ggplot2 and the subjectivity of “tidy”

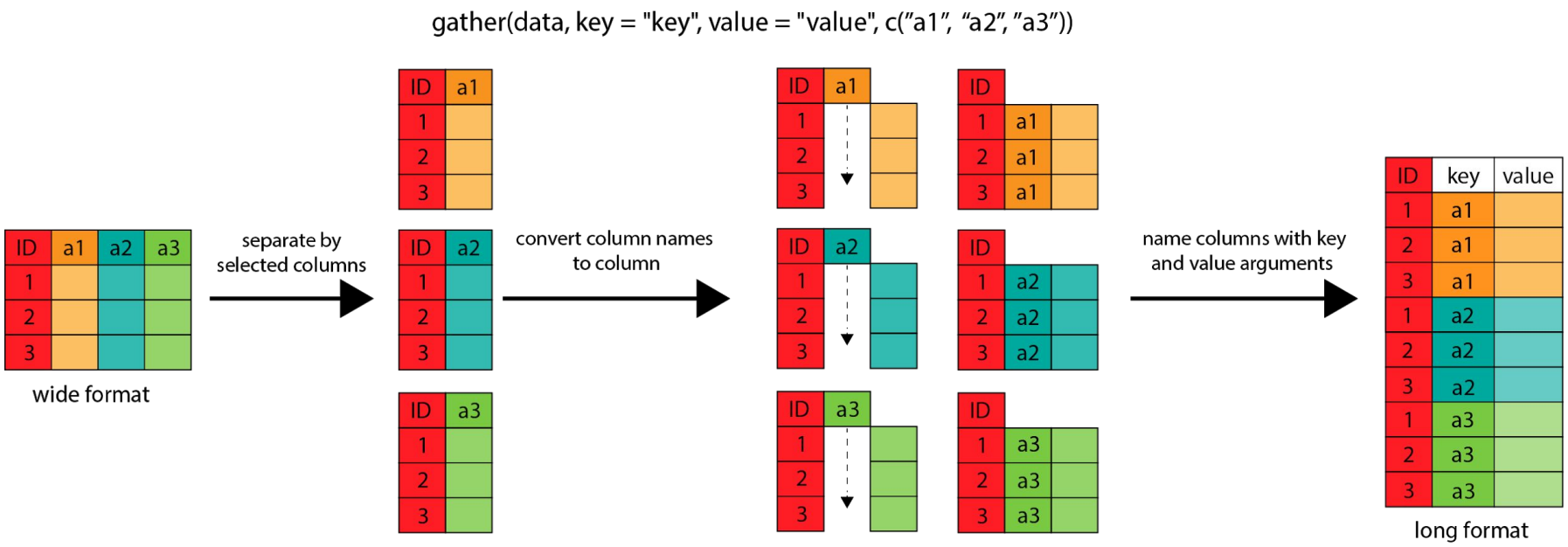
Marianna Foos

R-Ladies Boston

March 28, 2018

tidyr – your secret weapon for ggplot

- ggplot wants tidy data
- “tidy” means “long format”



tidyr – long data

- ggplot wants long data
- x = country, y = gdpPercap 👍

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Albania	Europe	1952	55.230	1282697	1601.0561
Algeria	Africa	1952	43.077	9279525	2449.0082
Angola	Africa	1952	30.015	4232095	3520.6103
Argentina	Americas	1952	62.485	17876956	5911.3151
Australia	Oceania	1952	69.120	8691212	10039.5956
Austria	Europe	1952	66.800	6927772	6137.0765
Bahrain	Asia	1952	50.939	120447	9867.0848
Bangladesh	Asia	1952	37.484	46886859	684.2442

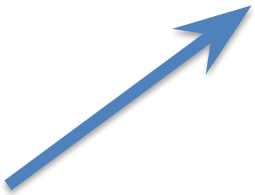
tidyr – wide data

- **ggplot wants long data**
- x = learner, y = ??? 😱

learner	test1	test2	test3	midterm	final
Katherine Johnson	99	83	89	85	100
Mae Jemison	84	97	86	85	97
Shirley Ann Jackson	99	96	92	86	82
Chanda Prescod-Weinstein	97	89	81	97	90
Alice Ball	91	90	86	81	87
Mamie Phipps Clark	95	92	81	93	98
Patricia Bath	95	87	93	84	82
Betty Harris	98	89	97	85	94
Jewel Plummer Cobb	97	96	84	82	80

tidyr::gather()

```
data %>%
  gather(key, value, [columns])
```



Name for new column containing labels

Name for new column containing values

Which columns to modify (or not)

tidyr – gather()

