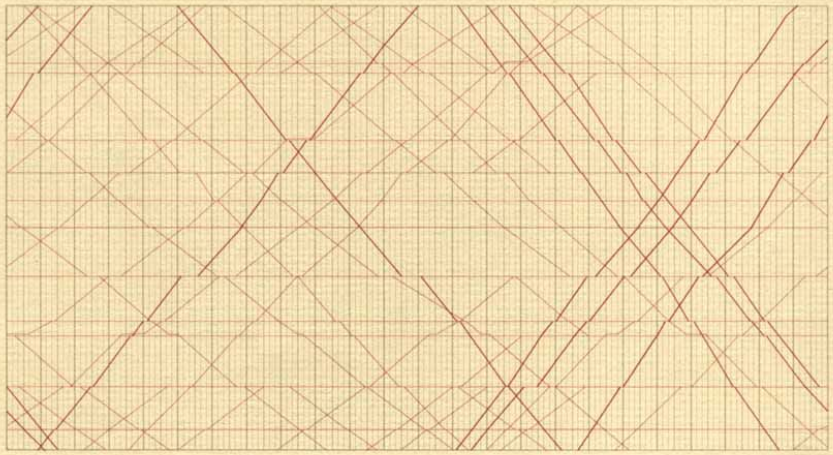
# Data Visualization

## Principles of good visualization

Habet Madoyan

American University of Armenia



# The Visual Display of Quantitative Information

EDWARD R. TUFTE

# Chartjunk

- The term is coined by Edward Tufte in his book - The Visual Display of Quantitative Information.
- Chartjunk refers to all visual elements in charts and graphs that are not necessary to comprehend the information represented on the graph, or that distract the viewer from this information.
- Thus, anything in the graph that does not convey information is a chartjunk.
- If there is something in the graph, that if removed, will not change the message of the graph, is a chartjunk and needs to be removed.

# Chartjunk - data-ink

Edward Tufte also coined a term data-ink - Data-ink is the ink that is used to present information. Thus if you remove some of the data-ink, the message/ information will change. Thus, data-ink is the unremovable part of the graph.

*A large share of ink on a graphic should present data-information, the ink changing as the data change. Data-ink is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented.*

# Chartjunk: data-ink ratio

Data-ink ratio is calculated in the following way:

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used in the graph}}$$

- Data-ink ratio = proportion of a graphic ink used to display non-redundant data information.
- Data-ink ratio = 1- proportion of the graph that can be erased without losing information.

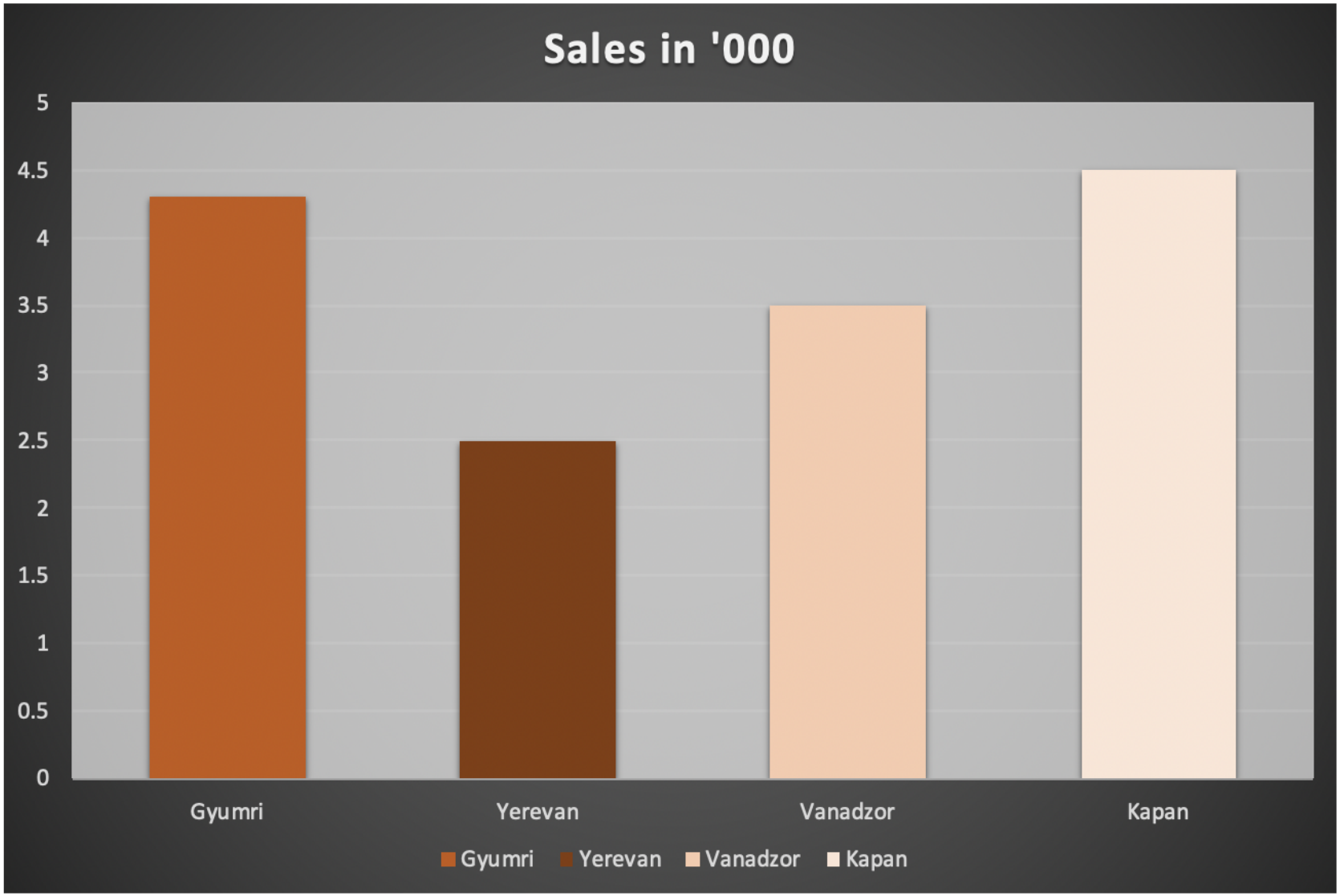
# Chartjunk

*Edward Tufte*

The interior decoration of graphics generates a lot of ink that does not tell the viewer anything new. The purpose of decoration varies—to make the graphic appear more scientific and precise, to enliven the display, to give the designer an opportunity to exercise artistic skills. Regardless of its cause, it is all non-data-ink or redundant data-ink, and it is often chartjunk.

