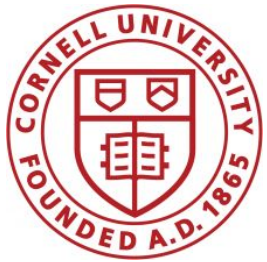


DSFA

Spring 2021

Lecture 7

Charts

Announcements

- Prelim dates finalized, rooms TBA
  - March 16, 8:30-10PM
  - April 20, 8:30-10PM
  - All Ithaca-resident students expected to show for in-person prelims/exam (whether online or not)
  - Non-resident prelim plans TBA
- HW 2 due Friday 5:59PM, 1 point bonus for Thursday submission
- Want to start using PollEverywhere more regularly Friday

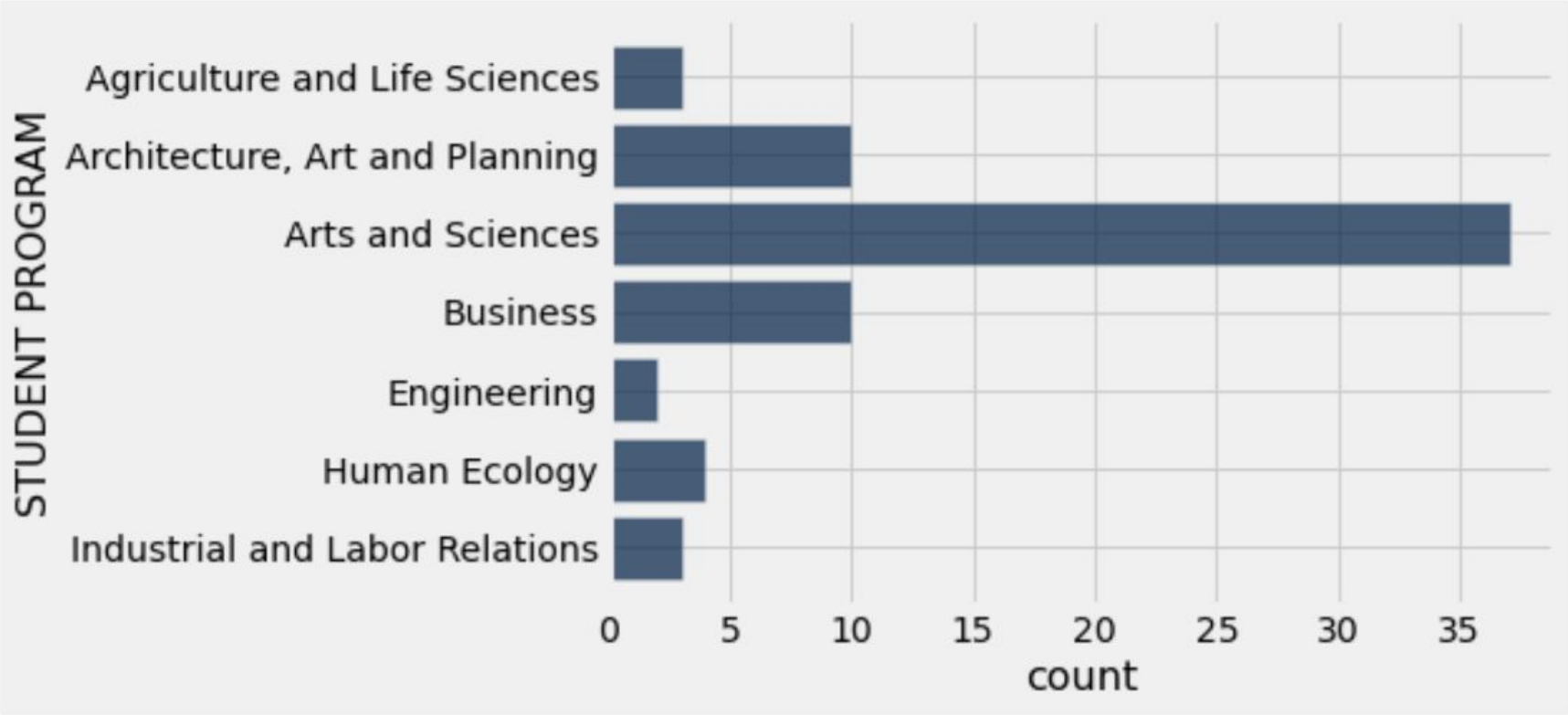
I'm registered for PollEverywhere

Yes

No

What actor/actress has made the most money per movie made?

How can we make a chart like this?



