190F Foundations of Data Science

Fall 2018

Lecture 7

Charts

Announcements

Data and Visualization

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

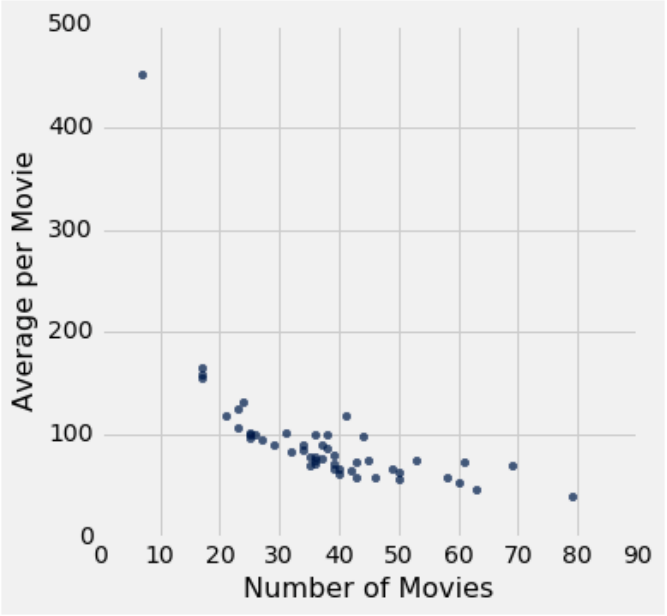
Just because the values are numbers, doesn’t mean the variable is numerical

