COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 6: Histograms and Bar Charts

Sept. 3, 2025

February 9, 2026

Logistics

- My office hours will be on **Wednesday** from 3-5 this week
  - They’re normally Monday 3-5 but I had a last minute conflict
  - Elena will still be having her 1:30-3 office hours today
- HW 1 is due Wednesday at 11:59pm
  - Please remember to submit it as a **.ipynb** file
  - HW 2 is out and due next week Wednesday

Last Time: Attributes

Types of Attributes

- Attributes are the names of columns in tables
- All values in a column should be the same type and comparable to each other
    - **Numerical:** Values are on a numerical scale (e.g., years)
        - Values are ordered
        - Differences are meaningful
    - **Categorical:** Each value is from a fixed inventory (e.g., material)
        - May not have an ordering
        - Categories are either the same or different

Numerical Attributes

Values that are numbers are not necessarily numerical

- Sometimes people use numbers instead of strings to represent categories
- Example: In US census data, `SEX` code is (0, 1, 2)
  - Arithmetic on these “numbers” is meaningless
  - The variable `SEX` is still categorical even though numbers were used for the categories

Is this a numerical or categorical attribute?

Speed Skating: The **time** to skate 3000m (in min and sec)

[3:54, 3:56, 3:59, 4:01]

Is this a numerical or categorical attribute?

Skating: **Events that figure skaters can perform in**

[‘single skating’, ‘single skating’, ’pair skating’, ‘ice dance’, ‘team event’]

Is this a numerical or categorical attribute?

Speed Skating: **Race distances (in meters)**

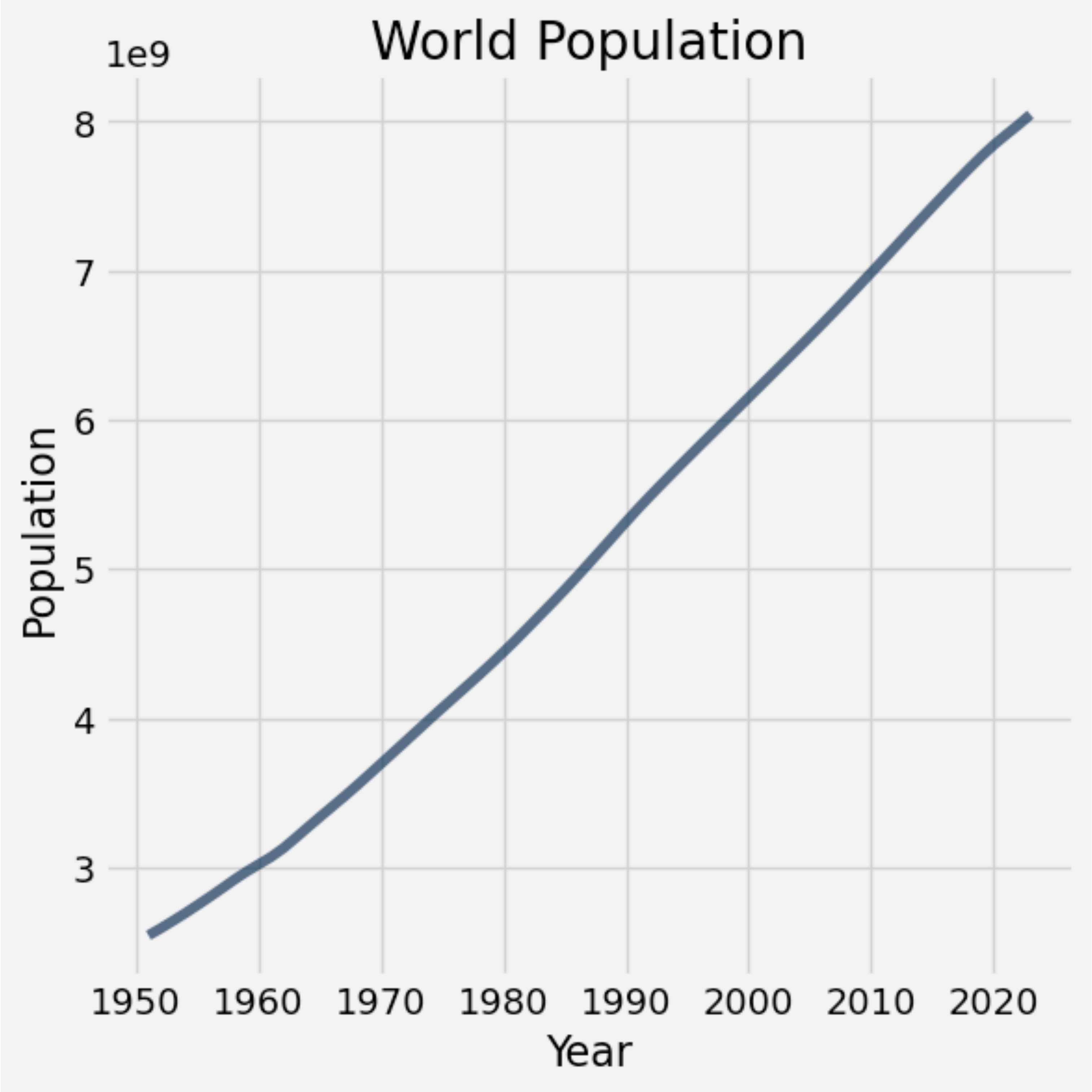
[500, 1000, 1500, 3000, 5000]

Visualizing Relationships between Numerical Values

Line Plots and Scatter Plots

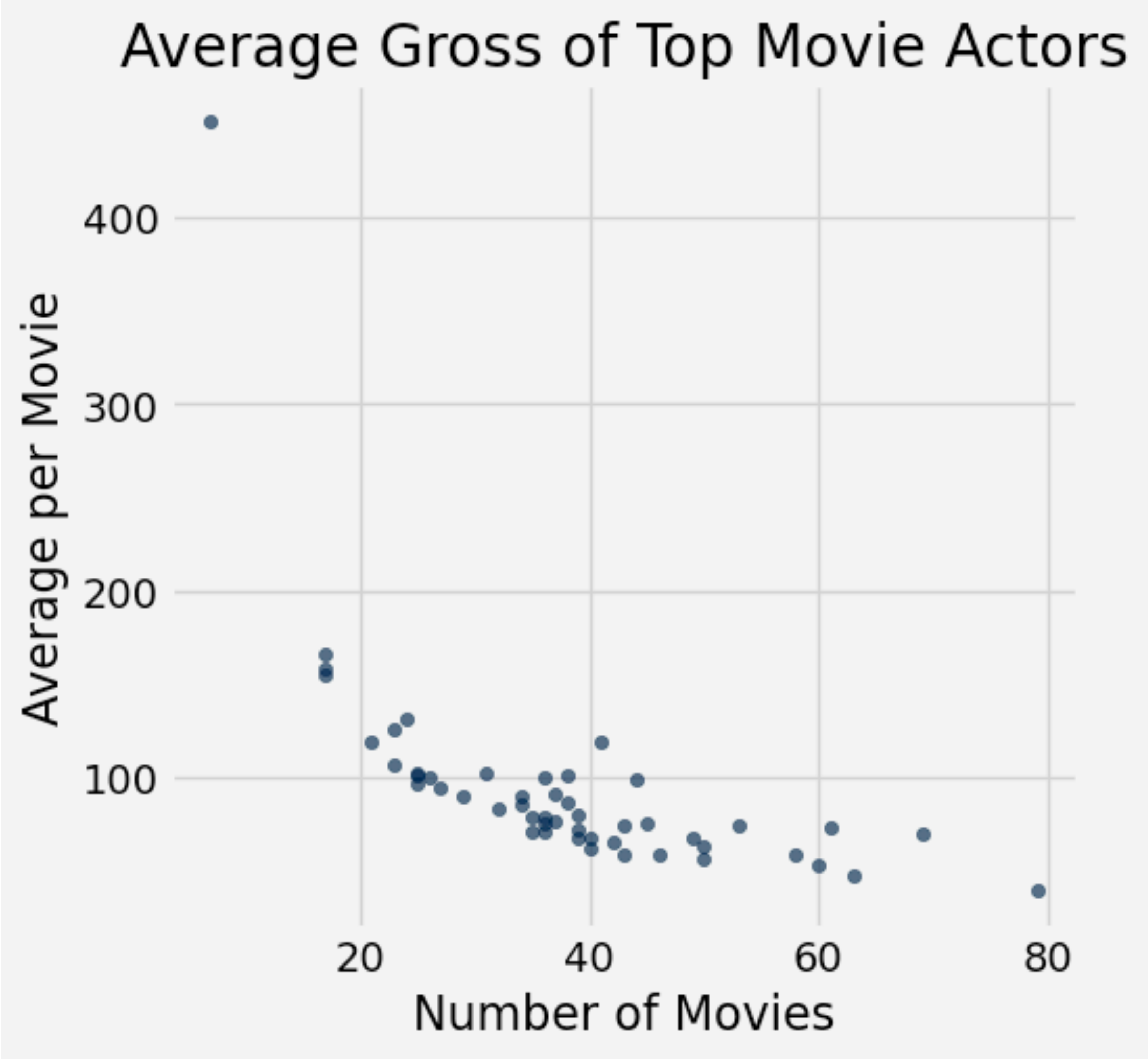
Line Plot

`plot`



Scatter Plot

`scatter`



Line Plots vs Scatter Plots

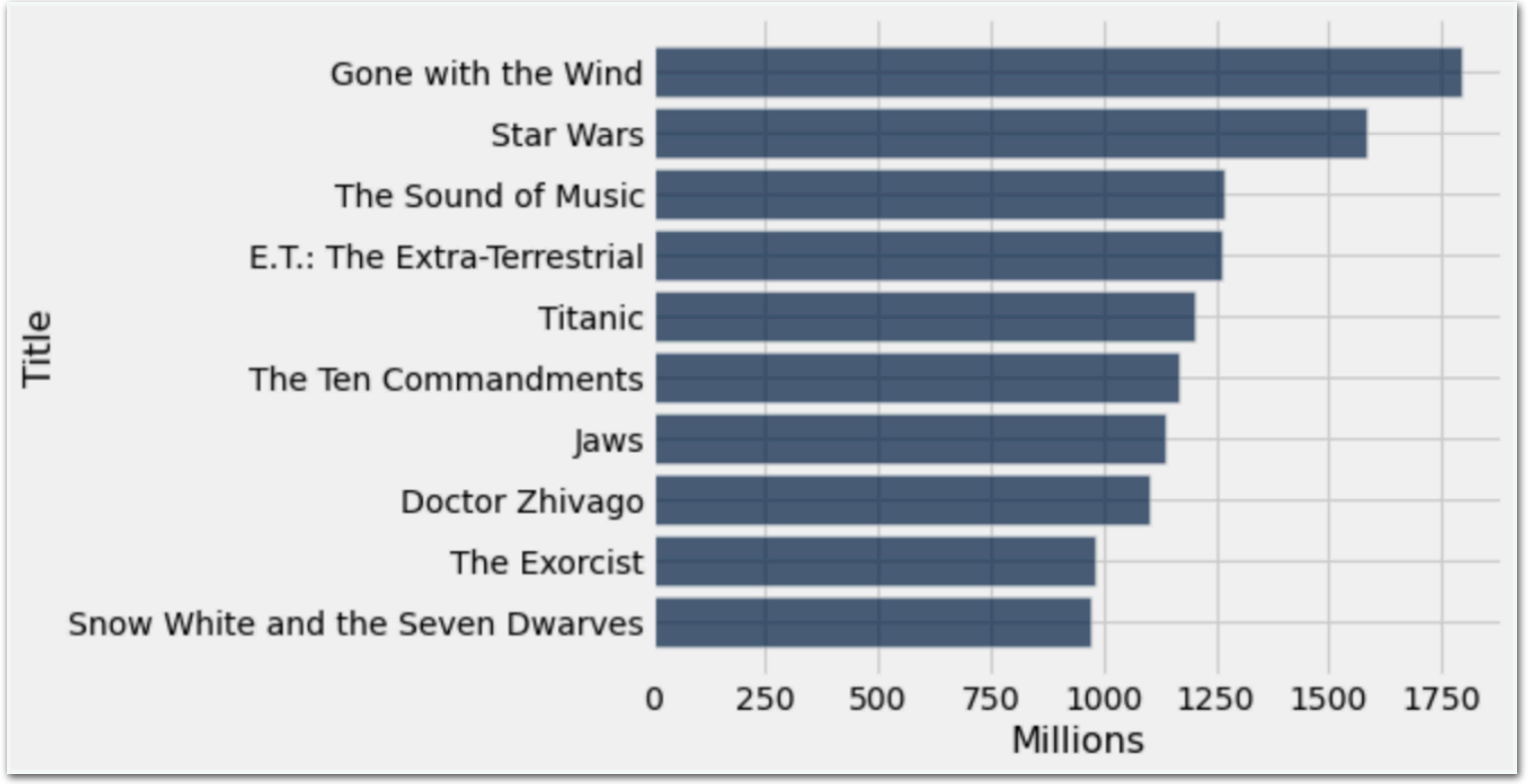
- **Line plots** are good for sequential data if
  - x-axis has an order (e.g., time, years, distance)
  - sequential differences in y value are meaningful
  - there's only one y-value for each x-value
- Use **scatter plot** for non-sequential quantitative data
  - great for looking for associations

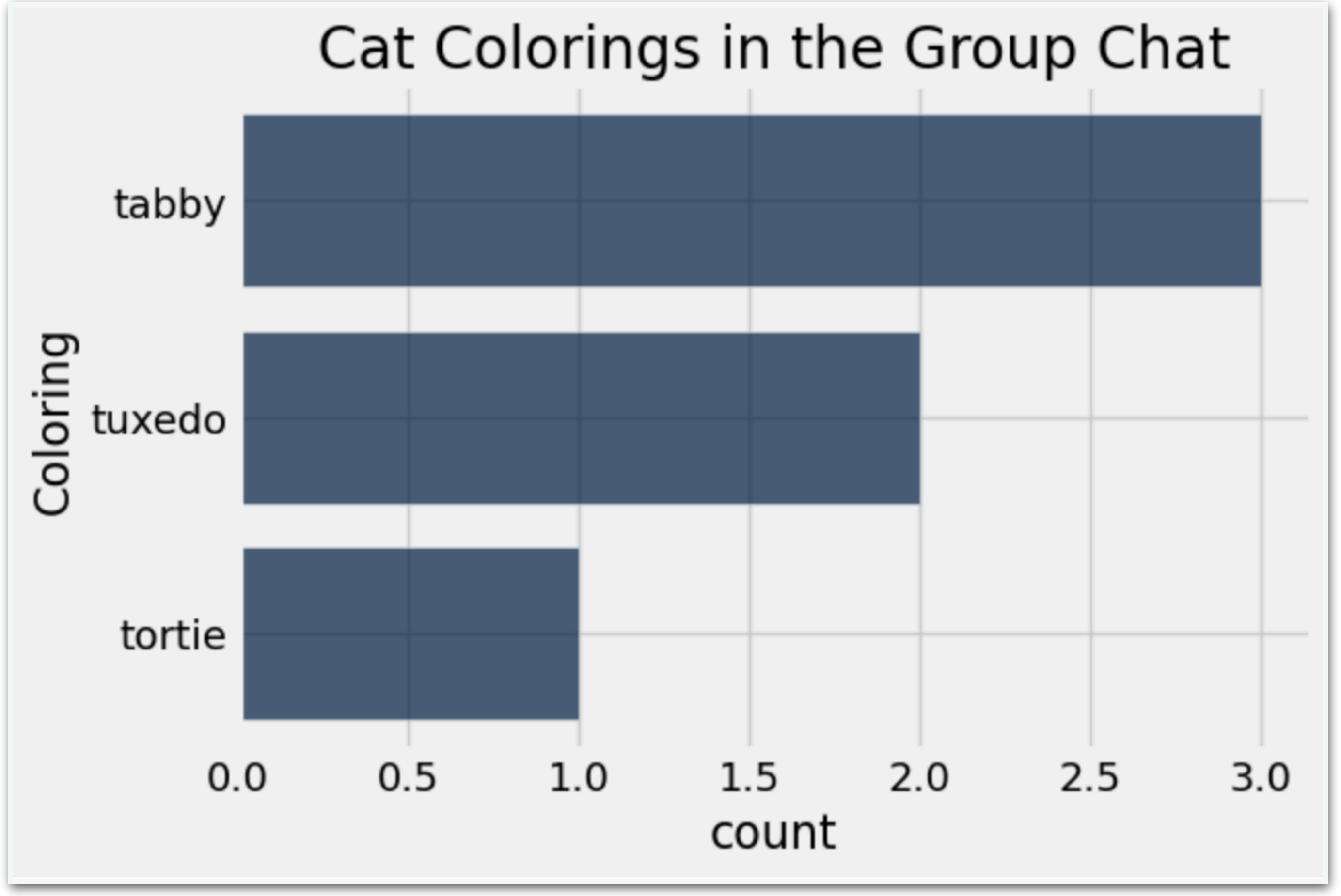
Bar Charts:
Categorical Distributions

Bar Plots

Bar charts are good for visualizing:

- relationships between categorical variable(s) and a numerical variable
- a categorical distribution
  - Distribution refers to the frequencies of each value
  - Note that individuals will only have exactly one value for each categorical variable

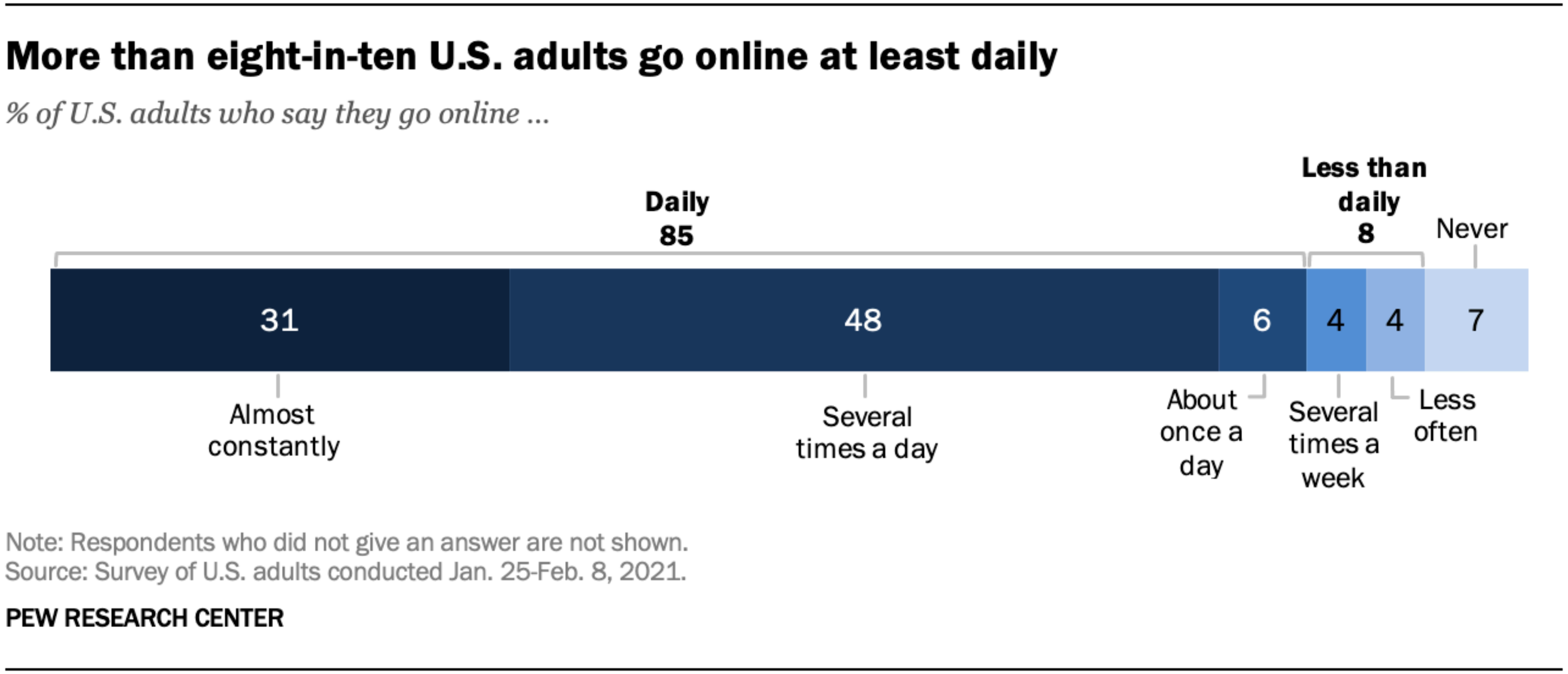




Categorical Distributions

Each individual has exactly one category

- Percents add to 100



More than eight-in-ten U.S. adults go online at least daily

% of U.S. adults who say they go online …

Note: Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

PEW RESEARCH CENTER

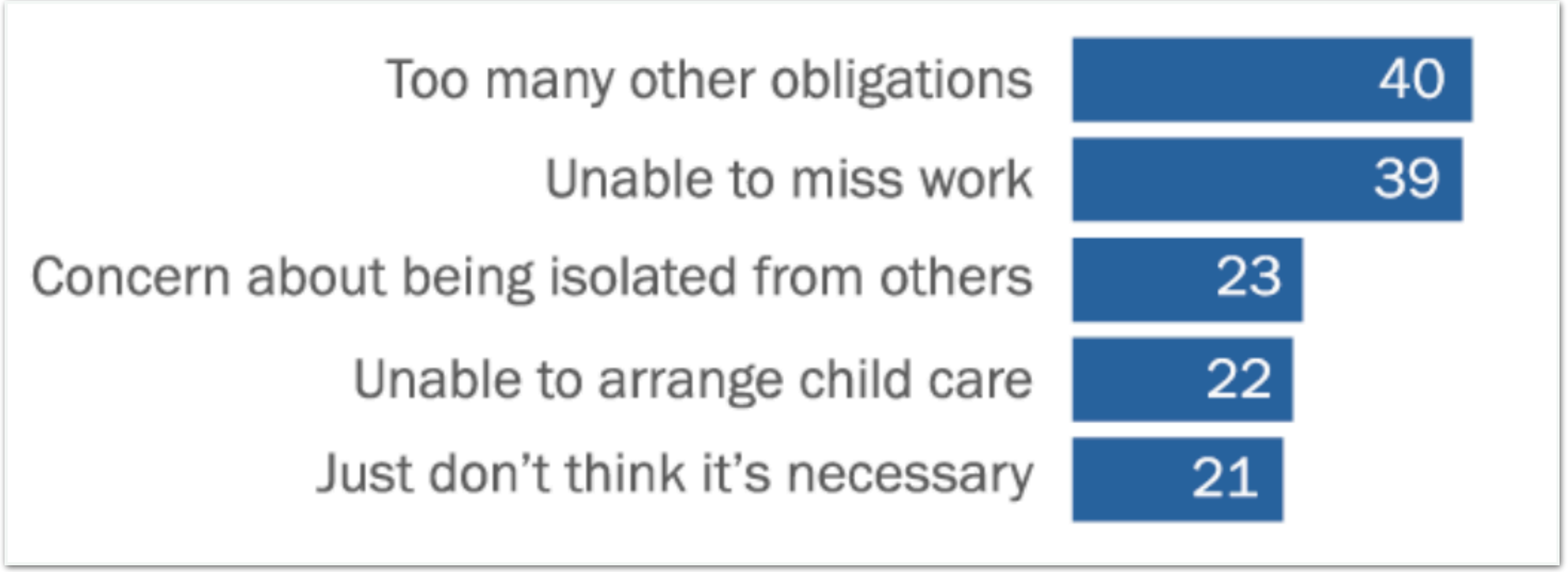
Source: Pew Research

Not a distribution

When individuals can pick more than one answer

- Percents don’t necessarily add up to 100

Survey question: “A major reason you would find it difficult to quarantine for 14 days”

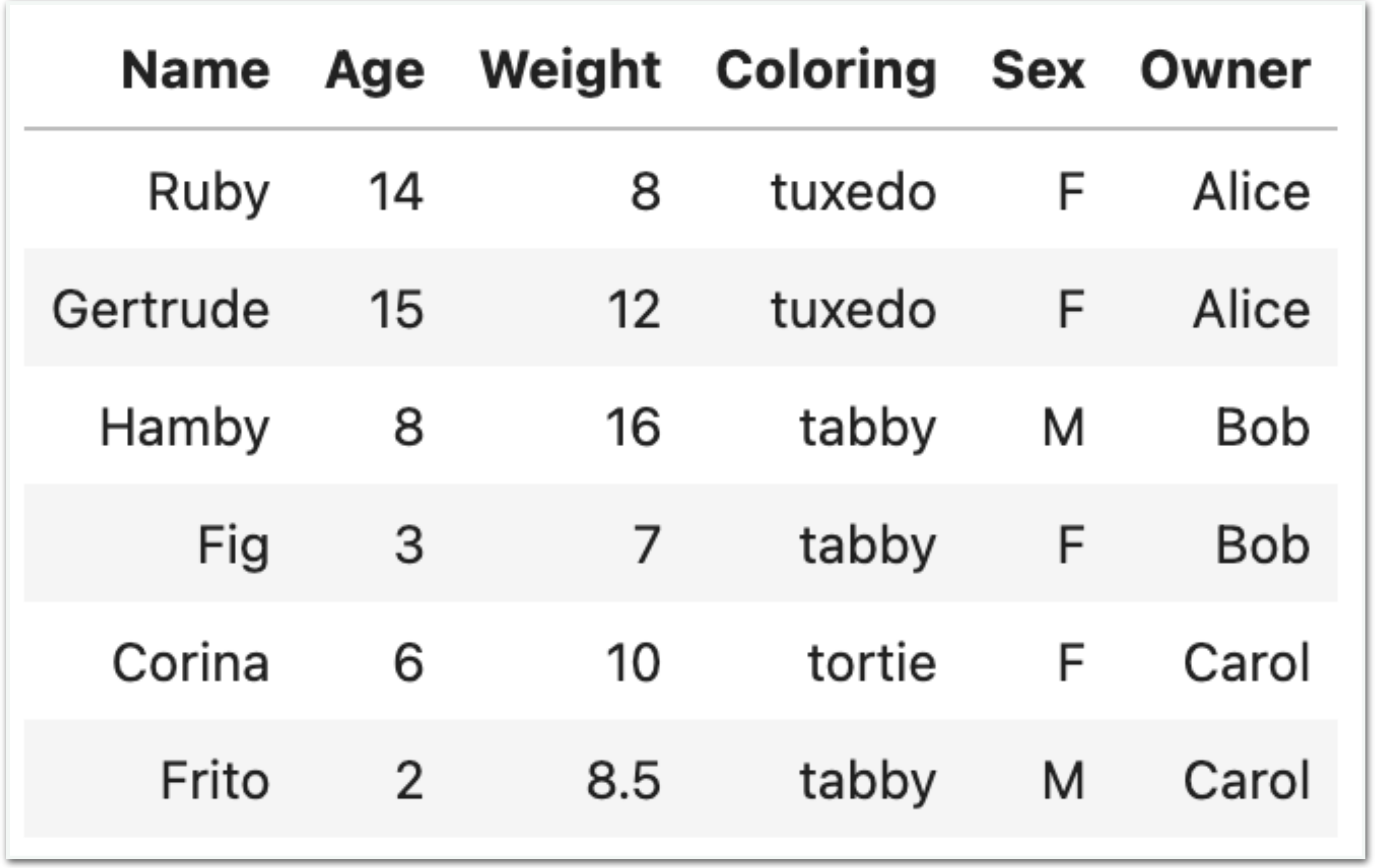


Source: Pew Research

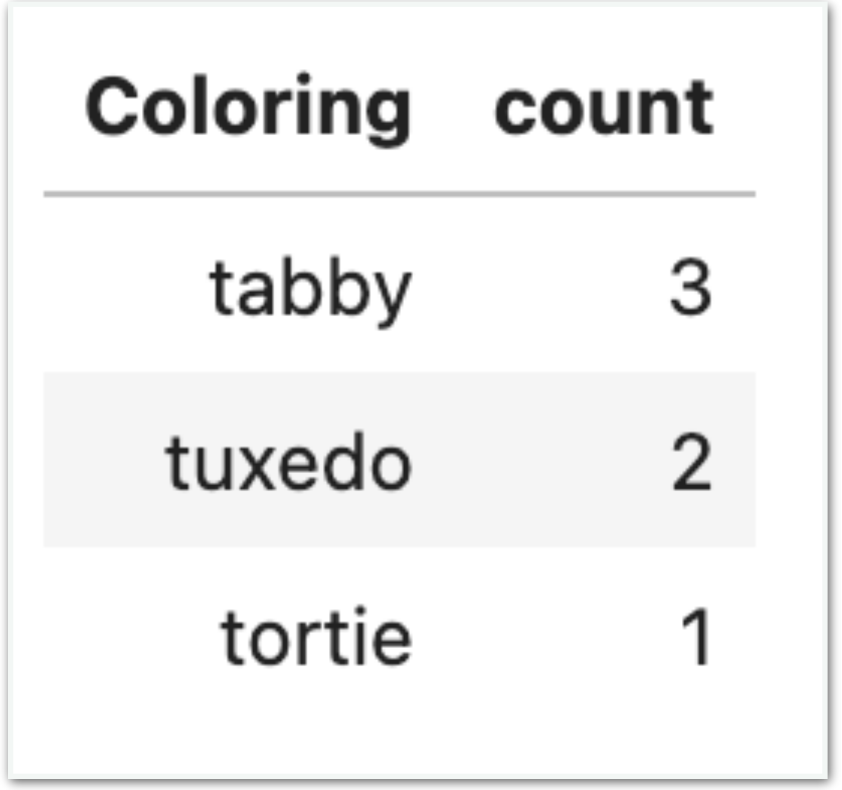
Grouping Categorical Data

Recall the data in the 2025 Cat Census

- If I want to plot the frequency of, say, the different colorings I need some way of aggregating data



Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol

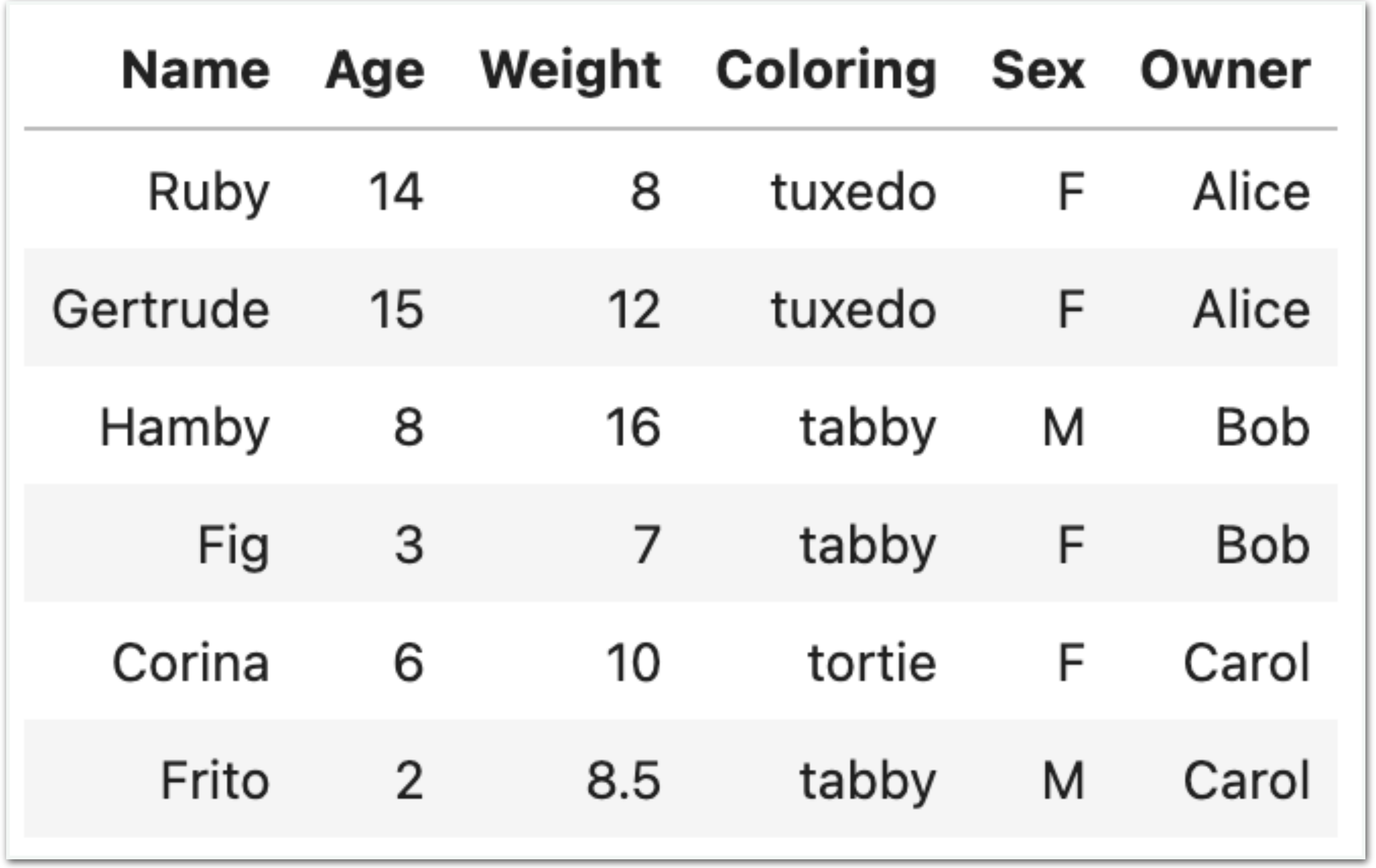


Coloring	count
tabby	3
tuxedo	2
tortie	1

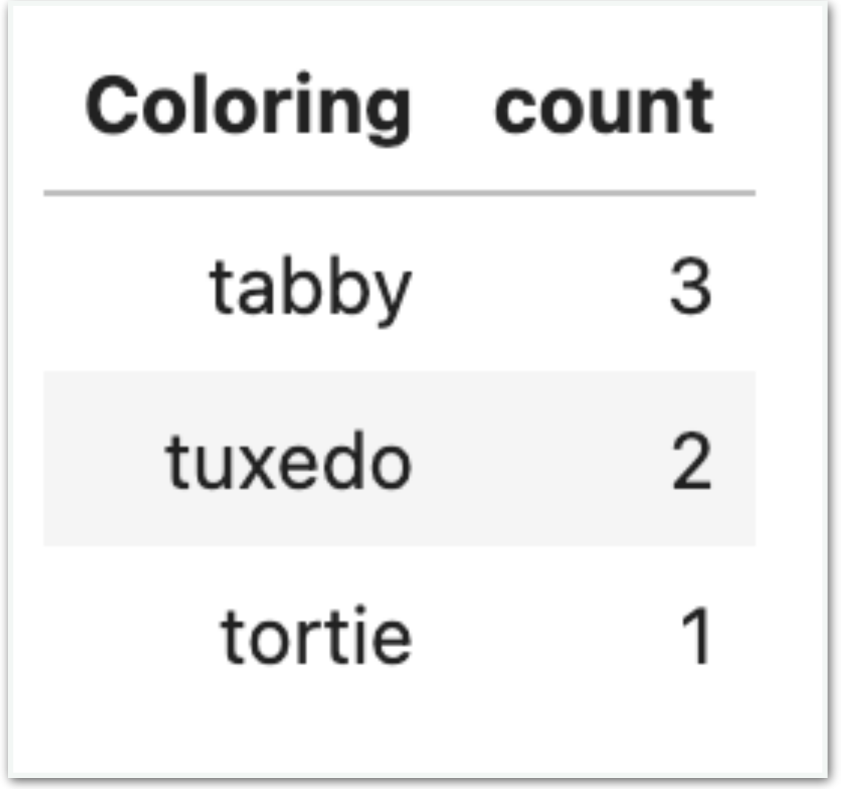
Grouping Categorical Data

Recall the data in the 2025 Cat Census

- If I want to plot the frequency of, say, the different colorings I need some way of aggregating data
- This is where `group` comes in handy
  - This lecture we'll go over grouping by a single column and cover multiple columns next time



Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol



Coloring	count
tabby	3
tuxedo	2
tortie	1

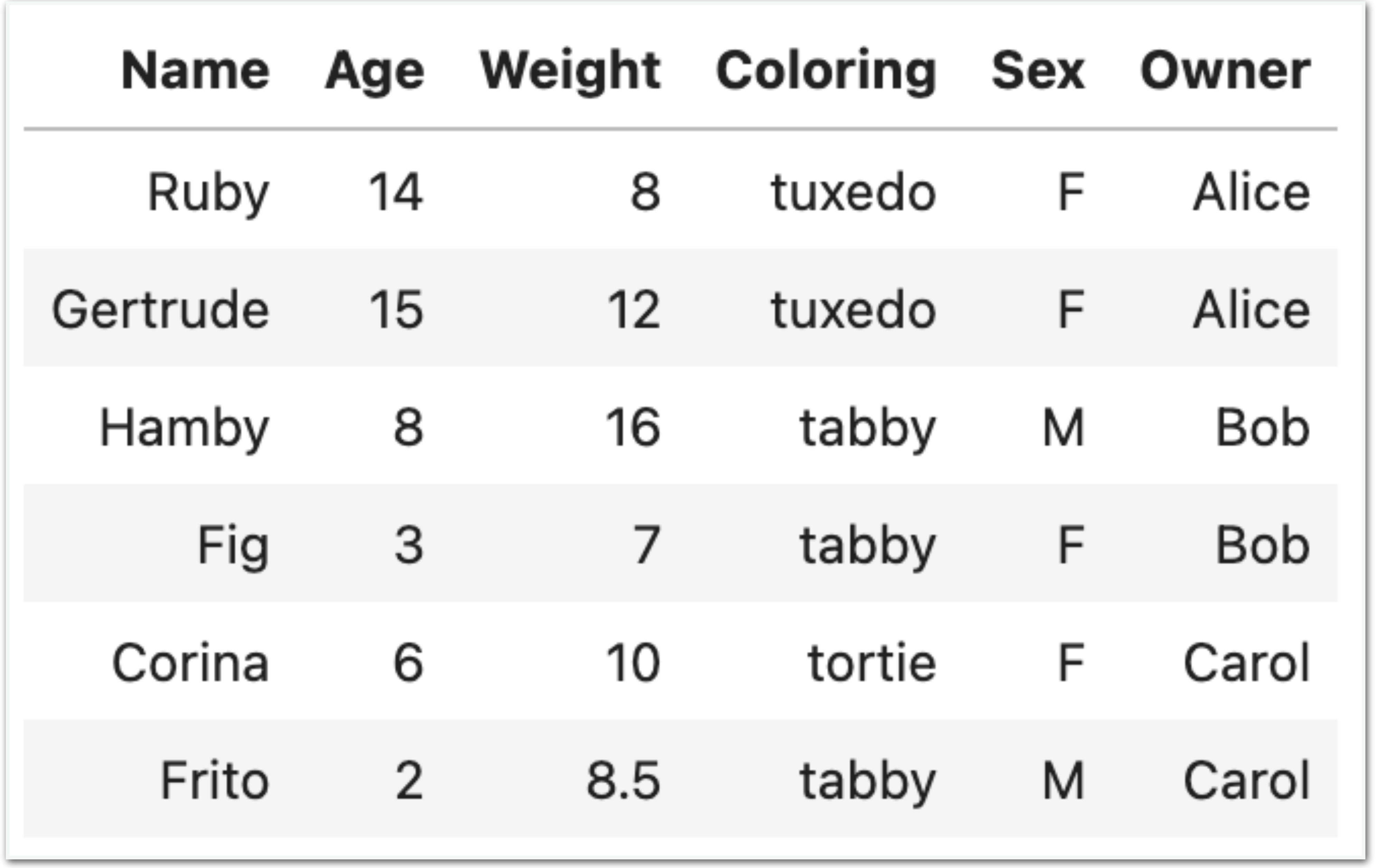
Grouping by a Single Column

The `group` method aggregates all rows with the same value in column `c`

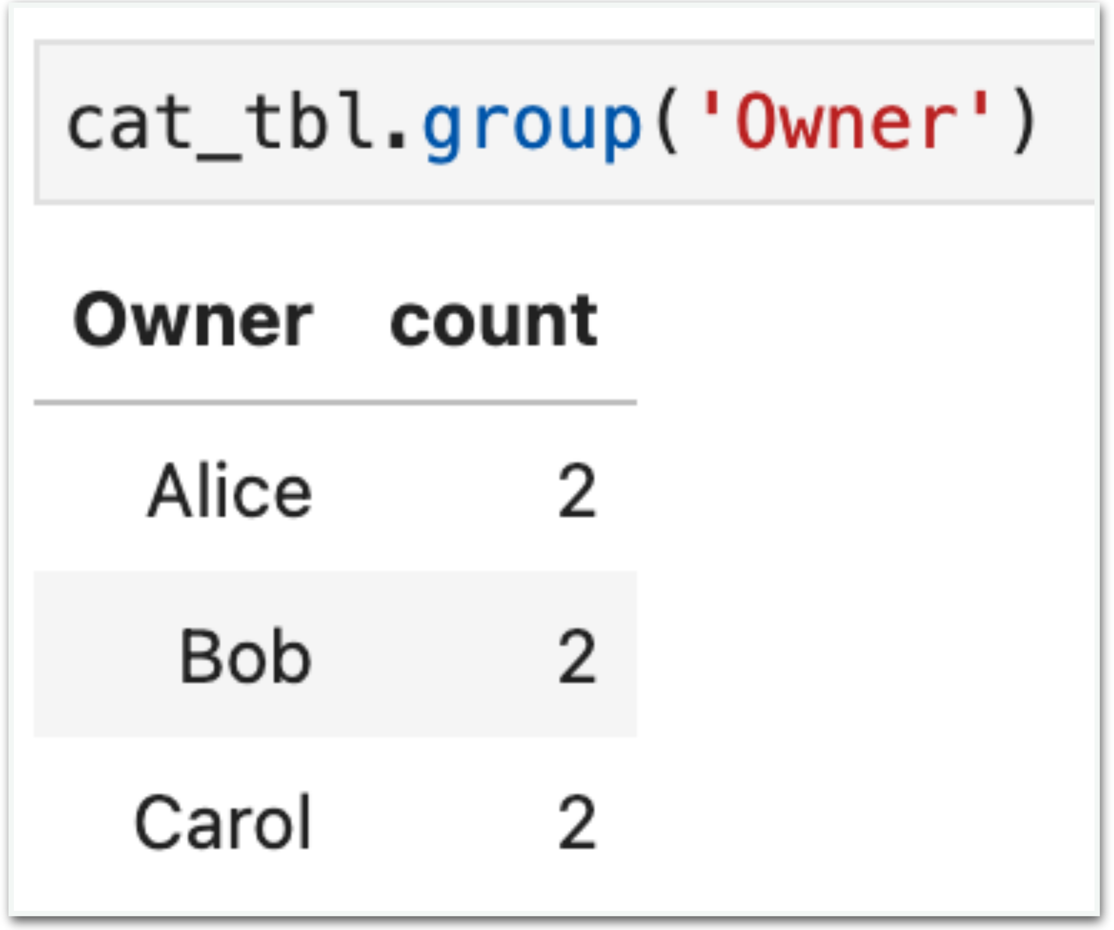
- `tbl.group(c)`
- `tbl.group(c, func)`

`group` can optionally apply `func` to grouped values, for example:

- `len`: count of grouped values (default)
- `list`: list of all grouped values
- `sum`: total of all grouped values



Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol



```
cat_tbl.group('Owner')
```

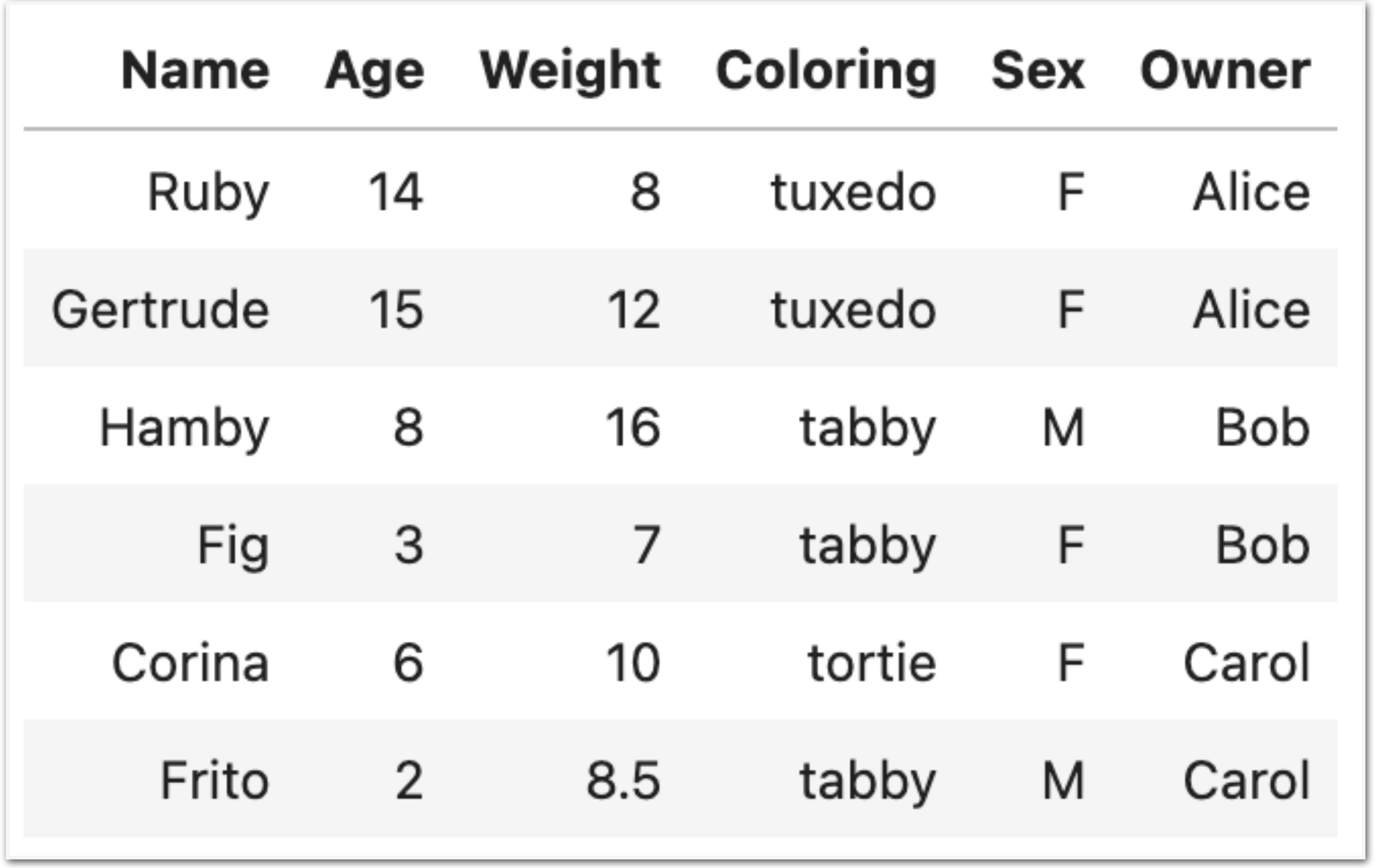
Owner	count
Alice	2
Bob	2
Carol	2

Grouping by a Single Column

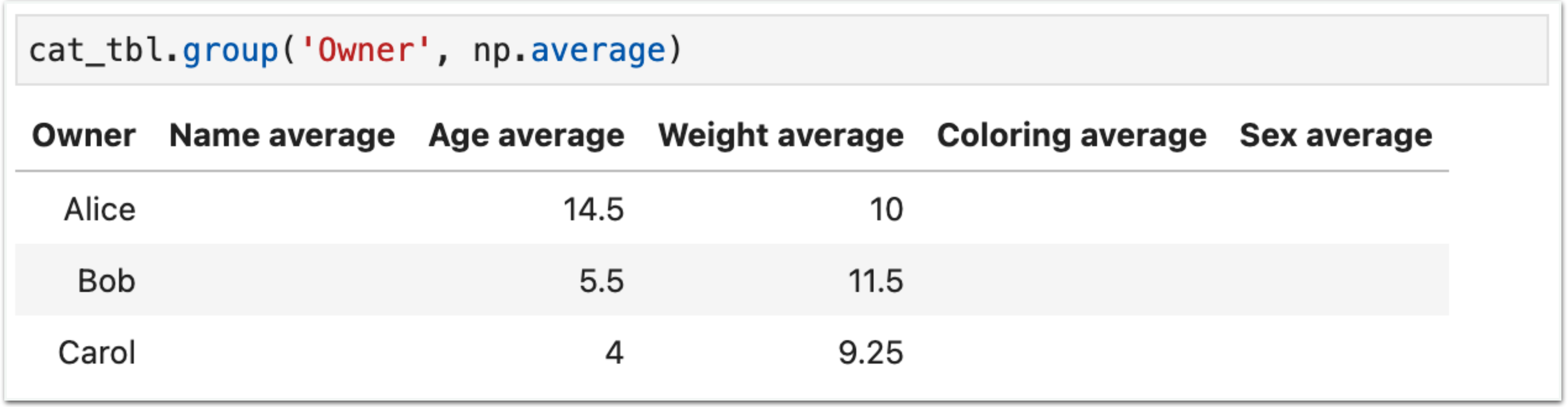
The `group` method aggregates all rows with the same value in column `c`

