Alex-Thinh-Nguyen-R-Sample1

Alex-Thinh Nguyen

2/9/2021

Forest Fires Assignment

```
df_forestfires = read.csv('forestfires(1).csv', na.string = "") #read csv file
library(ggplot2)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

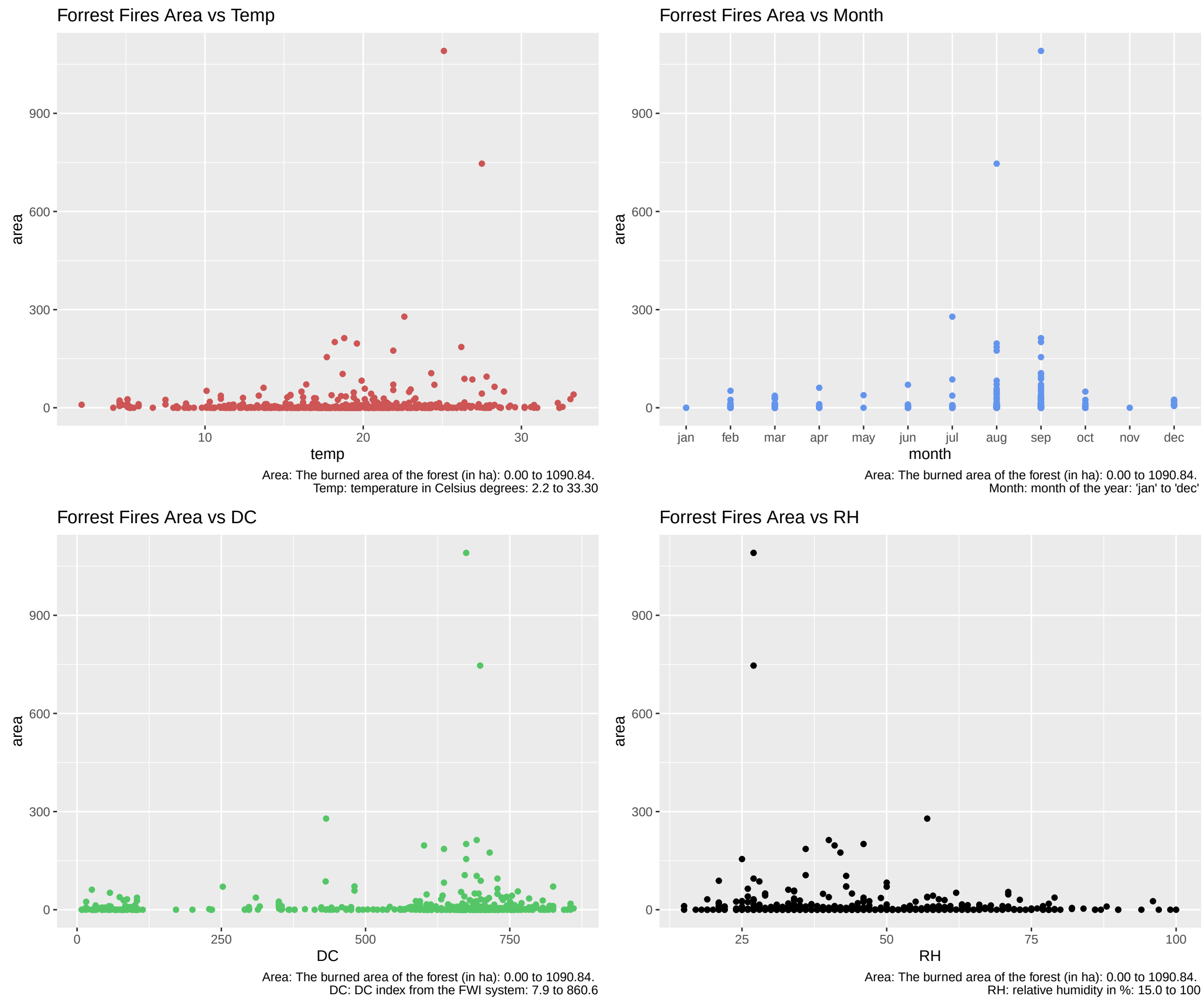
```
library(tidyr)
```

FOrest Fires - a

```
df_forestfires$month = factor(df_forestfires$month, levels = c("jan","feb" ,"mar", "apr","may","jun", "
area_temp_point = ggplot(data = df_forestfires,
                         mapping = aes(x = temp, y = area))  + geom_point(color = "indianred3") + labs(
                         Temp: temperature in Celsius degrees: 2.2 to 33.30")
area_month_point = ggplot(data = df_forestfires,
                         mapping = aes(x = month, y = area)) + geom_point(color = "cornflowerblue") + la
                         Month: month of the year: 'jan' to 'dec'")
area_DC_point = ggplot(data = df_forestfires,
                         mapping = aes(x = DC, y = area)) + geom_point(color = "#55C667FF") + labs(titl
                         DC: DC index from the FWI system: 7.9 to 860.6")
area_RH_point = ggplot(data = df_forestfires,
                         mapping = aes(x = RH, y = area)) + geom_point()+ labs(title = "Forrest Fires A
                         RH: relative humidity in %: 15.0 to 100")
library(gridExtra)
```

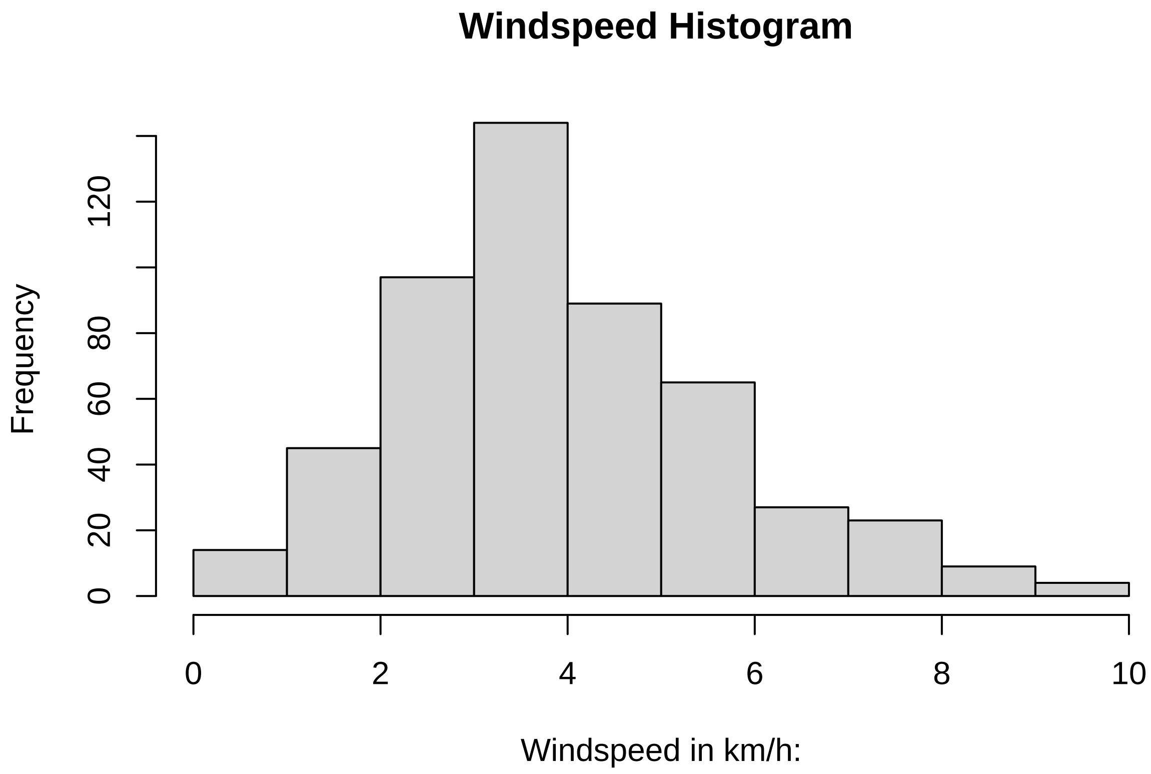
```
##
## Attaching package: 'gridExtra'