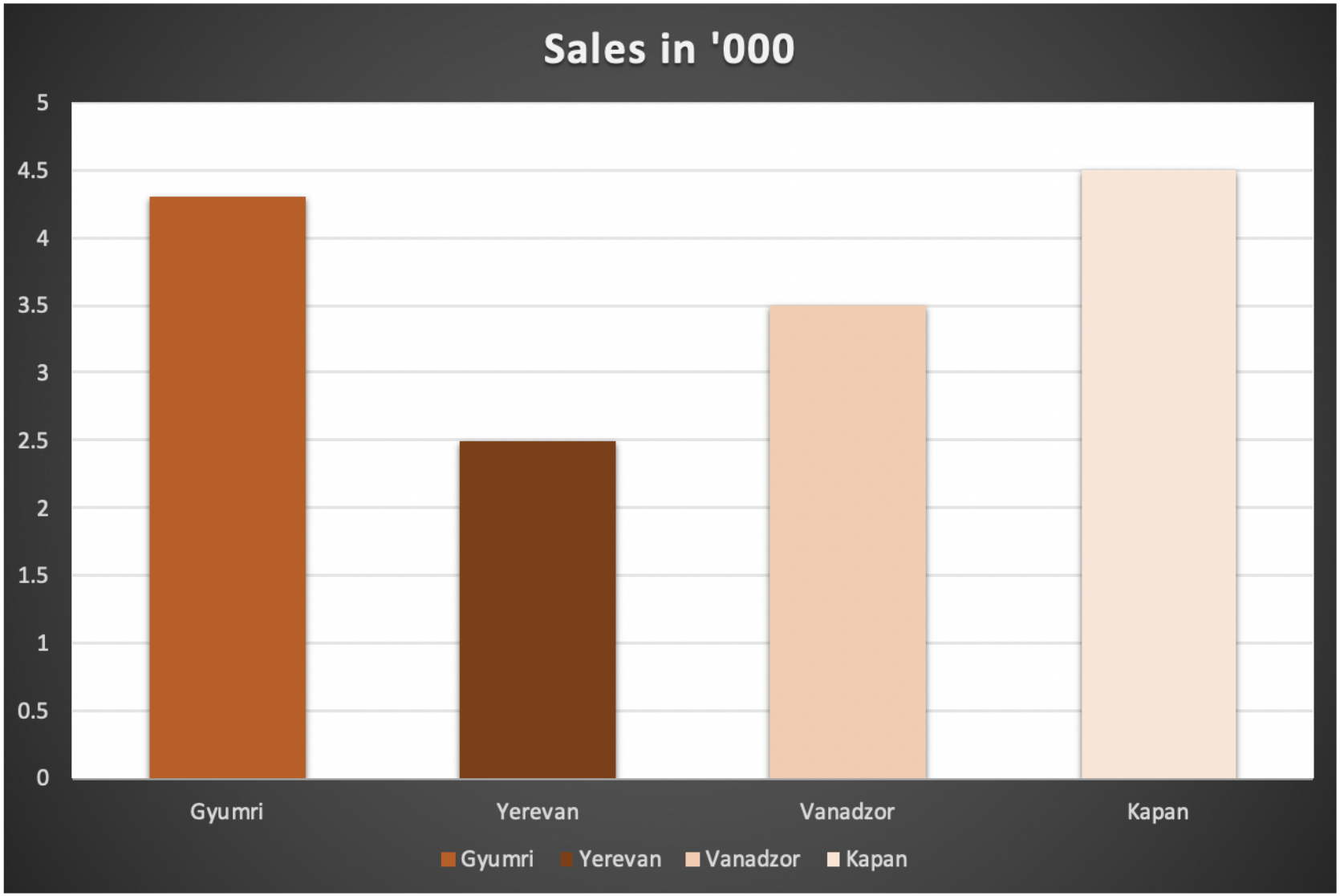
# Chartjunk

Look at the example below, how would you assess data-ink ratio here



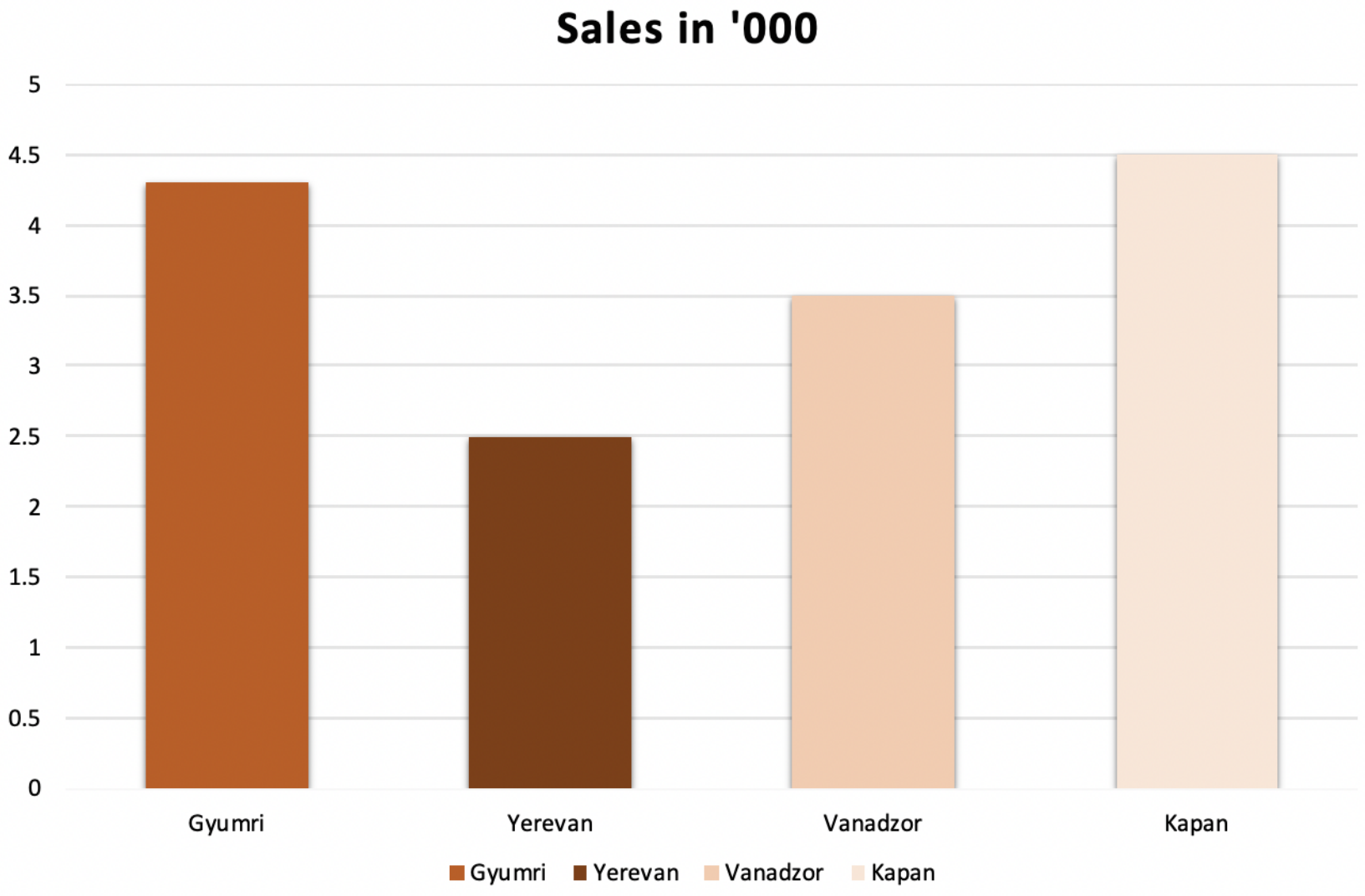
# Chartjunk

Lets remove the junk part by part Barplot background



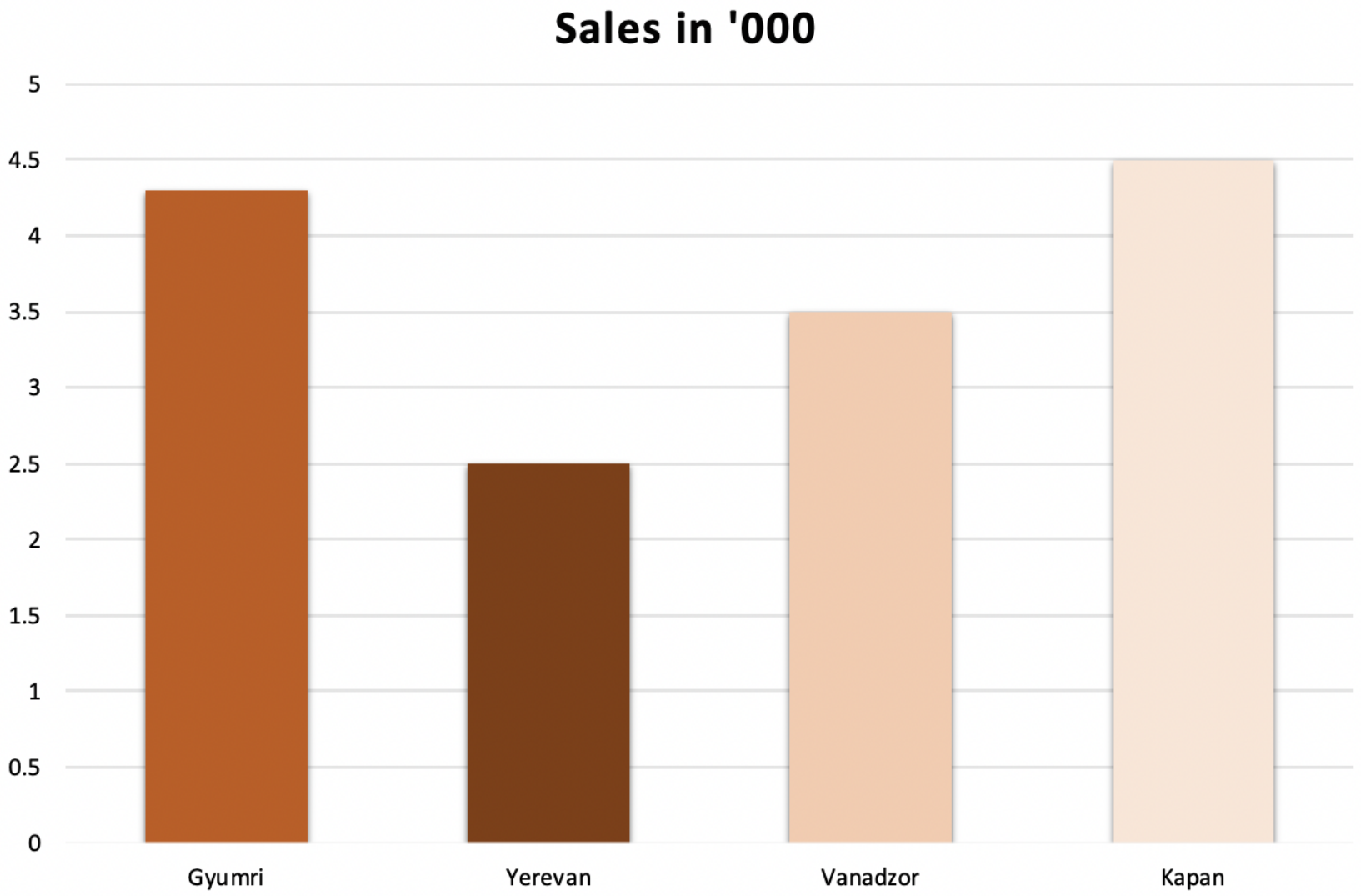
# Chartjunk

Background at all



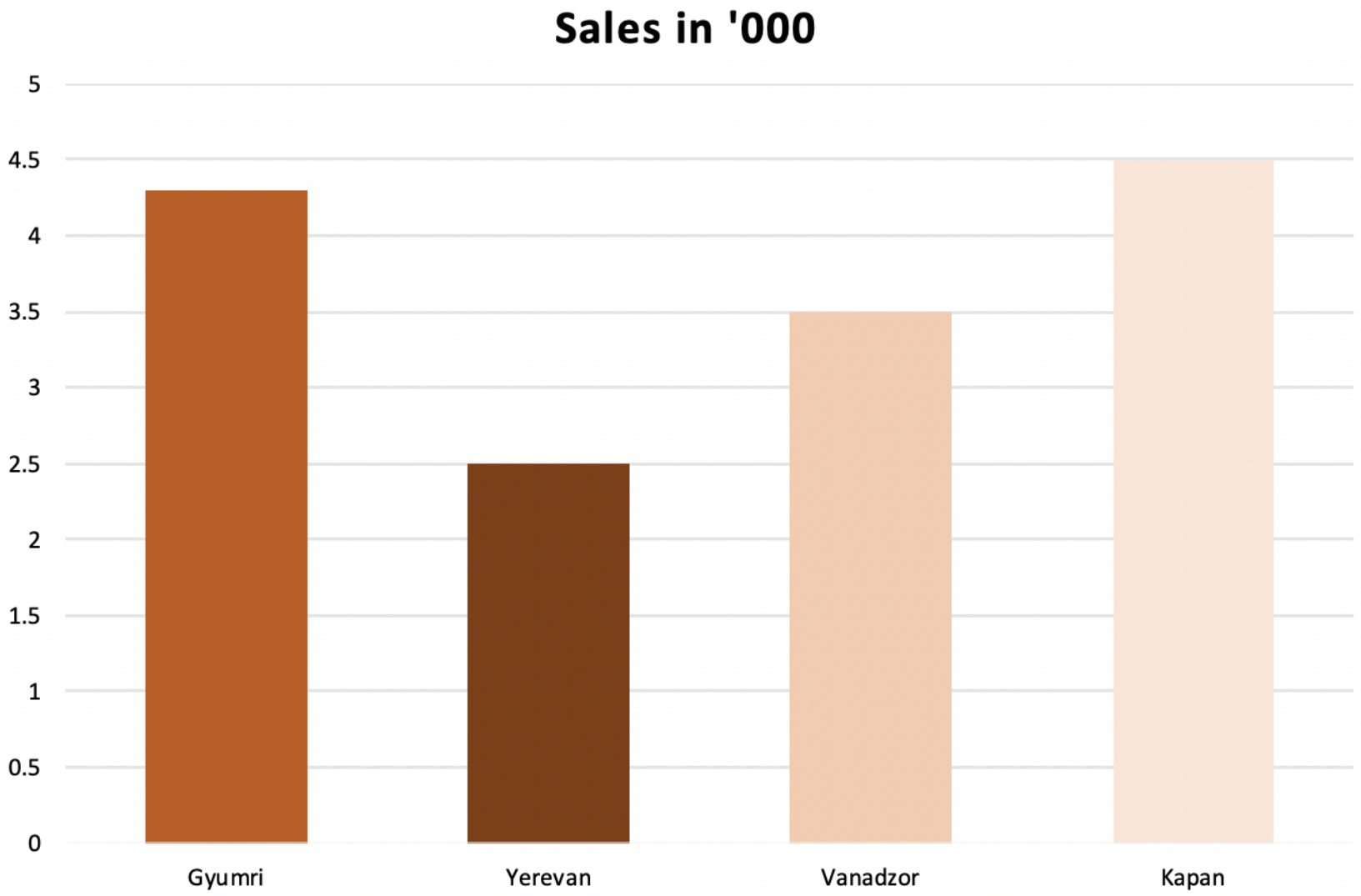
# Chartjunk

Redundant legends



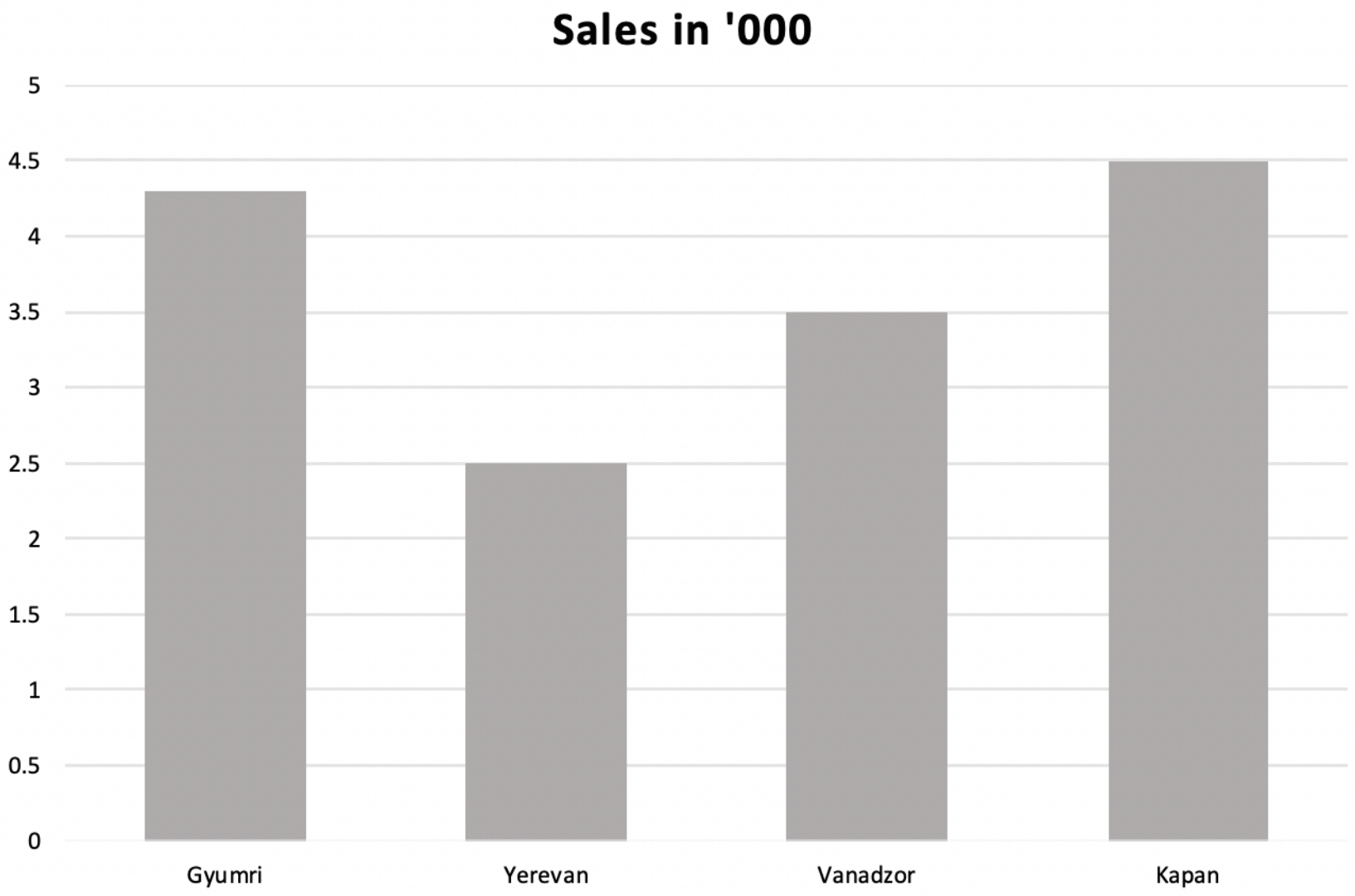
# Chartjunk

Remove shadings



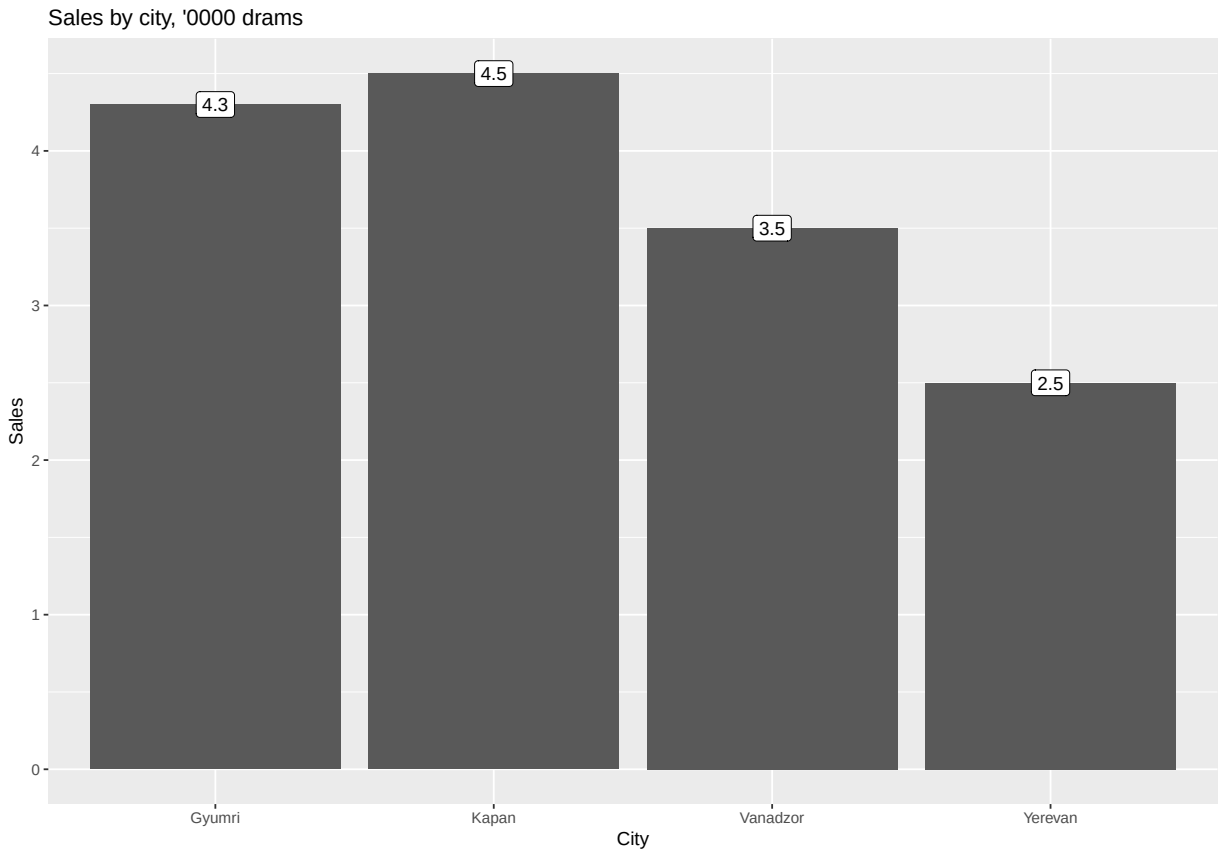
# Chartjunk

Remove colors of the bar



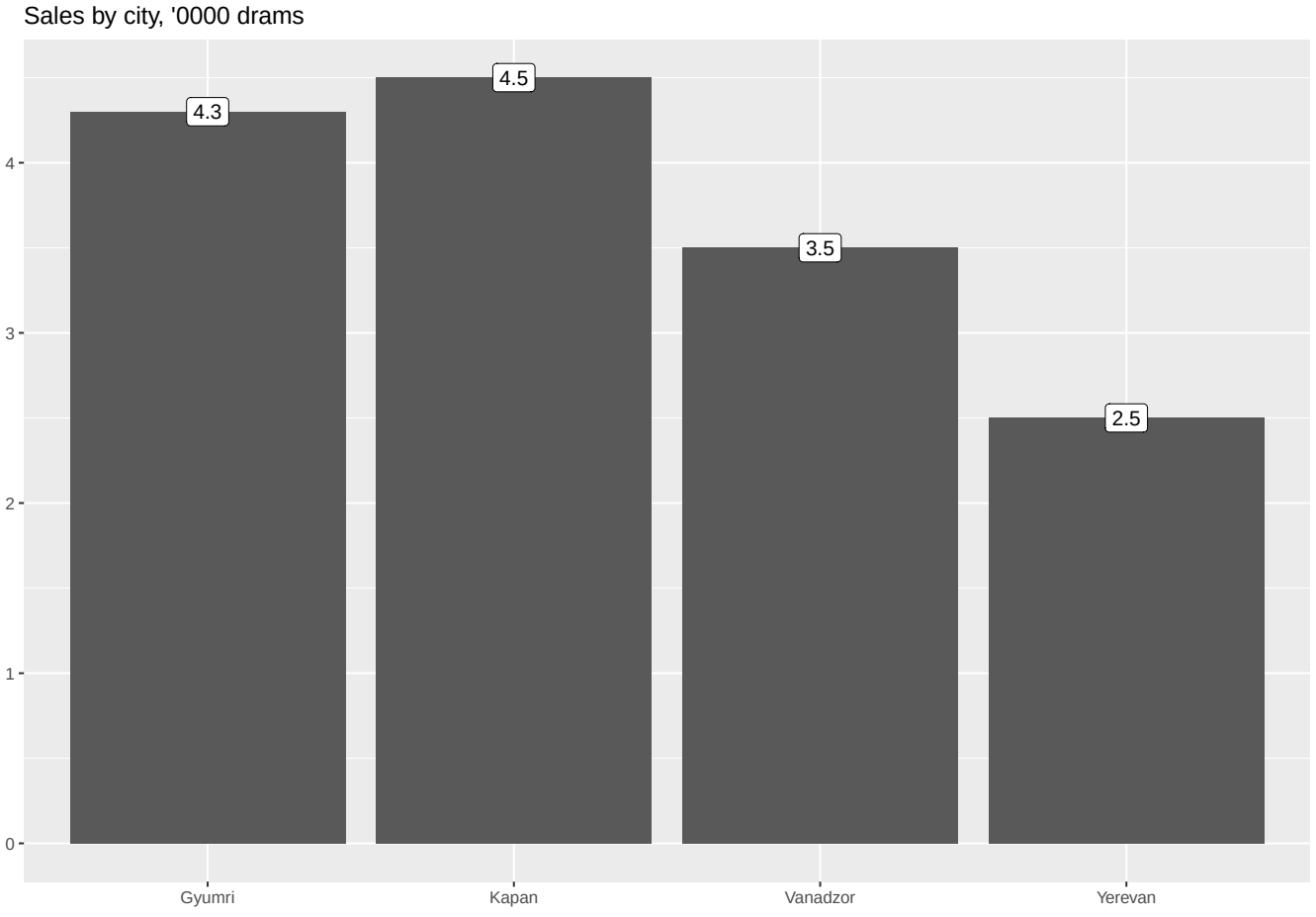
# Chartjunk

Example of the chart, how can you imporve the data-ink ratio ?



