MODULE 3

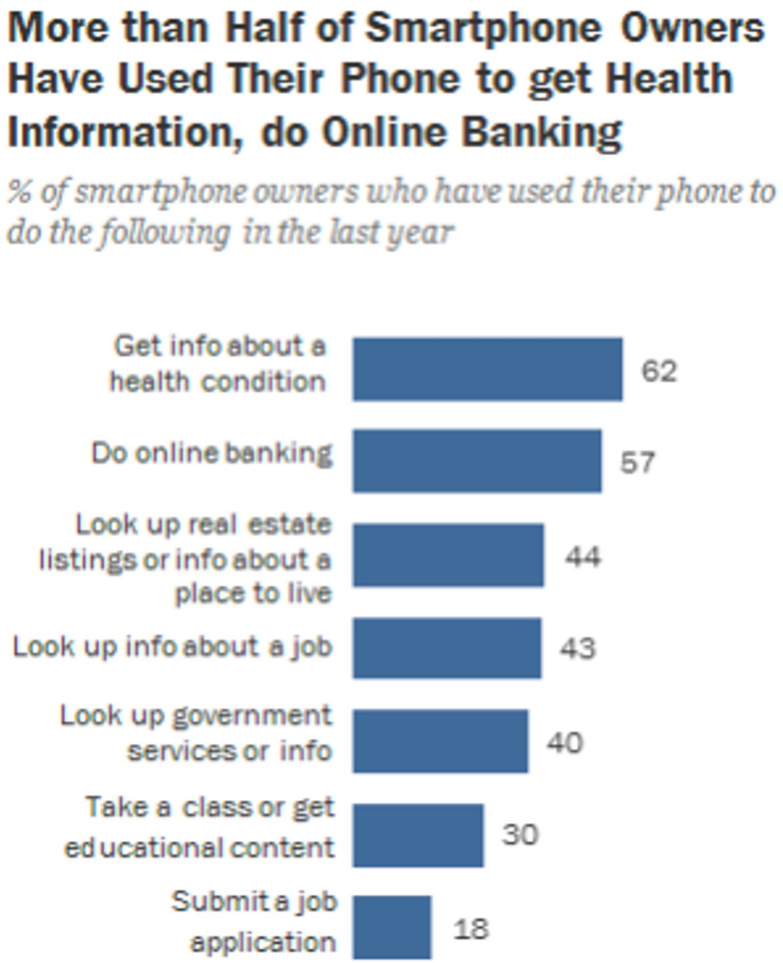
Data visualization

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?



More than Half of Smartphone Owners Have Used Their Phone to get Health Information, do Online Banking

% of smartphone owners who have used their phone to do the following in the last year

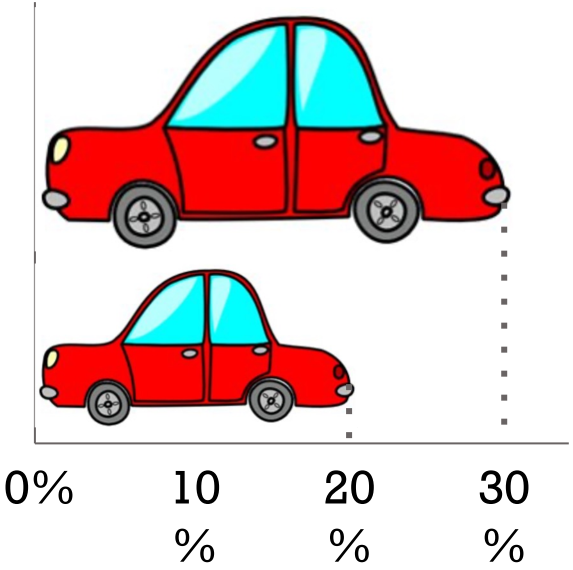
Pew research center, 2014

VISUALIZATION – BEYOND TABLES

- Tables are a powerful way of organizing and visualizing data.
- However, large tables of numbers can be difficult to interpret, no matter how organized they are.
- Sometimes it is much easier to interpret graphs than numbers.

