Exploratory Data Analysis

Reading and cleaning data

First, we input our stock data.

Our stock data consists of the following indices between 2000 and 2021:

- S&P500
- NASDAQ
- NYSE100

```
setwd("..")
sp<-read.csv("Data/sp500.csv")
ny<-read.csv("Data/nyse.csv")
nas<-read.csv("Data/nasdaq.csv")
```

Now we will change the 'caldt' column to the Date format in order to plot the time series for each index:

```
sp$caldt<-as.Date(sp$caldt, format="%d/%m/%Y")
ny$caldt<-as.Date(ny$caldt, format="%d/%m/%Y")
nas$caldt<-as.Date(nas$caldt, format="%d/%m/%Y")

str(sp)
```

```
## 'data.frame':    5032 obs. of  2 variables:
##  $ caldt : Date, format: "2001-01-02" "2001-01-03" ...
##  $ spindx: num  1283 1348 1333 1298 1296 ...
```

```
str(ny)
```

```
## 'data.frame':    5284 obs. of  2 variables:
##  $ caldt : Date, format: "2000-01-03" "2000-01-04" ...
##  $ spindx: num  1455 1399 1402 1403 1441 ...
```

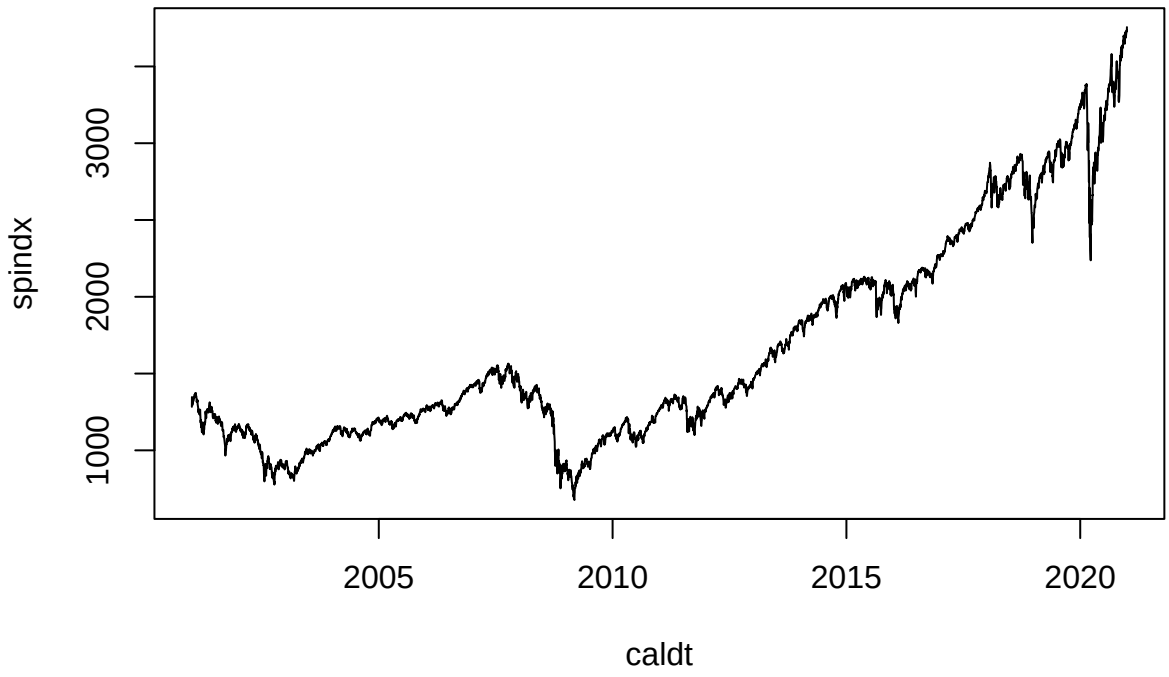
```
str(nas)
```

```
## 'data.frame':    5284 obs. of  2 variables:
##  $ caldt : Date, format: "2000-01-03" "2000-01-04" ...
##  $ ncindx: num  4131 3902 3878 3727 3883 ...
```

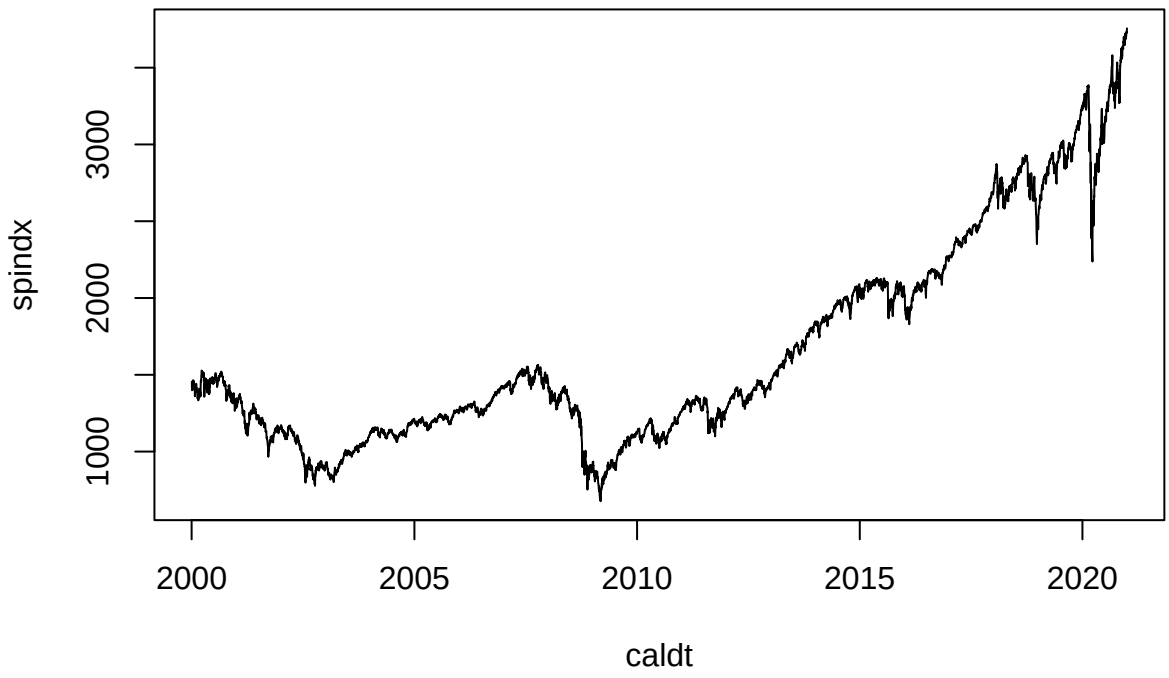
Initial Plots

We will start off by plotting a basic time series for each index to get an idea of what our data looks like:

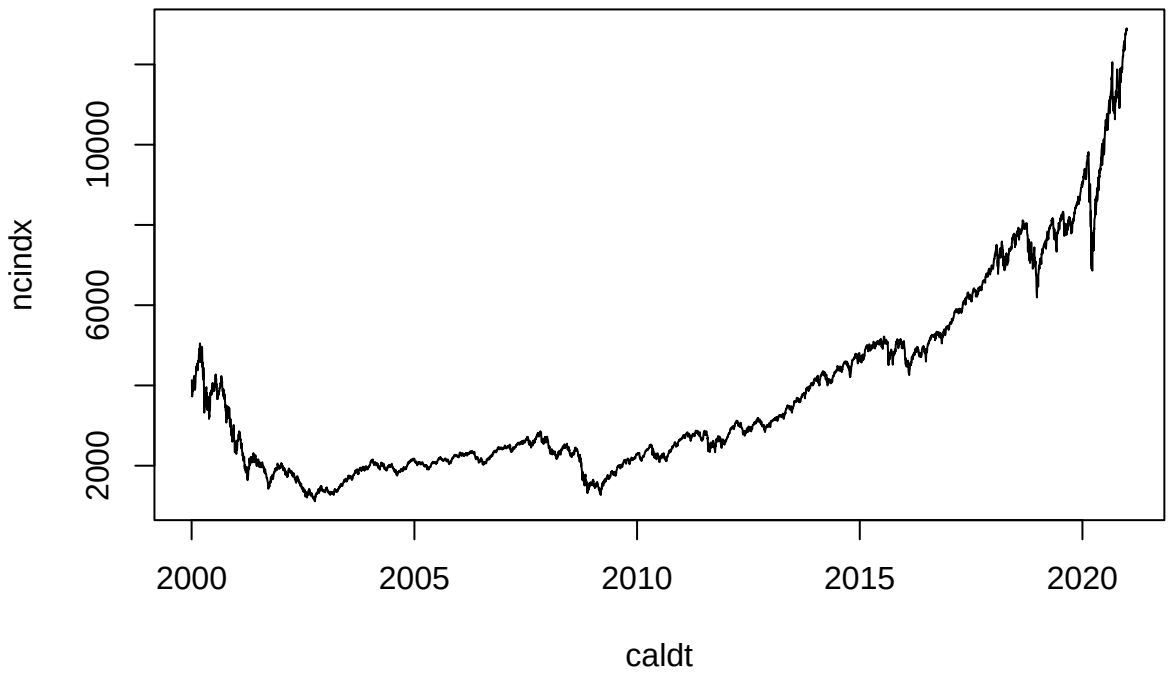
```
plot(sp, type='l')
```




```
plot(ny, type='l')
```



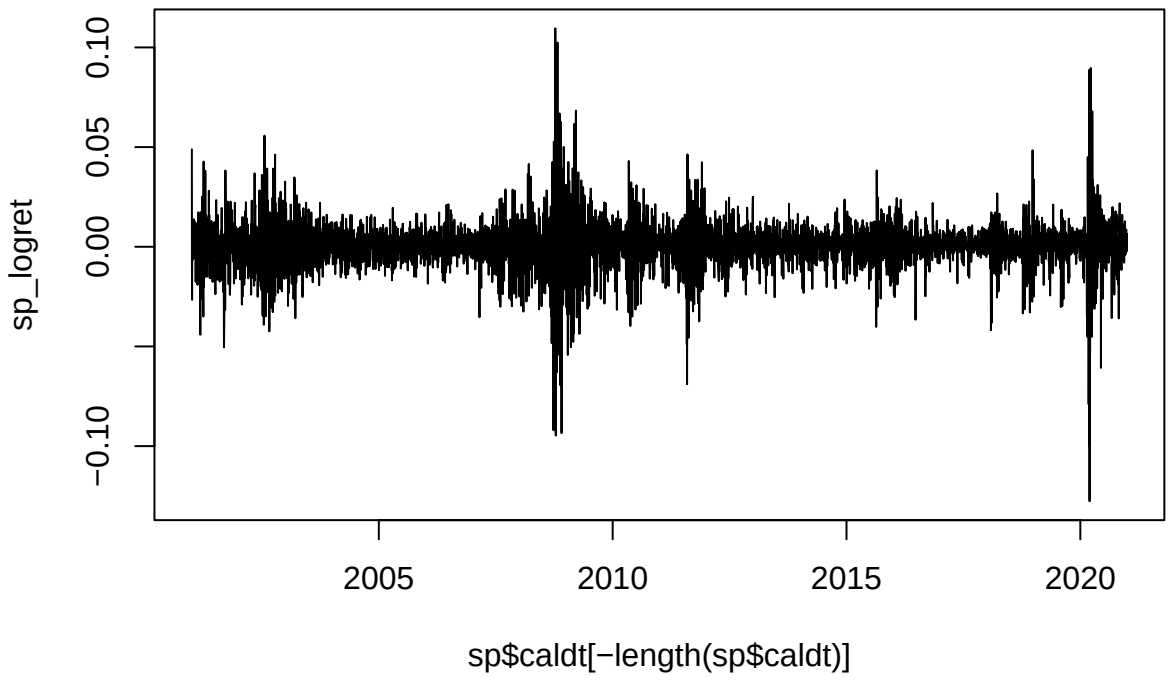

```
plot(nas, type='l')
```



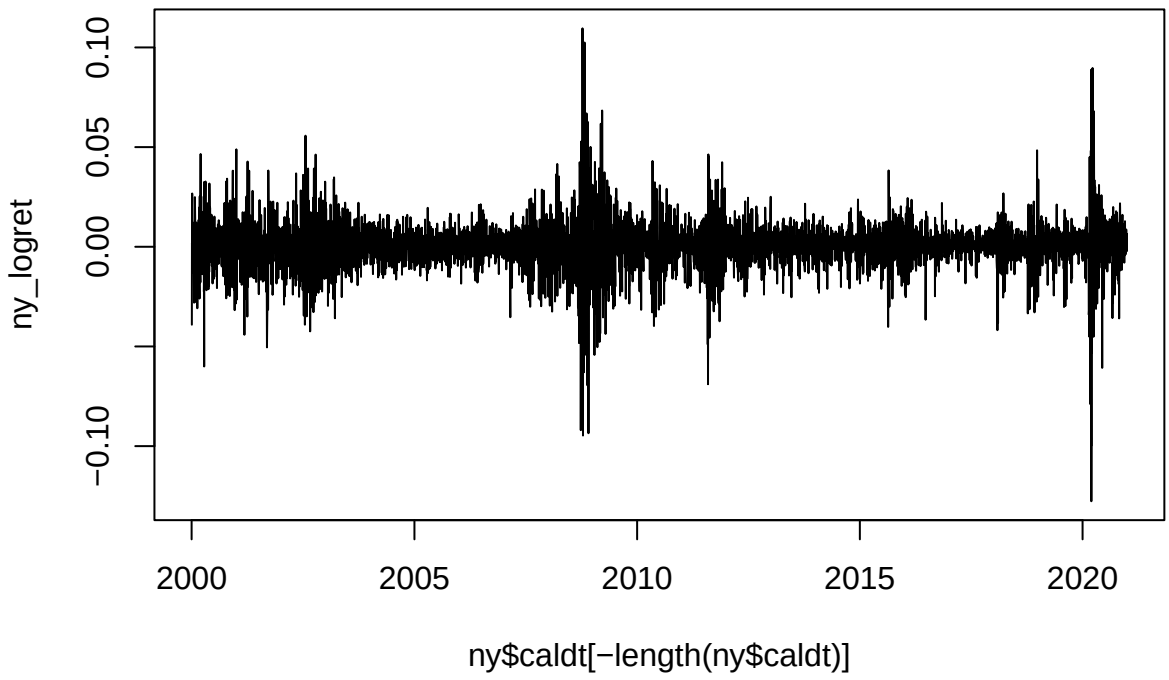
They all follow the same basic pattern, which is what we would expect, with the iconic fall in stock-price during the 2008-2009 period of the 'Great Recession'.

However, the stock price directly does not give us much information. Instead, we will take at the **daily log stock returns**.

