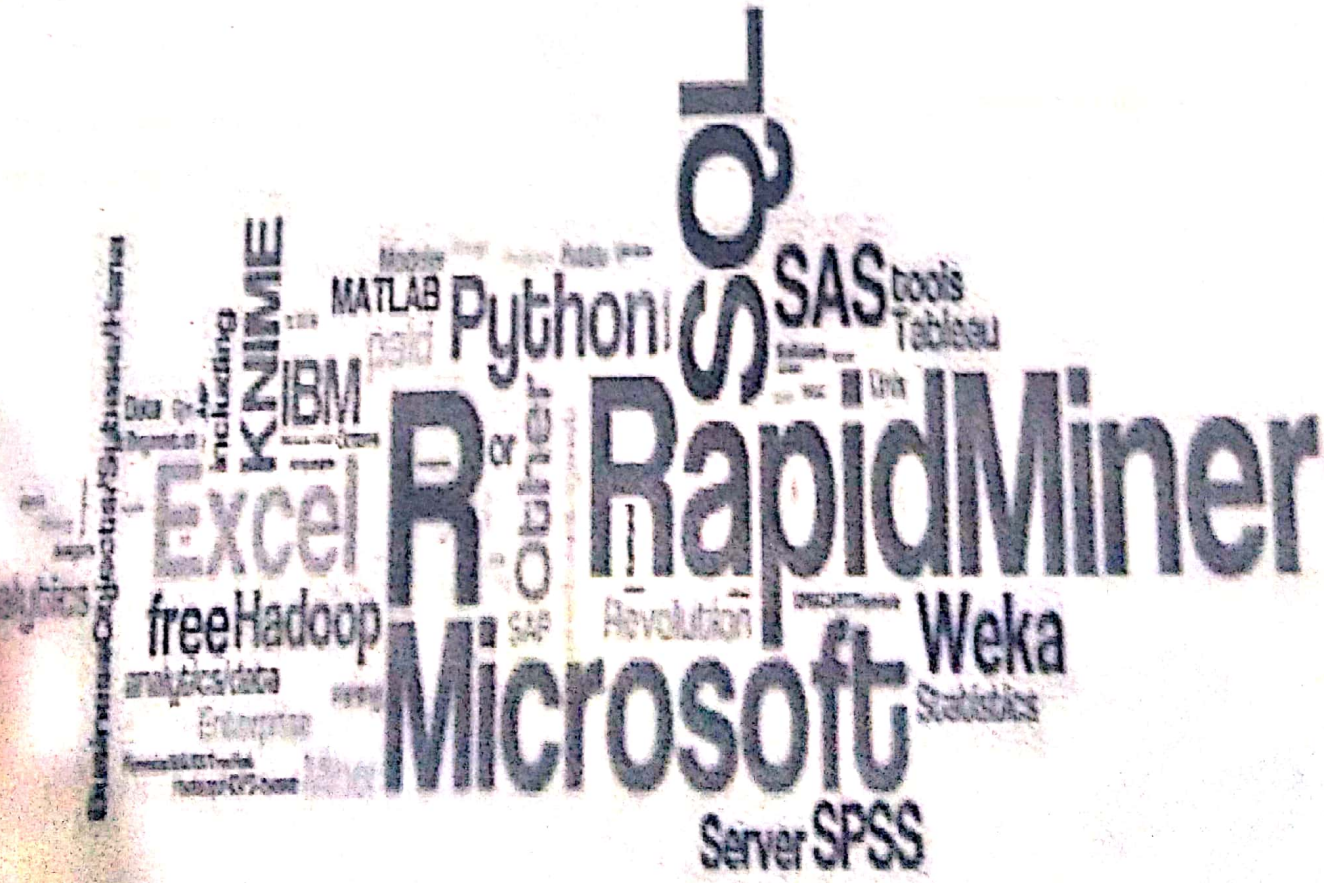
DMining

REVISION 1

2019



- **What is Data Mining?** Extraction of interesting (*non-trivial*, *implicit*, *previously unknown* and *potentially useful*) patterns or knowledge from huge amount of data.

