

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Departament d'Estadística i Investigació Operativa

ACADEMIC YEAR 2021-22

STATISTICAL INFERENCE AND MODELLING – MASTER OF DATA SCIENCE Introduction to R: EDA Introduction to R software Lecturer: Lídia Montero September 2021 – Version 1.1

MASTER OF DATA SCIENCE



1. LAB SESSIONS

- 2 hours every 1 week, in a PC's classroom.
- Practical assignments posted through ATENEA TASKS. Formative assessment will be given by the lecturer before the next laboratory session when deliverable is indicated.
- Guidelines for laboratory session posted in ATENEA Course webpage
- Datasets posted on ATENEA Course webpage.

FIRST SESSION: Introduction to R and R Studio statistical software

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.



⇒Davis data: davis.RData (data.frame) - Use comands in davis.R for basics

```
> library(car)
Loading required package: MASS
Loading required package: nnet
> data(Davis)
> ls()
[1] "Davis"
> attributes (Davis)
$names
              "weight" "height" "repwt"
[1] "sex"
                                          "repht"
$class
[1] "data.frame"
$row.names
  [1] "1"
             "2"
                   "3"
                                       "6"
                          "4"
                                "5"
                                             "7"
                                                    "8"
                                                           11 9 11
                                                                 "10" "11"
                                                                              "12"
 ...
[193] "193" "194" "195" "196" "197" "198" "199" "200"
```

- 2.1 Univariate descriptive analysis Numeric data
 - <u>Missing and Outliers might occur</u>
 - Numerical values
 - Measures of Central Tendency: Mean, Median, Mode
 - Measures of Dispersion: Variance, Standard Deviation, Quartiles, IQR, Maximum, Minimum.
 - Graphical Representations
 - Histogram, Cumulative Histogram. Absolute or relative.
 - BoxPlot.
 - Dotplot

- 2.1.1 Continuous Univariate Descriptive Analysis: Numeric statistics
- > summary(dataframe) # R command
- Mean $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Median: Value of the variable such that

50% Observations are < Median (Q2) & 50% Observations are > Median (Q2)

- Quartile Q1 of the 25% and quartile Q3 of the 75%: Values of the variable that
 - 25% Observations are < Q1 & 75% Observations are > Q1
 - 75% Observations are < Q3 & 25% Observations are > Q3
- Variance $\mathbf{S}_{\mathbf{X}}^{2} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{i} \overline{\mathbf{X}})^{2}$
- Standard Deviation S_x (square root of variance)



⇒Davis data: davis.RData (data.frame) - Use comands in davis.R for basics



• Missing data: Do not miss them! Track them.

NA: Not available - Missing data NaN: Not available for numerical reasons (divided by 0)

2.1.2 Continuous Univariate Analysis Description: Histogram



Prof. Lídia Montero ©

120

120

140

٠

160

160

UNITS 1 AND 2: EXPLORATORY DATA ANALYSIS (EDA)-UNIVARIATE

2.1.3 Continuous Univariate Analysis Description: Dotplot

Dotplot:

```
# Dotplot
par(mfrow=c(2,1))
stripchart(Davis$weight,method="stack")
                                                              40
                                                         60
                                                               80
                                                                    100
                                                            Dotplot Weight in Davis dataset
stripchart(Davis$weight,method="stack"
    ,xlab="weight",pch=19,
    col=3,
    main="Dotplot Weight in Davis dataset")
```

140

40

60

80

100

weight

2.1.4 Continuous Univariate Analysis Description: Boxplot

Boxplot: Basic implementation in boxplot() method, recommended Boxplot() method in car library "Five issues Summary" (Min, Q1, Me, Q3, Max) for Univariate EDA, useful to detect the presence of outliers.









2.2 Univariate descriptive analysis - Categorical data

Description of categorical variables: only 'missings' might occur. Graphical representations:

- barplot (a) absolute or relative (proportions)
 b) density or accumulated.
 - Suitable for graphical description of discrete-qualitative data (factor) with a few levels or categories.
- Pie Chart.

```
table(Davis$sex)
margin.table(table(Davis$sex))
prop.table(table(Davis$sex))

par(mfrow=c(2,2))
barplot(table(Davis$sex))
barplot(table(Davis$sex),col=rainbow(2))
pie(table(Davis$sex),col=rainbow(2))
pie(table(Davis$sex),col=rainbow(2))
```









3. INTRODUCTION TO R

- *RStudio* for Windows: Basic Input/Output (R is case sensitive)
- Select working directory in R console window (Change dir / Cambiar dir ...)





INTRODUCTION TO R: LOAD WORKSPACE (RETRIEVE PREVIOUS USED DATA)

RStudio														
File Edit Code View Plots	Session Build D	ebug Profile	Tools Help				5	elecionar image	e para cargar				? ×	
New File	•	+ Addins +						Buscar en:	C LABORAT	TORI	- G) 🏚 📂 📖)-	遂 Project: (None) 🔻
New Project		u_Sessió4_PCA	Rmd × 💽 Lab23Ma	r2018_PCA.Rmd ×	IntroR.Rmd* ×		_ [RAmericanVo	'oter.RData	R glmzme.RData	R sessio10	_copen.RData	
Open File	Ctrl+O					😳 Inse	t• ☆ ⊕ ⊟ <mark> </mark> 2	_ <u>(</u>	Anscombe7	73.RData	R jsp3.RData	R sessio10	_womenIf.RData	🗏 List 🕶 🛛 🎯
Reopen with Encoding							-	Documentos recientes	ceb.RData	ta a	R jsp.RData	💦 spa_clain	ns.RData	Q
Recent Files	*								R copenhagu	uen.RData	R marefill.RData			
Open Project								Escritorio	crossedeffe	ects.RData a	Moore.RData			
Open Project in New Session	n							•	R davis.RDat	ta	R nested.RData			
Recent Projects	•							>	dobson.RD	Data	R petrol.RData			
Import Dataset	•							Mis documentos	duncan2.R	RData	R sessio6_moore.RData			
Save	Ctrl+S								efectesmix	ktes.RData	R sessio7_moore.RData			
Save As								MiPC	R Eleccions_9	92.RData	R sessio9_Dobson.RData			
Save with Encoding									R exemples.F	RData	R sessio9_Elecc92.RData			
Save All	Ctrl+Alt+S	w(15))							•			1	F	
Knit Document	Ctrl+Shift+K							Mis sitios de red	Nombre:			•	Abrir	
Publish]		Tipo:	R images (*.RD	Data)	•	Cancelar	
Print				=										💁 Publish 👻 🥃
Close	Ctrl+W	.0 Max	. :124.00	Max.	:200.0				•		Llists	arom of l	Devietureight	
Close All	Ctrl+Shift+W	NA'	s :17	NA's	:17				н		HISLO	gram or i	Davisşweign	L
Close All Except Current	Ctrl+Alt+Shift+W													
Close Project		t" "rep	ht"						Q	_				
Quit Session	Ctrl+Q								e G					
[1] "data framo"	•								<u> </u>					

• Open (load) and Save Workspace - File Menu (Archivo)

File/Archivo → Cargar área de trabajo (load workspace) File/Archivo → Guardar área de trabajo (save workspace)

Example: Open/Load Davis.RData from a Workspace.

• To exit: File/Archivo \rightarrow Salir or quit() command in R Console



INTRODUCTION TO R: SCRIPTS

From Archivo (File) menu: you can open, close, save, create a new, save as scripts.

Scripts are text files containing R command. Always use them to track lab session commands

Markdown documents are dynamic documents combining ordinary text and R commands. They can be interpreted to produce an output: html, pdf or word.





