Charts & Histograms

Prof. Abby Stylianou

Logistics

- Homework 2 out on Tuesday, due the following Tuesday

- Quiz today -- on paper!

Types of Data

- What is the difference between *numerical* and *categorical* data?

Types of Data

- What is the difference between *numerical* and *categorical* data?
  - **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

Types of Data

- Are ZIP codes categorical or numerical?

Types of Data

- Are ZIP codes categorical or numerical?

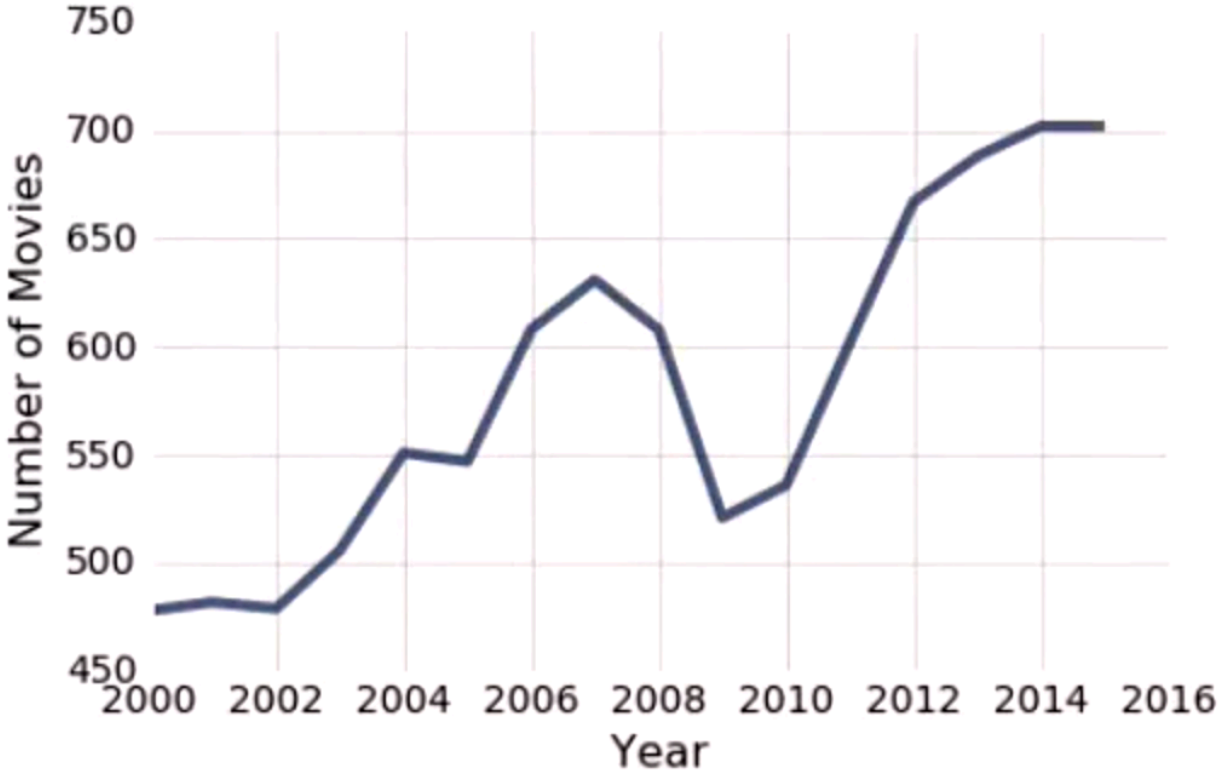
- What about ‘unsatisfied’, ‘satisfied’, ‘very satisfied’?

Types of Data

- Are ZIP codes categorical or numerical?
- What about ‘unsatisfied’, ‘satisfied’, ‘very satisfied’?
- What about class grades?