Census Continued

(Demo)

Data 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

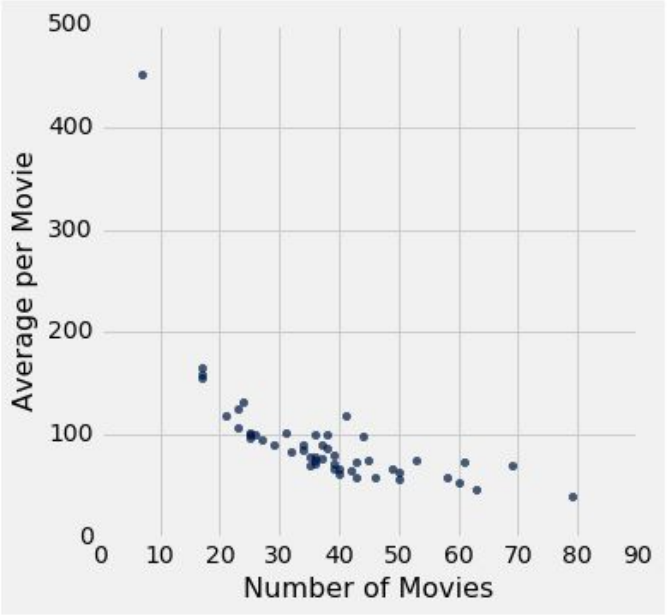
- Census example had numerical `SEX` code (0, 1, and 2)
- It doesn’t make sense to perform arithmetic on these “numbers”, e.g. 1 - 0 or (0+1+2)/3 are nonsense here
- The variable `SEX` is still categorical, even though numbers were used for the categories

Terminology

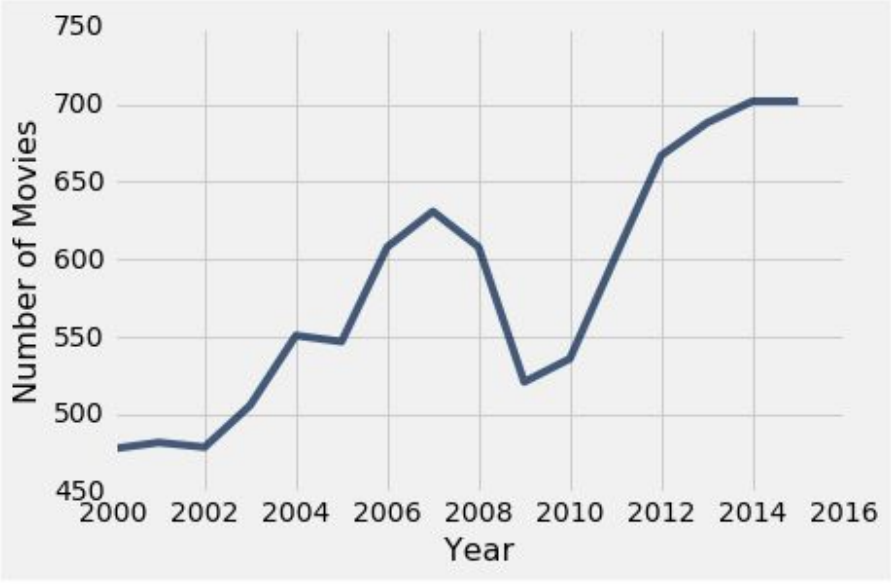
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- Often:
  - Individual = row
  - Variable or feature = column
- **Distribution**: 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`



Numerical Data

(Demo)

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

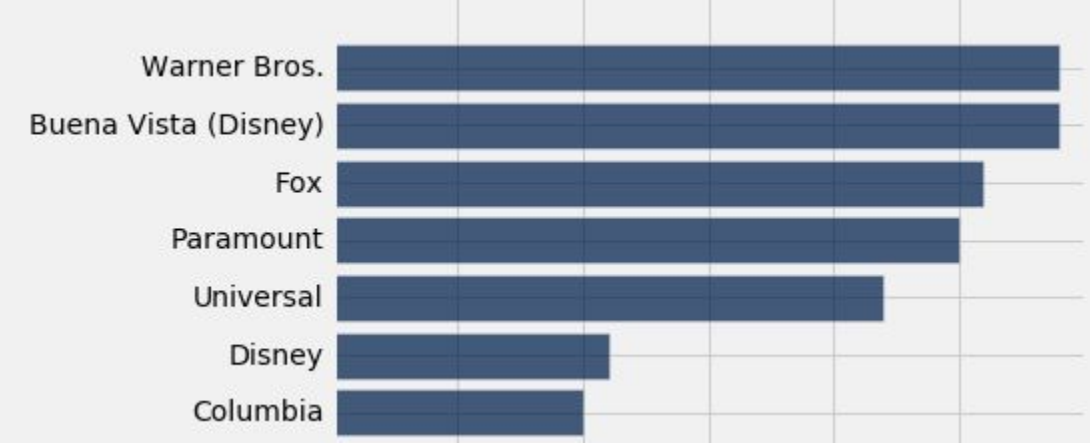
Bar charts can display the distribution of categorical values