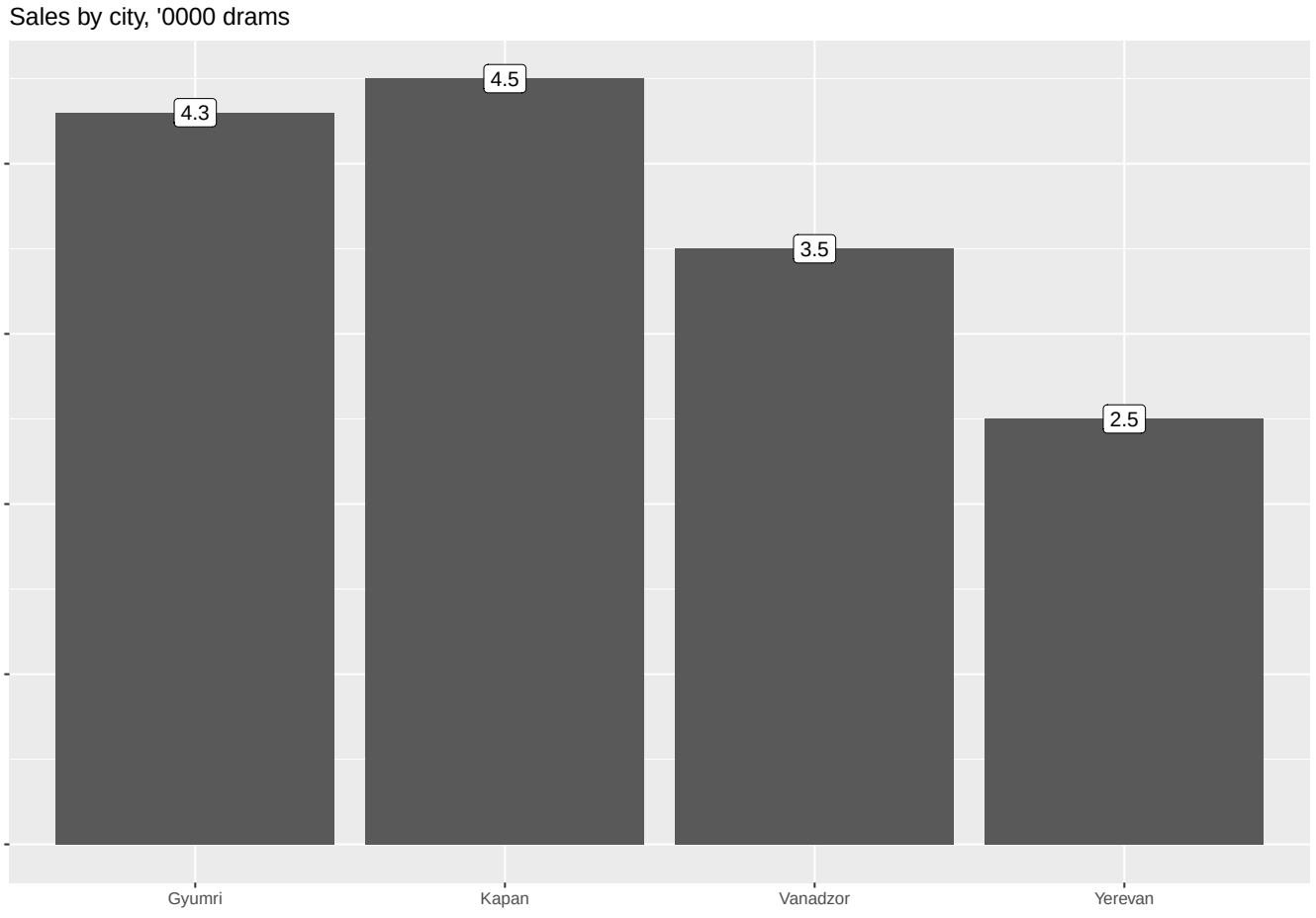
# Chartjunk

What is it that we dont need? -Redundant axis lables



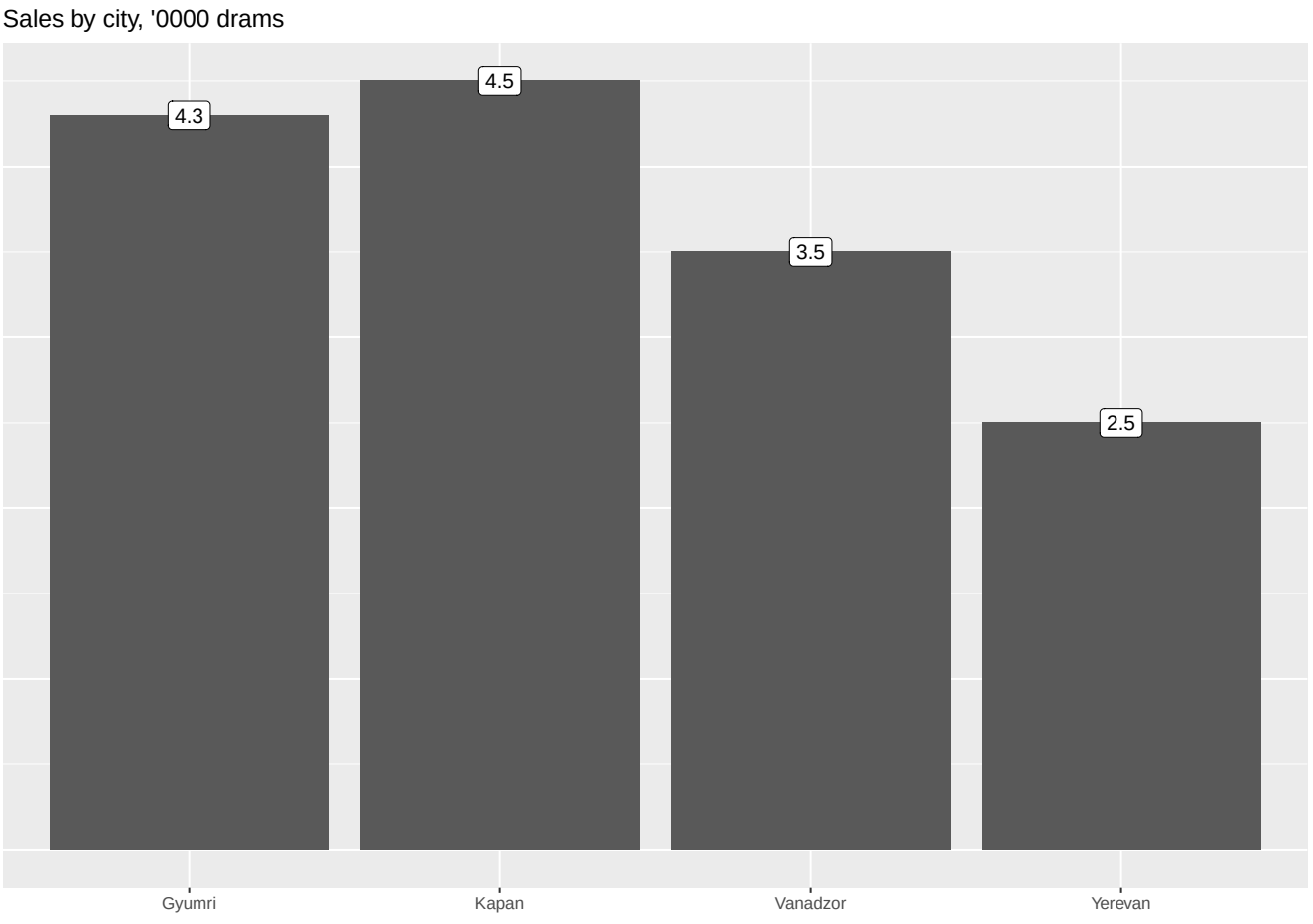
# Chartjunk

## Take out Y axis text



# Chartjunk

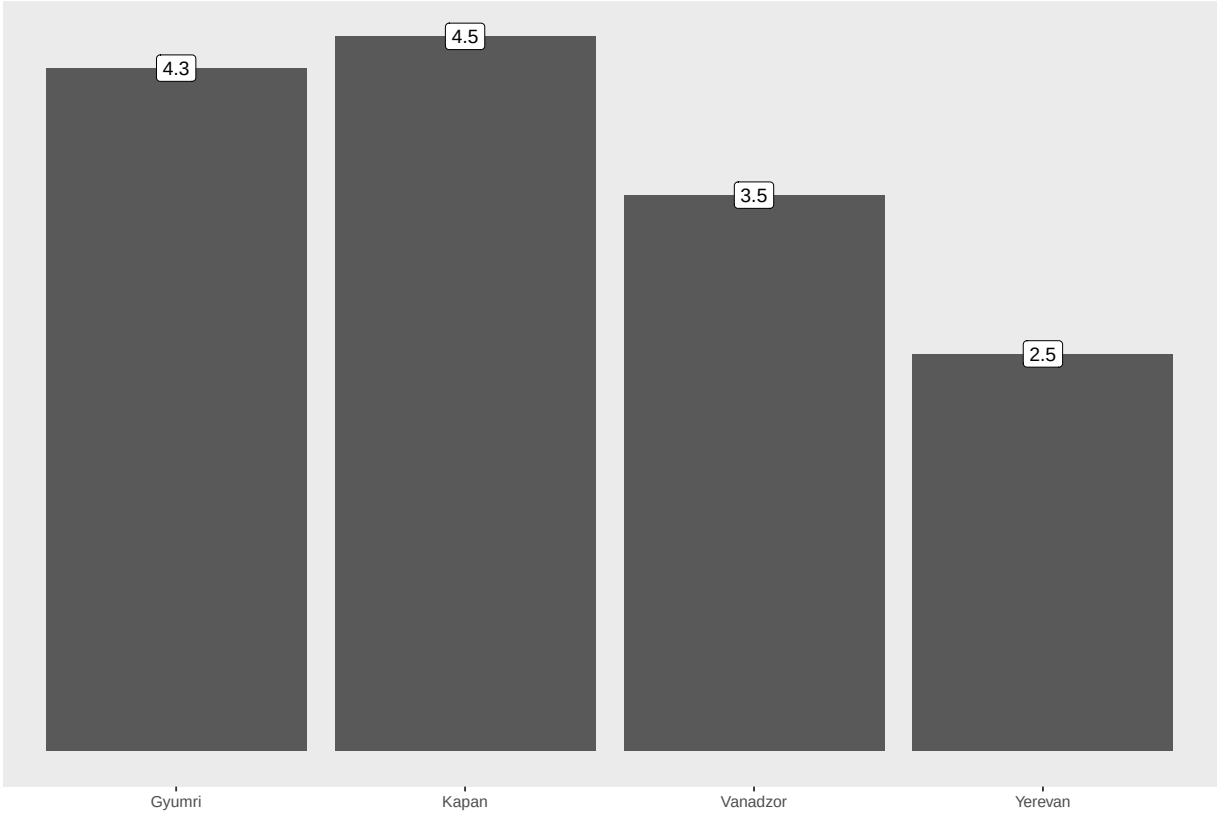
## Ticks add no information to the graph



# Chartjunk

## Do you really need gridlines?

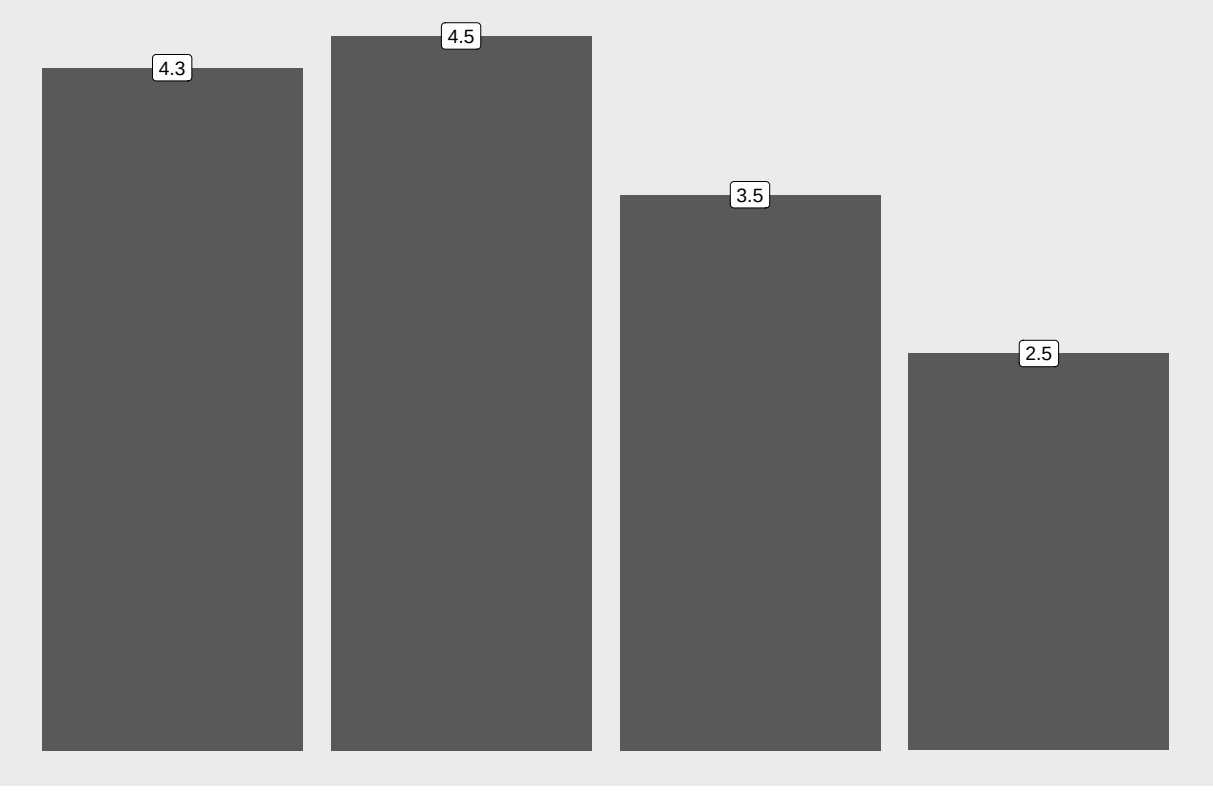
Sales by city, '0000 drams



# Chartjunk

## Tick marks on x axis ?

Sales by city, '0000 drams



# Chartjunk

Do we need the background ?

