Name:	Ayush Gutpa
Roll No:	12
Class/Sem:	TE/V
Experiment No.:	3
Title:	Tutorial on: a) Data Exploration b) Data pre-processing
Date of Performance:	
Date of Submission:	
Marks:	
Sign of Faculty:	

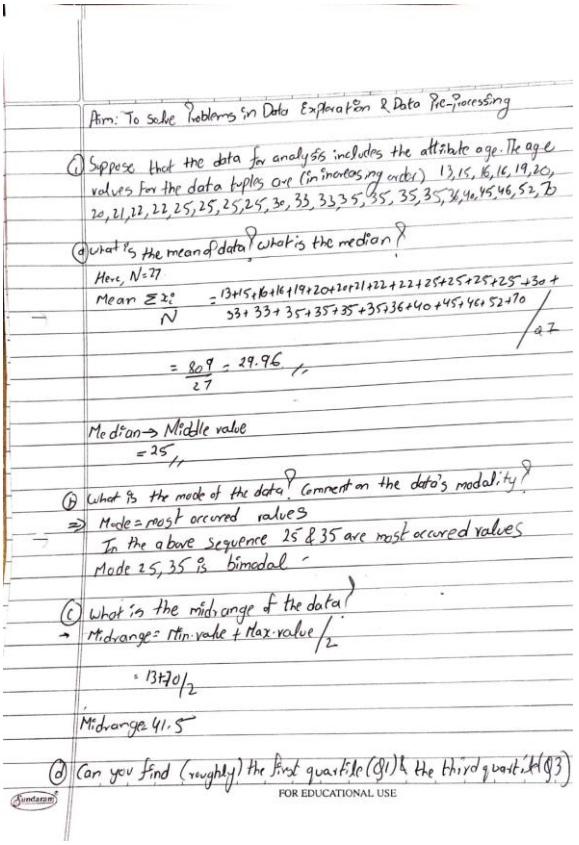
Aim: To solve problems in Data Exploration and Data Pre-processing.

Objective: To enable students to effectively identify sources of data and process it for data mining.

1. Suppose that 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.

a. What is the mean of the data? What is the median?
b. What is the mode of the data? Comment on the data's modality (i.e., unimodal, bimodal, trimodal, etc.).
c. What is the midrange of the data?
d. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
e. Give the five-number summary of the data.
f. Show a boxplot of the data.

Solution:


Aim: To solve Problems in Data Exploration & Data Pre-processing

(1) Suppose that 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

(a) What is the mean of data? what is the median?

Here, N=27

Mean $\frac{\Sigma x}{N}$ = 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 / 27

= 809/27 = 29.96 //

Median → Middle value

= 25 //

(b) What is the mode of the data? Comment on the data's modality?

⇒ Mode = most occured values

In the above sequence 25 & 35 are most occured values

Mode 25, 35 is bimodal.

(c) What is the midrange of the data?

