Group 11 - Homework 1

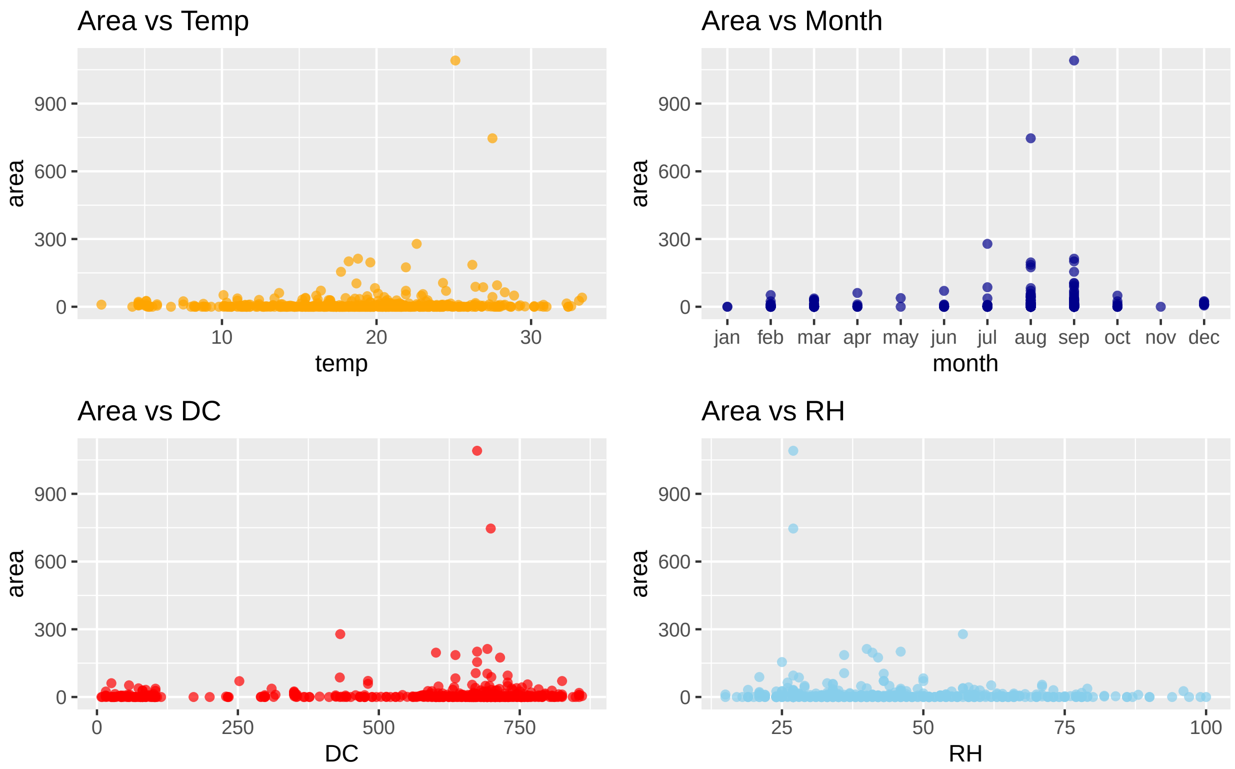
Kumar Sri Chandra Bhaskar Adabala, NUID - 001083381
Abhinash Ambati, NUID - 001023924

Problem 1

Task a

Plot area vs.temp, area vs. month, area vs. DC, area vs. RH for January through December combined in one graph.

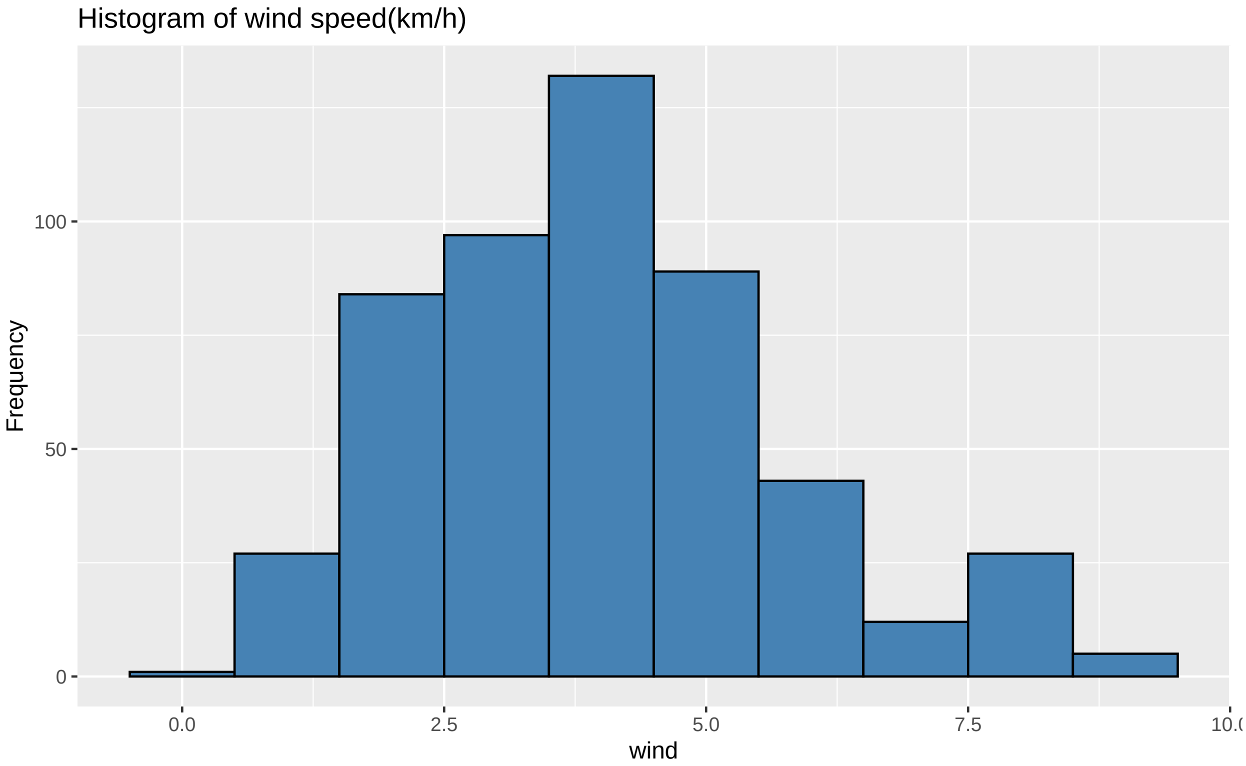
```r
fig1 <- ggplot(forestfires) + geom_point(aes(x = temp, y = area), alpha = 0.7,color='orange') +
  ggtitle("Area vs Temp")
fig2 <- ggplot(forestfires) +
  geom_point(aes(x = factor(month, levels =
                          c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"),
                          ordered = T), y = area), alpha = 0.7, color='darkblue') +
  ggtitle("Area vs Month") +
  xlab("month")
fig3 <- ggplot(forestfires) + geom_point(aes(x = DC, y = area),  alpha = 0.7,color='red') +
  ggtitle("Area vs DC")
fig4 <- ggplot(forestfires) + geom_point(aes(x = RH, y = area),  alpha = 0.7,color='skyblue') +
  ggtitle("Area vs RH")

gridExtra::grid.arrange(fig1, fig2, fig3, fig4, ncol=2)
```