_ 0 X

INTRODUCTION TO R: FILE MENU

Knit to produce R Markdown output:

RStudio

	File Edit Code View Plots Session Build Debug Profile Tools Help	D/Dropbox/DOCENCIA/FIB-ADEI/LABORATORY/IntroRhtml	
	🔍 🔹 🤕 🔹 🔜 🚔 🕐 Go to file/function	IntroR.html D Open in Browser O, Find	📀 Publish 👻 🛞 Project: (None)
Critical Flements	● DRT-PassDA_v4.Rmd × ● ODEMIT_KINEO_v3.Rmd × ● Meu_Sessi64_PCA.Rmd × ● Lab23Mar2018_PCA.Rmd × ◆ ○ ○ ○ ○ • • • • • • • • • • • • • • • • • • •	IntroR	
	28 + Item 2b 29	Lídia Montero	
: D.	30 31 # Useden 1	7 de septiembre de 2018	
	32 - ## Header 1	output: html_document: toc: true toc_depth: 3 number_sections: true editor_options: chunk_output_type: console	
	33 34 35 - ```{r}	Presentation	=
 Expressions 	36 library(car) 37 data(Davis)	R Markdowns document	
	38 summary(Davis) 39 attributes(Davis)	This is an R Markdown document. We are showing some examples of EDA. Markdown is a simple formatting syntax for authoring HTML, PDF,	
and Objects	40 hist(Davis\$weight,15,col=rainbow(15))	and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com. Use * to provide emphasis such as italics and bold.	
and Objects	41 42	Create lists: Unordered * and + or ordered 1. 2.	
ý l	43 411 @ Chunk1 ÷	1. Item 1	-
(escalars	Console R Markdown ×	2. Item 2 • Item 2a	😏 Publish 🔻 🌀
(00001010),	~/ Ø	∘ Item 2b	
vectors	> attributes (Davis) \$names	Header 1	
(00)0,0,	<pre>[1] "sex" "weight" "height" "repwt" "repht"</pre>	Header 2	
mathicad	\$class		
marrices,		library (car)	
lists atc)	\$row.names [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" [19] "19" "20" "21" "22" "23" "24" "25" "26" "27" "28"	## Warning: package 'car' was built under R version 3.4.4	
11313, 610)	[37] "37" "38" "39" "40" "41" "42" "43" "44" "45" "46" [55] "55" "56" "57" "58" "59" "60" "61" "62" "63" "64"	## Loading required package: carData	
T I I ·	[73] "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" [91] "91" "92" "93" "94" "95" "96" "97" "98" "99" "100"	<pre>## Warning: package 'carData' was built under R version 3.4.4</pre>	
• The basic	[109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" [127] "127" "128" "129" "130" "131" "132" "133" "134" "135" "136"	data (Davie)	
	[145] "145" "146" "147" "148" "149" "150" "151" "152" "153" "154" [163] "163" "164" "165" "166" "167" "168" "169" "170" "171" "172"	summary (Davis)	
object is a	$ \begin{bmatrix} 105 \\ 181 \end{bmatrix} \begin{bmatrix} 181 \\ 182 \end{bmatrix} \begin{bmatrix} 182 \\ 183 \end{bmatrix} \begin{bmatrix} 184 \\ 184 \end{bmatrix} \begin{bmatrix} 185 \\ 185 \end{bmatrix} \begin{bmatrix} 105 \\ 105 \\ 185 \end{bmatrix} \begin{bmatrix} 105 \\ 187 \end{bmatrix} \begin{bmatrix} 171 \\ 172 \\ 188 \end{bmatrix} \begin{bmatrix} 171 \\ 172 \\ 189 \end{bmatrix} \begin{bmatrix} 171 \\ 189 \\ 189 \end{bmatrix}$	žž sev vejcht bejcht renut renht	
UDJECT 15 U	[199] "199" "200"	## F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. :148.0	160
	<pre>> hist(Davis\$weight,15,col=rainbow(15)) > </pre>	## Median : 63.0 Median :169.5 Median : 63.00 Median :168.0	-
list: list().			
	🚱 🜍 📗 MSCTM 🚺 MSCTM R RStudio 🔃 D:/Dro	🗧 🧐 Inbox 🔇 🥝 Google 🏋 IEEE-T 🤔 2011 fr 2011 Fr 2014 fr 🐨 TrRQua ES	▲ 📭 📶 🕩 😻 13:33 07/09/2018
			07703/2010

• Data matrix -

rows are individuals and columns are variables: data.frame.



INTRODUCTION TO R: CONSOLE, DEVICES AND SCRIPTS

Available ones:

R Console (to write command and obtain results) As many script windows as you want.

Data

Command win.graph() to create a new graphic device.

e Edit Code View Plots Session Build Debug Profile Tools Help				
🔹 🞯 🔹 🔜 🛛 🚔 🖉 Go to file/function 💦 🛛 🔯 🔹 Addins 👻				(No
DRT-PassDA_v4.Rmd × 🐑 ODEMIT_KINEO_v3.Rmd × 🖤 Meu_Sessió4_PCA.Rmd × 🖤 Lab23Mar2018_PCA.Rmd × 🖤 IntroR.Rmd ×	-0	Environment History		_
🗁 🔊 🔒 💞 💁 💉 Knit + 🐵 +	🖓 Insert 🔹 🏠 😓 Run 🔹 🍲 🗮	🞯 🔒 📑 Import Dataset	- 🖌	🗏 List 👻
28 + Item 2b	A	🐴 Global Environment 🝷		Q,
29		Data		
30		ODavis	200 obs. of 5 variables	
31- # Header 1 32- ## Header 2				
33				
34				
35 - ```{r}	© ≍ ►			
37 data(Davis)				
38 summary(Davis)				
39 attributes(Davis)				
40 hist(Davis\$weight,15,col=rainbow(15))	E			
41 42				
43	-	Filer Diete Dackager He	ala Mauar	
41:1 G Chunk 1 0	R Markdown 🗘	Thes Flots Fackages The	ng viewei	G Dublich -
onsole R Markdown ×	_			- TODIAT
class 1] "data.frame"		8 7		
row. names				
[1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "	13" "14" "15" "16" "17" "18"	ω.		
[19] "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30" " [37] "37" "38" "39" "40" "41" "43" "43" "44" "45" "46" "46" "47" "48" "	31" "32" "33" "34" "35" "36"	> 2 -		
[57] "57 - 58 - 59 - 40 - 41 - 42 - 45 - 44 - 45 - 46 - 47 - 48 - 7 [55] "55" "56" "57" "58" "59" "60" "61" "62" "63" "64" "65" "66" "i	67" "68" "69" "70" "71" "72"	en o		
[73] "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83" "84" "	85" "86" "87" "88" "89" "90"	edn 30		
[91] "91" "92" "93" "94" "95" "96" "97" "98" "99" "100" "101" "102" "	103" "104" "105" "106" "107" "108"	μ ^μ		
109] "109" "110" "111" "112" "113" "114" "115" "116" "11/" "118" "119" "120" ". 127] "127" "128" "120" "130" "131" "132" "133" "134" "136" "136" "137" "138" ".	121" "122" "123" "124" "125" "126" 139" "140" "141" "142" "143" "144"	- 3		
145] "145" "146" "147" "148" "149" "150" "151" "152" "153" "154" "155" "154" "155" "156" "2	157" "158" "159" "160" "161" "162"			
163] "163" "164" "165" "166" "167" "168" "169" "170" "171" "172" "173" "174" "	175" "176" "177" "178" "179" "180"	ę -		
181] "181" "182" "183" "184" "185" "186" "187" "188" "189" "190" "191" "192" ": 1993 "1994" "2994"	193" "194" "195" "196" "197" "198"			
199] 199 200		o J 		
hist(Davis\$weight,15,co]=rainbow(15))	=			110 100
hist(Davis\$weight,15,col=rainbow(15))	8.0°	40	60 80 100 120	140 160
win, graph()			Davis\$weight	

Graphic Devices: *R graphics* has a matrix structure that allows to obtain several figures: for ex. 2 rows and 2 columns

par(mfrow=c(2,2))



