DataMiningRajeevMoetwani

Problem 1

Question a

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ tidyv
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(gridExtra)
```

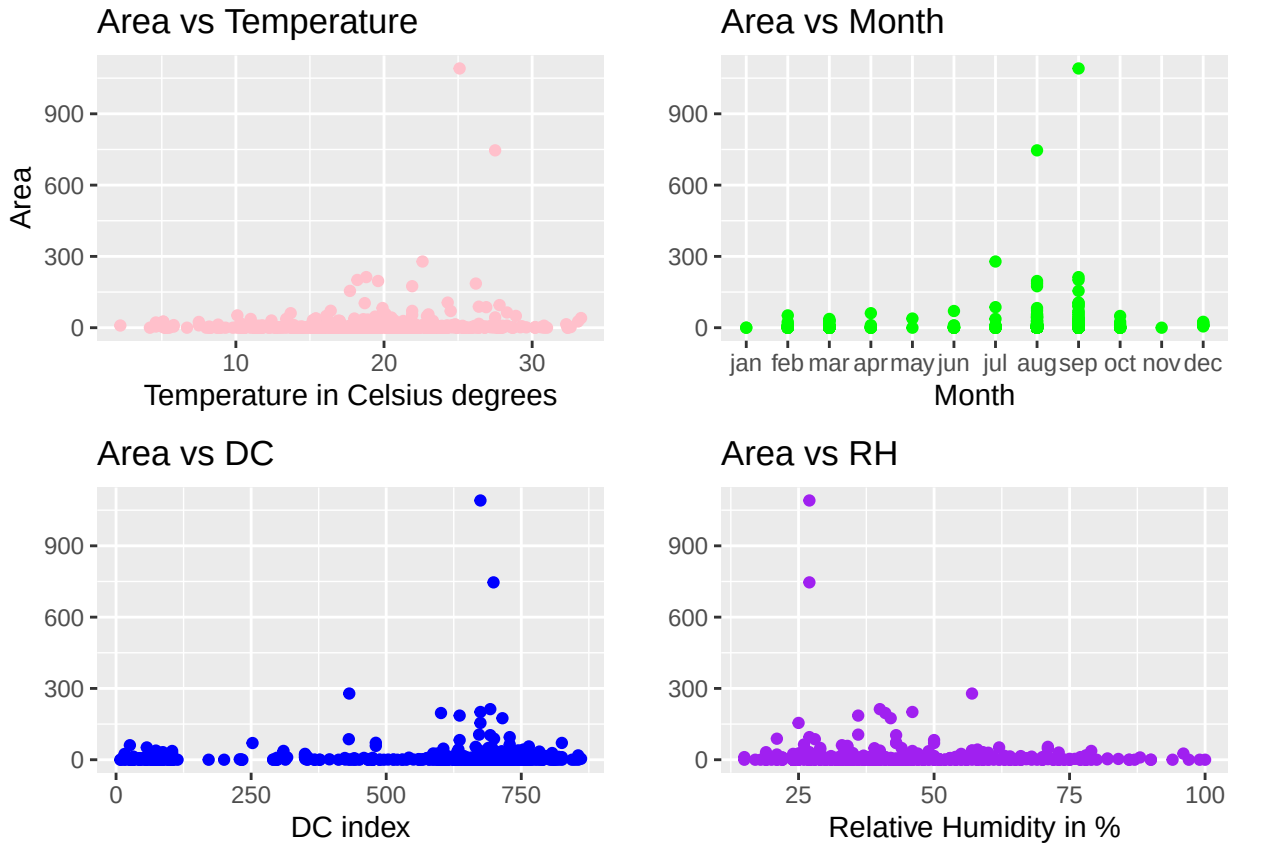
```
##
## Attaching package: 'gridExtra'
```

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

```
forestFires_data <- read.csv("forestfires.csv")
```

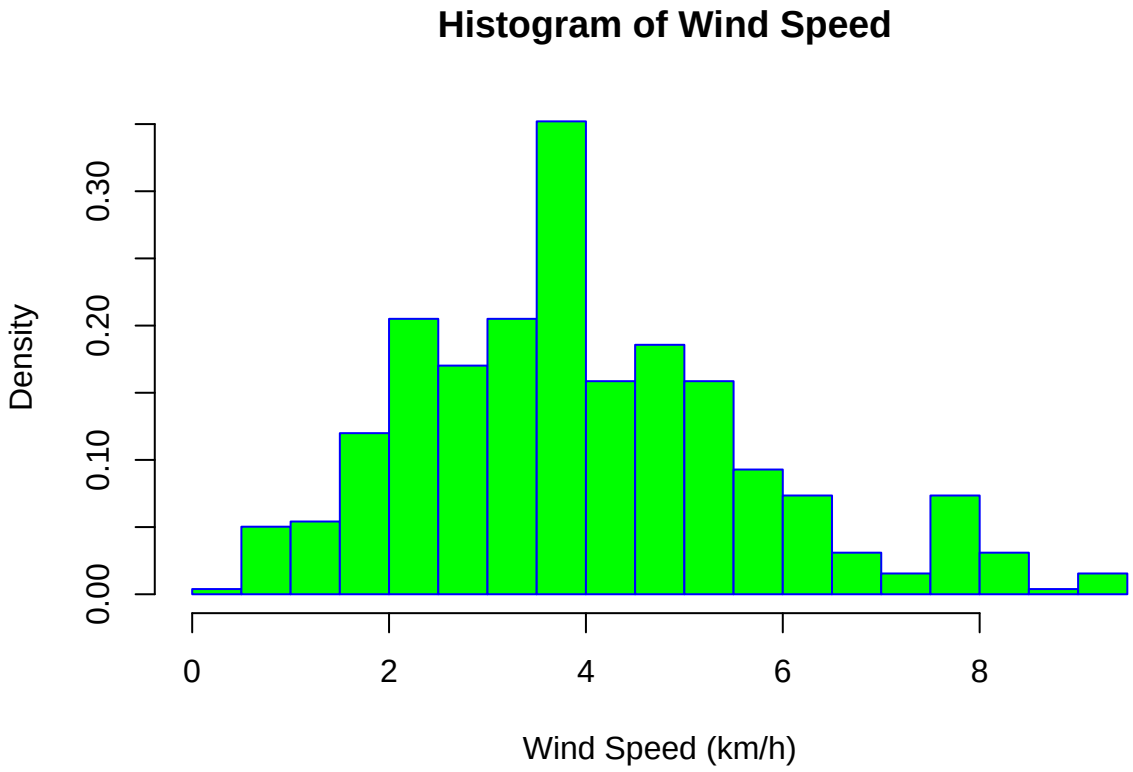
```
forestFires_data$month <-
  forestFires_data$month %>% factor(
    levels = c(
      "jan",
      "feb",
      "mar",
      "apr",
      "may",
      "jun",
      "jul",
      "aug",
      "sep",
      "oct",
      "nov",
      "dec"
    )
  )

