

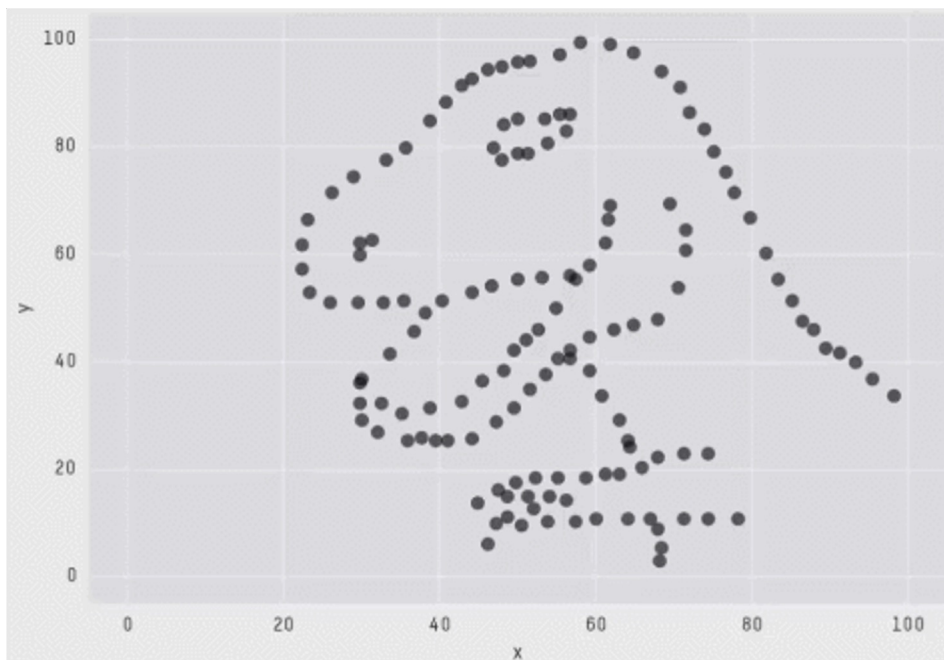
Lies, Damned Lies, and Statistics

A dark blue, solid-colored shape that starts from the bottom left corner and extends diagonally upwards towards the right, covering the bottom half of the slide.

General Rules

1. Look at the data
1. Distrust anything given without context
1. Actually, maybe don't even trust the context
1. Just don't trust anyone

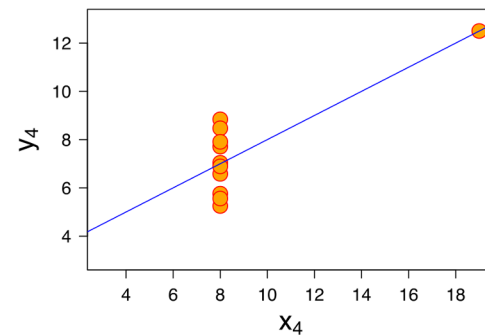
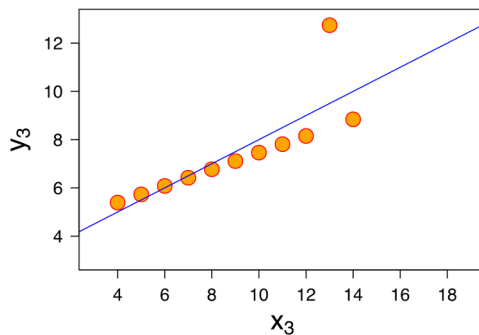
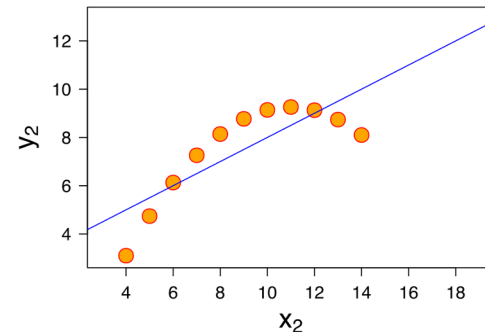
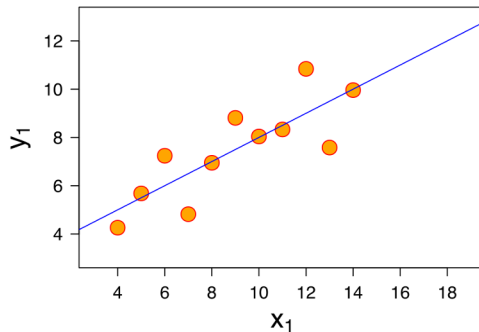
Why Visually Inspecting the Data Matters



