13. Normality and outliers treatment

Outliers treatment

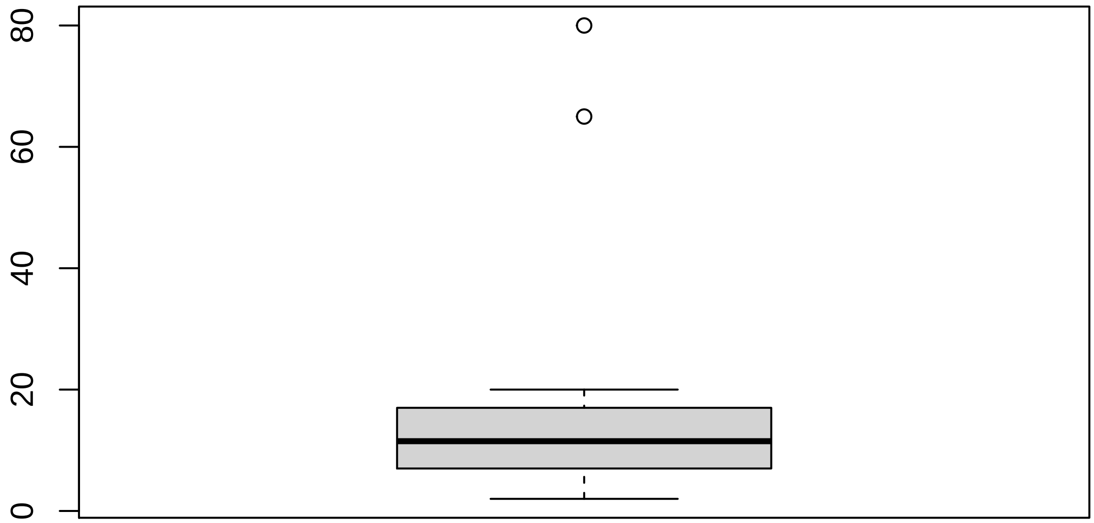
```
data=c(sample(x=1:20, size=40,replace = T),65,80)
data
```

```
##  [1]  8 19  6 19 19  4 15  6  9 17  5 15 14 14  7 15 12  8  2 11 20  8 16 15  2
## [26]  3 20  5  7  9  6 19 14  6 20 16 11 10  9 20 65 80
```

```
summary(data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    7.00   11.50   14.43   16.75   80.00
```

```
boxplot(data)
```



### Descarding outliers from the dataset

```r
length(data)
```

```
[1] 42
```

```r
quantile(data,c(0.75))
```

```
  75%
16.75
```

```r
bench=quantile(data,c(0.75))+1.5*IQR(data)
#bench=Q3+1.5*IQR(data) (upper value)
#bench=Q1-1.5*IQR(dat) (lower value)
bench
```

```
   75%
31.375
```

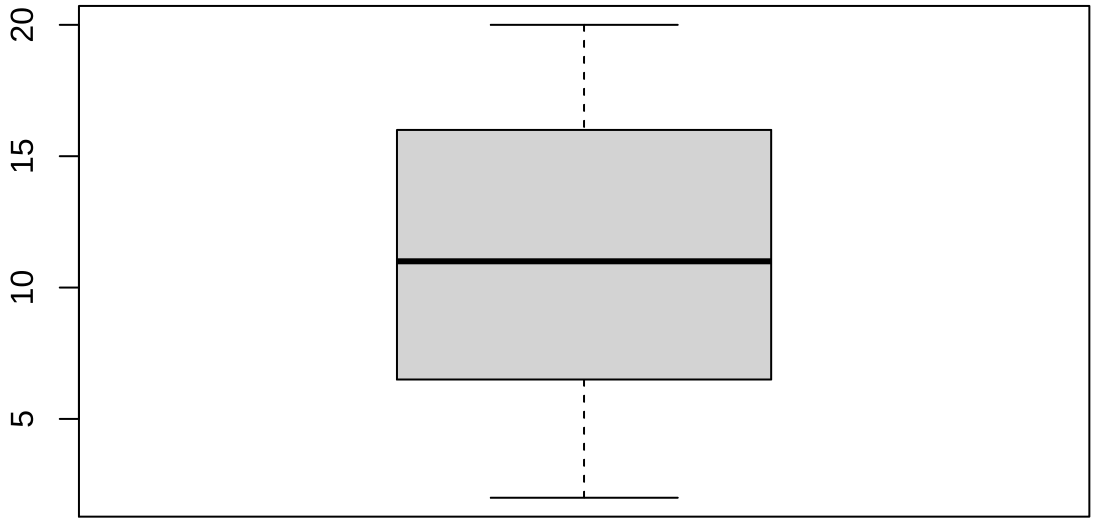
```r
data=data[data<bench]
data
```

```
 [1]  8 19  6 19 19  4 15  6  9 17  5 15 14 14  7 15 12  8  2 11 20  8 16 15  2
[26]  3 20  5  7  9  6 19 14  6 20 16 11 10  9 20
```

```r
summary(data)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    6.75   11.00   11.53   16.00   20.00
```

```r
boxplot(data)
```



```
length(data)
```

```
[1] 40
```

