**Homework1**

Group 6

Akshay Rao Raghavendra Udipi

Rasagnya Reddy

857-260-9324 (Tel of Akshay Rao)

857-891-1132 (Tel of Rasagnya)

[udipiragavendrarao.a@husky.neu.edu](mailto:udipiragavendrarao.a@husky.neu.edu)

[avala.r@husky.neu.edu](mailto:avala.r@husky.neu.edu)

Percentage of Effort Contributed by Akshay Rao: **50%**

Percentage of Effort Contributed by Rasagnya: **50%**

Signature of Akshay Rao: **Akshay Rao**

Signature of Rasagnya: **Rasagnya**

Submission Date: 02-11-2019

## Problem 1 (Forest Fires) [40 points] The file forestfires.xlsx includes data from Cortez and Morais (2007). The output area was first transformed with a ln(x+1) function. Then, several data mining methods were applied. After fitting the models, the outputs were post-processed with the inverse of the ln(x+1) transform. Four different input setups were used. The experiments were conducted using a 10-fold (cross-validation) ⊆ 30 runs. Two regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed with only 4 direct weather conditions (temp, RH, wind and rain) obtained the best MAD value: 12.71 ± 0.01 (mean and confidence interval within 95% using a t-student distribution). The best RMSE was attained by the naive mean predictor. An analysis to the regression error curve (REC) shows that the SVM model predicts more examples within a lower admitted error. In effect, the SVM model predicts better small fires, which are the majority. Number of instances and attributes are 517 and 13 respectively.

#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.

forestfires.df <- read.csv("C:\\Users\\arjun\\OneDrive - Northeastern University\\data mining\\forestfires.csv")

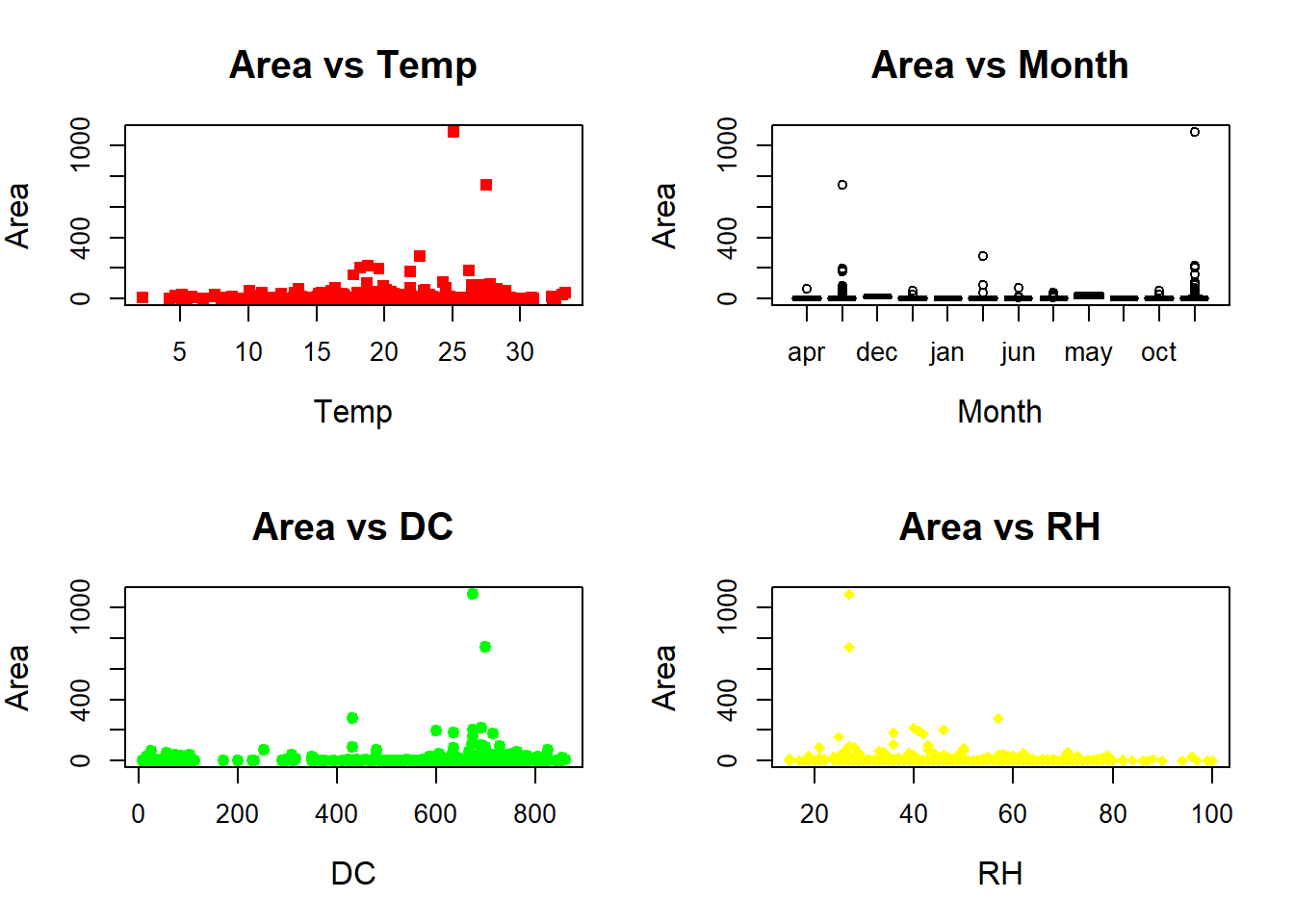
par(mfrow = c(2,2))

plot(forestfires.df$temp,forestfires.df$area,main="Area vs Temp", xlab="Temp", ylab="Area", col="red", cex.main=1.5,cex.lab=1.25, cex.axis=1.0,pch=15)

plot(forestfires.df$month,forestfires.df$area, main="Area vs Month", xlab="Month", ylab="Area", cex.main=1.5,cex.lab=1.25,cex.axis=1.0, col="blue")

plot(forestfires.df$DC,forestfires.df$area,main="Area vs DC", xlab="DC", ylab="Area", cex.main=1.5,cex.lab=1.25, cex.axis=1.0, col="green", pch=19)

plot(forestfires.df$RH,forestfires.df$area, main="Area vs RH", xlab="RH", ylab="Area", cex.main=1.5,cex.lab=1.25,cex.axis=1.0, col="yellow", pch=18)

****

library("dplyr")

library(ggplot2)

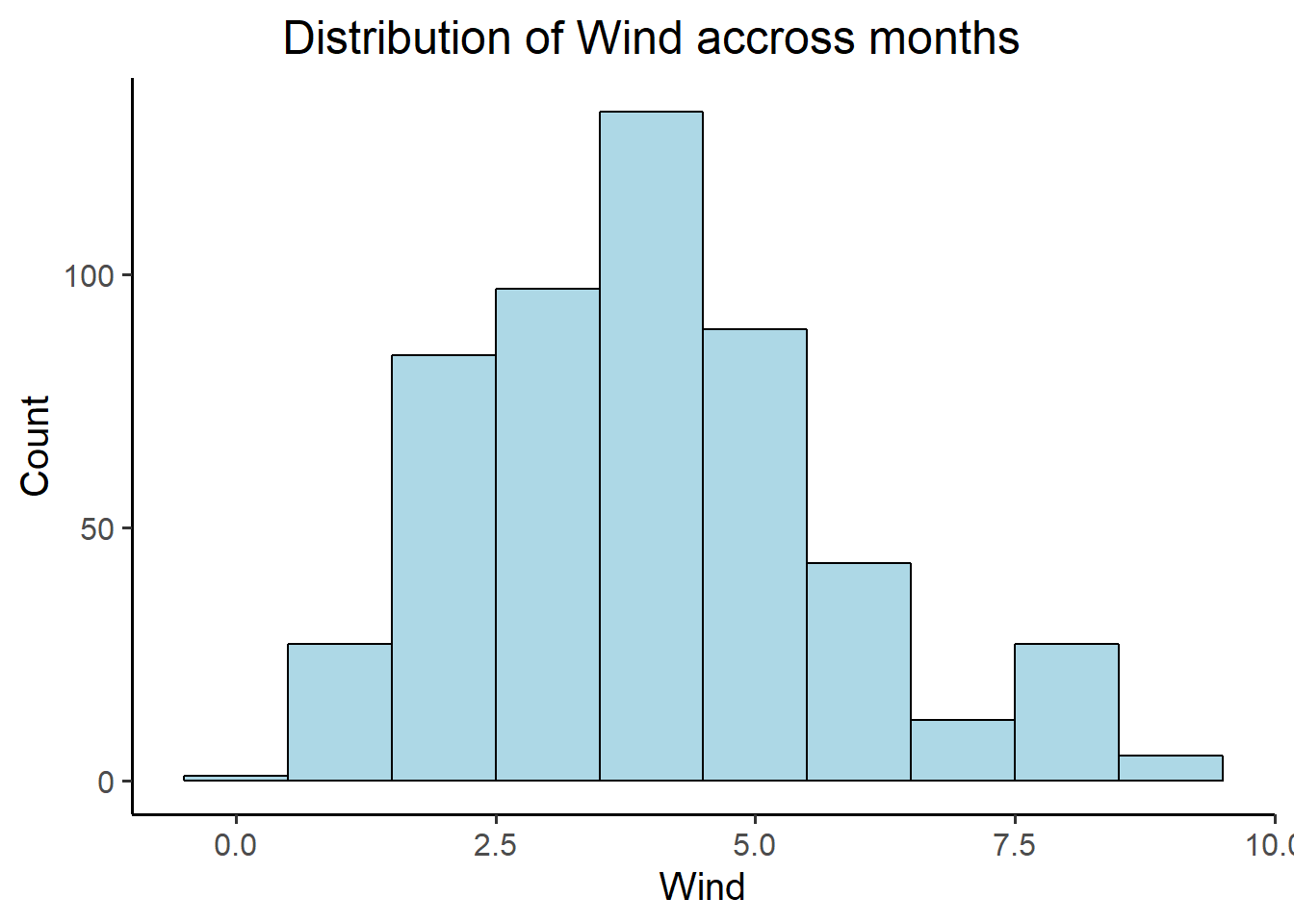
#b)Plot the histogram of wind speed (km/h) from January through December.

