```
### 인천 3년 평균 인구수 EDA ###
## set working directory
setwd('C:/Users/wndy4/Desktop/Project_DEMA')
## csv file load
dat_pop = read.csv('정동호/인천광역시인구수/인천광역시 월별 구별 인구수(new).csv',header=T,stringsAsFactors=F)

# 기초정보
dim(dat_pop)
```

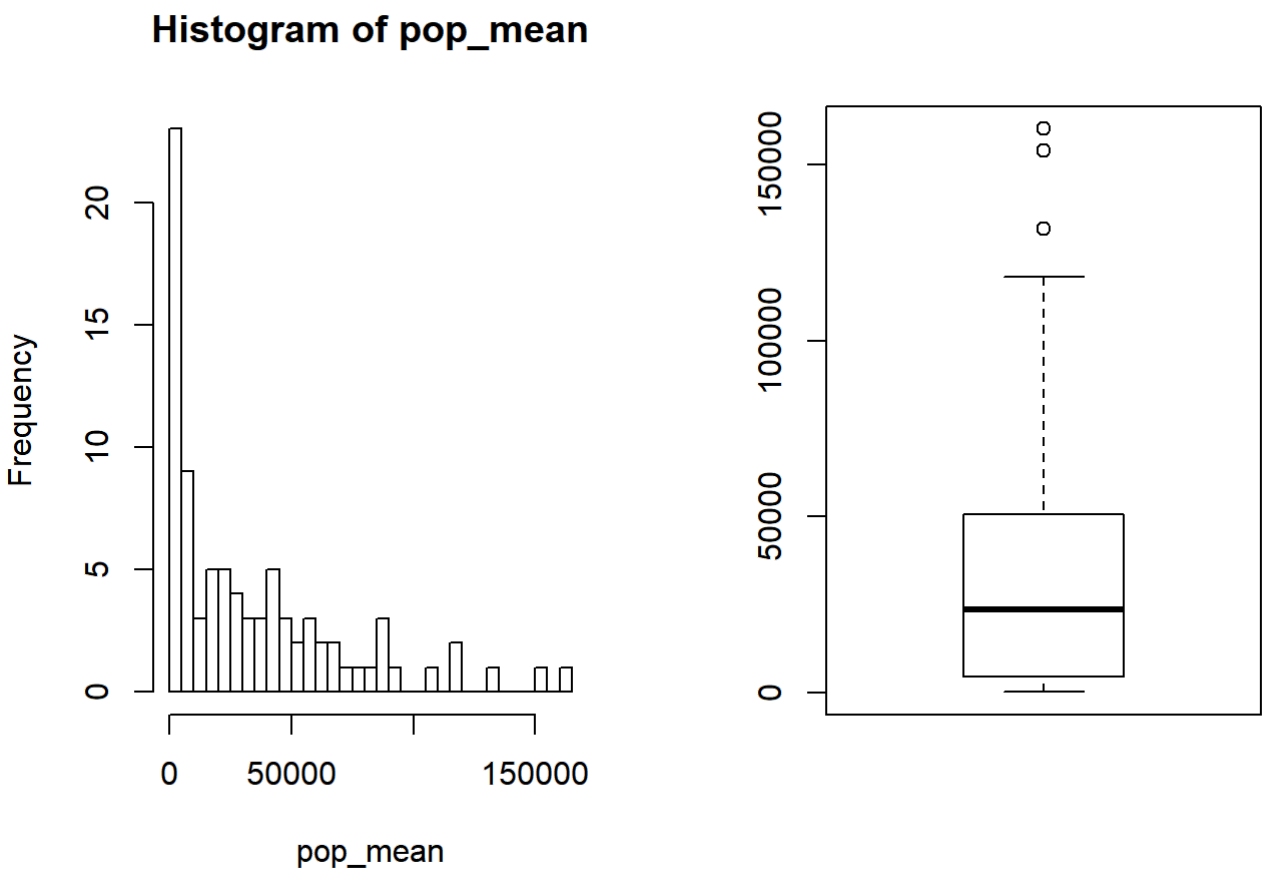
```
## [1] 95 39
```

```
