

Module No : 02

Introduction to data mining

Data Exploration

Attributes

- An **attribute** is a data field, representing a characteristic or feature of a data object.
- The nouns *attribute*, *dimension*, *feature*, and *variable* are often used interchangeably in the literature.
- The term *dimension* is commonly used in data warehousing.
- Machine learning literature use the term *feature*, while statisticians prefer the term *variable*.
- Data mining and database professionals commonly use the term *attribute*
- Attributes describing a customer object can include: *customer ID*, *name*, and *address*.

Types of Attributes [UQ](#)

- 1. Nominal attributes**
- 2. Binary attributes**
- 3. Ordinal attributes**
- 4. Numeric attributes**
 - a) Interval-scaled attributes
 - b) Ratio-scaled attributes

☑ Types of Attributes 1. Nominal attributes

- Nominal means “relating to names.” The values of a **nominal attribute** are symbols or *names of things*.
- Each value represents some kind of category, code, or state
- The values do not have any meaningful order.
- Example 1: *hair color* and *marital status* are two attributes describing *person* objects, then possible values for *hair color* are *black, brown, blond, red, gray, and white*.
- The attribute *marital status* can take on the values *single, married, divorced, and widowed*.
- Both *hair color* and *marital status* are nominal attributes.
- Example 2: *occupation*, with the values *teacher, dentist, programmer, farmer, and so on*.

☑ Types of Attributes 2. Binary attributes

- A **binary attribute** is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present.
- Binary attributes are referred to as **Boolean** if the two states correspond to *true* and *false*.
- Example: The attribute *medical test* is binary, where a value of 1 means the result of the test for the patient is positive, while 0 means the result is negative.
- there is no preference on which outcome should be coded as 0 or 1 attribute *gender* having the states *male* and *female*.

☑ Types of Attributes 3. Ordinal attributes

- An **ordinal attribute** is an attribute with possible values that have a meaningful order or *ranking* among them, but the magnitude between successive values is not known.
- Example: *drink size* corresponds to the size of drinks available at a fast-food restaurant. This attribute has three possible values: *small*, *medium*, and *large*.
- *grade* (e.g., A++, A+, A, B++, B+, B, C++ and so on)

☑ Types of Attributes 4. Numeric attributes

- A **numeric attribute** is *quantitative*; that is, it is a measurable quantity, represented in integer or real values.
- Numeric attributes can be *interval-scaled* or *ratio-scaled*.

a) Interval-Scaled Attributes

- **Interval-scaled attributes** are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative
- such attributes allow us to compare and quantify the *difference* between values.
- Example 1: temperature (20 degree Celsius is five degrees higher than a temperature of 15 degree Celsius)
- Example 2: calendar dates (the years 2002 and 2010 are eight years apart)

☑ Types of Attributes 4. Numeric attributes

b) Ratio-Scaled Attributes

- A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point i.e. if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value
- the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode
- Example 1: year_of_experience
- Example 2: no-of-words (in a document)
- Example 3: weight, height, latitude and longitude

☑ Statistical Description of data

- For data pre-processing to be successful, it is essential to have an overall picture of your data.
- Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
- Following are the different ways to describe data statistically
 1. Mean
 2. Median
 3. Mode
 4. Midrange
 4. Range
 5. Quartiles
 6. Interquartile Range
 7. Five-Number Summary
 8. Boxplots
 9. Outliers
 10. Variance
 11. Standard Deviation
 12. Histograms
 13. Scatter Plots
 14. Data Correlation

☑ Statistical Description of data : 1. Mean (Average Value)

Let X_1, X_2, \dots, X_n be a set of N values or observations, such as for some numeric attribute X . The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}.$$

Example

Mean. Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq.

$$\begin{aligned}\bar{x} &= \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} \\ &= \frac{696}{12} = 58.\end{aligned}$$

Thus, the mean salary is \$58,000.

☑ Statistical Description of data : 2. Median (middle value)

Let X_1, X_2, \dots, X_n be a set of N values or observations, such as for some numeric attribute X , like salary. The median of this set of values is

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})_l}{\text{freq}_{\text{median}}} \right) \text{width},$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum \text{freq})_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $\text{freq}_{\text{median}}$ is the frequency of the median interval, and width is the width of the median interval.

Example

Median. Let's find the median of the data from Example

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

The data are already sorted in increasing order. There is an even number of observations (i.e., 12); therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the average of the two middlemost values as the median; that is, $\frac{52+56}{2} = \frac{108}{2} = 54$. Thus, the median is \$54,000.

☑ Statistical Description of data : 3. Mode (most common value)

Let X_1, X_2, \dots, X_n be a set of N values or observations, such as for some numeric attribute X .

Data set : 30,36,47,50,52,52,56,60,63,70,70,110

The mode of this set of values is : (the values repeating maximum times)

This set of data is bimodal i.e. there are two modes 52 and 70

(30,36,47,50,**52,52**,56,60,63,**70,70**,110)

☑ Statistical Description of data : 4. Midrange

Let X_1, X_2, \dots, X_n be a set of N values or observations, such as for some numeric attribute X .

Data set : 30,36,47,50,52,52,56,60,63,70,70,110

The midrange is the average of largest and smallest values in the set.

The midrange of this set of values is $(30+110)/2=70$

(**30**,36,47,50,52,52,56,60,63,70,70,**110**)

Statistical Description of data : 5. Range

Let X_1, X_2, \dots, X_n be a set of N values or observations, such as for some numeric attribute X .

The range of the set is the difference between the largest (max) and smallest (min) values.

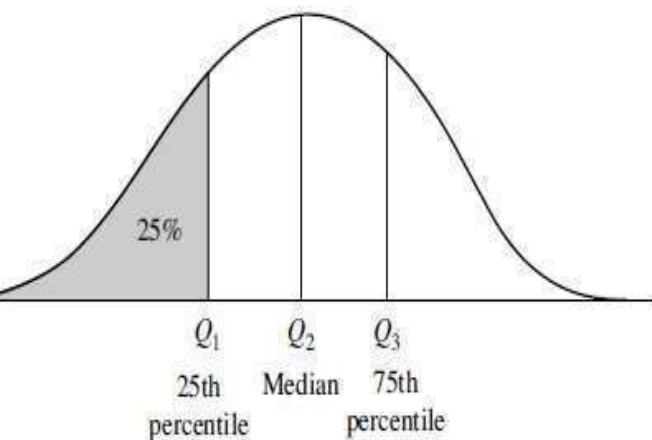
e.g. (**30**, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, **110**)