ggplot(forestfires.df, aes(x=forestfires.df$wind)) +

geom\_histogram(color="Black",fill="light blue", binwidth = 1) +

labs(title=" Distribution of Wind accross months", x="Wind", y = "Count") +

theme\_classic(base\_size = 15)

****

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

summary(forestfires.df$wind)

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

## 0.400 2.700 4.000 4.018 4.900 9.400

#d)Density curve to histogram

ggplot(forestfires.df, aes(x=forestfires.df$wind)) +

geom\_histogram(aes(y =..density..),

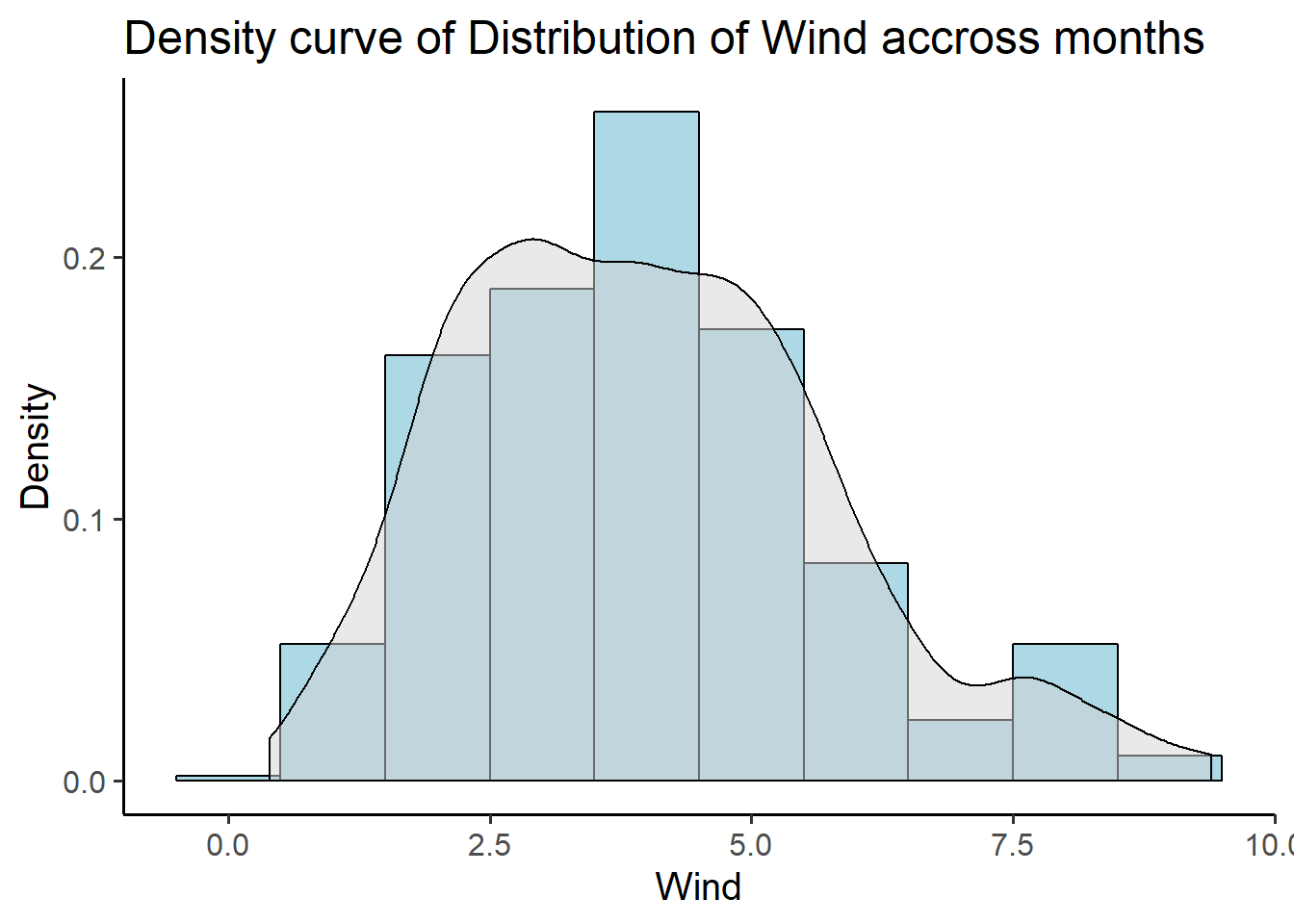
binwidth=1,

colour="black", fill="light blue") +

geom\_density(alpha=.5, fill="light grey") +

labs(title="Density curve of Distribution of Wind accross months", x="Wind", y="Density") +

theme\_classic(base\_size = 15)

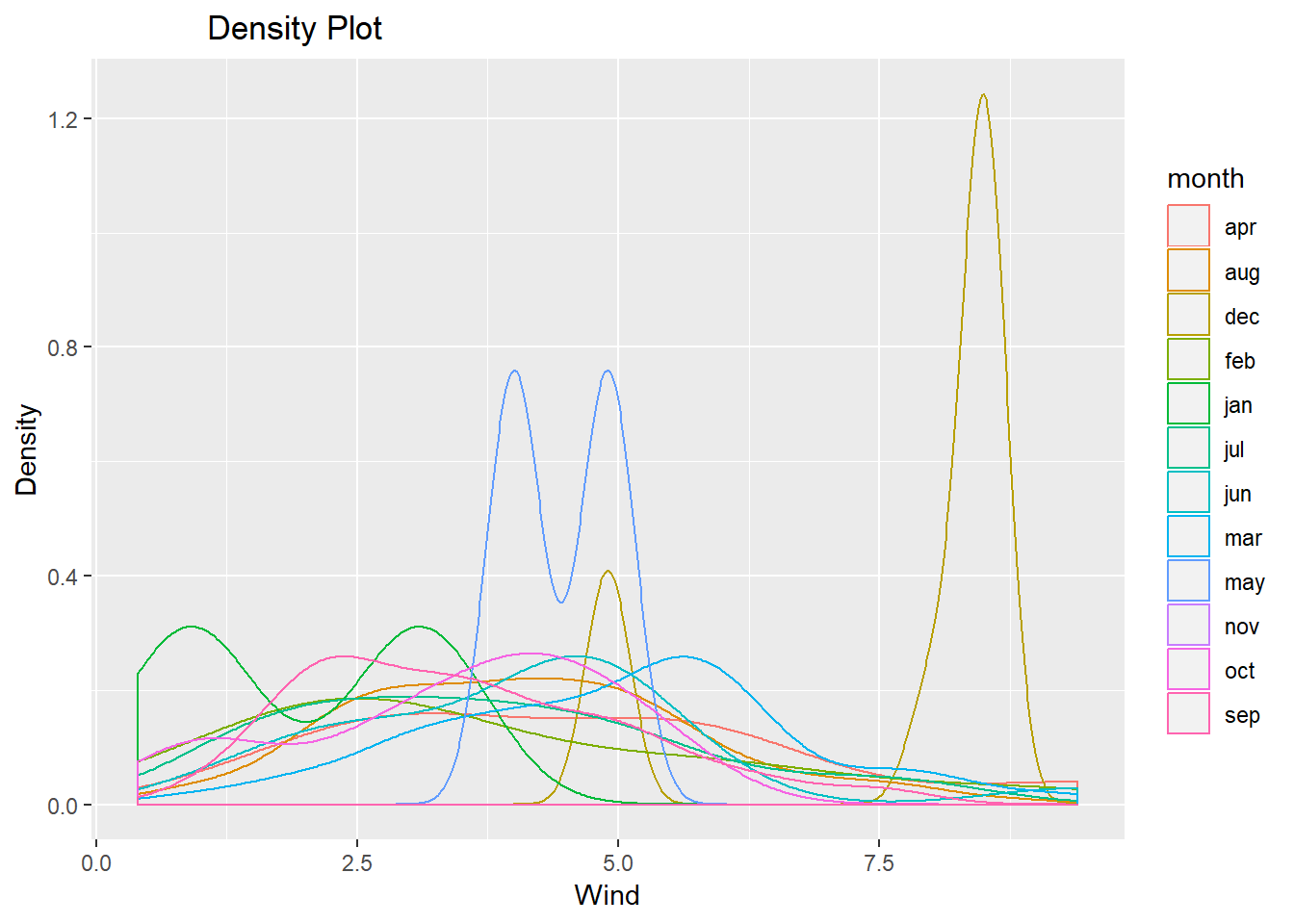
****

#e) Plot the density function of months. Use different colors in the graph to interpret your result clearly.

