ACADEMIA DE STUDII ECONOMICE BUCUREȘTI

FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

SPECIALIZAREA MASTERULUI: ANALIZA AFACERILOR ȘI CONTROLUL PERFORMANȚEI ÎNTREPRINDERII

Știința Datelor în Afaceri

Proiect

Realizat în *RStudio*

Prof. MAER MATEI MONICA MIHAELA

Student GROSU MARILENA

Grupa 1108

Seria A

București

Ianuarie, 2020

1) Aplicația 1 - DASHBOARD

Scopul aplicației

Scopul acestei aplicații este acela de a crea o interfață grafică, numită tablou de bord, care să descrie într-un mod atractiv și intuitive indicatorii cheie aferenți datelor analizate.

Se cunoaște, de astfel, faptul că cel mai bine informația se înțelege din grafice, tabele, fiind organizată și sintetizată. În cadrul acestei analize, voi utiliza și eu mai multe tipuri de grafice.

Descrierea datelor

Datele pentru care voi realiza DashBoard-ul se referă la nivelul de dezvoltare economică a țărilor. Astfel, pentru 28 de țări membre Uniunii Europene, am ales analiza a 3 indicatori macroeconomici pentru anul 2018 și anume:

- Rata șomajului
- Rata inflației
- Rata de creștere economică

Sursa datelor

Datele sunt preluate de pe EUROSTAT, din mai multe tabele, uninnd informația necesară mie ulterior într-un Excel. Link-urile aferente sunt:

- https://ec.europa.eu/eurostat/databrowser/view/tps00203/default/table?lang=en
- https://ec.europa.eu/eurostat/databrowser/view/tec00118/default/table?lang=en
- https://ec.europa.eu/eurostat/databrowser/view/tec00115/default/table?lang=en

Aplicația propriu-zisă

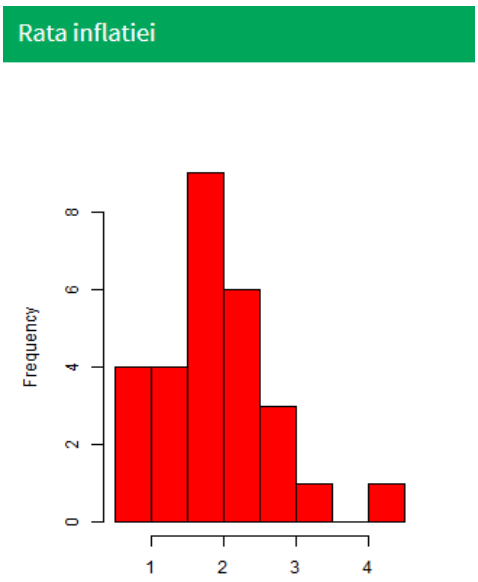
Înainte de crearea DashBoard-ului, am încărcat setul de date. Dimensiunea acestuia fiind de 28 x 3 (28 de țări și 3 indicatori macroeconomici). Astfel, output-ul cu datele încărcate este următorul:

	Unemployment.Rate	Inflation.Rate	Growth.Rate
Belgium	6.0	2.3	1.5
Bulgaria	5.2	2.6	3.1
Czechia	2.2	2.0	3.0
Denmark	5.1	0.7	2.4
Germany	3.4	1.9	1.5
Estonia	5.4	3.4	4.8
Ireland	5.8	0.7	8.2
Greece	19.3	0.8	1.9
Spain	15.3	1.7	2.4
France	9.1	2.1	1.7
Croatia	8.4	1.6	2.7
Italy	10.6	1.2	0.8

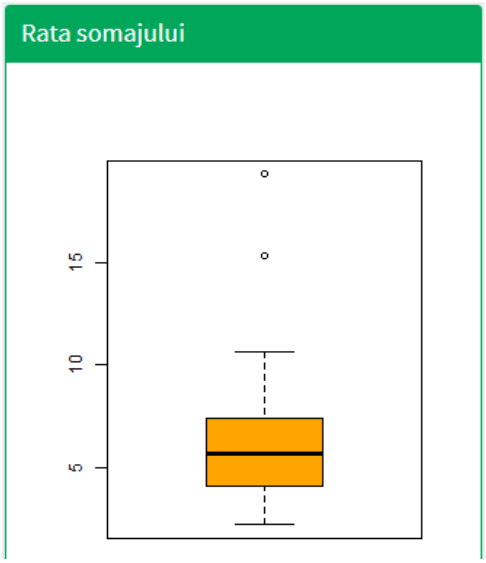
Mai departe, am luat indicatorii pe rând și i-am analizat, construindu-le câte un grafic. Pentru creșterea economică am ales BAR Chart-ul. Nuanța cea mai închisă de albastru simbolizează un nivel scăzut al ratei de creștere economică, iar cu cât se deschide, ajungând la galben, nivelul indicat crește.

Astfel, privind atât nuanța culorilor, cât și pe Ox și Oy, se constată faptul că o singură țară din eșantionul ales are o rată de creștere foarte mare, de 8.2, și anume: Irlanda. Tot o rată bună de creștere se remarcă și la Malta, în valoare de 7. Pe de altă parte, cea mai scăzută rată de creștere o are Italia și anume de 0.8.

Pentru rata inflației am ales histograma pentru identificarea formei distribuției datelor. Se constată faptul că forma histogramei este ușor alungită, ceea ce înseamnă că valorile se strâng în jurul mediei, având de-a face cu o formă leptocurtică. De asemenea, în partea dreapta datele sunt ușor asimetrice, ceea ce indică faptul că rata inflației este preponderant moderată, însă există și câteva țări pentru care rata inflației este una mare.



Pentru rata șomajului am ales boxplot-ul pentru a identifica cum sunt distribuite valorile și dacă există outlieri. Se constată faptul că poziționarea datelor este aproape simetrică, remarcându-se doar 2 valori foarte mari și anume pentru Grecia (19.3) și Spania (15.3).



Mai jos, am atașat imaginea de ansamblu a DashBoard-ului construit.

SCRIPT

```
# APLICATIA 1

# ~~~~~~~~~~~~~~~~~~~~~~~~~~DASHBOARD~~~~~~~~~~~~~~~~~~~~~~~~~~

#install.packages("shiny")

#install.packages ("flexdashboard")

#install.packages("shinydashboard")

#install.packages("plotly")

#install.packages("ggplot2")

library(ggplot2)

library(shiny)

library(flexdashboard)

library(shinydashboard)

library(plotly)
```

```
# Importul datelor in R
date <- read.csv("C:/Users/Marii/Desktop/Proiect/date DASHBOARD.csv",row.names=1)
date
attach(date)
View(date)
dim(date)
ui <- dashboardPage(
  ############# Header
  dashboardHeader(title = "UE Economic Growth Indicators"),
  ############# Sidebar
  dashboardSidebar(
    sidebarMenu(
      menuItem("DASHBOARD", tabName = "dashboard", icon = icon("dashboard"))
    )
  ),
  ############# Dashboard body
  dashboardBody(
    # Boxes need to be put in a row (or column)
    fluidRow(
      box(title = "Alegeti din urmatoarea lista", status = "success", solidHeader = T, width = 12,
          fluidPage(
            fluidRow(
              column(2, offset = 0, style = 'padding:1px;',
                     selectInput(inputId = "unit",
                                 label = "Tara",
                                 choices = row.names(date)))
            )
          )
      ),
      column(width=8,
             box(title="Rata de crestere economica",status="success",solidHeader = T,
plotlyOutput("plot0"),width=NULL),
```

```
            box(title = "Rata inflatiei", status="success", solidHeader = T,plotOutput("plot1")),
            box(title="Rata somajului", status="success", solidHeader=T, plotOutput("plot2"))
      ),    ) ) )
#######
####### Server
#######
server <- function(input, output, session) {
  output$plot0 <- renderPlotly({
    plot_ly(date, x = ~Growth.Rate, text = paste("Tara: ", row.names(date)),
            type = 'bar', mode = "markers", color = ~Growth.Rate)
      })
    output$plot1 <- renderPlot({
    hist(date$Inflation.Rate,col="red",main="",xlab="")
  })
  output$plot2 <- renderPlot({
    boxplot(date$Unemployment.Rate,col="orange",main="",xlab="")
  })
}
shinyApp(ui, server)
```

2) Aplicația 2 - *REGRESIE LOGISTICĂ* - Estimarea unui model pentru analiza înclinației către antreprenoriat

3) Aplicația 3 - WORDCLOUD

Scopul aplicației:

Pornind de la selecția datelor și până la operarea lor, scopul acestei părți a proiectului este aceea de a forma un nor a celor mai semnificative cuvinte cu privire la tema aleasă. De asemenea, se urmăresc asocierile cele mai puternice din text, dar și cele mai frecvente cuvinte utilizate în descrierea subiectului abordat.

Descrierea datelor:

Datele colectate sunt de tip text si abordează subiectul materiilor prime. Documentul este alcătuit din 11 paragrafe, preluate de pe site-uri diferite, separate prin Enter. Pentru o mai bună operare a acestora, datele preluate sunt în limba engleză.

