

A plot tells a thousand words

UNDERSTANDING DATA VISUALIZATION



Richie Cotton

Data Evangelist at DataCamp

What you'll learn

- How do you choose an appropriate plot?
- How do you interpret common types of plots?
- What are best practices for drawing plots?

Three ways of getting insights

Calculating summary statistics

mean, median, standard deviation

Running models

linear and logistic regression

Drawing plots

scatter, bar, histogram

The Datasaurus Dozen

away_x	away_y	bullseye_x	bullseye_y	...	x_shape_x	x_shape_y
32.33	61.41	51.20	83.34	...	38.34	92.47
53.42	26.19	58.97	85.50	...	35.75	94.12
63.92	30.83	51.87	85.83	...	32.77	88.52
70.29	82.53	48.18	85.05	...	33.73	88.62
34.12	45.73	41.68	84.02	...	37.24	83.72
67.67	37.11	37.89	82.57	...	36.03	82.04

¹ Matejka, J., & Fitzmaurice, G. (2017) <https://www.autodeskresearch.com/publications/samestats>

Mean of x for each dataset

dataset	mean(x)
away	54.27
bullseye	54.27
circle	54.27
dino	54.26
dots	54.26
h_lines	54.26
high_lines	54.27

dataset	mean(x)
slant_down	54.27
slant_up	54.27
star	54.27
v_lines	54.27
wide_lines	54.27
x_shape	54.26

Mean of x and y for each dataset

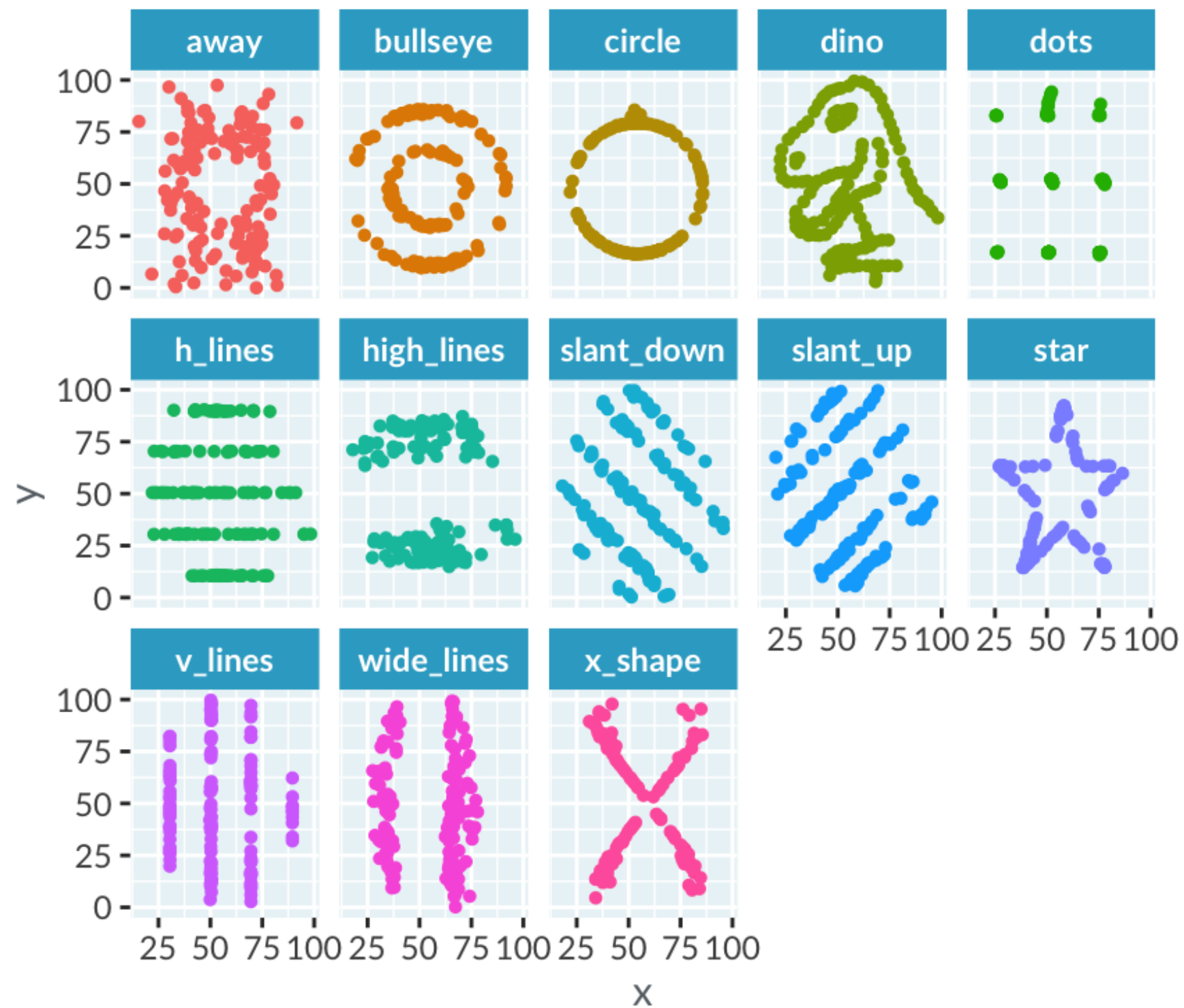
dataset	mean(x)	mean(y)
away	54.27	47.83
bullseye	54.27	47.83
circle	54.27	47.84
dino	54.26	47.83
dots	54.26	47.84
h_lines	54.26	47.83
high_lines	54.27	47.84

dataset	mean(x)	mean(y)
slant_down	54.27	47.84
slant_up	54.27	47.83
star	54.27	47.84
v_lines	54.27	47.84
wide_lines	54.27	47.83
x_shape	54.26	47.84

Standard deviations for each dataset

dataset	std_dev(x)	std_dev(y)
away	16.77	26.94
bullseye	16.77	26.94
circle	16.76	26.93
dino	16.77	26.94
dots	16.77	26.93
h_lines	16.77	26.94
high_lines	16.77	26.94

dataset	std_dev(x)	std_dev(y)
slant_down	16.77	26.94
slant_up	16.77	26.94
star	16.77	26.93
v_lines	16.77	26.94
wide_lines	16.77	26.94
x_shape	16.77	26.93



Continuous and categorical variables

Continuous: usually numbers

- heights, temperatures, revenues

Continuous and categorical variables

Continuous: usually numbers

- heights, temperatures, revenues

Categorical: usually text

- eye colors, countries, industry

Continuous and categorical variables

Continuous: usually numbers

- heights, temperatures, revenues

Categorical: usually text

- eye colors, countries, industry

Can be either

- age is continuous, but age group is categorical
- time is continuous, month of year is categorical

Let's practice!

UNDERSTANDING DATA VISUALIZATION

Histograms

UNDERSTANDING DATA VISUALIZATION



Richie Cotton

Data Evangelist at DataCamp

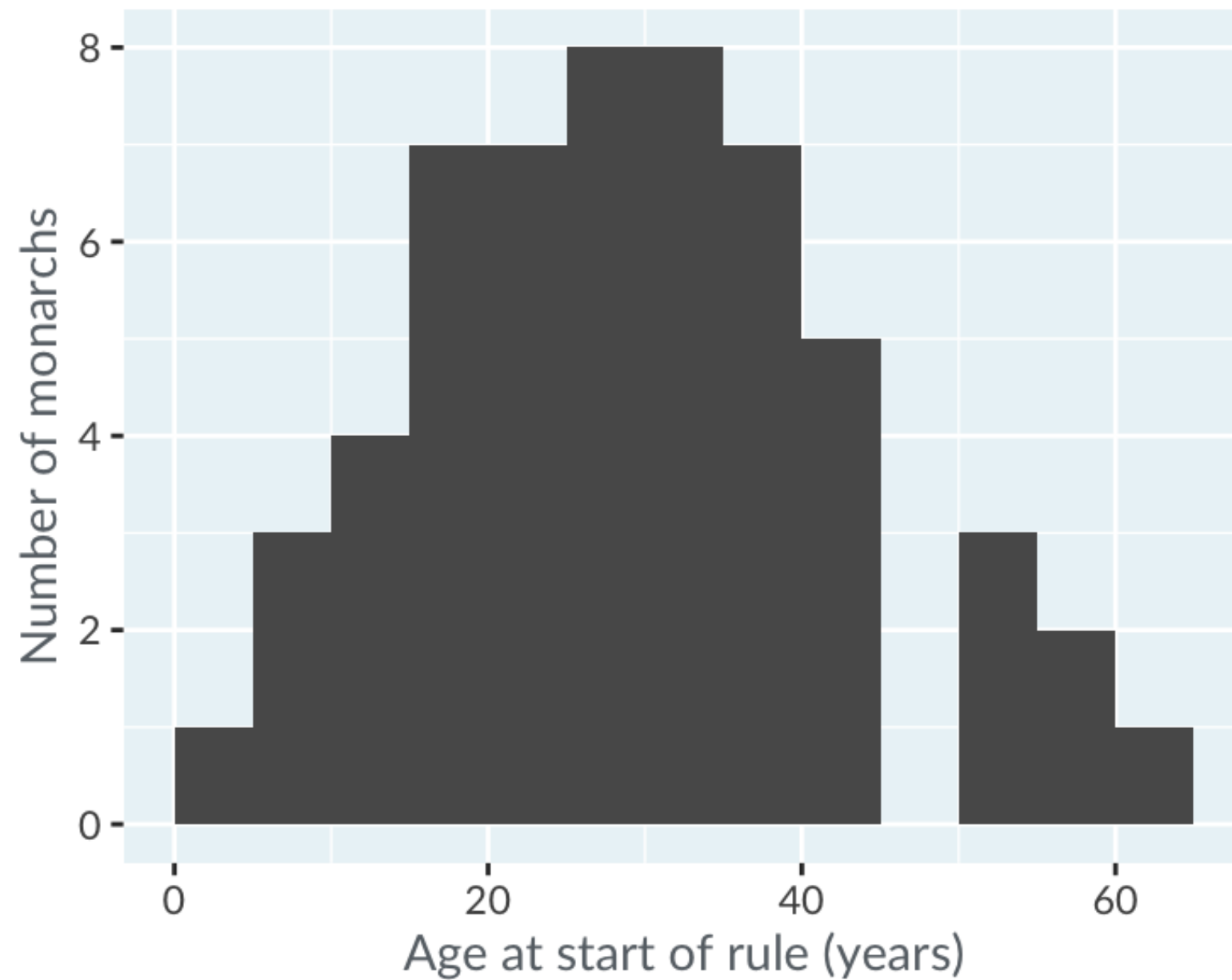
When should you use a histogram?

