Lecture 3

Visualization

Tables Review

Manipulating Rows

- `t.sort(column)` sorts the rows in increasing order
- `t.take(row_numbers)` keeps the numbered rows
  - Each `row` has an index, starting at 0
- `t.where(column, are.condition)` keeps all rows for which a column's value satisfies a condition
- `t.where(column, value)` keeps all rows for which a column's value equals some particular value

Discussion Questions

The table **`nba`** has columns **`NAME`**, **`POSITION`**, and **`SALARY`**.

a) Create an array containing the names of all point guards (**`PG`**) who make more than $15M/year

```
nba.where(1, 'PG').where(2, are.above(15)).column(0)
```

b) After evaluating these two expressions in order, what's the result of the second one?

```
nba.with_row(['Mahdi Zareei', 'Mascot', 100])
nba.where('POSITION', are.containing('Mascot'))
```

Census Data

The Decennial Census

- Every ten years, the Census Bureau counts how many people there are in the U.S.
- In between censuses, the Bureau estimates how many people there are each year.

Analyzing Census Data

Leads to the discovery of interesting features and trends in the population

(Demo)

Census Table Description

Sort order of observations: SEX and AGE

Data fields (in order of appearance):

VARIABLE	DESCRIPTION
SEX	Sex
AGE	Age
CENSUS2010POP	4/1/2010 resident Census 2010 population
ESTIMATESBASE2010	4/1/2010 resident population estimates base
POPESTIMATE2010	7/1/2010 resident population estimate
POPESTIMATE2011	7/1/2011 resident population estimate
POPESTIMATE2012	7/1/2012 resident population estimate
POPESTIMATE2013	7/1/2013 resident population estimate
POPESTIMATE2014	7/1/2014 resident population estimate
POPESTIMATE2015	7/1/2015 resident population estimate

The key for SEX is as follows:

0 = Total
1 = Male
2 = Female

AGE is single-year of age (0, 1, 2, ...99, 100+ years) and 999 is used to indicate total population.

http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.pdf

Census Table Description

- Values have column-dependent interpretations
  - The SEX column: 1 is *Male*, 2 is *Female*
  - The POPESTIMATE2010 column: *7/1/2010 estimate*
- In this table, some rows are sums of other rows
  - The SEX column: 0 is *Total* (of *Male* + *Female*)
  - The AGE column: 999 is *Total* of all ages
- Numeric codes are often used for storage efficiency
- Values in a column have the same type, but are not necessarily comparable (AGE 12 vs AGE 999)

