Data Transformation by Normalization :→

The Measurement unit used can affect the data Analysis. For ex:- Changing Measurement units from meters to inches for height, or from Kilograms to Pounds for weight, may lead to very different results. In General, expressing an attribute in smaller units will lead to larger range for that attribute, and thus tend to give such an attribute greater effect or "weight" to. To avoid dependence on the choice of M/s units, the data should be normalized or standardized. this involves transforming the data to fall within a smaller or common range such as [-1,1] or [0.0, 1.0].

Normalization of data attempts to give all attribute an equal weight. Normalization is particularly useful for classification Algo. Involving neural networks or distance Measurements such as nearest neighbor classification and clustering. if using the neural network back Propagation algorithm for classification Mining, normalizing the I/P values for each attribute measured in the Training tuples will help speed up the learning Phase.

For distance based Methods, normalization helps prevent attributes with initially large ranges (e.g Income) from outweighing attributes with initially smaller ranges (e.g binary attributes) it is also useful when given no Prior Knowledge of data.

There are many Methods for data normalization. We study min max normalization, z-score normalization, and normalization by decimal scaling. For our discussion, let A be a numeric attribute with n observed values, $v_1, v_2 \dots v_n$

Min-Max normalization:→ it performs a linear transformation on the original data. Suppose 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

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

min-max normalization preserves the relationship among the original data values. it will encounter an "out of bounds" error if a future I/P case for normalization falls outside of the original data range for A.

Ex:→ 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

$$\text{Transformed to } \frac{73600 - 12000}{98000 - 12000}(1.0 - 0) + 0 = 0.716$$

in z-score normalization (or zero Mean normalization), the value for an attribute, A, are normalized based on the mean and SD of A. A value, v_i of A is normalized to v_i by computing

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

where $\bar{A}$ and σ_A are the Mean of Standard deviation (SD), respectively of Attribute A. this Method of normalization is useful when the actual Minimum and Maximum of attribute A are (unknown), or when there are outliers that are dominate the min-max normalization.

Ex:- Z-Score normalization:- Suppose that the Mean and SD of the values for the attribute Income are $54000 and $16000, respectively. with Z-Score normalization, a value of 73600 $ for income is transformed to $\frac{73600-54000}{16000} = 1.225$

The Mean absolute deviation of A

$$S_A = \frac{1}{n}\left((|v_1-\bar{A}| + |(v_2-\bar{A})| + (|v_3-\bar{A}) + \cdots + |v_n-\bar{A}|\right)$$

Thus Z-Score normalization using the Mean absolute deviation is $v_i' = \frac{v_i-\bar{A}}{S_A}$

The Mean Absolute deviation, S_A is more robust to outlier than the SD, σ_A, Here effects of outlier are ~~more~~ somewhat reduced because the deviation from the Mean (i.e $|x_i-\bar{x}|$) are not squred.

Normalization by decimal Scaling :→ it normalizes by moving the decimal Point of values of attribute A. The number of decimal Points Depends on the maximum absolute value of A. A value v_i, of A is normalized to v_i' by computing.

$$v_i' = \frac{v_i}{10^j}$$ where j is the smallest integer such that $\max(|v_i'|) < 1$

Suppose that a recorded values of A ranges 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

Question:- Use these Methods to normalize the following group of data:

200, 300, 400, 600, 1000

(a) min-max normalization by setting min=0, max=1
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation.
(d) normalization by decimal scaling.

Sol:- (1) Min max normalization

$$v_i' = \frac{v_i - \min A}{\max A - \min A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}$$

$\min A = 200$, $\max A = 1000$

$$\bar{A} = \frac{200+300+400+600+1000}{5} = \frac{2500}{5} = 500$$

$v_1' = 0$

$v_2' = \frac{300-200}{1000-200}(1-0)+0 \Rightarrow \frac{100}{800} = 0.125$

$v_3' = \frac{400-200}{800} = 200/800 = 0.25$

$v_4' = \frac{600-200}{800}(1)+0 \Rightarrow \frac{400}{800} = 0.5$

$v_5 = \frac{1000-200}{800}(1)+0 = \frac{800}{800} = 1$

and soon.