INTRODUCTION TO R: COMMAND STRUCTURE

- R Command structure:
- > Command parameters <CR>
- > Command parameters ; Command parameters <CR>

RStudio	
File Edit Code View Plots Session Build Debug Profile Tools Help	
🝳 🔹 😭 📄 🔚 🦾 Go to file/function 📄 🔯 🔹 Addins 👻	
DRT-PassDA_v4.Rmd × ODDEMIT_KINEO_v3.Rmd × ODMeu_Sessió4_PCA.Rmd × OLab23Mar2018_PCA.Rmd × ODIntroR.Rmd ×	
	🍓 Insert • 🖓 🖑 📑 Run • 🤹 • 🛎
28 + Item 2b	Run Selected Line(s) Ctrl+Enter
30	Run Current Chunk Ctrl+Shift+Enter Run Next Chunk Ctrl+Alt+N
31-# Header 1 32-## Header 2	Bun Setun Chunk
33	The All Churks Above Ctria Alta P
34 25 ```(n)	Run All Chunks Below
36 library(car)	Run All Ctrl+Alt+R
37 data(Davis)	
38 summary(Davis) 39 attributes(Davis)	
40 hist(Davis\$weight,15,col=rainbow(15))	=
41	
42 43	*
41:1 Chunk1 ÷	R Markdown 🗘
Console R Markdown ×	
~/ 🛱	
[1] "sex" "weight" "height" "repwt" "repht"	^
Sclass [1] "data frame"	
\$row.names	1141 1151 1161 1171 1191
$\begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 2 \\ 20 \end{bmatrix} \begin{bmatrix} 2 \\ 21 \end{bmatrix} \begin{bmatrix} 2 \\ 22 \end{bmatrix} \begin{bmatrix} 23 \\ 23 \end{bmatrix} \begin{bmatrix} 24 \\ 25 \end{bmatrix} \begin{bmatrix} 26 \\ 25 \end{bmatrix} \begin{bmatrix} 26 \\ 27 \end{bmatrix} \begin{bmatrix} 28 \\ 28 \end{bmatrix} \begin{bmatrix} 29 \\ 30 \end{bmatrix} \begin{bmatrix} 30 \\ 31 \end{bmatrix}$	"32" "33" "34" "35" "36"
[37] "37" "38" "39" "40" "41" "42" "43" "44" "45" "46" "47" "48" "49"	"50" "51" "52" "53" "54"
[55] "55" "56" "57" "58" "59" "60" "61" "62" "63" "64" "65" "66" "67"	"68" "69" "70" "71" "72" "86" "87" "88" "80" "00"
$\begin{bmatrix} 2 & 5 \end{bmatrix} = \begin{bmatrix} 75 & 74 & 75 \end{bmatrix} = \begin{bmatrix} 76 & 77 & 78 & 79 \end{bmatrix} = \begin{bmatrix} 80 & 81 & 82 & 85 \\ \hline 91 \end{bmatrix} = \begin{bmatrix} 92 & 93 & 94 \end{bmatrix} = \begin{bmatrix} 95 & 96 & 97 & 98 \end{bmatrix} = \begin{bmatrix} 98 & 99 & 100 \\ \hline 91 & 101 & 102 \end{bmatrix} = \begin{bmatrix} 102 & 103 \\ \hline 103 & 103 \end{bmatrix}$	"104" "105" "106" "107" "108"
[109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120" "121"	"122" "123" "124" "125" "126"
$\begin{bmatrix} 127 \end{bmatrix}$ "127" "128" "129" "130" "131" "132" "133" "134" "135" "136" "137" "138" "139" $\begin{bmatrix} 145 \end{bmatrix}$ "145" "146" "147" "148" "149" "150" "151" "152" "152" "154" "155" "156" "157"	"140" "141" "142" "143" "144"
$ \begin{bmatrix} 143 \end{bmatrix} \ 143 \ 144 \ 145 \ 146 \ 147 \ 148 \ 150 \ 151 \ 152 \ 153 \ 154 \ 155 \ 156 \ 157 \ 168 \ 169 \ 170 \ 171 \ 172 \ 171 \ 172 \ 173 \ 174 \ 175 \ 176 \ 177 \$	"176" "177" "178" "179" "180"
[181] "181" "182" "183" "184" "185" "186" "187" "188" "189" "190" "191" "192" "193"	"194" "195" "196" "197" "198"
[199] "199" "200"	
<pre>> hist(Davis\$weight,15,col=rainbow(15))</pre>	Ξ
<pre>> hist(Davis\$weight,15,col=rainbow(15)) > win graph()</pre>	
> writ, graph()	
	·
🚱 🧿 📔 MSCTM 🚺 MSCTM 🔞 RStudio 🧤 🚳 Inbox 🔇 실	Google

To be written in R console or any script or inside a chunk in R Markdown.

To execute a command line included in a script: press <ctrl- Enter>.

To execute several command lines: select and <ctrl- Enter>.

To execute one or several chunks use R Studio menu.

INTRODUCTION TO R: SEQUENCES ...

Example: create a vector with 4 integer elements





INTRODUCTION TO R - BASIC OBJECTS:

Important Objects: lists, vectors, matrices and arrays

R Gui Archivo Editar Paquetes Ventanas Ayuda Image: Image	• Matrices are arrays of 2 dimensions.
<pre>> x<-1:24 > dim(x)<-c(6,4) > x<-matrix(1:24, nrow=6) > rownames(x) <- letters[1:6] > colnames(x)<-c("A","B","C","D") > colnames(x)<-list("A","B","C","D") > y<-x</pre>	 Matrices and arrays of dimension greater than a 2 are allowed.
<pre>> dim(y) <-c(4,3,2) > x A B C D a 1 7 13 19 b 2 8 14 20 c 3 9 15 21 d 4 10 16 22 e 5 11 17 23</pre>	•Related commands: rownames(), colnames(), dim() to check dimensions.
f 6 12 18 24 > y	 To create matrices:
<pre> [, 1] [, 2] [, 3] [1,] 1 5 9 [2,] 2 6 10 [3,] 3 7 11 [4,] 4 8 12 , , 2 [, 1] [, 2] [, 3] [1,] 13 17 21 [1,] 13 17 21 [2,] 14 18 22 [3,] 15 19 23 [4,] 16 20 24] </pre> REVITIONALIDALIDALIDALIDALIDALIDALIDALIDALIDALID	<pre>> x<-matrix(1:24, nrow=6) > rownames(x) <- letters [1:6] > colnames(x)<- c("A","B","C","D") > colnames(x)<- list("A","B","C","D")</pre>

INTRODUCTION TO R - FUNCTIONS AND ARGUMENTS



Functions and arguments:

- An R function might be a mathematic or statistical function, as log(x), but there are additional functions as plot(height, weight).
- Functions have actual parameters (actual arguments) and formal parameters (formal arguments).

• Most arguments have default values and can be omitted.

• R functions arguments can be positionally matched (positional matching) or by name (keyword matching). You can mix positional matching with matching by name.