Before:

learner	test1	test2	test3	midterm	final
Katherine Johnson	99	83	89	85	100
Mae Jemison	84	97	86	85	97
Shirley Ann Jackson	99	96	92	86	82
Chanda Prescod-Weinstein	97	89	81	97	90

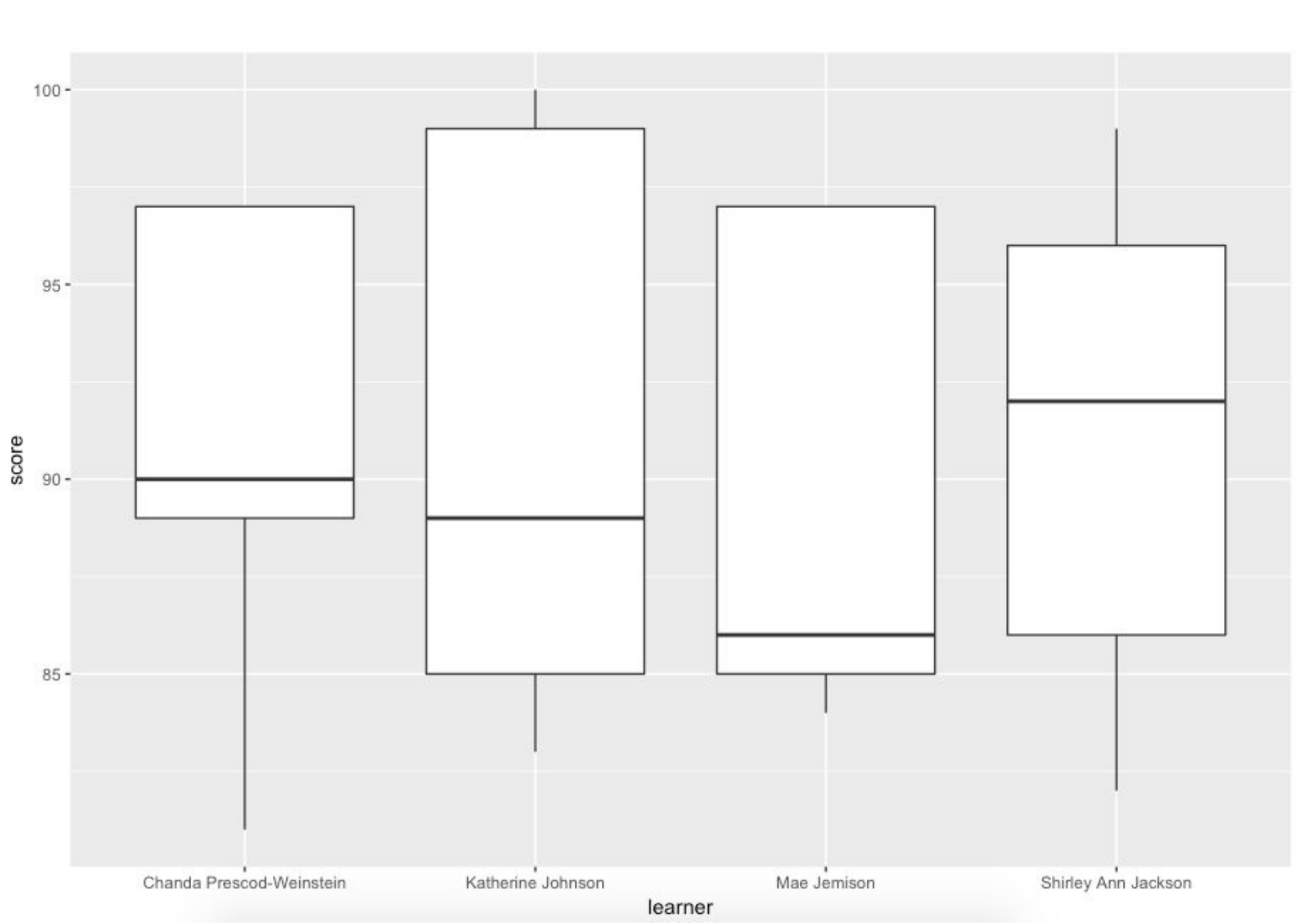
```
grades %>%
   gather(key = "exam", value = "score", -learner)
```

After:

learner	exam	score
Katherine Johnson	test1	99
Mae Jemison	test1	84
Shirley Ann Jackson	test1	99
Chanda Prescod-Weinstein	test1	97
Katherine Johnson	test2	83
Mae Jemison	test2	97
Shirley Ann Jackson	test2	96
Chanda Prescod-Weinstein	test2	89
Katherine Johnson	test3	89

x = learner

y = score


100
95
90
85
score
Chanda Prescod-Weinstein
Katherine Johnson
Mae Jemison
Shirley Ann Jackson
learner