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

```
grid.arrange (area_temp_point,
              area_month_point,
              area_DC_point,
              area_RH_point,
              ncol = 2)
```



Forest Fires - b

```
windspeed = pull(select(df_forestfires,wind))
hist(windspeed, breaks = 12, xlab = "Windspeed in km/h:", main = "Windspeed Histogram ")
```



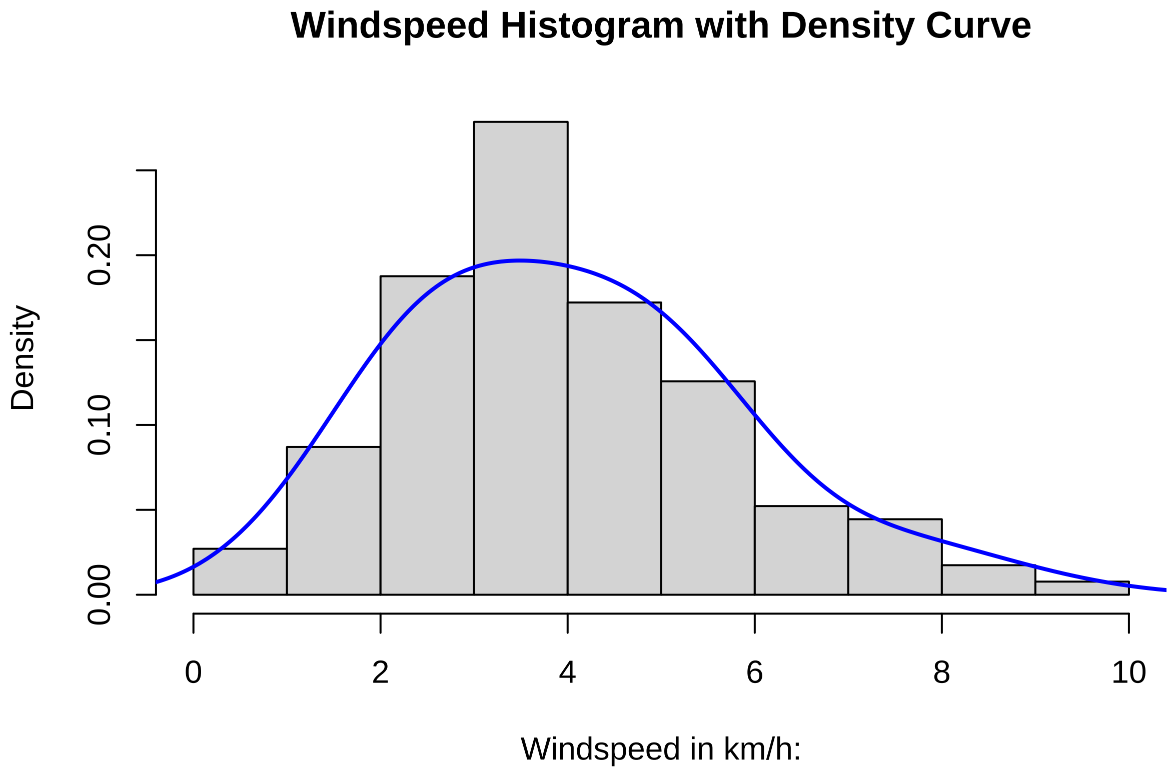
# Forest Fires - c

```
summary(windspeed)
```

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

Forrest Fires - d

```
hist(windspeed, probability = T, xlab = "Windspeed in km/h:", main = "Windspeed Histogram with Density (
lines(density(windspeed, adjust = 2), lwd = 2, col = "blue")
```