Open normality dataset

```
normality=read.csv(file.choose())
summary(normality)
```

```
     Gender           Day1             Day2             Day3       
 Min.   :1.000   Min.   : 0.020   Min.   :0.0000   Min.   :0.0200  
 1st Qu.:1.000   1st Qu.: 1.305   1st Qu.:0.4100   1st Qu.:0.4400  
 Median :2.000   Median : 1.790   Median :0.8200   Median :0.7600  
 Mean   :1.619   Mean   : 1.794   Mean   :0.9718   Mean   :0.9739  
 3rd Qu.:2.000   3rd Qu.: 2.232   3rd Qu.:1.3500   3rd Qu.:1.5250  
 Max.   :2.000   Max.   :20.000   Max.   :3.4400   Max.   :3.4100  
                                  NA's   :538      NA's   :677     
```

```
library(moments)
skewness(normality$Day1)
```

```
[1] 8.836643
```

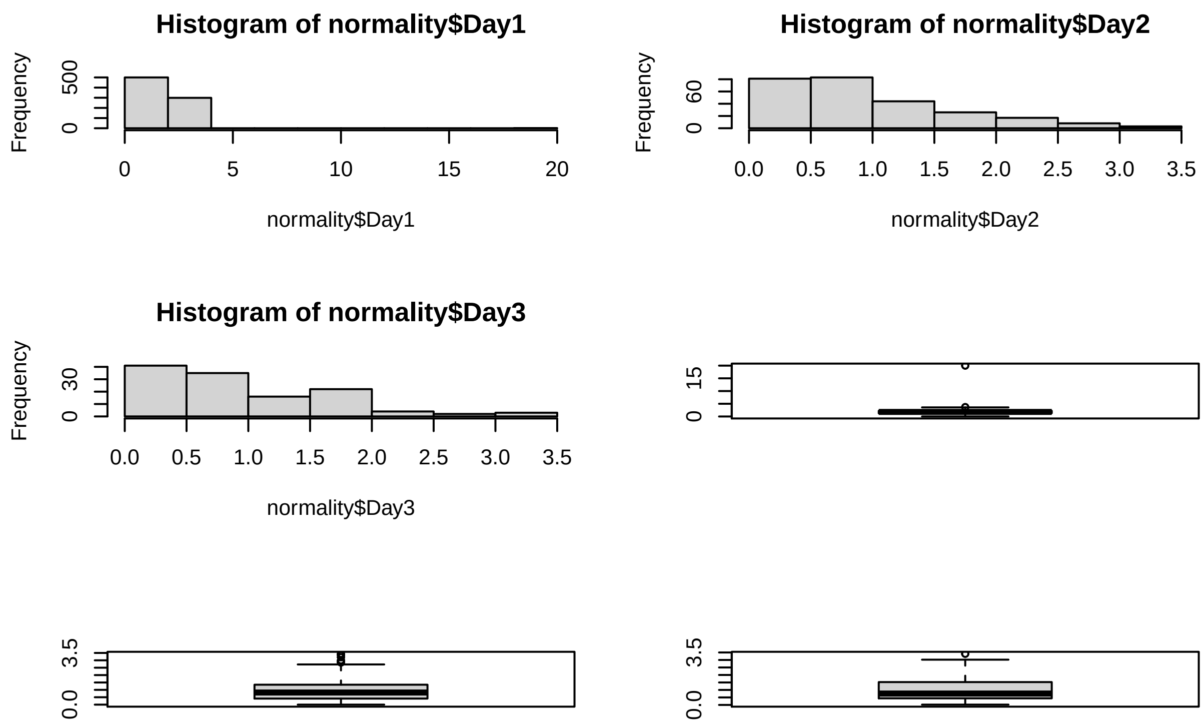
```r
skewness(normality$Day2,na.rm = T)
```

```
[1] 1.062469
```

```r
skewness(normality$Day3,na.rm = T)
```

```
[1] 1.017236
```

```r
par(mfrow=c(3,2))
hist(normality$Day1)
hist(normality$Day2)
hist(normality$Day3)
boxplot(normality$Day1)
boxplot(normality$Day2)
boxplot(normality$Day3)
```



```r
match(c(20),normality$Day1) #match(normality$Day1>4,normality$Day1)
```

```
[1] 672
```

