R Statistics Training.

Sin

2022-09-10

Episode 4

- Transforming Skewed Data.

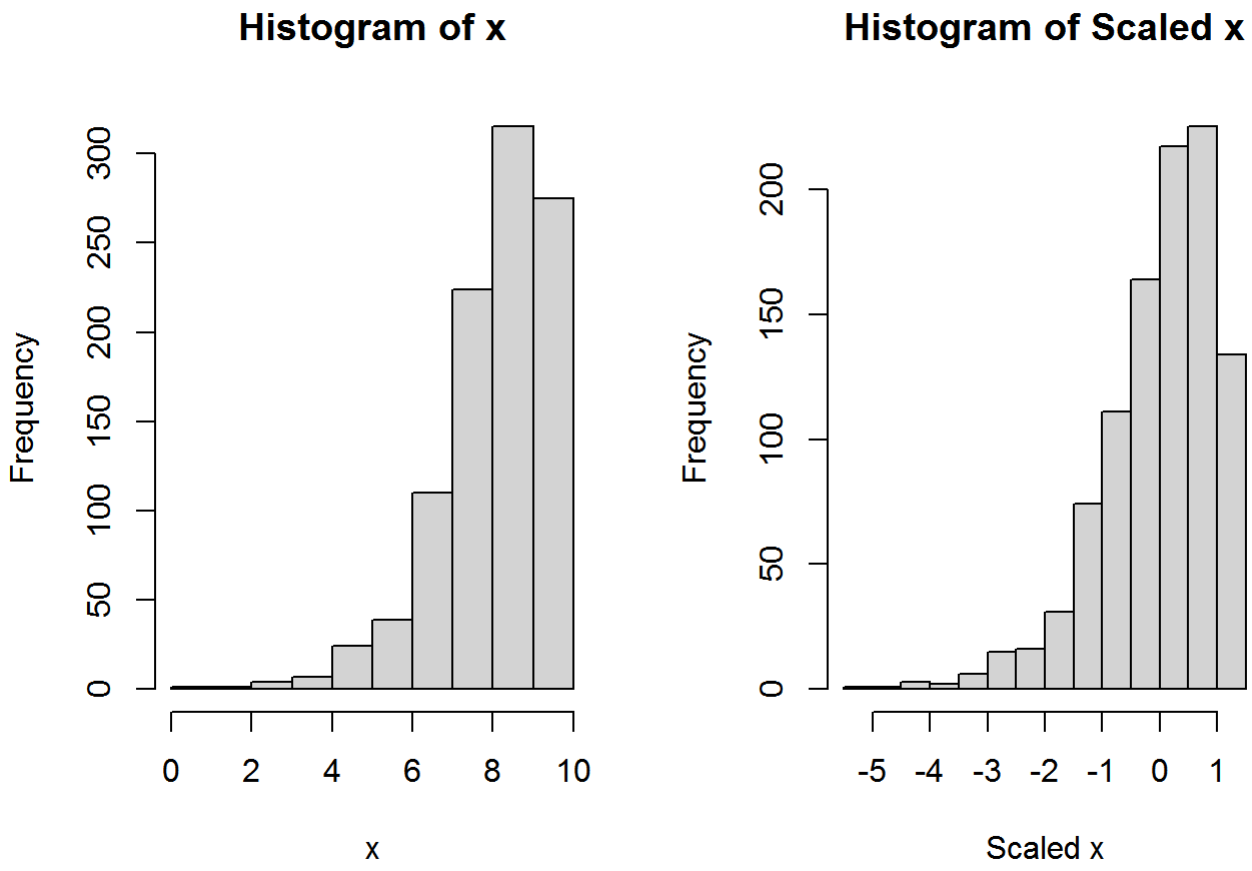
Scaling our dataset

> Histograms

```
skew <- read.csv("xskew.csv", header = T)
head(skew)
```

```
##   X        x
## 1 1 9.997147
## 2 2 7.806174
## 3 3 8.665934
## 4 4 8.888980
## 5 5 7.258421
## 6 6 8.949910
```