Task b

Plot the histogram of wind speed (km/h).

```
ggplot(forestfires, aes(wind)) +
  geom_histogram(bins = 10, color = 'Black', fill = 'steelblue') + ylab("Frequency") +
  ggtitle("Histogram of wind speed(km/h)")
```



Task c

Compute the summary statistics (min, 1Q, mean, median, 3Q, max,) of part b.

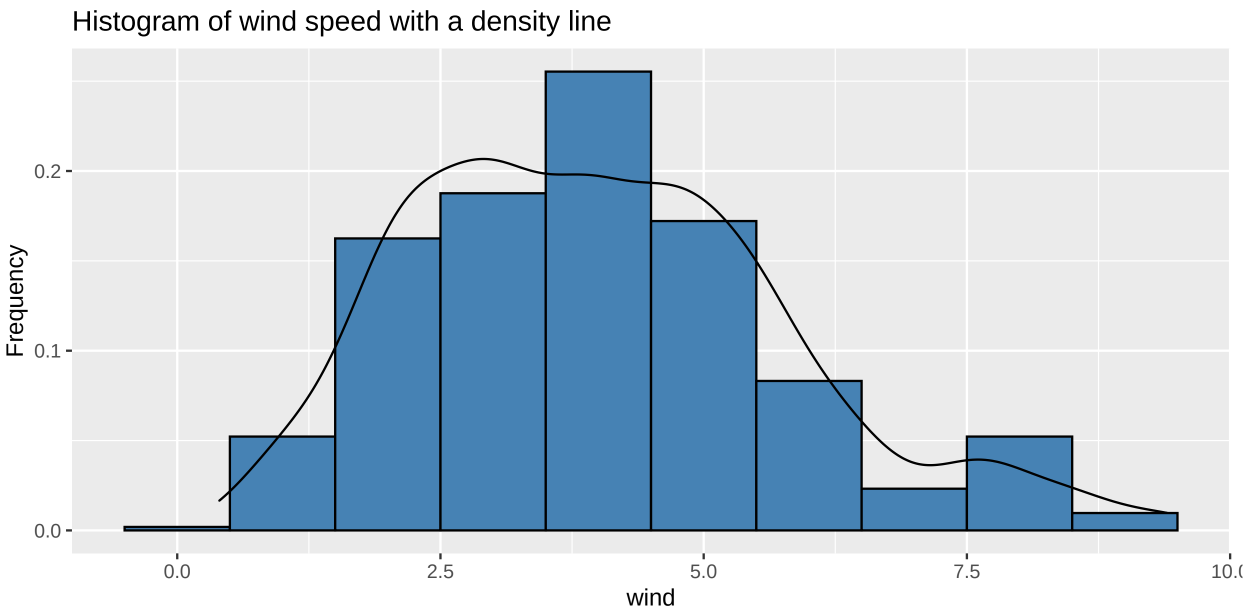
```
summary(forestfires$wind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.400   2.700   4.000   4.018   4.900   9.400
```

Task d

Add a density line to the histogram in part b.

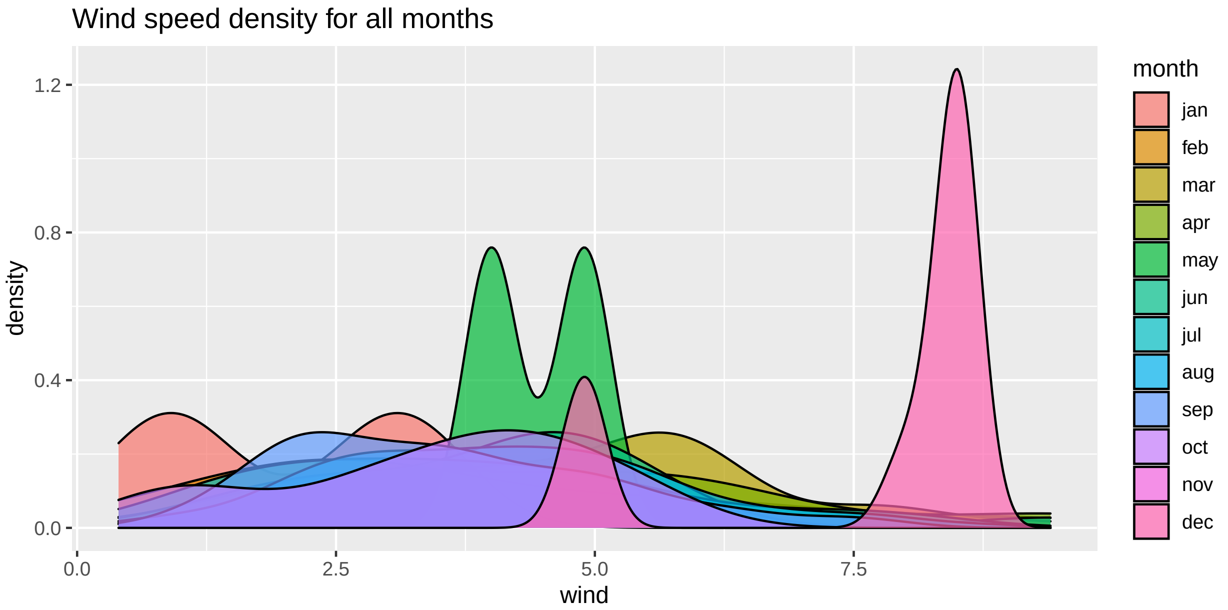
```
ggplot(forestfires,aes(x =wind, y = ..density..))  +
  geom_histogram(bins = 10, color = 'Black', fill = 'steelblue') +
  geom_density() + ylab("Frequency") +
  ggtitle("Histogram of wind speed with a density line")
```



