STAT GR5241 HW1_Q3_mjs2364

Mathieu Sauterey - UNI: MJS2364

January 29th, 2017

Problem 3 - Principal Component Analysis

1. For each of the 30 stocks in the Dow Jones Industrial Average, download the closing prices for every trading day from January 1, 2017 to January 1, 2018.

```
library(XML)
urlt <- "http://money.cnn.com/data/dow30/"  #URL of website containing data
doc.html = htmlTreeParse(urlt, useInternal = TRUE) #captures HTML code
tables <- readHTMLTable(doc.html,as.data.frame=FALSE) #extracts table from it
tables <- readHTMLTable(doc.html,stringsAsFactors=FALSE,which = 2) #get
tickers

# Credits for the section of code above go to stackoverflow
(questions/29736119)

seg <- sapply(FUN=strsplit, tables$Company, split="\\s")
seg <- sapply(FUN=head, seg, n=1)
#seg <- unname(seg)
ticker <- sort(seg)

library(quantmod)

getSymbols(ticker, auto.assign = T, from = "2017-01-01", to = "2018-01-01")

##  [1] "AAPL" "AXP"  "BA"   "CAT"  "CSCO" "CVX"  "DIS"  "DWDP" "GE"   "GS"
## [11] "HD"   "IBM"  "INTC" "JNJ"  "JPM"  "KO"   "MCD"  "MMM"  "MRK"  "MSFT"
## [21] "NKE"  "PFE"  "PG"   "TRV"  "UNH"  "UTX"  "V"    "VZ"   "WMT"  "XOM"

Prices <- do.call(merge, lapply(ticker, function(x) Cl(get(x))))

# Credits for the last line of code go to stackoverflow (questions/5574595)
```

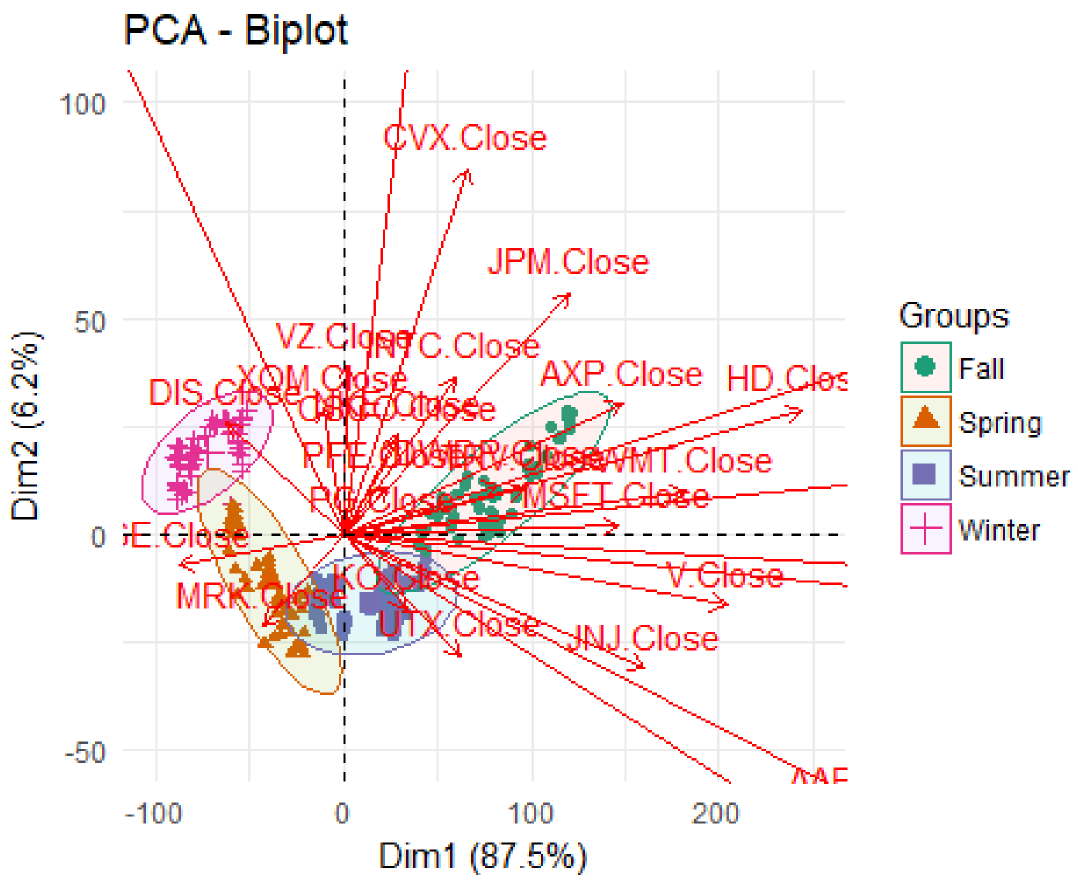
```
head(Prices,3)
```

```
##            AAPL.Close AXP.Close BA.Close CAT.Close CSCO.Close CVX.Close
## 2017-01-03     116.15     75.35   156.97     93.99      30.54    117.85
## 2017-01-04     116.02     76.26   158.62     93.57      30.10    117.82
## 2017-01-05     116.61     75.32   158.71     93.00      30.17    117.31
##            DIS.Close DWDP.Close GE.Close GS.Close HD.Close IBM.Close
## 2017-01-03    106.08      57.60    31.69   241.57   134.31    167.19
## 2017-01-04    107.44      58.06    31.70   243.13   135.50    169.26
## 2017-01-05    107.38      57.80    31.52   241.32   133.90    168.70
##            INTC.Close JNJ.Close JPM.Close KO.Close MCD.Close MMM.Close
## 2017-01-03      36.60    115.84     87.23    41.80    119.62    178.05
## 2017-01-04      36.41    115.65     86.91    41.65    119.48    178.32
## 2017-01-05      36.35    116.86     86.11    41.75    119.70    177.71
##            MRK.Close MSFT.Close NKE.Close PFE.Close PG.Close TRV.Close
## 2017-01-03     60.15      62.58     51.98     33.00    84.20    120.90
## 2017-01-04     60.13      62.30     53.07     33.29    84.50    120.25
## 2017-01-05     60.11      62.30     53.06     33.61    85.06    118.33
##            UNH.Close UTX.Close V.Close VZ.Close WMT.Close XOM.Close
## 2017-01-03    161.45    110.83   79.50    54.58     68.66     90.89
## 2017-01-04    161.91    110.90   80.15    54.52     69.06     89.89
## 2017-01-05    162.18    111.35   81.09    54.64     69.21     88.55
```

