Introduction to the Tidyverse
Day 2

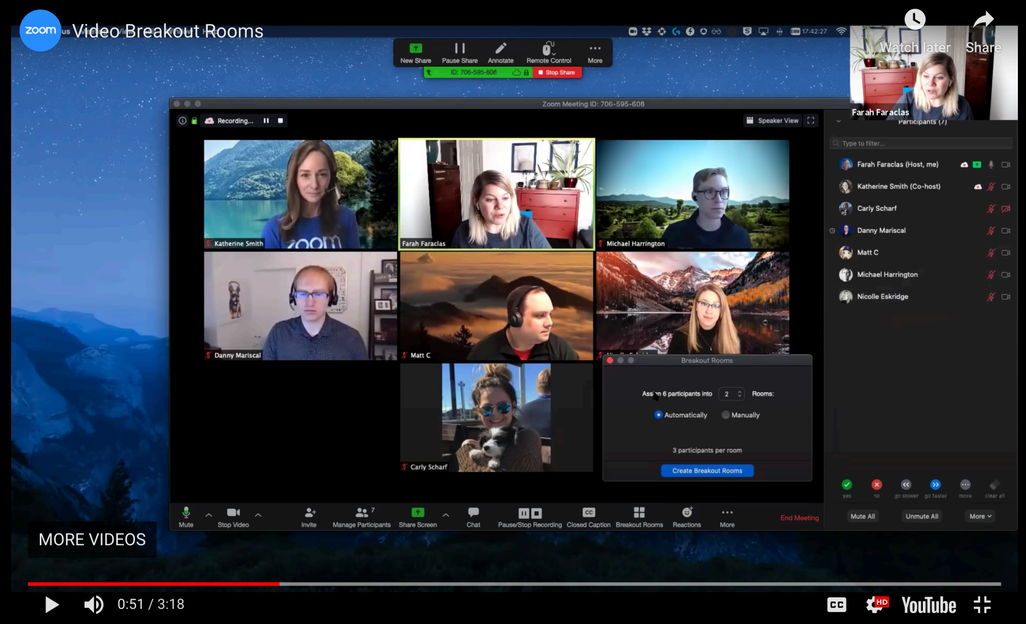
Ari Lamstein

Exercise: What do you already know?

- In the Chat Window type:

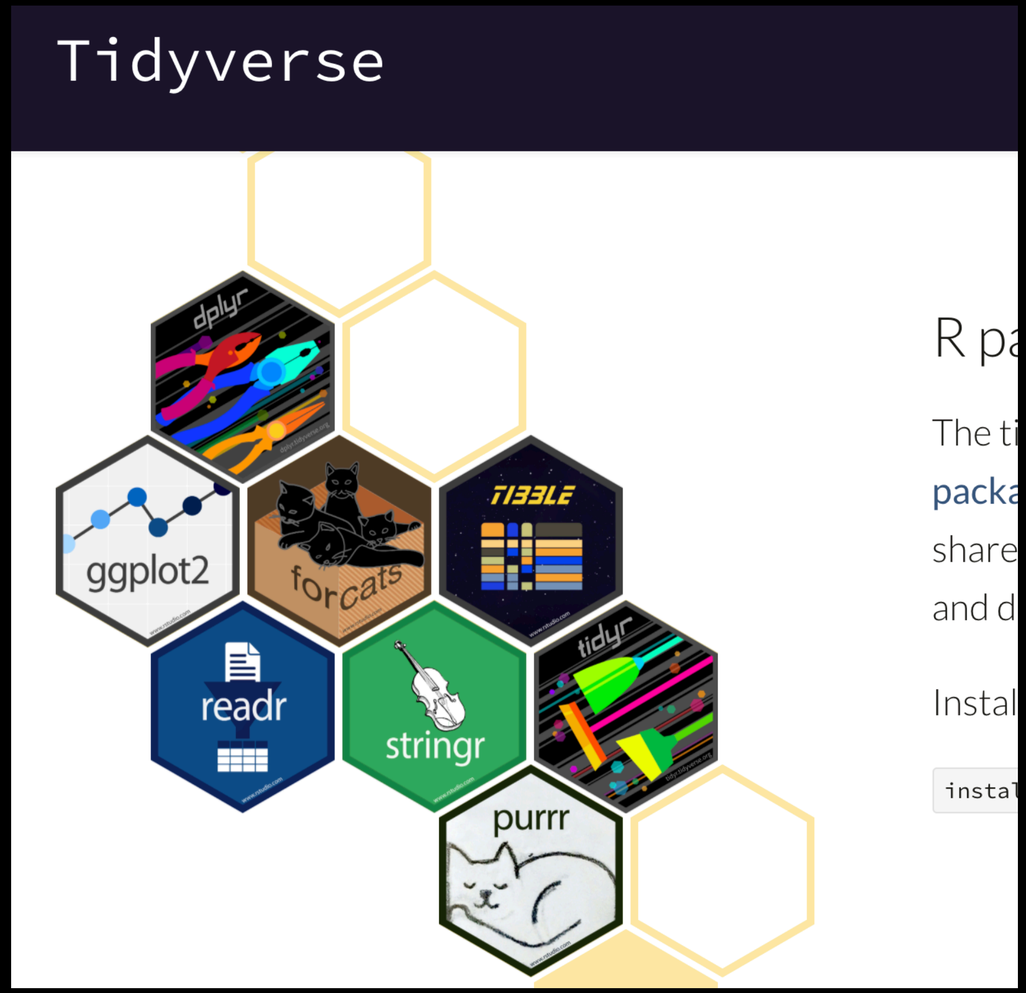
  1. What is one thing you remember / found useful from Day 1?

