A Walk on the R Side

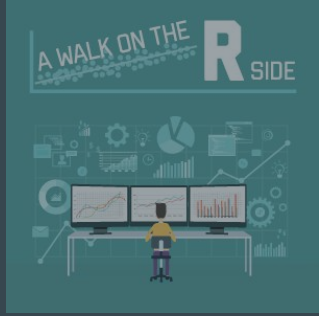
Online edition!

an introduction to R for data analysis

•••

GW Libraries Workshop
January 27, 2021

go.gwu.edu/rworkshop



Logistics

- Schedule (approximate)

  12:45-2:45(ish) R, with 1 ☕ break

  ~~ 15 minute break ~~

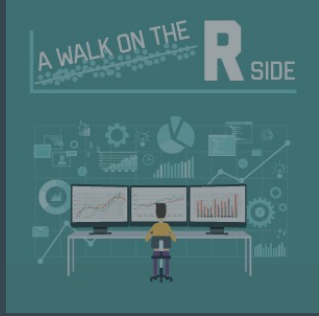
  3:00-5:00 R, with 1 ☕ break
- Ese can provide individual help
  - We can use Zoom breakout rooms if needed
- Collaborative Notes document: bit.ly/r_jan27

Goals

Learning Objectives

[Hopefully] You will learn how to do some of the following:

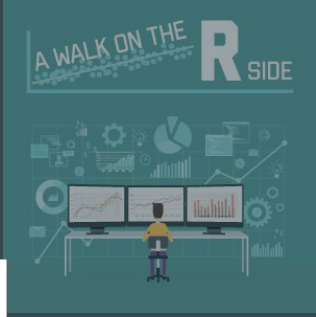
- Set up your laptop with R & RStudio (done!)
- Write and run an R program in RStudio
- Use variables of different types in R
- Use vectors and data frames in R to represent data
- Import & export data files
- "Wrangle" data in R
- Explore data in R with basic statistics and data visualizations
- Learn how to look for help to overcome obstacles



Agenda

- About R and RStudio
- Along the way: How to get help
- 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
  - R Markdown / reports
- Resources for further learning

Acknowledgments



software carpentry

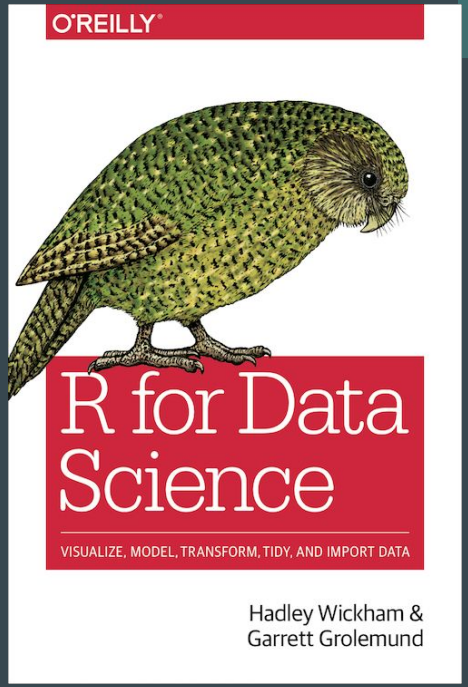
Teaching basic lab skills for research computing

DATA CARPENTRY

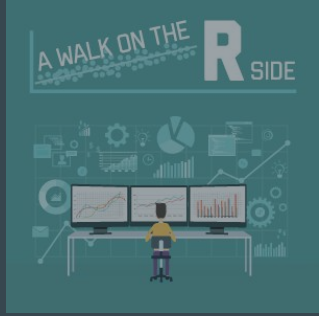
BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

R Tutorial

An R Introduction to Statistics



Workshop Housekeeping

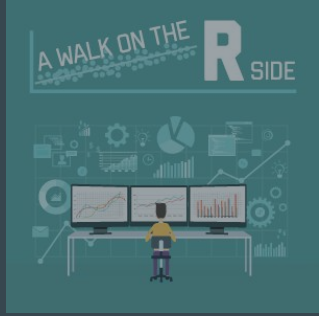


Ask questions! Either via voice or chat

Use chat to help each other out

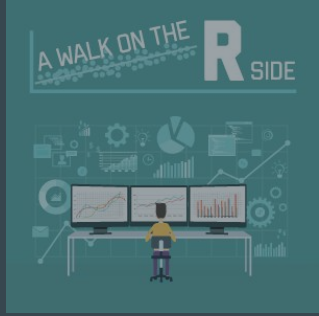
If something is confusing in the workshop, let us know.

