Exploratory Data Analysis

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

Initial Data setup.

```
library(quantmod)
```

```
## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 4.0.5
```

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

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

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

```
## Loading required package: TTR
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
getSymbols(c("^GSPC","CSCO","C","CVX","AMZN"), from = "1991-01-01", to = "2021-02-01")
```

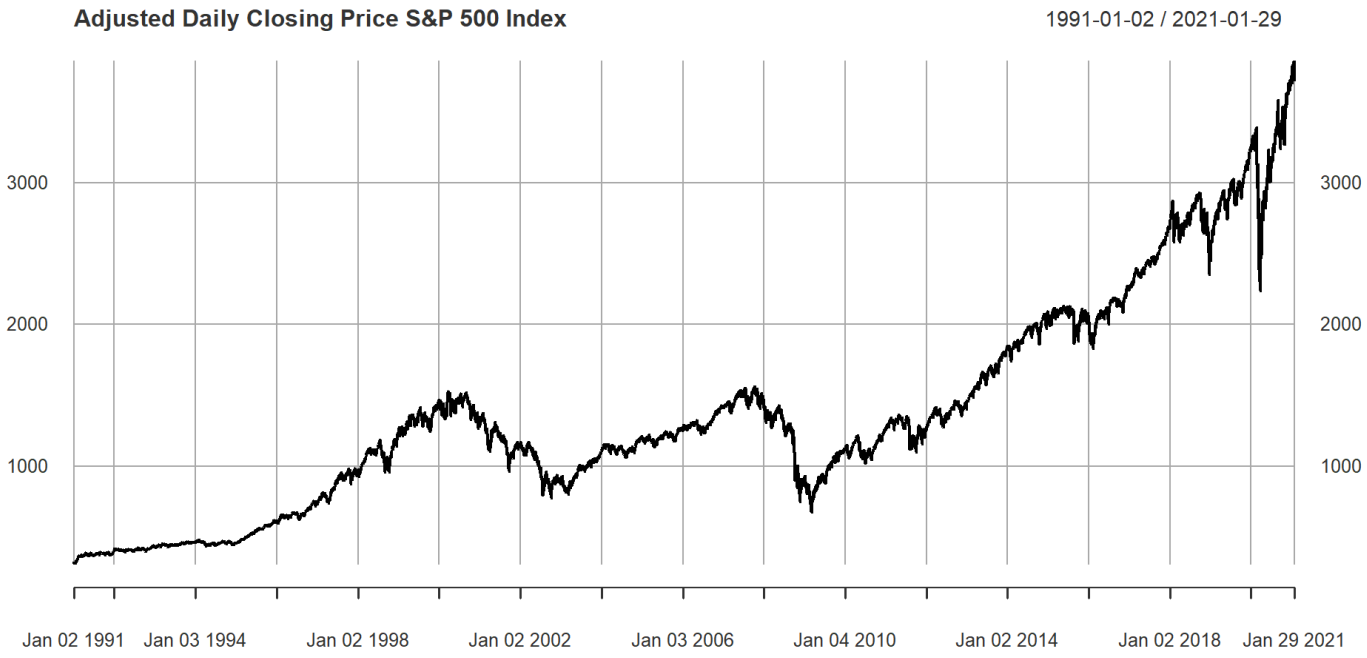
```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.
```

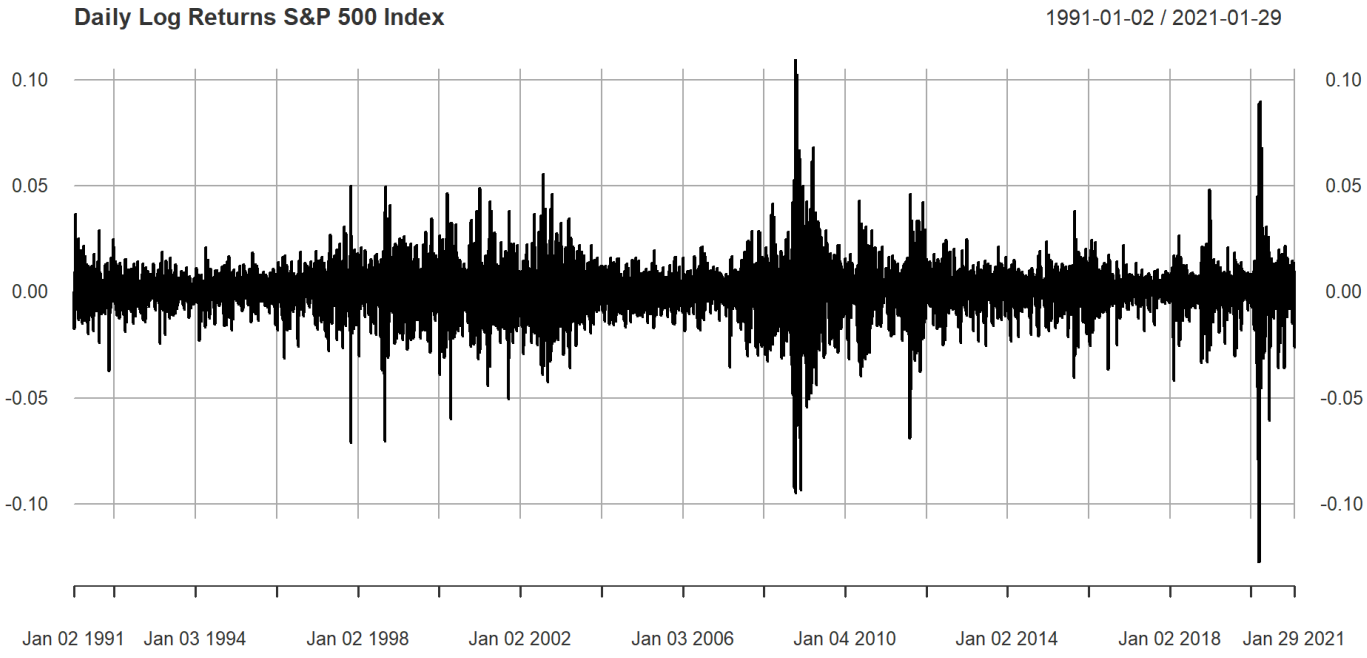
```
## [1] "^GSPC" "CSCO"  "C"     "CVX"   "AMZN"
```

Question 1: S&P 500 Index

Part a: Plot adjusted closing price and daily log returns.

```
par(mfrow   = c(2, 1))
plot(Ad(GSPC), main="Adjusted Daily Closing Price S&P 500 Index")
plot(dailyReturn(Ad(GSPC), type="log"), main="Daily Log Returns S&P 500 Index")
```





Part b: Describe the two plots accounting for macroeconomic events and compare to NASDAQ.