② Z-Score Normalization :-

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

$$\sigma_A = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2} \Rightarrow \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$\bar{A} = 500$

$$\sigma_A = \frac{1}{5}\left(200^2 + 300^2 + 400^2 + 600^2 + 1000^2\right) - 500^2$$

$$\Rightarrow \frac{1}{5}\left(40000 + 90000 + 160000 + 360000 + 1000000\right) - 250000$$

$$\Rightarrow \frac{1}{5}\left(1650000\right) - 250000$$

29
136
165

$$\Rightarrow 330000 - 250000$$

$$\Rightarrow \sqrt{80000} \Rightarrow 282.84$$

(b) $v_4' = \frac{600 - 500}{\sqrt{80000}} = \frac{100}{80000} = \frac{100}{282.84} = 0.353$

③ Mean absolute deviation

$$s_A = \frac{1}{N}\left(|200 - 500| + |300 - 500| + |400 - 500| + |600 - 500| + |1000 - 500|\right)$$

$$\Rightarrow \frac{1}{N}\left((300 + 200 + 100 + 100 + 500)\right)$$

$$\Rightarrow \frac{1}{5}(1200) \Rightarrow 240$$

$$v_4' = \frac{600 - 500}{240} \Rightarrow \frac{100}{240} = 0.416$$

(4) Normalization by decimal Scaling :→

$$v_i' = \frac{600}{10^3} \Rightarrow \underline{0.6}$$

Graphic Displays of Basic Statistical Descriptions of data:-

These include quantile plots, quantile-quantile plots, Histograms and Scatter plots. Such graphs are helpful for visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distribution (i.e data for one attribute), while scatter plots show bivariate distribution (i.e involving two attributes).

Histograms:-

Histograms (or frequency histograms) are atleast a century old and widely used method. "Histos" means Pole or Mast, and "gram" means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summerizing the distribution of a given attribute, X. If X is nominal, such as automobile_model or item_type, Then a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e. count) of that x-value.

The resulting graph is more commonly known as a bar chart.

If X is numeric, the term histogram is preferred. The range of value for X is partitioned into disjoint consecutive subranges. The subranges, referred to as bucket or bins, are disjoint subsets of the data distribution for X. The range of a bucket is known as width. Typically, the bucket are of equal width.

for Example, a Price attribute with a value range of $1 to $200 can be partitioned into subranges 1 to 20, 21 to 40, 40 to 60

For each subrange, a bar is drawn with a height that represents the total count of items observed within the subrange.

