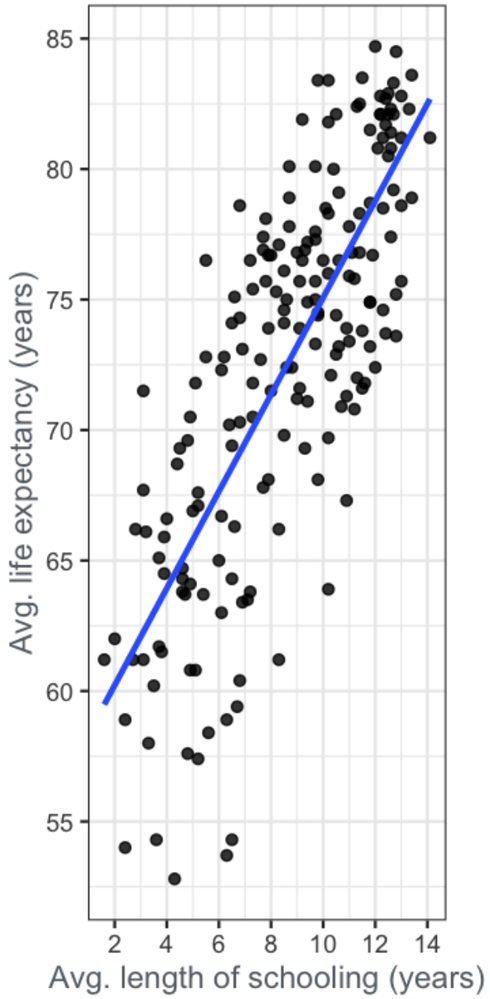
Higher dimensions

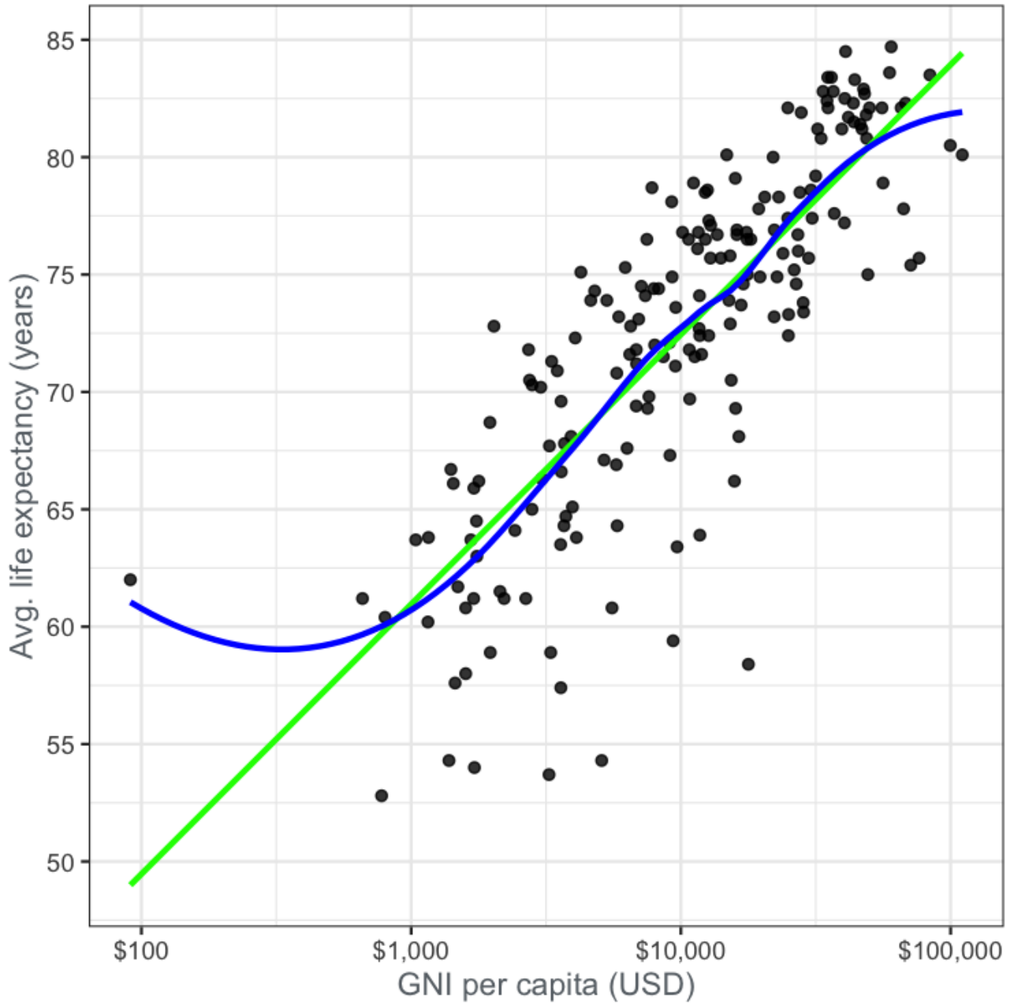
UNDERSTANDING DATA VISUALIZATION

Richie Cotton
Data Evangelist at DataCamp

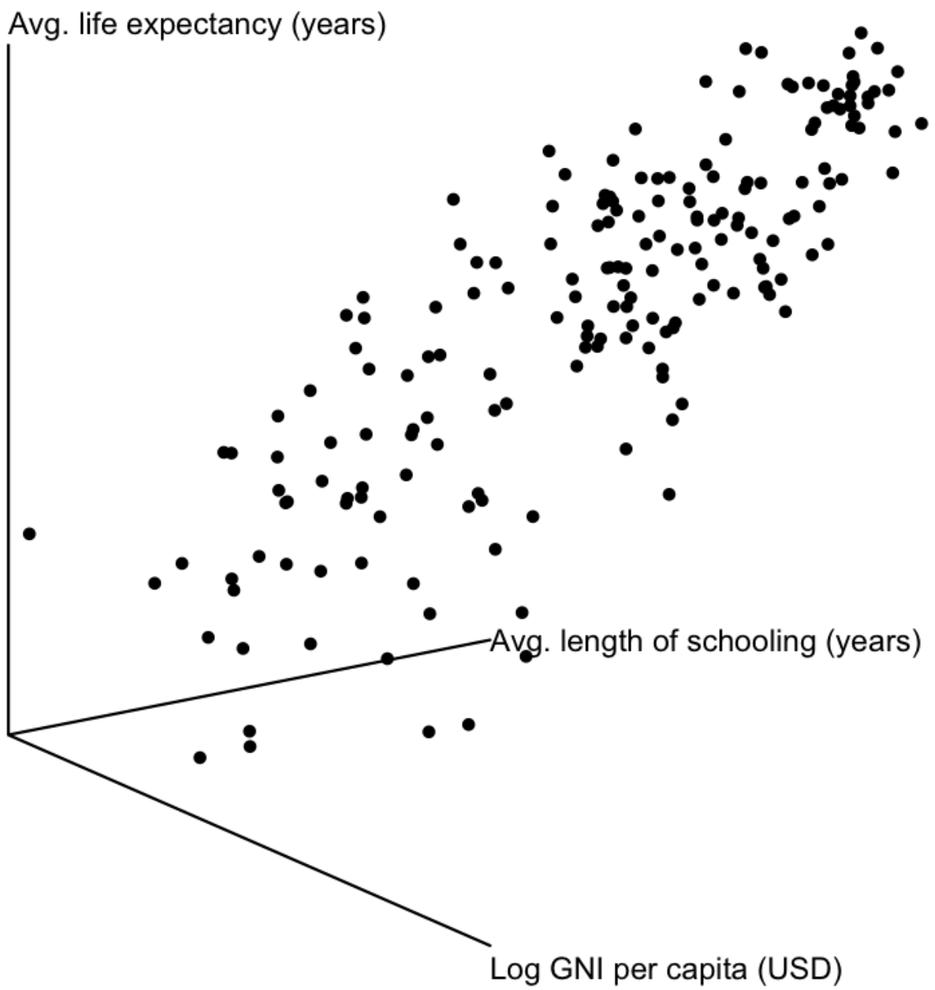
datacamp

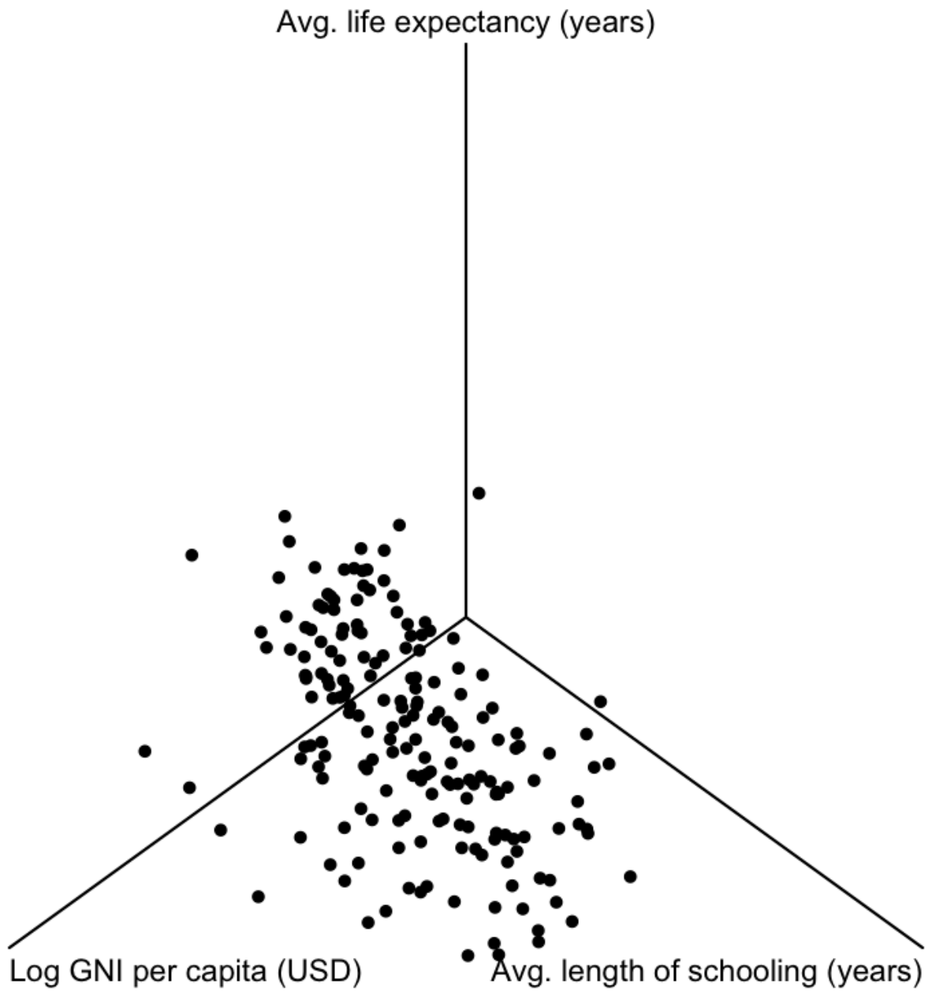
The UN life expectancy scatter plots





3D scatter plots



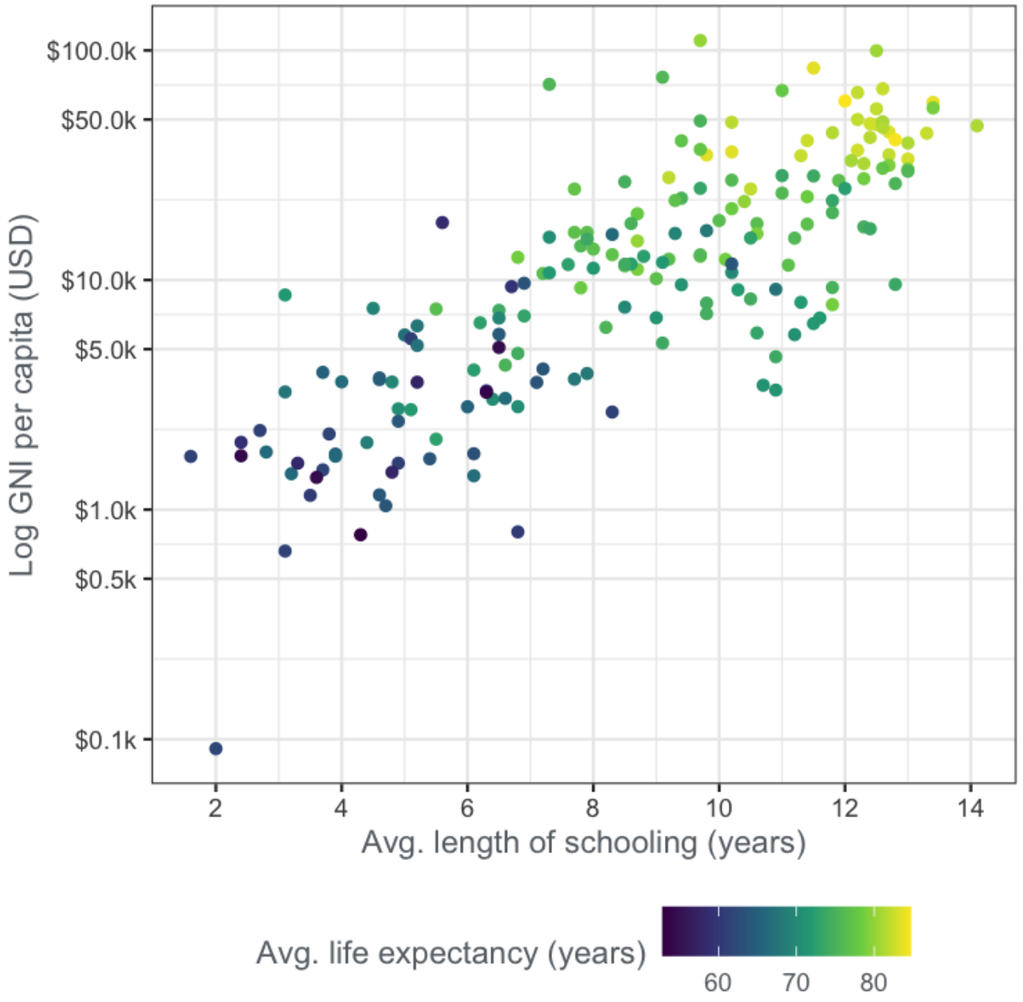


x and y are not the only dimensions

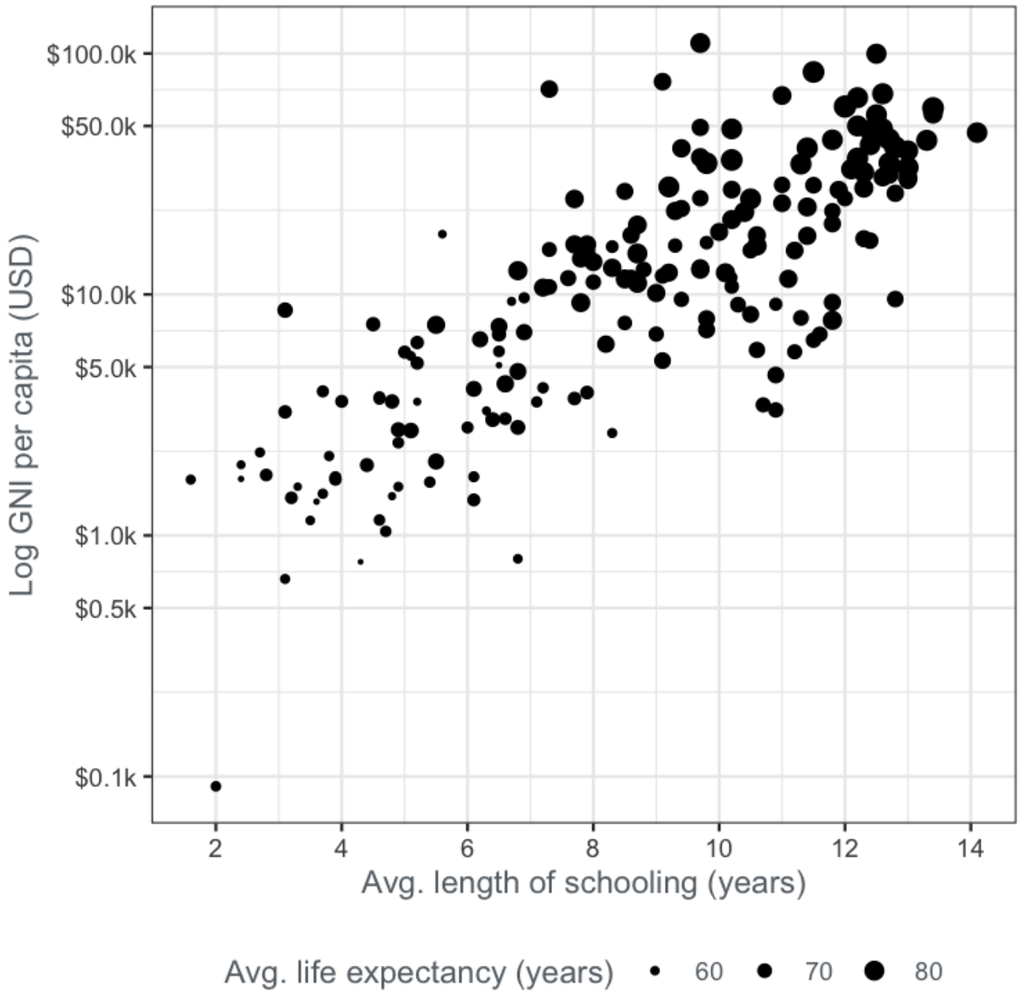
Points also have these dimensions