- `tbl.group(c)`
- `tbl.group(c, func)`

`group` can optionally apply `func` to grouped values

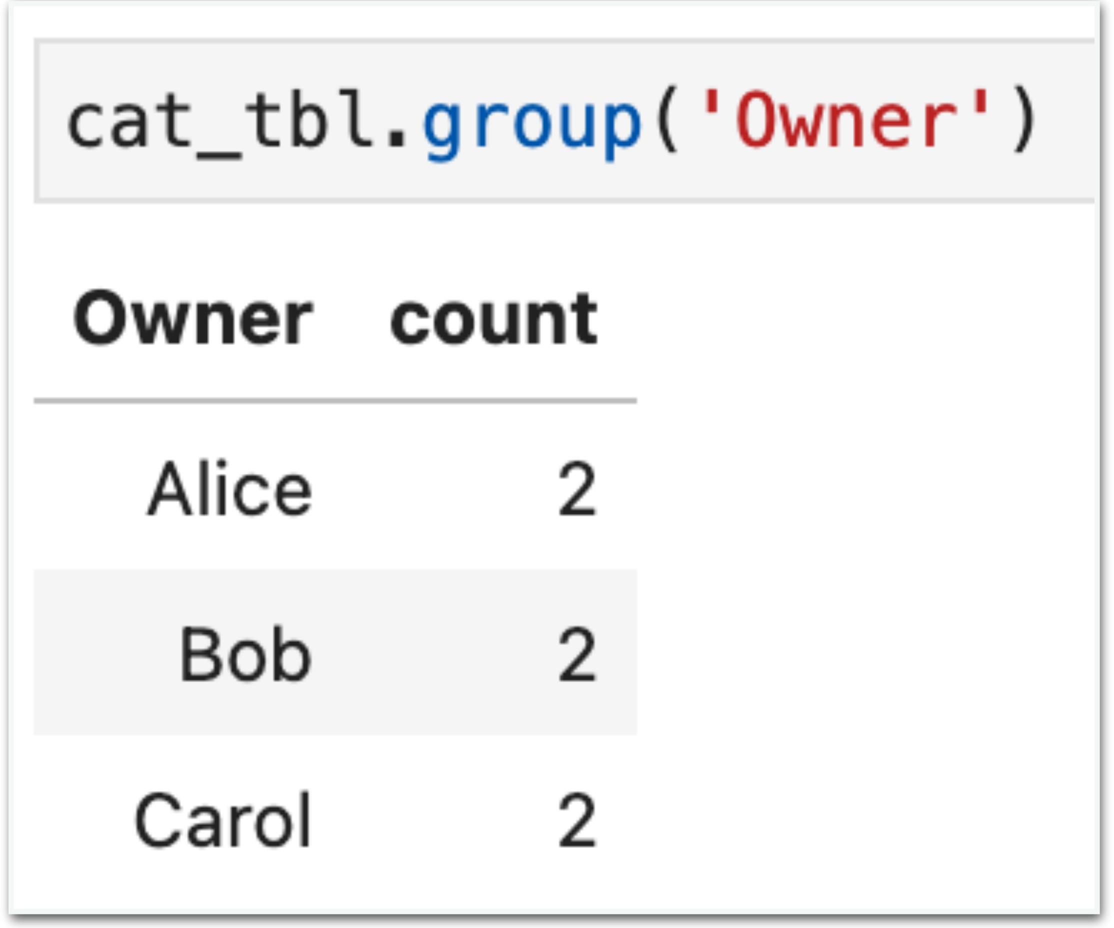


Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol



```
cat_tbl.group('Owner', np.average)
```

Owner	Name average	Age average	Weight average	Coloring average	Sex average
Alice		14.5	10		
Bob		5.5	11.5		
Carol		4	9.25		



```
cat_tbl.group('Owner')
```

Owner	count
Alice	2
Bob	2
Carol	2

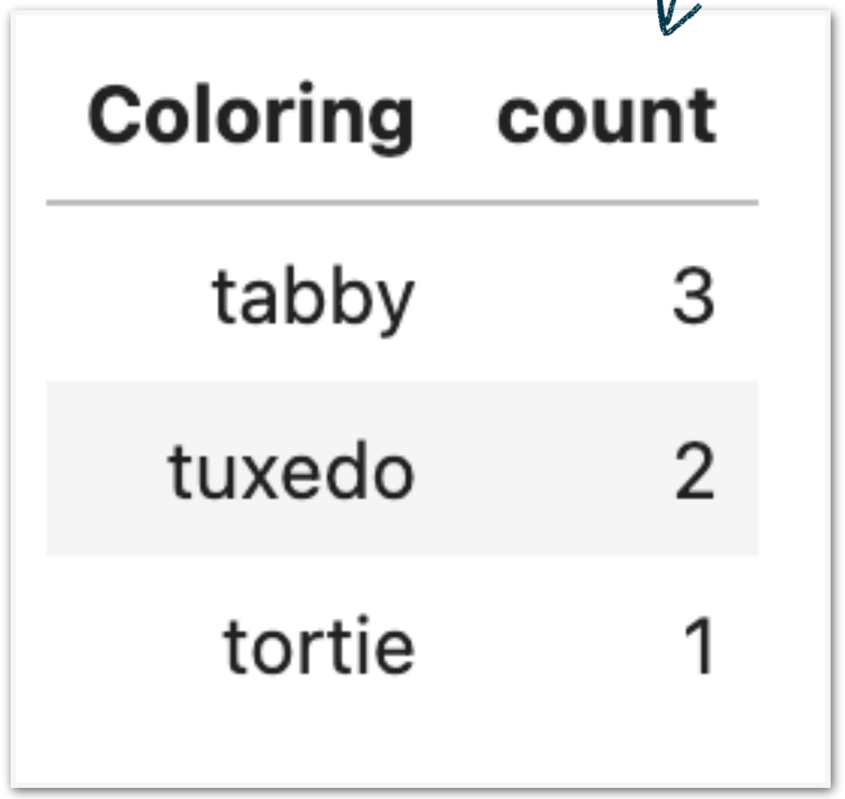
Bar Chart: Categorical Distributions

Bar charts: display categorical variables and frequencies

- One bar for each category
- Ordering of the bar can be specified (e.g. `.sort`)
- Length of bar is proportional to the frequency of the category

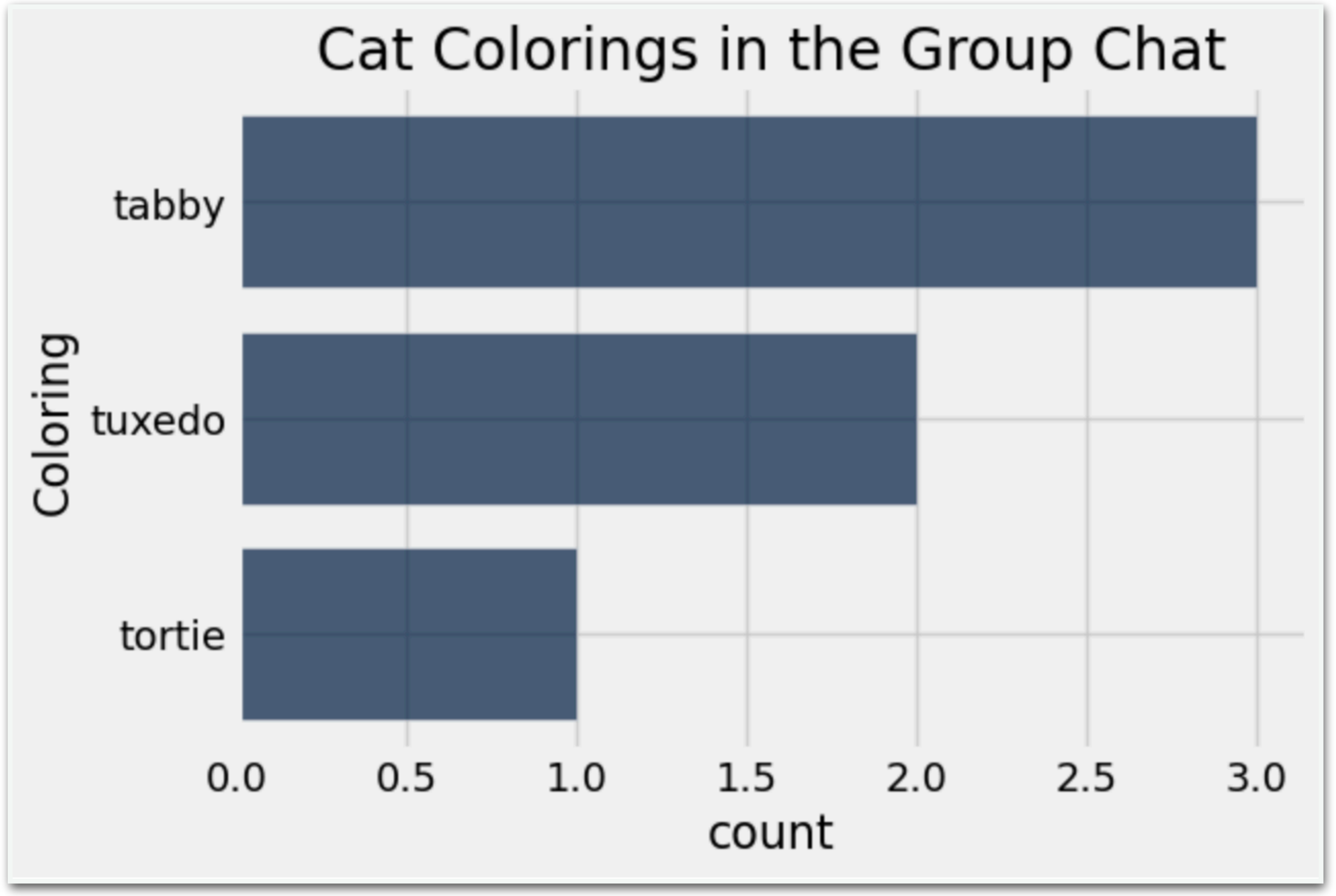
category

frequency



Coloring	count
tabby	3
tuxedo	2
tortie	1

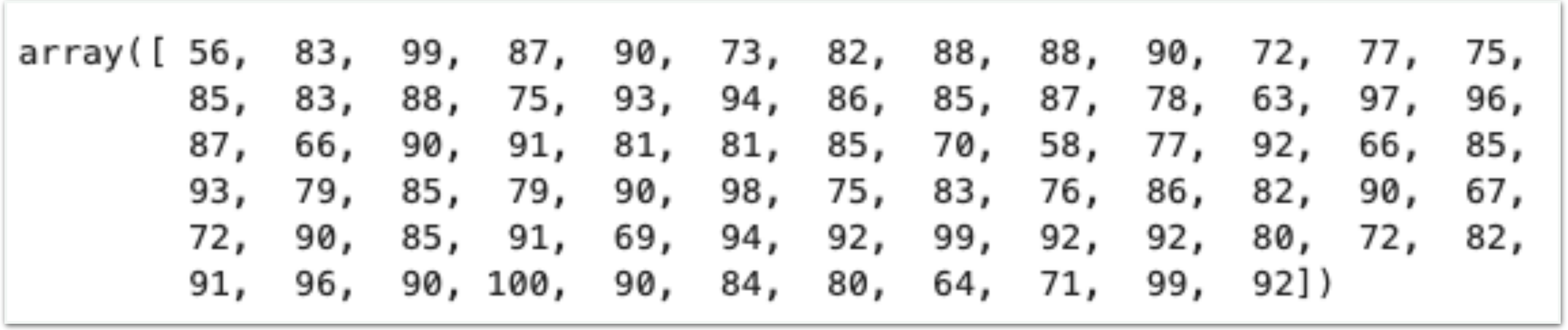
```
tbl.barh(category_label, freq_label)
– cat_tbl.barh(‘Coloring’, ’count’)
– cat_tbl.barh(‘Coloring')
```



Histograms:
Visualizing Numerical Distributions

Visualizing Numerical Distributions

Let’s say we have a data set containing grades students scored on an exam:



```
array([ 56,  83,  99,  87,  90,  73,  82,  88,  88,  90,  72,  77,  75,
        85,  83,  88,  75,  93,  94,  86,  85,  87,  78,  63,  97,  96,
        87,  66,  90,  91,  81,  81,  85,  70,  58,  77,  92,  66,  85,
        93,  79,  85,  79,  90,  98,  75,  83,  76,  86,  82,  90,  67,
        72,  90,  85,  91,  69,  94,  92,  99,  92,  92,  80,  72,  82,
        91,  96,  90, 100,  90,  84,  80,  64,  71,  99,  92])
```

What if we want to know generally how students on the exam?

- How many students got between 90 and 100?
- What range of values did the majority of students fall into?

Visualizing Numerical Distributions

Histograms display the distribution of a numerical value

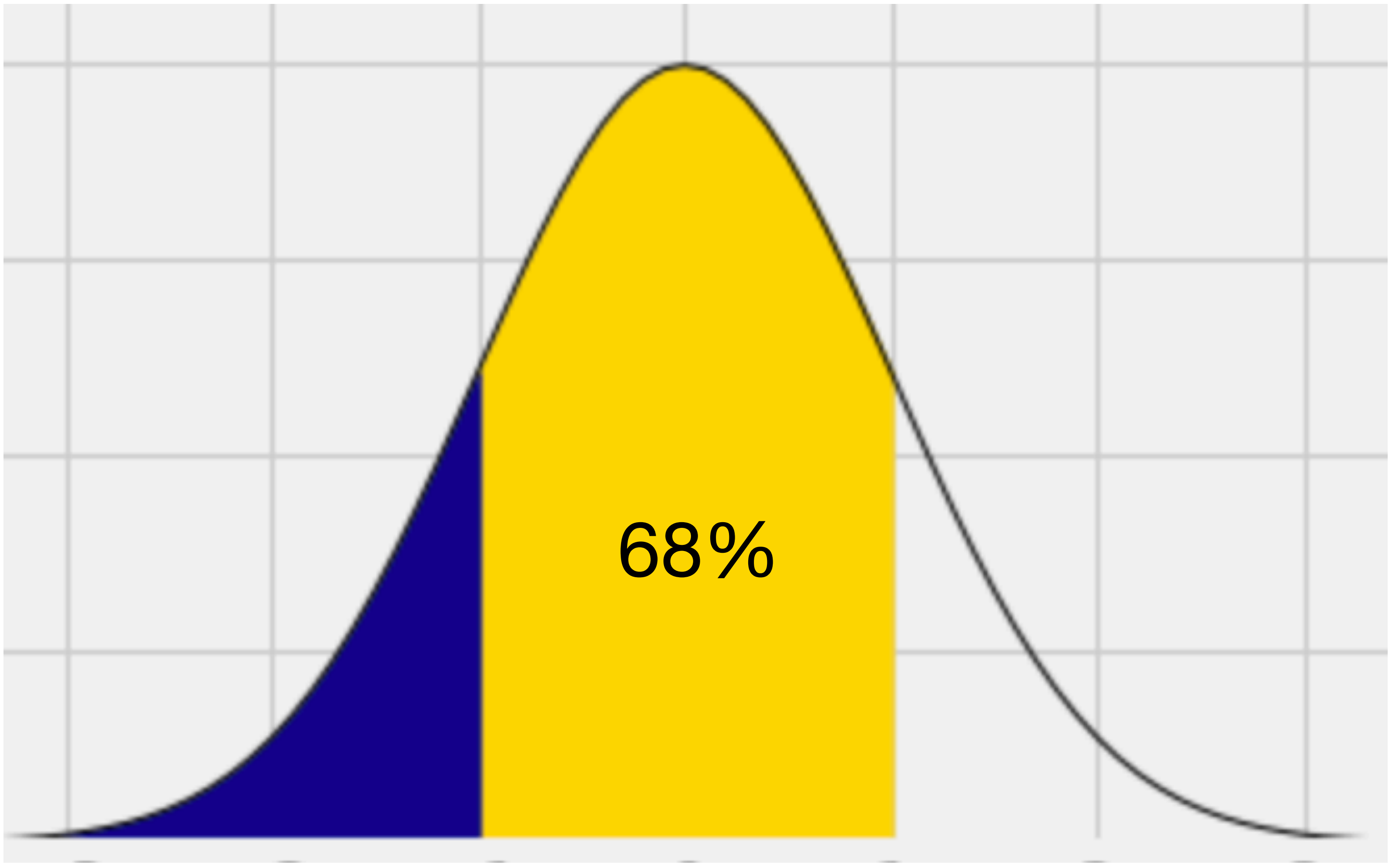
- Makes use of **bins (**each bar corresponds to an individual bin)
  - A **bin** refers to a range of numerical values and **binning** counts the number of values that lie within that range
  - Binning converts a numerical distribution into a categorical distribution
- Makes use of the **area principle**

