Data Munging

Closer Look into Data Visualisation

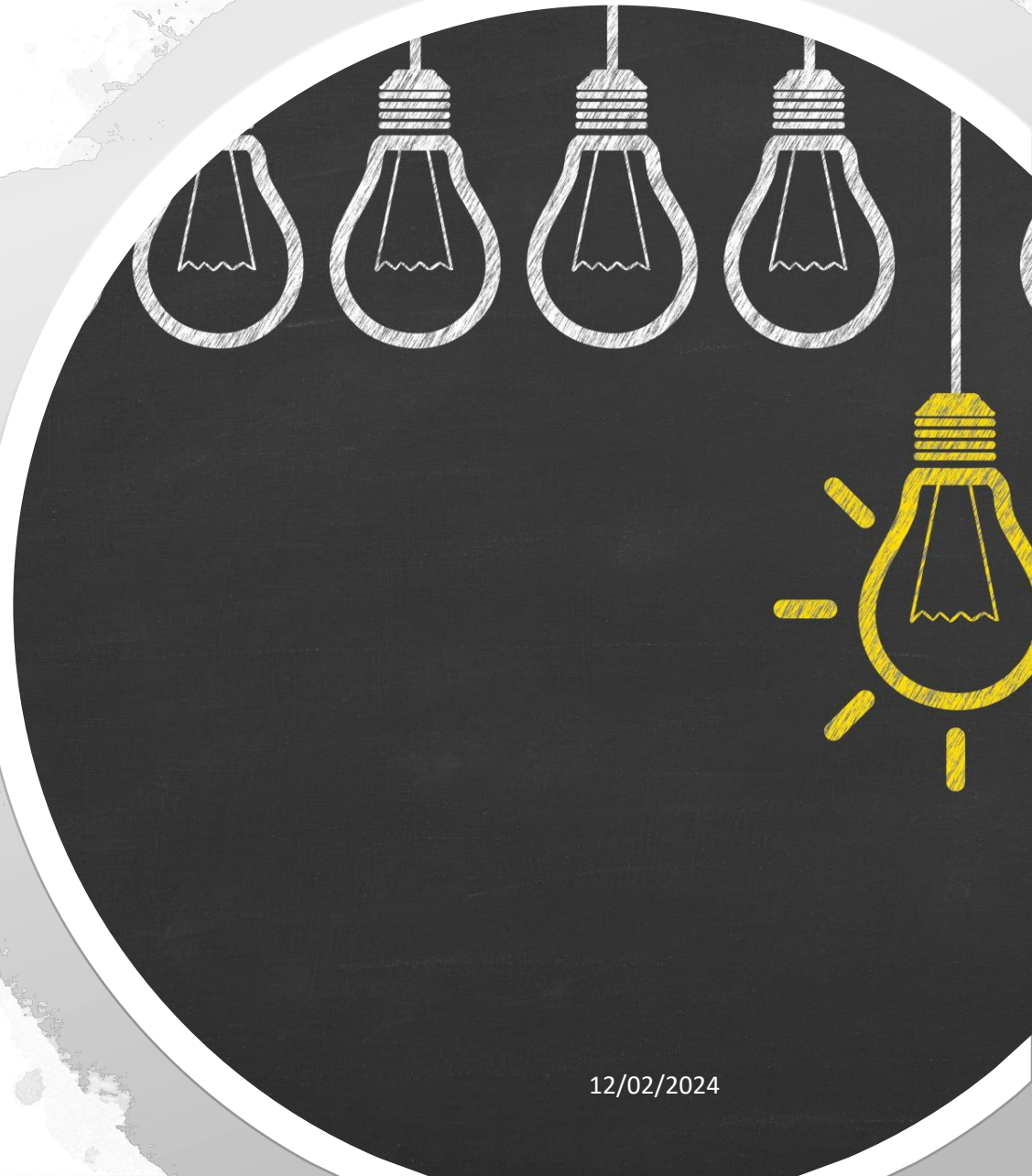
Areej Alasiry

Lecture Outline

- EDA via Visualisation
- Developing a Visualisation Aesthetic
- Chart Types
- Great Visualisations
- Reading Graphs

