08/01/2025

DMW

2310080078
Mythri

Compare Data Information and Knowledge.

DATA	INFORMATION	KNOWLEDGE
* Data is collection of raw facts or figures.	* Data that has been processed, organized and structured.	* Knowledge is derived from analysing Information.

Differentiate Nominal and Ordinal attributes with example.

NOMINAL	ORDINAL
* Distinct and unordered categories	* Distinct categories with a meaningful order.
* No inherent order or ranking.	* Natural hierarchy or sequence among categories
* Qualitative and descriptive without numerical significance.	* Qualitative with ranking order, but not equal intervals.
* chi-square test is statistical approach.	* Kruskal-Wallis test
ex: Gender classifications	ex: Feedback methods

2310080078

Consider the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

i) Find the mean of the data.

$$(\text{Mean})\ \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

$$\bar{x} = \frac{809}{27} = 29.96.$$

ii) Determine the median.

$N = 27$ - odd

$$\text{Median} = \frac{n+1}{2} = \frac{28}{2} = 14$$

14th term 25.

2310080078

iii) Find the mode of the data. Comment on the data's modality (i.e., bimodal, trimodal, etc.).

Mode - most repeated values

25, 35.

Modality - Bimodal.

iv) Calculate the range and midrange of the data.

Range = $\cancel{\frac{70 - 13}{2}}$ = 57

Range = (Max value - Min value)

Midrange = $\frac{\text{Max value + Mid value}}{2}$

$= \frac{70 + 13}{2} = \frac{83}{2}$

$= 41.5$.

v) Find Q_1, Q_2, Q_3 and IQR values of the above data.

Q_2 = median of data set

$Q_2 = 25$

Q_1 = mid of first 13 terms

7th term is 20

$Q_1 = 20$,

$Q_3 =$ mid to 2^{nd} 13 terms

7^{th} term of 2^{nd} half. is 35

$Q_3 = 35$

$IQR = Q_3 - Q_1$

$= 35 - 20$

$= 15$

(vi) Find the five-number summary of the given data.

Minimum	— 13
Q_1	— 20
Median	— 25
Q_3	— 35
Maximum	— 70

Define Noisy data. Describe Binning methods with the following data

13, 15, 16, 16, 19, 20, 20, 21, 22, 22,

25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35,

36, 40, 45, 46, 52, 70.

2310080078

Noisy data: Noise is a random error or variance in a measured variable.

Bin size = 9

Partitions:

Bin 1: 13, 15, 16, 16, 19, 20, 20, 21, 22

Bin 2: 22, 25, 25, 25, 25, 30, 33, 33, 35

Bin3: 35, 35, 35, 36, 40, 45, 46, 52, 70

Smoothing by bin Means:

Bin1 Mean: ~~162~~ 162/9 = 18

Bin2 Mean: 253/9 = 28.1

Bin3 Mean: 394/9 = 43.7

Bin1: 18, 18, 18, 18, 18, 18, 18, 18, 18

Bin2: 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1

Bin3: 43.7, 43.7, 43.7, 43.7, 43.7, 43.7, 43.7, 43.7, 43.7.

Smoothing by bin Median:

Bin1 Median: 5th term 19

Bin2 Median: 25

Bin3 Median: 40

Bin 1: 19, 19, 19, 19, 19, 19, 19, 19, 19

Bin 2: 25, 25, 25, 25, 25, 25, 25, 25, 25

Bin 3: 40, 40, 40, 40, 40, 40, 40, 40, 40

Smoothing by bin Boundaries :

Bin 1: 13, 13, 13, 13, 22, 22, 22, 22, 22

Bin 2: 22, 22, 22, 22, 22, 35, 35, 35, 35

Bin 3: 35, 35, 35, 35, 35, 35, 35, 35, 70

Define Data Cleaning. Describe the methods to fill the missing values for an attribute in date cleaning.

Data Cleaning attempts to fill in missing values, smooth out noisy values while identifying outliers and correcting inconsistencies in the data.

METHODS TO FILL THE MISSING VALUES FOR AN ATTRIBUTE :

* Ignore the tuple.

