# Covid-Data analysis Draft

## **Abstract**

This paper is about data analysis to extract some insight about data provided with R using a number of packages. The aim of this analysis is to establish some relationship with socio and economic variables like Age, Industry, Activity level, Gender etc.

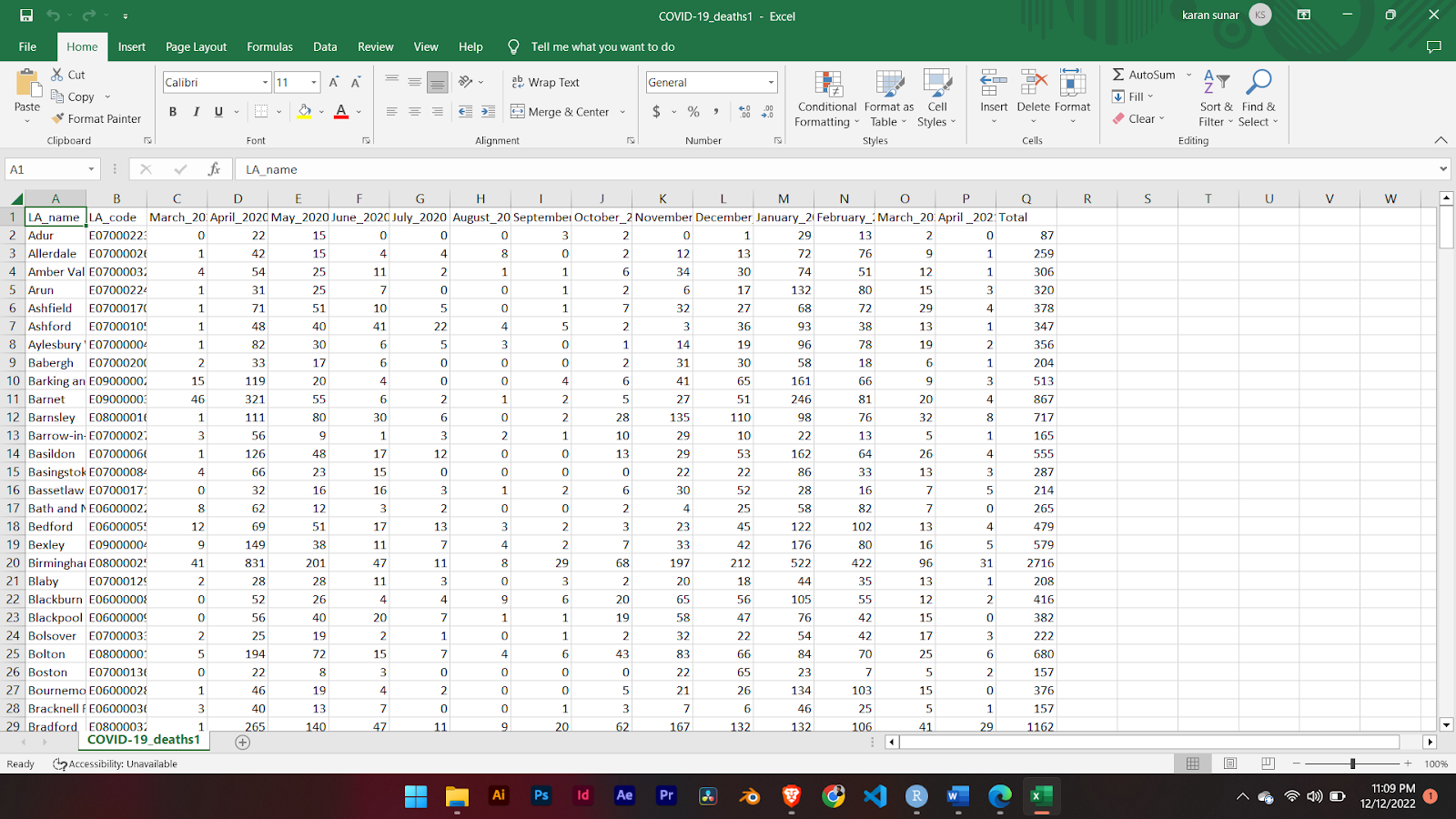
We can use the data set to find the correlations between different variables, further we can build a model using regression for e.g., we can find how many people had died in covid within certain age groups who were active or not.

## **Introductions**

We have been provided with some data about covid-19 death, we will try to choose the different variables and find correlations among these variables using exploratory data analysis such that it will establish some kind of relationships.

The algorithms which are to be used in this analysis are linear regression, multiple regressions and various hypothesis techniques like parametric and non-parametric tests for example t-test, Chi-squared, Anova etc.

## **Covid-19 Data set**



## **Data set selected**

I have selected the three themes for now, variables might be changed based on the analysis.

**A.)** **Age**

It consists of the age group for the populations for the local authority by districts. It contains the various age group populations.

**B.)** **Gender:** It consists of the male and female populations data

**C.)**  **Activity level**

It consists of people with different activity levels on a daily basis like Day-to-Day activities limited a lot, Day-to-Day activities not limited and Day-to-Day activities limited little.

**Dependent Variable**:

 I have taken the **Total\_covid\_Death** as the dependent variable.

**Independent Variables**:

                    I have taken three themes with a number of variables into independent variables.

a. **Age variable** is grouped into four categories

·   **Children**: It is the integer and number of populations being children.

·   **Young**: It is the integer and number of populations being young.

·   **Adults**: It is the integer and number of populations being adults.

·   **Pensioner**: It is the integer and number of populations being pensioners/old.

b. **Gender variable** contains male and Female population data separately. Example:

·   **Male**: It is the integer and number of populations being male

·   **Female**: It is the integer and number of populations being females

c.  **Activity level on daily basis** into three categories

·   **Day-to-Day activities limited a lot**: It is the integer and number of populations whose daily activities are limited a lot.

·       **Day-to-Day activities limited little:** It is the integer and number of populations whose daily activities are limited little only.

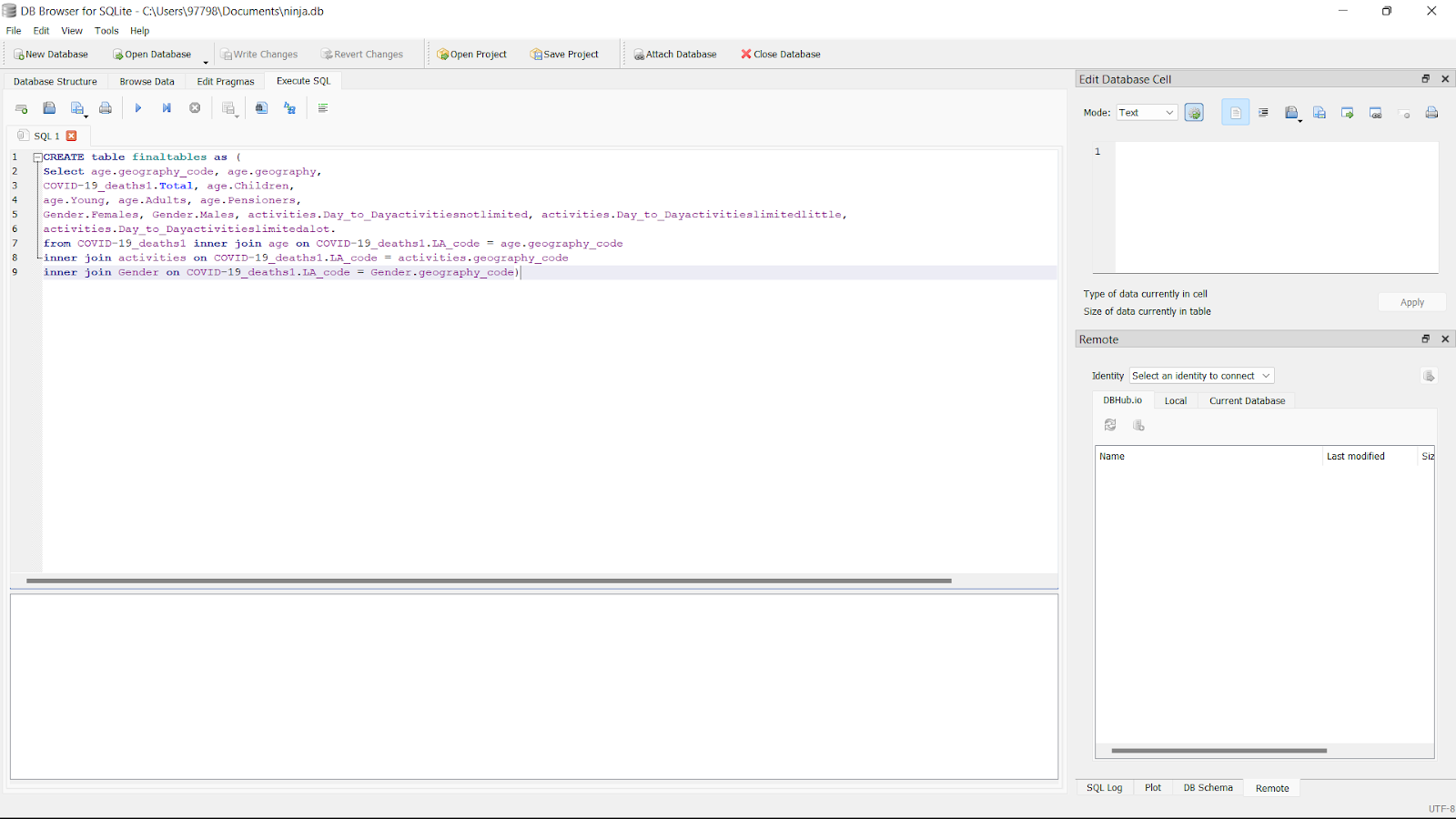
·   **Day-to-Day activities not limited:** It is the integer and number of populations whose daily activities are not limited.

**Note:** Number of variables might change at the end of the project with respect to analysis and to develop some crucial relationships.

## Data pre-processing and Cleaning

         As data seems to be clean already. It doesn’t have any missing values but the data we are provided and the variables from themes don't seem to match up with respect to geography\_code and rows number.

We joined the two dataset with sqlite3 and created a new csv file as our final data set.

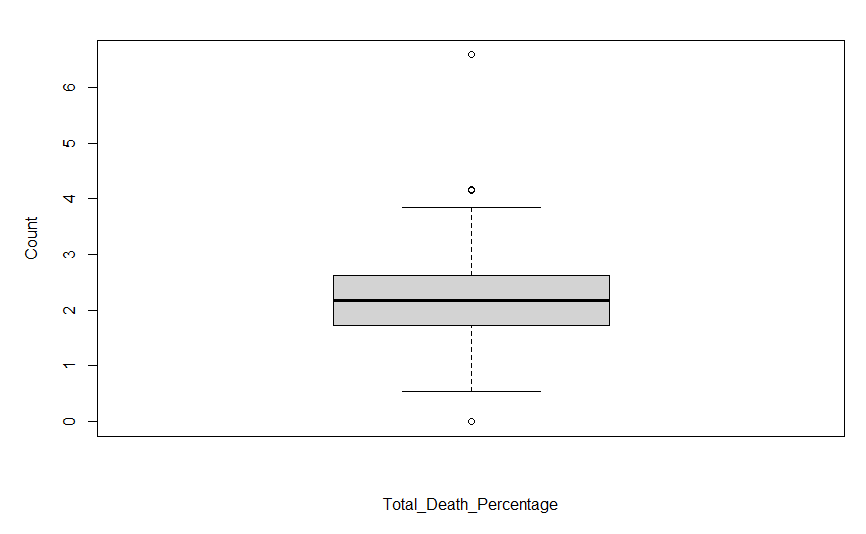


And then we exported this **finaltable** as a csv file for further explorations.

## Data Analysis in R

At first, we have checked the normality of different dependent and independent variables.

I have calculated the percentage of population of the **total\_death** variable. The boxplot as shown below in the figure **Fig (1.2**) shows some outliers in the data.



**Fig (1.2): Boxplot for Total\_Death\_percentage**

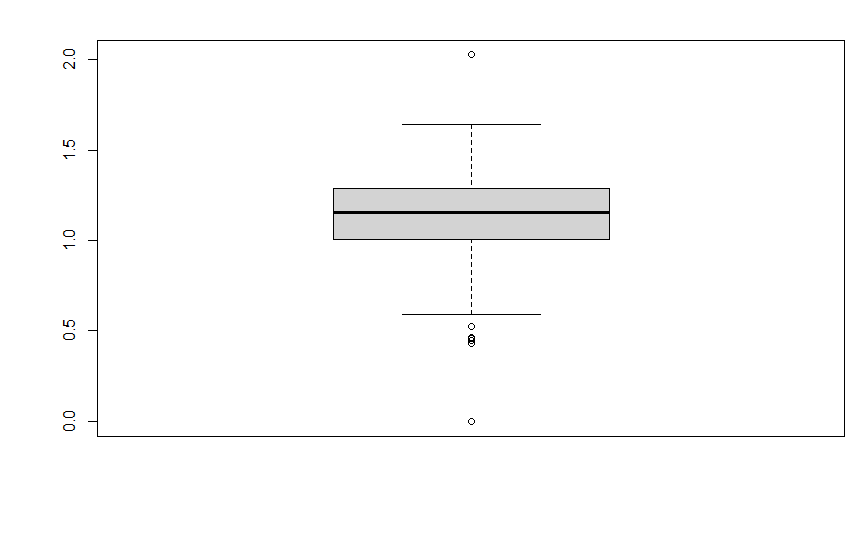
### **Normality Checking**

There are lots of techniques to check the normality. It can be graphical or mathematical. Under graphical I am going to use histogram and Q-Q plots.

