

Visual Analytics

Giuseppe Santucci

2 – Visualizing numbers - Introduction

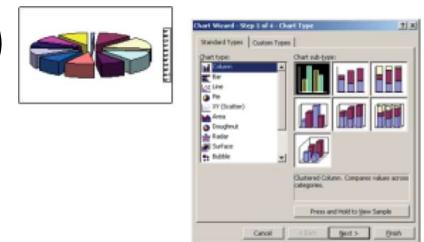
Thanks to Ross Ihaka

Outline

- An introductory example
- Good and bad graphs

Number visualization ?

- Information visualization is, in general, about numbers
- In some cases, however, the numerical values are the main focus of the visualization, and the use of tables, graphs and other visual means to communicate quantitative information is commonplace in business today (pie chart, diagrams, boxplots, scatterplots, etc.)
- Actual software applications allows for easy (?) development of different typologies of charts
- I will discuss the basic relationships and the logical steps that allow for moving from quantitative data to suitable visualizations



A starting example : a lotto game

- Lotto games are played world-wide and many people have theories about how to make money at the game
- **User task ? ---> Money !!!**
- We will examine a particular lotto game, to see whether it might be possible to play it profitably
- The game we'll look at is the daily pick-it lottery run by the state of New Jersey in the USA

Lotto rules

- Each player selects a number between 000 and 999
- A winning number is selected by independently picking three digits between 0 and 9 at random
- All players that hold the winning number split the prize money for the game
- The size of the prize depends on the number of players who choose the winning number

Available data

- The results of the games (winning number and winning amount) are publicly available
- We will use the results of 254 consecutive games to look for a profitable strategy

The data (254 values, 100 shown)

(winning number, winning amount)

- (810, \$190.0), (156, \$120.5), (140, \$285.5), (542, \$184.0), (507, \$384.5),
- (972, \$324.5), (431, \$114.0), (981, \$506.5), (865, \$290.0), (499, \$869.5),
- (020, \$668.5), (123, \$83.0), (356, \$188.0), (015, \$449.0), (011, \$289.5),
- (160, \$212.0), (507, \$466.0), (779, \$548.5), (286, \$260.0), (268, \$300.5),
- (698, \$556.5), (640, \$371.5), (136, \$112.5), (854, \$254.5), (069, \$368.0),
- (199, \$510.0), (413, \$102.0), (192, \$206.5), (602, \$261.5), (987, \$361.0),
- (112, \$167.5), (245, \$187.0), (174, \$146.5), (913, \$205.0), (828, \$348.5),
- (539, \$283.5), (434, \$447.0), (357, \$102.5), (178, \$219.0), (198, \$292.5),
- (406, \$343.0), (079, \$332.5), (034, \$532.5), (089, \$445.5), (257, \$127.0),
- (662, \$557.5), (524, \$203.5), (809, \$373.5), (527, \$142.0), (257, \$230.5),
- (008, \$482.5), (446, \$512.5), (440, \$330.0), (781, \$273.0), (615, \$171.0),
- (231, \$178.0), (580, \$463.5), (987, \$476.0), (391, \$290.0), (267, \$176.0),
- (808, \$195.0), (258, \$159.5), (479, \$296.0), (516, \$177.5), (964, \$406.0),
- (742, \$182.0), (537, \$164.5), (275, \$137.0), (112, \$191.0), (230, \$298.0),
- (310, \$110.0), (335, \$353.0), (238, \$192.5), (294, \$308.5), (854, \$287.0),
- (309, \$203.5), (026, \$377.5), (960, \$211.5), (200, \$342.0), (604, \$259.0),
- (841, \$231.0), (659, \$348.0), (735, \$159.0), (105, \$130.5), (254, \$176.0),
- (117, \$128.5), (751, \$159.0), (781, \$290.0), (937, \$335.0), (020, \$514.0),
- (348, \$191.0), (653, \$304.5), (410, \$167.0), (468, \$257.0), (077, \$640.0),
- (921, \$142.0), (314, \$146.0), (683, \$356.0), (000, \$96.0), (963, \$295.0),

Available data

- Does this data contain information which will enable us to choose a profitable strategy for this game?
 - (810, \$190.0), (156, \$120.5), (140, \$285.5), (542, \$184.0), (507, \$384.5), - (972, \$324.5), (431, \$114.0), (981, \$506.5), (865, \$290.0), (499, \$869.5), - (020, \$668.5), (123, \$83.0), (356, \$188.0), (015, \$449.0), (011, \$289.5), - (160, \$212.0), (507, \$466.0), (779, \$548.5), (286, \$260.0), (268, \$300.5), - (698, \$556.5), (640, \$371.5), (136, \$112.5), (854, \$254.5), (069, \$368.0), - (199, \$510.0), (413, \$102.0), (192, \$206.5), (602, \$261.5), (987, \$361.0), - (112, \$167.5), (245, \$187.0), (174, \$146.5), (913, \$205.0), (828, \$348.5), - (539, \$283.5), (434, \$447.0), (357, \$102.5), (178, \$219.0), (198, \$292.5), - (406, \$343.0), (079, \$332.5), (034, \$532.5), (089, \$445.5), (257, \$127.0), - (662, \$557.5), (524, \$203.5), (809, \$373.5), (527, \$142.0), (257, \$230.5), - (008, \$482.5), (446, \$512.5), (440, \$330.0), (781, \$273.0), (615, \$171.0), - (231, \$178.0), (580, \$463.5), (987, \$476.0), (391, \$290.0), (267, \$176.0), - (808, \$195.0), (258, \$159.5), (479, \$296.0), (516, \$177.5), (964, \$406.0), - (742, \$182.0), (537, \$164.5), (275, \$137.0), (112, \$191.0), (230, \$298.0), - (310, \$110.0), (335, \$353.0), (238, \$192.5), (294, \$308.5), (854, \$287.0), - (309, \$203.5), (026, \$377.5), (960, \$211.5), (200, \$342.0), (604, \$259.0), - (841, \$231.0), (659, \$348.0), (735, \$159.0), (105, \$130.5), (254, \$176.0), - (117, \$128.5), (751, \$159.0), (781, \$290.0), (937, \$335.0), (020, \$514.0), - (348, \$191.0), (653, \$304.5), (410, \$167.0), (468, \$257.0), (077, \$640.0), - (921, \$142.0), (314, \$146.0), (683, \$356.0), (000, \$96.0), (963, \$295.0),
- Using which method?
- Viz?
- Statistics?
- Data mining?
- Opinion?

Visualizing the data

- Humans can really only make sense of three or four numbers at a time
- By representing the values in a graphical form we make it easier to handle large numbers of values
- Using visualizations should make it possible to learn more about this data
- We have NOT to **lie** or make **noise !!!**

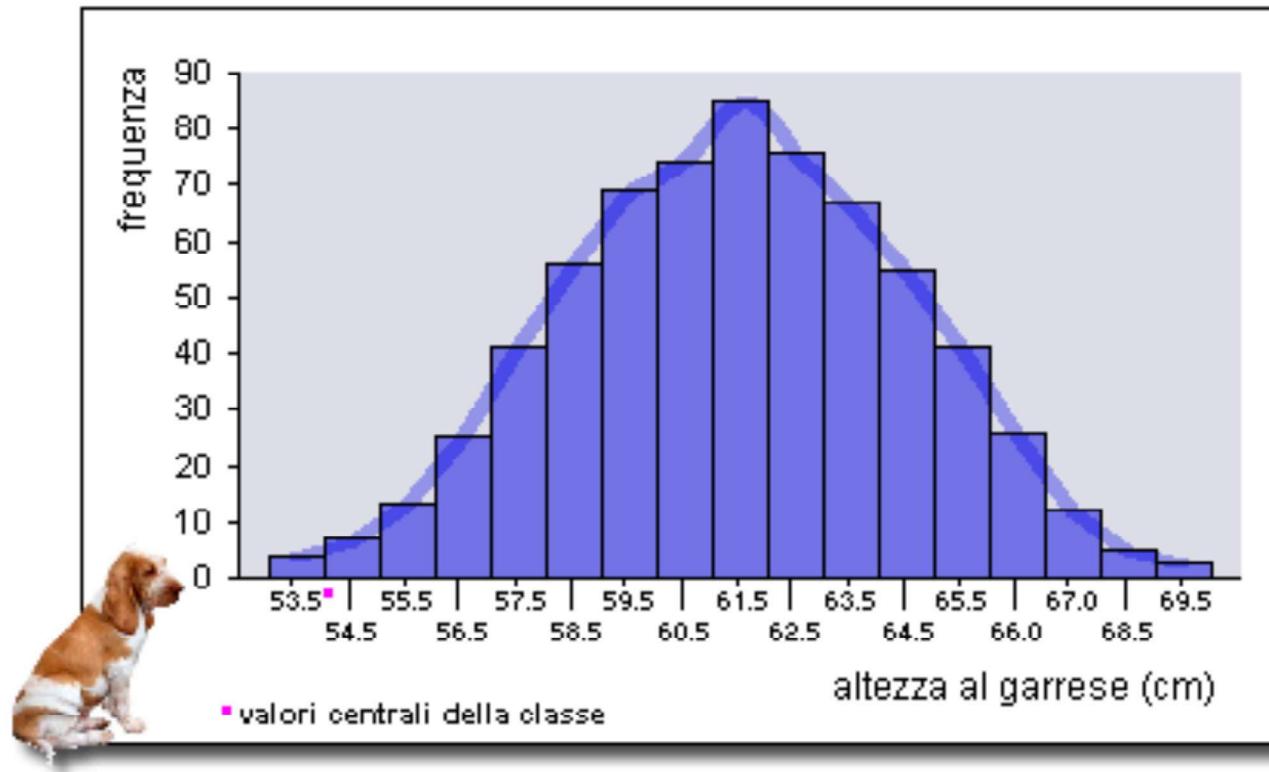
User task and visualization

- One approach to making money at “Pick It” is to try to select numbers which are **more likely to win**
 - This is why casinos change the roulette wheel on regular basis...
- Since we have data on the winning numbers we can look at the distribution of the winning numbers and see whether some (ranges of) values are more likely to produce a win than others
- One way to do this is to produce a histogram of the winning numbers

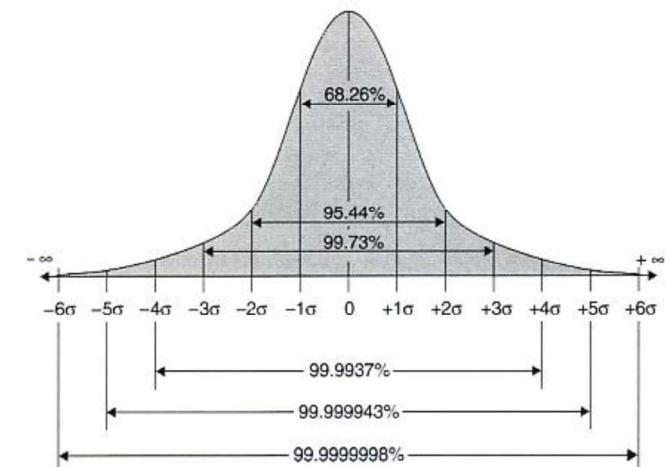


Histogram example

Altezza al garrese di 659 cani di razza "Bracco italiano". Iistogramma.



bin



Excel and histograms

Microsoft Excel - histogram.xls

File Edit View Insert Format Tools Data Window Help Acrobat REC Spelling... F7 Error Checking... Speech Share Workbook... Track Changes Compare and Merge Workbooks... Protection Online Collaboration Goal Seek... Scenarios... Formula Auditing Solver... Tools on the Web... Macro Add-Ins... AutoCorrect Options... Customize... Options... Data Analysis...

1. Create a column of bin right hand endpoints. Bin widths must be equal. There should be no sample values before the first endpoint or beyond the last endpoint.

2. Select Tools: Data Analysis...

3. Select Histogram, OK

4. Insert data and bin endpoints

5. Select left hand top cell of output

6. Tick Chart Output and OK

7. A table of frequencies and a chart purporting to be a histogram appears.

	Bin	Frequency
25	3.5	0
26	4.5	2
27	5.5	1
28	6.5	3
29	7.5	4
30	8.5	3
31	9.5	2
32	10.5	4
33	11.5	0
34	12.5	1
35	More	0

Histogram

Frequency

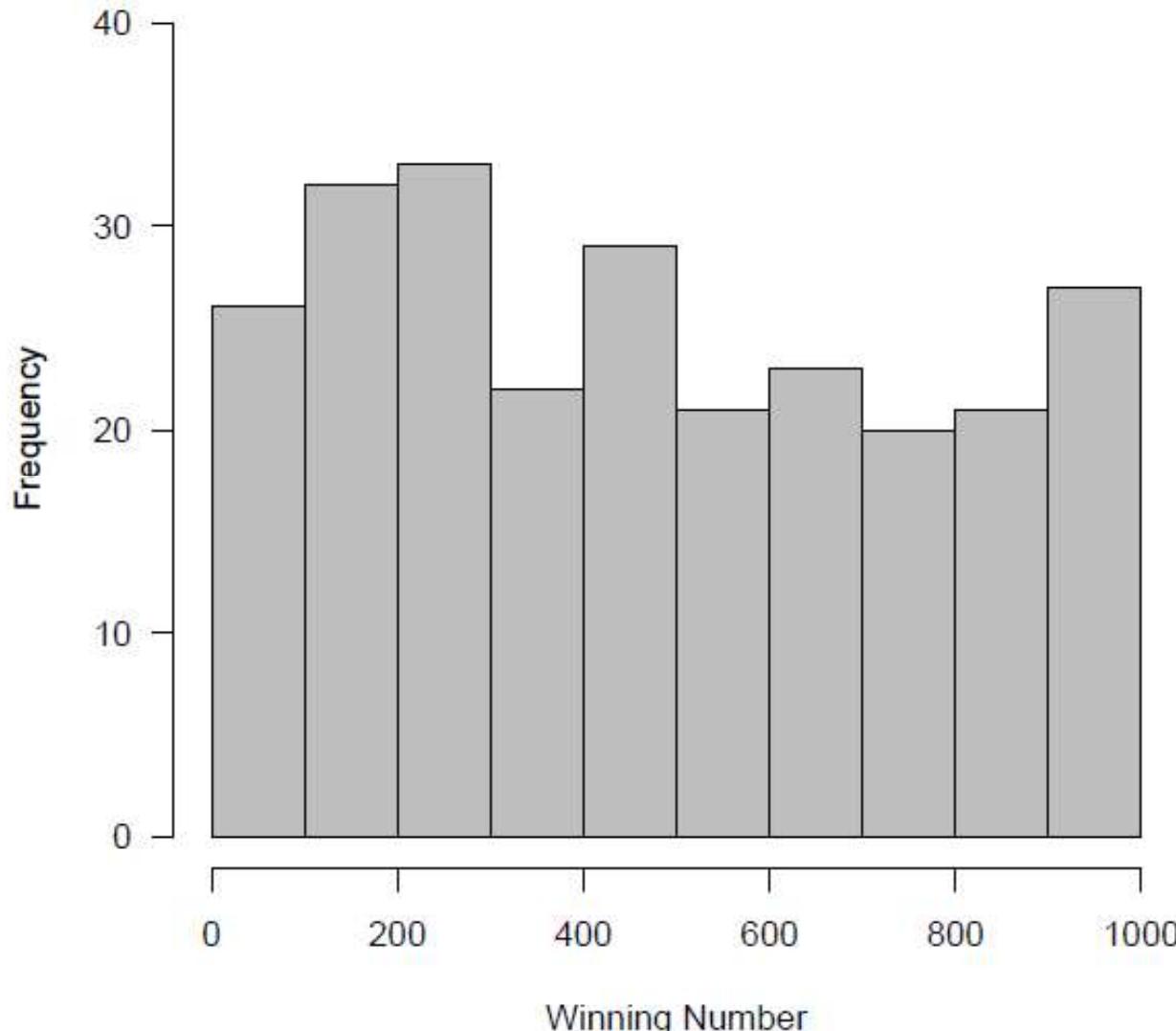
Bin

3.5 4.5 5.5 6.5 7.5 8.5 9.5 10.5 11.5 More

Frequency

The histogram displays the frequency of data points across bins of width 1.0. The x-axis is labeled 'Bin' with ticks at 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5, 10.5, 11.5, and 'More'. The y-axis is labeled 'Frequency' with ticks at 0, 2, 4, and 6. The bars represent the frequency of data falling within each bin: (3.5, 4.5] has freq 2; (4.5, 5.5] has freq 1; (5.5, 6.5] has freq 3; (6.5, 7.5] has freq 4; (7.5, 8.5] has freq 3; (8.5, 9.5] has freq 2; (9.5, 10.5] has freq 4; (10.5, 11.5] has freq 0; (11.5, More] has freq 1.

