Data Visualization & Exploratory Data Analysis

Hrachya Asatryan

Why This Lecture Exists

- Every data science pipeline starts with data
- Models only optimize what you give them
- Wrong assumptions propagate silently
- EDA is where most real-world mistakes are caught

The Most Common Data Science Failure

> **“The model trained successfully, but the results don’t make sense.”**

Typical causes:

- misunderstood variables
- hidden data leakage
- invalid comparisons
- aggregation masking structure

This is almost never a modeling issue.

What EDA Is NOT

EDA is **not**:

- running .plot() to see something
- making dashboards
- pressing PCA/t-SNE because they exist
- checking boxes in a notebook

EDA is *reasoning*.

What EDA IS

EDA is:

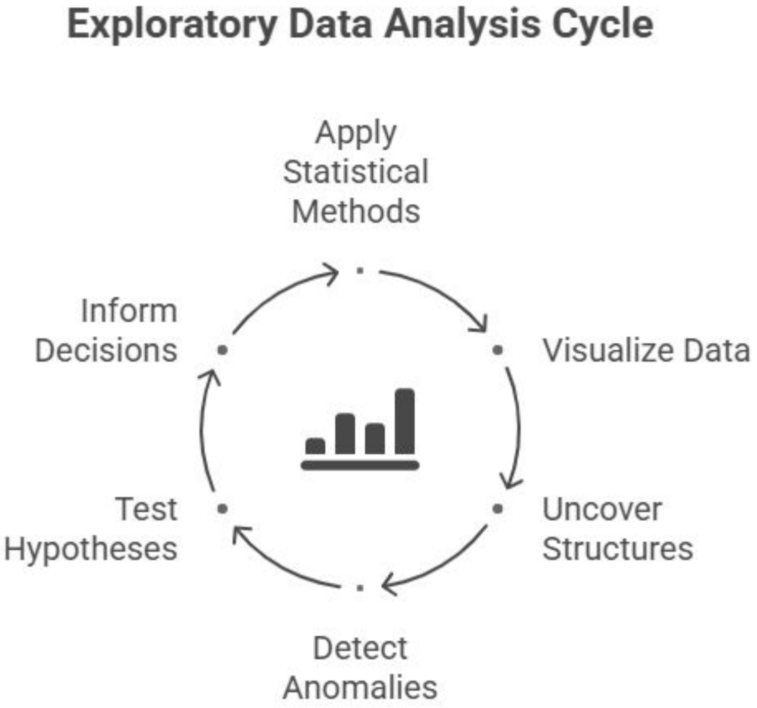
- asking precise questions
- validating assumptions
- building intuition
- discovering surprises

If nothing surprises you, you’re not doing EDA.

The EDA Loop

1. Look at the data
2. Ask a concrete question
3. Visualize or summarize
4. Update your mental model
5. Repeat

EDA is iterative, not linear.



Data Is Not Neutral

Every dataset reflects:

- how it was collected
- who designed the measurement
- what was *not* recorded
- constraints of sensors/surveys

EDA starts *before* code.

Rows and Columns Lie

Tabular data suggests:

- independence between rows
- equal importance of columns
- flat structure

Reality:

- repeated measurements
- hierarchical structure
- dependencies everywhere

What a Column Actually Represents

A column is:

- a measurement
- of a concept
- under assumptions
- with noise

Numbers ≠ truth.

Variable Types (Critical)

- Numerical
    - continuous (temperature)
    - discrete (counts)
- Categorical
    - nominal (country)
    - ordinal (ratings)
- Binary
- Temporal
- Spatial

Variable type determines valid operations.

When Numbers Are Not Numeric

Examples:

- IDs
- ZIP codes
- product codes
- encoded categories

Averaging these is meaningless.

Before Any Plot: Basic Questions

For every column:

- What does it measure?
- What is a valid range?
- Can it be missing?
- Can it be zero?
- Can it be negative?

EDA starts **here**.

Why Univariate Analysis Comes First

Before relationships:

- understand scale
- understand spread
- detect anomalies
- detect encoding issues

You can’t interpret interactions otherwise.

Distribution Is the First Story

Every variable has:

- center (mean/median)
- spread (variance/IQR)
- shape (skewness)
- tails (outliers)

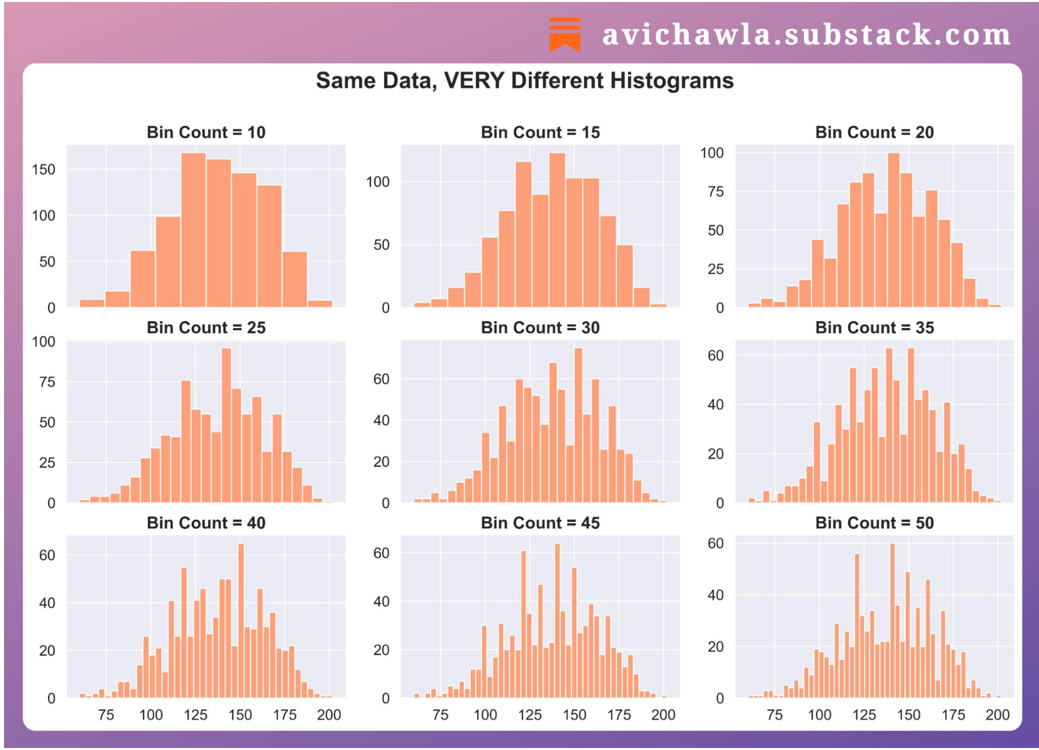
Mean alone is almost useless.