Sursa datelor:

Datele au fost preluate de pe internet: Investopedia, Dividend, Reuters, Seekingalpha, Fool, Wika, Valuepenguin. Link-urile aferente sunt:

- https://www.investopedia.com/terms/b/basic_materials.asp
- https://www.dividend.com/dividend-stocks/basic-materials/#tm=3-sector-stocks&r=ES%3A%3ADividendStock%3A%3AStock%23WLK--NYSE&f_1=basic-materials
- https://www.reuters.com/article/us-brazil-environment-mining/brazil-pushes-plans-for-mining-on-tribal-lands-to-european-diplomats-idUSKBN1Z92C2
- https://www.reuters.com/article/usa-minerals-recycling/apple-pushes-recycling-with-robot-but-mined-metals-still-needed-idUSL1N298151
- https://seekingalpha.com/article/4270282-basic-materials-you-sufficient-exposure
- https://www.fool.com/investing/general/2014/08/06/basic-materials-investing-essentials.aspx
- https://www.wika.ro/industries_basic_materials_en_co.WIKA
- https://www.valuepenguin.com/sectors/materials#chemicals
- https://www.valuepenguin.com/sectors/materials#construction
- https://www.valuepenguin.com/sectors/materials#metals
- https://www.valuepenguin.com/sectors/materials#paper

Aplicația propriu-zisă

Înainte de prelucrarea datelor, le-am încărcat și le-am salvat într-un vector sursă.

1	The basic materials sector is a category of stocks for compa...
2	Basic materials companies are involved in the exploration, d...
3	Brazil's government on Friday said it was pushing ahead wit...
4	Apple Inc is trying to change the way electronics are recycle...
5	The basic materials sector is often overlooked by many inve...
6	Some companies specialize in producing gold, silver, platinu...
7	Measurement technology for the basic materials industry Th...
8	Chemicals The chemicals industry includes basic or "commo...
9	Construction Materials Companies within the construction ...
10	Metals and Mining The metals and mining industry includes...
11	Paper and Forest Products Companies in this industry produ...

Mai departe, am creat o colecție de documente ce conține textul original în limba engleză. Asupra colecției, am efectuat următoarele prelucrări:

- transformarea majusculelor în minuscule
- eliminarea punctuației
- eliminarea spațiilor albe multiple
- eliminarea numerelor
- eliminarea prepozițiilor

Continuând, am eliminat cuvintele care nu aduc niciun plus de informație, precum: "basic", "materials", "sector", "companies", "also", "many", "within" și am transformat colecția de documente într-o matrice cu (n,m) dimensiuni, unde n reprezintă numărul de paragrafe, iar m numărul de cuvinte din paragrafe.

Name	Type	Value
dtm	list [11 x 556] (S3: DocumentTern	List of length 6
i	integer [683]	1 1 1 1 1 1 ...
j	integer [683]	1 2 3 4 5 6 ...
v	double [683]	1 1 1 1 1 1 ...
nrow	integer [1]	11
ncol	integer [1]	556
dimnames	list [2]	List of length 2

Se observă în output-ul alăturat că matricea creată are 11 linii și 556 coloane.

```
> dim(dtm)
[1]  11 556
```

Cele mai frecvente 10 cuvinte utilizate în cele 11 paragrafe sunt cele din output-ul următor.

“products” a fost folosit de 16 ori, urmat de “chemicals”, “mining” și “industry”.

```
> frequency[1:10] #cele mai frecvente cuvinte
    products    chemicals       mining     industry       metals   processing
          16           14           11           11            7            6
      stocks     minerals      process construction
           6            6            6            5
```

Pentru a putea vedea ce cuvinte caracterizează cel mai bine **industriile** de ”**chemicals**” și ”**metals**”, din sectorul **MATERII PRIME**, am utilizat o funcție, findAssocs( ) , care găsește toate asocierile peste o anumită proporție din totalul acestora.

Astfel, conform primul output de mai jos, cele mai reprezentative cuvinte care se asociază cu ”chemicals”, într-o măsură mai mare de 80% sunt: „gases”, „oxygen”, „helium” ceea ce indică faptul că, preponderent, în industria chimică se folosesc gaze naturale, dar și “agricultural”, unul dintre domeniile de bază în care are aplicabilitate această industrie pentru dezvoltarea recoltei.

Pe de altă parte, conform celui de-al doilea output, cele mai reprezentative cuvinte care se asociază cu ”metals”, într-o măsură mai mare de 50% sunt: “steel”, “gold”, “aluminum”, “gemstones” care indică principalele elemente cu care se operează în această industrie.

```
> findAssocs(dtm, "chemicals", 0.8)
$chemicals
          activities       classification           classified
                0.98                 0.98                 0.98
         cropgrowing                 four                gases
                0.98                 0.98                 0.98
              helium                 make               oxygen
                0.98                 0.98                 0.98
            purposes            specialty                three
                0.98                 0.98                 0.98
           typically                  use \u0093commodity\u0094
                0.98                 0.98                 0.98
          industrial          agricultural               groups
                0.95                 0.95                 0.85
```

```
> findAssocs(dtm, "metals", 0.5)
$metals
        steel          gold        africa       african       finance     gemstones
         0.85          0.81          0.76          0.76          0.76          0.76
interestingly       located      reserves         south       variety      aluminum
         0.76          0.76          0.76          0.76          0.76          0.64
```

În final, am ajuns și la "Norul de cuvinte" (WORD CLOUD) ce indică vizual importanța anumitor termeni în textul ales. Cu cât dimensiunea lor este mai mare, iar culorile sunt mai puternic evidențiate, cu atât aceștia descriu mai bine tema abordată.

Asfel, în sectorul materiilor prime cuvintele cheie sau grupurile de cuvinte cheie sunt următoarele: industria produselor chimice ("industry" of "chemicals" "products"), minerit ("mining"), procesarea metalelor ("metals" "processing") etc.

SCRIPT

```
# APLICATIA 3

############################## Obiectivul: NOR DE CUVINTE - date: text

#install.packages("tm")

#install.packages("wordcloud")
```

```
#install.packages("SnowballC")

#install.packages("RColorBrewer")

library(tm)

library(SnowballC)

library(wordcloud)

library(RColorBrewer)

a<-readLines("C:/Users/Marii/Desktop/Proiect/date - aplicatia 3.txt")

a

View(a)

sursa<-VectorSource(a)

sursa

corpus<-Corpus(sursa)

corpus<-tm_map(corpus, content_transformer(tolower))

corpus<-tm_map(corpus, removePunctuation)

corpus<-tm_map(corpus, stripWhitespace)

corpus<-tm_map(corpus, removeNumbers)

corpus<-tm_map(corpus, removeWords, stopwords("english")) #sterge prepozitiile

corpus

#corpus<-tm_map(corpus, content_transformer(gsub), pattern="statistical", replacement="statistics")

#stergere<-c("performance", "contribuie", "work") !!! etc cuvinte care nu au importanta, care nu aduc un plus de informatie

#corpus<-tm_map(corpus, removeWords, stergere)

stergere<-c("basic", "materials", "sector", "companies", "also", "many", "within", "used", "includes", "said", "including", "daisy")# !!! etc cuvinte care nu au importanta, care nu aduc un plus de informatie

corpus<-tm_map(corpus, removeWords, stergere)

#Corpus=colectie de documente

#MAtrix transforma corpus intr-o matrice -> Linii(Nr de paragrafe) si coloane(toate cuvintele)

dtm<-DocumentTermMatrix(corpus)

dtm

View(dtm)

dim(dtm)

dtm2<-as.matrix(dtm)
```

```
frequency<-colSums(dtm2)
frequency<-sort(frequency, decreasing=TRUE)
frequency[1:10] #cele mai frecvente cuvinte
### DAca vrem sa vedem cu ce se asociaza un cuvant
findAssocs(dtm, "chemicals", 0.8)
findAssocs(dtm, "metals", 0.5)
set.seed(1) #sa nu se schimbe ordinea cuvintelor din grafic
wordcloud(names(frequency), frequency, min.freq=10)
windows()
wordcloud(names(frequency[1:10]), frequency[1:10], colors=brewer.pal(6, "Dark2"))
```

4) Aplicația 4 – MODELUL LUI SHARPE, IMPUTARE VALORI NA, DETECTARE OUTLIERI

Scopul aplicației:

Această aplicație este structurată pe 3 părți și anume:

1. Modelul de piață a lui Sharpe ce descrie rentabilitățile și riscul investițiilor

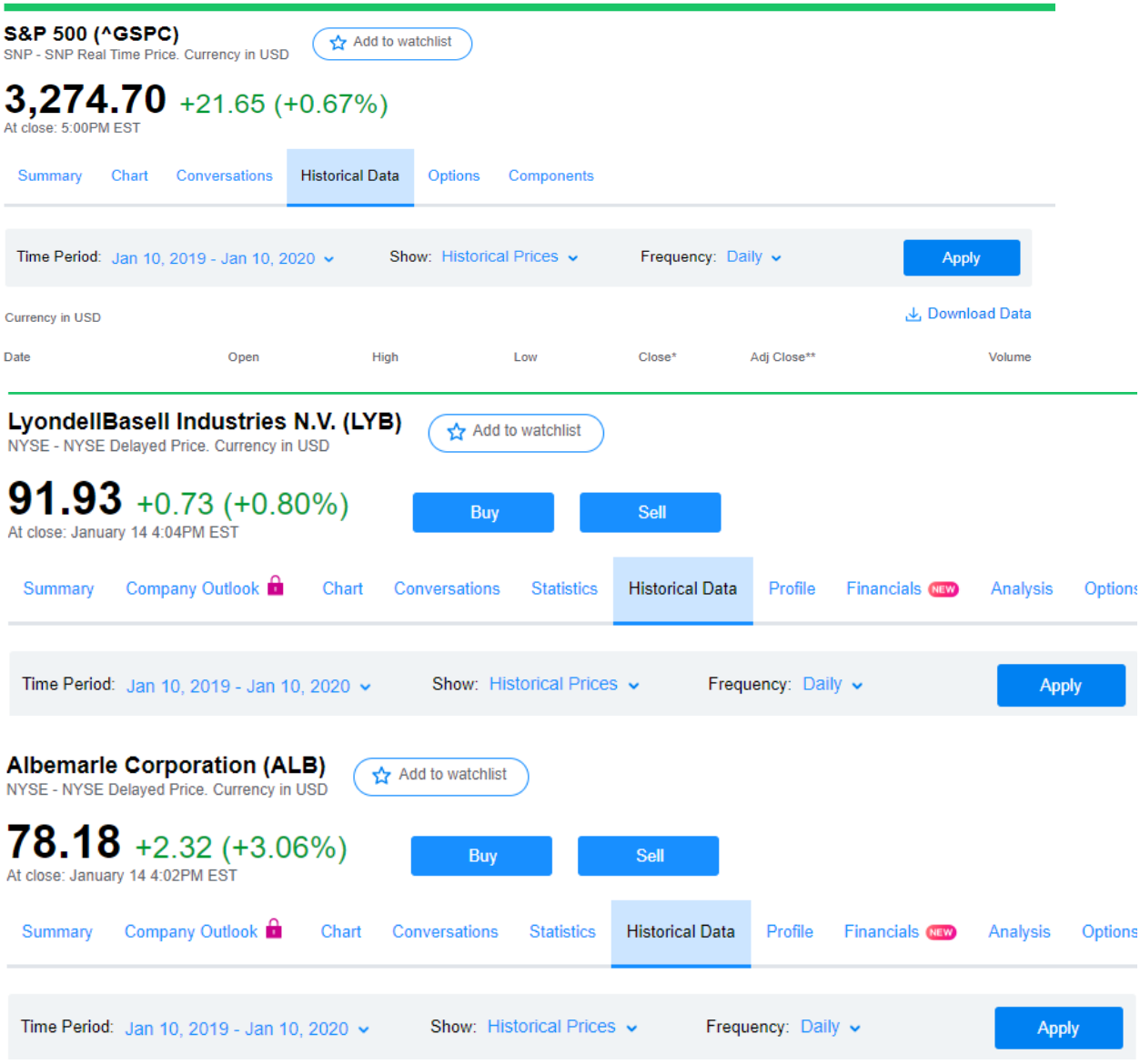
2. Imputarea valorilor lipsă prin intermediul a 3 algoritmi:

- ***MICE***
- ***HMISC***
- ***AMELIA***

3. Detectarea outlierilor prezenți în date

Descrierea datelor:

Pentru această aplicație, am ales prețurile de închidere ajustate (***Adj. Close***) pentru 2 companii din sectorul **MATERII PRIME,** industria **PETROLIERĂ** și pentru un indice de piață S&P 500. Perioada de timp pentru care se efectuează analiza este de 1 an: 10.01.2019-10.01.2020, iar frecvența datelor este zilnică.



Sursa datelor:

Datele au fost preluate de pe **Yahoo! Finance**, categoria **Industries** -> **Basic Materials.**

- https://finance.yahoo.com/quote/%5EGSPC?p=^GSPC&.tsrc=fin-srch
- https://finance.yahoo.com/quote/LYB/history?period1=1547078400&period2=1578614400&interval=1d&filter=history&frequency=1d
- https://finance.yahoo.com/quote/ALB/history?period1=1547078400&period2=1578614400&interval=1d&filter=history&frequency=1d

1. *Modelul de piață a lui Sharpe ce descrie rentabilitățile și riscul investițiilor*

Aspecte teoretice

Acest model are rol în descrierea rentabilităților și riscurilor unei investiții. Ideea de bază a modelului este că variația anumitor titluri sau acțiuni este influențată de piață.

Ecuația modelului se poate scrie sub forma:

$$R_{j,t} = \alpha_j + \beta_j R_{M,t} + \varepsilon_{j,t}$$

unde :

$R_{j,t}$ = rata de rentabilitate a acțiunii j, în perioada t ;

$R_{M,t}$ = rata de rentabilitate a pieței, în perioada t ;

β_j = parametru propriu fiecărei acțiuni, care indică relația care există între fluctuațiile acțiunii j și fluctuațiile pieței; se mai numește coeficient de volatilitate sau simplu beta ;

$\varepsilon_{j,t}$ = variabilă specifică acțiunii j, care însumează alți factori de influență asupra titlului j, înafară de piață;

α_j = parametru care arată locul de intersecție a dreptei de regresie cu axa ordonatei, reprezentând rentabilitatea care ar putea fi obținută de titlul j, în condițiile în care rentabilitatea pieței este 0 .

β este egal cu covarianța dintre rentabilitatea titlului j și rentabilitatea pieței, raportată la varianța ratei de rentabilitate a pieței, dupa expresia:

$$\beta_j = \frac{\sigma_{jM}}{\sigma_M^2}$$

Conform teoriei moderne a portofoliilor, β este elementul central pentru că el măsoară riscul sistematic al acelui titlu sau portofoliu. În funcție de valoarea pe care o ia acesta, acțiunile se pot împărți în mai multe categorii :

- acțiuni cu volatilitate unitară: variază în același sens și în aceeași proporție cu piața; achiziționarea unei astfel de acțiuni presupune expunerea investitorului exact la riscul pieței ;
- acțiuni cu volatilitate subunitară (nevolatile): variază în același sens dar într-o proporție mai mică ca piața; expunerea la riscul pieței este mai mică, ele fiind acțiunile „defensive";
- acțiuni cu volatilitate supraunitară (volatile): variază în același sens dar într-o proporție mai mare ca piața; sunt acțiunile „ofensive" care amplifică variația pieței și sunt atractive când se anticipează o tendință ascendentă a pieței .

Riscul sistematic este egal cu beta înmulțit cu abaterea medie pătratică a pieței: $\beta\sigma(R_M)$.

Riscul specific este egal cu abaterea medie pătratică a factorului rezidual : $\sigma(\varepsilon_j)$, aceasta fiind măsura variabilității proprii titlului.

Riscul unui portofoliu depinde de trei factori:

- riscul fiecărui titlu inclus în portofoliu;
- covarianța dintre randamentele acțiunilor din portofoliu;
- numarul de titluri din portofoliu.

Un portofoliu va fi cu atât mai riscant cu cât titlurile care-l conțin vor avea un β mai mare. Gradul de interdependență a variațiilor de curs între ele au o mare importanță în reducerea riscului portofoliului. În general, două acțiuni nu vor varia de o manieră total independentă. Covarianța lor este în general mai mare de 0. În acest caz, reducerea riscului nu este așa de mare ca și în cazul în care cele doua acțiuni vor varia independent. Componenta de piață a unui portofoliu va varia de o manieră „sistematică" dată de incertitudinile pieței. Este imposibil de a elimina acest risc și orice investitor și-l va asuma mai mult sau mai puțin. Componenta independentă a portofoliului dată de factorii specifici societaților cotate poate fi eliminată ușor prin diversificarea portofoliului.

