Homework 1

Group 5 Charu Aggarwal

1. Mei-Chun Hung
2. Sukanya Aswini Dutta
3. Vaibhavi Gaekwad
4. Wasinee Sriapha
5. Yufei Wang

206-880-9298 (Tel of Student 1)
206-941-3762 (Tel of Student 2)
425-624-5609 (Tel of Student 3)
206-380-6328 (Tel of Student 4)
206-319-8422(Tel of Student 5)

Percentage of Effort Contributed by Student 1:_____20%_____
Percentage of Effort Contributed by Student 2:_____20%_____
Percentage of Effort Contributed by Student 3:_____20%_____
Percentage of Effort Contributed by Student 4:_____20%_____
Percentage of Effort Contributed by Student 5:_____20%_____

Signature of Student 1:____MH_____
Signature of Student 2:____SAD_____
Signature of Student 3:____VG_____
Signature of Student 4:____WOS_____
Signature of Student 5:____YW_____

Submission Date:_____02/29/20_____

IE7275 HW1 Group 5

Group 5

2/25/2020

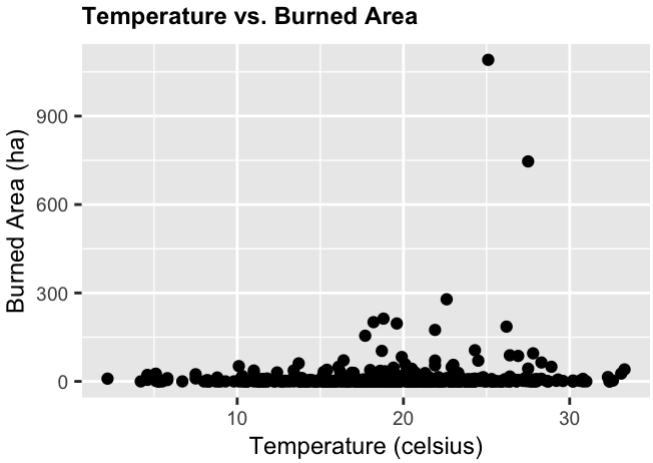
```
#install.packages("webshot")
#webshot::install_phantomjs()
library(webshot)
```

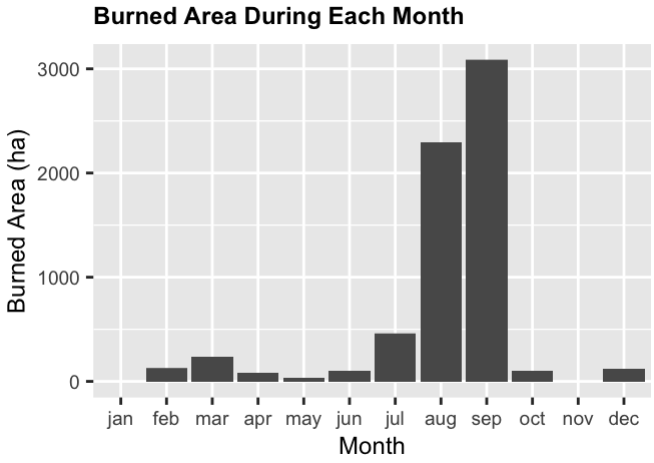
Problem 1 (Forest Fires)

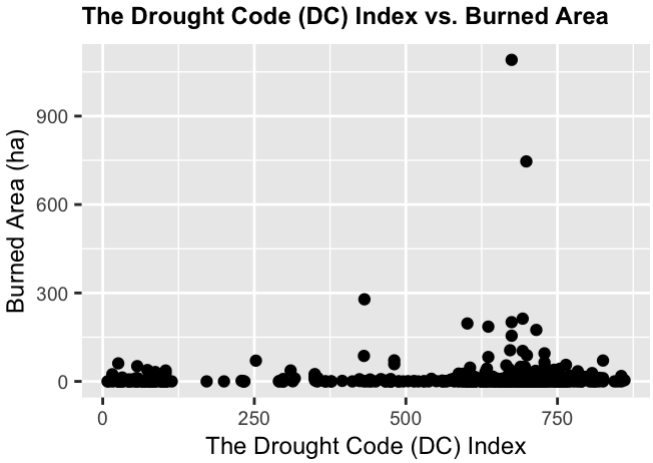
```
library(tidyverse)
df_f <- read_csv(file="/Users/mc.hung9298/Desktop/IE7275/forestfires.csv")
```

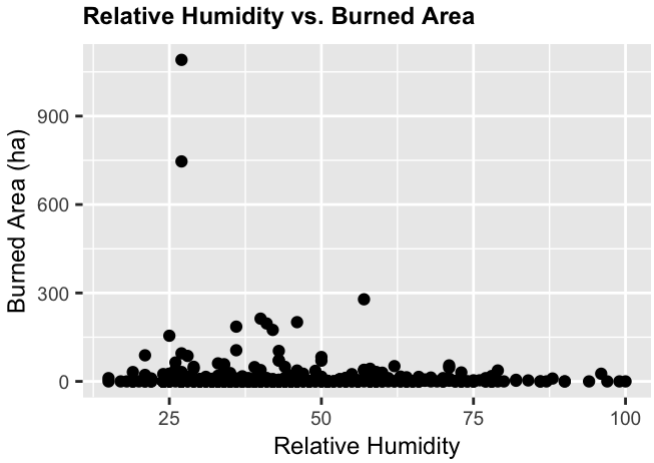
a. Plot are vs. temp, area vs. month, area vs. DC, area vs. RH for Jan to Dec combined in 1 graph.

```
#install.packages("cowplot")
library(ggplot2)
library(cowplot)
p1<- ggplot(df_f)+
  geom_point(aes(x=temp, y=area)) +
  xlab("Temperature (celsius)") +
  ylab("Burned Area (ha)") +
  ggtitle("Temperature vs. Burned Area") +
  theme(text = element_text(size=9), plot.title = element_text( size = 9, face = "bold"
))

p2<- ggplot(df_f)+
  geom_bar(aes(x=month, y=area), stat = "identity") +
  scale_x_discrete(limits=c("jan","feb","mar","apr","may","jun",
                            "jul","aug","sep","oct","nov","dec"))+
  xlab("Month") +
  ylab("Burned Area (ha)") +
  ggtitle("Burned Area During Each Month") +
  theme(text = element_text(size=9), plot.title = element_text( size = 9, face = "bold"
))

p3<- ggplot(df_f)+
  geom_point(aes(x=DC, y=area)) +
  xlab("The Drought Code (DC) Index") +
  ylab("Burned Area (ha)") +
  ggtitle("The Drought Code (DC) Index vs. Burned Area") +
  theme(text = element_text(size=9), plot.title = element_text( size = 9, face = "bold"
))


p4<- ggplot(df_f)+
  geom_point(aes(x=RH, y=area)) +
  xlab("Relative Humidity") +
  ylab("Burned Area (ha)") +
  ggtitle("Relative Humidity vs. Burned Area") +
  theme(text = element_text(size=9), plot.title = element_text( size = 9, face = "bold"
))

plot_grid(p1, p2, p3, p4, ncol=2, nrow=2)
```

