pFrom_Home	5.232465e-04	1.312641e-09	1.030084e-06	5.424810e-16	2.166318e-14	4.155957e-06	4.836288e-01	8.651571e-01	0.000000e+00	3.841339e-02	1.240280e-02	9.118356e-04
pPart_Time	2.594092e-03	6.503264e-08	2.817077e-07	2.526712e-01	1.728821e-03	6.889216e-05	2.222287e-03	3.328661e-01	3.841339e-02	0.000000e+00	1.628964e-06	4.325851e-03
pFemale	2.850187e-01	4.075990e-01	1.445693e-07	1.647917e-16	9.718344e-34	3.023660e-34	1.531017e-17	4.062411e-04	1.240280e-02	1.628964e-06	0.000000e+00	4.760889e-01
pMixed	3.014171e-02	5.810371e-07	6.021377e-01	2.557745e-30	3.635764e-01	8.205978e-04	6.423091e-10	1.971137e-05	9.118356e-04	4.325851e-03	4.760889e-01	0.000000e+00

3.9 Selected IVs' Normality Check

In addition to the dependent variable, the normality of dependent variables was also were checked using the Kolmogorov-Smirnov Test (KS-Test) through defining the following hypothesis:

- H_0: the sample follows normal distribution;
- H_a: the sample does not follow normal distribution;

The table below, summarizes the result of the normality test for each variable:

Theme	Variable	Test	Test Value	P-Value	Statically Significant?	H0	Normality
Age	pChild_Age	KS-Test	0.057354	0.2385	No	Confirmed	Yes
	pTeen_Age	KS-Test	0.098663	0.003716	Yes	Rejected	No
Disability	pHigh_Disability	KS-Test	0.072778	0.06531	No	Confirmed	Yes
Distance Travel to Work	pDist_0to5	KS-Test	0.10715	0.001203	Yes	Rejected	No
	pDist_5t020	KS-Test	0.044909	0.5326	No	Confirmed	Yes
	pDist_20to40	KS-Test	0.055416	0.2744	No	Confirmed	Yes
	pDist_40orOver	KS-Test	0.69406	2.2e-16	Yes	Rejected	No
	pFrom_Home	KS-Test	0.074666	0.05456	No	Confirmed	Yes
Hours Worked	pPart_Time	KS-Test	0.039422	0.697	No	Confirmed	Yes
	pFemale	KS-Test	0.046843	0.4778	No	Confirmed	Yes
Ethnicity	pMixed	KS-Test	0.1717	1.071e-08	Yes	Rejected	No

4 Data Analysis

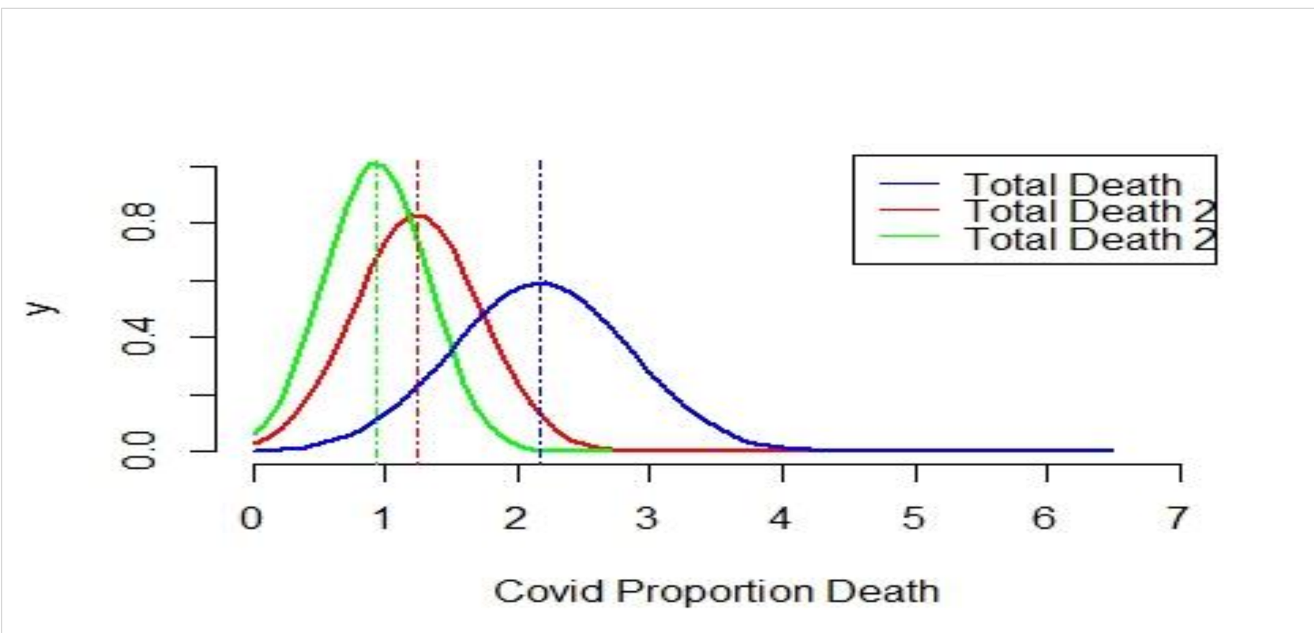
So far, we have a number of selected IVs from different selected themes, and DVs from the Covid-19 deaths. The Data Exploration phase provided some useful information about the associations of IVs and DVs, and also about the distribution of each variable. In this part, a number of statistical process are performed on the selected variables. These process include hypothesis test of some DVs and IVs, as well as Clustering and Factor Analysis.

4.1 Hypothesis Testing

The parametric T-Test used to do hypothesis test on all the dependent variables, considering the statistical characteristics of the sample variable. The statistical characteristics and the associated test results are summarized in the table below:

Variable	Data Type	Type	Variance	Normality	Test	Test Value	P-Value	H_0
pTotal_Death	Ratio	Paired	0.4565492	Yes	Parametric T-Test	42.515	2.2e-16	Rejected
pTotal_Death2020			0.2312726	Yes				
pTotal_Death	Ratio	Paired	0.4565492	Yes	Parametric (T-Test)	46.411	2.2e-16	Rejected
pTotal_Death2021			0.1544893	Yes				
pTotal_Death2020	Ratio	paired	0.2312726	Yes	Parametric (T-Test)	9.994	2.2e-16	Rejected
pTotal_Death2021			0.1544893	Yes				

The distributions graph of the dependent variables is also visually confirming the different between means of the samples:



4.1.1 DV(pTotal_Death) and IVs (pChild_Age, pHigh_Disability, pMixed):

In addition to dependent variables, hypothesis tests were performed on a number of independent variables which the summary of the test results are depicted in the below table:

Variable	Data Type	Type	Variance	Normality	Appropriate Test
pTotal_Death	Ratio	Independent		Yes	Parametric (T-Test)
pChild_Age				Yes	
pHigh_Disability	Ratio	independent		Yes	Wilcoxon-Test
pTeen_Age				No	
pTotal_Death	Ratio	independent		Yes	Wilcoxon-Test
pMixed				No	

Test Results:

Samples	Test	Test Value	p-value	H_0	
pTotal_Death	T-Test	-139.25	2.2e-16	Rejected	
pChild_Age					
pHigh_Disability	Wilcoxon-Test	98435	2.2e-16	Rejected	
pTeen_Age					
pTotal_Death	Wilcoxon -Test	104307	2.2e-16	Rejected	
pMixed					

4.2 Factor Analysis

The main goal of factor analysis is to reduce the number of variables or dimension of measurement by identifying a common structure associating a number of variables in the data set (Fabrigar & Duane , 2012). The first step in factor analysis process is to examine whether the variables included in the analysis are suitable for factor analysis. To do so, the Kaiser-Meyer-Olkin (KMO) statistic is used to measure the adequacy of sample for factory analysis.