Task e

Plot the wind speed density function of all months in one plot. Use different colors for different months in the graph to interpret your result clearly.

```
forestfires$month <-
  factor(forestfires$month,
         levels = c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"))

ggplot(forestfires) + geom_density(aes(x = wind,fill = month),ordered = T, alpha = 0.7) +
  ggtitle("Wind speed density for all months")
```

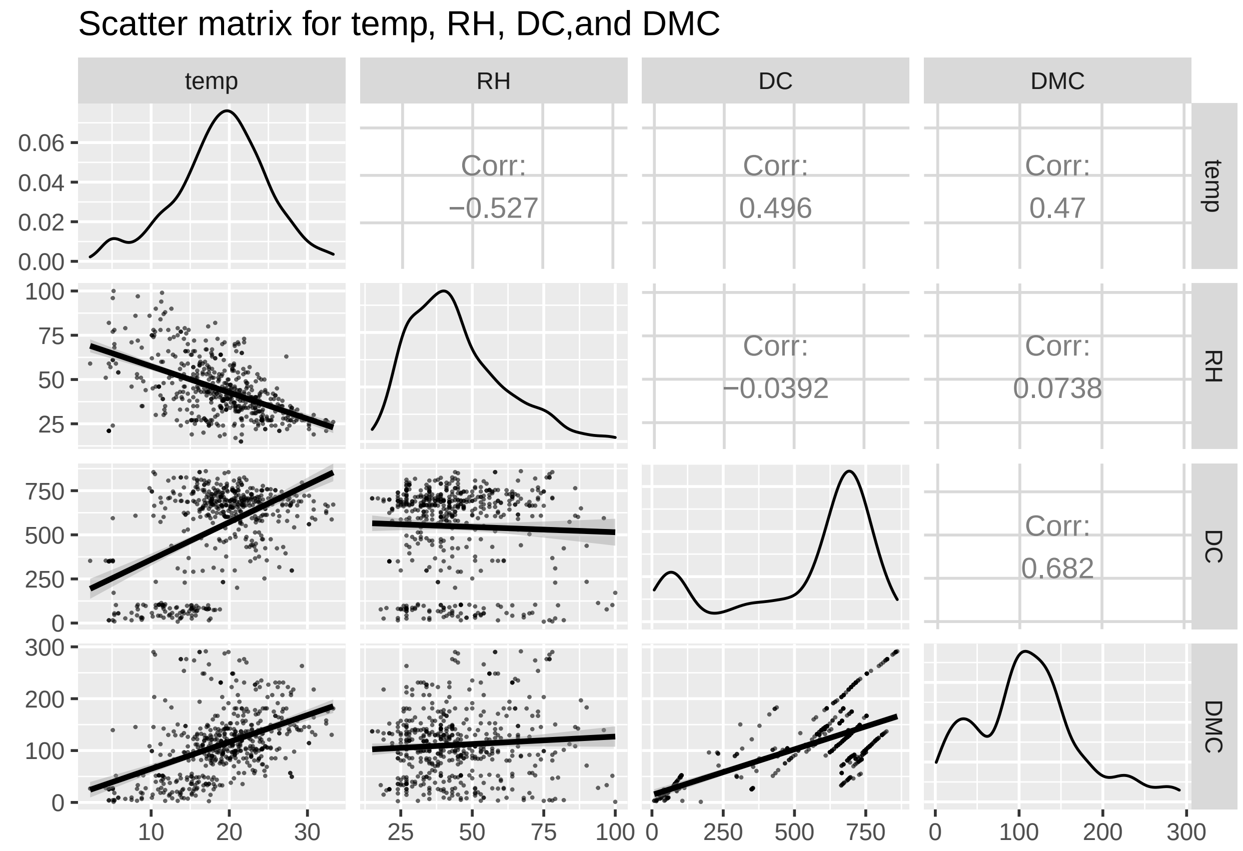


```
#used alpha = 0.7 to make the colors transparent so that we can see all background months densities
```

Task f

Plot the scatter matrix for temp, RH, DC and DMC. How would you interpret the result in terms of correlation among these data?

```r
ggpairs(forestfires[,c(9,10,7,6)], title = "Scatter matrix for temp, RH, DC,and DMC",
        lower = list(continuous = wrap("smooth", alpha = 0.6, size=0.1)))
```

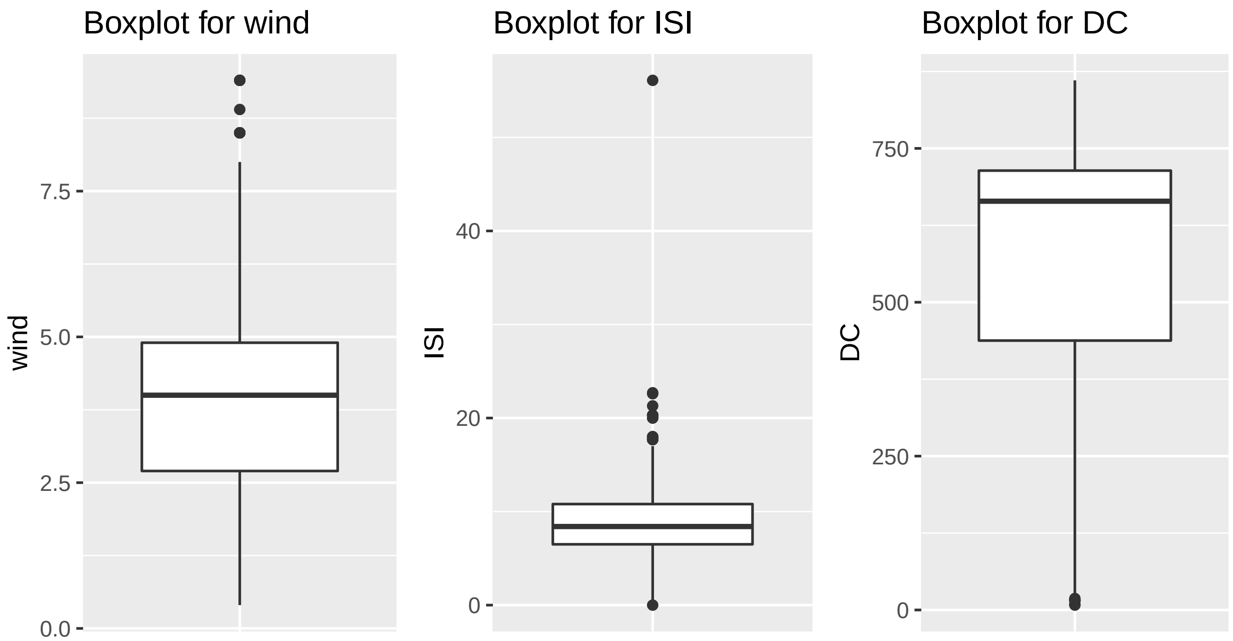


- From above scatter matrix plot we can understand that:
  - There is a downward trend between temp and RH so, we can conclude that there is a negative correlation.
  - There is no relation between temp and DC and it is same between DMC and RH.
  - There is a strong upward trend between DC and DMC so, we can conclude that there is a positive correlation.
  - There is a little upward trend between temp and DMC but cannot consider it as a positive correlation.

