**HOMEWORK 1**

**IE 7275 Data Mining in Engineering**

|  |  |
| --- | --- |
| Submitter by | Ashwin Muthiah Murugappan, Muhammad Rehman |
| Study group | 9 |

**Problem 1 (Forest Fires)**

**1.a.** Plot area vs. temp, area vs. month, area vs. DC, area vs. RH for January through December combined in 1 graph. Hint: Place area on Y axis and use 2x2 matrix to place the plots adjacent to each other.

***Code:***

*ffire.df <- read.csv("forestfires.csv", header = TRUE)*

*attach(ffire.df)*

*opar <- par(no.readonly=TRUE)*

*par(mfrow=c(2,2))*

*plot(temp,area)*

*plot(month,area)*

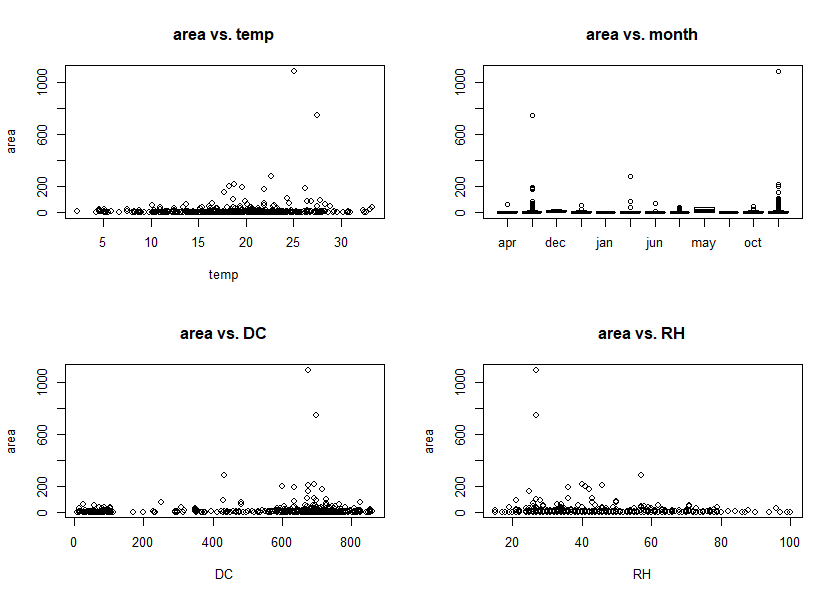
*plot(DC,area)*

*plot(RH,area)*

*par(opar)*

*detach(ffire.df)*

**Result:**

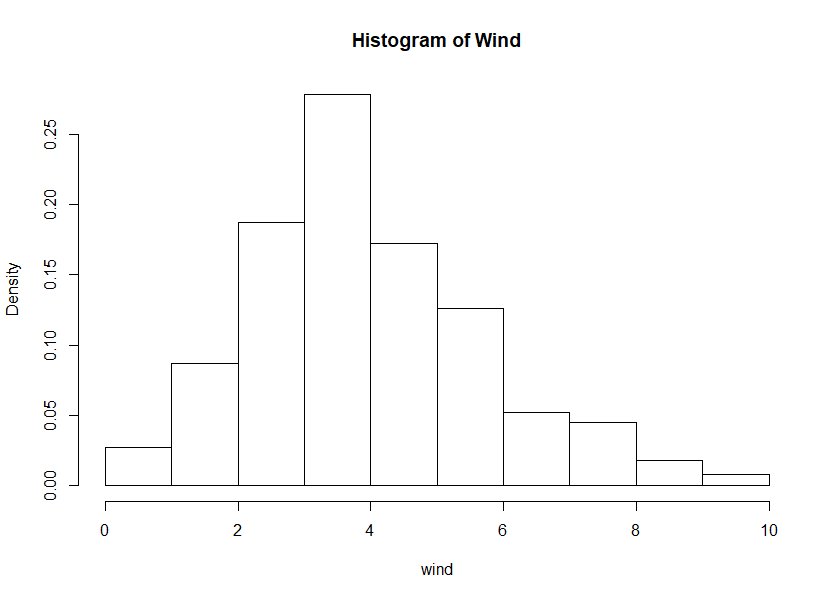


**1.b.** Plot the histogram of wind speed (km/h)

***Code:***

*hist(ffire.df$wind , freq = FALSE,xlab ="wind",main = "Histogram of Wind")*

**Result:**

**

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

**Code:**

*summary(ffire.df$wind)*

**Result:**

Min. 1st Qu. Median Mean 3rd Qu. Max.

