DATA 8

Spring 2022

Lecture 7

Charts

Announcements

- Lab 1 regrades due tonight
- HW 2 is due Thursday, 2/3
  - Submit tonight for extra credit
- Lab 3 is due Friday at 5pm PT
- Check out #389 to sort out exam conflicts!

Weekly Goals

- Monday
  - Table review
  - Working with Census data
- **Today**
  - Visualizing data
  - Distributions
- Friday
  - Visualizing two kinds of distributions
  - Proportions as areas

Attribute Types

Types of Attributes

All values in a column of a table 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” Attributes

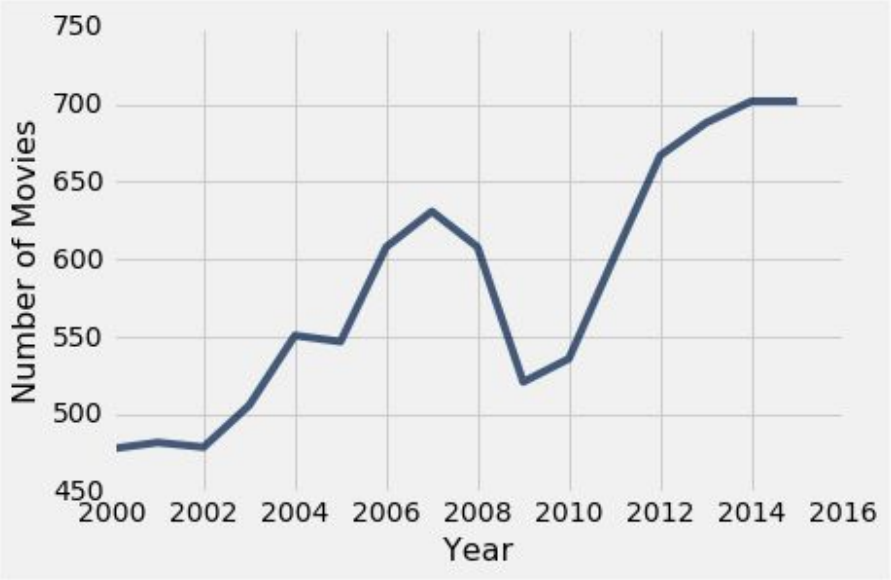
Just because the values are numbers, doesn’t mean the attribute is numerical

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

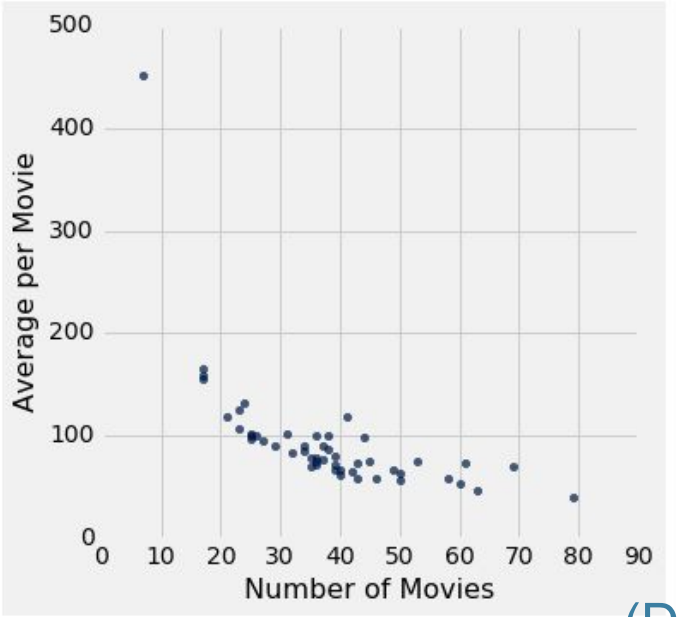
Numerical Data

Plotting Two Numerical Variables

Line plot: `plot`



Scatter plot : `scatter`



(Demo)

Anthony Daniels,
actor

https://en.wikipedia.org/wiki/C-3PO

Line vs Scatter Plot

- `t.plot(x_label, y_label)`
- `t.scatter(x_label, y_label)`

- Use line plots for sequential quantitative 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
  - Often: x-axis is **time** or **distance**
- Use scatter plots for non-sequential quantitative data
  - If you are looking for associations

