DSC 10, Spring 2018
Lecture 7

Charts

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

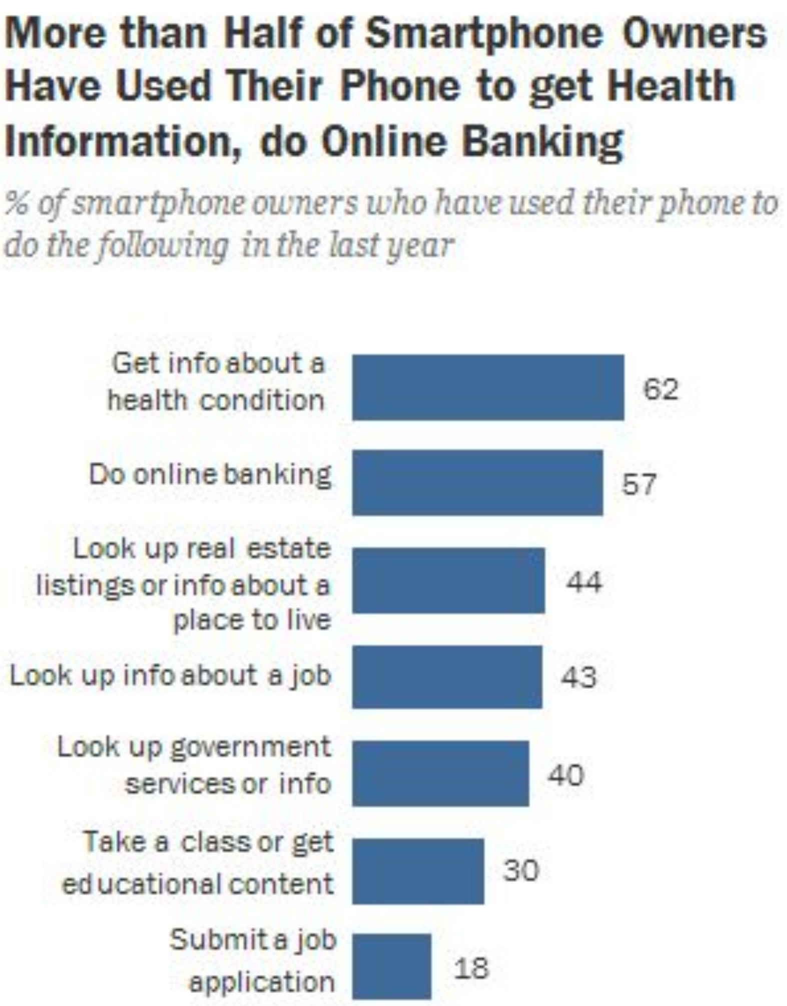
Data Visualization

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?



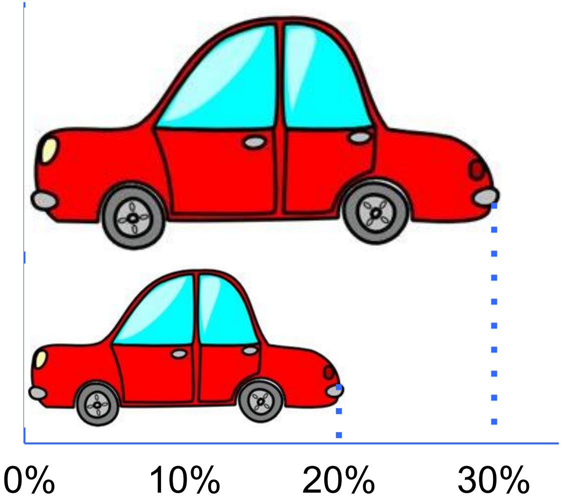
More than Half of Smartphone Owners Have Used Their Phone to get Health Information, do Online Banking

