The basics of working in R

The objective of the lecture:

You will learn how to work with the basic R tools needed to work in R.
Objectives of the lecture:
Access R packages
Effectively organize your workspace

learn the methods and rules for loading data into R

Packages:

1. Package Overview
2. Installing packages in R and RStudio
3. Use of packages

Рекомендуемая литература:

1. Мастицкий С., Шитиков В. Статистический анализ и визуализация данных с помощью R. ДМК Пресс, 2015. - 496 с.
2. Роберт И. Кабаков. R в действии. Анализ и визуализация данных на языке R. ДМК Пресс, 2014. – 588 с.
3. An Introduction to R. интернет-источник: https://cran.r-project.org/doc/manuals/r-release/R-intro.html
4. Пакеты в R. Основы программирования на R. Видео (10 мин)

https://www.youtube.com/watch?v=DXzHCVEkFz8&list=PLu5flfwrnSD7wxKXFgsiuxrMKLfFHm6CD&index=10

1. Package Overview

A package is a collection of functions created to perform a specific class of tasks, or a collection of tables with data

Getting package information

1. not installed - the package was not installed using the install.packages function. You can get a list of such packages with the following command:

```
>setdiff(row.names(available.packages()), .packages(all.available = TRUE))
```

2. installed but not connected - the package was installed using the install.packages function, but not connected using the library function. You can get a list of such packages with the following command:

```
>setdiff(.packages(all.available = TRUE), (.packages()))
```

3. installed and connected - the package was installed using the install.packages function and connected using the library function. You can get a list of such packages with the following command

```
>(.packages())
```

2. Installing packages in R

Installing a new package (Internet connection required):

```
> install.packages("package_name")
```

3. Using Packages

Download an already installed package:

>library(package)

or

>require(имя_установленного_пакета)

When downloaded, the package may report various diagnostic information. You can suppress the output of these messages with the suppressPackageStartupMessages () function.

>suppressPackageStartupMessages(library(rvest))

The exercise

Connect the ggplot2 package and apply its qplot function:

```
>library(ggplot2)

> qplot(carat, price, data = diamonds)
```

Getting help that comes with the package A package can come with accompanying documentation (help), you can get it like this:

```
>help(package = "имя_пакета")
```

s

Package removal

```
>remove.packages("имя_пакета")
```

For example:

```
>remove.packages(«ggplot2")
```

Пакеты

Other functions for working with packages:

.libPaths() # returns the directory where the packages are installed

library() # listing installed packages
search() # listing downloaded packages

1. Preparing data for R

Data can be entered from the keyboard, imported from text files, from Microsoft Excel and Access.

1. Подготовка данных для R

Microsoft Excel is one of the most common programs for preparing data for R.

Before uploading to R, the Excel file is usually saved as a text file .txt or .csv

Some data preparation rules

- ✔ No empty cells – missing values are denoted as NA
- ✔ Assign a name to each variable:
- ✔ No spaces in names
- ✔ Names must not start with dots or numbers
- ✔ The file should be placed in the current working folder

	A	B	C	D
1	Treatment	Barrel	Length	Weight
2	Control	Control3	29.28	0.992
3	Control	Control3	29.83	0.772
4	Control	Control3	31.93	0.894
5	Control	Control3	26.63	0.822

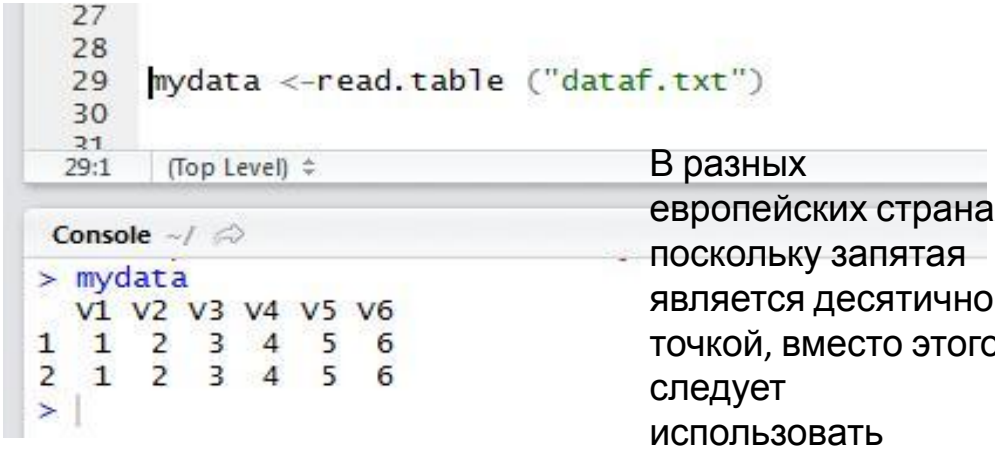
1. Подготовка данных для R

Рассмотрим чтение данных из текстового документа:

R может читать данные, сохраненные в текстовом (ASCII) файле.
Для этого используются три функции: read.table() (которая имеет два варианта: read.csv() и read.csv2(),
scan() и read.fwf().