  2. What is something about the Tidyverse (or R) you wish you knew more about?



Two Modest Learning Goals

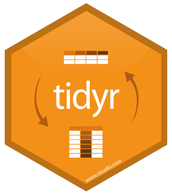
1. Review key concepts of the Tidyverse and Tidy Data
2. Learn today's Agenda



Tidy data

A data set is **tidy** iff:

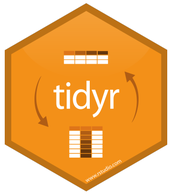
1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**
3. Each **value** is in its own **cell**



Tidy tools

country	year	cases	pop
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

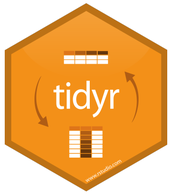
```
filter(df, year == 2000)
```



Tidy tools

country	year	cases	pop
Afghanistan	2000	2666	20595360
Brazil	2000	80488	174504898
China	2000	213766	1280428583

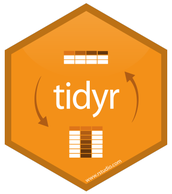
```
filter(df, year == 2000)
select(df, -year)
```



Tidy tools

country	cases	pop	rate
Afghanistan	2666	20595360	0.00013
Brazil	80488	174504898	0.00046
China	213766	1280428583	0.00017

```
filter(df, year == 2000)
select(df, -year)
mutate(df, rate = cases / pop)
```

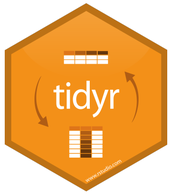


Tidy tools

country	cases	pop	rate
Afghanistan	2666	20595360	0.00013
Brazil	80488	174504898	0.00046
China	213766	1280428583	0.00017

avg
0.00025

```
filter(df, year == 2000)
select(df, -year)
mutate(df, rate = cases / pop)
summarise(df, avg = mean(rate))
```



Tidy tools

country	cases	pop	rate
Afghanistan	2666	20595360	0.00013
Brazil	80488	174504898	0.00046
China	213766	1280428583	0.00017

avg
0.00025

```
df %>%
  filter(year == 2000) %>%
  select(-year) %>%
  mutate(rate = cases / pop) %>%
  summarise(avg = mean(rate))
```

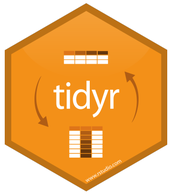
Non-Tidy R

Lists

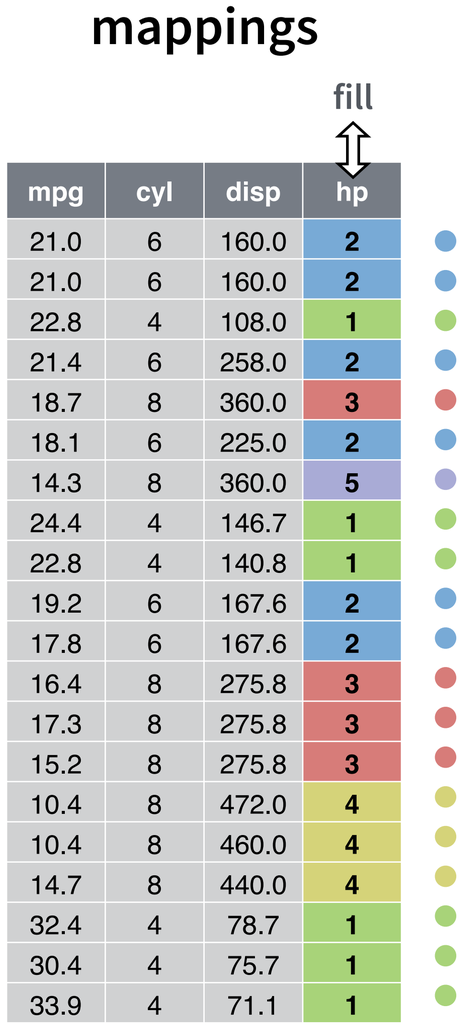
```
$city
[1] "New York" "New York" "London"
[4] "London"   "Beijing"  "Beijing"

$size
[1] "large" "small" "large" "small"
[5] "large" "small"

$amount
[1]  23  14  22  16 121 121

attr(,"row.names")
[1] 1 2 3 4 5 6
```