Data distribution



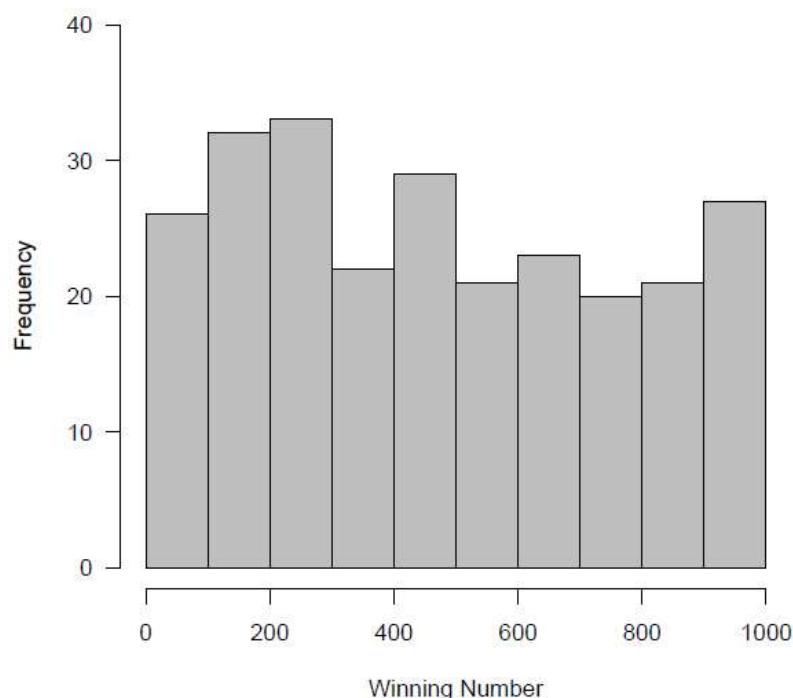
What can we
infer from this
histogram?

Opinions?

By the way, is the bin size ok?

Analysis

- It looks there are more winners in the region from 200 to 300 than in other regions
- This suggests that we might be best to choose numbers in this range



Do you agree ?

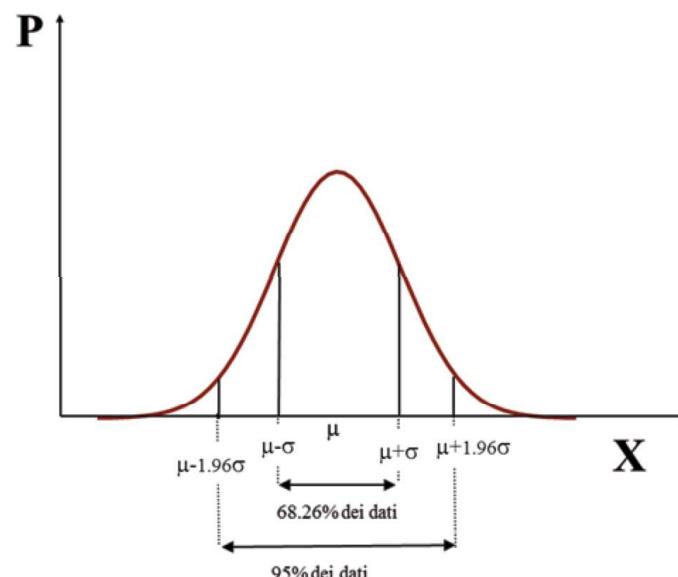
We are telling lies...

(wrong number understanding)

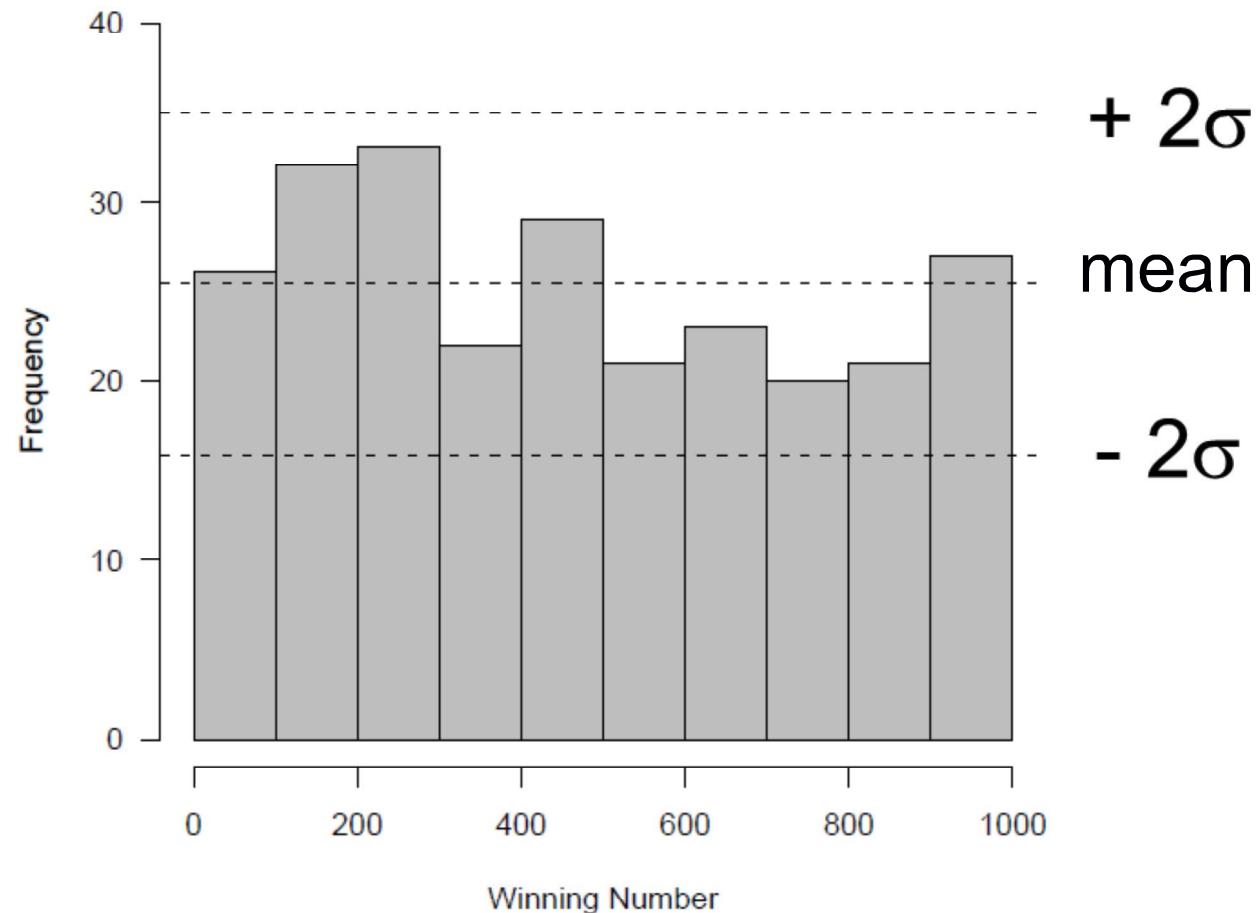
- Even if the winning numbers are chosen randomly, we can expect some “random variability” in a sample
- To judge the significance of what we see in the histogram we have to recall some formal statistical theory

The mean is not enough !

- There are 254 values. We would expect the number of values in each cell to be approximately: $25.4 = 254/10$
- Such a number is a random variable as well, with normal distribution
- 95% of the observations fall within $+/- 2\sigma$



Better number visualization



- Variance analysis AND visualization

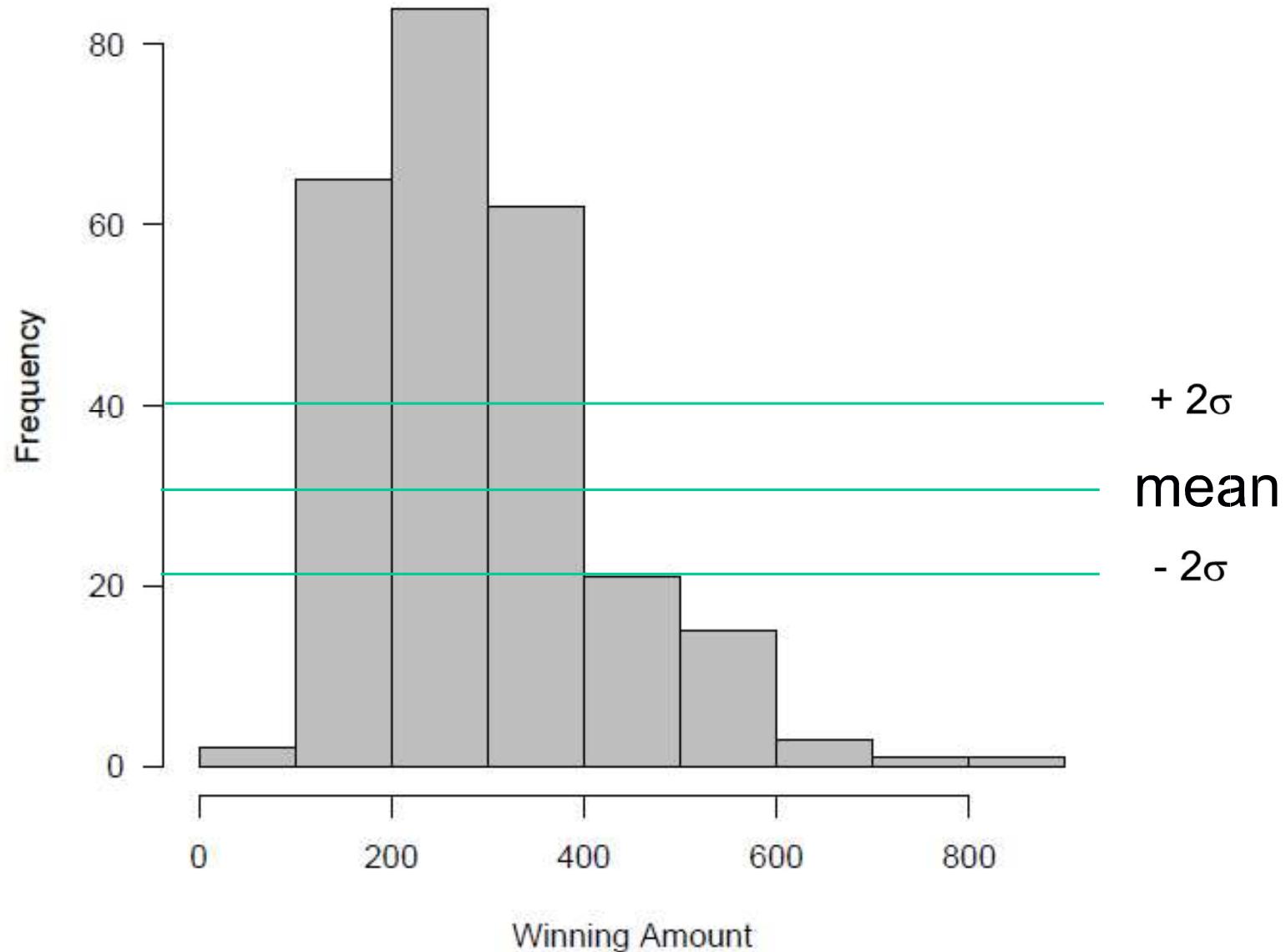
Conclusions and new task

- Winning numbers are totally random (according to the actual sample...)
- It makes no sense to look for a "lucky" number
- However, we can change our task:

–to increase the won amount !

- So we study the distribution of winning amount

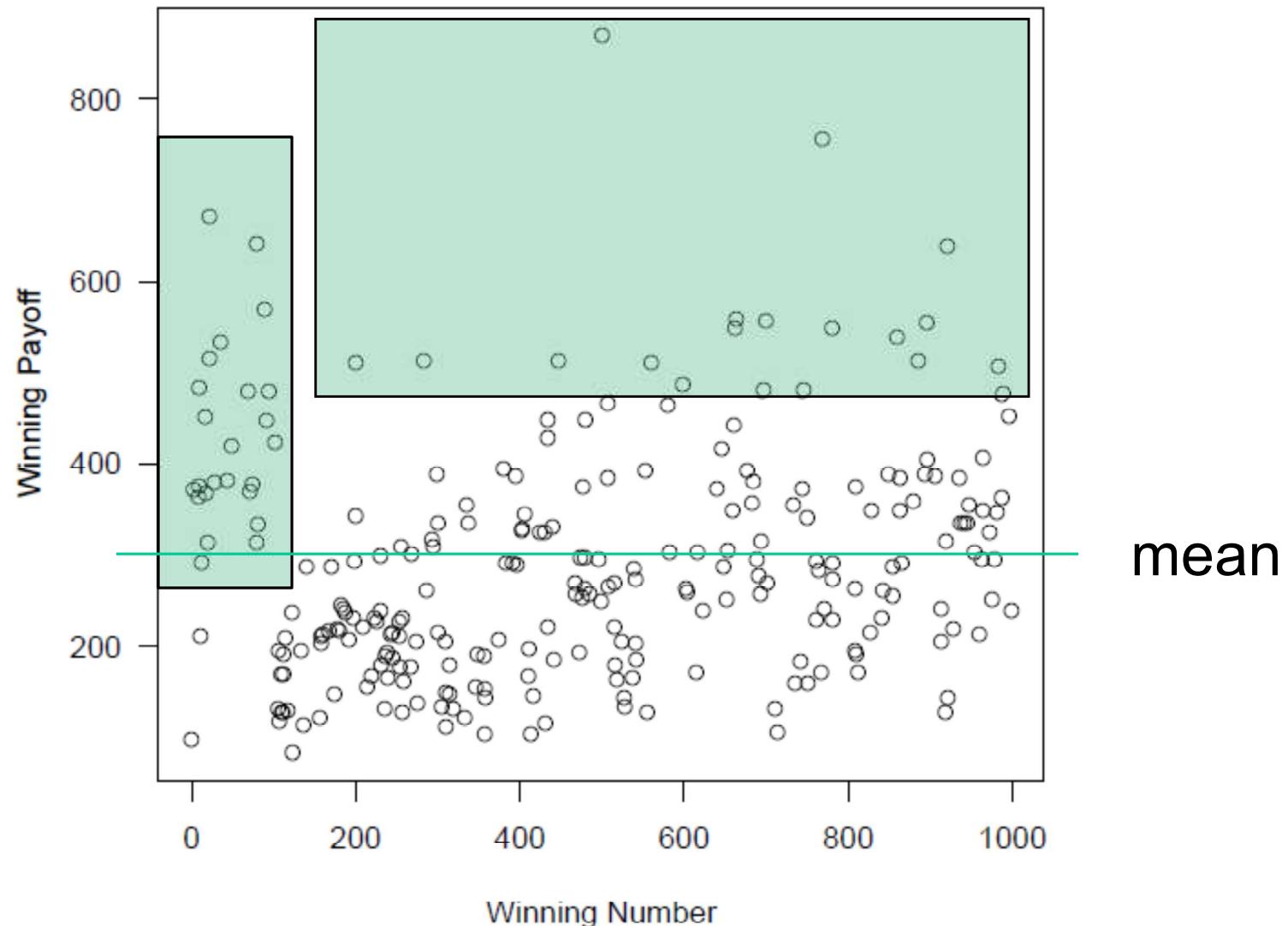
New visualization



Looking for new insights

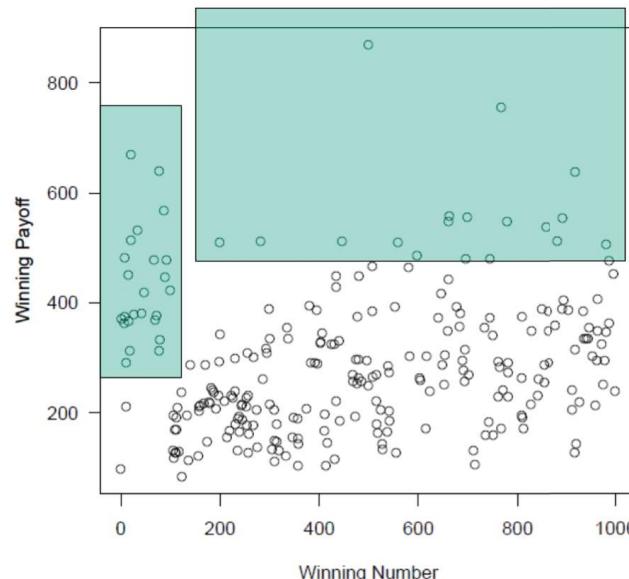
- The histogram shows that there is a wide (more than 2σ) range amounts won in the game
- It might be possible to choose the numbers which win larger amounts
- We search for relationship between ticket number and winning amount
- A scatter plot is the natural way to look for such a relationship.