a = which(dat_pop$행정구역 == '합계')
# 합계 제외한 3년 평균 인구수 데이터
pop_mean = dat_pop[-a,39]

# 평균, 분산
mean(pop_mean);sd(pop_mean)
```

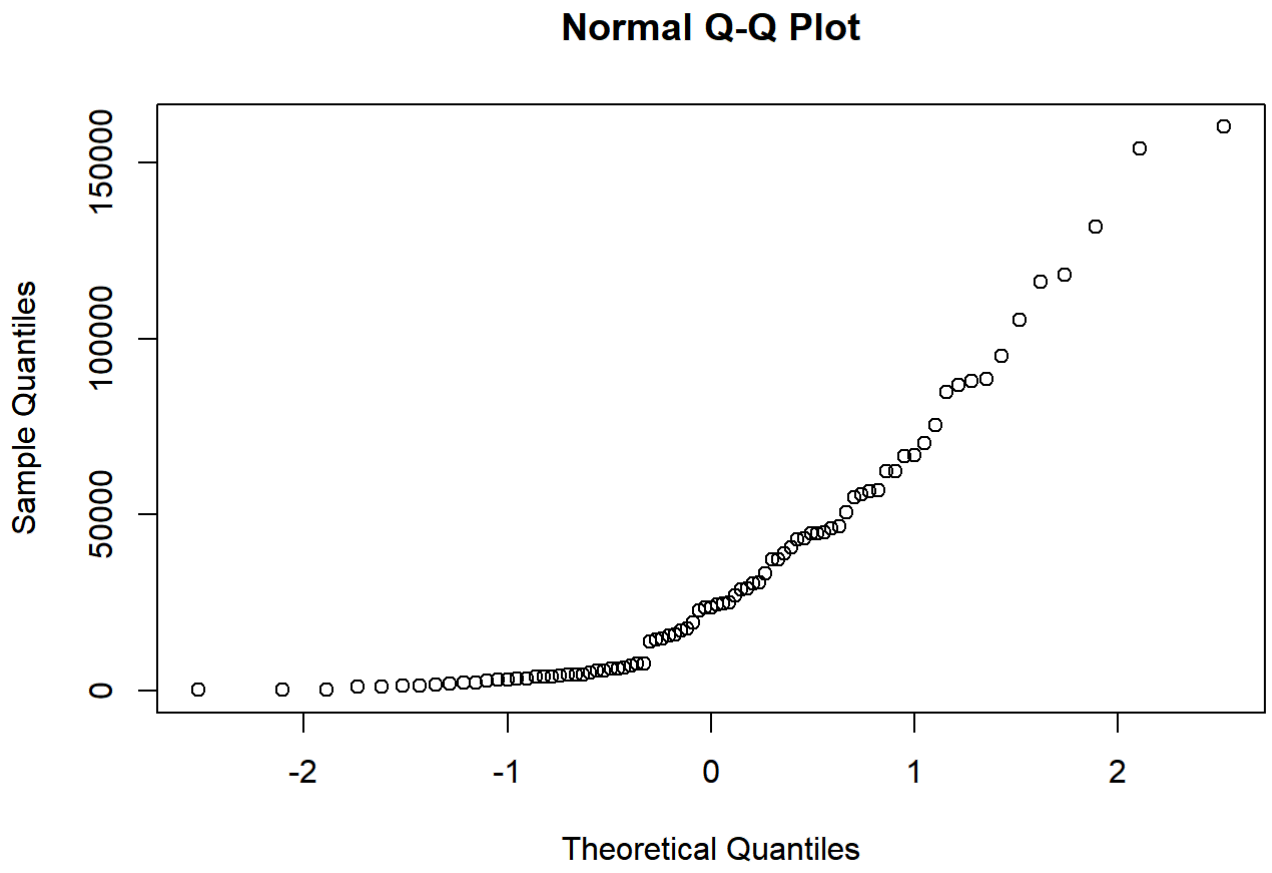
```
## [1] 34649.31
```

```
## [1] 37217.64
```

```
graphics.off()
# 그래프
windows()
par(mfrow=c(1,2))
hist(pop_mean,breaks = 40)
boxplot(pop_mean)
```




