NIEKTORÉ UŽITOČNÉ PRÍKAZY JAZYKA R PRI MANIPULÁCII S ÚDAJMI

Michal Páleš

ÚVOD

Cieľom príspevku je rozšíriť učebnicu [1] o niektoré v praxi veľmi užitočné príkazy, ktoré môže používateľ (aktuár) využiť pri manipulácii s údajmi. Tieto môžu byť nápomocné pre analýzy opísané v predmetnej publikácii a budú zakomponované v jej rozšírenom 2. vydaní (pripráva do tlače s odhadom v roku 2019). Dodatok príspevku sa venuje prostrediu ***RStudia***, ktoré predstavuje populárny editor kódu jazyka R.

Štruktúra príspevku pozostáva zo stručnej formulácie úlohy a jej riešeniu na modelových údajoch, resp. ilustračných príkladoch v jazyku R. Uvádzame celý kód bez vynechania medzi krokov avšak bez obšírneho sprievodného textu. Je dôležité si uvedomiť, že skoro všetky príkazy v jazyku R vieme vykonať viacerými príkazmi, resp. rôznymi postupmi príkazov alebo metódami, z ktorých niektoré sú jednoduchšie iné zložitejšie, niektoré využívajú špecifické knižnicé iné nie. Je teda na autorovi (programátorovi), aký postup zvolí. Tieto štúdie uvádzame najmä ako pomôcku pre študentov k predmetu *Softvérové aplikácie pre aktuárov*.

PRÍKLADY

☐ Rozdelenie textového reťazca spojeného znakom „_“ do samostatných stĺpcov.

```
before <- data.frame(attr = c(1,30,4,6),
type=c('foo_and_bar','foo_and_bar_2'))
```

```
  attr          type
1    1   foo_and_bar
2   30 foo_and_bar_2
3    4   foo_and_bar
4    6 foo_and_bar_2
```

```
out <- strsplit(as.character(before$type),'_and_')
after <- with(before, data.frame(attr = attr))
after <- cbind(after, data.frame(t(sapply(out, `[`))))
names(after)[2:3] <- paste("type", 1:2, sep = "_")
```

```
  attr type_1 type_2
1    1    foo    bar
2   30    foo  bar_2
3    4    foo    bar
4    6    foo  bar_2
```

☐ Rozdelenie po sebe zapísaných údajov v jednom stĺpci do samostatných stĺpcov.

```
x<-read.csv2("D:/Multiple3.csv",header=F); x
```

```
     V1                 8     x1                16  0,94
1       y               9    338                17  0,96
2   28179              10    264                18  0,88
3   11645              11    278                19  0,93
4    7754              12    352                20  0,92
5   15548              13    320                21  0,64
6   11658              14    260
7   18614              15     x2
```

```
x<-as.matrix(x)
x<-as.data.frame(matrix(x,ncol =3,byrow = F))
as.numeric(gsub(".", ",", x))
x<-x[-c(1),]
x # v stĺpci V3 majú byť desatinné "."
```

```
     V1  V2   V3
2 28179 338 0,94
3 11645 264 0,96
4  7754 278 0,88
5 15548 352 0,93
6 11658 320 0,92
7 18614 260 0,64
```

```
y<-as.numeric(as.matrix(x[1]))
x1<-as.numeric(as.matrix(x[2]))
x2<-scan(text=as.matrix(x[3]), dec=",", sep=".")
x2
```

```
 [1] 0.94 0.96 0.88 0.93 0.92 0.64
```

☐ Transformácia textového reťazca na formát *numeric*.

```
# s oddeľovačom tisícov („.“)
var1 <- c("50,0", "72,0", "960,0", "1.920,0", "50,0", "50,0",
"960,0")
```

```
[1] "50,0"    "72,0"    "960,0"   "1.920,0" "50,0"    "50,0"    "960,0"
```

```
var2<-as.numeric(gsub(",", ".", gsub("\\.", "", var1)))
```

```
[1]   50   72  960 1920   50   50  960
```

```
# bez oddeľovača tisícov
var1 <- c("50,0", "72,0", "960,0", "1920,0", "50,0", "50,0",
"960,0")

var2<-scan(text=var1, dec=",", sep=".")
```

☐ Pridanie znakov do reťazca.

```
library(stringi)
tract<-c(1,11,101,1001,10001,100001)
stri_pad_right(tract, 6, "0")
```

```
[1] 100000 110000 101000 100100 100010 100001
```

☐ Nastavenie desatinných miest čísiel v tabuľke.

```
data.frame(a=c(20.222,50),b=c(8,4.04),c=c(3.1,25.2))
```

```
       a    b    c
1 20.222 8.00  3.1
2 50.000 4.04 25.2
```

```
a<-sprintf("%.3f",c(20.222,50))
b<-sprintf("%.3f",c(8,4.04))
c<-sprintf("%06.3f",c(3.1,25.2))
data.frame(a,b,c)
```

```
       a     b      c
1 20.222 8.000 03.100
2 50.000 4.040 25.200
```

☐ Výpočet súčtu (sumy) početností v kontigenčnej tabuľke.

```
N<-rpois(23589,0.14422)
as.data.frame(table(N))
```

```
  N  Freq
1 0 20442
2 1  2943
3 2   196
4 3     8
```

```
sum(as.data.frame(table(N))$Freq)
```

```
[1] 23589
```

☐ Práca s databázou údajov (oddelených čiarkami).

	A	B	C	D
1	POHLAVIE,VEK,SITUACIA,VYSKA			
2	M,22,C7,1875			
3	F,22,C11,3992.12			
4	M,25,C1,187			
5	F,32,M2,92.11			
6	M,48,F6,3897			
7	F,33,C7,87.112			

```
data<-read.csv("D:/Separ.csv")
```

```
  POHLAVIE VEK SITUACIA    VYSKA
1        M  22       C7 1875.000
2        F  22      C11 3992.120
3        M  25       C1  187.000
4        F  32       M2   92.110
5        M  48       F6 3897.000
6        F  33       C7   87.112
```

```
subset(data,VEK==22,c(VYSKA,SITUACIA))
```

```
    VYSKA SITUACIA
1 1875.00       C7
2 3992.12      C11
```

```
subset(data,SITUACIA=="C7",c(VEK,VYSKA))
```

```
  VEK    VYSKA
1  22 1875.000
6  33   87.112
```

```
attach(data); v<-VYSKA
```

```
[1] 1875.000 3992.120  187.000   92.110 3897.000   87.112
```

☐ Doplnenie špecifického súčtového riadku do tabuľky údajov.

```
data<-read.csv("D:/Separ.csv")
data<-as.data.frame(data)
rbind(data,  data.frame(POHLAVIE='Spolu',  VEK="-",  SITUACIA="-
", VYSKA = sum(data[4])))
```

```
  POHLAVIE VEK SITUACIA     VYSKA
1        M  22       C7  1875.000
2        F  22      C11  3992.120
3        M  25       C1   187.000
4        F  32       M2    92.110
5        M  48       F6  3897.000
6        F  33       C7    87.112
7    Spolu   -        - 10130.342
```

☐ Spojenie dvoch matíc.

```
A<-matrix(6,ncol=2,nrow=1)
```

```
     [,1] [,2]
[1,]    6    6
```

```
B<-matrix(5,ncol=2,nrow=2)
```

```
     [,1] [,2]
[1,]    5    5
[2,]    5    5
```

```
merge(A,B,all=T)
```

