

**INDIAN INSTITUTE OF TECHNOLOGY ROORKEE**

**NPTEL**

**NPTEL ONLINE CERTIFICATION COURSE**

**Marketing Research**

**Lec -31**

**Cluster Analysis - II**

**Dr. Jogendra Kumar Nayak  
Department of Management Students  
Indian Institute of Technology Roorkee**

Welcome friends let us continue with the section on cluster analysis so in the last section we had discussed the requirement the need behind cluster analysis and the use of cluster analysis so we saw that cluster analysis has got a large application very large application in almost every field right for example in management in marketing for example we use it for segmenting the market to target the right costumers right.

And in for example in biology or in other fields you use it classification so since age is classification as been one of the most important research areas for human beings right so in that aspect if you see cluster analysis as got a very large application and now a day's cluster analysis is being used for other areas also like image processing and all where people are we can you know we are assembling people are creating clusters of people may be through certain attributes certain criteria of the looks are you know the shape of the human face and all.

And may be from that we can may be create some kind of behavioral pattern some kind of you know we can use them for certain purposes right we do not know we but for sure we can use it for doing different kinds of analysis and studies interesting studies okay so while doing this we had said how do we find a similarity so we said that in cluster analysis we use the distance as a measure instead of the correlation which was used a factor analysis right.

Ann when you do a distance you calculate the distance so we said there are so many ways of calculating the euclidean distance the monatomic distance and the modular distance and all but we are basically generally we are using the euclidean distance so once you do that how do you

the next question was how do you identify the clusters so when we did this study we found the initial solutions if you remember this values which had come between.

(Refer Slide Time: 02:29)

### How do we form clusters?

Step	AGGLOMERATIVE PROCESS		CLUSTER SOLUTION		
	Minimum Distance Unclassified Observations*	Observation Pair	Cluster Membership	Number of Clusters	Overall Similarity Measure (Average Within-Cluster Distance)
	Initial Solution		(A)(B)(C)(D)(E)(F)(G)	7	0
1	1.414	E-F	(A)(B)(C)(D)(E-F)(G)	6	1.414
2	2.000	E-G	(A)(B)(C)(D)(E-F-G)	5	2.152
3	2.000	C-D	(A)(B)(C-D)(E-F-G)	4	2.144
4	2.000	B-C	(A)(B-C-D)(E-F-G)	3	2.234
5	2.236	B-E	(A)(B-C-D-E-F-G)	2	2.896
6	3.342	A-B	(A-B-C-D-E-F-G)	1	3.420

In steps 1,2,3 and 4, the OSM does not change substantially, which indicates that we are forming other clusters with essentially the same heterogeneity of the existing clusters.  
When we get to step 5, we see a large increase. This indicates that joining clusters (B-C-D) and (E-F-G) resulted a single cluster that was markedly less homogenous.

The minimum distance which we have observed right so then we added the clusters right we the observation the pairs which where clubbed together right for example E and F where clubbed together to form one so here initially there was 7 clusters now it has reduced to 6 by 1 by to 6 clusters right okay similarly then we go on reducing by adding up the closeted clusters and forming the finally the only one final cluster which is the combination of all the total together A, B, C, D, E, F, G okay.

All the variables sorry responders so now after doing this you need this over all similarity measure now what is this overall similarity measure the overall similarity measure is nothing but the average within the cluster distance now what do you mean by the average within the cluster distance how do we measure it for example let us look at the 1<sup>st</sup> case if you look at the 1<sup>st</sup> case which is when it is 1.414 now how you have archived this 1.414 now let us look at this table.

(Refer Slide Time: 03:35)



### How do we measure similarity?

Proximity Matrix of Euclidean Distance Between Observations

Observation	Observations						
	A	B	C	D	E	F	G
A	---						
B	3.162	---					
C	5.099	2.000	---				
D	5.099	2.828	2.000	---			
E	5.000	2.236	2.236	4.123	---		
F	6.403	3.606	3.000	5.000	1.414	---	
G	3.606	2.236	3.606	5.000	2.000	3.162	---

$$d_{\text{Euclidean}}(A, B) = \sqrt{(V_{1(A)} - V_{1(B)})^2 + (V_{2(A)} - V_{2(B)})^2}$$

$$d_{\text{Euclidean}}(A, B) = \sqrt{(3 - 4)^2 + (2 - 5)^2} = 3.162$$

Right now if you look at this table E and F I think the pair was E and F if you look at the combination it is 1.414. And the next one if you see

(Refer Slide Time: 03:45)

## How many groups do we form?

- Therefore, the three – cluster solution of Step 4 seems the most appropriate for a final cluster solution, with two equally sized clusters, (B-C-D) and (E-F-G), and a single outlying observation (A).

*This approach is particularly useful in identifying outliers, such as Observation A. It also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.*

Is 2.192 another question comes to your mind is how does 2.192 come.