Table 1:— A set of unit price data for items sold at a branch of AllElectronics

Unit Price $	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

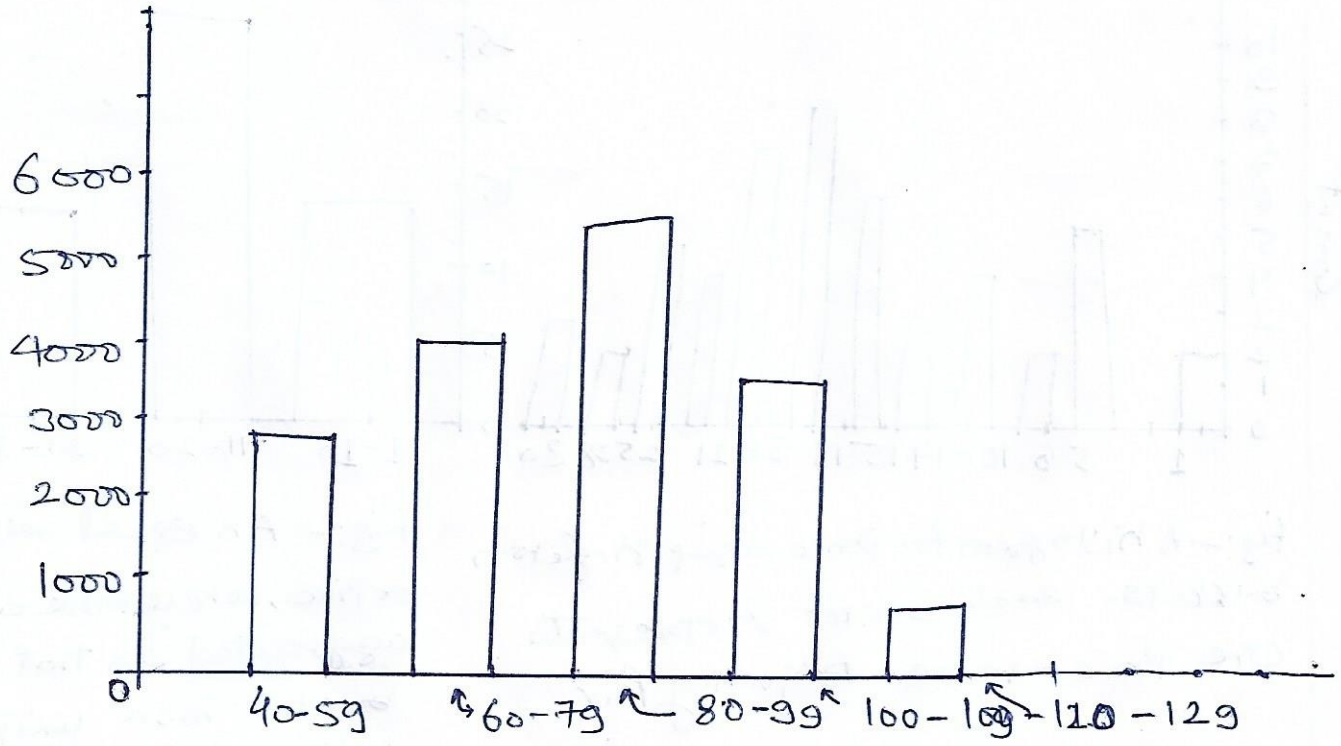


fig:- A histogram for above data table.

Histograms use binning to approximate data distributions and are a popular form of data reduction. A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets referred to as bucket or bins. If each bucket represents only a single attribute-value/frequency pair, the bucket are called singleton buckets, bucket instead represents continuous ranges for the given attribute.

for Ex:- 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, 21,21,21,21, 25,25,25,25,25, 28,28, 30,30,30.

Fig (a) shows a histogram for the data using singleton buckets. To further reduce the data, it is common to have each bucket denote a continuous value range for the given attribute. in Fig (b) each bucket represents a different 10$ range of Price.

