Time Series clustering

Aayush Agrawal

October 11, 2017

1) Installing required packages

```
library(Quandl) ## Pulling data from Quandl
```

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

```
library(ggplot2) ## For visualization
library(gridExtra) ## Visualizing in grid
library(ggdendro) ## For dendograms visualization
library(zoo) ## Time series data manipulation
library(TSclust) ## Time series clustering and distance calculation
```

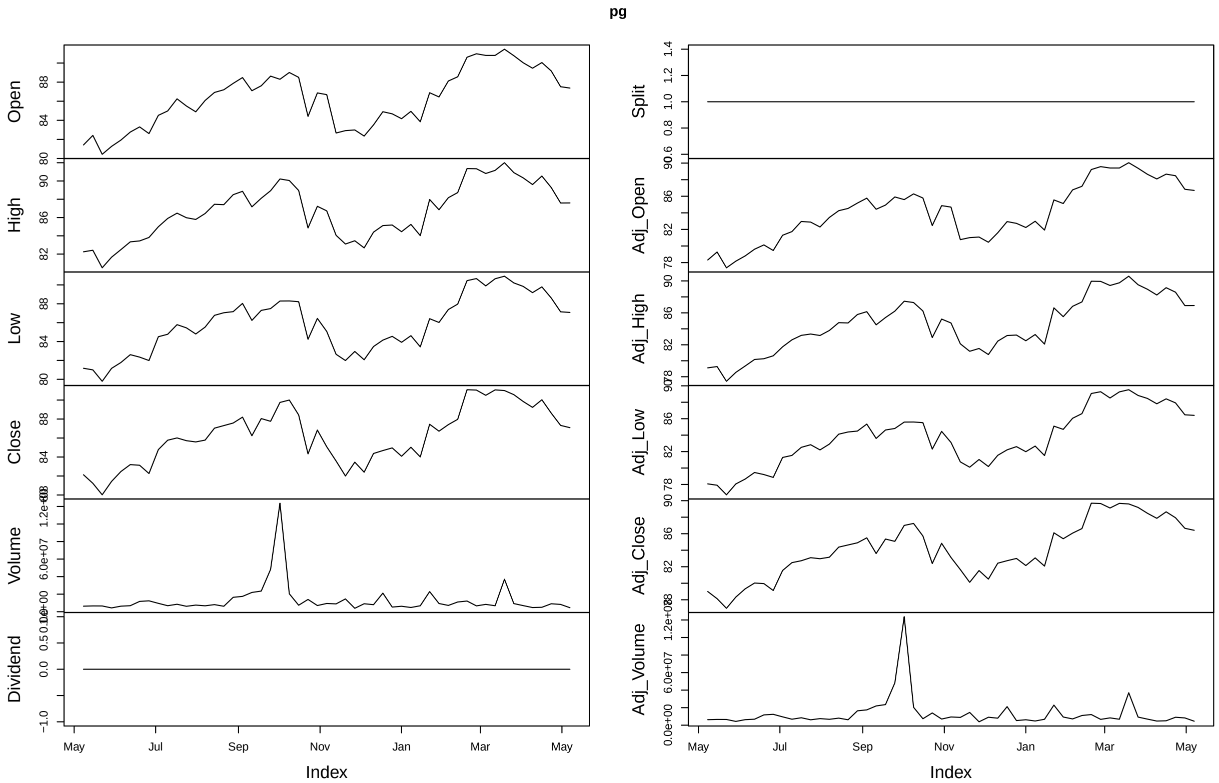
```
## Loading required package: wmtsa

## Loading required package: pdc

## Loading required package: cluster
```

2) Pulling stock prices data from Quandl

```
#Quandl.api_key('api here')
pg <- Quandl('EOD/PG', start_date="2016-05-01", end_date='2017-05-01',
             collapse='weekly', type='zoo')
apple <- Quandl('EOD/AAPL', start_date="2016-05-01", end_date='2017-05-01',
                collapse='weekly', type='zoo')
visa <- Quandl('EOD/V', start_date="2016-05-01", end_date='2017-05-01',
               collapse='weekly', type='zoo')
uhg <- Quandl('EOD/UNH', start_date="2016-05-01", end_date='2017-05-01',
              collapse='weekly', type='zoo')
cocacola <- Quandl('EOD/KO', start_date="2016-05-01", end_date='2017-05-01',
                   collapse='weekly', type='zoo')
goldmansach <- Quandl('EOD/GS', start_date="2016-05-01", end_date='2017-05-01',
                      collapse='weekly', type='zoo')
walmart <- Quandl('EOD/WMT', start_date="2016-05-01", end_date='2017-05-01',
                  collapse='weekly', type='zoo')
merk <- Quandl('EOD/MRK', start_date="2016-05-01", end_date='2017-05-01',
               collapse='weekly', type='zoo')
```