About R



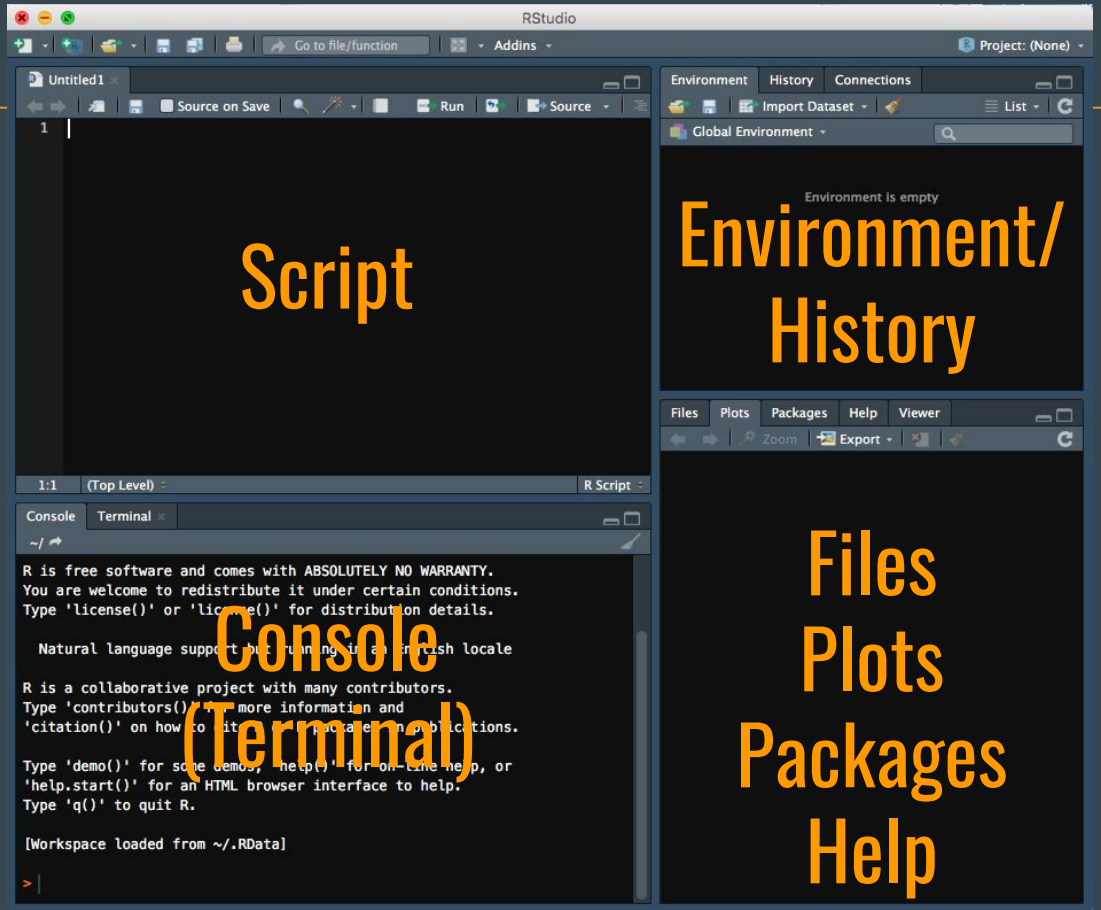
- Free/Open source
- Cross-platform (Mac, Windows, Linux)
- For statistical computing (and data visualization)
- CRAN - r-project.org
  - R packages
  - R journal