Range of this data set is $110 - 30 = 80$

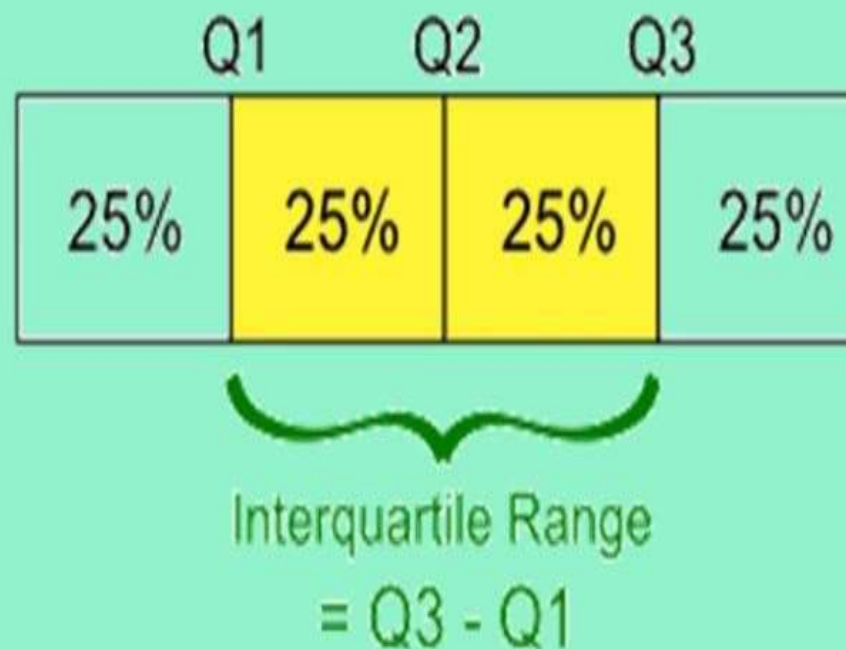
☑ Statistical Description of data : 5. Quartiles

Let X_1, X_2, \dots, X_n be a set of N values or observations, such as for some numeric attribute X .

We can pick certain data points so as to split the data distribution into equal-size consecutive sets, as shown in figure



ot of the data distribution for some attribute X . The quantiles plotted are quartiles. The
e quartiles divide the distribution into four equal-size consecutive subsets. The second
tile corresponds to the median.



☑ Statistical Description of data : 5. Quartiles

Q1=Lower Quartile

Q2= Median

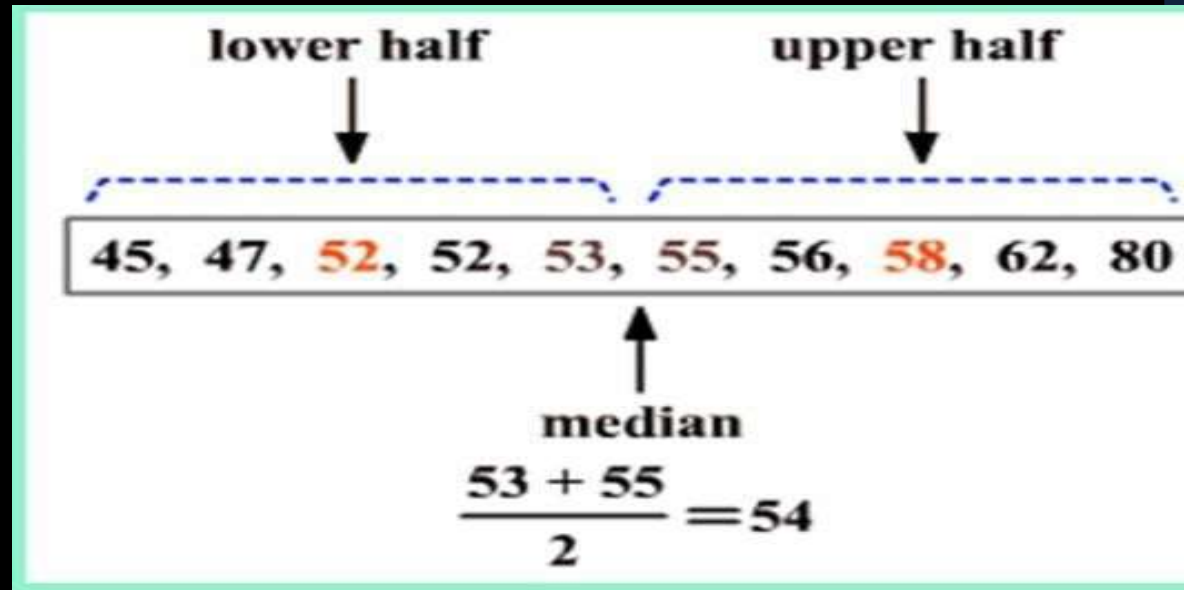
Q3=Upper Quartile

In this e.g.

Q1=52

Q2=54

Q3=58



- The first quartile, denoted by Q1, is the 25th percentile. It cuts off the lowest 25% of the data.
- The third quartile, denoted by Q3, is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data.
- The second quartile is the 50th percentile. As the median, it gives the centre of the data distribution.

☑ Statistical Description of data : 6. Interquartile Range(IQR)

- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data.
- This distance is called the **interquartile range (IQR)** and is defined as $IQR = Q3 - Q1$

☑ Statistical Description of data : 6. Interquartile Range(IQR)

2, 3, 3, 4, 5, 6, 8, 9

Finding the Upper Quartile:

2, 3, 3, 4, 5, 6, 8, 9

Median of the Upper Half

~~5~~, 6, 8, ~~9~~

$$\frac{6 + 8}{2} = \frac{14}{2} = 7$$

Finding the Lower Quartile:

2, 3, 3, 4, 5, 6, 8, 9

Median of the Lower Half

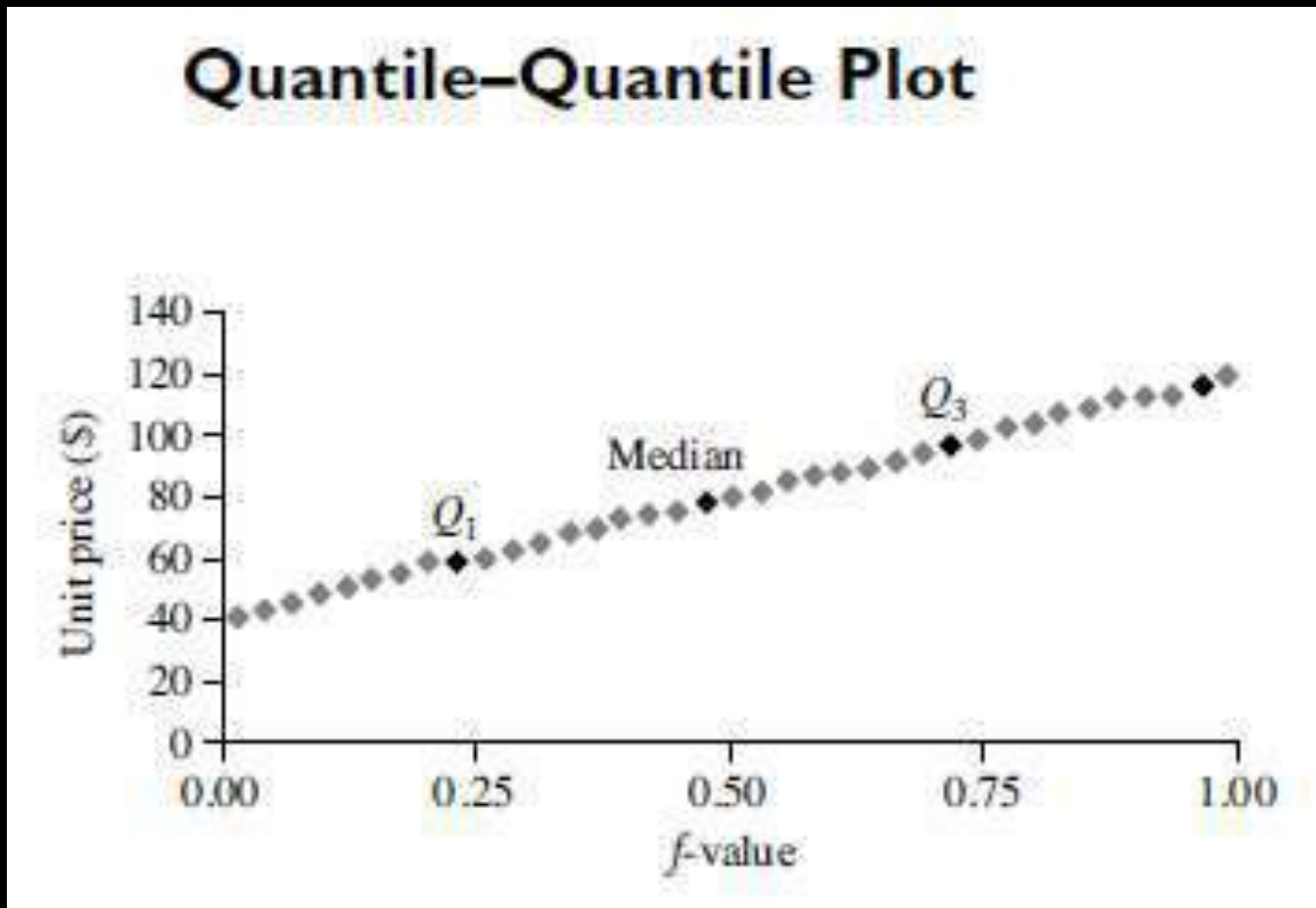
~~2~~, 3, 3, ~~4~~

$$\frac{3 + 3}{2} = \frac{6}{2} = 3$$

IQR = Upper Quartile - Lower Quartile

$$7 - 3 = 4$$

☑ Statistical Description of data :



☑ Statistical Description of data : 7. Five-Number Summary

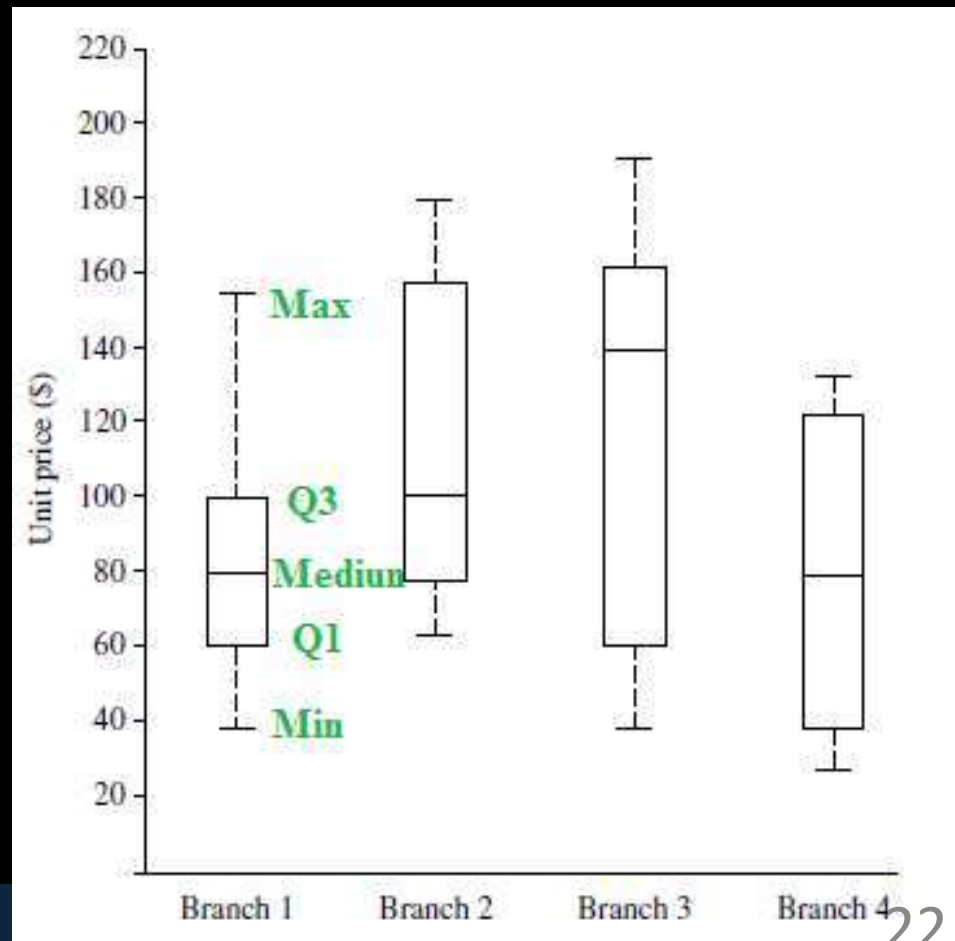
- The **five-number summary** of a distribution consists of the median (Q_2), the quartiles Q_1 and Q_3 , and the smallest and largest individual observations, written in the order of *Minimum, Q_1 , Median, Q_3 , Maximum*.
- E.g. 2,3,3,4,5,6,8,9
- Minimum = 2
- $Q_1 = 3$
- Median = 4.5
- $Q_3 = 7$
- Maximum = 9

☑ Statistical Description of data : 8. Boxplots

- **Boxplots** are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:
- E.g. 2,3,3,4,5,6,8,9

The five-number summary

- Minimum = 2
- Q1 = 3
- Median = 4.5
- Q3 = 7
- Maximum = 9



☑ Exercise

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 (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.

☑ Exercise (answers)

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

The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = 809/27 = 30$ (Equation 2.1). The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

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

This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

- (c) What is the *midrange* of the data?

