Elective in Software and Services
(Complementi di software e servizi per la società dell'informazione)

Section **Information Visualization**

Numbers of credit : 3

Giuseppe Santucci

2 – Visualizing numbers - Introduction

Thanks to Ross Ihaka

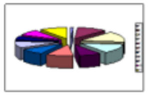
Outline

- An introductive example
- Good and bad graphs

Number visualization ?

- Obviously information visualization is, in general, about numbers

- In some cases, however, the numerical part is relevant, 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.)



Clustered Column. Compares values across categories.

- Actual software applications allows for easy (?) development of different typologies of charts

- I will discuss the basic relationships and the logical steps that allows for moving from quantitative data to suitable visualizations.

A starting example : a lotto game

- Forms of lotto 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
- Does this data contain information which will enable us to choose a profitable strategy for this game?
- We will use the results of 254 consecutive games to look for a profitable strategy

The data (254 values)

(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),

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** !!!

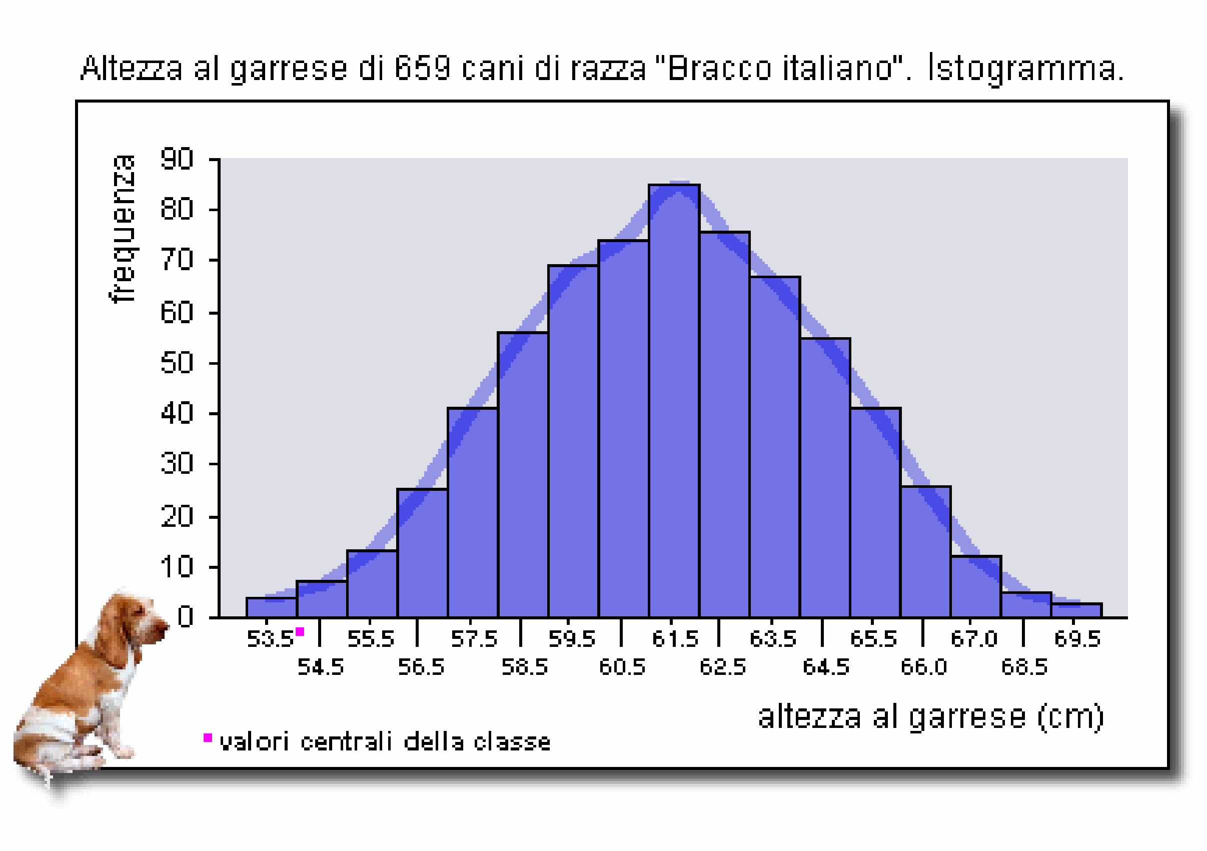
User task and visualization

- One approach to making money at “Pick It” is to try to select numbers which are more likely to win

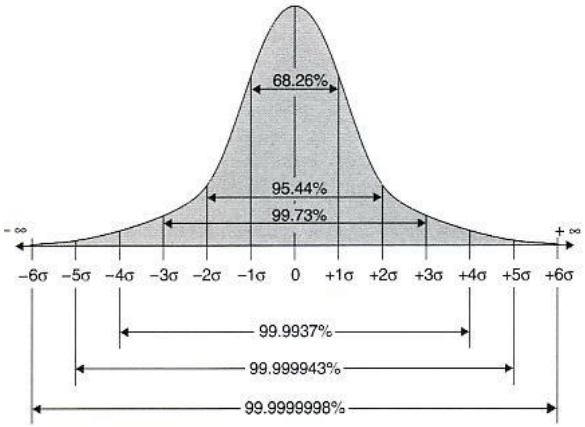
- Since we have data on the winning numbers we can look at the distribution of the winning numbers and see whether some ranges of values are more like to produce a winner than others

- One way to do this is to produce a histogram of the winning numbers

Histogram example

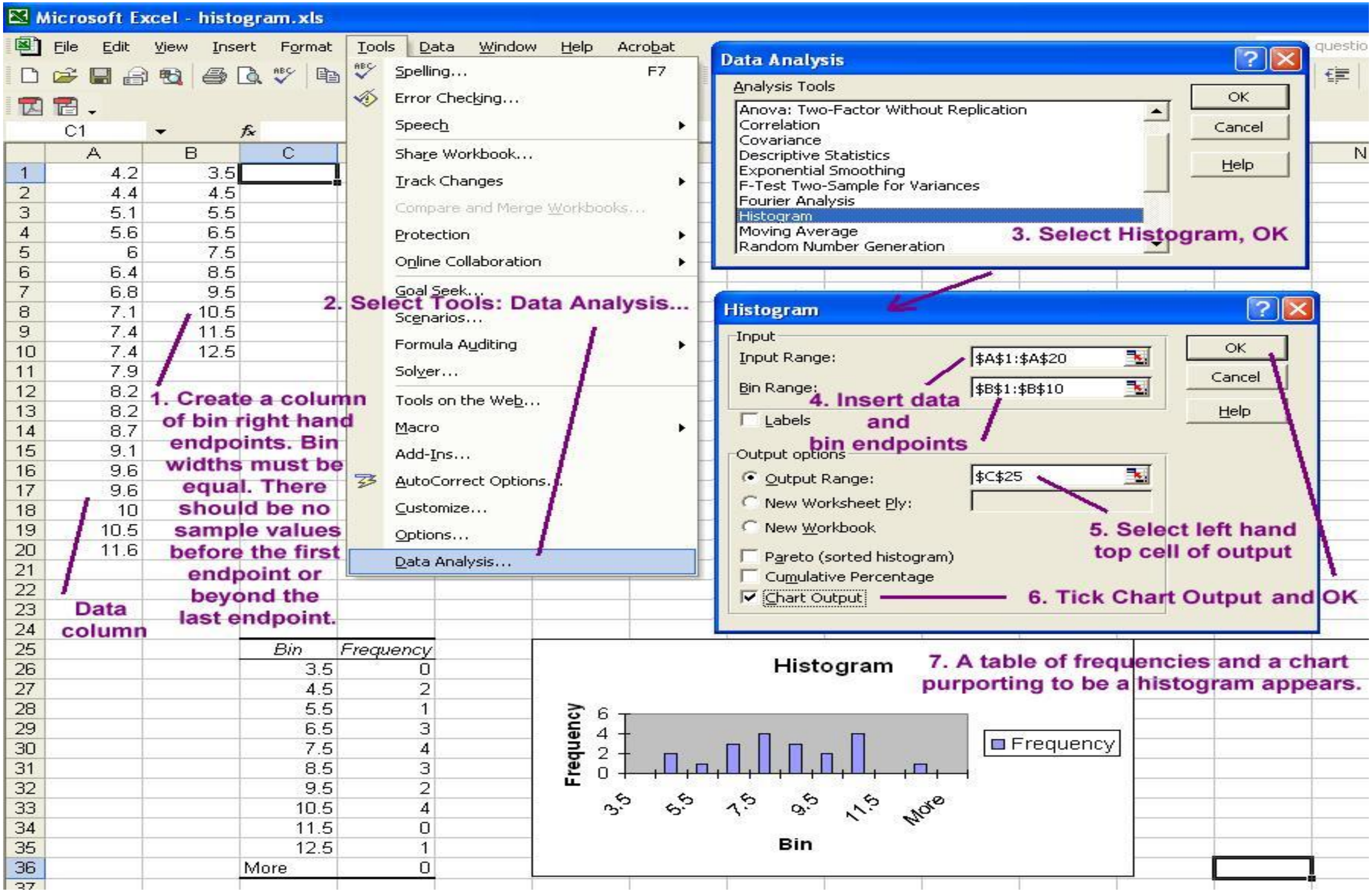


Altezza al garrese di 659 cani di razza "Bracco italiano". Istogramma.



bin

Excel and histograms



	A	B
1	4.2	3.5
2	4.4	4.5
3	5.1	5.5
4	5.6	6.5
5	6	7.5
6	6.4	8.5
7	6.8	9.5
8	7.1	10.5
9	7.4	11.5
10	7.4	12.5
11	7.9	
12	8.2	
13	8.2	
14	8.7	
15	9.1	
16	9.6	
17	9.6	
18	10	
19	10.5	
20	11.6	

Data column

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

Input Range: $A$1:$A$20

Bin Range: $B$1:$B$10

☐ Labels

Output options

● Output Range: $C$25

○ New Worksheet Ply:

○ New Workbook

☐ Pareto (sorted histogram)

☐ Cumulative Percentage

☑ Chart Output

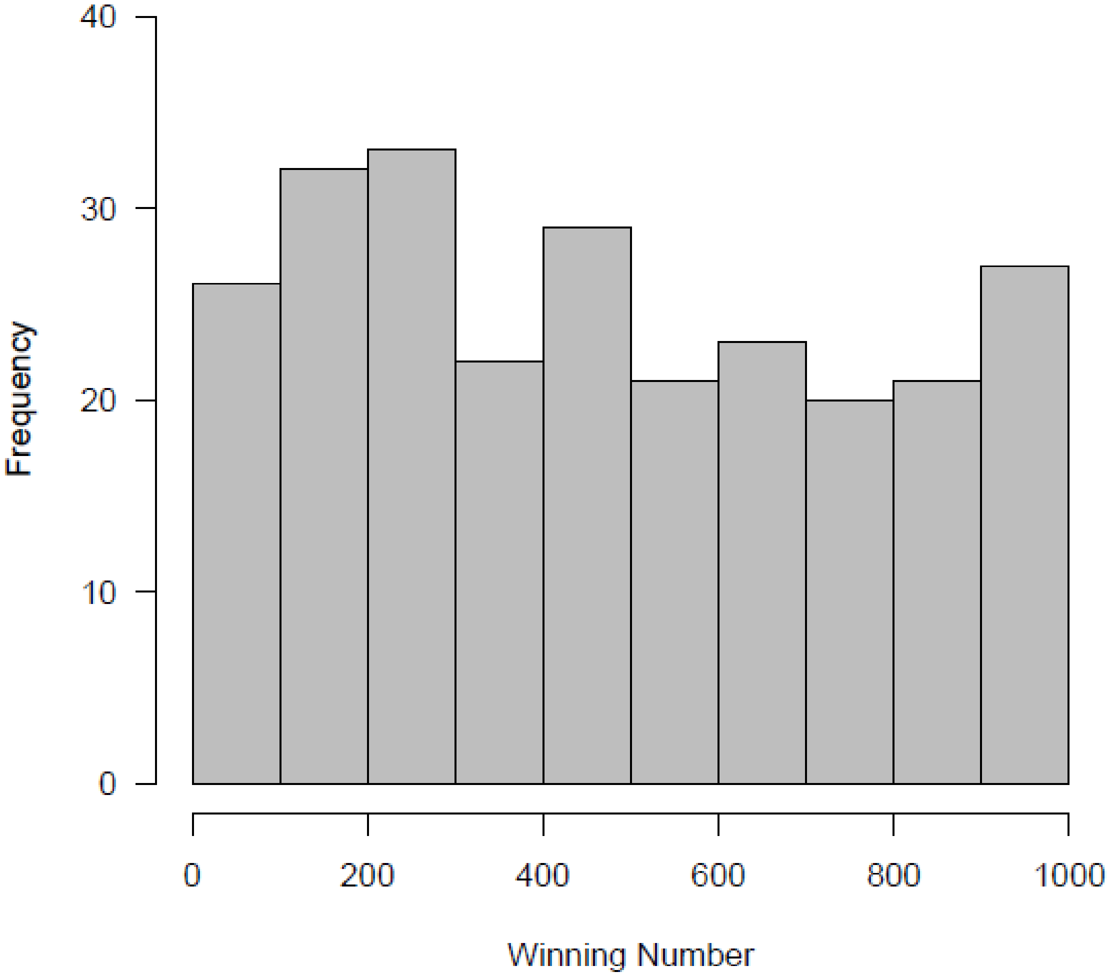
5. Select left hand top cell of output

6. Tick Chart Output and OK

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

Bin	*Frequency*
3.5	0
4.5	2
5.5	1
6.5	3
7.5	4
8.5	3
9.5	2
10.5	4
11.5	0
12.5	1
More	0

Data distribution



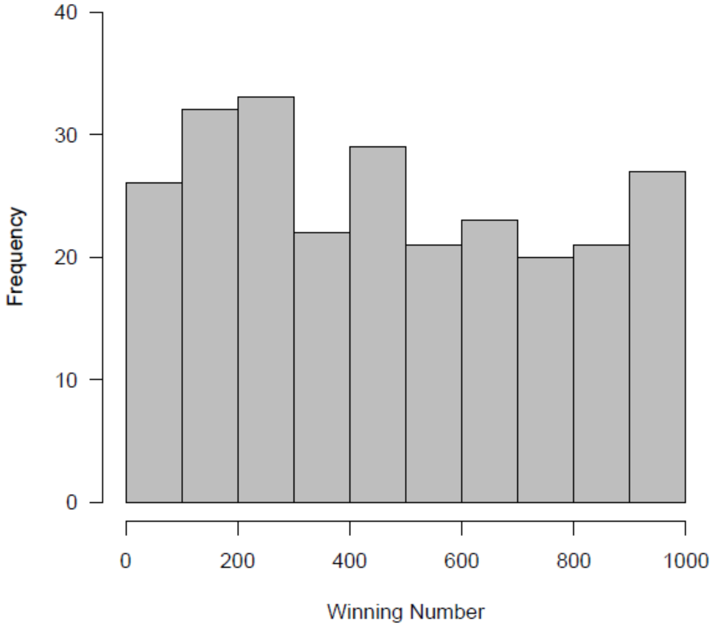
What can we infer from this histogram?

