CS1070: Taming Big Data

Intro to Data Visualization

Logistics

- Coronavirus quiz today
- Homework 2 due Thursday

Types of Data

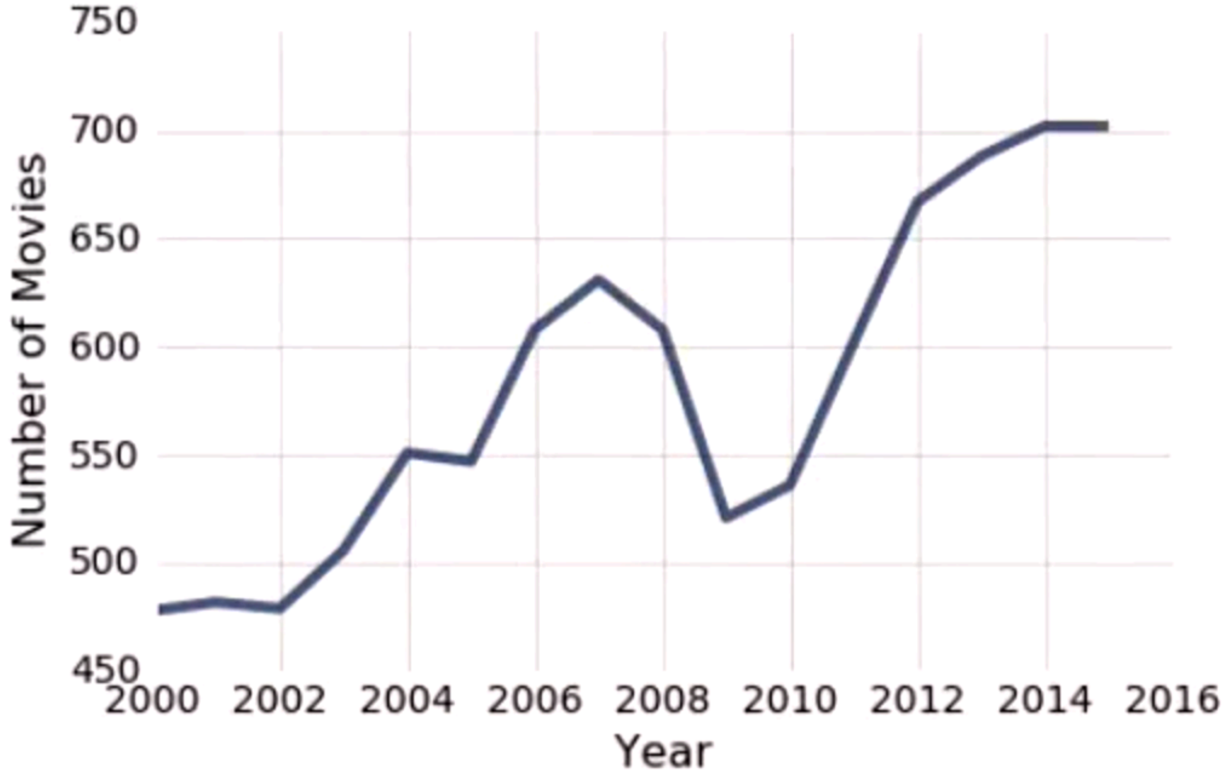
- What is the difference between *numerical* and *categorical* data?

