Capítulo 2

Data Visualization

AMG, JFO (v8 – 2017)
adaptado de: *Estatística*,
Rui Campos Guimarães,
José A. Sarsfield Cabral

Slide 2a.-1

Conteúdo

Slide 2a.0

2.1 Information Visualisation

Slides based in: "Lecture 1", Statistics 120 – Information Visualisation, by Ross Ithaka

Dictionary Definition

- The action or fact of visualising; the power or process of forming a mental picture or vision of something not actually present to the sight; a picture thus formed
- The action or process of rendering visible

Operational Definition

- The construction of images which represent important aspects of some situation or process
- Synonyms for image are: plot, graph, diagram, picture

Slide 2a.1

Why Visualise?

- The human visual cortex is arguably the most powerful computing system we have access to
- Visualisation allows us to put information into a form which allows us to use the power of this computing system
- By harnessing some of the capabilities of our visual system we can free other parts of our brains to work on problems

How Visualisation Can Be Useful

Communication – visualisation provides a quick way to communicate a very rich message

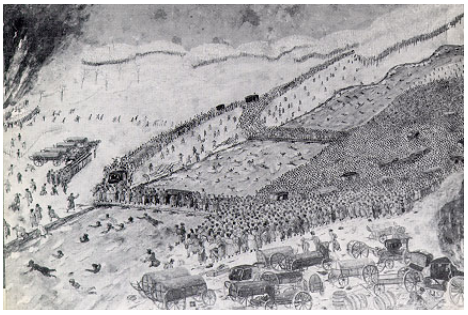
Discovery – visualisation provides a way of displaying a large amount of information so we can uncover new facts and relationships

Insight – visualisation provides a way to obtain better insight into things we already know

Slide 2a.2

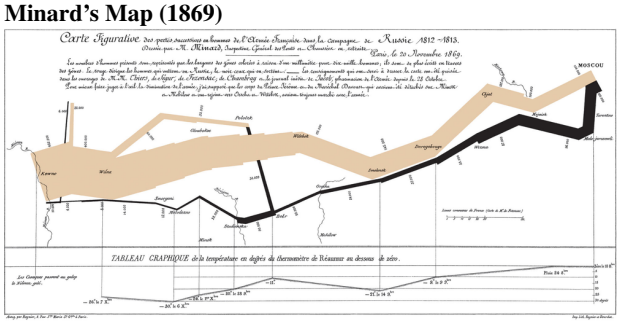
2.1.1 Napoleon's 1812 Campaign

- French Emperor Napoleon invaded Russia with an army of 500,000 men
- The French occupied Moscow, but had to retreat due to the Russian winter, the lack of supplies and a scorched earth policy by the Russians
- During the retreat, temperatures fell as low as -30^{o}C. The bitter weather, together with guerrilla attacks by Russian forces, decimated the French
- The crossing of the Bérézina river posed the last challenge, due to a frozen river thawing. Of 40,000 men, 25,000 were lost
- Just 10,000 men returned from Russia



Slide 2a.3

Minard's Map (1869)



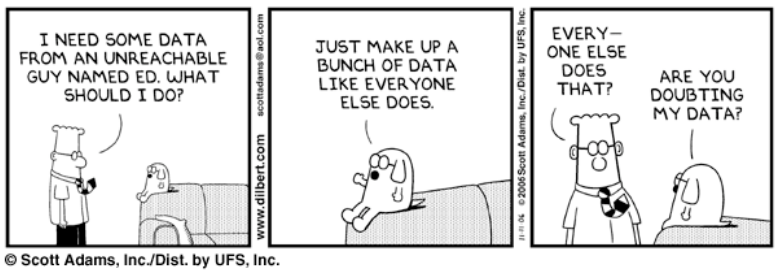
- This map has been termed "the best graph ever produced"
- It shows the spatial location, marching direction and size of the invading army as well as the temperature during the retreat
- The complexity is handled by abstracting out the most important data features and presenting them in a simple graphical form

Slide 2a.4

2.2 Statistical Graphs

Slides based in:

- "Lecture 3", Statistics 120 – Information Visualisation, by Ross Ithaka
- "A Bad Graph Hall of Shame", Course Notes for Beginning and Intermediate Statistics, by Carl Schwarz
- "A discussion of Misleading Graphs", by Mark Harbison



© Scott Adams, Inc./Dist. by UFS, Inc.

Slide 2a.5

- There are three kinds of lies, *"lies, damned lies, and statistics"*. This is never more true than when poor statistical graphs are drawn
- A picture is worth a thousand words, and a graph worth a thousand numbers only if it is *clear, concise and correct*
- As a general rule graph your data – often a properly chosen graph will obviate the need for any further analysis. Don't lose sight of the purpose of the graph when you are drawing it
- Small amounts of data don't need to be graphed
- Besides presenting the data graphically, raw data should also be provided (as a table, in appendices, online, . . .)

