Homework 1

Group 2
Harshish Vataliya
Sagar Pancholi

617-982-4228
857-763-8362

vataliya.h@husky.neu.edu
pancholi.sa@husky.neu.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Harshish

Signature of Student 2: Sagar.

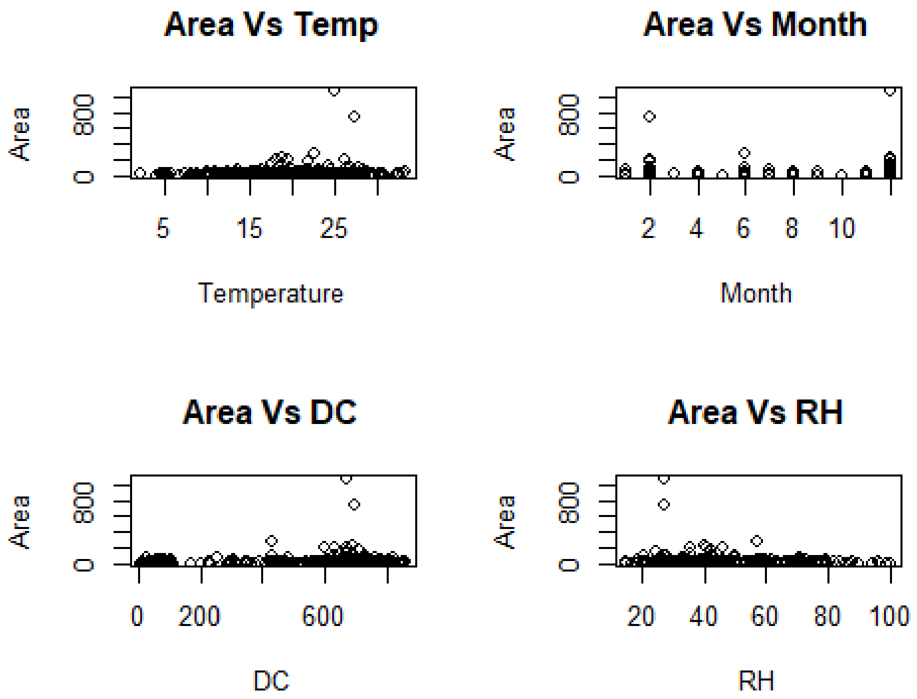
Submission Date: 10/04/2019

Problem 1 (Forest Fires)

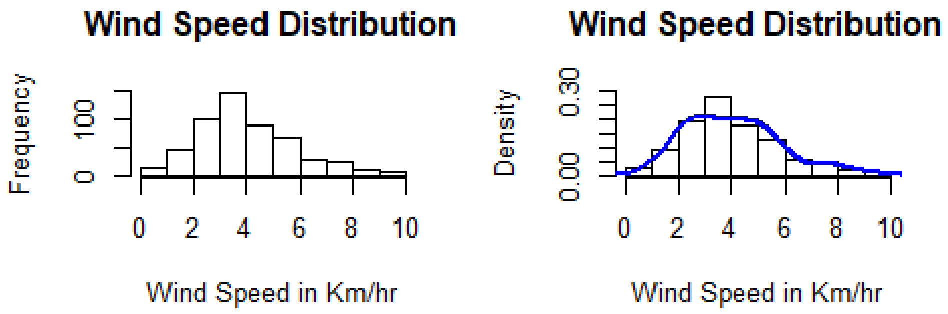
```r
data <- read.csv("forestfires.csv")
str(data)
```

```
## 'data.frame':    517 obs. of  13 variables:
##  $ X    : int  7 7 7 8 8 8 8 8 8 7 ...
##  $ Y    : int  5 4 4 6 6 6 6 6 6 5 ...
##  $ month: Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12
12 ...
##  $ day  : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...
##  $ FFMC : num  86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...
##  $ DMC  : num  26.2 35.4 43.7 33.3 51.3 ...
##  $ DC   : num  94.3 669.1 686.9 77.5 102.2 ...
##  $ ISI  : num  5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...
##  $ temp : num  8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...
##  $ RH   : int  51 33 33 97 99 29 27 86 63 40 ...
##  $ wind : num  6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...
##  $ rain : num  0 0 0 0.2 0 0 0 0 0 0 ...
##  $ area : num  0 0 0 0 0 0 0 0 0 0 ...
```

```r
data$month <- as.numeric(data$month)

#a


par(mfrow=c(2,2))
plot(data$temp,data$area, xlab = "Temperature",ylab = "Area", main = "Area Vs
Temp" )
plot(data$month,data$area, xlab = "Month",ylab = "Area", main = "Area Vs
Month" )
plot(data$DC,data$area, xlab = "DC",ylab = "Area", main = "Area Vs DC" )
plot(data$RH,data$area, xlab = "RH",ylab = "Area", main = "Area Vs RH" )
```



