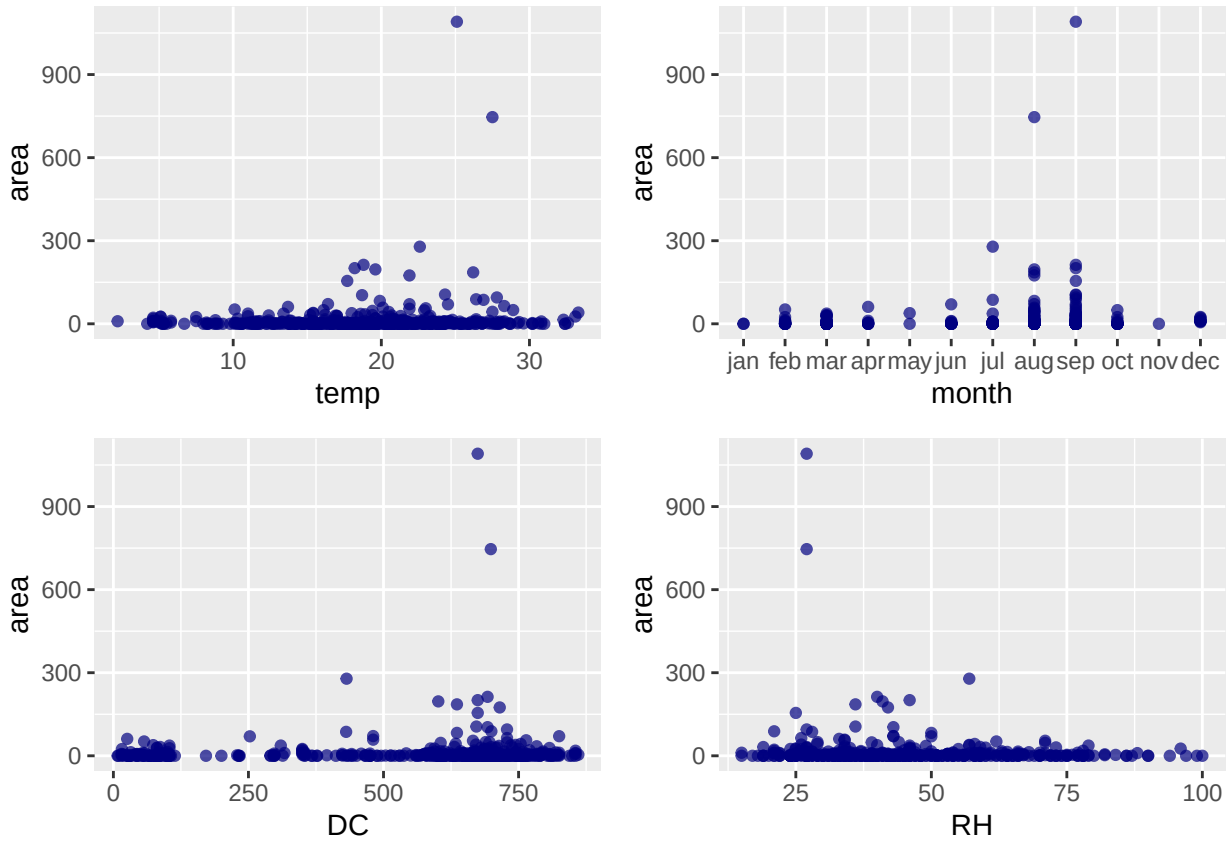
Group10

Chenxuan He, Zejin Kong

5/12/2020

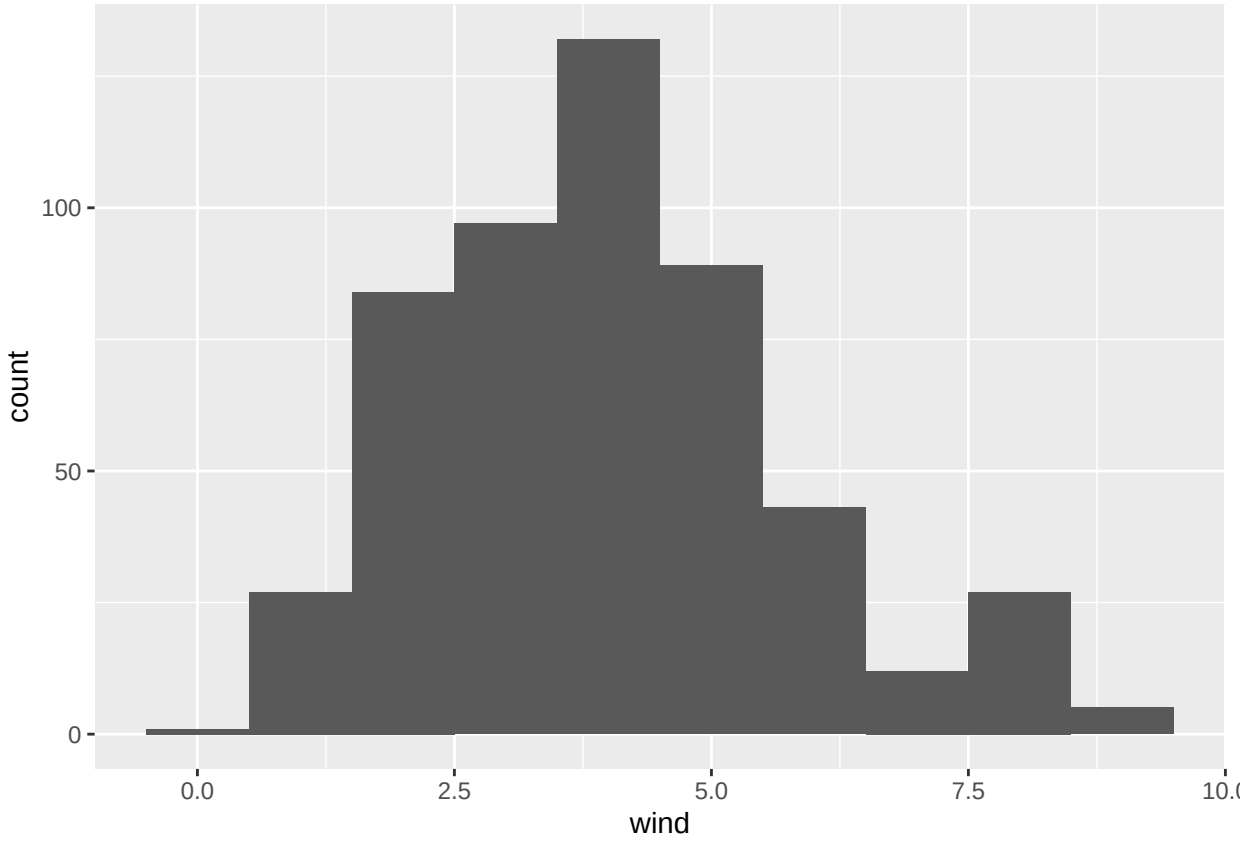
1.a Plot area vs.temp, area vs. month, area vs. DC, area vs. RH for January- through December combined in one graph. Hint: Place area on Y axis and use 2x2 matrix to place the plots adjacent to each other.

```
forestfires <- read.csv("C:/Users/hcx07/Desktop/7275/HW1/forestfires.csv")
#forestfires <- read.csv("forestfires.csv")
library(ggplot2)
p1<-ggplot(forestfires)+geom_point(aes(x=temp, y= area), color = "navy",alpha = 0.7)
forestfires$month <-factor(forestfires$month,levels =
c("jan", "feb", "mar", "apr", "may", "jun","jul", "aug", "sep", "oct", "nov", "dec"))
p2<-ggplot(forestfires,aes(month,area))+geom_point( color = "navy",alpha = 0.7)
p3<-ggplot(forestfires)+geom_point(aes(x=DC, y= area), color = "navy",alpha = 0.7)
p4<-ggplot(forestfires)+geom_point(aes(x=RH, y= area), color = "navy",alpha = 0.7)
library("gridExtra")
grid.arrange(p1,p2,p3,p4,ncol = 2, nrow = 2)
```