The overall measure of sample adequacy (MSA = 0.26) shows that the final selected IVs are not adequately supporting factory analysis. It is predictable, because, the selected IVs come through collinearity resolution process which drastically reduced their coviances.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(final.IVs))
Overall MSA =  0.26
MSA for each item = 
      pChild_Age         pTeen_Age pHigh_Disability      pDist_0to5km     pDist_5to20km 
            0.72              0.52             0.33              0.14              0.15 
  pDist_20to40km    pDist_40orOver        pFrom_Home        pPart_Time           pFemale 
            0.16              0.19             0.31              0.39              0.24 
          pMixed 
            0.37 
```

4.2.1 Factory Analysis on All IVs

As shown above, the selected variables are not factorable. However, there is no choice, unless all the variables from the data set were chosen to test their factor abilities; The KMO overall measure of sample adequacy (MSA = 0.5) shows that the overall sample variables are relatively suitable for defining factor analysis.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(total.IVs))
Overall MSA =  0.5
MSA for each item =
      pChild_Age        pTeen_Age       pYoung_Age         pOld_Age pHigh_Disability  pLow_Disability
             0.5              0.5              0.5              0.5              0.5              0.5
  pNo_Disability      pTotal_Dist     pDist_0to5km    pDist_5to20km   pDist_20to40km   pDist_40orOver
             0.5              0.5              0.5              0.5              0.5              0.5
      pFrom_Home        pHW_Total       pPart_Time       pFull_Time       pOver_Time           pMales
             0.5              0.5              0.5              0.5              0.5              0.5
         pFemale           pWhite           pMixed           pAsian           pBlack     pOther_Ethnic
             0.5              0.5              0.5              0.5              0.5              0.5
```

4.3 Defining Number of Factors

The two commonly used technique such as the Eigenvalue greater than one rule, and Scree plot (Fabrigar & Duane , 2012), were used to determine the number of appropriate factors. In addition to these two methods, the Cumulative Proportion Eigenvalue plot also used as complementary tools for identifying the number of factors.

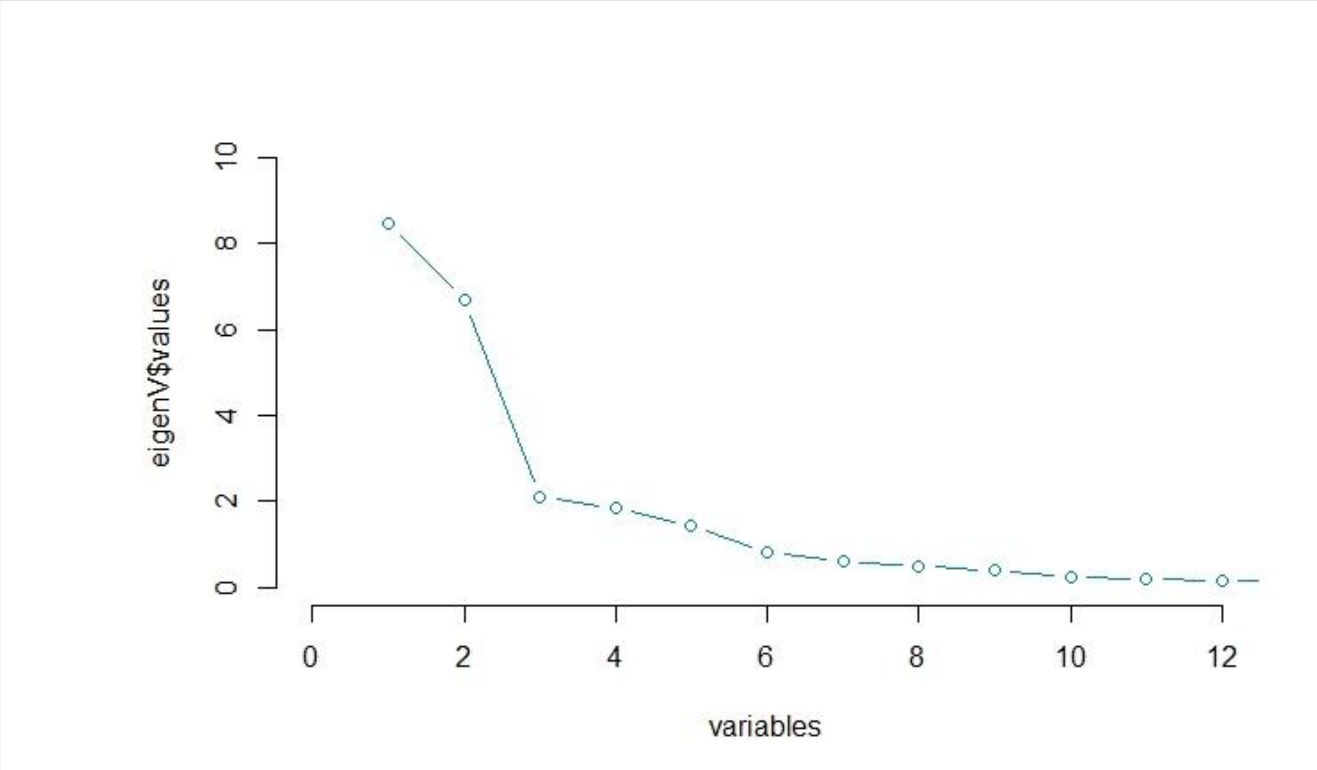
Eigenvalues: the following eigenvalues calculated from sample show that five of them have values greater than one. The value (0.804) is close to one, but it was ignored. Therefore, based on this procedure, the factor which is needed for the sample is 5.

```
all.IVs <- DS7006E[,c("pChild_Age","pTeen_Age","pYoung_Age","pOld_Age","pHigh_Disability",
                       "pLow_Disability","pNo_Disability","pTotal_Dist","pDist_0to5km",
                       "pDist_5to20km","pDist_20to40km","pDist_40orOver","pFrom_Home","pHW_Total",
                       "pPart_Time","pFull_Time","pOver_Time","pMales","pFemale","pWhite","pMixed",
                       "pAsian","pBlack","pOther_Ethnic")]

KMO(cor(all.IVs))
# get eigenvalues: eigen() uses a correlation matrix
eigenV <- eigen(cor(all.IVs))
round(eigenV$values,3)
```

```
8.466 6.705 2.110 1.850 1.455 0.804 0.614 0.499 0.391 0.256 0.207 0.171 0.160 0.137 0.080 0.059 0.030 0.005
0.000 0.000 0.000 0.000 0.000 0.000
```