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

Pew research center, 2014

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

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

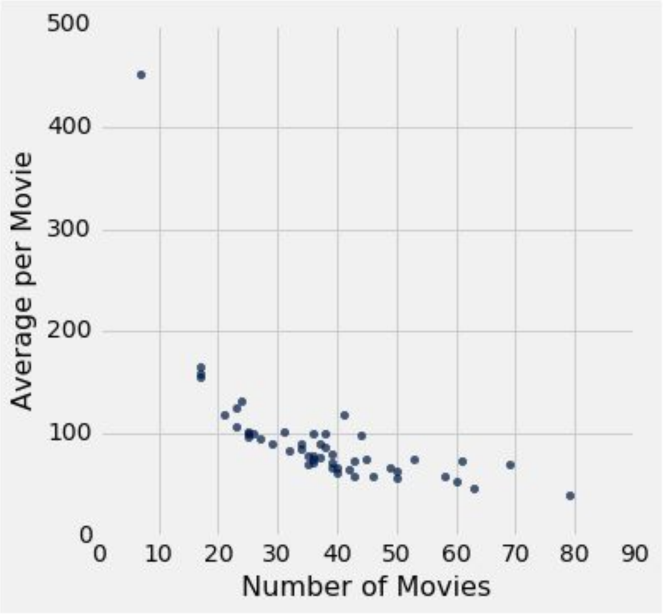
- **Numerical** — Each value is from a fixed scale
    - Numerical measurements are ordered
    - Differences are typically meaningful
- **Categorical** — Each value is from a fixed inventory
    - May or may not have an ordering
    - Categories are either the same or different

Terminology

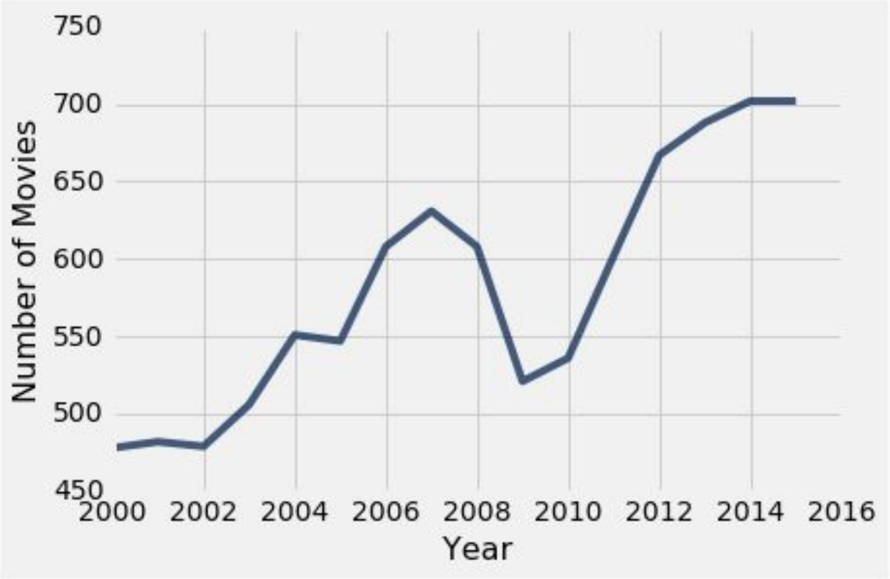
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the frequency of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



(Demo)

Categorical Data

Numerical or Categorical?

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.

Bar Charts

Compare some quantity across categories

- % of smartphone owners who have used their phone for the following in the last year: online banking, job search, etc.
- Gross ticket sales for individual movies

(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 with 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

Question

Suppose we execute this code:

```
aged = top.with_column("Age", 2017-top.column('Year'))
aged.group('Age').barh(‘Age’)
```

What type of bar graph will be produced?

A. A bar for each movie.
The length of the bar is the age of the movie.

B. A bar for each age.
The length of the bar is the number of movies of that age.

C. A bar for each year.
The length of the bar is the age of movies made that year.

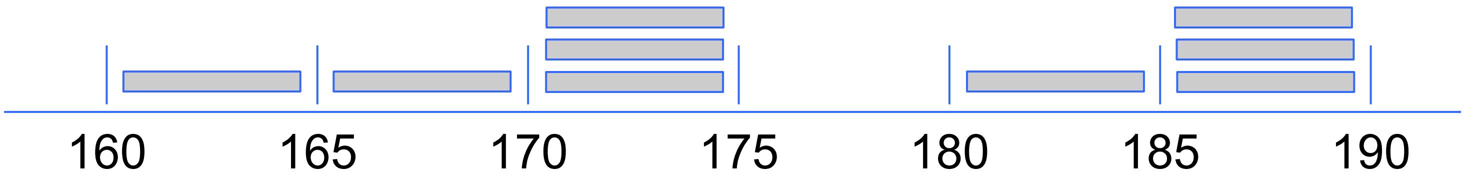
Binning

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

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



The [185,190) bin

Histogram

Chart displaying the distribution of numerical values using bins

(Demo)