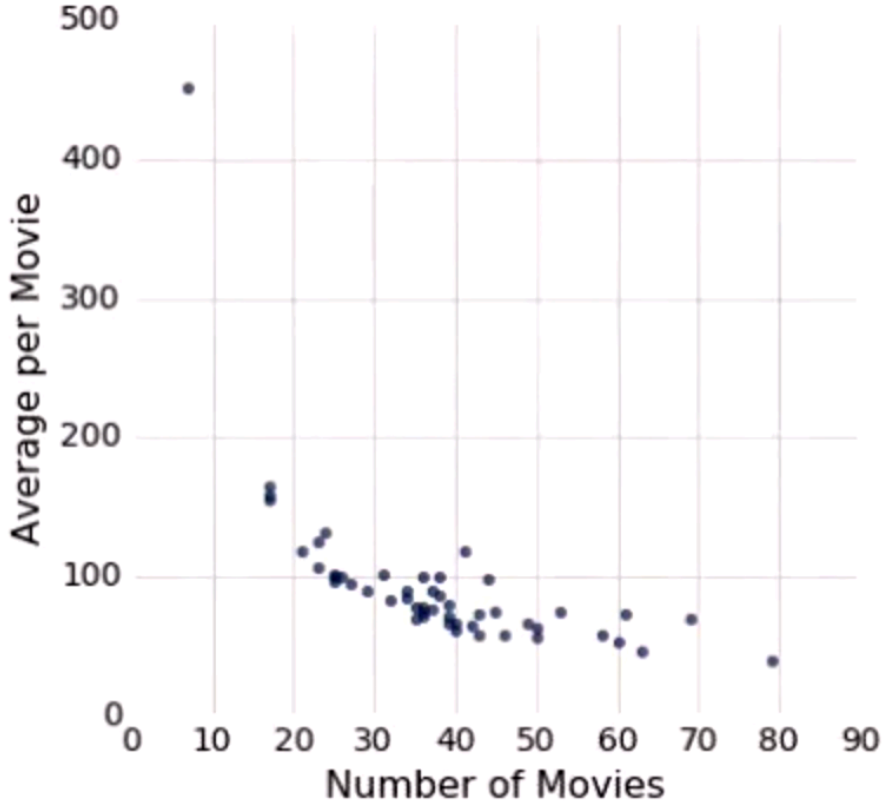
Plotting Two Numerical Variables

Line graph: `plot`



How something changes as the X-axis changes (often chronologically)

Scatter plot: `scatter`



Comparing two numerical variables

Terminology

- **Individuals**: those whose features are recorded

Terminology

- **Individuals**: those whose features are recorded
- **Variable**: a feature or attribute
  - Variables have different values
  - Values can be categorical or numerical (and many sub-types within these)

Terminology

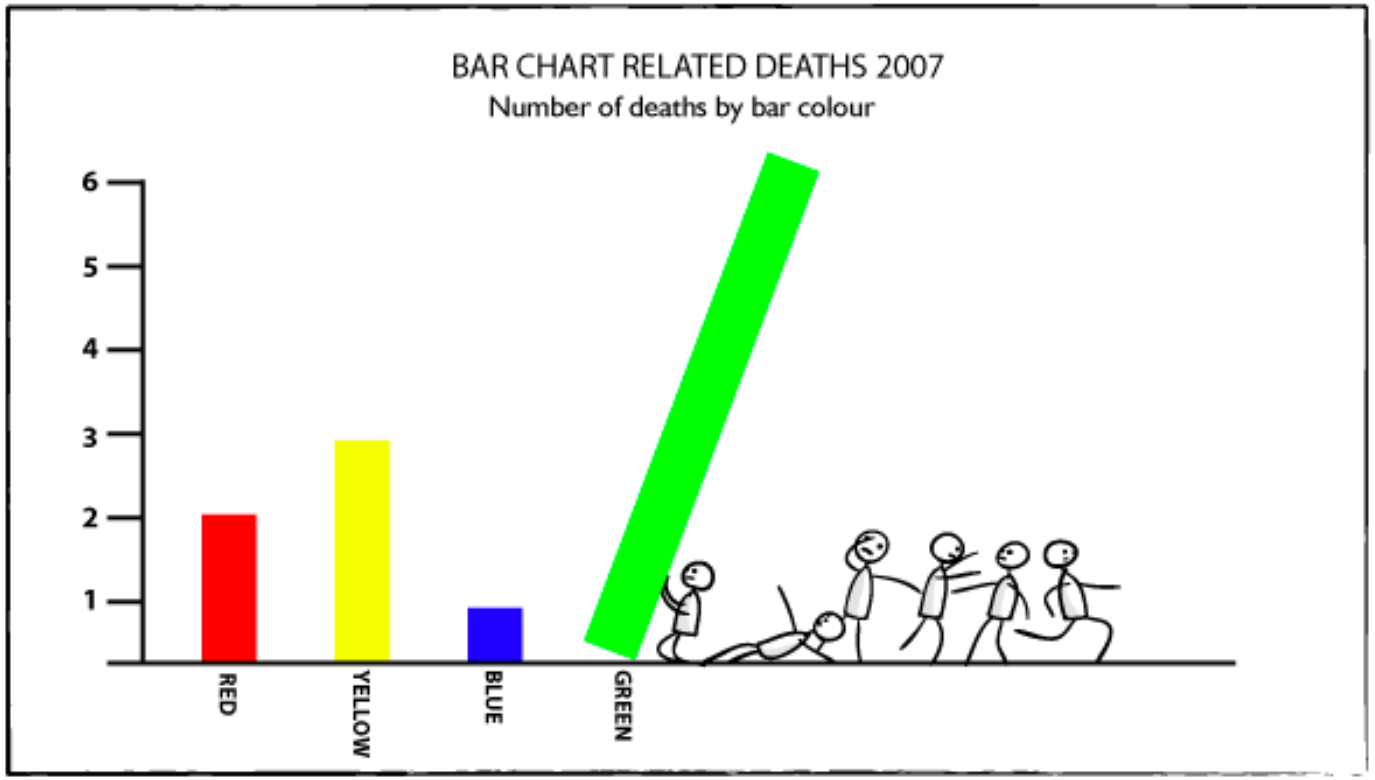
- **Individuals**: those whose features are recorded
- **Variable**: a feature or attribute
  - Variables have different values
  - Values can be categorical or numerical (and many sub-types within these)
- Each individual has ***one*** value of the variable

