Data Normalization

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling

Min-Max Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]
  - Then \$73,000 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Z-score Normalization

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

- Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Normalization by Decimal Scaling

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

- **Example**:
- Data ranges from -986 to 917.
- Maximum absolute value is 986.
- Normalize by dividing by 1000 (since j=3).
- After normalization, -986 becomes -0.986 and 917 becomes 0.917.