The adjusted closing price plot at the top corresponds to major macroeconomic events with the index peaking just before major crises. The first peak in early 2000 followed by a steep price decline corresponds to the 2000 dot com technology stock bust and national recession. Similarly, the peak in late 2007 and subsequent price decline

corresponds to the 2008 financial crisis and recession (housing overvaluations, fraud, bankruptcies of Lehman Bros. and Bear Sterns). Another sharp market drop in early 2020 corresponds to the current COVID pandemic and recession. Volatility in returns is significantly higher in the three crisis periods.

Compared to the NASDAQ, the S&P 500 index was more affected by the 2008 and 2020 crises than the 2000 dot com bubble. The NASDAQ was more affected by the 2000 dot com bubble, with a maximum log return loss of 30%. The NASDAQ was also affected by the 2008 financial crisis and the 2020 pandemic, but these events had less impact as compared to the dot com bubble and compared to the impact these events had on the S&P 500.

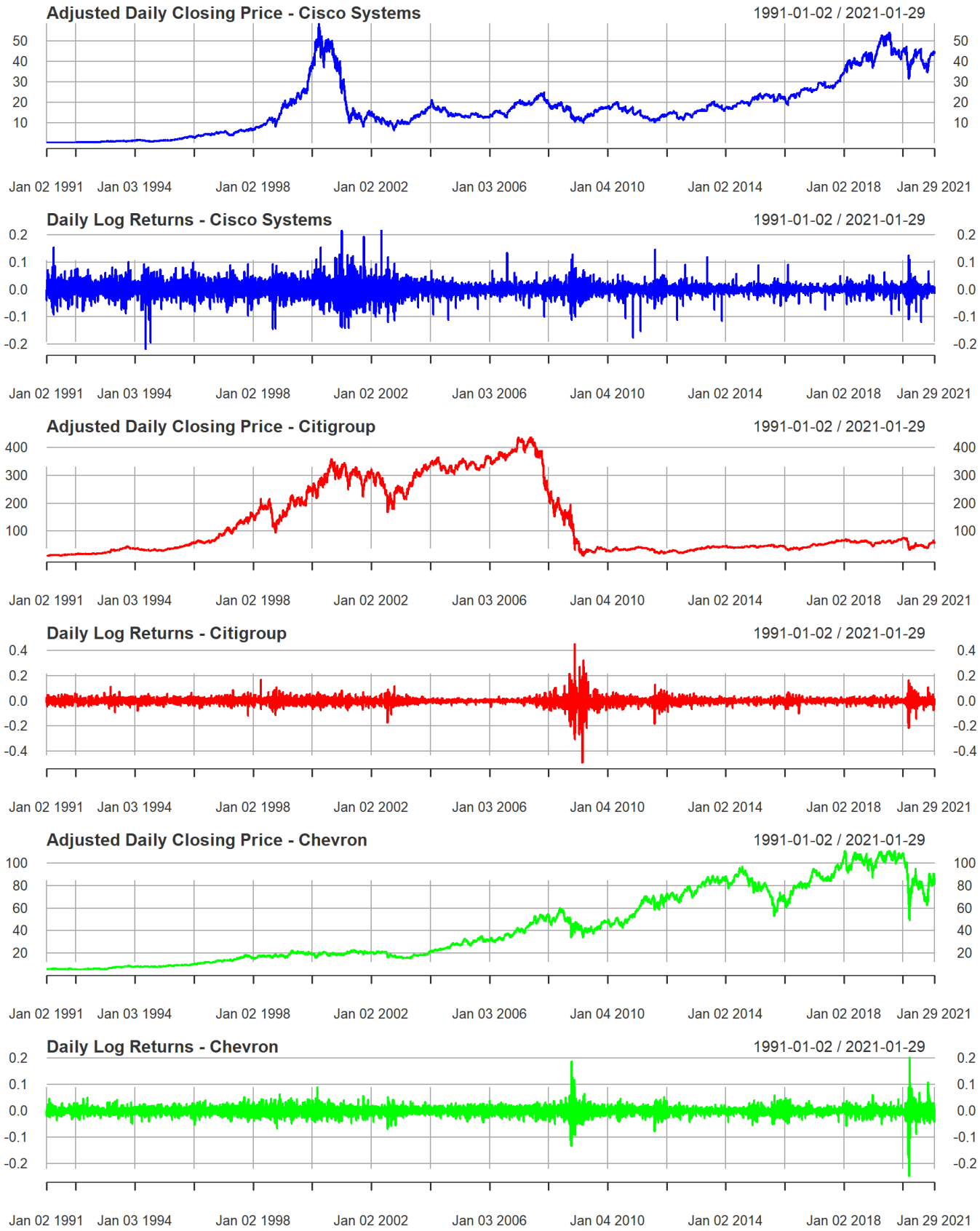
The S&P 500 Index shares some of the stlyized facts with the NASDAQ:
Stationarity: index prices don't appear stationary, but expansionary post-crisis following economic growth, federal and monetary stimulus and technology advancement. Returns do appear stationary however, following a mean daily return of close to 0%.

Asymmetery: returns again appear negatively skewed with downturns steeper than recoveries, but more short lived. The negative skew here seems slighltly less pronounced compared to the NASDAQ.