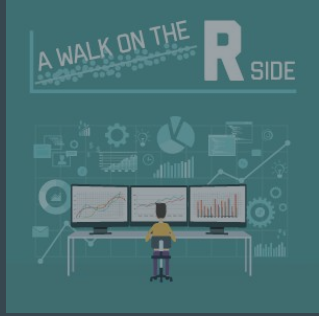
Reasons ReseaRcheRs pRefeR R

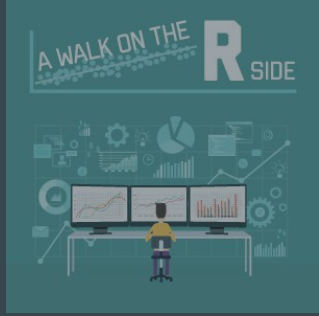


- Scripted language (vs. point/click)
- Features built around working with data
- Reproducibility
- Interdisciplinary
- Extensible
- Beautiful data visualization
- Community - RStudio Community, Stack Overflow

R Studio


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Zoom
Export
Console
Terminal
~/
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.
[Workspace loaded from ~/.RData]
>
Script
Console
(Terminal)
Environment/
History
Files
Plots
Packages
Help


A WALK ON THE R SIDE



Variables/Objects

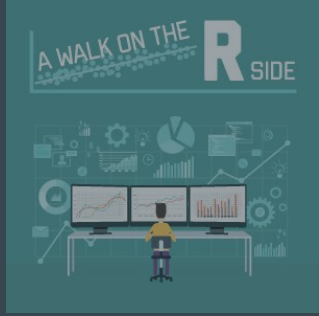
"Binding" data to a named object/variable allows you to store data in memory and access it later.

```
x <- 5

y <- c("Washington", "Chicago", "Washington", "Boston")

z <- data.frame(pt_id = c("A001", "B204"), bpm = c(60, 72))
```

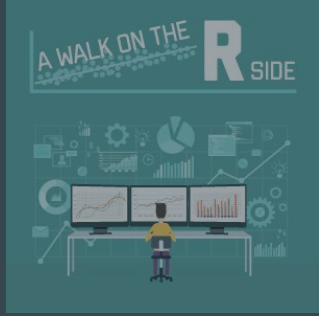
Variables

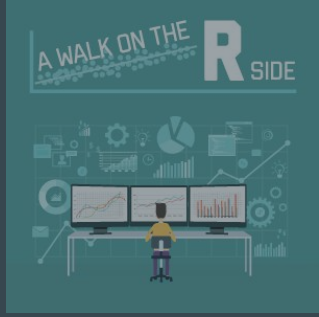


- Try using R as a "calculator" in the Console
  - Try some mathematical functions, too
- Create some variables
  - variable naming
  - `<-` for assigning values to variables (Option - on Mac, Alt - on Win)
  - numeric, character, logical
  - Watch the Environment pane!
  - `typeof()`
  - Coercion w/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

Logical Expressions

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





Basic Data Structures

Atomic Vector

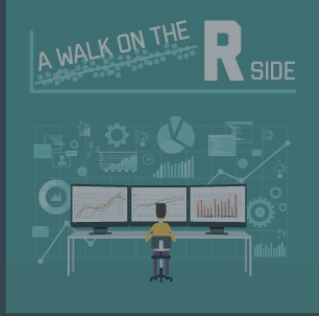
10.2

Vector

1	10.2
2	11.3
3	11.5
4	12.0

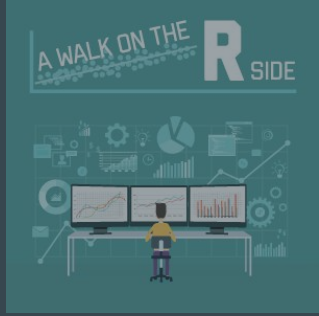
Data Frame

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

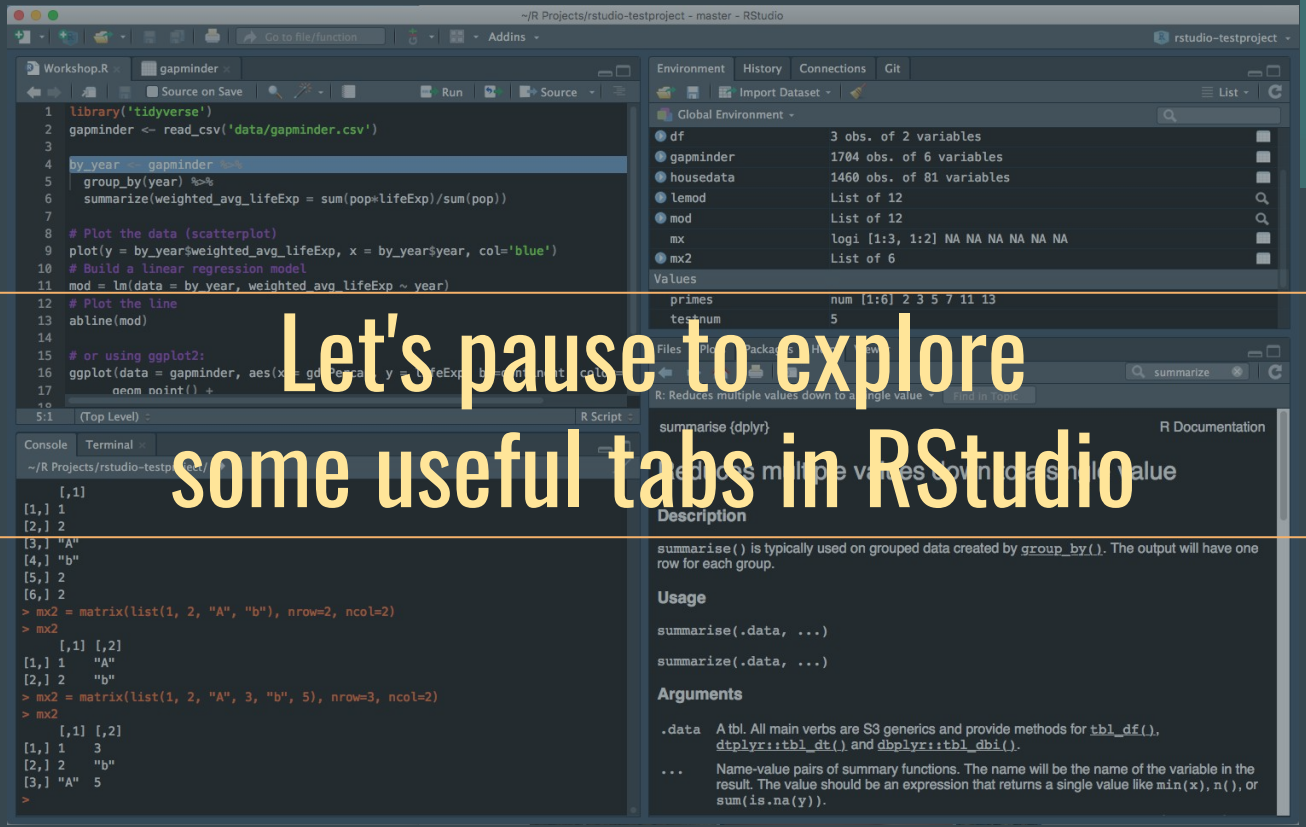


Vectors

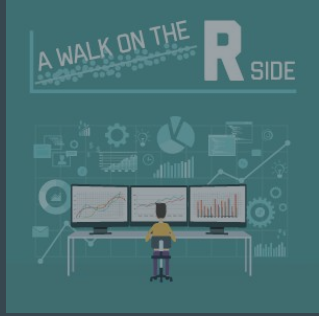
Vectors



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

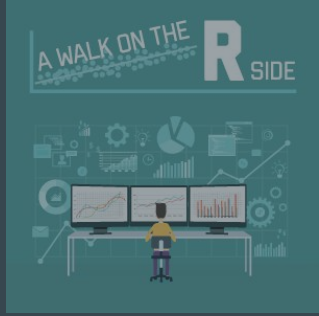


Let's pause to explore some useful tabs in RStudio

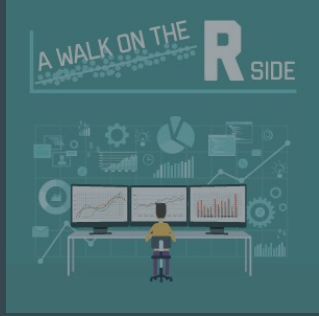


Data Frames

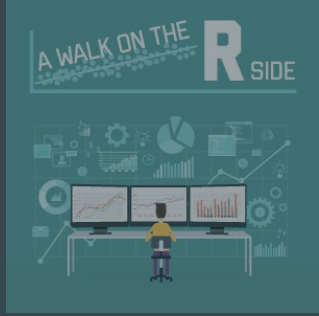
Data Frames



- 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`**

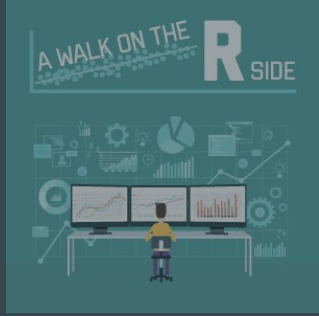


A brief word on `list` and `matrix`



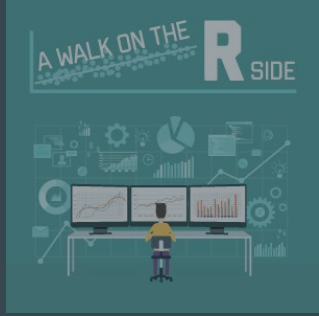
Projects in RStudio

Projects in RStudio

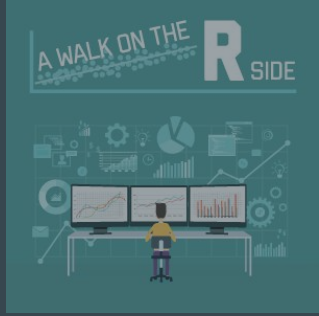


Recommendations:

- 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

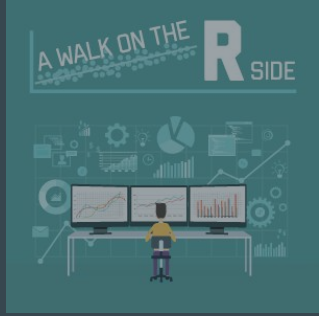