The midrange (average of the largest and smallest values in the data set) of the data is: $(70 + 13)/2 = 41.5$

- (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: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

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

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.

☑ Statistical Description of data : 9. Outliers

- **Outliers** are an extremely high or extremely low values in the data set. We can identify an outliers by following
- The values greater than $Q3 + 1.5(IQR)$
- The values less than $Q1 - 1.5(IQR)$

Phone calls
recieved 

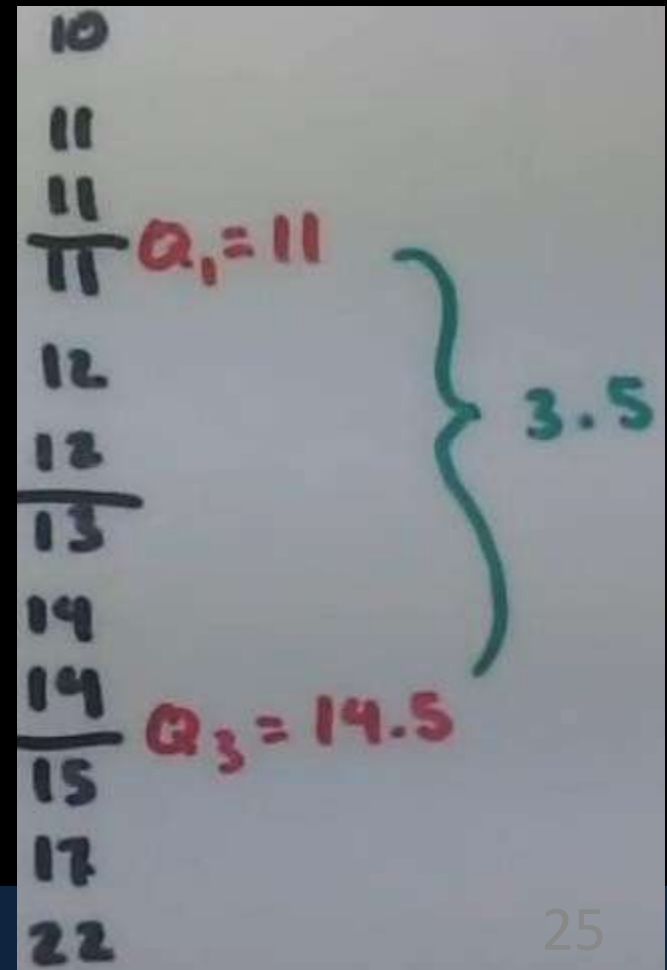
10	12	11	15
11	14	13	17
12	22	14	11

$Q3 + 1.5(IQR)$ = higher outlier

$$14.5 + 1.5(3.5) = 19.75$$

$Q1 - 1.5(IQR)$ = lower outlier

$$11 - 1.5(3.5) = 5.75$$



☑ Statistical Description of data : 10. Variance and Standard deviation

data set --> 17 15 23 7 9 13

sample variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

sample variance

$$s^2 = \frac{166}{6-1} = 33.2$$

sample standard deviation

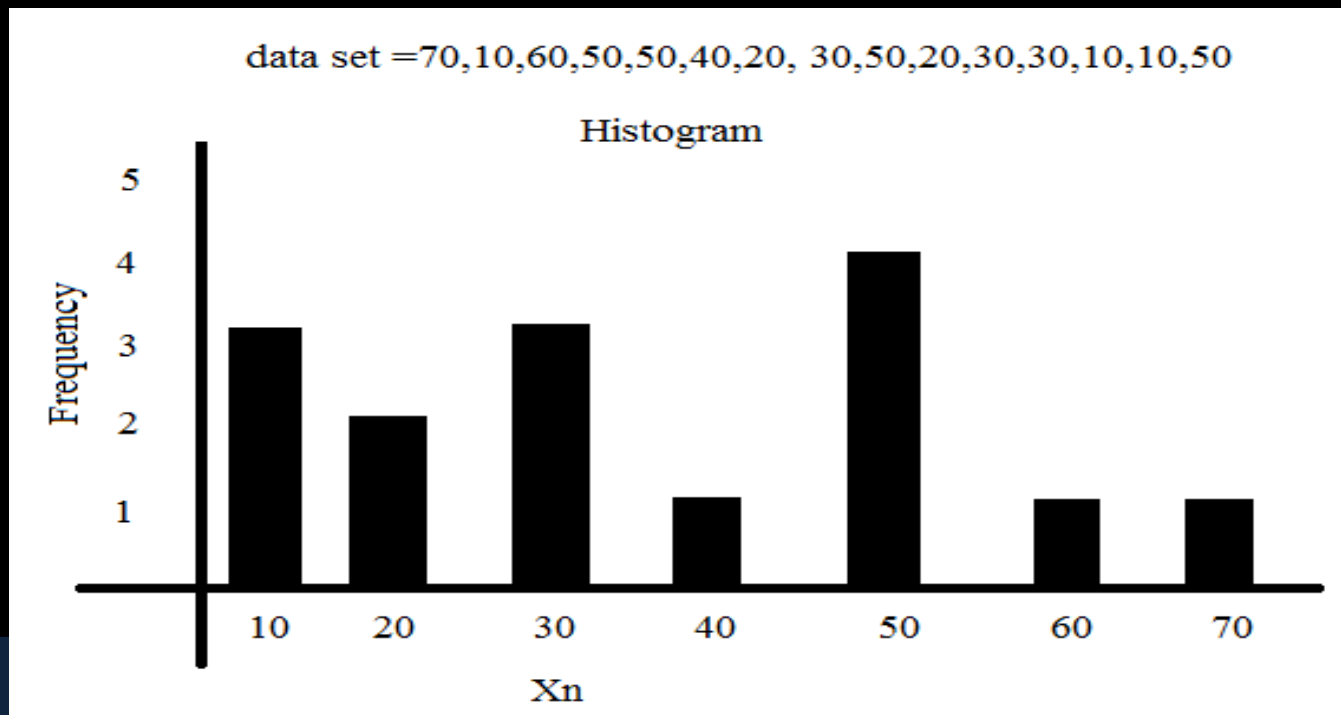
$$s = 5.76 \text{ (square root of 33.2)}$$

x	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
17	14	3	9
15	14	1	1
23	14	9	81
7	14	-7	49
9	14	-5	25
13	14	-1	1
84		0	166
			$\sum (x_i - \bar{x})^2$