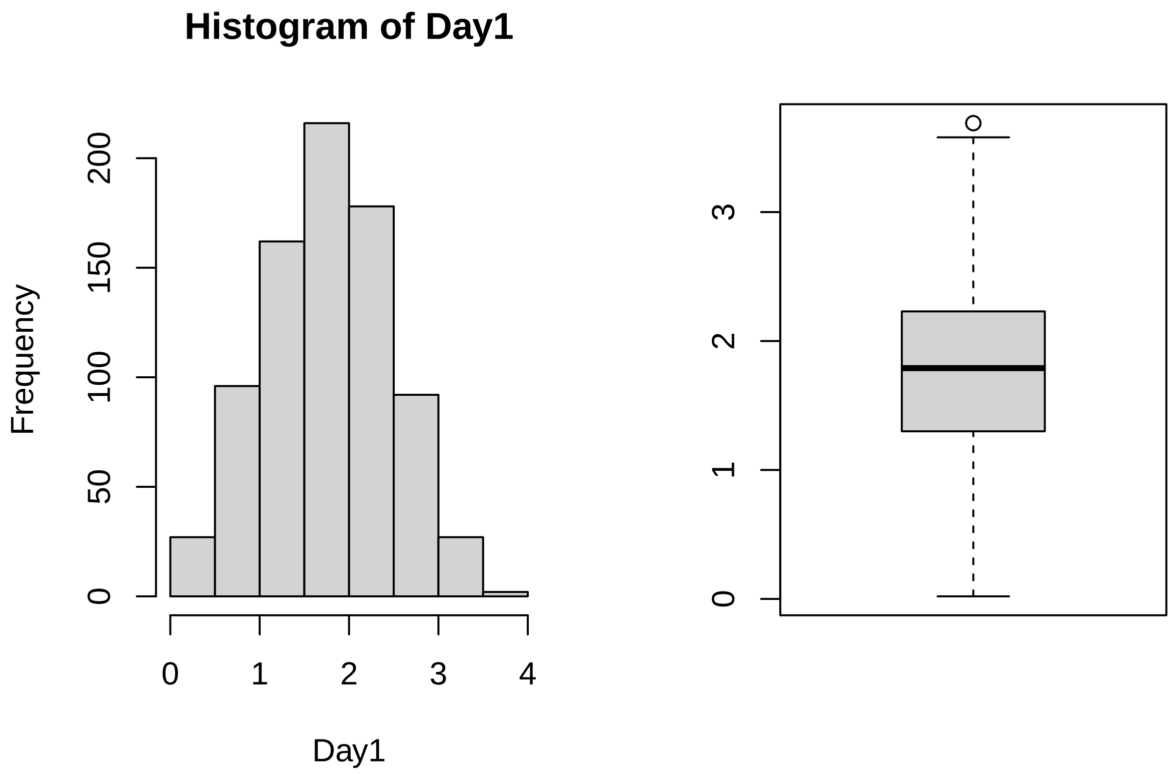
```r
normality$Day1[672]=2
summary(normality)
```

```
     Gender           Day1            Day2             Day3       
 Min.   :1.000   Min.   :0.020   Min.   :0.0000   Min.   :0.0200  
 1st Qu.:1.000   1st Qu.:1.305   1st Qu.:0.4100   1st Qu.:0.4400  
 Median :2.000   Median :1.790   Median :0.8200   Median :0.7600  
 Mean   :1.619   Mean   :1.772   Mean   :0.9718   Mean   :0.9739  
 3rd Qu.:2.000   3rd Qu.:2.230   3rd Qu.:1.3500   3rd Qu.:1.5250  
 Max.   :2.000   Max.   :3.690   Max.   :3.4400   Max.   :3.4100  
                                 NA’s   :538      NA’s   :677     
```

```r
attach(normality)
skewness(Day1)
```

```
[1] -0.003379654
```

```r
par(mfrow=c(1,2))
hist(Day1)
boxplot(Day1)
```