Task g

Create boxplot for wind, ISI and DC. Are there any anomalies/outliers? Interpret your result.

```r
fig1 <- ggplot(forestfires) + geom_boxplot(aes(x = "",y = wind)) + xlab("") +
  ggtitle("Boxplot for wind")+
  theme(axis.ticks.x = element_blank())
fig2 <- ggplot(forestfires) + geom_boxplot(aes(x = "",y = ISI))+ xlab("") +
  ggtitle("Boxplot for ISI")+
  theme(axis.ticks.x = element_blank())
fig3 <- ggplot(forestfires) + geom_boxplot(aes(x = "",y = DC))+ xlab("") +
  ggtitle("Boxplot for DC")+
  theme(axis.ticks.x = element_blank())

gridExtra::grid.arrange(fig1,fig2,fig3, ncol=3)
```

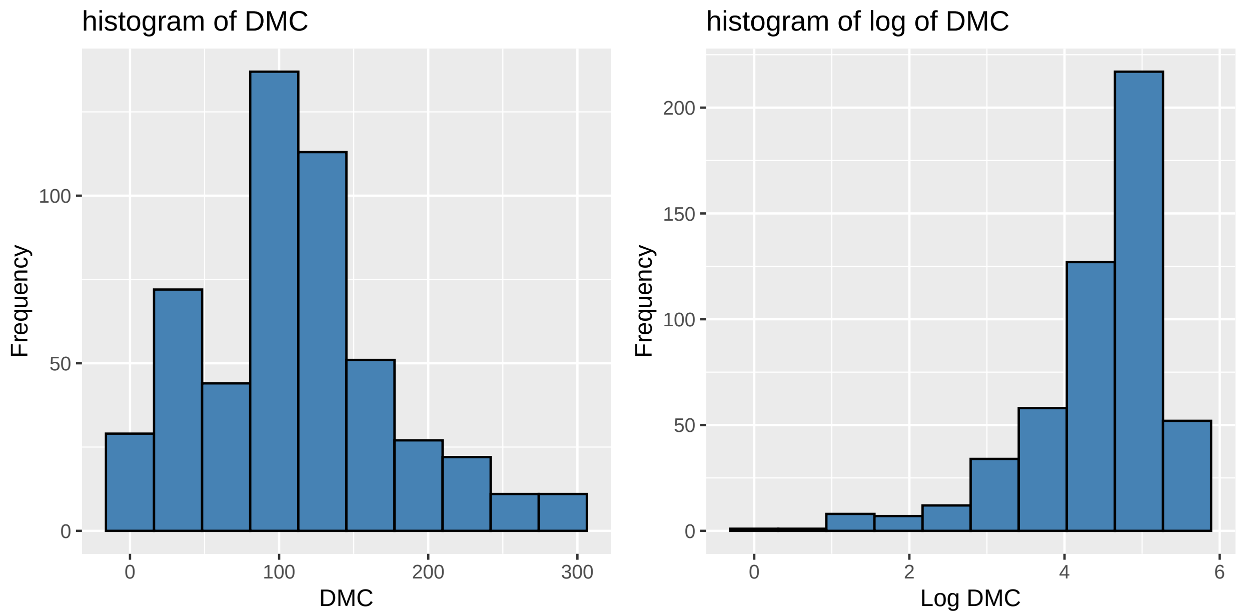


- Interpretation:
  - From the boxplot for wind we can understand that there are few outliers above the maximum value.
  - From the boxplot for ISI we can understand that there are many outliers above the maximum value and few below minimum value.
  - From the boxplot for DC we can understand that there are few outliers below minimum value and the interquartile range is close to maximum value.

Task h

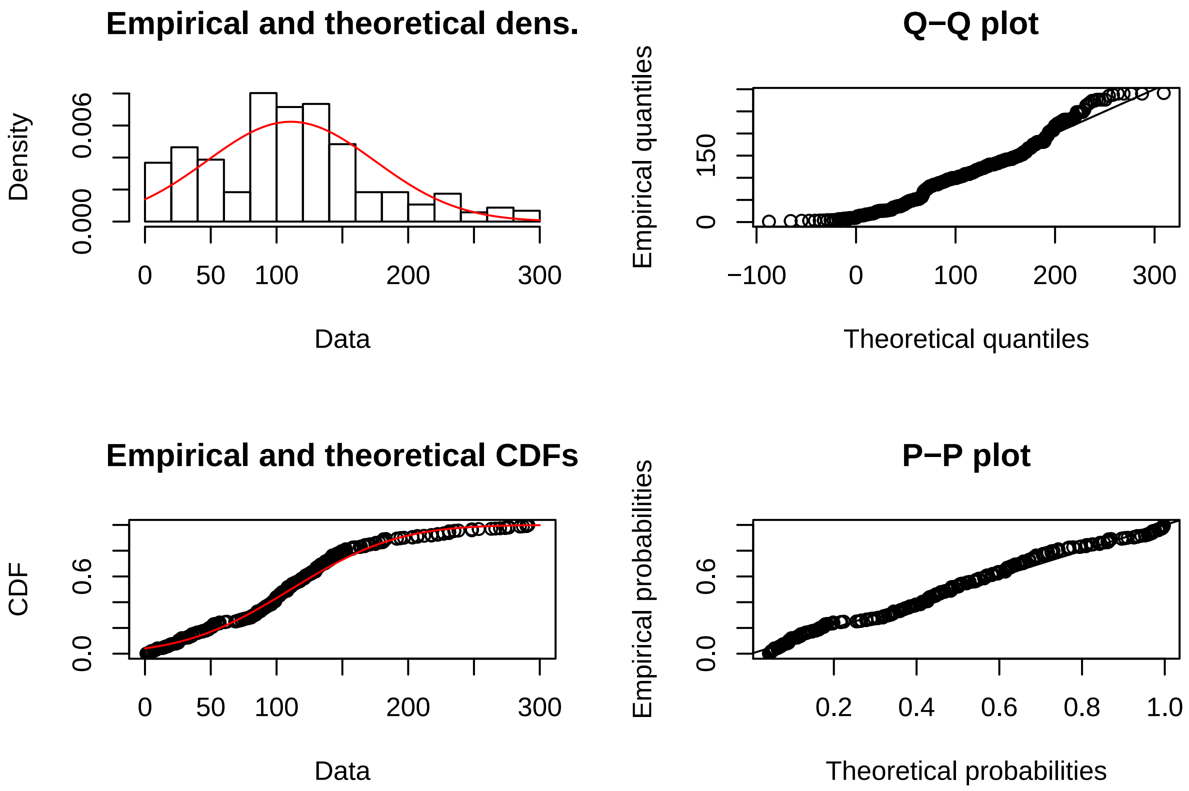
Create the histogram of DMC. Create the histogram of log of DMC. Compare the result and explain your answer.

```
fig1 <- ggplot(forestfires,aes(x =DMC))  +
  geom_histogram(bins = 10, color = 'Black', fill = 'steelblue')+ ylab("Frequency") +
  ggtitle("histogram of DMC")
fig2 <- ggplot(forestfires,aes(x = log(DMC)))  +
  geom_histogram(bins = 10, color = 'Black', fill = 'steelblue') + xlab("Log DMC") + ylab("Frequency")+
  ggtitle("histogram of log of DMC")

gridExtra::grid.arrange(fig1,fig2, ncol=2)
```