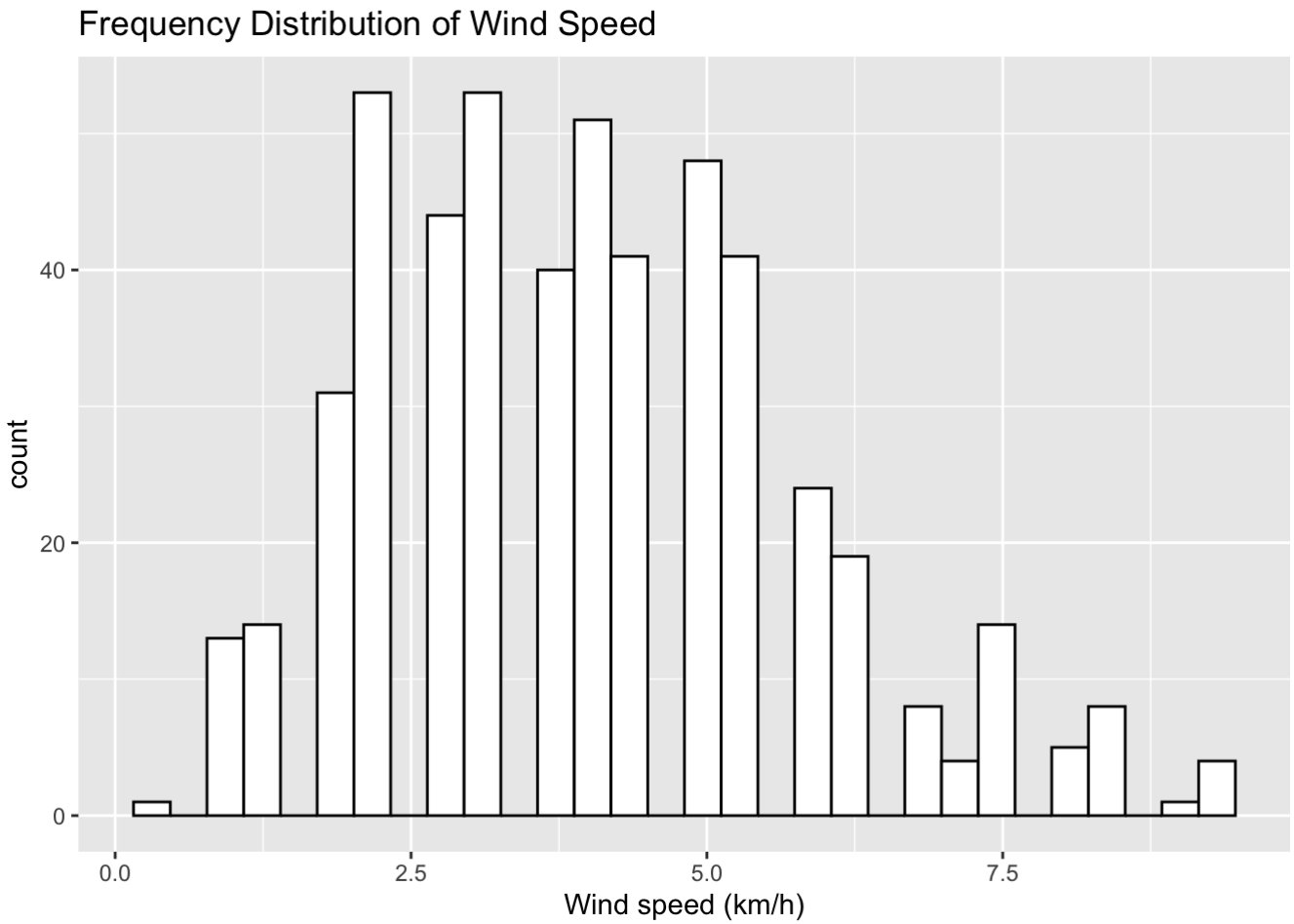









```
ggplot(df_f)+
  geom_histogram(aes(x=wind),color="black", fill="white",position="identity") +
  xlab("Wind speed (km/h)") +
  ggtitle("Frequency Distribution of Wind Speed")
```



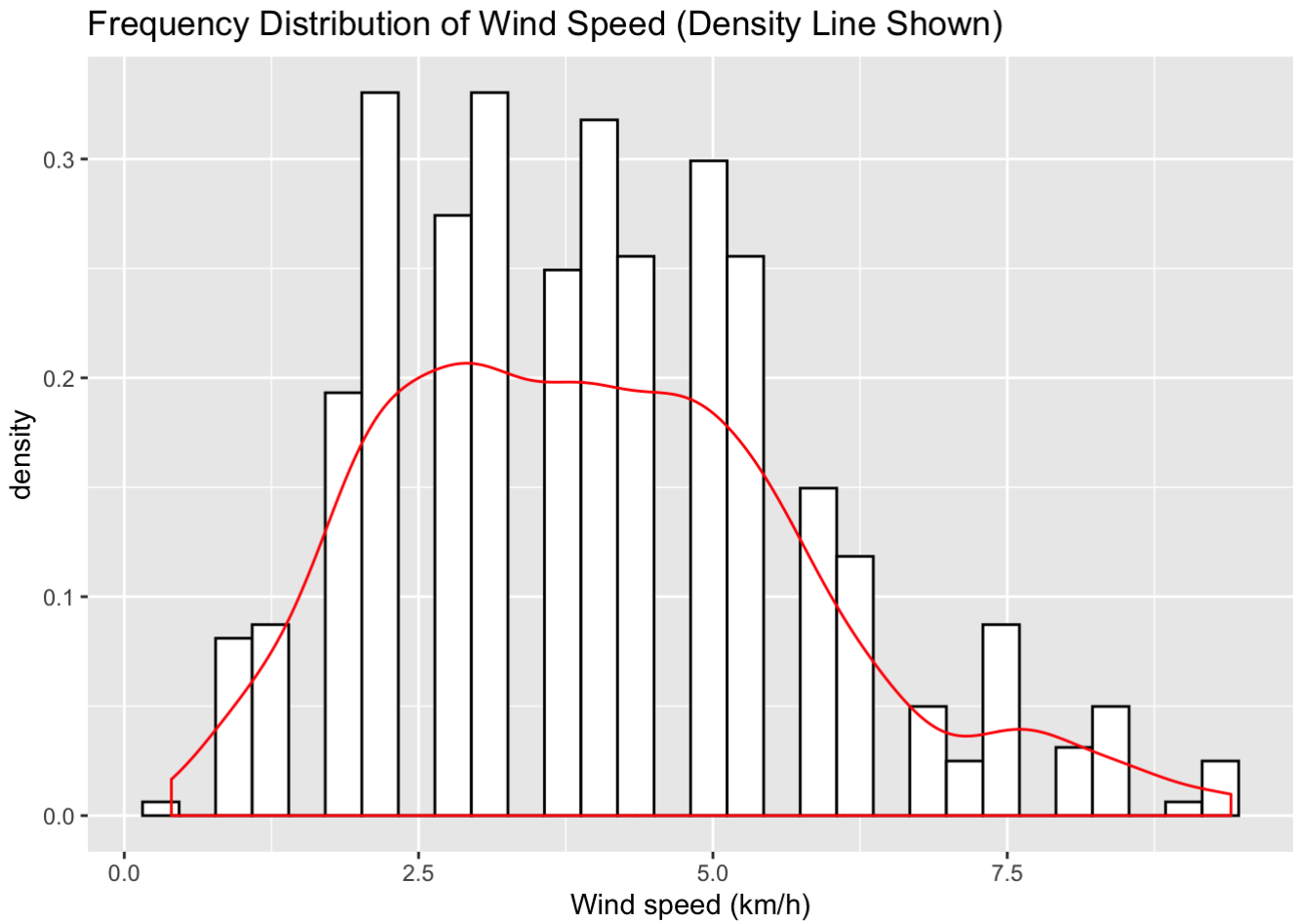
c. Compute the summery statistics (min, 1Q, mean, median, 3Q, max,) of part b

```
df_f_wind <- df_f[,11]
summary(df_f_wind)
```