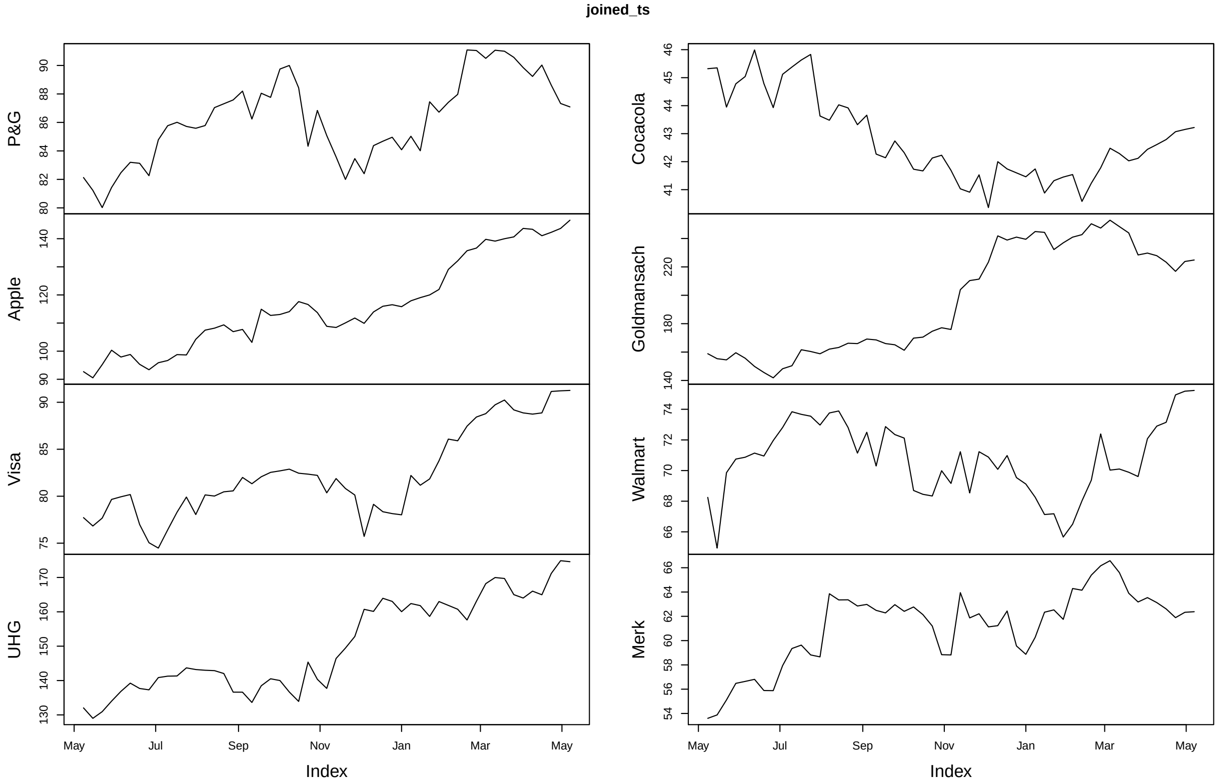
Plot P&G's data to observe what we downloaded from Quandl

```r
# Plot the time series for Procter and Gamble
plot(pg)
```



