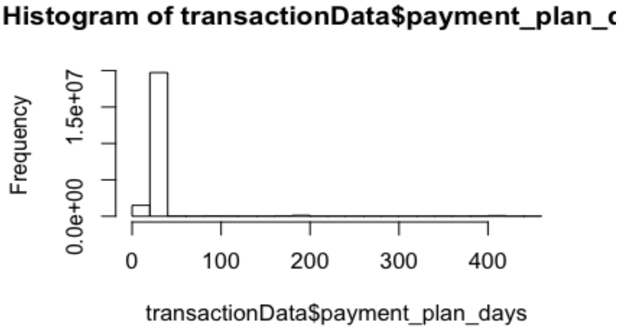
DATA ANALYTICS WITH R

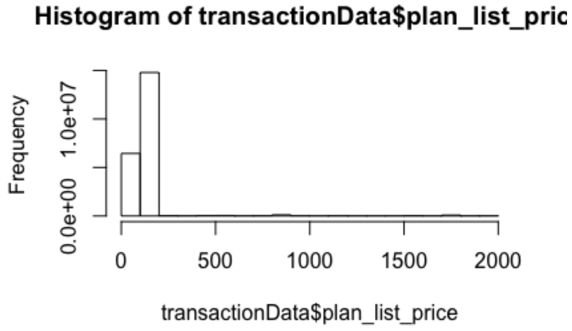
HOMEWORK 3

JASNEEK SINGH CHUGH
JC2433

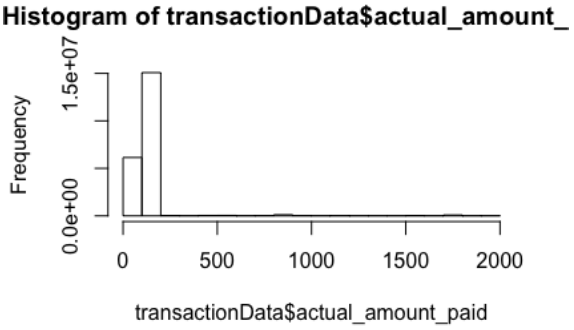
```
#Loading Required Packages
library(UsingR)
library(ggplot2)

#A

transactionData <- read.csv(file="/Users/jasneekchugh/Desktop/DS_NJIT/CS636-\ DA\ with\
R/Assignment/Assignment3/transactions.csv")

#1
#Numerical data: Data that takes numerical values and for which arithmetic operations
make sense
#In transaction data, numerical variables are: payment_plan_days, plan_list_price,
actual_amount_paid.
#Categorical data: Qualitative data that is limited to number of distinct categories.
#In transaction data, Categorical variables are: payment_method_id , is_auto_renew(0,1),
is_cancel(0,1)

 #2
 hist(transactionData$payment_plan_days)
```



```
 hist(transactionData$plan_list_price)
```



```
 hist(transactionData$actual_amount_paid)
```



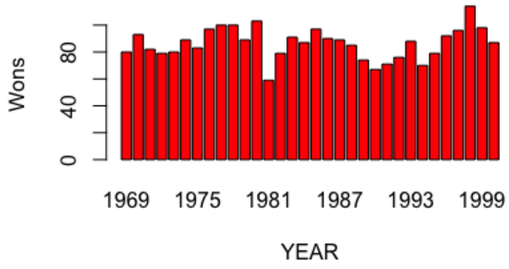
```
#3
> table(transactionData$payment_method_id)

     1       2       3       4       5       6       7       8      10 
    12      52     210      15     474     466    1094     657    1326 
    11      12      13      14      15      16      17      18      19 
  2129    3834    6571   13621    1479   11064    7437   16177   32073 
    20      21      22      23      24      25      26      27      28 
 28278   22883   20130   42386   16196   13780    4591   62525   95733 
    29      30      31      32      33      34      35      36      37 
113885  160957  252342  146481  411164  731539  541399  855115 1007689 
    38      39      40      41 
1703590 1466655 2225283 11526454 

>  table(transactionData$is_auto_renew)

     0       1 
3189796 18357950 

>  table(transactionData$is_cancel)

       0        1 
20690895   856851 

> #B
> #2.4
> weather1<- table(central.park$WX)
> weather1

 1 18 
10 11 
> weather2<- table(central.park$WX, exclude = FALSE)
> weather2

   1   18 <NA> 
  10   11   10 
```

Table 2 (weather2) is better because it shows the number of NAs as well present in the "WX" column. So it gives more clarity of the data.

