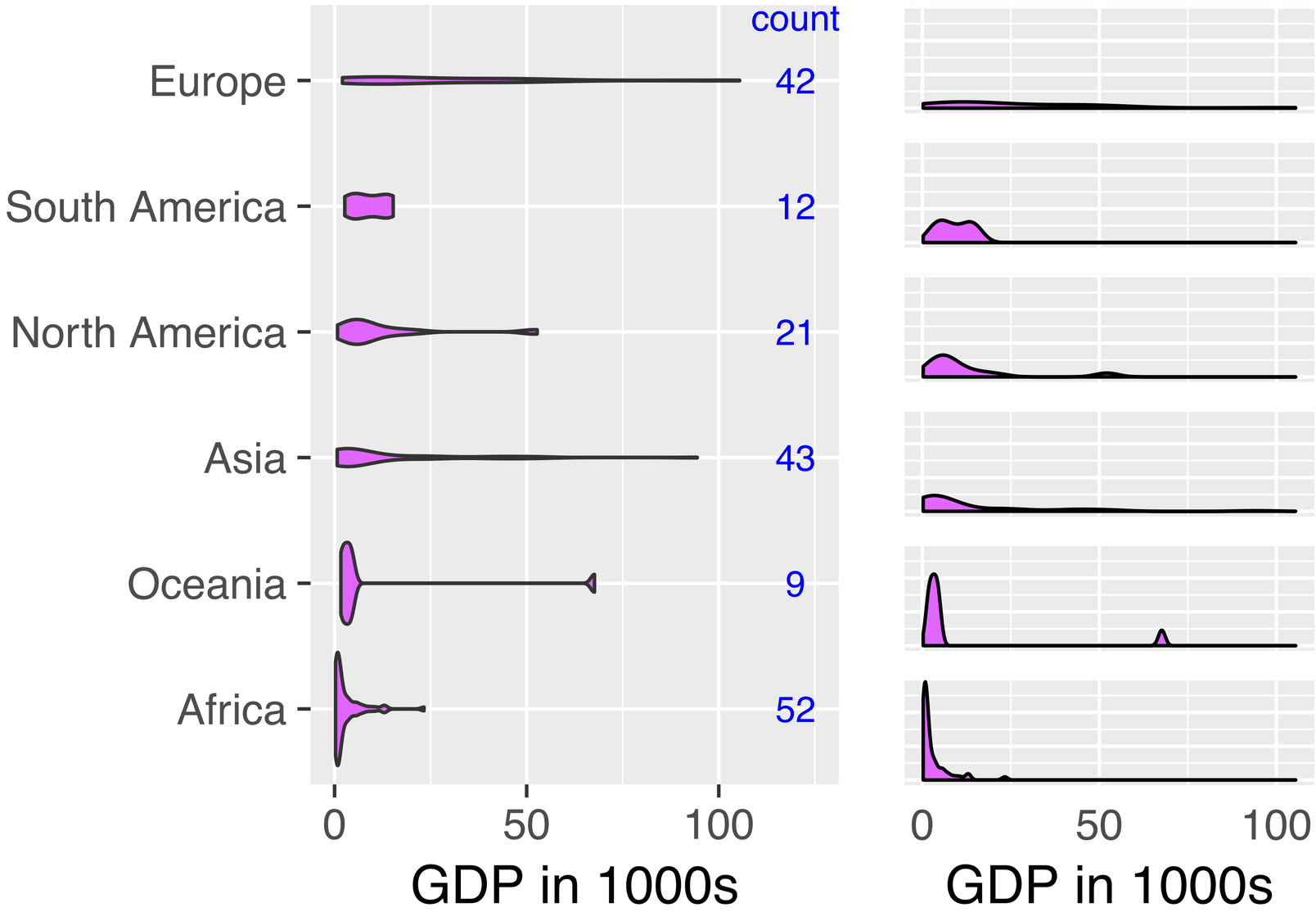
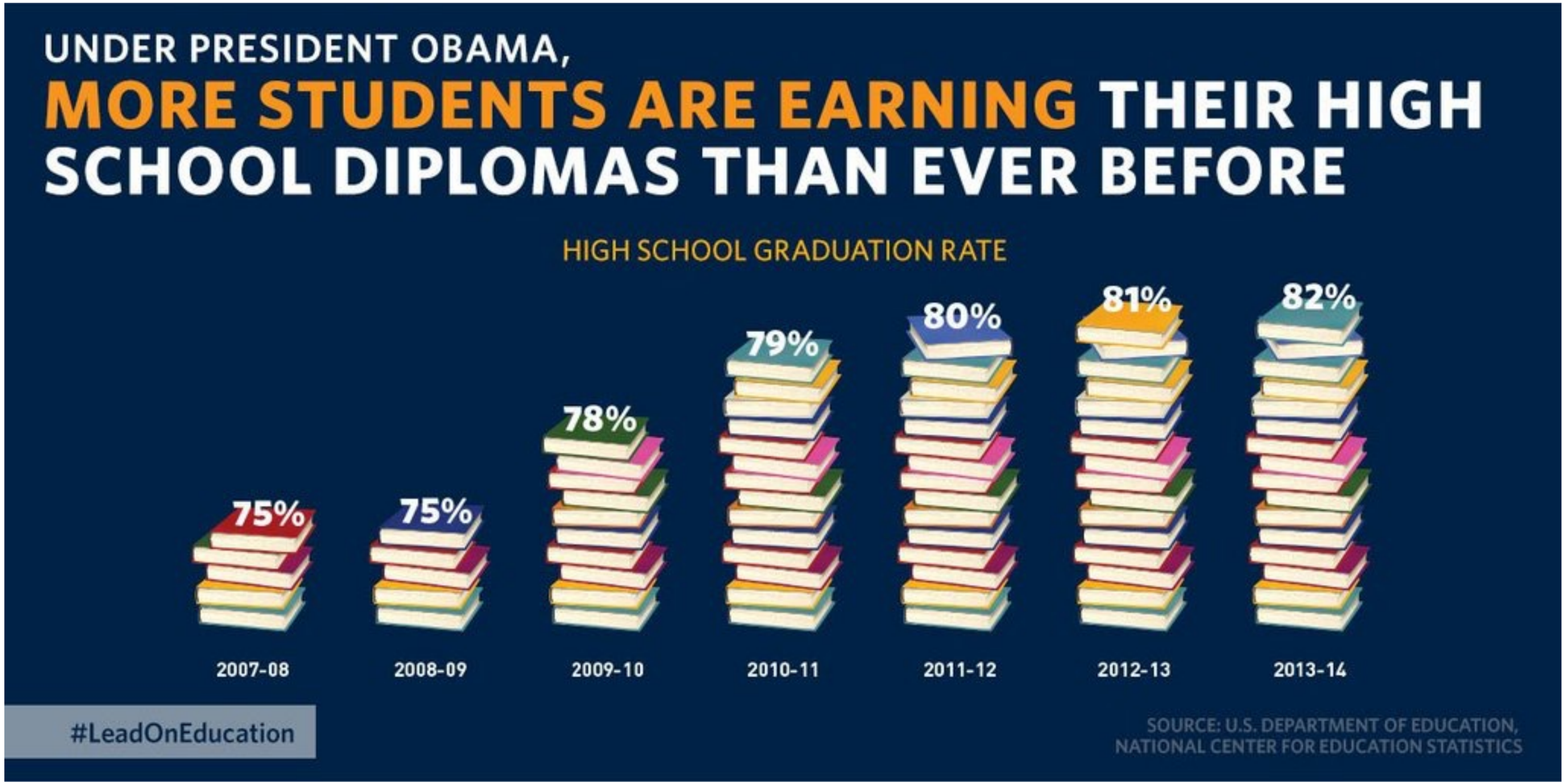
GR5702
Exploratory Data Analysis and Visualization

