

Tidying Data

Tidy data

Hadley Wickham, "Tidy Data", *Journal of Statistical Software* 59 (10), August 2014.

- <http://vita.had.co.nz/papers/tidy-data.pdf>
- <https://www.jstatsoft.org/article/view/v059i10>

<https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>

Tidy data

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

Tidy data

Tidy datasets

messy dataset

~~Happy families~~ are all alike; every ~~unhappy family~~ is
~~unhappy~~ in its own way — Hadley Wickham
messy

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


Principles of tidy data

- Every column is a variable.
- Every row is an observation.
- Every cell is a single value.

messy

		id	city	hwy
1		car1	19	24
2		car2	20	30
3		car3	29	35

tidy

		id	roadtype	mpg
1		car1	city	19
2		car2	city	20
3		car3	city	29
4		car1	hwy	24
5		car2	hwy	30
6		car3	hwy	35

Step 1: identify "keep as is"

	id	city	hwy
1	car1	19	24
2	car2	20	30
3	car3	29	35

Step 2: create new columns

	id	roadtype	mpg
1	car1	city	19
2	car2	city	20
3	car3	city	29
4	car1	hwy	24
5	car2	hwy	30
6	car3	hwy	35

Everything else just happens:

messy column names are moved to "name" column

values are moved to a single "value" column

	id	city	hwy
1	car1	19	24
2	car2	20	30
3	car3	29	35

	id	roadtype	mpg
1	car1	city	19
2	car2	city	20
3	car3	city	29
4	car1	hwy	24
5	car2	hwy	30
6	car3	hwy	35

Step 1: identify "keep as is"

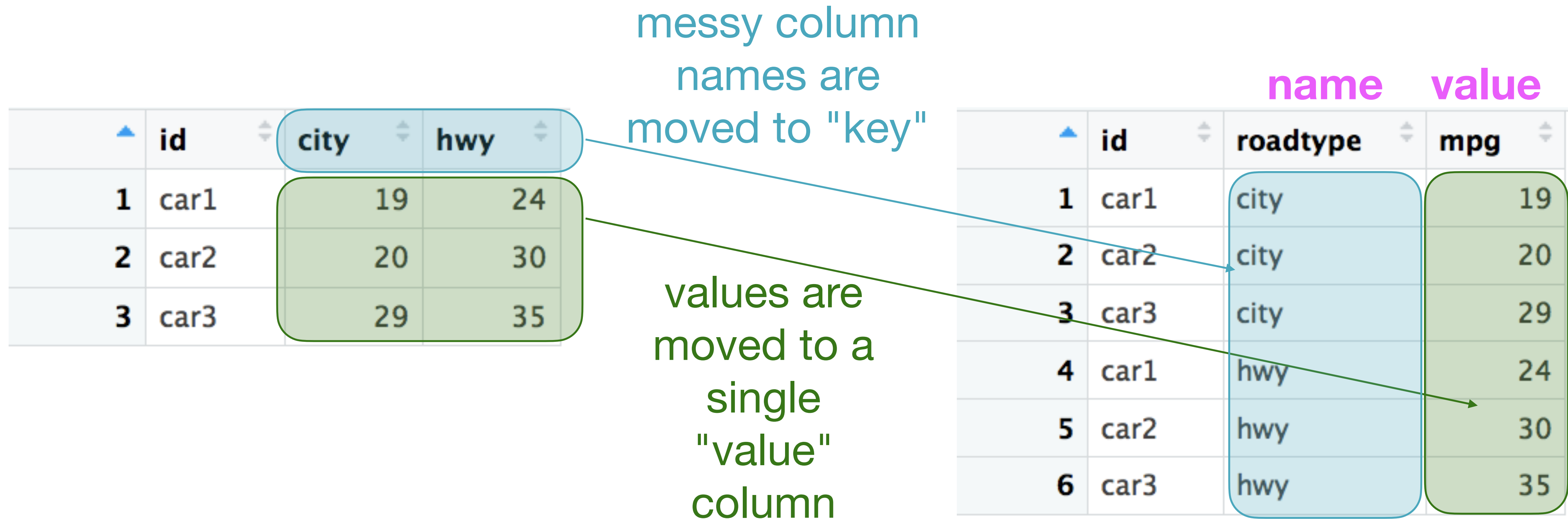
	id	city	hwy
1	car1	19	24
2	car2	20	30
3	car3	29	35

Step 2: create
new
columns

	id	roadtype	mpg
1	car1	city	19
2	car2	city	20
3	car3	city	29
4	car1	hwy	24
5	car2	hwy	30
6	car3	hwy	35

```
tidydata <- messydata %>%  
  pivot_longer(cols = !id, names_to = "roadtype",  
               values_to = "mpg")
```

Everything else just happens:



```
tidydata <- messydata %>%  
  pivot_longer(cols = !id, names_to = "roadtype",  
               values_to = "mpg")
```

If there is no id column

```
> head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

...create one