Effective Data Visualization

- Exploratory Data Analysis
- Error Detection
- Communication



Areej Alasiry

12/02/2024

Visualizing Data for EDA

Exploratory Data Analysis via Visualization

- Hypothesis driven vs data driven
- EDA is the search of patterns and trends in the dataset
- Using visualization, EDA helps in
  - Error detection
  - Finding violations of statistical assumption
  - Suggest interesting hypothesis

Confronting a New Dataset

- Answer the basic questions
  - Who constructed the dataset, when, and why?
  - How bias is it?
  - What are the attributes/fields/features represents or means?
- Examine a subset of the records
- Get the statistical summary of the dataset
- Pairwise correlation
- Class breakdown

Example (Pairwise Correlation)

	Age	Weight	Height	Leg Length	Arm Length	Arm Circum	Waist
Age	1.000						
Weight	0.017	1.000					
Height	−0.105	0.443	1.000				
Leg_Len	−0.268	0.238	0.745	1.000			
Arm_Len	0.053	0.583	0.801	0.614	1.000		
Arm_Circ	0.007	0.890	0.226	0.088	0.444	1.000	
Waist	0.227	0.892	0.181	−0.029	0.402	0.820	1.000

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Example (Statistical Summary Vs Visualisation)

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	

Figure 6.2: Four data sets with identical statistical properties. What do they look like?

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Example (Statistical Summary Vs Visualisation)