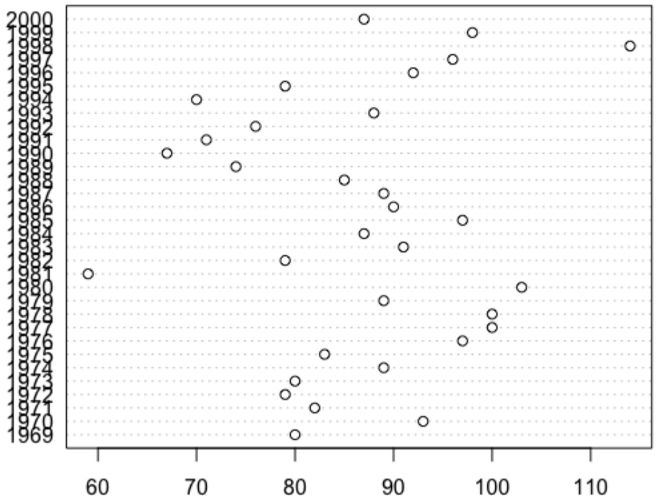
```
> #2.8
> attach(npdb)
> head(npdb)
> state<- sort(table(npdb$state))
> state[which.max(state)]

  CA
1566
> detach(npdb)
>   #CA has the maximum malpractice awards i.e. 1566
```

```
#2.9
table(ID)
sort(table(ID))
#the command table(ID) returns the frequency of each element in column ID. It's is interesting because we can quickly see if there are any duplicates values in the column ID. Since there should be only one ID for each doctor.
```

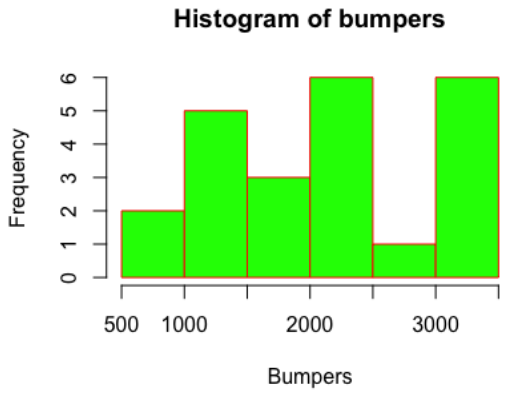
```
> #2.10
> data(MLBattend)
> attach(MLBattend)
> win =  wins[franchise == "NYA"]
> names(win) = c(1969:2000)
> detach(MLBattend)
> barplot(win,xlab = "YEAR", ylab = "Wons", col="red")
```



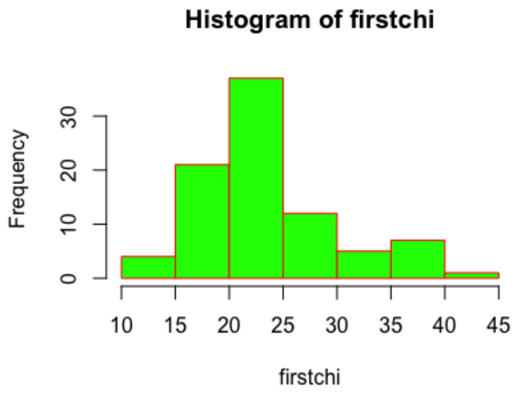
```
> dotchart(win)
```



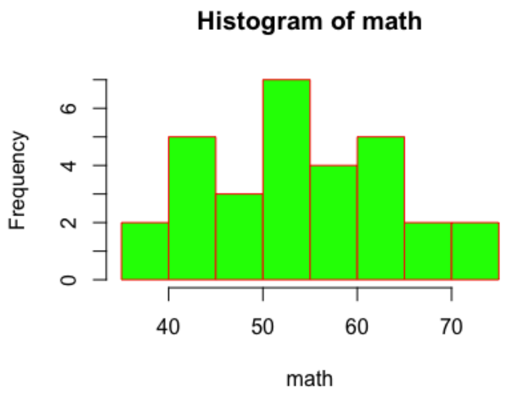
```
> #2.16
>  #1) It's a proportion so we divide it by the total number of elements in rivers dataset
> length(rivers[rivers < 500])/length(rivers)
[1] 0.5815603
>  #2)
> length(rivers[rivers < mean(rivers)])/length(rivers)
[1] 0.6666667
>  #3) We find quartile using the summary function. 0.75 is the third quartile
> rivers_summary<-summary(rivers)
> rivers_summary[5]
3rd Qu.
   680

> #2.23
> mean(npdb$amount)
[1] 166257.2
> median(npdb$amount)
[1] 37500
> quantile(npdb$amount)
      0%      25%      50%      75%     100%
      50     8750    37500   175000 25000000
>  # So from output of quantile(npdb$amount) we can say that percentile of mean lies
between 50th and 75th percentile approx. 74.85 percentile.
>  #It might be because the data is right skewed and Mean is greater than Median.

> #2.30
> hist(bumpers,xlab = "Bumpers",col = "green",border = "red")
```