- color
- size
- transparency
- shape

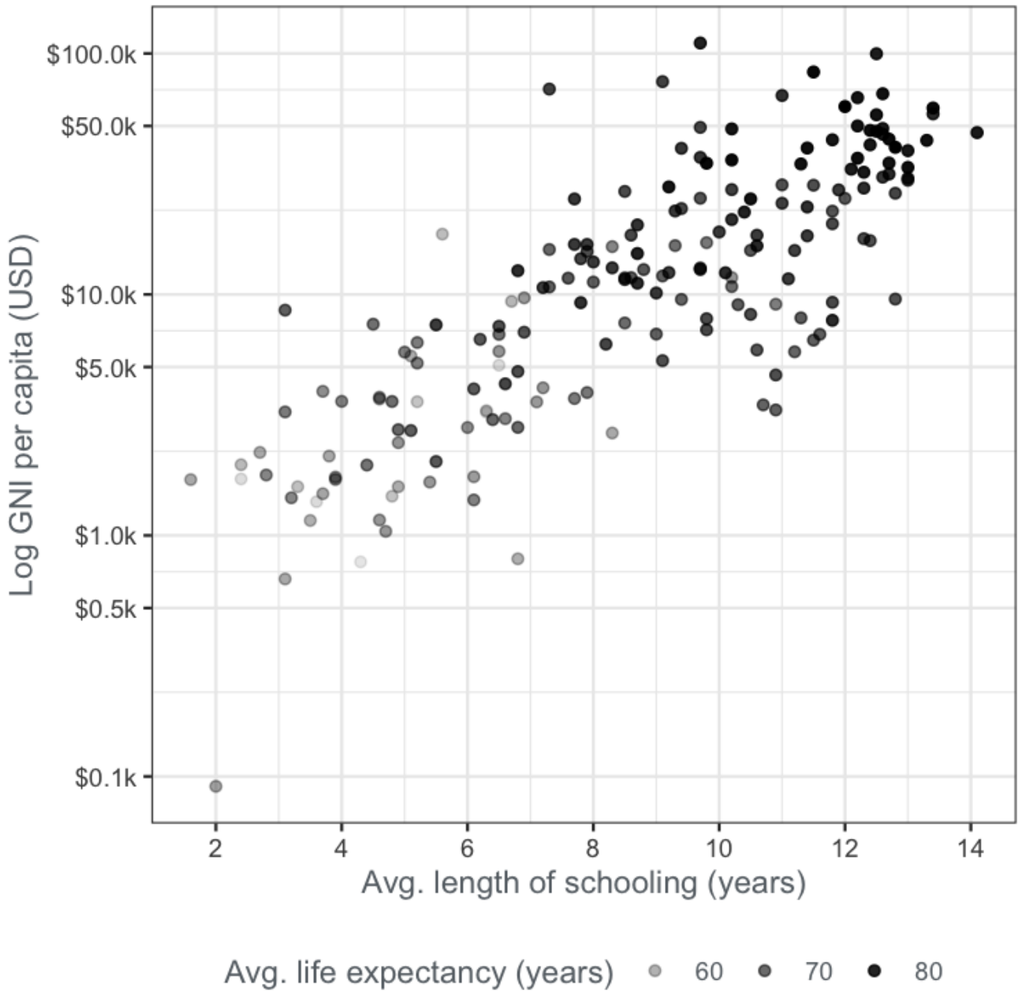
Color


$100.0k
$50.0k
$10.0k
$5.0k
$1.0k
$0.5k
$0.1k
Log GNI per capita (USD)
2
4
6
8
10
12
14
Avg. length of schooling (years)
Avg. life expectancy (years)
60
70
80

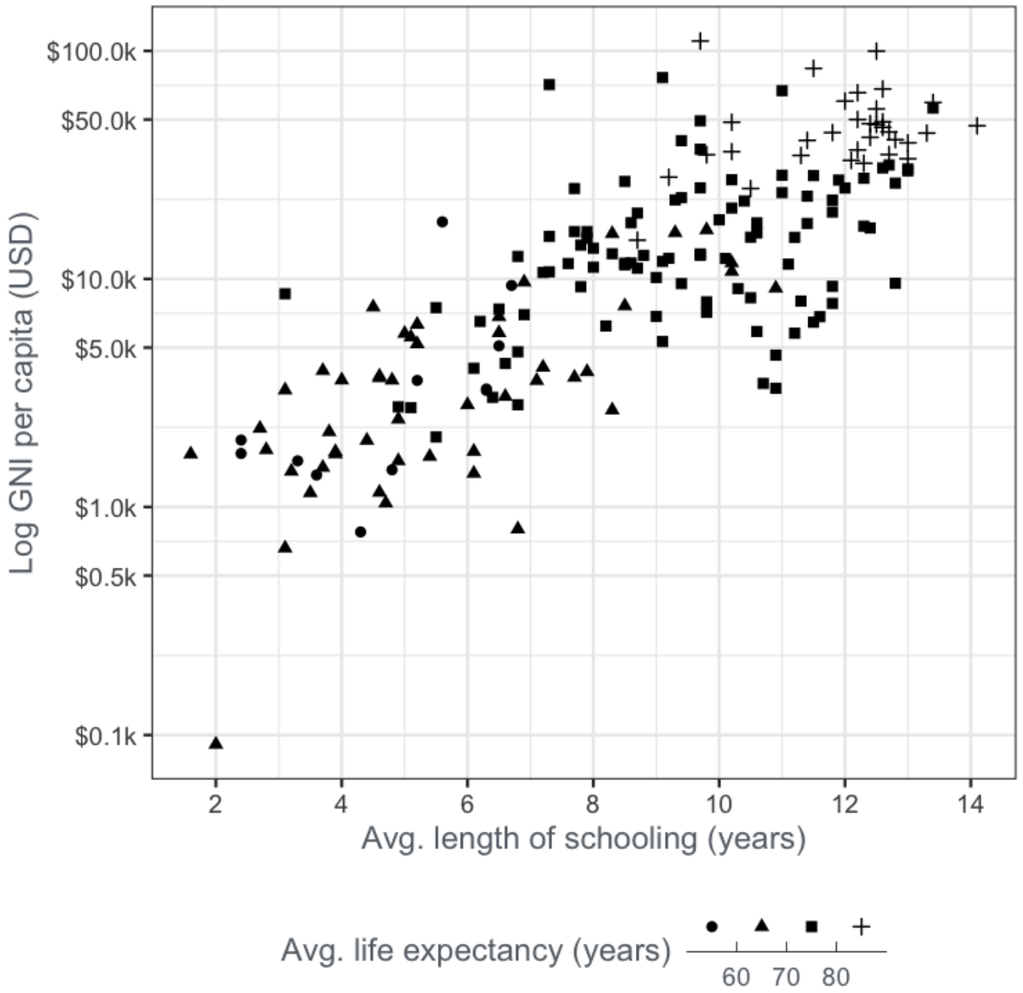
Size


$100.0k
$50.0k
$10.0k
$5.0k
$1.0k
$0.5k
$0.1k
Log GNI per capita (USD)
2
4
6
8
10
12
14
Avg. length of schooling (years)
Avg. life expectancy (years)
60
70
80