Area Principle

Areas should be *proportional* to the values they represent

In a histogram, the area of each bar is the percent of individuals in the corresponding bin

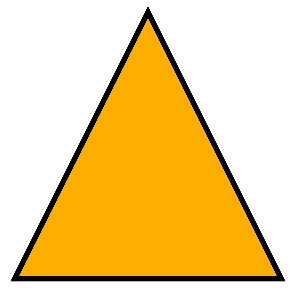
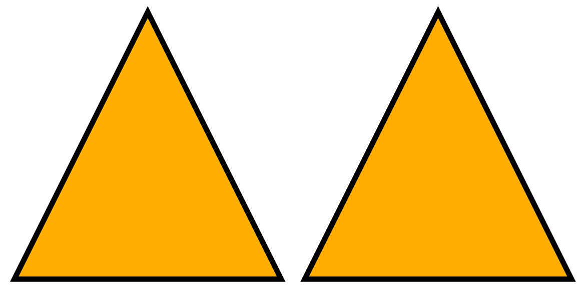
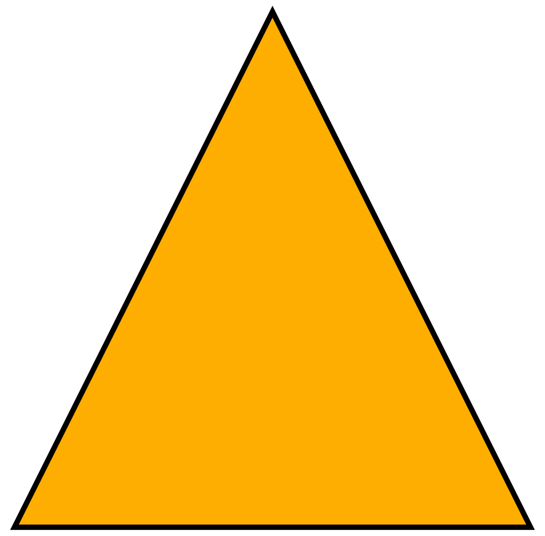
(Later on in the course, we will approximate histograms with smooth curves)



Area Principle

Areas should be *proportional* to the values they represent

For example

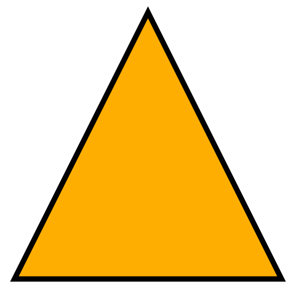
- If you represent 20% of a population by: 
- Then 40% can be represented by: 
- But not by: 

Area Principle

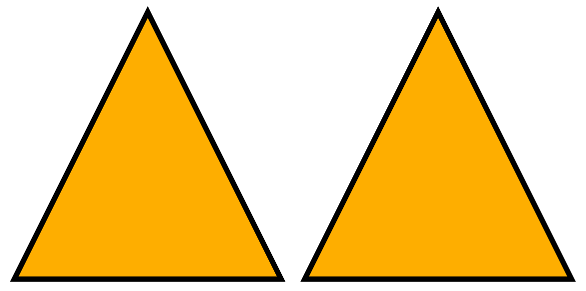
Areas should be *proportional* to the values they represent

For example

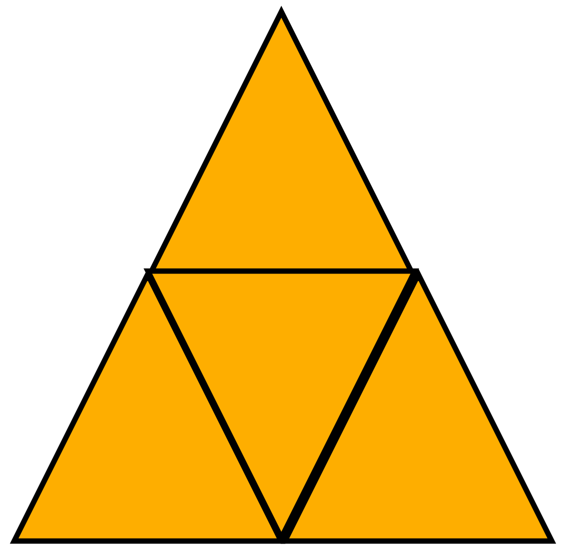
- If you represent 20% of a population by:



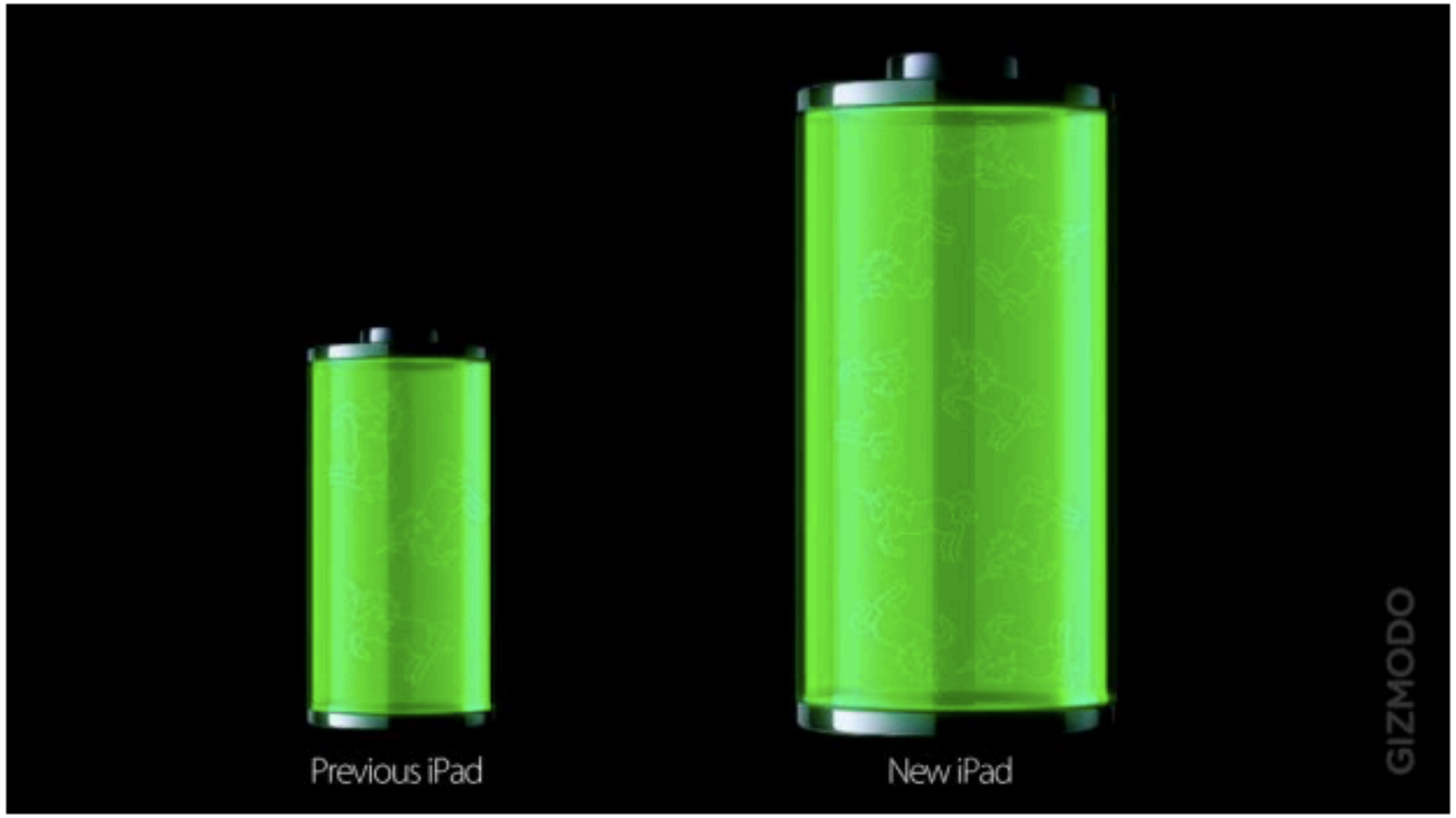
- Then 40% can be represented by:



- But not by:

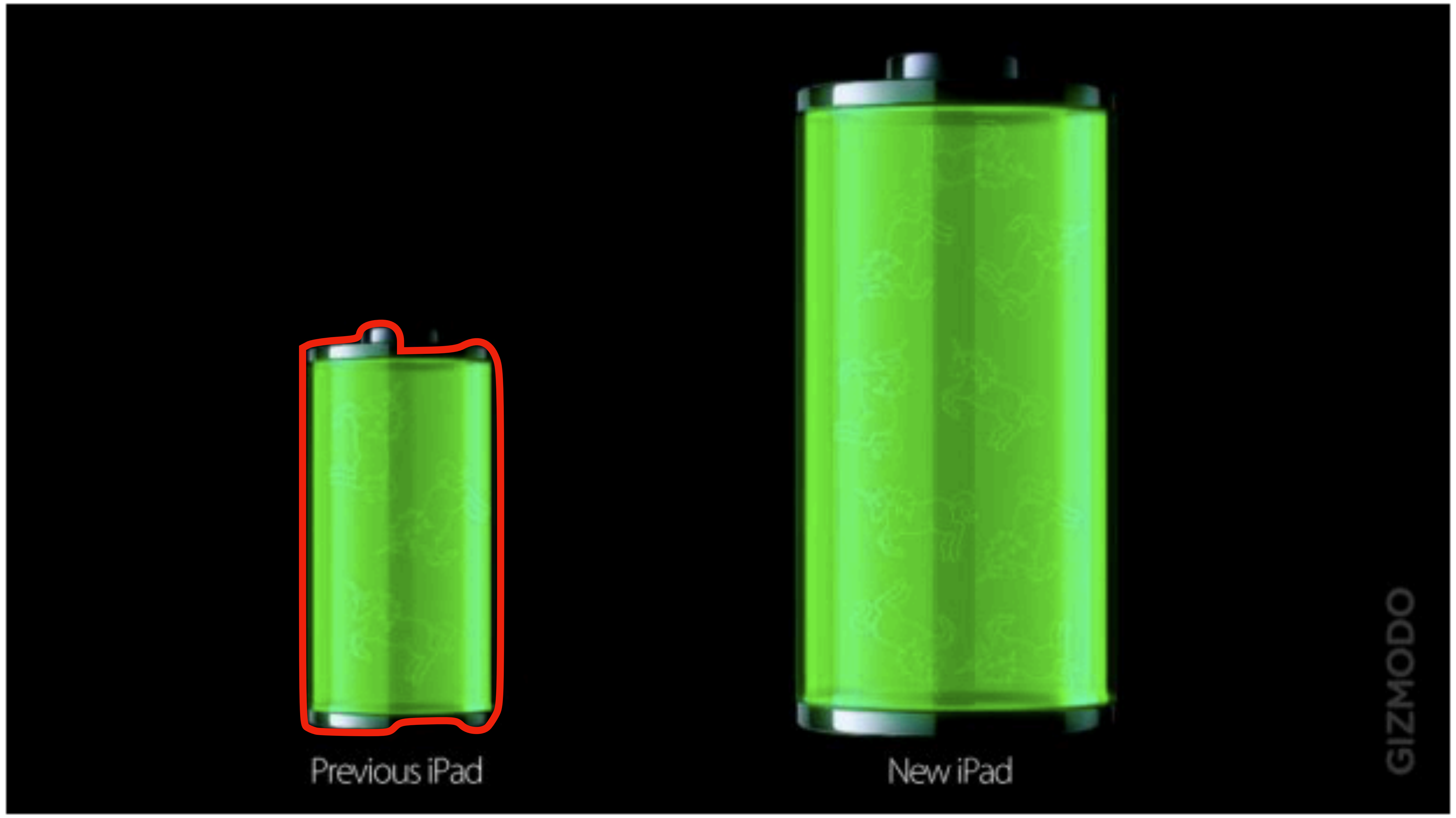


Area Principle



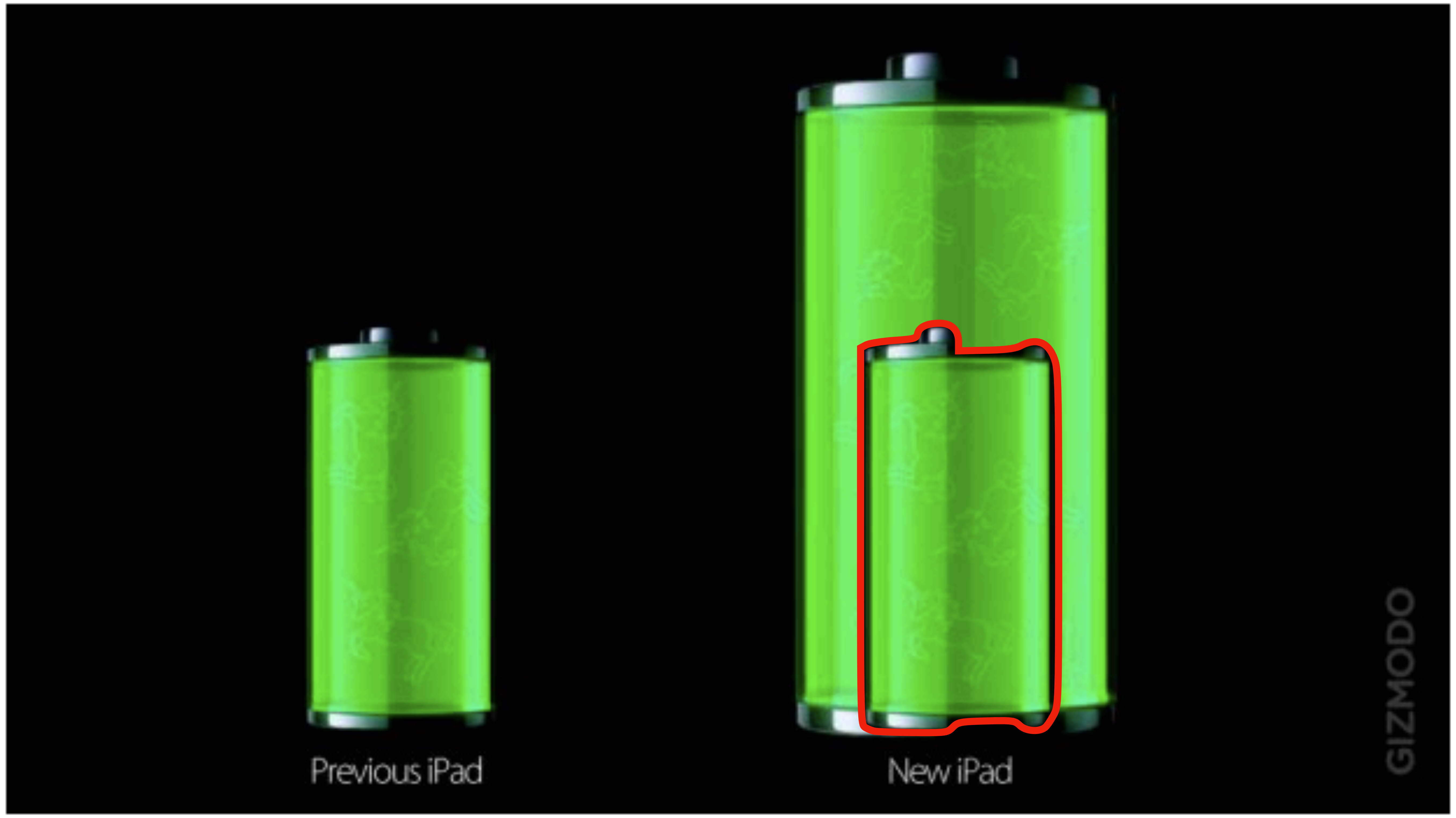
From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

Area Principle



From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

Area Principle



From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

Grouping Numerical Values into Bins

Binning is counting the number of numerical values that lie within a range (which is called bins)

- Bins are defined by their lower bounds (inclusive)
- The upper bound is the lower bound for the next bin)
    - Example: [10, 20)

Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

Binning: Example

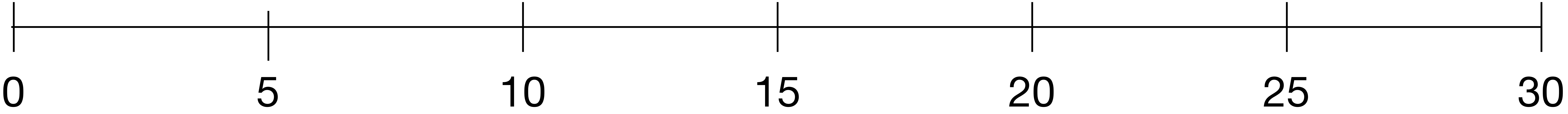
[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

Let’s say we decide to use a bin size of 5

Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

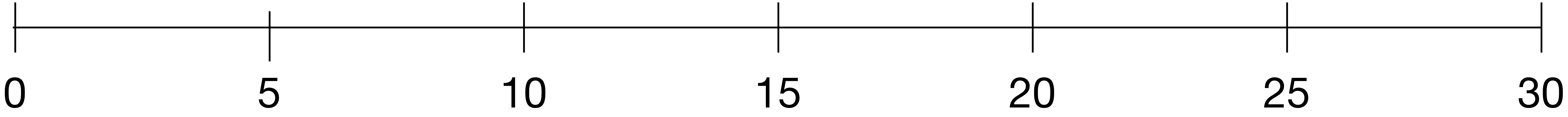
Let’s say we decide to use a bin size of 5



Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

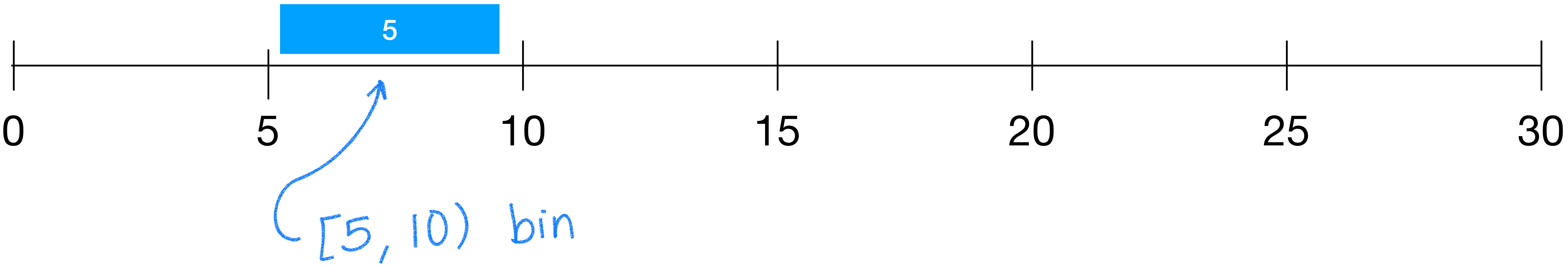
What bin would 5 fall into?



Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

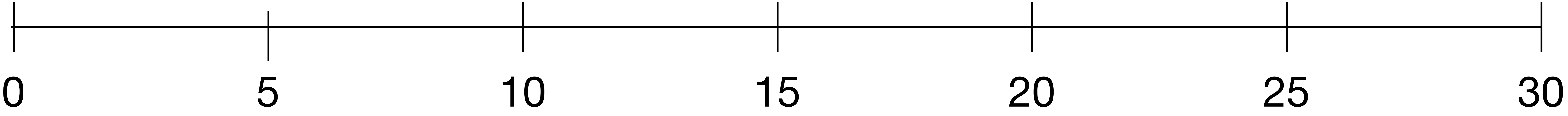
What bin would 5 fall into?



Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

How many individuals fall into bin 15-20?

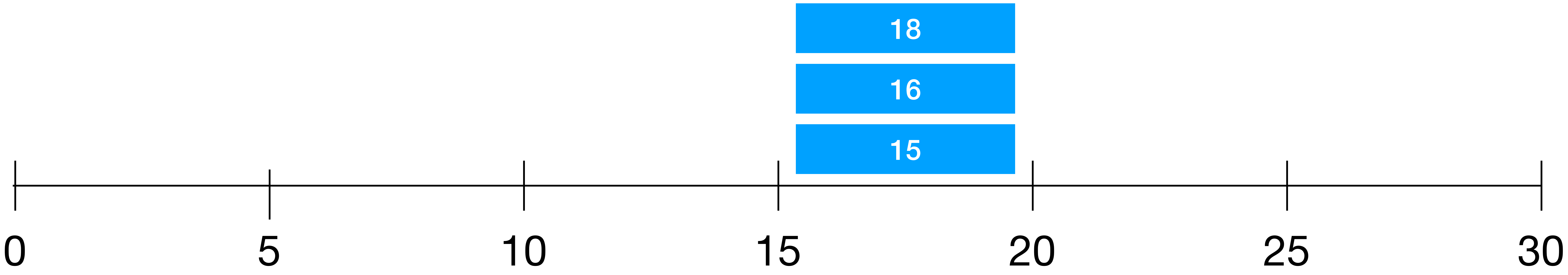


Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]

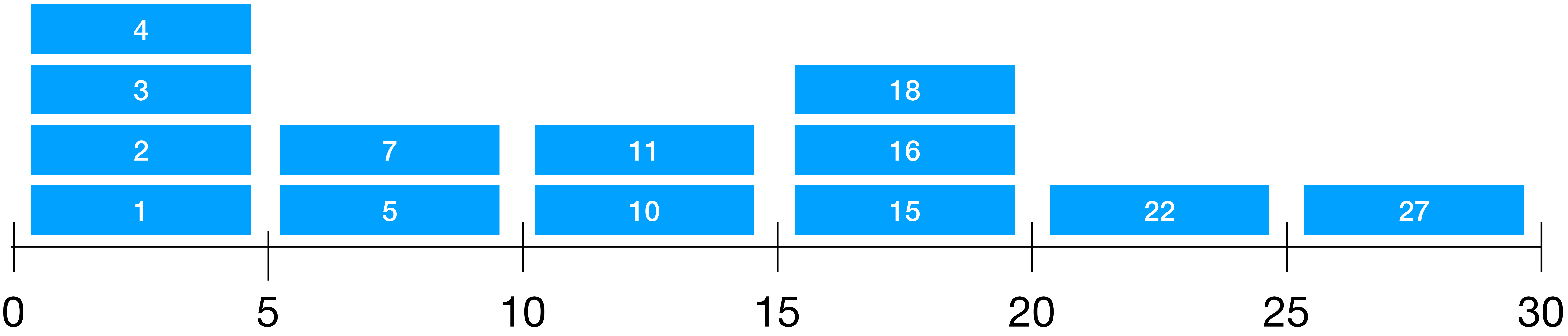
How many individuals fall into bin 15-20?

3



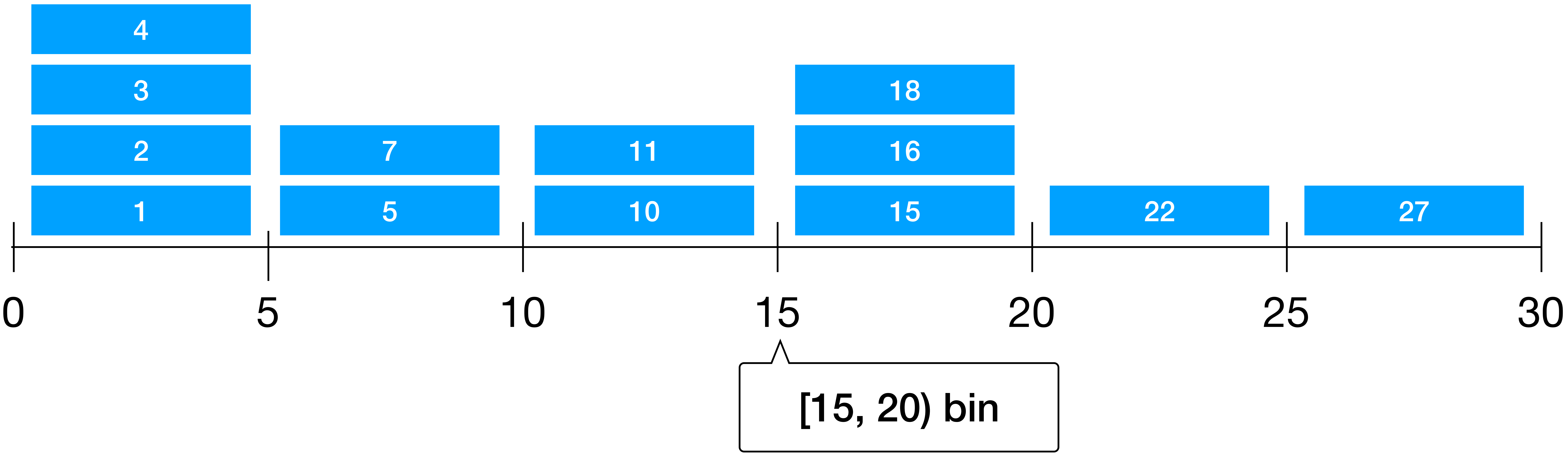
Binning: Example

[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]



Binning: Example

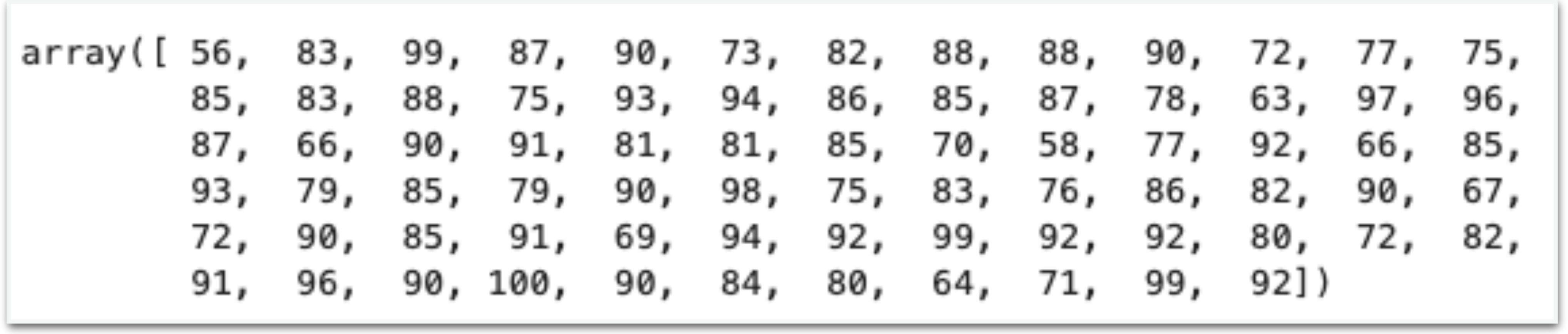
[1, 5, 7, 3, 2, 11, 18, 16, 15, 10, 22, 27, 4]



[15, 20) bin

Choosing Bin Size

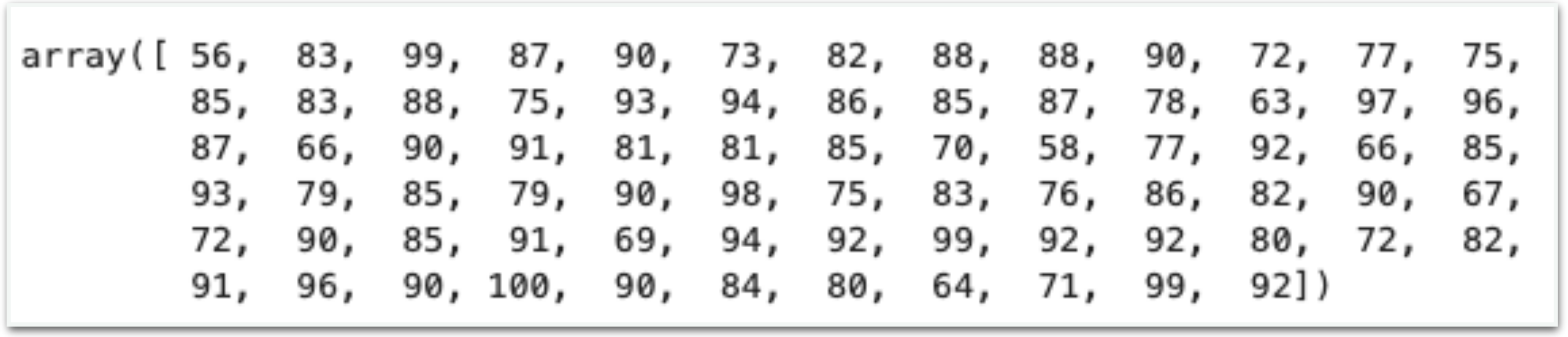
Let’s go back to our data from before:



```
array([ 56,  83,  99,  87,  90,  73,  82,  88,  88,  90,  72,  77,  75,
        85,  83,  88,  75,  93,  94,  86,  85,  87,  78,  63,  97,  96,
        87,  66,  90,  91,  81,  81,  85,  70,  58,  77,  92,  66,  85,
        93,  79,  85,  79,  90,  98,  75,  83,  76,  86,  82,  90,  67,
        72,  90,  85,  91,  69,  94,  92,  99,  92,  92,  80,  72,  82,
        91,  96,  90, 100,  90,  84,  80,  64,  71,  99,  92])
```

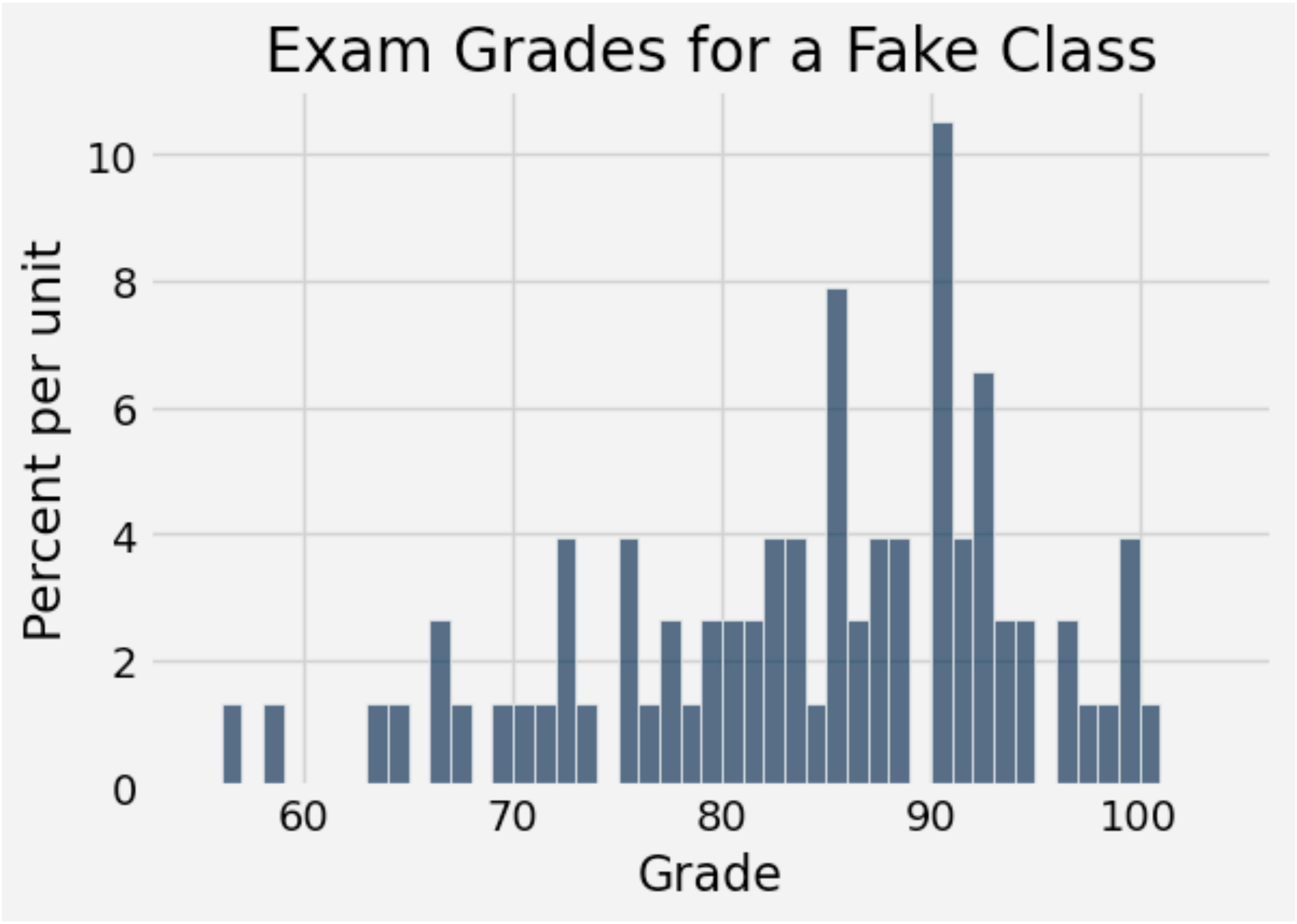
Choosing Bin Size

Let’s go back to our data from before:

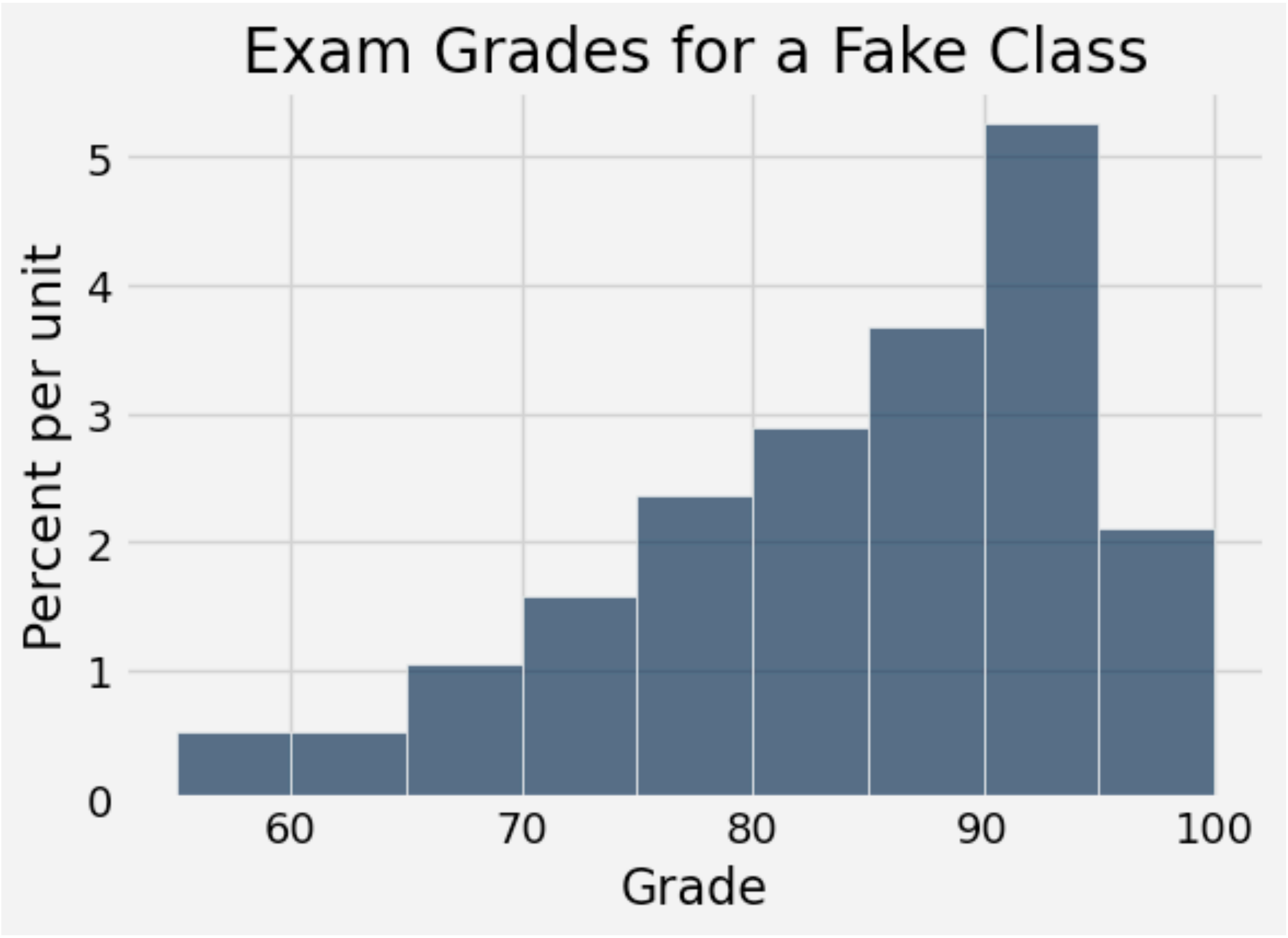


```
array([ 56,  83,  99,  87,  90,  73,  82,  88,  88,  90,  72,  77,  75,
        85,  83,  88,  75,  93,  94,  86,  85,  87,  78,  63,  97,  96,
        87,  66,  90,  91,  81,  81,  85,  70,  58,  77,  92,  66,  85,
        93,  79,  85,  79,  90,  98,  75,  83,  76,  86,  82,  90,  67,
        72,  90,  85,  91,  69,  94,  92,  99,  92,  92,  80,  72,  82,
        91,  96,  90, 100,  90,  84,  80,  64,  71,  99,  92])
```

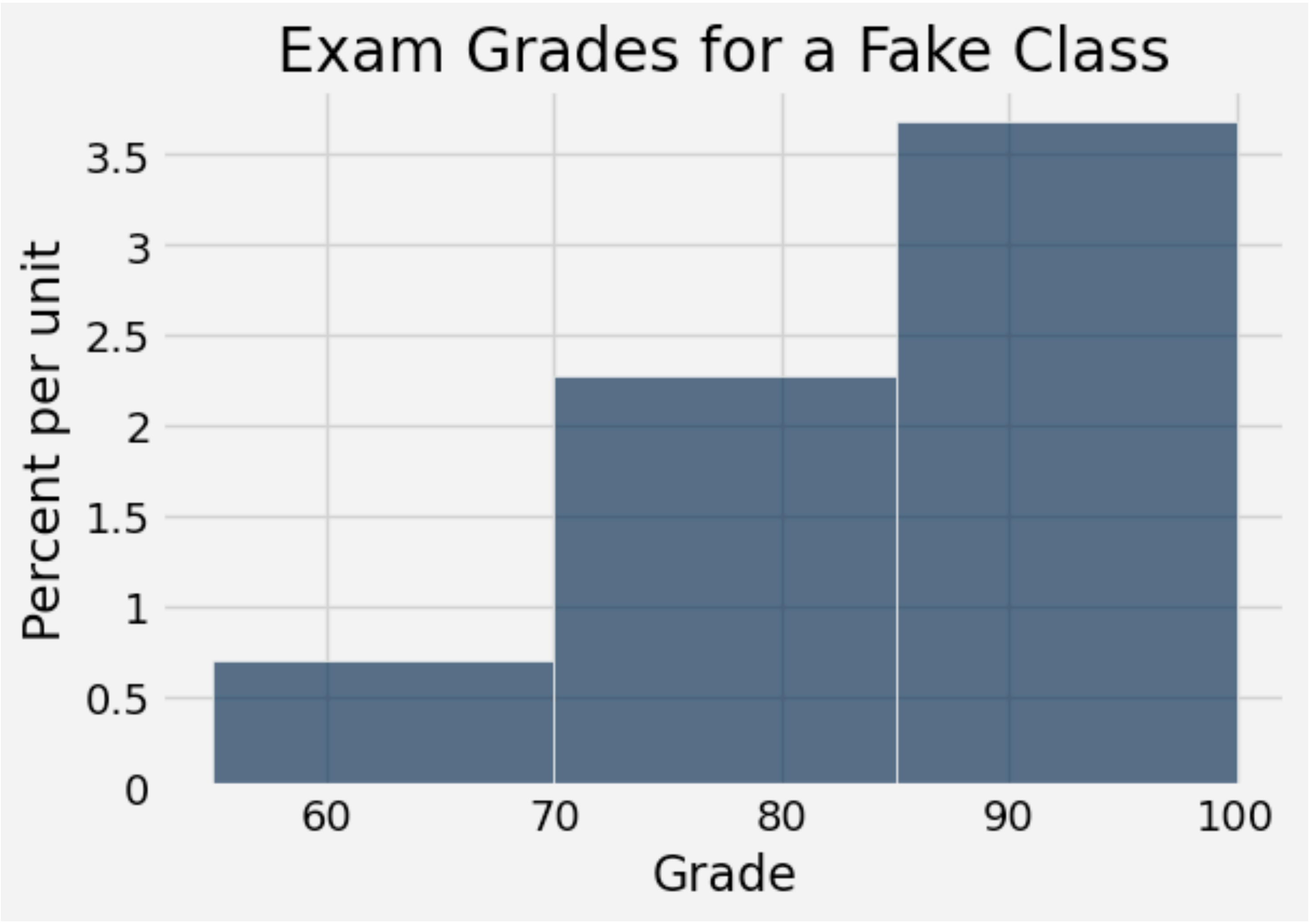
Bin size = 1



Bin size = 5

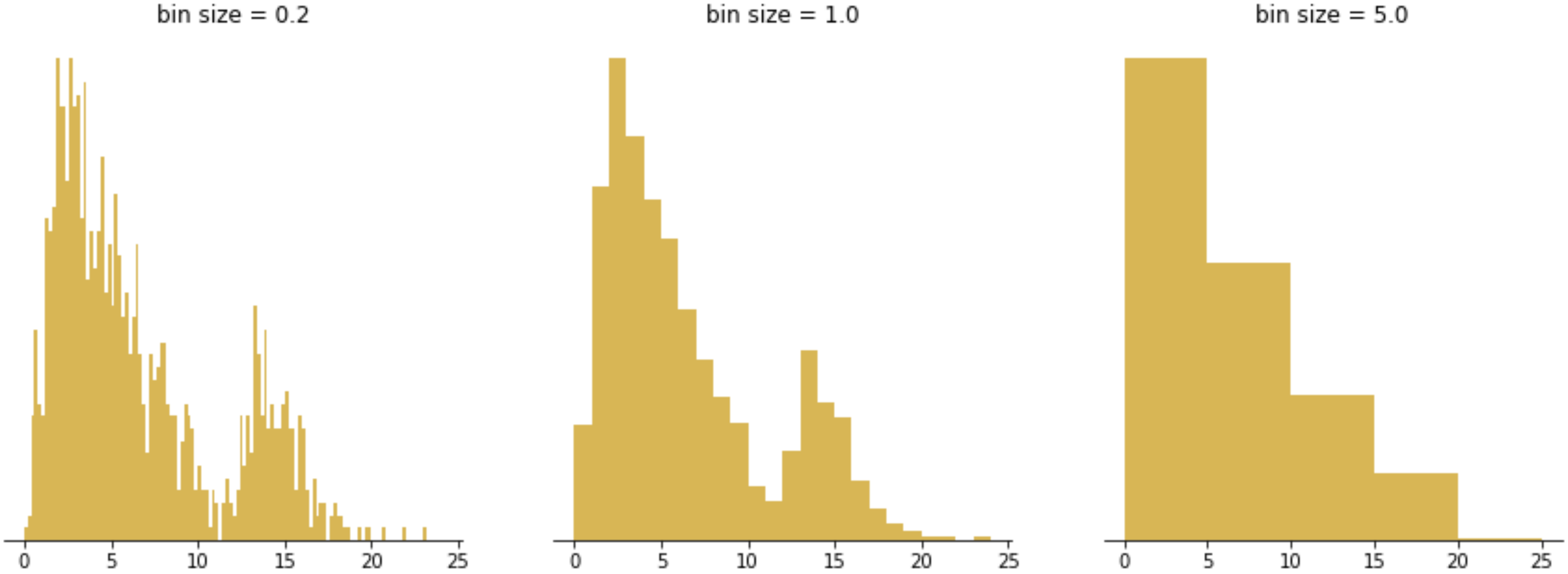


Bin size = 15



Choosing Bin Size

Choose so that it's representative of your data



https://www.atlassian.com/data/charts/histogram-complete-guide

`bin`

Group values in column `c` into 10 equally sized intervals:

- `tbl.bin(c)`

Create `n` equally wide bins:

- `tbl.bin(c, bins=n)`

Create bins of size `step` from `start` to `end`:

- `tbl.bin(c, bins=np.arange(start, end, step))`

`hist`

Create a histogram of numerical values in column `c` with 10 equal bins:

- `tbl.hist(c)`

Create a histogram with `u` as the x-axis:

- `tbl.hist(c, unit=u)`

Create a histogram with specified bins:

- `tbl.hist(c, bins=np.arange(start, end, step))`

Create a histogram with x-axis `u` and specified bins:

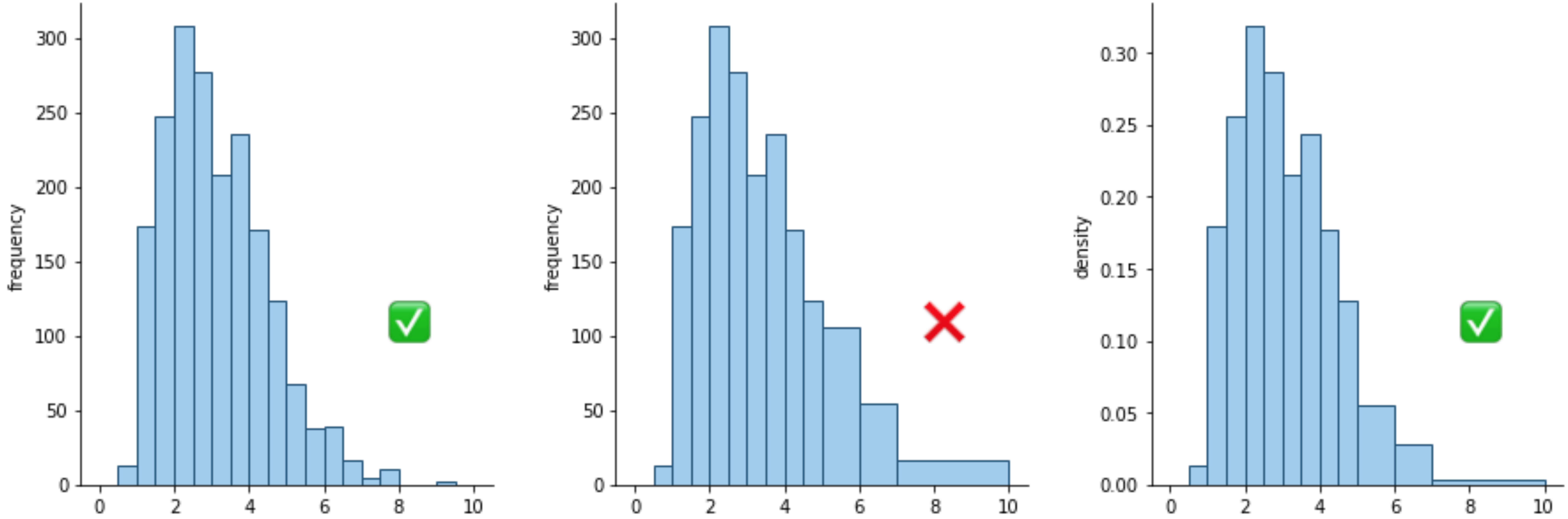
- `tbl.hist(c, unit=u, bins=np.arange(start, end, step))`

Histogram Notebook Demo - Bins

Unequal Bin Sizes

Bin sizes don’t need to be equal - unequal bin size is often used for better representing tails

For unequal bin sizes - vertical axis now represents ***density*** rather than frequency