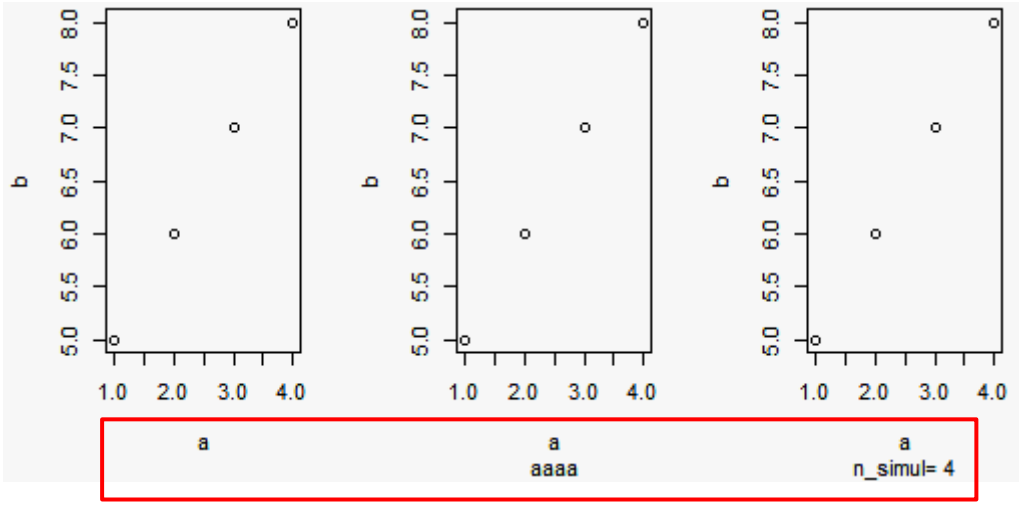
```
  V1 V2
1  5  5
2  5  5
3  6  6
```

```
merge(A,B,all=T,sort=F)
```

```
  V1 V2
1  6  6
2  5  5
3  5  5
```

☐ Vloženie textu a hodnoty definovanej premennej do poznámky (titulku) grafu.

```
a<-c(1:4)
b<-c(5:8)
par(mfrow=c(1,3))
plot(a,b)
plot(a,b,sub="aaaa")
n<-length(a);plot(a,b,sub=paste("n_simul=",n))
```



☐ Filtrovanie v údajoch uložených do stĺpcov.

```
set.seed(100)
(P<-replicate(5,round(rnorm(10,0,1),2)))
```

```
      [,1]  [,2]  [,3]  [,4]  [,5]
 [1,] -0.50  0.09 -0.44 -0.09 -0.10
 [2,]  0.13  0.10  0.76  1.76  1.40
 [3,] -0.08 -0.20  0.26 -0.14 -1.78
 [4,]  0.89  0.74  0.77 -0.11  0.62
 [5,]  0.12  0.12 -0.81 -0.69 -0.52
 [6,]  0.32 -0.03 -0.44 -0.22  1.32
```

```
 [7,] -0.58 -0.39 -0.72  0.18 -0.36
 [8,]  0.71  0.51  0.23  0.42  1.32
 [9,] -0.83 -0.91 -1.16  1.07  0.04
[10,] -0.36  2.31  0.25  0.97 -1.88
```

```
subset(P[,1],P[,1]>=0) # iba z 1. stĺpca
```

```
[1] 0.13 0.89 0.12 0.32 0.71
```

```
P[P[,1] >= 0,] # ako filter podľa prvého
```

```
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.13  0.10  0.76  1.76  1.40
[2,] 0.89  0.74  0.77 -0.11  0.62
[3,] 0.12  0.12 -0.81 -0.69 -0.52
[4,] 0.32 -0.03 -0.44 -0.22  1.32
[5,] 0.71  0.51  0.23  0.42  1.32
```

```
P[P[,1] >= 0] # len údaje ako filter podľa 1. stĺpca
```

```
[1]  0.13  0.89  0.12  0.32  0.71  0.10  0.74  0.12 -0.03  0.51  0.76  0.77
[13] -0.81 -0.44  0.23  1.76 -0.11 -0.69 -0.22  0.42  1.40  0.62 -0.52  1.32
[25]  1.32
```

```
subset(P,P[,1]>=0&P[,2]>=0) # ako filter, kde 1. a 2. stĺpec má kladné hodnoty
```

```
     [,1] [,2]  [,3]  [,4]  [,5]
[1,] 0.13 0.10  0.76  1.76  1.40
[2,] 0.89 0.74  0.77 -0.11  0.62
[3,] 0.12 0.12 -0.81 -0.69 -0.52
[4,] 0.71 0.51  0.23  0.42  1.32
```

```
subset(P,P[,1]>=0|P[,2]>=0) # ako filter kde 1. a 2. stĺpec nemá súčasne záporné hodnoty
```

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] -0.50  0.09 -0.44 -0.09 -0.10
[2,]  0.13  0.10  0.76  1.76  1.40
[3,]  0.89  0.74  0.77 -0.11  0.62
[4,]  0.12  0.12 -0.81 -0.69 -0.52
[5,]  0.32 -0.03 -0.44 -0.22  1.32
[6,]  0.71  0.51  0.23  0.42  1.32
[7,] -0.36  2.31  0.25  0.97 -1.88
```

```
# subset(P,P>=0) # nie je možné využiť
```

```
Pr<-sapply(P,function(x)subset(x,x>0));
as.vector(na.omit(as.numeric(Pr)))
```

```
[1] 0.13 0.89 0.12 0.32 0.71 0.09 0.10 0.74 0.12 0.51 2.31 0.76 0.26 0.77
[15] 0.23 0.25 1.76 0.18 0.42 1.07 0.97 1.40 0.62 1.32 1.32 0.04
```

☐ Nahradenie konkrétnej hodnoty v údajoch uložených do stĺpcov.

```
df <- replicate(20, sample(c(1,2,3,99), 4))
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    2    1    2    1    1    2    2    1    1    99
[2,]    3    2    3    3    2    3    3    2   99     3
[3,]   99   99   99    2   99   99   99   99    3     1
[4,]    1    3    1   99    3    1    1    3    2     2
```

```
dfc <- df
dfc[dfc == 99] <- NA
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    2    1    2    1    1    2    2    1    1    NA
[2,]    3    2    3    3    2    3    3    2   NA     3
[3,]   NA   NA   NA    2   NA   NA   NA   NA    3     1
[4,]    1    3    1   NA    3    1    1    3    2     2
```

```
colMeans(dfc, na.rm = TRUE) # priemer v hodnôt v stĺpcoch
```

```
[1] 2 2 2 2 2 2 2 2 2 2
```

☐ Usporiadanie údajov do matice požadovaného tvaru (typu).

```
set.seed(1)
trunc(rnorm(15,5,2))
```

```
[1] 3 5 3 8 5 3 5 6 6 4 8 5 3 0 7
```

```
set.seed(1)
as.matrix(trunc(rnorm(15,5,2)))
```

```
      [,1]
 [1,]    3
 [2,]    5
 [3,]    3
 [4,]    8
 [5,]    5
 [6,]    3
 [7,]    5
 [8,]    6
 [9,]    6
[10,]    4
[11,]    8
[12,]    5
[13,]    3
[14,]    0
[15,]    7
```

```
set.seed(1)
matrix(trunc(rnorm(15,5,2)),5,3)
```

```
     [,1] [,2] [,3]
[1,]    3    3    8
[2,]    5    5    5
[3,]    3    6    3
[4,]    8    6    0
[5,]    5    4    7
```

```
set.seed(1)
C<-matrix(trunc(rnorm(15,5,2)),6,3)
C[16:length(C)] <- NA; C
```

```
     [,1] [,2] [,3]
[1,]    3    5    3
[2,]    5    6    0
[3,]    3    6    7
[4,]    8    4   NA
[5,]    5    8   NA
[6,]    3    5   NA
```

☐ Vymazanie riadkov s hodnotou „NA“ v databáze.