Question 2: Cisco, Citigroup & Chevron

Part a: Plot adjusted closing price and daily log returns.

```
par(mfrow   = c(6, 1))
plot(Ad(CSCO), main="Adjusted Daily Closing Price - Cisco Systems", col="blue")
plot(dailyReturn(Ad(CSCO), type="log"), main="Daily Log Returns - Cisco Systems", col="blue")
plot(Ad(C), main="Adjusted Daily Closing Price - Citigroup", col="red")
plot(dailyReturn(Ad(C), type="log"), main="Daily Log Returns - Citigroup", col="red")
plot(Ad(CVX), main="Adjusted Daily Closing Price - Chevron", col="green")
plot(dailyReturn(Ad(CVX), type="log"), main="Daily Log Returns - Chevron", col="green")
```



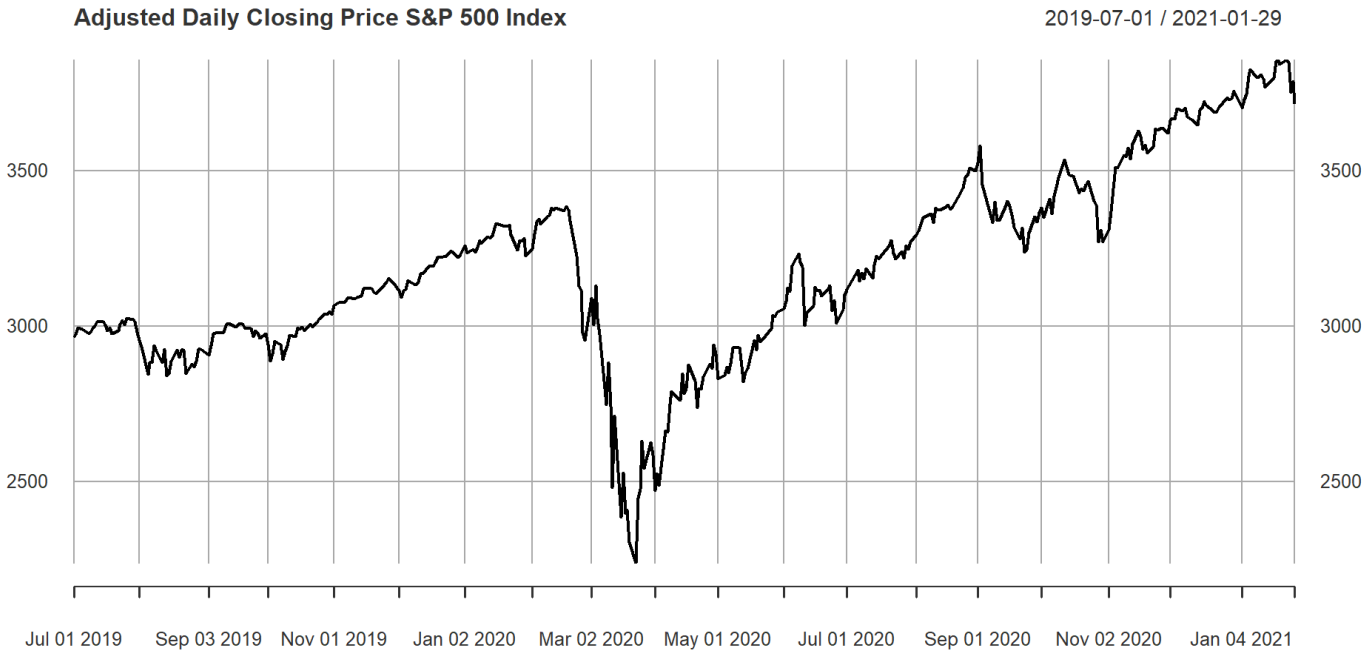
From the above plots, it seems that each of the three companies experienced different impacts following each of the three major crises previously discussed. Cisco, a technology company, saw it's widest retun volatility and largest price decline during the dot com recession. Citigroup, a banking and financial services company experienced the most difficulty in the 2008 financial crisis. Finally, Chevron, an energy (mainly petroleum) company experienced the most volatility and price declines during the 2020 COVID pandemic when petroleum/gas

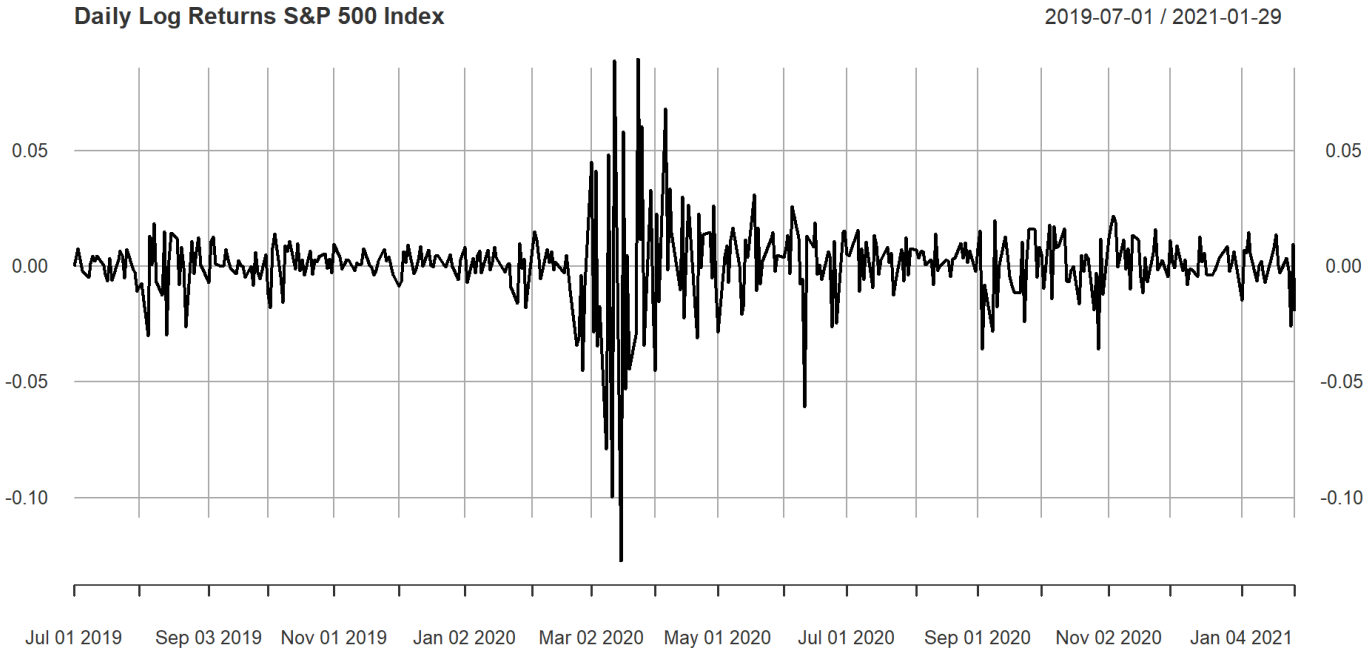
demand decreased dramatically as people around the world curtailed air and auto travel. Another interesting stylized fact is that each of these three companies could be seen as a safe investment shelter in at least one type of crisis. The impact of the dot com recssion didn't have a significant impact on Chevron's or Citigroup's share prices. Cisco and Chevron were only moderately affected by the financial crisis. Finally, Citigroup and Cisco have weathered the pandemic crisis fairly well as compared to Chevron, whose stock price has suffered and experienced high volatility.