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

```
ggplot(forestfires,aes(wind,))+geom_histogram(binwidth = 1)
```



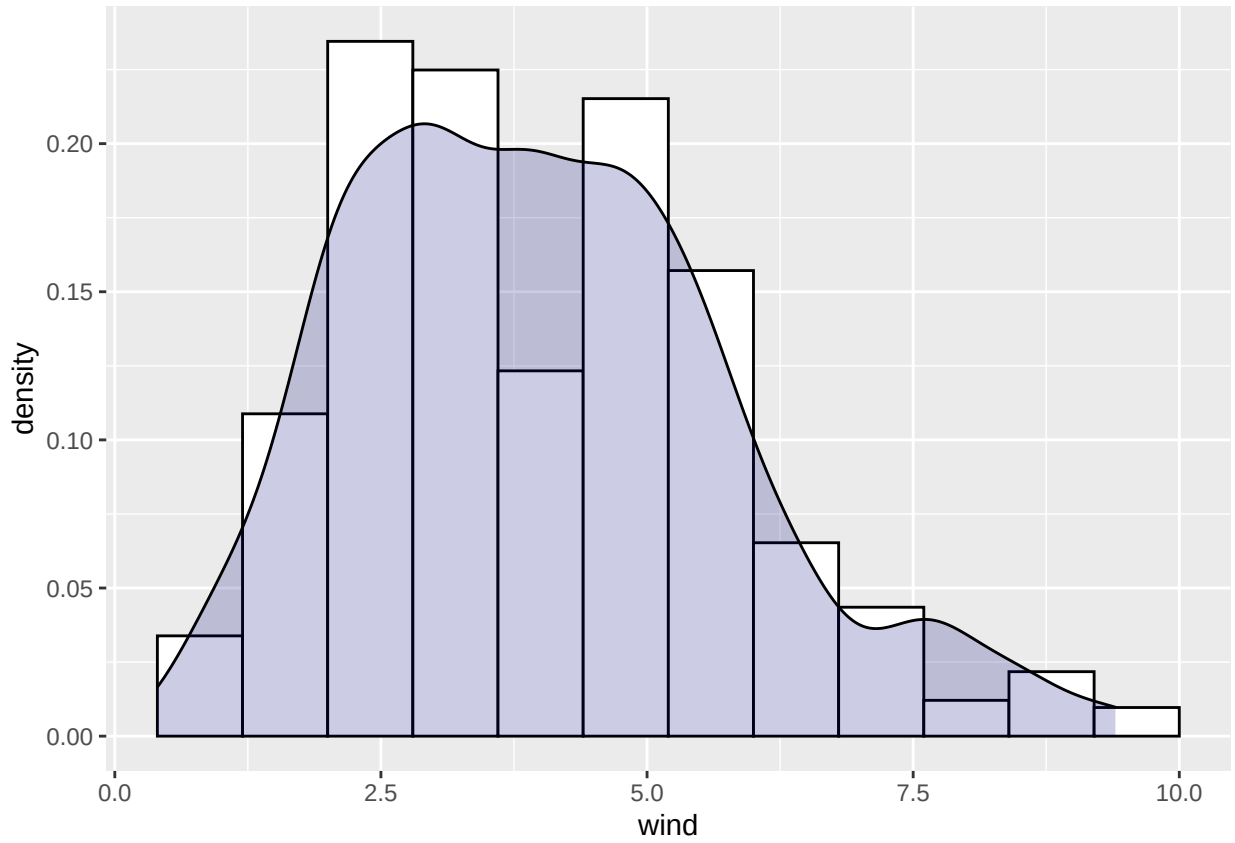
1.c Compute the summery statistics (min, 1Q, mean, median, 3Q, max,) of part b.

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

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

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

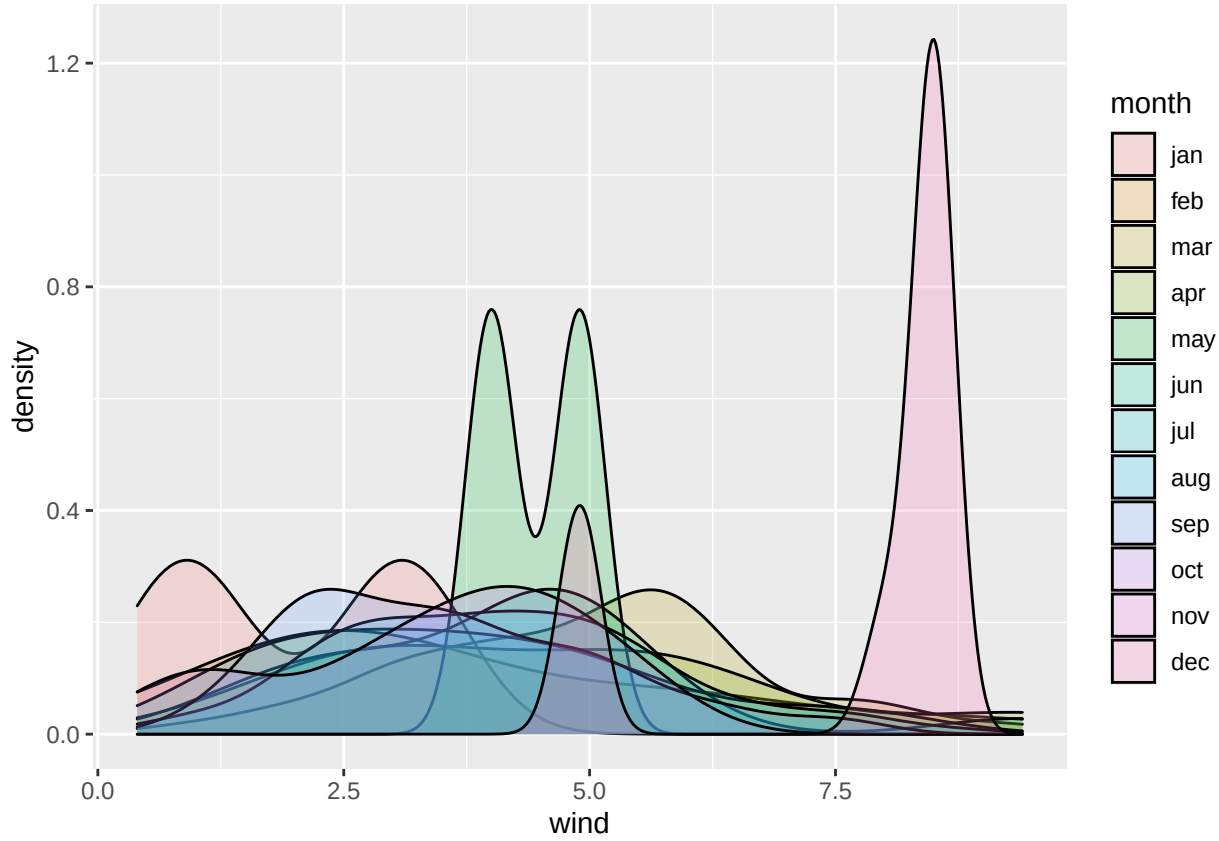
```
ggplot(forestfires,aes(wind,))+ geom_histogram(aes(y=..density..), binwidth = .8, colour="black", fill=
```



