Forbes2000 Data Analysis

Hyeongjin Joo

2023-01-14

데이터에 대한 summary

```
Forbes<-read.csv("Forbes2000.csv",header=T)
dim(Forbes)
```

```
## [1] 2000    8
```

```
names(Forbes)
```

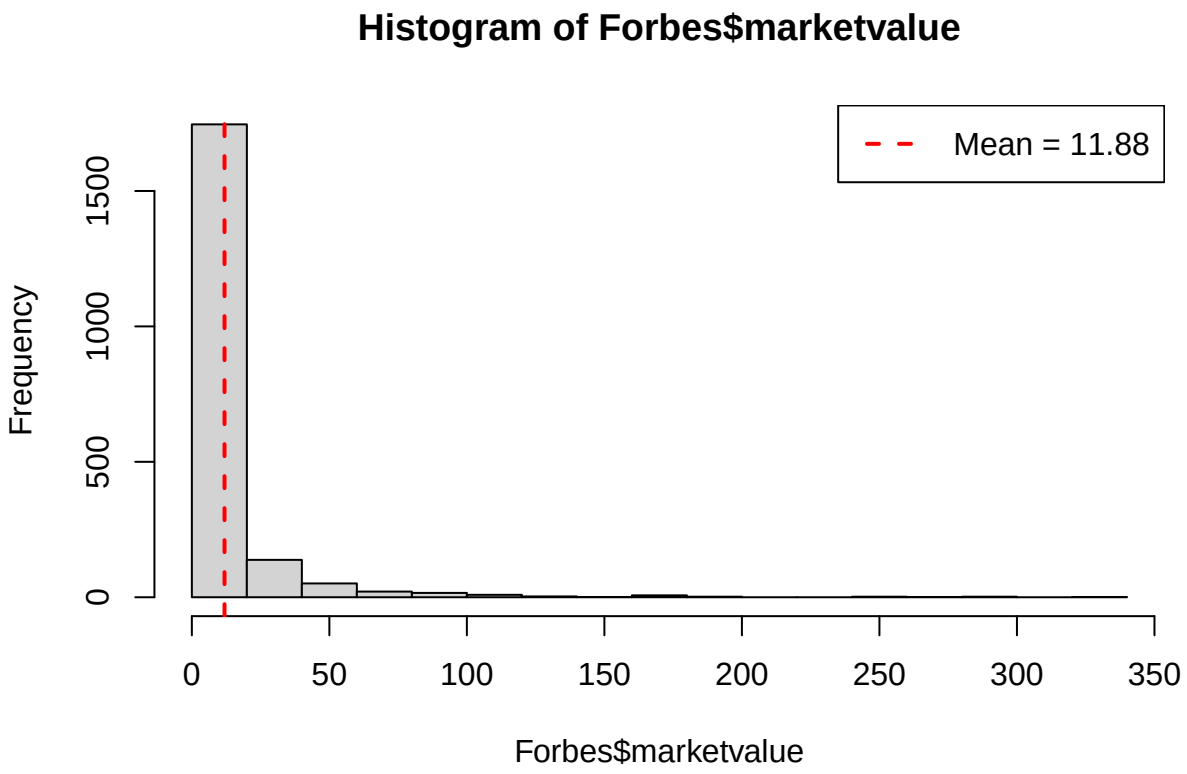
```
## [1] "rank"        "name"        "country"     "category"    "sales"
## [6] "profits"     "assets"      "marketvalue"
```

Marketvalue에 대한 분석

Graph

Histogram

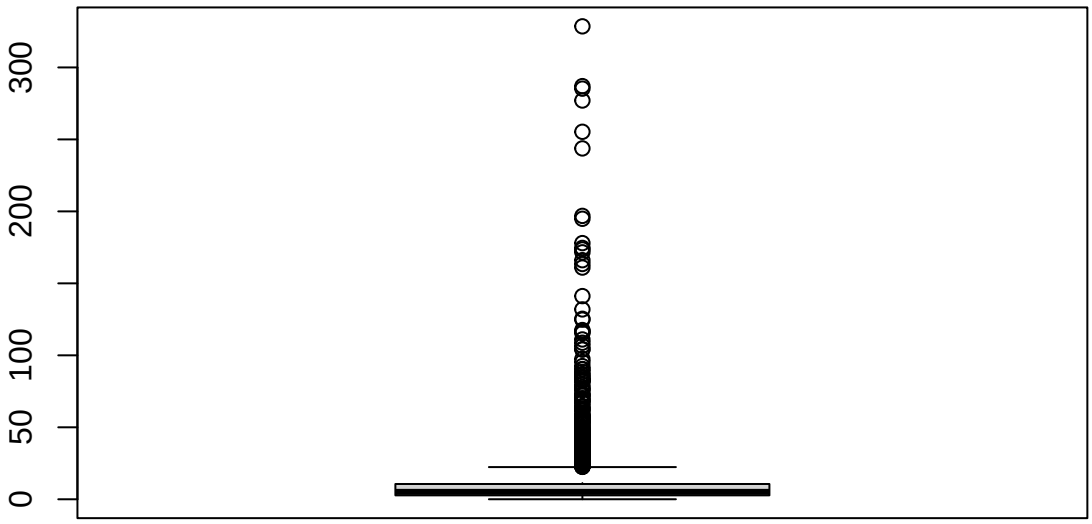
```
hist(Forbes$marketvalue)
abline(v = mean(Forbes$marketvalue), col="red", lwd=2, lty="dashed")
legend("topright", c("Mean = 11.88"),col="red", lwd=2, lty = "dashed")
```