```
data<-read.csv2("D:/diabetes2.csv",header=T)
```

```
   pocet_umrti vek pohlavie populacia       miera
1            3   A      muz   1141100 0.000002630
2            0   B      muz    485571 0.000000000
3           12   C      muz    504312 0.000023800
4           25   D      muz    447315 0.000055900
5           61   E      muz    330902 0.000184345
6          130   F      muz    226403 0.000574197
7          192   G      muz    130527 0.001470960
8          102   H      muz     29785 0.003424543
9            2   A     zena   1086408 0.000001840
10           1   B     zena    489948 0.000002040
11           3   C     zena    504030 0.000005950
12          11   D     zena    445763 0.000024700
13          30   E     zena        NA 0.000092700
14          63   F     zena        NA 0.000260883
15         174   G     zena        NA 0.000968356
16         159   H     zena        NA 0.002365966
```

```
attach(data)
na.omit(populacia) # len predmetný stĺpec
```

```
[1] 1141100  485571  504312  447315  330902  226403  130527   29785 1086408
[10]  489948  504030  445763
```

```
# alternatíva: populacia<-populacia[!is.na(populacia)]
```

```
na.omit(data) # celý riadok s NA
```

```
   pocet_umrti vek pohlavie populacia       miera
1            3   A      muz   1141100 0.000002630
2            0   B      muz    485571 0.000000000
3           12   C      muz    504312 0.000023800
4           25   D      muz    447315 0.000055900
5           61   E      muz    330902 0.000184345
6          130   F      muz    226403 0.000574197
7          192   G      muz    130527 0.001470960
8          102   H      muz     29785 0.003424543
9            2   A     zena   1086408 0.000001840
10           1   B     zena    489948 0.000002040
11           3   C     zena    504030 0.000005950
12          11   D     zena    445763 0.000024700
```

☐ Generovanie špecifického reťazca (napr. pre kódovanie, heslá, ID v tabuľkách, ...).

```
library(stringi)
n<-10
stri_paste(
stri_rand_strings(n, 2, '[A-Z]'),
stri_rand_strings(n, 5, '[0-9]'))
```

```
[1] "GK34621" "RA34676" "HM22974" "WH61567" "TY67913" "LJ87631" "VM90108"
[8] "NR25025" "YL49169" "VI29719"
```

☐ Generovanie špecifického reťazca (*text „ID“ + 5 ciferné kladné celé č.*) do tabuľky.

```
library(stringi)
n<-1000
x<-stri_paste(
stri_paste("ID"),
stri_rand_strings(n, 5, '[0-9]'))
DF<-
data.frame(cislo_zmluvy=unique(x),poistne=sample(1:500,length(
unique(x)),replace=T)) # špecificky možné upraviť opakovanie
head(DF)
```

```
  cislo_zmluvy poistne
1      ID94086      95
2      ID29972      67
3      ID38370       2
4      ID46030      41
5      ID75741      30
6      ID29582      10
```

☐ Pridanie marginálnych súčtov do tabuľky údajov.

```
A <- c("a","b","c","d"); B <- c(33,19,27,48)
C <- c(1.8,1,2.5,2); D <- c(82,91,58,81)
tab <- data.frame(A,B,C,D, stringsAsFactors=F)
tab <- transform(tab,E=B+C)
tab <- transform(tab,SPOLU=rowSums(tab[1:4,2:5]))
rbind(tab, c("SPOLU", colSums(tab[,2:6])))
```

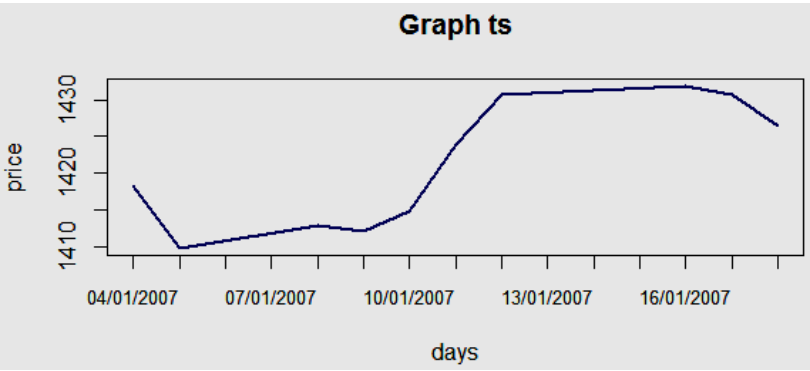
```
      A   B   C   D     E SPOLU
1     a  33 1.8  82  34.8 151.6
2     b  19   1  91    20   131
3     c  27 2.5  58  29.5   117
4     d  48   2  81    50   181
5 SPOLU 127 7.3 312 134.3 580.6
```

☐ Pridanie časových údajov na os *x* do grafu časového radu.

```
ts<-read.csv2("D:/ts.csv",header=T)
```

```
        Date Date_code   Price
1   4.1.2007     39086 1418.34
2   5.1.2007     39087 1409.71
3   8.1.2007     39090 1412.84
4   9.1.2007     39091 1412.11
5  10.1.2007     39092 1414.85
6  11.1.2007     39093 1423.82
7  12.1.2007     39094 1430.73
8  16.1.2007     39098 1431.90
9  17.1.2007     39099 1430.62
10 18.1.2007     39100 1426.37
```

```
d<-as.matrix(ts[,2])
ts<-as.matrix(ts[,3])
d<-as.Date(d, origin = "1899-12-30") # is MS Excel date code
plot(d,ts,xlab="days",ylab="price",main="Graph
ts",type="l",col="darkblue",lwd=2,xaxt="n")
axis.Date(1,at=seq(d[1],d[length(ts)],"days"),
format="%d/%m/%Y",cex.axis=0.8)
```



☐ Vymazanie riadkov s hodnotou „NA“ v databáze, podľa špecifických požiadaviek.

```
df1<-read.csv2("D:/Missing.csv")
```

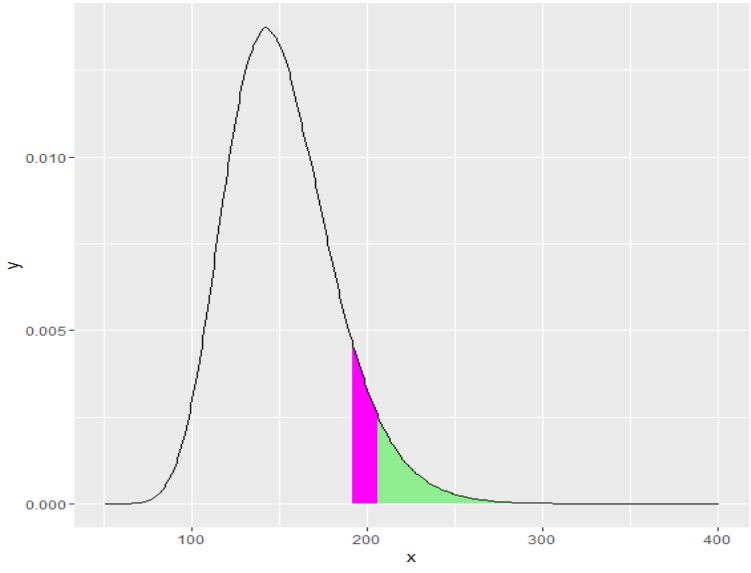
```
Name Lacation profits loss sales address revenue stock
1   AA   London      20   30     2 Lheigts      54    45
2   BB   Boston      NA   NA    NA    KicK      NA    NA
3   CC   Mumbai      NA    2    NA     New      43    NA
```

```
df1[rowSums(!is.na(df1[c(3:5,7:8)]))!=0,]
# vymazať riadky, kde sú všetky ukazovatele NA
```

```
  Name Lacation profits loss sales address revenue stock
1   AA   London      20   30     2 Lheigts      54    45
3   CC   Mumbai      NA    2    NA     New      43    NA
```

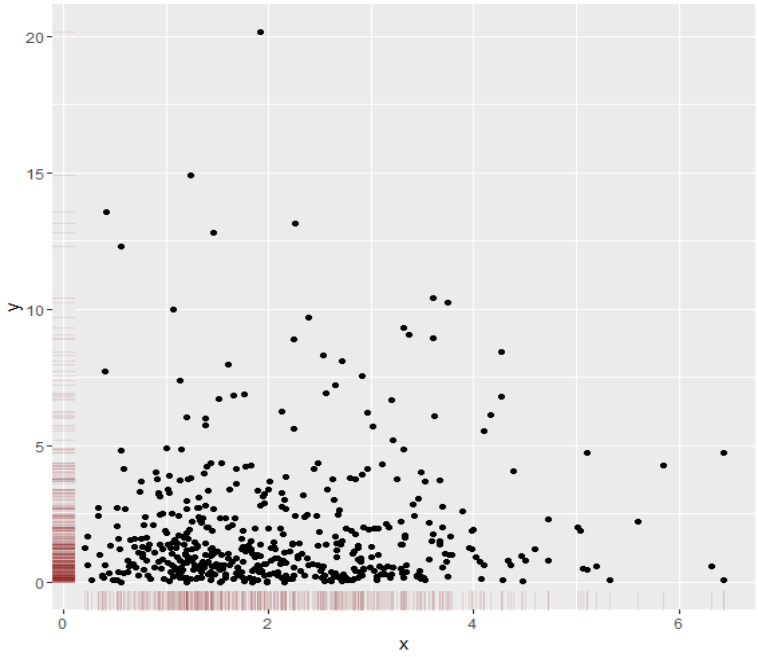
☐ Znázornenie plochy kvantilov na grafe hustoty.

