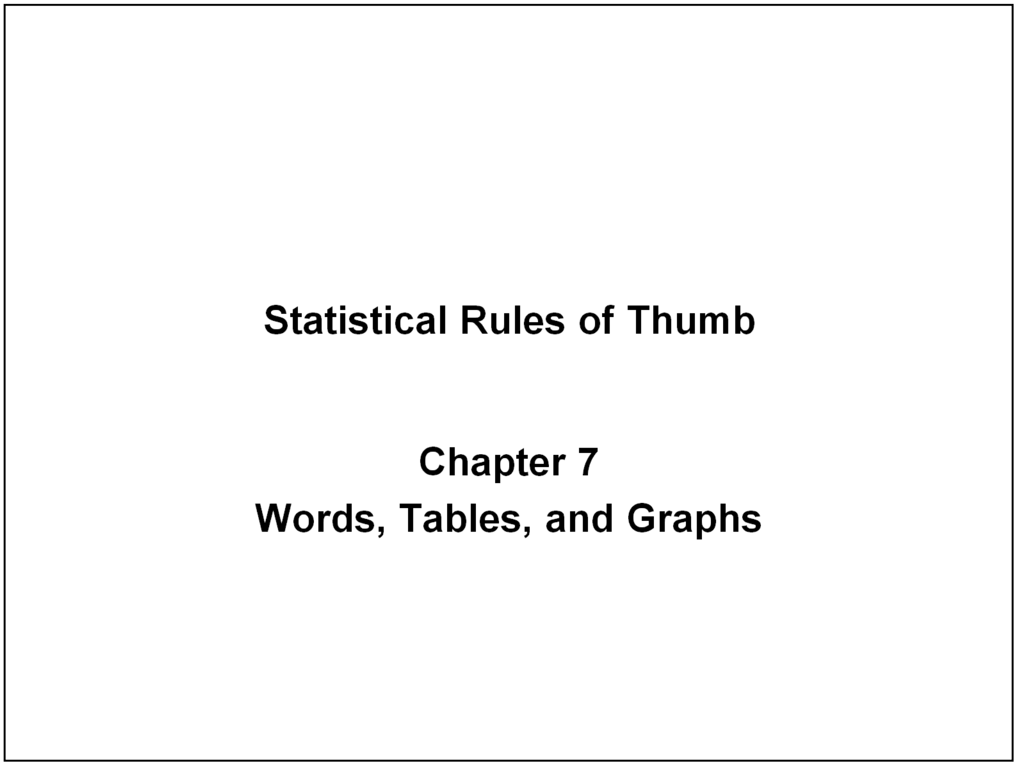
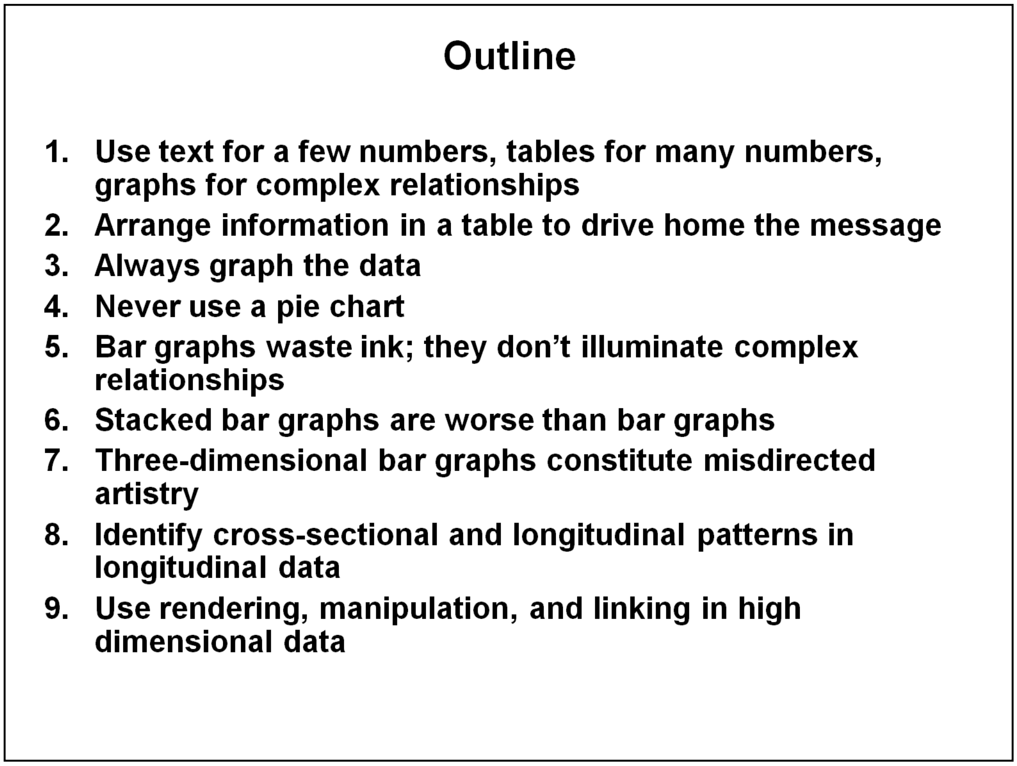


Statistical Rules of Thumb

Chapter 7
Words, Tables, and Graphs

Summary of “Chapter 7: Words, Tables, and Graphs” in *Statistical Rules of Thumb* by Gerald van Belle



Outline

1. Use text for a few numbers, tables for many numbers, graphs for complex relationships
2. Arrange information in a table to drive home the message
3. Always graph the data
4. Never use a pie chart
5. Bar graphs waste ink; they don’t illuminate complex relationships
6. Stacked bar graphs are worse than bar graphs
7. Three-dimensional bar graphs constitute misdirected artistry
8. Identify cross-sectional and longitudinal patterns in longitudinal data
9. Use rendering, manipulation, and linking in high dimensional data

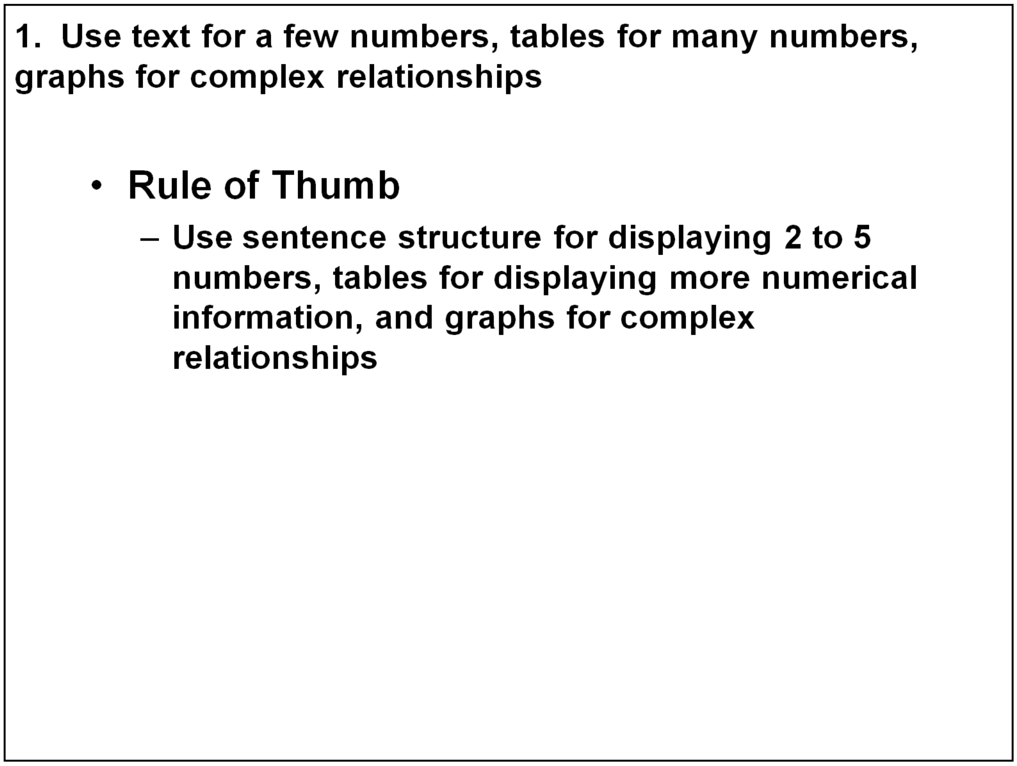
This section provides an organizational structure for summarizing numerical information by means of words, tables and graphs.

The choice between words and tables is primarily a function of the amount of numerical data that has to be described.

The choice between a table and a graph is determined whether a relationship is to be assessed.

This suggests limited or no usefulness for pie charts and bar graphs since they are very poor at displaying relationships.

Additional graphs of relationships can be found in the succeeding sections.



1. Use text for a few numbers, tables for many numbers, graphs for complex relationships

- Rule of Thumb
  - – Use sentence structure for displaying 2 to 5 numbers, tables for displaying more numerical information, and graphs for complex relationships

Introduction

Numerical information is primarily displayed in sentences, tables and graphs (Tufte 1983, page 178).

When should these be used?

Either singly or in combination?

The following sections provide some rules of thumb.

