IE 522 HW02

Yue Ma

9/5/2020

Install the ISLR library in R. Smarket is a data frame from this library that contains returns of S&P 500 in the five year period from 1/10/2001 to 12/30/2005. The following shows the first and last three rows of the data frame:

```
library(ISLR)
n=nrow(Smarket)
Smarket[c(1:3,(n-2):n),]
```

```
##      Year   Lag1   Lag2   Lag3   Lag4   Lag5  Volume  Today Direction
## 1    2001  0.381 -0.192 -2.624 -1.055  5.010 1.19130  0.959        Up
## 2    2001  0.959  0.381 -0.192 -2.624 -1.055 1.29650  1.032        Up
## 3    2001  1.032  0.959  0.381 -0.192 -2.624 1.41120 -0.623      Down
## 1248 2005 -0.955  0.043  0.422  0.252 -0.024 1.54047  0.130        Up
## 1249 2005  0.130 -0.955  0.043  0.422  0.252 1.42236 -0.298      Down
## 1250 2005 -0.298  0.130 -0.955  0.043  0.422 1.38254 -0.489      Down
```

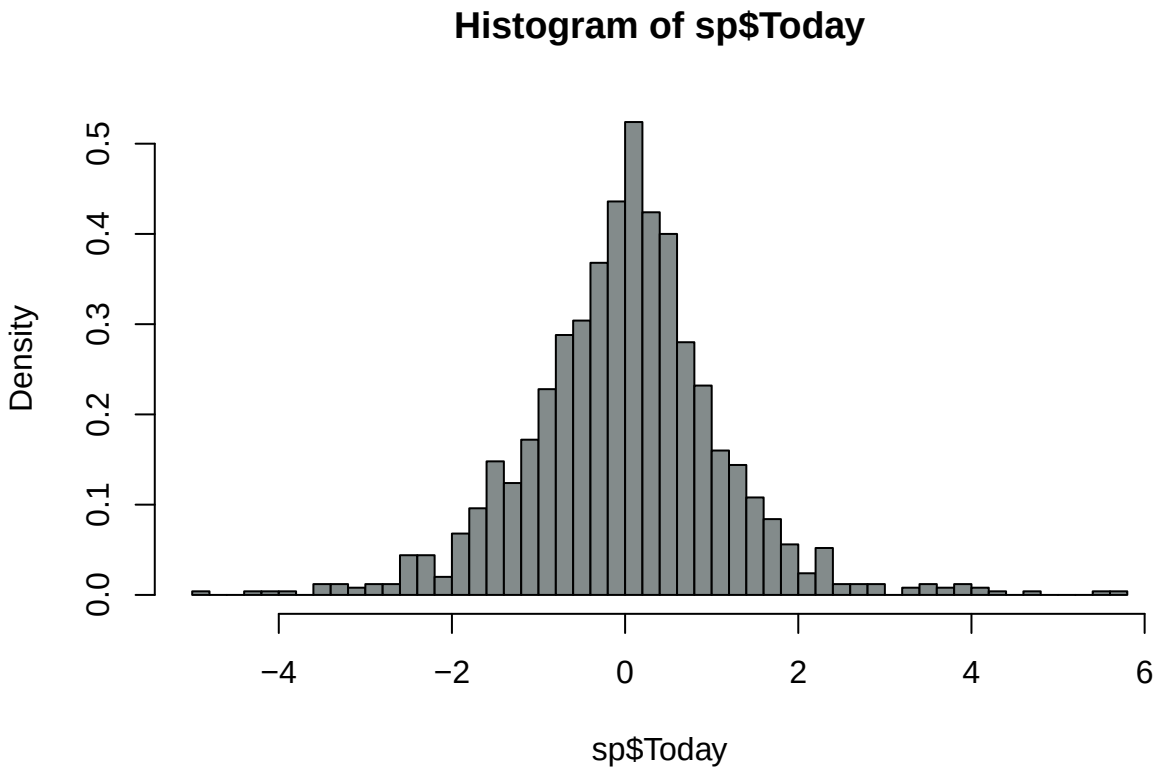
For each date, *Today* is the percentage return of the day. *Direction* indicates whether S&P 500 was going up or down during the day. *Volume* is the trading volume on the previous day (in billions). *Lag1* to *Lag5* are the percentage returns in the previous 5 days. From compass2g, download ISLRSmarketDates.csv and put it in your R working directory. It contains the corresponding dates. Replace the first column of Smarket by these dates.

```
setwd("/home/mavy731/Documents/IE522/HW2")
dates=read.csv("ISLRSmarketDates.csv",header=TRUE)
sp=data.frame(dates,Smarket[,-1])
n=nrow(sp)
sp[c(1:3,(n-2):n),]
```

```
##            Date   Lag1   Lag2   Lag3   Lag4   Lag5  Volume  Today Direction
## 1     1/10/2001  0.381 -0.192 -2.624 -1.055  5.010 1.19130  0.959        Up
## 2     1/11/2001  0.959  0.381 -0.192 -2.624 -1.055 1.29650  1.032        Up
## 3     1/12/2001  1.032  0.959  0.381 -0.192 -2.624 1.41120 -0.623      Down
## 1248 12/28/2005 -0.955  0.043  0.422  0.252 -0.024 1.54047  0.130        Up
## 1249 12/29/2005  0.130 -0.955  0.043  0.422  0.252 1.42236 -0.298      Down
## 1250 12/30/2005 -0.298  0.130 -0.955  0.043  0.422 1.38254 -0.489      Down
```

