Up your data game: How to use R to wrangle, analyze, and visualize data faster and better

MERL Tech DC 2018

Charles Guedenet, MEL Technical Advisor, IREX
Jonathan Seiden, Learning Research Specialist, Save the Children

Please: Plug in your laptop and check its wifi
download your dataset and r files here: **http://bit.ly/merl-r**

Objectives

By the end of this session, you should:

1. Have R and Rstudio setup on your computer
2. Have a better understanding of what R programming is and what it can do for you
3. Learn about useful R tools (functions and packages) that you can use in your own work
4. Feel intrigued (and excited?) enough about R to pursue further learning

What is R?

A programming language for data manipulation

Command-line driven vs. Not point-and-click

Who uses it?

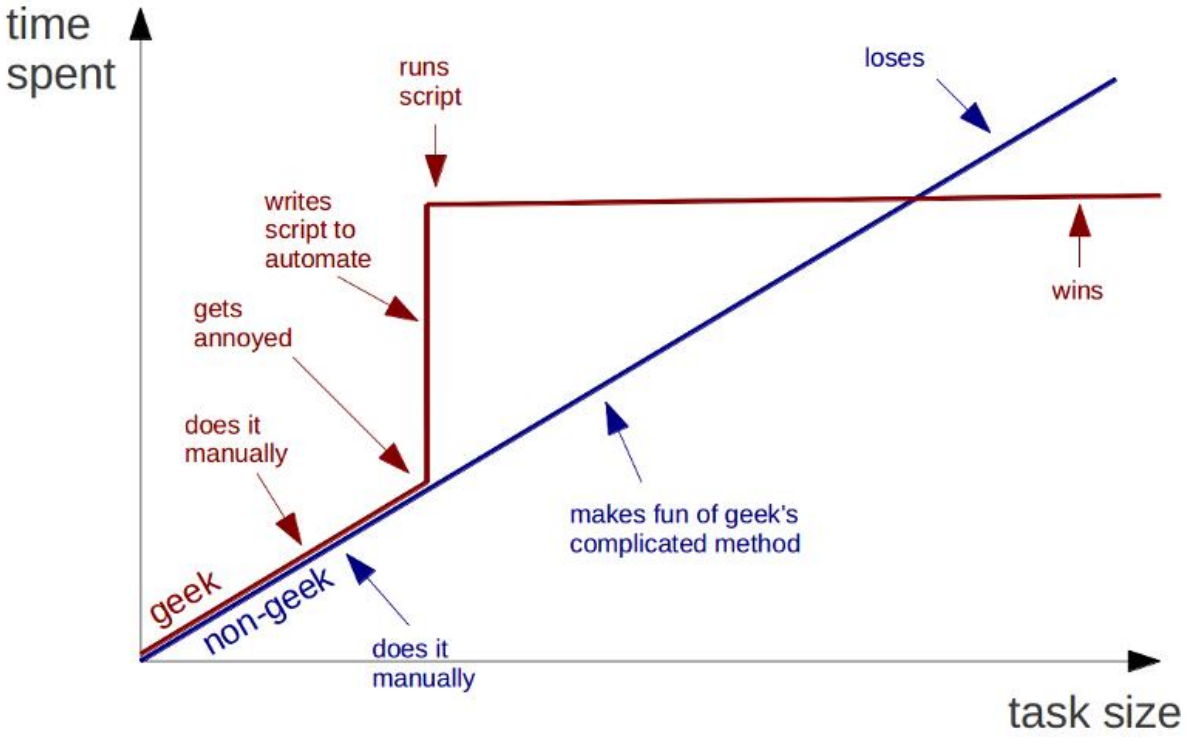
- Academics, journalists, statisticians, open data enthusiasts

Who uses R?

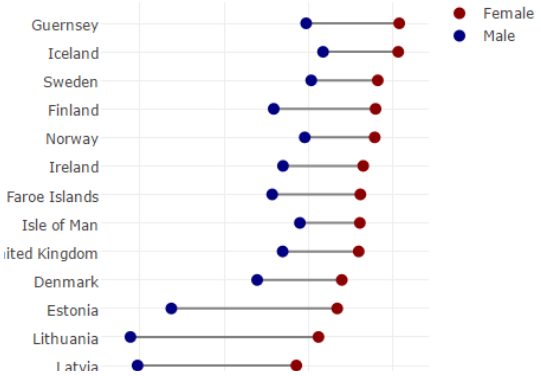
Capgemini
ORBITZ
McKinsey&Company
KPMG
Gartner
Zillow
trulia
Ford
wipro
Mindtree
Welcome to possible
PNC
Bing
NOVARTIS
genpact
UBER
iGATE
KeyBank
paytm
Thomas Cook
TechCrunch
Infosys
Fractal
Morgan Stanley
The New York Times
MSD
TATA
ANZ
TATA CONSULTANCY SERVICES
J.P.Morgan
Deloitte.
accenture
High performance. Delivered.
BAIN & COMPANY
snapdeal
Booz | Allen | Hamilton
strategy and technology consultants
Cognizant
hp
JOHN DEERE
BCG
The Boston Consulting Group
Tech Mahindra
HSBC

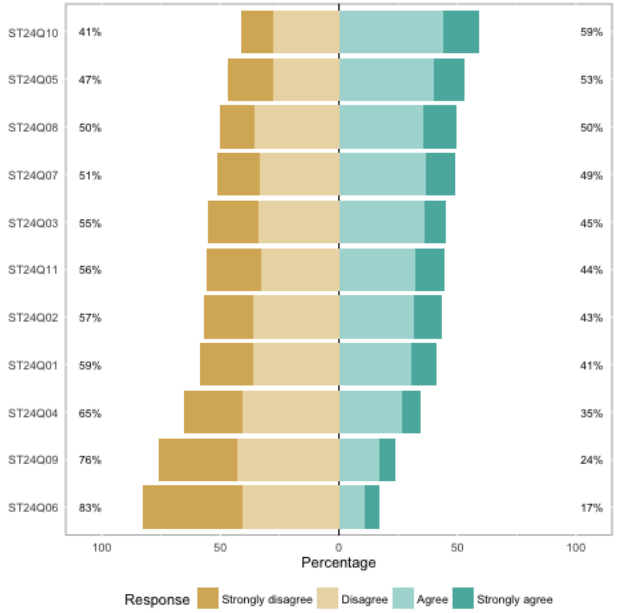
Pro #1: a reproducible workflow!

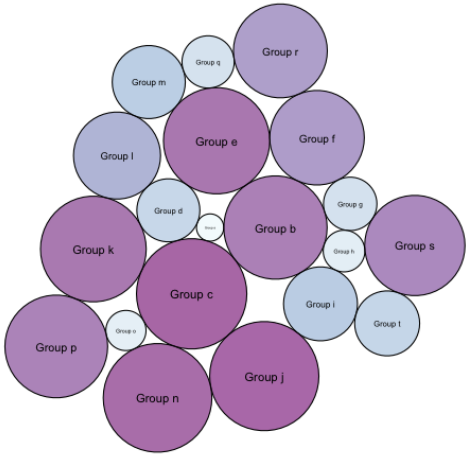
Geeks and repetitive tasks

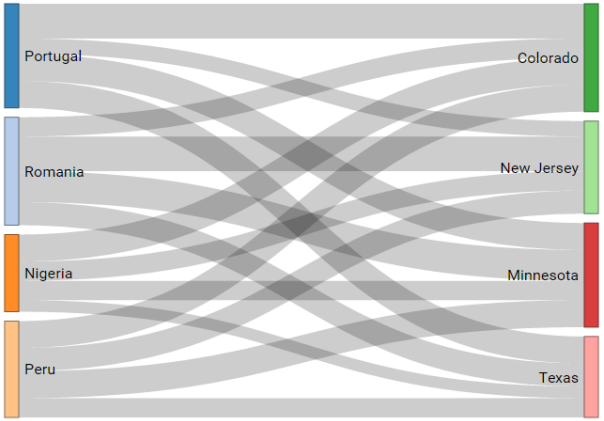


Pro #2: top notch data visualization!









https://www.r-graph-gallery.com/

Pro #3: Flexible & Comprehensive

Work with data across the data life cycle

Pro #4: Large & active community

- ✓ **Tutorials, blogs, websites**
  R-bloggers.com – news and tutorials by 750+ bloggers
  Stats.stackexchange.com

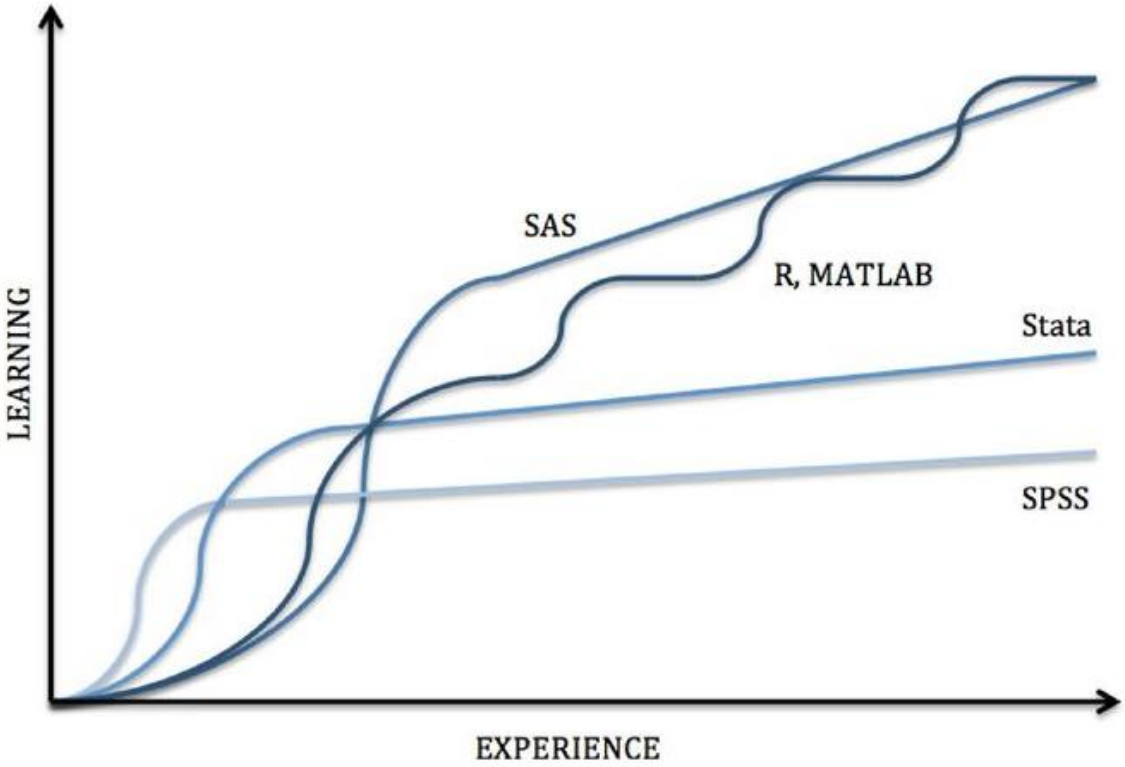
- ✓ **So much free code! Copy + paste**
  Kaggle.com
  Github.com

- ✓ **A package for everything - +13k packages**
  www.r-pkg.org

Pro #5: It's free!

Compare with:	ANNUAL Cost
SPSS	$1,200 (statistics only) + $$ for addons
SAS	$8,700 first year (basic Analytics Pro)
STATA	$595 - $1,500

Con #1: Steep learning curve



Source: https://sites.google.com/a/nyu.edu/statistical-software-guide/summary

Pros and Cons

Pros

- ✓ Programming language = reproducible work & huge efficiency gains
- ✓ Top notch data visualization capabilities
- ✓ Flexible & comprehensive
- ✓ Active R community
- ✓ Free and open source

Cons

- ✓ A steep learning curve for programming newbies
- ✓ Colleagues/friends may still prefer STATA SPSS users

Getting setup

1. Install R on your computer
http://lib.stat.cmu.edu/R/CRAN/

Windows or MacOS - choose one of the *precompiled binary* distributions (i.e., ready-to-run applications) linked at the top of the R Project's webpage.

2. Install RStudio
https://www.rstudio.com/products/rstudio/download/

RStudio Desktop
Open Source License

FREE

DOWNLOAD

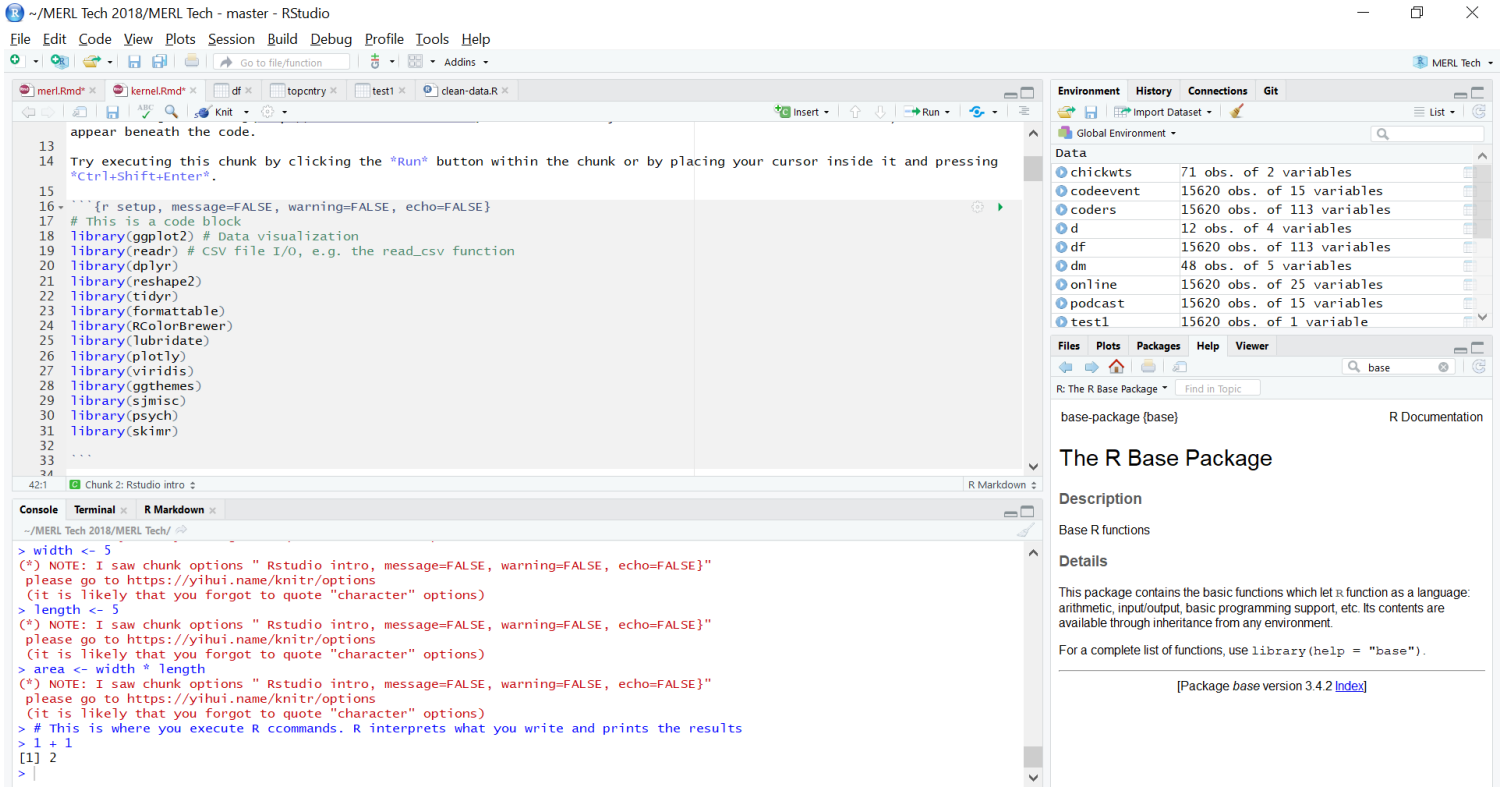
Learn More

Introduction to Rstudio

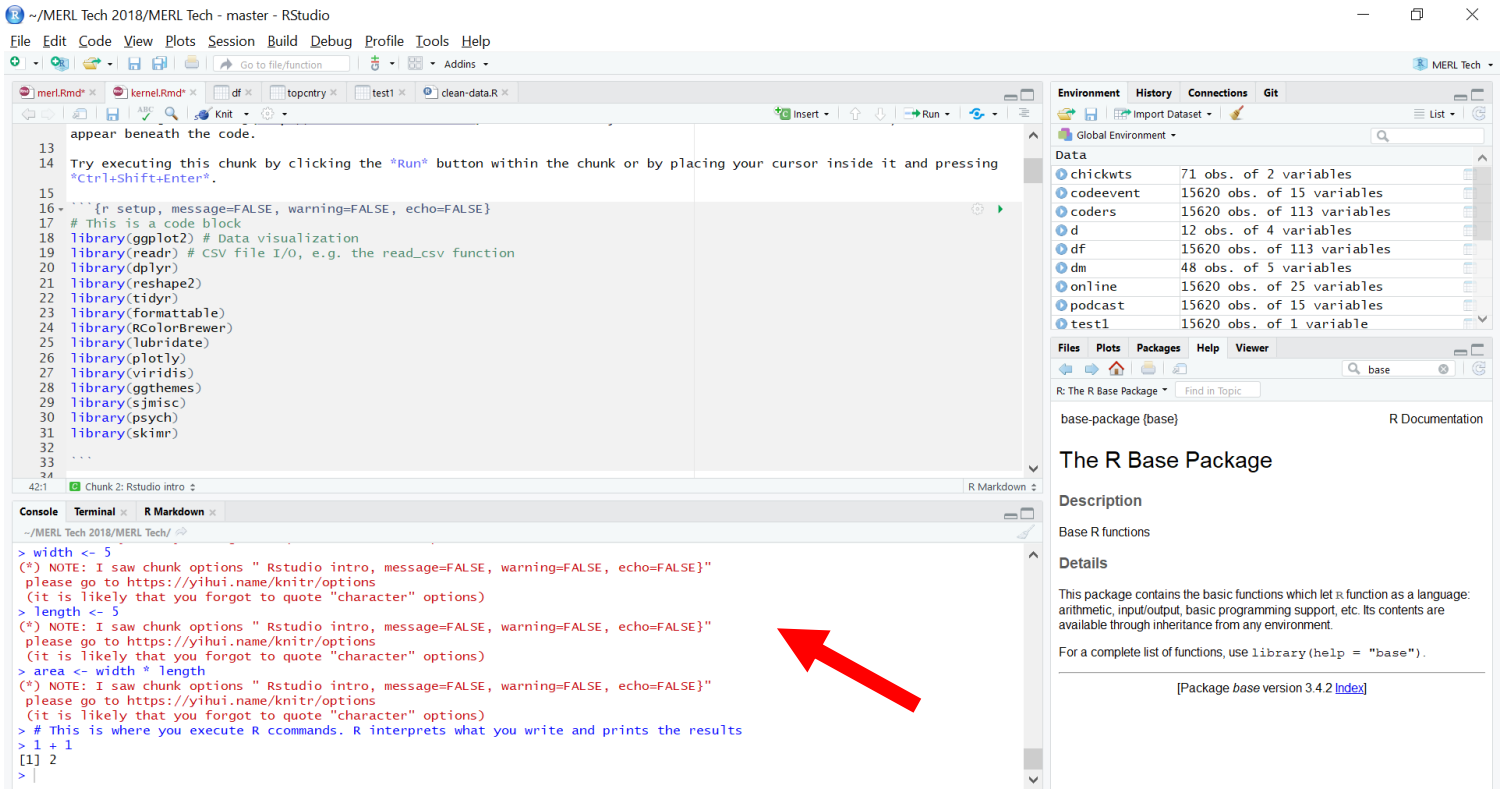
Go to your Start menu and in programs start Rstudio by clicking on the Rstudio icon

Rstudio basics

This is what the Rstudio interface looks like



Rstudio basics: the console

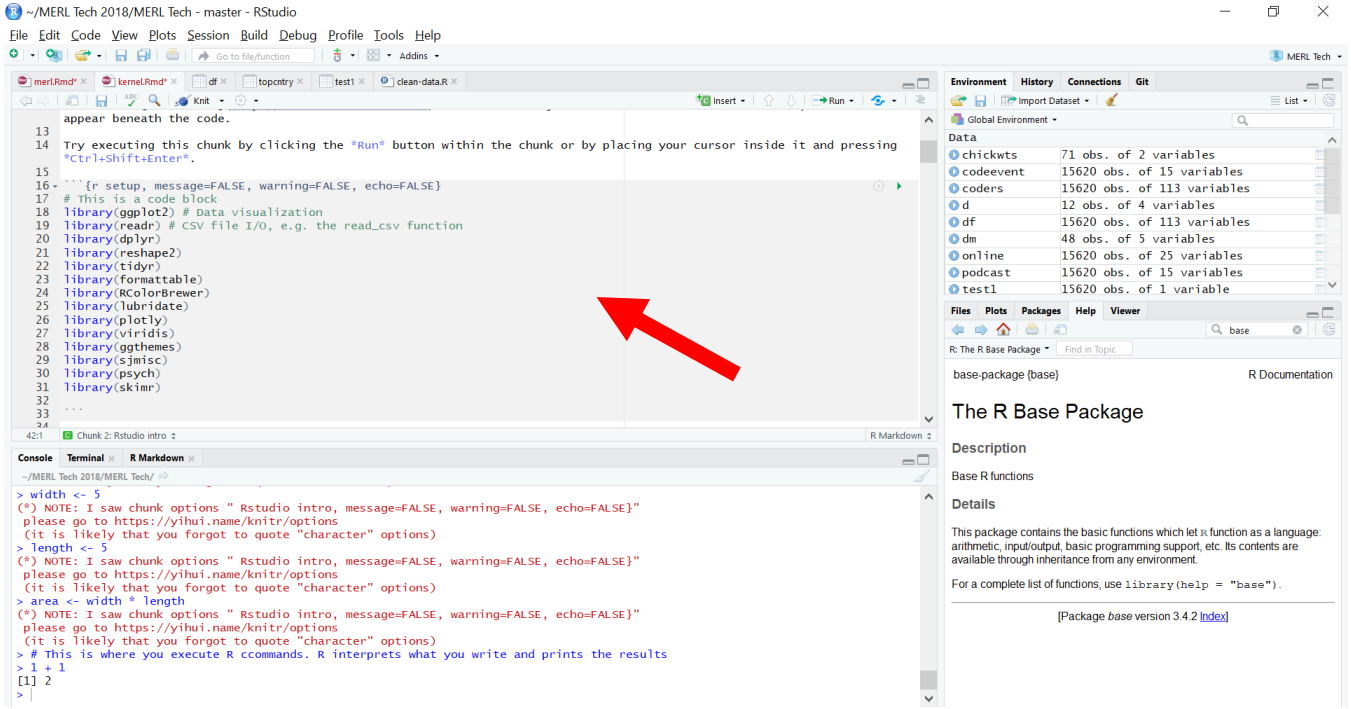


Introduction to Rstudio: Source tabs

This is a built-in text editor

Open an empty script File → New File → R Script

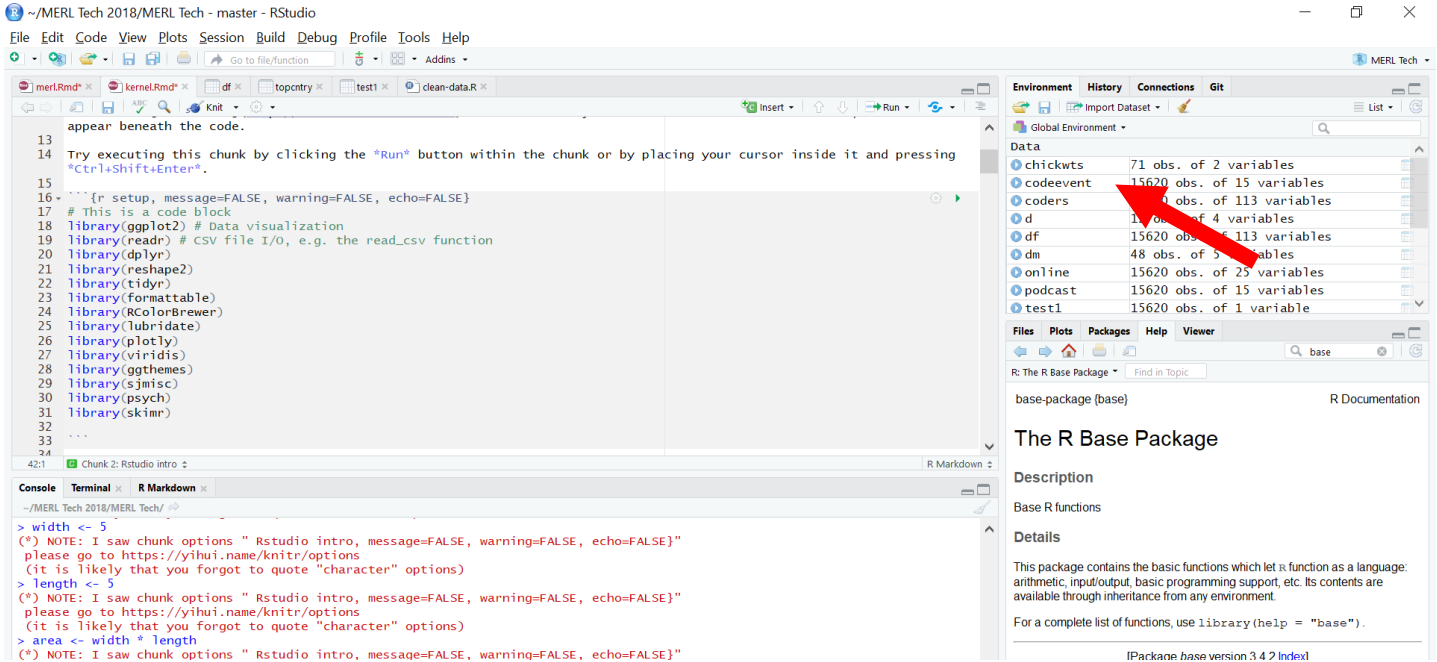
You can write a script and then execute it in the console using Ctrl Shift Enter



Global Environment & History

Environment tab - where you can see the values and functions that you've created or imported

