project 5

Eya MAHDHAOUI , Manon CASTANET, Aline RIVERA, Pedro José TRILLOS

2022-11-14

# Framework :

In this project, an economic and statistical factor analysis will be carried out in order to determine whether or not there are any unexplained factors that may contribute to the co-variation among the stocks.

We have chosen ten stocks from various industries, including the energy sector(ExxonMobil (NYSE:XOM) and Chevron (NYSE:CVX)), the healthcare sector(UnitedHealth Group (NYSE:UNH) and Johnson & Johnson (NYSE:JNJ)), the financial sector(Warren Buffett’s Berkshire Hathaway (NYSE:BRK-A) and JPMorgan Chase (NYSE:JPM)), the information technology sector(Apple (NASDAQ:AAPL) and Microsoft (NASDAQ:MSFT)), and the real estate sector(American Tower (NYSE:AMT) and major shopping mall owner and operator Simon Property Group (NYSE:SPG)).

To conduct the economic factor analysis, we will consider macroeconomic factors such as the monthly inflation rate and unemployment rate in the United States from 2014/01 to 2022/10 . And,for the statistical factor analysis we will use the daily close prices of the selected stocks from 2014-01-01 to 2022-10-31 retrieved from YahooFinance.

Factor Analysis is a statistical method basically used to reduce the number of variables among a particular data and present it with a lower number of factors.

## Import packages:

library(tidyquant)

## Le chargement a nécessité le package : lubridate

##   
## Attachement du package : 'lubridate'

## Les objets suivants sont masqués depuis 'package:base':  
##   
## date, intersect, setdiff, union

## Le chargement a nécessité le package : PerformanceAnalytics

## Le chargement a nécessité le package : xts

## Le chargement a nécessité le package : zoo

##   
## Attachement du package : 'zoo'

## Les objets suivants sont masqués depuis 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attachement du package : 'PerformanceAnalytics'

## L'objet suivant est masqué depuis 'package:graphics':  
##   
## legend

## Le chargement a nécessité le package : quantmod

## Le chargement a nécessité le package : TTR

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

library(ggplot2)

## Warning: le package 'ggplot2' a été compilé avec la version R 4.2.2

library(reshape2)

## Warning: le package 'reshape2' a été compilé avec la version R 4.2.2

library(viridis)

## Warning: le package 'viridis' a été compilé avec la version R 4.2.2

## Le chargement a nécessité le package : viridisLite

library(psych)

##   
## Attachement du package : 'psych'

## Les objets suivants sont masqués depuis 'package:ggplot2':  
##   
## %+%, alpha

library (GPArotation)  
library(readxl)

## Warning: le package 'readxl' a été compilé avec la version R 4.2.2

library(xts)

options("getSymbols.warning4.0"=TRUE)  
options("getSymbols.yahoo.warning"=TRUE)  
tickers = c("XOM","CVX","UNH","JNJ","BRK-A","JPM","AAPL","MSFT","AMT","SPG")  
close\_prices = NULL  
for (stock in tickers) {  
 close\_prices <- cbind(close\_prices, getSymbols(stock,from = '2014-01-01', to = '2022-10-31', warnings = FALSE, auto.assign = F)[,4])  
}

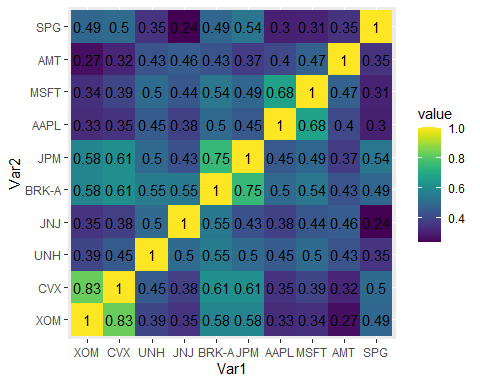
portfolio\_ret = diff(log(close\_prices))[-1,]  
df\_portfolio\_ret = as.data.frame(portfolio\_ret)  
colnames(df\_portfolio\_ret) = tickers  
head(df\_portfolio\_ret,4)