☑ Statistical Description of data : 12. Histograms

- —Histos means pole, and —gam means chart, so a histogram is a chart of poles
- Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X

data set = 70,10,60,50,50,40,20, 30,50,20,30,30,10,10,50



☑ Statistical Description of data : 12. Histograms

Draw histogram for THE following data set

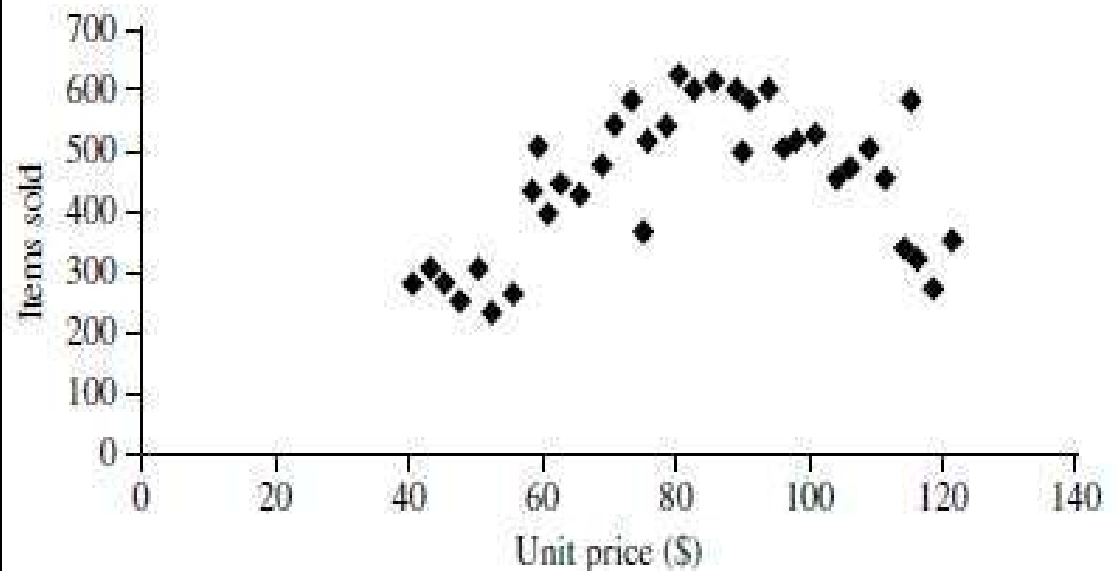
Transaction ID	Items Brought
T1	F,A,D,B
T2	D,A,C,E,B
T3	C,A,B,E
T4	B,A,D

☑ Statistical Description of data : 13. Scatter Plots

- A scatter plot is one of the most effective graphical method for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
- To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

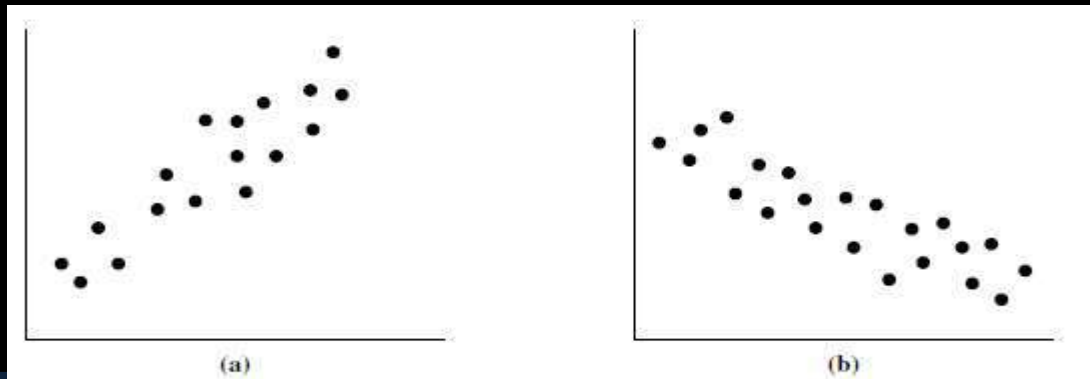
Unit price (\$)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



A scatter plot for the data set.

☑ Statistical Description of data : 14. Data Correlations

- Two attributes, X , and Y , are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated)
- Figure shows examples of positive and negative correlations between two attributes.
- If the plotted points pattern slopes from lower left to upper right, this means that the values of X increase as the values of Y increase, suggesting a positive correlation
- If the pattern of plotted points slopes from upper left to lower right, the values of X increase as the values of Y decrease, suggesting a negative correlation



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

☑ Statistical Description of data : 14. Data Correlations

Correlation Coefficient (r)

An agricultural research organization tested a particular chemical fertilizer to try to find out whether an increase in the amount of fertilizer used would lead to a

Corresponding increase in the food supply.

Fertilizer (lbs)	2	1	3	2	4	5	3
Bushels of Beans	4	3	4	3	6	5	5

x
 y

x	y	xy	x ²	y ²
2	4	8	4	16
1	3	3	1	9
3	4	12	9	16
2	3	6	4	9
4	6	24	16	36
5	5	25	25	25
3	5	15	9	25

Σx	Σy	Σxy	Σx^2	Σy^2
20	30	93	68	136

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{7(93) - (20)(30)}{\sqrt{[7(68) - (20)^2][7(136) - (30)^2]}}$$

$$= \frac{651 - 600}{\sqrt{[476 - 400][952 - 900]}}$$

$$= \frac{51}{\sqrt{[76][52]}} = \frac{51}{\sqrt{3952}} = \frac{51}{62.8649} = 0.811$$

✓ Exercise

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	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.
- (d) Calculate the *correlation coefficient* (Person's product moment coefficient). Are these two variables positively or negatively correlated?

✓ Exercise: Answers

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

For the variable *age* the mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable *%fat* the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

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

See Figure 2.1.