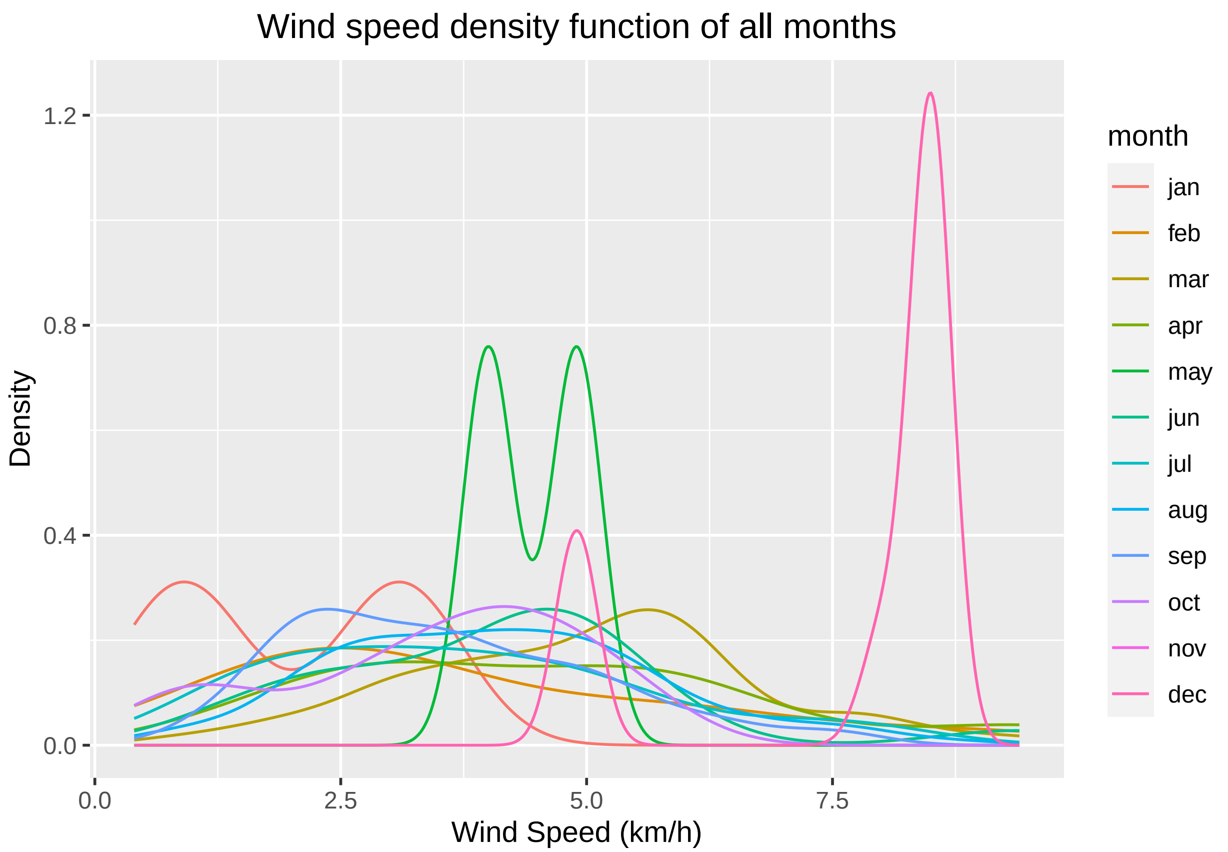


## Forest Fires - e

```
ggplot(df_forestfires, aes (x = wind, color = month)) + geom_line(stat = "density") + labs(title = "Winc
```

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

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

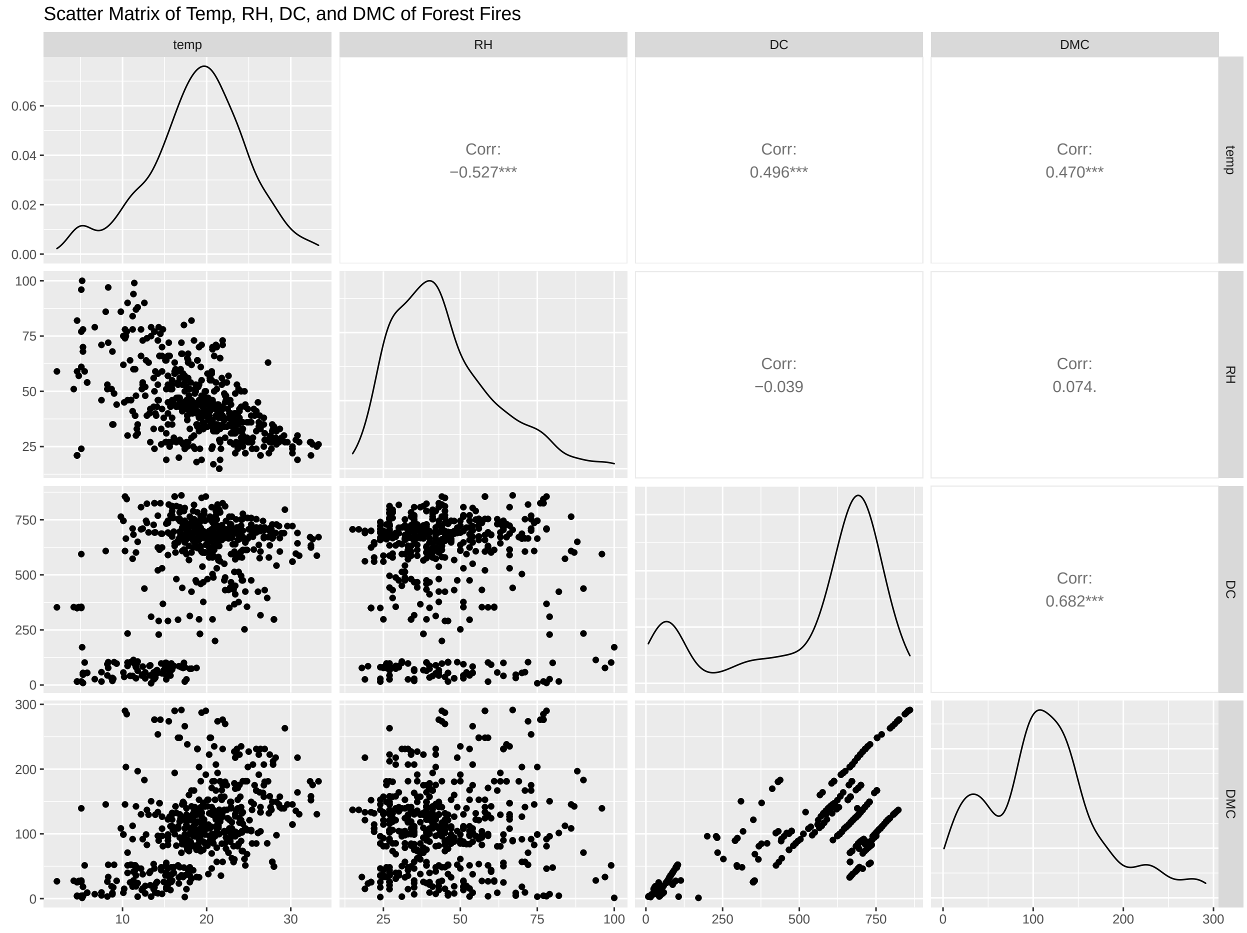


Forest Fires - f

```
library(GGally)
```

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

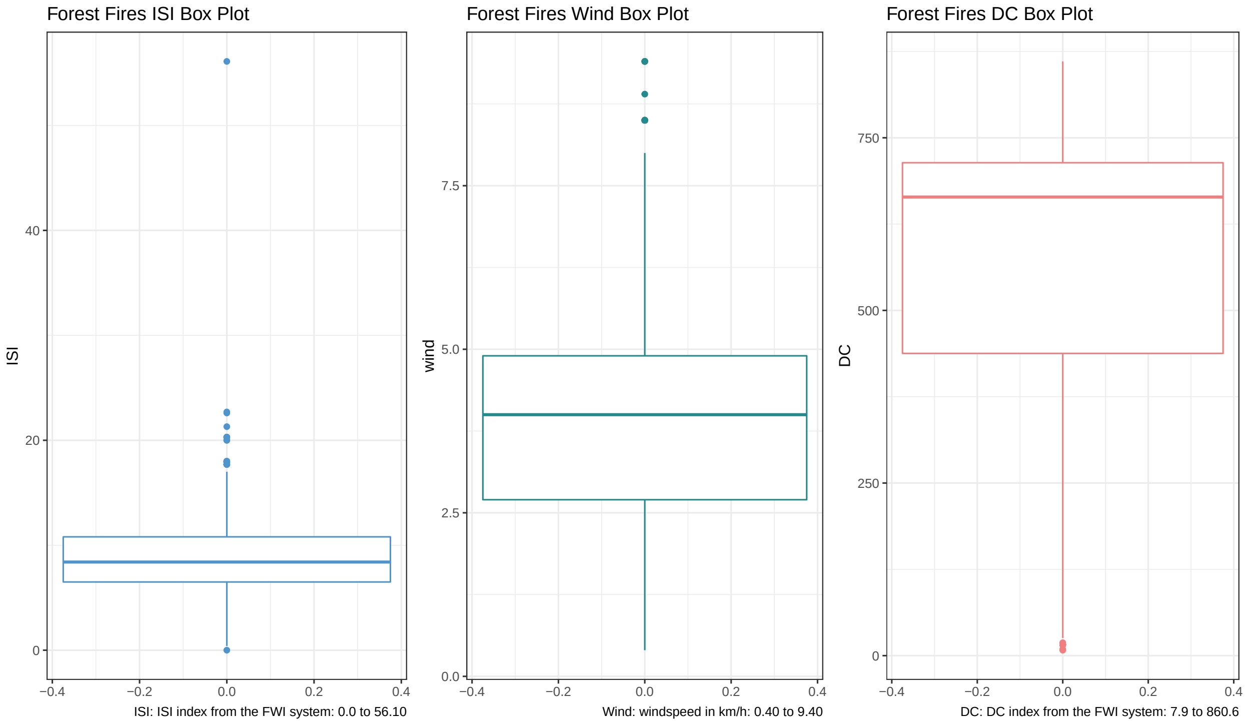
```
ggpairs(select(df_forestfires, temp, RH, DC, DMC)) + labs (title = "Scatter Matrix of Temp, RH, DC, and
```