## XOM CVX UNH JNJ BRK-A  
## 2014-01-03 -0.002408894 0.001690201 0.007082251 0.008967678 9.073998e-05  
## 2014-01-06 0.001506271 -0.002657335 -0.011517471 0.005212348 -1.046652e-02  
## 2014-01-07 0.014048913 0.008430762 0.030118416 0.021006013 -1.749380e-03  
## 2014-01-08 -0.003270426 -0.014334269 -0.011700639 -0.001379645 -5.243494e-03  
## JPM AAPL MSFT AMT SPG  
## 2014-01-03 0.007700919 -0.022210665 -0.006750397 0.002890758 0.0121553624  
## 2014-01-06 0.005779380 0.005438193 -0.021358944 0.006380230 0.0067327238  
## 2014-01-07 -0.011592356 -0.007177270 0.007719889 0.013747210 -0.0007099017  
## 2014-01-08 0.009386518 0.006312851 -0.018013568 0.007841243 -0.0034279214

## Correlation between the stocks:

After calculating the differentiated log return of each stock, we will see and measure the correlation between them.

correlation = cor(df\_portfolio\_ret)  
correlation\_melt <- melt(correlation)  
   
ggplot(correlation\_melt, aes(x = Var1,  
 y = Var2,  
 fill = value))+ geom\_tile()+scale\_fill\_viridis(discrete = FALSE) + geom\_text(aes(label = round(value, 2)))



As shown in the heatmap above, where dark,light green and yellow indicate that the variables are highly correlated, we can see that there is a substantial correlation between the majority of the stocks. We can also see that the majority of stock pairs from the same sector are highly correlated, however SPG and AMT are low correlated with a value of 0.35.

# Economic Factor Analysis:

To conduct the economic factor analysis we will use monthly inflation and unemployment rates in the united states.

inflation = data.frame(read\_excel('inflation\_monthly\_us.xlsx'))  
unemployment = data.frame(read\_excel('unemployment\_monthly\_us.xlsx'))

factors = cbind(inflation$Rate,unemployment$rate)  
colnames(factors) = c('inflation\_rate', 'unemployment\_rate')  
head(factors,5)

## inflation\_rate unemployment\_rate  
## [1,] 0.0158 0.066  
## [2,] 0.0113 0.067  
## [3,] 0.0151 0.067  
## [4,] 0.0195 0.062  
## [5,] 0.0213 0.063

We need to change the returns of the stocks from daily to monthly values.

monthly\_log\_return<- to.monthly(portfolio\_ret, indexAt = "last", OHLC = FALSE)

monthly\_log\_return = monthly\_log\_return[-107,]

## Building the model:

We will apply an auto Regression model (AR) to extract the residuals that represent the unexpected shocks related to the inflation and unemployment rates.

ar.mod <- ar(factors)  
res = ar.mod$resid[3:106,]

monthly\_log\_return = monthly\_log\_return[3:106,]

After ensuring that all variables have the same length, we can apply the linear regression model to each of the stock’s returns, defining each macroeconomic variable’s residuals as independent variables.

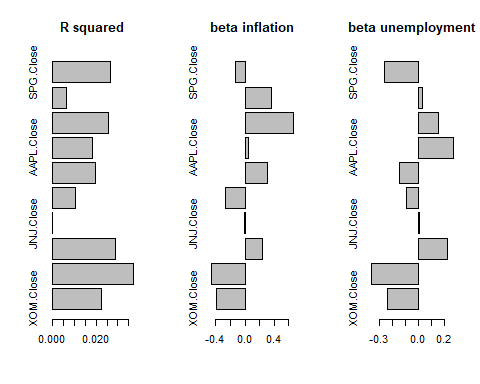
lmfit = lm(monthly\_log\_return ~res[,1]+res[,2])  
summary(lmfit)