```
library(ggplot2)
X<-rlnorm(1000000,5,0.2)
k1<-quantile(X,0.9)
k2<-quantile(X,0.95)
myd = data.frame(xvar=X,yvar=X)
    xd <- data.frame(density(myd$xvar)[c("x", "y")])
    p <- ggplot(xd, aes(x, y)) +

      geom_area(data = subset(xd, x > k1), fill = "magenta") +
      geom_area(data = subset(xd, x > k2), fill = "lightgreen") +

      geom_line()

p
```



☐ Zobrazenie fragmentu histogramu na osiach grafu závislostí (*scatter plot*).

```
library(ggplot2)
library(actuar)
x<-rgamma(500,3,1.5)
y<-rpareto(500,4,6)
xy<-data.frame(x,y)
qplot(x,y, data=xy)  +
scale_x_continuous(limits=c(min(x),max(x))) +
scale_y_continuous(limits=c(min(y),max(y))) +
geom_rug(col=rgb(.5,0,0,alpha=.1))
```



☐ Formátovanie tabuľky.

```
library(formattable)
df1<-read.csv2("D:/Missing.csv")
formattable(df1)
```

Name	Lacation	profits	loss	sales	address	revenue	stock
AA	London	20	30	2	Lheigts	54	45
BB	Boston	NA	NA	NA	KicK	NA	NA
CC	Mumbai	NA	2	NA	New	43	NA

```
library(formattable) # zdroj [2]
DF <- data.frame(Ticker=c("", "", "", "IBM", "AAPL", "MSFT"),
                 Name=c("Dow Jones", "S&P 500", "Technology",
                        "IBM", "Apple", "Microsoft"),
                 Value=accounting(c(15988.08, 1880.33, NA,
                                    130.00, 97.05, 50.99)),
                 Change=percent(c(-0.0239, -0.0216, 0.021,
                                  -0.0219, -0.0248, -0.0399)))
formattable(DF, list(
  Name=formatter(
    "span",
    style = x ~ ifelse(x == "Technology",
                       style(font.weight = "bold"), NA)),
  Value = color_tile("white", "orange"),
  Change = formatter(
    "span",
    style = x ~ style(color = ifelse(x < 0 , "red", "green")),
    x ~ icontext(ifelse(x < 0, "arrow-down", "arrow-up"), x)))
)
```

Ticker	Name	Value	Change
	Dow Jones	15,988.08	-2.39%
	S&P 500	1,880.33	-2.16%
	Technology	NA	2.10%
IBM	IBM	130.00	-2.19%
AAPL	Apple	97.05	-2.48%
MSFT	Microsoft	50.99	-3.99%

```
library(formattable) # zdroj [2]
df <- data.frame(
  id = 1:10,
  name = c("Bob", "Ashley", "James", "David", "Jenny",
           "Hans", "Leo", "John", "Emily", "Lee"),
  age = c(48, 47, 40, 28, 29, 29, 27, 27, 31, 30),
  test1_score = c(18.9, 19.5, 19.6, 12.9, 11.1, 7.3, 4.3, 3.9,
2.5, 1.6),
  test2_score = c(9.1, 9.1, 9.2, 11.1, 13.9, 14.5, 19.2, 19.3,
19.1, 18.8),
  stringsAsFactors = FALSE)

formattable(df, list(
  age = color_tile("white", "orange"),
  test1_score = color_bar("pink",fun = "proportion"),
  test2_score = color_bar("pink")
))
```

id	name	age	test1_score	test2_score
1	Bob	48	18.9	9.1
2	Ashley	47	19.5	9.1
3	James	40	19.6	9.2
4	David	28	12.9	11.1
5	Jenny	29	11.1	13.9
6	Hans	29	7.3	14.5
7	Leo	27	4.3	19.2
8	John	27	3.9	19.3
9	Emily	31	2.5	19.1
10	Lee	30	1.6	18.8

☐ Duplikovanie hodnôt pre prázdne hodnoty v tabuľke údajov.

```
df<-data.frame("event" = c(1,NA,2,NA,3,NA,5),"other" = 1:7)
```

```
  event other
1     1     1
2    NA     2
3     2     3
4    NA     4
5     3     5
6    NA     6
7     5     7
```

```
temp <- rle(df$event)
temp$values[is.na(temp$values)] <-
temp$values[which(is.na(temp$values))-1]
df$event <- inverse.rle(temp)
df # kód platí len za obmedzených podmienok
```

```
  event other
1     1     1
2     1     2
3     2     3
4     2     4
5     3     5
6     3     6
7     5     7
```

```
# využitie knižnice purrr (pre číselné hodnoty)

library(purrr)
df$event <- accumulate(.x = df$event, .f = function(x, y) {
if(is.na(y)) x else y })
df
```

```
# využitie knižnice purrr (pre text)

library(purrr)
df <- data.frame("event" = c("Adam", NA, NA, NA, "Peter", NA,
NA), "other" = 1:7)
df
```

```
  event other
1  Adam     1
2    NA     2
3    NA     3
4    NA     4
5 Peter     5
6    NA     6
7    NA     7
```

```
df$event <- accumulate(.x = df$event, .f = function(x, y) {
if(is.na(y)) x else y })
df
```

```
  event other
1  Adam     1
2  Adam     2
3  Adam     3
4  Adam     4
5 Peter     5
6 Peter     6
7 Peter     7
```

☐ Prepísanie konkrétnych prvkov matice podľa zadefinovanej podmienky.

```
set.seed(100)
df <-
data.frame(a=round(rnorm(1000),2),b=round(rnorm(1000),2))
head(df)
```

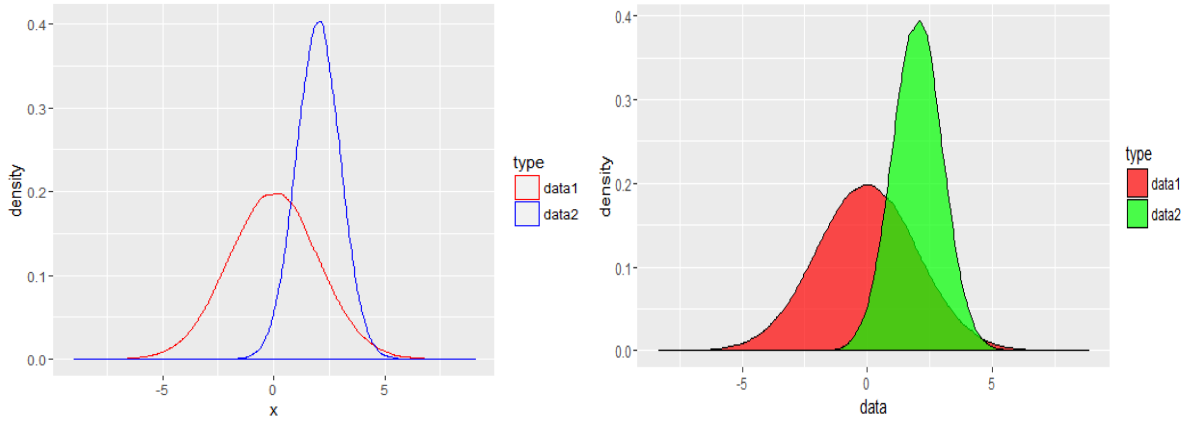
```
      a    b
1 -0.50 1.10
2  0.13 1.18
3 -0.08 0.59
4  0.89 1.08
5  0.12 1.14
6  0.32 0.76
```

```
m <- as.matrix(df)
m[m<0] <- 0
df <- as.data.frame(m)
head(df)
```

```
     a    b
1 0.00 1.10
2 0.13 1.18
3 0.00 0.59
4 0.89 1.08
5 0.12 1.14
6 0.32 0.76
```

☐ Zobrazenie dvoch kriviek hustoty na jednom grafe s využitím knižnice `ggplot2`.