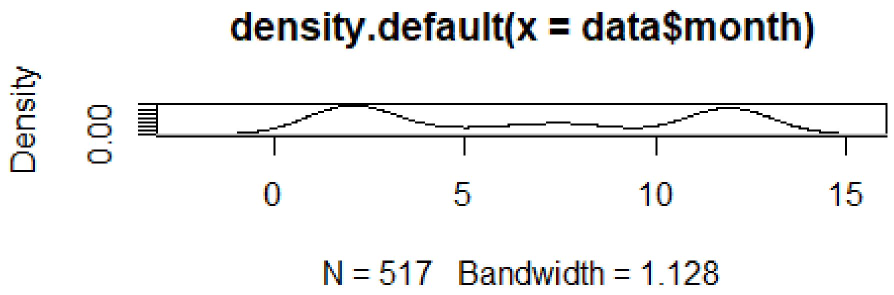
```
#b
hist(data$wind, freq = TRUE, main="Wind Speed Distribution", xlab = "Wind
Speed in Km/hr",ylim = c(0,150))

#c
summary(data$wind)

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

#d
hist(data$wind, freq = FALSE, main="Wind Speed Distribution", xlab = "Wind
Speed in Km/hr",ylim = c(0,0.30))
lines(density(data$wind),col="Blue",lwd=2)

#e
par(mfrow=c(2,1))
```

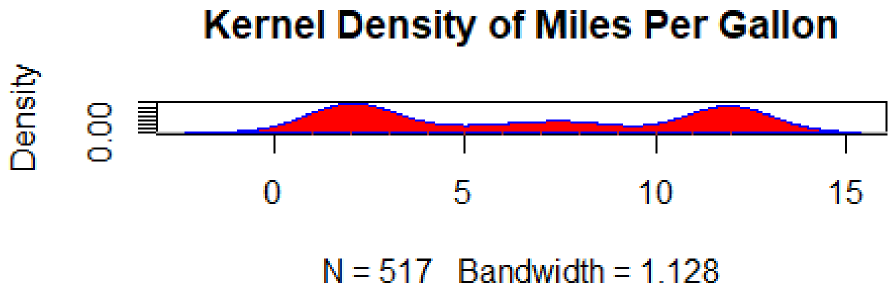


```r
d <- density(data$month)
plot(d)
d <- density(data$month)
plot(d, main="Kernel Density of Miles Per Gallon")
polygon(d, col="red", border="blue")
rug(data$month, col="brown")

#f
#install.packages("GGally")
#install.packages("ggplot2")
library(GGally)
```