- 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`**



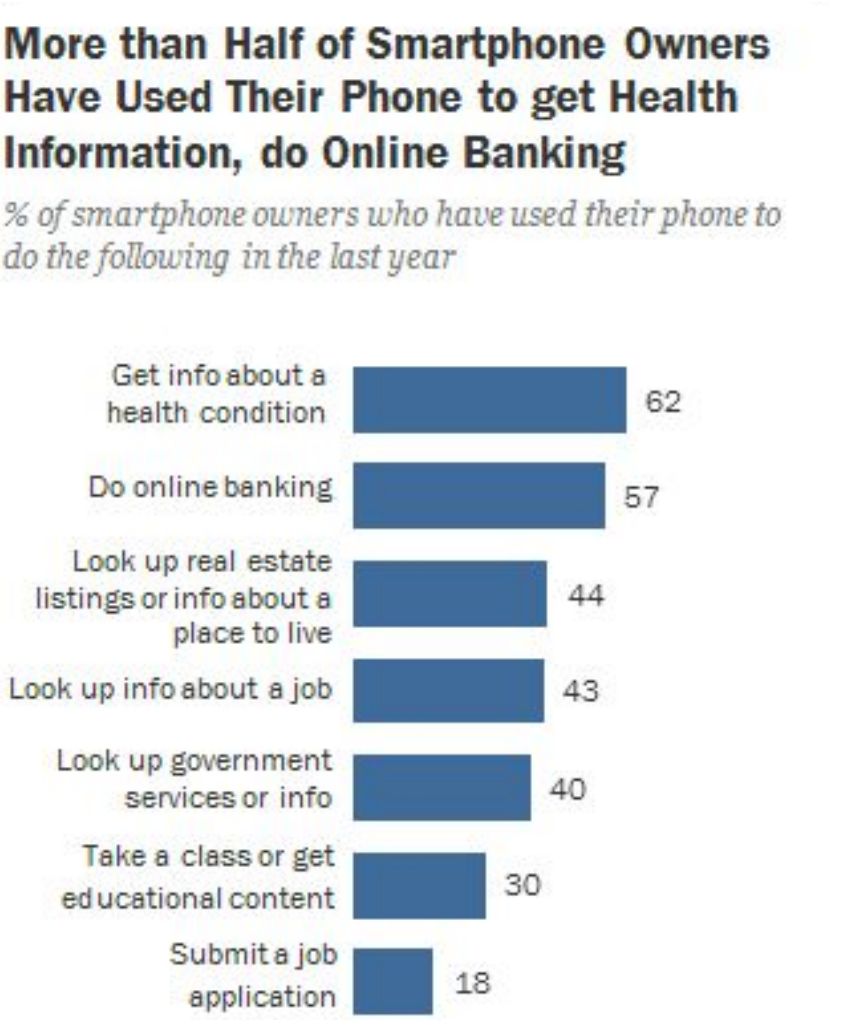
Displays a categorical distribution

Discussion Question

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

Among survey responders...

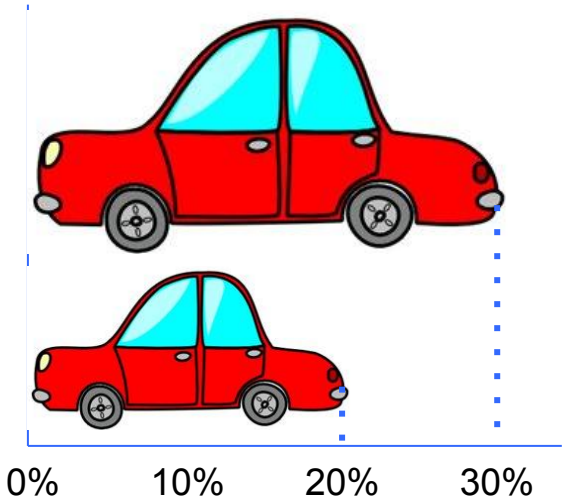
- What proportion did **not** use their phone for online banking?
- What proportion either used their phone for online banking or to look up real estate listings?
- Did everyone use their phone for at least one of these activities?
- Did anyone use their phone for both online banking and real estate?



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