temp <- ggplot(forestFires_data, aes(y = area)) +
  geom_point(aes(x = temp), color = "pink") + ggtitle("Area vs Temperature") +
  xlab("Temperature in Celsius degrees") +
  ylab("Area")
month <- ggplot(forestFires_data, aes(y = area)) +
```

```
  geom_point(aes(x = month), color = "green") + ggtitle("Area vs Month") +
  xlab("Month") +
  ylab("")
DC <- ggplot(forestFires_data, aes(y = area)) +
  geom_point(aes(x = DC), color = "blue") + ggtitle("Area vs DC") +
  xlab("DC index") +
  ylab("")
RH <- ggplot(forestFires_data, aes(y = area))+
  geom_point(aes(x = RH), color = "purple") + ggtitle("Area vs RH") +
  xlab("Relative Humidity in %") +
  ylab("")

grid.arrange(temp, month, DC, RH,ncol = 2)
```



Question b

```
hist(x = forestFires_data$wind, prob = T,
     main  = " Histogram of Wind Speed",
     breaks = 30,
     xlab = "Wind Speed (km/h)",
     col = "green",
     border = "blue")
```



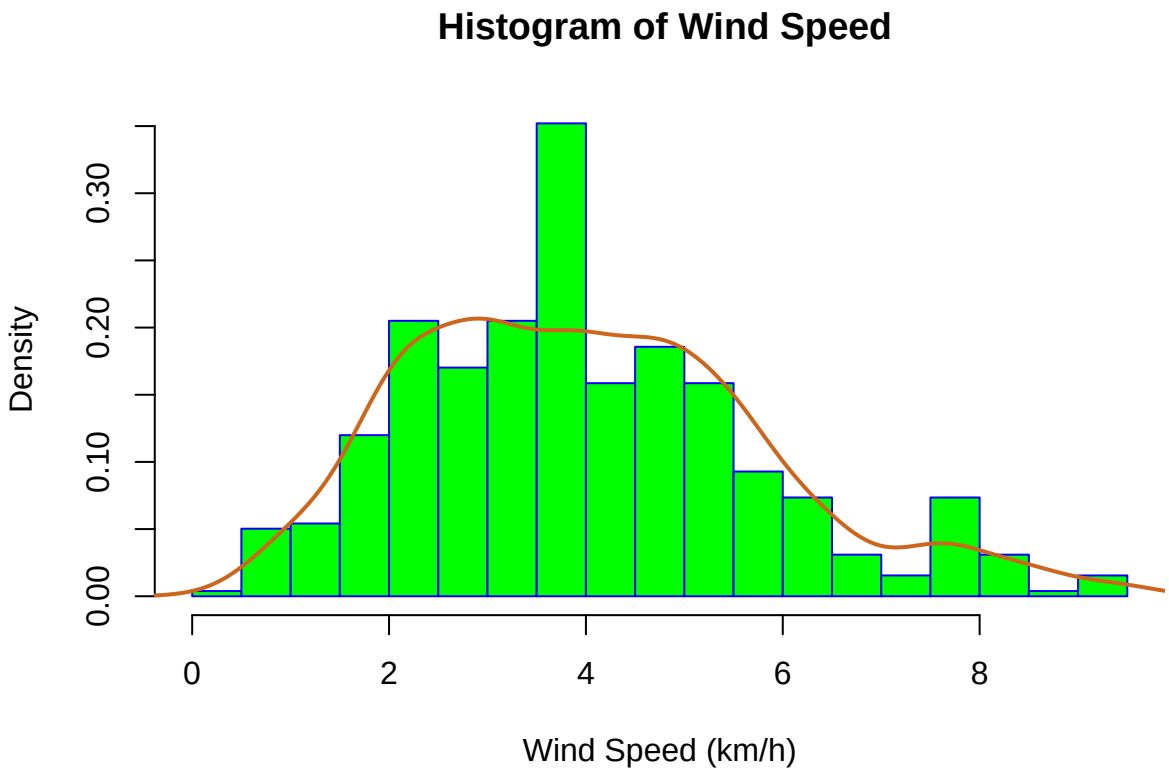
Question c