Scatter Matrix of Temp, RH, DC, and DMC of Forest Fires

```
#Conclusion: medium negative correlation between temp and RH. Temp also hold low positive correlation w
```

Forest Fires - g

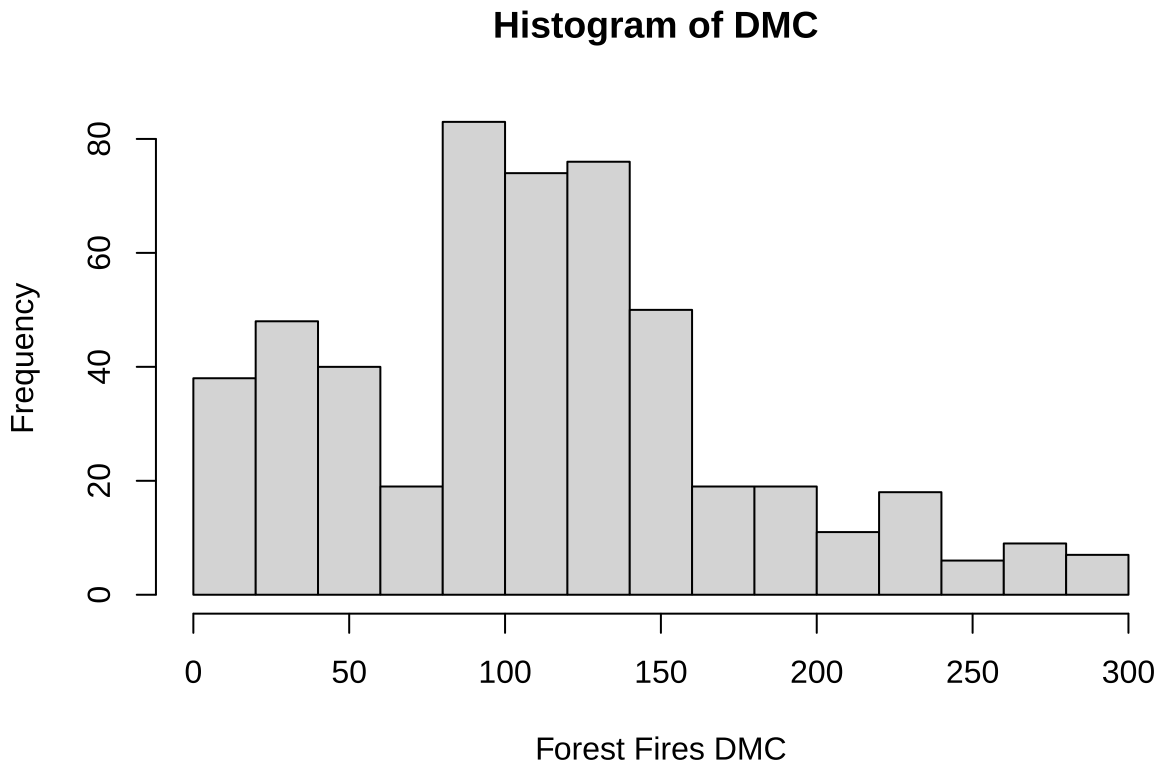
```
#Box plot for Forest Fires Wind
wind = select(df_forestfires,wind)
boxplot_wind = ggplot(data = wind, aes (y = wind)) + geom_boxplot(color = "#238A8DFF") + labs(title = "

#Box plot for Forest Fires ISI
ISI = select(df_forestfires,ISI)
boxplot_ISI = ggplot(data = ISI, aes (y = ISI)) + geom_boxplot(color = "steelblue3") + labs(title = "Fo

#Box plot for Forest Fires DC
DC = select(df_forestfires,DC)
boxplot_DC = ggplot(data = DC, aes (y = DC)) + geom_boxplot(color = "lightcoral") + labs(title = "Fores

grid.arrange(boxplot_ISI, boxplot_wind, boxplot_DC, ncol = 3)
```



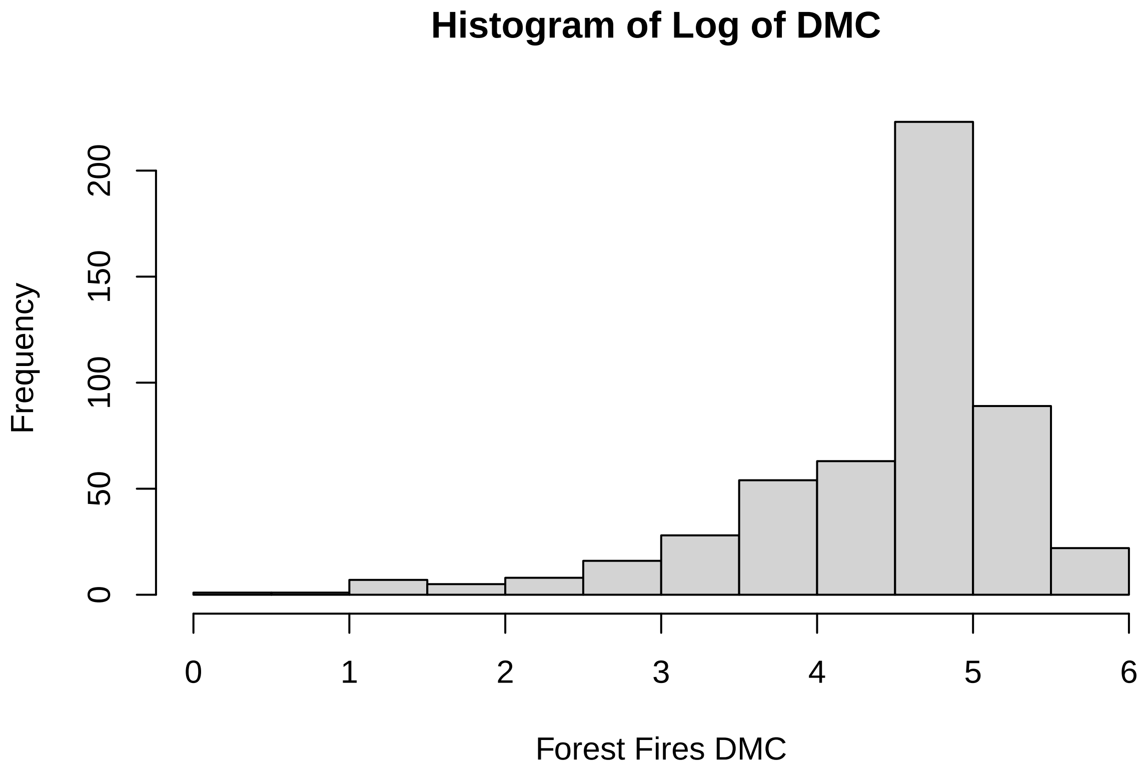
```
#All have outliers. For Wind and ISI, there are some data above 1.5 IQR and for DC, some outliers on bo
```