Histograms: Powerful but Opinionated

Histograms depend on:

- bin width
- bin alignment

Same data → different conclusions.



Skewness Is the Norm

Most real-world data is:

- right-skewed
- heavy-tailed
- non-Gaussian

Examples:

- income
- waiting times
- errors

Outliers Are Questions

An outlier might be:

- measurement error
- rare event
- regime change
- most informative sample

We shouldn’t delete without explanation.

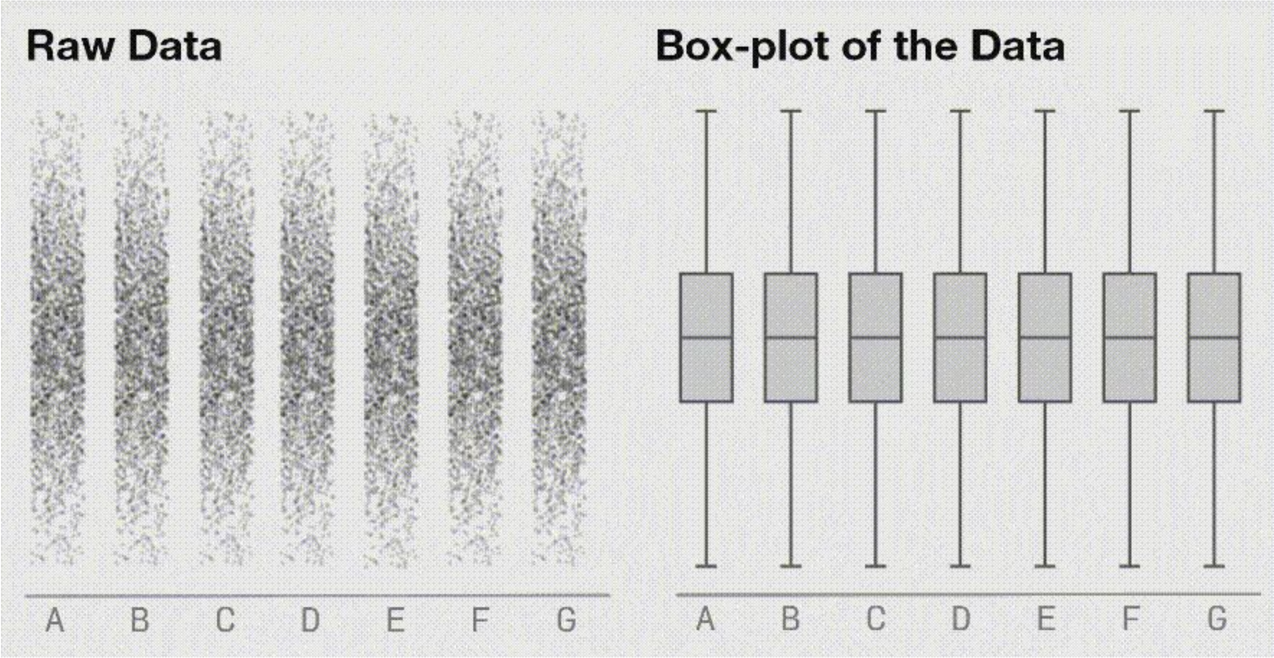
Boxplots: What They Show and Hide

They show:

- median
- spread
- outliers

They hide:

- multimodality
- distribution shape

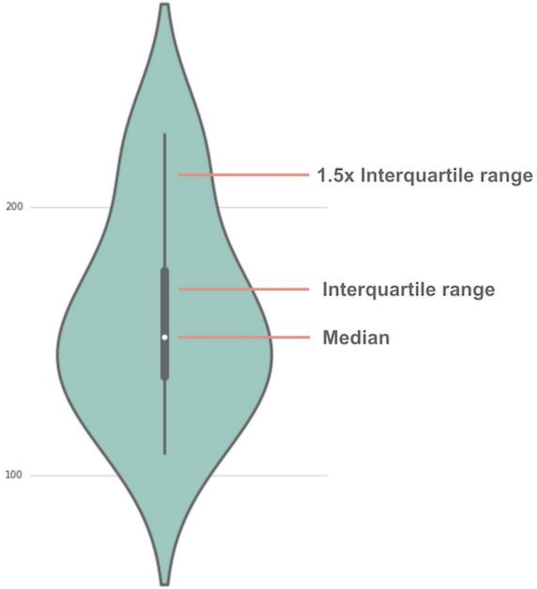


Violin Plots: The Best of Both Worlds

A violin plot combines a box plot with a kernel density estimate (KDE).

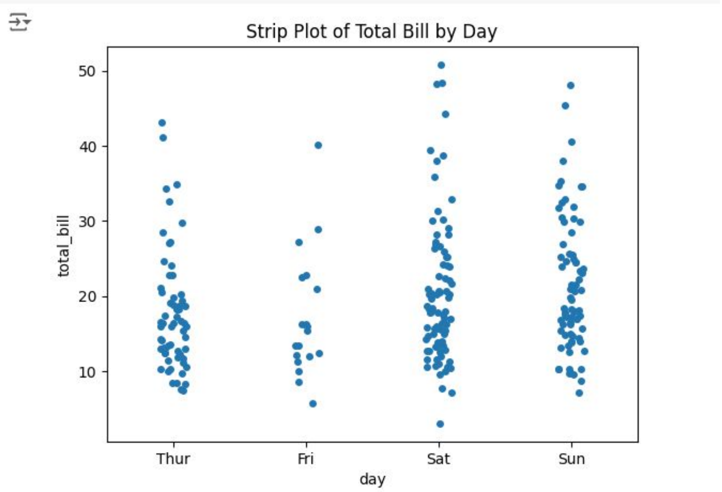
What it shows: It shows the median and IQR (like a boxplot) but *also* the probability density of the data at different values.

