190F Foundations of Data Science

Spring 2020

Lecture 5

Charts and Histograms

Announcements

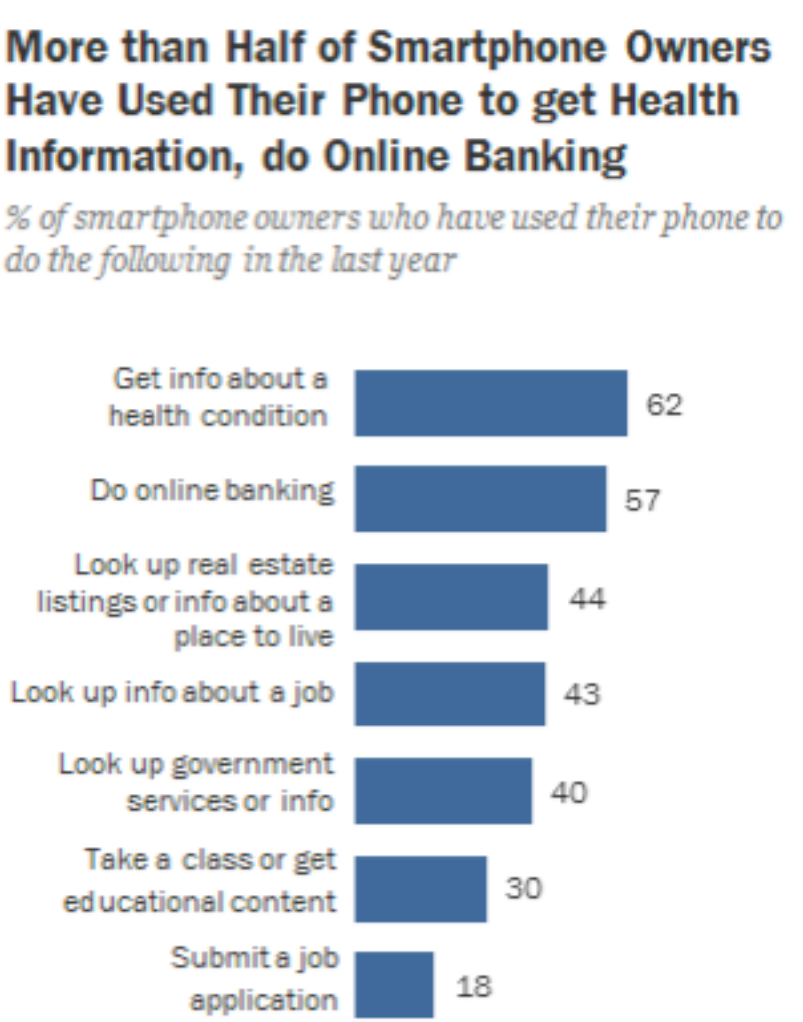
Data Visualization

Discussion Question

Which of the following questions can be answered by this chart?

Among survey responders...

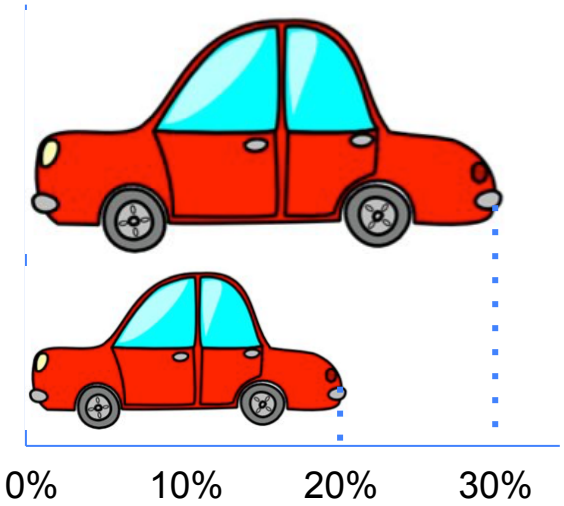
- What proportion did **not** use their phone for online banking?
- What proportion either used their phone for online banking or to look up real estate listings?
- Did everyone use their phone for at least one of these activities?
- Did anyone use their phone for both online banking and real estate?



