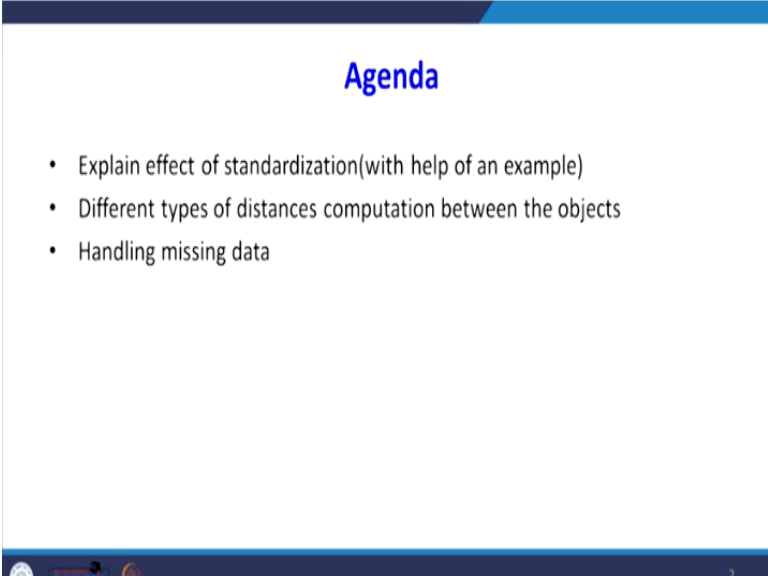


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 50
Clustering Analysis: Part II

In my previous lecture, we have started about introduction to cluster analysis, and I have explained how to handle interval types of data. Then I have started about the importance of standardization. In this lecture we will see that what is the effect of standardization because sometime standardization may mislead your clustering structure and I will explain different types of distances computation between the objects.

(Refer Slide Time: 00:56)



Agenda

- Explain effect of standardization(with help of an example)
- Different types of distances computation between the objects
- Handling missing data

Because for different types of data set, there are different ways to compute the distances, so that I will explain the many time when we collect the data. It is not necessary that we will collect all the data some time there may be a missing data. If the data is missed how to hold to handle that, that also will cover in this lecture.

(Refer Slide Time: 01:13)

Example

- Lets take four persons A, B,C, D with following age and height:

Person	Age (yr)	Height (cm)
A	35	190
B	40	190
C	35	160
D	40	160

TABLE: 1

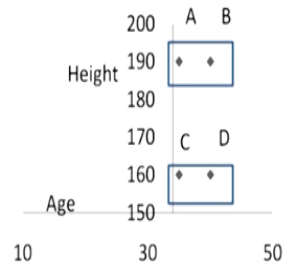


FIGURE: 1

Finding Groups in Data: An Introduction to Cluster Analysis
Author(s): [Leonard Kaufman](#), [Peter J. Rousseeuw](#)
March 1990, John Wiley & Sons, Inc.

Now let us see the effect of standardization I have taken one simple problem with a numerical example. This problem is taken from this book Finding Groups in Data: An Introduction to Cluster Analysis by Leonard Kaufman and Peter Rousseeuw; it is a John Wiley publishers. There are 4 persons and your age yet in terms of year and height in terms of centimeter is given. Suppose if you take age on horizontal axis and height on vertical axis, you can mark this all four persons A, B, C, D.

So what you were able to understand that is a distinct cluster is there. Because A, B is one group one cluster C, D in another cluster. Now the same data let us standardize after standardizing again we will go for clustering let us see how it appears.

(Refer Slide Time: 02:08)

Example

- In Figure 1 we can see distinct clusters
- Let us standardize the data of Table 1
- The mean age equals $m_1 = 37.5$ and the mean absolute deviation of the first variable works out to be $s_1 = (2.5 + 2.5 + 2.5 + 2.5)/4 = 2.5$
- Therefore, standardization converts age 40 to +1 $((40 - 37.5)/2.5 = 1)$ and age 35 $((35 - 37.5)/2.5 = -1)$ to -1
- Analogously, $m_2 = 175$ cm and $s_2 = (15 + 15 + 15 + 15)/4 = 15$ cm, so 190 cm is standardized to +1 and 160 cm to -1

In figure 1 we can see the distinct clusters, let us standardize the data of table 1. For standardizing we should know the mean and standard deviation, standard deviation otherwise mean absolute deviation. So that mean of age equals to $m_1 = 37.5$ just by adding all the ages and divided by the number of data set and the mean absolute deviation is not standard deviation it is mean absolute deviation of the first variable works out to be $S_1 = 2.5$.

