Unsupervised learning- New York Stock Exchange

Lorenzo de Sario

17/8/2020

S&P500 companies - Fundamental indicators

Abstract

How the market will be tomorrow morning? Could we reduce the relevant components of a balance sheet? Which companies are more similar if the objective is to build a robust security portfolio? Most of these questions need differents techniques chosen among supervised and unsupervised learning and differents types of data. Indeed, while supervised learning requires to build models in order to get the closest prediction to a given response (e.g. the tomorrow's price of AAPL ticker), unsupervised learnings consider just the inputs

$$X_1, X_2, ..., X_p$$

without taking into account if there exists output. Thus with unsupervised methods is not possible to check for accuracy measures of our results, but is included in the analysis for the purpose of identify possible patterns among the data, such as unknown subgrups and/or as a part of the exploratory data analysis with the scope of reducing dimensions of the n*p dataset before supervised methods are applied.

Datasets and goals

The data of fundamentals metrics of the S&P500 companies are got from Kaggle at the link https://www.kaggle.com/dgawlik/nyse which makes use of the Nasdaq Financials data extended by fields of the annual annual report (Form 10-K) of the U.S. Securities and Echange Commission which provides company's financial performance. The dataset provided looks at the second crisis of the 2000s (2012-2016) and is composed by 1781 observations and 77 continuous variables containing different missing values represented as NA and zeros. Each of these missings are replaced by the mean values for each variable, and for each of the 448 stocks is taken the mean among different years of each measured variable in order to check which stocks reaches similarity during the entire period taken into account. The objective of this analysis is to complete an exploratory data analysis so to prepare the data for further analysis, such as the prediction of the return on equity (ROE) for the year 2017, thus with the aim of generalize the models to future data.

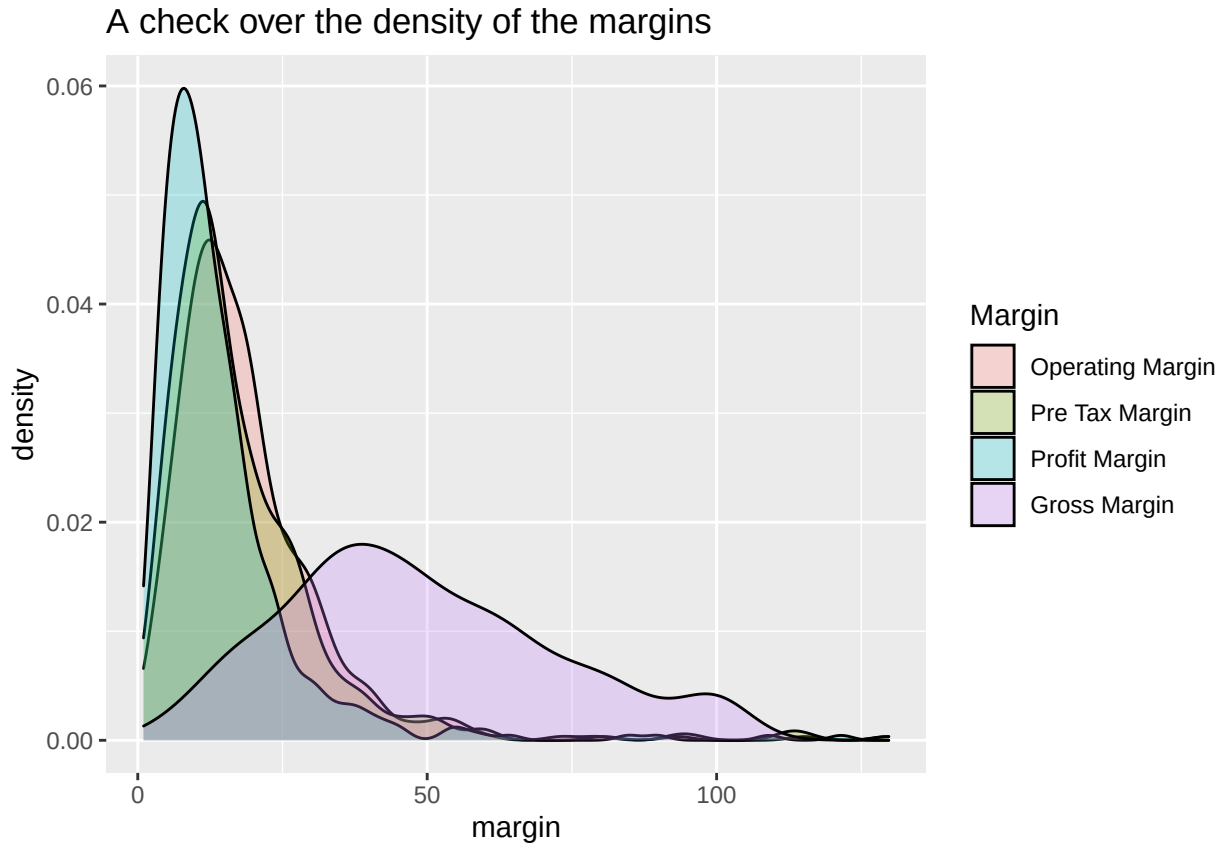
- GOALS
  - Check the density of the margins gained by the companies as a way to assess the expected relation Gross Margin > Operating Margin > Pre Tax Margin > Profit Margin ;
  - Determine the minimum number of principal components that are able to explain the larger cumulative proportion of variance explained by each component (PVE);
  - Evaluate similarity between companies in terms of profittability;
  - Compare Hierarcal clustering with differents measure of distances and different linkage methods and look at patterns when the dendogram is cutted;
  - Compare the k-means clustering total within sum of squares of the deviations for differents random subsets.

```
## [1] 1781   79
```

```
##  Operating.Margin Pre.Tax.Margin   Profit.Margin     Gross.Margin
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  9.00   1st Qu.:  8.00   1st Qu.:  6.00   1st Qu.: 29.00
##  Median : 15.00   Median : 14.00   Median : 10.00   Median : 43.00
##  Mean   : 18.18   Mean   : 17.75   Mean   : 13.96   Mean   : 46.76
##  3rd Qu.: 23.00   3rd Qu.: 22.00   3rd Qu.: 17.00   3rd Qu.: 64.00
##  Max.   :437.00   Max.   :442.00   Max.   :369.00   Max.   :100.00
##
##    Cash.Ratio      Current.Ratio     Quick.Ratio      Pre.Tax.Margin.1
##  Min.   :   0.00   Min.   :  17.0   Min.   :  10.00   Min.   :  0.00
##  1st Qu.:  17.00   1st Qu.: 109.0   1st Qu.:  77.25   1st Qu.:  8.00
##  Median :  41.00   Median : 152.0   Median : 115.00   Median : 14.00
##  Mean   :  74.46   Mean   : 186.8   Mean   : 146.95   Mean   : 17.75
##  3rd Qu.:  90.00   3rd Qu.: 226.0   3rd Qu.: 180.00   3rd Qu.: 22.00
##  Max.   :1041.00   Max.   :1197.0   Max.   :1197.00   Max.   :442.00
##  NA's   :299       NA's   :299      NA's   :299
##  Earnings.Before.Interest.and.Tax  Pre.Tax.ROE
##  Min.   :-2.793e+10               Min.   :   0.00
##  1st Qu.: 5.852e+08               1st Qu.:  13.00
##  Median : 1.139e+09               Median :  22.00
##  Mean   : 2.710e+09               Mean   :  59.65
##  3rd Qu.: 2.586e+09               3rd Qu.:  36.00
##  Max.   : 7.905e+10               Max.   :9089.00
##
```