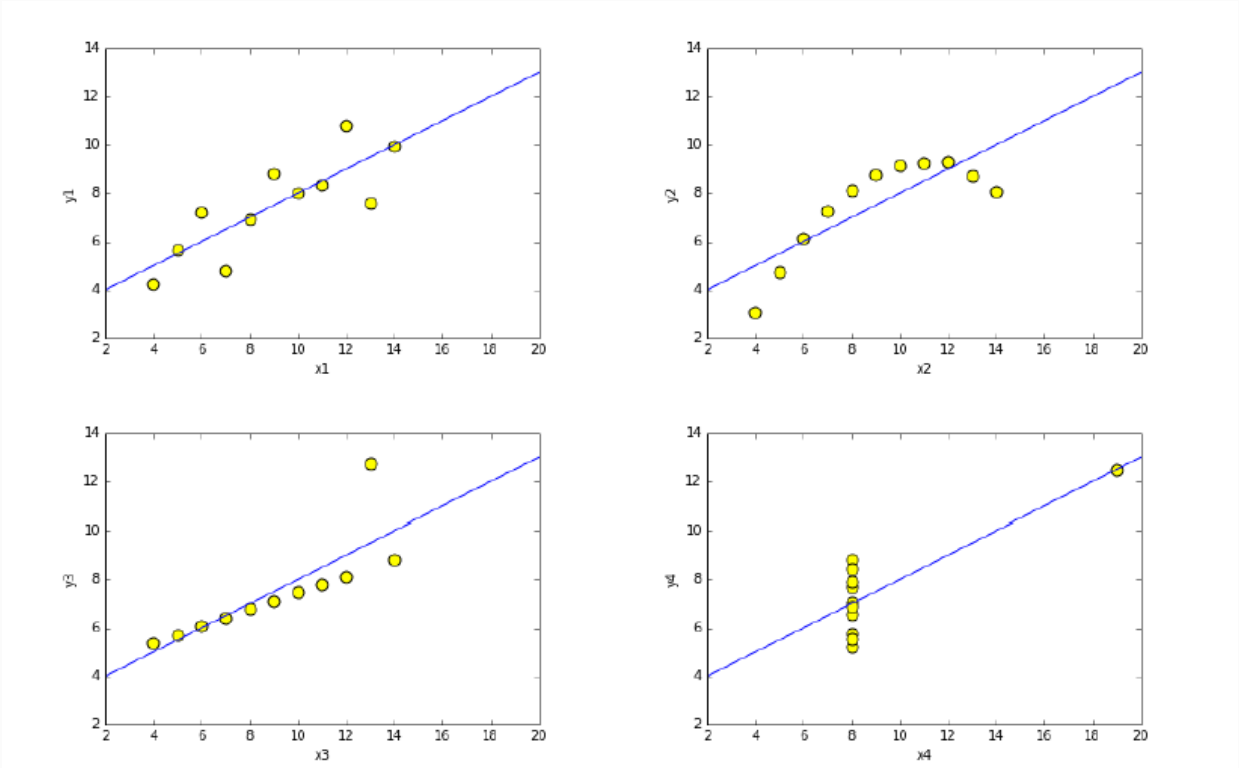


Figure 6.3: Plots of the Ascombe quartet. These data sets are all dramatically different, even though they have identical summary statistics.

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Developing a Visualization Aesthetic

Design Aesthetic

- What makes a graph or a chart *informative?*
- Maximize data-ink ratio
- Minimize the lie factor
- Minimize chart junk
- Use proper scales and clean labels
- Make effective use of colour
- Exploit the power of repetition

Example
Data-Ink Ratio

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

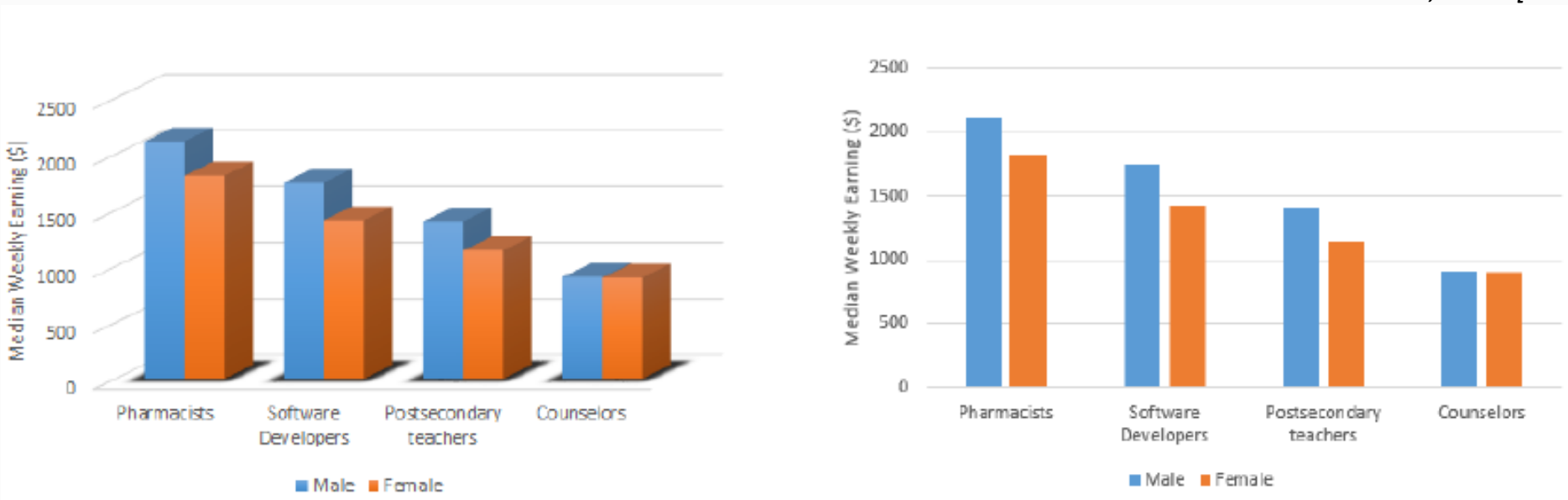


Figure 6.5: Three-dimensional monoliths casting rendered shadows (l) may look impressive. But they really are just chartjunk, which serves to reduce the clarity and data-ink ratio from just showing the data (r).

