SOC245: Visualizing Data
Lecture 4: Graphical Excellence and Association

Chris Felton and Vikram Ramaswamy

Freshman Scholars Institute
Princeton University

July 18, 2022

Outline

Where we are and where we're going

- Understanding data with one variable: central tendency and spread

Where we are and where we're going

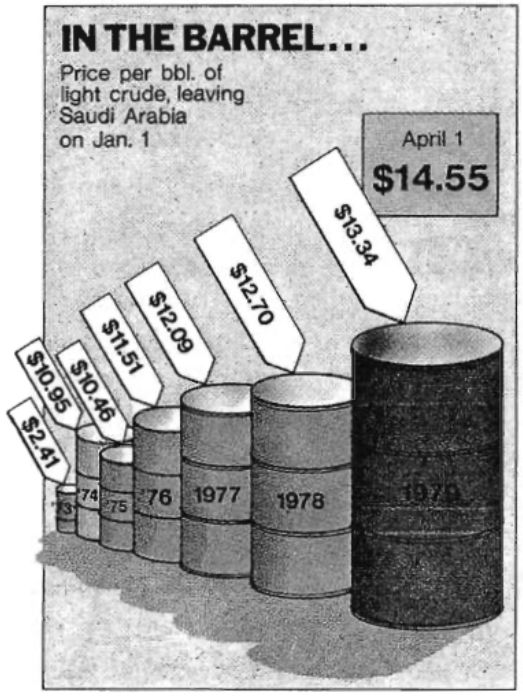
- Understanding data with one variable: central tendency and spread
- Looking at frequency distributions to better understand these variables.

Where we are and where we're going

- Understanding data with one variable: central tendency and spread
- Looking at frequency distributions to better understand these variables.
- Next: How can we use visualizations to best convey our takeaways?

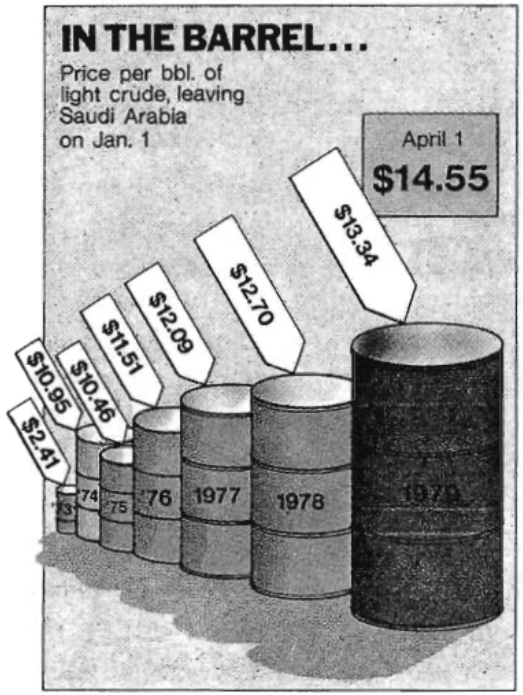
Outline

What's wrong with this graph?



Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 62

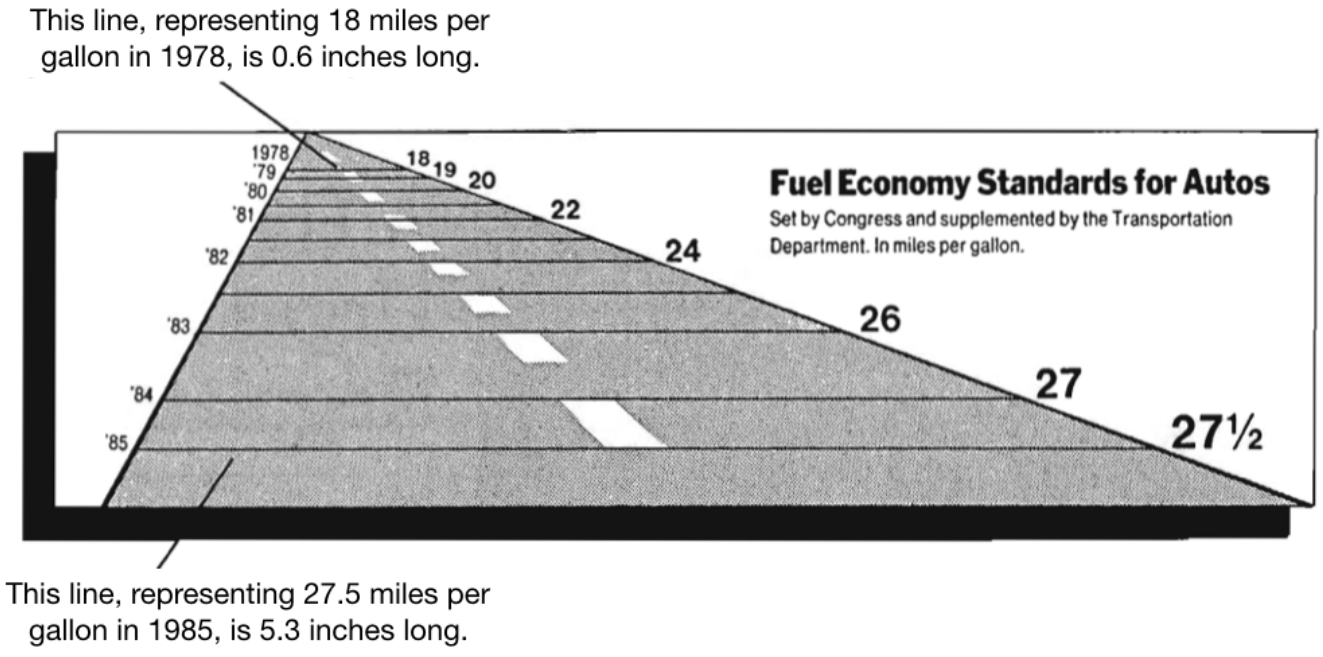
What's wrong with this graph?



Impossible to quickly understand how much the price has actually changed without looking at the numbers.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 62

What's wrong with this graph?



Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 57

The Lie Factor

- **Lie factor**: intuitively, this is how much graphs distort measurements.

The Lie Factor

- **Lie factor**: intuitively, this is how much graphs distort measurements.
- Formally,

$$\text{Lie factor} = \frac{\text{size of effect in the graph}}{\text{size of effect in data}}$$

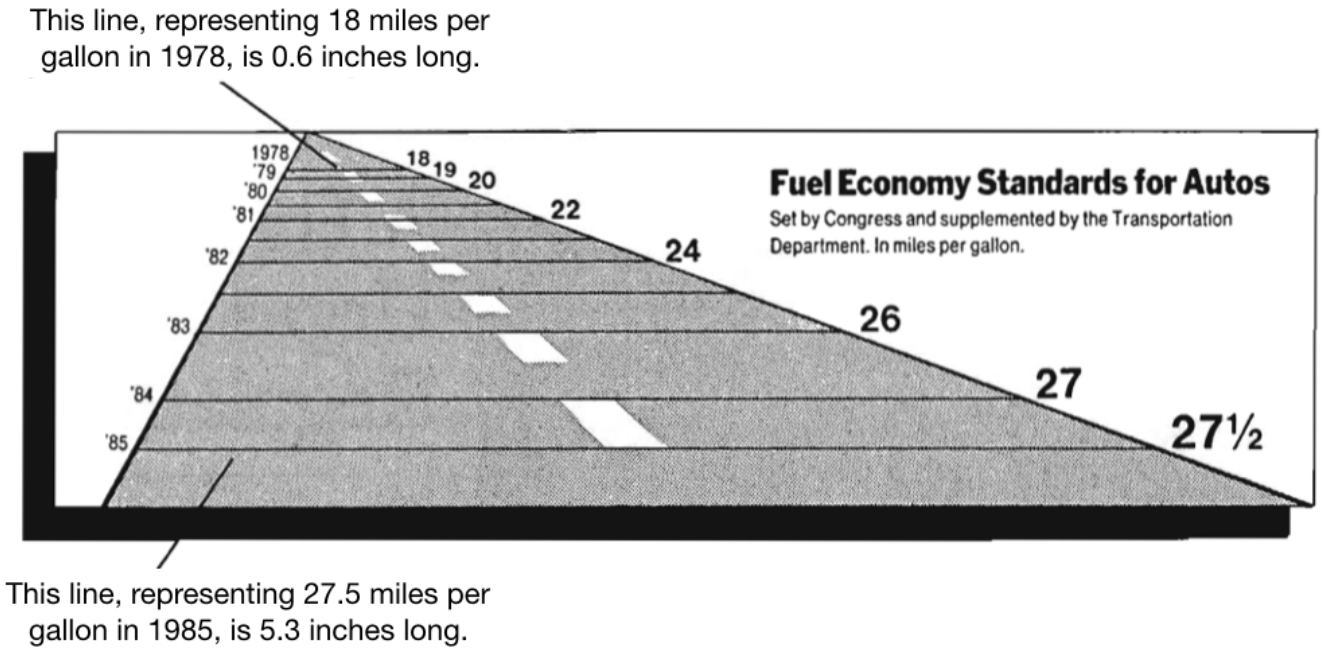
The Lie Factor

- **Lie factor**: intuitively, this is how much graphs distort measurements.
- Formally,

$$\text{Lie factor} = \frac{\text{size of effect in the graph}}{\text{size of effect in data}}$$

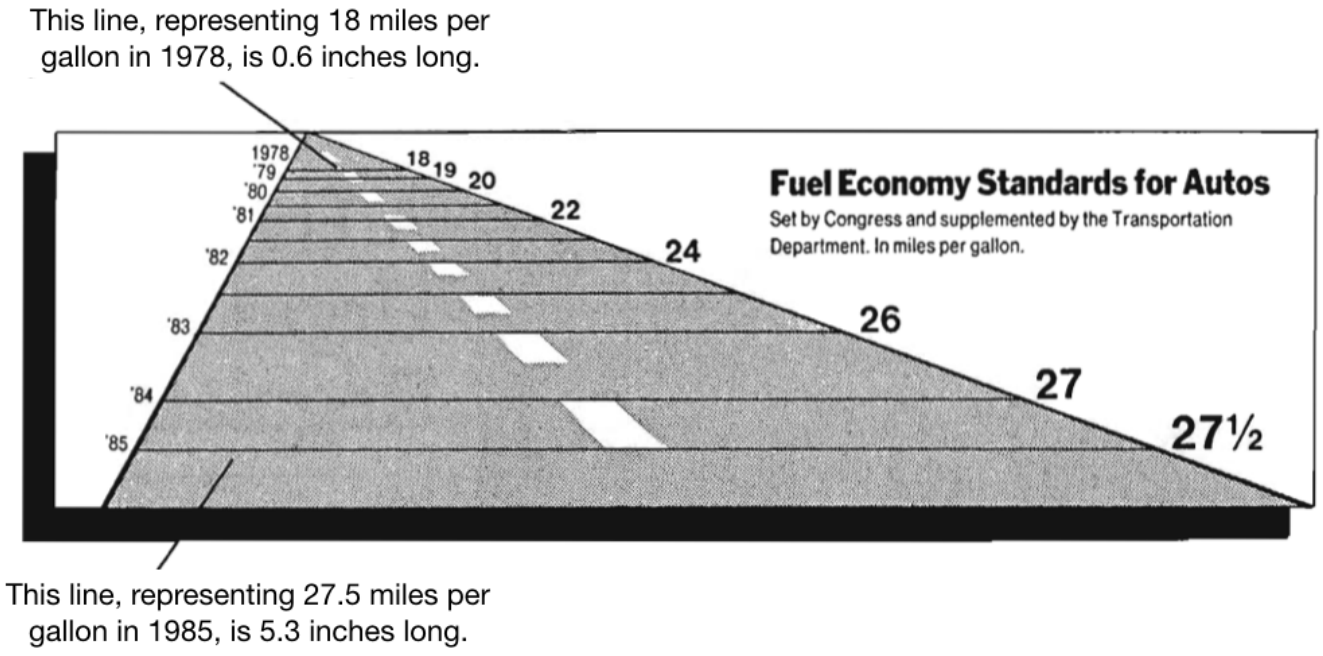
- Ideally, lie factor should be 1. Anything above 1.05 or less than 0.95 is substantial distortion.

What's wrong with this graph?



Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 57

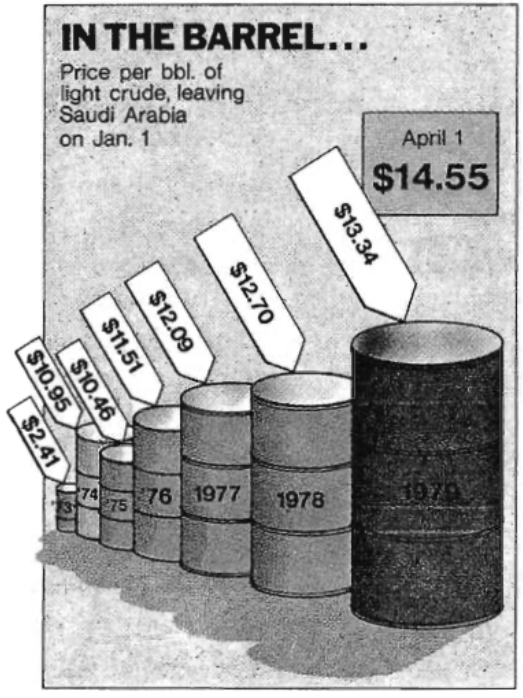
What's wrong with this graph?



$$\text{Lie factor} = \frac{\text{size of effect in the graph}}{\text{size of effect in data}} = \frac{\frac{5.3-0.6}{0.6}}{\frac{27.5-18.0}{18}} = 14.8$$

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 57

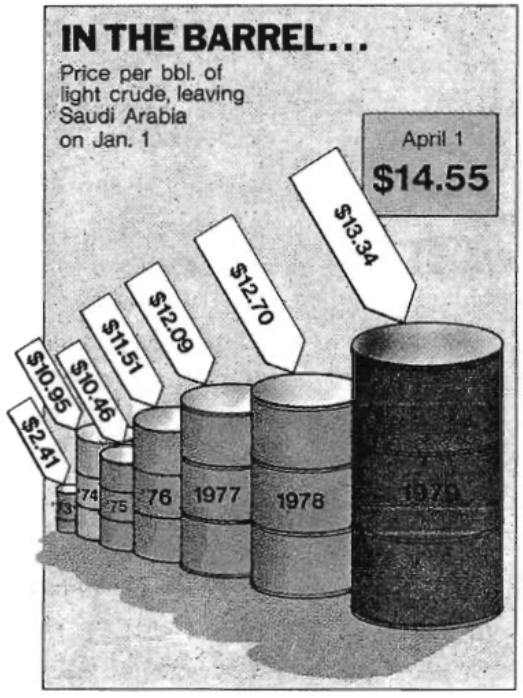
What's wrong with this graph?



