1. R essentials

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

Announcement: Regarding Homework 01

- **Extended deadline** Friday Oct 6 11:59pm
- Make sure to look at the **.html** output of your `Knit` command :
    - Make sure the text **Solution x:** appears at the top of each of your solutions
    - Include narrative in your own words.
    - Reflect on your work in the worksheet and include learning gains in the last exercise.
    - Use office hours and HW clinics today.

Lecture 0 Summary

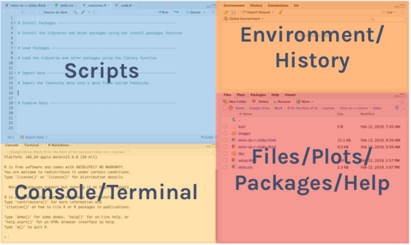
- Three parts to the course - All in `R` programming language
  - Programming
  - Statistics and Probability
  - Databases
- Accessing Rstudio instance for the course (Refer to slides and videos)
- Created a Data Science project report for UN votes.
- Course overview and Brief Syllabus walk through
- Rmarkdown essentials.(Completed in section 1, Homework 1)

Post Lecture 1 to-do for you

- Read syllabus carefully
- Note down important dates, and final exams
- Get familiar with Course site on Canvas
- Go to Section.
- Use office hours to get help
- Practice will make it perfect for you!

Have a great start to the quarter!

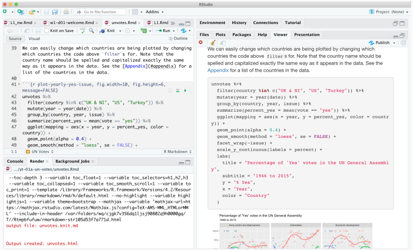
1.1 Last time: Rstudio and Rmd



```
# Install Packages ------------------------------------------------------------

# Install the tidyverse and skimr packages using the install.packages function


# Load Packages ---------------------------------------------------------------

# Load the tidyverse and skimr packages using the library function


# Import Data -----------------------------------------------------------------

# Import the faketucky data into a data frame called faketucky.



# Examine Data ----------------------------------------------------------------
```

```
~/Google Drive/Work/R for the Rest of Us/courses/intro-to-r-course/
Platform: x86_64-apple-darwin15.6.0 (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.

>
```

Name	Size	Modified
..		
Icon	0 B	Feb 12, 2019, 7:20 AM
images		
intro-to-r-slides.html	15.1 KB	Feb 13, 2019, 9:39 AM
intro-to-r-slides.Rmd	9.5 KB	Feb 13, 2019, 9:39 AM
libs		
setup.Rmd	[illegible]	Feb 12, 2019, 2:57 PM
style.css	2.3 KB	Feb 12, 2019, 3:07 PM



~~~
We can easily change which countries are being plotted by changing
which countries the code above `filter`s for. Note that the
country name should be spelled and capitalized exactly the same
way as it appears in the data. See the [Appendix](#appendix) for a
list of the countries in the data.

```{r plot-yearly-yes-issue, fig.width=10, fig.height=6,
message=FALSE}
unvotes %>%
  filter(country %in% c("UK & NI", "US", "Turkey")) %>%
  mutate(year = year(date)) %>%
  group_by(country, year, issue) %>%
  summarize(percent_yes = mean(vote == "yes")) %>%
  ggplot(mapping = aes(x = year, y = percent_yes, color =
country)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
~~~

```
.../yt-01a-un-votes/unvotes.Rmd
 --toc-depth 3 --variable toc_float=1 --variable toc_selectors=h1,h2,h3
 --variable toc_collapsed=1 --variable toc_smooth_scroll=1 --variable to
c_print=1 --template /Library/Frameworks/R.framework/Versions/4.2/Resour
ces/library/rmarkdown/rmd/h/default.html --no-highlight --variable highl
ightjs=1 --variable theme=bootstrap --mathjax --variable 'mathjax-url=ht
tps://mathjax.rstudio.com/latest/MathJax.js?config=TeX-AMS-MML_HTMLorMM
L' --include-in-header /var/folders/wq/cjgk7v356dq1ljsj90802q9h0000gq/
T//Rtmpbfufuw/rmarkdown-str105d53f7a771d.html
output file: unvotes.knit.md


Output created: unvotes.html
```

We can easily change which countries are being plotted by changing which countries the code above `filter`s for. Note that the country name should be spelled and capitalized exactly the same way as it appears in the data. See the Appendix for a list of the countries in the data.

```
unvotes %>%
  filter(country %in% c("UK & NI", "US", "Turkey")) %>%
  mutate(year = year(date)) %>%
  group_by(country, year, issue) %>%
  summarize(percent_yes = mean(vote == "yes")) %>%
  ggplot(mapping = aes(x = year, y = percent_yes, color = countr
y)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~issue) +
  scale_y_continuous(labels = percent) +
  labs(
    title = "Percentage of 'Yes' votes in the UN General Assembl
y",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```

Percentage of 'Yes' votes in the UN General Assembly
1946 to 2015

Today: Get started with R : Console, Environment panes, R essentials

What the heck *is* `R`?