Including Plots

With the scope of determine the varacity of the data analyzed, are plotted the densities of the different margins gained by the companies. Indeed, the density of the Gross Margin with respect to the Operating Margin possess much more observations for higher margins. This is due to the fact that the gross profit margin includes also the operating expenses while the gross margin consider just the direct costs (thus, Gross Margin is greather than Operating Margin). Moreover, is possible to observe that the Pre Tax Margin and the Pre Tex Margin at most overlaps (on average, taxes aren't an heavy budget item if compared with the operating costs) and retain higher margins with respect to the Profit ones for which all the budget costs are taken into account.



Since the data are in different unit of measures ($ for indicators and percentage points for indexes) the data are normalized that is transformed in such a way that the variable's mean is 0 and its standard deviation (SD) is equal to 1. Moreover this task need to be performed when due to large dimensionality of the dataset, we want to reduce the number of variables with Principal Component Analysis (PCA). In particular, this is done because PCA produces projection of the data on new axis based on the SD of the features. Thus in order to let contribute all the variables with the same weight, thus with the same SD, for the calculation of the axis, a normalization is required.

Principal component analysis

PCA is the unsupervised technique that allows to reduce the dimension of the dataset and let it be rapresented by those variables that explain its most variability along each dimension. The first Principal component, as a linear combination of p features, is given as

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \phi_{31}X_3 + ... + \phi_{p1}X_p$$

and captures the largest variance of the dataset, with p loadings obtained from a maximization problem constrained so as their sum of squares is equal to one (normalization). To solve this maximization problem, an eigen decomposition (singular value decomposition approach) is required. After the first principal component has been calculated, it is possible to determine the second ones as a linear combination of p variables which catch the remaining variance and is uncorrelated with

$$Z_1$$

, thus the two components will be orthogonal. Then the following principal components, up to min(n-1, p), for a n * p dataset, can be calculated in order to capture the remaining variance.

For further development, since the principal component analysis is frequently adopted before a predictive model is engaged, the validation set approach is required , but so to complete the prediction the PCA need to be applied to the training set and to the test set taking into account the same n components that are choosen from the training PCA. For each variable of the performed principal component on the training set is got the mean and the standard deviation and a total of (min(448-1, 76) =) 76 principal components with respective loading vectors are calculated starting from a correlation matrix. Moeover, is obtained the proportion of variance explained by each principal component (PVE) and the cumulative proportion of variance explained.

```
##           Accounts.Payable        Accounts.Receivable
##               4.784022e+09              -7.138991e+07
## Add.l.income.expense.items              After.Tax.ROE
##               8.328501e+07               4.367769e+01
##       Capital.Expenditures            Capital.Surplus
##              -1.297553e+09               6.257641e+09
##                 Cash.Ratio  Cash.and.Cash.Equivalents
##               7.499317e+01               8.480560e+09
##     Changes.in.Inventories              Common.Stocks
##              -1.052851e+08               1.712134e+09
```

```
##           Accounts.Payable        Accounts.Receivable
##               1.395070e+10               4.401707e+08
## Add.l.income.expense.items              After.Tax.ROE
##               4.691315e+08               1.353977e+02
##       Capital.Expenditures            Capital.Surplus
##               2.921929e+09               1.098934e+10
##                 Cash.Ratio  Cash.and.Cash.Equivalents
##               8.612734e+01               5.375028e+10
##     Changes.in.Inventories              Common.Stocks
##               3.087731e+08               9.140769e+09
```