- ***Is everything "data mining"?***
  - Simple search and query processing (**as it is** ....) -----No
  - (Deductive) expert systems (no new information)-----No

- If the mean is equal to the median then this might be an indication that the data is what? ***Symmetric data (normal distribution)***

- If the mean is larger than the median then this might be an indication that the data is what? ***Positive skewed data***

- If the mean is smaller than the median then this might be an indication that the data is what? ***Negative skewed data***

- If you have 100 values in your data and I add 5.0 to all of the values, then how will this change the standard deviation? ***The same***

- SD =0 when all observation are .....(***equal to mean***)

- If you have 100 values in my data and you add 5.0 to all of the values, then how will this change the median?

- If the mean, median and mode of a distribution are 5, 6, 7 respectively, then the distribution is:

1. ***skewed negatively***
2. not skewed
3. skewed positively
4. symmetrical
5. bimodal.

- Which of the following measures of central tendency tends to be most influenced by a extreme score?

a. median

b. mode

c. mean

➢ Which of the following is not a measure of central tendency?

a. mean
b. median
c. mode
d. standard deviation
e. none of these

➢ In a group of 12 scores, the largest score is increased by 36 points.What effect will this have on the mean of the scores?

a. it will be increased by 12 points
b. it will remain unchanged
c. it will be increased by 3 points
d. it will increase by 36 points
e. there is no way of knowing exactly how many points the mean will be increased.

➢ True or False? Generally, a small standard deviation implies that the measurements are clustered close to the mean. ***true***

➢ If you know that the Q1- 1.5 IQR > min. This means what? ***Outlier exists***

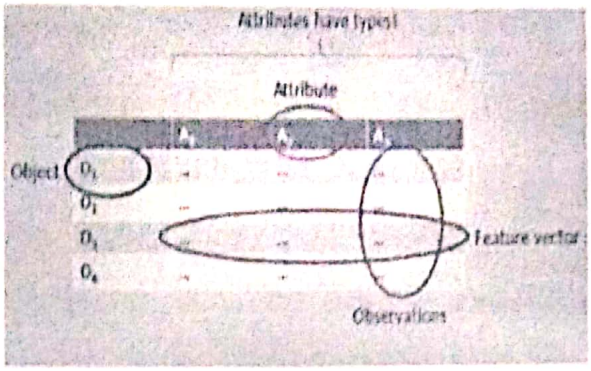
➢ If you know that the Q3+ 1.5 IQR< max. This means what? ***Outlier exist***

➢ List the steps of knowledge discovery (5)

➢ To measure objects similarity/dissimilarity for objects with numerical attributes we use ----- - ***(Manhattan, Euclidean, minkoski)***

➢ True or false and why: All continuous variables are ratio (f)

➢ True or false and why: Attributes are sometimes called variables and objects are sometimes called observations (f)



- True or false and why: Binary variables are sometimes continuous ***false***
- True or false and why: Database mining refers to the process of deriving high-quality information from text. ***False***, ***text mining***
- True or false and why: Dissimilarity matrix stores n data objects that have p attributes as an n-by-p matrix ***false***
- True or false and why: Multimedia Mining is the application of data mining techniques to discover patterns from the Web. ***False ,web mining***
- True or false and why: Data matrix stores a collection of proximities for all pairs of n objects as an n-by-n matrix (***False***)
- True or false and why: Median is a value that occurs most frequently in the attribute values ***false mode***
- True or false and why: Mode is a middle value in set of ordered values ***false median***
- What is the five numbers summery of the data? How is it represented graphically?
- What is the mode of the data? What is the mean of (bimodal, trimodel)
- What is the problem that related to calculate the mean? How you can fix it?
- ..... is an essential process where intelligent methods are applied to extract da patterns.