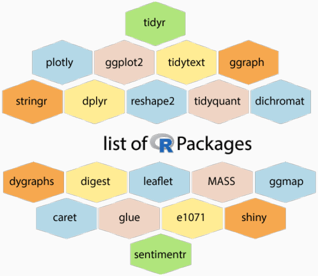
- R is an open-source statistical **programming language**
- R is also an environment for statistical computing and **graphics**
- It's easily extensible with *packages* (more on this later)
- `R` is based on the `S` language, which was developed by Bell laboratories in the 90's
- Home page: http://www.r-project.org

R and Rstudio

R: Engine	RStudio: Dashboard

- R is a programming language.
- RStudio is a convenient interface for R called an **IDE** (integrated development environment), e.g. *“I write R code in the RStudio IDE”*

R packages

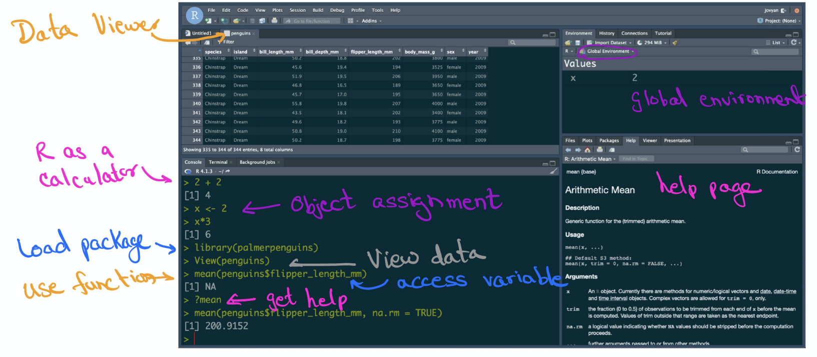


- **Packages** are the fundamental units of reproducible R code.
- There are over 18,000 R packages available on **CRAN** (the Comprehensive R Archive Network)1

1 Community contributed packages are stored at CRAN Comprehensive R Archive Network

R essentials: In the console pane

1. **R as a calculator:**
2. **Object Assignment:** `<-`
3. **Load** the `palmerpenguins` **package** so that we have access to all functions and data in this package.
4. **Check out** information (aka man page, help page) about `palmerpengunis` package
5. **View** the `penguins` **dataset**
6. Take a look at the `flipper_length` variable
7. **Find the average** `flipper_length`
8. **Get help** on the mean **function**



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
335	Chinstrap	Dream	50.2	18.8	202	3800	male	2009
336	Chinstrap	Dream	45.6	19.4	194	3525	female	2009
337	Chinstrap	Dream	51.9	19.5	206	3950	male	2009
338	Chinstrap	Dream	46.8	16.5	189	3650	female	2009
339	Chinstrap	Dream	45.7	17.0	195	3650	female	2009
340	Chinstrap	Dream	55.8	19.8	207	4000	male	2009
341	Chinstrap	Dream	43.5	18.1	202	3400	female	2009
342	Chinstrap	Dream	49.6	18.2	193	3775	male	2009
343	Chinstrap	Dream	50.8	19.0	210	4100	male	2009
344	Chinstrap	Dream	50.2	18.7	198	3775	female	2009

Showing 335 to 344 of 344 entries, 8 total columns

```
> 2 + 2
[1] 4
> x <- 2
> x*3
[1] 6
> library(palmerpenguins)
> View(penguins)
> mean(penguins$flipper_length_mm)
[1] NA
> ?mean
> mean(penguins$flipper_length_mm, na.rm = TRUE)
[1] 200.9152
>
```

Values

x 2

mean {base} R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

x An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects. Complex vectors are allowed for trim = 0, only.

trim the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.

na.rm a logical value indicating whether NA values should be stripped before the computation proceeds.

... further arguments passed to or from other methods.

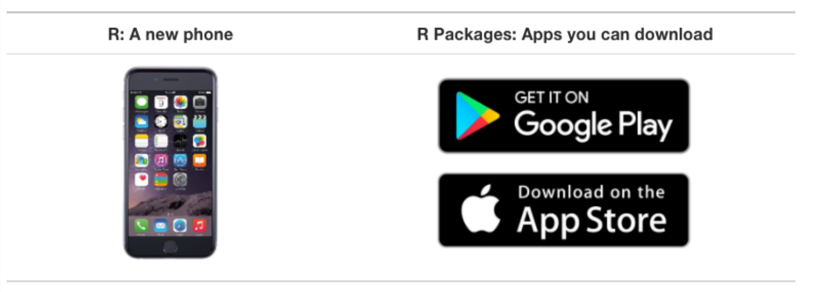
Review this

Easier to reproduce this code from an **.Rmd document** than the console.