As above boxplot shows some outliers we are going to check for its data for normal distributions. Here log is used for normalization.

### **For Dependent Variable**

#### **Boxplot**



**Fig (1.3): Boxplot of normalized Total\_Death\_Percentage variable**

As from figure **Fig (1.3)** above we can clearly see there are some outliers but it is more normalized. We can see the summary of this data as given below

**Total\_Death\_percentage**

**Min. :0.000**

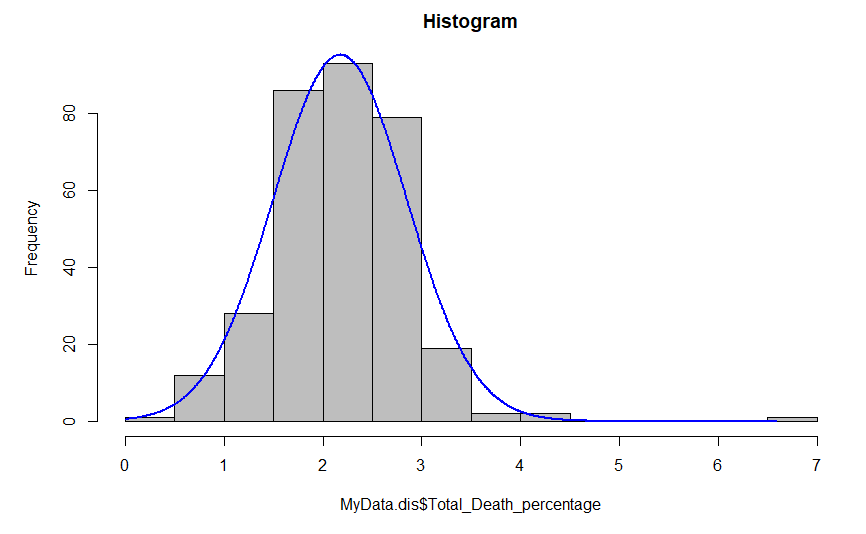
**1st Qu.:1.004**

**Median :1.155**

**Mean :1.132**

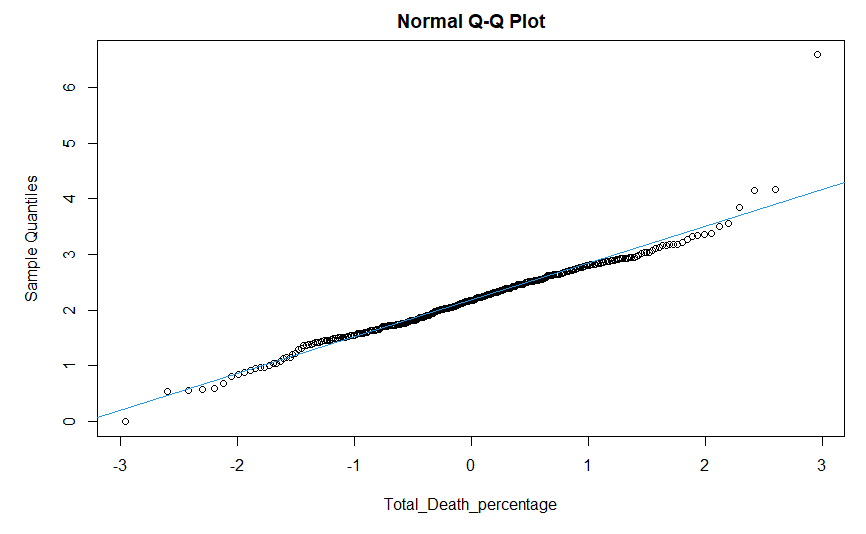
**3rd Qu.:1.287**

**Max. :2.026**



The above histogram shows that the data is kind of normalized with a little right skewed.

#### **Q-Q plot**



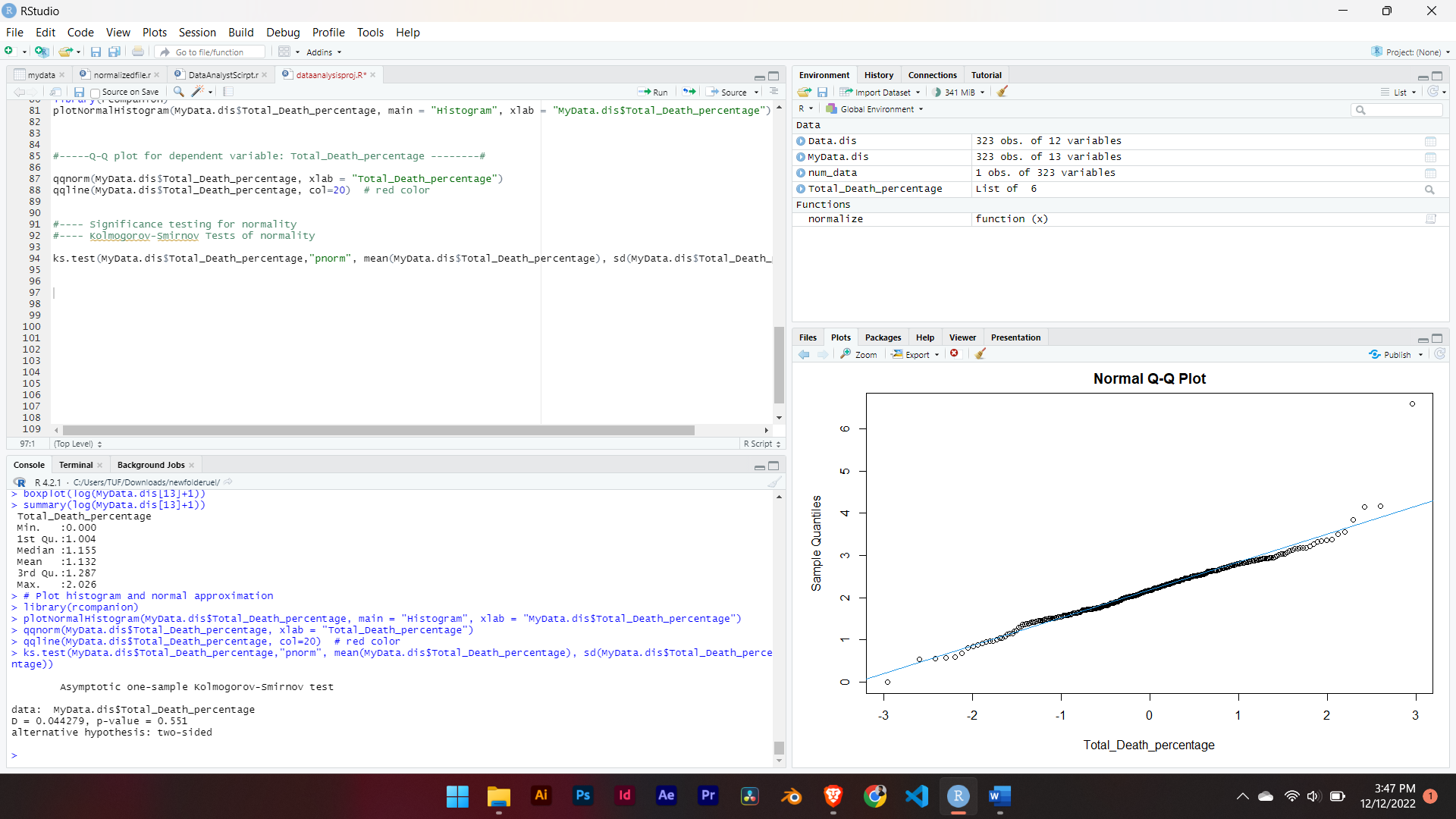
**Fig (1.4): Q-Q normalization plots for Total\_Death\_percentage**

This is the graphical way of checking for normality. The above figure shows some outliers at sample quantiles of 4 and 5.

#### **Significance Test**

**Null hypothesis**: The data comes from the normal distributions

**Alternative hypothesis**: The data doesn’t come from the normal distributions.



**Fig (1.5): Kolmogorov-Smirnov test**

The above figure **Fig (1.5)** illustrates the significance test on normalized value of **Total\_Death\_Percentage** variable.

The p value is greater than 0.05 So we can accept the null hypothesis and say that the data is normally distributes and significant.

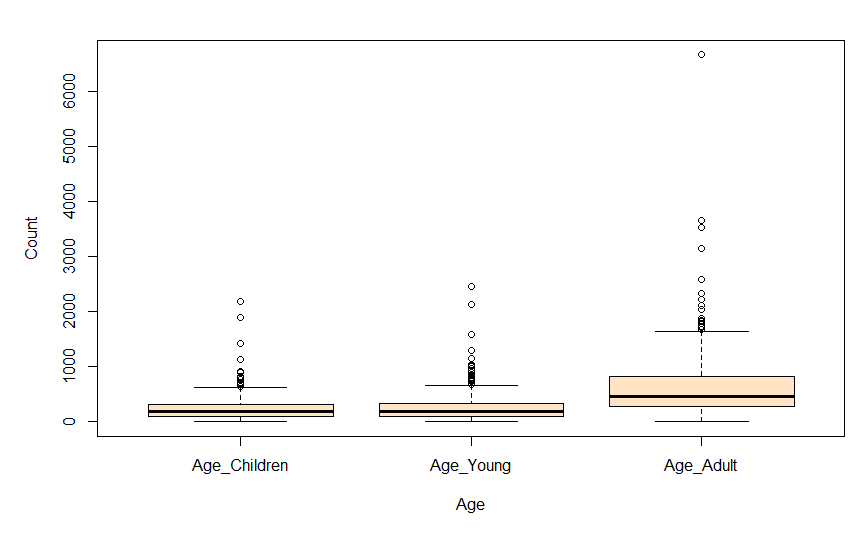
### **For independent variable**

At first, we calculate the percentage of populations of independent variables. We will check and analyze the data pattern and distributions through boxplot and Q-Q plot for each of the independent variables.

#### **Age**

Children, Young, Adult, Pensioners

##### **Boxplot for Age**



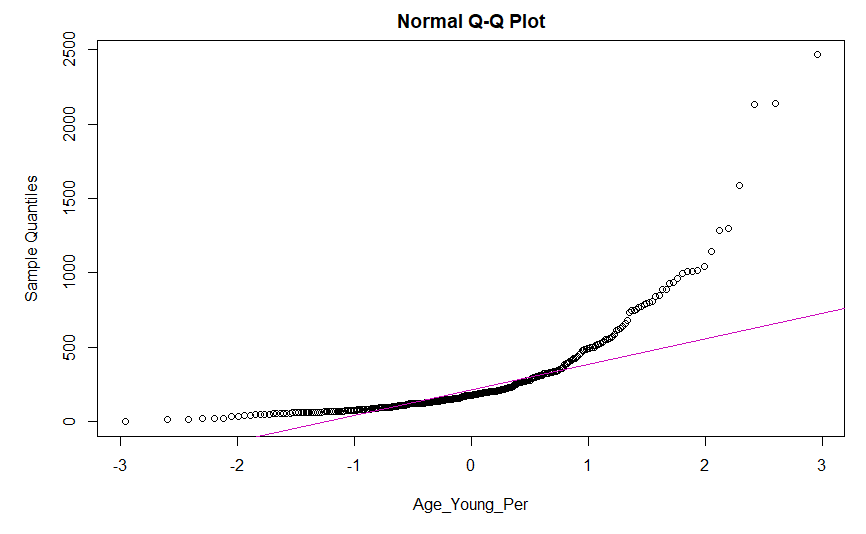
**Fig (1.6): Box plot for independent variable: Age (Children, Young, Adult)**

##### **Q-Q Plots for Age variable**

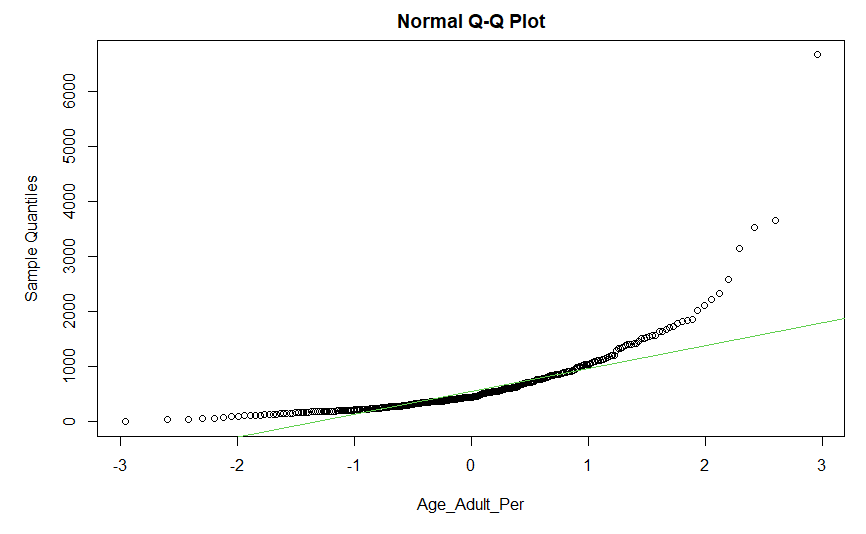
**(Children, Young, Adults, Pensioners)**



**Fig (1.7a): Q-Q plot for variable Age: Children**



**Fig (1.7b): Q-Q plot for variable Age: Young**



**Fig (1.7c): Q-Q plot for variable Age: Adults**

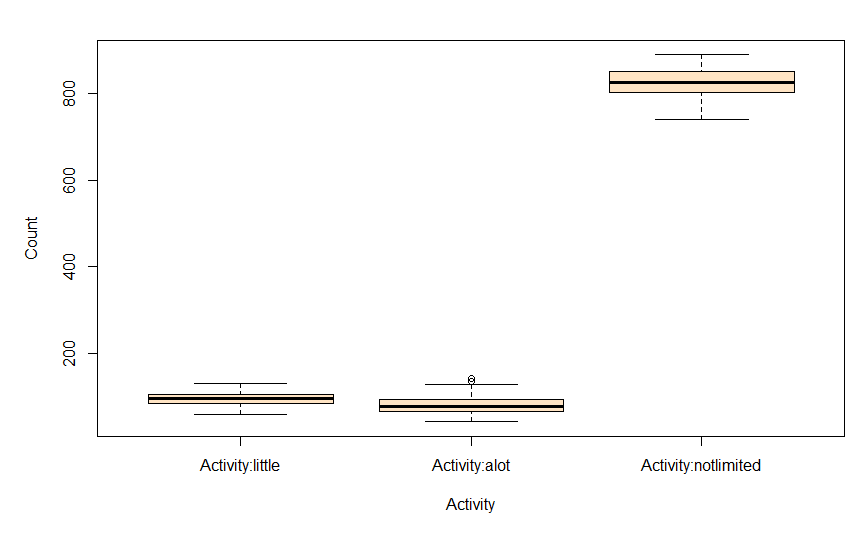
##### ***Interpretations:***

From the above Box plot and Q-Q plots shown in figure Fig (1.6) and Fig (1.7) we can say that the data are not normally distributed.