```
# 왼쪽으로 쏠려 있는 모습이며, outlier가 있는 것으로 보인다.

# 정규확률그림
windows()
qqnorm(pop_mean)
```




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

# 기술통계량
summary(pop_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     253    4623   23627   34649   50723  160052
```

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

# 왜도 및 첨도
# 왜도, 첨도
library(moments)
```

```
## Warning: package 'moments' was built under R version 3.4.4
```

```
skewness(pop_mean)
```

```
## [1] 1.416518
```

```
kurtosis(pop_mean) # 정규분포와 비슷한 꼬리를 갖는다고 하기 힘들다
```

```
## [1] 4.637511
```

```
# Agostino skewness test
# 왜도가 0인지 검정
agostino.test(pop_mean)
```

```
##
##  D'Agostino skewness test
##
## data:  pop_mean
## skew = 1.4165, z = 4.5287, p-value = 5.934e-06
## alternative hypothesis: data have a skewness
```

```
# p-value가 매우 작으므로 유의수준 0.05에서 왜도가 0이라고 하기 힘들다

# 첨도가 3인지 검정
anscombe.test(pop_mean)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pop_mean
## kurt = 4.6375, z = 2.4368, p-value = 0.01482
## alternative hypothesis: kurtosis is not equal to 3
```

```
# p-value가 작으므로 유의수준 0.05에서 첨도가 0이라고 하기 힘들다

# 정규성 검정
shapiro.test(pop_mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pop_mean
## W = 0.83224, p-value = 2.101e-08
```

```
# p-value가 매우 작으므로 유의수준 0.05에서 귀무가설 기각(귀무가설 : 정규분포를 따른다)
# 정규분포를 따른다고 하기 힘들다

jarque.test(pop_mean)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  pop_mean
## JB = 37.923, p-value = 5.824e-09
## alternative hypothesis: greater
```

```
library(nortest)

# kolmogorov-smirnov test
lillie.test(pop_mean)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pop_mean
## D = 0.17769, p-value = 5.327e-07
```

```
ad.test(pop_mean)
```

```
##
##  Anderson-Darling normality test
##
## data:  pop_mean
## A = 4.2991, p-value = 9.066e-11
```