Demo

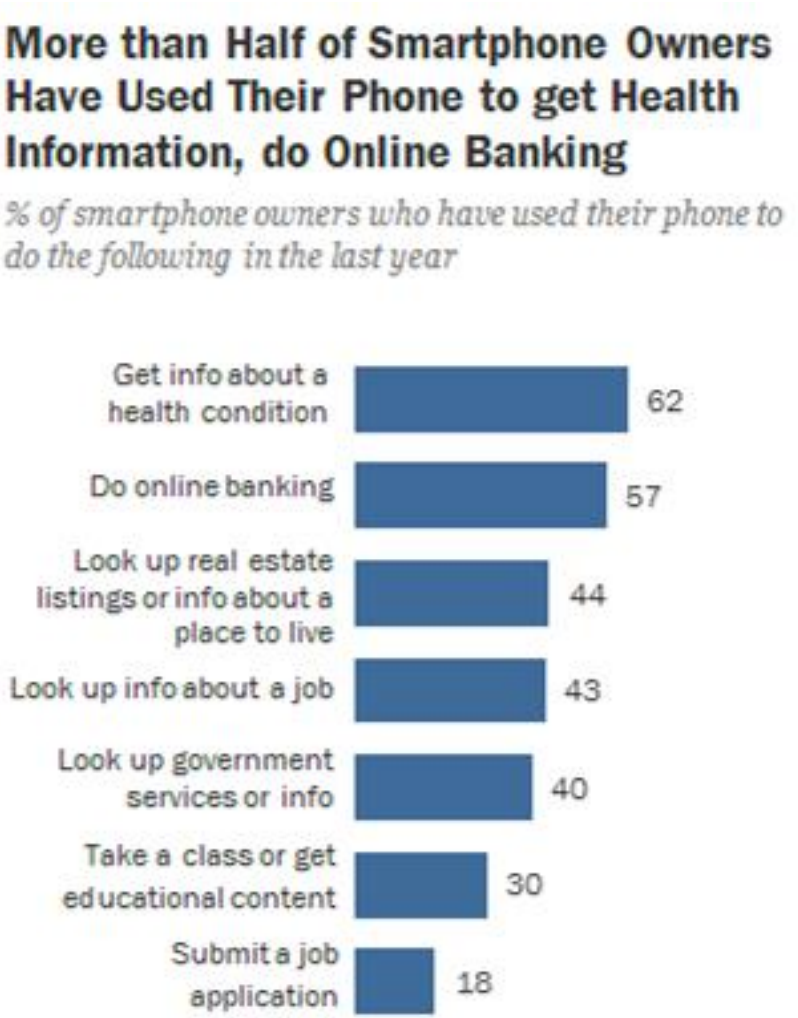
Data Visualization

Discussion Question

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

Among survey responders...

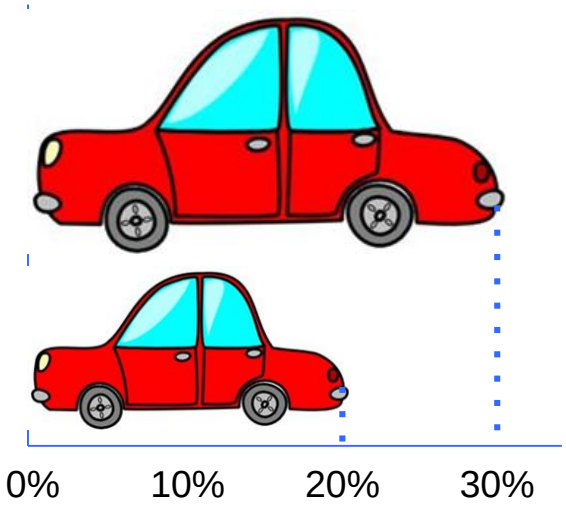
- What proportion did **not** use their phone for online banking?
- What proportion either used their phone for online banking or to look up real estate listings?
- Did everyone use their phone for at least one of these activities?
- Did anyone use their phone for both online banking and real estate?



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

Visualization

A picture is worth a thousand numbers...

(Demo)

Numerical Data

Types of Data

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

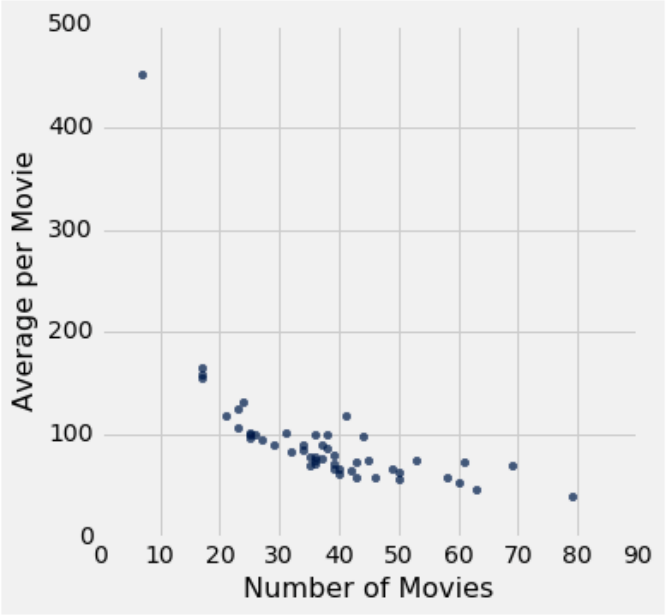
- **Numerical** — Each value is from a fixed scale
  - Numerical measurements are ordered
  - Differences are typically meaningful
- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering
  - Categories are the same or different

