

DSFA

Spring 2018

Lecture 7

Charts

Announcements

- Lab 3 is posted
 - Clicker questions start Friday
 - Register your clicker on Blackboard
 - Bring your clicker to class on Friday
 - Homework 3 will be released Friday, due next Thursday
-

Census Continued

(Demo)

Data Visualization

Discussion Question

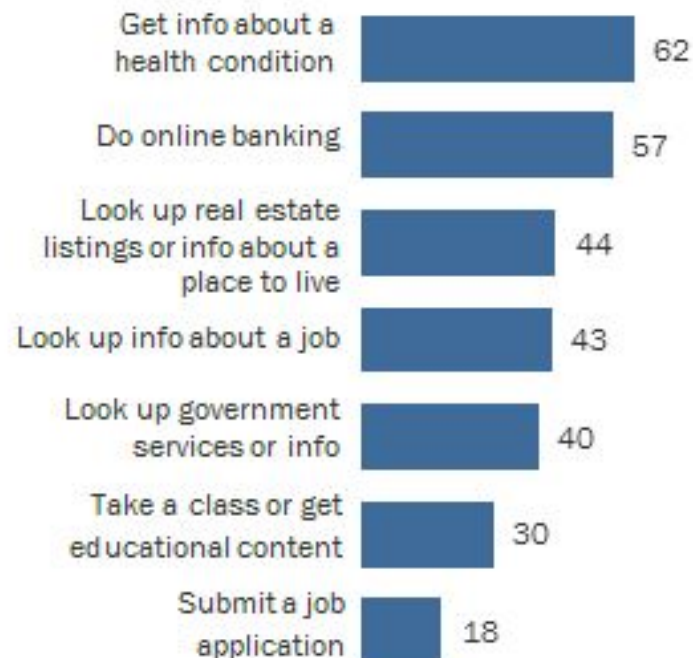
Which of the following questions can be answered by this chart?

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**?

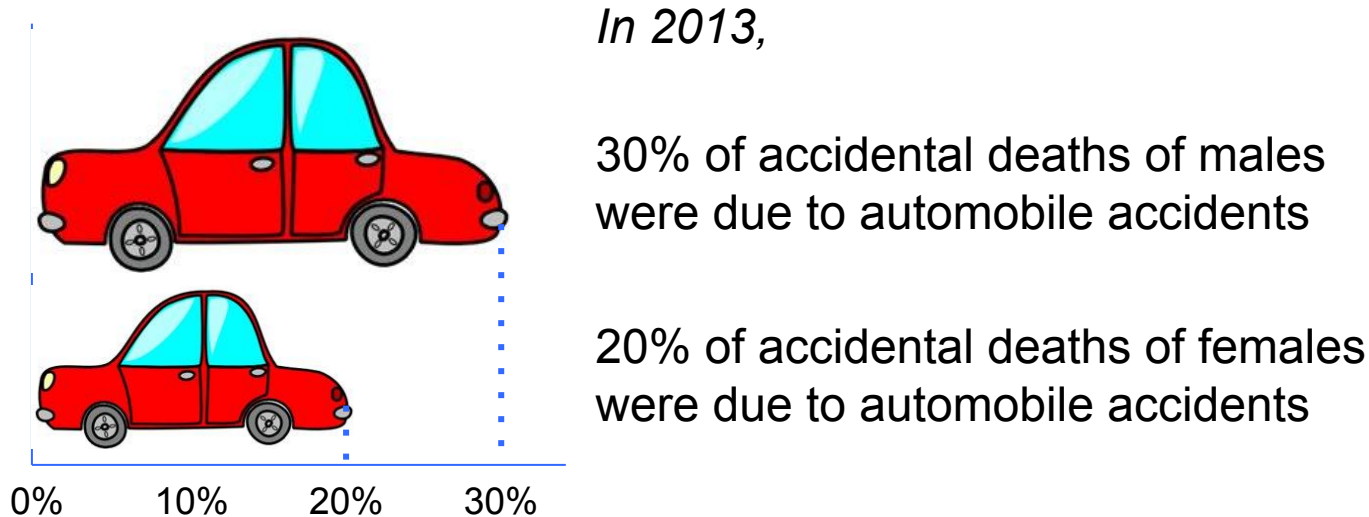
More than Half of Smartphone Owners Have Used Their Phone to get Health Information, do Online Banking