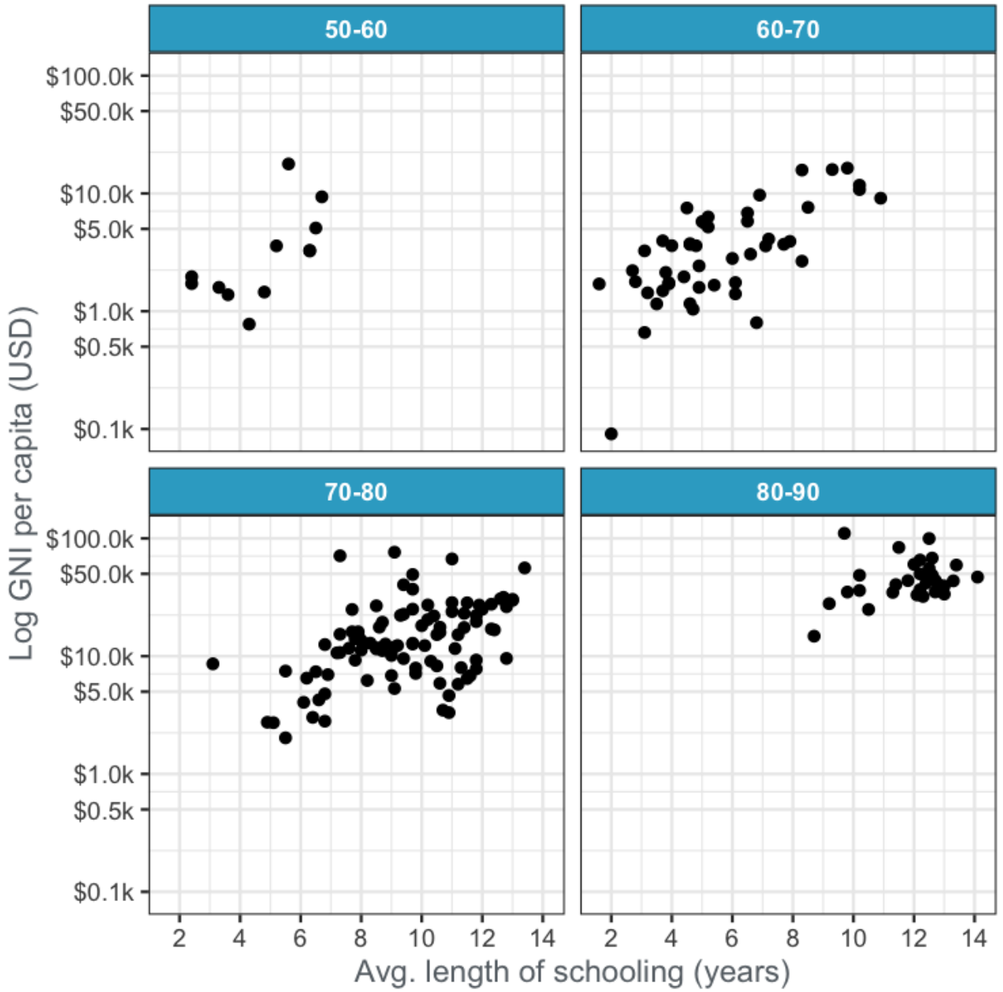
Transparency



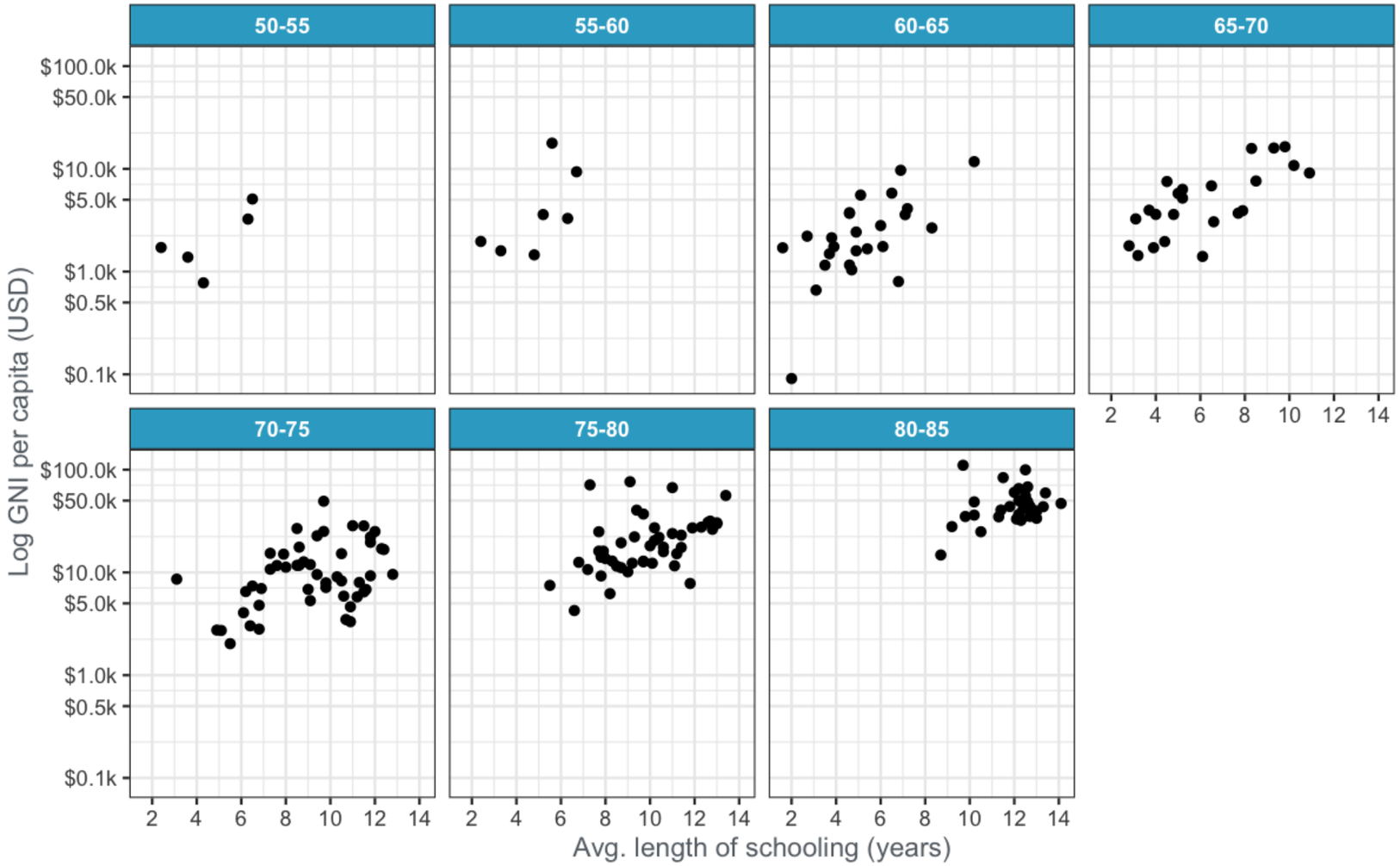
Shape


$100.0k
$50.0k
$10.0k
$5.0k
$1.0k
$0.5k
$0.1k
Log GNI per capita (USD)
2
4
6
8
10
12
14
Avg. length of schooling (years)
Avg. life expectancy (years)
60
70
80