AREA PRINCIPLE

Areas should be proportional to the values they represent



In 2013,

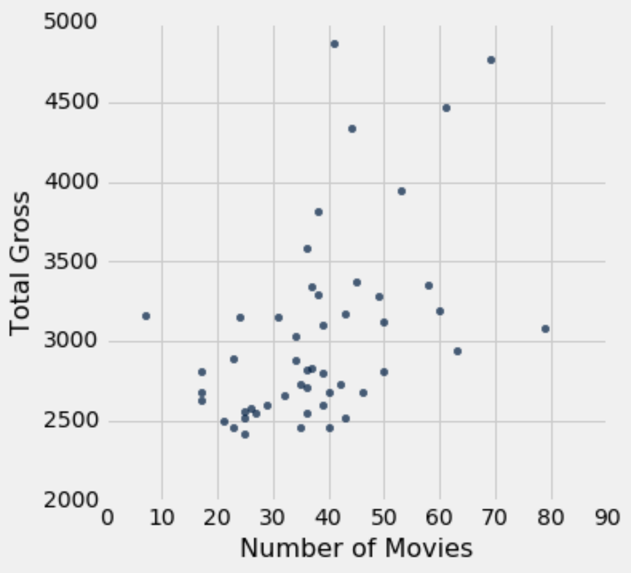
30% of accidental deaths of males were due to automobile accidents

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

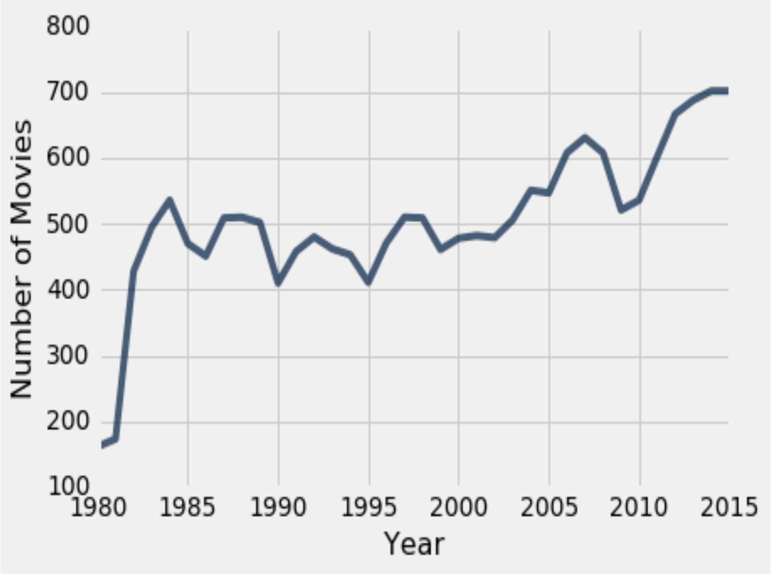
Example from Tian Zheng

EXAMPLES OF CHART TYPES

- Scatter plots



- Line graphs



(Demo) – notebook 3.1 Charts

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

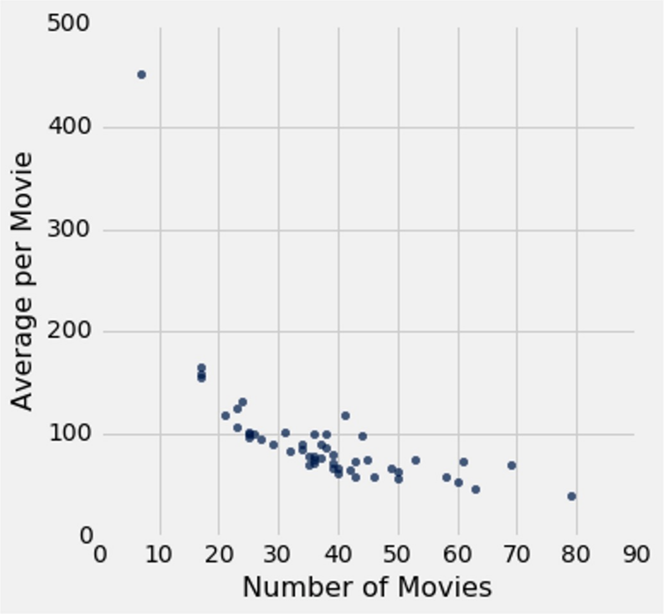
- Census example had numerical SEX code (0, 1, and 2)
- It doesn't make sense to perform arithmetic on these "numbers", e.g. 1 - 0 or (0+1+2)/3 are nonsense here
- The variable SEX is still categorical, even though numbers were used for the categories