Or use select helpers!

learner	test1	test2	test3
Katherine Johnson	99	83	89
Mae Jemison	84	97	86
Shirley Ann Jackson	99	96	92
Chanda Prescod-Weinstein	97	89	81

```
tests %>%
    gather(key = "exam", value = "score", starts_with("test"))
```

- `starts_with()`: starts with a prefix
- `ends_with()`: ends with a prefix
- `contains()`: contains a literal string
- `matches()`: matches a regular expression
- `num_range()`: a numerical range like
- `one_of()`: variables in character vector.
- `everything()`: all [remaining] variables

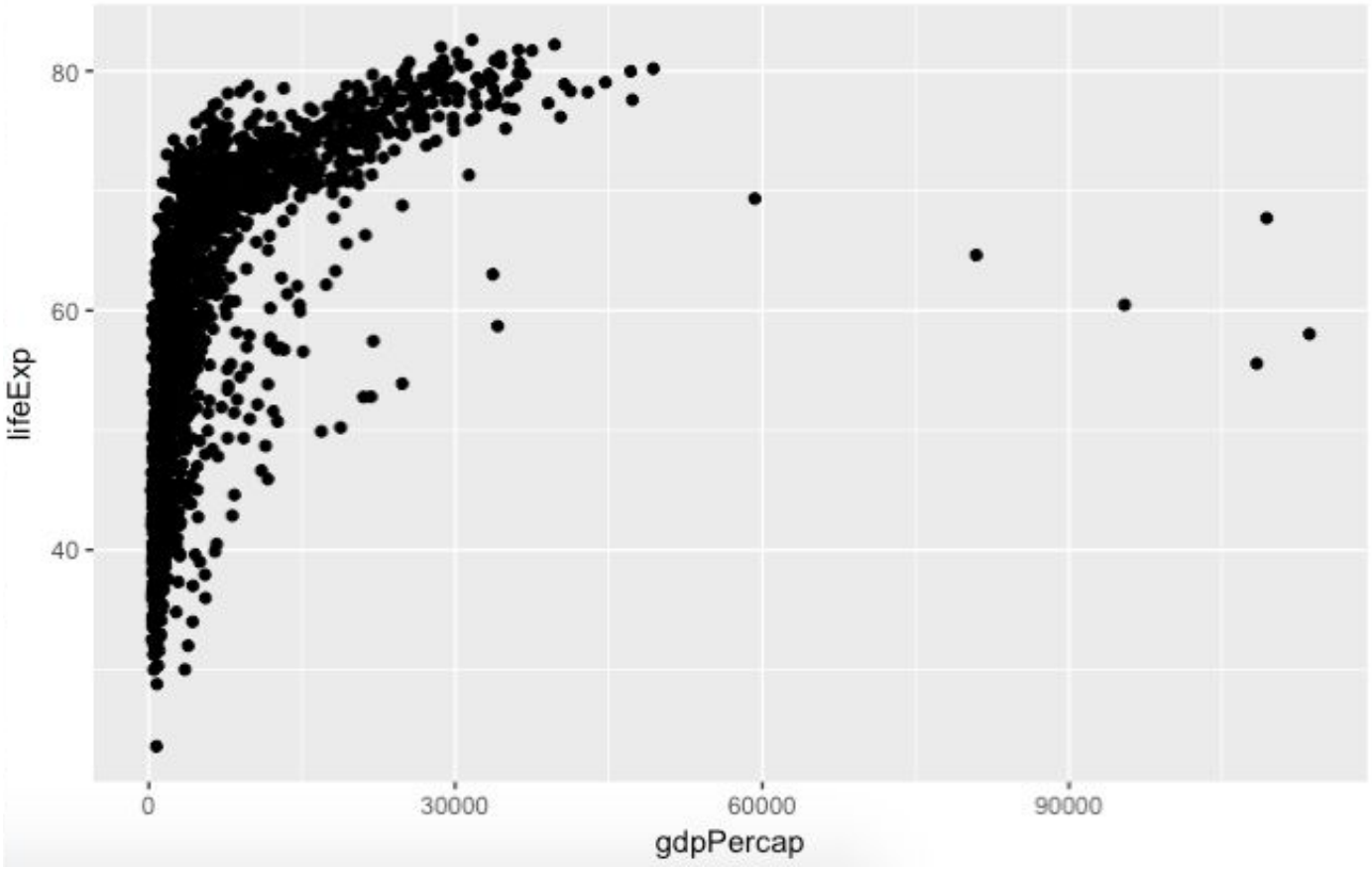
ggplot2 basics

1. Data
2. Aesthetic mapping
3. Geomzzz

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Albania	Europe	1952	55.230	1282697	1601.0561
Algeria	Africa	1952	43.077	9279525	2449.0082
Angola	Africa	1952	30.015	4232095	3520.6103
Argentina	Americas	1952	62.485	17876956	5911.3151
Australia	Oceania	1952	69.120	8691212	10039.5956
Austria	Europe	1952	66.800	6927772	6137.0765
Bahrain	Asia	1952	50.939	120447	9867.0848
Bangladesh	Asia	1952	37.484	46886859	684.2442