Is the bin size ok?

Analysis

- It looks there tend to be more winners in the region from 100 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...

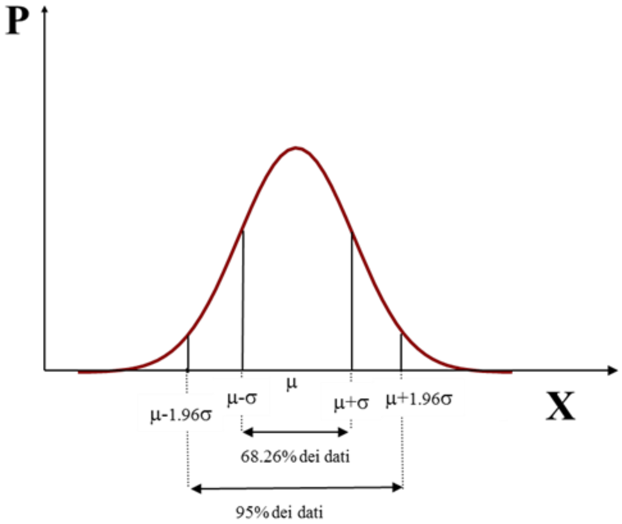
(wrong number understanding)

- Even if the winning numbers are chosen randomly, we can expect some “random variability” in a sample

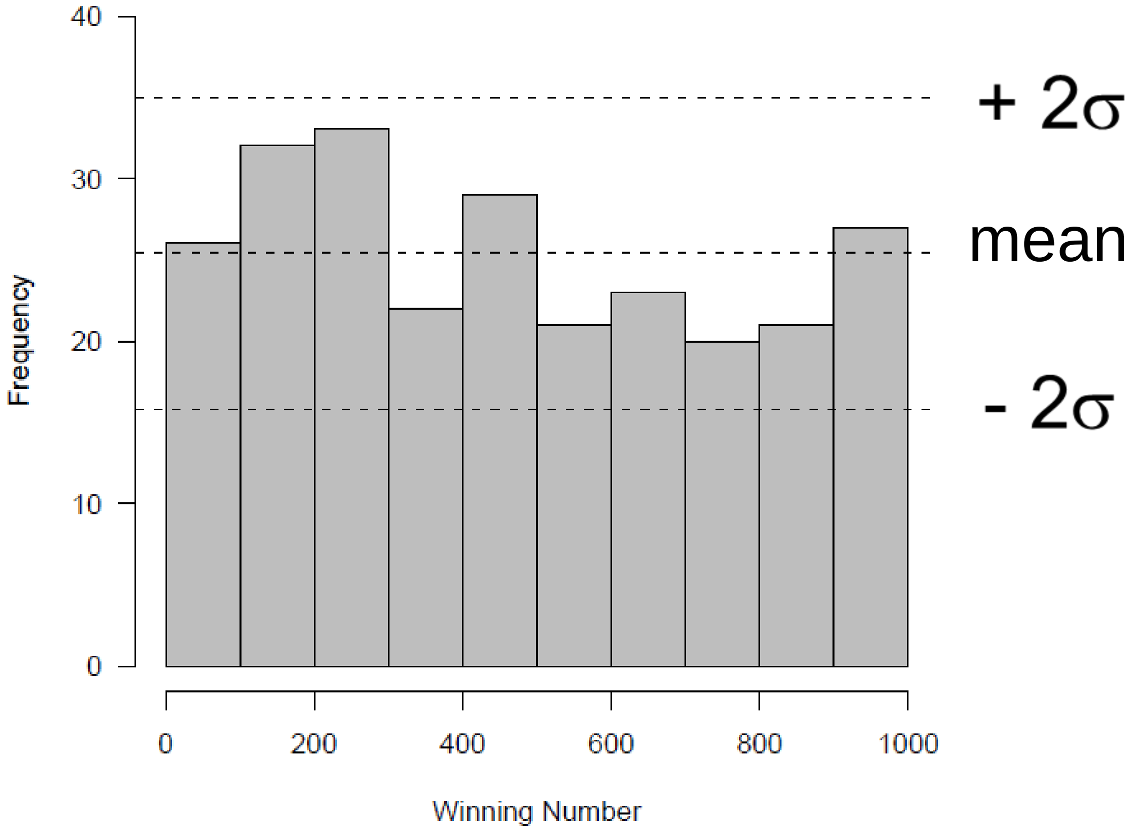
- To judge the significance of what we see in the histogram we have to recall some formal statistical theory

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 +/- 2σ



Better number visualization

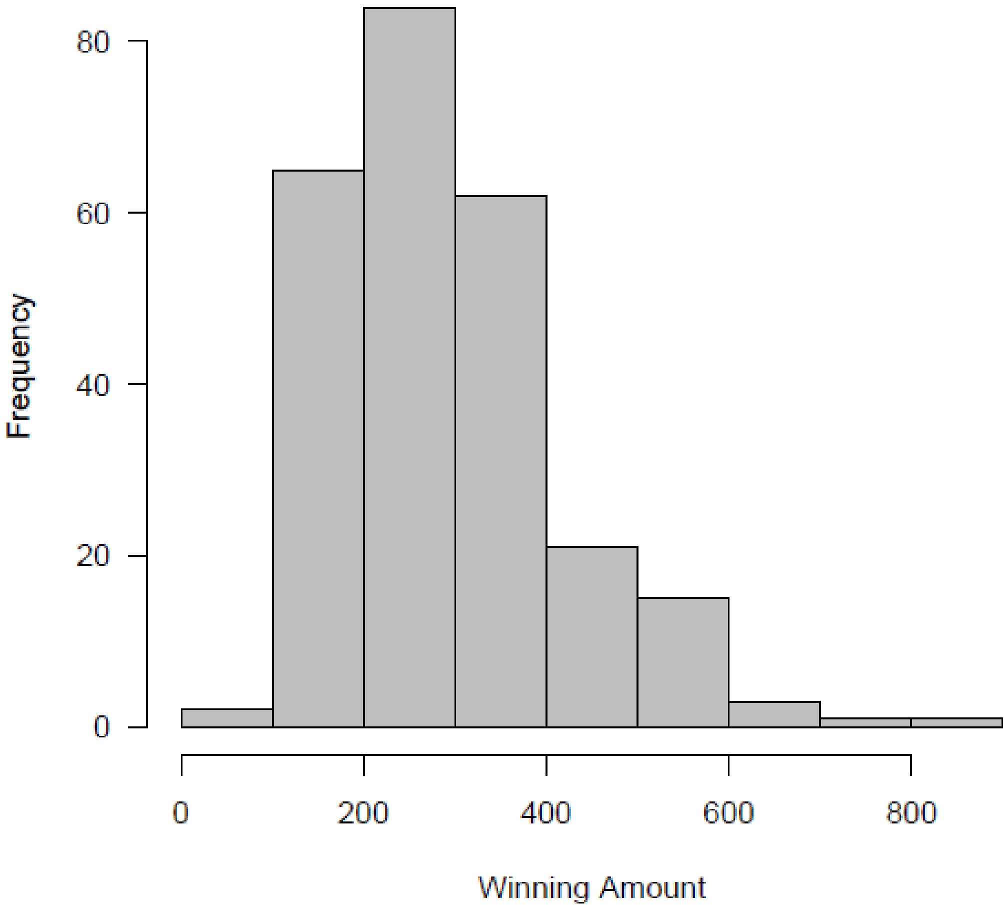


- Variance analysis AND visualization

Conclusions and new task

- Winning numbers are totally random
- It makes no sense to look for a " lucky " number
- However, we can change our task:
  - to increase the amount won !
- So we study the distribution of winning amount

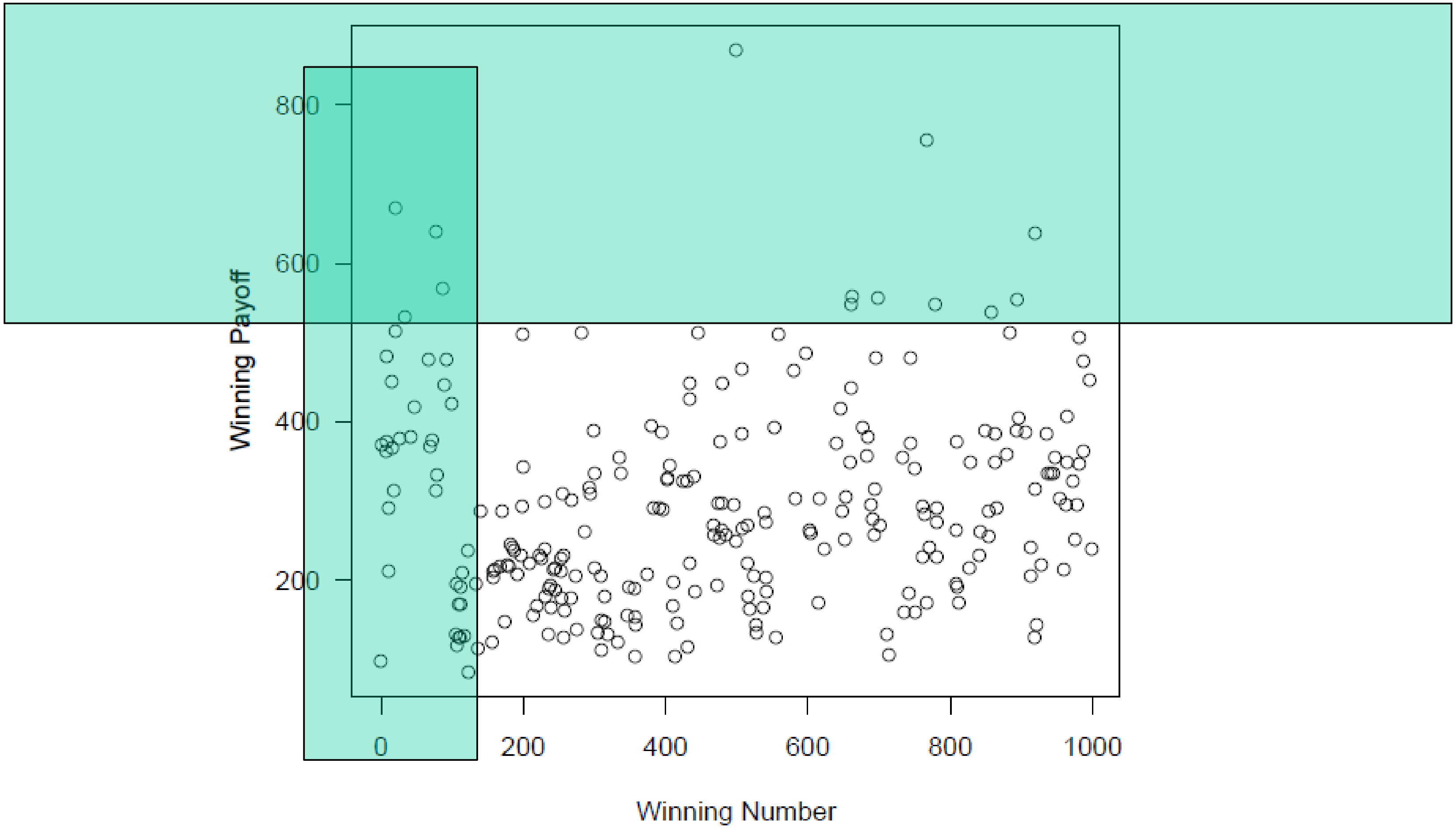
New visualization



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.