- 0과 20 사이에 대부분이 관측값이 모여 있다. - 왼쪽으로 많이 쏠려 있다. - 오른쪽에 outlier들이 있는 것 같다.

Boxplot

```
boxplot(Forbes$marketvalue)
```



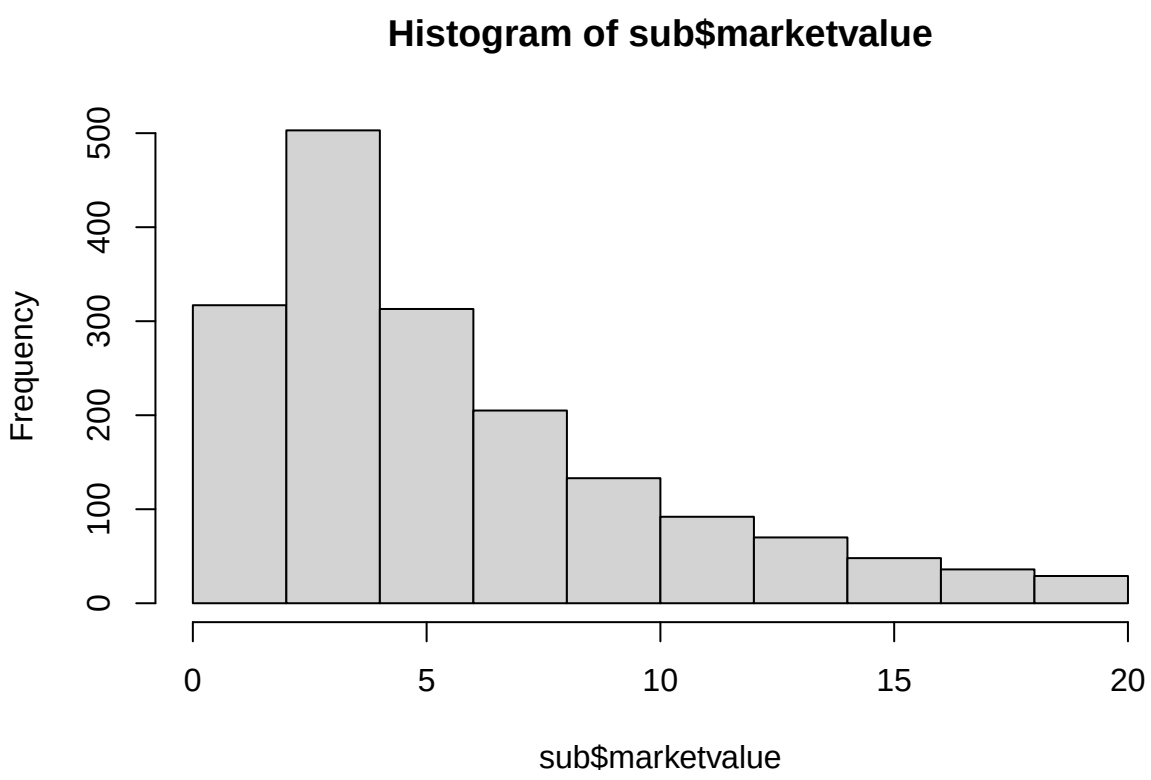
-히스토그램과 같이 왼쪽으로 많이 쏠려 있다는 것을 확인할 수 있다.
-75% 이상이 20보다 작은 쪽에 있는 것으로 보이고 오른쪽으로 긴 꼬리를 가지고 있는 것으로 보인다.