- 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

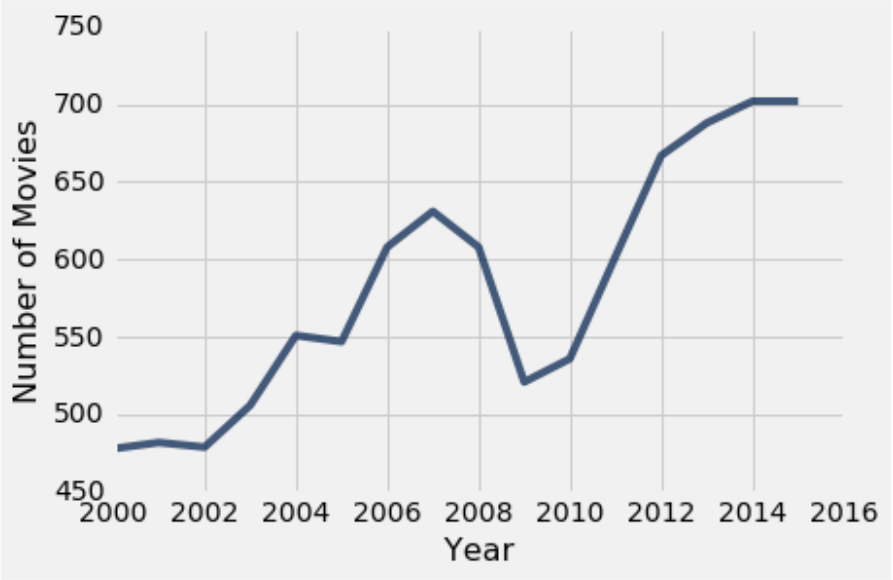
Visualizing Numerical Data

Plotting Two Numerical Variables

Scatter plot: `scatter`



Line graph: `plot`



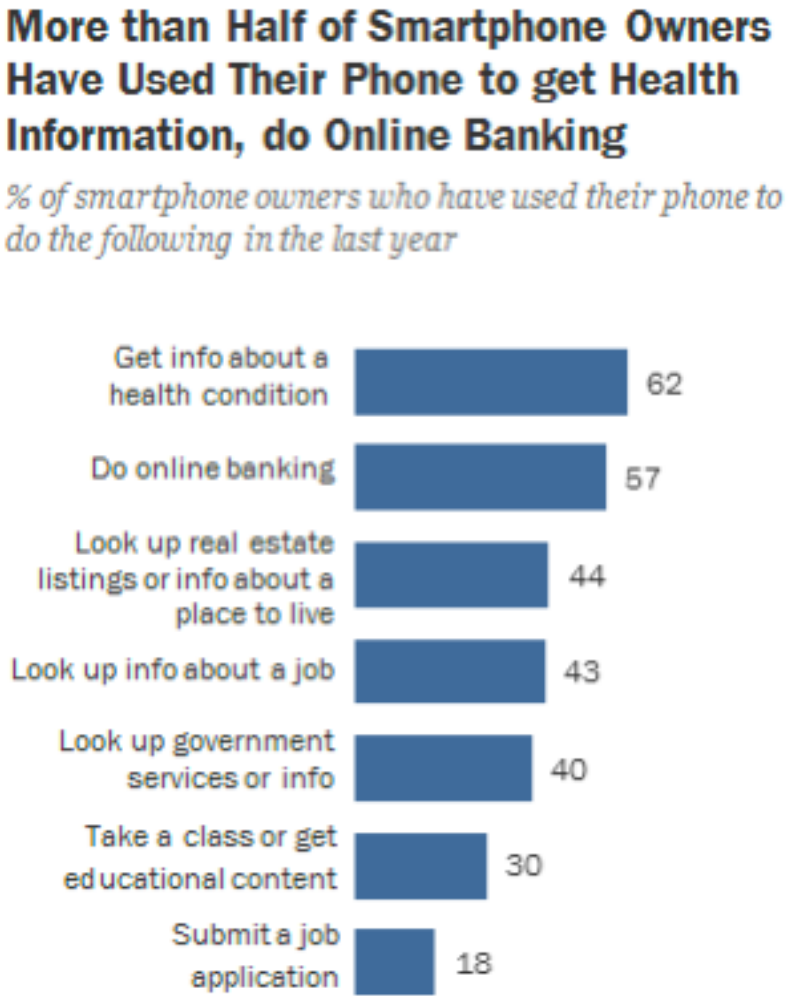
Visualizing Categorical Data

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

Bar Charts of Counts

Distributions:

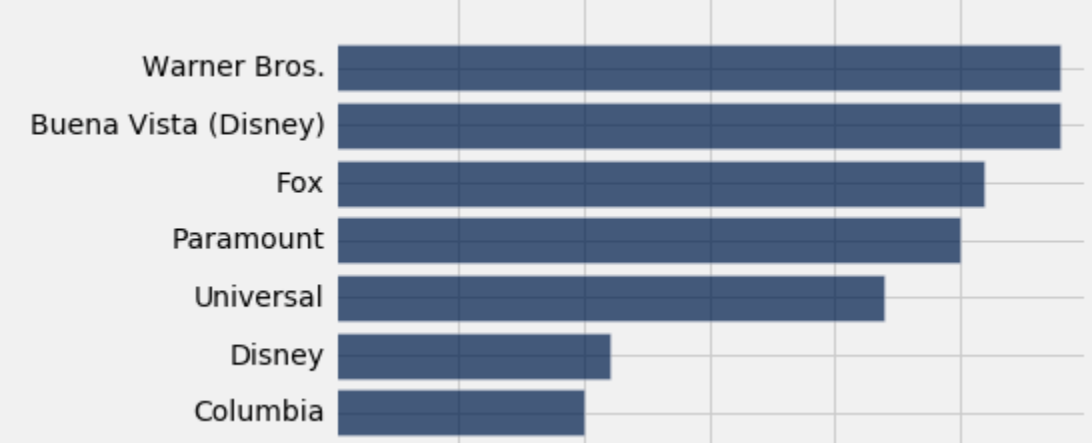
- The distribution of a variable (a column) describes the frequency of its different values
- The `group` method counts the number of rows for each value in a column

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

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.)