Here, the lie factor is 9.4.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 62

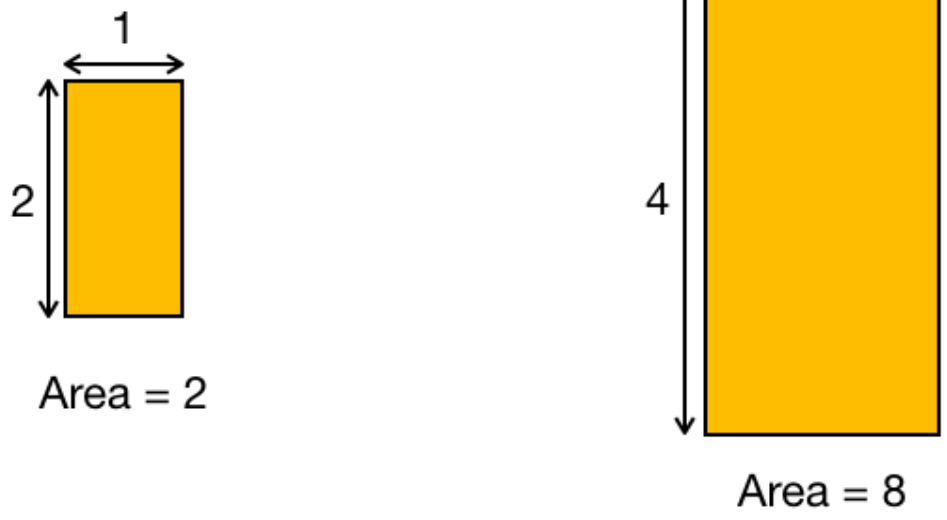
What's wrong with this graph?



The number of variable dimensions used should not exceed the number of dimensions: here, we are depicting increase in price (single dimension) with area of the barrels (2 dimensions).

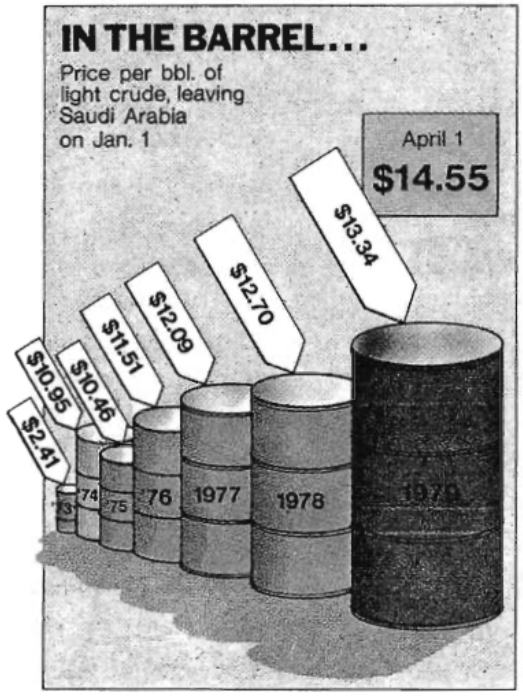
Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 62

Varying the number of dimensions



- Notice that if both the width and length of the bar changes, the area changes by the product of that.
- Humans are better are perceiving area than lengths, so increasing both width and length makes the new rectangle seem a lot bigger.

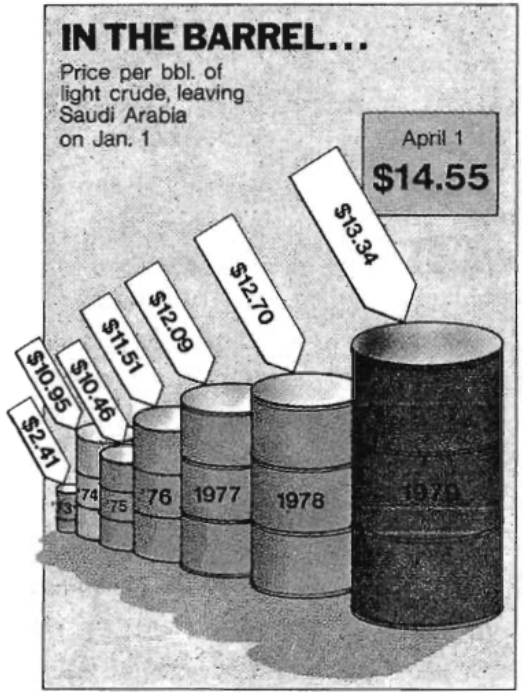
What's wrong with this graph?



The number of variable dimensions used should not exceed the number of dimensions: here, we are depicting increase in price (single dimension) with area of the barrels (2 dimensions).

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 62

What's wrong with this graph?

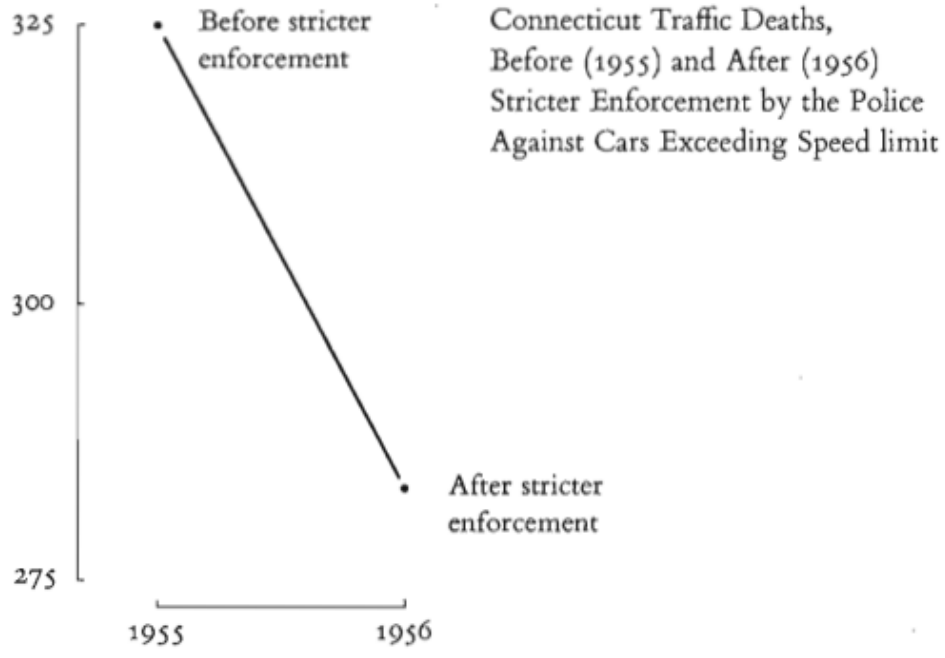


The number of variable dimensions used should not exceed the number of dimensions: here, we are depicting increase in price (single dimension) with area of the barrels (2 dimensions). Our eyes are drawn to the **size** of the barrel, not the **height**.

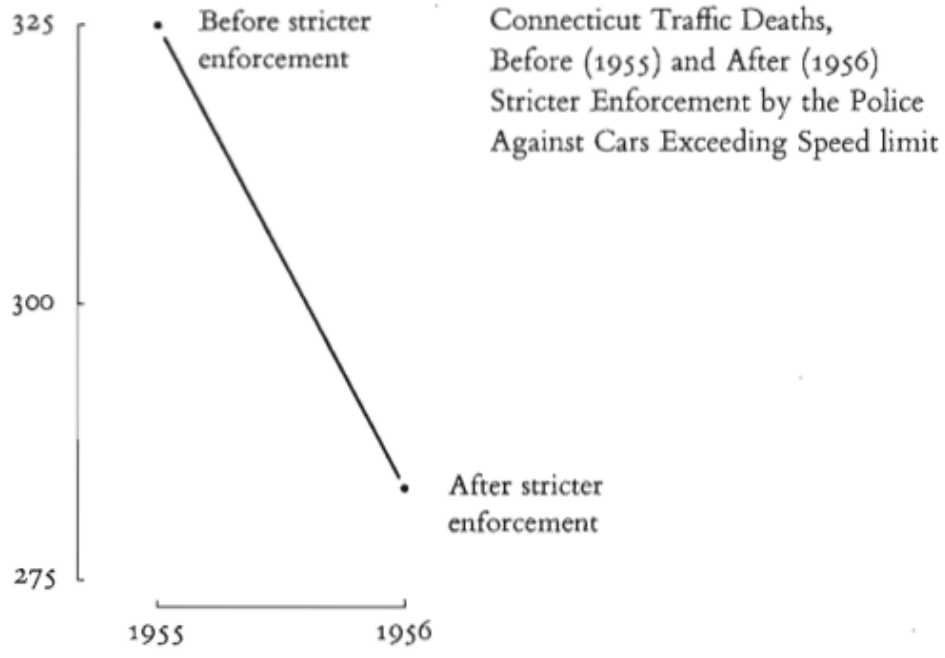
Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 62

Show data variation, not design variation

Compared to what?



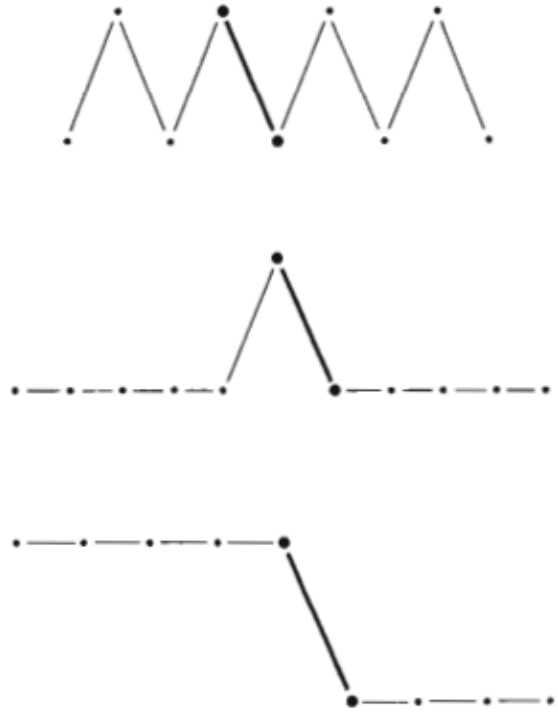
Compared to what?



Can we conclude that stricter law enforcement reduces traffic deaths?

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 74

Compared to what?

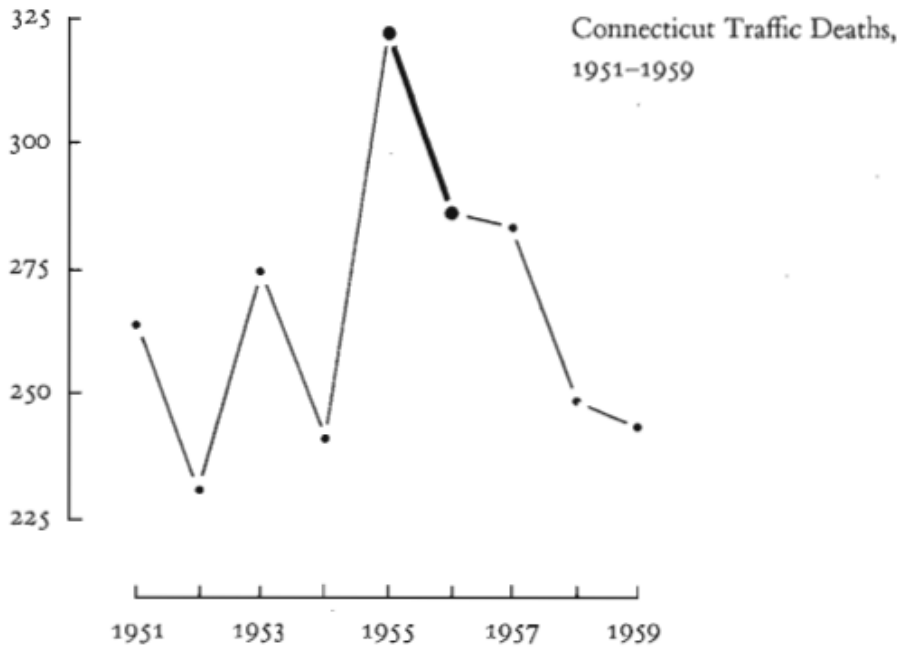


Without context, impossible to say which plot these points could be a part of.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 75

Compared to what?

A few more data points add immensely to the account:



Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 74

Graphics must be quoted in context

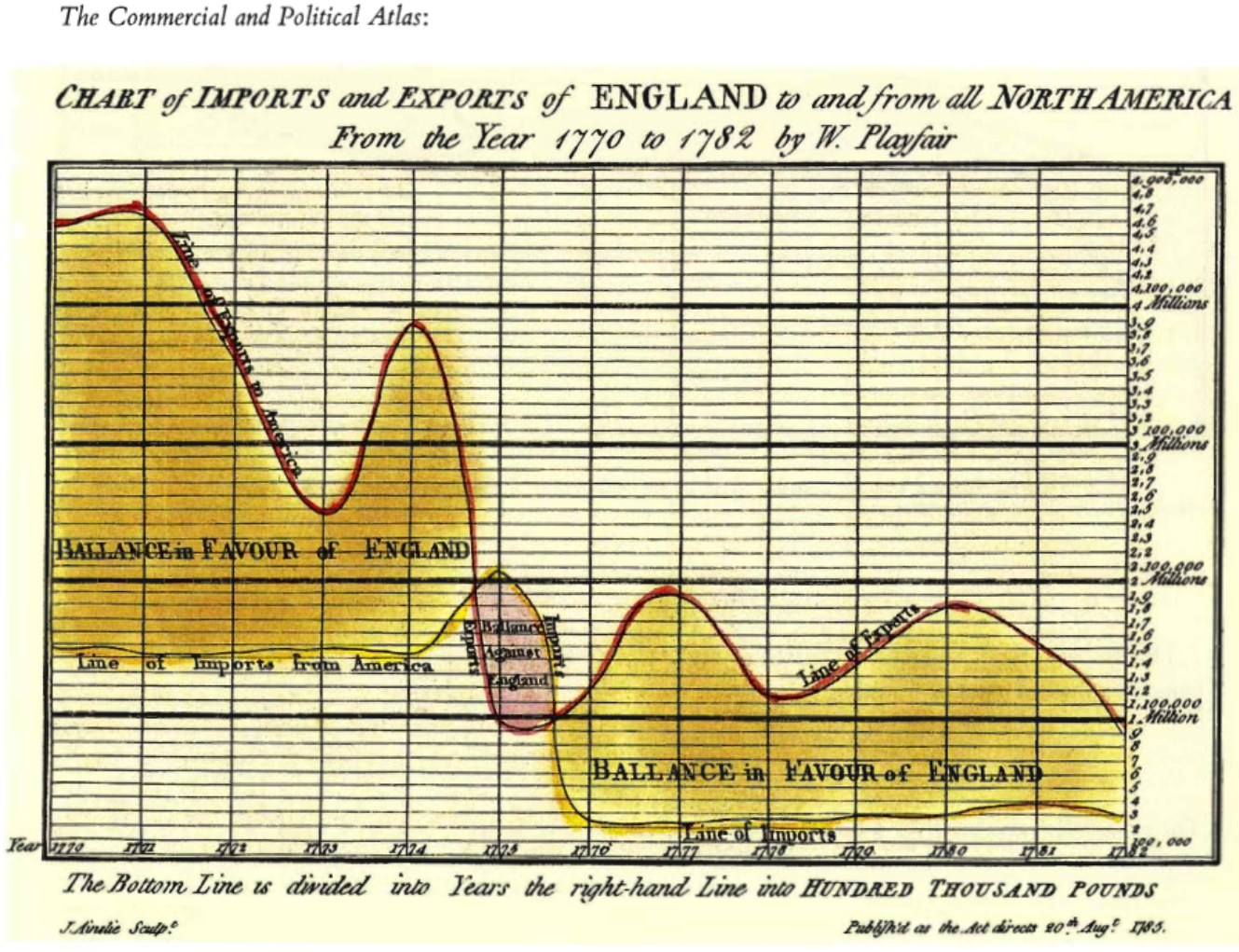
Outline

What should be on a graph?

