The Data Foundation

Why EDA, Preprocessing & Visualization Define Machine Learning Success

The "Garbage In, Garbage Out" Problem

Raw, Messy Data

Real-world data is chaotic. It's filled with missing values, mixed formats (text and numbers), human errors, and extreme outliers. It's unusable in its raw state.

What Algorithms Need

Machine Learning algorithms are powerful but rigid. They are mathematical functions that require clean, complete, numerical, and well-structured data to learn patterns.

Our Three-Step Solution

1. Exploratory Data Analysis (EDA)

The "detective" phase. We investigate the data to find patterns, spot anomalies, and understand its structure.

2. Data Preprocessing

The "cleaning" phase. We fix the problems found in EDA, such as handling missing data, scaling features, and encoding text.

3. Visualization

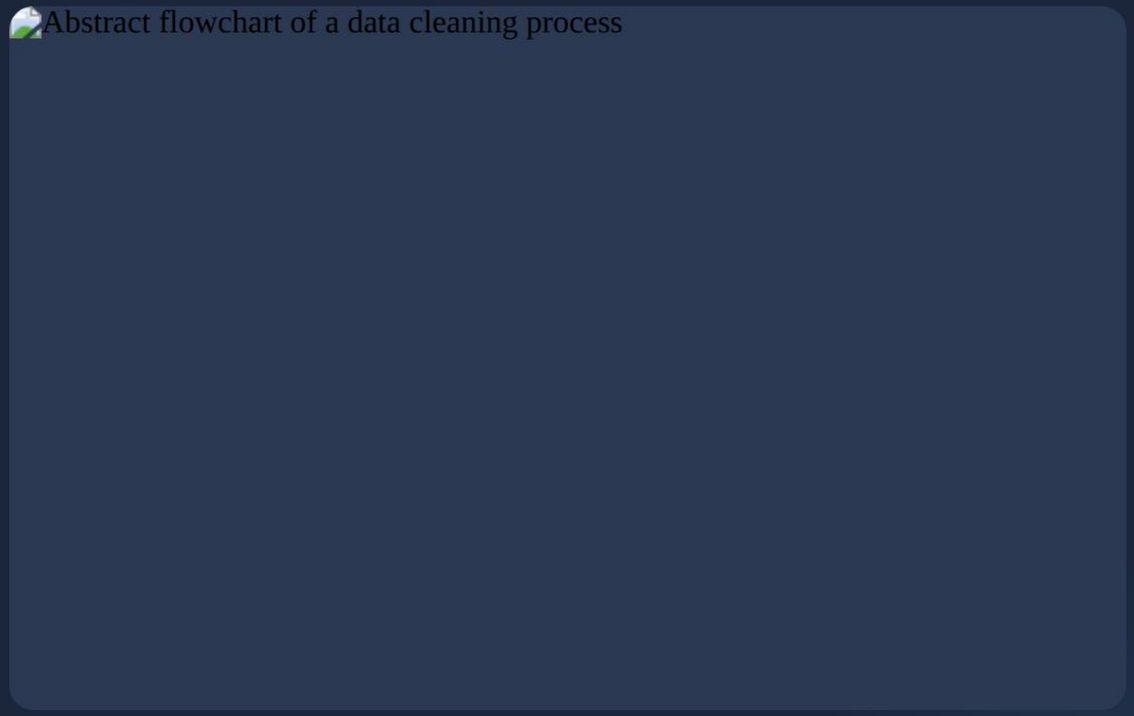
The "storytelling" phase. We use charts and graphs to understand complex relationships and communicate our findings.

Part 1: Exploratory Data Analysis (EDA)

- **Goal:** To understand the data's "personality" before modeling.
- **Action:** Use summary statistics (mean, median, mode) to get a baseline.
- **Action:** Identify all data types (e.g., numerical, categorical, boolean).
- **Action:** Spot anomalies like outliers (a person's age listed as 200) or typos.
- **Action:** Form initial hypotheses (e.g., "Is variable A related to variable B?").

Part 2: Data Preprocessing

This is often the most critical and time-consuming phase. We transform the raw data into a clean, complete, and suitable format that an ML model can consume. This step directly impacts model performance.