How we are finding mean absolute deviation that variable minus mean for example $35 - 37.5$ for second variable $40 - 37.5$ we have to take only the positive value. There are four data set, so the mean absolute deviation is 2.5. Therefore, the standardization convert 40 to +1 how we got to 40 is converted standardized to 1 we know that this is $(x - \mu)$ divided by S . So x is 40 μ that is m is 37.5 divided by mean absolute deviation 2.5 , $= 1$.

And same way age 35 is standardized to -1 how we got the -1, $35 - \text{mean}$ divided by mean absolute deviation. So it is -2.5 divided by 2.5 it is -1 the same way for the variable m_2 the mean is 175 and mean absolute deviation for variable 2 is 15. So each variable in the second column also standardized for example 190 centimeter is standardized to +1 and same way 160 centimeters is standardized to -1.

(Refer Slide Time: 03:53)

Example

- The resulting data matrix, which is unitless, is given in Table 2
- Note that the new averages are zero and that the mean deviations equal 1

Table 2

Person	Variable 1	Variable 2
A	1	1
B	-1	1
C	1	-1
D	-1	-1

- Even when the data are converted to very strange units standardization will always yield the same numbers



The result data matrix which is unitless because below standardized is given in the table 2. Note that the new averages are 0 and the mean deviations equal to 1. So this table 2 shows that these standardized. Table for each variable is variable 1 and variable 2. Even when the data are converted into various strange units standardization will always yield the same numbers that is the advantage of standardization.

(Refer Slide Time: 04:25)

Example

- Plotting the values of Table 2 in Figure 2 does not give a very exciting result
- Figure 2 shows no clustering structure because the four points lie at the vertices of a square
- One could say that there are four clusters, each consisting of a single point, or that there is only one big cluster containing four points
- Here standardizing is no solution

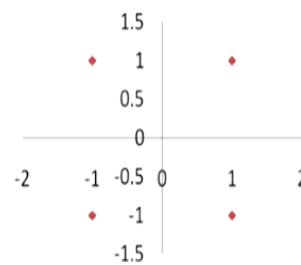


FIGURE: 2

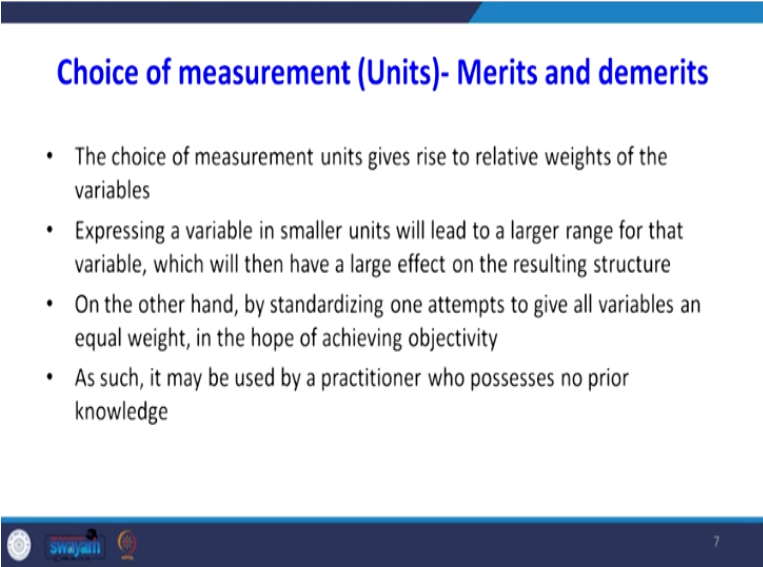


Now plotting the values of table 2 in the figure 2 does not give any very exciting result. So what do you have done? In the previous table we have the standardized values for both the variables. So when you plot it there are 4 points are appearing. So this points is not giving any useful result so figure 2 shows no clustering structure because 4 points lay out the vertices of a square. One

could say that there are 4 clusters; each consisting of single point are that there is only one big cluster containing 4 points.

