JAN BISCHOFF
DATA SCIENCE & VISUALIZATION

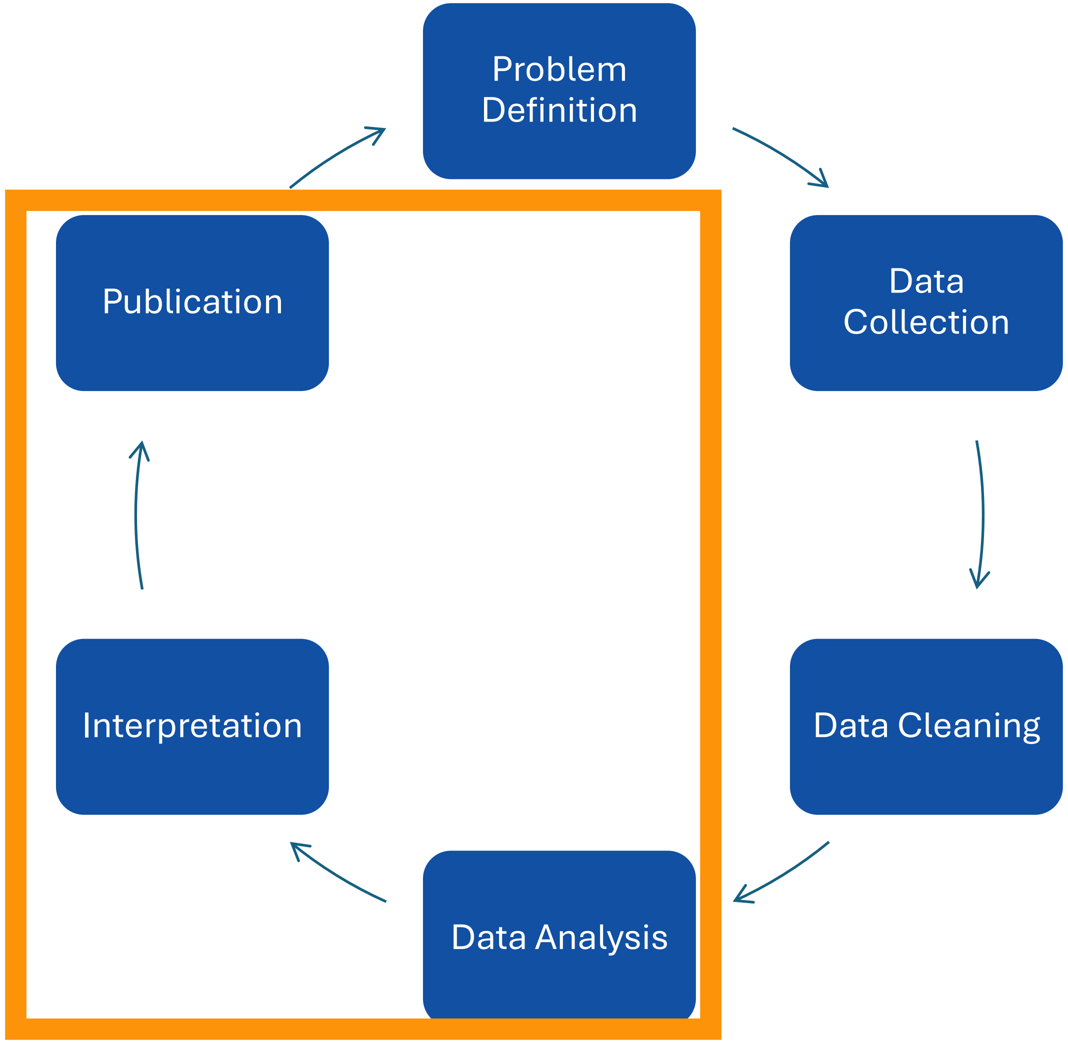
Advanced Data Visualization with Python

Organisation of the course

- Alternation between theory sessions and hands-on programming sessions in Google Colab
- Q&A after each programming session
- Coffee break in the afternoon

The Research Lifecycle

Focus of this course



Purpose of Data Visualization

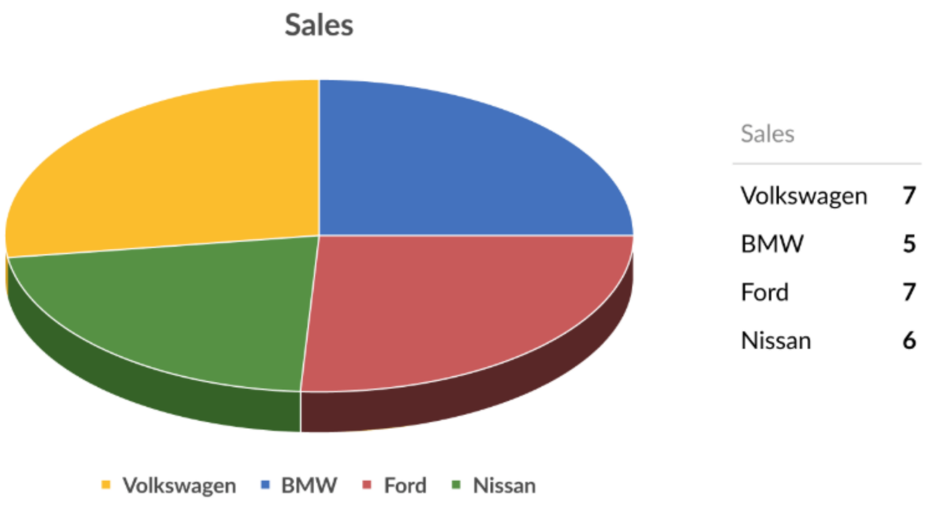
- **Exploration**: Finding patterns in high-dimensional data
- **Communication**: Reducing "cognitive load" for the reader
- **Confirmation:** To verify statistical assumptions (normality, linearity)