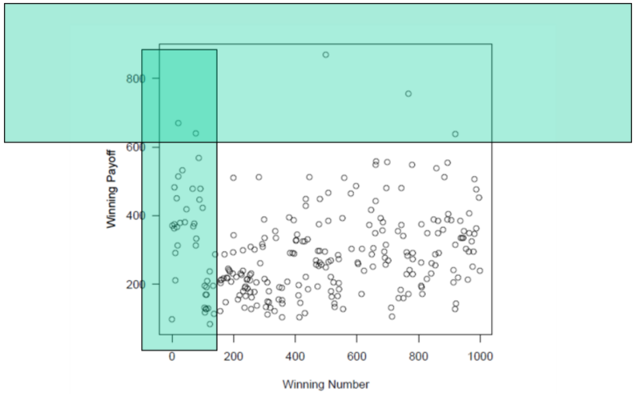
New visualization



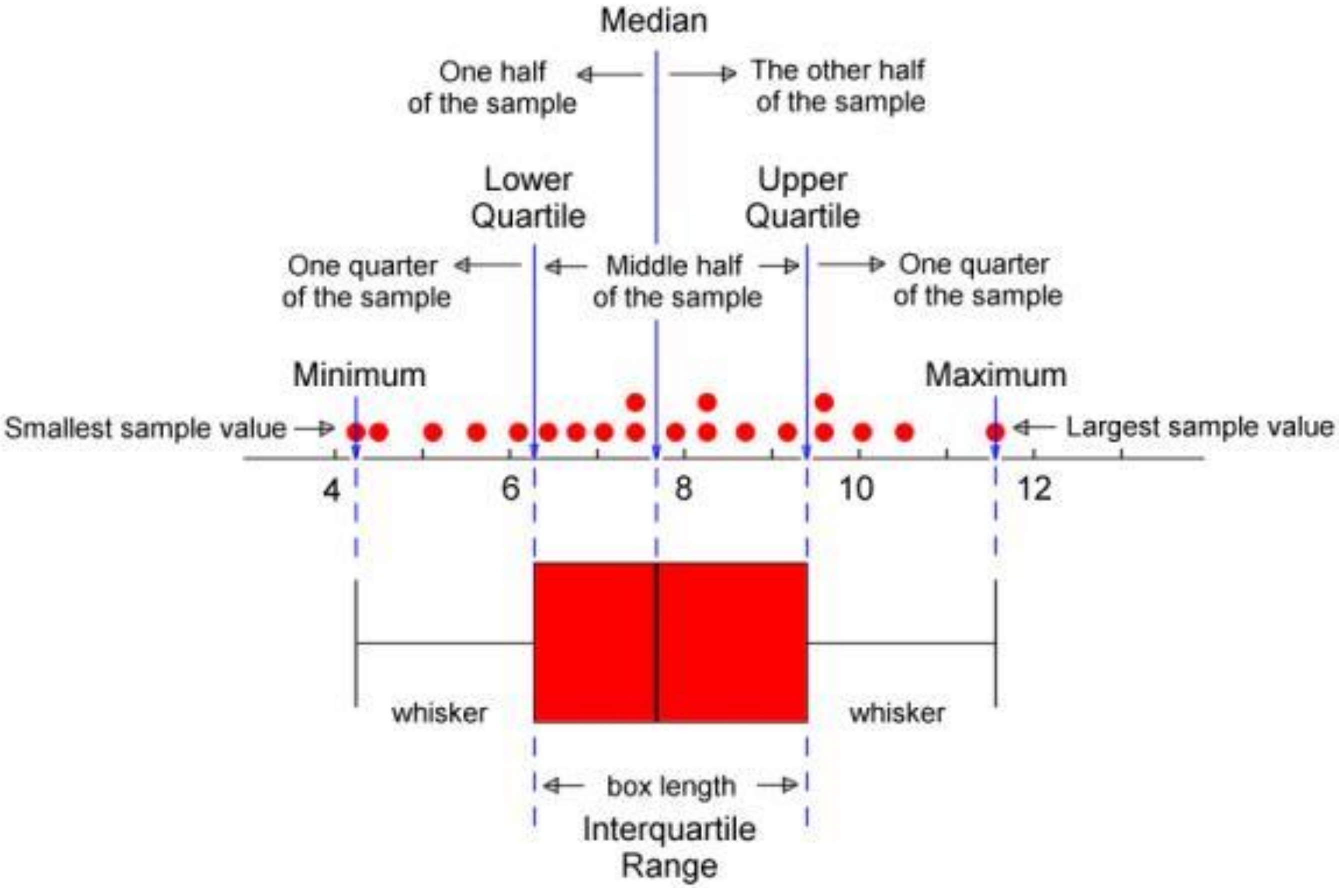
Insights from the scatterplot

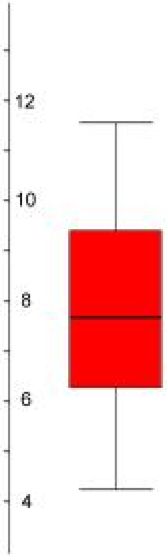
- The winning amounts in a band to the left of the plot appear to generally be higher than those in the rest of the plot

- 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

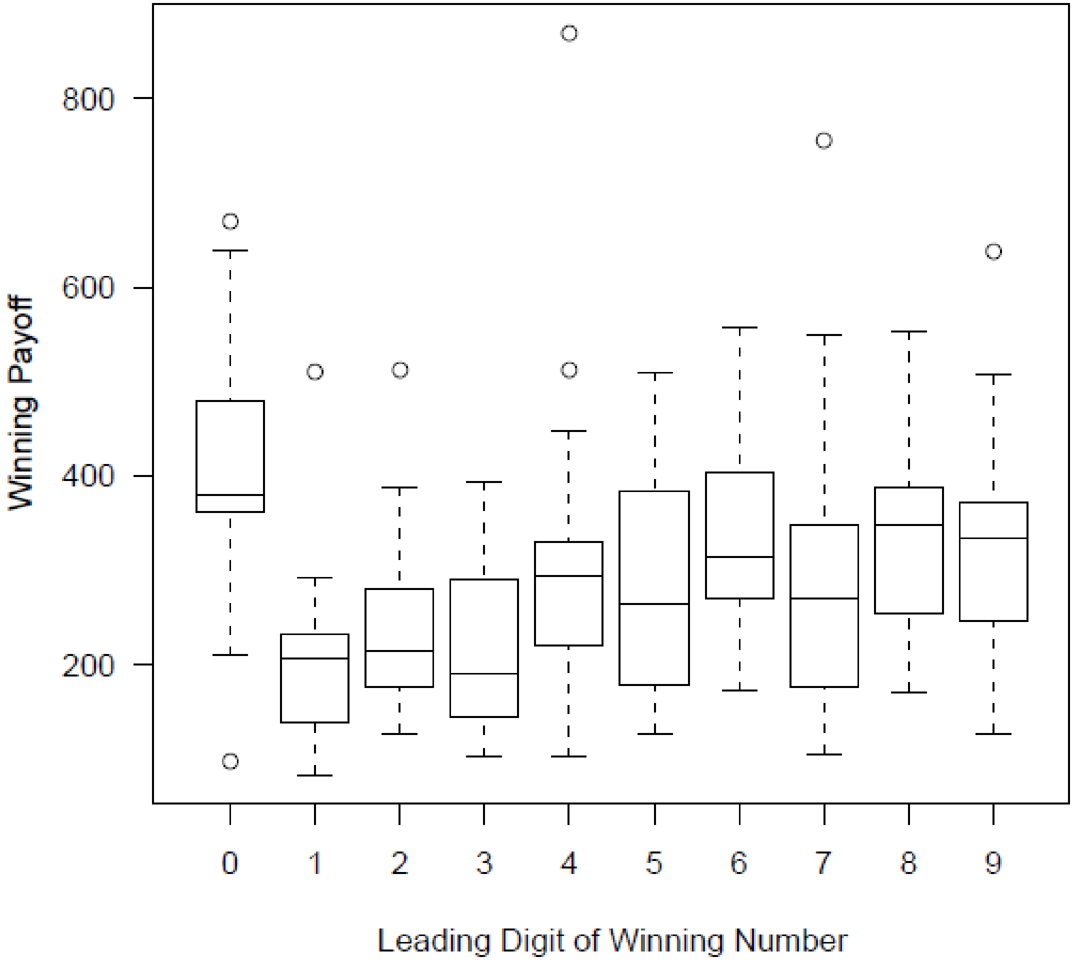


Boxplot


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
Middle half of the sample
One quarter of the sample
Minimum
Maximum
Smallest sample value
Largest sample value
4
6
8
10
12
whisker
whisker
box length
Interquartile Range

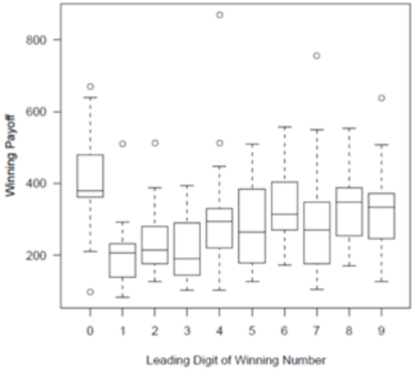

12
10
8
6
4

Lottery's boxplots

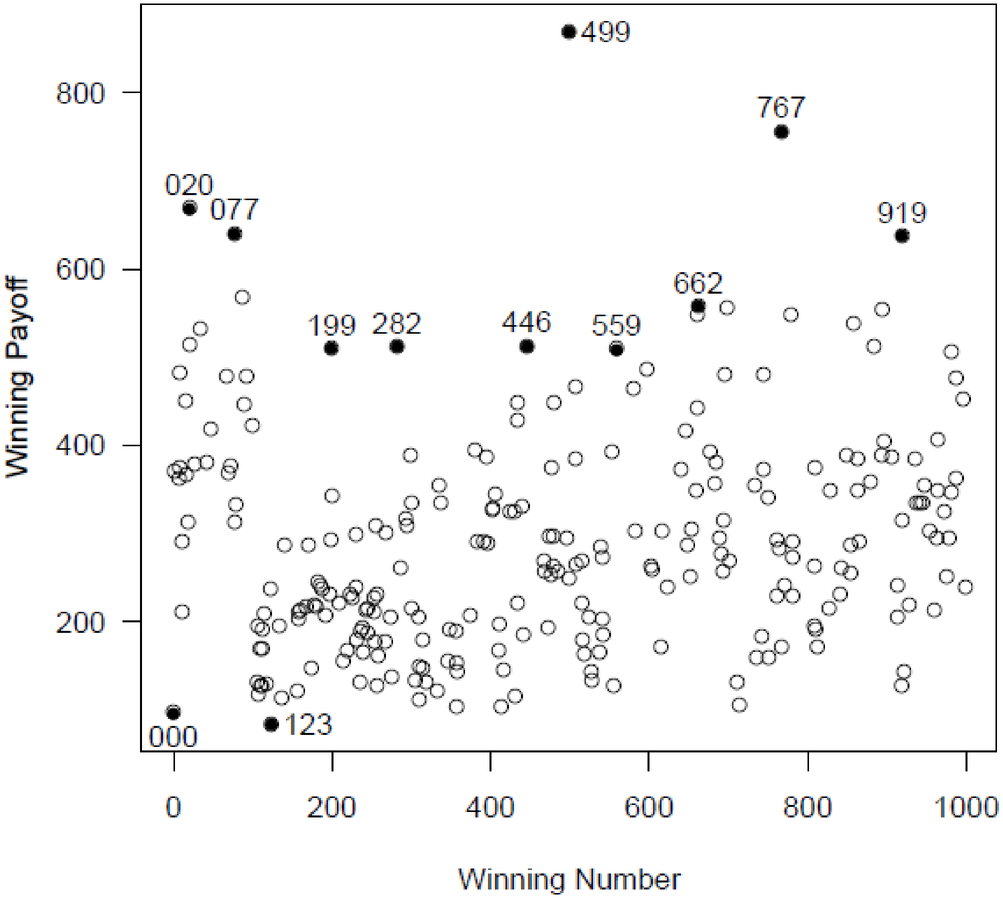


New insights

- 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, 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 introductive 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

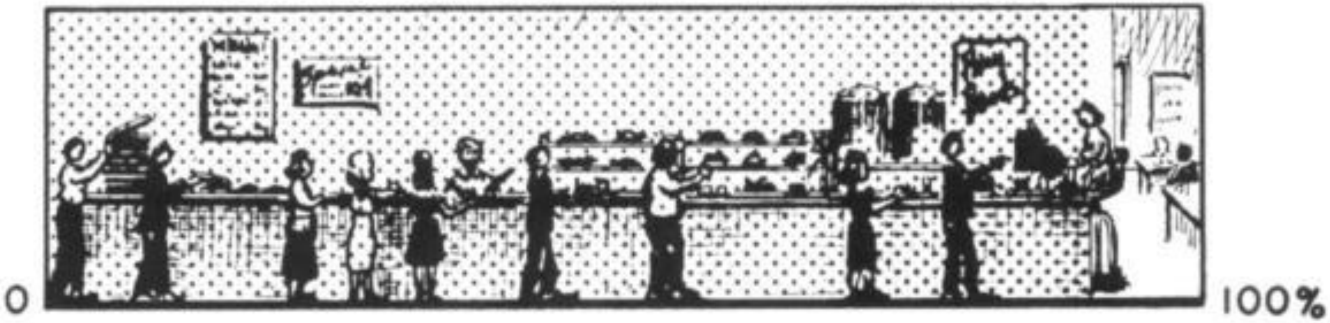
- Next lectures will cover these issues in a more formal way

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

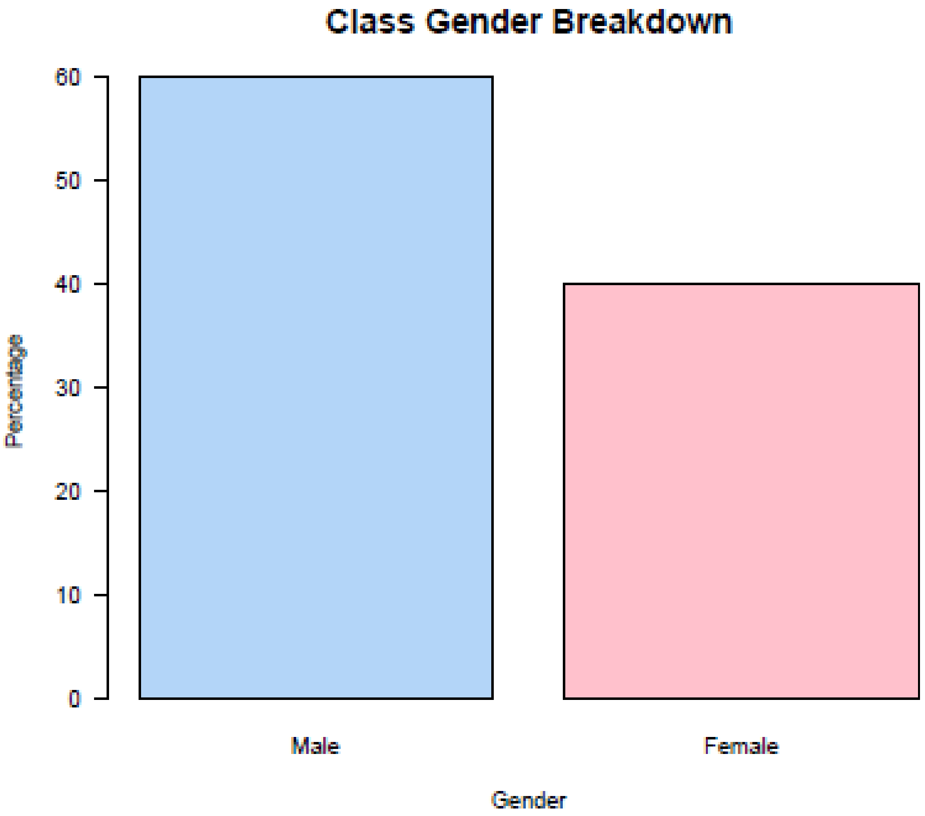
Rule 0 violation (and also rule 2)

