PDXDataSciStkMkt

Charles Howard

October 21, 2017

The raw data

I read the data in using read.csv with defaults:

```
library(xts)
```

```
## Warning: package 'xts' was built under R version 3.4.2
```

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

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

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

```
datadir<-"C:/Users/Charles/Documents/StockStuff/PDXDataSci/"
dat<-read.csv(paste(datadir,"stocks-us-adjClose.csv",sep=""))
# get column wise count of NA's
nacnt<-sapply(1:dim(dat)[2],function(n){length(which(is.na(dat[,n])))})
# looks like I just need to change the name of col 1 to date
names(dat)[1]<-"Date"
dat$Date<-as.Date(as.character(dat$Date))
```

The data provided to the group consists of 12,032 rows by 711 columns.

```
dim(dat)
```

```
## [1] 12032   711
```

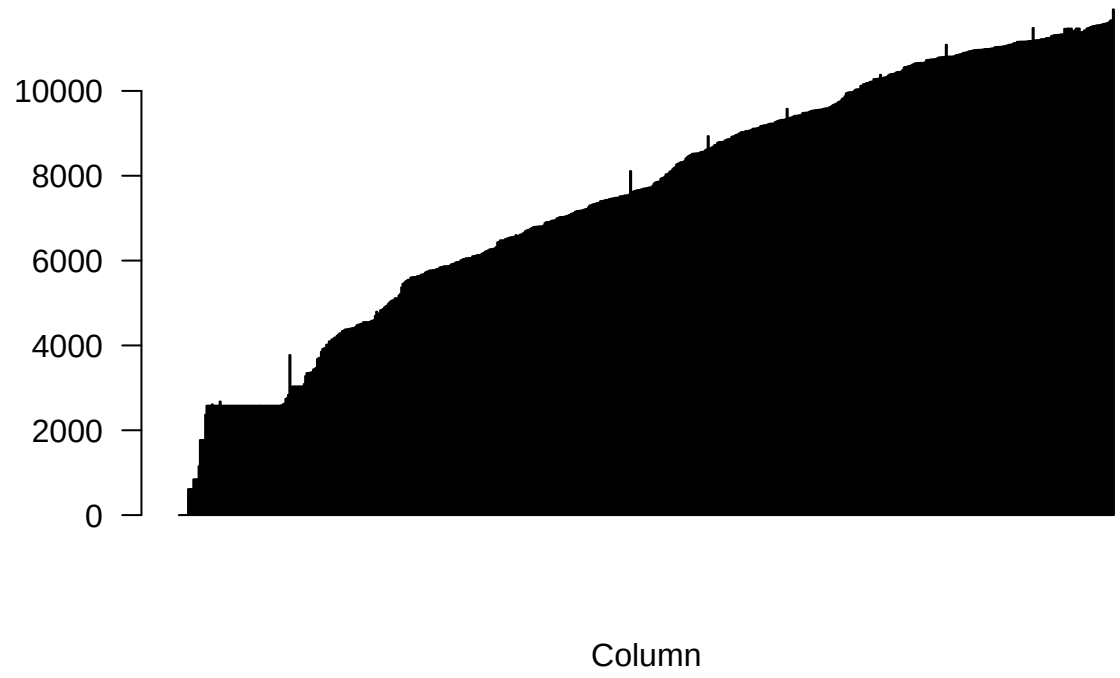
The first column is a date and the remaining 710 columns are adjusted closing prices for securities. The date range is:

```
range(dat$Date)
```

```
## [1] "1970-01-02" "2017-09-08"
```

Due to the ~ 47 year date range, there is a large variance in number of empty cells column-to-column.

```
barplot(nacnt,main="Empty Cell Count by Column",las=1,xlab="Column")
```



Identifying the Securities

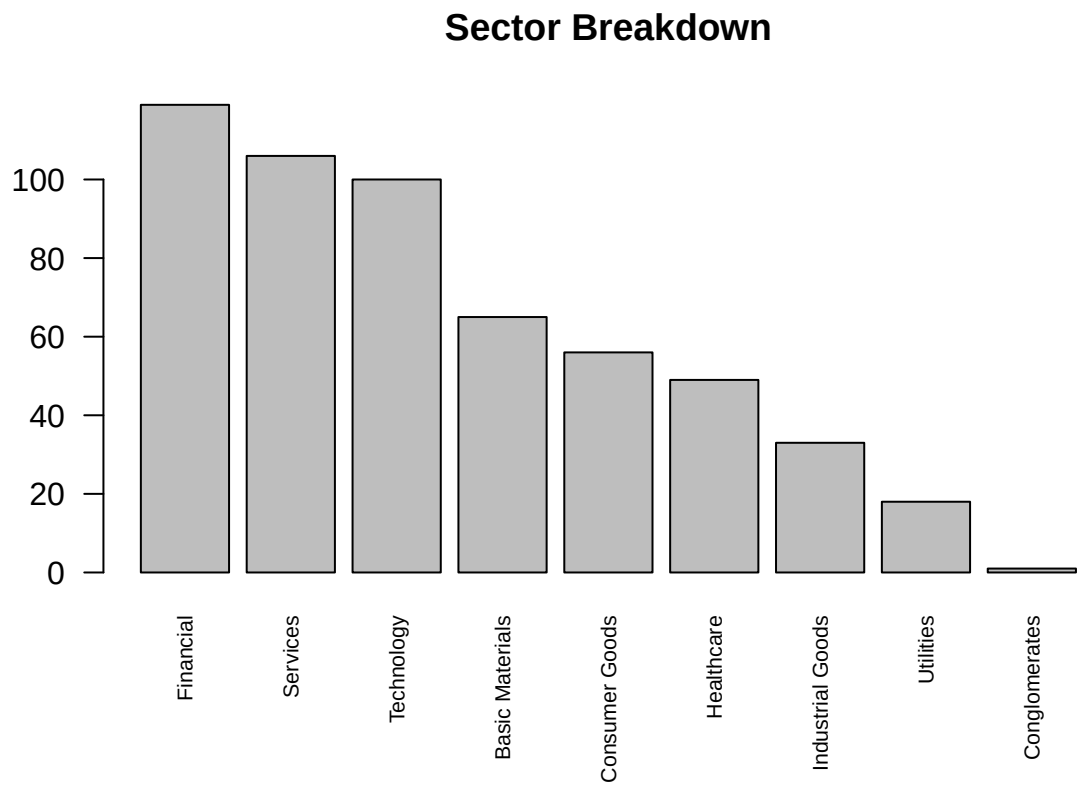
Next I gathered some basic information about each of the securities (Company name, Market Sector, and Industry) from Yahoo Finance:

```
library(rvest)
# lookup each symbol and retrieve the sector and industry
# two problems scraping Yahoo Finance.
# 1.) intermittent HTTP error 503
# 2.) successful read_html, but no data on the company
sybls<-names(dat)[2:length(names(dat))]
compdata<-c()
for(i in sybls){
  errflag<-0
  Sys.sleep(0.001)
  e<-simpleError("HTTP error 503.")
  url<-paste("https://finance.yahoo.com/quote/",i,"/profile?p=",i,sep="")
  test<-try(webpg<-read_html(url))
  if(class(test) %in% "try-error"){
    cmpy_name<-NA
    stk_sector<-NA
    stk_industry<-NA
    compdata<-rbind(compdata,cbind(i,cmpy_name,stk_sector,stk_industry))
    next} else
#  webpg<-read_html(url)
  cmpy_name<-html_text(html_nodes(webpg,"div h3[class='Mb(10px)']"))
  strspans<-html_nodes(webpg,"strong , span")
  strspans_text<-html_text(strspans)
```

```
  stk_sector<-strspans_text[which(strspans_text %in% "Sector")+1]
  stk_industry<-strspans_text[which(strspans_text %in% "Industry")+1]
  if(length(cmpy_name)==0){cmpy_name<-NA}
  if(length(stk_sector)==0){stk_sector<-NA}
  if(length(stk_industry)==0){stk_industry<-NA}
  compdata<-rbind(compdata,cbind(i,cmpy_name,stk_sector,stk_industry))
}
# there are 111 sybls have 0 length character objects in
# stk_sector and stk_industry
compdata[which(nchar(compdata[,3])==0),3]<-NA
compdata[which(nchar(compdata[,4])==0),4]<-NA
compdata_df<-as.data.frame(compdata)
names(compdata_df)<-c("Symbol","Company.Name","Sector","Industry")
```

Breakdown of Sectors:

```
oldpar<-par()
nwmar<-par("mar")+c(1,0,0,0)
par(mar=nwmar)
sectortb<-tapply(compdata_df$Sector,compdata_df$Sector,length)
sectortb<-sectortb[order(sectortb,decreasing = TRUE)]
barplot(sectortb,main="Sector Breakdown",las=2,cex.names = 0.7)
```