B)***Data-mining***

C)Text-mining

D) Data selection

➢ ***Data mining can also applied to other forms such as .....***

i)Data-streams ii)Sequence-data iii)Networked-data iv)Text-data v) Spatial data

A)i, ii, iii and v only

B) ii, iii, iv and v only

C) i, iii, iv and v only

D) ***All i, ii, iii, iv and v***

➢ ***The various aspects of data mining methodologies is/are.....***

i) Mining various and new kinds of knowledge
ii) Mining knowledge in multidimensional space
iii) Pattern evaluation and pattern or constraint-guided mining.
iv) Handling uncertainty, noise, or incompleteness of data
A) i, ii and iv only
B) ii, iii and iv only
C) i, ii and iii only
D***) All i, ii, iii and iv***

➢ The full form of KDD is.....

A) Knowledge Database
B) ***Knowledge Discovery Database***
C) Knowledge Data House
D) Knowledge Data Definition

➢ The output of KDD is.....

A) Data
B) Information
C) Query
D) ***Useful information***

➢ If the mean is larger than the median then the data is positively skewed (***true***)

➢ For nominal attributes ,use eqlidean to measure similiarity between objects (***false***)

- Data mining is the process of extracting trivial ,implicit and previously known and potential useful pattern or knowledge from large amount of data (***false***)
- Boxplot used for data smoothing (f)
- The five number summary include mean, median, mode, min, max (***false***)
- True or false and why: Computing the total sales of a company. Is a data mining task?

- **1. Discuss whether or not each of the following activities is a data mining task.**

a. Dividing the customers of a company according to their gender.

ANS This activity is ***not a data mining*** task because it can be done by using a simple database query.

b. Dividing the customers of a company according to their profitability.

ANS This activity is ***not a data mining***. If profitability of each customer is one of the attributes in customer records, using a threshold can divide the customers according to their profitability

c. Computing the total sales of a company.

ANS This activity is ***not a data mining*** task because the total sales can be computed by using simple calculations.

d. Sorting a student database based on student identification numbers.

ANS This activity is ***not a data mining task*** because it is a simple database algorithm.

e. Predicting the outcomes of tossing a (fair) pair of dice.

ANS This activity ***is not a data mining task*** because predicting the outcome of tossing a fair pair of dice is a probability calculation

f. Predicting the future stock price of a company using historical records.

ANS This activity ***is a data mining task***. Historical records of stock price can be used to create a predictive model

g- Monitoring the heart rate of a patient for abnormalities.

ANS This activity is a *data mining task called* anomaly detection.

h. Monitoring seismic waves for earthquake activities.

ANS This activity is a *data mining task.*

i-Extracting the frequencies of a sound wave

ANS This activity is *not a data mining task*

- Know your data useful for data preprocessing
- You can apply mean to nominal attribute (*false*)
- states are not equally important in symmetric binary attributes (*false*)
- Ordinal attributes are qualitative (*true*)
- Can apply median to ordinal attributes (*true*)
- Ratio-scaled have true zero point and cannot be expressed as multiplies (*false*)
- Mean problem is outlier and can be solved by Trimmed mean
- No mode for data means that everyone not repeated (*false*)
- When mean is greater than mode then data expressed as positive and when equal expressed as symmetric normal distribution
- Two lines outside the box extended to Minimum and Maximum in boxplot called whiskers
- Low SD →data observations tend to be very close to the mean (*true*)
- The data matrix is often called a two-mode matrix. The dissimilarity matrix contains one kind of entity (dissimilarities) and so is called a one-mode matrix.
- Dissimilarity matrix for both symmetric and asymmetric binary attributes are the sam (*false*)
- Phone number consider *nominal*

- Concept hierarchies used to compress data (***true***)
- Concept hierarchy can be automatically formed for both numeric and nominal data, but sometimes it is not accurate. ***True***