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

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

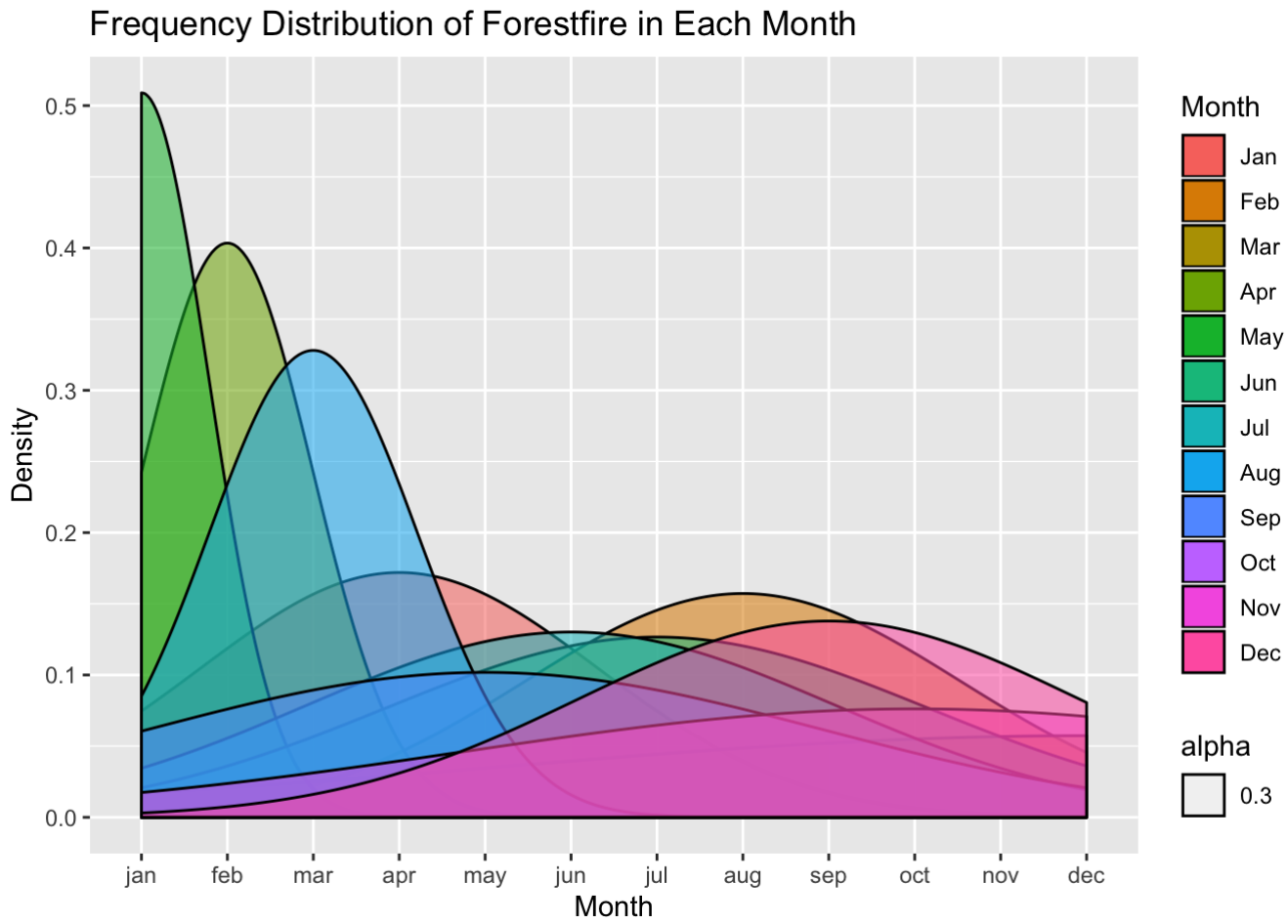
```
ggplot(df_f, aes(x=wind))+
  geom_histogram(aes(y=..density..), color="black",
                 fill="white", position="identity")+
  geom_density(color = "red") +
  xlab("Wind speed (km/h)") +
  ggtitle("Frequency Distribution of Wind Speed (Density Line Shown)")
```



e. Plot the density function of each month of the 12 months, possibly on one plot.

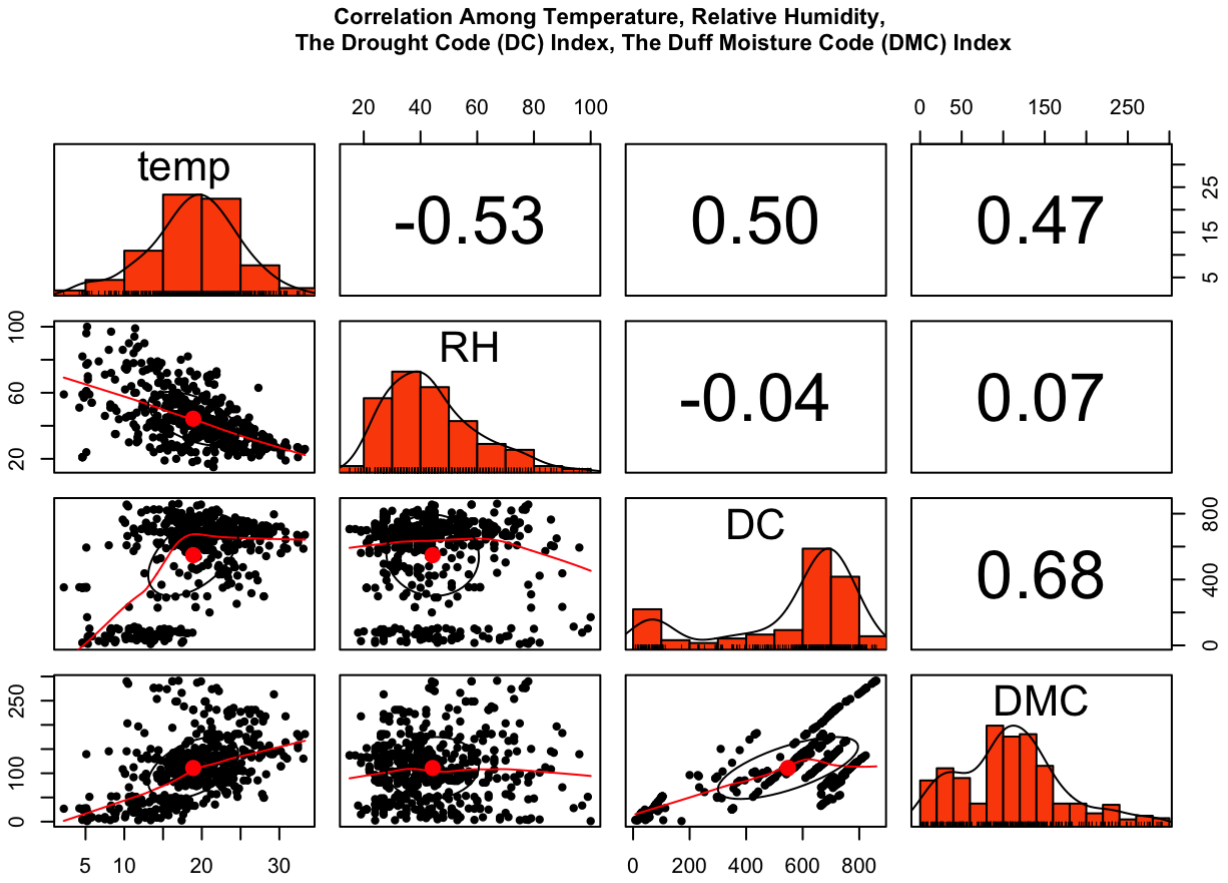
```
qplot(month, data=df_f, geom="density", fill=month, alpha = 0.3) +
scale_x_discrete(limits=c("jan","feb","mar","apr","may","jun",
                          "jul","aug","sep","oct","nov","dec")) +
  scale_fill_discrete(name="Month",
                      labels=c("Jan","Feb","Mar","Apr","May","Jun",
                               "Jul","Aug","Sep","Oct","Nov","Dec")) +
  xlab("Month") + ylab("Density") +
  ggtitle("Frequency Distribution of Forestfire in Each Month")
```

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



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

```
#install.packages("psych")
library(psych)
pairs.panels(df_f[,c(9,10,7,6)],
             method = "pearson", # correlation method
             hist.col = "#FC4E07",
             density = TRUE, # show density plots
             ellipses = TRUE, # show correlation ellipses
             main = "Correlation Among Temperature, Relative Humidity,
             The Drought Code (DC) Index, The Duff Moisture Code (DMC) Index",
             cex.main= 0.7
)
```