1. (1 point) Construct a histogram for *Today*. Make the vertical axis density instead of frequency. Set the number of bins to 50.

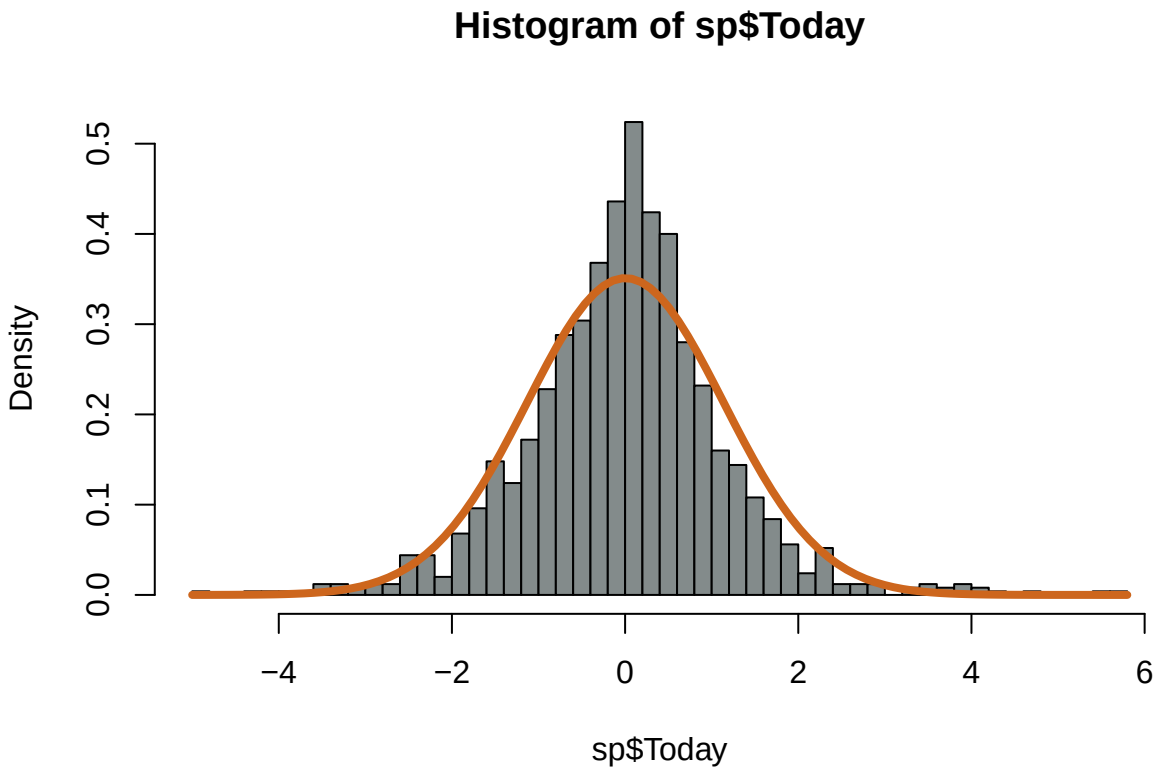
```
hist(sp$Today,breaks=50,col='azure4',prob=TRUE)
```