Prof. Joyce Robbins

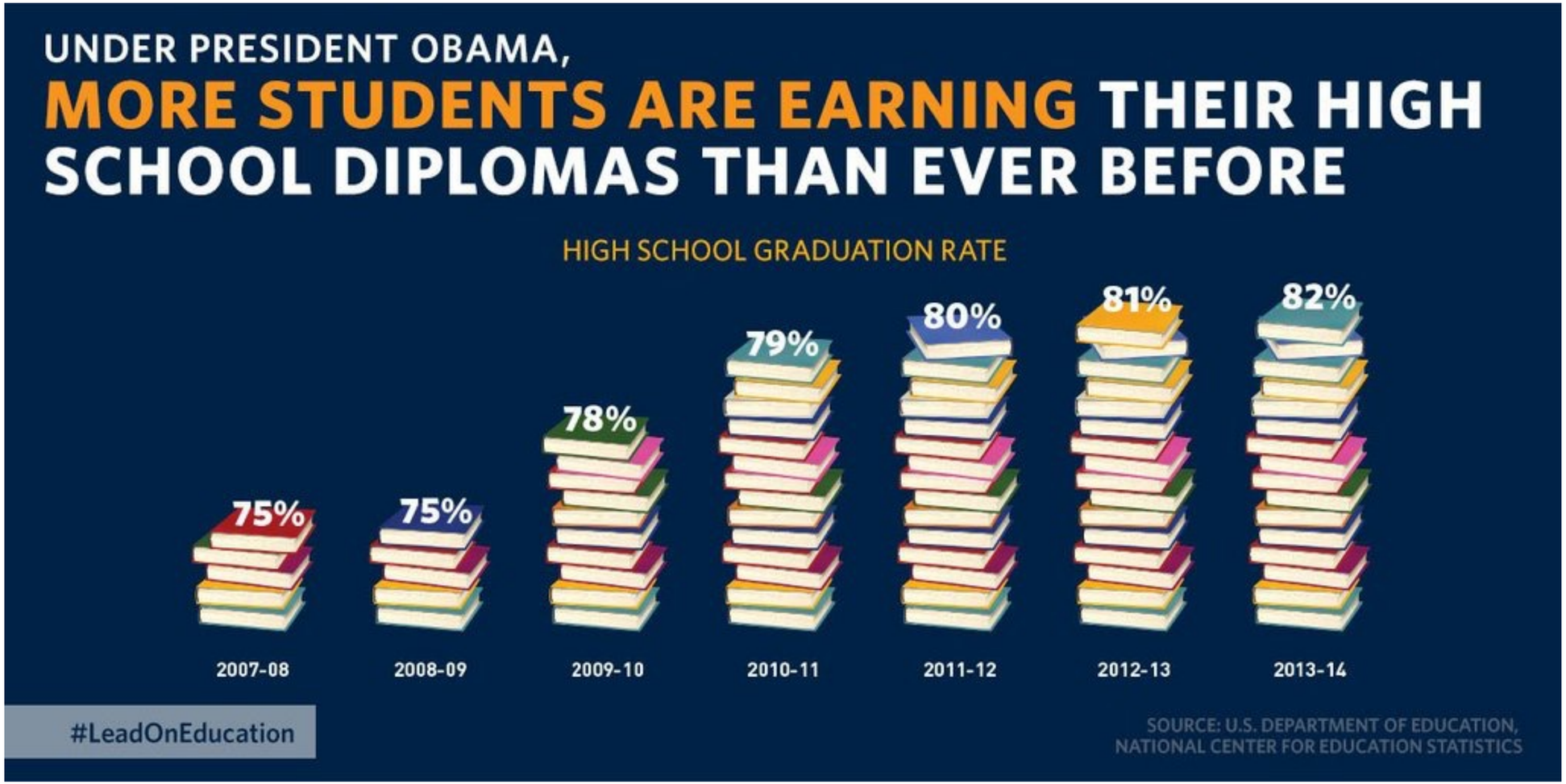
Violin plots vs. facetted histograms



Bad graphs

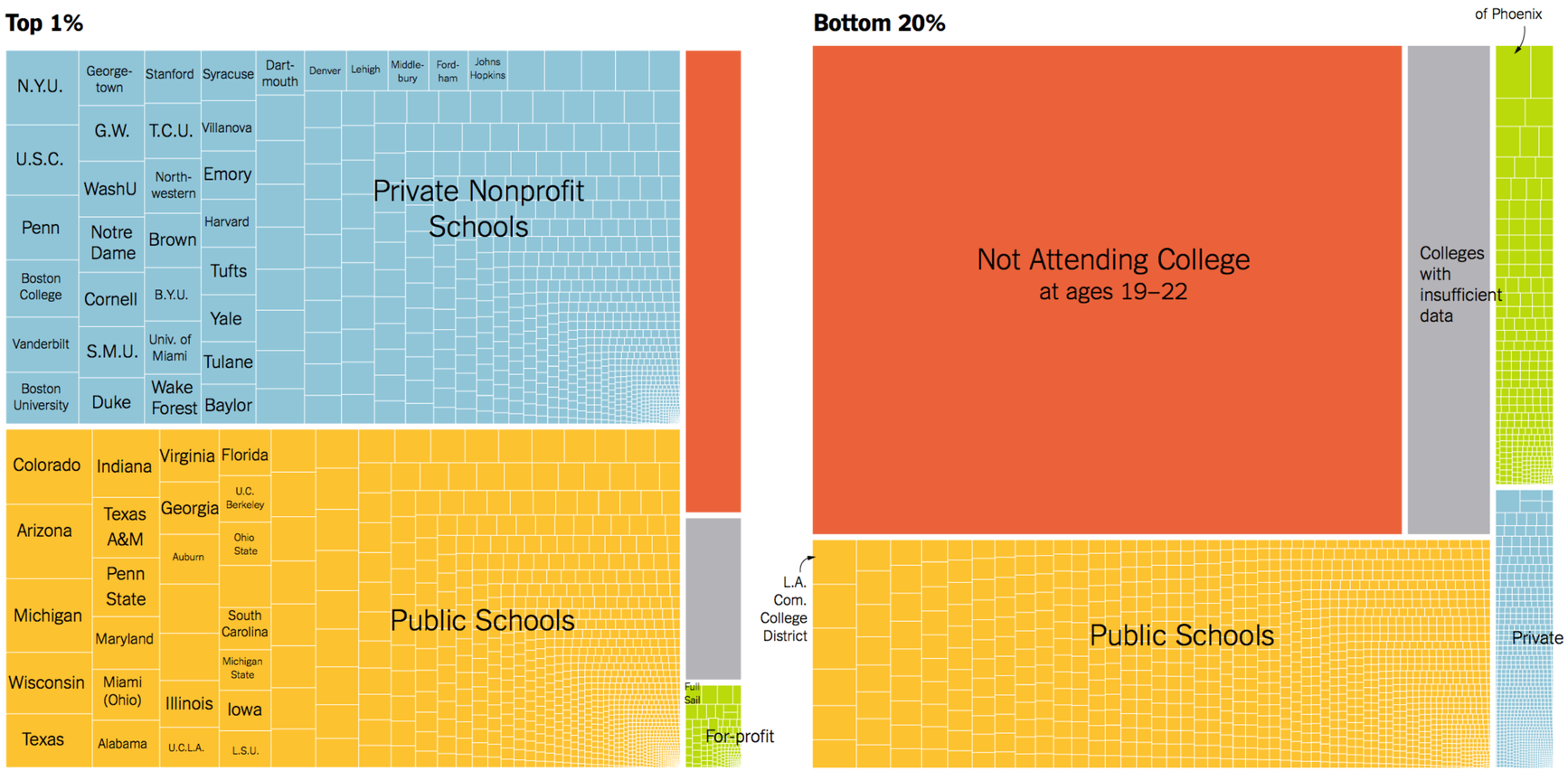


https://twitter.com/ObamaWhiteHouse/status/677189256834609152



... but they are not studying dataviz!

https://twitter.com/ObamaWhiteHouse/status/677189256834609152


Top 1%
N.Y.U.
George-town
Stanford
Syracuse
Dart-mouth
Denver
Lehigh
Middle-bury
Ford-ham
Johns Hopkins
U.S.C.
G.W.
T.C.U.
Villanova
WashU
North-western
Emory
Penn
Notre Dame
Brown
Harvard
Boston College
Cornell
B.Y.U.
Tufts
Vanderbilt
S.M.U.
Univ. of Miami
Yale
Boston University
Duke
Wake Forest
Tulane
Baylor
Private Nonprofit Schools
Colorado
Indiana
Virginia
Florida
Arizona
Texas A&M
Georgia
U.C. Berkeley
Penn State
Auburn
Ohio State
Michigan
Maryland
South Carolina
Wisconsin
Miami (Ohio)
Michigan State
Illinois
Iowa
Texas
Alabama
U.C.L.A.
L.S.U.
Public Schools
Full Sail
For-profit
Bottom 20%
of Phoenix
Not Attending College
at ages 19–22
Colleges with insufficient data
L.A. Com. College District
Public Schools
Private

Anatomy of a Winning TED Talk

1%

Sophisticated Visual Aids

We're not sure who puts the D in TED—most of the best presentations favor tepid PowerPoint slide shows (sorry, Brené Brown), Pictionary-quality drawings (really, Simon Sinek?), or no props at all.

5%

Opening Joke

Remember the one about the shoe salesmen who went to Africa in the 1900s? That's how Benjamin Zander opened his talk—which turned out to be about classical music.

5%

Spontaneous Moment

Don't overprepare. Tease the guy in the front row ("You could light up a village with this guy's eyes"). Commend the stagehand who handles the human brain you brought.

5%

Statement of Utter Certainty

People come for answers—give 'em what they want, as Shawn Achor did: "By training your brain … we can reverse the formula for happiness and success."

12%

Snappy Refrain

The TED equivalent of "I have a dream." Example: "People don't buy what you do; they buy why you do it." Repeat 7x.

23%

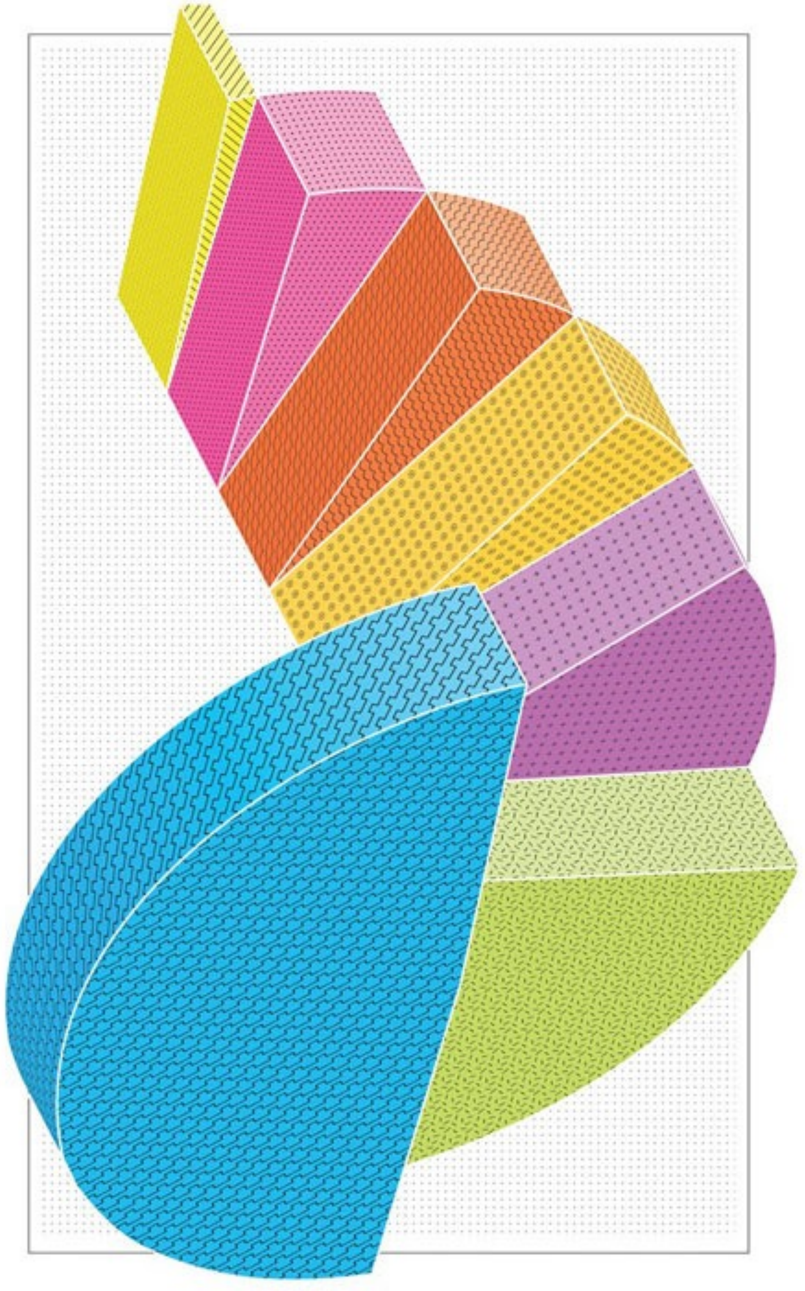
Personal Failure

Be relatable.We want to know about that nervous breakdown. Or at least the time you didn't fit in at summer camp.

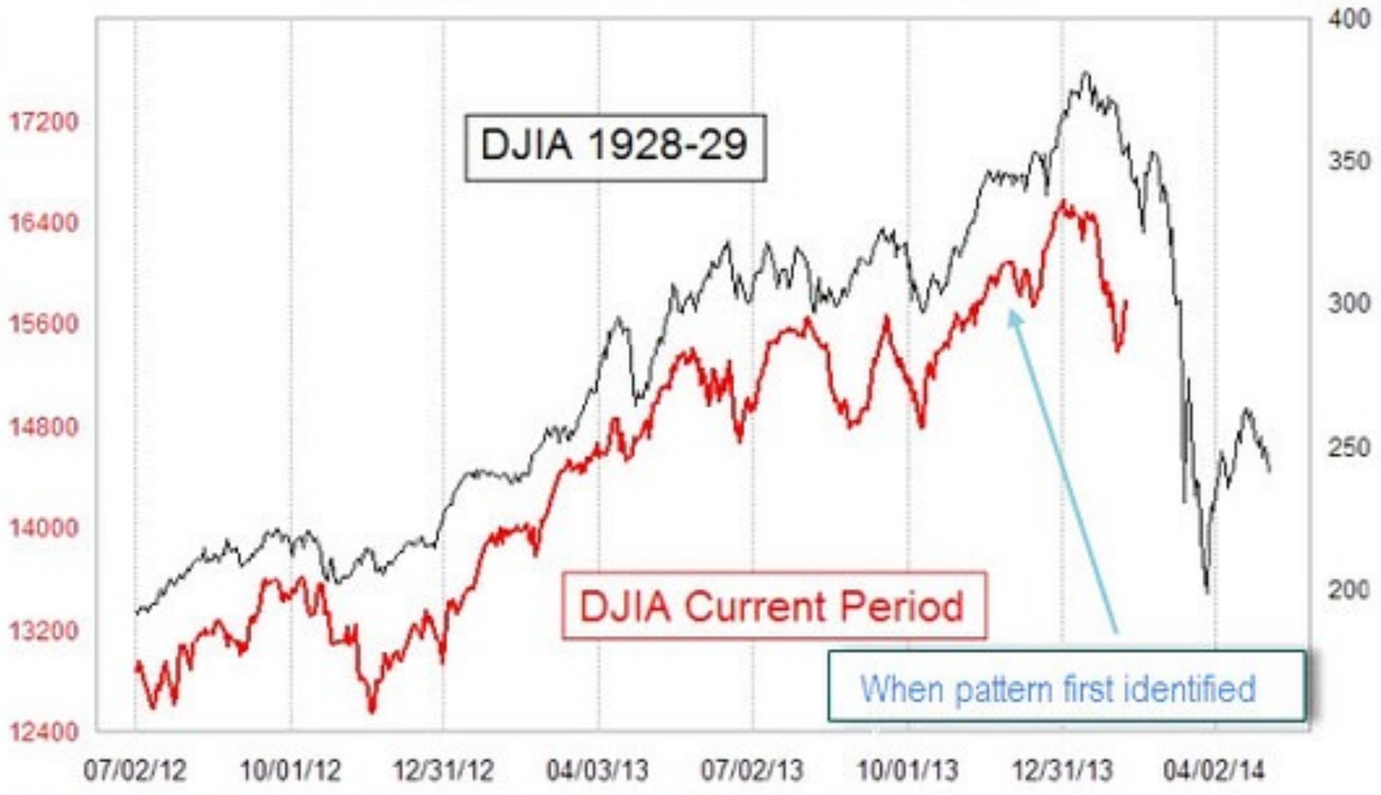
49%

Contrarian Thesis

Wait a sec—we should be playing *more* videogames? The more choices we have, the worse off we are? TED is where conventional wisdom goes to die.