Slide 2a.6

2.2.1 Principles of Good Graphical Design

- *There should be a high data to chart ratio*
  Data points should be clearly visible and should not be overwhelmed by axes, titles, reference lines, and chart junk
- *Use the appropriate graph for the appropriate purpose*
  There are only a few basic types of graphs:

  Trend graphs – to emphasize the trend in a time series, a line chart is better than a series of side-by-side bars

  Relative size graphs – side-by-side bar graphs are best; all bars must be anchored at zero; all bars should be equal width

  Composition graphs – stacked bar-charts with the most important segments at the top or the bottom of the bar
- *Make sure that the graph is complete*
  All axes must be labelled; there should be a title on the graph

Slide 2a.7

Think about the overall presentation of the graph:

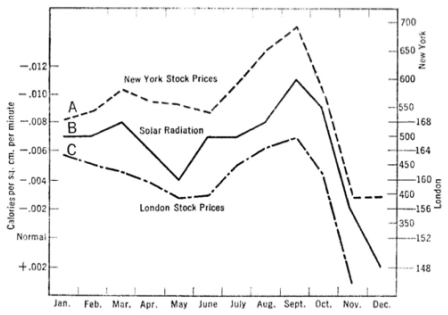
- The points on a plot should be spread over the area of the graph without being shoved into one corner
- The axes scales should be appropriate. In some cases, a log-scale is a better representation of data that spans several orders of magnitude
- Where is the 0 point on a graph. In particular, bar charts should always be anchored at zero
- Use different plotting symbols or line-types to differentiate among groups on the graph
- The independent variables is usually plotted on the X-axis; the dependent variable usually on the Y-axis.

⇒ *The best graph is one that is self-explanatory*

Slide 2a.8

2.2.2 Examples of Bad Graphs

Data Relevance

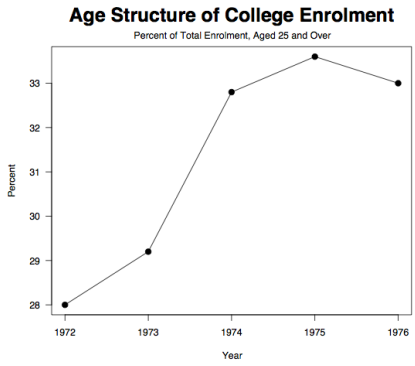


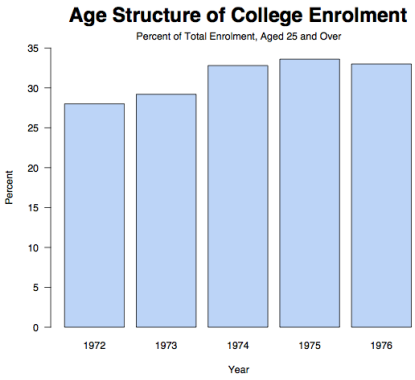
- Graphs are only as good as the data they display
- No amount of creativity can produce a good graph from dubious data