## Response XOM.Close :  
##   
## Call:  
## lm(formula = XOM.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## XOM.Close  
## Min -0.048709  
## 1Q -0.008206  
## Median 0.001717  
## 3Q 0.009892  
## Max 0.047673  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.003967 0.001601 -2.477 0.0149 \*  
## res[, 1] -0.390352 0.479252 -0.815 0.4173   
## res[, 2] -0.241683 0.164659 -1.468 0.1453   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01619 on 101 degrees of freedom  
## Multiple R-squared: 0.02247, Adjusted R-squared: 0.003112   
## F-statistic: 1.161 on 2 and 101 DF, p-value: 0.3174  
##   
##   
## Response CVX.Close :  
##   
## Call:  
## lm(formula = CVX.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## CVX.Close  
## Min -0.054208  
## 1Q -0.008935  
## Median 0.001106  
## 3Q 0.009844  
## Max 0.085988  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.002573 0.001788 -1.439 0.1534   
## res[, 1] -0.453683 0.535162 -0.848 0.3986   
## res[, 2] -0.358800 0.183868 -1.951 0.0538 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01807 on 101 degrees of freedom  
## Multiple R-squared: 0.03715, Adjusted R-squared: 0.01808   
## F-statistic: 1.948 on 2 and 101 DF, p-value: 0.1478  
##   
##   
## Response UNH.Close :  
##   
## Call:  
## lm(formula = UNH.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## UNH.Close  
## Min -0.0283202  
## 1Q -0.0067703  
## Median 0.0004102  
## 3Q 0.0075559  
## Max 0.0352125  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.002529 0.001285 -1.968 0.0518 .  
## res[, 1] 0.233741 0.384587 0.608 0.5447   
## res[, 2] 0.228844 0.132135 1.732 0.0863 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01299 on 101 degrees of freedom  
## Multiple R-squared: 0.02896, Adjusted R-squared: 0.009733   
## F-statistic: 1.506 on 2 and 101 DF, p-value: 0.2267  
##   
##   
## Response JNJ.Close :  
##   
## Call:  
## lm(formula = JNJ.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## JNJ.Close  
## Min -0.033677  
## 1Q -0.006013  
## Median 0.000578  
## 3Q 0.007522  
## Max 0.023638  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0026242 0.0010458 -2.509 0.0137 \*  
## res[, 1] -0.0009813 0.3129773 -0.003 0.9975   
## res[, 2] 0.0054009 0.1075313 0.050 0.9600   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01057 on 101 degrees of freedom  
## Multiple R-squared: 2.832e-05, Adjusted R-squared: -0.01977   
## F-statistic: 0.00143 on 2 and 101 DF, p-value: 0.9986  
##   
##   
## Response BRK.A.Close :  
##   
## Call:  
## lm(formula = BRK.A.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## BRK.A.Close  
## Min -0.0280903  
## 1Q -0.0073427  
## Median -0.0006556  
## 3Q 0.0053658  
## Max 0.0349374  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.002108 0.001061 -1.986 0.0497 \*  
## res[, 1] -0.262306 0.317616 -0.826 0.4108   
## res[, 2] -0.091501 0.109125 -0.838 0.4037   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01073 on 101 degrees of freedom  
## Multiple R-squared: 0.01053, Adjusted R-squared: -0.009059   
## F-statistic: 0.5377 on 2 and 101 DF, p-value: 0.5858  
##   
##   
## Response JPM.Close :  
##   
## Call:  
## lm(formula = JPM.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## JPM.Close  
## Min -3.988e-02  
## 1Q -7.783e-03  
## Median -4.634e-05  
## 3Q 9.098e-03  
## Max 4.065e-02  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.004537 0.001429 -3.176 0.00198 \*\*  
## res[, 1] 0.305748 0.427580 0.715 0.47622   
## res[, 2] -0.143483 0.146906 -0.977 0.33105   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01444 on 101 degrees of freedom  
## Multiple R-squared: 0.01977, Adjusted R-squared: 0.0003582   
## F-statistic: 1.018 on 2 and 101 DF, p-value: 0.3648  
##   
##   
## Response AAPL.Close :  
##   
## Call:  
## lm(formula = AAPL.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## AAPL.Close  
## Min -0.056942  
## 1Q -0.008914  
## Median -0.001501  
## 3Q 0.007468  
## Max 0.097225  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.0004763 0.0019781 0.241 0.810  
## res[, 1] 0.0472840 0.5919604 0.080 0.936  
## res[, 2] 0.2715023 0.2033829 1.335 0.185  
##   
## Residual standard error: 0.01999 on 101 degrees of freedom  
## Multiple R-squared: 0.01831, Adjusted R-squared: -0.001134   
## F-statistic: 0.9417 on 2 and 101 DF, p-value: 0.3934  
##   
##   
## Response MSFT.Close :  
##   
## Call:  
## lm(formula = MSFT.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## MSFT.Close  
## Min -0.041226  
## 1Q -0.008288  
## Median -0.001313  
## 3Q 0.008652  
## Max 0.053966  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.0004835 0.0014723 -0.328 0.743  
## res[, 1] 0.6634538 0.4405907 1.506 0.135  
## res[, 2] 0.1567723 0.1513761 1.036 0.303  
##   
## Residual standard error: 0.01488 on 101 degrees of freedom  
## Multiple R-squared: 0.02568, Adjusted R-squared: 0.00639   
## F-statistic: 1.331 on 2 and 101 DF, p-value: 0.2688  
##   
##   
## Response AMT.Close :  
##   
## Call:  
## lm(formula = AMT.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## AMT.Close  
## Min -0.053892  
## 1Q -0.006307  
## Median 0.002928  
## 3Q 0.008167  
## Max 0.038599  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.0009975 0.0015083 -0.661 0.510  
## res[, 1] 0.3684309 0.4513755 0.816 0.416  
## res[, 2] 0.0338341 0.1550814 0.218 0.828  
##   
## Residual standard error: 0.01524 on 101 degrees of freedom  
## Multiple R-squared: 0.006556, Adjusted R-squared: -0.01312   
## F-statistic: 0.3333 on 2 and 101 DF, p-value: 0.7174  
##   
##   
## Response SPG.Close :  
##   
## Call:  
## lm(formula = SPG.Close ~ res[, 1] + res[, 2])  
##   
## Residuals:  
## SPG.Close  
## Min -0.047994  
## 1Q -0.008497  
## Median 0.002386  
## 3Q 0.010817  
## Max 0.038717  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.004971 0.001545 -3.218 0.00174 \*\*  
## res[, 1] -0.130154 0.462261 -0.282 0.77886   
## res[, 2] -0.261910 0.158821 -1.649 0.10224   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01561 on 101 degrees of freedom  
## Multiple R-squared: 0.0266, Adjusted R-squared: 0.007328   
## F-statistic: 1.38 on 2 and 101 DF, p-value: 0.2562