작은 쪽의 분포

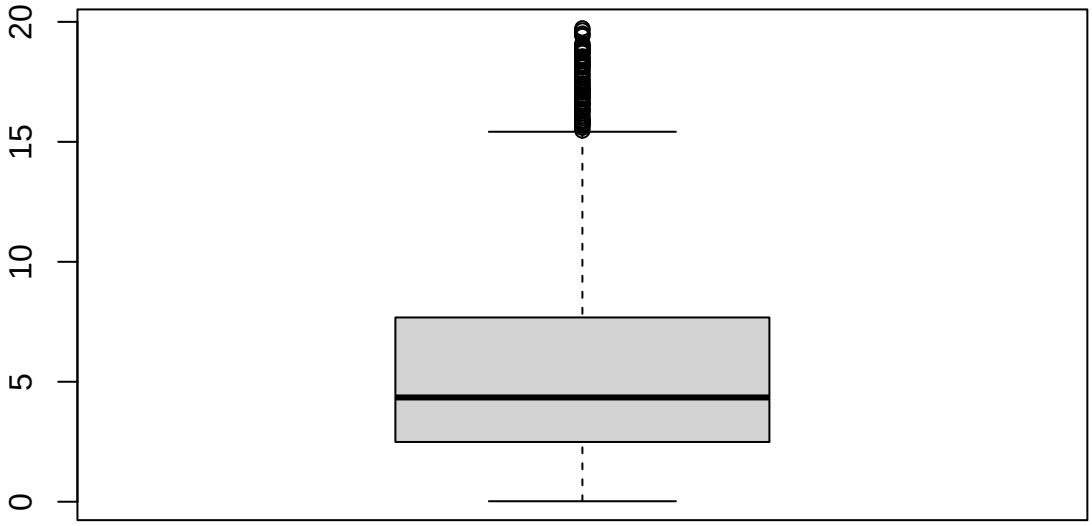
```
# marketvalue  20
sub<-Forbes[which(Forbes$marketvalue<=20),]
dim(sub)
```

```
## [1] 1746    8
```

```
hist(sub$marketvalue)
```



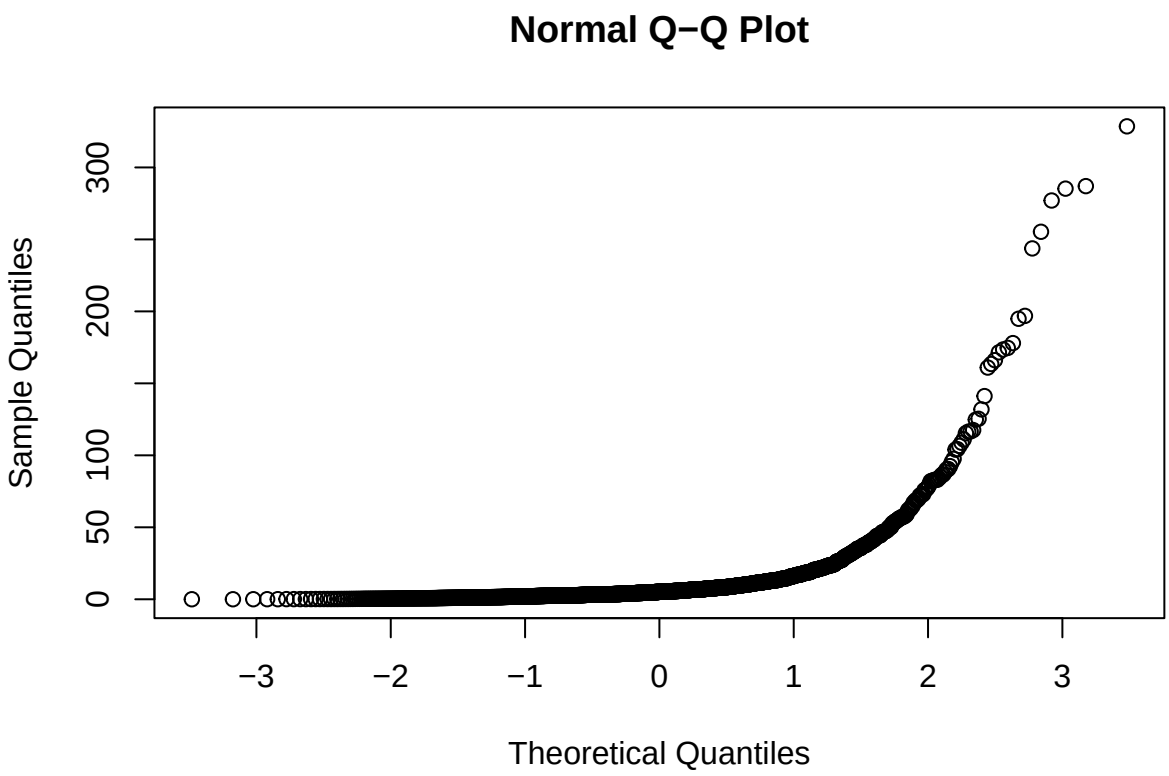

```
boxplot(sub$marketvalue)
```



- 왼쪽으로 쏠린 분포를 갖는다. - 전체적으로 왼쪽으로 쏠린 분포를 가지고 있다.