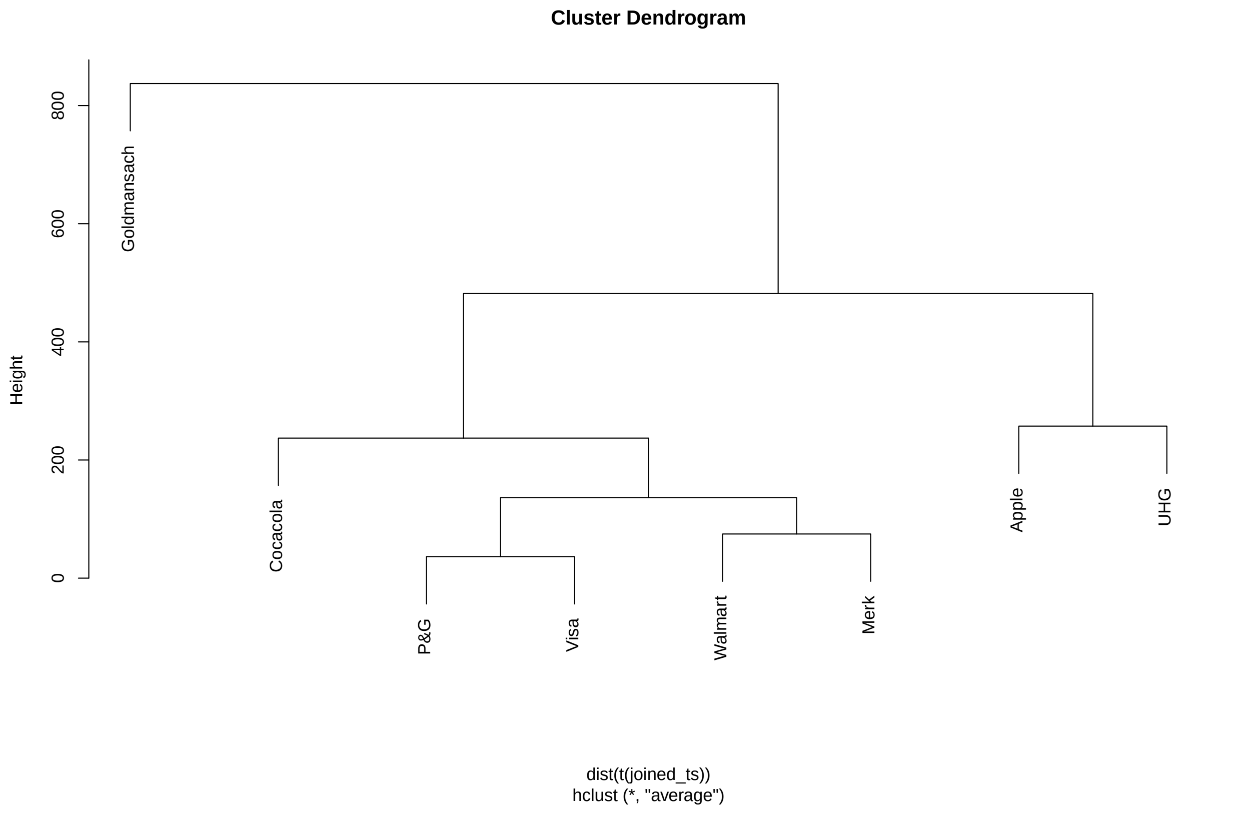
We are actually only interested in looking at Weekly closing prices of stock. So let's take weekly closing price of each stock and stack them together.

```r
# Merge and plot the time series (just the closing price)
joined_ts <- cbind(pg[,4], apple[,4], visa[,4], uhg[,4], cocacola[,4], goldmansach[,4], walmart[,4], mer
names(joined_ts) <- c('P&G', 'Apple', 'Visa', 'UHG', 'Cocacola', 'Goldmansach', 'Walmart', 'Merk')
plot(joined_ts)
```

