DATA 8

Fall 2016

Lecture 8, September 12

Histograms

Slides created by Ani Adhikari and John DeNero

Announcements

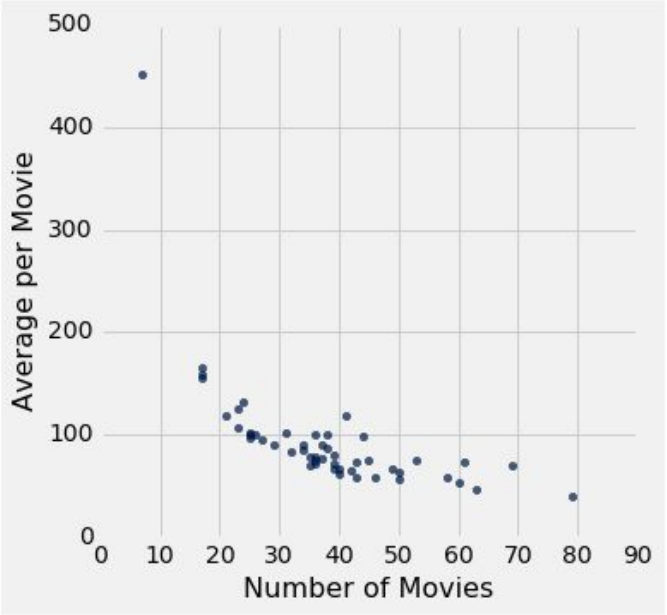
- **Waitlisted students:** I'm in touch with everyone and we're working on it.
- **No late work.** If you joined the class late, please do current work. We'll prorate based on when you joined.
- **Technical problems? Questions about hw/lab credit?** Please email your GSI, not me. Addresses are on the Staff Contact page in data8.org.
- **Python Playground: NEW!** There's a link on Piazza. A place for you to experiment with Python and the class datasets.

Terminology

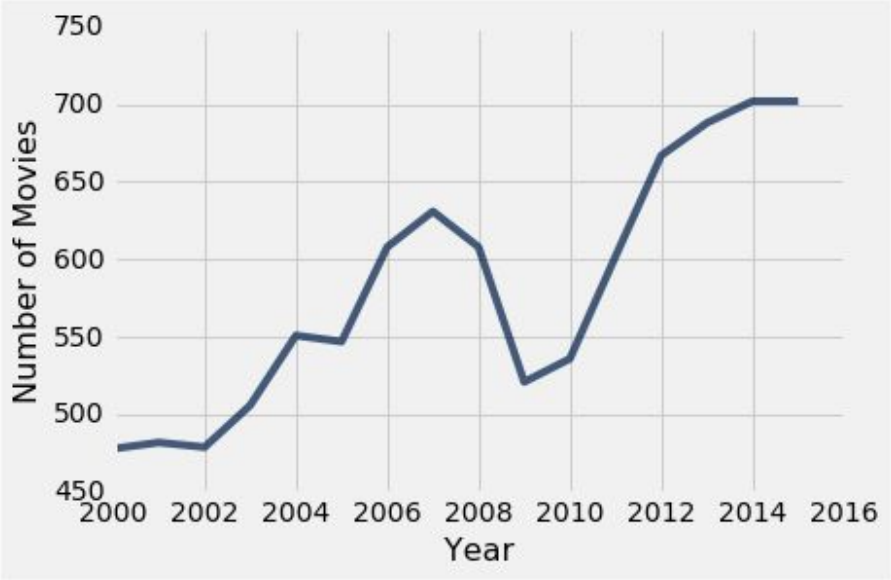
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the proportion of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

Plotting Two Numerical Variables

Scatter plot: `scatter`

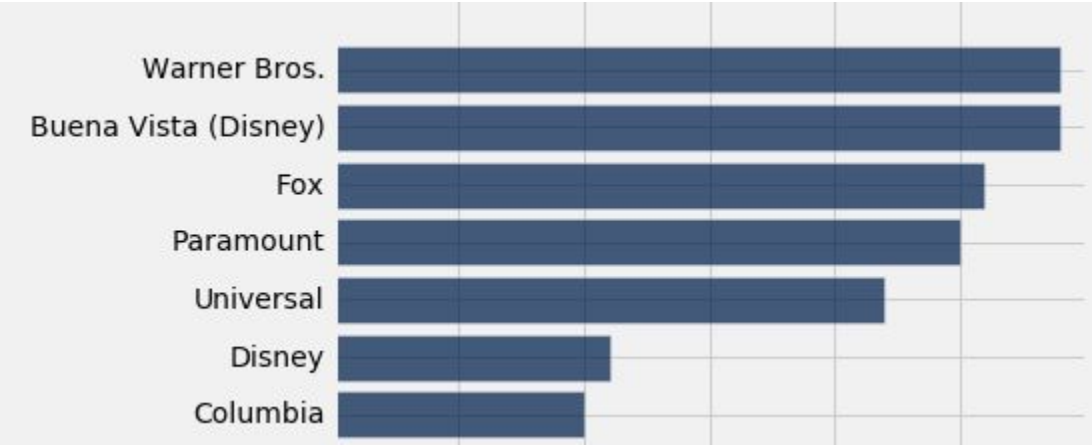


Line graph: `plot`



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.)

“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).
- Doesn’t make sense to do 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 as codes.

Numerical Data

A **histogram** displays the distribution of a numerical variable.

(Demo)

How to Calculate Height

The [300, 400) bin contains 81 out of 200 movies.

- “81 out of 200” is 40.5%
- The bin is 400 - 300 = 100 million dollars wide

$$\text{Height of bar} = \frac{40.5\ \%}{100 \text{ million dollars}}$$

$$= 0.405\ \% \text{ per million dollars}$$

(Demo)

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.

Bar Chart vs. Histogram

Bar Chart

- Categorical data
- Bars have arbitrary (but equal) widths and spacings
- Height (or length) of bars proportional to percent of individuals

Histogram

- Numerical data
- Horizontal axis is numerical, hence to scale with no gaps
- Height measures density; areas are percents

Overlaid Graphs

For visually comparing two populations

(Demo)