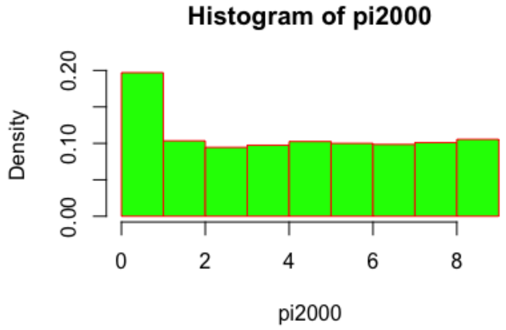
```
> mean(bumpers)
[1] 2122.478
> median(bumpers)
[1] 2129
> sd(bumpers)
[1] 798.4574
> hist(firstchi,xlab = "firstchi",col = "green",border = "red")
```



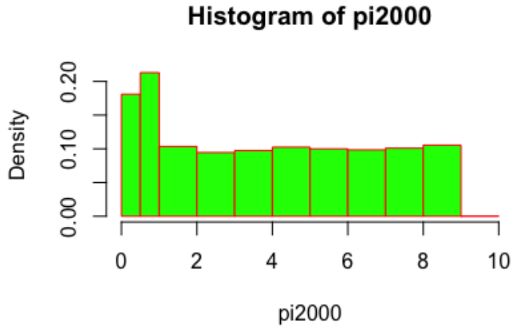
```
> mean(firstchi)
[1] 23.97701
> median(firstchi)
[1] 23
> sd(firstchi)
[1] 6.254258
> hist(math,xlab = "math",col = "green",border = "red")
```



```
> mean(math)
[1] 54.9
> median(math)
[1] 54
> sd(math)
[1] 9.746264

> #2.32
> hist(pi2000, prob=TRUE, col = "green",border = "red")
```

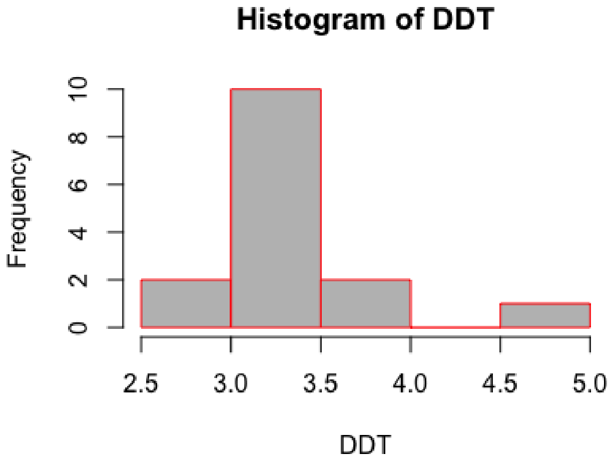


```
> #The above distribution looks flat as all digits are equal So we subtract 0.1. So that the
bins for 0 and 1 do not gets combined, by using the argument breaks=0:10-.5.
> hist(pi2000,breaks=c(0:10,.5), prob=TRUE, col = "green",border = "red")
```

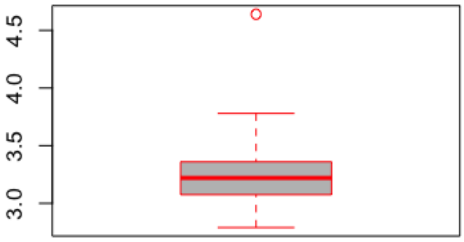


```
> #2.34
> DDT
 [1] 2.79 2.93 3.22 3.78 3.22 3.38 3.18 3.33 3.34 3.06 3.07 3.56 3.08 4.64 3.34
```

```
> histogram(DDT, col = "grey",border = "red")
```

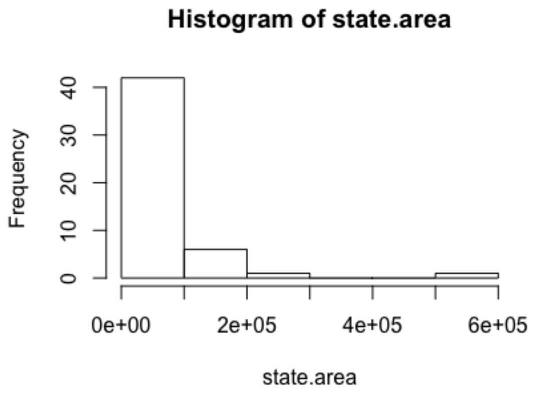


```
> boxplot(DDT, col = "grey",border = "red") #the small circle point highlights the outlier in the data
```