2. Perform a PCA on the closing prices and create the biplot (call function princomp() and use cor=FALSE). Do you see any structure in the biplot, perhaps in terms of the types of stocks? How about the screeplot - how many important components seem to be in the data?

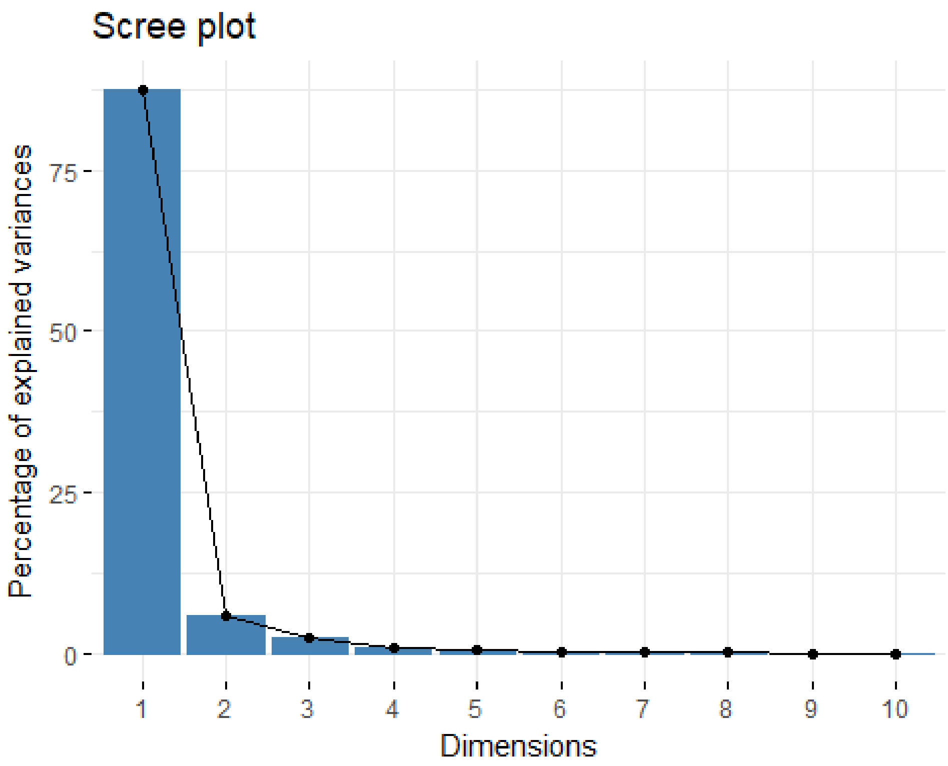
```
# Loads the required packages to perform an efficient PCA
library("FactoMineR")

library("devtools")

library("factoextra")

# Adds a column indicating each day's season
Season <- c(rep("Winter",times = 54),rep("Spring", each=64),rep("Summer",
each=64),rep("Fall", each=69))
Season <- factor(Season)
Prices2 <- data.frame(Prices,Season)

# Performs the PCA using the covariance matrix
pca.1 <- PCA(Prices2, quali.sup=31, scale.unit = FALSE)
```

```
# Plots the PCA Biplot and zooms in to better observe the results
fviz_pca_biplot(pca.1,