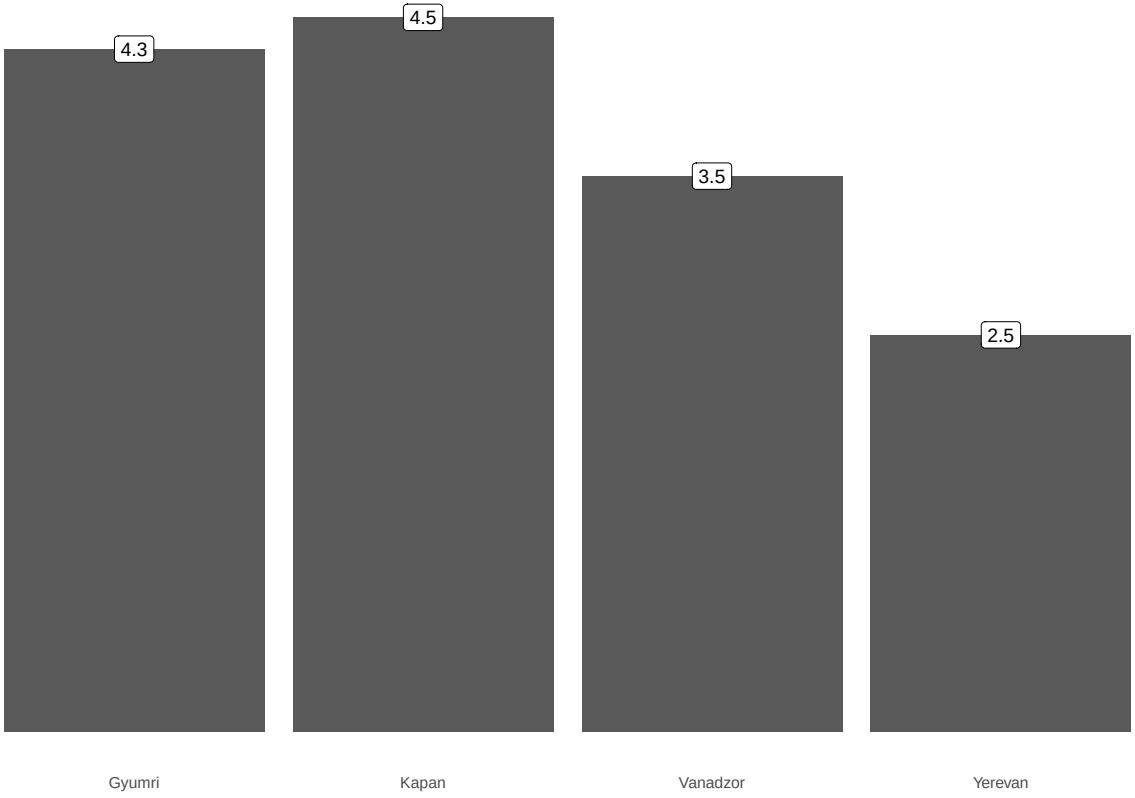
**Hadley Wickham on background color and gridlines**

. . . theme_gray(), uses a very light grey background with white gridlines. This follows from the advice of Tufte (1990, 1997, 2001, 2006) and Brewer (1994a); Carr (1994, 2002); Carr and Sun (1999). We can still see the gridlines to aid in the judgement of position (Cleveland, 1993b), but they have little visual impact and we can easily "tune" them out. The grey background gives the plot a similar colour (in a typographical sense) to the remainder of the text, ensuring that the graphics fit in with the flow of a text without jumping out with a bright white background. Finally, the grey background creates a continuous field of colour which ensures that the plot is perceived as a single visual entity. *H.Whickham ggplot2: Elegant Graphics for Data Analysis*

# Chartjunk

## Overdoing ?

Sales by city, '0000 drams



# Principles of good visualization

According to Edward Tufte (The visual display of Quantitave Information)

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
3. Show data variation, not design variation.
4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units. (We will skip this).
5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
6. Graphics must not quote data out of context.

# Principles of good visualization: Principle 1

Distortion in Data graphic

- The graphic does not distort if the visual representation of the data is consistent with the numerical representation.
- Thus the visual perception of the data (area, height, length) should match the data itself.
- People perceive the length, width, area differently.
- The experiments show that there is a power law relationship between the actual area of the circle and perceived area of the circle.

$$\text{percieved area} = (\text{actual area})^{x}, \text{where } x = 0.8 \pm 0.3$$

# Principles of good visualization: Principle 1

- As an example, the visual perception of length of the line depends on the context and what other people think.

Look at the Solomon Ash Experiment.

https://www.youtube.com/watch?v=iRh5qy09nNw

# Principles of good visualization: Principle 1

Lie factor:

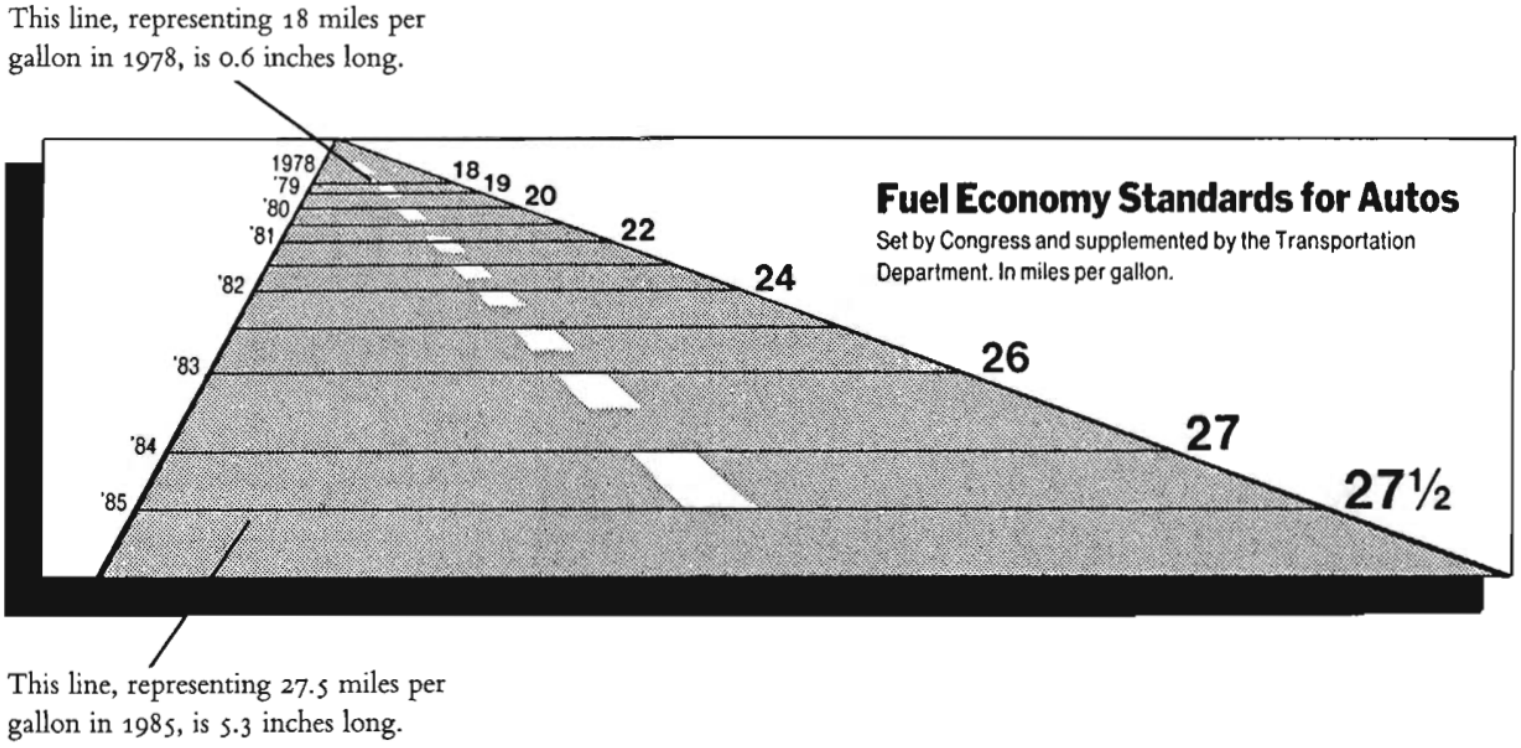
Edward Tufte defines Lie factor in a following way

$$\text{Lie factor} = \frac{\text{Size of effect shown in graph}}{\text{Size of effect in data}}$$

The Lie factor should be equal to 1 with the negligent variation

# Principles of good visualization: Principle 1

Example from the book The visual display of quantitative information



# Principles of good visualization: Principle 1

Size of effect in the data:

- In 1978 the MPG was 18
- In 1985 the MPG was 27.5

$$\text{The size of effect} = \frac{27.5 - 18}{18} * 100 = 53\%$$

# Principles of good visualization: Principle 1

The magnitude of the change from 1978 to 1985 is given with the relative lengths of the lines.

- Length of the line for 1978 - 0.6 inches.
- Length of the line for 1985 - 5.3 inches

$$\text{Size of effect shown} = \frac{5.3 - 0.6}{0.6} * 100 = 783\%$$

# Principles of good visualization: Principle 1

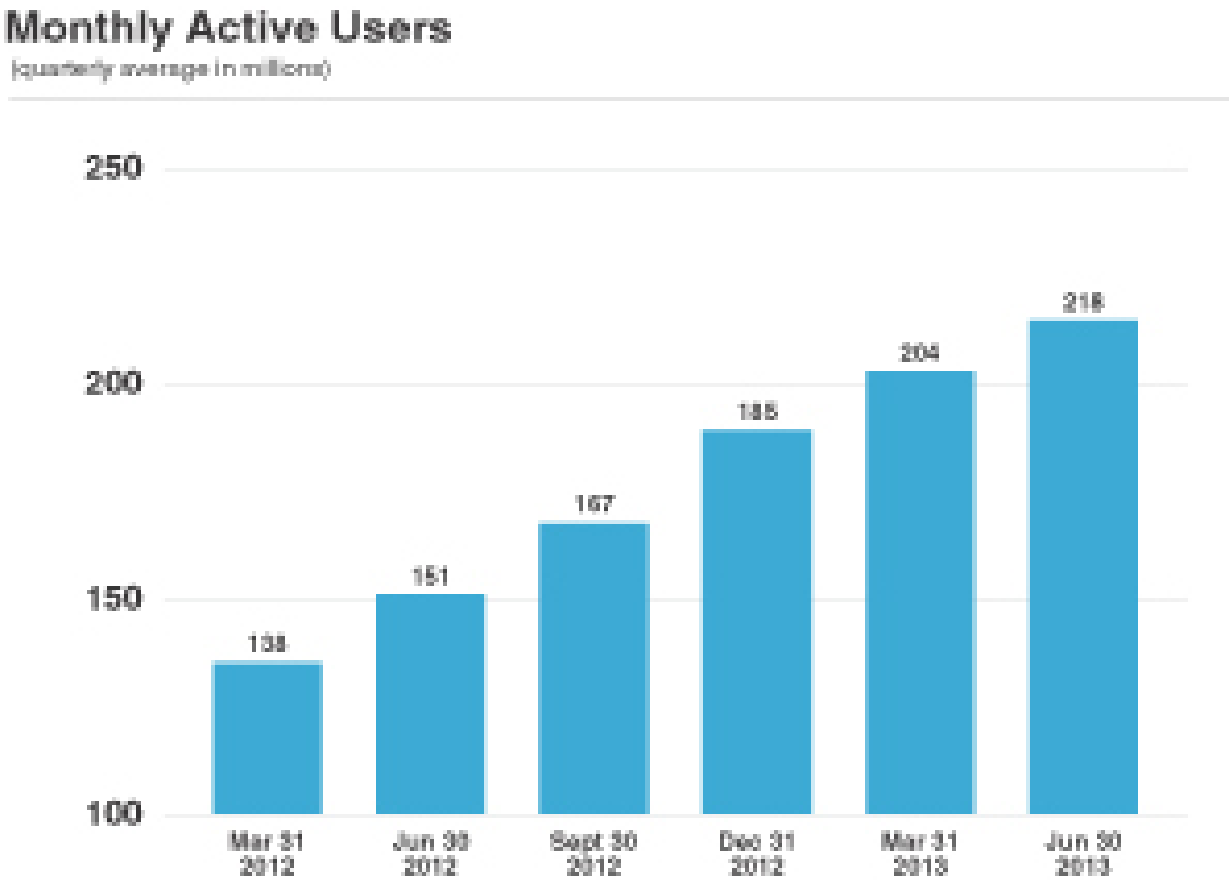
$$\text{Lie Factor} = \frac{783}{53} = 14.7$$