정규확률그림 (qqplot)

```
qqnorm(Forbes$marketvalue)
```



- 정규확률그림이 아래로 볼록한 형태이므로 데이터의 분포가 - 왼쪽으로 쏠려있는 형태라는 것을 알 수 있다.

기술통계량

Mean, sd, Quantile

```
mean(Forbes$marketvalue)
```

```
## [1] 11.87766
```

```
sd(Forbes$marketvalue)
```

```
## [1] 24.4602
```

```
summary(Forbes$marketvalue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.02    2.72    5.15   11.88   10.60  328.54
```

- 평균이 중앙값보다 크므로 왼쪽으로 쏠려 있는 분포라는 것을 알 수 있다.

왜도, 첨도

```
# PKG
library(moments)

#  (Skewness)
skewness(Forbes$marketvalue)
```

```
## [1] 6.432051
```

```
#  (Kurtosis)
kurtosis(Forbes$marketvalue)
```

```
## [1] 58.73137
```

- 왜도가 6으로 매우 크므로 왼쪽으로 상당히 쏠린 분포인 것을 알 수 있다.
- 첨도가 58.7로 3보다 매우 크므로 정규분포와 비슷한 꼬리를 갖는다고 하기 힘들다.

왜도가 0인지 검정

```
agostino.test(Forbes$marketvalue)
```

```
##
##  D'Agostino skewness test
##
## data:  Forbes$marketvalue
## skew = 6.4321, z = 41.4627, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

- p-value가 매우 작으므로 유의수준 0.05에서 marketvalue는 왜도가 0이라고 하기 힘들다.

```
anscombe.test(Forbes$marketvalue)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  Forbes$marketvalue
## kurt = 58.731, z = 26.573, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

- p-value가 매우 작으므로 유의수준 0.05에서 marketvalue는 첨도가 3이라고 하기 힘들다.

정규성 검정

```
shapiro.test(Forbes$marketvalue)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Forbes$marketvalue
## W = 0.40421, p-value < 2.2e-16
```

- p-value가 매우 작으므로 유의순준 0.05에서 귀무가설 기각 정규분포를 따른다고 하기 힘들다.

```
# Jarque test
jarque.test(Forbes$marketvalue)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  Forbes$marketvalue
## JB = 272623, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
# Kolmogorov-Smirnov test
library(nortest)
lillie.test(Forbes$marketvalue)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Forbes$marketvalue
## D = 0.31392, p-value < 2.2e-16
```

```
ad.test(Forbes$marketvalue)
```

```
##
##  Anderson-Darling normality test
##
## data:  Forbes$marketvalue
## A = 352.75, p-value < 2.2e-16
```

- 모든 검정에서의 p-value가 0.05보다 작으므로 데이터가 정규분포로부터 나왔다는 귀무가설을 채택하기 힘들다.
- 즉, 데이터가 정규분포로부터 나왔다고 하기 힘들다.