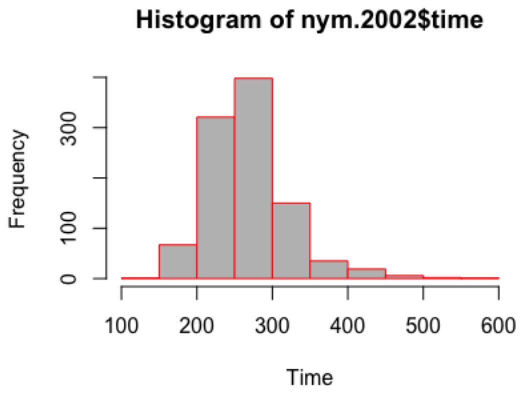


```
> mean(DDT)
[1] 3.328
> sd(DDT)
[1] 0.4371531

> #2.35
> names(state.area) <-state.abb
> length(state.area)
[1] 50
> length(state.area[state.area < state.area['NJ']])/50*100 #percentage of states with area less than NJ
[1] 8
> length(state.area[state.area < state.area['NY']])/50*100 #percentage of states with area less than NY
[1] 40
```

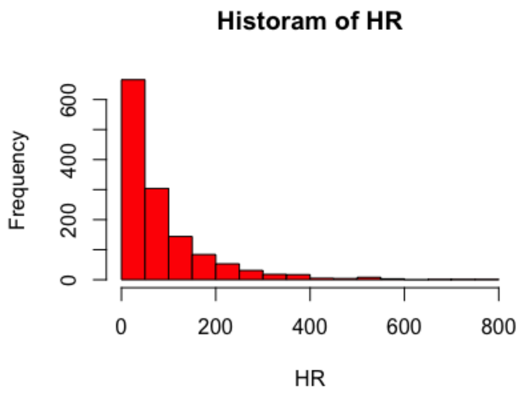
```
> hist(state.area)
```



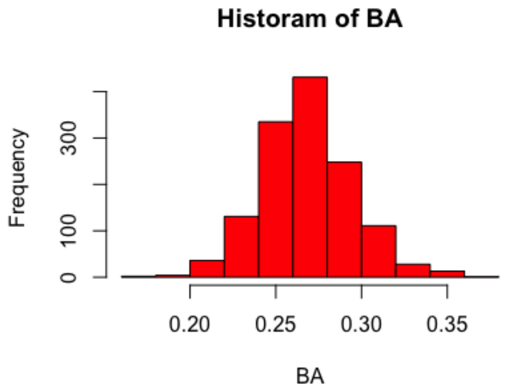
```
> state.area[state.area > 5e5]#Outlier
   AK
589757

> #2.36
> hist(nym.2002$time, xlab="Time",col = "grey",border = "red")
```



```
> mean(nym.2002$time)
[1] 268.5707
> median(nym.2002$time)
[1] 262.8417
#Mean is greater than median therefore, the data is right skewed and the tail of the
histogram is towards right. The bulk data is represented in the middle by median.

> #2.39
> #Histograms are very useful to understand the pattern of variability in the data.
> hist(hall.fame$HR,xlab="HR", main="Historam of HR",col = "red") #Shape: Right Skewed
```

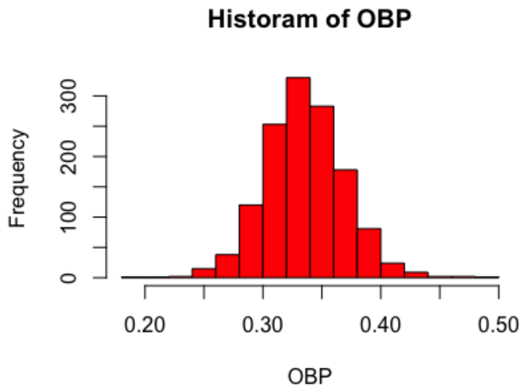


```
> mean(hall.fame$HR)
[1] 85.1097
> median(hall.fame$HR)
[1] 51
> hist(hall.fame$BA,xlab="BA", main="Historam of BA",col = "red")#Shape: Symmetric
```



(Mean and Median are almost equal), Smooth curve

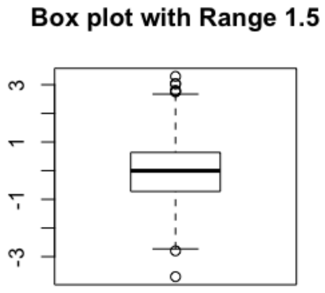
```
> mean(hall.fame$BA)
[1] 0.2687739
> median(hall.fame$BA)
[1] 0.267
> hist(hall.fame$OBP,xlab="OBP", main="Historam of OBP",col = "red") #Shape: Symmetric
```