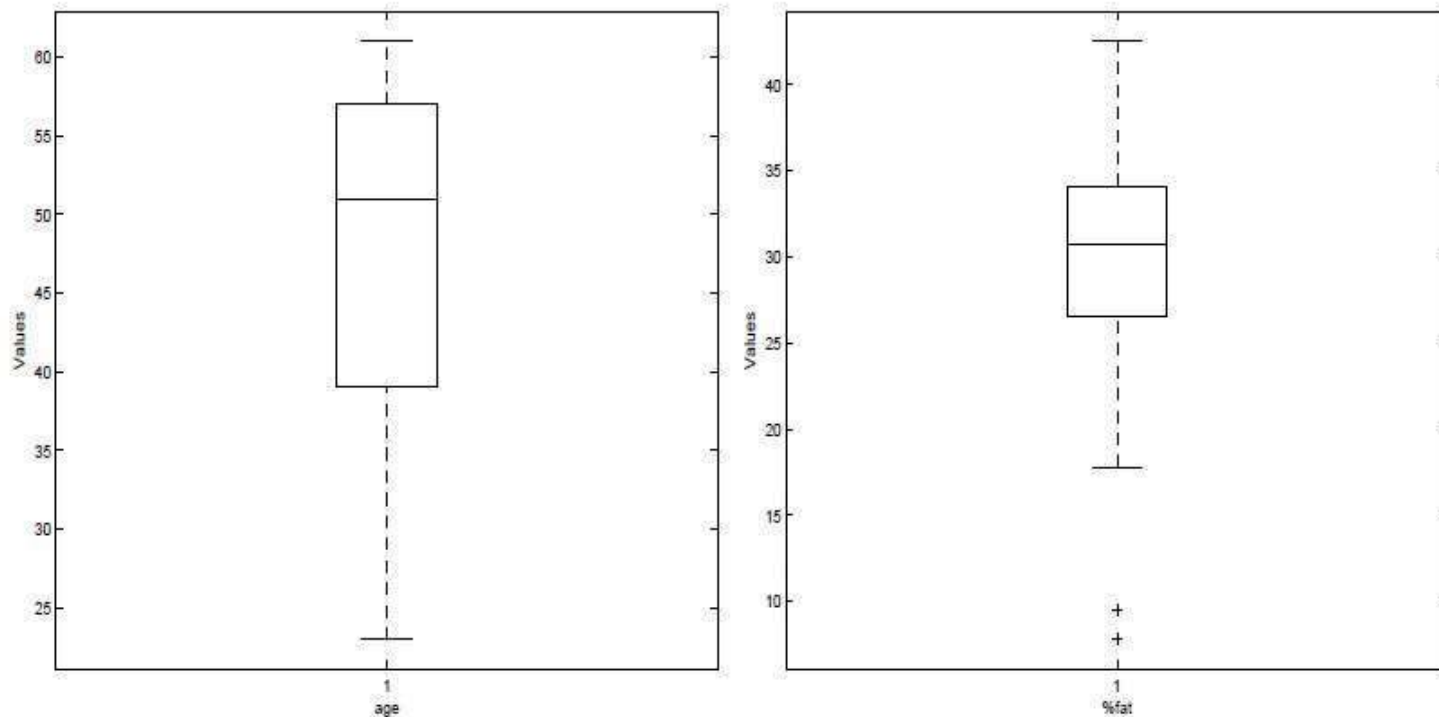


Figure 2.1: A boxplot of the variables *age* and *%fat* in Exercise 2.9.

✓ Exercise: Answers

- (c) Draw a *scatter plot* and a *q-q plot* based on these two variables.
See Figure 2.2.

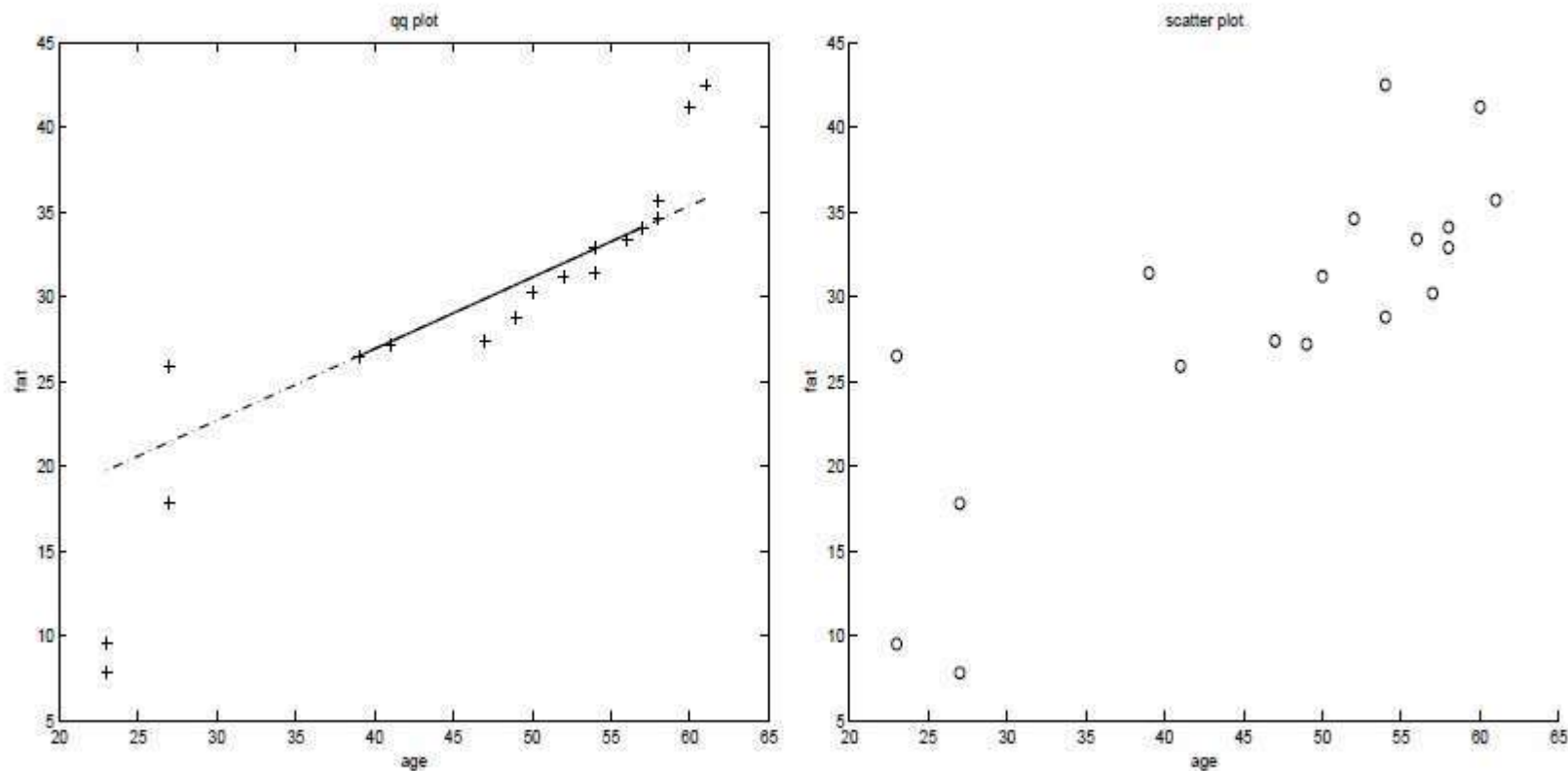


Figure 2.2: A *q-q plot* and a *scatter plot* of the variables *age* and *%fat* in Exercise 2.9.

Exercise: Answers

- (e) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

The *correlation coefficient* is 0.82. The variables are positively correlated.