Problems

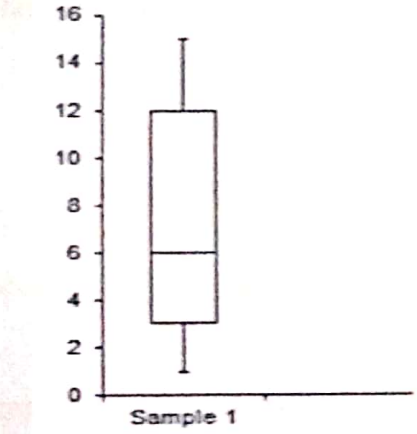
Question 1

Draw the boxplot For the following set

1
1
3
5
6
6
7
8
12
13
15

Min=1
Q1=3
Median=6
Q3=12
Max=15

IQR=12-3=9
1.5×IQR=13.5
Q1- 1.5×IQR=-10.5
Q3+1.5×IQR=25.5



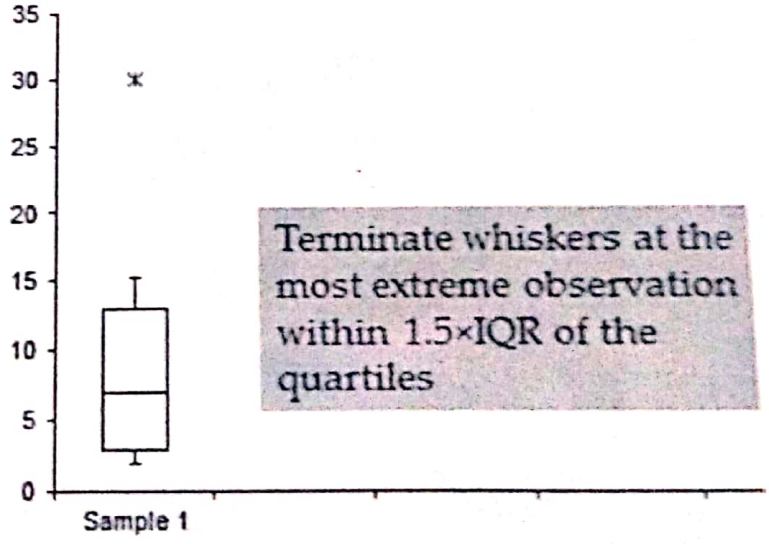
Question 2

Draw the boxplot for the following set

2
3
3
6
7
7
7
9
13
15
30

Min=2
Q1=3
Median=7
Q3=13
Max=30

IQR=10
1.5×IQR=15
Q1- 1.5×IQR=-12
Q2+1.5×IQR=28



Terminate whiskers at the most extreme observation within 1.5×IQR of the quartiles

---

Question 3

Suppose that the data for analysis includes the attribute grade. The grade values for the data tuples are:

4, 5, 9, 11, 12, 13, 13, 13, 13, 14, 15, 15, 16, 17, 18, 18, 19, 20

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

- the mean = 13.61
- The median = (13+14)/2 = 13.5

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

- The mode of the data is 13, the mode is only one value so it's called *unimodal*.

(c) What is the midrange of the data?