Example
Minimizing the Lie Factor

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

- Presenting means without variance
- Presenting interpolations without the actual data
- Distortions of scale
- Eliminating tick labels from numerical axes
- Hide the origin point from the plot

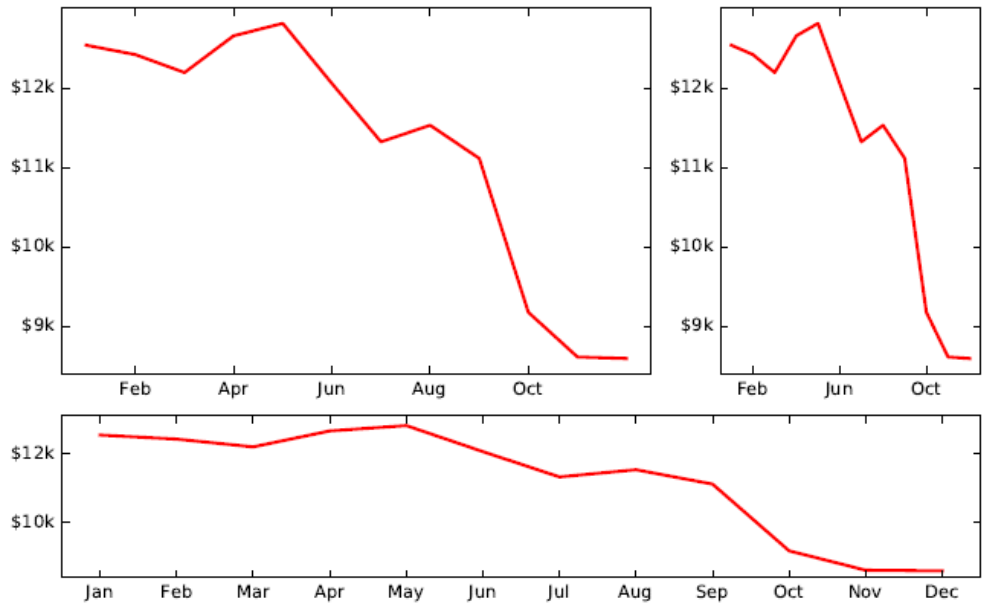


Figure 6.6: Three renderings of the same financial time series. Which most accurately represents the situation?

Example
Minimizing Chart Junk

- How can we improve the chart?

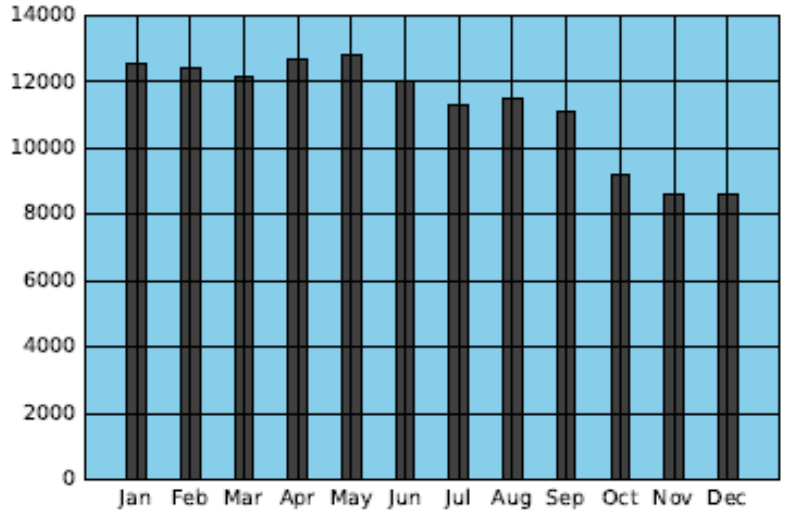


Figure 6.7: A monthly time series of sales. How can we improve/simplify this bar chart time series?