TERMINOLOGY

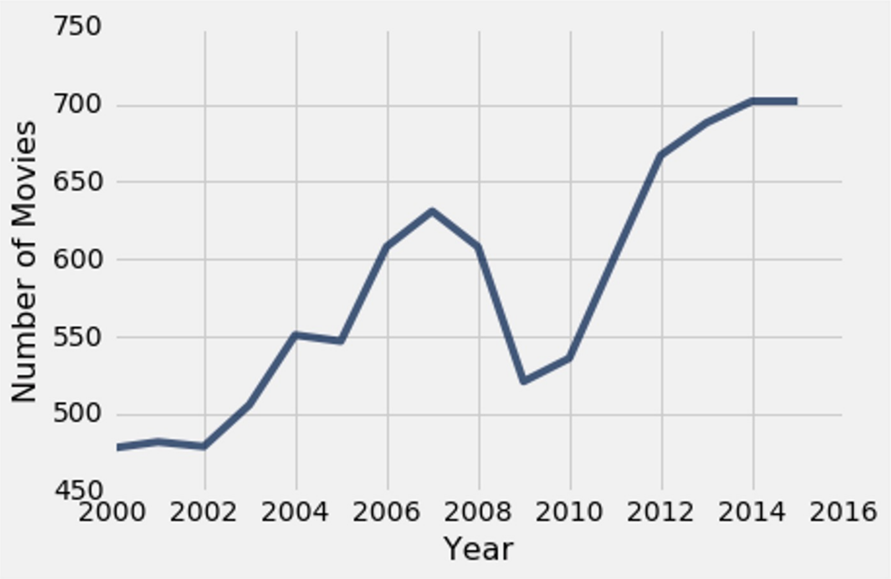
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the frequency of 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 DISTRIBUTION

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

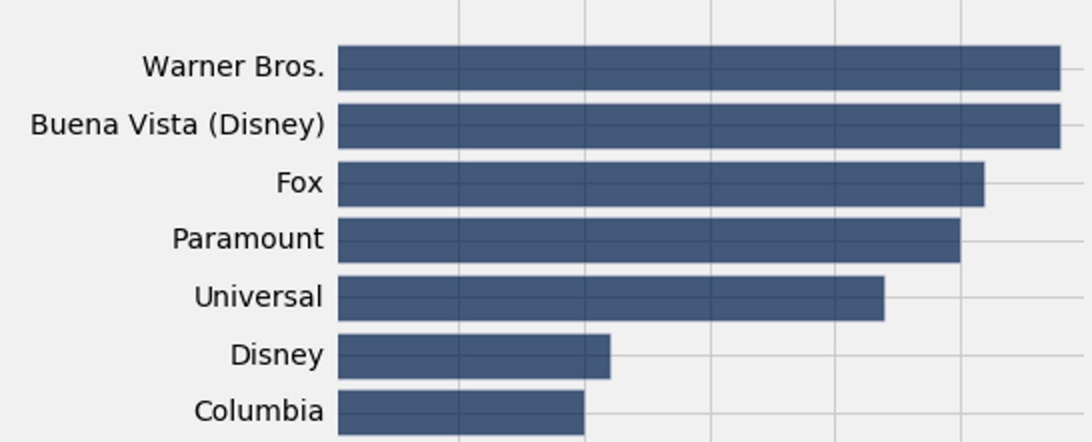
Bar charts can display the distribution of categorical values

- Ex. 1: Proportion of how many US residents are male or female
- Ex. 2: Count of how many top movies were released by each studio

