федеральное государственное бюджетное образовательное учреждение высшего образования "Волгоградский государственный технический университет"

Факультет электроники и вычислительной техники

Кафедра «Системы автоматизированного проектирования и поискового конструирования»

**Лабораторная работа № 1: Загрузка данных в R**

1. **Цель работы**

* Приобрести основные навыки работы с данными в R
* Освоить наиболее распространенные способы загрузки данных в рабочую среду R

1. **Задачи**

* Углубить и закрепить знания по основным типам данных и научить загрузить и изучить свойства данных.
* Совершенствовать навыки самостоятельной работы.
* ...

1. **План выполнения работы**

**3.1 Загрузка различных данных**

* Классический вариант: загрузка данных форматов .txt или .csv в среду R: для этого воспользуемся командами:

*data <- read.csv("путь к файлу/имя файла.csv", sep = ";")*

*data <- read.table("путь к файлу/имя файла.txt",sep = "\t")*

* Загрузка файлов в форматах .xls или .xlsx (форматы программы Excel) :

**Способ 1:** Сохранение excel files в формате .csv, потом использование базовой функции *read.csv*

**Способ 2:** Используя пакеты **xlsx, readxl gdata, xlsReadWrite, XLConnect,…**

library(readxl)

M3C <- read\_excel("GitHub/LAB/LAB 1/data/M3C.xls")

View(M3C)

* Загрузка данных из других пакетов R: к данным доступа из определенного пакета используйте параметр пакета, например:

*data(package="FORA")*

*data(FORAYear, package="FORA")*

* Загрузка данных из интернета: для этого воспользуемся пакетом **Rcurl**

Например: Загрузка набора данных AirPassengers – количество перевозки авиапассажиров 1949-1960.

> library(RCurl)

> myfile <- getURL('https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/datasets/AirPassengers.csv', ssl.verifyhost=FALSE, ssl.verifypeer=FALSE)

> data <- read.csv(textConnection(myfile), sep = ",", header = TRUE)

> head(data)

X time AirPassengers

1 1 1949.000 112

2 2 1949.083 118

3 3 1949.167 132

4 4 1949.250 129

5 5 1949.333 121

6 6 1949.417 135

**3.2 Свойства данных**

Предлагаем, что мы уже загрузили данные в виде data frame c именем data1:

*data1 <- read.csv("FORA.csv", sep = ";")*

Для получения свойства загруженных данных воспользуемся следующими командами:

* head(x) — посмотреть начальные строки объекта
* tail(x) — посмотреть последние строки объекта
* dput(x) — получить представление объекта в R-синтаксисе
* summary(object) — общая информация об объекте
* str(object) — внутренняя структура объекта
* dim(x) — размерности объекта
* dimnames(x) — имена размерностей объекта
* names(x) — имена объекта
* length(x) — число элементов

Например:

* *head(data1)*
* *summary(data1)*

**3.3 Работа с Data Frame**

Data frame - это гибридный тип представления, одномерный список из векторов одинаковой длины. Таким образом, каждая data frame - это список колонок, причём внутри одной колонки все данные должны быть одного типа. Проиллюстрируем это на примере:

> name <- c("Kolia", "Jenhia", "Sasha", "Kastia", "Vasia", " Jora")

> weight <- c(60, 68, 71, 87, 67, 93)

> height <- c(174, 168, 178, 188, 165, 172)

> size <- c("L", "S", "XL", "XXL", "S", "M")

> sex <- c("male", "female", "male", "male", "female", "male")

> data <- data.frame(name, weight, height, size, sex)

> data

name weight height size sex

1 Kolia 60 174 L male

2 Jenhia 68 168 S female

3 Sasha 71 178 XL male

4 Kastia 87 188 XXL male

5 Vasia 67 165 S female

6 Jora 93 172 M male

Очень часто бывает нужно отобрать несколько конкретных колонок. Это можно сделать разными способами (исключаем столбец name и отобрать столбец weight):

> data[,-1]

weight height size sex

1 60 174 L male

2 68 168 S female

3 71 178 XL male

4 87 188 XXL male

5 67 165 S female

6 93 172 M male

> data[,2:4]

weight height size

1 60 174 L

2 68 168 S

3 71 178 XL

4 87 188 XXL

5 67 165 S

6 93 172 M

> data$weight

[1] 60 68 71 87 67 93

Отобрать из нашей таблицы только данные, относящиеся к женщинам? Вот один из способов:

> data[data$sex == "female",]

name weight height size sex

2 Jenhia 68 168 S female

5 Vasia 67 165 S female

Или отобрать только нужные данные с помощью команды subset:

> data1 <- subset(data, sex == "male")

> data1

name weight height size sex

1 Kolia 60 174 L male

3 Sasha 71 178 XL male

4 Kastia 87 188 XXL male

6 Jora 93 172 M male

> data2 <- subset(data, height >= 168 & sex == "male")

> data2

name weight height size sex

1 Kolia 60 174 L male

3 Sasha 71 178 XL male

4 Kastia 87 188 XXL male

6 Jora 93 172 M mal

Более сложным случаем селекции является сортировка таблиц данных. Для сортировки таблиц данных достаточно применить команду sort(). Например, если мы хотим отсортировать наши данные сначала по полу, а потом по росту, приходится выполнить следующую команду:

> data[order(data$sex, data$height), ]

name weight height size sex

5 Vasia 67 165 S female

2 Jenhia 68 168 S female

6 Jora 93 172 M male

1 Kolia 60 174 L male

3 Sasha 71 178 XL male

4 Kastia 87 188 XXL male