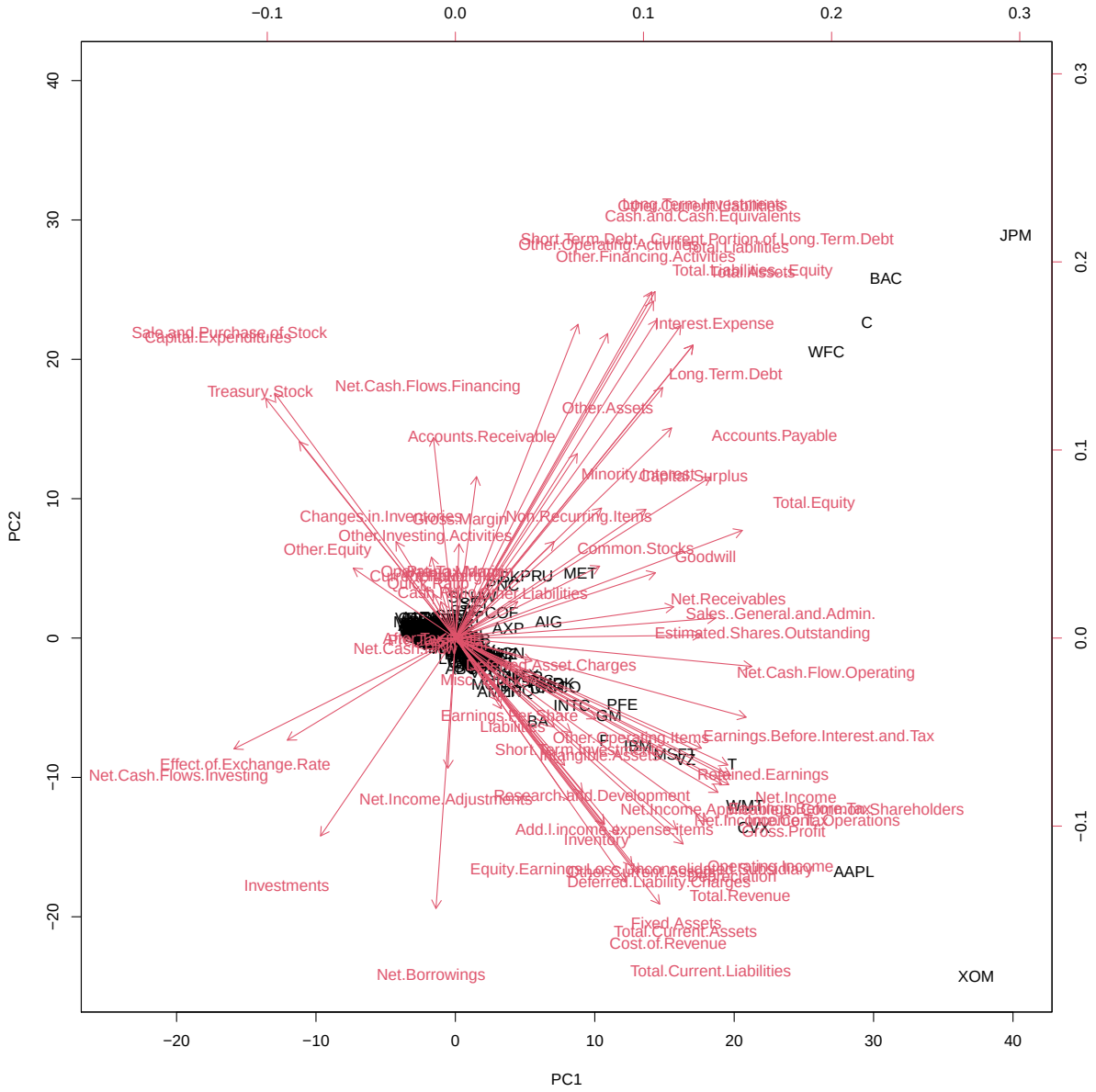
```
##                                    PC1          PC2         PC3          PC4
## Accounts.Payable             0.169680077  0.107017019  0.04321594 -0.04050443
## Accounts.Receivable          0.014068191  0.107253655  0.03874558  0.10092099
## Add.l.income.expense.items   0.084607774 -0.102478528 -0.06285373  0.05669238
## After.Tax.ROE               -0.011813237 -0.000175547 -0.06705175  0.05812154
## Capital.Expenditures        -0.126082548  0.159368048  0.01320481 -0.10214000
## Capital.Surplus              0.126626732  0.085586280  0.01905398  0.09564072
## Cash.Ratio                  -0.009708654  0.024234442 -0.27243205  0.09834206
## Cash.and.Cash.Equivalents    0.131609544  0.223890906 -0.03582208 -0.07837813
## Changes.in.Inventories      -0.039453927  0.064050587 -0.06625036  0.04116537
## Common.Stocks                0.095876634  0.047785956 -0.02408209 -0.04117361
##                                     PC5          PC6         PC7          PC8
## Accounts.Payable             0.009448083 -0.07441460  0.03727731 -0.007003443
## Accounts.Receivable          0.208060612  0.04011109 -0.07524884 -0.166035694
## Add.l.income.expense.items   0.148777330 -0.10259631 -0.22486450  0.025152609
## After.Tax.ROE                0.022500559 -0.03225748  0.03222986  0.190396933
## Capital.Expenditures        -0.149043503 -0.12204334  0.06015622 -0.090906859
## Capital.Surplus             -0.015545754  0.20914214  0.08037110 -0.060342398
## Cash.Ratio                  -0.186995894 -0.07652379 -0.23971524 -0.224738548
## Cash.and.Cash.Equivalents    0.038755438 -0.06051953  0.01698859  0.040073489
## Changes.in.Inventories       0.014956851  0.06608834  0.06583350 -0.063542412
## Common.Stocks               -0.201614833 -0.17130748 -0.01089573  0.220869226
##                                     PC9         PC10
```

```
## Accounts.Payable            0.05715362 -0.13667121
## Accounts.Receivable         0.06842324  0.17754190
## Add.l.income.expense.items -0.05795743  0.03407158
## After.Tax.ROE               0.49464860 -0.01495234
## Capital.Expenditures        0.10272857  0.05387618
## Capital.Surplus            -0.05491357 -0.04691121
## Cash.Ratio                 -0.06925207 -0.15589781
## Cash.and.Cash.Equivalents  -0.02211916 -0.04670180
## Changes.in.Inventories     -0.13246938  0.23433586
## Common.Stocks              -0.14872536  0.17071544
```

```
##                             PC1      PC2      PC3     PC4      PC5      PC6
## Standard deviation     4.861468 3.182743 2.060821 2.01040 1.962771 1.701884
## Proportion of Variance 0.310970 0.133290 0.055880 0.05318 0.050690 0.038110
## Cumulative Proportion  0.310970 0.444260 0.500140 0.55332 0.604010 0.642120
##                             PC7      PC8      PC9     PC10     PC11    PC12
## Standard deviation     1.605608 1.573531 1.468153 1.407125 1.322709 1.24199
## Proportion of Variance 0.033920 0.032580 0.028360 0.026050 0.023020 0.02030
## Cumulative Proportion  0.676040 0.708620 0.736980 0.763040 0.786060 0.80635
##                            PC13    PC14     PC15     PC16     PC17      PC18
## Standard deviation     1.129638 1.10929 1.082278 1.044779 1.020782 0.9375793
## Proportion of Variance 0.016790 0.01619 0.015410 0.014360 0.013710 0.0115700
## Cumulative Proportion  0.823140 0.83933 0.854750 0.869110 0.882820 0.8943900
##                             PC19     PC20      PC21      PC22      PC23
## Standard deviation     0.8543903 0.835861 0.7928738 0.7739773 0.7309773
## Proportion of Variance 0.0096100 0.009190 0.0082700 0.0078800 0.0070300
## Cumulative Proportion  0.9039900 0.913180 0.9214600 0.9293400 0.9363700
##                             PC24      PC25      PC26      PC27      PC28
## Standard deviation     0.6833496 0.6312471 0.6102934 0.5790373 0.5670808
## Proportion of Variance 0.0061400 0.0052400 0.0049000 0.0044100 0.0042300
## Cumulative Proportion  0.9425100 0.9477600 0.9526600 0.9570700 0.9613000
##                             PC29      PC30      PC31      PC32      PC33
## Standard deviation     0.5207918 0.5137194 0.4972029 0.4616803 0.4254097
## Proportion of Variance 0.0035700 0.0034700 0.0032500 0.0028000 0.0023800
## Cumulative Proportion  0.9648700 0.9683400 0.9715900 0.9744000 0.9767800
##                             PC34      PC35      PC36      PC37      PC38
## Standard deviation     0.3999409 0.3923064 0.3850594 0.3603823 0.3536607
## Proportion of Variance 0.0021000 0.0020300 0.0019500 0.0017100 0.0016500
## Cumulative Proportion  0.9788800 0.9809100 0.9828600 0.9845700 0.9862200
##                            PC39      PC40      PC41      PC42      PC43
## Standard deviation     0.335353 0.3232086 0.2988041 0.2949866 0.2863298
## Proportion of Variance 0.001480 0.0013700 0.0011700 0.0011400 0.0010800
## Cumulative Proportion  0.987700 0.9890700 0.9902400 0.9913900 0.9924700
##                             PC44      PC45     PC46      PC47      PC48
## Standard deviation     0.2505948 0.2473542 0.234358 0.2193744 0.2091335
## Proportion of Variance 0.0008300 0.0008100 0.000720 0.0006300 0.0005800
## Cumulative Proportion  0.9932900 0.9941000 0.994820 0.9954600 0.9960300
##                             PC49     PC50      PC51     PC52      PC53
## Standard deviation     0.1966706 0.183291 0.1769282 0.170718 0.1641468
## Proportion of Variance 0.0005100 0.000440 0.0004100 0.000380 0.0003500
## Cumulative Proportion  0.9965400 0.996980 0.9973900 0.997780 0.9981300
##                             PC54      PC55      PC56      PC57      PC58
## Standard deviation     0.1532733 0.1409205 0.1228166 0.1098343 0.1015432
## Proportion of Variance 0.0003100 0.0002600 0.0002000 0.0001600 0.0001400
```