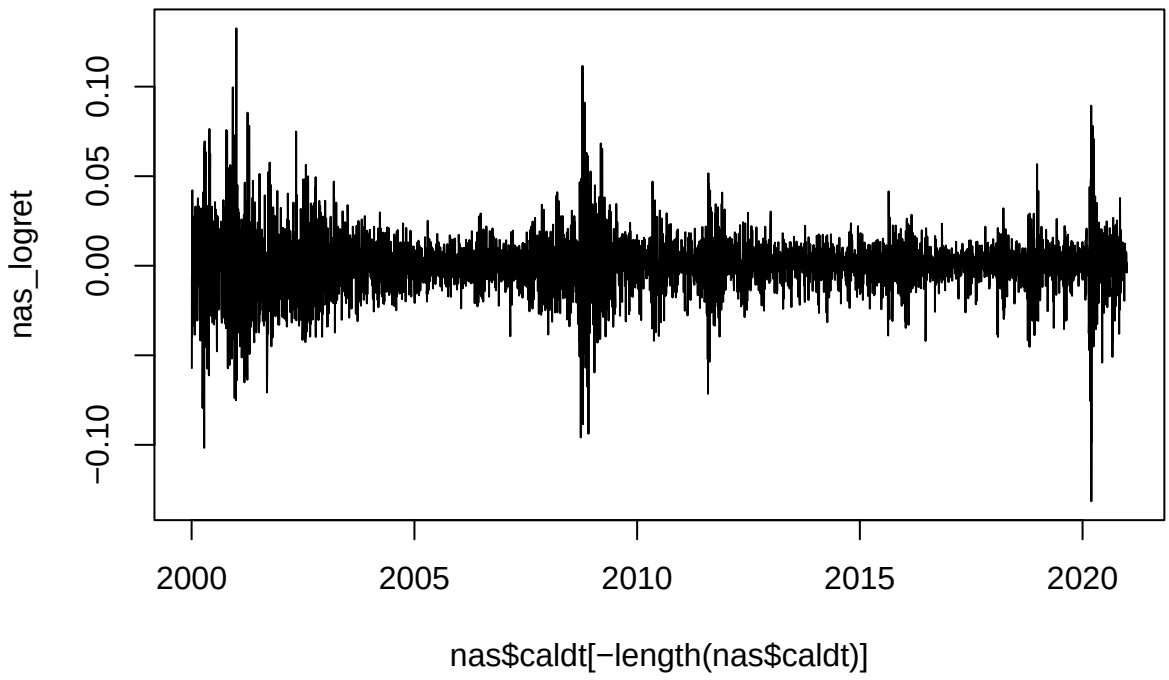
```
sp_logret <- diff(log(sp$spindx))
ny_logret <- diff(log(ny$spindx))
nas_logret <- diff(log(nas$ncindx))

plot(sp$caldt[-length(sp$caldt)],sp_logret,type='l')
```




```
plot(ny$caldt[-length(ny$caldt)],ny_logret, type='l')
```



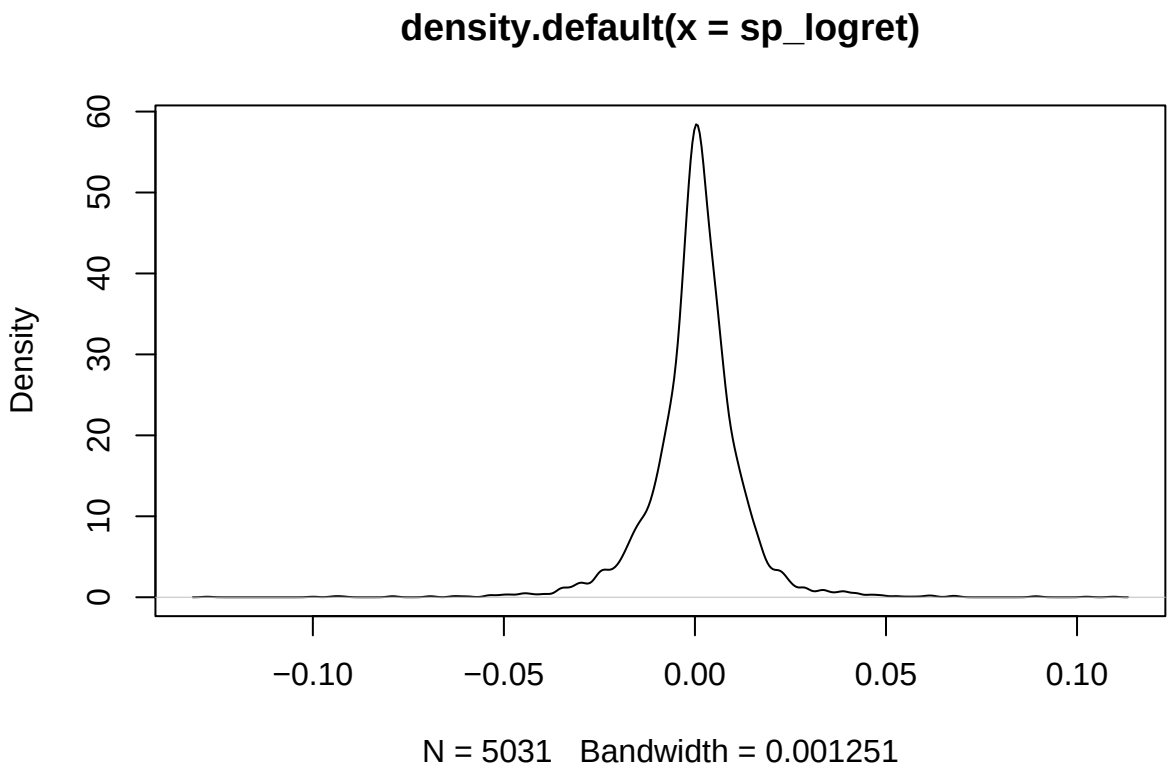

```
plot(nas$caldt[-length(nas$caldt)],nas_logret, type='l')
```



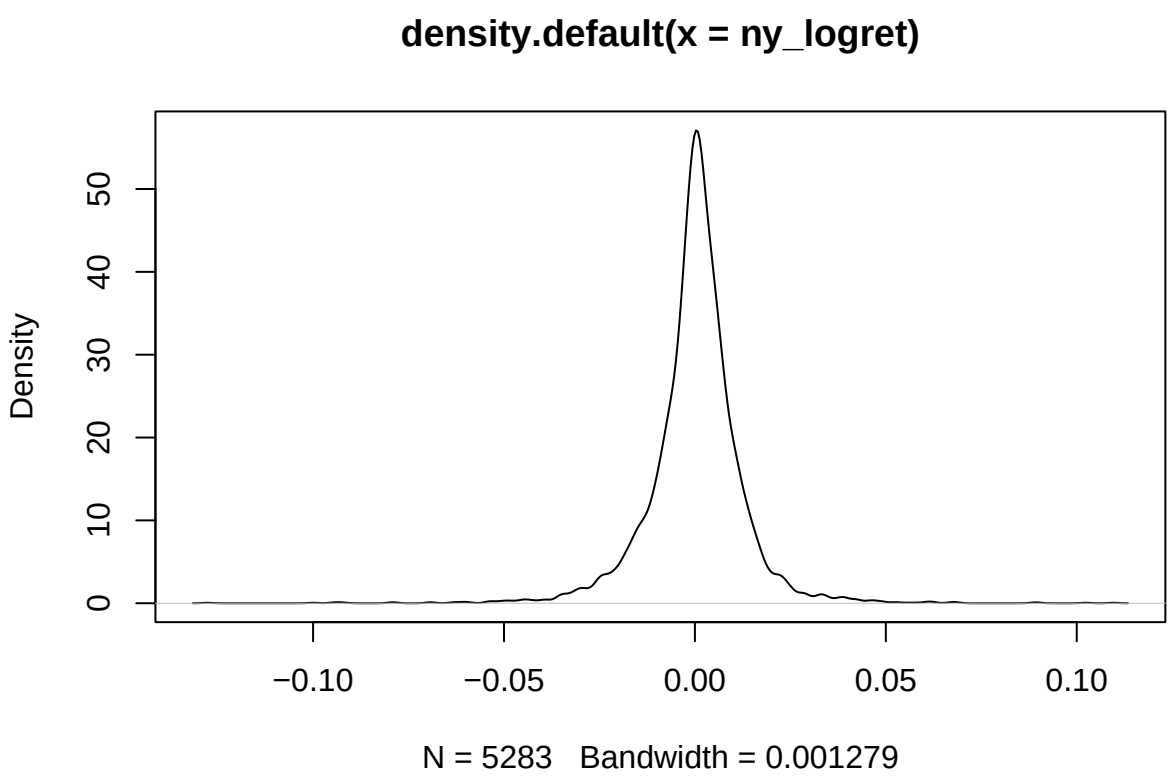
We can see that the returns average around 0% with very high variability during 2008-2009 (caused by the Great Recession) and during 2020 (caused by COVID-19).