From the above histrogram of DMC we can see that the distribution is right skewed and from the histogram of log of DMC we can see that the distribution is left skewed. Let us see whether this is a normal distribution or not.

```
fit_normal <- fitdist(forestfires$DMC,"norm")
plot(fit_normal)
```



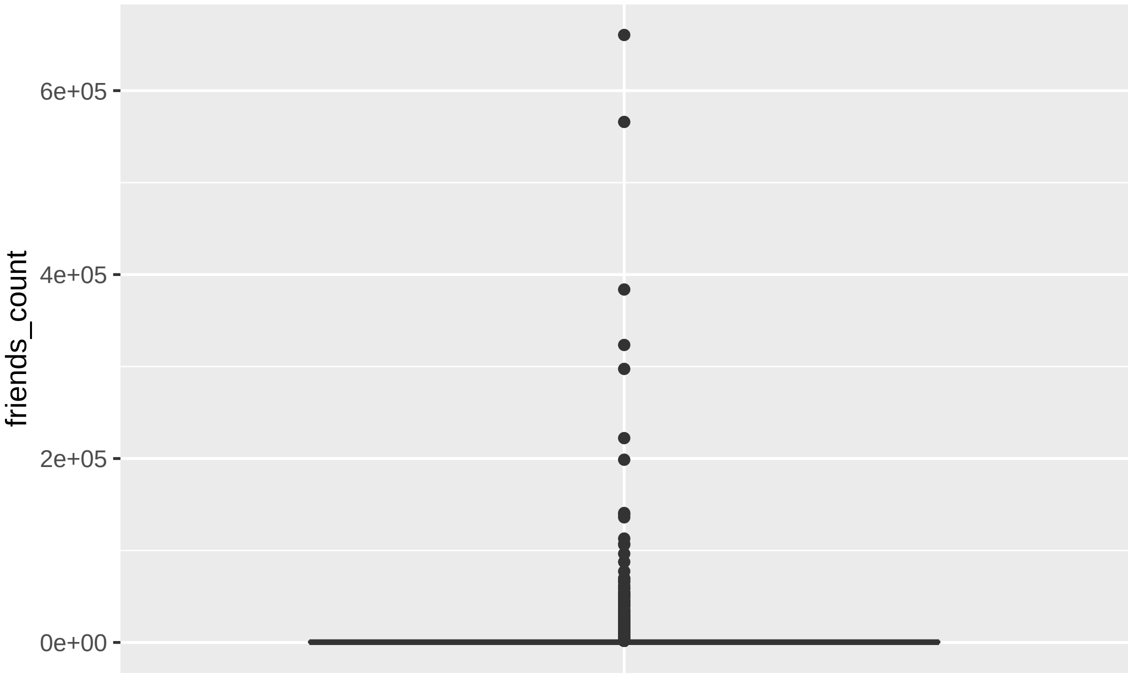
From this we can say that the normal distribution fits well with this DMC data.

Problem 2

Task a

How are the data distributed for friend_count variable?

```
ggplot(M01_quasi_twitter, aes(x= "",y = friends_count)) + geom_boxplot( ) +
  theme(axis.ticks.x = element_blank()) + xlab("")
```



From the above plot, we can understand that the most of the data is in the bottom, and there are too many outliers above the maximum value. Also the difference between outliers is very high.

Task b

Compute the summery statistics (min, 1Q, mean, median, 3Q, max) on friend_count.

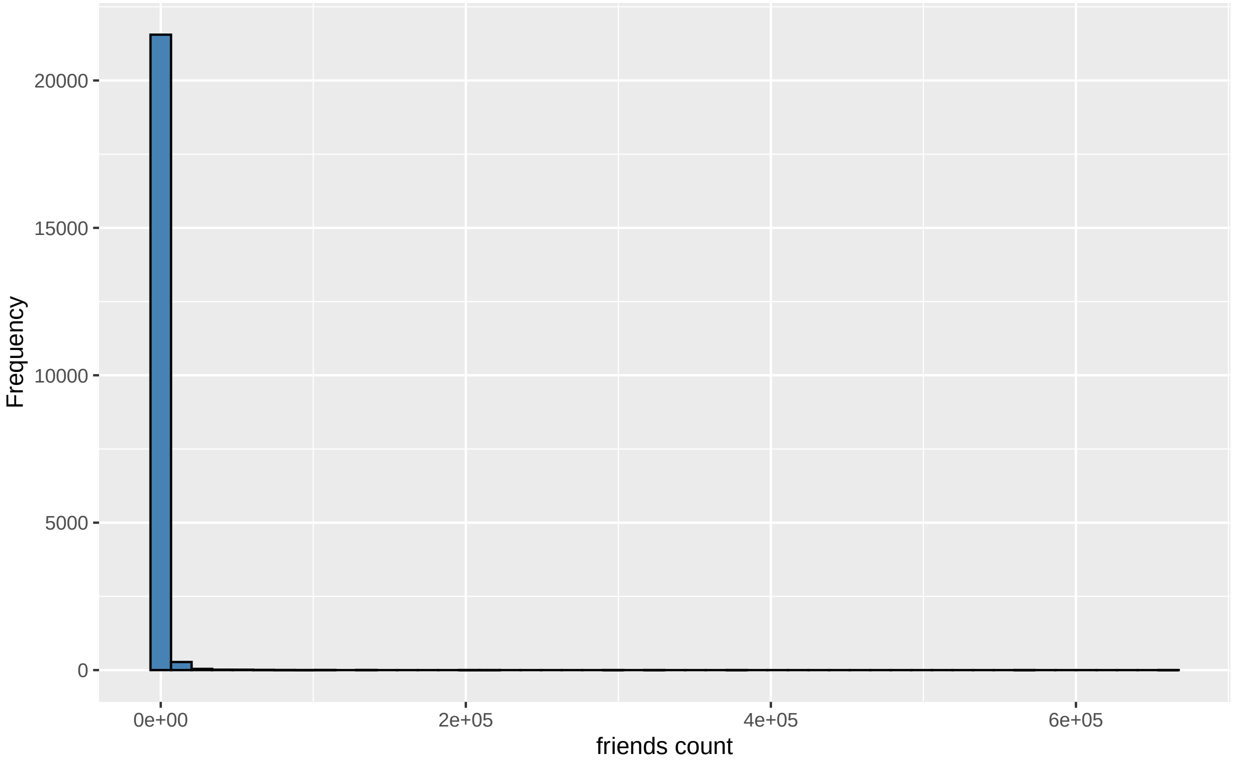
```
summary(M01_quasi_twitter$friends_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     -84     123     324    1058     849  660549
```

