The Data Prep Playbook

From Messy Notebooks to ML-Ready Data

The "Garbage In, Garbage Out" Problem

The Problem: Raw Data

Machine Learning models don't understand text, missing values, or different scales. Feeding them raw, messy data leads to...

❌ Inaccurate predictions
❌ Biased results
❌ Complete model failure

The Goal: ML-Ready Data

We must act as "data translators." Our job is to clean, transform, and structure the data into a format that a model can actually learn from.

✅ Clean (No missing values)
✅ Numerical (All numbers)
✅ Normalized (Common scale)

Our "Patient": The Raw Student Dataset

Student ID	Name	Subject	Marks	Grade	Attendance %
S024	Laura	History	40	D	75
S010	Emma	Math	77	B	91
S038	Zane	English	95	A	54
S045	Bella	Math	35	D	50
S032	Fiona	Math	89	A	97

Step 1 (EDA): The First Look

What is it? Using `.head()` and `.info()` to inspect the data's structure, data types (like 'object' or 'int64'), and look for missing values.

Real-World Analogy: It's like a doctor reading a new patient's chart. Before you can treat them, you need to know their vital signs, allergies, and history.

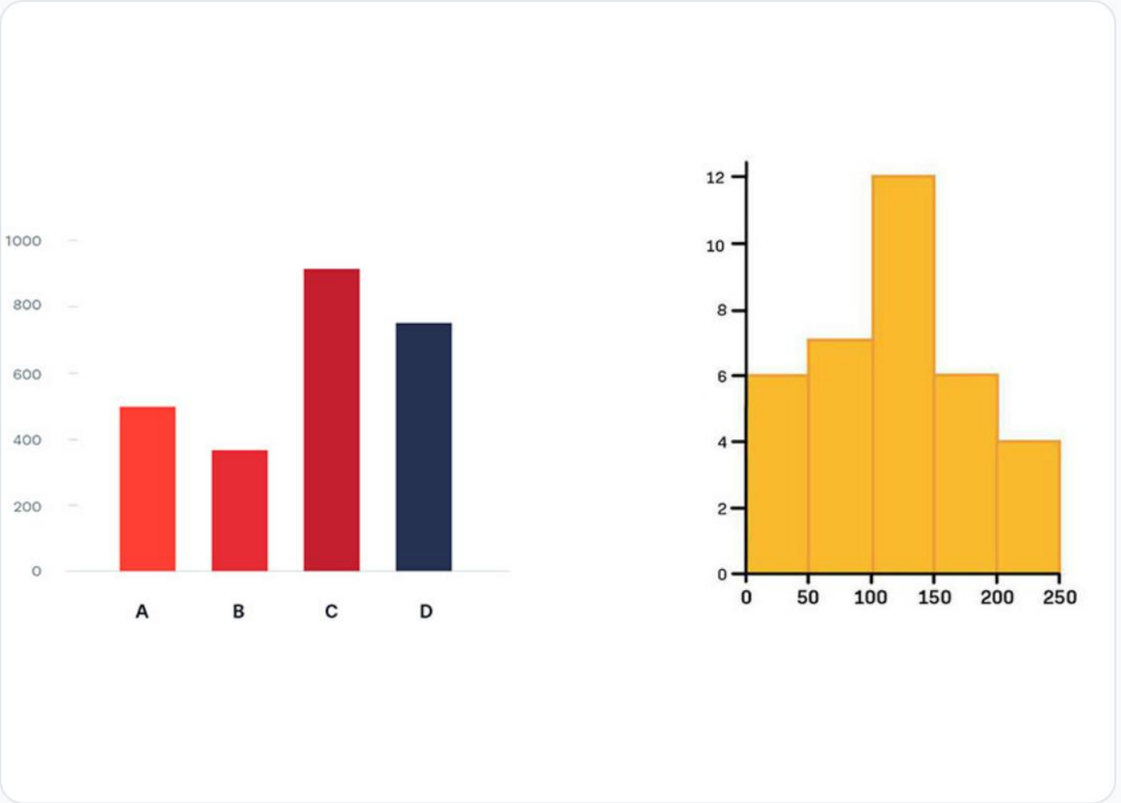
Step 2 (EDA): Understanding Distributions

What are Histograms & Count Plots?

