Data Preprocessing Pipeline for Clustering

Raw Data

- Missing values
- Different scales
- Outliers

Example:

`[NaN, 100, 0.5, 1000]`

Cleaning

- Fill missing
- Remove duplicates
- Fix errors

↓

`[50, 100, 0.5, 1000]`

↓

Scaling

- Standardization
- Normalization
- Min-Max scaling

`[-0.5, 0.8, -1.2, 1.5]`

↓

`[PC1, PC2, PC3]`

Feature Engineering

- Create new features
- Combine features
- Transform features

↓

Ready!

Ready for ML

- Clean data
- Scaled features
- Optimized for clustering