Lots of panels



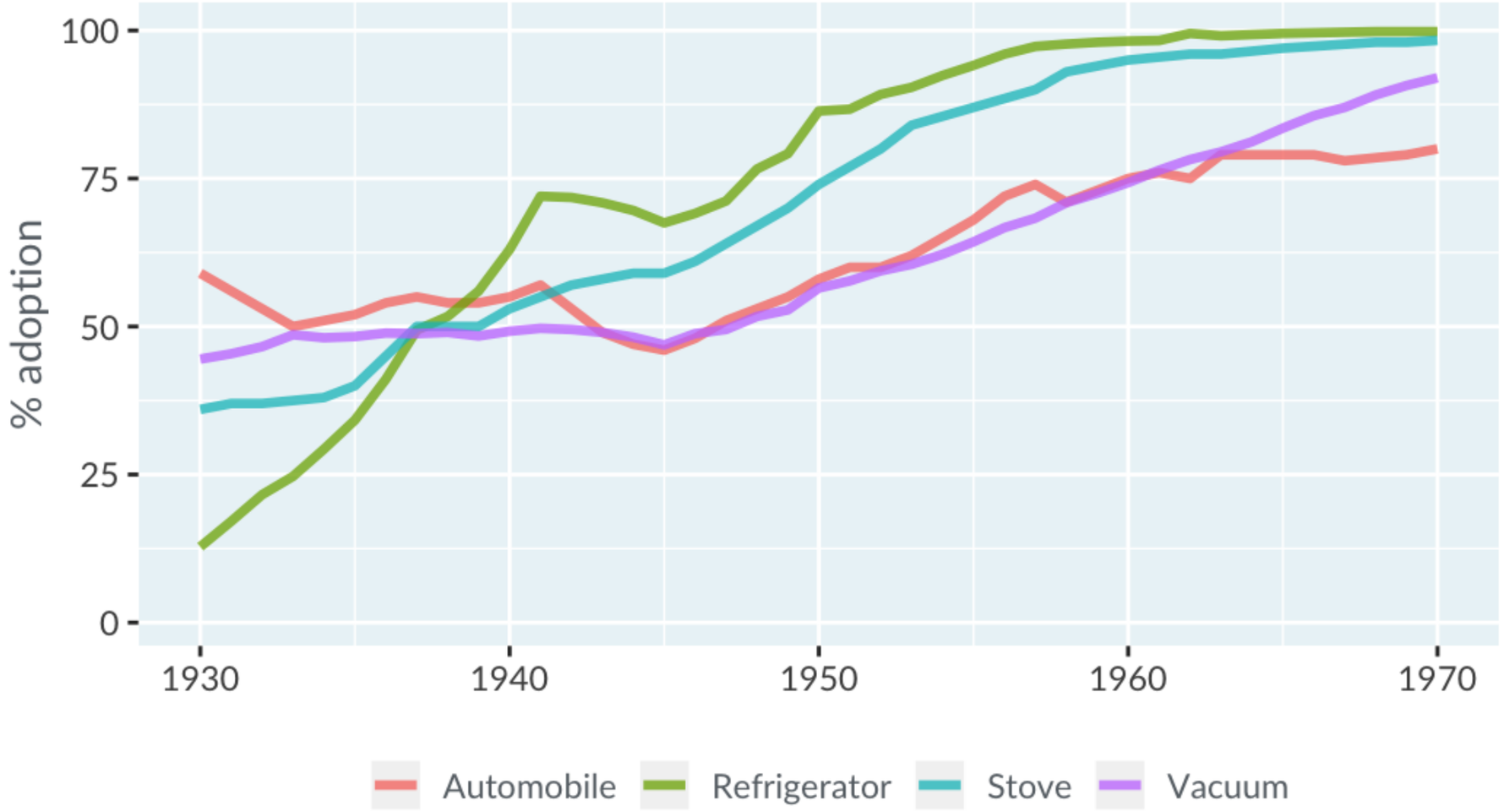
Even more panels



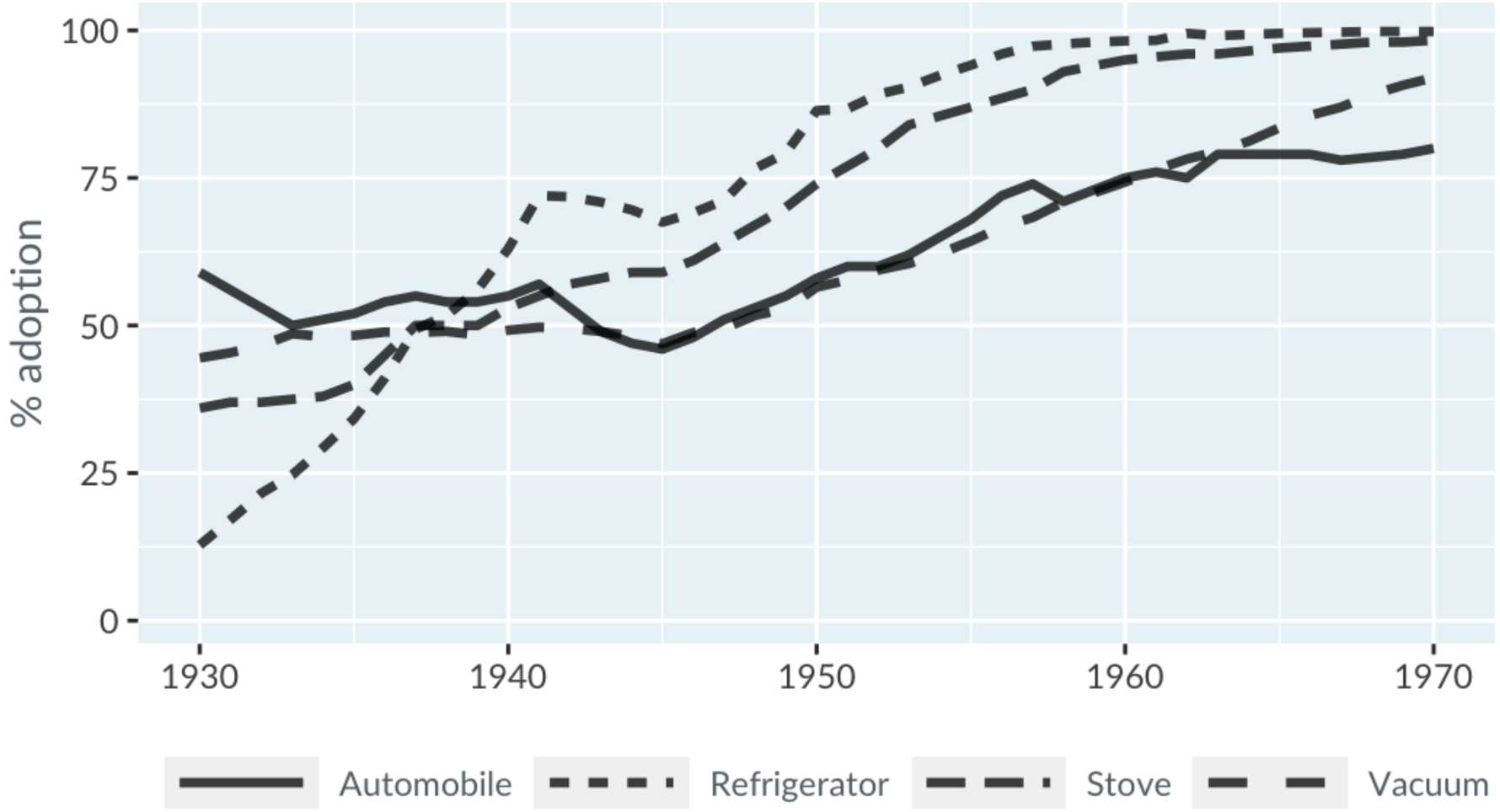
Other dimensions for line plots

- color
- thickness
- transparency
- line type (solid, dashes, dots)

Color


100
75
50
25
0
% adoption
1930
1940
1950
1960
1970
Automobile
Refrigerator
Stove
Vacuum

Linetype


100
75
50
25
0
% adoption
1930
1940
1950
1960
1970
Automobile
Refrigerator
Stove
Vacuum

Let's practice!

UNDERSTANDING DATA VISUALIZATION

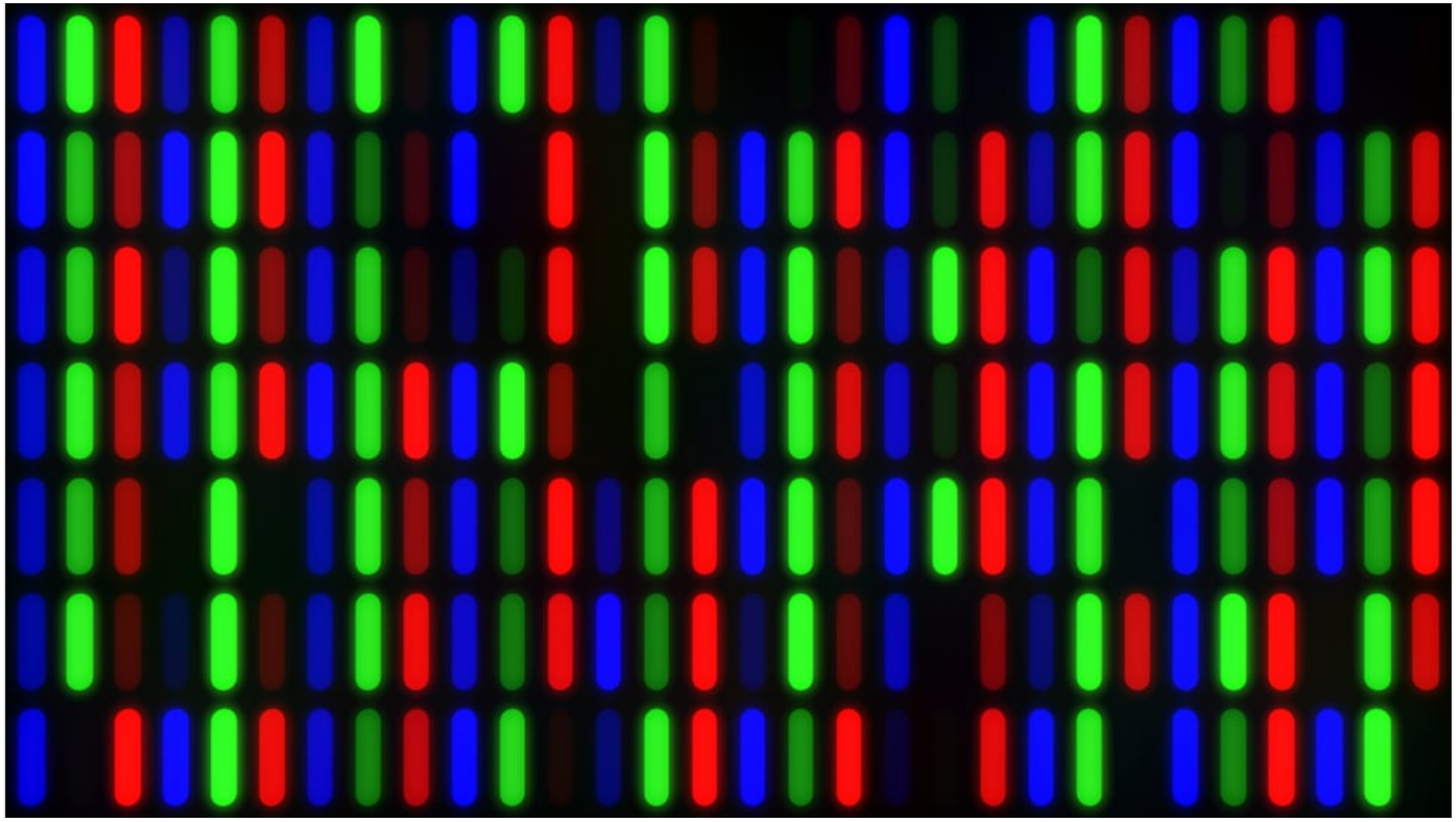
Using color

UNDERSTANDING DATA VISUALIZATION

Richie Cotton
Data Evangelist at DataCamp

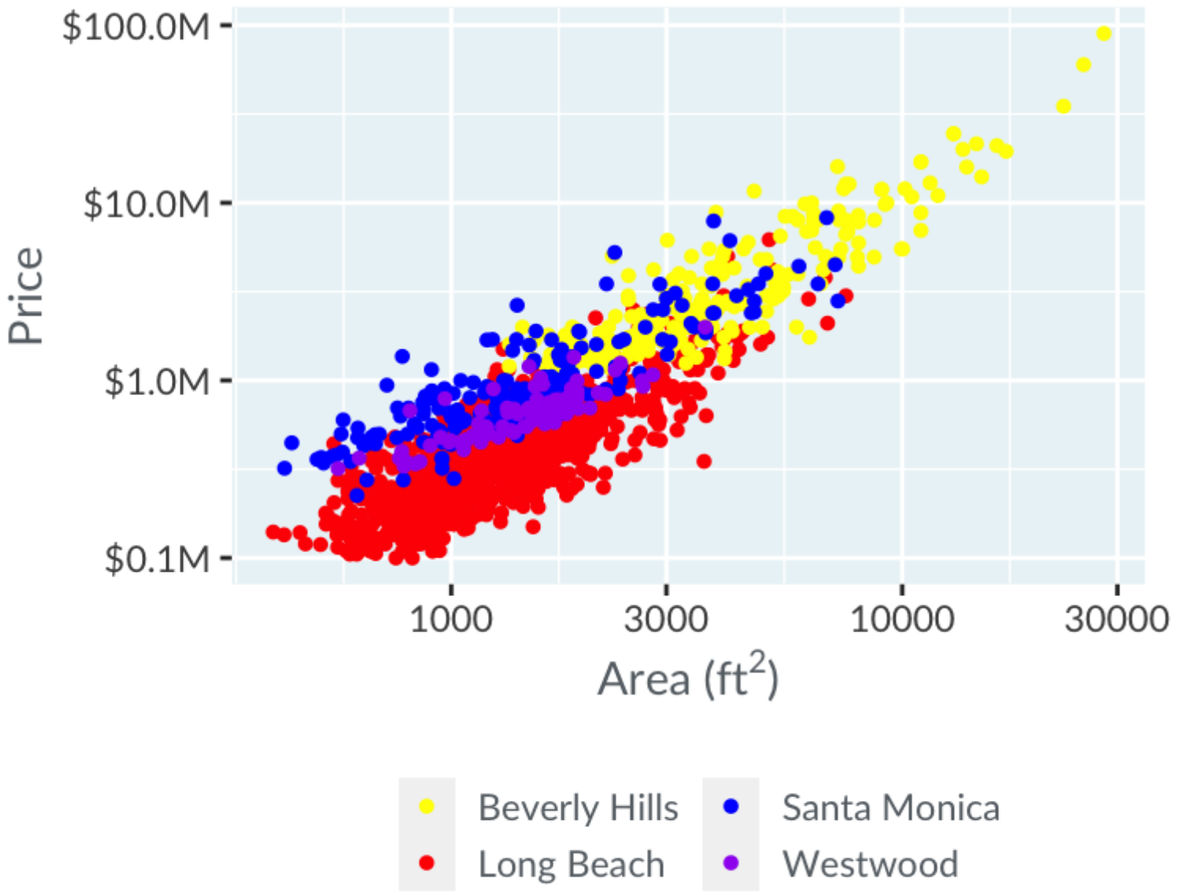
datacamp

Colorspaces: Red-Green-Blue



Colorspaces: Cyan-Magenta-Yellow-blacK

Choosing a plotting palette



- Usually, each color should stand out as much as other colors.
- The perceptual distance from one color in the plot to the next should be constant.

