YData: An Introduction to Data Science

Lecture 07: Charts

Jessi Cisewski-Kehe
Statistics & Data Science, Yale University
Spring 2020

Credit: data8.org

Announcements

Census Review

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)

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

(DEMO)

Growth Rate

- Growth rate = g (for example 3%, or 0.03)

- Initial value x, final value y after t periods of time

Value after 1 period	= x + xg	= x * (1+g)
Value after 2 periods	= x(1+g)(1+g)	= x * (1+g) ** 2
Value after t periods	= y	= x * (1+g) ** t

So (1+g) ** t = y/x and so 1+g = (y/x) ** (1/t)

So **g = (y/x) ** (1/t) - 1**

Data Visualization

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

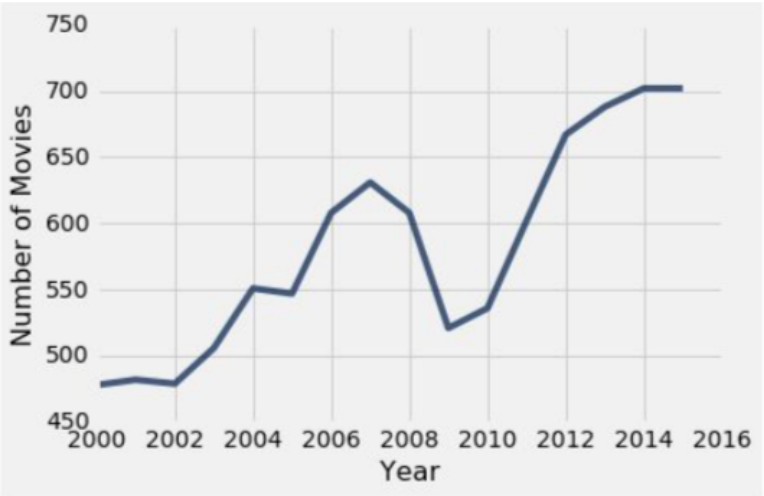
"Numerical" Data

Just because the values are numbers, doesn't mean the variable is numerical

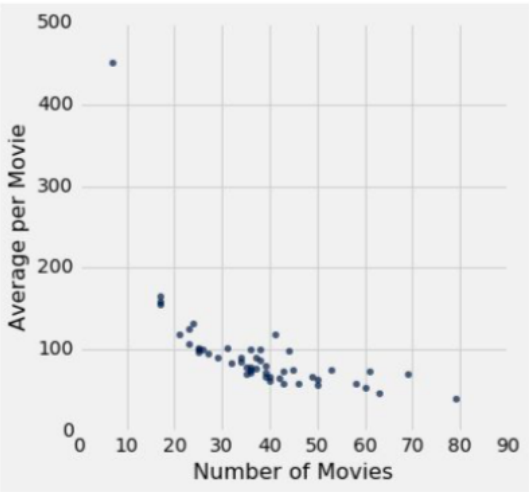
- Census example had numerical `SEX` code (0, 1, and 2)
- It doesn't make sense to perform arithmetic on these "numbers", e.g. 1 - 0 or (0+1+2)/3 are meaningless
- The variable `SEX` is still categorical, even though numbers were used for the categories

Plotting Two Numerical Variables

Line graph: `plot`



Scatter plot: `scatter`



(DEMO)