Terminology

- **Individuals**: those whose features are recorded
- **Variable**: a feature or attribute
  - Variables have different values
  - Values can be categorical or numerical (and many sub-types within these)
- Each individual has ***one*** value of the variable
- **Distribution**: for each different value of the variable, what is the frequency of individuals that have that value

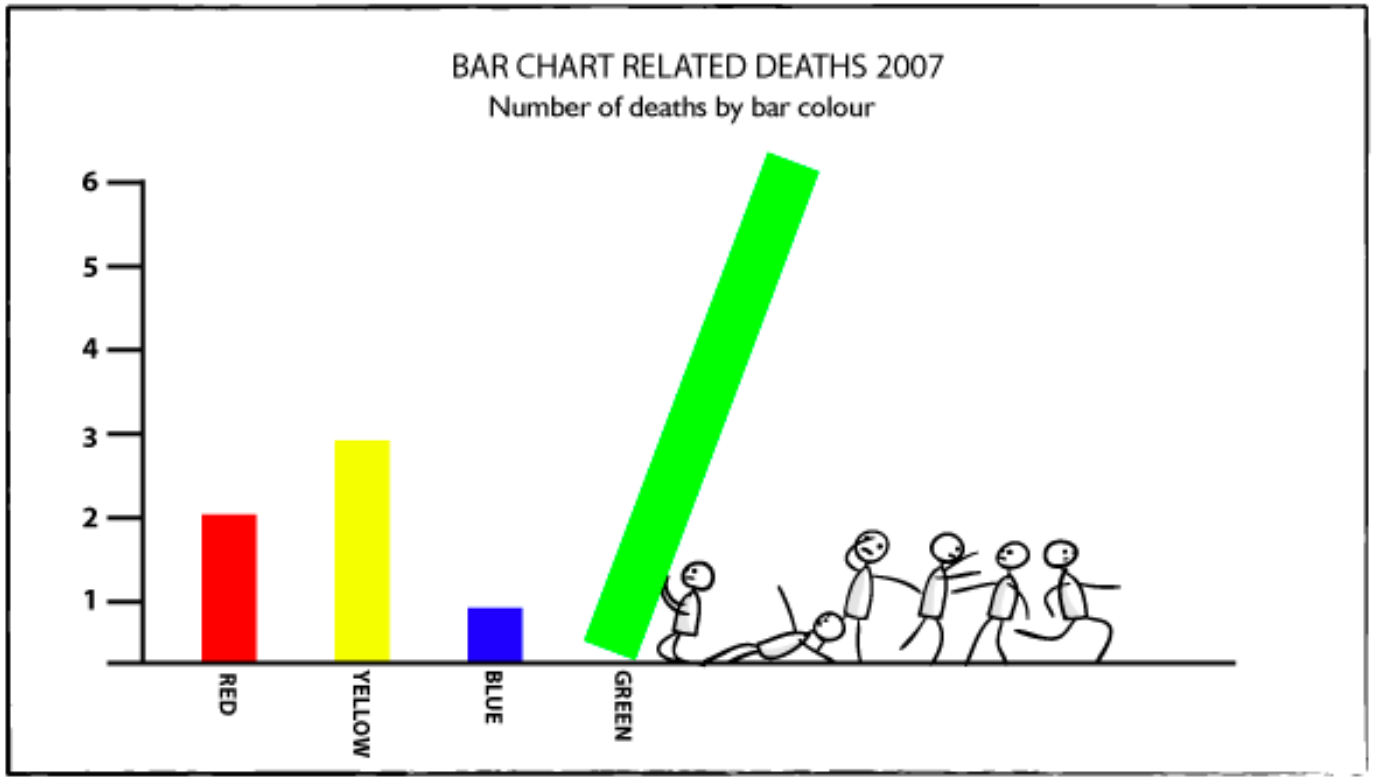
Categorical Visualization

- Bar charts!
  - One axis is categorical, one is numerical



Categorical Visualization

- Bar charts!
  - One axis is categorical, one is numerical



Demo → cs1070.com → cs1070_materials → demos

Numerical Visualization

- For categorical data, visualization of distribution is easy → plot # of individuals in a category

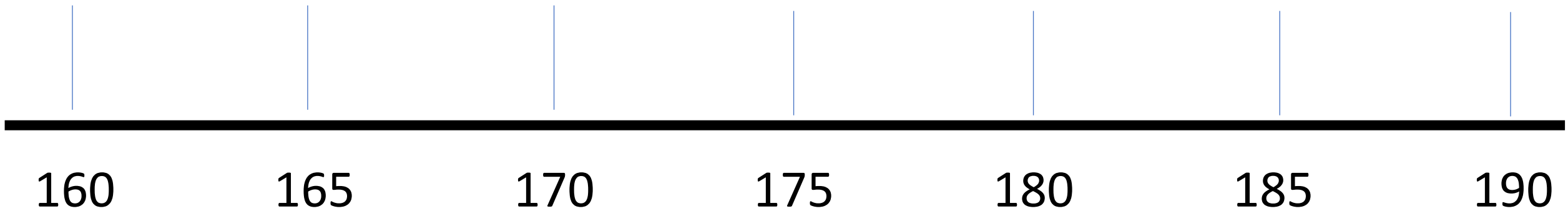
- What about for numerical data?
  - E.g., height (person A is 68.3” tall, person B is 68.4” tall, person C is 61” tall, person D is 61.5” tall, etc.)

Binning Numerical Values

- Count the number of numerical values that lie within a range or bin
  - Typical convention: Bins are defined by their lower bounds (inclusive)
  - The upper bound is the lower bound of the next bin