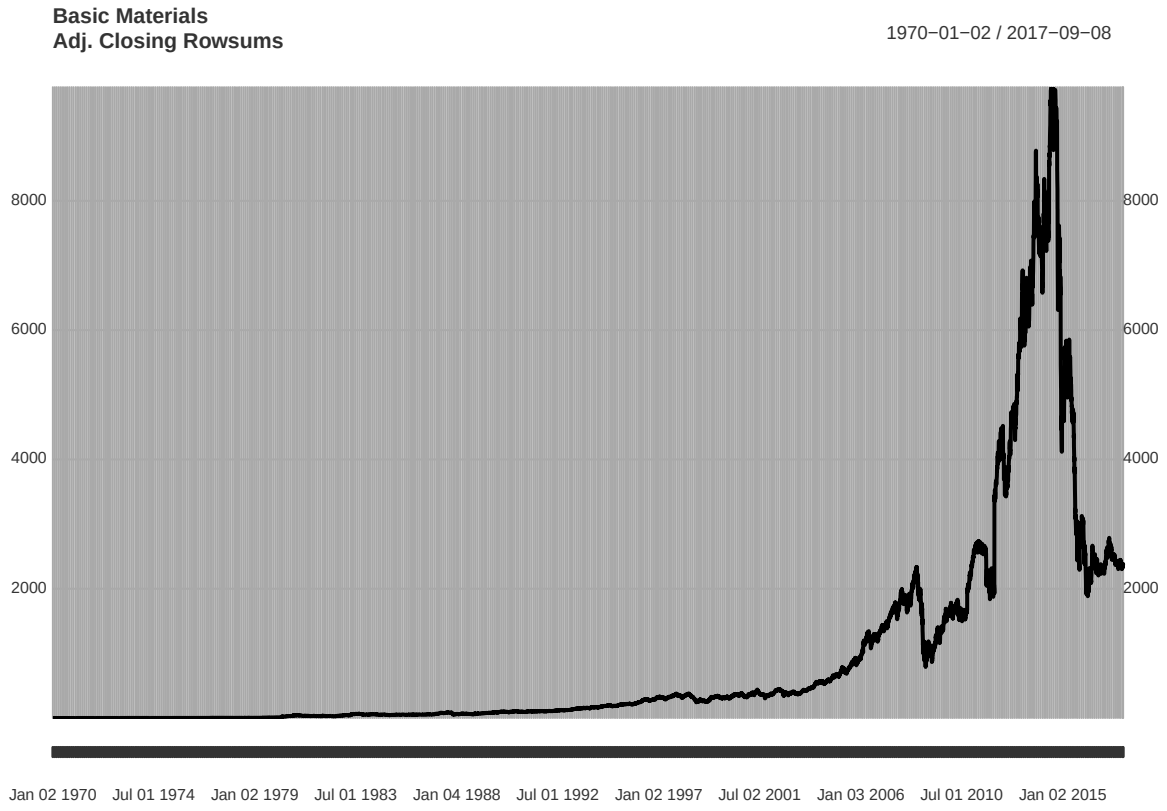
Sector Analysis

Created a list of time series for each sector and plotted various quantities.

```
library(xts)
sector_rowsums_ls<-lapply(levels(compdata_df$Sector),function(n){
  symb<-compdata_df$Symbol[which(compdata_df$Sector %in% n)]
  cols<-which(names(dat) %in% symb)
  rsm<-unlist(sapply(1:dim(dat)[1],function(n){sum(dat[n,cols],na.rm=T)}))
  rsm_ts<-xts(rsm,order.by = dat$Date)
  zeroes<-which(rsm_ts==0)
  if(length(zeroes)==0){rsm_ts} else
    {rsm_ts<-rsm_ts[-zeroes]}
})
```

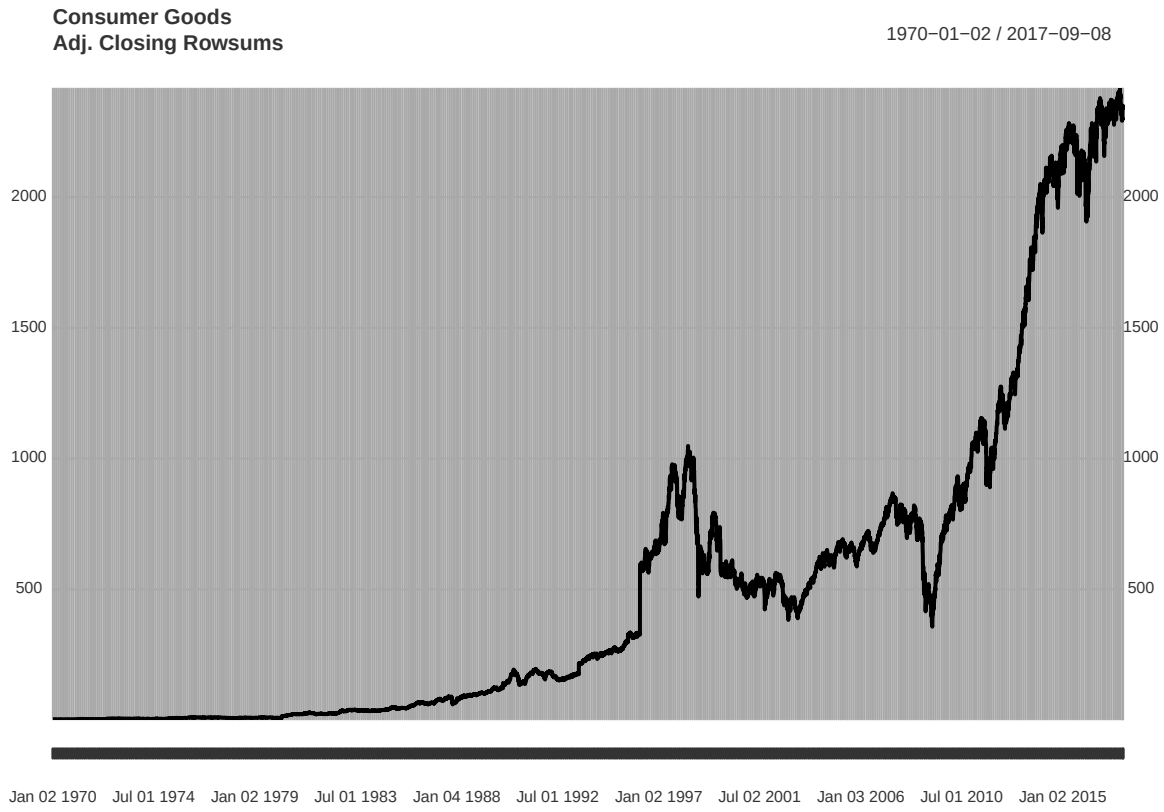
Plots of daily adjusted closing price row sums for each sector. Note: Rows containing zeroes have been removed.

```
plot(sector_rowsums_ls[[1]],main=paste(levels(compdata_df$Sector)[1],
                                       "\nAdj. Closing Rowsums",sep=""))
```



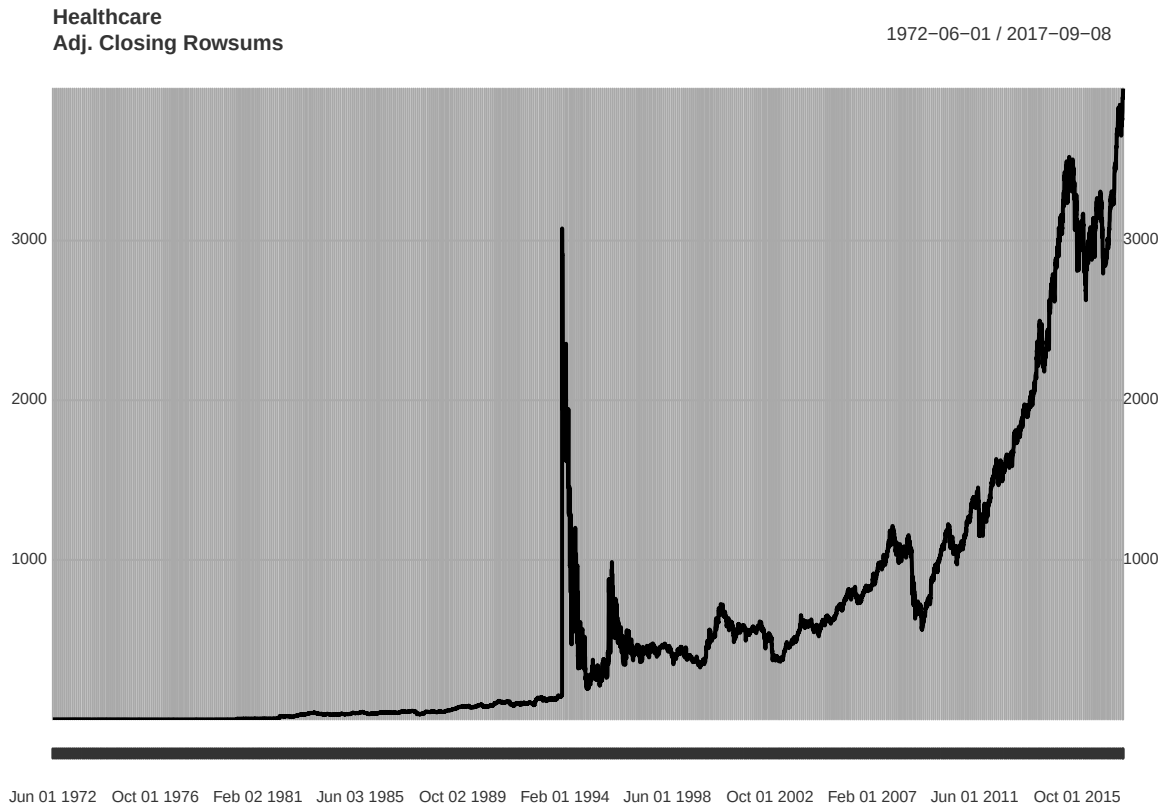

```
plot(sector_rowsums_ls[[2]],main=paste(levels(compdata_df$Sector)[2],
                                       "\nAdj. Closing Rowsums",sep=""))
```


```
plot(sector_rowsums_ls[[3]],main=paste(levels(compdata_df$Sector)[3],
                                       "\nAdj. Closing Rowsums",sep=""))
```

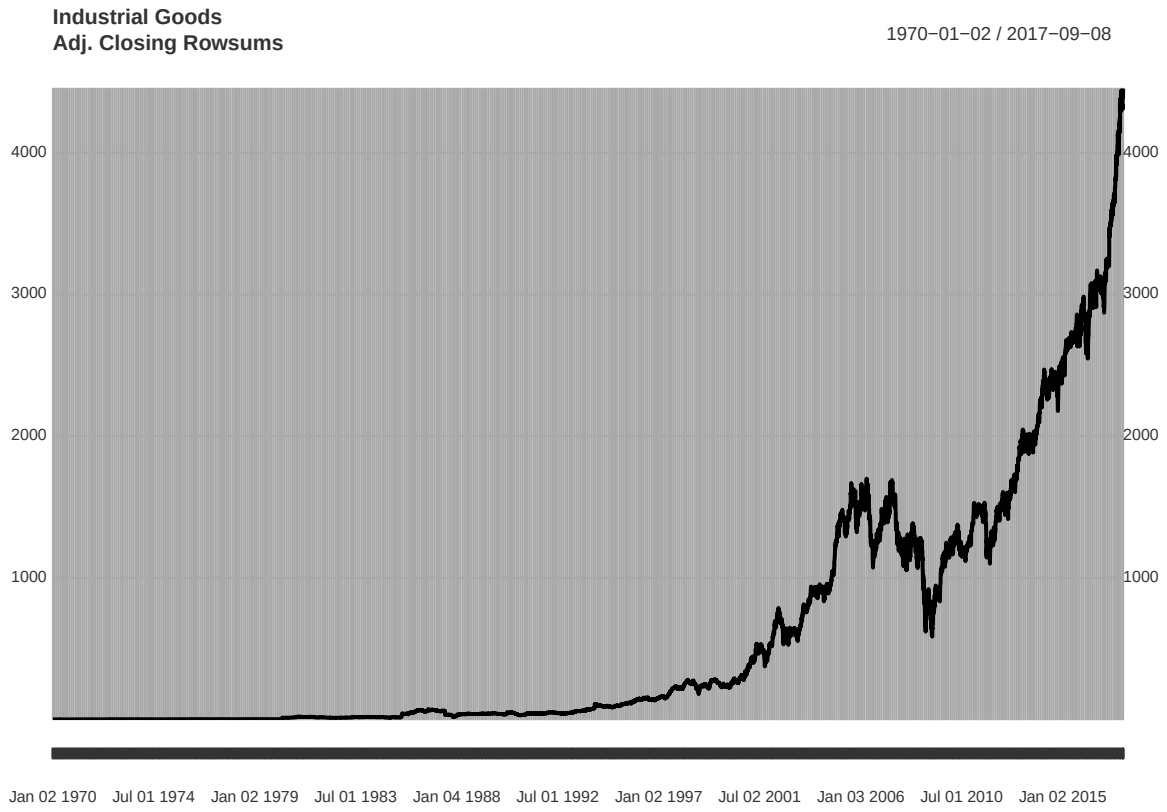



```
plot(sector_rowsums_ls[[4]],main=paste(levels(compdata_df$Sector)[4],
                                       "\nAdj. Closing Rowsums",sep=""))
```