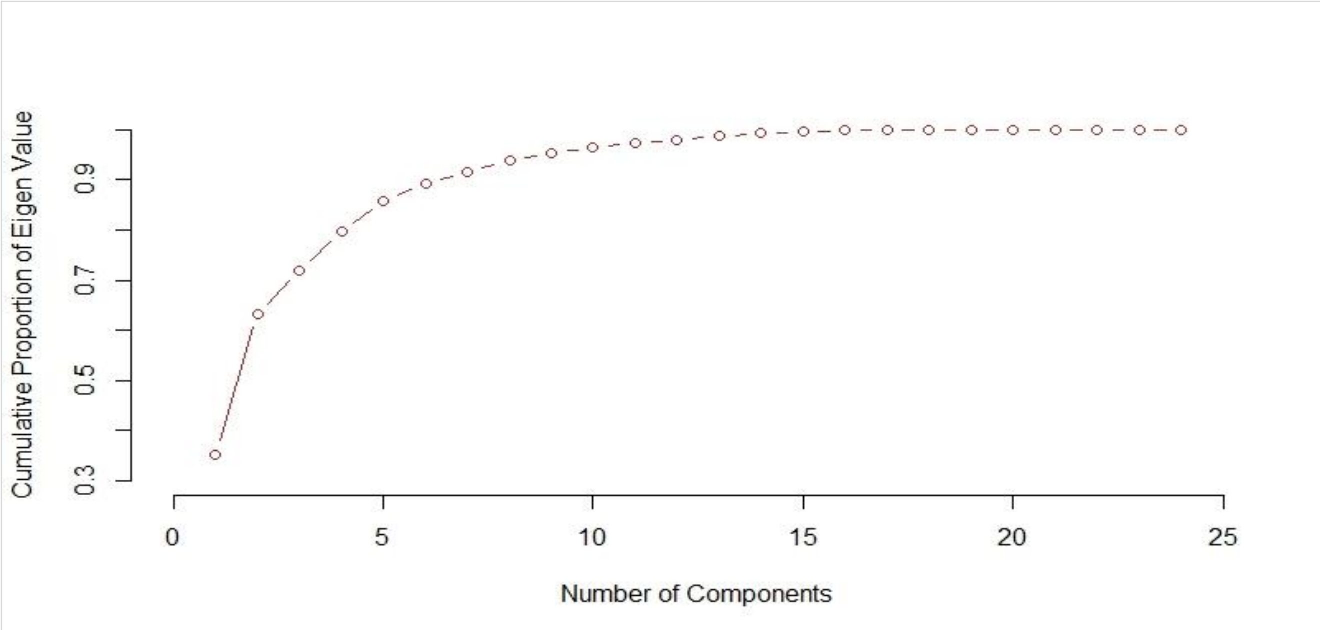
Eigenvalue Scree Plot: from the scree plot of the eigenvalue, it can be concluded that the best decision would be to define 5 factors for the sample.

```
# plot a scree plot of eigenvalues
op <- par(mar = c(5, 8, 4, 2) + 0.1)
plot(eigenV$values, type="b", frame.plot = FALSE, col="cyan4",xlim = c(0,12),xlab="variables", ylim = c(0,10))
```



Cumulative Proportion of Eigenvalue: the cumulative proportion of eigenvalues indicate that the 5 factors are able to account for nearly 90% of the sample covariance. Therefore, the 5 factors explain above the 85% of the sample variances, and it is acceptable.

```
# calculate cumulative proportion of eigenvalue and plot
eigenV.sum<-0
for(i in 1:length(eigenV$value)){
  eigenV.sum<-eigenV.sum+eigenV$value[i]
}
eigenPv.List1<-1:length(eigenV$value)
for(i in 1:length(eigenV$value)){
  eigenPv.List1[i]=eigenV$value[i]/eigenV.sum
}
eigenCv.List2<-1:length(eigenV$value)
eigenCv.List2[1]<-eigenPv.List1[1]
for(i in 2:length(eigenV$value)){
  eigenCv.List2[i]=eigenCv.List2[i-1]+eigenPv.List1[i]
}
eigenCv.List2

plot (eigenCv.List2, type="b", col="brown",xlim = c(0,25),ylim = c(0.3,1), xlab="Number of Components",
      frame.plot = FALSE,ylab ="Cumulative Proportion of Eigen Value")

# principal() uses a data frame or matrix of correlations
PCA <- principal(total.IVs, nfactors=5, rotate="varimax")
```



4.4 Principal Component Analysis (PCA)

Through the PCA all the variables loaded into 5 factors as indicated below. The loading values are assessed based on the 0.5 threshold, and those variables having above the 0.5 loading values are captured by the factors and considered significant as the table illustrate it.

```
# principal() uses a data frame or matrix of correlations
PCA <- principal(all.IVs, nfactors=5, rotate="varimax")
```

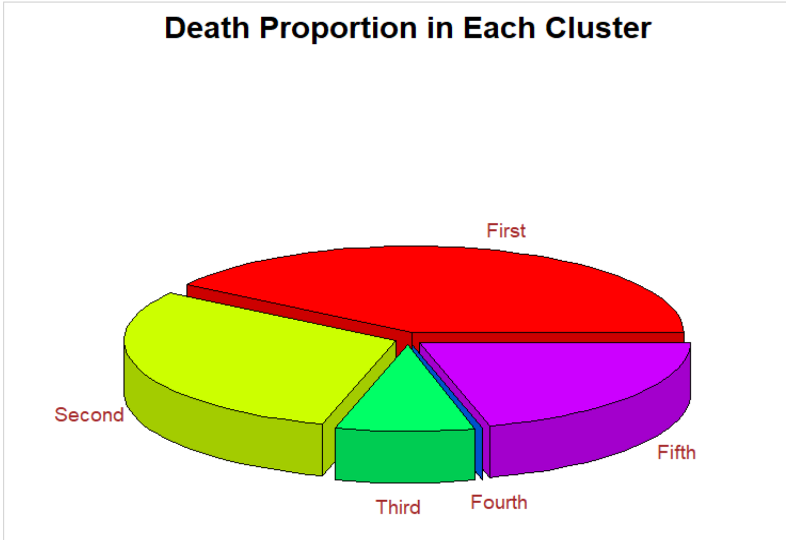
Factor	Theme	IV	Loading	
RC1	Age Groups	pYoung_Age	0.83	
		pOld_Age	-0.85	
	Distance to Work	pPart_Time	-0.60	
	Ethnic Groups	pWhite	-0.96	
		pMixed	0.92	
		pAsian	0.83	
		pBlack	0.84	
		pOther_Ethnic	0.86	
RC2	Disability	pHigh_Disability	-0.78	
		pLow_Disability	-0.67	
		pNo_Disability	0.77	
	Distance to Work	pTotal_Distance	0.97	
	Hours Worked	pHW_Total	0.97	
		pFull_Time	0.84	
		pMales	0.92	
		pFemale	0.92	
RC3	Age Groups	pChild_Age	-0.66	
		pTeen_Age	-0.80	
	Distance to Work	pFrom_Home	0.61	
	Hours Worked	pOver_Time	0.67	