qplot(wind, geom = 'density', color = month, data = forestfires.df) +

labs(title=" Density Plot", x="Wind", y = "Density")

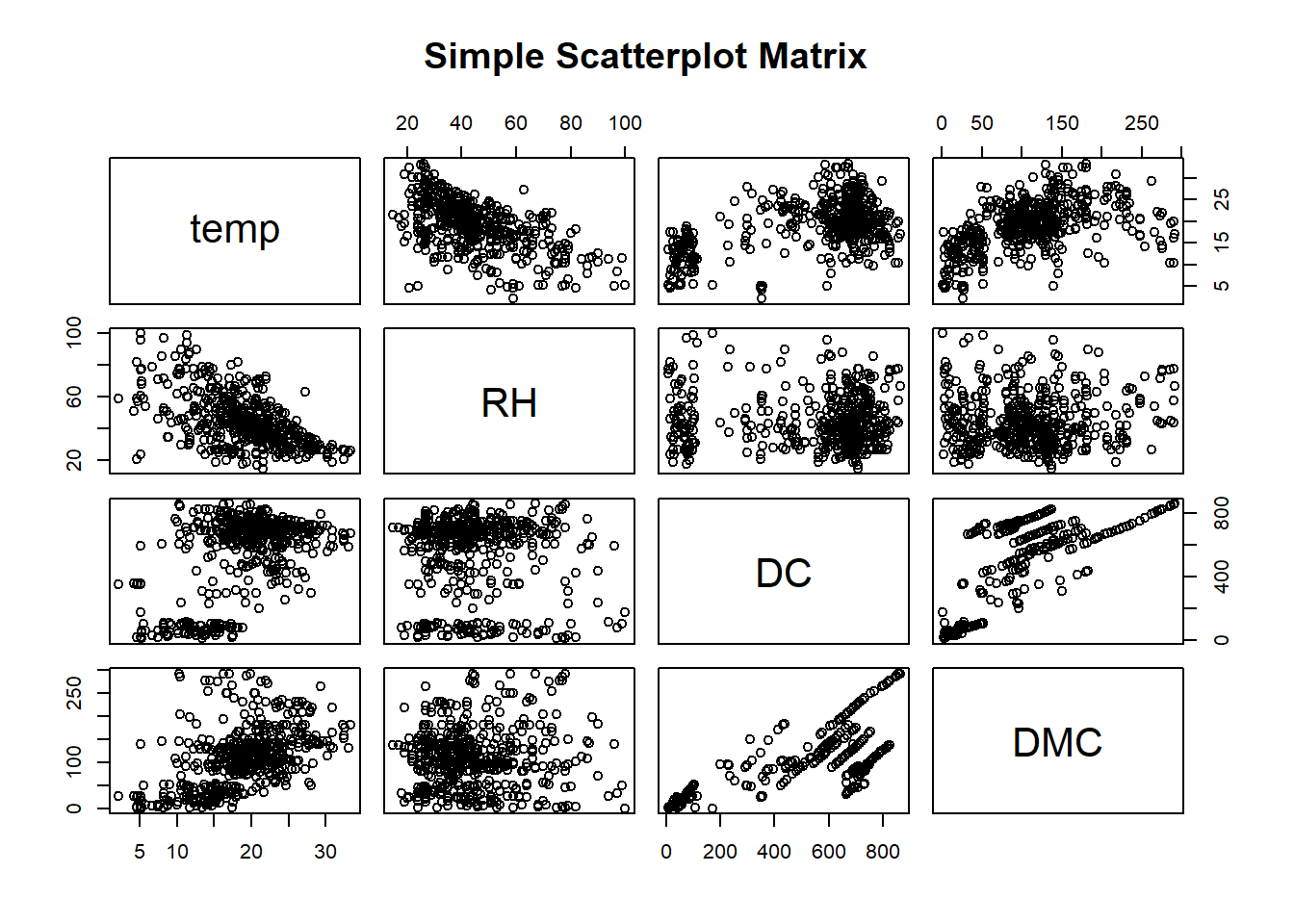
## Warning: Groups with fewer than two data points have been dropped.

****

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

pairs(~temp+RH+DC+DMC,data=forestfires.df,

main="Simple Scatterplot Matrix")

****Inferences : 1) In the graph Temp versus RH , they are Inversely correlated. The Values of them ranges from less than 0 to nearly -1.

1. In the graph Temp Versus DMC, they are inversely correlated as well and their value is less than 0 to -1.
2. In the Graph DC vs DMC, we see a positive correlation and the correlation coefficient ranges from greater than 0 to close to 1.
3. RH Vs DMC, this graph shows no correlation and these attributes are independent.
4. Similarly, we see RH Vs DC , RH vs DC , Temp vs DC , Temp vs DMC has no correlation between them.

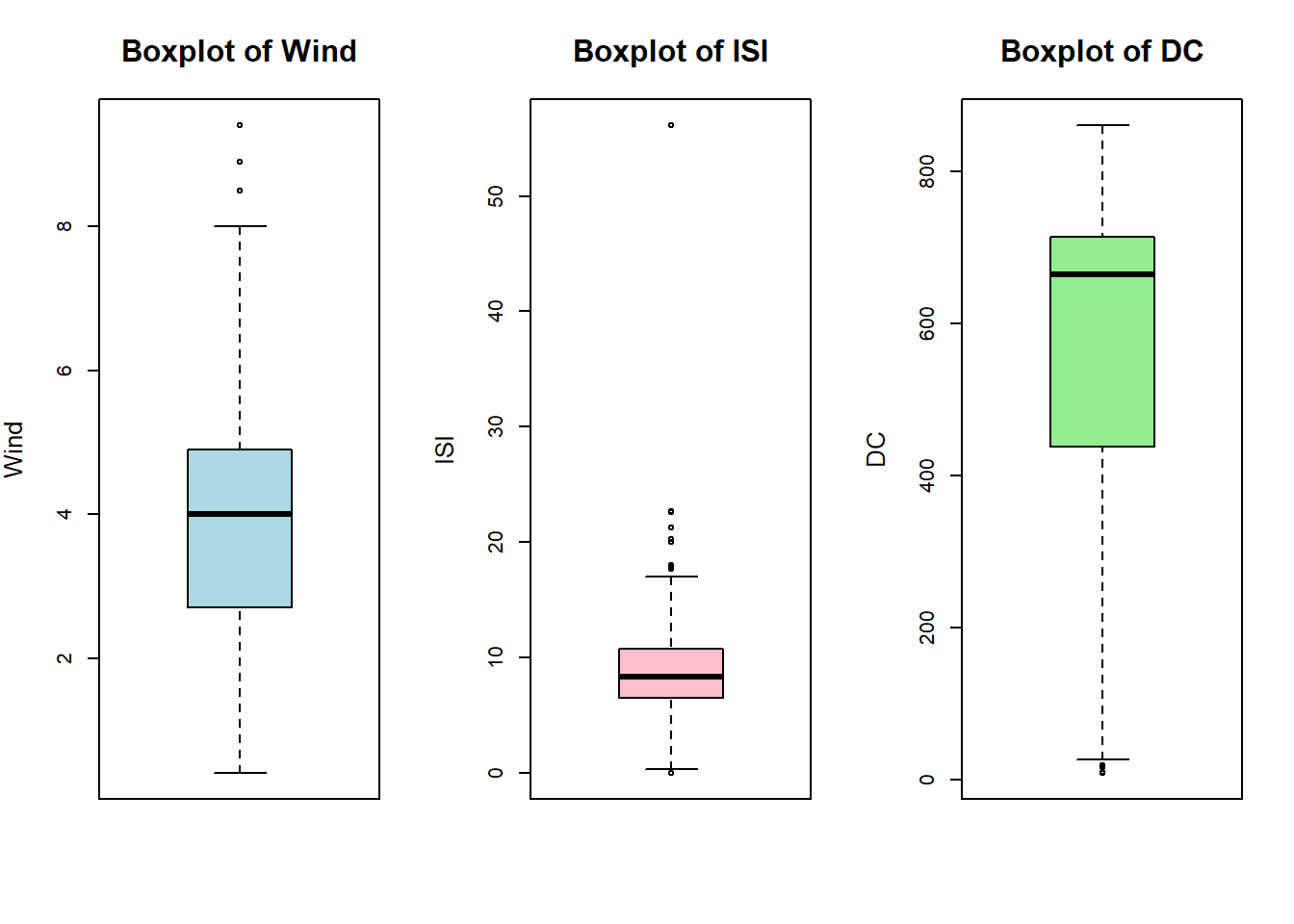
#g)Create boxplot for wind, ISI and DC. Are there anomalies/outliers. Interpret your result.

par(mfcol=c(1,3))

boxplot(forestfires.df$wind, col="light blue", main="Boxplot of Wind", ylab="Wind", cex.main=1.5,cex.lab=1.25)

boxplot(forestfires.df$ISI, col="pink", main="Boxplot of ISI", ylab="ISI", cex.main=1.5,cex.lab=1.25)

boxplot(forestfires.df$DC, col="light green", main="Boxplot of DC", ylab="DC", cex.main=1.5,cex.lab=1.25)

****Interpretation:

1. Yes, there are outliers, We will have t treat the outliers by eliminating them, this can affect the accuracy of the model.
2. There are more outliers in ISI.
3. Attributes Wind and ISI has data evenly distibuted over its mean, while data in DC is skewed.

