Data Cleaning

Problem Solving - 2

Data Normalization

- Min-max Normalization
- Z-Score Normalization
- Decimal Scale Normalization

Min-max Normalization

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v'_i in the range $[new\_min_A, new\_max_A]$ by computing

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \quad (3.8)$$

Example

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range $[0.0, 1.0]$. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$. ■

Z-score Normalization

In **z-score normalization** (or *zero-mean normalization*), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \tag{3.9}$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A. The

Example

z-score normalization. Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600 - 54,000}{16,000} = 1.225$. ■

Decimal Scale Normalization

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i}{10^j}, \tag{3.12}$$

where j is the smallest integer such that $max(|v_i'|) < 1$.

Example

Decimal scaling. Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that -986 normalizes to -0.986 and 917 normalizes to 0.917. ■

Exercise

Use these methods to normalize the following group of data:

200, 300, 400, 600,1000

(a) min-max normalization by setting min = 0 and max = 1

(b) z-score normalization

(c) normalization by decimal scaling

Answers Min-max Normalization

Given data is A = **200, 300, 400, 600,1000**

New_minA = 0
New_maxA = 1

minA = 200
maxA = 1000

Formula:

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

$$V_i = 200$$

$$V'_i = \frac{(200-200)}{(1000-200)} \times (1 - 0) + 0$$

Answers z-score Normalization

Given data is A = **200, 300, 400, 600,1000**

Formula:

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

Mean ($\bar{A}$) = 500

Standard Deviation (σ_A) = 316.22

$V_i = 200$

$$V'_i = \frac{(200-500)}{(316.22)} = -0.94$$

Answers Decimal Scale Normalization

Given data is A = **200, 300, 400, 600,1000**

Formula: $$v_i' = \frac{v_i}{10^j}$$

$V_i = 200$

$$V'_i = \frac{(200)}{(1000)} = 0.2$$

Histograms

- Histograms use **binning** to approximate data distributions
- a popular form of data reduction
- partitions the data distribution of an attribute A into disjoint subsets, referred to as **buckets** or **bins**.
- Buckets are 3 types
  1. Singleton
  2. Equal Width
  3. Equal frequency

Histograms. The following data are a list of *AllElectronics* prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Example