Above all else, show the data

What should be on a graph?

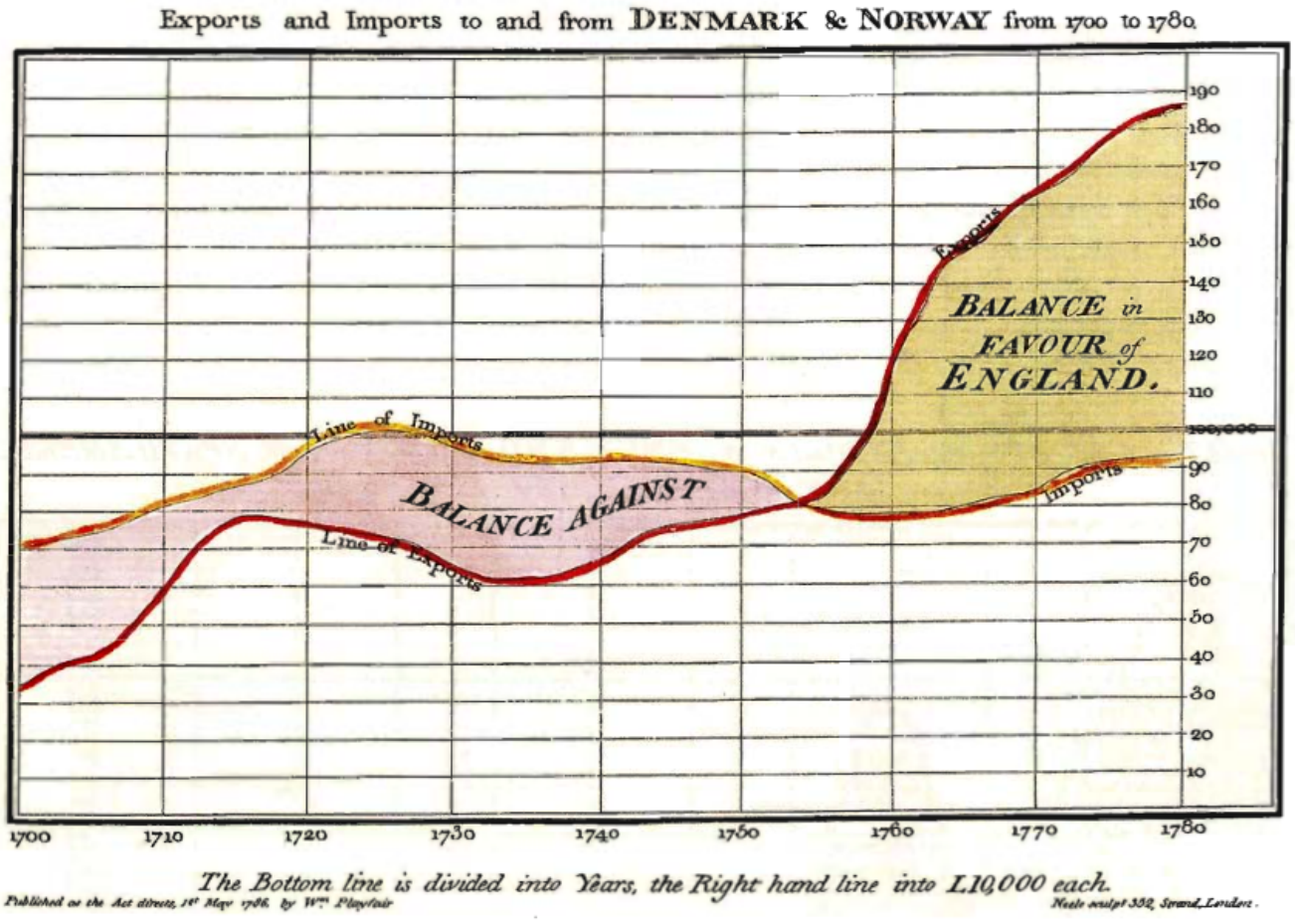
Above all else, show the data



Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 91-92

What should be on a graph?

Above all else, show the data



Data Ink ratio

- **Data-ink** is the amount of ink that cannot be removed without removing infomation.

Data Ink ratio

- **Data-ink** is the amount of ink that cannot be removed without removing infomation.

$$\begin{aligned}\text{Data-Ink ratio} &= \frac{\text{data-ink}}{\text{total ink used in the graphic}} \\ &= \text{Proportion of graphic's ink used} \\ &\quad \text{to display non-redundant data information} \\ &= 1 - \text{erasable proportion of the graphic}\end{aligned}$$

Data Ink ratio

- **Data-ink** is the amount of ink that cannot be removed without removing infomation.

$$\begin{aligned}\text{Data-Ink ratio} &= \frac{\text{data-ink}}{\text{total ink used in the graphic}} \\ &= \text{Proportion of graphic's ink used} \\ &\quad \text{to display non-redundant data information} \\ &= 1 - \text{erasable proportion of the graphic}\end{aligned}$$

Maximise the data-ink, within reason.

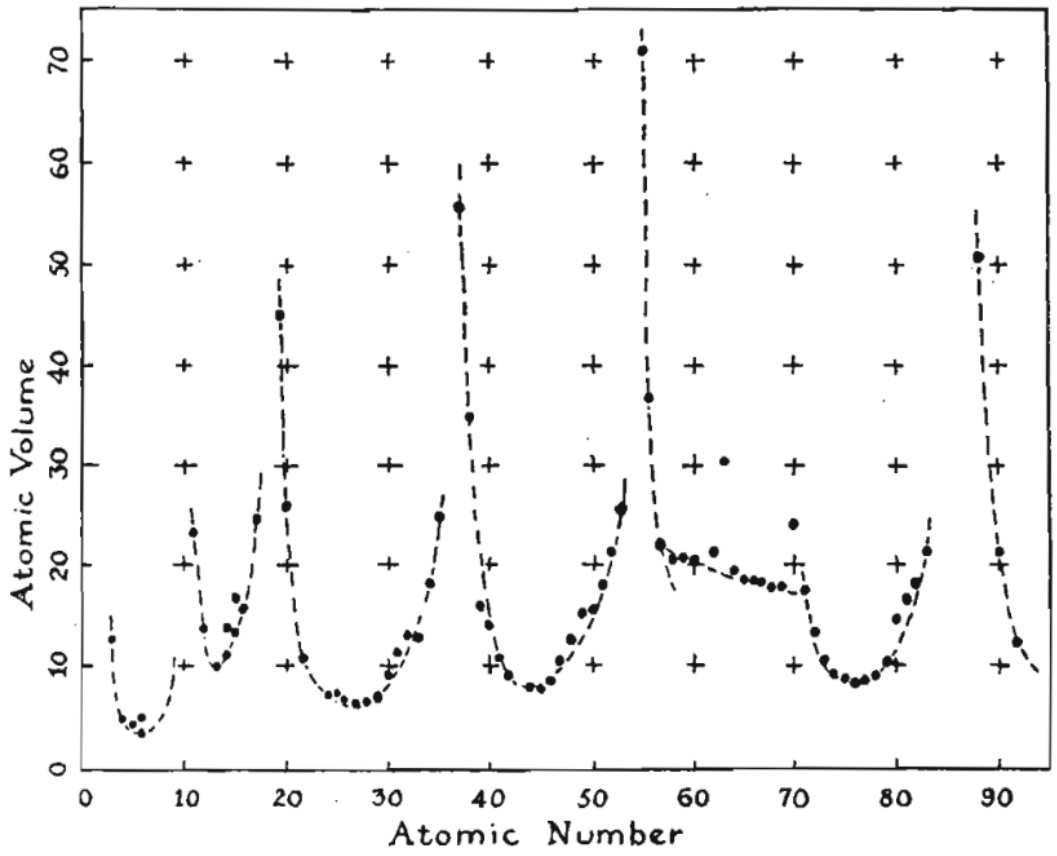
Data Ink ratio

- **Data-ink** is the amount of ink that cannot be removed without removing infomation.

$$\begin{aligned}\text{Data-Ink ratio} &= \frac{\text{data-ink}}{\text{total ink used in the graphic}} \\ &= \text{Proportion of graphic's ink used} \\ &\quad \text{to display non-redundant data information} \\ &= 1 - \text{erasable proportion of the graphic}\end{aligned}$$

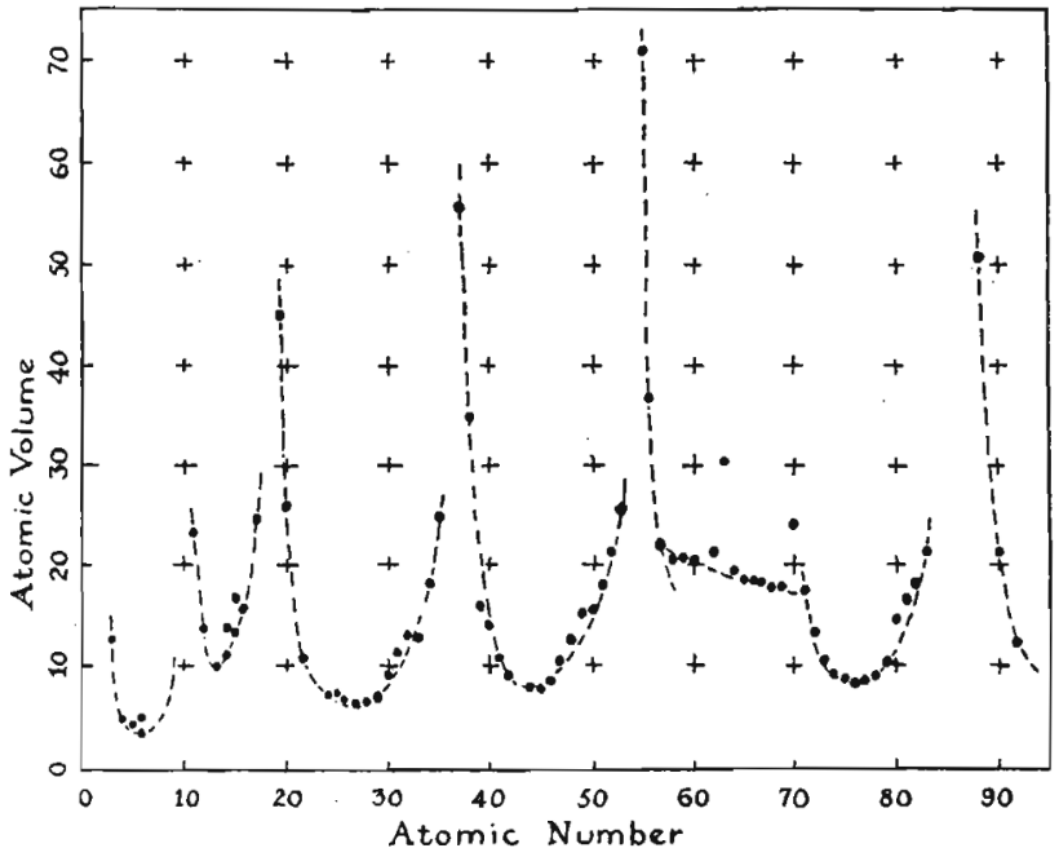
Maximise the data-ink, within reason. Minimize the non-data ink.

Example of maximizing data-ink



- Consider this figure that plots the volume of an atom of an element to its atomic number.

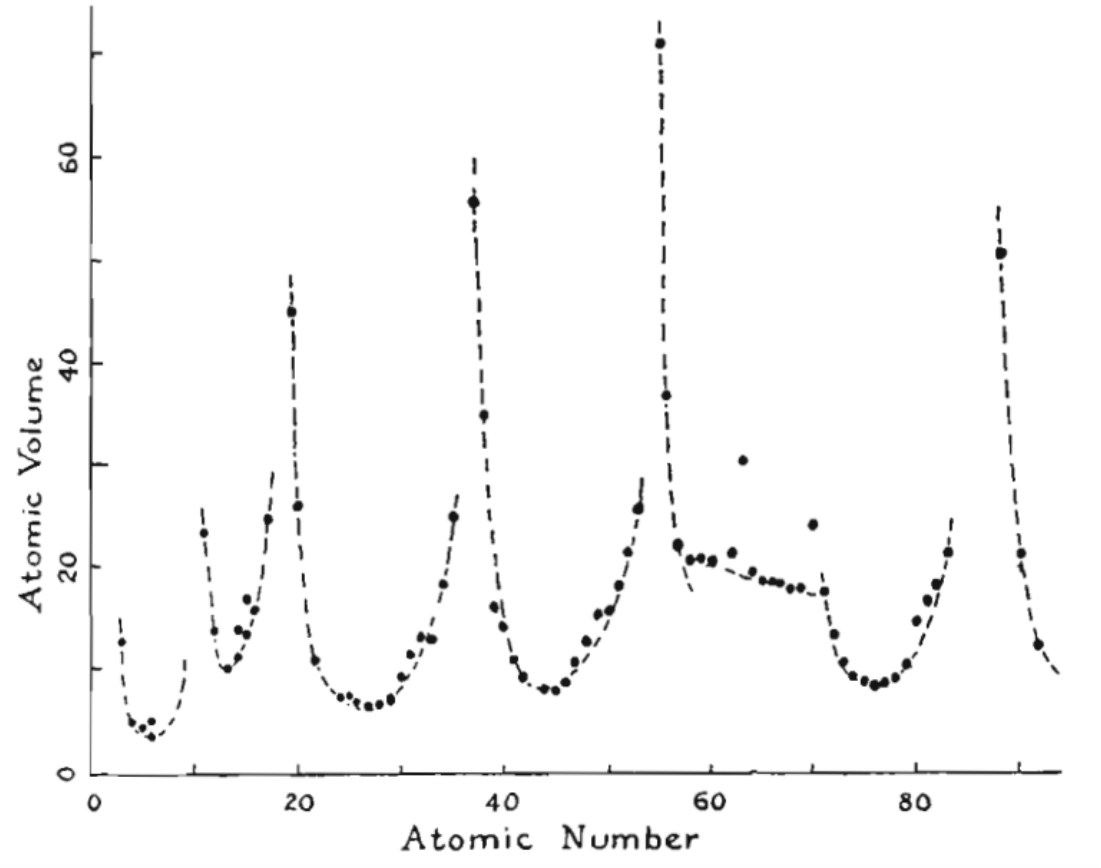
Example of maximizing data-ink



- Consider this figure that plots the volume of an atom of an element to its atomic number.
- The periodicity is obscured by the grid.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 102-105

Example of maximizing data-ink



- Removing the grid helps us see the trend better.