Например, если мы имеем файл data.txt, то для того чтобы его прочитать можно набрать:

mydata <-read.table ("dataf.txt")



В разных европейских странах, поскольку запятая является десятичной точкой, вместо этого следует использовать функцию read.csv2

Функция read.table()

"Рабочая лошадка" для загрузки данных
Основные аргументы:

- **File** = "*имя*.txt": имя файла (или URL-ссылка)
- **Header** = TRUE : есть ли в файле заголовки столбцов
- **Sep** = = "\t" или sep = "," : разделитель значений в файле

An example of LOADING DATA

Iris Dataset
(archive.ics.uci.edu/ml/datasets/Iris)

download.file() – downloading file

read.csv() – reading data in csv

Upload the file to R

```
>fileUrl <- "http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"

>download.file(fileUrl, destfile="./iris.csv")
```

```
>iris.data <- read.csv("./iris.csv")  #   iris.data became data frame
```

Первичный анализ в R

```
>head(iris.data, 1)
```

```
      X5.1      X3.5      X1.4      X0.2 Iris.setosa
1    4.9  3.0  1.4  0.2   Iris-setosa
```

```
colnames(iris.data) <- c("Sepal.Length", "Sepal.Width",
"Petal.Length", "Petal.Width", "Species")
```

Saving a workspace

```
> save.image(file =
  "pH_experiment.rda")
```

Downloading a file from the Internet

Birth data for boys and girls from 1940 to 2002 in the United States

```
>source("http://www.openintro.org/stat/data/present.R")
>str(present)
>head(present)
>summary(present)
```

4. The treatment of missing values

Consider the following example: suppose we have the result of a survey of the same seven employees. They were asked: how many hours they sleep on average, while one of the respondents refused to answer, another said "I do not know", and the third at the time of the survey was simply not in the office. So there was a missing data:

```
>h <- c(8, 10, NA, NA, 8, NA, 8)
```

```
> h [1] 8 10 NA NA 8 NA 8
```

From the example you can see that NA should be entered without quotes, and R is not at all embarrassed that among the numbers there is a" like " text

4. The treatment of missing values

If we try to calculate the average value (the mean () function), we get:

```
>mean(h)
[1] NA
```

To calculate the average value not including NA, you can use one of two ways:

```
>mean(h, na.rm=TRUE)
>[1] 8.5

>mean(na.omit(h))
>[1] 8.5
```

4. Обработка пропущенных значений

Часто возникает ещё одна проблема: как сделать подстановку пропущенных данных, скажем, заменить все NA на среднюю по выборке.

Распространённое решение примерно следующее:

```
>h[is.na(h)] <- mean(h, na.rm=TRUE)

>h

>[1] 8.0 10.0 8.5 8.5 8.0 8.5 8.0
```

В левой части первого выражения осуществляется индексирование, то есть выбор нужных значений h таких, которые являются пропущенными (is.na()).

После того, как выражение выполнено, «старые» значения исчезают навсегда.

Вопросы для самопроверки

1. Какие источники данных для R вам известны?
2. Как в R считать текстовые файлы?
3. Как в R считать файлы из MS Excel?
4. Как в R считать интернет- файлы?
5. Как в R считать файлы баз данных?
6. Как в R привести исходные данные к аккуратному виду, пригодному для анализа?

Выводы по лекции 4

МЫ УЗНАЛИ:

- ✔ Какие источники данных можно использовать в R
- ✔ Какие данные считаются пригодными к анализу в R
- ✔ Как привести данные к аккуратному виду

МЫ НАУЧИЛИСЬ:

- ✔ Как скачать данные из файлов *.txt, Excel, Интернета и баз данных
- ✔ Как работать с пропущенными значениями
- ✔ Как задавать имена столбцам и строкам

Что такое пакет в R?

A. это программа, необходимая для установки языка R на компьютер

B. это набор драйверов для управления ресурсами в среде R

C. это коллекция функций, созданных для выполнения определенного класса задач, или коллекция таблиц с данными

D. это набор системных команд для управления ядром языка R

Как подключить установленный в R пакет?

A.
>(.packages())

B.
>install.packages(name, repos = uri)

C.
>library(package)

D.
>insert.packages(name, repos = uri)

Спасибо за внимание