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

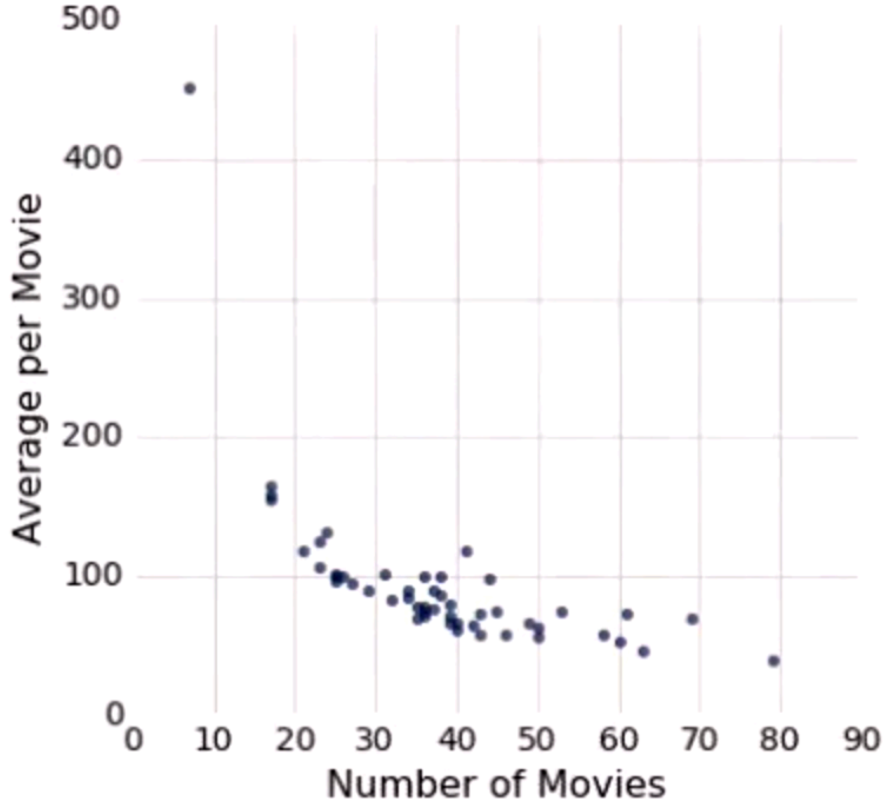
Plotting Two Numerical Variables