Course content

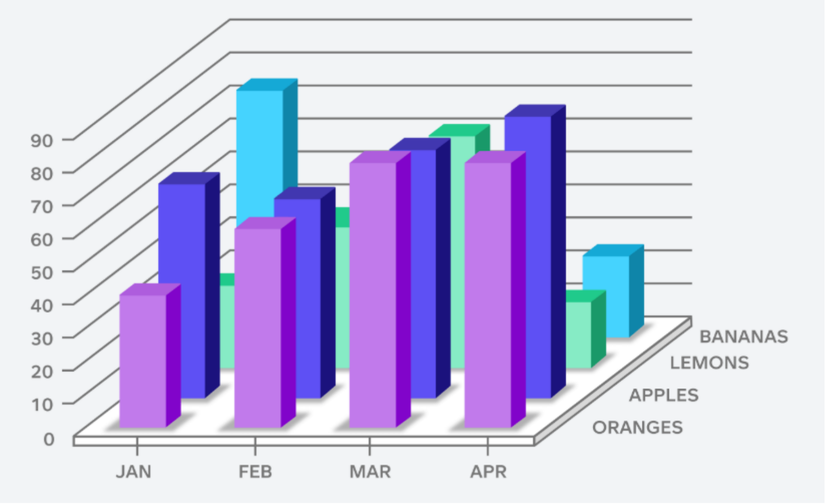
1. Common Pitfalls
2. Fundamentals of Data Visualizations
3. Multivariate Exploration
4. Distributions, Uncertainty & Modeling