Key Tasks Include:

- Handling Missing Values (Imputation)
- Encoding Categorical Data
- Feature Scaling (Normalization)
- Treating or Removing Outliers

How This Helps Algorithms (1/2)

Problem: Categorical Data

An algorithm can't understand text like **["Red", "Green", "Blue"]**. It's a mathematical function, and it only understands numbers.

Solution: One-Hot Encoding

We convert categories into a binary format.
`Red` -> `[1, 0, 0]`
`Green` -> `[0, 1, 0]`
`Blue` -> `[0, 0, 1]`

Why it helps: This enables algorithms (like Linear Regression, Neural Networks) to use non-numeric data in their calculations.

How This Helps Algorithms (2/2)

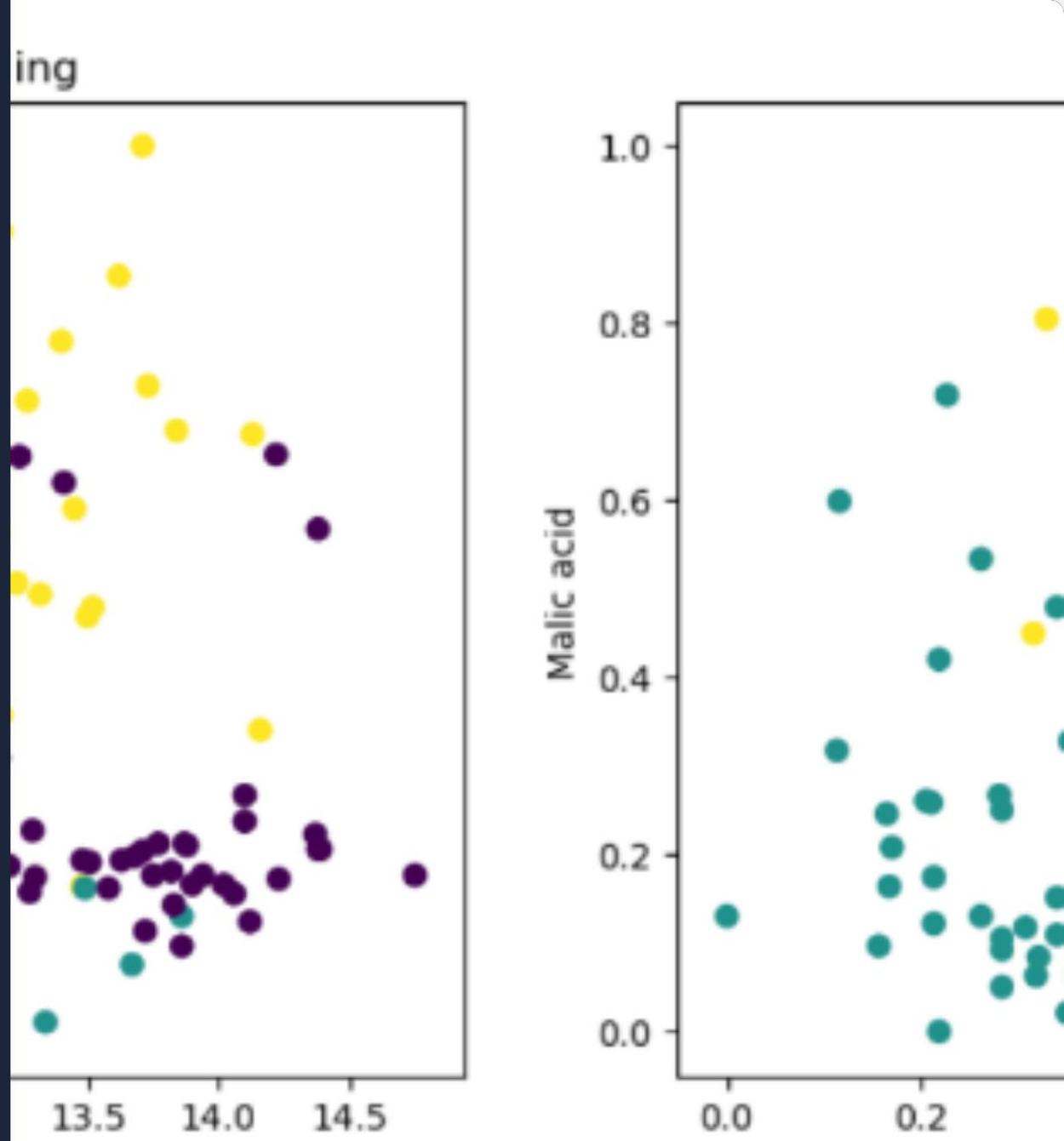
Problem: Different Scales

Imagine two features: **Age (0-100)** and **Salary (0-1,000,000)**. The `Salary` feature will unfairly dominate the model's calculations, making `Age` seem unimportant.

Solution: Feature Scaling

We rescale both features to a similar range (e.g., 0 to 1). This ensures both features are treated equally by the algorithm.