Line graph: `plot`

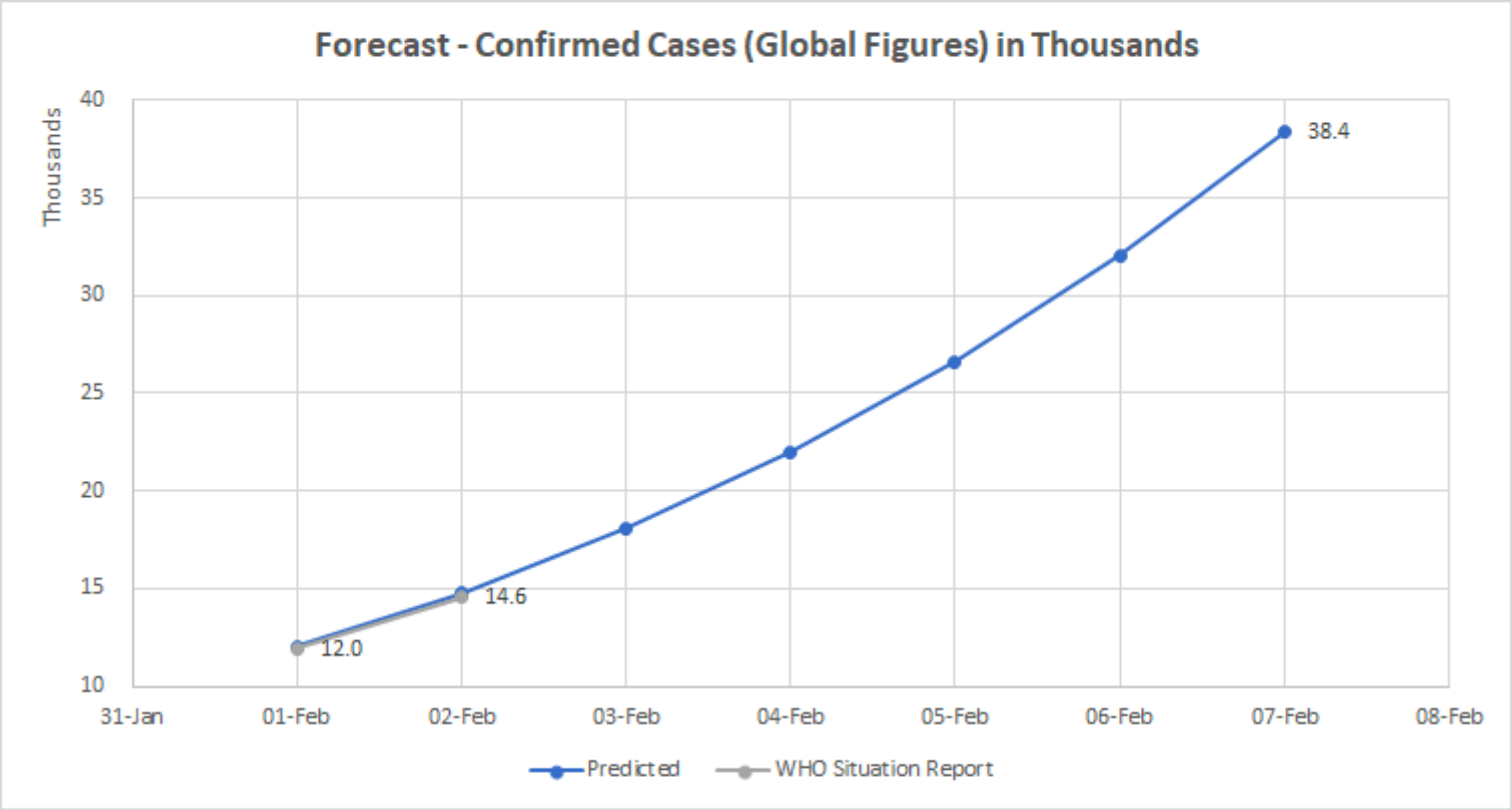


How something changes as the X-axis changes (often chronologically)