Task c

How is the data quality in friend_count variable? Interpret your answer.

```
ggplot(M01_quasi_twitter,aes(x =friends_count))  +
  geom_histogram(bins = 50, color = 'Black', fill = 'steelblue') + xlab("friends count") + ylab("Frequency")
```

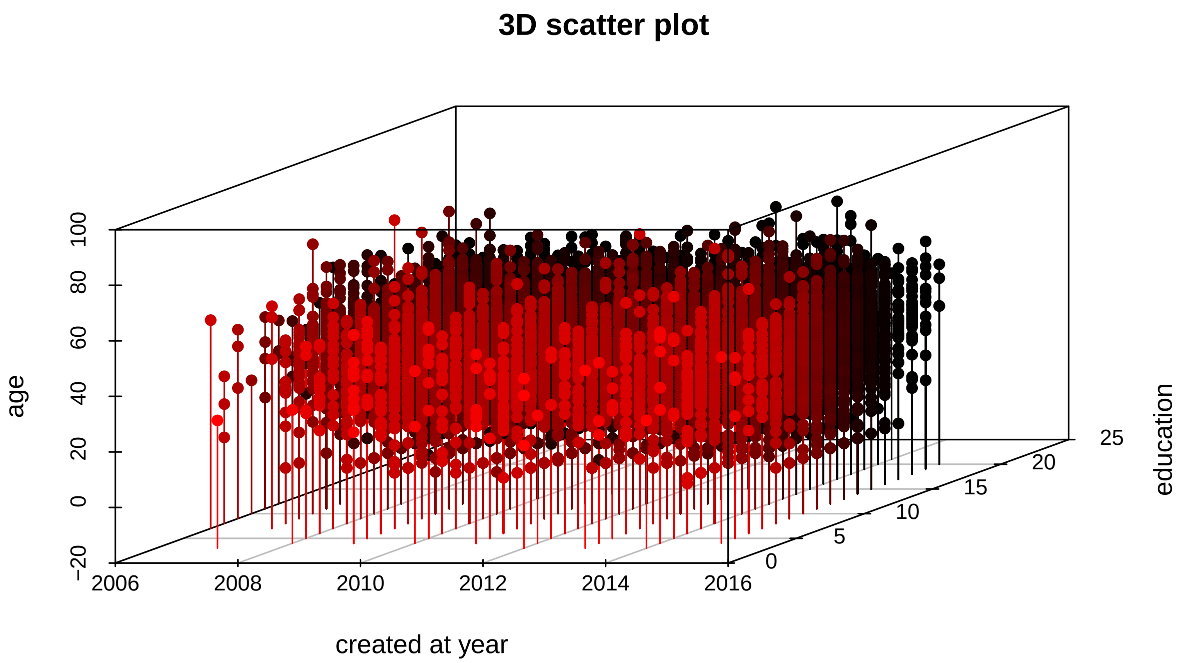


From the above histogram we can say that the data quality in friend_count variable is not good because there are many outliers and the data is completely right tailed. Even after using 50 bins we cannot see that the data spread evenly.

Task d

Produce a 3D scatter plot with highlighting to impression the depth for variables below on M01_quasi_twitter.csv dataset. created_at_year, education, age. Put the name of the scatter plot "3D scatter plot".

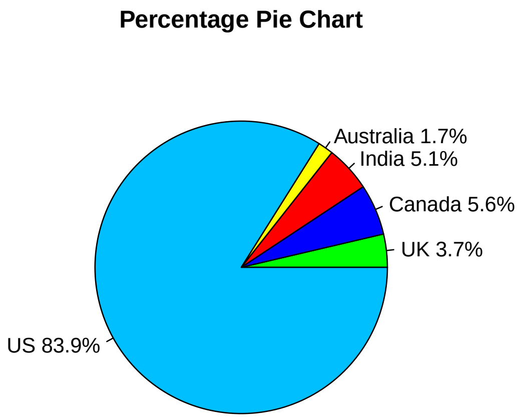
```
scatterplot3d(M01_quasi_twitter$created_at_year, M01_quasi_twitter$education, M01_quasi_twitter$age,
              main = "3D scatter plot",   pch = 16, type = "h", highlight.3d = TRUE,
              xlab = "created at year", zlab = "age",ylab = "education")
```

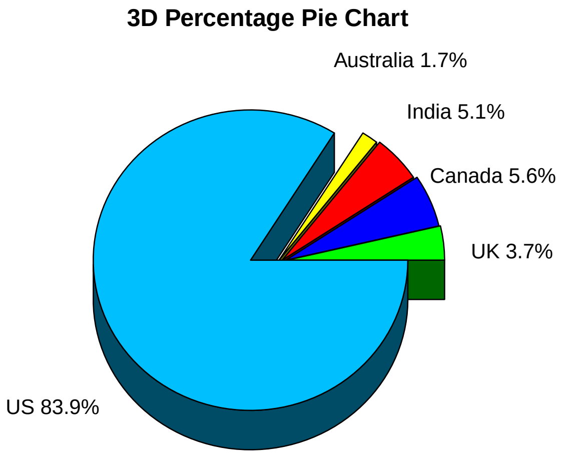


Task e

Consider 650, 1000, 900, 300 and 14900 tweeter accounts are in UK, Canada, India, Australia and US, respectively. Plot the percentage Pie chart includes percentage amount and country name adjacent to it, and also plot 3D pie chart for those countries along with the percentage pie chart.

```
accounts <- c(650, 1000, 900, 300, 14900)
Countries <- c("UK", "Canada", "India", "Australia", "US")
percentage <- round(accounts/sum(accounts)*100,1)
Countries <- paste(Countries, percentage)
Countries <- paste(Countries,"%",sep="")
colors_pie <- c("green","blue","red","yellow","deepskyblue1")
par(mfcol = c(1,2))
pie(labels = Countries,accounts, main = "Percentage Pie Chart", col = colors_pie,radius = 0.8)
pie3D(labels = Countries,accounts, main = "3D Percentage Pie Chart",
      col = colors_pie, explode = 0.1, theta= 1.5, shade = .4, labelcex = 1, radius = 0.8)
```