#### **Activity level on a daily basis:**

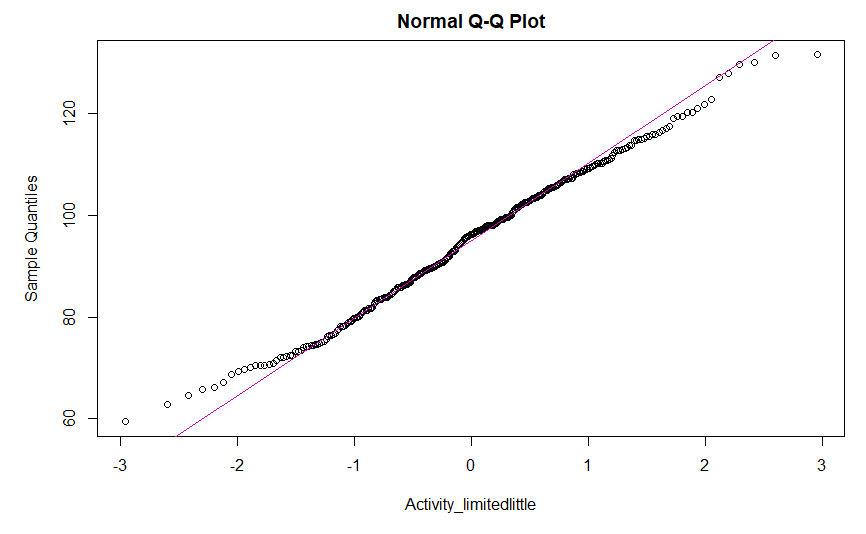
limited a little, limited a lot, not limited

##### **Boxplot for Activity level on a daily basis**

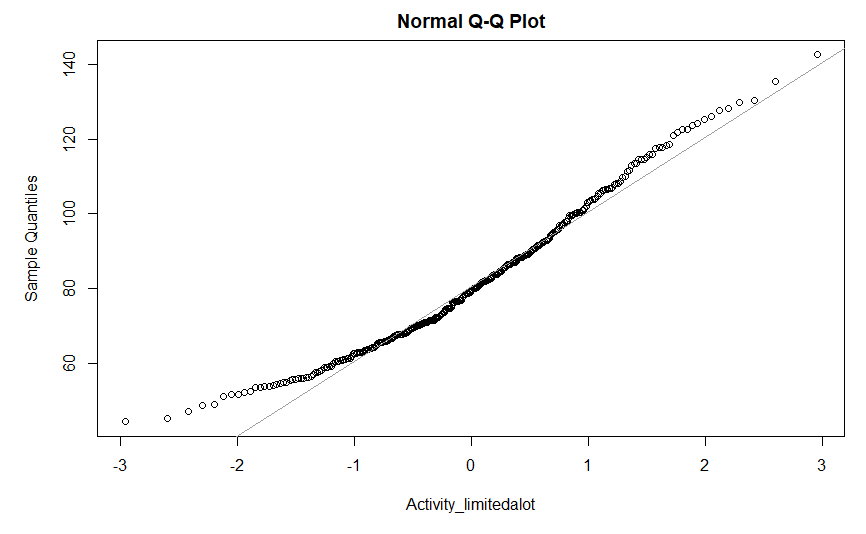
****

**Fig (1.8): Boxplot for activity levels**

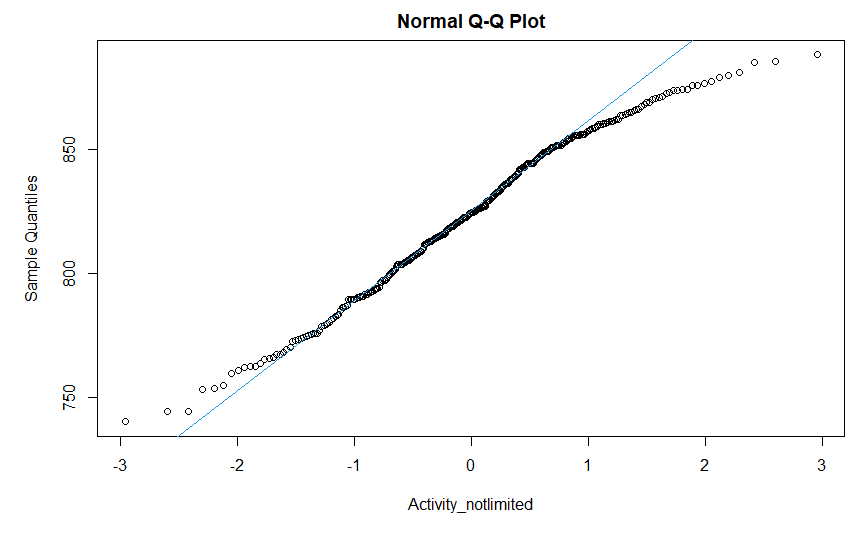
##### **Q-Q plots**

****

**Fig (1.9a): Q-Q plot for Activity level: limited little**

****

**Fig (1.9b): Q-Q plot for Activity level: limited a lot**



**Fig (1.9c): Q-Q plot for activity level: not limited**

##### ***Interpretations:***

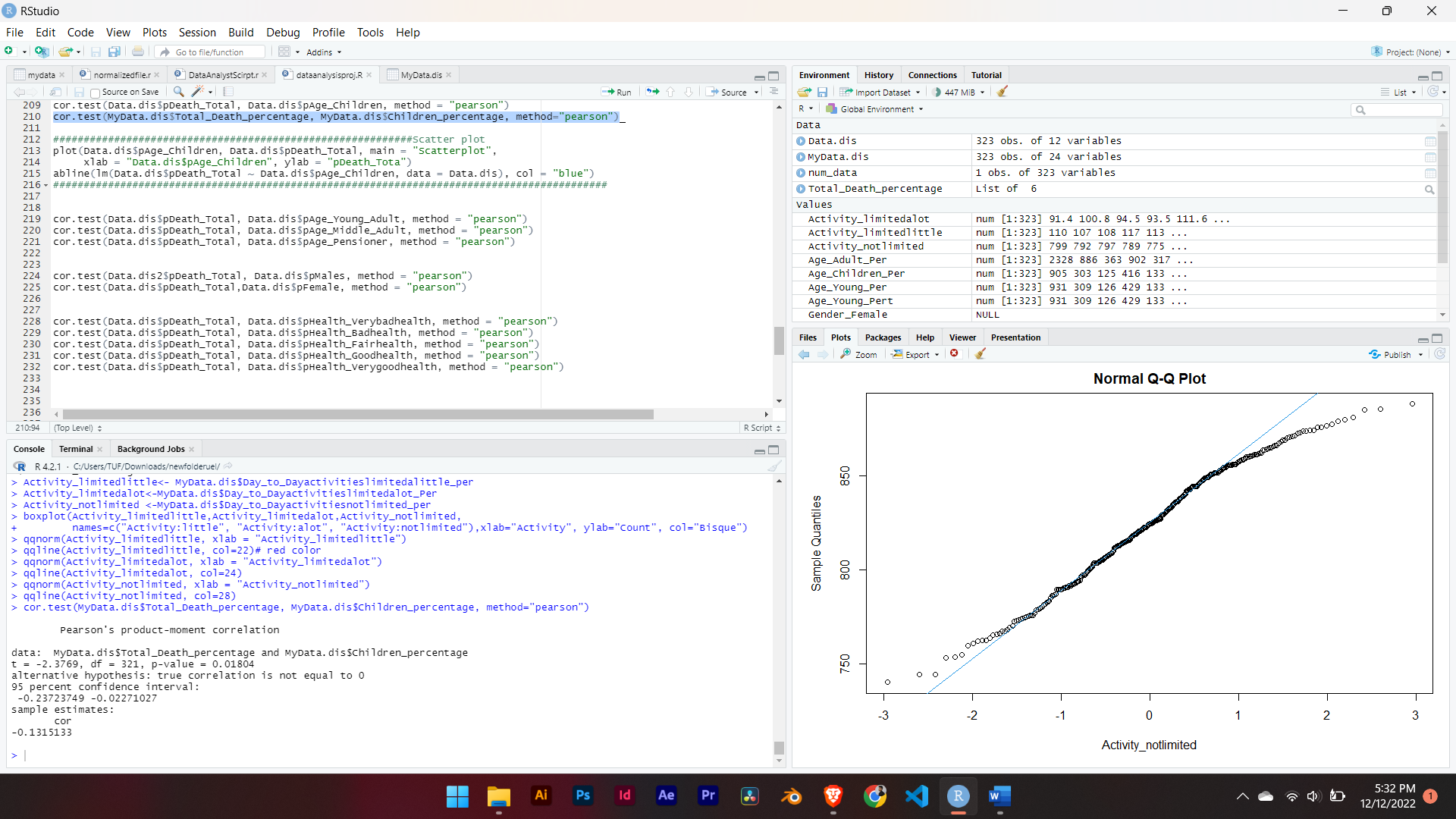
From the figures above Fig (1.9) we can clearly see many outliers.

#### **Analysis**

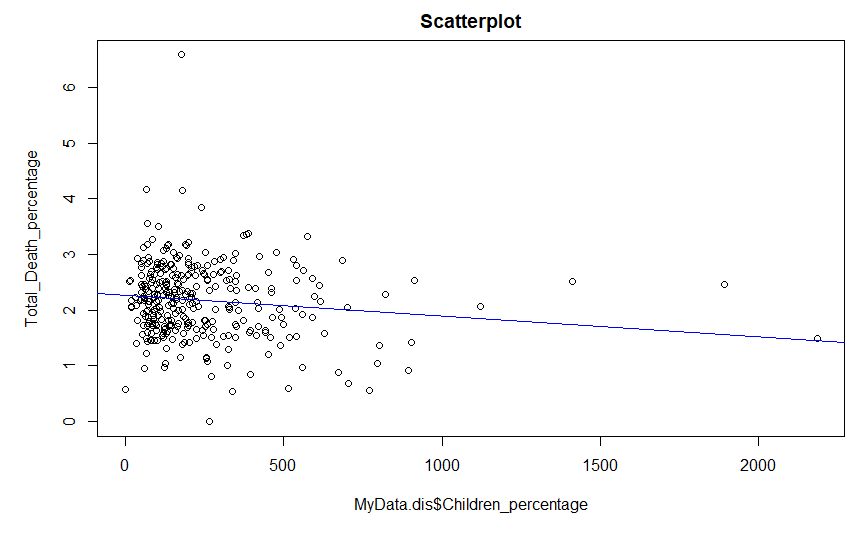
Now we are going to the correlations between the independent variables and dependent variables as one dependent variable is normalized.

