YData: An Introduction to Data Science

Lecture 07: Charts

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University

Spring 2021

Credit: data8.org

Announcements

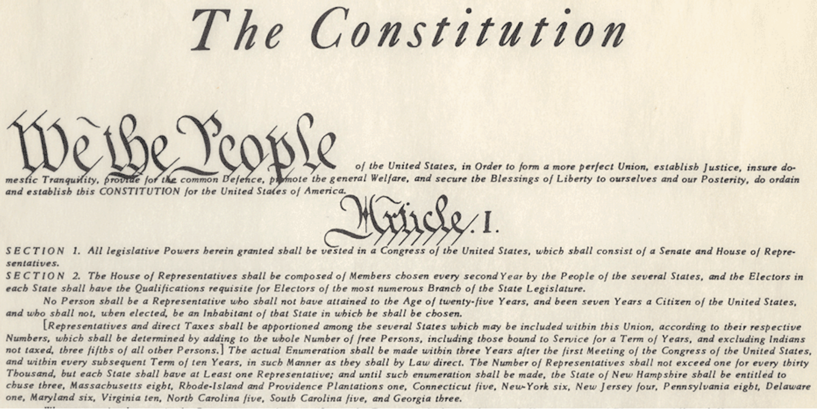
- Hw01 currently being graded
- Hw02 is due on Thursday
- Please complete the survey on Anaconda install
- Updated instructions on YCRC cluster coming

Census Review

The Decennial Census

- Every ten years, the Census Bureau counts how many people there are in the U.S.
- In between censuses (censi?), 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 ...”



The Constitution

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common Defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this CONSTITUTION for the United States of America.

Article. I.

SECTION 1. All legislative Powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.

SECTION 2. The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in each State shall have the Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

No Person shall be a Representative who shall not have attained to the Age of twenty-five Years, and been seven Years a Citizen of the United States, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

[Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons.] The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at Least one Representative; and until such enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode-Island and Providence Plantations one, Connecticut five, New-York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, South Carolina five, and Georgia three.

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

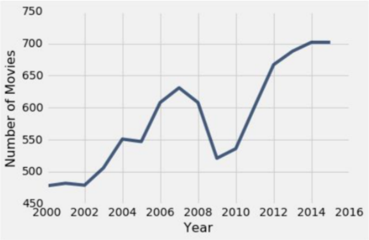
“Numerical” Data

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

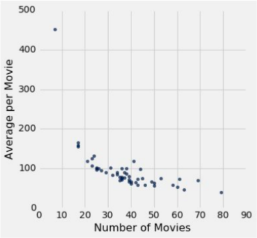
- Census example had numerical `SEX` code (0, 1, and 2)
- It doesn’t make sense to perform arithmetic on these “numbers”, e.g. 1 - 0 or (0+1+2)/3 are meaningless
- The variable `SEX` is still categorical, even though numbers were used for the categories

Plotting Two Numerical Variables

Line graph: `plot`



Scatter plot: `scatter`



(DEMO)