Slide 2a.9

Complexity

Age Structure of College Enrolment (1972-1976), American Education Magazine

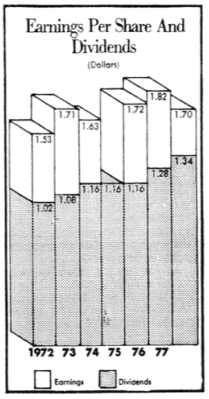




Slide 2a.10

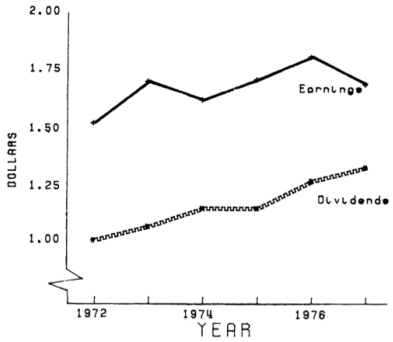
Share Results

The Washington Post, 1979



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

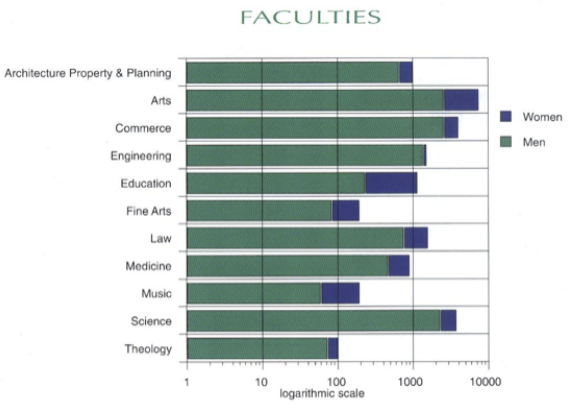
Slide 2a.11



- a clear version of share results graph
- y-axis break correctly shown

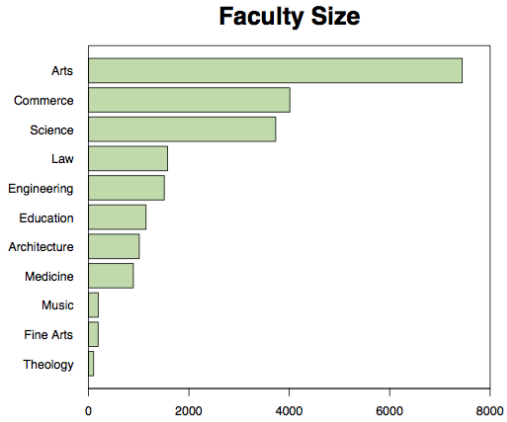
Slide 2a.12

Distortion

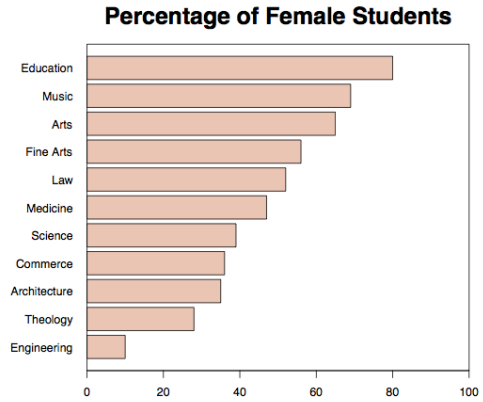


- Graphs should not provide a distorted picture of the values they portray
- Distortion can be either deliberate or accidental

Slide 2a.13



Slide 2a.14



Slide 2a.15

2.2.3 The "Lie Factor"

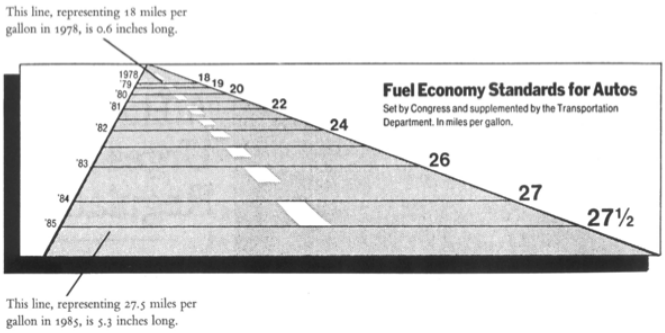
The "Lie Factor"

- Eduard Tufte of Yale University has defined a "lie factor" as a measure of the amount of distortion in a graph. (Don't take this too seriously – i.e don't learn it for the exam).
- The lie factor is defined to be:

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect shown in data}}$$

- If the lie factor of a graph is greater than 1, the graph is exaggerating the size of the effect
- If the lie factor of a graph is smaller than 1, the graph is diminishing the size of the effect
- "Lie factors" smaller than 0.95 or greater than 1.05 $\Longrightarrow$ Distortion
- "Lie factors" between 2 and 5 aren't rare

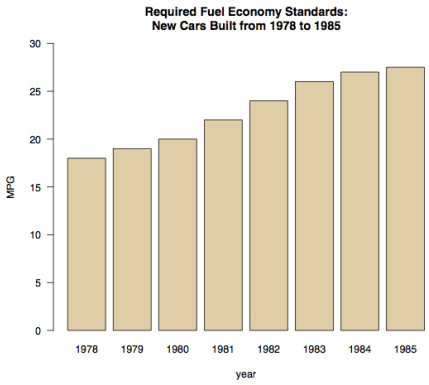
Slide 2a.16



$$\text{Data Effect} = \frac{27.5 - 18}{18} = 0.53, \qquad \text{Graph Effect} = \frac{5.3 - 0.6}{0.6} = 7.83$$

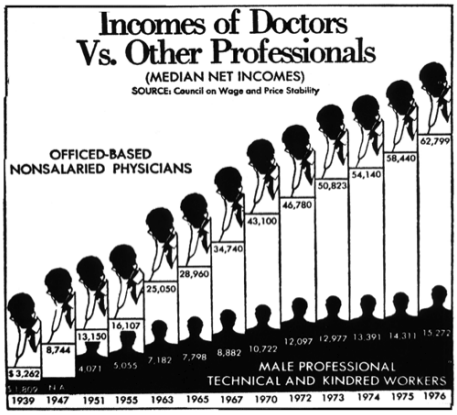
$$\text{Lie Factor} = \frac{7.83}{0.53} = 14.8$$