Forest Fires - h

```
DMCHist = pull(select(df_forestfires,DMC))
hist(DMCHist, xlab = "Forest Fires DMC", main = "Histogram of DMC ")
```



```r
hist(log(df_forestfires$DMC), xlab = "Forest Fires DMC", main = "Histogram of Log of DMC ")
```



```
#Conclusion: histogram is left-skewed so DMC is not a perfect normal distribution. Log of DMC is, on th
```

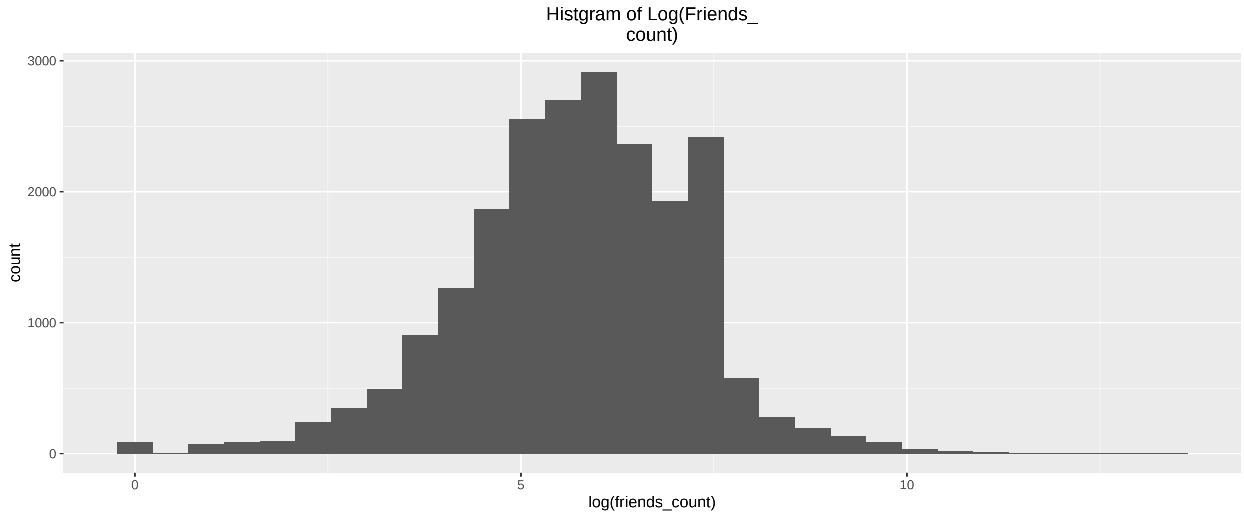
Tweeter Account Assignment

```
df_tweeter = read.csv('M01_quasi_twitter(1).csv', na.string = "")
```

Tweeter Account - a

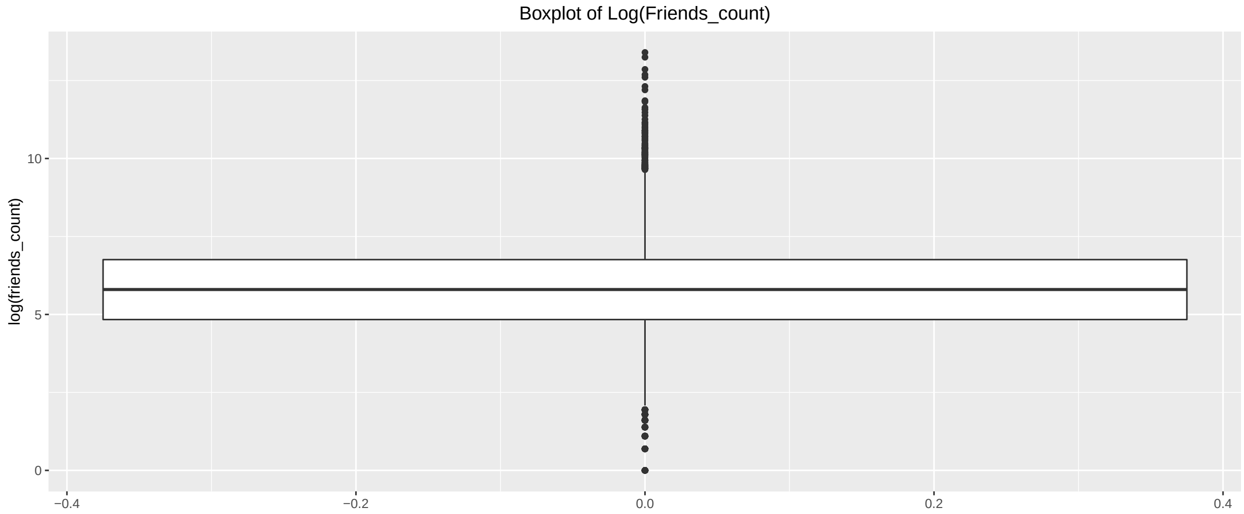
```
val_friend_count = df_tweeter$friends_count

ggplot(df_tweeter,aes(x=log(friends_count)))+ geom_histogram(bins = 30)+ labs(title = "Histgram of Log(
count)")+theme(plot.title=element_text(hjust=0.5))
```

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

## Warning in log(friends_count): NaNs produced

## Warning: Removed 221 rows containing non-finite values (stat_bin).
```



```
ggplot(df_tweeter,aes(y = log(friends_count))) + geom_boxplot() + labs(title = "Boxplot of Log(Friends_c
```

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

## Warning in log(friends_count): NaNs produced

## Warning: Removed 221 rows containing non-finite values (stat_boxplot).
```