☑ Measuring data similarity and dissimilarity

- In data mining applications(like clustering, classifications etc.) We are interested in comparison of objects on the basis of their similarities and dissimilarities
- Similarities and dissimilarities can be measured by using following ways
 1. Data Matrix
 2. Dissimilarity Matrix
 3. Minkowski Distance
 - a) Manhattan(City block) Distance
 - b) Euclidean Distance
 - c) Supremum Distance
 4. Cosine similarity

☑ Measuring data similarity and dissimilarity

Data Matrix:

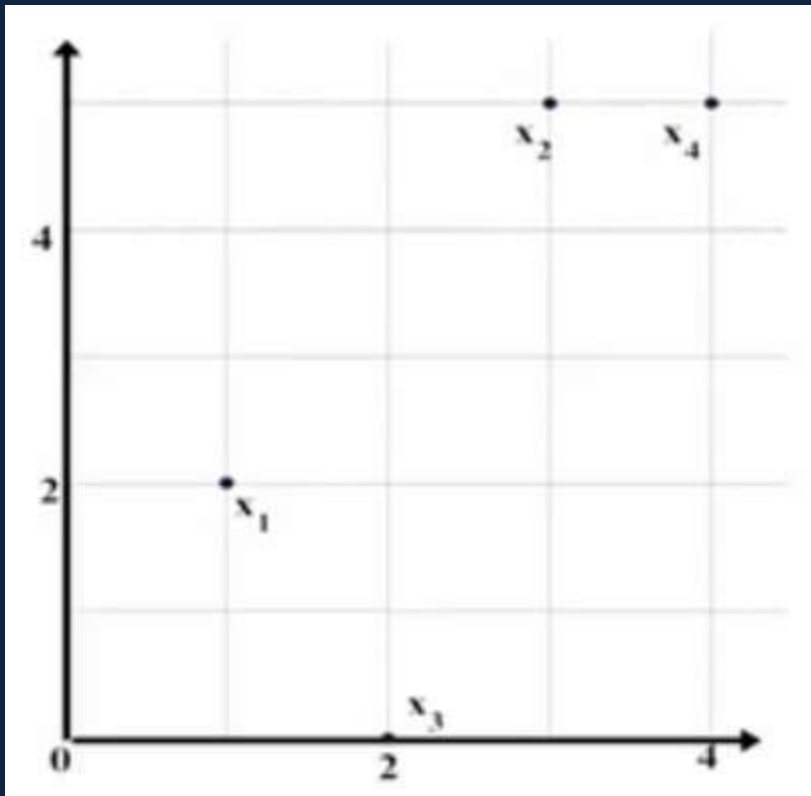
- This structure stores the n data objects in the form of a relational table, or n -by- p matrix (n objects, p attributes)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

☑ Measuring data similarity and dissimilarity

Data Matrix:

- Data points $X_1(1,2)$, $X_2(3,5)$, $X_3(2,0)$, $X_4(4,5)$



Data Matrix

point	attribute1	attribute2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

☑ Measuring data similarity and dissimilarity

Dissimilarity Matrix:

- This structure stores a collection of proximities that are available for all pairs of n objects.
- It is often represented by an n -by- n table

$$\begin{bmatrix} 0 & & & & \\ d(2, 1) & 0 & & & \\ d(3, 1) & d(3, 2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n, 1) & d(n, 2) & \dots & \dots & 0 \end{bmatrix},$$

Dissimilarity Matrix:

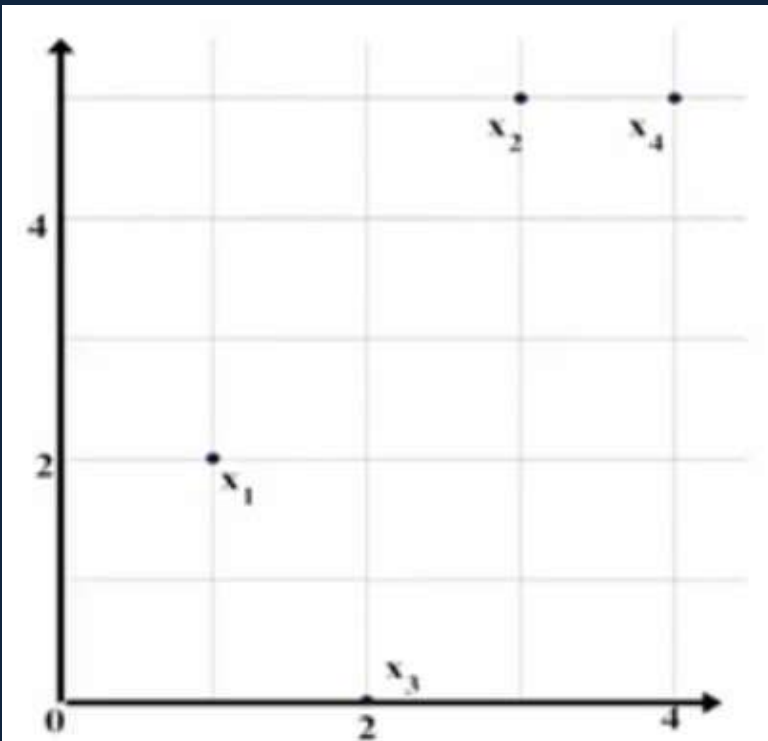
- Data points $X1(1,2), X2(3,5), X3(2,0), X4(4,5)$

Euclidean distance Let $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ be two objects described by p numeric attributes. The Euclidean distance between objects i and j is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}.$$

Data Matrix

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



$x_1 = (1, 2)$ and $x_2 = (3, 5)$

The Euclidean distance = $\sqrt{2^2 + 3^2} = 3.61$.

Dissimilarity Matrix (by Euclidean Distance)

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

Minkowski Distance

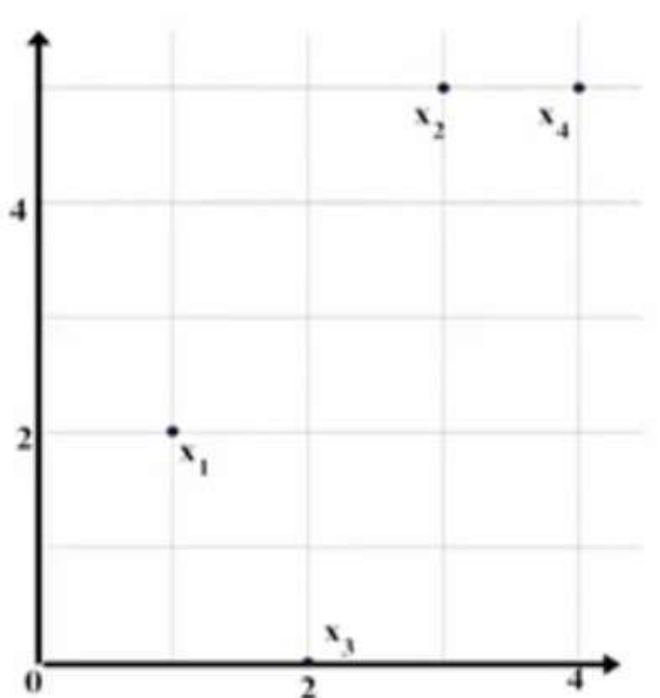
Data points X1(1,2), X2(3,5),X3(2,0),X4(4,5)

a)Manhattan(City block) Distance b)Euclidean Distance c)Supremum Distance

Manhattan (or city block) distance,

Let $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ be two objects described by p numeric attributes.

