DATA SCIENCE

R for Data Science I

Tomas Karpati MD

tc.datascience@gmail.com

054-2002430

מבוא ל-R במדעי הנתונים

מה נלמד?

- למה R
- תכנות בסיסית ב-R
- תכנות מתקדמת ב-R, יצירת פונקציות
- ניהול ומניפולציה של נתונים עם R
- ויזואליזציה של נתונים ב-R

ספר מומלץ ללימוד

R Programming for Data Science (https://bookdown.org/rdpeng/rprogdatascience/)

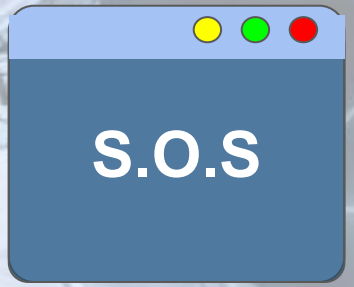
למה R ?

1. קוד פתוח - אפשר להשתמש בחינם וניתן לשנות את הקוד מתי שרוצים
2. תמיכה רחבה של קהילת R סביב לעולם
3. כל תכנות הסטטיסטיקה של החברות הגדולות (SAS, IBM/SPSS, וכו') תומכות בריצת קוד R בתוך הסביבה שלהם
4. היא הסטנדרט (יחד עם פייטון) בעולם ה-ML
5. יותר מ-15,208 ספריות רשמיות שונות במחסן ה-CRAN
6. ניתן להשתמש בה בפיתוח מוצרים בכל מערכות הפעלה (כולל אנדרואיד!)

R CRAN

איפה לקבל עזרה ב-R ?

1. help(ggplot2) או ?ggplot2
2. גוגל
3. Quick-R
4. StackOverflow
5. StackExchange
6. רשימות תפוצה - R-help / R-devel
7. Kaggle



ממשק העבודה ב RStudio: R-

File Edit Code View Plots Session Build Debug Profile Tools Help

karpati

Go to file/function | Addins | Rintro

Console | **Terminal**

~/Rintro/

```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

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.

>
```

Environment | **History** | **Connections**

Import Dataset | List

Global Environment

Environment is empty

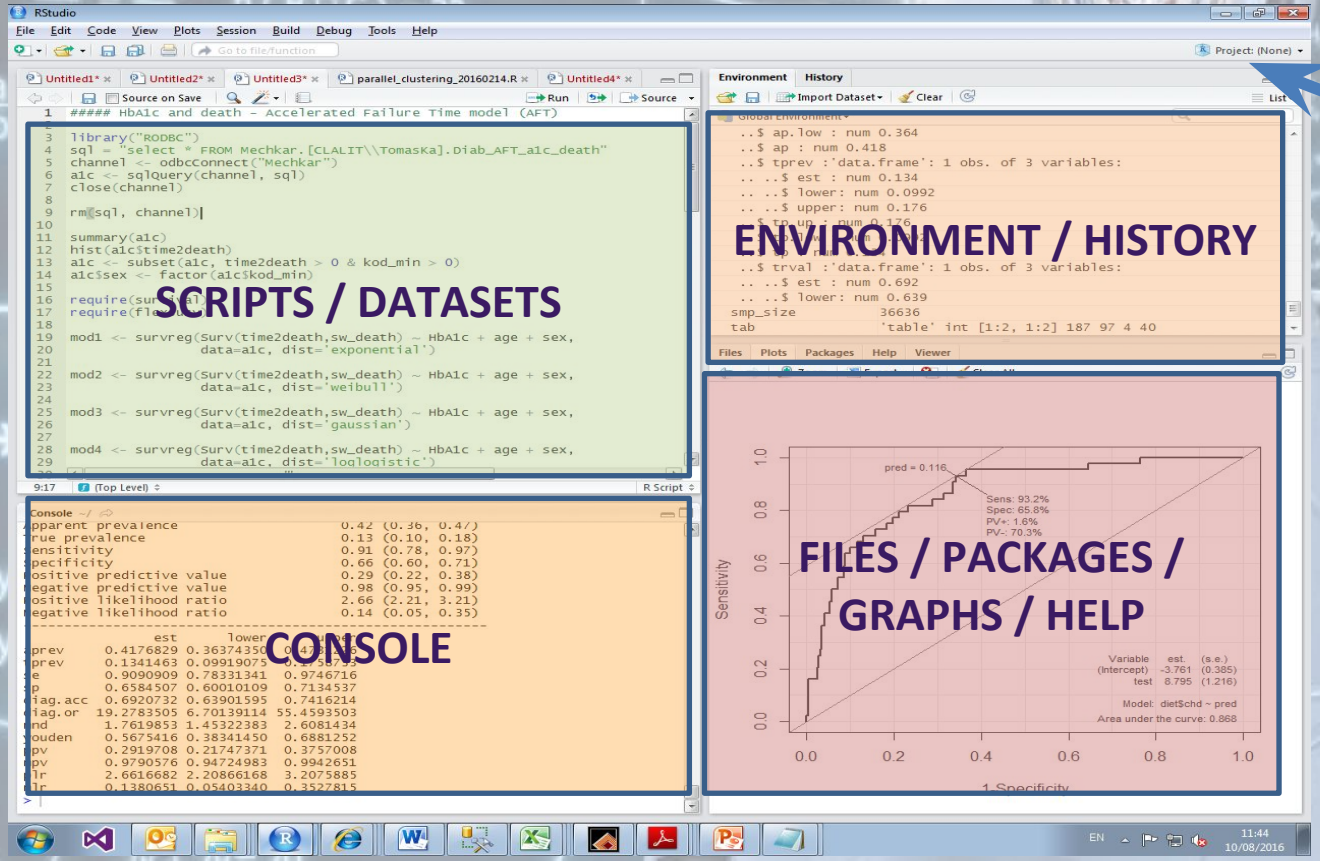
Files | **Plots** | **Packages** | **Help** | **Viewer**

New Folder | Upload | Delete | Rename | More

Home > Rintro

Name	Size	Modified
..		
Rintro.Rproj	205 B	Jul 8, 2018, 7:15 AM

RStudio IDE Functionality



R: עקרונות השפה

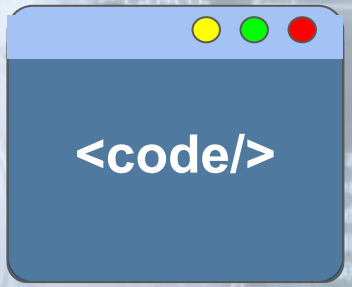
- Vector
- Matrix
- Array
- List
- Dataframe

a. נומרי (integer,numeric)
b. מחרוזות (character)
c. קטגוריות (factor)
d. תאריכים (Date)
e. בוליאני (TRUE/FALSE)

תרגול

תרגיל:

1. לפתוח את ה-RStudio
2. ליצור פרוייקט חדש בשם "R-class-exercise.R"
3. לפתוח R-script חדש
4. ליצור משתנה 'a' הכולל מספרים מ-10 עד 20
5. ליצור משתנה 'b' הכולל אותיות מ-d עד m.
6. ליצור משתנה בשם 'f' ולהוסיף בו שלושה 1 וחמישה 0
7. להפוך 'f' לפקטור ולהגדיר 0 כ-"YES" ו-1 כ-"NO"
8. לחפש עזרה עבור "objects"
9. להשתמש בפקודה המתוארת בעזרה. מה מקבלים?
10. לשמור את הקובץ
11. לסגור את הפרויקט



R: עקרונות השפה

- Vector
- **Matrix**
- Array
- List
- Dataframe

Define a 3x4 matrix
matrix(x, nrow=3, ncol=4)

[10, 15, 20, 25,
20, 20, 20, 20,
18, 17, 16, 15]

Define a 1x3 matrix
matrix(x, nrow=1, ncol=3)

[3, 5, 8]

R: עקרונות השפה

Addition of matrices

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 1 & -3 \end{bmatrix} = \begin{bmatrix} 1+2 & 2+0 \\ 3+1 & 4+-3 \end{bmatrix}$$

2 x 2 2 x 2 2 x 2

A+B=B+A

Subtraction of matrices: same way

R: עקרונות השפה

Multiplication of matrices

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 0 \\ 1 & -3 \end{bmatrix} = \begin{bmatrix} (1\times 2)+(2\times 1) & (1\times 0)+(2\times -3) \\ (3\times 2)+(4\times 1) & (3\times 0)+(4\times -3) \end{bmatrix}$$

$$\begin{bmatrix} 4 & -6 \\ 10 & -12 \end{bmatrix}$$

2 x 2 2 x 2 2 x 2

R: עקרונות השפה

Multiplication of matrices

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 0 \\ 1 & -3 \end{bmatrix} = \begin{bmatrix} 4 & -6 \\ -9 & -12 \end{bmatrix}$$

2 x 2   2 x 2   2 x 2

$$A.B \neq B.A$$

$$\begin{bmatrix} 2 & 0 \\ 1 & -3 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} (2x1)+(0x3) & (2x2)+(0x4) \\ (1x1)+(-3x3) & (1x2)+(-3x4) \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ -8 & -10 \end{bmatrix}$$

2 x 2   2 x 2   2 x 2

R: עקרונות השפה

Multiplication of matrices

$$\begin{bmatrix} 1 & 2 & -2 \\ 3 & -4 & -1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ -1 & -3 \\ -2 & -1 \end{bmatrix} = \begin{bmatrix} (1\text{x}1)+(2\text{x}-1)+(-2\text{x}-2) & (1\text{x}0)+(2\text{x}-3)+(-2\text{x}-1) \\ (3\text{x}1)+(-4\text{x}-1)+(-1\text{x}-2) & (3\text{x}0)+(-4\text{x}-3)+(-1\text{x}-1) \end{bmatrix}$$

$$\begin{bmatrix} 3 & -10 \\ 6 & 13 \end{bmatrix}$$

2 x 3   3 x 2   2 x 2

(3 x 2) x (3 x 2) = non-conformable

(3 x 4) x (4 x 2) = (3 x 2)

(4 x 1) x (4 x 1) = non-conformable

(2 x 5) x (5 x 5) = ?