Scatter plot: `scatter`

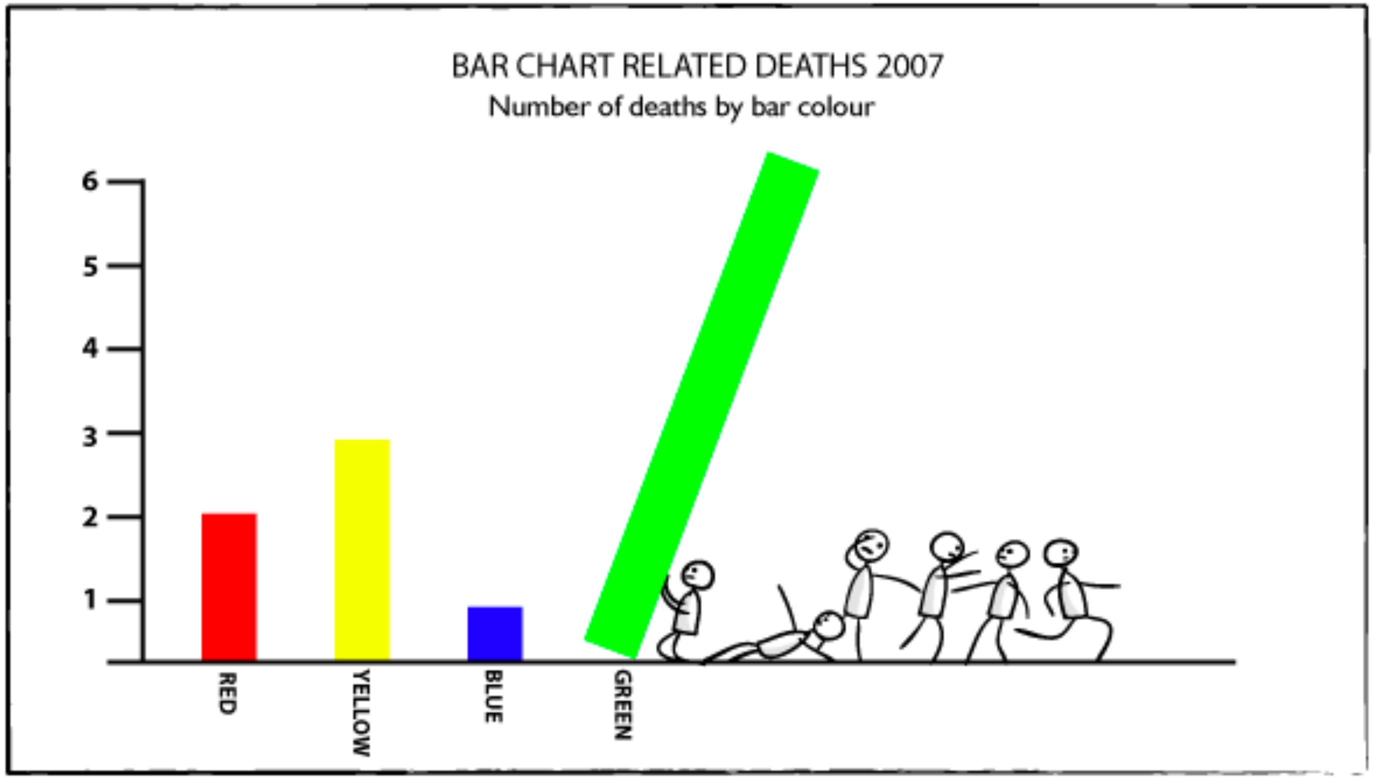


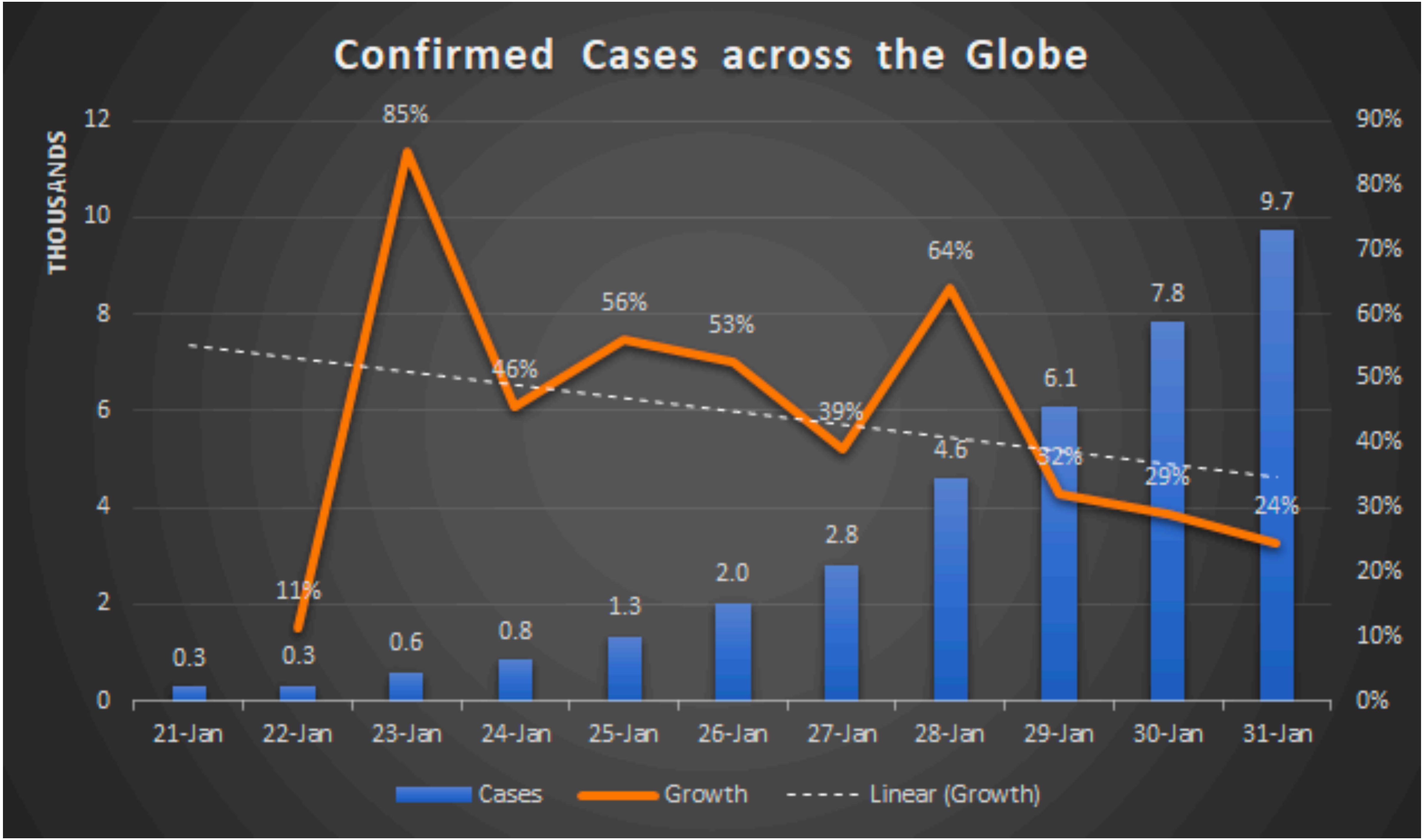
Comparing two numerical variables


Forecast - Confirmed Cases (Global Figures) in Thousands
Thousands
40
35
30
25
20
15
10
31-Jan
01-Feb
02-Feb
03-Feb
04-Feb
05-Feb
06-Feb
07-Feb
08-Feb
12.0
14.6
38.4
Predicted
WHO Situation Report

Categorical Visualization