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



Source: COMPANY REPORTS

Rule 0 violation

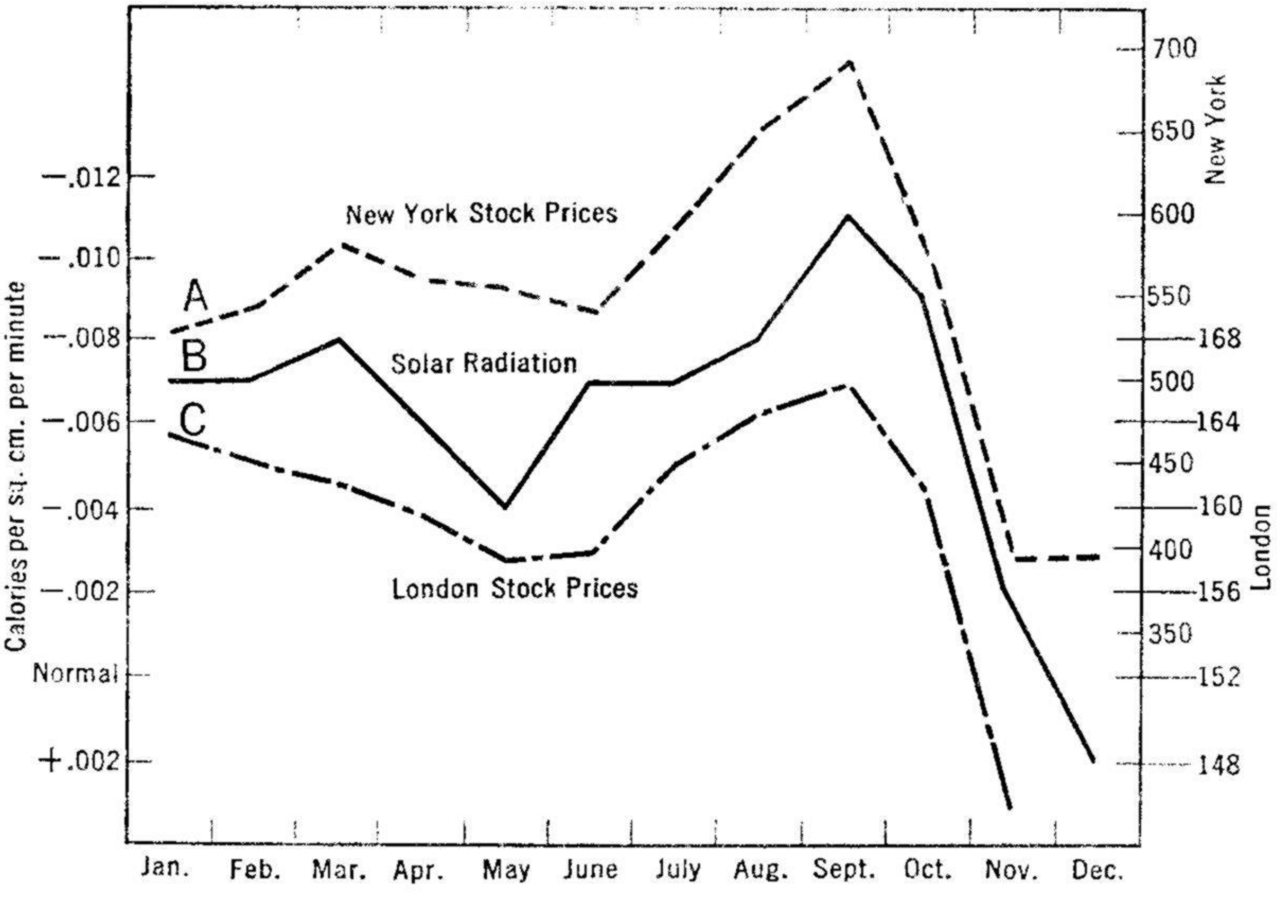


Male	60%
Female	40%

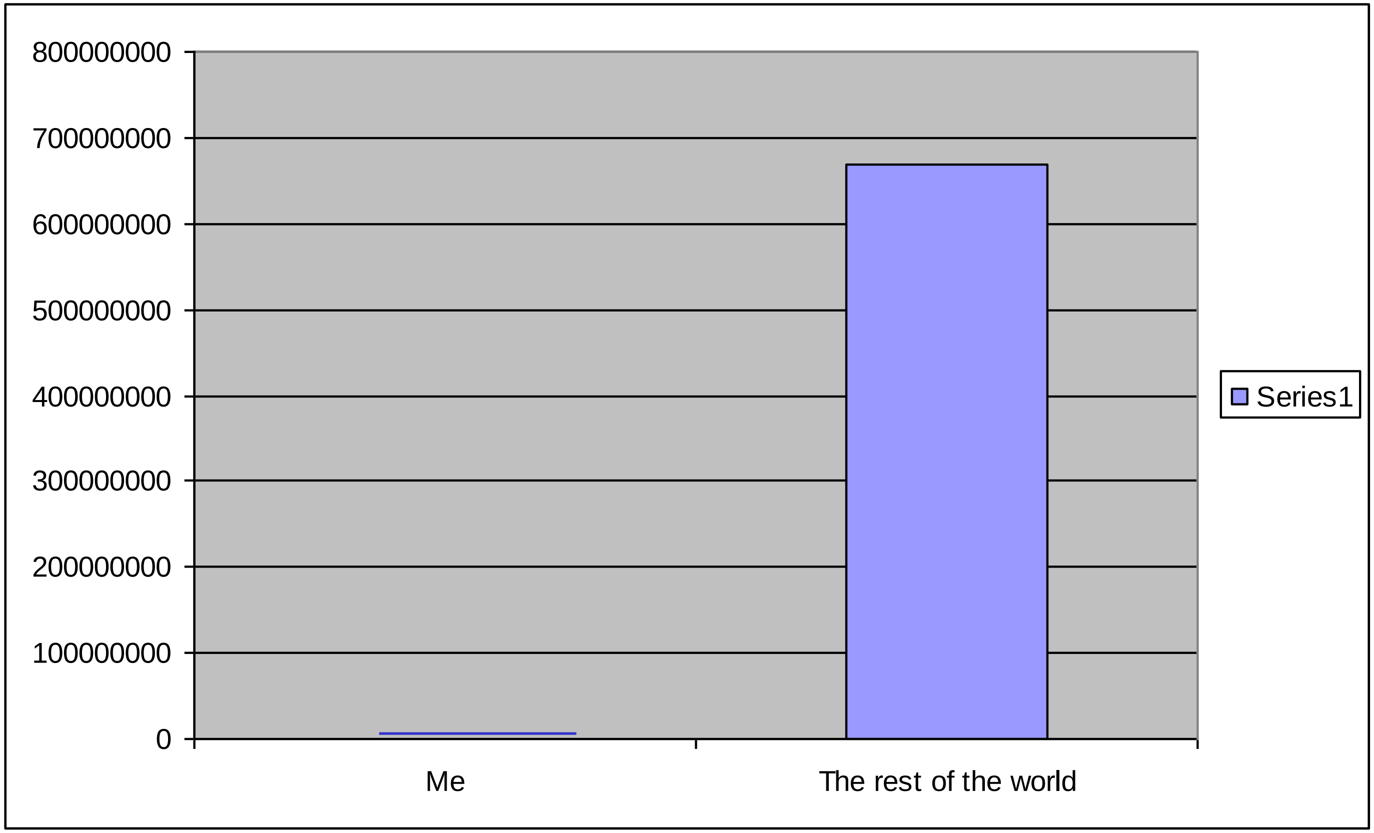
Role 1

- **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 good example of distortion

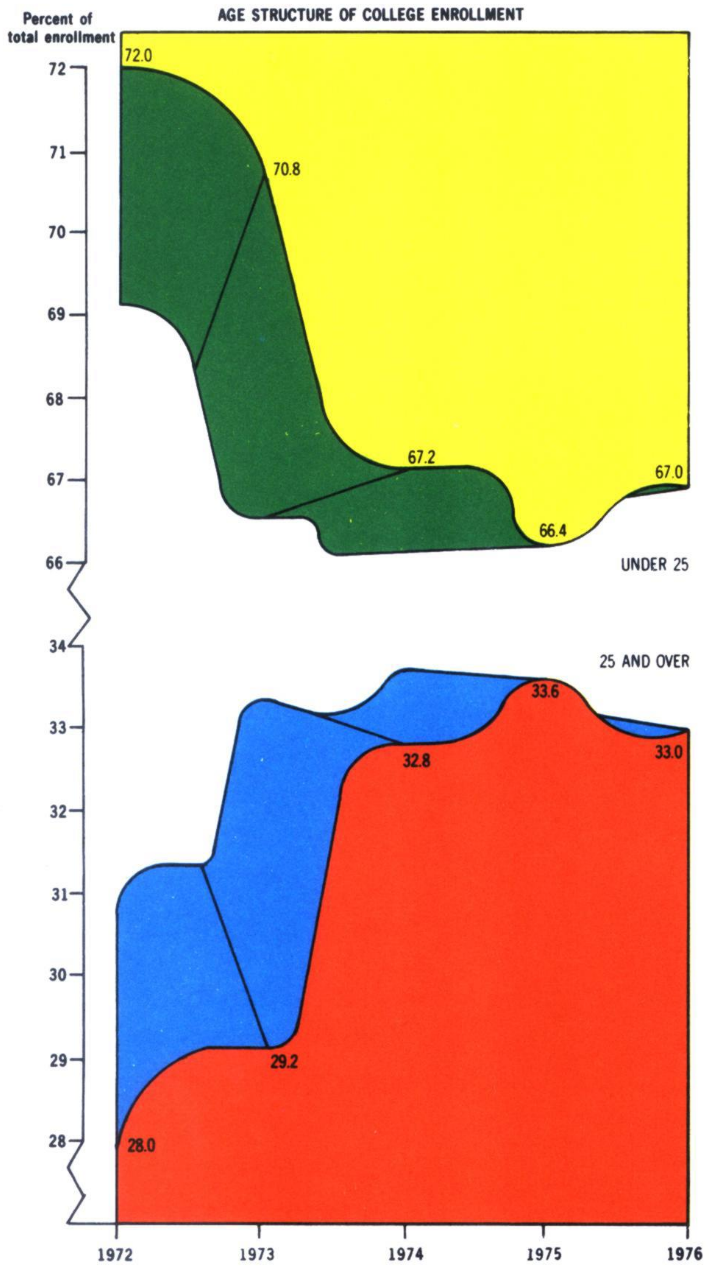
Rule 2: Insure chart simplicity

- Graphs should be no more complex than the data which they portray

- Unnecessary complexity can be introduced by
  - irrelevant decorations
  - colors
  - 3d effects
  - ...
- These are collectively known as “chartjunk”

- For a very comprehensive set of chartjunk effects look at Microsoft Excel
  - the later the version the larger the set !

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

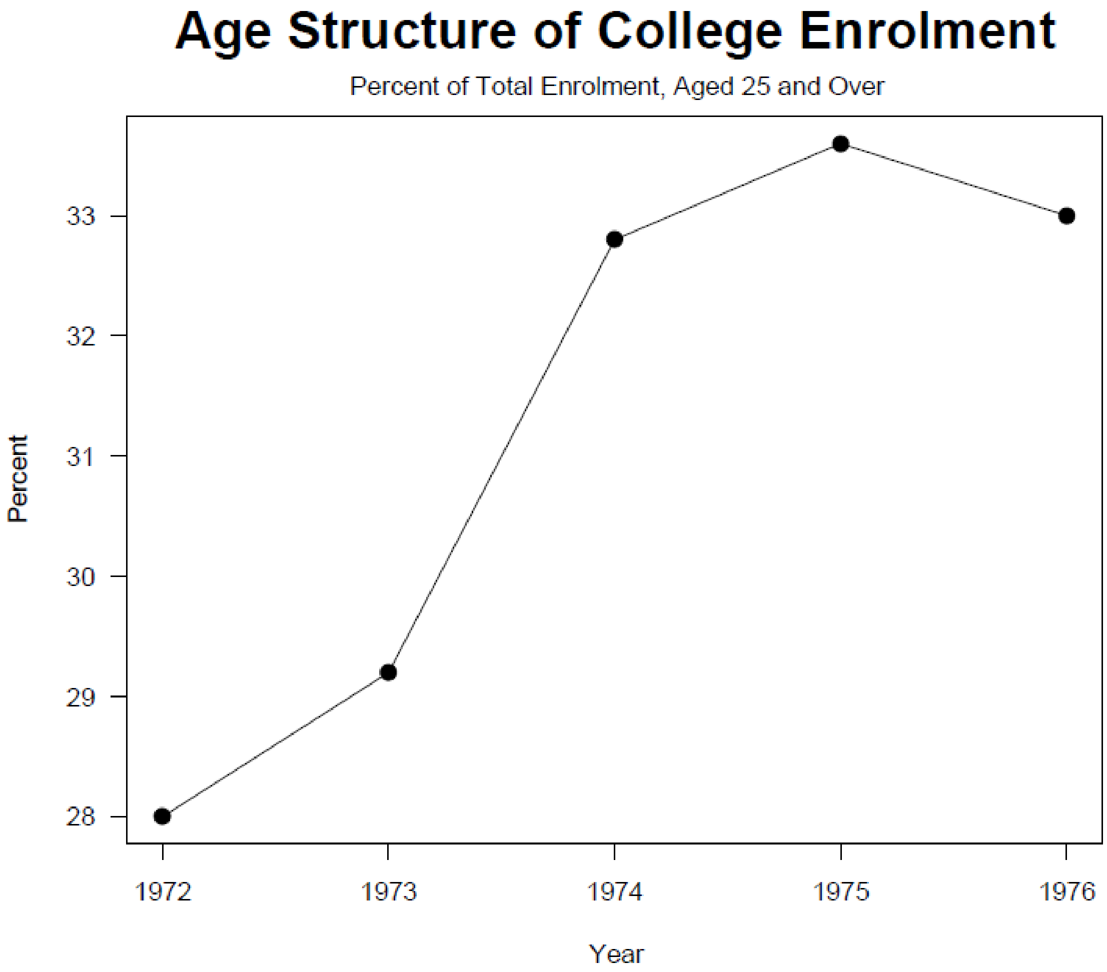


Role 2 violation (and also rule 3)

- 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
  - random interpolation
- Designers of this graph are now working in the Microsoft Excel's team, inspiring the new Excel's versions ...