#1.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.

```
ggplot(forestfires,aes(x=wind,group=month,fill=month))+ geom_density(alpha=.2)
```

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

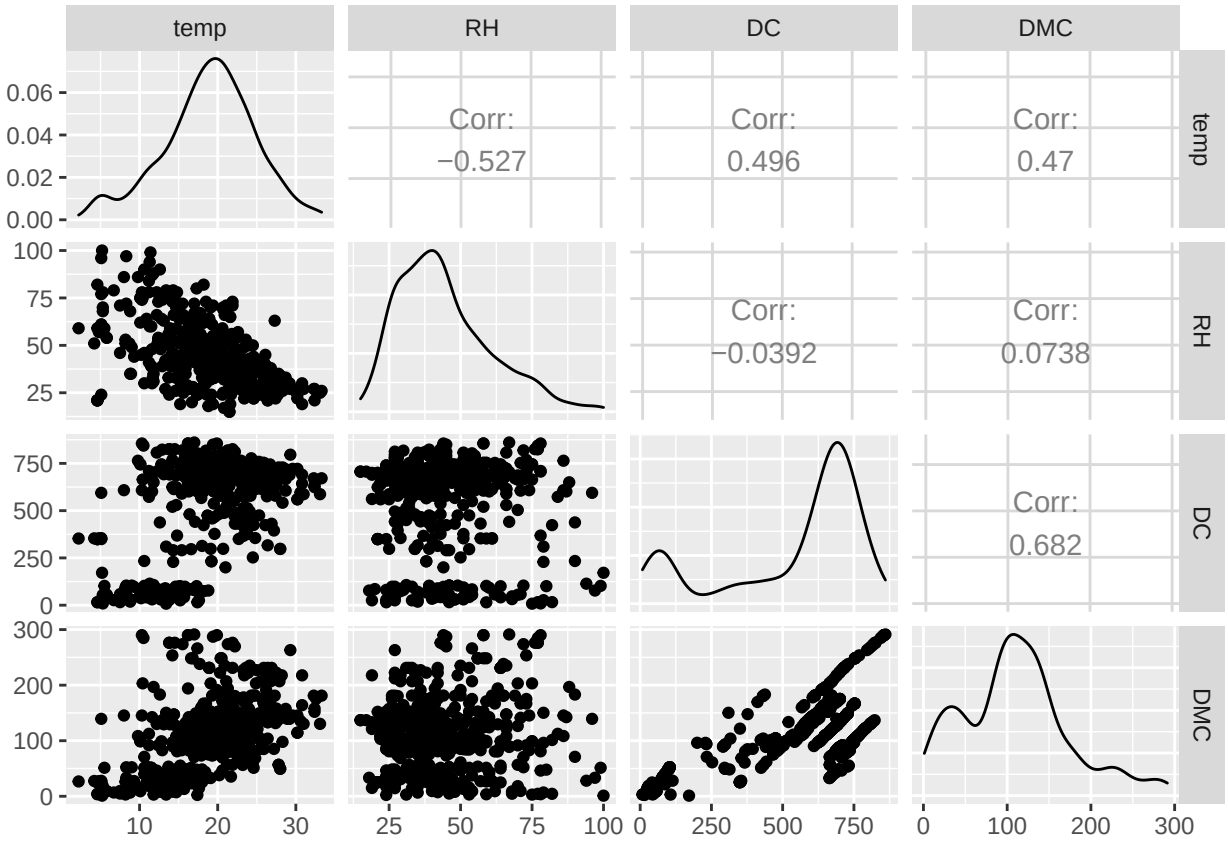


1.f Plot the scatter matrix for temp, RH, DC and DMC. How would you interpret the result in terms of correlation among these data?

```
trdd<- forestfires[c("temp","RH","DC","DMC")]
library("GGally")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(trdd)
```

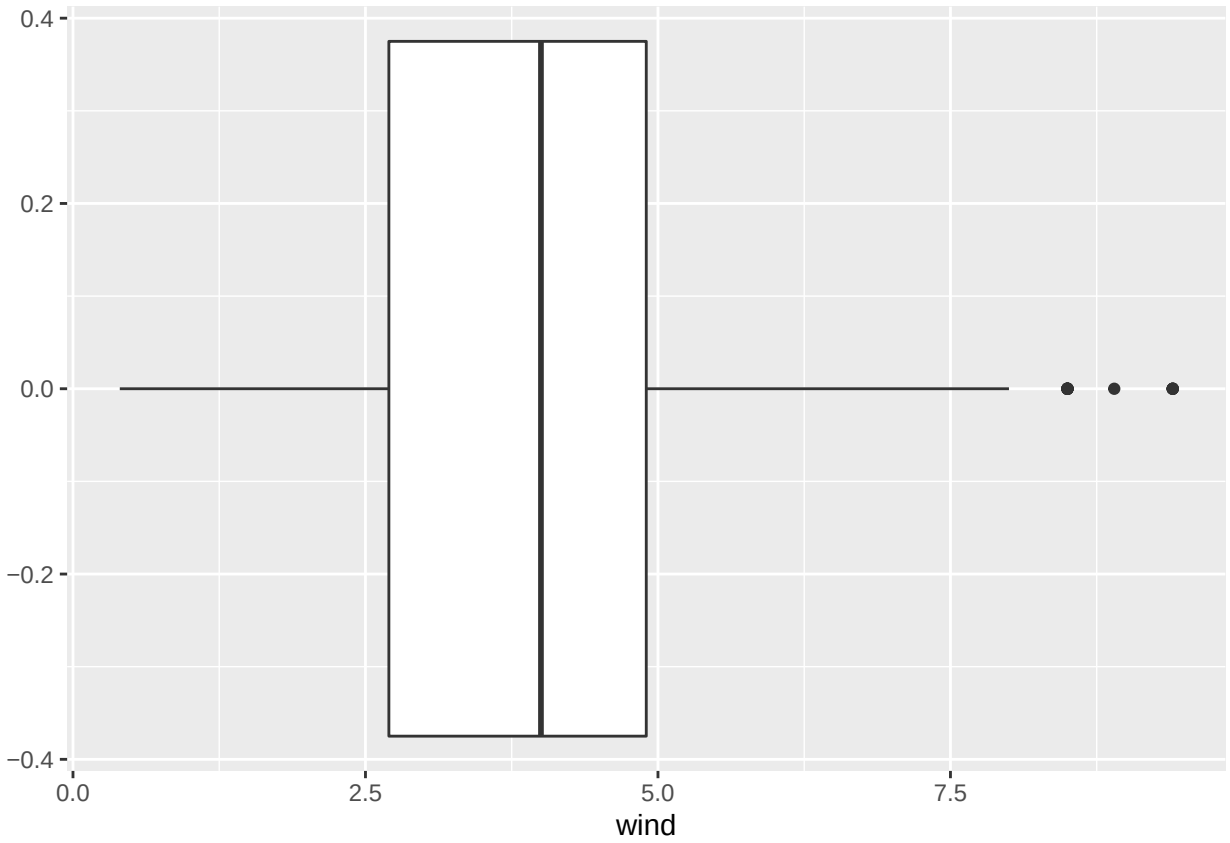


temp has a low negative correlation wtih RH, and a low positive correlation with DC and DMC. There is no correlation betwen RH and DC or RH and DMC.
DC has a low positive correlation with DMC