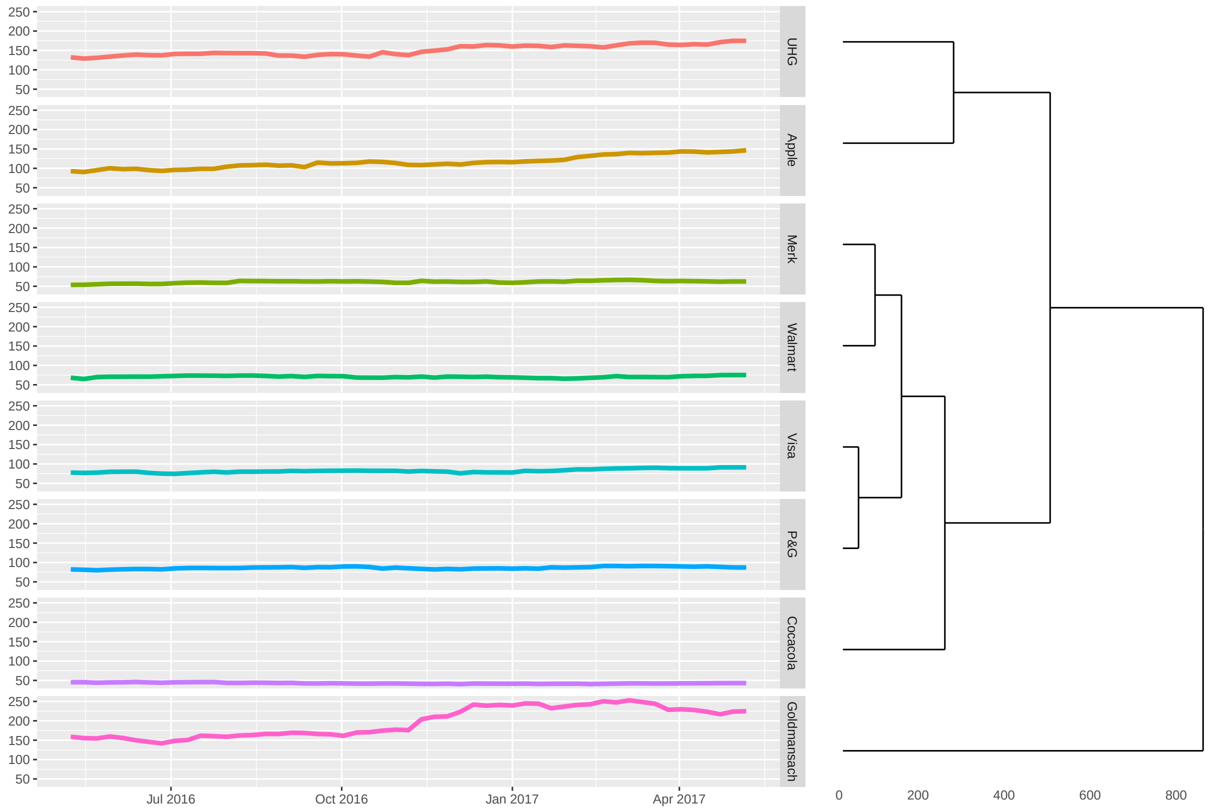


3) Case 1 : The problem with time series data

```
## Hierarchical clustering with average linkage
hc <- hclust(dist(t(joined_ts)), "ave")
## Plotting dendogram
plot(hc)
## colour the tree at different levels by changing the h value
colours_hc <- cutree(hc, h=2)

## Plot
hcdata <- dendro_data(hc)
names_order <- hcdata$labels$label
## Use the folloing to remove labels from dendogram so not doubling up - but good for checking
hcdata$labels$label <- ''
p1 <- ggdendrogram(hcdata, rotate=TRUE, leaf_labels=FALSE)

new_data <- joined_ts[,rev(as.character(names_order))]
p2 <- autoplot(new_data, facets = Series ~ . ) +
  aes(colour=as.character(rep(colours_hc,each=53)), linetype = NULL) +geom_line(size=1.5) +
  xlab('') + ylab('') + theme(legend.position="none")

gp1<-ggplotGrob(p1)
```



```
gp2<-ggplotGrob(p2)


grid.arrange(gp2, gp1, ncol=2, widths=c(4,2))
```



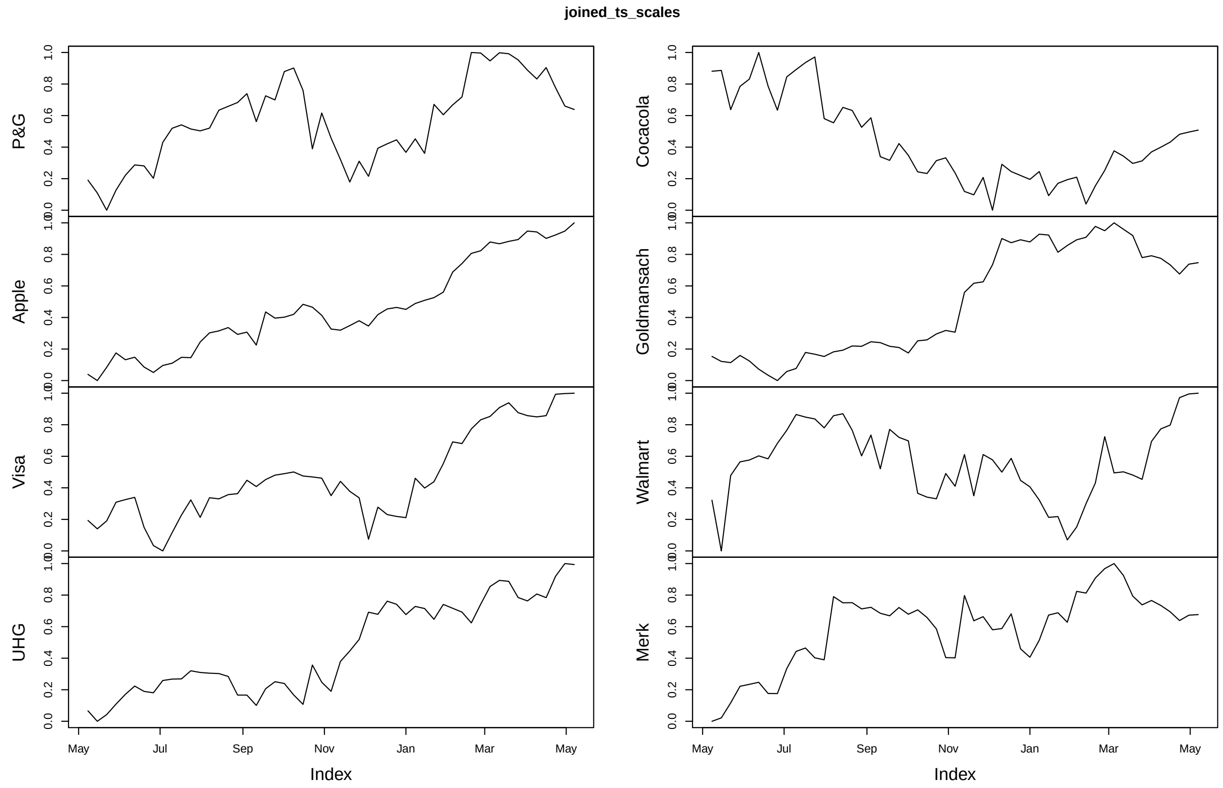
As we can see UHG and Apple got clustered together because there stock price is between 100-150 USD. Merk, Walmart, Visa, P&G, Cocala got clustered together because their stock price is between 50-100 USD. Goldmansach is clustered separately because it's stock price varies between 150 to 250 USD.So the problem of scaling effects hierarchical clustering to not capture trending information together.

