# Data Wrangling I

- Wrangling = data surgery → no real strict definition, but to me is getting the data into an analyzable state
  - Cleaning/EDA
  - Joining/Filtering
  - Transforming/Normalizing/Scaling if necessary
- Today we are going to cover:
  - Transformations and Normalizations/Scaling

# Transforming

- What it does: Changes the shape of the distribution.
- Examples: log, square root, reciprocal, Box-Cox, sigmoid.
- Why: Fix skewness, stabilize variance, reveal hidden linearity.
- Effect on correlation: Can change Pearson correlation by altering relationships between variables.
- Critical for some modeling assumptions (mostly statistical)

# Normalizing

- In statistics / ML preprocessing: Often used interchangeably with “scaling,” meaning “rescale to a standard range or distribution” (e.g., z-score standardization).
- Changes the **scale** of the data, but preserves the **shape** of the distribution.
- Example: z-score, mean = 0, std = 1.
  - Critical to realize you can normalize/scale over different values (subjects/time)

# Scaling

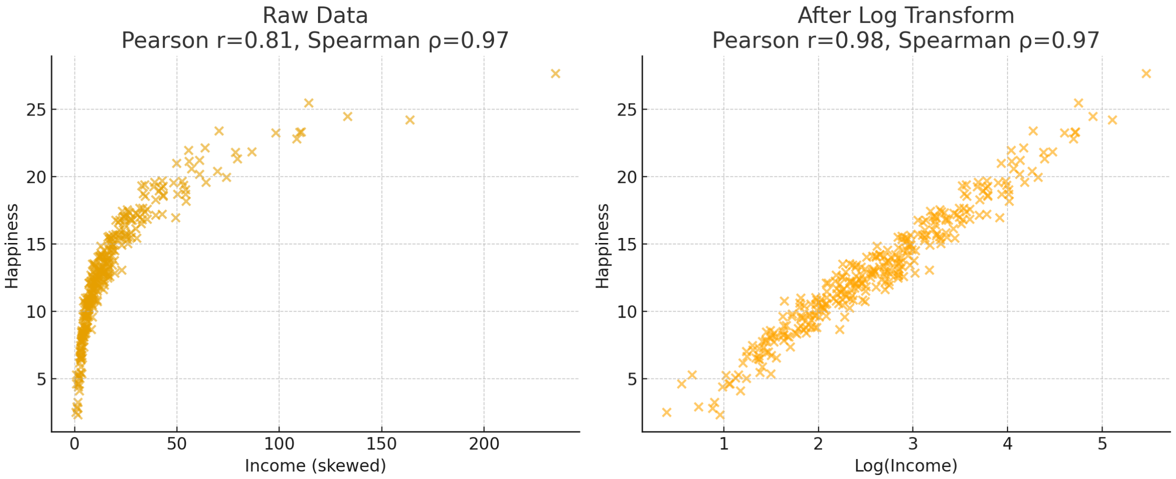
- What it does: Changes the range or unit of a variable, but **not its shape**.
- Examples: min–max scaling to [0,1], dividing by max, unit norm scaling.
- Why: Put different variables on comparable ranges (important for distance-based ML).
- Effect on correlation: Does not change Pearson correlation (linear rescaling leaves r the same).

# Why Transform Data?

- Skewness, scaling differences, and nonlinear patterns distort relationships
- Correlation may appear weak even when relationships exist
- Transformations and normalizations help reveal structure

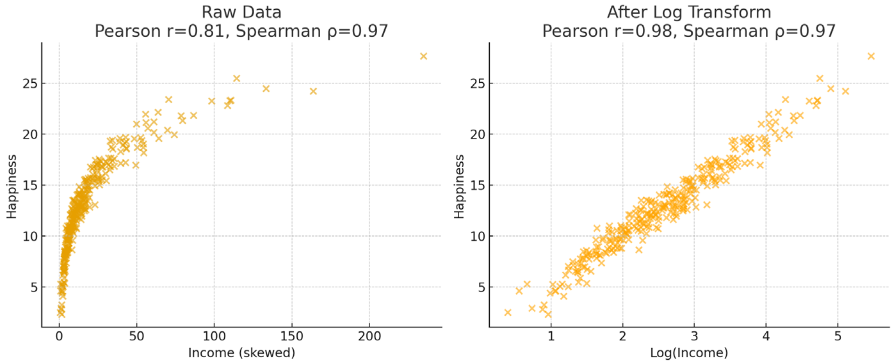
# Example: Income vs Happiness

- Scatterplot: nonlinear relationship
- Pearson correlation is low
- But visually, a relationship is clear



# General Transformation Idea

- Raw Income–Happiness:
  - Pearson correlation (r ≈ 0.81)
  - Spearman correlation strong (ρ ≈ 0.97)
  - After log transform: Pearson correlation improves (r ≈ 0.98)
  - If Spearman already captures the relationship, why bother transforming?



# Transform vs Technique

- No need to force a log transform (negative numbers)
- Captures monotonic trend without extra preprocessing
- Simple, nonparametric → fewer assumptions
- Doesn’t Spearman have an inherent transformation?

# Transform vs Technique

- **Transform** when:
  - Relationship is monotonic but nonlinear (log, sqrt help).
  - Skew or outliers distort correlation.
  - Model assumptions (e.g., normality, homoscedasticity) need to be met.
  - You want to reveal hidden linear structure.
- **Change Technique** when:
  - Relationship is inherently nonlinear or non-monotonic (U-shape, quadratic).
  - No reasonable transformation improves correlation.
  - Ordinal/rank information matters more than exact values → switch to Spearman/Kendall.
  - Binary or categorical outcomes → logistic regression, classification, or nonparametric methods.
- Rule of thumb: *If a simple monotonic transformation improves linear correlation, transform. If the pattern stays nonlinear or non-monotonic, switch to a different analytic technique.*
- Key Message: Both are valid — judgment depends on your analytic goal