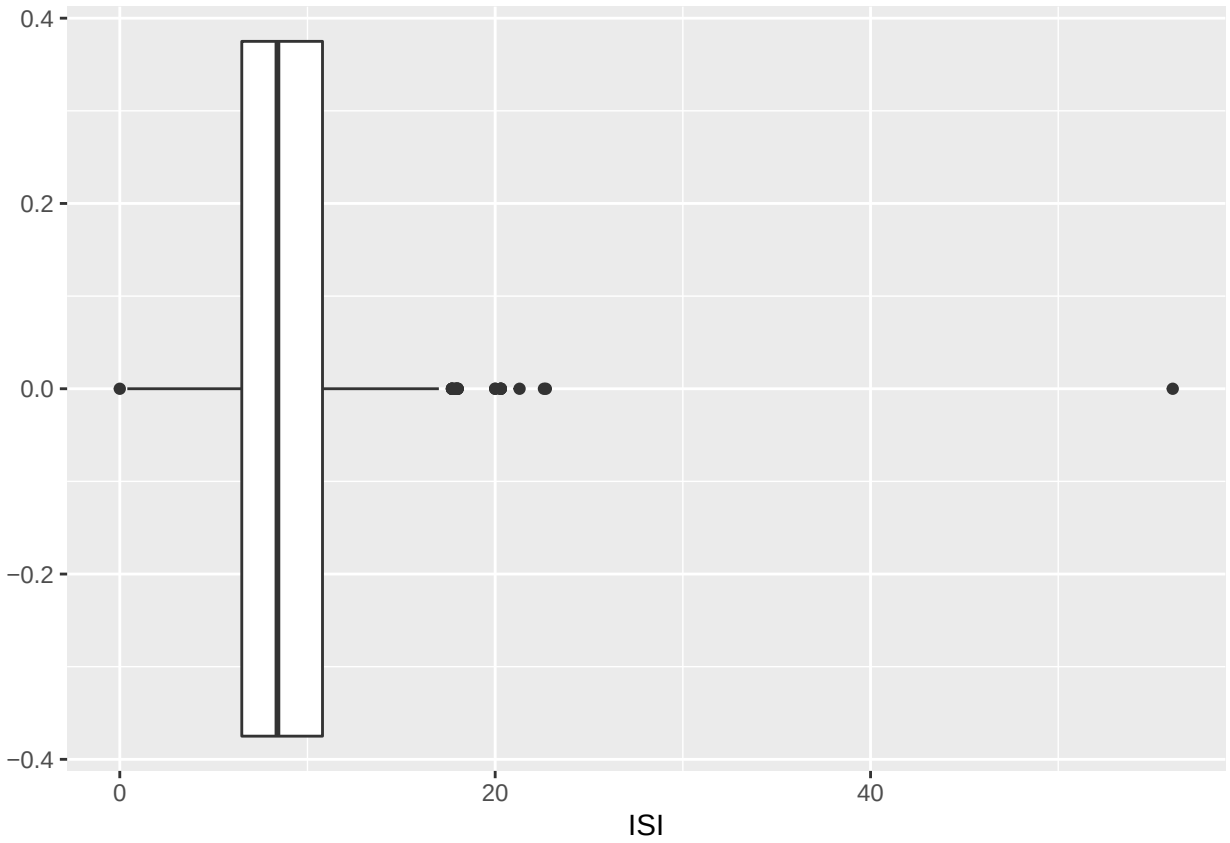
1.g Create boxplot for wind, ISI and DC. Are there any anomalies/outliers? Interpret

your result.

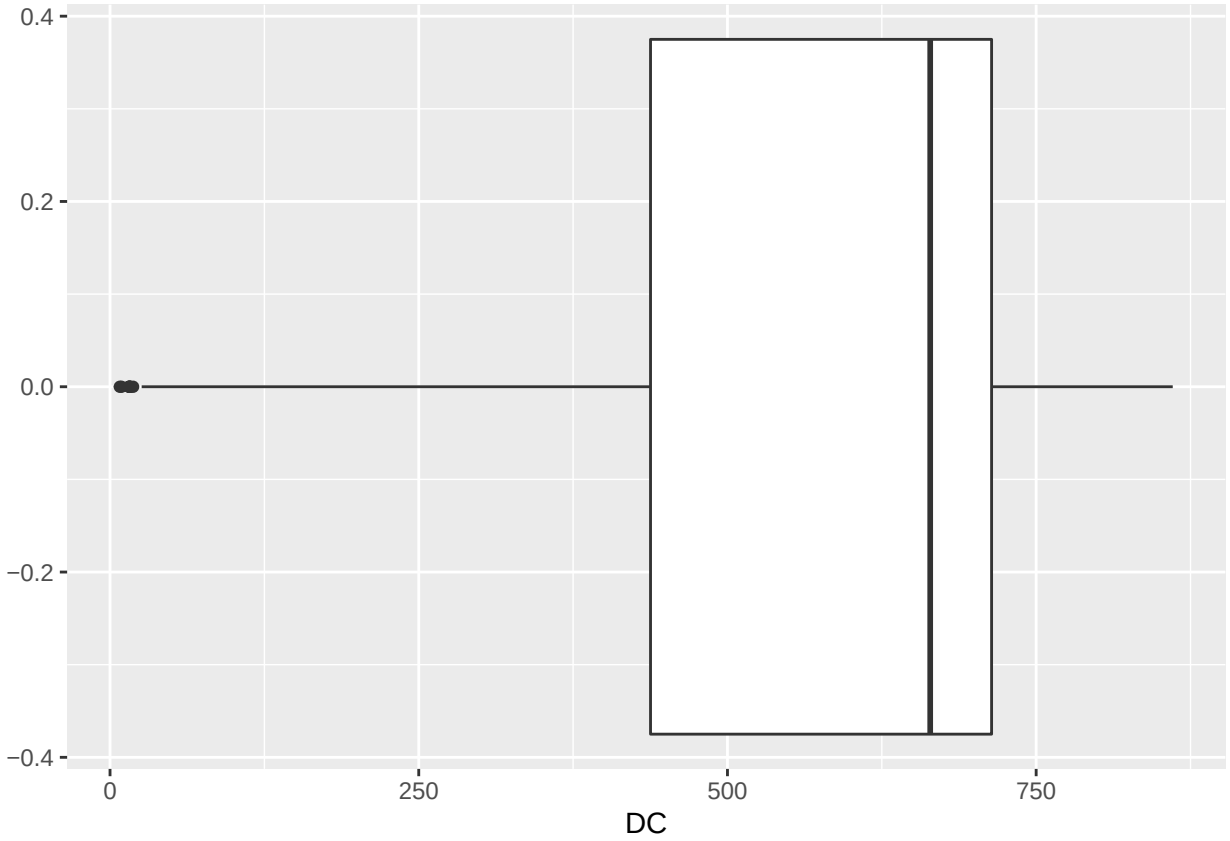
```
ggplot(forestfires,aes(x=wind))+geom_boxplot()
```




```
ggplot(forestfires,aes(x=ISI))+geom_boxplot()
```




```
ggplot(forestfires,aes(x=DC))+geom_boxplot()
```