Question 3: COVID 19 Impact

Part a: Plot adjusted closing price and daily log returns for the S&P 500 Index from 2019 onwards.

```
par(mfrow   = c(2, 1))
plot(Ad(GSPC["2019-07-01::2021-02-01"]), main="Adjusted Daily Closing Price S&P 500 Index")
plot(dailyReturn(Ad(GSPC["2019-07-01::2021-02-01"]), type="log"), main="Daily Log Returns S&P 50
0 Index")
```

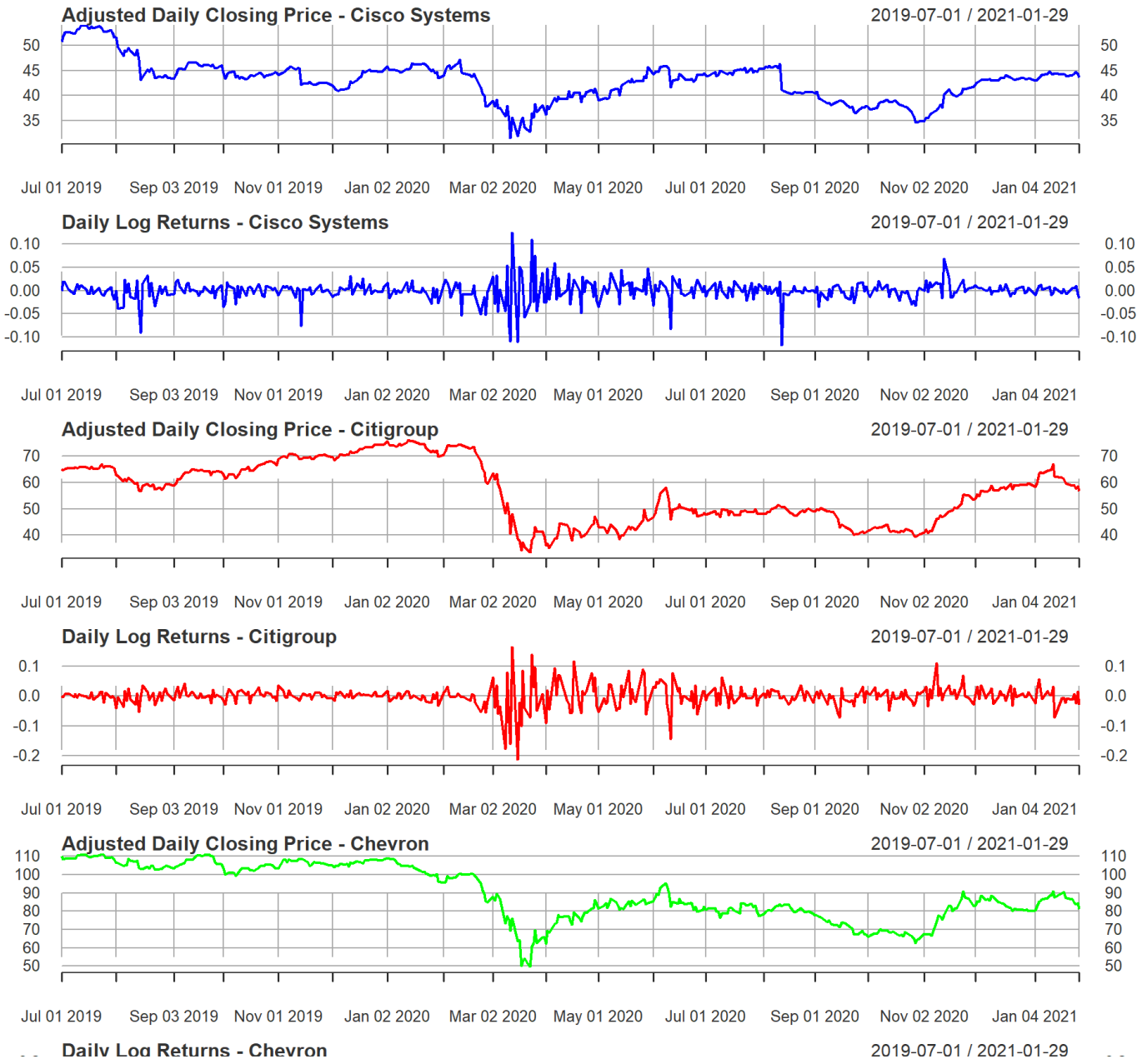


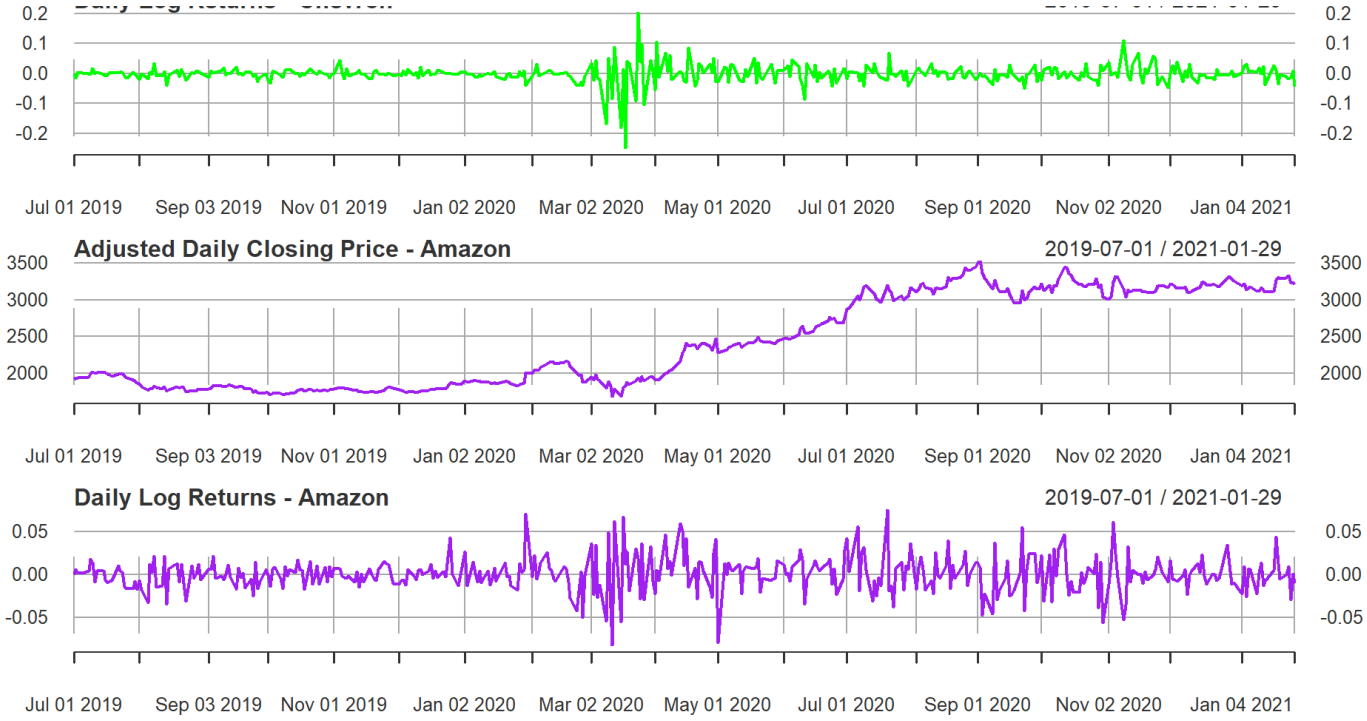


The plots indicate high return volatility and ongoing price declines at the start of the COVID 19 pandemic in the spring of 2020. This decline was then offset by a period of upward trending price appreciation and periods of lower return volatility as compared to the start of the pandemic. However, volatility remains higher during the pandemic than it was in the year preceeding the pandemic.

Part b: Plot the 4 companies (now including Amazon). Compare performance vs S&P during the pandemic.

```
par(mfrow   = c(8, 1))
plot(Ad(CSCO["2019-07-01::2021-02-01"]), main="Adjusted Daily Closing Price - Cisco Systems", co
l="blue")
plot(dailyReturn(Ad(CSCO["2019-07-01::2021-02-01"]), type="log"), main="Daily Log Returns - Cisc
o Systems", col="blue")
plot(Ad(C["2019-07-01::2021-02-01"]), main="Adjusted Daily Closing Price - Citigroup", col="red"
)
plot(dailyReturn(Ad(C["2019-07-01::2021-02-01"]), type="log"), main="Daily Log Returns - Citigro
up", col="red")
plot(Ad(CVX["2019-07-01::2021-02-01"]), main="Adjusted Daily Closing Price - Chevron", col="gree
n")
plot(dailyReturn(Ad(CVX["2019-07-01::2021-02-01"]), type="log"), main="Daily Log Returns - Chevr
on", col="green")
plot(Ad(AMZN["2019-07-01::2021-02-01"]), main="Adjusted Daily Closing Price - Amazon", col="purp
le")
plot(dailyReturn(Ad(AMZN["2019-07-01::2021-02-01"]), type="log"), main="Daily Log Returns - Amaz
on", col="purple")
```