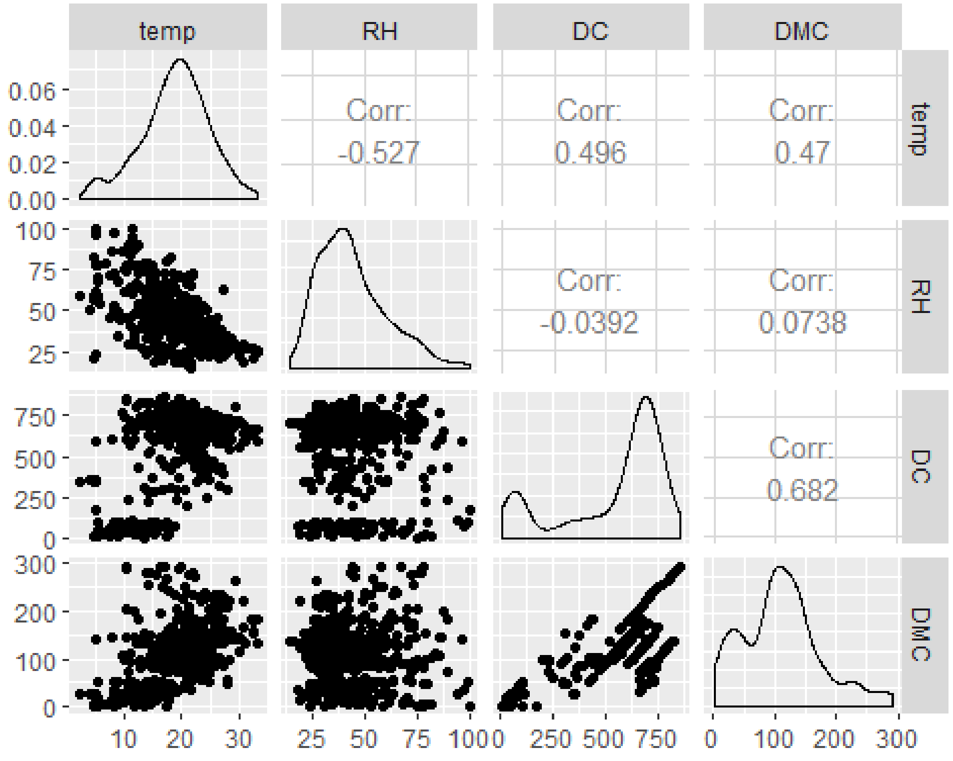
```
## Loading required package: ggplot2

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

>> Based on the scatterplot matrix of Temp, RH, DC and DMC, it can be said that Temp is moderately correlated with DC, DMC and RH (Inversely). DC and DMC are somewhat strongly correlated where as other combinations are not correlated with each other.

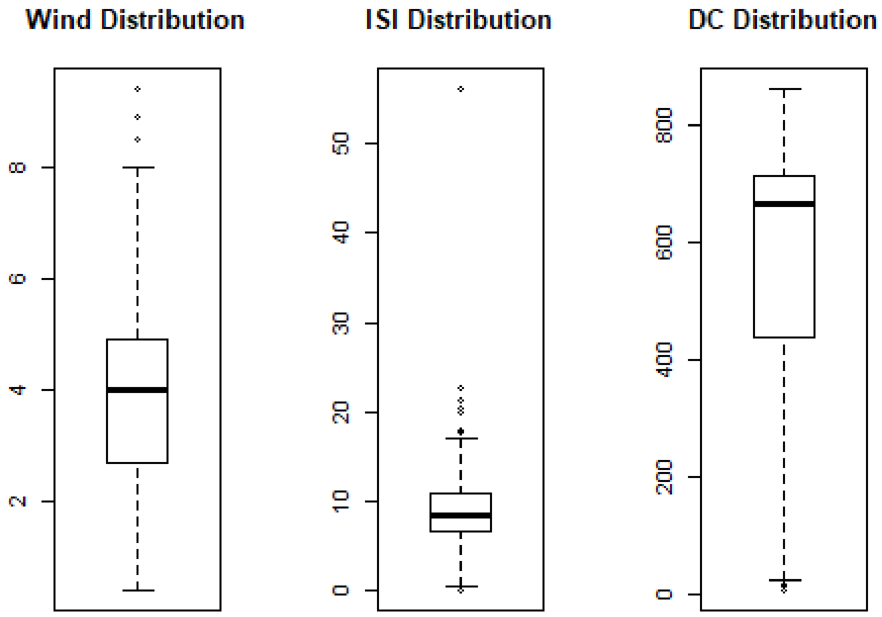