Here standardization is no solution. So what we have seen many times when you go for standardization, the standardization may not give the useful result that is what this example shows.

(Refer Slide Time: 05:16)



Choice of measurement (Units)- Merits and demerits

- The choice of measurement units gives rise to relative weights of the variables
- Expressing a variable in smaller units will lead to a larger range for that variable, which will then have a large effect on the resulting structure
- On the other hand, by standardizing one attempts to give all variables an equal weight, in the hope of achieving objectivity
- As such, it may be used by a practitioner who possesses no prior knowledge

swayam 7

Now let us look at the choice of measurements. Here the measurement means that units of that variable. What is the merits and demerits? The choice of measurement units gives rise to relative weight of variables, expressing a variable in smaller units will lead to large range for that variable, which will then have a large effect on the resulting structure. So what will happen if variable is in smaller units? So, that will give a larger effect in the; your clustering result.

On the other hand, by standardizing one attempts to give all variables an equal weight in the hope that achieving objectivity. As such it may be used for practitioners who possesses no prior knowledge. So the benefit of standardization is that anybody those who are not having any prior knowledge about the problem also can do with the help of standardized variables. They can do the cluster analysis because there is a unitless.

(Refer Slide Time: 06:18)

Choice of measurement- Merits and demerits

- However, it may well be that some variables are intrinsically more important than others in a particular application, and then the assignment of weights should be based on subject-matter knowledge
- On the other hand, there have been attempts to devise clustering techniques that are independent of the scale of the variables

However, it may well be that some variables are intrinsically more important than others in a particular application and then the assignment of weight should be based on the subject matter knowledge. Every time because standardization is giving equal weight some time some variables are more important. So for that variable with the help of experts, we can give a higher weightage for that variable.

On the other hand, there have been attempts to devise clustering techniques that are independent of scale of the variables. There are many techniques people are trying to come with a different clustering model.

(Refer Slide Time: 06:55)

Distances computation between the objects

- The next step is to compute distances between the objects, in order to quantify their degree of dissimilarity
- It is necessary to have a distance for each pair of objects i and j .
- The most popular choice is the Euclidean distance:

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

- When the data are being standardized, one has to replace all x by z in this expression
- This Formula corresponds to the true geometrical distance between the points with coordinates (x_{i1}, \dots, x_{ip}) and (x_{j1}, \dots, x_{jp})

Distances computation between objects. The next step is to compute distances between the objects in order to quantify their degree of dissimilarity. It is necessary to have a distance for each pair of objects i and j . The most popular choice is the Euclidean distance. What is this Euclidean distance? The distance between variable i , $j = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$ up to $(x_{ip} - x_{jp})^2$.

When the data are being standardized one has to replace all x by z in this expression if you are standardizing instead of x you have to use z . This formula corresponds to the true geometrical distance between points with the coordinates x_{i1} up to x_{ip} and x_{j1} up to x_{jp} .

(Refer Slide Time: 07:55)

Example

- let us consider the special case with $p = 2$ (Figure 3)
- Figure shows two points with coordinates (x_{i1}, x_{i2}) and (x_{j1}, x_{j2})
- It is clear that the actual distance between objects i and j is given by the length of the hypotenuse of the triangle, yielding expression in previous slide by virtue of Pythagoras' theorem

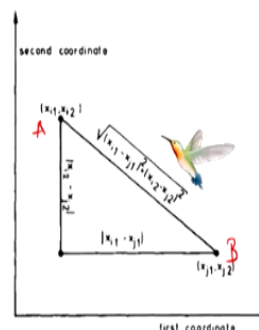


Figure 3: Illustration of the Euclidean distance formula

See the Euclidean distance suppose if you want to move from point A to B see this is point A and B let us find out the concept behind the Euclidean distance. Suppose if we want to move if you want to point A to B you can directly you can fly from one point to because the birds will fly from A to B. So that distances called Euclidean distance. Let us consider the special case with $p = 2$ where there are only two variable.