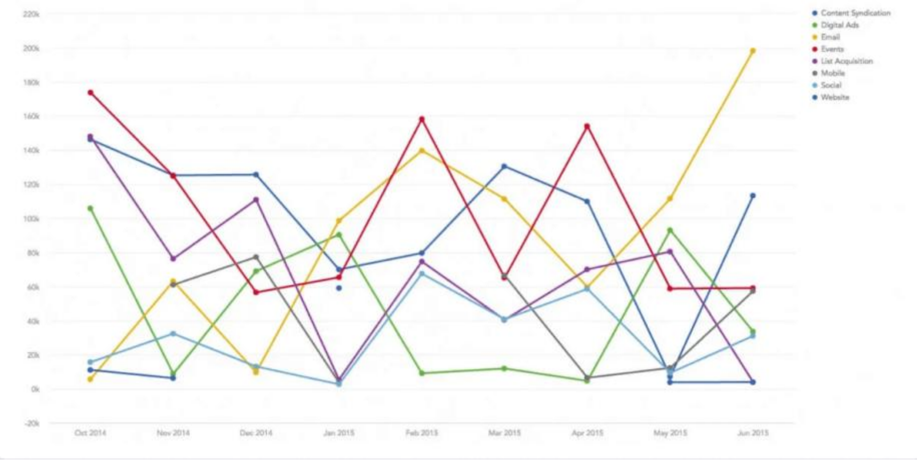
1. Common Pitfalls

Wrong Chart Type

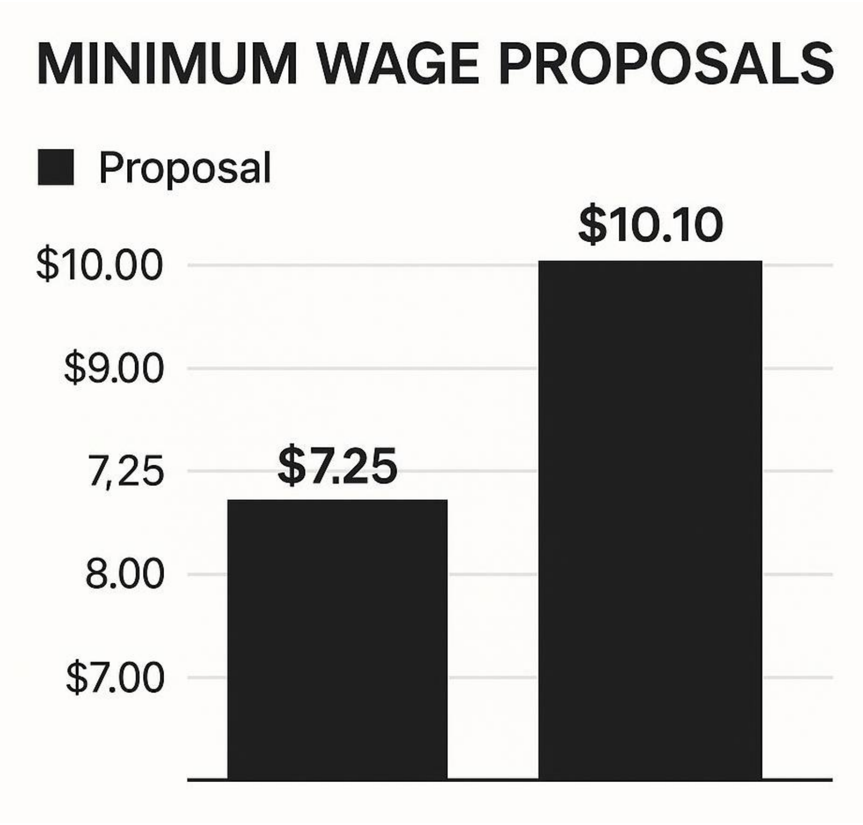


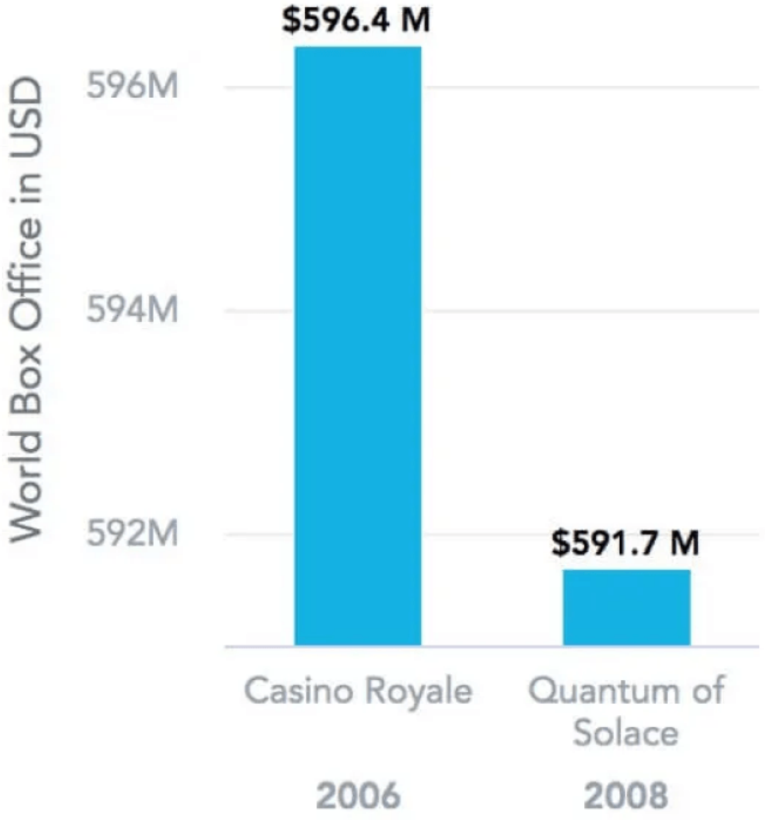
Sales	
Volkswagen	7
BMW	5
Ford	7
Nissan	6