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





Confirmed Cases across the Globe

Numerical Visualization

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

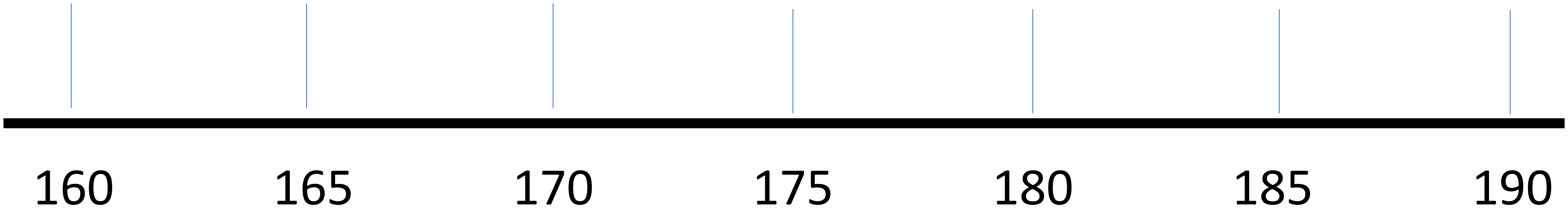
- What about for numerical data?
  - E.g., height (person A is 68.3” tall, person B is 68.4” tall, person C is 61” tall, person D is 61.5” tall, etc.)

