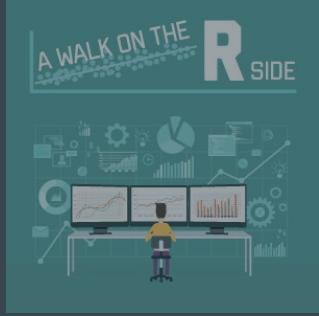
# A Walk on the R Side

## an introduction to R for data analysis

•••

GW Libraries/STEMWorks Workshop
Spring 2018

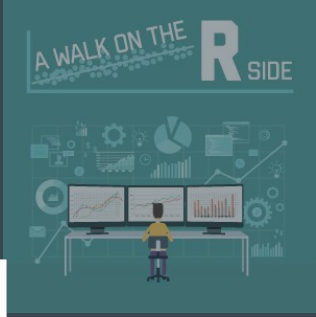
go.gwu.edu/gwlibrworkshop




# 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




software carpentry

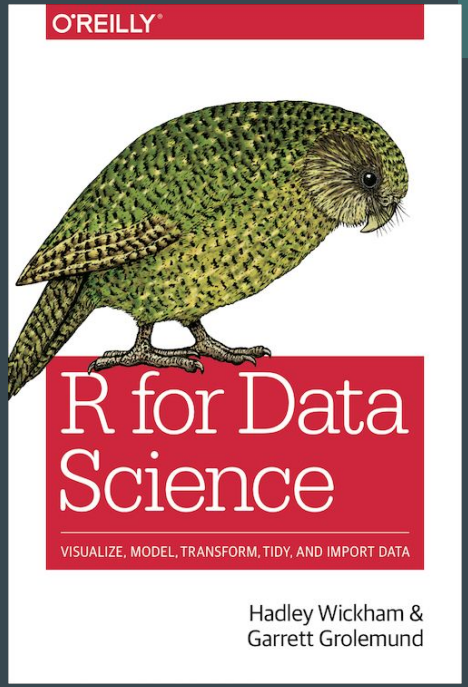
Teaching basic lab skills for research computing

DATA CARPENTRY

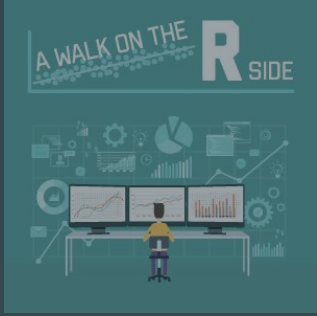
BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

R Tutorial

An R Introduction to Statistics



# Workshop Housekeeping




Ask questions!

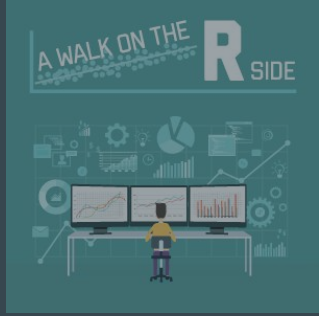
Respect every question and person asking the question

Help each other out!

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

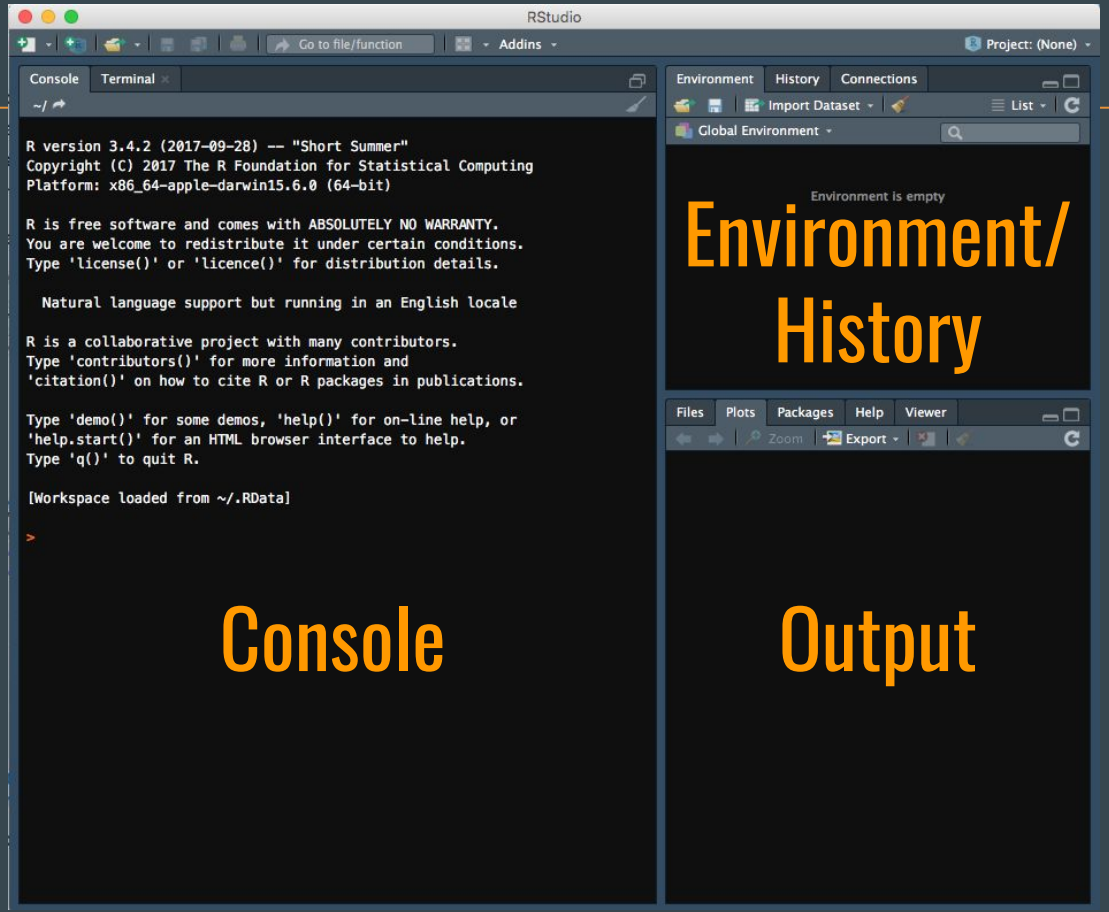
Stay as long as you like

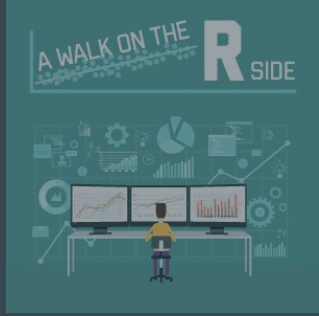
# About R



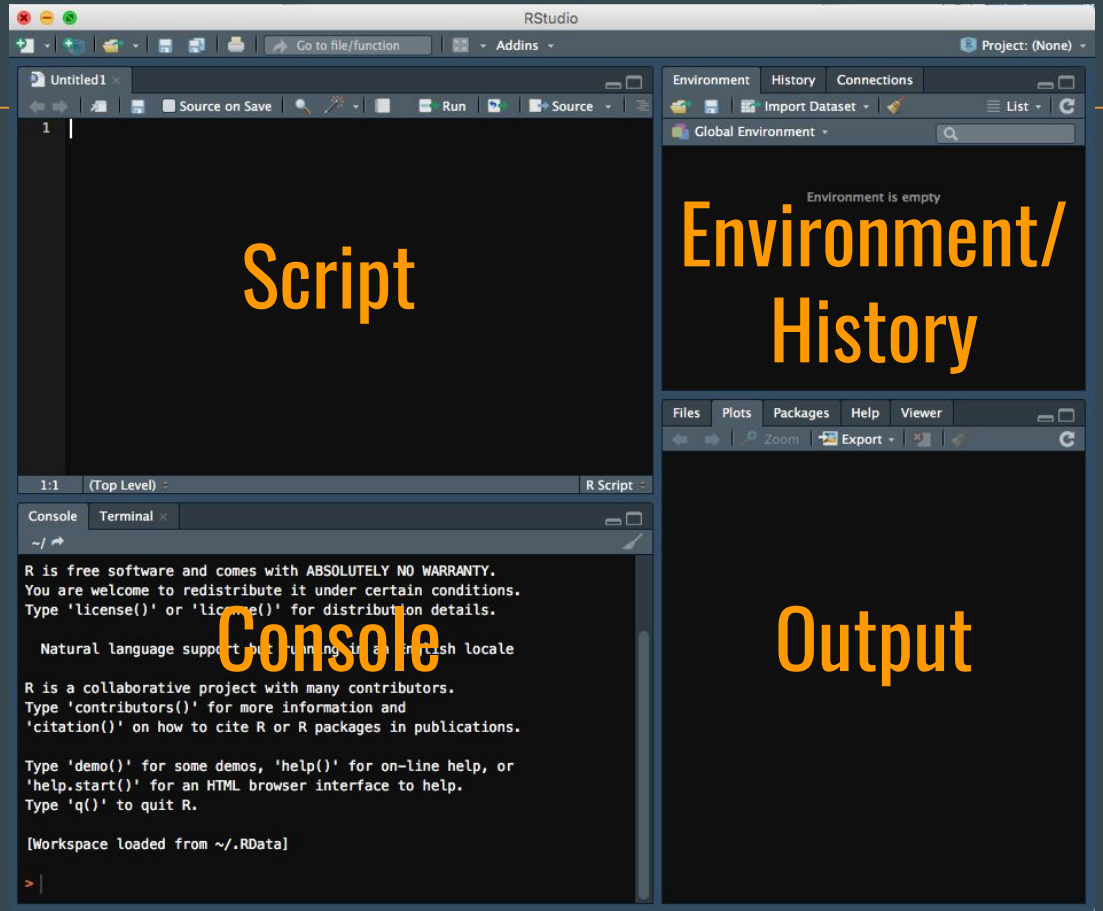

- Free/Open source
- For statistical computing and graphics
- CRAN - r-project.org
  - R packages
  - R journal
  - ...

# R Studio

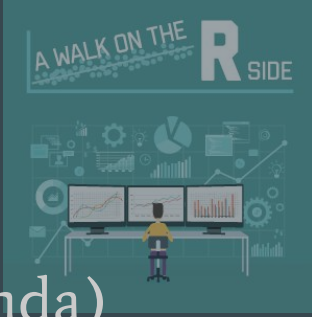

RStudio
Go to file/function
Addins
Project: (None)
Console
Terminal
~/
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to 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.
[Workspace loaded from ~/.RData]
>
Console
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Environment/
History
Files
Plots
Packages
Help
Viewer
Zoom
Export
Output


A WALK ON THE R SIDE

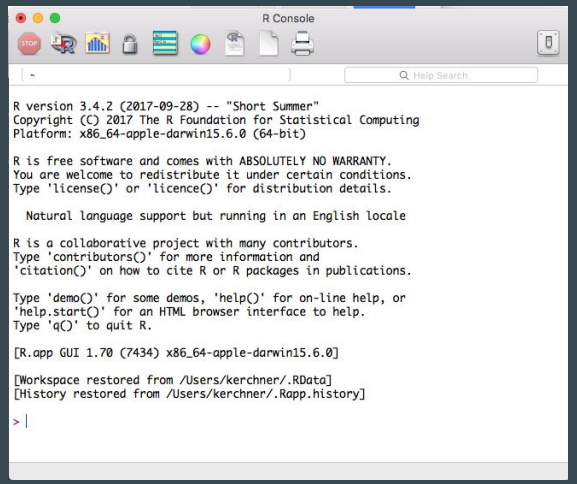
# R Studio


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
Script
1:1
(Top Level)
R 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
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Environment/
History
Files
Plots
Packages
Help
Viewer
Zoom
Export
Output

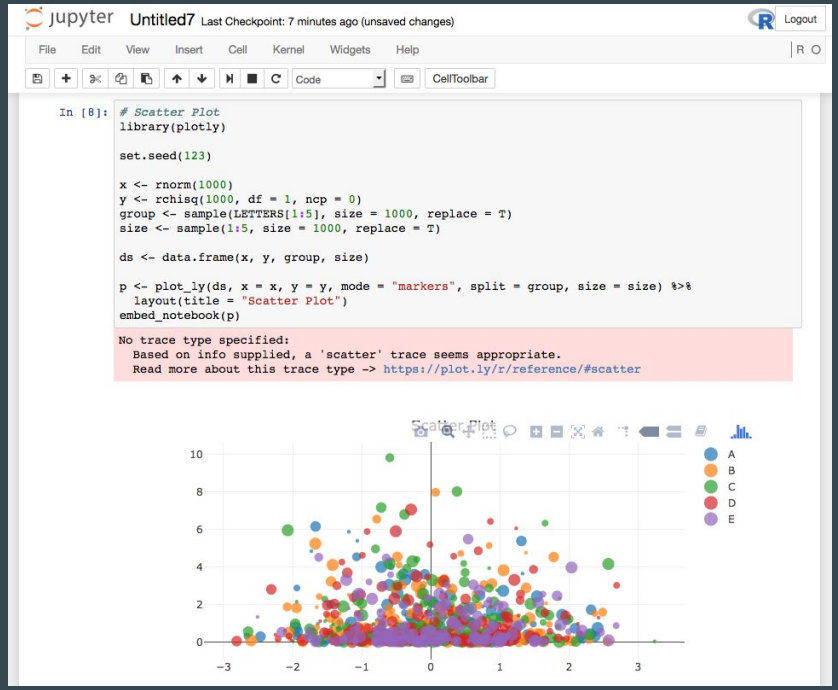
# Other Ways To Use R

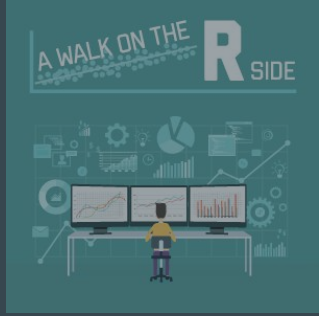



## Plain R console



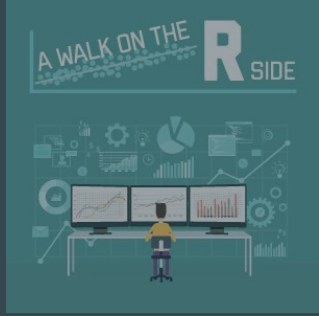

## Jupyter Notebook (e.g. in Anaconda)






# Let's tRy it!

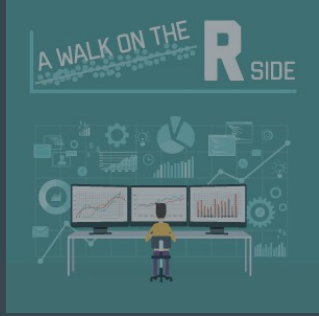
# Variables

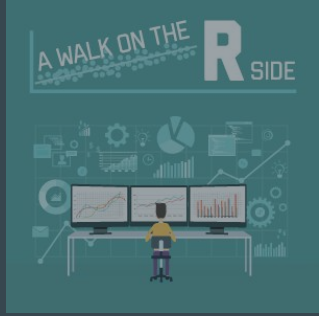



- Try using R as a "calculator" in the Console
  - Try some mathematical functions, too
- Create some variables
  - variable naming
  - `<-` for assigning values to variables (Option - on Mac, Alt - on Win)
  - numeric, character, logical
  - Watch the Environment pane!
  - `class()`
  - Coercion w/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

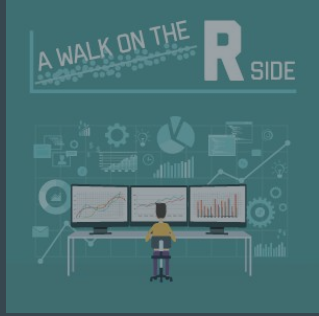
# Logical Expressions

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





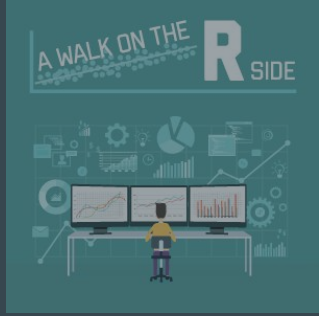

# Vectors




# Vectors

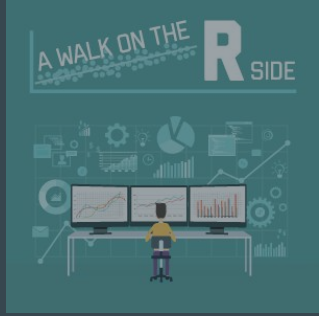
- A vector is
  - A sequence of data elements (components) all of the same type.
- Create vectors with c()  (short for "combine")

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

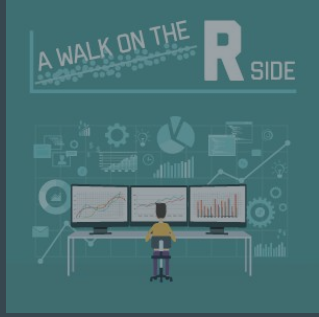



# Data Frames

# 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

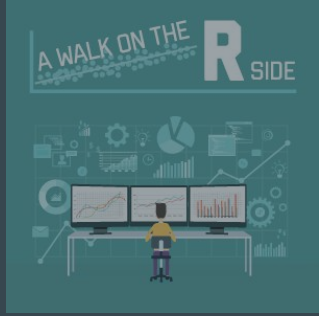
| |
|---|
| 10.2 |

## Vector

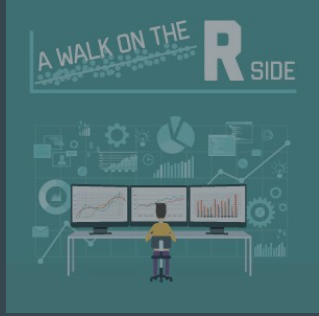
| | |
|---|---|
| 1 | 10.2 |
| 2 | 11.3 |
| 3 | 11.5 |
| 4 | 12.0 |

## 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 |

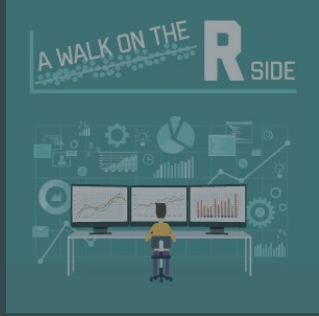



# A brief word on `list` and `matrix`



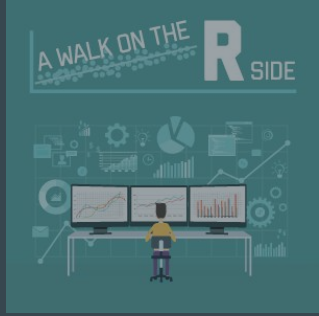

# Projects in RStudio

# Projects in RStudio

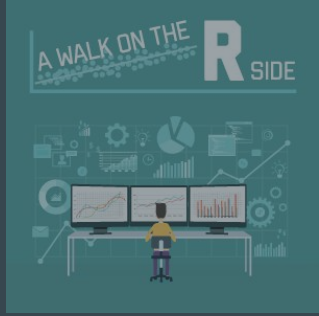



Recommendations:

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




# It's time to import some data!




# Data Importing

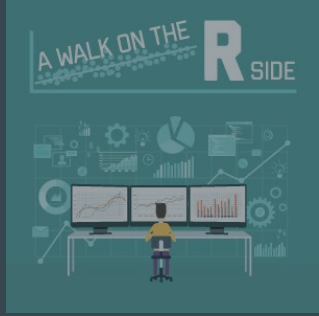
- Prepare data as "tidy"
  - rectangular
  - one table per file
  - rows are observations, columns are variables

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

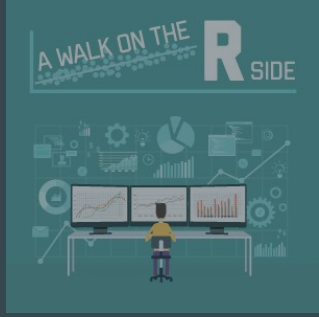
- A word about "big data" (consider **`data.table`**)

# 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

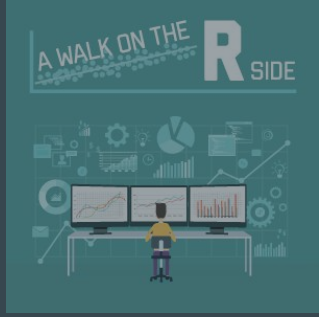



# Tidyverse Core Packages

tidyverse.org

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

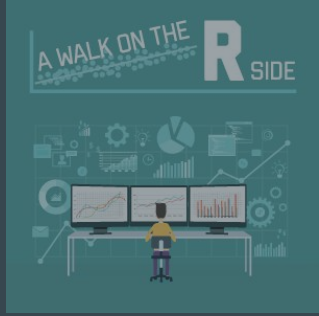
# Other often-used R packages




| | |
|---|---|
| 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 |

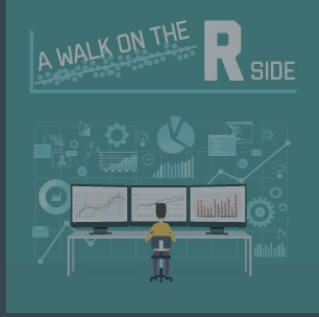
# Exploring Data

- head, tail
- subsetting
- slicing and dicing



# Data Wrangling

flickr.com/photos/thewomensmuseum/3607075017




# Data Transformation using the dplyr package

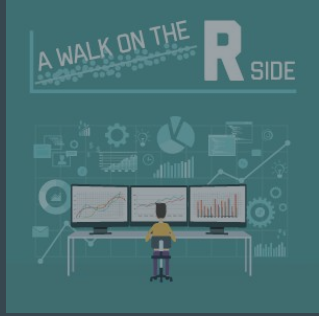
- filter()
- arrange()
- select()
- mutate()
- summarize()
- group_by()
- ...

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

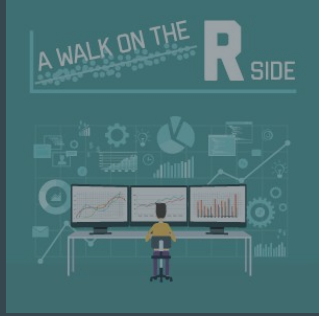
**(shortcut: control-shift-M)**

# Data Tidying with dplyr

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

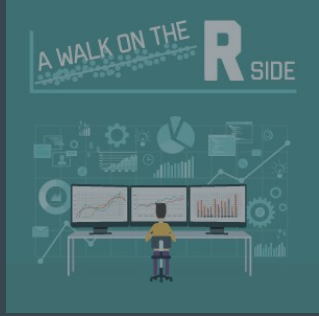


# Joining with dplyr



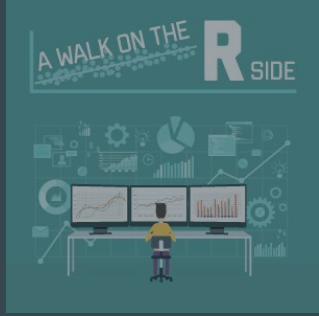

"Merges" tables together

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



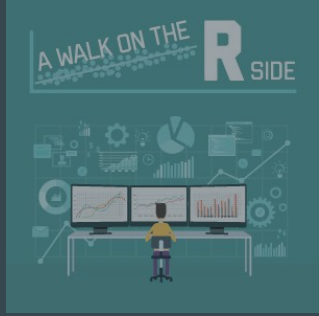

# Data Visualization

# Data Visualization

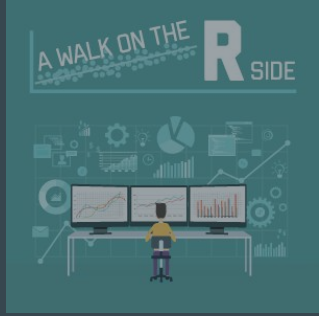



3 main packages:

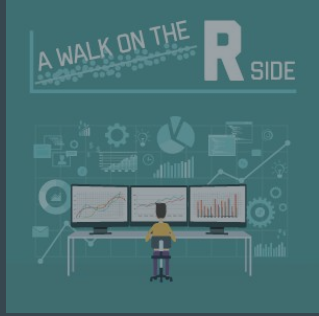
- "base R"
- lattice
- ggplot2



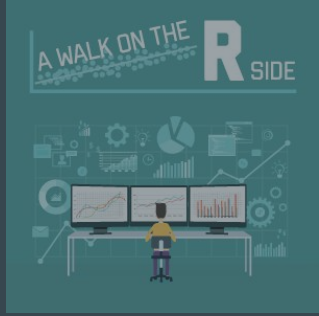

# Data Analysis



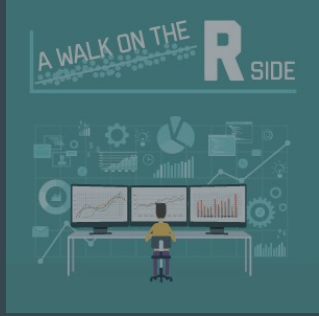

# Functions



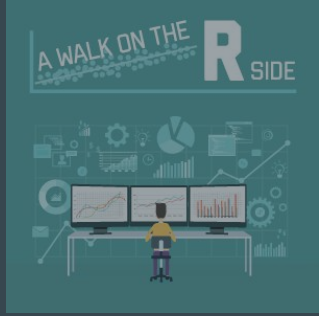

# R Markdown




# R Shiny




# Some Handy R Links




# Tutorials

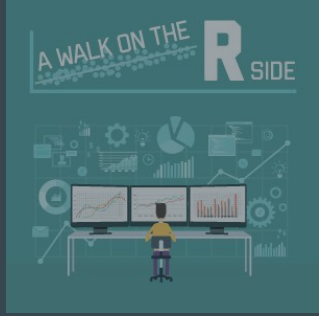
- Software Carpentry:
  - http://swcarpentry.github.io/r-novice-inflammation
  - http://swcarpentry.github.io/r-novice-gapminder
- Data Carpentry:
  - http://datacarpentry.github.io/R-ecology-lesson/
  - http://www.datacarpentry.org/R-genomics/
- Lynda.com lynda.it.gwu.edu - 3 video courses (~12 hours)
- r-tutor.com/r-introduction
  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:

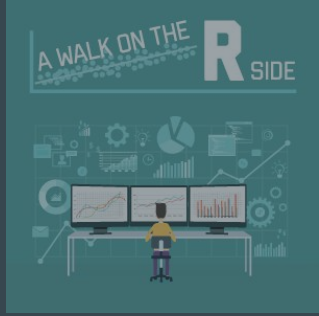
- PSC 2102 - Fall 2018 (and possibly Summer II, 2018)
  Visualizing and Modeling Politics
  Prof. Eric Lawrence
- PPPA 6085 - Data Visualization
- TBD - Public Health, Prof. Helmchen - Fall 2018

# Reference Links




- r-project.org
- R search engine: rseek.org
- rstudio.com
  - Cheat Sheets
- stackoverflow.com

# Thanks!





- Dan Kerchner kerchner@gwu.edu
- Dr. Kes Schroer schroerk@gwu.edu
- Vishwesh Haldevanekar vishwesh_s_h@gwu.edu


These slides: go.gwu.edu/gwlibrworkshop

R or Statistics Appointments with Vishwesh:
calendly.com/vishwesh_s_h