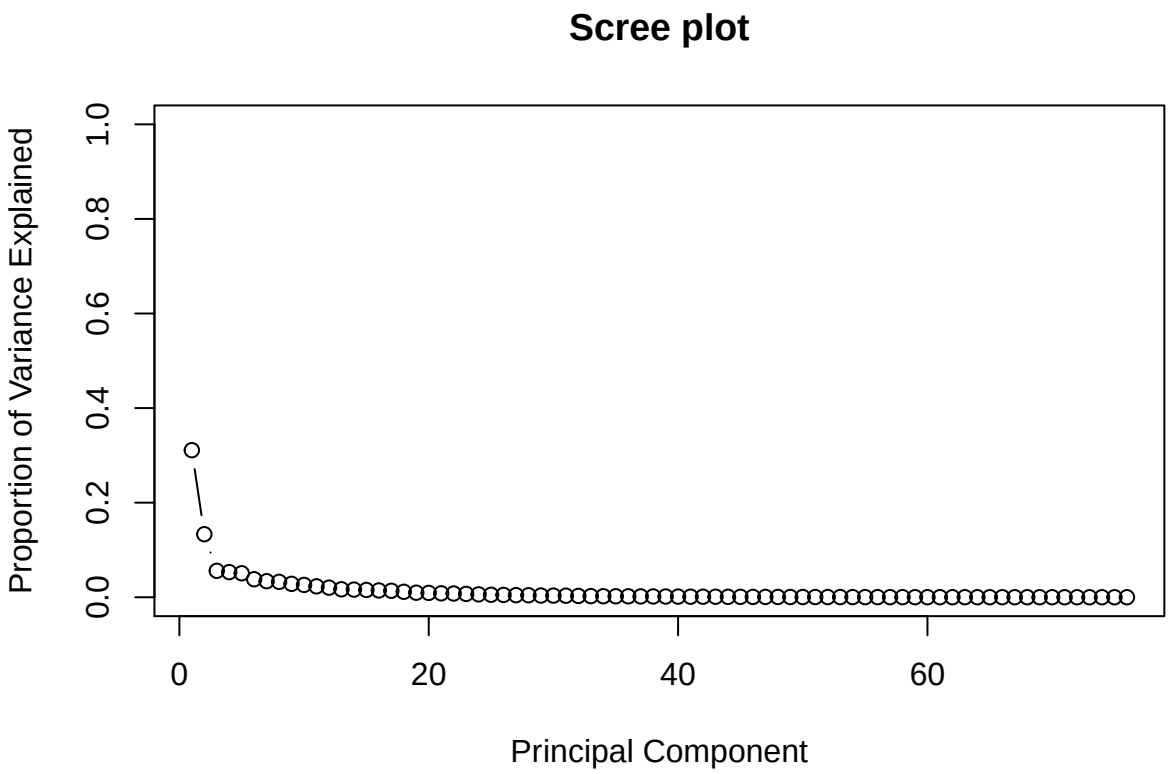
```
## Cumulative Proportion  0.9984400 0.9987000 0.9989000 0.9990600 0.9991900
##                              PC59       PC60       PC61       PC62       PC63
## Standard deviation     0.09943385 0.09368466 0.08868384 0.07951173 0.07234381
## Proportion of Variance 0.00013000 0.00012000 0.00010000 0.00008000 0.00007000
## Cumulative Proportion  0.99933000 0.99944000 0.99954000 0.99963000 0.99970000
##                              PC64       PC65       PC66       PC67       PC68
## Standard deviation     0.07009989 0.06381487 0.05878796 0.05728737 0.05182251
## Proportion of Variance 0.00006000 0.00005000 0.00005000 0.00004000 0.00004000
## Cumulative Proportion  0.99976000 0.99981000 0.99986000 0.99990000 0.99994000
##                              PC69      PC70       PC71       PC72        PC73
## Standard deviation     0.04621745 0.0403308 0.02622842 0.01286618 0.006410623
## Proportion of Variance 0.00003000 0.0000200 0.00001000 0.00000000 0.000000000
## Cumulative Proportion  0.99997000 0.9999900 1.00000000 1.00000000 1.000000000
##                               PC74         PC75         PC76
## Standard deviation     0.005324755 0.0005739672 4.453682e-16
## Proportion of Variance 0.000000000 0.0000000000 0.000000e+00
## Cumulative Proportion  1.000000000 1.0000000000 1.000000e+00
```