```
summary(forestFires_data$wind)
```

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

Question d

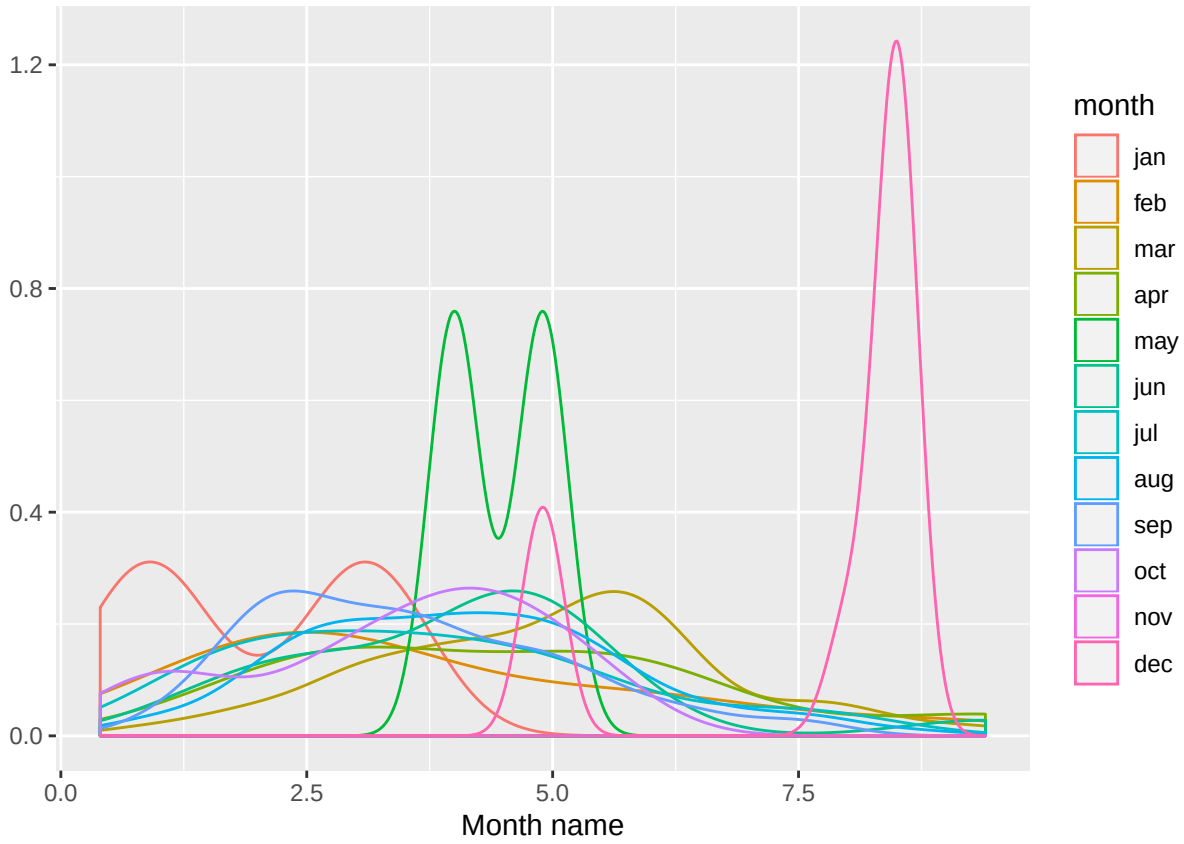
```
hist(x = forestFires_data$wind, prob = T,
     main  = " Histogram of Wind Speed",
     breaks = 30,
     xlab = "Wind Speed (km/h)",
     col = "green",
     border = "blue")
lines(density(forestFires_data$wind), col = "chocolate3", lwd = 2)
```



Question e

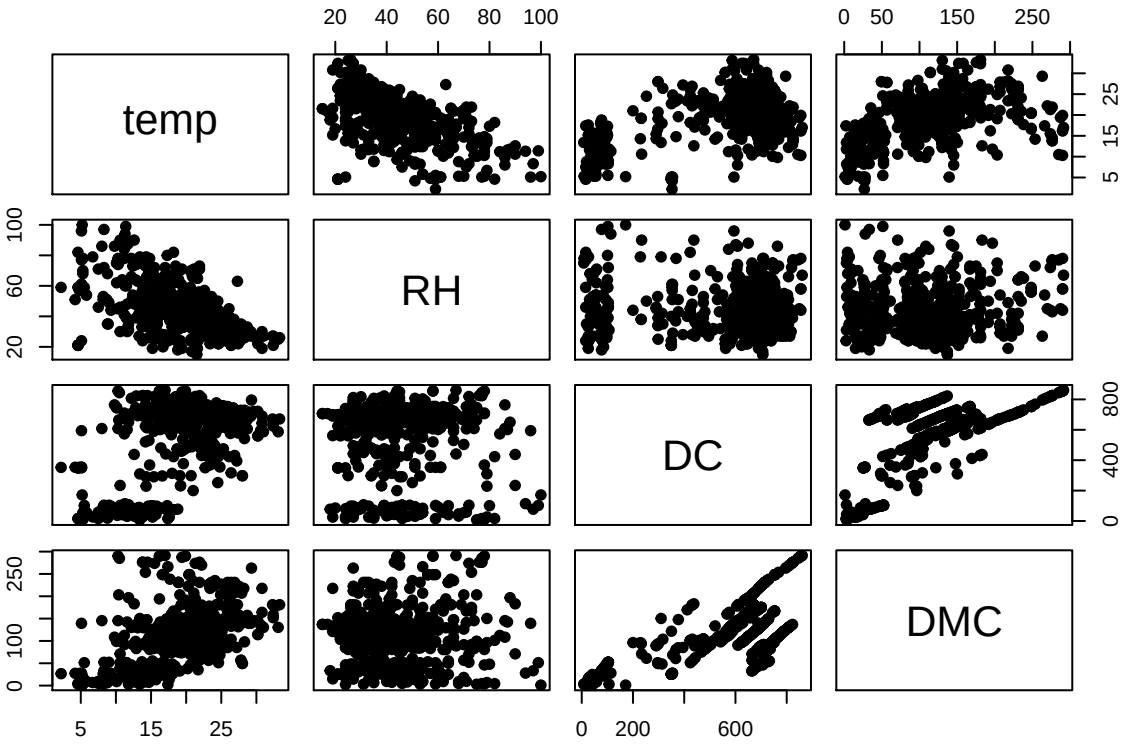
```
qplot(forestFires_data$wind, geom = "density", col = forestFires_data$month) + xlab("Month name") + labs
```

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



Question f

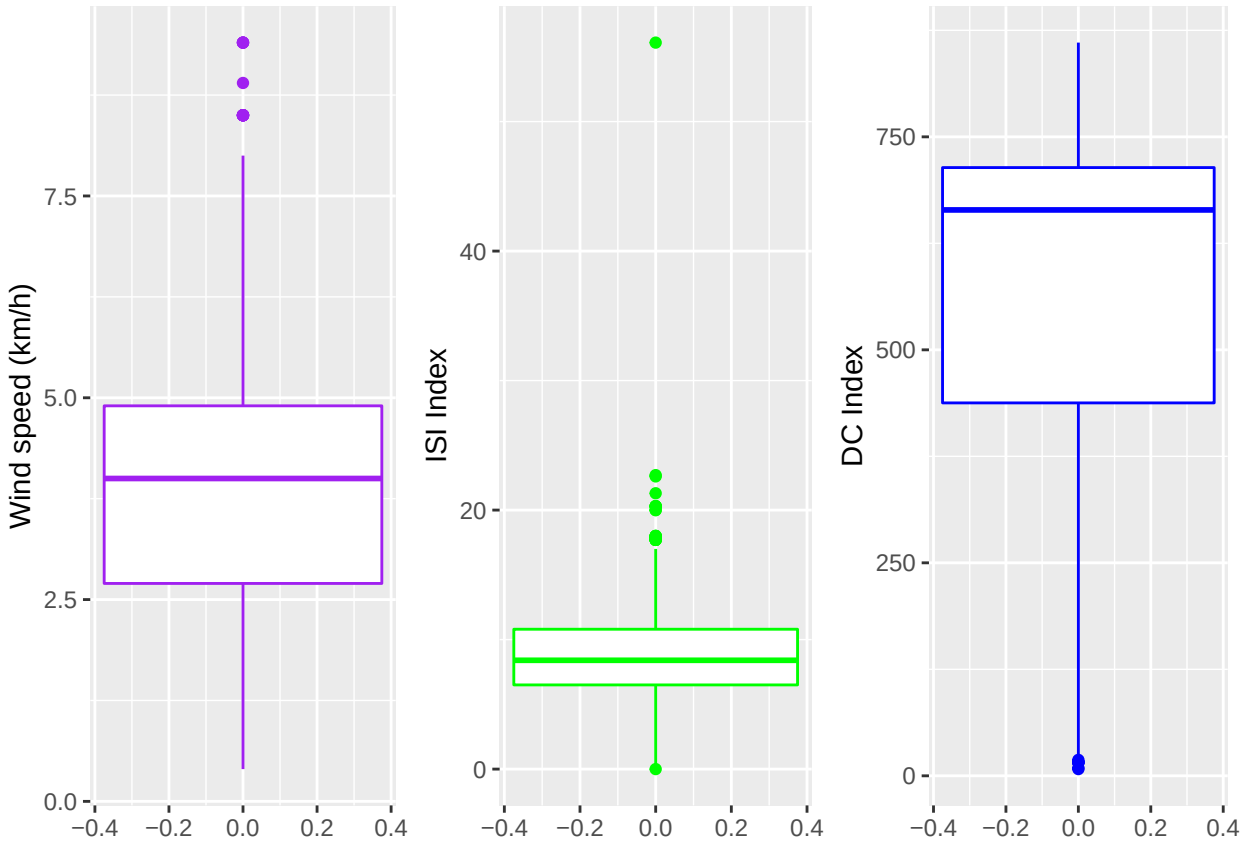
```
pairs(forestFires_data[,c("temp","RH","DC","DMC")], pch = 19)
```



