Assignment 1

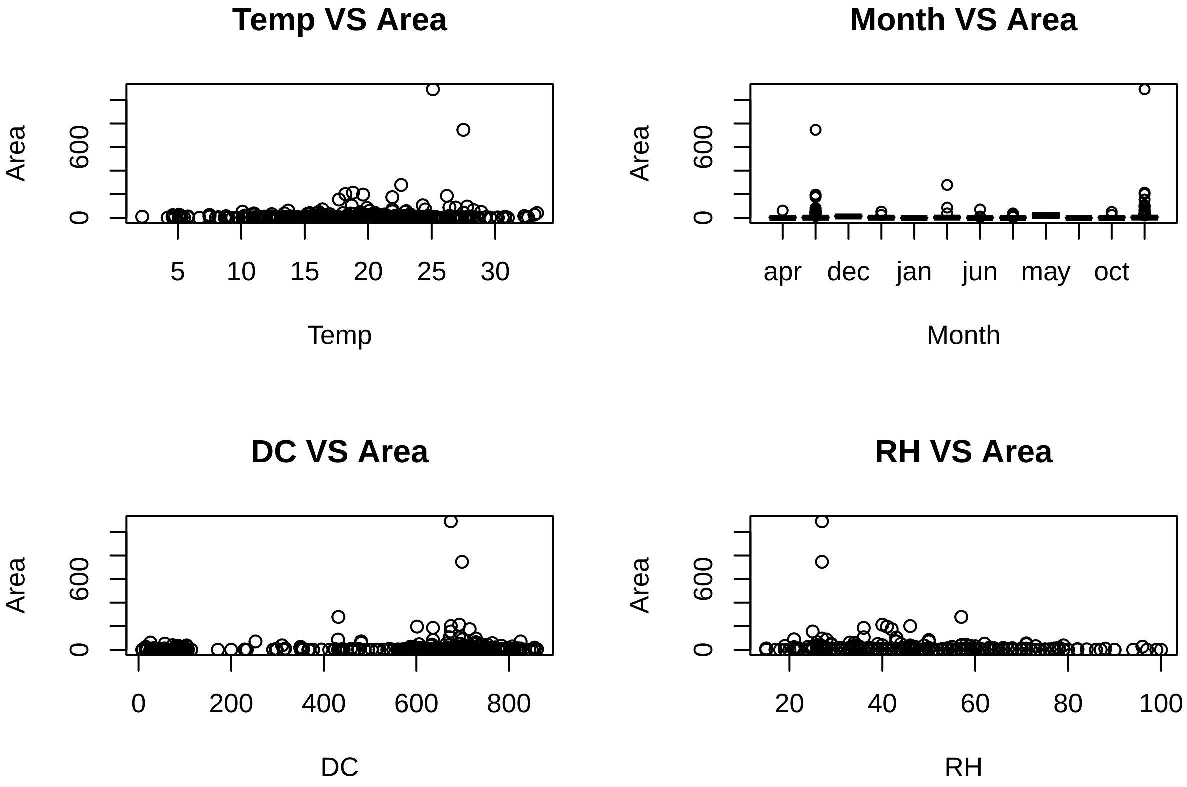
Group 24 :Pratik Mante and Jiaqi Wang

1/21/2020

Problem 1. Forest fire plroblem:

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

```
forestfires <- read.csv("~/Downloads/forestfires.csv")
opar<-par(no.readonly = TRUE)
par(mfrow=c(2,2))
plot(forestfires$temp,forestfires$area,xlab = "Temp",ylab = "Area",main = "Temp VS Area")
plot(forestfires$month,forestfires$area,xlab = "Month",ylab = "Area",main = "Month VS Area")
plot(forestfires$DC,forestfires$area,xlab = "DC",ylab = "Area",main = "DC VS Area")
plot(forestfires$RH,forestfires$area,xlab = "RH",ylab = "Area",main = "RH VS Area")
```



```
cor(forestfires$temp,forestfires$area)
```

```
## [1] 0.09784411
```

```
cor(forestfires$DC,forestfires$area)
```

```
## [1] 0.04938323
```

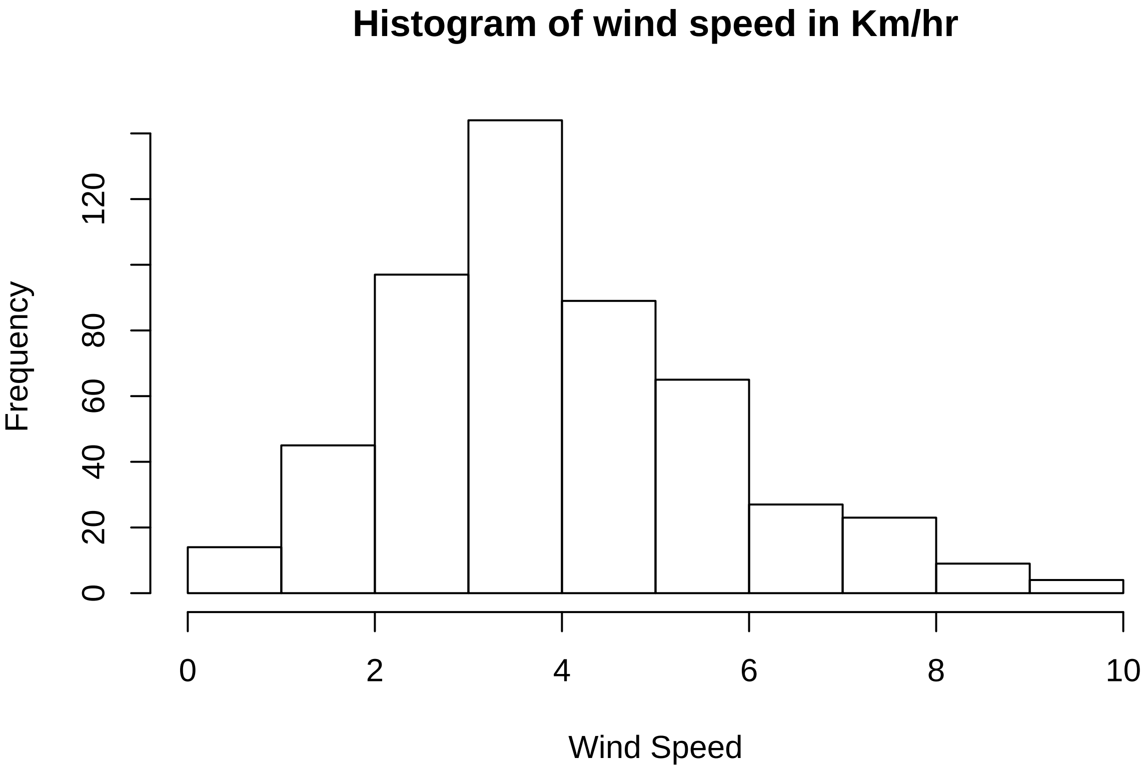
```
cor(forestfires$RH,forestfires$area)
```

```
## [1] -0.07551856
```

```
par(opar)
```

