Assignment - 01

(Q1) What is Data Minning? Describe the steps involved in data mining when Viewed as a process of Knowledge discovery.

Data Mining refers to extracting or mining Knowledge from large amount of data. It is also called as Knowledge discovery from data.

Steps involved in data mining when viewed as a process of knowledge discovery are:-

(i) Data Cleaning: To remove noise and inconsistent data.

(ii) Data integration: Where multiple data sources maybe combined and stored in data warehouse.

(iii) Data Selection: where data relevant to the analysis task are retrieved from the database.

(iv) Data Transformation: Where data are transformed or consolidated into appropriate forms for mining by performing sealing.

(v) Data Mining: An essential process where intelligent methods are applied in order to extract data patterns.

(vi) Pattern evaluation: To identify truely interesting patterns representing Knowledge based on measures.

(vii) Knowledge Presentation: It is used to present the mind Knowledge to the user.

(Q2) Define predictive Analytics and explain its relationship with Data Mining.

- Predictive Analytics is the process of extracting information from Large dataset in order to make prediction and estimate about future outcomes.
- Data Mining is the broader process of discovering hidden patterns, correlations and relationships in large datasets. Whereas predictive Analytics is a specific application of data mining that focuses on forecasting future events reathere than just finding existing patterns.

For example, Data Mining is discover that Customers who buy baby products often buy diapers. Whereas predictive Analytics is that use, that pattern to predict which customers are likely to buy diapers next week.

Q3) For each of the following meetings, explain which phase in the CRISP-DM process is represented:

(a) Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is:-

Evaluation phase.

(b) The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.

Data understanding phase.

(c) The data mining consultant meets with the production line supervisor, to discuss implementation of changes and improvements.

Business understanding phase

d) The data mining project manager meets with the production line supervisor, to discuss implementation of changes and improvements.

Deployment phase

e) The analysts meet to discuss whether the neural network or decision tree models should be applied.

Modelling phase

4) What is an outlier? Why do we need to treat outliers carefully?

An outlier is a data point that is significantly different from the rest of the data values. It lies far away from the average range of dataset.

We need to treat outliers carefully because:-

) It affect model accuracy.

i) It may represent errors.

i) It may contain valuable information.

v) It can disort statistical results.

(Q5) Use the following stock price data (in dollars) to compute mean, median, mode and standard deviation.

Stock Price	10	7	20	12	75	15	9	18	4	12	8	14

dataset = 4, 7, 8, 9, 10, 12, 12, 14, 15, 18, 20, 75

$$\text{Mean} = \frac{4+7+8+9+10+12+12+14+15+18+20+75}{12}$$

$$= 17$$

$$\text{Median} = \frac{12+12}{2} = \frac{24}{2} = 12$$

Mode = 12

$$\sigma^2 = \frac{1}{12}\Big[(4-17)^2+(7-17)^2+(8-17)^2+(9-17)^2+(10-17)^2+(12-17)^2+(12-17)^2+(14-17)^2+(15-17)^2+(18-17)^2+(20-17)^2+(75-17)^2\Big]$$

$$= \frac{1}{12}\Big[(-13)^2+(-10)^2+(-9)^2+(-8)^2+(-7)^2+(-5)^2+(-5)^2+(-3)^2+(-2)^2+(1)^2+(3)^2+(58)^2\Big]$$

$$= \frac{1}{12}\Big[169+100+81+64+49+25+25+9+4+1+9+3364\Big]$$

$$= \frac{1}{12}\big[3900\big] = 325$$

$$\text{Standard deviation } (\sigma) = \sqrt{\sigma^2} = \sqrt{325} \approx 18.03$$

(Q6) Use the following stock price data (in dollars) to answer the following questions.

Stock Price	10	7	20	12	75	15	9	18	4	12	8	14

(a) Find the min-max normalized stock price for the stock worth \$20.

Min = 4, max = 75, $x = 20$

$$\text{Min-Max normalisation} = \frac{x-\text{min}}{\text{max}-\text{min}} = \frac{20-4}{75-4} = \frac{16}{71}$$

$$= 0.225$$

(b) Compute the midrange stock price.

$$\text{Midrange} = \frac{\max + \min}{2} = \frac{75+4}{2} = \frac{79}{2} = 39.5$$

(c) Compute the z-score standardized stock price for the stock worth \$20.

$x = 20$, $\mu = 17$ (mean), $\sigma = 18.03$

$$\text{Z-Score} = \frac{x-\mu}{\sigma} = \frac{20-17}{18.03} = \frac{3}{18.03} \approx 0.167$$

(d) Compute the decimal scaling stock price for the stock worth \$20. .

$x = 20$, $d = 2$

$$\text{Decimal scaling} = \frac{X}{10^d} = \frac{20}{10^2} = \frac{20}{100} = 0.2$$

(e) Compute the skewness for the stock price data.

mean = 17 , median = 12 , Standard deviation = 18.03

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{Standard deviation}}$$

$$= \frac{3(17-12)}{18.03} = \frac{3(5)}{18.03} \approx 0.832$$

Q7) Use the given dataset for the following questions:
111337

(a) Bin the data into three bins of equal width (width = 3).

Range = max − min = 7 − 1 = 6

Bin 1 = 1,1
Bin 2 = 1,3
Bin 3 = 3,7

Bin no.	range	Value falling in bin
Bin1	1-3	1,1,1,3,3
Bin2	4-6	–
Bin3	7-9	7

∴ Bin 1 = 1,1,1,3,3
Bin 2 = empty
Bin 3 = 7

(b) Bin the data into three bins of two records each.