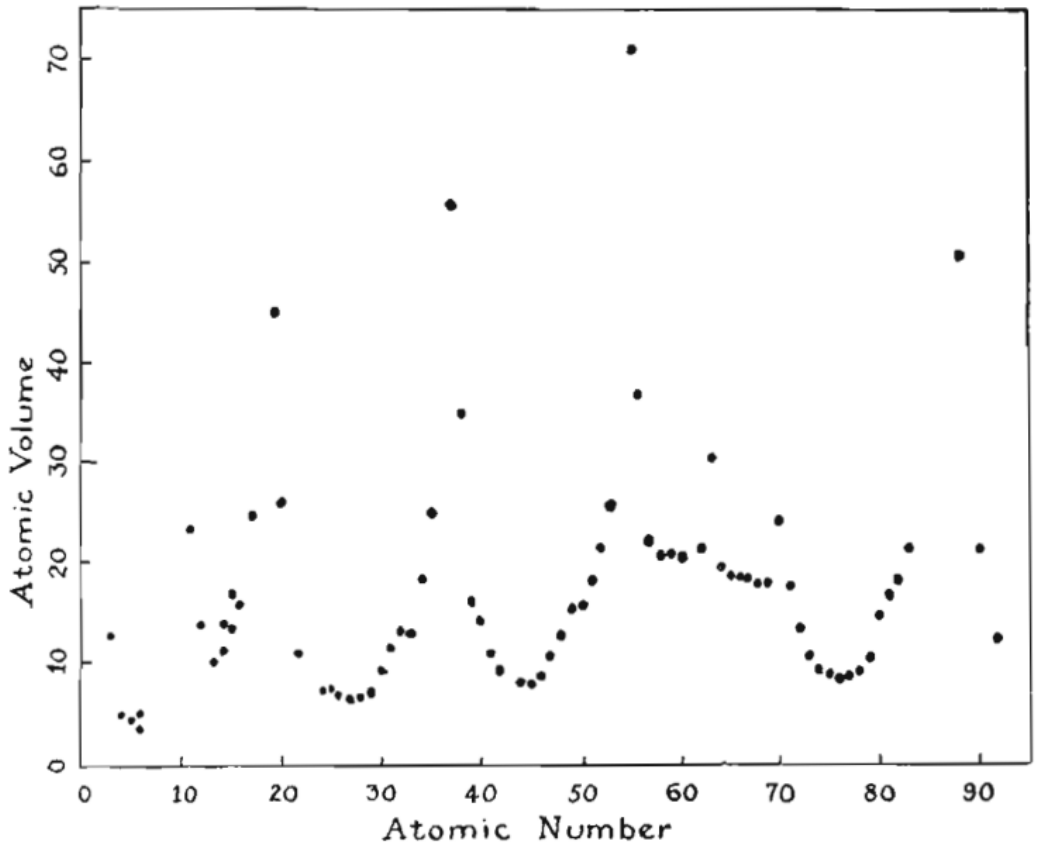
Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 102-105

Example of maximizing data-ink

Do we need the reference curves?

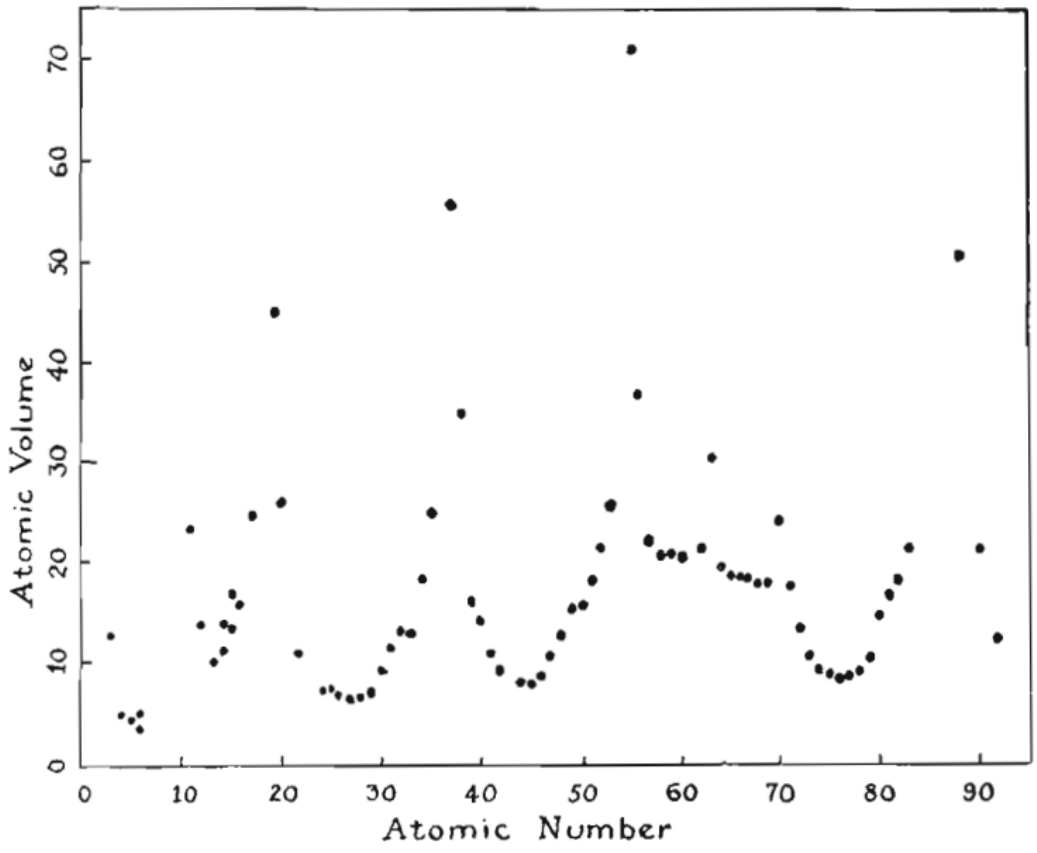
Example of maximizing data-ink

Do we need the reference curves?



Example of maximizing data-ink

Do we need the reference curves?



Yes! They help us see the periodicity better

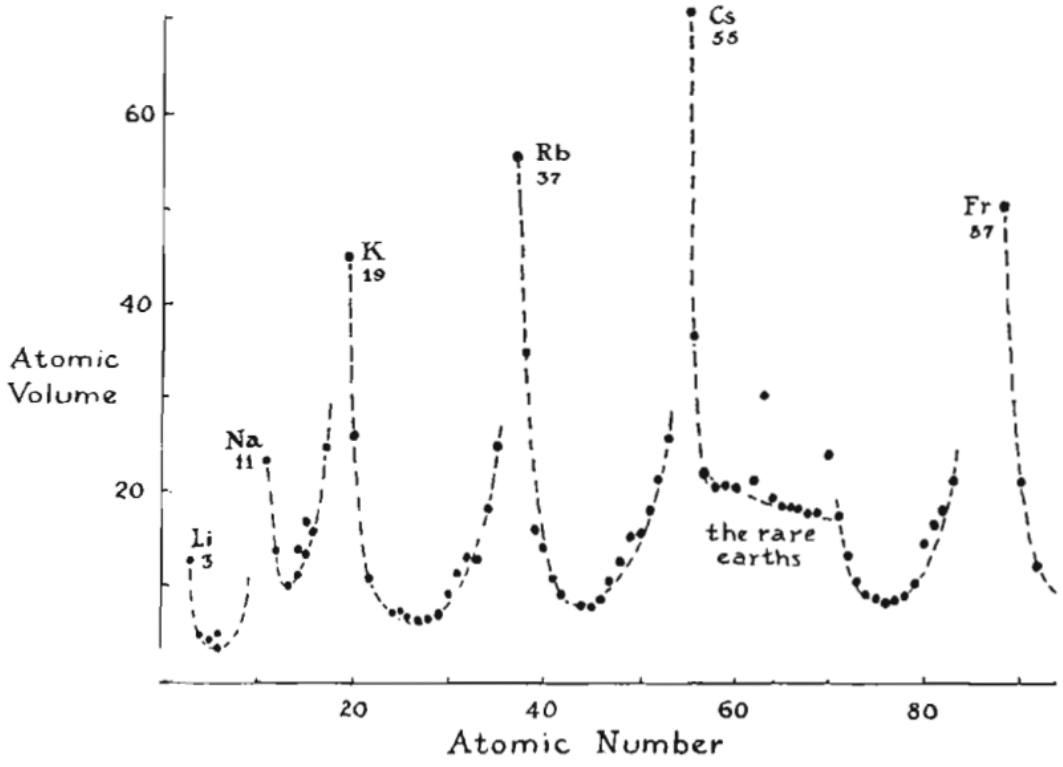
Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 102-105

Example of maximising data-ink

We can now add more information to the graph, for example the alkali metal at the beginnning of each cycle.

Example of maximising data-ink

We can now add more information to the graph, for example the alkali metal at the beginnning of each cycle.



Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 102-105

Steps to follow

- Above all else, show the data

Steps to follow

- Above all else, show the data
- Maximise the data-ink ratio

Steps to follow

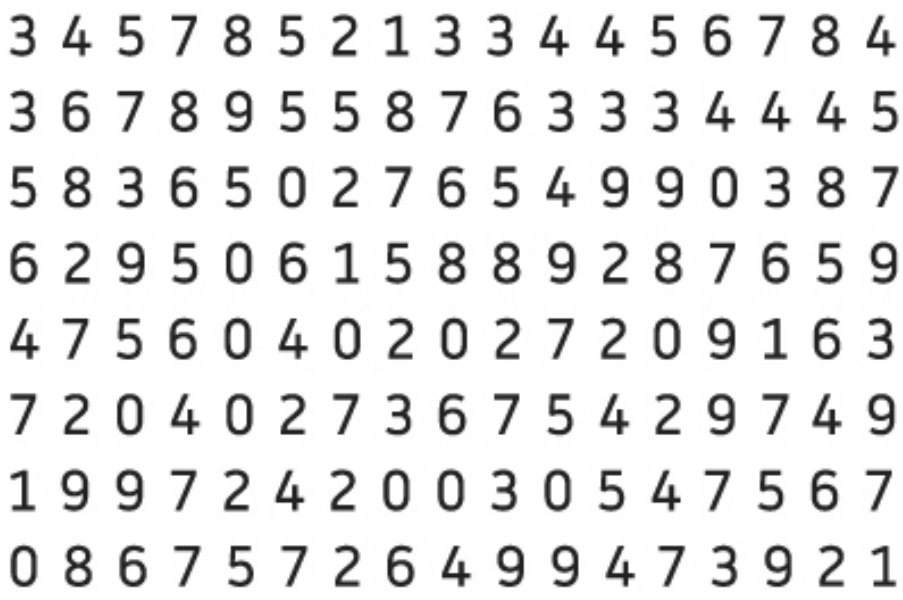
- Above all else, show the data
- Maximise the data-ink ratio
  - Erase non-data ink

Steps to follow

- Above all else, show the data
- Maximise the data-ink ratio
  - Erase non-data ink
  - Erase redundant data-ink

Steps to follow

- Above all else, show the data
- Maximise the data-ink ratio
  - Erase non-data ink
  - Erase redundant data-ink
- Revise and edit.



Count the number of 7's in the table of numbers.

https://datascience.aero/brain-data-visualization

How about now?

How about now? The change in color makes the recognition instant.

https://datascience.aero/brain-data-visualization

Pre-attentive processing

- Humans can distinguish between differences in line length, shape orientation and color without a lot of effort.

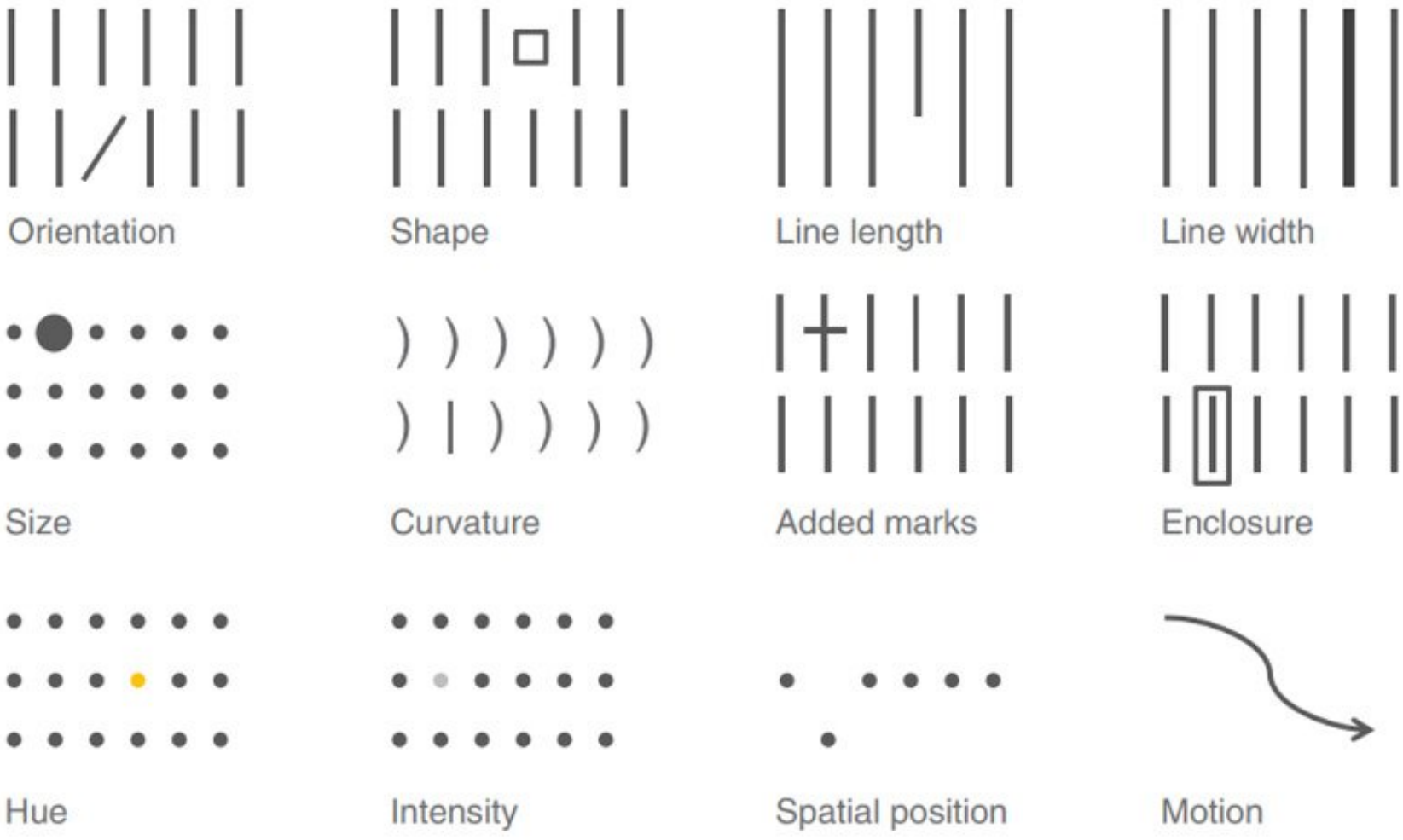
Pre-attentive processing

- Humans can distinguish between differences in line length, shape orientation and color without a lot of effort.
- These are called **pre-attentive attributes**

