R Bootcamp

Sandeep Gangarapu
11 Sep, 2019

Adapted from Scott Schanke's presentation for R Bootcamp 2018

About Me

- 4th Year PhD student in Information and Decision Sciences
- Research on Online Experimentation, Targeting strategies for Marketing
- Expert in R, Python, SQL
- Instructor for IDSc 3103 (Data Modeling and Databases - Spring 2020)
- Previously worked at Google Inc. in Fraud Prevention team as a Data Scientist

How many of you use Microsoft Excel?

Problems with Excel

- Speed & Scale
  - Excel can handle a maximum of 16k columns and 1M rows, and will drastically slow down with complex calculations
  - R is practically limited only by the amount of RAM that your computer has, and easily handles multi-million row datasets and calculations
- Reproducibility
  - In Excel, you have one file which contains your data as well as any manipulations of it
  - In R, you have a data file and a R Code file. The data file stays the same, while the R Code specifies the analysis
  - If you share your R Code with a friend and they run it, they will get the same results, every time.

R Language

What is R and Why is it important?

“R is an integrated suite of software facilities for data manipulation, calculation and graphical display”

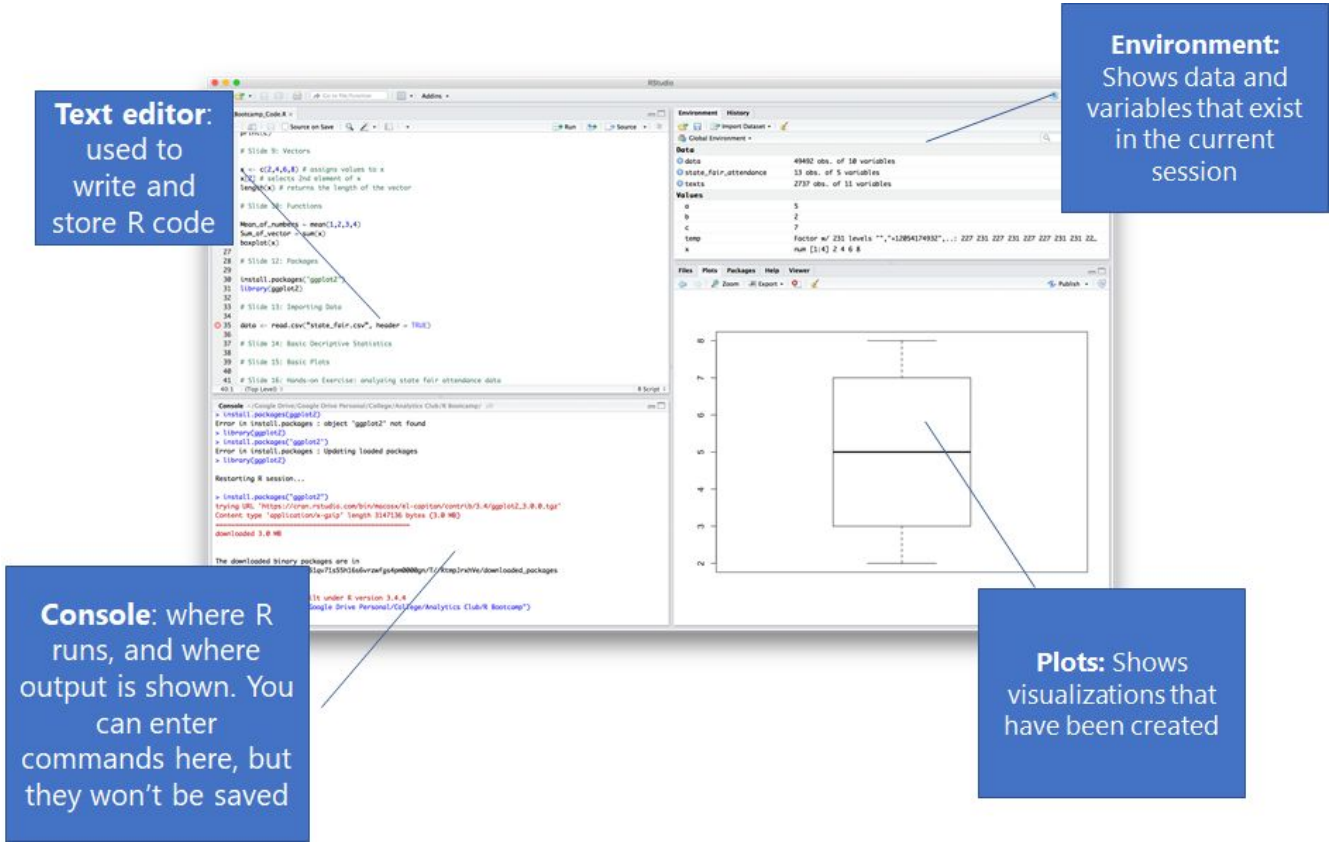
What does this mean?