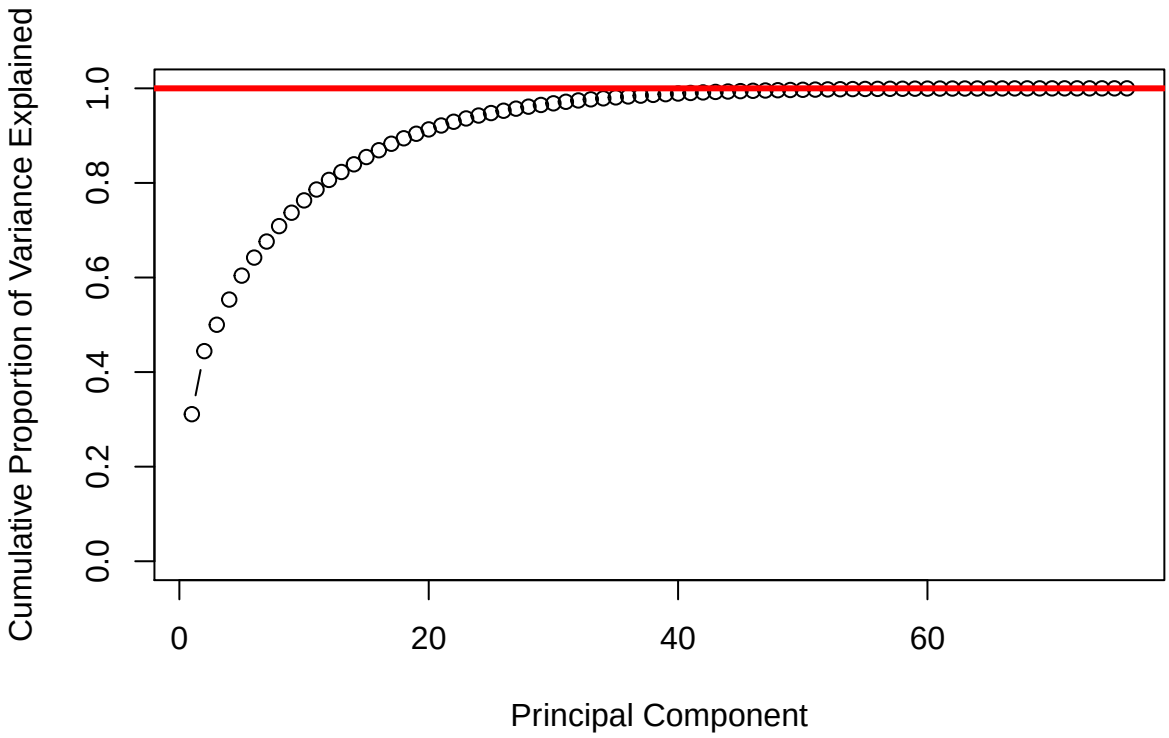
By plotting the calculated principal components and the loading vectors in the two dimensional space, thus PC1 on the horizontal axis and PC2 on vertical, is possible to obtain the so called biplot. This plot shows how variables possess weight over the principal components (e.g Estimated.Shares.Outstanding and Sales.General.and.Administration positively correlated influence PC1, while Net.Borrowing and Investments influences PC2). Moreover is possible to notice that variables such as Net.Borrowind and Estimated.Shares.Outstanding are not correlated while Treasury.Stock and Total.Current.Liabilities presents negative correlation. So as to find the principal components that explain togheter the gratest amount of variance of the dataset, thus to capture as much informations from the data, is computed first the variance explained by each principal component by squaring the standard deviations and then those results are divided by the total variance explained by all the principal components in order to get proportion of variance explained by each component (PVE) (e.g. the first principal component explain the 31.5% of the variance, the second ones the 13.6%). Thus, by plotting the PVE explained by the components against the number of principal components is obtained a scree plot, which tell us that $\sim$ 34 components are able to explains the 98.3% of the variance in the dataset. The results are confirmed by plotting the cumulative proportion of variance explained against the number of principal components. Thus for a possible prediction in terms of supervised methods the PCA should be executed also on the test set with the first 34 components.




```
##  [1] 23.634 10.130  4.247  4.042  3.852  2.896  2.578  2.476  2.155  1.980
## [11]  1.750  1.543  1.276  1.231  1.171  1.092  1.042  0.879  0.730  0.699
```

```
##  [1] 0.31097 0.13329 0.05588 0.05318 0.05069 0.03811 0.03392 0.03258 0.02836
## [10] 0.02605 0.02302 0.02030 0.01679 0.01619 0.01541 0.01436 0.01371 0.01157
## [19] 0.00961 0.00919
```


Scree plot
Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
Principal Component



Clustering Clustering is the unsupervised technique adopted with the aim of discover clusters, that are partitions of the dataset into distintic groups whithin each observation appear to be similar. The main two clustering methods that are applied are the hierarchical clustering and the K-means clustering. This two methods differes for the fact that the K-means clustering requires to establish the needed number of clusters K to which assign each observations while is not required for the hierarchical procedure for which a dendogram shows how many clusters has been generated for each possible number of clusters, from 1 to n. In details, the k-means clustering generates non-nested clusters for which each observation is assigned to any cluster by minimizing the within-cluster variation (closest centroid). The within-cluster variation could be computed with measures of distances (e.g. Minkowsky distance