How can we Solve this?

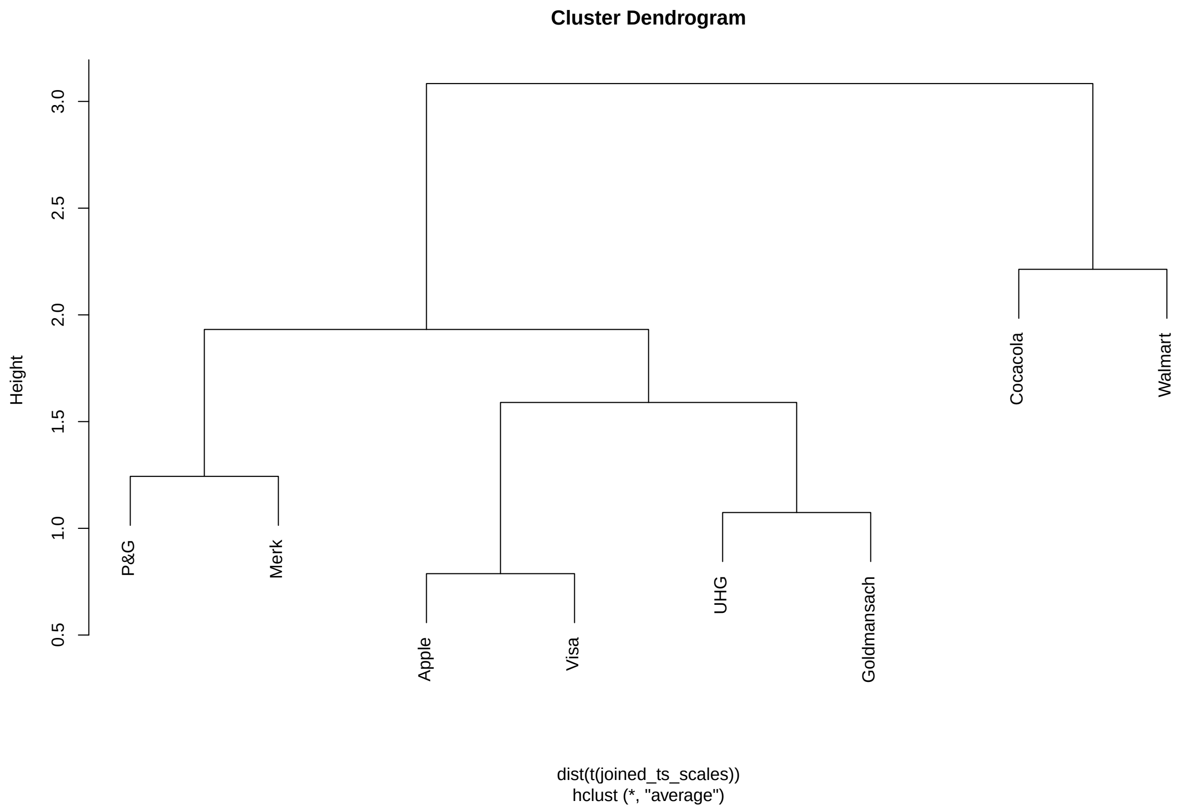
4) Case 2 : The min max scaling solution

Let's try preprocessing data with min-max scaling within each companie's weekly closing stock prices.