```
library(ggplot2)
data1 <- rnorm(100000, 0, 2)
data2 <- rnorm(100000, 2, 1)
plot1  <- ggplot(
   rbind(
     data.frame(x=data1,type="data1"),
     data.frame(x=data2,type="data2")),
   aes(x=x,color=type))+
   geom_density()+
   scale_color_manual(
     values = c(
       "data1" = "red",
       "data2" = "blue"))
plot1

# vo formáte s výplňou (grafický výstup vpravo)

library(ggplot2)
data1 <- rnorm(100000, 0, 2)
data2 <- rnorm(100000, 2, 1)
plot2  <- ggplot(
   rbind(
     data.frame(x=data1,type="data1"),
     data.frame(x=data2,type="data2")),
   aes(x=x,fill=type))+
   geom_density(alpha=0.7)+
   labs(x="data")+
      scale_fill_manual(
     values = c(
       "data1" = "red",
       "data2" = "green"))
plot2
```



☐ Pridanie prázdnej hodnoty do stĺpca tabuľky.

```
x<-c(1,2,5,10,12)
y<-diff(x) # výpočet 1. diferencie
```

```
[1] 1 3 5 2
```

```
data.frame(x=x,difx=y) # počet riadkov x a difx sa nerovná
y<-c("",diff(x)) # y<-c("",c(1,3,5,2))
data.frame(x=x,difx=y)
```

```
   x difx
1  1
2  2    1
3  5    3
4 10    5
5 12    2
```

```
data.frame(x=x,difx=as.numeric(y))
```

```
   x difx
1  1   NA
2  2    1
3  5    3
4 10    5
5 12    2
```

☐ Tvorba jednoduchého cyklu (*loop*) `for` pre výpočet 1. diferencie.

```
y<-NULL
x<-c(1,2,5,10,12)
for(i in 1:length(x))
{
y[i]<-x[i]-x[i-1]
}
```

```
data.frame(x=x,difx=y)
```

```
  x  difx
1  1    NA
2  2     1
3  5     3
4 10     5
5 12     2
```

```
na.omit(data.frame(x=x,difx=y)) # odstránenie riadku s NA
```

```
   x difx
2  2     1
3  5     3
4 10     5
5 12     2
```

```
df<-data.frame(x=x,difx=y) # zmena hodnoty NA na "-"
dfc <- df
dfc[is.na(dfc)] <- "-"
dfc
```

```
   x difx
1  1     -
2  2     1
3  5     3
4 10     5
5 12     2
```

☐ Využitie funkcie `paste`.

```
paste("A", 1:6, sep = ":")
```

```
[1] "A:1" "A:2" "A:3" "A:4" "A:5" "A:6"
```

```
paste("Today is", date())
```

```
"Today is Wed Jul 26 16:18:09 2017"
```

☐ Zobrazenie, resp. odstránenie riadkov v tabuľke údajov.

```
df<-data.frame("cl"=paste("CL",c(1:5)),"we"=c(0.5,1,1,0.5,2))
```

```
   cl  we
1 CL 1 0.5
2 CL 2 1.0
3 CL 3 1.0
4 CL 4 0.5
5 CL 5 2.0
```

```
# zobrazenie prvých dvoch riadkov tabuľky

head(df,2)
```

```
    cl  we
1 CL 1 0.5
2 CL 2 1.0
```

```
# odstránenie posledných dvoch riadkov tabuľky

head(df,-2)
```

```
    cl  we
1 CL 1 0.5
2 CL 2 1.0
3 CL 3 1.0
```

```
# vymazanie 2. a 4. riadka

df[-c(2,4),]
```

```
    cl  we
1 CL 1 0.5
3 CL 3 1.0
5 CL 5 2.0
```

☐ Vymazanie duplicitných riadkov tabuľky v závislosti na jednej hodnote.

```
df<-
data.frame("ID"=c(1,1,2,2,2),"Test"=c("POS","POS","NEG","NEG",
"NEG"),"Year"=c(2015,2014,2013,2012,2008))
```

```
  ID Test Year
1  1  POS 2015
2  1  POS 2014
3  2  NEG 2013
4  2  NEG 2012
5  2  NEG 2008
```

```
df[!duplicated(df$ID),]
```

```
  ID Test Year
1  1  POS 2015
3  2  NEG 2013
```

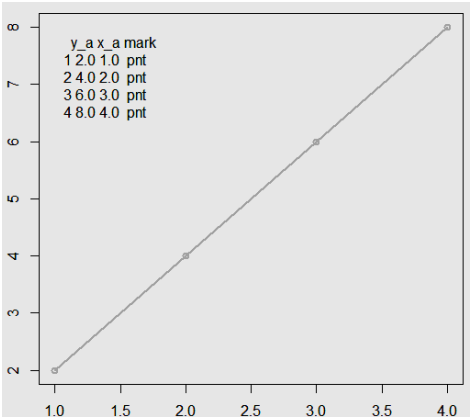
☐ Vloženie výstupu do grafu.

```
x<-c(1,2,3,4)
y<-c(2,4,6,8)
```

```
df<-
data.frame("y_a"=sprintf("%.1f",y),"x_a"=sprintf("%.1f",x),"ma
rk"=c("pnt","pnt","pnt","pnt"))
```

```
  y_a x_a mark
1 2.0 1.0  pnt
2 4.0 2.0  pnt
3 6.0 3.0  pnt
4 8.0 4.0  pnt
```

