

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?

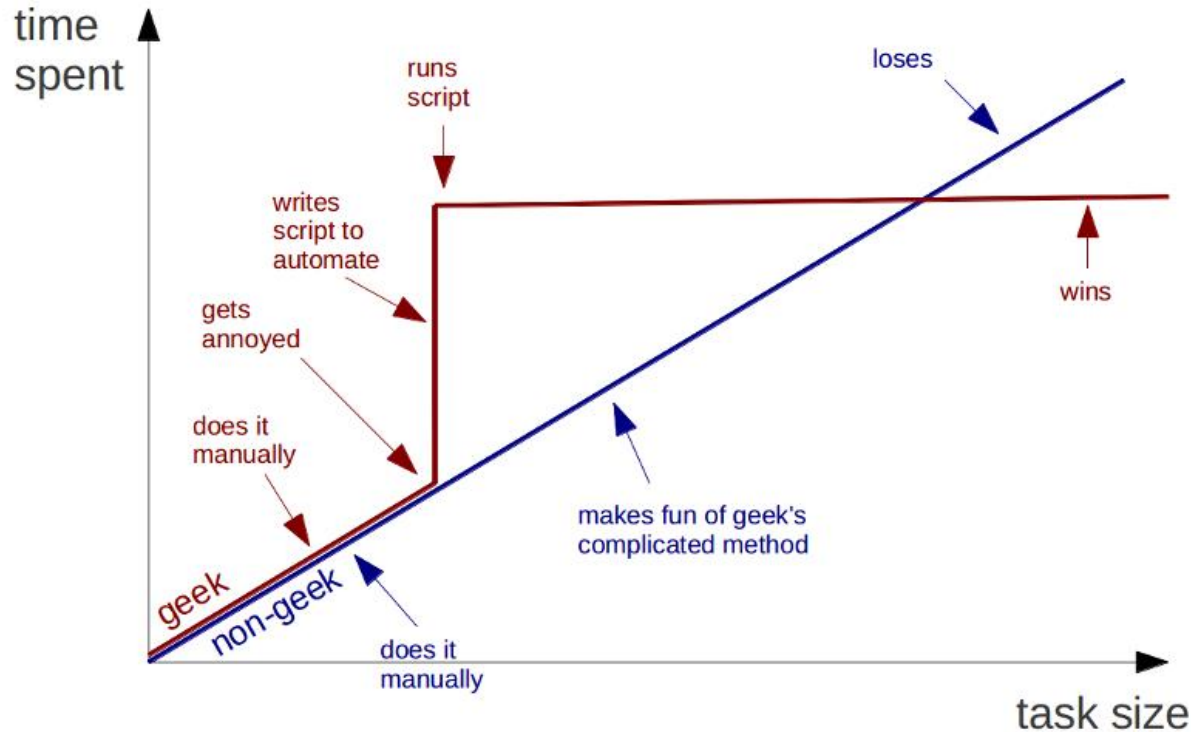
- Academics, journalists, statisticians, open data enthusiasts

Who uses R?

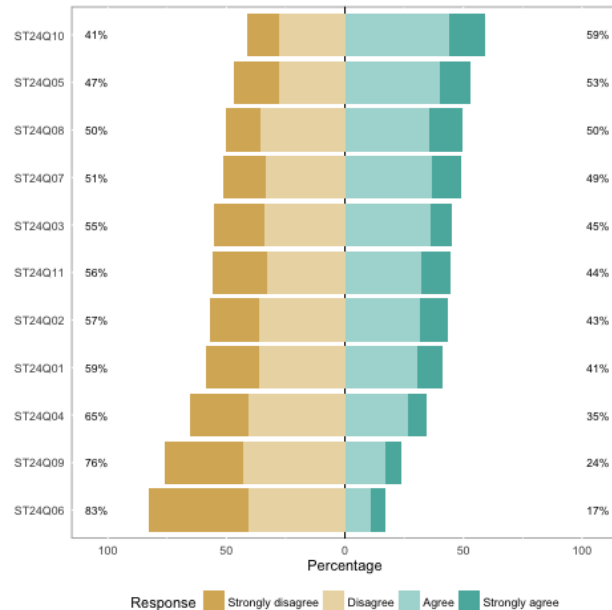
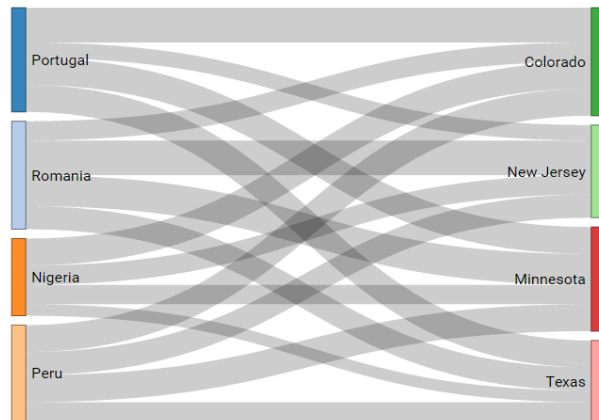
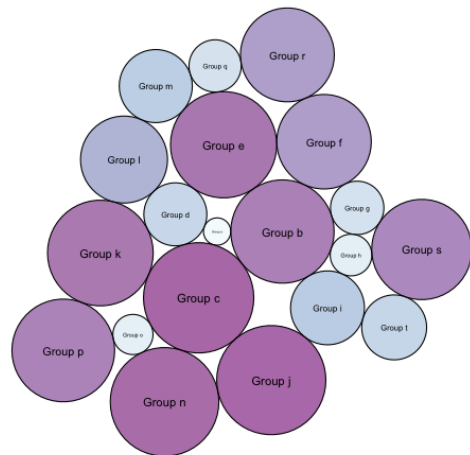
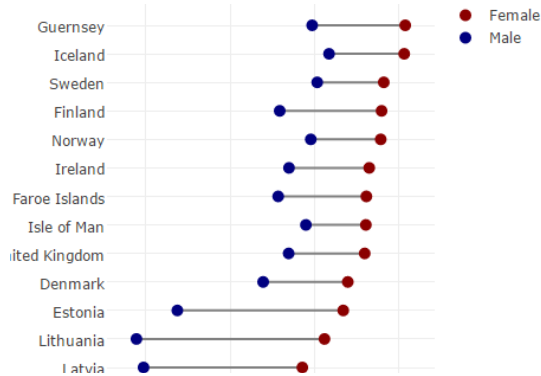


Pro #1: a reproducible workflow!

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

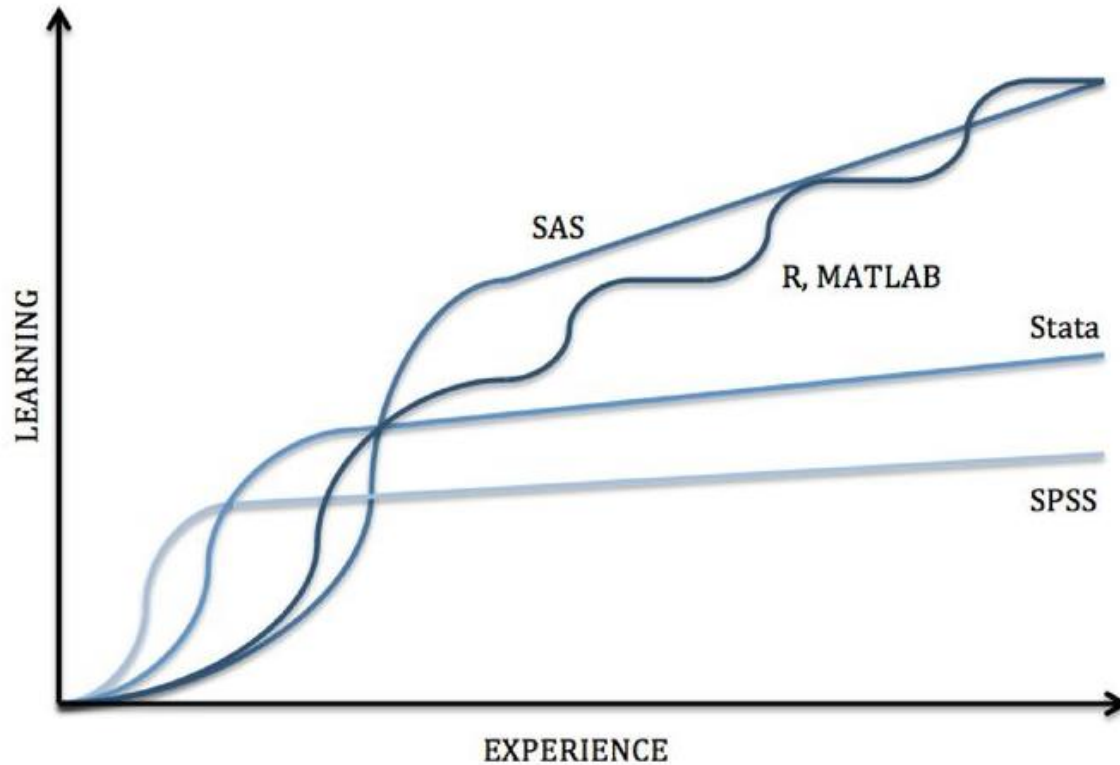
SAS

\$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/>

RStudio Desktop
Open Source License

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.

FREE

2. Install RStudio

<https://www.rstudio.com/products/rstudio/download/>

DOWNLOAD

Learn More

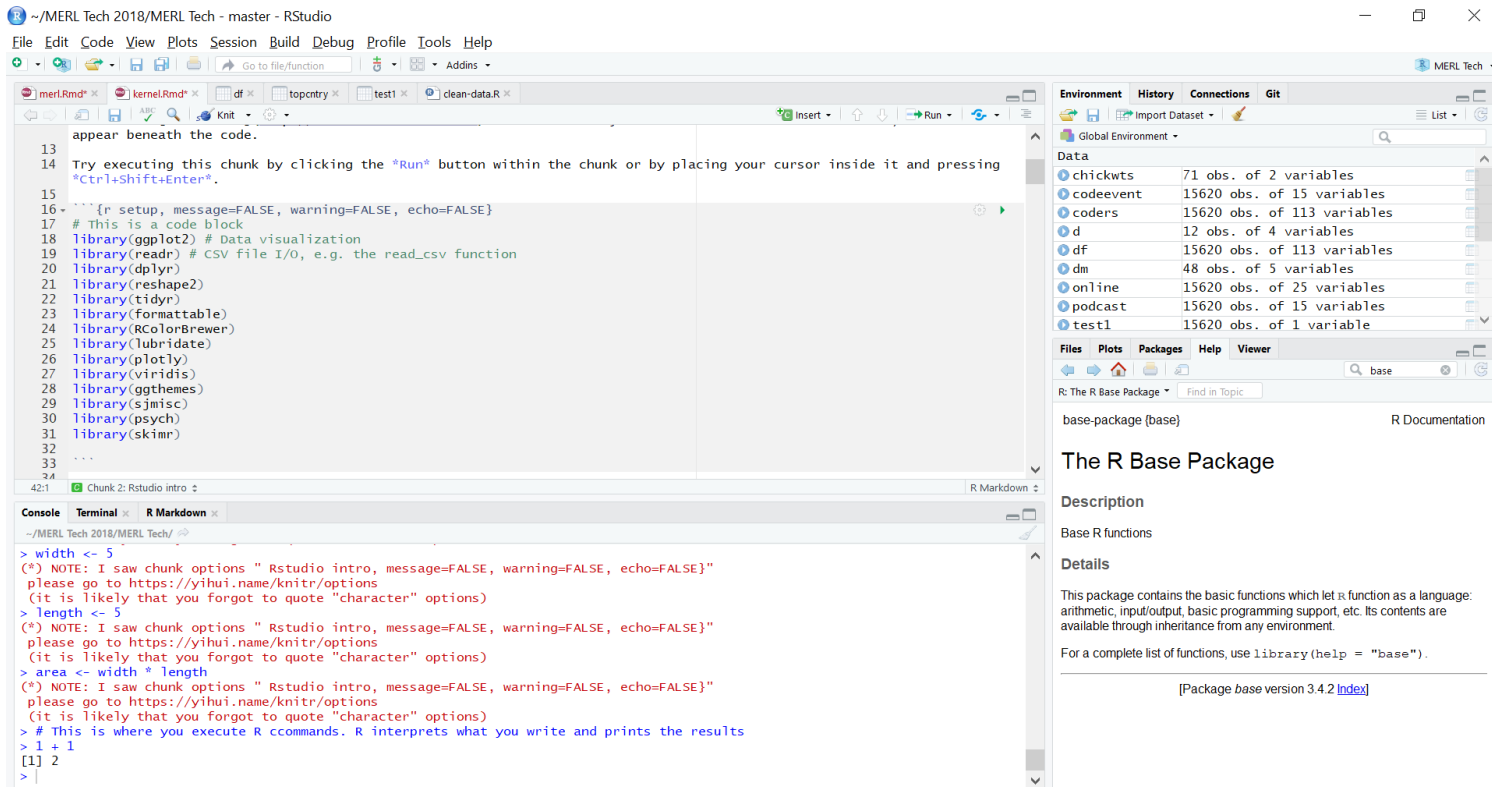
Introduction to Rstudio

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

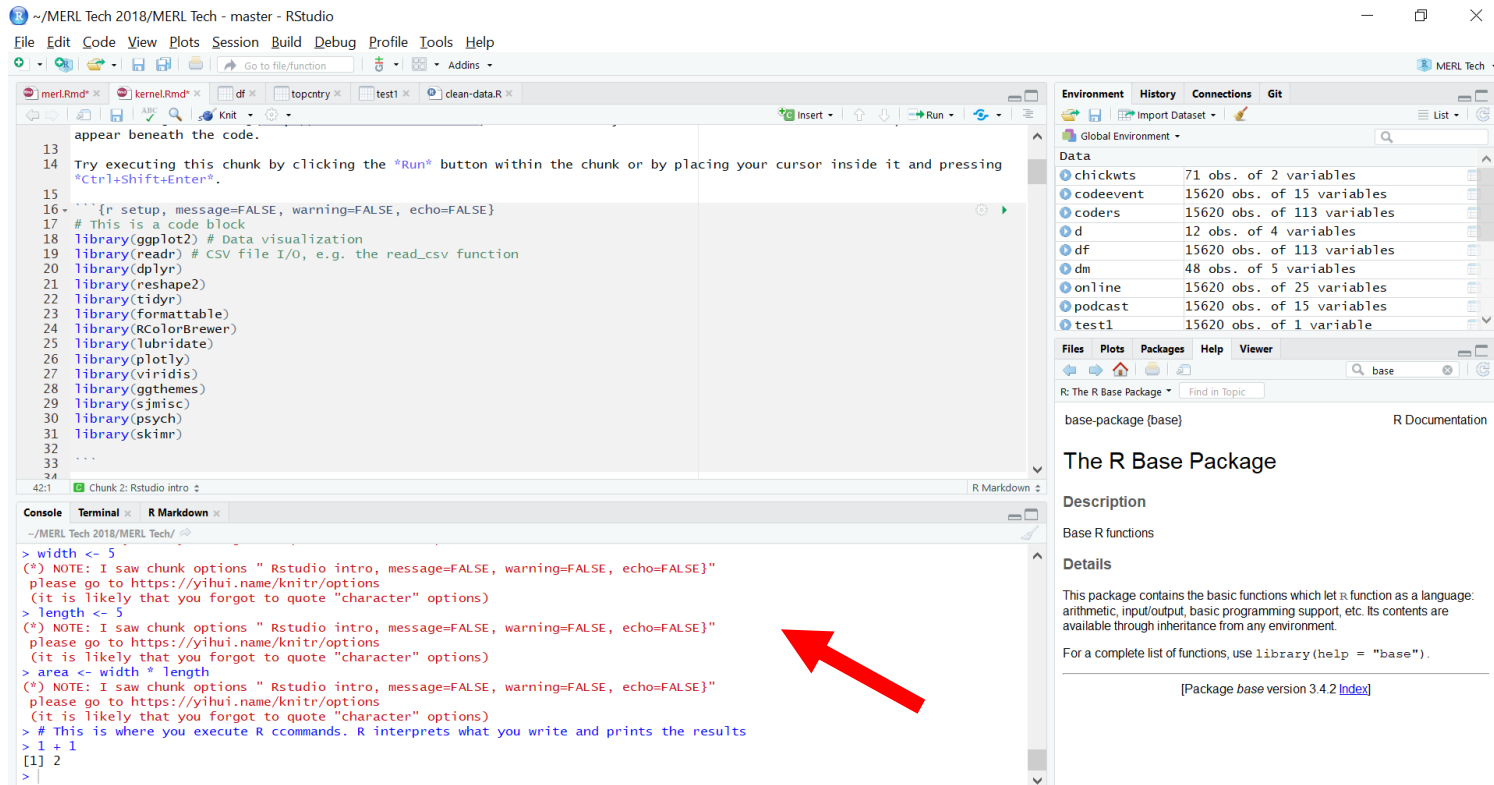


Rstudio basics

This is what the Rstudio interface looks like



Rstudio basics: the console



The screenshot displays the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar shows icons for file operations, running code, and other functions. The main editor window contains a code chunk with the following R code:

```
13 appear beneath the code.
14 Try executing this chunk by clicking the "Run" button within the chunk or by placing your cursor inside it and pressing
15 "Ctrl+Shift+Enter".
16 ---{r setup, message=FALSE, warning=FALSE, echo=FALSE}
17 # This is a code block
18 library(ggplot2) # Data visualization
19 library(readr) # CSV file I/O, e.g. the read_csv function
20 library(dplyr)
21 library(reshape2)
22 library(tidyr)
23 library(formattable)
24 library(RColorBrewer)
25 library(lubridate)
26 library(plotly)
27 library(viridis)
28 library(ggthemes)
29 library(sjmisc)
30 library(psych)
31 library(skimr)
32 ...
33
34
```

The console window at the bottom shows the output of the code chunk, including several notes and the results of the calculations:

```
> width <- 5
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> length <- 5
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> area <- width * length
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> # This is where you execute R ccommands. R interprets what you write and prints the results
> 1 + 1
[1] 2
>
```

A red arrow points to the console output.

The right sidebar shows the Environment pane with a table of objects in the Global Environment:

Object	Size	Variables
chickwts	71 obs.	2 variables
codeevent	15620 obs.	15 variables