Pew research center, 2014

Area Principle

Areas should be proportional to the values they represent



In 2013,

30% of accidental deaths of males were due to automobile accidents

20% of accidental deaths of females were due to automobile accidents

Example from Tian Zheng


$41 BILLION IN SALES
BILLIONS OF DOLLARS
GDP OF AFGHANISTAN
$21 BILLION
$4.1 BILLION IN SALES
$4.3 BILLION IN SALES
$8 BILLION IN SALES
$8.2 BILLION IN SALES
$9.4 BILLION IN SALES
$11.3 BILLION IN SALES
40
35
30
25
20
15
10
5
Pizza Hut
KFC
Wendy's
HAMBURGERS
BURGER KING

WHERE YOUR FOOD COMES FROM

Imports of foods have doubled in a decade and now account for a fifth of what Americans eat.

A large percentage of these foods that Americans eat are imported.

HONEY
61%

LAMB
52%

FRUIT AND NUTS
51%

FRESH VEGETABLES/ MELONS
20%

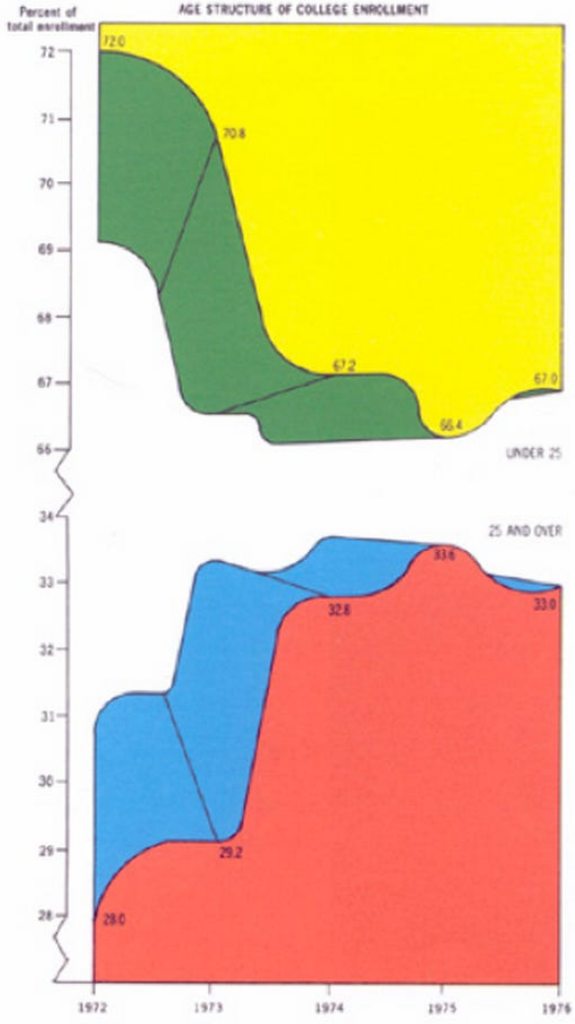
SEAFOOD
88%

These countries are the largest exporters of each food.