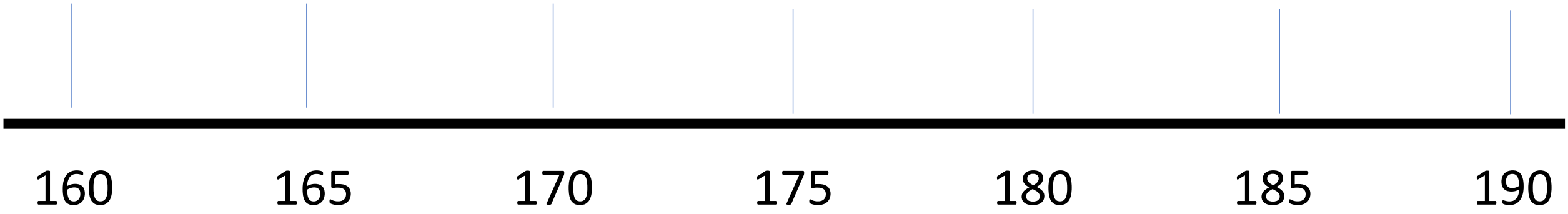
Binning Numerical Values

188, 170, 189, 163, 183, 171, 185, 168, 173, …



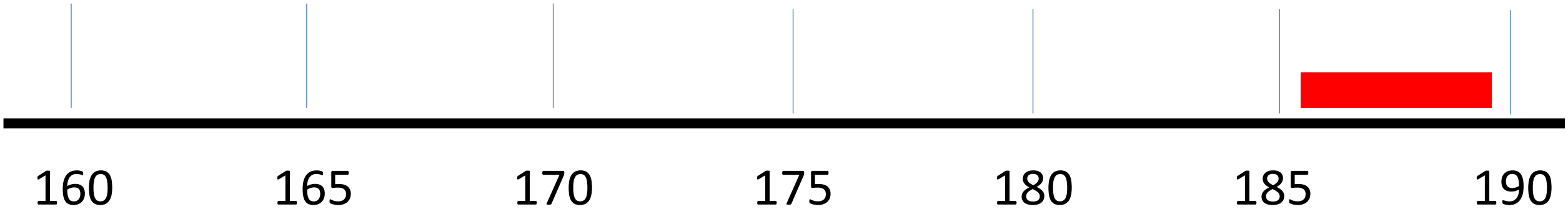
Binning Numerical Values

188, 170, 189, 163, 183, 171, 185, 168, 173, …



Binning Numerical Values

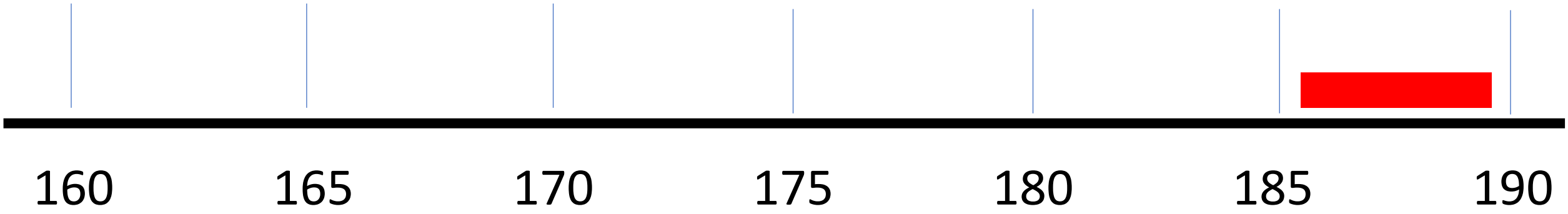
188, 170, 189, 163, 183, 171, 185, 168, 173, …