# Types of Transforms

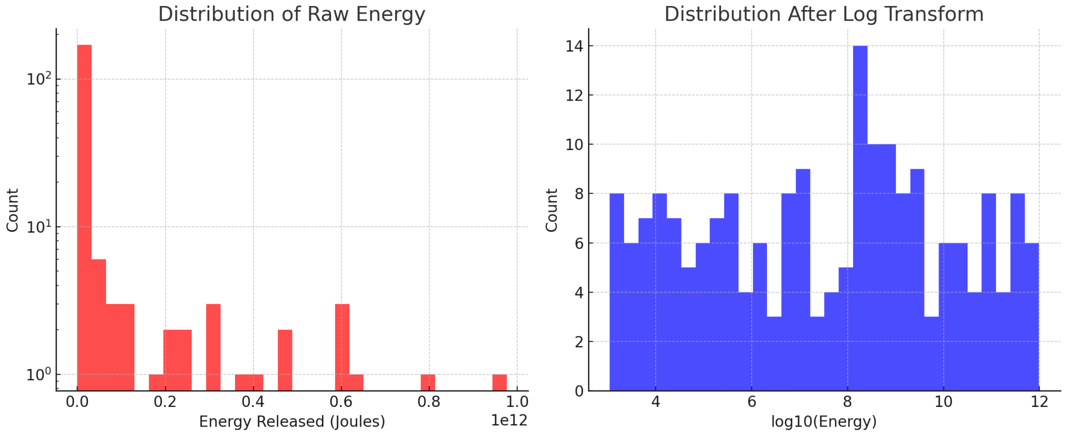
- Equation based – e.g. log, ln, sqrt, inverse
- Thresholding/Quantiling – segmenting data into buckets
- <u>Transformations change the distributions of variables, use with caution!</u>

# Log Transform: Intro

- Purpose: reduce skewness, stabilize variance
- Common for incomes, population sizes, growth data, frequency data
- Wherever you have large ranges (orders of magnitude) differences in values

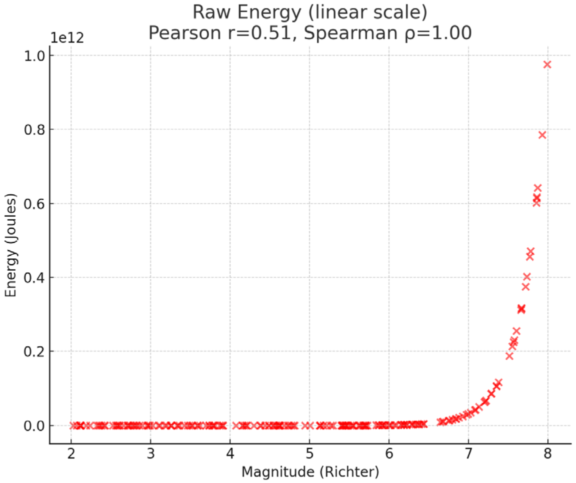
# Log Transform: Visual

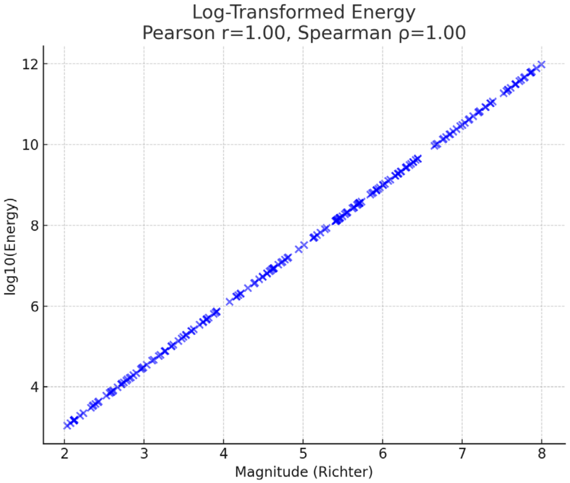
- Suppose you’re a seismologist studying earthquakes. The energy released by earthquakes varies enormously: tiny tremors release barely anything, while large quakes release millions of times more energy. If we plot energy on a raw scale, the big ones dominate and the small ones disappear. To compare patterns across all magnitudes, we use a log transform.



# Log Transform Correlation

- Raw: Pearson correlation weak or unstable (dominated by outliers)
- After log: Pearson correlation strong, linear, interpretable
- Spearman correlation was always high (monotonic) → log transform brings Pearson in line


Raw Energy (linear scale)
Pearson r=0.51, Spearman ρ=1.00

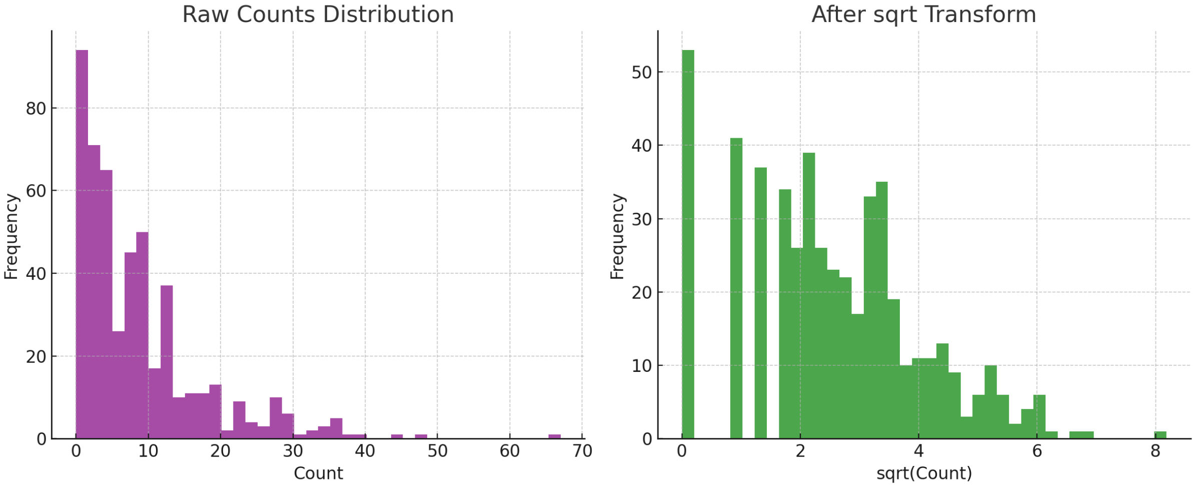

Log-Transformed Energy
Pearson r=1.00, Spearman ρ=1.00

# Square Root Transform: Intro

- Purpose: moderate skew reduction
- Common for counts/frequencies
- Gentle correction for right skew

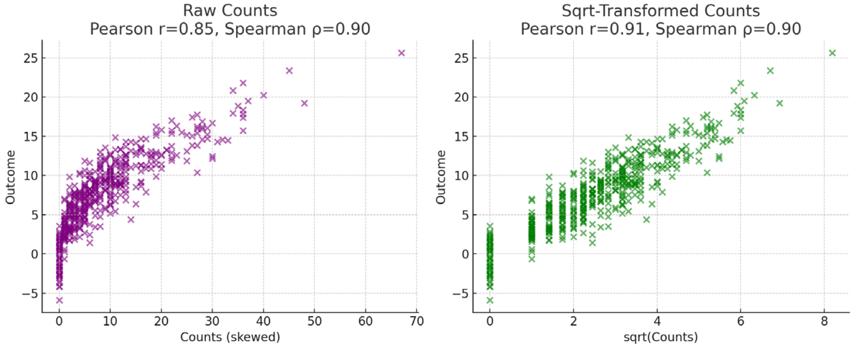
# Square Root Transform: Visual

- Histogram of variable before and after square root transform
- Distribution or scale changes visibly