Colorspaces: Hue-Chroma-Luminance

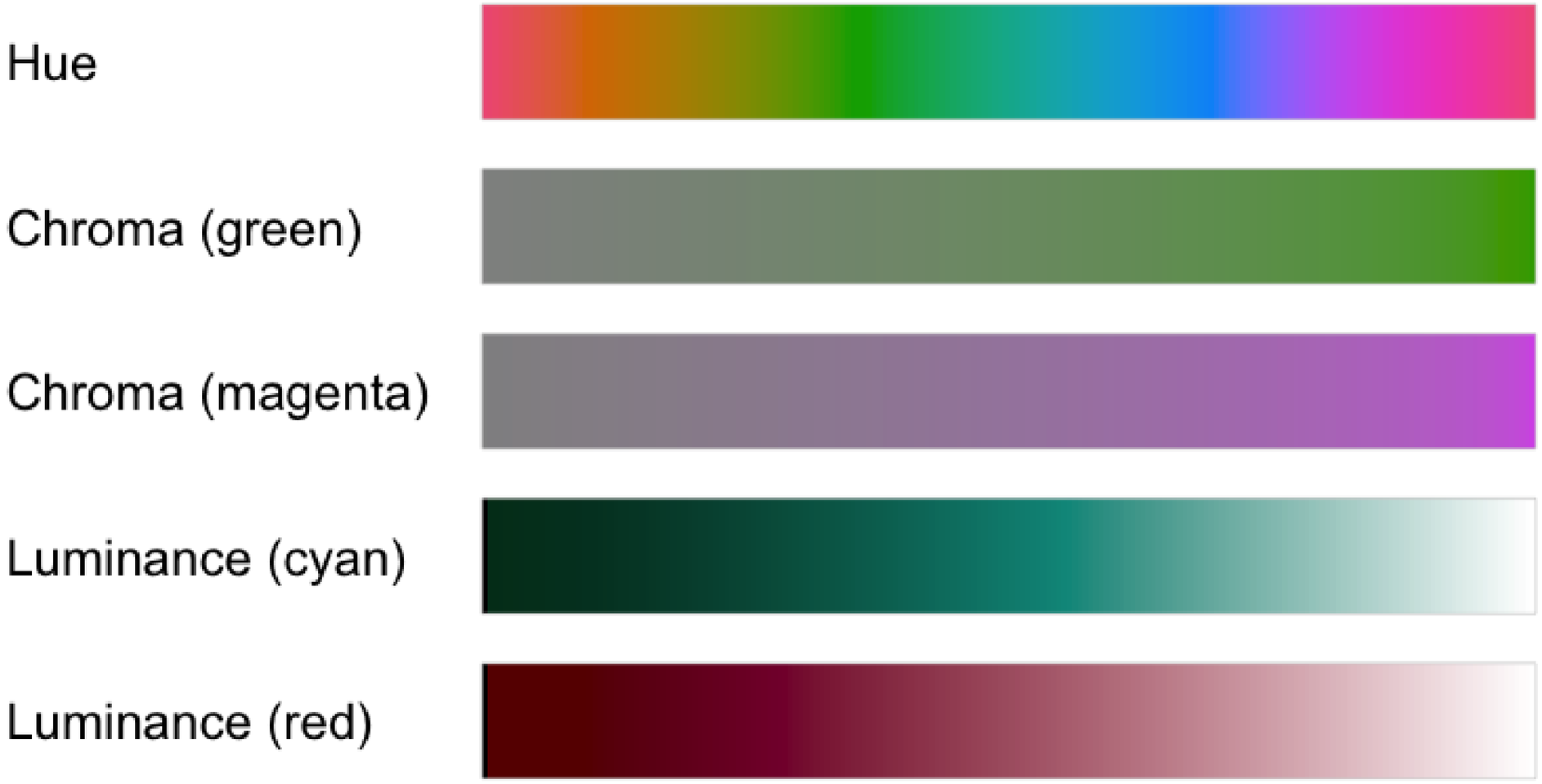
Colorspaces: Hue-Chroma-Luminance

Hue

Chroma (green)

Chroma (magenta)

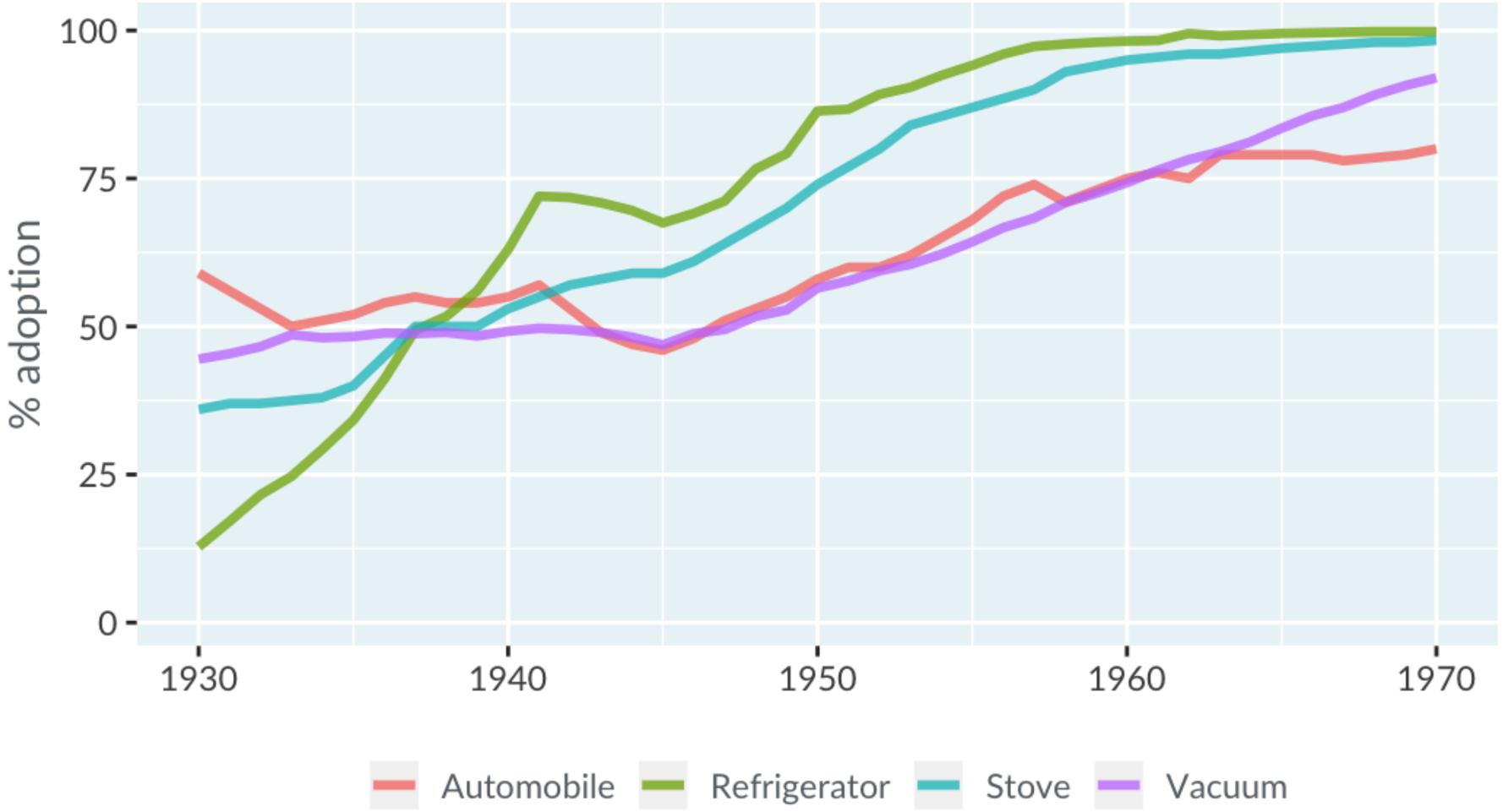
Colorspaces: Hue-Chroma-Luminance



Three types of color scale: qualitative

Type	Purpose	What to vary
qualitative	Distinguish unordered categories	hue

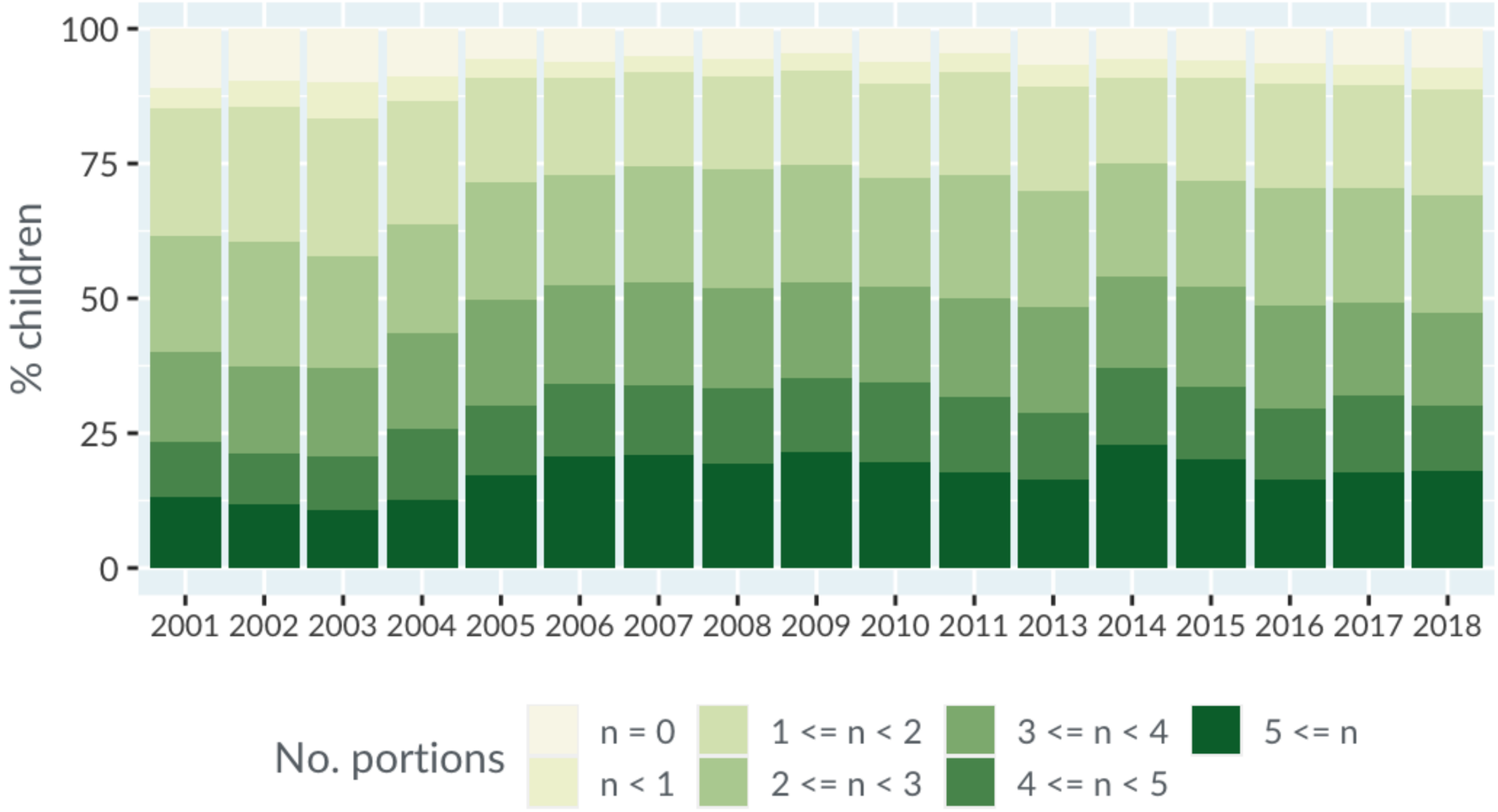
Qualitative palette example



Three types of color scale: sequential

Type	Purpose	What to vary
sequential	Show ordering	chroma or luminance

Sequential palette example

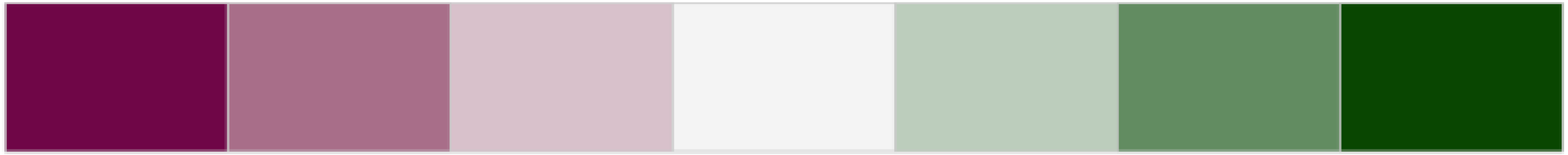


Another sequential palette example

[1] Viridis color scale: https://bids.github.io/colormap

Three types of color scale: diverging

Type	Purpose	What to vary
diverging	Show above or below a midpoint	chroma or luminance, with 2 hues