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

```
hist(forestfires$wind,xlab = "Wind Speed",main = "Histogram of wind speed in Km/hr")
```



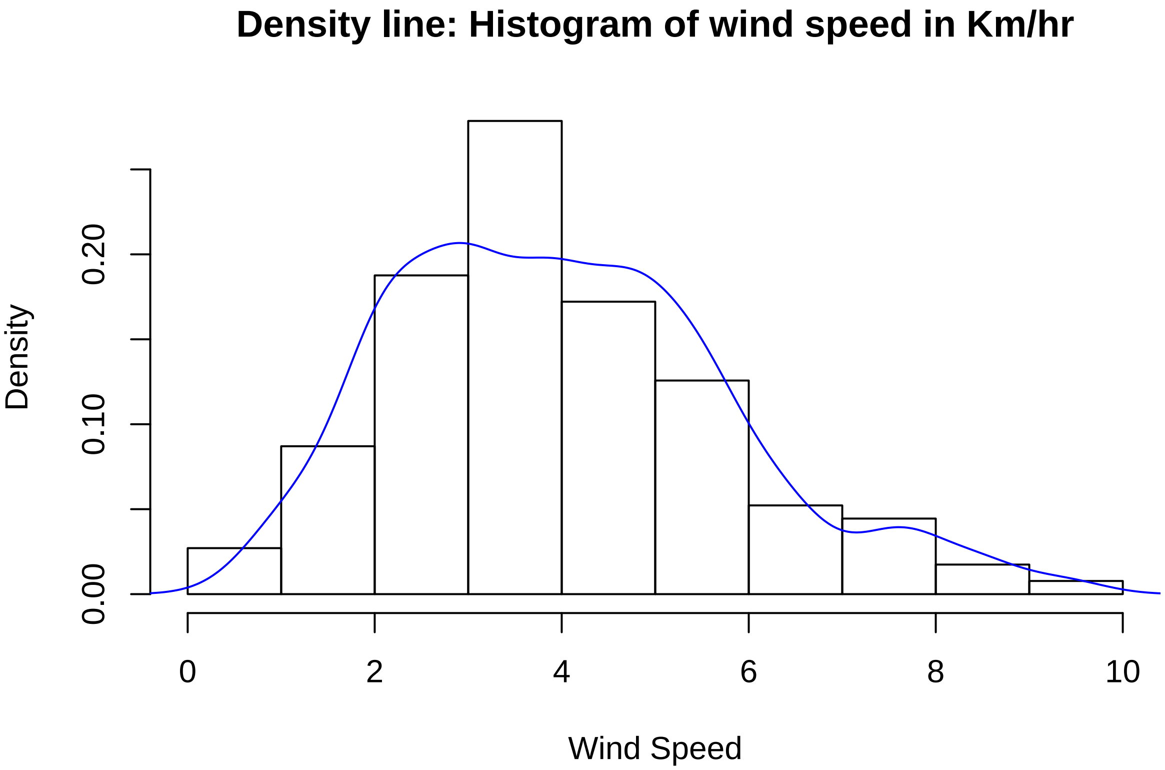
c. Compute the summary statistics (min, 1Q, mean, median, 3Q, max,) of part b.

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

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

d. Add a density line to the histogram in part b.

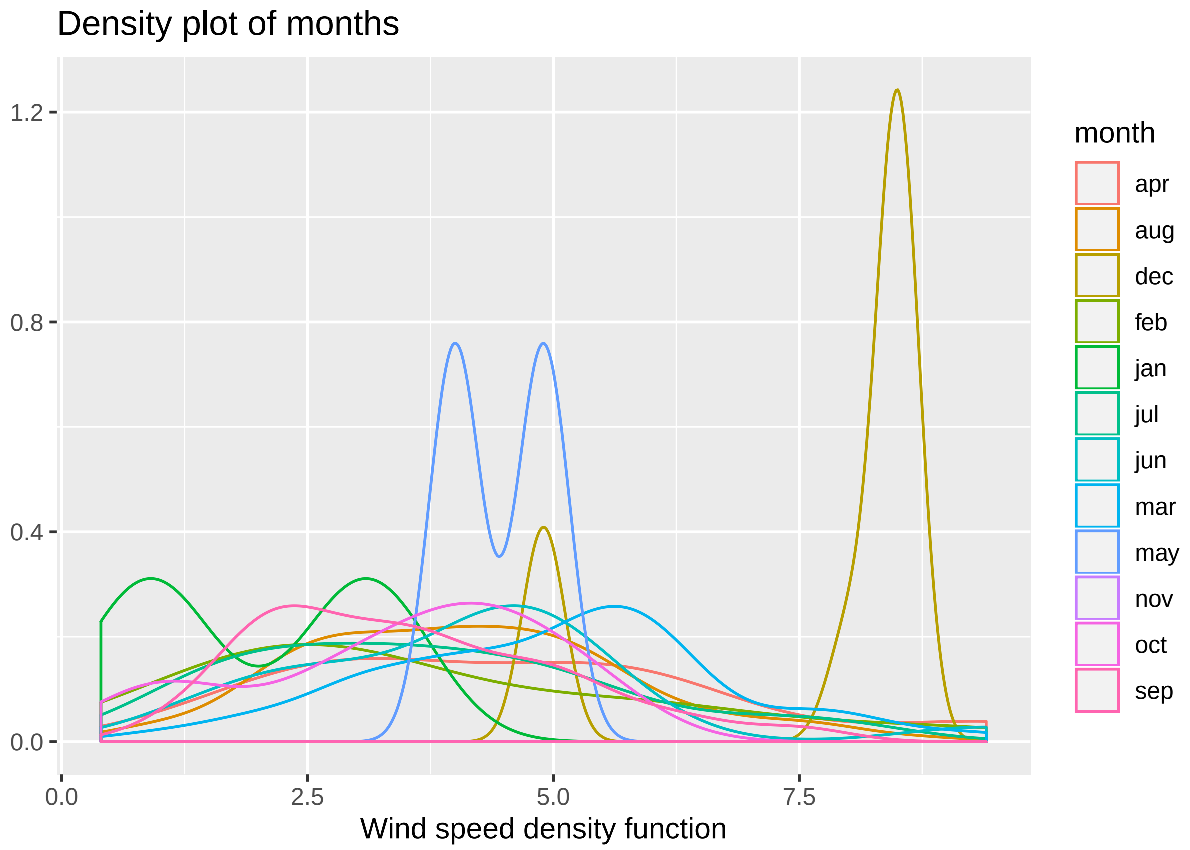
```
hist(forestfires$wind,xlab = "Wind Speed",main = "Density line: Histogram of wind speed in Km/hr",freq =
lines(density(forestfires$wind),col="blue")
```



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.