As shown in the summary of the regression model, R squared is very low for all the stocks which means that the shocks of the macroeconomic factors does not have a significant impact on each of the stock’s returns.

rsq = rep(0,10)  
slmfit = summary(lmfit)  
for (i in 1:10){rsq[i]= slmfit[[i]][[8]]}   
beta\_inflation = lmfit$coef[2,]   
beta\_unemployment = lmfit$coef[3,]

par(mfrow=c(1,3))   
barplot(rsq,horiz=T,names=names(beta\_inflation),main="R squared")   
barplot(beta\_inflation,hori=T,main="beta inflation")   
barplot(beta\_unemployment,hori=T,main="beta unemployment")

 As shown in the the three barplots above, following the outcome of the summary of the economic model, we can conclude that the macroeconomic factor model is not significant.

# Statistical factor model:

## Evaluate the factorability of the data:

After detecting the correlation between the stocks, which identifies the existence of redundant variables that might be eliminated and presented by a smaller number of factors, we will do the following tests to confirm the factorability of our dataset:

* Bartlett’s test of sphericity
* Kaiser-Meyer-Olkin measure of sampling adequacy

cortest.bartlett(df\_portfolio\_ret);

## R was not square, finding R from data

## $chisq  
## [1] 12031.89  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 45

Bartlett’s test was utilized to determine whether or not variables intercorrelate by comparing the observed correlation matrix to a “identity matrix.” If this test is not statistically significant, a factor analysis should not be performed.  
Bartlett’s test is statistically significant in our case, indicating that the observed correlation matrix among the items is not an identity matrix. Which confirms that at least some of the variables are correlated with each other.