From the above scatter plot matrix we can clearly see a linear relation between DMC vs DC indexes while higher temperature results in lower RH

Question g

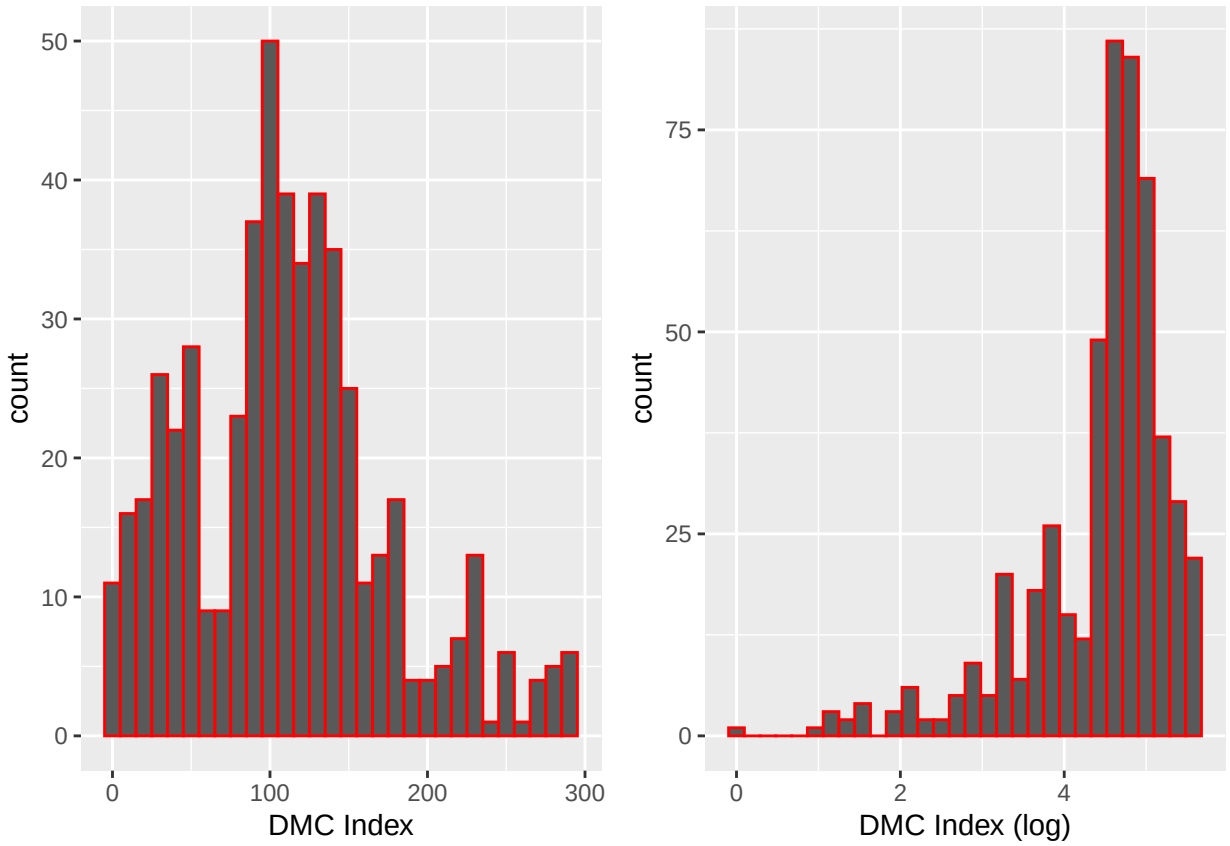
```
wind <- ggplot(forestFires_data)+
  geom_boxplot(aes(y = wind) , color = "purple") + ylab("Wind speed (km/h)")
ISI <- ggplot(forestFires_data)+
  geom_boxplot(aes(y = ISI), color = "green") + ylab("ISI Index")
DC <- ggplot(forestFires_data)+
  geom_boxplot(aes(y = DC), color = "blue") + ylab("DC Index")
grid.arrange(wind,ISI,DC, ncol = 3)
```



There are outliers for ISI which is denoted by the dots in the above boxplot, From the boxplot we can deduce that the median lies between 2.5 - 5 for wind, for ISI the median lies between 0-20 wiht the highest outlier above 40, fir DC the median lies between 500 - 750 with outliers near 0.

Question h