UNIVERSITAT POLITÈCNICA DE CATALUNYA

BARCELONATECH

nvestigació Operativa



INTRODUCTION TO R - FUNCTIONS AND ARGUMENTS

R Help	for package graphics						For example:
Ocultar Contenido	Atrás Imprimir Opcjones						· · · · · · · · · · · · · · · · · · ·
	? matplot	plot.default(graph	ics)				• plot(neight, weight) is a positional
	? matpoints			The Defa	ult Scatternl	ot Eupotic	
	? mosaicplot ? mosaicplot.default			The Dera	dit scatterpi	OU FUICCIO	matchina call
	? mosaicplot.formula	Descriptio	n				materning can.
	? pairs	-					<pre>> plat(baight waight cal=2) adda a</pre>
	? pairs.default ? pairs.formula	Draw a scatter plo	ot with "decorations" such as	axes and titles in the active	e graphics window.		• plot(neight, weight, col-z) adds a
	2 panel.smooth	Usage					
	? par ? persp						keyword matching argument.
	? persp.default	## Default S3	method:	- NULL TIM - NULL			
	? pie ? piechart	log = "",	main = NULL, sub = N	ULL, xlab = NULL, yl	ab = NULL,		a plat(y=waight y=baight gal=2) allows
	? plot	ann = par	("ann"), axes = TRUE,	frame.plot = axes,			• plot(y=weight, x=height, col=2) allows
	2 plot.data.frame 2 plot.Date	panel.fii	st = NULL, panel.last	. = NULL, asp = NA, .	•••)		
	plot.default	Arguments					arouments in any order (all arouments
	2) plot.design 2) plot.factor	-					a guillents in any order (an a guillents
	2 plot.formula	×, y th	e x and y arguments provide	the x and y coordinates fo	r the plot. Any reasonable way	y of defining the coordi	
-	2 plot.function	de	tails.				by Keyword matching).
-	? plot.new	type 1-	character string giving the typ	pe of plot desired. The follo	owing values are possible, for	details, see <u>plot</u> : "p" i	
	? plot.POSIXct	lin	es, "b", "c") for (empty if " lines	'e") points joined by lines,	"s" and "S" for stair steps an	id "h" for histogram-lik	
-	2 plot.PUSIXIt 2 plot.table	xlim th	unce. e vlimite (v1. v2) of the nlot	Note that with a wall of allow	med and leads to a "reversed.	amie"	
	? plot.window	vlim th	e a limite (a1, a2) of the plot. e a limite of the plot	140te ulat x1 > x2 is allo	wed and leads to a Teversed.	anio .	
-	2 plot.xy 2 points	log	c y minis or inc pior. character string which contai	ne "v" if the v avie is to he	logarithmic "w" if the v avis is	to be logarithmic and	Check narameters in:
	? points.default	main at	main title for the plot-see als		roganamine, y nancyaansa	to be logarithmic and	• check parameters in
-	2 points.formula 2 polugon	sub a	wh title for the plot	• <u>ercre</u> .			
	? rect	xlab al	ahel for the v avis, defaults t	o a description of v			held(blot)
-	? rug	vlab al	abel for the v axis, defaults t	o a description of w			
-	? segments	ann al	orical value indication wheth	er the default annotation (ti	tle and v and v avis labels) sho	uld appear on the plot	anag(plat dafault)
	?) spineplot	axes at	ogical value indicating wheth	er both axes should be dra	wn on the plot. Use graphical	parameter "xaxt." or	urys(pior.ueraurr)
	 spineplot. derault spineplot. formula 	frame.plot a	ogical indicating whether a h	ox should be drawn around	d the plot.	FILLIOPOI AGAO OI	
	? split.screen	panel.first an	expression to be evaluated :	after the plot axes are set u	p but before any plotting takes	s place. This can be use	
	? stars ? stem	panel.last an	expression to be evaluated :	after plotting has taken plac	р		
	? strheight	asp th	v/x aspect ratio, see plot.	. window.			
	? stripchart ? stripplot	ot	her graphical parameters (se	e par and section 'Details'	below).		
	? strwidth				,		
-	2 sunflowerplot	Details					
	? text						
	2) text.default 2) title	Commonly used g	raphical parameters are:				
	? xinch	col					
	2 xyinch	The colors	for lines and points. Multiple	colors can be specified so	that each point can be given it	s own color. If there at	
📗 🗉 🔶 F	Package graphics: Titles	standard fa	shion. Lines will all be plotted	d in the first colour specifie	d.		
		pa .	1 -	1	[_	
🟄 Inicio	🕑 😂 🜉 🔤 🔁 🕒 🥭 🗖	1 🧼 🤻 💋	CUR506-07	🔄 migz	wissio0_sol.doc - Micros	💦 RGui	



INTRODUCTION TO R - FACTORS

R RGui	
Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda	
	Factors:
<pre>> opinio <- sample(seq(1:5), 20, replace = TRUE) > opinio [1] 3 3 2 2 2 5 3 5 1 5 2 5 3 2 4 4 5 2 4 5 > summary(opinio) Min. 1st Qu. Median Mean 3rd Qu. Max. 1.00 2.00 3.00 3.35 5.00 5.00 > > opiniol<- factor(opinio, labels=c("molt desacord","desacord","no sap","d'acord","molt d' > summary(opinio1) molt desacord desacord no sap d'acord molt d'acord 1 6 4 3 6 > > opinio2<-factor(opinio) > opinio2<-factor(opinio) > opinio2<-factor(opinio) > opinio3 [1] 3 3 2 2 2 5 3 5 1 5 2 5 3 2 4 4 5 2 4 5 Levels: 1 2 3 4 5 > opinio3</pre>	 Vectors to represent qualitative variables. Ordered or not. Consider values as levels o labels. To convert labels into numeric values: as.numeric(factor)
<pre>> levels(opinio3)=c("molt desacord","desacord","no sap","d'acord","molt d'acord") > opinio3 [1] no sap no sap desacord desacord desacord molt d'acord no [8] molt d'acord molt desacord molt d'acord desacord molt d'acord no sap d [15] d'acord d'acord molt d'acord desacord d'acord molt d'acord</pre>	io sap lesacord
<pre>Levels: molt desacord desacord no sap d'acord molt d'acord > summary(opinio3) molt desacord desacord no sap d'acord molt d'acord</pre>	<pre>factor(x = character(),</pre>
bg	as.factor(x)
🏄 Inicio 🛛 🙆 🏹 🖾 🏂 💿 🧔 🗂 🦻 🦹 🥑 👔 💋 🔰 🏠 CURSO6-07 🔹 🖓 mlgz 🖉 Sessio0_sol.doc - Micros 🕅 🥂 R.Gui	as.ordered(x)

INTRODUCTION TO R - NEW VARIABLES

💕 R Help for package gra	aphics		Manipulation of data matrices
🖅 🕩 🎒 Ocultar Atrás Imprim	ir Opcjones		manipulation of data matrices
Contenido <u>Ín</u> dice <u>B</u> úsque	eda	plot.default(graphics)	(data.frame) :
? matpoints ? mosaicplot		The Default Scatterplot Function	
? mosaicplot.def ? mosaicplot.form ? mtext	ault nula	Description	 Create a new variable from existent
? pairs ? pairs.default ? pairs.formula		Draw a scatter plot with "decorations" such as axes and titles in the active graphics window.	variables in the current workspace
2 panel.smooth 2 par		Usage	· · · · · · · · · · · · · · · · · · ·
? persp. default ? pie		<pre>## Default S3 method: plot(x, y = NULL, type = "p", xlim = NULL, ylim = NULL,</pre>	using mathematic functions:
piechart plot plot plot.data.frame plot.Date		ann = par("ann"), axes = TRUE, frame.plot = axes, panel.first = NULL, panel.last = NULL, asp = NÅ,)	 For example, y<- log(x)+z+4.5 (x and z
? plot.default ? plot.design ? plot.faster		Arguments	aviatant vactora
2 plot.formula 2 plot.function		x, y the x and y arguments provide the x and y coordinates for the plot. Any reasonable way of defining the coordinates are the states of the	existent vectors.
plot.histogram plot.new plot.POSIXct plot.POSIXIt		type 1-character string giving the type of plot desired. The following values are possible, for details, see <u>plot</u> : "p" 1 lines, "b", "c") for (empty if "c") points joined by lines, "s" and "S" for stair steps and "h" for histogram-lik or lines.	 In a data.frame: attach(Davis)
Plot.tablePlot.window		x1im the x limits $(x1, x2)$ of the plot. Note that $x1 > x2$ is allowed and leads to a "reversed axis".	
? plot.xy ? points ? points.default		$\begin{array}{ll} 1 \text{ or } \\ 1 \text{ or } $	 weight2 <- weight^2 new variable not
2 points.formula 2 polygon		sub a sub title for the plot.	included in Davis data frame
? rect ? rug		xlab a label for the x axis, defaults to a description of x.	included in Duvis duta. // dijie.
? segments ? spineplot ? spineplot.defau	ılt	ann a logical value indicating whether the default annotation (title and x and y axis labels) should appear on the plot axes a logical value indicating whether both axes should be drawn on the plot. Use graphical parameter "xaxt" or	 Davis\$weight2 <- weight^2 new
? spineplot.formu ? split.screen	la	frame.plot a logical indicating whether a box should be drawn around the plot. panel.first an expression to be evaluated after the plot axes are set up but before any plotting takes place. This can be use	vaniable included in Davis data frame
? stars ? stem ? strheight		panel.last an expression to be evaluated after plotting has taken place.	variable included in Davis data. Trame,
? stripchart ? stripplot		 asp the y/x aspect ratio, see <u>plot.window</u>. other graphical parameters (see <u>par</u> and section 'Details' below). 	but a detach(Davis) and new attach(
? strwidth ? sunflowerplot		Details	
? text ? text ? text.default		Commonly used graphical parameters are:	Davis).
2 title 2 xinch 2 xyinch 2 yinch 1 • • • Package graphics:	Titles 💌	col The colors for lines and points. Multiple colors can be specified so that each point can be given its own color. If there ar standard fashion. Lines will all be plotted in the first colour specified.	 Remove an object: rm(object-name).
🏄 Inicio 📔 🎯 🏹 🧱 🛛		i bg	