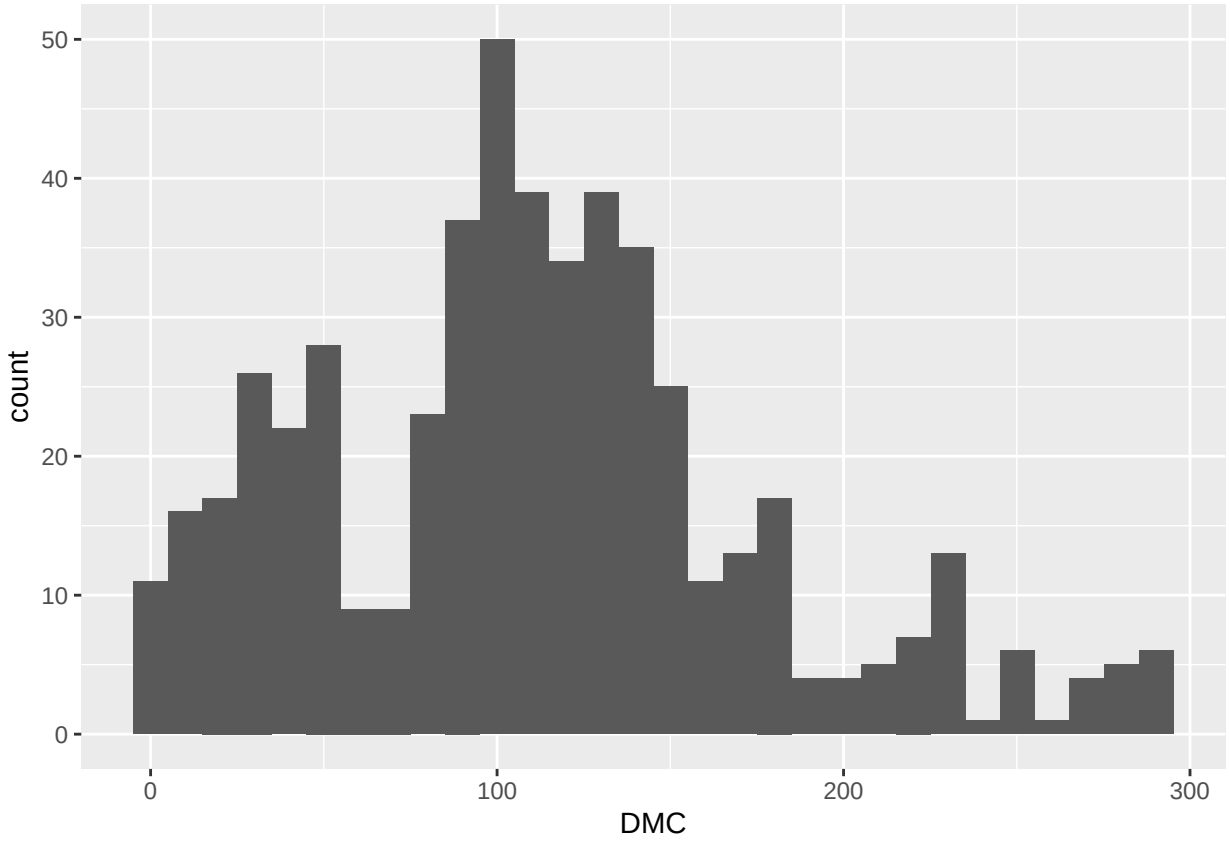


There are outliers in all three boxplots

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

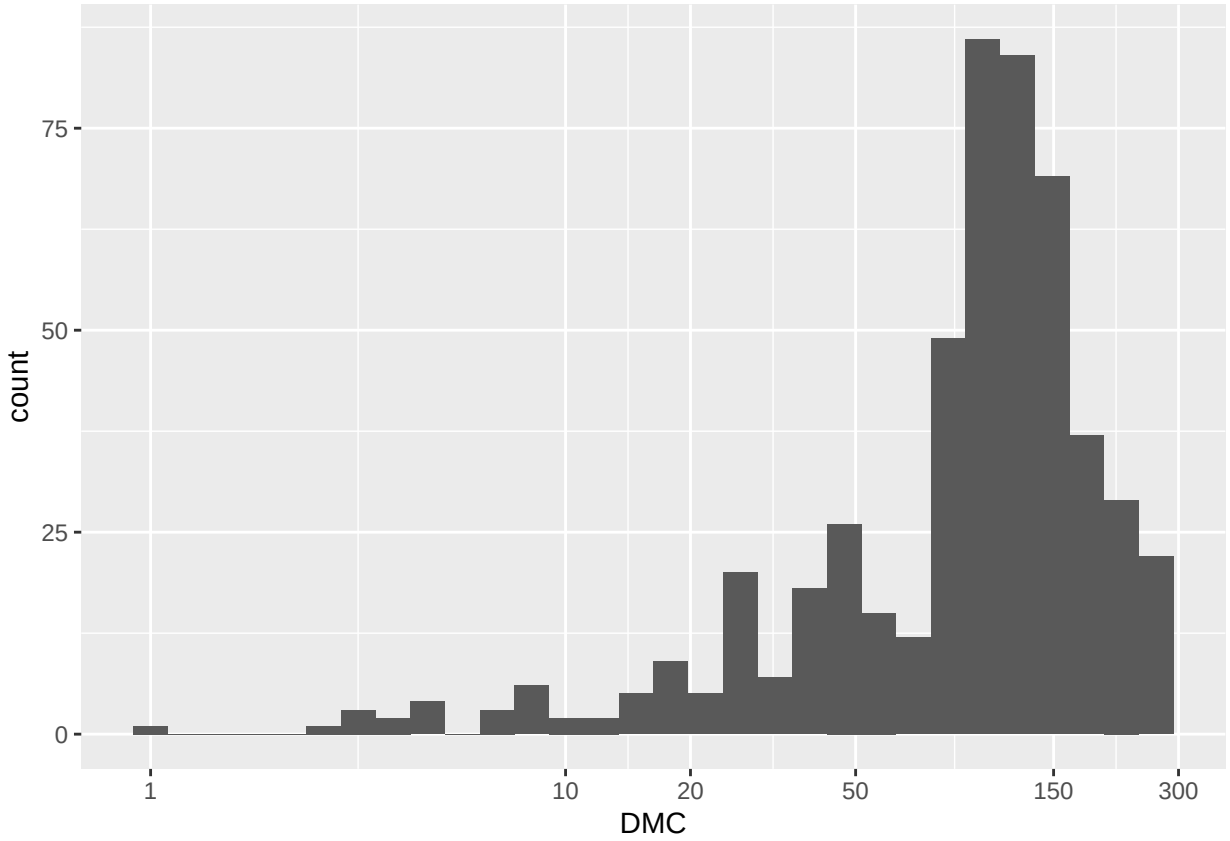
```
ggplot(forestfires, aes(x=DMC)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(forestfires, aes(x=DMC)) + geom_histogram() + scale_x_log10(breaks=c(1,10,20,50,150,300))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

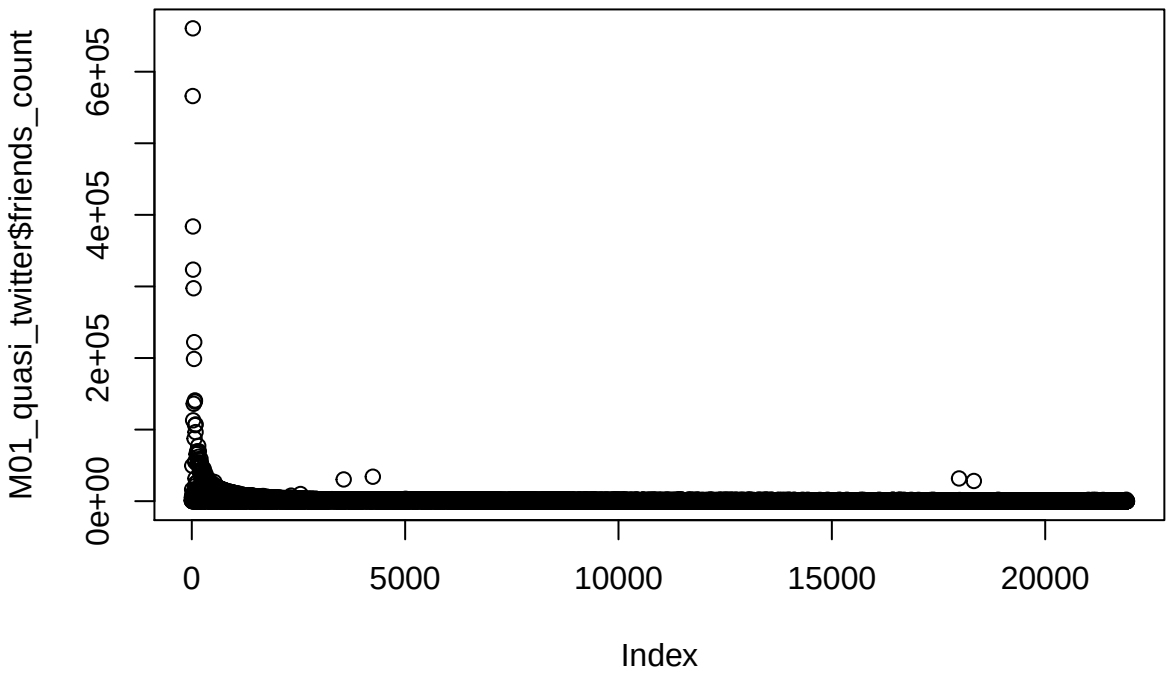


Taking logs resulting in the plot having a distorted shape rather than twin peaks which it was.

2.a. How are the data distributed for friend_count variable?