SCARY PARALLEL



Source: McClellan Market Report, based on pattern discovered by Tom DeMark

Good graphs

Obama
ELECTORAL VOTES
243
Needs 27 to win

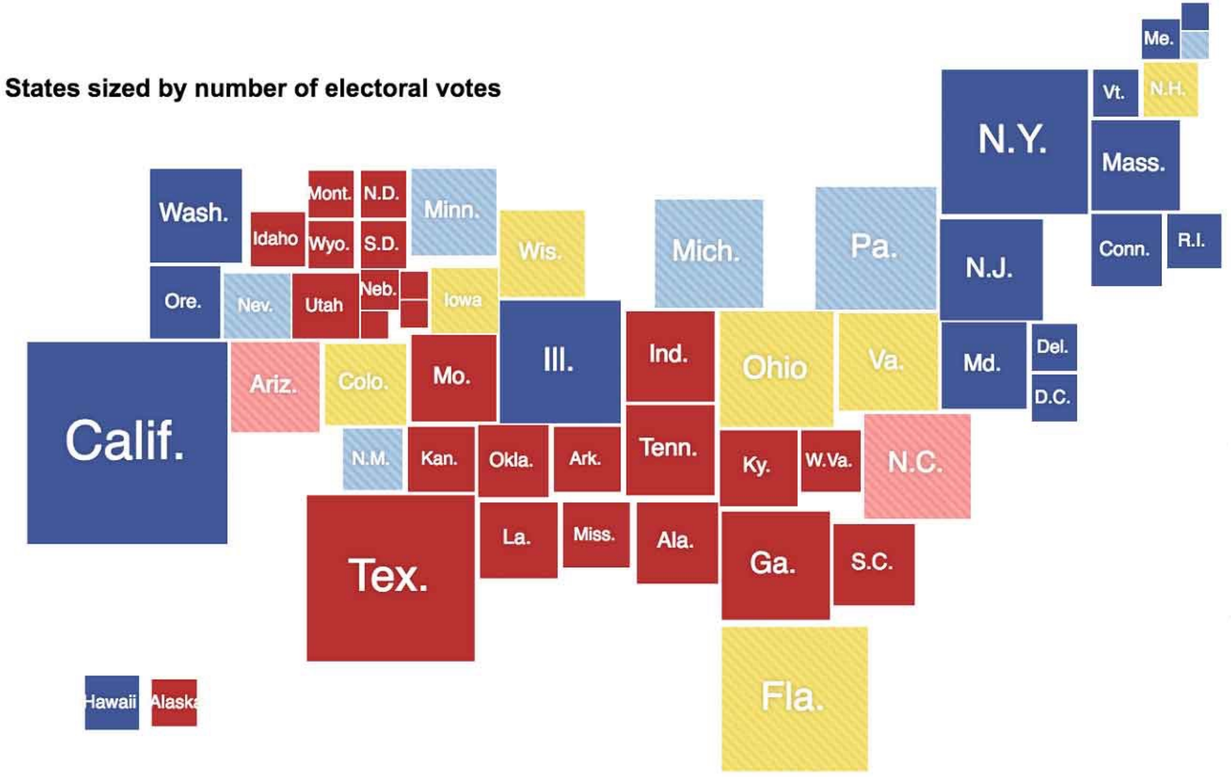
Needs 64 to win
206
Romney
ELECTORAL VOTES

185 Solid Obama | 58 Leaning Oba... | 89 Tossup Votes | 26 Le... | 180 Solid Romney

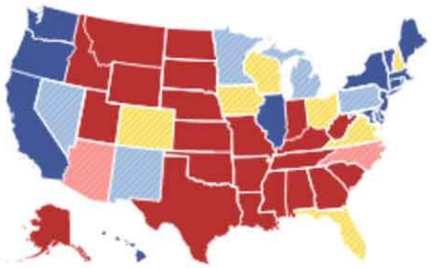
270 needed to win

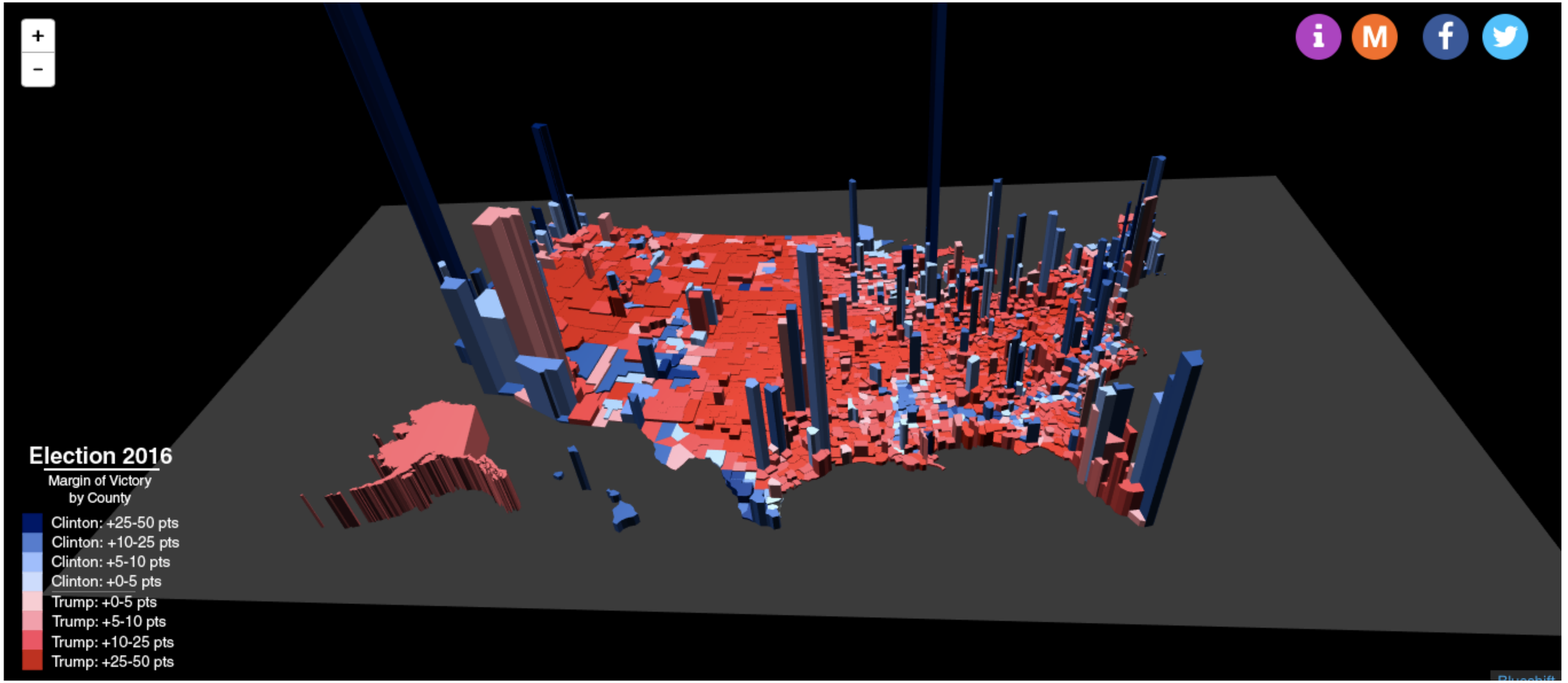
Maine and Nebraska give two electoral votes to the statewide winner and allocate the rest by congressional district.