Let us now plot the density of the returns to try to understand the distribution which will be helpful when we will try to model the returns later one:

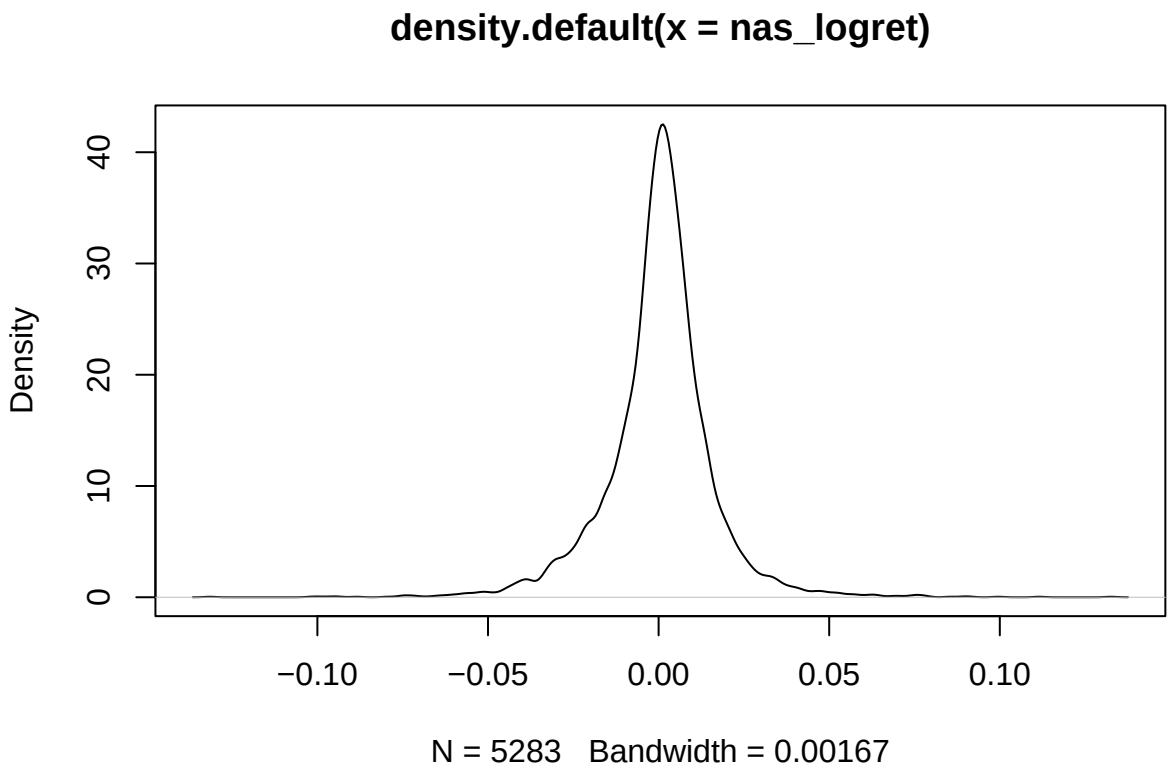
```
plot(density(sp_logret))
```




```
plot(density(ny_logret))
```

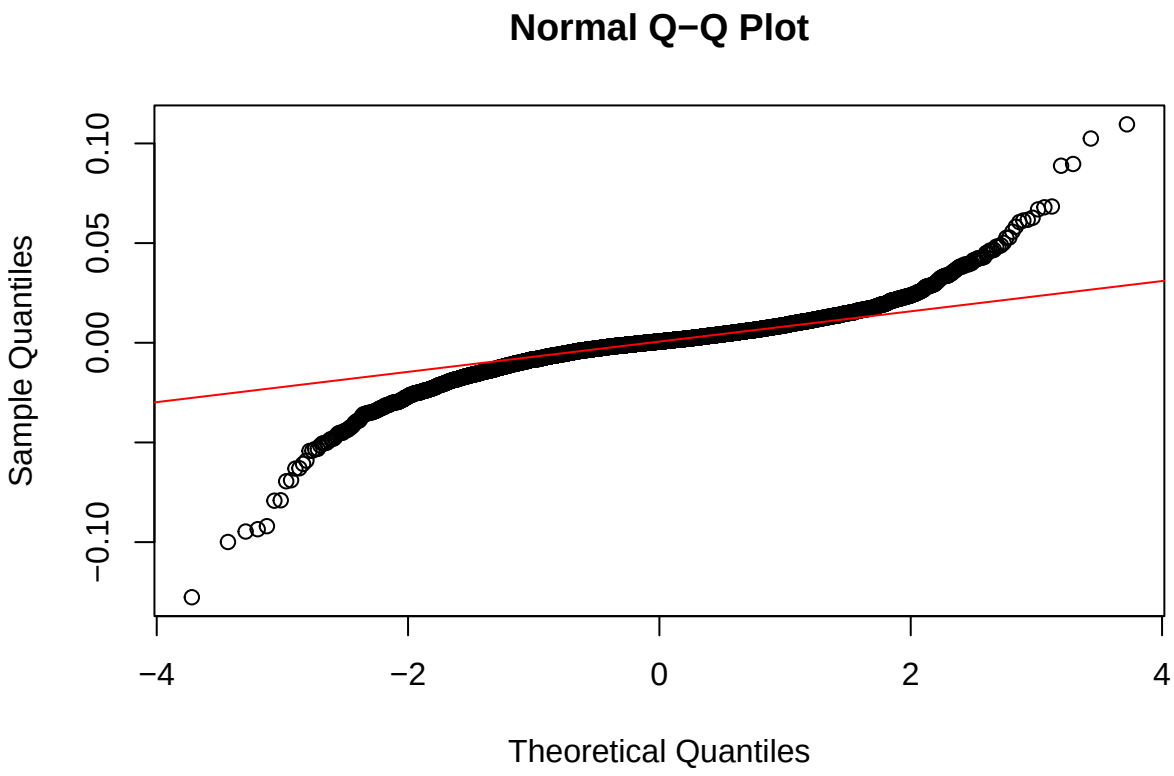



```
plot(density(nas_logret))
```

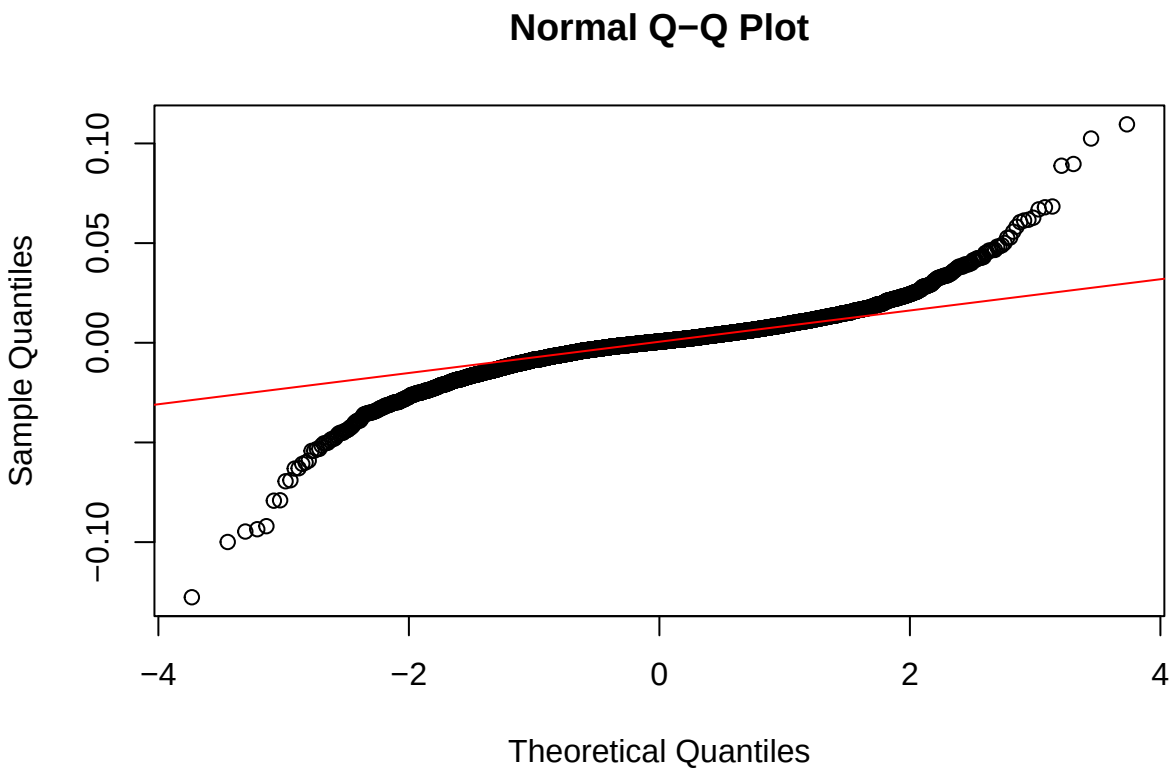


The returns look like they follow a normal distribution. So, we will make qq-plots to further confirm this:

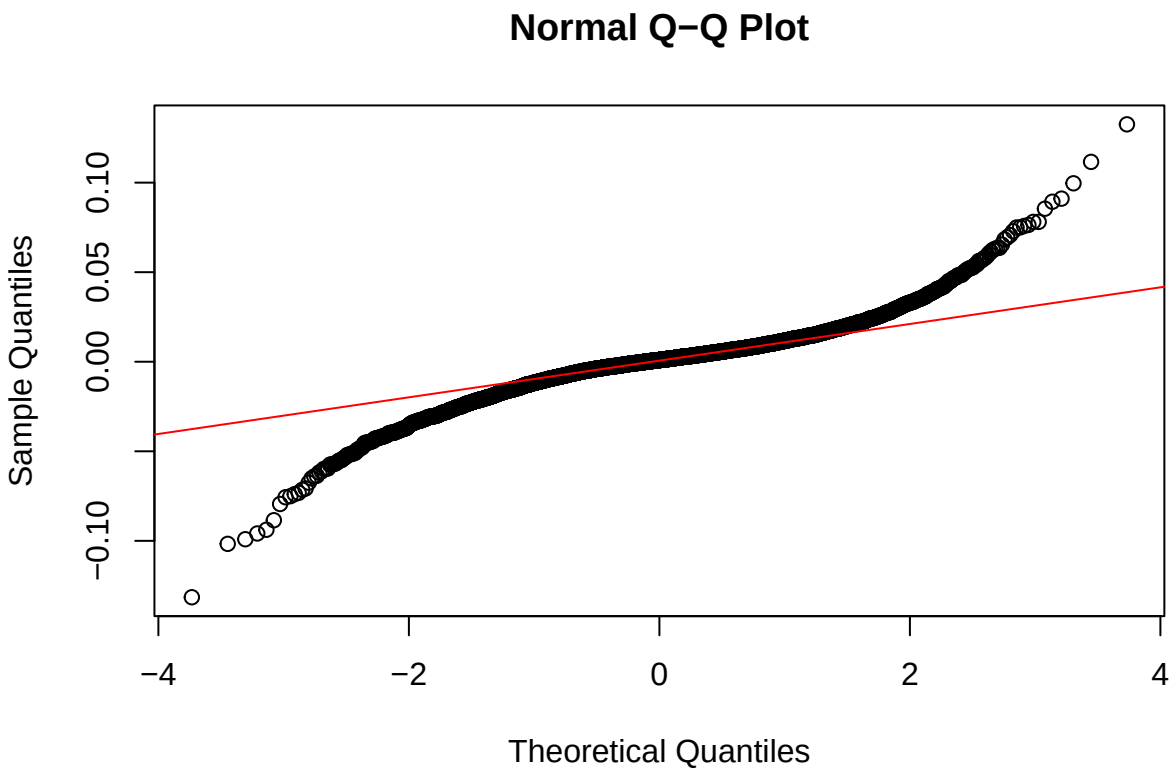
```
qqnorm(sp_logret)
qqline(sp_logret,col='red')
```




```
qqnorm(ny_logret)
qqline(ny_logret,col='red')
```

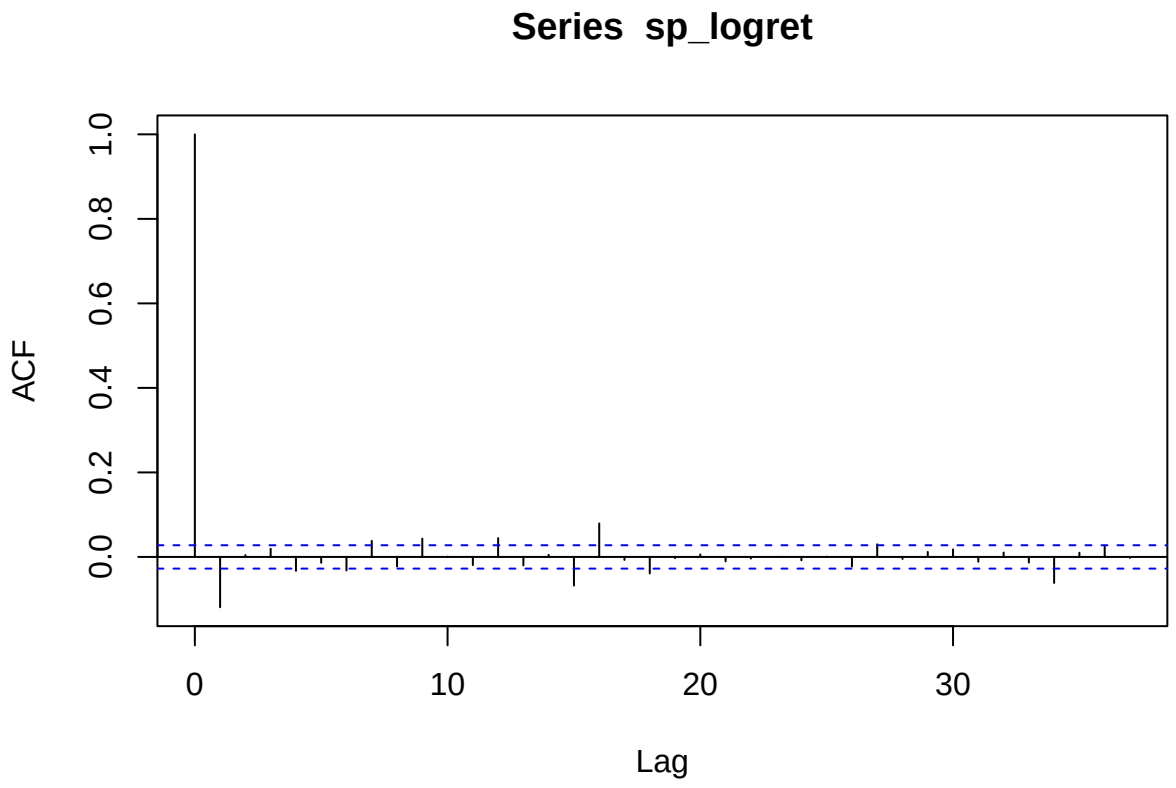



```
qqnorm(nas_logret)
qqline(nas_logret,col='red')
```