Pre-attentive processing

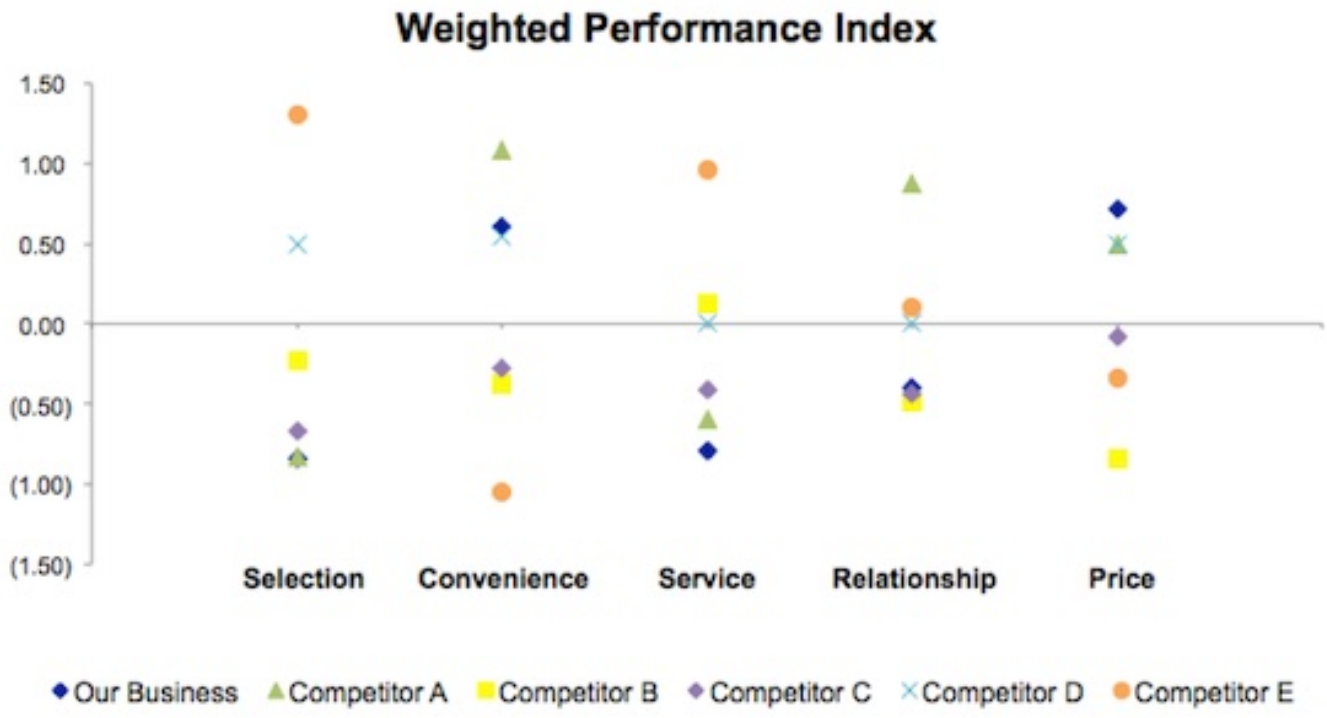
- Humans can distinguish between differences in line length, shape orientation and color without a lot of effort.
- These are called **pre-attentive attributes**
- We can use these to make more effective graphics!

Examples of pre-attentive graphics



Nussbaumer Knaflic's Examples of Preattentive Attributes

Using pre-attentive graphics to make strong visualizations



- Chaotic, no clear focus to the graph; hard to get full scope of data

Credit: Cole Nussbaumer Knaflic

Using pre-attentive graphics to make strong visualizations

- Used color and position to make clear what the graphic conveys.

Credit: Cole Nussbaumer Knaflic

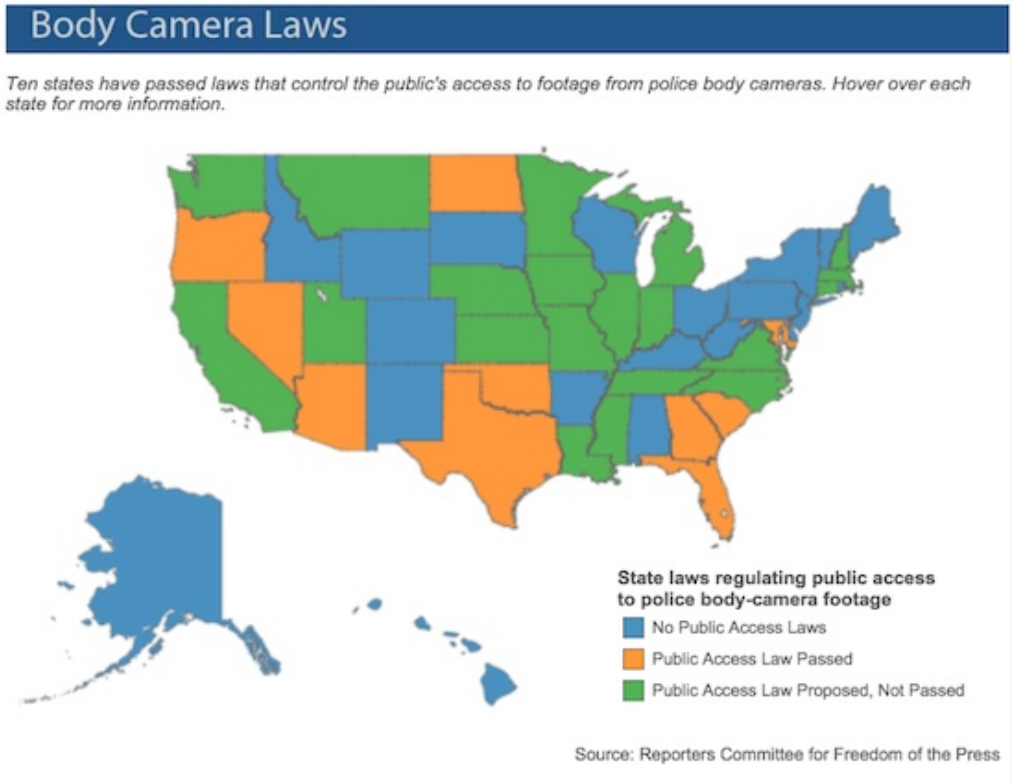
What colors should we use?

- Depends on the data!

What colors should we use?

- Depends on the data!

For distinct groups, use categorical colors.

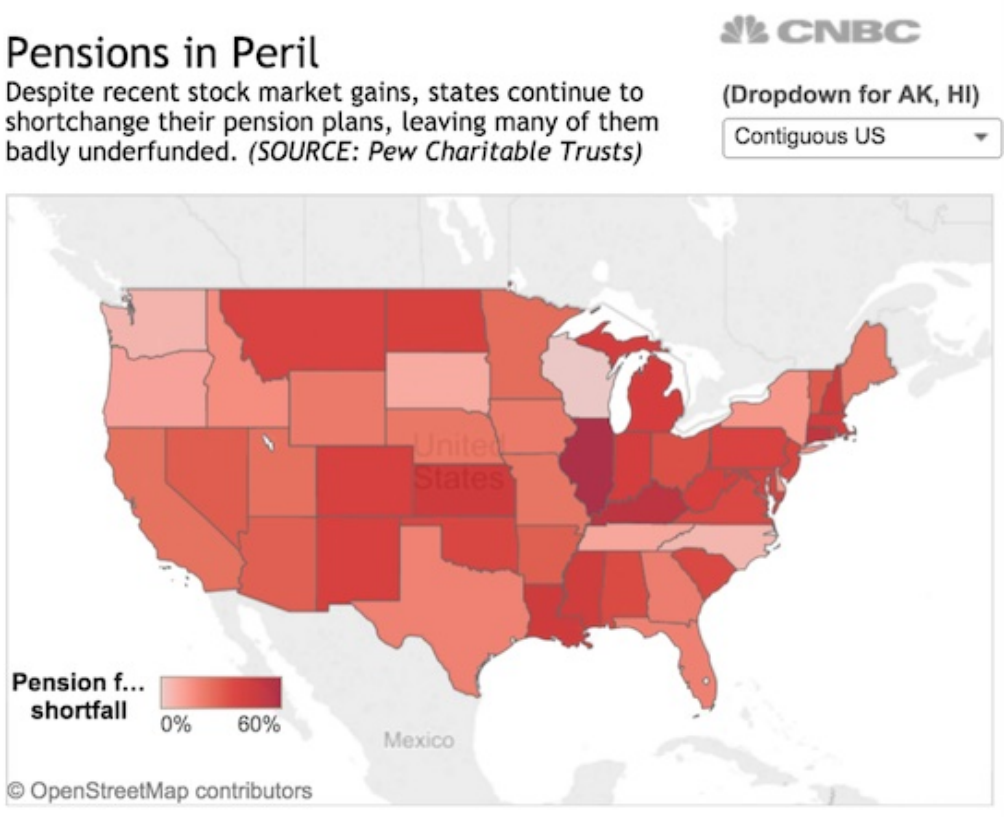


`http://mediashift.org/2016/02/checklist-does-your-data-visualization-say-what-you-think-it-says/`

What colors should we use?

- Depends on the data!

For continuous values, use sequential colors

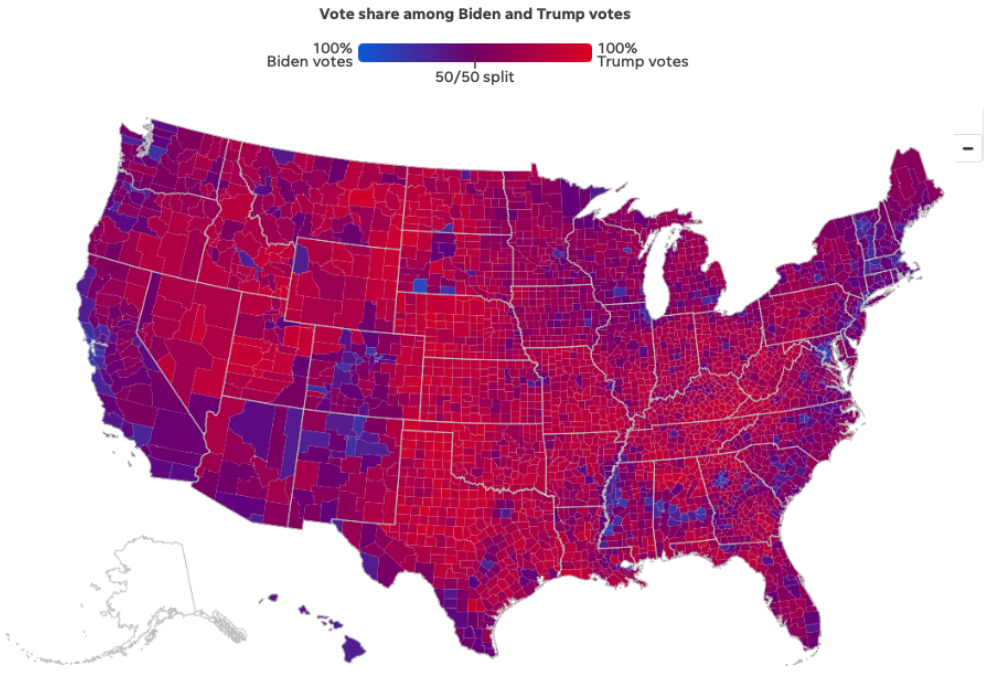


http://mediashift.org/2016/02/checklist-does-your-data-visualization-say-what-you-think-it-says/

What colors should we use?

- Depends on the data!

For ordered values that have a clear midpoint, we use diverging colors



https://www.usatoday.com/in-depth/graphics/2020/11/10/election-maps-2020-america-county-results-more-voters/6226197002/

A note about using color

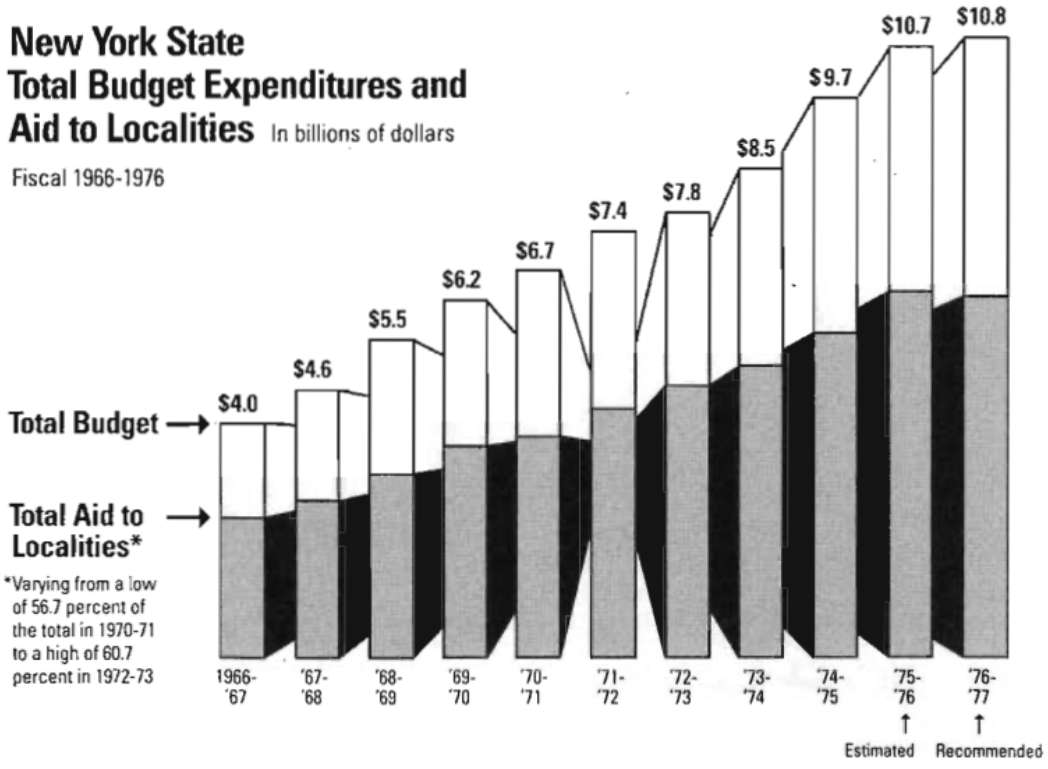
- A large fraction of the population is colorblind.