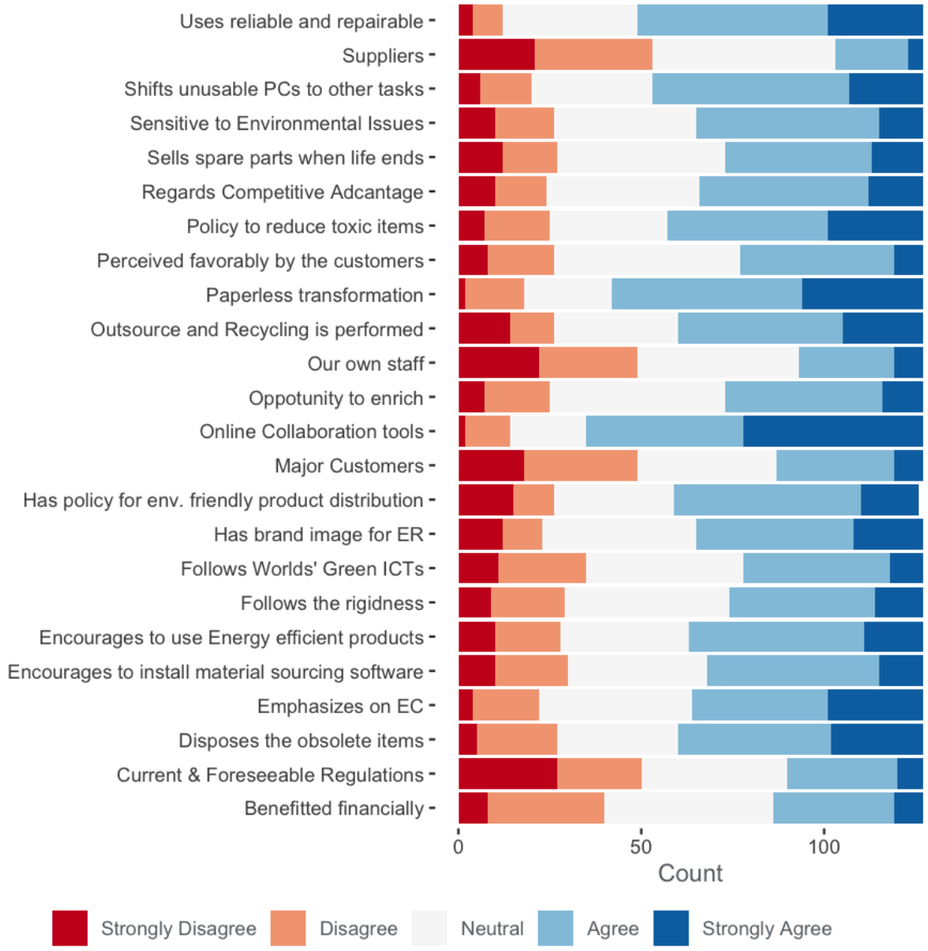


Green Tech in Malaysia survey dataset

question	response	n
Uses reliable and repairable	Strongly Disagree	4
Uses reliable and repairable	Disagree	8
Uses reliable and repairable	Neutral	37
Uses reliable and repairable	Agree	52
Uses reliable and repairable	Strongly Agree	26
...	...	...

[1] Islam et al. (2019) http://dx.doi.org/10.17632/wggvryfhsk.1

Diverging palette example



Let's practice!

UNDERSTANDING DATA VISUALIZATION

Plotting many variables at once

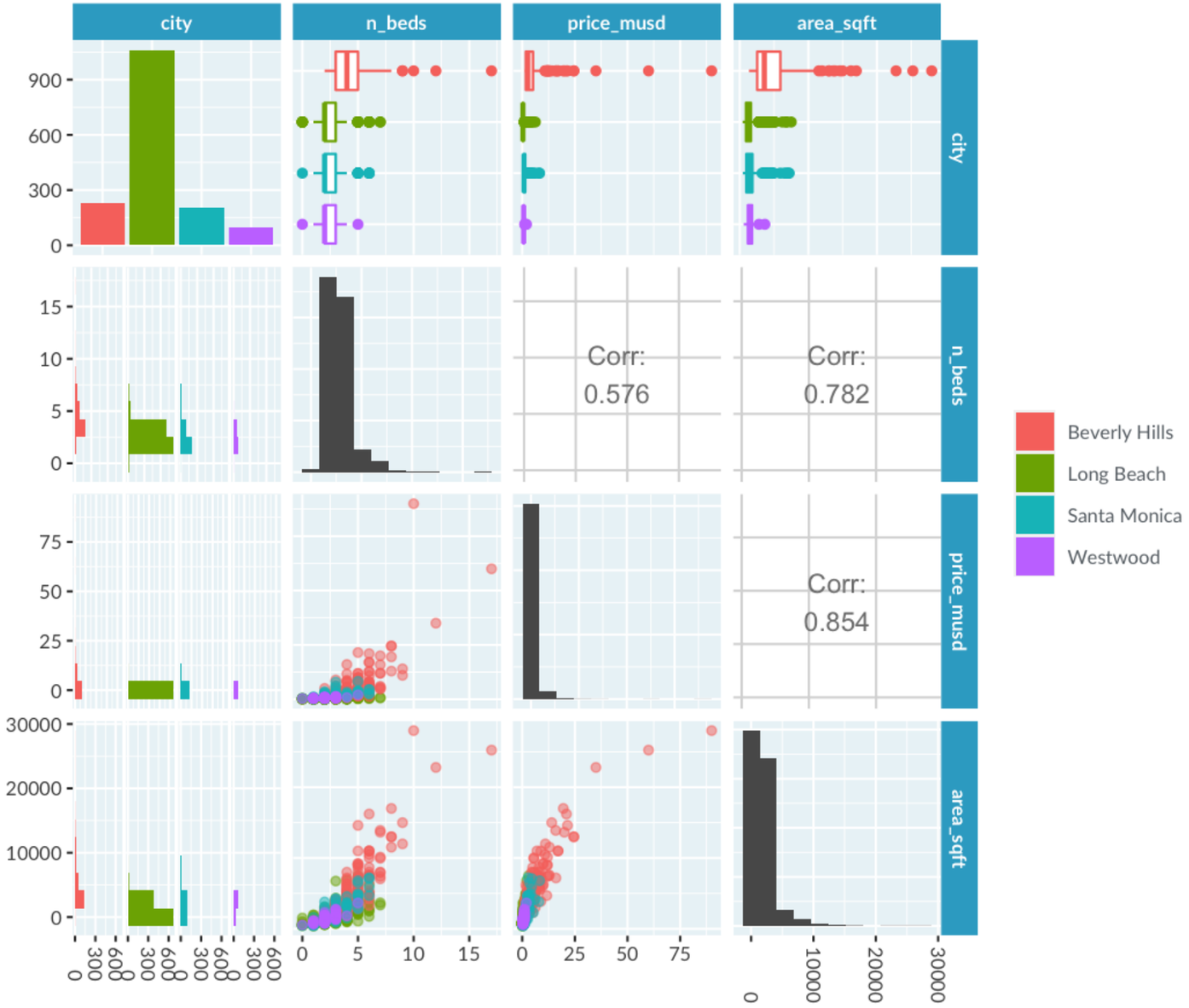
UNDERSTANDING DATA VISUALIZATION

Richie Cotton
Data Evangelist

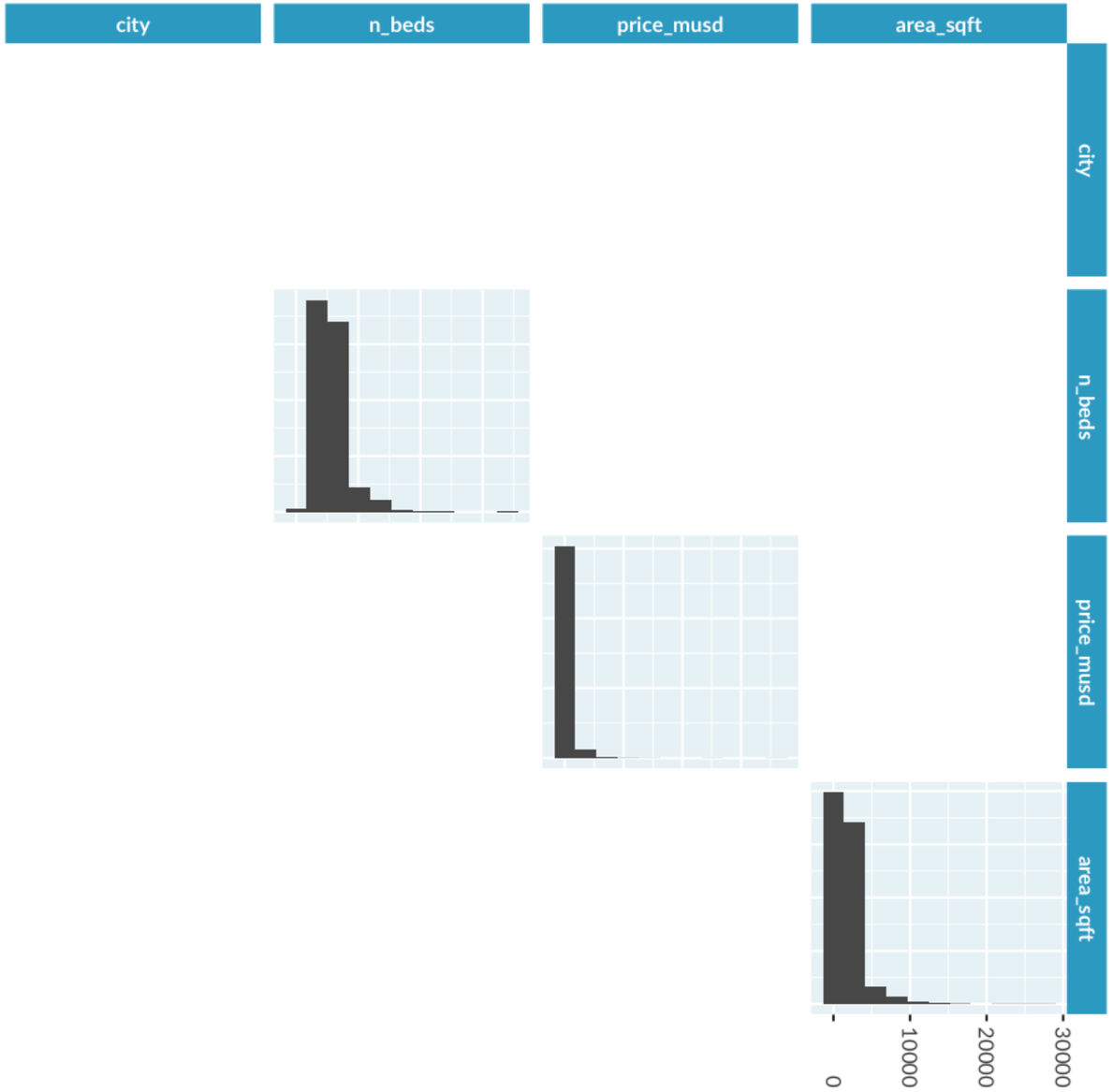
datacamp

When should you use a pair plot?

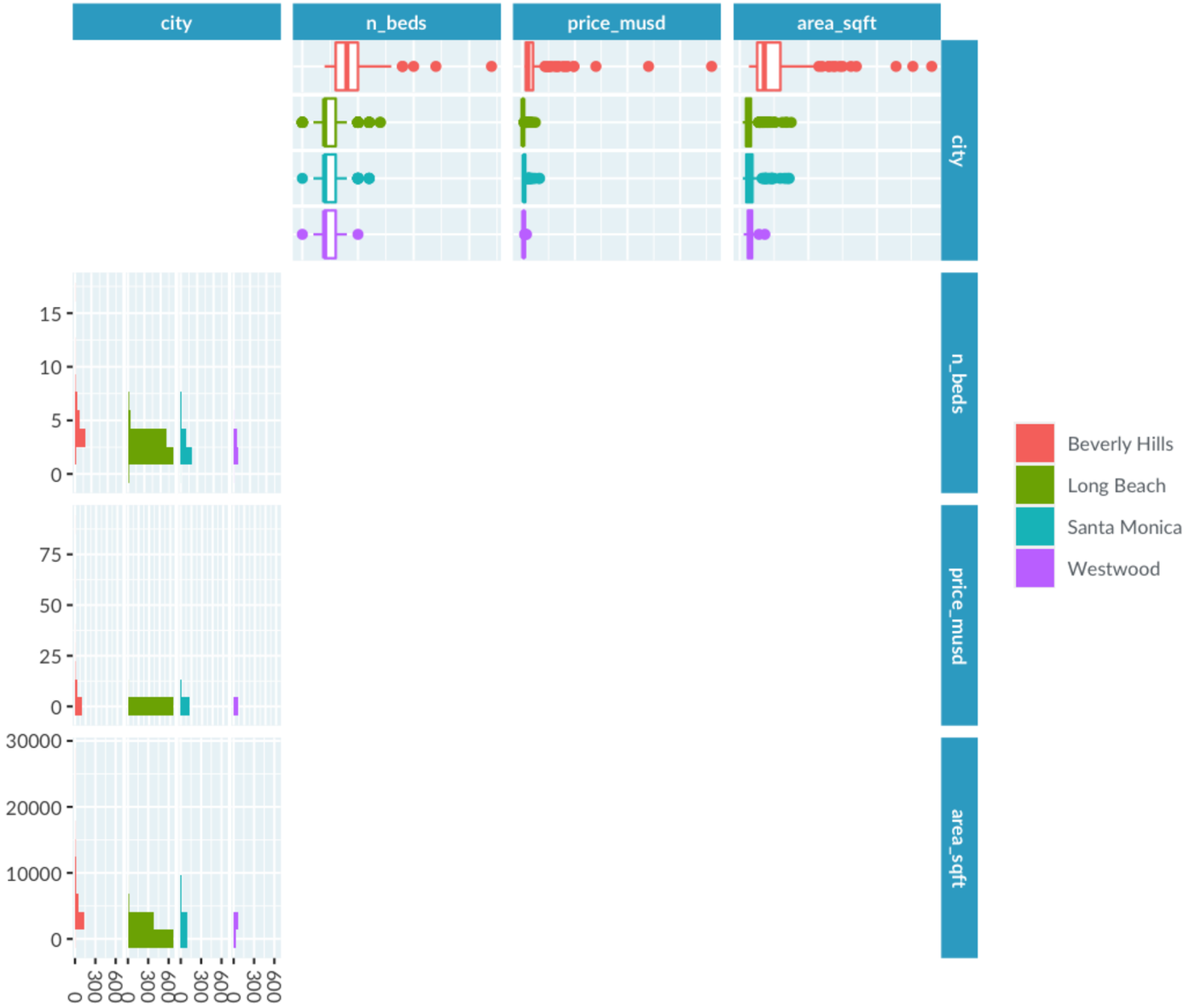
- You have up to ten variables (either continuous, categorical, or a mix).
- You want to see the distribution for each variable.
- You want to see the relationship between each pair of variables.

