A Walk on the R Side

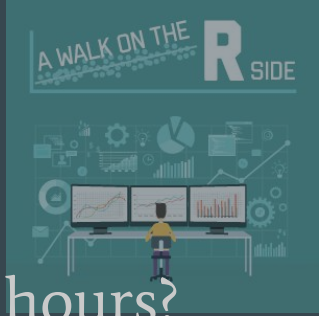
an introduction to R for data analysis

• • •

GW Libraries Workshop
Spring 2024

go.gwu.edu/rworkshop

FAQ



Q: Will you sign my form for Professional Enhancement hours?
A: Yes, email me!

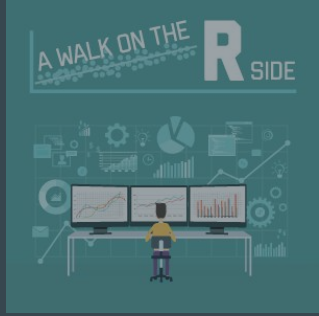
Q: Can I get a copy of your R code?
A: Yes, email me

Q: Will this workshop be recorded?
A: No, so hang on for the ride!

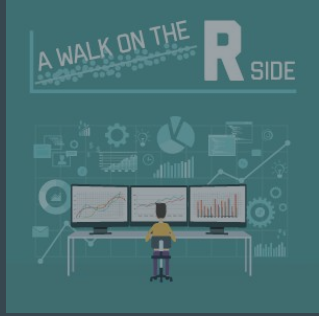
Logistics

- Schedule

  9:30- 2:15 with a ~1 hr break for lunch



Upcoming R workshops



- Feb. 1 (Thurs., 9:30am-2:15pm)

  This workshop, AGAIN! ← *tell your friends!*

- Feb. 6 (Tues., 1-3:30pm)

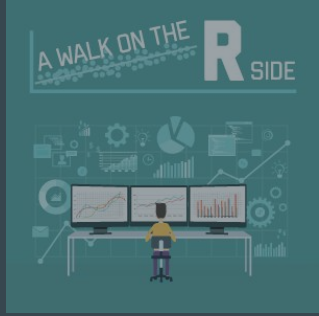
  Farther into R: *More* R for Data Analysis

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

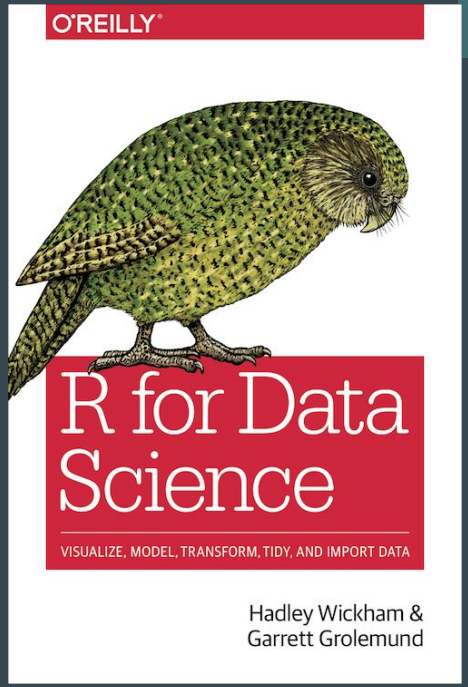
DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

R Tutorial

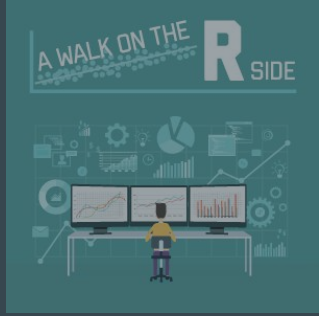
An R Introduction to Statistics

r-tutor.com



r4ds.had.co.nz

Workshop Housekeeping

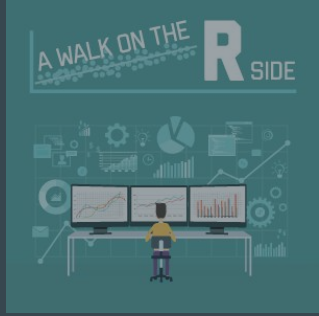


Ask questions! Either via voice or chat

Use chat to help each other out

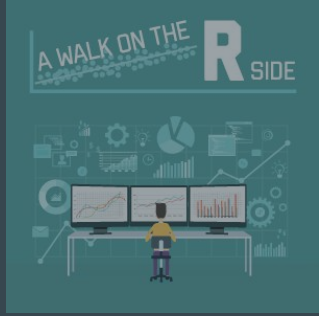
If something is confusing in the workshop, let us know.

About R



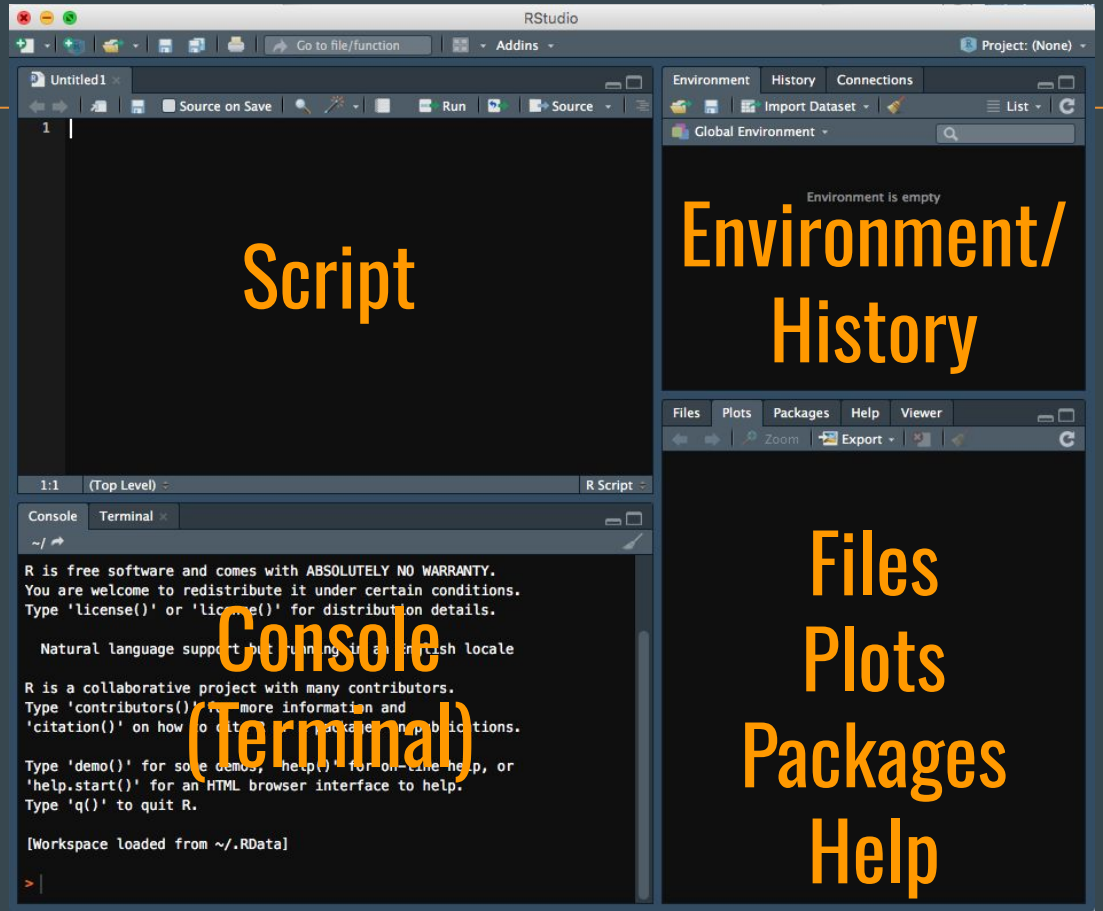
- Free/Open source
- Cross-platform (Mac, Windows, Linux)
- For statistical computing (and data visualization)
- CRAN - r-project.org
  - R packages
  - R journal

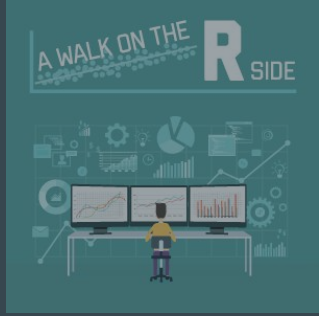
Reasons researchers prefer R



- Scripted language (vs. point/click)
- Features built around working with data
- Reproducibility
- Interdisciplinary
- Extensible
- Beautiful data visualization
- RStudio (Posit) is a well-liked R development app
- 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
Script
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]
>
Console
(Terminal)
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Environment/
History
Files
Plots
Packages
Help
Viewer
Zoom
Export
Files
Plots
Packages
Help



Variables/Objects

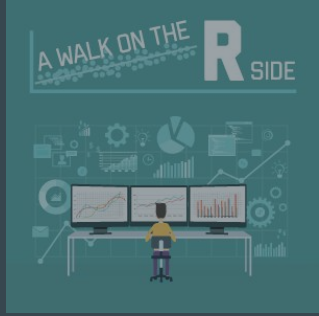
"Binding" data to a named object/variable allows you to store data in memory and access it later.

```
x <- 5
```

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

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

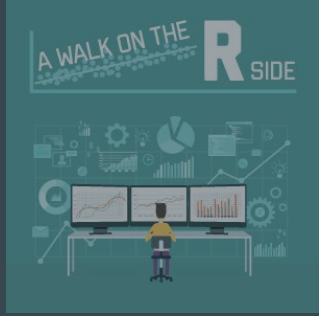
Variables

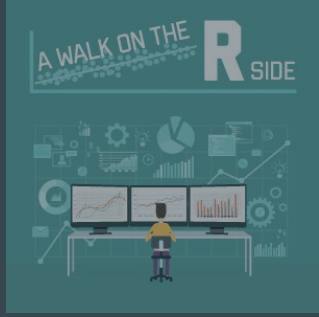


- Try using R as a "calculator" in the Console
  - Try some mathematical functions, too
- Create some variables
  - variable naming
  - `<-` for assigning values to variables (Option - on Mac, Alt - on Win)
  - numeric, character, logical
  - Watch the Environment pane!
  - `typeof()`
  - Coercion w/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

Logical Expressions

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





Basic Data Structures

Atomic Vector

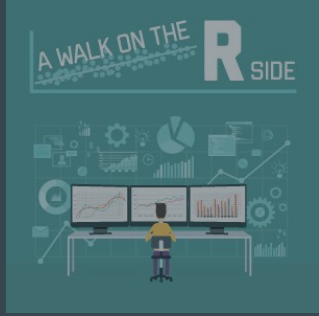
10.2

Vector

1	10.2
2	11.3
3	11.5
4	12.0

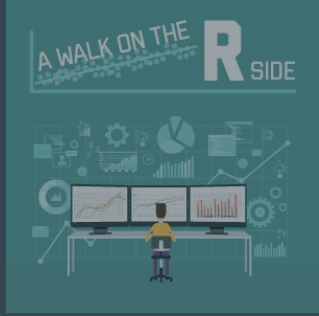
Data Frame

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

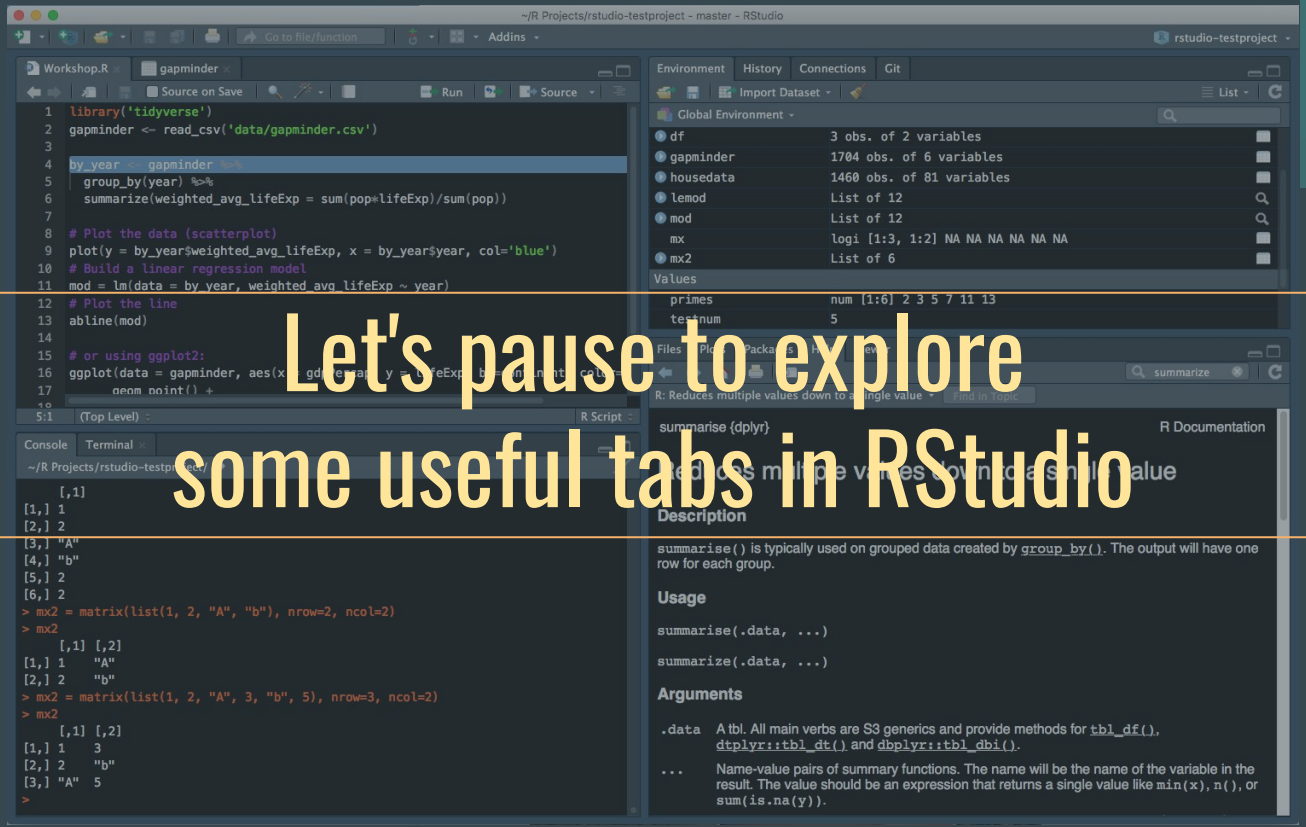


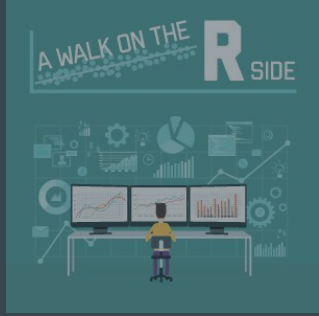
Vectors

Vectors

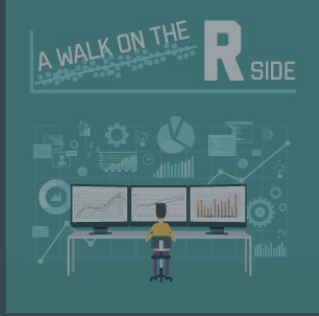


- A vector is
  - A sequence of data elements (components) all of the same type.
- Create vectors with `c()` (short for "combine")


~/R Projects/rstudio-testproject - master - RStudio
Workshop.R
gapminder
Source on Save
Run
Source
library('tidyverse')
gapminder <- read_csv('data/gapminder.csv')
by_year <- gapminder %>%
group_by(year) %>%
summarize(weighted_avg_lifeExp = sum(pop*lifeExp)/sum(pop))
# Plot the data (scatterplot)
plot(y = by_year$weighted_avg_lifeExp, x = by_year$year, col='blue')
# Build a linear regression model
mod = lm(data = by_year, weighted_avg_lifeExp ~ year)
# Plot the line
abline(mod)
# or using ggplot2:
ggplot(data = gapminder, aes(x
geom_point() +
Environment
History
Connections
Git
Import Dataset
List
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
summarize
summarise {dplyr}
R Documentation
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)).
Console
Terminal
[,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
>
Let's pause to explore
some useful tabs in RStudio

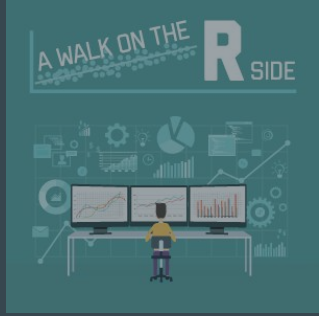


Data Frames

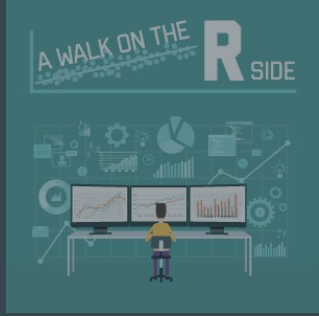


Data Frames

- 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`**

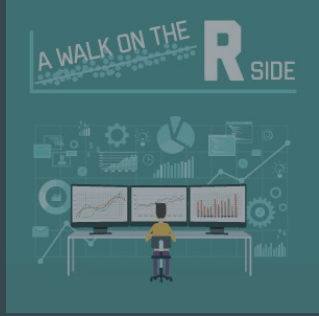


A brief word on `list` and `matrix`



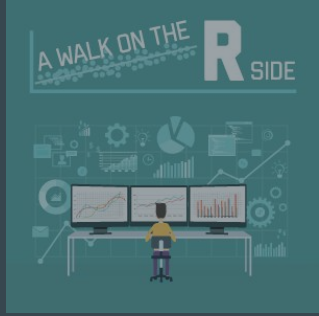
Projects in RStudio

Projects in RStudio

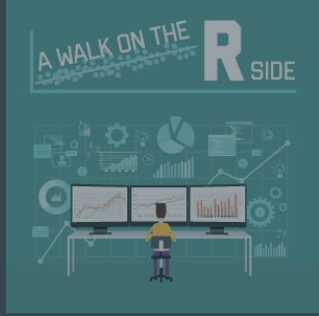


Recommendations:

- 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`**)

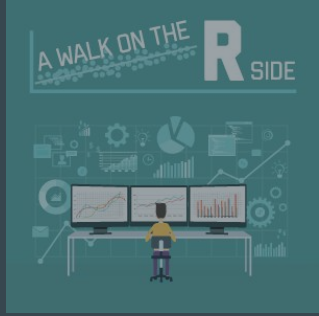
- A word about "big data" (consider **`data.table`**)

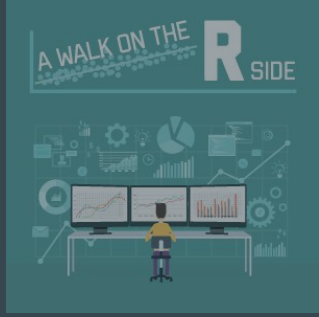
R Packages

flickr.com/photos/philocrites/405250216

Installing and loading R packages

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





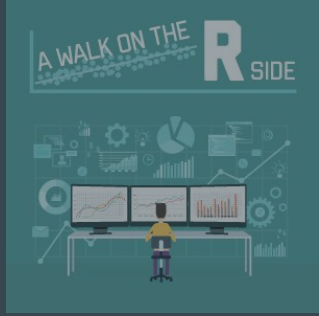
Tidyverse Core Packages

tidyverse.org

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

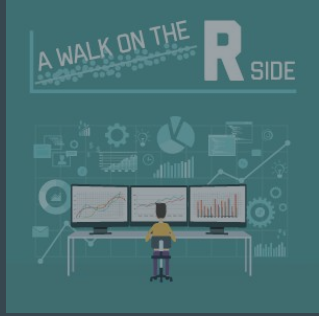
Other often-used R packages

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

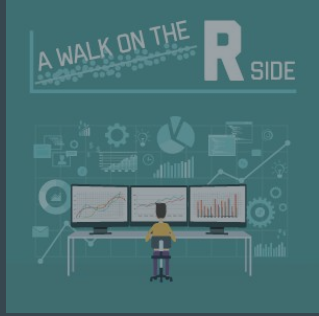
- head, tail
- subsetting
- slicing and dicing



Data Wrangling

flickr.com/photos/thewomensmuseum/3607075017

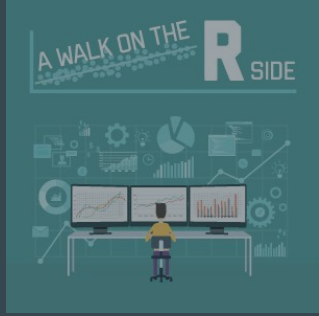
Data Transformation using the dplyr package



- select() # keep only certain columns
- filter() # keep only certain rows
- mutate() # add/modify variables
- group_by() %>% summarize()
  # compute summary statistics per group
- arrange() # order by a variable
- dropna() # drop rows with NAs in specified vars.

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

(shortcut: control-shift-M)



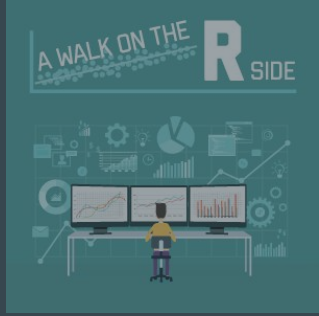
Joining with dplyr

"Merge" tables together

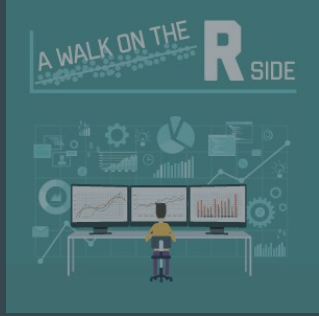
- left_join()
- right_join()
- ...

Data Tidying/Reshaping with tidyr

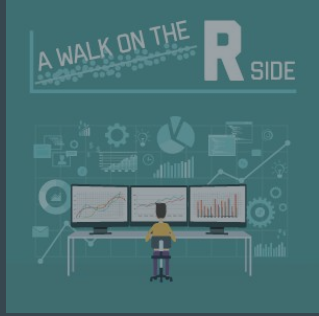
- pivot_wider()
- pivot_longer()
- ...



