Mid_term exam solution

Question 1

a - What is data Mining? Mention steps involved in data Mining when ie viewed as a process of Knowledge discovery

- Data Mining refers to extracting or "Mining" knowledge from large amounts of data « Knowledge mining of data".

- Extraction of interesting Pattern or Knowledge from huge amount of data

- other names : - Knowledge discovery (mining) in databases.
  - Knowledge extraction
  - Data / Pattern analysis
  - Information harvesting
  - Business Intelligence.

Knowledge discovery is an iterative sequence of the following steps:

1- Data Cleaning remove noise and inconistent data.
2- Data Integration where multiple data sources may be combined.
3- Data Selection where data relevant to the analysis task are retreived from the data warehouse.
4- Data Transformation where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.
5- Data Mining an essential process where intelligent methods are applied in order to extract data patterns, knowledge discovery.
6- Pattern evalution to identify the truly interesting patterns representing knowledge based on some interesting measures.

7- Knowledge Presentation where visualizing and knowledge representation techniques are used to present the mined knowledge to the users.

b- Suppose that this data shows the marks from a math exam...

Marks	13	12	9	11	14	12	10	15	11	10	7

i- what is the mean & median of the data.

mean :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad , \frac{124}{11} = 11.27 \quad \simeq 11$$

median : Middle value if odd number of values, or average of the middle two values otherwise.

7 9 10 10 11 (11) 12 12 13 14 15

median = 11

ii- what is the mode of the data

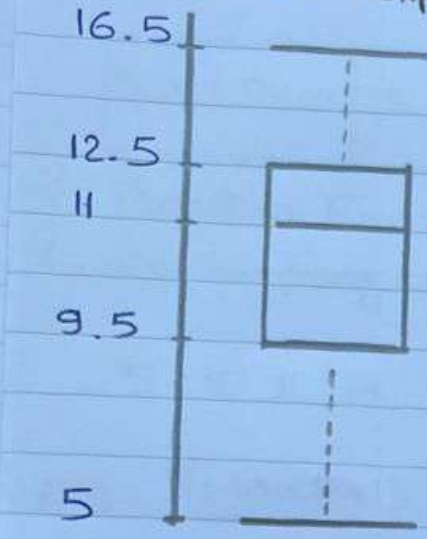
-value that occurs most frequently in the data

→ 10, 11, 12

iii- find roughly Q1 and Q3 of the data.

- Give the five number Summary of the data.

v. Draw the boxplot of the data.



Question-2- In Real world, Tuples with missing values for some attributes are a common occurance. Describe various methods for handling this problem.

Various methods to handle missing data:

- Ignore the Tuple - usually done when class label is missing
  - not effective when the % of values per attribute varies considerably or contains several attributes with missing values

- Fill in Data/value missing manually.
  - This approach is time consuming and may not be responsible task for large data sets with many missing values, specially when the value to be filled in is not easily determined.

- Fill in automatically with:
  - a global constant
  - the attribute mean

#	Value	Count	Percentage	Quantile
1	7	1	1/11 9.09%	
2	9	1	2/11 18.18%	← Q1 25%
3	10	2	4/11 36.36%	
4	11	2	6/11 54.54%	← Q2 50%
5	12	2	8/11 72.72%	← Q3 75%
6	13	1	9/11 81.81%	
7	14	1	10/11 90.90%	
8	15	1	11/11 100%	

- $Q1 = \frac{9+10}{2} = 9.5$, $Q2 = 11$, $Q3 = \frac{12+13}{2} = 12.5$
- $IQR = Q3 - Q1 = 12.5 - 9.5 = 3$
- Maximum Acceptable $= Q3 + 1.5\,IQR$
  $= 12.5 + 1.5\,(3)$
  $= 16.5$
- Minimum Acceptable $= Q1 - 1.5\,IQR$
  $= 9.5 - 1.5\,(3)$
  $= 5$
- outlier $= Max + Min + IQR$, $= 16.5 + 5 + 3$, $= 24.5$

b- Suppose a group of 12 records has been stored as following:

Sales price	5	10	11	13	15	35	50	55	72	92	204	215

i- Partition them into 3 bins by equal frequency.

ii- use Smoothing by bin means to smooth these data partitions.

5 10 11 13	15 35 50 55	72 92 204 215
Avg $\frac{5+10+11+13}{4}$	$\frac{15+35+50+55}{4}$	$\frac{72+92+204+215}{4}$
$= 9.75$	$= 38.75$	$= 145.75$

iii - use min max Normalization to normalize these data onto the range [0.0, 1.0]

$$v' = \frac{v - \min A}{\max A - \min A} * (\text{new } \overset{1}{\max} A - \text{new } \overset{0}{\min} A) + \text{new } \overset{0}{\min} A$$

$v'_5 = \frac{5-5}{215-5} * 1 = 0$

$v'_{10} = \frac{10-5}{215-5} * 1 =$

$v'_{11} = \frac{11-5}{215-5} * 1 =$

$v'_{13} = \frac{13-5}{215-5} * 1 =$

$v'_{15} = \frac{15-5}{215-5} * 1 =$

$v'_{35} = \frac{35-5}{215-5} * 1 =$

$v'_{50} = \frac{50-5}{215-5} * 1 =$

$$V_{55} = \frac{55-5}{215-5} * 1 =$$

$$V_{72} = \frac{72-5}{215-5} * 1 =$$

$$V_{92} = \frac{92-5}{215-5} * 1 =$$

$$V_{204} = \frac{204-5}{215-5} * 1 =$$

$$V_{215} = \frac{215-5}{215-5} * 1 = 1$$

Question _3_

a- explain 3 of The TYPical oLAP operations or multidimensional data.

- Roll up (drill_up): Summarize data
- Drill down (Roll down): reverse of roll_up.
- Slice and dice: Project and select.
- Pivot (rotate)
- other operations: drill across → involving more Than one fact table

  drill Through → Through The bottom level of The cube to its backend relational tables