Why it helps: This is critical for algorithms like **k-NN, SVM, and PCA** that are based on distance calculations.



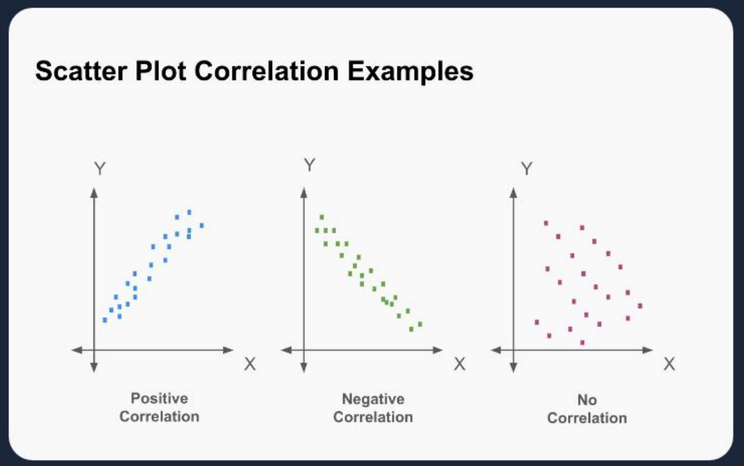
Part 3: The Power of Visualization

> The greatest value of a picture is when it forces us to notice what we never expected to see.

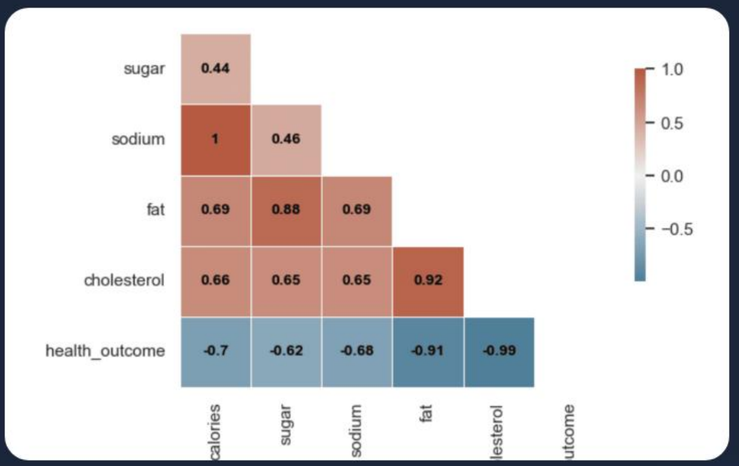
— John Tukey (Pioneer of Exploratory Data Analysis)

Visualization: Seeing the Story

Histogram: To understand the distribution and skew of a single variable.



Scatter Plot: To see the relationship (or correlation) between two variables.



Heatmap: To find correlations between all variables at once.

Real World: Financial Fraud Detection

99.9%

of transactions are legitimate.

The Problem: Imbalanced Data

In fraud detection, 0.1% of data is "fraud." A model could be 99.9% accurate by just guessing "not fraud" every time, which is useless.

The Solution (Preprocessing)

We use techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** to create more "fraud" samples for the model to learn from.