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



Area Principle

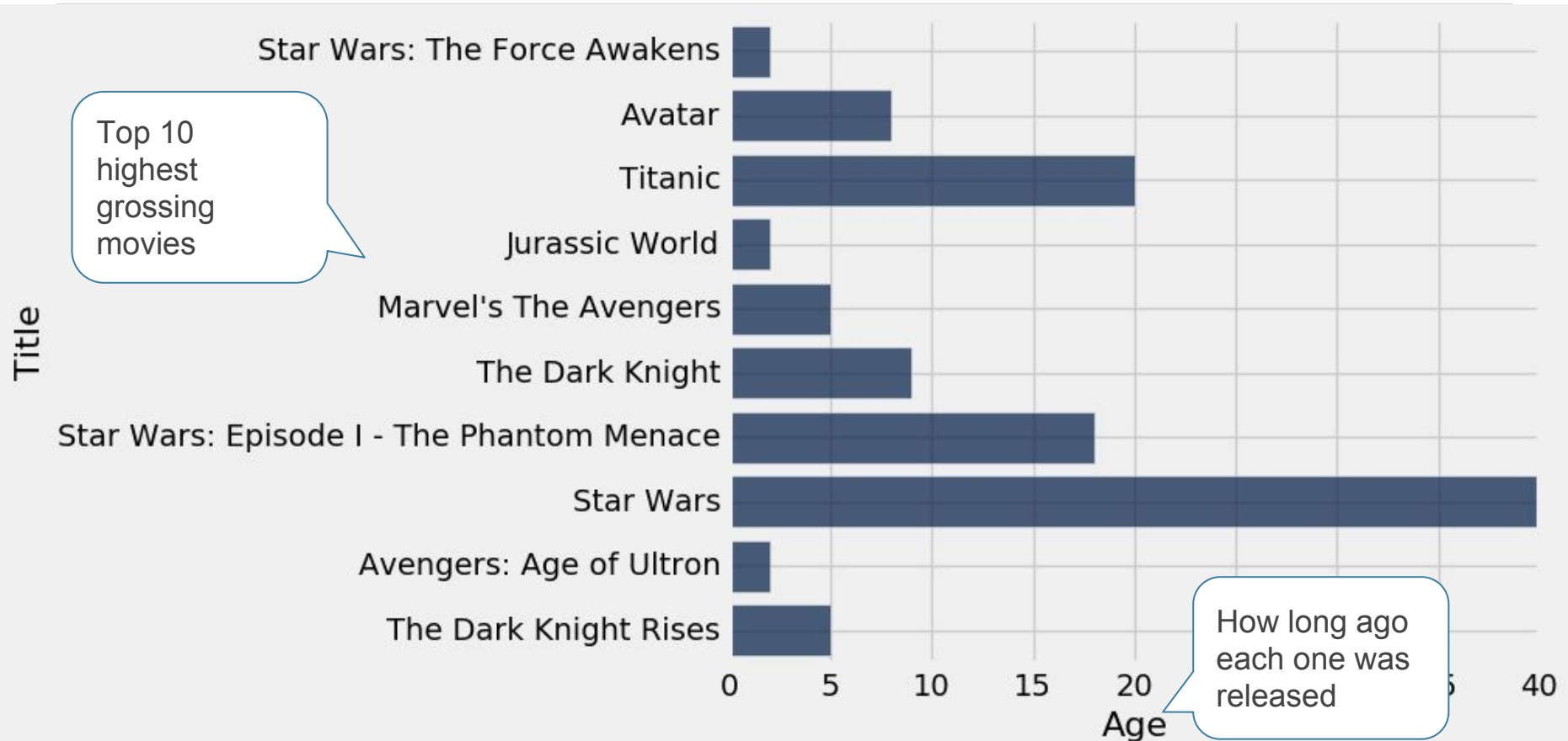
Areas should be proportional to the values they represent



Numerical Data

(Demo)

How Do You Generate This Chart?