- R is a scripting language that lets you work with data much like you would in Excel
- However, all the work on the data is done through code, rather than clicking on menus
- R is widely used in both business and academic settings because it allows for reproducibility

R and R Studio: better together

- R is the underlying programming language
- RStudio is an application with added features that makes it easier to work with R
- Both are open-source, meaning they are freely available and can be used for any purpose

An overview of RStudio



Basic calculations in R

- R can be used as a basic calculator by entering expressions directly into the command line
- Parenthesis can be used to modify the usual order of operations

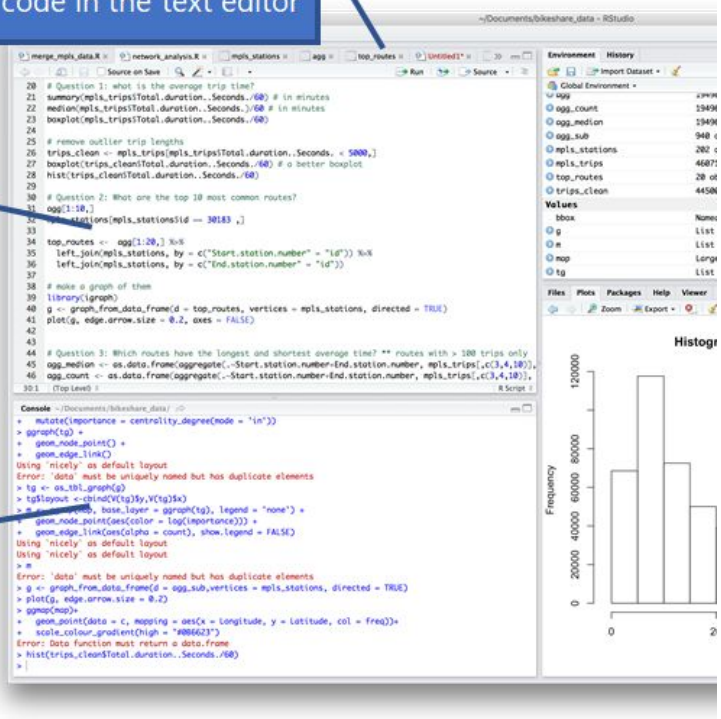
```
Input: 1+1
Output: 2
Input: 4*(3+2)
Output: 20
```

Saving your commands with the text editor

Run button: runs the current line or selection of code in the text editor

Text editor: Code can be written here and saved in .R files. Code can be run line-by-line or all at once

Console: where R runs, and where output is shown. You can enter commands here, but they won't be saved



Assigning values to a variable

- Values can be assigned to a variable using the assignment operator ‘<-’
- You may see the ‘=‘ operator used for assignment. It operates similarly but is not best practice.

```
a <- 5
b <- 2
c <- a+b
print(c)
```

Vectors: variables that can hold multiple values

- Vectors are a special type of variable that can hold multiple values, however, they must be all of the same type
- Square brackets [] are used to select a specific element based on where it is located

```
x <- c(2,4,6,8) # assigns values to x
   x[3] # selects 3rd element of x
length(x) # returns the length of the
                vector
```

Anatomy of a function

- Functions take in variables, do something to them, then output the result

```
mean(x)

Sum_of_vector <- sum(x)

boxplot(x) # some functions make
           charts
```

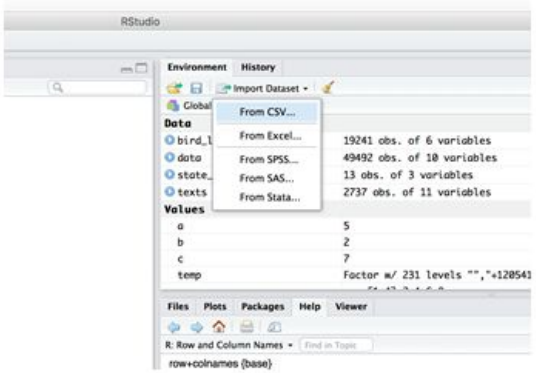
Packages

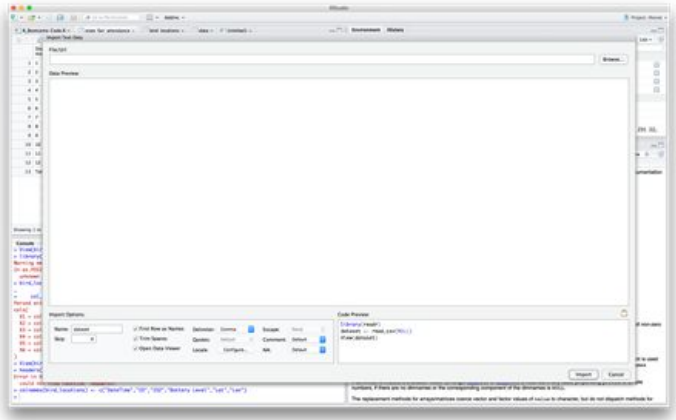
- Packages are collections of functions that others have created
- They are freely accessible and help you easily accomplish complex tasks without writing the code yourself
- Some commonly used packages:
  - ggplot2: used to create complex graphics
  - caret: contains functions for machine learning
  - tidyverse: a collection of data science packages with common grammar and data structures

```
install.packages("ggplot2")
library(ggplot2)
```

Importing data: Method 1

- The easiest way to get data into RStudio is to use the built-in ‘Import Dataset’ tool:
  - Select ‘Import Dataset’ from the Environment Menu
  - 2.Select the type of data you are trying to import
  - 3.In the window that appears, browse for the file you want to import
  - 4.Rstudio automatically generates and runs the code needed to import your file
  - 5.Your data is now available in R, in a dataframe with the same name as your file





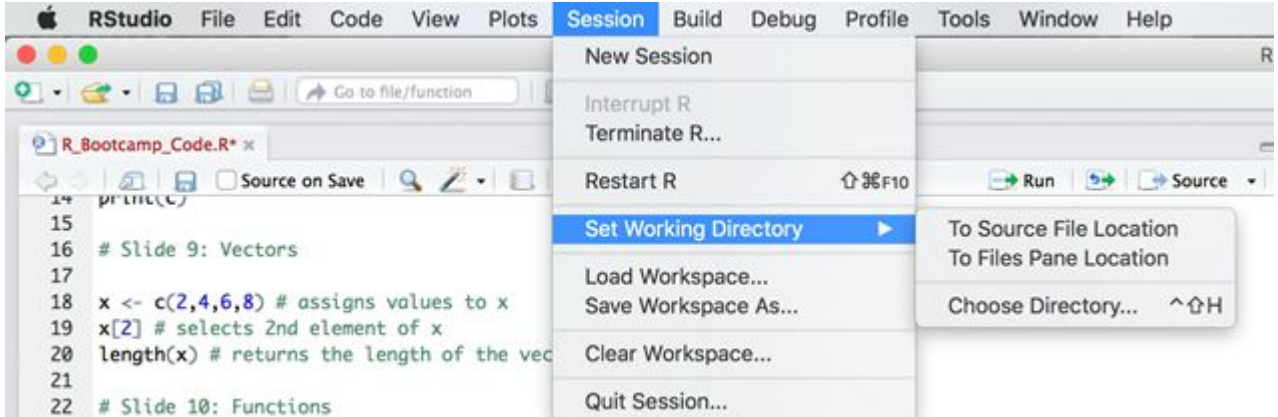
Importing data: Method 2

- More advanced users may prefer to import data programmatically
- The function ‘read.csv()’ is used to import data from a CSV (comma separated values) file
- If your data is in Excel (.xlsx) format, you can either save it as CSV or use another function
- Make sure that your working directory is set appropriately

```
data <- read.csv(”state_fair.csv”, header = TRUE)
```

Working directories

- The working directory (often abbreviated as ‘wd’) is the folder where the R code you write is executed
- This main benefit of this is that you just have to type the file name when importing a file, rather than the entire file path
- You can set the wd by going to Session > Set Working Directory
- Alternatively, you can use the command ‘setwd()’



Data frames

- A data frame is a table that can hold different types of data
- It is one of the most common data structures used in R

	Day of the Fair	2018	Record
1	1	122695	119145
2	2	108059	141023
3	3	222194	202126
4	4	184716	209969
5	5	124438	144504
6	6	120209	133595
7	7	144940	128966
8	8	156764	155183
9	9	179402	187066
10	10	NA	260374
11	11	NA	242759
12	12	NA	178867
13	Total	1363417	1997320

To select a single value:

```
dataframe_a[row,column]
```

To select an entire column:

```
dataframe_a$column_name
```

Basic plots

- The most basic way to plot in R is with the ‘plot’ command, which requires an X and Y argument
- Additional arguments can be set to customize the chart (titles, colors, etc.)
- The library ggplot2 allows for much more customization

```
plot(x = state_fair$day_of_fair, y = state_fair$attendance_2019)
```

Best practices for using R in class

- Use logical names for variables, not ‘variable1’
- Use very clear, simple names
- Add comments to your code using ’#’
  - Comments should make it easy for someone else to understand how your code works
- Add a header to your R code with your name, date, and the course (use ‘#’)
- Use blank lines and comments to separate your code into sections

Hands-on exercise: analyzing State Fair attendance data

- Download the ‘state_fair.csv’ file from the UBAC drive folder and import it into R
- Use R to answer the following questions:
  - What was the total number of visitors to the State Fair this year?
  - Which day had the most visitors this year?
  - Were any records broken this year? How many?
  - Create a plot of attendance by day for 2019
  - Create a plot of attendance by day for 2019, and overlay the record attendance numbers

Additional resources

https://www.datacamp.com/

Thanks

ganga020@umn.edu

https://sandeepgangarapu.com/