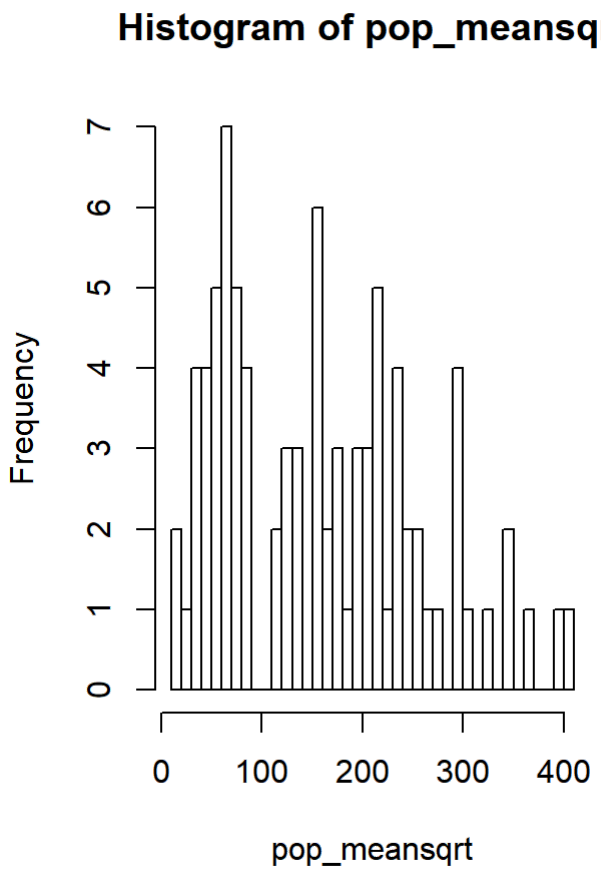
```
# 모든 검정에서의 p-value가 0.05보다 작으므로 데이터가
# 정규분포로부터 나왔다는 귀무가설을 채택하기 힘들다.
# 즉, 데이터가 정규분포로부터 나왔다고 하기 힘들다

# 변수 변환
# 변수변환전 범위 확인
min(pop_mean)
```

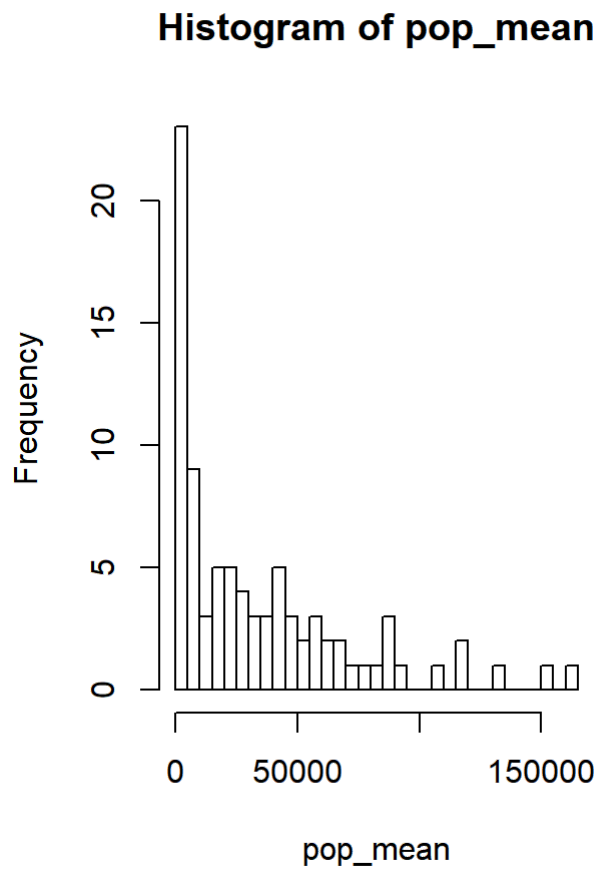
```
## [1] 253
```

```
max(pop_mean)
```

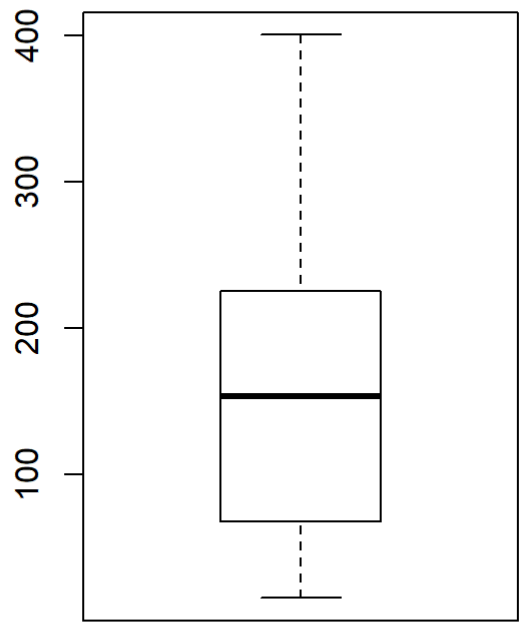
```
## [1] 160052
```

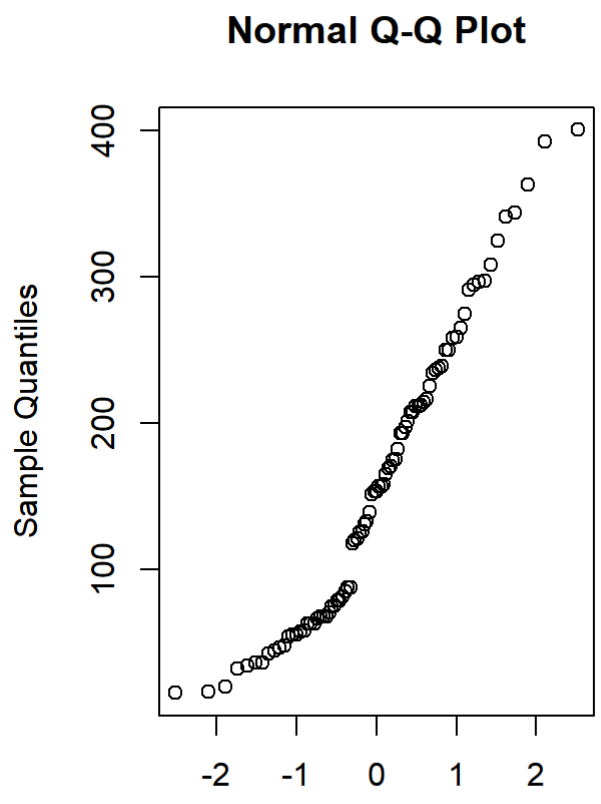
```
# 변수변환 : sqrt, log

pop_meansqrt = sqrt(pop_mean)
pop_meanlog = log(pop_mean)

# sqrt 분석
windows()
par(mfrow=c(1,2))
hist(pop_meansqrt,breaks = 40)
hist(pop_mean,breaks = 40)
```






```
windows()
par(mfrow=c(1,2))
boxplot(pop_meansqrt)
qqnorm(pop_meansqrt)
```






```
skewness(pop_meansqrt)
```

```