```
plot(x,y,type="o",lwd=2,col="grey",xlab="",ylab="")
text(1,7, paste(capture.output(df),collapse='\n'), pos=4)
```



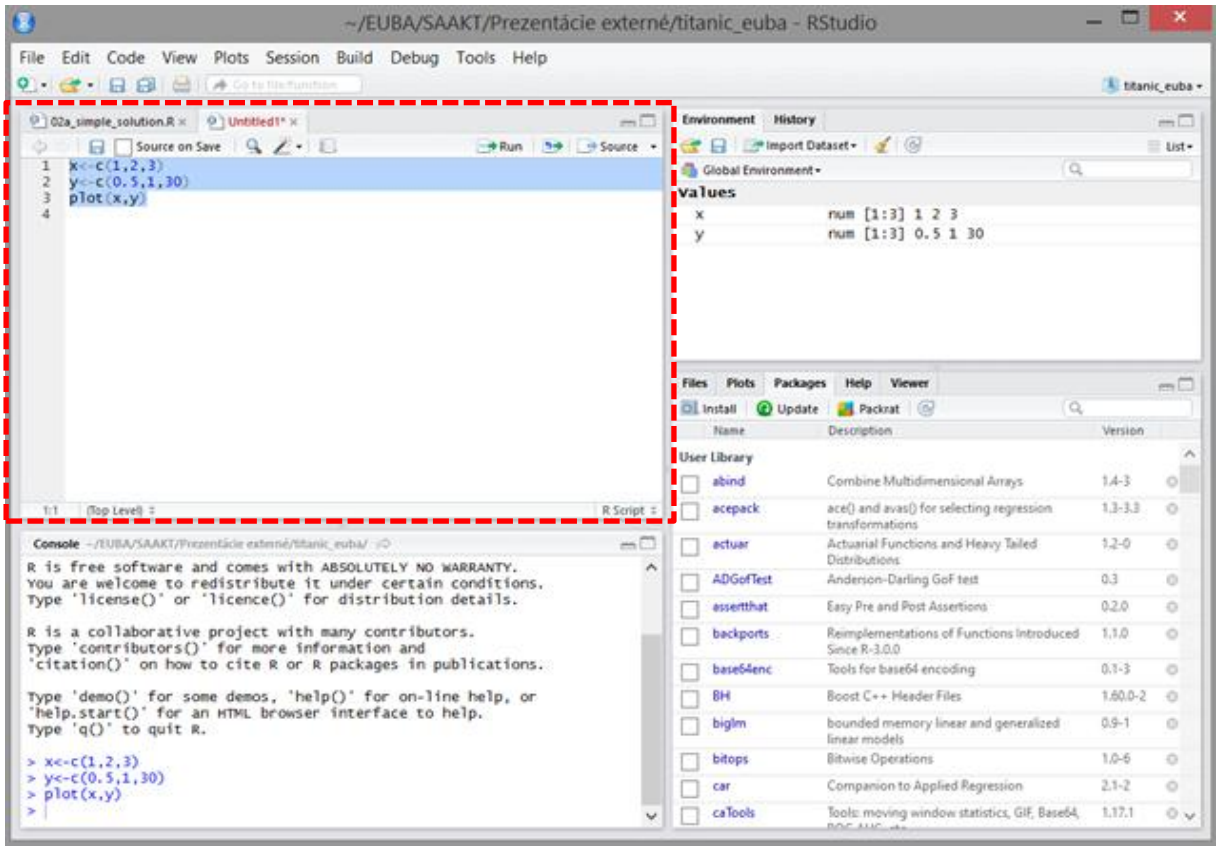
DODATOK

RStudio je bezplatné a open source integrované vývojárske prostredie (IDE) pre jazyk R. Predstavuje súbor nástrojov, ktoré zjednodušujú, zlepšujú a rozširujú základné možnosti jazyka R. Viac informácií používateľ nájde na: https://www.rstudio.com/. Medzi hlavné výhody oproti štandardnému rozhraniu R patrí napríklad:

- možnosť tvorby a priameho spúšťania skriptov, či už ako celku alebo po častiach,
- integrovaná pomoc pri jednotlivých príkazoch, bez nutnosti zobrazenia pomoci v externom okne prehliadača,
- zlepšenie grafických možností,
- lepší ladenie kódu za pomoci vyznačenia chybnej syntaxe, a pod.

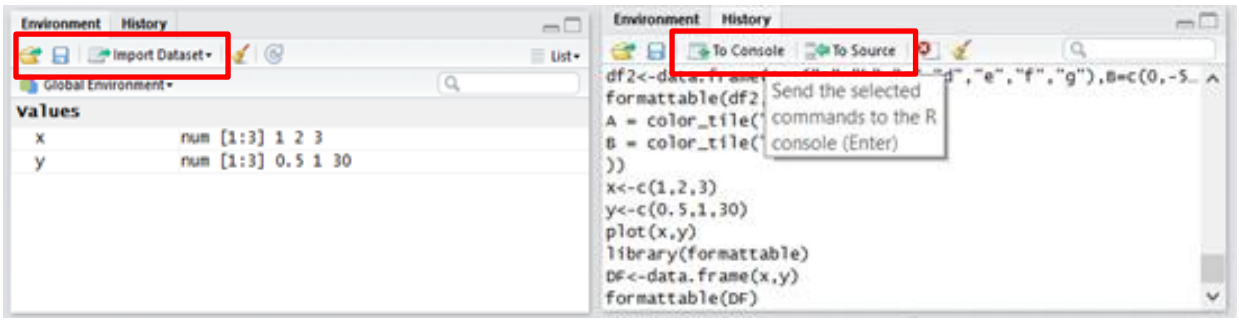
Vo všeobecnosti a zjednodušene môžeme užívateľské prostredie *RStudia* rozdeliť do 4 kvadrantov (obr. 1 tohto dodatku):

- **Údaje/História (1.)**
- **Skript (2.)**
- **Konzola (3.)**
- **Súbory/Grafy/Knižnice (Balíčky)/Nápoveda/Náhľad (4.)**



Obr. D1

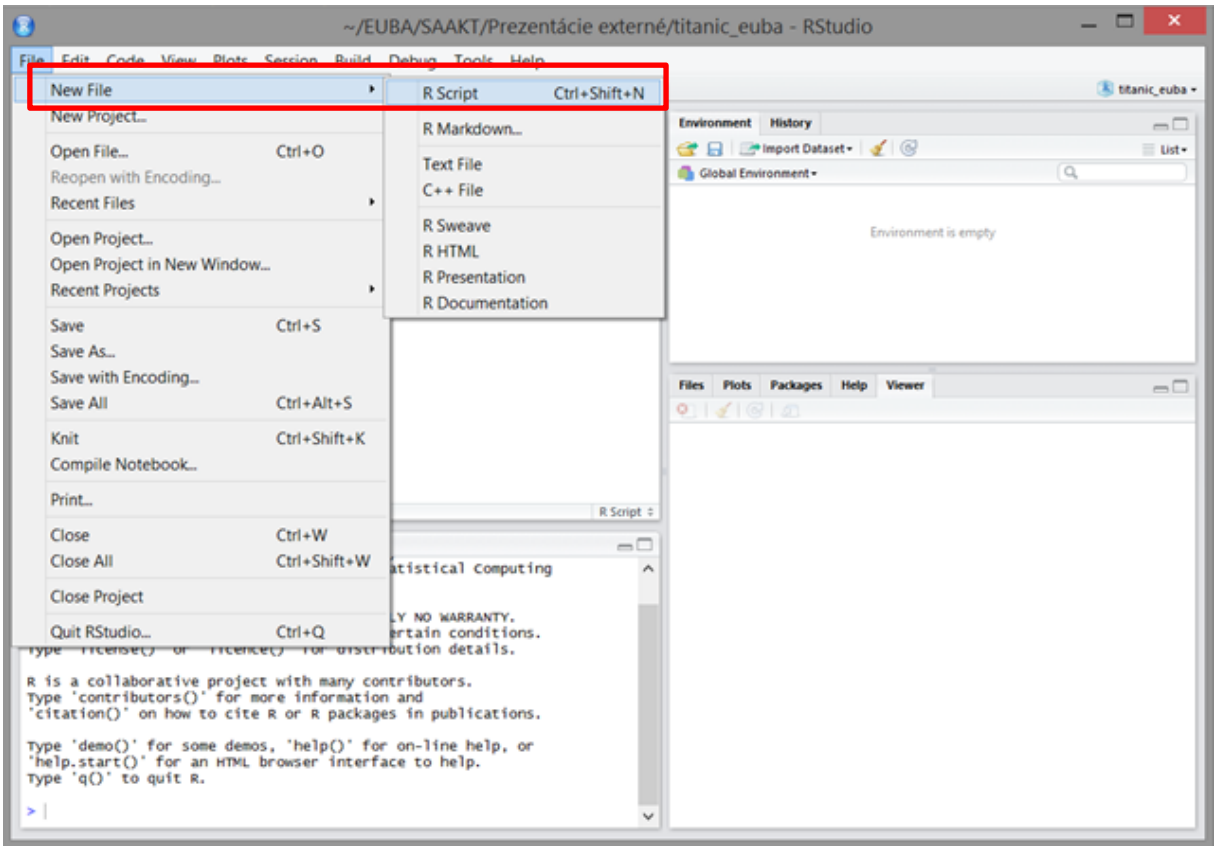
V 1. kvadrante (obr. 2) môžeme načítavať a ukladať **objekty** (údaje), ktoré v R využívame (ako *.RData) priamo cez kontextové menu (pozn. pre ukladanie objektov je potrebné poznať možnosti ukladania objektov v R, resp. rôzne funkcie (resp. knižnice) ukladania a načítavania údajov). Rovnako po kliknutí na ďalšiu záložku v tomto kvadrante môžeme sledovať **históriu** realizovaných príkazov a túto následne využiť v rámci konzoly



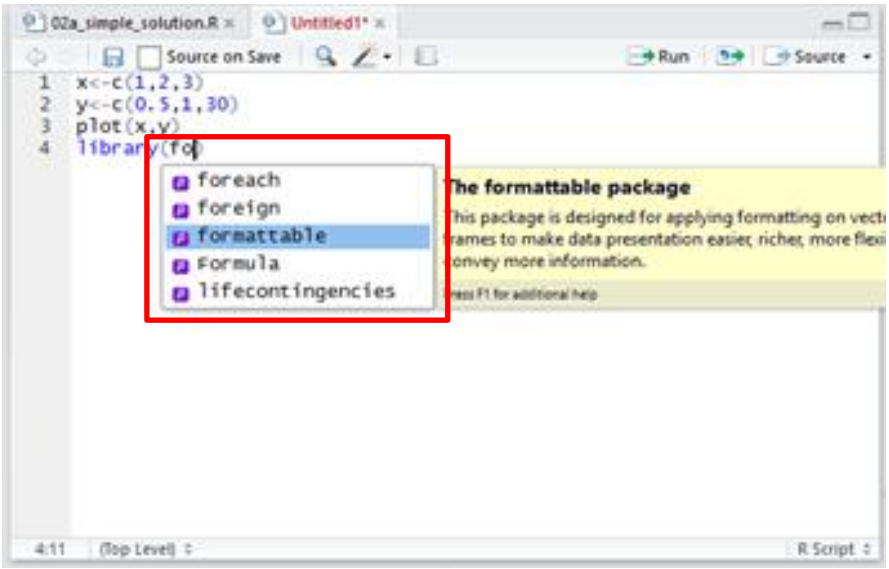
Obr. D2

Druhý kvadrant slúži na písanie **skriptov** (nový skript; obr. 3), pričom *RStudio* pracuje pri písaní skriptov ako pokročilý a efektívny editor kódu (automatické dopĺňanie textu, farebné rozlíšenie textu podľa častí kódu (funkcia, komentár, knižnica, objekt a pod., pozri obr. 4).

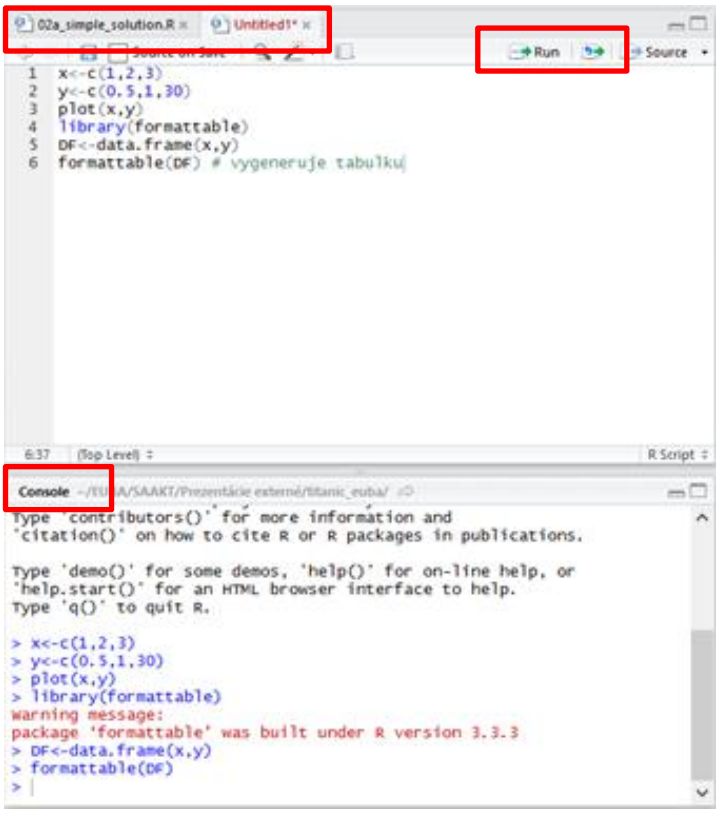
Skript spúšťame do **konzoly** (3. kvadrant) najčastejšie pomocou tlačidla „Run“ (obr. 5). Konzola jazyka R pracuje ako v štandardnom rozhraní. Viaceré skripty sa zobrazujú na otvorených kartách v záhlaví okna kvadrantu. Skript môžeme uložiť (*File/Save*) ako súbor s koncovkou *.R.


~/EUBA/SAAKT/Prezentácie externé/titanic_euba - RStudio
File Edit Code View Plots Session Build Debug Tools Help
New File
New Project...
Open File... Ctrl+O
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Window...
Recent Projects
Save Ctrl+S
Save As...
Save with Encoding...
Save All Ctrl+Alt+S
Knit Ctrl+Shift+K
Compile Notebook...
Print...
Close Ctrl+W
Close All Ctrl+Shift+W
Close Project
Quit RStudio... Ctrl+Q
R Script Ctrl+Shift+N
R Markdown...
Text File
C++ File
R Sweave
R HTML
R Presentation
R Documentation
titanic_euba
Environment History
Import Dataset
Global Environment
Environment is empty
Files Plots Packages Help Viewer
R Script
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Obr. D3

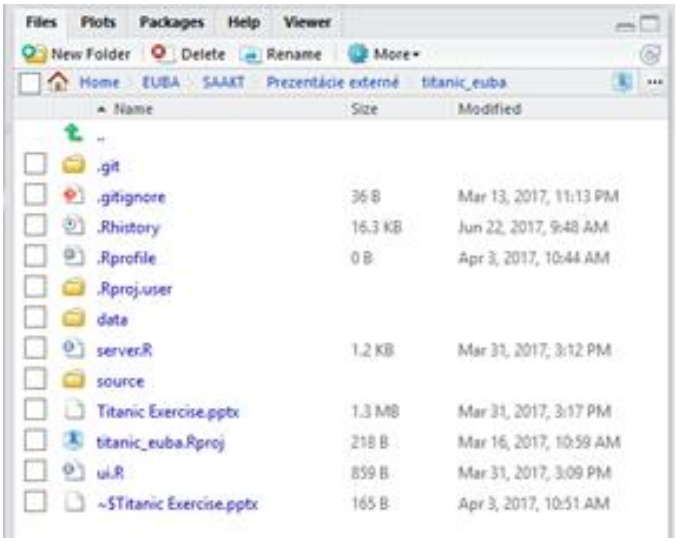