```
hist_wind <- ggplot(forestFires_data, aes(x = DMC))+
  geom_histogram(bins = 30, colour = "red") + xlab("DMC Index")
hist_wind_log <- ggplot(forestFires_data, aes(x = log(DMC)))+
  geom_histogram(bins = 30, colour = "red") + xlab("DMC Index (log)")
grid.arrange(hist_wind, hist_wind_log,ncol = 2)
```



Applying log on the wind prevents spread of the large values in the data and thus results in a much more skewed graph as seen after applying log(DMC)

Problem 2

Question a

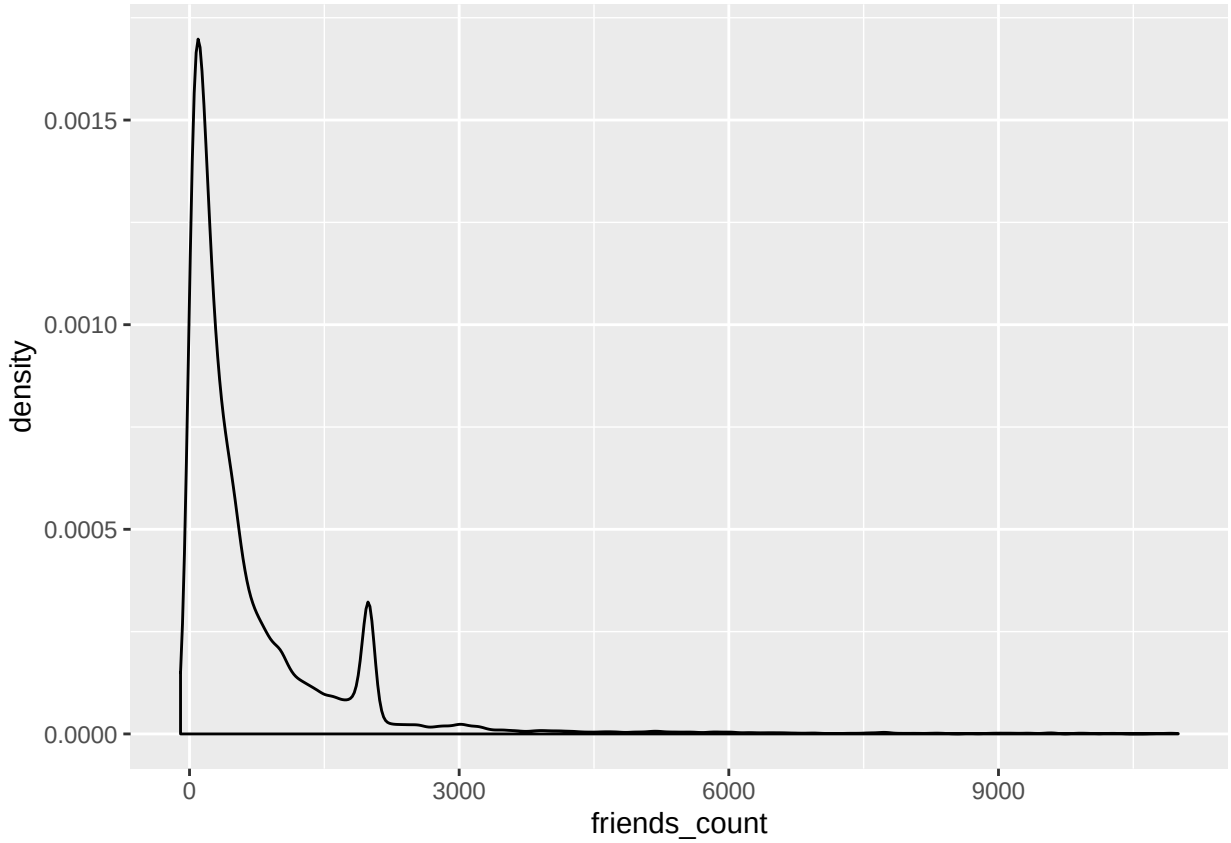
```
library(tidyverse)
library(ggplot2)
twitter_data <- read_csv("M01_quasi_twitter.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   screen_name = col_character(),
##   country = col_character(),
##   location = col_character(),
##   gender = col_character(),
##   race = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
ggplot(twitter_data) +
  geom_density(aes(x = friends_count)) +
  xlim(-100, 11000)
```