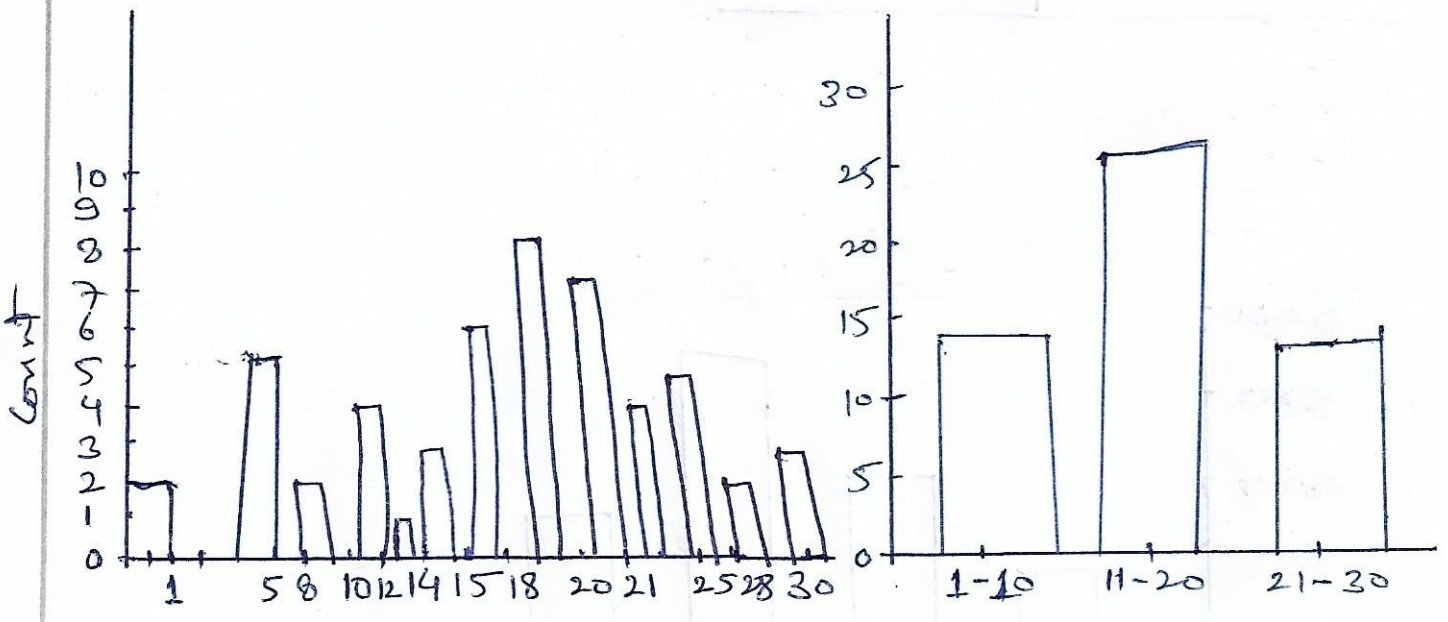


Fig:- A histogram for Price using singleton buckets- each bucket represents one Price-value frequency pair

Fig:- An equal width histogram for Price, where value are aggregated so that each bucket has a uniform width of $10.

Scatter plots and Data Correlation :→ A Scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, Pattern, or trend between two numeric attributes. To construct a scatter plot each pair of values is treated as a pair of co-ordinates in an algebraic sense and plotted as Points in the plane. Fig(a) shows a scatter plot for the set of data in table.

The Scatter plot is a useful method for providing a first look at bivariate data to see clusters of Points and outliers, or to explore the possibility of correlation relationship.

Two attributes, X, and Y are correlated if one attribute implies the other. Correlations can be +ve, -ve or null (uncorrelated).

Fig (b) shows examples of +ve and -ve correlations b/w two attributes. If the plotted Points Pattern slopes from lower left to upper right, this means that the value of X increase as the values of Y increase, suggesting a Positive correlation b(i). If the pattern of plotted Points slopes from upper left to lower right, the values of X increase as the value of Y decrease, suggesting a -ve correlation (Fig b(ii))