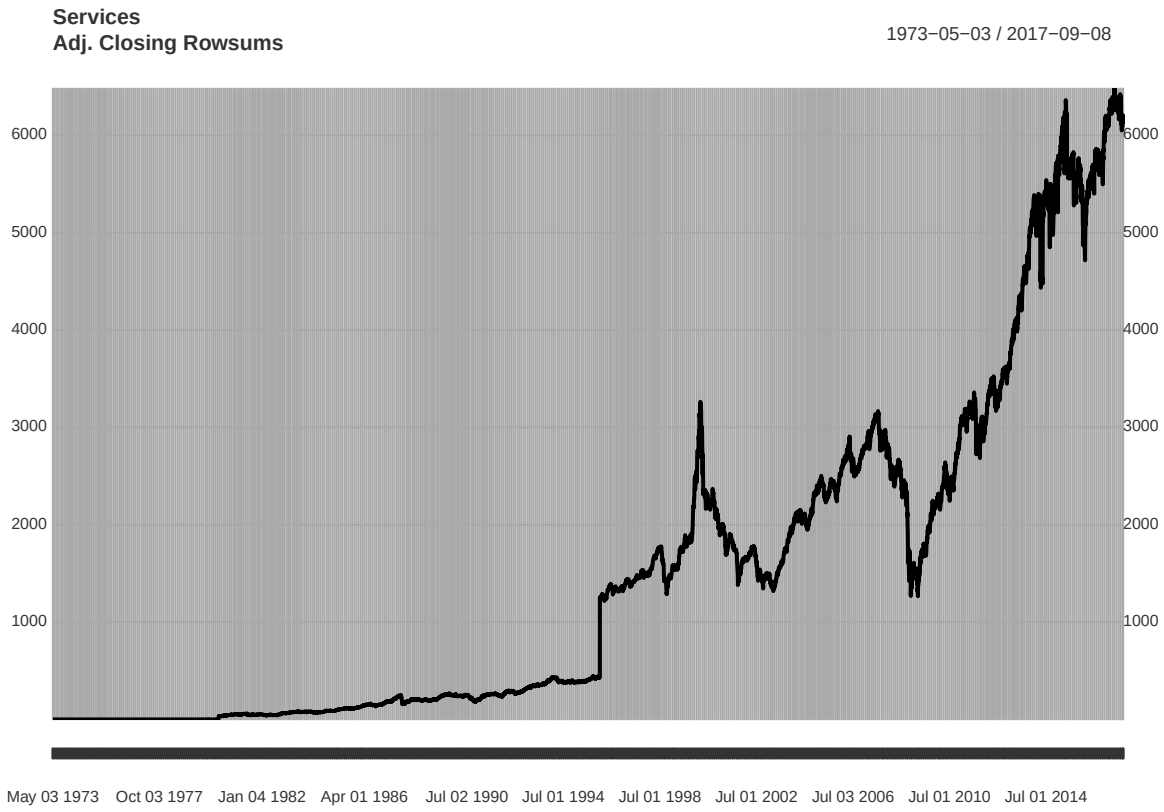

```
plot(sector_rowsums_ls[[5]],main=paste(levels(compdata_df$Sector)[5],
                                       "\nAdj. Closing Rowsums",sep=""))
```



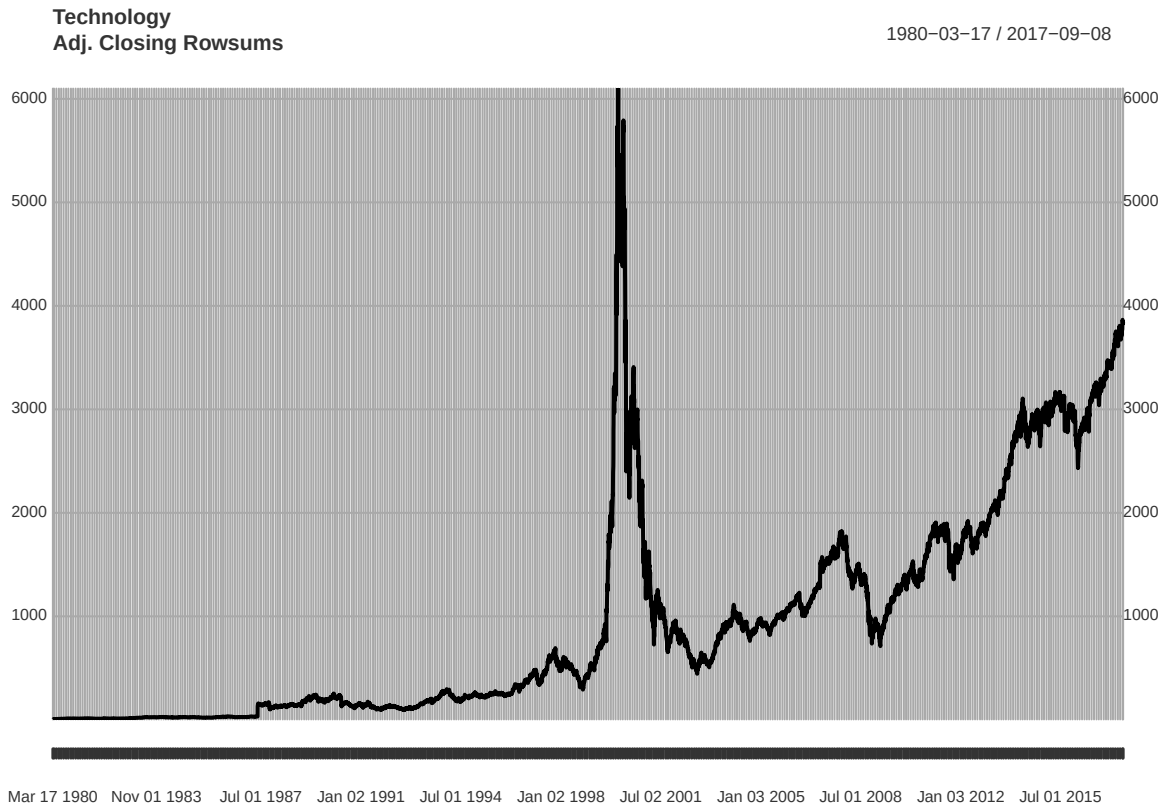

```
plot(sector_rowsums_ls[[6]],main=paste(levels(compdata_df$Sector)[6],
                                       "\nAdj. Closing Rowsums",sep=""))
```



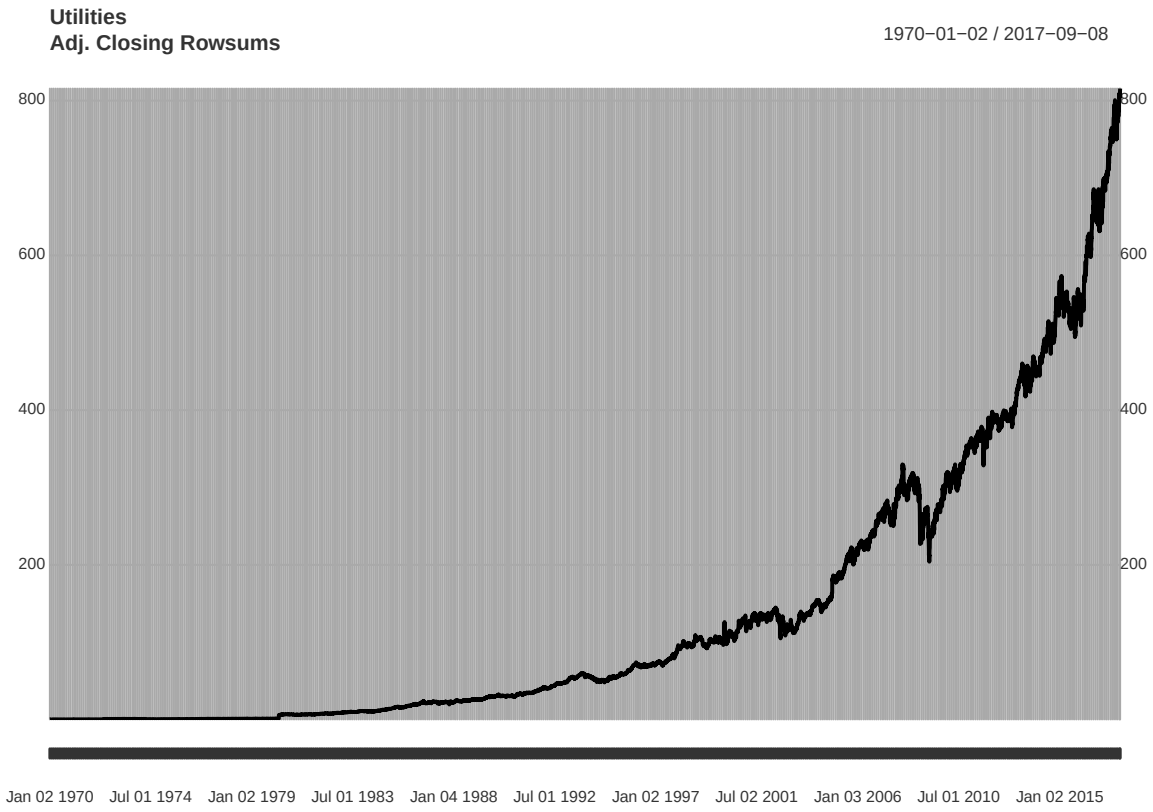

```
plot(sector_rowsums_ls[[7]],main=paste(levels(compdata_df$Sector)[7],
                                       "\nAdj. Closing Rowsums",sep=""))
```




```
plot(sector_rowsums_ls[[8]],main=paste(levels(compdata_df$Sector)[8],
                                       "\nAdj. Closing Rowsums",sep=""))
```

