Name:	Priyanka N Bhandari
Roll No:	02
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:	

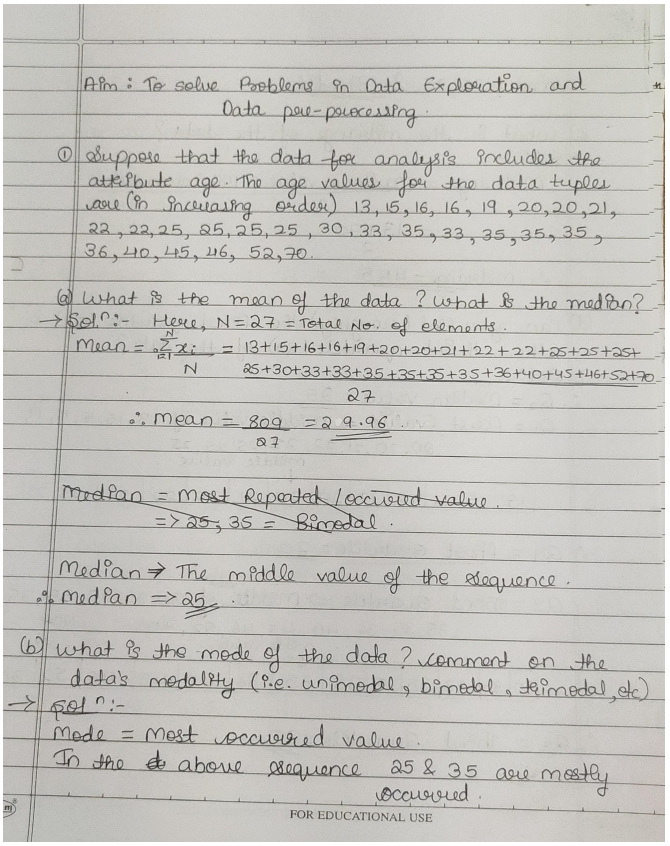
Aim: To solve problems in Data Exploration and Data Pre-processing.

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 on Data Exploration and 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 the data ? what is the median?

→ Sol^n :- Here, N = 27 = Total No. of elements.

$$\text{Mean} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70}{27}$$

$$\therefore \text{mean} = \frac{809}{27} = 29.96$$

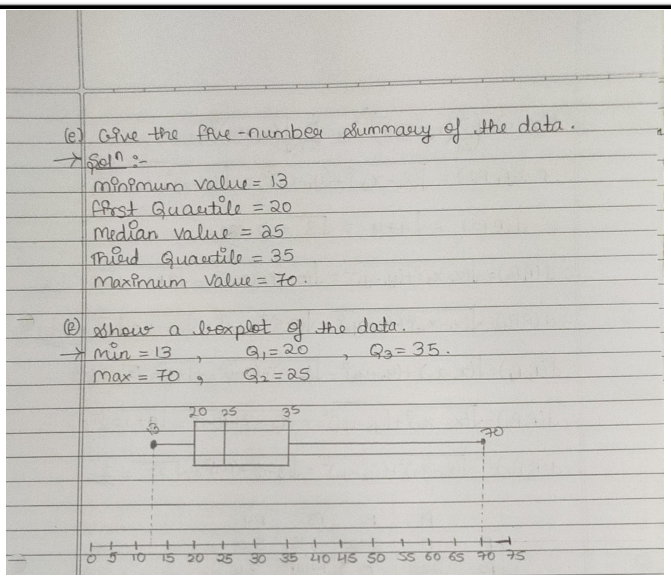
Median = most Repeated / occurred value.
=> 25, 35 = Bimodal.

Median → The middle value of the sequence.
∴ Median => 25.

(b) what is the mode of the data ? comment on the data's modality (i.e. unimodal, bimodal, trimodal, etc)

→ Sol^n :-
Mode = Most occurred value.
In the above sequence 25 & 35 are mostly occurred.

FOR EDUCATIONAL USE

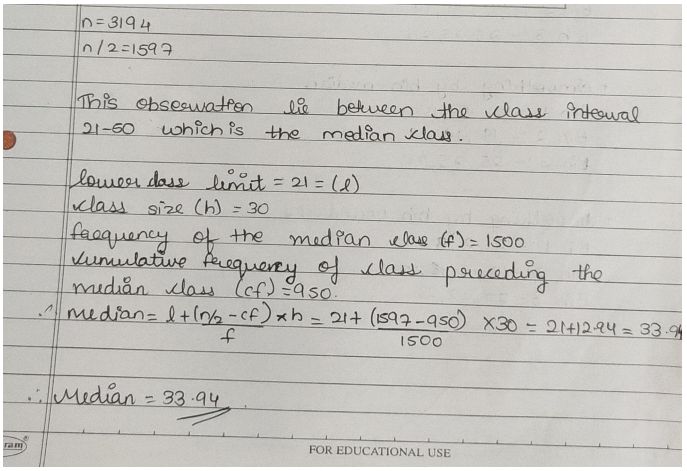


2. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

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

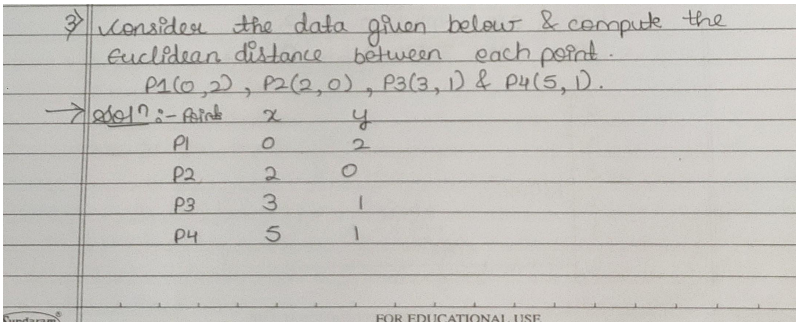
Compute an approximate median value for the data.

