

# A Walk on the Side

an introduction to R for data analysis

...

GW Libraries/STEMWorks Workshop  
Spring 2018

[go.gwu.edu/gwlibrworkshop](https://go.gwu.edu/gwlibrworkshop)



# Goals



# Agenda

- About R and RStudio
- Hands-on:
  - variables
  - logical expressions
  - values, vectors, and data frames
  - R Studio projects
  - reading in data
  - exploring data
  - data wrangling:  
cleaning and reshaping
  - data visualization
  - data analysis
  - functions
  - reports
- Resources for further learning



# Acknowledgments



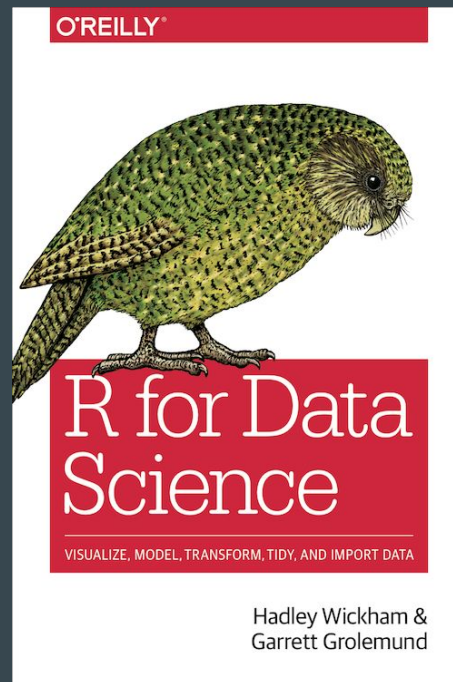
Teaching basic lab skills  
for research computing

**DATA CARPENTRY**

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

**R Tutorial**

An R Introduction to Statistics



# Workshop Housekeeping

Ask questions!

Respect every question and person asking the question

Help each other out!

If something is confusing in the workshop,  
it probably needs improvement; let us know.