Sursă: **scribd.com – Andreea Elena Hategan**

(https://www.scribd.com/document/63561631/Modelul-de-pia%C5%A3%C4%83-a-fost-dezvoltat-de-Sharpe)

Aplicația propriu-zisă

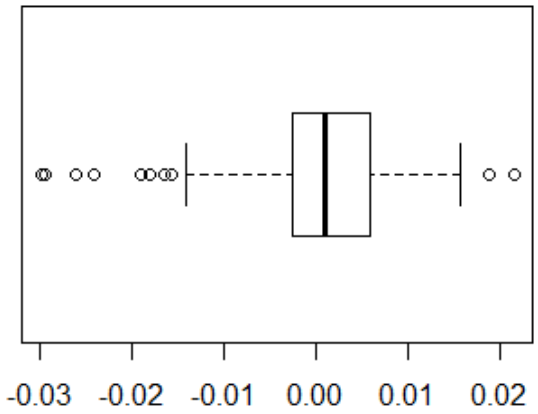
După formarea setului de date, am calculate în Excel rentabilitățile, atât pentru companii, cât și pentru indice, după formula:

$$R = \frac{\text{Adj.Close t} - \text{Adj.Close t}-1}{\text{Adj.Close}_{t-1}}$$

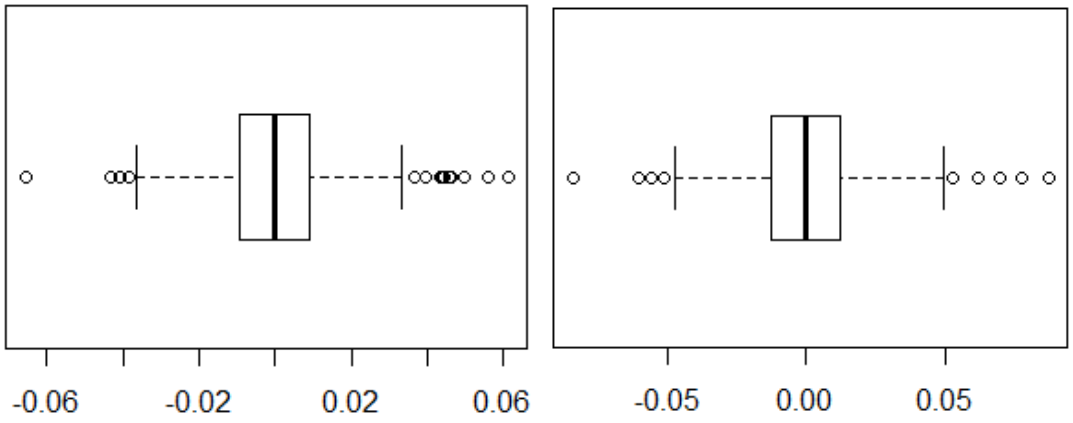
	Date	LYB	ALB	S.P.500	Ra1	Ra2	Rm
1	1/10/2019	83.15873	75.71443	2596.64	*NA*	*NA*	*NA*
2	1/11/2019	81.82668	75.85162	2596.26	-0.01602	0.00181	-0.00015
3	1/14/2019	81.55074	75.42048	2582.61	-0.00337	-0.00568	-0.00526
4	1/15/2019	80.47559	72.15751	2610.30	-0.01318	-0.04326	0.01072
5	1/16/2019	80.66587	72.44167	2616.10	0.00236	0.00394	0.00222
6	1/17/2019	81.68395	73.77429	2635.96	0.01262	0.01840	0.00759
7	1/18/2019	83.59642	75.08732	2670.71	0.02341	0.01780	0.01318
8	1/22/2019	81.63638	73.85269	2632.90	-0.02345	-0.01644	-0.01416
9	1/23/2019	80.33286	72.43188	2638.70	-0.01597	-0.01924	0.00220
10	1/24/2019	79.88567	72.52985	2642.33	-0.00557	0.00135	0.00138
11	1/25/2019	82.02647	75.03833	2664.76	0.02680	0.03459	0.00849
12	1/28/2019	81.95987	75.53806	2643.85	-0.00081	0.00666	-0.00785
13	1/29/2019	82.86377	76.74330	2640.00	0.01103	0.01596	-0.00146
14	1/30/2019	83.93894	78.59525	2681.05	0.01298	0.02413	0.01555

Întrucât orizontul de timp începe de la 1/10/2019, pentru această zi, nu am putut calcula rentabil itățile activelor și a indicelui, nefiind o zi anterioară de raportare. Prin urmare, această înregistrare va fi ște arsă din setul de date. Mai departe, am creat boxplot-uri pentru fiecare Rm, Ra1 și Ra2 cu scopul de a ved ea tendința centrală și forma distribuției lor.

Rentabilitatea indicelui are o distribuție ușor asimetrică și platicurtică, 16epend16 desemnează f aptul că rentabilitățile înregistrează, în medie, valori mari, însă sunt și zile când acestea sunt mai mici. Out lierii sunt într-un număr redus, făcându-și apariția mai mult spre stânga Boxplot-ului.

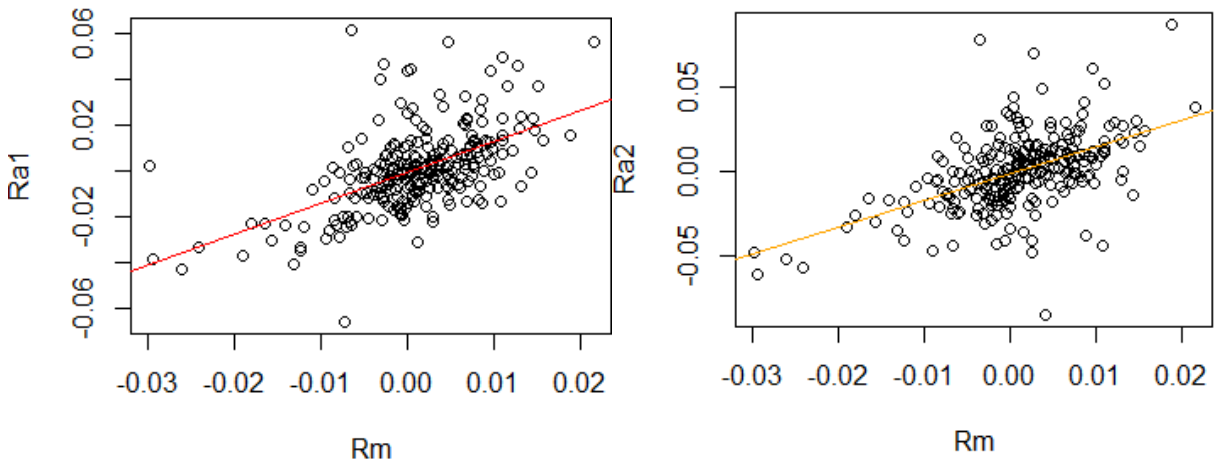


Rentabilitățile companiilor LYB (stânga) și ALB (dreapta) au o distribuție 17epend, rentabilităț ile înregistrând, în medie, valori apropiate de 0. Astfel, din punct de vedere economic, firmele au atât zile când sunt profitabile, cât și zile în care sunt mai puțin profitabile. Outlierii sunt într-un număr redus, făcân du-și apariția mai mult spre dreapta Boxplot-ului, a valorilor mari.

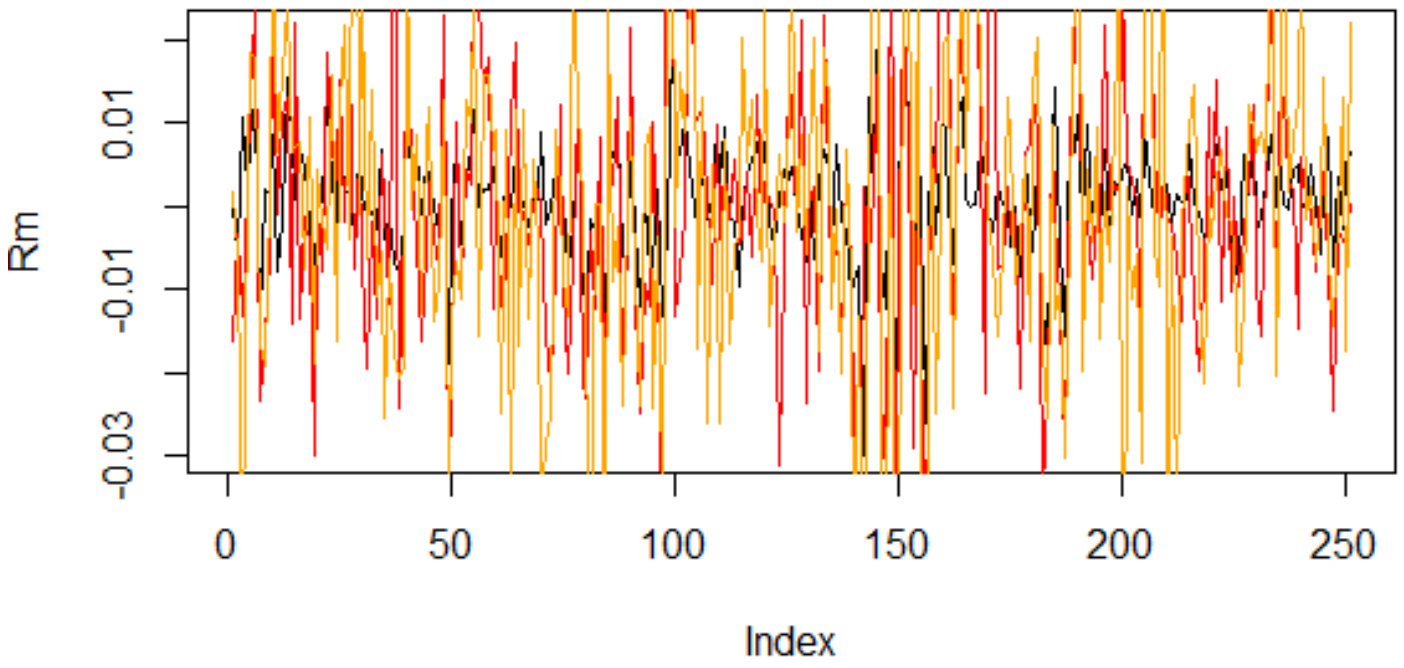


Pasul următor din analiză este acela de a vedea dacă există dependență între piață și active. Con form Scatterplot-urilor și coeficienților de corelație de mai jos (~0.56 și ~0.53), se observă că ambele activ e 17epend într-o proporție destul de mare de piață, în mod linear și pozitiv. Astfel, mișcările pieței influen țează moderat spre ridicat evoluția activelor.

```
> cor(Rm, Ra1)
[1] 0.5569331
> cor(Rm, Ra2)
[1] 0.5286853
```