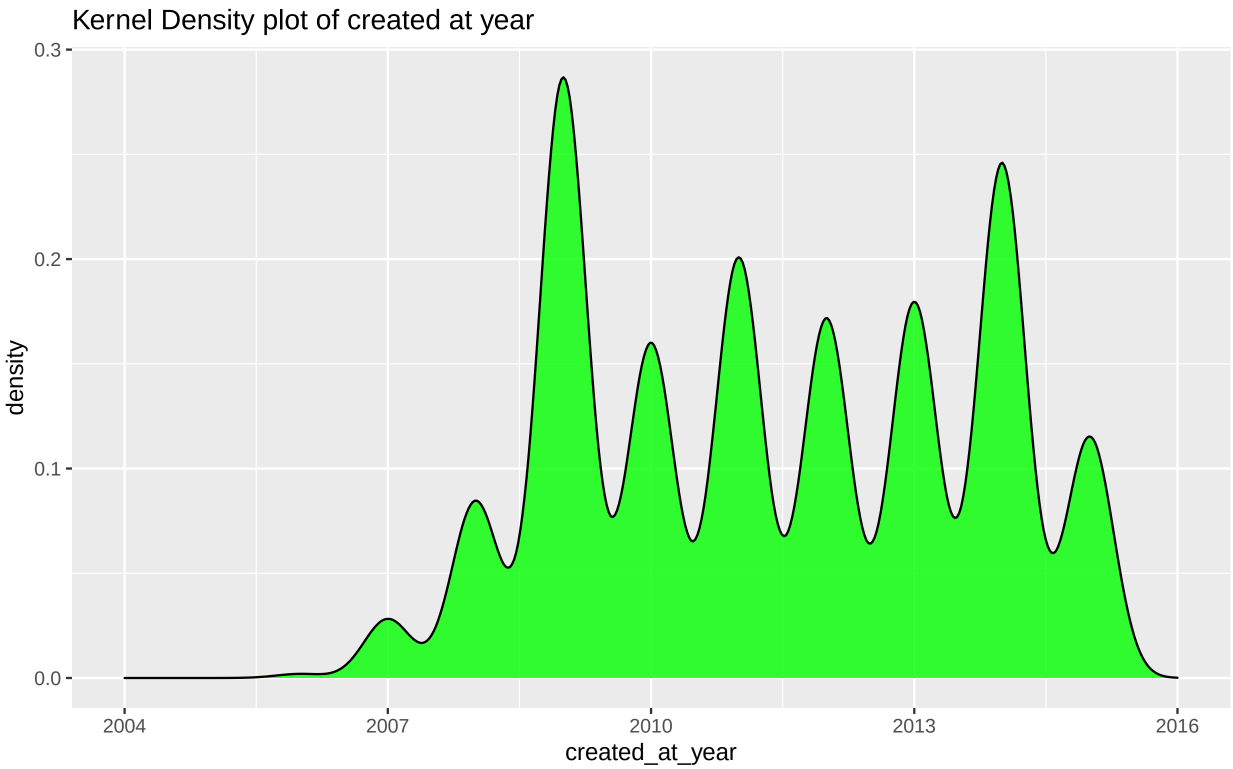




Task f

Create kernel density plot of created__at__year variable and interpret the result.

```
ggplot(M01_quasi_twitter, aes(created_at_year)) +
  geom_density(fill = "green", alpha = 0.8) +
  ggtitle("Kernel Density plot of created at year") + xlim(c(2004,2016))
```



From the above kernal density plot we can understand that there is a high density of accounts created in the year 2009 and low in year 2007. Also, we can understand they are following a pattern by shifting from low to high and high to low.

Problem 3

Task a

Standardize the data and create new dataset with standardized data and name it Ndata.

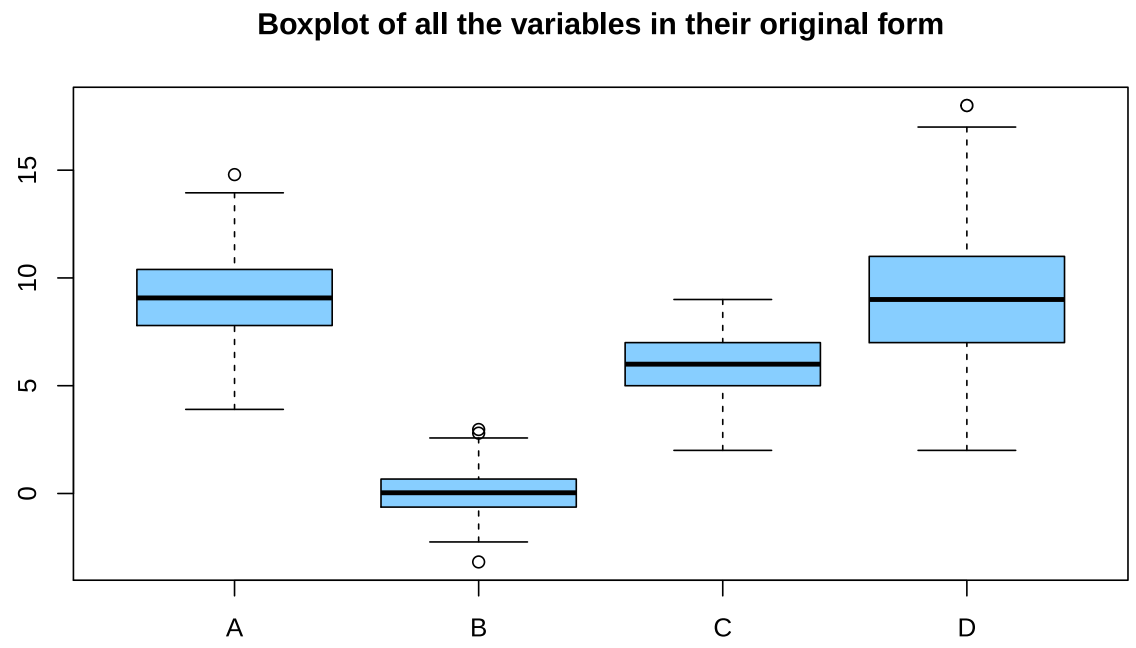
```
# We set center as True to subract with "x bar" and scale as True to divide by sigma
Ndata <- as.data.frame(scale(raw_data, center = TRUE, scale = TRUE))
head(Ndata)
```

```
##             A          B          C          D
## 1 -0.46047167 -0.6870000 -0.2019694 -0.2931233
## 2  0.82780052 -0.7467798  0.4705888 -0.2931233
## 3 -0.18769316  0.7693173  0.4705888 -1.2500845
## 4 -1.41378095  1.5532638 -0.2019694  0.3448509
## 5  0.15837732  0.9970078  0.4705888 -0.2931233
## 6 -0.03285735  0.6893851  0.4705888  0.9828251
```

Task b

Create the boxplot of all the variables in their original form.