A note about using color

- A large fraction of the population is colorblind.
- Important to use color palettes that are colorblind friendly.

And finally . . .

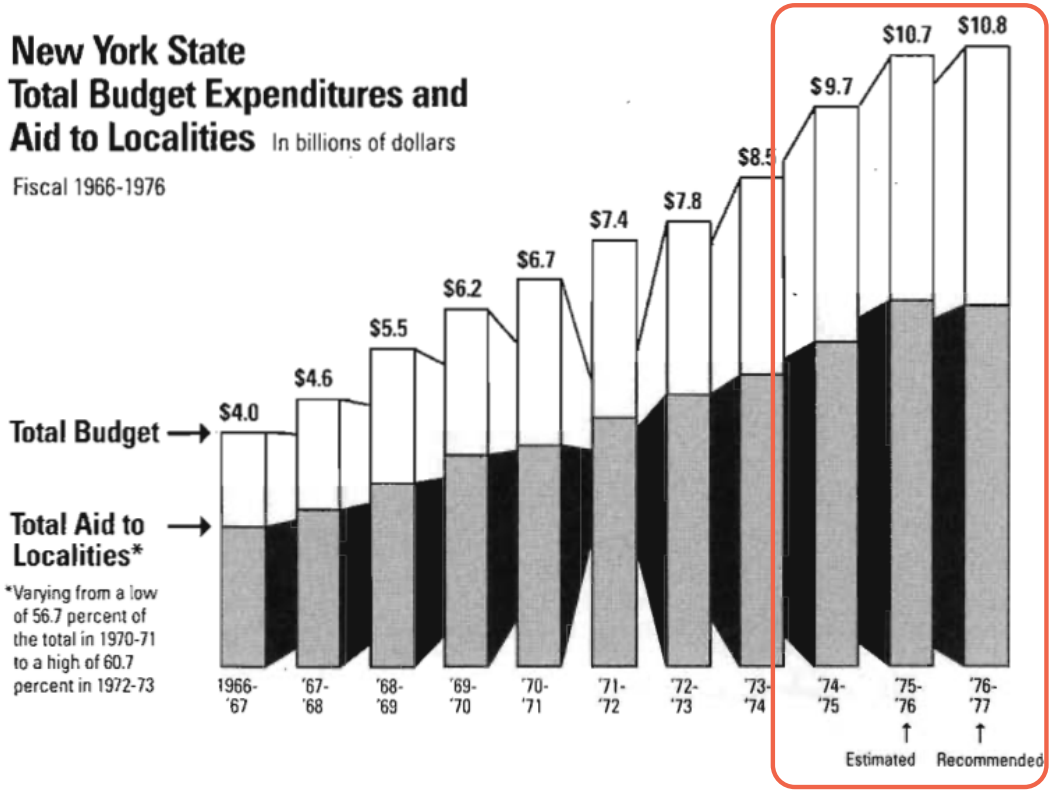
What's wrong with this graph?



New York Times, February 1, 1976, p. IV–6.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 66-67

What's wrong with this graph?

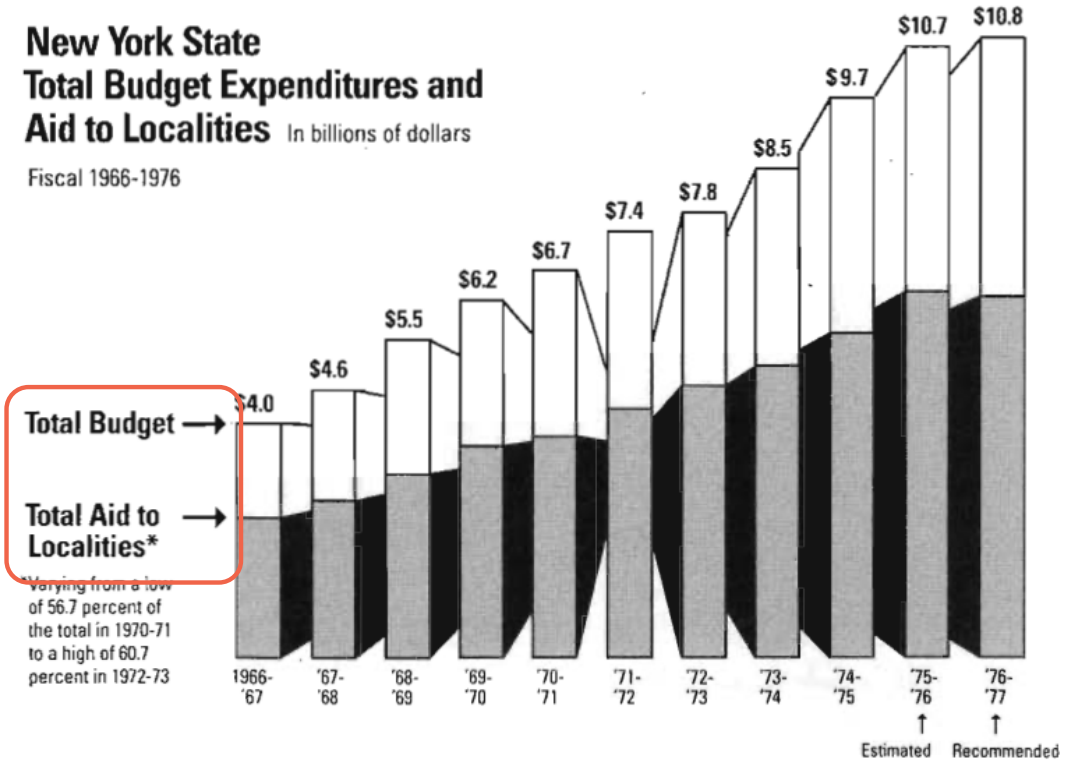


New York Times, February 1, 1976, p. IV-6.

- These three cuboids are not in the same plane as the remaining 8. This makes it seem like these are larger than the remaining.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 66-67

What's wrong with this graph?

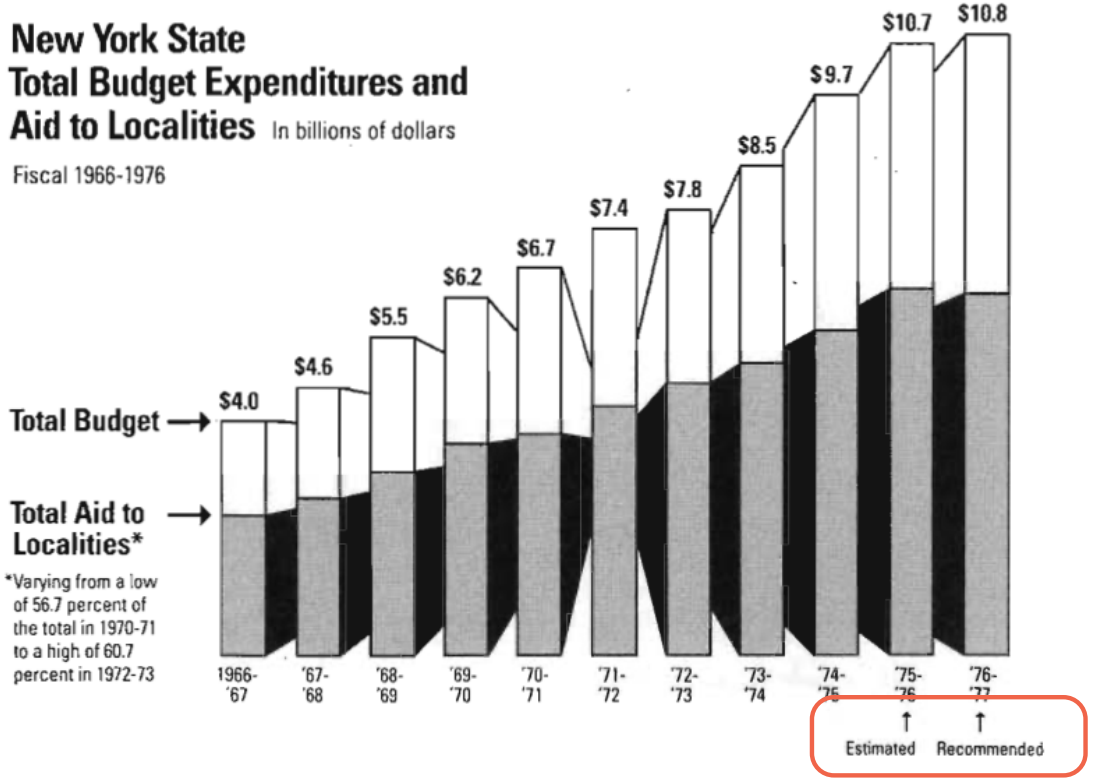


New York Times, February 1, 1976, p. IV-6.

- Text here along with the horizontal arrows make it seem like the value from 1966-1967 is stable, and values after that have shot up.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 66-67

What's wrong with this graph?

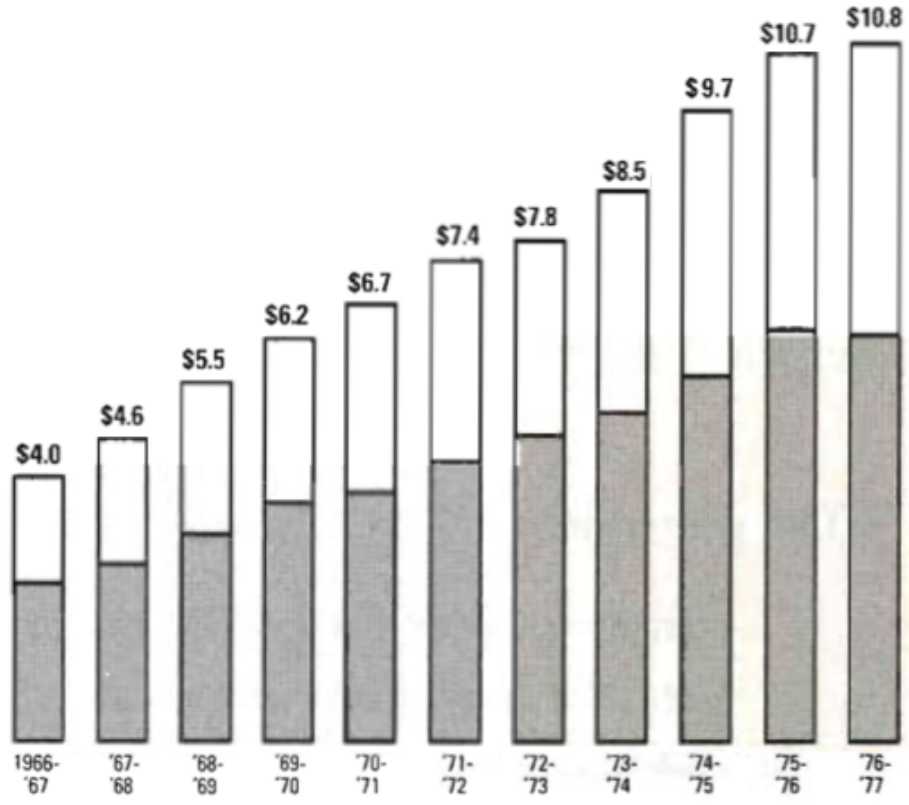


New York Times, February 1, 1976, p. IV-6.

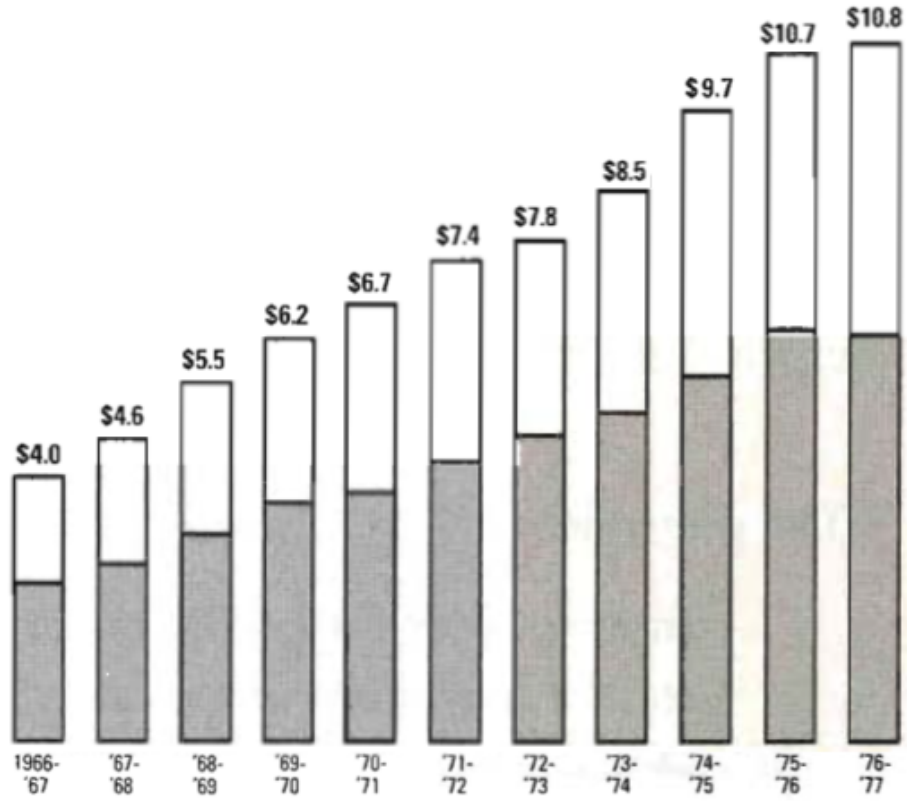
- Similarly the vertical arrows suggest that these cuboids are larger than they actualy are.

Tufte, E. R. *The Visual Display of Quatitative Information*. pgs 66-67

What if we remove these distractions?