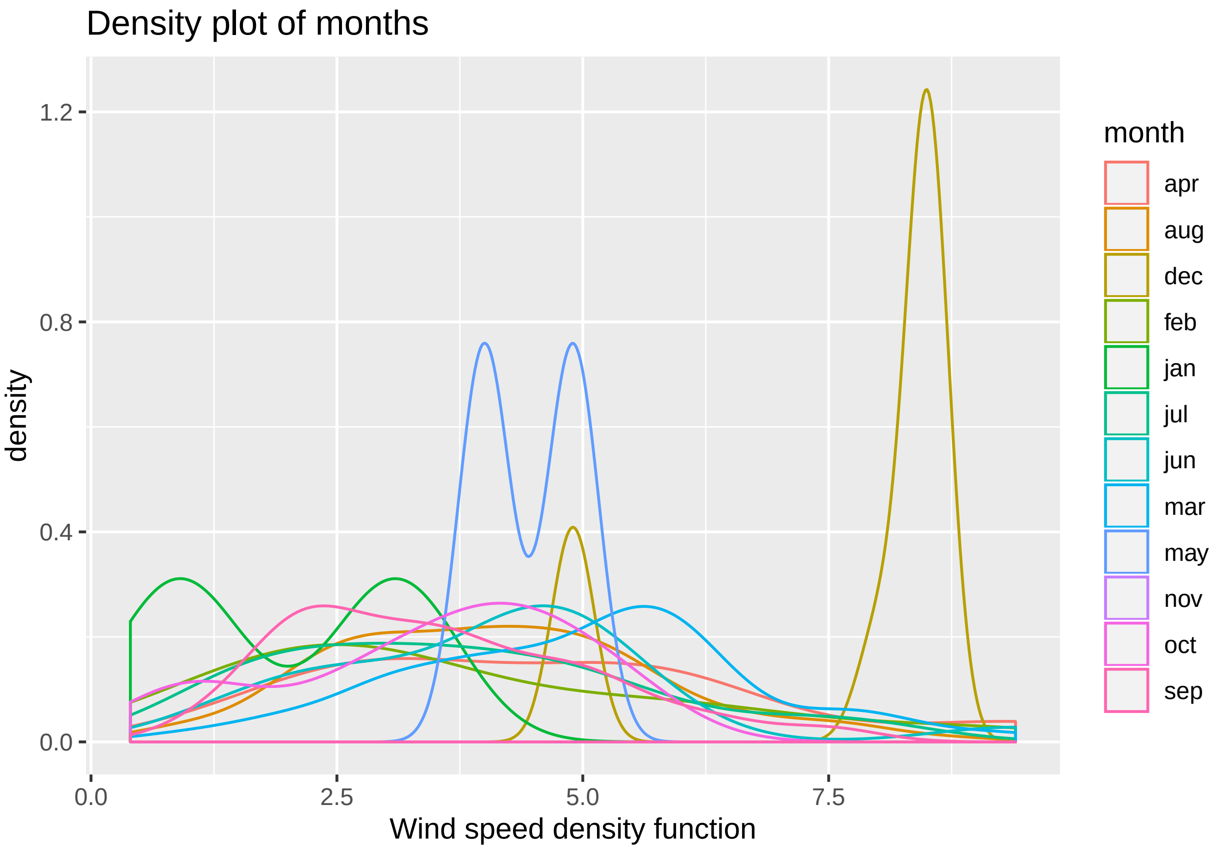
```
library(ggplot2)
qplot(forestfires$wind,data = forestfires,geom = "density",color= month,xlab = "Wind speed density func
```

```
## Warning: Groups with fewer than two data points have been dropped.
```



```
ggplot(forestfires,aes(x=forestfires$wind,color=month))+geom_density(position = "identity")+scale_x_con
```

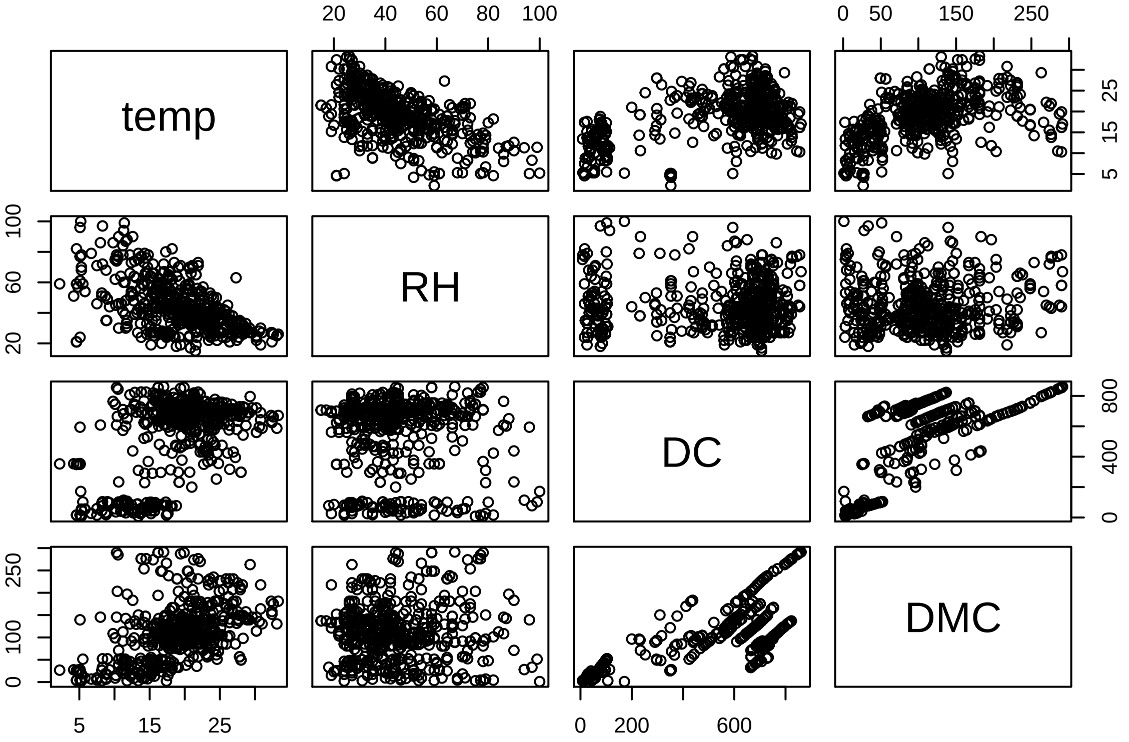
```
## Warning: Groups with fewer than two data points have been dropped.
```



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

Answer: None of the variables have 0 correlation and most of the relation is concentrated at the sides.