Figure shows two points with the coordinates x_{i1} , x_{i2} and x_{j1} , x_{j2} . It is clear that the actual distance between objects i and j is given by the length of the hypotenuse of the triangle yielding expression in previous slide by virtue of Pythagoras theorem. So this formula is nothing but the

hypotenuse. So this is as per the Pythagoras theorem so square of adjacent side and square of opposite side equal to square of hypotenuse.

(Refer Slide Time: 09:02)

Distances computation between the objects

- Another well-known metric is the city block or Manhattan distance, defined by:

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

Let us go to the next distance measures that is Manhattan distance. It is a another well-known metric is the city block or Manhattan distance. It is given by x modulus value of $x_{i1} - x_{j1} + x_{i2} - x_{j2}$ modulus value only the positive values up to $x_{ip} - x_{jp}$. Suppose this is city map if you want to move from point A to B right. There are two way one way is directly you can suppose we are if you are a bird or you are move if you want to go A to B you can fly. Otherwise, the flight goes from point A to point B.

But if there is a fire suppose a fire engine it has to move. It has to follow a rectangular distance. So because there are different streets, so this distance is nothing but your Manhattan distance. You see that the distance or Manhattan distance will be larger than the Euclidean distance the green one is Euclidean distance. The blue one is nothing but the Manhattan distance.

(Refer Slide Time: 10:10)

Interpretation

- Suppose you live in a city where the streets are all north-south or east-west, and hence perpendicular to each other
- Let Figure 3 be part of a street map of such a city, where the streets are portrayed as vertical and horizontal lines

Let us interpret the meaning of Manhattan distance. Suppose you live in a city where the streets are all north-south or east-west and hence perpendicular to each other. Let figure 3 be the part of street map of such a city where the streets are portrayed as a vertical and horizontal lines. So if you want to move from point A to point B, you cannot directly you cannot go by shortest path you have to take a rectangular distance. Another name for this Manhattan distance is rectilinear distance.

(Refer Slide Time: 10:41)

Interpretation

- Then the actual distance you would have to travel by car to get from location i to location j would total $|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$
- This would be the shortest length among all possible paths from i to j
- Only a bird could fly straight from point i to point j, thereby covering the Euclidean distance between these points

Then the actual distance you would have to travel by a car or fire engine to get from location i to location j would be $|x_{i1} - x_{j1}|$ modulus value + $|x_{i2} - x_{j2}|$ modulus value. This would be the shortest length among all possible paths from i to j. Only a bird could fly straight from point i to j

thereby covering Euclidean distance between these points. So the example of the bird, which covers point A to B is the example for your Euclidean distance.

(Refer Slide Time: 11:21)

Mathematical Requirements of a Distance Function

- Both the Euclidean metric and the Manhattan metric satisfy the following mathematical requirements of a distance function, for all objects i , j , and h :
- (D1) $d(i, j) \geq 0$
- (D2) $d(i, i) = 0$
- (D3) $d(i, j) = d(j, i)$
- (D4) $d(i, j) \leq d(i, h) + d(h, j)$
- Condition (D1) merely states that distances are nonnegative numbers and (D2) says that the distance of an object to itself is zero
- Axiom (D3) is the symmetry of the distance function
- The triangle inequality (D4) looks a little bit more complicated, but is necessary to allow a geometrical interpretation
- It says essentially that going directly from i to j is shorter than making a detour over object h



The mathematical requirements of a distance function, both Euclidean metric and Manhattan metric, satisfy the following mathematical requirements of a distance function for all objects i , j and h . The first property is D1 $d(i, j) \geq 0$, $d(i, i) = 0$, $d(i, j) = d(j, i)$, $d(i, j) \leq d(i, h) + d(h, j)$ condition D1 merely states that the distances are non-negative numbers and D2 says that the distance of an object itself is 0 because $d(i, i) = 0$.

Axiom D3 is the symmetry of the distance function. The triangle inequality axiom D4 looks a little bit more complicated, but it is necessary to allow a geometrical interpretation. It says essentially that going directly from i to j is shorter than making a detour over object h . For example, suppose this is i this is j , and this is h so what says moving point i to j this will be shorter than moving i to h and h to j that is your triangular inequality.