Divide the data into **Equal width** and

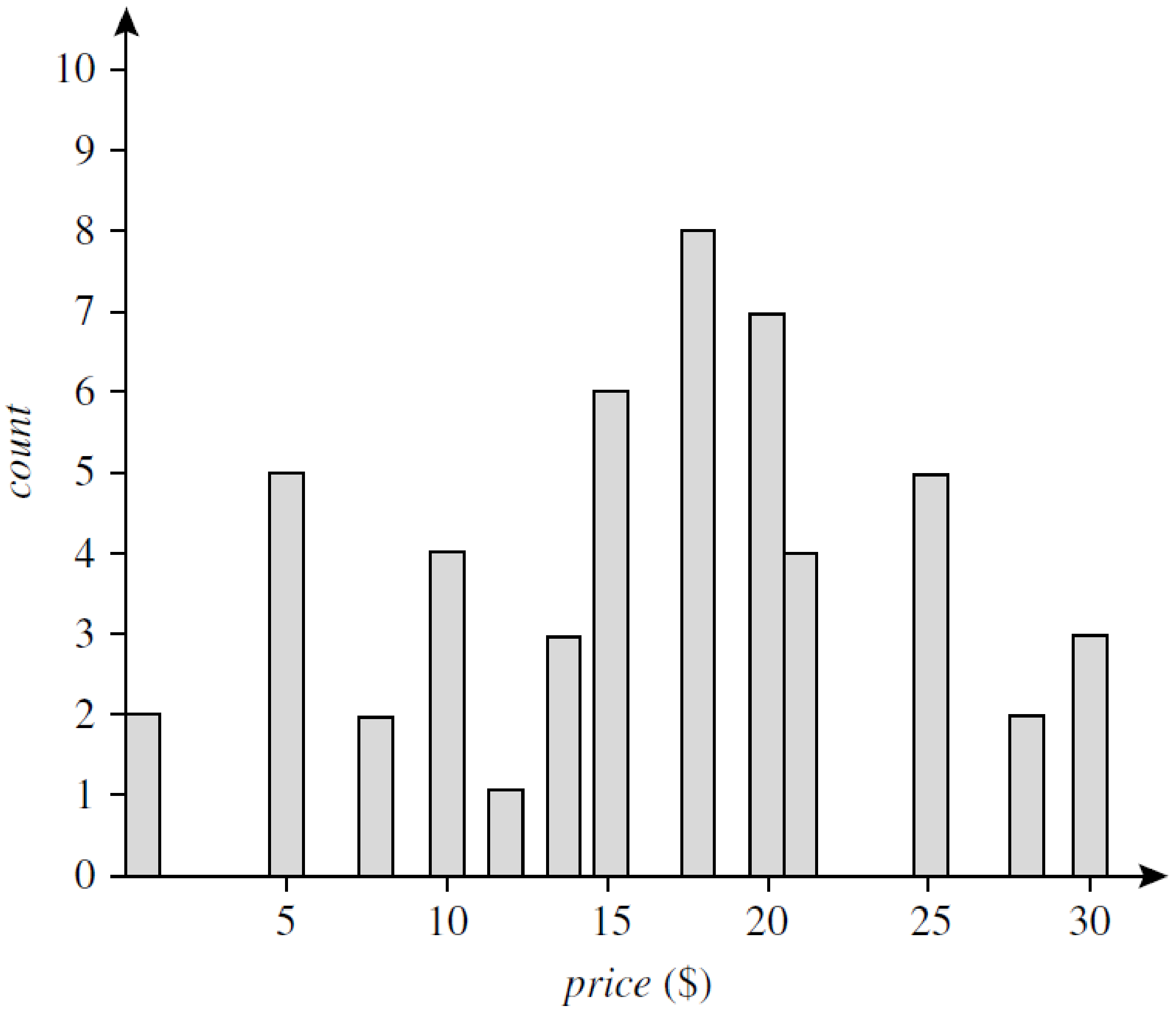
Equal Frequency Bins and

Draw the Histograms

Answers Singleton Histograms

Singleton Histogram

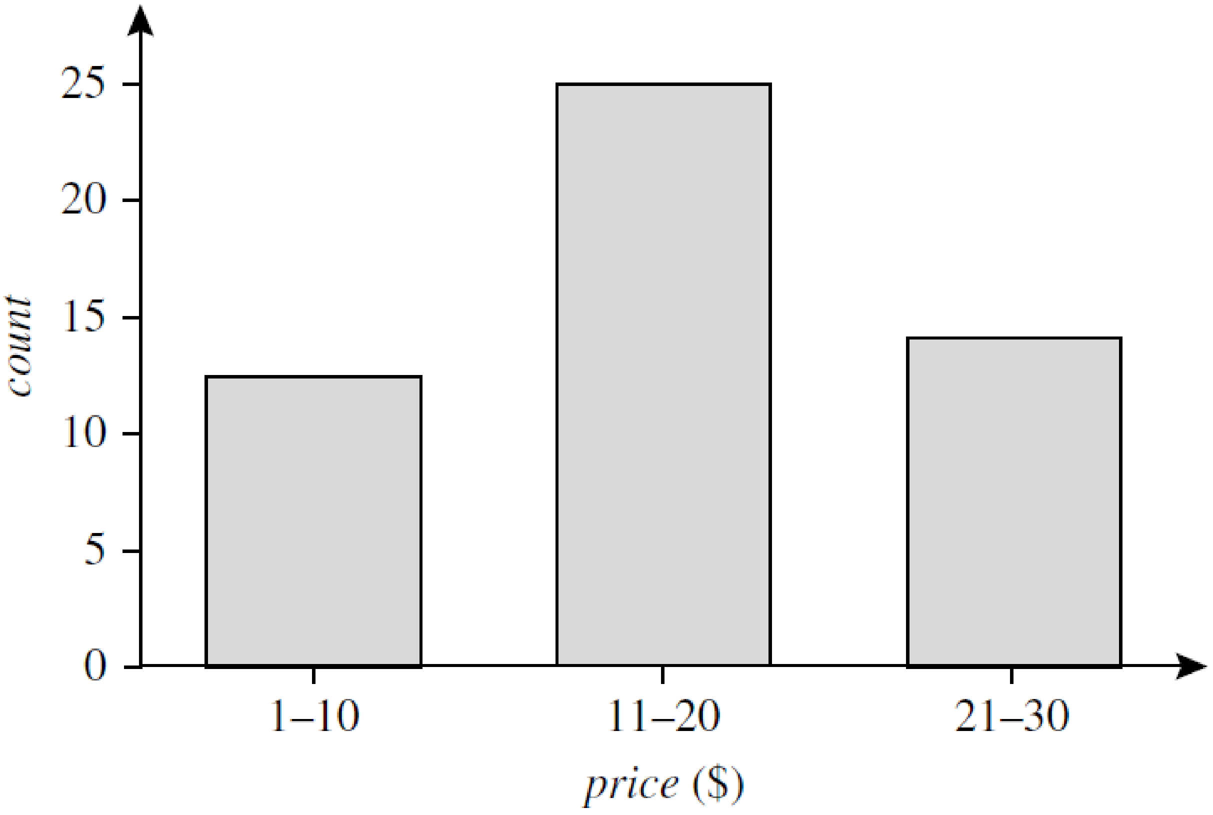
- Find the frequency of each value and draw the Histogram