States sized by number of electoral votes

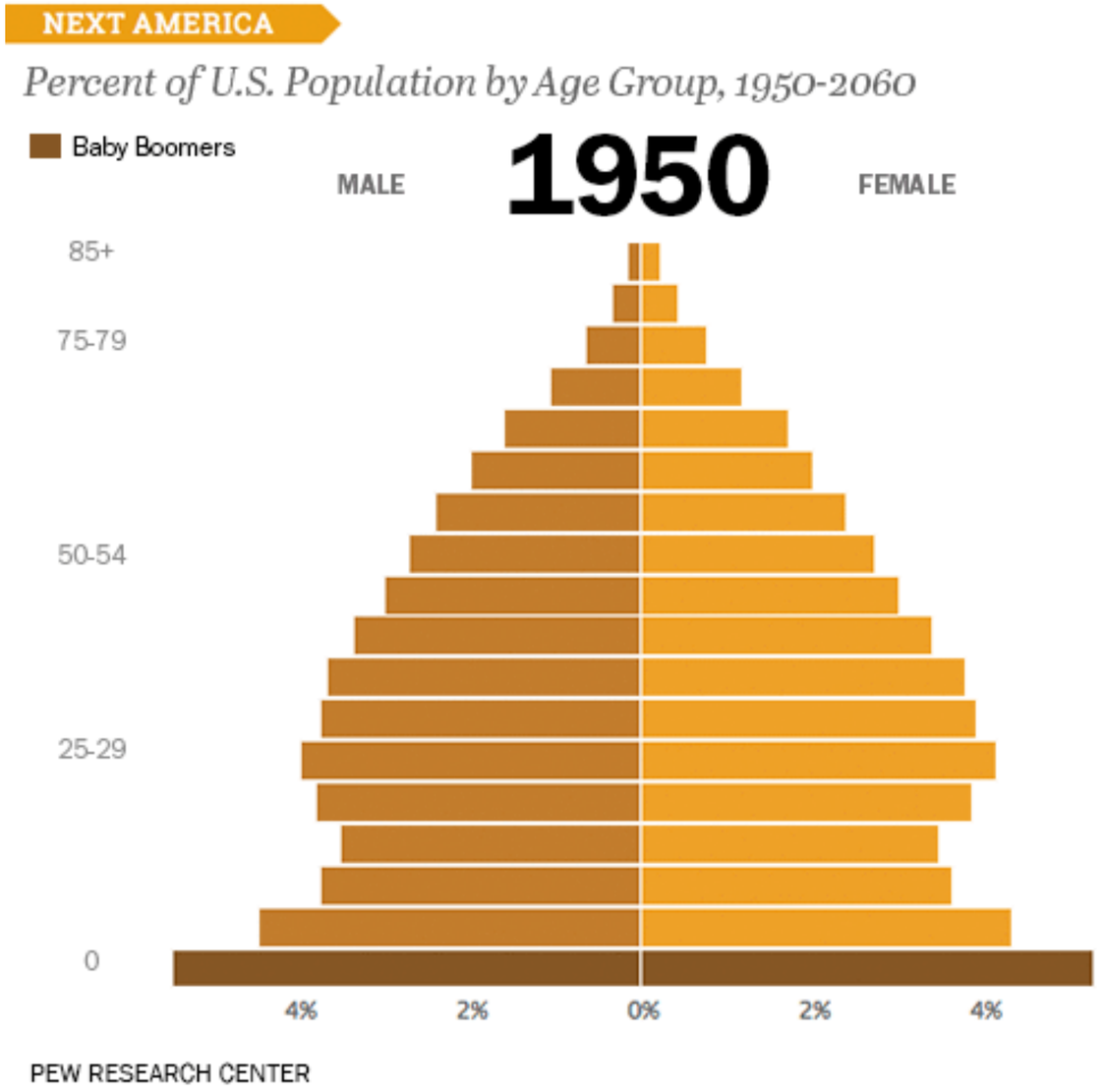


Geographic View

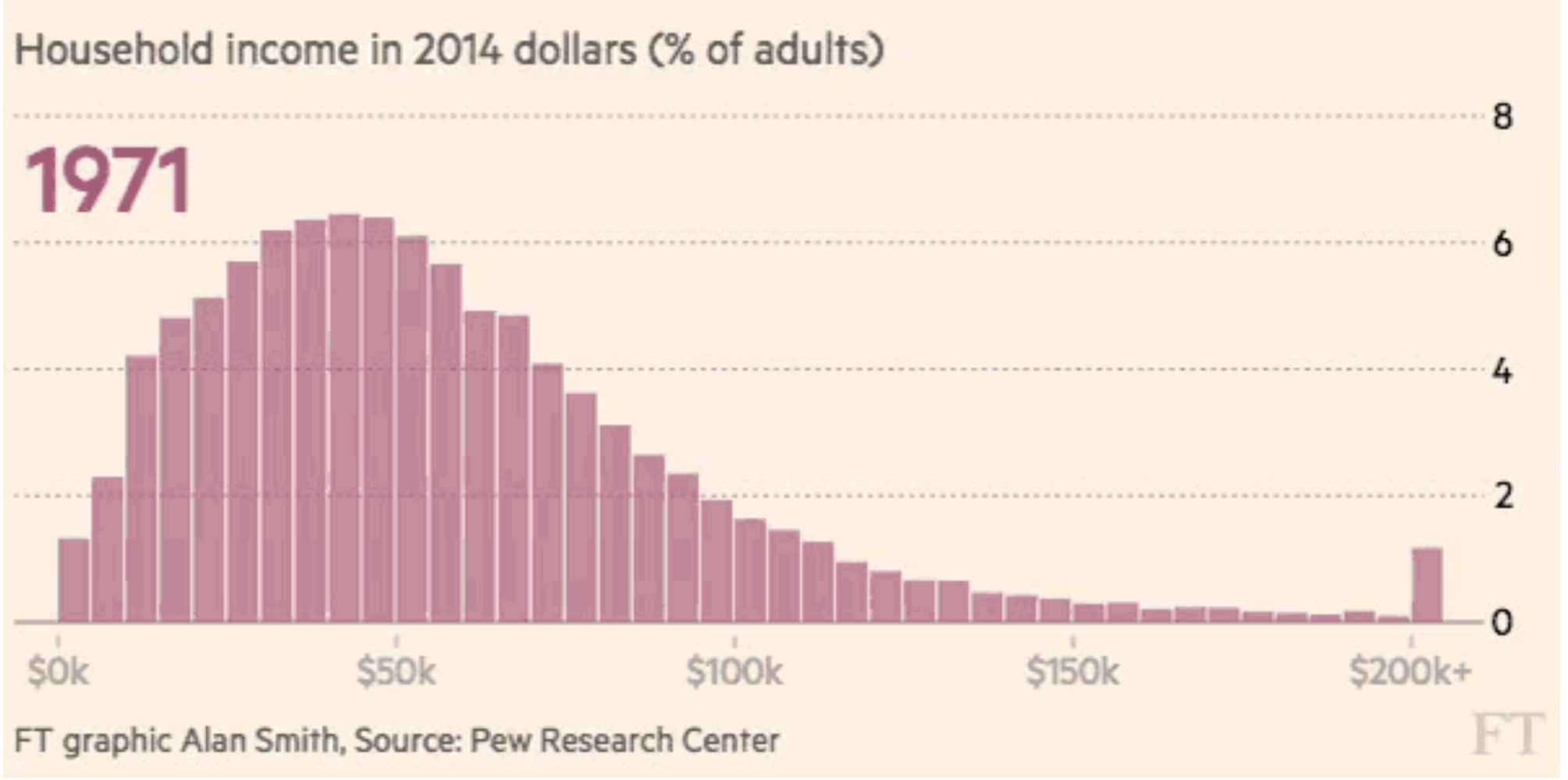




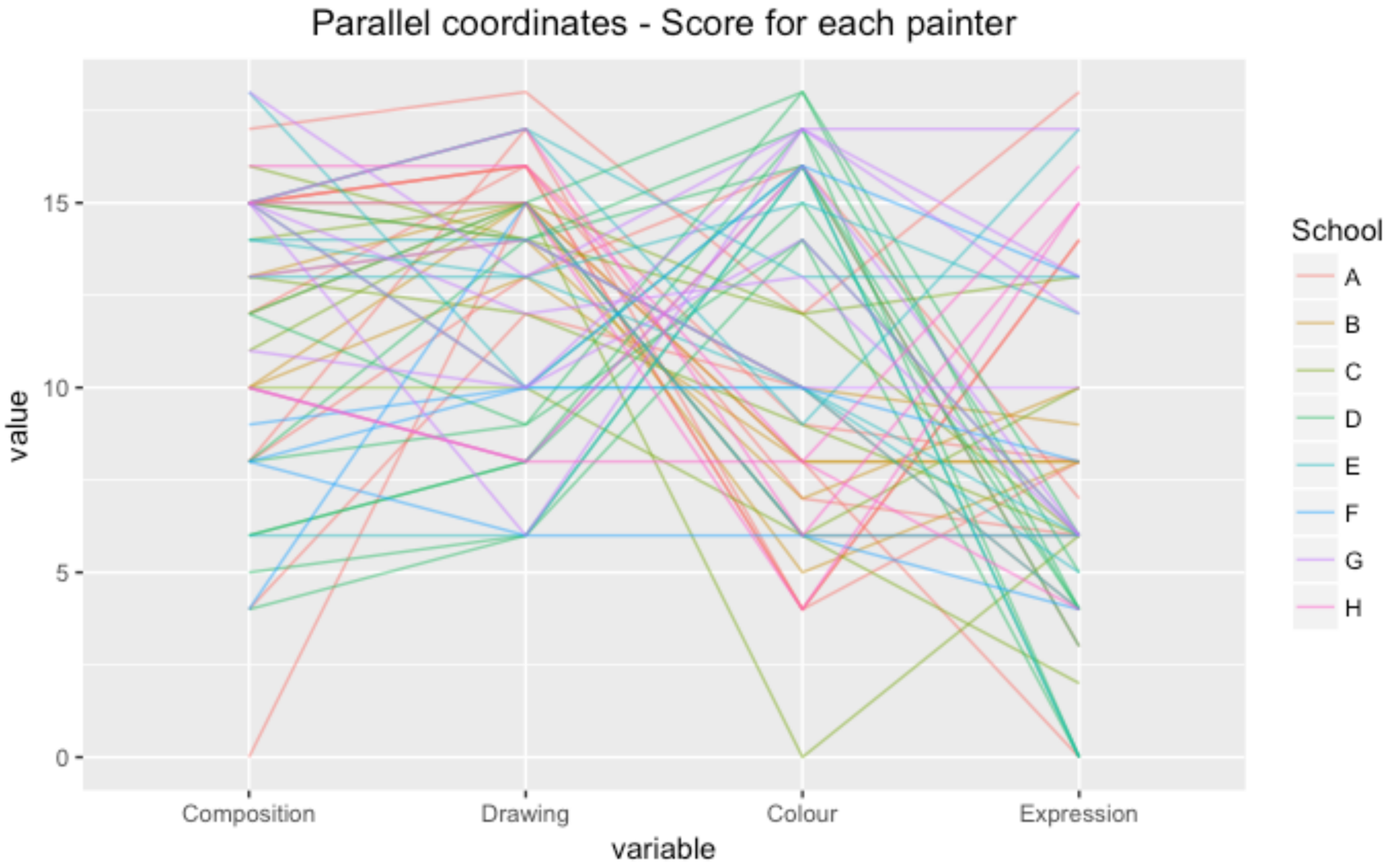
https://blueshift.io/election-2016-county-map.html