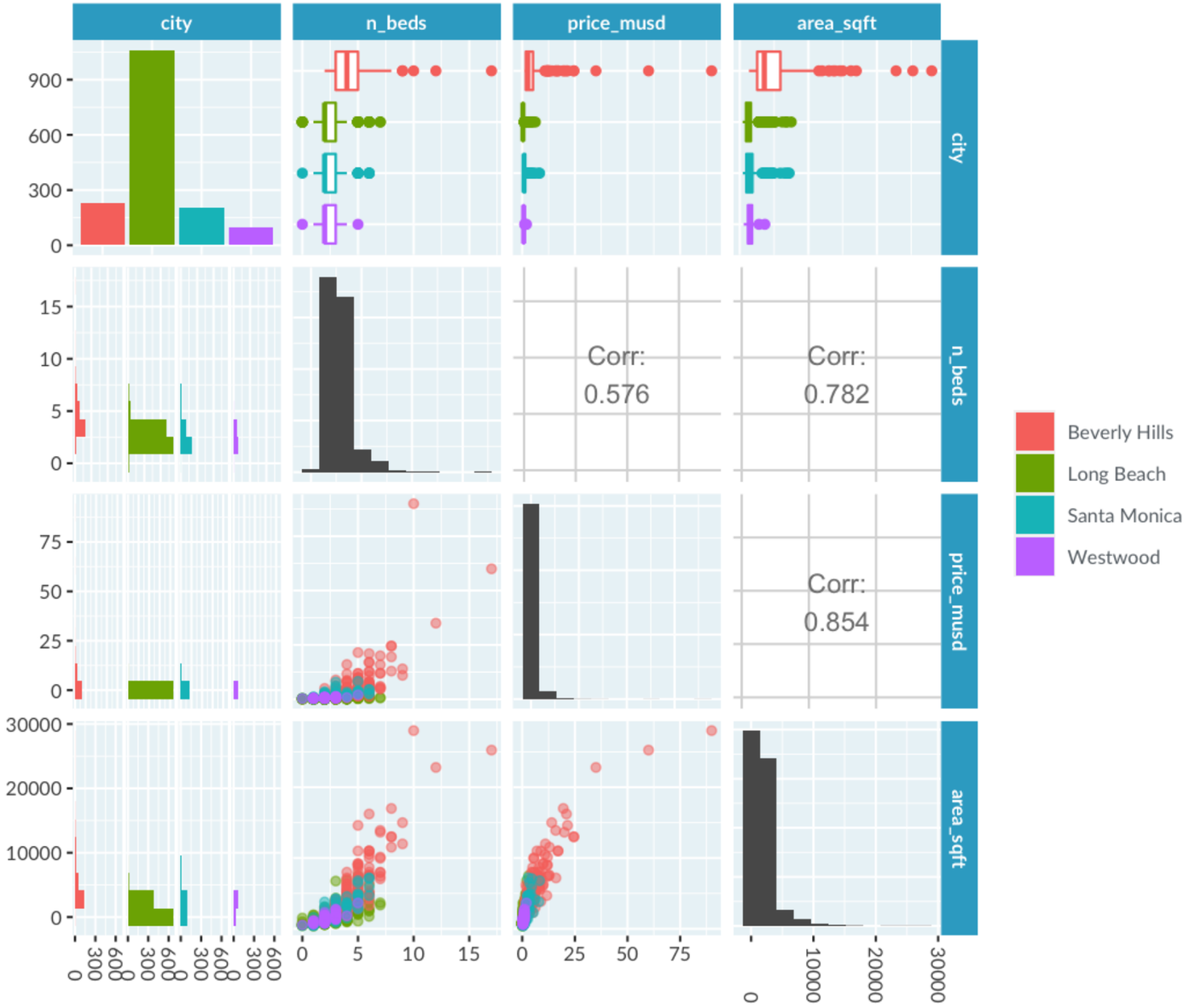

city
n_beds
price_musd
area_sqft
Corr:
0.576
Corr:
0.782
Corr:
0.854
Beverly Hills
Long Beach
Santa Monica
Westwood

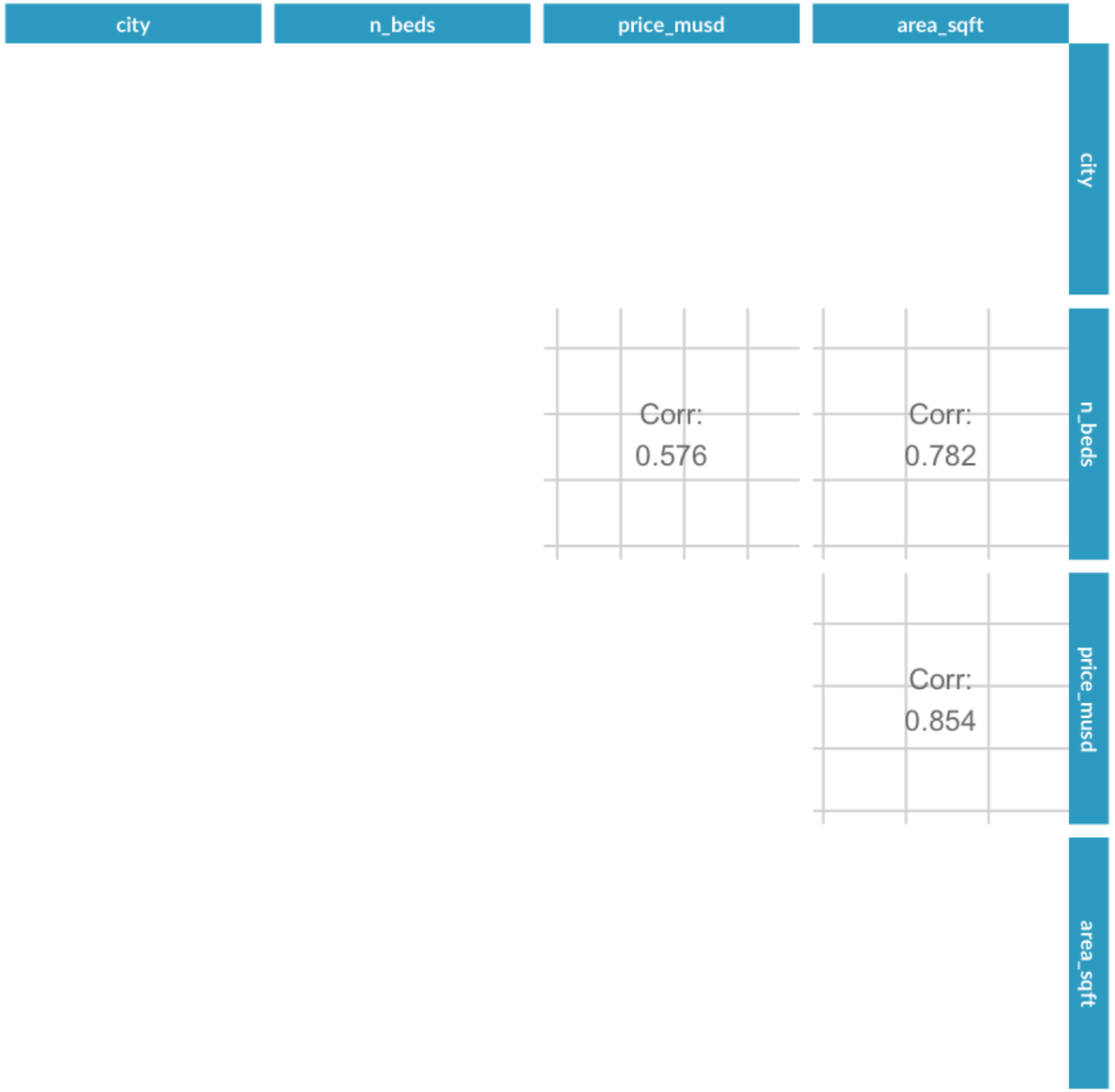
city
n_beds
price_musd
area_sqft
900
600
300
0
city
n_beds
price_musd
area_sqft
Beverly Hills
Long Beach
Santa Monica
Westwood


city
n_beds
price_musd
area_sqft
city
n_beds
price_musd
area_sqft
0
10000
20000
30000

city
n_beds
price_musd
area_sqft
Corr:
0.576
Corr:
0.782
Corr:
0.854
0
5
10
15
0
25
50
75
Beverly Hills
Long Beach
Santa Monica
Westwood