변수변환 into Log, sqrt

```
#
min(Forbes$marketvalue)
```

```
## [1] 0.02
```

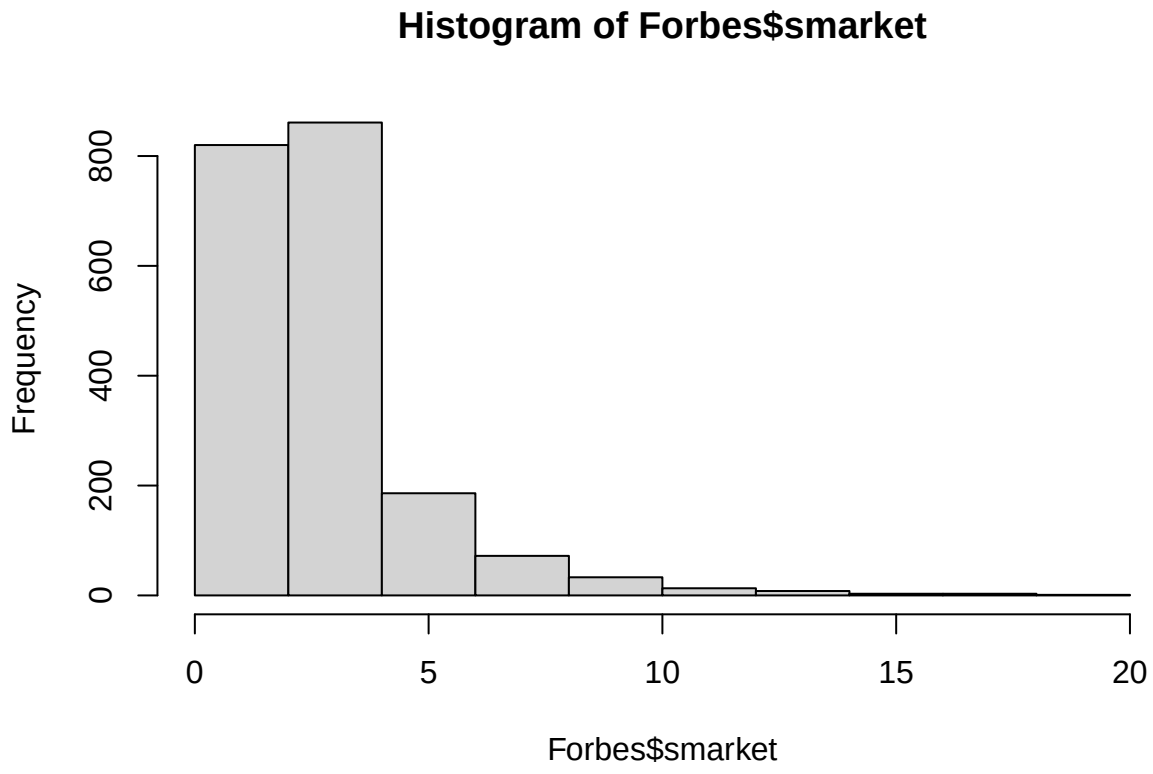
```
#
max(Forbes$marketvalue)
```

```
## [1] 328.54
```

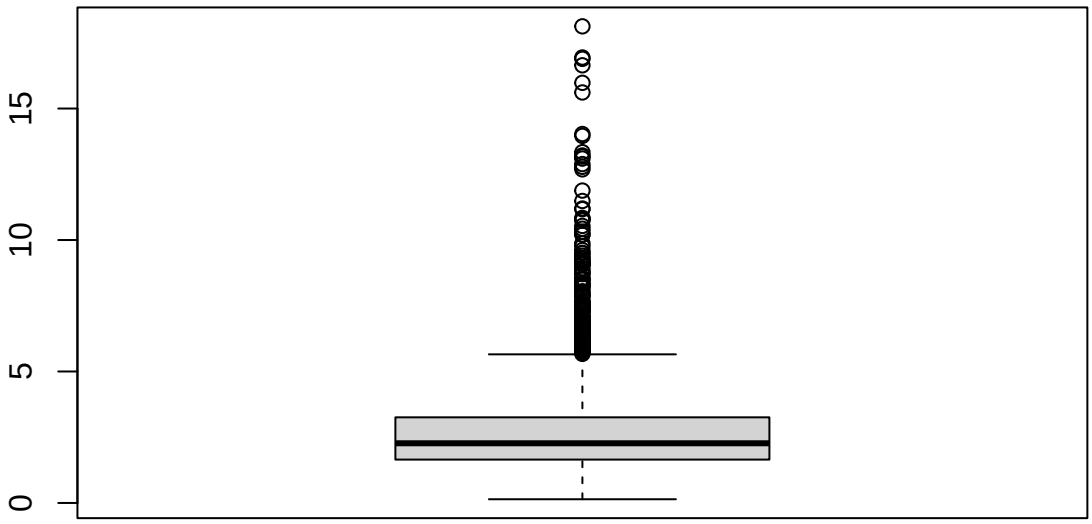
```
#   : sqrt
Forbes$smarket <- sqrt(Forbes$marketvalue)
#   : sqrt
Forbes$lmarket <- log(Forbes$marketvalue)
```