  habillage = Prices2$Season, addEllipses = TRUE,
  col.var = "red",
  label = "var",xlim = c(-100, 250),ylim = c(-50, 100)) +
  scale_color_brewer(palette="Dark2")+
  theme_minimal()

# Source for code above http://www.sthda.com/english/wiki/print.php?id=202
```



We observe that most of stock prices are correlated. In this case their loading arrows are pointing in the same direction (to the right along the First Principal Component) with small angles between them. This means that the prices also tend to be correlated. As we broke down the data into clusters representing the 4 seasons, we notice that most stocks had a higher price in Fall 2017 and lower price in Winter (January - March) and Spring 2017. This result is expected because stock prices tend to grow over time.

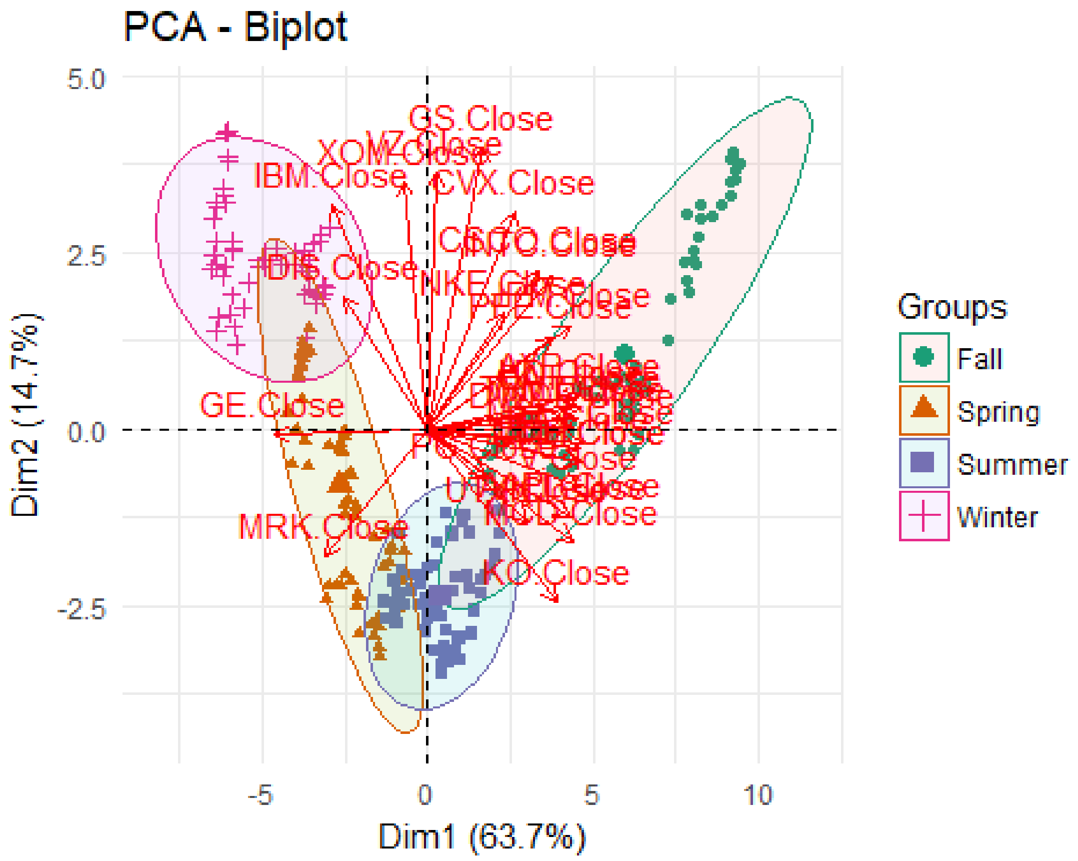
```
#Plots the PCA screeplot
fviz_screeplot(pca.1, ncp=10)
```



