Data Analytics
CS301
Chapter 2,
Intro to R, Workflows

Week 1: 8th July
Summer 2021
Oliver BONHAM-CARTER
For Your Own Analysis?

- There exist tools to help you with analysis of data
- These tools were developed others for specific activities
- BUT! What if you need your own tools for your own specific investigations? (You may have to create your own software)

Develop Your Own Tools!!
We will be using the Book

- Note the chapters between the book and the website are not numbered identically!

- **Book:**
  - Chap 1: Data Visualization with ggplot
  - **Chap 2: Workflow; Basics**

- **On the web site:**
  - http://r4ds.had.co.nz/
  - Chap 3: Data Visualization
  - **Chap 4: Workflow; Basics**
The R Programming Language

• https://www.r-project.org/

• What is the R language?
  – An open source, well-developed programming platform for work in statistics, mathematics and data analytics
  – Cross platform; runs on major OSs
  – Popular programming skill among Big Data analysts, and data scientists

• Community Blogs:
  – https://www.r-bloggers.com/
  – https://twitter.com/rstudiotips/
  – https://towardsdatascience.com/
R: The Most Popular Data Mining Tool

http://blog.revolutionanalytics.com/2013/10/r-usage-skyrocketing-rexer-poll.html
R is Exploding in Growth

http://blog.revolutionanalytics.com/2013/10/r-usage-skyrocketing-rexer-poll.html
Most users are satisfied with R

http://blog.revolutionanalytics.com/2013/10/r-usage-skyrocketing-rexer-poll.html
## Ranking To Others: IEEE 2017

<table>
<thead>
<tr>
<th>Language Rank</th>
<th>Types</th>
<th>Spectrum Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Python</td>
<td>🌐💻</td>
<td>100.0</td>
</tr>
<tr>
<td>2. C</td>
<td>📱💻</td>
<td>99.7</td>
</tr>
<tr>
<td>3. Java</td>
<td>🌐📱</td>
<td>99.4</td>
</tr>
<tr>
<td>4. C++</td>
<td>📱💻</td>
<td>97.2</td>
</tr>
<tr>
<td>5. C#</td>
<td>🌐📱</td>
<td>88.6</td>
</tr>
<tr>
<td>6. R</td>
<td>📱💻</td>
<td>88.1</td>
</tr>
<tr>
<td>7. JavaScript</td>
<td>🌐📱</td>
<td>85.5</td>
</tr>
<tr>
<td>8. PHP</td>
<td>🌐</td>
<td>81.4</td>
</tr>
<tr>
<td>9. Go</td>
<td>🌐💻</td>
<td>76.1</td>
</tr>
<tr>
<td>10. Swift</td>
<td>📱💻</td>
<td>75.3</td>
</tr>
</tbody>
</table>

Find more amazing studies about R:
Getting Help in R

- Online help: place a “?” in front of a keyword
  - Ex: ?print

Please take notes!!
We will be coding together.
Variable Names

- Variable Names:
  - Begin with a letter, and can only include letters, numbers, periods and hyphens.
  - Hyphens: “-”
  - Periods: “.”

- SnakeCase (recommended by book)
  - val_of_height,
  - val_of_length,
  - val_of_width
Basic Math

• Mathematics
  - Addition: 1 + 1
  - Subtraction: 1 - 1
  - Multiplication: 3 * 7
  - Division: 1 / 4

• More complicated math, var assignments:
  - 4*(7+3)/10+1  Note: watch the order of operations!
  - Parameter of circle (C = 2 * pi * r)
    • R <- 4, Note the “<” means equal in R.
    • C <- 2 * pi * R = 2 * 3.1415 * 4
    • C is 25.13274
Variable Names

- **CamelCase:**
  - valOfHeight,
  - valOfLength,
  - valOfWidth

- **PeriodCase**
  - Val.of.height,
  - Val.of.length,
  - Val.of.width

- **What-EVER.Case**
  - Val.ofHEIGHT,
  - Val.Of_Length,
  - Val.oF.Width
Assign a variable
- \( x = 1 \), or
- \( x \leftarrow 1 \)
- \( y = 3 \)
- \( y \leftarrow 3 \)
- Run:
  \[ x + y \]

- \( myNum \leftarrow -2 \)
- \( myOtherNum \leftarrow -4 \)
- Run:
  \[ myNum + myOtherNum \]
Variables and Assignments

- $A <- 3$
- You could also use “$A=3$” (but this is not traditional programming in R)
- Hypotenuse ($C$) defined by $\sqrt{A^2 + B^2}$
- $A <- 3$
- $B <- 4$
- $C <- \sqrt{A^2 + B^2}$
- $C$ is ??
Logical Operations

- Booleans: Returning True or False:

3 > 4, 3 < 4,
2 + 4 == 6,
2 + 3 == 4 + 1
T == TRUE
F == FALSE
3 + 4 != 5
3 + 4 == 7
5 * 2 != 11
Try some of These in R!