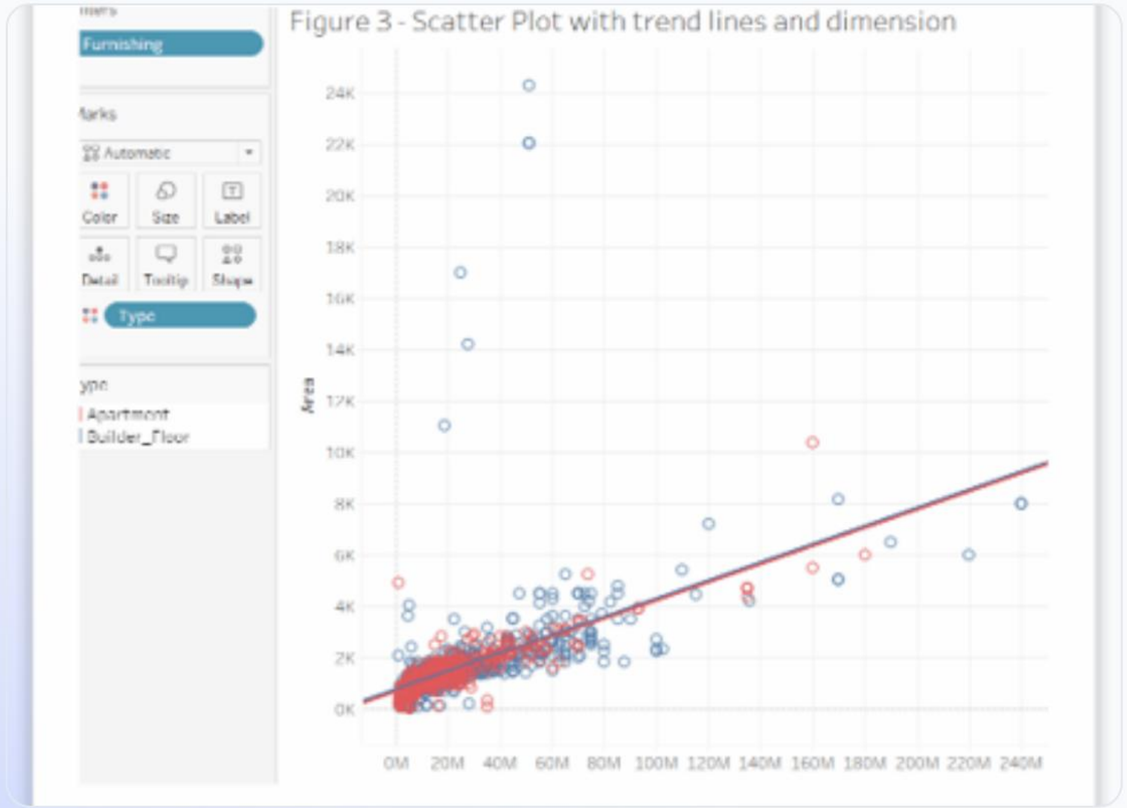
These charts show us the *shape* and *balance* of our data. We ask questions like:

- What is the most common grade? (Count Plot)
- Are most marks high or low? (Histogram)
- Are all subjects equally represented? (Count Plot)

Real-World Scenario: An e-commerce site uses this to see the distribution of 1-star vs. 5-star reviews. If 99% are 5-stars, their "average" rating is misleading!



Step 3 (EDA): Finding Relationships



What is a Scatter Plot?

This chart helps us find relationships between two numerical variables. We look for a pattern.

In Our Data: We plot 'Attendance %' vs. 'Marks'. We see a clear positive trend: as attendance goes up, marks tend to go up. This tells us 'Attendance' is a very important feature!

Real-World Scenario: A business plots 'Ad Spend' vs. 'Sales' to see if their marketing is working. If the plot is flat, they have a problem.

Step 4 (Prep): Cleaning & Ordinal Encoding

Cleaning: Dropping Columns

We drop `Student ID` and `Name`. Why?

They are **identifiers**, not features. A model trained on "Laura" (ID S024) can't make a prediction for "Bob" (ID S011). They just add noise and violate privacy.

Ordinal Encoding: (A > B > C)

We convert `Grade` (A, B, C, D) into numbers (4, 3, 2, 1). This is **Ordinal Encoding**.

We do this because the letters have a clear, built-in order. 'A' is better than 'B'.

Real-World Scenario: Encoding T-shirt sizes (S, M, L, XL) or survey answers (Bad, Neutral, Good).

Step 5 (Prep): One-Hot Encoding

The Problem (Nominal Data)

What about `Subject`? (Math, History, English). There is no order. Is Math > History?

We can't use 1, 2, 3. The model would learn a false relationship.

The Solution (One-Hot)

We create new *binary* columns: 'is_Math', 'is_History', etc. If a student took Math, 'is_Math' is 1 and the others are 0.

This is **One-Hot Encoding**.