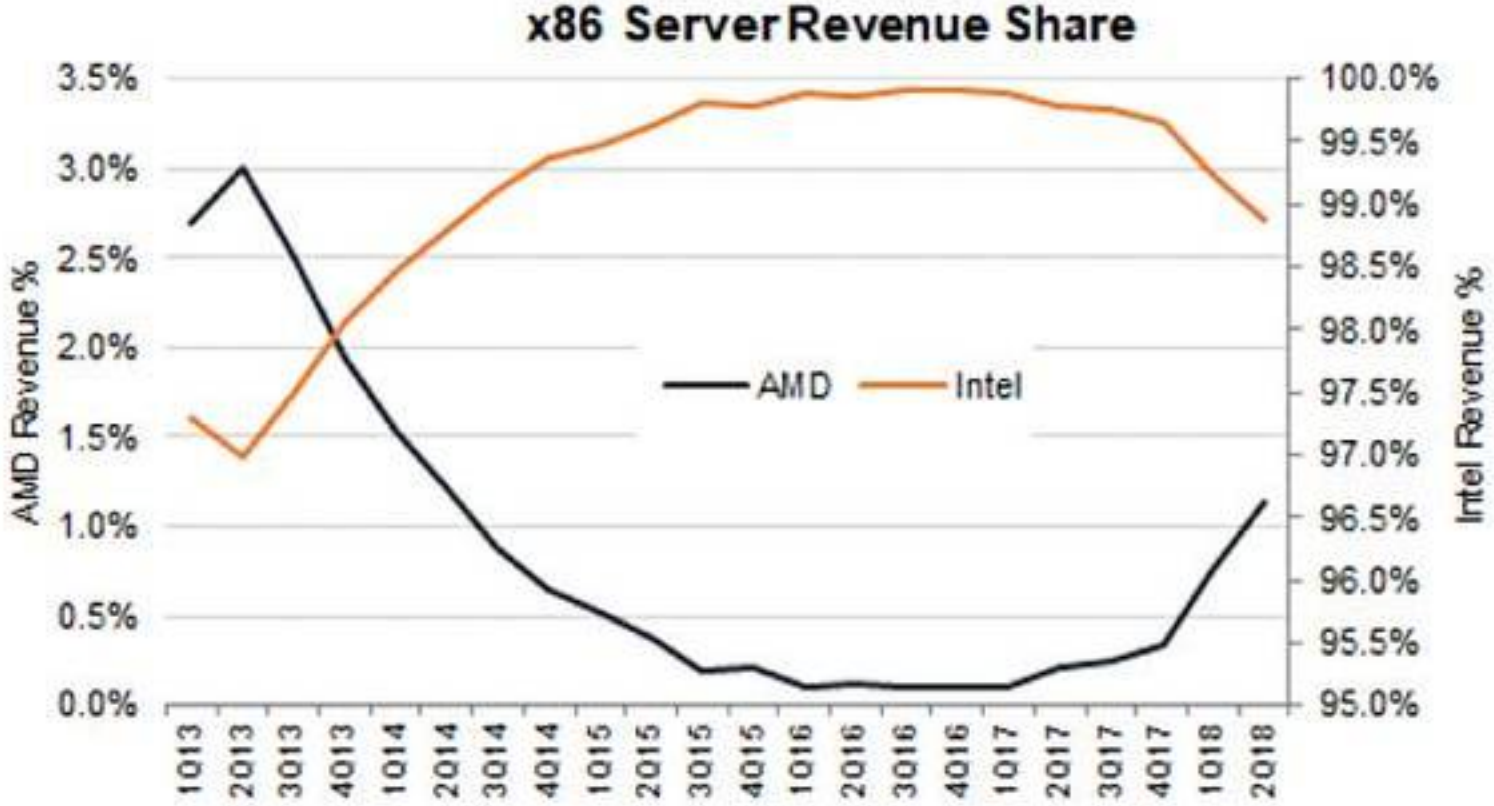
FRUIT		VEGETABLES		SEAFOOD	
MEXICO	CHILE	MEXICO	CANADA	CHINA	VIETNAM
26% SUPPLIED	13%	50%	22%	16%	6.5%

Source: USDA Economic Research Service

CHART: KANG KIM; STYLIST: LAURIE RAAB


AGE STRUCTURE OF COLLEGE ENROLLMENT
Percent of total enrollment
72
71
70
69
68
67
66
34
33
32
31
30
29
28
72.0
70.8
67.2
66.4
67.0
UNDER 25
25 AND OVER
33.6
32.8
33.0
29.2
28.0
1972
1973
1974
1975
1976


JOB LOSS BY QUARTER
15 MIL
13.5 MIL
9 MIL
7 MIL
DEC '07
SEPT '08
MARCH '09
JUNE '10
FOX NEWS .COM
SOURCE: BLS
AMERICA'S NEWSROOM
FAIRFAX, VA... BYRD WAS ADMITTED TO THE H
S&P 3.08