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Example
Minimizing Chart Junk

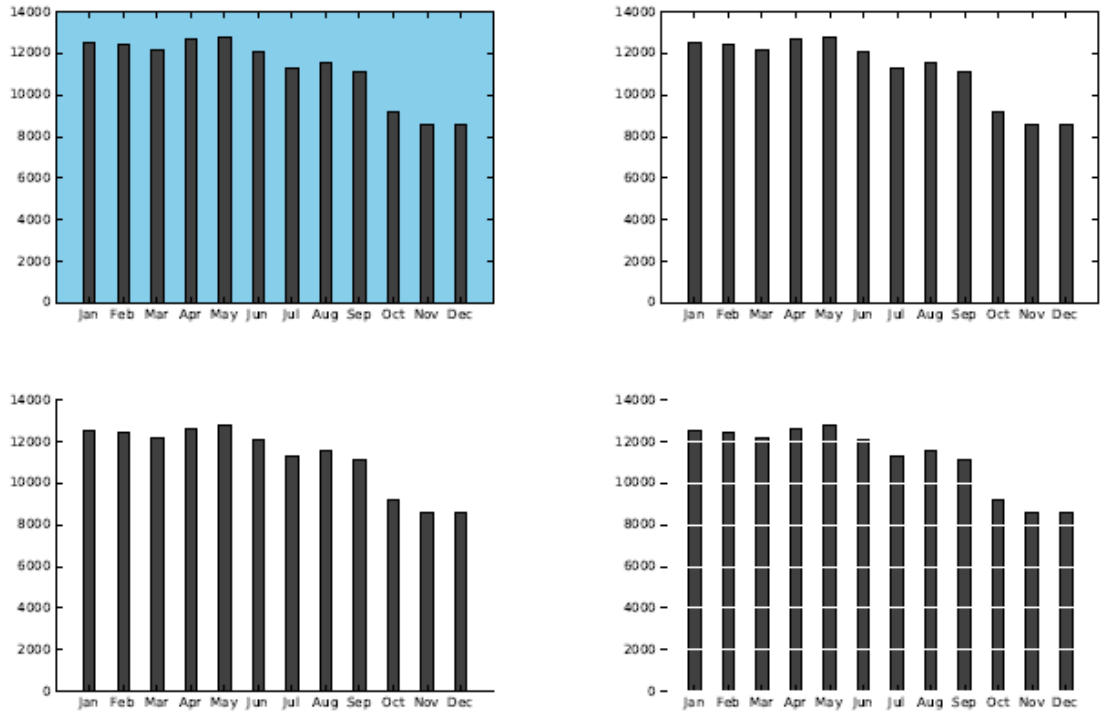


Figure 6.8: Four successive simplifications of Figure 6.7, by removing extraneous non-data elements.

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Example
Proper Scaling and Labeling