```
## Warning: Removed 226 rows containing non-finite values (stat_density).
```



Question b

```
summary(twitter_data$friends_count)
```

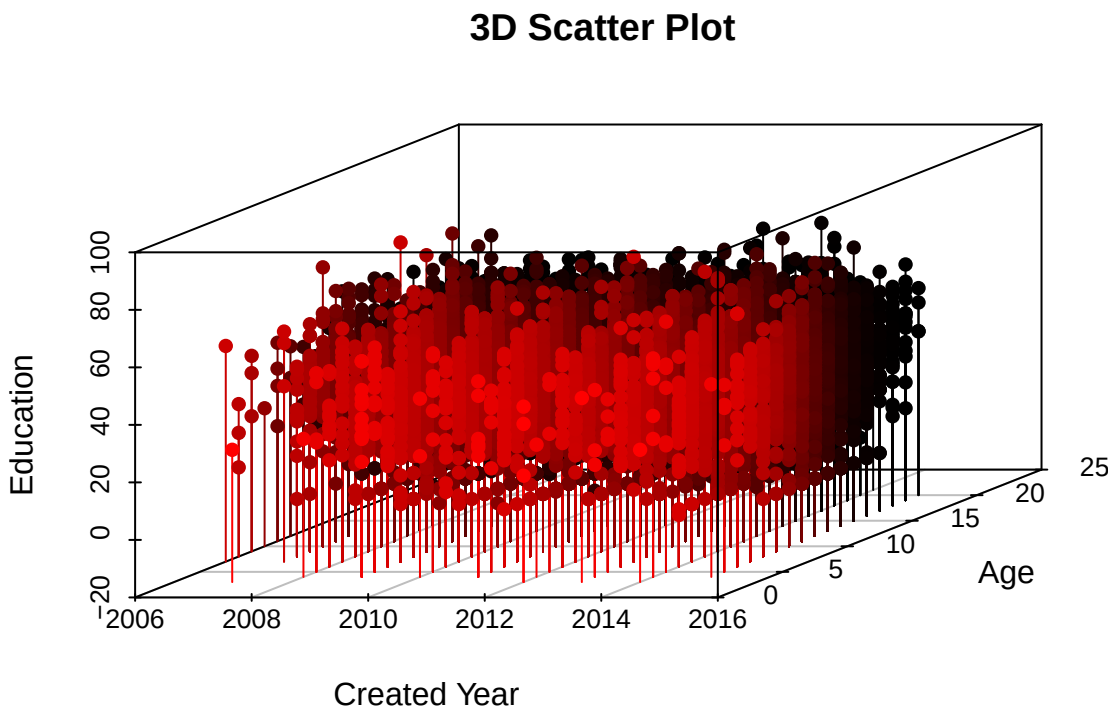
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     -84     123     324    1058     849  660549
```

Question c

The Data is pretty right skewed with one negative value of -84 and a median of 324. A large number of entries consists of 0 friends which explains the peak at 0 in the above graph. Outliers were removed to depict the distribution amongst the values.

Question d

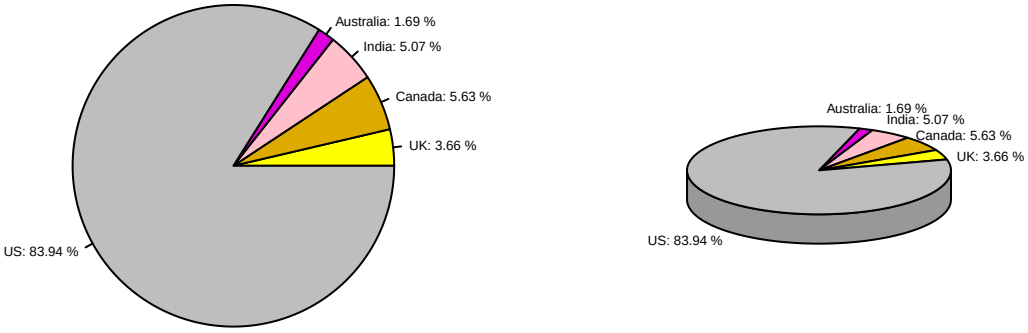
```
library("scatterplot3d")
scatterplot3d(twitter_data$created_at_year,twitter_data$education,twitter_data$age,
main="3D Scatter Plot",highlight.3d = TRUE,xlab='Created Year',ylab='',zlab='Education',type='h',pch=16)
dims<-par('usr')
x <- dims[1]+ 0.9*diff(dims[1:2])
y <- dims[3]+ 0.08*diff(dims[3:4])
text(x,y,expression('Age'),srt=0)
```



Question e

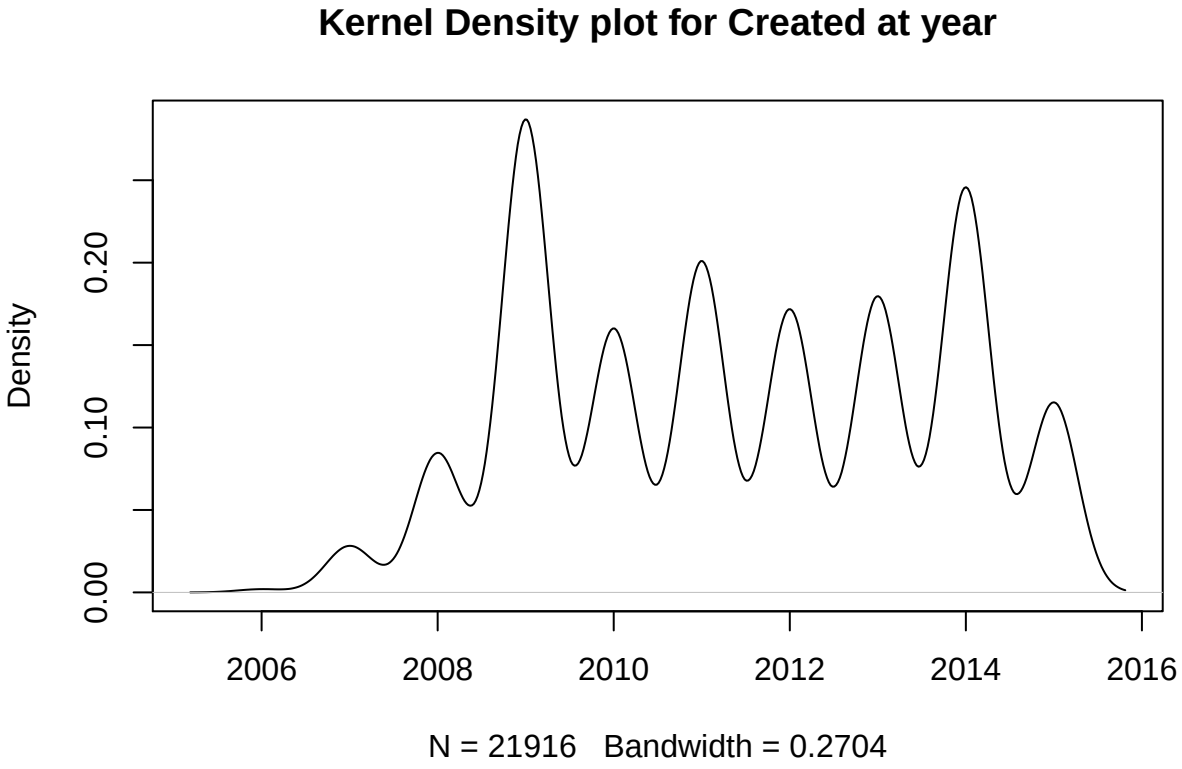
```
library(tidyverse)
library(plotrix)
tweeter_accounts <-
  data.frame(
    Country = c('UK', 'Canada', 'India', 'Australia', 'US'),
    number_of_accounts = c(650, 1000, 900, 300, 14900)
  )

tweeter_accounts <-
  tweeter_accounts %>% mutate(pie_percent  =
                                paste(round(number_of_accounts / sum(tweeter_accounts$number_of_accounts
  unite(label_names, Country, pie_percent, sep = ": ")
cols <- c("yellow","#ddaa00","pink","#dd00dd","grey")

par(mfrow = c(1,2))
pie(tweeter_accounts$number_of_accounts, labels = tweeter_accounts$label_names,
    col = cols, radius = 0.9, main = "Simple Pie Chart", cex = 0.4)
pie3D(tweeter_accounts$number_of_accounts,
      radius = 0.9,
      labels = tweeter_accounts$label_names,
      col=cols,
      #shade = 0.6,
      labelcex = 0.4,
      start = 0.25
      )
```