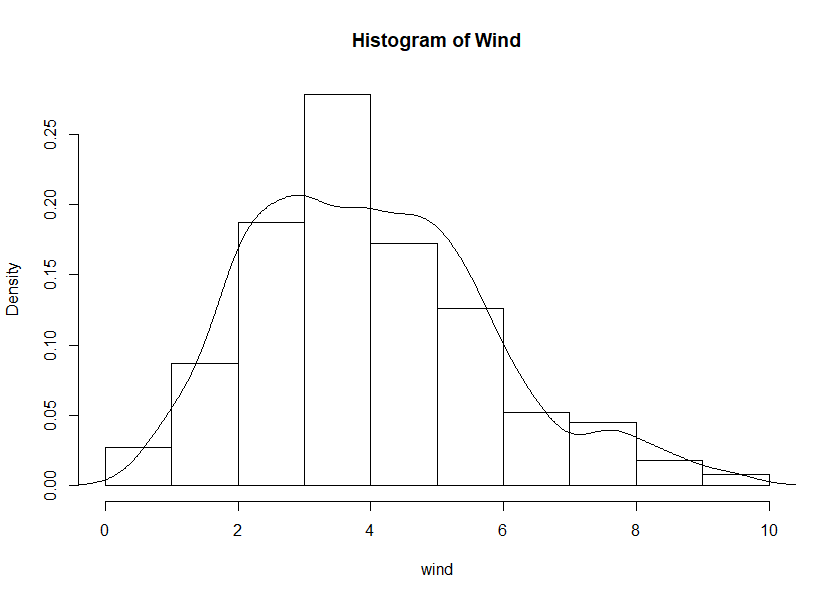
0.400 2.700 4.000 4.018 4.900 9.400

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

**Code:**

*lines(density(ffire.df$wind))*

**Result:**



**1.e**. Plot the wind speed density function of months. Use different colors in the graph to interpret your result clearly. [Hint: use qplot(geom=density)]

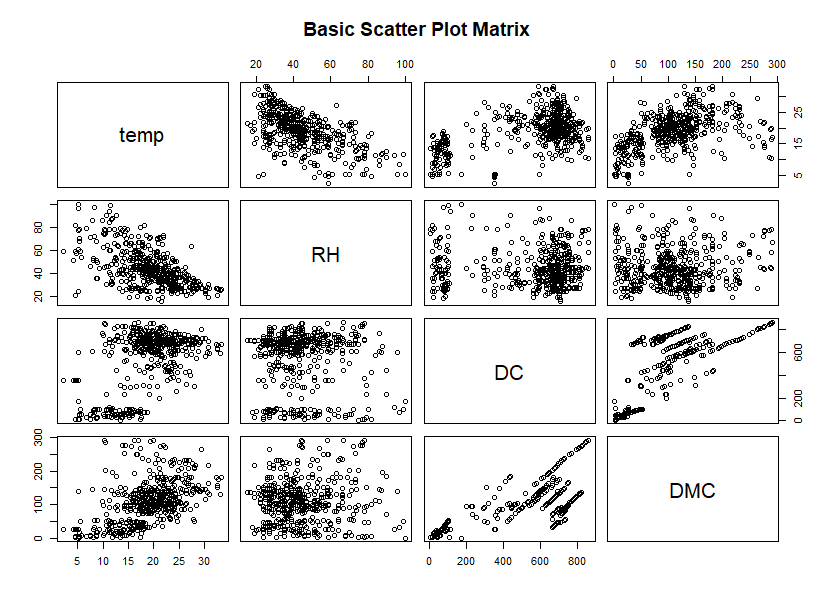
**1.f.** Plot the scatter matrix for temp, RH, DC and DMC. How would you interpret the result in terms of correlation among these data?

