

# Simple Discretization Methods: Binning

## 1. Equal-Width (Distance) Partitioning

Formula:

$$W = \frac{B - A}{N}$$

where:

- $W$  = Width of each interval
- $A$  = Minimum value in the dataset
- $B$  = Maximum value in the dataset
- $N$  = Number of bins

**Example:** Given the dataset:

$\{1, 3, 7, 10, 15, 18, 22, 25\}$

Let  $N = 3$ ,  $A = 1$ ,  $B = 25$ .

**Step 1: Compute Interval Width**

$$W = \frac{25 - 1}{3} = \frac{24}{3} = 8$$

**Step 2: Define Bins**

- **Bin 1:**  $[1, 9) \rightarrow$  contains  $\{1, 3, 7\}$
- **Bin 2:**  $[9, 17) \rightarrow$  contains  $\{10, 15\}$
- **Bin 3:**  $[17, 25] \rightarrow$  contains  $\{18, 22, 25\}$

This method is **simple and intuitive**, but **outliers** and **skewed data** can make it less effective.

## 2. Equal-Depth (Frequency) Partitioning

**Concept:** Each bin should have **approximately the same number of elements**.

**Example:** Given the dataset:

$\{1, 3, 7, 10, 15, 18, 22, 25\}$

Let  $N = 3$ . Since there are **8 elements**, each bin should have about  $\frac{8}{3} \approx 3$  elements.

**Step 1: Sort Data**

$\{1, 3, 7, 10, 15, 18, 22, 25\}$

**Step 2: Assign Bins**

- **Bin 1:**  $\{1, 3, 7\}$
- **Bin 2:**  $\{10, 15, 18\}$
- **Bin 3:**  $\{22, 25\}$

This method ensures **equal distribution** of data points in each bin, making it **better for skewed data**.

# Bin Processing Methods

## 1. Bin Means

**Concept:** Replace all values in a bin with the **mean** (average) of that bin.

**Example:** Using the same bins from equal-width partitioning:

- **Bin 1:**  $\{1, 3, 7\} \rightarrow \text{Mean} = \frac{1+3+7}{3} = 3.67 \rightarrow \{3.67, 3.67, 3.67\}$
- **Bin 2:**  $\{10, 15\} \rightarrow \text{Mean} = \frac{10+15}{2} = 12.5 \rightarrow \{12.5, 12.5\}$
- **Bin 3:**  $\{18, 22, 25\} \rightarrow \text{Mean} = \frac{18+22+25}{3} = 21.67 \rightarrow \{21.67, 21.67, 21.67\}$

**Pros:** Reduces variance. **Cons:** May hide important variations in data.

## 2. Bin Medians

**Concept:** Replace all values in a bin with the **median** of that bin. The **median** is the middle value when the numbers are sorted. If there is an even number of elements, the median is the average of the two middle values.

**Example:**

- **Bin 1:**  $\{1, 3, 7\} \rightarrow \text{Sorted: } \{1, 3, 7\}, \text{Median} = 3 \rightarrow \{3, 3, 3\}$
- **Bin 2:**  $\{10, 15\} \rightarrow \text{Sorted: } \{10, 15\}, \text{Median} = \frac{10+15}{2} = 12.5 \rightarrow \{12.5, 12.5\}$
- **Bin 3:**  $\{18, 22, 25\} \rightarrow \text{Sorted: } \{18, 22, 25\}, \text{Median} = 22 \rightarrow \{22, 22, 22\}$

**Pros:** Handles outliers better than means. **Cons:** May not capture overall distribution well.

## 3. Bin Boundaries

**Concept:** Replace each value with the nearest boundary value in the bin.

**Example:**

- **Bin 1:**  $\{1, 3, 7\} \rightarrow \text{Boundaries: } 1, 7 - 1 \rightarrow 1, 3 \rightarrow 1, 7 \rightarrow 7 \rightarrow \{1, 1, 7\}$
- **Bin 2:**  $\{10, 15\} \rightarrow \text{Boundaries: } 10, 15 - 10 \rightarrow 10, 15 \rightarrow 15 \rightarrow \{10, 15\}$
- **Bin 3:**  $\{18, 22, 25\} \rightarrow \text{Boundaries: } 18, 25 - 18 \rightarrow 18, 22 \rightarrow 25, 25 \rightarrow 25 \rightarrow \{18, 25, 25\}$