INTRODUCTION TO R - NEW VARIABLES

- To remove a variable included in a data.frame: Davis\$weight2<-NULL.
- Remove all objects in the current workspace: rm(list=ls()).
- Remove all objects in the current workspace beginning with 'la': rm(list=ls(pattern="la")).
- R can deal with multiple datasets at the same time.
 - You just need to specify the name of the dataset and a "\$" symbol before each variable name.
 - \odot If you don't want to write again

and again the name of the dataset as a prefix for each variable, you can use attach()

ls()																	
l] "davis" # M'intere # Vec les	"la ssa la cla columnes c	ast.warn Asse de Jue cont	ing" "d davis é (cara	opinio' (és un acterí:	" data.∷ stique:	opin" frame o de le	niol" o matr es obs	"(iu de d ervacio	o <mark>pinio</mark> dades) ons)	2"	"opir	nio3"					
attribute: names	(davis)																
] "id"	"sex"	"we	ight"	"hei	ght"	"r_we:	ight"	"r_hei	ght"								
ow.names																	
[1] "1"	"2" "3"	"4"	"5"	"6"	"7"	"8"	"9"	"10"	"11"	"12"	"13"	"14"	"15"	"16"	"17"	"18"	
19] "19"	"20" "21'	"22"	"23"	"24"	"25"	"26"	"27"	"28"	"29"	"30"	"31"	"32"	"33"	"34"	"35"	"36"	
37] "37"	"38" "39'	"40"	"41"	"42"	"43"	"44"	"45"	"46"	"47"	"48"	"49"	"50"	"51"	"52"	"53"	"54"	
55] "55"	"56" "57'	' "58"	"59"	"60"	"61"	"62"	"63"	"64"	"65"	"66"	"67"	"68"	"69"	"70"	"71"	"72"	
73] "73"	"74" "75'	"76"	"77"	"78"	"79"	"80"	"81"	"82"	"83"	"84"	"85"	"86"	"87"	"88"	"89"	"90"	
91] "91"	"92" "93'	' "94"	"95"	"96"	"97"	"98"	"99"	"100"	"101"	"102"	"103"	"104"	"105"	"106"	"107"	"108"	
09] "109"	"110" "111	." "112"	"113"	"114"	"115"	"116"	"11/"	"118"	"119"	"120"	"121"	"122"	"123"	"124"	"125"	"126"	
27] "127"	"128" "129	9" "130" N N140N	"131"	"132"	.133.	"134"	.135.	.136.	"137"	"138"	.139.	"140"	1411	"142"	"143"	"144"	
45] "145" 621 #162#	"146" "14.	1148° 1148°	"149"	"150"	"151"	"152"	.123.	"154"	"155"	11741	"15/" "175"	11761	11220	"160"	.101.	"162"	
63] "163" 611 #161#	"104" "10;	0" "100" 01 11041	11051	1100	11071	11001	11001	11001	11011	11021	11021	11041	11051	11061	11071	"100" "100"	
991 "199"	"200") 104	105	100	107	100	105	190	191	192	195	194	195	190	191	190	
lass																	
] "data.fr	rame"																
# O només	les columr	nes															
names(davi	.s)																
] "id"	"sex"	"we	ight"	"heig	ght"	"r_we:	ight"	"r_hei	ght"								
# Quin és	el nb d'ob	servaci	ons: di	imensi	5 file:	5											
dim(davi:	5)																
] 200 6																	
dim(davi:	5)[1]																
] 200																	
# NO es po	ou crear el	. quadra	r aet l	pes com	n a va:	гтарте	no es	VISID.	те								
pesz <- we	erdur.5 So "mojapti	. no ong	ontrod	_													
ror: objet # La varia	blo ág mig	ible ro	ferenc	J iada d:	ing do:	ta fro	no dott	ie									
т ша valia nes2 <= da	wisŚweicht	-v5 -v5	rerent.	raua u.	ua ua	ca.rrd	ae uev	10									
$\rho c \rho z \sim u c$		- 4															
180																	