##### **With independent variable age groups**

###### **With Age: Children**



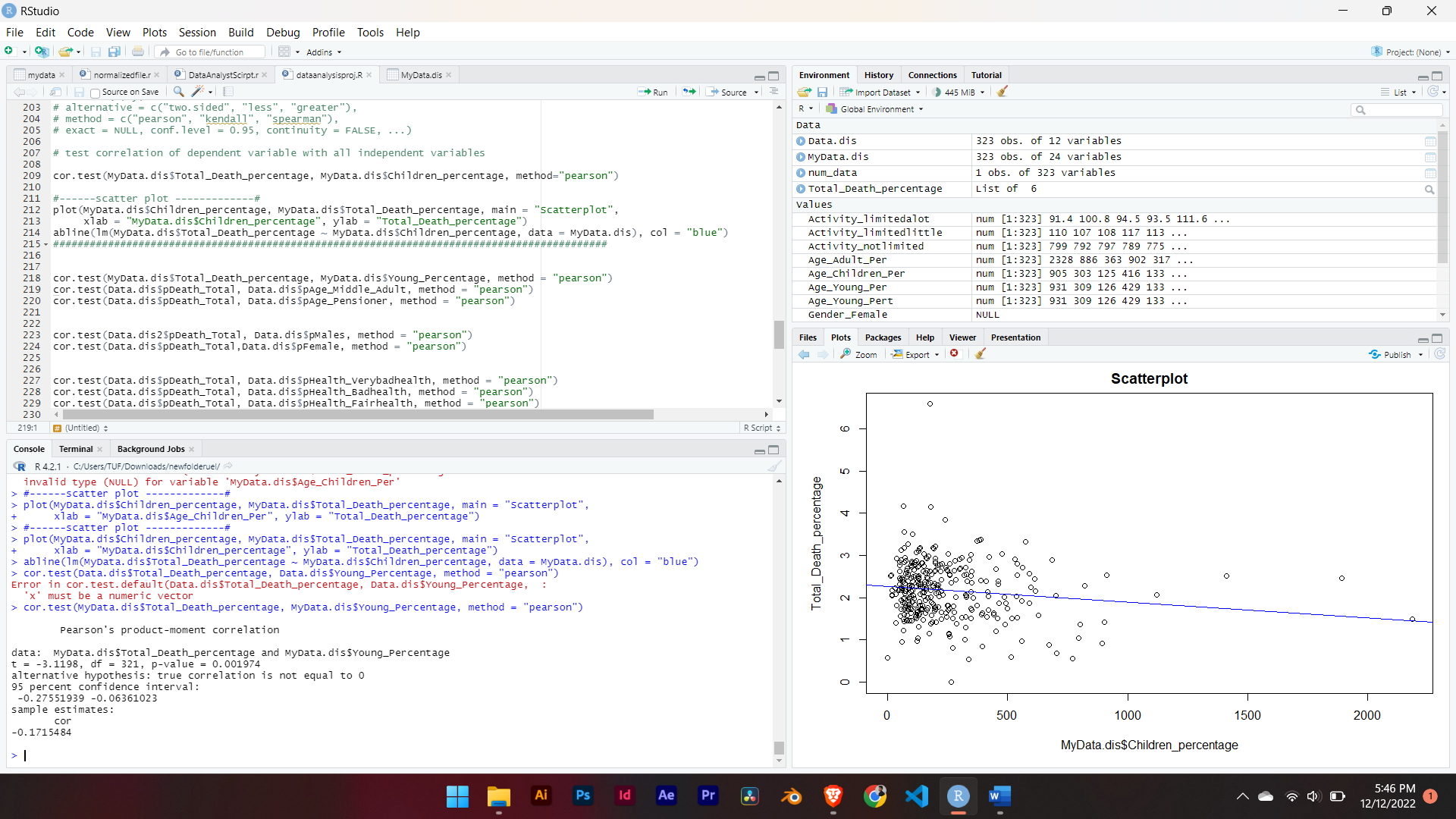
**Fig (2.1a): Pearson correlation calculation between total death and age group of children**

****

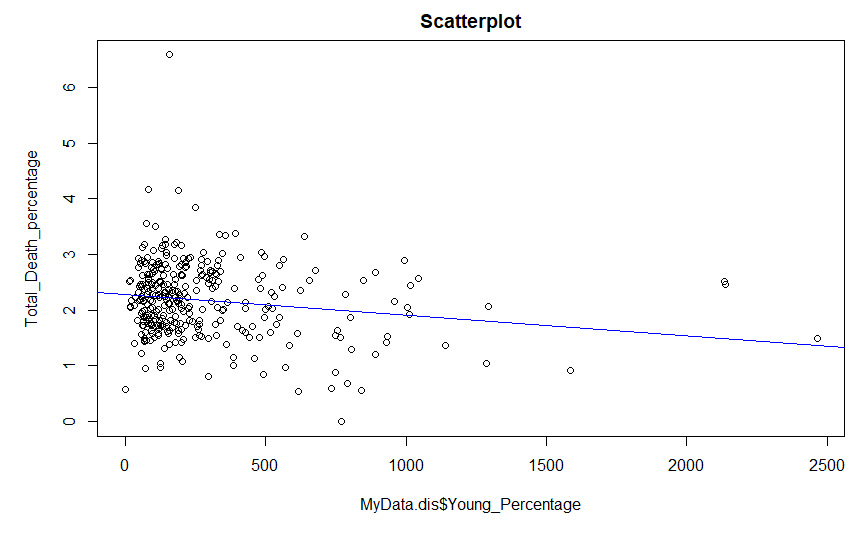
**Fig (2.1b): Scatter plot for the age: children and total\_death**

From the figure above Fig (2.1a), we can see that it shows weak negative correlations, thus it is not statistically significant and no relationships with dependent variable **Total\_Death**

###### **With Age: Young**



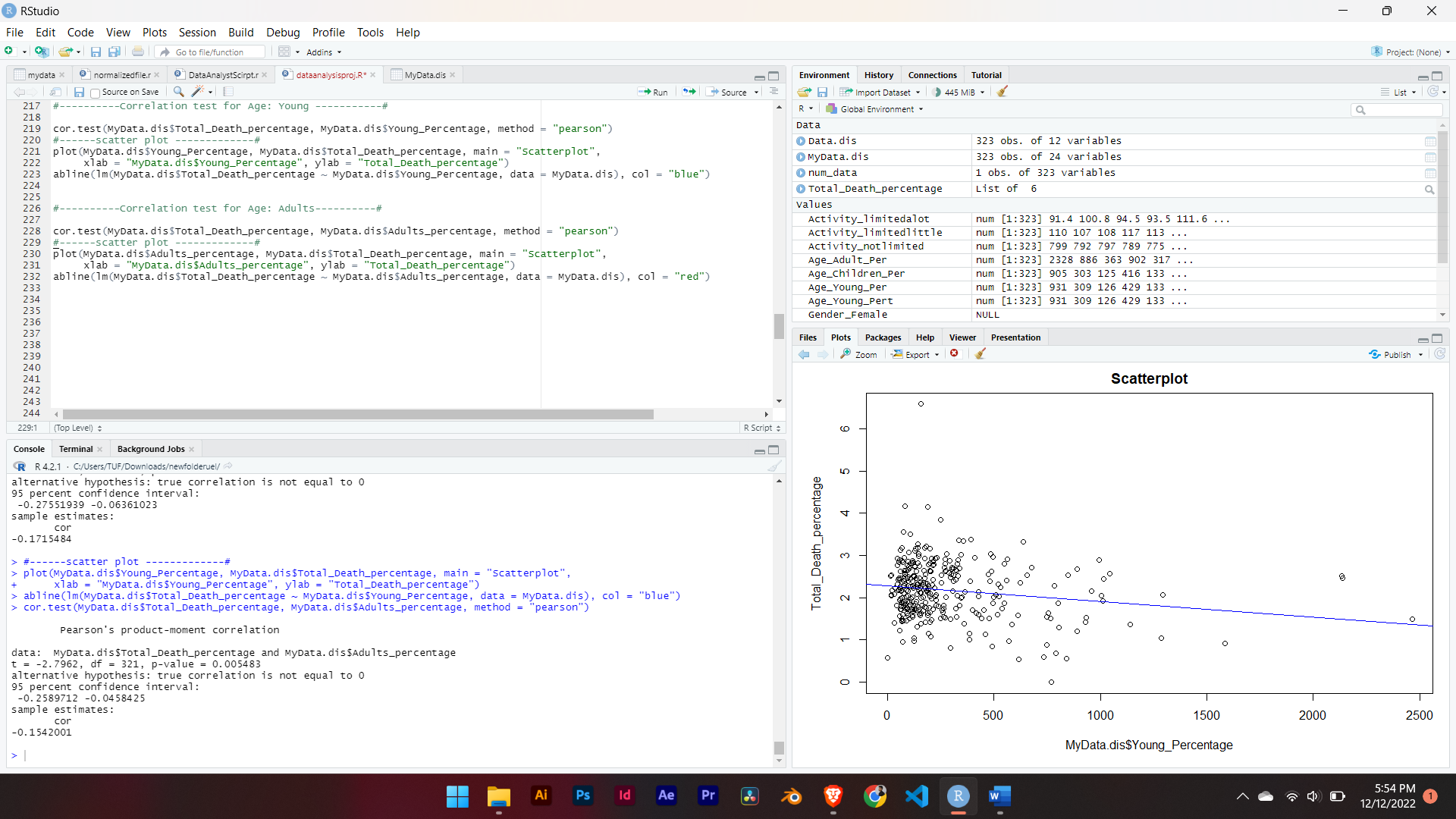
**Fig (2.2a):** **Pearson correlation calculation between total death and age group Young**



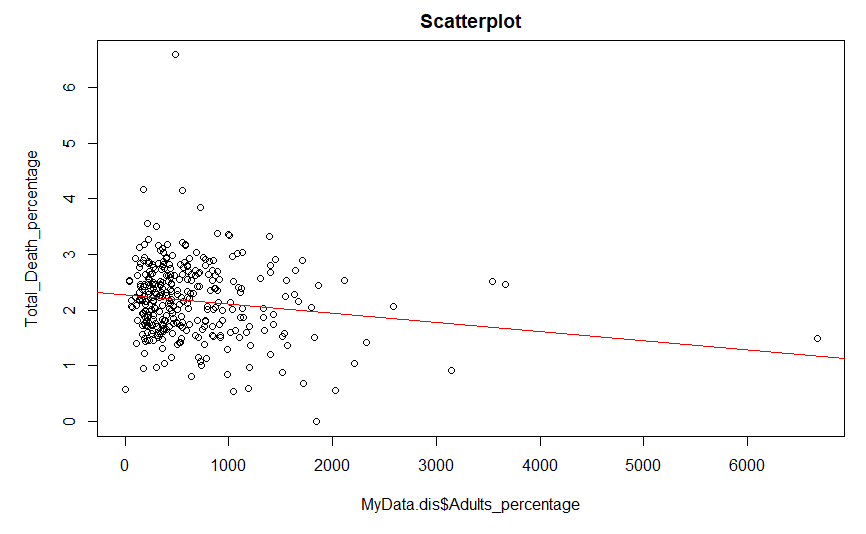
**Fig (2.2b): Scatter plot for the age: Young and total\_death**

From the figure above Fig (2.2a), we can see that it shows negative correlations between age group Young and p-value of 0.01804 which is less than 0.05, thus we reject the null hypothesis and it is not statistically significant.