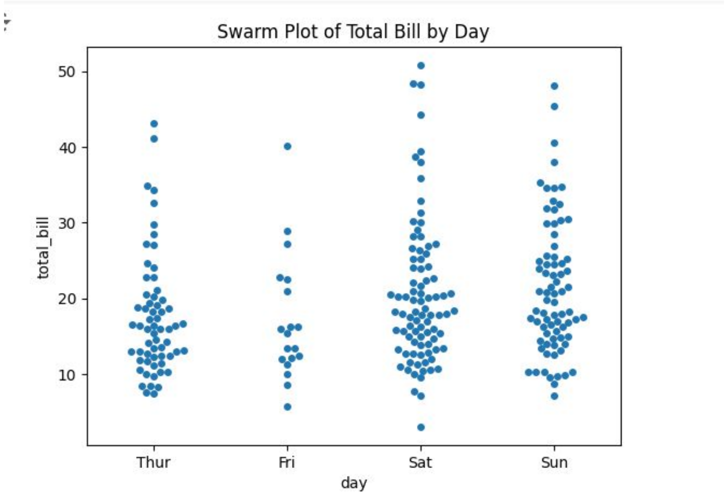
When to use: When you need to compare distributions between groups and check for multimodality (which boxplots hide).



Strip & Swarm Plots

Concept: Plotting every single data point.
Why: Boxplots and Violins are summaries. Sometimes the sample size is small (n<50), and summarizing it is misleading.
Rule of Thumb: "Show the raw data if you can."





From One Variable to Two

Bivariate analysis asks:

- does X relate to Y?
- how strongly?
- linearly or nonlinearly?

Relationships create insight.

Scatter Plots Reveal Structure

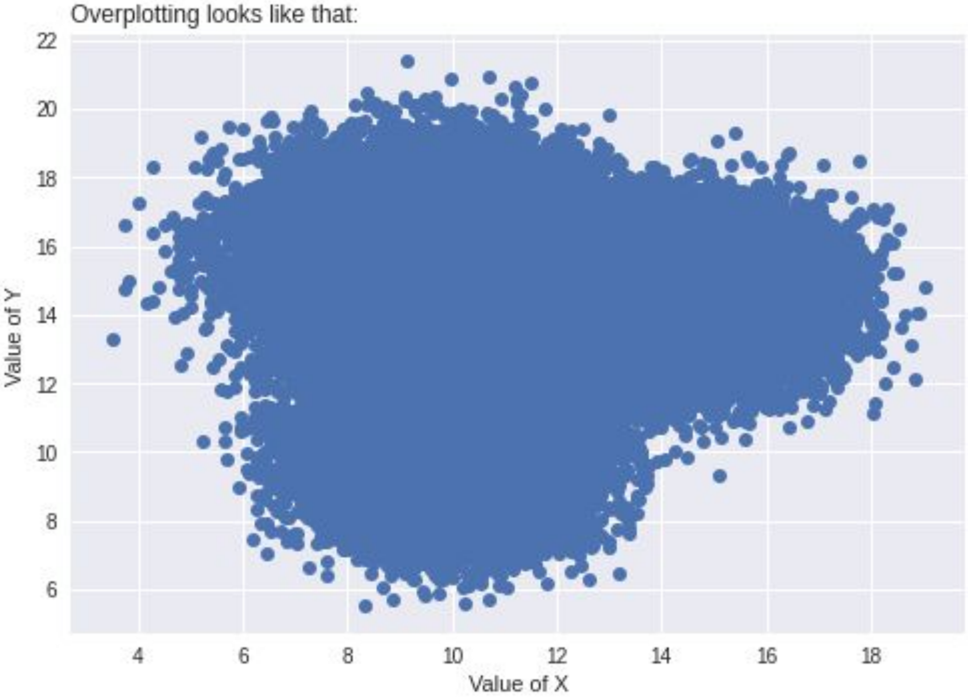
Scatter plots show:

- trends
- clusters
- noise
- nonlinearity

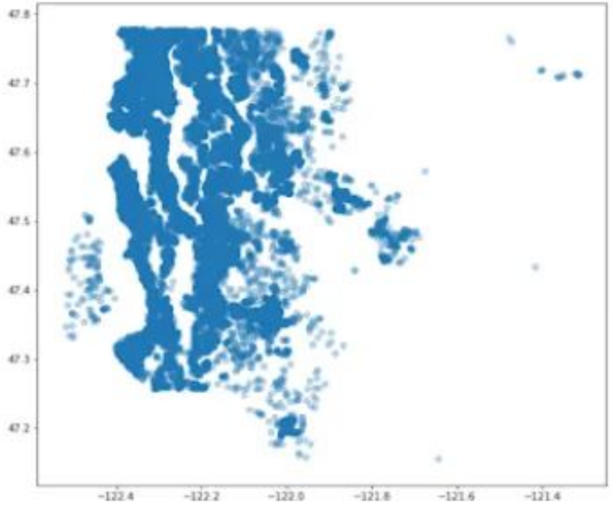
Statistics summarize.
Plots explain.

Overplotting

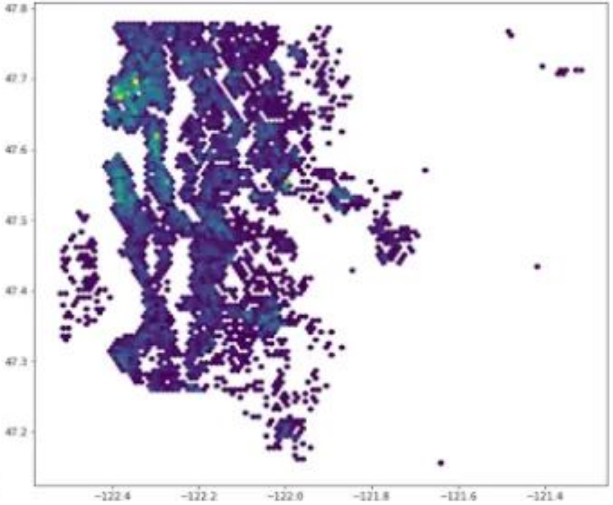
If dealing with too much data, scatterplots can cause **overplotting.**



