

Welcome to Session 3!

1. Make sure you have the session 3 practice materials downloaded from the webpage
<https://tinyurl.com/rbootcamp2024>
2. Sign-in here: <https://tinyurl.com/bootcamp3-signin> or here
3. Open up `s3_starter_code_2024.R` in Rstudio and get started with the warm-up (we will start at 2:40!)










Session 3: Data processing

Emily & Sierra
6/27/2024

Today's agenda

-  Warm-up
-  What is tidy data
-  Demo - Introducing tidyverse and the pipe operator
-  Individual practice - Data organization and processing
-  Discussion

What is “tidy data”?

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”

—HADLEY WICKHAM

Organizing a dataset this way makes it easy to interpret

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

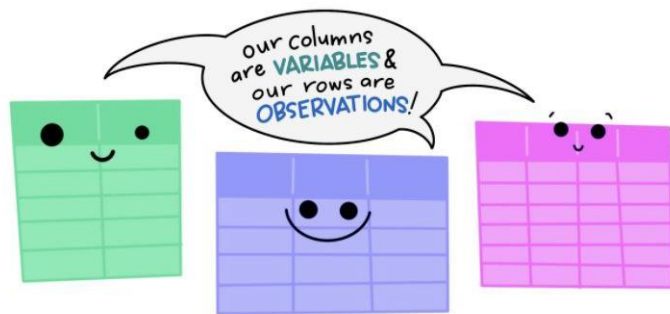
each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

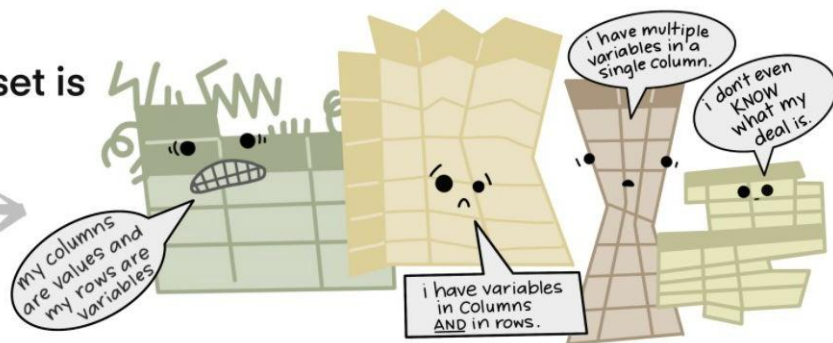
Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

The standard structure of tidy data means that
"tidy datasets are all alike..."



"...but every messy dataset is
messy in its own way."

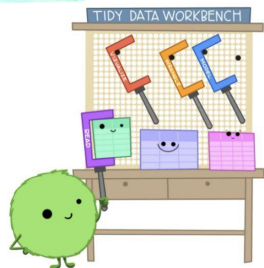
—HADLEY WICKHAM



Why do we want tidy data?

1. Reproducible code (fewer errors)!

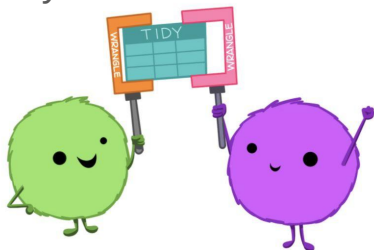
When working with tidy data, we can use the **same tools** in similar ways for different datasets...



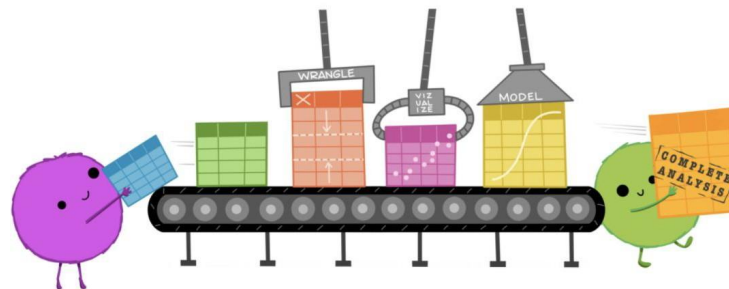
...but working with untidy data often means reinventing the wheel with **one-time** approaches that are **hard to iterate or reuse**.