###### **For Age: Adults**



**Fig (2.2c):** **Pearson correlation calculation between total death and Adults**

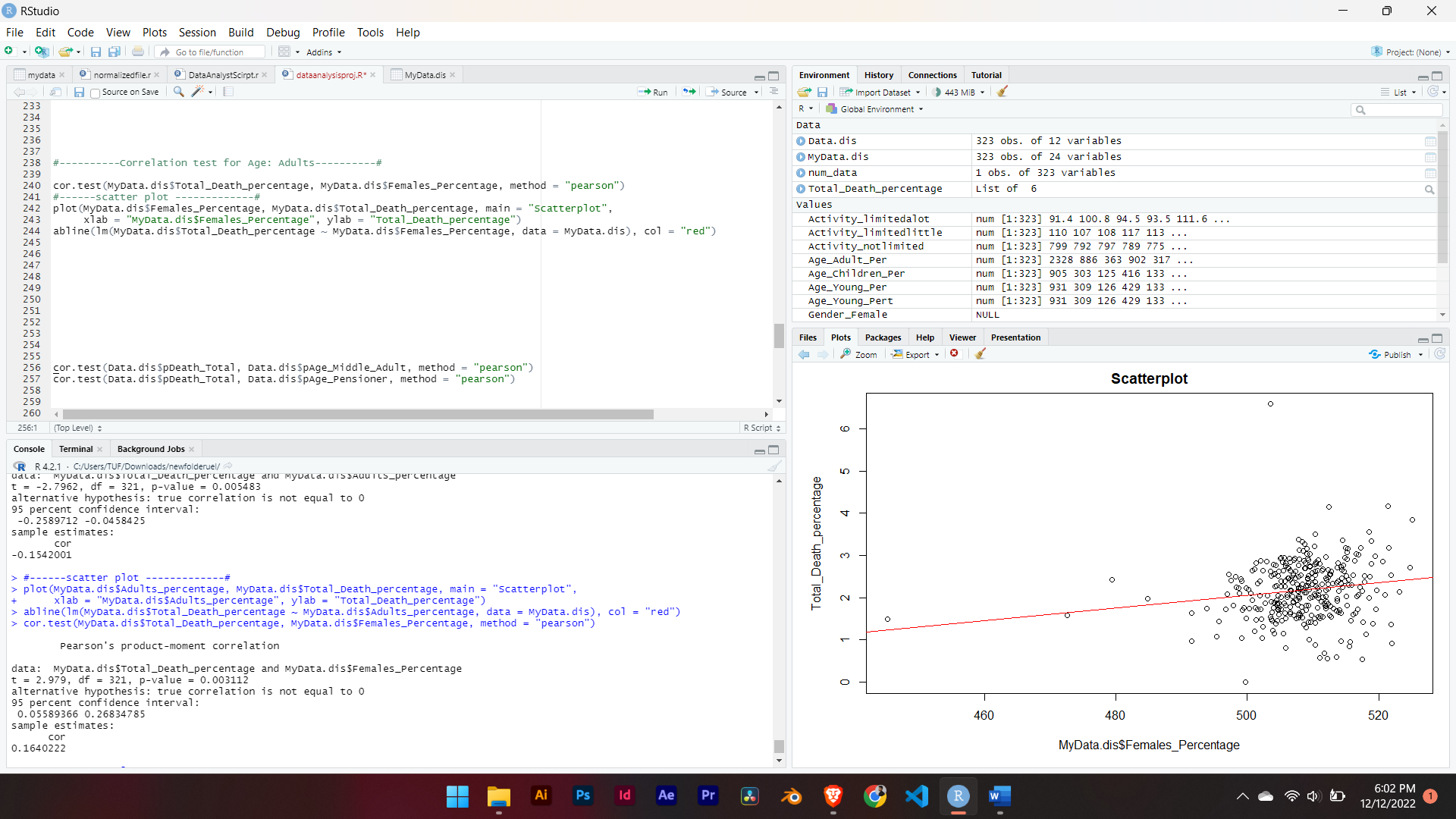


**Fig (2.2c): Scatter plot for the age: Adults and total\_death**

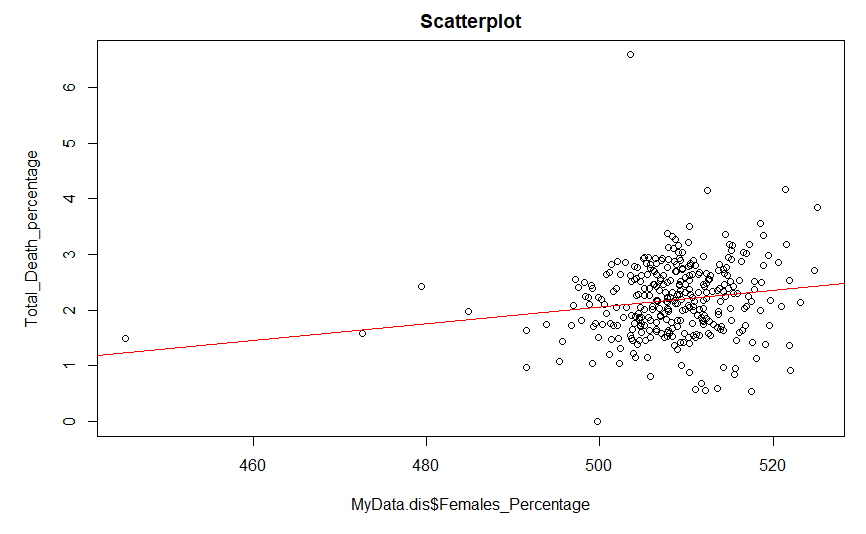
From the figure above Fig (2.3a), we can see that it shows weak negative correlations and p-value of 0.005483 which is less than 0.05, thus we reject the null hypothesis and it is not statistically significant.

##### **With independent variables Gender**

###### **With Gender: Females**

****

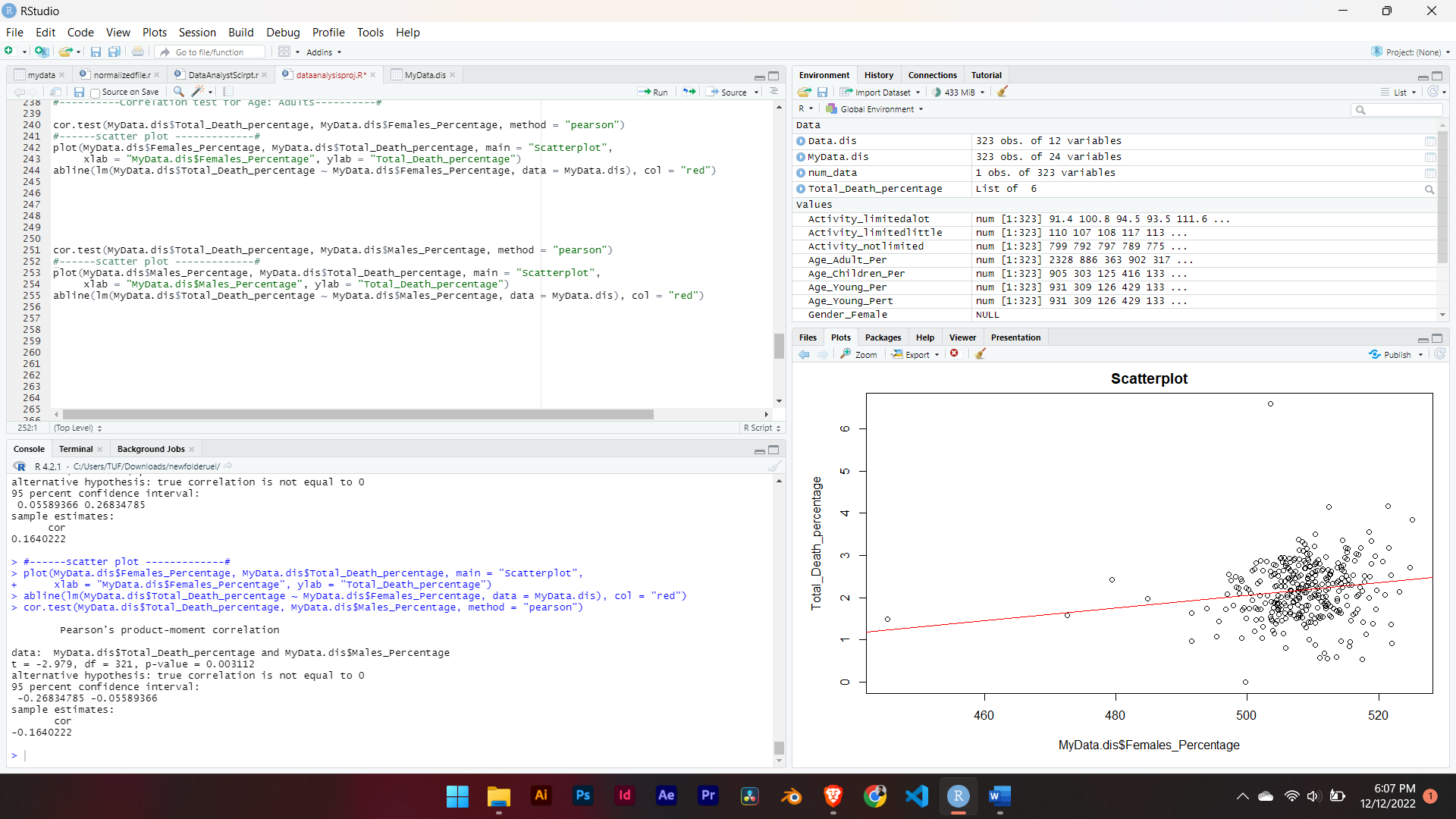
**Fig (2.3a):** **Pearson correlation calculation between total death and Females**



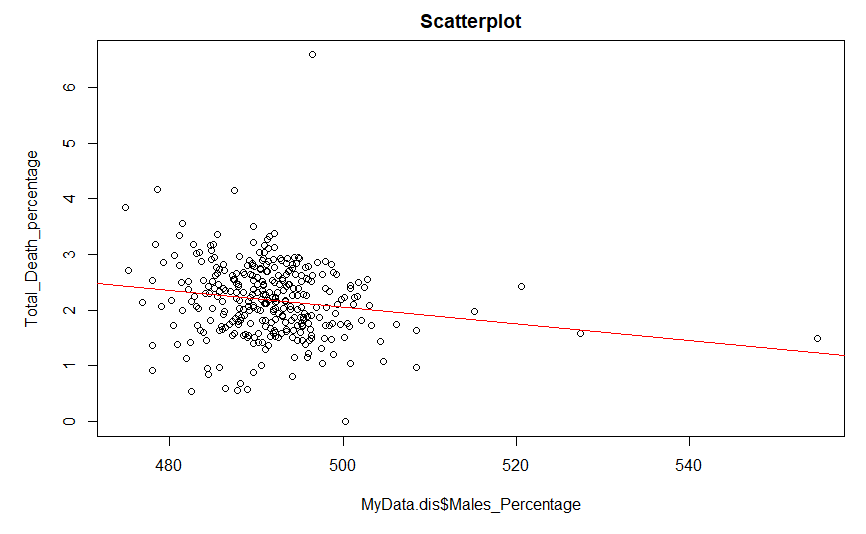
**Fig (3.1b): Scatter-Plot for the Gender: Females\_percentage**

From the figure above Fig (2.3a), we can see that it shows positive correlations and but still the p-value is less than 0.05.

###### **With Gender: Males**

****

**Fig (2.4a): Pearson correlation between Total\_Death\_per and Males**

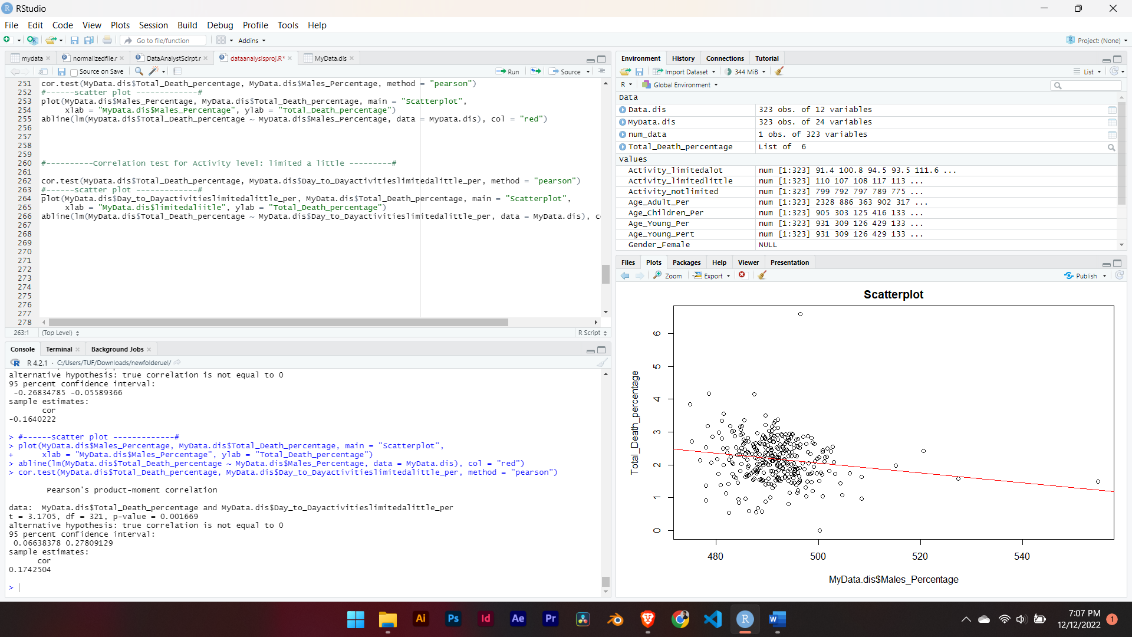


**Fig (2.4b): Scatter-Plot for the Gender: Males\_percentage**

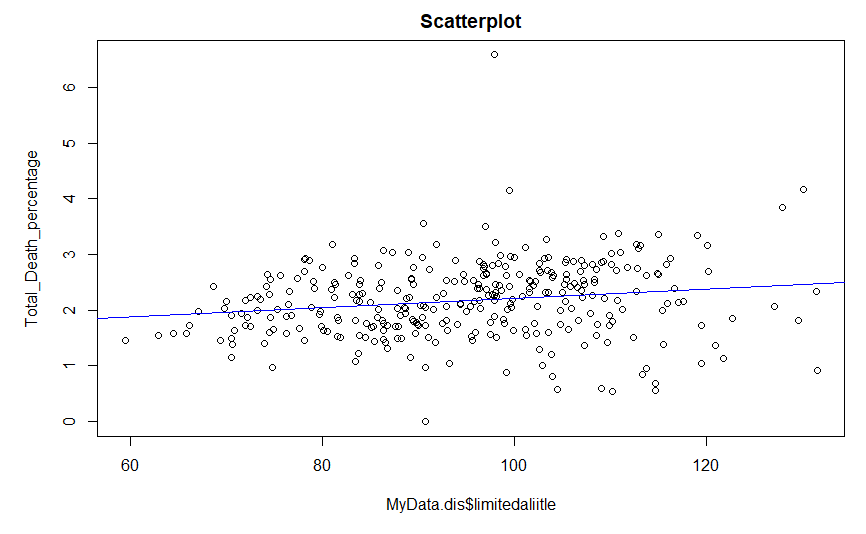
From the figure above **Fig (2.4a),** we can see that it shows negative correlations and p-value of 0.003112 which is less than 0.05, thus we reject the null hypothesis and it is not statistically significant.