- (20+ 4) / 2 = 12

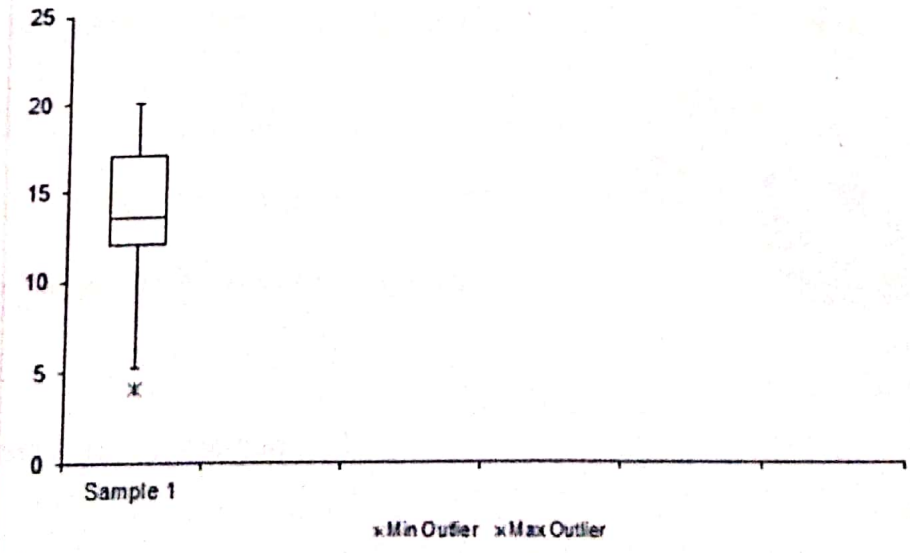
(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

- The first quartile (corresponding to the 25th percentile) of the data is: 12.
  The third quartile (corresponding to the 75th percentile) of the data is: 17.

(e) Give the five-number summary of the data.

- 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: 4,12,13.5,17,20

(f) Show a boxplot of the data.



Question 4

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are

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?

- ***mean*** of the data is: = 809/27 = 30. The ***median*** is: 25.

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

- Bimodal.
- 25 and 35.

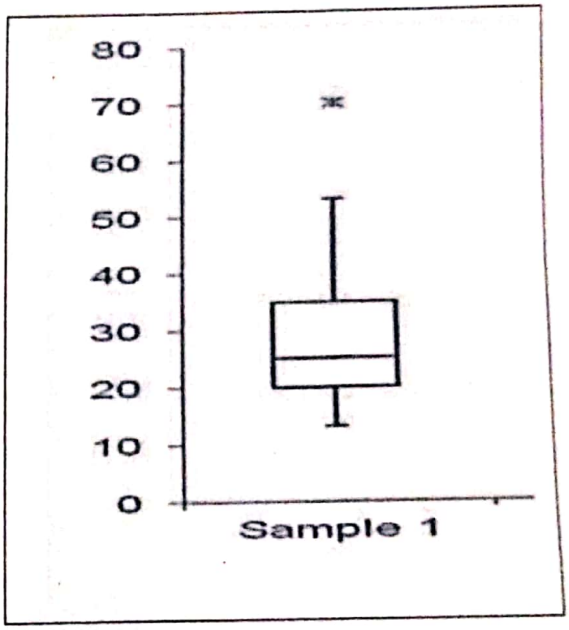
(c) What is the midrange of the data?

- (70+13)=2 = 41.5

(e) Give the five-number summary of the data.

- 13, 20, 25, 35, 70.

(f) Show a boxplot of the data.



Question 5

Age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%Fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	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.

- ***Age***
  - mean = 46.44, median = 51,

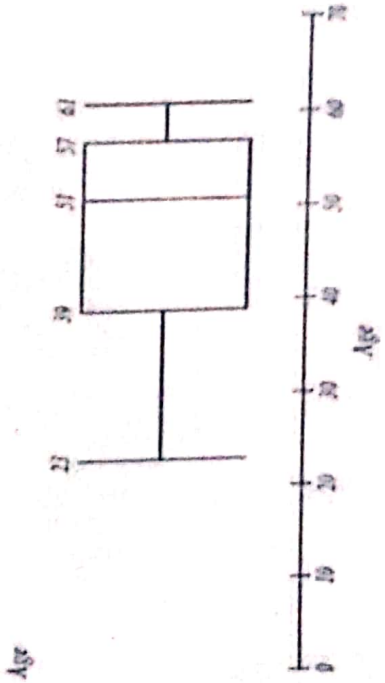
$$\text{Variance } \sigma^2 = \frac{1}{18}(23^2+23^2+27^2+27^2+39^2+41^2+47^2+49^2+50^2+52^2$$
$$+54^2+54^2+56^2+57^2+58^2+58^2+60^2+61^2) - (46.44)^2 = 165.024$$