A new visualization

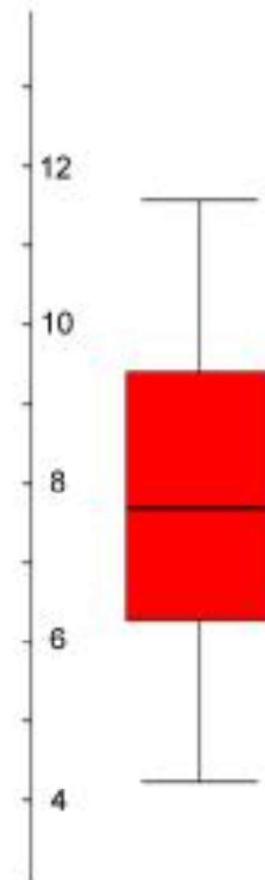
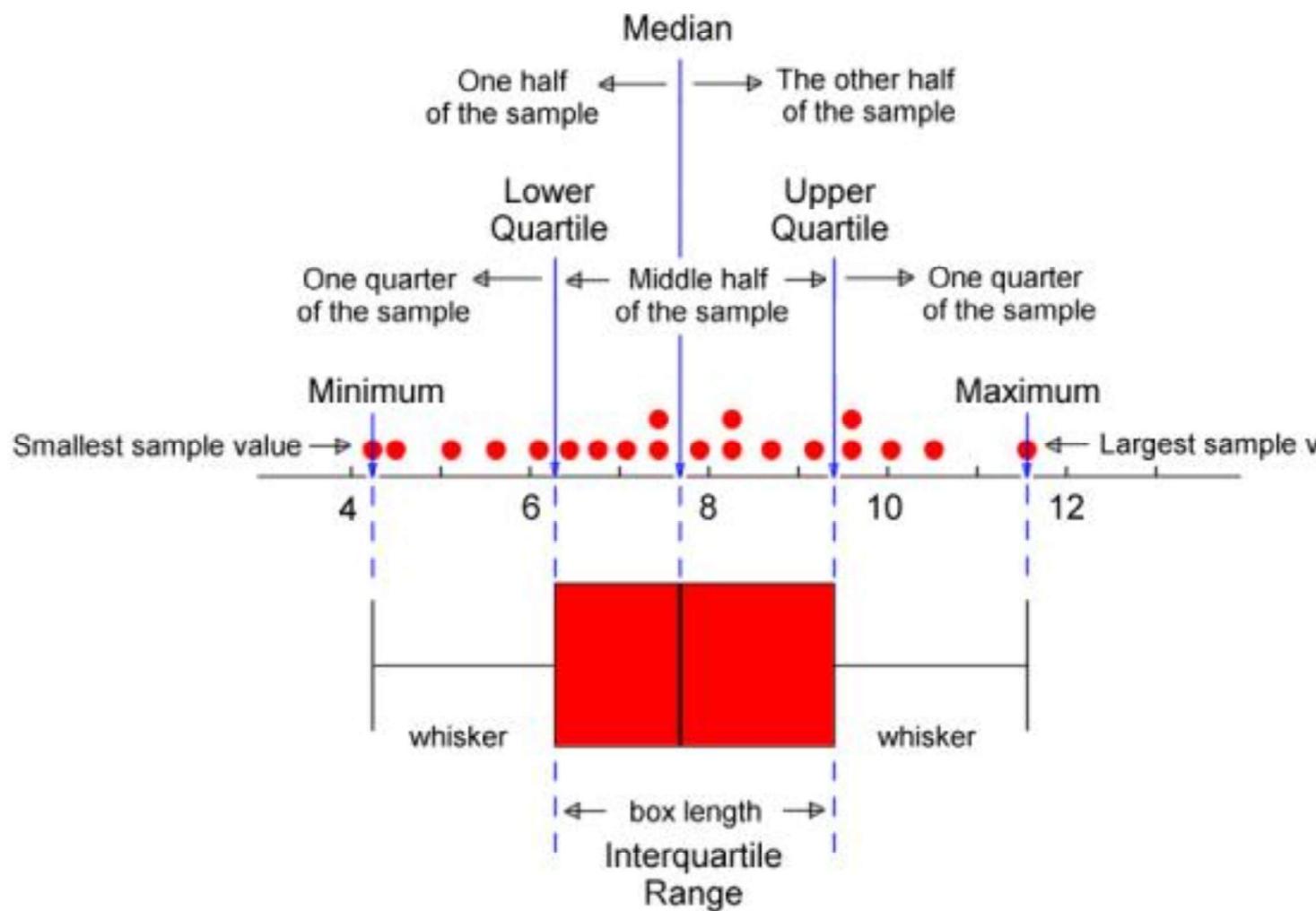


Insights from the scatterplot

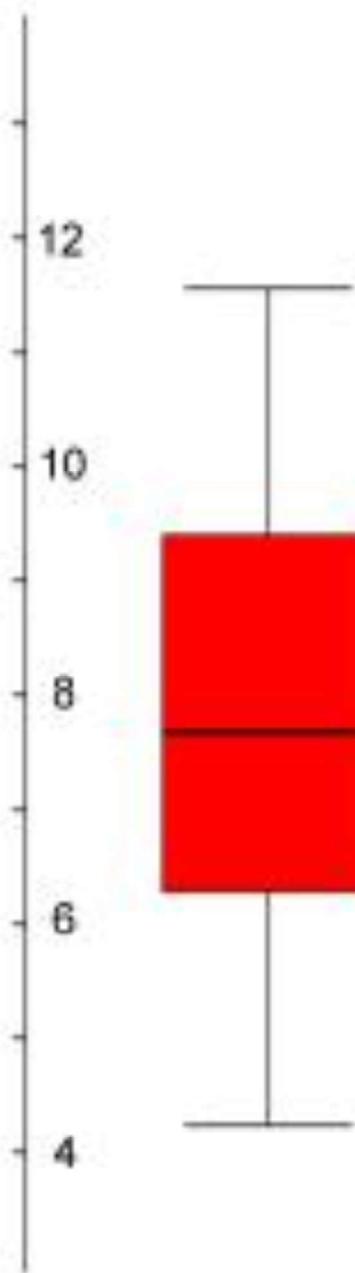
- The winning amounts in a band to the left of the plot appear to generally be higher than those in the rest of the plot
- There are some outliers in the top band
- We can investigate this further by separating the numbers into groups according to the **first digit** of the ticket number and drawing box plots for each group



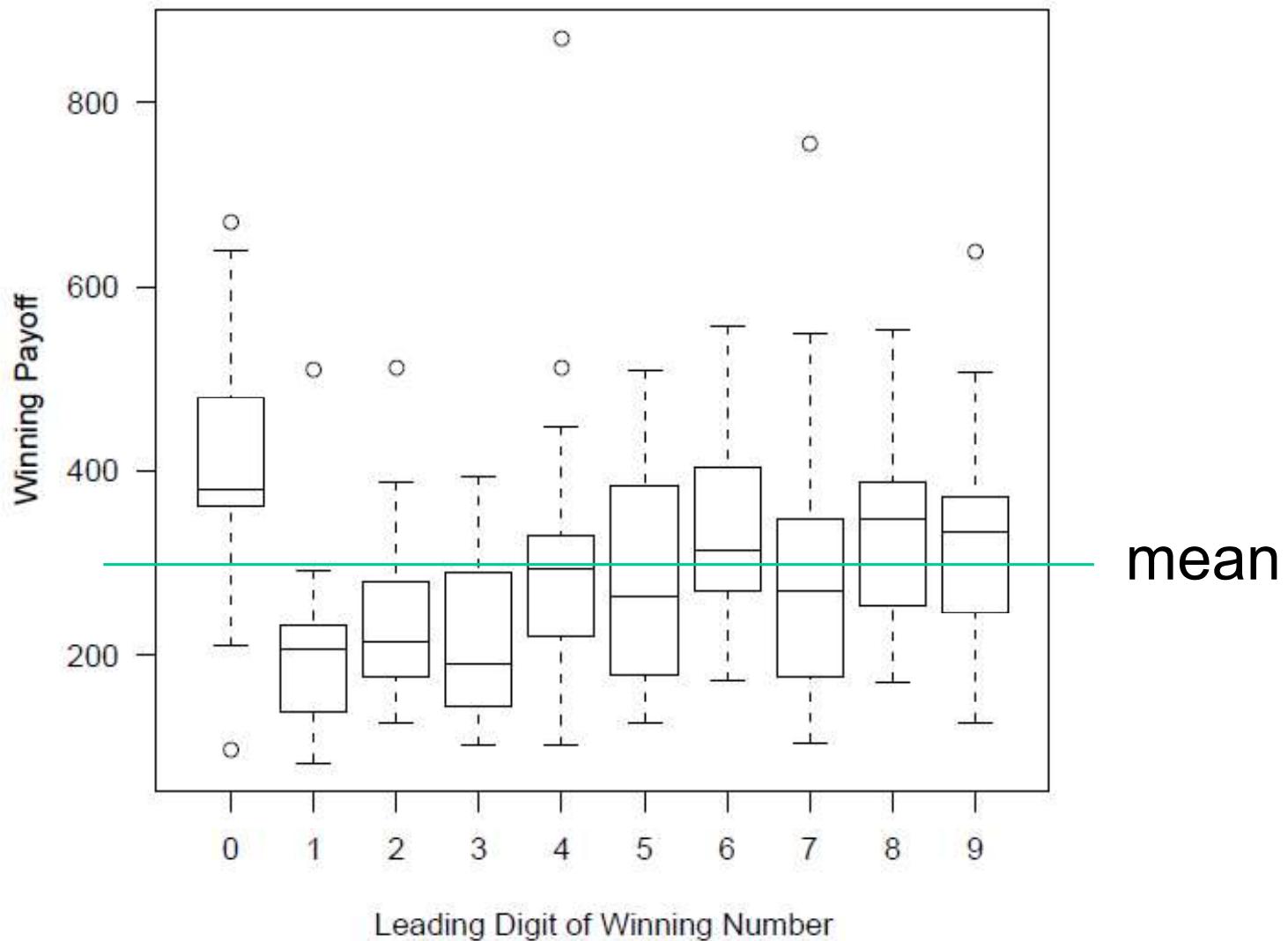
Boxplot



Boxplot

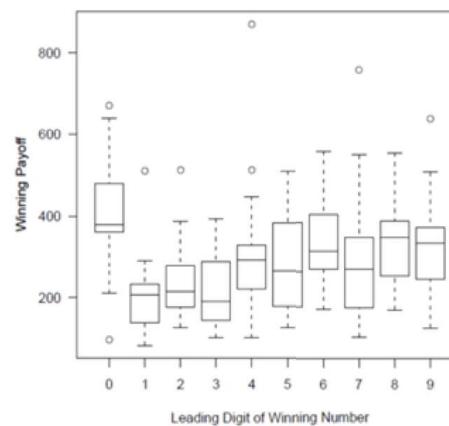


Lottery's boxplots

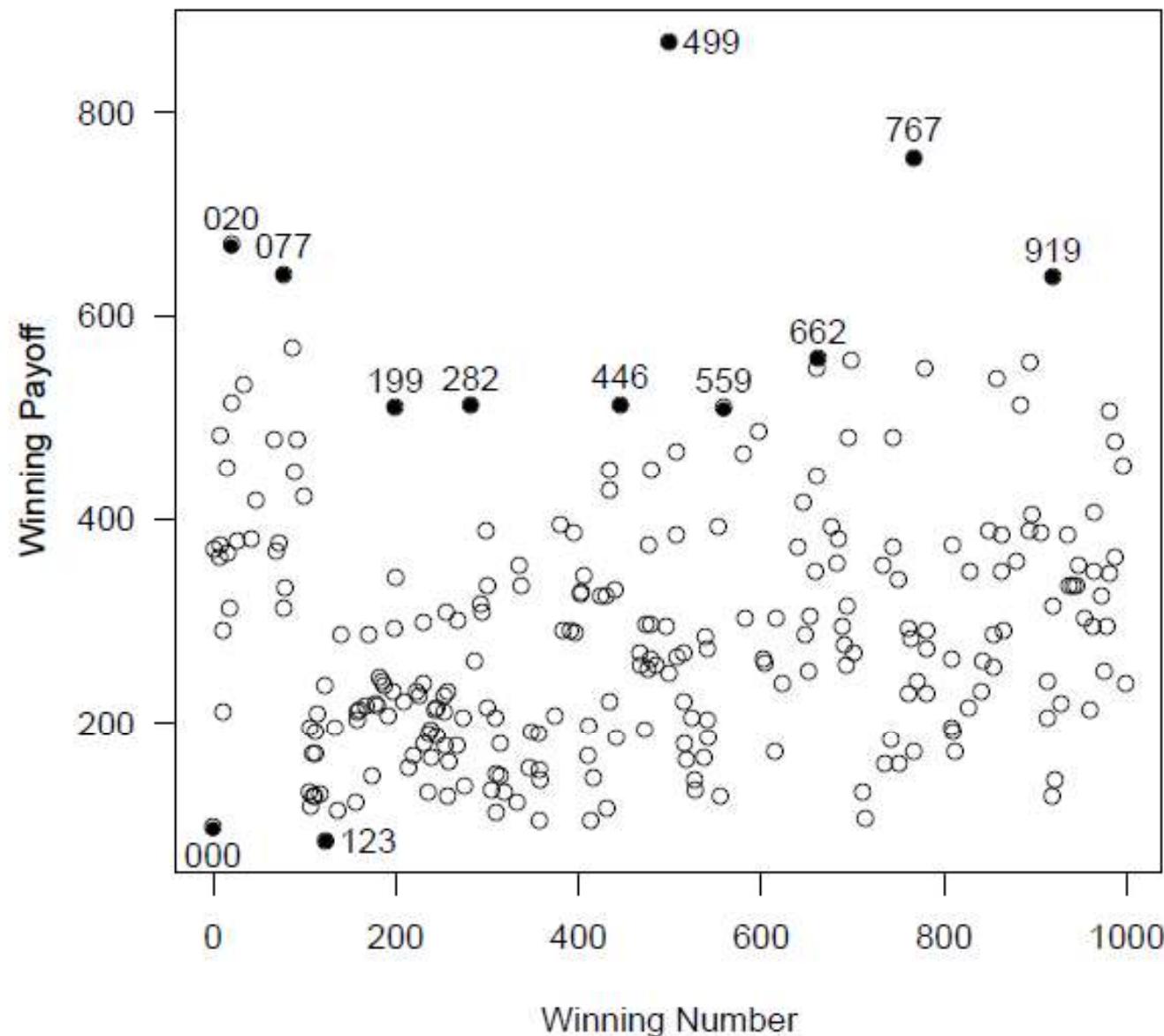


New insights

- Tickets with a leading zero digit clearly tend to produce larger winnings
- It is also apparent that there are some very large and some very small winning amounts
- It is probably of interest to identify the ticket numbers corresponding to these extremes



High and low winning numbers



Lotto strategy

- While winning numbers are non predictable, players' choices are!
- Choose numbers which are less likely to be chosen by other players
- Then, when you win (if), you will tend to win more
- Possible ways to choose:
 - Choose a number with a leading zero
 - Choose a number with repeated digits
 - Avoid “obvious” numbers like, e.g. 000, 123, 246, . . .

Lessons learned

- Define clearly the task
- Use basic visualizations
 - bar charts
 - scatterplots
 - boxplots
- Be ready to switch among them
- Look for precise values when needed
- Do not lie !