Real-World Scenario

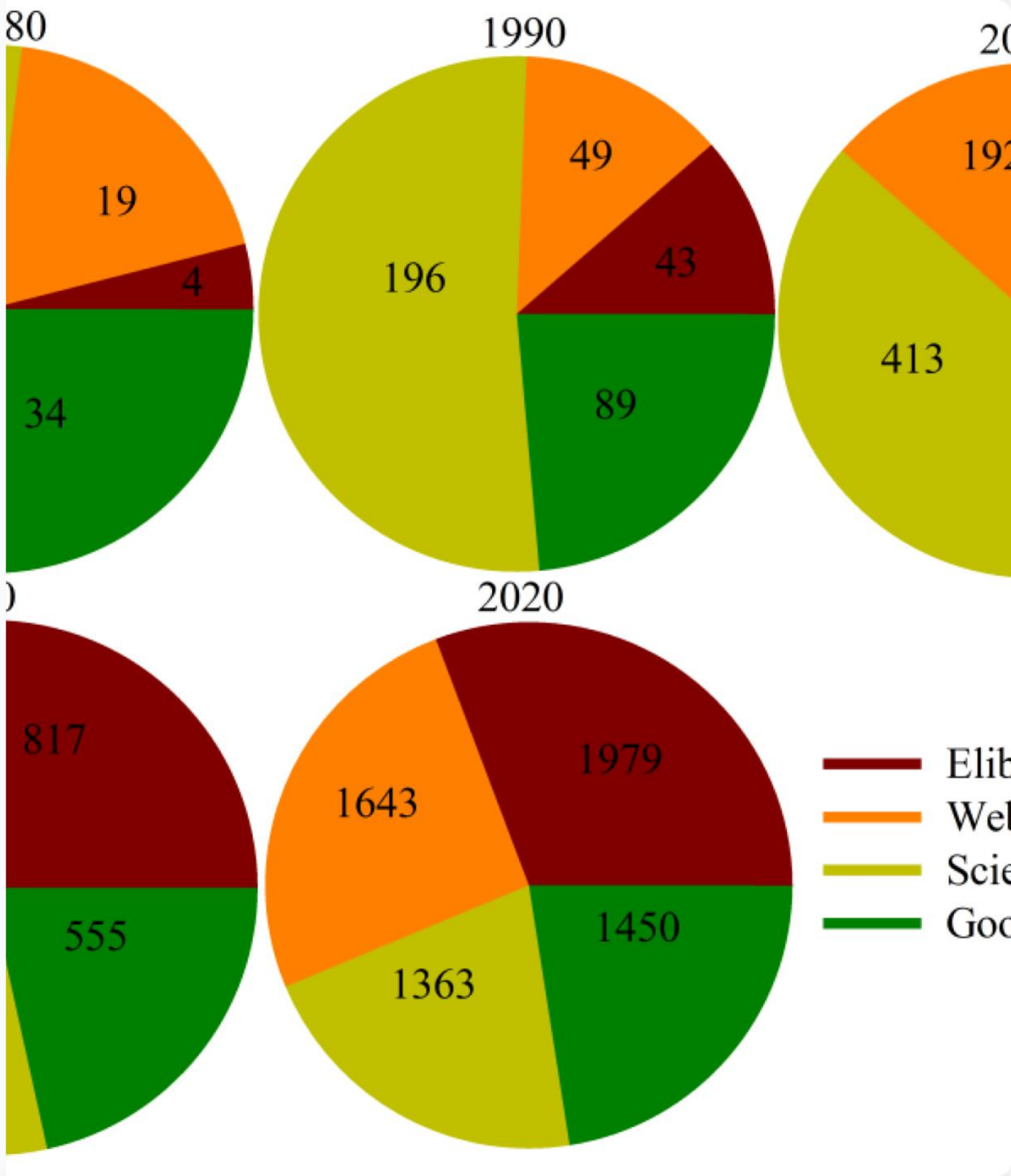
Used for any nominal data, like 'Product Category' (Electronics, Clothing) or 'City' (New York, London, Tokyo).

Step 6 (Prep): Feature Scaling

The "Apples & Oranges" Problem

What if we had 'Age' (0-100) and 'Salary' (50,000 - 1,000,000)? The model will think 'Salary' is 10,000x more important just because the number is bigger!

The Solution: We use `StandardScaler` to put all features on a common scale (e.g., mean of 0, std dev of 1). This ensures the model judges features by their *predictive power*, not their *magnitude*.