```
M01_quasi_twitter <- read.csv("C:/Users/hcx07/Desktop/7275/HW1/M01_quasi_twitter.csv")


plot(x=M01_quasi_twitter$friends_count)
```

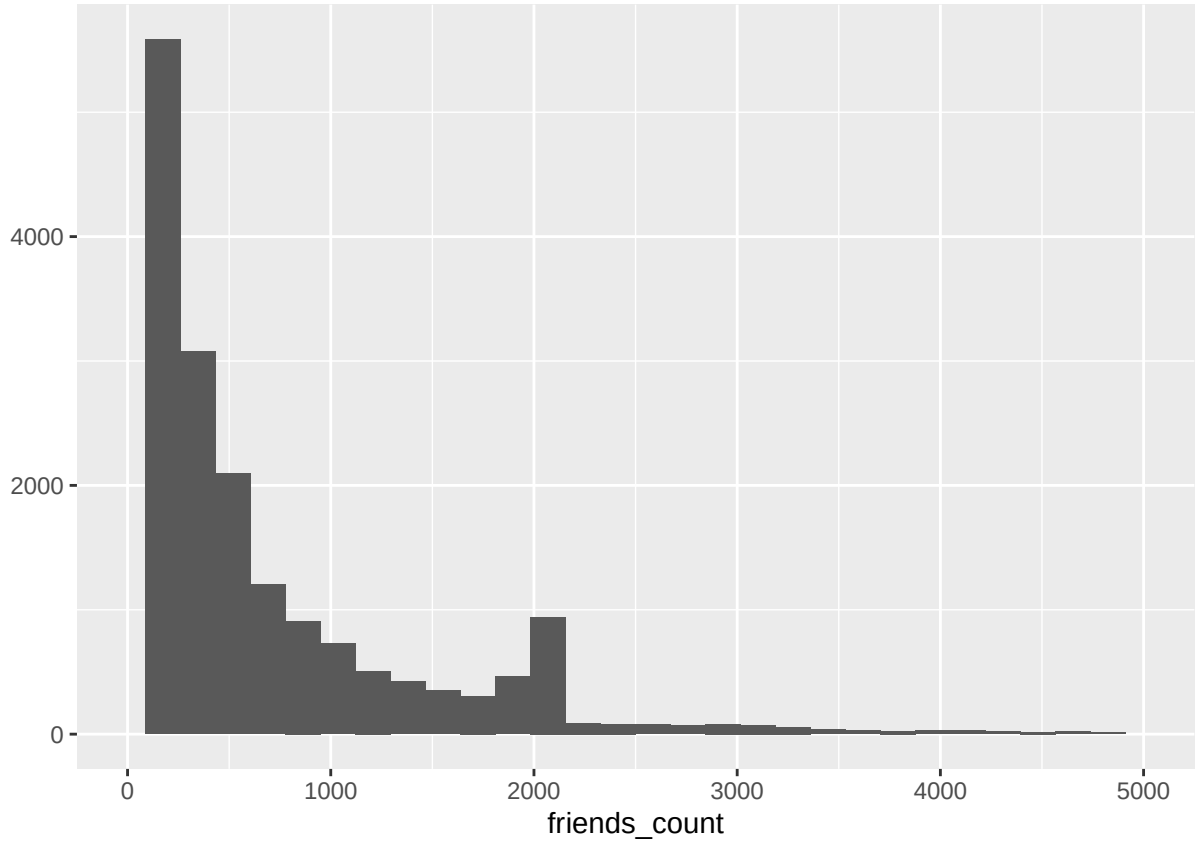



```
qplot(data = M01_quasi_twitter, x = friends_count) +  xlim(c(1, 5000))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 733 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Since the outliers made the range of whole sample too wide, we can not tell the distribution directly form the histogram. Then we cut out the outliers to see the further changes.

After removing 733 outliers, the new dataset is close to lognormal distribution.

2.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
```

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

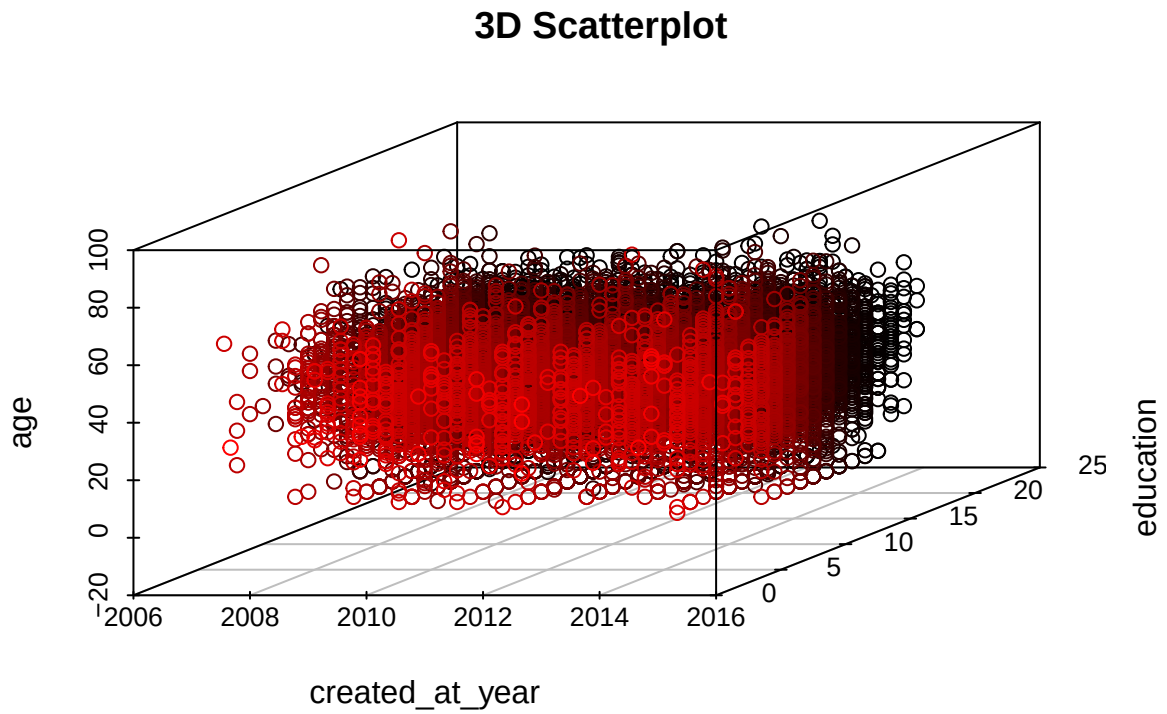
```
sd(M01_quasi_twitter$friends_count)
```