Outline

- An introductory example
- Good and bad graphs

Informal approach

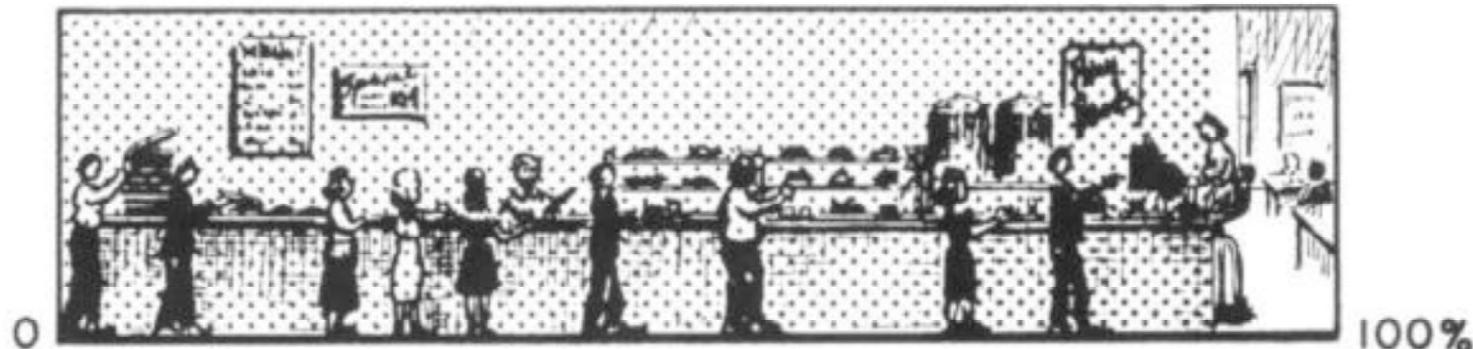
- In this lecture we will try to set down some basic rules for drawing good graphs
- We will do this by showing that violating the rules produces bad graphs
- Next lectures will cover these issues in a more formal way

Rule 0

- **Do not use diagrams when handling few numbers**
- It does not make sense to use graphs to display very small amounts of data
- The human brain is quite capable of grasping one two, or even three values

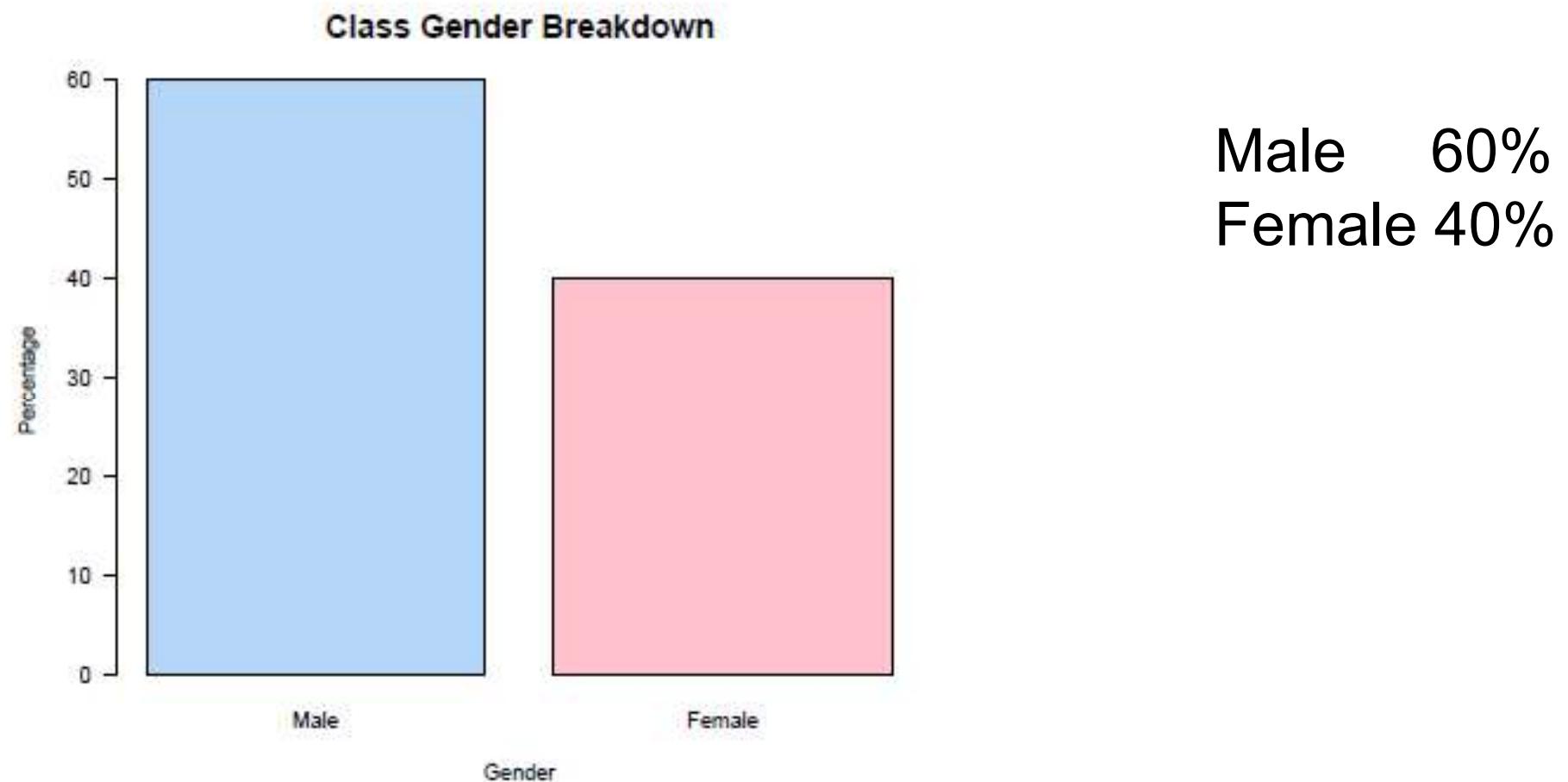
Rule 0 violation (and also rule 2)

The Company Cafeteria was used by 9 Out of 10 Employees during the Fiscal Year 1949



Source: COMPANY REPORTS

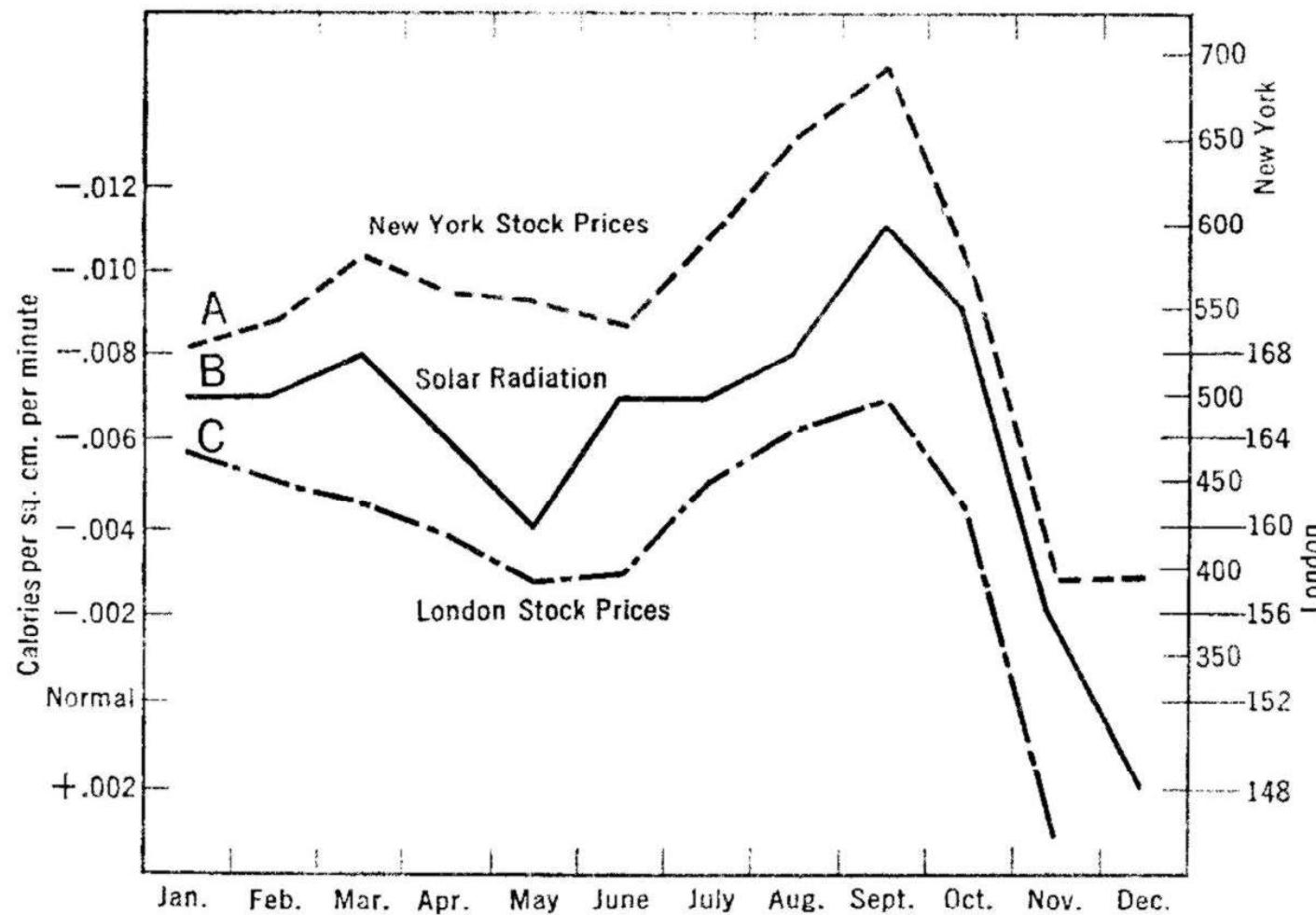
Rule 0 violation



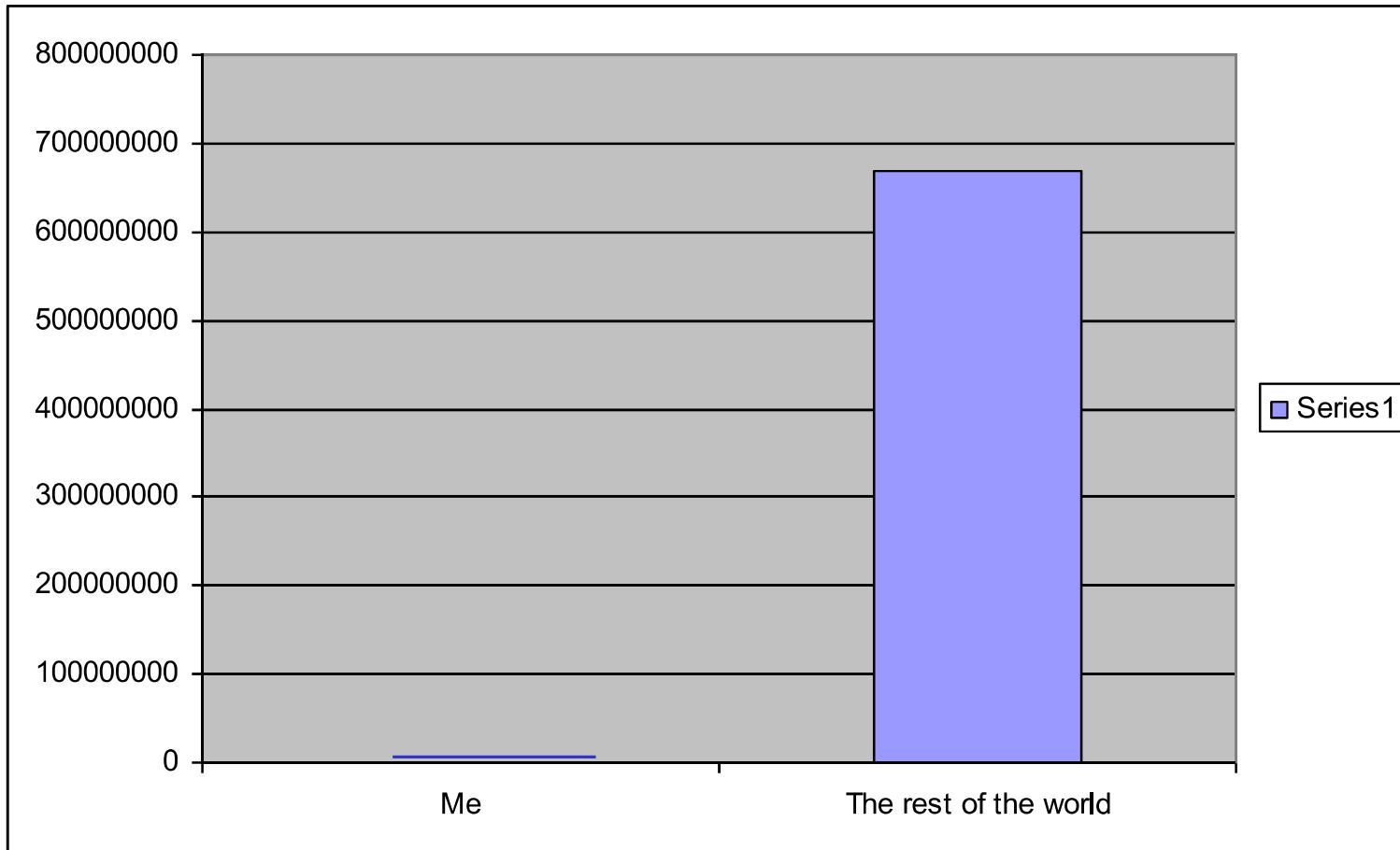
Role 1

- **Insure data quality / significance**
- Graphs are only as good as the data they display
- No amount of creativity can produce a good graph from dubious or non relevant data

Role 1 violation



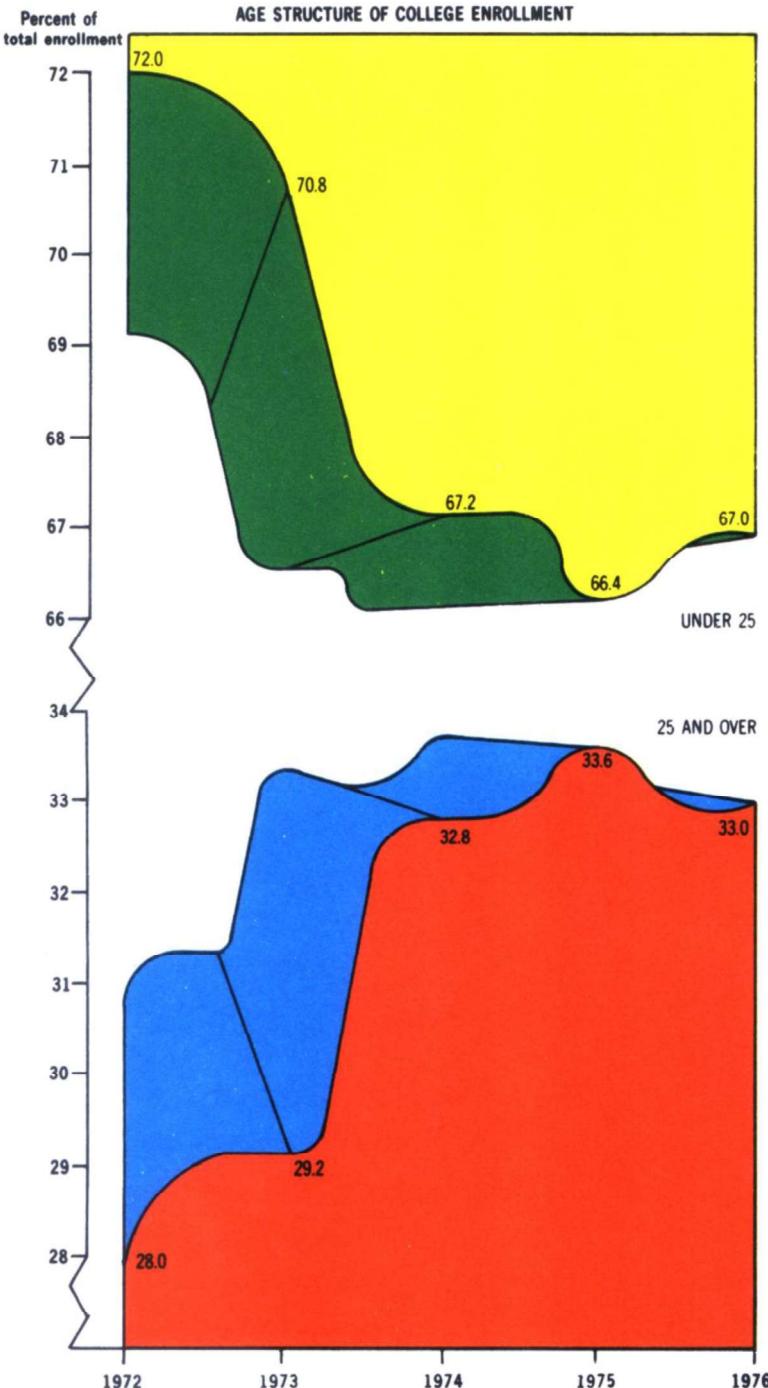
Role 1 violation (and also rule 0)



Not very significant data but a good example of distortion

Rule 2: Insure chart simplicity

- Graphs should be no more complex than the data which they portray
- Unnecessary complexity can be introduced by
 - irrelevant decorations
 - colors
 - 3d effects
 - ...
- They are collectively known as “chart junk”
- For a very comprehensive set of chart junk effects look at Microsoft Excel
 - the more recent the version the larger the set !