Based on the screeplot, there is one important Principal Component which explains 87.51% of the variance in the data. This agrees with the biplot where most loadings pointed along the first principal component axis which thus explains most of the variance in these loadings.

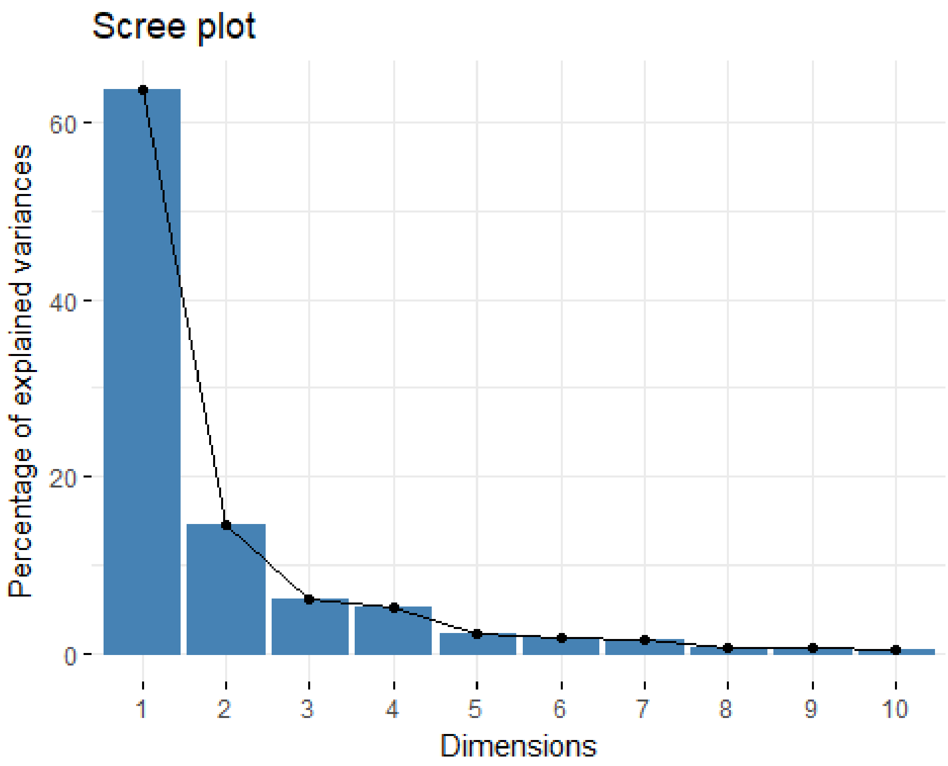
3. Repeat part 2 with cor=TRUE. This is equivalent to scale each column of the data matrix.

```r
# Performs the PCA using the correlation matrix
pca.2 <- PCA(Prices2, quali.sup=31, scale.unit = TRUE)

# Plots the PCA Biplot
fviz_pca_biplot(pca.2,
  habillage = Prices2$Season, addEllipses = TRUE,
  col.var = "red",
  label = "var") +
  scale_color_brewer(palette="Dark2")+
  theme_minimal()
```



Here again, we observe that most of stock prices are correlated, with the loading arrows pointing to the right along the first principal component. Yet now we observe that some loading arrows like Goldman Sachs(GS) Verizon (VZ) and Exxon Mobil (XOM) are also pointing upward along the second principal component. This means that the variation of prices in these stock is somewhat correlated (depending on how large the angle is between each upward arrow) and best explained by the Second Principal Component. The seasonal trend of seeing stock prices grow as the year unfolds still holds here.

