R wstÄ™p

dr Magdalena Guzowska

11 sierpnia 2018

Dlaczego R

1. darmowe środowisko oblicznieowe - **open source**
2. pracuje na wszystkich platformach - Windows, MAC OSX and Linux
3. szerokie spektrum analiz statystycznych oraz możliwości graficznego przedstawiania danych
4. możliwość wielokrotnej i ujednoliconej analizy danych poprzez stosowanie skryptów
5. zrzesza wielu użytkowników (nauka i biznes) - prężne fora !!!
6. poszerzanie możliwości R dzięki tysiącom dobrze udokumentowanych rozszerzeń - pakietów R (zastosowanie w biologii, ochronie zdrowia, różnych dziedzinach nauki, sektorze finansowym itp.)
7. łatwość tworzenia własnych skryptów i pakietów R, służących do rozwiązywania specyficznych problemów

Kilka zastosowań R

1. szeroki wachlarz klasycznych testów statystycznych np.:
   - t-test Student'a m.in. do porównywania średnich 2 grup
   - test Wilcoxon'a - nieparametryczna alternatywa testu T
   - ANOVA - porównanie średnich więcej niż 2 grup
   - test Chi^2 porównujący proporcje/ dystrybucje
   - analiza korelacji (ocena relacji pomiędzy zmiennymi)
2. tworzenie klasyfikacji grup danych np.:
   - PCA (ang: principal component analysis)
   - klastrowanie
3. wiele typów prostych i skomplikowanych wykresów (ogromne możliwości zmian wygladu i zawartości) np.: wykres pudełkowy, histogram, wykres gęstości, wykres punktowy, wykres liniowy, wykres słupkowy, i wiele innych...

CRAN, Bioconductor, GitHub

CRAN - *Comprehensive R Archive Network* - to źródło dokumentacji oraz pakietów do języka R
https://cran.r-project.org/

Bioconductor - Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.
https://www.bioconductor.org/

GitHub – hostingowy serwis internetowy przeznaczony dla projektów programistycznych wykorzystujących system kontroli wersji Git. Serwis działa od kwietnia 2008 roku.
https://github.com/

Wszystkie te serwisy są źródłem pakietów rozszerzających możliwości R: **CRAN** - pakiety o różnorodnych zastosowaniach, **Bioconductor** - pakiety o zastosowaniach biologicznych, **Github** wersje pakietów w trakcie budowy.

Instalacja i korzystanie z pakietów R

pakiet R - rozszerzenie funkcjonalności R o zbiory danych i specyficzne funkcje - cel:odpowiedź na specyficzne pytania

R w podstawowej wersji - pakiety umożliwiające wykonywanie podstawowych analiz statystycznych i graficznej wizualizacji danych (wykresy) oraz dostarcza podstawowe zestawy danych.

Możliwość pobrania wielu innych pakietów z ww. serwisów

Podczas każdej sesji **niezbędne jest załadowanie odpowiedniego pakietu, aby móc z niego korzystać !!!**

Instalacja i korzystanie z pakietów R

1. Funkcja **install.packages("NAZWA_PAKIETU")** służy do instalacji pakietów z CRAN. Możliwa jest instalacja wielu pakietów jednocześnie:
   install.packages(c("NAZWA_PAKIETU1", "NAZWA_PAKIETU2"))

2. Do instalowania pakietów z Bioconductor'a używa się następującego polecenia:
   source("https://bioconductor.org/biocLite.R"), biocLite("NAZWA_PAKIETU")

3. Do instalacji z GitHub, korzysta się z pakietu **"devtools"** (Hadley'a Wickham'a)
   install.packages("devtools"), devtools::install_github("ŚCIEŻKA")

Instalacja i korzystanie z pakietów R

Tworzenie listy zainstalowanych pakietów - ***installed.packages()***

```
head(installed.packages(), n=2)
```

```
        Package   LibPath
abind   "abind"   "C:/Users/Magda/Documents/R/win-library/3
acepack "acepack" "C:/Users/Magda/Documents/R/win-library/3
        Priority Depends         Imports           LinkingTo
abind   NA       "R (>= 1.5.0)" "methods, utils" NA
acepack NA       NA             NA                NA
        Enhances License                 License_is_FOSS
abind   NA       "LGPL (>= 2)"           NA
acepack NA       "MIT + file LICENSE"    NA
        License_restricts_use OS_type MD5sum NeedsCompilati
abind   NA                    NA      NA     "no"
acepack NA                    NA      NA     "yes"
```

W RStudio lista zainstalowanych pakietów jest widoczna w jednym z paneli

Instalacja i korzystanie z pakietów R

Pakiety instalowane są w podkatalogu o nazwie library. Funkcja ***.libPaths()*** służy do wypisania ścieżki dostępu

```
.libPaths()
```

```
[1] "C:/Users/Magda/Documents/R/win-library/3.5"
[2] "C:/Program Files/R/R-3.5.1/library"
```

Uruchomione/ załadowane pakiety podczas danej sesji wypisywane są po użyciu funkcji ***search()***

```
search()
```

```
[1] ".GlobalEnv"        "package:stats"     "package:graphi
[4] "package:grDevices" "package:utils"     "package:datase
[7] "package:methods"   "Autoloads"         "package:base"
```

Instalacja i korzystanie z pakietów R

Najlepiej używać aktualnych wersji pakietów -> należy je uaktuaniać !!! Służy do tego funkcja:

update.packages()

Możliwe jest uaktualnianie tylko wybranych pakietów umieszczając w nawiasie funkcji ich nazwy np.:

update.packages(oldPkgs = c("readr", "ggplot2"))

