YaleNUSCollege

# YSC2239 Lecture 4

# Today’s class

- Census data
- Charts
- Histograms

- Reading: Chapter 6.3, 6.4, 7

# 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.
- Article 1, Section 2 of the Constitution:
  - “Representatives and direct Taxes shall be apportioned among the several States … according to their respective Numbers …”

# 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)

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

# Analyzing Census Data

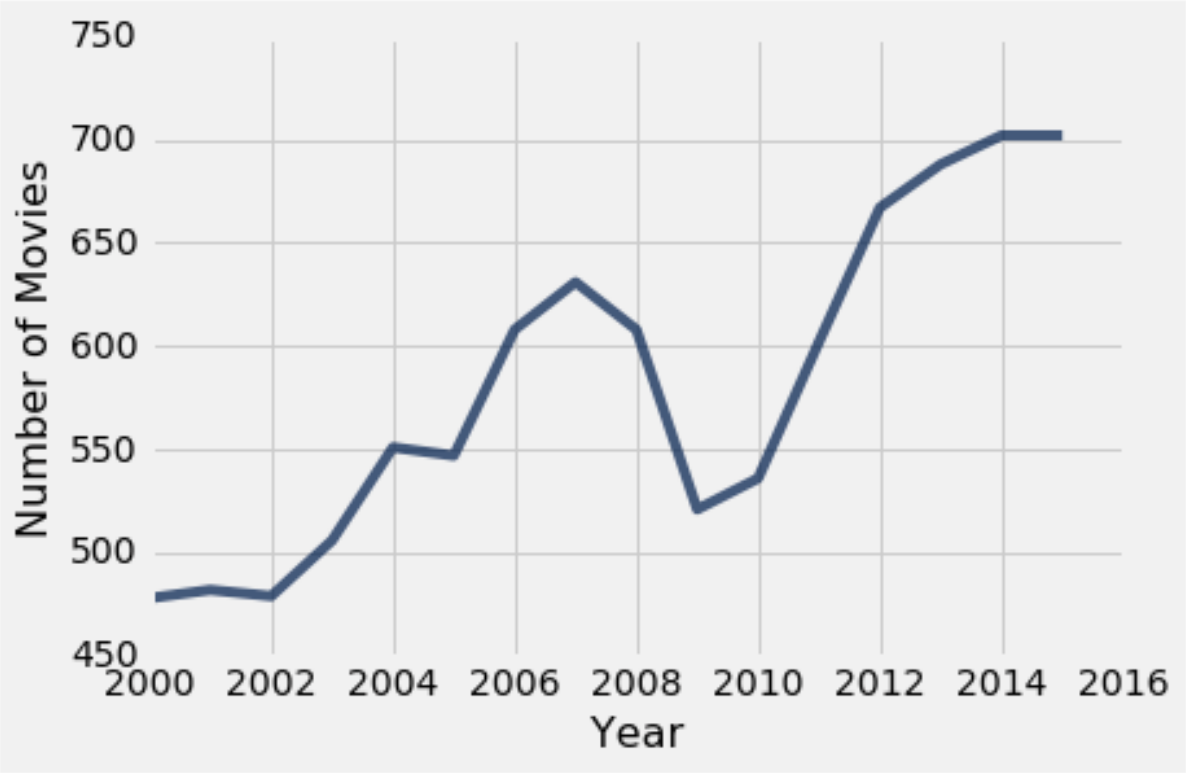
Leads to the discovery of interesting features and trends in the population

(Demo)