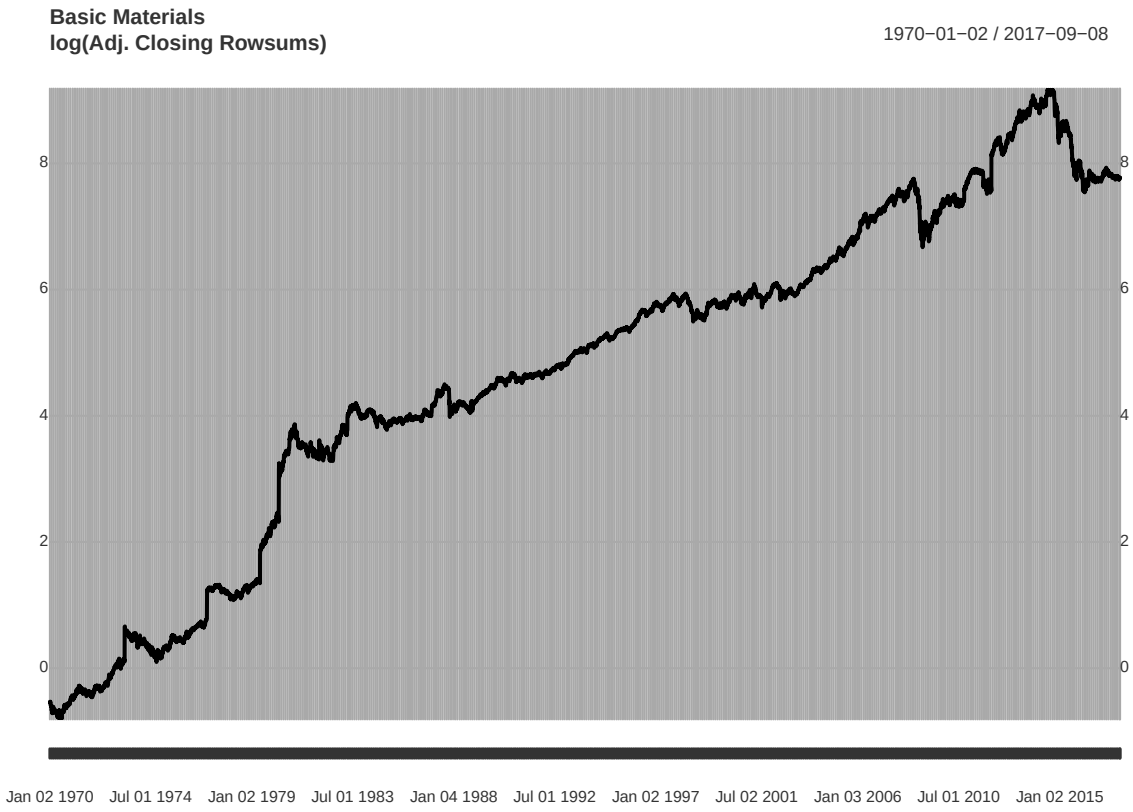



```
plot(sector_rowsums_ls[[9]],main=paste(levels(compdata_df$Sector)[9],
                                       "\nAdj. Closing Rowsums",sep=""))
```



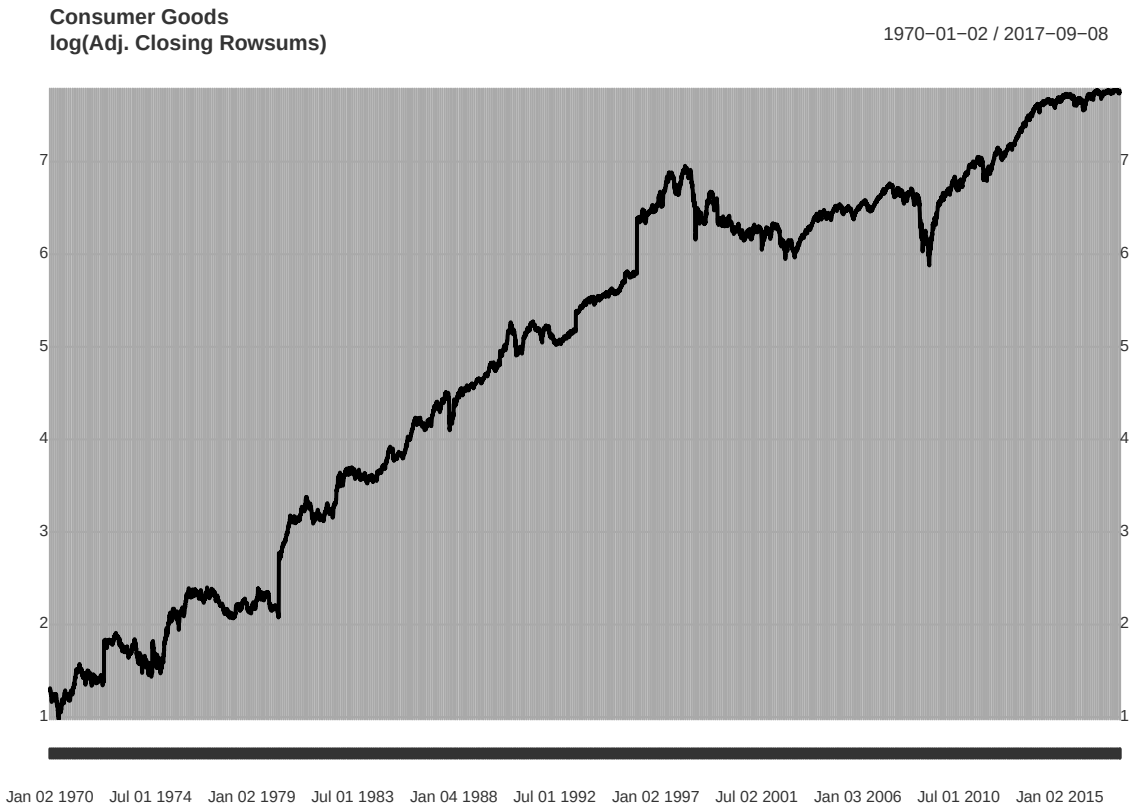
Plots of the log of daily adjusted closing price row sums for each sector. Note: Rows containing zeroes have been removed.

```
plot(log(sector_rowsums_ls[[1]]),main=paste(levels(compdata_df$Sector)[1],
                                        "\nlog(Adj. Closing Rowsums)",sep=""))
```

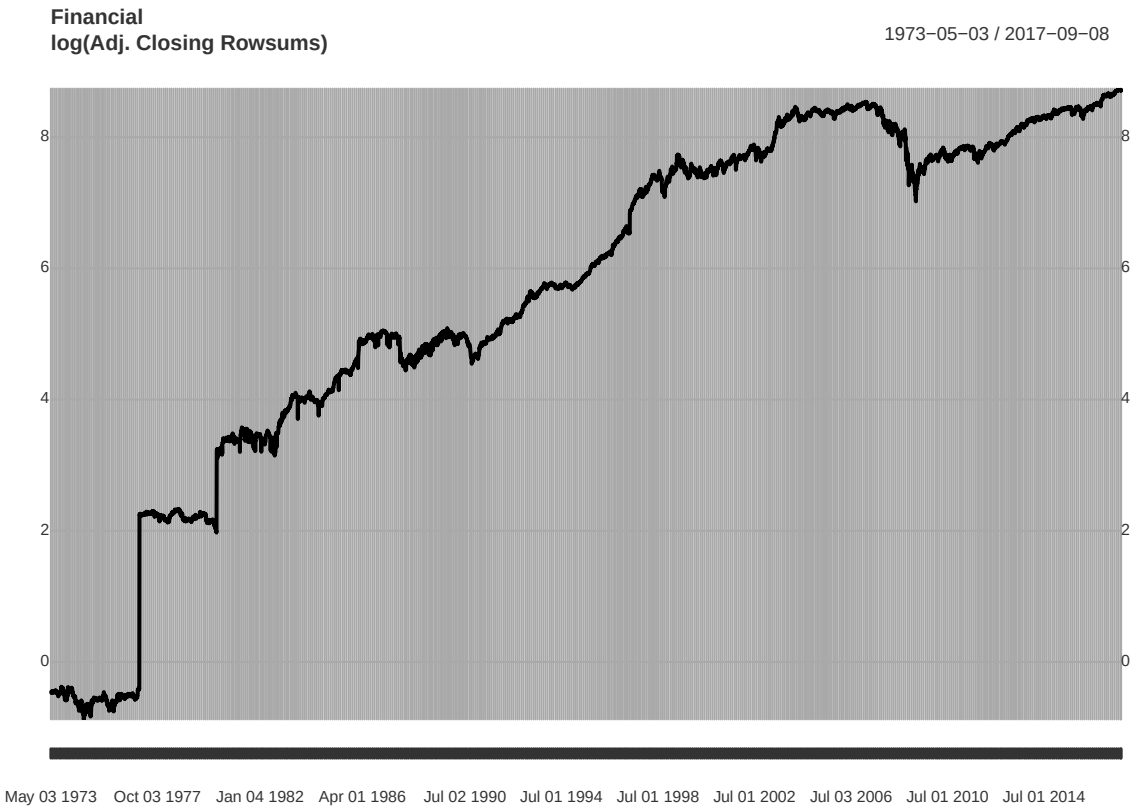



```
plot(log(sector_rowsums_ls[[2]]),main=paste(levels(compdata_df$Sector)[2],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```

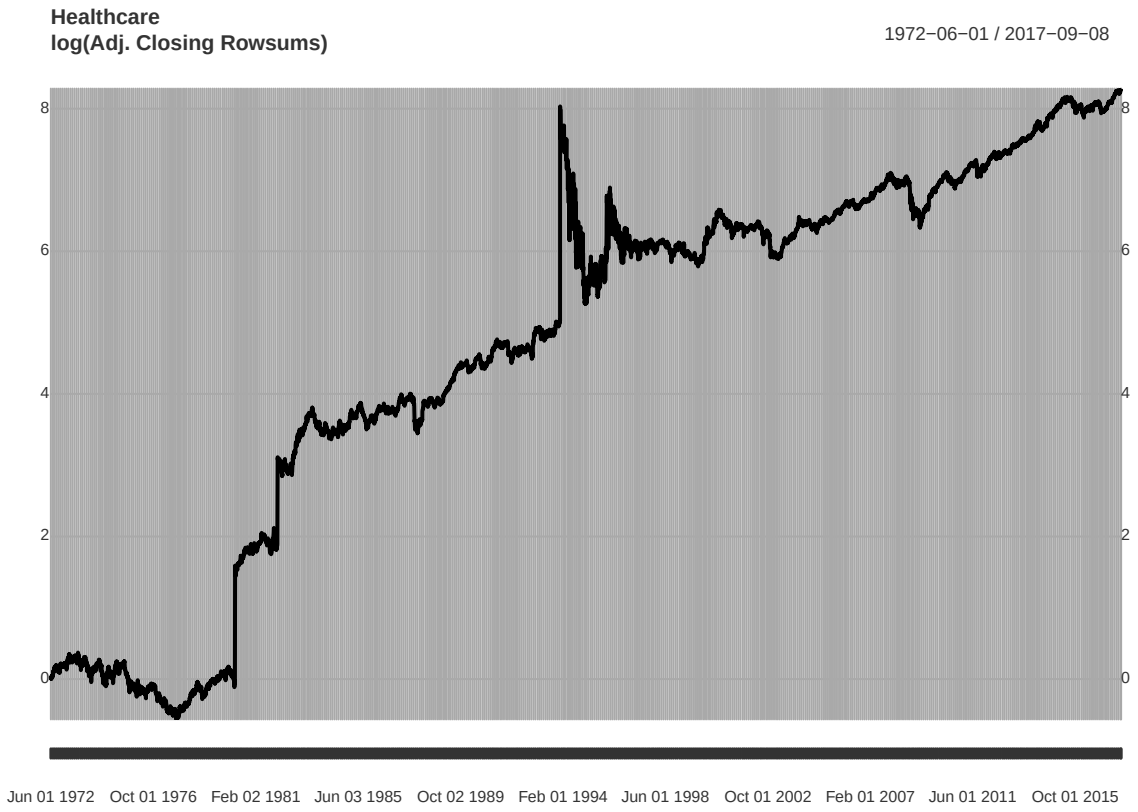