(Demo) – notebook 3.1
categorical distributions

CATEGORICAL DISTRIBUTIONS

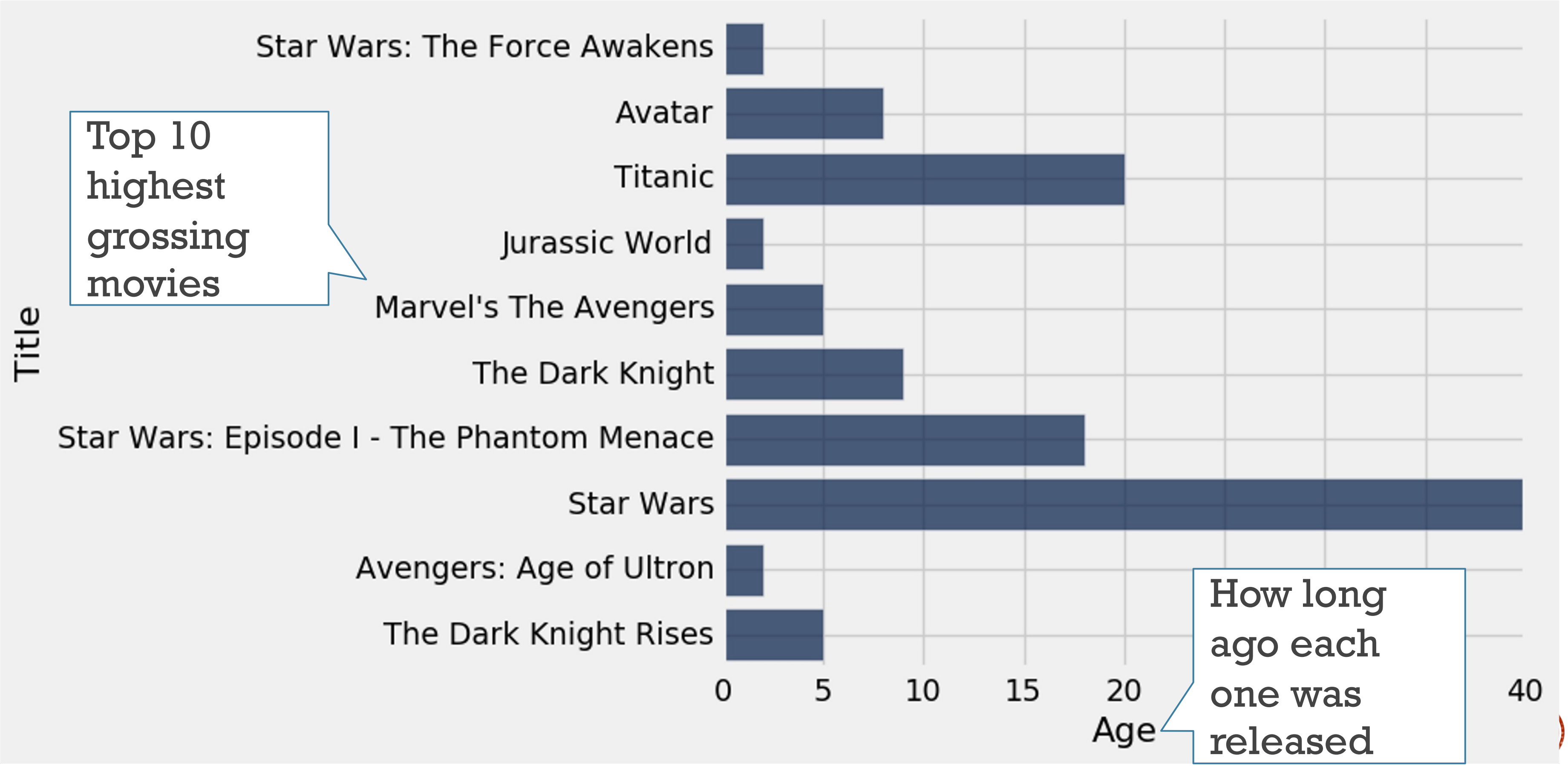
bar chart: `barh`



Displays a categorical distribution

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

HOW DO YOU GENERATE THIS CHART?



Top 10 highest grossing movies

How long ago each one was released

NUMERICAL DISTRIBUTION

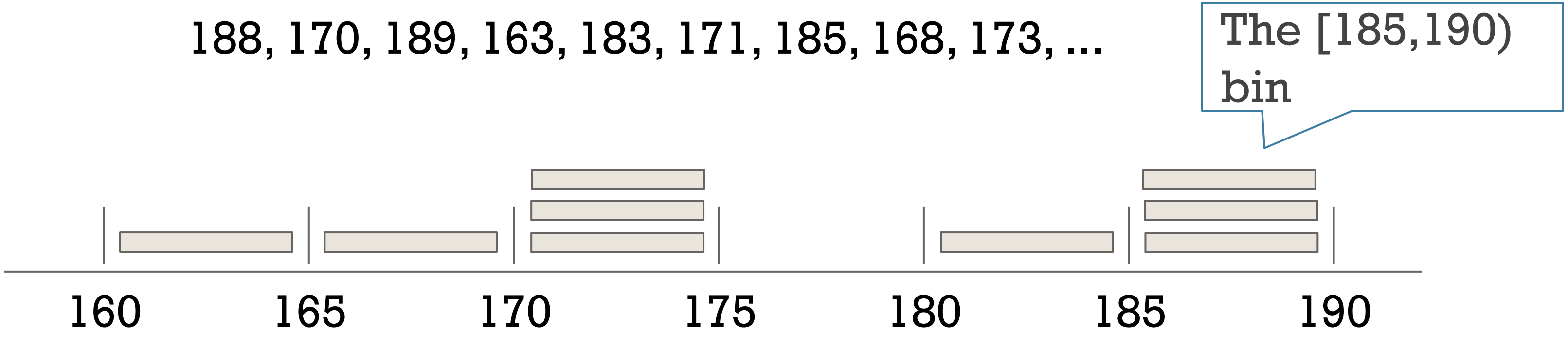
BINNING

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

188, 170, 189, 163, 183, 171, 185, 168, 173, ...