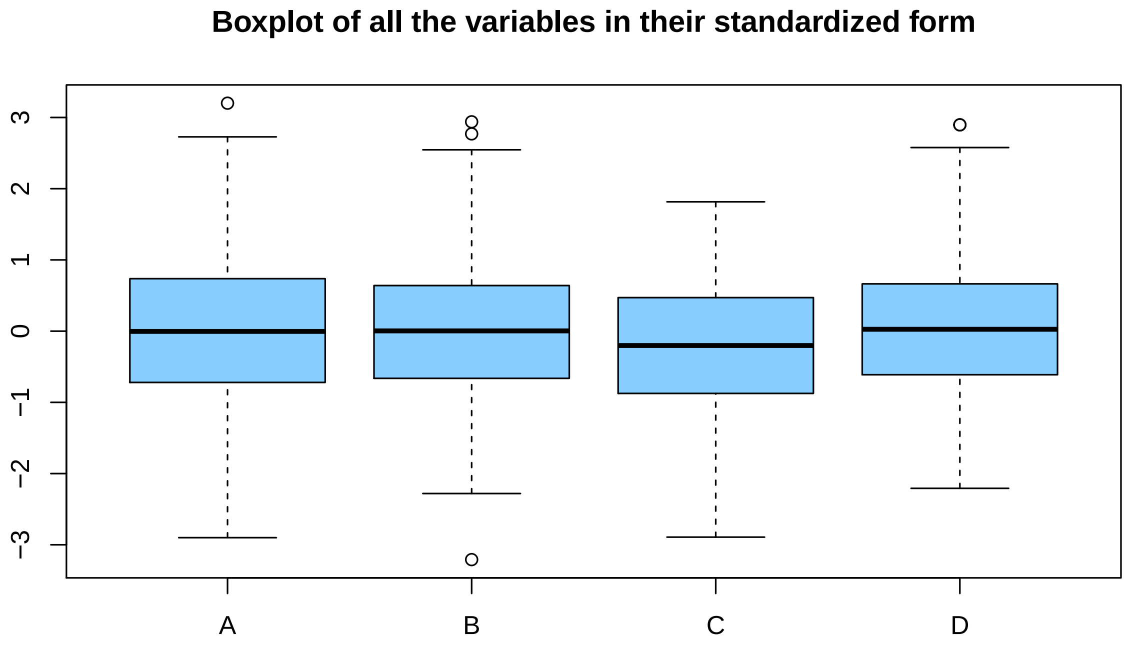
```
boxplot(raw_data, main = "Boxplot of all the variables in their original form", col = "skyblue1")
```



Task c

Create boxplot of all the variables in their standardized form.

```
boxplot(Ndata, main =  "Boxplot of all the variables in their standardized form", col = "skyblue1")
```



Task d

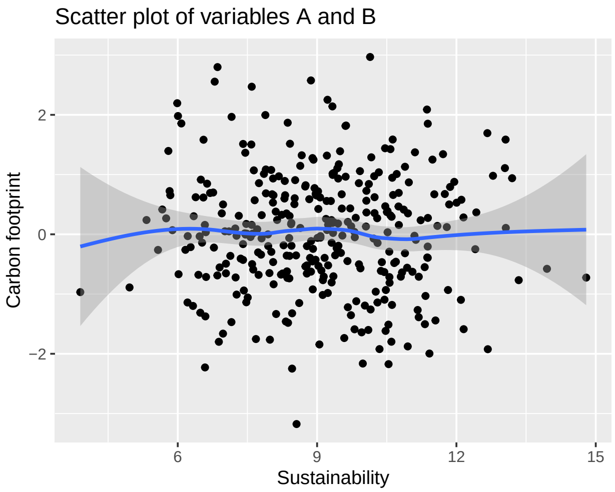
Compare the result of part b and part c; interpret your answer.

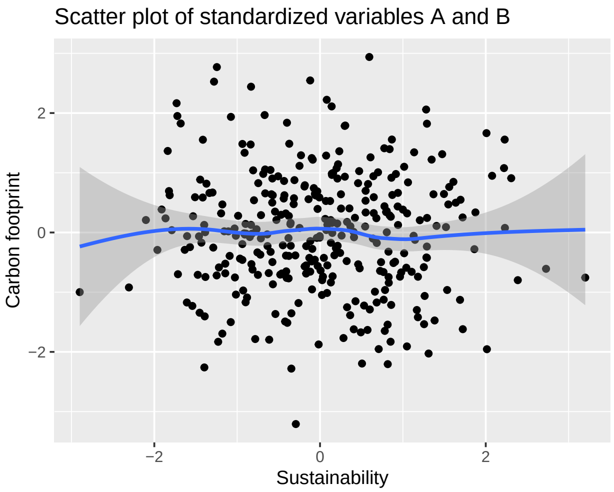
Although the boxplots look similar for the data before and after standardizing it is much easier to understand and compare all variables after standardizing because of negligible scale difference but it is difficult to interpret and compare all variables before standardizing because of the major scale difference.

Task e

Prepare scatter plot of variables A and B. How are the data correlated in these variables? Interpret your answer.

```
fig1 <- ggplot(raw_data, aes(A,B)) + geom_point() + geom_smooth() +
  ggtitle("Scatter plot of variables A and B") + xlab("Sustainability") +
  ylab("Carbon footprint")
fig2 <- ggplot(Ndata, aes(A,B)) + geom_point()+ geom_smooth() +
  ggtitle("Scatter plot of standardized variables A and B") + xlab("Sustainability") +
  ylab("Carbon footprint")
gridExtra::grid.arrange(fig1,fig2,ncol = 2)
```





By comparing both scatterplots we can interpret that there is no correlation between Sustainability and Carbon footprint or it is close to zero. The data is spread randomly without following any patterns.