```
library(ggplot2)
ggpairs(data[,c(9,10,7,6)])
```

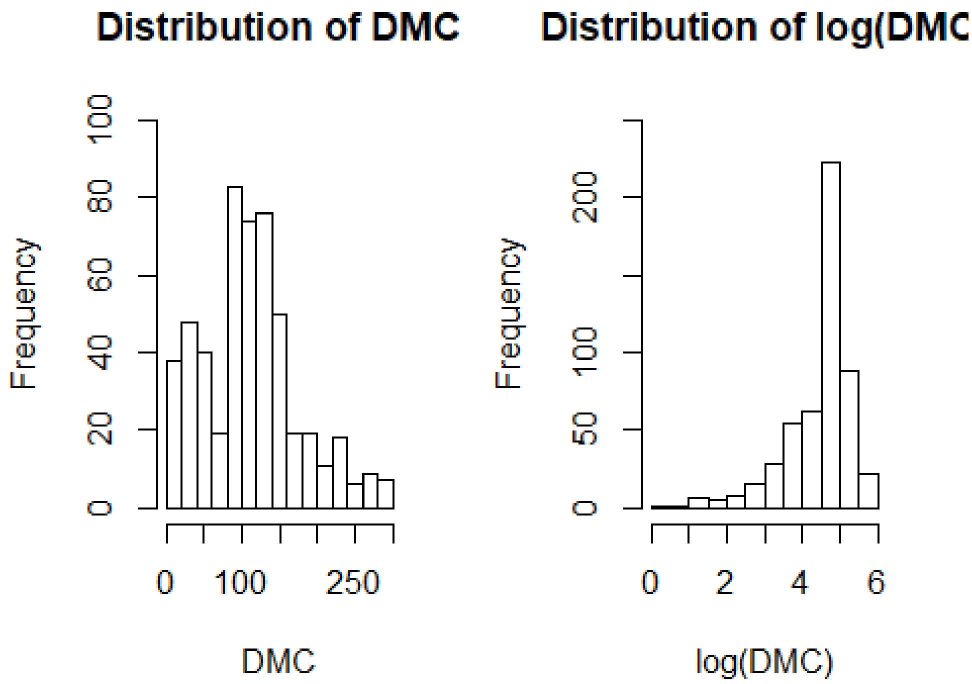


```
#g
par(mfrow=c(1,3))
boxplot(data$wind,main="Wind Distribution")
boxplot(data$ISI,main="ISI Distribution")
boxplot(data$DC,main="DC Distribution")
```

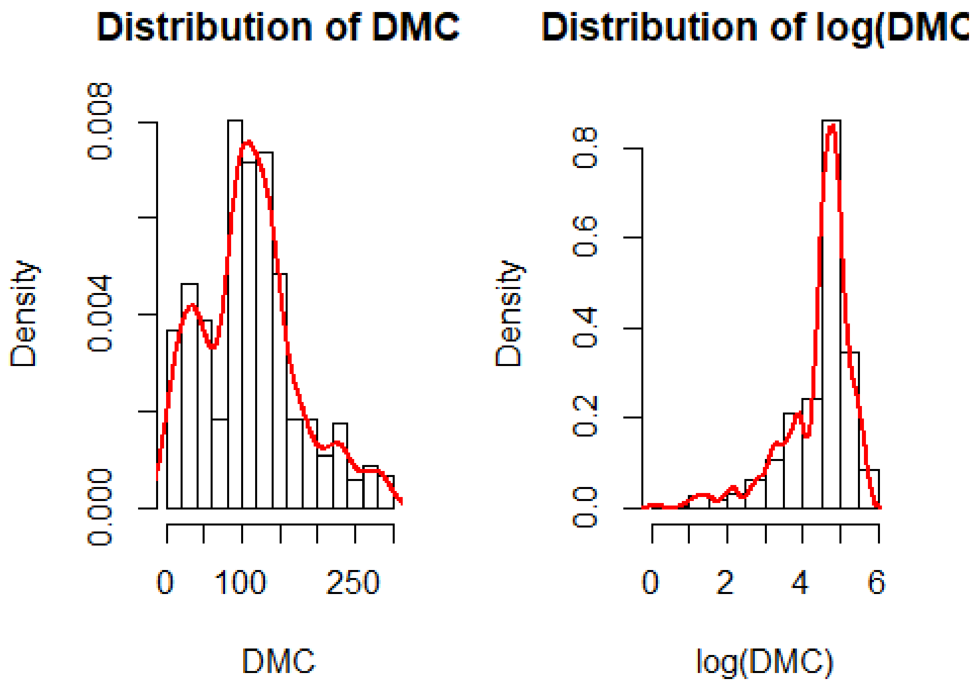


>> Box plot of Wind, ISI and DC shows that there are anomalies/outliers present in the data. Wind data is equally spread and has outliers on the upper end only where as ISI and DC both has outliers on both ends. ISI has one outlier which lies far above the other data which shows that there may be some error in data collection for that record.