# Principles of good visualization: Principle 1

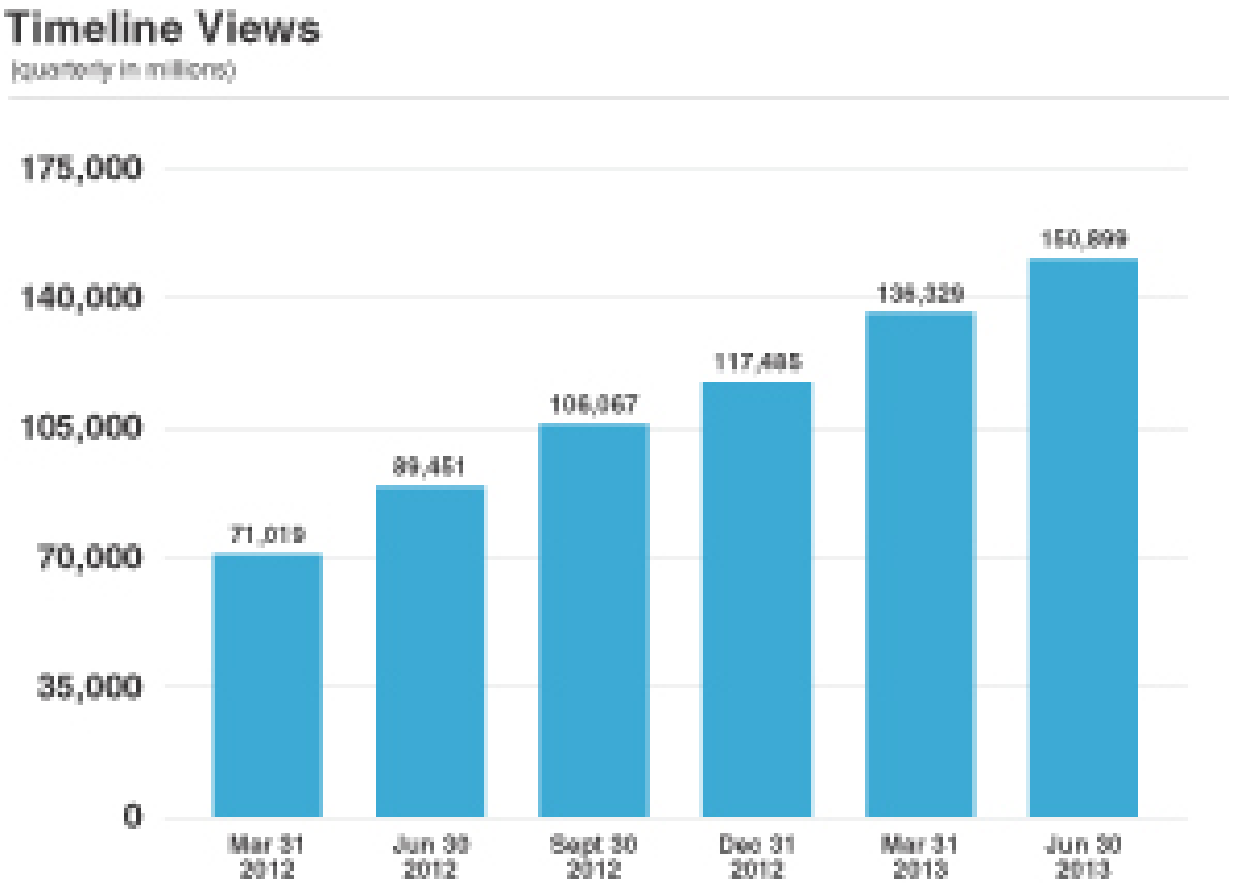
Lie Factor with time series data

In 2013 Twitter filled for IPO. Here are two charts from their SEC filling

# Principles of good visualization: Principle 1



# Principles of good visualization: Principle 1

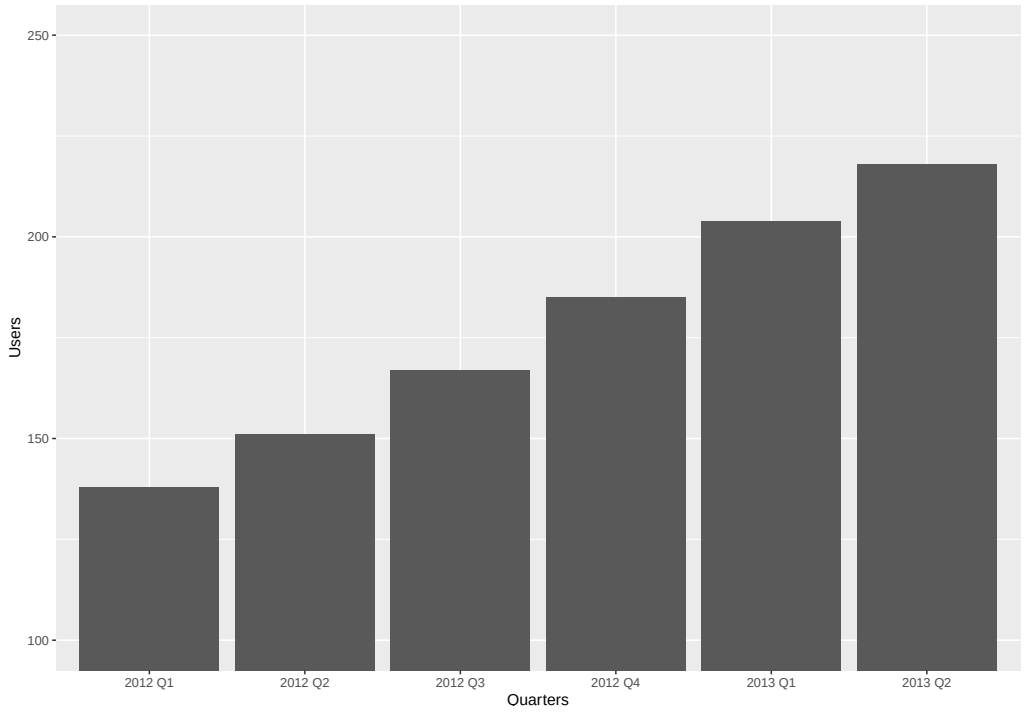


# Principles of good visualization: Principle 1

Recreate the first graph in R

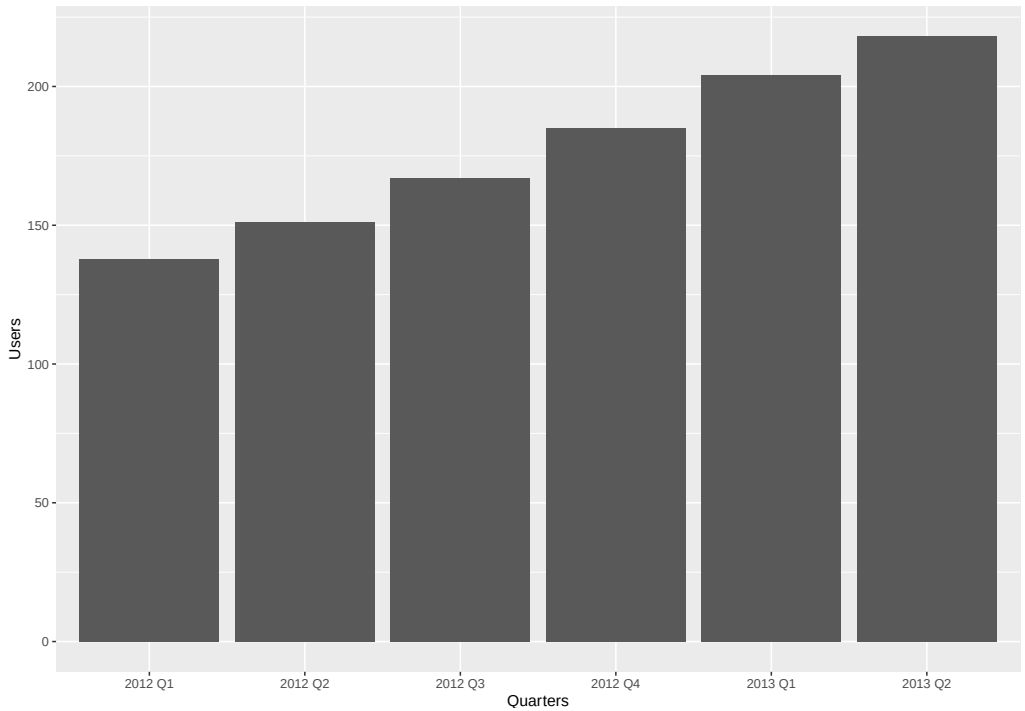
What is the Lie Factor ?

# Principles of good visualization: Principle 1



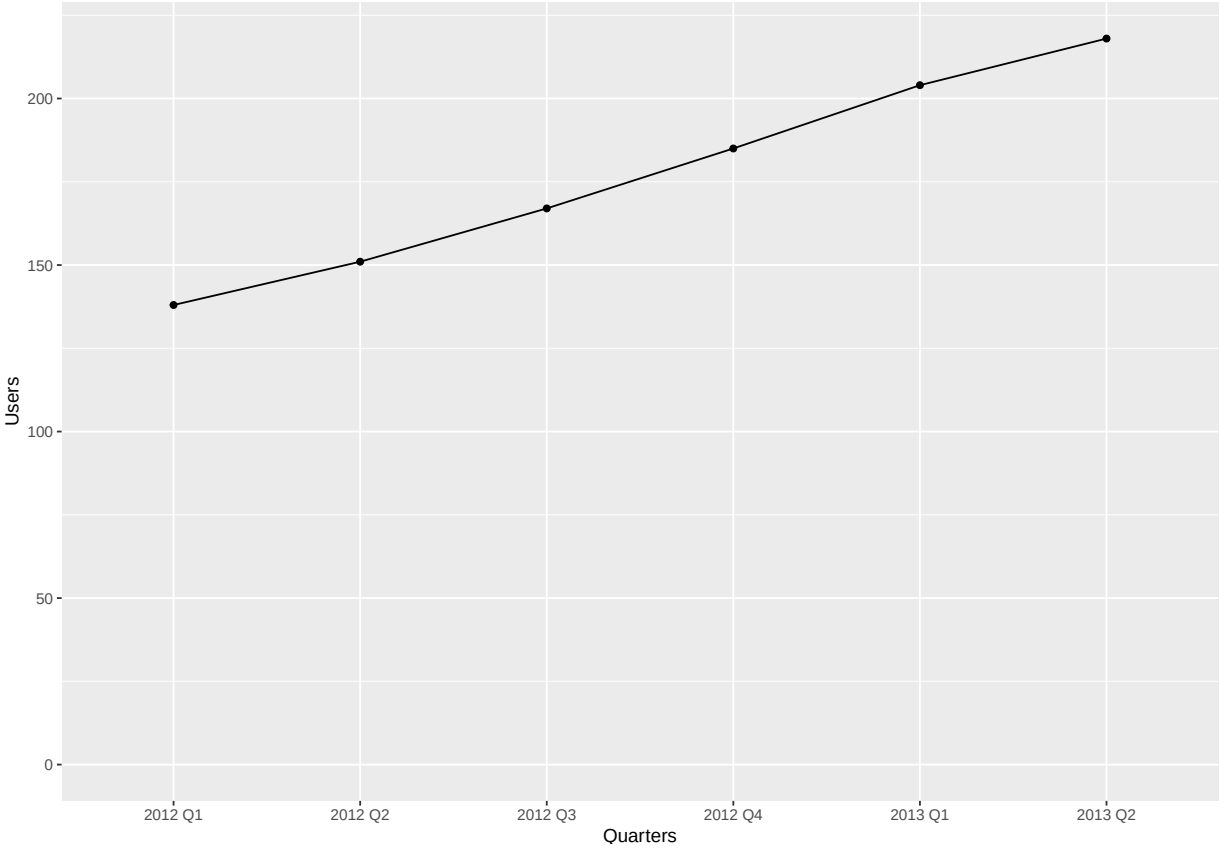
# Principles of good visualization: Principle 1

What if the Y axis starts from 0 ?



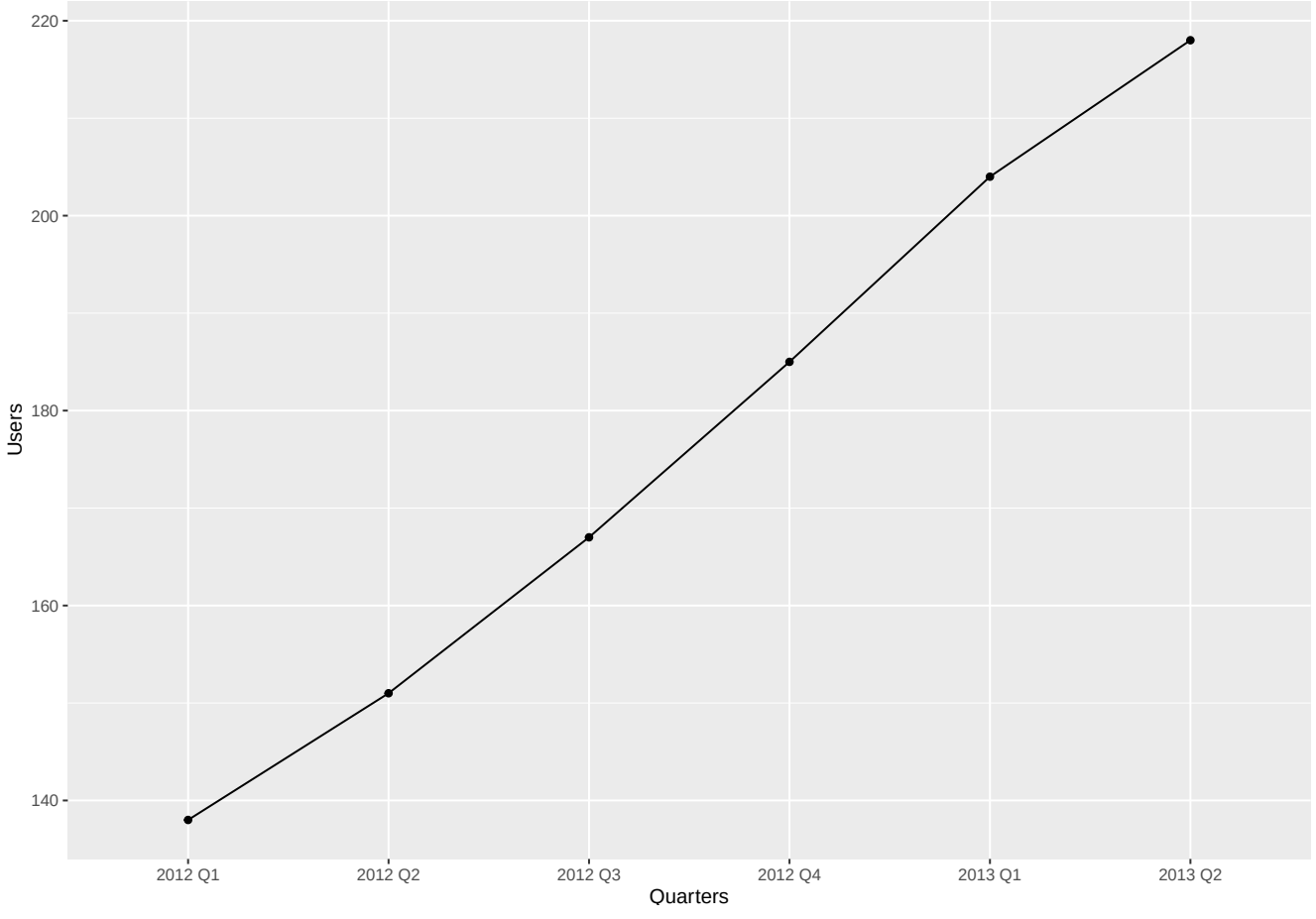
# Principles of good visualization: Principle 1

Use line charts



# Principles of good visualization: Principle 1

According to Tufte, you can have origin different from 0, when you have line chart. Why ?