Înainte de estimarea modelului Sharpe, testez, vizual, staționaritatea seriilor de timp: LYB, ALB și S.F.500. Se poate observa faptul că acest grafic este asemănător cu cel al zgomotului alb (”White Noise ”) pentru toate cele 3 serii. Astfel, se constantă faptul că există staționaritate în setul de date. Acest lucru e ra de așteptat, întrucât e vorba de o volatilitate foarte mare pe piața activelor.



```
> ###Estimare model Sharpe
> model1<-lm(Ra1~Rm)
> summary(model1)

Call:
lm(formula = Ra1 ~ Rm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.055079 -0.008998 -0.001034  0.005737  0.070649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0007815  0.0009492  -0.823    0.411
Rm           1.3447676  0.1270908  10.581   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01492 on 249 degrees of freedom
Multiple R-squared:  0.3102,    Adjusted R-squared:  0.3074
F-statistic:   112 on 1 and 249 DF,  p-value: < 2.2e-16
```

Ecuația modelului construit pe rentabilitatea activelor companiei LYB și rentabilitatea pieței se scrie sub următoarea formă:

$$\mathbf{R_{LYB} = 0.0007 + 1.3447\ R_{S.P.500}}$$

Coeficientul pieței, în valoare de 1.3447, cu p-value <2e-16 ~ 0.00000..2, foarte apropiat de 0 e ste mai mic decat 0.05. Astfel, pentru un nivel de încredere de 95%, acesta e semnificativ din punct de ved ere statistic.

R-squared = Variația explicată de variabilele independente / Variația totală

Valoarea Adjusted R-squared de 0.3074 indică faptul că modelul explică în proporție de 30% va riația datelor în jurul valorii medii.

F-critic (0.05:1:249) este în valoare **de 3.89.**

Conform output-ului de mai sus, F-statistic 112 este mai mare decât valoarea critică de 3.89 pentru un nivel de întredere de 95%, ceea de indică validitatea modelului, fapt confirmat și de p-value al modelului de 2.2e-16~ 0.00000..22, foarte apropiat de 0, deci mai mic decat 0.05.

```
> model2<-lm(Ra2~Rm)
> summary(model2)

Call:
lm(formula = Ra2 ~ Rm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.089425 -0.009719 -0.000935  0.008490  0.084928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001426   0.001208  -1.180    0.239
Rm           1.589835   0.161759   9.828   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01898 on 249 degrees of freedom
Multiple R-squared:  0.2795,    Adjusted R-squared:  0.2766
F-statistic:  96.6 on 1 and 249 DF,  p-value: < 2.2e-16
```

Ecuația modelului construit pe rentabilitatea activelor companiei ALB și rentabilitatea pieței se scrie sub următoarea formă:

R_{ALB} = 0.0014 +1.589835 $R_{S.P.500}$

În ceea ce privește coeficientul pieței, în valoare de 1.589835, cu p-value <2e-16 ~ 0.00000..2, f oarte apropiat de 0 este mai mic decat 0.05. Astfel, pentru un nivel de încredere de 95%, acesta e semnific ativ din punct de vedere statistic.

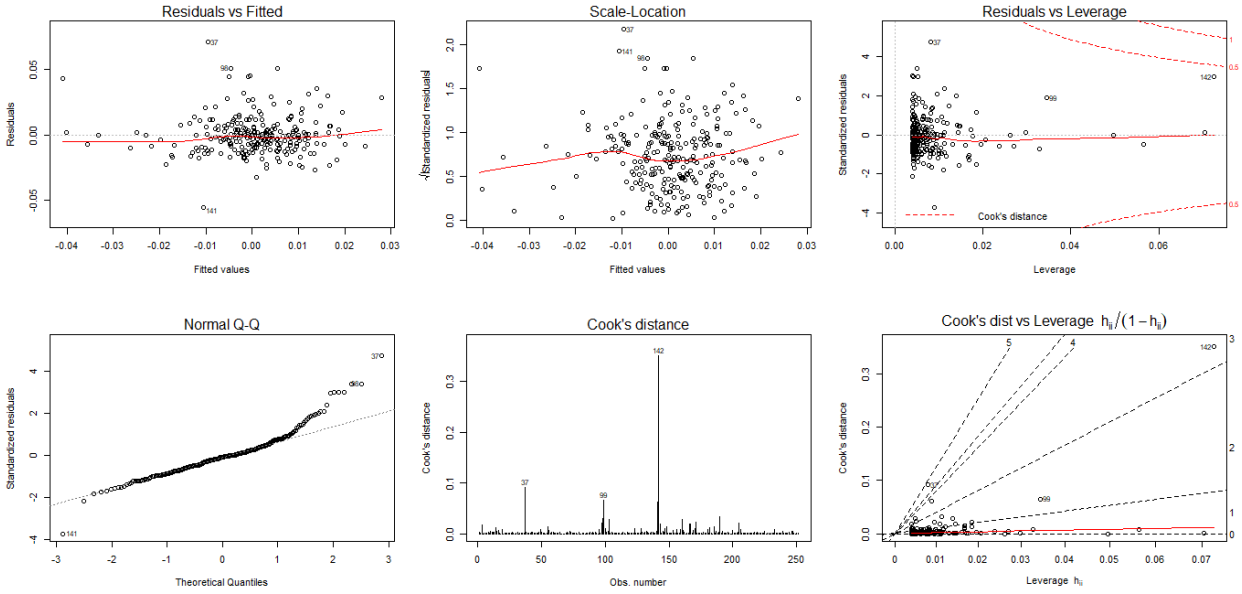
R-squared = Variația explicată de variabilele independente / Variația totală

Valoarea Adjusted R-squared de 0.2766 indică faptul că modelul explică în proporție de 27.66% ~ 28% variația datelor în jurul valorii medii.

F-critic (0.05:1:249) este în valoare **de 3.89.**

La fel ca și în cazul anterior, F-statistic 96.6 este mai mare decât valoarea critică de 3.89 pentru un nivel de întredere de 95%, ceea de indică validitatea modelului, fapt confirmat și de p-value al modelului de 2.2e-16~ 0.00000..22, foarte apropiat de 0, deci mai mic decat 0.05.

Întrucât cele două modele sunt foarte asemănătoare, voi explica cele mai reprezentative grafice corespunzătoare relației dintre rentabilitatea activelor LYB și rentabilitatea piației.



I. Residuals vs Fitted

Graficul indică existența sau inexistența reziduurilor non-liniare în model. Se observă faptul că reziduurile urmează linia orizontală roșie în mod uniform, deci modelul nu are reziduuri non-liniare.

II. Scale Location

Graficul arată dacă reziduurile sunt distribuite egal de-a lungul intervalelor de predictori, dar re prezintă și un test vizual al homoscedasticității. Se observă, în output, faptul că reziduurilor sunt uniform d istribuite, de-a lungul liniei roșii.

III. Residuals vs Leverage & Cook's distance

Distanța Cook indică ce valori sunt considerate outlieri și necesită verificare și, eventual, elimin area lor din analiză. Astfel, rezultatele regresiei ar putea fi îmbunătățile.

Conform output-ului de mai sus, nu există valori care să depășească distanța lui Cook, însă sunt câteva destul de apropiate de acestea și trebuiesc verificare și anume: 37, 99, 142.

IV. Normal Q-Q

Acest tip de graphic arată cum sunt aranjate reziduurile față de quantile teoretice. În modelul stu diat, valorile reziduurilor se distanțează ușor spre moderat de capete.

ne arată repartiția rezidurilor față de cuantilele teoretice. In graficul modelului nostru, putem observa cateva valori usor departate fata de capete.

2. Imputarea valorilor lipsă

Aspecte teoretice

Există mai multe tipologii de valori lipsă:

- Valori lipsă complet aleatorii
- Valori lipsă aleatorii
- Valori lipsă care depind de predictori neobservați
- Valori lipsă care depind de valoarea lipsă în sine

Tratarea valorilor lipsă se poate realiza cu ajutorul următorilor algoritmi:

- MICE - Multivariate Imputation via Chained Equations

  Metoda presupune că valorile lipsesc în mod aleator și determină pe rând pentru fiecare variabilă valori estimate pe baza celorlalte variabile disponibile. Metodele pe care le folosește diferă în funcție de tipul de date. În cazul de față, voi aplica PMM (Predictive Mean Matching) întrucât e vorba de variabile numerice.