# Square Root Transform: Correlation

- Imagine you’re analyzing data from a company’s customer support center. Each data point is a single customer, and the variable Counts is the number of times they called the help line in a month. Most customers never call or only call once. A few have 20, 30, or even 50 calls — the ‘frequent flyers’ who swamp support.
- When you plot the raw counts, the distribution is extremely skewed — almost everyone is bunched near zero, and those few heavy users stretch the scale.
- Now suppose we want to study how call frequency relates to customer satisfaction. On the raw scale, the skew drowns out most of the variation. But if we take the square root of counts, the frequent callers get compressed, the low callers spread out, and the relationship with satisfaction becomes clearer and more linear.

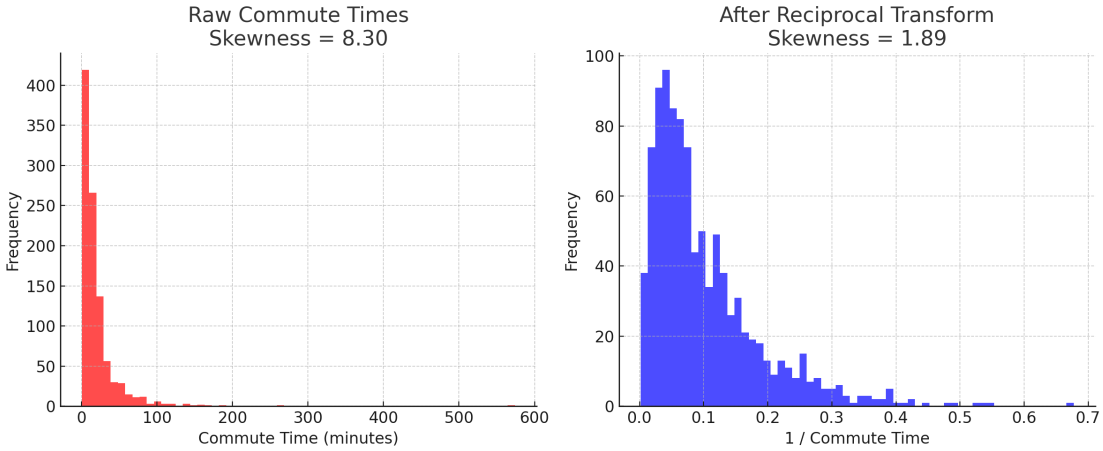


# Reciprocal Transform: Intro

- Purpose: invert large/small values
- Useful for ratios, travel times, diminishing effects
- Can linearize hyperbolic trends

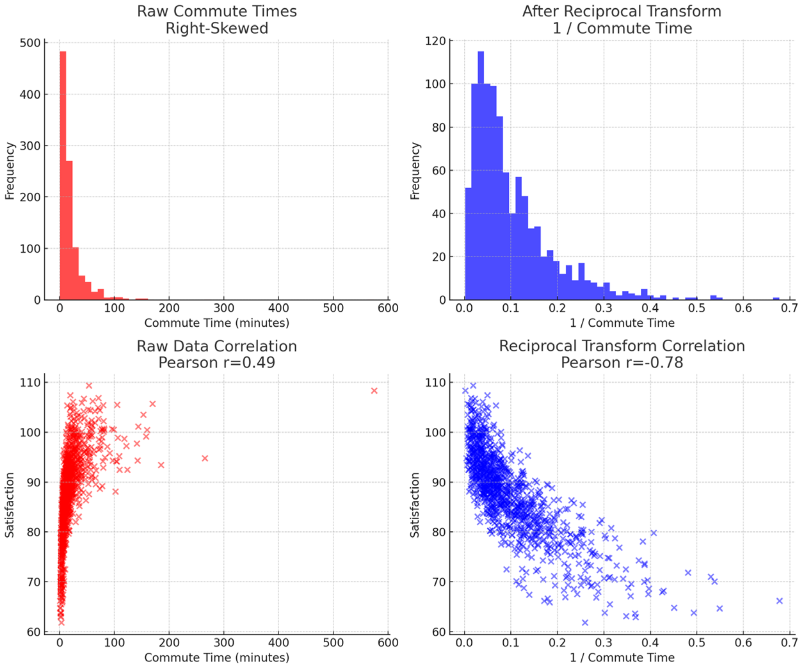
# Reciprocal Transform: Visual

- Histogram/Scatter of variable before and after reciprocal transform
- Distribution or scale changes visibly
- Example improves interpretability



# Reciprocal Transform: Correlation

- Suppose we're studying how commute time affects job satisfaction. In the raw data, long commutes dominate the scale, and the relationship looks nonlinear.
- If we apply a reciprocal transform (1/time), the very long commutes get compressed, short commutes are spread out, and the relationship becomes more linear and interpretable.
- What did it do to the sign of correlation?

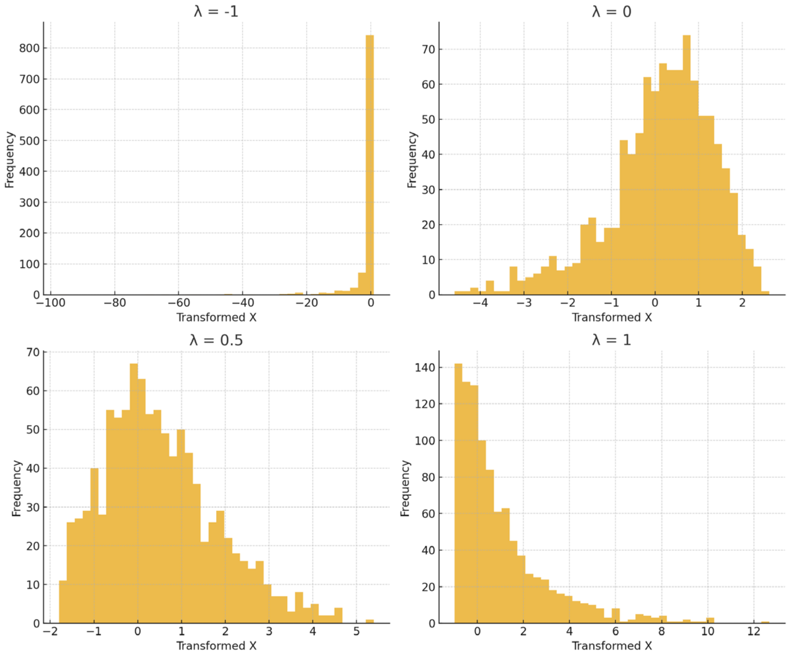


# Box-Cox / Yeo-Johnson: Intro

- Flexible power transforms
- Adjust shape to approximate normality
- Broader than log/sqrt for skew correction

