Math 441- Chp0

Introductory Statistics with R-Studio (and R)
[By M-C Koissi]

Installing R-Studio:

R-Studio is user-friendly interface of R, and it's open-source (just as R)
First, you need to install R, before you can use R-Studio

To install R:

- Go to Google.com and enter "R software"
- Choose: The R Project for Statistical Computing
  https://www.r-project.org/
- On the left side, under "Download", select "CRAN"
- Choose the location closest to you
- If you are in WI, go under USA and choose something like
  https://mirror.las.iastate.edu/CRAN/ Iowa State University, Ames, IA

Follow the steps, and you should be able to install R easily.

To install RStudio:

- Go to https://www.rstudio.com/
- Select: Rstudio > Download
- On the page "Choose Your Version of RStudio", choose "Studio Desktop" FREE > Download > "Installers for Supported Platforms" (Choice the appropriate plateform)

Short overview of R-Studio screen:

- Console, workspace, history, files, plots, packages, help, etc,…
- Start:
  - Set / change your working directory
  - Open a new script: File -> New -> R Script

Part I - Using R as a Calculator

1) Basic Arithmetic with R

Operator	Description
+	addition
-	subtraction
*	multiplication
/	division
^ or **	exponential
x % % y	modulus (x mod y) 5 % %2 is 1
x % / % y	integer division 5 % / %2 is 2

To Do: Compute the difference between 2015 and the year you started at UWEC and divide this by the difference between 2015 and the year you were born (use 2014 if you start here in 2015). Multiply this with 100 to get the percentage of your life you have spent at this university. Use brackets if you need them.

2) Vectors and matrices

2.1. Vectors and functions

R has many built-in functions. Most R functions work on vectors (example "diff", "mean", "range", "max", "min", "length", "sort", "cumsum"). Some functions create vectors.

- "**seq**" function produces arithmetic sequences.
  > seq (from = …., to = …., by = ….)
- **"rep"** function repeats the value x .
  > rep (x, times)

To do:

a) Write the script to create the following vector v1=(103, 104, 105, 106, 107, 108, 109, 110)

b) Write the script to create the following sequence: v2=(13, 11, 9, 7, 5, 3, 1)

c) Write a script to create the following vector v3 = (4 4 4 4 4 1 3 5 7 9 11)

d) Find the mean of the sum of these three vestors.

To sort a sequence of numbers in increasing / decreasing order

-Sort(u) gives the vector u with component in increasing order

-Sort (u, decreasing =TRUE) gives the vector with component in decreasing order

-The "rev" function provides a reversed version of its argument.

So "rev(sort(u))" will also give the vector with component in decreasing order

Elements in vectors can be addressed by standard [i]

2.2. Matrices

Matrices can be seen as 2-dimensional vectors. To define a matrix, use the function matrix:

Example : mat=matrix(data=c(9,2,3,4,5,6),ncol=3)
The argument data specifies which numbers should be in the matrix. Use either *ncol* to specify the number of columns or *nrow* to specify the number of rows.

To Do : Put the numbers 31 to 60 in a vector named P and in a matrix with 6 rows and 5 columns named Q.

2.3. Data frames

A data frame is a matrix with names above the columns.