Stay as long as you like

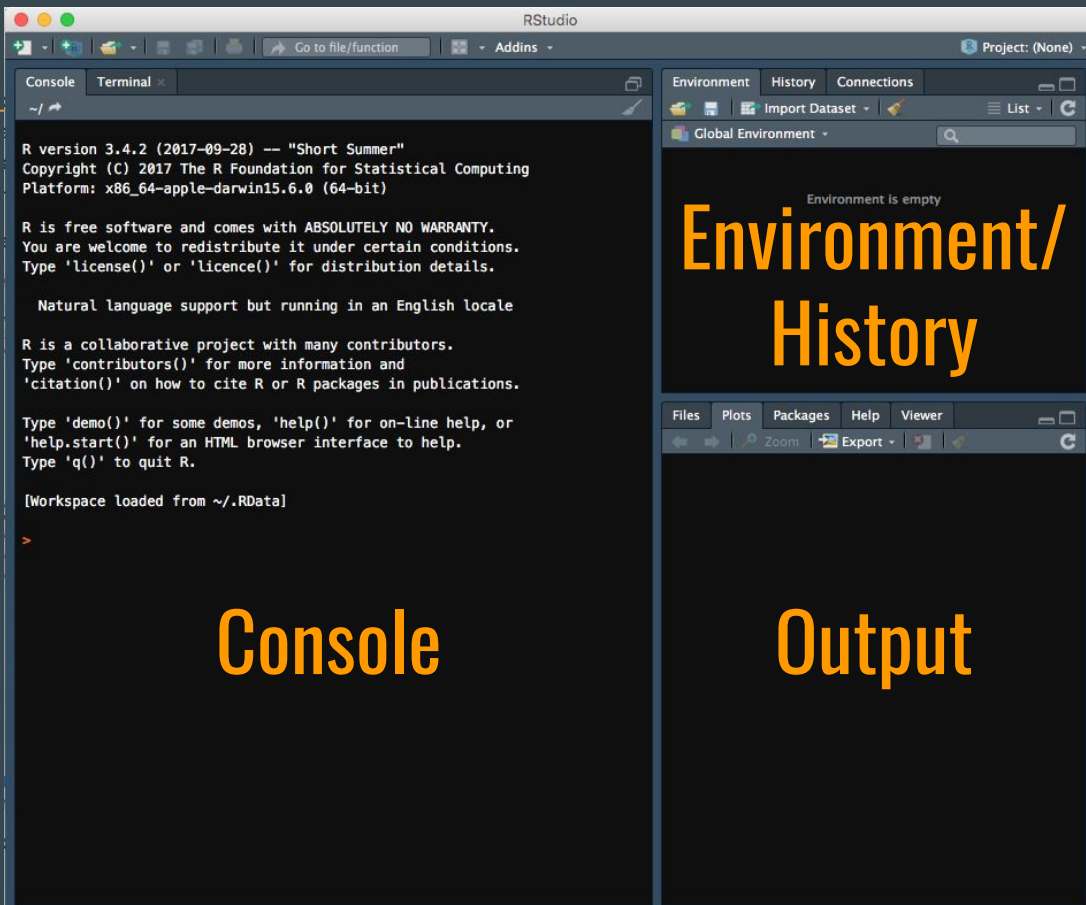


# About R

- Free/Open source
- For statistical computing and graphics
- CRAN - [r-project.org](https://r-project.org)
  - [R packages](#)
  - [R journal](#)
  - ...