2. (1 point) Add a normal fit to the histogram you obtain in #1. Is the normal distribution fitting the peak well?

The normal distribution can approximately fit the shape of the histogram, but does not fit the peak well.

```
hist(sp$Today,breaks=50,col='azure4',prob=TRUE)
TdMean=mean(sp$Today, na.rm=TRUE)
TdStd=sd(sp$Today, na.rm=TRUE)
curve(dnorm(x,TdMean,TdStd),add=TRUE,col="chocolate3",lwd=4)
```



3. (1 point) Use the dlaplace(x,μ,b) function in the VGAM library (dlaplace(x,μ,b) is the pdf of a Laplace distribution with location parameter μ and scale parameter b), add a Laplace fit to the histogram you obtain in #1. Is the Laplace distribution fitting the peak better than normal?

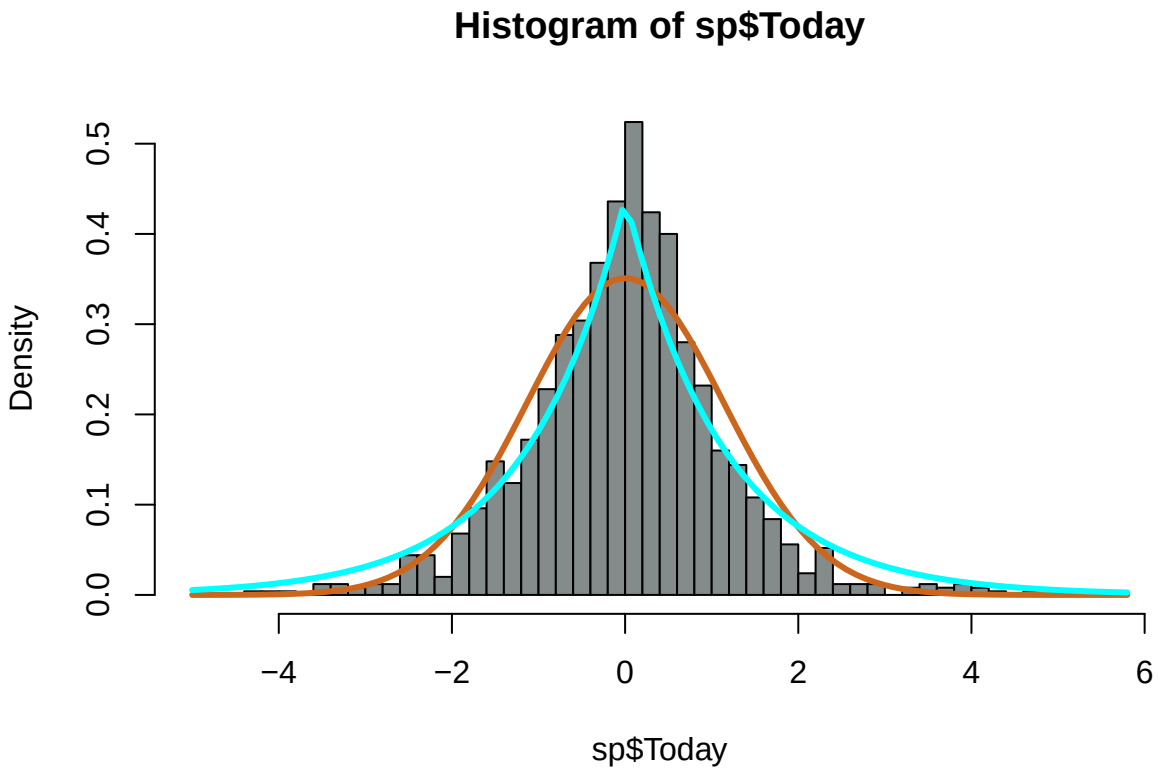
Yes, from the diagram, the Laplace distribution fits the peak better.