KMO(df\_portfolio\_ret)

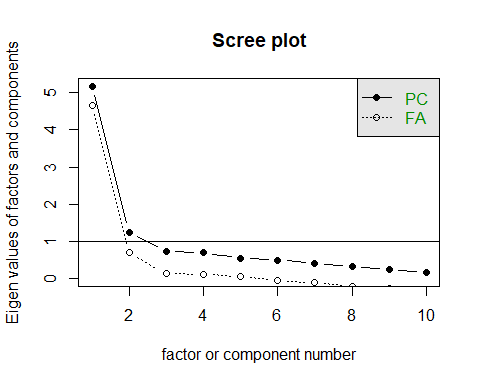
## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = df\_portfolio\_ret)  
## Overall MSA = 0.88  
## MSA for each item =   
## XOM CVX UNH JNJ BRK-A JPM AAPL MSFT AMT SPG   
## 0.81 0.82 0.95 0.90 0.90 0.90 0.86 0.86 0.91 0.93

The Kaiser-Meyer-Olkin (KMO) test determines whether the partial correlations in the data are not close enough to zero to indicate the presence of at least one latent factor underlying the variables.  
The lowest acceptable value is 0.50 that allows us to perform a factor analysis. In our case, the overall MSA=0,88 > 0,5 which allows us to conduct a factor analysis.

## Determining the number of factors to extract:

To determine the number of factors we should use to perform the factor model, we plot the scree plot of the returns. The scree presents the eigenvalues for all of our factors.To extract the optimal number of factors,we see where eigenvalues drop off sharply.

scree(df\_portfolio\_ret)



The scree plot above suggests that the number of factors is equal to three. After three factors the variability decreases.

fa\_none = factanal(df\_portfolio\_ret[,1:10], 3, rotation = 'none')  
print(fa\_none,cutoff = 0.1)

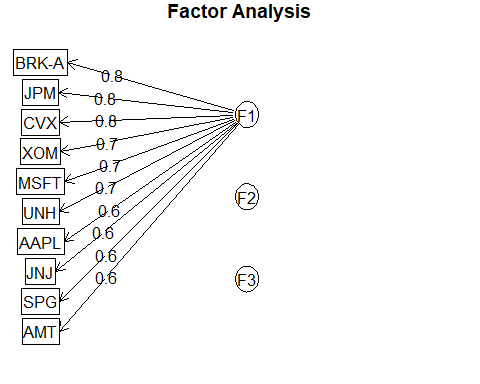
##   
## Call:  
## factanal(x = df\_portfolio\_ret[, 1:10], factors = 3, rotation = "none")  
##   
## Uniquenesses:  
## XOM CVX UNH JNJ BRK-A JPM AAPL MSFT AMT SPG   
## 0.180 0.165 0.556 0.620 0.193 0.310 0.412 0.245 0.665 0.640   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## XOM 0.809 -0.391 -0.115   
## CVX 0.840 -0.341 -0.112   
## UNH 0.621 0.240   
## JNJ 0.557 0.247   
## BRK-A 0.835 0.154 0.294   
## JPM 0.792 0.242   
## AAPL 0.583 0.450 -0.213   
## MSFT 0.642 0.520 -0.267   
## AMT 0.489 0.309   
## SPG 0.588 0.101   
##   
## Factor1 Factor2 Factor3  
## SS loadings 4.718 0.988 0.307  
## Proportion Var 0.472 0.099 0.031  
## Cumulative Var 0.472 0.571 0.601  
##   
## Test of the hypothesis that 3 factors are sufficient.  
## The chi square statistic is 370.58 on 18 degrees of freedom.  
## The p-value is 1.22e-67

Uniquenesses are the variance in each item that is not explained by the three factors. For example, 67% of the variance of AMT return is not explained by the 3 factors.