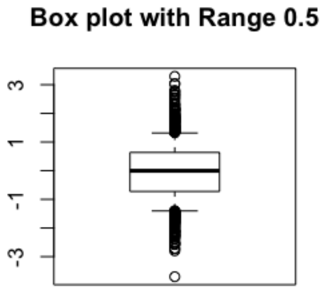


(Mean and Median are almost equal), Smooth curve

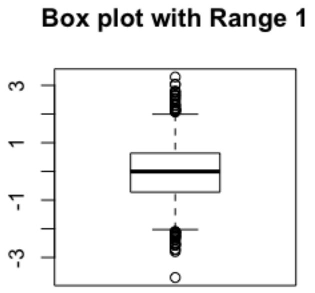
```
> mean(hall.fame$OBP)
```

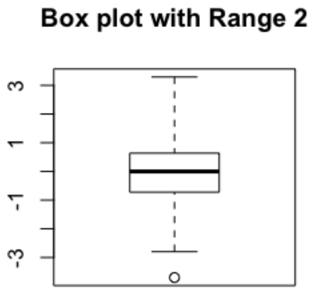
```
[1] 0.3360642
> median(hall.fame$OBP)
[1] 0.335

> #2.41
> x=rnorm(1000)
> boxplot(x,range=1.5,main="Box plot with Range 1.5")
> boxplot(x,range=0.5,main="Box plot with Range 0.5")
```



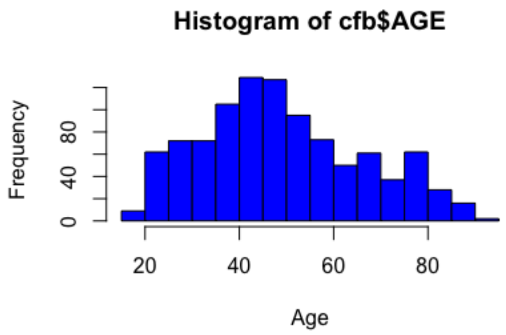


```
> boxplot(x,range=1,main="Box plot with Range 1")
> boxplot(x,range=2,main="Box plot with Range 2")
```

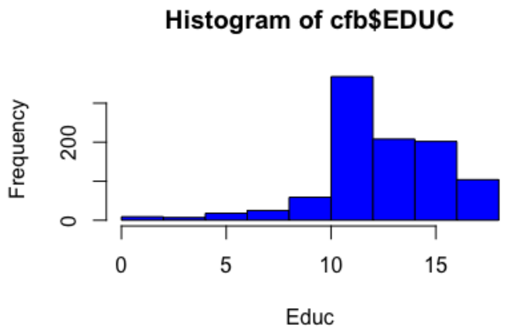




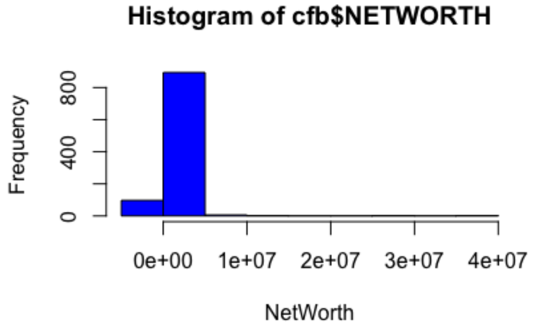
```
>  #Boxplots whiskers are chosen with factor of 1.5 because, the width of the notches is
proportional to the IQR and inversely proportion to square root of the size of the sample.
The box plot has equal whisker length of 1.5IQR for both whiskers.
>  #Upper whisker- Above third quartile: Q3+1.5*IQR
>  #Lower whisker- Below first quartile: Q1-1.5*IQR

> #2.42
> Mode <- function(x) {
+   uq <- unique(x)
+   uq[which.max(tabulate(match(x, uq)))]
+ }
>   #Age
> hist(cfb$AGE,xlab="Age", col="blue")
```