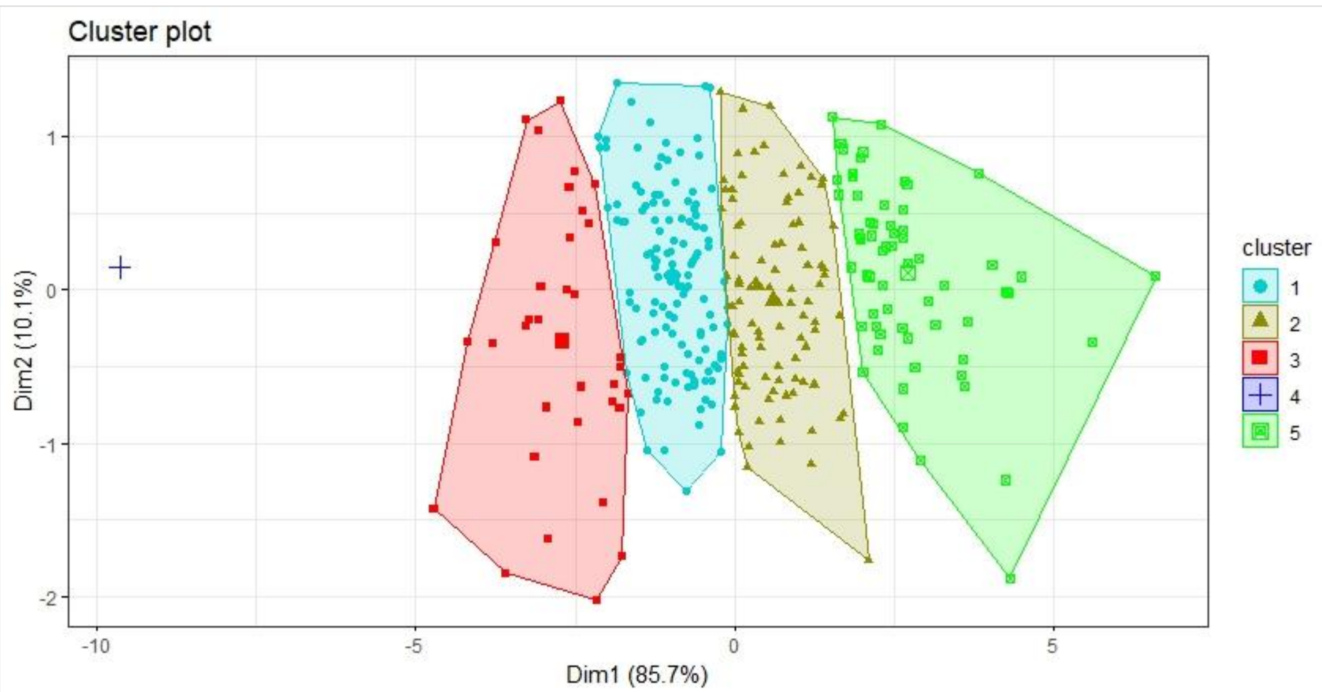
```
Principal Components Analysis
Call: principal(r = total.IVs, nfactors = 5, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                  RC1   RC2   RC3   RC4   RC5   h2    u2 com
pChild_Age       0.61  0.12 -0.66  0.07 -0.04 0.83 0.173 2.1
pTeen_Age       -0.02 -0.28 -0.80  0.20  0.30 0.85 0.151 1.7
pYoung_Age       0.83  0.22 -0.07 -0.39 -0.17 0.92 0.077 1.7
pOld_Age        -0.85 -0.18  0.32  0.29  0.11 0.95 0.049 1.7
pHigh_Disability -0.33 -0.78 -0.04 -0.25 -0.39 0.93 0.072 2.1
pLow_Disability -0.67 -0.67  0.16 -0.07 -0.12 0.94 0.058 2.2
pNo_Disability   0.50  0.77 -0.04  0.18  0.29 0.96 0.040 2.2
pTotal_Dist     -0.03  0.97  0.20  0.01  0.01 0.99 0.010 1.1
pDist_0to5km     0.11  0.04  0.00 -0.97  0.06 0.95 0.050 1.0
pDist_5to20km    0.24  0.23 -0.03  0.58 -0.67 0.90 0.096 2.5
pDist_20to40km  -0.34  0.36 -0.16  0.47  0.30 0.58 0.415 3.9
pDist_40orOver  -0.37  0.29 -0.07 -0.01  0.60 0.59 0.409 2.2
pFrom_Home      -0.27  0.38  0.61  0.32  0.41 0.86 0.138 3.6
pHW_Total       -0.03  0.97  0.20  0.01  0.01 0.99 0.010 1.1
pPart_Time      -0.60 -0.04  0.02  0.21  0.40 0.56 0.442 2.0
pFull_Time       0.13  0.84 -0.30 -0.19 -0.32 0.95 0.054 1.8
pOver_Time       0.11  0.60  0.67  0.12  0.18 0.86 0.135 2.3
pMales           0.08  0.92  0.23 -0.01  0.10 0.92 0.077 1.2
pFemale         -0.17  0.92  0.14  0.03 -0.10 0.91 0.092 1.1
pWhite          -0.96  0.04  0.09 -0.04  0.05 0.94 0.059 1.0
pMixed           0.92  0.14  0.03 -0.04 -0.10 0.87 0.128 1.1
pAsian           0.83 -0.11 -0.20  0.02  0.06 0.74 0.256 1.2
pBlack           0.84  0.03  0.01  0.11 -0.20 0.75 0.248 1.2
pOther_Ethnic    0.86  0.00  0.27 -0.02 -0.09 0.83 0.174 1.2

                       RC1  RC2  RC3  RC4  RC5
SS loadings           7.30 6.92 2.44 2.08 1.83
Proportion Var        0.30 0.29 0.10 0.09 0.08
Cumulative Var        0.30 0.59 0.69 0.78 0.86
Proportion Explained  0.35 0.34 0.12 0.10 0.09
Cumulative Proportion 0.35 0.69 0.81 0.91 1.00

Mean item complexity =  1.8
Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is  0.04
 with the empirical chi square  248.34  with prob <  3.6e-05

Fit based upon off diagonal values = 0.99
```

RC4	Distance to Work	pDist_0to5km	-0.97	
RC5	Distance to Work	pDist_5to20km	-0.67	
		pDist_40toOver	0.60	

4.5 Clustering

The England Local Authority were clustered based on the total covid-19 deaths, Total Young population, part time work, and work from home variables which are very significant in explaining the covid-19 deaths number in each local authority. All the 323 local authorities divided in 5 cluster based on K-Mean clustering algorithm. As the pie chart indicates that the majority of geographical areas are defined in cluster first. The cluster fourth seems to be an outlier which has no commonality with any areas.





5 Data Modeling

First, three multi-regression models were created using three groups of independent variables as regressors. These groups of variables are the PCA factors, all independent variables, and the selected variables. Then each models went under a series of refinement and comparison process to reach the final best model.

5.1 Model Based on PCA Factors

During the process of factor analysis in previous section, the IVs loaded on to five common factors based on the principal component analysis. Now, PCA factors are utilized as regressors for the model, as shown below:

```
all.IVs.PCA <- DS7006E[,c("pChild_Age","pTeen_Age","pYoung_Age","pOld_Age","pHigh_Disability",
                          "pLow_Disability","pNo_Disability","pTotal_Dist","pDist_0to5km","pDist_5to20km",
                        "pDist_20to40km","pDist_40orOver","pFrom_Home","pHW_Total","pPart_Time","pFull_Time",
                        "pOver_Time","pMales","pFemale","pWhite","pMixed","pAsian","pBlack","pOther_Ethnic")]

#Model based on the PCA factors
PCA <- principal(all.IVs.PCA, nfactors=5, rotate="varimax")

modelPCA.a <- lm(pTotal_Death~.,data.frame(PCA$scores))
```

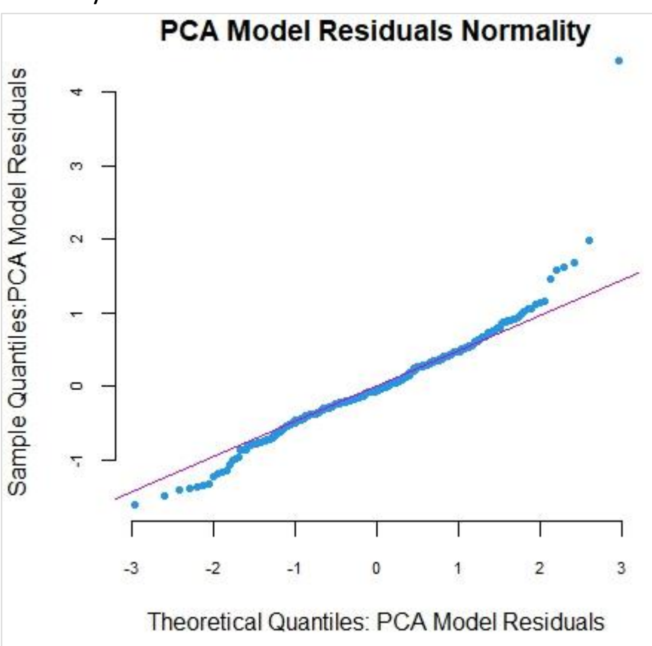
5.1.1 Collinearity Check:

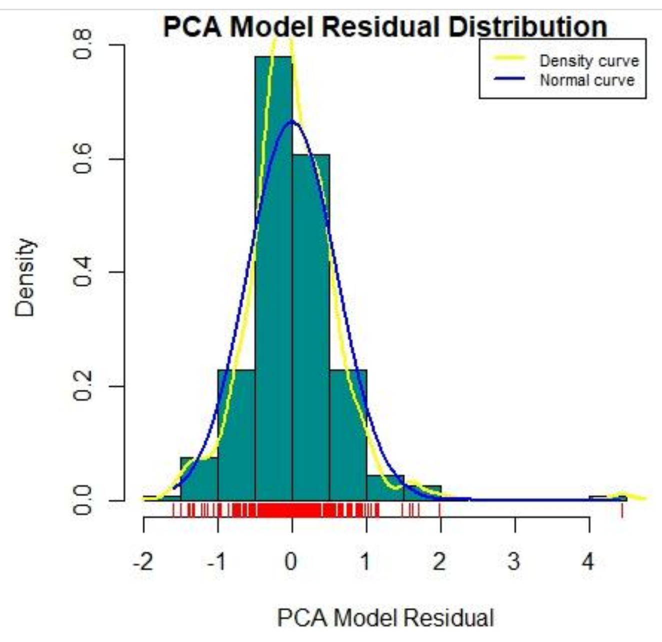
Calculating the variance inflation factors by the ***vif()*** function, and comparing the VIF square root with a benchmark, it shows that there are no collinearities between the PCA factors in the model.

```
> vif(modelPCA.a)
     RC1      RC2      RC3      RC4      RC5
3.098817 2.548061 1.896800 1.325306 2.086168
> sqrt(vif(modelPCA.a)) > 2
  RC1   RC2   RC3   RC4   RC5
FALSE FALSE FALSE FALSE FALSE
```

5.1.2 Residual Normality Check

The normality of the model residuals distribution was checked through visualization by histogram and Q-Q plot, and formal statistical test using the KS-Test. As have shown below, the model residuals are relative distributed normally across the model line.





```
ks.test(modelPCA.a$residuals,"pnorm",mean(modelPCA.a$residuals),sd(modelPCA.a$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  modelPCA.a$residuals
D = 0.061784, p-value = 0.1698
alternative hypothesis: two-sided
```

5.2 Model Based on All Variables

Although we know from the Exploration phase that there are multi-collinearities between the independent variables, still a model was created using all these variables. The summary of the model shows that before checking the collinearity, a number of covariates need to be removed. After removing a number of variables from set of regressors, the following model was created.

```
#Model based on the overall covariates
all.IVs.a <- DS7006E[,c("pChild_Age","pTeen_Age","pYoung_Age","pHigh_Disability","pLow_Disability",
                     "pDist_0to5km","pDist_5to20km","pDist_20to40km","pDist_40orOver","pFrom_Home",
                      "pPart_Time","pFull_Time","pOver_Time","pFemale","pWhite","pMixed",
                     "pAsian","pBlack")]

modelAll.a <- lm(pTotal_Death ~.,data = data.frame(all.IVs.a))
```

5.2.1 Model Collinearity Check

The VIF indicate multi-collinearity between most of the model's regressors. It was predictable, because in the Data Exploration section it was checked.

```
> vif(modelAll.a)
     pChild_Age        pTeen_Age       pYoung_Age pHigh_Disability  pLow_Disability     pDist_0to5km
       4.806476         3.969209        12.477611        12.064125        21.514874        80.686517
  pDist_5to20km   pDist_20to40km   pDist_40orOver       pFrom_Home       pPart_Time       pFull_Time
      69.253383        28.880870        11.995935        19.322550         9.677373        37.363414
     pOver_Time          pFemale           pWhite           pMixed           pAsian           pBlack
      26.406901        20.237124       477.371638        17.719877       187.878706        62.192770
> sqrt(vif(modelAll.a)) > 2
     pChild_Age        pTeen_Age       pYoung_Age pHigh_Disability  pLow_Disability     pDist_0to5km
           TRUE            FALSE             TRUE             TRUE             TRUE             TRUE
  pDist_5to20km   pDist_20to40km   pDist_40orOver       pFrom_Home       pPart_Time       pFull_Time
           TRUE             TRUE             TRUE             TRUE             TRUE             TRUE
     pOver_Time          pFemale           pWhite           pMixed           pAsian           pBlack
           TRUE             TRUE             TRUE             TRUE             TRUE             TRUE
```

```
ks.test(modelAll.a$residuals,"pnorm",mean(modelAll.a$residuals),sd(modelAll.a$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  modelAll.a$residuals
D = 0.067592, p-value = 0.1045
alternative hypothesis: two-sided
```

5.3 Creating Best Model out of Existing One

Instead delving into the tedious process of resolving collinearities in the model created based on the all data set variables, it would be good to create the best model out of it using the ***stepwise()*** function:

```
#creating the best model out of modelAll.a
modelAll.best.a <- stepwise(modelAll.a, direction = "forward")
#checking for collinearity between the model regressors
vif(modelAll.best.a)
sqrt(vif(modelAll.best.a)) > 2
```

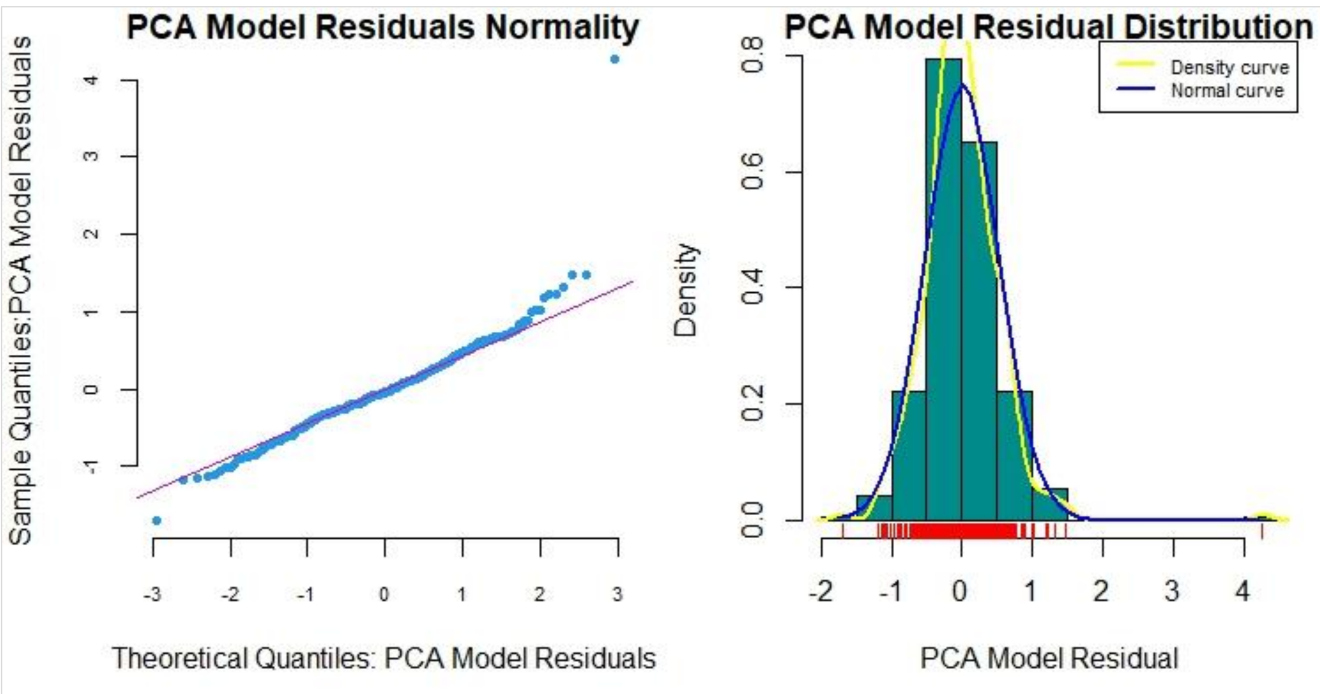
5.3.1 Collinearity Check:

The following result of vif() function shows that there are no collinearity in the model regressors.

```
> vif(modelAll.best.a)
   pFrom_Home    pYoung_Age    pPart_Time        pWhite pDist_20to40km
     1.319330      2.981858      1.781084      2.427410       1.262731
> sqrt(vif(modelAll.best.a)) > 2
   pFrom_Home    pYoung_Age    pPart_Time        pWhite pDist_20to40km
        FALSE         FALSE         FALSE         FALSE          FALSE
```

5.3.2 Residuals Normality

The visual graph, and the formal statistical normality test show that the model residuals have relatively normally distributed across the model.




```
ks.test(mabResidual,"pnorm",mean(mabResidual),sd(mabResidual))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  mabResidual
D = 0.060664, p-value = 0.1854
alternative hypothesis: two-sided
```

5.4 Model Based on Selected Variables

Resolving the collinearity issue between the covariates in the 'Exploration' phase, ended up with a number of selected variables which have no strong correlation coefficients with each other. The following model was created using these variable as regressors.

```
selected.IVs.a <- DS7006E[,c("pChild_Age","pTeen_Age","pDist_0to5km","pDist_5to20km",
                             "pDist_20to40km","pDist_40orOver","pFrom_Home","pHigh_Disability",
                             "pPart_Time","pFemale","pMixed")]

modelSIV.a <- lm(pTotal_Death~.,data = data.frame(selected.IVs.a))
```

5.4.1 Collinearity Check

The collinearity check through vif(), and comparing the square root of the model variance inflation factors with a benchmark, indicate that there are multi-collinearity between the regressors. Therefore, before heading to the next step, this issue need to be resolved by removing some regressor having high variance inflation factors.

```
> vif(modelSIV.a)
      pChild_Age        pTeen_Age     pDist_0to5km    pDist_5to20km   pDist_20to40km   pDist_40orOver       pFrom_Home
        2.790472         2.617824        22.118461        21.803688         9.271791         4.198765         7.224312
pHigh_Disability       pPart_Time          pFemale           pMixed
        5.570688         2.881691         9.706049         4.073318
> sqrt(vif(modelSIV.a)) > 2
      pChild_Age        pTeen_Age     pDist_0to5km    pDist_5to20km   pDist_20to40km   pDist_40orOver       pFrom_Home
           FALSE            FALSE             TRUE             TRUE             TRUE             TRUE             TRUE
pHigh_Disability       pPart_Time          pFemale           pMixed
            TRUE            FALSE             TRUE             TRUE
```

5.5 Refined Model

After removing the collinearity in the regressors, the refined model for the selected variables was created.

```
selected.IVs.b <- DS7006E[,c("pChild_Age","pTeen_Age","pDist_5to20km","pDist_20to40km","pFrom_Home",
                             "pHigh_Disability","pPart_Time","pFemale","pMixed")]
modelSIV.b <- lm(pTotal_Death~.,data = data.frame(selected.IVs.b))
```

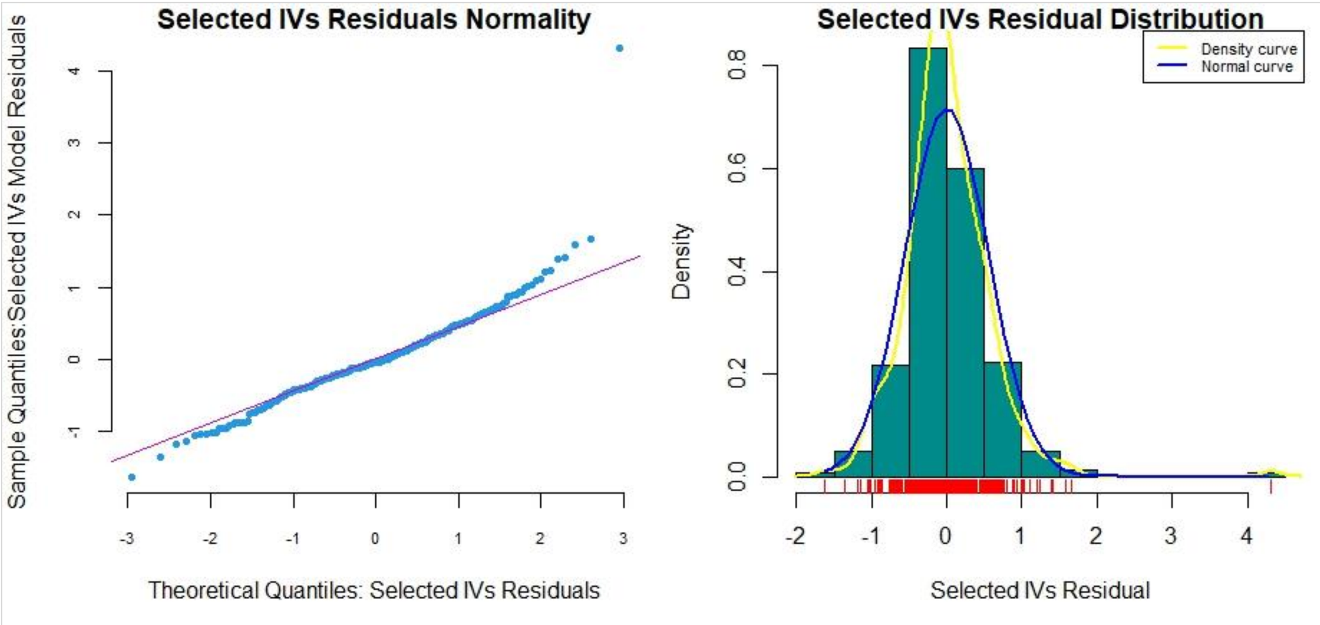
5.5.1 Collinearity Check

The collinearity check through the vif() function and calculating the square root of vif, indicate that there are no collinearity between the model's regressors.

```
> vif(modelSIV.b)
      pChild_Age        pTeen_Age    pDist_5to20km   pDist_20to40km       pFrom_Home pHigh_Disability       pPart_Time
        2.782178         2.491235         1.248626         1.653048         2.848585         3.898701         2.251298
         pFemale           pMixed
        2.622860         2.988634
> sqrt(vif(modelSIV.b)) > 2
      pChild_Age        pTeen_Age    pDist_5to20km   pDist_20to40km       pFrom_Home pHigh_Disability       pPart_Time
           FALSE            FALSE            FALSE            FALSE            FALSE            FALSE            FALSE
         pFemale           pMixed
           FALSE            FALSE
```

5.5.2 Residuals Normality Check

The histogram, Q-Q plot of the model residuals, and the KS-Test for normality confirm that the model residuals have relatively normally distributed.




```
ks.test(msivResidual,"pnorm",mean(msivResidual),sd(msivResidual))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  msivResidual
D = 0.065225, p-value = 0.128
alternative hypothesis: two-sided
```

5.6 Creating Best Model

Using the *stepwise()* function, the best model was created from the model created based on the selected variables.

```
modelSIV.best <- stepwise(modelSIV.b, direction = "forward")
```