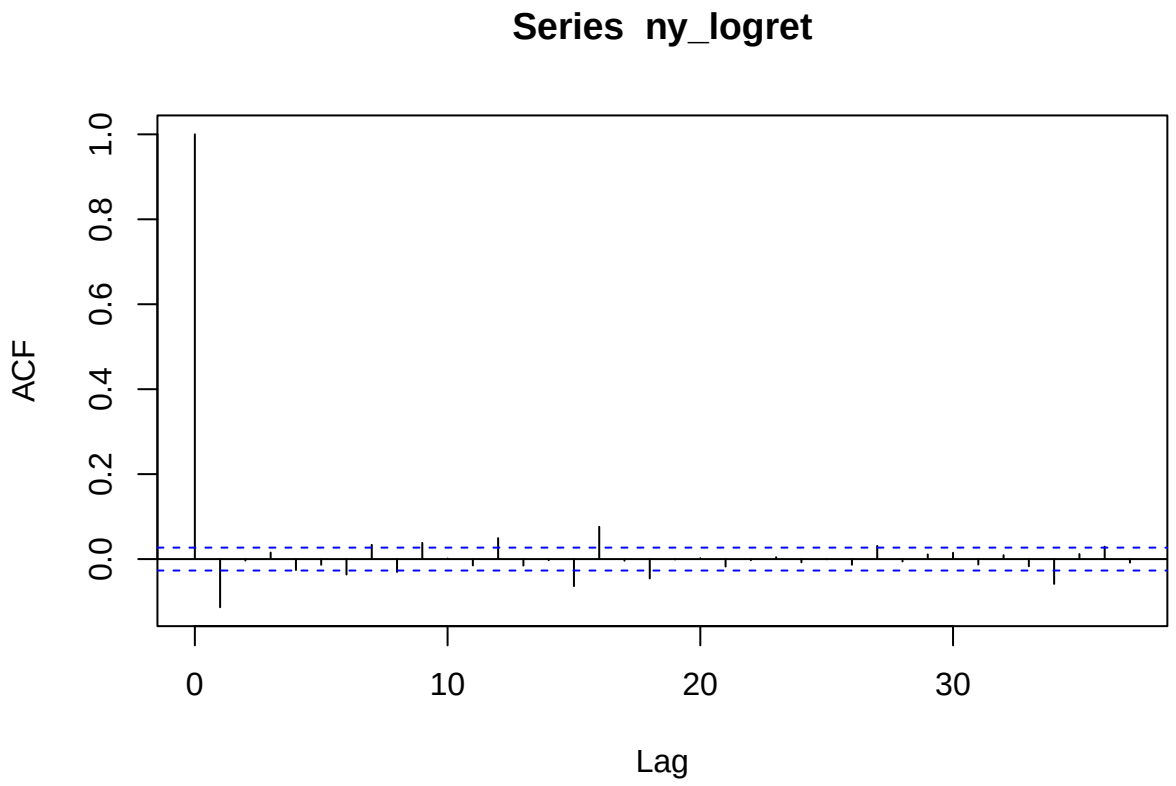


The log-returns have much heavier tails than the normal distribution, which suggests that it might follow a Student's t-distribution. Let's explore this further:

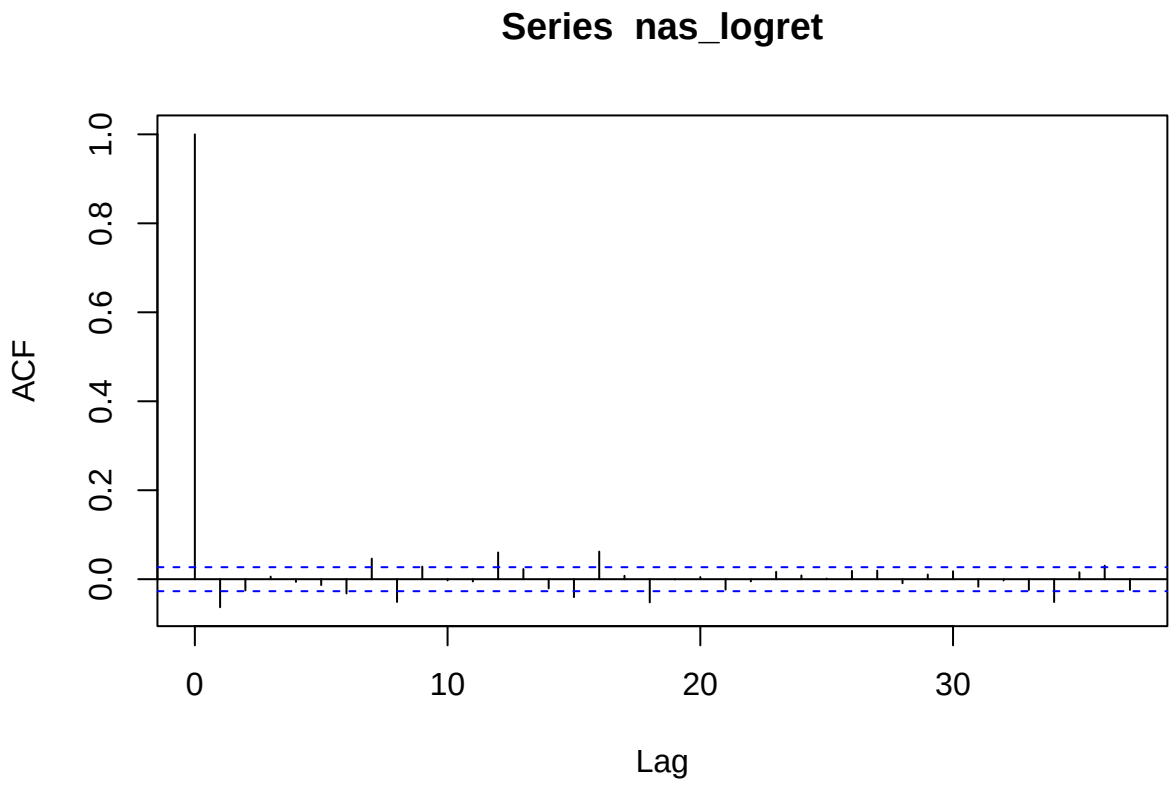
```
acf(sp_logret)
```



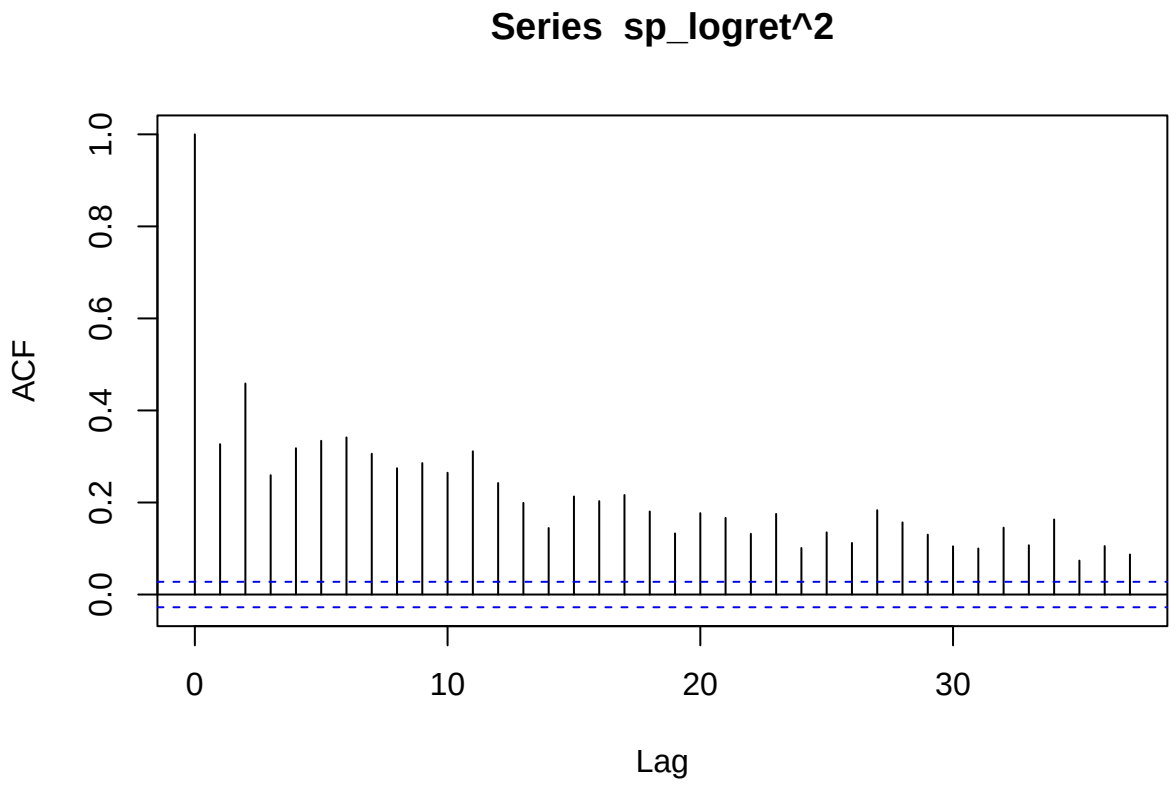

```
acf(ny_logret)
```



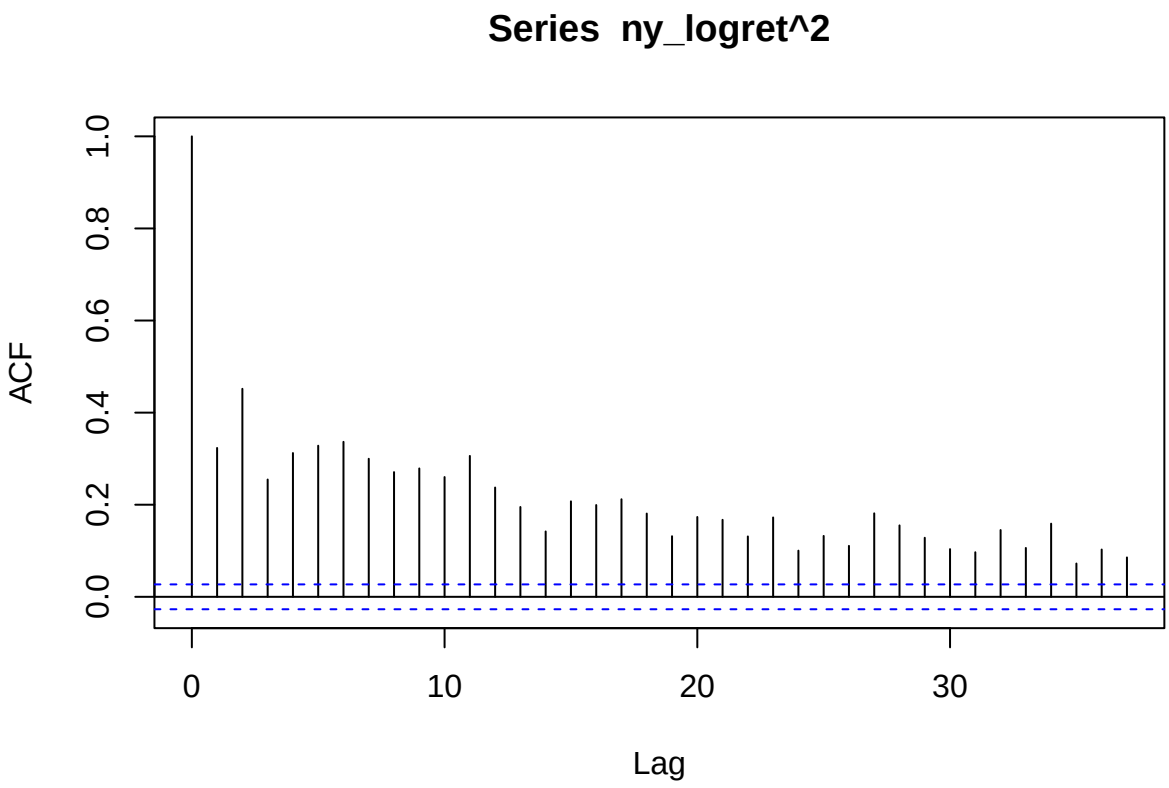

```
acf(nas_logret)
```



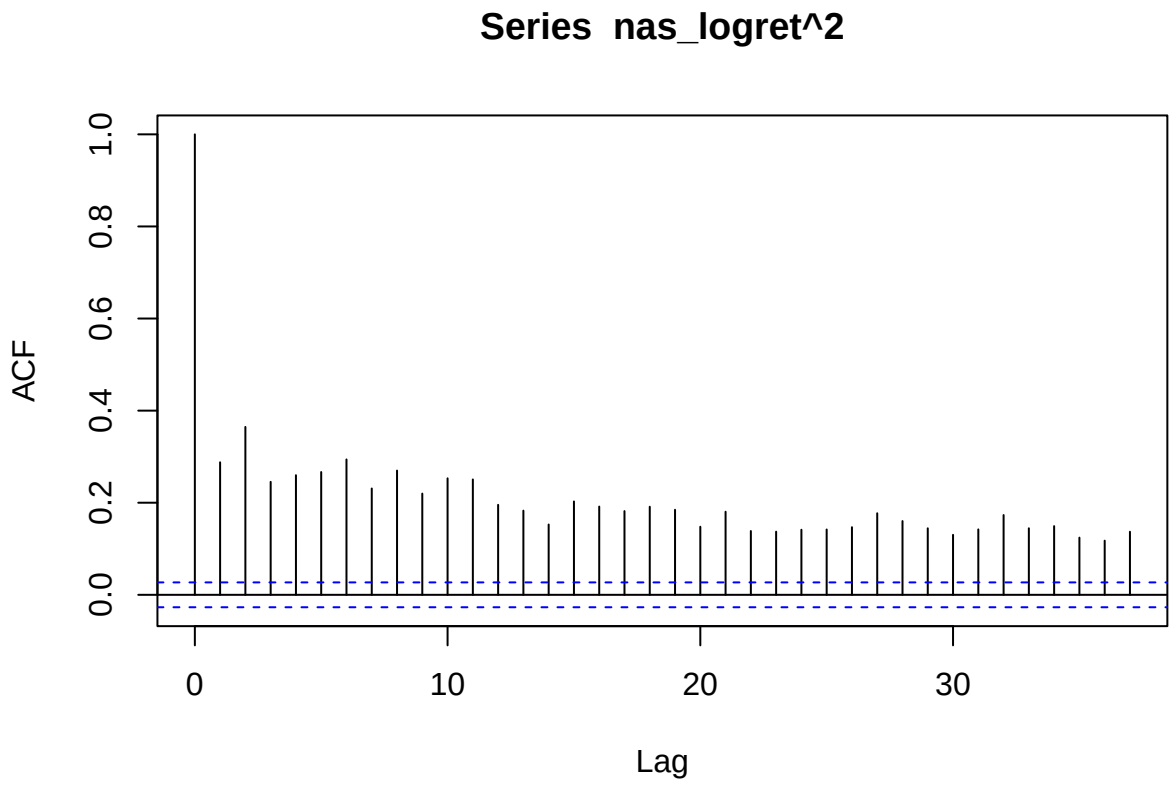

```
acf(sp_logret^2)
```



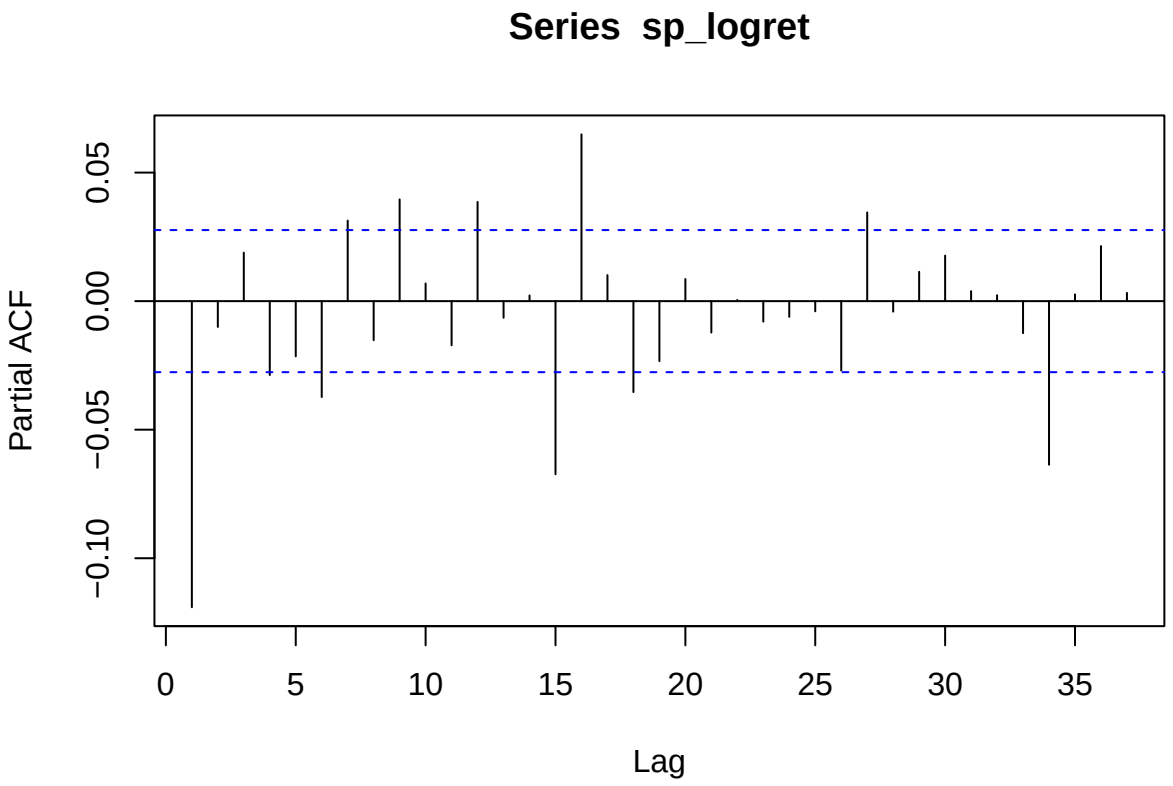