1. If you have a single continuous variable.
2. You want to ask questions about the shape of its distribution.

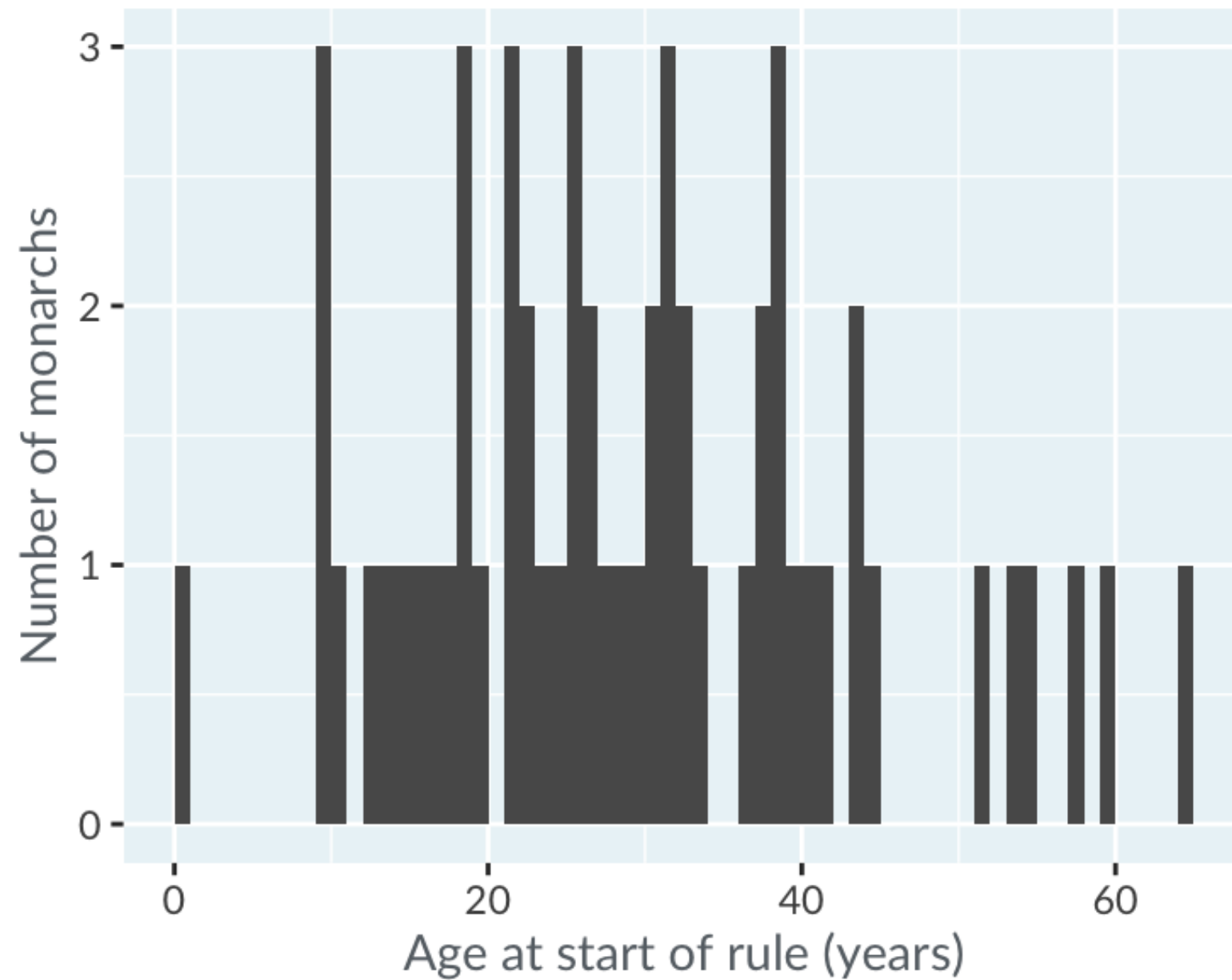
Kings and Queens of England & Britain

official_name	house	birth_date	start_of_rule	age_at_start_of_rule
Charles III	Windsor	1948-11-14	2022-09-08	73.86575
Elizabeth II	Windsor	1926-04-21	1952-02-06	25.79603
George VI	Windsor	1895-12-14	1936-12-11	40.99110
Edward VIII	Windsor	1894-06-23	1936-01-20	41.57426
...
Eadred	Wessex	0923-07-01	0946-05-26	22.90212
Edmund I	Wessex	0921-07-01	0939-10-27	18.32170
Aethelstan	Wessex	0894-07-01	0924-07-01	29.99863