Slide 2a.17



Slide 2a.18

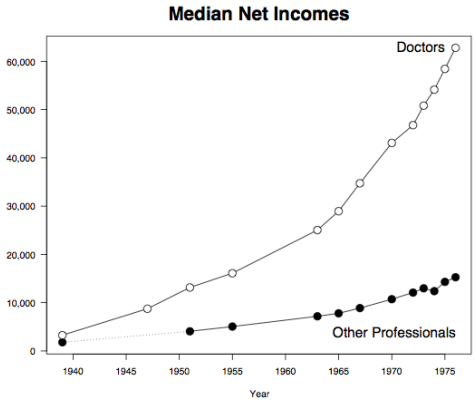
Deliberate Distortion



(Washington Post)

- Changing scale in mid-axis to make exponential growth linear
- Usually these are an attempt to hide some feature of the data

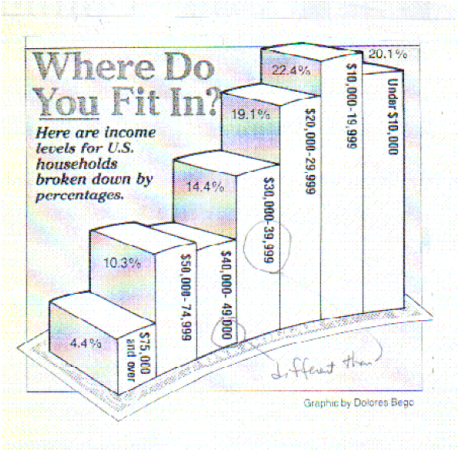
Slide 2a.19



Slide 2a.20

2.2.4 Additional Examples of Bad Graphs

Income Levels



(*Schwarz, 2006*)

- The 3-D effects make it difficult to read the bars (do you look at the front, the side, or the back of each bar?)
- The on-horizontal scale artificially increases the lower-income bars compared to the upper-income bars
- Some of the bars are missing the percentage figure
- The interval sizes change (it goes by by $10,000 than by $25,000 which artificially increases the 50-75,000 bar)
- Why use 29,999 rather than 30,000?

Slide 2a.21

Post vs. News

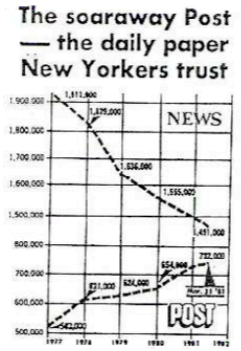


FIGURE 2-9
The *Post's* spectacular climb from obscurity. (*New York Post*, 1981, via Wainer, 1984.)

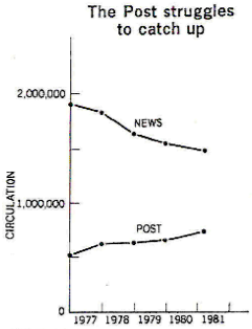
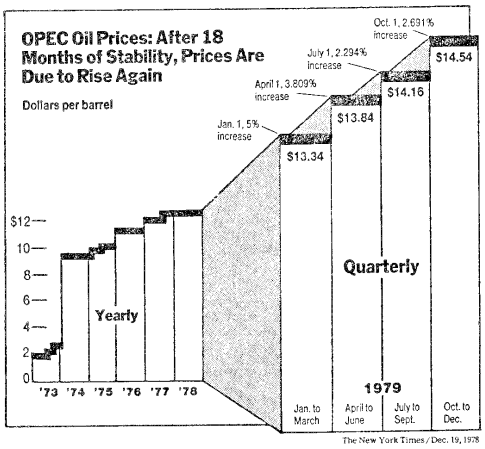


FIGURE 2-10
Restoring the complete *y*-axis in Figure 2-9 shows a much more modest performance for the *Post*, with the *News* still well in the lead.

Slide 2a.22

(*Nóvoa, 2008*)

Oil Prices

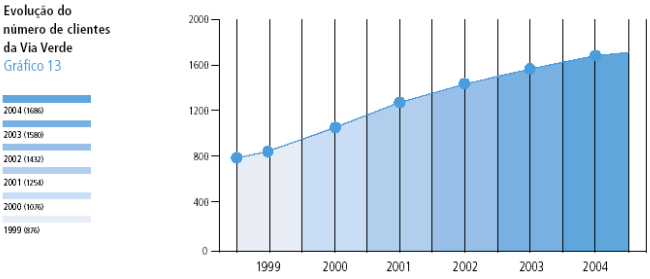


(*Tufte, 1983*)

- 1979 prices in a different plane suggests a higher increase
- unnecessary and misleading *design* variation