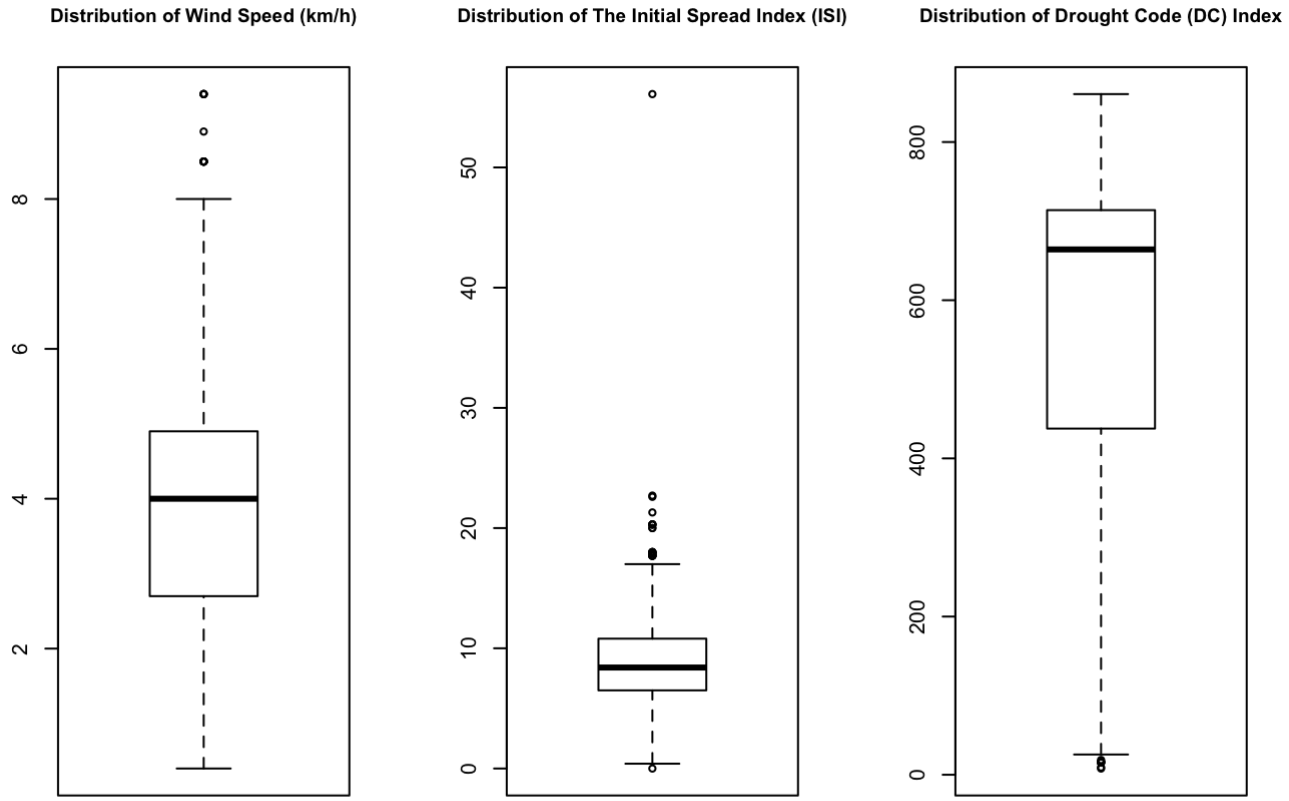


Interpretation

Both correlation coeffients between “Temperature & Drought Code (DC)” and “Temperature & Duff Moisture Code (DMC)” are close to 0.5, which indicate a mild degree of linear relationship. The correlation between “Temperature & Relative Humidity (RH)” of -0.5 has similar degree of correlation, but in a negative direction. This indicates an inverse correlation between two variables. In other words, if one increases, the other will decrease. The correlation coeffients between “RH & DC” and “RH & Duff Moisture Code(DMC)” are close to 0, which indicate low-to-none linear relationship between them. The strongest correlation among all pairs in this dataset is that of “DC & DMC”. However, the value is only 0.68, which is by no means a high degree of correlation.

g. Create boxplot for wind, ISI, and DC. Are there anomalies/outliers.

```
par(mfrow=c(1,3))
boxplot(df_f[,11], main="Distribution of Wind Speed (km/h)", cex.main= 0.9)
boxplot(df_f[,8], main="Distribution of The Initial Spread Index (ISI)", cex.main= 0.9)
boxplot(df_f[,7], main="Distribution of Drought Code (DC) Index", cex.main= 0.9)
```

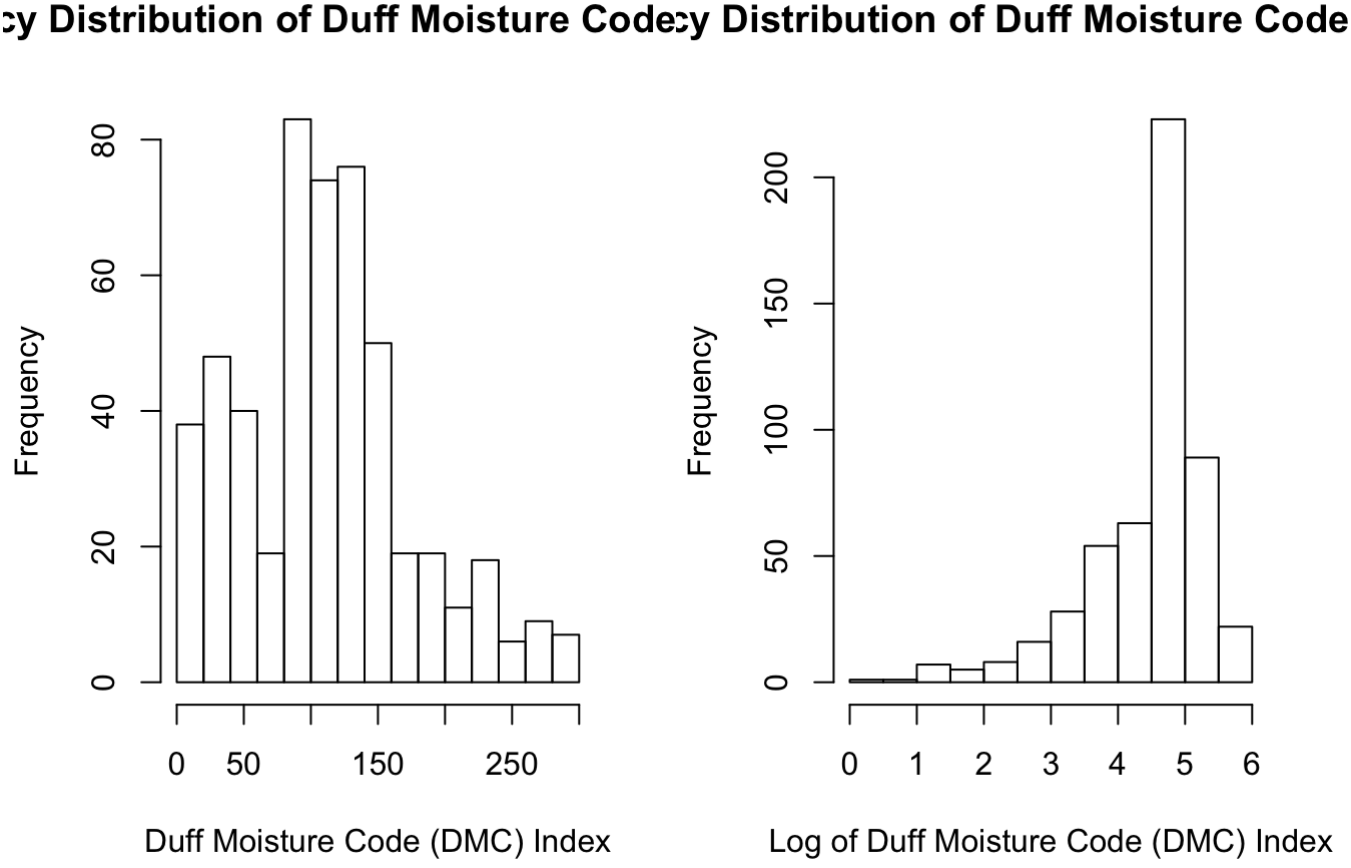


Interpretation

There are outliers for all three variables, with ISI containing the most outliers and DC with the fewest outliers. Wind Speed and Initial Spread Index plots show outliers above their Q3. This implies that the values of the outliers are higher than the 1.5*IQR higher than third quartile Q3. On the other hand, outliers of Drought Code (DC) Index plot are below Q1. This implies that the values of the two outliers are lower than the 1.5*IQR below the first/lower quartile.