- Logical **AND**
  - (&&)
    - F && F is F
    - F && T is F
    - T && F is F
    - T && T is T

- Logical **OR**
  - (||)
    - F || F is F
    - F || T is T
    - T || F is T
    - T || T is T

- Logical **NOT**
  - (!)
    - !F is T
    - !T is F

Truth Tables:
https://en.wikipedia.org/wiki/Truth_table

De Morgan’s Laws:
https://en.wikipedia.org/wiki/De_Morgan%27s_laws
Simple Strings

- **Strings**
  - “Hello World”

- **Concatenation of strings**
  - `H <- "Hello"`
  - `W <- "world"`
  - `paste(H,W, sep = " ")`

  • What is the result here??

- **You try: print your full name!**
  - `first <- "Sherlock"`
  - `last <- "Holmes"`
  - `paste(first,last, sep =" ")`
Built-in Functions

- R has a large collection of built-in functions:
  - function_name(arg1 = val1, arg2 = val2, …)
  - seq(from, to), ex: seq(0,10)
    - Gives a sequence, S =\{0,…,10\}
    - What happens when you press TAB after typing, “seq”?
- Use the sum( ) function to add two numbers
  - sum(1,10)
    - Adds 1 and 10
- Add all elements in a vector, v
  - v <- 0:10
  - sum(v)
    - Adds: 0 + 1 + … + 9 + 10 = 55
Simple Plots

- `x <- seq(1,100)` # assign `x` to the sequence 1 to 100
- `plot(x)` # plot this sequence
- `plot(sin(x))` or `plot(x,sin(x))` # see left plot below
- `plot(sin(x))` or `plot(x,sin(x), type = "l")` # see right plot below
Now, You Try

• Use R to write a command that...
  - Finds the **sum** of all numbers, 0 through 100
  - Finds the **sum** of all numbers, 0 through 10000
    *(now, set a variable equal to the sequence first)*

• Using the plot function, `plot(x,y,type = "l")` to plot a line of the function, \( f(x) = \sin(x) \) for \( x \) in \{0, \ldots, 30\}
  - \( x <- 0:10 \)
  - `plot(x, sin(x), type = "l")`

Exiting R: `q()`

Now try `cos()` and `tan()`!
Explore the Data
Of Your World

"Data exploration is the art of looking at your data, rapidly generating hypotheses, testing them, then repeating again and again..."

Import: Bringing in the raw data to work on it
Tidy: Cleaning it up so that numbers are numbers and etc.
Transform: Converting the data into something more convenient to use
Visualize: Finding general trends in data
Model: Testing phases, learning how to predict from the data.
Communicate: Publish and change the world!
Tidyverse’s Packages

The steps of the Tidyverse canonical data science workflow, as well as, the individual packages that the steps involve.

- **Import**
  - readr
  - readxl
  - haven
  - httr
  - rvest
  - xml2

- **Tidy**
  - tibble
  - tidyr

- **Program**
  - purrr
  - magrittr

- **Transform**
  - dplyr
  - forcats
  - hms
  - lubridate
  - stringr

- **Visualise**
  - ggplot2

- **Model**
  - broom
  - modelr
Data and Plotting

The Tidyverse library in R: a coherent system of packages for data manipulation, exploration and visualization
Data and Plotting

• For the first use, you need to **install** the library to your computer with,
  - `Install.packages(tidyverse)`

• Once installed, you only need to **call** (or **load**) the library with,
  - `library(tidyverse)`
Exploring Sun-Spot Data

- Sunspots – magnetic disturbances on the sun that can be observed from Earth
- Spots cycles are noted to repeatedly increase and then decrease over time
Articulating the Research Question

- Is there something predictable about the sunspot data?
- Can we collect some evidence of a pattern in the data?
- Could we use this pattern to predict?
- What does a pattern look like in the data?
Load and Plot Sunspot Data

```r
# Load library
library(tidyverse)

# find your sandbox file
sunData <- read.table(file.choose(), header = TRUE, sep = ",")

# See what the data looks like
View(sunData)

# Plot the data:
ggplot(data = sunData) + geom_point(mapping = aes(x = fracOfYear, y = sunspotNum))

# Save a file to the Desktop/ (or wherever) if you want...
ggsave("~/Desktop/myplot.png")
```

file: sandbox/sunspots.r
Code for a Simple GGPlot

- `install.packages("tidyverse")` # install as necessary
- `library(tidyverse)` # call installed library
- `ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))`

- Establish the *canvas* (where the plot is shown)
  - `ggplot()`

- Link to the data (set is called, ‘mpg’)
  - `ggplot(data = mpg)`

- Compute the geometry of point placement on canvas
  - `geom_point(mapping = ...)`

- Compute the aesthetics of the plot (titles, color, point type, etc)
  - `aes(x = displ, y = hwy)`