Tour de tidyverse

R-Ladies Austin | Caitlin Hudon

What is the `tidyverse`?

+ Collection of R packages based on tidy data principles
+ Designed to work together
+ An easier way to code!
+ AKA “Hadleyverse” (most packages written by Hadley Wickham)

```
install.packages("tidyverse")
```

What is the tidy data?

1. Each variable is a column
2. Each observation is a row
3. Each type of observational unit is a table

id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
6	98^0	Give Me Just One Nig...	3:24
7	A*Teens	Dancing Queen	3:44
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

The tidyverse

Components

dplyr

Grammar of data manipulation

+ `mutate()` to create new variables from existing ones
+ `select()` picks variables based on their names
+ `filter()` allows pointed selection based on given criteria
+ `summarise()` reduces multiple values down to a single summary
+ `arrange()` changes the ordering of rows

magrittr

Simplifying R code with pipes (%>%)

+ Easy way to pass data through functions without nesting
+ First argument of each function is “piped” in to reduce redundancy

dplyr + magrittr :: example

before

```
summarise(
  group_by((filter
            (babynames, name == "Caitlin")
            ),
           year
           ),
  total = sum(n))
```

after

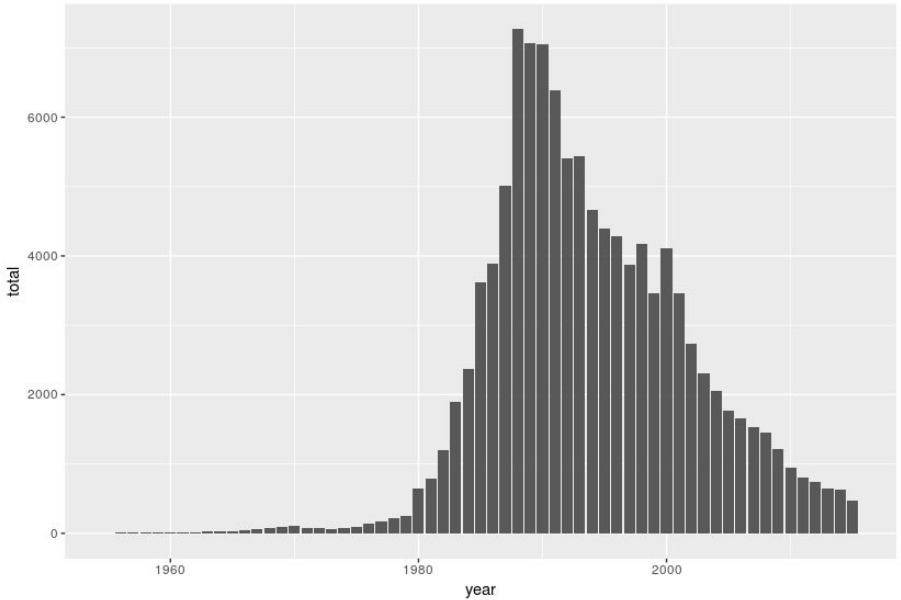
```
babynames %>%
  filter(name == "Caitlin") %>%
  group_by(year) %>%
  summarise(total = sum(n))
```

ggplot2

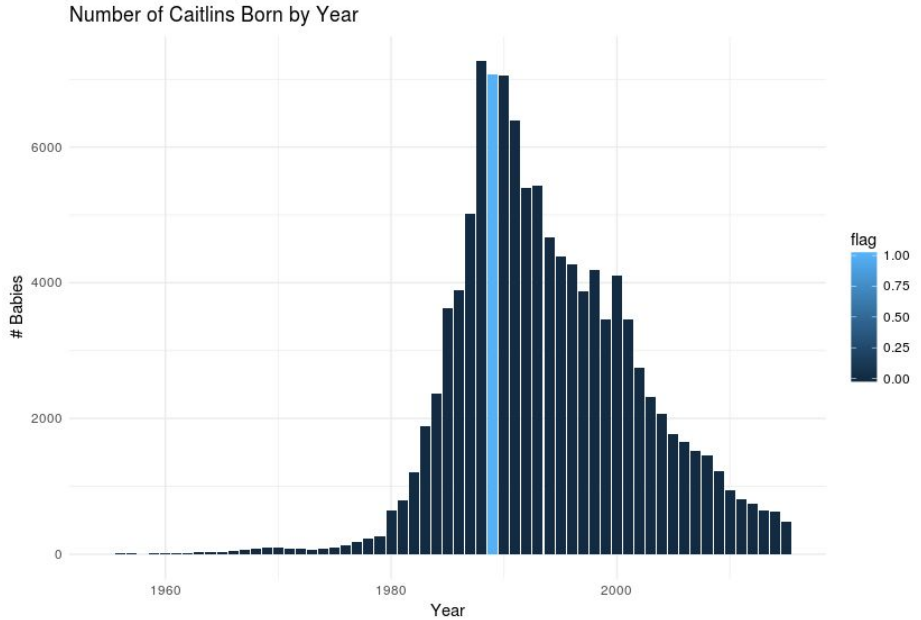
Grammar of graphics

+ Add or pipe plot elements together to create a graphic
+ Easy to create histograms, boxplots, scatterplots, violin plots + many more!