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

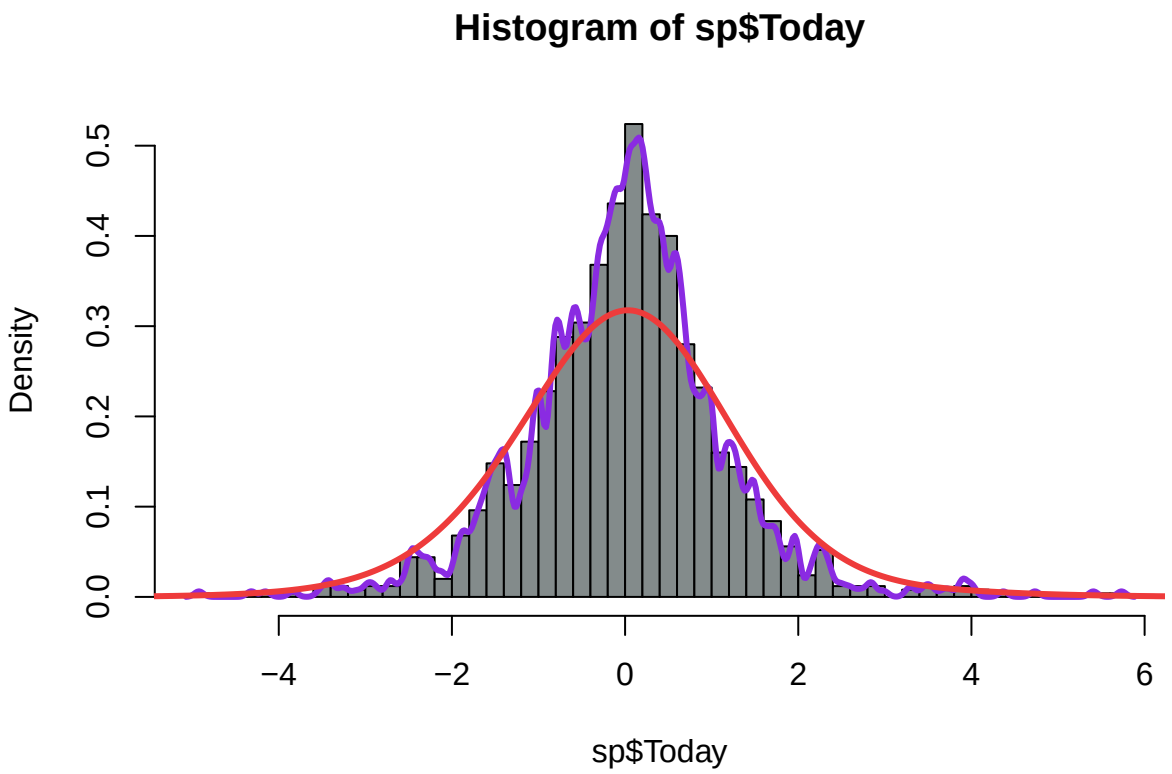
```
hist(sp$Today,breaks=50,col='azure4',prob=TRUE)
TdMean=mean(sp$Today, na.rm=TRUE)
TdStd=sd(sp$Today, na.rm=TRUE)
curve(dnorm(x,TdMean,TdStd),add=TRUE,col="chocolate3",lwd=3)
curve(dlaplace(x,TdMean,TdStd),add=TRUE,col='cyan',lwd=3)
```



4. (1 point) Add KDEs to the histogram you obtain in #1 using 1/4 of the default bandwidth and four times the default bandwidth. Which fits the data better? Which is less oscillating and smoother?