```
#h
par(mfrow=c(1,2))
hist(data$DMC,main = "Distribution of DMC",xlab = "DMC",ylim = c(0,100))
hist(log(data$DMC),main = "Distribution of log(DMC)", xlab = "log(DMC)", ylim
= c(0,250))
```



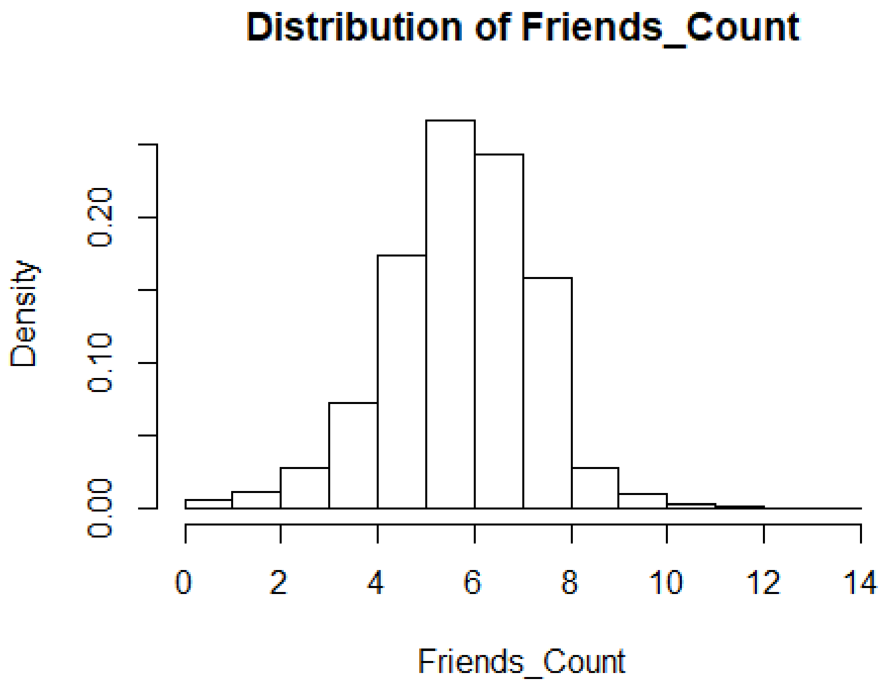
```
par(mfrow=c(1,2))
hist(data$DMC,main = "Distribution of DMC",xlab = "DMC",freq = FALSE)
lines(density(data$DMC),col="Red",lwd=2)
hist(log(data$DMC),main = "Distribution of log(DMC)", xlab = "log(DMC)",freq
= FALSE)
lines(density(log(data$DMC)),col="Red",lwd=2)
```



>> It is difficult to speak about skewness of the data from histogram of DMC whereas log transformation of the data makes it clear that the data is skewed towards right end.

Problem 2 (Tweeter Accounts)

```
#2
#a
data2 <- read.csv("M01_quasi_twitter.csv")

par(mfrow=c(1,1))
hist(log(data2$friends_count),freq = FALSE, main="Distribution of
Friends_Count", xlab = "Friends_Count")