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))
hist(x=df_f$DMC,
     main = "Frequency Distribution of Duff Moisture Code (DMC) Index",
     xlab = "Duff Moisture Code (DMC) Index")
hist(log(df_f$DMC),
     main = "Frequency Distribution of Duff Moisture Code (DMC) Index",
     xlab = "Log of Duff Moisture Code (DMC) Index")
```



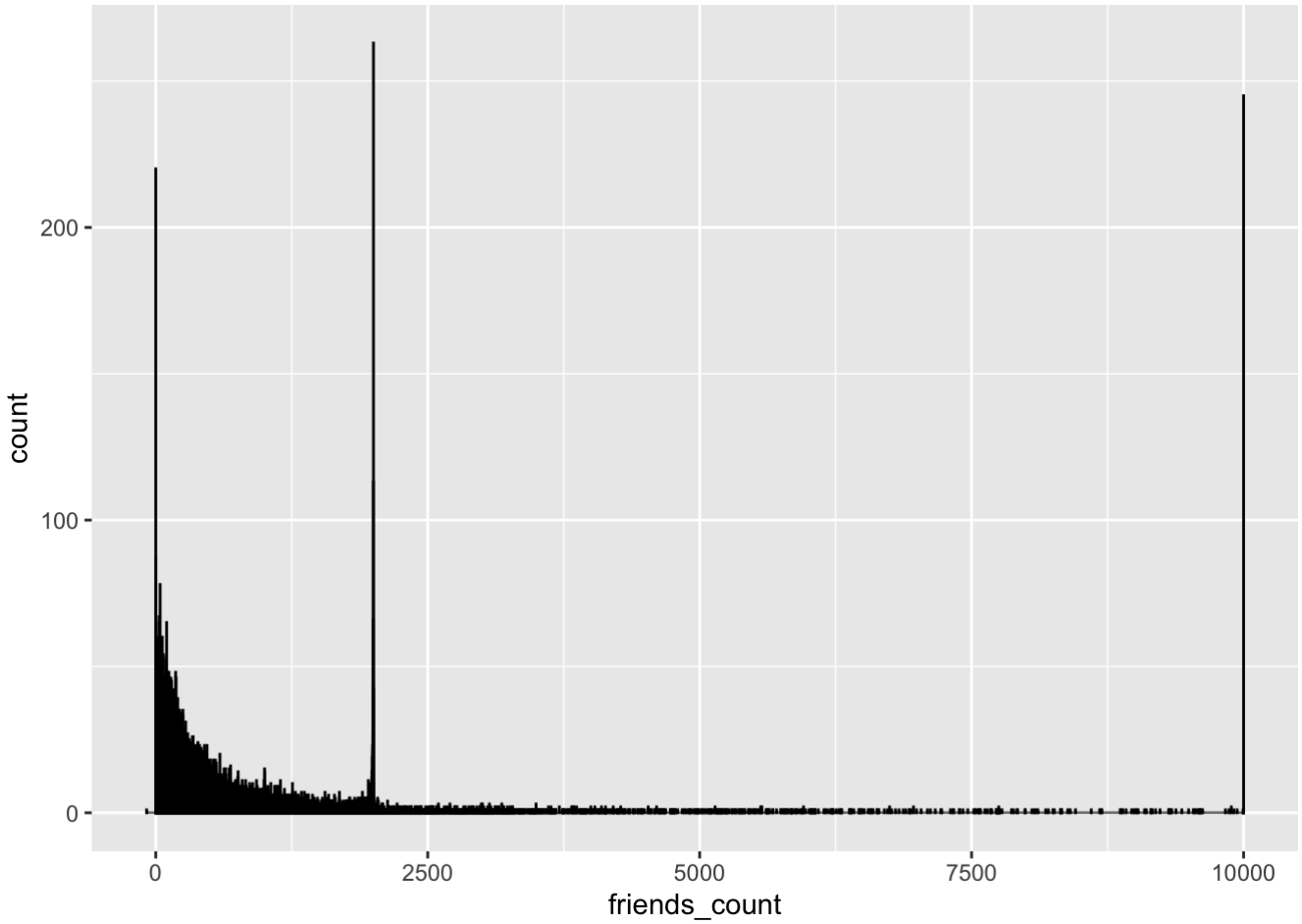
Compare the results and explain

The histogram of DMC Frequency Distribution is right-skewed with values mostly loacated between 0 to 150 on the x-axis. The mean and median are difficult to estimate from visually observing the plot. On the other hand, the histogram of log of DMC Frequency Distribution is left-skewed showing the long tail on the left. Without performing any calculation, mean and median can be visually estimated after the log transformation.

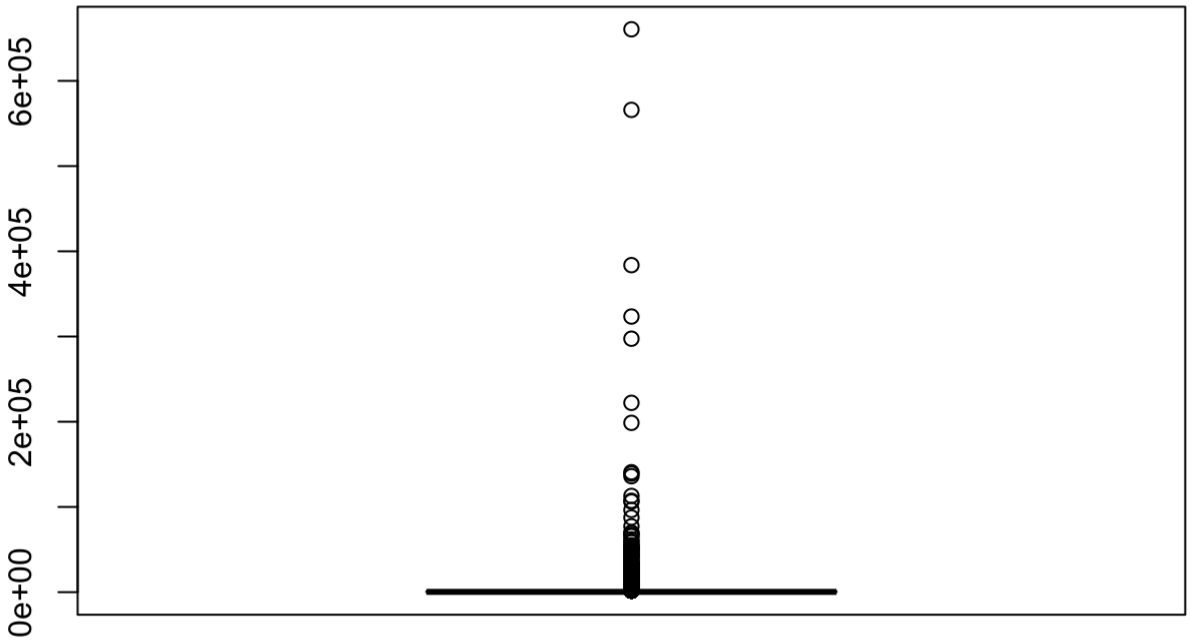
Problem 2 (Tweeter Accounts)

a. How are the data distributed for friend_count variable?

```
#install.packages("tidyverse")
library(tidyverse)
library(ggplot2)
twitter_data <- read_csv(file="/Users/mc.hung9298/Desktop/IE7275/M01_quasi_twitter.csv")
#twitter_data
#par(mfrow=c(1,2))
twitter_data %>%
  mutate(friends_count = ifelse(friends_count > 10000, 10000, friends_count)) %>%
  ggplot(aes(friends_count)) +
  geom_histogram(binwidth = .1, col = "black", fill = "cornflowerblue")
```




```
boxplot(twitter_data$friends_count)
```




```
#barplot(twitter_data$friends_count)
#hist(twitter_data$friends_count)
#plot(twitter_data$friends_count)
#summary(twitter_data$friends_count)
```

The values of “Friend_count” are right-skewed as we can see from the distributed graph of the variable. Also, there are a few outliers with extremly high values. Most of the values are quite similar to each other and have small values.

b. Compute the summary statitstics (min, 1Q, mean, median, 3Q, max) on friend_count

```
non_negative <- twitter_data$friends_count
non_negative[non_negative < 0] <- NA

#Computing Summary statistics
#min(twitter_data$friends_count, na.rm = T)
#quantile(twitter_data$friends_count, 0.25, na.rm = T)
#mean(twitter_data$friends_count, na.rm = T)
#quantile(twitter_data$friends_count, 0.75, na.rm = T)
#max(twitter_data$friends_count, na.rm = T)

summary(twitter_data$friends_count)
```

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

