Course: **Data Mining [A]** 10-Jan-2023 **(Fall 2022)**

Resource Person: **Dr. Muhammad Faheem** **QUIZ-1 (Getting to know your data)**

Total Points: 20 **Time Allowed: 20 Minutes**

Solution

Question 1: [5]

Explain the following concepts:

- Differentiate between characterization and discrimination with examples.
- Define support and confidence in association analysis with examples.

a) Data entries can be associated with classes or concepts and such descriptions can be extracted using data characterization or data discrimination.

Data Characterization:
Summary or features of a target class is called data characterization. For example, summarize the characteristics of customer who purchase more than 20 products a year.

Data Discrimination:
Comparison of target class data objects and contrasting class(es) data objects on the basis of general features is called data discrimination. For example, compare the characteristics of customers who buy 20 products per years and those who purchase less than 5 products annually.

Refer to the book section 1.4.1 for better understanding of the topic.

b) An association rule such as **computer ➔ software** [1%, 50%] states that there is 50% confidence (i.e. certainty) that if a customer buys computer he/she also purchase software. The rule also suggests that 1% of the transactions under analysis show that computer and software are bought together.

Refer to the book section 1.4.2 for better understanding of the topic.

Question 2: [15]

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?
(c) What is the *midrange* of the data?
(d) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data?
(e) Give the *five-number summary* of the data.
(f) Show a *boxplot* of the data.
(g) Draw the *quantile plot*?

Mean = 29.962962963

Median = 25

Mode = 23, 35

Midrange = (13+70)/2 = 41.5

Q1 = 20

Q2 = 25

Q3 = 35

IQR = 15

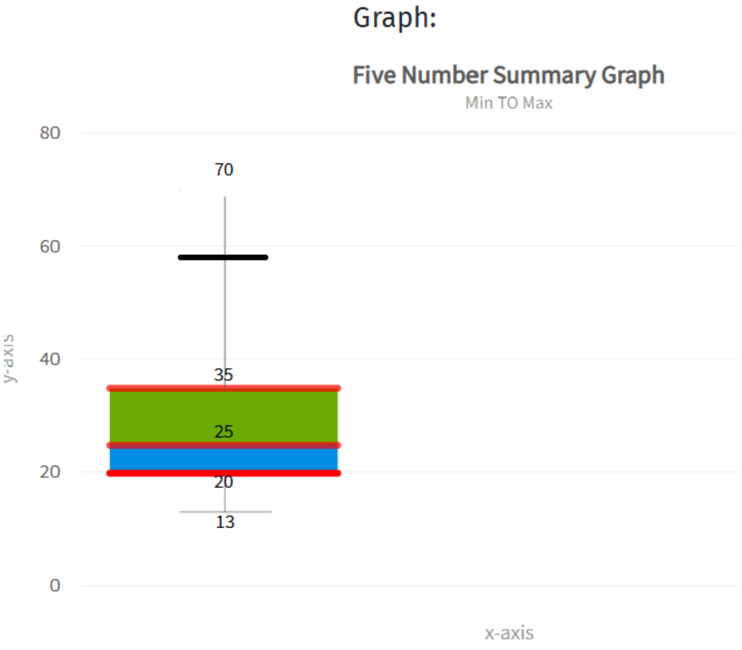
Outliers exists before Q1: Q1 – 1.5*IQR = 20 – 22.5 = -2.5

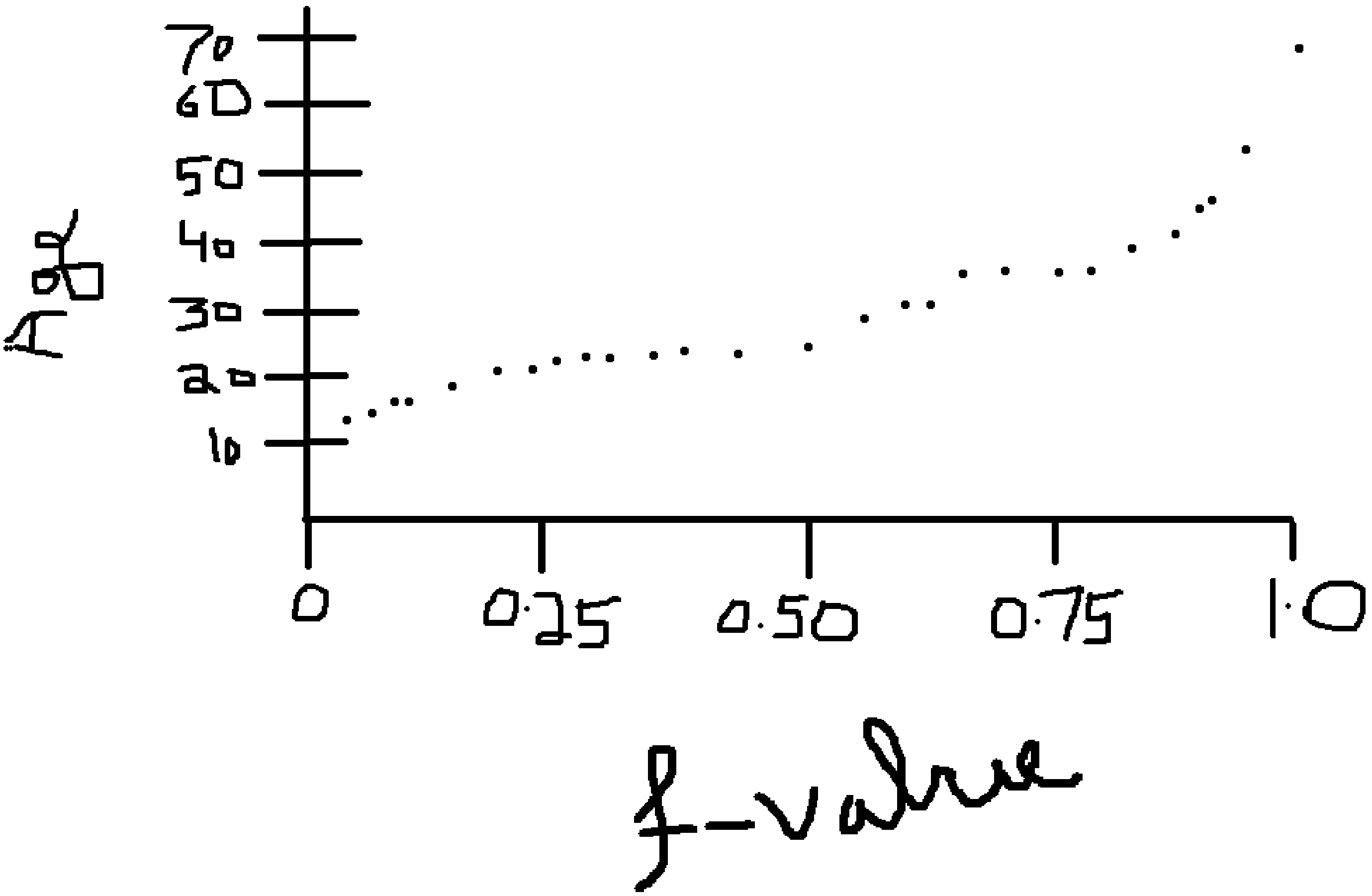
Min = 13 (The whisker Line)

Outlier exists after Q3: Q3 + 1.5*IQR = 35 +22.5 = 57.5

Max = 70 (but its outlier and whisker line plotted at 57.5)

Refer to the book section 2.2 for plotting the boxplot




70
60
50
40
30
20
10
Age
0
0.25
0.50
0.75
1.0
f-value