```
plot(log(sector_rowsums_ls[[3]]),main=paste(levels(compdata_df$Sector)[3],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```



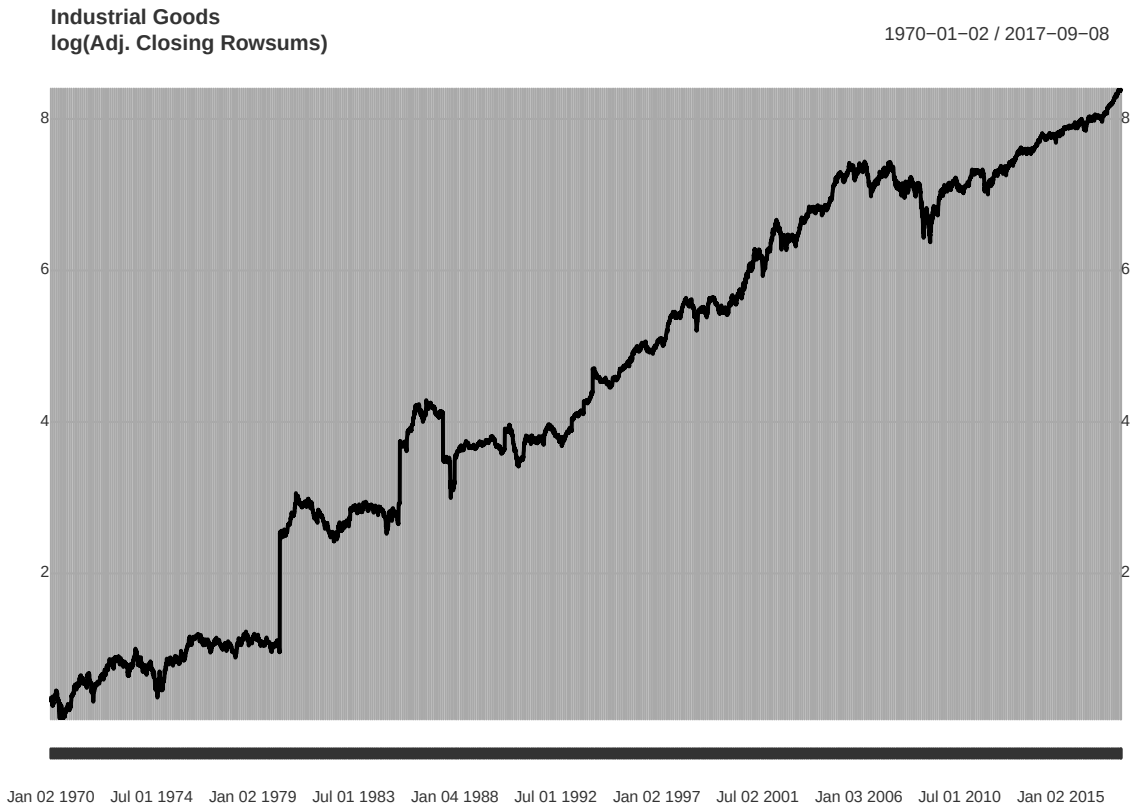

```
plot(log(sector_rowsums_ls[[4]]),main=paste(levels(compdata_df$Sector)[4],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```



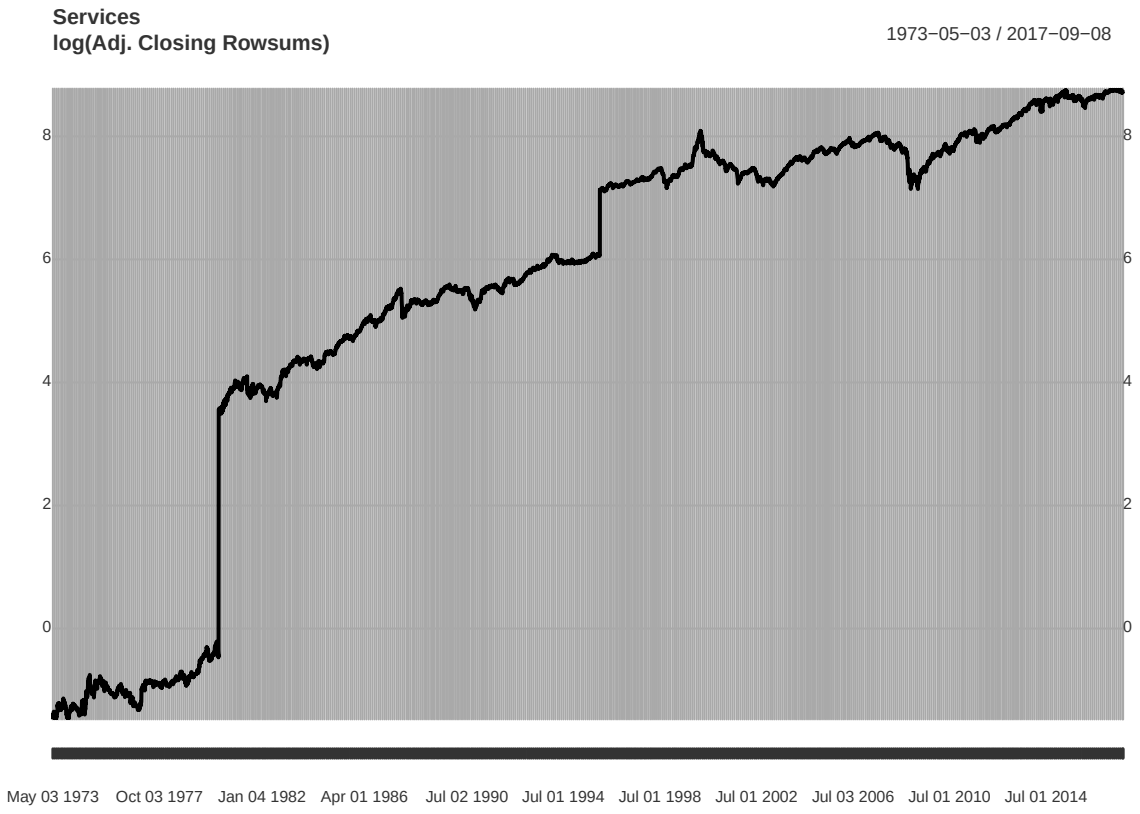