- Formats: CSV, TSV, Excel, Fixed-Width, JSON... and with the right packages: Stata, SPSS, SAS... (using **`rio`** or **`haven`**)

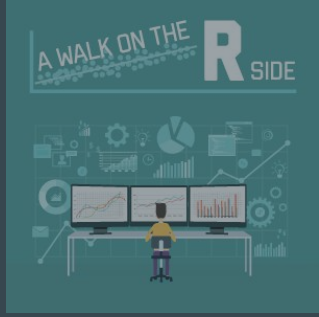
- A word about "big data" (consider **`data.table`**)

R Packages

Installing and loading R packages

- `install.packages('mypackage')`
- `library(mypackage)`





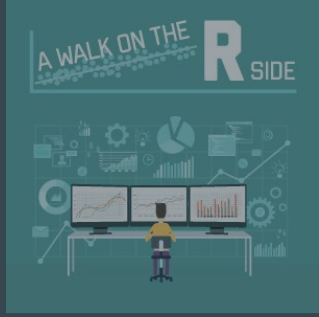
Tidyverse Core Packages

tidyverse.org

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

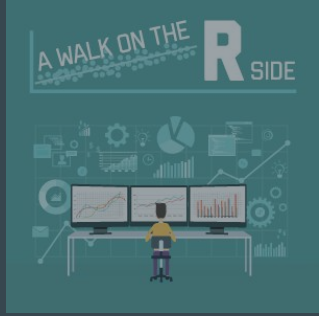
Other often-used R packages

Loading in various data file types	haven, readxl
Mapping	rgdal, tmap, 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
Web scraping	rvest



Exploring Data

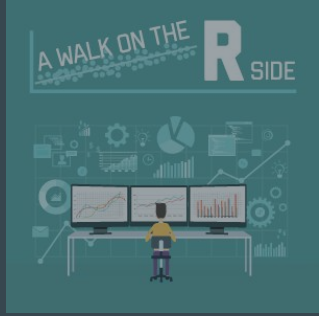
- head, tail
- subsetting
- slicing and dicing



Data Wrangling

flickr.com/photos/thewomensmuseum/3607075017

Data Transformation using the dplyr package



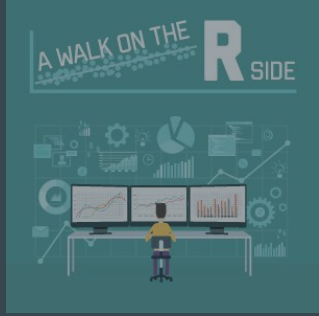
- filter()
- arrange()
- select()
- mutate()
- summarize()
- group_by()
- ...