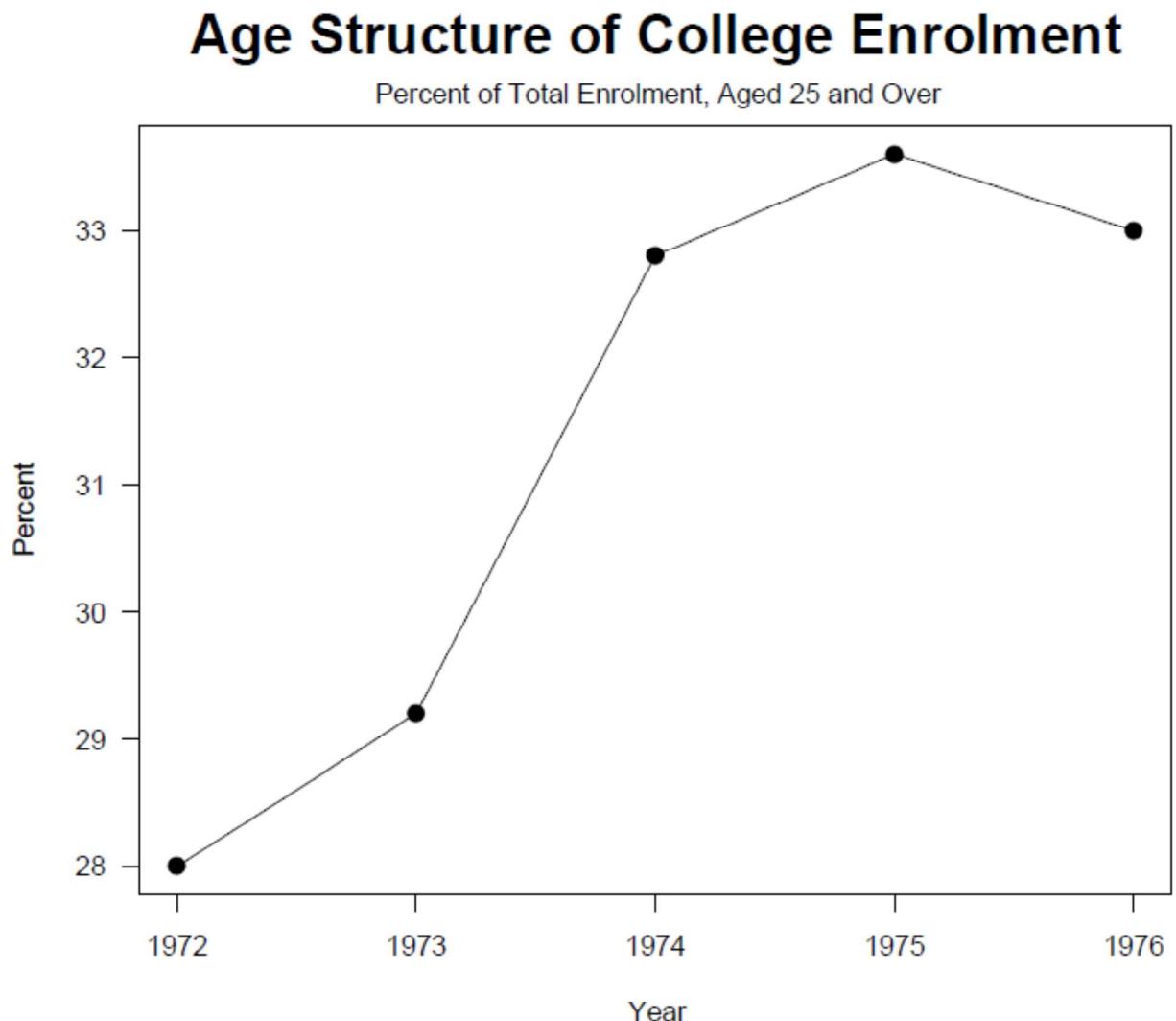


Age structure of College enrollment
(percentage of enrolled people above 25 years)

Role 2 violation (and also rule 3)

- A very good bad example!
- only 5 (!) numbers on it but
 - 4 meaningless colors
 - useless 3D
 - useless axes split
 - confusing and wrong visual attributes (size)
 - split y axis
 - odd interpolation
- Designers of this graph are now working in the Microsoft Excel team, inspiring the new Excel versions ...

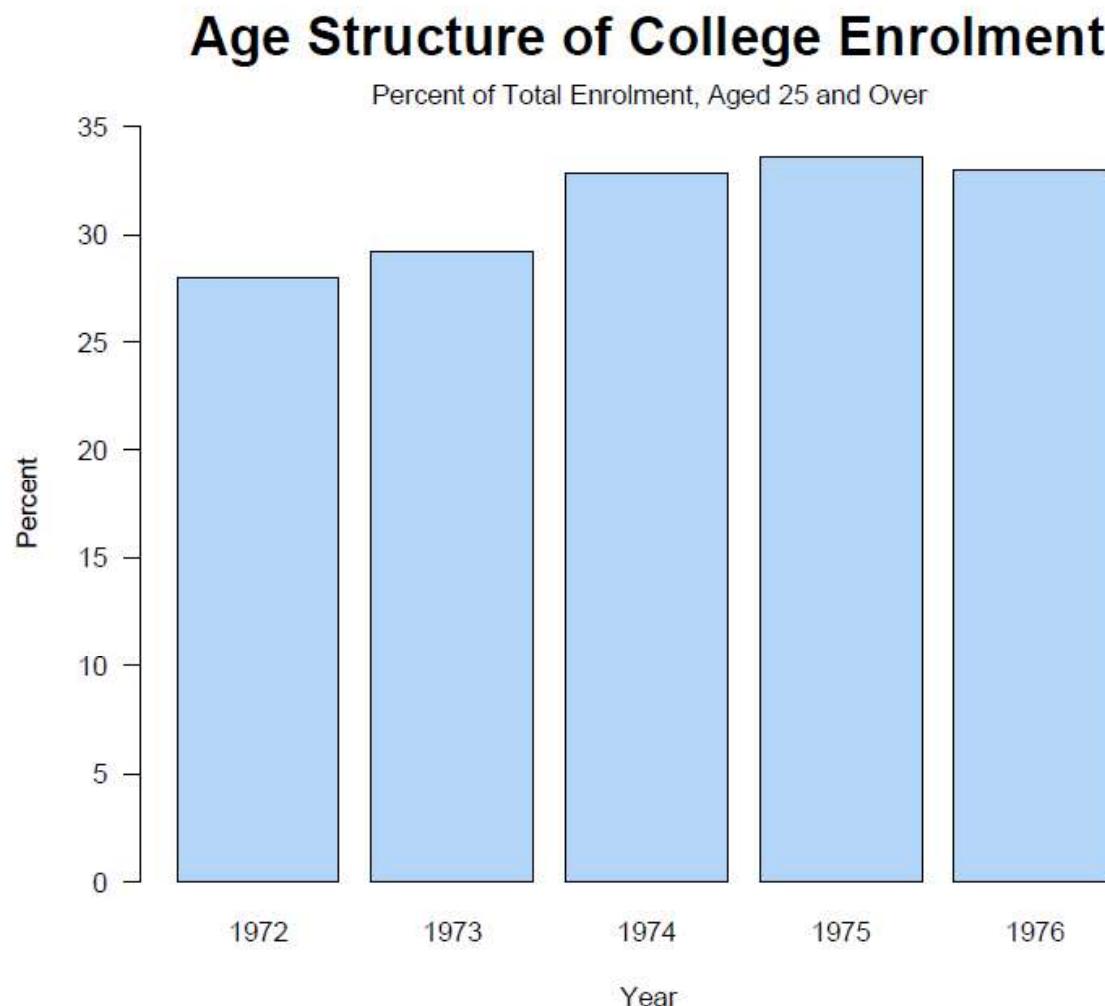
Same data...

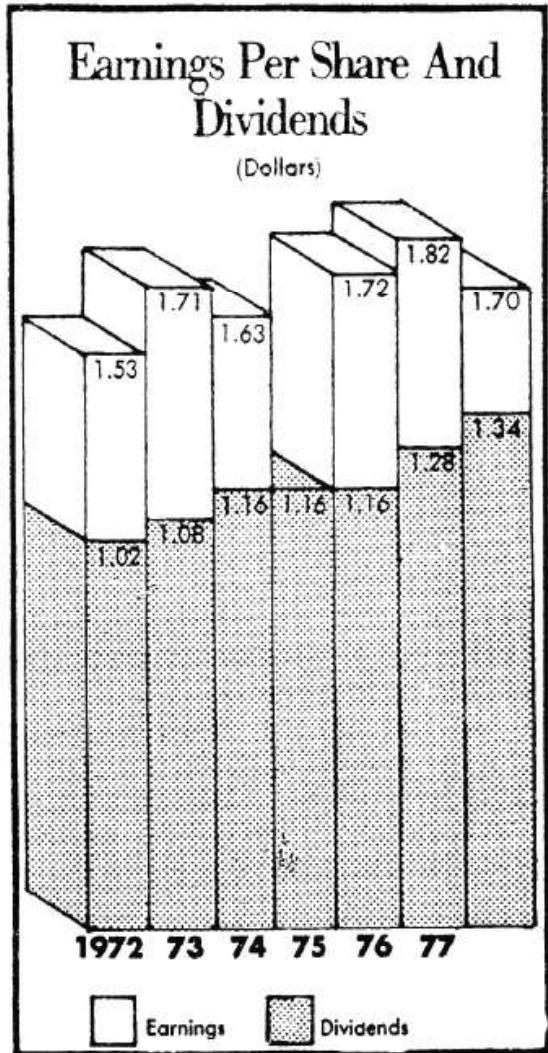


The same data...

Year	Percentage above 25
1972	28.0
1973	29.2
1974	32.8
1975	33.6
1976	33.0

Same data...





tion

Why 3D?

The extra dimension used in this graph has confused even the person who created it..

The Washington Post, 1979

The same data...

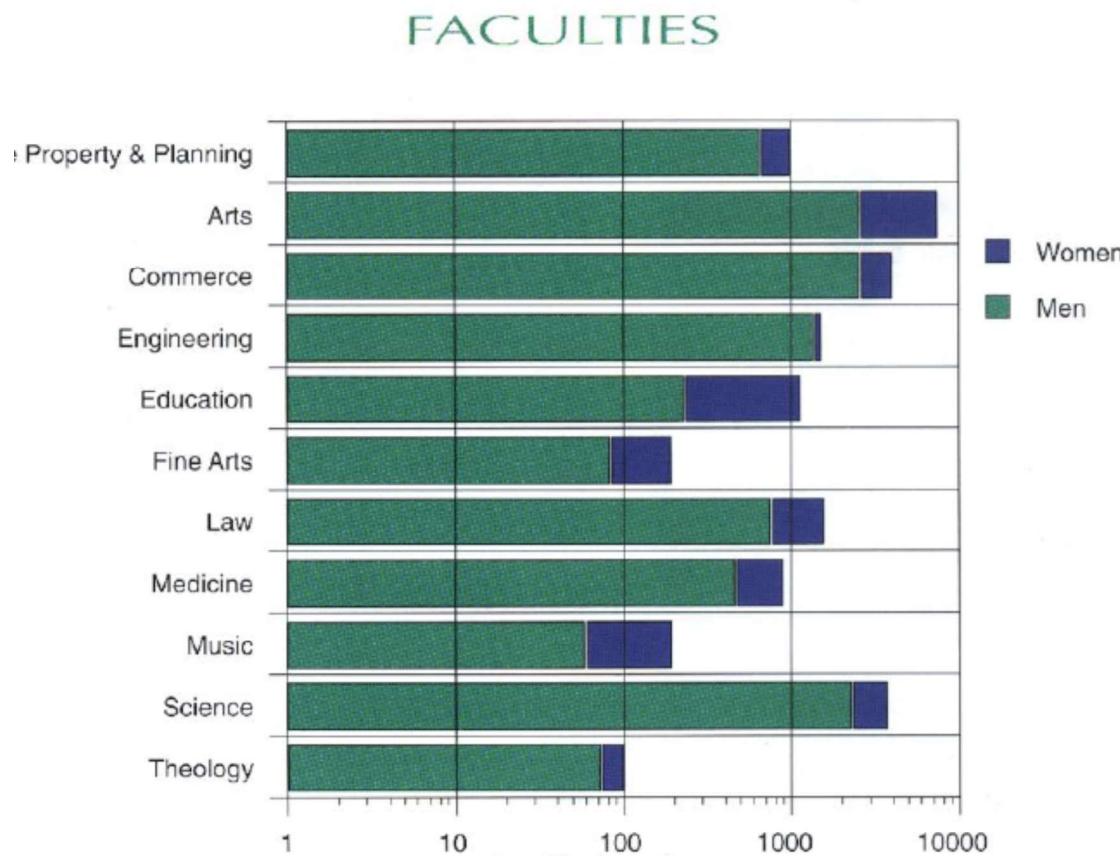
Earnings Per Share and Dividends



Role 3

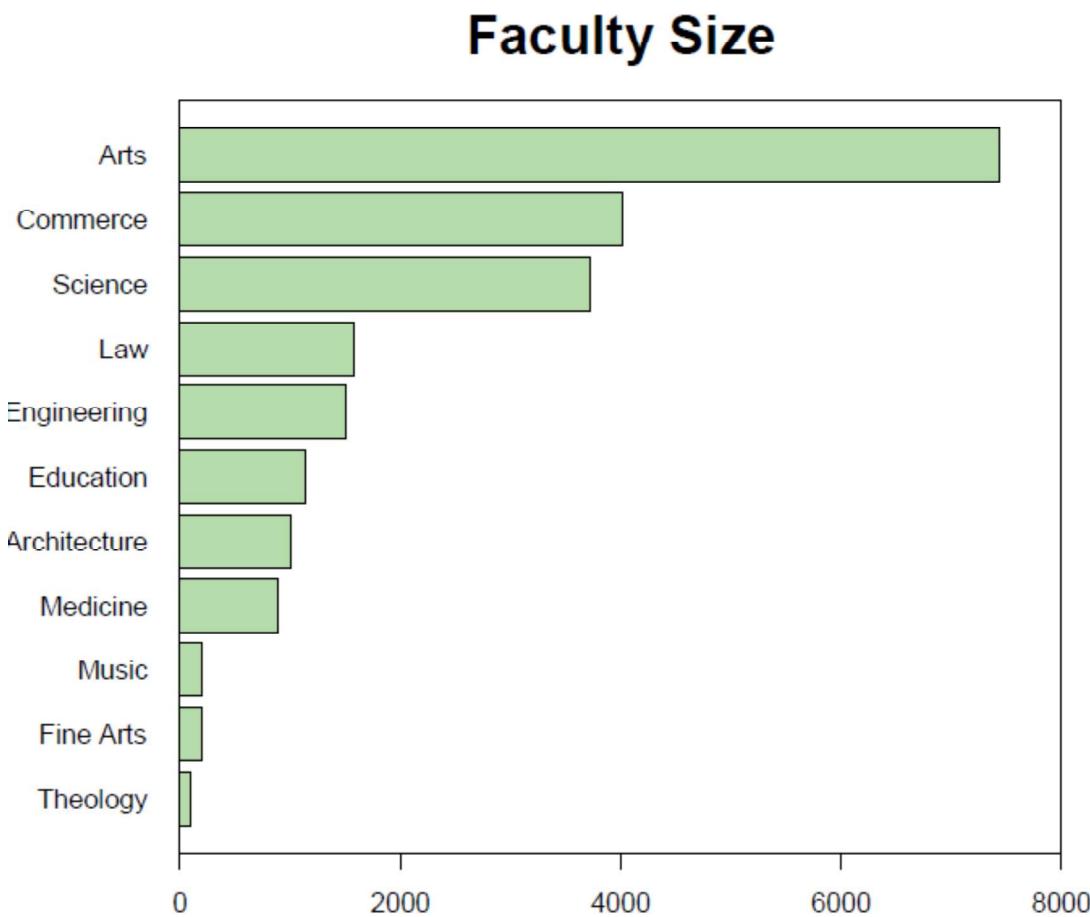
- **Do not distort data in a confusing way**
- Graphs should not provide a distorted picture of the values they portray
- Distortion can be either deliberate or accidental
- Of course, it could be useful to know how to produce a graph which bends the truth...