INTRODUCTION TO R - SCOPE OF VISIBILITY: ATTACH COMMAND

<pre>Note that the second of t</pre>										
<pre>> # Podem fer visibles totes les variables d'un data.frame > attach(davis) > # Ara es pot crear una nova variable fora del data.frame > pes2 <- weight^2 > ls() [1] "davis" "last.warning" "opinio" "opinio1" "opinio2" "opinio3" "pes2" > detach(davis) > # Si es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis) id sex weight height r_weight r_height pes2 Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. : 148.0 Min. : 1521 lst Qu:: 50.75 M: 88 lst Qu:: 55.0 lst Qu::164.0 lst Qu: 55.00 lst Qu::160.5 lst Qu:: 3025 Median :100.50 Median : 63.0 Median : 169.5 Median : 63.00 Median : 168.0 Median : 3969 Mean : 100.50 Mean : 65.8 Mean ::17.00 Mean : 65.62 Mean : 168.5 Mean : 44566 3rd Qu::150.25 3rd Qu:: 74.0 3rd Qu::177.2 3rd Qu:: 73.50 3rd Qu::175.0 3rd Qu:: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 NA's : 17.00 NA's : 17.00 > # o bé: > davis\$pes2<- davis\$weight > # Esborrar una columna d'un data.frame > davis\$pes2<- nuLL > # Noms de les caracteristiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	Reduce Reconsole	Ventanas Avuda								
<pre>> # Podem fer visibles totes les variables d'un data.frame > attach(davis) > # Ara es pot crear una nova variable fora del data.frame > pes2 <- weight^2 > ls() [1] "davis" "last.warning" "opinio" "opinio1" "opinio2" "opinio3" "pes2" > detach(davis) > # Si es vol crear dins del data.frame davis > davis<- transform(davis, pes2=weight^2) > summary(davis) id sex weight height r_weight r_height pes2 Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. : 148.0 Min. : 1521 lst gu.: 50.75 M: 88 lst gu.: 55.0 lst gu.:164.0 lst gu.: 55.00 lst gu.: 160.5 lst gu.: 3025 Median :100.50 Median : 63.0 Median : 63.00 Median : 63.00 Median : 168.0 Median : 3969 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd gu.: 150.25 3rd gu.: 74.0 3rd gu.:177.2 3rd gu.: 73.50 3rd gu.: 157.0 Max. :200.00 Max. :166.0 Max. :197.0 Max : 124.00 Max. :27556 NA's : 17.00 NA's : 17.0 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > m(pes2) > # Esborrar objecte de l'espai de treball: rm() > # f Sborrar una columna d'un data.frame > davis\$pes2<= NULL > # noms de les caracteristiques (variables) d'un data.frame > names (davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>										
<pre>> # Podem fer visibles totes les variables d'un data.frame > attach(davis) > # Ara es pot crear una nova variable fora del data.frame > pes2 <- weight^2 > ls() [1] "davis" "last.warning" "opinio" "opinio1" "opinio2" "opinio3" "pes2" > detach(davis) > # si es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis) id sex weight height r_weight r_height pes2 Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. :148.0 Min. : 1521 lst qu.: 50.75 M: 88 lst qu.: 55.0 lst qu.:164.0 lst qu.: 55.00 lst qu.:160.5 lst qu.: 3025 Median :100.50 Median : 63.0 Median :169.5 Median : 63.00 Median : 168.5 Mean : 4556 3rd qu.:150.25 3rd qu.: 74.0 3rd qu.:177.2 3rd qu.: 73.50 3rd qu.: 175.0 3rd qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 Na's : 17.00 Na's : 17.00 > # o bé: > davis\$pes2<- davis\$weight > # Esborrar object de l'espai de treball: rm() > rm(pes2) > # Bsborrar object de l'espai de treball: rm() > # fo sborrar una columna d'un data.frame > davis\$pes2<- NULL > # moms de les caracteristiques (variables) d'un data.frame > names (davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>										
<pre>> # totach (davis) > # trach (davis) > # Ara es pot crear una nova variable fora del data.frame > pes2 <- weight^2 > ls() [1] "davis" "last.warning" "opinio" "opiniol" "opinio2" "opinio3" "pes2" > detach (davis) # \$i es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis) id sex weight height r_weight r_height pes2 Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. :148.0 Min. : 1521 lst Qu.: 50.75 M: 88 lst Qu.: 55.0 lst Qu.:164.0 lst Qu.: 55.00 lst Qu.: 100.5 Median :100.50 Median : 63.0 Median : 63.0 Median : 168.0 Median : 3969 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd Qu.:150.25 3rd Qu.: 74.0 3rd Qu.:177.2 3rd Qu.: 73.50 3rd Qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 NA's : 17.00 NA's : 17.0 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) # # Seborrar una columna d'un data.frame > dawis\$pes2<- NULL > # Noms de les caracteristiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> # Podem fer visibles totes les variables d'un data frame									
<pre>> # Attach(davis) > # Atta es pot crear una nova variable fora del data.frame > pes2 <- weight^2 > ls() [1] "davis" "last.warning" "opinio" "opinio1" "opinio2" "opinio3" "pes2" > detach(davis) > # Si es vol crear dins del data.frame davis > davis<transform()<br="" davis,="" pes2="weight^2">> summary(davis)</transform(></pre>	> attach(davis)									
<pre>> # Ref Construction of the new of the</pre>	> # Ara es not crear una nova variable fora del data frame									
<pre>> ls() [1] "davis" "last.warning" "opinio" "opinio1" "opinio2" "opinio3" "pes2" > detach(davis) > # \$i es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis)</pre>	<pre>> # Ara es por crear una nova variabre rora der data.rralle > pes2 <= weight^2</pre>									
<pre>[1] "davis" "last.warning" "opinio" "opinio1" "opinio2" "opinio3" "pes2" > detach(davis) > # Si es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis)</pre>	> 1s()									
<pre>> detach(davis) > # Si es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis) id sex weight height r_weight r_height pes2 Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. : 1140.0 Min. : 1521 1st qu.: 50.75 M: 88 1st qu.: 55.0 1st qu.:164.0 1st qu.: 55.00 1st qu.:160.5 1st qu.: 3025 Median :100.50 Median : 63.0 Median : 163.5 Median : 63.00 Median : 168.0 Median : 3369 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd qu.:150.25 3rd qu.: 74.0 3rd qu.:177.2 3rd qu.: 73.50 3rd qu.:175.0 3rd qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 NA's : 17.00 NA's : 17.0 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	[1] "davis" "la	st.warning" "opinio	o" "opinio"	l" "opinio2"	"opinio3"	"pes2"				
<pre>> # Si es vol crear dins del data.frame davis > davis<-transform(davis, pes2=weight^2) > summary(davis)</pre>	> detach(davis)	5 1	1	1	1	1				
<pre>> davis<-transform(davis, pes2=weight^2) > summary(davis)</pre>	> # Si es vol crear di	ns del data.frame d	lavis							
<pre>> summary(davis) id sex weight height r_weight r_height pes2 Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. :148.0 Min. : 1521 1st qu.: 50.75 M: 88 1st qu.: 55.0 1st qu.:164.0 1st qu.: 55.00 1st qu.:160.5 1st qu.: 3025 Median :100.50 Median : 63.0 Median :169.5 Median : 63.00 Median :168.0 Median : 3969 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd qu.:150.25 3rd qu.: 74.0 3rd qu.:177.2 3rd qu.: 73.50 3rd qu.:175.0 3rd qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 NA's : 17.00 NA's : 17.0 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> davis<-transform(dav	vis, pes2=weight^2)							
<pre>id sex weight height r_weight r_height min. :</pre>	> summary(davis)									
<pre>Min. : 1.00 F:112 Min. : 39.0 Min. : 57.0 Min. : 41.00 Min. :148.0 Min. : 1521 1st gu.: 50.75 M: 88 1st gu.: 55.0 1st gu.:164.0 1st gu.: 55.00 1st gu.:160.5 1st gu.: 3025 Median :100.50 Median : 63.0 Median :169.5 Median : 63.00 Median :168.0 Median : 3969 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd gu.:150.25 3rd gu.: 74.0 3rd gu.:177.2 3rd gu.: 73.50 3rd gu.:175.0 Max. :27556 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	id sex	weight	height	r_weight	r_height	pes2				
<pre>1st Qu.: 50.75 M: 88 1st Qu.: 55.0 1st Qu.:164.0 1st Qu.: 55.00 1st Qu.:160.5 1st Qu.: 3025 Median :100.50 Median : 63.0 Median :169.5 Median : 63.00 Median :168.0 Median : 3969 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd Qu.:150.25 3rd Qu.: 74.0 3rd Qu.:177.2 3rd Qu.: 73.50 3rd Qu.:175.0 3rd Qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	Min. : 1.00 F:11	2 Min. : 39.0	Min. : 57.0	Min. : 41.00	Min. :148.0	Min. : 1521				
Median :100.50 Median : 63.0 Median :169.5 Median : 63.00 Median :168.0 Median : 3969 Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd Qu.:150.25 3rd Qu.: 74.0 3rd Qu.:177.2 3rd Qu.: 73.50 3rd Qu.:175.0 3rd Qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 > # 0 bé: NA's : 17.00 NA's : 17.00 NA's : 17.0 > # Esborrar objecte de l'espai de treball: rm() Na's : 17.00 NA's : 17.0 > m(pes2) # Esborrar una columna d'un data.frame davis\$pes2<- NULL	1st Qu.: 50.75 M: 8	3 1st Qu.: 55.0	1st Qu.:164.0	1st Qu.: 55.00	1st Qu.:160.5	1st Qu.: 3025				
Mean :100.50 Mean : 65.8 Mean :170.0 Mean : 65.62 Mean :168.5 Mean : 4556 3rd Qu.:150.25 3rd Qu.: 74.0 3rd Qu.:177.2 3rd Qu.: 73.50 3rd Qu.:175.0 3rd Qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 > # 0 bé: NA's : 17.00 NA's : 17.00 NA's : 17.0 > # constant objecte 0 1'espaid true :17.00 NA's : 17.0 > # Esborrar objecte 0 'espaid true :m() :	Median :100.50	Median : 63.0	Median :169.5	Median : 63.00	Median :168.0	Median : 3969				
3rd Qu.:150.25 3rd Qu.: 74.0 3rd Qu.:177.2 3rd Qu.: 73.50 3rd Qu.:175.0 3rd Qu.: 5476 Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 > # 0 bé: NA's : 17.00 NA's : 17.00 NA's : 17.0 > # 0 bé: vis\$pes2<- davis\$weight	Mean :100.50	Mean : 65.8	Mean :170.0	Mean : 65.62	Mean :168.5	Mean : 4556				
<pre>Max. :200.00 Max. :166.0 Max. :197.0 Max. :124.00 Max. :200.0 Max. :27556 NA's : 17.00 NA's : 17.0 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	3rd Qu.:150.25	3rd Qu.: 74.0	3rd Qu.:177.2	3rd Qu.: 73.50	3rd Qu.:175.0	3rd Qu.: 5476				
<pre>NA's : 17.00 NA's : 17.0 > # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	Max. :200.00	Max. :166.0	Max. :197.0	Max. :124.00	Max. :200.0	Max. :27556				
<pre>> # 0 bé: > davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>				NA's : 17.00	NA'S : 17.0					
<pre>> davis\$pes2<- davis\$weight > # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> # 0 bé:									
<pre>> # Esborrar objecte de l'espai de treball: rm() > rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> davis\$pes2<- davis\$w	eight								
<pre>> rm(pes2) > # Esborrar una columna d'un data.frame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> # Esborrar objecte de	e l'espai de trebal	LL: rm()							
<pre>> # Esporrar una columna d'un data.Irame > davis\$pes2<- NULL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> rm(pes2)									
<pre>> davis\$pes2<- NoLL > # Noms de les característiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> # Esborrar una columna d'un data.frame									
<pre>> # Woms de les caracteristiques (variables) d'un data.frame > names(davis) [1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> davis\$pes2<- NULL									
<pre>[1] "id" "sex" "weight" "height" "r_weight" "r_height"</pre>	> # Noms de les caracteristiques (variables) d'un data.frame									
Li ia sex weight height i_weight i_height	<pre>> names(davis) [1] "id" "could be install "boight" "boight" "boight""</pre>									
	(I) IU SEX	werdur uner	rduc r_werduu	r_nerduc						