You will want to use a "pipe": %>%

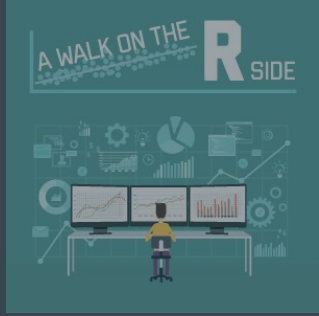
(shortcut: control-shift-M)

Data Tidying with dplyr

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

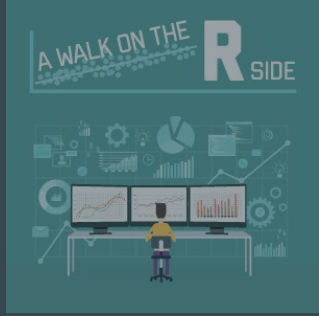


Joining with dplyr

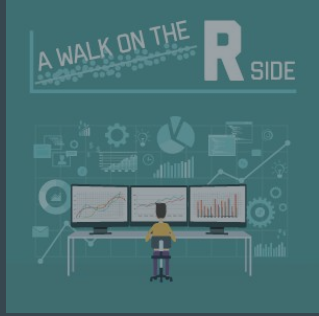


"Merges" tables together

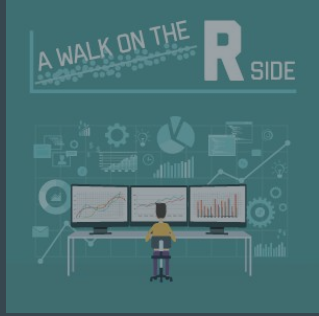
- left_join()
- right_join()
- ...



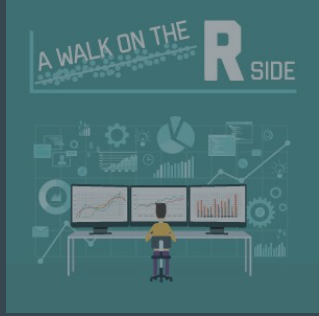
Data Visualization with "base R" and `ggplot`



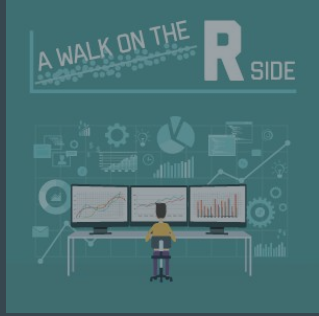
Data Analysis



Functions



R Markdown

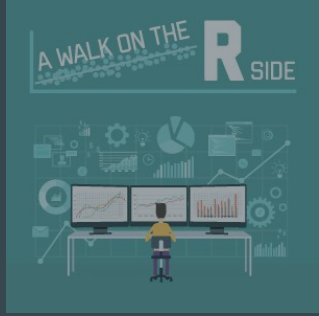


R Markdown

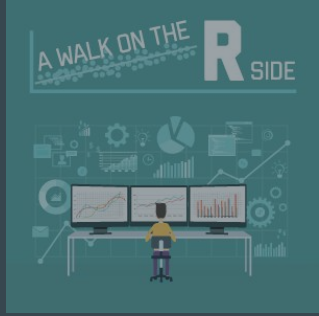
- A format for writing reproducible, dynamic reports with R (as HTML, PDF, MS Word, and more)
- rmarkdown.rstudio.com
- # Header 1
  ## Header 2
  Italic ****bold****
- Insert R code directly into your document

  ```
  ```{r setup}
  # your R code goes here
  ```
  ```