It is defined as $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$.



$x_1 = (1, 2)$ and $x_2 = (3, 5)$

The Manhattan distance between X1 and X2 is

$$\begin{aligned} d(X1, X2) &= |1-3| + |2-5| \\ &= |-2| + |-3| = 2 + 3 = 5 \end{aligned}$$

Manhattan (L_1)

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

Minkowski Distance

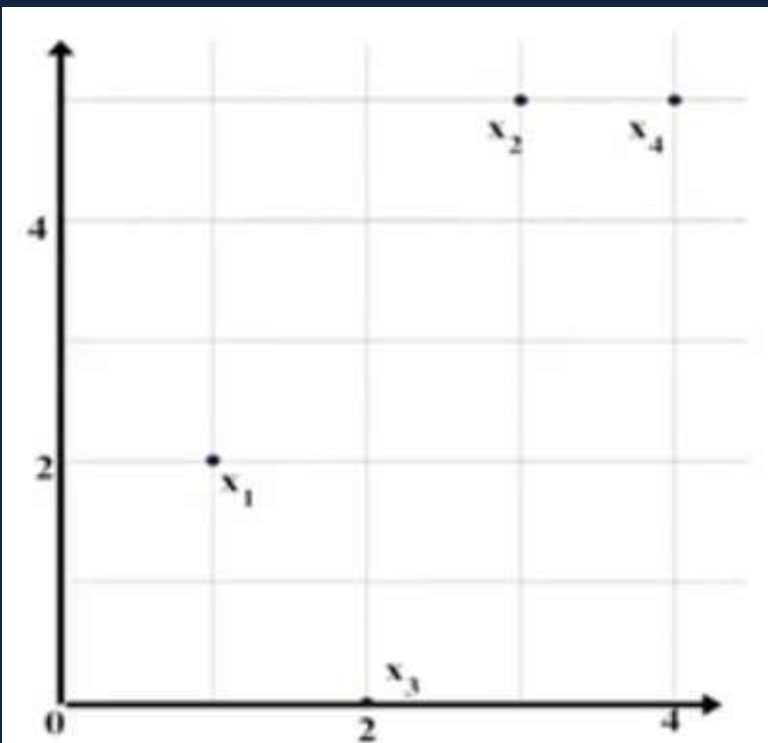
a) Manhattan(City block) Distance b) Euclidean Distance c) Supremum Distance

Euclidean distance Let $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ be two objects described by p numeric attributes. The Euclidean distance between objects i and j is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}.$$

Data Matrix

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



$x_1 = (1, 2)$ and $x_2 = (3, 5)$

The Euclidean distance = $\sqrt{2^2 + 3^2} = 3.61$.

Dissimilarity Matrix (by Euclidean Distance)

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

Minkowski Distance

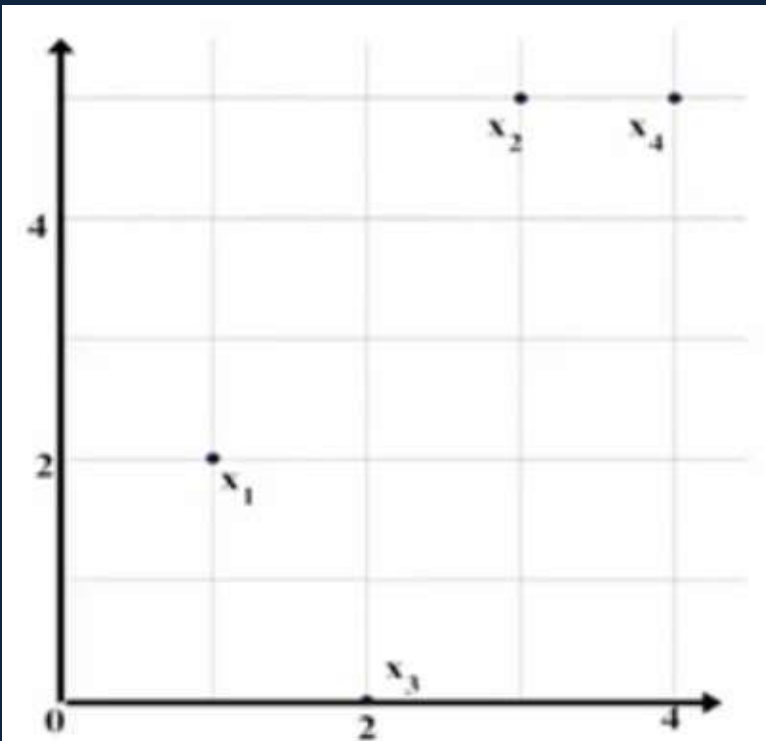
a) Manhattan (City block) Distance b) Euclidean Distance c) Supremum Distance

The supremum distance

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|.$$

Data Matrix

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



$x_1 = (1, 2)$ and $x_2 = (3, 5)$

The supremum distance between x_1 and x_2 is =
maximum of 2 and 3 = 3

Supremum (L_{∞})

L_{∞}	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3	0		
$x3$	2	5	0	
$x4$	3	1	5	0

☑ Measuring data similarity and dissimilarity

Cosine similarity

Cosine similarity

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|},$$

where $\|x\|$ is the Euclidean norm of vector $x = (x_1, x_2, \dots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$. Similarly, $\|y\|$

Example

Cosine similarity between two term-frequency vectors. Suppose that x and y are the first two term-frequency vectors. That is, $x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ and $y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$. How similar are x and y ?

$$\begin{aligned} x^t \cdot y &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 \\ &\quad + 0 \times 0 + 0 \times 1 = 25 \end{aligned}$$

$$\|x\| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$\|y\| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

$$\text{sim}(x, y) = 0.94$$

☑ Measuring data similarity and dissimilarity

Cosine similarity

Cosine similarity

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|},$$

where $\|x\|$ is the Euclidean norm of vector $x = (x_1, x_2, \dots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$. Similarly, $\|y\|$

X=(2,1,3,2,4,5,3) ; Y=(4,3,4,3,6,5,5) how similar are x and y?

Exercise : Minkowski Distance

a) Manhattan (City block) Distance b) Euclidean Distance c) Supremum Distance

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

- (a) Compute the *Euclidean distance* between the two objects.
- (b) Compute the *Manhattan distance* between the two objects.
- (c) Compute the *supremum distance* between the two objects.