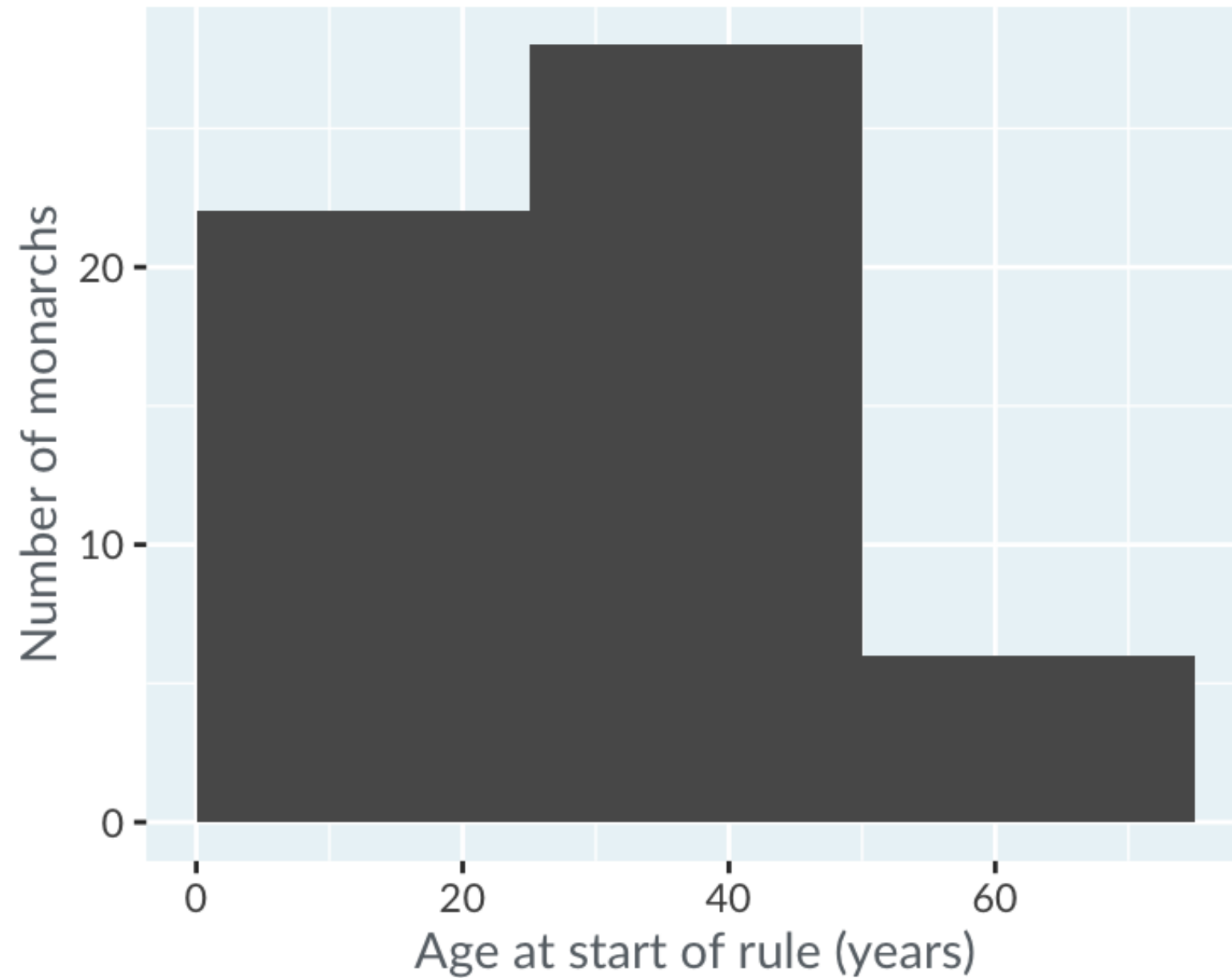
Histogram of age at start of rule



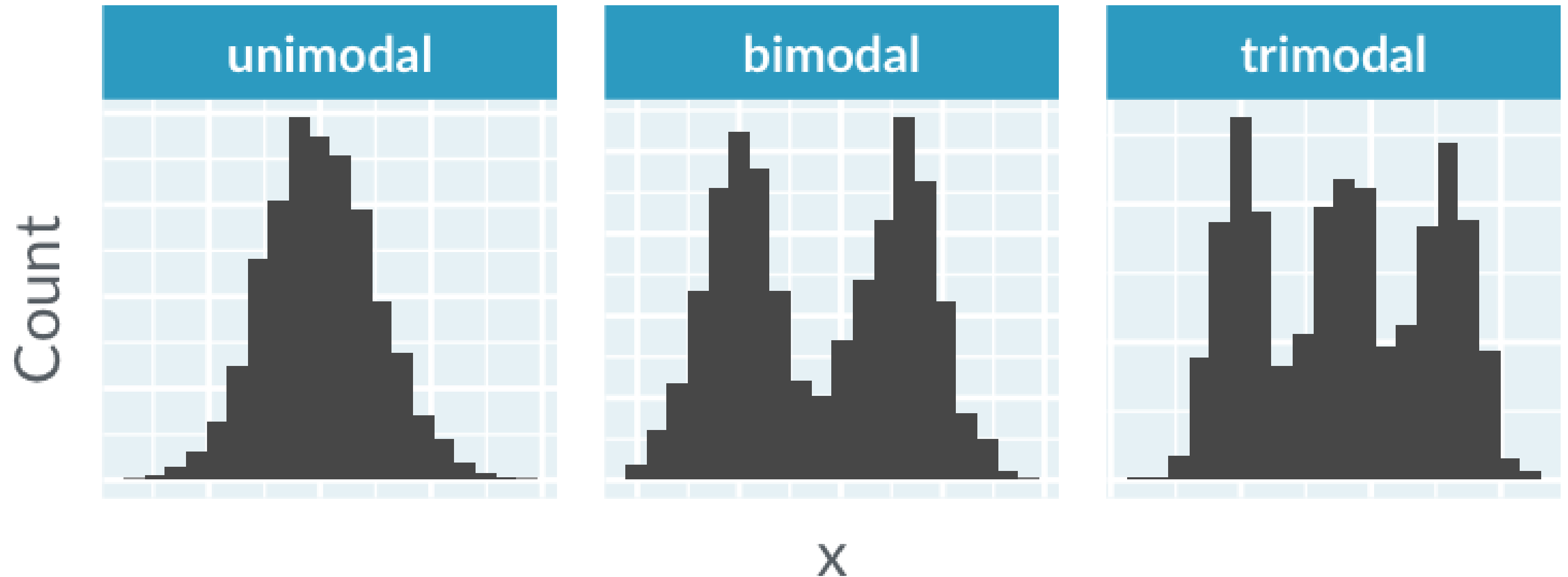
Choosing binwidth: 1 year



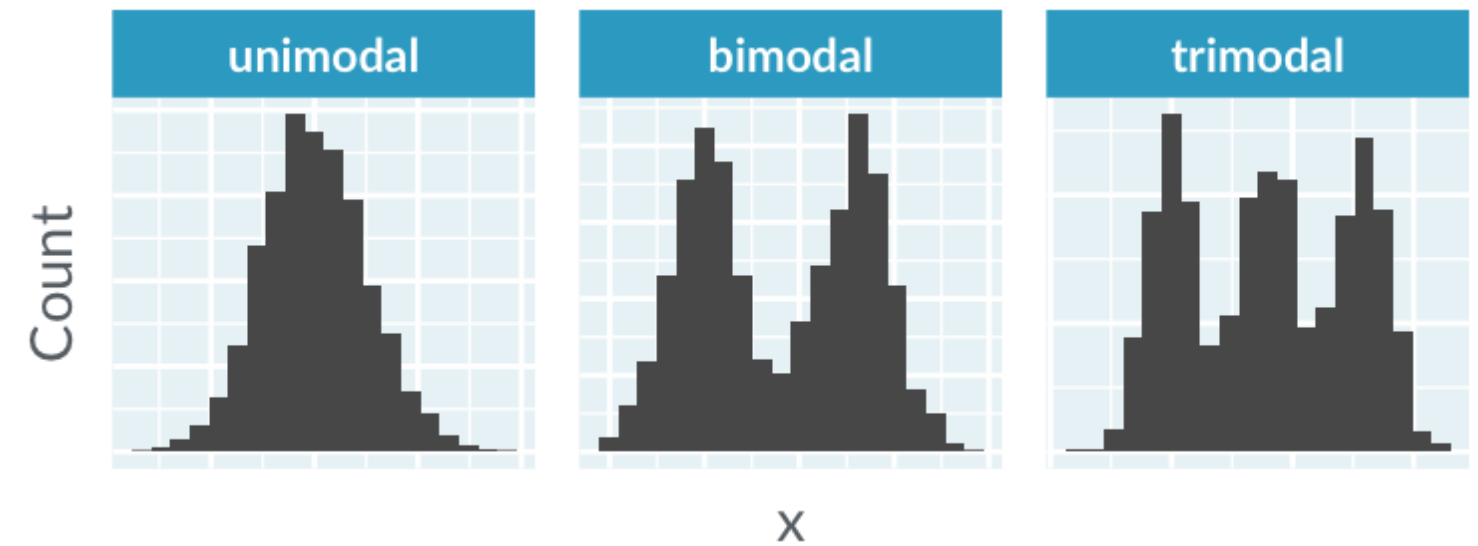
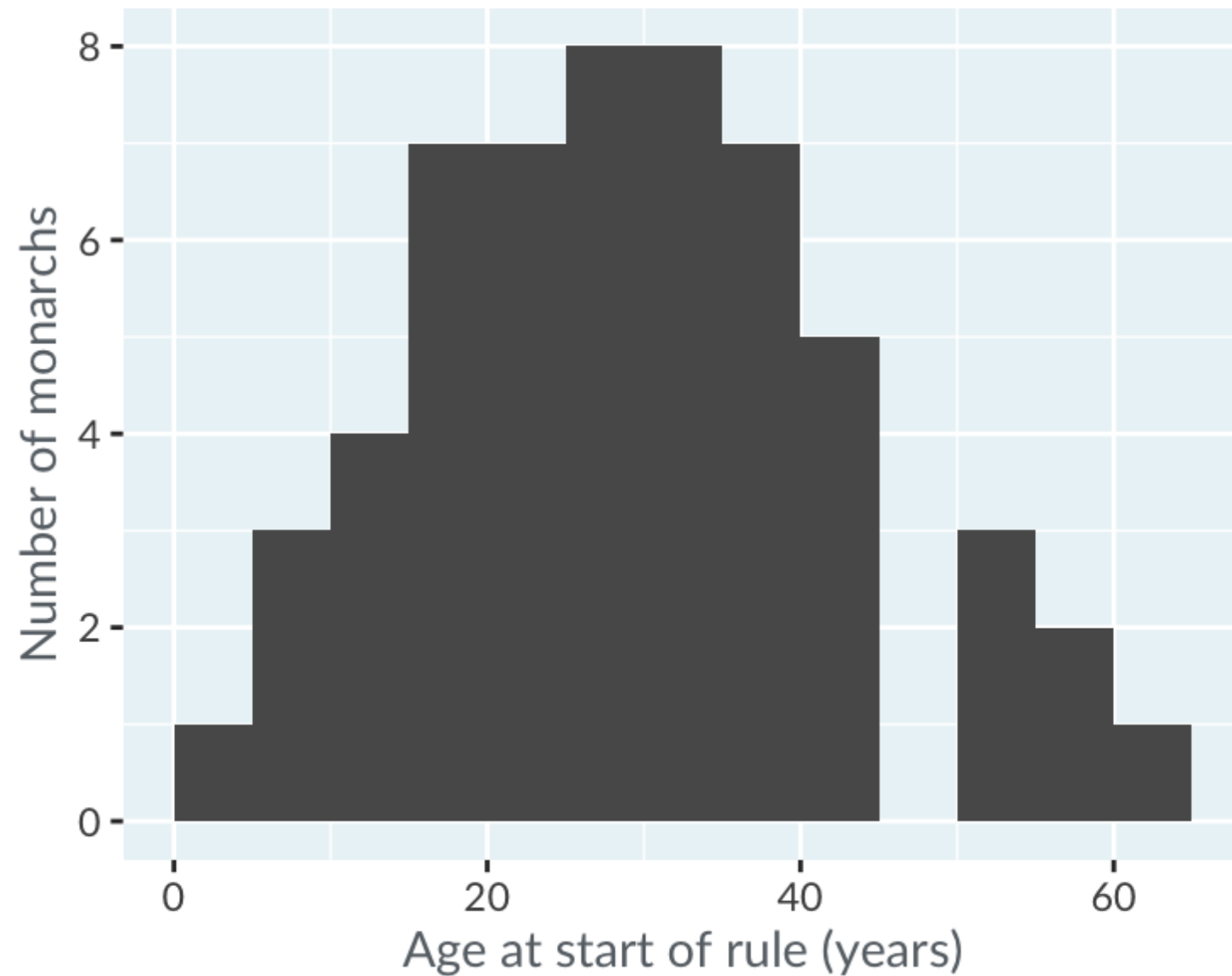
Choosing binwidth: 25 years



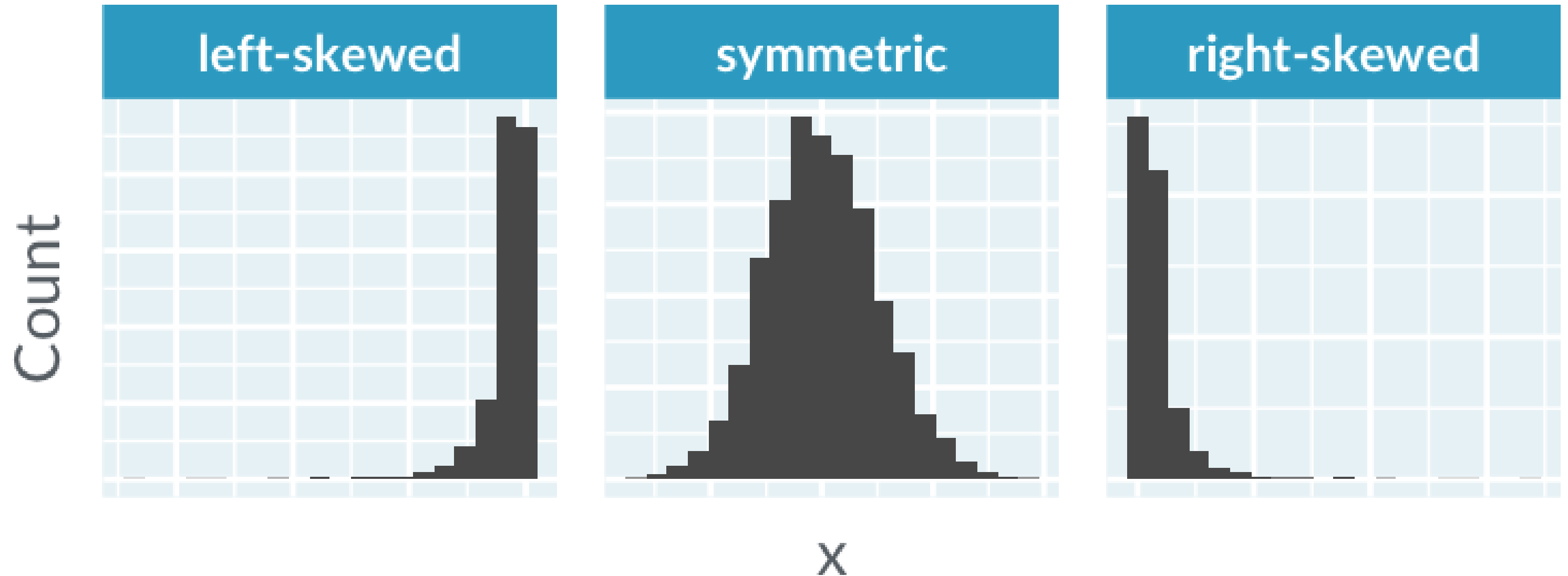
Modality: how many peaks?



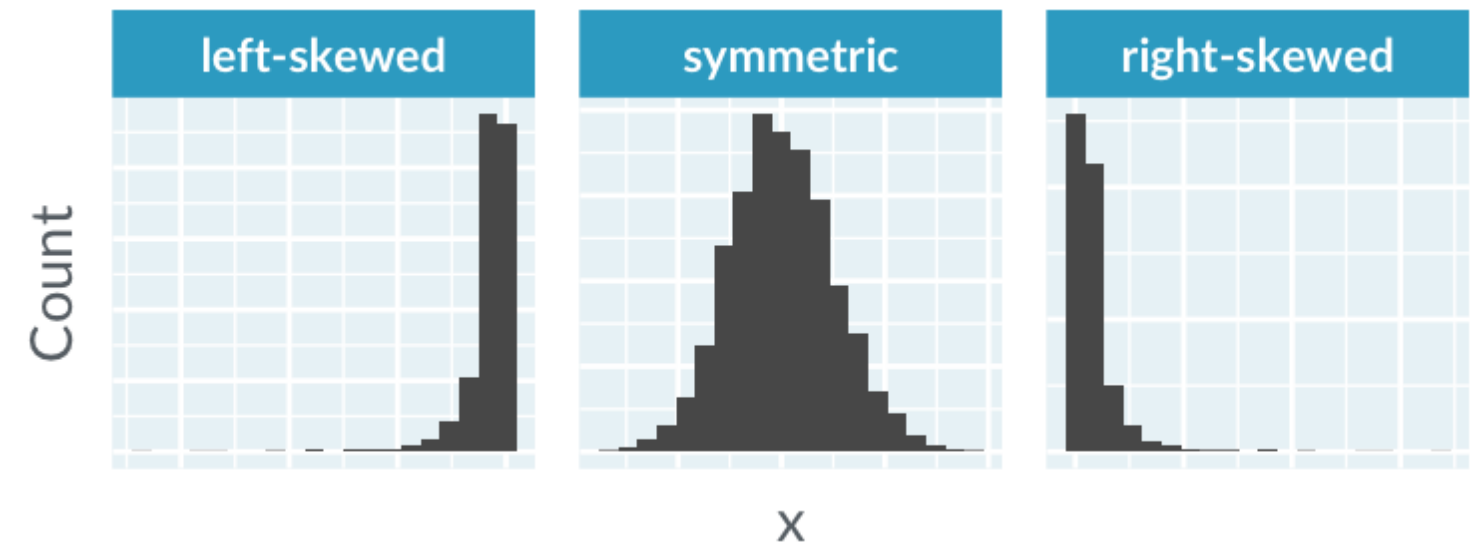
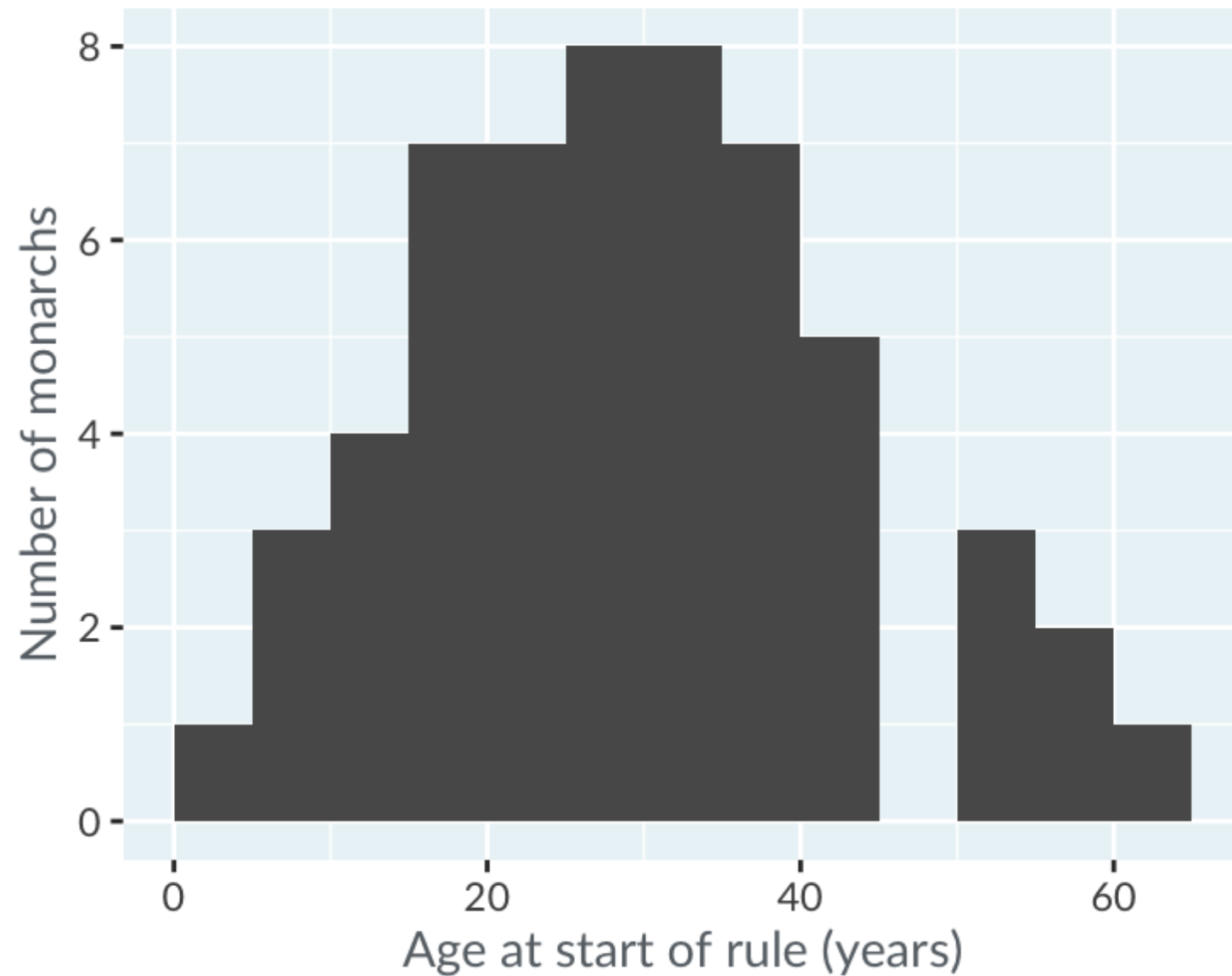
Modality: how many peaks?