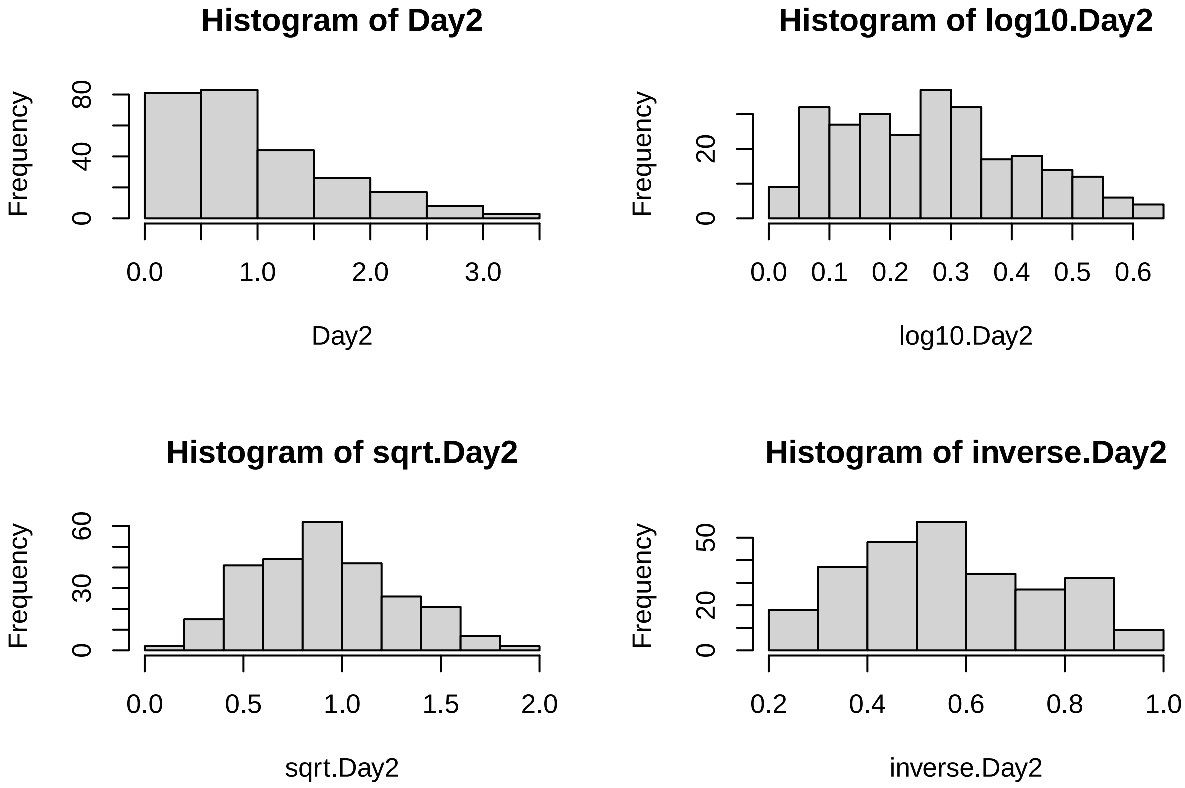
```r
#Checking normality of Day1
#H0:The distribution of the sample is not significantly different from a normal distribution.
#H1:The distribution is significantly different from a normal distribution.
#If p-value > 0.05, H0 may be accepted.
shapiro.test(Day1)
```

```

    Shapiro-Wilk normality test

data:  Day1
W = 0.99592, p-value = 0.03416
```

```
#convert Day2 into normality
log10.Day2=log10(Day2+1)
sqrt.Day2=sqrt(Day2)
inverse.Day2=1/(Day2+1)
par(mfrow=c(2,2))
hist(Day2)
hist(log10.Day2)
hist(sqrt.Day2)
hist(inverse.Day2)
```