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Anscombe's Quartet

	X	Y
Mean	9	7.50
Sample Variance	11	4.125
Correlation	0.816	
Linear regression line	$Y = 3.00 + 0.500x$	
R^2	0.67	



Averages

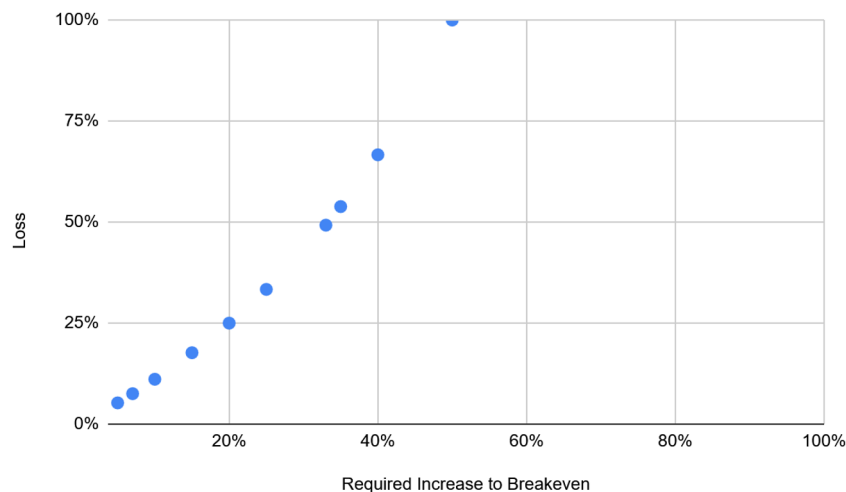
- Never trust an average without context
- Example
 - 9 employees, each making \$75,000/year
 - 1 CEO, making \$1,000,000
 - CEO can claim that “the average worker earns almost \$160,000 a year!”

Sampling

- 9 out of 10 is 90%
- 900 out of 1000 is 90%

- Claims
 - 90% of dentists recommend ____
 - 75% of voters support ____
 - 100% of Senior Data Analysts are GOATed
 - ____% of doctors smoke Marlboro
- Reality: If you run enough samples, you'll find one that fits your agenda
- Survey size
- Survey questions
 - 100% of political surveys
- Phrasing results
 - Toothpaste surveys had dentists rank toothpastes
 - Colgate then turned those rankings into "recommendations"
 - So *even if* 90% of dentists ranked Colgate higher than "most" other brands, they may not have ranked it #1 overall

Tangent: Percents



- Volatility Drag
- Percentage Change Asymmetry
 - Asymmetry of returns
- You buy a share of an ETF for \$100
- It decreases 5%, to \$95
- To get back to your breakeven price of \$100, the price needs to rise by 5.26%

- Real world story:
- Union workers had their pay cut by 20%
- The following year, they received a 5% pay increase
- NYT reported they got $\frac{1}{4}$ of their pay back
- Reality: they got $\frac{1}{5}$ of the pay cut back

Correlation -> Causation?

- Correlation does not imply causation
- Offenders **will** be mocked
- r
 - Pearson correlation coefficient
 - Range: -1.00 to 1.00
 - Measures the degree to which two variables move together, and in what direction
- r^2
 - Coefficient of determination
 - The percent of change in one variable that is predictable based on the change in the other variable
 - Range: 0 to 100%
- Spurious Correlation - Random

Simpson's Paradox

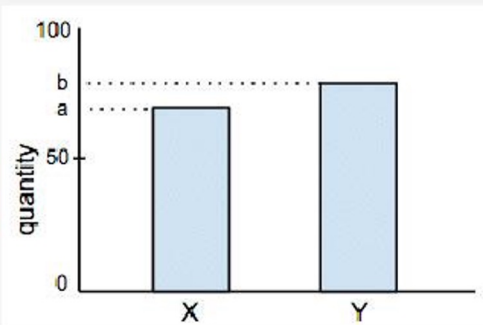
Batter	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	.310
David Justice	104/411	.253	45/140	.321	149/551	.270

- "...a trend appears in several groups of data but disappears or reverses when the groups are combined."
- Gender wage gap
- Wine with dinner