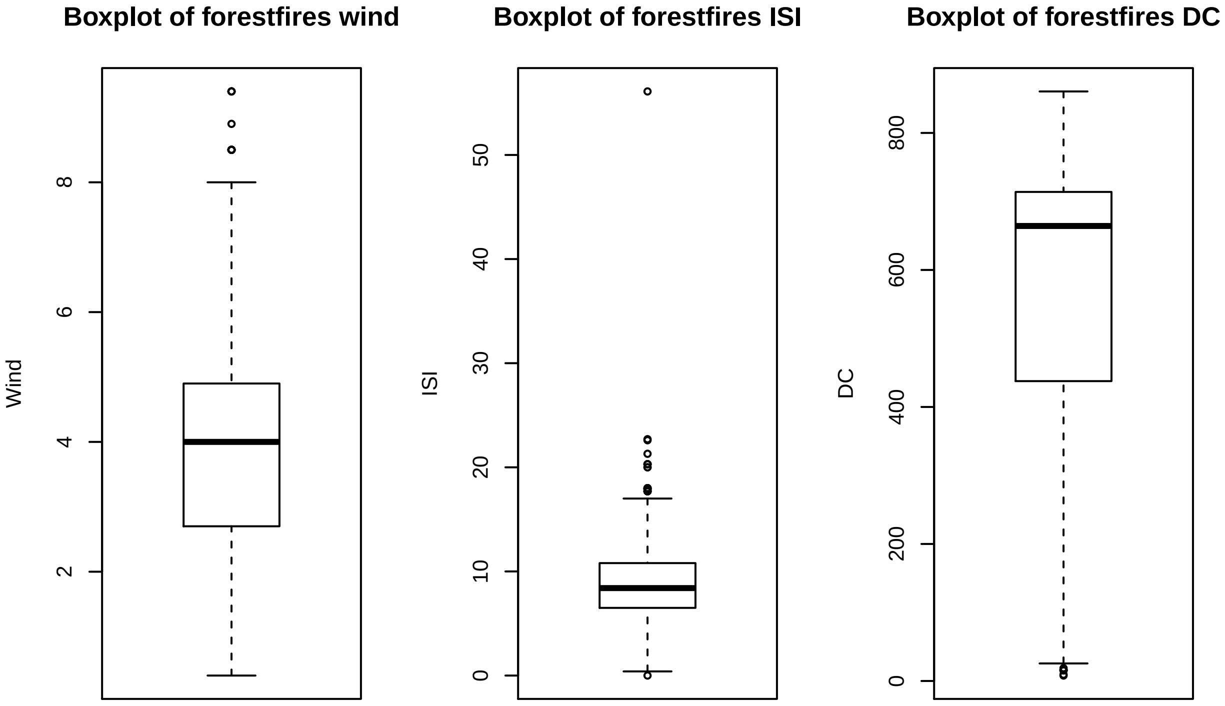
```
pairs(~temp+RH+DC+DMC,data = forestfires)
```



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

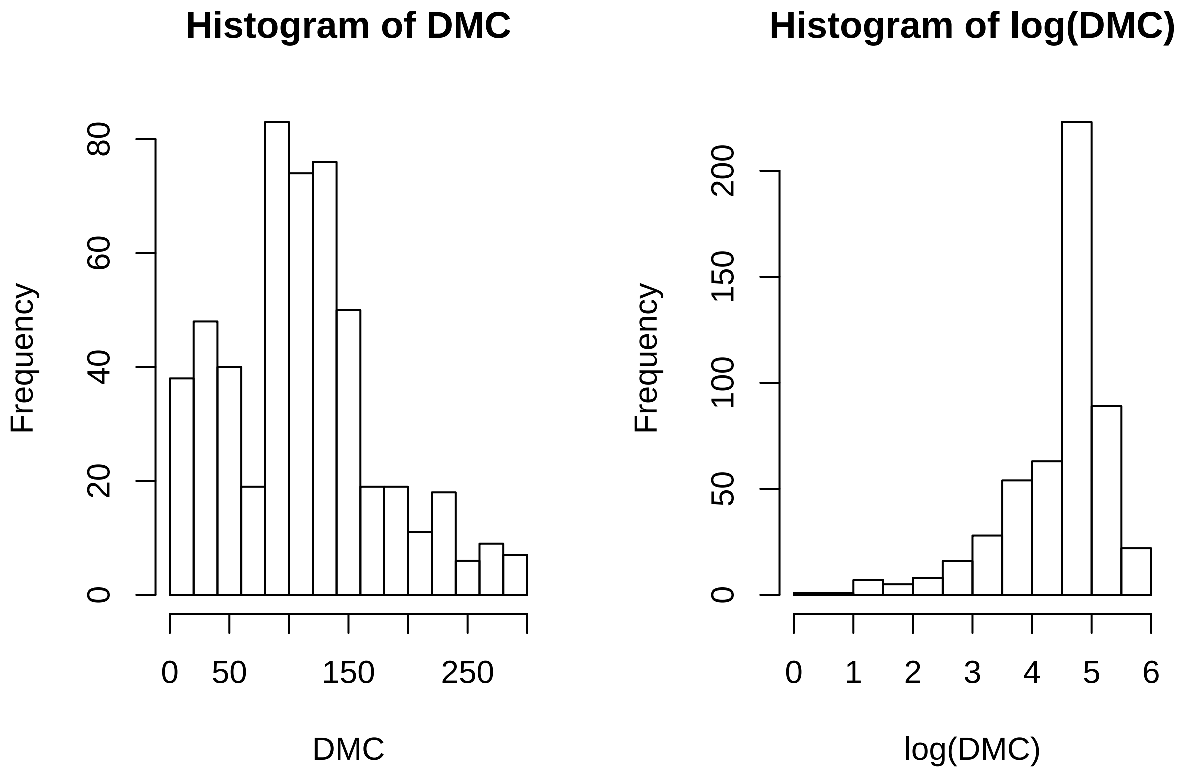
Answer: Outliers are present in all three boxplots. From the boxplot it can also be observed that the data is skewed: data for foresfires wind and ISI is skewed to right and for DC it is skewed to left.

```
par(mfrow=c(1,3))
boxplot(forestfires$wind,ylab = "Wind",main = "Boxplot of forestfires wind")
boxplot(forestfires$ISI,ylab="ISI",main="Boxplot of forestfires ISI")
boxplot(forestfires$DC,ylab="DC",main="Boxplot of forestfires DC")
```



