Dataset Overview

- **Rows:** 10,000
- **Columns:** 14
- **Memory Usage:** ~1.1 MB
- **No missing values** – all columns are fully populated.

🏷 Columns Breakdown

Column	**Type**	**Description**
UDI	int64	Unique identifier for each row (just an index).
Product ID	object	Unique ID for the manufactured product (e.g., M14860).
Type	object	Product type (categorical: likely L, M, H).
Air temperature [K]	float64	Air temperature in **Kelvin**.
Process temperature [K]	float64	Process (internal machine) temperature in **Kelvin**.
Rotational speed [rpm]	int64	Rotational speed of the machine in **RPM**.
Torque [Nm]	float64	Torque applied on the machine in **Newton-meters**.
Tool wear [min]	int64	Tool wear measured in **minutes**.
Machine failure	int64	Binary label (0 = No failure, 1 = Failure).
TWF	int64	Tool Wear Failure (0/1).
HDF	int64	Heat Dissipation Failure (0/1).
PWF	int64	Power Failure (0/1).
OSF	int64	Overstrain Failure (0/1).
RNF	int64	Random Failure (0/1).

- The **target column** for this ML task would be Machine failure.
- The other failure columns (TWF, HDF, PWF, OSF, RNF) are **subcategories** of machine failure (helpful for root-cause classification).
- Features include **environmental (temperature)**, **operational (speed, torque)**, and **wear indicators**.

Here’s the **Exploratory Data Analysis (EDA)** results for The Maintenance dataset:

□ 1. Data Quality Check

✅ **No missing values** were found in any of the 14 columns.
✅ **No obvious erroneous values** (e.g., negative RPM, torque, or temperature) — all data ranges are reasonable.

📊 2. Descriptive Statistics (Key Highlights)

Feature	Mean	Std	Min	Max	Insight
Air temperature [K]	300.0	2.0	295.3	304.5	Stable environment (small variation).
Process temperature [K]	310.0	1.48	305.7	313.8	Slightly higher & stable than air temp.
Rotational speed [rpm]	1539	179	1168	2886	Most values near 1400–1600 rpm.
Torque [Nm]	39.99	9.97	3.8	76.6	Wide spread; some low/high torque outliers.
Tool wear [min]	108 min	64 min	0	253	Even distribution across low-to-high wear.

Feature	Mean	Std	Min	Max	Insight
Machine failure	**3.39% failures**	-	-	-	Imbalanced dataset (failures are rare).

📉 3. Failure Analysis

- **Failure Rate:** Only **3.39%** of records have Machine failure = 1.
- Failures are **rare events**, so this is an **imbalanced classification problem**.
- Subtypes (TWF, HDF, PWF, OSF, RNF) have even lower rates (below 1.1%).

🧪 4. Correlation Analysis

- **Process temperature** and **air temperature** are moderately correlated (as expected).
- **Torque** and **rotational speed** show some negative correlation (higher speed = lower torque).
- **Machine failure** shows **weak correlation** with individual numeric features → failures are likely caused by a **combination of factors**, not just a single feature.

⚠ 5. Anomalies

- **Torque [Nm]** has some very low values (~3.8 Nm) which may be normal but could indicate special cases.
- **Rotational speed [rpm]** has a few unusually high values near 2800 (might be worth checking if these are outliers or normal high-speed operations).

✅ Extra EDA Steps I think must be Included

1. Failure vs. Non-Failure Feature Comparison

Instead of just plotting all data together, compare:

- **Distributions** of torque, speed, temperatures for failed vs. non-failed machines (boxplots or KDE plots).
- This will show which features are most predictive of failure.

2. Categorical Feature Analysis

- Analyze Type (L/M/H):
  - Countplot showing how failures are distributed across product types.
  - Check if certain product types fail more often.

3. Outlier Detection

- Use **IQR method** or **Z-scores** to formally flag extreme values for:
  - Torque
  - Rotational speed
  - Tool wear
- Decide whether to keep them (if they are valid rare cases) or treat them.

4. Pairwise Feature Relationships

- Use **pairplots** or **scatter plots** (colored by failure) to see:
  - If failures cluster in certain regions (e.g., low torque + high speed).
  - Whether there's interaction between multiple features leading to failures.

5. Feature Correlation with Target

- Calculate **Point-biserial correlation** (or use feature importance from a quick decision tree) to rank features by relevance to Machine failure.