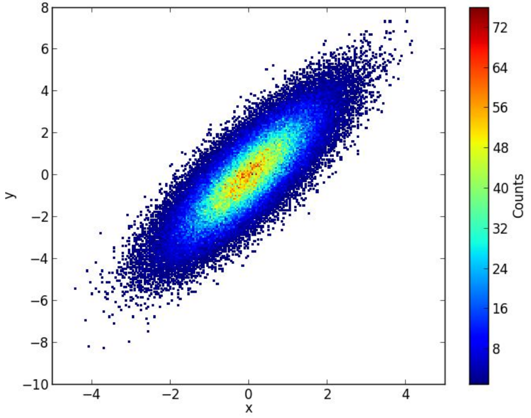
Handling Overplotting (Hexbins & 2D Histograms)



Scatterplot



Hexbin



2D histogram

Correlation

Correlation measures:

- linear association
- direction
- strength

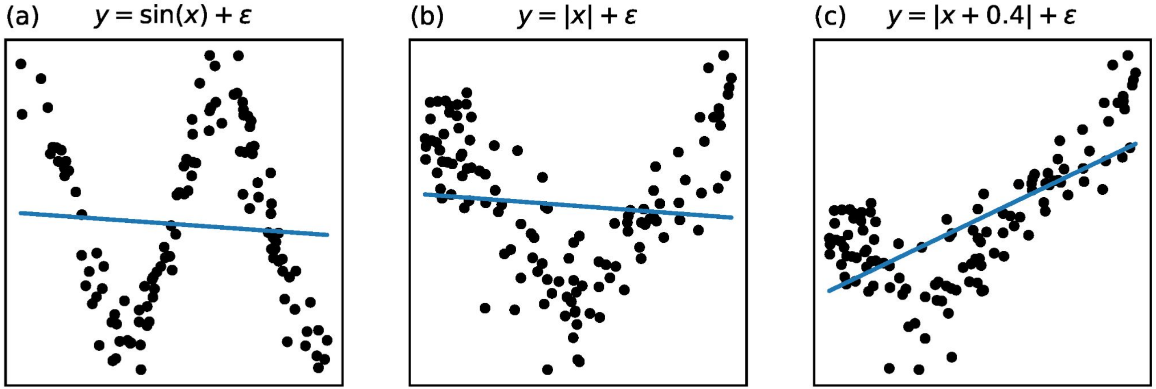
Correlation does NOT imply:

- causation
- nonlinear dependence
- importance

When Correlation Misleads

Correlation fails when:

- relationship is nonlinear
- groups behave differently
- scale dominates
- outliers dominate



Original: −0.07
Transformed: 0.96
MI: 0.94

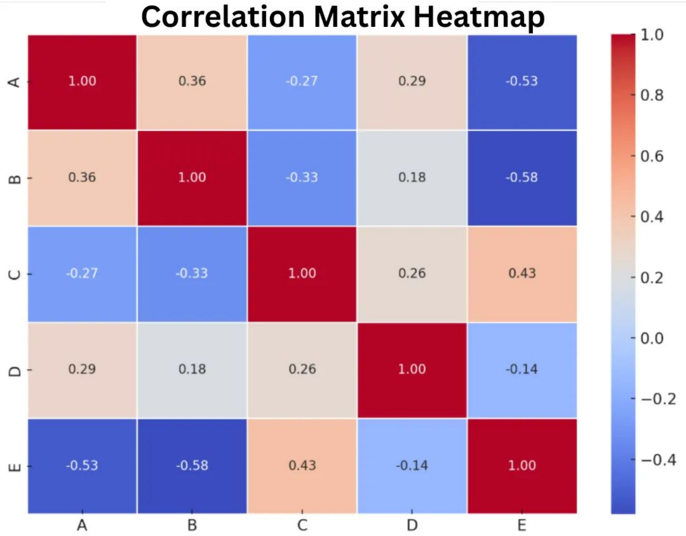
Original: −0.10
Transformed: 0.85
MI: 0.85

Original: 0.70
Transformed: 0.89
MI: 0.89

The Correlation Heatmap

Concept: A grid showing the correlation coefficient between every pair of variables, colored by intensity (Red=Positive, Blue=Negative).

Why: It allows you to spot "clusters" of correlated features instantly.



Correlation Matrix Heatmap

	A	B	C	D	E
A	1.00	0.36	-0.27	0.29	-0.53
B	0.36	1.00	-0.33	0.18	-0.58
C	-0.27	-0.33	1.00	0.26	0.43
D	0.29	0.18	0.26	1.00	-0.14
E	-0.53	-0.58	0.43	-0.14	1.00

Grouped Relationships

Relationships can change across groups:

- Simpson’s paradox
- Confounding variables

Always check stratified plots.

The Villain — What is a Confounding Variable?

Concept: A confounding variable is a "lurking" third variable that influences both the input and the output, making it look like there is a relationship when there isn't (or hiding a real one).

The "Ice Cream & Shark Attacks" Analogy:

- **The Data:** If you plot *Ice Cream Sales* vs. *Shark Attacks*, you will find a strong positive correlation.
- **The Wrong Conclusion:** "Buying ice cream causes shark attacks."
- **The Truth (The Confounder): Temperature**.
    - When it’s hot, people buy ice cream.
    - When it’s hot, people swim in the ocean (more shark attacks).
    - Temperature causes *both*. Ice cream has nothing to do with sharks.

