

# Въвеждане на данни и извеждане на графики

## Статистическа обработка на данни с R

Пламен Петров и Тодор Балабанов

Център за обучение  
Институт по информационни и комуникационни технологии  
Българската академия на науките

*p.petrov@iit.bas.bg todorb@iinf.bas.bg*

12.V.2020

## Acknowledgments

These teaching materials are funded by Velbazhd Software LLC and it is partially supported by the Bulgarian Ministry of Education and Science (contract D01-205/23.11.2018) under the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", approved by DCM # 577/17.08.2018.

# Съдържание

- 1 Редна директория
- 2 Проверка и установяване
- 3 Структурирани източници на данни
  - CSV файлове
  - Четене на данни от JSON формат
  - Excel файлове
  - SQL бази данни
- 4 Бинарни източници на данни
  - Други статистически програми като източници на данни
  - Бинарни файлове на R
  - Данни достъпни директно от R
- 5 Визуализация на данни от статистически анализ
  - Примерни данни
  - Хистограма
  - Диаграма на разсейване
  - Графики тип кутия
- 6 Заключение
  - Дискусия

# Работна директория

## Организация на файловете

### Път във файловата система

```
getwd()
[1] "/Users/todorbalkanov"
setwd("/Desktop")
getwd()
[1] "/Users/todorbalkanov/Desktop"
```

# Структурирани източници на данни

```
df <- read.table(file=
" http://raw.githubusercontent.com/TodorBalabanov/Statistical-
Data-Processing-with-R/master/data/tomato.csv",
header=TRUE, sep=",")
```

```
head( df )
tail( df )
supply(df, class)
```

# JavaScript Object Notation

## Четене на JSON данни

```
library( jsonlite )  
pizza <- fromJSON(  
  "https://raw.githubusercontent.com/TodorBalabanov/Statistical-  
  Data-Processing-with-R/master/data/pizza.json")  
class( pizza )  
head(pizza, n=3)
```



# Електронни таблици

- Четенето на Excel файлове в R не е толкова лесно
- Microsoft Excel е комерсиален софтуер
- Бинарните му файлови формати не са с отворен лиценз
- Има опити за четене с пакетите gdata, XLConnect, xlsReadWrite

## Релационни бази данни

### Сваляне на файл с данни

```
download.file("https://github.com/TodorBalabanov/Statistical-Data-Processing-with-R/blob/master/data/diamonds.db?raw=true",  
destfile="./diamonds.db", mode="wb")
```

## Процедура за достъп

### Връзка към базата данни

```
library(RSQLite)
driver <- dbDriver( "SQLite" )
class( driver )
connection <- dbConnect(driver, "./diamonds.db")
class( connection )
```

### Комадна за прекъсване на връзката

```
dbDisconnect( connection )
```

# Структуриране

## Изследване на базата данни

```
dbListTables( connection )  
dbListFields(connection, name="diamonds")  
dbListFields(connection, name="DiamondColors")
```

## Извадки от редове на една таблица

```
diamondsTable <- dbGetQuery(connection, "SELECT * FROM
diamonds", stringsAsFactors = FALSE)
colorTable <- dbGetQuery(connection, "SELECT * FROM
DiamondColors", stringsAsFactors = FALSE)
```

## Извадки от редове на две таблици

```
diamondsJoin <-dbGetQuery(connection, "SELECT * FROM
diamonds, DiamondColors WHERE diamonds.color =
DiamondColors.Color", stringsAsFactors = FALSE)
```

# Бинарни източници на данни

Други статистически програми като източници на данни

## Списък с файлови формати

Функция	Файлов формат
read.spss	SPSS
read.ssd	SAS
read.ocatave	Octave
read.dta	Stata
read.systat	Systat
read.mtp	Minitab





## Запис в RDS файл

```
saveRDS(c(21,04,1979), "object.rds")
```

## Четене от RDS файл

```
readRDS("object.rds")
```

## Демонстрация на функциите по пакетите

### Зареждане на примерни данни

```
data(diamonds, package="ggplot2")  
head(diamonds, n=3)
```

# Визуализация на данни от статистически анализ

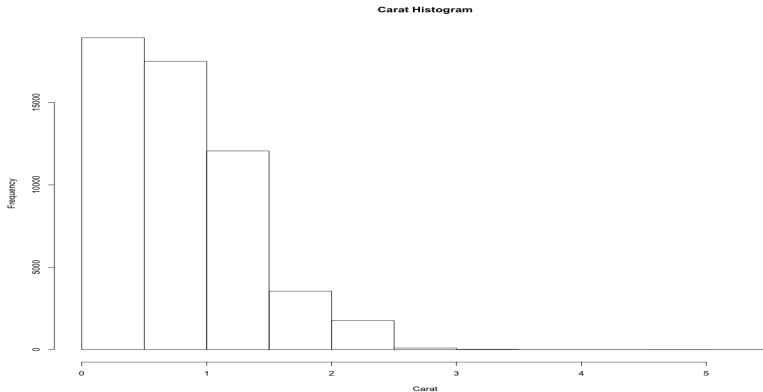
# Характеристики на диамантите

Характеристика	Значение
carat	Тегло на камъка (дробно число)
cut	Качество на среза (изброимо множество)
color	Цвят на камъка (изброимо множество)
clarity	Чистота на камъка (изброимо множество)
depth	Дълбочина на камъка (проценти)
table	Ширина на горната част, спрямо най-широката част (дробно число)
price	Цена (щатски долари)
x	Дължина (милиметри)
y	Ширина (милиметри)
z	Дълбочина (милиметри)

```
hist(diamonds$carat, main="Carat Histogram", xlab="Carat",
nclass=100)
```

## Хистограма

# Хистограма на каратите





## Отношение между два признака

Генериране на диаграма на разпръскване

```
plot(price~carat, data=diamonds)
```

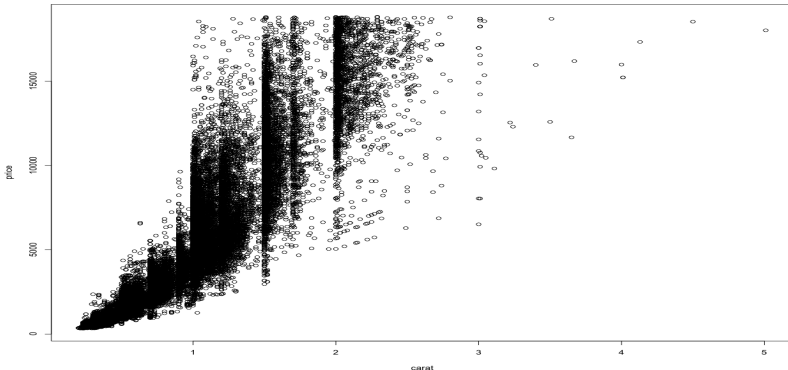
Алтернативен запис

```
plot(diamonds$carat, diamonds$price)
```



## Диаграма на разсейване

# Диаграма на разпръскване за камъните според отношението тегло към цена



# Квартили

Генериране на графика от тип кутия

```
boxplot( diamonds$carat )
```



# Заклучение

## Въпроси и отговори

Благодаря за вниманието!