Answers Equal Width Histograms

Equal-width Histogram

- Here the width of each bucket range is uniform (e.g., the width of $10 for the buckets in the below figure)



Equal-frequency (or equal-depth) Histograms

Answers
Equal Frequency
Histograms

- Here, the buckets are created so that, roughly, the frequency (depth) of each bucket is **constant**
- Refer the Binning techniques for dividing the elements in to equal depth and applying smoothing techniques.
- After applying smoothing techniques, find frequency of each element and draw the histogram

Discretization

- Here the raw values of a numeric attribute are replaced by interval variables or conceptual labels

- Ex: *age values* can be replaced by

  interval labels 0–10, 11–20, etc.

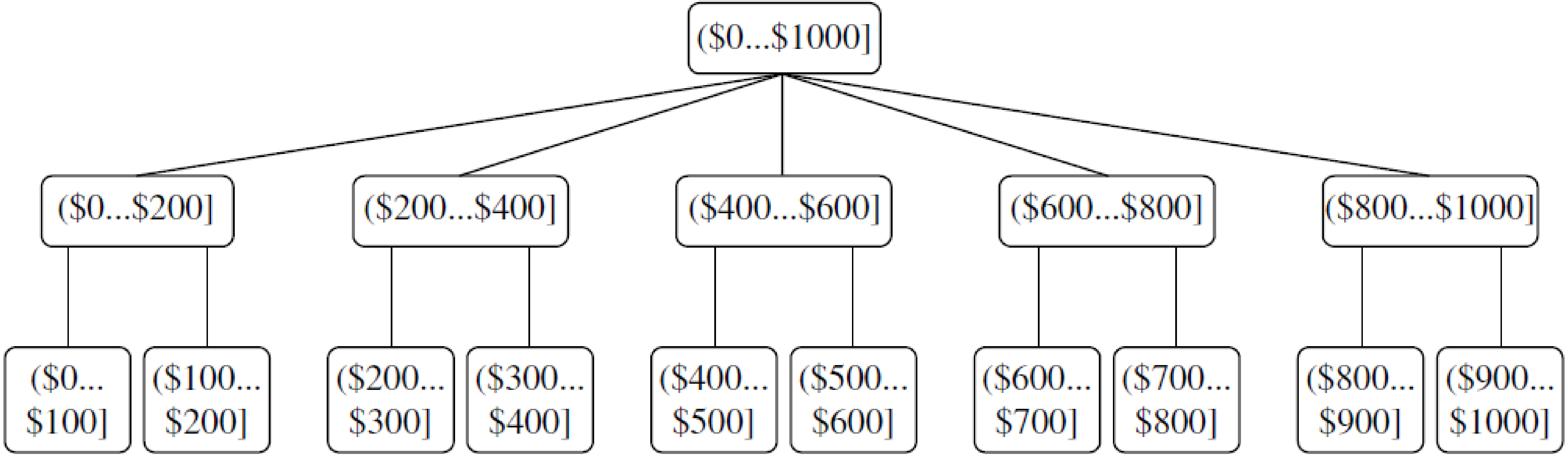
  or

  conceptual labels (e.g., *youth*, *adult*, *senior*)

Concept Hierarchy Generation

- The labels of an attribute can be recursively organized into higher-level concepts, resulting in a ***concept hierarchy*** for the numeric attribute.

- Ex: Price values of various items can be organized into concept hierarchy for easy analysis



A concept hierarchy for the attribute *price*, where an interval ($\$X \ldots \$Y$] denotes the range from $\$X$ (exclusive) to $\$Y$ (inclusive).