NAME :- KRUNAL RANK

ROLL NO :- U18CO081

CLASS :- BTECH 4TH YEAR

SEMESTER :- 7

DIVISION :- B

Data Warehousing and Data Mining
Tutorial 4

Ans 1:- A scatterplot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis and other variable appears on vertical axis. Each individual in the data appears as a point on graph.

A scatterplot helps in identifying the relation between the used two attributes and roughly helps us determine whether the two variables are related or not.

Ans 2:- Data transformation is required even after data preprocessing as it offers the following benefits :-

- Data is transformed to make it better organised.
- Properly formatted and validated data improves data quality and supports compatibility, integration and indexing.
- Transformed data can be used for multiple purposes that are not yet intended but can be thought of in near future.

Ans 3:- Data Smoothing is required as :-

- it reduces noise and irregularities
- it allows human and computer analysis to be done in a better manner.
- it allows us to identify patterns
- it allows us to predict trends.

Data smoothing by binning can be done in one of the following ways:-

- After designated each record to a specific bin, the values of the bin can be replaced ~~with~~ by mean value of that particular bin.
- The values can also be replaced by boundary values based on which boundary value is nearer.
- The values can be replaced by median if the data is skewed ~~in~~ as a whole.

Ans 4:-

a) Equal frequency binning:-

bin 1 :- 5, 10, 11, 13

bin 2 :- 15, 35, 50, 55

bin 3 :- 72, 92, 204, 215

b) Equal width binning :-

Interval size = $\frac{215 - 5}{3} = 70$ Intervals :- [5, 75], [75, 145], [145, 215]

bin 1 :- 5, 10, 11, 13, 15, 35, 50, 55, 72

bin 2 :- 92

bin 3 :- 204, 215

c) Clustering :-
bin 1 :- 5, 10, 11, 13, 15, 35,
bin 2 :- 50, 55, 72, 92
bin 3 :- 204, 215

Ans 5 :- a) Using bin depth 3,

	Mean Smoothing
bin 1 :- 13, 15, 16	14.66, 14.66, 14.66
bin 2 :- 16, 19, 20	18.33, 18.33, 18.33
bin 3 :- 20, 21, 22	21, 21, 21
bin 4 :- 22, 25, 25	24, 24, 24
bin 5 :- 25, 25, 30	26.66, 26.66, 26.66
bin 6 :- 33, 33, 35	23.66, 23.66, 23.66
bin 7 :- 35, 35, 35	35, 35, 35
bin 8 :- 36, 40, 45	40.33, 40.33, 40.33
bin 9 :- 46, 52, 70	56, 56, 56

b) Outliers can be detected using box plot of the given data. Inter Quartile range can be used to identify outliers as well.

c) Some other methods to smoothen out data are :-
→ Boundary smoothing
→ Exponential smoothing
→ Random walk
→ Replacing by median

Ans 6 :-
a) M
M
b)
c)
d) F
m
10

Ans 6.:

a) Max = 70
Min = 13

Normalised value for 35 = $\frac{35-13}{70-13}$

$= \frac{22}{57} = 0.38596$

b) Given, $\sigma = 12.94$

$\mu = 809/27 = 29.962$

Normalised value for 35 = $\frac{35-\mu}{\sigma} = \frac{35-29.962}{(12.94)}$

$= 0.3892$

c) Normalisation by decimal scaling = $\frac{35}{10^2} = 0.35$

d) For the given data, normalisation by decimal scaling is more suitable because data contains integral values between 10 and 100 and there is no problem of vanishing values.