Slide 2a.23

Via Verde

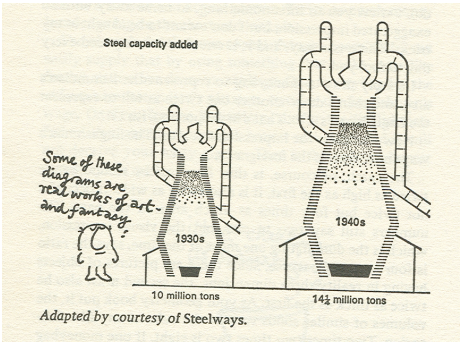


(*Oliveira, 2007*)

- design variation (*shading*) artificially suggests a bigger increase

Slide 2a.24

Steel capacity

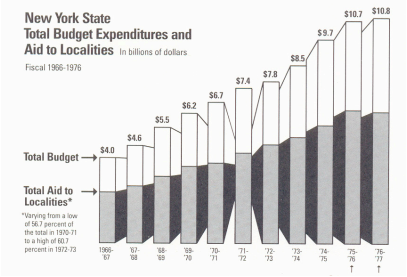


(*Huff, 1973*)

- increasing the image (width and height increasing) $\longrightarrow$ suggest bigger variation than the real one (difference in heights)

Slide 2a.25

NY State Total Budget



Suggestion of a big increase in the last 3 years:

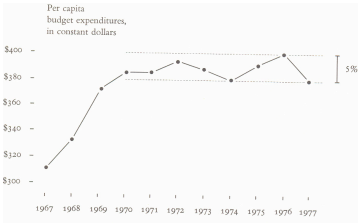
- 3D effect
- 3 last columns in a different plane

Slide 2a.26

Need to adjust values for:

- population increase
- inflation

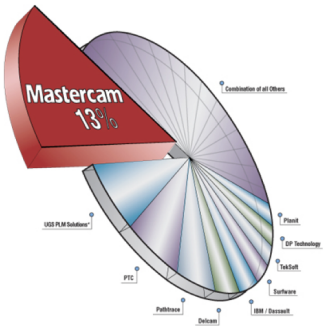
(*Tufte, 1983*)

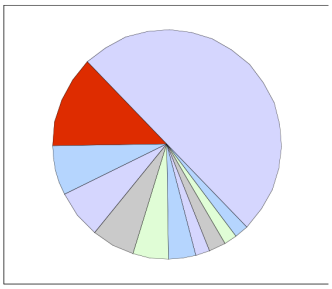


Pie Chart

(speedbump.com, 2007-08-24. This image is copyright protected.)

Slide 2a.27

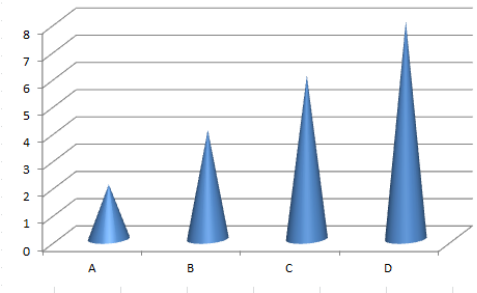




- As general rule avoid pie charts
- If used they should be plain 2-D to preserve angle measures
- Do not use 3-D or "exploding" effects
- Pie-charts are acceptable to represent categoric data

Slide 2a.28

Conic Excel graph



- Bad Graph Contest: What Software Produces the Worst Graph? (`http://www.forbes.com/sites/naomirobbins/2012/05/16/bad-graph-contest-what-software-produces-the-worst-graph/` `http://www.forbes.com/sites/naomirobbins/2012/05/30/winner-of-the-bad-graph-contest-announced-2/`)
- The height is underestimated when the gap depth is not zero (the numbers plotted are 2, 4, 6, and 8)
- Avoid 3-D and other visual effects, use simple 2-D bar chart

Slide 2a.29

2.2.5 Final Comments and Bibliography

Drawing Good Graphs

- 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.

Creating statistical graphs with Excel requires special care:

- horizontal axis in line and bar graphs don't have a proper numerical scale

- default format options for most of the graphs are *bad choices*

Slide 2a.30

Bibliography

- Eduard Tufte (1983), "The Visual Display of Quantitative Information", Graphics Press (http://www.edwardtufte.com/tufte/books_vdqi)
- Eduard Tufte (1990), "Envisioning Information", Graphics Press (http://www.edwardtufte.com/tufte/books_ei)
- Eduard Tufte (1997), "Visual Explanations", Graphics Press (http://www.edwardtufte.com/tufte/books_visex)
- John Tukey (1972), "Some Graphic and Semigraphic Displays", T.A. Bancroft, ed., Statistical Papers in Honor of George W. Snedecor (http://www.edwardtufte.com/tufte/tukey)
- Darrell Huff (1973), "How to Lie With Statistics", Penguin Books (https://archive.org/details/HowToLieWithStatistics)