```
mtcars %>%  
  rownames_to_column("carname") %>%  
  head()
```

	carname	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
2	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
3	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
4	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
5	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
6	Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
7	Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4

painters (MASS package)

painters x						
← → 📄 🔍 Filter						
	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	

What should the new columns be?

Current columns:

Composition

Colour

Drawing

Expression

School

What should the new columns be?

Current columns:

Composition	Colour	Drawing	Expression	School
-------------	--------	---------	------------	--------

New columns:

		(name)	(value)
(rownames) →	Name	School	Skill Score

What should the new data look like?

Name	School	Skill	Score
Da Udine	A	Composition	10
Da Vinci	A	Composition	15
Del Piombo	A	Composition	8
Del Sarto	A	Composition	12
⋮	⋮	⋮	⋮

Tidyr code?

Current columns:

Composition Colour Drawing Expression School

New columns:

(rownames) → Name School ^(name) Skill ^(value) Score

```
tidypaint <- painters %>%  
  rownames_to_column("Name") %>%  
  pivot_longer(cols = !c(Name, School),  
               names_to = "Skill",  
               values_to = "Score")
```

*Columns
specified
match new
columns*

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

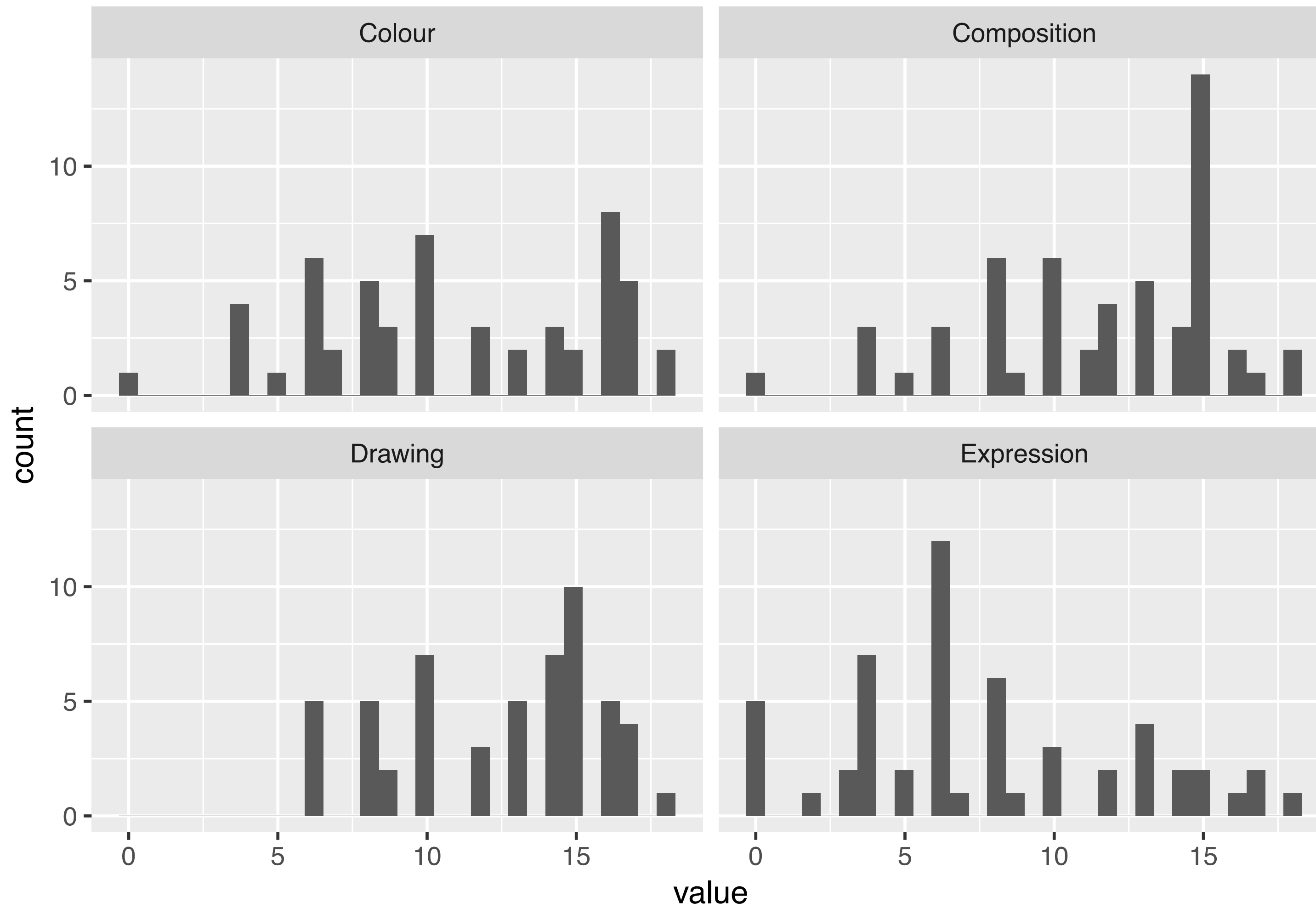
tidy definition:
1 variable per column
1 observation per row



... depends on the
use case

"Tidy" works better for some

```
ggplot(tidypaint, aes(Score)) + geom_histogram() +  
  facet_wrap(~School)
```



...and "Messy" for others

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