Simple Pie Chart



Question f

```
density_plot <- density(twitter_data$created_at_year)
plot(density_plot, main = 'Kernel Density plot for Created at year')
```



From the density plot we can deduce that most of the twitter accounts were created after 2006 and it was which is correctly explained as twitter was invented in 2006, it peaked between 2008-2010 and 2014. It decreased gradually after 2014 which is clearly seen in the above graph.

Problem 3

```
library(tidyverse)
library(scatterplot3d)
library(plotrix)
```

```
##### Reading the data

pc<-read_csv('raw_data.csv')
```

```
## Parsed with column specification:
## cols(
##   A = col_double(),
##   B = col_double(),
##   C = col_double(),
##   D = col_double()
## )
```

```
head(pc)
```

```
## # A tibble: 6 x 4
##       A      B     C     D
##   <dbl>  <dbl> <dbl> <dbl>
## 1  8.26 -0.656     6     8
## 2 10.6  -0.716     7     8
## 3  8.74  0.800     7     5
## 4  6.56  1.58      6    10
## 5  9.36  1.03      7     8
## 6  9.02  0.720     7    12
```

Question a

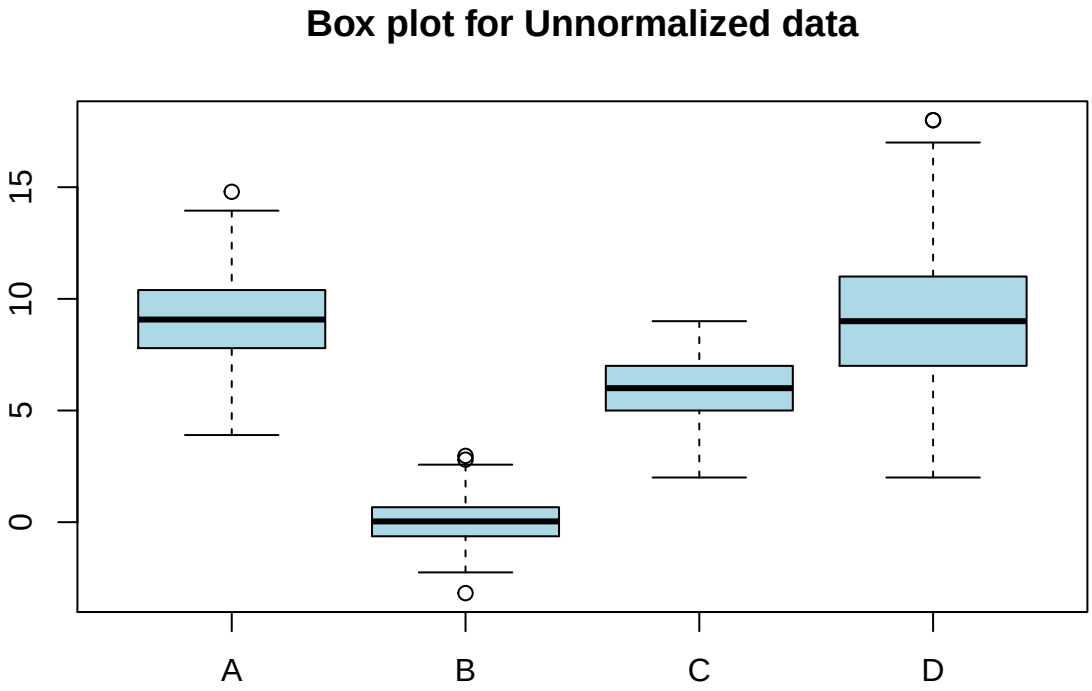
```
##### Using min-max Normalization method

normalize<-function(x){
  return ((x-min(x))/(max(x)-min(x)))
}
```

```
##### Normalizing data

Ndata<-normalize(pc)
Ndata<-as.data.frame(Ndata)
head(Ndata)
```

```
##           A         B         C         D
## 1 0.5399150 0.1190059 0.4333251 0.5277709
## 2 0.6485377 0.1161841 0.4805480 0.5277709
## 3 0.5629147 0.1877476 0.4805480 0.3861022
## 4 0.4595352 0.2247518 0.4333251 0.6222167
## 5 0.5920942 0.1984951 0.4805480 0.5277709
## 6 0.5759700 0.1839746 0.4805480 0.7166625
```