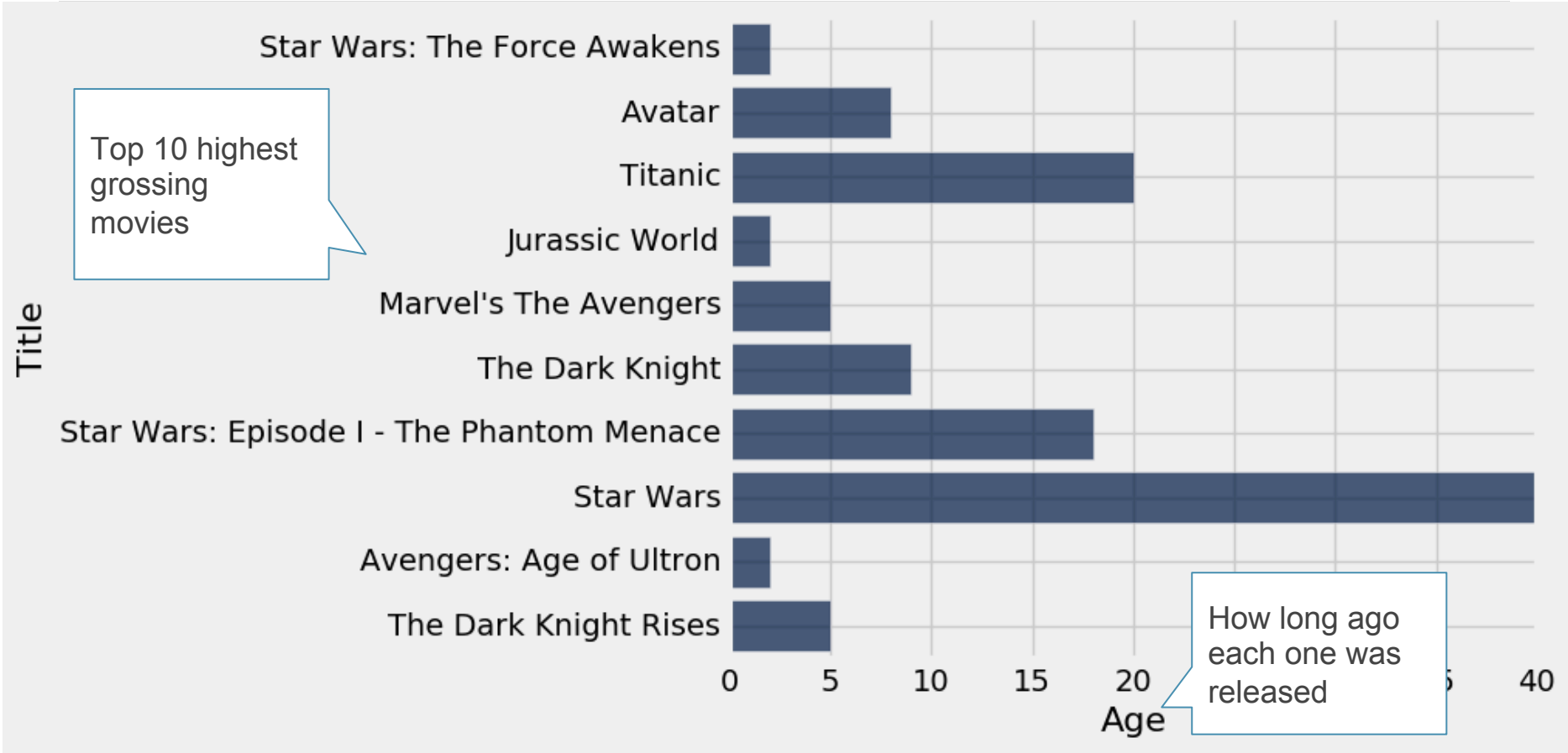


Source: Mercury Research; Wells Fargo Securities, LLC

Numerical Data

(Demo)

How Do You Generate This Chart?



Types of Data

All values in a column should be both the same type **and** be comparable to each other in some way

- **Numerical** — Each value is from a numerical scale
  - Numerical measurements are ordered
  - Differences are meaningful
- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering
  - Categories are the same or different

“Numerical” Data

Just because the values are numbers, doesn’t mean the variable is numerical

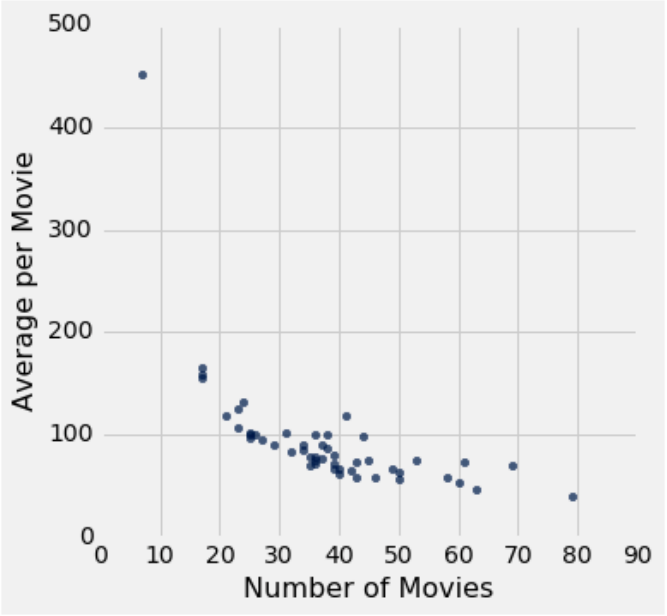
- Census example had numerical `SEX` code (0, 1, and 2)
- It doesn’t make sense to perform arithmetic on these “numbers”, e.g. 1 - 0 or (0+1+2)/3 are nonsense here
- The variable `SEX` is still categorical, even though numbers were used for the categories