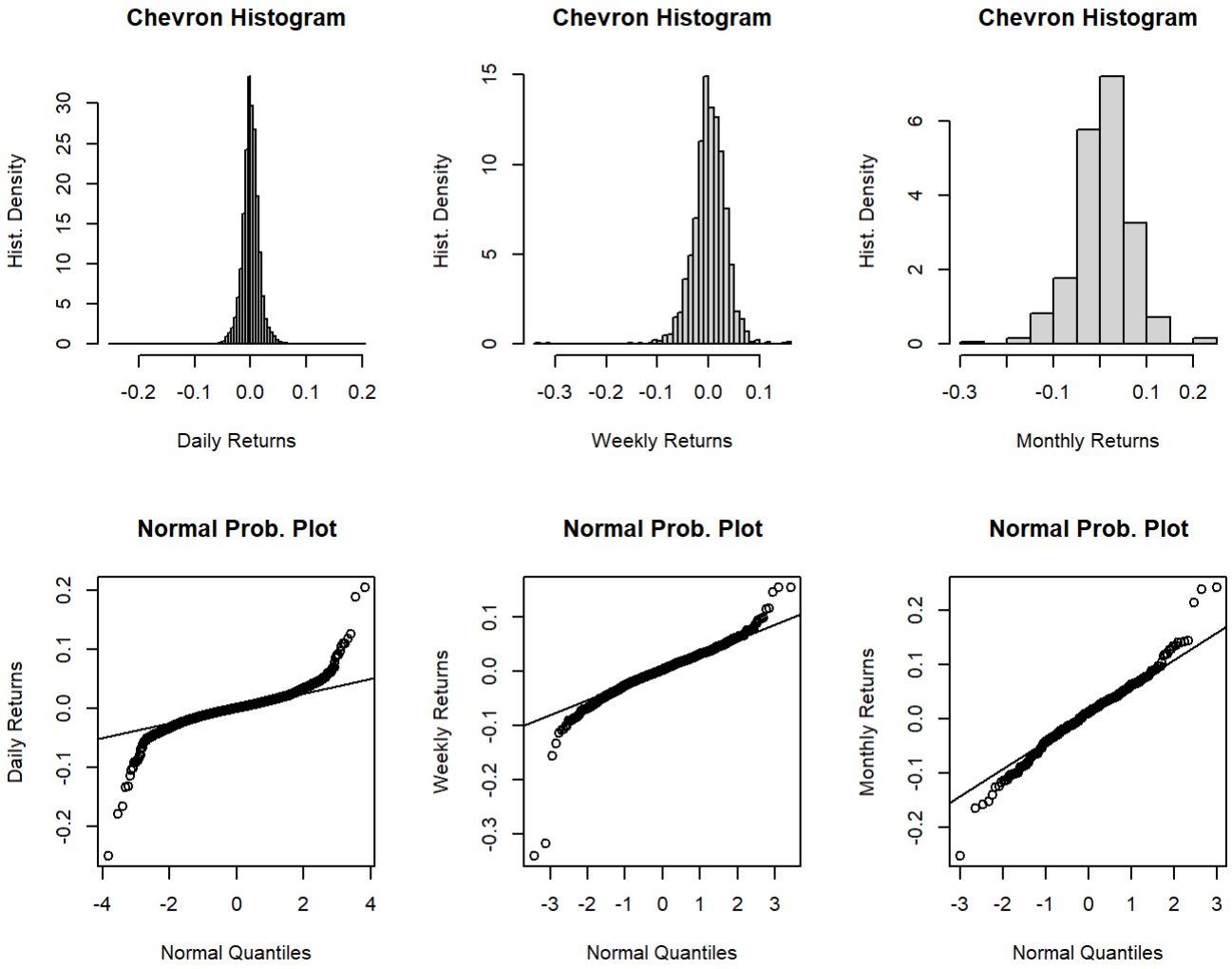




Similar to the S&P 500 Index, all 4 stocks saw the largest period of volatility at the start of the pandemic and each of the 4 companies have higher volatility throughout 2020 than they saw in 2019. The exception here is Amazon where the level of volatility experienced at the start of the pandemic has continued and not decreased. Amazon has also fully recovered it's price losses at the start of the pandemic and the stock price has grown significantly during the pandemic. This reflects the higher demand that the company faces as most people are turning to online retailers to obtain goods during the pandemic. Amazon's share price growth is similar to that of the S&P index. The other 3 companies have not seen share price growth beyond the share price at the start of the pandemic.

Question 4: Chevron

Part a: Historgram.



Part b: Comment on Plots.

The daily return histogram and normal probability plot indicate that the daily return data is not normally distributed. The daily returns have a much higher peak and longer tails than the standard normal curve. The deviation from the straight line on the normal probability plot indicate longer left and right tails. The weekly returns are slightly more normal but still have longer tails at both sides and thus do not appear to be sufficiently normal. The monthly returns are much closer with only some slight deviation from the line in the normal probability plot (again indicating slightly longer tails), but this would likely be close enough to be considered approximately normal.

Part c: Calculate skewness and kurtosis. Run Shapiro-Wilkes and Jarque-Bera tests.

```
Sk.fun <- function(x) { ## function to compute skewness from HO 3
  mean((x-mean(x))^3/sd(x)^3)
}
Kur.fun <- function(x){ ## function to compute kurtosis
  mean((x-mean(x))^4/sd(x)^4)
}
JB.test <- function(x) { ## Jarque-Bera Test