02a_simple_solution.R
Untitled1*
Source on Save
Run
Source
1 x<-c(1,2,3)
2 y<-c(0.5,1,30)
3 plot(x,y)
4 library(fo
foreach
foreign
formattable
formula
lifecontingencies
The formattable package
4:11 (Top Level)
R Script

Obr. D4

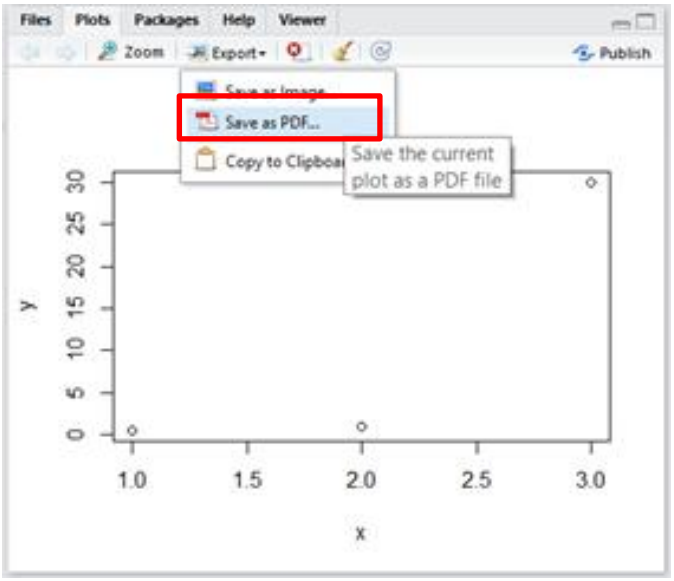


Obr. D5

Štvrtý kvadrant má päť užitočných kariet, ktoré môže používateľ využiť pre prácu s adresárom (**súbormi**) – najčastejšie s adresárom, kde je uložený skript a súbory s ktorými sa práve pracuje, ak používame *RProjects* (obr. 6). Ďalej okno pre **grafiku**, ktoré umožňuje pohodlné ukladanie grafických objektov (napr. aj do formátu PDF, obr. 7). Zvláštne okno, ktoré zobrazuje nainštalované **knižnice** a príkazy ktoré umožňujú manažovanie týchto knižníc (obr. 8). Ďalšie je okno pre ovládanie **nápovedy** a posledné okno **náhľadu**, kde sa napr. zobrazujú a dajú exportovať tabuľky údajov (*data.frame*), ak používame napr. knižnicu `formatttable` (obr. 9).



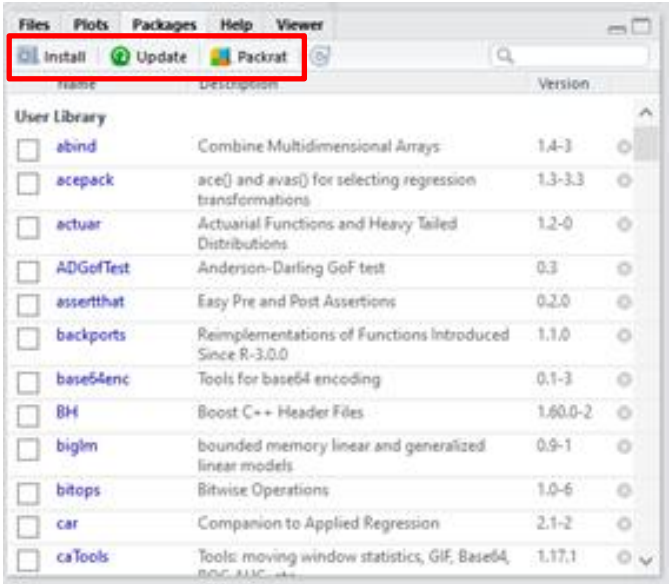
Obr. D6



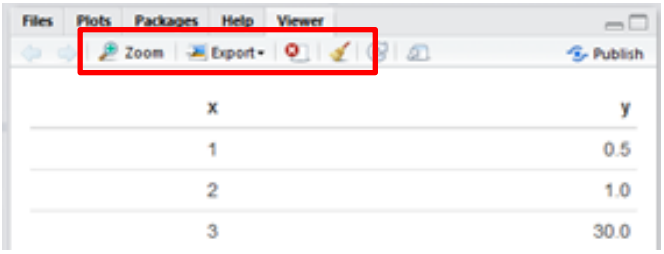
Obr. D7

Pre používanie **R projektov** (*RProjects*) je vhodné, aby si používateľ naštudoval túto oblasť pomocou napr. odkazu *Using Projects – RStudio Support*. Tieto sa používajú výhradne v editore kódu *RStudia* najmä pre zdieľanie práce pre ostatných používateľov. Umožňujú automaticky nastaviť pracovný adresár do adresára projektu, keď ho používateľ otvorí, pamätajú si aké súbory sú otvorené a poskytujú ďalšie preferencie súvisiace s editovaním kódu. Neslúžia však na ukladanie údajov z prostredia jazyka R.

Ako sme už uviedli vyššie *.RData je súbor objektov jazyka R. Vytvorením takého súboru môžeme ukladať objekty z R (nielen z *RStudia*) a neskôr ich načítať späť do pracovného prostredia. Takisto možno uložiť všetky objekty pracovného prostredia (pre viac informácií pozri `?save`, resp. `?save.image`). V *RStudiu* je potrebné rozlišovať medzi ikonou „*Uložiť*“ v rôznych oknách, kde napr. ikona v hornej časti prvého kvadrantu ukladá iba kód, ktorý bol napísaný ako skript. Až ikona v časti *Enviroment* uloží objekty R.



Obr. D8



Obr. D9

V nasledujúcej tabuľke veľmi stručne zhrňme s akými príponami súborov sa najčastejšie môže používateľ pri práci s jazykom R stretnúť:

***.R**	skript v R
***.RData**	uloženie súboru objektov v R
***.Rda**	skrátený názov pre to isté
***.Rds**	uloženie jediného objektu v R
***.Rproj**	R projekt (použitie len v *RStudiu*)
***.Rhistory**	súbor pre uloženie histórie
csv	je jednoduchý súborový formát vo forme čistého textu určený na ukladanie tabuľkových dát, pozostáva z ľubovoľného počtu záznamov (riadkov), oddelených znakom nového riadka, každý záznam obsahuje stĺpce, ktoré sú oddelené iným znakom, prevažne čiarkou (,) alebo tabulátorom, pre viac informácií o ukladaní v R pozri [1]. V MS Excel ukladáme údaje v zošite ako *.csv príkazom: *Uložiť ako/CSV (MS-DOS)*.

Rozdiel medzi súborom *Rds* a *Rda* uvádza príklad popísaný (v AJ) na obr. 10 (zdroj: [3]).