R i RStudio

1. **R** służy do manipulowania danymi, obliczeń i graficznego przedstawiania danych.
2. Prosty interface graficzny pozwala na:
   - wprowadzanie i zapisywanie danych, manipulacje i obliczenia
   - stoswanie wszystkich narzedzi dostarczanych przez pakiety
   - dobrze rozwinięty język programowania, stosujący warunki, pętle, funkcje itp.
3. **RStudio** wykorzystuje środowisko graficzne komputera w celu ułatwienia współpracy z R i zawiera min.:
   - konsolę do wpisywania kodu
   - okno danych wprowadzonych z zewnątrz i wyników obliczeń R
   - okno do podglądu wprowadzonych i uzyskanych danych oraz okno podglądu wykresów
   - inne okna w formie zakładek np.: pakiety, historia, pomoc

R i RStudio - porównanie

1. RStudio ma więcej możliwości i zastosowań. Pozwala na bezpośrednią interakcję z kodem R
2. Standardowy interface R i RStudio równie dobrze radzą sobie z podglądem danych, ale nie dają możliwości ich prostej edycji (kopiowanie/wklejanie, ręczne wpisywanie)
3. RStudio sprawdza się lepiej w projektach, w których trzeba bezpośrednio ingerować w kod i podczas pracy ze złożonymi danymi

RStudio ustawienia i możliwości

1. Program darmowy !!!
2. Możliwość dostosowania wyglądu i widocznych paneli **[Opcje]**
3. Łatwość instalacji pakietów, wyszukiwania pomocy, wyszukiwania dokumentów na komputerze
4. Możliwość pracy bez zewnętrznego edytora
5. Podgląd wszystkich wykonanych podczas sesji wykresów
6. Podgląd wprowadzonych danych w formie podobnej do arkusza Excel lub klasycznej tabeli
7. Liczne, funkcjonalne dodatki m.in. Markdown i Git
8. Developer Tools - narzędzia do tworzenia i sprawdzania własnych funkcji i pakietów

Pomoc dotycząca funkcji R

Aby przeczytać więcej o danej funkcji używa się komendy **help()**
np.: dla funkcji *mean*
help(mean), *lub* ***?mean***

Aby zapoznać się z przykładami zastosowań funkcji używa się komendy **example(NAZWA FUNKCJI)** np.:
example(mean)

Funkcja **apropos()** zwraca listę obiektów zawierających sekwencję liter, której szukamy. Jest to przydatne, kiedy nie znamy nazwy funkcji np.:

```
apropos("med")
```

```
[1] "elNamed"          "elNamed<-"        "median"           "med
[5] "medpolish"        "runmed"
```

Pomoc dotycząca funkcji R

Inną funkcja do wyszukiwania pomocy jest **help.search()** (lub **??**), która zwraca on-line listę funkcji zawierającą wyszukiwany termin, z krótkim opisem

```
help.search("mean")
```

```
starting httpd help server ... done
```

Dane wbudowane

data() *wypisuje wszystkie dostępne dane dostarczone przez R*
data(mtcars) *- umożliwia korzystanie z danych o nazwie umieszczonej w nawiasie*

```
head(mtcars, 6)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0

Dane wbudowane **"mtcars"**

```
summary(mtcars)
```

```
      mpg              cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 5
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 9
 Median :19.20   Median :6.000   Median :196.3   Median :12
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :14
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:18
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :33
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0
 Median :3.695   Median :3.325   Median :17.71   Median :0
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
```

Dane wbudowane **“mtcars”**

```
nrow(mtcars)
```

```
[1] 32
```

```
ncol(mtcars)
```

```
[1] 11
```

Aby dowiedzieć się więcej o którymkolwiek z pakietów wykorzystuje się komendę **?[*nazwa pakietu*]** np.:

?mtcars

Najczęściej używane wbudowane dane R

1. **iris** - zestaw danych pomiarów w centymetrach zróżnicowania długości i szerokości płatków 50 kwiatów irysów z 3 różnych gatunków - *Iris setosa*, *I. versicolor* i *I. virginica*.
2. **ToothGrowth** - zestaw danych przedstawiających wpływ podawania wit. C na wzrost zębów 60 świnek morskich. Każde zwierzę otrzymywało jedną z trzech dawek witaminy C (0.5, 1, lub 2 mg/dzień) jednym z dwóch źródeł - pochodzące z soku pomarańczowego (OC) lub kwasu askorbinowego (VC).
3. **PlantGrowth** - wyniki eksperymentu mającego na celu porównanie plonów (suchej masy) uprawianych na dwa różne sposoby.
4. **USArrests** - statystyki dotyczące brutalnych przestępstw w USA z podziałem na stany.

Instalacja niezbędnego podczas zajęć oprogramowania

Na zajęciach korzystać będziemy z: R i dodatkowych pakietów R, RStudio, Notepad++ (opcjonalnie), Git (opcjonalnie)

1. Dostęp do internetu - możliwe jest pobieranie i instalowanie oprogramowania na bieżąco.
2. Kolejność instalacji - R, RStudio, pakiety (zgodnie z aktualnym zapotrzebowaniem - na początek konieczne będzie zainstalowanie **“knitr”**)
3. Brak dostępu do internetu - przygotowana wersja instalacyjna zawiera aktualną wersję R, RStudio oraz niezbędnych pakietów. Dodatkowo zawiera też Notpad++ i Git (do jego wykorzystywania niezbędne jest połączenie z internetem)