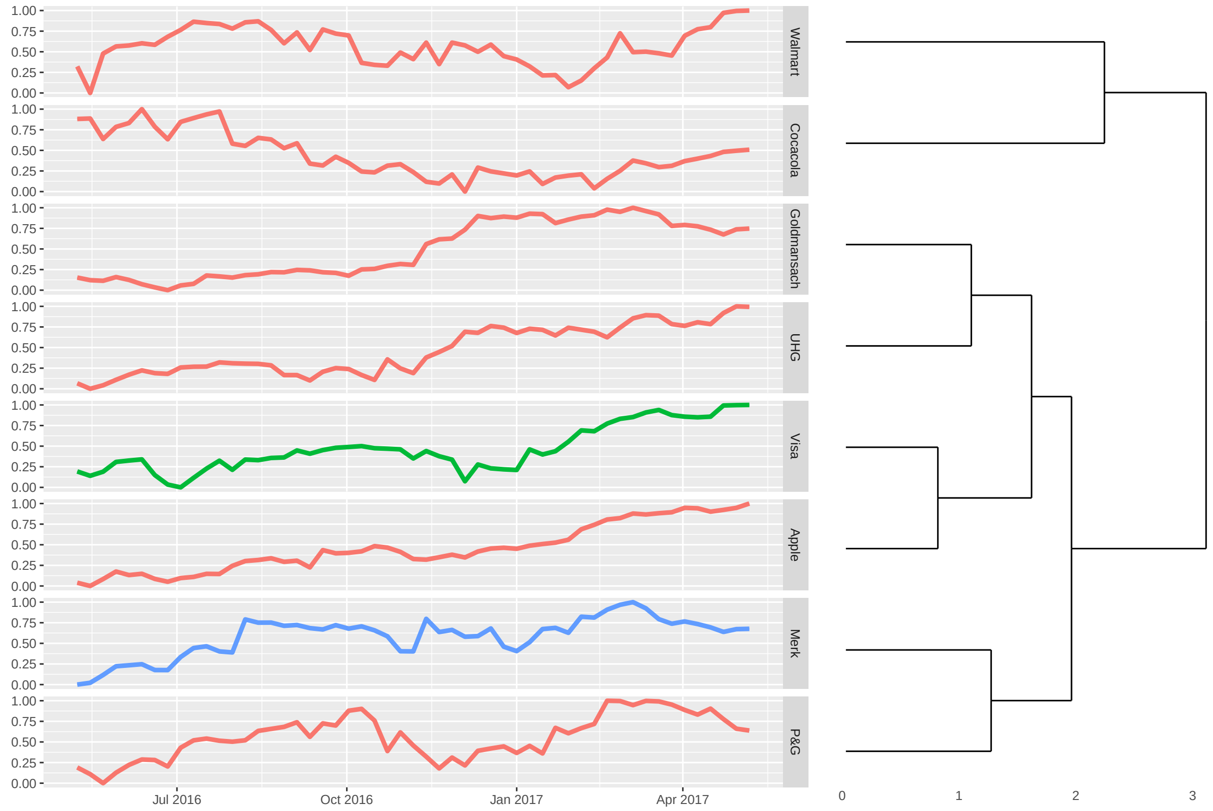
```r
# Scale the time series and plot
maxs <- apply(joined_ts, 2, max)
mins <- apply(joined_ts, 2, min)
joined_ts_scales <- scale(joined_ts, center = mins, scale = maxs - mins)
plot(joined_ts_scales)
```



```
## Hierarchical clustering with average linkage
hc <- hclust(dist(t(joined_ts_scales)), "ave")
## Plotting dendogram
plot(hc)
## colour the tree at different levels by changing the h value
colours_hc <- cutree(hc, h=2)

### Plot
hcdata <- dendro_data(hc)
names_order <- hcdata$labels$label
# Use the folloing to remove labels from dendogram so not doubling up
hcdata$labels$label <- ''
p1 <- ggdendrogram(hcdata, rotate=TRUE, leaf_labels=FALSE)

new_data <- joined_ts_scales[,rev(as.character(names_order))]
p2 <- autoplot(new_data, facets = Series ~ .) +
  aes(colour=as.character(rep(colours_hc,each=53)), linetype = NULL) + geom_line(size=1.5) +
  xlab('') + ylab('') + theme(legend.position="none")

gp1<-ggplotGrob(p1)
```



```
gp2<-ggplotGrob(p2)


grid.arrange(gp2, gp1, ncol=2, widths=c(4,2))
```