```
> x <- 1:5
> save(x, file="x.Rda")
> saveRDS(x, file="x.Rds")
> rm(x)

## ASSIGN USING readRDS
> new_x1 <- readRDS("x.Rds")
> new_x1
[1] 1 2 3 4 5

## 'ASSIGN' USING load -- note the result
> new_x2 <- load("x.Rda")
loading in to  <environment: R_GlobalEnv>
> new_x2
[1] "x"
# NOTE: `load()` simply returns the name of the objects loaded. Not the values.
> x
[1] 1 2 3 4 5
```

Obr. D10

ABECEDNÝ ZOZNAM RIEŠENÝCH PROBLÉMOV

☐ ***Doplnenie špecifického súčtového riadku do tabuľky údajov.***

☐ ***Duplikovanie hodnôt pre prázdne hodnoty v tabuľke údajov.***

☐ ***Filtrovanie v údajoch uložených do stĺpcov.***

☐ ***Formátovanie tabuľky.***

☐ ***Generovanie špecifického reťazca (napr. pre kódovanie, heslá, ID v tabuľkách, ...).***

☐ ***Generovanie špecifického reťazca (text „ID“ + 5 ciferné kladné celé č.) do tabuľky.***

☐ ***Nahradenie konkrétnej hodnoty v údajoch uložených do stĺpcov.***

☐ ***Nastavenie desatinných miest čísiel v tabuľke.***

☐ ***Práca s databázou údajov (oddelených čiarkami).***

☐ ***Prepísanie konkrétnych prvkov matice podľa zadefinovanej podmienky.***

☐ ***Pridanie časových údajov na os x do grafu časového radu.***

☐ ***Pridanie marginálnych súčtov do tabuľky údajov.***

☐ ***Pridanie prázdnej hodnoty do stĺpca tabuľky.***

☐ ***Pridanie znakov do reťazca.***

☐ ***Rozdelenie po sebe zapísaných údajov v jednom stĺpci do samostatných stĺpcov.***

☐ ***Rozdelenie textového reťazca spojeného znakom „_“ do samostatných stĺpcov.***

☐ ***Spojenie dvoch matíc.***

☐ ***Transformácia textového reťazca na formát numeric.***

☐ ***Tvorba jednoduchého cyklu (loop) for pre výpočet 1. diferencie.***

☐ ***Usporiadanie údajov do matice požadovaného tvaru (typu).***

☐ ***Vloženie textu a hodnoty definovanej premennej do poznámky (titulku) grafu.***

☐ ***Vloženie výstupu do grafu.***

☐ ***Vymazanie duplicitných riadkov tabuľky v závislosti na jednej hodnote.***

☐ ***Vymazanie riadkov s hodnotou „NA“ v databáze, podľa špecifických požiadaviek.***

☐ ***Vymazanie riadkov s hodnotou „NA“ v databáze.***

☐ ***Výpočet súčtu (sumy) početností v kontigenčnej tabuľke.***

☐ ***Využitie funkcie paste.***

☐ ***Znázornenie plochy kvantilov na grafe hustoty.***

☐ ***Zobrazenie dvoch kriviek hustoty na jednom grafe s využitím knižnice ggplot2.***

☐ ***Zobrazenie fragmentu histogramu na osiach grafu závislostí (scatter plot).***

☐ ***Zobrazenie, resp. odstránenie riadkov v tabuľke údajov***

ZDROJE

[1] PÁLEŠ, M.: *Jazyk R v aktuárskych analýzach*. Bratislava : Vydavateľstvo EKONÓM, 2017.
[2] https://www.r-bloggers.com/formatting-table-output-in-r/
[3] https://stackoverflow.com/
[4] https://sk.wikipedia.org/

KONTAKTNÉ ÚDAJE

Páleš, Michal, Ing., PhD., Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35 Bratislava, tel. +421 2/672 95 841, e-mail: pales.euba@gmail.com