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

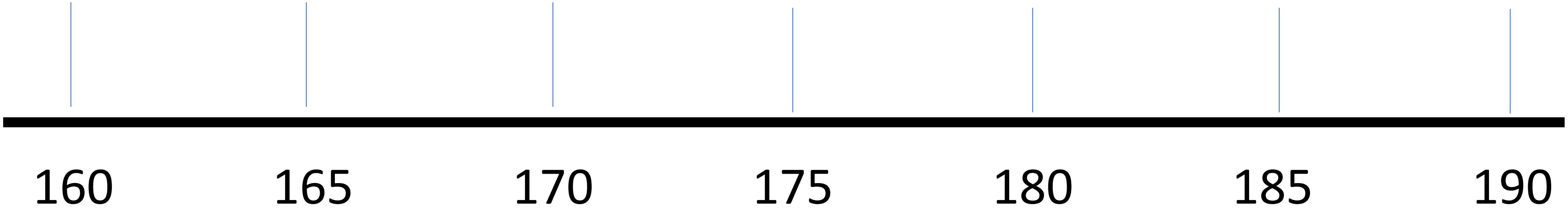
Binning Numerical Values

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



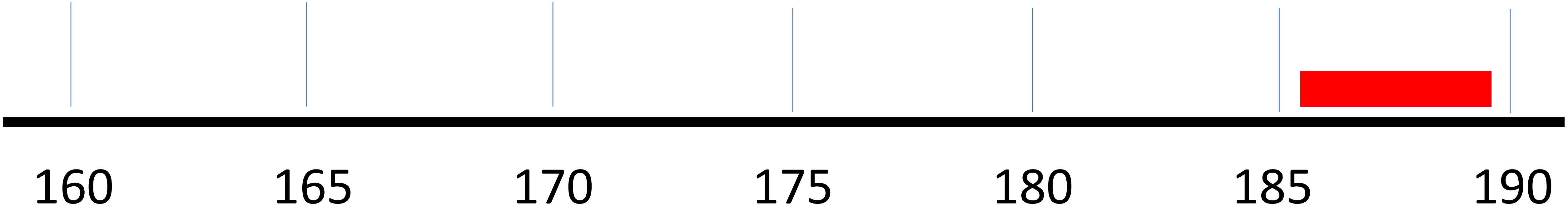
Binning Numerical Values

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



Binning Numerical Values

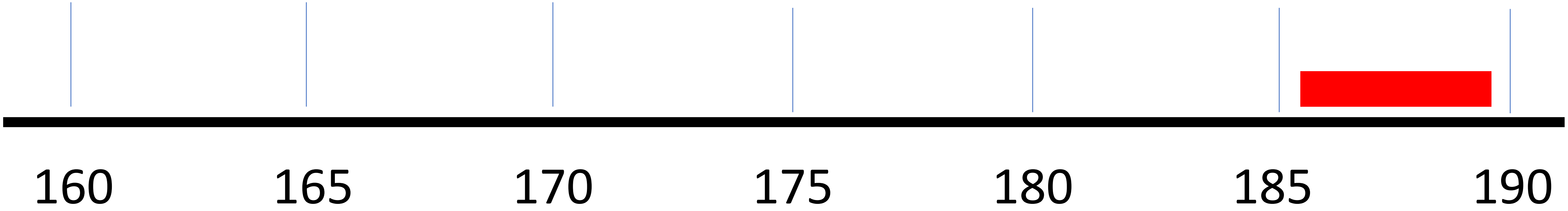
188, 170, 189, 163, 183, 171, 185, 168, 173, …