ggplot2 :: example




```
ggplot(caitlins, (aes(x = year, y = total))) +
  geom_histogram(stat = "identity")
```




```
ggplot(caitlins, (aes(x = year, y = total, fill=flag))) +
  geom_histogram(stat = "identity") + theme_minimal() +
  xlab("Year") + ylab("# Babies") +
  ggtitle("Number of Caitlins Born by Year")
```

lubridate

Makes it easier to convert and work with dates and times

+ Also helps with adding and subtracting dates
+ Functions include: `hour()`, `day()`, `week()`, `month()`, `year()`

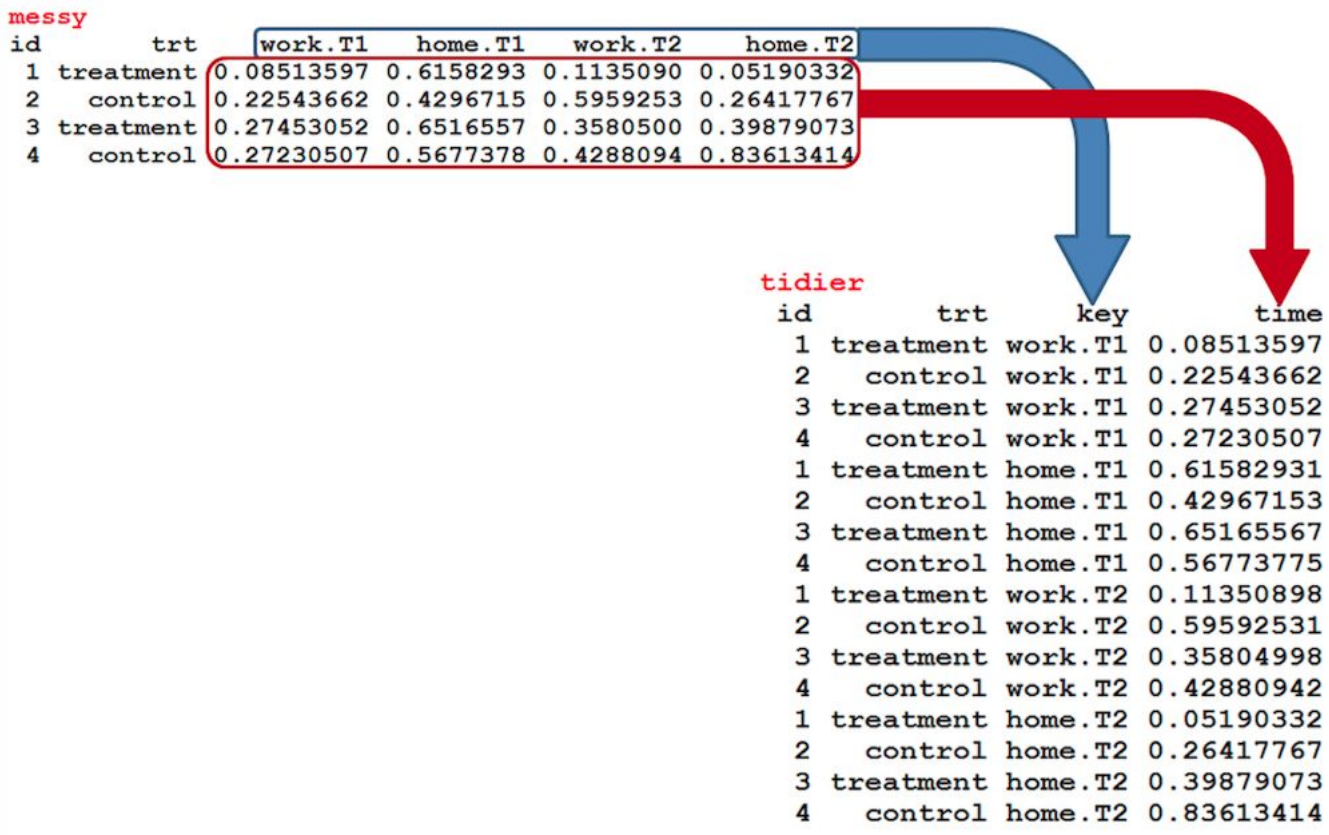
```
week('2017-01-01')
```

tidyr

Makes it easier to tidy your data by changing shape of datasets

+ Two most common functions: `spread()` and `gather()`
+ `spread()` takes two columns and spreads to multiple columns (this makes “long” data wider)
+ `gather()` takes multiple columns and gathers them to key-value pairs (this makes “wide” data longer)

tidyr :: example



messy

id	trt	work.T1	home.T1	work.T2	home.T2
1	treatment	0.08513597	0.6158293	0.1135090	0.05190332
2	control	0.22543662	0.4296715	0.5959253	0.26417767
3	treatment	0.27453052	0.6516557	0.3580500	0.39879073
4	control	0.27230507	0.5677378	0.4288094	0.83613414

tidier

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

… and the rest!

+ **`forcats`**: suite of tools to solve common problems with factors
+ **`haven`**: enables R to read and write data from various statistical packages
+ **`purrr`**: tools for working with function and vectors
+ **`readr`**: makes it easy to read rectangular data into R (like csv files)
+ **`readxl`**: makes it easy to get data from Excel into R with no dependencies
+ **`stringr`**: provides fast implementations of common string manipulations
+ **`tibble`**: reimagining of a df that forces you to be more explicit

Learn more:

Read Hadley Wickham's paper on tidy data

DataCamp has great intro courses:

+ Data Manipulation in R with dplyr
+ Data Visualization with ggplot2

Check out the official tidyverse website

Come to our R For Data Science workshop series this Fall!