Terminology

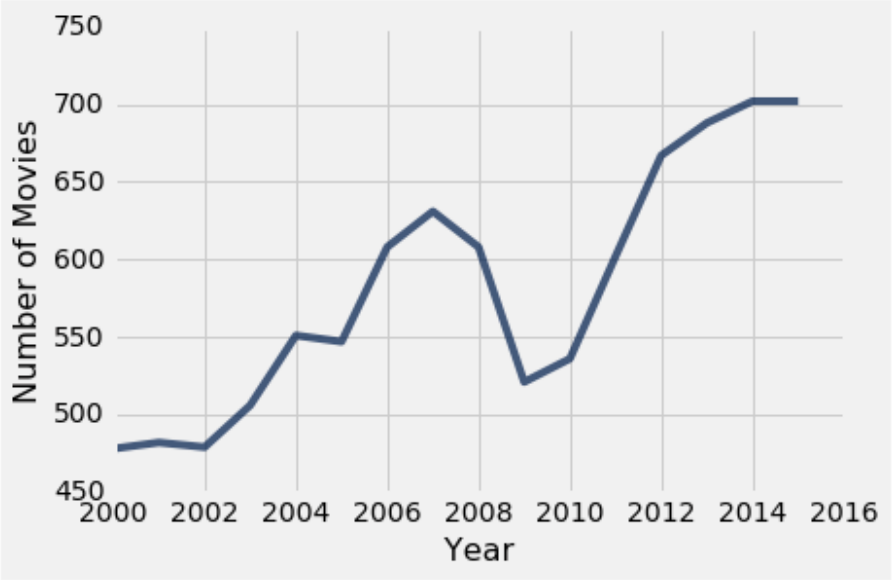
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the frequency of individuals that have that value
- Frequency is measured in counts. Later we will use proportions or percents.

Plotting Two Numerical Variables

Scatter plot: `scatter`



Line graph: `plot`



Categorical Data

(Demo)

Bar Charts of Counts

Distributions:

- The distribution of a variable (a column) describes the frequency of its different values
- The `group` method counts the number of rows for each value in a column

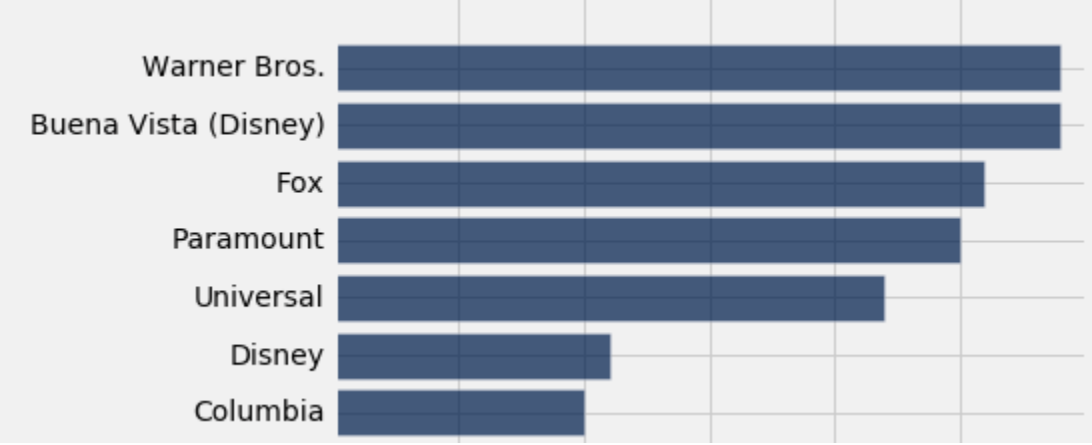
Bar charts can display the distribution of categorical values

- Proportion of how many US residents are male or female
- Count of how many top movies were released by each studio

(Demo)