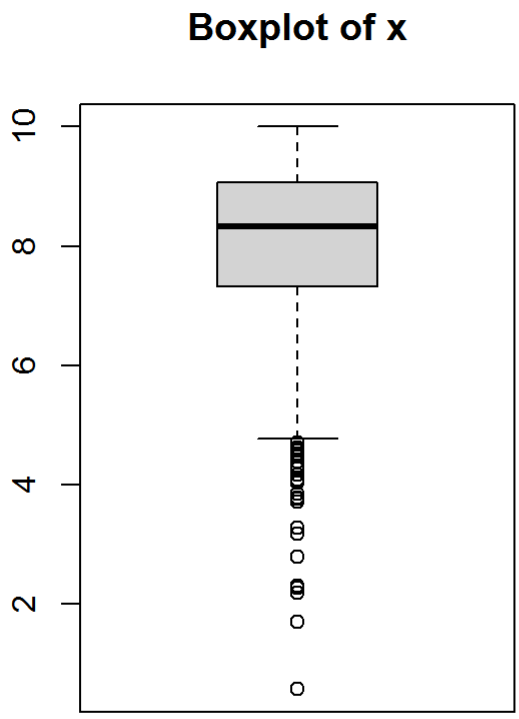
```
attach(skew)
skew.scale.x <- scale(x)
par(mfrow = c(1,2))
hist(x,
     main = "Histogram of x"
    )
hist(skew.scale.x,
     main = "Histogram of Scaled x",
     xlab = "Scaled x"
     )
```

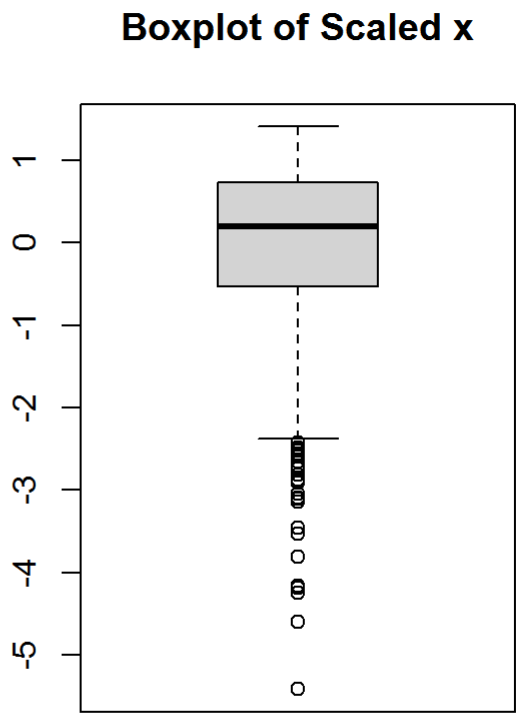


We can see that both histograms are similar; both are negatively skewed.

Boxplots

```
par(mfrow = c(1,2))
boxplot(x,
        main = "Boxplot of x")
boxplot(skew.scale.x,
        main = "Boxplot of Scaled x")
```