h. Create the histogram of DMC. Create the histogram of log of DMC. Compare the result and explain your answer. Answer: Log of DMC decreases the range of x-axis as compared to the x-axis of DMC. Histogram of log of DMC is right skewed whereas the histogram of DMC is neither skewed nor normalised.

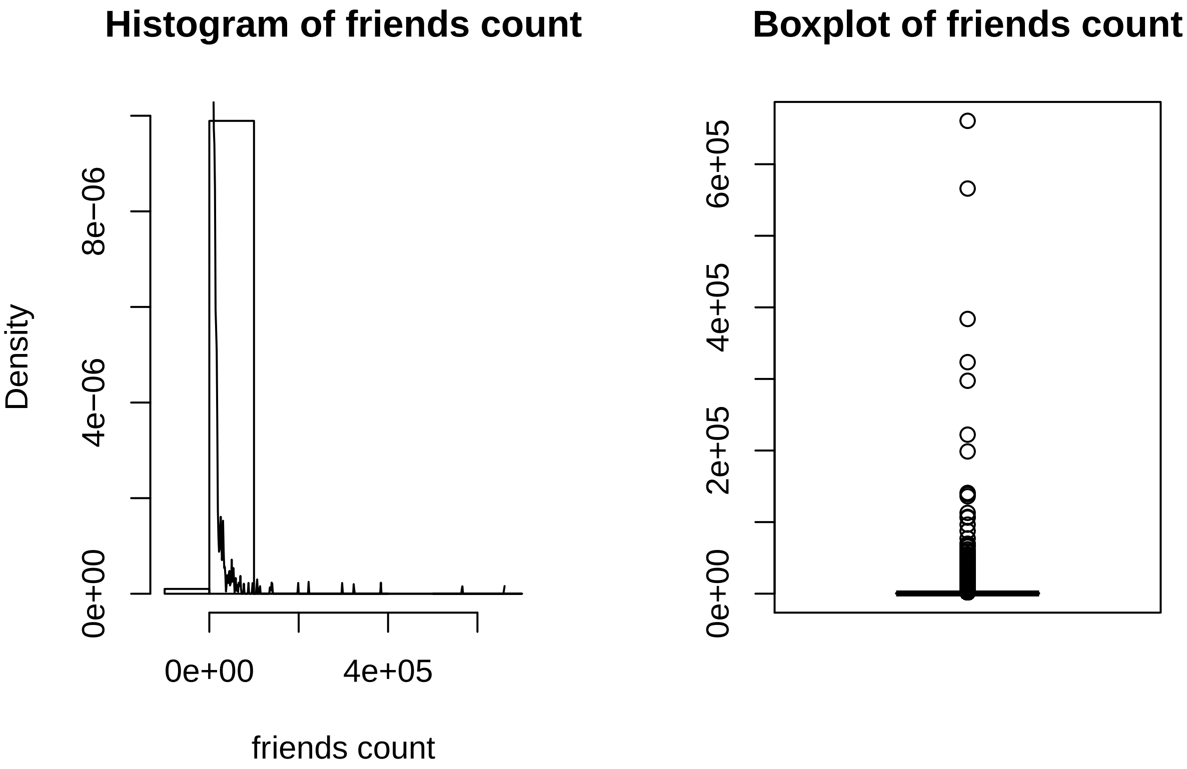
```r
par(mfrow=c(1,2))
hist(forestfires$DMC,xlab = "DMC",main = "Histogram of DMC")
hist(log(forestfires$DMC),xlab = "log(DMC)",main = "Histogram of log(DMC)")
```



Problem 2. Twitter accounts problem

a. How are the data distributed for friend_count variable? Answer: Data distributed is left skewed.

```r
twitterdata <- read.csv("~/Downloads/M01_quasi_twitter.csv")
par(mfrow=c(1,2))
hist(twitterdata$friends_count,breaks=5,freq = FALSE,xlab = "friends count",main = "Histogram of friends
lines(density(twitterdata$friends_count))
boxplot(twitterdata$friends_count,main = "Boxplot of friends count")
```



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

```
summary(twitterdata$friends_count)
```