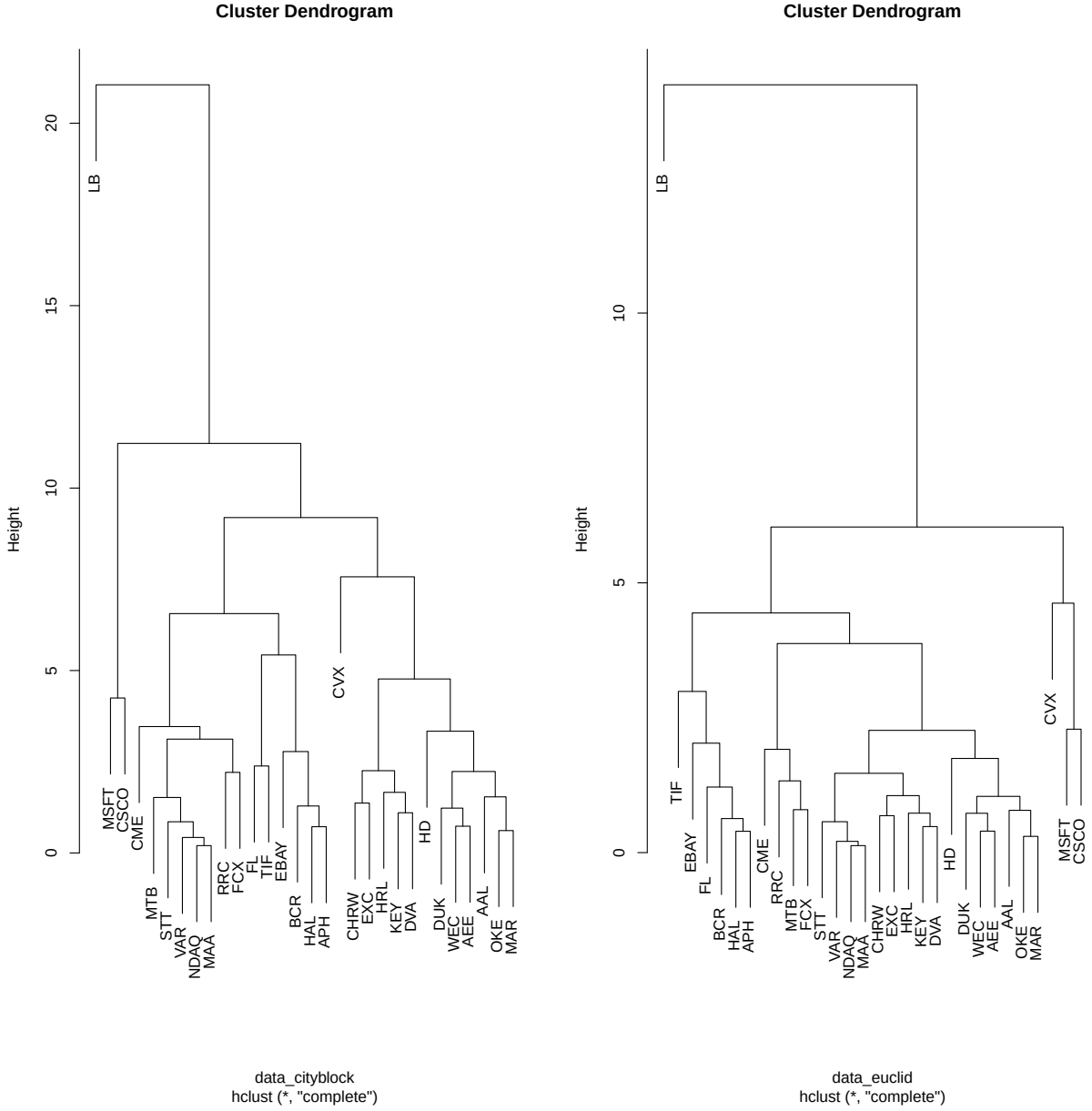
$$(\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$$

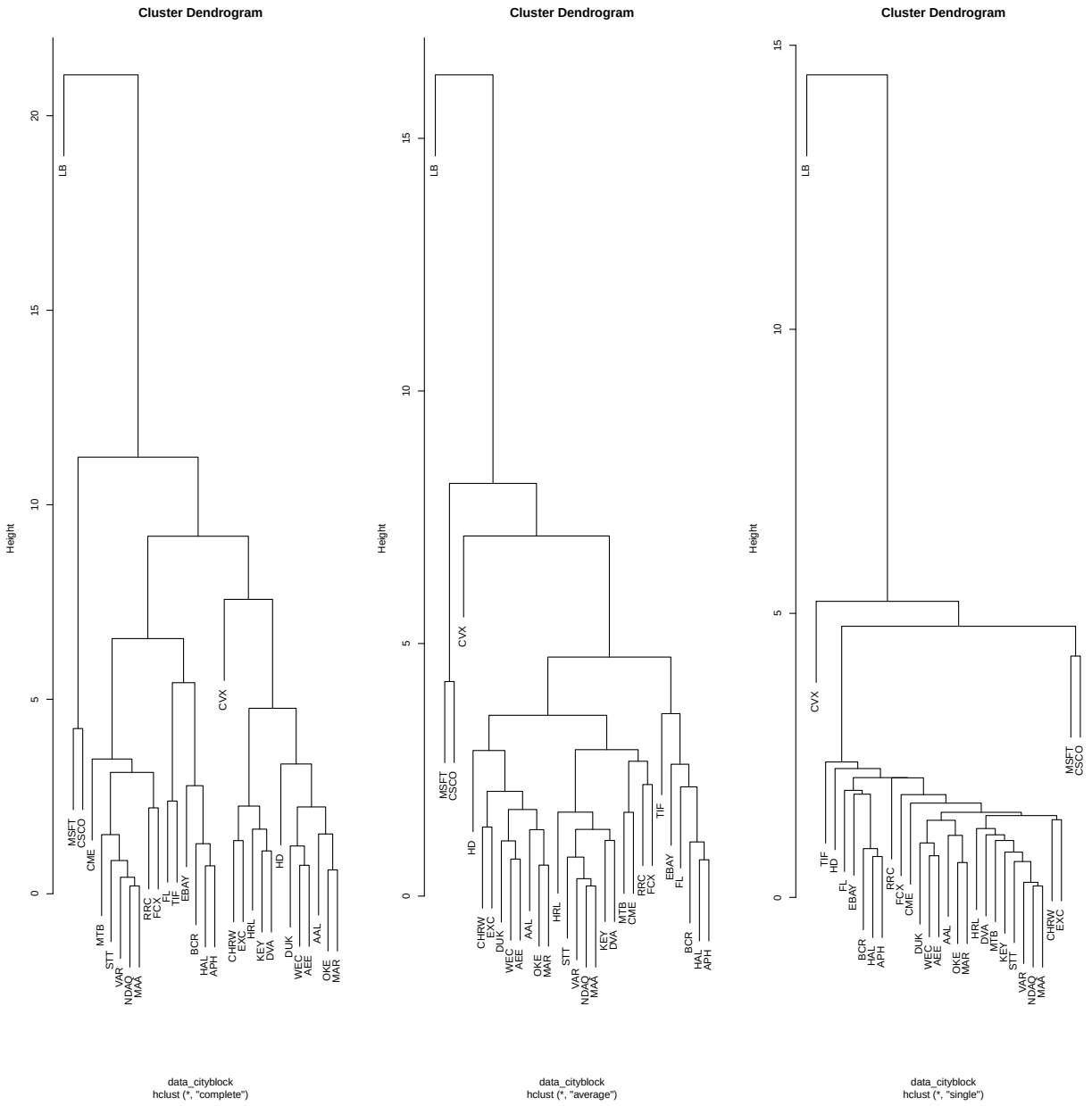
and its applications for k = 2: Euclidean Distance, and k = 1 : City Block) between data points for numerical variables, while is needed a similiraty index (e.g. Jaccard, Russel Rao for qualitative variables) for qualitative predictors.