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

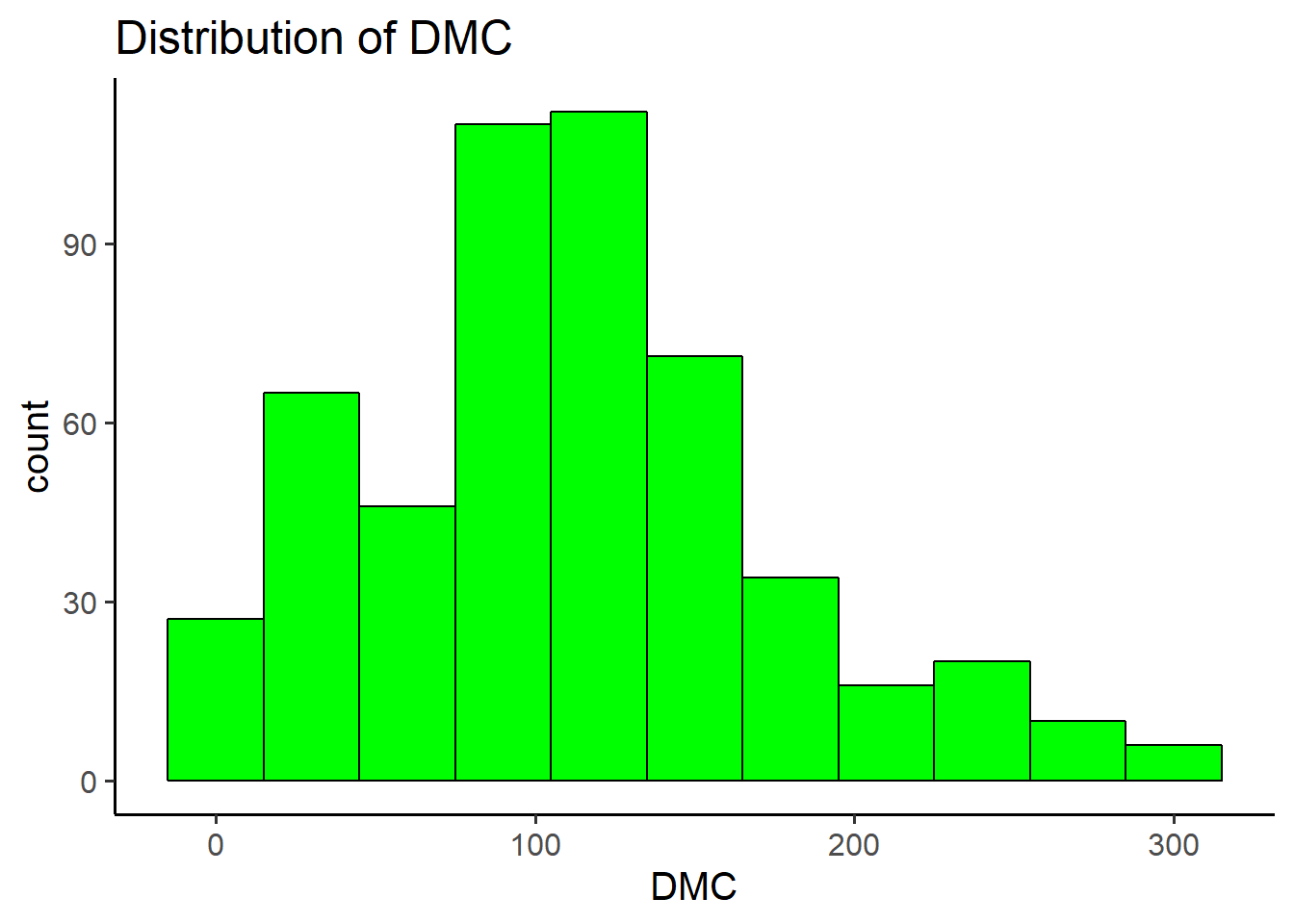
par(mfrow=c(1,2))

ggplot(forestfires.df, aes(x=forestfires.df$DMC)) +

geom\_histogram(binwidth=30, color="black", fill="green") +

labs(title="Distribution of DMC", x="DMC") +

theme\_classic(base\_size = 15)

****

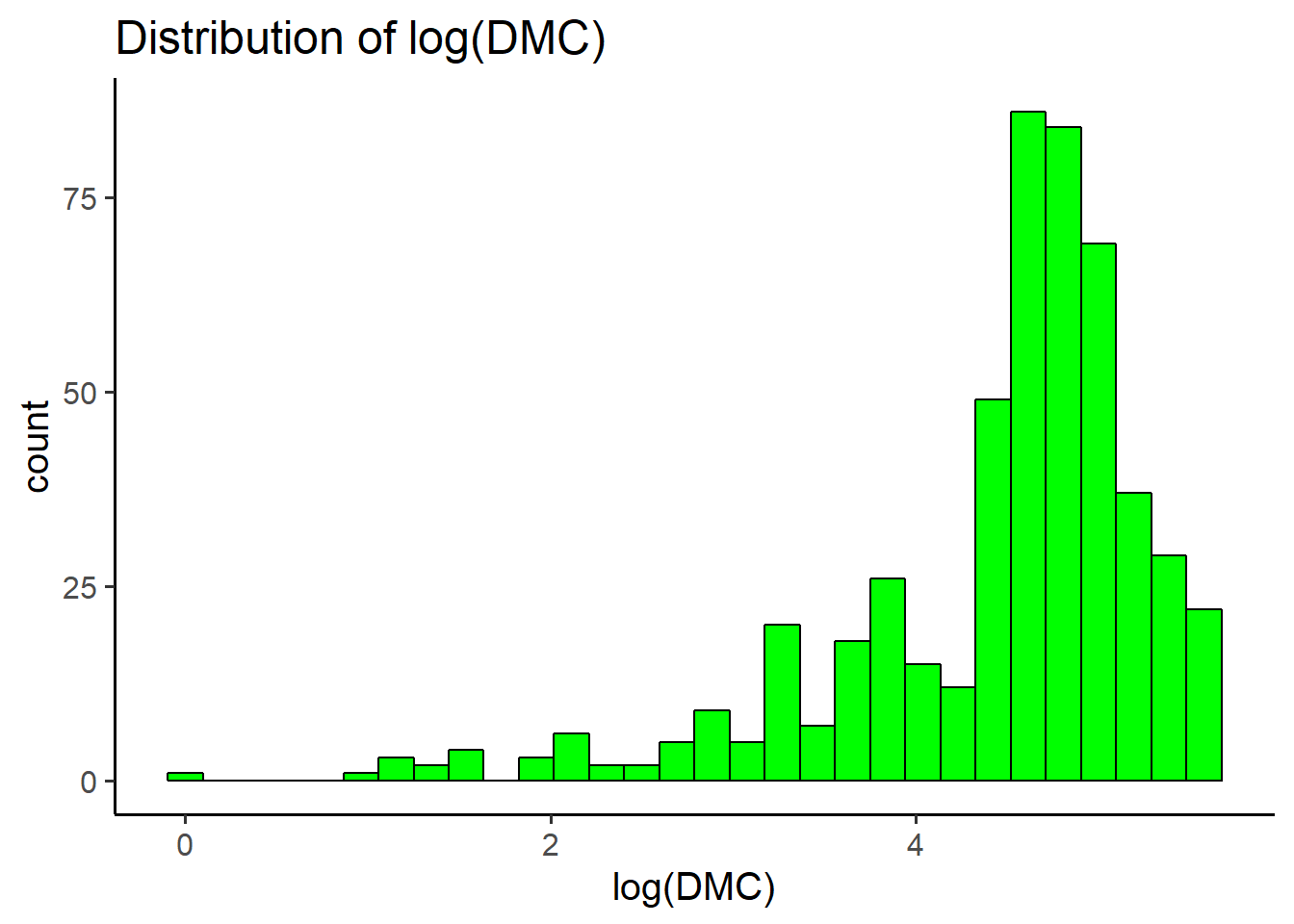
ggplot(forestfires.df, aes(x=log(forestfires.df$DMC))) +

geom\_histogram(color="black", fill="green") +

labs(title="Distribution of log(DMC)", x="log(DMC)") +

theme\_classic(base\_size = 15)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

****

Interpretation:

DM - Duff Moisture Code.