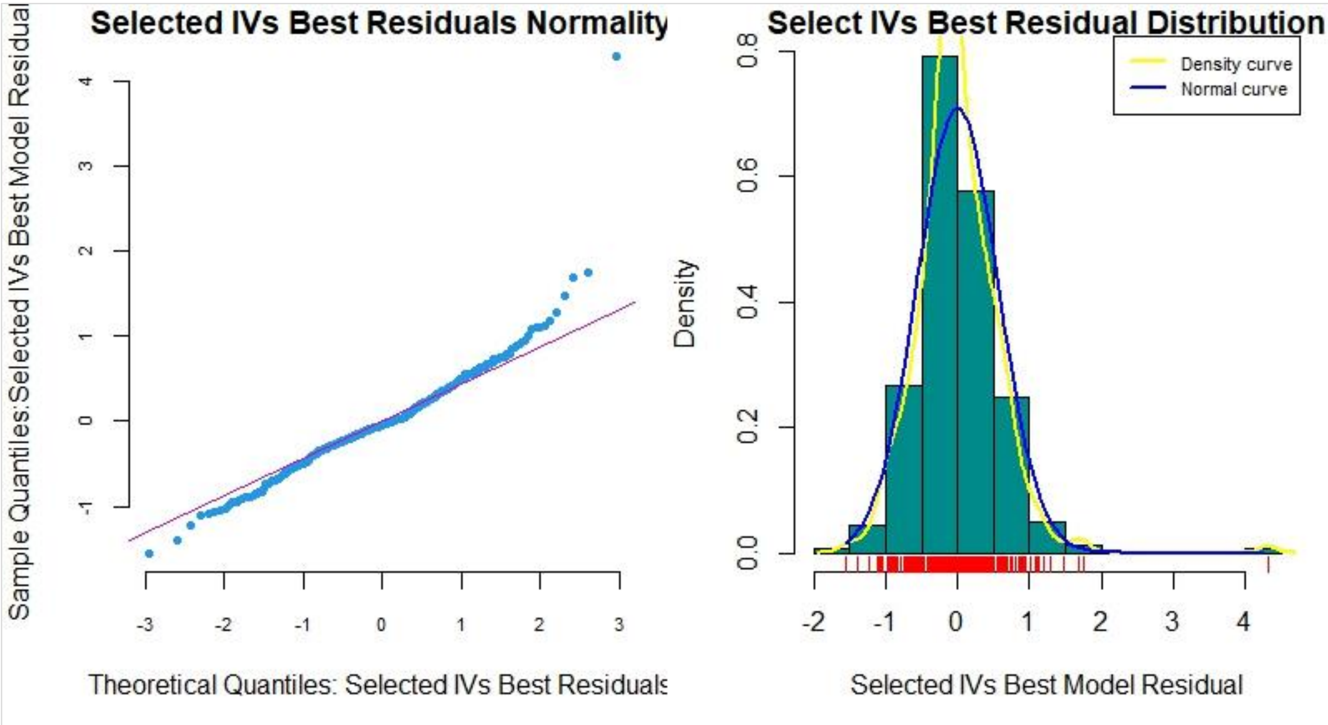
5.6.1 Check for Collinearity

The variance inflation factors indicate that there is no collinearity in the model regressors.

```
> vif(modelSIV.best)
      pFrom_Home    pDist_20to40km pHigh_Disability        pPart_Time         pTeen_Age
        2.438188          1.312503         1.423792          1.563030          1.624603
> sqrt(vif(modelSIV.best)) > 2
      pFrom_Home    pDist_20to40km pHigh_Disability        pPart_Time         pTeen_Age
           FALSE             FALSE            FALSE             FALSE             FALSE
```

5.6.2 Residuals Normality Check

The residuals of the model created in a stepwise forward process, have not been normally distributed, as the following graph and statistical test show.




```
ks.test(msivbResidual,"pnorm",mean(msivbResidual),sd(msivbResidual))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  msivbResidual
D = 0.078788, p-value = 0.03626
alternative hypothesis: two-sided
```

5.7 Checking and Selecting Models

The models created based on different groups of regressors, were compared based on a number of factors. These factors are the R^2, Residuals normality, Collinearity, AIC, Complexity in terms of number of regressors. The summary of calculating all the mentioned evaluation factors are listed in the table below.

Name	R^2	Normality Test	p-value	Collinearity	AIC	No. Var	Residual Distribution
modelPCA.a	0.2138	KS-Test	0.1698	No	598.6753	5	Noraml
modelAll.a	0.4316	KS-Test	0.1045	Yes	519.9101	18	Normal
modelAll.best.a	0.3738	KS-Test	0.1854	No	525.2040	5	Noraml
modelSIV.b	0.3214	KS-Test	0.128	No	559.1607	9	Normal
modelSIV.best	0.3083	KS-Test	0.03626	No	557.3200	5	Un-normal

Taking all the evaluation factors into account, the '***modelAll.best.a***' appears to be the best one among all. Therefore, it was selected to be considered relatively the best model, and will go through further refinement to make the final model out of it.

5.8 Final Model

To refine the 'modelAll.best.a' further, it is needed to know about the relative importance of each regressors in the model, and think about removing some unimportant regressors. This task was done using the calc.relimp() function as follow:

```
calc.relimp(modelAll.best.a, type = c("lmg"), rela = TRUE)
```

```
Relative importance metrics:

                    lmg
pFrom_Home     0.52432972
pYoung_Age     0.21690916
pPart_Time     0.17875565
pWhite         0.04838277
pDist_20to40km 0.03162271
```

The '*Relative importance metrics'* shows that three regressors are more important. Therefore, to have a parsimonious model, the variable 'pWhite', and 'pDist_20to40km' is removed from the regressors list, and the final model is created as following:

```
finalRegressors <- DS7006E[,c("pPart_Time","pFrom_Home","pYoung_Age")]
finalModel <- lm(pTotal_Death~., data = data.frame(finalRegressors))
```

```
lm(formula = pTotal_Death ~ ., data = data.frame(finalRegressors))

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6863 -0.2736 -0.0544  0.3081  4.2823 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.0528561  0.6567621  12.261  < 2e-16 ***
pPart_Time  -0.0220023  0.0036273  -6.066 3.72e-09 ***
pFrom_Home  -0.0170464  0.0018190  -9.371  < 2e-16 ***
pYoung_Age  -0.0057780  0.0006864  -8.417 1.32e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5526 on 319 degrees of freedom
Multiple R-squared:  0.3374,	Adjusted R-squared:  0.3312 
F-statistic: 54.15 on 3 and 319 DF,  p-value: < 2.2e-16
```

$$covid_{deaths} = -0.22 \times work_{part\ time} - 0.017 \times work_{from\ home} - 0.00577 \times Young_{age} + 8.05$$

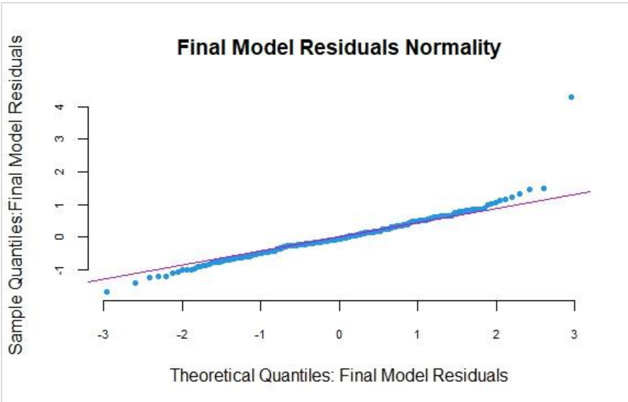
Defining mathematical annotation for each dependent and independent variable like *Covid-19 Deaths (Y*), *Part-Time (X1), Work from Home (X2), and Young Age (X3),* the model takes the following mathematical form:

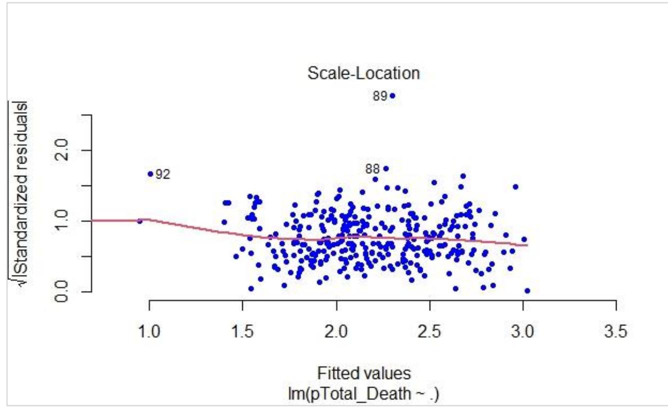
$$\boldsymbol{Y = -0.22X_1 - 0.017X_2 - 0.00577X_3 + 8.05} \qquad (1)$$

To be sure about the quality of the model, there are some assumptions need to be checked. These assumptions are the residuals normality, linearity between fitted values and residuals, and homogeneity of variance. Each of this assumption are checked through visualization tools.