```
summary(non_negative)
```

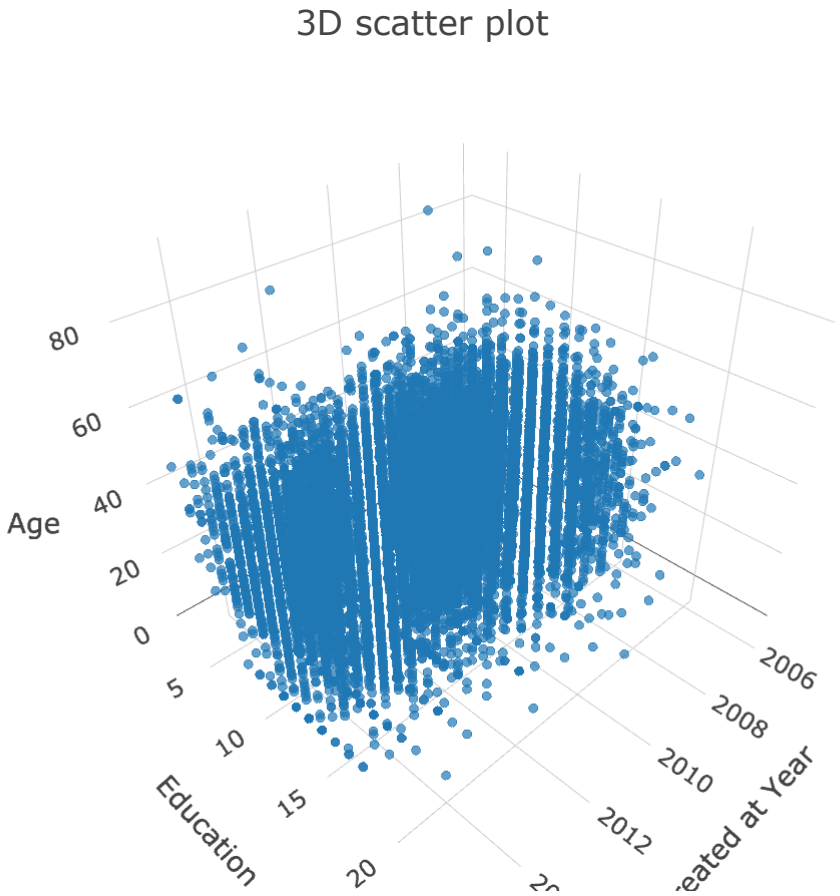
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     123     324    1058     849  660549       1
```

c. How are the data quality in friend_count variable?

The data is heavily skewed towards right showing that it has a lot of bias. There are many outliers. The variable is 'friends count' which cannot be in negative; but here we found a negative value that had to be removed.

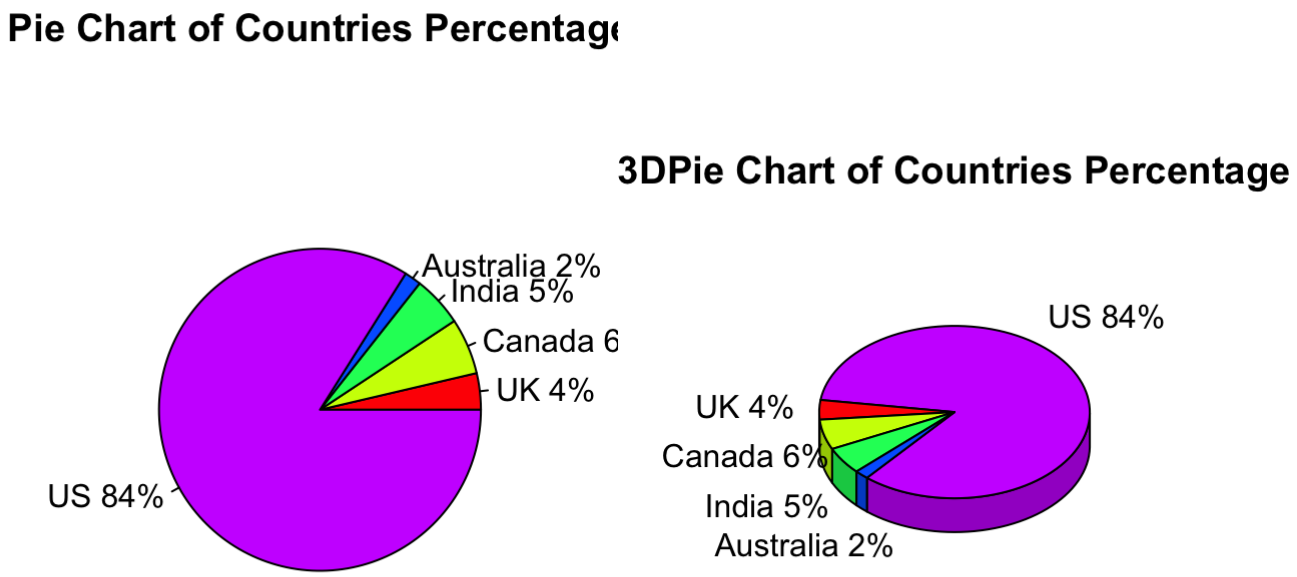
d. Product a 3D scatter plot with highlighting to impression the depth for variables below on dataset. created_at_year, education, age.

```
#install.packages("plotly")
library(plotly)


#colors <- c('#4AC6B7', '#1972A4', '#965F8A', '#FF7070', '#C61951')
plot_ly(data = twitter_data, x=twitter_data$created_at_year, y=twitter_data$education, z
=twitter_data$age, type="scatter3d", mode="markers",
        name ="3D scatter plot", size = 2)  %>%
  layout(title = '3D scatter plot',
scene = list(xaxis = list(title = 'Created at Year'), yaxis = list(title = 'Education'),
zaxis = list(title = 'Age')))
```



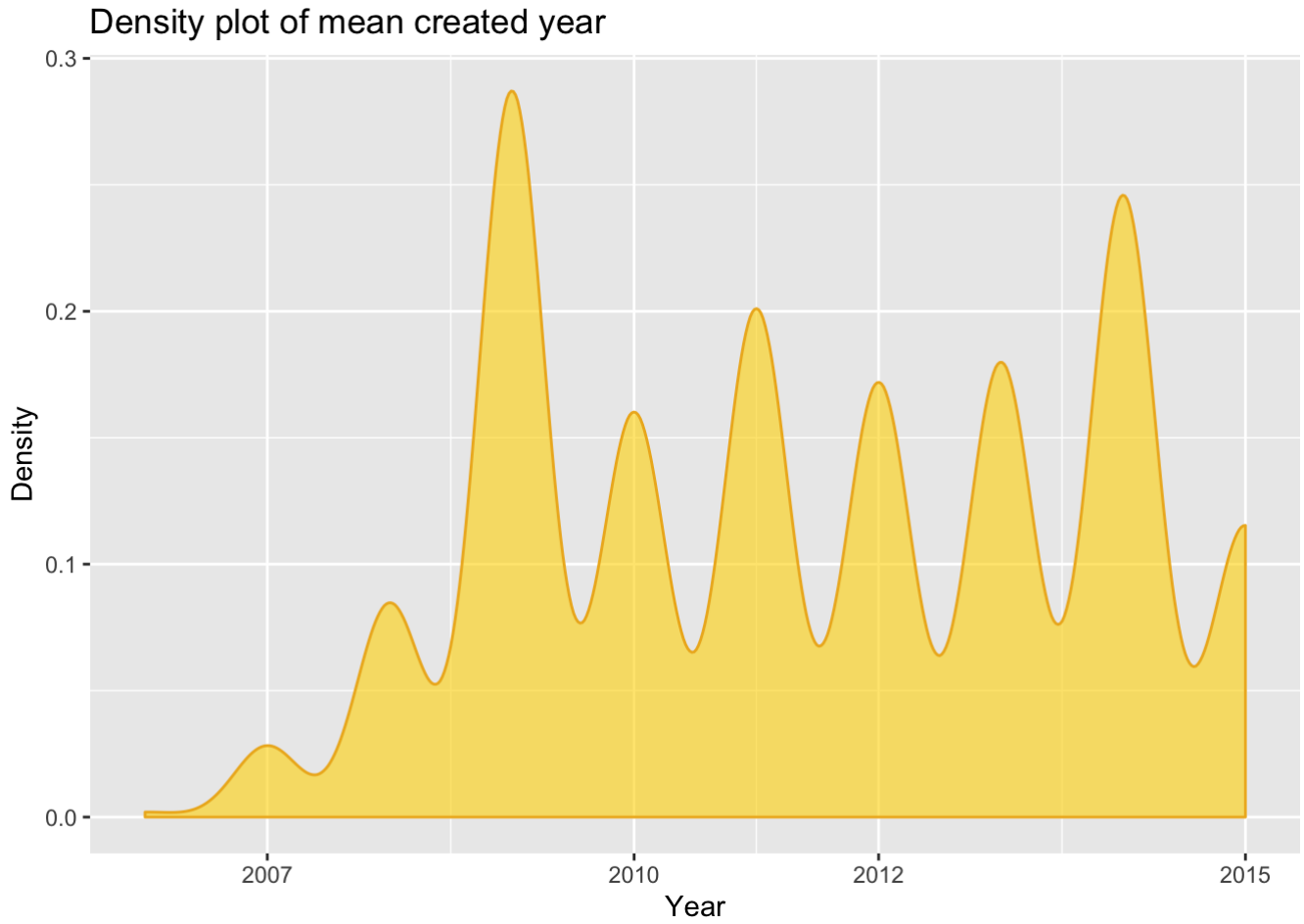
e. Consider 650, 1000, 900, 300 and 14900 tweeter acounts are in Uk, Canada, India, Australia and US respectively. Plot the percentage Pie chart includes percentage amount and country name adjacent to it, and also plot 3D pie chart for those countries along with the percentage pie chart.

```
#install.packages("plotrix")
library(plotrix)
par(mfrow=c(1, 2))
account <- c(650, 1000,900,300,14900)
country <- c("UK", "Canada","India", "Australia", "US")
#pie(account, labels = country, main="Simple Pie Chart")
pct <- round(account/sum(account)*100)
lbls2 <- paste(country, " ", pct, "%", sep="")
pie(account, labels=lbls2, col=rainbow(length(lbls2)), main="Pie Chart of Countries Perc
entage")
a<-c(1:5)
a[1]<- 3.05
a[2]<- 3.25
a[3]<- 3.80
a[4]<- 3.75
a[5]<- 7.05
pie3D(account,labels = lbls2, labelcex=1, explode = 0,
      main = "3DPie Chart of Countries Percentage",theta=1,radius = 0.8,start=3,labelpos
= a)
```