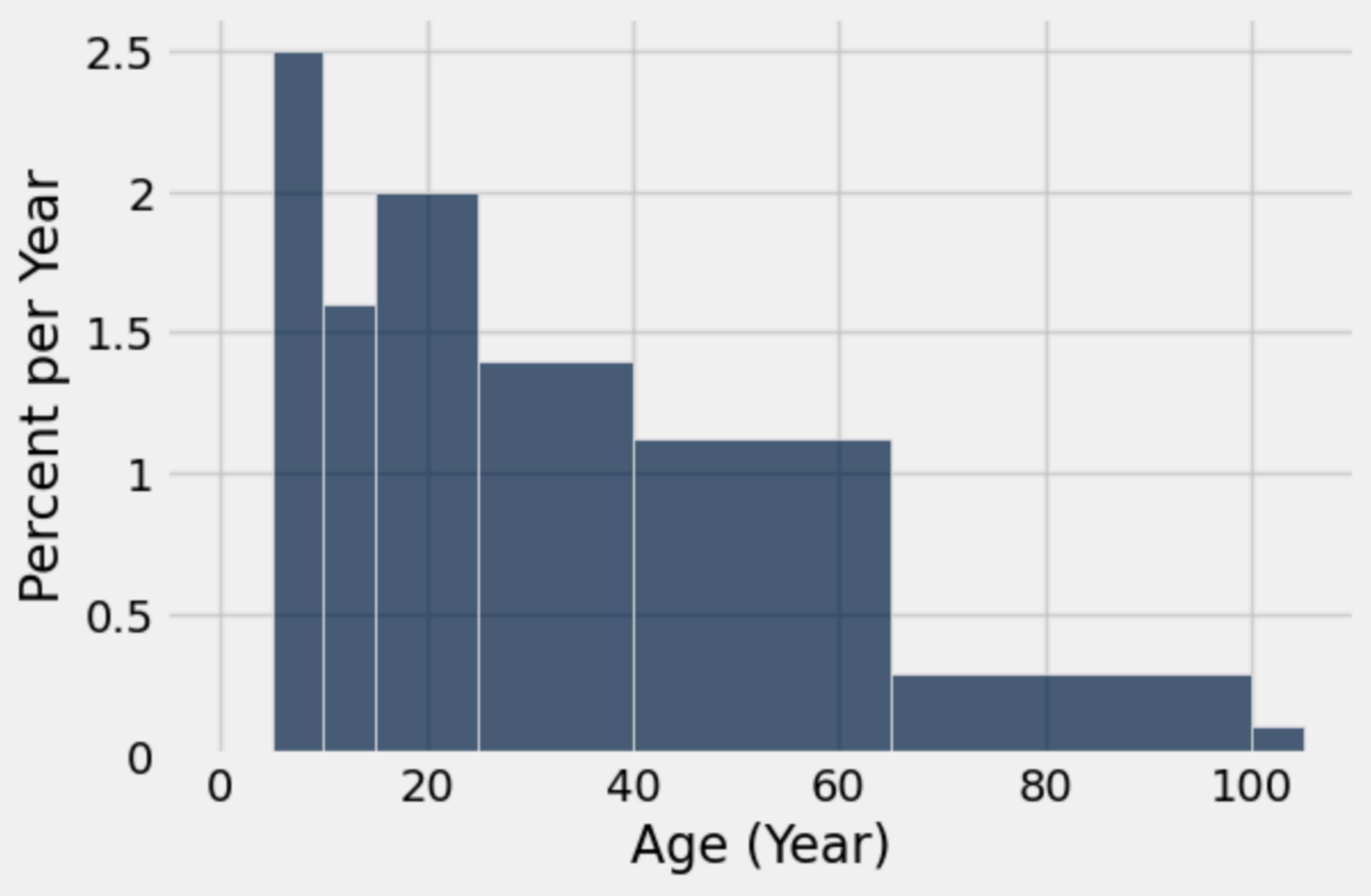
https://www.atlassian.com/data/charts/histogram-complete-guide

Histograms

The area of each bar is a percentage of the whole

The horizontal axis is a numerical distribution
the bins don’t need to be of equal size

The vertical axis is a rate
(e.g., percent/year) - density

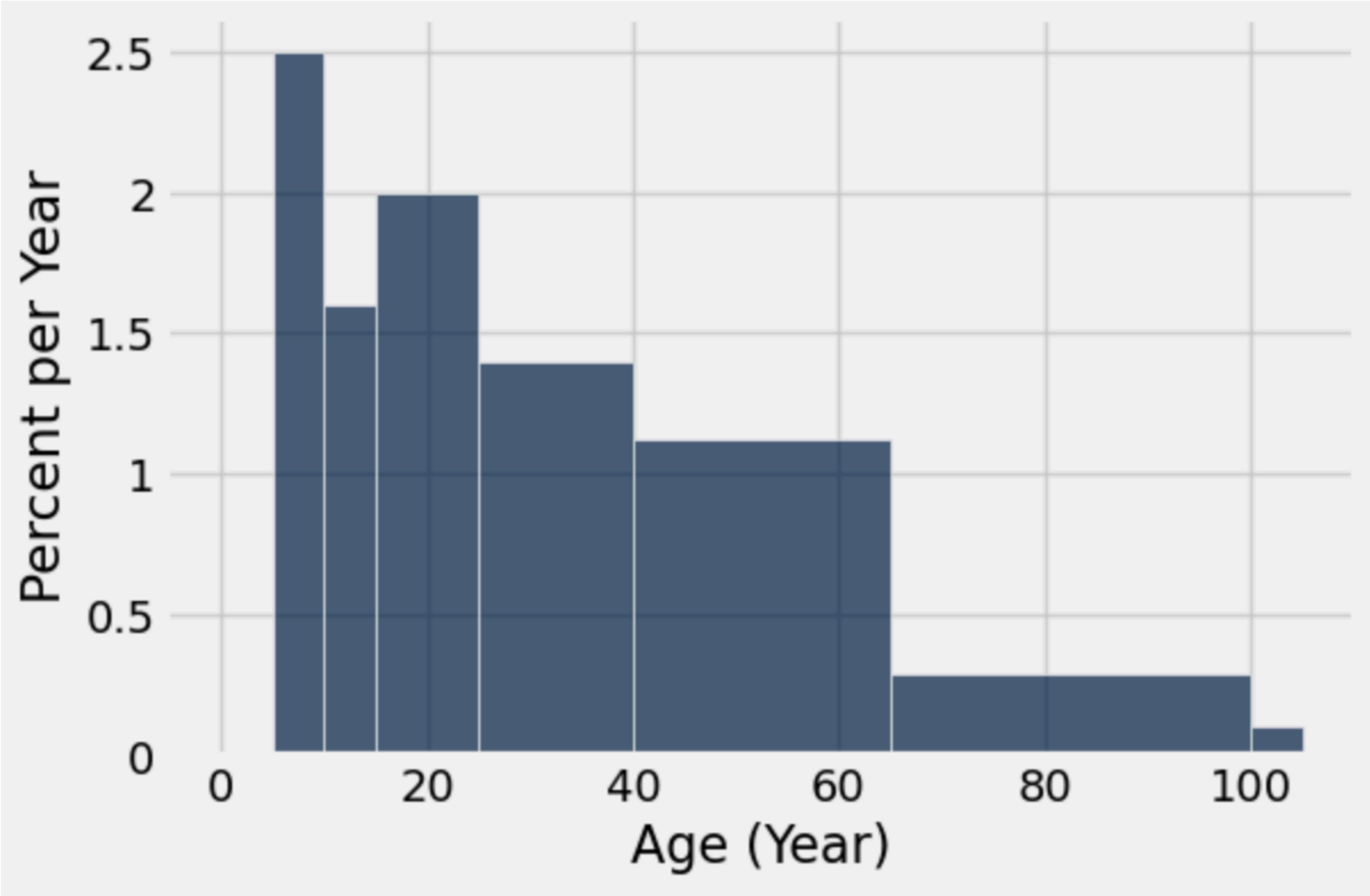


Histogram Formulas

The area of each bar is a percentage of the whole

$$\begin{aligned}\text{area of bar} &= (\text{height of bar}) \times (\text{width of bin}) \\ &= \text{percent of entries in bin}\end{aligned}$$

$$\begin{aligned}\text{height of bar} &= \frac{\text{percent of entries in bin}}{\text{width of bin}} \\ &= \frac{\text{area of bar}}{\text{width of bin}}\end{aligned}$$

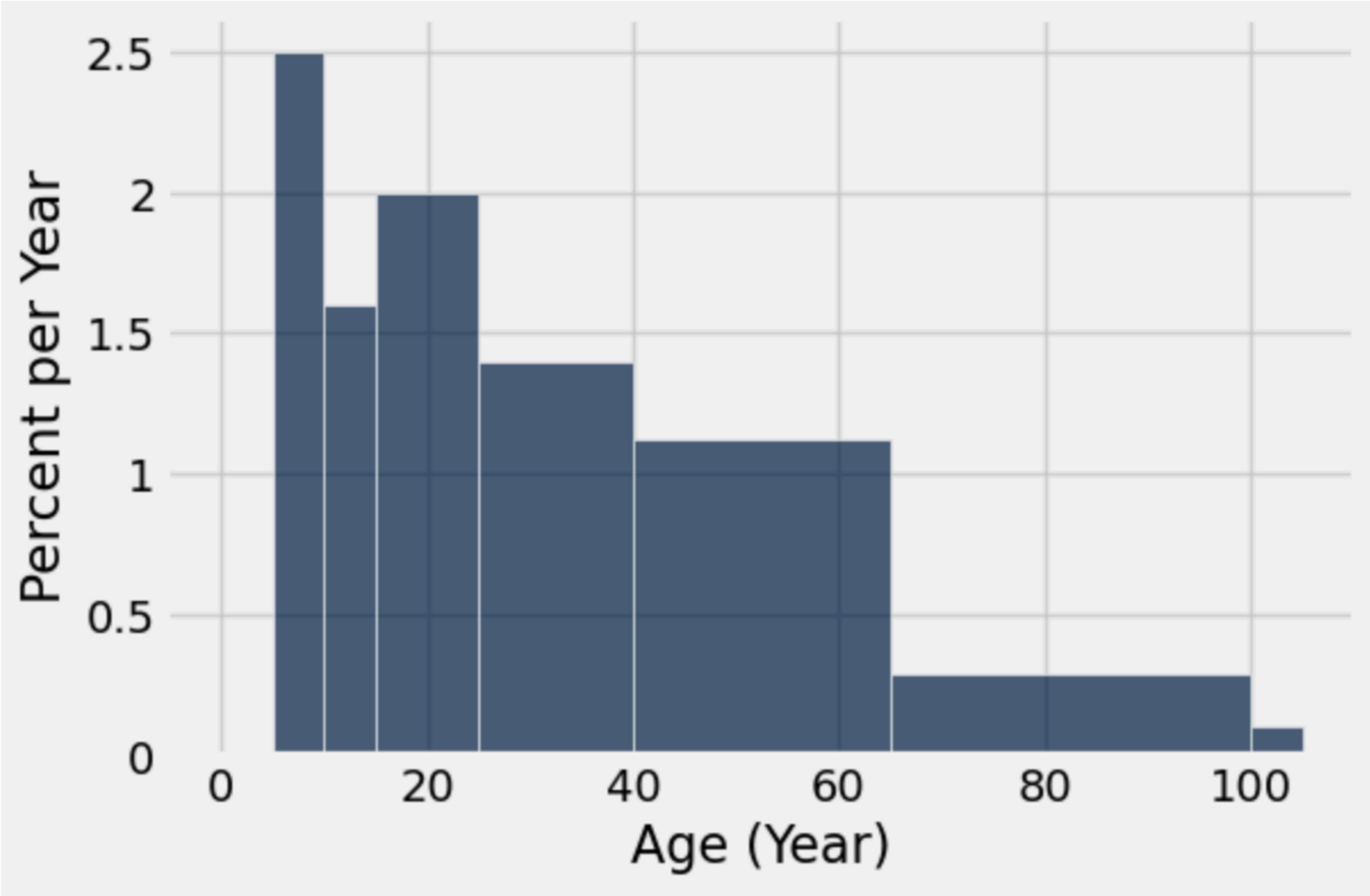


Calculating Heights

The [40, 65) bin contains 56/200 items

- The bin is 28% (56/200) of the whole
- The bin width is 65-40 = 25 years
- Height = $\dfrac{\text{28 percent}}{\text{25 years}}$

= 1.12% per year



Area Notebook Demo

Bar Chart vs Histogram

Bar Chart

- Distribution of categorical variable
- Length of bars is proportional to the frequency / percent of individuals

Histogram

- Distribution of numerical variable
- Horizontal axis is numerical, bins can be unequal
- Area of bars is proportional of percent of individuals, height measures density

Charts Summary

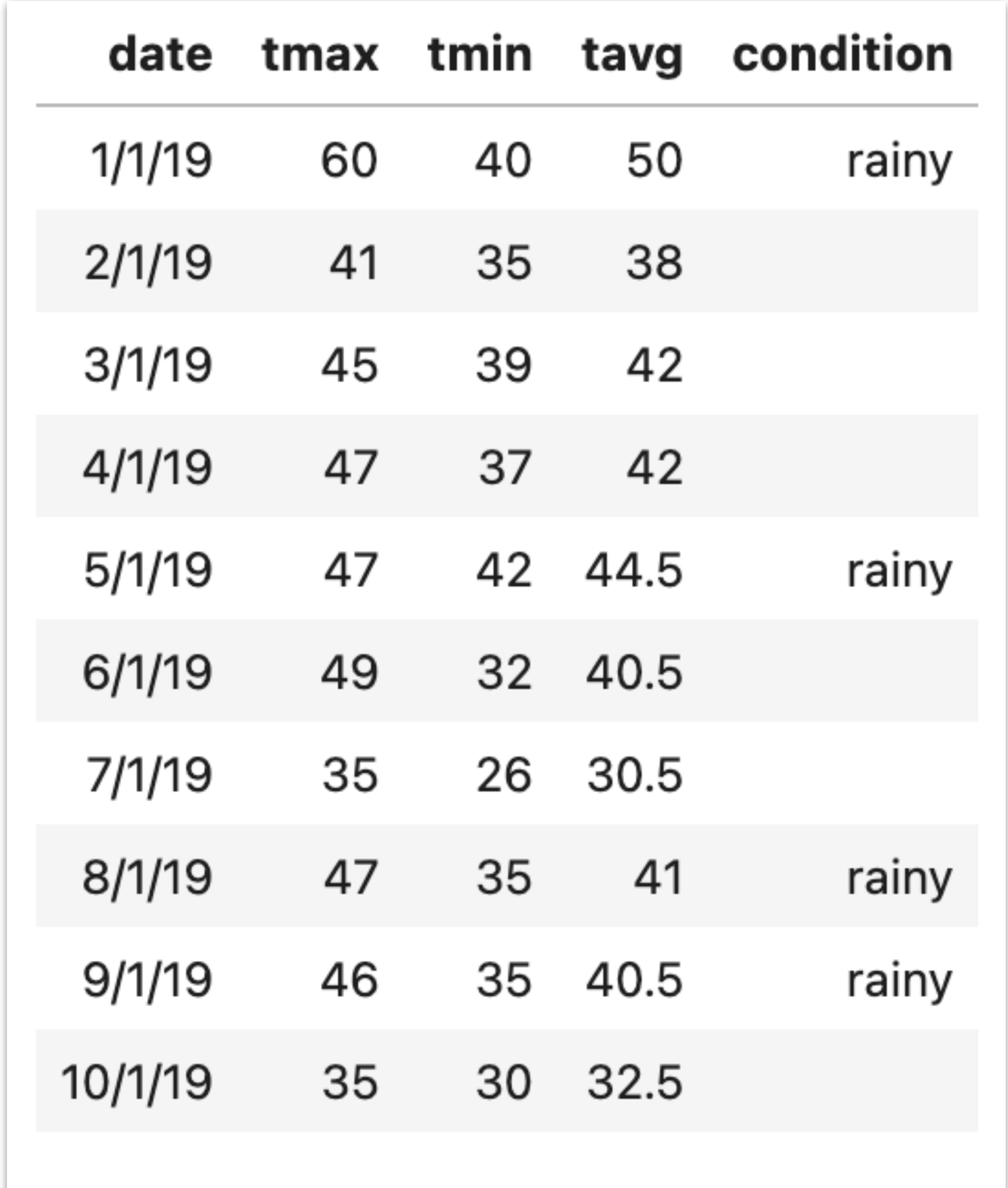
Type	Syntax	Description
Line graph	`.plot(x_axis, y_axis)`	Sequential numerical data
Scatter Plot	`.scatter(x_axis, y_axis)`	Relation between two numerical values
Bar Chart	`.barh(column_label)`	Distribution of one categorical variable (already grouped)
Histogram	`.hist(column_label, unit, bins)`	Distribution of one numerical variable

Chart Selection Exercise

We have NYC weather data from 2019 as shown below (from Kaggle)

Which type of chart (line, scatter, bar, histogram) would best help you answer to each question?

- Do days with hotter highs also tend to have hotter lows?
- How do the number of rainy days compare with the number of snowy days?
- What percent of days have a high of at least 75 degrees?



date	tmax	tmin	tavg	condition
1/1/19	60	40	50	rainy
2/1/19	41	35	38	
3/1/19	45	39	42	
4/1/19	47	37	42	
5/1/19	47	42	44.5	rainy
6/1/19	49	32	40.5	
7/1/19	35	26	30.5	
8/1/19	47	35	41	rainy
9/1/19	46	35	40.5	rainy
10/1/19	35	30	32.5	

Next Class

- Today
  - Histograms and Bar Charts
- Wednesday
  - Functions, Groups, Pivots, and Joins