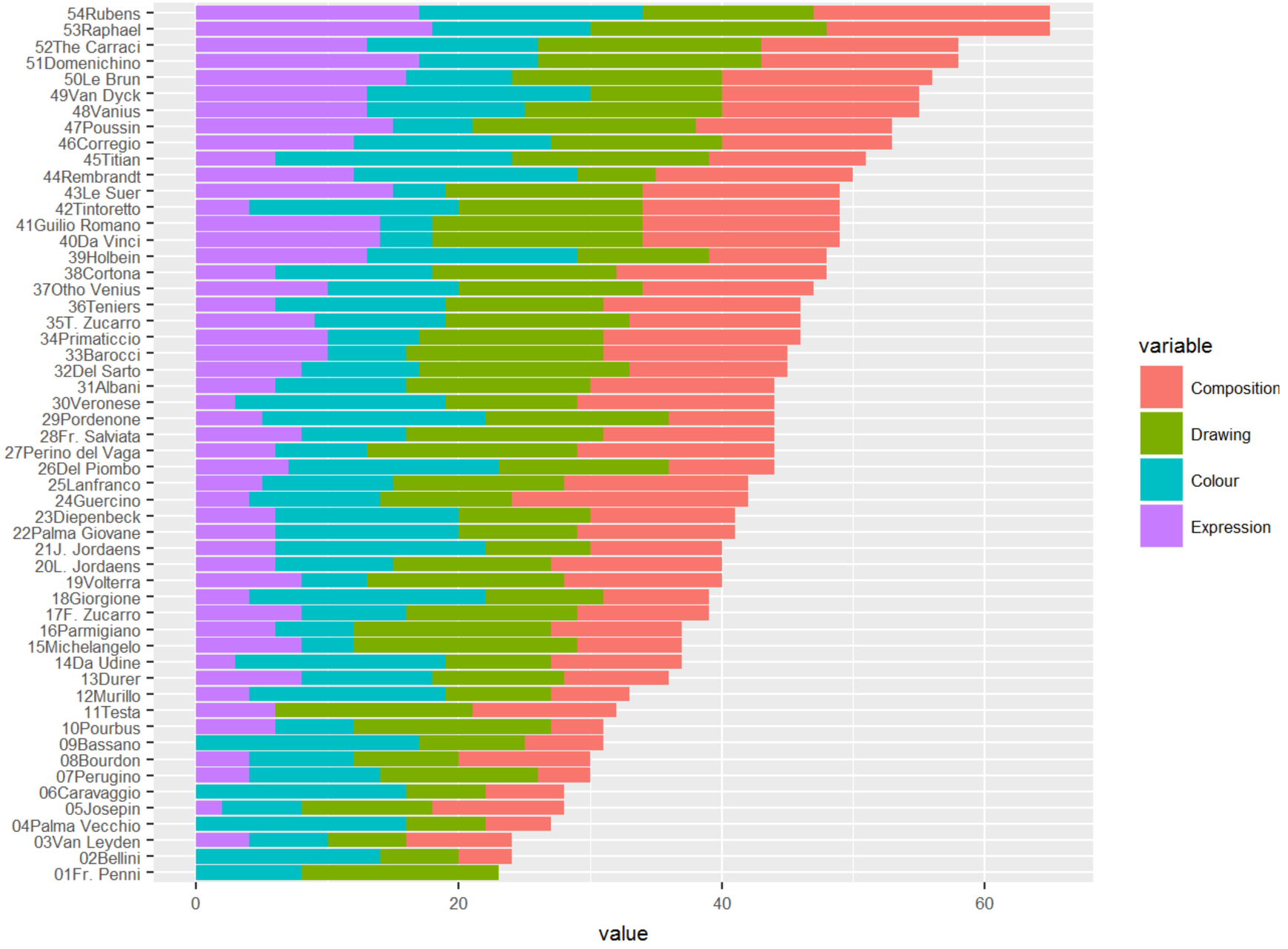


http://www.pewresearch.org/next-america/#Two-Dramas-in-Slow-Motion

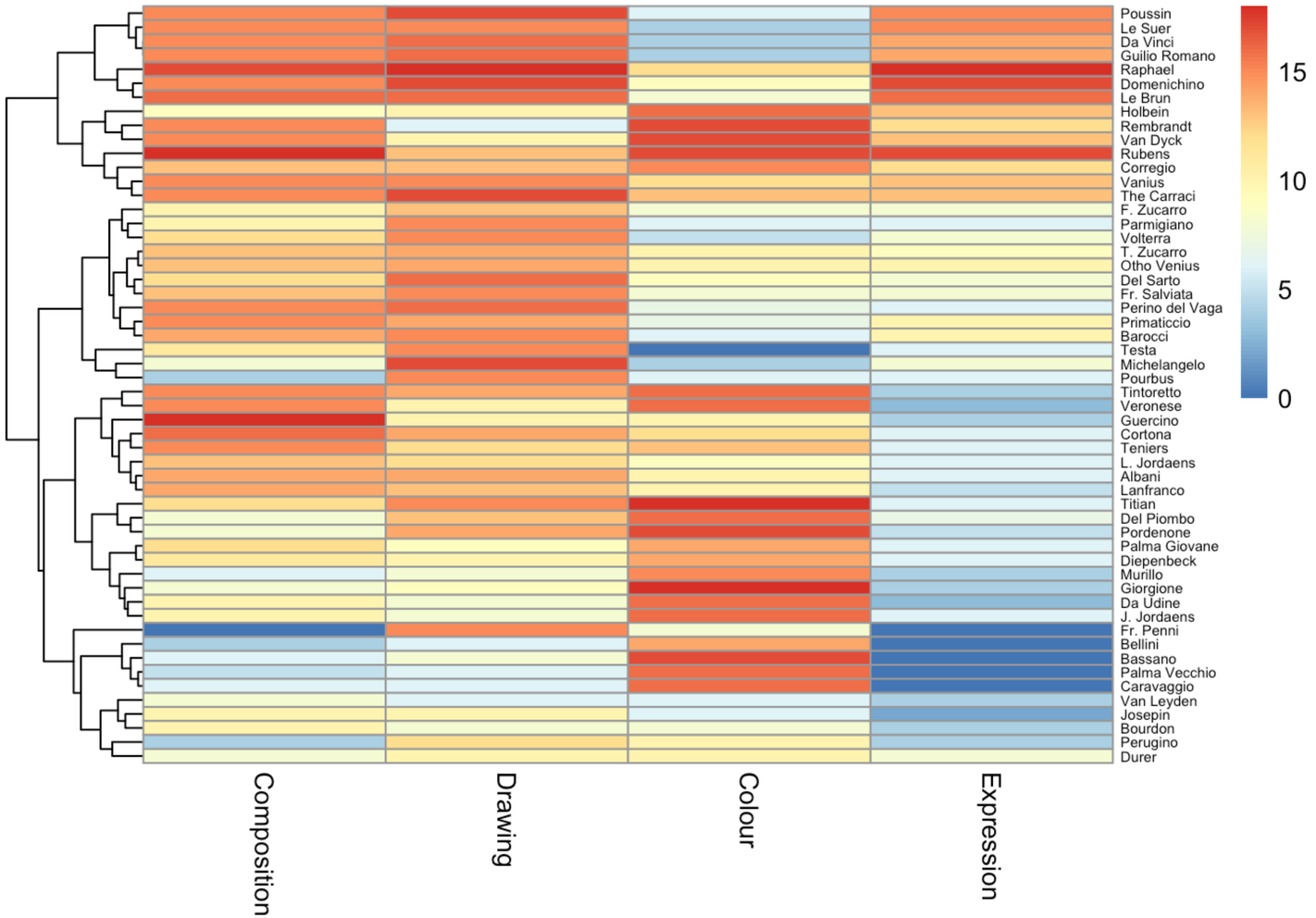


http://www.latimes.com/business/hiltzik/la-fi-hiltzik-ft-graphic-20160320-snap-htmlstory.html

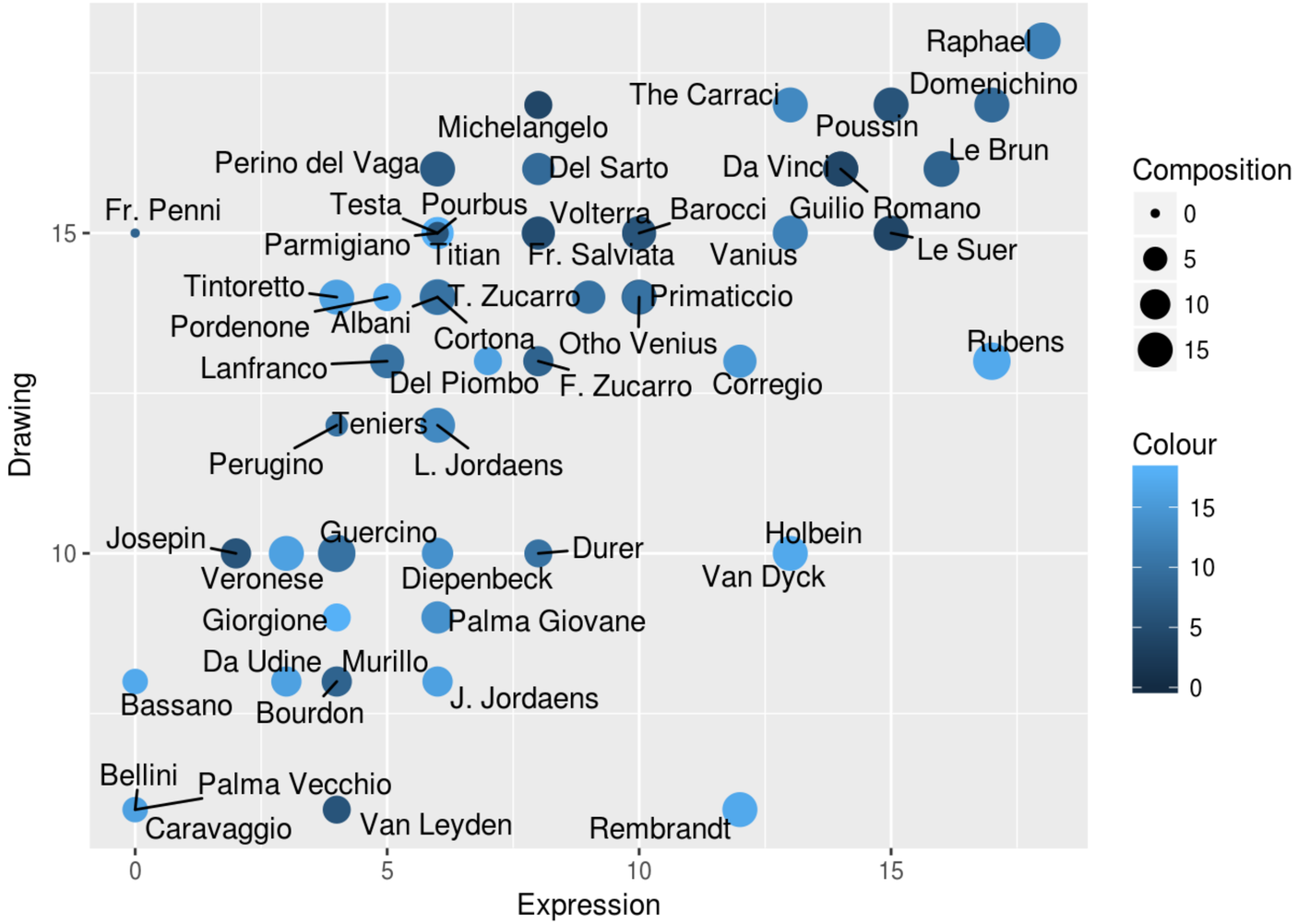

Parallel coordinates - Score for each painter
value
15
10
5
0
Composition
Drawing
Colour
Expression
variable
School
A
B
C
D
E
F
G
H


54Rubens
53Raphael
52The Carraci
51Domenichino
50Le Brun
49Van Dyck
48Vanius
47Poussin
46Corregio
45Titian
44Rembrandt
43Le Suer
42Tintoretto
41Guilio Romano
40Da Vinci
39Holbein
38Cortona
37Otho Venius
36Teniers
35T. Zucarro
34Primaticcio
33Barocci
32Del Sarto
31Albani
30Veronese
29Pordenone
28Fr. Salviata
27Perino del Vaga
26Del Piombo
25Lanfranco
24Guercino
23Diepenbeck
22Palma Giovane
21J. Jordaens
20L. Jordaens
19Volterra
18Giorgione
17F. Zucarro
16Parmigiano
15Michelangelo
14Da Udine
13Durer
12Murillo
11Testa
10Pourbus
09Bassano
08Bourdon
07Perugino
06Caravaggio
05Josepin
04Palma Vecchio
03Van Leyden
02Bellini
01Fr. Penni
0
20
40
60
value
variable
Composition
Drawing
Colour
Expression

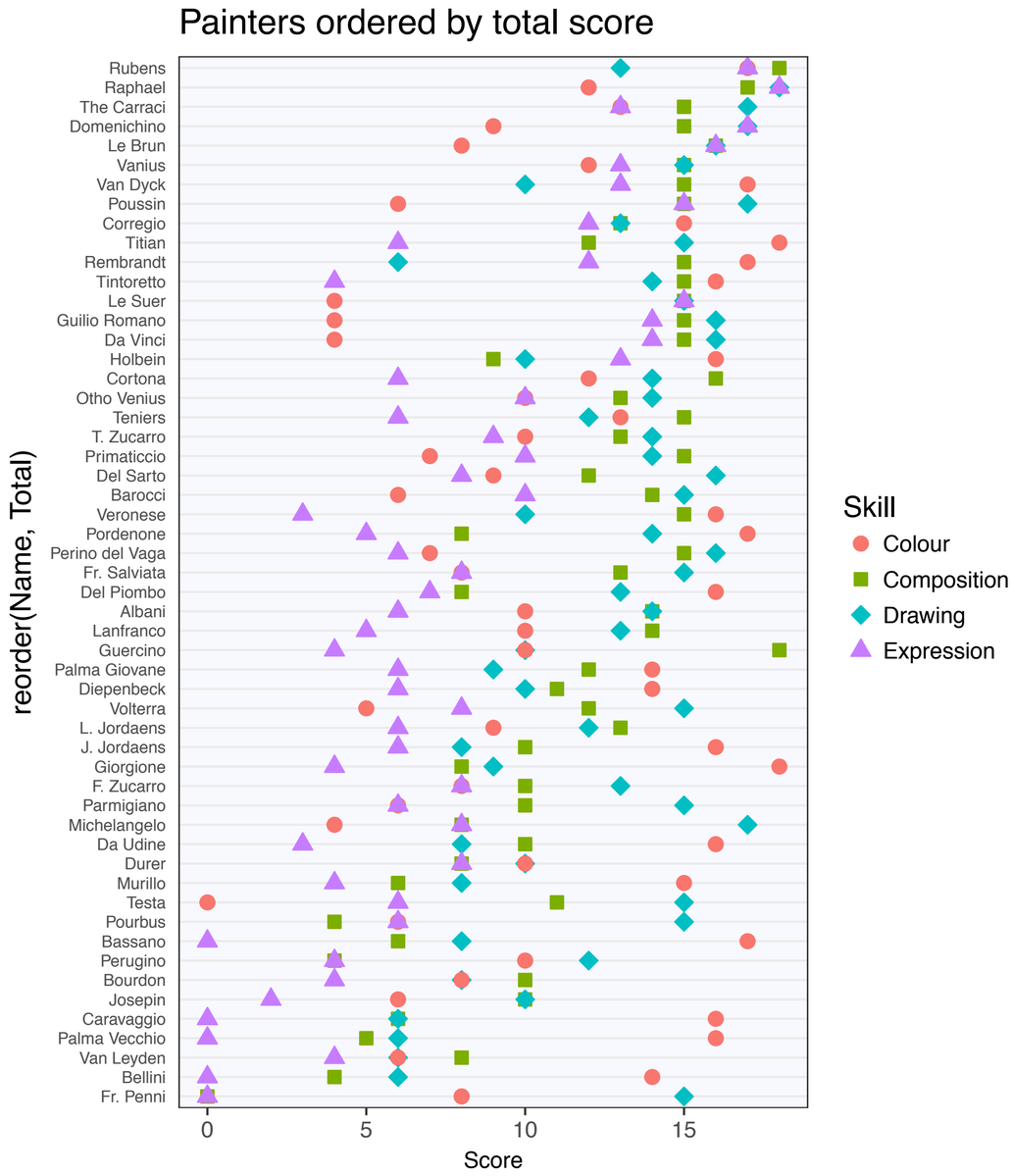
```r
data(painters)
library(pheatmap)
pheatmap(painters[,1:4], fontsize_row = 5.5, cluster_cols = F)
```