- attach() command can be dangerous. Use detach() as soon as possible.
- Suggested command: Evaluate an R expression in an environment constructed from

data, possibly modifying the original data:

with(Davis,{boxplot(height);summary(height)})



INTRODUCTION TO R - NEW VARIABLES

RGui - [R Console] R Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda C C C C C C C C C C C C C C C C C C C		e × e ×
<pre>> names(davis) [1] "id" "sex" "weight" "height" > # Vull un data.frame reduit sense pes i alça > davis1 <- davis[,1:4] > attributes(davis1) Cremes</pre>	"r_weight" "r_height" da reportat	
<pre>snames [1] "id" "sex" "weight" "height"</pre>	Access to columns in a data frame as if th	ere were matri
\$class [1] "data.frame"		
<pre>\$row.names [1] "1" "2" "3" "4" "5" "6" "7" [19] "19" "20" "21" "22" "23" "24" "25" [37] "37" "38" "39" "40" "41" "42" "43" [55] "55" "56" "57" "58" "59" "60" "61" [73] "73" "74" "75" "76" "77" "78" "79" [91] "91" "92" "93" "94" "95" "96" "97" [109] "109" "110" "111" "112" "113" "114" "115 [127] "127" "128" "129" "130" "131" "132" "133 [145] "145" "146" "147" "148" "149" "150" "151 [163] "163" "164" "165" "166" "167" "168" "169 [181] "181" "182" "183" "184" "185" "186" "187</pre>	"8" "9" "10" "11" "12" "13" "14" "15" "16" "17" "18" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35" "36" "44" "45" "46" "47" "48" "49" "50" "51" "52" "53" "54" "62" "63" "64" "65" "66" "67" "68" "69" "70" "71" "72" "80" "81" "82" "83" "84" "85" "86" "87" "88" "89" "90" "98" "99" "100" "101" "102" "103" "104" "105" "106" "107" "108" "116" "117" "118" "119" "120" "121" "122" "123" "124" "125" "126" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144" "152" "153" "154" "155"	
<pre>> names(davis1) [1] "id" "sex" "weight" "height" > # Vull un data.frame pels homes sex == M i u > > homes <- (davis\$sex=='M') > davisM <- davis[homes,] > #davisM > davisF <- davis[davis\$sex=='F',]</pre>	n altre per dones amb totes les característiques	
<pre>> dim(davisM) [1] 88 6 > dim(davisF) [1] 112 6 > # Per quedar-me amb les observacions 20 a 11 > davis2<- davis[c(20:110,119), 1:4] > dim(davis2) [1] 92 4</pre>	0 més 119 i les 4 columnes:	



INTRODUCTION TO R - INDEXING VARIABLES

- Indexing vectors?: weight2[29] position 29 in weight2 vector.
- Indexing matrices?: Davis[2,4] observation 2 and variable in 4th column.
- Row number 2 in a data.frame: Davis[2,].
- Column number 4 in a data.frame : Davis[, 4] (height is a vector with 200 observations).
- A set of columns: Davis[, c(1,3:4)].
- A set of rows (observations):
 - Davis[1:100,] observations 1, 2, 3 ... 100
 - Davis[seq(1,100,2),] ... observations 1, 3, 5, 7 ...
 - Davis[sample(100:200,50,rep=T),] 50 random rows contained in row numbers 1 tp 100.
 - Davis[rep(c(1,2),10),]
 observations (repeated)

1212121212121212121212

```
> davis3<- Davis[ sample(100:200,10,rep=T), ]
> table(Davis3$id)
104 105 141 173 174 175 177 180 194
1 1 1 1 1 1 2 1 1
```





INTRODUCTION TO R - RECODIFICATION OF VARIABLES

R Console

```
> summary( davis$weight )
   Min. 1st Qu. Median
                            Mean 3rd Qu.
                                            Max.
   39.0
           55.0
                   63.0
                            65.8
                                    74.0
                                           166.0
> davis$tipus <- factor(cut(davis$weight, 4)) # Discretització en 4 intervals</p>
> table(davis$tipus)
(38.9,70.7) (70.7,102)
                           (102, 134]
                                       (134, 166)
        142
                     55
                                   2
                                               1
> summary( davis$tipus |
(38.9,70.7] (70.7,102]
                           (102, 134]
                                       (134,166)
        142
                      55
                                   -2
> tapply( davis$weight, davis$tipus, median)
(38.9,70.7] (70.7,102]
                           (102, 134]
                                       (134,166)
         58
                      80
                                 111
                                              166
> # Discretització per 4 guartils
> davis$tipus <- factor(cut(davis$weight, quantile(davis$weight,c(0,1/4,2/4,3/4,1))))</pre>
> table(davis$tipus)
 (39,55] (55,63] (63,74] (74,166]
      52
               50
                         48
                                  49
> tapply( davis$weight, davis$tipus, median)
 (39,55] (55,63] (63,74] (74,166]
      52
               59
                         68
                                  82
> # Discretització en 4 intervals triats per l'usuari
> davis$tipus <- factor(cut(davis$weight, breaks=c(-1,55,65,75,200)))</p>
> table(davis$tipus)
 (-1,55] (55,65] (65,75] (75,200]
      53
                         38
               64
                                  45
> tapply( davis$weight, davis$tipus, median)
 (-1,55] (55,65] (65,75] (75,200]
      52
               61
                         69
                                  83
> levels(davis$tipus)<-paste("TYPE",levels(davis$tipus), sep=":")</pre>
> summary(davis$tipus)
TYPE: (-1,55] TYPE: (55,65] TYPE: (65,75] TYPE: (75,200]
           53
                          64
                                        38
                                                       45
> levels(davis$tipus) <- c("prim","normal","sobrepes","obes")</p>
> summary(davis$tipus)
    prim normal sobrepes
                                obes
      53
               64
                         38
                                  45
>
```

Recodification: Create a new variable from an existent numerical one.