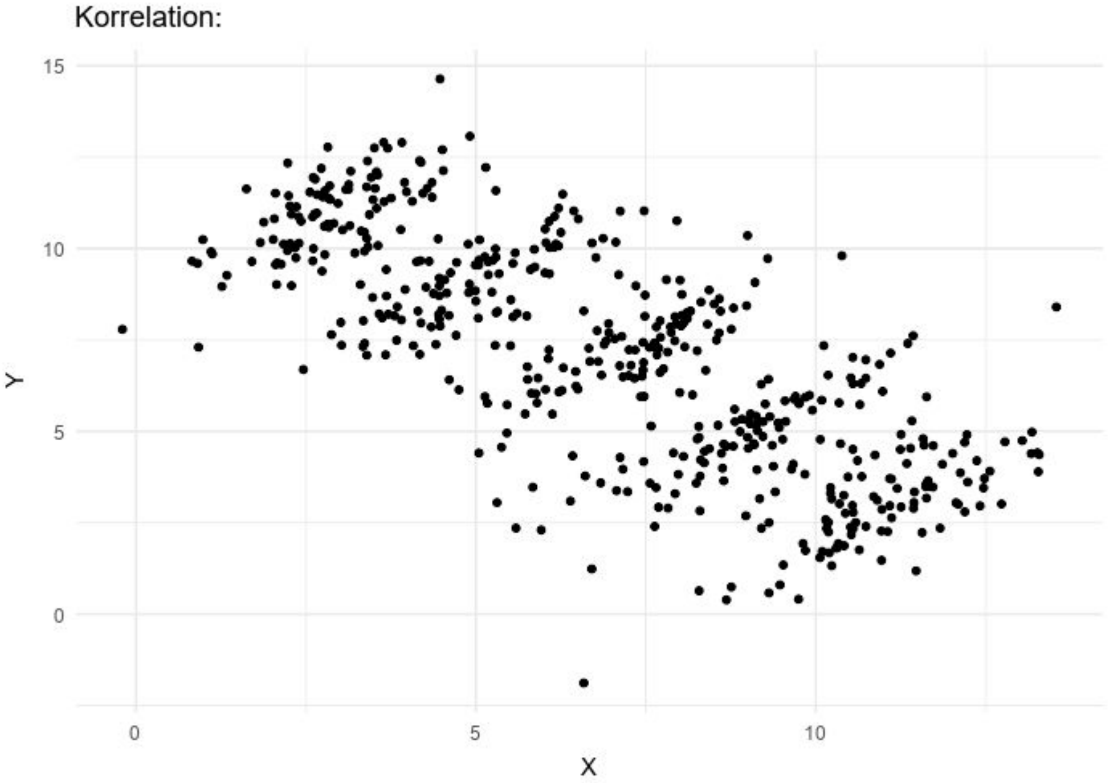
The Mystery — What is Simpson’s Paradox?

Concept: Simpson's Paradox is a specific type of statistical error where a trend appears in different groups of data but **disappears or reverses** when these groups are combined.
The "Kidney Stone" Example (The Reversal): Imagine comparing two treatments for kidney stones, Treatment A and Treatment B.

- **Group 1 (Small Stones):** Treatment A is better.
- **Group 2 (Large Stones):** Treatment A is better.
- **Combined (Total):** Treatment **B** looks better.

How is this possible? Treatment A was mostly used on "Large Stone" cases (which are harder to cure), dragging its average down. Treatment B was mostly used on easy cases.

Simpson’s Paradox: Visualization



The Solution — What are Stratified Plots?

Concept: "Stratification" just means **slicing**. A stratified plot is where you split your data into subgroups (strata) based on a categorical variable (like Age, Gender, or "Stone Size") rather than lumping everyone together.

The Rule: "Always color by category."

- **Aggregate Plot:** A scatter plot of X vs Y (All black dots). Shows a positive trend.
- **Stratified Plot:** The same scatter plot, but dots are colored by Group Z. Now you see three distinct negative trends.

The Takeaway:

- If you *don't* stratify, you see the "Total" (which might be a lie).
- If you *do* stratify, you see the "Subgroups" (the truth).
- **EDA Rule:** Never trust a summary statistic or a plot until you have checked it against major categories (Sex, Region, Time, etc.).

Line Plots Require Order

Valid x-axis:

- time
- ordered sequence

Invalid:

- categories
- shuffled data

Line plots imply continuity.

Bar Charts Are Aggregation

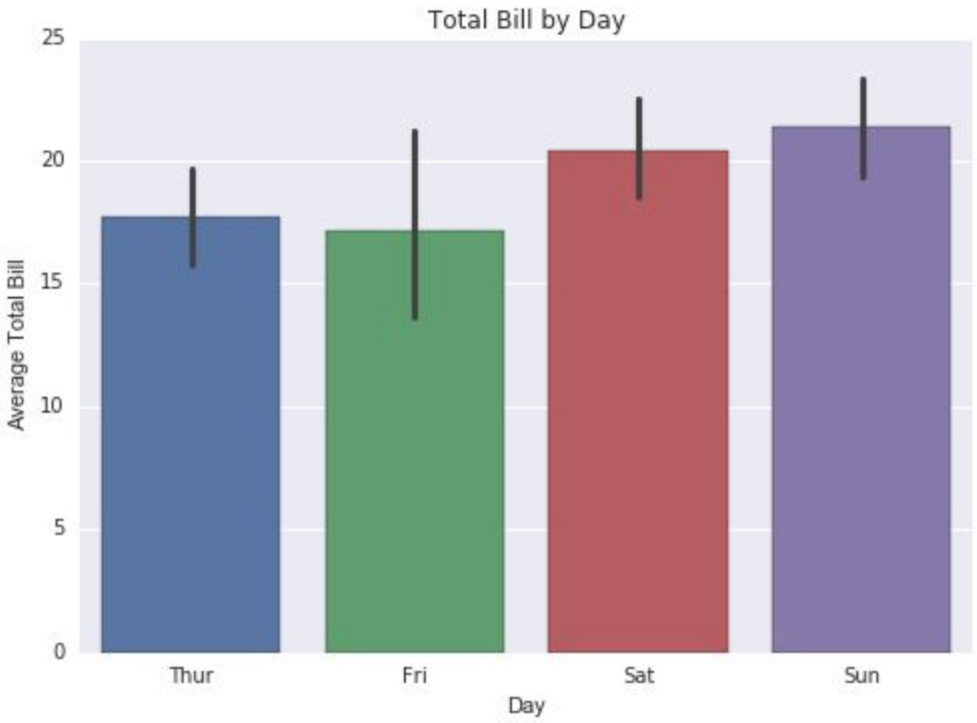
Bar charts show:

- means
- counts
- summaries

They hide:

- variance
- distribution
- outliers

Aggregation is a decision.



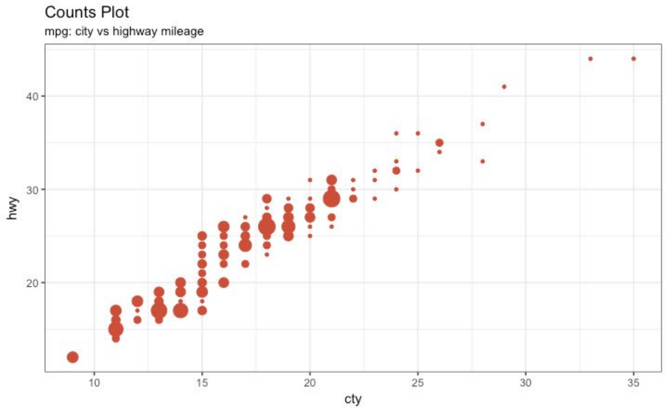
Count Plots (Frequency)

Concept: The categorical equivalent of a histogram.