**Code:**

*pairs(~temp+RH+DC+DMC, data=ffire.df, main="Basic Scatter Plot Matrix")*

**Result:**

|  |  |  |
| --- | --- | --- |
| Temp and Rh - negative correlation |  |  |
| Temp and DC - No correlation | RH and DC – No correlation |  |
| Temp and DMC - Positive Correlation | RH and DMC – No correlation | DC and DMC – Positive correlation |



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

**Code:**

*par(mfrow=c(1,3))*

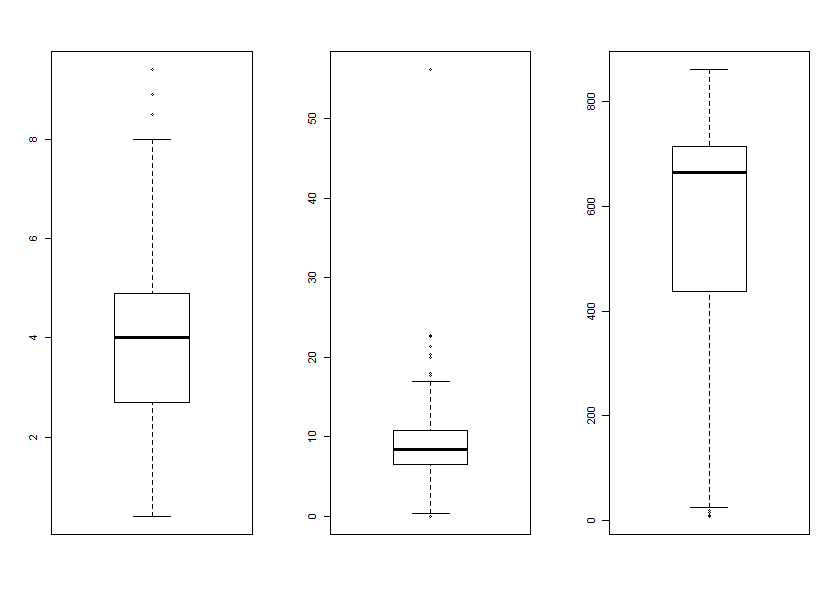
*boxplot(ffire.df$wind)*

*boxplot(ffire.df$ISI)*

*boxplot(ffire.df$DC)*

**result:**

There are a few outliers for variable wind , many outliers for ISI and few outliers for DC



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

**Code:**

*par(mfrow=c(1,2))*

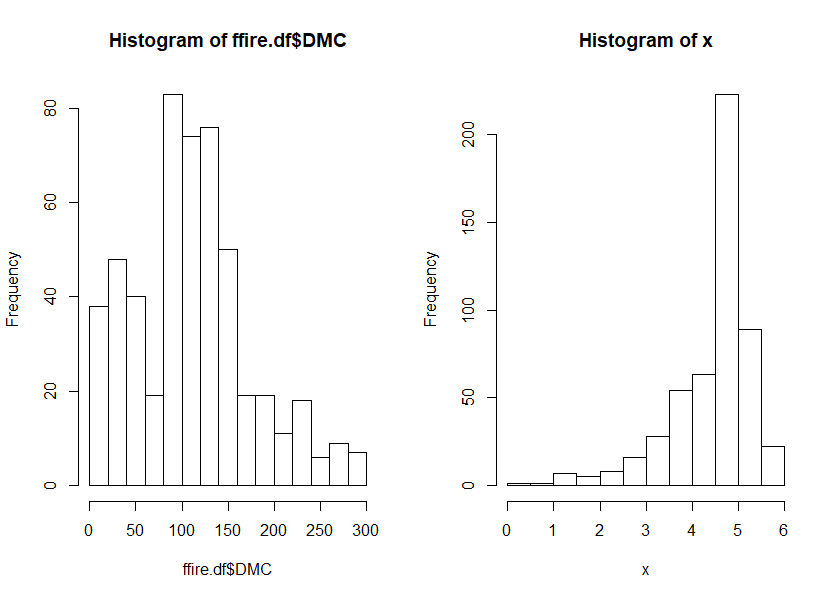
*hist(ffire.df$DMC)*

*x<-log(ffire.df$DMC)*

*hist(x)*

**Result:**

Changing the scale in a display can enhance the plot and illuminate relationships., we see the effect of changing axes to logarithmic (log) scale. Whereas the original plots (left) are hard to understand, the patterns become visible in log scale (right). The rescaling allows a better view of the relationship between the two log-scaled variables.