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

```
fill <- "gold1"
line <- "goldenrod2"
twitter_data$created_at_year <- as.numeric(twitter_data$created_at_year)
kernel_plot <- ggplot(twitter_data,
                      aes(x = twitter_data$created_at_year)) +
  geom_density(fill = fill,
               colour = line,
               alpha = 0.6) +
  scale_x_continuous(name = "Year", breaks =c(2007,2010,2012,2015)) +
  scale_y_continuous(name = "Density") +
  ggtitle("Density plot of mean created year")
kernel_plot
```




```
summary(twitter_data$created_at_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2006    2009    2011    2011    2013    2015
```

Interpretation

The desity of each year is the percentage of the amount of account created in that year in comparison to all years. The peak number of account is in 2009.

Problem 3 (Insurance Claims)

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

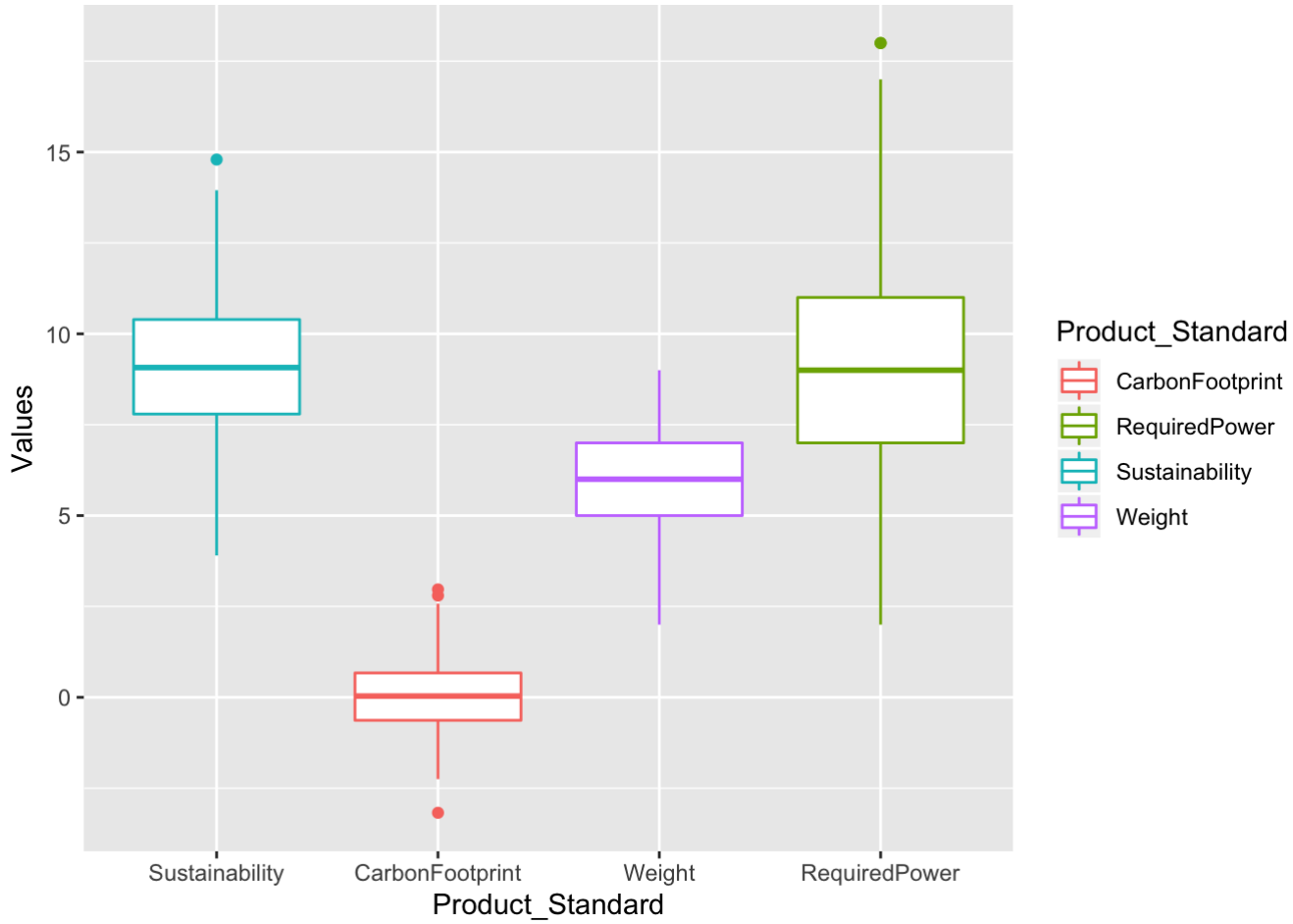
```
library(tidyverse)
library(dplyr)
#upload the dataset
raw_data <- read_csv(file="/Users/mc.hung9298/Desktop/IE7275/raw_data.csv")
names(raw_data) <- c("Sustainability","CarbonFootprint","Weight","RequiredPower")

#tide the dataset
raw_data2<- raw_data %>% gather("Sustainability","CarbonFootprint","Weight","RequiredPow
er",key="Product_Standard",value="Values")
#raw_data2
```

```
#get the normalized dataset
NNN <- lapply(raw_data, function(x) round((x-min(x))/(max(x)-min(x)), 1))
Ndata <- data.frame(NNN)

#tidy the dataset
Ndata1 <- Ndata %>% gather("Sustainability","CarbonFootprint","Weight","RequiredPower",k
ey="Product_Standard",value="Values")
head(Ndata1, n=20)
```

```
##    Product_Standard Values
## 1    Sustainability    0.4
## 2    Sustainability    0.6
## 3    Sustainability    0.4
## 4    Sustainability    0.2
## 5    Sustainability    0.5
## 6    Sustainability    0.5
## 7    Sustainability    0.7
## 8    Sustainability    0.5
## 9    Sustainability    0.5
## 10   Sustainability    0.6
## 11   Sustainability    0.5
## 12   Sustainability    0.7
## 13   Sustainability    0.7
## 14   Sustainability    0.3
## 15   Sustainability    0.5
## 16   Sustainability    0.3
## 17   Sustainability    0.5
## 18   Sustainability    0.8
## 19   Sustainability    0.6
## 20   Sustainability    0.3
```