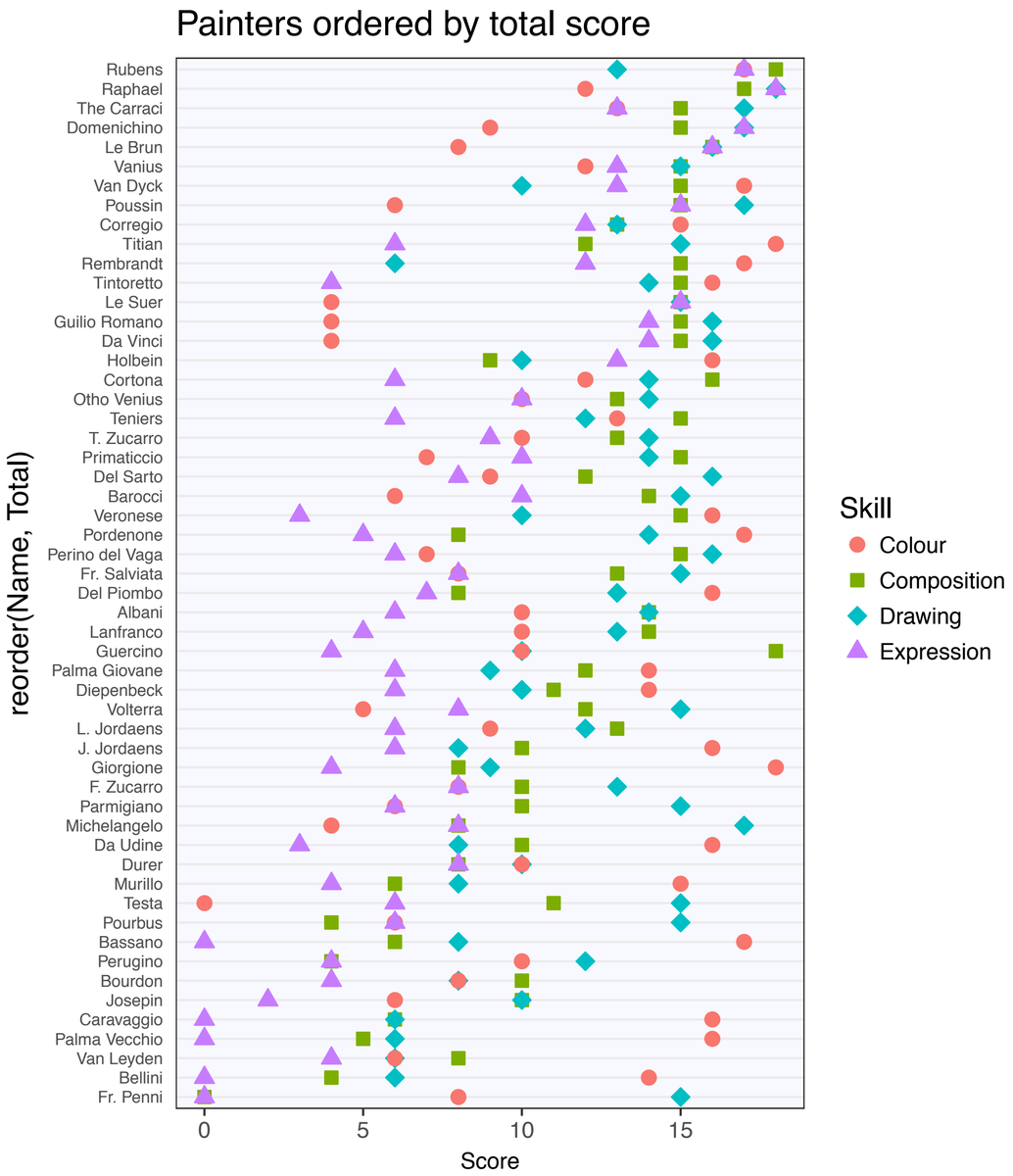


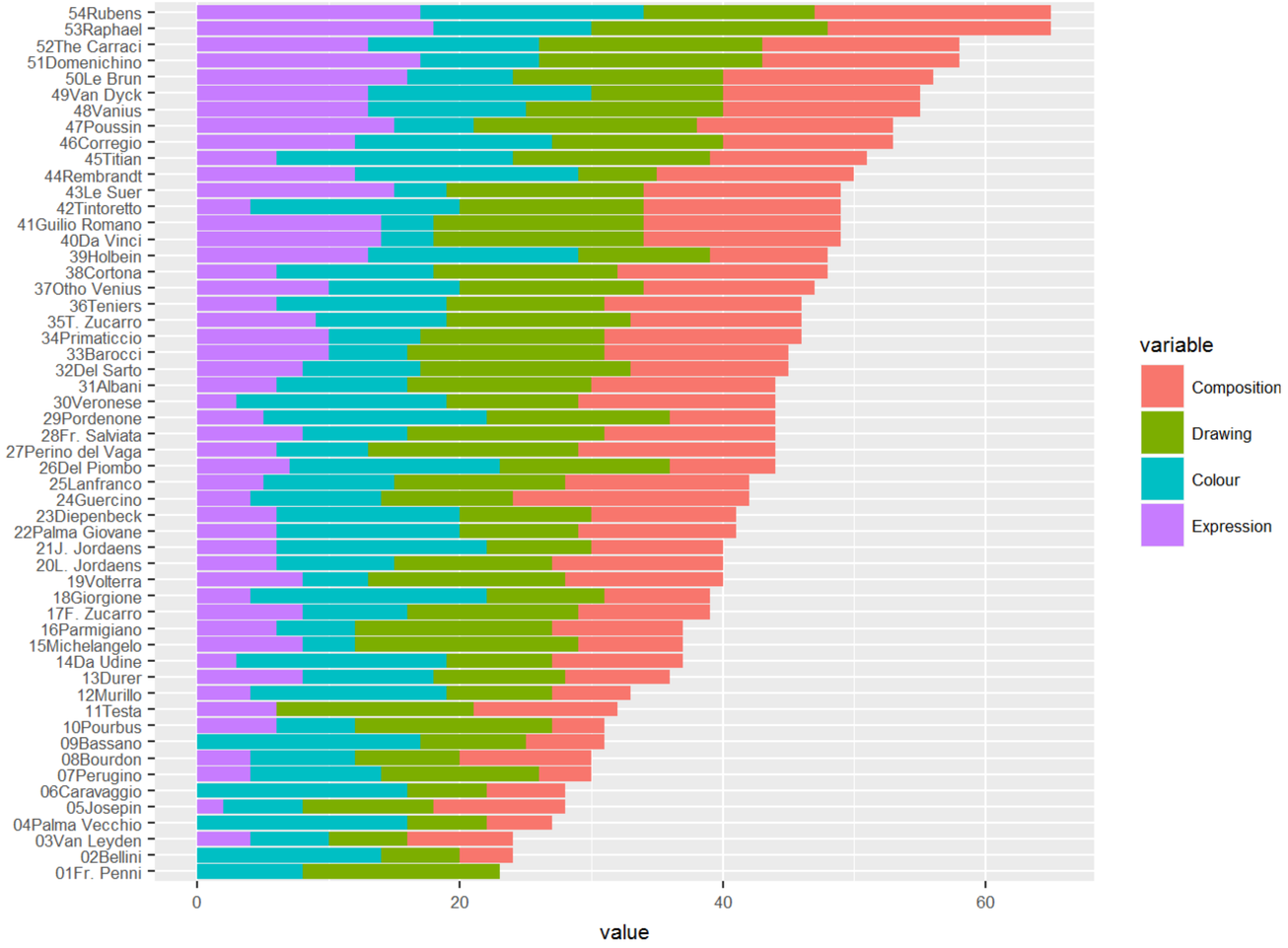
```
ggplot(painters, aes(Expression, Drawing)) +
    geom_point(aes(size = Composition, color = Colour)) +
    geom_text_repel(aes(label = rownames(painters)))
```



```
g +ggtitle("Painters ordered by total score")
```



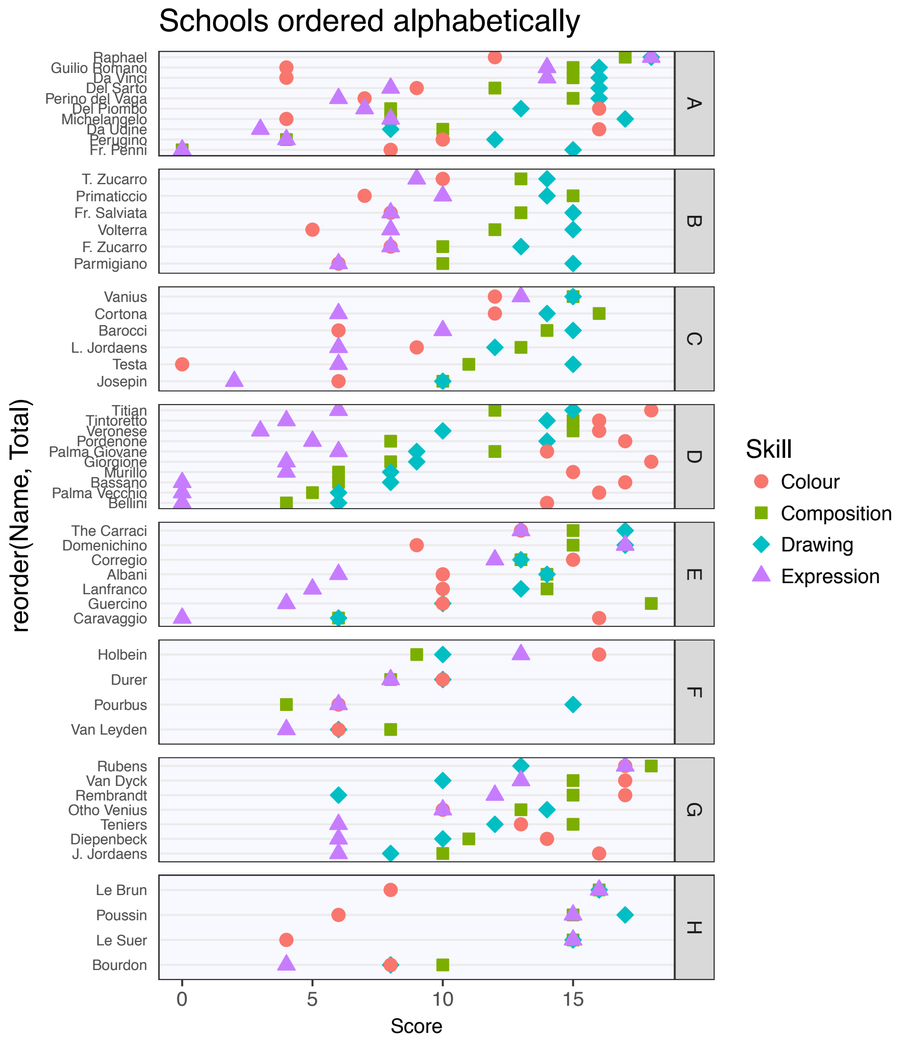

Painters ordered by total score
reorder(Name, Total)
Rubens
Raphael
The Carraci
Domenichino
Le Brun
Vanius
Van Dyck
Poussin
Corregio
Titian
Rembrandt
Tintoretto
Le Suer
Guilio Romano
Da Vinci
Holbein
Cortona
Otho Venius
Teniers
T. Zucarro
Primaticcio
Del Sarto
Barocci
Veronese
Pordenone
Perino del Vaga
Fr. Salviata
Del Piombo
Albani
Lanfranco
Guercino
Palma Giovane
Diepenbeck
Volterra
L. Jordaens
J. Jordaens
Giorgione
F. Zucarro
Parmigiano
Michelangelo
Da Udine
Durer
Murillo
Testa
Pourbus
Bassano
Perugino
Bourdon
Josepin
Caravaggio
Palma Vecchio
Van Leyden
Bellini
Fr. Penni
0
5
10
15
Score
Skill
Colour
Composition
Drawing
Expression