American Education Magazine

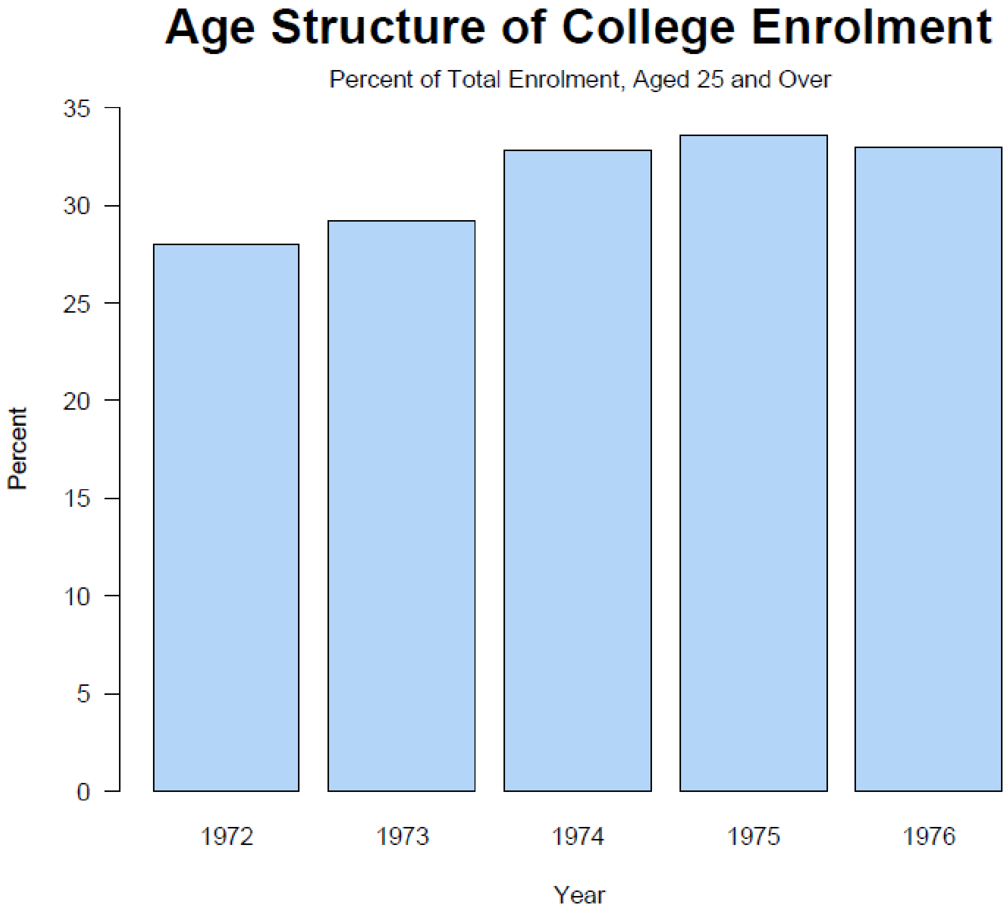
Same data...


Age Structure of College Enrolment
Percent of Total Enrolment, Aged 25 and Over
Percent
33
32
31
30
29
28
1972
1973
1974
1975
1976
Year

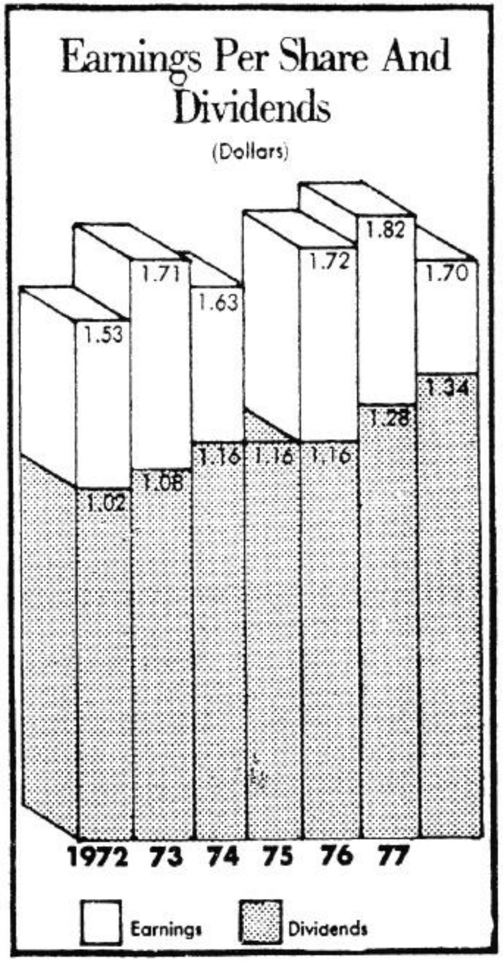
The same data...

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

Same data...


Age Structure of College Enrolment
Percent of Total Enrolment, Aged 25 and Over
Percent
0
5
10
15
20
25
30
35
1972
1973
1974
1975
1976
Year

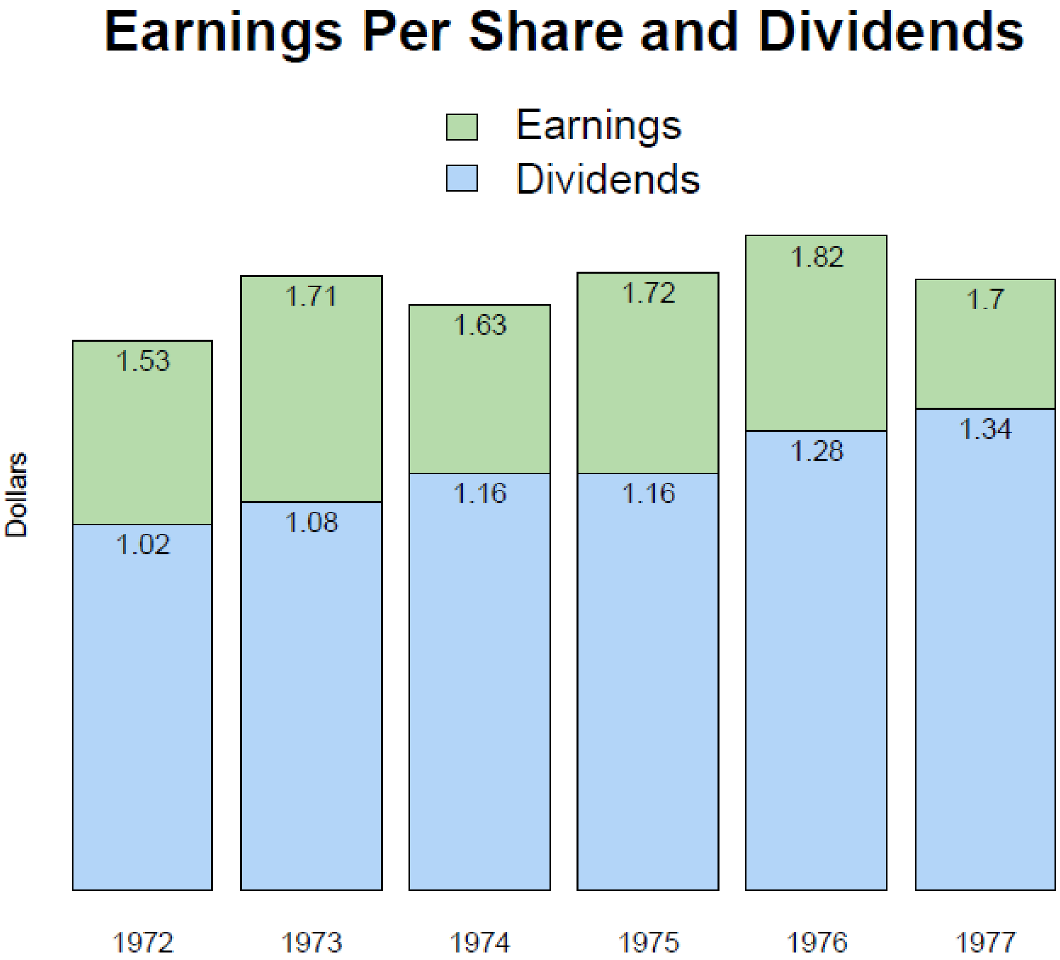
Role 2 violation



- Why 3D?
- The extra dimension used in this graph has confused even the person who created it..

The Washington Post, 1979

The same data...

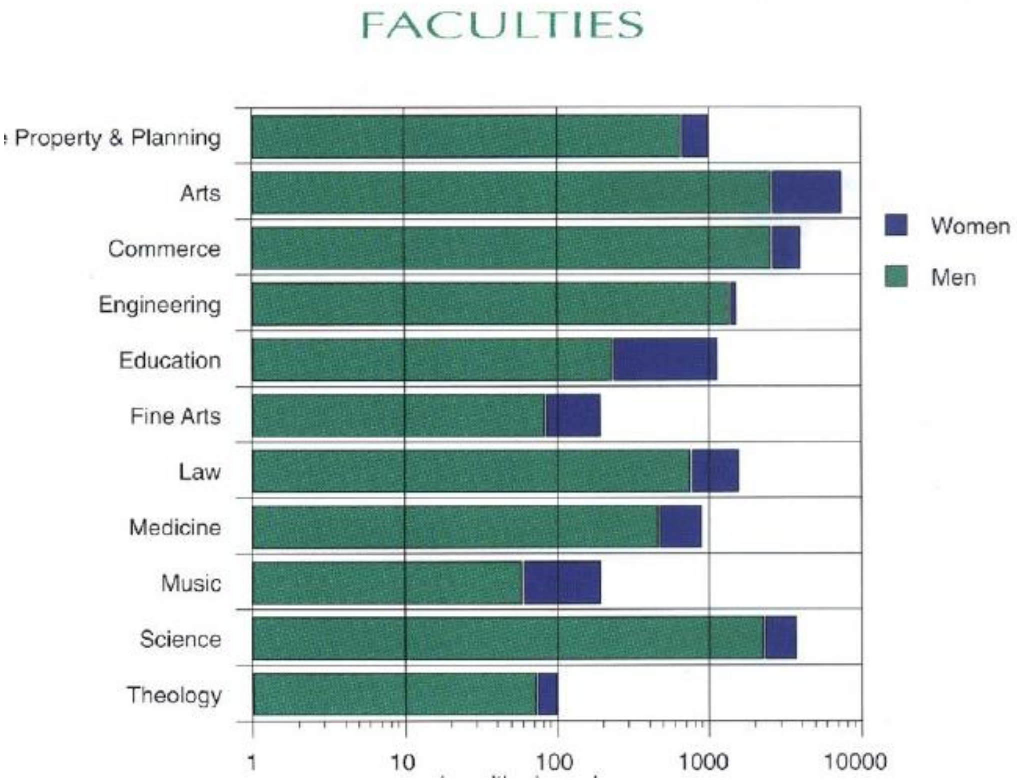


Role 3

- **Do not distort data in a confusing way**

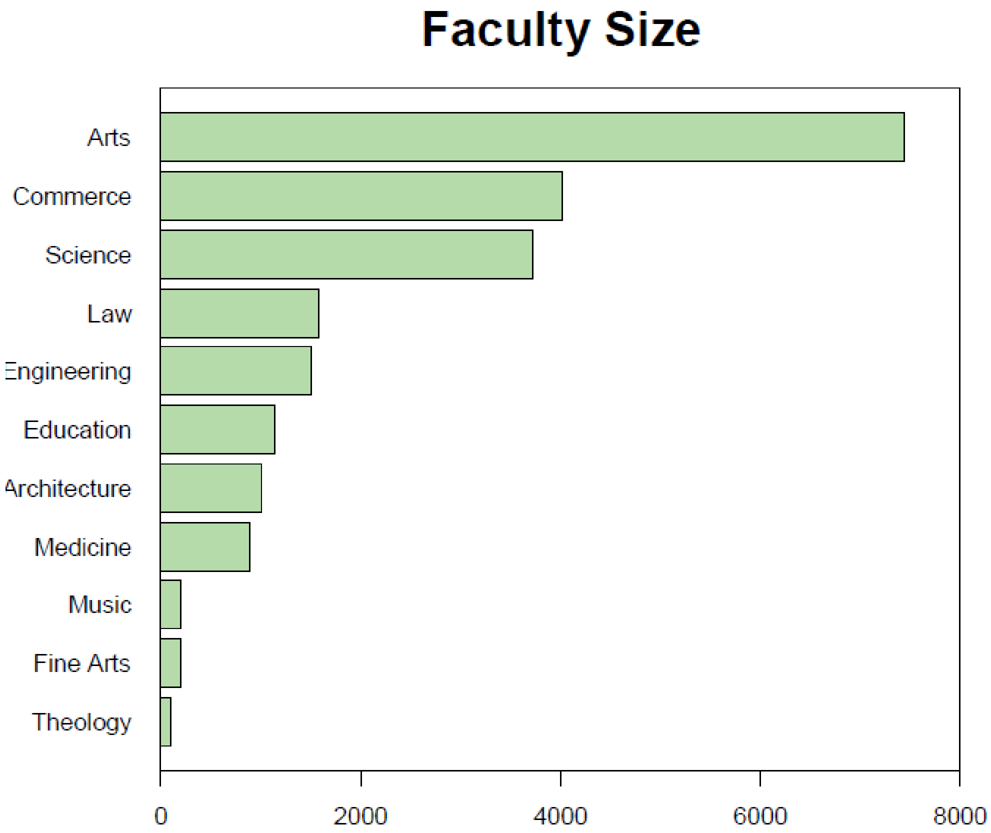
- Graphs should not provide a distorted picture of the values they portray
- Distortion can be either deliberate or accidental
- Of course, it could be useful to know how to produce a graph which bends the truth...