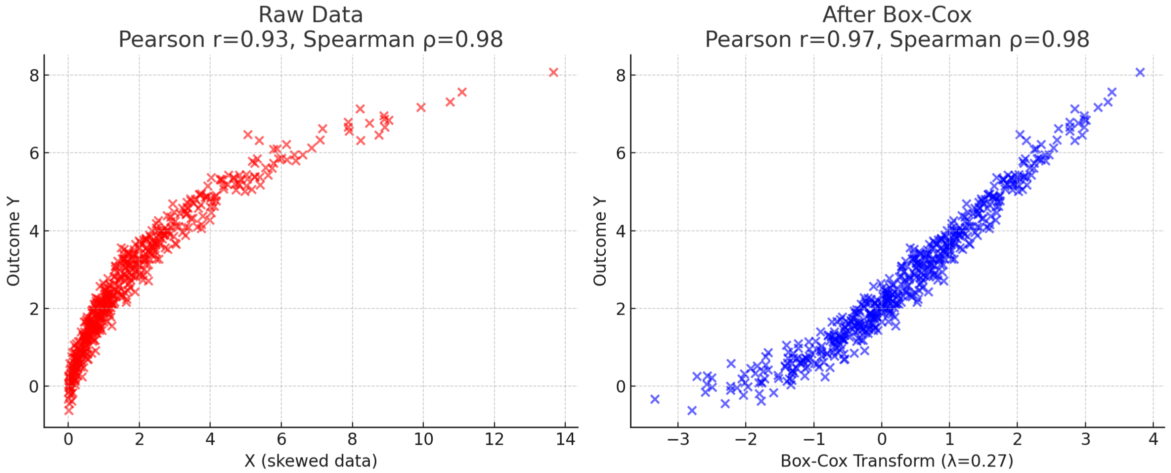
# Box-Cox / Yeo-Johnson: Visual

- Dataset: skewed income or wait times.
- Try different λ values:
  - λ = 1 → no change.
  - λ = 0 → log transform.
  - λ ≠ 0,1 → stretches/compresses differently.
- Typically handled by optimization — λ is chosen to maximize log-likelihood under a normality assumption.
- Yeo-Johnson extends to handle negative values.



# Box-Cox / Yeo-Johnson: Correlation

- Raw data: skew distorts linear correlation.
- Box-Cox/Yeo-Johnson: optimized transform reveals linearity.
- Pearson correlation improves; Spearman remains stable.
- Scatterplot with outcome variable:
  - Left = raw skewed predictor vs. outcome (weak Pearson).
  - Right = Box-Cox transformed predictor vs. outcome (stronger Pearson).

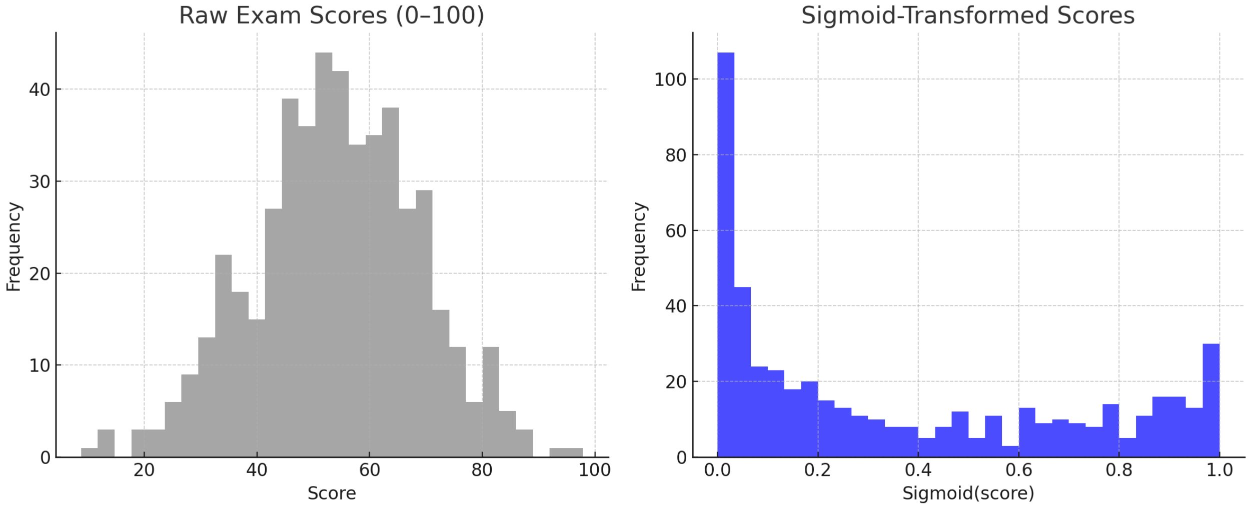


# Sigmoid Transform: Intro

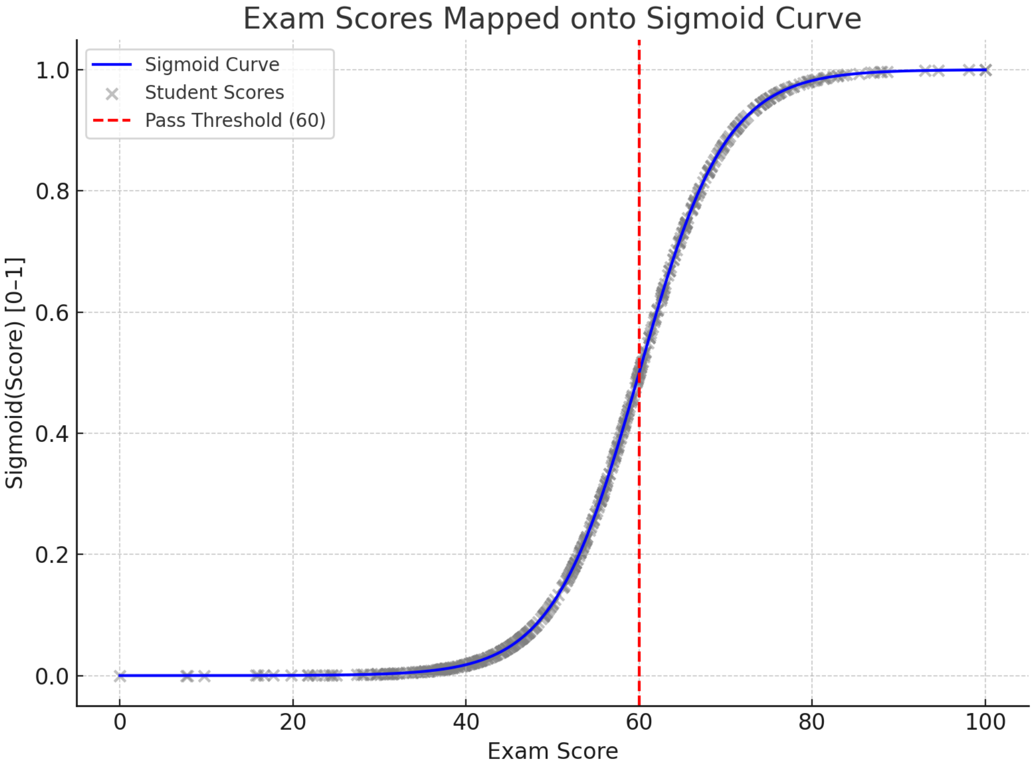
- Formula: σ(x) = 1 / (1+e^-x)
- Squashes values into (0,1) range
- Foreshadows logistic regression