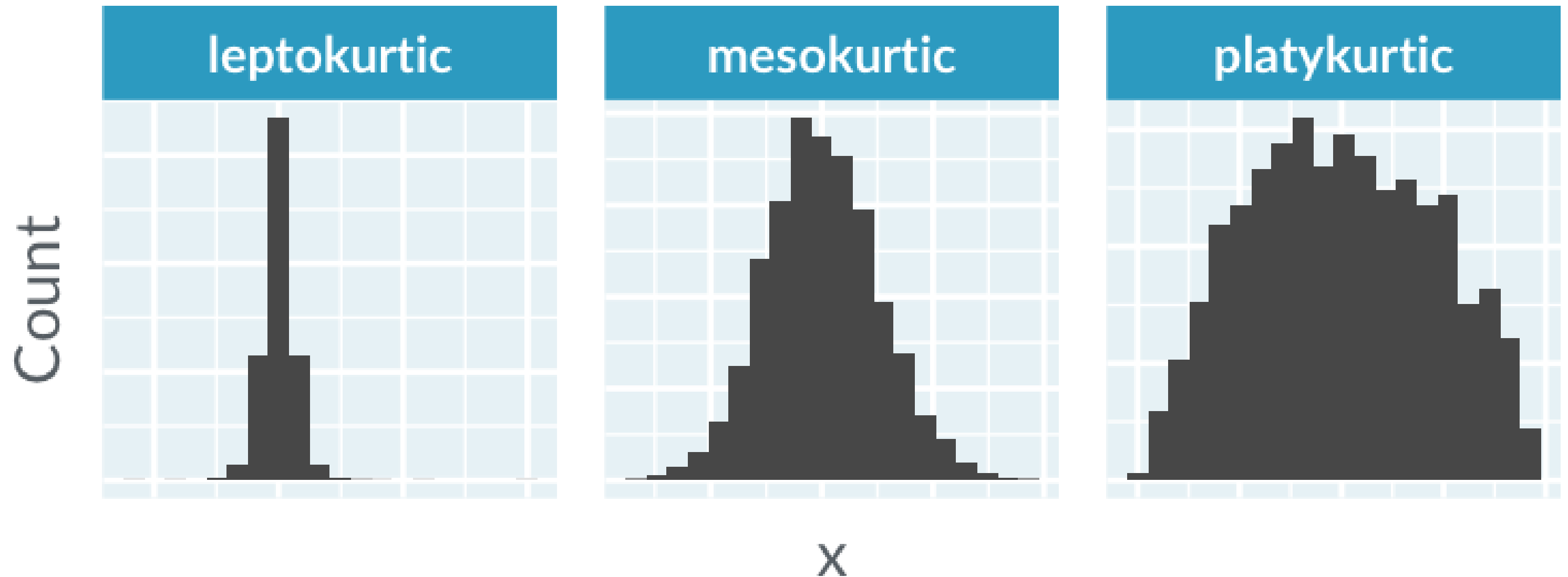
Skewness: is it symmetric?



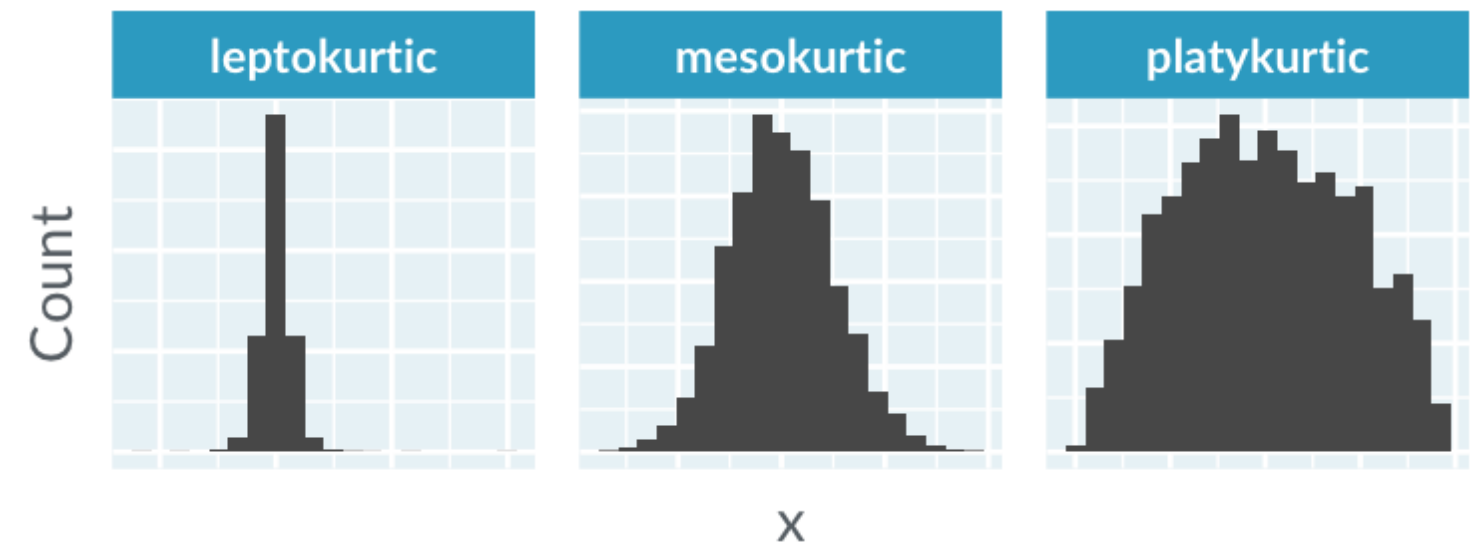
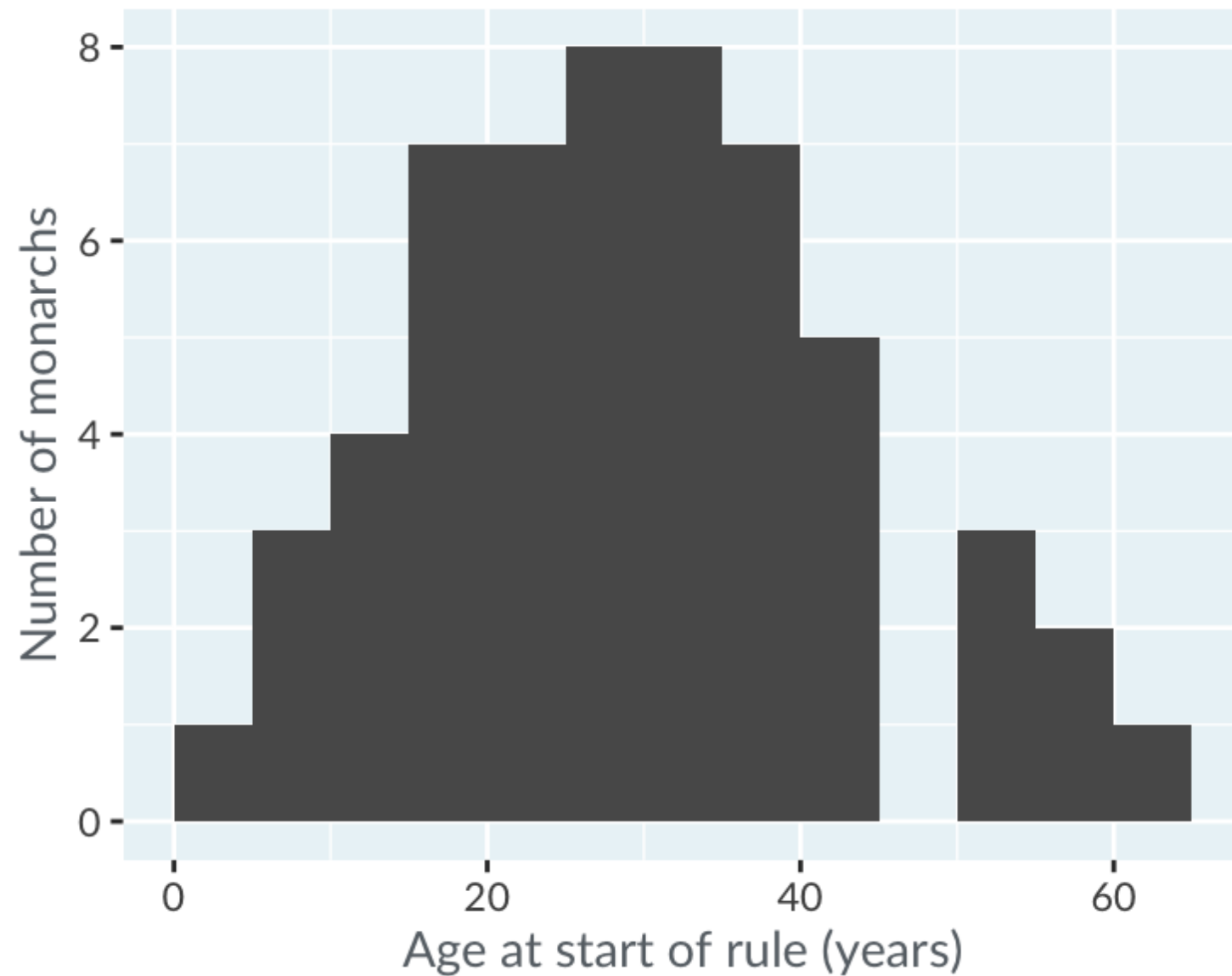
Skewness: is it symmetric?



Kurtosis: how many extreme values?



Kurtosis: how many extreme values?



Let's practice!