54Rubens
53Raphael
52The Carraci
51Domenichino
50Le Brun
49Van Dyck
48Vanius
47Poussin
46Corregio
45Titian
44Rembrandt
43Le Suer
42Tintoretto
41Guilio Romano
40Da Vinci
39Holbein
38Cortona
37Otho Venius
36Teniers
35T. Zucarro
34Primaticcio
33Barocci
32Del Sarto
31Albani
30Veronese
29Pordenone
28Fr. Salviata
27Perino del Vaga
26Del Piombo
25Lanfranco
24Guercino
23Diepenbeck
22Palma Giovane
21J. Jordaens
20L. Jordaens
19Volterra
18Giorgione
17F. Zucarro
16Parmigiano
15Michelangelo
14Da Udine
13Durer
12Murillo
11Testa
10Pourbus
09Bassano
08Bourdon
07Perugino
06Caravaggio
05Josepin
04Palma Vecchio
03Van Leyden
02Bellini
01Fr. Penni
0
20
40
60
value
variable
Composition
Drawing
Colour
Expression

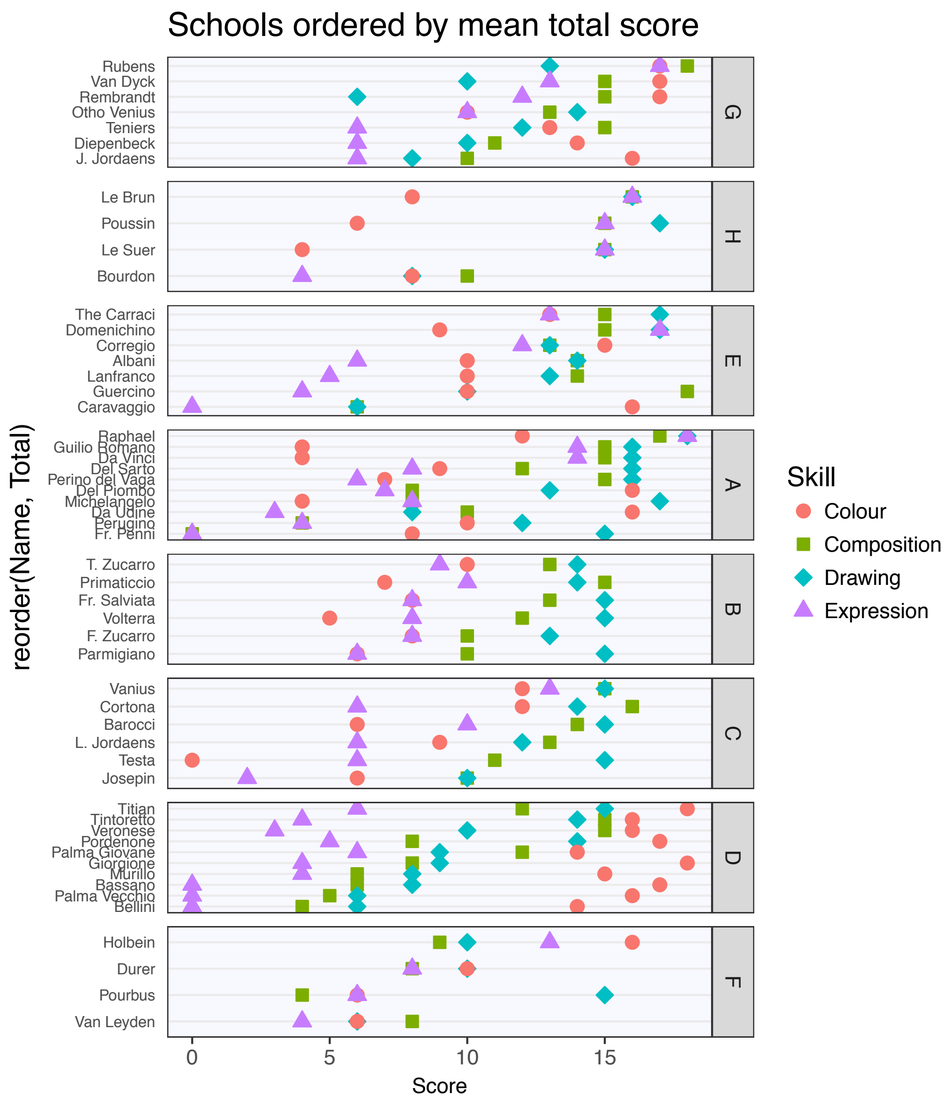
```
library(MASS)
library(tidyverse)

tidypaint <- painters %>% rownames_to_column("Name") %>%
    mutate(Total = Composition + Drawing + Colour +
                   Expression) %>%
    gather(key = Skill, value = Score, -Name, -School, -Total)

g <- ggplot(tidypaint, aes(x = Score, y = reorder(Name, Total),
                           color = Skill, shape = Skill,
                           fill = Skill)) +
    geom_point(size = 3) +
    scale_shape_manual(values = c(21, 22, 23, 24)) +
    theme_dotplot + theme(panel.background =
              element_rect(fill = "ghostwhite"))
```

```
g1 + ggtitle("Open shapes")
```

```
g + facet_wrap(~School, ncol = 1, scales = "free_y",
               strip.position = "right") +
    ggtitle("Schools ordered alphabetically")
```



```
g2 + ggtitle("Schools ordered by mean total score")
```

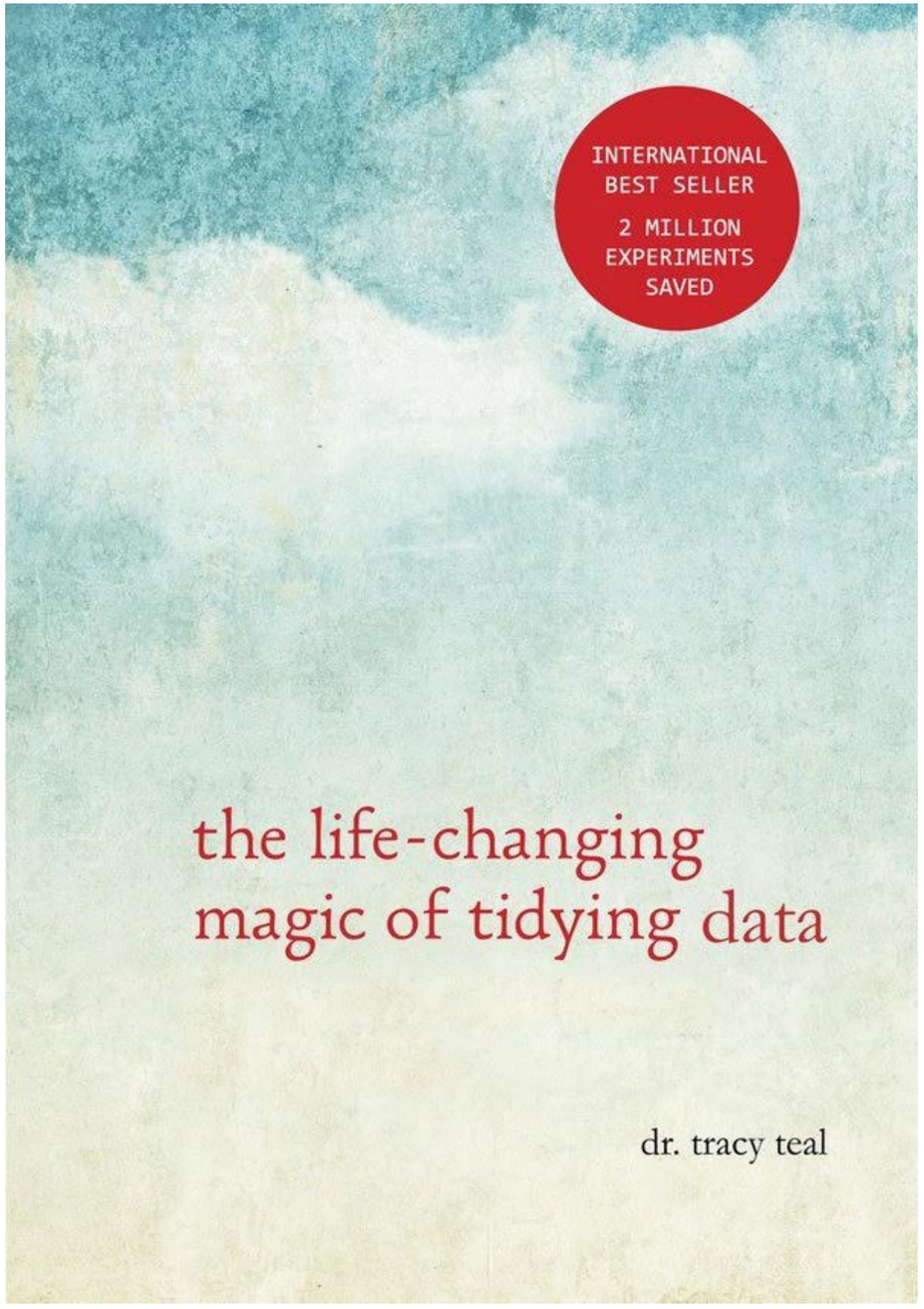


```
schoolorder <- tidypaint %>% group_by(School) %>%
    summarize(mean = mean(Total)) %>% arrange(desc(mean))
tidypaint$School <- factor(tidypaint$School, levels =
                                 schoolorder$School)
g2 <- ggplot(tidypaint, aes(x = Score, y = reorder(Name, Total),
                        color = Skill, shape = Skill,
                        fill = Skill)) +
    geom_point(size = 3) +
    scale_shape_manual(values = c(21, 22, 23, 24)) +
    facet_wrap(~School, ncol = 1, scales = "free_y",
                strip.position = "right") +
    theme_dotplot + theme(panel.background =
                element_rect(fill = "ghostwhite"))
```

Tidy Data

"Happy families are all alike; every unhappy family is unhappy in its own way."
--Leo Tolstoy

"Tidy datasets are all alike but every messy dataset is messy in its own way."
--Hadley Wickham


INTERNATIONAL
BEST SELLER
2 MILLION
EXPERIMENTS
SAVED
the life-changing
magic of tidying data
dr. tracy teal

Messy 1

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Messy 2

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Tidy

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Messy or Tidy?

Da Vinci	15	16	4	14	A
Del Piombo	8	13	16	7	A
Del Sarto	12	16	9	8	A
Fr. Penni	0	15	8	0	A
Guilio Romano	15	16	4	14	A
Michelangelo	8	17	4	8	A
Perino del Vaga	15	16	7	6	A
Perugino	4	12	10	4	A
Raphael	17	18	12	18	A
F. Zucarro	10	13	8	8	B
Fr. Salviata	13	15	8	8	B
Parmigiano	10	15	6	6	B
Primaticcio	15	14	7	10	B
T. Zucarro	13	14	10	9	B

tidy definition:

1 variable per column

1 observation per row

Messy or Tidy?