Factor loading is basically the correlation coefficient for the variable and factor. It shows the variance explained by the variable on that particular factor. If the value is missing and not shown it means that it is basically less than 0,1.

The chi-square statistic and p-value are testing the hypothesis that the model fits the data perfectly. When the p value is low, as it is here in our case, we can reject this hypothesis - which means that the 3 factor model does not fit the data perfectly.

factor.model <- fa(df\_portfolio\_ret[,1:10], nfactors = 3, fm="ols", max.iter = 100, rotate = "none")  
fa.diagram(factor.model)



As shown in the diagram above, the majority of the variances are explained by the first factor by 47%.

## Oblimin rotation:

Now, we we will use the oblim rotation. It is a type of oblique rotation which means it allows the factors to correlate. As opposed to orthogonal rotations (Varimax rotation). Instead of avoiding the relationship between our factors and splitting the sensitivity of their loadings over all of the factors, we find it more appropriate to employ the oblimin rotation because it allows correlation between the factors.

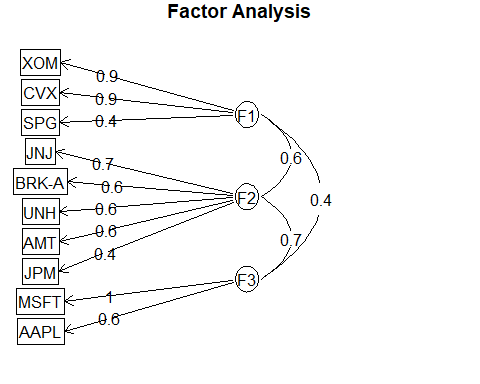
fa\_oblimin = factanal(df\_portfolio\_ret[,1:10],3, rotation = 'oblimin')  
print(fa\_oblimin,cutoff = 0.1)

##   
## Call:  
## factanal(x = df\_portfolio\_ret[, 1:10], factors = 3, rotation = "oblimin")  
##   
## Uniquenesses:  
## XOM CVX UNH JNJ BRK-A JPM AAPL MSFT AMT SPG   
## 0.180 0.165 0.556 0.620 0.193 0.310 0.412 0.245 0.665 0.640   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## XOM 0.917   
## CVX 0.877   
## UNH 0.389 0.310   
## JNJ 0.472 0.236   
## BRK-A 0.908   
## JPM 0.770 0.103   
## AAPL 0.740   
## MSFT 0.872   
## AMT 0.278 0.388   
## SPG 0.394 0.265   
##   
## Factor1 Factor2 Factor3  
## SS loadings 2.026 1.701 1.613  
## Proportion Var 0.203 0.170 0.161  
## Cumulative Var 0.203 0.373 0.534  
##   
## Factor Correlations:  
## Factor1 Factor2 Factor3  
## Factor1 1.000 0.438 -0.731  
## Factor2 0.438 1.000 -0.687  
## Factor3 -0.731 -0.687 1.000  
##   
## Test of the hypothesis that 3 factors are sufficient.  
## The chi square statistic is 370.58 on 18 degrees of freedom.  
## The p-value is 1.22e-67

After using the oblimin rotation, the variability among the loadings changed. The first factor is basically explained by the stocks coming from the healthcare sector, financial sector and real state sector. The second factor is mostly explained by the energy sector stocks and the third factor is mostly explained by the stocks coming from the technology sector ( apple and microsoft ) .

But, p value remains low( 1.22e-67 < 0,05), in this case, the 3 factor model does not fit the data perfectly.

factor.model <- fa(df\_portfolio\_ret[,1:10], nfactors = 3, fm="ols", max.iter = 100, rotate = "oblimin")  
fa.diagram(factor.model)



As shown in the diagram above, there is a high correlation between F1 and F2, also between F2 and F3. But almost a weak correlation between F1 and F3.

# Conclusion:

Following the completion of both an economic and a statistical factor analysis, we are able to draw the conclusion that the statistical analysis provided us with a more useful outcome in comparison to the economical factor analysis. This is the case even though the p value was not significant to validate the selection of the appropriate number of factors suggested by the scree plot, without and with oblimin rotation.