UNDERSTANDING DATA VISUALIZATION

Box plots

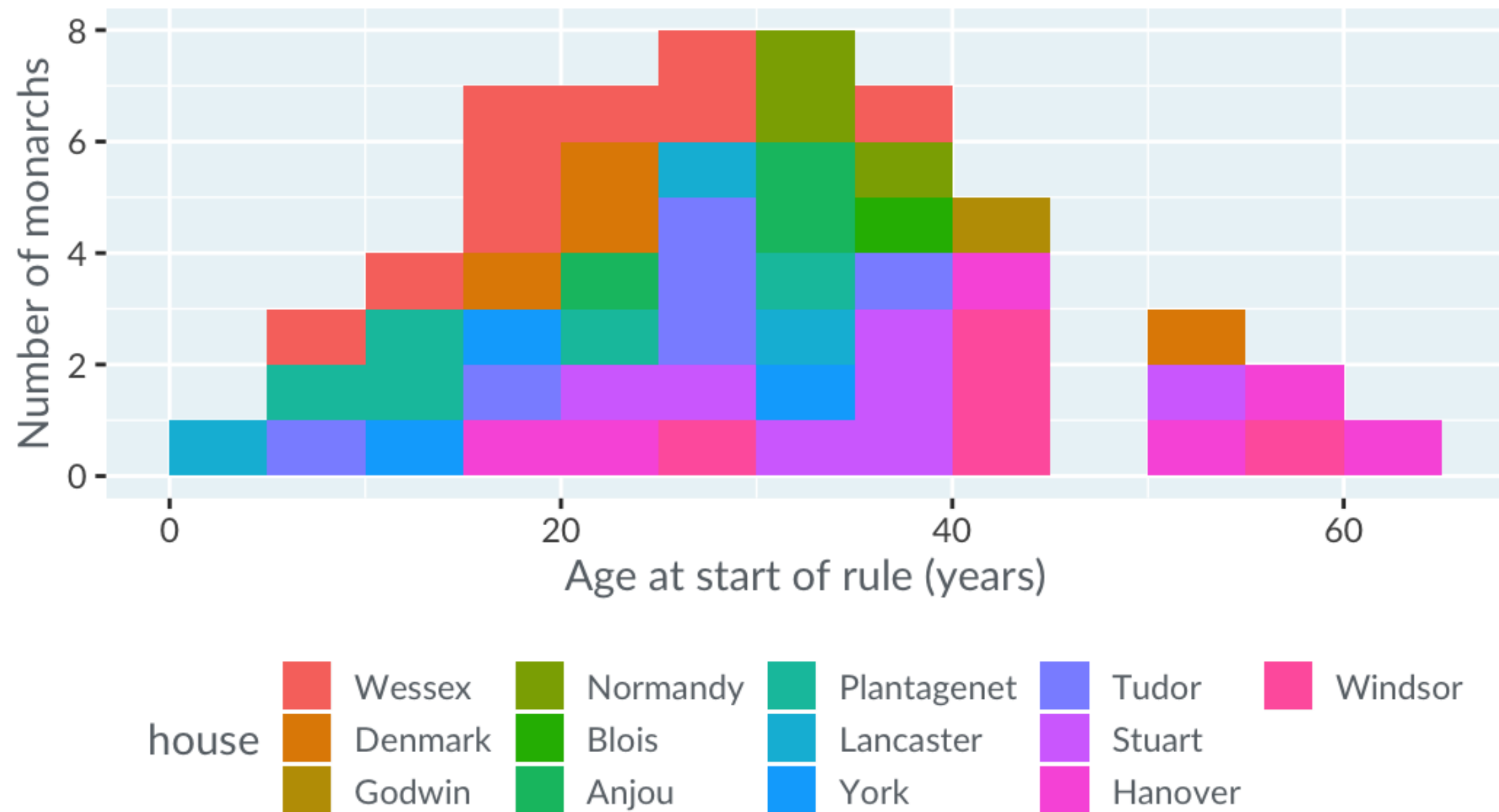
UNDERSTANDING DATA VISUALIZATION



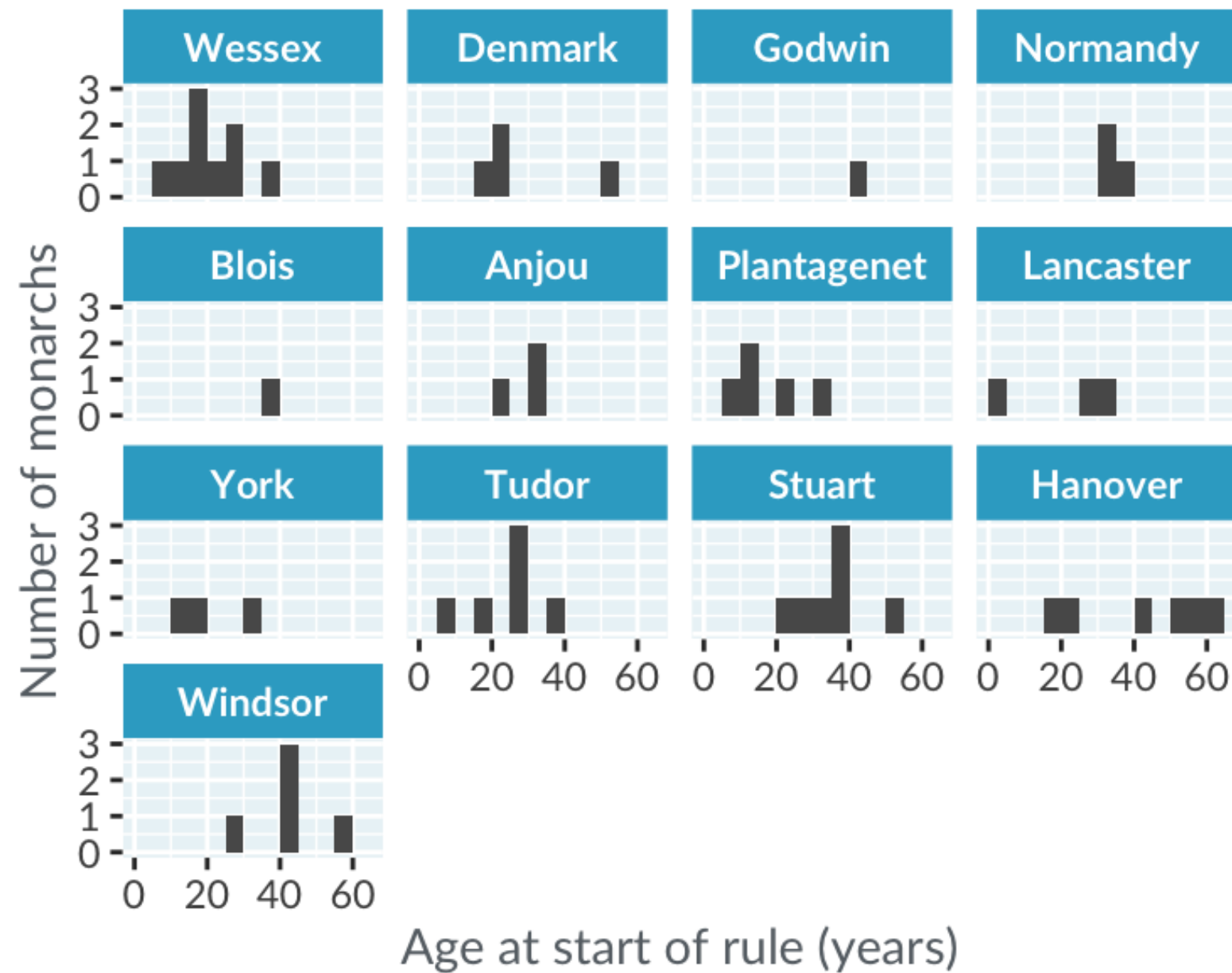
Richie Cotton

Data Evangelist at DataCamp

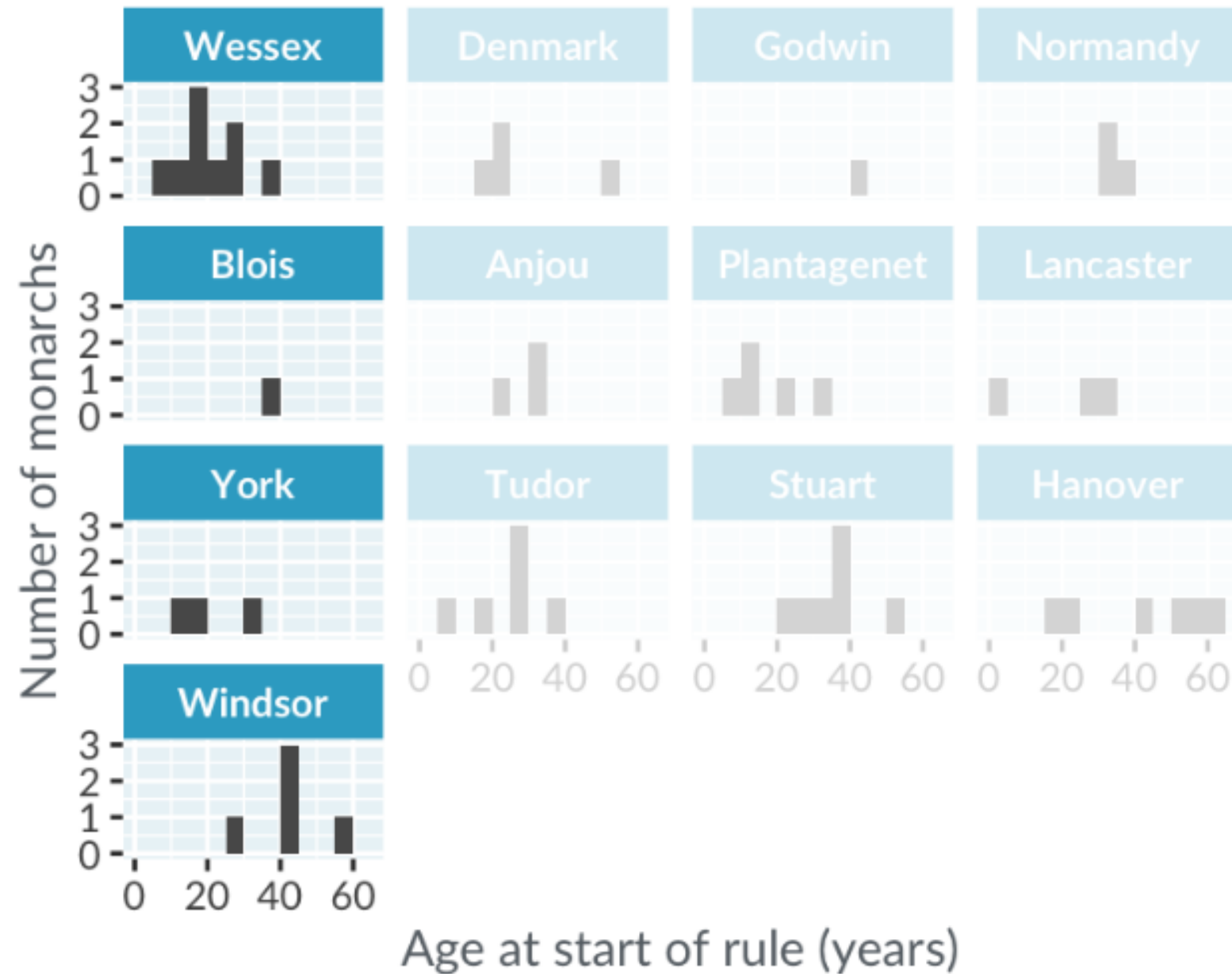
You can't just color in histograms