  JB <-length(x)*(Sk.fun(x)^2/6+(Kur.fun(x)-3)^2/24)
  list(stat = JB, pvalue = 1-pchisq(JB,2))
}

resultsP4 = matrix(rep(0,18),ncol=3)  # set up a results matrix to gather calcs

for (i in 1:3) {
  resultsP4[1,i] = round(Sk.fun(as.vector(Rt[[i]])),4)  # skewness calc
  resultsP4[2,i] = round(Kur.fun(as.vector(Rt[[i]])),4) # kurtosis calc
  resultsP4[3,i] = ifelse(length(as.vector(Rt[[i]]))<=5000,round(as.numeric(shapiro.test(as.vect
or(Rt[[i]]))[[1]]),4),NA)
  resultsP4[4,i] = ifelse(length(as.vector(Rt[[i]]))<=5000,round(as.numeric(shapiro.test(as.vect
or(Rt[[i]]))[[2]]),4),NA)
  resultsP4[5,i] = round(as.numeric(JB.test(Rt[[i]])[1]),4)
  resultsP4[6,i] = round(as.numeric(JB.test(Rt[[i]])[2]),4)
}
```

	Daily Log Returns	Weekly Log Returns	Monthly Log Returns
Skewness	-0.372	-1.2738	-0.0528
Kurtosis	22.7796	15.3981	4.8651
SW Statistic	NA	0.9274	0.9778
SW p-value	NA	0	0
JB Statistic	1.236906310^{5}	1.047317810^{4}	52.3453
JB p-value	0	0	0

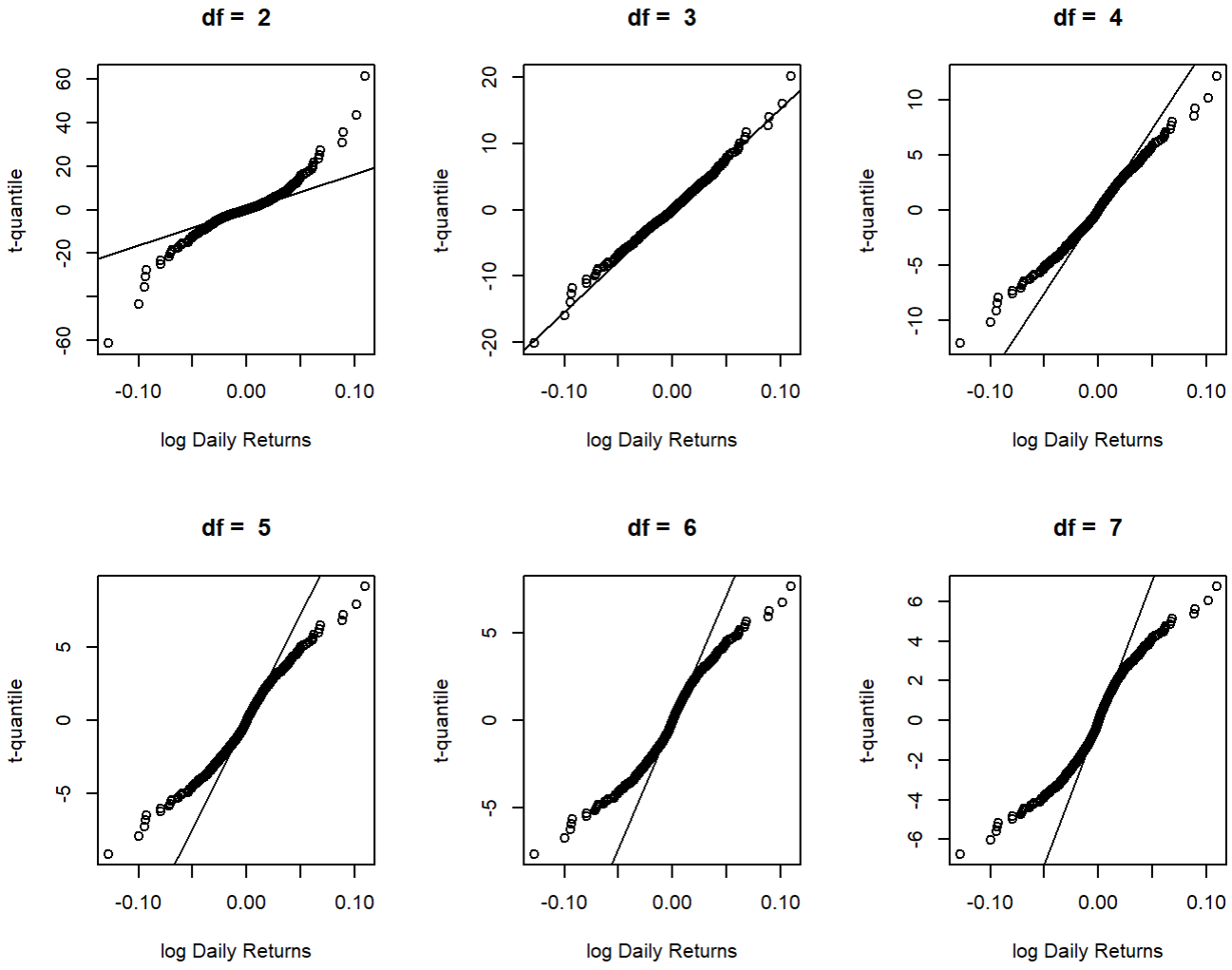
Of the three log return series, the only one close to being approximately normal is monthly returns. Log monthly returns have close to 0 skewness and a kurtosis of 4.86, which is similar to the skewness and kurtosis of a normal distribution (0 and 3, respectively). The Shapiro-Wilk test statistic is close to 1 at 0.9778 which also indicates an approximately normal distribtuion. However, the p-value is still very low, indicating there is significant evidence to reject the null hypothesis that the monthly return data is normally distributed. A similar conclusion is reached from the Jarque-Bera test, which results in a high test statistic of 52 and a near 0 p-value, again indicating that there's significant evidence to reject the null hypothesis that the log monthly returns has skewness and kurtosis levels comparable to a normal distribution. So, contrary to the answer from part b, none of the 3 returns are approximately normally distributed.