Role 3 violation

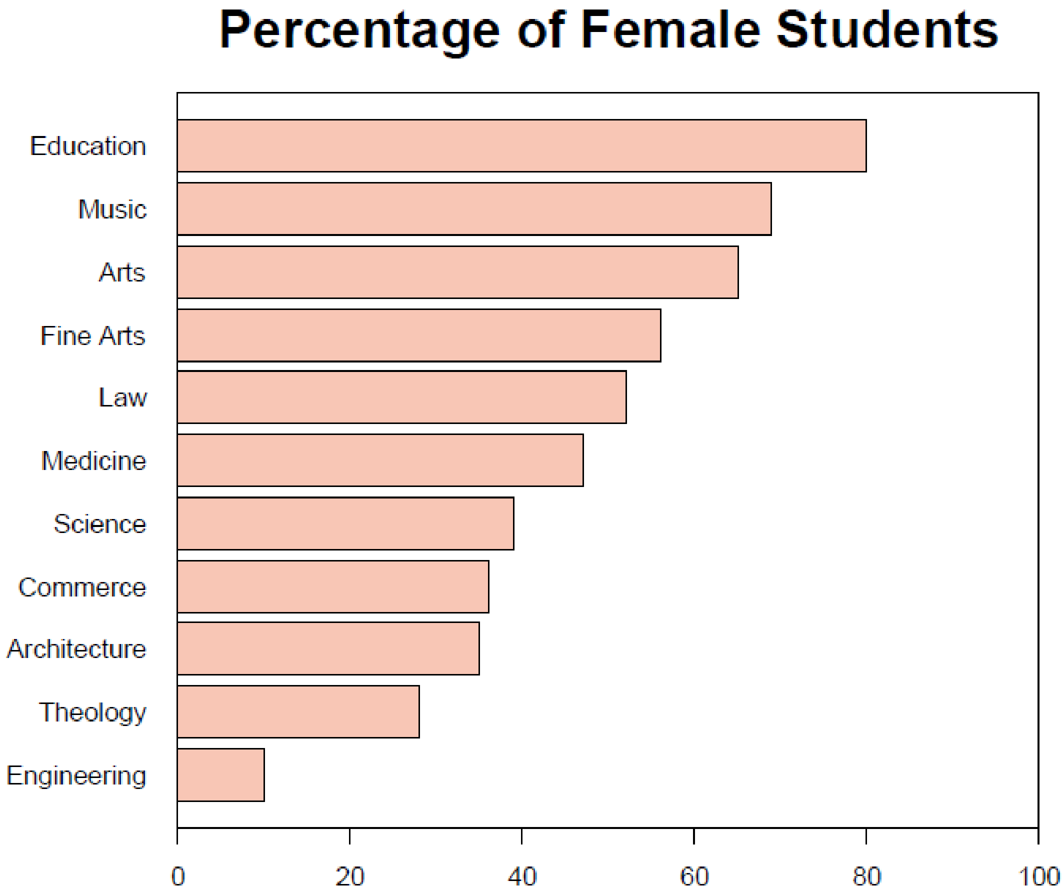


- At a very quick glance:
  - balanced faculty population
  - most male students
- The X scale is logarithmic!

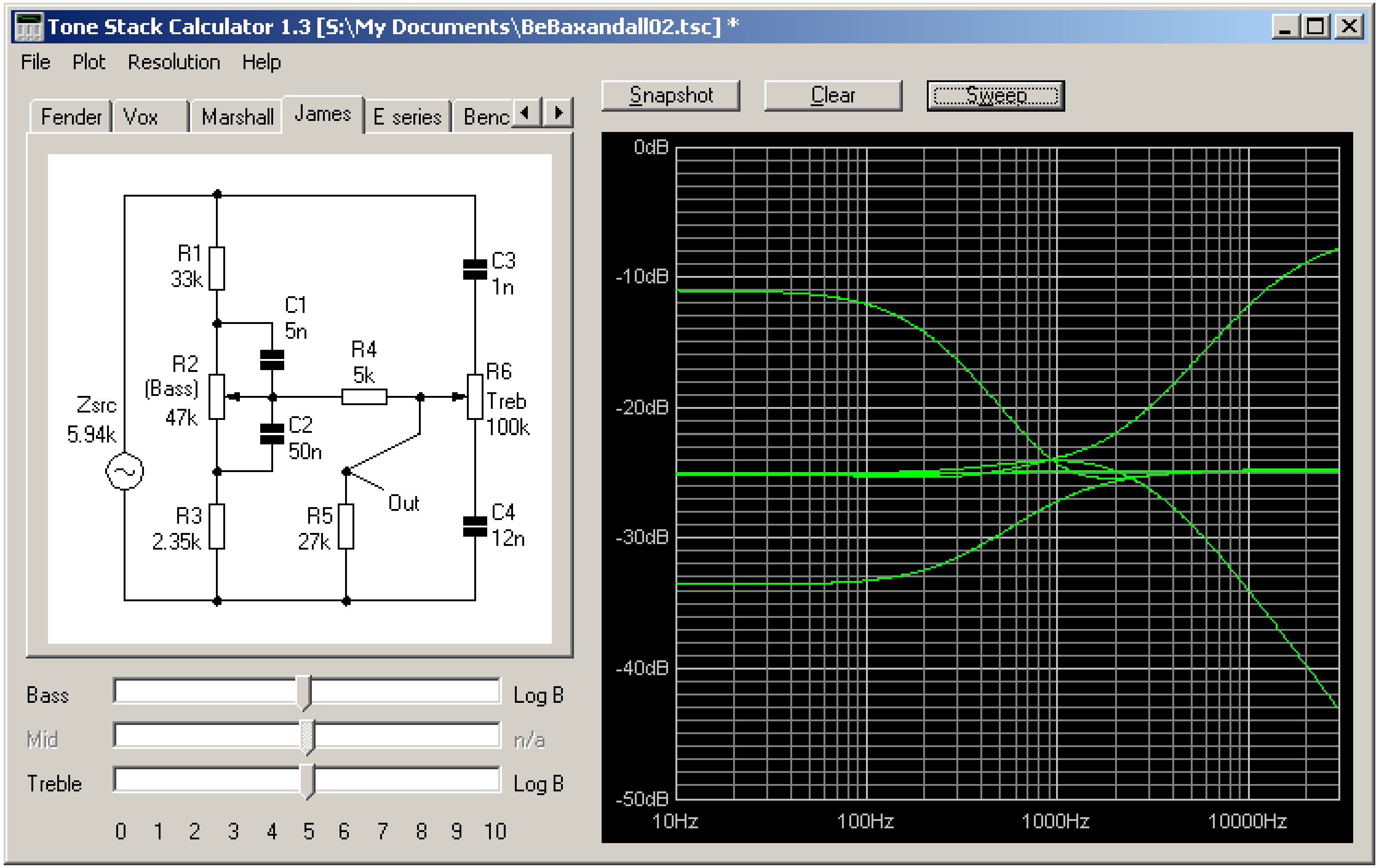
The truth : population size



The truth : female /male ratio



In other cases distortion is ok...



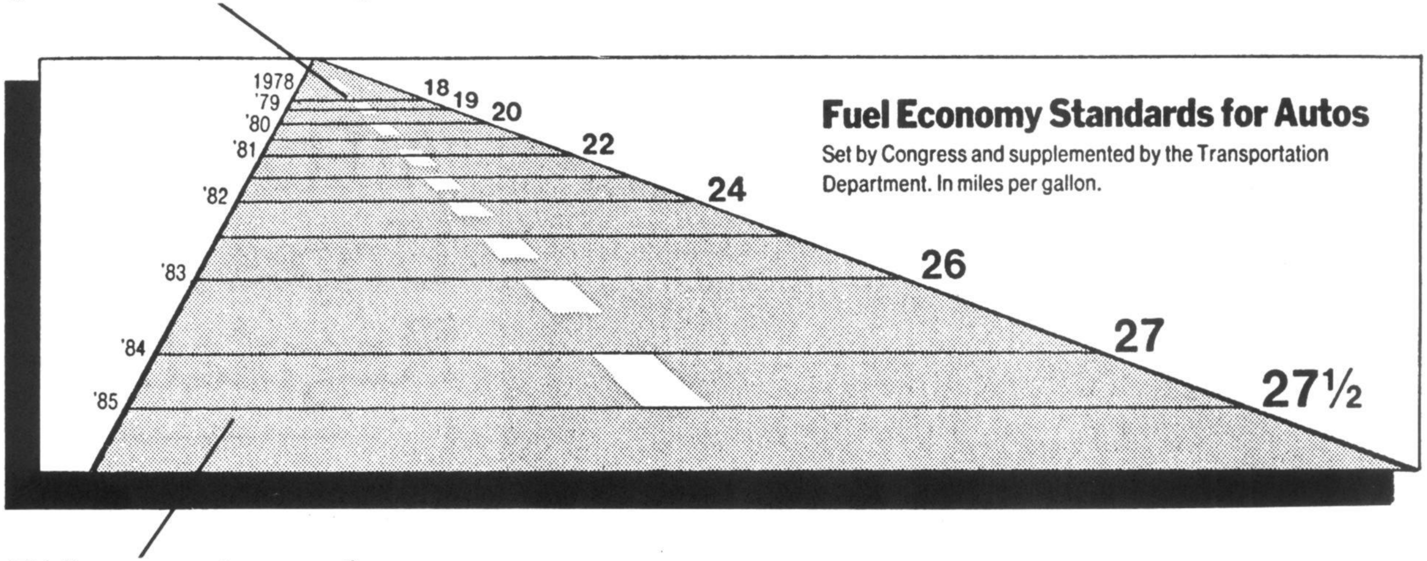
The lie factor

- 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 size of the effect

Measuring distortion through the lie factor

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

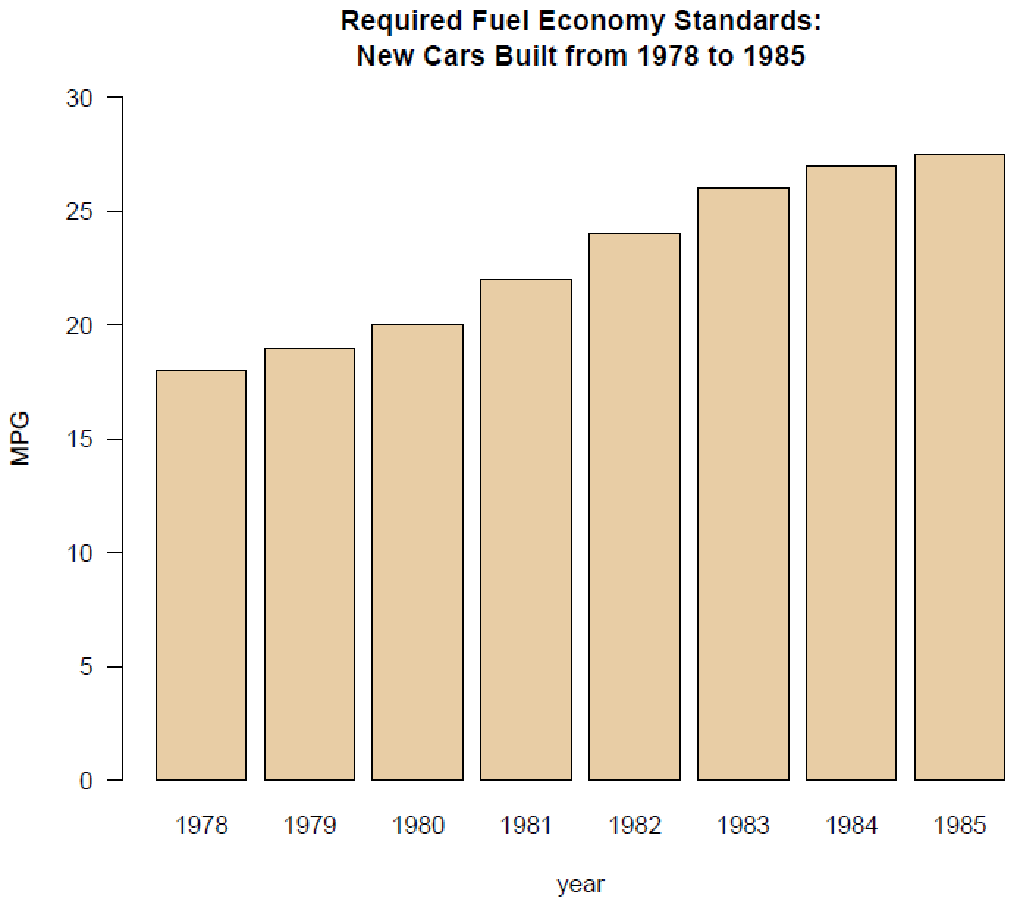


This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

$$\text{Data Effect} = \frac{27.5 - 18}{18} = 0.53, \qquad \text{Graph Effect} = \frac{5.3 - .6}{.6} = 7.83,$$

$$\text{Lie Factor} = 14.8$$

The same data with lie factor=1

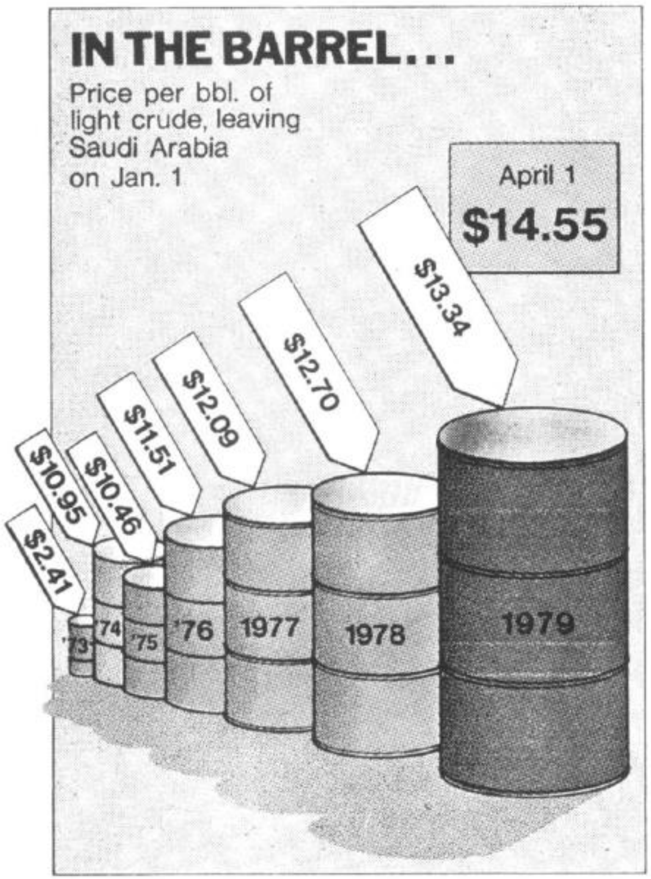


Common Sources of Distortion

- The use of 3 dimensional “effects” 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