Goes into the
[185, 190) bin

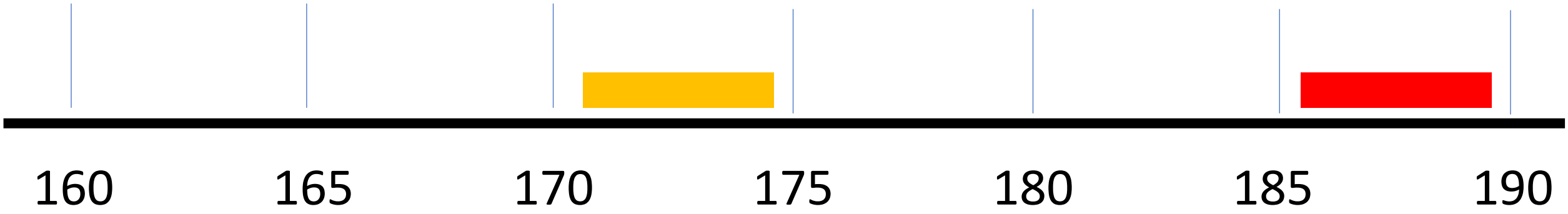
Binning Numerical Values

188, 170, 189, 163, 183, 171, 185, 168, 173, …



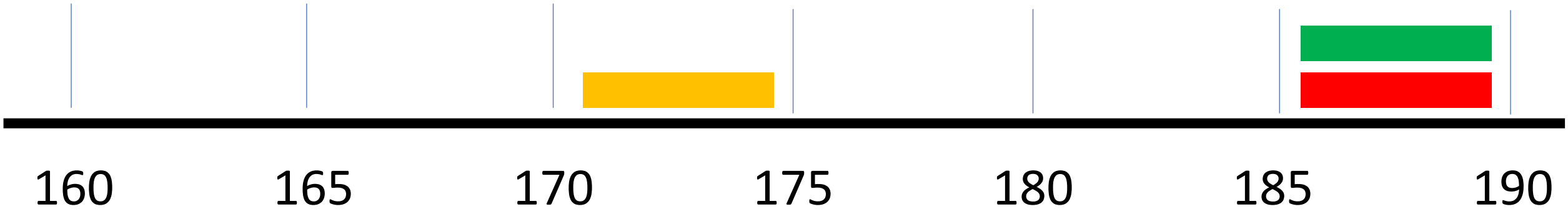
Binning Numerical Values

188, 170, 189, 163, 183, 171, 185, 168, 173, …



Binning Numerical Values

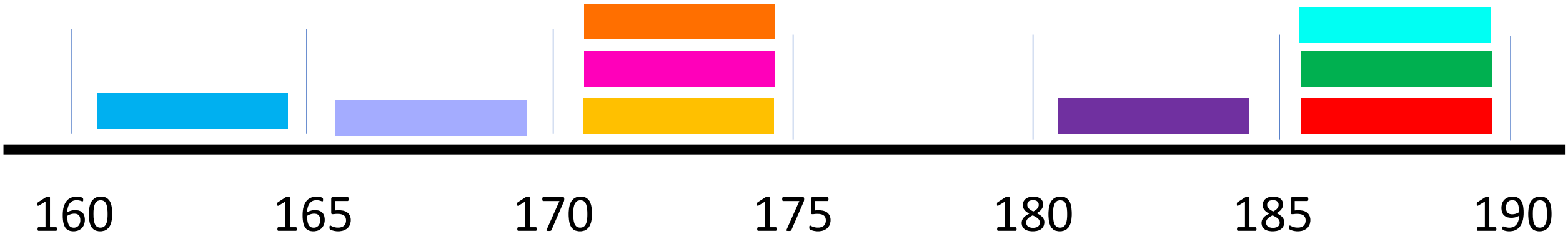