5.8.1 Normality of Residuals, and Homogeneity of Variance:

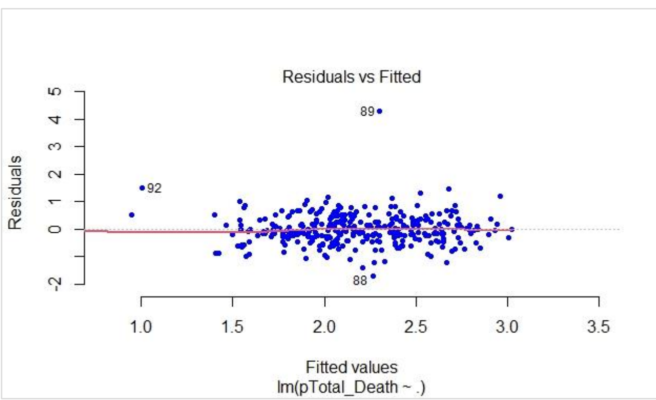
Across the mass of the data points, the homogeneity line is horizontally flatted, except at the end of left side of the data points it has been curved upward. However, the variance preserve homogeneity across the mass of the data.

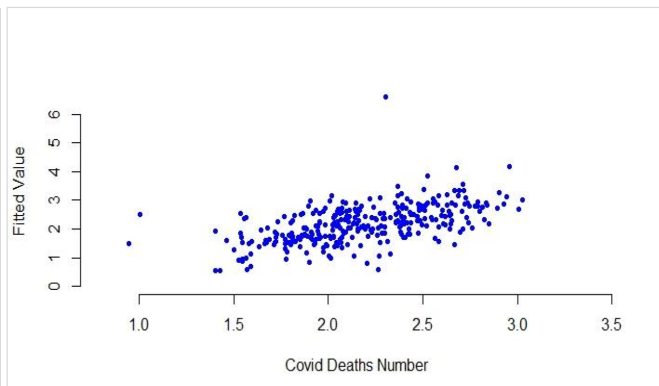




5.8.2 Linearity Between Fitted and Observed Values, and Fitted Values & Residuals

The scatter plots were drawn from the fitted and observed values, as well as fitted values and residuals indicate good linearity. Except a few number of outliers, other data points are linearly homogenous across data distribution direction.





6 Discussion and Conclusion

The quantitative analysis of covid-19 data and a number of variables such as age groups, disabilities, traveling distance to work, work type, and ethnicities were taken from the census data of England local authority, reveals that among all the mentioned socioeconomic variables, “*work from home*”, and “*part-time work*” play significant role in the reduction of Covid-19 deaths. As independent variables, both “work from home” and “part-time work” are associated with the negative coefficients with dependent variable (Covid-19 Deaths) in the model. It means that by increasing work from home and doing part-time job, the number of Covid-19 Deaths are declined, and the people would be relatively safe.

In the context of Covid-19 pandemic, work from home, and part-time job means reduction of mobility and movement of the people which have been, according to previous studies, effective in containing the outbreak. The finding of this research project is in line with findings of other researches saying that “reduction of work-related mobility was accompanied by a nearly linear benefit in outbreak containment” (Vinceti, et al., 2022), and (Fadinger & Schymik, 2020).

The research ended with associating a number of socioeconomic variables from the target population census data and Covid-19 deaths; this association formulated as mathematical equation which satisfactorily fulfil the main objective of the research, and answer questions regarding the association of effective socioeconomic factors with the pandemic deaths.

The research conducted on independent variables extracted from 2011 census data in England local authorities, and the covid-19 from 2020, and 2021. The difference in time periods between dependent and independent variables, may negatively affect the consistency between dependent and independent variables, and consequently the result, which can be considered the main limitation and weakness of the research.

Although, the research findings are confirmed by other research performed in the similar context; however, researches on census data containing similar socioeconomic variables from pandemic time period will contribute to further validation of the finding. Therefore, there are avenues for further research on similar context and socioeconomic variables, especially with no inconsistency of the time period between dependent and independent variables.

7 References

Alfano, V. & Ercolano, S., 2020. The Efficacy of Lockdown Against COVID-19: A Cross-Country Panel. *Applied Health Economics and Health Policy,* Volume 18, pp. 509-517.

Cifuentes-Faura, J., 2021. Factors influencing the COVID-19 mortality rate in the European Union: importance of medical professionals. *Public Health,* Volume 200, pp. 1-3.

Ciotti, M. et al., 2020. The COVID-19 pandemic. *Critical Reviews in Clinical Laboratory Sciences,* 57(6), pp. 365-388.

Dormann, C. F. et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography,* Volume 36, pp. 27-46.

Fabrigar, L. & Duane , W., 2012. *Exploratory Factory Analysis.* New York: Oxford University Press, Inc..

Fadinger, H. & Schymik, J., 2020. The Effects of Working from Home on Covid-19 Infections and Production: A Macroeconomic Analysis for Germany. *Covid Economics,* 9(24), pp. 107-139.

Goujon, A. et al., 2020. Age, gender, and territory of COVID-19 infections and fatalities. *Luxembourg: Publications Office of the European Union,* p. 26.

Ieno, E. N. & Zuur, A. F., 2015. *A Beginner's Guide to Data Exploration and Visualization with R.* 1st ed. Newburgh: Highland Statistics Ltd..

Mishra, P. et al., 2019. *National Institute of Health (NIH).* [Online]
Available at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/#:~:text=The%20two%20well%2Dknown%20tests,the%20normality%20of%20the%20data.
[Accessed 30 November 2023].

Navarro, D., 2015. *Learning Statistics with R: A Tutorial for Psychology Students and other Beginners.* 6 ed. Adelaide: Creative Commons BY-SA.

Padhan, R. & Prabheesh, K., 2021. The economics of COVID-19 pandemic: A survey. *Economic Analysis and Policy,* Volume 70, pp. 220-237.

Pokhrel, S. & Chhetri, R., 2021. A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. *Sage Journals,* 8(1), p. 1330141.

Shakespeare, T., Ndagire, F. & Seketi, Q. E., 2021. Triple jeopardy: disabled people and the COVID-19 pandemic. *The Lancent,* 397(10282), pp. 1331-1333.

Tarkar, P., 2020. Impact Of Covid-19 Pandemic On Education System. *International Journal of Advanced Science and Technology,* 29(9), pp. 3812-3814.

Vinceti, M. et al., 2022. Substantial impact of mobility restrictions on reducing COVID-19 incidence in Italy in 2020. *Journal of Travel Medicine ,* pp. 1-10.

Zuur, A., Ieno, E. & Elphick, C., 2010. A Protocol for Data Exploration to Avoid Common Statistical Problem. *British Ecological Society,* Volume 1, pp. 3-14.

8 Appendix

DS7006_Code_File

DS7006_Data_Set