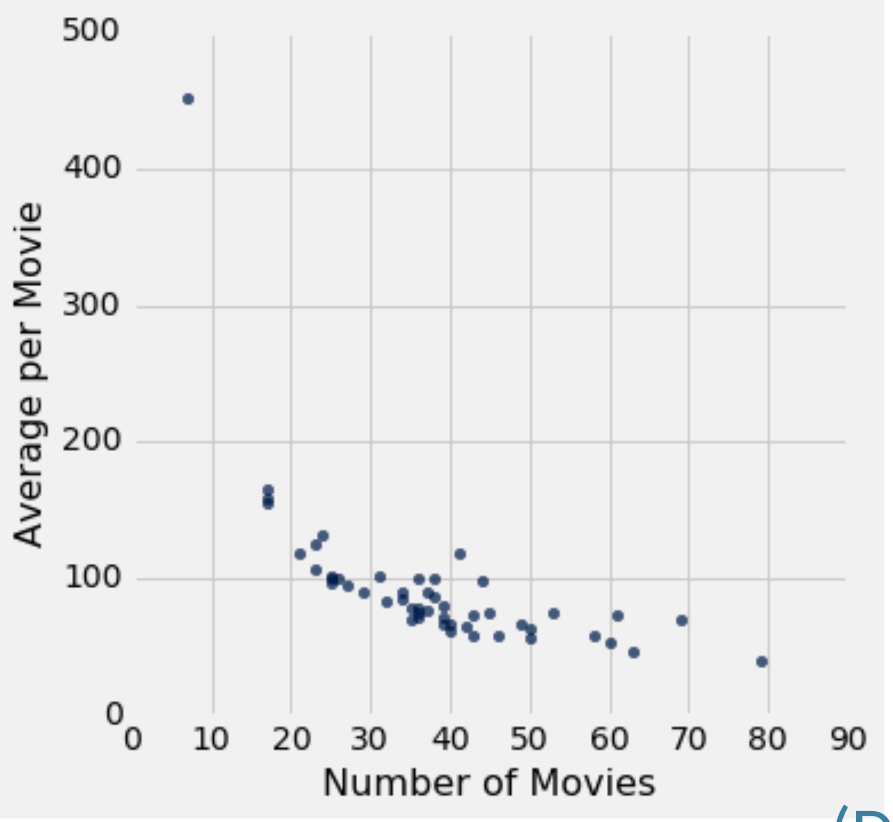
# Numerical Data

# Plotting Two Numerical Variables

Line graph: `plot`




Scatter plot : `scatter`




(Demo)

# When to use a line vs scatter plot?

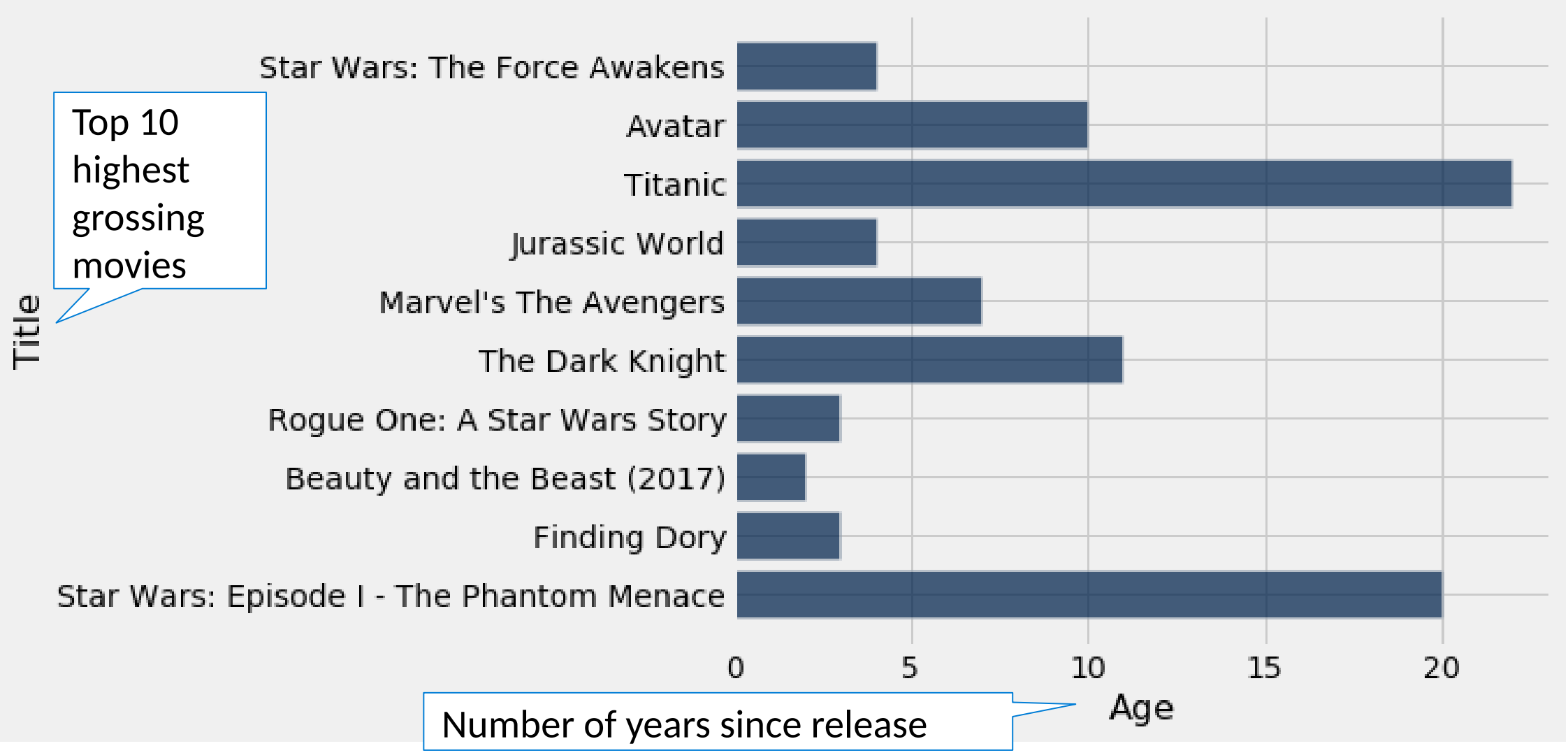
- Use line plots for sequential data: if…
  - …your x-axis has an order
  - …sequential differences in y values are meaningful
  - …there’s only one y-value for each x-value
  - Usually: x-axis is **time** or **distance**