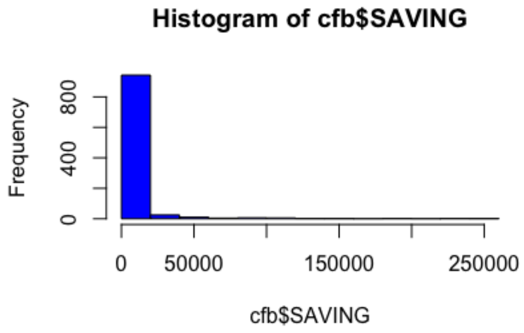
```
>
> Mode(cfb$AGE) #We can also see from the histogram that the bar of the histogram is at
the peak.
[1] 40
> mean(cfb$AGE)
[1] 49.635
> median(cfb$AGE)
[1] 48
>  #For Age, Mean is greater than the Median. Therefore, the data is skewed to the right.
Tail
>  #is also towards the right of the histogram.
>
>  #EDUC
> hist(cfb$EDUC,xlab="Educ",col="blue")
```



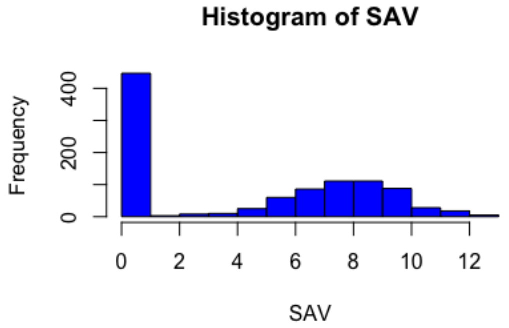
```
> Mode(cfb$EDUC)
[1] 12
> mean(cfb$EDUC)
[1] 13.075
> median(cfb$EDUC)
[1] 13
>  #For EdDUC, Mean and Median are almost the same. However, as per the histogram the
data is skewed left. Tail of the histogram is on the left side.
>
>  #NETWORTH
> as.numeric(cfb$NETWORTH)
> hist(cfb$NETWORTH,xlab="NetWorth",col="blue")
```



```
> mean(cfb$NETWORTH)
[1] 376993.4
> median(cfb$NETWORTH)
[1] 89947.5
> Mode(cfb$NETWORTH)
[1] 0
> #we can see that the data is so much skewed to the right. So we can transform the NETWORTH using the logarithmic function to improve the skewness.
>
>  #SAVINGS
> as.numeric(cfb$SAVING)
> hist(cfb$SAVING,col="blue")
```



```
> SAV<-log(cfb$SAVING+1)
> hist(SAV,col="blue")
```

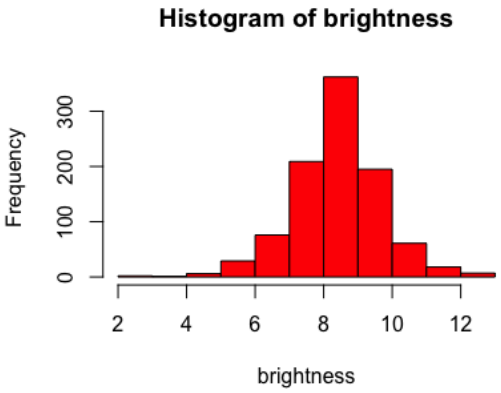


```
> mean(SAV)
[1] 4.24621
> median(SAV)
```

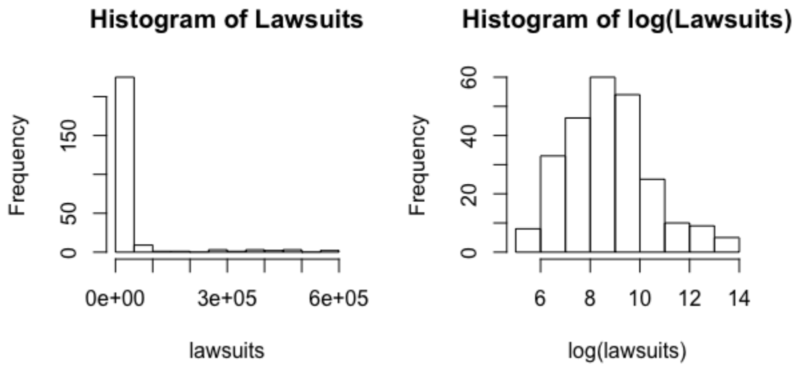
```
[1] 5.303305
> Mode(SAV)#most of the people have savings of Zero and that is why mode is zero.
[1] 0
>  #For SAVINGS, median is greater than mean, therefore data is skewed to left.
>  #Tail of the histogram is also towards left.
```

I was able convince myself for the charts they have for all the 4 variables. However, for "EDUC" variable after looking at the mean and median almost close, I was expecting the histogram to be nearly normally distributed however, the histogram is left skewed.