Basis of the Rule

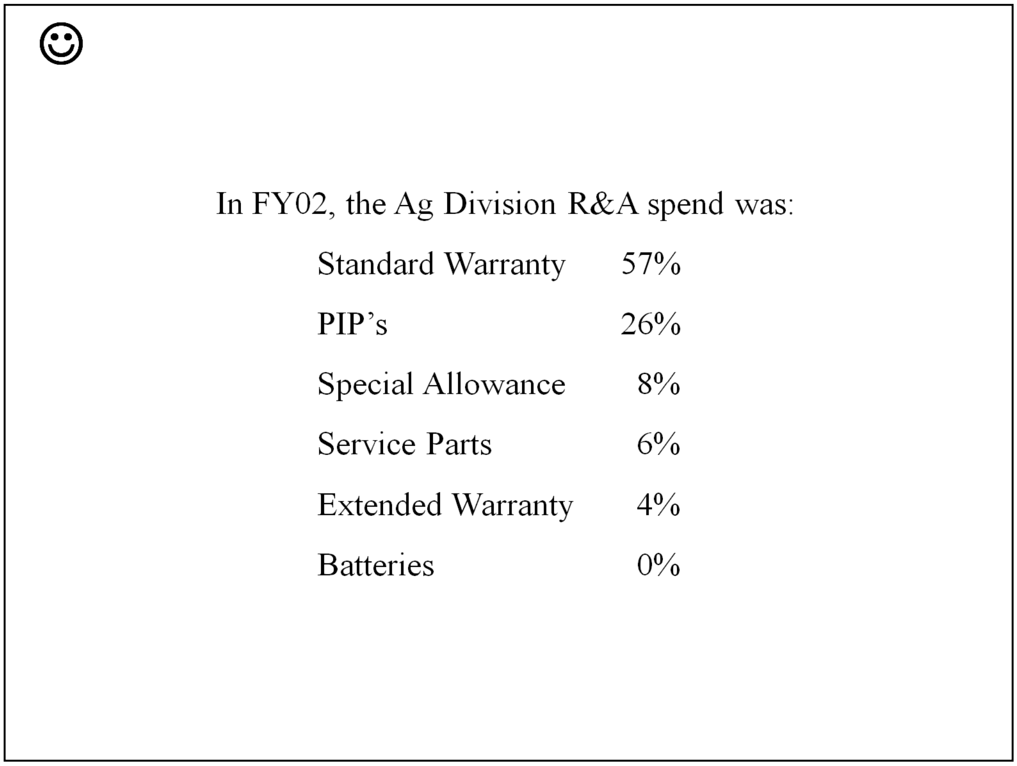
The content and context of the numerical data determines the most appropriate mode of presentation.

A few numbers can be listed, many numbers require a table.

Relationships among numbers can be displaced by statistics.

However, statistics, of necessity, are summary quantities so they cannot fully display the relationships, so a graph can be sued to demonstrate them visually.

The attractiveness of the form of the representation is determined by word layout, data structure, and design.



☺

In FY02, the Ag Division R&A spend was:

Standard Warranty	57%
PIP’s	26%
Special Allowance	8%
Service Parts	6%
Extended Warranty	4%
Batteries	0%

Illustration

Consider the sentence: “In FY02, the Ag Division R&A spend was approximately 57%, 26%, 8%, 6%, 4% and 0% for Standard Warranty, PIP’s, Special Allowances, Service Parts, Extended Warranty, and Batteries, respectively.”

This is a pretty bad sentence.

The reader has to go to the end of the sentence to understand the structure, then go back and forth.

Use of “respectively” is usually not a good idea, and not even necessary in this case.

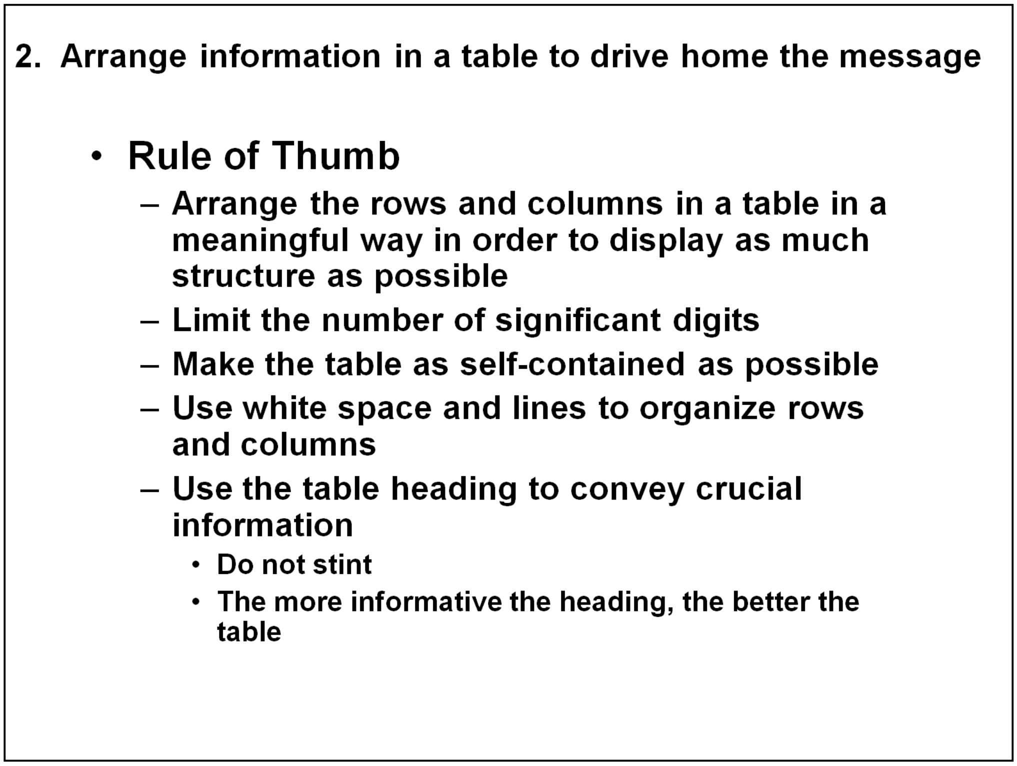
The sentence could have been re-written: “In FY02, the Ag Division R&A spend was approximately 57% Standard Warranty, 26% PIP’s, 8% Special Allowance, 6% Service Parts, 4% Extended Warranty and 0% Batteries.”

This structure immediately associates expense category with frequency.

But a simple sentence table, as shown above, is still better.

Note that the categories have been ordered, not by alphabet, but by the more meaningful characteristic, frequency.

The sentence has been displayed so that the information about expense category is clearly set off.



2. Arrange information in a table to drive home the message

- Rule of Thumb
  - Arrange the rows and columns in a table in a meaningful way in order to display as much structure as possible
  - Limit the number of significant digits
  - Make the table as self-contained as possible
  - Use white space and lines to organize rows and columns
  - Use the table heading to convey crucial information
    - Do not stint
    - The more informative the heading, the better the table

Introduction

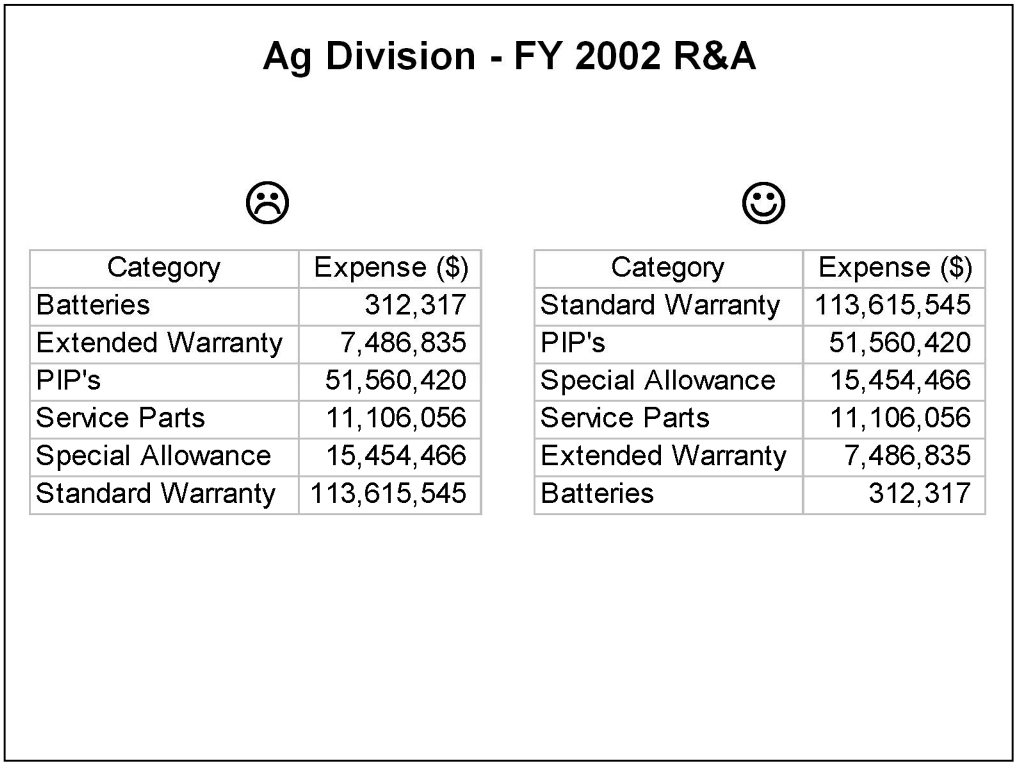
There is some flexibility in table structure and content. Structure deals with labeling, content with the values associated with the labels.

Basis of the Rule

Columns and rows of tables represent dimensions and must be used creatively and efficiently.

Labels may be arbitrary and an ordering based on labels may not display data structure.

(Consider what table on left on the next slide would have looked like with categories in Spanish.)