So standard deviation = 12.846

- The five-number summary → min = 23, Q1 = 39, median = 51, Q3 = 57, max = 61.
  - IQR = 18 → 1.5*IQR = 27

    Q1 - 1.5*IQR = 39 - 27 = 12 Q3 + 1.5*IQR = 57 + 27 = 84

    So no outliers here



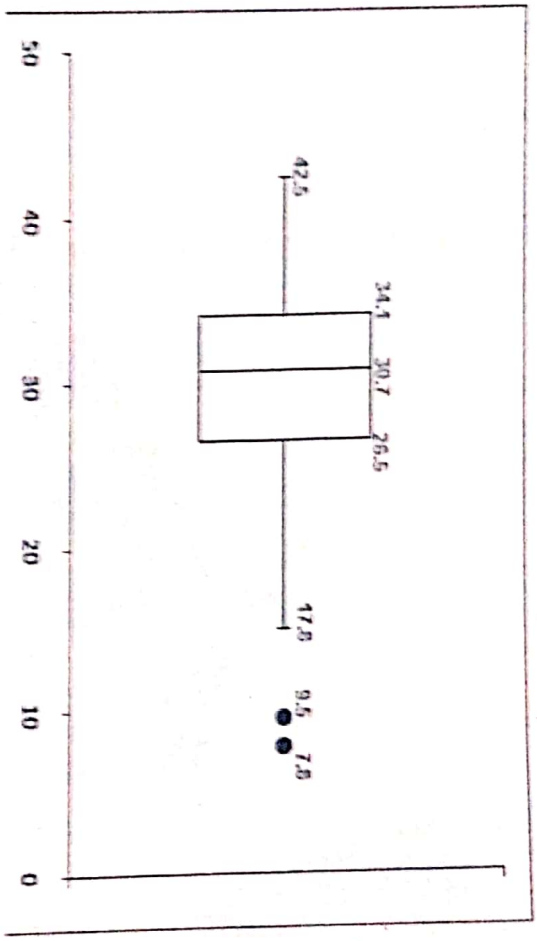
- **%fat**
  - mean = 28.78, median = 30.7,

$$\text{Variance } \sigma^2 = \frac{1}{18}(9.5^2+26.5^2+7.8^2+17.8^2+31.4^2+25.9^2+27.4^2+27.2^2+31.2^2+34.6^2$$
$$+28.8^2+42.5^2+33.4^2+30.2^2+34.1^2+32.9^2+41.2^2+35.7^2) - (28.78)^2 = 80.885$$

So standard deviation = 8.995

- The five-number summary for %fat → min = 7.8, Q1 = 26.5, median = 30.7, Q3 = 34.1, max = 42.5

- IQR = 7.6 → 1.5*IQR = 11.4
- Q1 - 1.5*IQR = 15.1 Q3 + 1.5*IQR = 45.5
- So we have 2 outliers below 15.1 which are 9.5 and 7.8 so boxplot MIN will be 17.8



Question 6

Suppose that a patient record table contains the attributes name, gender, fever, cough, test-1, test-2, test-3, and test-4, where name is an object identifier, gender is a symmetric attribute, and remaining attributes are asymmetric. "**Assume only asymmetric attributes**" calculate dissimilarity matrix.(binary)

name	gender	fever	cough	Test-1	Test-2	Test-3	Test-4
Ali	M	Y	N	P	N	N	N
Adam	M	Y	Y	N	N	N	N
Salma	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Symmetric *Asymmetric*

- Y and P → 1
- N and N → 0
- **Assuming we use only asymmetric attributes**

$$d(Ali, Adam) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Ali, Salma) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Adam, Salma) = \frac{1+2}{1+1+2} = 0.75$$