Slide 2a.31

- C. Kelleher, T. Wagener (2011), "Ten guidelines for effective data visualization in scientific publications", Environmental Modelling & Software, Vol. 26, pp 822–827 (https://doi.org/10.1016/j.envsoft.2010.12.006)
- NP Rougier, M. Droettboom, PE Bourne PE (2014), "Ten Simple Rules for Better Figures", PLOS Computational Biology, Vol. 10(9): e1003833 (https://doi.org/10.1371/journal.pcbi.1003833)
- W. S. Cleveland; R. McGill (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods", Journal of the American Statistical Association, Vol. 79, pp. 531–554 (https://doi.org/10.1080/01621459.1984.10478080)
- W. S. Cleveland; R. McGill (1987), "Graphical Perception:The Visual Decoding of Quantitative Information on Graphical Displays of Data", Journal of the Royal Statistical Society. Series A (General), Vol. 150, pp. 192–229 (http://doi.org/10.2307/2981473)

Slide 2a.32

2.3 Caso de Estudo: "Analyzing Tipping Behavior"

Adapted from: "Interactive and Dynamic Graphics for Data Analysis — With R and GGobi", Dianne Cook and Deborah F. Swayne © 2007 Springer Science+Business Media, LLC

Slide 2a.33

Description

One waiter recorded information about each tip he received over a period of a few months working in one restaurant. He collected several variables:

- tip (i.e., gratuity) in US dollars
- bill (the cost of the meal) in US dollars
- sex of the bill payer
- whether the party included smokers
- day of the week
- time of day
- size of the party

In all he recorded 244 tips. The data was reported in a collection of case studies for business statistics:

> Bryant, P. G. and Smith, M. A. (1995), "Practical Data Analysis: Case Studies in Business Statistics", Richard D. Irwin Publishing, Homewood, IL

The primary question suggested by the data is this: *What are the factors that affect tipping behavior?*

Slide 2a.34

Id	*Total Bill*	*Tip*	*Sex*	*Smoker*	*Day*	*Time*	*size*
1	16.99	1.01	Female	No	Sun	Dinner	2
2	10.34	1.66	Male	No	Sun	Dinner	3
3	21.01	3.5	Male	No	Sun	Dinner	3
4	23.68	3.31	Male	No	Sun	Dinner	2
5	24.59	3.61	Female	No	Sun	Dinner	4
6	25.29	4.71	Male	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
240	29.03	5.92	Male	No	Sat	Dinner	3
241	27.18	2	Female	Yes	Sat	Dinner	2
242	22.67	2	Male	Yes	Sat	Dinner	2
243	17.82	1.75	Male	No	Sat	Dinner	2
244	18.78	3	Female	No	Thur	Dinner	2

Slide 2a.35

Typical approach:

- Start by gathering knowledge about tipping behavior
  - Cultural issue (which is the usual tipping behavior in US?)
  - Service quality
  - Percentage of total bill (new variable: percentage = tip/bill)
  - Party size (service charge)
  - ...
- Analyze (main) variables individually (univariate analysis)
  - Quantitative and Qualitative variables
  - Summarize variables
    - quantitative: compute statistics and create statistical graphs
    - qualitative: count occurrences and create tables and statistical graphs
  - Statistics (average, median, quartiles, max/min, standard deviation, skewness, kurtosis. ...)
  - Statistical graphs (histograms, bar and line charts, boxplots, ...)
  - Look for patterns, (a)symmetries, extreme values (outliers?)

Slide 2a.36

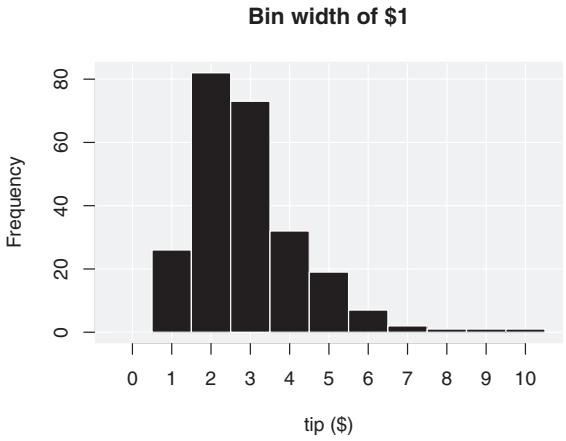
- Analyze quantitative variables by groups (qualitative variables)
  - Use qualitative variables to create groups (Male/Female, ...)
  - Repeat univariate analysis by groups
- Relationships between pairs of quantitative variables (bivariate analysis)
  - Linear relationship (other types of relationships)
  - Covariance and correlation
  - Scatter plots (XY graphs)
- Relationships by groups
  - ...
- Report / Poster / ...
  - Intro / Methodology / Data / Results / Discussion / Conclusions