For the hierarchical clustering, insted, each observation is treated as its own cluster and are calculated all the pairwise between-cluster dissimilarities among each i cluster. By choosing the least dissimilar pair of clusters and joining them is possible to compute the i-1 remaining dissimilarities between clusters. Note that the dissimilarity between each joined clusters represent the height in the dendogram and that the dissimilarity between two clusters for which one of them, or both, contains multiple observations, is evaluated with the linkage methods. Clusters are generated from a sample of 30 stocks for demonstrative purposes on the basis of the variables that refers to profittability and to liquidity of the different companies. Moreover, for the purpose of this analysis, are compared the results between the complete linkage (maximum intercluster dissimilarity- largest dissimilarity between observations in the two clusters)and average linkage (average intercluster dissimilarity- average dissimilarity between observations in the two clusters) which provides more balanced dendograms; single linkage (minimum intercluster dissimilarity- smallest dissimilarity between

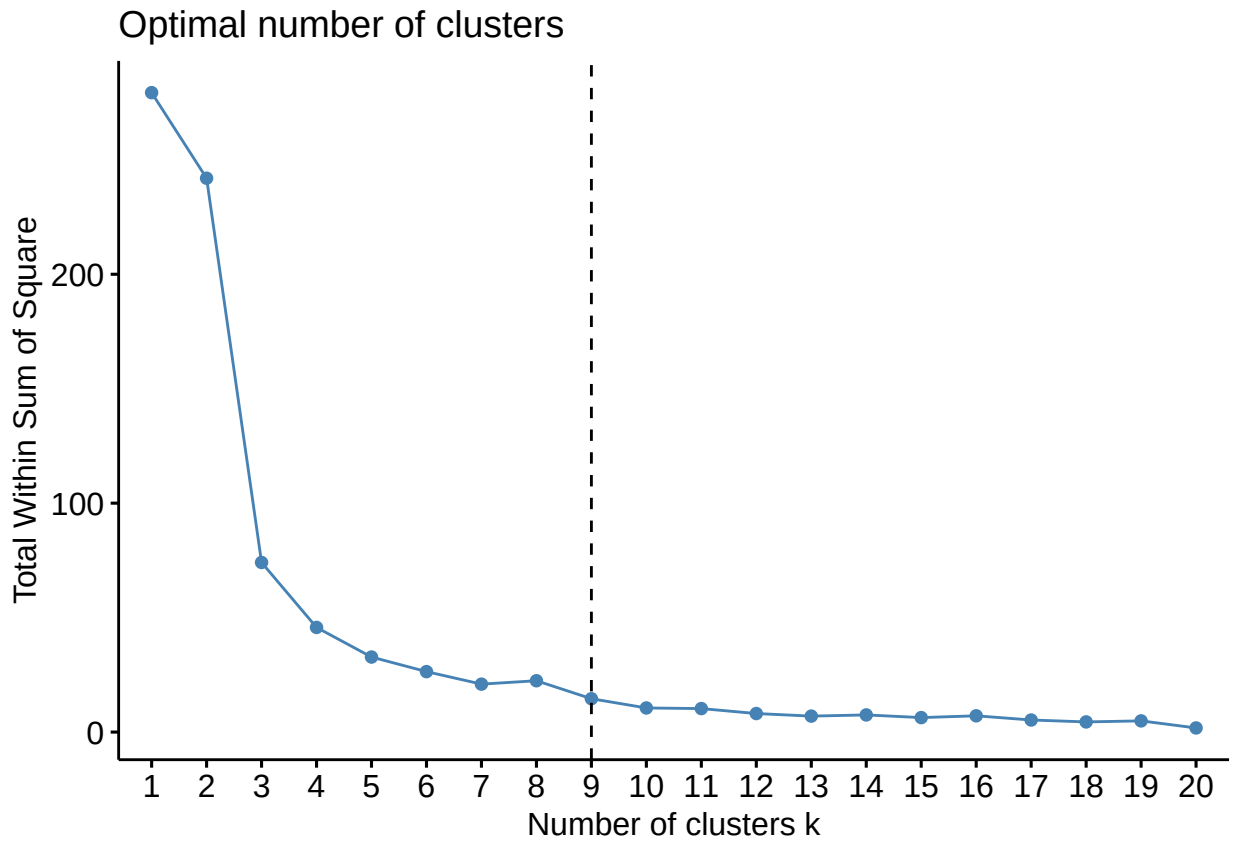
observations in the two clusters). Thus, based on the profittability and liquidity variables, is possible to affirm that for different measure of distances (e.g. city block against Euclidean distance), the hierarchical clustering with complete linkage, shows that firms such as Microsoft (MSFT) and CISCO (CSCO) or also Amphenol (APH), an american producer of electronic and optic cables, and Halliburton (HAL), a oil company, presents similarity in terms of profittability and liquidity. On the numbers of clusters could be said that, with the city block distances, at an height of 5 there exists 6 clusters, while the euclidean distance claims for 2 clusters at the same height. By cutting the tree in order to obtain 6 clusters CSCO and MSFT will not belong to the same group, while APH and HAL will.