Models

```
Call:
lm(formula = lifeExp ~ year, data = gapminder)

Residuals:
   Min      1Q  Median      3Q     Max
-39.949  -9.651   1.697  10.335  22.158

Coefficients:
                  Estimate Std. Error t value   Pr(>|t|)
(Intercept) -585.65219   32.31396  -18.12   <2e-16 ***
year                0.32590     0.01632   19.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.63 on 1702 degrees of freedom
Multiple R-squared:  0.1898,      Adjusted R-squared:  0.1893
F-statistic: 398.6 on 1 and 1702 DF,  p-value: < 2.2e-16
```



To make a graph



1. Pick a **data** set

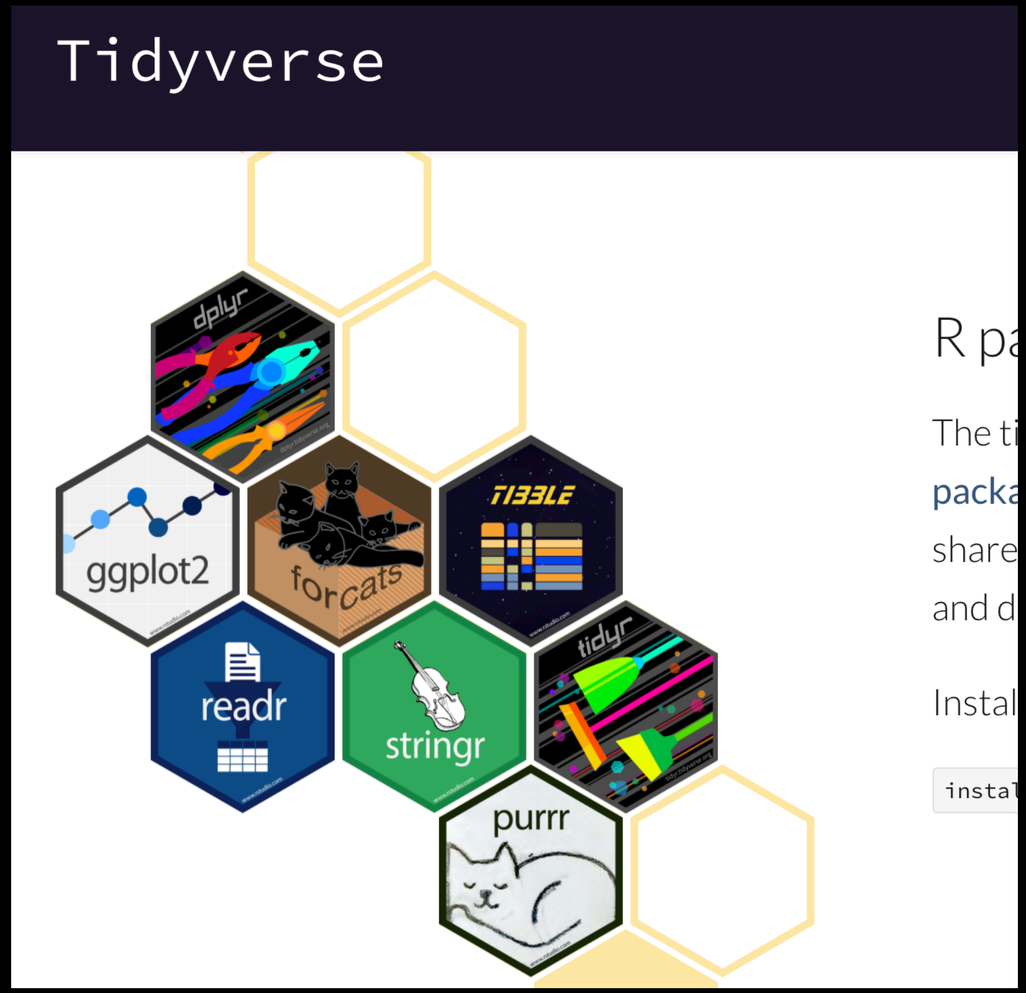
```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

2. Choose a **geom** to display cases

3. **Map** aesthetic properties to variables

Two Modest Learning Goals

1. Review key concepts of the Tidyverse and Tidy Data
2. Learn today's Agenda



Today

Functions for specific types of data.

strings

factors

dates

times

Train-the-Trainer