Statistics

Count	244
Mean	2.998279
Median	2.9
Mode	2
Geometric Mean	2.725189
Harmonic Mean	2.482253
Q1	2
Q3	3.5625
Minimum	1
Maximum	10
Range	9
IQR	1.5625
Variance	1.914455
Standard Deviation	1.383639
Skewness	1.465451
Kurtosis	3.648376

Count	244
Mean	3.00
Median	2.9
Mode	2
Geometric Mean	2.73
Harmonic Mean	2.48
Q1	2
Q3	3.56
Minimum	1
Maximum	10
Range	9
IQR	1.56
Variance	1.91
Standard Deviation	1.38
Skewness	1.47
Kurtosis	3.65

- We look first at the variable of greatest interest to the waiter: *tip*.
- A common graph for looking at a single variable is the histogram (data values are binned and the count is represented by a rectangular bar)

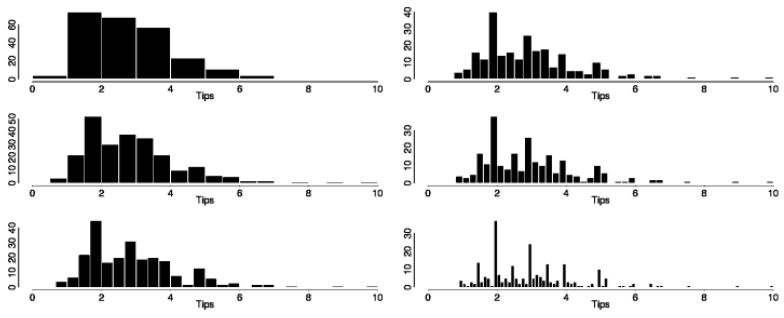


Slide 2a.39

- We choose an initial bin width of $1 and produce the above histogram
- The distribution appears to be unimodal; that is, it has one peak, the bar representing the tips greater than $1.50 and less than or equal to $2.50.
- There are very few tips of $1.50 or less.
- The number of larger tips trails off rapidly, which suggests that this is not a very expensive restaurant.

Slide 2a.40

Changing bin width

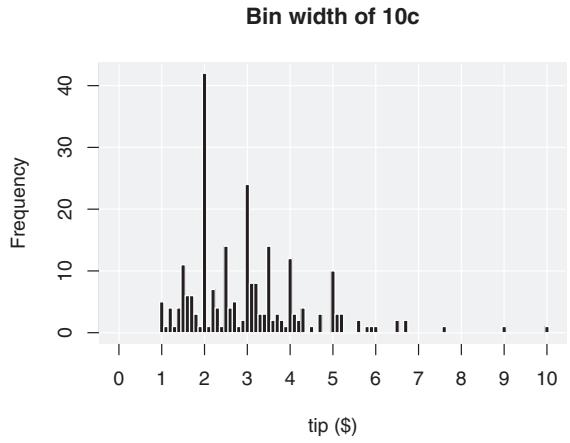


The small multiples are generated from the data by using several bin widths, ranging from $1 at top to 10c at bottom

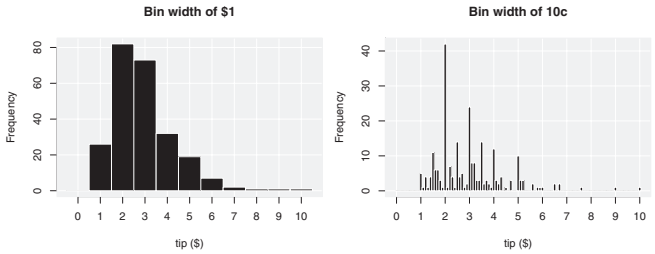
Slide 2a.41

- The multiple layout illustrates how different information that is gained from examining the data at different resolutions like this
- The largest bin width $1 shows tip to have a unimodal and skewed distribution, which for the data means that most tips are of the smaller amounts with less and less larger tips
- As the bin width is reduced the shape of the distribution becomes multimodal
- At the smallest bin width 10c it is clear that there are modes at the full dollar and fifty cent amounts
- This means that the customers tend to round the tip amount to the nearest fifty cents or dollar
- It is important to emphasize that the salient features are not found with one ideal bin width but one must use multiple bin widths to extract different features of data

Slide 2a.42



Slide 2a.43



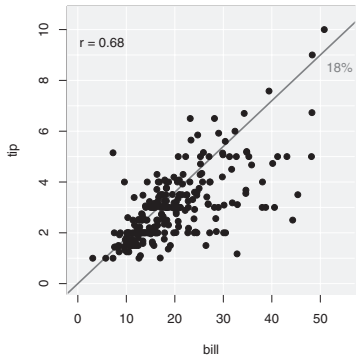
- This type of observation occurs frequently when studying histograms:
  - A large bin width smooths out the graph and shows rough or global trends,
  - Whereas a smaller bin width highlights more local features.
- Since the bin width is an example of a graph parameter, experimenting with bin width is an example of exploring a set of related graphs.
- Exploring multiple related graphs can lead to insights that would not be apparent in any single graph.