# Sigmoid Transform: Visual

- Imagine you’re looking at exam scores from 0 to 100. Professors don’t care about differences between a 10 and a 20 — both are failing. But near the pass mark (say, 60), small differences matter a lot.
- Logistic regression models this idea with a sigmoid transform: it squashes raw scores into values between 0 and 1, where the middle region shows the biggest changes



# Sigmoid Transform: Correlation



# Thresholding/Quantiling: Intro

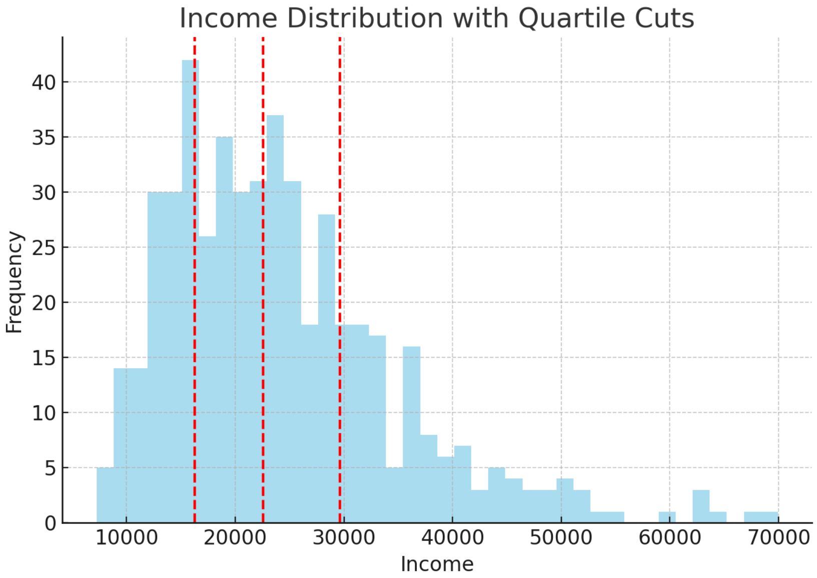
- Turns continuous variable into binary
- Useful in domains with natural cutoffs (medical, legal)
- Risk: loss of nuance

# Thresholding/Quantiling

- Splits continuous data into quantiles or bins
- Useful for skewed data, group comparisons
- Risk: arbitrary cutoffs, information loss
- Continuous ages -> binary: Age > 65 = 1, else 0, or smaller: Ages 0 > 40 = 1, Ages >45 <65 = 2, etc.
- Pearson weakens with thresholding
- Point-biserial/chi square correlation may be more appropriate

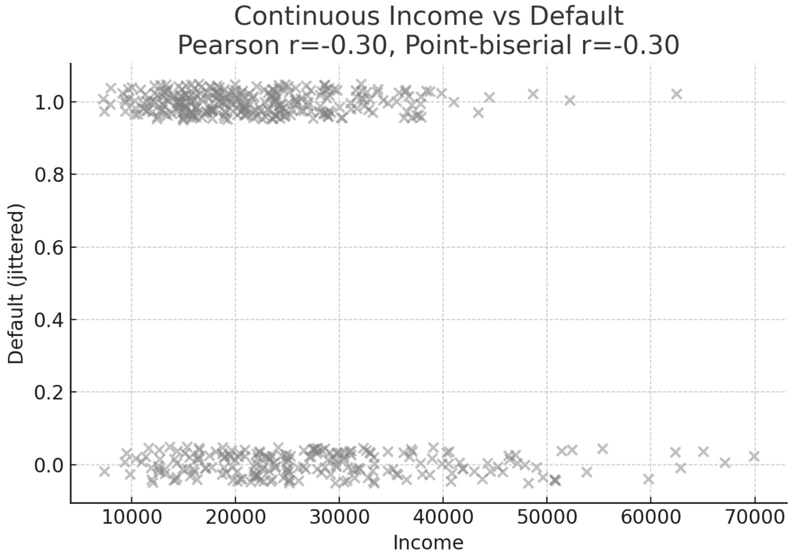
# Thresholding/Quantile Example

- Suppose we want to understand how income relates to whether someone defaults on a loan.
- If we use income as a continuous predictor, Pearson correlation might be weak.
- But if we chop the income distribution into quantiles, the relationship becomes clear:
  - default rates are much higher in the lowest group and drop off in the higher groups.



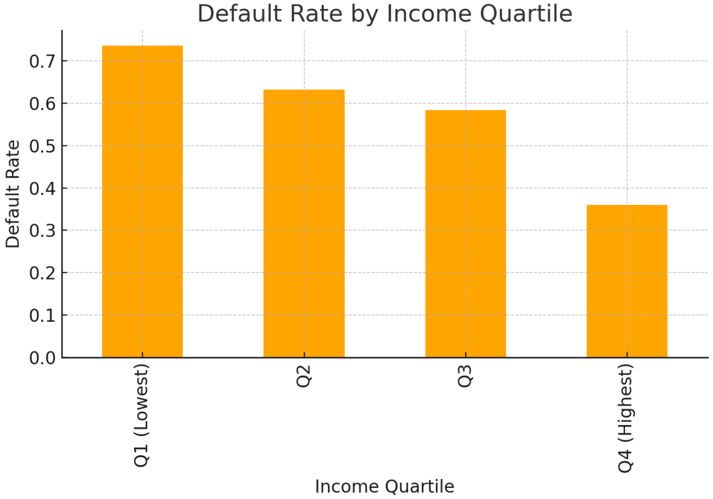
# Raw Data

- Continuous income vs. default (jittered).Pearson r ≈ –0.30 (weak linear correlation).
- Point-biserial r is the same here (since binary outcome).