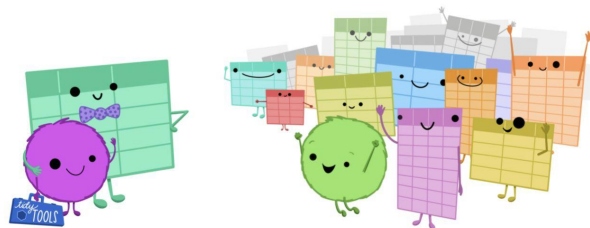
2. Easy collaboration



3. Automated pipelines (efficient and consistent!)



4. Data sharing (easy to interpret and combine with other data)



Is our data tidy?

Open `penguins.csv`

Check the basic structure

- Is every column a variable ?
- Is every row an observation ?
- Is every cell one value ?

An observation might mean something different for different data! Here each penguin is an observation.

We're in good shape but there is still more processing to do to get the data we want for our analyses.

Open `penguins_cleaned.csv`

What are some of the differences between these two dataframes?

1. Selected only a few of the variables

Changes to columns

2. Filtered observations by a specific year

Changes to rows

3. Removed subjects with missing values

4. Changed the values of cells in the sex column

Changes to cells

It's often useful to follow this hierarchy when *removing* data

There is an easy way to do all this in R!

Introducing our favorite library: Tidyverse!

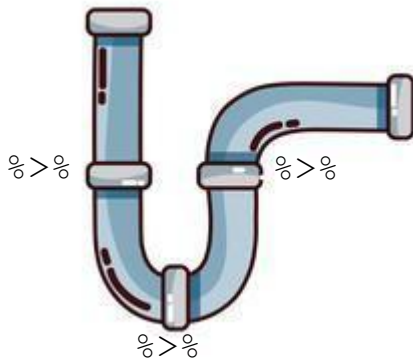
Blast off into the...



- A *library* is an organized collection of code and functions written by other members of the R community.
- `Tidyverse` is a library created specifically for organizing and processing your data
 - Includes *dplyr*, *ggplot*, etc.
- Install `tidyverse` and unlock a whole new world of functions and commands.

A new operator: Pipes %>%

- Once you have installed tidyverse you have access to a new symbol: %>%
- The pipe operator (%>%) allows you to string together many functions on the same data frame.
- This lets you make a workflow of tasks that you perform sequentially on a dataframe.



Let's remember the steps we want to perform on the penguins dataset

In R we can combine these steps using the %>% operator and save it all as a new dataframe.

```
penguins <- read.csv("penguins.csv")
```

1. Select only a few of the variables

2. Filter observations by a specific year

3. Remove NAs

4. Change the values of cells in the gender columns.

New df
penguins_final <- penguins %>% *then*

"Select certain columns" %>% *then*

"Filter by a specific year" %>% *then*

"Remove missing values" %>% *then*

"Change the cell values for the gender variable"

Note: this is called "*pseudo code*". We'll replace the highlighted sections with real tidyverse commands in R

Pipes help make your code:

- *Reproducible*
- *Readable*
- *Easy to automate*



Tidy data and happy collaborators



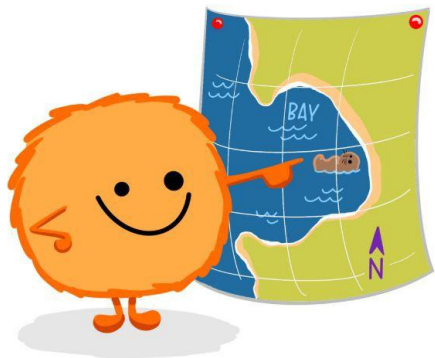
Now let's venture into the tidyverse...