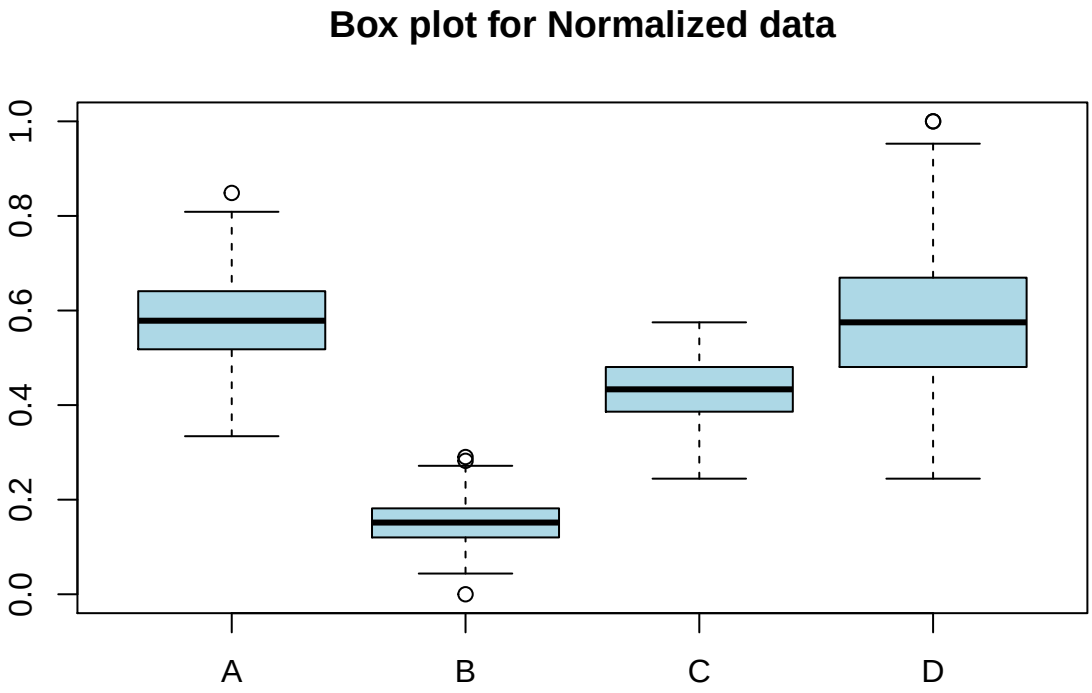
Question b.

```
#### Box plot for unnormalized data
boxplot(pc[,],main='Box plot for Unnormalized data',col='light blue')
```



Question c

```
##### Boxplot for Normalized data

boxplot(Ndata[,],main='Box plot for Normalized data',col='light blue')
```



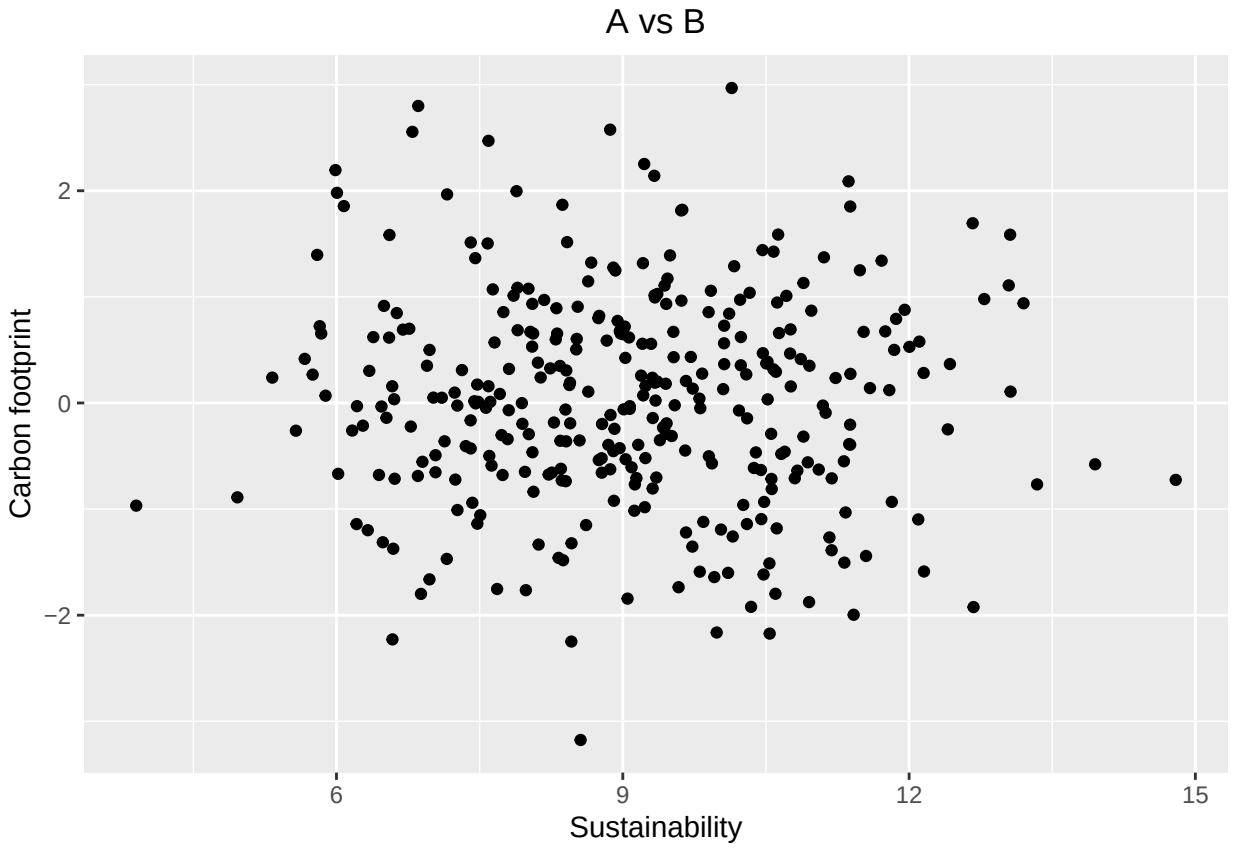
Question d

Interpreting the results

The difference between two box plot is the range of data points, the normalized data ranges between 0 and 1 for all the variables,while the unnormalized data has different range of data points for all variables. Normalization helps in better visualization of datas and makes optimization algorithm to run faster.

Question e.

```
##### Plotting scatter plot

  ggplot(pc,aes(x=`A`,y = `B`))+geom_point()+xlab('Sustainability')+ylab('Carbon footprint')+ggtitle('A
```



Interpretting the results:

The variables Sustainability and Carbon Footprints(A and B) have no association. The data points are spreaded randomly in the plot indicating there is no correlation between variables.