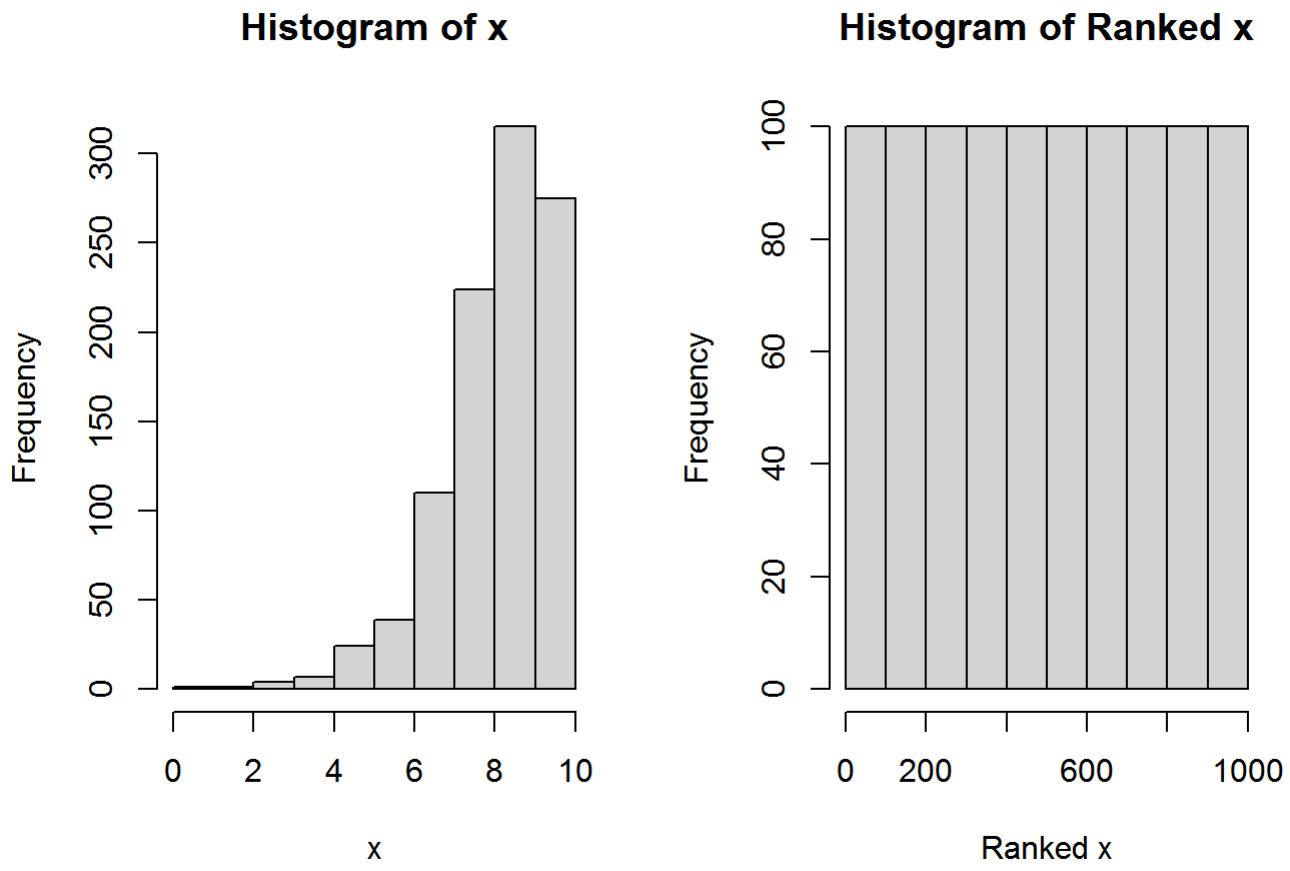


There's no difference in the boxplots too.

Ranking our Dataset

Histograms

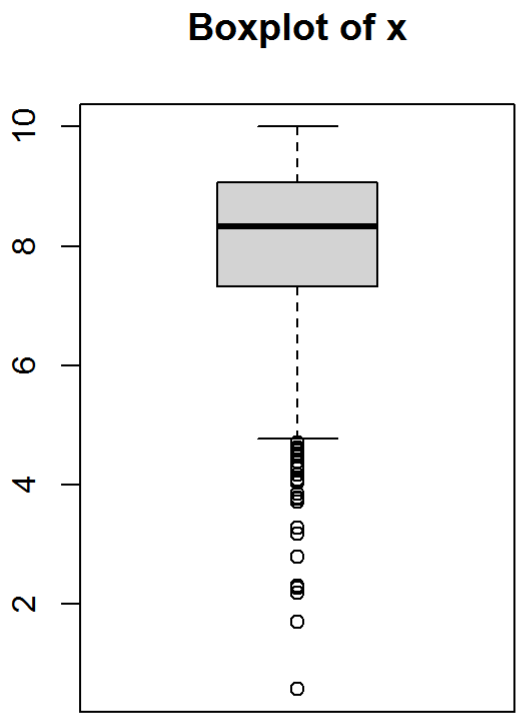
```
skew.rank.x <- rank(x)
par(mfrow = c(1,2))
hist(x)
hist(skew.rank.x,
     main = "Histogram of Ranked x",
     xlab = "Ranked x"
     )
```