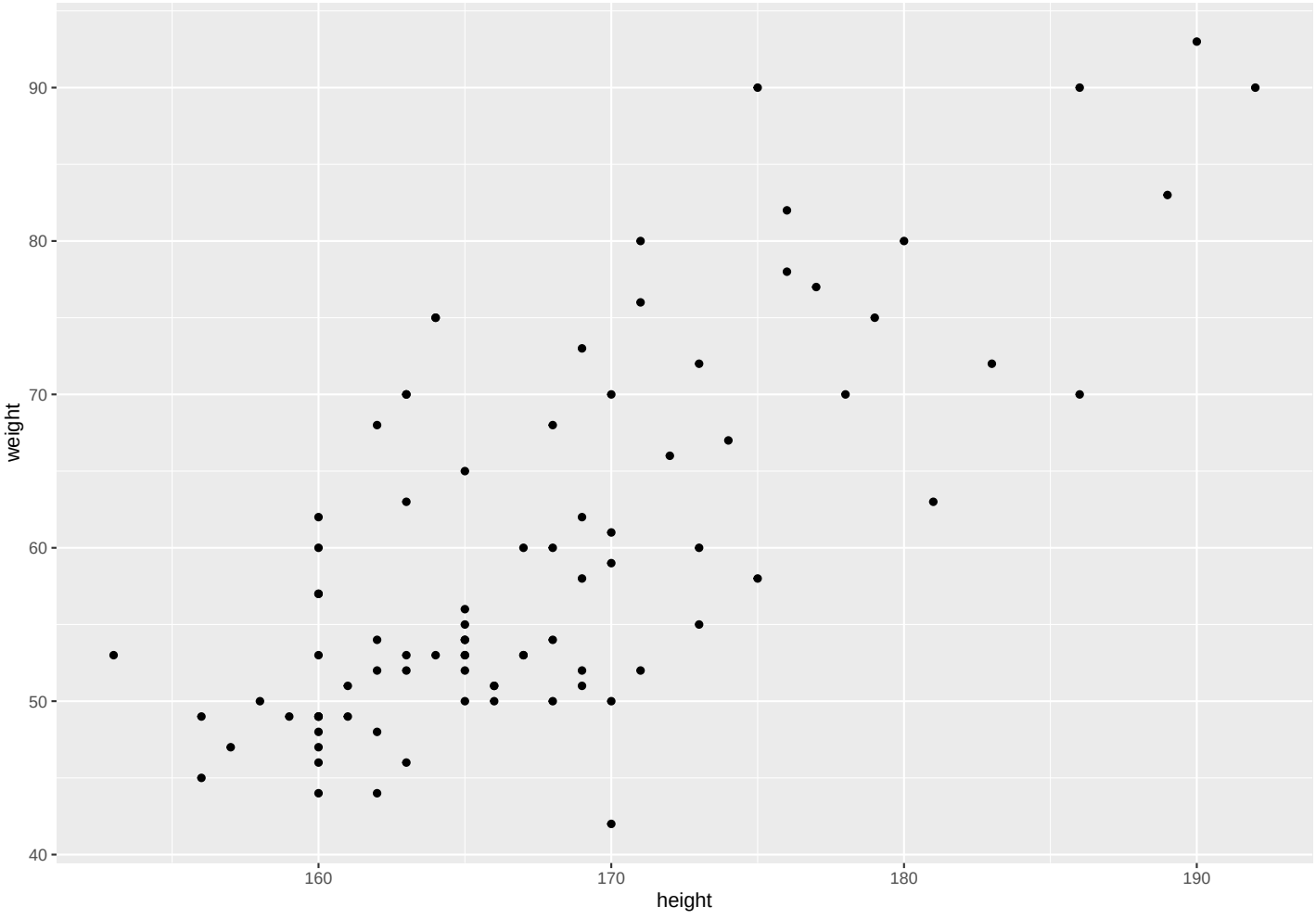


# Principles of good visualization: Principle 2

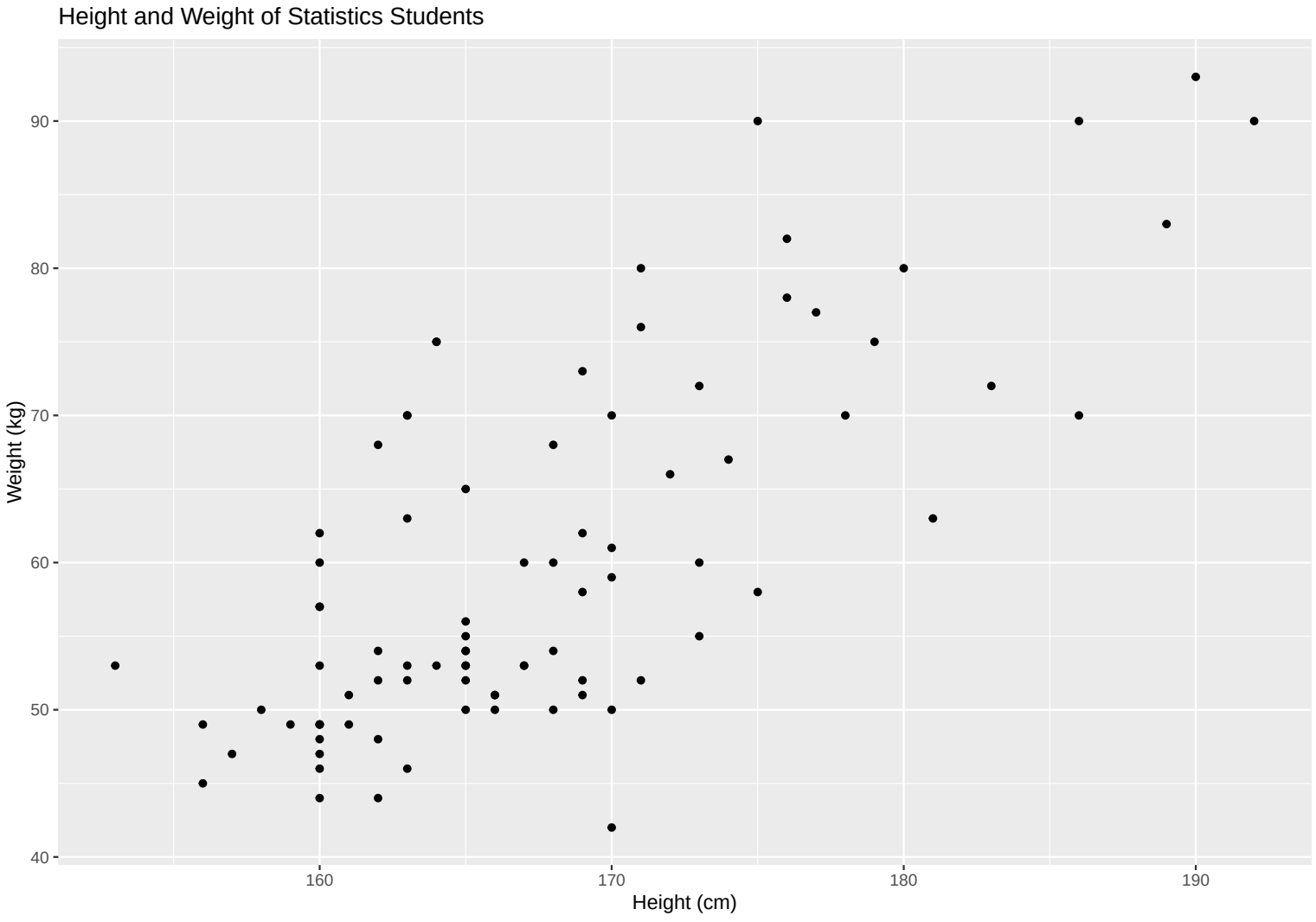
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data. Your graphs need to have meaningful axis labels, title etc. No redundancies in the information. If it is on the graph it should be explained.

# Principles of good visualization: Principle 2

Height and weight data

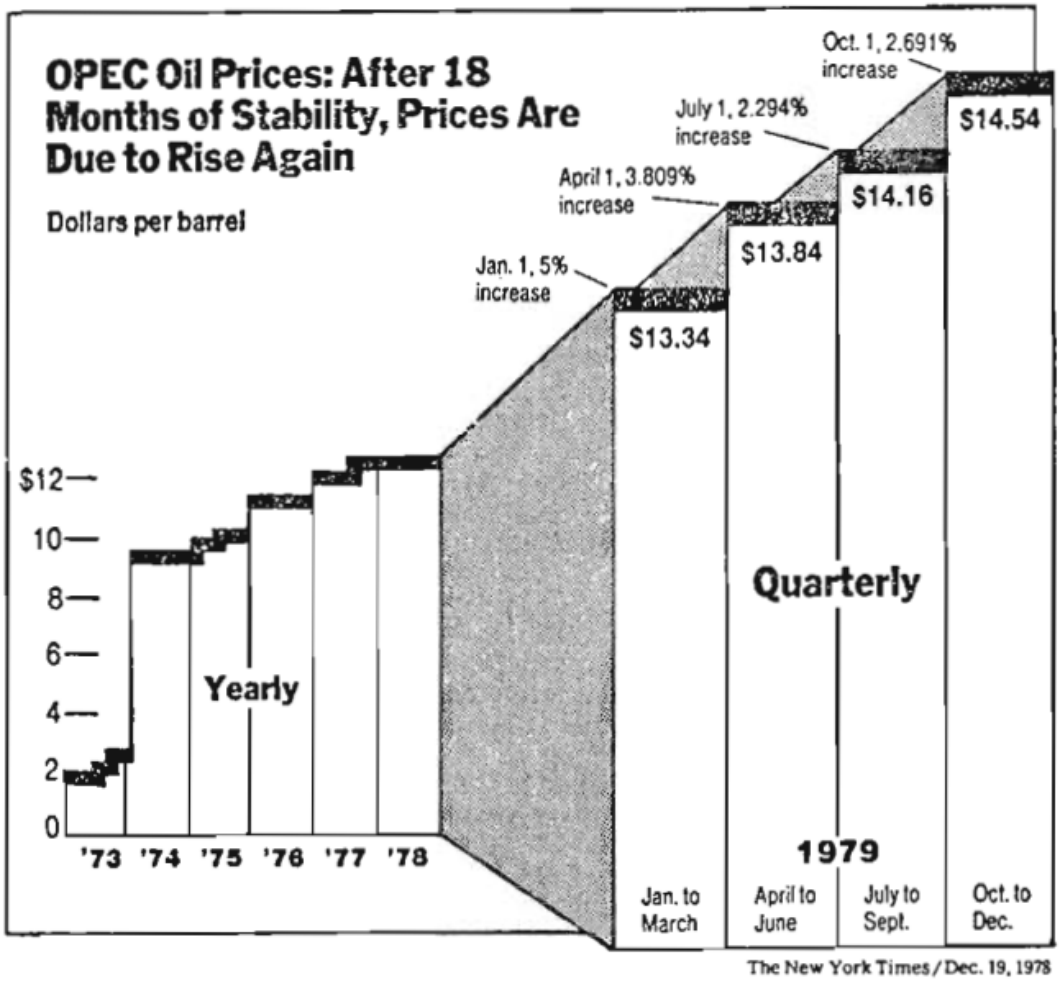


# Principles of good visualization: Principle 2



# Principles of good visualization: Principle 3

Show data variation, not design variation.



# Principles of good visualization: Principle 3

Lies and deceptions

Five different vertical scales show the price:

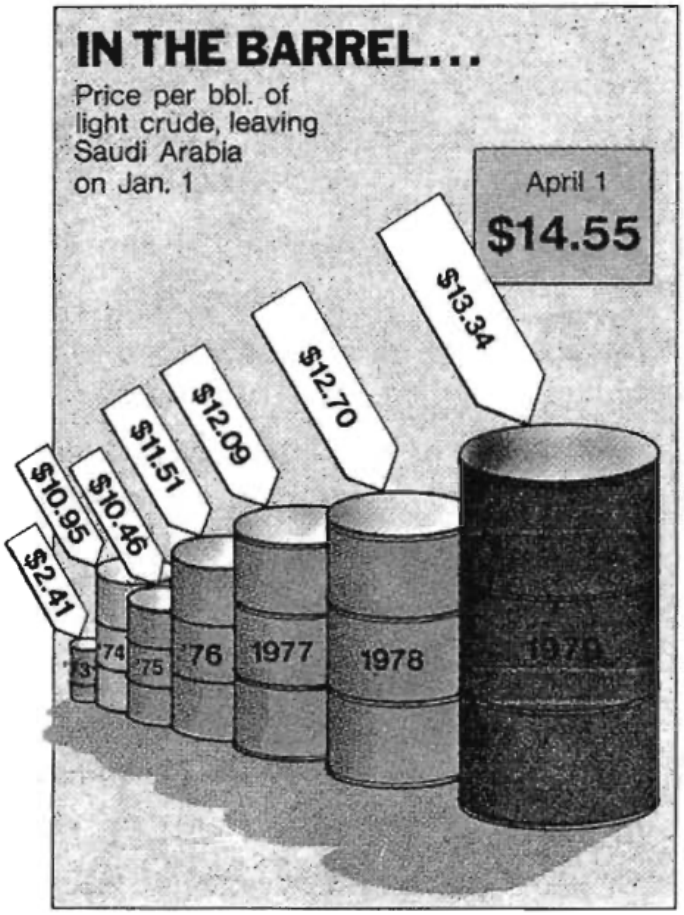
| During this time | one vertical inch equals |
|---|---|
| 1973–1978 | $8.00 |
| January–March 1979 | $4.73 |
| April–June 1979 | $4.37 |
| July–September 1979 | $4.16 |
| October–December 1979 | $3.92 |

And two different horizontal scales show the passage of time:

| During this time | one horizontal inch equals |
|---|---|
| 1973–1978 | 3.8 years |
| 1979 | 0.57 years |

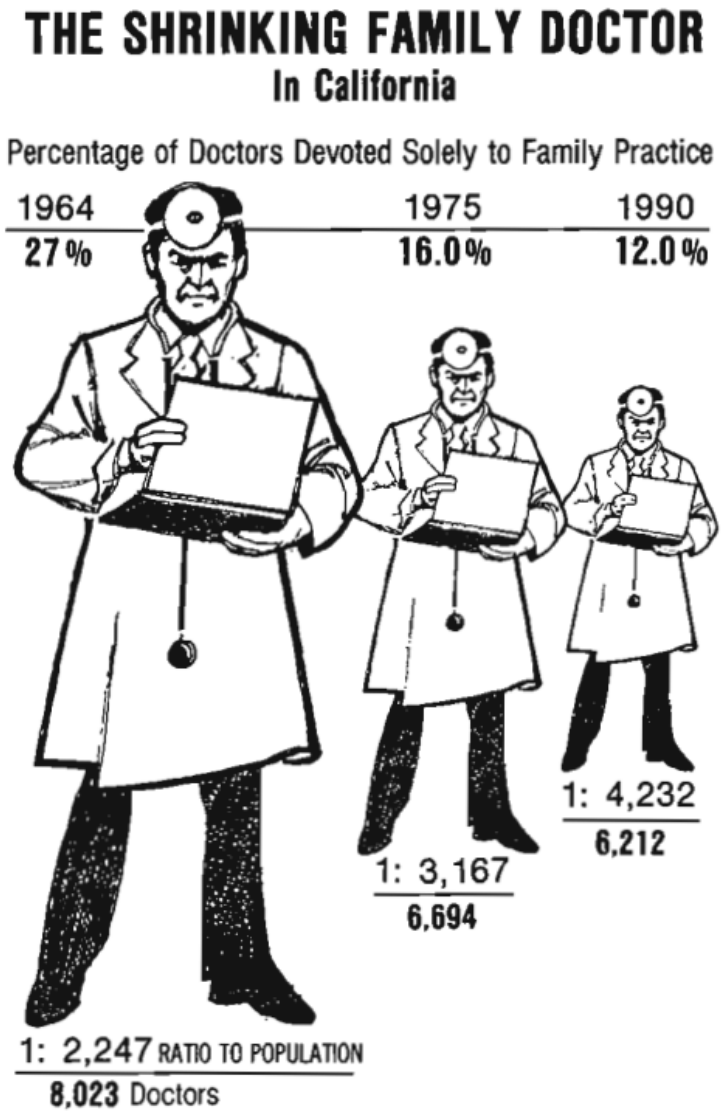
# Principles of good visualization: Principle 3

Avoid 3D graphs



# Principles of good visualization: Principle 3

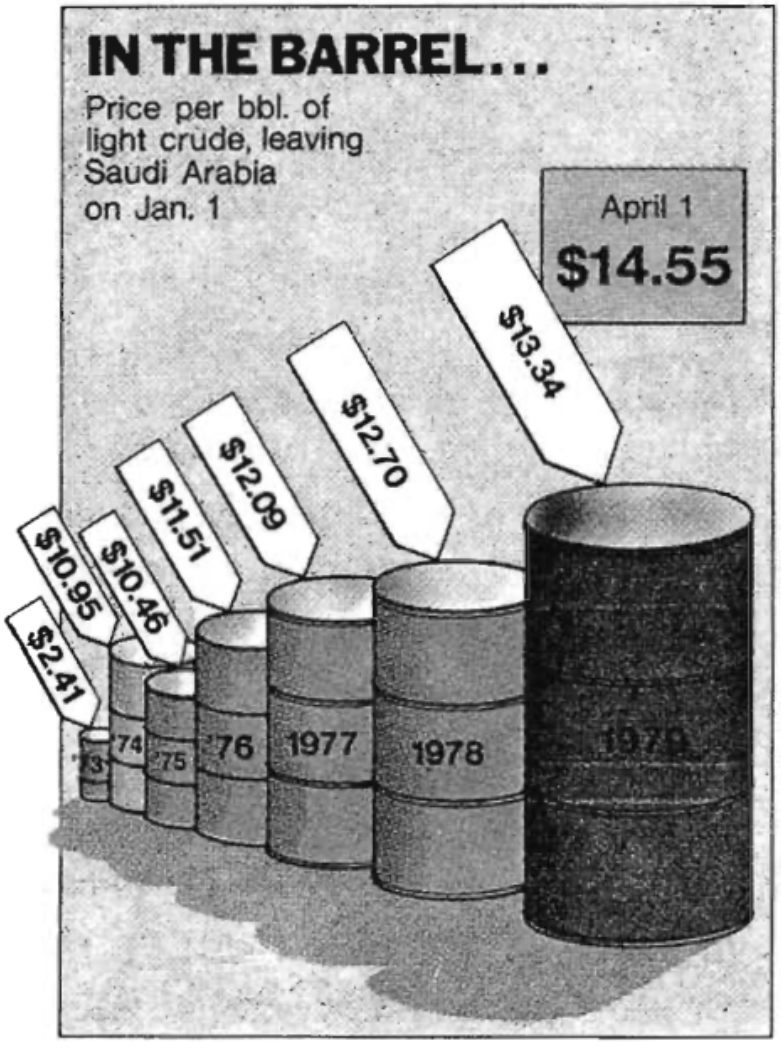
Avoid 3D graphs



# Principles of good visualization: Principle 5

5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

# Principles of good visualization: Principle 5




* If you just take the surface of the barrel than the Lie Factor of the chart will be 9.4
* If you take the volume of the barrel, than the Lie factor will be 59.4 (Tufte, page 71)

