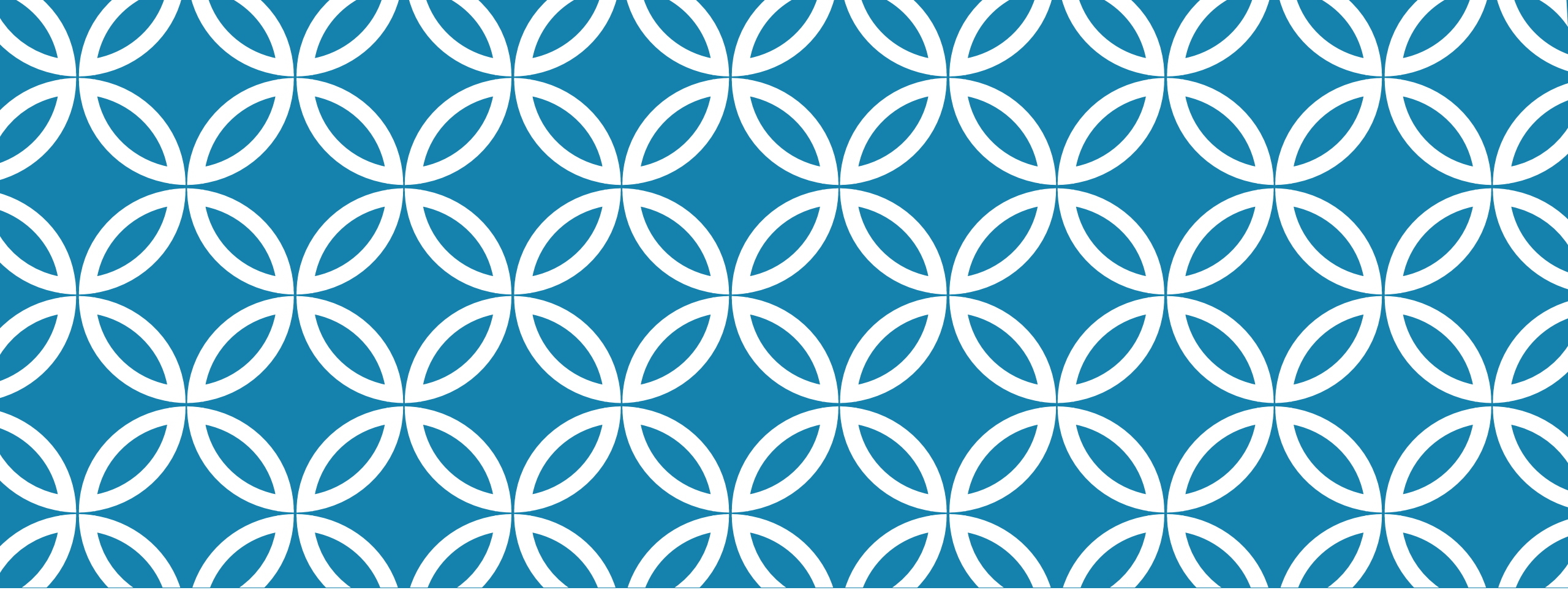


Introduction to R Workshop

Session 2

Reproducible Reporting

May 6, 2019

7:00 am - 8:00 am	BREAKFAST - BALLROOM LOBBY - 2ND FLOOR
8:00 am - 8:10 am	Instructor Instructions
8:10 am - 9:50 am	Introduction to R and RStudio for Reproducible Reporting
9:50 am - 10:10 am	REFRESHMENT BREAK - BALLROOM LOBBY - 2ND FLOOR
10:10 am - 11:50 am	Data Wrangling
12:00 pm - 1:00 pm	LUNCH - BALLROOM LOBBY - 2ND FLOOR
1:00 pm - 2:50 pm	Data Understanding
2:50 pm - 3:10 pm	REFRESHMENT BREAK - BALLROOM LOBBY - 2ND FLOOR
3:10 pm - 5:00 pm	Exploratory Data Analysis

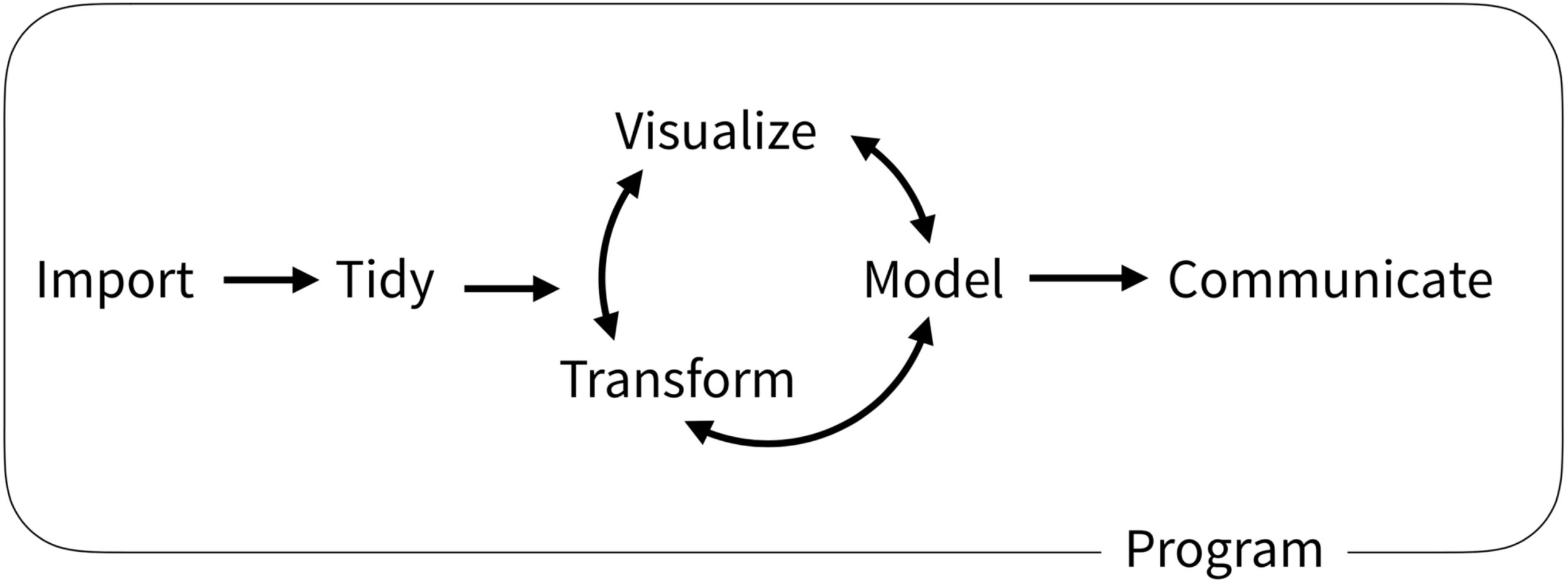
Goals

1. Learn to work within R Markdown for reproducible reports
2. Practice initial steps of the data import and analysis process

Objectives

1. Understand why reproducible reporting is important
2. Create an R Markdown document and generate different types of output files
3. Import data from files into your R environment
4. Create basic summaries of data via tabulation and simple plots

Typical Data Science Pipeline



From *R for Data Science* (https://r4ds.had.co.nz/introduction.html)

Why is reproducibility important?

Replication vs Reproduction

- Replication: other people collect new data
  - Scientific gold standard
  - Difficult and time-consuming

- Reproduction: other people analyze the same data
  - Does not by itself validate the analysis …
  - Has been proposed as a minimal standard

Case

37 y/o M informatician with PMH of email overload disorder

Request from informatics staff:

> "Please provide detailed data from your 2 year old analysis of total departmental effort spent performing test cancellations for a SBAR calling out the need to invest effort in duplicate checking rules for the ongoing EHR implementation project"

Analysis is in an Excel file but original raw data is nowhere to be found

Consider the above scenario, but with someone else performing the original analysis

Would it be less work to start from scratch and rewrite the analysis?

Point-and-Click Is Not Reproducible

- Interactive tools do not record user actions
- Manual documentation is error-prone
- Manual analyses cannot be repeated on new data sets or shared with collaborators

Computer code can precisely document each step of the analysis

Why YOU Should Do Data Analysis Reproducibly

"Can we redo the analysis with this month's data?"

"Why do the data in Table 1 not seem to agree with Figure 2?"

"Why did I decide to omit these six samples?"

YOUR CLOSEST COLLABORATOR IS **YOU** FROM 6 MONTHS AGO
(BUT YOU DON'T ANSWER E-MAILS)

Using R for Reproducibility

Programming in R (or another language) allows one to reproduce analysis steps exactly or perform same analysis on new data

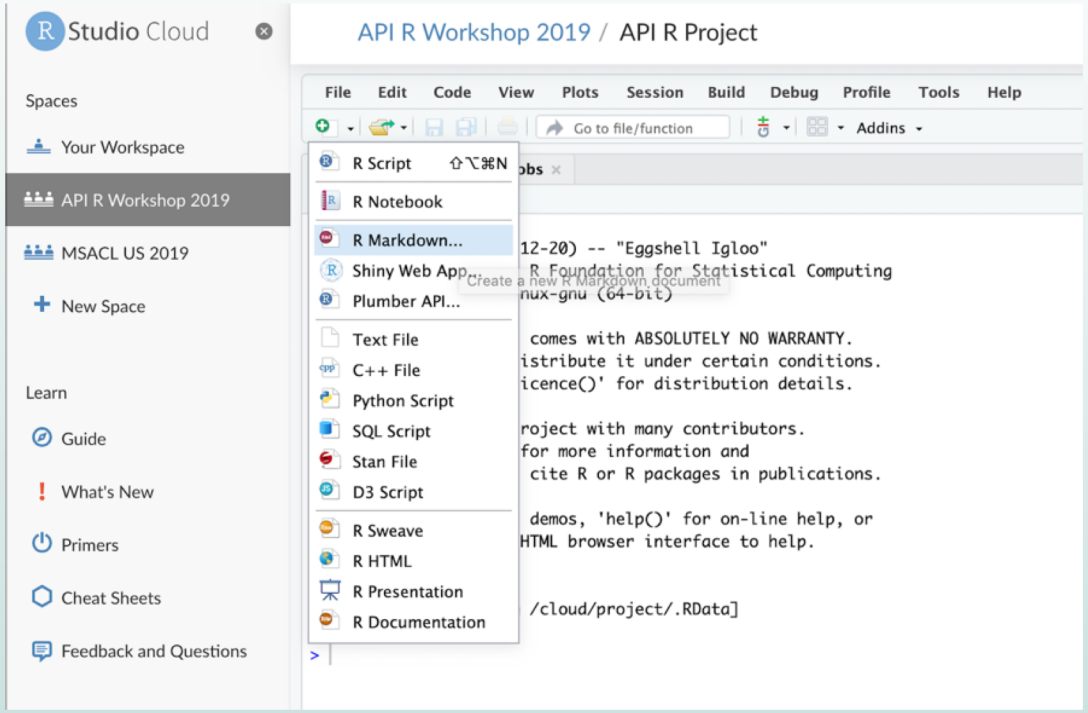
Better practice is to create documentation about analysis to accompany and explain code

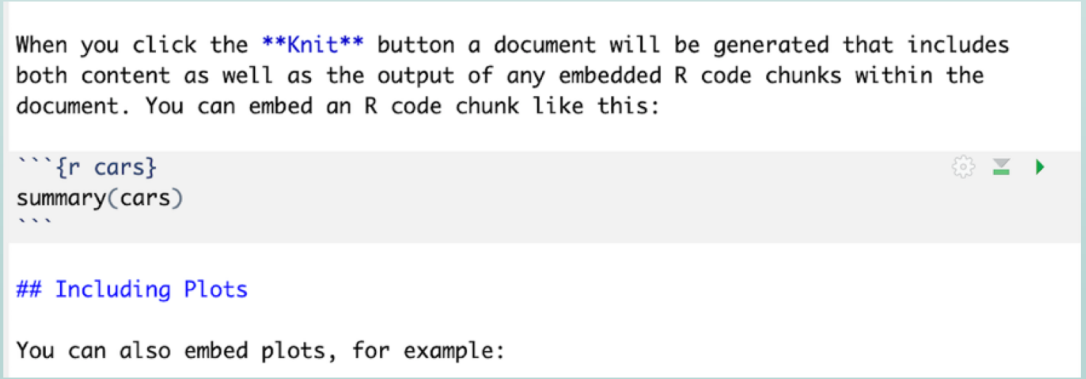
Best practice is include documentation and code in one place

Using R Markdown to Support Reproducibility

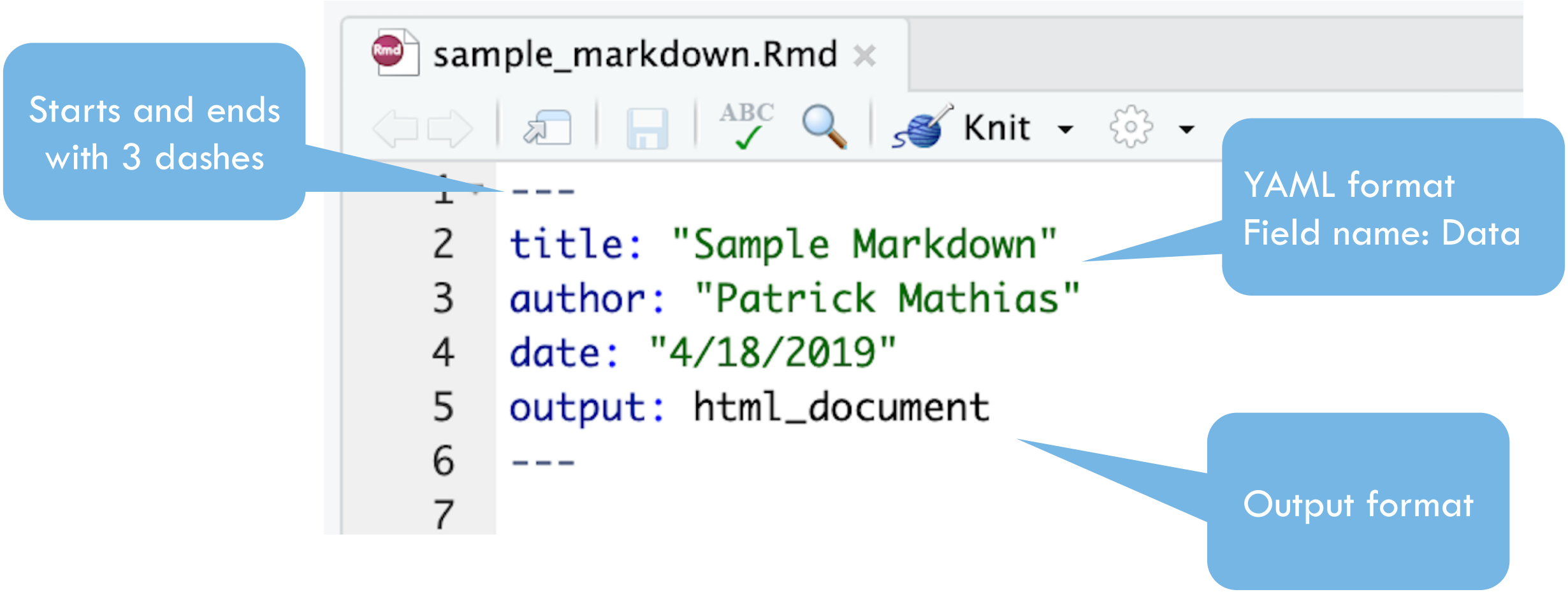
Your Turn #1

1. Open a new R Markdown document within RStudio Cloud
2. Enter a Title and Author and leave the output format as HTML
3. Click each of the buttons in the top right corner of the grey areas and observe what happens
4. “Knit” the document by clicking the Knit button and save the file



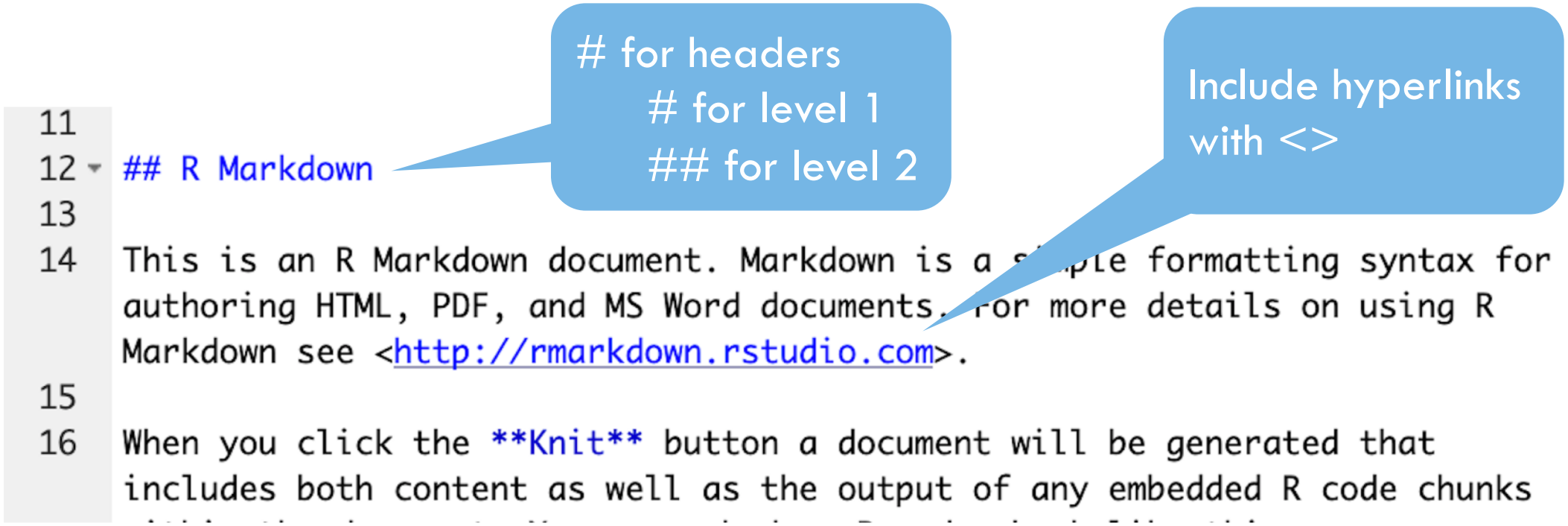


Header


sample_markdown.Rmd
Knit
---
title: "Sample Markdown"
author: "Patrick Mathias"
date: "4/18/2019"
output: html_document

Starts and ends with 3 dashes
YAML format
Field name: Data
Output format

Text



1 asterisk for *italics* (*italics*)

2 asterisks for **bold** (**bold**)

Hyphens (-bullet 1) for bullet points

R Markdown cheatsheet

syntax

```
Plain text
End a line with two spaces to start a new paragraph.
*italics* and _italics_
**bold** and __bold__
superscript^2^
~~strikethrough~~
[link](www.rstudio.com)

# Header 1

## Header 2

### Header 3

#### Header 4

##### Header 5

###### Header 6

endash: --
emdash: ---
ellipsis: ...
inline equation: $A = \pi*r^{2}$
image: ![](path/to/smallorb.png)

horizontal rule (or slide break):

***
```

becomes

Plain text
End a line with two spaces to start a new paragraph.
italics and *italics*
bold and **bold**
superscript^2
~~strikethrough~~
link

Header 1
Header 2
Header 3
Header 4
Header 5
Header 6

endash: –
emdash: —
ellipsis: …
inline equation: $A = \pi * r^2$
image: 

horizontal rule (or slide break):

```
> block quote

* unordered list
* item 2
    + sub-item 1
    + sub-item 2

1. ordered list
2. item 2
    + sub-item 1
    + sub-item 2

Table Header  | Second Header
------------- | -------------
Table Cell    | Cell 2
Cell 3        | Cell 4
```

> block quote

- unordered list
- item 2
  - sub-item 1
  - sub-item 2

1. ordered list
2. item 2
   - sub-item 1
   - sub-item 2

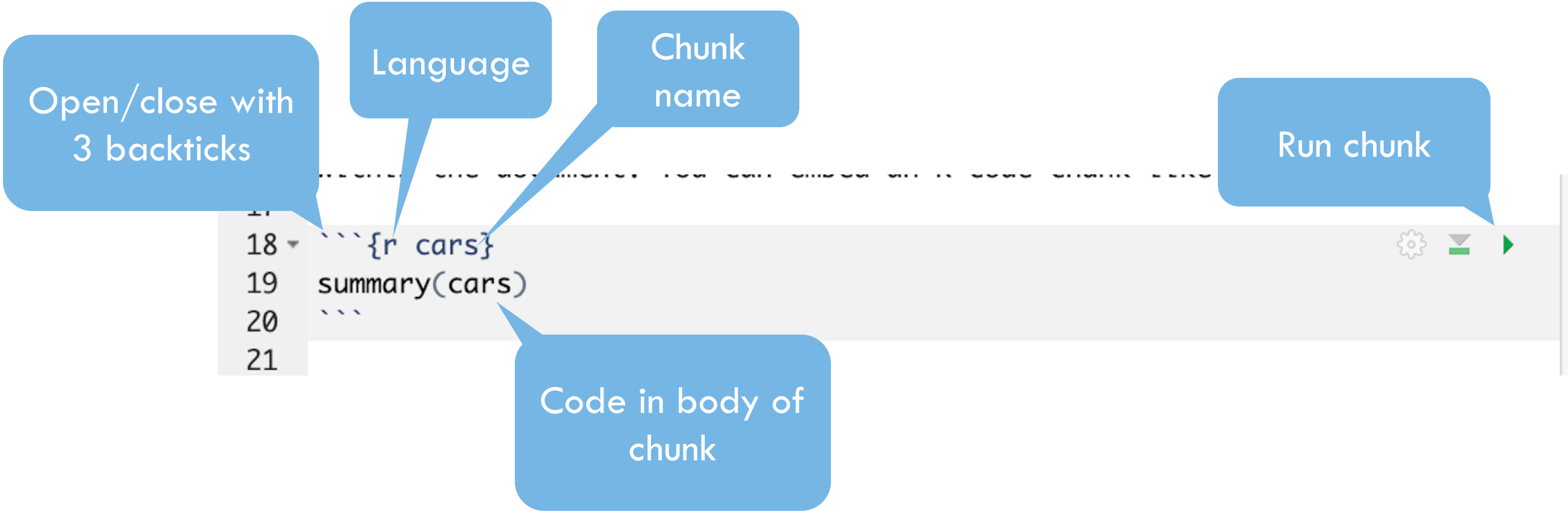
Table Header	Second Header
Table Cell	Cell 2
Cell 3	Cell 4

https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

Your Turn #2

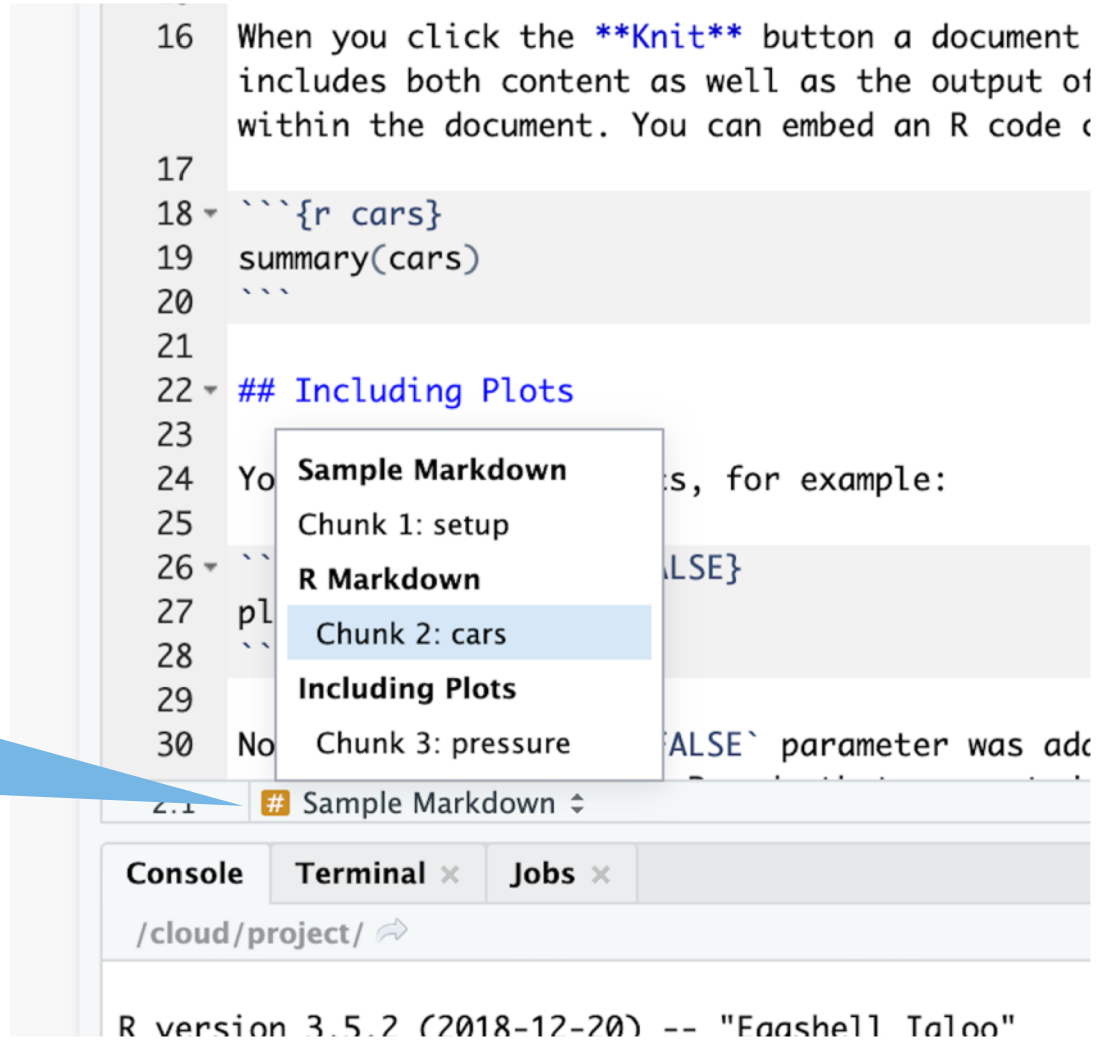
1. Within your R Markdown document, add another level 2 heading at the end called “Document Info”
2. Under your new heading add a level 3 heading called “Author Info”
3. Add your name in bold under the new heading
4. Add your job title in italics
5. Knit the document

Code chunks


Open/close with 3 backticks
Language
Chunk name
Run chunk
18 ```{r cars}
19 summary(cars)
20 ```
21
Code in body of chunk

Why name Code Chunks?

Jump between chunks



“Setup” Chunk

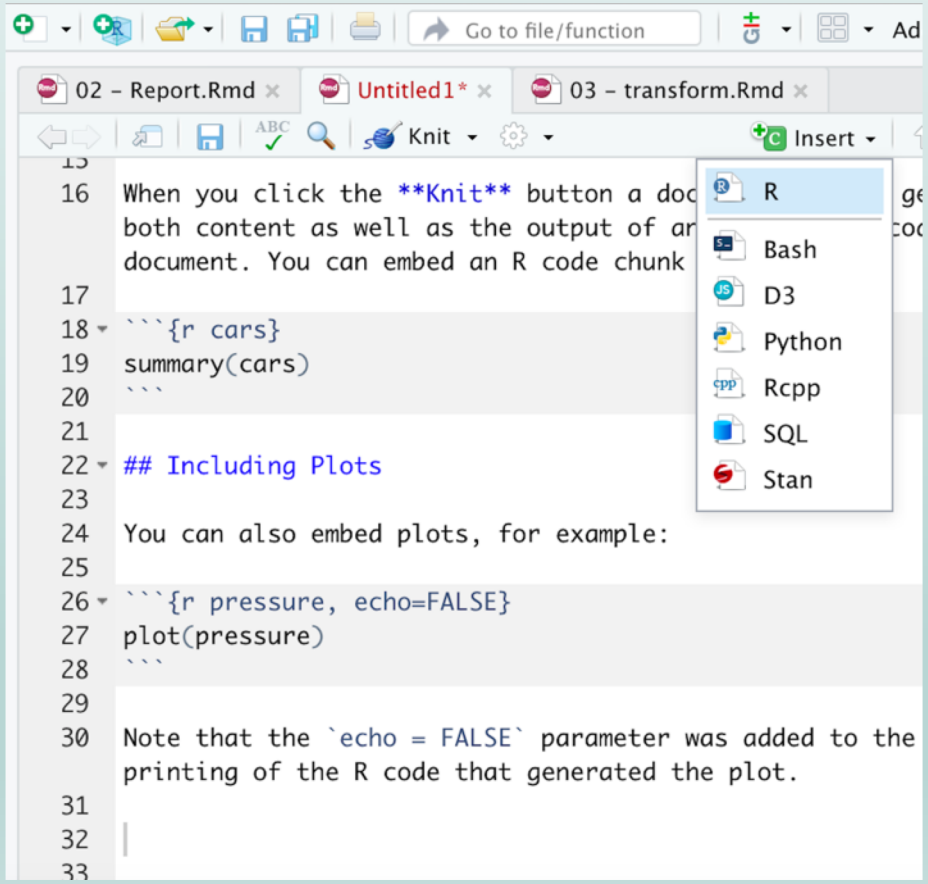
chunk name
(optional)

“chunk option”
don’t show code in rendered document

````
```{r setup, include=FALSE}
library(tidyverse)
library(lubridate)
```
````

for dealing
with dates

Your Turn #3



1. Insert a code chunk into white space within your open R Markdown document using Insert button on top right of code window

2. Add the following to your new code chunk:

    mean(c(10, 20, 30))

3. Execute code chunk by pressing Run button on top right of code chunk

4. Include the following within any text area (white space) and knit:

    `r mean(c(10, 20, 30))`

Keyboard Shortcuts

Insert a code chunk into white space within your open R Markdown document using:

- Windows: CTRL+ALT+i
- Mac: COMMAND+OPTION+I

Execute using shortcuts:

- Windows: CTRL+SHIFT+ENTER
- Mac: COMMMAND+SHIFT+ENTER

Importing Files

Working with R Markdown for this course

Each lesson has an R Markdown file

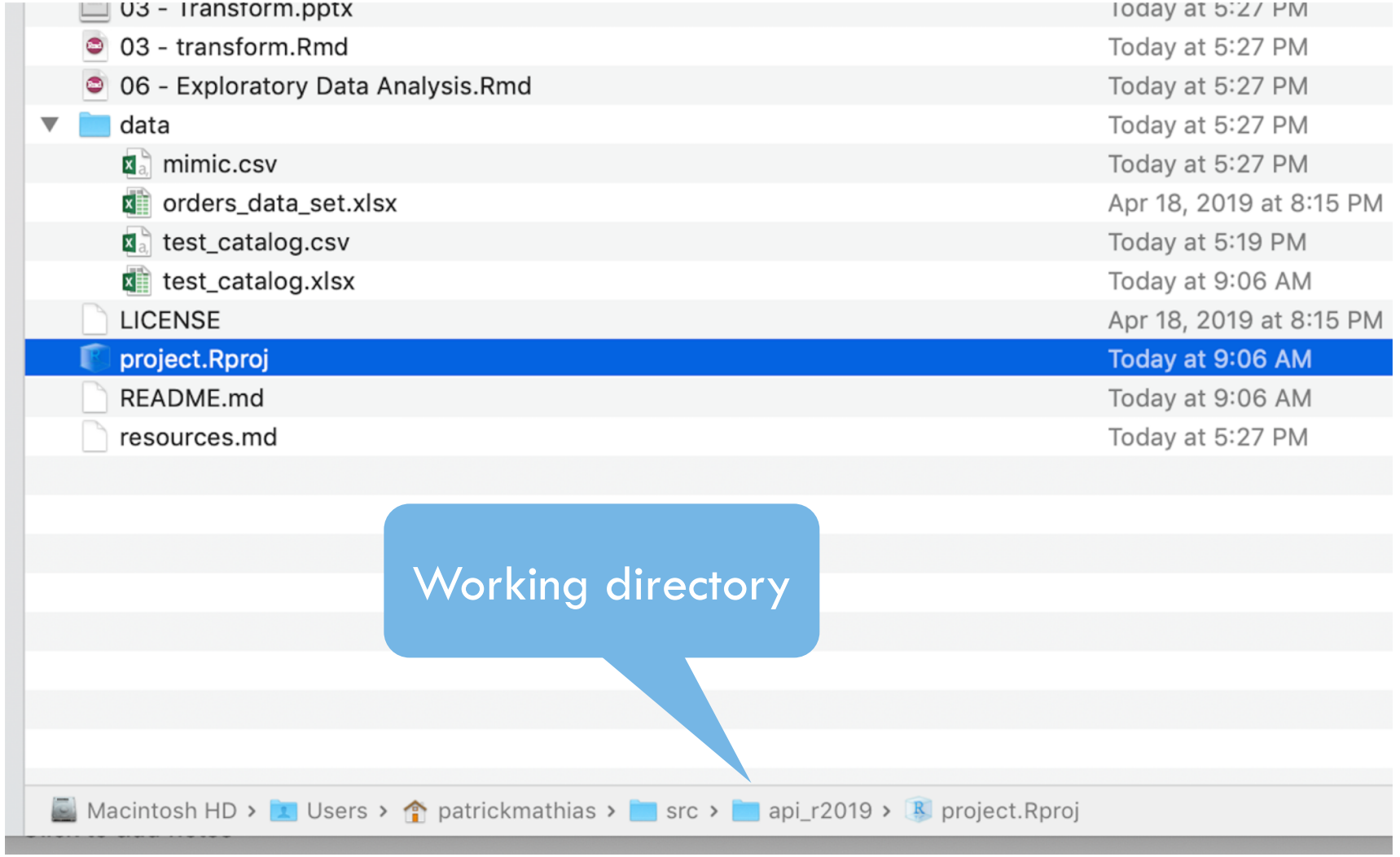
- Executable examples
- Exercises

Files used as “notebooks”: can document, execute, and iterate

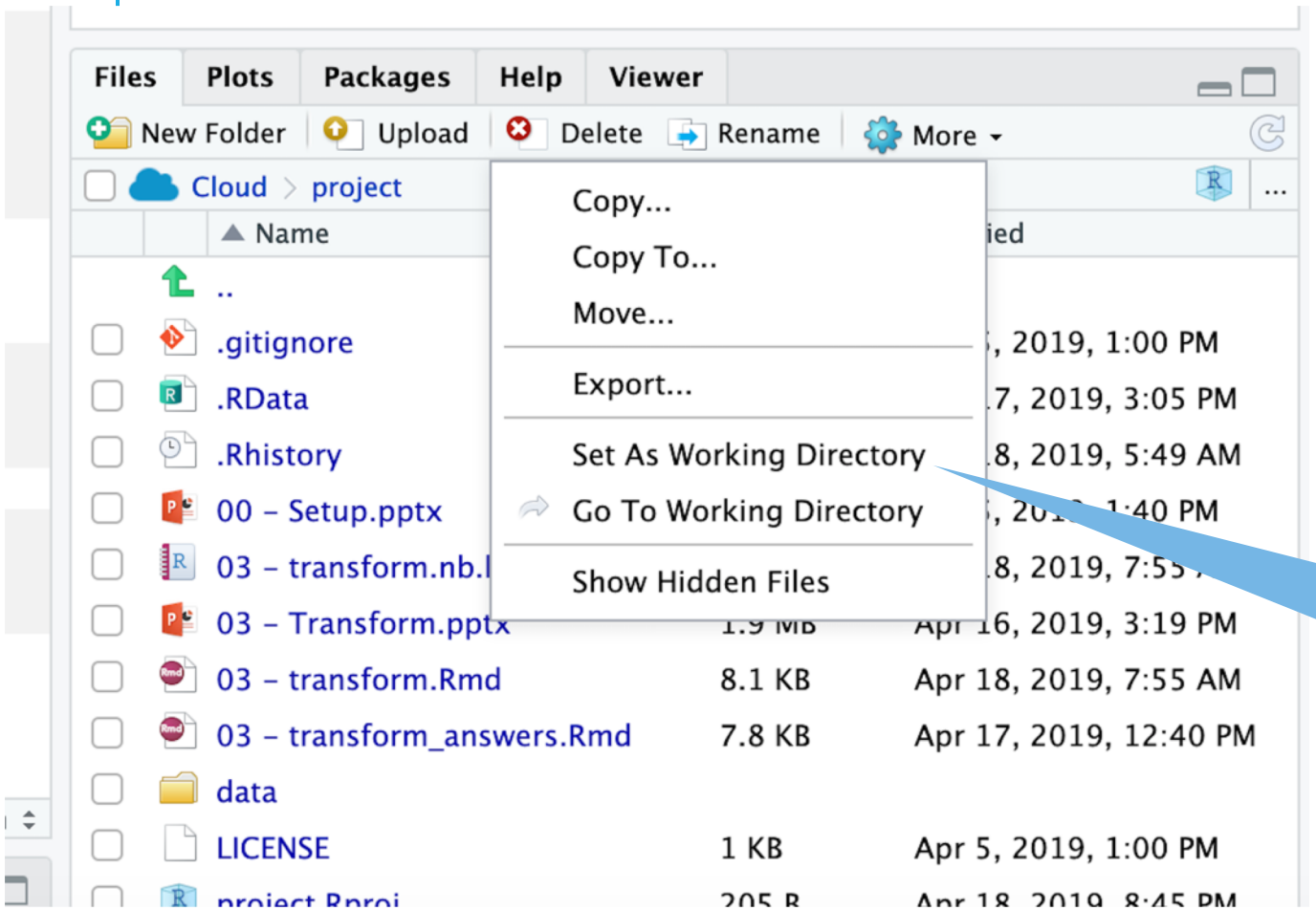
Best practice: work from notebook rather than console

Open “02 – Report.Rmd” and run the setup chunk.

Working with Your file system



Where am I? Your working directory



Navigate folder and file structure on your computer from Rstudio

getwd() function will tell you which folder you're in

ew working

Navigate folders and set working directory

Reading comma separated or tab delimited files


```
read_csv("data/test_catalog.csv")
```

File type (csv, tsv, delim for non-standard delimiters)

File path (if file in a folder within working directory)

File name

Reading Excel files

File name

```
read_excel("data/excel_file.xlsx",
           sheet = 1)
```

Specify sheet by number or "name"

Can also extract arbitrary rows and columns using "range = " argument

Your Turn #4

order_id	patient_id	description	proc_code	order_class_c_descr	lab_status_c	lab_status_c_descr	order_status_c	order_status_c_descr
19766	511388	PROTHROMBIN TIME	PRO	Normal			4	Canceled
88444	511388	BASIC METABOLIC PANEL	BMP	Normal			4	Canceled
40477	508061	THYROID STIMULATING HORMONE	TSH	Normal	3	Final result	5	Completed
97641	508061	T4, FREE	T4FR	Normal	3	Final result	5	Completed
99868	505646	COMPREHENSIVE METABOLIC PANEL	COMP	Normal	3	Final result	5	Completed

1. Find and import the orders data set and store in an object called "orders"
2. View the data
3. Run the summary function on the data

View a quick summary

```
order_status_c_descr reason_for_canc_c reason_for_canc_c_descr
Length:45002         Min.   :   1.0    Length:45002
Class :character     1st Qu.:  11.0    Class :character
Mode  :character     Median :  11.0    Mode  :character
                     Mean   : 437.2
                     3rd Qu.:1178.0
                     Max.   :1178.0
                     NA's   :37794

  order_time                    result_time
Min.   :2017-08-13 11:59:00   Min.   :2017-06-15 00:00:00
1st Qu.:2017-09-05 11:16:00   1st Qu.:2017-09-07 12:51:00
Median :2017-09-27 08:48:00   Median :2017-09-29 14:06:30
Mean   :2017-09-27 09:39:30   Mean   :2017-09-30 17:11:17
3rd Qu.:2017-10-19 13:45:00   3rd Qu.:2017-10-23 18:39:30
Max.   :2017-11-11 19:49:00   Max.   :2017-12-29 07:37:00
                              NA's   :7152
```

summary( ) function outputs quick statistical summaries for numerical and timestamp fields

Provides limited data (only a count) for character fields

Will provide counts of different categories for factor fields (categorical)

What is NA?

"Not available"

- Equivalent to NULL in some other languages

Represents missing values

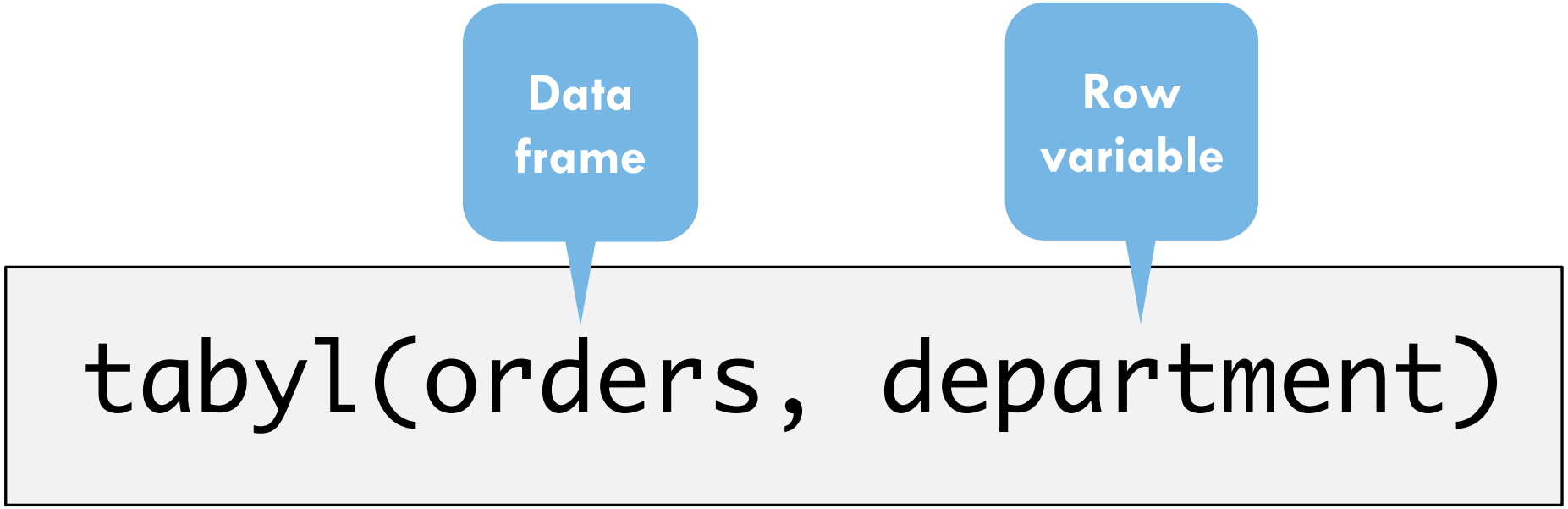
Tabulating Data

Orders Data Set

- Data set of outpatient laboratory orders
- 45,000 rows x 17 columns
- Queried from Epic orders data (Clarity)
- Deidentified and time shifted
- Includes timestamps for lab order, result, and provider review

Variable	Description
order_id	Key for order
patient_id	Key for patient
description	Text description of lab test
proc_code	Procedure code for lab test
order_class_c_descr	Setting test is intended to be performed in (eg. Normal = regular blood draw)
lab_status_c	Code for status of laboratory result
lab_status_c_descr	Status of laboratory result
order_status_c	Code for status of order
order_status_c_descr	Status of order
reason_for_canc_c	Code for cancellation reason (if applicable)
reason_for_canc_c_descr	Cancellation reason (if applicable)
order_time	Timestamp for time of original test order
result_time	Timestamp for most recent result in the record
review_time	Timestamp for provider acknowledgment of review of result
department	Clinic associated with test order
ordering_route	Structure/menu in health record from which order was placed
pref_list_type	Category of preference list (if applicable)

Create a one variable table



Two variable tables can help answer simple questions quickly

Which clinic cancelled the highest number of lab orders?

Data frame

Row variable

```
tabyl(orders, department,
  order_status_c_descr)
```

Column variable

Your Turn #5

Which 3 departments ordered the highest number of labs using “Provider Preference Lists”?

What happens if you add another variable to the tabyl() function (3 variables)?

Prettifying Table Output

```
kable(tabyl(orders, department, pref_list_type))
```

department	Clinic Preference List	None	Provider Preference List
BEHAVIORAL HEALTH CLINIC	40	463	0
CARDIOLOGY CLINIC	888	33	105
ENDOCRINOLOGY CLINIC	613	94	779
FAMILY MEDICINE CLINIC	3135	365	1
GASTROENTEROLOGY CLINIC	618	89	74

Basic Plotting

To make **any** kind of graph:

1. Choose a "tidy" **data** frame

```
ggplot(data = data_frame) +
  geom_function(mapping = aes(mappings))
```

2. Pick a "**geom**" function

3. Write aesthetic **mappings**

Creating a Histogram – Orders Over Time

initialize a plot with `ggplot()`

data frame

+ sign (at end of line)