Question 5: t-Plots

Setup.

```
par(mfrow = c(2, 3))
logR      = as.vector(dailyReturn(Ad(GSPC), type="log"))  # set up daily log return vector used
 in code
n         = length(logR)
 q_grid   = (1:n) / (n + 1)
 df_grid  = 2:7

for(df in df_grid)  {
  qqplot(logR, qt(q_grid,df),
  main = paste("df = ", df), ylab="t-quantile", xlab="log Daily Returns" )
  abline(lm(qt(c(0.25, 0.75), df = df) ~ quantile(logR, c(0.25, 0.75))))
}
```



Problem 4: What does the code q_grid = (1:n) / (n + 1) do? What does qt(q.grid, df = df[j]) do? What does paste do?

The q_grid object calculates the endpoints for equally spaced intervals on the 0 to 1 line segment given the total number of log daily returns we have.

The qt object then uses these endpoints to calculate t-quantiles for each interval endpoint given a specified value of degrees of freedom. This sets up the q-q plot for a t-distribution.

Paste converts and combines the arguments of the function into a character string for the plot titles (here the # of deg. freedom).

Problem 5: state which choice of the degrees of freedom parameter gives the best-fitting t-distribution and explain why.
The t-distribution with df=3 provides the best fit as the empirical t-quantiles match up well to the straight line theoretical t-quantiles. For df=2, the log return data has significantly longer tails than the theoretical t-distribution. For df=4 and above, the log retrun data has significantly shorter tails than the theoretical t-distribution.

Question 6: Density Plots

Setup.

```
library("fGarch")
```

```
## Loading required package: timeDate
```

```
## Loading required package: timeSeries
```

```
##
## Attaching package: 'timeSeries'
```

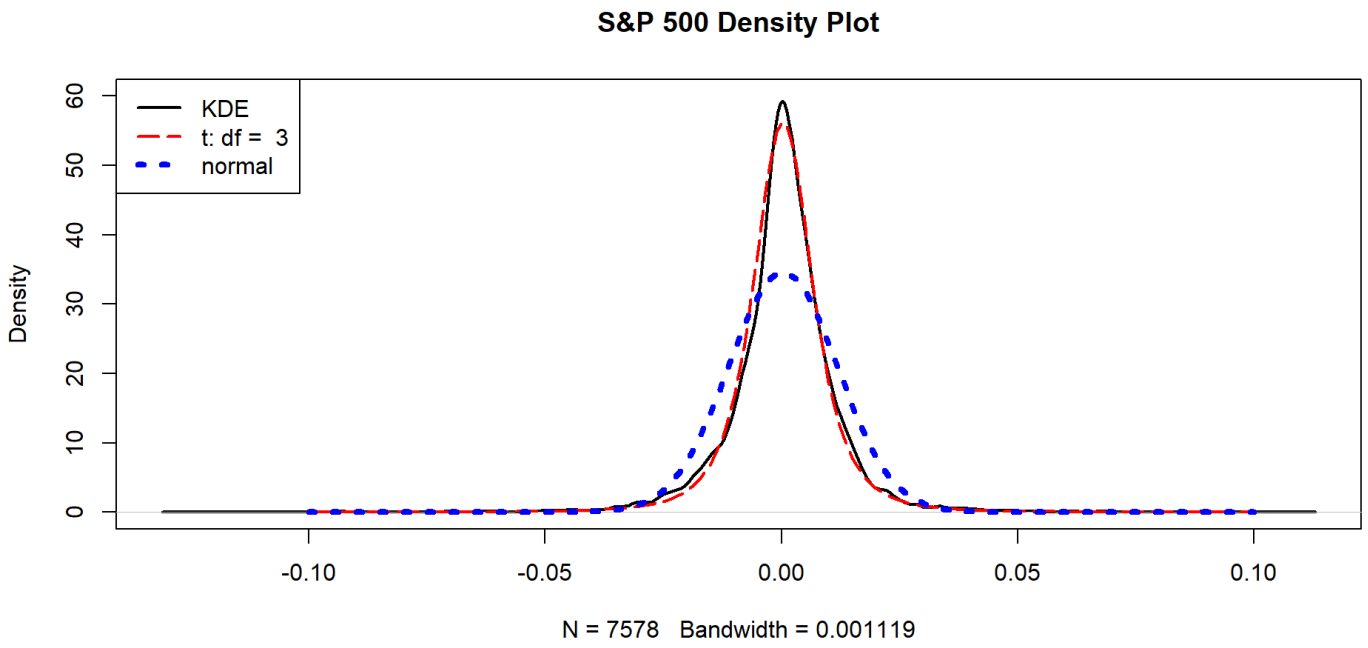
```
## The following object is masked from 'package:zoo':
##
##     time<-
```

```
## Loading required package: fBasics
```