# R Studio

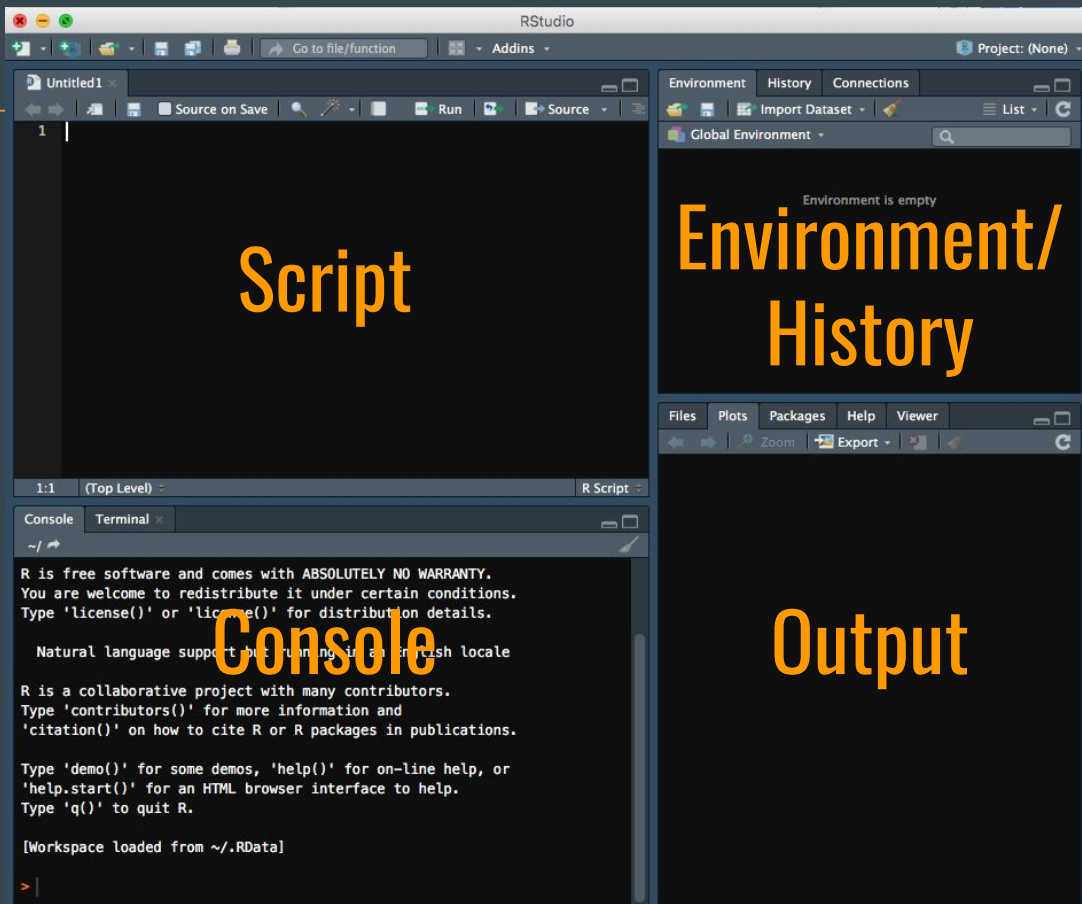


Console

Output



# R Studio



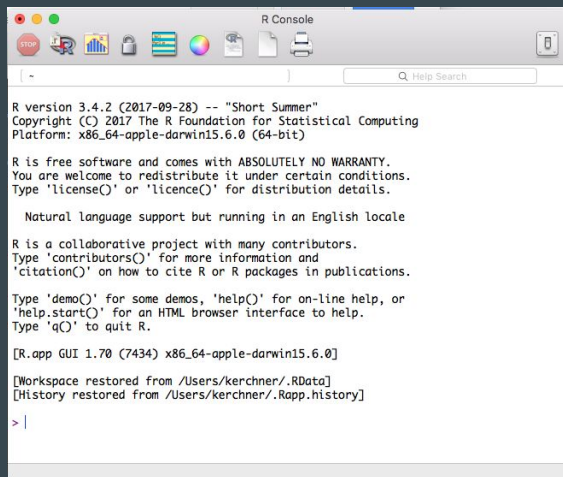
A WALK ON THE R SIDE



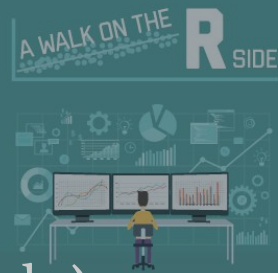
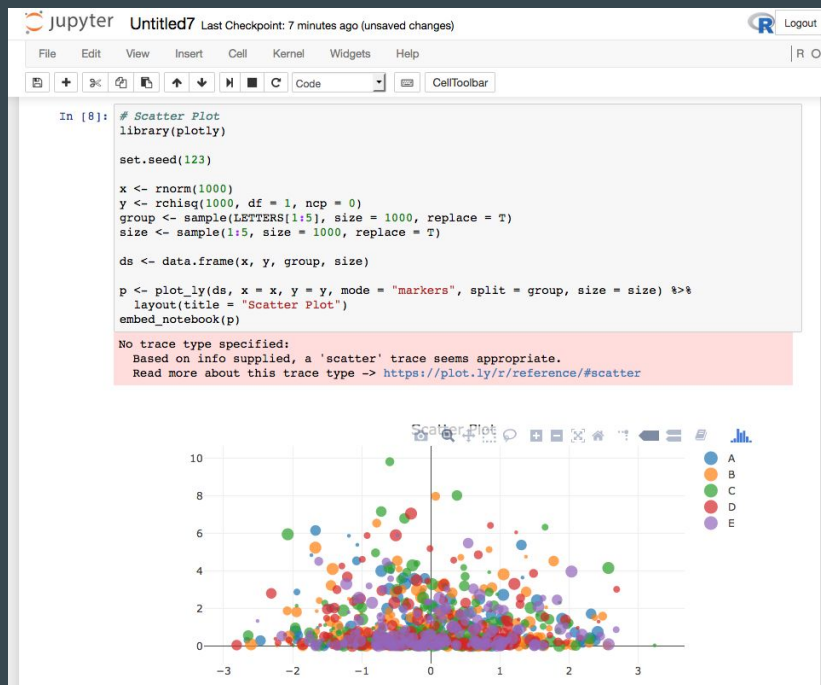


# Other Ways To Use R

Plain R console



Jupyter Notebook (e.g. in Anaconda)





Let's tRy it!



# Variables

- Try using R as a "calculator" in the Console
  - Try some mathematical functions, too
- Create some variables
  - variable naming
  - `<-` for assigning values to variables (Option - on Mac, Alt - on Win)
  - numeric, character, logical
  - Watch the Environment pane!
  - `class()`
  - Coercion w/ `as.integer`, `as.character`, `as.logical`, `as...`

# Logical Expressions

- Operators include:  
==, <, >, ! (not), & (and), | (or), etc.