Role 3 violation



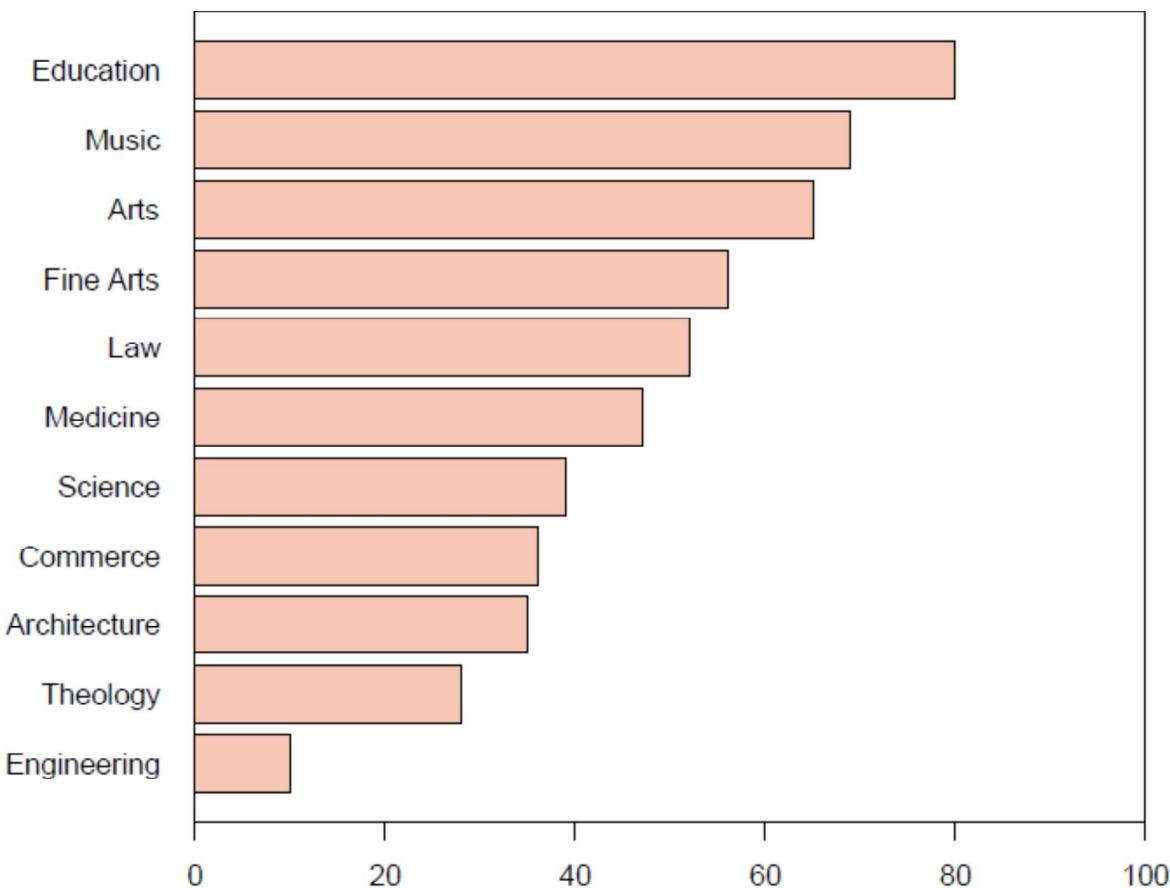
- At a very quick glance:
 - balanced faculty population
 - most male students
- What is wrong with this graph?
- The X scale is logarithmic!

The truth : population size

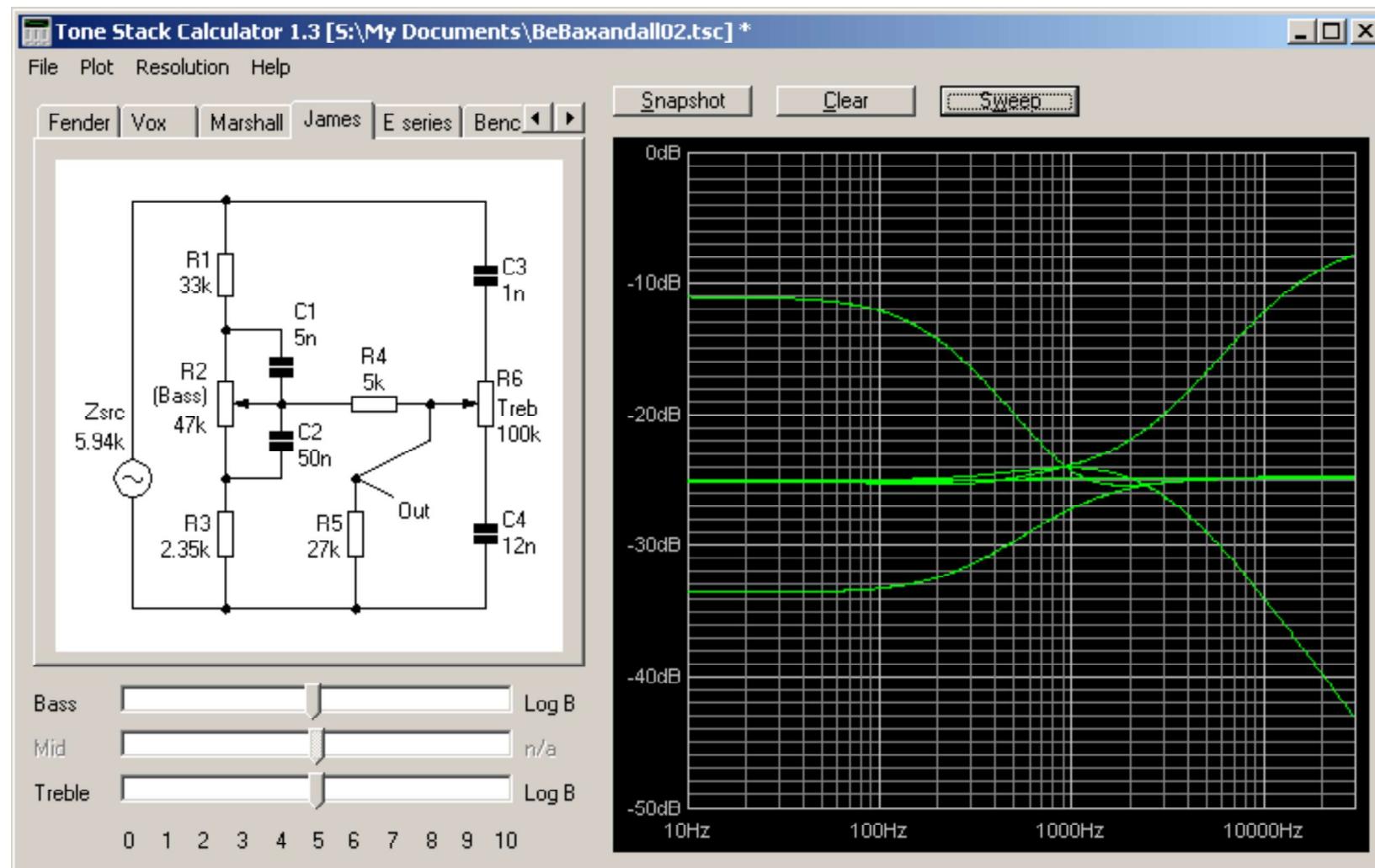


The truth : female /male ratio

Percentage of Female Students



In other cases distortion is ok...

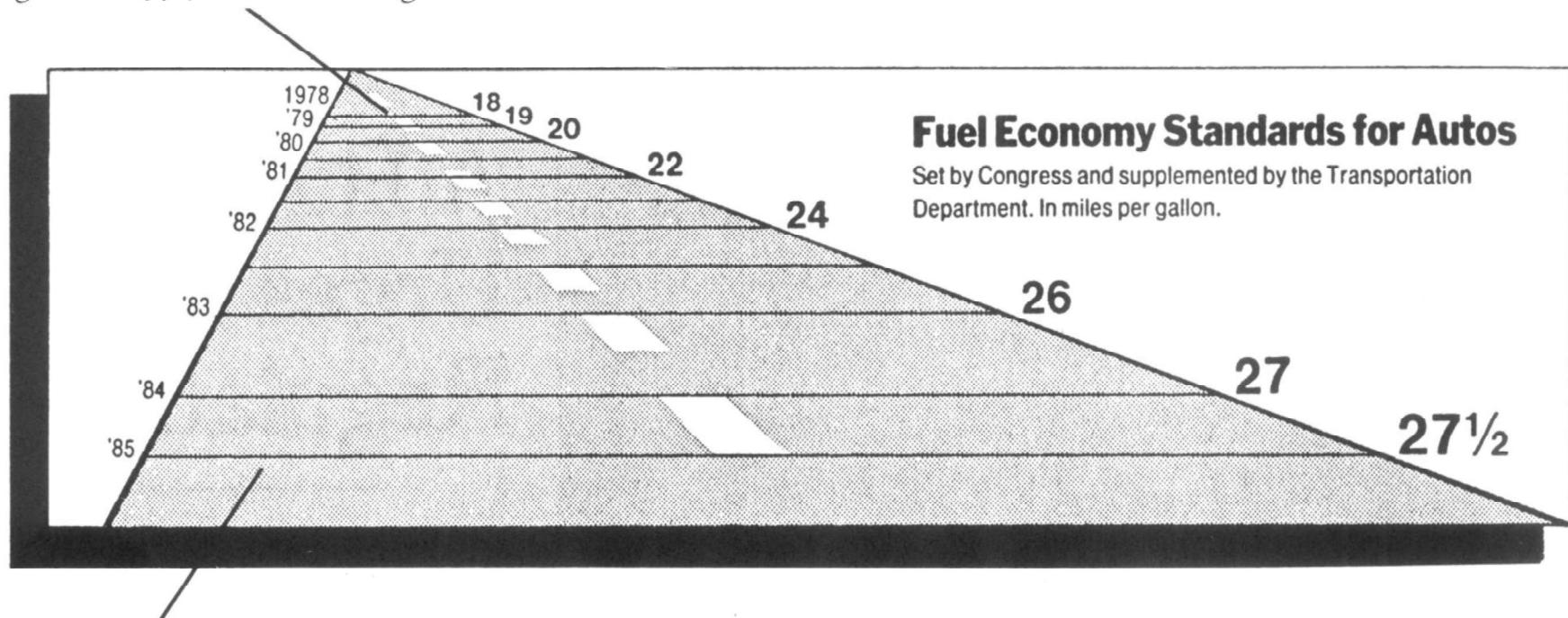


The lie factor

- The visual pioneer Ed Tufte of Yale University has defined a “lie factor” as a measure of the amount of distortion in a graph
- The lie factor is defined to be:
- Lie Factor = size of effect in graphic / size of effect in data
- If the lie factor of a graph is greater than 1, the graph is exaggerating the data effect
- If the lie factor of a graph is lower than 1, the graph is minimizing the data effect

Measuring distortion through the lie factor

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



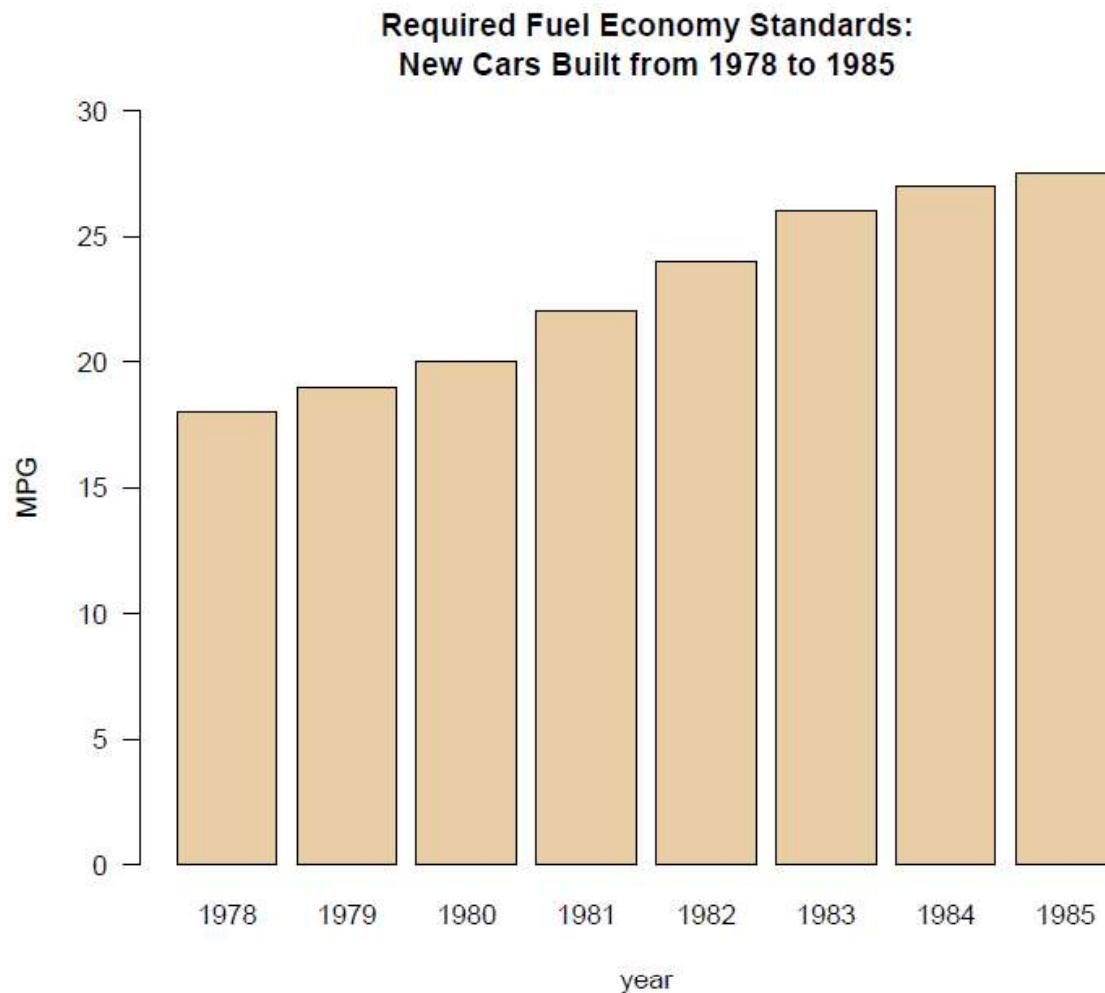
This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

$$\text{Graph effect} = 5.3/0.6=8.8$$

$$\text{Data effect} = 27.5/18=1.52$$

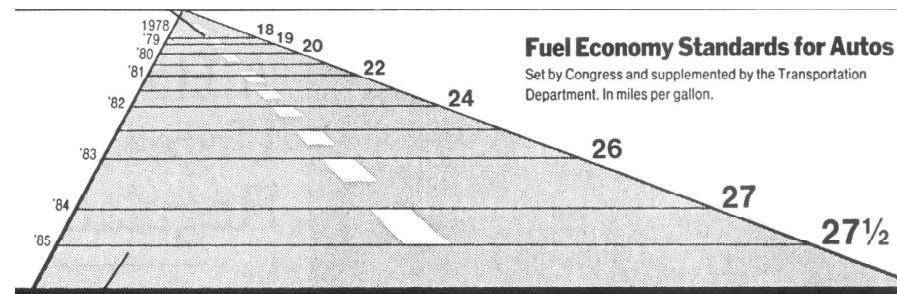
$$\text{Lie Factor} = 8.8/1.52 = 5.8$$

The same data with lie factor=1

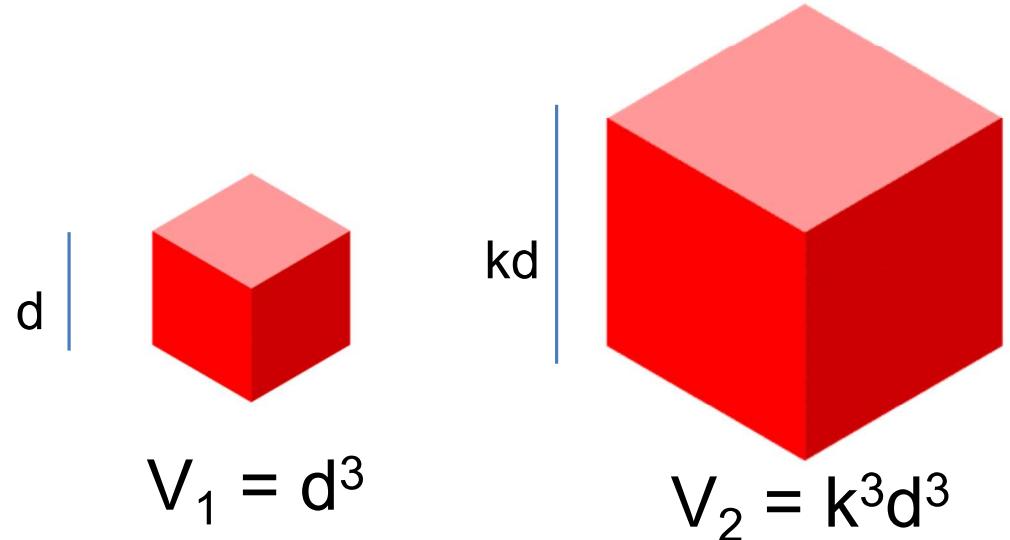
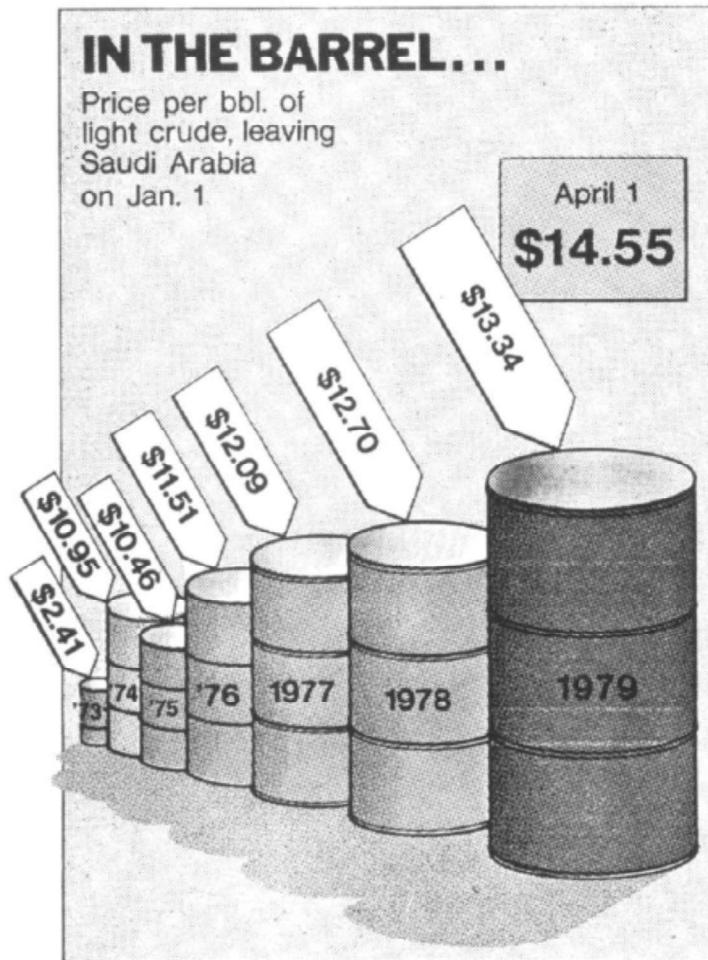