→ Midrange = Min value + Max value / 2

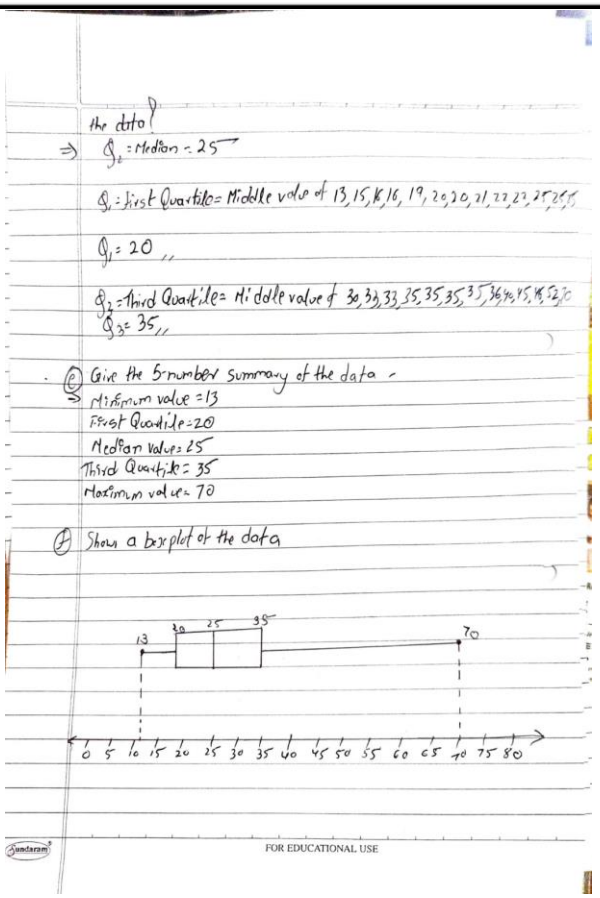
= 13+70/2

Midrange 41.5

(d) Can you find (roughly) the first quartile (Q1) & the third quartile (Q3)

Sundaram

FOR EDUCATIONAL USE



2. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Compute an approximate median value for the data.

Solution:

n = 3194
n/2 = 1597

This observation lie between the class interval 21-50 which is the median class.

lower class limit = 21 = (l)
class size (h) = 30
frequency of the median class (f) = 1500
cumulative frequency of class preceding the median class (cf) = 950.

$\therefore$ median $= l + \frac{(n/2 - cf)}{f} \times h = 21 + \frac{(1597 - 950)}{1500} \times 30 = 21 + 12.94 = 33.94$

$\therefore$ Median = 33.94

FOR EDUCATIONAL USE

3. Consider the data given below and compute the Euclidean distance between each point. P1 (0,2), P2(2,0), P3(3,1) and P4(5,1).

Solution :

3) Consider the data given below & compute the Euclidean distance between each point.
P1(0,2), P2(2,0), P3(3,1) & P4(5,1).

→ Soln :-

Point	x	y
P1	0	2
P2	2	0
P3	3	1
P4	5	1

FOR EDUCATIONAL USE

$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2} = \sum_{i=1}^{n}\sqrt{(x_i - y_i)^2}$

$\therefore d(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(0-2)^2 + (2-0)^2}$

$d(P_1, P_2) = \sqrt{4+4} = \sqrt{8} = 2.828$

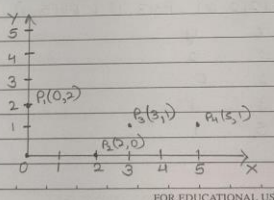
$\therefore d(P_1, P_3) = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} = \sqrt{(0-3)^2 + (2-1)^2} = \sqrt{9+1} = \sqrt{10} = 3.16$

$\therefore d(P_1, P_4) = \sqrt{(x_1 - x_4)^2 + (y_1 - y_4)^2} = \sqrt{(0-5)^2 + (2-1)^2} = \sqrt{25+1} = \sqrt{26} = 5.09$

$\therefore d(P_2, P_3) = \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2} = \sqrt{(2-3)^2 + (0-1)^2} = \sqrt{1+1} = \sqrt{2} = 1.414$

$\therefore d(P_3, P_4) = \sqrt{(x_3 - x_4)^2 + (y_3 - y_4)^2} = \sqrt{(3-5)^2 + (1-1)^2} = \sqrt{2^2} = 2$

	P1	P2	P3	P4
P1	0	2.828	3.16	5.09
P2	2.828	0	1.4114	3.16
P3	3.16	1.2414	0	2
P4	5.09	3.16	2	0



FOR EDUCATIONAL USE

4. Suppose that the minimum and maximum values for the attribute income are $12,000 and

$98,000 respectively. Normalize income value $73,600 to the range [0.0, 1.0] using min-max normalization method.

5. Partition the given data into bins of size 3 using equi-depth binning method and perform smoothing by bin mean, bin median and bin boundaries. Consider the data: 2, 10, 18, 18, 19, 20, 22, 25, 28.

Solution: