BINF 5003: Data Mining, Modeling, and Biostatistics

Week 3

Module 2 – Data Wrangling

Overview

- Writing functions

- Structure of dataframes and indexing for subseting data

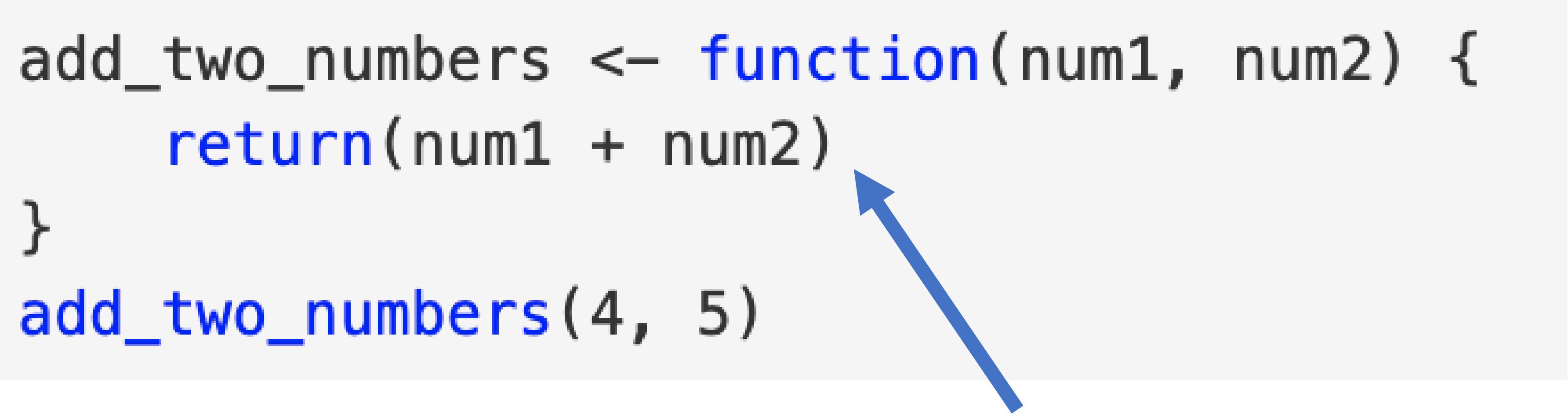
- Two methods of completing the same task (of subseting)
  - Base R functions
  - Tidyverse functions

- Tidyverse-specific syntax

Writing custom functions

Function name

Defined function arguments

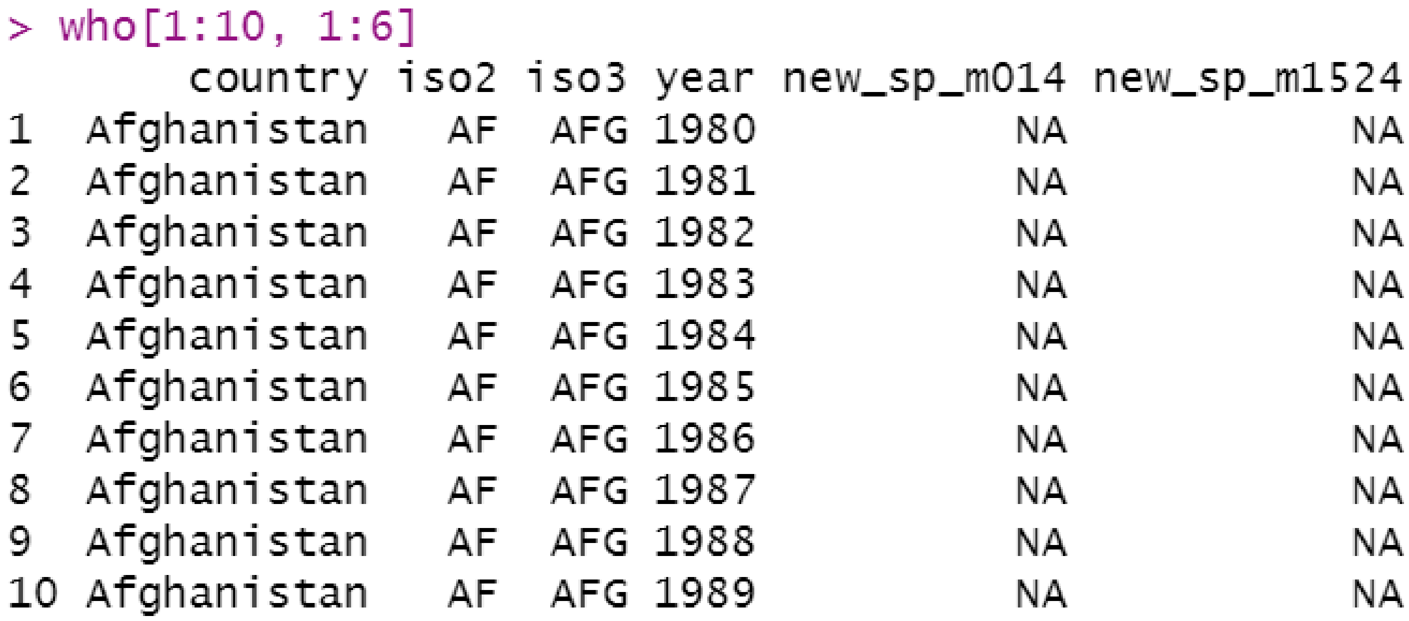


```
add_two_numbers <- function(num1, num2) {
    return(num1 + num2)
}
add_two_numbers(4, 5)
```

Commands that uses defined arguments

Indexing with Base R

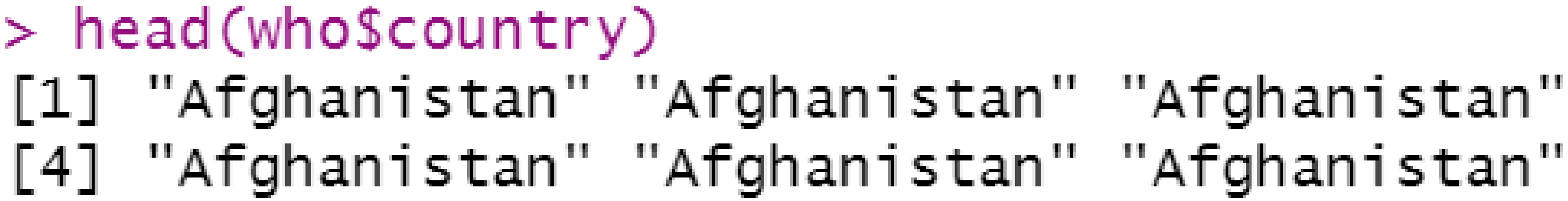
- Square brackets for indexing
- Data frames require two numeric values separated by a comma
  - First position: rows
  - Second position: columns



```
> who[1:10, 1:6]
       country iso2 iso3 year new_sp_m014 new_sp_m1524
1  Afghanistan   AF  AFG 1980          NA           NA
2  Afghanistan   AF  AFG 1981          NA           NA
3  Afghanistan   AF  AFG 1982          NA           NA
4  Afghanistan   AF  AFG 1983          NA           NA
5  Afghanistan   AF  AFG 1984          NA           NA
6  Afghanistan   AF  AFG 1985          NA           NA
7  Afghanistan   AF  AFG 1986          NA           NA
8  Afghanistan   AF  AFG 1987          NA           NA
9  Afghanistan   AF  AFG 1988          NA           NA
10 Afghanistan   AF  AFG 1989          NA           NA
```

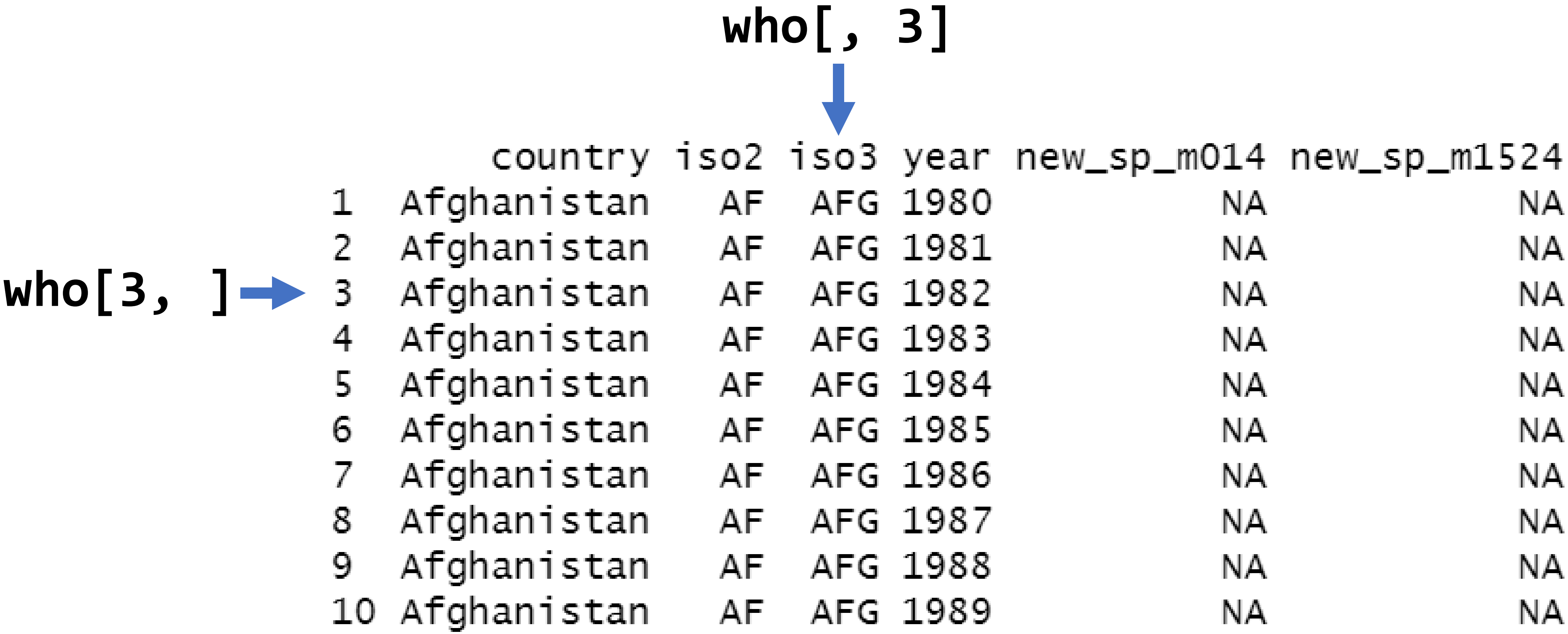
Indexing with Base R

- $ string operator can be used to select columns



```
> head(who$country)
[1] "Afghanistan" "Afghanistan" "Afghanistan"
[4] "Afghanistan" "Afghanistan" "Afghanistan"
```

Indexing with Base R



`who[, 3]`

`who[3, ]`

```
      country iso2 iso3 year new_sp_m014 new_sp_m1524
1  Afghanistan  AF  AFG 1980          NA           NA
2  Afghanistan  AF  AFG 1981          NA           NA
3  Afghanistan  AF  AFG 1982          NA           NA
4  Afghanistan  AF  AFG 1983          NA           NA
5  Afghanistan  AF  AFG 1984          NA           NA
6  Afghanistan  AF  AFG 1985          NA           NA
7  Afghanistan  AF  AFG 1986          NA           NA
8  Afghanistan  AF  AFG 1987          NA           NA
9  Afghanistan  AF  AFG 1988          NA           NA
10 Afghanistan  AF  AFG 1989          NA           NA
```

Strategies

Base R

- Isolated functions
- Column names are strings, require quotes
- Use more brackets
- Computationally (memory) lighter

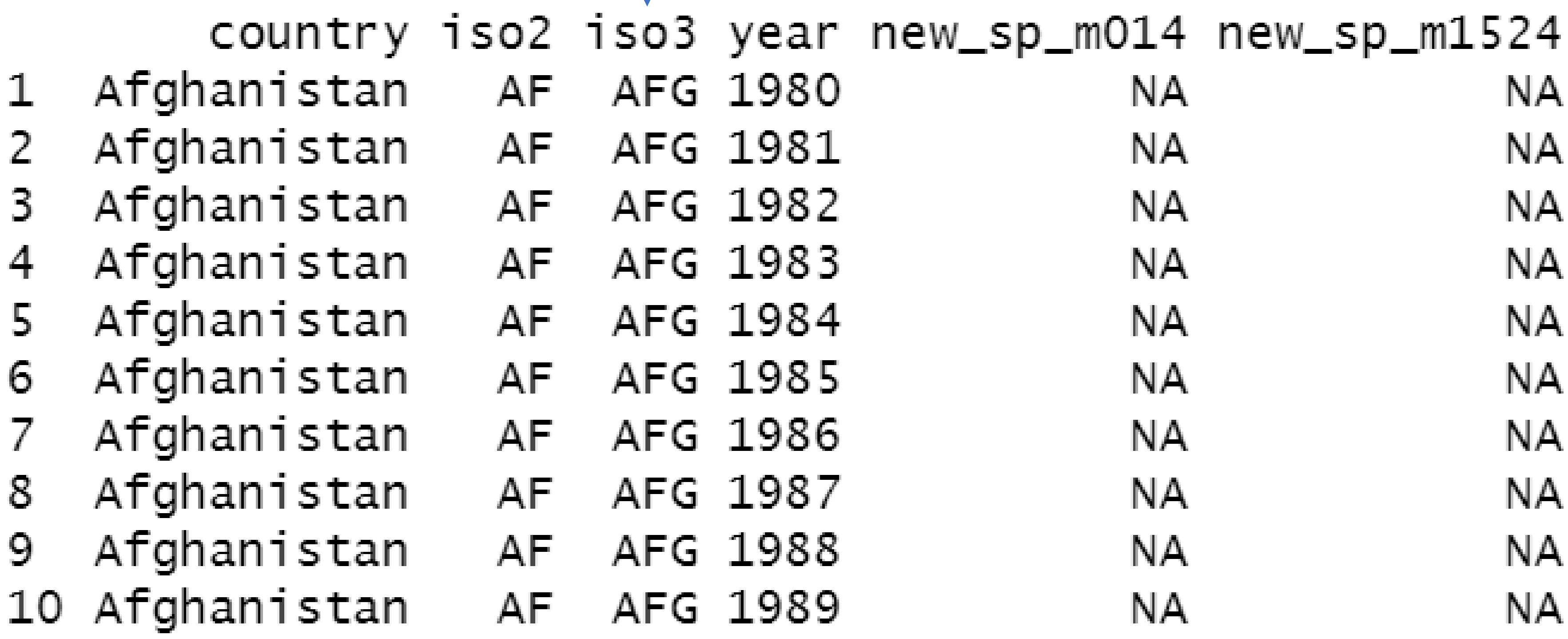
Tidyverse

- Connect multiple functions with \`+\` (ggplot) or \`%>%\` (pipe)
- Functions recognize column names, no quotes required
- More “readable”

Indexing with tidyverse

select

filter

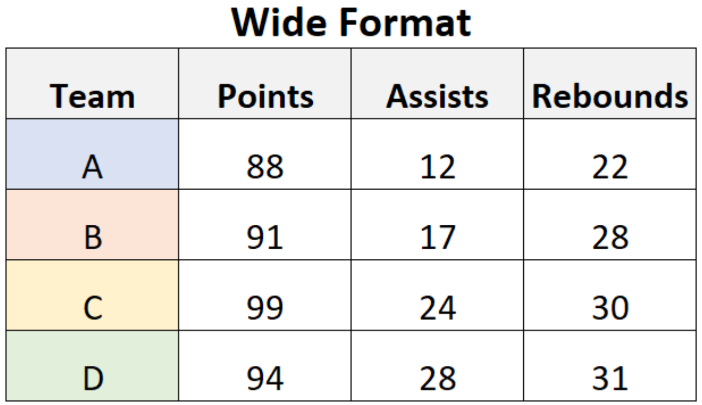


```
       country iso2 iso3 year new_sp_m014 new_sp_m1524
1  Afghanistan   AF  AFG 1980          NA           NA
2  Afghanistan   AF  AFG 1981          NA           NA
3  Afghanistan   AF  AFG 1982          NA           NA
4  Afghanistan   AF  AFG 1983          NA           NA
5  Afghanistan   AF  AFG 1984          NA           NA
6  Afghanistan   AF  AFG 1985          NA           NA
7  Afghanistan   AF  AFG 1986          NA           NA
8  Afghanistan   AF  AFG 1987          NA           NA
9  Afghanistan   AF  AFG 1988          NA           NA
10 Afghanistan   AF  AFG 1989          NA           NA
```

Tidyverse has different syntax

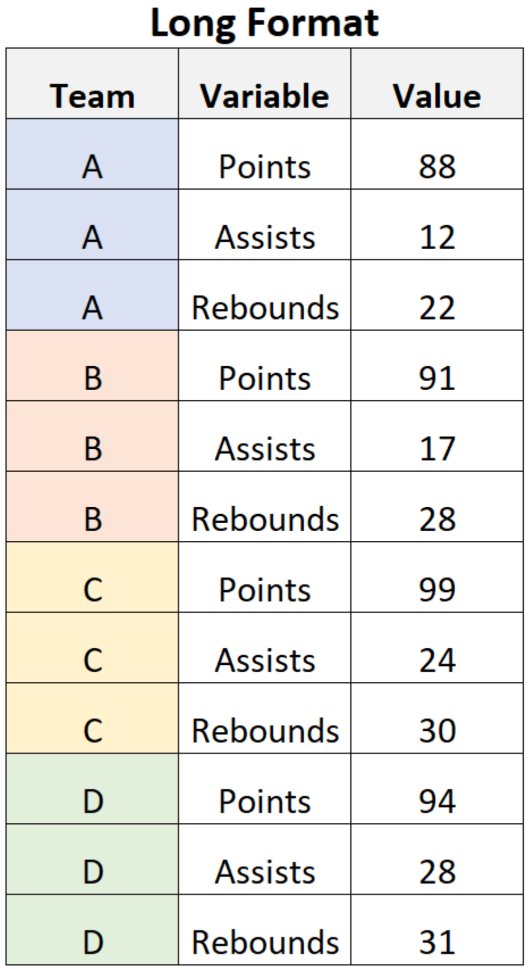
- *Select* for columns and *filter* for rows that match a criteria
- Column names in the arguments do not require (and should not use) quotations that commonly accompany strings in base R

Re-shaping data



Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31



Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

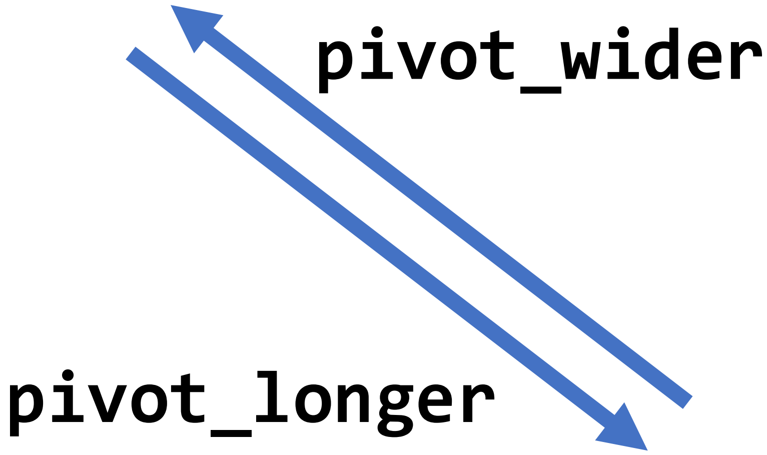
- Both tables are communicating the same data
- Which one do you prefer?

Re-shaping data

- One observation per row
    - All parts of the same observation in the same row
- Remember, we are trying to minimize repetition when cleaning up data for analysis

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31



pivot_wider

pivot_longer

Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

Why reshape data?

- Make multiple datasets more consistent
  - Match other datasets from the same study
  - Match an industry standard for storing or communicating data
  - Easier to communicate with others

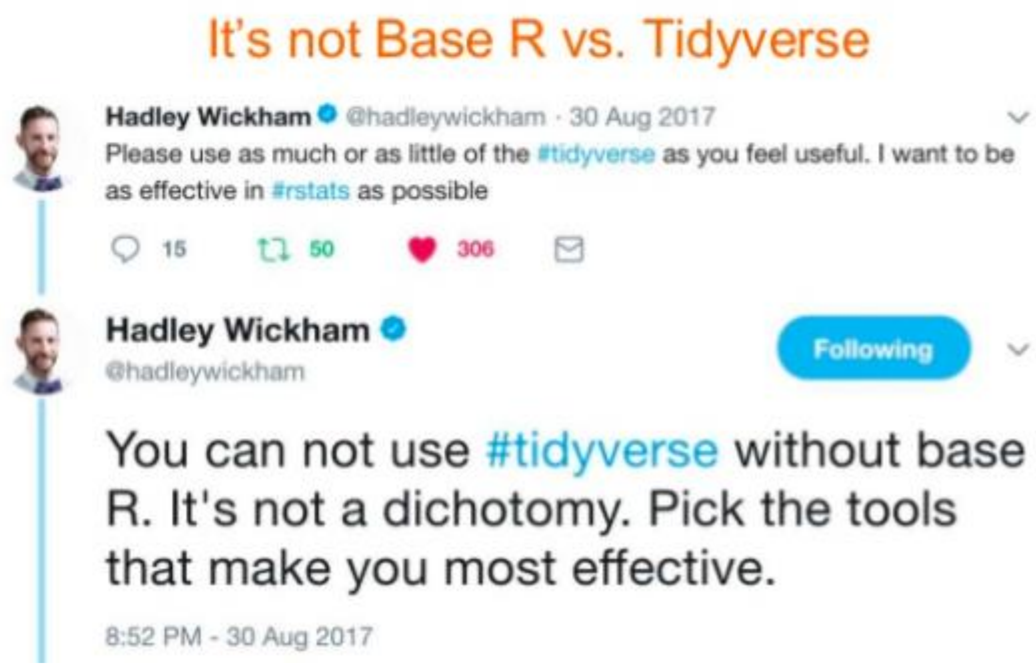
- Prepare data for a downstream application

Structure of tidy data (for computers!)

1. Variables make up the columns
2. Observations make up the rows
3. Values go into cells
4. Reduce redundancy

Which method should you pick?

- Both are valid strategies! They do not exist in isolation
- Each have their own pros and cons
- Depends what the next steps are going to be



It's not Base R vs. Tidyverse

Hadley Wickham @hadleywickham · 30 Aug 2017

Please use as much or as little of the #tidyverse as you feel useful. I want to be as effective in #rstats as possible

Hadley Wickham
@hadleywickham

You can not use #tidyverse without base R. It's not a dichotomy. Pick the tools that make you most effective.

8:52 PM - 30 Aug 2017

Tips for data wrangling

- Spend time understanding the data before you get started

- Understand what format you need the data at the end of the process before you begin

- Consider what tools you have, rather than coding solutions from scratch

Connecting tidyverse functions with a pipe

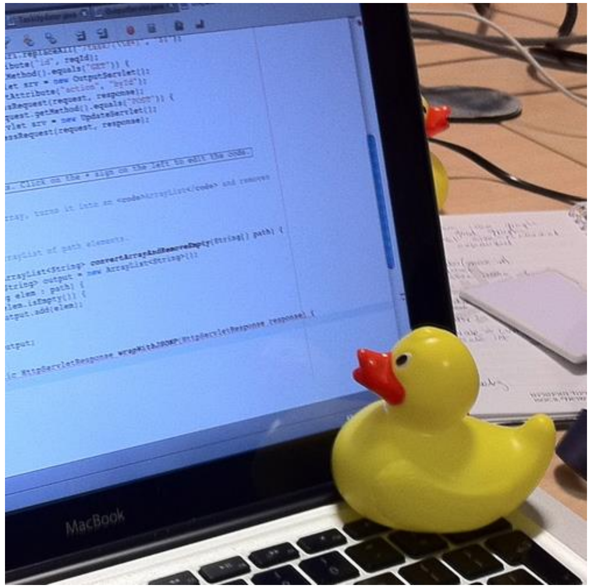
- %>%

- Concept adapted from terminal scripting

- Tidyverse’s alternative to base R nesting multiple codes with brackets
    - Brackets require reading from inside to out, can be more difficult to edit because you need to find the complementary open and close brackets
    - Pipes lay out the steps from top to bottom, minimize the use of brackets or intermediate objects

Tips for writing more complex code

- Separate the tasks of
  - What code to write
  - How to write the code
- Use markdown plain text or hashtag comments to plan out your steps before you start writing code
- Rubber Ducky Debugging!
- Work with a partner to brainstorm how to solve the task, and have a partner watch you code to catch typos



Is there a relationship between head size and brain weight?

Need computational tools to analyze large datasets!

Head Size(cm^3)	Brain Weight(grams)
4512	1530
3738	1297
4261	1335
3777	1282
4177	1590
3585	1300
3785	1400
3559	1255
3613	1355
3982	1375
3443	1340
3993	1380

. .

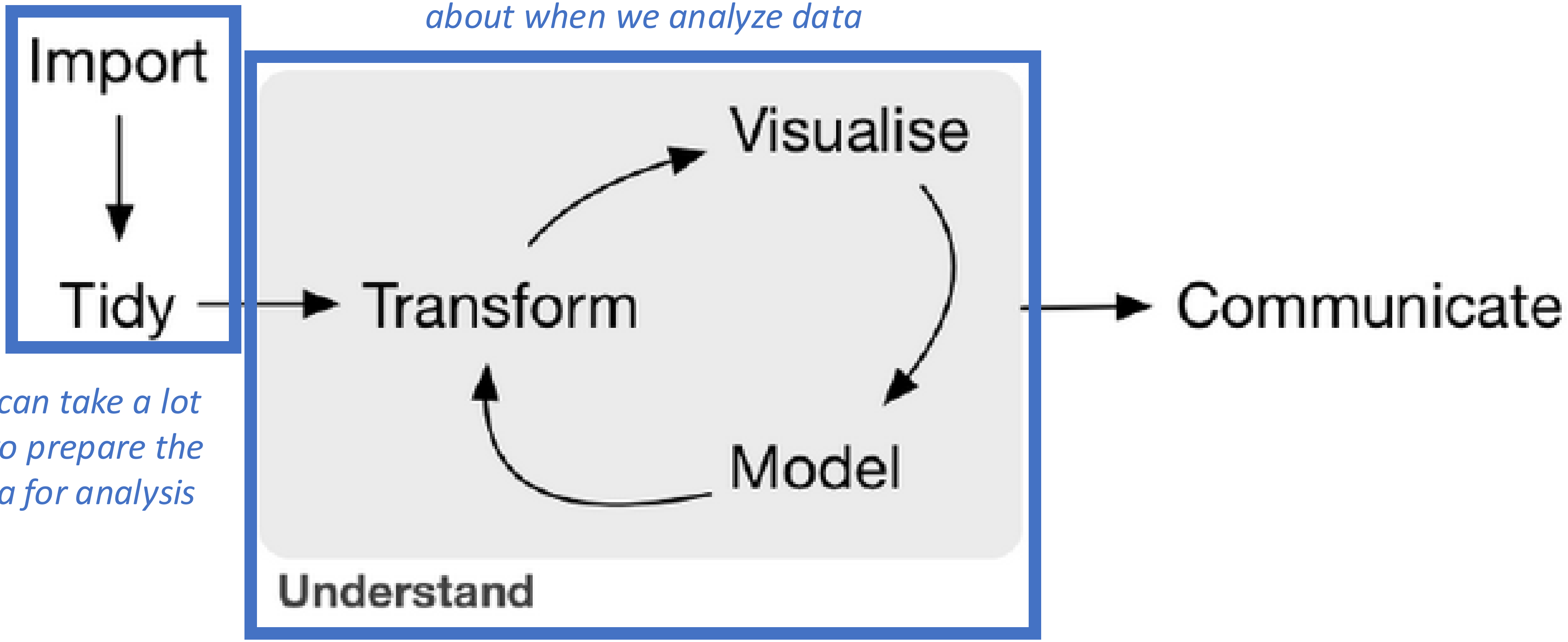
. .

. .

https://www.injurymap.com/free-human-anatomy-illustrations

Finding patterns in data is fun!