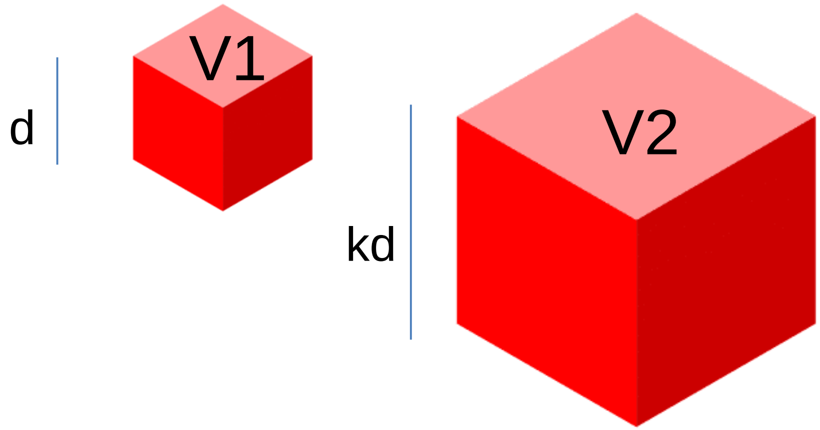


$V1=d^3$

$V2=k^3d^3$

$V2/V1 = k^3$

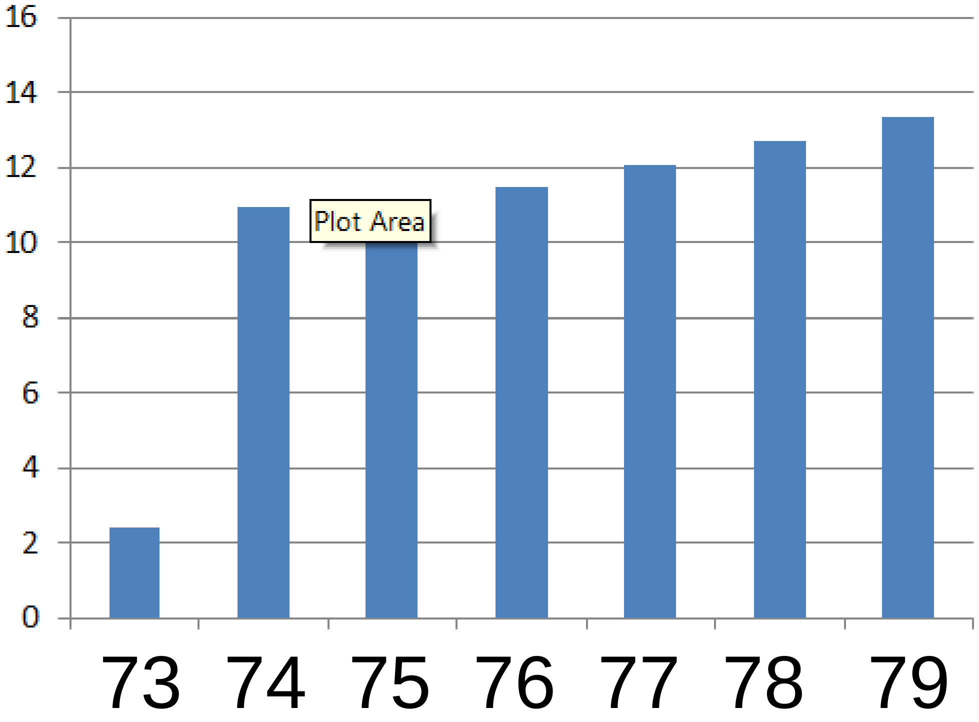
$kd/d \quad = k$

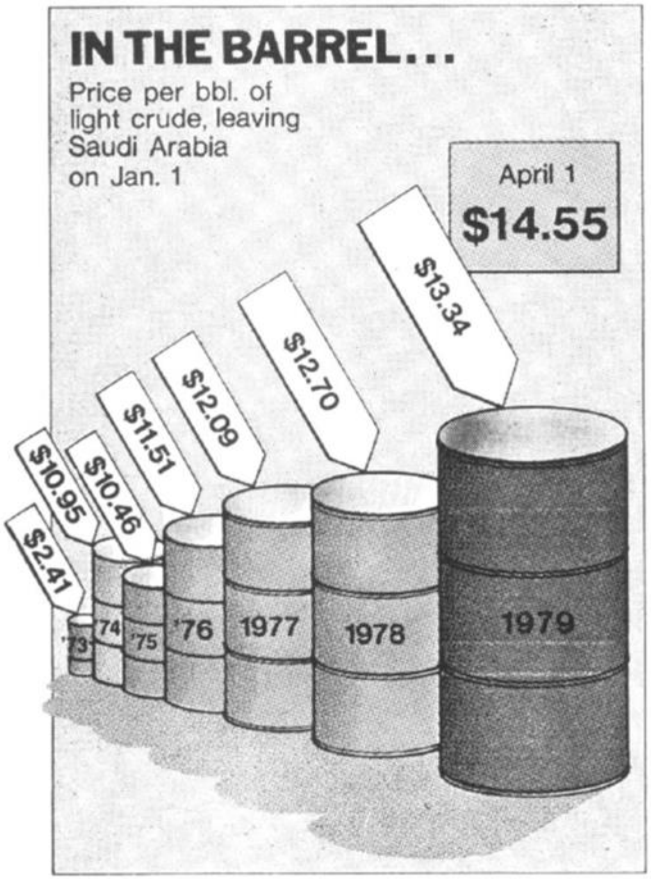


Lie factor ~= $k^3/k = k^2 =$
size_of_effect_in_data^2

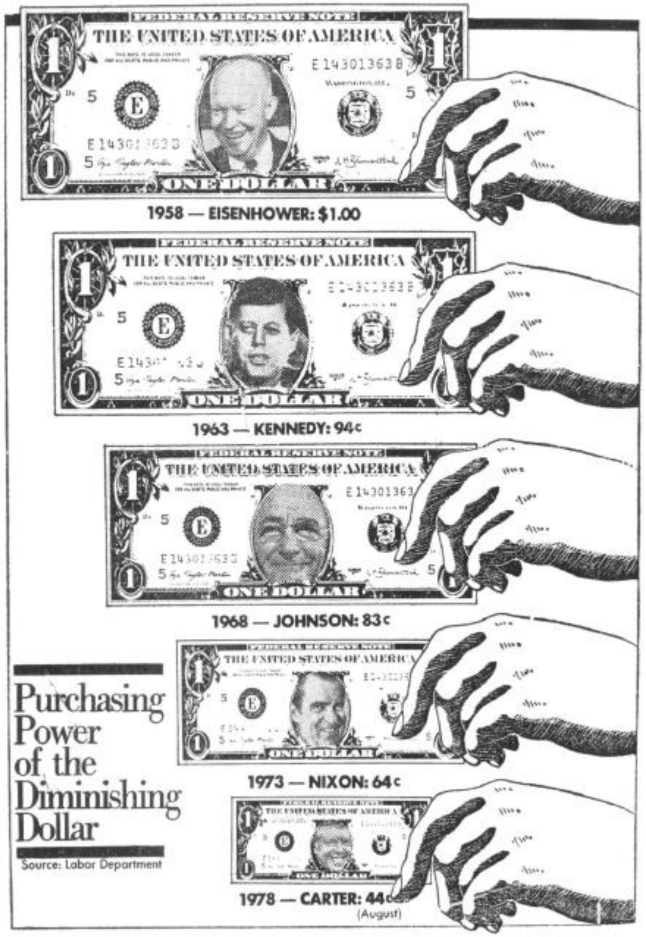
Lie factor= ~9

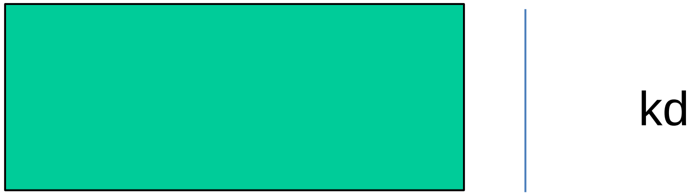
The same data



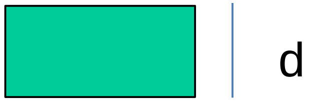


Distortion through areas





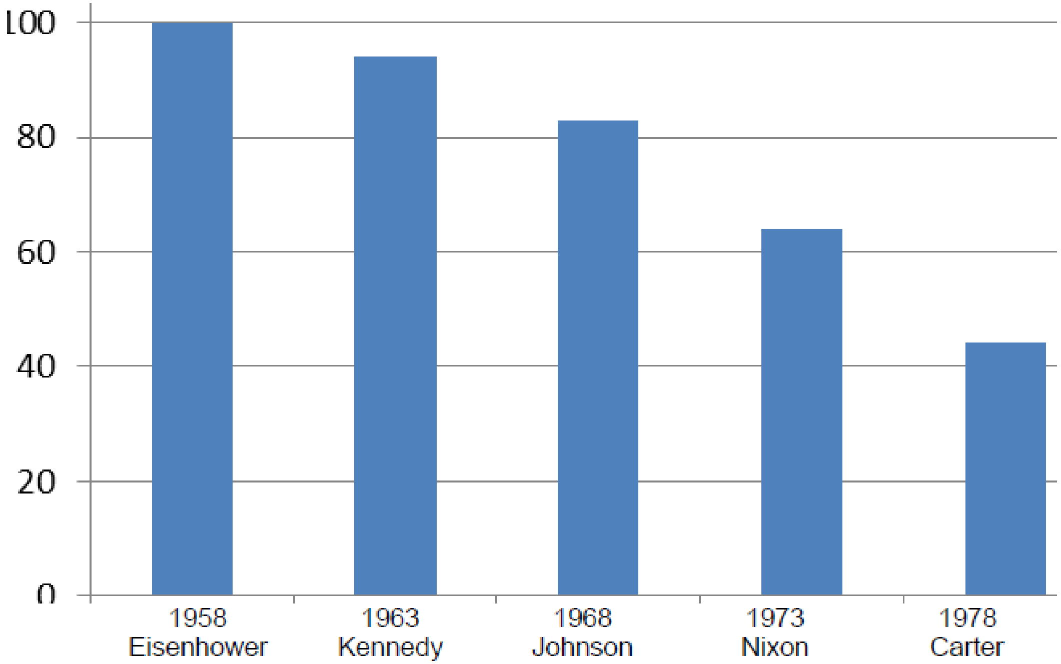
Lie factor ~= k^2/k = k = **size_of_effect_in_data**