Question: "How many rows belong to each category?"

Example: Number of passengers per Class (1st, 2nd, 3rd) on the Titanic.

Tip: Always sort the bars (highest to lowest) unless there is a natural ordering (like months).



Multivariate Reality

Real datasets:

- have many features
- are redundant
- hide structure in combinations

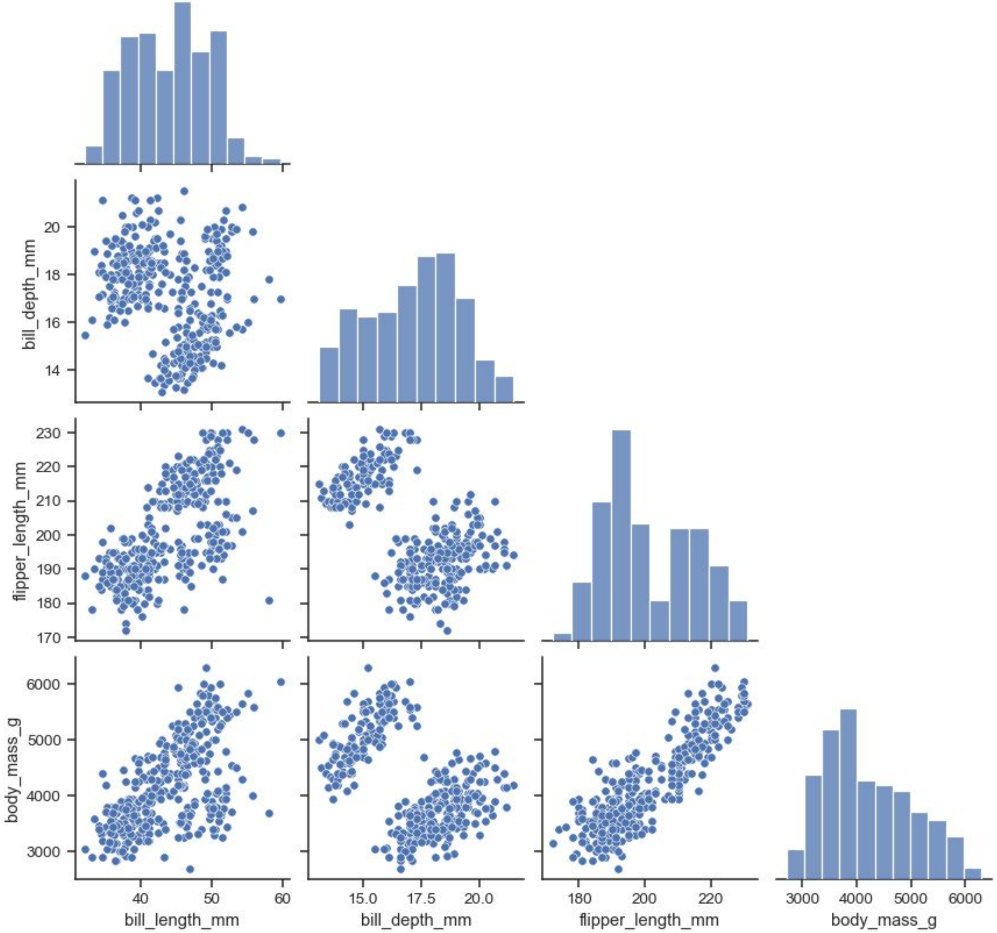
EDA helps reduce complexity mentally.

Pairwise Exploration

Pair plots help:

- detect redundancy
- spot relationships
- identify useless features

They don’t replace thinking.



Scale Changes Interpretation

Different scales affect:

- distances
- dominance
- visualization

Scaling is part of understanding, not just preprocessing.