dplyr::filter()

KEEP ROWS THAT
satisfy
your CONDITIONS

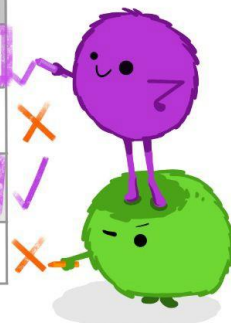
keep rows from... this data... ONLY IF... type is "otter" AND site is "bay"

```
filter(df, type == "otter" & site == "bay")
```



type	food	site
otter	urchin	bay
shark	seal	channel
otter	abalone	bay
otter	crab	wharf

@allison_horst





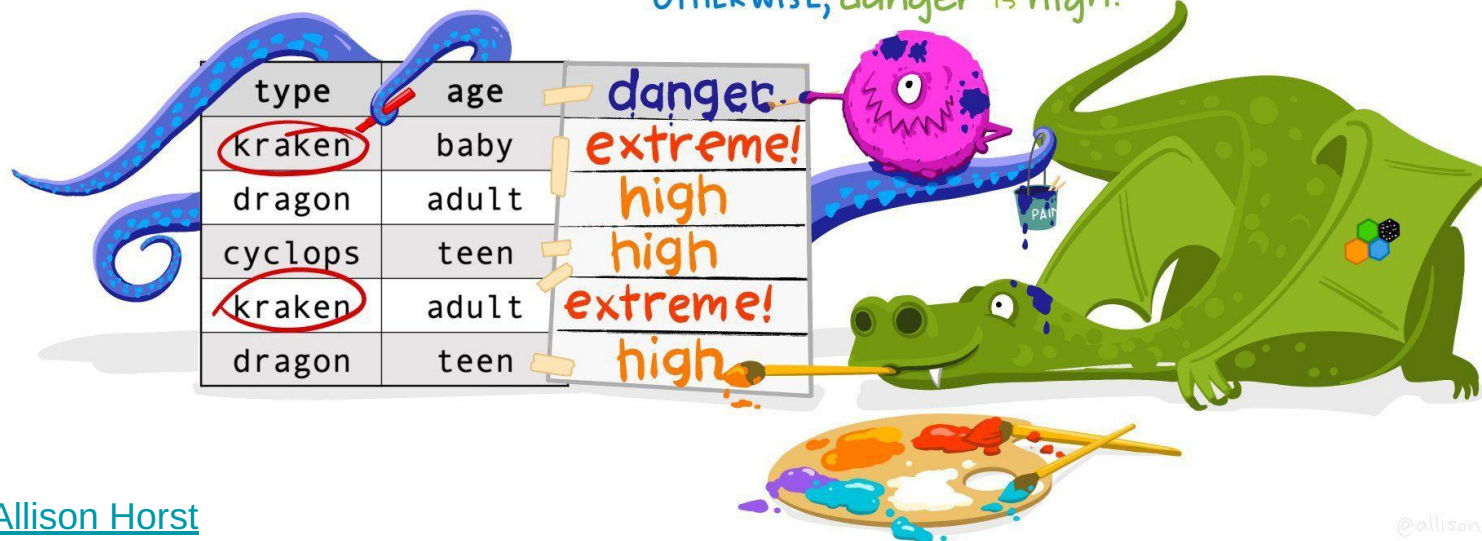
dplyr::case_when()

IF ELSE...
(but you love it?)

```
df %>% ADD COLUMN 'danger'  
  mutate(danger = case_when(type == "kraken" ~ "extreme!",  
                             T ~ "high"))
```

danger is
extreme!

OTHERWISE, danger is high.



Reminders!

- Session 4 is in *2 weeks* on July 11! (Not next Thursday)
- Office hours next week are being adjust to accommodate the July 4 holiday:
 - Sierra's office hours will only be held on Zoom on Tuesday (2 PM)
 - Emily's office hours are moving to...