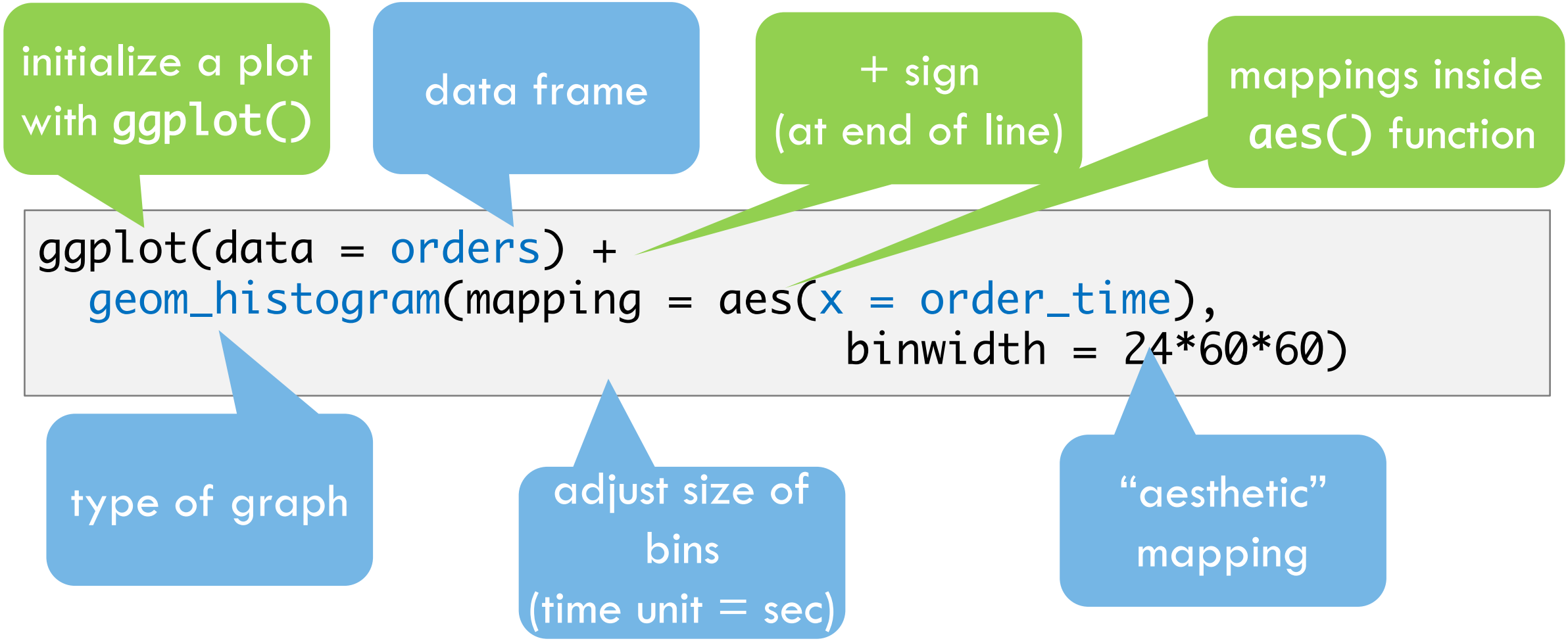
```
ggplot(data = orders) +
  geom_histogram(mapping = aes(x = order_time))
```

type of graph

mappings inside `aes()` function

"aesthetic" mapping

Creating a Histogram – controlling bins



Creating a Histogram - Counts

initialize a plot with ggplot()

data frame

+ sign (at end of line)

mappings inside aes() function

```
ggplot(data = orders) +
  geom_histogram(mapping = aes(x = order_status_c_descr),
                               stat = "count")
```

type of graph

statistical transformation

"aesthetic" mapping

Your Turn #6

1. Plot the distribution of review times by day.
2. Plot the distribution of order reviews by week instead of by day.

Knitting Other Document Types

Multiple Output Formats are Available

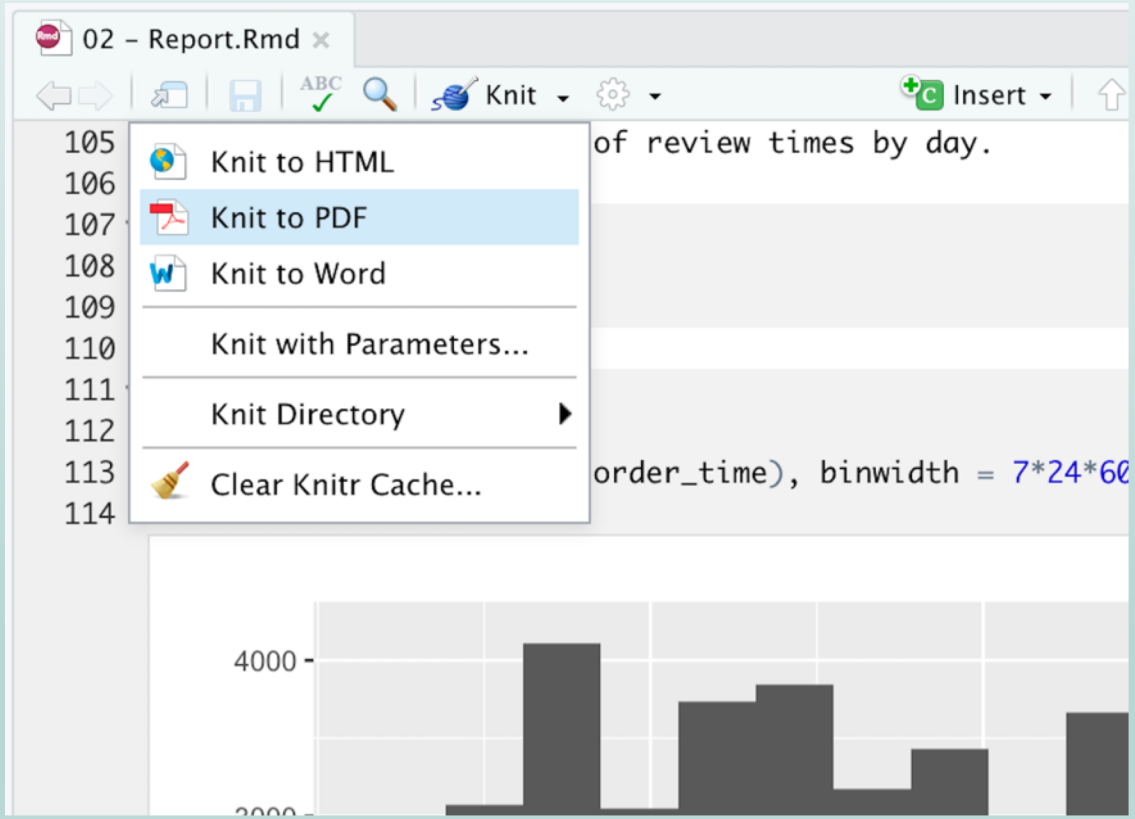
Pandoc universal document converter can create multiple document types:

- html
- pdf
- Docx

Can also create presentations and dashboards

- Including Powerpoint (most recent version of RStudio)

Your Turn #7



Create a pdf of your lesson 2 markdown document using "Knit to pdf"

01:00

Your Turn #7

Create a new R Markdown document but select “Presentation” instead of “Document” from menu

Select one of the HTML options and knit the document