Ag Division - FY 2002 R&A

☹

Category	Expense ($)
Batteries	312,317
Extended Warranty	7,486,835
PIP's	51,560,420
Service Parts	11,106,056
Special Allowance	15,454,466
Standard Warranty	113,615,545

☺

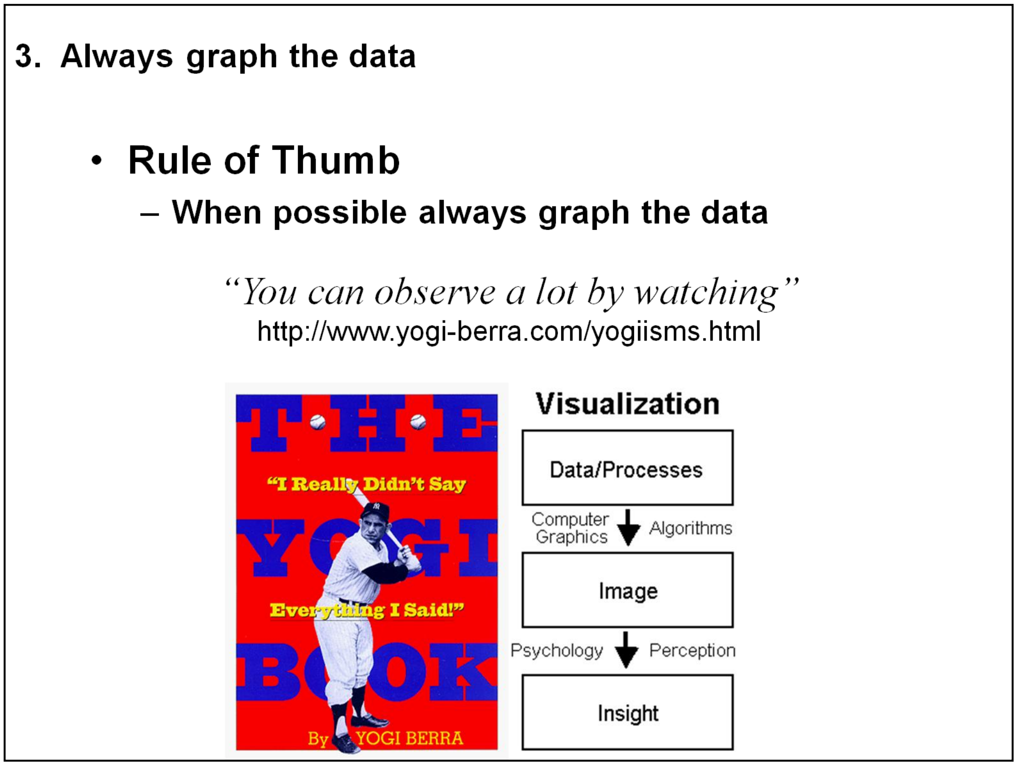
Category	Expense ($)
Standard Warranty	113,615,545
PIP's	51,560,420
Special Allowance	15,454,466
Service Parts	11,106,056
Extended Warranty	7,486,835
Batteries	312,317

Illustration

The table on the left is sorted alphabetically in ascending order by Category. This is a common output from Excel's Pivot Table feature. The Batteries category is at top of the list.

The table on the right is sorted numerically in descending order by Expense. The Batteries category in at the bottom of the list.

Note: we probability don't need that many significant digits either.



3. Always graph the data

- Rule of Thumb
  - When possible always graph the data

"You can observe a lot by watching"
http://www.yogi-berra.com/yogiisms.html

Visualization

Data/Processes

Computer Graphics ↓ Algorithms

Image

Psychology ↓ Perception

Insight

Introduction

The virtues of good graphics have already been touted.

A good graph illustrates patterns, identifies outliers, and shows relationships that were perhaps unanticipated.

There are any number of horror stories associate with ungraphed statistical analyses.

Tufte (1983) provides three criteria for graphical excellence.

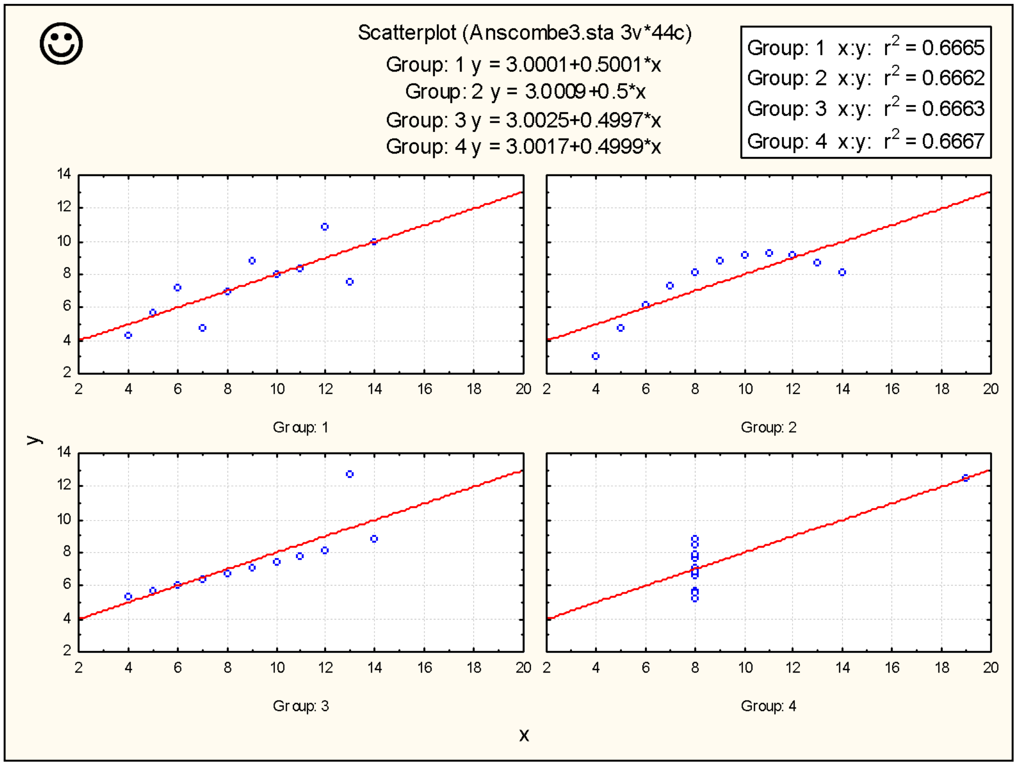
1. "Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, of *statistics*, and of *design*."

2. "Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency."

3. "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."

Basis of the Rule

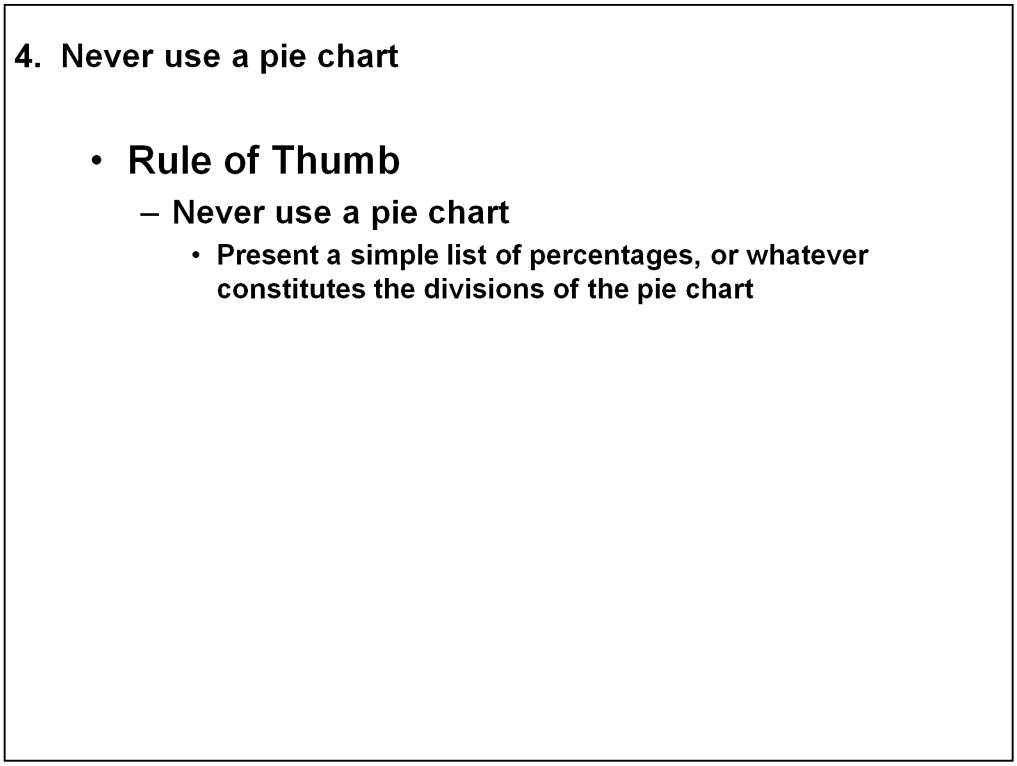
A good graph displays relationships and structures that are difficult to detect by merely looking at the data.



Illustration

Regression data with the property that for every set the means are equal, the regression lines are Y = 3 + 0.5*X, the standard errors of estimate of slope are 0.118, and the R-squared values are 0.667 (correlation coefficients are 0.82). Data from Anscombe (1977).

If one just looked at summary statistics from computer output, one would conclude that each group has the same relationship between X and Y. But looking at the graphs reveals that this is not true.