## [1] 0.4863035
```

```
agostino.test(pop_meansqrt)
```

```
##
##  D'Agostino skewness test
##
## data:  pop_meansqrt
## skew = 0.4863, z = 1.8847, p-value = 0.05946
## alternative hypothesis: data have a skewness
```

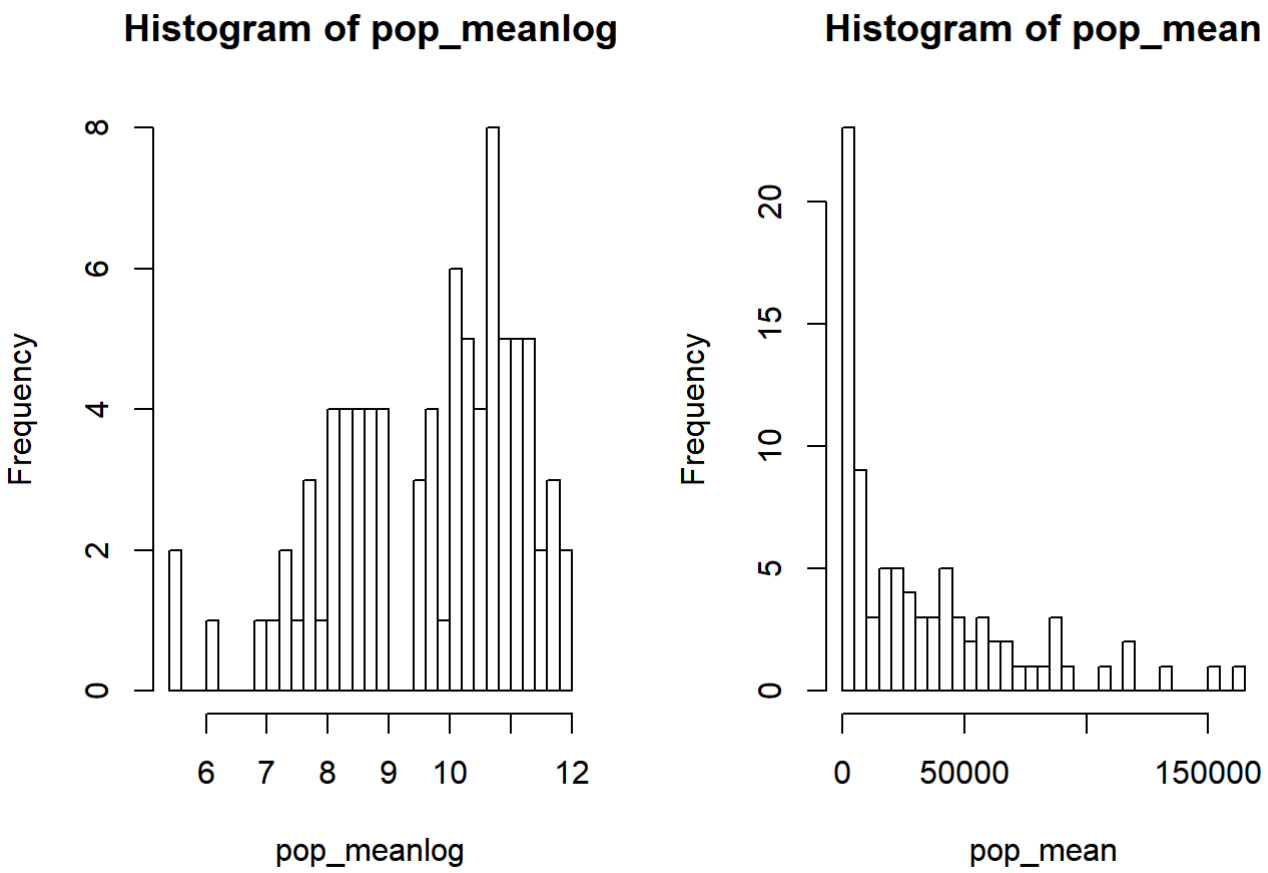
```
ad.test(pop_meansqrt)
```

```
##
##  Anderson-Darling normality test
##
## data:  pop_meansqrt
## A = 1.3226, p-value = 0.00185
```

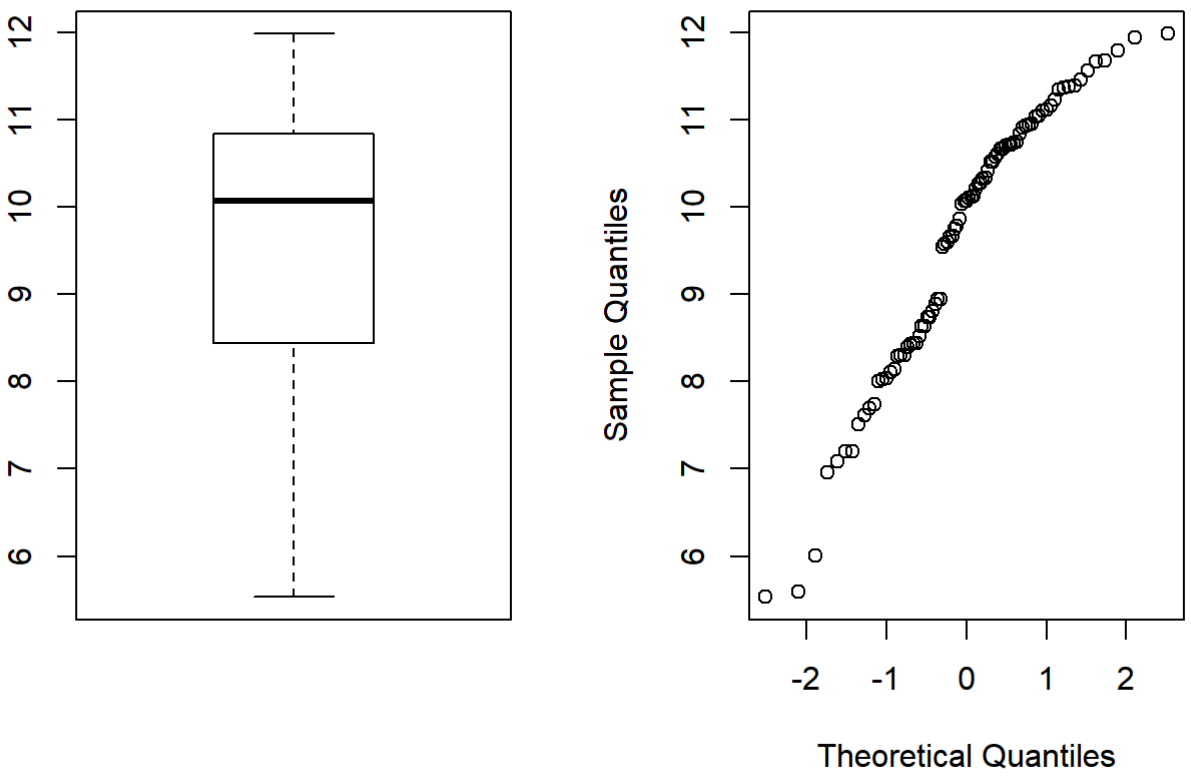
```
lillie.test(pop_meansqrt)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pop_meansqrt
## D = 0.14033, p-value = 0.0002803
```

```
# 히스토그램과 상자그림에서 변환 전보다 쏠린 정도가 완화되었지만
# 아직 왼쪽으로 쏠린 분포를 보이고 있다.
# 왜도도 변환전보다 0에 가까운 값을 갖고 왜도가 0인지에대한 검정에서 p-value가
# 0.05보다 크므로 왜도가 0이라고 할 수 있다.
# 정규성 검정에서 역시 p-value가 0.05보다 작으므로
# 데이터가 정규분포로부터 나왔다고 하기 힘들다.

# log 분석
windows()