1. The first Histogram has large bin size which makes it hard for the data interpreation of how the data is distributed and the density of DMC between 100 and 200.
2. Log of DMC gives a clear picture as to the properties of moisture content in atmosphere with relatively interprtable bins and their distributions.

## Problem2 (Tweeter Accounts) [40 **points**] Twitter is a social news website. It can be viewed as a hybrid of email, instant messaging and sms messaging all rolled into one neat and simple package. It’s a new and easy way to discover the latest news related to subjects you care about. This is the data set crawled on July, 2009. BlogCatalog is a social blog directory website. This contains the friendship network crawled. For easier understanding, all the contents and variables are organized in CSV file format.

twitter=read.csv("C:\\Users\\arjun\\OneDrive - Northeastern University\\data mining\\M01\_quasi\_twitter.csv")

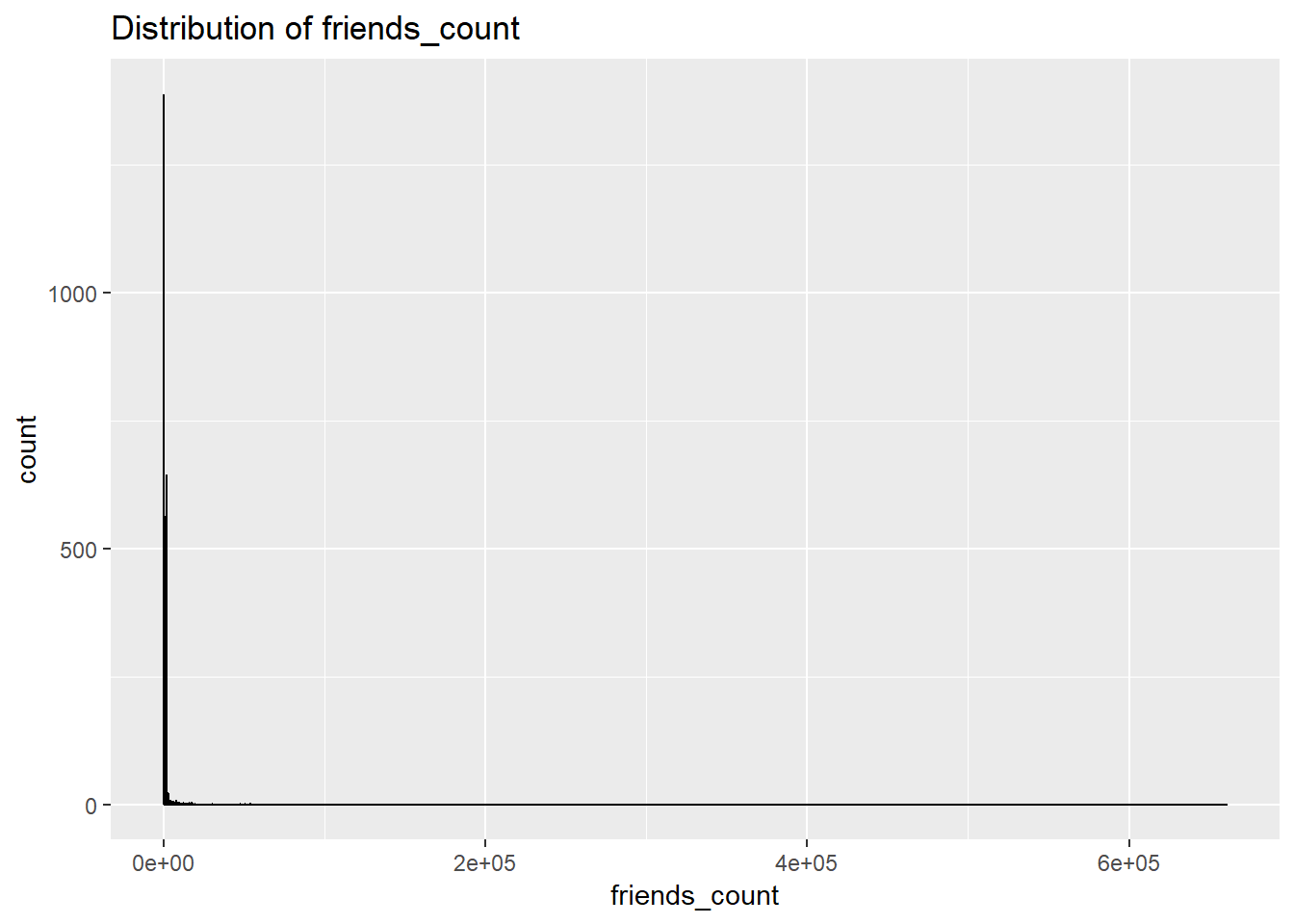
#a)How are the data distributed for friends\_count variable?

library(ggplot2)

ggplot(twitter, aes(x=twitter$friends\_count)) +

geom\_histogram(color="black", fill="light blue",binwidth=30) +

labs(title="Distribution of friends\_count", x="friends\_count")# +

****

#theme\_classic(base\_size = 15)

Cannot interpret anything hence plotting the descdist.

library(fitdistrplus)

## Loading required package: MASS

##

## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':

##

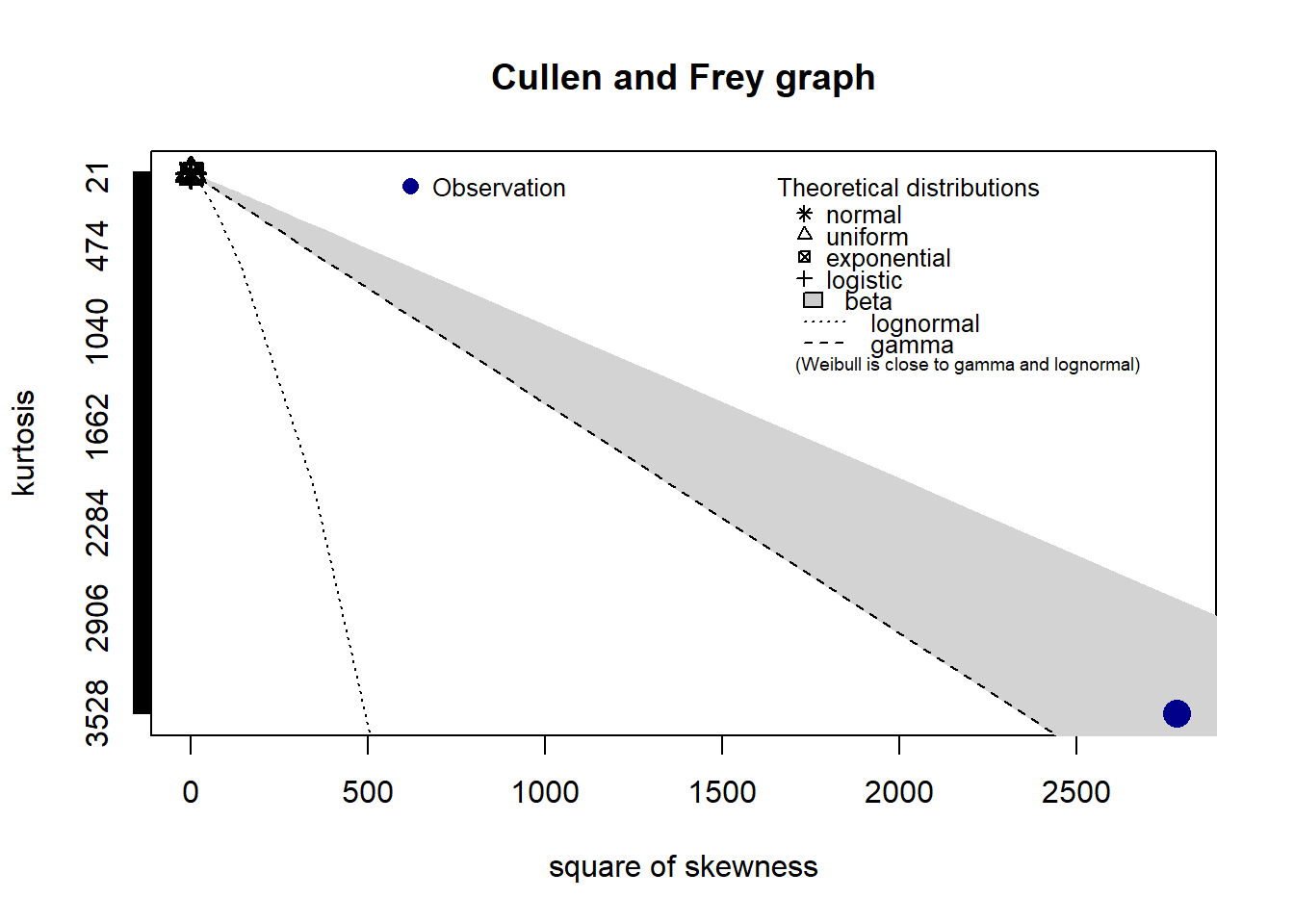
## select

## Loading required package: survival

## Loading required package: npsurv

## Loading required package: lsei

descdist(twitter$friends\_count, discrete = F)

****

## summary statistics

## ------

## min: -84 max: 660549

## median: 324

## mean: 1057.911

## estimated sd: 8125.054

## estimated skewness: 52.73123

## estimated kurtosis: 3527.241

It Can be interpretted that the data in Friends\_count is Log-normally Distributed.