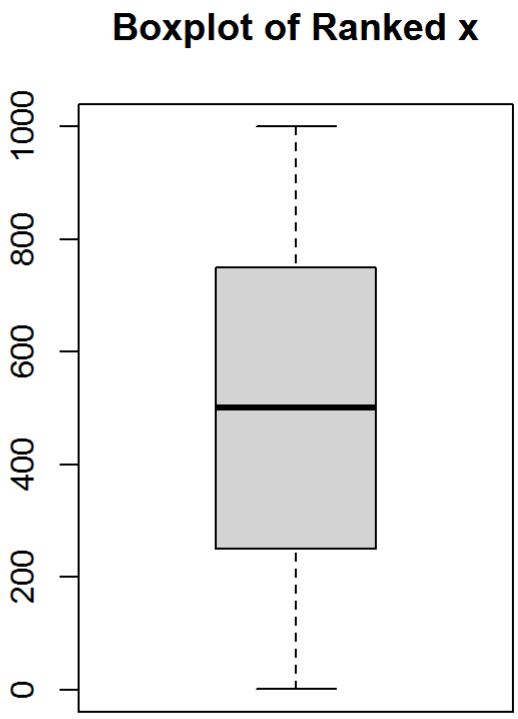


Now, there's a visible difference in our histograms; while the histogram of x is negatively skewed, histogram of ranked x is flat, meaning that there's no more outlier.

Boxplots

```
par(mfrow = c(1,2))
boxplot(x,
        main = "Boxplot of x"
        )
boxplot(skew.rank.x,
        main = "Boxplot of Ranked x"
     )
```



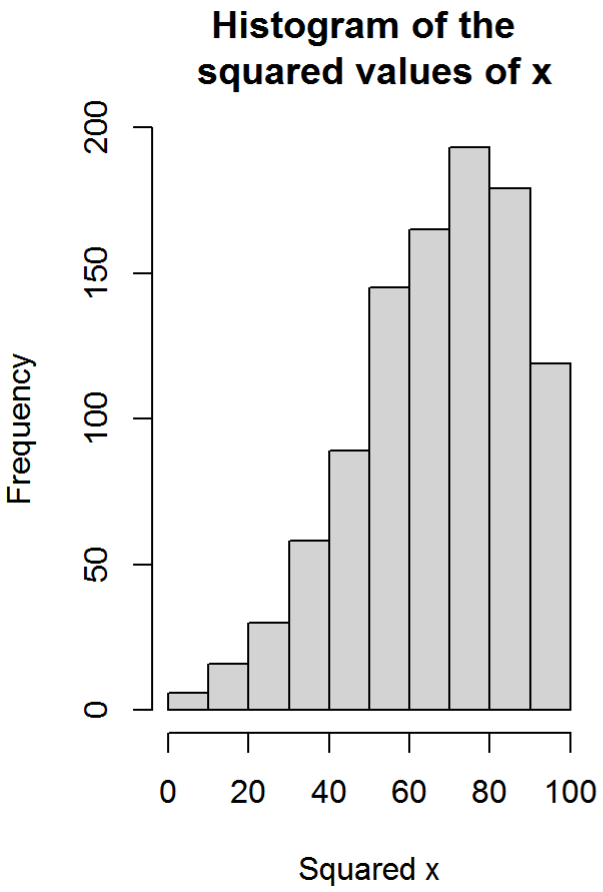


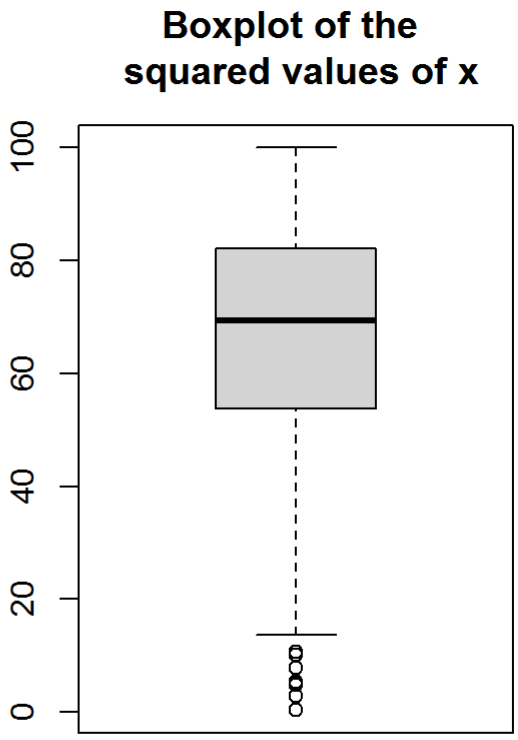
Ranking x has transformed the outliers.

Squaring our dataset

> Second Power

```
par(mfrow = c(1, 2))
squared.x <- x^2
hist(squared.x,
     main = "Histogram of the \n squared values of x",
     xlab = "Squared x")
boxplot(squared.x,
        main = "Boxplot of the \n squared values of x"
        )
```

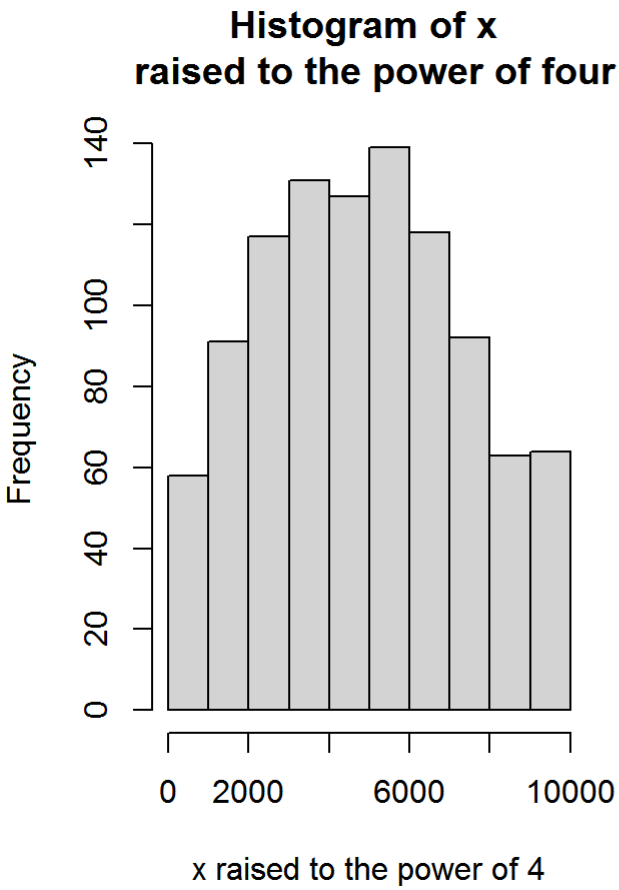


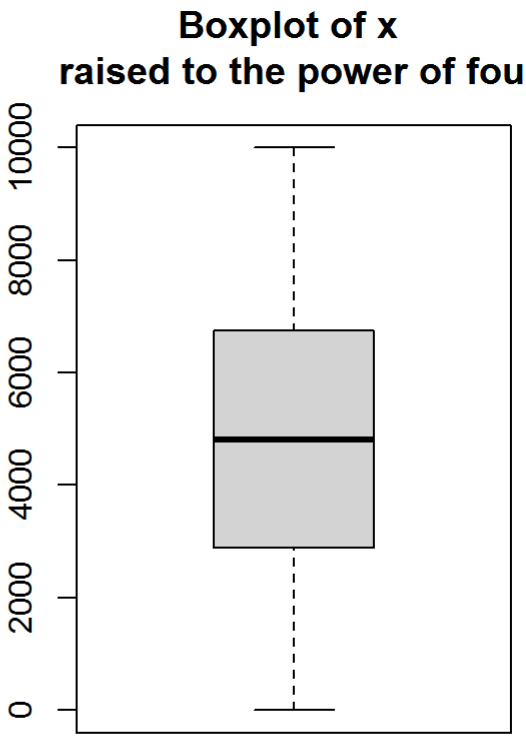


After raising x to the power of 2, we can see from our histogram that the skewness has reduced. The outliers have decreased too.

Fourth Power

```
par(mfrow = c(1, 2))
fourth.x <- x^4
hist(fourth.x,
     main = "Histogram of x \n raised to the power of four",
     xlab = "x raised to the power of 4")
boxplot(fourth.x,
        main = "Boxplot of x \n raised to the power of four"
        )
```





Now, x is neither skewed nor have any outlier.

Episode 5

- Working with the Data File.

```
require(datasets)
data("ToothGrowth")
attach(ToothGrowth)
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

While len is a quantitative variable, supp is a categorical variable.

```
aggregate(len ~ supp, FUN = mean)
```

```
##   supp      len
## 1   OJ 20.66333
## 2   VC 16.96333
```

This is the mean of “len” for each of the group in “supp”

```
aggregate(len ~ supp, FUN = median)
```

```
##   supp  len
## 1   OJ 22.7
## 2   VC 16.5
```

This is the median of “len” for each of the group in “supp”