The KDE with bandwidth adjusted by 0.25 fits the histogram better, while the KDE with bandwidth adjusted by 4 is smoother and less oscillating.

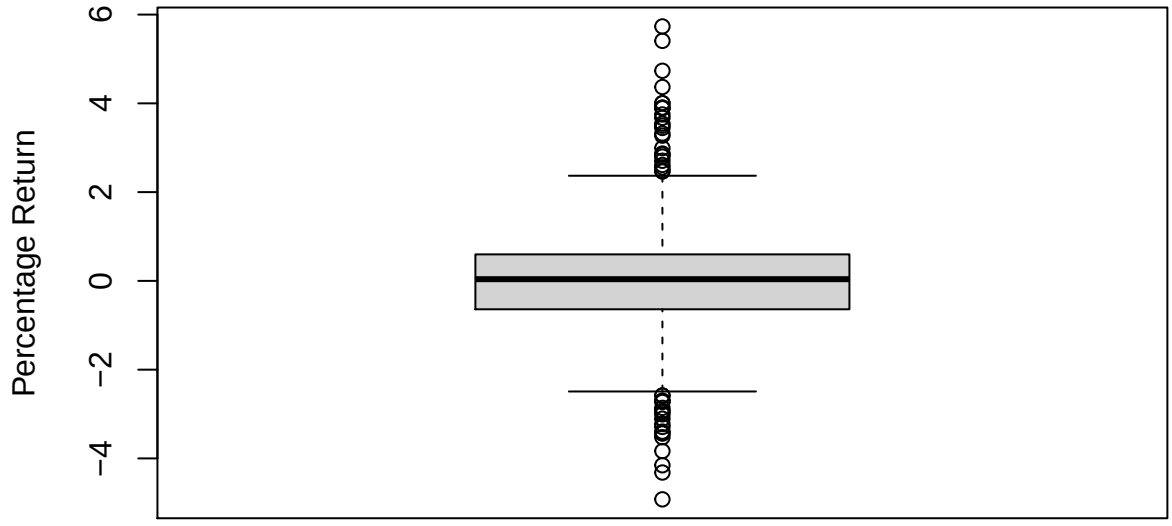
```
hist(sp$Today,breaks=50,col='azure4',prob=TRUE)
lines(density(sp$Today,bw="nrd0",adjust=0.25,na.rm=TRUE),col="blueviolet",lwd=3)
lines(density(sp$Today,bw="nrd0",adjust=4,na.rm=TRUE),col="brown2",lwd=3)
```



5. (1 point) Construct a boxplot for *Today*. Investigate the most negative outlier. What caused this extreme negative return?

From the boxplot, the outliers can be seen as legitimate for their continuity.
Following the worst terrorist attack in U.S history on Sep.11, NYSE has been shut down for one week and was re-open on 9/17/2001. The extreme drop of the S&P500 showed the market shock toward that tragic event.

```
set.seed(2)
boxplot(sp$Today,ylab='Percentage Return')
```




```
outliers=boxplot(sp$Today,plot=FALSE)$out
sp[sp$Today %in% min(outliers),]
```

```
##          Date  Lag1   Lag2   Lag3   Lag4   Lag5 Volume  Today Direction
## 169 9/17/2001 0.623 -1.864 -2.239 -0.106 -0.056 1.2766 -4.922      Down
```

6. (1 point) Report the correlation matrix for all the variables excluding Date and Direction. Do you see any strong linear relationship?

No strong linear relationship found.