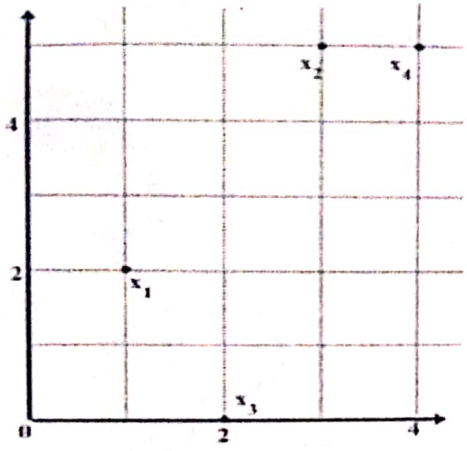
	Ali	Adam	Salma
Ali	0		
Adam	.76	0	
Salma	.33	0.75	0

Measurements suggest that Adam and Selma are unlikely to have a similar disease because they have the highest dissimilarity

---

Question 7

(numeric)



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Distance Matrix (Manhattan)

	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Distance Matrix (Euclidean)

	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Question 8

- Suppose that we have the sample data of Table except the object identifier where test-1 and test-2 are nominal. Let's compute the dissimilarity matrix

ID	Test-1	Test-2
1	Code A	Excellent
2	Code B	Good
3	Code C	Good
4	Code A	Excellent

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 0.5 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Question 9

- Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

- Compute the Euclidean distance between the two objects.
- Compute the Manhattan distance between the two objects.
- Compute the Minkowski distance between the two objects, using h= 3.

• (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects.

$$= \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}$$

$$= \sqrt{45}$$

$$= 6.7082$$

• (22, 1, 42, 10) and (20, 0, 36, 8):

(b) Compute the Manhattan distance between the two objects

$$= |22-20| + |1-0| + |42-36| + |10-8|$$

$$= 11$$

• (22, 1, 42, 10) and (20, 0, 36, 8):

(c) Compute the Minkowski distance between the two objects, using h = 3

$$\sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3}$$

$$= \sqrt[3]{233}$$

$$= 6.1534$$

Determine which of the following Data Mining is:

- Looking up phone number in phone directory. (NO)
- Query processing. (NO)
- Searching for 'Data Mining books' in a search engine. (NO)
- Finding more prevalent names in certain locations. (YES)

ايجاد الاسماء الاكثر انتشارا فى مكان معين

Classify the following attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio), also classify them as binary, discrete, and continuous:

The answers are between the brackets:

- Age in years.*[Discrete, quantitative, ratio]*
- Outlook for weather data (sunny, overcast, rainy). *[Discrete, qualitative, nominal]*
- Temperature in weather data. (Hot, mild, cool)*[Discrete, qualitative, ordinal]*
- Angles as measured in degrees between 0 and 360.*[Continuous, quantitative, interval]*
- Bronze, Silver, and Gold medals as awarded in the Olympics.*[Discrete, qualitative, ordinal]*
- Height above sea level.*[Continuous, quantitative, interval]*
- Number of patients in a hospital.*[Discrete, quantitative, ratio]*
- Calendar dates.*[Discrete, quantitative, interval]*
- Ability to pass light in terms of the following values: opaque, translucent, and transparent.*[Discrete, qualitative, ordinal]*

- Distance from the center of campus.*[Continuous, quantitative, ratio]*
- Density of a substance in grams per cubic centimeter.*[Continuous, quantitative, ratio]*
- Coat check number.*[Discrete, qualitative, nominal]*
- Street Numbers.*[Discrete, qualitative, nominal; unless they can be used for ordering then it would be ordinal].*

Question 7

From those two given tables, give example for each type of attribute with a brief description. Show how may the same attribute have different types based on how it is being represented.

Table 1:

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	True	Yes

Table 2:

Outlook	Temperature (F)	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	73	80	True	Yes

Nominal: Outlook and windy columns in both tables.
Ordinal: Temperature and humidity in table 1.
Interval: Temperature and humidity in table 2.
Ratio: None.

Is data mining another hype (***false***)

Question

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.