- * Fill in the missing value manually
- * Use a global constant to fill in the missing value
- * Use a measure of central tendency for the attribute (eg., the mean or median) to fill in the missing value.
- * Use the attribute mean or median for all samples belonging to the same class as the given tuple.
- * Use the most probable value to fill the missing value,

Define Data Integration and mention the issues during data integration.

Data integration combines data from multiple sources into a unified view, ensuring consistency and usability for analysis or reporting.

ISSUES DURING DATA INTEGRATION 2310080078

* Schema Integration
* Entity Resolution
* Data Redundancy
* Semantic Conflicts
* Data Quality Issues
* Heterogeneous Formats
* Scalability
* Timeliness
* Privacy and Security
* Source Availability.

List Graphic displays of basic statistical descriptions of data.

The graphical displays help for visual inspection of data useful for pre-processing.

* Quantile plots
* Quantile - Quantile plots
* Histograms
* Scatter plots
* Box plots

2310080078

Suppose that a hospital test the age and body fat for 18 randomly selected adults with following results:

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Calculate the mean, median and standard deviation of age and % fat.

$$\text{Mean} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_n}{N}$$

mean age = 836/18 = 46.44

mean % fat = 518.1/18 = 28.7833

Median = n/2 th term = 18/2 = 9th term

median age = 50

median % fat = 31.2

$$\text{standard deviation} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

2310080078

age	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
23	-23.44	549.4.
23	-23,44	549.4
27	-19.44	377.9
27	-19.44	377.9
39	-7.44	55.35
41	-5.44	29.59
47	0.56	0.31
49	2.56	6.55
50	3.56	12.67
52	5.56	30.91
54	7.56	57.1
54	7.56	57.1
56	9.56	91.39
57	10.56	111.51
58	11.56	133.63
58	11.56	133.63
60	13.56	183.87
61	14.56	211.99
		2970.2

$$\text{age} = \sqrt{2970.2/18} = \sqrt{165.01}$$

standard deviation of age = 12.84

2310080078

% fat	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
9.5	-19.28	372.72
26.5	-2.28	5.20
7.8	-20.98	440.14
17.8	-10.98	120.64
31.4	2.62	6.86
25.9	-2.88	8.29
27.4	-1.38	1.90
27.2	-1.58	2.50
31.2	2.42	5.86
34.6	5.82	33.88
42.5	13.72	188.20
28.8	0.02	0.00
33.4	4.62	21.36
30.2	1.42	2.02
34.1	5.32	28.30
32.9	4.12	16.98
41.2	12.42	154.28
35.7	6.92	47.92
		1457.05

$$fat = \sqrt{1457.05/18} = \sqrt{80.9472}$$

standard deviation of fat = 8.99

2310080078

Define binary and nominal variables.

BINARY	NOMINAL
* Binary attributes are referred to as Boolean if the two states are true/false	* Each value represents category code or state sybols of things
* Nominal means relating to names	* Binary means two categories
ex: Smoker - 1 or 0 Diseased - 1 or 0	eg: Gender - M, F Hair colour - black, brown, gray, white.

How can the data be pre-processed in order to help improve the quality of the data and consequently of the mining results?

* Data processing techniques are applied before mining so that it improves the overall quality of the patterns mined and time required for the actual mining.

* Data Pre-processing Methods
  → Data cleaning
  → Data Integration

2310080078

→ Data Reduction

→ Data Transformation

Find the Outliers if any in the following data

2, 4, 5, 6, 8, 12, 19, 22, 28, 30, 100.

Median = $\frac{n+1}{2} = \frac{11+1}{2} = \frac{12}{2} = 6$ th term

= 12

$Q_1 = 5$

$Q_3 = 28$.

$IQR = Q_3 - Q_1 = 23$

Lower Bound = $Q_1 - 1.5 \times IQR$

$= 5 - 1.5 \times 23$

$= -29.5$

Upper Bound = $Q_3 + 1.5 \times IQR$

$= 28 + 1.5 \times 23$

$= 62.5$

∴ Outlier = 100.