(4 x 6) x (3 x 6) = ?

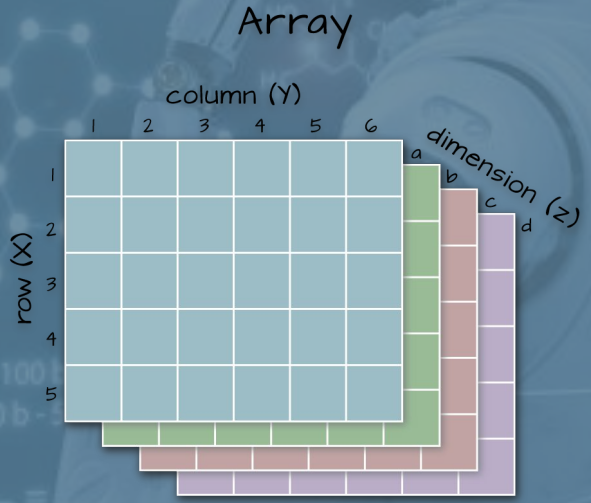
(3 x 2) x (2 x 1) = ?

R: עקרונות השפה

- Vector
- Matrix
- **Array**
- List
- Dataframe

מערכים (Arrays) הם כמו מטריצות, רק שמכילים יותר משני-מימדים

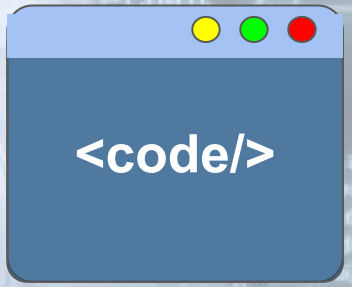
array(data=c(x,y,z), dim=c(5,6,4))



תרגול

תרגיל:

1. לפתוח את ב-RStudio את הפרויקט בשם: "R-class-exercise.R"
2. ליצור מטריצה עם שלוש שורות ושמונה עמודות.
3. ליצור מטריצה עם שתי שורות וחמש עמודות.
4. ליצור array עם 16 שורות, 16 עמודות ושלושה מימדים.



R: עקרונות השפה

Vector

Matrix

Array

List

Dataframe

רשימות הם עצמים שבהם ניתן לאחסן כל סוג של עצמים אחרים, כולל גם רשימות.

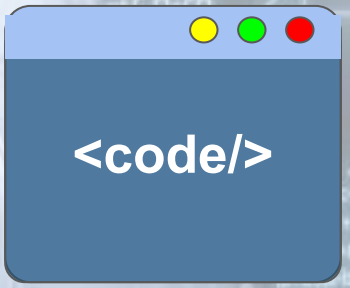
```
n <- list(a=c(1,2,3,4,5), b=c("a","b","c","d"),
          q=matrix(1:6,ncol=2),z=FALSE)

n$a
n[[1]]
n["a"]]
```

תרגול

תרגיל:

1. ליצור רשימה לפי הטבלה הבאה:

Rank	Peak	Title	Worldwide gross	Year
1	1	*Avatar*	$2,787,965,087	2009
2	1	*Titanic*	$2,187,463,944	1997
3	3	*Star Wars: The Force Awakens*	$2,068,223,624	2015
4	4	*Avengers: Infinity War* †	$1,844,894,638	2018
5	3	*Jurassic World*	$1,671,713,208	2015



1. לחלץ את השם של סרט השני ברשימה שיצרתם לפי
   1. השם של הפרמטר של הכותרת
   2. המיקום (רמה) של הפרמטר ברשימה

עקרונות השפה R:

Vector

Matrix

Array

List

Dataframe

מסגרת נתונים (Data frame) הינה טבלה עם סוגים שונים של וקטורים מסודרים בעמודות (משתנים). כל שורה הינה רשומה עם ערכים עבור כל אחד של המשתנים.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3

תרגול

תרגיל:

1. לבנות מסגרת נתונים (data frame) מבוסס על הטבלה הבאה:

name	age	is.married	city	has.pet
Avi	31	FALSE	Jerusalem	TRUE
Ben	25	FALSE	Jerusalem	FALSE
Gad	28	TRUE	Haifa	FALSE
Dan	28	FALSE	Jerusalem	FALSE
Harel	33	TRUE	Haifa	TRUE
Vered	27	TRUE	Tel Aviv	FALSE
Zelig	32	FALSE	Tel Aviv	TRUE

1. מה הגיל של המקרה השלישי?
2. כמה אנשים נשואים?
3. מה ממוצע הגיל של הקבוצה?
4. לכמה אנשים שאינם גרים בירושלים יש חיית מחמד?
5. תמחקו מהטבלה כל מי שמעל גיל 30.