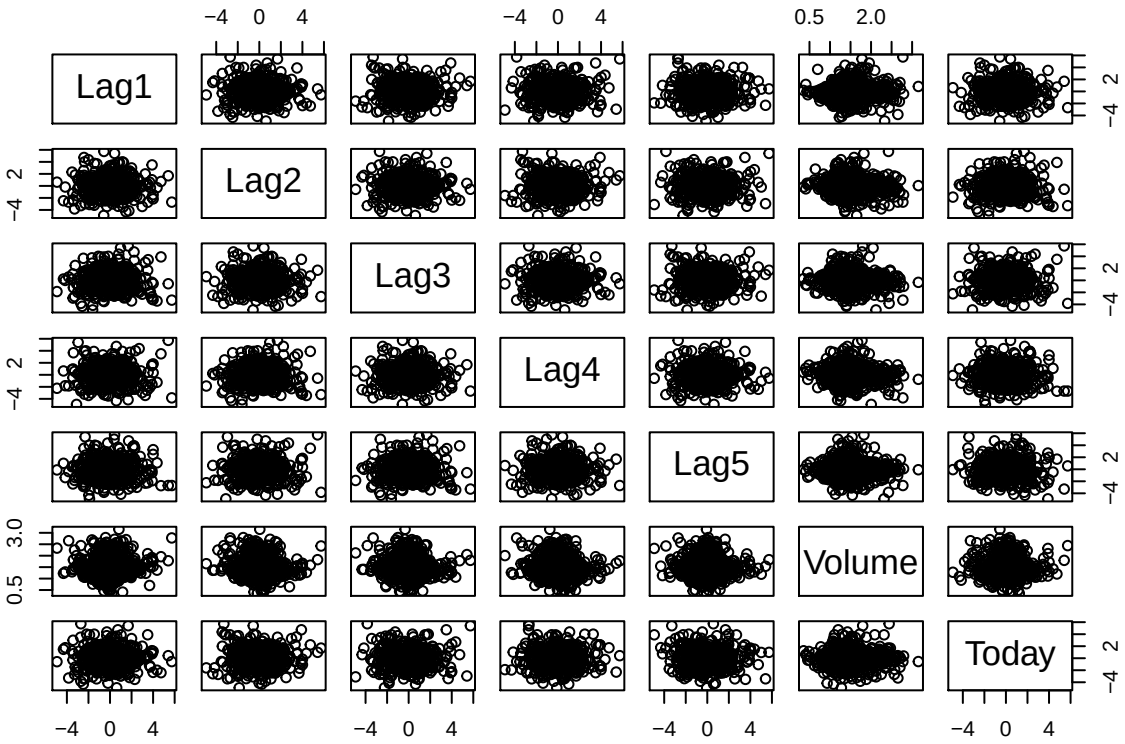
```
round(cor(sp[,2:8],use="complete.obs"),digits=2)
```

```
##         Lag1  Lag2  Lag3  Lag4  Lag5 Volume Today
## Lag1    1.00 -0.03 -0.01  0.00 -0.01   0.04 -0.03
## Lag2   -0.03  1.00 -0.03 -0.01  0.00  -0.04 -0.01
## Lag3   -0.01 -0.03  1.00 -0.02 -0.02  -0.04  0.00
## Lag4    0.00 -0.01 -0.02  1.00 -0.03  -0.05 -0.01
## Lag5   -0.01  0.00 -0.02 -0.03  1.00  -0.02 -0.03
## Volume  0.04 -0.04 -0.04 -0.05 -0.02   1.00  0.01
## Today  -0.03 -0.01  0.00 -0.01 -0.03   0.01  1.00
```

7. (1 point) Construct a scatterplot matrix for all variables excluding Date and Direction. Do you see any significant linear or nonlinear relationship?

No significant linear or nonlinear relationship found.

```
pairs(sp[,2:8])
```



8. (1 point) Construct a time series plot for Volume. What do you observe near the end of each year?

By the end of each year, the trading volume of the market usually will have a big drop.

```
 library(xts)
```

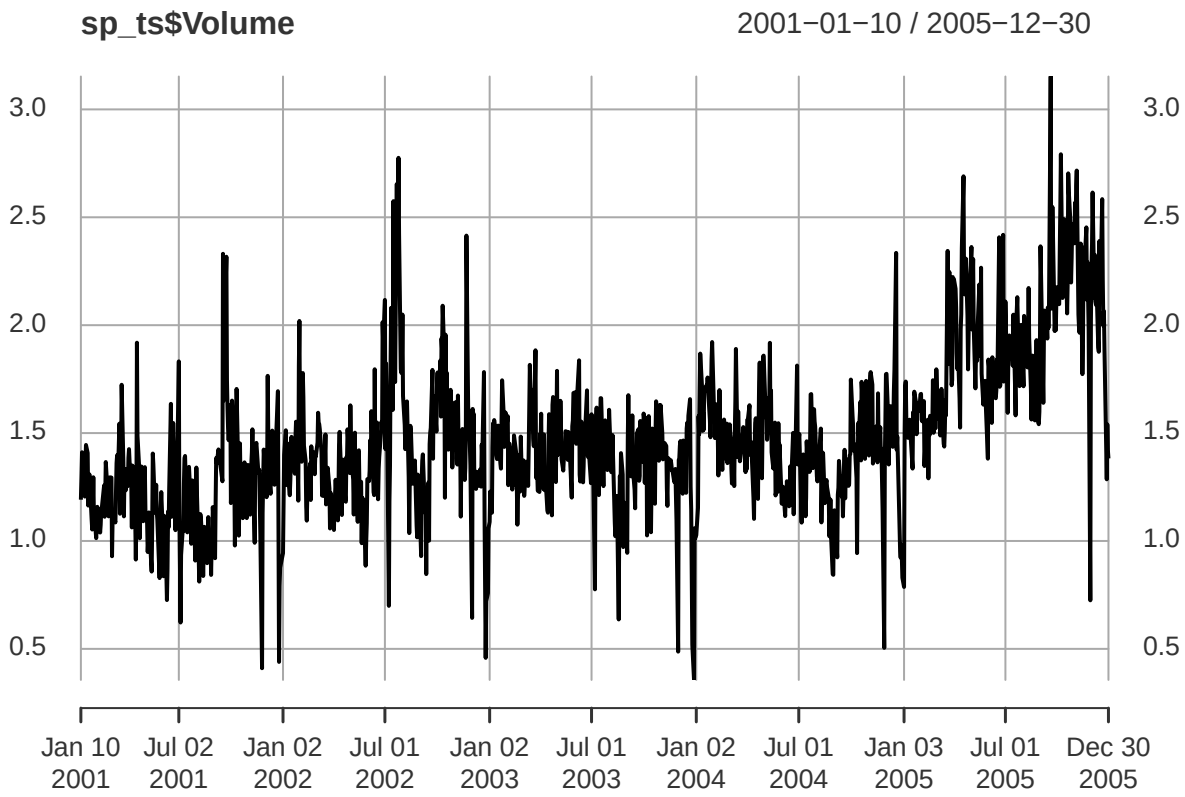
```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
 library(zoo)
d<-as.Date(sp$Date,"%m/%d/%Y")
sp_ts <-xts(sp[2:8],order.by=d)

plot(sp_ts$Volume)
```



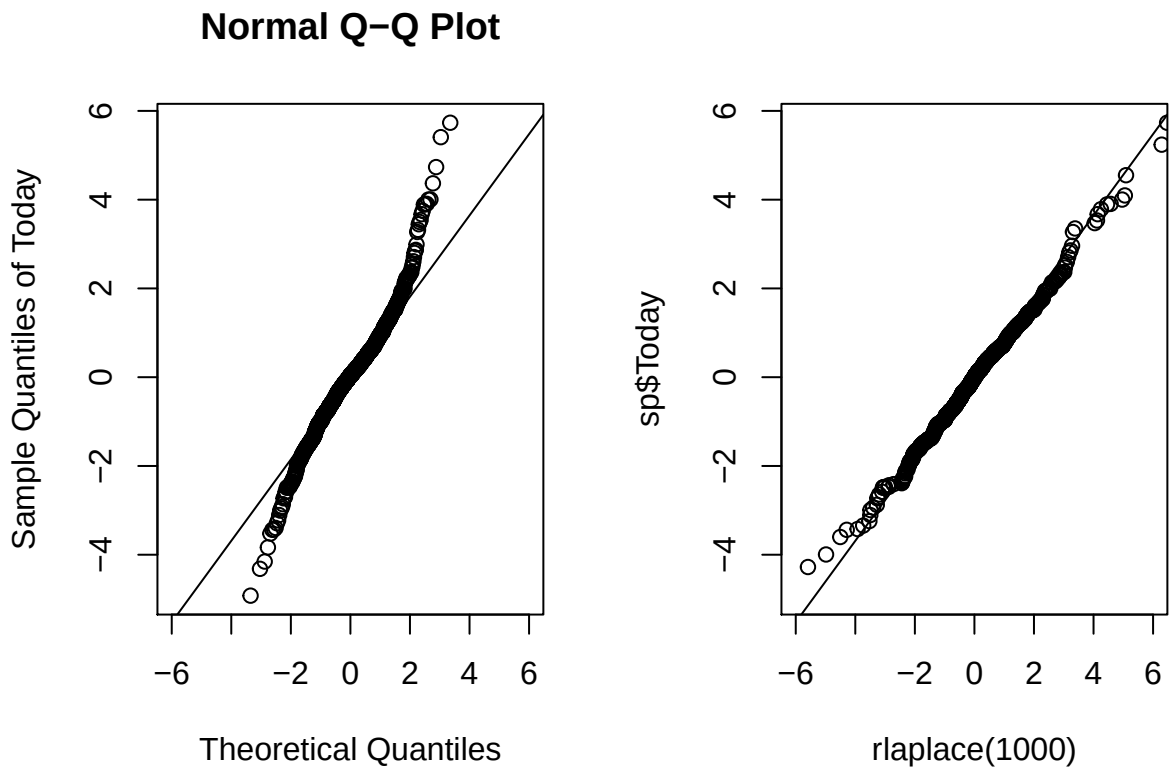
9. (1 point) Construct a normal plot and a Laplace plot for *Today*, side by side. Make sure that the ranges of the x-axis are the same for both plots. Between the normal distribution and the Laplace distribution, which is fitting the data better?

From the two plots, the Laplace distribution fits the data better for being more close to a straight line.

```
normaldata<-rnorm(1000,mean=TdMean,sd=TdStd)
par(mfrow=c(1,2))

qqnorm(sp$Today,ylab="Sample Quantiles of Today",xlim=c(-6,6))
qqline(sp$Today)
```

```
qqplot(rlaplace(1000),sp$Today,xlim=c(-6,6))
qqline(sp$Today)
```



10. (1 point) Given a random sample $\{X_1, \cdots, X_n\}$ from a Laplace distribution with location parameter μ and scale parameter b, using the method of moments, find the point estimator for μ. Show that the corresponding point estimator for b is

$$\hat{b} = \frac{1}{\sqrt{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

Point estimator for μ: Let

$$\lambda = X - \mu$$

$$\exp(X) = \exp(\lambda + \mu) = \exp(\lambda) + \mu$$

$$= \frac{1}{2b}\int_{-\infty}^{0} \lambda \exp(\frac{-\lambda}{b})d\lambda + \frac{1}{2b}\int_{0}^{\infty} \lambda \exp(\frac{\lambda}{b})d\lambda + \mu = \mu$$

$$\exp(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\hat{\mu} = \bar{X}_n$$

$$\exp(X^2) = \exp((\lambda + \mu)^2) = exp(\lambda^2 + 2\mu\lambda + \lambda^2) = \exp(\lambda^2) + \mu^2$$

$$= \frac{1}{2b}\int_{-\infty}^{0} \lambda^2 \exp(\frac{-\lambda}{b})d\lambda + \frac{1}{2b}\int_{0}^{\infty} \lambda^2 \exp(\frac{\lambda}{b})d\lambda + \mu^2$$

$$= 2b^2 + \mu^2$$

$$Var(X) = \exp(X^2) - (\exp(X))^2 = 2b^2$$

$$\hat{b} = \frac{1}{\sqrt{2}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$$