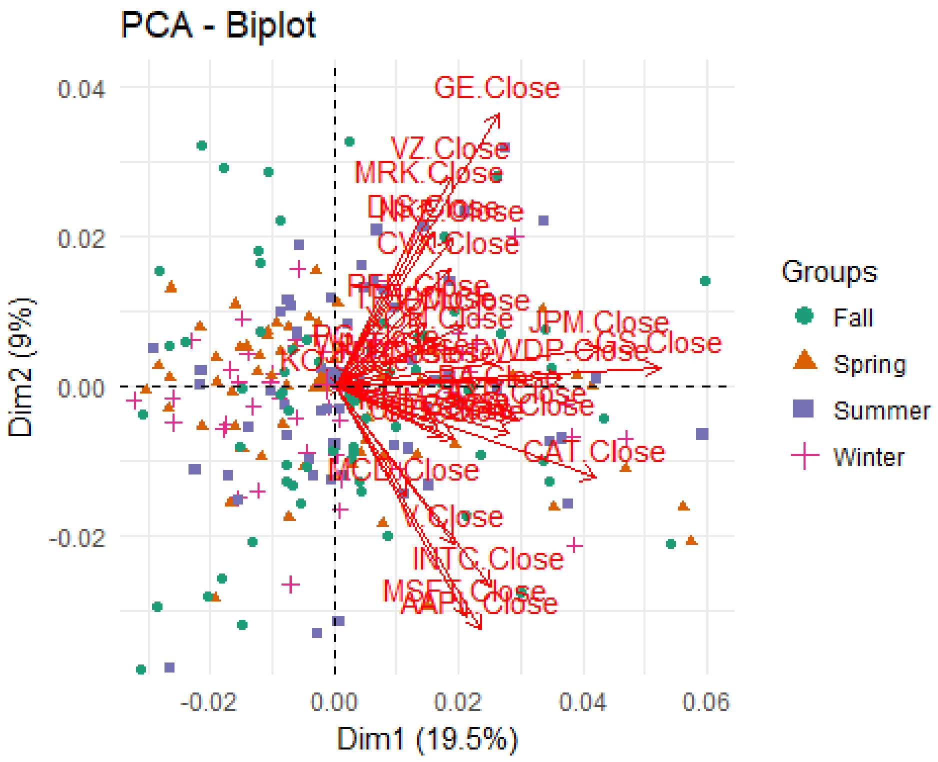
```
#Plots the PCA screeplot
fviz_screeplot(pca.2, ncp=10)
```



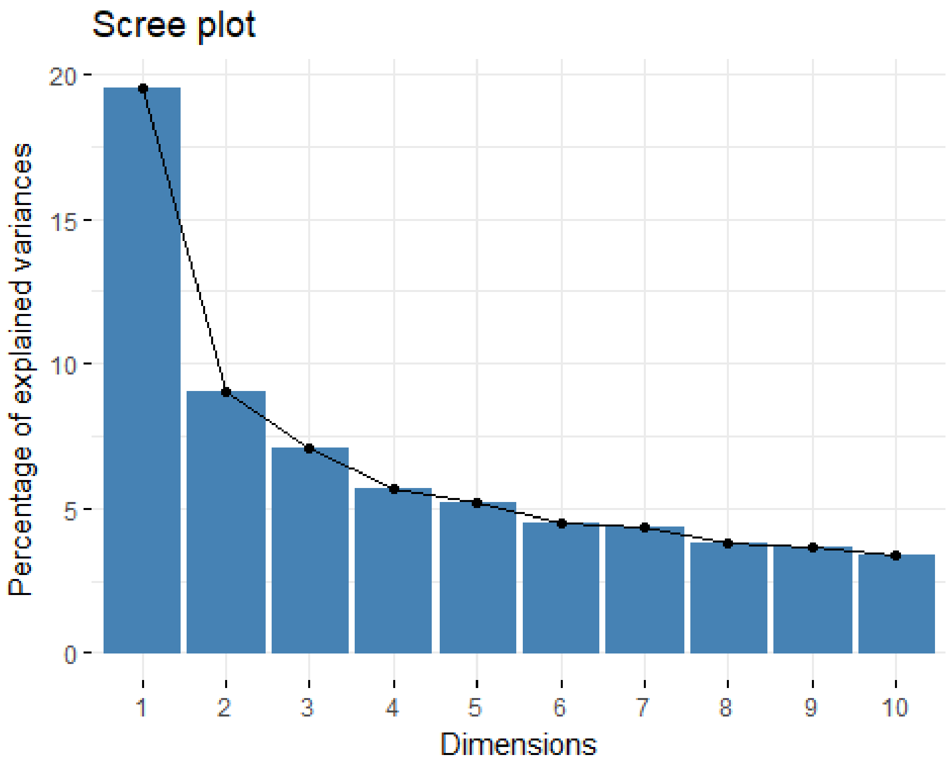
Based on the screeplot, there are two important Principal Components which both together explain 78.4% of the variance in the data. This agrees with the biplot where we described that some loadings like Verizon and Exxon Mobile pointed along the Second Principal Component which thus best explains variations for these stocks.

