Normalization & Standardization Tutorial (with Math + Python)

Step 1: Create the Dataset

We create a simple dataset with Age and Salary using Python:

```
import pandas as pd

data = {
    "Age": [25, 30, 35, 40, 45],
    "Salary": [30000, 40000, 50000, 60000, 70000]
}

df = pd.DataFrame(data)
print("Original Dataset:\\n", df)
```

Output:

```
   Age  Salary
0   25   30000
1   30   40000
2   35   50000
3   40   60000
4   45   70000
```

Step 2: Normalization (Min-Max Scaling)

Formula: `x' = (x - min(x)) / (max(x) - min(x))`

Step 2a: Manual Normalization (Mathematics)

Normalize Age:

Min(Age) = 25, Max(Age) = 45
Age_normalized = (Age - 25) / (45 - 25) = (Age - 25)/20

Age	Calculation	Normalized
25	(25-25)/20	0.0
30	(30-25)/20	0.25
35	(35-25)/20	0.5
40	(40-25)/20	0.75
45	(45-25)/20	1.0

Normalize Salary:

Min(Salary) = 30000, Max(Salary) = 70000
Salary_normalized = (Salary - 30000) / (70000 - 30000) = (Salary - 30000)/40000

Salary	Calculation	Normalized
30000	(30000-30000)/40000	0.0
40000	(40000-30000)/40000	0.25
50000	(50000-30000)/40000	0.5
60000	(60000-30000)/40000	0.75
70000	(70000-30000)/40000	1.0

Step 2b: Normalization using Library (MinMaxScaler)

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
df[["Age_Norm_lib", "Salary_Norm_lib"]] = scaler.fit_transform(df[["Age", "Salary"]])
```

Step 3: Standardization (Z-Score Scaling)

Formula: `x' = (x - μ) / σ`

Step 3a: Manual Standardization (Mathematics)

Standardize Age:

Mean(Age) = (25+30+35+40+45)/5 = 35
Std(Age) = sqrt(((25-35)^2 + (30-35)^2 + (35-35)^2 + (40-35)^2 + (45-35)^2)/5) = sqrt(50) ≈ 7.071
Age_standardized = (Age - 35) / 7.071

Age	Calculation	Standardized
25	(25-35)/7.071	-1.414
30	(30-35)/7.071	-0.707
35	(35-35)/7.071	0
40	(40-35)/7.071	0.707
45	(45-35)/7.071	1.414

Standardize Salary:

> Mean(Salary) = (30000+40000+50000+60000+70000)/5 = 50000
> Std(Salary) = sqrt(((30000-50000)^2 + (40000-50000)^2 + ... + (70000-50000)^2)/5) = sqrt(200000000) ≈ 14142.14
> Salary_standardized = (Salary - 50000) / 14142.14

Salary	Calculation	Standardized
30000	(30000-50000)/14142.14	-1.414
40000	(40000-50000)/14142.14	-0.707
50000	(50000-50000)/14142.14	0
60000	(60000-50000)/14142.14	0.707
70000	(70000-50000)/14142.14	1.414

Step 3b: Standardization using Library (StandardScaler)

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df[["Age_Std_lib", "Salary_Std_lib"]] = scaler.fit_transform(df[["Age", "Salary"]])
```

Step 4: Comparison Table

Method	Age	Salary
Original	25-45	30000-70000
Min-Max	0-1	0-1
Z-Score	-1.414 → 1.414	-1.414 → 1.414

Step 5: Summary

- **Normalization:** Scales data to 0-1, preserves distribution, sensitive to outliers.
- **Standardization:** Centers data to mean=0 and std=1, handles outliers better, assumes Gaussian distribution.
- Manual calculation and library functions give the same results.