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

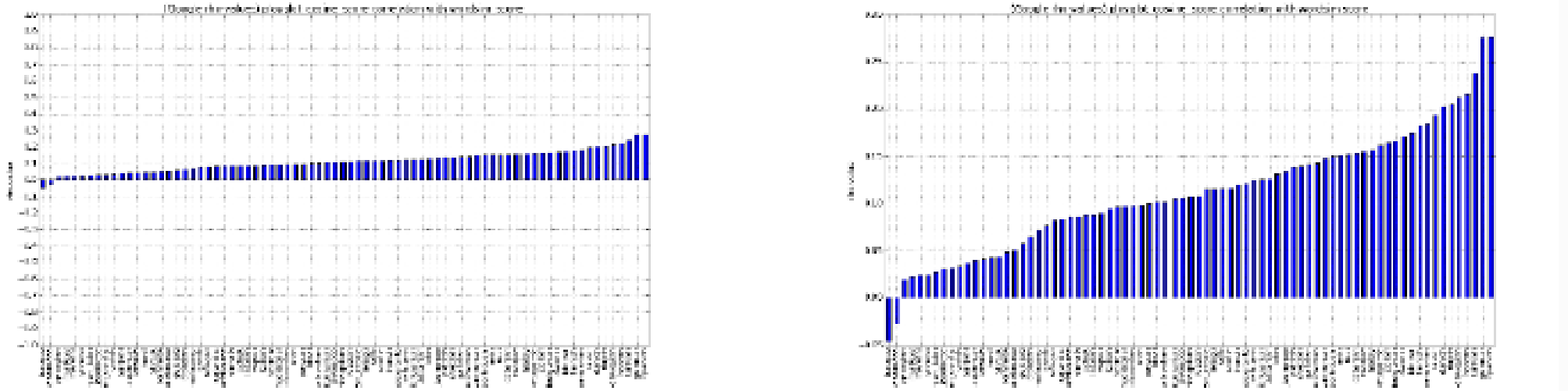
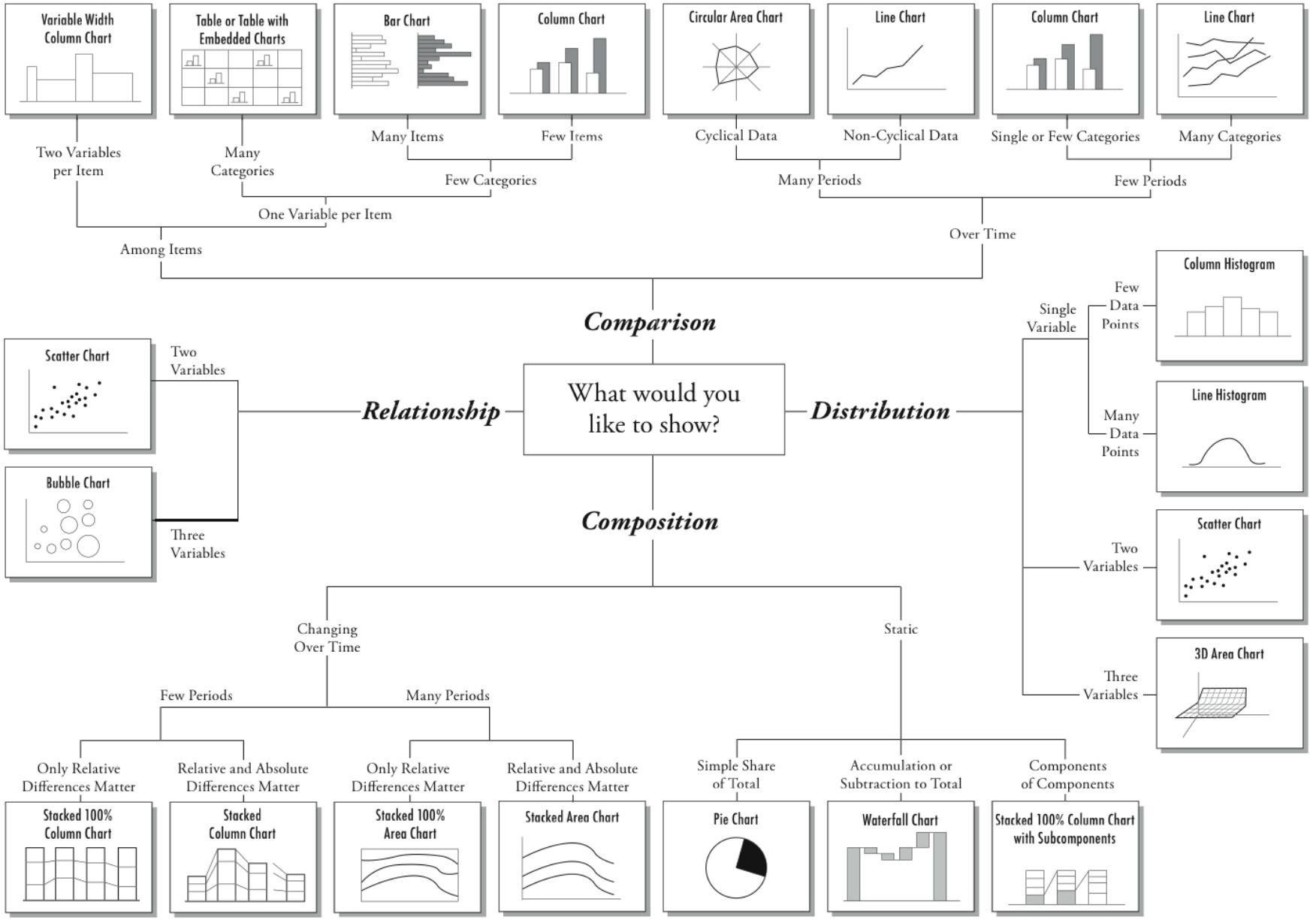


Figure 6.9: Scaling over the maximum possible range (left) is silly when all it shows is white space. Better scaling permits more meaningful comparisons (right).