4. Never use a pie chart

- Rule of Thumb
  - – Never use a pie chart
    - Present a simple list of percentages, or whatever constitutes the divisions of the pie chart

Introduction

The most ubiquitous graph is the pie chart.

It is a staple of the business world.

Basis of the Rule

The pie chart has very low data density.

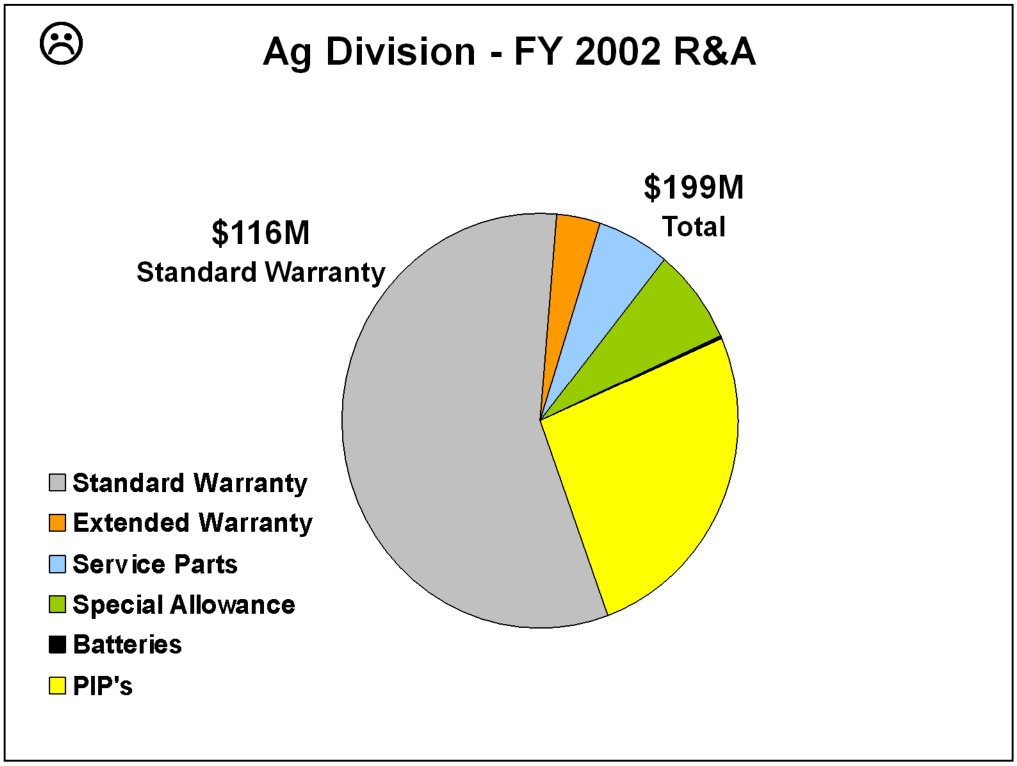
In the example above there are eight data points (actually seven since the sum of the percentages adds up to 100).

These points can be presented much better as a table.

Edward Tufte, in his classic *The Visual Display of Quantitative Information*, has this to say about pie charts:

“A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies... Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.”

http://www.krazydad.com/bestiary/bestiary_piechart.html

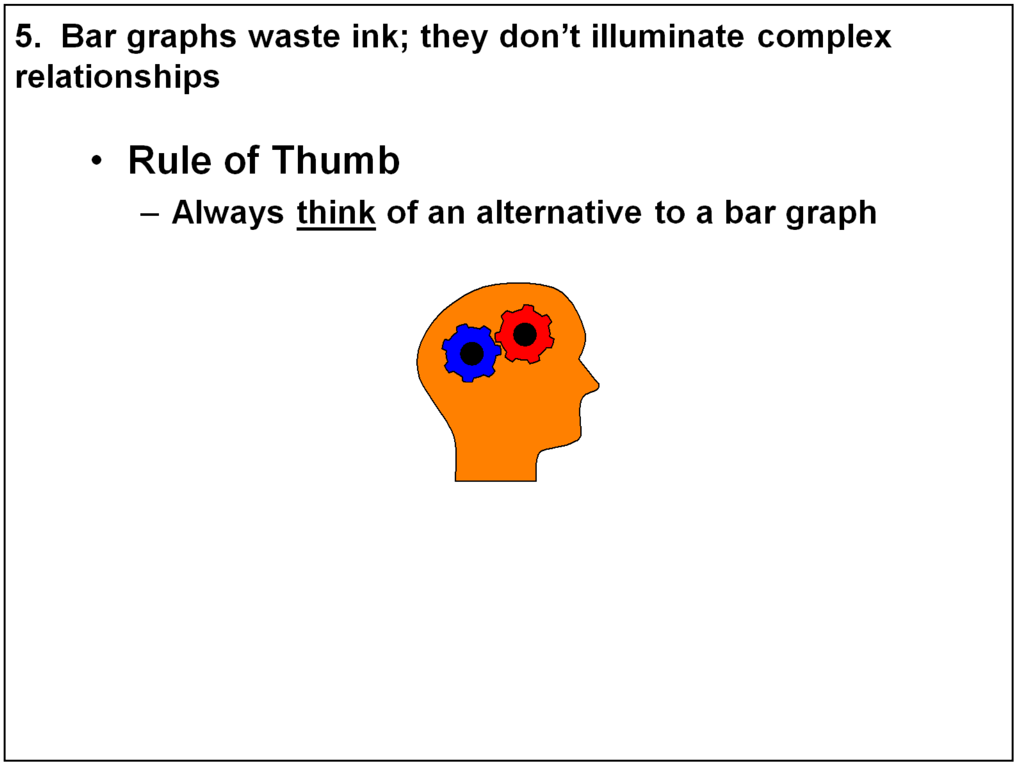


Illustration

The data presented in the pie chart above can be simply presented with a table:

Category	Expense ($)	Percent
Standard Warranty	113,615,545	57%
PIP's	51,560,420	26%
Special Allowance	15,454,466	8%
Service Parts	11,106,056	6%
Extended Warranty	7,486,835	4%
Batteries	312,317	0%
Total	199,535,639	100%

Note: we probability don't need that many significant digits either.



5. Bar graphs waste ink; they don’t illuminate complex relationships

- Rule of Thumb
  - Always think of an alternative to a bar graph

Introduction

Every meeting has presentations that involve bar graphs.

Basis of the Rule

The bar graph is far removed from the original data and there are now better ways of presenting data.

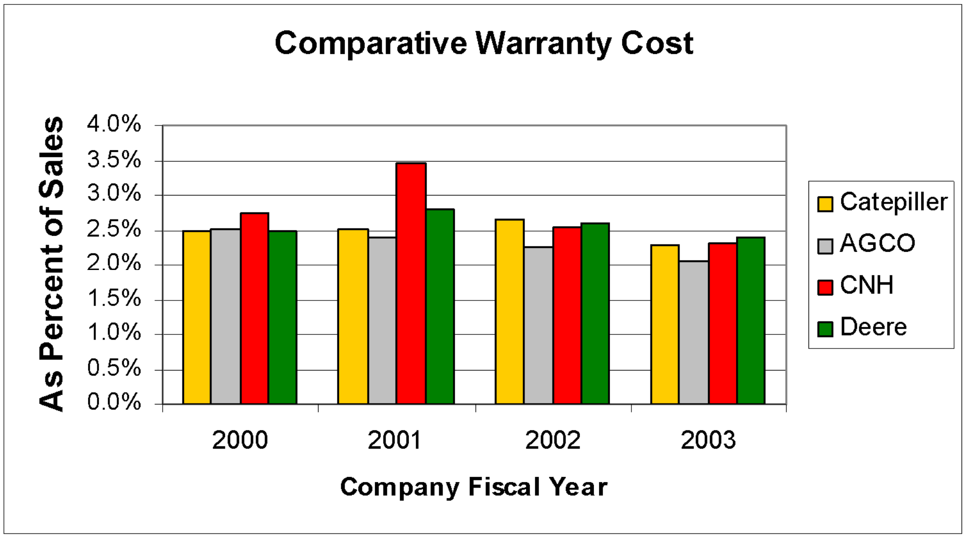
Alternatives take more thinking but it clearly pays off.

Bar graphs are prime examples of “chart junk” defined by Tufte as unnecessary use of ink.

The bars take up a great deal of ink, yet the only purpose is to indicate height.

Furthermore, a bar graph frequently ignores the underlying structure of the data, as the example indicates.