```
plot(log(sector_rowsums_ls[[5]]),main=paste(levels(compdata_df$Sector)[5],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```




```
plot(log(sector_rowsums_ls[[6]]),main=paste(levels(compdata_df$Sector)[6],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```

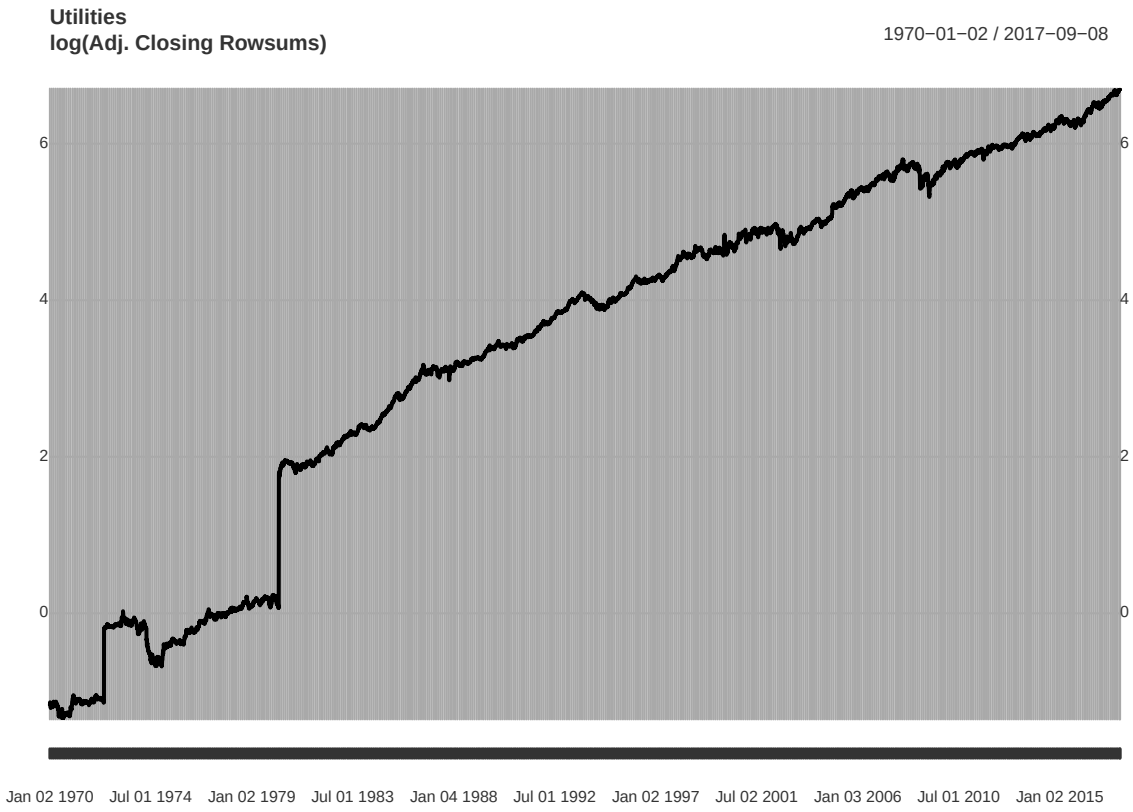



```
plot(log(sector_rowsums_ls[[7]]),main=paste(levels(compdata_df$Sector)[7],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```



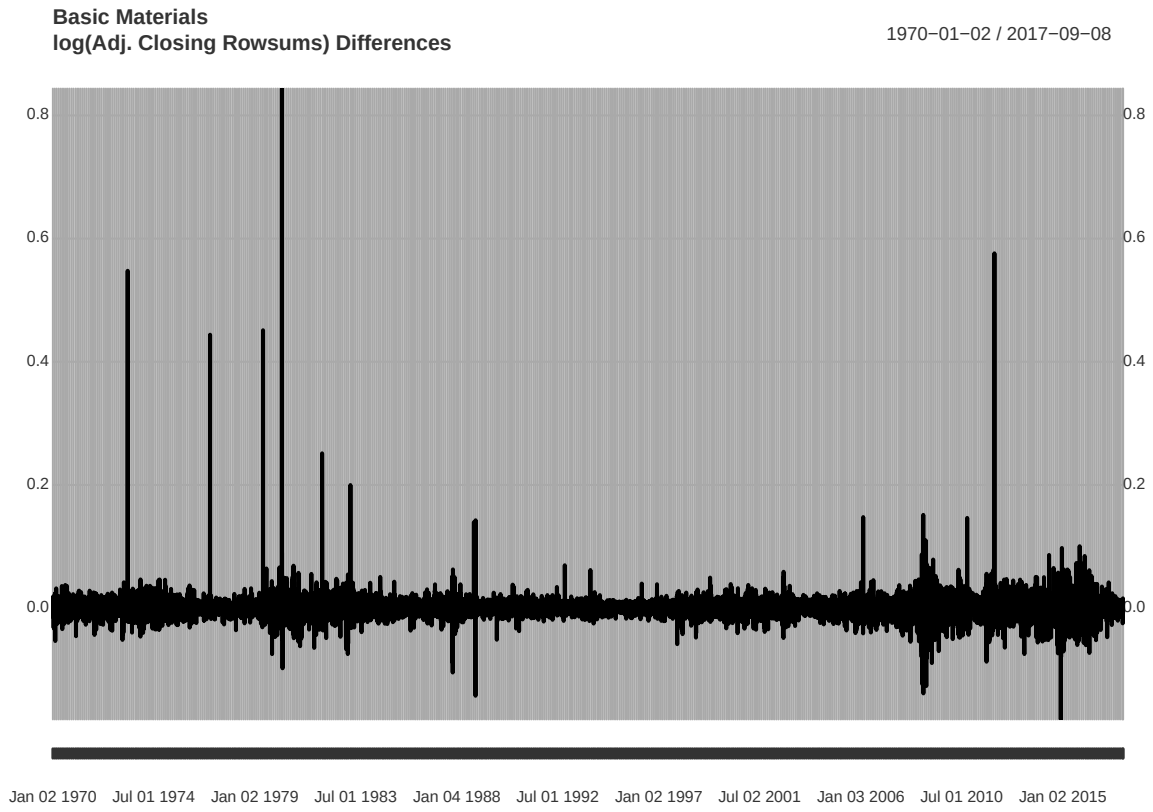

```
plot(log(sector_rowsums_ls[[8]]),main=paste(levels(compdata_df$Sector)[8],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```


```
plot(log(sector_rowsums_ls[[9]]),main=paste(levels(compdata_df$Sector)[9],
                                    "\nlog(Adj. Closing Rowsums)",sep=""))
```

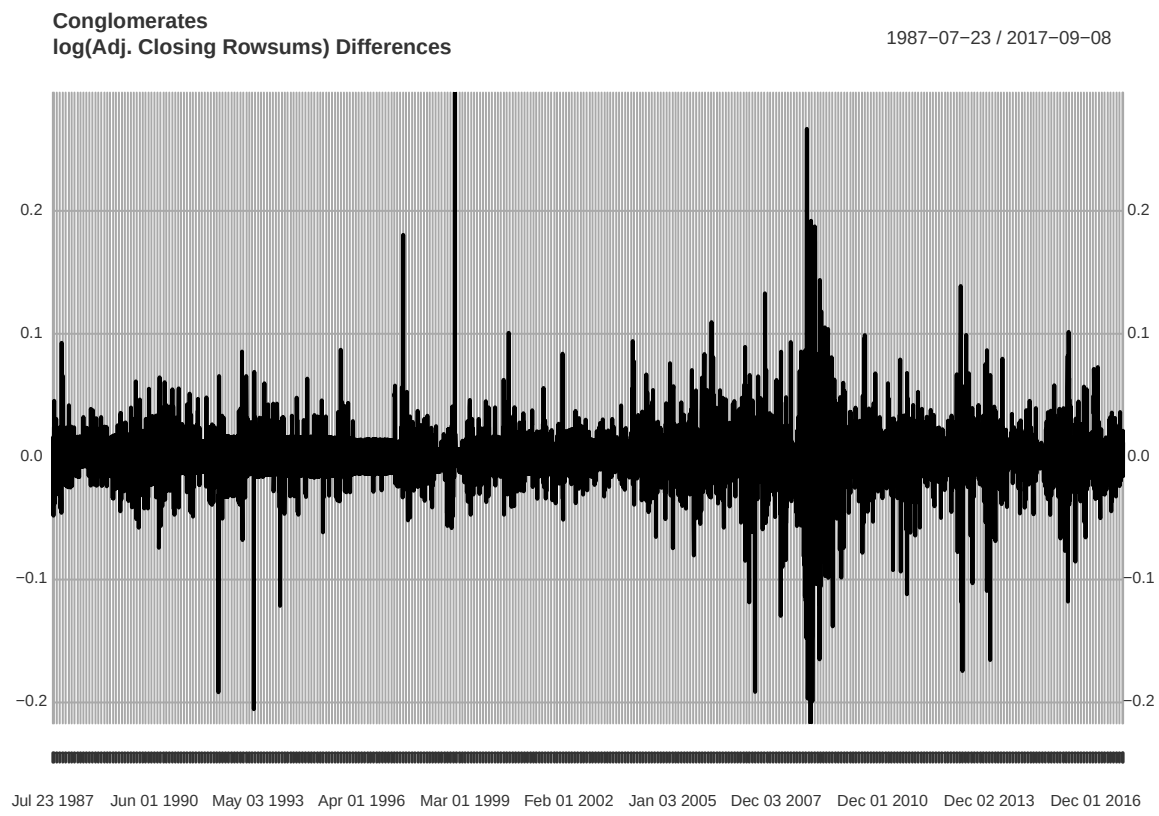


Plots of daily log-differences by sector.

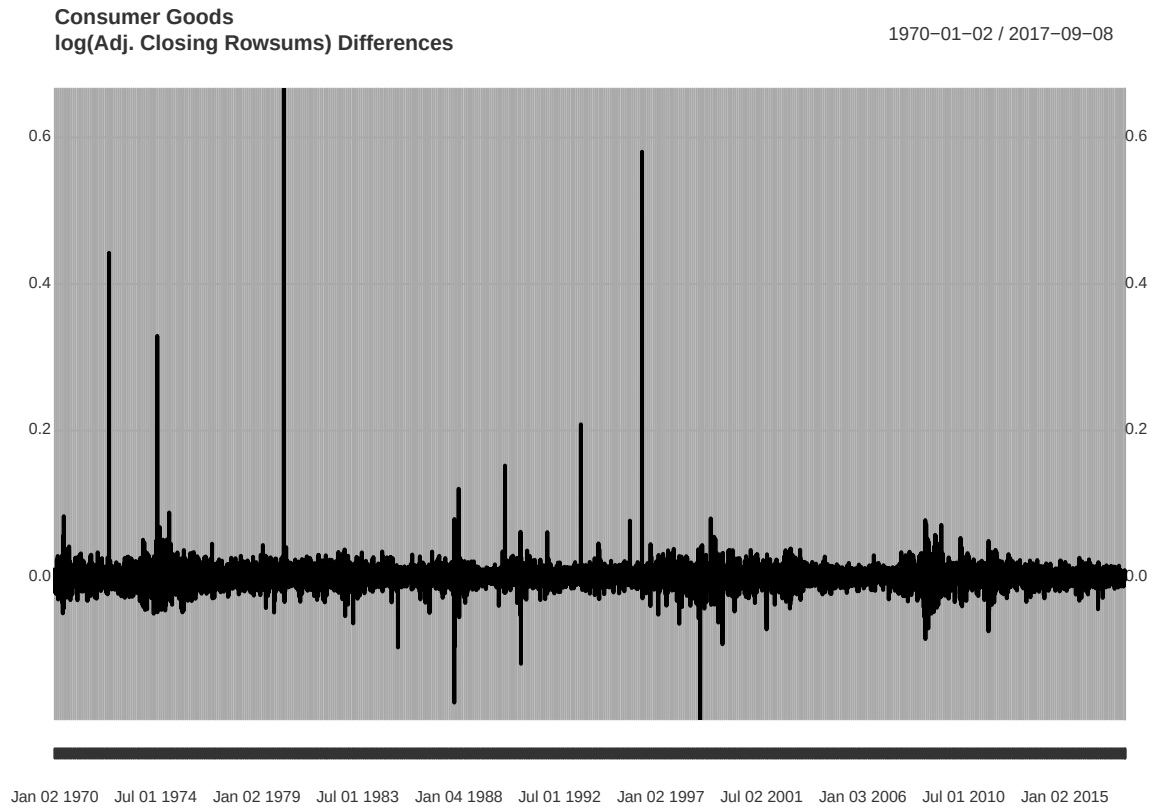
```
plot(diff.xts(log(sector_rowsums_ls[[1]])),main=paste(levels(compdata_df$Sector)[1],
                                        "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```