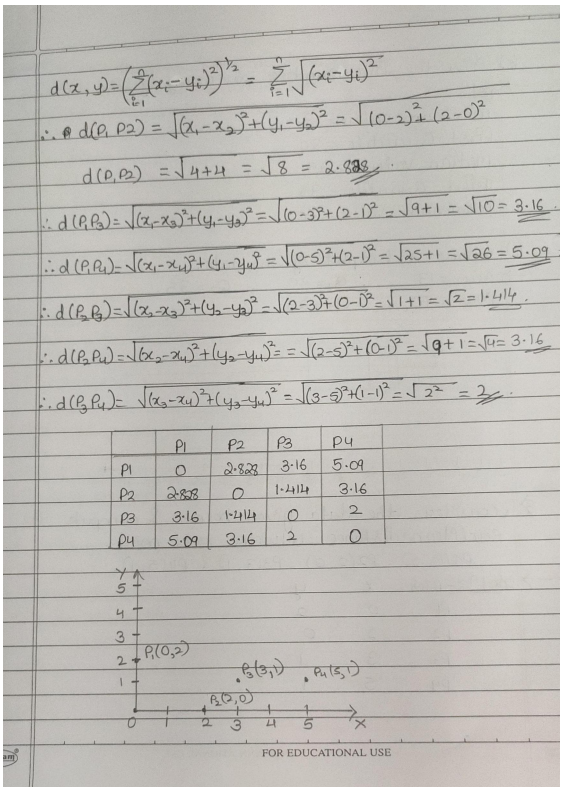
Solution:



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 :

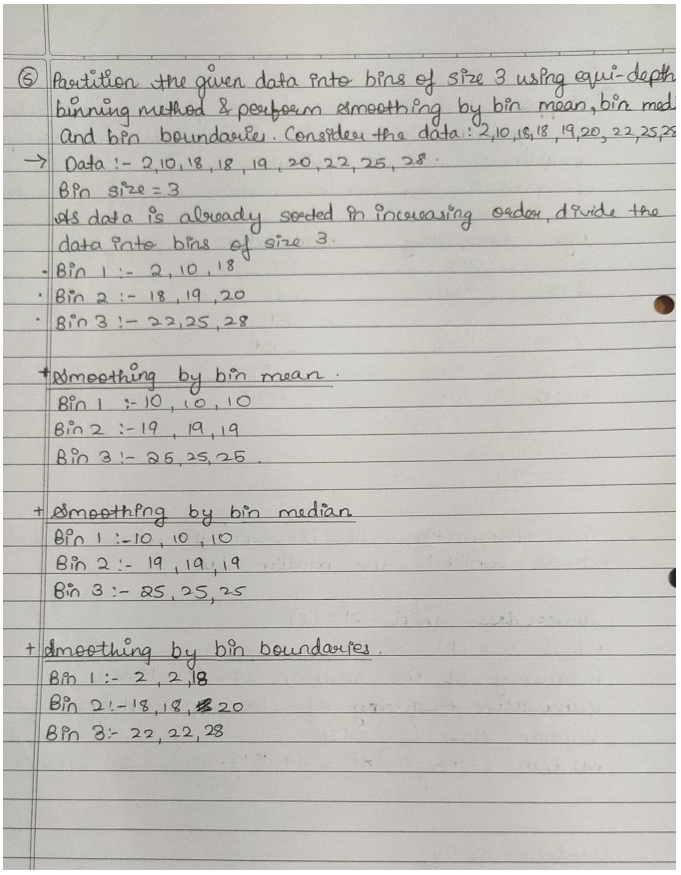




4. Suppose that the minimum and maximum values for the attribute income are $12,000 and $98,000 respectively. Normalize income value $73,600 to the range [0.0, 1.0] using min-max normalization method.

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:



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

→ Data :- 2,10,18,18,19,20,22,25,28.

Bin size = 3

As data is already sorted in increasing order, divide the data into bins of size 3.

- Bin 1 :- 2,10,18
- Bin 2 :- 18,19,20
- Bin 3 :- 22,25,28

+ Smoothing by bin mean.

Bin 1 :- 10,10,10

Bin 2 :- 19,19,19

Bin 3 :- 25,25,25.

+ Smoothing by bin median

Bin 1 :- 10,10,10

Bin 2 :- 19,19,19

Bin 3 :- 25,25,25

+ Smoothing by bin boundaries.

Bin 1 :- 2,2,18

Bin 2 :- 18,18,20

Bin 3 :- 22,22,28