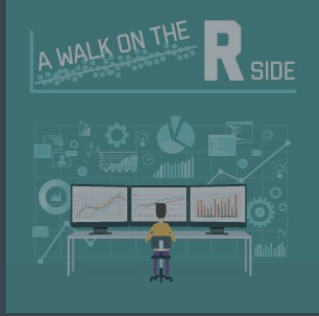
- Include LaTeX code with $ or $$



R Shiny

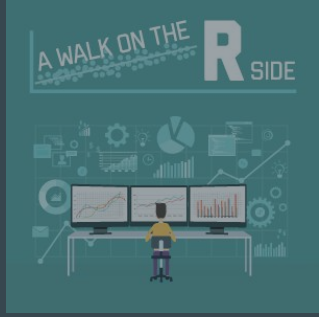


Some Handy R Links



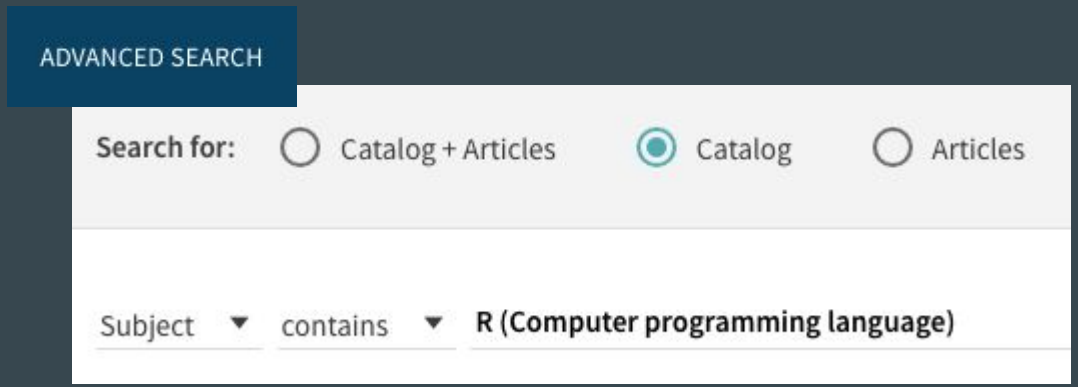
Tutorials

- RStudio R paths: education.rstudio.com/learn/
- Data Carpentry & Software Carpentry:
  - datacarpentry.org/R-ecology-lesson/
  - datacarpentry.org/r-socialsci/
  - swcarpentry.github.io/r-novice-inflammation
  - swcarpentry.github.io/r-novice-gapminder
- Linkedin Learning @ GW: go.gwu.edu/linkedinlearning
- r-tutor.com/r-introduction & r-tutor.com/elementary-statistics
- R Graph Gallery (w/code): r-graph-gallery.com

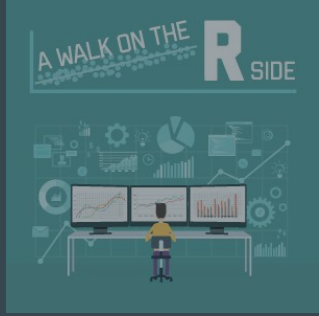


Books you can access for free

- Free books online - Hadley Wickham:
  - R for Data Science r4ds.had.co.nz
  - Advanced R adv-r.hadley.nz/
- Through your GW library privileges:

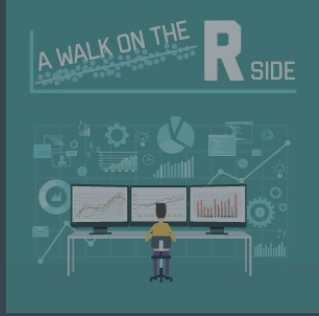


Reference Links



- R language (CRAN): r-project.org
- R search engine: rseek.org
- rstudio.com
  - Cheat Sheets! rstudio.com/resources/cheatsheets
- stackoverflow.com

Thanks!



Dan Kerchner kerchner@gwu.edu

These slides: go.gwu.edu/rworkshop

R or Statistics Appointments: calendly.com/statistical-consulting-gw

Appointments with me: calendly.com/kerchner

Coding consultations (Python, git, etc.): calendly.com/gwul-coding/