Step 7 (Final Check): The Heatmap

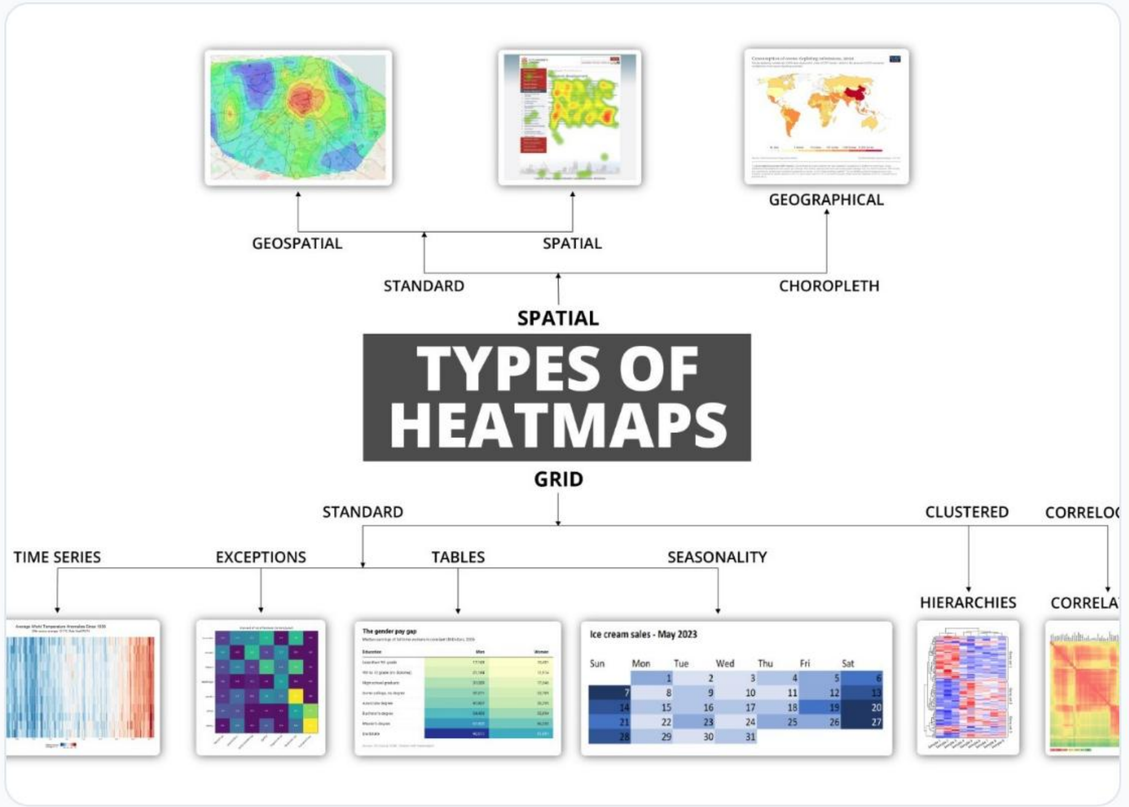
What is a Correlation Heatmap?

After all our work, we run one last visualization on the *processed* data.

This heatmap shows the correlation (relationship) between all our new features and the target (`Grade_Encoded`).

What We Look For:

1. **Good:** 'Marks' and 'Attendance' are still highly correlated with 'Grade'.
2. **Bad:** Two *input* features are highly correlated with *each other* (this is called multicollinearity).



The Transformation: Before vs. After

Before (Raw Data)

Name	Subject	Marks	Grade	Attendance %
Laura	History	40	D	75
Emma	Math	77	B	91
Zane	English	95	A	54

After (ML-Ready Data)

Marks	Attendance %	Grade_Encoded	Subject_History	Subject_Math
-1.05	-0.10	1	1	0
0.65	0.78	3	0	1
1.50	-1.30	4	0	0

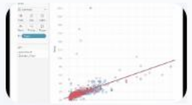
Questions?

The data is now clean, numerical, and scaled. It's finally ready for model training!

Image Sources

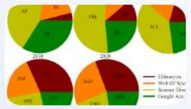
https://wpdatatables.com/wp-content/uploads/2024/03/bar-chart-vs-histogram-feat.jpg

Source: wpdatatables.com



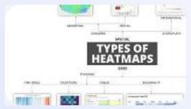
https://businessanalyst.techcanvass.com/wp-content/uploads/2021/11/SP-New-7-1.png

Source: businessanalyst.techcanvass.com



https://pub.mdpi-res.com/sensors/sensors-24-01209/article_deploy/html/images/sensors-24-01209-g001.png?1707894158

Source: www.mdpi.com



https://inforiver.com/wp-content/uploads/image-17.jpeg

Source: inforiver.com