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

summary(twitter$friends\_count)

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

## -84 123 324 1058 849 660549

sd(twitter$friends\_count)

## [1] 8125.054

c) How are the data quality in friend\_count variable? Interpret your answer?

Ans : After Interpreting the summary and standard deviation for the attribute Friends count in Twitter Data, we can conclude that the data is has a large variation. Data is right skewed and has a large number of outliers.

#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(twitter)

scatterplot3d(twitter$created\_at\_year, twitter$education, twitter$age, pch=16, main="3D Scatter Plot",

xlab="Created\_at\_year",

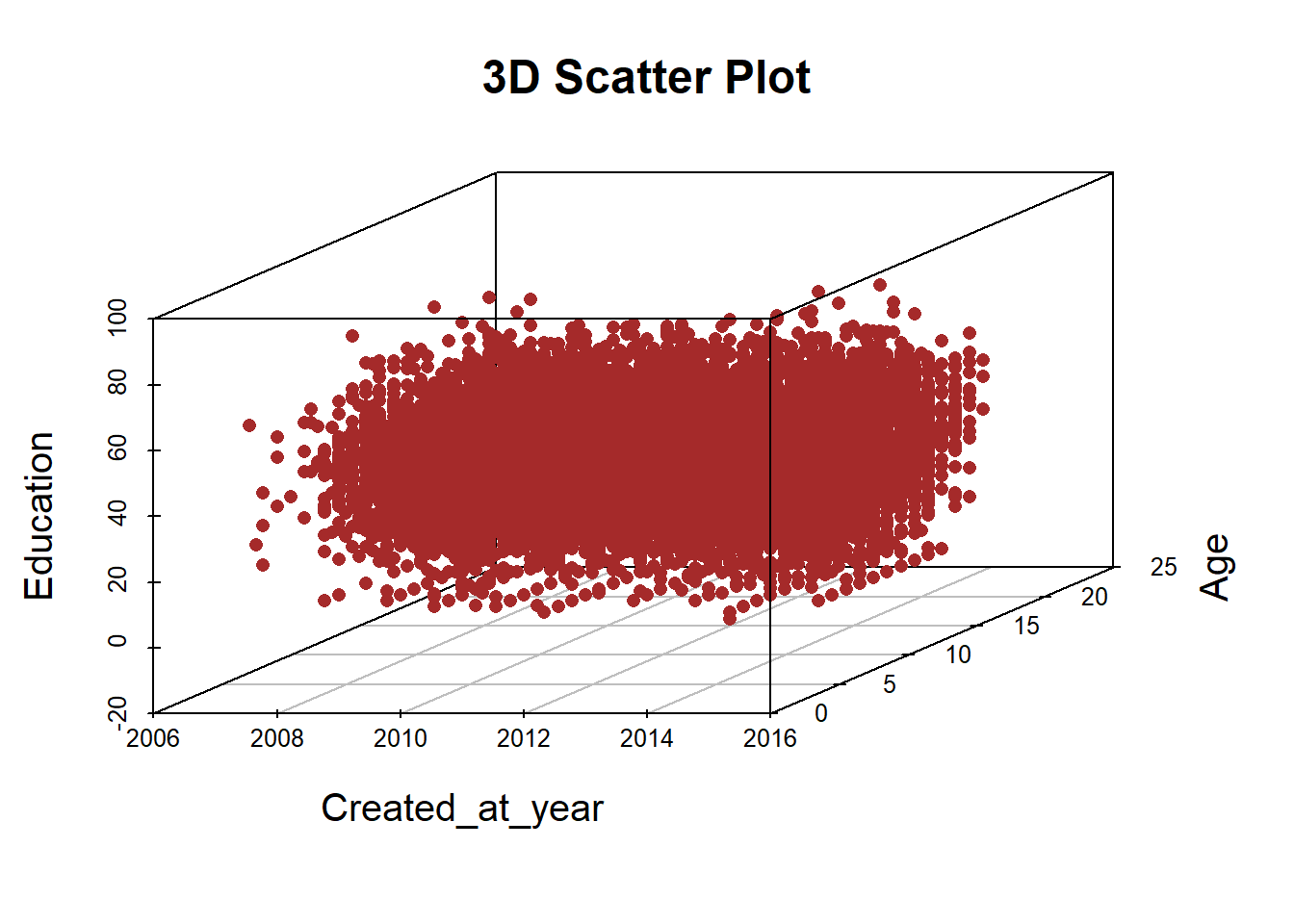
ylab="Age",

zlab="Education",

color="brown",

cex.main=1.5,

cex.lab=1.25)

****

#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

par(mfrow=c(1,2))

slices <- c(650,1000,900,300,14900)

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

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

lbls <- paste(lbls, pct) # add percents to labels

lbls <- paste(lbls,"%",sep="") # ad % to labels

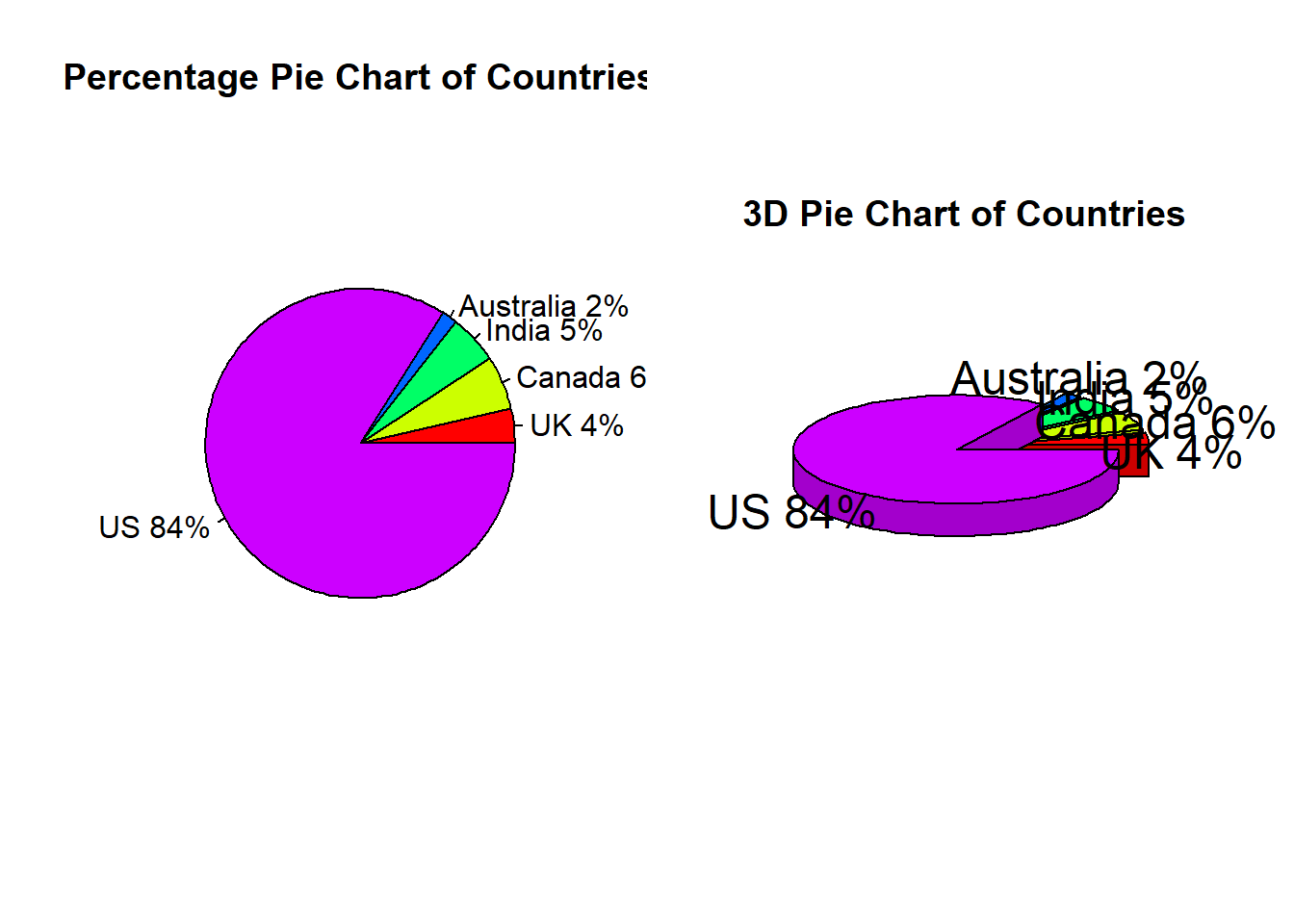
pie(slices,labels = lbls, col=rainbow(length(lbls)),