On the same dataset is applied the k-mean clustering for which by adopting the minimization of the total within sum of square methods (sum of the squared deviations for each observation from the cluster centroid) are arbitrary choosen 10 clusters. Moreover, taking k = 3, the within sum of square is minimized when 50 random sets are chosen.






```
##  KEY   HD  CVX   LB  DUK  DVA  AAL   FL  STT  OKE CHRW  MAR  BCR  WEC  VAR  TIF
##    1    1    2    3    1    1    1    4    1    1    1    1    4    1    1    4
##  EXC NDAQ  HRL  HAL  AEE EBAY MSFT  MTB  RRC  APH  CME CSCO  FCX  MAA
##    1    1    1    4    1    4    5    1    1    4    1    6    1    1
```




```
## [1] 9.93
```

```
## [1] 14
```

Take aways-Conclusions.

As a result from the density plots, the relation between the margin is verified, thus is possible to affirm the veracity of the data. Moreover, by performing a PCA on the training datasets it result that out of 76 components, 34 are enough to explain the most of the comulative proportion of the variance explained. By the hierarchical clustering analysis notice that companies such as Microsoft (MSFT) and CISCO (CSCO) or also Amphenol (APH) are similar in terms of profittability with the city block distance either with the Euclidean distance. So as to state that the complete linkage method is the more balanced with respect to the single and average linkage notice that at an height of 5 still exists 6 clusters thus avoiding forcing observations to be closed each others. WIth respect to the k- mean clustering, an higher values of random sets will produce smallers total Within sum of squares. After this exploratory data analysis is now possible to make predictions with any supervised learning methods and evantually try to generalize this analysis to more recent data so to build a strong portfolio for long-term investments.

Appendix

```
stock = read.csv("C:/Users/Lorenzo de Sario/Desktop/unsupervised/fundamentals.csv",
                 header = TRUE, sep = ",")
```

```
dim(stock)
columns = c('Operating.Margin','Pre.Tax.Margin','Profit.Margin',
            'Gross.Margin', 'Cash.Ratio', 'Current.Ratio',
            'Quick.Ratio', 'Pre.Tax.Margin',
            'Earnings.Before.Interest.and.Tax', 'Pre.Tax.ROE')
summary(stock[columns])

#replacing missing values with the mean for each column
stock[stock==0] = NA
for(i in 1:ncol(stock)){
  stock[is.na(stock[,i]), i] <- mean(stock[,i], na.rm = TRUE)
}

#group by Ticker symbol and take mean, removing inuseful columns
stock = aggregate(stock[,0:ncol(stock)], list(stock$Ticker.Symbol), mean )
stock = stock[, -which(names(stock)
                       %in% c('Period.Ending', 'X', 'Ticker.Symbol','Year'))]
names(stock)[1] = 'Ticker'

#Exploratory data analysis
library(ggplot2)
library(reshape2)
Margin = data.frame(stock$Operating.Margin,
                    stock$Pre.Tax.Margin,
                    stock$Profit.Margin,
                    stock$Gross.Margin)

Margin = melt(Margin)
ggplot(Margin,aes(x=value, fill=variable)) +
  geom_density(alpha=0.25)+
  labs(title = 'A check over the density of the margins', x='margin')+
  scale_fill_discrete(name = 'Margin', labels = c('Operating Margin', 'Pre Tax Margin',
                                                  'Profit Margin', 'Gross Margin'))
#margins differs according to the rule

#assign ticker symbol to rowname
row.names(stock) = (stock$Ticker)
stock<-stock[,-1]

##PRINCIPAL COMPONENT ANALYSIS
#split in train and test since PCA is useful in the
#exploratory data analysis and thus it could be followed by predictive models
stock_train = stock[1:nrow(stock),]
stock_test = stock[-(1:nrow(stock)),]
#Since there are indexes and economic values($):
#Normalize the data within 0 mean and 1 sd
principal <- prcomp(stock_train, scale = TRUE)
principal$center[1:10] #mean 0 of the variables
principal$scale[1:10] #standard deviation 1 of the variables
principal$rotation[1:10, 1:10]

b = summary(principal)
b$importance#in order to get a cumulative portion of variance explained
```

```r
#by the components of 95%  consider the first 34 components.

#switch off scientific notation
options(scipen=999, digits = 3)

## biplot
biplot(principal, scale = 0) #scale = 0 ensure arrows are scaled so as to represent loadings
#calculating the percentage of variance explained
pr.var=principal$sdev^2
pr.var[1:20]
pve=pr.var/sum(pr.var)
pve[1:20]
#scree plot
plot(pve, main = 'Scree plot', xlab="Principal Component",
     ylab="Proportion of Variance Explained", ylim=c(0,1),type='b')
#cumulative scree plot
plot(cumsum(pve), xlab="Principal Component",
     ylab="Cumulative Proportion of Variance Explained", ylim=c(0,1),type='b')
abline(h=1, lwd=3, col="red")

#CLUSTERING
#hierarchical clustering
set.seed(346)
#sampling 10 stock from standardized data and selecting measures of profittability
sample_stock = data[sample(nrow(stock_train), 30), ]
sample_stock = subset(sample_stock, select = c('Cash.Ratio', 'Current.Ratio', 'Quick.Ratio',
  'Pre.Tax.Margin', 'Earnings.Before.Interest.and.Tax', 'Pre.Tax.ROE'))

#cityblock vs euclidean distance- complete linkage
data_cityblock = dist(sample_stock, upper = TRUE, method="manhattan")
data_euclid = dist(sample_stock, upper = TRUE, method = 'euclidean')
par(mfrow = c(1, 2))
hc_b_c <- hclust(data_cityblock, method = 'complete')
he_d <- hclust(data_euclid)
plot(hc_b_c)
plot(he_d)

#cityblock distance with average, single and complete dissimilarity
#citiblock comparison since is the more robust measure of distance
par(mfrow = c(1, 3))
hc_b_a <- hclust(data_cityblock, method = 'average')
hc_b_s <- hclust(data_cityblock, method = 'single')
plot(hc_b_c)
plot(hc_b_a)
plot(hc_b_s)
cutree(hc_b_a, 6)

#number of clusters for k-mean
fviz_nbclust(sample_stock, kmeans, method = "wss") +
  geom_vline(xintercept = 7, linetype = 2)+
  labs(subtitle = "Elbow method")
km.out = kmeans(sample_stock, 10, nstart = 50)
km.out$tot.withinss
```

```r
km.out1 = kmeans(sample_stock, 10, nstart = 1)
km.out1$tot.withinss
```