Goes into the
[185, 190) bin

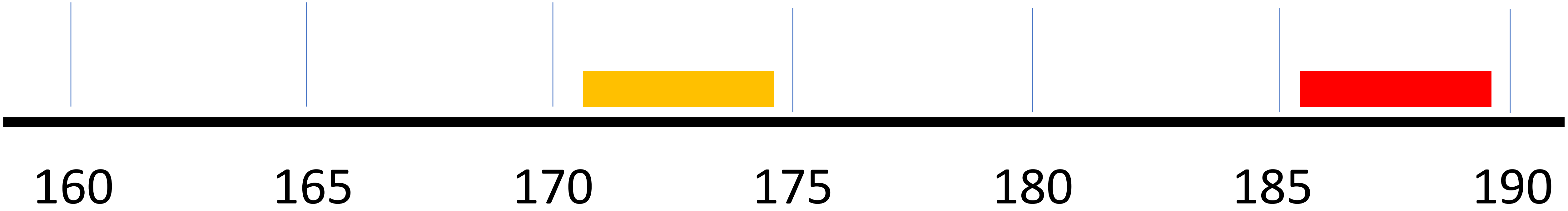
Binning Numerical Values

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



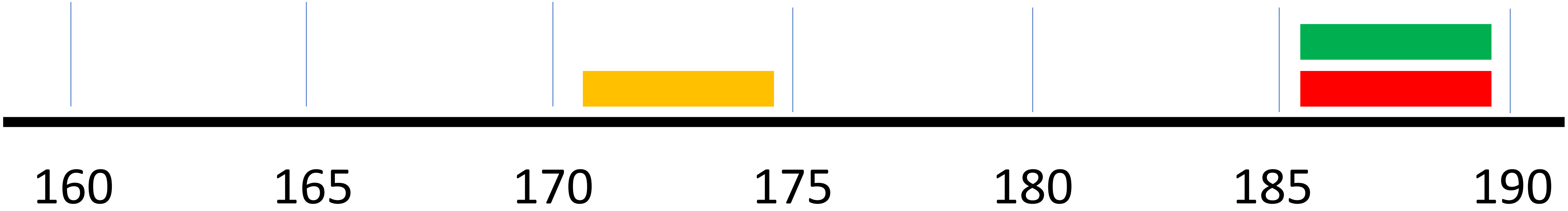
Binning Numerical Values

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



Binning Numerical Values

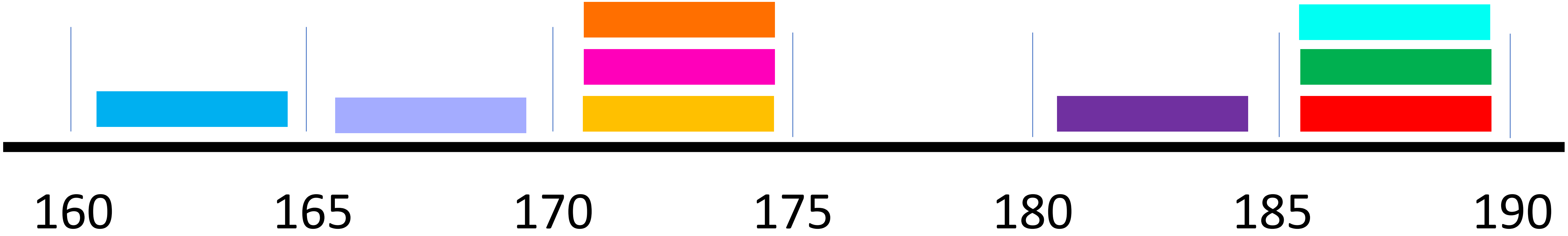
188, 170, 189, 163, 183, 171, 185, 168, 173, …



Finish with you neighbors!

Binning Numerical Values

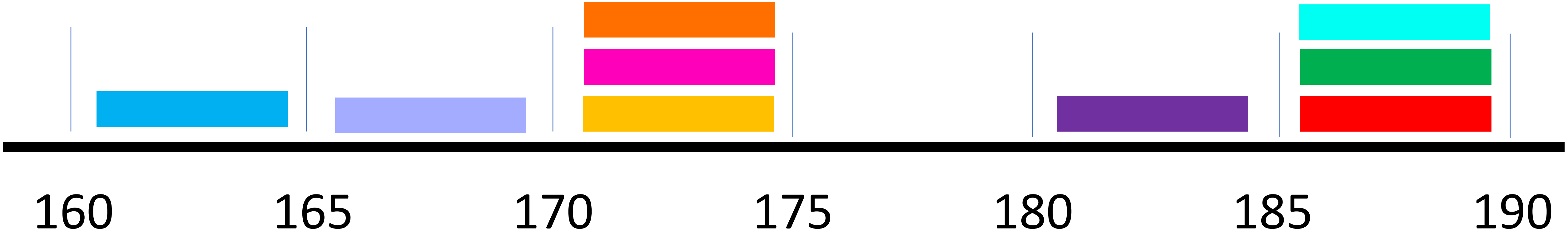
188, 170, 189, 163, 183, 171, 185, 168, 173, ...