- Use scatter plots for non-sequential data
  - When you’re looking for associations

# Categorical Data

(Demo)

# How Do You Generate This Chart?



# Terminology

- **Individuals**: those whose features are recorded
- **Variable**: a feature, an attribute
- A variable has different **values**
- Values can be **numerical** or **categorical**, and of many sub-types within these
- Each **individual has exactly one value** of the variable

- **Distribution**: For each different value of the variable, the frequency of individuals that have that value

# Categorical Distributions

# Visualization

- Bar charts are commonly used to visualize categorical distributions
- One axis is categorical, one numerical

(Demo)

# Displaying a Categorical Distribution

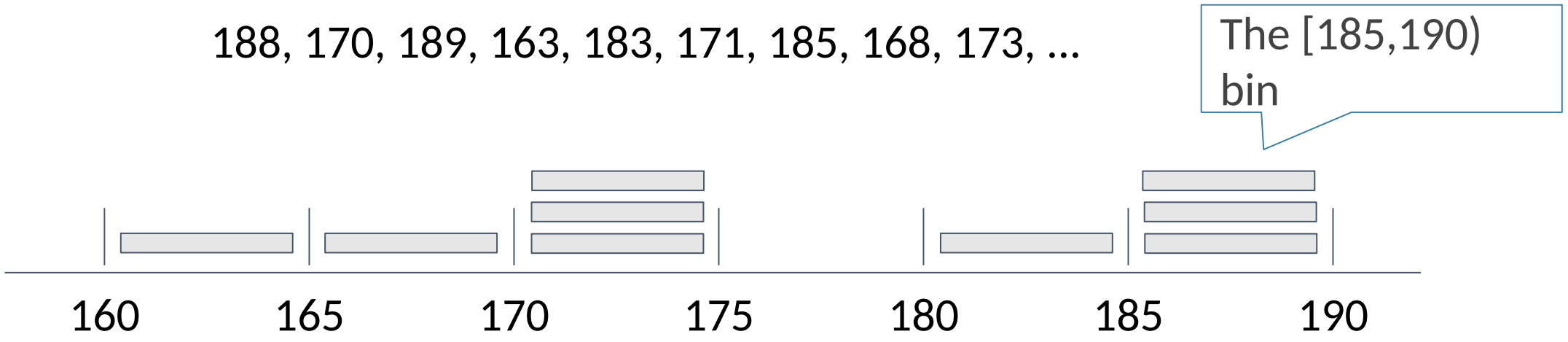
- The distribution of a variable (a column, e.g. Studios) describes the frequencies of its different values
- The `group` method counts the number of rows for each value in the column (e.g. the number of top movies released by each studio)
- Bar charts can display the distribution of a categorical variable (e.g. studios):
  - One bar for each category
  - Length of bar is the count of individuals in that category
  - You can choose the order of the bars

# Binning a Numerical Variable

# 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



# Area Principle

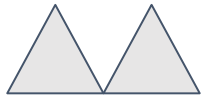
# Area Principle

**Areas** should be proportional to the values they represent.

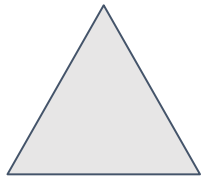
For example

- If you represent 20% of a population by

- Then 40% can be represented by:



- But not by:



# Drawing Histograms

# Histogram

- Chart that displays the distribution of a numerical variable
- Uses bins; there is one bar corresponding to each bin
- Uses the area principle:
  - The ***area*** of each bar is the percent of individuals in the corresponding bin

(Demo)

# Bar Chart or Histogram?

To display a distribution:

## Bar Chart

- Distribution of categorical variable
- Bars have arbitrary (but equal) widths and spacings
- height (or length) of bars proportional to the percent of individuals

## Histogram

- Distribution of numerical variable
- Horizontal axis is numerical: to scale, no gaps, bins can be unequal
- Area of bars proportional to the percent of individuals; height measures density

# Summary: charts

- **Scatter plot**: relation between numerical variables
- **Line graph**: sequential data (over time, etc.)
- **Bar chart**: distribution of categorical data
- **Histogram**: distribution of numerical data

# To do

- Assignment 1 due on Friday
- No Lecture on next Monday (due to CNY)
- No Lab next week (due to CNY)
- Assignment 2 will be uploaded before this weekend and due on Next Friday