## Warning in log(data2$friends_count): NaNs produced
```

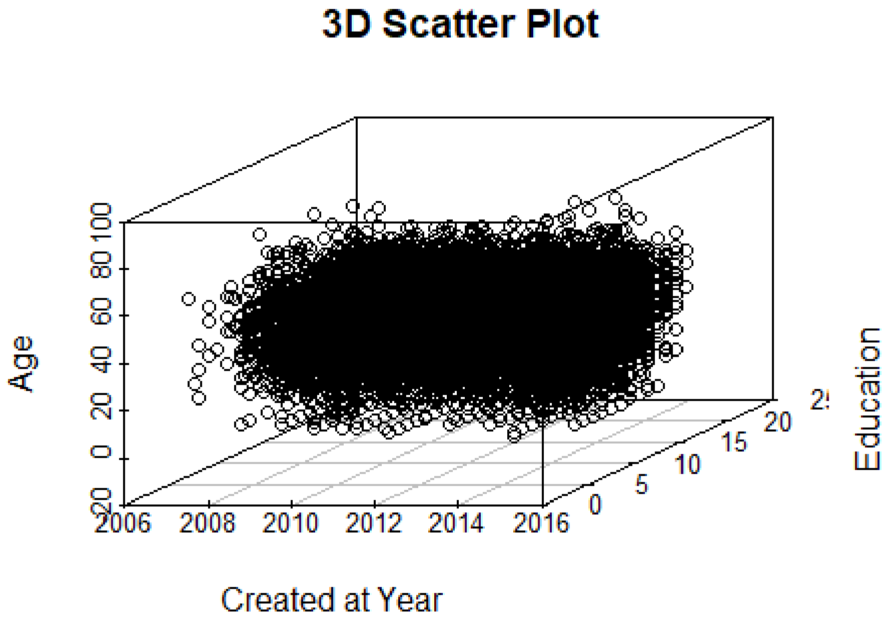


>> Data Distribution of Friends Count variable is uniform and symmetric about its mean and it approximately follows normal distribution, but data is having many outliers.

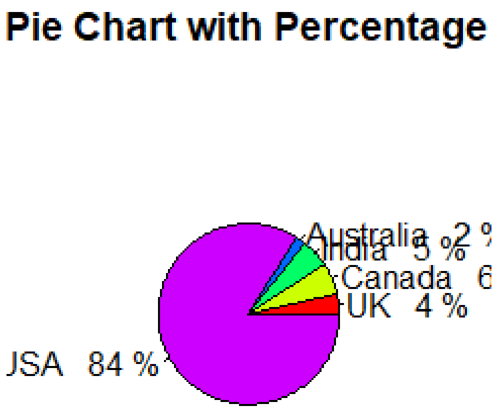
```
#b
summary(data2$friends_count)

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

#c
a<- sum(is.na(data2$friends_count))
a

## [1] 0
```

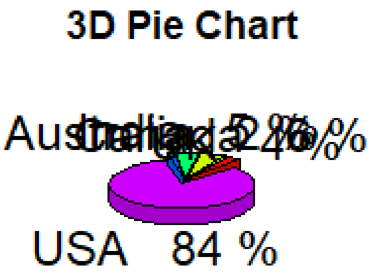
>> Data is free from NA values which shows that data quality is good

```
#d
#install.packages("scatterplot3d")
require("scatterplot3d")

## Loading required package: scatterplot3d

scatterplot3d(data2$created_at_year , data2$education , data2$age, xlab =
"Created at Year", ylab =  "Education", zlab = "Age" ,main = "3D Scatter
Plot" )
```



```
#e
#pie3D(Accounts,labels = labls,explode = 0.1,main="3D Pie Chart",radius =
0.9,start = 0.785 )
#install.packages("plotrix")
library(plotrix)
Countries <- c("UK","Canada","India","Australia","USA")
Accounts <- c(650,1000,900,300,14900)
percentage <- round(Accounts/sum(Accounts)*100)
labls <- paste(Countries," ",percentage,"%",sep = " ")
par(mfrow= c(1,2)) > pie(Accounts,labels = labls,col =
rainbow(length(Countries)),main = "Pie Chart with Percentage")