History tab –where you can see a list of key strokes you've entered in console

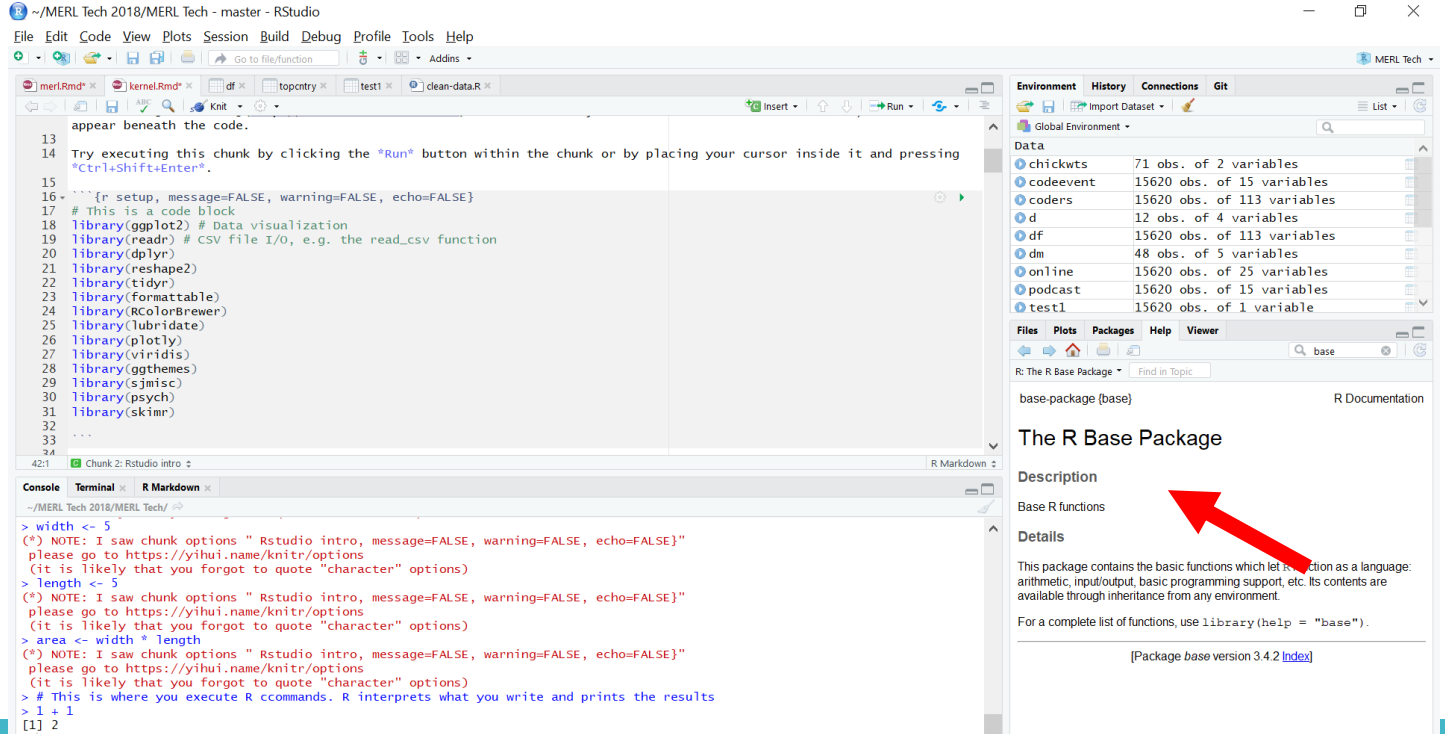


Files, plots, packages, help

Files – navigate your computer's files

Packages – find and install packages

Help – search and find help with functions and packages



Files, plots, packages, help

A **function** is a set of statements (or instructions) organized to perform a specific task.
e.g. sqrt(), sd(), mean()

Packages are collections of R functions, data, and compiled code in a well-defined format.

The **library** is where packages are storied on your computer.

Basic tips

1. To run a command/function, **click the “Run” button or press Ctrl + Enter**
2. **R is case sensitive.** Make sure your spelling and capitalization are correct.
3. **The $ symbol** is used to select a particular column within a table (e.g., table$column).
4. **The # symbol:** Any text that you do not want R to act on (such as comments, notes, or instructions) needs to be preceded by the # symbol (a.k.a. hash-tag, comment, pound, or number symbol). R ignores the remainder of the script line following #.

From <http://ncss-tech.github.io/stats_for_soil_survey/chapters/1_introduction/1_introduction.html>