Charts Types

Chart Suggestions—A Thought-Starter



Great Visualization Examples

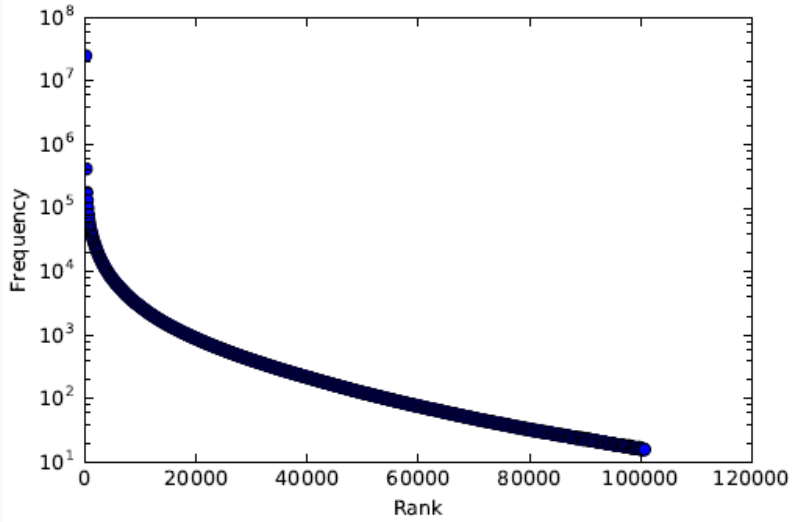
Snow's Cholera Map

Figure 6.28: Cholera deaths center around a pump on Broad street, revealing the source of the epidemic.

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Reading Graphs

Obscure Distribution



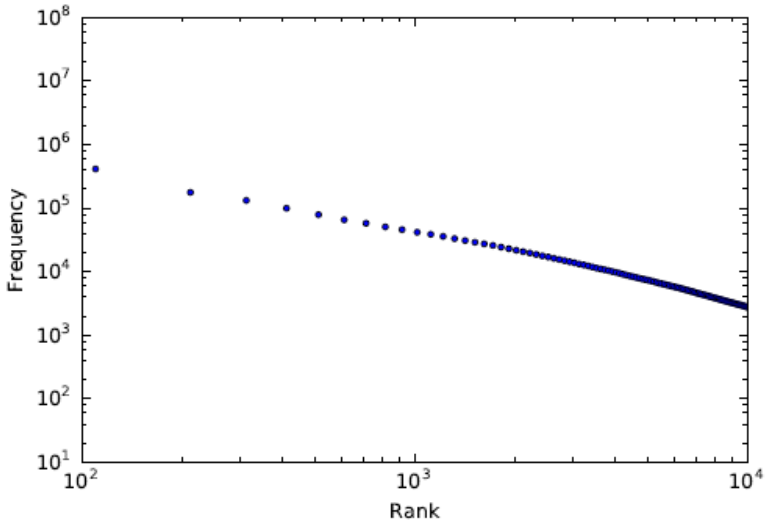


Figure 6.31: Frequency of 10,000 English words. Plotting the word frequencies on a log scale (left), or even better on a log-log scale reveals that it is power law (right).