Chart Crimes

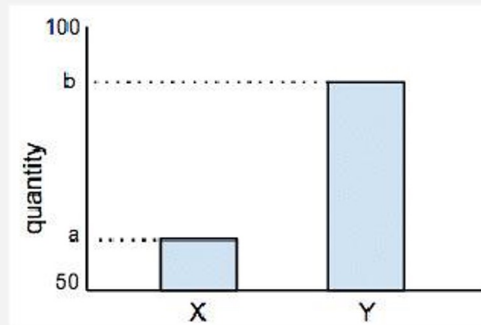
[r/ChartCrimes](#)

This is a whole other presentation, but the same basic rule applies:
context, context, context



Control

vs.



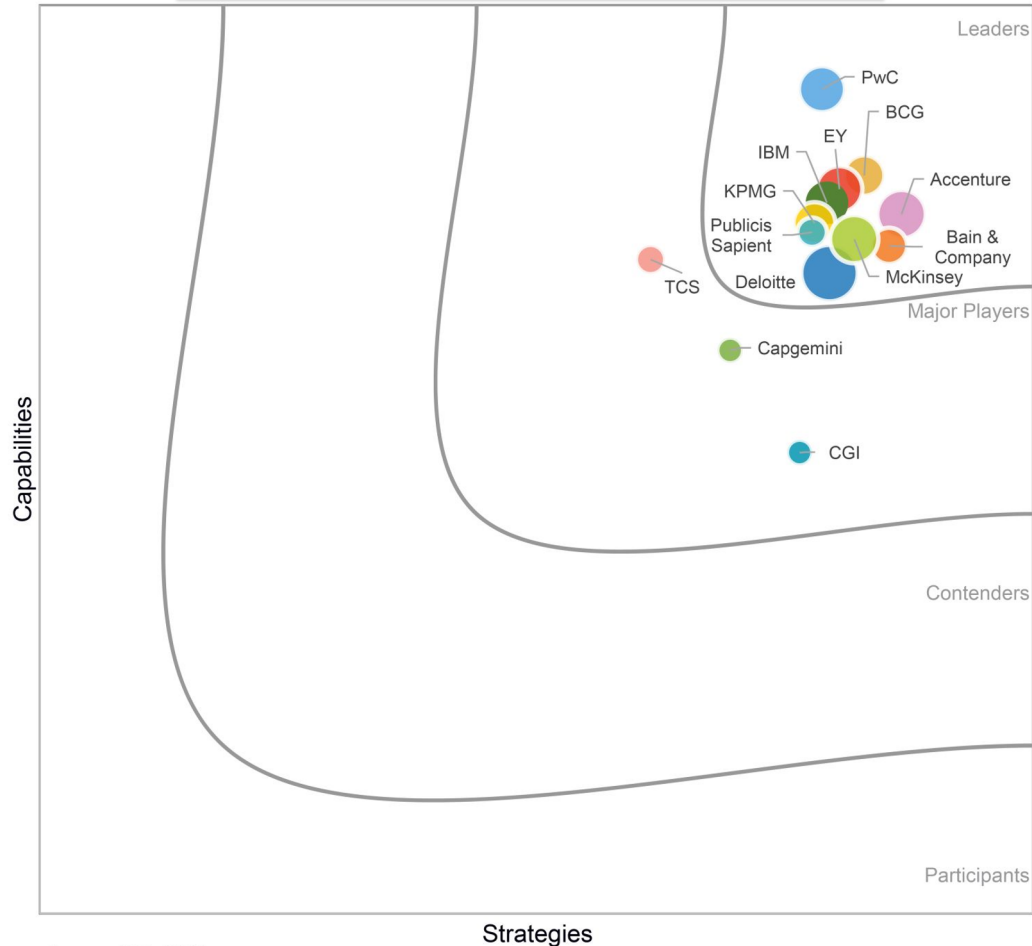
Deceptive

Cumulative



Straight Up Made Up

IDC MarketScape: Worldwide Digital Strategy Consulting Services, 2021



Source: IDC, 2021

Sources & Further Reading

- [Datasets from the Datasaurus Dozen](#)
- [Anscombe's quartet - Wikipedia](#)
- [Spurious Correlations](#)
- [Misleading Statistics Can Be Dangerous \(Some Examples\)](#)
- [How to Lie with Statistics](#)