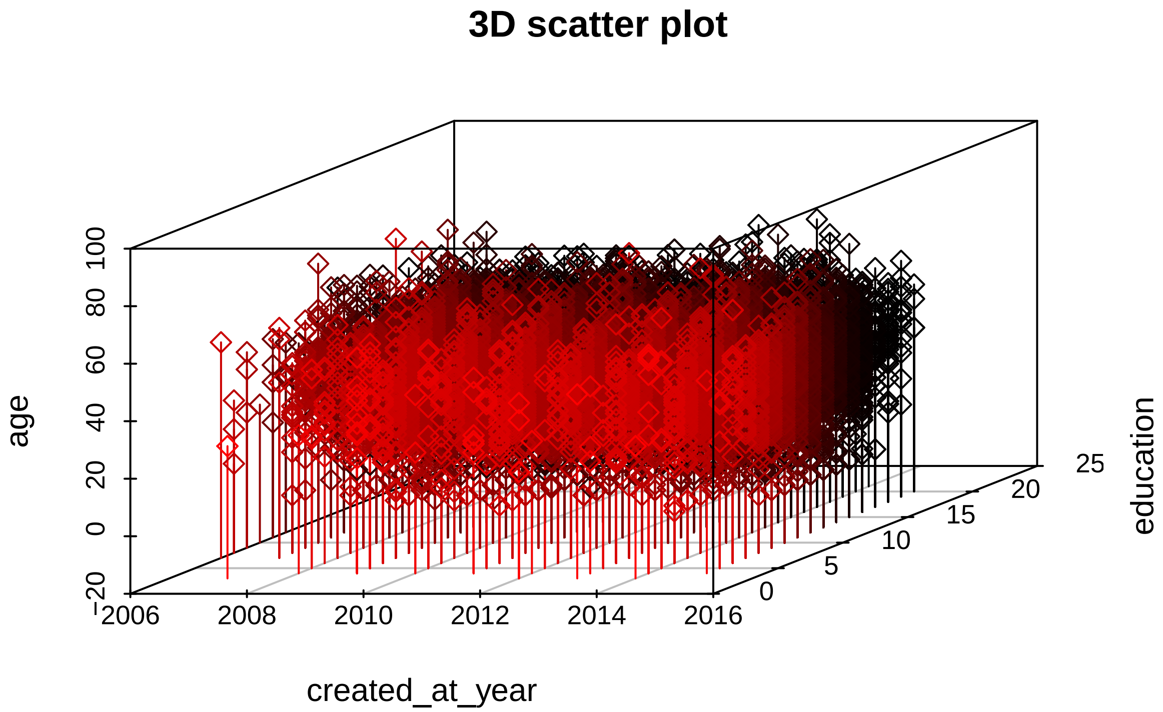
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     -84     123     324    1058     849  660549
```

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

Answer: The quality of the data is not good because it is concentrated in one region.

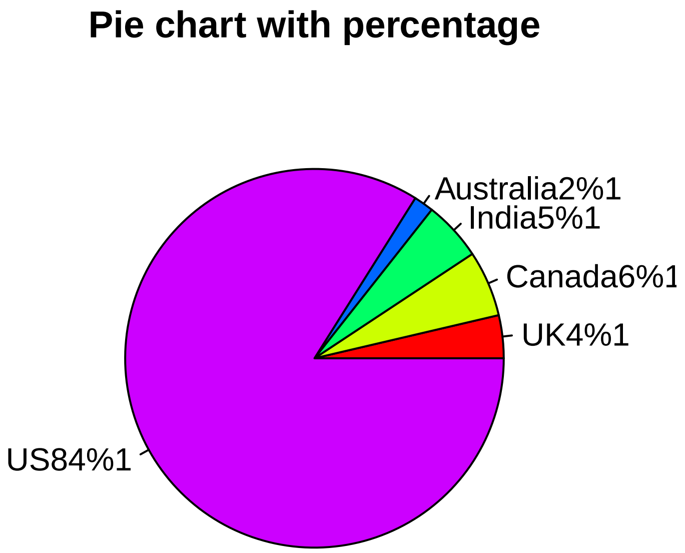
d. Produce a 3D scatter plot with highlighting to impression the depth for variables below on M01_quasi_twitter.csv dataset. created_at_year, education, age. Put the name of the scatter plot "3D scatter plot".

```
library(scatterplot3d)
attach(twitterdata)
par(mfrow=c(1,1))
plot3d <- scatterplot3d(created_at_year,education,age,pch = 5,highlight.3d = TRUE,type = "h",main = "3D
```

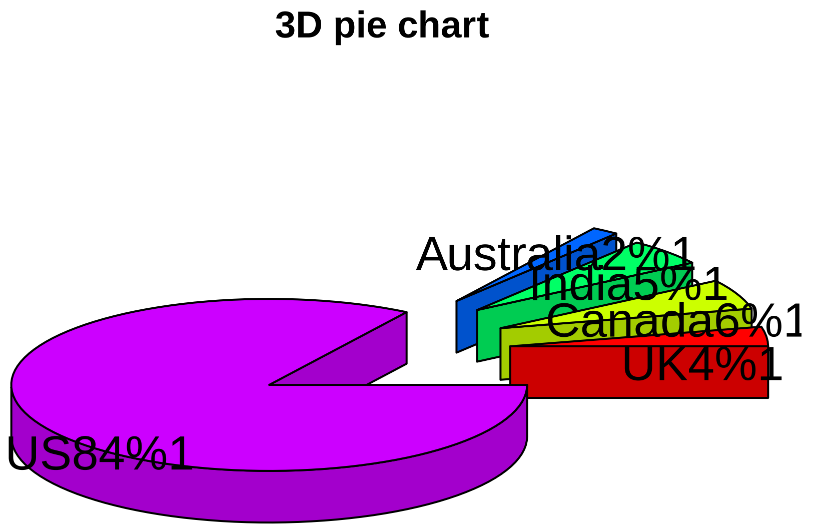


e. Consider 650, 1000, 900, 300 and 14900 tweeter accounts are in UK, Canada, India, Australia and US, respectively. Plot the percentage Pie chart includes percentage amount and country name adjacent to it, and also plot 3D pie chart for those countries along with the percentage pie chart. Hint: Use C=(1, 2) matrix form to plot the charts together.