# Default Rates by Income Quartile

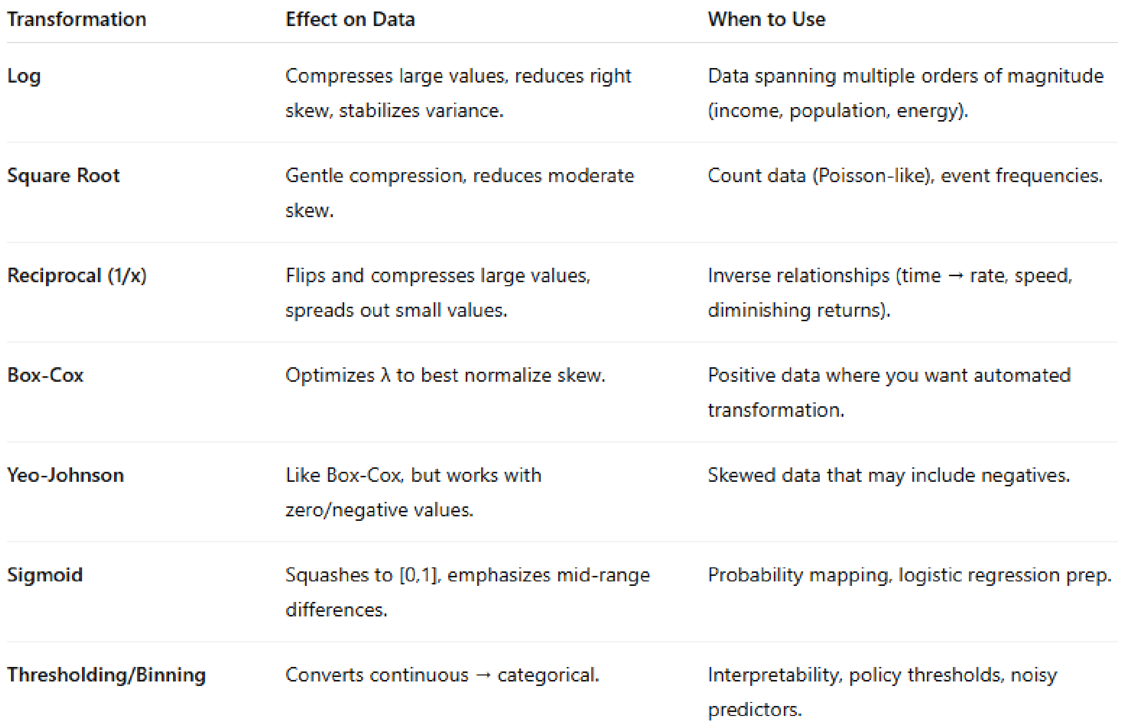
- Clear monotonic trend:
  - Q1 (lowest income): 73.6% default
  - Q2: 63.2% defaultQ3: 58.4% default
  - Q4 (highest income): 36.0% default
- Much stronger relationship than raw Pearson showed.



# Why to Threshold/Quantile

- Continuous Pearson correlation understated the relationship.
- Binning (quantiles) revealed a categorical-level trend.
- Good for:
  - Noisy continuous predictors with categorical outcomes.
  - When interpretability (“low vs high risk”) matters.
  - Use point-biserial or Chi-square after binning for stronger associations.
  - Determine appropriate thresholds

# Transformations



| Transformation | Effect on Data | When to Use |
|---|---|---|
| **Log** | Compresses large values, reduces right skew, stabilizes variance. | Data spanning multiple orders of magnitude (income, population, energy). |
| **Square Root** | Gentle compression, reduces moderate skew. | Count data (Poisson-like), event frequencies. |
| **Reciprocal (1/x)** | Flips and compresses large values, spreads out small values. | Inverse relationships (time → rate, speed, diminishing returns). |
| **Box-Cox** | Optimizes λ to best normalize skew. | Positive data where you want automated transformation. |
| **Yeo-Johnson** | Like Box-Cox, but works with zero/negative values. | Skewed data that may include negatives. |
| **Sigmoid** | Squashes to [0,1], emphasizes mid-range differences. | Probability mapping, logistic regression prep. |
| **Thresholding/Binning** | Converts continuous → categorical. | Interpretability, policy thresholds, noisy predictors. |

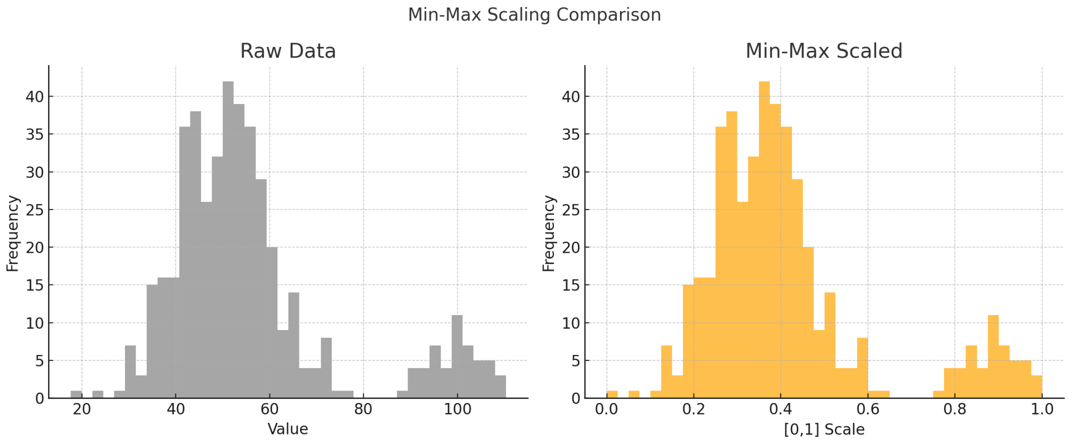
Transformations change the distributions of variables, use with caution!

# Why Normalize or Scale Data?

- Different features may be on very different scales (e.g., Income in $100,000s vs. Hours Worked in single digits)
- Distance-based methods (kNN, clustering) can be dominated by large-scale variables
- Gradient-based models (e.g., neural nets) converge faster on normalized data
- Visual comparability improves (heatmaps, scatterplots look balanced)
- Key idea: Normalization changes scale, not relationships (Pearson correlation is invariant under linear rescaling)

# Min-Max Scaling

- Rescales values to [0,1].
- Preserves shape of distribution, compresses outliers into narrow bands.
- Often used in neural networks and when absolute range matters.
- Histogram of income before and after min-max scaling (same shape, but squashed into [0,1]).