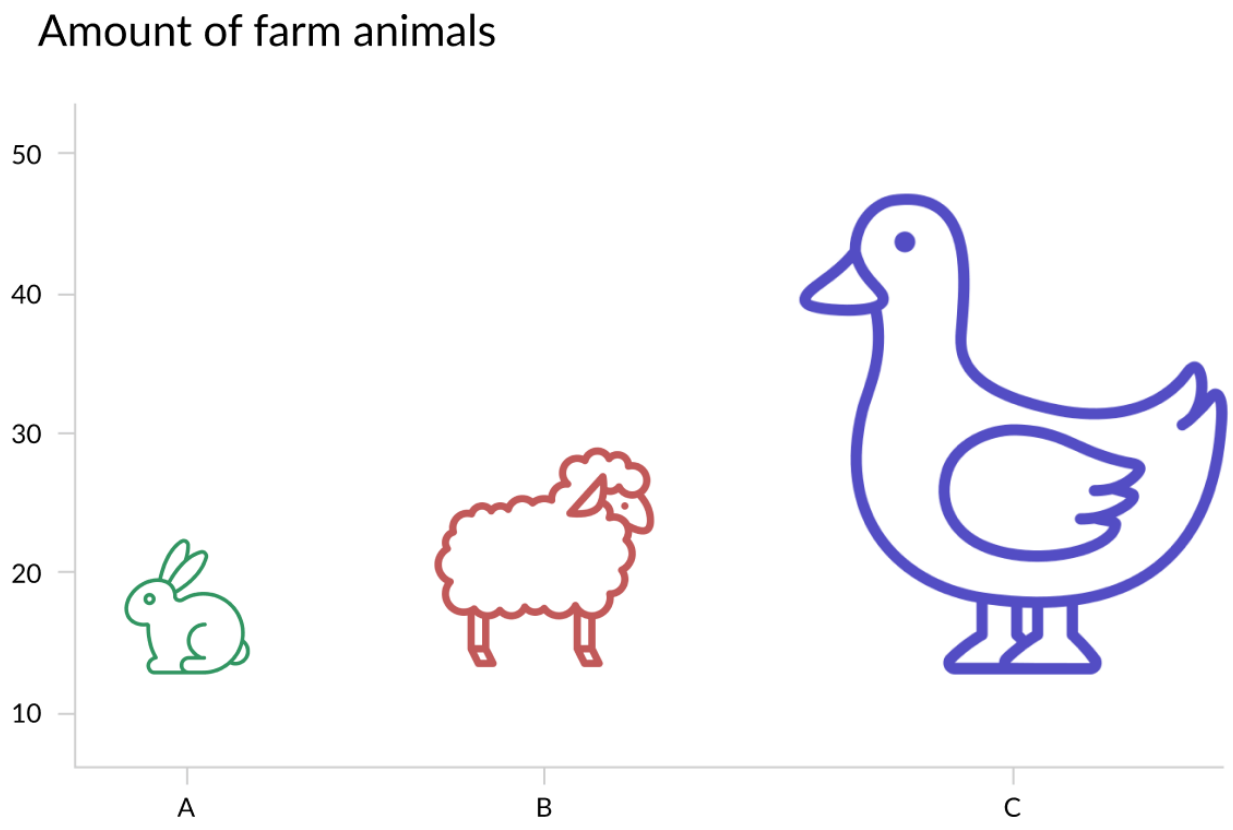


Truncated Y-Axis

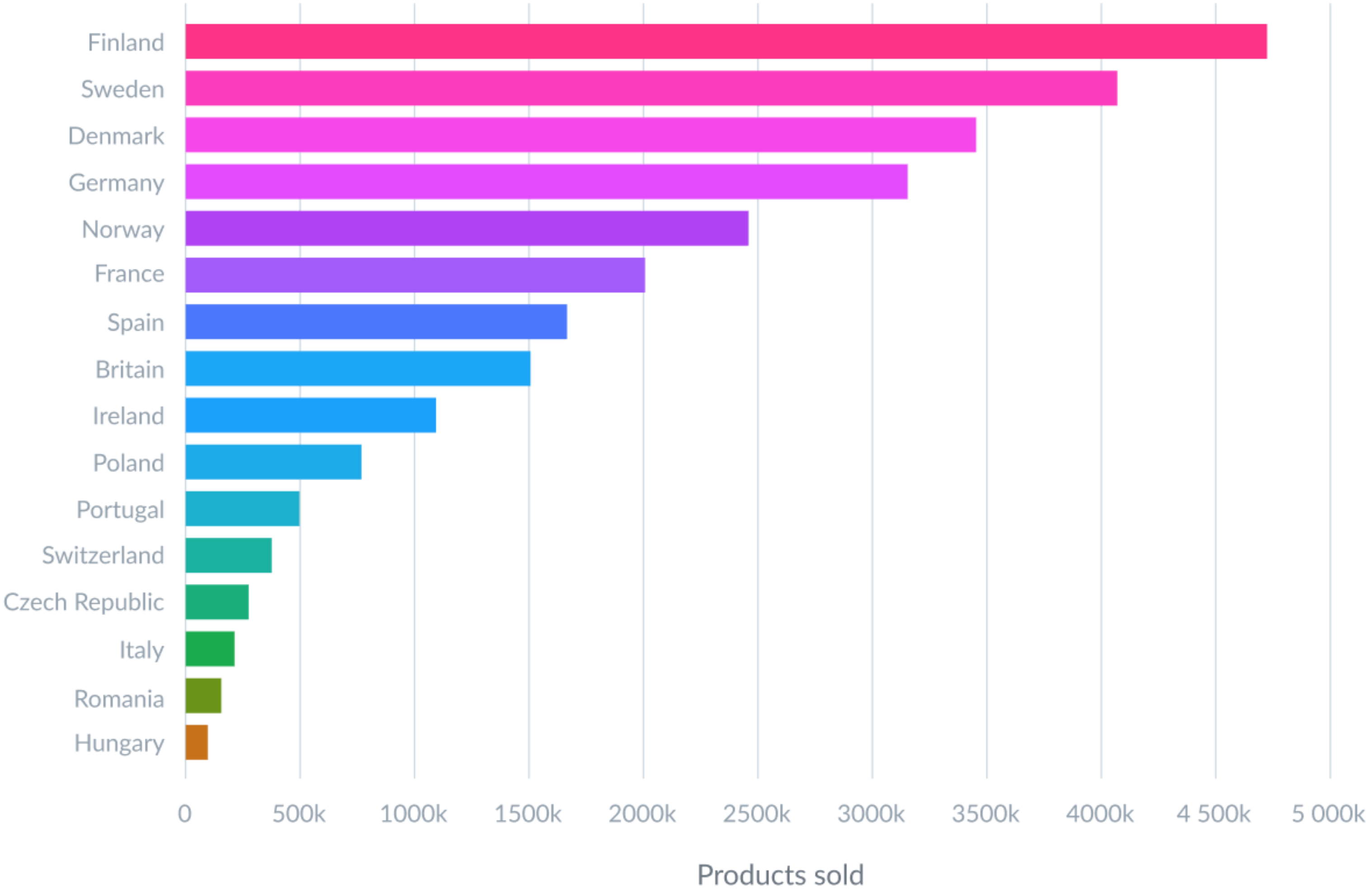




Misleading Scale



Misuse of Colors

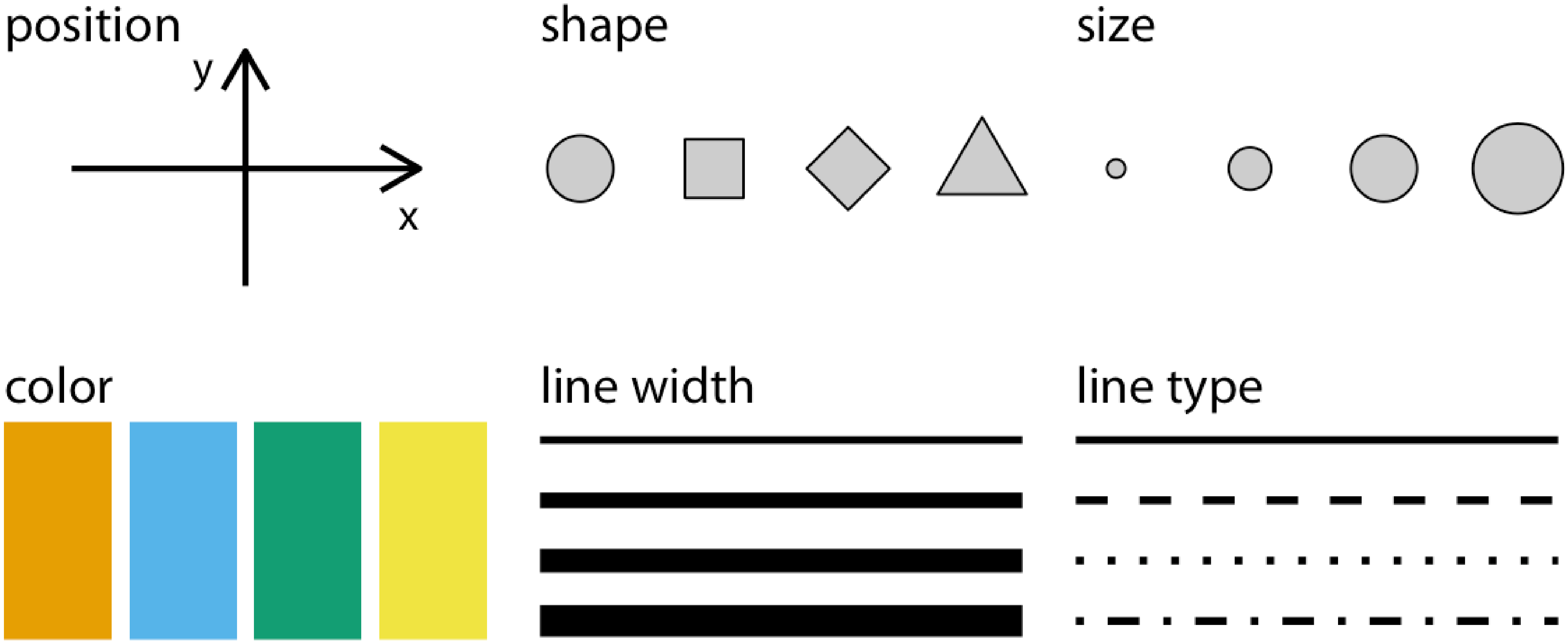


2. Fundamentals & Best Practices

Data Types

Type of Variable	Example	Scale
Quantitative / numerical continuous	1.4, -3.5, 5.2 x 10^2	Continuous
Quantitative / numerical discrete	1,2,3,4,5	Discrete
Qualitative / categorical unordered (nominal)	Math, Physics, Economics	Discrete
Qualitative / categorical ordered (ordinal)	Good, better, best	Discrete

Aesthetics



The Data-to-Ink Ratio (Tufte)

- Definition: Data-Ink Ratio = (Ink used for data) / (Total ink used on graphic)
- Maximize Data-Ink:
    - Erase background colors, heavy gridlines, 3D effects, borders
    - De-emphasize non-data elements (axes, ticks)
    - Emphasize the data points themselves

Accessible & Accurate Color Palettes

- Sequential: For magnitude (Low to High) use viridis, plasma
- Diverging: For deviations from a midpoint (Negative to Zero to Positive) use bwr, coolwarm
- Qualitative: For distinct categories use colorblind friendly sets like tab10 or Set1
- Accessibility: 1 in 12 men are colorblind. Avoid Red/Green contrasts

Titles, Captions, and Labels

- Titles: Should be descriptive ("Figure 1: Effect of Drug A on Growth") not generic ("Scatter Plot")
- Axis Labels: Must always include Units (e.g., "Time (s)“)
- Captions: The figure should be understandable without reading the main text
- Font Size: Ensure axis text are readable

Export Formats

- Raster (PNG, JPG, TIFF)
  - Made of pixels
  - Gets blurry when zoomed
- Vector (PDF, SVG, EPS)
  - Made of mathematical paths
  - Infinite zoom
  - Use for: Line charts, bar charts, and final manuscript figures
- Resolution: Always set dpi=300 or higher for raster exports

Practice time

Let’s get hands-on and practice String methods!

Open your Google Colab exercise notebook.

Solve exercises 1 – 3.