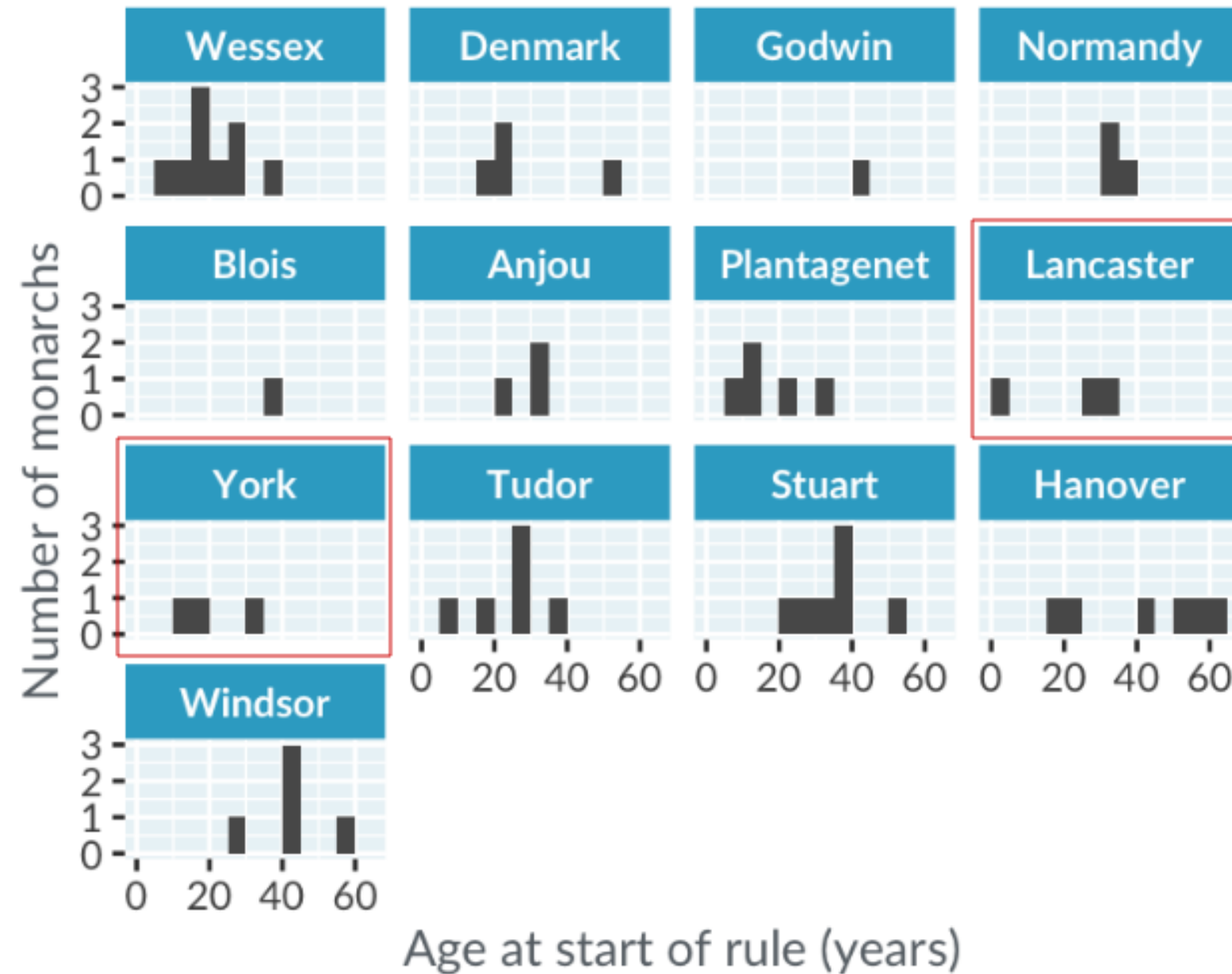
Draw each histogram in its own panel

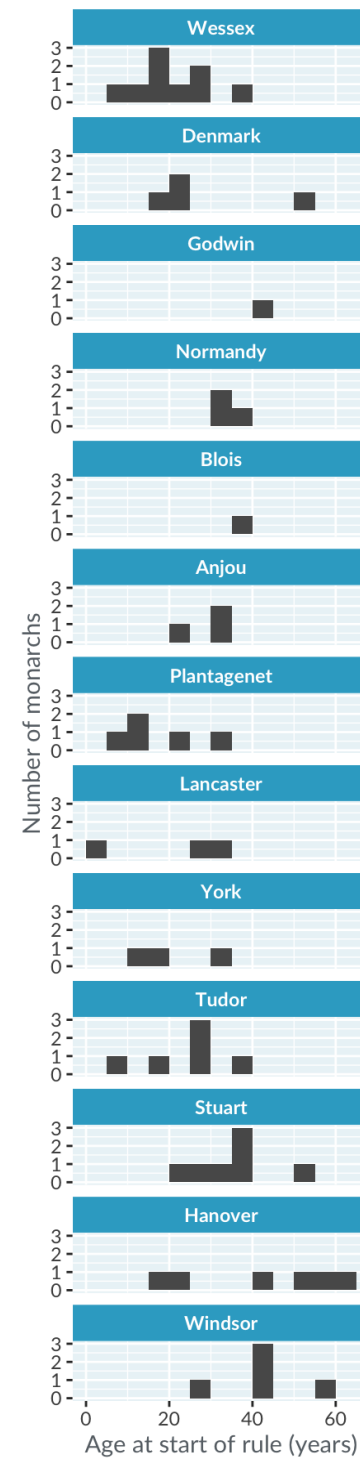


Draw each histogram in its own panel



Draw each histogram in its own panel

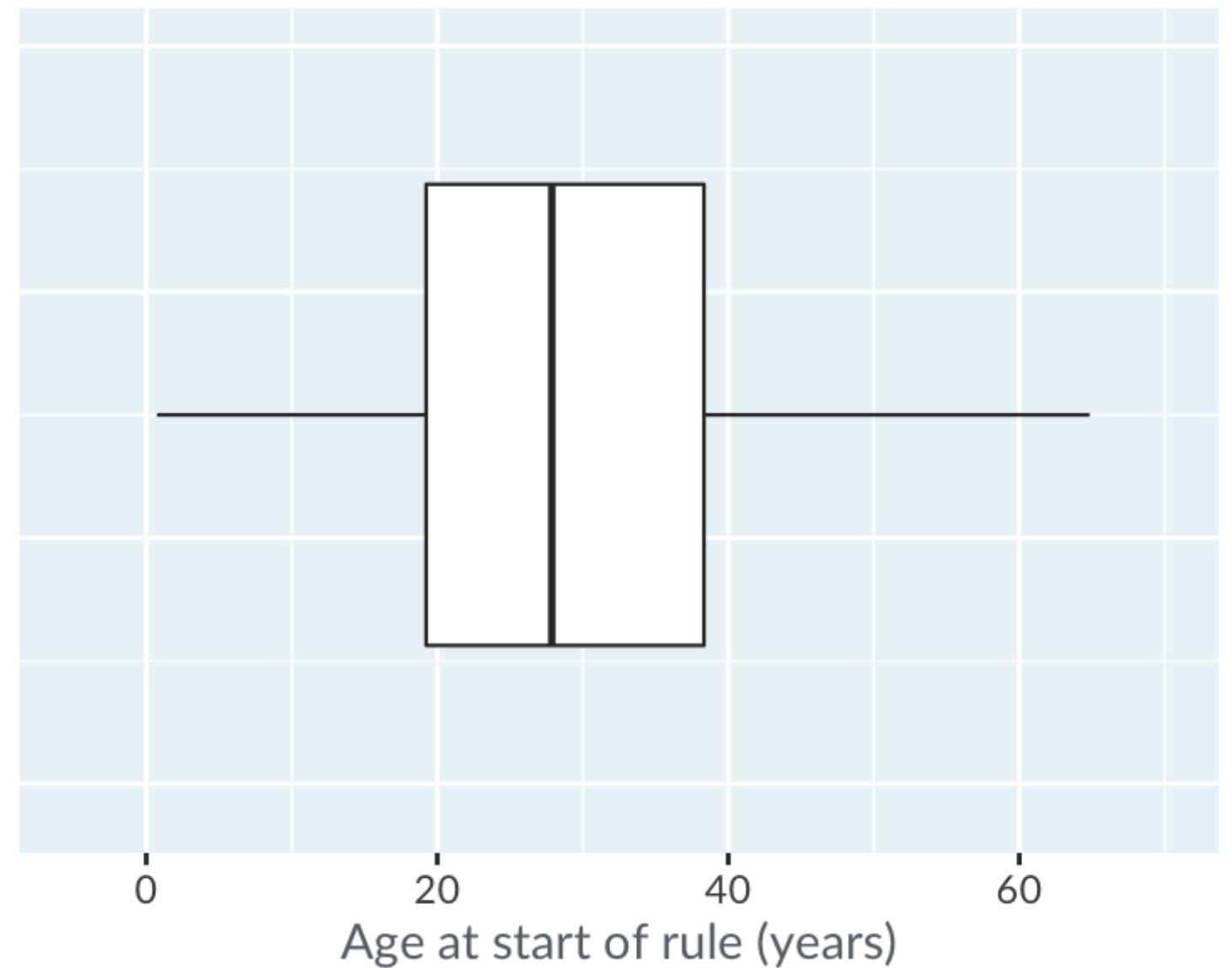
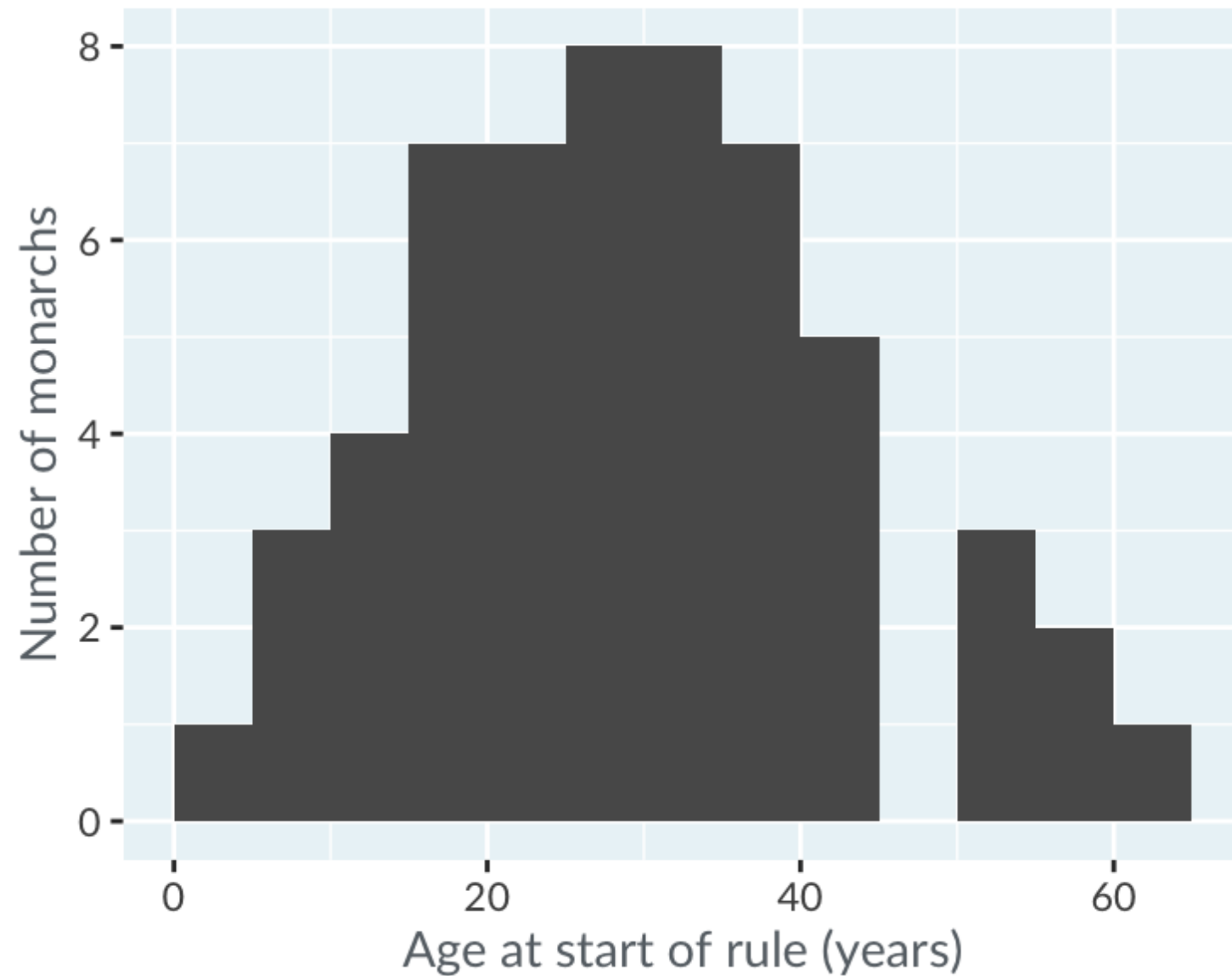