- `your_workingfiles-> Lecture01 -> YT01 -> R-essentials.Rmd` section # 1. and 2.
- Remember to save any files in **your workingfiles folder, in subdirectories where you can locate it**, not in **content** folder

R essentials : Packages (aka libraries)

Packages add functionality that is not present in base R.



R: A new phone

R Packages: Apps you can download

```r
# don't forget quotes, install once on the computer
install.packages("package_name")

# no need for quotes, load once per session when needed
library(package_name)
```

The packages we will use in this class

- base R
- datasets
- graphics
- stats
- and a few others that are already **installed** for you in our PSTAT10 Rstudio instance.

You may need to download these special packages if you are working on an Rstudio instance you downloaded on your personal machine

OYO: On your own!

What command will **load** these packages into your `R` session?

Are these packages loaded already?

R essentials : Assignment operator <-

is used to assign values to objects.

`OBJECT <- VALUE`

```
x <- 2  # note the change in the environment
```

"Say: Create an object x and assign it the value 2"

R essentials : Accessing variables

- By name of a (scalar) variable
- Columns (variables) in data frames are accessed with $:

```
x
dataframe$var_name
```

R essentials : Help/Documentation

```
?function_name
```

```
?mean
```

Each function help file has the following useful sections:

- **Description:** overview of the function
- **Usage:** syntax, with list of arguments in particular order
- **Arguments:** description of arguments
- **Details:** in depth description of function's operation
- **Value:** output of the function
- **Examples:** copy-and-pasteable examples

FTC: For the curious: Check this stackover flow page for a write up of more ways to get help in R

R essentials : Comments

#

- use comments often and appropriately
- do this for the grader (and your grade) as well
- do this for your future self!

```r
x <- 2  # note the change in the environment

# A longer comment spanning a few lines
# can be written like this
# and is helpful to provide documentation
# for code
```

R essentials: Naming conventions for Objects in R

- Use names that convey information about the object. Descriptive names are best.
- Use names that are concise and meaningful (this is not easy!)
- Generally, variable names should be nouns and function names should be verbs

- Use lower case letters and numbers
- Use underscores (_) so called snake case to separate words within a name. eg. `flipper_length`
- Letters, digits and dot (period) can all be used.
- Must not start with a digit.
- Avoid names that start with a period. They have a special meaning in `R`

E.g. `welcome_msg <- "Welcome to PSTAT10"`

.

R essentials : Coding style

canyoureadthissentence?

```r
age <- c(6, 9, 15)

data.frame(age_kid = age)
```

- After function names do not leave any spaces.
- Before and after operators (e.g. <-, =) leave spaces.
- Put a space after a comma, **not** before.
- Object names are all lower case, with words separated by an underscore (snake_case)

- Have a style
- Stick to it

Caution! : Environments

The environment of your R Markdown document is separate from the Console!

Remember this, and expect it to bite you a few times as you begin learning to work with R Markdown!

Let’s take a look

Go to `your_workingdirectory -> Lecture01 -> YT01 -> R-essentials.Rmd section ## 3. R essentials : Defining objects`

R essentials: summary

- Console and Environment Panes, Command Prompt
- Objects
  - Variables: nouns
  - Functions: verbs
  - Naming conventions
- Packages: ready made functions and datasets from others
  - Install once
  - Load every time you need it
- Help : ?
- Assignment Operator : <-
  - printing objects
- Comments: #
  - **use them!** for yourself, the grader
- Coding style : **have one** and be consistent
  - See chapters 1-3 of the tidyverse style guide
- Environment

Next we will see...

- Data Types : What type of data does R use?
- Data Structures : What structure can be used to store our data?

Some extras FTC!

Downloading R

Go to: https://cran.r-project.org/

Chose from:

- Download R for (Mac) OS X
- Download R for Windows

Mac users choose Mac download

Windows users choose Windowns download

Downloading RStudio

1. Download and install R first.
2. Go to https://posit.co/download/rstudio-desktop/

Know Your RStudio Environment

There are a *lot* of keyboard shortcuts in RStudio. To view all the options, you must engage the keyboard shortcut that rules them all:

- Windows: `Alt` + `Shift` + `K`
- macOS: `Option` + `Shift` + `K`

Some favorites

1. Autocomplete command.
    - Both: `Tab`
2. Run the current line, selection from the editor.
    - Windows: `Ctrl` + `Enter`
    - macOS: `Cmd` + `Return`
3. Run the current code chunk from the editor.
    - Windows: `Ctrl` + `Shift` + `Enter`
    - macOS: `Cmd` + `Shift` + `Return`