Lesson 02: What are the three rules of data analysis?

Q: Why do you think it is essential to know the W's and H of the data?

The Three Rules of Data Analysis

The three rules of data analysis won't be difficult to remember:

- Make a picture—things may be revealed that are not obvious in the raw data. These will be things to *think* about.
- Make a picture—important features of and patterns in the data will *show* up. You may also see things that you did not expect.
- Make a picture—the best way to *tell* others about your data is with a well-chosen picture.

What pictures?

What kind of pictures?

Are there different pictures for different type of variables?

11:40 on the night of April 14, 1912	
WHO	People on the *Titanic*
WHAT	Survival status, age, sex, ticket class
WHEN	April 14, 1912
WHERE	North Atlantic
HOW	A variety of sources and Internet sites
WHY	Historical interest

1. Make piles - Frequency Tables

- We can "pile" the data by counting the number of data values in each category of interest.

Coin

- We can organize these counts into a frequency table, which records the totals and the category names.

Titanic Passengers

Class	Count
First	325
Second	285
Third	706
Crew	885

2. Relative Frequency Table - gives the percentages (instead of counts) for each category.

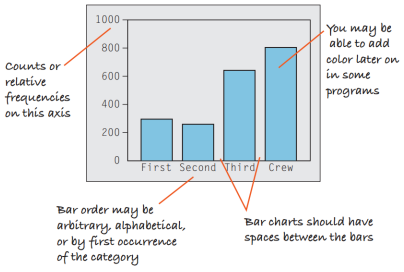
Titanic Passengers

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

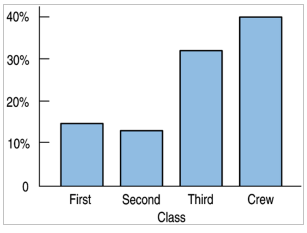
Distribution:

The frequency tables give us the distribution of the categorical variables. They name the possible categories and tell us how frequently each occurs.

3. Bar Chart - shows a bar whose area represents the count (or percentage) of observations for each category of the categorical variables.



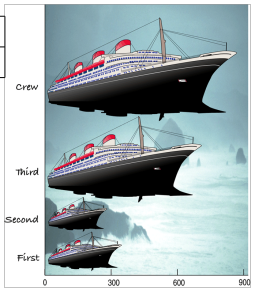
Relative Frequency Bar Chart - replace counts with percentages



Advantages and Disadvantages of Bar Charts

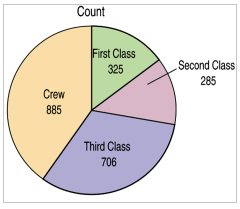
Advantages	Disadvantages
Summarize large data set in visual form	Can be manipulated to yield false impressions (via arrangement of bars for example)
Clarify trends better than do tables	Can fail to reveal key patterns
Estimate key values at a glance	
Can compare two or three data sets	
Be easily understood	

What's wrong with this picture?
We notice area instead of length



Area Principle: The area occupied by a part of the graph should correspond to the magnitude of the value it represents.

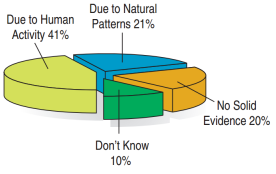
4. Pie Charts - shows how a "whole" divides into categories. The area of each wedge of the circle corresponds to the proportion in each category.



* Notice how a different display creates a different focus for your eyes and brain.

Advantages	Disadvantages
Summarize large data set in visual form	Can be manipulated to yield false impressions (other category, Total unknown, slanted pie)
Visually simpler than other graphs	no exact numerical data
Be easily understood	Too many categories are confusing
	Small or categories of similar size are a problem too. Dominating categories attract attention.

Global Warming. The Pew Research Center for the People and the Press (http://people-press.org) has asked a representative sample of U.S. adults about global warming, repeating the question over time. In January 2007, the responses reflected an increased belief that global warming is real and due to human activity. Here's a display of the percentages of respondents choosing each of the major alternatives offered:

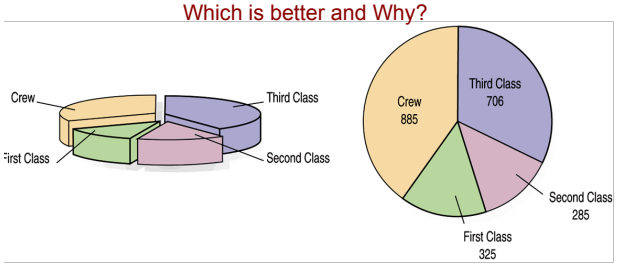


List the errors in this display.

Sample Response

Perhaps the most obvious error is that the percentages in the pie chart only add up to 92%,when they should, of course, add up to 100%. Furthermore, the three-dimensional perspective view distorts the regions in the graph, violating the area principle. The regions corresponding to No Solid Evidence and Due to Natural Patterns should be roughly the same size, at 20% and 21% of respondents, respectively. However, the angle for the 21%region looks much bigger. Always use simple, two-dimensional graphs.

Which is better and Why?



easier to compare parts of a whole in 2D

This plot of the percentage of high-school students who engage in specified dangerous behaviors has a problem. Can you see it?