# Vectors

# Vectors

- A vector is
  - A sequence of data elements (components) all of the same type.
- Create vectors with `c()` (short for "combine")





# Let's pause to explore some useful tabs in RStudio

~/R Projects/rstudio-testproject - master - RStudio

Workshop.R x gapminder x

Source on Save Go to file/function Addins

```
1 library('tidyverse')
2 gapminder <- read_csv('data/gapminder.csv')
3
4 by_year <- gapminder %>%
5   group_by(year) %>%
6   summarize(weighted_avg_lifeExp = sum(pop*lifeExp)/sum(pop))
7
8 # Plot the data (scatterplot)
9 plot(y = by_year$weighted_avg_lifeExp, x = by_year$year, col='blue')
10 # Build a linear regression model
11 mod = lm(data = by_year, weighted_avg_lifeExp ~ year)
12 # Plot the line
13 abline(mod)
14
15 # or using ggplot2:
16 ggplot(data = gapminder, aes(x= gdpPercap, y = lifeExp, base = continent, color = continent))
17   geom_point() +
18   # ...
```

5:1 (Top Level) R Script

~/R Projects/rstudio-testproject

```
[,1]
[1,] 1
[2,] 2
[3,] "A"
[4,] "b"
[5,] 2
[6,] 2
> mx2 = matrix(list(1, 2, "A", "b"), nrow=2, ncol=2)
> mx2
      [,1] [,2]
[1,] 1    "A"
[2,] 2    "b"
> mx2 = matrix(list(1, 2, "A", 3, "b", 5), nrow=3, ncol=2)
> mx2
      [,1] [,2]
[1,] 1    3
[2,] 2    "b"
[3,] "A"  5
>
```

Environment History Connections Git

Global Environment

- df 3 obs. of 2 variables
- gapminder 1704 obs. of 6 variables
- housedata 1460 obs. of 81 variables
- lemod List of 12
- mod List of 12
- mx logi [1:3, 1:2] NA NA NA NA NA NA
- mx2 List of 6

Values

primes num [1:6] 2 3 5 7 11 13

testnum 5

Files Packages Help

R: Reduces multiple values down to a single value

summarise (dplyr)

R Documentation

Reduces multiple values down to a single value

Description

summarise() is typically used on grouped data created by `group_by()`. The output will have one row for each group.

Usage

```
summarise(.data, ...)
```

summarize(.data, ...)

Arguments

- .data A tbl. All main verbs are S3 generics and provide methods for `tbl_df()`, `dtplyr::tbl_dt()` and `dbplyr::tbl_dbi()`.
- ... Name-value pairs of summary functions. The name will be the name of the variable in the result. The value should be an expression that returns a single value like `min(x)`, `n()`, or `sum(is.na(y))`.



# Data Frames



# Data Frames

- A `data.frame` stores a data table
- Comprised of **vectors** of equal length. Vectors become columns.
- Columns and rows can have names.
- `tibble` (from the `tibble` package) has some advantages over `data.frame`



# To summarize...



## Value

10.2

## Vector

1	10.2
2	11.3
3	11.5
4	12.0

## Data Frame

	time	temp	boiling
1	51	10.2	FALSE
2	58	11.3	FALSE
3	63	11.5	FALSE
4	70	12.0	TRUE



# A brief word on **list** and **matrix**



# Projects in RStudio

# Projects in RStudio

## Recommendations:

- Use [Github for] **version control!**
- Create **folders** to keep things organized





It's time to **import** some data!



# Data Importing

- Prepare data as "tidy"
  - rectangular
  - one table per file
  - rows are observations, columns are variables
- Formats: CSV, TSV, Excel, Fixed-Width, JSON... and with the right packages: Stata, SPSS, SAS... (using **rio** or **haven**)
- A word about "big data" (consider **data.table**)



# R Packages





# Installing and loading R packages

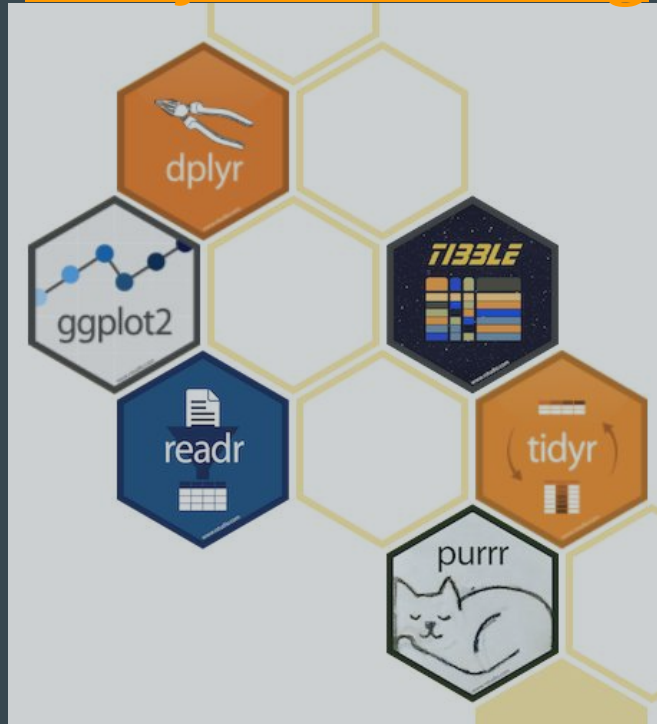
- `install.packages('mypackage')` - or press the Install button on the Packages tab in RStudio
- `library('mypackage')` -- or check the box on the Packages tab in RStudio