coders	15620 obs.	113 variables
d	12 obs.	4 variables
df	15620 obs.	113 variables
dm	48 obs.	5 variables
online	15620 obs.	25 variables
podcast	15620 obs.	15 variables
test1	15620 obs.	1 variable

The Files pane shows the R Base Package and its documentation.

The R Base Package

Description

Base R functions

Details

This package contains the basic functions which let R function as a language: arithmetic, input/output, basic programming support, etc. Its contents are available through inheritance from any environment.

For a complete list of functions, use `library(help = "base")`.

[Package base version 3.4.2 [Index](#)]

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

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and execution. The main Source editor window shows an R script with the following content:

```
appear beneath the code.  
13  
14 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing  
   *Ctrl+Shift+Enter*.  
15  
16 ```{r setup, message=FALSE, warning=FALSE, echo=FALSE}  
17 # This is a code block  
18 library(ggplot2) # Data visualization  
19 library(readr) # CSV file I/O, e.g. the read_csv function  
20 library(dplyr)  
21 library(reshape2)  
22 library(tidyr)  
23 library(formattable)  
24 library(RColorBrewer)  
25 library(lubridate)  
26 library(plotly)  
27 library(viridis)  
28 library(ggthemes)  
29 library(sjmisc)  
30 library(psych)  
31 library(skimmr)  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42
```

A red arrow points to the Run button (a green play icon) in the toolbar above the Source editor. The Environment pane on the right shows the Global Environment with a list of objects: chickwts, codeevent, coders, d, df, dm, online, podcast, and test1, each with its dimensions. The Console pane at the bottom shows the output of the R script, including the execution of the library() functions and the calculation of the area of a rectangle.