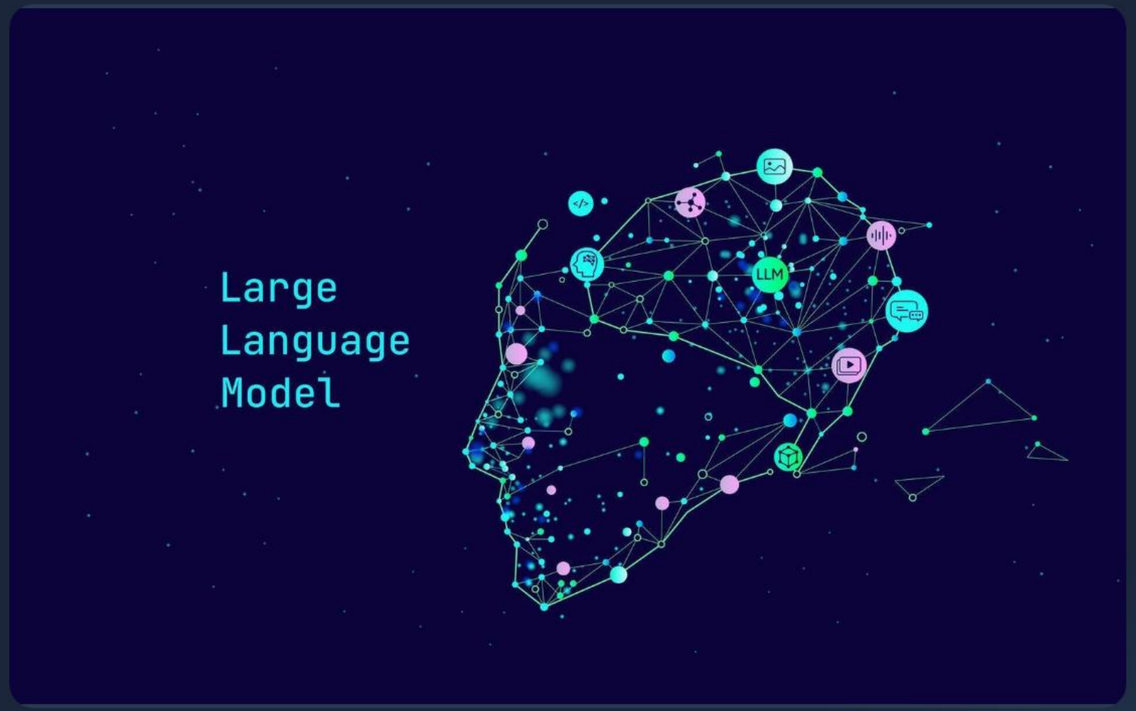
Real World: E-commerce Recommendation

The Problem: Unstructured Text

How does a model understand customer reviews? It can't tell the difference between "This product was amazing!" and "This was an amazing waste of money."

The Solution (NLP Preprocessing)

- **Tokenization:** Splitting sentences into individual words.
- **Stop-word removal:** Removing common words like "a", "is", "the".
- **Vectorization:** Turning the final, clean words into numbers.



Questions?

Thank you.

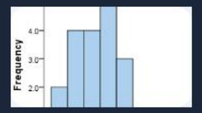
Image Sources

https://www.slideteam.net/wp/wp-content/uploads/2023/03/Organizational-Data-Cleansing-Process-Flow-Chart.jpeg

Source: www.slideteam.net

https://datasciencedojo.com/wp-content/uploads/scatter-plot-from-the-wine-dataset.png

Source: datasciencedojo.com



https://i.sstatic.net/DZrkR.png

Source: stackoverflow.com

https://files.planyway.com/strapi-uploads/assets/Scatter_plot_422d3c3f82.png

Source: planyway.com

https://www.quanthub.com/wp-content/uploads/correlation_heatmap_food_health.png

Source: quanthub.com

https://static.vecteezy.com/system/resources/previews/045/992/276/non_2x/illustration-of-abstract-stream-information-with-cyan-line-and-dot-big-data-technology-ai-data-transfer-data-flow-large-language-model-natural-language-processing-llm-nlp-vector.jpg

Source: www.vecteezy.com