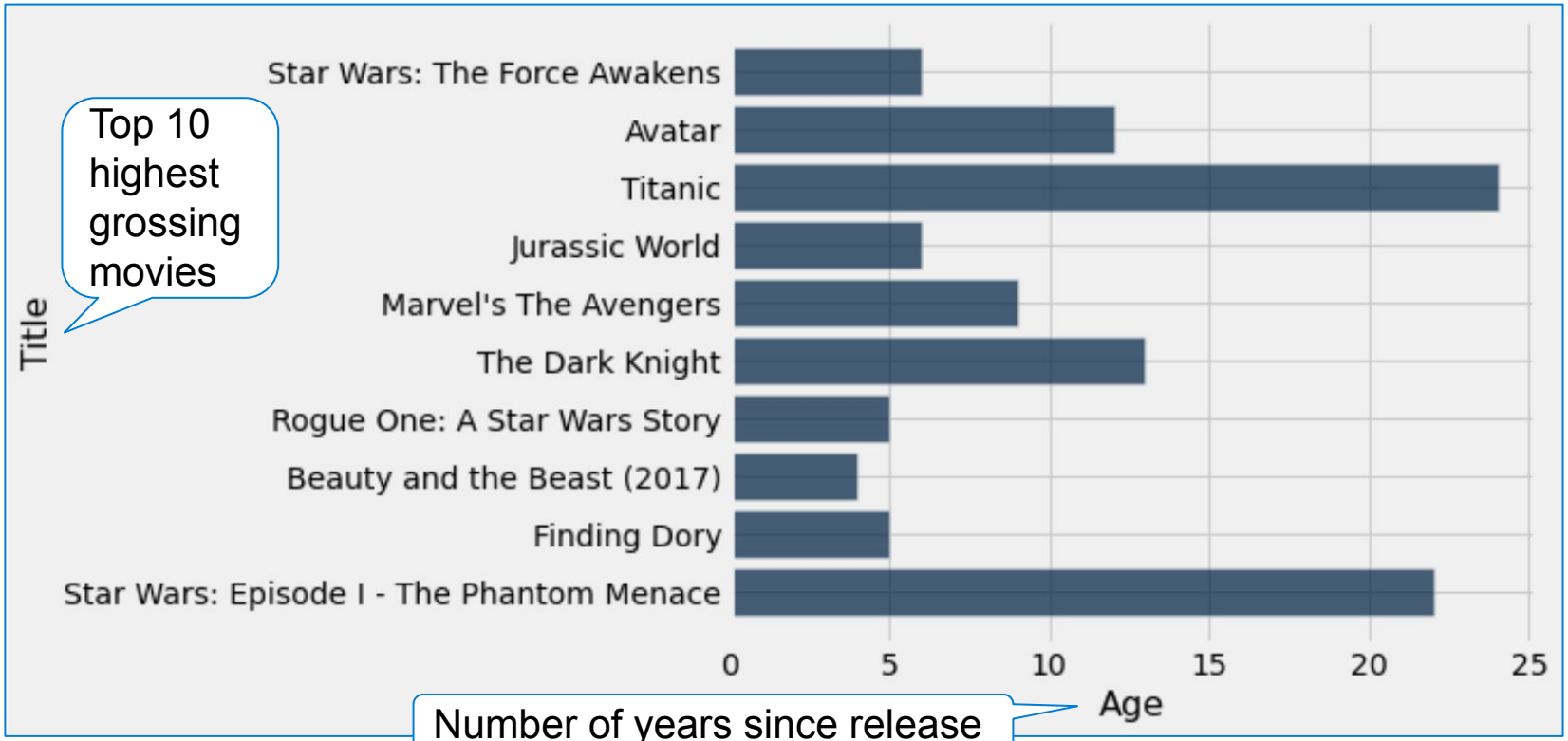
Categorical and Numerical Variables

Highest Grossing Movies as of 2017

Title	Studio	Gross	Gross (Adjusted)	Year
Gone with the Wind	MGM	198676459	1796176700	1939
Star Wars	Fox	460998007	1583483200	1977
The Sound of Music	Fox	158671368	1266072700	1965
E.T.: The Extra-Terrestrial	Universal	435110554	1261085000	1982
Titanic	Paramount	658672302	1204368000	1997