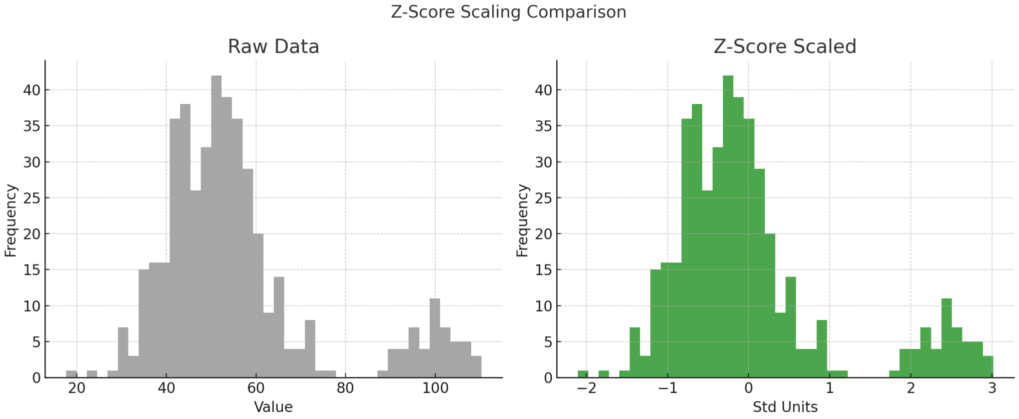


$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

What are the units?

# Z-Score Standardization

- Centers data at mean = 0, scales to standard deviation = 1.
- Distributions keep shape but are re-centered and rescaled.
- Useful for comparability when units differ (e.g., height vs weight).
- Visual: Histogram showing shift to mean 0, spread 1. Overlay normal curve.

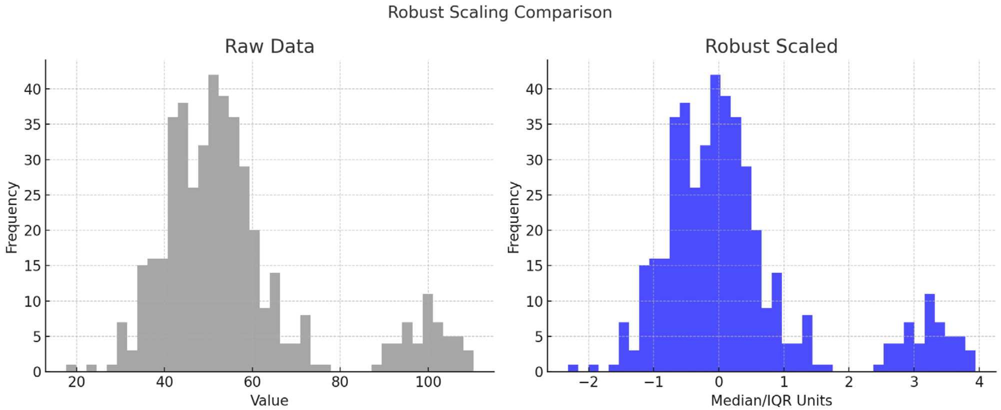


$$x' = \frac{x - \mu}{\sigma}$$

What are the units?

# Robust Scaling

- Uses median and IQR instead of mean and standard deviation.
- Resistant to outliers — “middle 50% of the data” defines the scale.
- Good for messy, heavy-tailed data.
- Boxplots before/after scaling with a few extreme outliers.