Tweeter Account - b & c

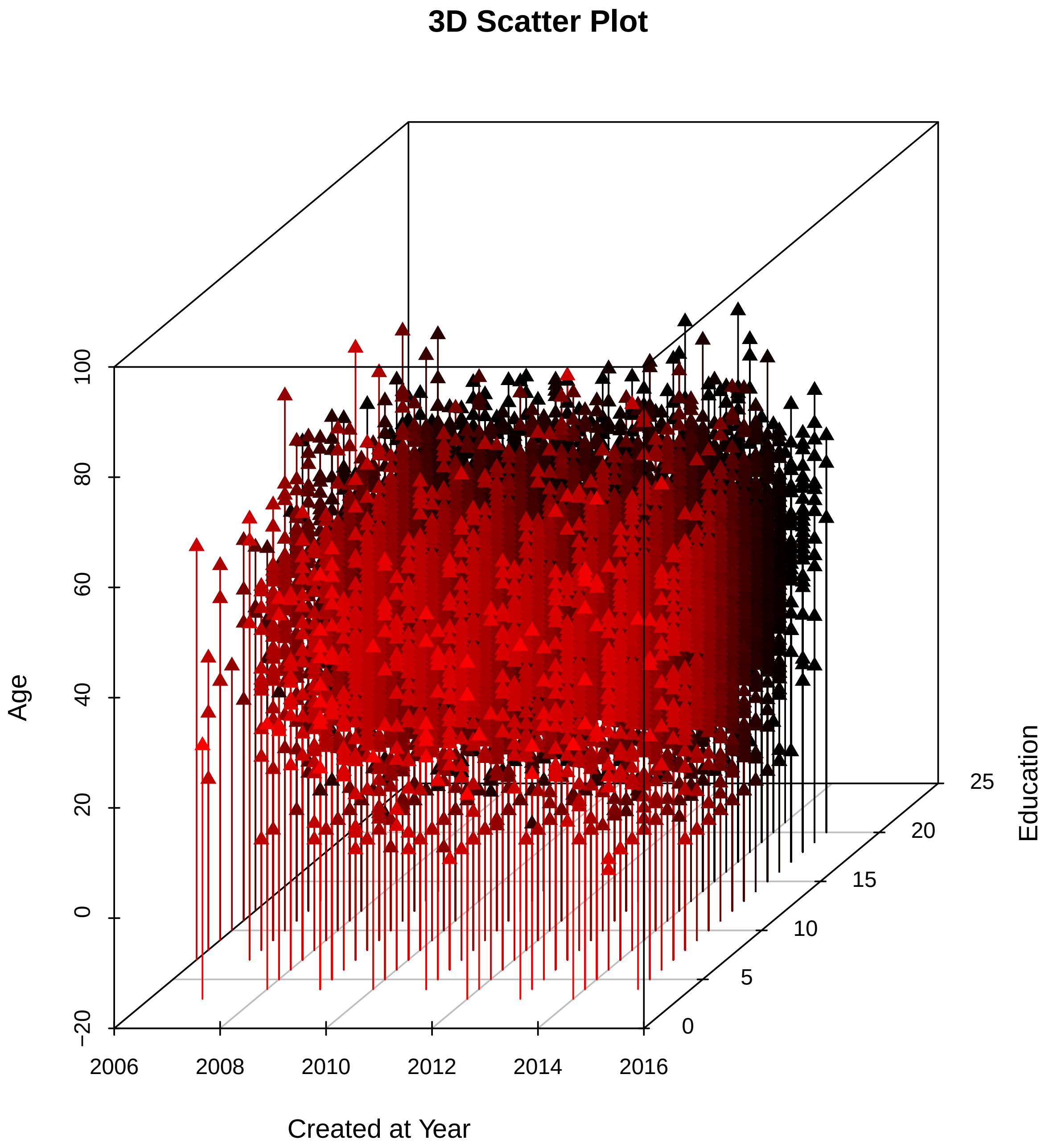
```
friends_count = pull(select(df_tweeter,friends_count))
summary(friends_count)
```

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

```
#Conclusion: not normally distributed. All ranges (including IQR) are massive. Distance between major m
#Outliers certainly affect interpretations.The data quality is not good. From above, we can see that th
```

Tweeer Account - d

```
library(scatterplot3d)
scatterplot3d(df_tweeter$created_at_year,df_tweeter$education ,df_tweeter$age,
              pch = 17,
              highlight.3d = TRUE,
              type="h",
              main="3D Scatter Plot",
              xlab="Created at Year",
              ylab = "Education",
              zlab = "Age")
```

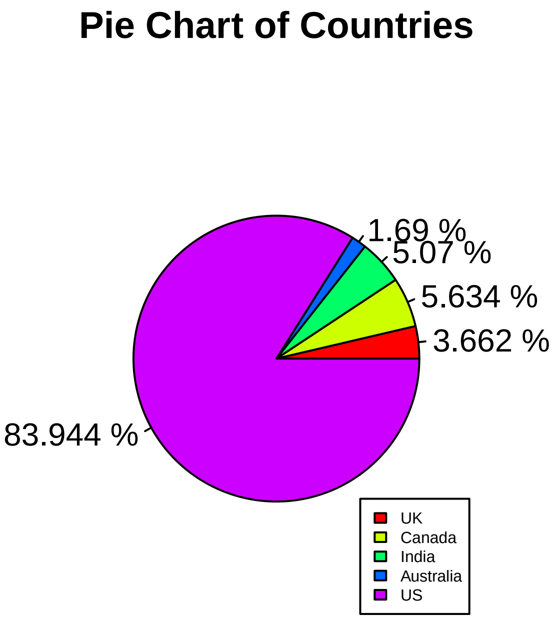


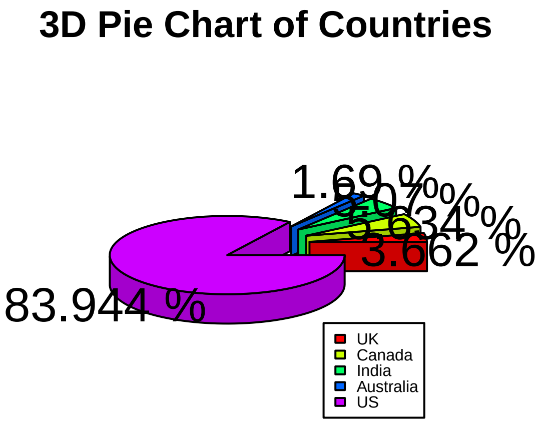
```
split.screen(c(1,2))
```

```
## [1] 1 2
```

```
screen(1)
country = data.frame(country = c("UK","Canada","India","Australia","US"), num = c(650,1000,900,300,14900
piepercent = paste(round(100*country$num/sum(country$num),3), "%")
```