☹

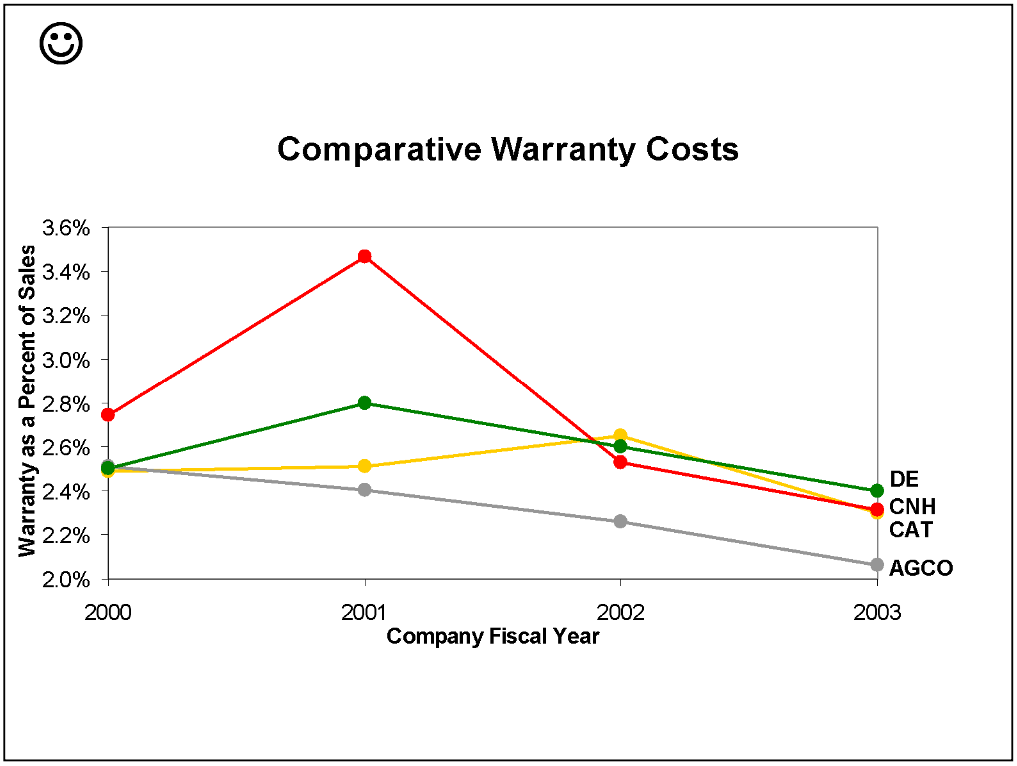


Illustration

Since the SEC adopted FASB Interpretation No. 45, publicly traded companies report their warranty spend and their reserves.

This chart shows warranty as a percent of sales for 2000 to 2003 for Deere and three similar companies.

Whenever there is a time dimension, an XY Chart would be a better choice than a bar chart.



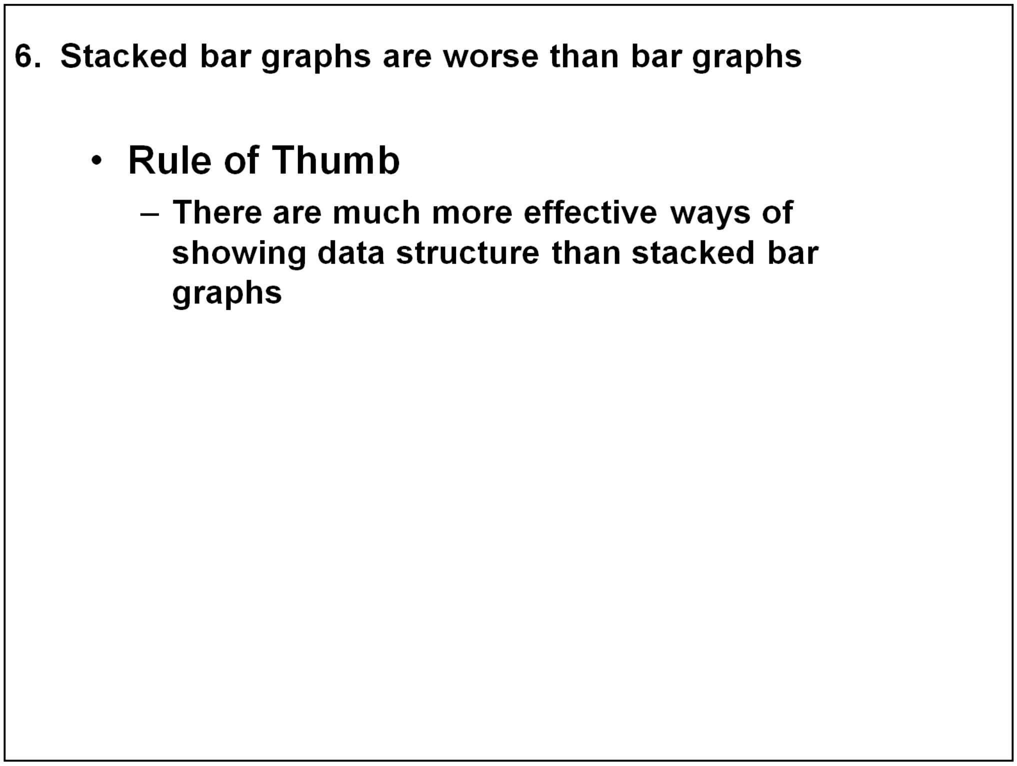
☺

Comparative Warranty Costs

Illustration (continued)

Here is the same data presented in a XY Chart with Time on the X-axis and Warranty on the Y-axis.

What do you notice now?



6. Stacked bar graphs are worse than bar graphs

- Rule of Thumb
  - There are much more effective ways of showing data structure than stacked bar graphs

Introduction

Stacked bar graphs are seen even more frequently than bar graphs.

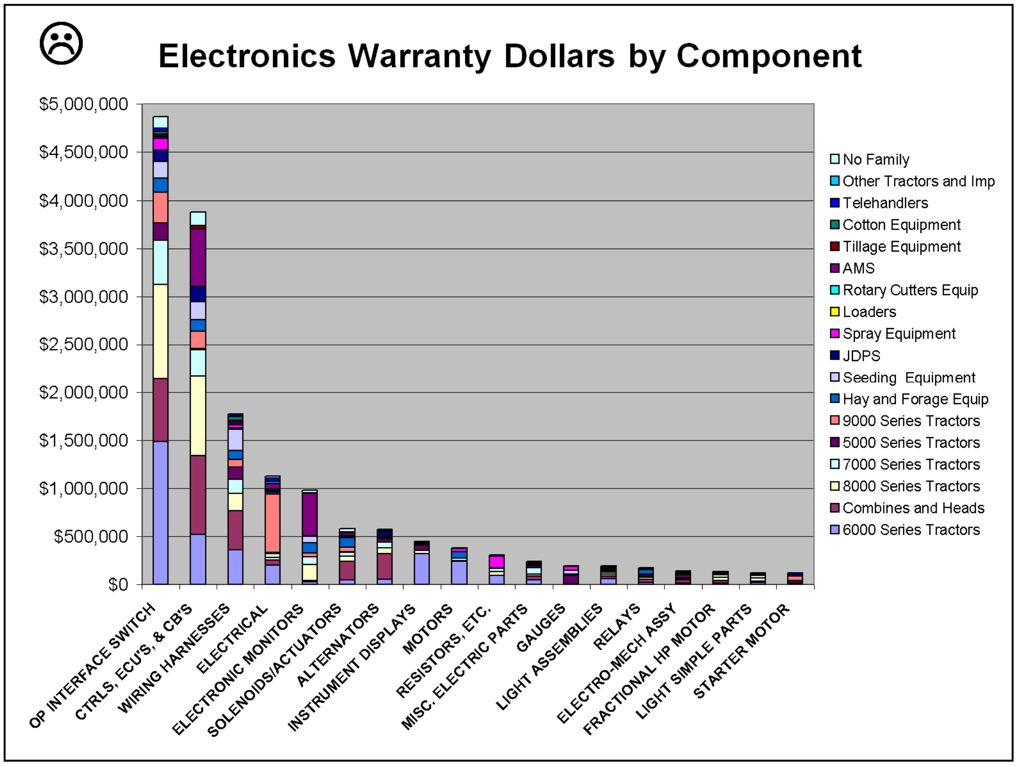
Their putative purpose is to show structure in the data.

Basis of the Rule

Stacked bar graphs do not show data structure well.

A trend in one of the stacked variables has to be deduced by scanning along the vertical bars.

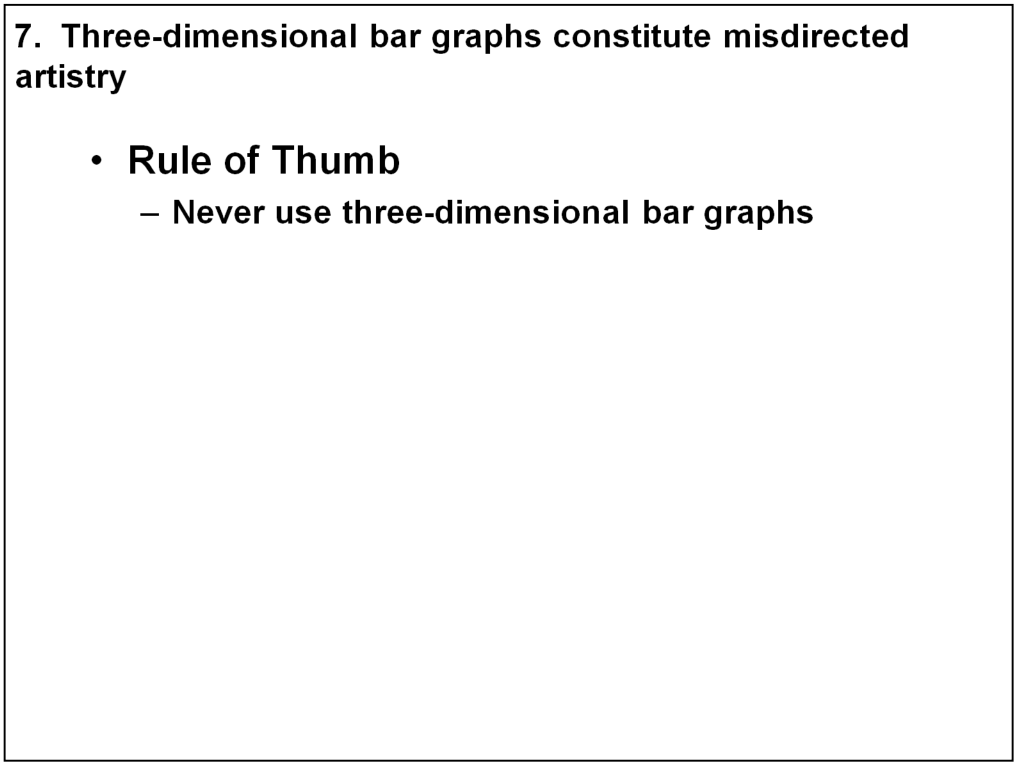
This becomes especially difficult when the categories do not move in the same direction.