Categorical Distributions

bar chart: **`barh`**



Displays a categorical distribution

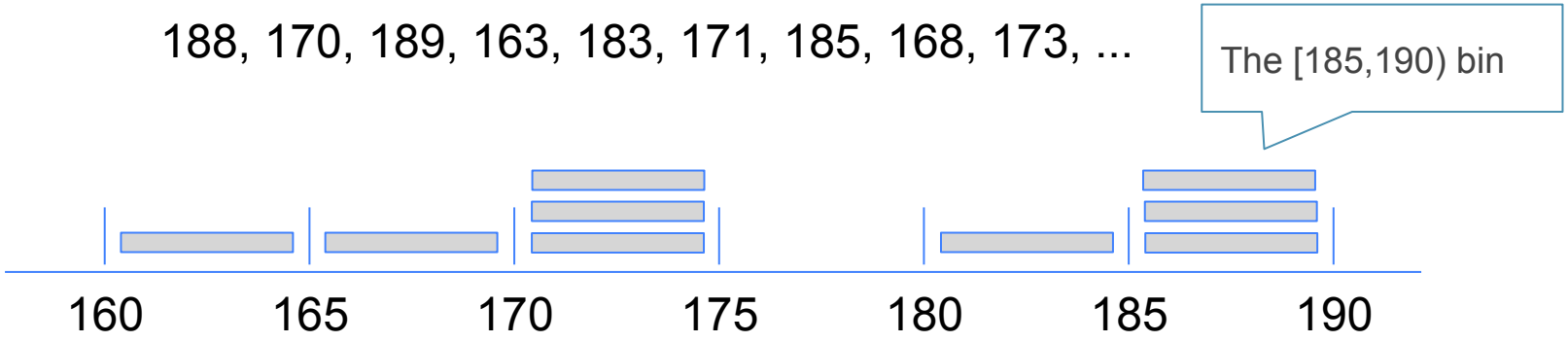
(But when the values of the variable have a rank ordering, or fixed sizes relative to each other, more care might be needed.)

Binning

Binning Numerical Values

Binning is counting the number of numerical values that lie within ranges, called bins.

- Bins are defined by their lower bounds (inclusive)
- The upper bound is the lower bound of the next bin



Histogram

Chart to display the distribution of numerical values using bins

(Demo)

The Density Scale

Histogram Axes

By default, `hist` uses a scale (`normed=True`) that ensures the area of the chart sums to 100%

- The horizontal axis is a number line (e.g., years)
- The vertical axis is a rate (e.g., percent per year)
- The area of a bar is a percentage of the whole

(Demo)

How to Calculate Height

The [20, 40) bin contains 59 out of 200 movies

- “59 out of 200” is 29.5%
- The bin is 40 - 20 = 20 years wide

$$\text{Height of bar} = \frac{29.5 \text{ percent}}{20 \text{ years}}$$

$$= 1.475 \text{ percent per year}$$

Height Measures Density

$$\text{Height} = \frac{\%\text{ in bin}}{\text{width of bin}}$$

- The height measures the percent of data in the bin ***relative to the amount of space in the bin*.**
- So height measures crowdedness, or **density**.

(Demo)

Area Measures Percent

Area = % in bin = Height x width of bin

- “How many individuals in the bin?” Use area.
- “How crowded is the bin?” Use height.

Discussion Question

What's the height of each bar in these two histograms?

```
actress.hist(1, bins=[0,15,25,85])
```

```
actress.hist(1, bins=[0,15,35,85])
```

What are the vertical axis units?

Name	2016 Income (millions)
Jennifer Lawrence	61.7
Scarlett Johansson	57.5
Angelina Jolie	40
Jennifer Aniston	24.75
Anne Hathaway	24
Melissa McCarthy	24
Bingbing Fan	20
Sandra Bullock	20
Cara Delevingne	15
Reese Witherspoon	15
Amy Adams	15
Kristen Stewart	12
Amanda Seyfried	10.5
Tina Fey	10.5
Julia Roberts	10
Emma Stone	10
Natalie Portman	8.5
Margot Robbie	8
Meryl Streep	6
Mila Kunis	4.5

Chart Types

Bar Chart Versus Histogram

Bar Chart

- 1 categorical axis & 1 numerical axis
- Bars have arbitrary (but equal) widths and spacings
- For distributions: height (or length) of bars are proportional to the percent of individuals

Histogram

- Horizontal axis is numerical, hence to scale with no gaps
- Height measures density; areas are proportional to the percent of individuals