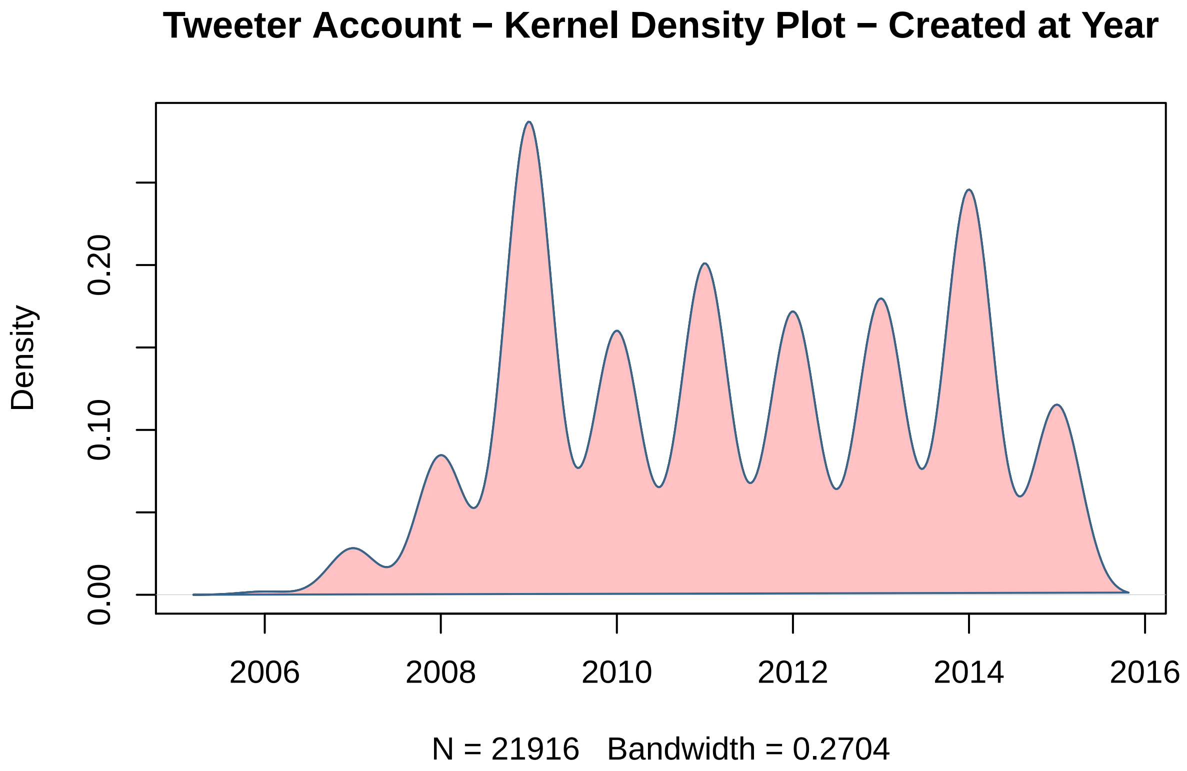
```
piechart = pie(x=country$num, labels =piepercent, main = "Pie Chart of Countries",radius =0.8,col = rai
legend("bottomright",country$country,fill =rainbow(length(country$country)),xpd= TRUE,cex=0.5 )
screen(2)
library("plotrix")
piepercent = paste(round(100*country$num/sum(country$num),3), "%")
pie3D(x=country$num, labels =piepercent , main = "3D Pie Chart of Countries",radius=0.8,height=0.1,
explode = 0.3,
col = rainbow(length(country$country)))
legend('bottomright',country$country,fill = rainbow(length(country$country)),xpd= TRUE,cex=0.5 )
```





Tweeter Account - f

```
kernel_year = density(df_tweeter$created_at_year)
plot(kernel_year, main = "Tweeter Account - Kernel Density Plot - Created at Year")
polygon(kernel_year, col = "rosybrown1", border = "steelblue4")
```



```
## From the above graph, we can see that the kernel density plot provides a more smoothing way to see t
## Also ,we can see that the period between 2008 and 2017 is a peak of the user creating their account.
```

Insurance Claim Assignment

```
df_insurance = read.csv(’raw_Data(1).csv’, na.string = "")
head(df_insurance) #Before
```

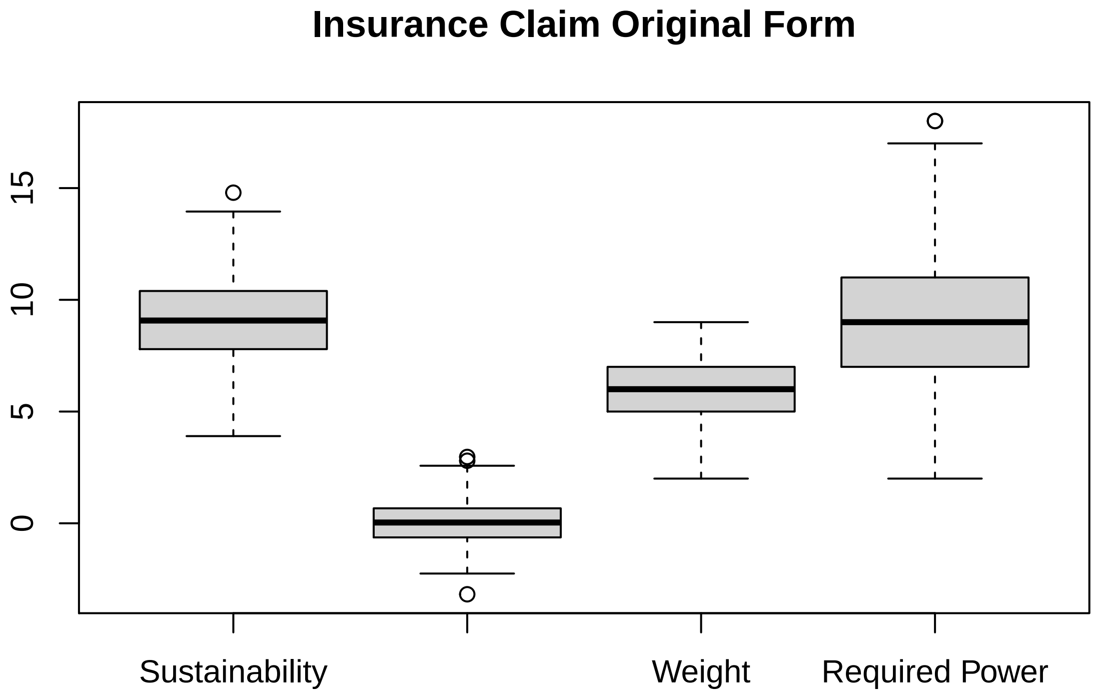
```
##           A          B C  D
## 1  8.257164 -0.6560755 6  8
## 2 10.557378 -0.7158294 7  8
## 3  8.744211  0.7996106 7  5
## 4  6.555028  1.5832173 6 10
## 5  9.362121  1.0272024 7  8
## 6  9.020671  0.7197130 7 12
```

```
Ndata = df_insurance
names(Ndata)[1] = "Sustainaility"
names(Ndata)[2] = "Carbon_Footprint"
names(Ndata)[3] = "Weight"
names(Ndata)[4] = "Required_Power"
Ndata = as.data.frame(scale(Ndata[,1:4]))
head(Ndata) #After
```