main="Percentage Pie Chart of Countries")

library(plotrix)

pie3D(slices,labels=lbls,explode=0.1,

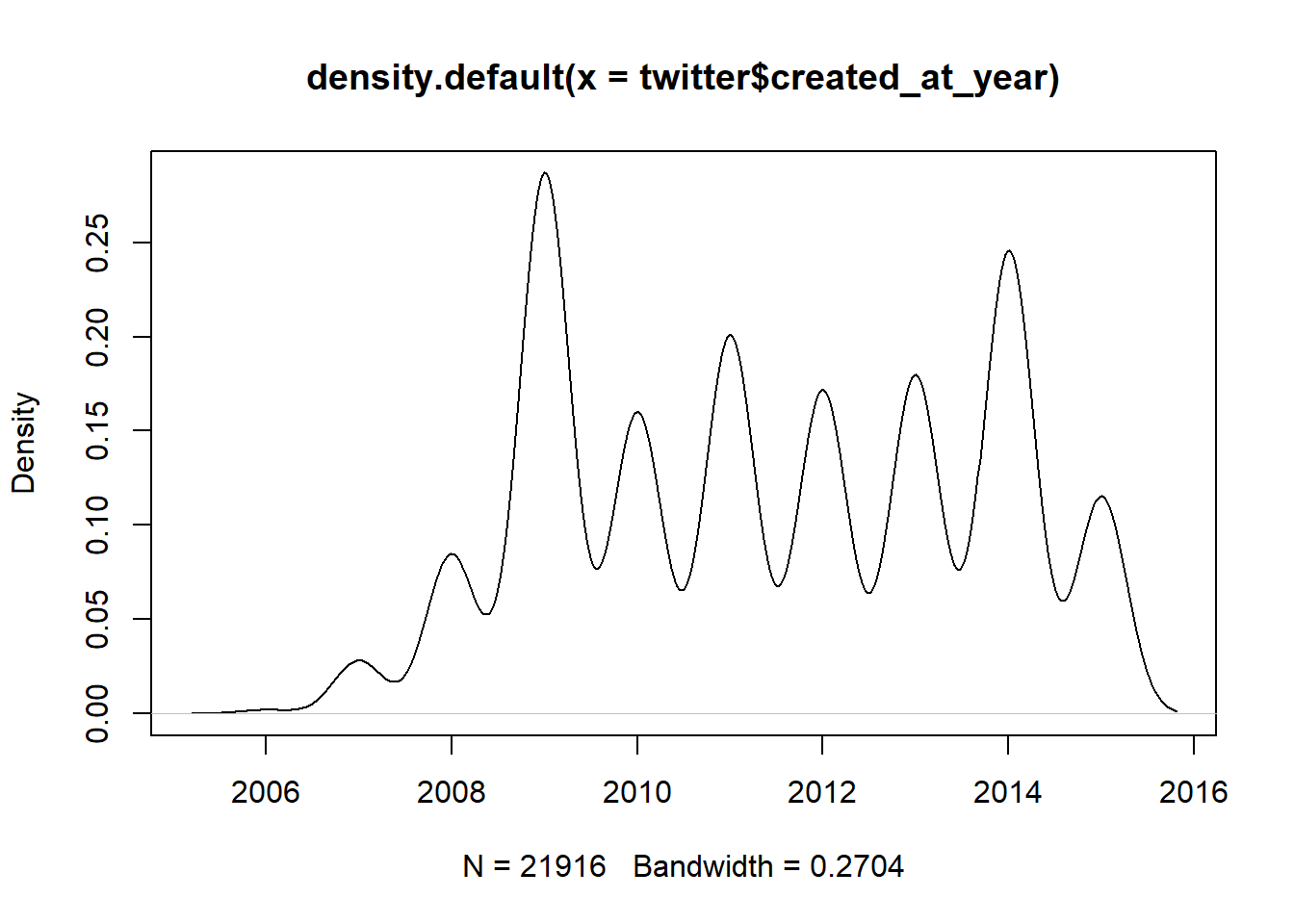
main="3D Pie Chart of Countries ")

****

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

d=density(twitter$created\_at\_year)

plot(d)

****Interpretation : It Can be concluded from the plot that most of the users signed up in 2009 in the given data sample the second highest number of user signed up in 2014. It can also interpreted that the sample has high number of data of users who signed up to Twitter from late 2008 to start of 2015.

## Problem3 (Insurance Claims) [20 points] Consider that we need to rate a product based on four different aspects Sustainability, Carbon footprint, weight and required power to be built. Those variables are gathered into raw\_data.csv spreadsheet in columns A, B, C and D respectively.

data=read.csv("C:\\Users\\arjun\\OneDrive - Northeastern University\\data mining\\raw\_data.csv")

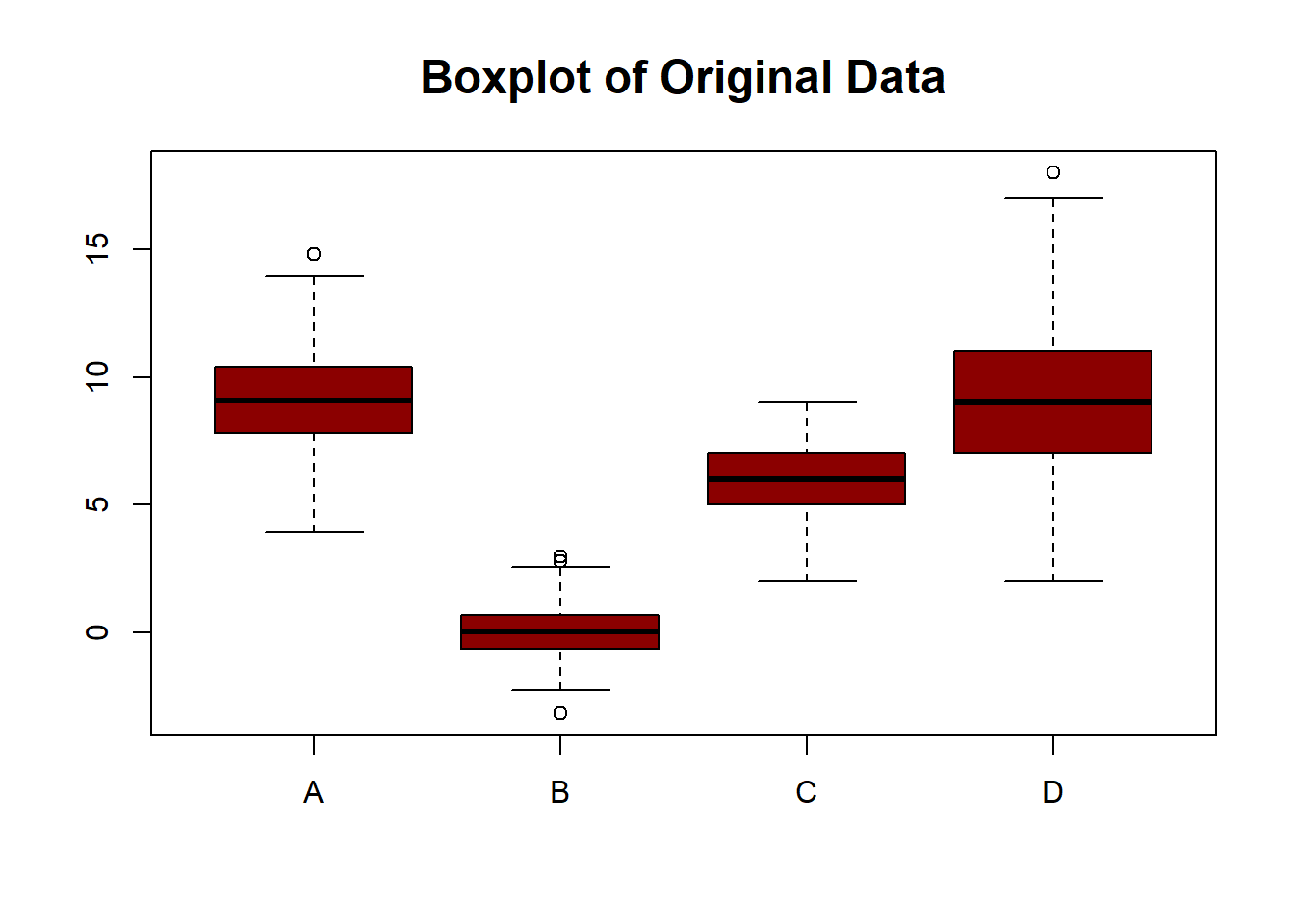
#a)Normalize the data and create new dataset with normalized data and name it Ndata.