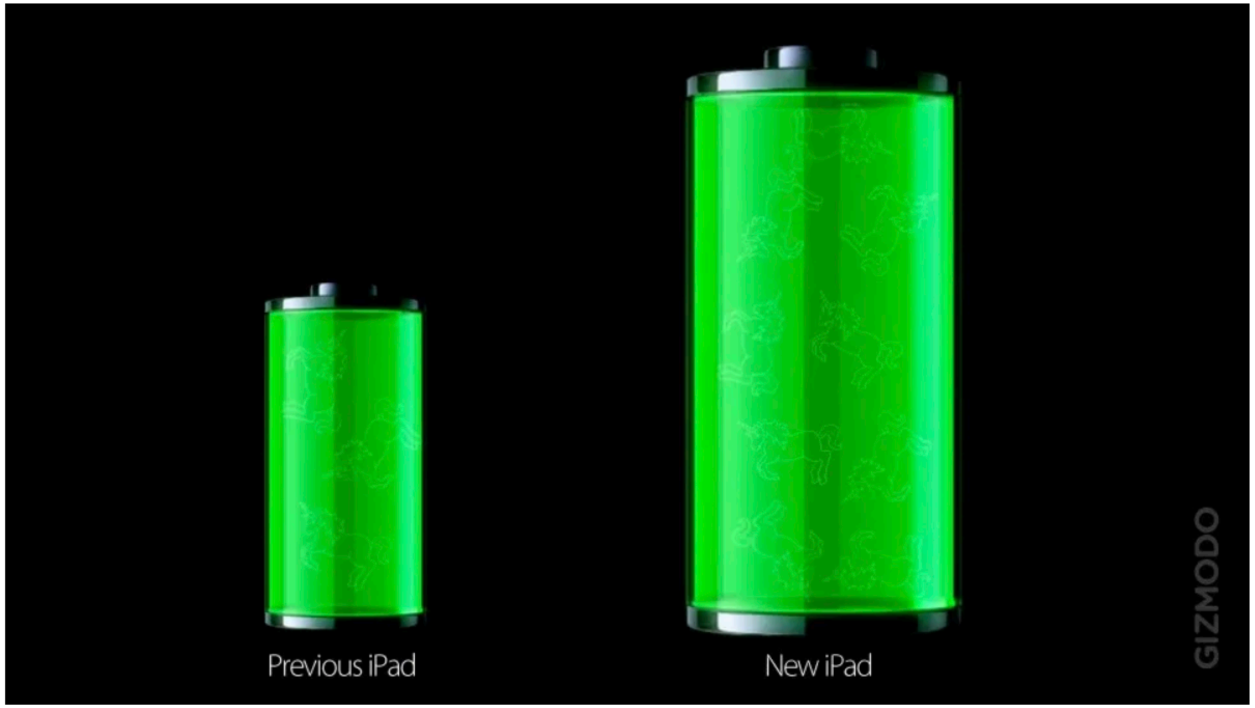
Finish with you neighbors!

Binning Numerical Values

This looks a lot like a bar chart!



What is wrong with this picture?



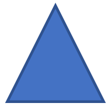
From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

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

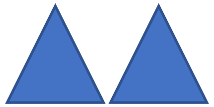
Area Principle

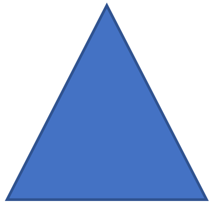
- Areas should be proportional to the values they represent

20% of the population



Which of these can be 40%?

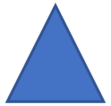




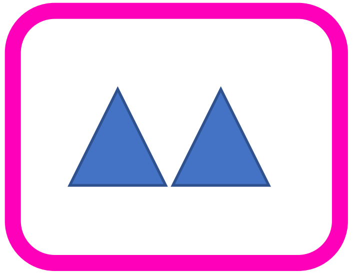
Area Principle

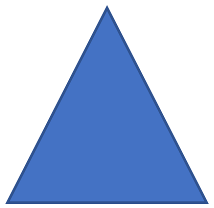
- *Areas* should be proportional to the values they represent (not length and width)

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