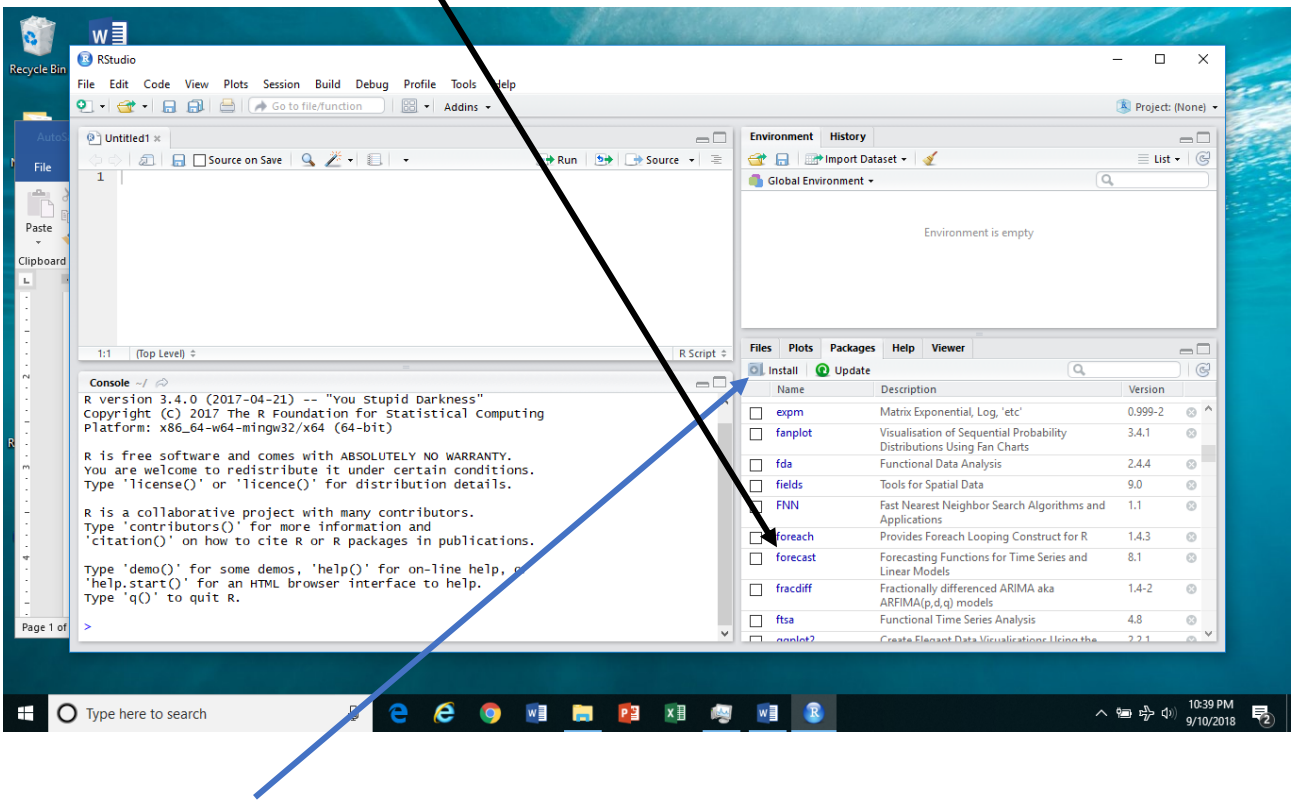
Part 2: Reading a dataset

A data set can be read from several sources: either from the computer where the data has been saved or from an open source (R has a lot of freely available dataset).

1) **Reading data from R sources**: Packages such as MASS or ggplot are addition to R, and are freely available.

Example: [from package "MASS] dataset "Boston"

i. MASS is a package that usually comes with R. [check the list of packages in the low right corner



If the library MASS is not listed, it can easily be installed:

- Click on "install" and
- enter MASS under "Packages"
- click on "Install dependencies", to install MASS

We want, for example, to use the data set "Boston" from the library (package) MASS

```
> head(Boston) # this gives you the first six lines of your data
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
```

 # First look at data and basic calculations

```
library(MASS)          #to get the package MASS
head(Boston)           #to see the first rows of the data
Boston$crim            #  to look a the column "crim" from the data set Boston
mean(Boston$crim)      # to find the mean number of crimes in Boston
```

To Do:

a. Read the data set " Cars93" from the MASS library
b. Show the first lines of the data.
c. How many variables (= columns) does the data have ?
d. Compute the mean weight of the cars in this data set.

2) Reading data from various formats (txt, csv, SPSS, Stata, SAS,…)

i. read.OPTIONS ("filename.OPTIONS", header=TRUE)
   OPTIONS can be replaced by
   - table (for txt files),
   - csv (for excels files), spss, …

   you can also use the following code to read a file.

ii. read.OPTIONS(choose.files(), header=TRUE) or

Example:

# To read the data as a txt file: read.table("filename.txt", header=TRUE) or read.table(choose.files(), header=TRUE, sep=",")	# To read the data as an texcels file read.csv("filename.csv", header=TRUE) or read.csv(choose.files(),header=TRUE)

To Do:

a. Read the data set " Cars93" from the MASS library
b. Show the first lines of the data.
c. How many variables (= columns) does the data have ?
d. Compute the mean weight of the cars in this data set.

To Do:

a. Upload and Read the data set " Wisconsin lottery Sales" available on Canvas
b. Show the first lines of the data.
c. Compute the mean SALES.

To Do:

d. Read the data set " DECAY"(available on Canvas) from your textbook problem #1.46
e. Show the first lines of the data.
f. Compute the mean RATE.