- HMISC

  Metoda presupune utilizarea de tehnici statistice alese de utilizator (medie, mediana, random) prin funcția input, cât și folosind bootstraping, PMM prin funcția aregImpute. Prin bootstraping se realizează imputare multiplă: se estimează regresii pe eșantioanele extrase și apoi se aplică PMM pentru imputarea valorilor lipsă.

- AMELIA

  Metoda presupune că variabilele au o distribuție normală multivariată (variabilele trebuie sa urmeze această distribuție sau sa fie trasformate în așa fel încât sa se apropie cât mai mult de ea). Se extrag m esantioane pentru bootstraping și pe fiecare aplică algoritmul EMB (Expected Maximization Bootstrap). Seturile de estimatori obținute se folosesc pentru imputarea prin regresie a fiecărui set de observații lipsă.

  Comparativ cu MICE care realizează imputarea variabilă cu variabilă, Amelia realizează imputarea concomitentă, considerând distirubția normală multivariată.

Aplicația propriu-zisă

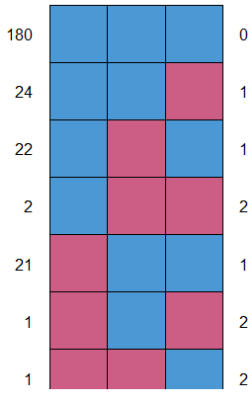
Primul pas în această parte a fost acela de a crea un subset de date format doar din rentabilități: rentabilitatea pieței și rentabilitatea activelor celor două companii LYB și ALB.

Pasul următor a constat în verificarea existenței valorilor nule. Întrucât nu existau valori nule, am generat 10% cu ajutorul funcției *prodNA( )*.

Următorul pas a fost identificarea numărului de valori nule pentru fiecare indicator. Cu ajutorul unui summary, dar și cu ajutorul unui grafic, am identificat pattern-ul valorilor lipsă. Astfel, în privința activelor LYB - 25 de valori sunt nule, în privința activelor ALB – 27 de valori nule, respectiv în privința pieței – 23, conform output-ului de mai jos.

În privința culorilor graficului alăturat, albastru reprezintă valorile prezente, iar mov reprezintă valorile nule generate.

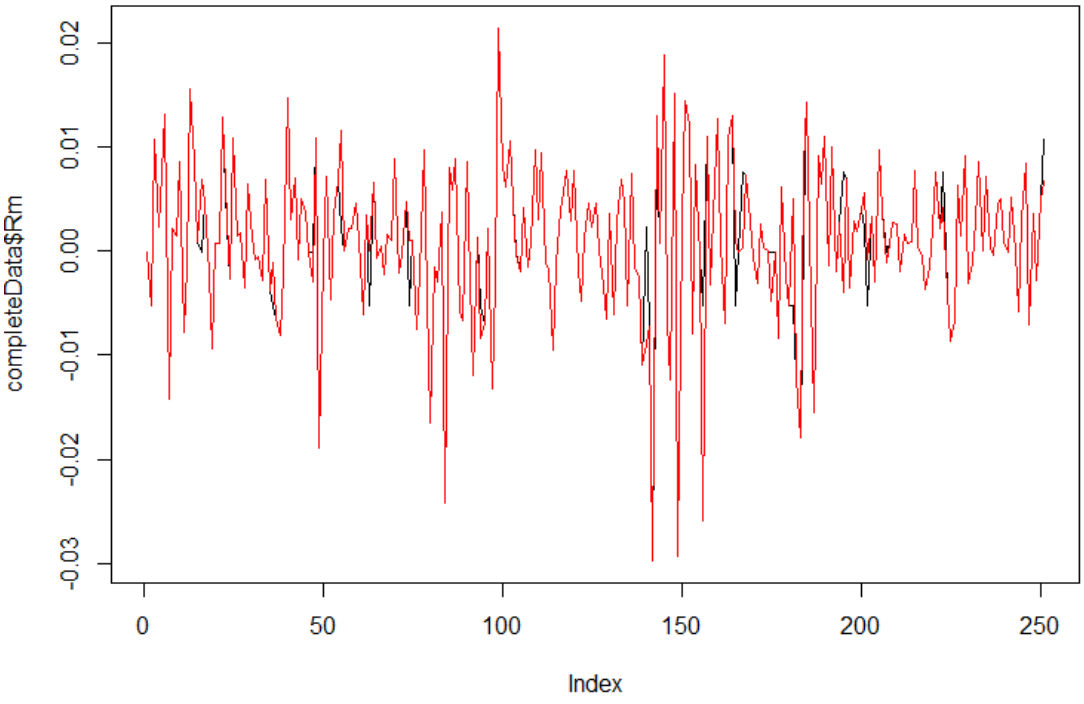
```
> summary(msf.mis)
      Ra1                 Ra2                  Rm          
 Min.   :-0.043110   Min.   :-0.084380   Min.   :-0.029780  
 1st Qu.:-0.009495   1st Qu.:-0.013145   1st Qu.:-0.002598  
 Median :-0.000500   Median :-0.000435   Median : 0.001030  
 Mean   : 0.000243   Mean   :-0.000707   Mean   : 0.001015  
 3rd Qu.: 0.008630   3rd Qu.: 0.011615   3rd Qu.: 0.005915  
 Max.   : 0.061100   Max.   : 0.087290   Max.   : 0.021430  
 NA's   :25          NA's   :27          NA's   :23         
```



1) MICE

Primul algoritm utilizat în completarea valorilor nule este MICE - Multivariate Imputation via Chained Equations. Așa cum am zis și mai sus, se consideră că valorile lipsesc în mod aleator. Prin intermediul acestei metode, se estimează valorile lipsă pe baza celor prezente cu ajutorul PMM-ului. (Predictive Mean Matching)

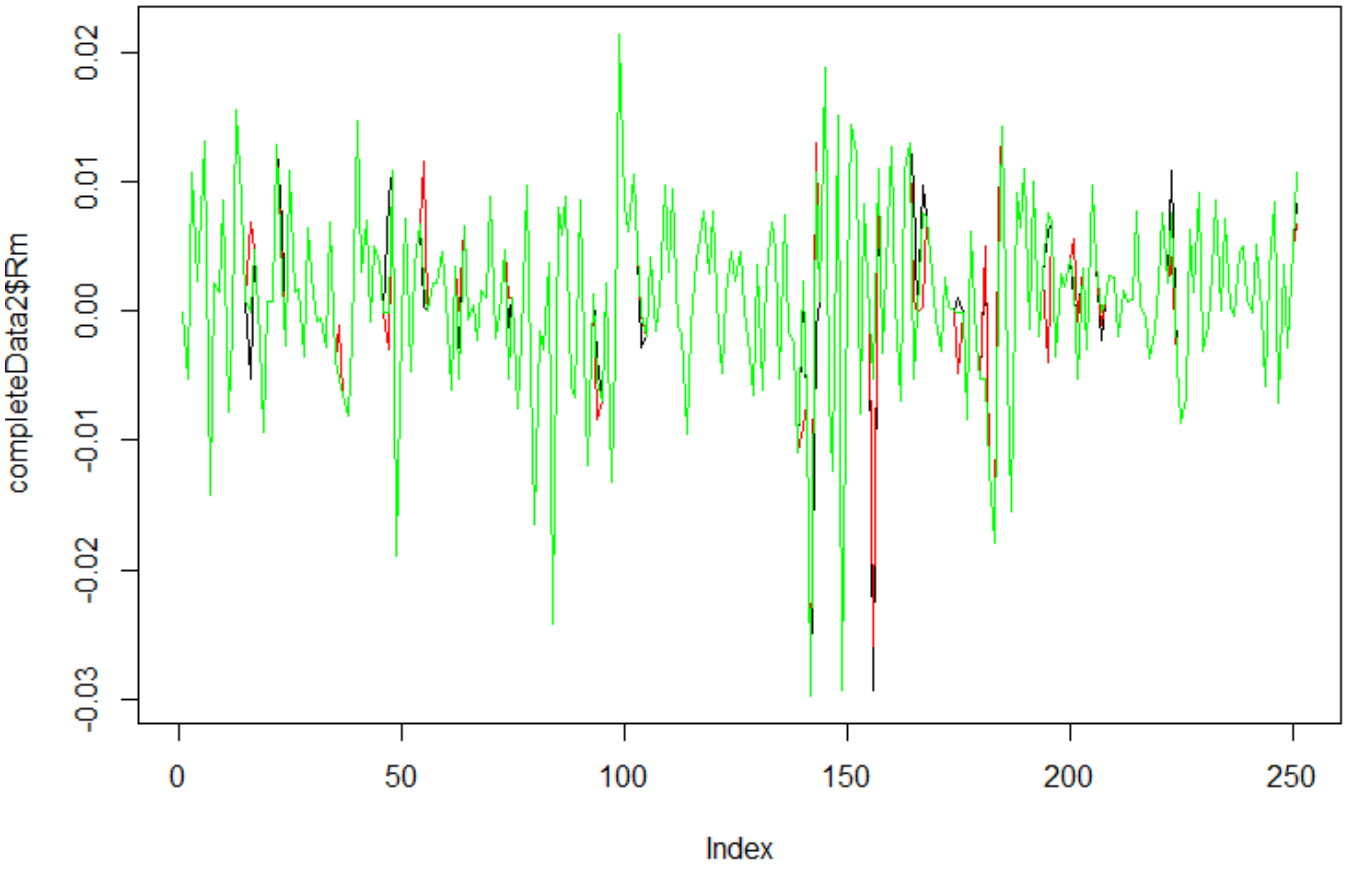
În graficul de mai de jos, completarea valorilor lipsă este evidențiată prin culoarea negru, în tim p ce valorile deja existente, înainte de completare celor nule, prin culoarea roșie. Se observă faptul că valorile estimate se încadrează foarte bine printre cele neestimate, ceea ce indică faptul că această met odă este foarte bună.