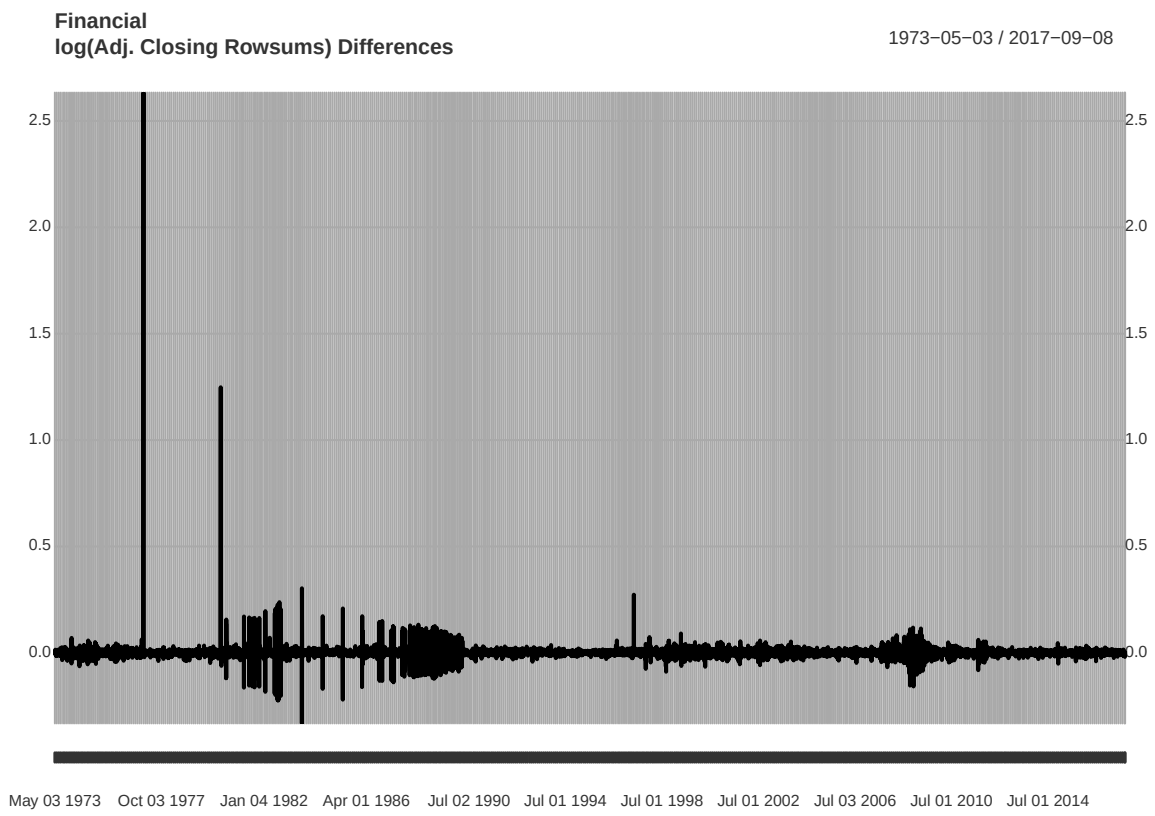

```
plot(diff.xts(log(sector_rowsums_ls[[2]])),main=paste(levels(compdata_df$Sector)[2],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```



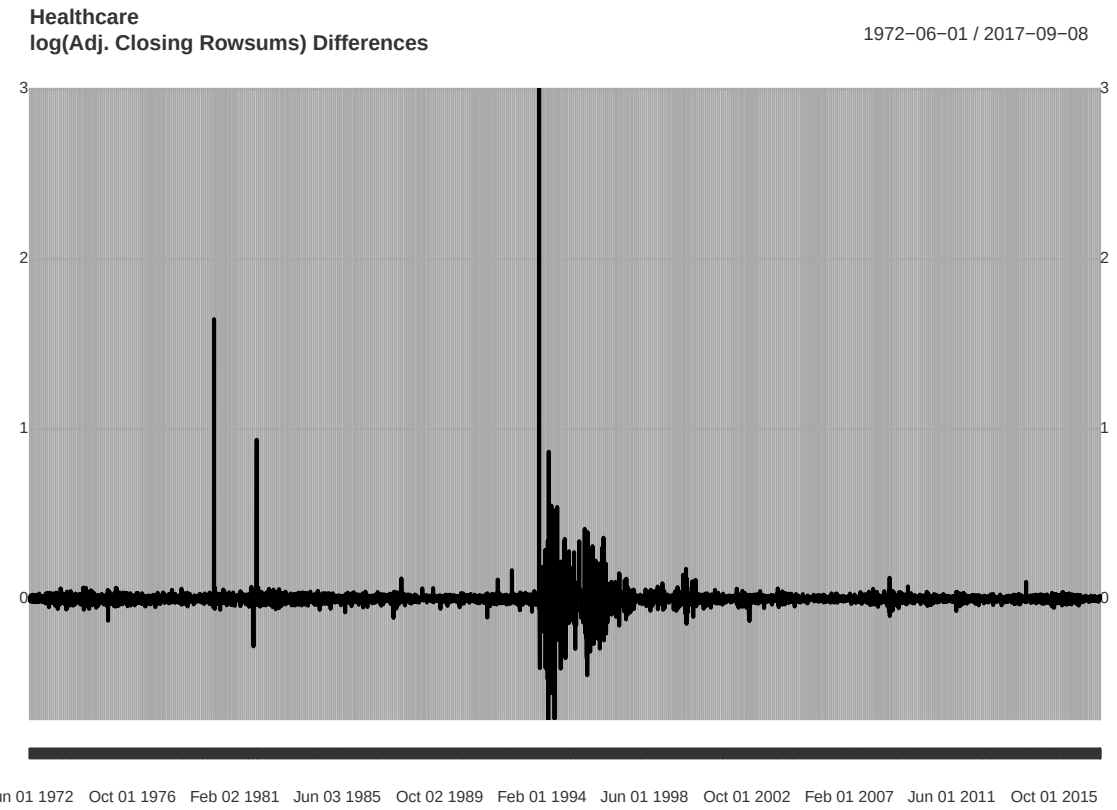

```
plot(diff.xts(log(sector_rowsums_ls[[3]])),main=paste(levels(compdata_df$Sector)[3],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```