library(standardize)

Ndata=data.frame(scale(data, center = TRUE, scale=TRUE))

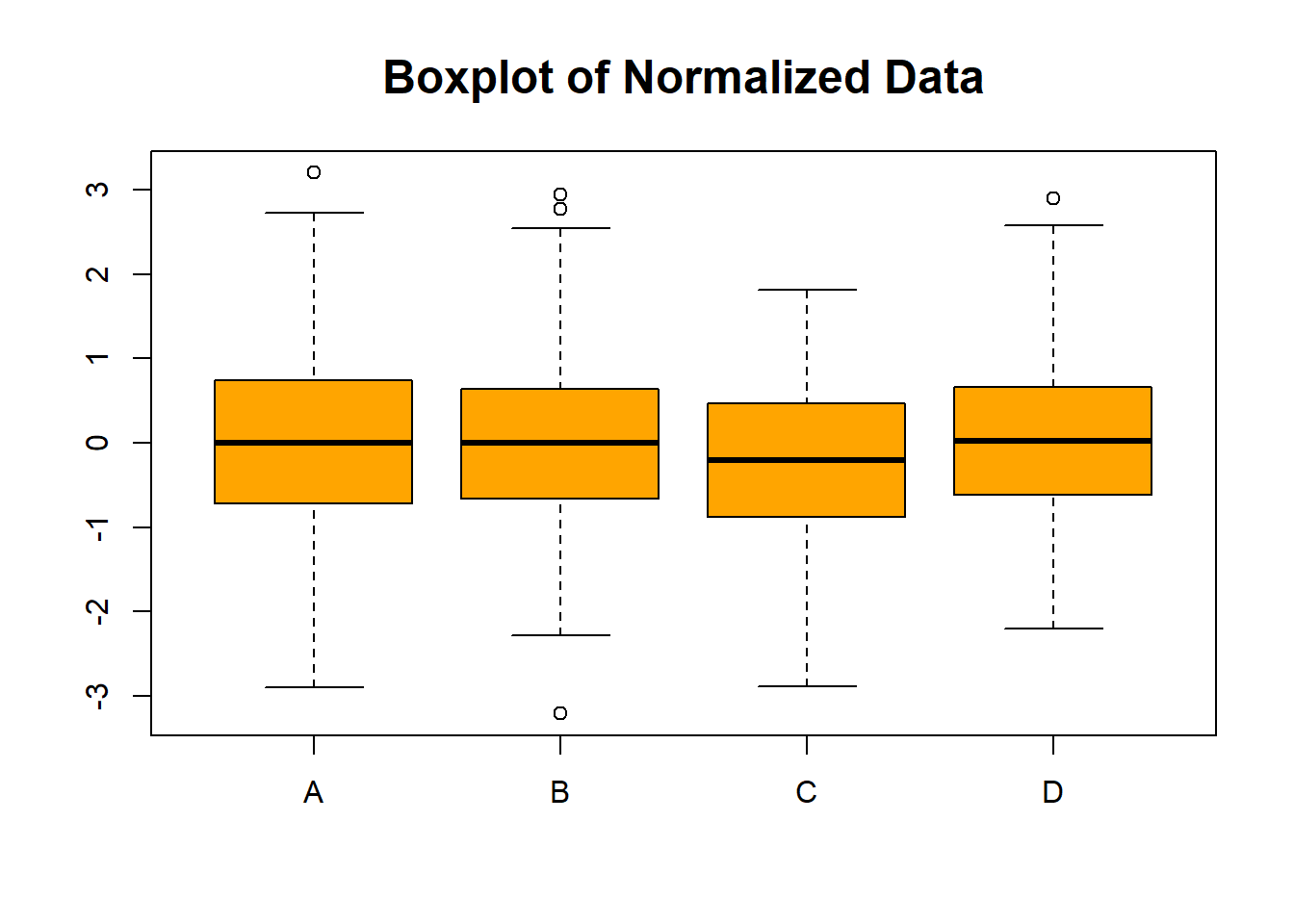
#b)Create the boxplot of all the variables in their original form.

boxplot(data,col="dark red",main="Boxplot of Original Data",cex.main=1.5, cex.lab=1.25)

****

#c)Create boxplot of all the variables in their normalized form.

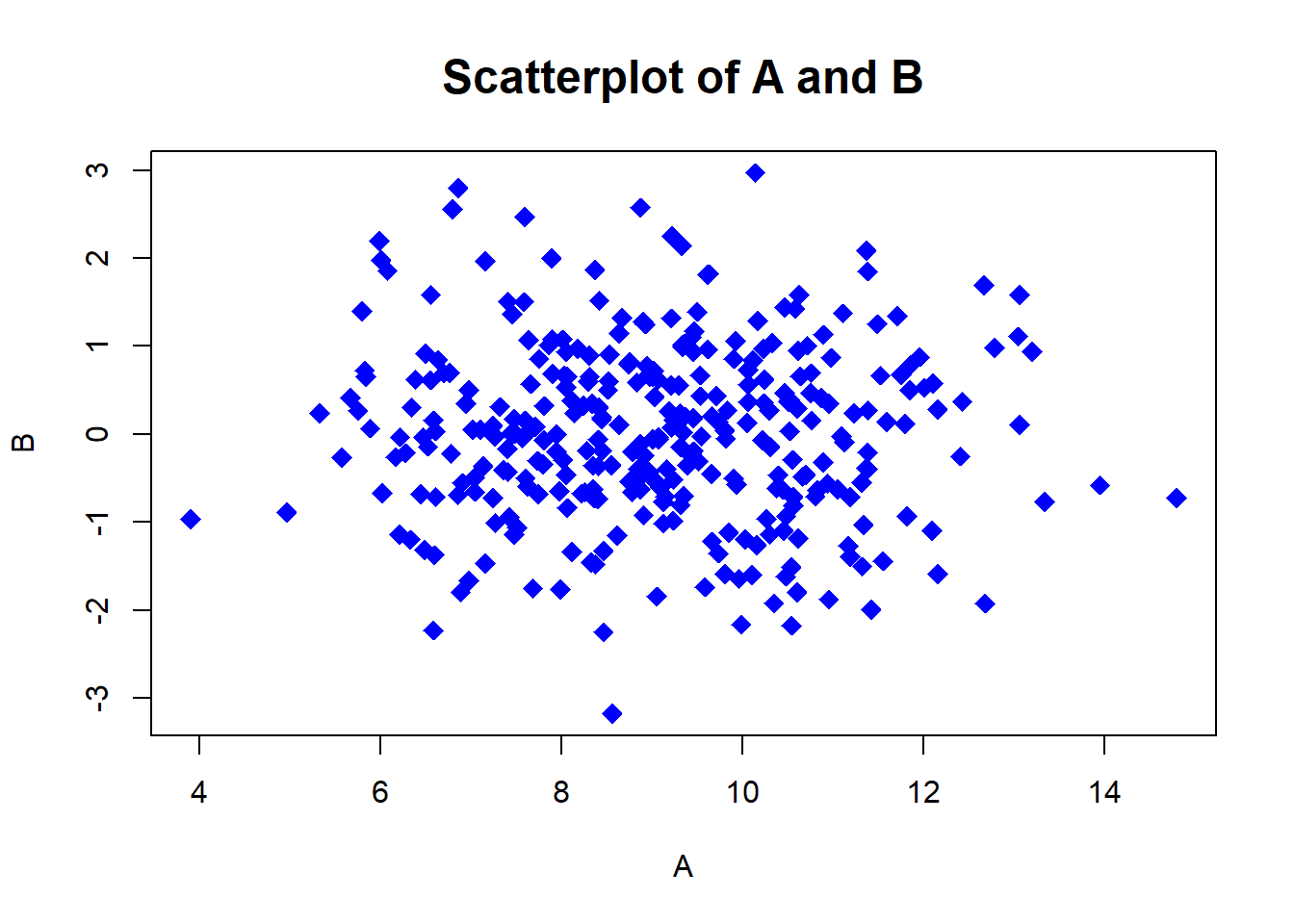
boxplot(Ndata,col="orange",main="Boxplot of Normalized Data",cex.main=1.5, cex.lab=1.25)

****Interpretations:

1. Sustainability, Carbon footprint, weight and required power to be built named A, B, C, D are measured in different units hence their values differ in ranges in the first plot whereas the standardized plot makes it easy for understanding as to how these variables are on the same scale and unit of measure.
2. The outliers are more distinctive , which makes it easy for interpretation for building further model.

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

plot(data$A,data$B,main="Scatterplot of A and B", xlab="A", ylab="B", cex.label=1.25, cex=1.5,cex.main=1.5, pch=18, col="Blue")

****

Interpretation:

There is no correlation between the two variables, i.e Sustainability and Carbon Footprint are independent to each other.