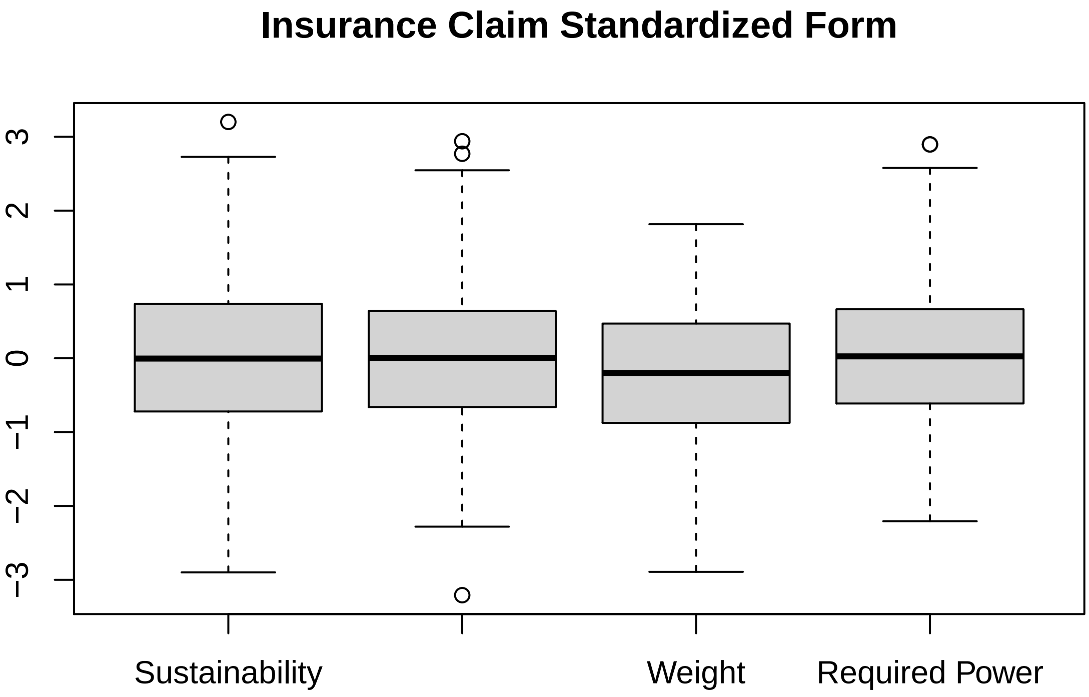
```
##   Sustainaility Carbon_Footprint     Weight Required_Power
## 1   -0.46047167       -0.6870000 -0.2019694     -0.2931233
## 2    0.82780052       -0.7467798  0.4705888     -0.2931233
## 3   -0.18769316        0.7693173  0.4705888     -1.2500845
## 4   -1.41378095        1.5532638 -0.2019694      0.3448509
## 5    0.15837732        0.9970078  0.4705888     -0.2931233
## 6   -0.03285735        0.6893851  0.4705888      0.9828251
```

Insurance Claim - b & c & d

```
insurance_box_1 = boxplot(df_insurance$A,df_insurance$B, df_insurance$C, df_insurance$D, names = c("Sus
```



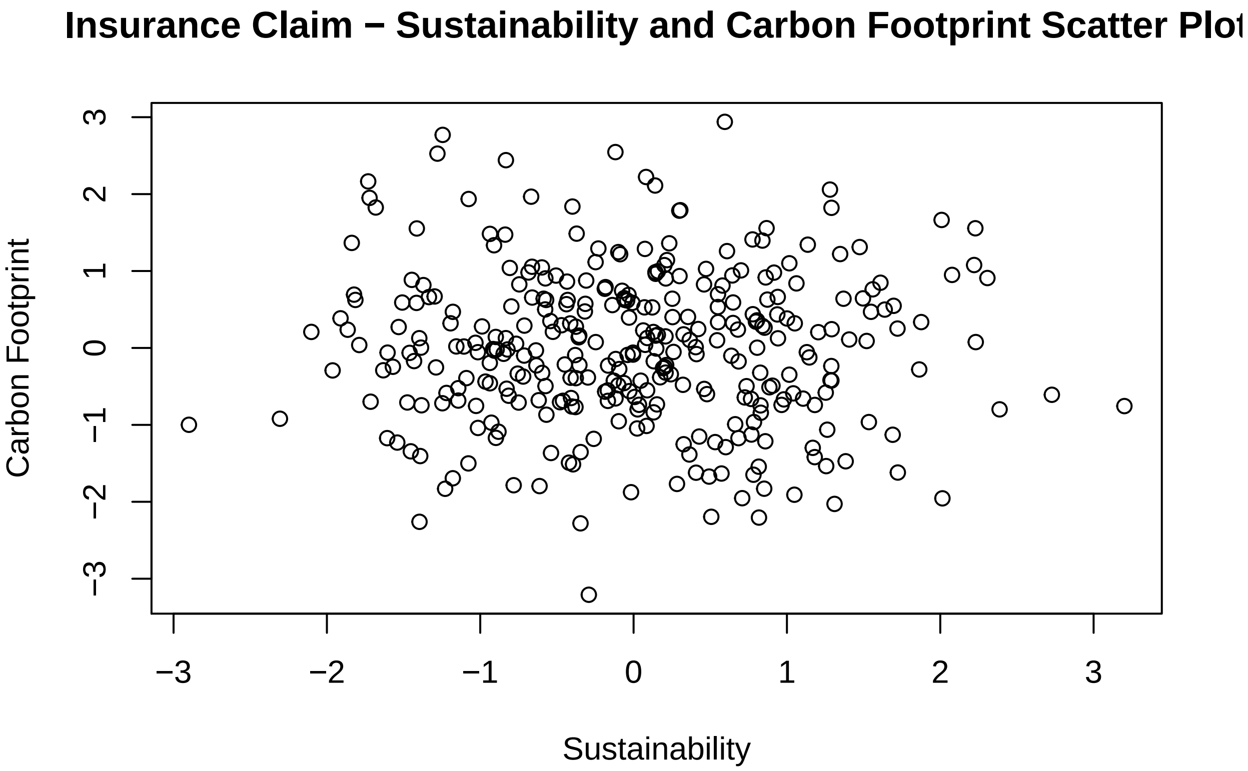
```
insurance_box_2 = boxplot(Ndata$Sustainaility,Ndata$Carbon_Footprint, Ndata$Weight, Ndata$Required_Powe
```



```
#Conclusion: Standardize form yield different range, median, mean, Q1, Q3, max though poses same number
```

Insurance Claim - e

```
plot(Ndata$Sustainaility, Ndata$Carbon_Footprint, xlab = "Sustainability", ylab = "Carbon Footprint", ma
```



```
cat("Correlation is: ", cor(Ndata$Sustainaility, Ndata$Carbon_Footprint)) #Calculate correlation
```

```
## Correlation is:  -0.03059086
```

```
cor.test(x = df_insurance$A , y = df_insurance$B)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_insurance$A and df_insurance$B
## t = -0.55681, df = 331, p-value = 0.578
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13761417  0.07713851
## sample estimates:
##         cor
## -0.03059086
```

```
cat("Covariance is : " ,cov(x = df_insurance$A , y = df_insurance$B))
```

```
## Covariance is :  -0.05459638
```

```
#Conclusion: weak negative correlation and covariance ~ almost 0 -> Sustainability and Carbon footprint
```