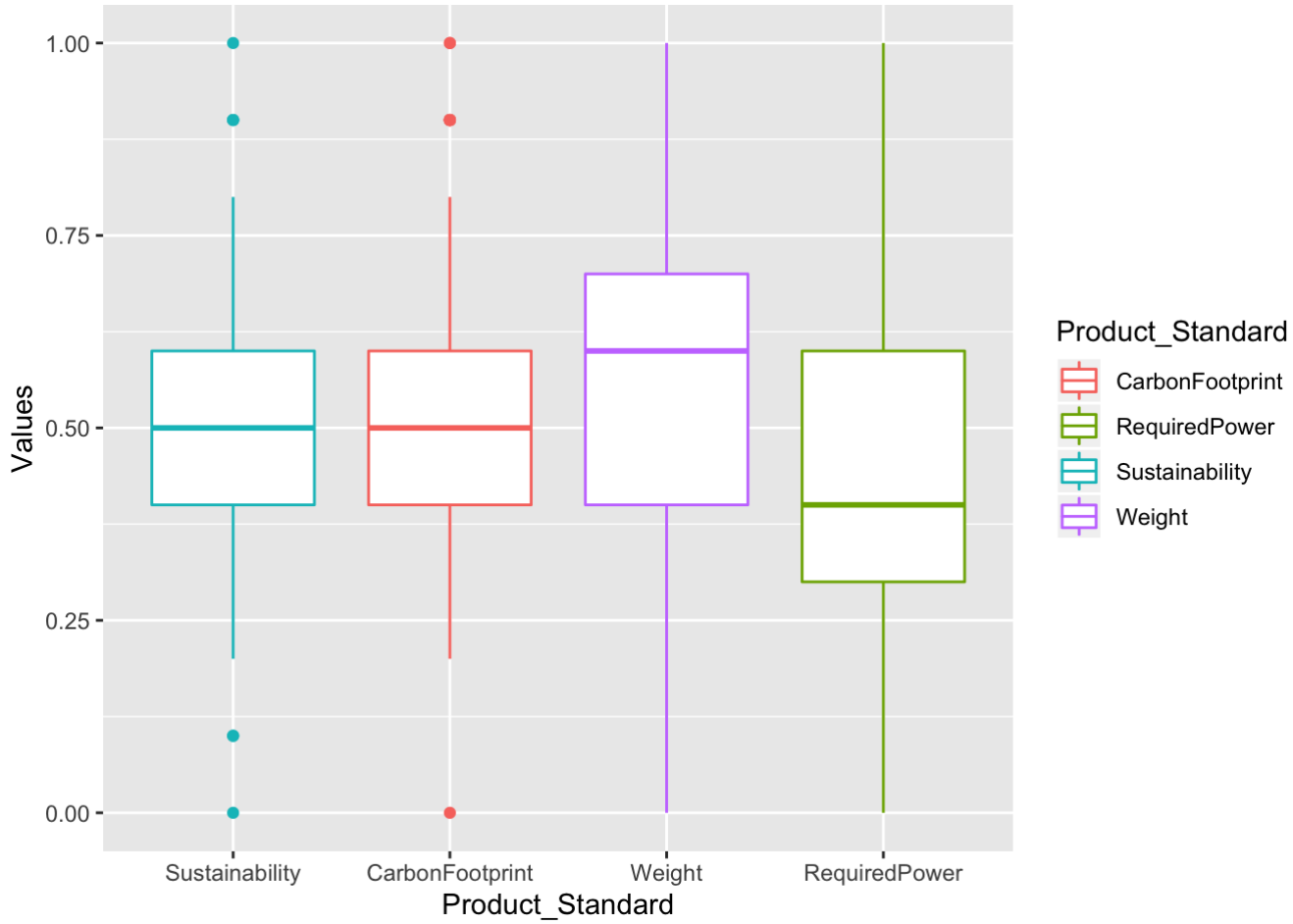
b. Create the boxplot from the original dataset

```
raw_data2 %>% ggplot()+
  geom_boxplot(mapping = aes(x=Product_Standard,y=Values,color=Product_Standard))+
  scale_fill_discrete()+
  scale_x_discrete(limits=c("Sustainability","CarbonFootprint","Weight","RequiredPower"
))
```



c. Create the boxplot from the normalized dataset

```
Ndata1 %>% ggplot()+
  geom_boxplot(mapping = aes(x=Product_Standard,y=Values,color=Product_Standard))+
  scale_fill_discrete()+
  scale_x_discrete(limits=c("Sustainability","CarbonFootprint","Weight","RequiredPower"
))
```



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

The first graph from the orginal dataset shows that the median values for four product standards are different. The sustainability and the required power have the similiar median, and the median valus is zero for the carbon footprint.The second graph from the normalized dataset shows that the normalized median values for sustainability and the carbon print are same, around 0.5, with outliers.

e. Scatter plot of A and B. How correlated the data are in these variables.

```
#install.packages("ggpubr")
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

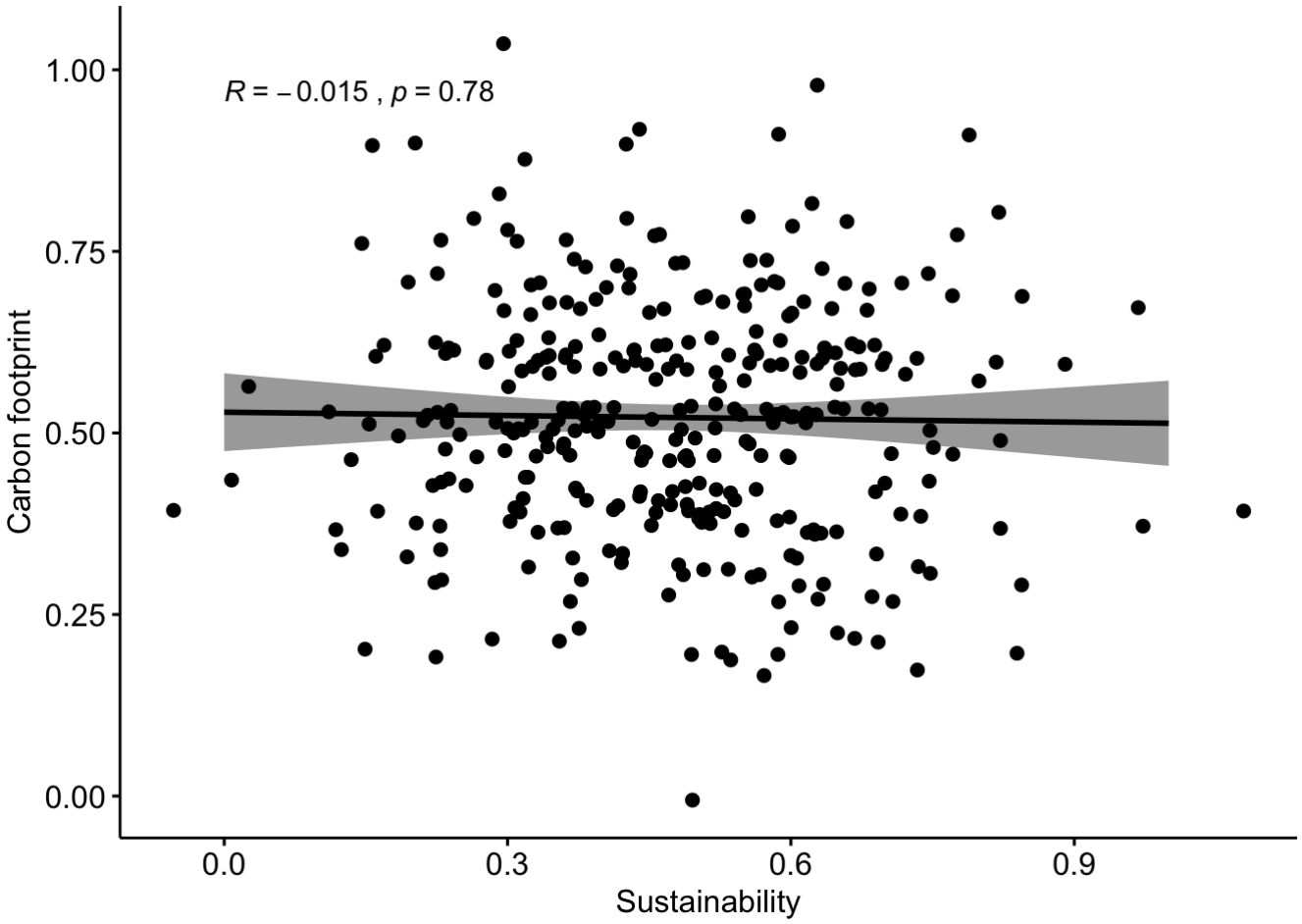
```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

```
ggscatter(Ndata, x = "Sustainability", y = "CarbonFootprint",
          add = "reg.line", conf.int = TRUE,
          position = position_jitter(0.1),
          cor.coef = TRUE,
          cor.method = "pearson",
          xlab = "Sustainability",
          ylab = "Carbon footprint")
```



Interpretation

The calucalted correlation coefficient (R) of Sustainability and Carbon Footprint is -0.015, as shown on the graph. The R value of approximately zero indicates that the points have no direction. This can be seen from the round shape of the cluster. In fact, the plotted line does not fit to the points on the graph. In addition, it can be noted that the calculated p-value is high. This indicates a weak correlations as the probability that the two variables have no relationship is very high.