## logical(0)

pie3D(Accounts,labels = labls,explode = 0.1,main="3D Pie Chart",radius =
0.9,start = 0.785 )
```



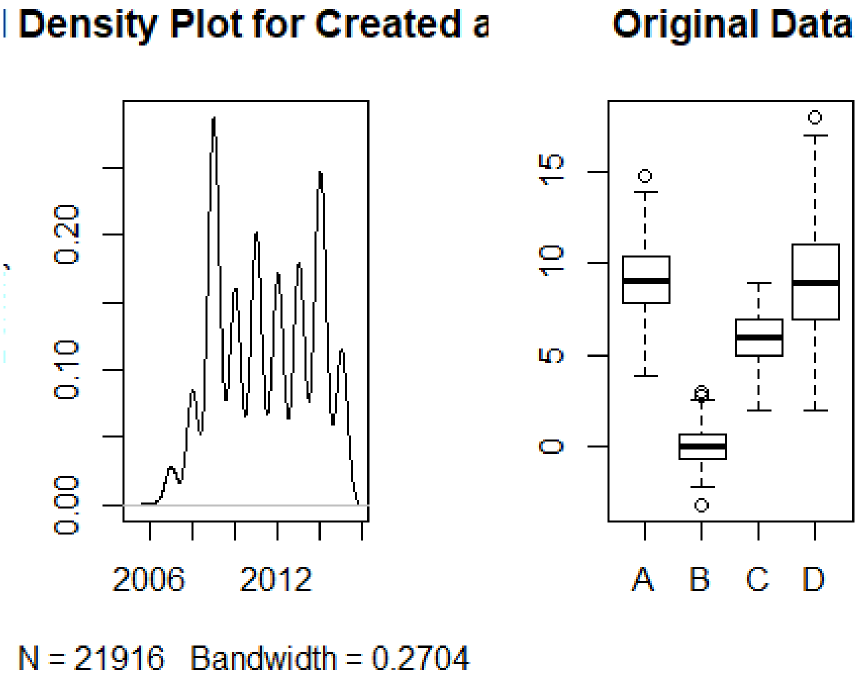


```r
#f
plot(density(data2$created_at_year), main ="Kernal Density Plot for Created
at Year")
```

>> Kernel density plot of the created_at_year variable shows that most of the accounts were created in the year 2009 and then it follows periodic rise and fall for the next years.

Problem 3 (Insurance Claims)

```r
#3
#a
data3<- read.csv("raw_data.csv")
Ndata <- scale(data3)

#b
boxplot(data3, main=" Original Data")
```

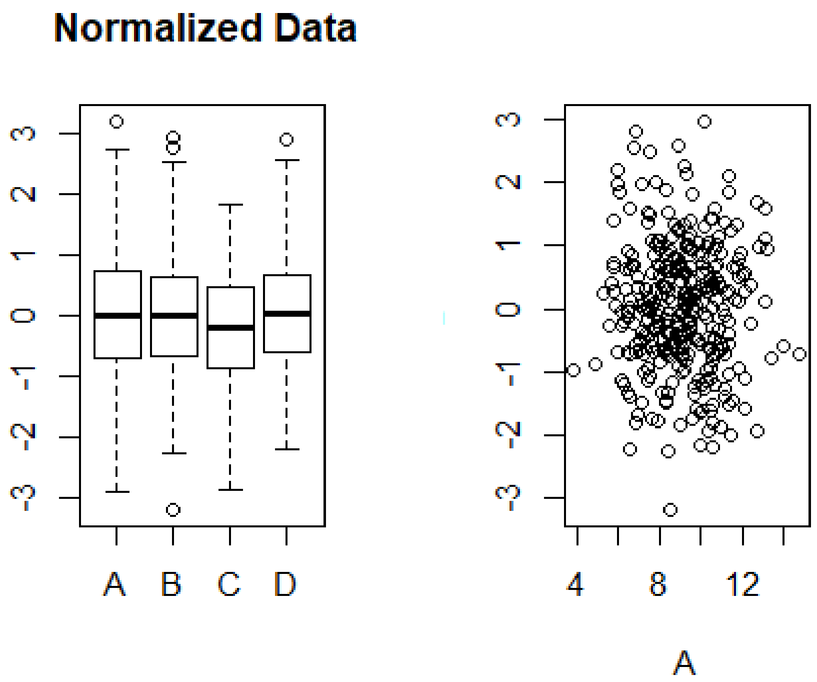


```
#c
boxplot(Ndata, main=" Normalized Data")

#d
```

>>It is very difficult to compare all four variables when plotted on box plot in the original form due to different value ranges. Normalization of the data makes it easier to compare all four variables on the same scale. Normalized Data Box plot shows that Variable A is having highest variance where as variable B has the least.

```
#e
plot(data3$A,data3$B, xlab = "A",ylab = "B")
```

>> From the scatter plot of variable A and B it can be said that, the two variables are not correlated as there is no correlation visible from the plotted data.


Normalized Data
3
2
1
0
-1
-2
-3
A
B
C
D
3
2
1
0
-1
-2
-3
4
8
12
A