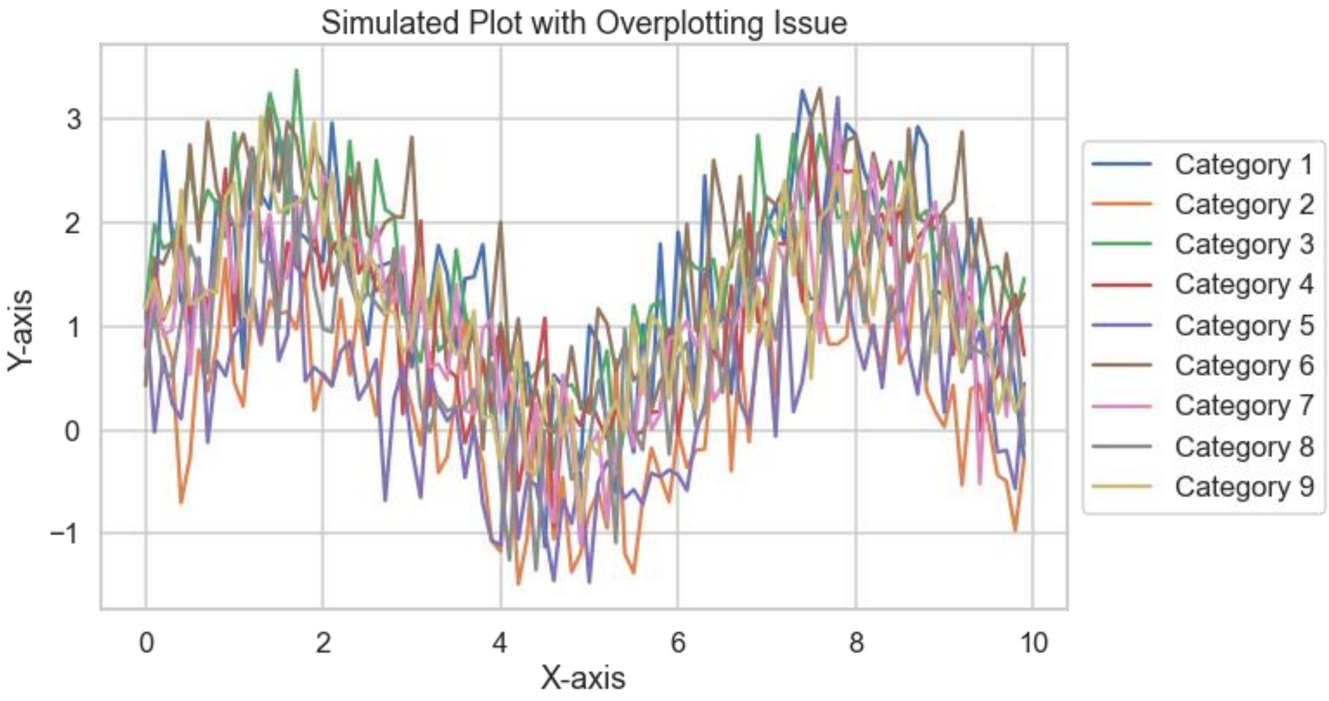
3. Multivariate Exploration

Handling High-Dimensionality

- We rarely study just two variables
- Encoding Semantics:
  - Variable A: X-Axis (Position)
  - Variable B: Y-Axis (Position)
  - Variable C: Size (Quantitative)
  - Variable D: Hue (Categorical)
- Caution: Too many encodings create "visual soup"

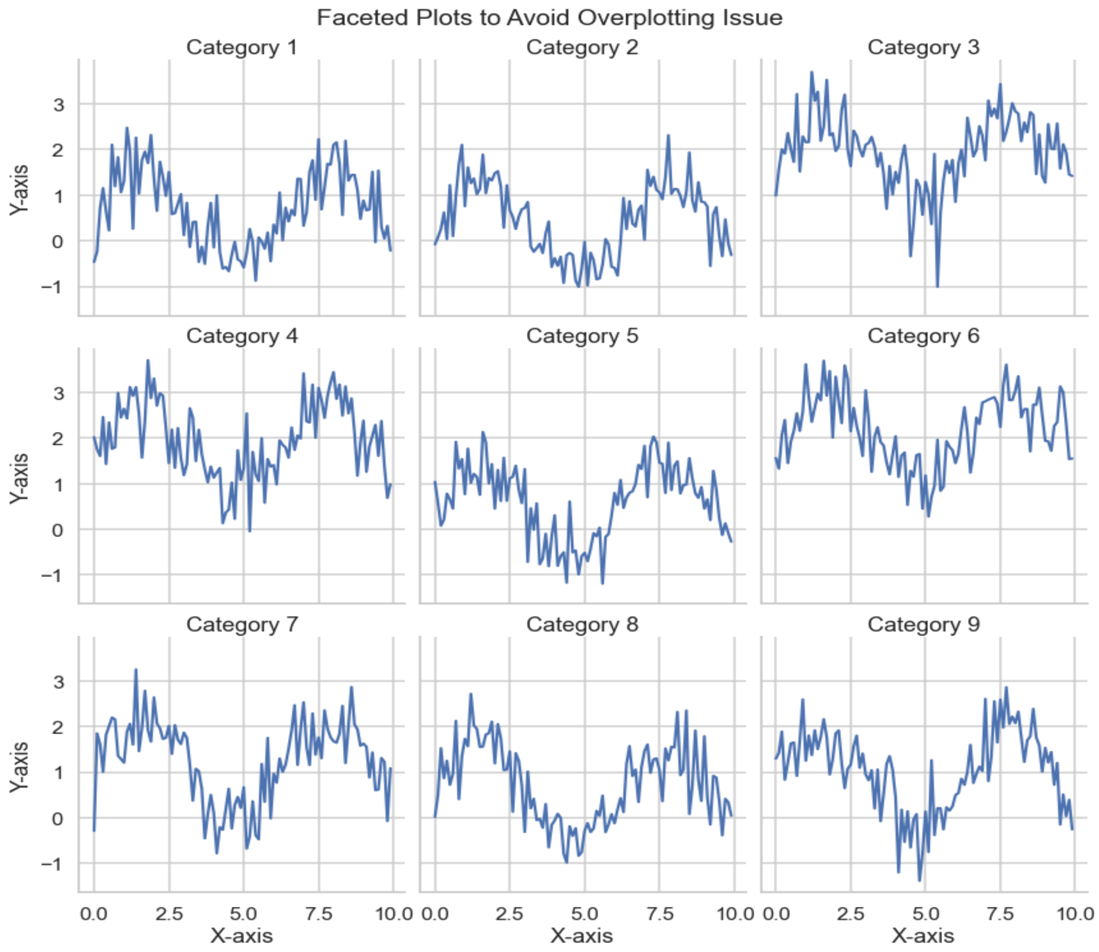
The Overplotting Problem

The Issue: Plotting multiple groups on a single set of axes makes tracing individual lines impossible



Faceting (Small Multiples)

- Splitting a chart into a grid of subplots based on a categorical variable
  - Reduces cognitive load
  - Allows direct comparison (if axes are shared/fixed)



Practice time

Let’s get hands-on and practice String methods!