Terminology

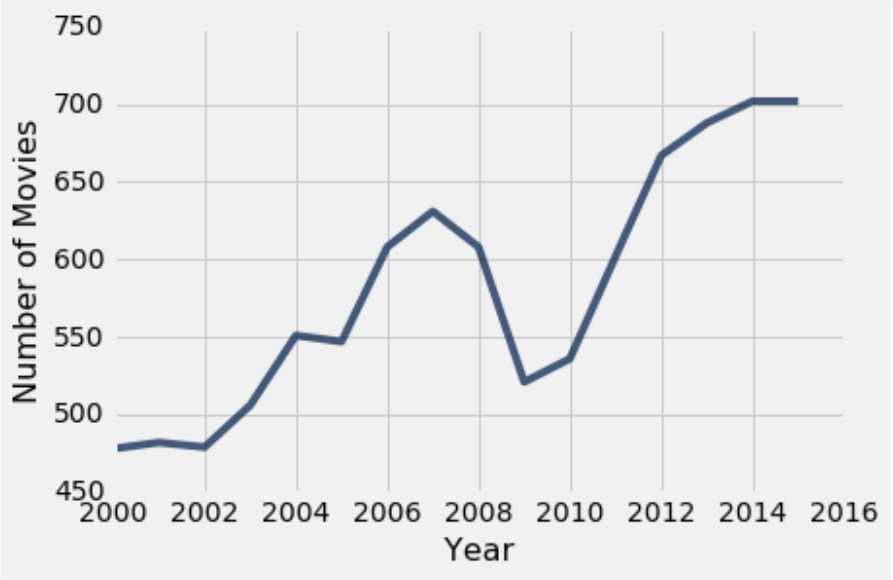
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the frequency of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

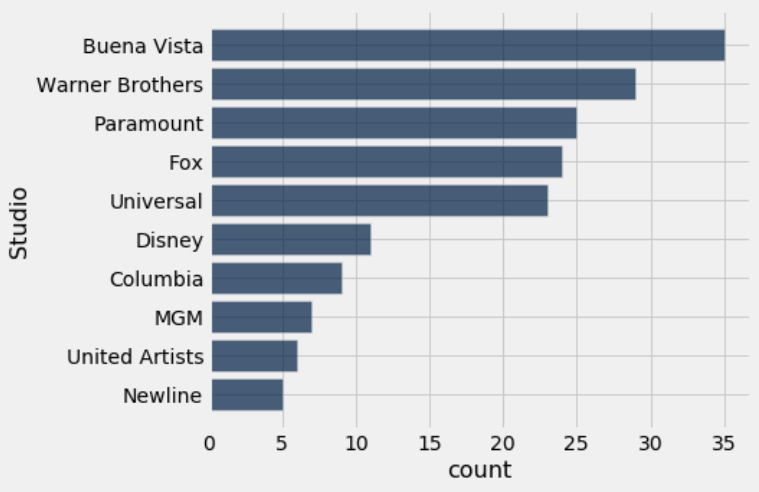
“Numerical” Data

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.