• Discretization of a numeric variable:

- Equal length intervals.
- Intervals selected by the users.
- Intervals defined by quantiles.



INTRODUCTION TO R - DEFINING FACTORS

Recoding: Creating a new variable by working with ranges.

• Grouping categories: create a new variable using ifelse() sentence.

```
as.numeric( davis$tipus )
  [1] 4 2 1 3 2
               44
                   3 3 2 3 4
                               2
                                    2
                                        2
                                           2 4 2 2 3 1 1 2 2
                                                             1 4 3 3 4 1 3 2 2
                                                                                   2
                                                                                               1 3 1 1
                                         4
                                                                               3
                                             4 2 3 1 2 1 1 1 2 3 3 2 1 1 2 2
[50] 3 4 2 3 4 2 2 3
                     3
                         2
                                                                            2 2 1 2
                                                                                     2
                                                                                       3
                       3
                           3
                             -3
                                         2
                                           1
                                                                                                   4 1
                                             3 2 4 3 3 1 3
                                                           1 1 1 3 1 2 4 2 1 4 1 2 2 4
[99] 1 1 1 2
             1 1
                 2
                   1
                     1
                       2
                         2
                           2
                             4
                                                                                       4
                                                                                         3
                                                                                           1 1 1
                                                                                                 2 1 1
[148] 2 3 4 2 1 1 1
                                     2 2 2 2 3 4 1 4 1 2
                                                           2 3 2 3 4 3 4
                                                                             23222
                   1
                       3 2 2 2 1 1
                                   1
                                                         - 4
                                                                         1 4
                                                                                       4 1 4 4
                                                                                              2123
[197] 4 4 4 4
> grup <- rep( 0, dim( davis )[1] )</pre>
  grup <- factor(ifelse( as.numeric(davis$tipus)>2,1,0))
  levels(grup) <- c("correcte","controlar")</pre>
> summary(grup)
correcte controlar
      117
                   83
```



INTRODUCTION TO R - EDA - BIVARIATE: NUMERIC VS FACTOR

TWO VARIABLES ARE INVOLVED:

RESPONSE VARIABLE IS NUMERIC, as Davis\$height EXPLANATORY VARIABLE IS A FACTOR, as Davis\$sex (max 5-6 levels)

Goal: Do groups defined by levels of the factor determine a difference profile in the numeric response.

- Do height and sex show an independent behavior/profile? Statistical question: Is the profile of height the same for both levels of factor sex?
- If height and sex don't show any relationship- Statistical statement: The profile of height is the same for both levels in sex factor ?

EDA for a numeric variable according to groups defined by factor. Particular analysis: ANOVA - Analysis of Variance

INTRODUCTION TO R - EDA - BIVARIATE: NUMERIC VS FACTOR

For example: Height (Y) vs Sex (A) - Formula expression in R: $Y \sim A$



library(car) data(Davis)	<pre>summary(Davis\$height) with(Davis.Boxplot(height~sex.pch=19.</pre>
names(Davis)	<pre>col=rainbow(2),main="Height vs Sex"))</pre>
head(Davis)	
<pre>with(Davis,tapply(height,sex,summary))</pre>	

4. INTRODUCTION TO R - EDA - BIVARIATE: NUMERIC VS FACTOR





5. EDA - BIVARIATE: 2 NUMERICS Y VS X

5.1 Numeric statistics to assess linear relationship between Y and X

Covariance, COV(y,x)=COV(x,y), defined as E(YX) - E(X)E(Y)

• Disadvantage: Depends on units, so not direct interpretation

Pearson's coefficient of correlation, suitable for assessment in normal data

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$
 and $\sigma_X = \sqrt{Var(X)} \sigma_Y = \sqrt{Var(Y)}$

- Advantage: Adimensional, no affected by units
 - $\circ
 ho(X, Y)$ range is $\begin{bmatrix} -1, 1 \end{bmatrix}$.
 - $\rho(X, Y) > 0$ means positive relationship X and Y.
 - $\rho(X, Y) < 0$ means negative relationship X and Y,.
 - $\rho(X, Y) = 0$ indicates uncorrelated variables, not equivalent to independence.
 - If Y = aX + b then $| \rho(X, Y) | = 1$.
- Spearman's coefficient of correlation, is a nonparametric measure of statistical dependence.

EDA - BIVARIATE: 2 NUMERICS Y VS X

In R, use var(Davis[,2:3]) or try with Census Data data("CPS1985") in library AER.

```
> library(AER)
> data("CPS1985")
> df<-CPS1985
> 1s()
[1] "CPS1985" "df"
> dim( df ) # dimensions: rows and columns
[1] 534 11
> summary( df )
                    education
                                     experience
                                                                       ethnicitv
                                                                                     region
                                                                                                  aender
                                                                                                                  occupation
      wage
                                                        age
                                                                                               malē :289
                                                                                                                       :156
       : 1.000
                        : 2.00
                                   Min. : 0.00
                                                           :18.00
                                                                            :440
                                                                                   south:156
                                                                                                             worker
 Min.
                  Min.
                                                   Min.
                                                                    cauc
 1st Qu.: 5.250
                  1st Qu.:12.00
                                   1st Qu.: 8.00
                                                   1st Qu.:28.00
                                                                    hispanic: 27
                                                                                   other:378
                                                                                               female:245
                                                                                                             technical :105
 Median : 7.780
                  Median :12.00
                                   Median :15.00
                                                   Median :35.00
                                                                    other
                                                                                                                       : 83
                                                                          : 67
                                                                                                             services
       : 9.024
                  Mean
                         :13.02
                                   Mean
                                          :17.82
                                                   Mean
                                                          :36.83
                                                                                                             office
                                                                                                                       : 97
 Mean
 3rd Ou.:11.250
                  3rd Qu.:15.00
                                   3rd Ou.:26.00
                                                   3rd Qu.:44.00
                                                                                                             sales
                                                                                                                       : 38
                        :18.00
Max.
       :44.500
                 Max.
                                  Max.
                                         :55.00
                                                  Max.
                                                         :64.00
                               married
           sector
                     union
                                                                                       Wage(Y) vs Education (X) | Race
                               no :184
 manufacturing: 99
                     no :438
 construction: 24
                     yes: 96
                               ves:350
                                                                        cauc
 other
              :411

    hispanic

                                                                     4

    other

> attach( df )
  # Bivariate analysis: 2 numeric variables
>
                                                                     8
> plot(education,wage,col=as.numeric(ethnicity)+1,
       main="Wage(Y) vs Education (X) | Race", pch=19)
                                                                  wage
> legend("topleft",legend=levels(ethnicity),col=2:4,
                                                                    8
        pch=19)
> cor(wage,education,method="spearman")
[1] 0.3813425
                                                                     9
> cor(wage,education,method="pearson") # The one defined i
[1] 0.3819221
                                                                     0
Nicer option: scatterplot, try in lab session
                                                                                 5
                                                                                                10
                                                                                                               15
                                                                                              education
> library(car)
  scatterplot(wage~education|ethnicity,main="Wage(Y) vs Education (X)
                                                                                        Race'', smooth=FALSE)
```



6. EDA - BIVARIATE: 2 FACTORS, A AND B

6.1 Numeric statistics to assess linear relationship A and B

Non-existent. Analysis of Contingency Tables and classical inference test to assess Independence of both factors using Chi-Squared Test: chisq.test() in R, arguments a contingency table.





EDA - BIVARIATE: 2 FACTORS, A AND B

Graphic display (default in R): mosaic plot

More than 2 dimensions: use xtabs() command in R