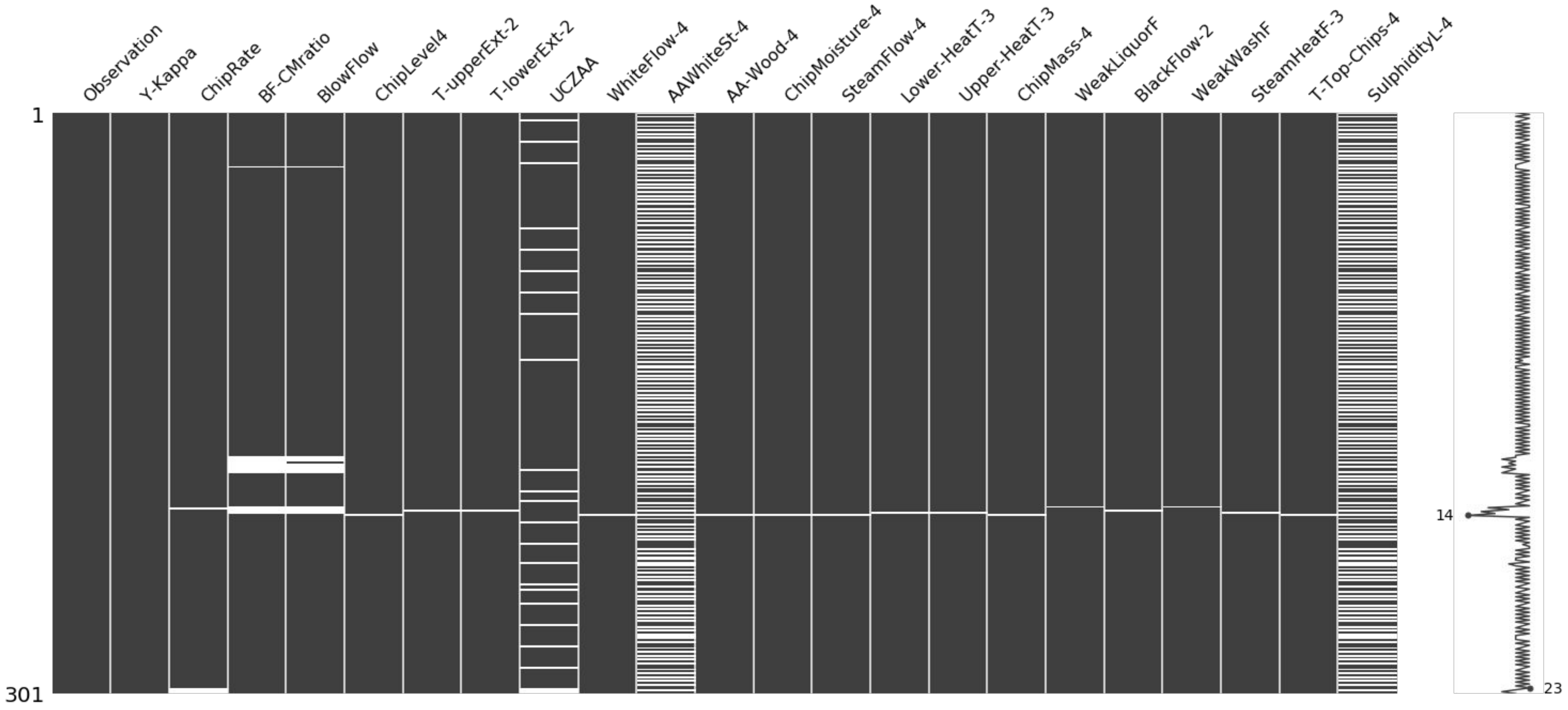
Visualizing Missingness

Concept: A "Missingness Matrix."

Visual: A chart where the X-axis is columns, Y-axis is rows, and pixels are black if data exists and white if missing.

Insight: Helps you see *patterns* in missing data (e.g., "Oh, whenever Age is missing, Income is also missing").

Missingness Matrix: example



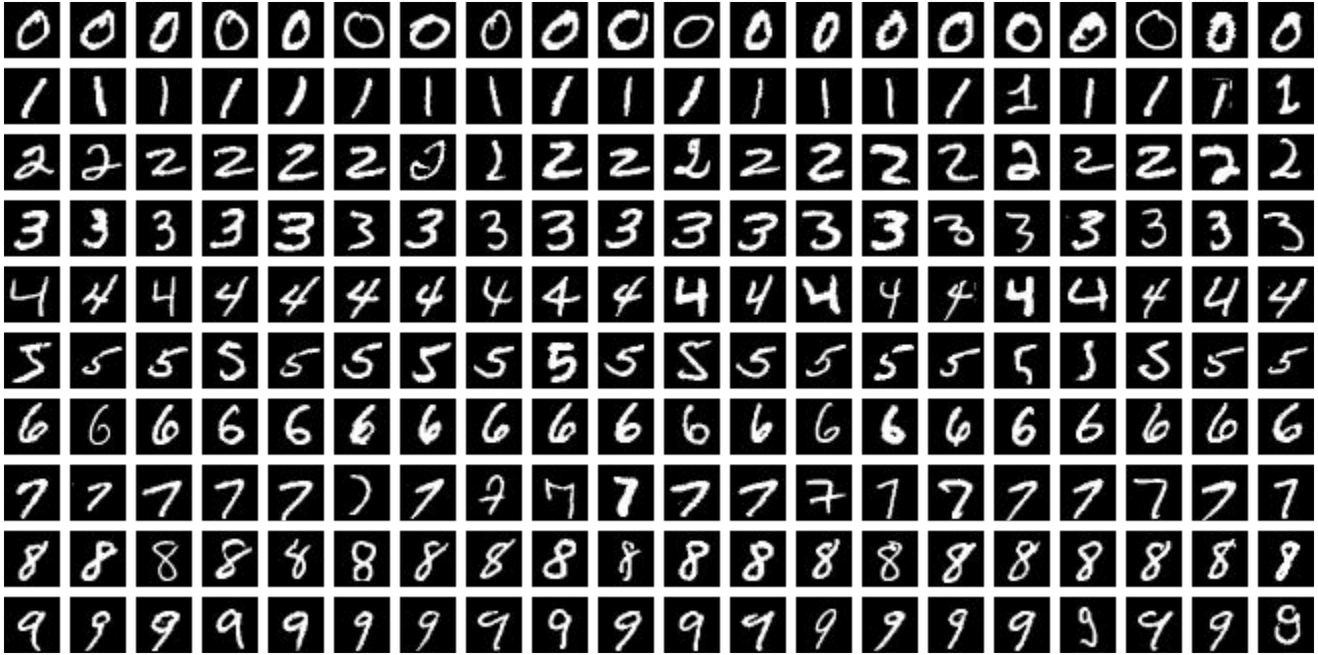
Why Dimensionality Reduction Exists

Humans can’t visualize >3D (we even struggle with 3D).

We reduce dimensions to:

- see structure
- compress information
- remove redundancy

Example: MNIST Dataset



PCA: Intuition

PCA:

- rotates axes
- maximizes variance
- preserves global structure

Useful for:

- compression
- noise reduction
- insight

PCA Limitations

PCA:

- is linear
- ignores labels
- mixes features

High variance ≠ importance.

t-SNE: Intuition

t-SNE:

- preserves local neighborhoods
- exaggerates separation
- destroys global distances

