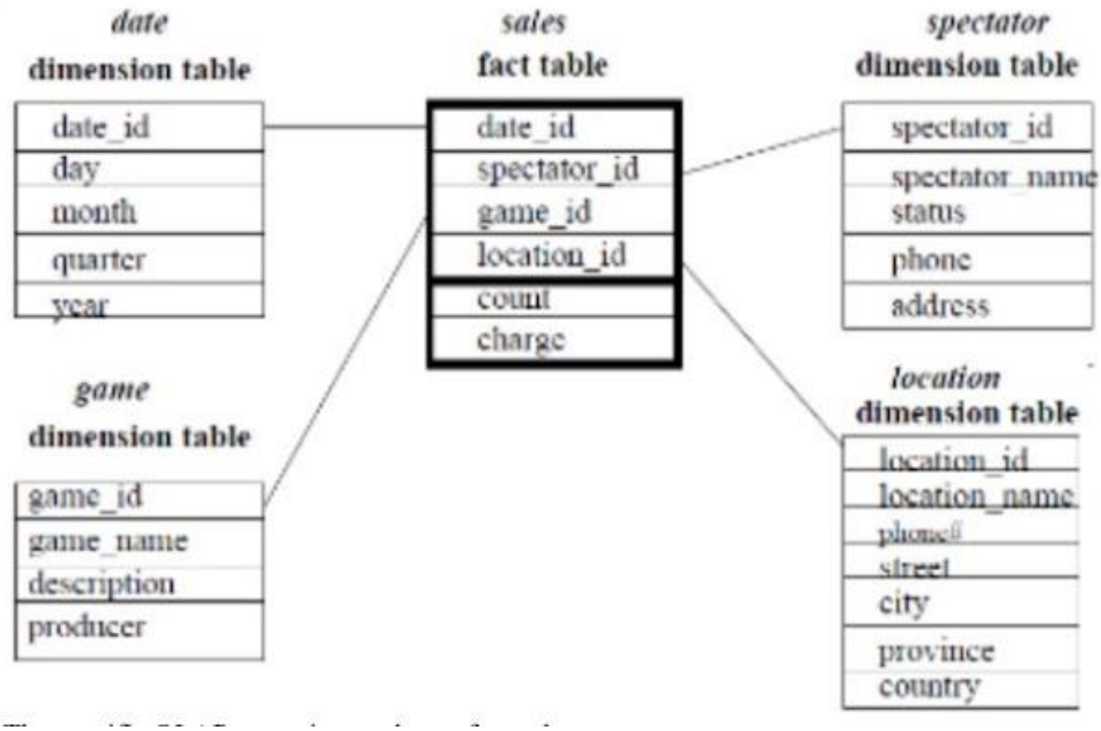
Q.5) Suppose that a data warehouse consists of the four dimensions date, spectator, location and game, and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(a) Draw a star schema diagram for the data warehouse.

(b) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should you perform in order to list the total charge paid by student spectators at GM Place in 2004?

Ans:

a)



b) The specific OLAP operations to be performed are:

1. Roll-up on date from date id to year.

2. Roll-up on spectator from spectator id to status.

3. Roll-up on location from location id to location name.

4. Roll-up on the game from game id to all.

5. Dice with status= “students”, location name= “GM Place”, and year=2004

```
SELECT SUM(Charge)
FROM DataWarehouse
WHERE Date.Year = 2004
  AND Spectator.Category = 'Students'
  AND Location.Name = 'GM Place';
```

Q.4) Suppose a group of 12 sales price records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods:

(a) equal-frequency (equal-depth) partitioning

(b) equal-width partitioning

(c) clustering

Ans:

(a) **Equal-Frequency (Equal-Depth) Partitioning:**

- Divide the data into three bins with an equal number of records in each bin.
- In this case, since there are 12 records, each bin would contain 4 records.

Bins:

- Bin 1: 5, 10, 11, 13
- Bin 2: 15, 35, 50, 55
- Bin 3: 72, 92, 204, 215

(a) **Equal-Frequency (Equal-Depth) Partitioning:**

- Divide the data into three bins with an equal number of records in each bin.
- In this case, since there are 12 records, each bin would contain 4 records.