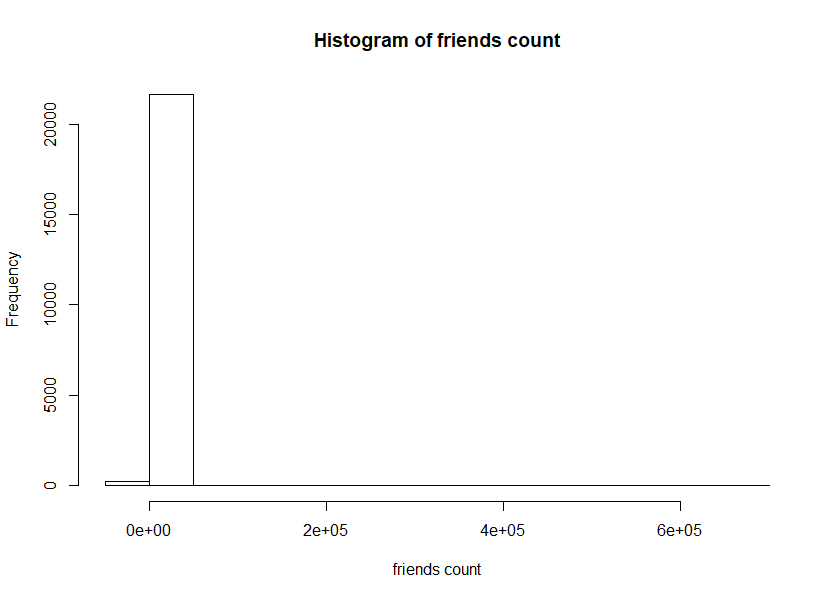
**Problem 2 (Tweeter Accounts)**

**a.** How are the data distributed for friend\_count variable?

**Code:**

*hist(twitter.df$friends\_count,xlab = "friends count",main = "Histogram of friends count")*

**result:**



**b.** Compute the summery statistics (min, 1Q, mean, median, 3Q, max) on friend\_count.

Code:

*summary(twitter.df$friends\_count)*

results:

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

The quality of Variable friends\_count bad as it has a mean of 1058 and median of 324 but the min and max values are very extreme (-84 and 660549). Outliers can be removed to make sense of the data.