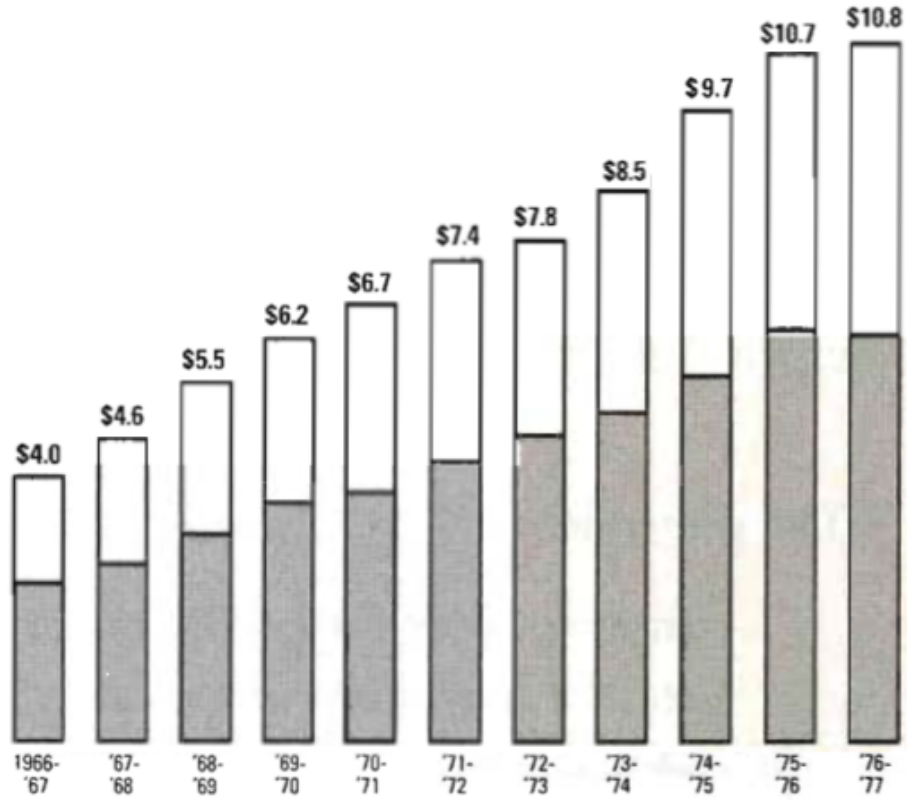
What if we remove these distractions?



- But population increased by 1.7 million during this time . . .

Tufte, E. R. *The Visual Display of Quatitative Information*. pg 67

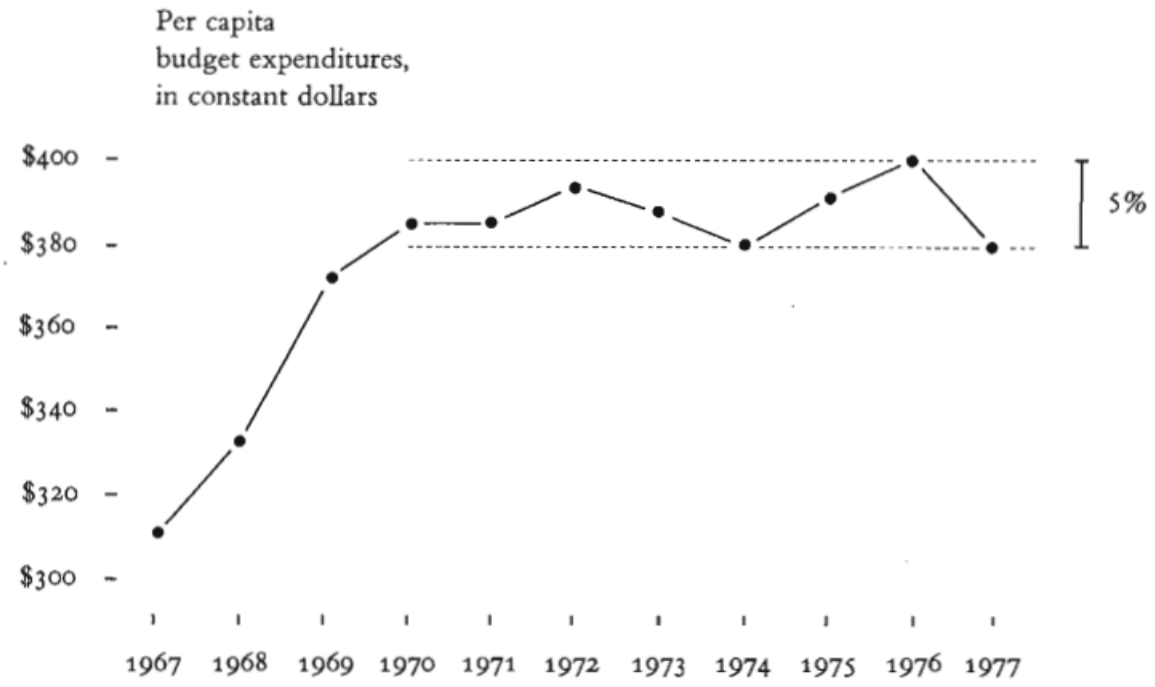
What if we remove these distractions?



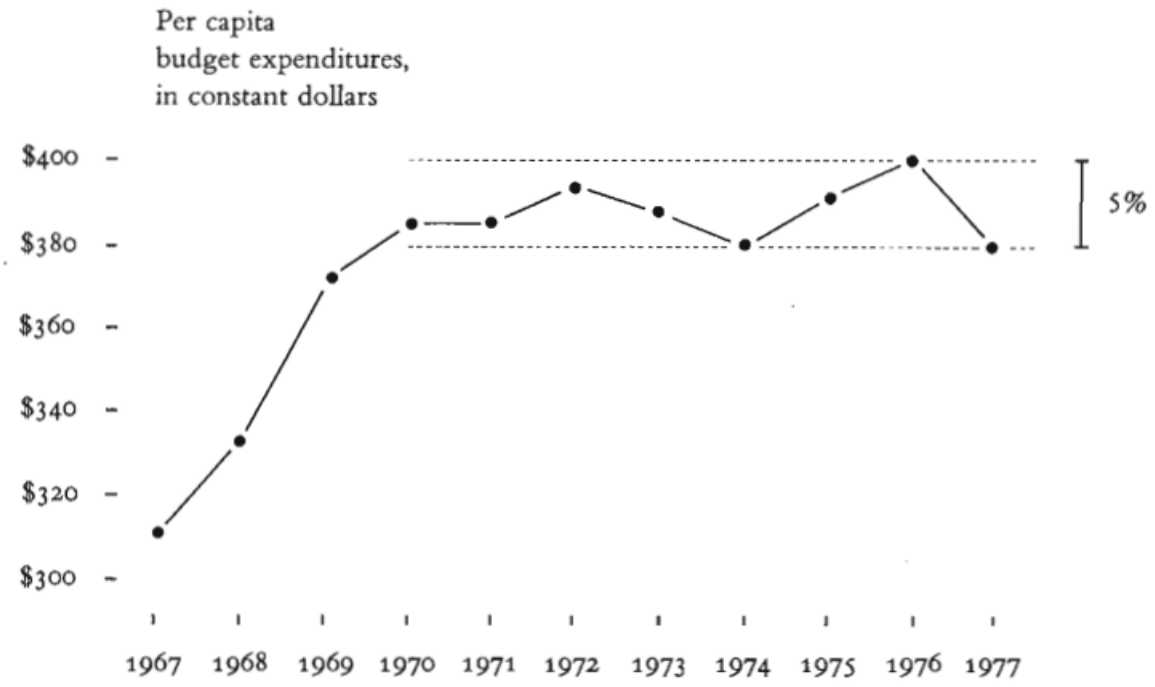
- But population increased by 1.7 million during this time . . .
- . . . and this was a time of large inflation: \$1 in 1967 could purchase goods and services that cost \$2.03 in 1977.

Tufte, E. R. *The Visual Display of Quatitative Information*. pg 67

A better plot



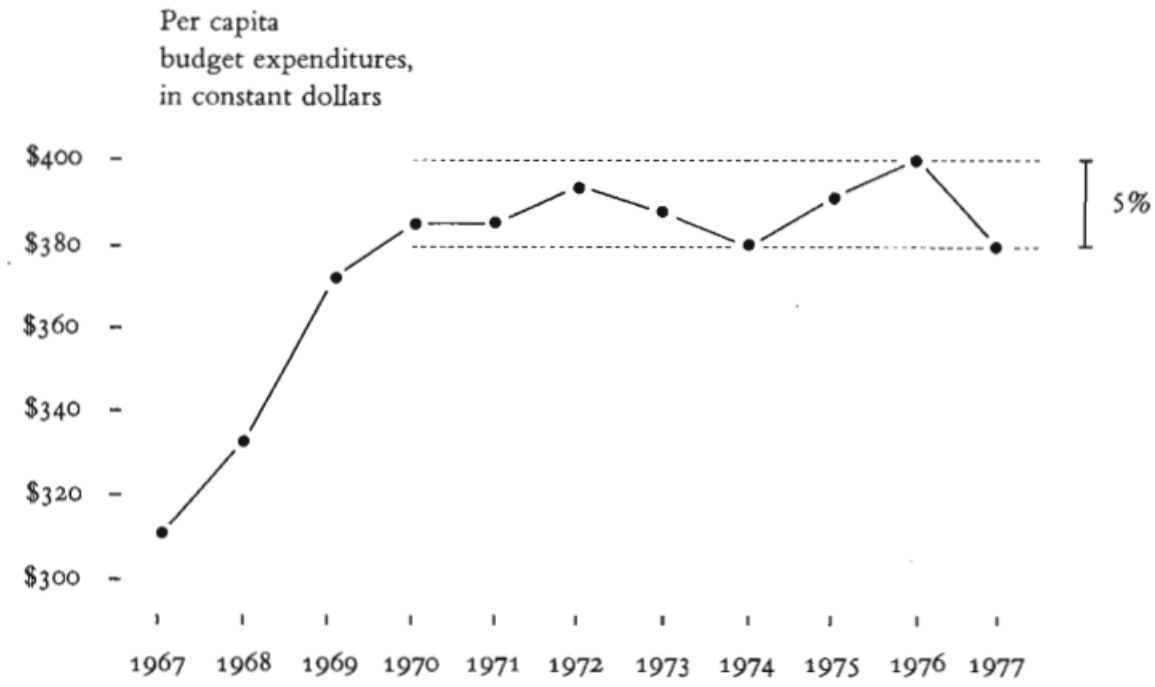
A better plot



- This presents a very different view: the budget increased from 1967 - 1970, and then remained stable.

Tufte, E. R. *The Visual Display of Quatitative Information*. pg 68

A better plot



- This presents a very different view: the budget increased from 1967 - 1970, and then remained stable.
- **No matter how good the graphical principles are, we still need to to visualize the "right" data**

Tufte, E. R. *The Visual Display of Quatitative Information*. pg 68

Outline

Where we are and where we're going

- So far, we visualized data with just one variable.

Where we are and where we're going

- So far, we visualized data with just one variable.
- Now (and for the rest of this class), we're going to consider data with multiple variables.

Motivating questions

How does subjective well-being vary with age?

How does income vary with education level?

How does global temperature vary with carbon dioxide levels in the atmosphere?

Associations

- Most data that we work with has multiple variables.

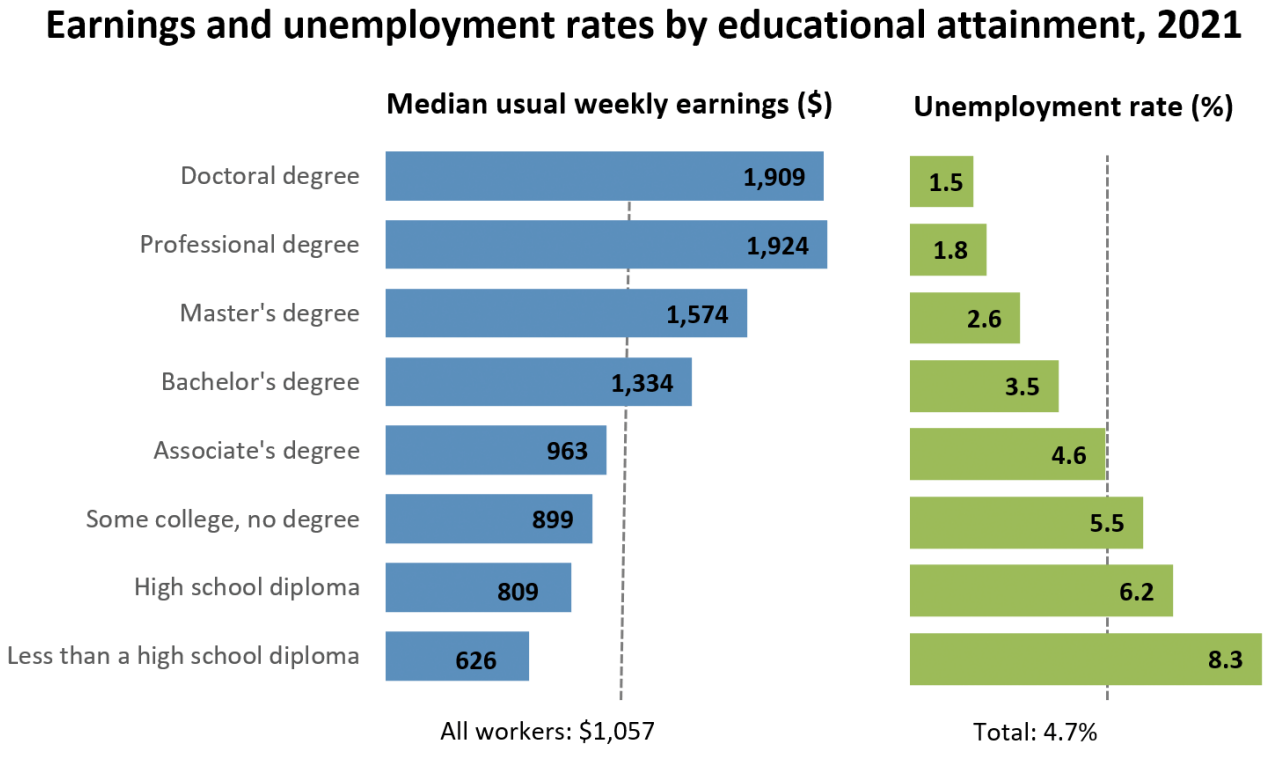
Associations

- Most data that we work with has multiple variables.
- Example: in the ACS dataset, for one person, we have their age, gender, household income, race, ...