Console Terminal x R Markdown x

```
~/MERL Tech 2018/MERL Tech/ > width <- 5  
(* NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE]"  
please go to https://yihui.name/knitr/options  
(it is likely that you forgot to quote "character" options)  
> length <- 5  
(* NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE]"  
please go to https://yihui.name/knitr/options  
(it is likely that you forgot to quote "character" options)  
> area <- width * length  
(* NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE]"  
please go to https://yihui.name/knitr/options  
(it is likely that you forgot to quote "character" options)  
> # This is where you execute R commands. R interprets what you write and prints the results  
> 1 + 1  
[1] 2  
>
```


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

The screenshot displays the RStudio interface. The main editor window shows an R script with comments and library loading code. The Environment tab on the right lists objects in the Global Environment, with a red arrow pointing to the 'df' object. The History tab below it shows the execution of the script. The Console window at the bottom shows the output of the script, including notes about chunk options and the calculation of width and length.

Environment tab:

Object	Value
chickwts	71 obs. of 2 variables
codeevent	15620 obs. of 15 variables
coders	15620 obs. of 113 variables
d	15620 obs. of 4 variables
df	15620 obs. of 113 variables
dm	48 obs. of 5 variables
online	15620 obs. of 25 variables
podcast	15620 obs. of 15 variables
test1	15620 obs. of 1 variable

History tab:

```
42:1 [1] width <- 5
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> length <- 5
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> area <- width * length
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE"
```

Files, plots, packages, help

Files – navigate your computer's files

Packages – find and install packages

Help – search and find help with functions and packages

The screenshot displays the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar contains icons for file operations and running code. The main script editor shows R code for setting up a message, loading libraries (ggplot2, readr, dplyr, reshape2, tidyr, formattable, RColorBrewer, lubridate, plotly, viridis, ggthemes, sjmisc, psych, skimr), and a chunk of R code for data visualization. The console at the bottom shows the execution of the code, including the output of the `width` and `length` functions. The right-hand pane is split into two sections: 'Data' and 'Help'. The 'Data' section lists various datasets with their respective observations and variables. The 'Help' section is currently displaying the documentation for 'The R Base Package', which includes a description of the package's purpose and a list of functions. A red arrow points to the 'Description' section of the help pane.

```
~/MERL Tech 2018/MERL Tech - master - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
merlRmd* kernelRmd* df* topcntry* test1* clean-data.R*
Insert Run
appear beneath the code.
13
14 Try executing this chunk by clicking the "Run" button within the chunk or by placing your cursor inside it and pressing
15 *Ctrl+Shift+Enter*.
16
17 {r setup, message=FALSE, warning=FALSE, echo=FALSE}
18 # This is a code block
19 library(ggplot2) # Data visualization
20 library(readr) # CSV file I/O, e.g. the read_csv function
21 library(dplyr)
22 library(reshape2)
23 library(tidyr)
24 library(formattable)
25 library(RColorBrewer)
26 library(lubridate)
27 library(plotly)
28 library(viridis)
29 library(ggthemes)
30 library(sjmisc)
31 library(psych)
32 library(skimr)
33
34
42:1 Chunk 2: Rstudio intro
R Markdown
Console Terminal R Markdown
~/MERL Tech 2018/MERL Tech/
> width <- 5
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE}"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> length <- 5
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE}"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> area <- width * length
(*) NOTE: I saw chunk options " Rstudio intro, message=FALSE, warning=FALSE, echo=FALSE}"
please go to https://yihui.name/knitr/options
(it is likely that you forgot to quote "character" options)
> # This is where you execute R ccommands. R interprets what you write and prints the results
> 1 + 1
[1] 2
```

Environment History Connections Git

Global Environment

Data

Dataset	Observations	Variables
chickwts	71 obs.	2 variables
codeevent	15620 obs.	15 variables
coders	15620 obs.	113 variables
d	12 obs.	4 variables
df	15620 obs.	113 variables
dm	48 obs.	5 variables
online	15620 obs.	25 variables
podcast	15620 obs.	15 variables
test1	15620 obs.	1 variable

Files Plots Packages Help Viewer

R: The R Base Package Find in Topic

base-package (base) R Documentation

The R Base Package

Description

Base R functions

Details

This package contains the basic functions which let R function as a language: arithmetic, input/output, basic programming support, etc. Its contents are available through inheritance from any environment.

For a complete list of functions, use `library(help = "base")`.

[Package base version 3.4.2 [Index](#)]

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 stored 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>