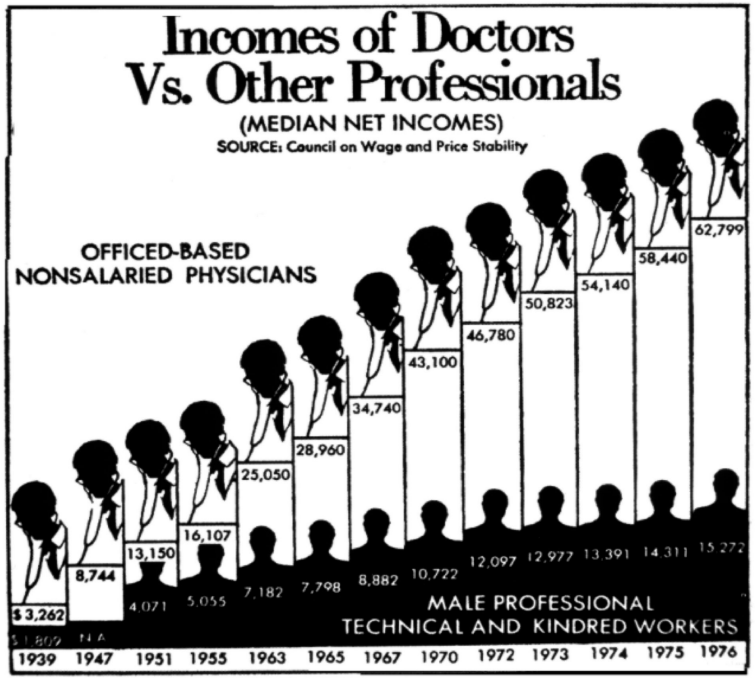
(Demo)

How Do You Generate This Chart?



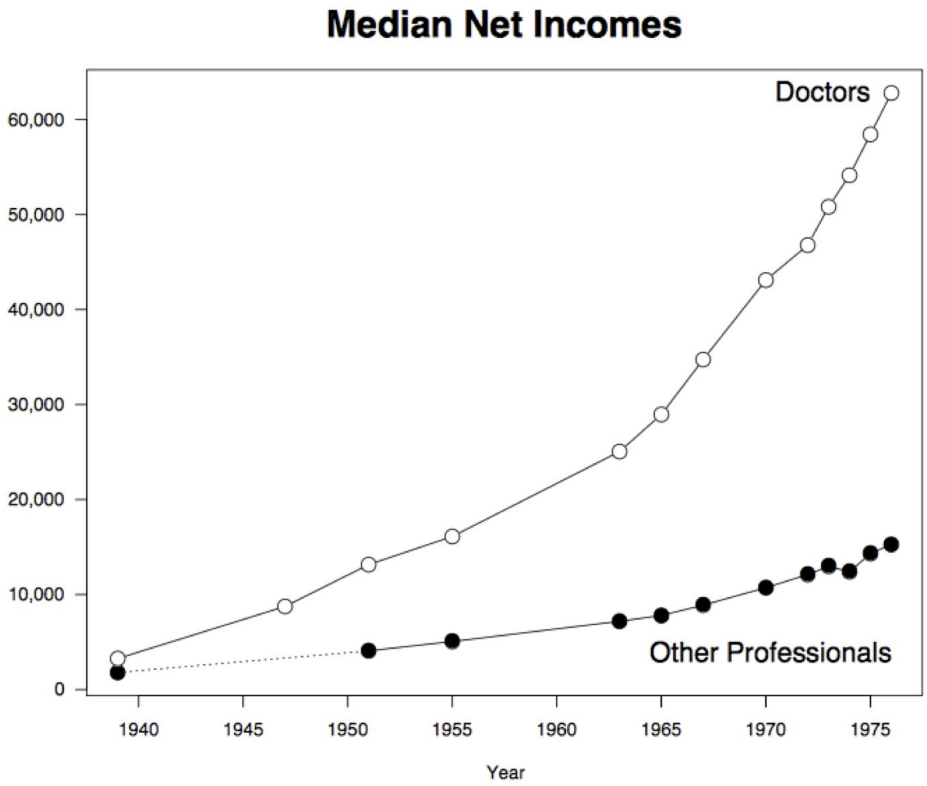
Visualization Fundamentals

Don’t Do This



Source: Washington Post

Do This Instead



Source: Ross Ihaka

Good Practices

- Less can be more
  - Minimize decoration
  - Choose colors carefully
    - Minimize the number of different colors
- If data are numerical, preserve their relative values and distances between them

See Edward Tufte's "The Visual Display of Quantitative Information"

Importance of the Y-Axis