**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”.

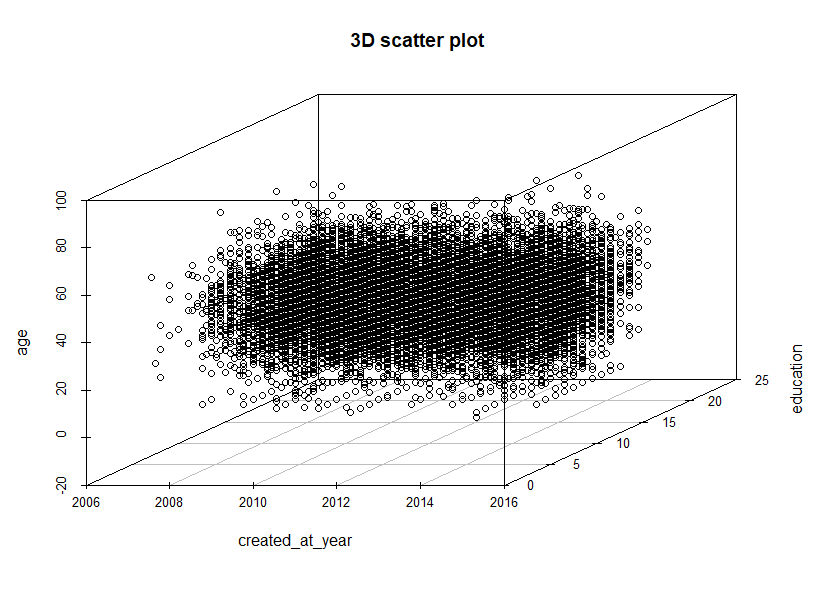
**Code:**

*library(scatterplot3d)*

*attach(twitter.df)*

*scatterplot3d(created\_at\_year, education, age , main = "3D scatter plot")*

**Result:**



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

**Code:**

*par(mfrow=c(1, 2))*

*slice<-c(650, 1000,900,300,14900)*

*labs<-c("UK", "Canada","India", "Australia","US")*

*pct <- round(slice/sum(slice)\*100)*

*lbls2 <- paste(labs, " ", pct, "%", sep="")*

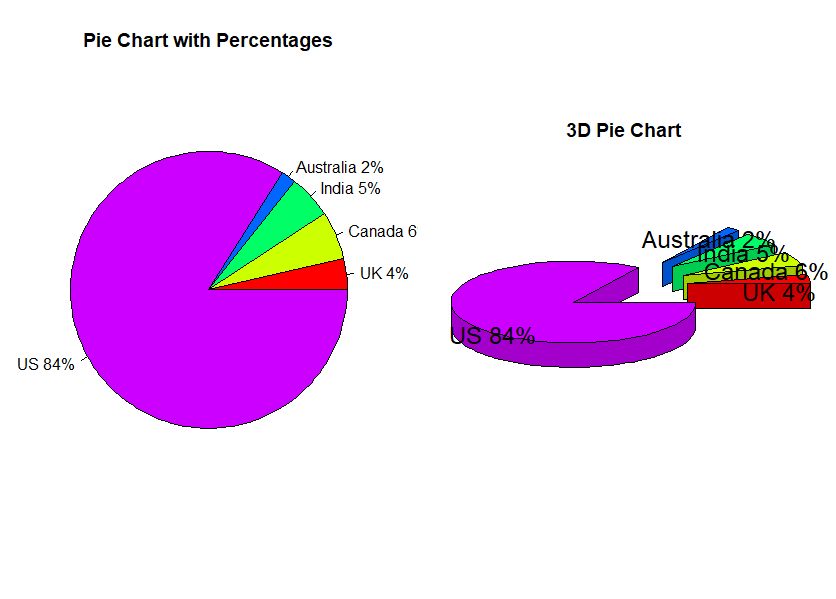
*pie(slice, labels=lbls2, col=rainbow(length(lbls2)),radius = 1,main="Pie Chart with Percentages")*

*library(plotrix)*

*pie3D(slice, labels=lbls2,explode=0.5,radius = 1,main="3D Pie Chart ")*

*par(opar)*

**Results:**



**f.** Create kernel density plot of created\_at\_year variable and interpret the result.

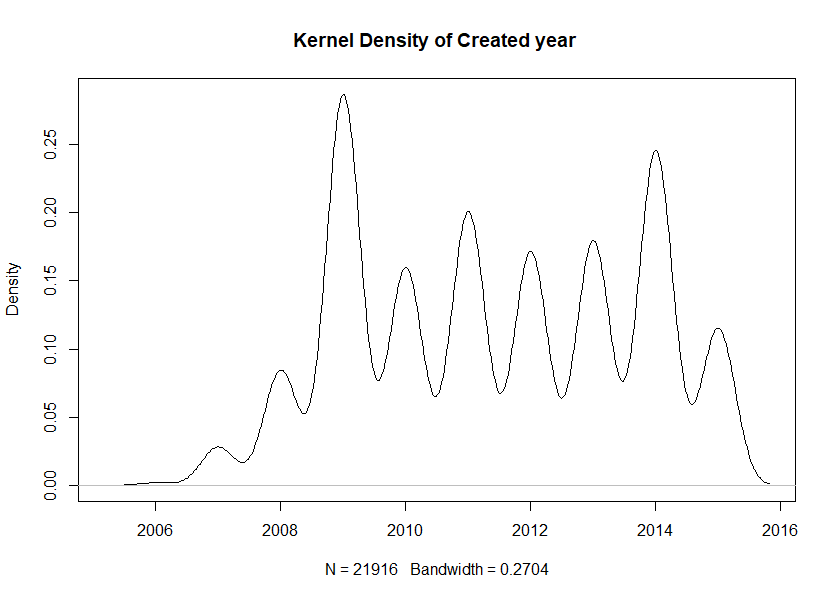
**Code:**

*d <- density(twitter.df$created\_at\_year)*

*plot(d, main="Kernel Density of Created year ")*

**Results:**

kernel density estimation is a nonparametric method for estimating the probability density function of a random variable .The probability of an account created in the year is highest in the year 2010 followed by 2014 .