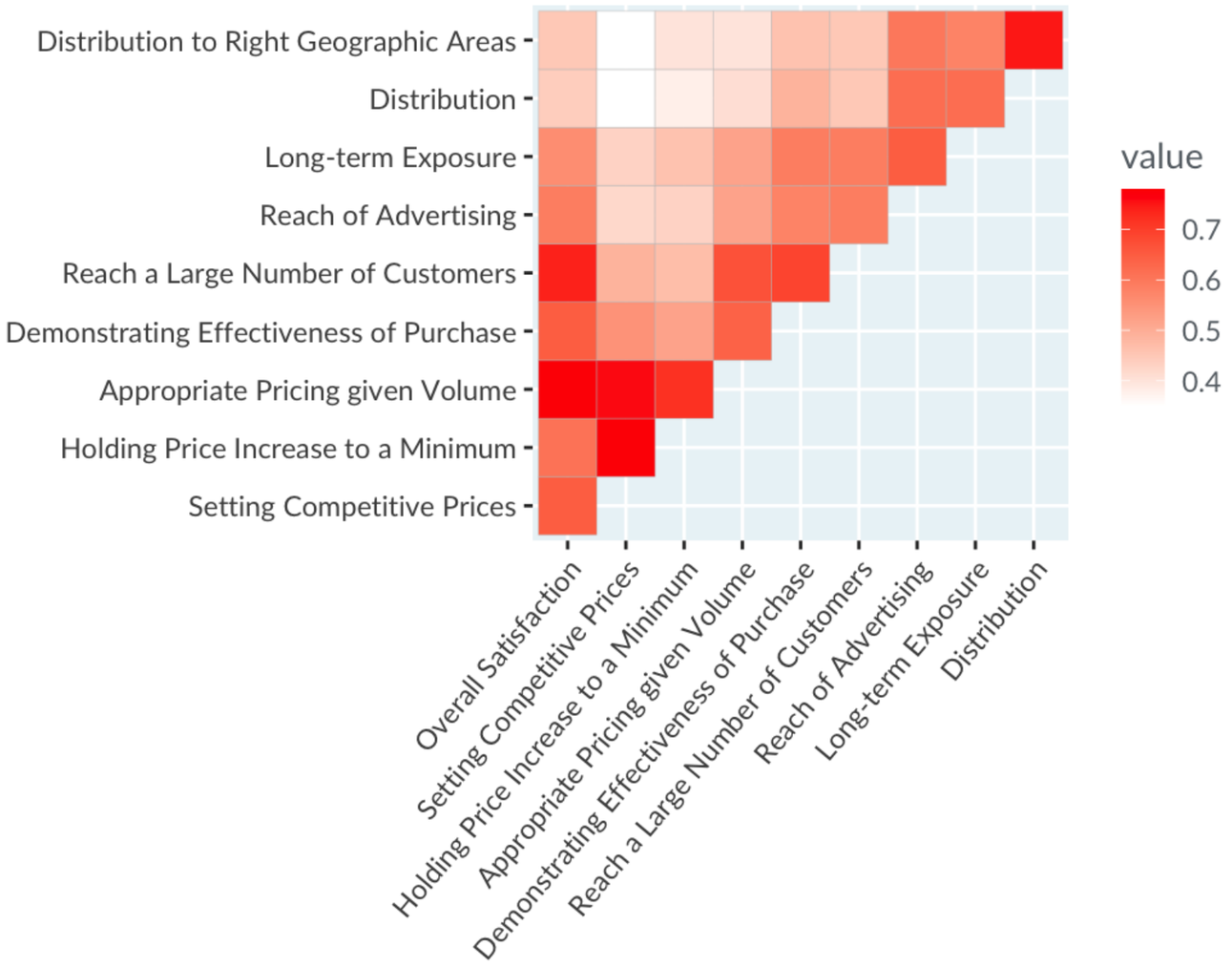

city
n_beds
price_musd
area_sqft
city
n_beds
price_musd
area_sqft
15
10
5
0
75
50
25
0
30000
20000
10000
0
0
300
600
300
600
300
600
300
600
Beverly Hills
Long Beach
Santa Monica
Westwood


city
n_beds
price_musd
area_sqft
Corr:
0.576
Corr:
0.782
Corr:
0.854
Beverly Hills
Long Beach
Santa Monica
Westwood


city
n_beds
price_musd
area_sqft
city
Corr:
0.576
Corr:
0.782
n_beds
Corr:
0.854
price_musd
area_sqft

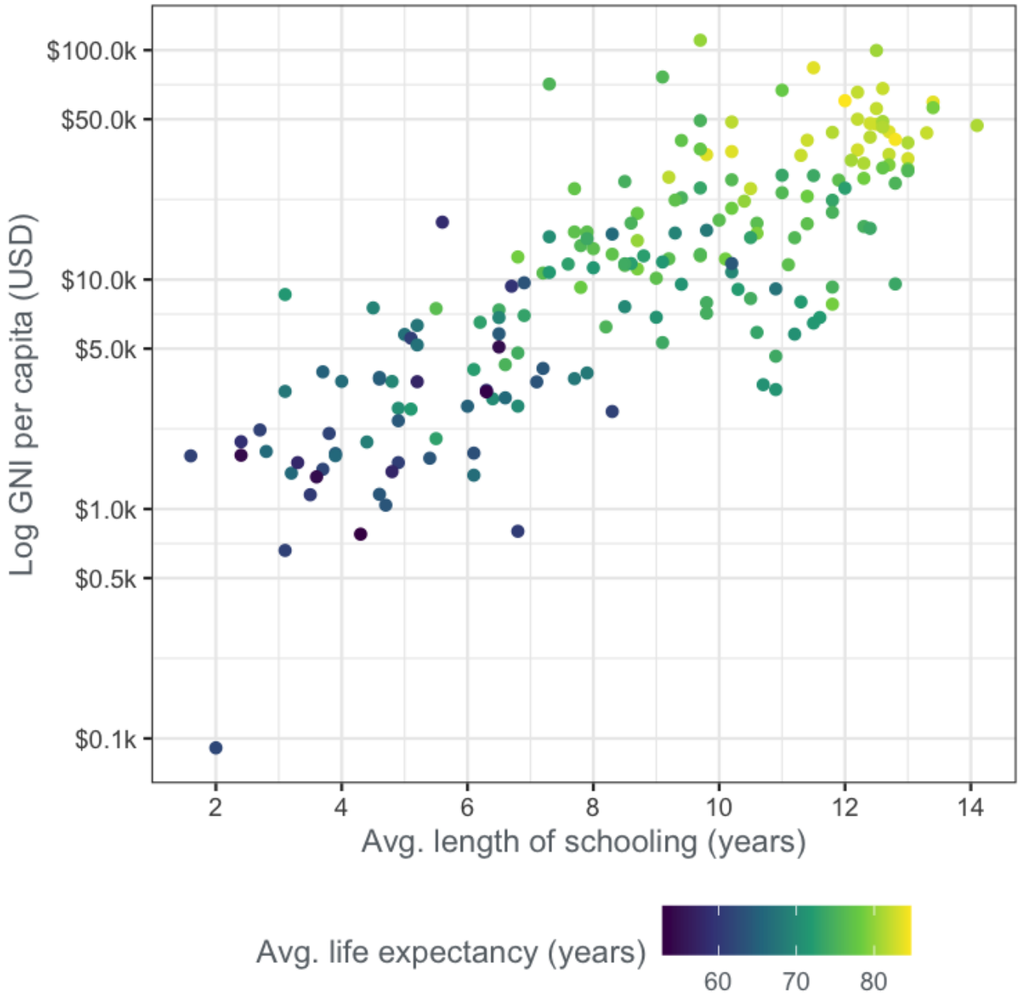
When should you use a correlation heatmap?

- You have lots of continuous variables.
- You want to a simple overview of how each pair of variables is related.



[1] Rossi, Allenby, and McCulloch (2005). Bayesian Statistics & Marketing

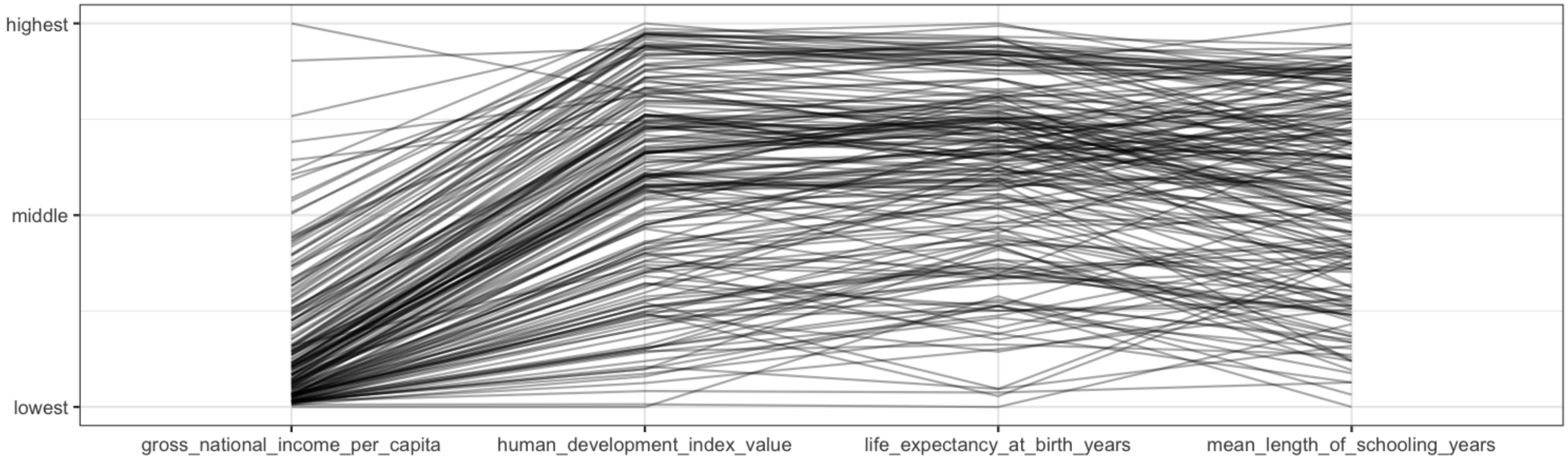
The United Nations dataset again

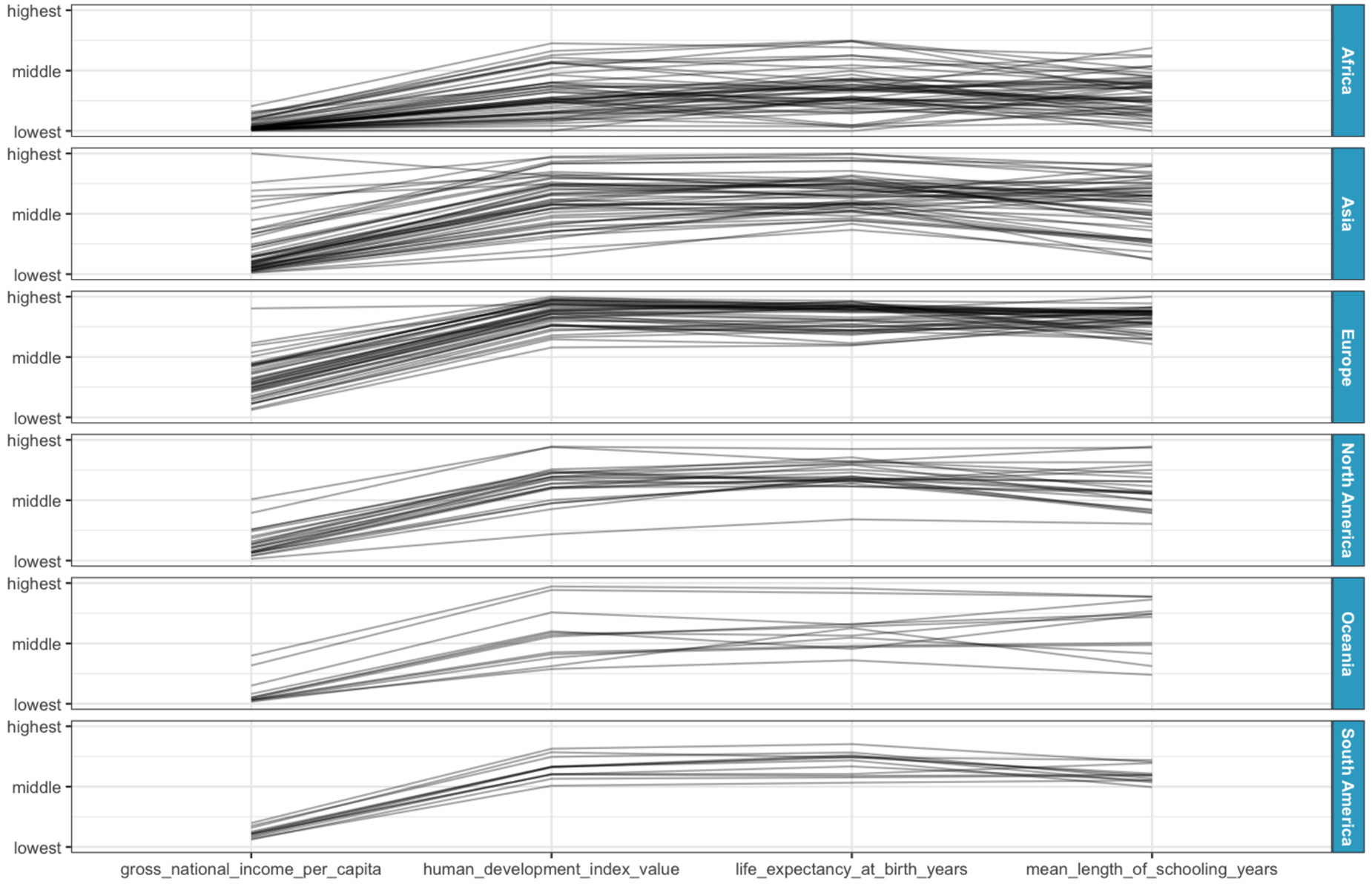


When should you use a parallel coordinates plot?

- You have lots of continuous variables.
- You want to find patterns across these variables, or
- You want to visualize clusters of observations.

A parallel coordinates plot




highest
middle
lowest
Africa
Asia
Europe
North America
Oceania
South America
gross_national_income_per_capita
human_development_index_value
life_expectancy_at_birth_years
mean_length_of_schooling_years

Let's practice!

UNDERSTANDING DATA VISUALIZATION