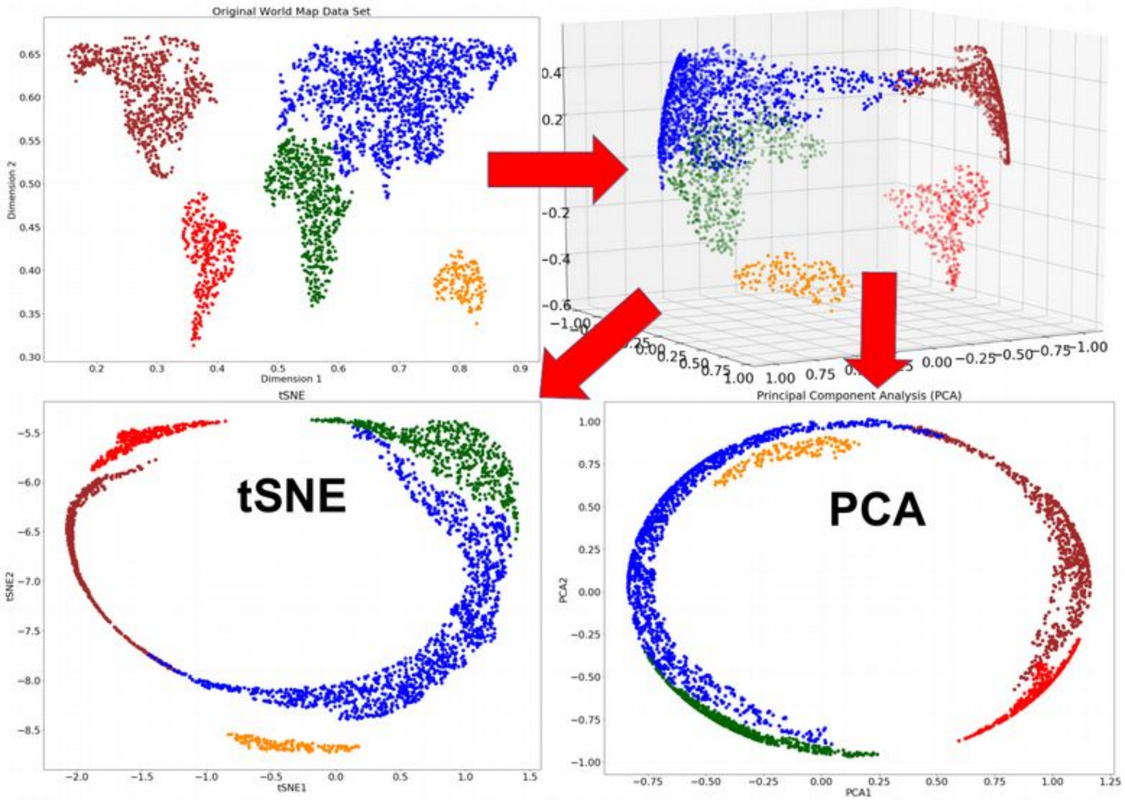
Great for visualization.
Terrible for inference.

PCA vs t-SNE (Mental Model)

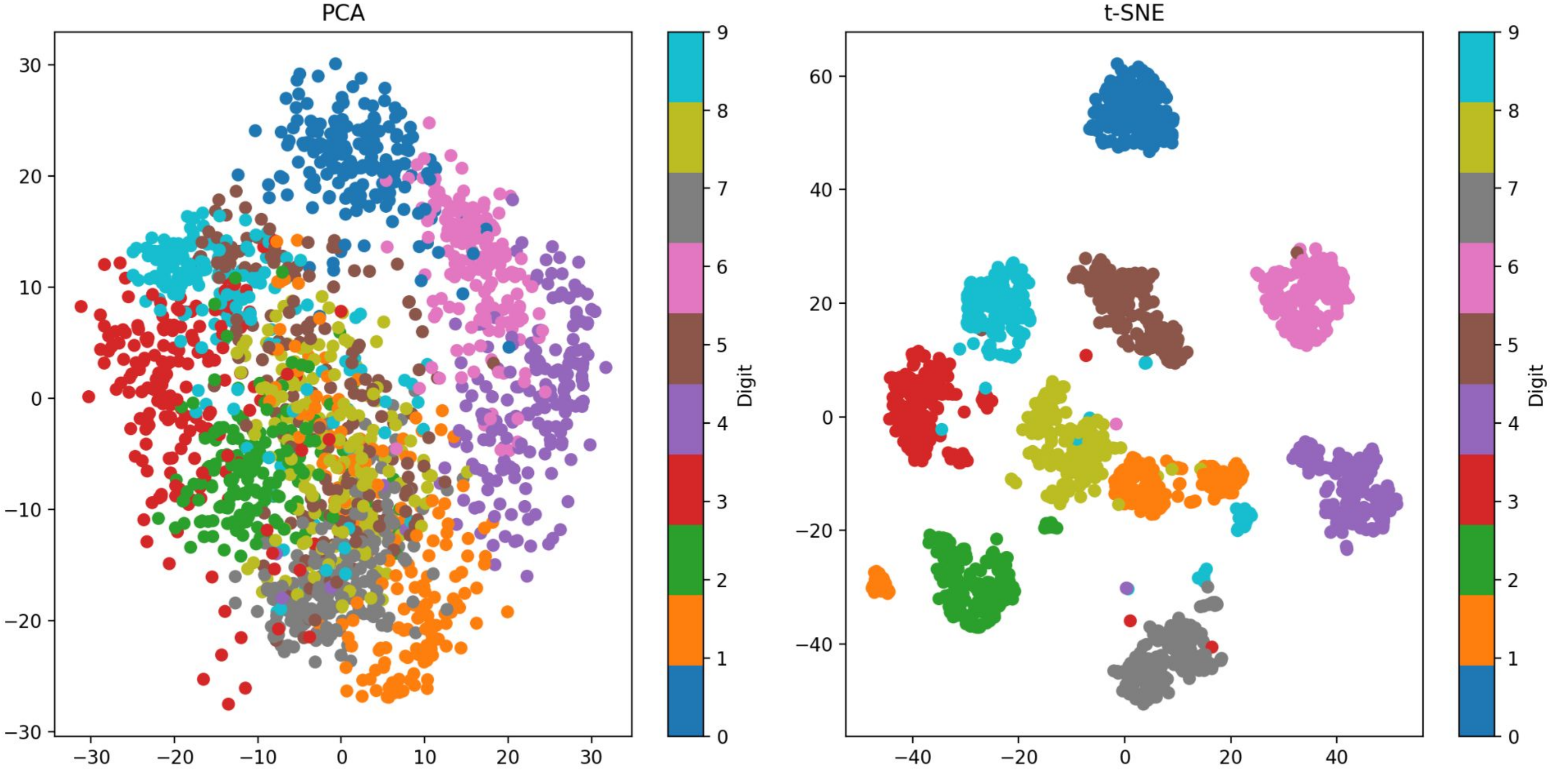
- PCA → explanation, structure
- t-SNE → intuition, clusters

Neither replaces EDA.

Dimensionality Reduction: Visualization



Dimensionality Reduction: Visualization



A Real EDA Workflow

1. Inspect structure & types
2. Univariate distributions
3. Bivariate relationships
4. Multivariate patterns
5. Summarize insights

EDA ends when surprises stop.

Writing Matters

EDA is incomplete without:

- written observations
- stated assumptions
- explicit decisions

If it’s not written, it didn’t happen.

Jupyter Notebook Demo

In this jupyter notebook, we will be demoing the core ideas of EDA and Data visualization: inspecting and questioning data.