Bins:

- Bin 1: 5, 10, 11, 13
- Bin 2: 15, 35, 50, 55
- Bin 3: 72, 92, 204, 215

(b) **Equal-Width Partitioning:**

- Determine the range of the data (max - min) and divide it by the number of desired bins.
- In this case, let's aim for three bins.

Range = 215 - 5 = 210 Width of each bin = 210 / 3 = 70

bin1:5,10,11,13,15,35,50,55,72	(all values between 5 and 75)
bin2:92	(all values between 75 and 145)
bin3:204, 215	(all values between 145 and 215)

- Use clustering algorithms to group similar values together.
- One simple approach might be to use a k-means clustering algorithm with k=3.

Clusters:

- Cluster 1: 5, 10, 11, 13, 15, 35
- Cluster 2: 50, 55, 72, 92
- Cluster 3: 204, 215

Q.2) 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., bimodal, trimodal, etc.).

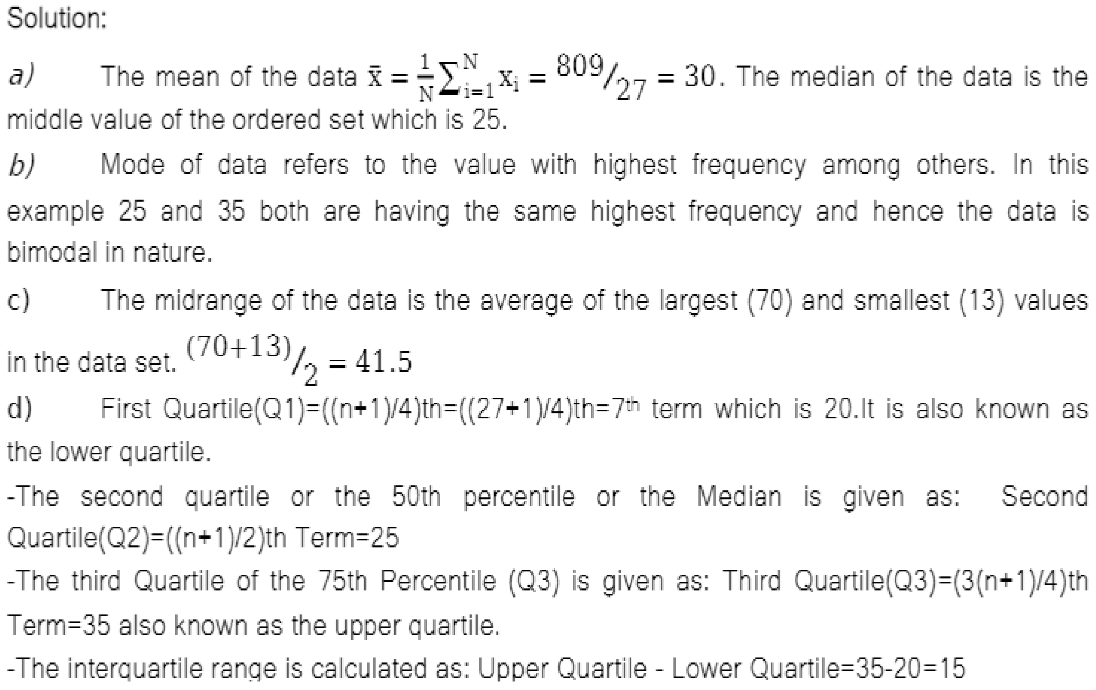
(c) What is the midrange of the data?

(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.

(g) How is a quantile–quantile plot different from a quantile plot?



Solution:

a) The mean of the data $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = {}^{809}/_{27} = 30$. The median of the data is the middle value of the ordered set which is 25.

b) Mode of data refers to the value with highest frequency among others. In this example 25 and 35 both are having the same highest frequency and hence the data is bimodal in nature.

c) The midrange of the data is the average of the largest (70) and smallest (13) values in the data set. ${}^{(70+13)}/_{2} = 41.5$

d) First Quartile(Q1)=((n+1)/4)th=((27+1)/4)th=7th term which is 20.It is also known as the lower quartile.