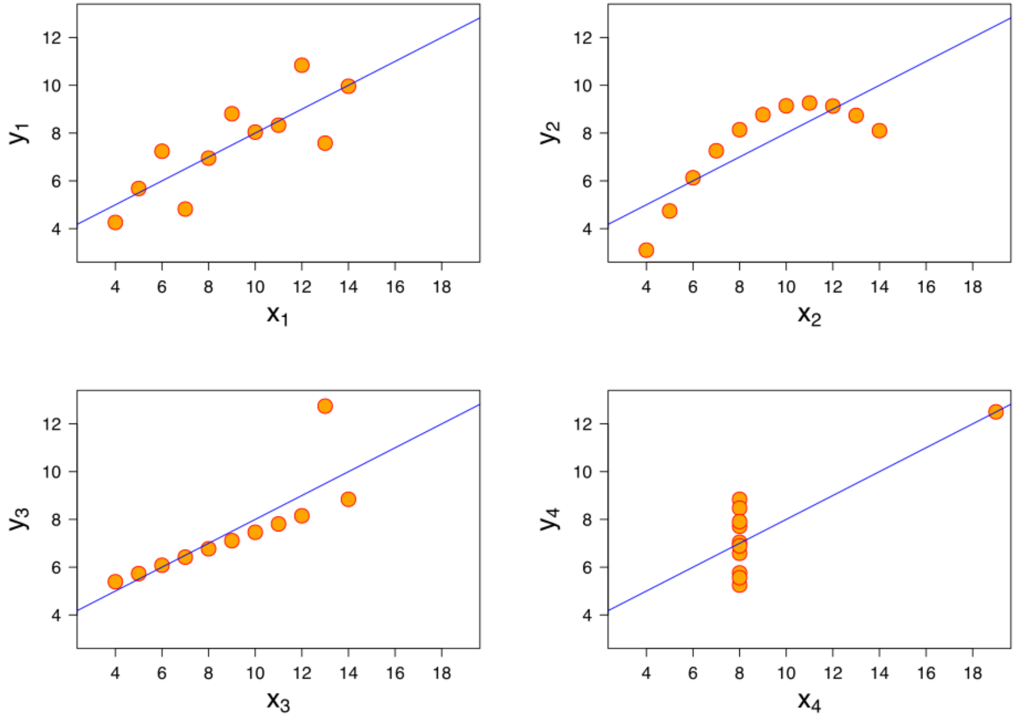
Open your Google Colab exercise notebook.

Solve exercises 4 – 6.

4. Distributions, Uncertainty & Modeling

The Limitation of Summary Stats

Property	Value	Accuracy
Mean of x	9	exact
Variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Variance of y	4.125	±0.003
Correlation	0.816	to 3 decimal places
Linear regression	$y = 3.00 + 0.500x$	to 2 and 3 decimal places
R^2	0.67	to 2 decimal places



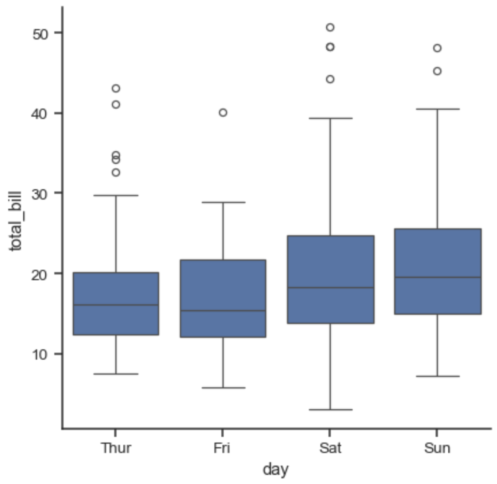
The Anatomy of Uncertainty

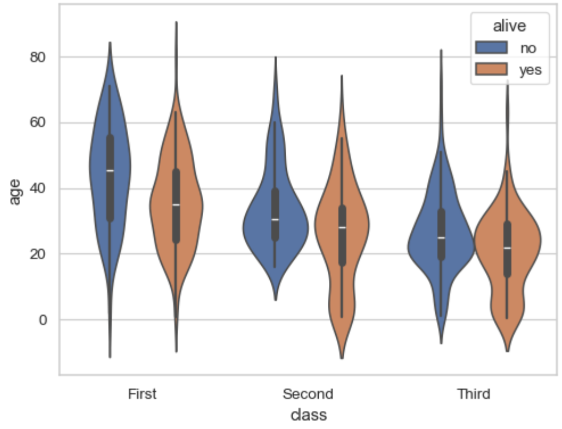
- Standard Deviation (SD): "How spread out is the data?" (Descriptive)
- Standard Error (SE): "How precise is our estimate of the mean?" (Inferential)
- Confidence Interval (CI): "If we repeated this experiment 100 times...“

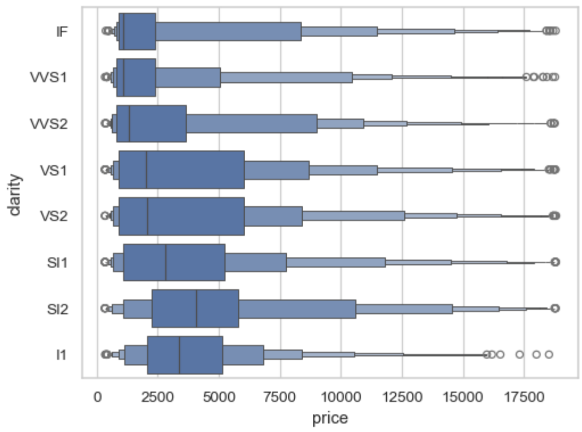
Golden Rule: Always explicitly state what your error bars represent in the caption!