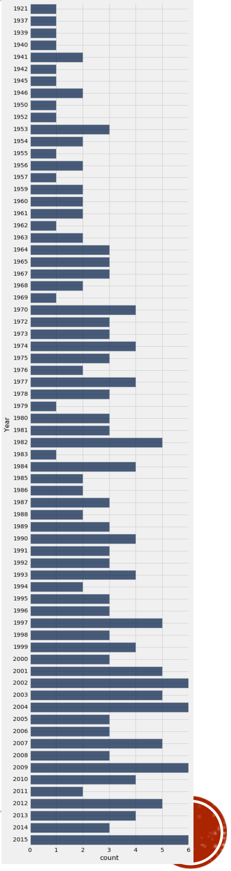


WHY BIN?

```
movies_and_years = top.select('Title', 'Year')
movies_and_years.group('Year').sort('count', descending=True).barh('Year')
```

The chart to the right has many issues:

1. The bars at 1921 and 1937 are just as far apart from each other as the bars at 1937 and 1939.
2. The bar chart doesn't show that none of the 200 movies were released in the years 1922 through 1936, nor in 1938.
3. Shows the bar chart is unsuitable for visualization such data (data that's ordered some way/has some rank)
4. The individual bars are too many, we need to group them somehow (hence binning)



HISTOGRAM

Chart to display the distribution of numerical values using bins

(Demo) – notebook 3.1
binning and histograms

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) – notebook 3.1
the density scale

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

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

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

HEIGHT MEASURES DENSITY

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

- 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**.

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

COMPARING HISTOGRAMS

OVERLAID GRAPHS

For visually comparing two populations

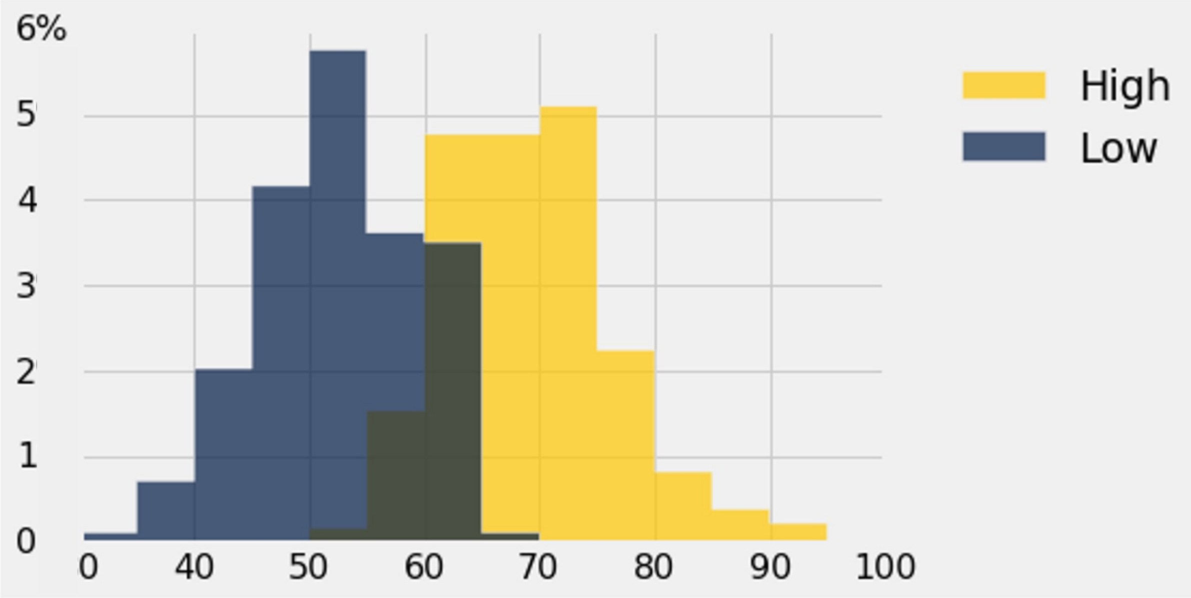
(Demo)

DISCUSSION QUESTION

This histogram describes a **year** of daily temperatures

Try to answer these questions:

- What proportion of days had a high temp in the range 60-69?
- What proportion had a low of 45 or more?
- How many days had a difference of more than 20 degrees between their high & low temperatures?



QUESTIONS?