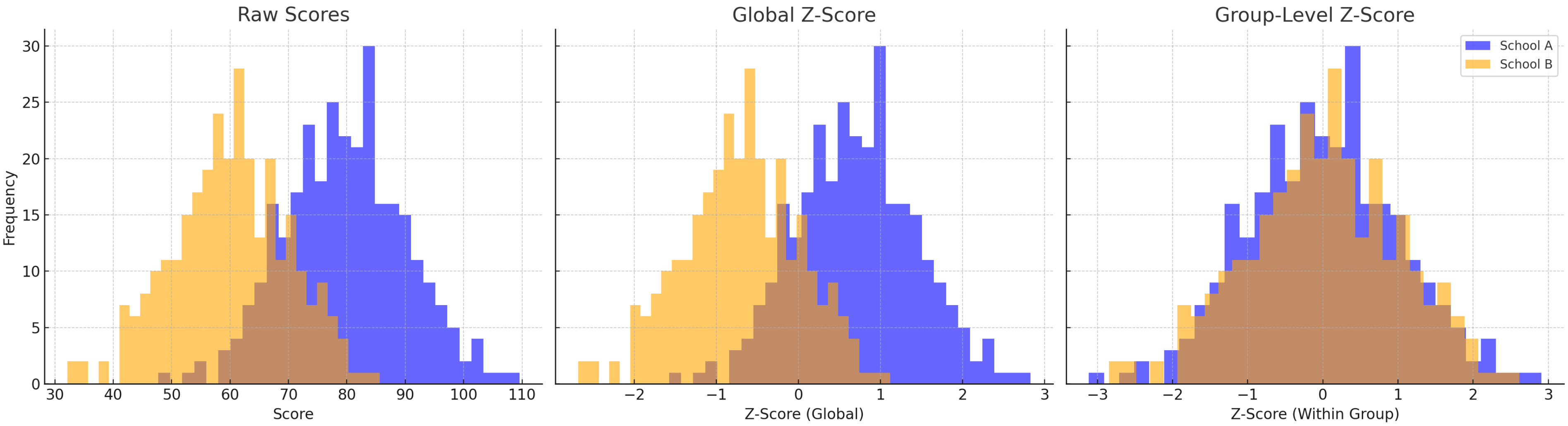
# BIG CAVEAT

- Global vs Group-Level Scaling

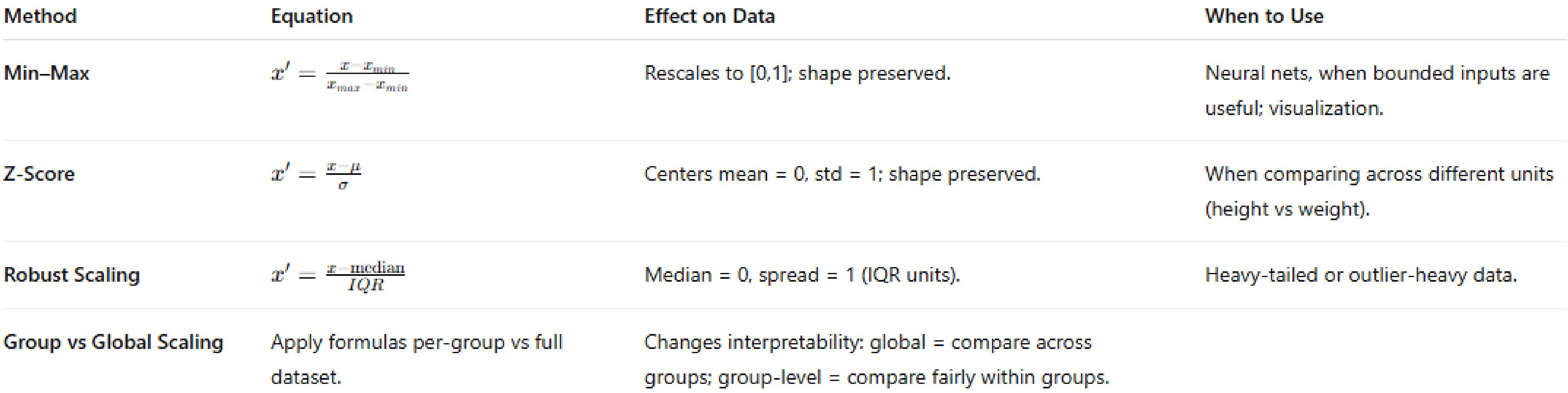
# Exam Scores Across Schools

- Suppose you’re comparing exam scores from two schools. School A has generally higher scores (mean ~80), while School B has lower scores (mean ~60).
- If you standardize globally (both schools together):
  - School A’s students will mostly appear above average.
  - School B’s students will mostly appear below average.
  - This is appropriate if your question is: ‘Who is stronger across the whole population?’
- If you standardize within each group (school-level z-score):
  - Each school’s mean becomes 0, and variance is scaled within that group.
  - School A and School B distributions now overlap perfectly.
  - This is appropriate if your question is: ‘Who is stronger relative to peers in their own school?’
- What questions would you be asking for both scenarios?
- REMEMEBER WHAT YOU DID AND WHY?!!?!?!

# Let’s walk through the distributions

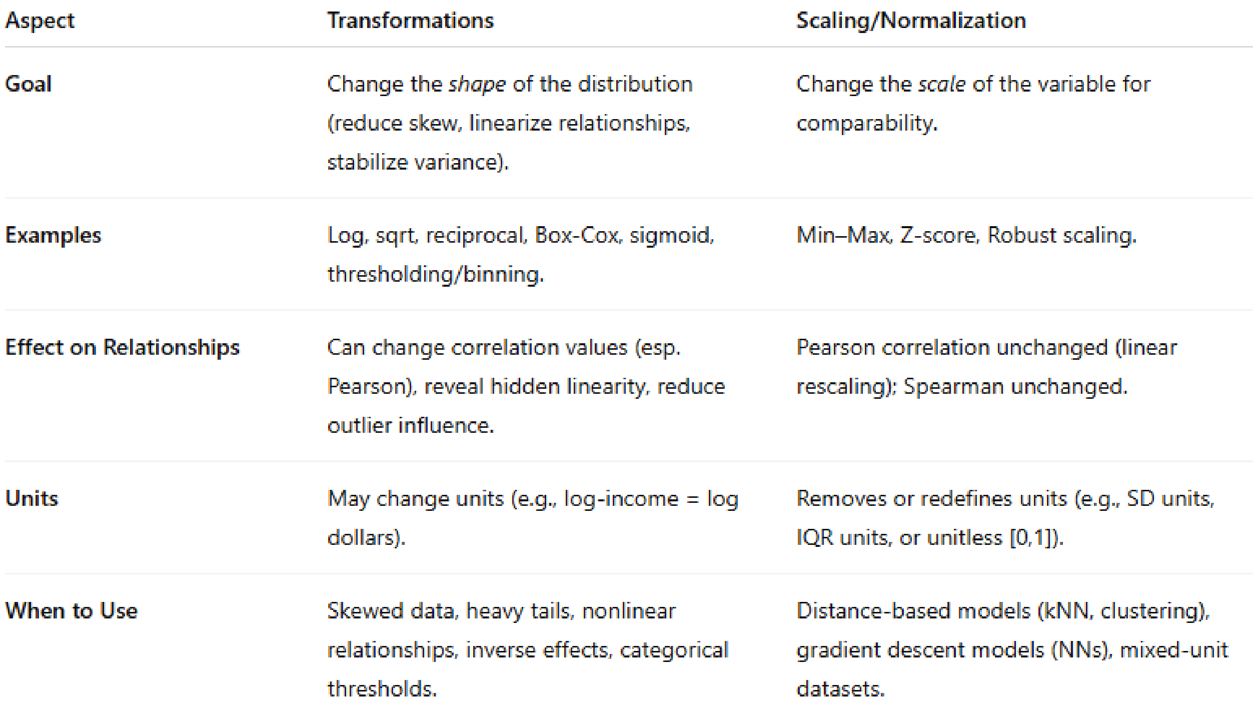


# Different Types of Scaling



| Method | Equation | Effect on Data | When to Use |
|---|---|---|---|
| **Min–Max** | $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$ | Rescales to [0,1]; shape preserved. | Neural nets, when bounded inputs are useful; visualization. |
| **Z-Score** | $x' = \frac{x - \mu}{\sigma}$ | Centers mean = 0, std = 1; shape preserved. | When comparing across different units (height vs weight). |
| **Robust Scaling** | $x' = \frac{x - \text{median}}{IQR}$ | Median = 0, spread = 1 (IQR units). | Heavy-tailed or outlier-heavy data. |
| **Group vs Global Scaling** | Apply formulas per-group vs full dataset. | Changes interpretability: global = compare across groups; group-level = compare fairly within groups. | |

# Transforms vs. Scaling



| Aspect | Transformations | Scaling/Normalization |
|---|---|---|
| **Goal** | Change the *shape* of the distribution (reduce skew, linearize relationships, stabilize variance). | Change the *scale* of the variable for comparability. |
| **Examples** | Log, sqrt, reciprocal, Box-Cox, sigmoid, thresholding/binning. | Min–Max, Z-score, Robust scaling. |
| **Effect on Relationships** | Can change correlation values (esp. Pearson), reveal hidden linearity, reduce outlier influence. | Pearson correlation unchanged (linear rescaling); Spearman unchanged. |
| **Units** | May change units (e.g., log-income = log dollars). | Removes or redefines units (e.g., SD units, IQR units, or unitless [0,1]). |
| **When to Use** | Skewed data, heavy tails, nonlinear relationships, inverse effects, categorical thresholds. | Distance-based models (kNN, clustering), gradient descent models (NNs), mixed-unit datasets. |