```
## [1] 8125.054
```

The friends_count shouold be a non-negetive interger, since the the min is -84, this should be an error. and this set of data with high std.

2.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)
scatterplot3d(M01_quasi_twitter[c("created_at_year","education","age")],
  highlight.3d=TRUE,
  main= "3D Scatterplot")
```

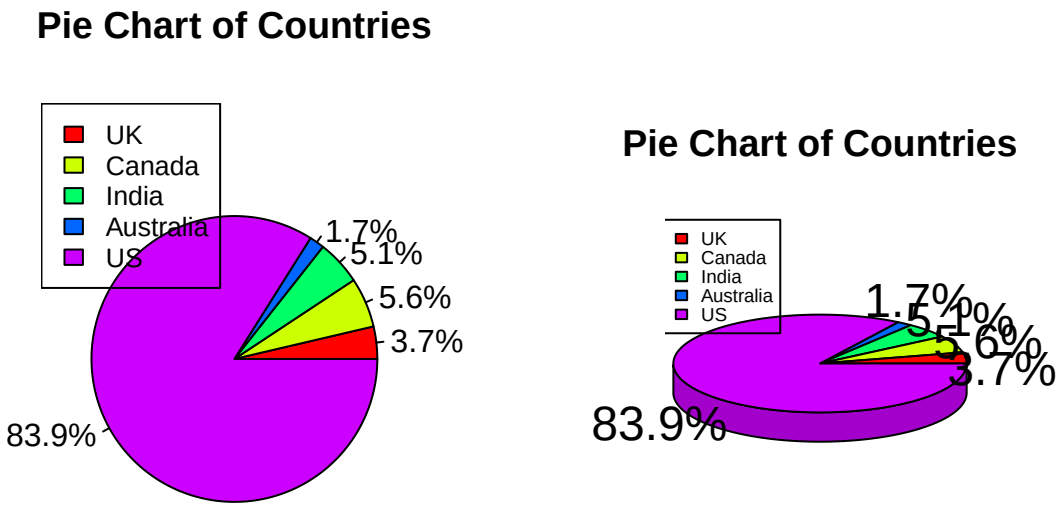


2.e. Consider 650, 1000, 900, 300 and 14900 tweeter accounts are in UK, Canada, India, Australia and US, respectively. Plot the percentage Pie chart includes percentage amount and country name adjacent to it, and also plot 3D pie chart for those countries along with the percentage pie chart. Hint: Use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

```
par(mfrow=c(1,2))
slices <- c(650,1000,900,300,14900)
countries <- c( "UK", "Canada", "India", "Australia","US")
piepercent<- round(100*slices/sum(slices),1)
pielabels<- paste(piepercent, "%", sep="")
```

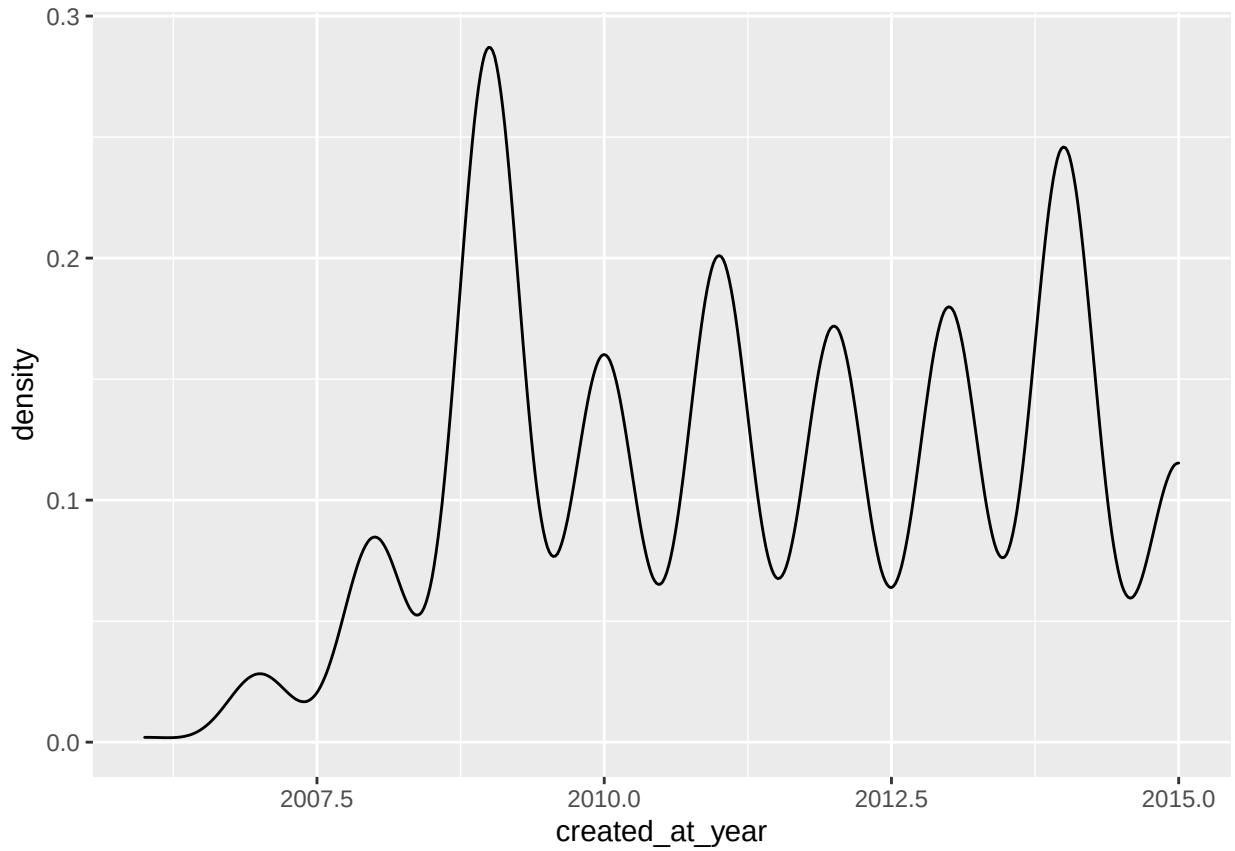
```
pie(slices, labels = pielabels, main="Pie Chart of Countries",col=rainbow(5))
legend("topleft", countries,cex=0.8,fill =rainbow(5))

library(plotrix)
pie3D(slices, labels = pielabels, main="Pie Chart of Countries")
legend("topleft", countries,cex=0.6,fill=rainbow(5))
```



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

```
ggplot(M01_quasi_twitter,aes(x=created_at_year))+geom_density()
```