-The second quartile or the 50th percentile or the Median is given as: Second Quartile(Q2)=((n+1)/2)th Term=25

-The third Quartile of the 75th Percentile (Q3) is given as: Third Quartile(Q3)=(3(n+1)/4)th Term=35 also known as the upper quartile.

-The interquartile range is calculated as: Upper Quartile - Lower Quartile=35-20=15

a) **Mean and Median:**

- Mean: Add up all the values and divide by the number of values. Mean = (13 + 15 + ... + 70) / 26 ≈ 28.5
- Median: Since there are 26 values, the median is the average of the 13th and 14th values. Median = (25 + 25) / 2 = 25

(b) **Mode and Modality:**

- Mode: 25 is the mode as it appears more frequently than any other value.
- Modality: The data is unimodal, meaning it has one mode.

(c) **Midrange:**

- Midrange is the average of the minimum and maximum values. Midrange = (13 + 70) / 2 = 41.5

(d) **First and Third Quartiles (Q1 and Q3):**

- To find Q1 and Q3, divide the data into quartiles. Q1 is the median of the lower half, and Q3 is the median of the upper half. Q1 ≈ 20, Q3 ≈ 35

(e) **Five-Number Summary:**

- Minimum: 13
- Q1: 20
- Median: 25
- Q3: 35
- Maximum: 70

(f) **Boxplot:**

- A boxplot visually represents the five-number summary, displaying the minimum, Q1, median, Q3, and maximum values.

(g) **Quantile-Quantile Plot vs. Quantile Plot:**

- A quantile-quantile (Q-Q) plot compares the quantiles of a dataset with the quantiles of a theoretical distribution (e.g., normal distribution). It helps assess if the data follows a particular distribution.
- A quantile plot typically shows the percentiles or quantiles of a dataset on one axis and the corresponding values from a standard distribution on the other. It helps visualize how the data deviates from a theoretical distribution

The given Data Points **= 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**

Total Data Points (n) **= 27**

a] Mean and Median of the Data Points

Mean -

- Mean is the arithmetic average of data points.

- This is the addition of the numbers of data points and dividing by the total data points.

Median -

- The median is the middle number in the data points when the numbers are listed in either ascending or descending order.
- If data points are not listed in any order then first arrange them in ascending order and then find out the Median.

Median when total Data Points (n) are ODD:

Median=(n+12)th number������=(�+12)�h ������

Median when total Data Points (n) are EVEN:

Median=(n2)thnumber+(n+12)thnumber2������=(�2)�h������+(�+
12)�h������2

- Here, the total data points are **27 which means ODD.**

Median of Data Points=(27+12)th number=14th positioned number�����
� �� ���� ������=(27+12)�h ������=14�h ���������
� ������

Therefore,

Median=25

b] Mode of Data Points

- The mode is the most frequently occurring number in the data points.
- Here, in the given data points numbers **25 and 35** are modes with the ***most frequent occurrence count of 4.***

Mode of Data Points=25 and 35 (Occurence Count 4)���� �� ���� �
�����=25 ��� 35 (��������� ����� 4)

c] Midrange of Data Points

- Midrange is the difference between the highest and lowest values in the data points.
- It shows the halfway between the minimum and maximum numbers of the data points.

Midrange of Data Points=Maximum Number in Data Points+Minimum Numbe
r in Data Points2�������� �� ���� ������=������� �

����� �� ���� ������+������� ������ �� ����
������2

Midrange of Data Points=70+132=832�������� �� ���� ����
��=70+132=832

Midrange of Data Points=41.5�������� �� ���� ������=41.5

d] Q1, Q3 of Data Points

- Q1 and Q3 represent the Quartiles.
- In statistical measure, a quartile, is one type of quantile of three points (Q1, Q2, & Q3) that divides sorted data points into four equal groups in terms of count of numbers, each representing a fourth of the distributed sampled population.
- There are three quartiles as follows:

The First Quartile (Q1) - *It is a 1st quartile or lower quartile that separates the lowest 25% of data from the highest 75%.*

Lower Quartile (Q1)=[(n+1)×14]thnumber����� �������� (�1)=[(
�+1)×14]�h������

Q1=(27+1)×14=284=7th positioned number�1=(27+1)×14=284=7�h �����
����� ������

Q1=20�1=20

The Second Quartile (Q2) - *It is a 2nd quartile or middle quartile also same as* ***Median*** *it divides numbers into 2 equal parts.*

Middle Quartile (Q2)=[(n+1)×24]thnumber������ �������� (�2)=
[(�+1)×24]�h������

Q2=(27+1)×24=564=14th positioned number�2=(27+1)×24=564=14�h ����
������ ������

Q2=25�2=25

The Third Quartile (Q3) - *It is a 3rd quartile or the upper quartile that separate the highest 25% of data from the lowest 75%.*

Upper Quartile (Q3)=[(n+1)×34]thnumber����� �������� (�3)=[(
�+1)×34]�h������

Upper Quartile (Q3)=(27+1)×34=844=21th positioned number����� ��
������ (�3)=(27+1)×34=844=21�h ���������� ������

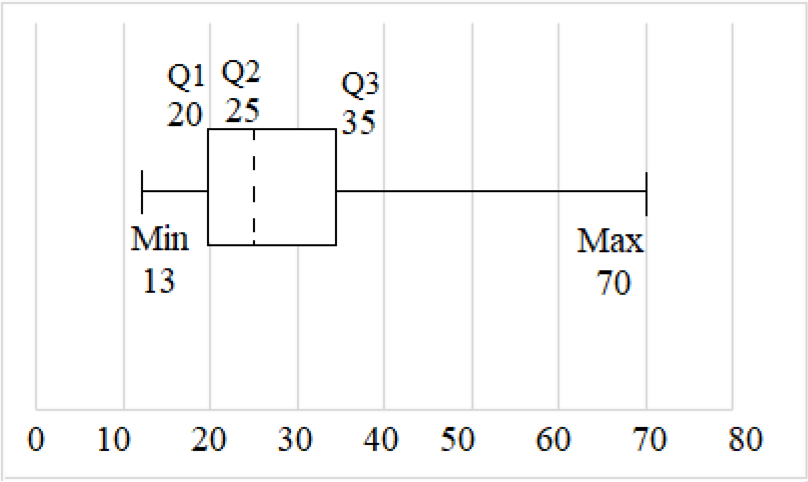
Q3=35�3=35

- Based on Q1 & Q3 values ***Interquartile Range*** also calculated as follows:

Interquartile Range=Q3−
Q1=35−20=15�������������� ������=�3−�1=35−20=15

e] Boxplot of Data Points

- Box plots represent the graphical image of the concentration of the data points.
- The box plot is created based on the 5 values as follows:
  - *The Minimum Value* = ***13***
  - *The First Quartile (Q1)* = ***20***
  - *The Median (Q2)* = ***25***
  - *The Third Quartile (Q3)* = ***35***
  - *The Maximum Value* = ***70***
- The box plot can be drawn either by vertically or horizontally.
- For the given data points Horizontal Box Plot can be drawn as follows:



Box Plot for the given Data Points

c) The midrange (average of the largest and smallest values in the data set) of the data is: (70+13) / 2 = 41:5

First quartile = 25% of the given data and the value of quartile1 is 20.25. Third quartile = 75% of the given data and the value of quartile3 is 35.

e)The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, 70.

g)A quantile plot is a graphical technique used to approximate the percentage of values below or equal to the independent variable in a univariate distribution. In this manner, it shows quantile information for all the data, where the qualities measured for the independent variable are plotted against their independent quantile. The quantile-

quantile or q-q plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a

Q.3) Use these methods to normalize the following group of data:

200, 300, 400, 600, 1000

(a) min-max normalization by setting min = 0 and max = 1

(b) z-score normalization

/*

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

*/

Q) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the 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

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

(b) Draw the boxplots for age and %fat.

(c) Draw a scatter plot and a q-q plot based on these two variables.