**Problem 3 (Insurance Claims)**

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

**Code:**

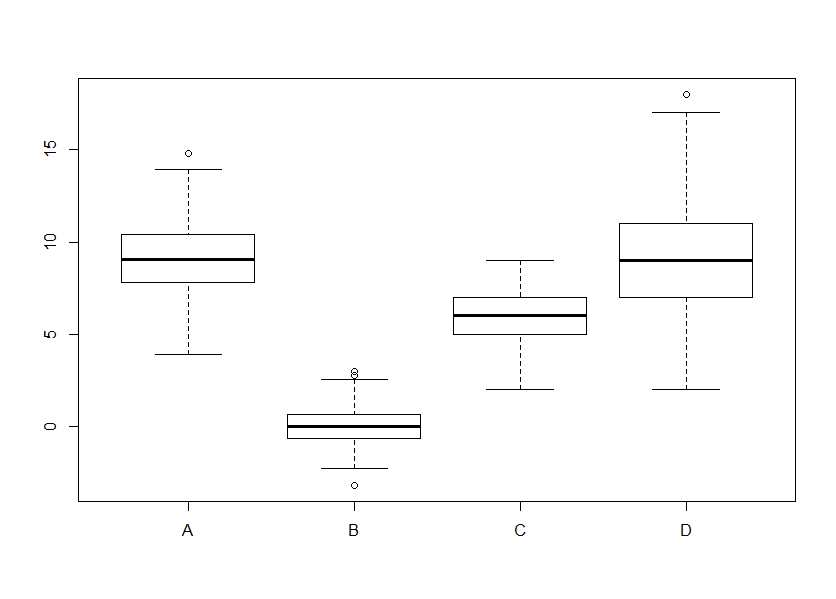
*Ndata<-scale(raw.df)*

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

**code:**

*boxplot(raw.df)*

**Result:**

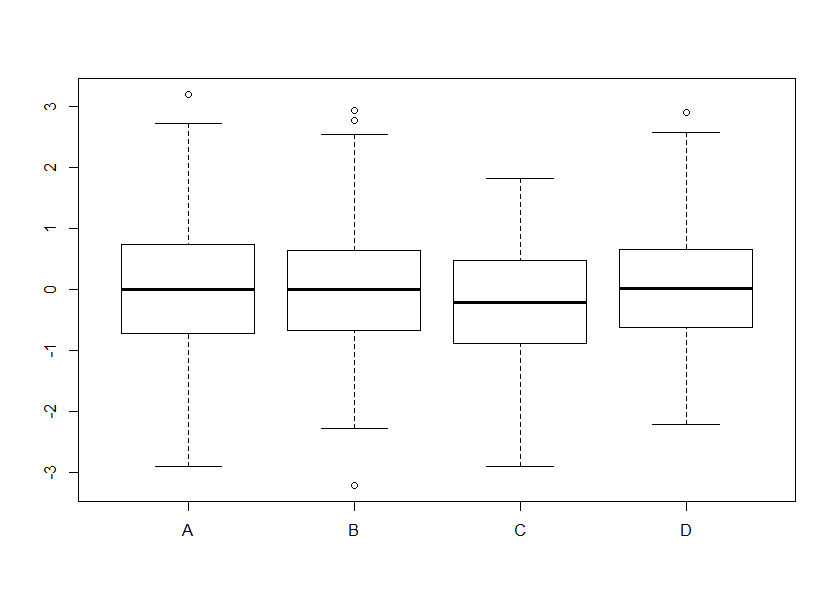


**c.** Create boxplot of all the variables in their standardized form.

**Code:**

*boxplot(Ndata)*

**Result:**



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

The data after normalization all the variables (A,B,C,D) are converted to a same scale and its easier to compare the data after normalization .

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

**code:**

*pairs(~A+B,data = raw.df, main="Basic Scatter Plot Matrix")*

**Results:**

Between the variables A and B there is no correlation The data are widely spread as we see below and there is no significant correlation.