# Principles of good visualization: Principle 6

6. Graphics must not quote data out of context.

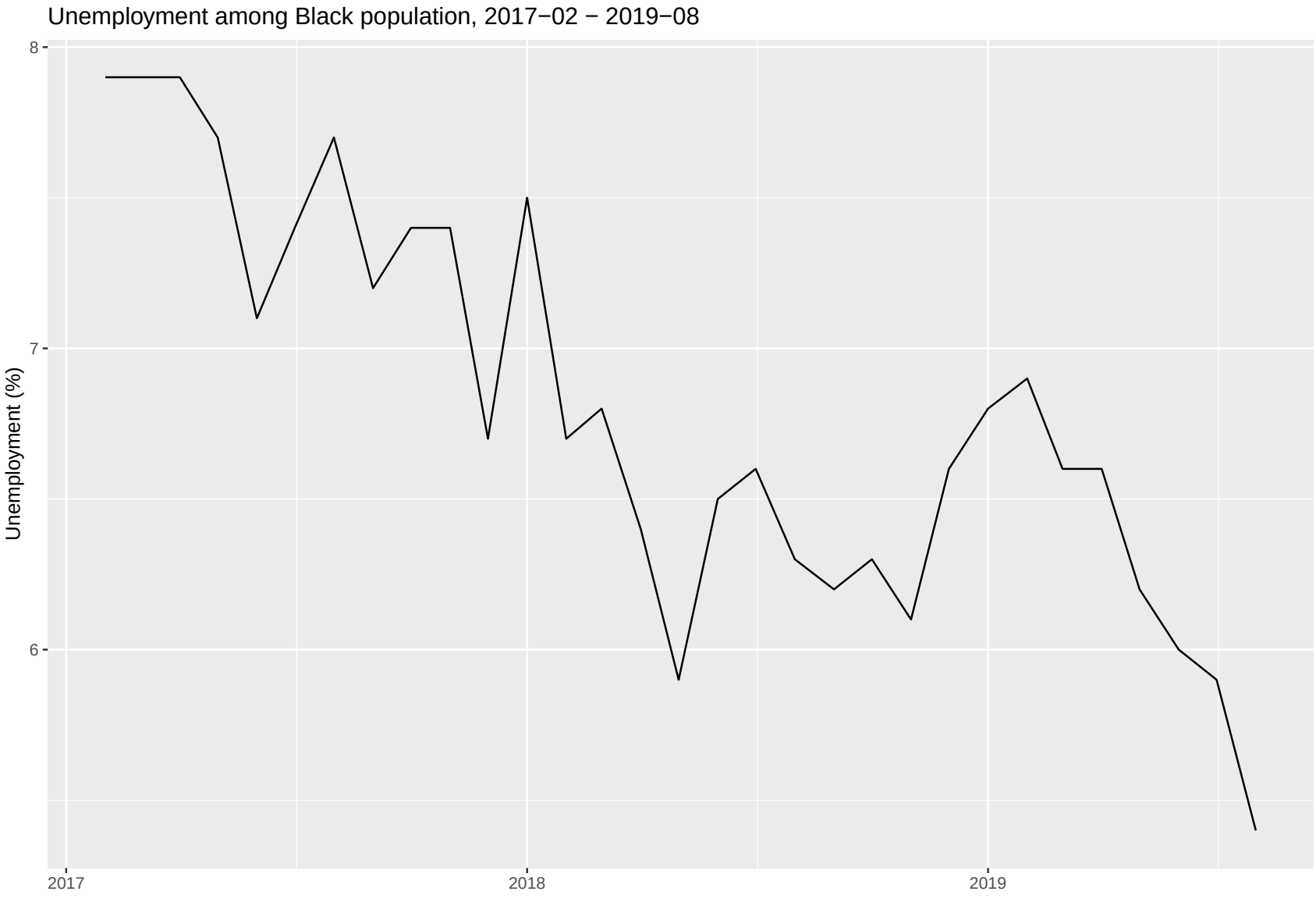
# Principles of good visualization: Principle 6

- In August 2019, the unemployment rate for the Black population hit the historic low - 5.4%.
- Is this the achievement of Trump administration ?

# Principles of good visualization: Principle 6

Trump Period

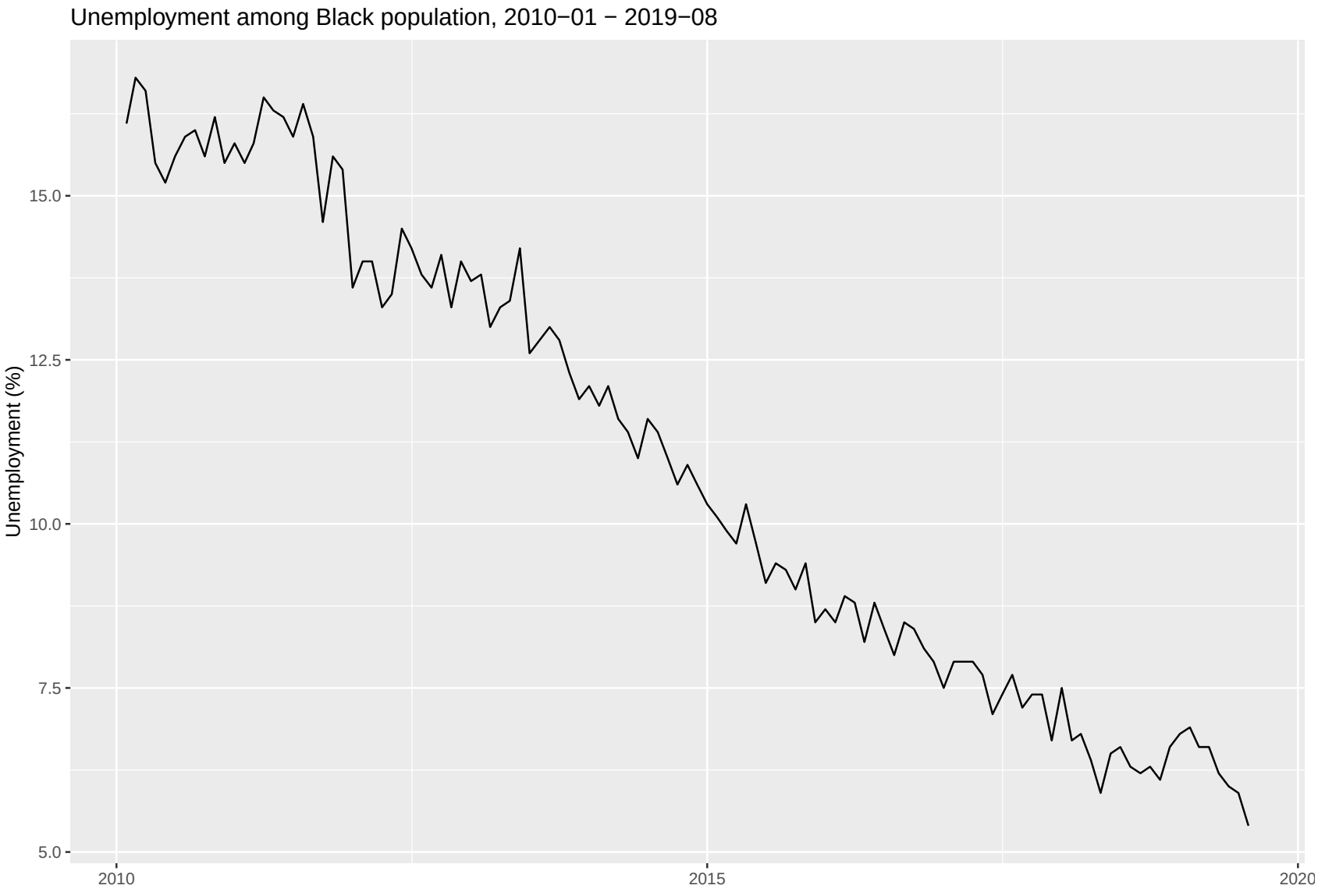
# Principles of good visualization: Principle 6



# Principles of good visualization: Principle 6

Can we say that it is because of Obama's policies to decrease unemployment rate among black population ?

# Principles of good visualization: Principle 6



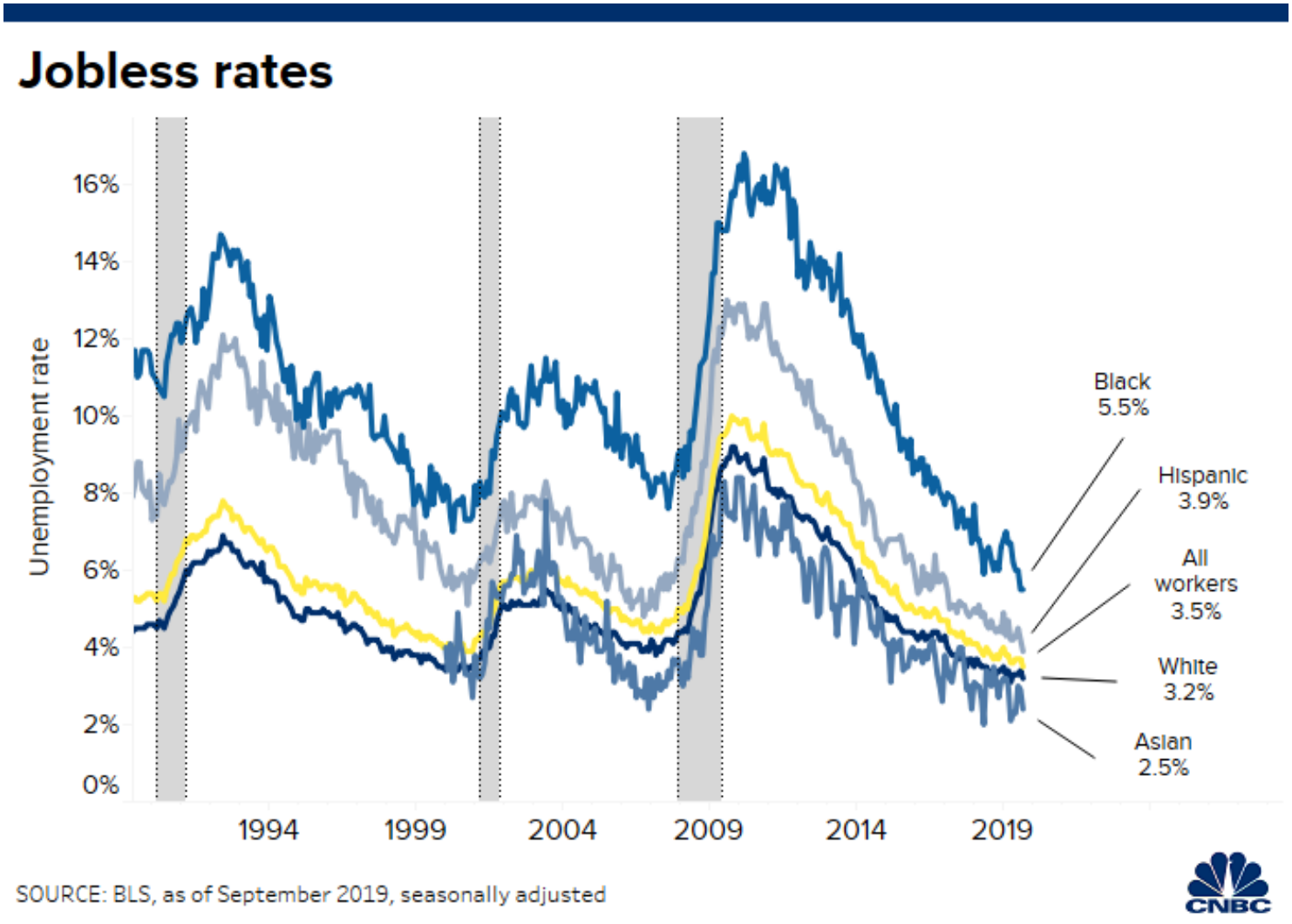
# Principles of good visualization: Principle 6

Unemployment for Black and white populations

# Principles of good visualization: Principle 6

# Principles of good visualization: Principle 6

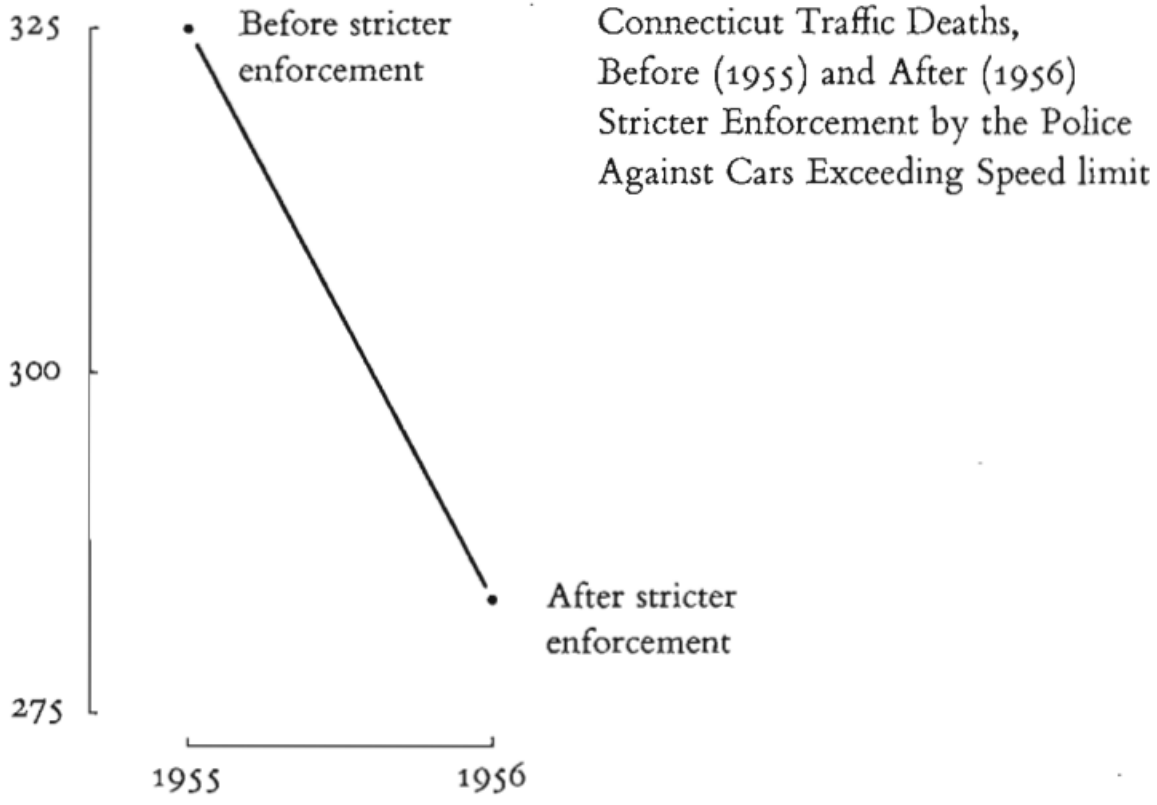
The big picture



# Principles of good visualization: Principle 6

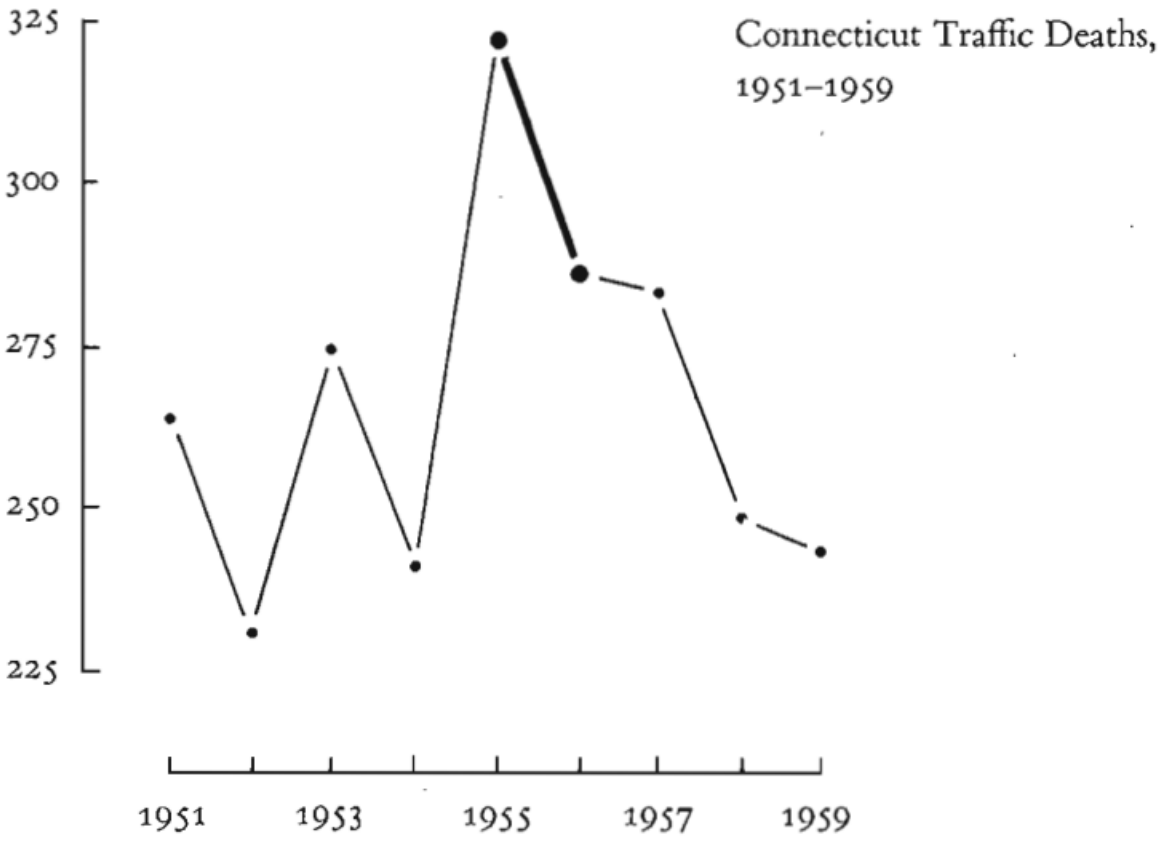
"... To be truthful and revealing, data graphics must bear on the question at the heart of quantitative thinking:"Compared to what?" The emaciated, data-thin design should always provoke suspicion, for graphics often lie by omission, leaving out data sufficient for comparision... "