sqrt에 대한 분석

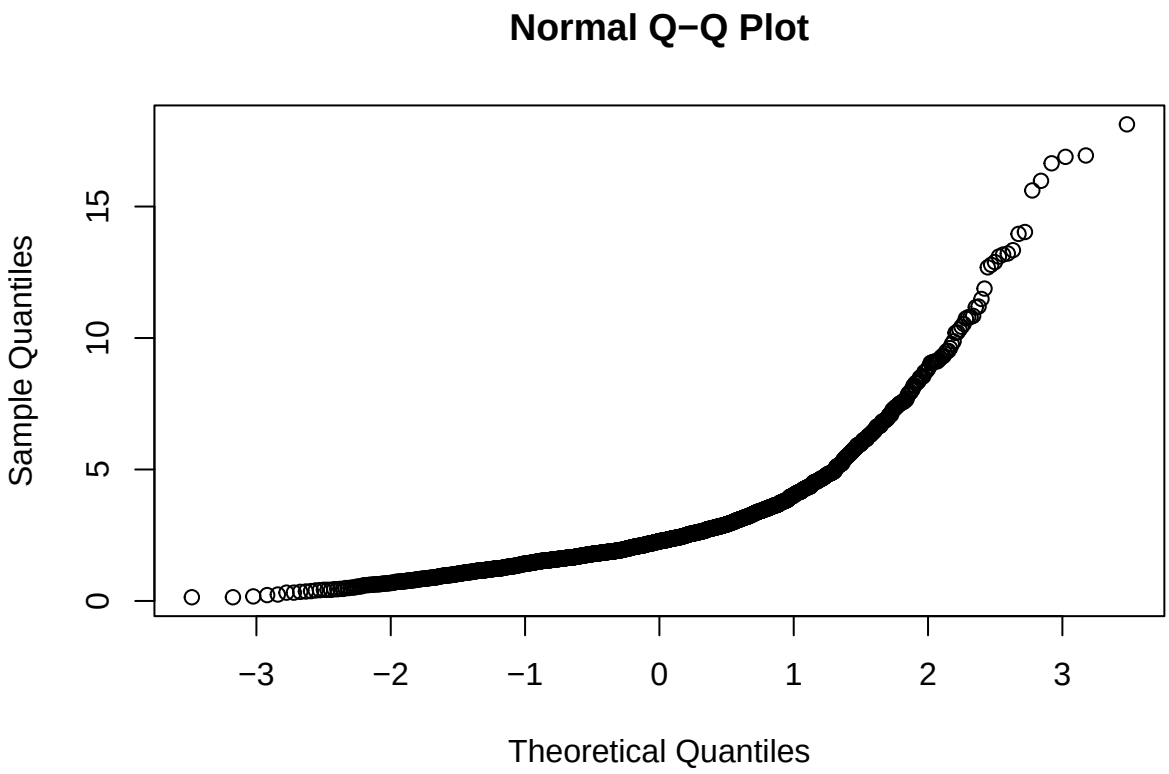
```
hist(Forbes$smarket)
```




```
boxplot(Forbes$smarket)
```




```
qqnorm(Forbes$smarket)
```



- 히스토그램과 상자그림에서 변환 전보다 쏠린 정도가 완화되었지만 아직 왼쪽으로 쏠린 분포를 보이고 있다.

```
#  (Skewness)
skewness(Forbes$smarket)
```

```
## [1] 2.788321
```

```
agostino.test(Forbes$smarket)
```

```
##
##  D'Agostino skewness test
##
## data:  Forbes$smarket
## skew = 2.7883, z = 29.1411, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

- 왜도는 2.79로 변환 전보다 0에 가까운 값을 갖지만 아직 0에 가깝지 않다.
- 왜도가 0인지에 대한 검정에서 p-value가 0.05보다 작으므로 왜도가 0이라고 하기 힘들다.