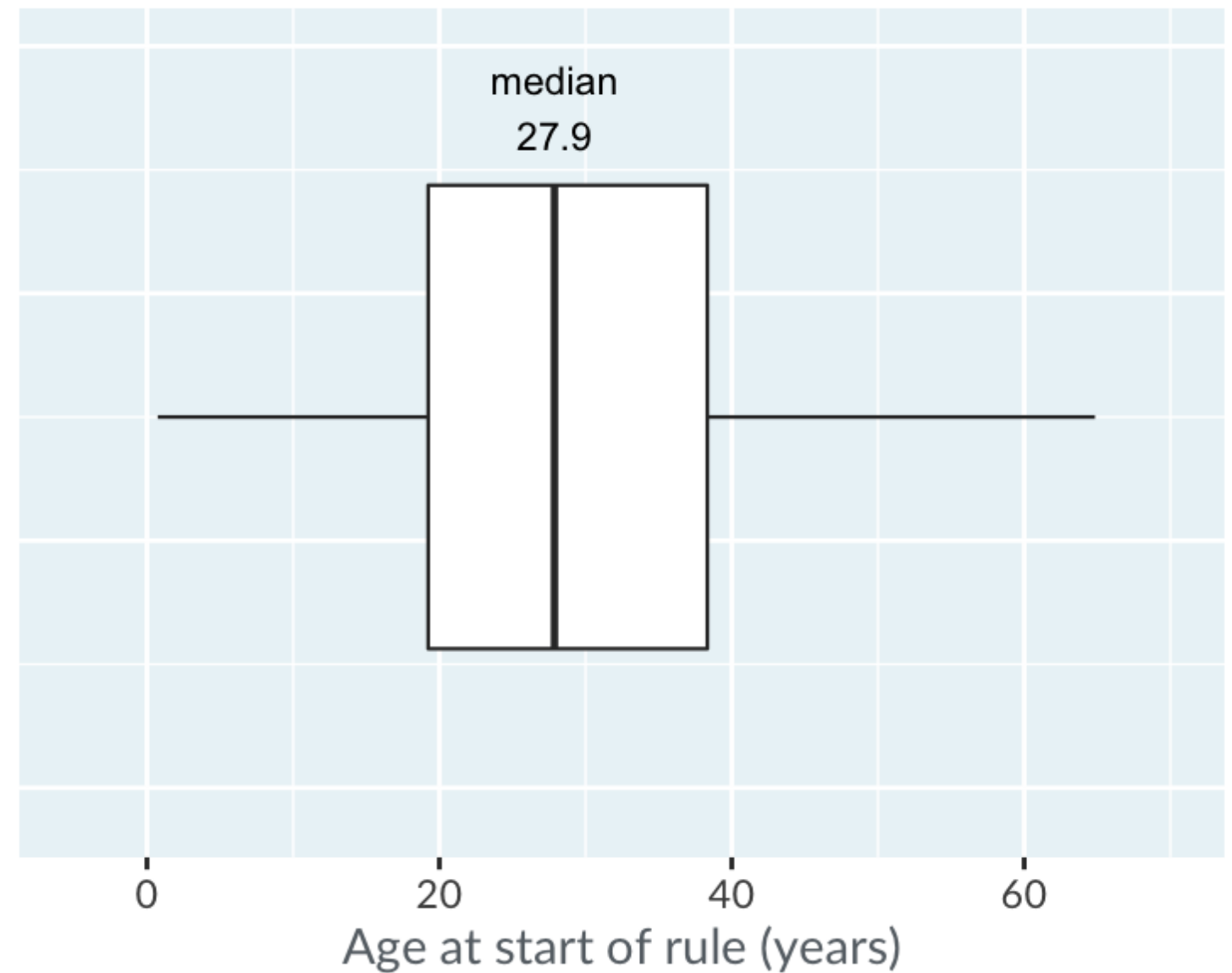
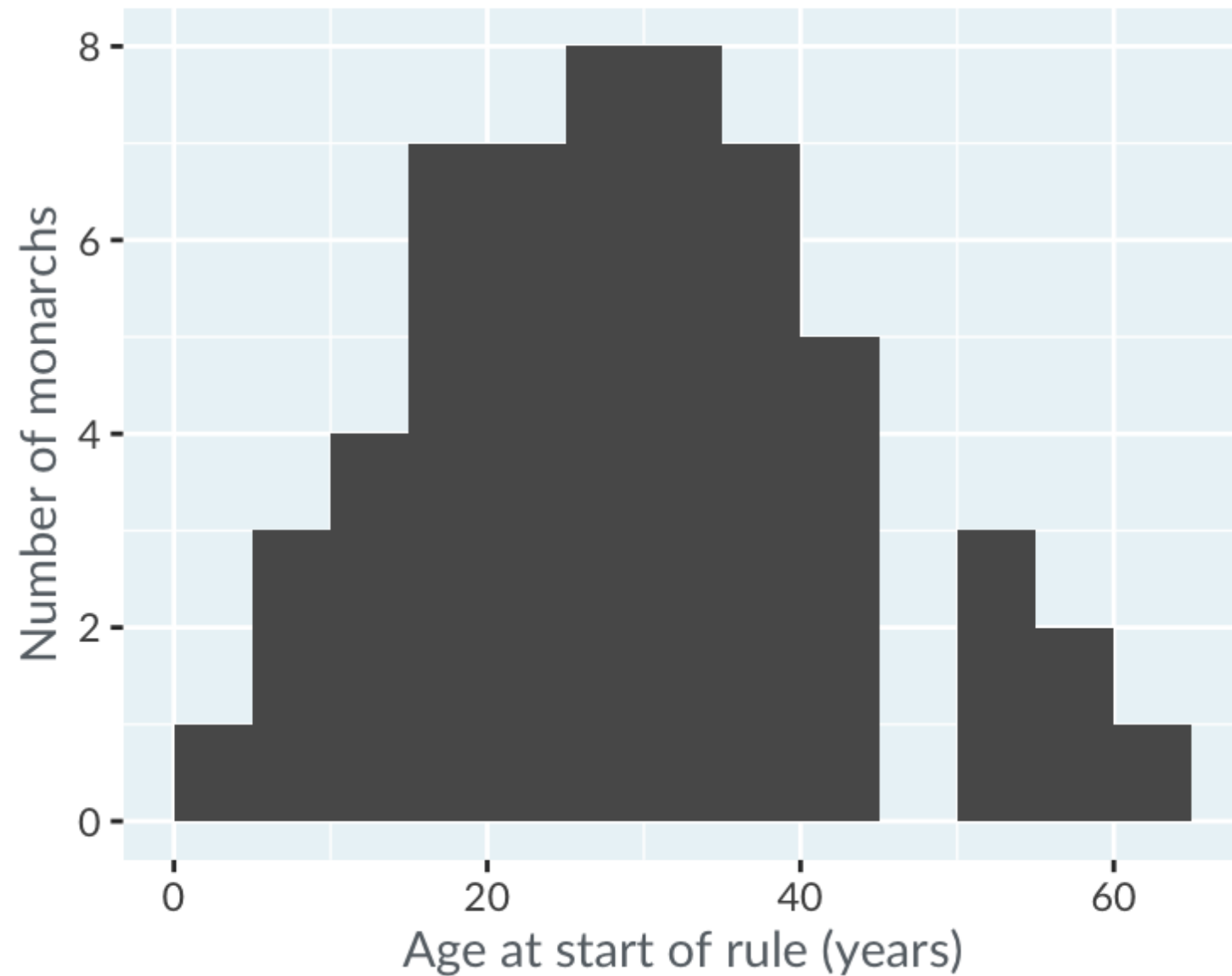
When should you use a box plot?

1. When you have a continuous variable, split by a categorical variable.
2. When you want to compare the distributions of the continuous variable for each category.

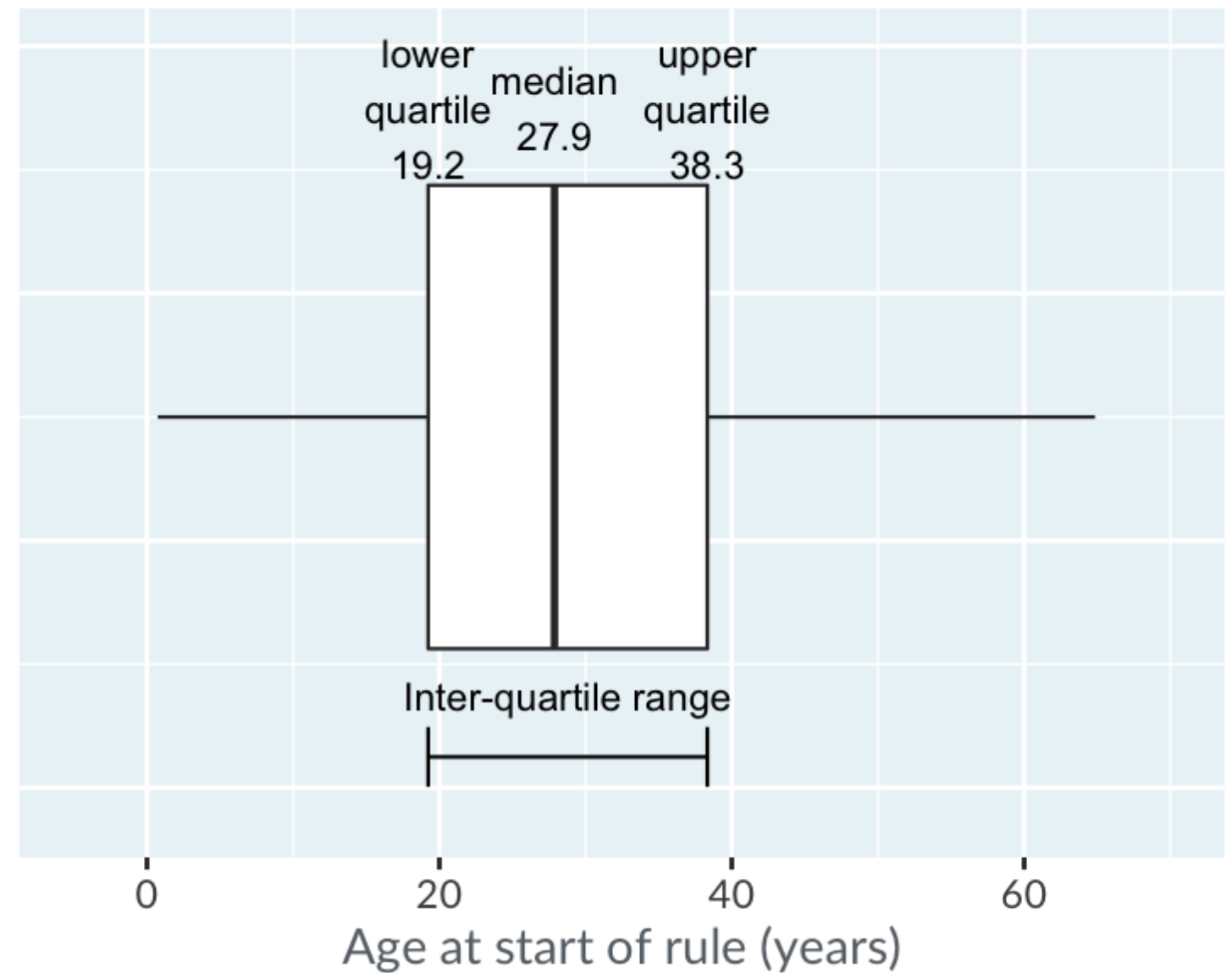
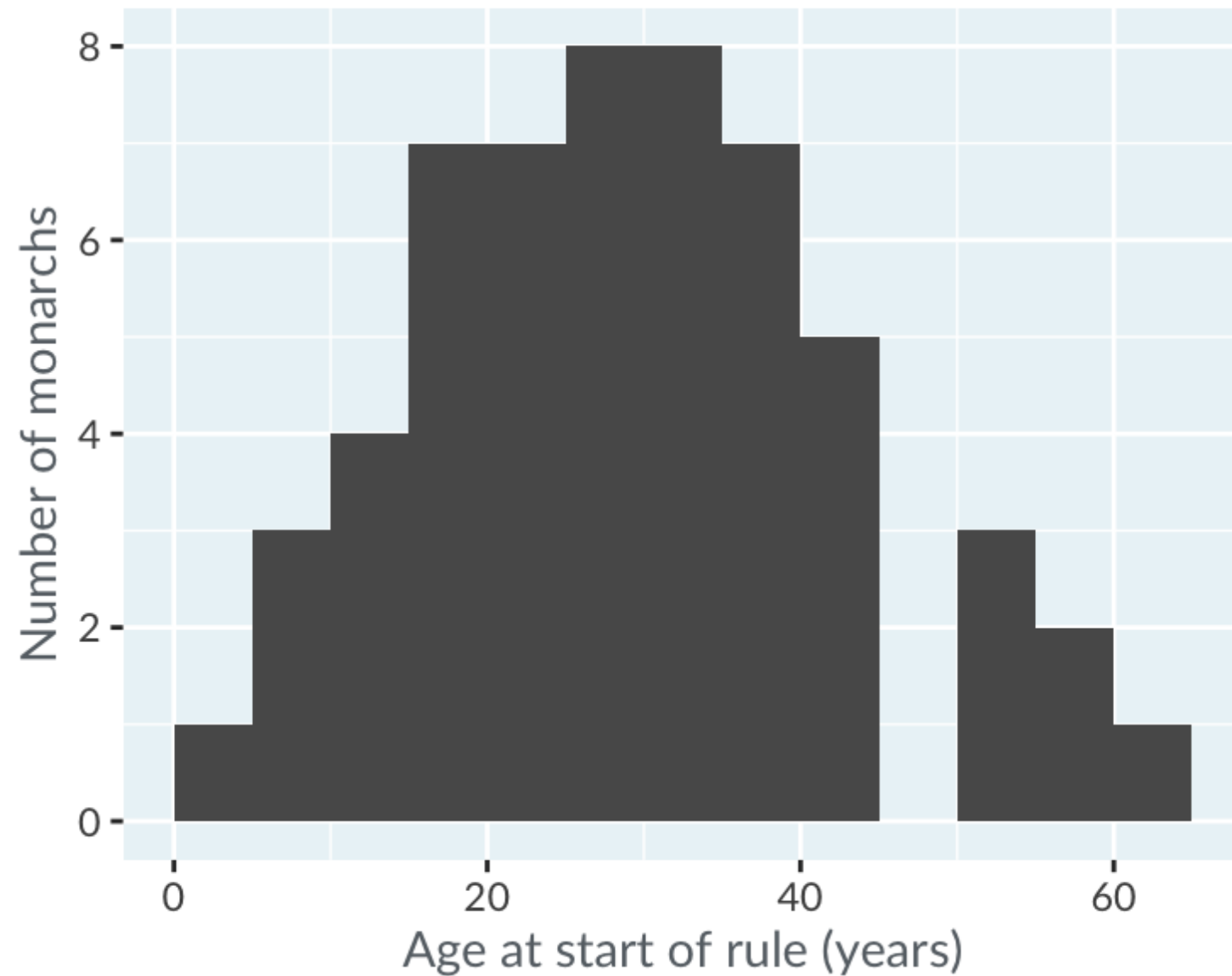
Histogram vs. box plot