# Principles of good visualization: Principle 6

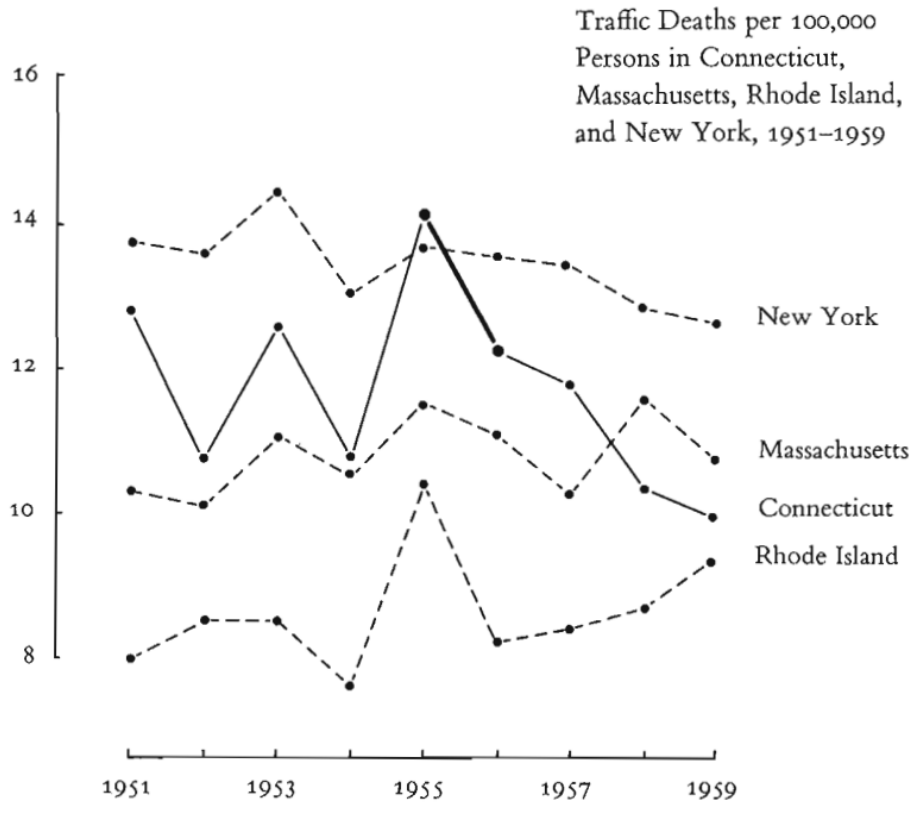


# Principles of good visualization: Principle 6

A few more data points add immensely to the account:

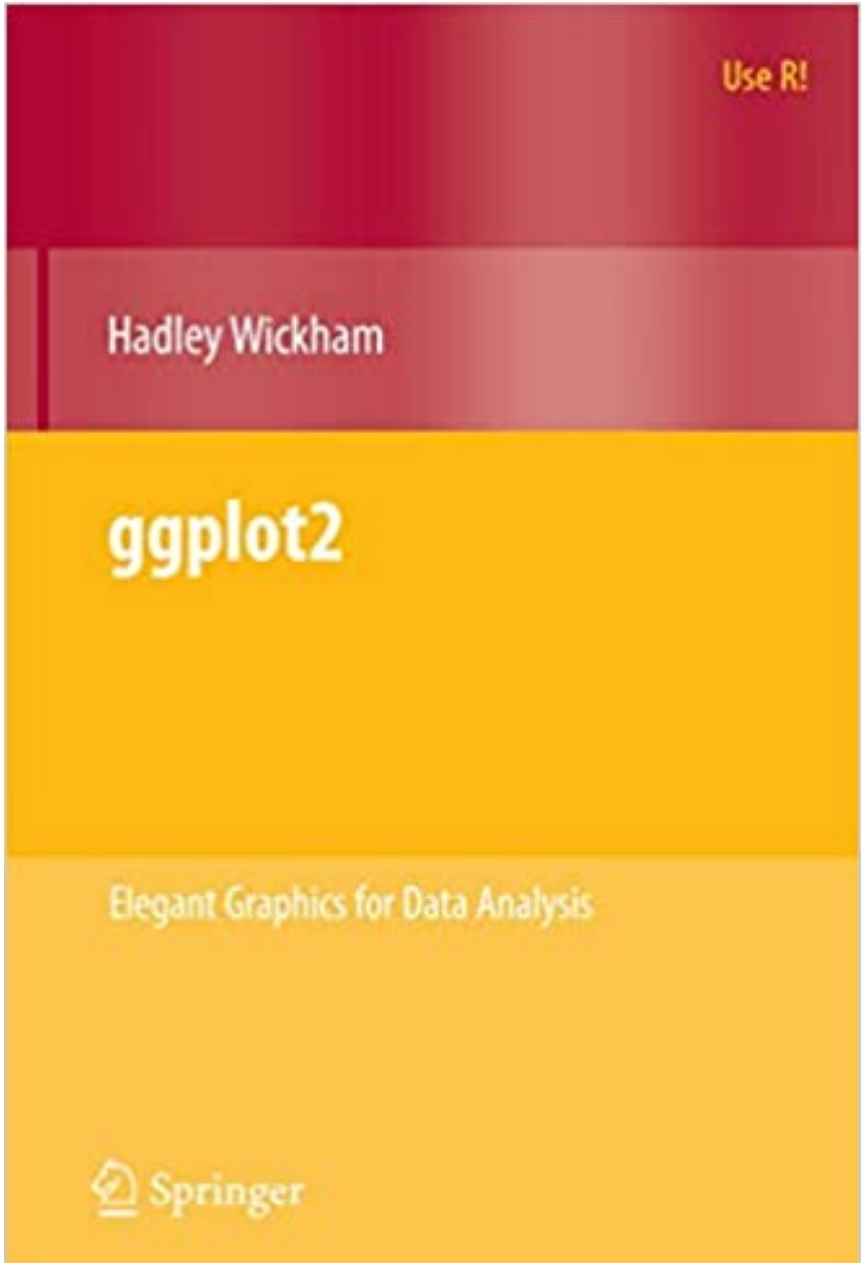


# Principles of good visualization: Principle 6

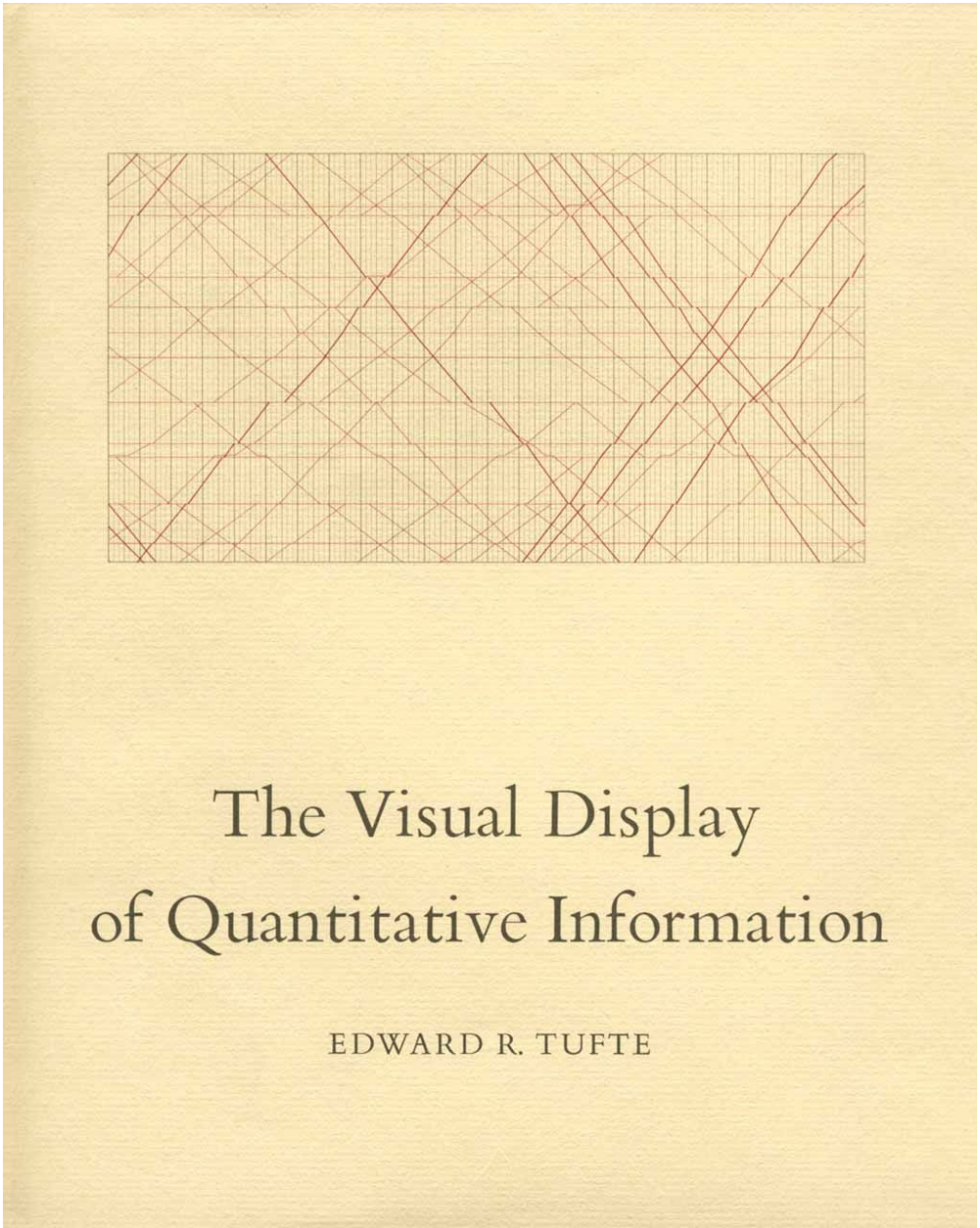



Donald T. Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis," in Edward R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass., 1970), 110–125.

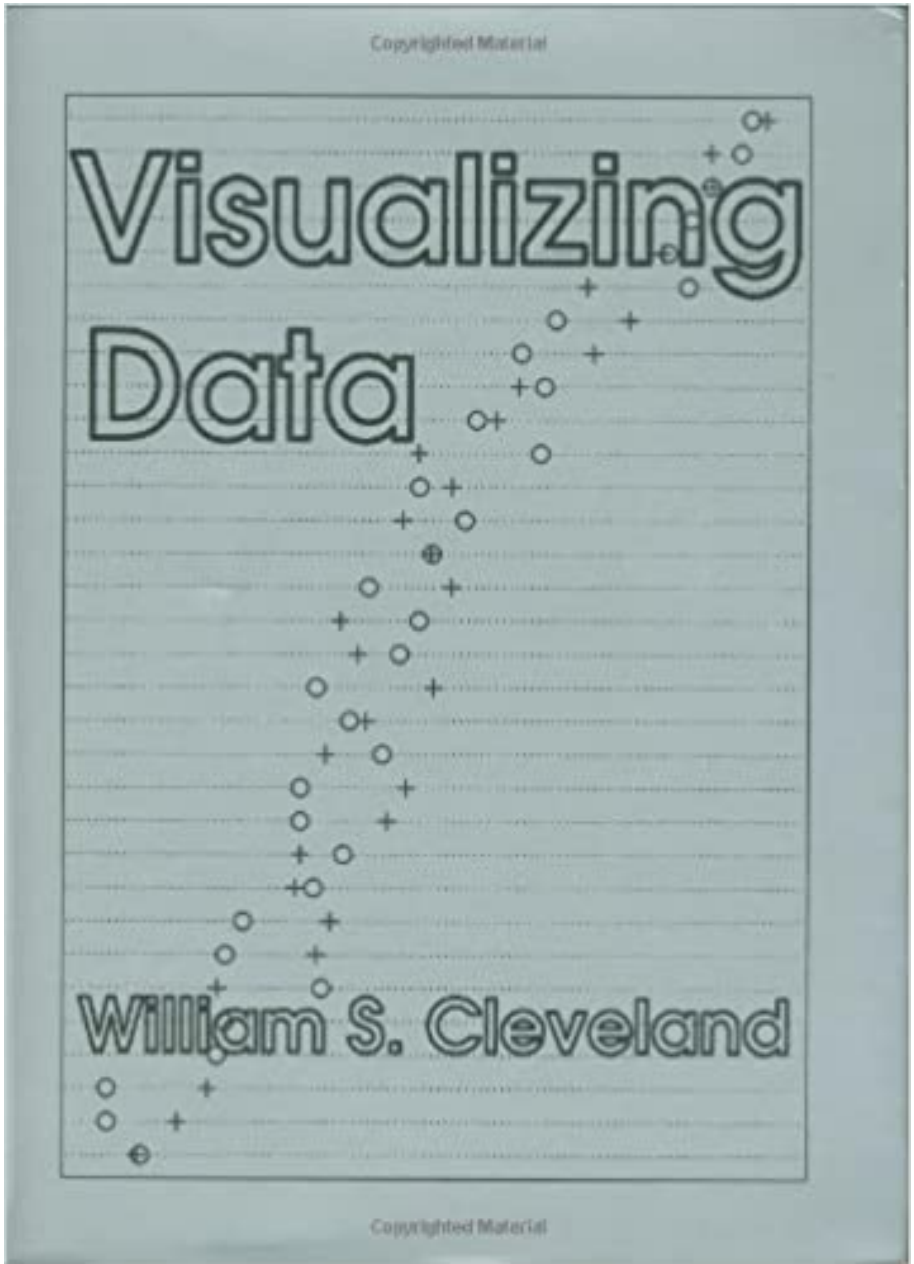
# Books


Use R!
Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Springer

# Books


The Visual Display
of Quantitative Information
EDWARD R. TUFTE

# Books


Copyrighted Material
Visualizing
Data
William S. Cleveland
Copyrighted Material