DATA 8

Spring 2020

Lecture 7

Charts

Announcements

- HW 2 due tomorrow night, 2/6
  - Turn in by 11:59PM **tonight** for a bonus point!
- New tutoring sections will open up this Friday ~11:30AM
- Midterm will be on **March 13, 7-9PM**

Weekly Goals

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

Table Review

Manipulating Rows

- `t.sort(column)` sorts the rows in increasing order
- `t.sort(column, descending=True)` sorts the rows in decreasing order
- `t.take(row_numbers)` keeps the numbered rows
  - Each row has an index, starting at 0
- `t.where(column, are.condition)` keeps all rows for which a column's value satisfies a condition
- `t.where(column, value)` keeps all rows for which a column's value equals some particular value
  - Same as `t.where(column, are.equal_to(value))`

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)

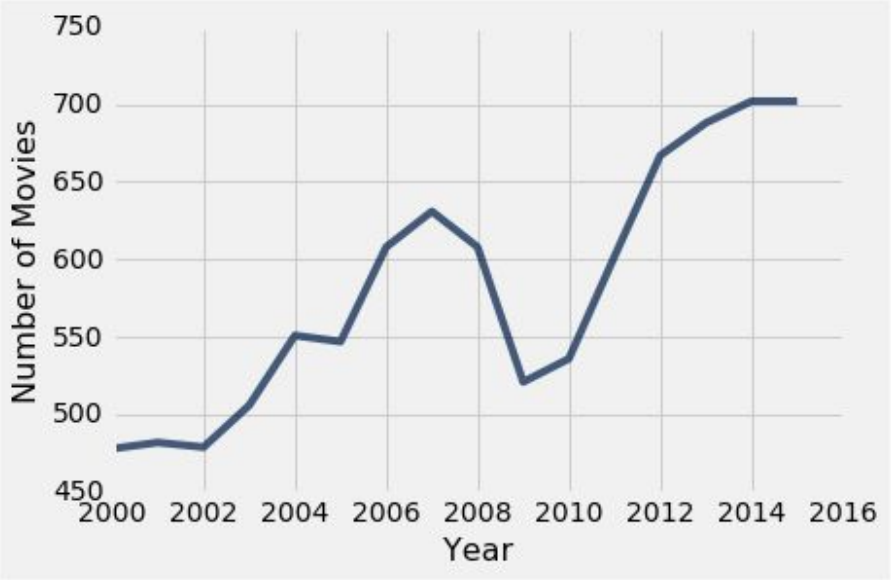
Numerical Data

Anthony Daniels,
actor

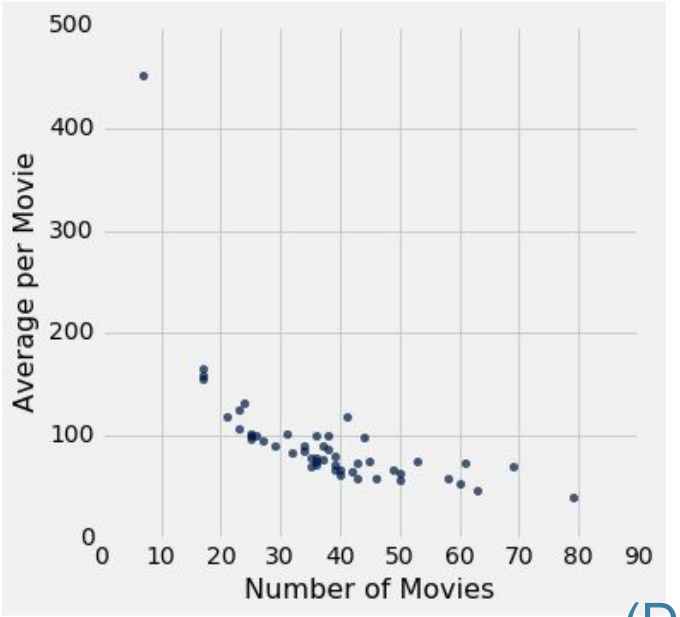
https://en.wikipedia.org/wiki/C-3PO

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?