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

Overinterpreting Variance

From: The Data Science Design MANUAL. Steven S. Skiena, 2017.[Chapter 6]

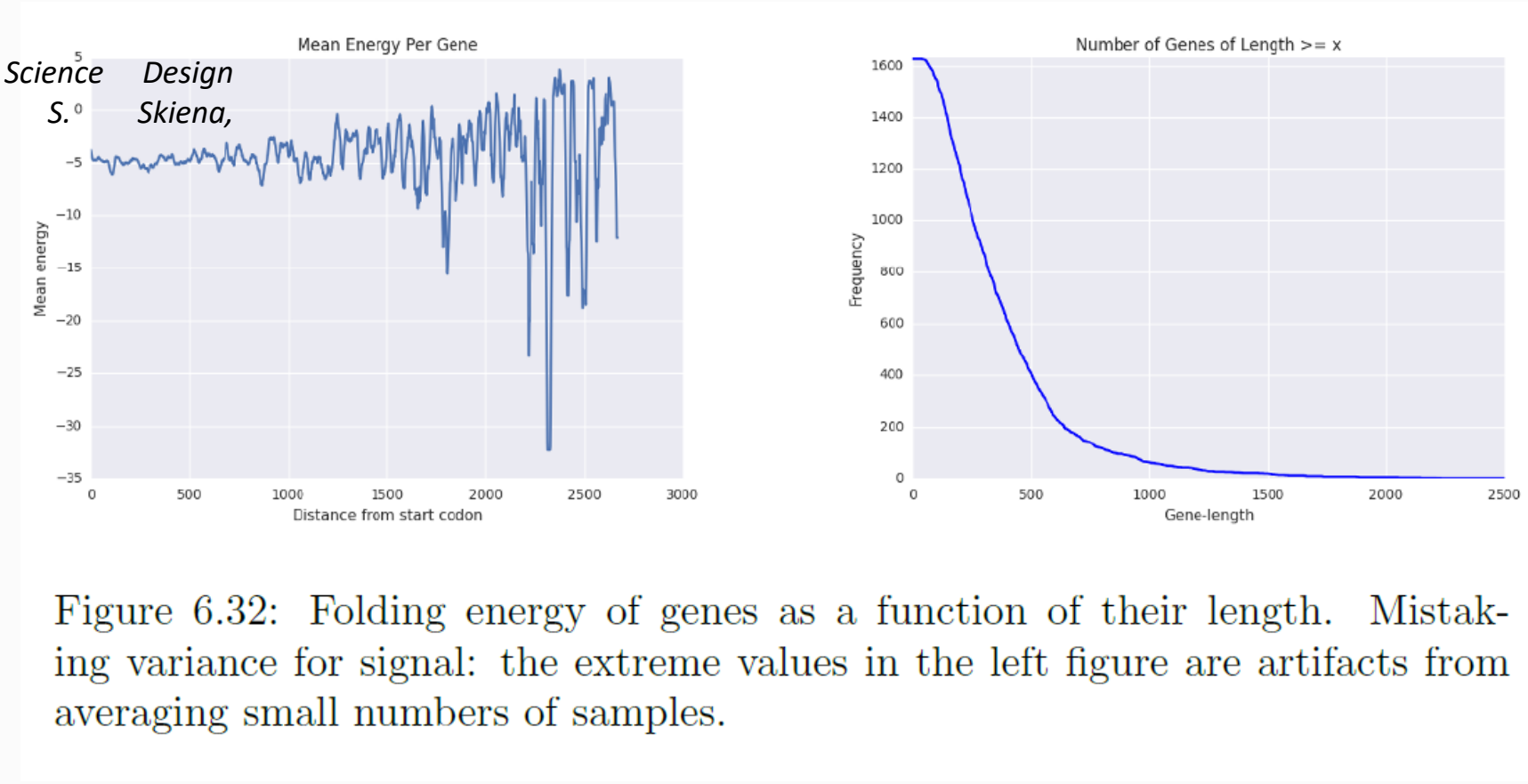


Figure 6.32: Folding energy of genes as a function of their length. Mistaking variance for signal: the extreme values in the left figure are artifacts from averaging small numbers of samples.

Session Mind Map

References

- *The Data Science Design MANUAL. Steven S. Skiena, ISBN: 978-3-319-55444-0 ©2017.[Chapter 6]*