##### **With Day-To-Day activity level**

###### **With Day-To-Day activity limited little:**

****

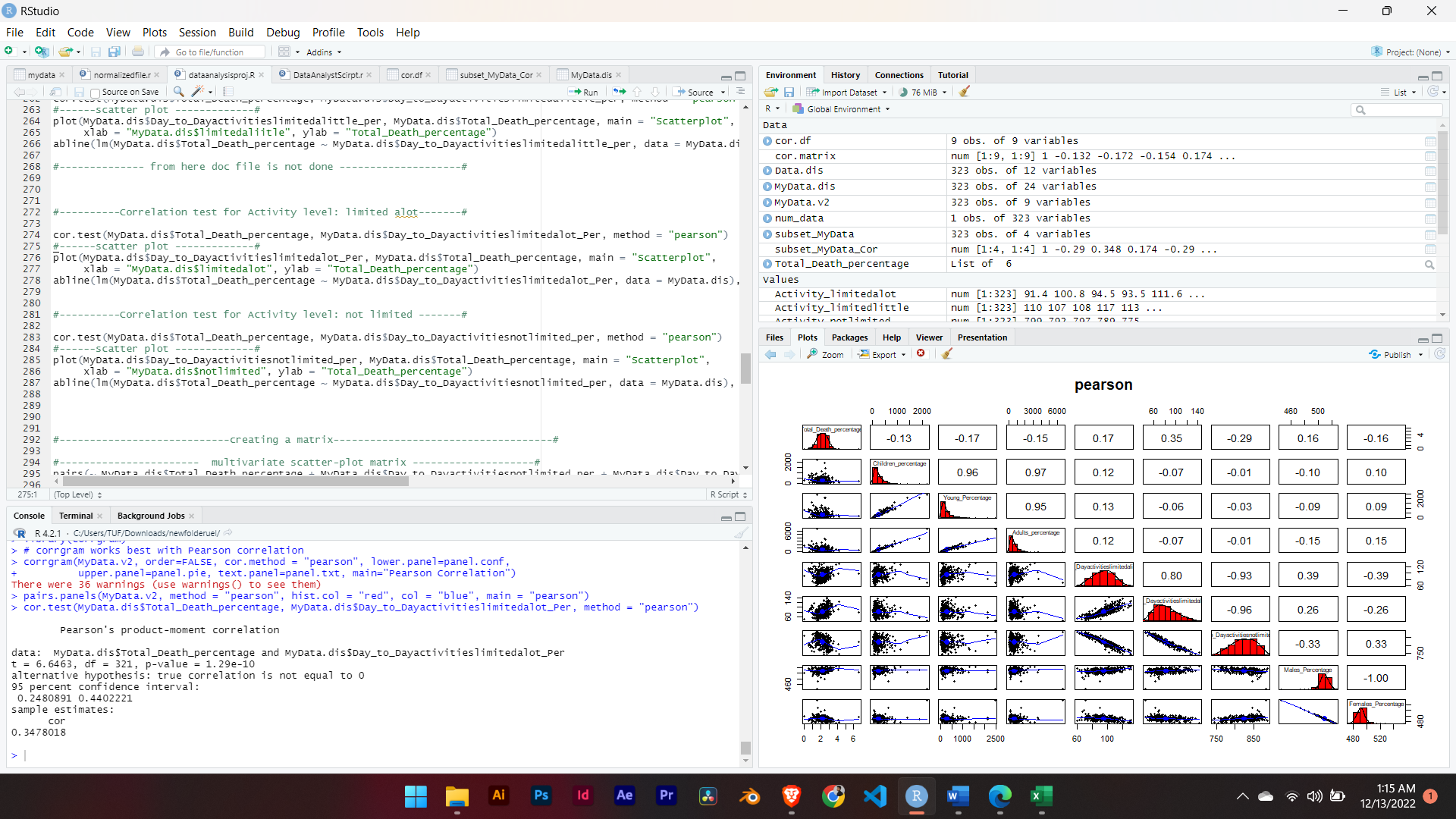
**Fig (2.5a): Pearson correlation between Total\_Death\_per and activity limited little**



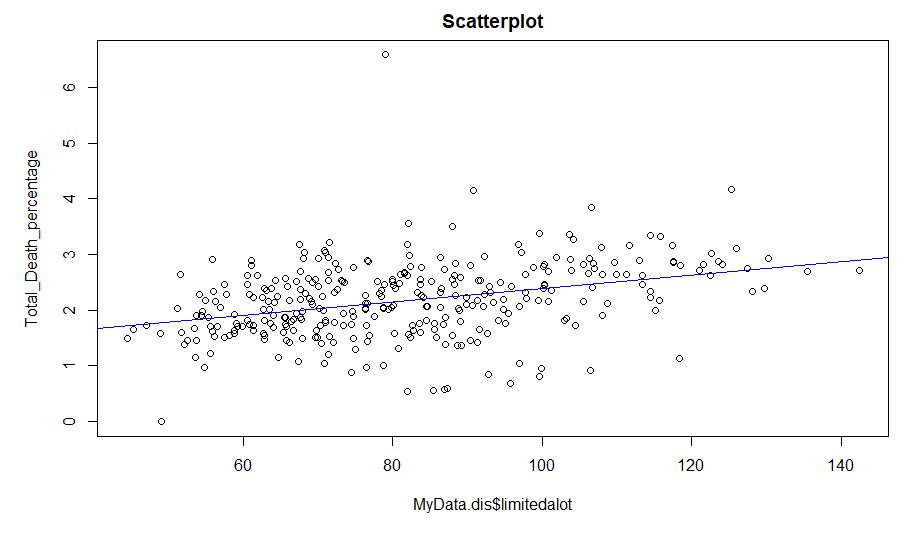
**Fig (4.1b): Scatter-Plot for the activity limited a little**

From the figure **Fig (2.5a)** we can see that it has positive correlations with Total\_Death\_per but has low p-value.

###### **With Day-To-Day activities limited a lot**

****

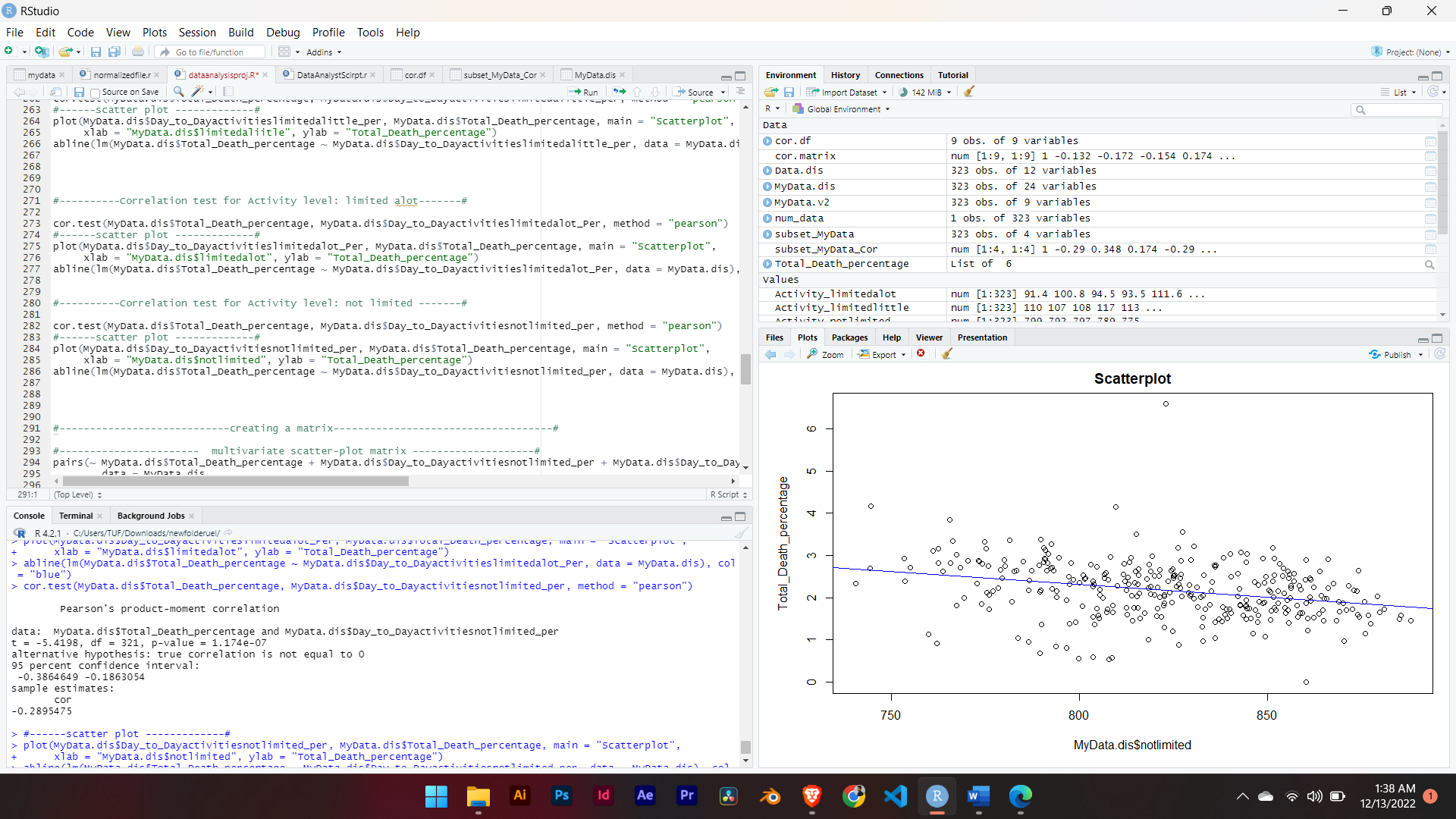
**Fig (2.6a): Pearson correlation between Total\_Death\_per and activity limited a lot**



**Fig (2.6b): Scatter-Plot for variable activity limited a lot**

From the figure **Fig (2.5a)** we can see that it has strong positive correlations with Total\_Death\_per.

###### **With Day-To-Day activity not limited**

****

**Fig (2.7a): Pearson correlation between Total\_Death and activities not limited**



**Fig (2.7b): Scatter-plot for Day-To-Day activities not limited**

From the above figures **Fig (2.7a)** we can see that there is strong negative correlation between **Total\_Death\_percentage** and Day-To-Day **activities not limited**.

##### **Multivariate scatter plot for activity levels**

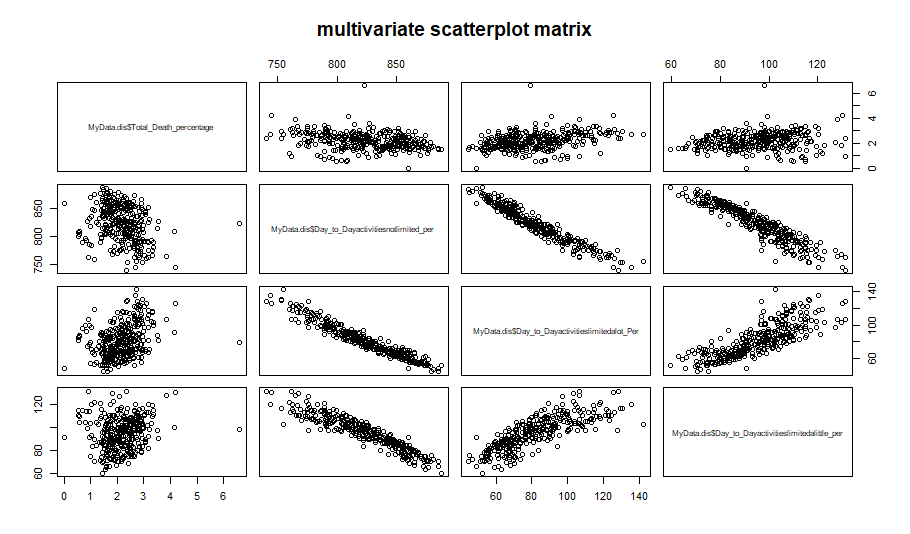
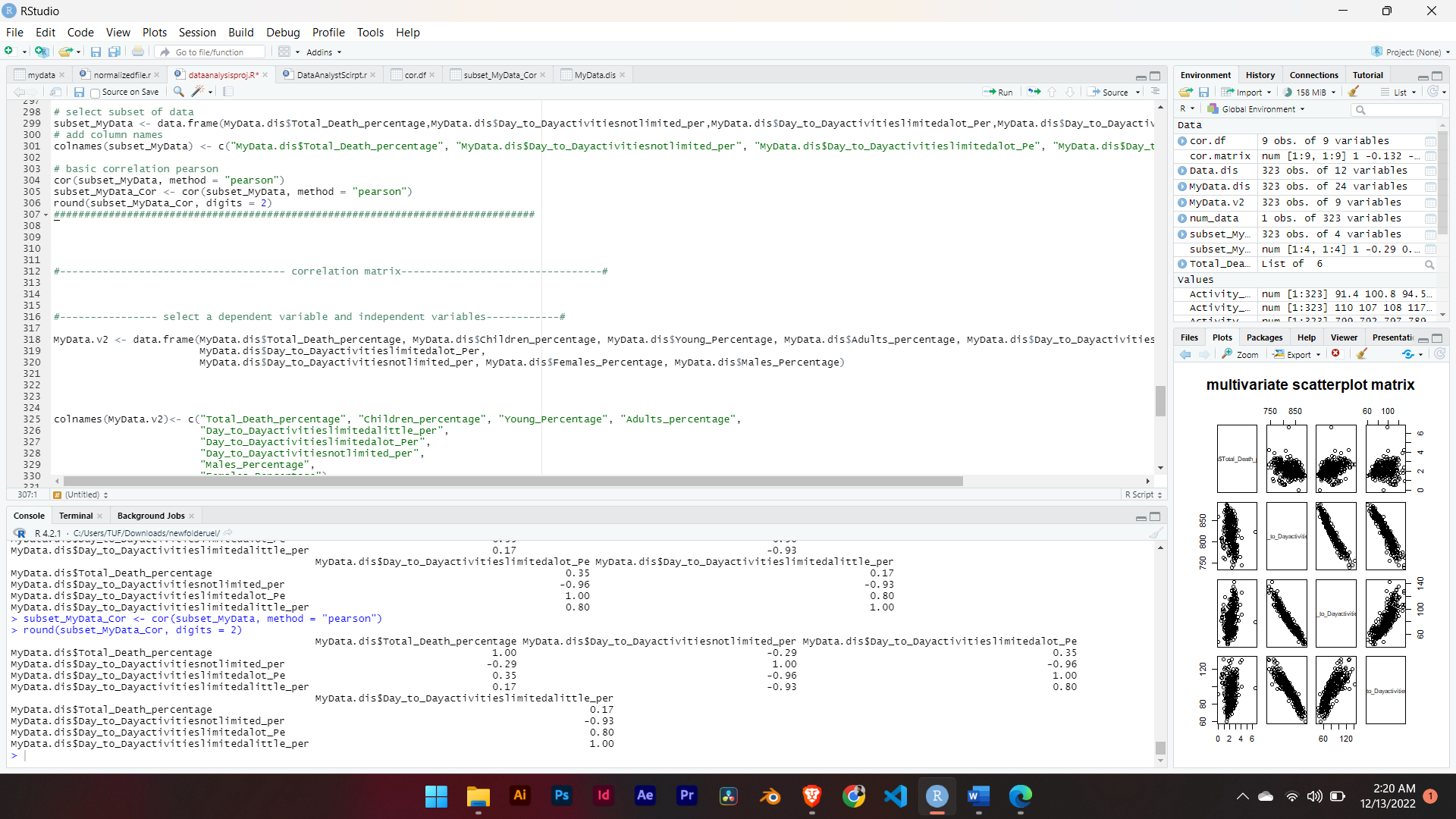