Illustration

Here is a typical stacked bar chart. What can you tell from it?

We will look at alternatives later.



7. Three-dimensional bar graphs constitute misdirected artistry

- Rule of Thumb
  - Never use three-dimensional bar graphs

Introduction

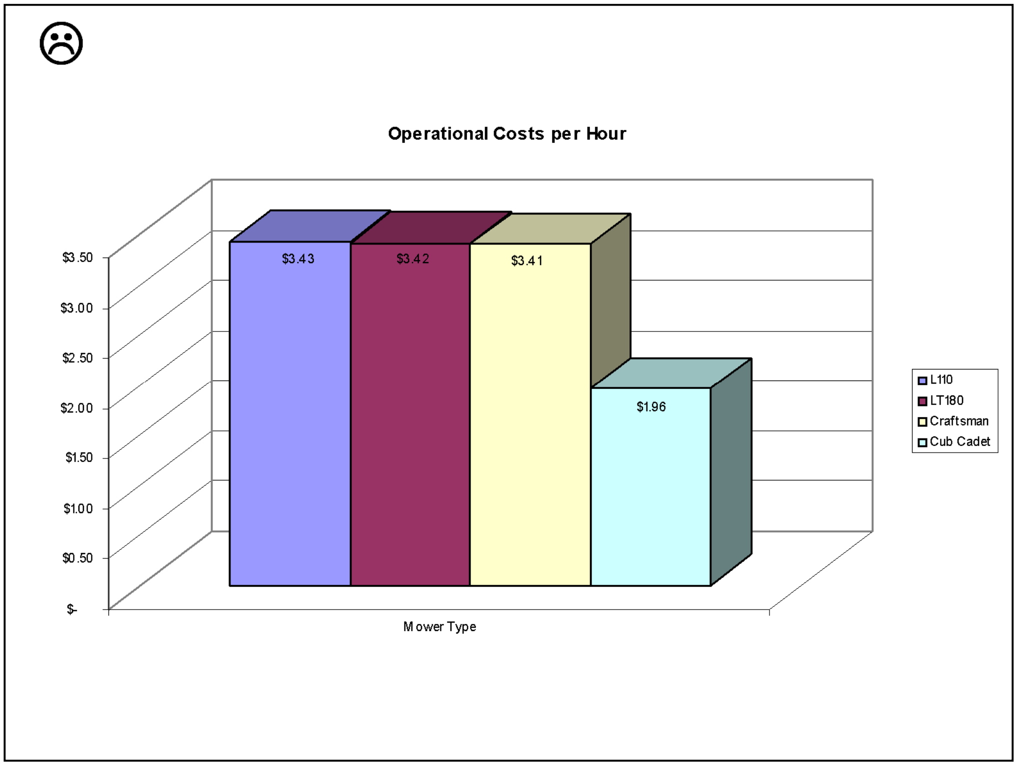
Graphs can be embellished in a variety of ways.

In Victorian England this might have included zephyrs blowing across the graph.

In the 21st century there are even worse embellishments.

Basis of the Rule

The embellishment of a third dimension creates confusion.

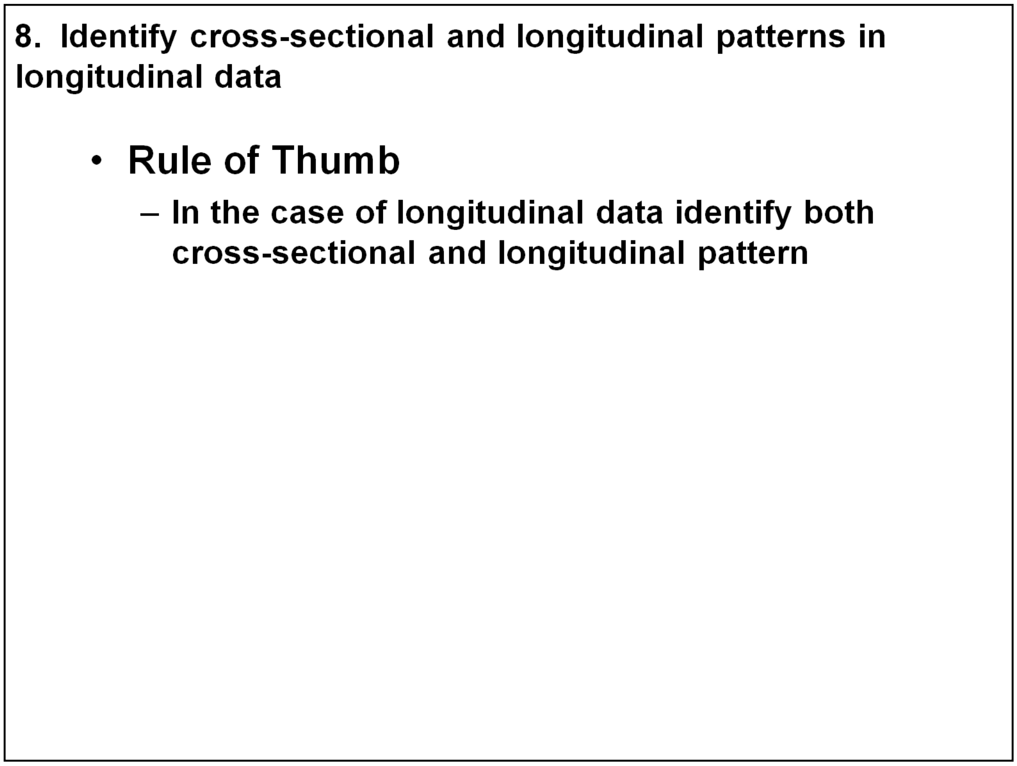


Illustration

This 3-D bar chart shows the operational cost per hour of 4 different mowers: L110, LT180, Craftsman and Cub Cadet.

Distortion is introduced by adding extra dimension to a bar graph.

Also, there is no reason to plot 4 numbers – a simple table will do.



8. Identify cross-sectional and longitudinal patterns in longitudinal data

- Rule of Thumb
  - In the case of longitudinal data identify both cross-sectional and longitudinal pattern

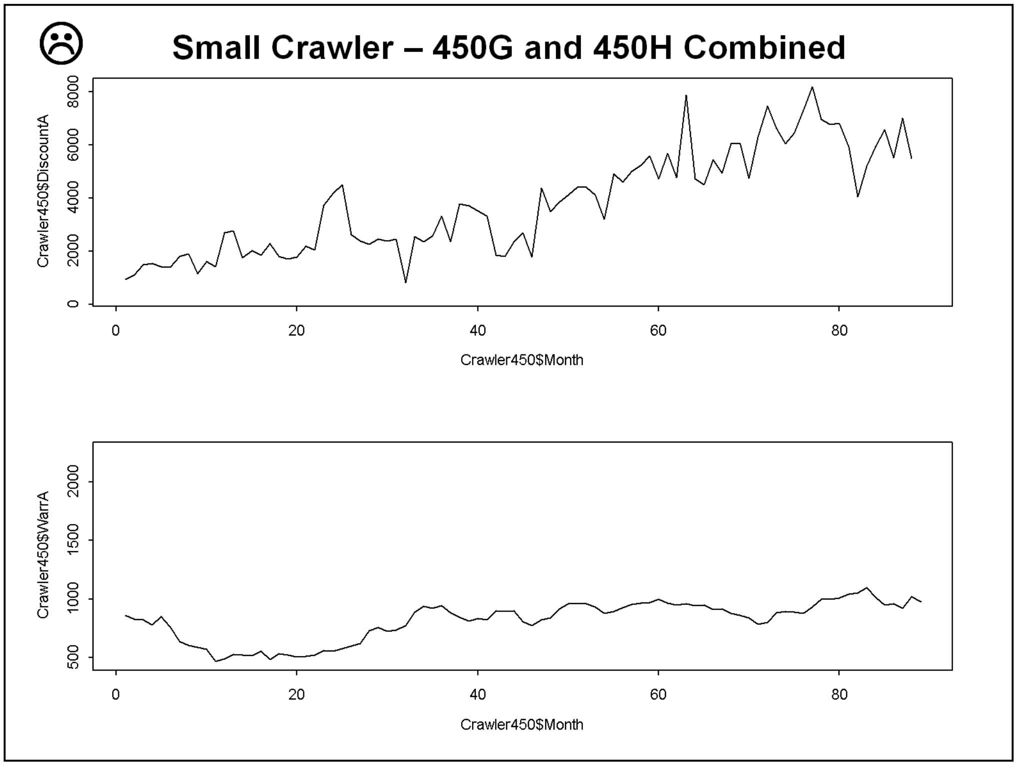
Introduction

The last twenty years have seen an explosion of methods for the analysis of longitudinal data, in part due to increased computational resources.

The graphical presentation of data presents special challenges because frequently there is both cross-sectional and longitudinal information.

Basis of the Rule

When there is more than one source of variation it is important to identify those sources.