2) HMISC

Cel de-al doilea algoritm utilizat în imputarea valorilor lipsă este HMISC). Cu ajutorul funcției aregImpute, am realizat imputarea multiplă: am estimat o regresie de forma: ~Ra1 + Ra2 + Rm și apoi am aplicat PMM pentru completarea valorilor absente.

În graficul de mai de jos, completarea valorilor lipsă prin această metodă este evidențiată prin c uloarea negru, în timp ce valorile deja existente, înainte de completare celor nule, prin culoarea verde . Se observă faptul că și de această data valorile estimate se încadrează foarte bine printre cele neestim ate, ceea ce indică faptul că această metodă este si ea foarte bună.

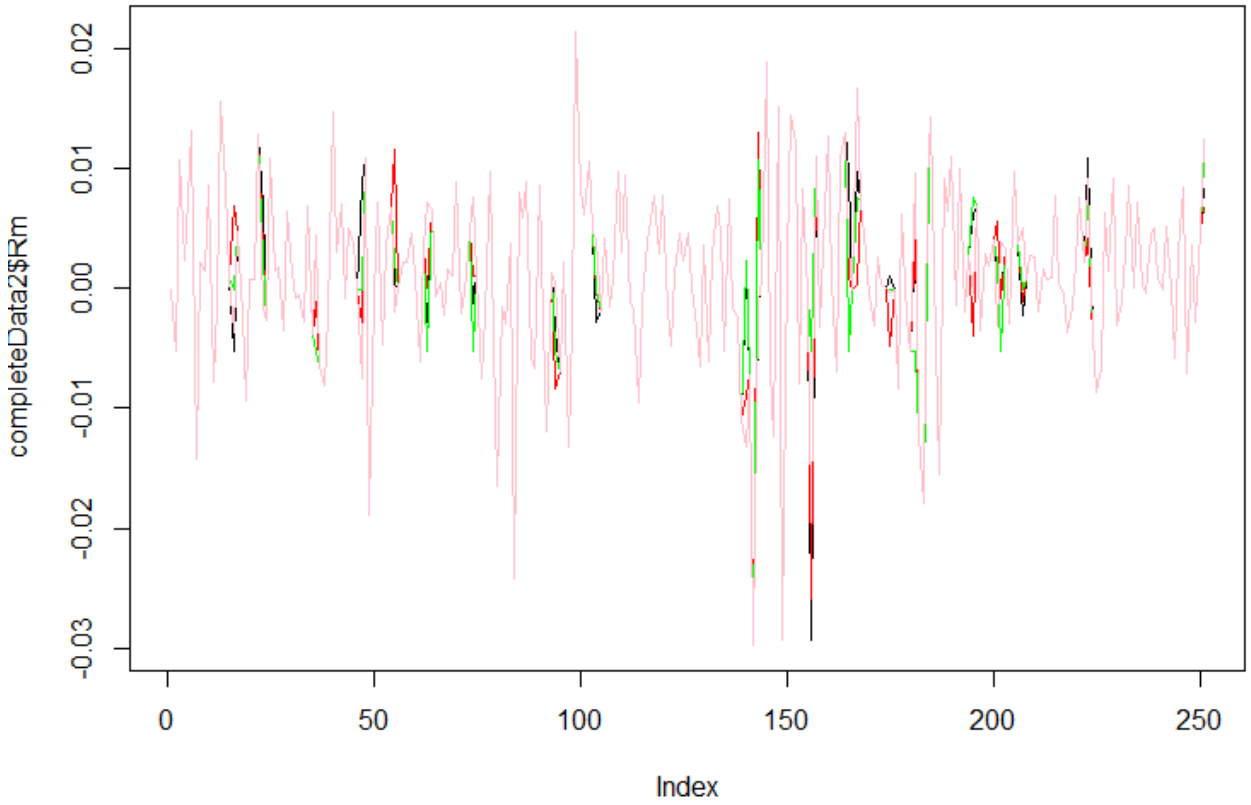


3) AMELIA

Cel de-al treilea algoritm utilizat în imputarea valorilor lipsă este Amelia. Spre deosebire de pri mul, MICE, care realizează imputarea variabilă cu variabilă, Amelia realizează imputarea concomitentă, c onsiderând distirubția normală multivariată.

În graficul de mai de jos, completarea valorilor lipsă prin această metodă este evidențiată prin c uloarea negru, în timp ce valorile deja existente, înainte de completare celor nule, prin culoarea roz. Î n plus: roșu – MICE, verde – HMISC. Se observă faptul că și de această data valorile estimate se înca drează bine printre cele neestimate.

După părerea mea, cel mai bun algoritm a fost MICE, valorile estimate încadrându-se cel mai bi ne cu celalalte.



3. Detectarea outlierilor

Aspecte teoretice

Valorile aberante (engleză outliers) sunt considerate valorile mai mari decât Q3 + 1,5xIQR sau valori mai mici decât Q1 - 1,5xIQR. Într-un boxplot, intervalul IQR este reprezentat grafic printr-un dreptunghi („cutie"). În interiorul său se află mediana reprezentată grafic prin o linie orizontală. Intervalele (Xmin , Q1) și (Q3 , Xmax) sunt reprezentate de câte o linie (engleză whisker =„mustață") trasată în continuarea dreptunghiului**.**

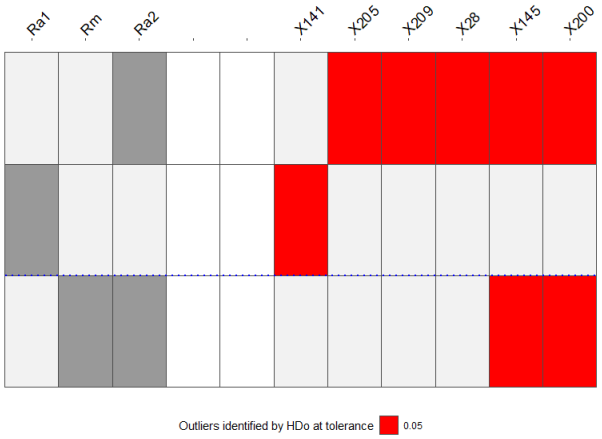
Sursă: ro.wikipedia.org (https://ro.wikipedia.org/wiki/Boxplot**)**

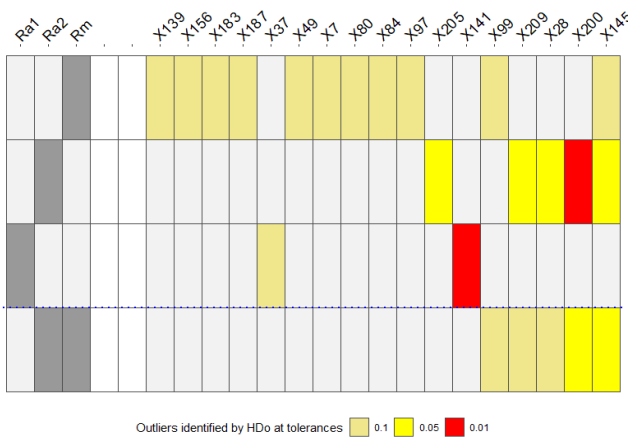
Aplicația propriu-zisă

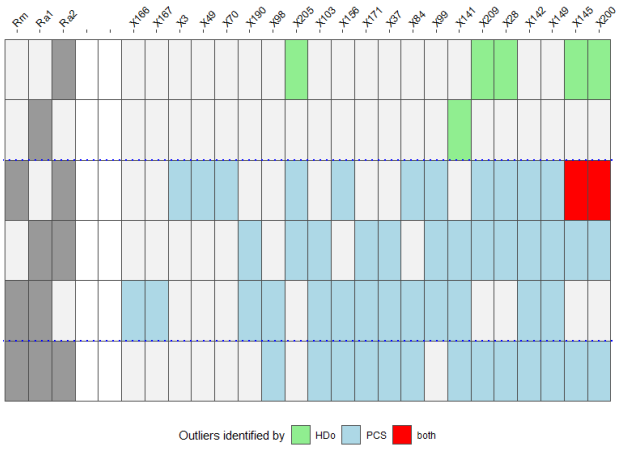
În cazul identificării outlierilor, aceștia trebuie foarte atent analizați întrucât, odată scoși din analiză, vor afecta rezultatele studiului. Astfel spus, unii outlieri sunt acceptați în funcție de contextual din care este privită problema.