Fig (3.1): Multivariate scatterplot matrix for activity level

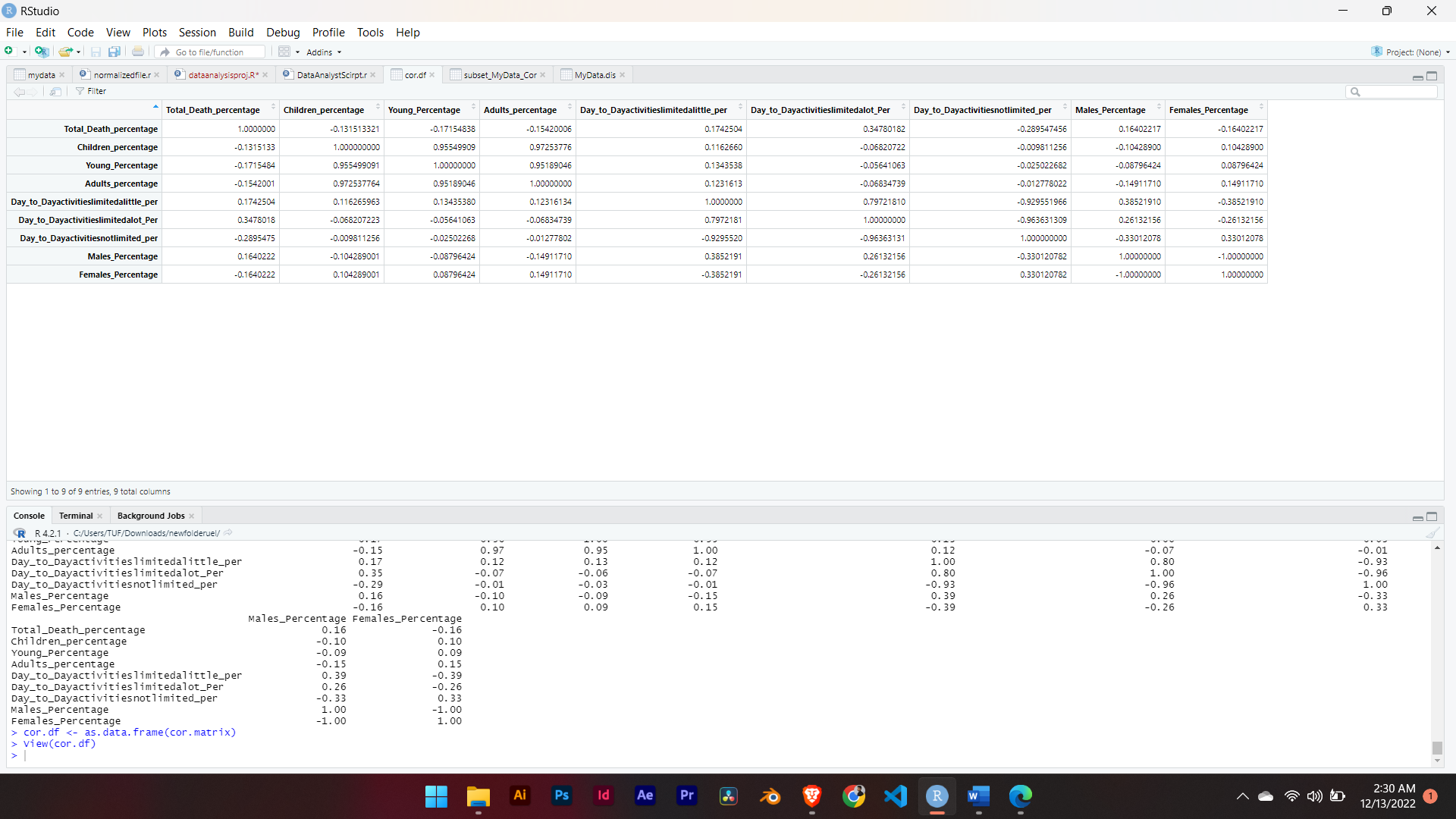
Now we are going to make subset of data from whole data set for activity level and find correlations among those variables.



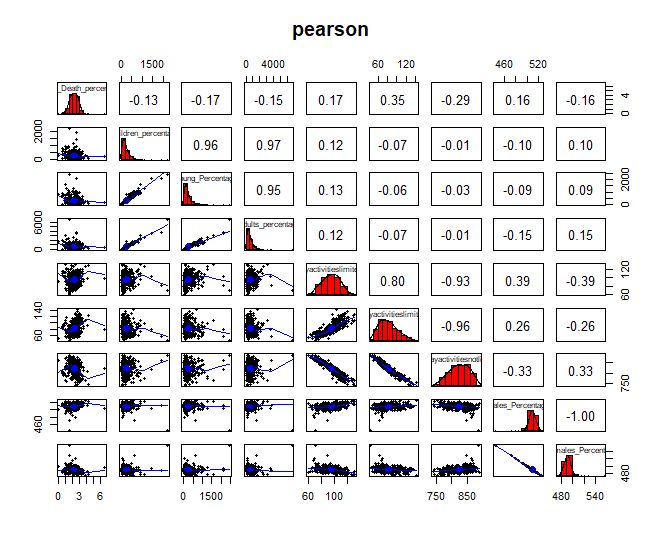
**Fig (3.2): Correlations among independent variables**

We are going to calculate the correlation between dependent and independent variables and construct a correlation matrix using Pearson and spearman separately.

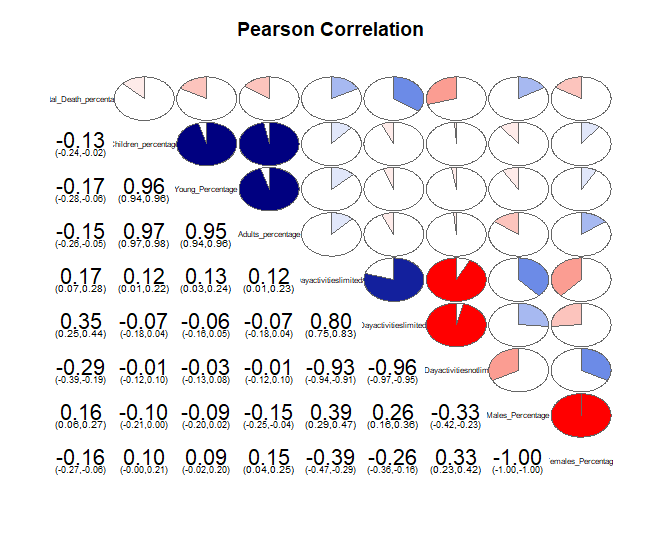
**Using Pearson method**



**Fig (3.3): Correlation matrix using Pearson method**



**Fig (3.4) Pearson using psych package**

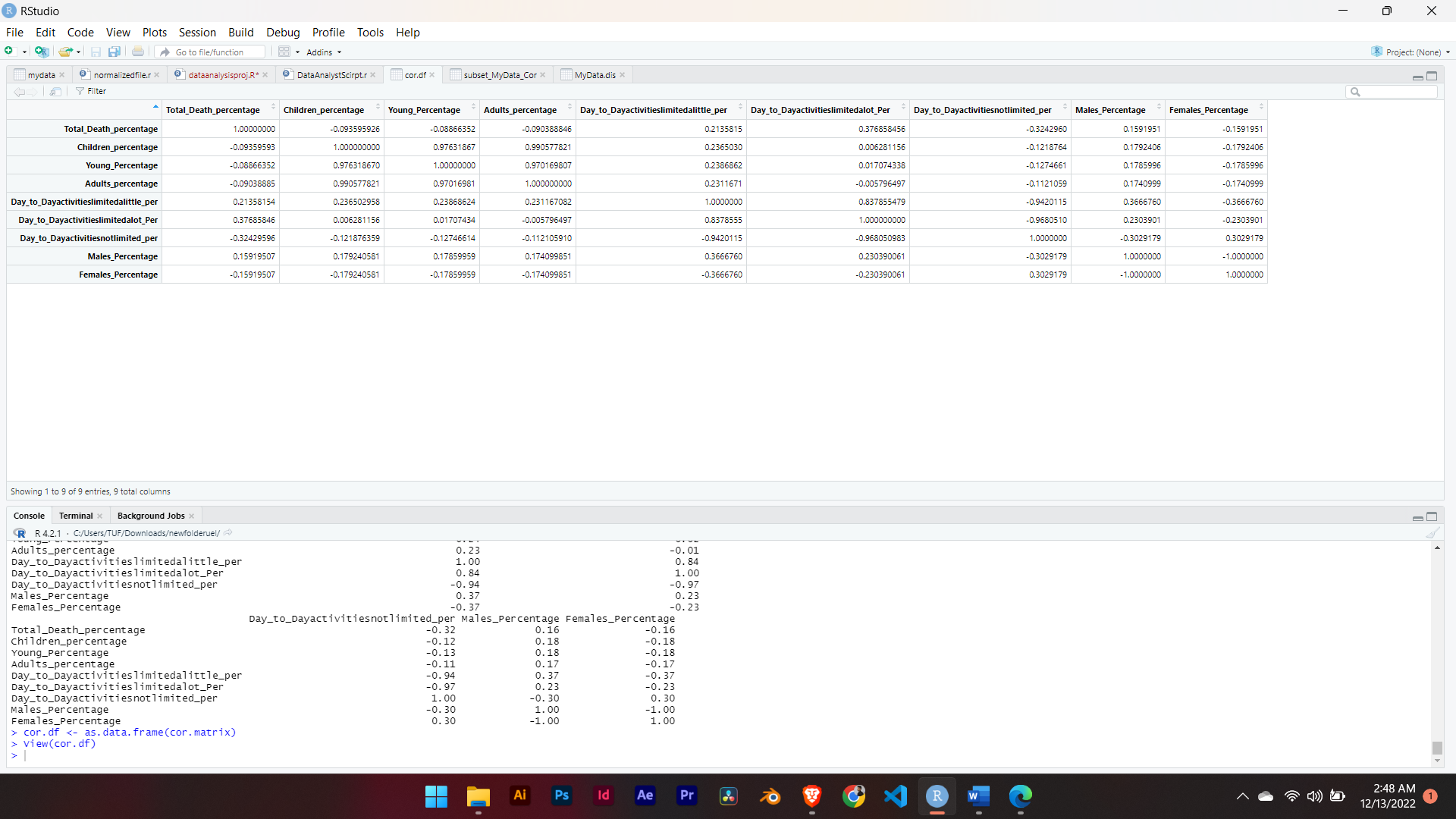


**Fig (3.5): Pearson using corrgram package**

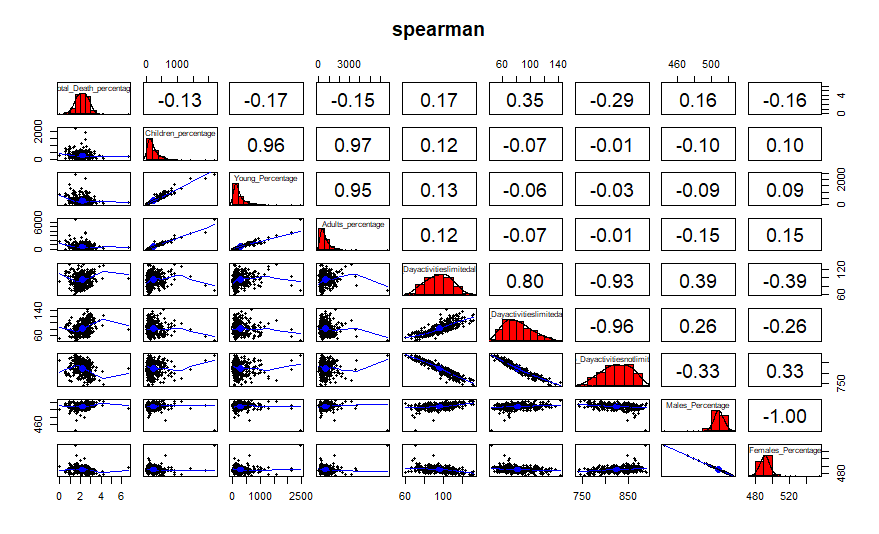
From the above figure **Fig (3.3)** we can see that **Day-To-Day activities are not limited** and have a strong negative correlation with **Total\_Death** which is -0.2895475. We can assume the hypothesis that the person who is more active throughout the day might be in good health and had lower chances of covid death.

Again, on the contrary **Day-To-Day activities are limited\_alot and** have strong positive correlations with **Total\_Death** which is 0.3478018. We can assume the hypothesis that the person whose day-to- day activities are more limited had contribution to covid death.