par(mfrow=c(1,2))
hist(pop_meanlog,breaks = 40)
hist(pop_mean,breaks = 40)
```




```
# 오히려 로그 변환은 기존보다 오른쪽으로 쏠리는 모양을 보인다.
windows()
par(mfrow=c(1,2))
boxplot(pop_meanlog)
qqnorm(pop_meanlog)
```




```
# 정규확률그림이 오목한 모양으로 보아 오른쪽으로 쏠린 모양을 유추할 수 있다.
skewness(pop_meanlog)
```

```
## [1] -0.6414402
```

```
agostino.test(pop_meanlog)
```

```
##
##  D'Agostino skewness test
##
## data:  pop_meanlog
## skew = -0.64144, z = -2.41980, p-value = 0.01553
## alternative hypothesis: data have a skewness
```

```
# 왜도가 -0.64이고 왜도가 0인지 검정하기 위한 가설검정에서도 p-value가 0.015으로 0.05보다 작으므로
# 귀무가설을 기각할 수 있다. 즉 왜도가 0이라고 할 수 없다.

kurtosis(pop_meanlog)
```

```
## [1] 2.696255
```

```
anscombe.test(pop_meanlog)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pop_meanlog
## kurt = 2.69630, z = -0.37097, p-value = 0.7107
## alternative hypothesis: kurtosis is not equal to 3
```

```
# 첨도가 약 2.7로 3보다 작지만, 첨도가 3인지에 대한 검정에서
#p-value가 0.7로 매우 크므로 첨도가 3이라고 할 수 있다.

ad.test(pop_meanlog)
```

```
##
##  Anderson-Darling normality test
##
## data:  pop_meanlog
## A = 1.549, p-value = 0.0005081
```

```
lillie.test(pop_meanlog)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pop_meanlog
## D = 0.13062, p-value = 0.001077
```

```
# p-value가 0.05보다 작으므로 귀무가설 기각
# 즉 데이터의 분포가 정규분포라 하기 힘들다
```

NA