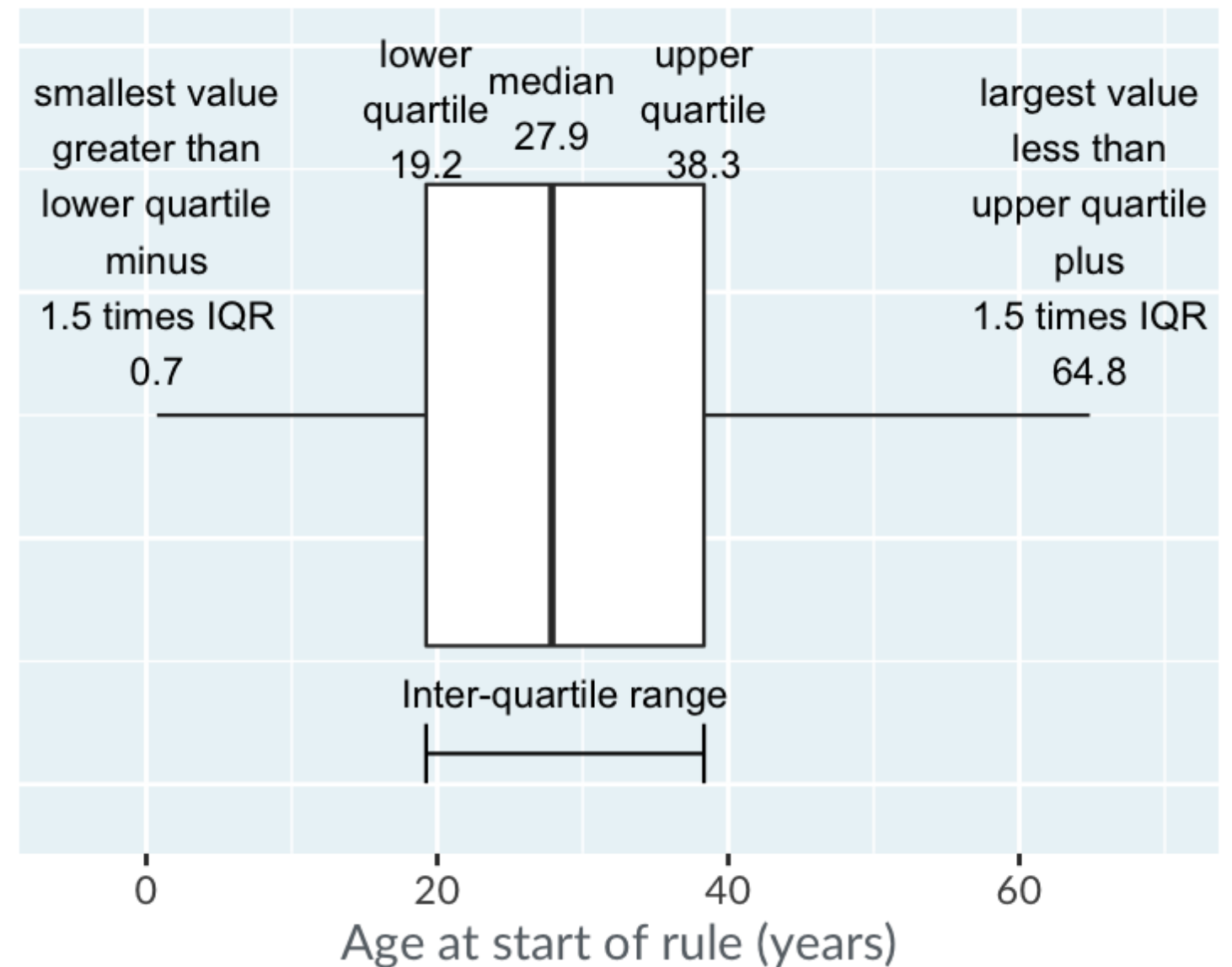
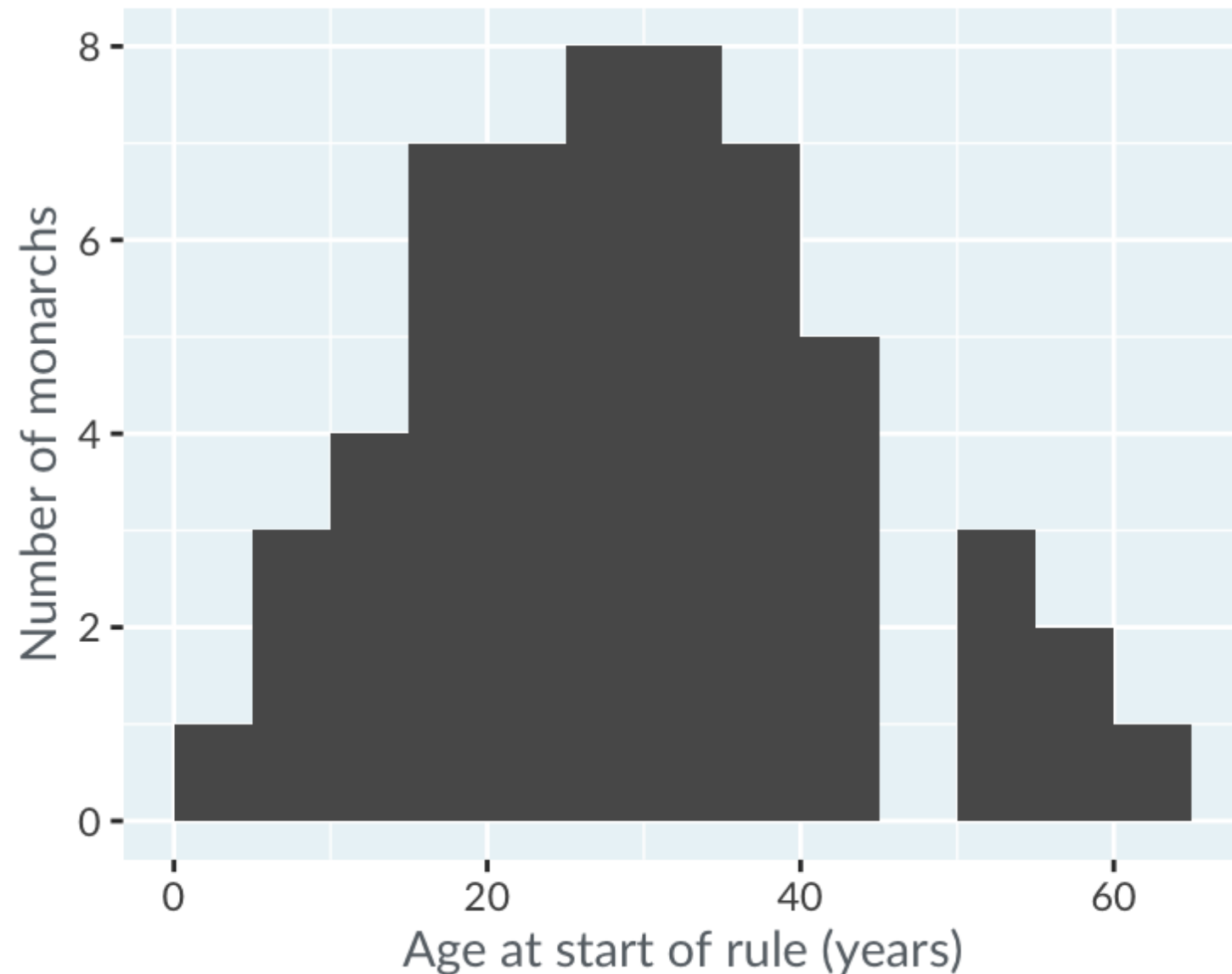
Histogram vs. box plot: mid-line



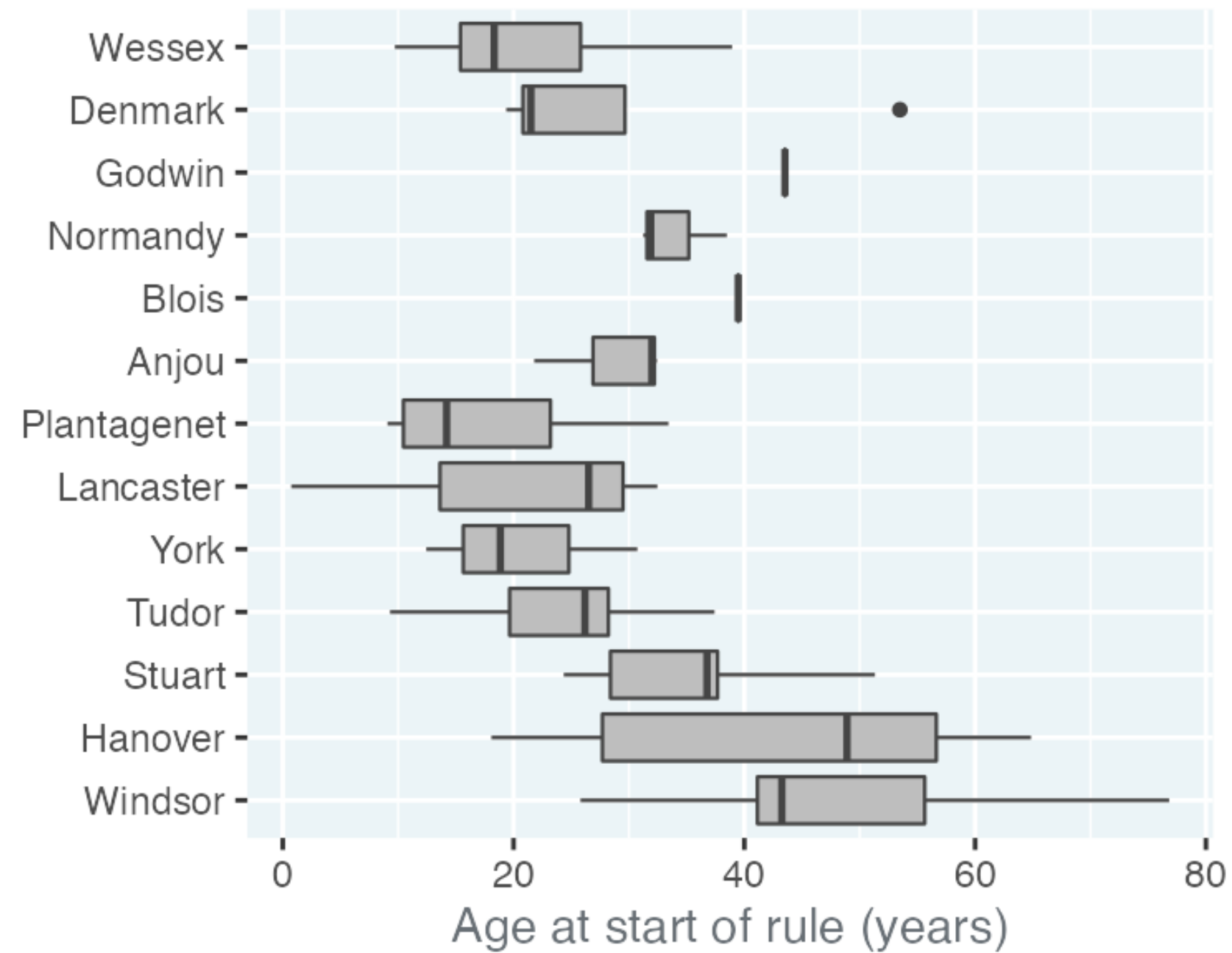
Histograms vs. box plot: the box



Histograms vs. box plots: the whiskers



Monarchs by house



Let's practice!

UNDERSTANDING DATA VISUALIZATION