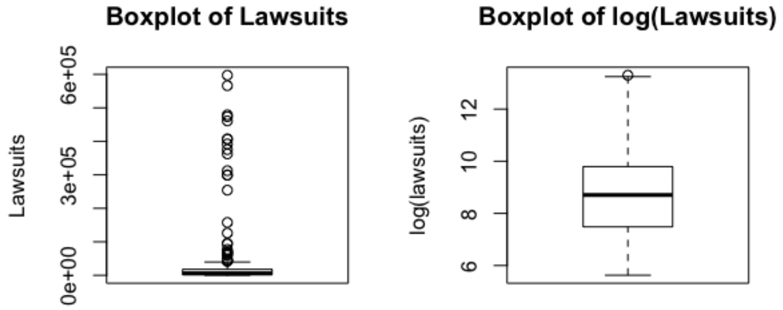
```
> #2.43
> summary(brightness)#it returns the useful summary like Min, Max, Median, of each column.
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.070   7.702   8.500   8.418   9.130  12.430 
> hist(brightness,col = "red")
```



```
> mean(brightness)
[1] 8.417743
> median(brightness)
[1] 8.5
> max(brightness)-min(brightness) #Range
[1] 10.36

> #2.44
> par(mfrow=c(1,2))
> hist(lawsuits,main="Histogram of Lawsuits")
> hist(log(lawsuits),main="Histogram of log(Lawsuits)")
```

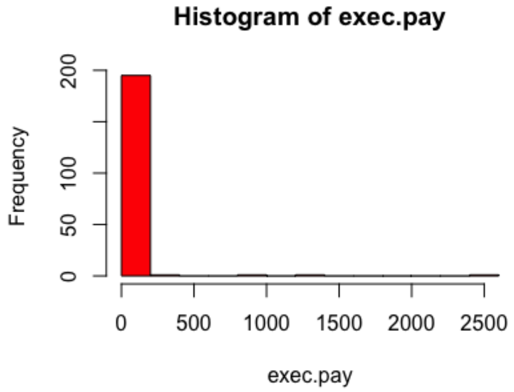


```
> boxplot(lawsuits,ylab="Lawsuits",main="Boxplot of Lawsuits")
> boxplot(log(lawsuits),ylab="log(lawsuits)",main="Boxplot of log(Lawsuits)")
```

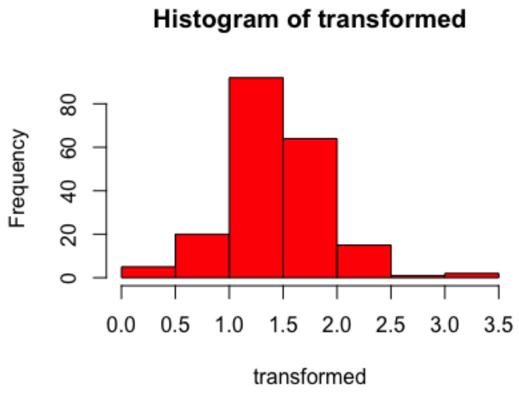


It would be bit difficult to guess where the middle 50% of the observations are, without the help of boxplot because the range of lawsuits is quite high. Hence, Boxplot helps us to see where actually the observations are. Also, after the logarithmic transformation it gives more clear picture of the data in the boxplot and show the outlier clearly.

```
> #2.45
> hist(exec.pay,col = "red")#the data is very much skewed to the right side
```