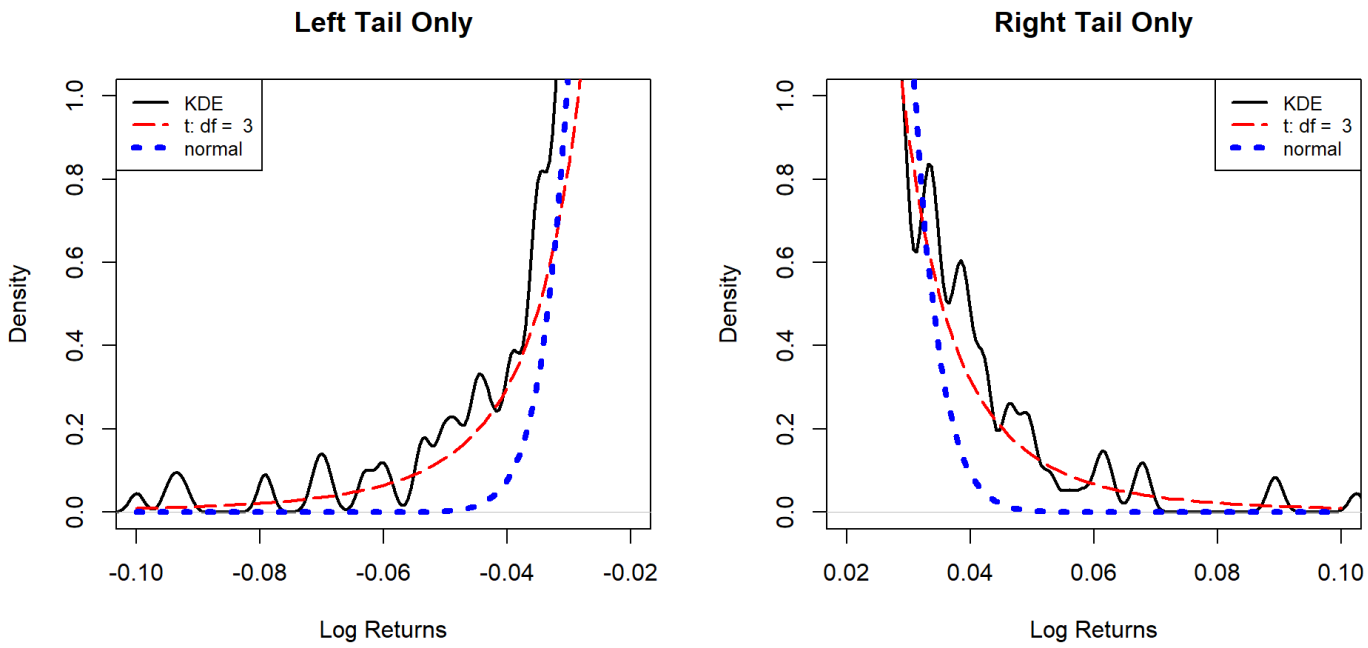
```
##
## Attaching package: 'fBasics'
```

```
## The following object is masked from 'package:TTR':
##
##     volatility
```

```
x=seq(-0.1, 0.1,by = 0.001)
par(mfrow = c(1, 1))
df = 3
mad_t = mad(logR, constant = sqrt(df / (df - 2)) / qt(0.75, df))
plot(density(logR), lwd = 2, ylim = c(0, 60), main="S&P 500 Density Plot")
lines(x, dstd(x, mean = mean(logR), sd = mad_t, nu = df), lty = 5, lwd = 2, col = "red")
lines(x, dnorm(x, mean = mean(logR), sd = sd(logR)), lty = 3, lwd = 4, col = "blue")
legend("topleft", c("KDE", paste("t: df = ",df), "normal"), lwd = c(2, 2, 4), lty = c(1, 5, 3),
col = c("black", "red", "blue"))
```




```
# zoom in plot, left tail
par(mfrow = c(1, 2))
plot(density(logR), lwd = 2, ylim = c(0, 1), xlim=c(-.1,-0.02), main="Left Tail Only", xlab = "L
og Returns")
lines(x, dstd(x, mean = mean(logR), sd = mad_t, nu = df), lty = 5, lwd = 2, col = "red")
lines(x, dnorm(x, mean = mean(logR), sd = sd(logR)), lty = 3, lwd = 4, col = "blue")
legend("topleft", c("KDE", paste("t: df = ",df), "normal"), lwd = c(2, 2, 4), lty = c(1, 5, 3),
 col = c("black", "red", "blue"), cex=.8)
# zoom in plot, right tail
plot(density(logR), lwd = 2, ylim = c(0, 1), xlim=c(0.02,0.1), main="Right Tail Only", xlab = "L
og Returns")
lines(x, dstd(x, mean = mean(logR), sd = mad_t, nu = df), lty = 5, lwd = 2, col = "red")
lines(x, dnorm(x, mean = mean(logR), sd = sd(logR)), lty = 3, lwd = 4, col = "blue")
legend("topright", c("KDE", paste("t: df = ",df), "normal"), lwd = c(2, 2, 4), lty = c(1, 5, 3),
col = c("black", "red", "blue"), cex=.8)
```



Problem 6: Do either of the parametric models provide a reasonably good fit to the S&P 500 index? Explain.

Yes, the t-distribution with 3 degrees of freedom provides a fairly good fit as the density plot matches up very well to the kernel density estimator (KDE). The KDE is only slightly more peaked than the t-distribution and the tails match up well. The Normal Distribution does not match up well at all, with lighter/shorter tails and much shorter and wider peak.

Problem 7: Which bandwidth selector is used as the default by density? What is the default kernel?

The default bandwidth selector is bw.nrd0 which is Silverman's rule-of-thumb for selecting the bandwidth of a gaussian estimator (here we get 0.0011). The default kernel is the gaussian kernel.