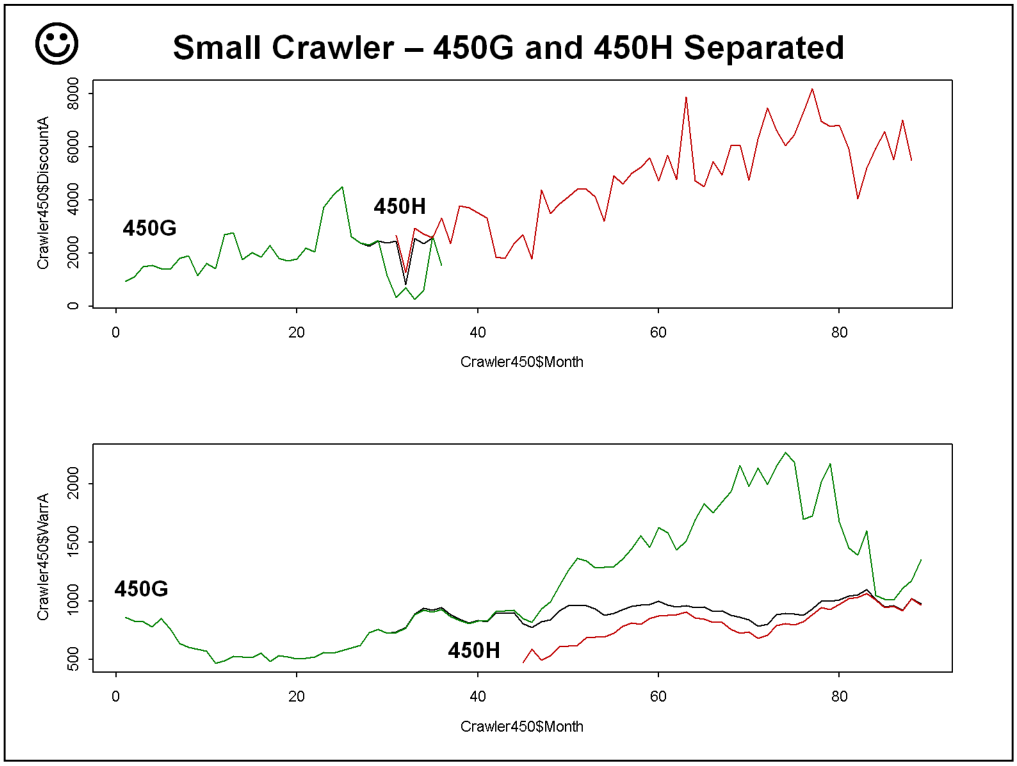
☹ Small Crawler – 450G and 450H Combined

Illustration

A longitudinal study is one that examines data over time [note: when money is involved, it is important to adjust for inflation].

Here are plots for an 8 year study of discounts and warranty for the 450 Small Crawler series, G and H combined, adjusted for inflation.

There doesn’t seem to be much of a relationship over time between discounts and warranty.



Small Crawler – 450G and 450H Separated

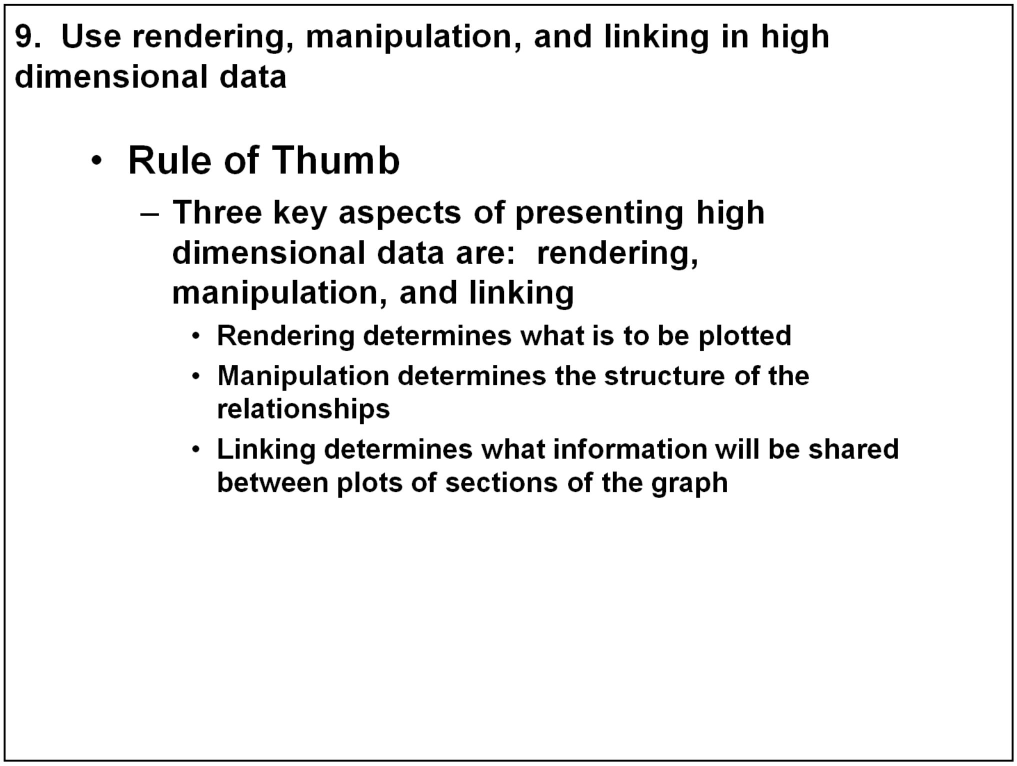
Illustration (cont.)

The green lines are for the 450G; the red lines are for the 450H; and the black lines are for the two combined (same as previous chart).

Once the G and H are separated, you can see something strange: H series discounts are rising as the G series warranty is rising.

This could be just a coincidence, or maybe the poor warranty on the G series is souring customers on the H series.

The point is that if you merge together data sets of non-homogenous subpopulations, you are likely to miss what is going on. This is a significant disadvantage of using Rolling 12 or “L18” warranty metrics.



9. Use rendering, manipulation, and linking in high dimensional data

- Rule of Thumb
  - Three key aspects of presenting high dimensional data are: rendering, manipulation, and linking
    - Rendering determines what is to be plotted
    - Manipulation determines the structure of the relationships
    - Linking determines what information will be shared between plots of sections of the graph

Introduction

A great deal of statistical data is high-dimensional – for example, multiple regression data with dozens of predictor variables.

Visualizing these kinds of data has intrigued and challenged statisticians for many years.

The confluence of high capacity computing, statistical methodological advances, and huge data sets has focused attention on graphical displays of high-dimensional data.

Since humans are restricted to visualizing at most three dimensions the challenge is how to approach higher–dimensional data.

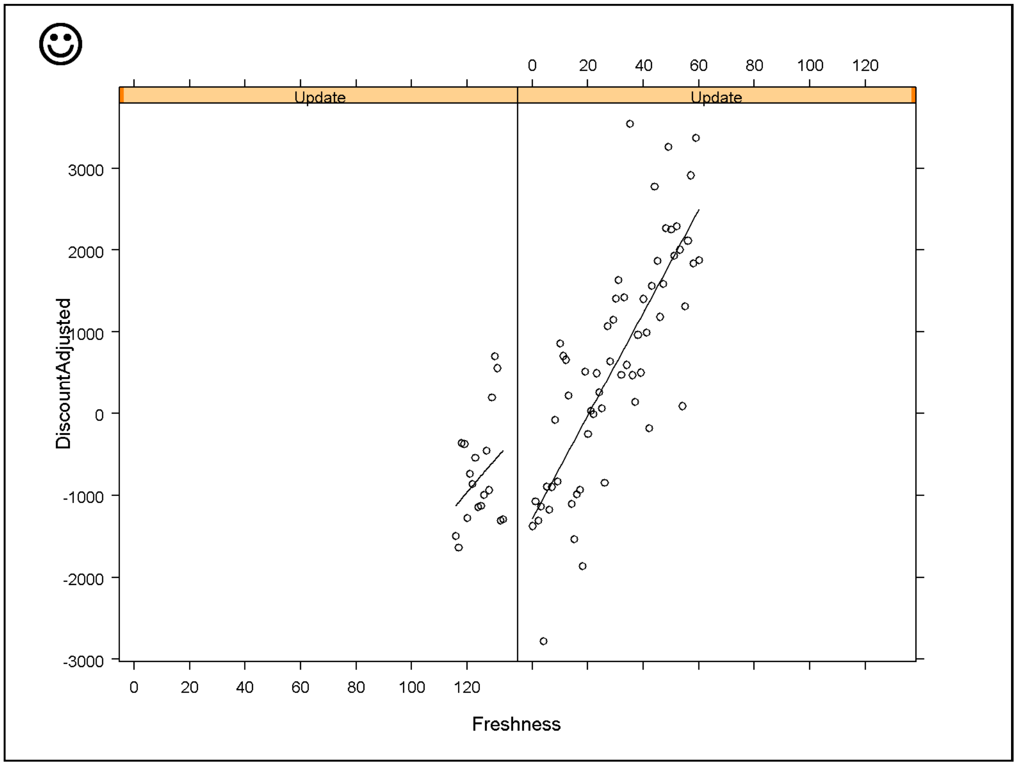
The ideas in this section are based on the work of Sutherland et. al. (2000).

Basis of the Rule

Rendering, manipulation, and linking are basic ways in which relationships can be categorized.

The choice of which aspects to render, manipulate, and link, are, of course, the hard thing to do.

But the rule provides a means of attacking the problem of graphing high-dimensional data.