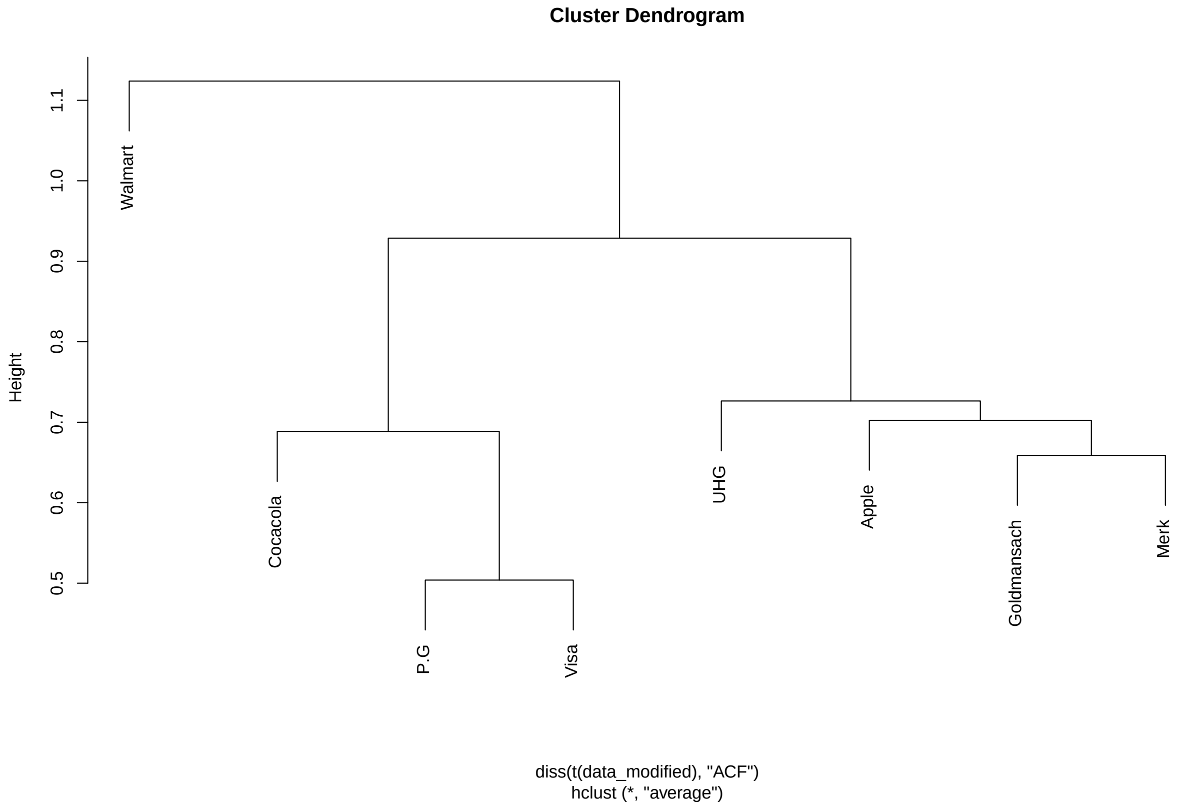
As we can see it improves our clustering solution but still Walmart and Cocacola getting grouped together is not right solution as Walmart has stable stock prices while Cocacola's share is declining. So what next?

4) Case 3 : Relative variation and autocorrelation (Recommeded)