# Tidyverse Core Packages

[tidyverse.org](https://tidyverse.org)

- ggplot2 - graphics
- dplyr - data manipulation
- tidyr - tidying data
- readr - reading in data
- tibble - modern data frame
- purrr - functional programming





## Other often-used R packages

Basic stats functions, like ANOVA ▶ MASS

Mapping ▶ tmap, tmaptools, leaflet

Analyzing 2D and 3D shapes ▶ geomorph

Genomic data ▶ bioconductor

Cluster analyses ▶ cluster

Time series data ▶ forecast

Text mining ▶ qdap, sentimentr, tidytext

graph/network analysis ▶ igraph, sna

Interactive web visualizations ▶ shiny

# Exploring Data

- head, tail
- subsetting
- slicing and dicing





# Data Wrangling

[flickr.com/photos/thewomensmuseum/3687975017/](https://www.flickr.com/photos/thewomensmuseum/3687975017/)



# Data Transformation using the dplyr package

- filter()
- arrange()
- select()
- mutate()
- summarize()
- group\_by()
- ...

You will want to use a "pipe": `%>%`  
(shortcut: **control-shift-M**)



# Data Tidying with dplyr

- `gather()`
- `spread()`
- `separate()`
- `unite()`



# Joining with dplyr

"Merges" tables together

- `left_join()`
- `right_join()`
- ...







# Data Visualization

# Data Visualization

3 main packages:

- "base R"
- lattice
- ggplot2





# Data Analysis



# Functions



# R Markdown



# R Shiny



## Some Handy R Links

# Tutorials



- Software Carpentry:
  - <http://swcarpentry.github.io/r-novice-inflammation>
  - <http://swcarpentry.github.io/r-novice-gapminder>
- Data Carpentry:
  - <http://datacarpentry.github.io/R-ecology-lesson/>
  - <http://www.datacarpentry.org/R-genomics/>
- Lynda.com [lynda.it.gwu.edu](http://lynda.it.gwu.edu) - 3 video courses (~12 hours)
- [r-tutor.com/r-introduction](http://r-tutor.com/r-introduction)  
[r-tutor.com/elementary-statistics](http://r-tutor.com/elementary-statistics)
- R for Data Science <http://r4ds.had.co.nz>





## Classes at GW that teach or use R

Aside from the Data Science and Business Analytics programs:

- PSC 2102 - Fall 2018 (and possibly Summer II, 2018)  
Visualizing and Modeling Politics  
Prof. Eric Lawrence
- PPPA 6085 - Data Visualization
- TBD - Public Health, Prof. Helmchen - Fall 2018

# Reference Links

- [r-project.org](https://r-project.org)
- R search engine: [rseek.org](https://rseek.org)
- [rstudio.com](https://rstudio.com)
  - Cheat Sheets
- [stackoverflow.com](https://stackoverflow.com)



# Thanks!

- Dan Kerchner [kerchner@gwu.edu](mailto:kerchner@gwu.edu)
- Dr. Kes Schroer [schroerk@gwu.edu](mailto:schroerk@gwu.edu)
- Vishwesh Haldevanekar [vishwesh\\_s\\_h@gwu.edu](mailto:vishwesh_s_h@gwu.edu)

These slides: [go.gwu.edu/gwlibrworkshop](https://go.gwu.edu/gwlibrworkshop)

R or Statistics Appointments with Vishwesh:  
[calendly.com/vishwesh\\_s\\_h](https://calendly.com/vishwesh_s_h)