Illustration

This plot shows several variables at the same time:

•Discounts adjusted for warranty on the Y-axis [DiscountA – 23(Warr2A9)]

•Model Freshness (months since introduction) on the X-axis

•Model Update = 0 (no update occurred - just 450 Crawler G series) on the left side, and model Update = 1 (updated occurred - 450 Crawler G and H series) on the right side

Regression equation is: DiscountA – 23(Warr2A9) = -8,521 + 63(Freshness) + 7,259(Update), $R^2 = 0.671$

This model suggest that:

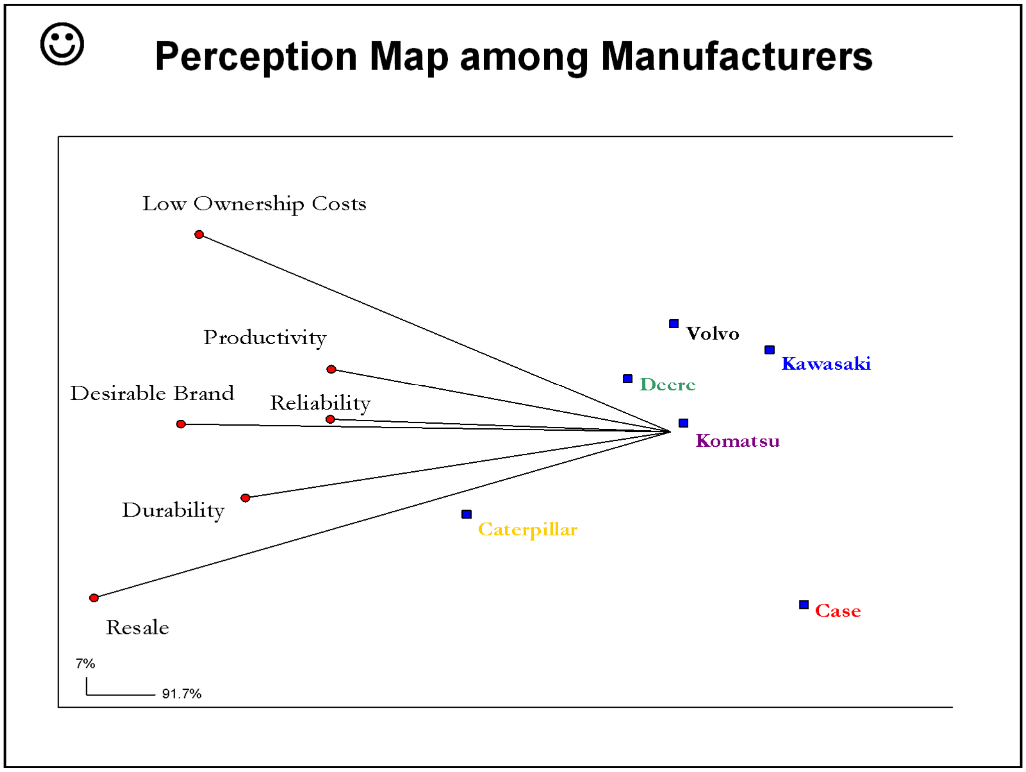
•Holding Update and Freshness constant, $1 in warranty (7-12 month, lagged by 9 months) increases discounts by $23

•Holding warranty and Update constant, each month the product line ages increases discounts by $63

•Holding warranty and Freshness constant, the Update increased discounts by $7,300

•If warranty, Freshness and update were all 0, price could be increased by $8,500 (i.e., discount of negative $8,500)

Just looking at the plot suggests that discounts at the introduction of the H series were the same as the end of the G series, but the discounts steadily increased since the introduction of the H series. Why? It probability has something to do with the economy and the competition.



☺ Perception Map among Manufacturers

Illustration

Perception maps are a common marketing tool that utilize multivariate statistics. This one represents 6 dimensions of data in 2 dimensional space.

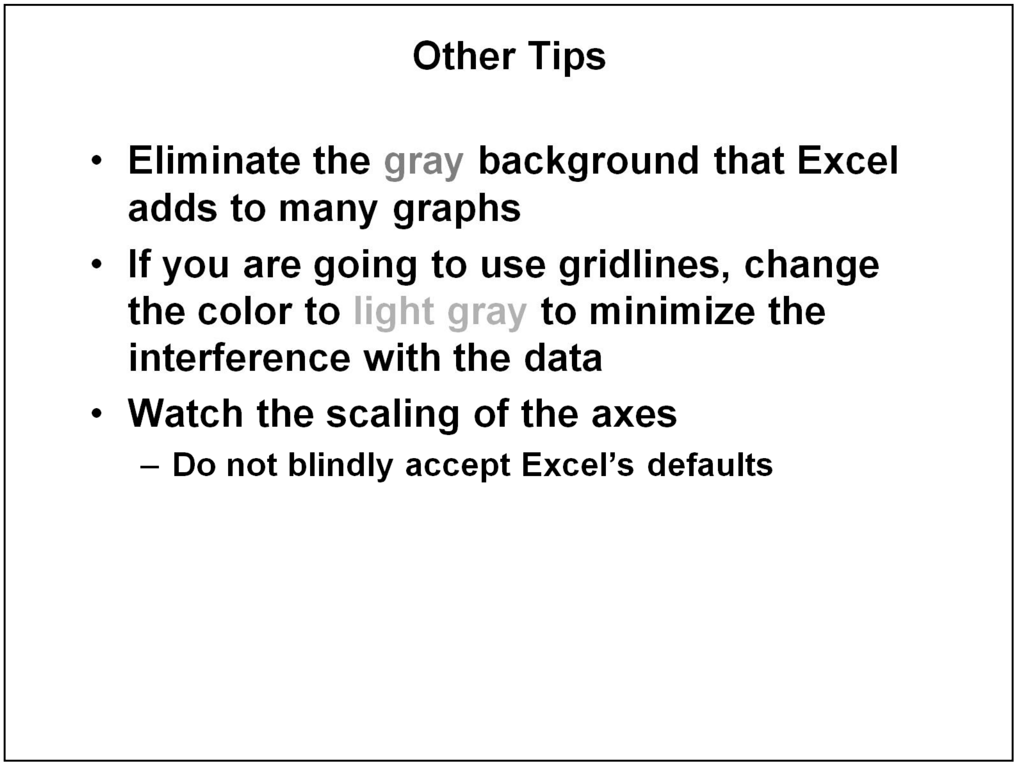
Some conclusions from the data:

•Durability and Resale Value are very highly correlated.

•Low Ownership Costs and Resale Value are not highly correlated.

•Reliability, Desirable Brand, and Productivity are somewhat correlated.

•Cat is consistently rated the highest in all categories while Deere is consistently rated the next highest.



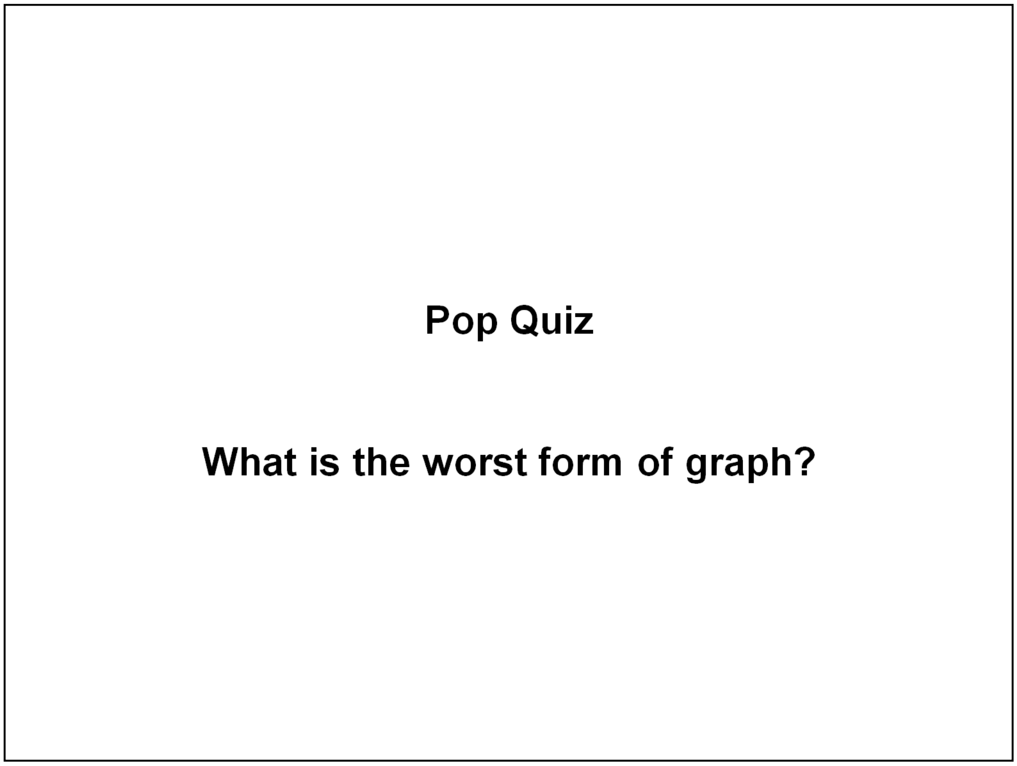
Other Tips

- Eliminate the gray background that Excel adds to many graphs
- If you are going to use gridlines, change the color to light gray to minimize the interference with the data
- Watch the scaling of the axes
  - Do not blindly accept Excel's defaults

Exercise: Explore the following web sites

http://www.math.sfu.ca/~cschwarz/Stat-301/Handouts/node8.html

http://www.math.yorku.ca/SCS/Gallery/



Pop Quiz

What is the worst form of graph?

Hint: It was not covered in this section, but it is a combination of two of the concepts covered.