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

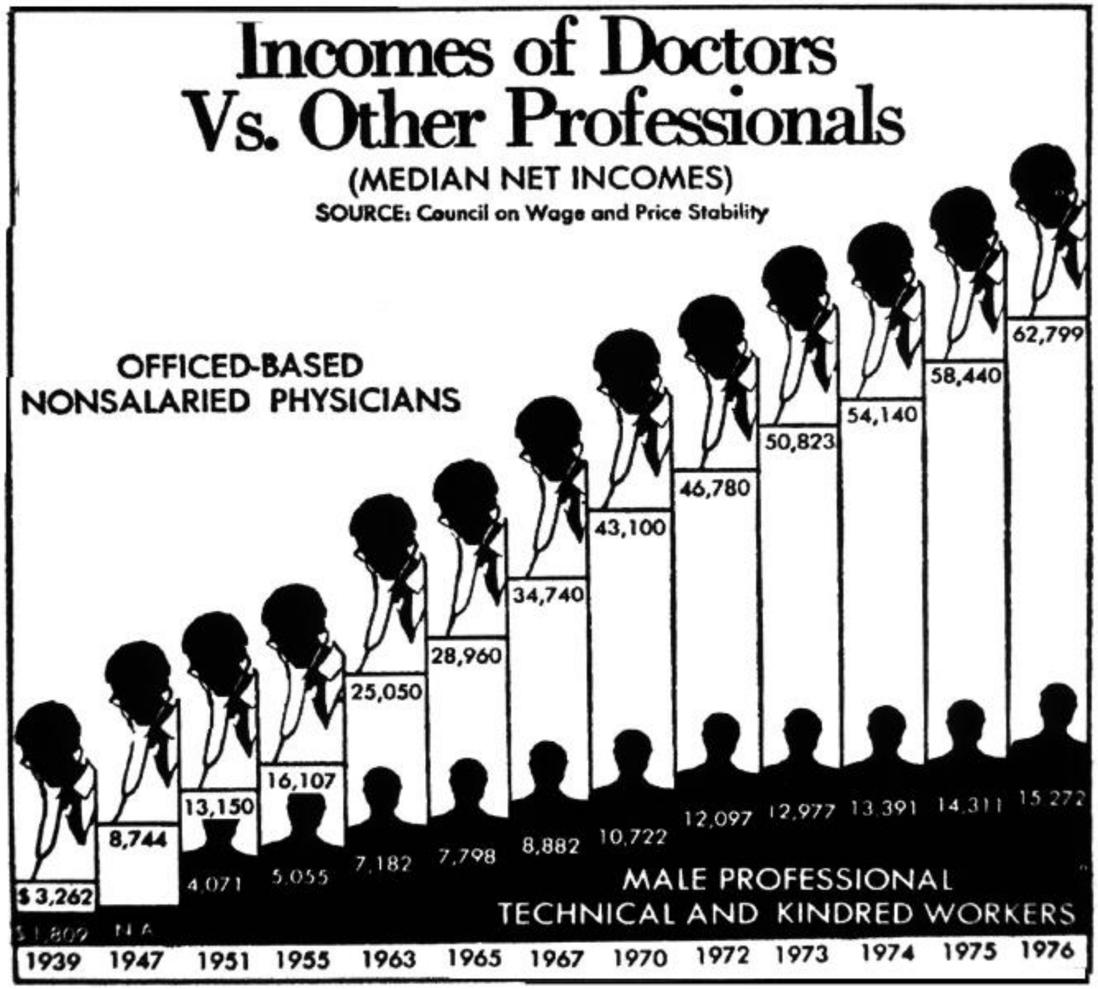
The same data with lie factor=1

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



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

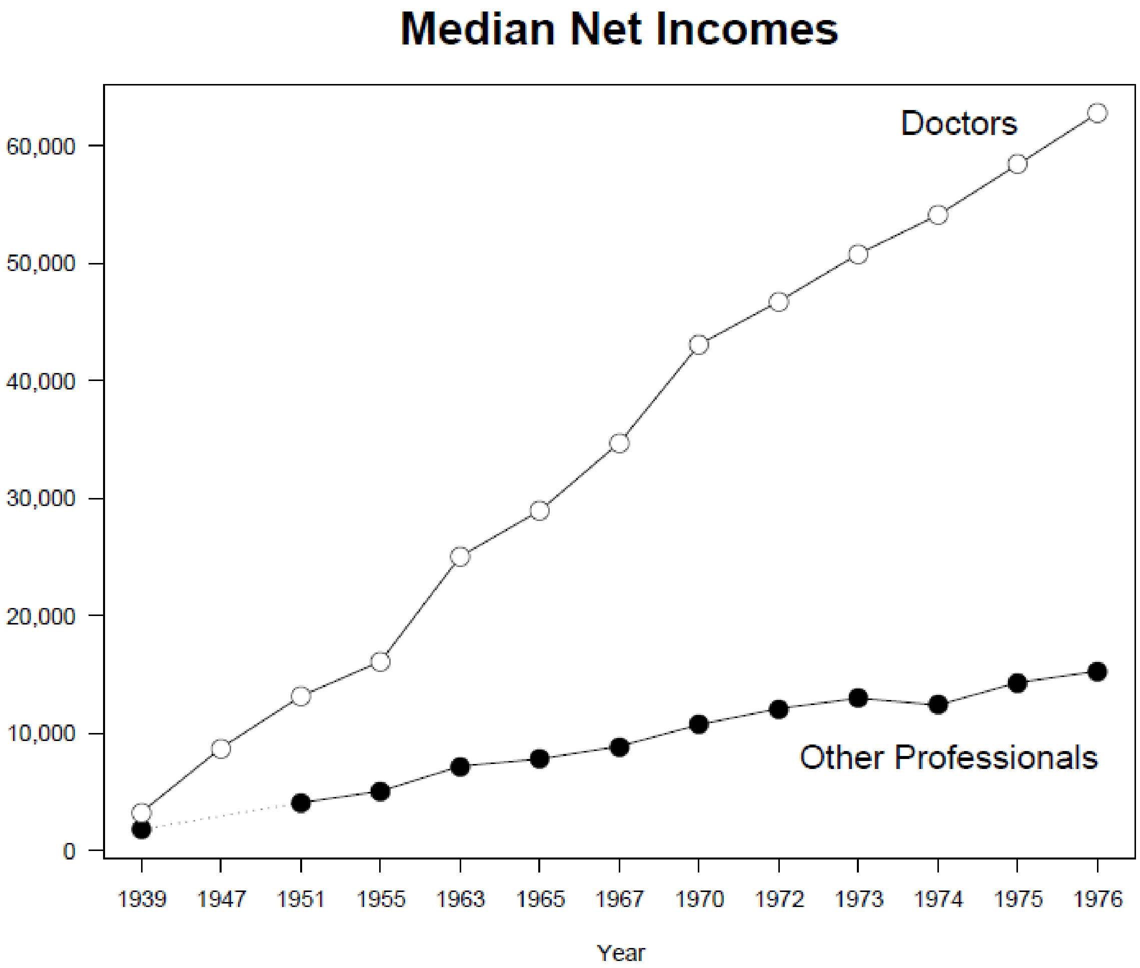
Distortion (deliberate?)



What's wrong with this graph?

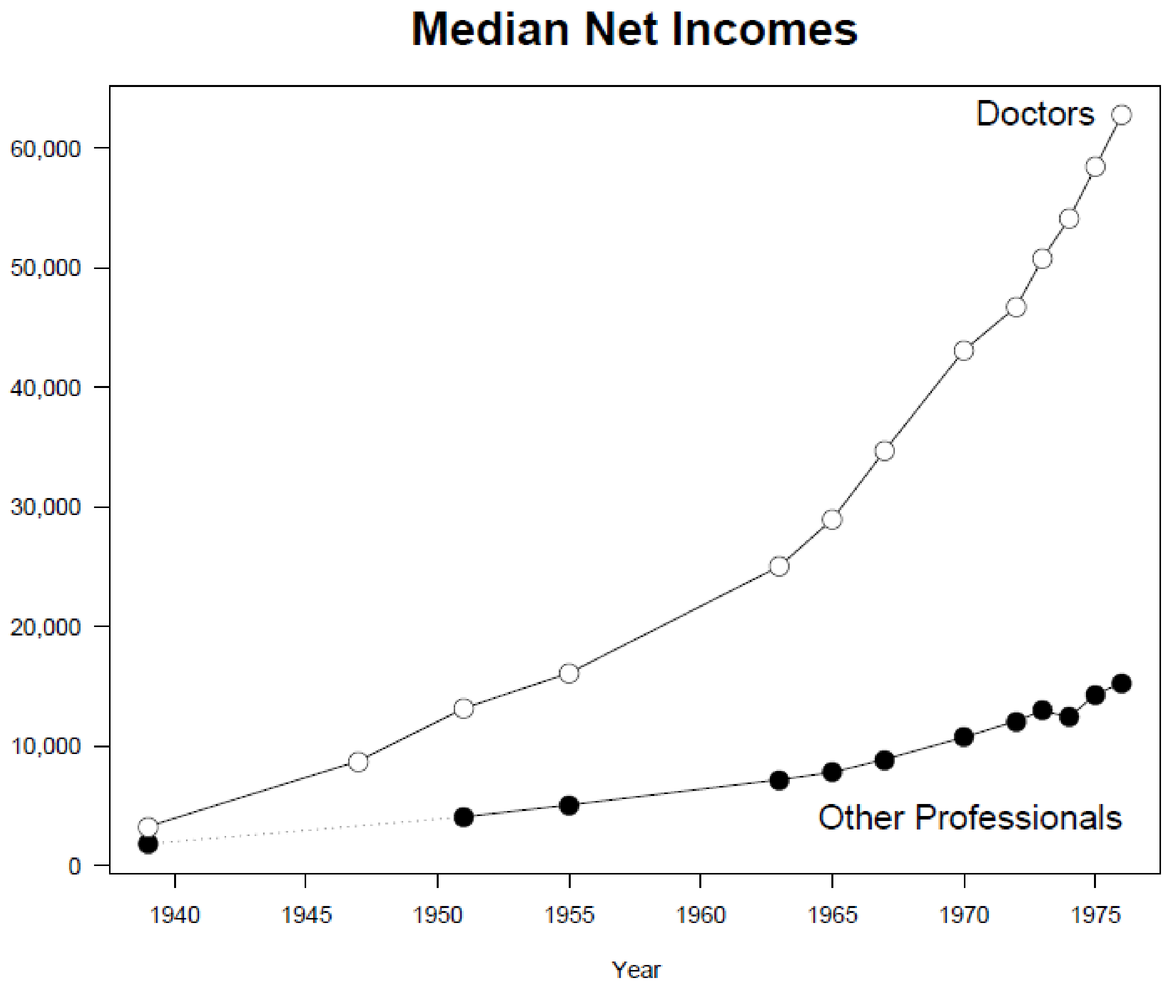
A part of the chartjunk

Presented data



It suggests
a linear trend

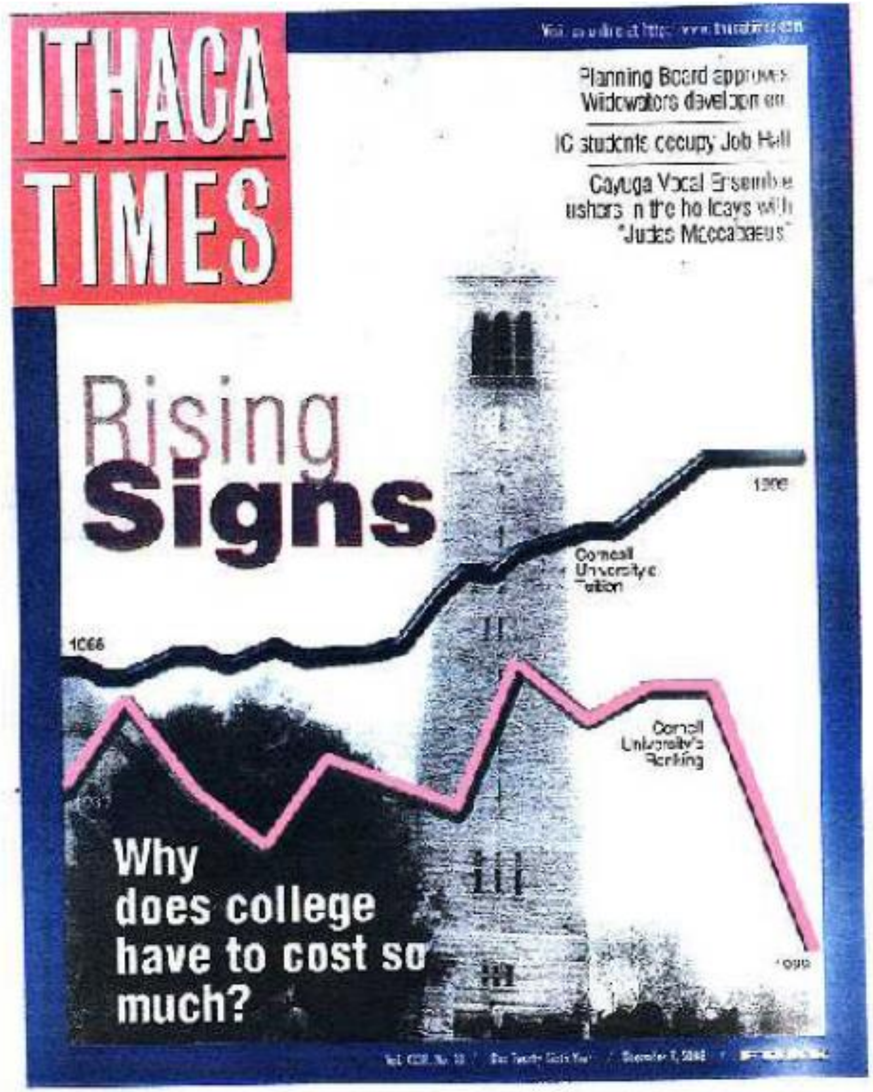
Real data...



The time scales were different!

Now it is clear the exponential trend

One of the best graph lie...



ITHACA TIMES

Visit us online at http://www.ithacatimes.com

Planning Board approves Widewaters development

IC students occupy Job Hall

Cayuga Vocal Ensemble ushers in the holidays with "Judas Maccabaeus"

Rising Signs

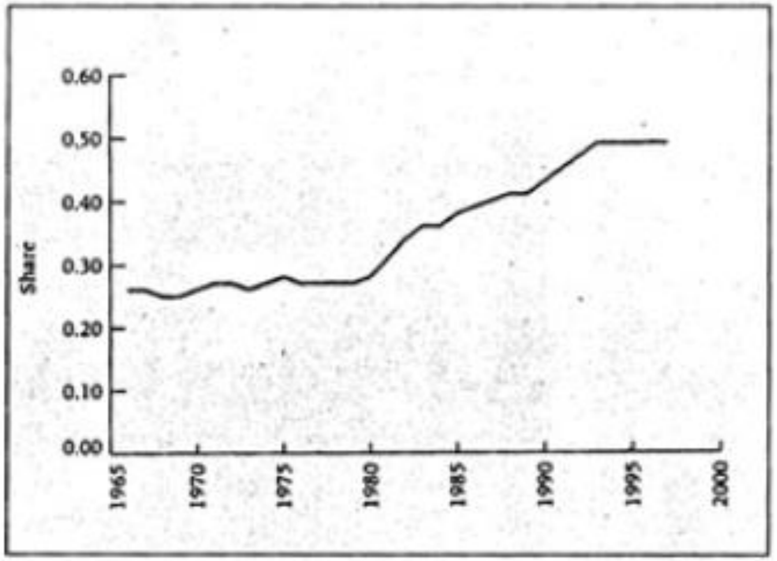
Why does college have to cost so much?

- 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

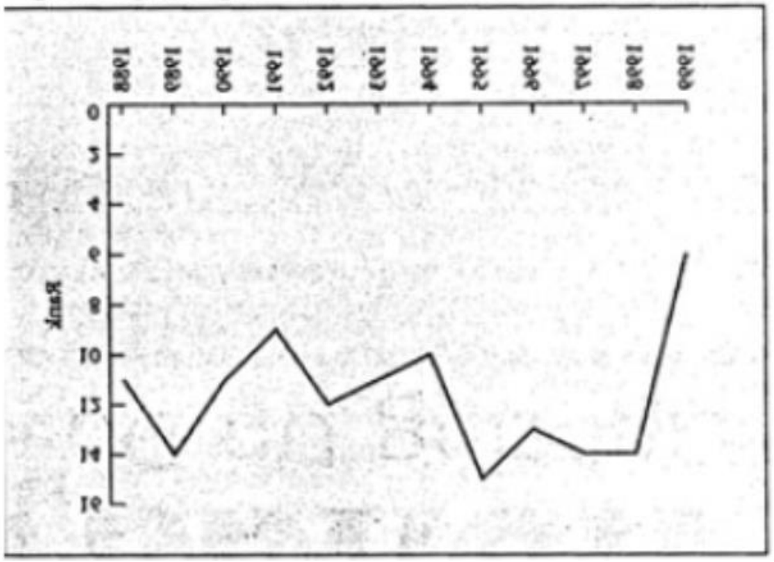
- More careful reading of the whole article (buried several pages into the paper) reveals a different story:
  - (1) The ranking graph covers an 11 year period, the tuition graph 35 years, yet they are shown simultaneously (the same apparent width) on the same horizontal "scale".
  - (2) The vertical 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) The differing time units are cleverly disguised by printing them rotated 90°.
  - (4) 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



By the Numbers: Over 35 years, Cornell's tuition has taken an increasingly larger share of its median student family income.

Pecking Order: Over 12 years, Cornell's ranking in US News & World Report has risen and fallen erratically.



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)