188, 170, 189, 163, 183, 171, 185, 168, 173, …



Finish with you neighbors!

Binning Numerical Values

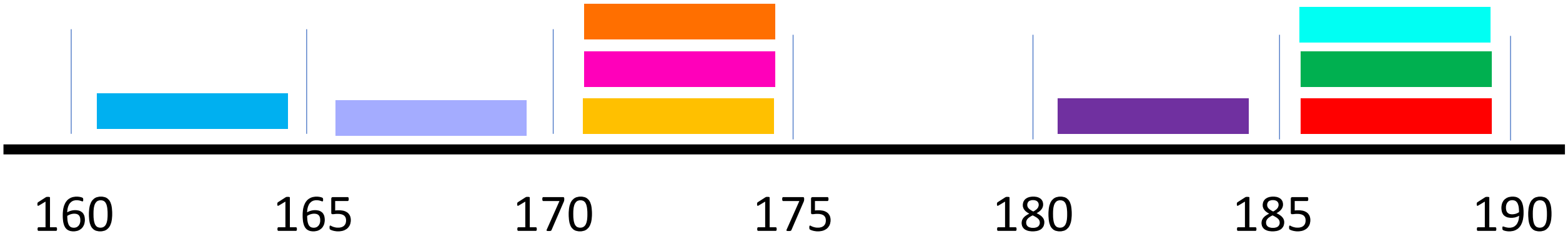
188, 170, 189, 163, 183, 171, 185, 168, 173, ...



Finish with you neighbors!

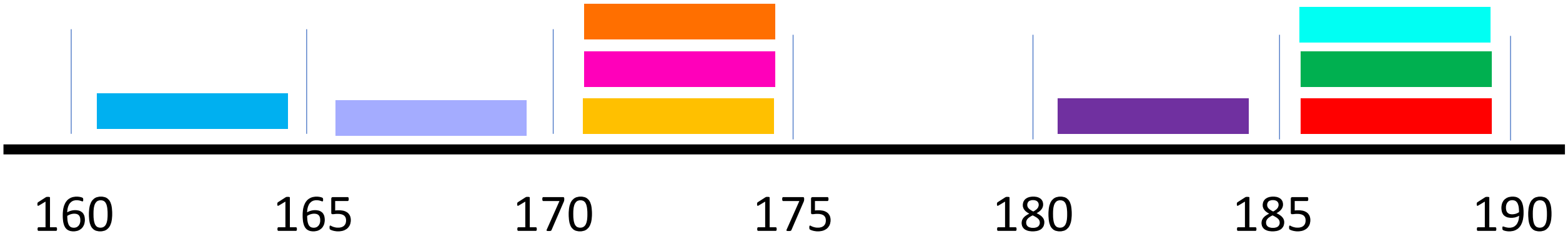
Binning Numerical Values

This looks a lot like a bar chart!