Associations

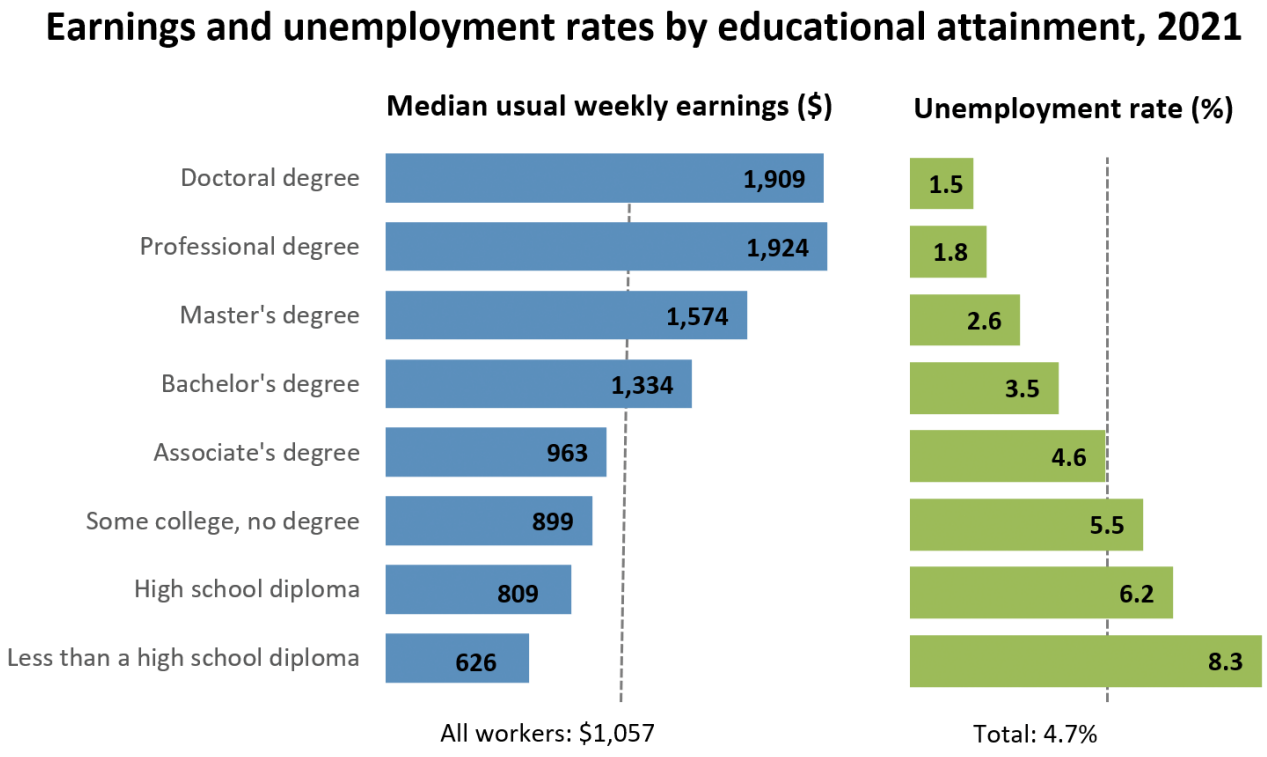
- Most data that we work with has multiple variables.
- Example: in the ACS dataset, for one person, we have their age, gender, household income, race, ...
- Loosely, associations measure **relationships** between variables.

Examples of associations: Education and income



https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

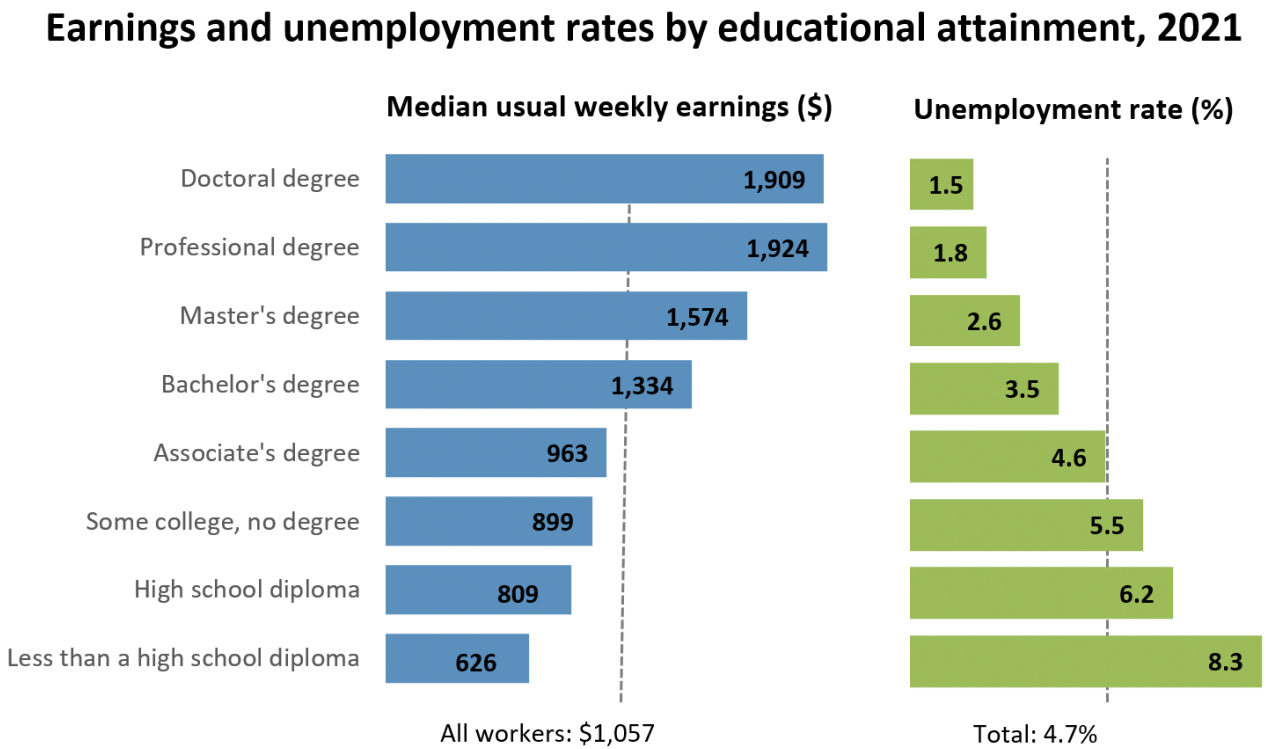
Examples of associations: Education and income



The median weekly earnings increases with the educational attainment.

https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

Examples of associations: Education and income

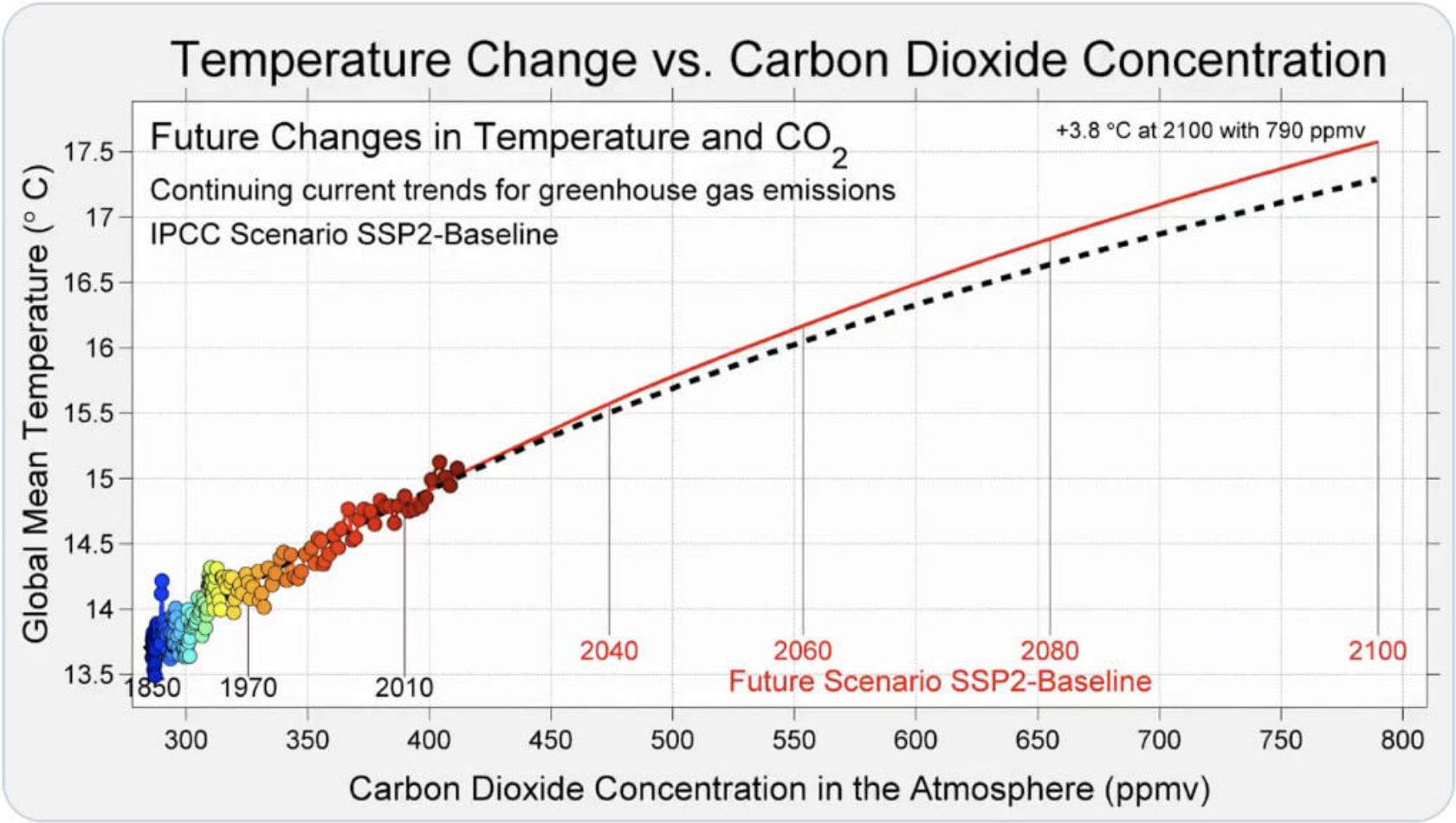


The median weekly earnings increases with the educational attainment.

This uses a **categorical, ordinal** variable and a **quantitative** variable

https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

Examples of associations: Global temperature and carbon dioxide levels in the atmosphere

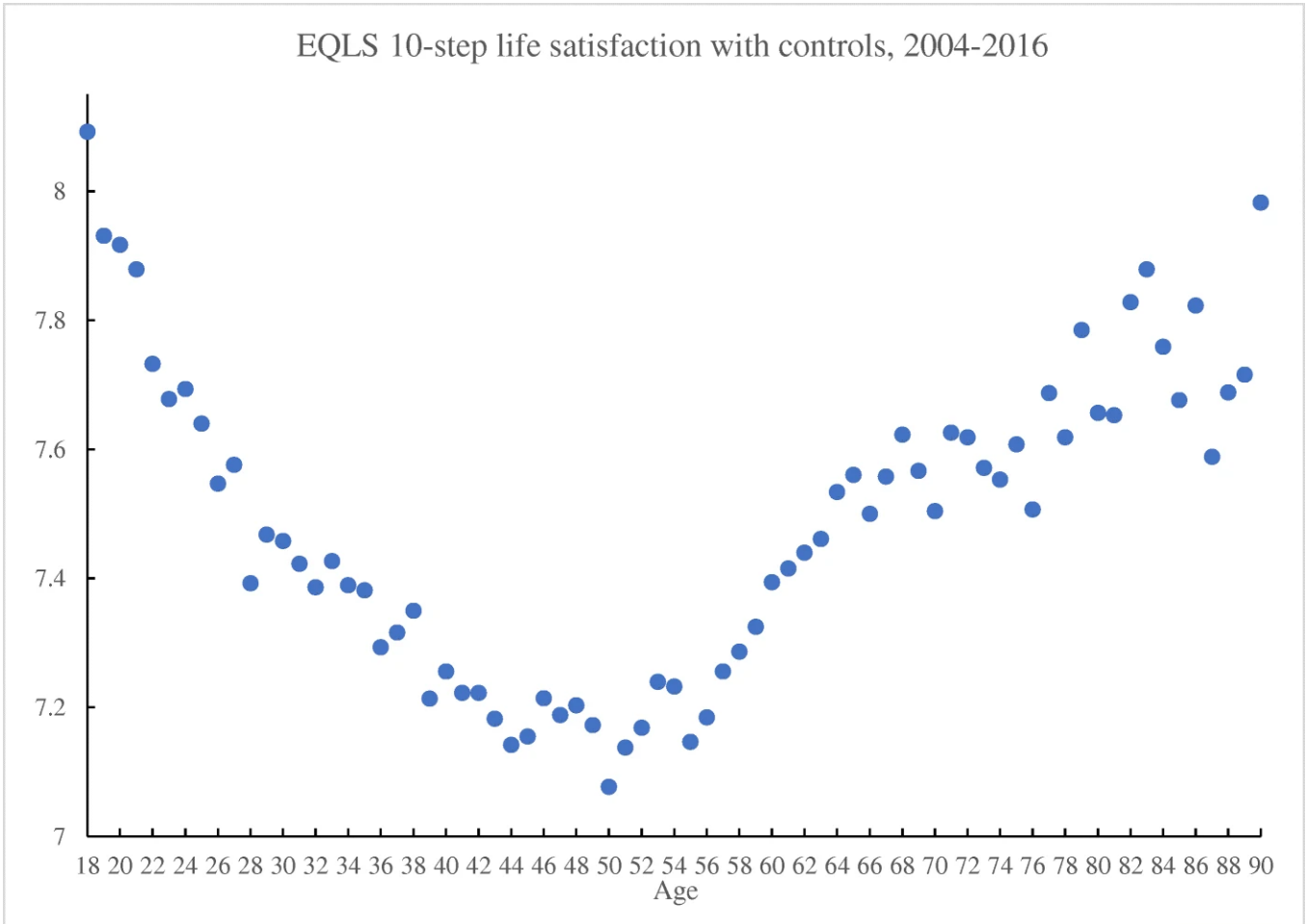


Here, both variables are quantitative (and continuous).

http://berkeleyearth.org/dv/temperature-change-vs-carbon-dioxide-concentration/

Examples of associations: Age and Happiness

Measuring perceived happiness among people from European countries based on age.



Blanchflower, D.G. *Is happiness U-shaped everywhere? Age and subjective well-being in 145 countries.* J Popul Econ 34, 575–624 (2021)

Main takeaway

Association lets us understand how well one variable predicts another in our sample

Where we are and where we're going

- Last week and before: we visualized data with just one variable.

Where we are and where we're going

- Last week and before: we visualized data with just one variable.
- Today: considered **associations**: How 2 variables might be related

Where we are and where we're going

- Last week and before: we visualized data with just one variable.
- Today: considered **associations**: How 2 variables might be related
- Wednesday: What are the types of associations? How can these associations be measured? What can we use associations for?

Where we are and where we're going

- Last week and before: we visualized data with just one variable.
- Today: considered **associations**: How 2 variables might be related
- Wednesday: What are the types of associations? How can these associations be measured? What can we use associations for?

  Homework 2 due tomorrow at 1:29pm