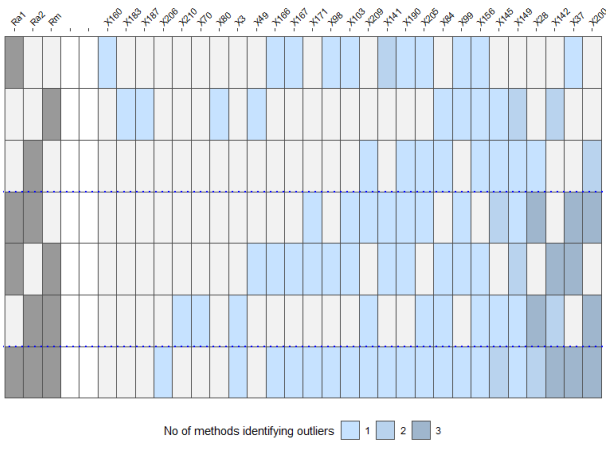
În output-urile de mai jos, în prima figură (stânga sus), am ales 1 nivel de toleranță, de 95%. Pentru cel de-al doilea nivel, am obținut câțiva outlieri (cei cu roșu). Pentru cea de-a doua figură, am optat pentru 3 nivele de toleranță: 90%, 95% și 99%.

Dacă aplicăm două metode de identificare și anume: HDoutliers si FastPCS, identificăm doar doi outlieri (stânga jos), însă la utilizarea tuturor celor 6 metode: HDo, PCS, BAC, adjOut, DDC și MCD. Cu ajutorul unui tabel, am sumarizat numărul de outlieri identificați cu ajutorul a unei metode, folosind două și apoi trei, în diferite combinații.









HDoutliers	FastPCS	mvBACON	adjOutlyingness	DetectDeviatingCells	covMCD
2	5	6	0	0	27

SCRIPT

```
# APLICATIA 4
#path<-"C:\\Users\Marii\Desktop\Data Science"
#msf<-read.table(file.path(path, 'Yahoo.txt'), sep='\t', header=TRUE, dec=".")
#Incarcarea datelor
msf<-read.table(file.choose(),sep='\t', header=TRUE)
msf
View(msf)
#Stergerea inregistrarilor nule
msf<-na.omit(msf)
msf
attach(msf)
names(msf)
boxplot(Rm, horizontal=TRUE)
boxplot(Ra1, horizontal=TRUE)
boxplot(Ra2, horizontal=TRUE)
###Dependenta dintre piata si active
plot(Rm, Ra1)
abline(lm(Ra1~Rm), col="red")
plot(Rm, Ra2)
abline(lm(Ra2~Rm), col="orange")
cor(Rm, Ra1)
cor(Rm, Ra2)
###Testarea stationaritatii
plot(Rm, type="l")
lines(Ra1, col="red")
lines(Ra2, col="orange")
###Estimare model Sharpe
model1<-lm(Ra1~Rm)
summary(model1)
model2<-lm(Ra2~Rm)
summary(model2)
```

```
windows()
layout(matrix(1:6, nrow=2))
plot(model1, which=1:6, ask=FALSE)
windows()
layout(matrix(1:6, nrow=2))
plot(model2, which=1:6, ask=FALSE)
#####Valori lipsa ---> prima metoda mis
#install.packages("mice")
#install.packages("VIM")
#install.packages("missForest")
#install.packages("Amelia")
#install.packages("Hmisc")
#install.packages("mi")
#install.packages("foreign")
library(mice)
library(missForest)
library(VIM)
library(Amelia)
library(Hmisc)
library(mi)
msf.mis<-subset(msf, select=-c(Date, LYB, ALB, S.P.500)) #le scot
msf.mis
msf.mis<-prodNA(msf.mis, noNA=0.1) # generare 10% valori lipsa pentru ca nu erau
summary(msf.mis)
#patternul valorilor lipsa
md.pattern(msf.mis) #--- albastru -> prezente, mov-absente
#sau
#mice_plot<-aggr(msf.mis, col=c('navyblue', 'yellow'), numbers=TRUE, sortvars=TRUE, labels=names(msf.mis),
cex.axis=.7, gap=3, ylab=c("Missing data", "Pattern"))
#mice_plot
#imputarea datelor
#predictive mean matching = pmm
```

```
imputed_msf_mice<-mice(msf.mis, m=5, maxit=50, method='pmm', seed=500)
summary(imputed_msf_mice)
#Check imputed values
imputed_msf_mice$imp$Rm
plot(imputed_msf_mice)
#get complete data(2nd out of 5)
completeData<-mice::complete(imputed_msf_mice, mean(1,2,3,4,5))
completeData
plot(completeData$Rm, type="l")
lines(msf$Rm, col="red")
##################Hmisc ### se efectueaza pentru valori lipsa - metoda 2!!!
# impute with mean value
imputed_Ra1hmean <- with(msf.mis, impute(Ra1, mean))
imputed_Ra1hmean
# impute with random value
imputed_Ra1hrand <- with(msf.mis, impute(Ra1, 'random'))
imputed_Ra1hrand
#similarly you can use min, max, median to impute missing value
#using argImpute
impute_arg <- aregImpute(~ Ra1 + Ra2 + Rm, data = msf.mis, n.impute = 5)
#argImpute() automatically identifies the variable type and treats them accordingly.
impute_arg
#check imputed variable Sepal.Length
impute_arg$imputed$Rm
completeData2 <- impute.transcan(impute_arg, imputation=mean(1,2,3,4,5), data=msf.mis,
list.out=TRUE,pr=FALSE, check=FALSE)
head(completeData2)
plot(completeData2$Rm,type="l")
lines(msf$Rm, col="red")
lines(completeData$Rm, col="green")
#######################Amelia
#specify columns and run amelia
```

```
amelia_fit <- amelia(msf.mis, m=5, p2s=1)
#access imputed outputs
amelia_fit$imputations[[1]]
amelia_fit$imputations[[2]]
amelia_fit$imputations[[3]]
amelia_fit$imputations[[4]]
amelia_fit$imputations[[5]]
#To check a particular column in a data set, use the following commands
amelia_fit$imputations[[5]]$Rm
amelia_fit$imputations[[mean(1,2,3,4,5)]]$Rm
plot(completeData2$Rm,type="l")
lines(msf$Rm, col="red") #adevarate
lines(completeData$Rm, col="green")
lines(amelia_fit$imputations[[mean(1,2,3,4,5)]]$Rm, col="pink")
#export the outputs to csv files
#write.amelia(amelia_fit, file.stem = "imputed_data_set")
#### o3 - vizualizare ouliers
install.packages("OutliersO3")
install.packages("ggplot2")
install.packages("GGally")
install.packages("dplyr")
library(OutliersO3)
library(ggplot2)
library(GGally)
library(dplyr)
msfo<-na.omit(subset(msf, select=-c(Date, LYB, ALB, S.P.500)))
sx <- O3prep(msfo)
O3x1 <- O3plotT(sx)
O3x1$gO3
## 'Analiza outliers cu metoda HDoutliers la 3 nivele diferite de toleranta, $0.1$ (khaki), $0.05$ (yellow), $0.01$ (red).'----
sy <- O3prep(msfo, method="HDo", tols=c(0.1, 0.05, 0.01), boxplotLimits=c(3, 6, 10))
```

```
O3sy <- O3plotT(sy)

O3sy$gO3

## Analiza outliers cu metoda HDoutliers si FastPCS.'----

sz <- O3prep(msfo, method=c("HDo", "PCS"), tolHDo=0.05, tolPCS=0.05)

O3sz <- O3plotM(sz, coltxtsize=10)

O3sz$gO3

## Analiza outliers cu toate cele 6 metode'----

sw <- O3prep(msfo, method=c("HDo", "PCS", "BAC", "adjOut", "DDC", "MCD"), tolHDo=0.01, tolPCS=0.01, tolBAC=0.01, toladj=0.01, tolDDC=0.01, tolMCD=0.01, boxplotLimits=6)

O3sw <- O3plotM(sw, coltxtsize=8)

O3sw$gO3

## tabel metode

cx <- data.frame(nOut=O3sw$nOut)

rownames(cx) <- c("HDoutliers", "FastPCS", "mvBACON", "adjOutlyingness", "DetectDeviatingCells", "covMCD")

cx

knitr::kable(t(cx), row.names=FALSE, caption="Total numbers of outliers identified for one or more of the variable combinations by each of six different methods.")
```