Categorical Distributions

bar chart: **`barh`**

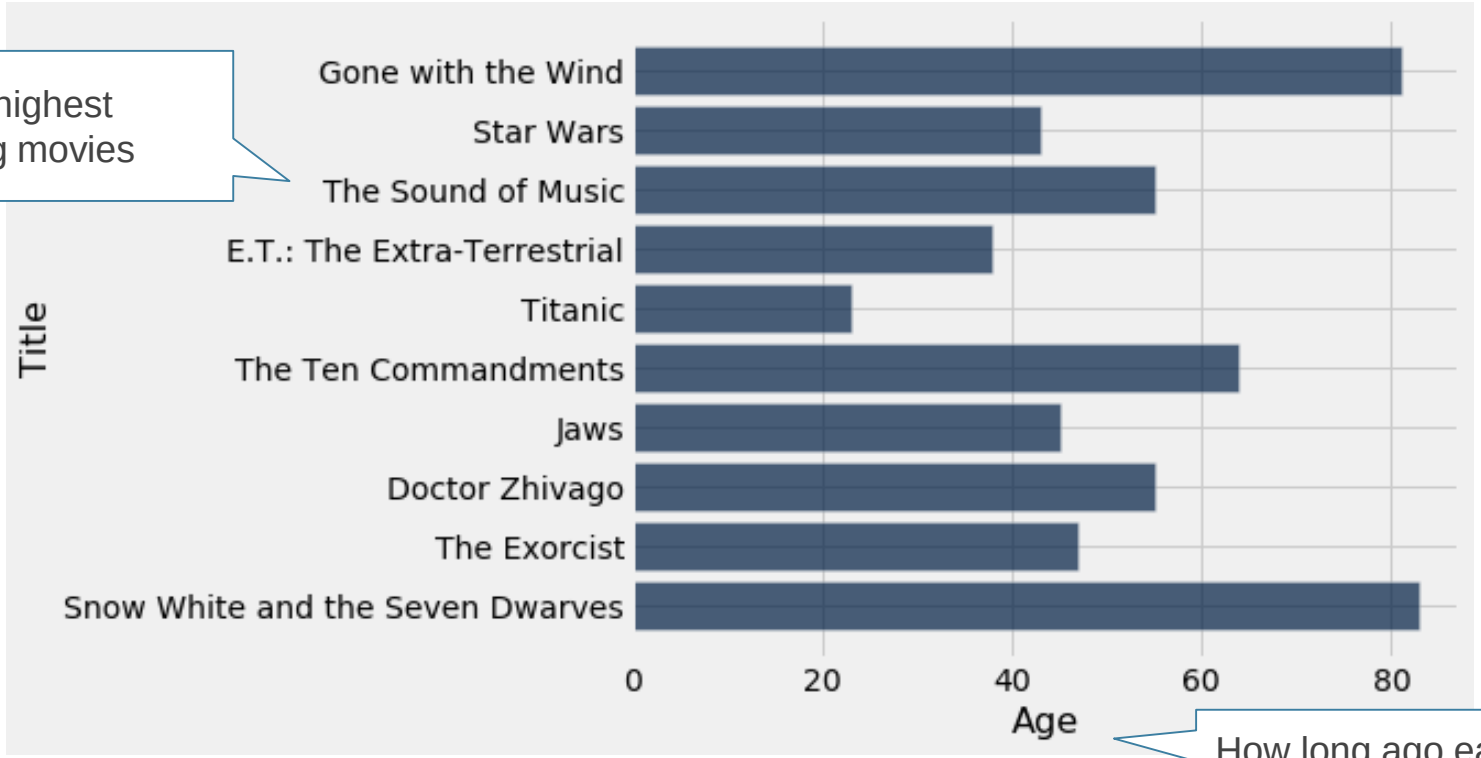


Displays a categorical distribution

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

Discussion Question

Top 10 highest grossing movies



How long ago each one was released

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

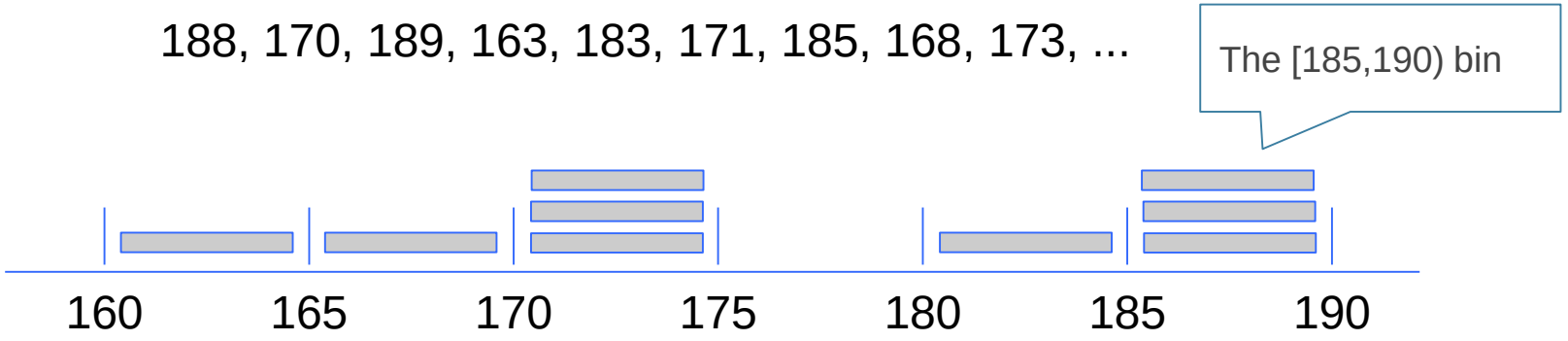
- Proportion of how many US residents are male or female
- Count of how many top movies were released by each studio

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.

Chart Types

Bar Chart vs. 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

Overlaid Graphs

For visually comparing two populations

(Demo)