Common Sources of Distortion

- The use of image perspective is a common source of distortions in graphs
- Another common source is the inappropriate (or deliberate?) use of linear scaling when using area or volume to represent values



Distortion through non linear volumes



$$\text{Graph effect} = V_2/V_1 = k^3 d^3/d^3 = k^3$$

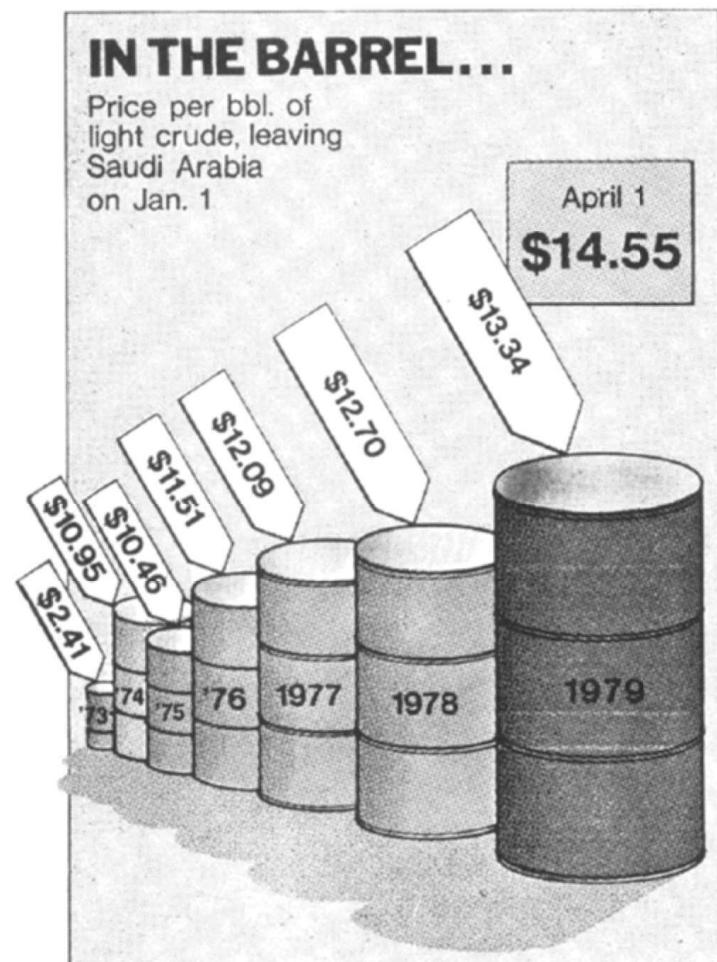
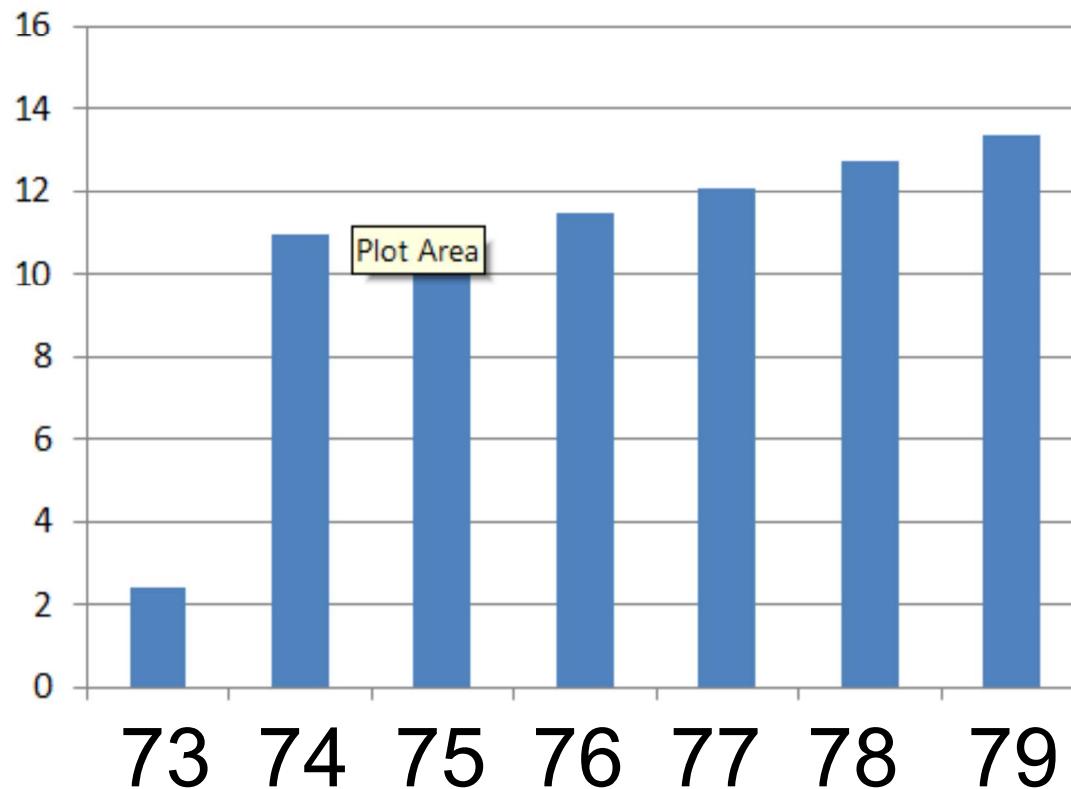
$$\text{Data effect} = kd/d = k$$

$$\text{Lie Factor} = k^3/k = k^2$$

$$\text{Lie Factor} = \text{Data effect}^2$$

$$\text{Lie factor} = (14.55/2.41)^2 = 6^2 = 36$$

The same data



Distortion through areas

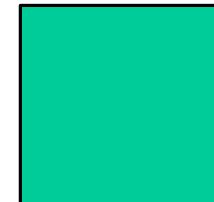
1

0.94

0.83

0.64

0.46



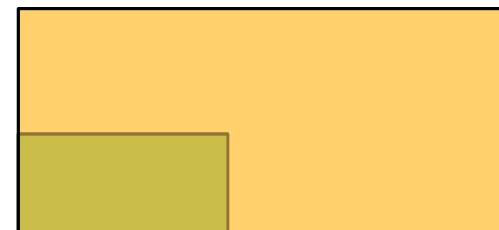
kd

Graph effect = $A_2/A_1 = k^2 d^2/d^2 = k^2$

Data effect = $kd/d = k$

Lie Factor = $k^2/k = k$

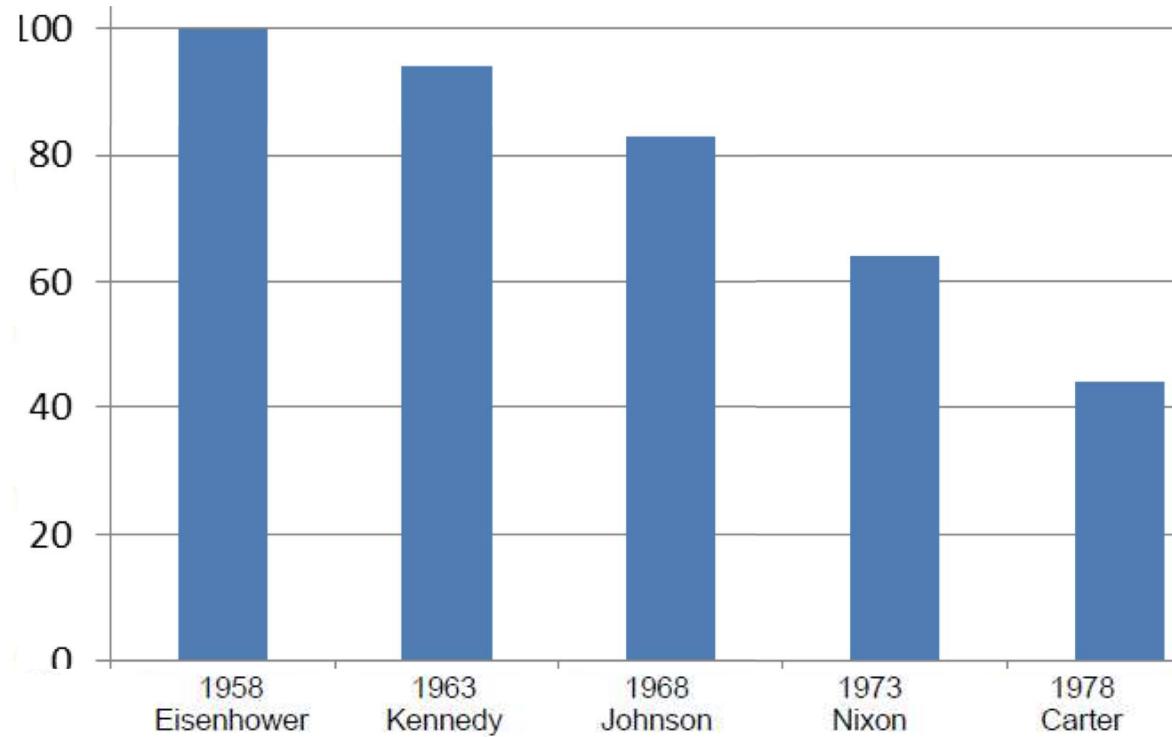
Lie factor = Data effect



Is the bottom dollar roughly half the size of the top one?

The same data with lie factor=1

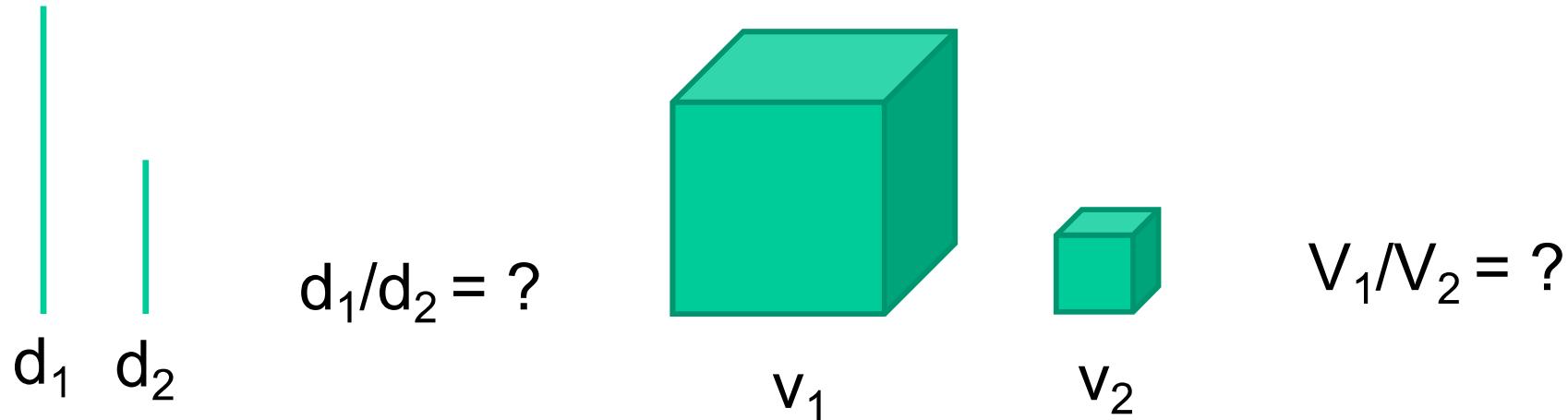
Note that in a histogram you are visually comparing **lengths**, not **areas**



This is why it is better to use thin bars...

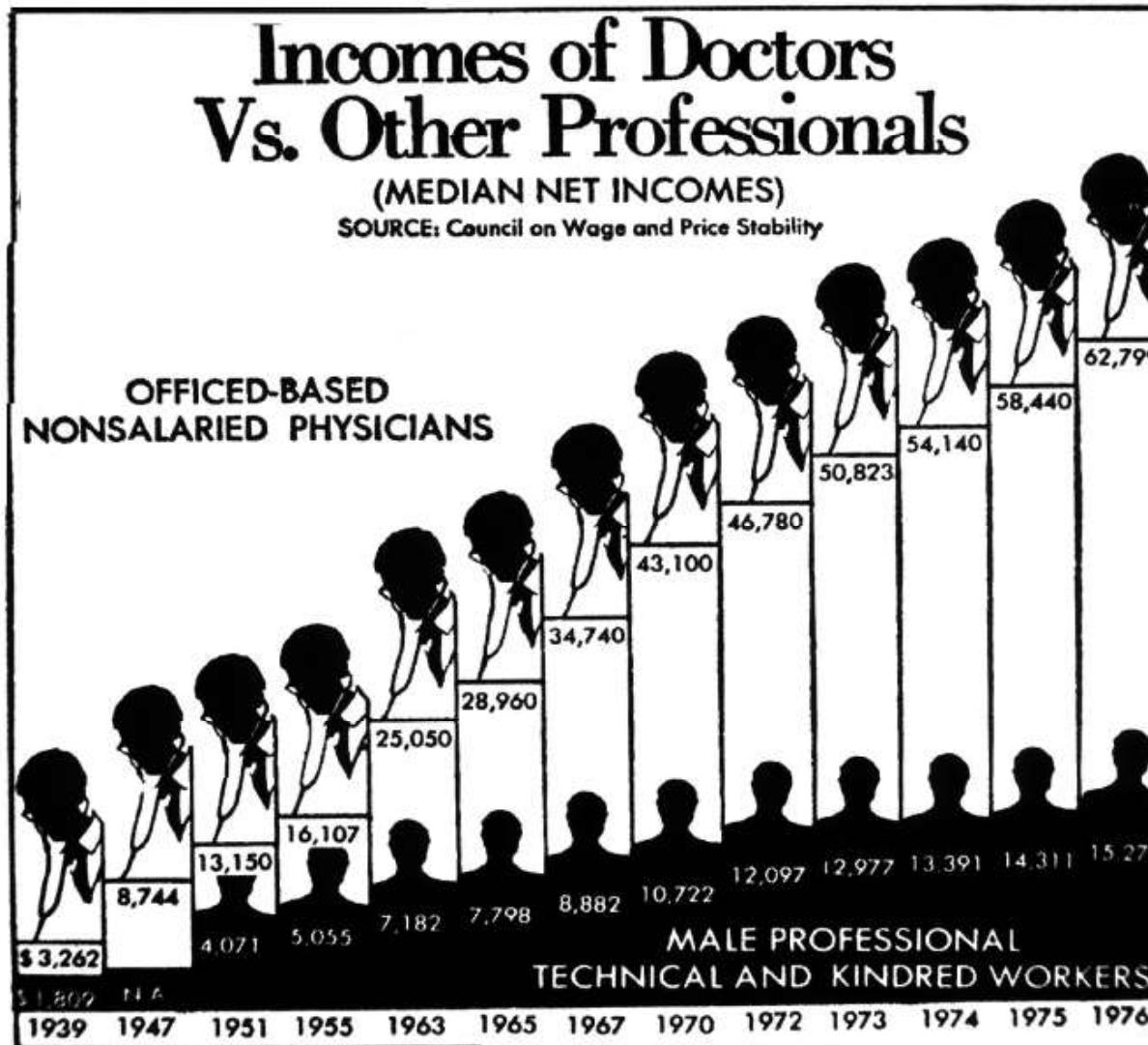
Encoding numerical values

- Human beings are better in comparing lengths than areas or volumes



- So, using volume or area **instead** of length is **wrong!**
- Or it is an intentional lie!