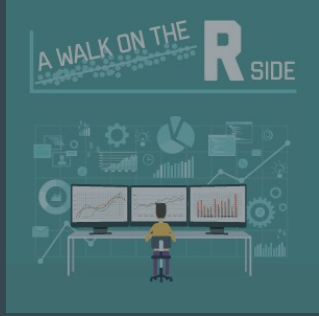
Data Visualization
with "base R" and `ggplot`



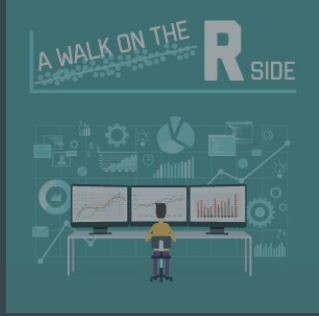
Data Analysis



Functions



R Markdown

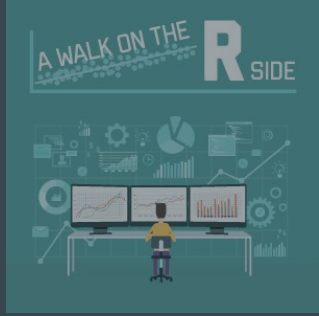


R Markdown

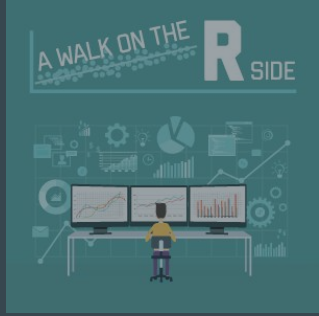
- A format for writing reproducible, dynamic reports with R (as HTML, PDF, MS Word, and more)
- rmarkdown.rstudio.com
- # Header 1
  ## Header 2
  Italic ****bold****
- Insert R code directly into your document

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

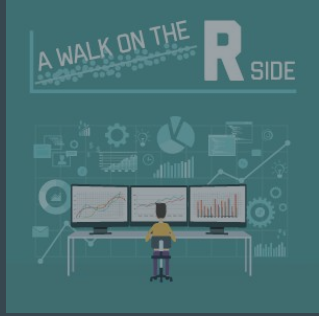
- Include LaTeX code with $ or $$



R Shiny

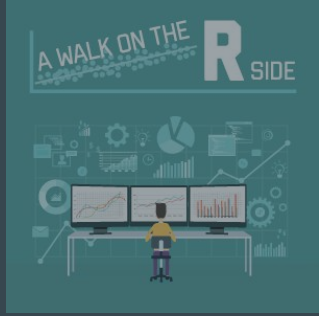


Parting thoughts



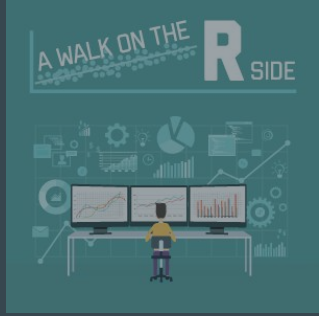
Recommended practices

- Use Projects in RStudio
  - Set up folders
- Use tidyverse packages (dplyr, tidyr, etc.) to wrangle your data
- Leave raw data raw
- 🧹 Empty out your variables, then make sure your script runs from the top
- Learn by finding and using working examples



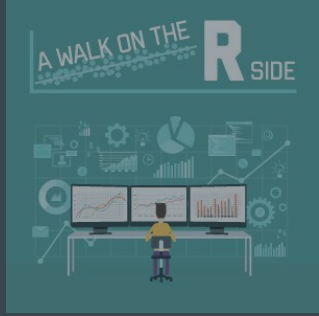
Some Handy R Links

NEW for 2024!! R "libguide"



Only the best R links:

libguides.gwu.edu/Rstats



Thanks!

Dan Kerchner kerchner@gwu.edu

These slides: go.gwu.edu/rworkshop

Statistics focused (+ R/Python/SAS/etc.) appointments w/graduate student consultants: go.gwu.edu/dataconsulting

Appointments with me: calendly.com/kerchner

Coding consultations (**R**, Python, HTML/CSS, etc.): calendly.com/gwul-coding