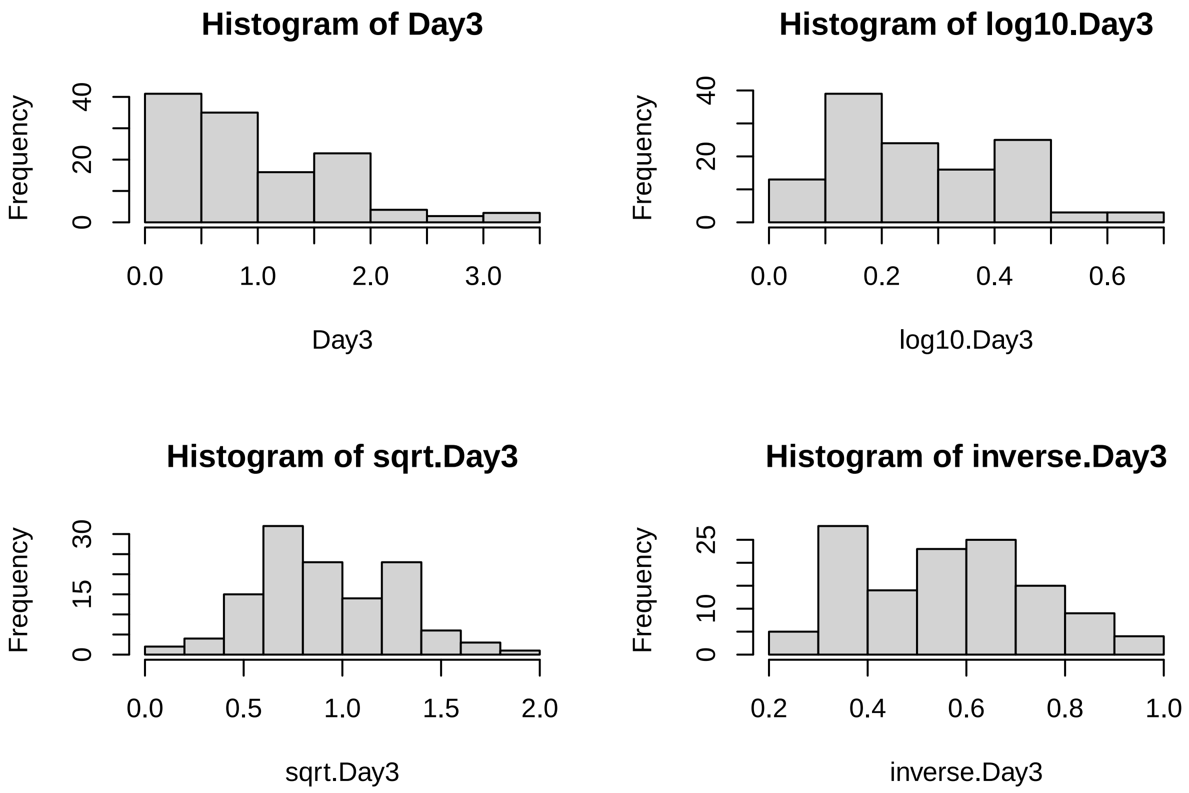
```
#Checking normality of Day2
#H0:The distribution of the sample is not significantly different from a normal distribution.
#H1:The distribution is significantly different from a normal distribution.
#If p-value > 0.05, H0 may be accepted.
shapiro.test(inverse.Day2)
```

```

    Shapiro-Wilk normality test
```

```
data:  inverse.Day2
W = 0.97421, p-value = 0.0001103
```

```
#convert Day2 into normality
log10.Day3=log10(Day3+1)
sqrt.Day3=sqrt(Day3)
inverse.Day3=1/(Day3+1)
par(mfrow=c(2,2))
hist(Day3)
hist(log10.Day3)
hist(sqrt.Day3)
hist(inverse.Day3)
```



```
#Checking normality of Day3
#H0:The distribution of the sample is not significantly different from a normal distribution.
#H1:The distribution is significantly different from a normal distribution.
#If p-value > 0.05, H0 may be accepted.
shapiro.test(inverse.Day3)
```

```
    Shapiro-Wilk normality test

data:  inverse.Day3
W = 0.9724, p-value = 0.0126
```