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

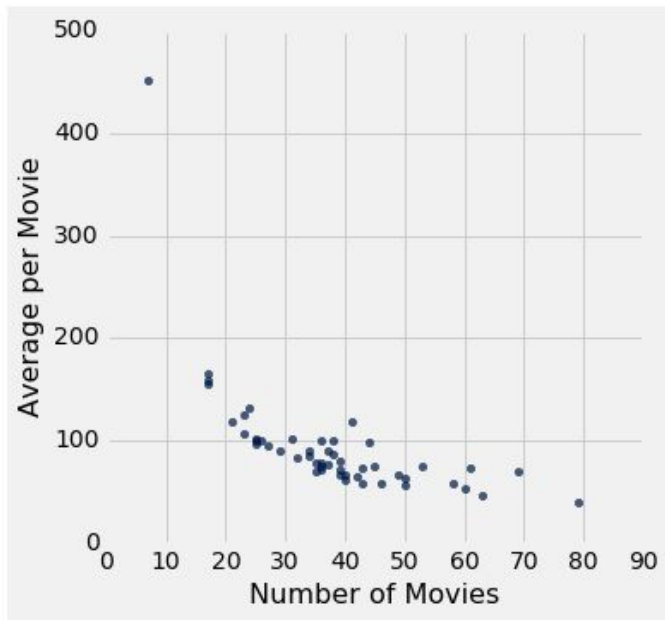
- Census example had numerical `SEX` code (0, 1, and 2)
 - It doesn't make sense to perform arithmetic on these “numbers”, e.g. $1 - 0$ or $(0+1+2)/3$ are nonsense here
 - The variable `SEX` is still categorical, even though numbers were used for the categories
-

Terminology

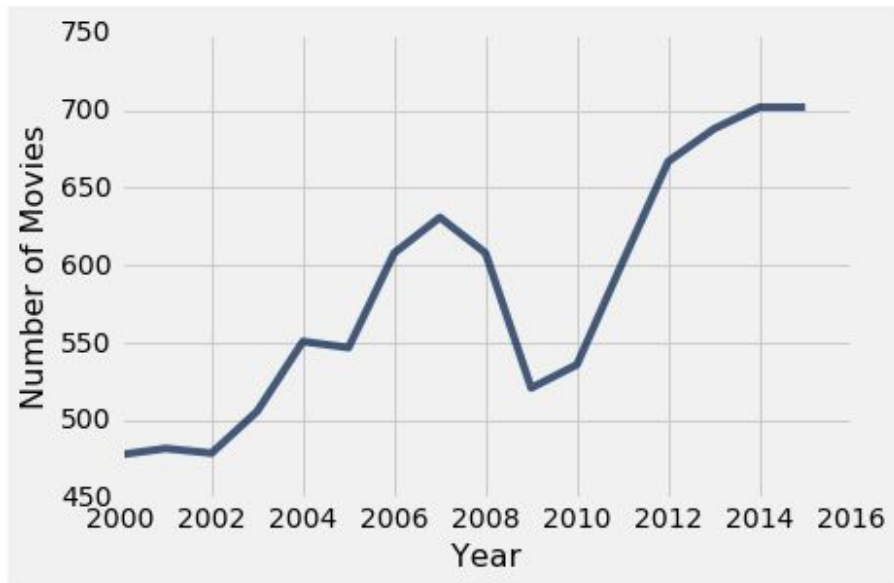
- **Individuals**: those whose features are recorded
 - **Variables**: features; these vary across individuals
 - Variables have different **values**
 - Values can be **numerical**, or **categorical**, or of many other types
 - **Distribution**: For each different value of the variable, the frequency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

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

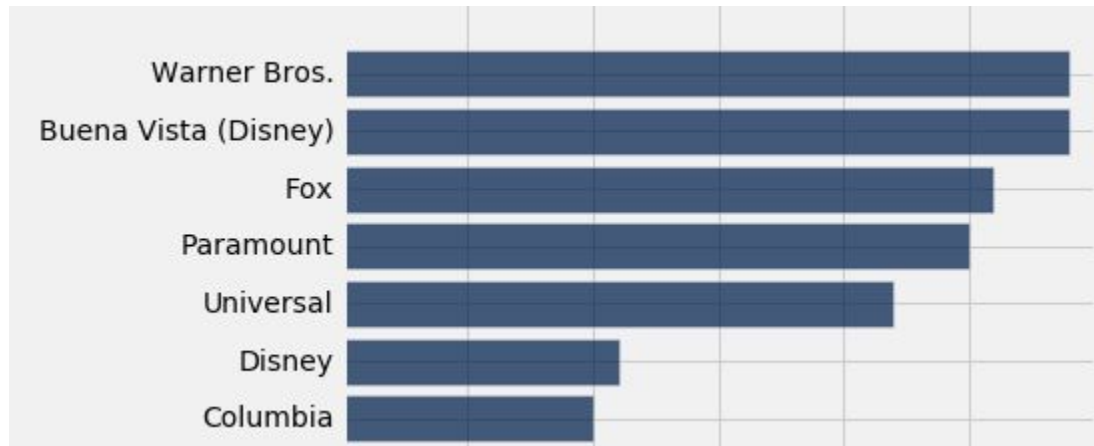
Bar charts can display the distribution of categorical values

- 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

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