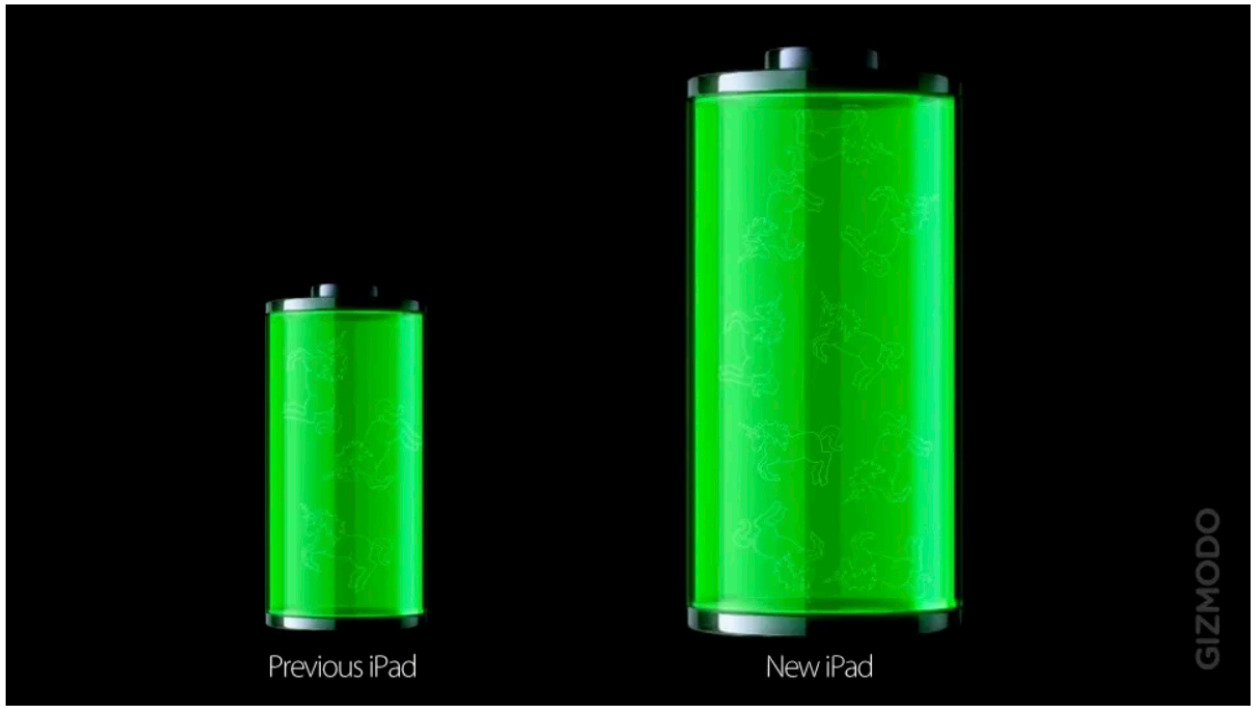
Binning Numerical Values

This looks a lot like a bar chart!



Back to jupyter notebook → let's get python to do the binning automatically!

What is wrong with this picture?



From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

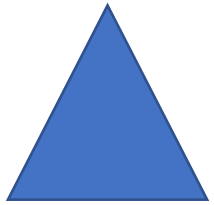
https://flowingdata.com/2012/03/16/new-ipad-battery-size-is-huge/

Area Principle

- Areas should be proportional to the values they represent

20% of the population

Which of these can be 40%?

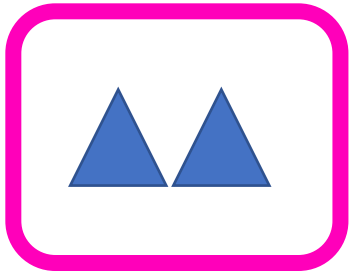


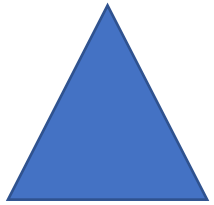
Area Principle

- ***Areas*** should be proportional to the values they represent (not length and width)

20% of the population

Which of these can be 40%?





Histograms

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

Histograms

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

Back to jupyter notebook

Histogram Axes

- By default, hist uses a scale (normed=True) that ensures the area of the chart sums to 100%
- The area of each bar is a percentage of the whole
- The horizontal axis is a number line (e.g., years0, and the bin sizes don’t have to be equal to each other
- Vertical axis is numerical

Next class → density! (and more)