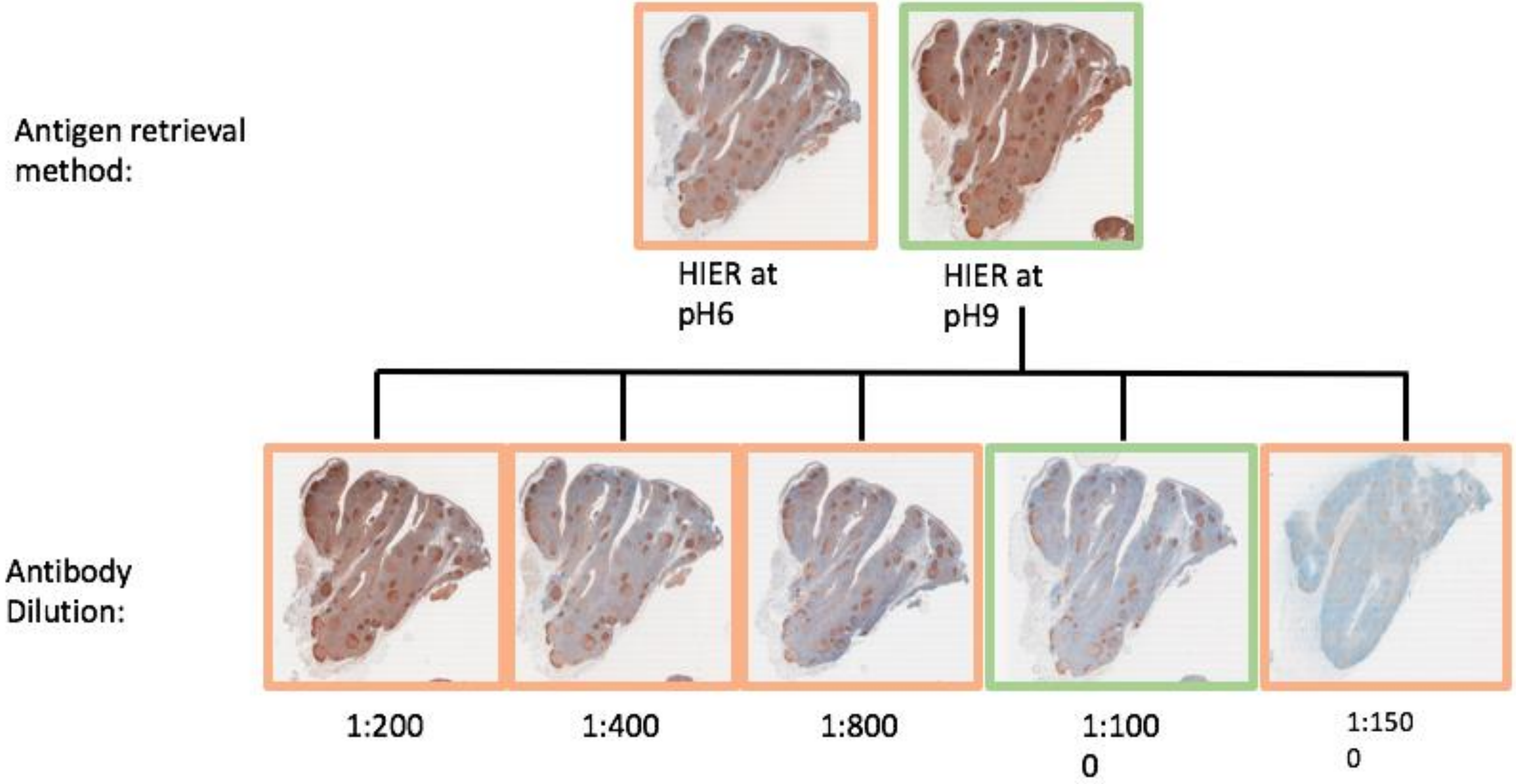
This is the part we normally think about when we analyze data



… but it can take a lot of time to prepare the raw data for analysis

"Playing the whole game": A data collection and analysis exercise with Google Calendar

Optimizing your workflow



Optimisation of chromogenic IHC for the Activation-Induced Cytidine Deaminase (AID) Antibody on Tonsil Tissue

https://www.cancer.ox.ac.uk/support/THL/Ab-list

Wrap-up

- You may need to write your own custom function in base R during data wrangling

- Data stored in dataframes can be accessed by columns or rows
  - Can use rules or pattern matching

- Redundancy is good! More options to achieve the same end result
  - Base R functions
  - Tidyverse functions

- Tidyverse mostly does not require quotations around column names (unlike base R) and can connect functions using the pipe (%>%)