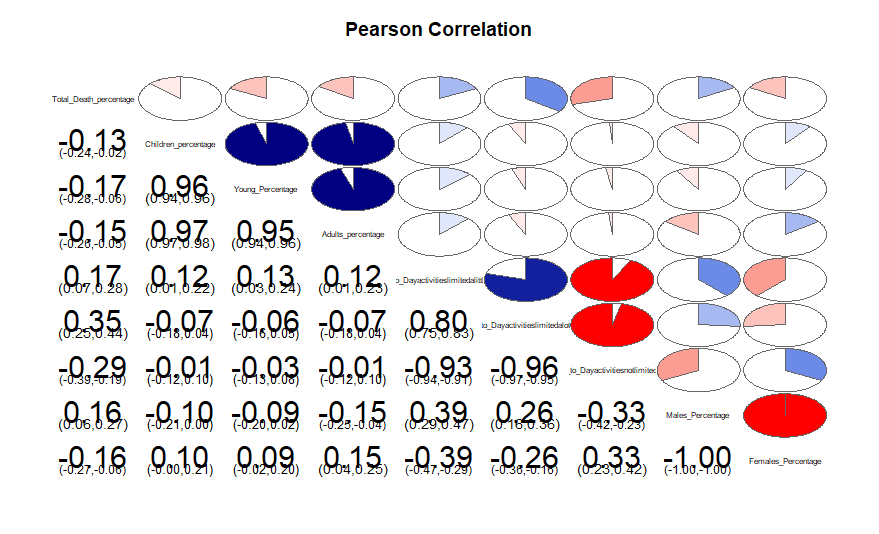
**Using Spearman’s method**

****

**Fig (3.6): Spearman’s correlation matrix**



**Fig (3.7): Spearman’s Correlation using psych package**



**Fig (3.8): Spearman’s Correlation matrix using corrgram package**

From the above figures: **Fig (3.6), Fig (3.7) and Fig (3.8),** we can see the same analysis as we have found from Pearson’s method.

## **Appendix I**

### **SQLite Code**

*Create table finaltable as*

*(Select age.geography\_code, age.geography,*

*COVID-19\_deaths1.Total, age.Children,*

*age.Young, age.Adults, age.Pensioners,*

*Gender.Females, Gender.Males, activities.Day\_to\_Dayactivitiesnotlimited, activities.Day\_to\_Dayactivitieslimitedlittle, activities.Day\_to\_Dayactivitieslimitedalot.*

*from COVID-19\_deaths1 inner join age on COVID-19\_deaths1.LA\_code = age.geography\_code*

*inner join activities on COVID-19\_deaths1.LA\_code = activities.geography\_code inner join Gender on COVID-19\_deaths1.LA\_code = Gender.geography\_code)*

### **R Source Code**

*#----- Read the csv file and print head -------------------------------------------#*

*# set working directory*

*setwd(dirname(file.choose()))*

*getwd()*

*# read in data from csv file*

*MyData.dis <- read.csv("draftcsv01.csv", stringsAsFactors = FALSE)*

*head(MyData.dis)*

*str(MyData.dis)*

*colnames(MyData.dis)*

*summary(MyData.dis)*

*#----------- missing data calculation ----------------------------#*

*apply(MyData.dis, margin=2, Fun = function(x) sum(is.na(x)))*

*library(Amelia)*

*missmap(MyData.dis, col = c("black", "grey"), legend = FALSE)*

*# creating percentage of all variables*

*#------- Dependent variable: Total\_Death columns ---------------------#*

*#percentage calculations*

*MyData.dis <- within (MyData.dis, Total\_Death\_percentage <- (Total/ All\_Usual\_Residents)\*1000)*

*#----- boxplot ------------#*

*Total\_Death\_percentage<- boxplot(MyData.dis$Total\_Death\_percentage, xlab="Total\_Death\_Percentage", ylab="Count")*

*MyData.dis[13]*

*#------------------Normality checking for Total\_death column-------#*

*boxplot(log(MyData.dis[13]+1))*

*summary(log(MyData.dis[13]+1))*

*hist(MyData.dis$Total\_Death\_percentage)*

*# Plot histogram and normal approximation*

*library(rcompanion)*

*plotNormalHistogram(MyData.dis$Total\_Death\_percentage, main = "Histogram", xlab = "MyData.dis$Total\_Death\_percentage")*

*#-----Q-Q plot for dependent variable: Total\_Death\_percentage --------#*

*qqnorm(MyData.dis$Total\_Death\_percentage, xlab = "Total\_Death\_percentage")*

*qqline(MyData.dis$Total\_Death\_percentage, col=20) # red color*

*#---- Significance testing for normality*

*#---- Kolmogorov-Smirnov Tests of normality*

*ks.test(MyData.dis$Total\_Death\_percentage,"pnorm", mean(MyData.dis$Total\_Death\_percentage), sd(MyData.dis$Total\_Death\_percentage))*

*#------------------ For Independent Variables --------------------------------#*

*### calculations of percentage of independent variables*

*MyData.dis <- within (MyData.dis, Children\_percentage <- (Children/ All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Young\_Percentage <- (Young/ All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Adults\_percentage <- (Adults/ All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Day\_to\_Dayactivitieslimitedalittle\_per <- (Day\_to\_Dayactivitieslimitedalittle/ All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Day\_to\_Dayactivitieslimitedalot\_Per <- (Day\_to\_Dayactivitieslimitedalot/ All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Day\_to\_Dayactivitiesnotlimited\_per <- ( Day\_to\_Dayactivitiesnotlimited / All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Males\_Percentage <- (Males/ All\_Usual\_Residents)\*1000)*

*MyData.dis <- within (MyData.dis, Females\_Percentage <- (Females/ All\_Usual\_Residents)\*1000)*

*#----------------- Creating boxplot qq plot for normalizations for different independent variables --------------#*

*#Age: Children, young, Adults*

*#assign names for vairable Age category*

*Age\_Children\_Per <- MyData.dis$Children\_percentage*

*Age\_Young\_Per <- MyData.dis$Young\_Percentage*

*Age\_Adult\_Per <- MyData.dis$Adults\_percentage*

*#-----------------boxplot for Age groups ----------------------------#*

*boxplot(Age\_Children\_Per,Age\_Young\_Pert,Age\_Adult\_Per,*

*names=c("Age\_Children", "Age\_Young", "Age\_Adult"),*

*xlab="Age", ylab="Count", col = "Bisque")*

*#-----------Q-Q plots--------------------#*

*qqnorm(Age\_Children\_Per, xlab = "Age\_Children\_Per")*

*qqline(Age\_Children\_Per, col=21) # red color*

*qqnorm(Age\_Young\_Per, xlab = "Age\_Young\_Per")*

*qqline(Age\_Young\_Per, col=24) # red color*

*qqnorm(Age\_Adult\_Per, xlab = "Age\_Adult\_Per")*

*qqline(Age\_Adult\_Per, col=27) # red color*

*#------ boxplot for variables Gender: Males and Females------#*

*Gender\_male\_per <- MyData.dis$Males\_percentage*

*Gender\_Female\_per <- MyData.dis$Females\_percentage*

*boxplot(Gender\_male\_per,Gender\_Female\_per,*

*names=c("Gender:Male", "Gender:Female"),*

*xlab="Gender", ylab="Count", col = "Bisque")*

*#-----Q-Q plot for variable Gender----#*

*qqnorm(Gender\_male\_per, xlab = "Gender\_Male\_per")*

*qqline(Gender\_male\_per, col=20) # red color*

*qqnorm(Gender\_Female\_per, xlab = "Gender\_Female\_per")*

*qqline(Gender\_Female\_per, col=23)# red color*

*#------ boxplot for variables Activity level : Activity limited little, alot and not limited ------#*

*Activity\_limitedlittle<- MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per*

*Activity\_limitedalot<-MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per*

*Activity\_notlimited <-MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per*

*boxplot(Activity\_limitedlittle,Activity\_limitedalot,Activity\_notlimited,*

*names=c("Activity:little", "Activity:alot", "Activity:notlimited"),xlab="Activity", ylab="Count", col="Bisque")*

*#-------------- Q-Q plots -------------------#*

*qqnorm(Activity\_limitedlittle, xlab = "Activity\_limitedlittle")*

*qqline(Activity\_limitedlittle, col=22)# red color*

*qqnorm(Activity\_limitedalot, xlab = "Activity\_limitedalot")*

*qqline(Activity\_limitedalot, col=24)*

*qqnorm(Activity\_notlimited, xlab = "Activity\_notlimited")*

*qqline(Activity\_notlimited, col=28)*

*#-------------------------------------------------------------------------------------------#*

*#-----------------------Pearson method--------------------------------------------#*

*#cor.test(x, y,*

*# alternative = c("two.sided", "less", "greater"),*

*# method = c("pearson", "kendall", "spearman"),*

*# exact = NULL, conf.level = 0.95, continuity = FALSE, ...)*

*# test correlation of dependent variable with all independent variables*

*#----------Correlation test for Age: Children ---------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Children\_percentage, method="pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Children\_percentage, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$Children\_percentage", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Children\_percentage, data = MyData.dis), col = "blue")*

*#----------Correlation test for Age: Young -----------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Young\_Percentage, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Young\_Percentage, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$Young\_Percentage", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Young\_Percentage, data = MyData.dis), col = "blue")*

*#----------Correlation test for Age: Adults----------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Adults\_percentage, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Adults\_percentage, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$Adults\_percentage", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Adults\_percentage, data = MyData.dis), col = "red")*

*#----------Correlation test for Age: Adults----------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Females\_Percentage, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Females\_Percentage, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$Females\_Percentage", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Females\_Percentage, data = MyData.dis), col = "red")*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Males\_Percentage, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Males\_Percentage, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$Males\_Percentage", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Males\_Percentage, data = MyData.dis), col = "red")*

*#----------Correlation test for Activity level: limited a little ---------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$limitedaliitle", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per, data = MyData.dis), col = "blue")*

*#----------Correlation test for Activity level: limited alot-------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$limitedalot", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per, data = MyData.dis), col = "blue")*

*#----------Correlation test for Activity level: not limited -------#*

*cor.test(MyData.dis$Total\_Death\_percentage, MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per, method = "pearson")*

*#------scatter plot -------------#*

*plot(MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per, MyData.dis$Total\_Death\_percentage, main = "Scatterplot",*

*xlab = "MyData.dis$notlimited", ylab = "Total\_Death\_percentage")*

*abline(lm(MyData.dis$Total\_Death\_percentage ~ MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per, data = MyData.dis), col = "blue")*

*#----------------------------creating a matrix------------------------------------#*

*#----------------------- multivariate scatter-plot matrix --------------------#*

*pairs(~ MyData.dis$Total\_Death\_percentage + MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per + MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per + MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per*

*, data = MyData.dis,*

*main = "multivariate scatterplot matrix")*

*# select subset of data*

*subset\_MyData <- data.frame(MyData.dis$Total\_Death\_percentage,MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per,MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per,MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per)*

*# add column names*

*colnames(subset\_MyData) <- c("MyData.dis$Total\_Death\_percentage", "MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per", "MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Pe", "MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per")*

*# basic correlation pearson*

*cor(subset\_MyData, method = "pearson")*

*subset\_MyData\_Cor <- cor(subset\_MyData, method = "pearson")*

*round(subset\_MyData\_Cor, digits = 2)*

*###############################################################################*

*#------------------------------------- correlation matrix---------------------------------#*

*#---------------- select a dependent variable and independent variables------------#*

*MyData.v2 <- data.frame(MyData.dis$Total\_Death\_percentage, MyData.dis$Children\_percentage, MyData.dis$Young\_Percentage, MyData.dis$Adults\_percentage, MyData.dis$Day\_to\_Dayactivitieslimitedalittle\_per,*

*MyData.dis$Day\_to\_Dayactivitieslimitedalot\_Per,*

*MyData.dis$Day\_to\_Dayactivitiesnotlimited\_per, MyData.dis$Females\_Percentage, MyData.dis$Males\_Percentage)*

*colnames(MyData.v2)<- c("Total\_Death\_percentage", "Children\_percentage", "Young\_Percentage", "Adults\_percentage",*

*"Day\_to\_Dayactivitieslimitedalittle\_per",*

*"Day\_to\_Dayactivitieslimitedalot\_Per",*

*"Day\_to\_Dayactivitiesnotlimited\_per",*

*"Males\_Percentage",*

*"Females\_Percentage")*

*#----Correlations between numeric variables using pearson -----#*

*cor.matrix <- cor(MyData.v2, use = "pairwise.complete.obs", method = "pearson")*

*round(cor.matrix, digits = 2)*

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

*View(cor.df)*

*library(psych)*

*pairs.panels(MyData.v2, method = "pearson", hist.col = "red", col = "blue", main = "pearson")*

*library(corrgram)*

*# corrgram*

*corrgram(MyData.v2, order=FALSE, cor.method = "pearson", lower.panel=panel.conf,*

*upper.panel=panel.pie, text.panel=panel.txt, main="Pearson Correlation")*

*#------------------------------------------------------------------------------------#*

*#----Correlations between numeric variables using Spearman -------#*

*cor.matrix <- cor(MyData.v2, use = "pairwise.complete.obs", method = "spearman")*

*round(cor.matrix, digits = 2)*

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

*View(cor.df)*

*library(psych)*

*pairs.panels(MyData.v2, method = "pearson", hist.col = "red", col = "blue", main = "spearman")*

*library(corrgram)*

*# corrgram works best with Pearson correlation*

*corrgram(MyData.v2, order=FALSE, cor.method = "pearson", lower.panel=panel.conf,*

*upper.panel=panel.pie, text.panel=panel.txt, main="Pearson Correlation")*