```
#   ad.test
ad.test(Forbes$smarket)
```

```
##
```

```
##  Anderson-Darling normality test
##
## data:  Forbes$smarket
## A = 124.28, p-value < 2.2e-16
```

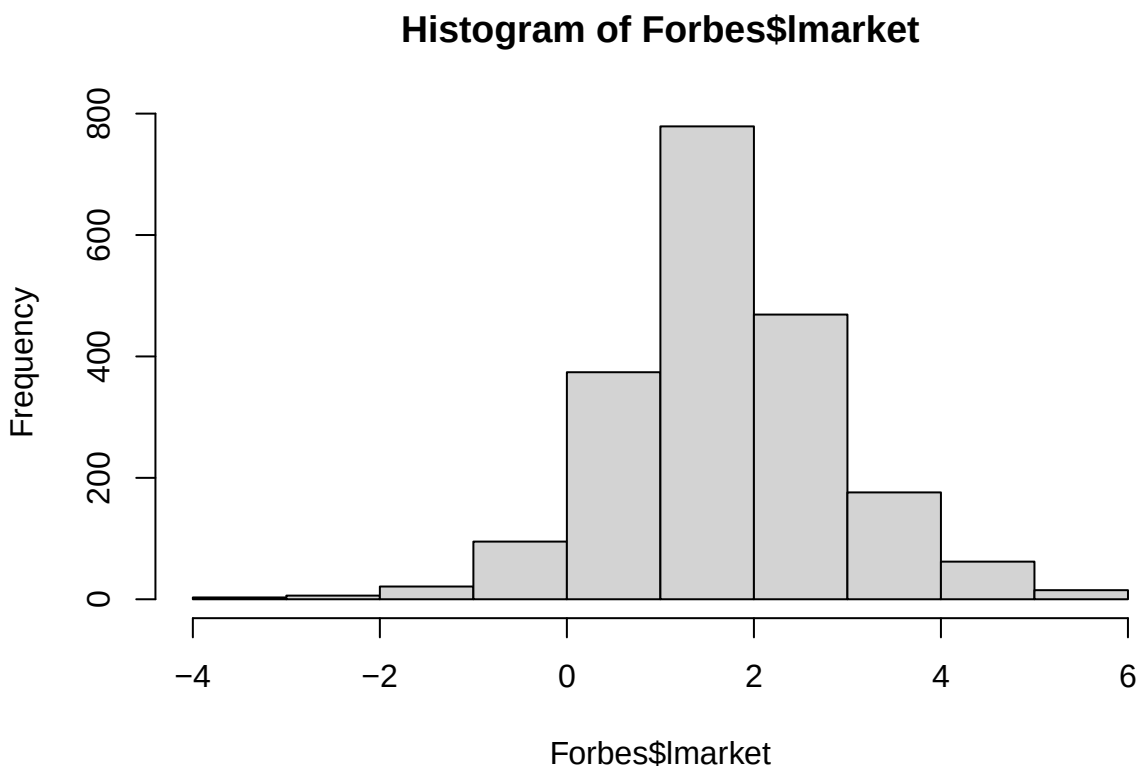
```
#   lillie.test
lillie.test(Forbes$marketvalue)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Forbes$marketvalue
## D = 0.31392, p-value < 2.2e-16
```

- 정규성 검정에서 역시 유의확률(p-value)가 0.05보다 작으므로 데이터가 정규분포로부터 나왔다고 하기 힘들다.

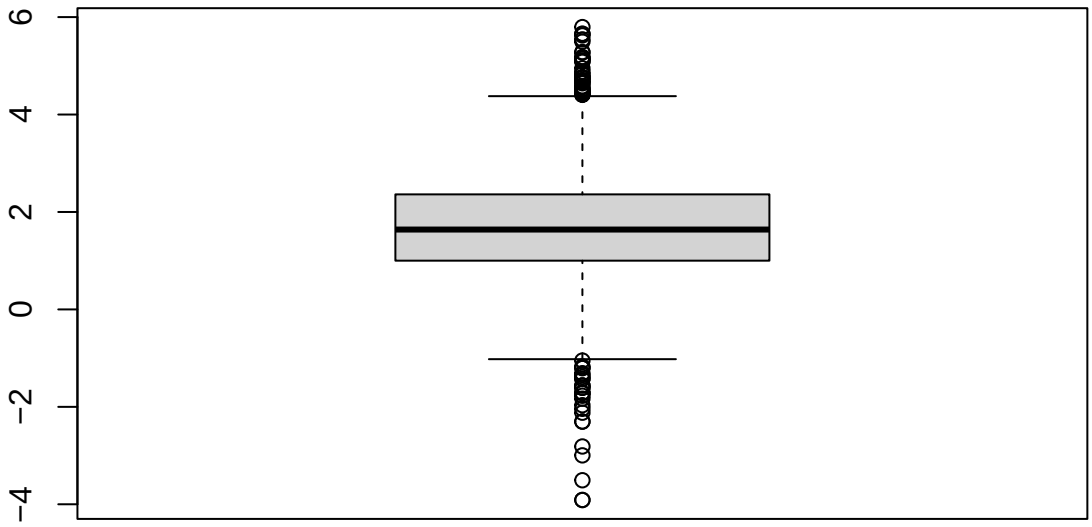
log에 대한 분석

```
hist(Forbes$lmarket)
```



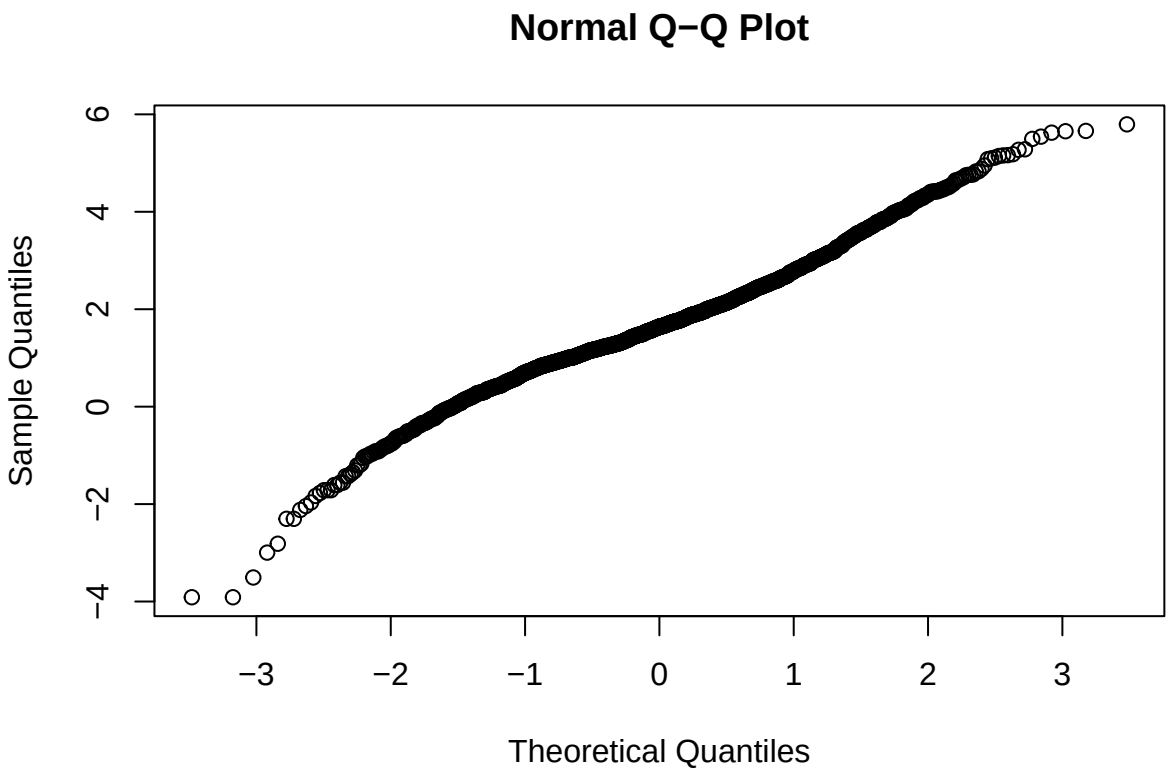
- 히스토그램으로부터 데이터의 분포가 좌우 대칭인 단봉형 분포인 것처럼 보인다.

```
boxplot(Forbes$lmarket)
```



- 분포가 좌우대칭으로 보이나 작은 쪽과 큰 쪽 모두 fence(경계점) 바깥에 많은 점들이 위치해 있는 것으로 보인다. - 꼬리가 두꺼울 수도 있다.

```
qqnorm(Forbes$lmarket)
```



- 직선에 가깝기는 하나 약간 역S자 형태를 보인다.

```
#  (Skewness)
skewness(Forbes$lmarket)
```

```
## [1] 0.03204996
```

```
agostino.test(Forbes$lmarket)
```

```
##
##  D'Agostino skewness test
##
## data:  Forbes$lmarket
## skew = 0.03205, z = 0.58718, p-value = 0.5571
## alternative hypothesis: data have a skewness
```

- 왜도가 0.03으로 0에 매우 가깝고 왜도가 0인지를 검정하기 위한 가설검정에서도 p-value가 0.56으로 0.05보다 크므로 귀무가설을 기각할 수 없다.
- 즉, 왜도가 0이라고 할 수 있다.

```
#  (Kurtosis)
kurtosis(Forbes$lmarket)
```

```
## [1] 4.352814
```

```
anscombe.test(Forbes$lmarket)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  Forbes$lmarket
## kurt = 4.3528, z = 7.7578, p-value = 8.645e-15
## alternative hypothesis: kurtosis is not equal to 3
```

- 첨도가 4.4로 3보다 크고 첨도가 3인지에 대한 검정에서 p-value가 8.645e-15로 매우 작으므로 첨도가 3이라고 할 수 없다.

```
#    : ad.test
ad.test(Forbes$lmarket)
```

```
##
##  Anderson-Darling normality test
##
## data:  Forbes$lmarket
## A = 8.9438, p-value < 2.2e-16
```

```
#    : lillie.test
lillie.test(Forbes$lmarket)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Forbes$lmarket
## D = 0.048933, p-value = 4.782e-12
```

- p-value가 0.05보다 작으므로 귀무가설 기각.
- 즉, 데이터의 분포가 정규분포라 하기 힘들다.
- 따라서, 데이터의 분포는 log를 취했을 경우 좌우대칭이고 꼬리가 정규분포보다 두꺼운 분포라고 할 수 있다.