Bin no.	records
Bin1	1,1
Bin2	1,3
Bin3	3,7

(Q8) Answer following questions:

(a) What is the graphical counterpart of a contigency table?.

The graphical counterpart of a contigency table is a mosaic. A mosaic plot visually displays the relationship between two categorical variable, where the size of each rectangle represents the frequency of each cell in the contigency table.

(b) What is the difference between taking row percentages and taking column percentages in a contigency table?

Row percentage	Column percentage
(i) Each cell's value is expressed as a percentage of the total for that row	(i) Each cell's value is expressed as a percentage of the total for that column.
(ii) To understand how the categories of one variable are distributed accross the levels of another variable within each row.	(ii) To understand how the categories of one variable are distributed accross the levels of another variable within each column.

(Q9) For each of the following descriptive methods, state whether it may be applied to categorical data, continuous, numerical data or both.

(a) Bar Chats: Categorical data

(b) Histograms: Continous numerical data

(c) Summary statistics: Continuous numerical data

(d) Cross-tabulations: Categorical data.

correlation analysis : Continous numerical data

scatter plots : Continous numerical data

web graphs : Both

Binning : Continous numerical data.

Find Q_1, Q_2, Q_3 for the following data set, and draw a box-and-whisker plot. {2, 6, 7, 8, 8, 11, 12, 13, 14, 15, 22, 23}.

2, 6, 7, 8, 8, 11, 12, 13, 14, 15, 22, 23

min = 2, max = 23.

$$\text{Median} = \frac{11+12}{2} = \frac{23}{2} = 11.5 = Q_2$$

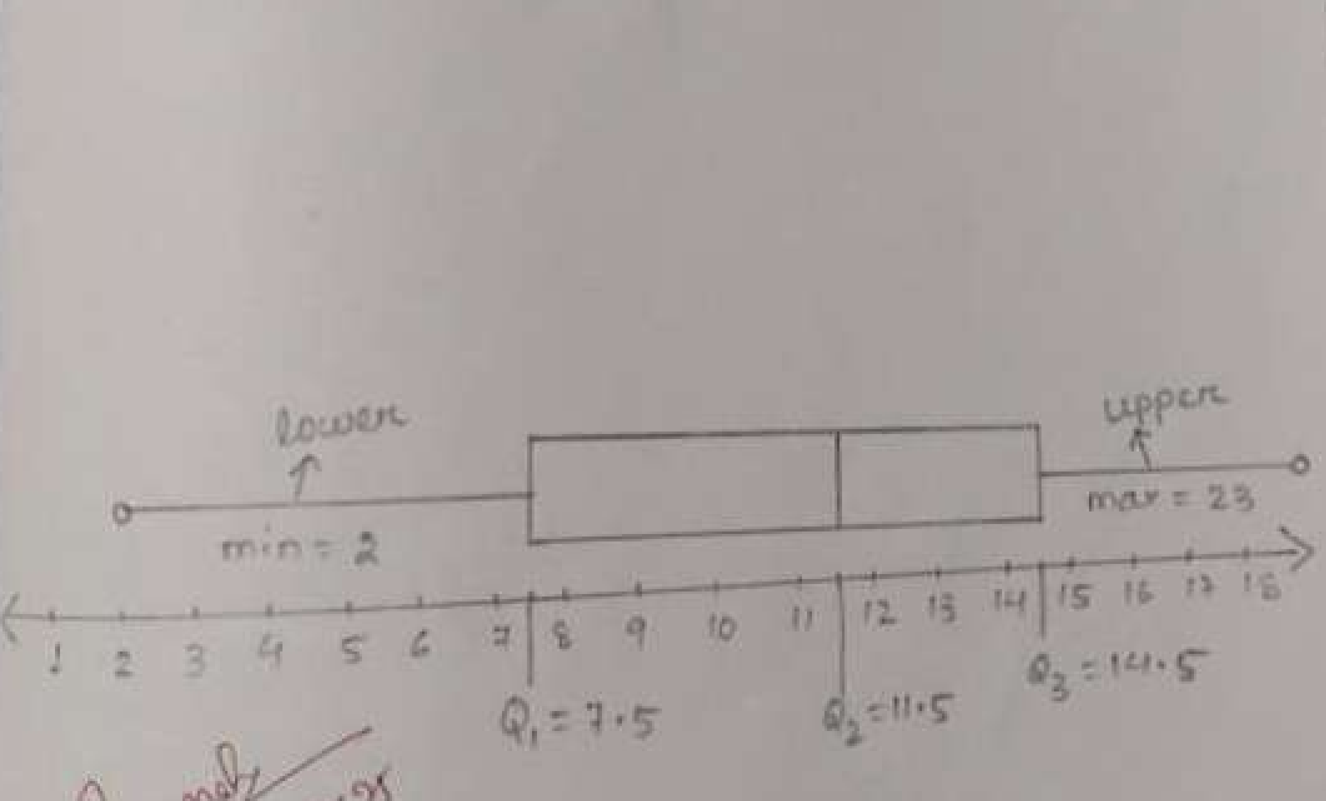
$$Q_1 = \frac{7+8}{2} = 7.5 \quad , \quad Q_3 = \frac{14+15}{2} = \frac{29}{2} = 14.5$$

$$\text{Lower} = Q_1 - 1.5 \times IQR = 7.5 - 1.5(7) = -3$$

$$\text{upper} = Q_3 + 1.5 \times IQR = 14.5 + 1.5(7) = 25$$

$$IQR = Q_3 - Q_1 = 14.5 - 7.5 = 7$$

Box - plot



25.10.25