Distributional Visualizations

- Boxplot
  - Shows median and quartiles
  - Good for summary, but hides multimodality
- Violin Plot
  - Boxplot + KDE (Kernel Density Estimate)
  - Shows the "shape" of data
- Boxen Plot
  - Enhanced boxplot for large N
  - Shows more quantiles (tails)

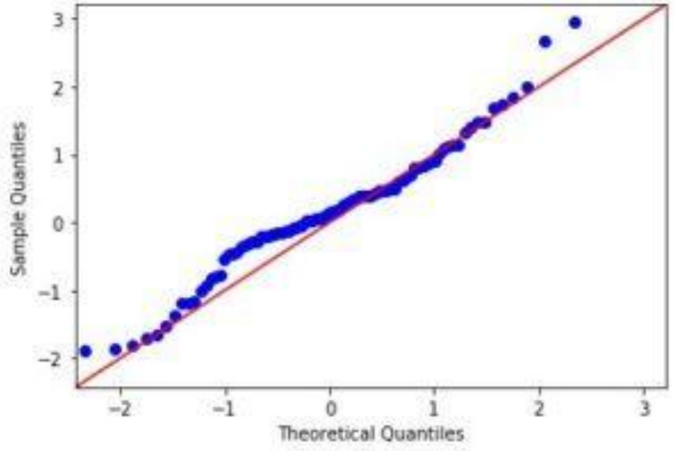


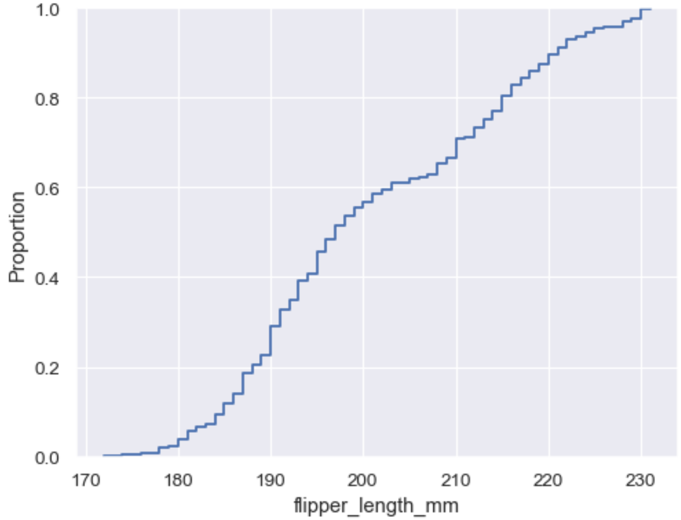




Validating Assumptions

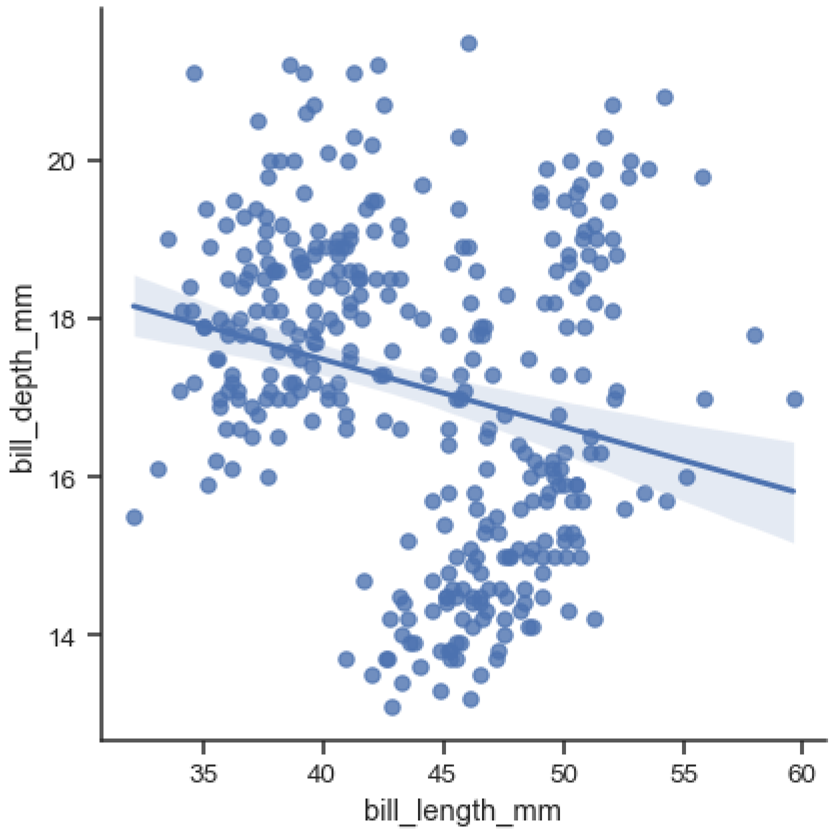
- Q-Q Plot (Quantile-Quantile):
  - Compares data to a theoretical distribution (usually Normal)
  - If points fall on the line: Data is Normal
- ECDF (Empirical Cumulative Distribution Function):
  - Shows the proportion of data less than or equal to x
  - Great for comparing distributions without "binning bias" (unlike histograms)





Visualizing Models

- The Line: The best fit model (e.g., linear regression)
- The Band: The 95% Confidence Interval (usually calculated via bootstrapping)
- Best Practice: Overlay the raw data points behind the model fit
- Advice: Use alpha=0.3 for points to highlight the density of the trend



Practice time

Let’s get hands-on and practice regular expressions!

Open your Google Colab exercise notebook.

Solve exercises 7 – 10.

Thank you very much for participating!

Sources

- https://clauswilke.com/dataviz/
- https://seaborn.pydata.org/index.html
- https://datanizant.com/examples-of-bad-data-visualization/
- https://www.gooddata.com/blog/bad-data-visualization-examples-that-you-can-learn-from/