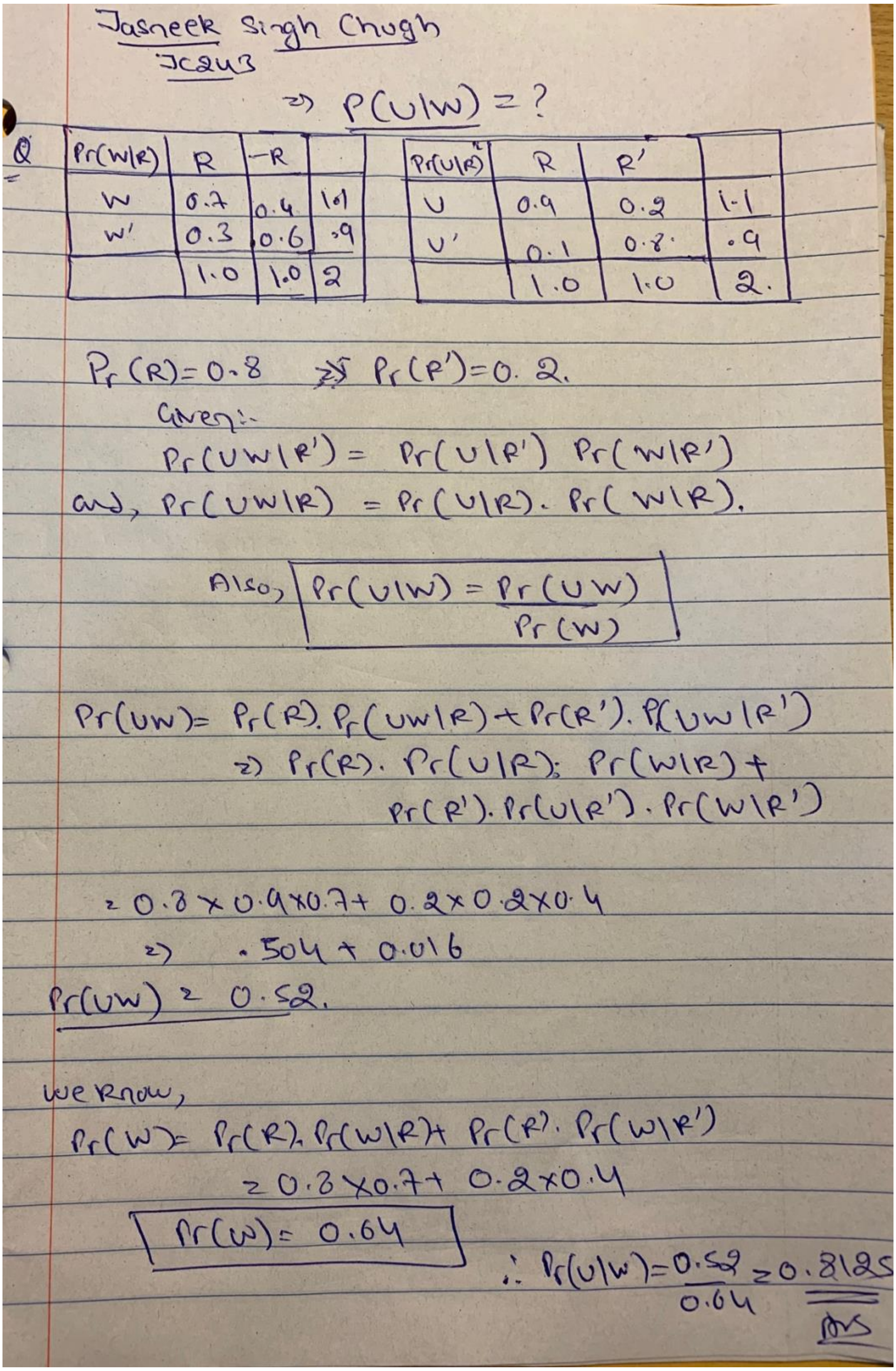


```
> mean(exec.pay)
[1] 59.88945
> median(exec.pay)
[1] 27
> transformed<-log(1+ exec.pay,10)
> hist(transformed = "red")
```



```
> mean(transformed)
[1] 1.438955
> median(transformed)
[1] 1.447158
> #after the transformation the data looks symmetric and the mean and the median of
> #the transformed data are also almost equal.
> #Histogram of transformed data is better because it looks symmetric and distributed
normally.
> #With transformed histogram we get to see more clear bars for each.
> #The mean and Median of transformed data are almost equal.
```

Probability Distribution



Jasneek Singh Chugh

JC243

$\Rightarrow P(U|W) = ?$

Q

Pr(W|R)	R	¬R	
W	0.7	0.4	1.1
W'	0.3	0.6	.9
	1.0	1.0	2

Pr(U|R)	R	R'	
U	0.9	0.2	1.1
U'	0.1	0.8	.9
	1.0	1.0	2.

$Pr(R) = 0.8 \Rightarrow Pr(R') = 0.2.$

Given:-

$Pr(UW|R') = Pr(U|R')\ Pr(W|R')$

and, $Pr(UW|R) = Pr(U|R).\ Pr(W|R).$

Also, $\boxed{Pr(U|W) = \frac{Pr(UW)}{Pr(W)}}$

$Pr(UW) = Pr(R).\ Pr(UW|R) + Pr(R').\ Pr(UW|R')$

$\Rightarrow Pr(R).\ Pr(U|R).\ Pr(W|R) + Pr(R').\ Pr(U|R').\ Pr(W|R')$

$= 0.8 \times 0.9 \times 0.7 + 0.2 \times 0.2 \times 0.4$

$\Rightarrow .504 + 0.016$

$Pr(UW) = 0.52.$

We know,

$Pr(W) = Pr(R).\ Pr(W|R) + Pr(R').\ Pr(W|R')$

$= 0.8 \times 0.7 + 0.2 \times 0.4$

$\boxed{Pr(W) = 0.64}$

$\therefore Pr(U|W) = \frac{0.52}{0.64} = 0.8125$

Ans