```
accounts <- c(650,1000,900,300,14900)
countries <- c("UK","Canada","India","Australia","US")
piechart <- round(accounts/sum(accounts)*100)
countries1 <- paste(countries,"",piechart,"%",sep = "",font.size=1)
pie(accounts,labels = countries1,col = rainbow(length(countries1)),main = "Pie chart with percentage")
```



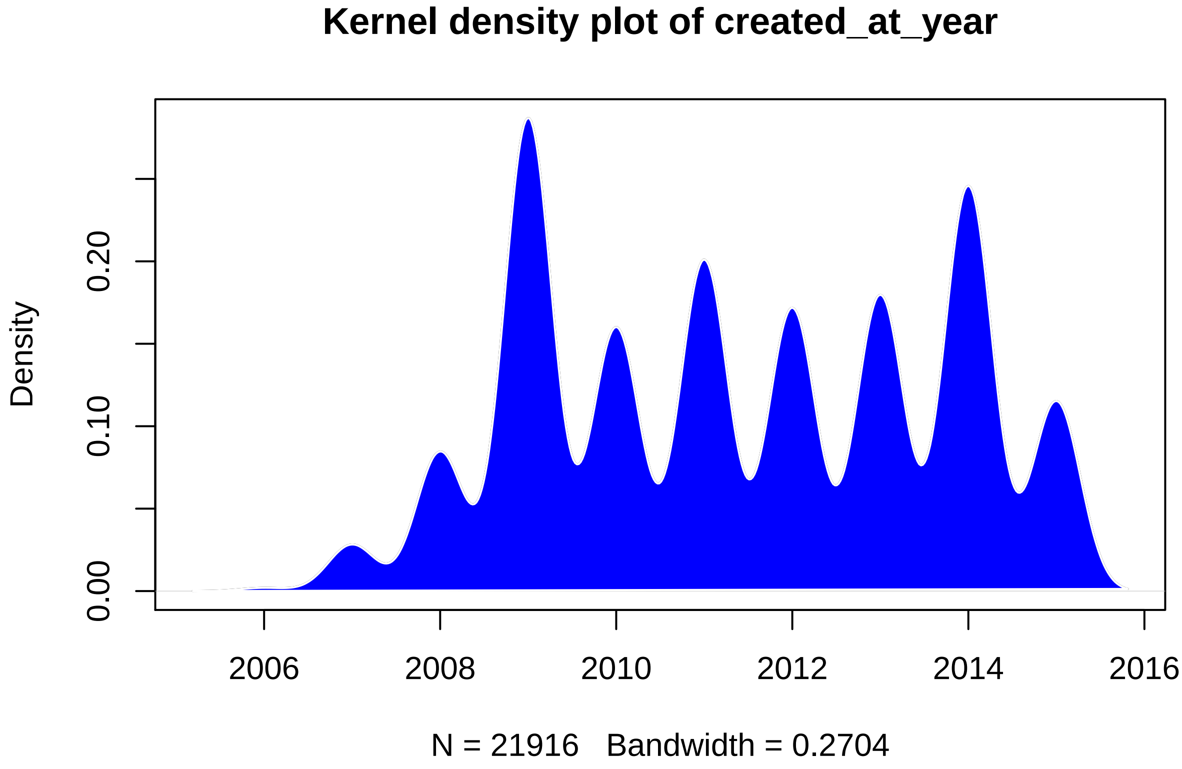
```
library(plotrix)
pie3D(accounts,labels = countries1,explode = 0.5,main = "3D pie chart")
```



f. Create kernel density plot of created_at_year variable and interpret the result.

Answer: Plot represents that the highest number of accounts were created in mid of year 2009 and from 2010 till 2013 the number of accounts created was similar but less than the number of accounts created in 2009. In 2014 there was rise in the number of accounts created but it again fell down in 2015.

```
plot(density(twitterdata$created_at_year),main = "Kernel density plot of created_at_year")
polygon(density(twitterdata$created_at_year),col = "blue",border = "white")
```



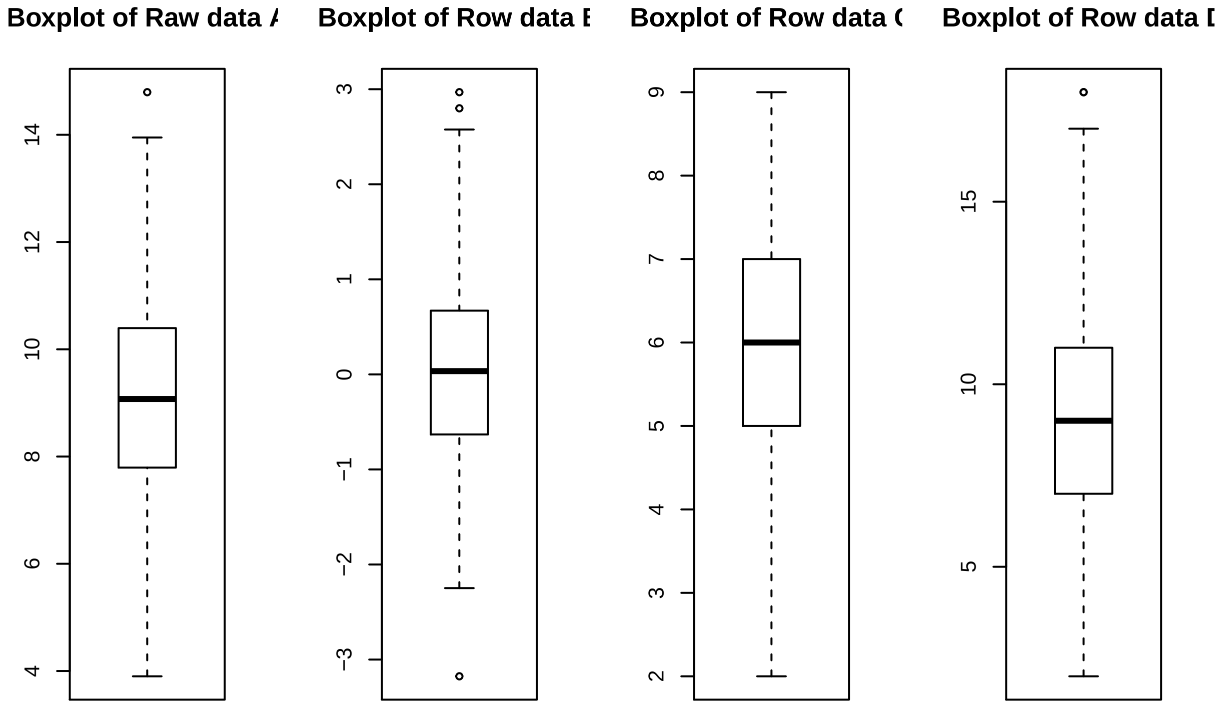
Problem 3. Insurance claims problem

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

```
insurance_data <- read.csv("~/Downloads/raw_data.csv")
Ndata <- scale(insurance_data,center = TRUE,scale = TRUE)
```

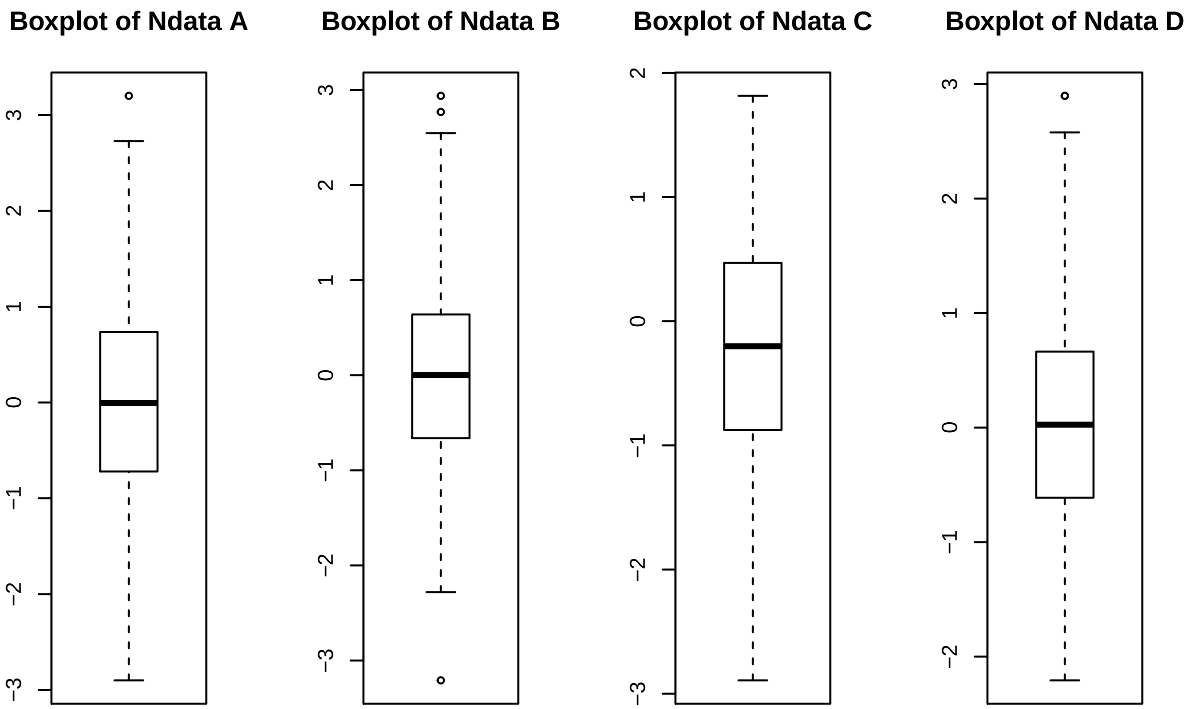
b. Create the boxplot of all the variables in their original form.