(Refer Slide Time: 12:59)

Distances computation between the objects

- If $d(i, j) = 0$ does not necessarily imply that $i = j$, because it can very well happen that two different objects have the same measurements for the variables under study
- However, the triangle inequality implies that i and j will then have the same distance to any other object h , because $d(i, h) \leq d(i, j) + d(j, h) = d(j, h)$ and at the same time $d(j, h) \leq d(j, i) + d(i, h) = d(i, h)$, which together imply that $d(i, h) = d(j, h)$

Distance computation between the objects if $d(i, j) = 0$ does not necessarily imply that $i = j$ because it can very well happen that two different objects have the same measurement for the variable under study. What is the meaning of this one is if the distance between object $i, j = 0$ it need not necessary that always it should be $i = j$. Sometimes there may be two objects which is not $i = j$ their distance also may be 0.

However, the triangle inequality implies that i and j will then have the same distance to any other object h because $d(i, h) \leq d(i, j) + d(j, h) = d(j, h)$ at the same time $d(j, h) \leq d(j, i) + d(i, h) = d(i, h)$ which together imply that $d(i, h) = d(j, h)$.

(Refer Slide Time: 14:13)

Minkowski distance



- A generalization of both the Euclidean and the Manhattan metric is the Minkowski distance given by:

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{in} - x_{jn}|^p)^{1/p},$$

Where p is any real number larger than or equal to 1

- This is also called the L_p metric, with the Euclidean ($p = 2$) and the Manhattan ($p = 1$) as special cases

The next measure of the distance is Minkowski distance. A generalization of both Euclidean and Manhattan metric is the Minkowski distance. It is given by $d_{ij} = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{in} - x_{jn}|^p}$ where p is any real number larger than or equal to 1. This is also called the L_p metric for the Euclidean distance $p = 2$ and for Manhattan distance $p = 1$ as a special case.

(Refer Slide Time: 14:58)

Example for Calculation of Euclidean and Manhattan Distance

- Let $x_1 = (1, 2)$ and $x_2 = (3, 5)$ represent two objects as in the given Figure. The Euclidean distance between the two is $\sqrt{2^2 + 3^2} = 3.61$. The Manhattan distance between the two is $2 + 3 = 5$.

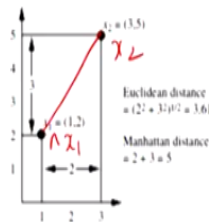


Figure: 4

Now let us take some example and calculate the Euclidean distance, Manhattan distance and Minkowski distance. Let $x_1 = (1, 2)$ and $x_2 = (3, 5)$ represent two objects. This is point 1 so call it as x_1 this is point x_2 . The Euclidean distance between these two points x_1 and x_2 is what you see that it is $2^2 + 3^2$, square root it is 3.61. The Manhattan distance between the two points is $2 + 3$. So this is your Euclidean distance, this is Manhattan distance.

You see that the Euclidean distance is smaller than the Manhattan distance because in Manhattan distance you cannot have a direct route, you have to take only a rectangular route that will be the larger. So this line represents Euclidean distance, move here then move here when you add that, that represents your Manhattan distance.

(Refer Slide Time: 16:13)

n- by- n Matrix

- For example, when computing Euclidean distances between the objects of the following Table can be obtain as next slide:

Person	Weight(Kg)	Height(cm)
A	15	95
<u>B</u>	49	156
C	13	95
D	45	160
<u>E</u>	85	178
F	66	176
G	12	90
H	10	78

- Euclidean distances between B and E:
- $((49 - 85)^2 + (156 - 178)^2)^{\frac{1}{2}} = 42.2$



Let us take another example n- by-n matrix. This is one of the input for a cluster analysis for example, when computing Euclidean distance between the objects of the following table can be obtained in the next slides. For example, there are 1, 2, 3, 4, 5, 6, 7, 8. There are 8 persons their weight and heights are given. Now let us find out how to make n- by-n matrix, by calculating the distance between each persons each objects.

So generally, if you want to know the Euclidean distance between B and E, for example, B and E that is nothing but $(49 - 85)$ whole square. Otherwise, $(85 - 49)$ because we are squaring it + $(156 - 178)$ total square you take square root that is 42.2. So the distance between B and E is 42.2. Like that for between A and B, A and C we can find out.