```
acf(ny_logret^2)
```



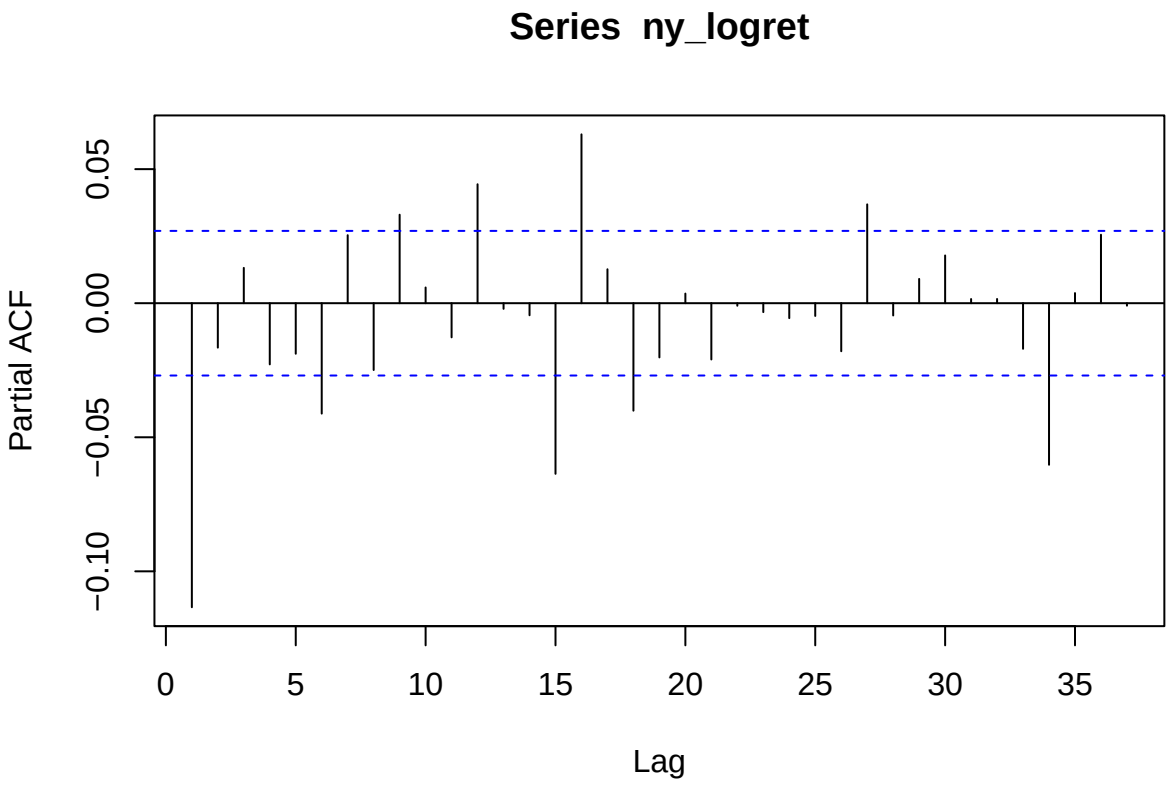

```
acf(nas_logret^2)
```



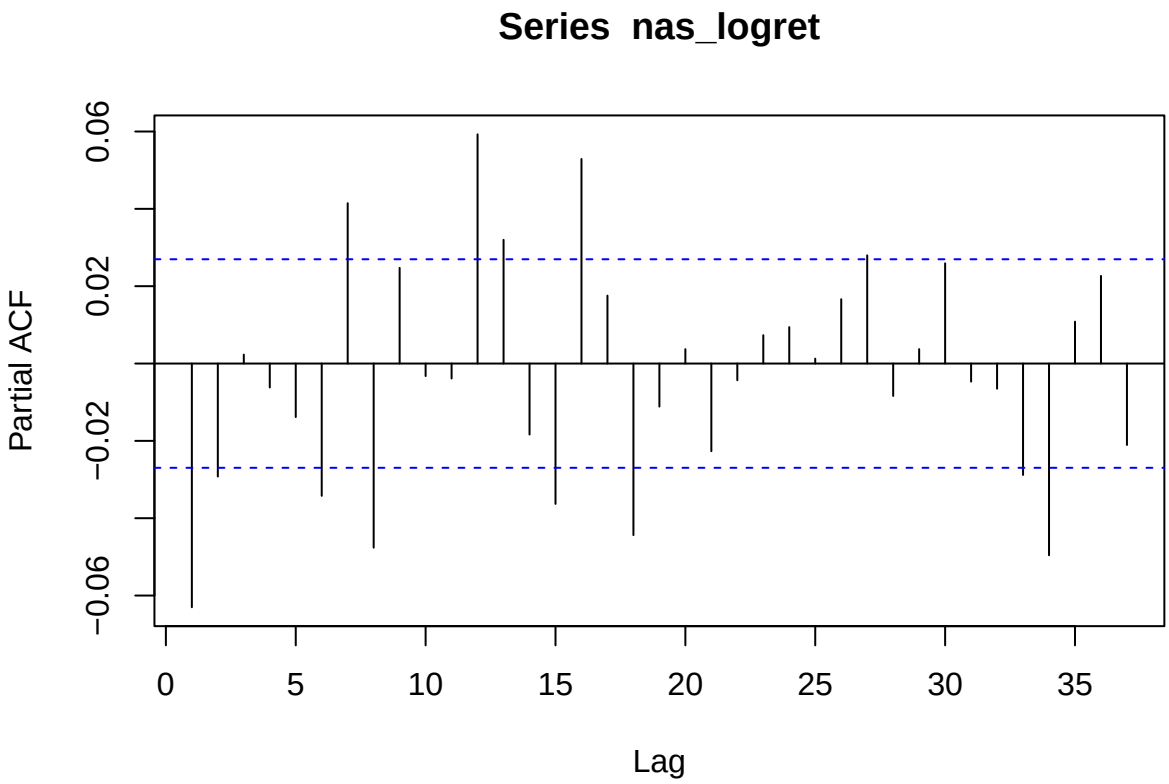

```
pacf(sp_logret)
```




```
pacf(ny_logret)
```




```
pacf(nas_logret)
```


Series nas_logret
Partial ACF
0.06
0.02
−0.02
−0.06
0
5
10
15
20
25
30
35
Lag