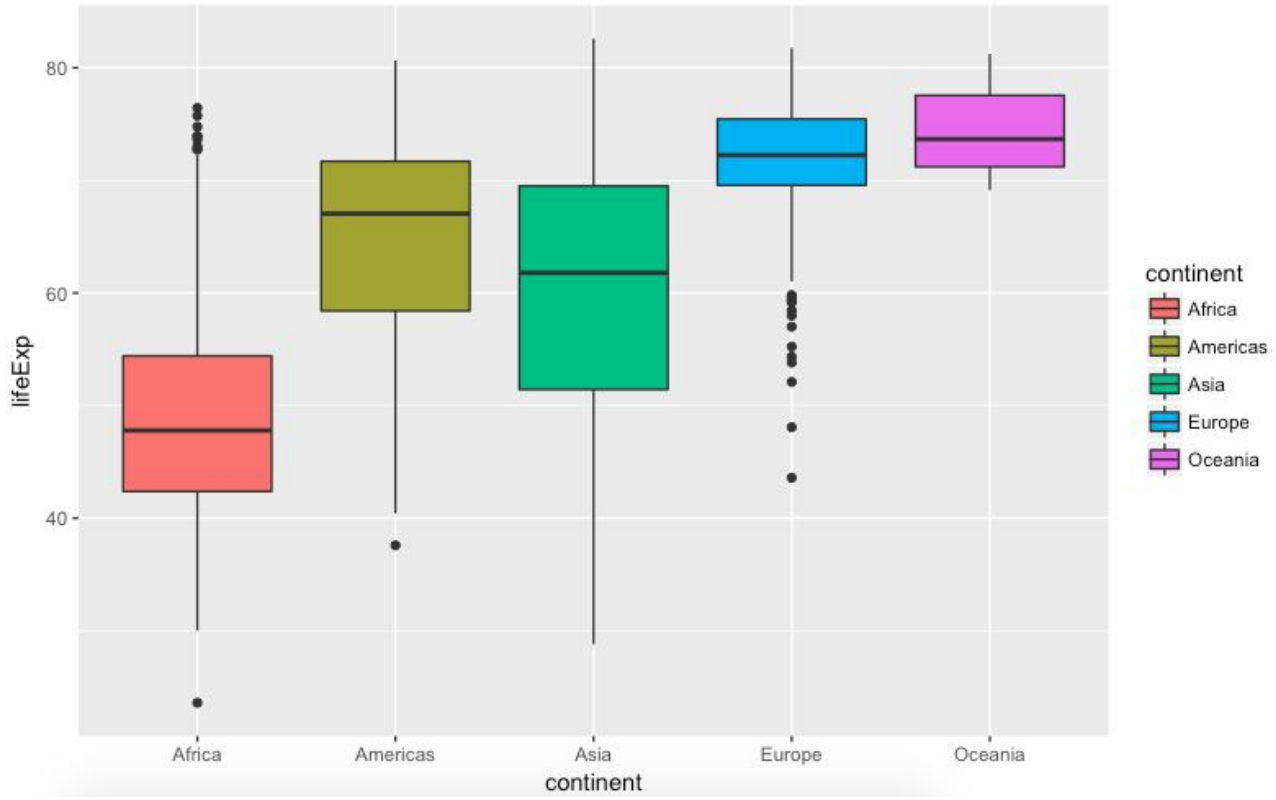
Example - scatter

```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp)) +
    geom_point()
```



Example - box

```
ggplot(gapminder,
aes(x = continent, y = lifeExp, fill = continent)) +
    geom_boxplot()
```



Resources

- Great ggplot2 slide deck from R-Ladies RTP: http://rpubs.com/eamcvey/218880
- Graphics chapter from R 4 Data Science: http://r4ds.had.co.nz/graphics-for-communication.html
- Data Carpentry's ggplot2 lesson: http://www.datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html
- Jenny Bryan's Secrets to Happy Graphing: http://stat545.com/block016_secrets-happy-graphing.html