4. Use the closing prices to calculate the return for each stock, and repeat part 3 on the return data. In looking at the screeplot, what does this tell you about the 30 stocks in the DJIA? If each stock were fluctuating up and down randomly and independent of all the other stocks, what would you expect the screeplot to look like?

```
#Calculates the return matrix
Prices <- data.matrix(Prices)
Returns <- diff(Prices)/Prices[-dim(Prices)[1],]
Returns2 <- data.frame(Returns, Season[1:250])


# Performs the PCA using the covariance matrix
pca.3 <- PCA(Returns2, quali.sup=31, scale.unit = FALSE)


# Plots the PCA Biplot and zooms in to better observe the results
fviz_pca_biplot(pca.3,
  habillage = Returns2$Season,
  col.var = "red",
  label = "var",xlim = c(-0.03, 0.06),ylim = c(-0.035, 0.04)) +
  scale_color_brewer(palette="Dark2")+
  theme_minimal()
```

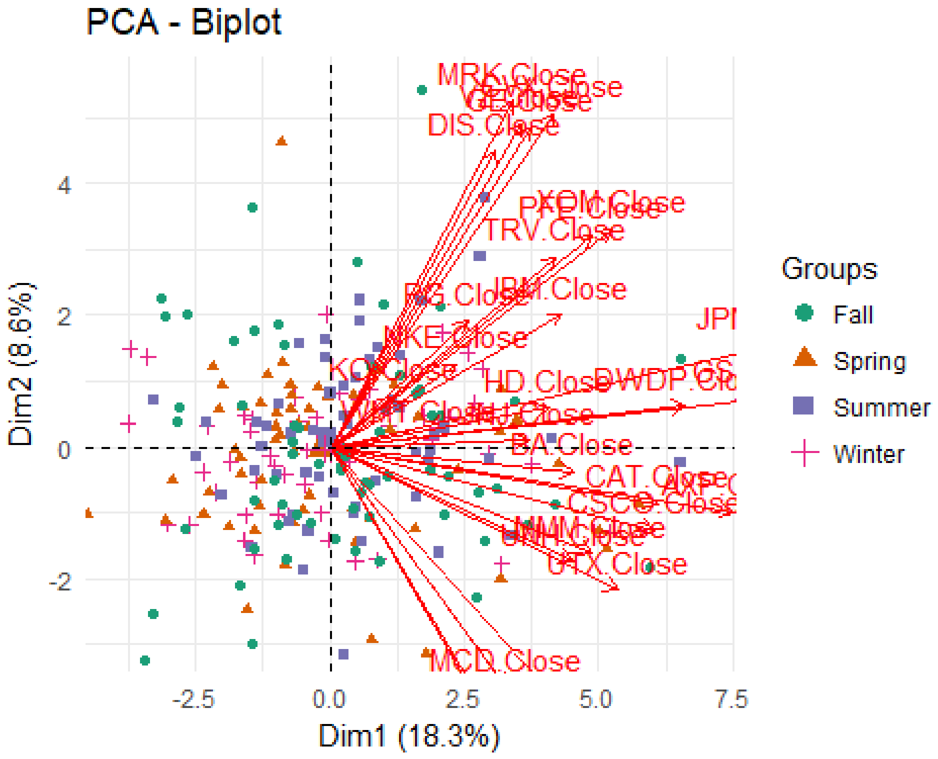


```
#Plots the PCA screeplot
fviz_screeplot(pca.3, ncp=10)
```