(Refer Slide Time: 17:17)

n- by- n Matrix

	A	B	C	D	E	F	G	H
A	0	69.8	2.0	71.6	108.6	95.7	5.8	17.7
B	69.8	0	70.8	5.7	42.2	26.3	75.7	87.2
C	2.0	70.8	0	72.5	109.9	96.8	5.1	17.3
D	71.6	5.7	72.5	0	43.9	26.4	77.4	89.2
E	108.6	42.2	109.9	43.9	0	19.1	114.3	125.0
F	95.7	26.3	96.8	26.4	19.1	0	101.6	112.9
G	5.8	75.7	5.1	77.4	114.3	101.6	0	12.2
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2	0

Do you see that n- by -n matrix the distance between in A and A is 0. You see that all the diagonal will be 0. The distance between A and B is 69.8 the distance between A and C is 2.0. So in my previous slides I have explained the distance between B and E is 42.2. So you see that this is symmetric see this upper triangle value is equal to your lower triangle value. So that is a replica, that is a mirror image of this value.

(Refer Slide Time: 17:54)

Interpretation

- The distance between object B and object E can be located at the intersection of the fifth row and the second column, yielding 42.2
- The same number can also be found at the intersection of the second row and the fifth column, because the distance between B and E is equal to the distance between E and B
- Therefore, a distance matrix is always symmetric
- Moreover, note that the entries on the main diagonal are always zero, because the distance of an object to itself has to be zero

Let us interpret this distance matrix. The distance between object B and E can be located at the intersection of the fifth row and the second column yielding 42.2. Now let us interpret that distance matrix. The distance between object B and E can be located at the intersection of fifth

row and second column. That was this one I am going to previous slide. The fifth row 1, 2, 3 ,4 fifth row second column this one 42.2.

The same number can be found at the intersection of 2nd row and 5th column because the distance between B and E is equal to the distance between E and B, therefore the distance matrix is always symmetric. Moreover, note that the entries of the main diagonal are always 0 because the distance of an object to itself has to be 0.

(Refer Slide Time: 18:54)

Distance matrix

- It would suffice to write down only the lower triangular half of the distance matrix

	A	B	C	D	E	F	G
B	69.8						
C	2.0	70.8					
D	71.6	5.7	72.5				
E	108.6	42.2	109.9	43.9			
F	95.7	26.3	96.8	26.4	19.1		
G	5.8	75.7	5.1	77.4	114.3	101.6	
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2

Now we have shown only the lower triangle it would be suffice to write down only the lower triangular half of the distance Matrix.

(Refer Slide Time: 19:04)

Selection of variables

- It should be noted that a variable not containing any relevant information (say, the telephone number of each person) is worse than useless, because it will make the clustering less apparent.
- The Occurrence of several such “trash variables” will kill the whole clustering because they yield a lot of random terms in the distances, thereby hiding the useful information provided by the other variables.
- Therefore, such non informative variables must be given a zero weight in the analysis, which amounts to deleting them

Now let us see the selection of the variables, because before doing cluster analysis we have to see whether we have to select all the variables of the variables, which is relevant to our problem. It should be noted that a variable not containing any relevant information say the telephone number of each person is worse than useless because it will make the clustering less apparent. The occurrence of several such trash variable will kill the whole clustering.

Because they yield a lot of random terms in the distances thereby hiding the useful information provided by the other variables. Therefore, such non-informative variables must be given you zero weight in the analysis, which amounts to deleting them. So any not important variable, you can give zero weightage so that that will not be taken into calculation.

(Refer Slide Time: 20:02)

Selection of variables

- The selection of “good” variables is a nontrivial task and may involve quite some trial and error (in addition to subject-matter knowledge and common sense)
- In this respect, cluster analysis may be considered an exploratory technique

So the selection of good variable is a non-trivial task and may involve quite some trial and error in addition to subject matter knowledge and common sense. In this respect so a cluster analysis may be considered as an exploratory technique. In this lecture we have seen the effect of standardization then calculation of different types of distances with the help of example. I have explained how to find out Euclidean distance, Manhattan Distance and Minkowski distance.

Then formulation and interpretation of n by n matrix. Then I have explained this is one of the input for cluster analysis there are n objects, n variables, how to find out the distance between these two variables or objects. Then I have explained how to select relevant variables for the cluster analysis.