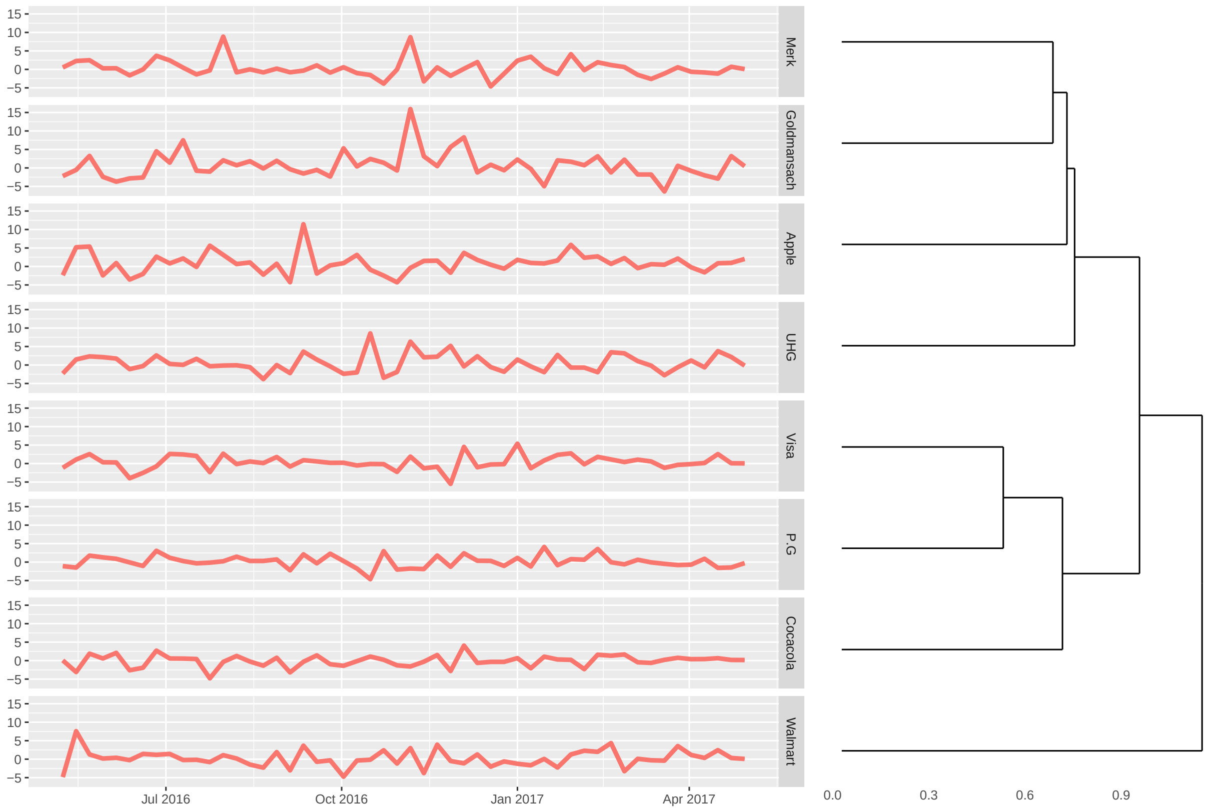
Let's try calculating the percentage change in weekly closing prices compared to last week and then use autocorrelation to calculate distance between different companies time series data.

```
## Copying the data
data <- data.frame(joined_ts)

## Reindexing the data
data_modified <- data
rownames(data_modified) = 1:nrow(data_modified)

## Calculating percentage change in weekly closing price compared to last week price
data_modified <- (data_modified[2:53,] - data_modified[1:52,])*100/data_modified[1:52,]

## Hierarchical clustering with average linkage
hc <- hclust(diss(t(data_modified),"ACF"), "ave")
## Plotting dendogram
plot(hc)

## colour the tree at different levels by changing the h value
colours_hc <- cutree(hc, h=2)

## Converting data back to zoo format for plotting
```

```
rownames(data_modified) <- rownames(data)[1:52]
data_modified <- as.matrix(data_modified)
data_modified <- xts(data_modified,as.POSIXct(rownames(data_modified)))

### Plot
hcdata <- dendro_data(hc)
names_order <- hcdata$labels$label

# Use the folloing to remove labels from dendogram so not doubling up - but good for checking
hcdata$labels$label <- ''
p1 <- ggdendrogram(hcdata, rotate=TRUE, leaf_labels=FALSE)

new_data <- data_modified[,rev(as.character(names_order))]
p2 <- autoplot(new_data, facets = Series ~ .) +
  aes(colour=as.character(rep(colours_hc,each=52)), linetype = NULL) + geom_line(size=1.5) +
  xlab('') + ylab('') + theme(legend.position="none")

gp1<-ggplotGrob(p1)
```



```
gp2<-ggplotGrob(p2)
grid.arrange(gp2, gp1, ncol=2, widths=c(4,2))
grid.arrange(gp2, gp1, ncol=2, widths=c(4,2))
```



As we can see now Merk, GoldmanSach, Apple & UHG which were always increasing stocks got clustered together. Visa, P&G and Cocala which has shown increase in Jan to May 2017 period got clustered together. Walmart whose stock has been relatively stable has been a separate cluster. So in this case we were truly able to capture the trend within the stock prices of different companies.