```
par(mfrow=c(1,4))
boxplot(insurance_data$A,main = "Boxplot of Raw data A")
boxplot(insurance_data$B,main = "Boxplot of Row data B")
boxplot(insurance_data$C,main = "Boxplot of Row data C")
boxplot(insurance_data$D,main = "Boxplot of Row data D")
```



c. Create boxplot of all the variables in their standardized form.

```
par(mfrow=c(1,4))
boxplot(Ndata[,1],main = "Boxplot of Ndata A")
boxplot(Ndata[,2],main = "Boxplot of Ndata B")
boxplot(Ndata[,3],main = "Boxplot of Ndata C")
boxplot(Ndata[,4],main = "Boxplot of Ndata D")
```



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

Answer: From part b and part c it can be compared that the data is already normalised but since we have standardized the data in part c the range of the data has been made more tight.

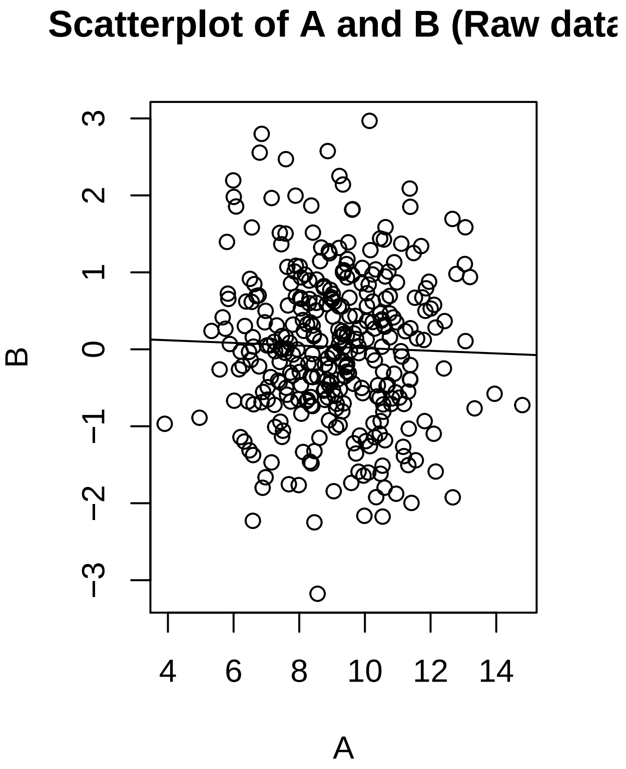
e. Prepare scatter plot of variables A and B. How are the data correlated in these variables? Interpret your answer.

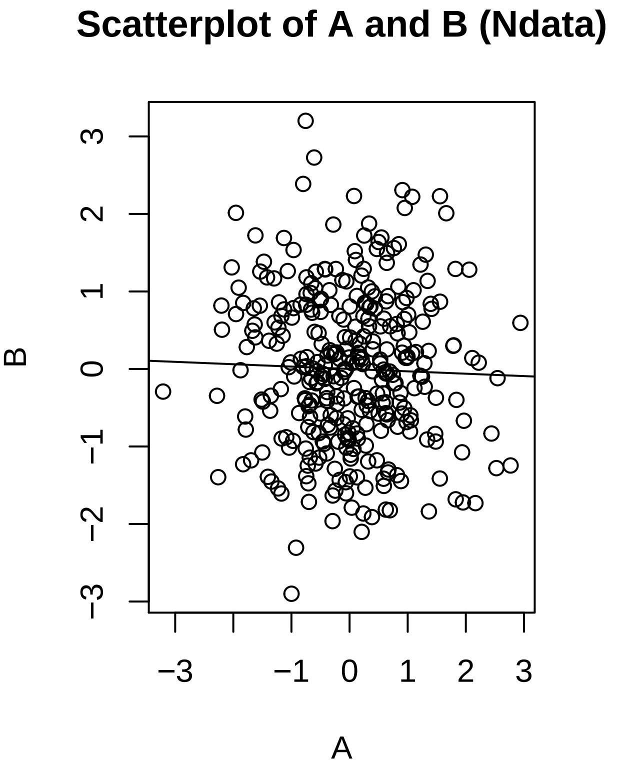
Answer: Scatterplot of A and B depicts that the data is concentrated towards the centre i.e. the data is normalised. From the scatterplots it can be inferred that the data has approximately 0 correlation.

```
par(mfrow = c(1,2))
plot(insurance_data$A,insurance_data$B,xlab = "A",ylab = "B",main = "Scatterplot of A and B (Raw data)"
abline(lm(insurance_data$B~insurance_data$A),col = "black")
cor(insurance_data$B,insurance_data$A)
```

```
## [1] -0.03059086
```

```
plot(Ndata[,2],Ndata[,1],xlab = "A",ylab = "B",main = "Scatterplot of A and B (Ndata)")
abline(lm(Ndata[,2]~Ndata[,1]),col = "black")
```


Scatterplot of A and B (Raw data


Scatterplot of A and B (Ndata)

```r
cor(Ndata[,2],Ndata[,1])
```

```
## [1] -0.03059086
```