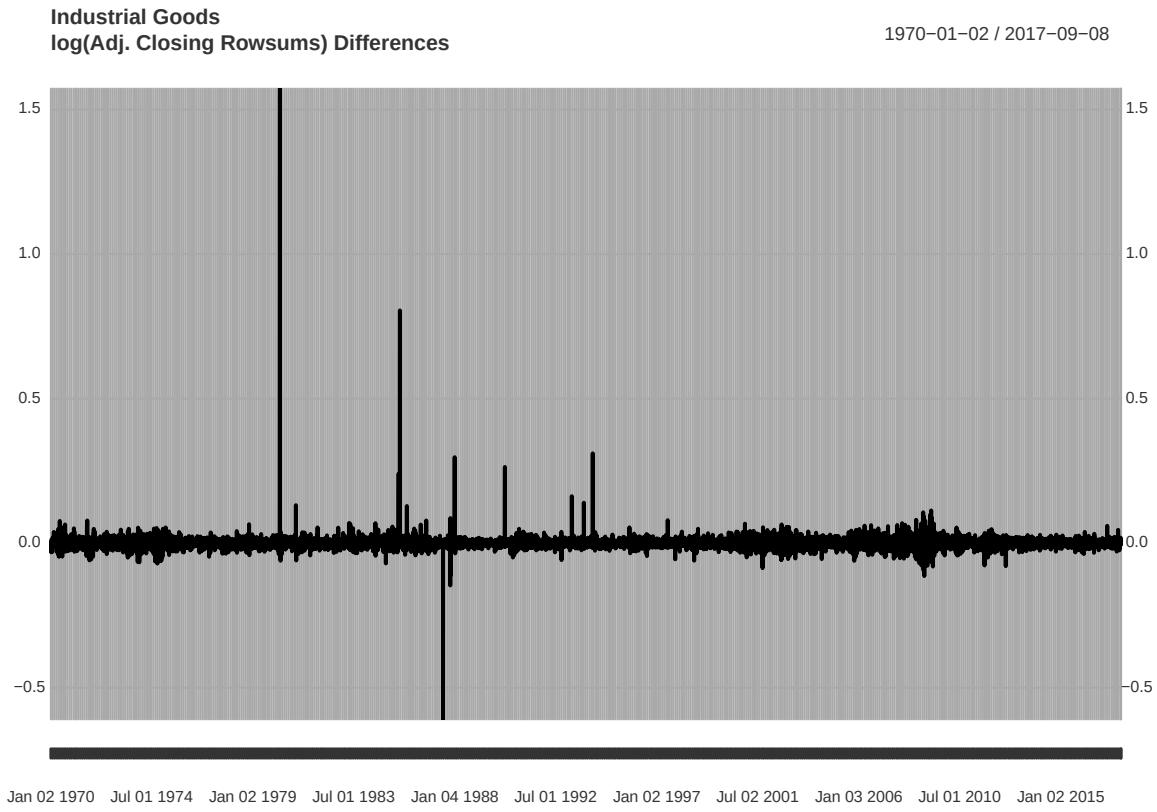

```
plot(diff.xts(log(sector_rowsums_ls[[4]])),main=paste(levels(compdata_df$Sector)[4],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```



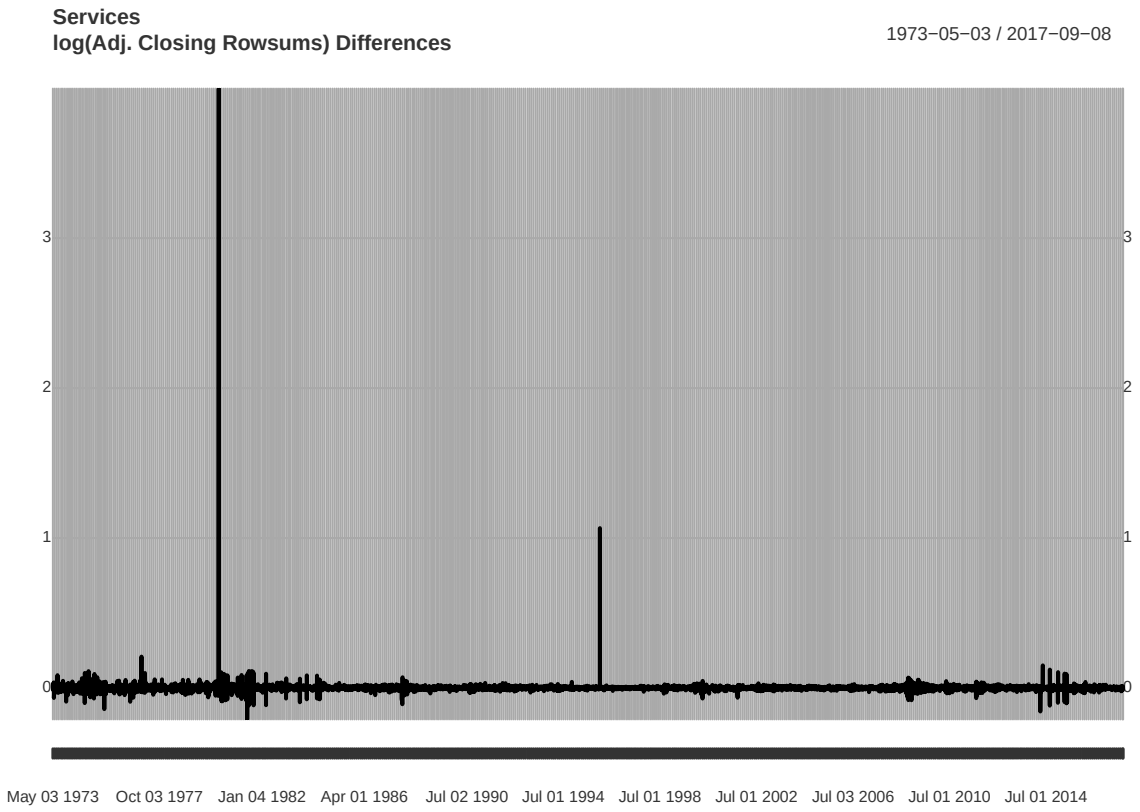

```
plot(diff.xts(log(sector_rowsums_ls[[5]])),main=paste(levels(compdata_df$Sector)[5],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```



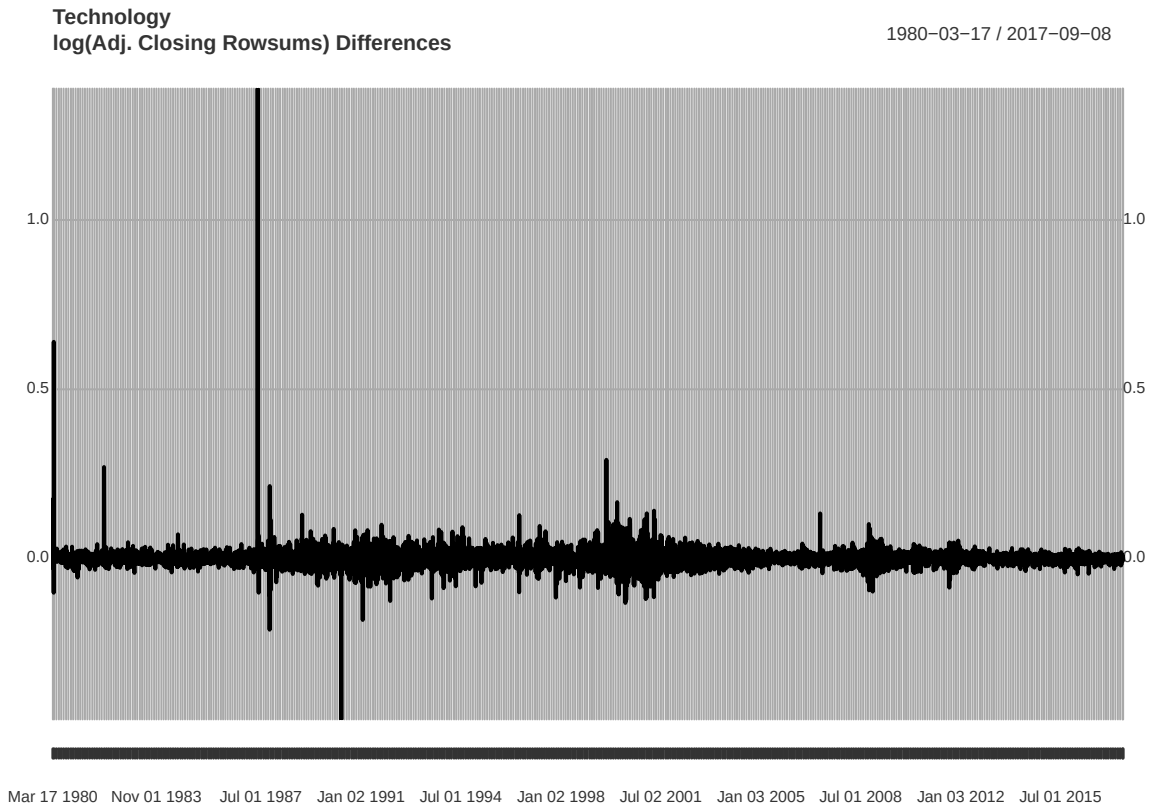

```
plot(diff.xts(log(sector_rowsums_ls[[6]])),main=paste(levels(compdata_df$Sector)[6],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```



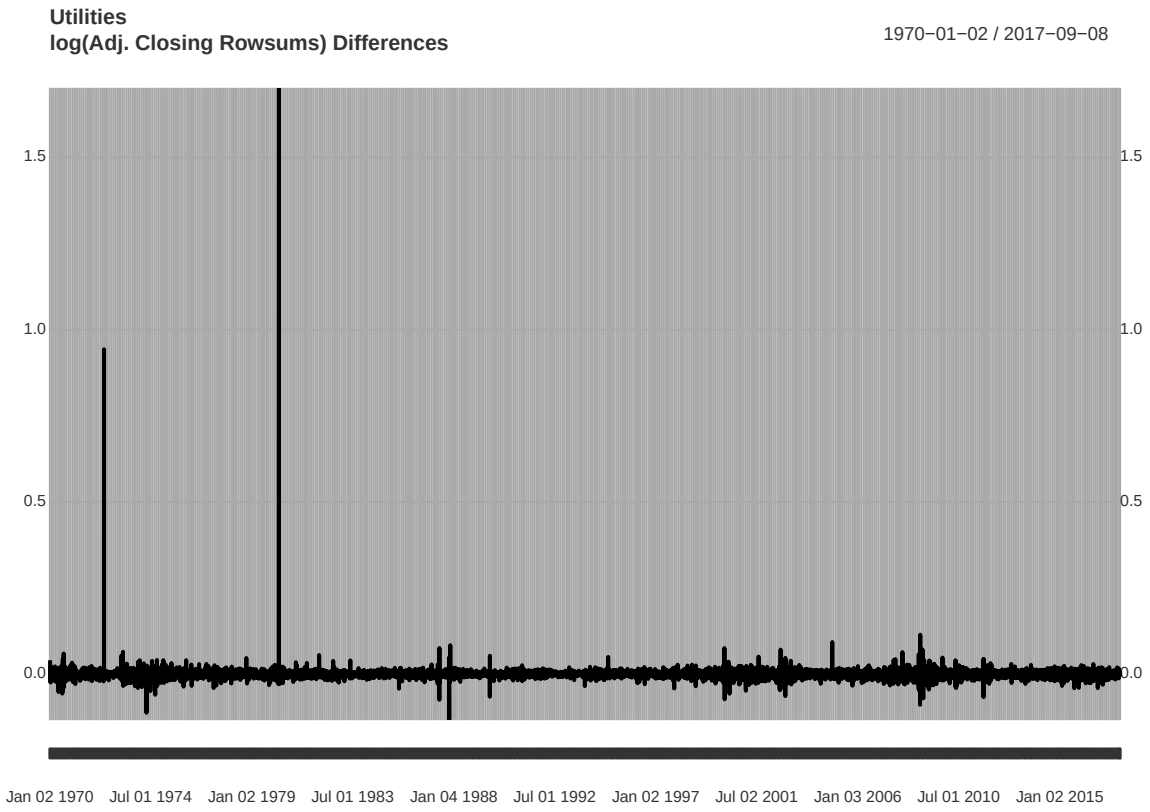

```
plot(diff.xts(log(sector_rowsums_ls[[7]])),main=paste(levels(compdata_df$Sector)[7],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```




```
plot(diff.xts(log(sector_rowsums_ls[[8]])),main=paste(levels(compdata_df$Sector)[8],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```




```
plot(diff.xts(log(sector_rowsums_ls[[9]])),main=paste(levels(compdata_df$Sector)[9],
                                                      "\nlog(Adj. Closing Rowsums) Differences",sep=""))
```


Utilities
log(Adj. Closing Rowsums) Differences
1970−01−02 / 2017−09−08
1.5
1.0
0.5
0.0
Jan 02 1970
Jul 01 1974
Jan 02 1979
Jul 01 1983
Jan 04 1988
Jul 01 1992
Jan 02 1997
Jul 02 2001
Jan 03 2006
Jul 01 2010
Jan 02 2015