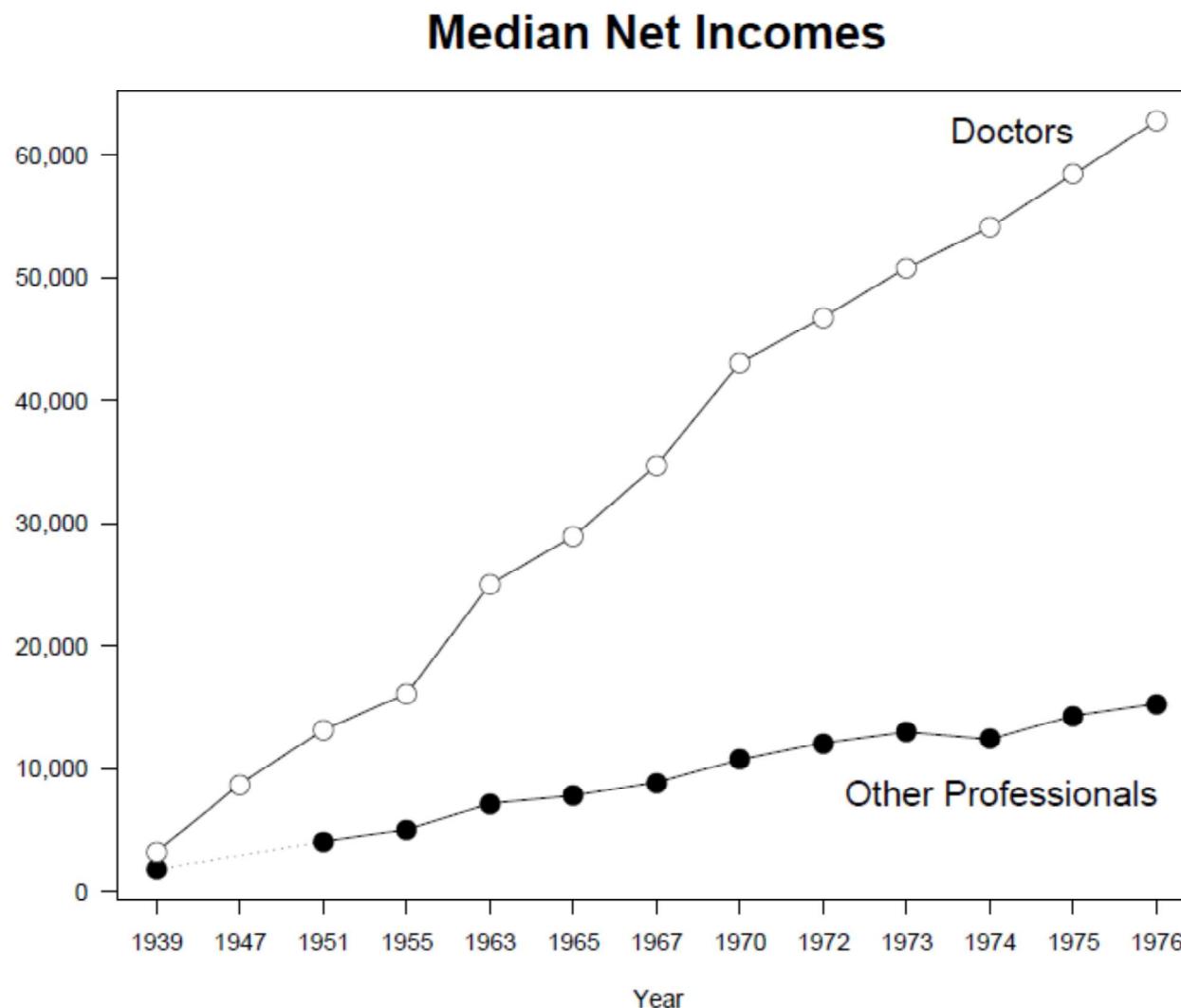
Distortion (deliberate?)



What's wrong
with this graph?

A part of the
chart junk

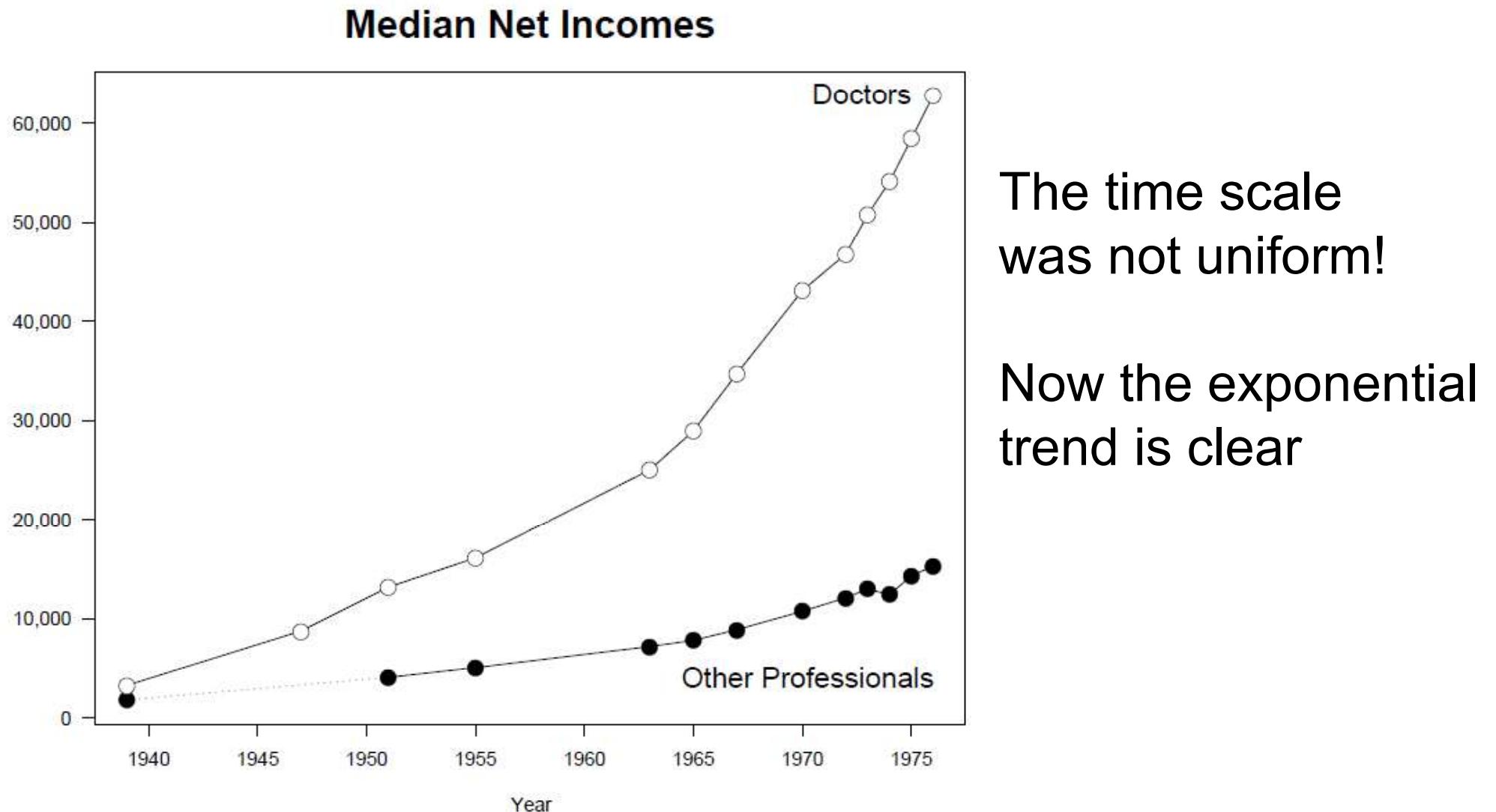
Presented data



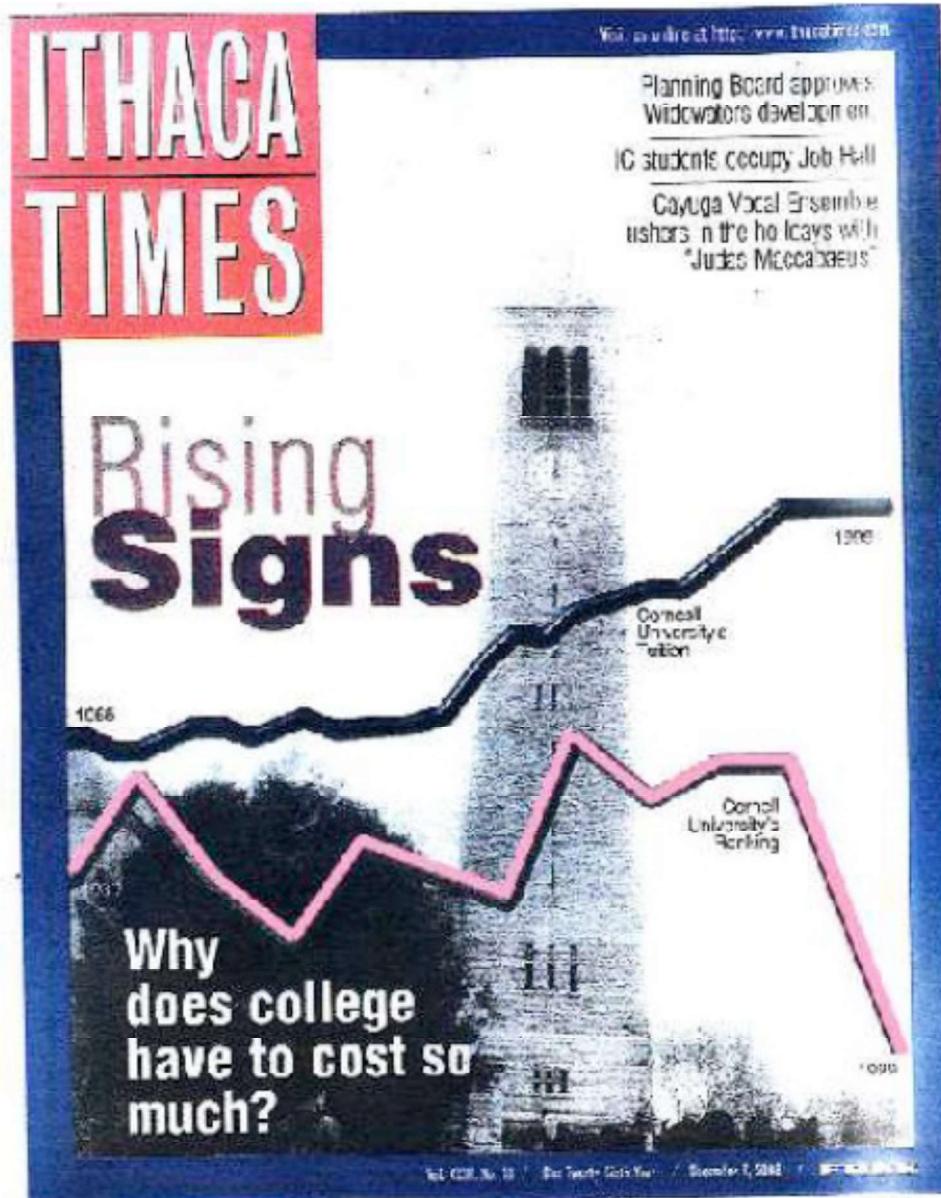
It suggests
a linear trend

What is wrong
with it?

Real data...



One of the best graph lie...

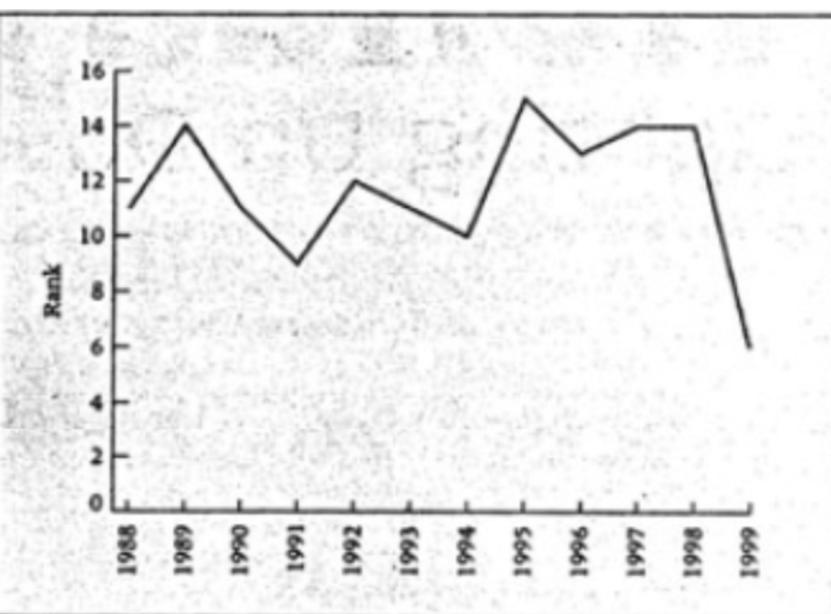


- The cover story, "Why does college have to cost so much?" shows a large graph superimposed on a scene from the Cornell campus. There are two jagged lines running across the graph
 - "Cornell's Tuition" = MONEY
 - "Cornell's Ranking"= QUALITY
- The Tuition graph shows a steady rise, and the ranking graph, after some early meandering, plummets to an all time low.
- The clear impression is that students are paying more for far less
- What is wrong with it?

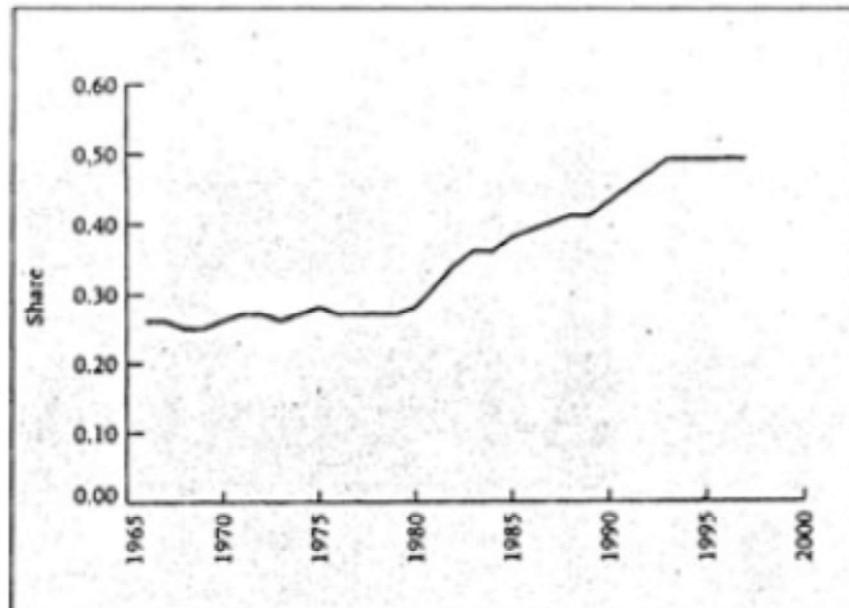
The lie

- More careful reading of the whole article (buried several pages into the paper) reveals a different story:
 - (1) The ranking graph covers an 11 year period, the tuition graph 35 years, yet they are shown simultaneously (the same apparent width) on the same horizontal "scale".
 - (2) The vertical scale for tuition and ranking could not possibly have common units, but the ranking graph is placed under the tuition graph creating the impression that cost exceeds quality.
 - (3) And here is the masterstroke: the sharp "drop" in the ranking graph over the past few years actually represents the fact that Cornell's rank has IMPROVED from 15th TO 6th ...

The real data



PECKING ORDER: OVER 12 YEARS, CORNELL'S RANKING IN US NEWS & WORLD REPORT HAS RISEN AND FALLEN ERRATICALLY.



BY THE NUMBERS: OVER 35 YEARS, CORNELL'S TUITION HAS TAKEN AN INCREASINGLY LARGER SHARE OF ITS MEDIAN STUDENT FAMILY INCOME.



Summarizing

- If the “story” is simple, keep it simple
- If the “story” is complex, make it look simple
- Tell the truth – don’t distort the data
 - (at least not by chance)