Slide 2a.44

- So far we have not addressed the primary question:

  What relationships exist between tip and the other variables?

  Since the tip is usually calculated based on the bill, it is natural to look first at a graph of tip and bill. A common graph for looking at a pair of continuous variables is the *scatterplot*

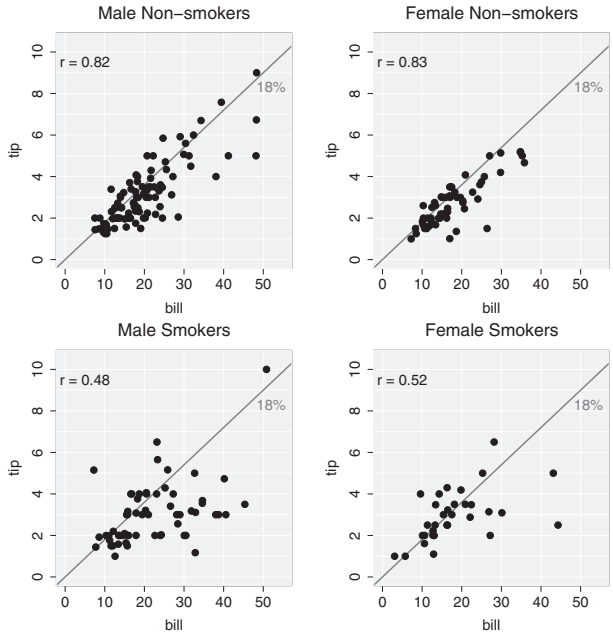


Slide 2a.45

- Highly correlated variables (r = 0.68), confirming that tip is calculated from the bill (the line represents a tip of 18% – US typical tip rate).
- The greater number of points far below the line indicates that there are more “cheap tippers” than generous tippers.
- There are a couple of notable exceptions, especially one party who gave a $5.15 tip for a $7.25 bill (a tip rate of about 70%).

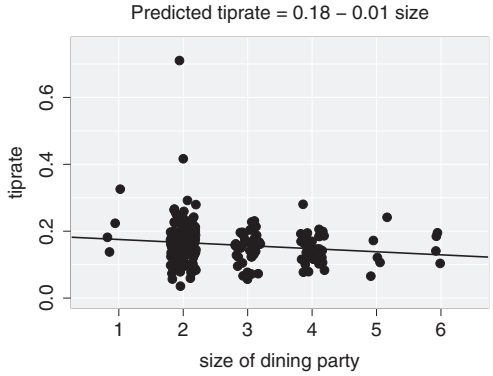
- We said earlier that an essential aspect of data visualization is capturing relationships among many variables: three, four, or even more. This dataset, simple as it is, illustrates the point.
- Let us ask, for example, how a third variable such as *sex* affects the relationship between tip and bill.
- As sex is categorical with two levels (i.e., binary), it is natural to divide the data into female and male payers and to generate two scatterplots of tip vs. bill.
- Let us go even further by including a fourth variable, *smoking*, which is also binary.
- We now divide the data into four parts and generate four scatterplots

  (The 18% tip guideline is included in each plot, and the correlation between the variables for each subset is in the top left of each plot.)



Scatterplot of tip vs. bill conditioned by sex and smoker

- For smoking parties, there is a lot less association between the size of the tip and the size of the bill;
- When a female non-smoker paid the bill, the tip was a very consistent percentage of the bill, with the exceptions of three dining parties;
- Larger bills were mostly paid by men.



Model fitting

- This scatterplot of tiprate vs. size shows the best model along with the data (jittered horizontally).
- There is a lot of variation around the regression line, showing very little signal relative to noise.
- In addition there are very few data points for parties of 1, 5, or 6 diners, so the model may not be valid at these extremes.

Reporting Conclusions

A report summarizing our findings from the study would include the following items:

- Statement of objective for study
- Description of study design and data collection procedures
- Numerical and graphical summaries of datasets
- Description of all inference methodologies:
  - Hypothesis Tests and Confidence Intervals
  - Verification that all necessary conditions for using inference techniques were satisfied
- Discussion of results and conclusions
- Interpretation of findings relative to previous studies
- Recommendations for future studies
- Listing of dataset + detailed calculations (worksheet)