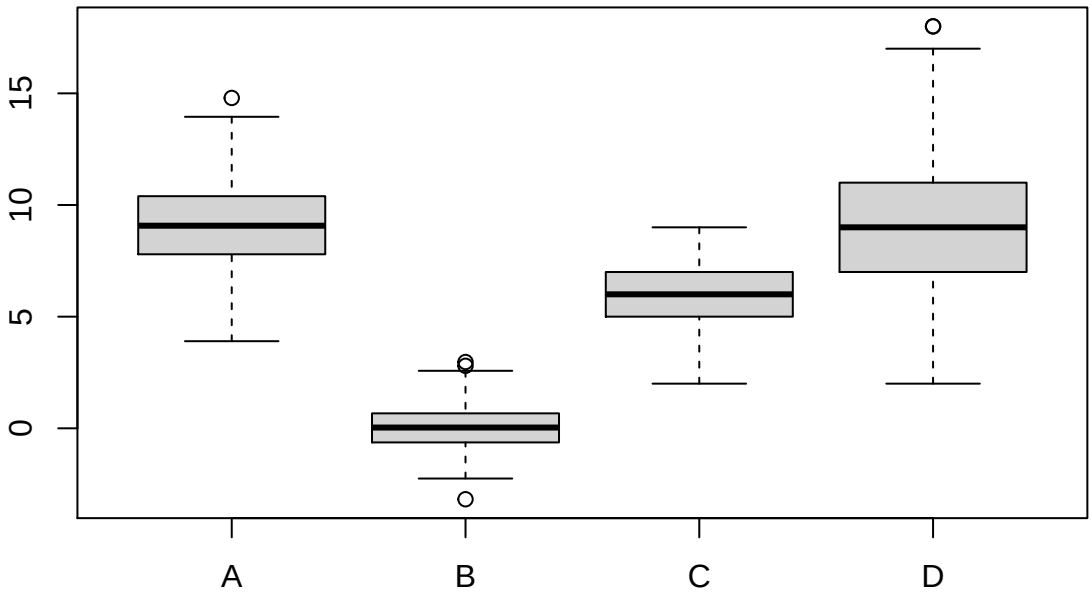

```
# The dataset is impacted seasonally because the peak comes every fixed period.
```

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

```
raw_data <- read.csv("C:/Users/hcx07/Desktop/7275/HW1/raw_data.csv")
Ndata <-scale(raw_data)
```

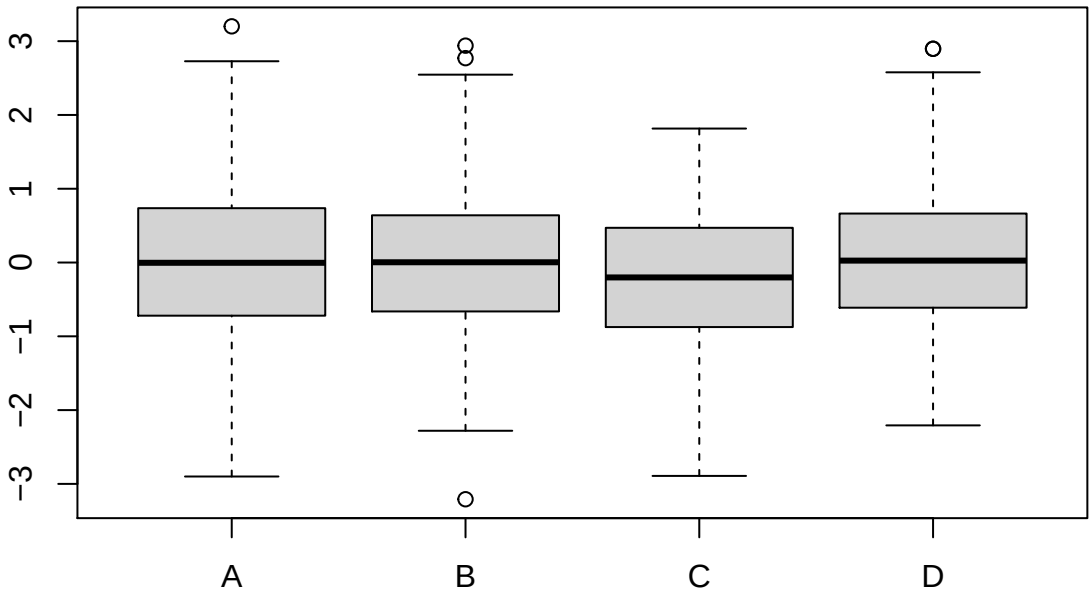
3.b. Create the boxplot of all the variables in their original form.

```
boxplot(raw_data)
```



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

```
boxplot(Ndata)
```

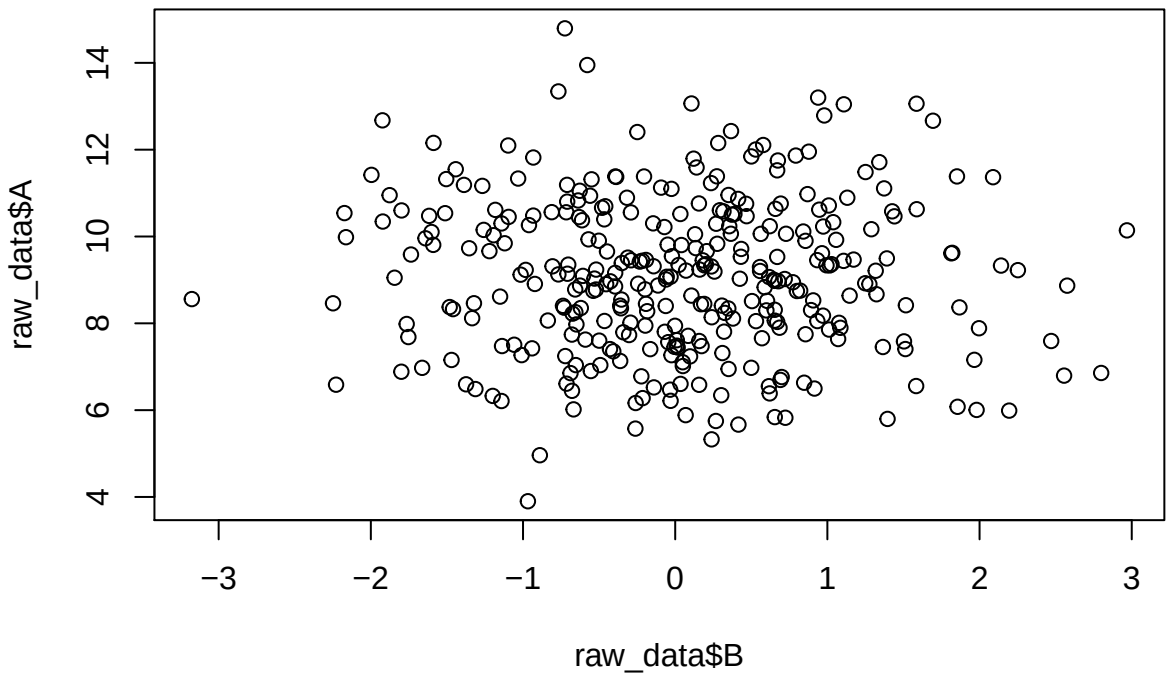


3.d. Compare the result of part b and part c; interpret your answer.

After standarization, we could find all the data subsets are in a similler distrubution, and they are all Symmetry.

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

```
plot(raw_data$A~raw_data$B)
```



A and B are no correlation, they are distributed ramdommly.