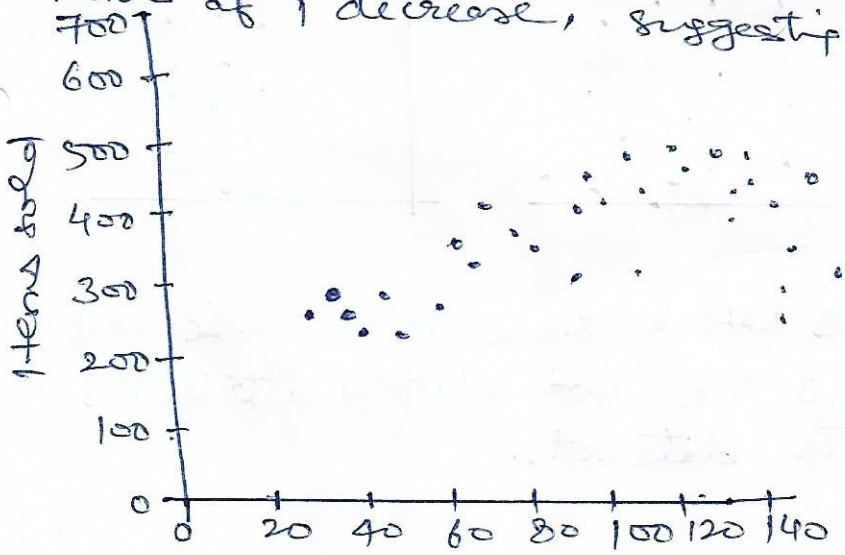
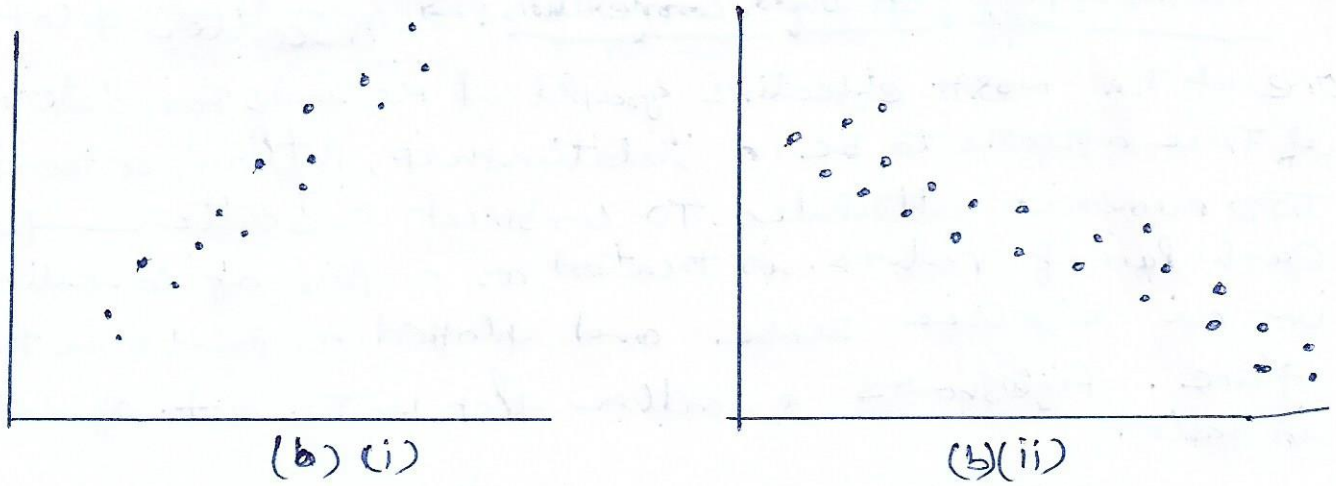


Fig (a) : A histogram for the table.



(b) (i) (b)(ii)

fig:- Scatter plots can be used to find (a) positive or (b) Negative Correlations b/w attributes

A line of best fit can be drawn to study the correlation between the variables.

fig (C) shows three cases for which there is no correlation / relationship between the two attributes in each of the given data sets.

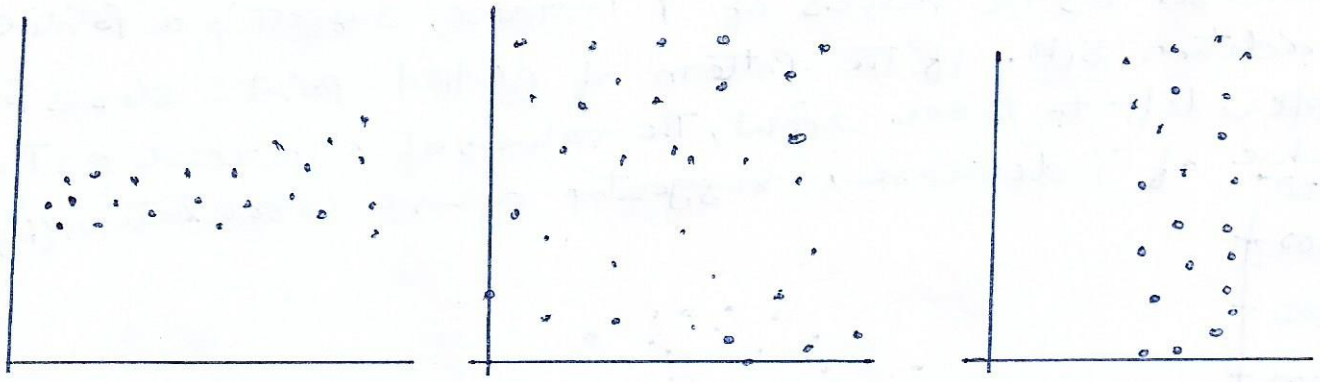


fig:- Three cases where there is no observed co-relation b/w the two plotted attributes in each of the datasets.