**Pros:** Good for preserving extreme values. **Cons:** Can distort middle-range values.

# Data Transformation: Normalization

## 1. Min-Max Normalization

**Formula:**

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

where:

- $v$  = Original value
- $\min_A$  = Minimum value of the original dataset
- $\max_A$  = Maximum value of the original dataset
- $\text{new\_min}_A$  = Minimum value of the new range
- $\text{new\_max}_A$  = Maximum value of the new range
- $v'$  = Normalized value

**Example:** Convert age = 30 to a range [0, 1], where min = 10 and max = 80:

$$v' = \frac{30 - 10}{80 - 10} = \frac{20}{70} = \frac{2}{7} \approx 0.2857$$

This transformation scales the value to fall between the new range, in this case, between 0 and 1.

## 2. Z-Score Normalization

**Formula:**

$$v' = \frac{v - \text{mean}_A}{\text{stand\_dev}_A}$$

where:

- $v$  = Original value
- $\text{mean}_A$  = Mean of the original dataset

$$\text{mean}_A = \frac{1}{n} \sum_{i=1}^n x_i$$

- $\text{stand\_dev}_A$  = Standard deviation of the original dataset

$$\text{Population: } \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \text{mean}_A)^2}$$

$$\text{Sample: } s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \text{mean}_A)^2}$$

- $v'$  = Normalized value

The Z-score normalization transforms the data to have a **mean of 0** and a **standard deviation of 1**, making it useful for comparing data from different distributions.

## 3. Normalization by Decimal Scaling

**Formula:**

$$v' = \frac{v}{10^j}$$

where  $j$  is the smallest integer such that:

$$\text{Max}(|v'|) < 1$$

This transformation scales the data by dividing each value by a power of 10, where the exponent  $j$  is chosen to ensure the maximum absolute value of  $v'$  is less than 1.

# Five-Number Summary

## 1. Even-Length Datasets

**Formula:**

Five-number summary = (Minimum,  $Q1$ , Median ( $Q2$ ),  $Q3$ , Maximum)

where:

- $Q1$  = Median of the first half
- $Q2$  = Overall median (average of middle two values)
- $Q3$  = Median of the second half

**Example 1:** Given the dataset: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

- **Minimum:** 1
- **Maximum:** 10
- **Median ( $Q2$ ):** Average of 5th and 6th values =  $\frac{5+6}{2} = 5.5$
- **$Q1$ :** Median of [1, 2, 3, 4, 5] = 3
- **$Q3$ :** Median of [6, 7, 8, 9, 10] = 8

**Five-number summary:** (1, 3, 5.5, 8, 10)

**Example 2:** Given the dataset: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

- **Minimum:** 1
- **Maximum:** 12
- **Median ( $Q2$ ):** Average of 6th and 7th values =  $\frac{6+7}{2} = 6.5$
- **$Q1$ :** Median of [1, 2, 3, 4, 5, 6] =  $\frac{3+4}{2} = 3.5$
- **$Q3$ :** Median of [7, 8, 9, 10, 11, 12] =  $\frac{9+10}{2} = 9.5$

**Five-number summary:** (1, 3.5, 6.5, 9.5, 12)

---

## 2. Odd-Length Datasets

**Example 3:** Given the dataset: [1, 2, 3, 4, 5, 6, 7, 8, 9]

- **Minimum:** 1
- **Maximum:** 9
- **Median ( $Q2$ ):** 5th value = 5
- **$Q1$ :** Median of [1, 2, 3, 4] =  $\frac{2+3}{2} = 2.5$
- **$Q3$ :** Median of [6, 7, 8, 9] =  $\frac{7+8}{2} = 7.5$

**Five-number summary:** (1, 2.5, 5, 7.5, 9)

**Example 4:** Given the dataset: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]

- **Minimum:** 1
- **Maximum:** 11
- **Median ( $Q2$ ):** 6th value = 6
- **$Q1$ :** Median of [1, 2, 3, 4, 5] = 3
- **$Q3$ :** Median of [7, 8, 9, 10, 11] = 9

**Five-number summary:** (1, 3, 6, 9, 11)