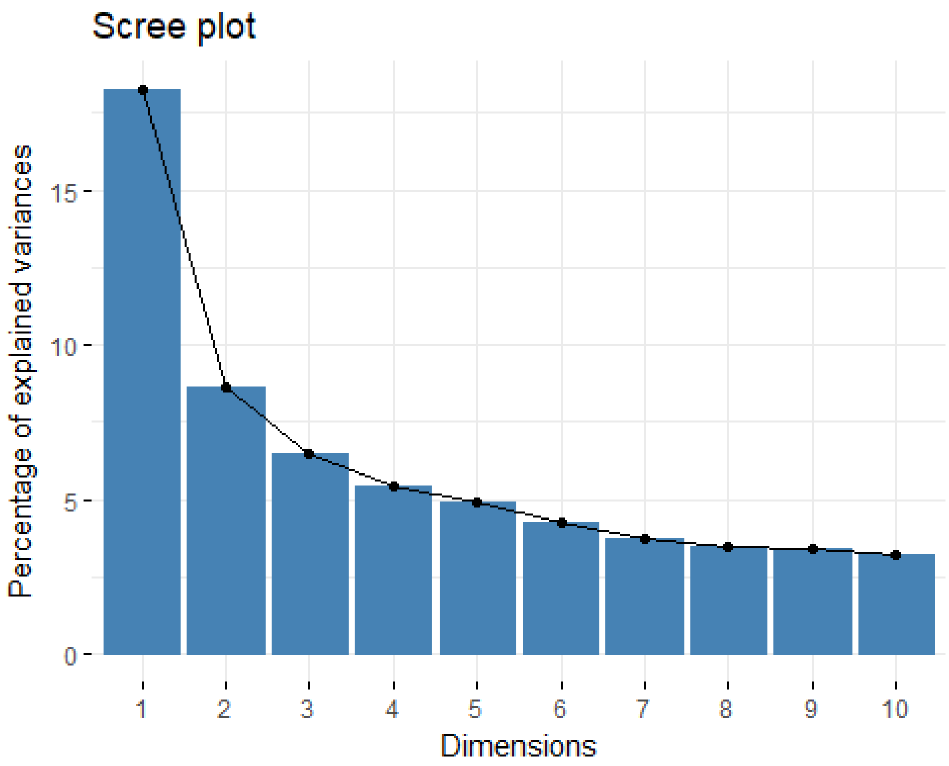


```
(Response to the question is at the end of the problem)
```

```
# Performs the PCA using the correlation matrix
pca.4 <- PCA(Returns2, quali.sup=31, scale.unit = TRUE)



# Plots the PCA Biplot and zooms in to better observe the results
fviz_pca_biplot(pca.4,
  habillage = Returns2$Season,
  col.var = "red",
  label = "var",xlim = c(-4, 7),ylim = c(-3, 5.5)) +
  scale_color_brewer(palette="Dark2")+
  theme_minimal()
```



```r
#Plots the PCA screeplot
fviz_screeplot(pca.4, ncp=10)
```



In both biplots we observe that the returns are centered around 0 and all the loading variables are pointing towards the right along the first principal component. This explains why this component has the largest explanatory power compared to all other principal components as seen on the screeplots.

Indeed, on both screeplots we observe one important principal component which explains 19.5% of the variance (for the covariance matrix) and 18.3% (for the correlation matrix), then the next principal component only explains about 8% of the variance. The remaining principal components slowly flatten out below 5%. This means that the returns of the 30 stocks listed on the Dow Jones move erratically in many different directions. Unlike prices, their variance is not well explained by just one principal direction and thus returns do not all significantly move in similar fashion.

If each stock were fluctuating up and down randomly and independent of all the other stocks, we would expect the screeplot to look flat. There would be no correlation between the stocks and therefore we wouldn't expect to find a main principal component that explains the variations.