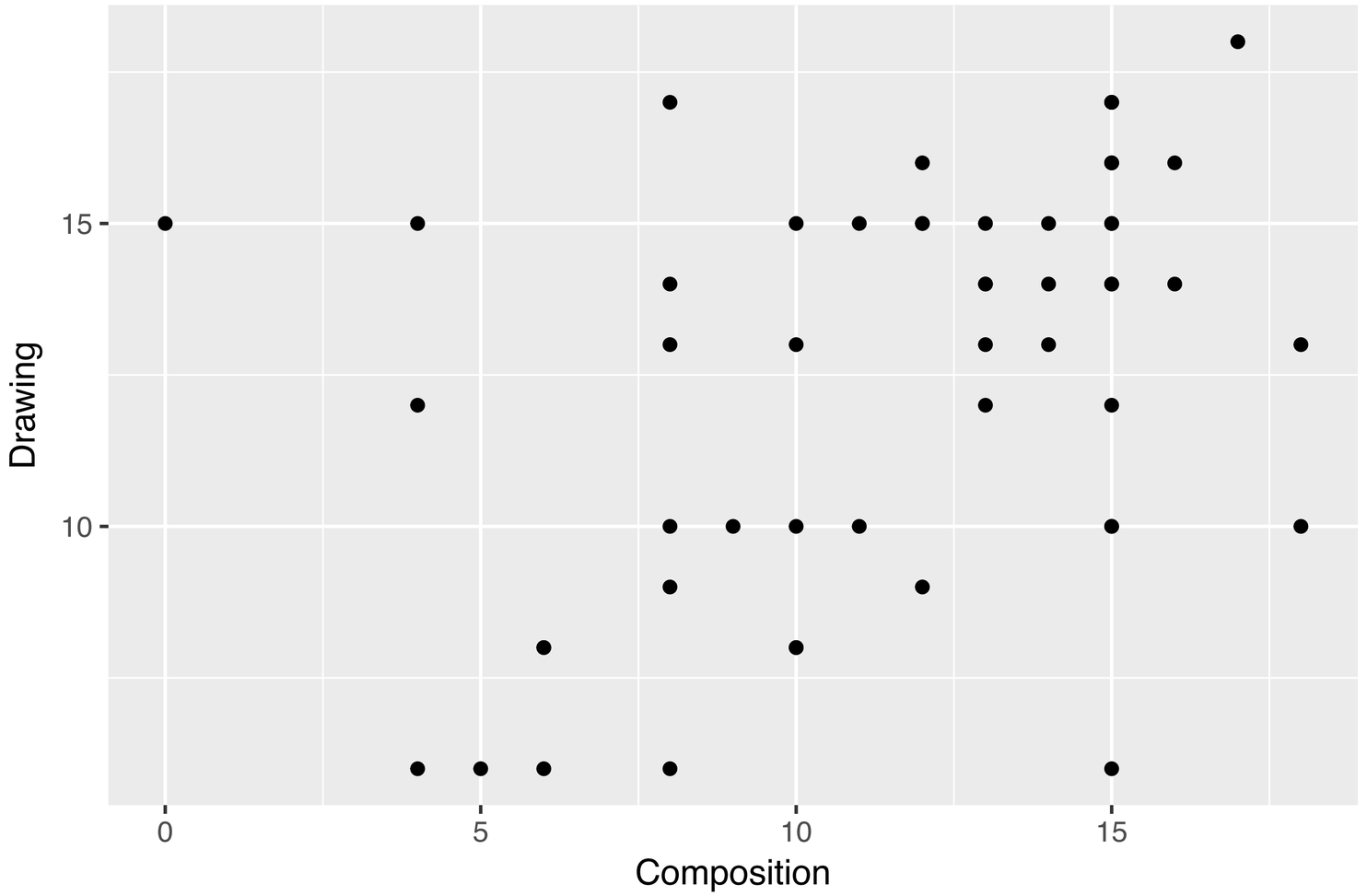
	Composition	Drawing	Colour	Expression	School
Da Udine	10	8	16	3	A
Da Vinci	15	16	4	14	A
Del Piombo	8	13	16	7	A
Del Sarto	12	16	9	8	A
Fr. Penni	0	15	8	0	A
Guilio Romano	15	16	4	14	A
Michelangelo	8	17	4	8	A
Perino del Vaga	15	16	7	6	A
Perugino	4	12	10	4	A
Raphael	17	18	12	18	A
F. Zucarro	10	13	8	8	B
Fr. Salviata	13	15	8	8	B
Parmigiano	10	15	6	6	B

tidy definition:

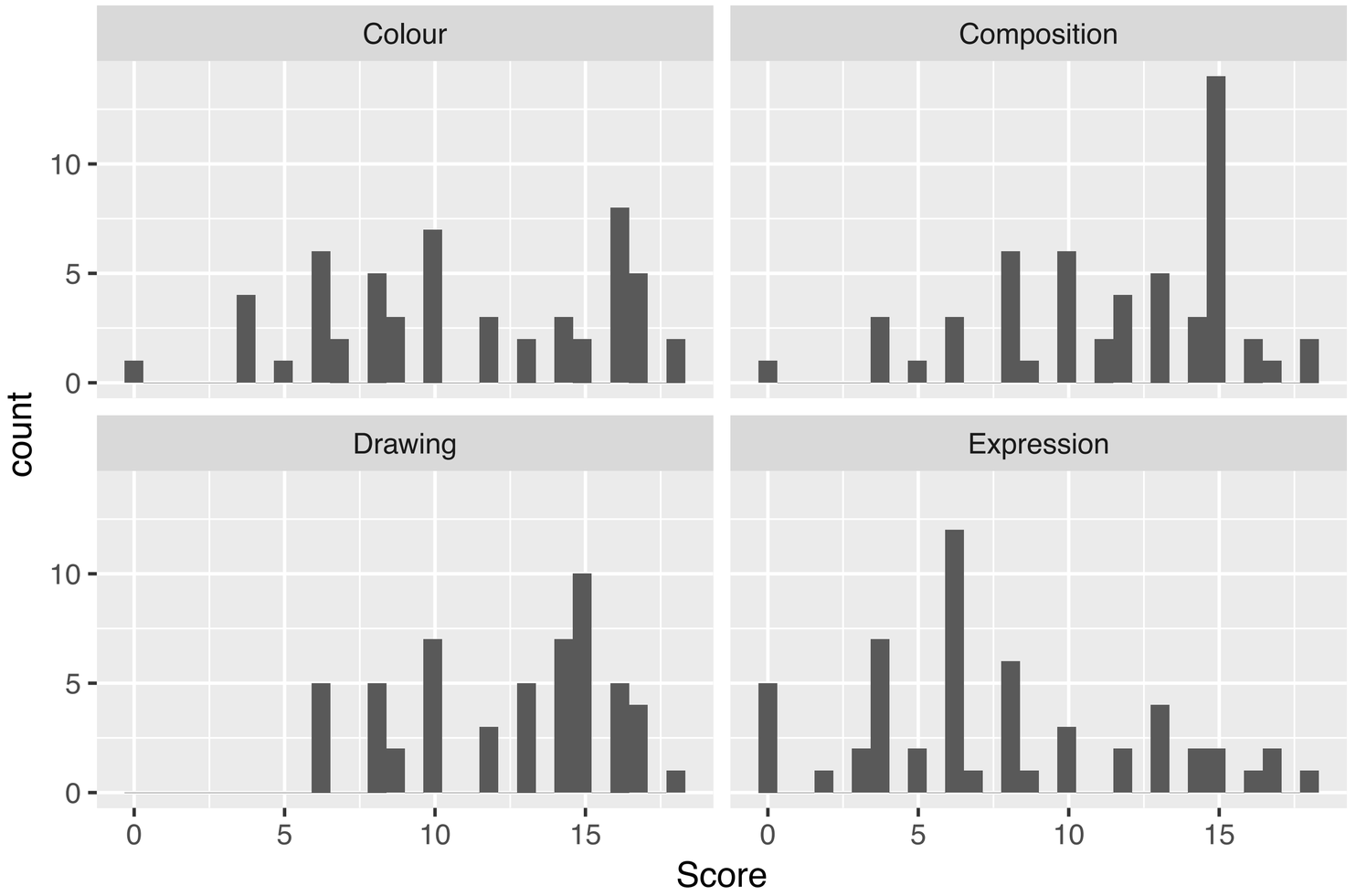
1 variable per column

1 observation per row

```
ggplot(painters, aes(Composition, Drawing)) + geom_point()
```



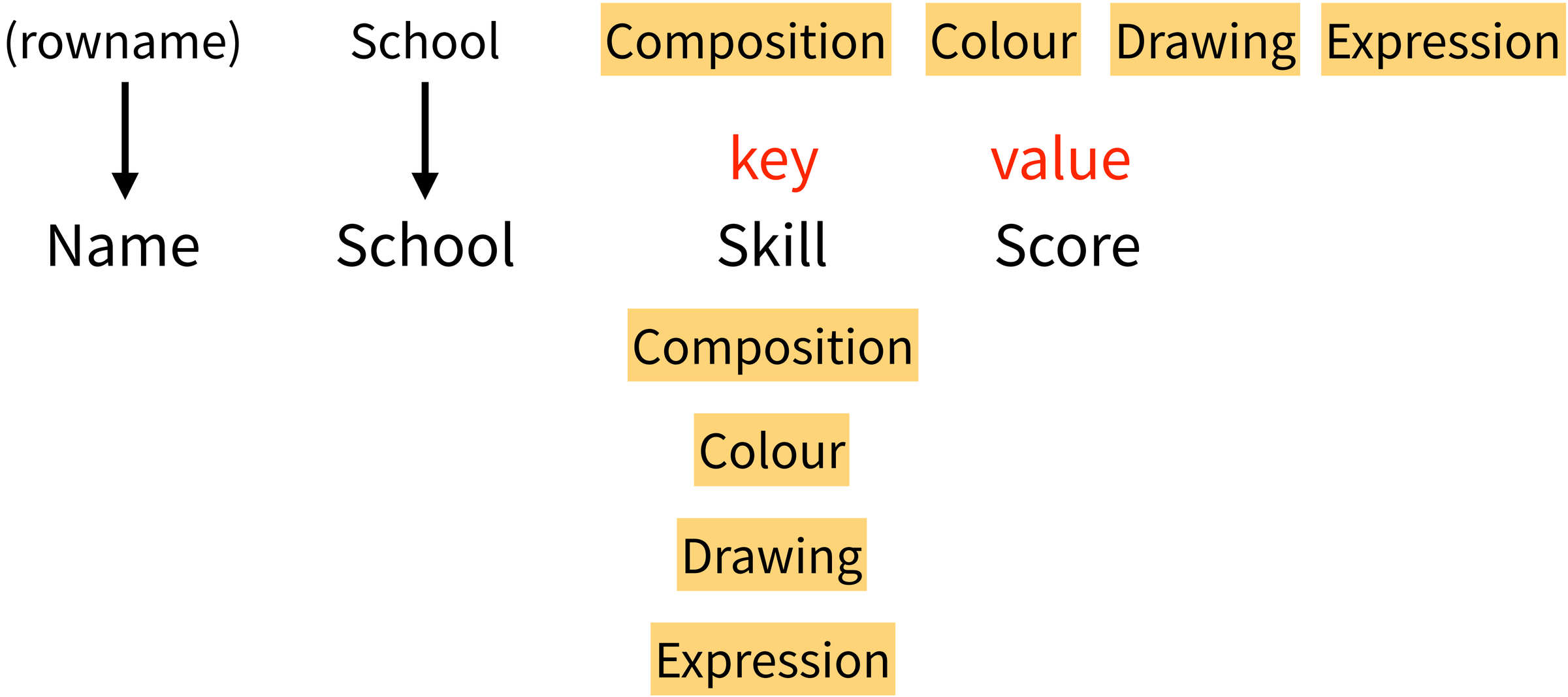
g3


Colour
Composition
Drawing
Expression
count
Score
0
5
10
15

```
library(tidyverse)
library(MASS)
tidypaint <- painters %>% rownames_to_column("Name") %>%
    gather(key = Skill, value = Score, -Name, -School)

g3 <- ggplot(tidypaint, aes(x =Score)) + geom_histogram() +
    facet_wrap(~Skill)
```

~~[,] $ subset cbind rbind~~

gather




```
tidypaint <- painters %>% rownames_to_column("Name") %>%
    gather(key = Skill, value = Score, -Name, -School)
```

Coding

- rownames
- %>% "pipe" (magrittr, dplyr)

Aesthetic specifications

Hadley Wickham

2016-03-01

This vignette summarises the various formats that grid drawing functions take. Most of this information is available scattered throughout the R documentation. This appendix brings it all together in one place.

Colour

Colours can be specified with:

- A **name**, e.g., `"red"`. R has 657 built-in named colours, which can be listed with `colours()`. The Stowers Institute provides a nice printable pdf that lists all colours: http://research.stowers-institute.org/efg/R/Color/Chart/.
- An **rgb specification**, with a string of the form `"#RRGGBB"` where each of the pairs `RR`, `GG`, `BB` consists of two hexadecimal digits giving a value in the range `00` to `FF`

  You can optionally make the colour transparent by using the form `"#RRGGBBAA"`.
- An **NA**, for a completely transparent colour.
- The munsell package, by Charlotte Wickham, provides a wrapper around the colour system designed by

http://docs.ggplot2.org/current/vignettes/ggplot2-specs.html