(Refer Slide Time: 03:52)

### How do we measure similarity?

Proximity Matrix of Euclidean Distance Between Observations

Observation	Observations						
	A	B	C	D	E	F	G
A	---						
B	3.162	---					
C	5.099	2.000	---				
D	5.099	2.828	2.000	---			
E	5.000	2.236	2.236	4.123	---		
F	6.403	3.606	3.000	5.000	1.414	---	
G	3.606	2.236	3.606	5.000	2.000	3.162	---

$$d_{Euclidean}(A, B) = \sqrt{(V_1(A) - V_1(B))^2 + (V_2(A) - V_2(B))^2}$$

$$d_{Euclidean}(A, B) = \sqrt{(3 - 4)^2 + (2 - 5)^2} = 3.162$$

WU WIRTSCHAFTS UNIVERSITÄT WIEN  
WU ONLINE CERTIFICATION COURSE

So if you look at E, F, G so the 3 variables which are connected are E, F, G so you need to take the distance between and E and F one pair E and G another pair and F and G find the distance between these all of those three and let us and then take the average so for example E and F is 1.414 right so 1.414 + E and G so G and E is 2 right.

(Refer Slide Time: 04:28)

$$\begin{array}{c}
 \text{E F G} \\
 \hline
 \text{EF} \quad \text{EG} \quad \text{FG} \\
 1.44 + 2.00 + 3.162 = \frac{6.6}{3} = 2.2
 \end{array}$$

So 2 and then + is F and G so F and G, G and F is 3.162 so 3.162 so when I am adding this 3 I am taking the mean so it becomes something like you know 3, 4, 5, 6, 6.4, 6.5 almost 6.6 let us say divided by 3 so something around 2.2 close to 2.2 the value should be there right now if you look at it so we have got a value of 2.192 similarly you can measure for others also so you can add up for CD.

(Refer Slide Time: 05:12)

### How do we form clusters?

Step	AGGLOMERATIVE PROCESS		CLUSTER SOLUTION		
	Minimum Distance (Uncolored Observations)	Observation Pair	Cluster Membership	Number of Clusters	Overall Similarity Measure (Average Within Cluster Distance)
	Initial Solution		{A,B,C,D,E,F,G}	7	0
1	1.414	E-F	{A,B,C,D,G}, {F,G}	6	1.414
2	2.000	E-G	{A,B,C,D,G}, {F,G}	5	2.182
3	2.000	C-D	{A,B,C,D,G}, {F,G}	4	2.144
4	2.000	B-C	{A,B,C,D,G}, {F,G}	3	2.234
5	2.234	B-E	{A,B,C,D,E,F,G}	2	2.896
6	3.362	A-B	{A,B,C,D,E,F,G}	1	3.620

In steps 1,2,3 and 4, the OSM does not change substantially, which indicates that we are forming other clusters with essentially the same heterogeneity of the existing clusters.

When we get to step 5, we see a large increase. This indicates that joining clusters (B-C-D) and (E-F-G) resulted a single cluster that was markedly less homogenous.

ERG and then this table you can find out all the values right now if you look at here in this position what is stays here in steps 1 2 3 and 4 right the overall similarity measure does not change substantially now this is what is basically we find some from an agglomeration schedule we say so when you the in agglomeration process how do you find out the question that was important to you was how do you find the number of clusters because if you say on clusters it does not make your sense.

Does not make your sense because the entire let say India becomes one single market then the company find s it very, very tough to exactly a position itself or targets the customer similarly if every individual becomes a cluster which is 1.2 billion people then it is also impossible to something so unit of value sometimes we need to divide the country into some clusters maybe in states, let say the states let say we have 30 states right, so the number of states could be a the way of understanding the cluster.

Right so at least if a company wants to get into India it says okay well I would like to you know get into only three states out of it maybe the southern states of India the normal states of India the eastern states whatever it is right, so by doing that it becomes a simpler for the company to target it is customers okay so how what would you do in this case you will look at the agglomeration schedule the values right.

(Refer Slide Time: 06:44)

$$\begin{array}{c}
 E \quad F \quad G \\
 \hline
 (EF) \quad (EG) \quad (FG) \\
 1.414 + 2.000 + 3.182 = \frac{6.6}{3} = 2.2
 \end{array}$$
  

$$\begin{array}{r}
 2.134 \\
 2.234 \\
 2.896 \\
 3.420
 \end{array}
 \left.
 \begin{array}{l}
 \\
 3 \\
 2 \\
 \end{array}
 \right\}
 \begin{array}{l}
 \text{minimum} \\
 \\
 3 \text{ clusters}
 \end{array}$$

So as it starts you have the total finally here which is somewhere around 3.420 right so if you look at this the change look at this change between the clusters right, so this was 3.420 then the other one it is not visible 2.8 2.234 2.896 2.234 now other values whichever above it there we found that the change was not much right, so it was 2.1 only 2.134 so the change from this to this is quite minimum so when you need to find out how many clusters the question comes how many clusters.

please go through the bottom of approach so what you do is this is let say the one clusters right now look at these two suppose there is a substantial difference between these two values the coefficients then what you can do is you assume okay that means this is good enough to be call another clusters so there are two clusters at least look at the difference between this and this if the difference between these two is also sufficiently large then we say there is a possibility of third cluster.

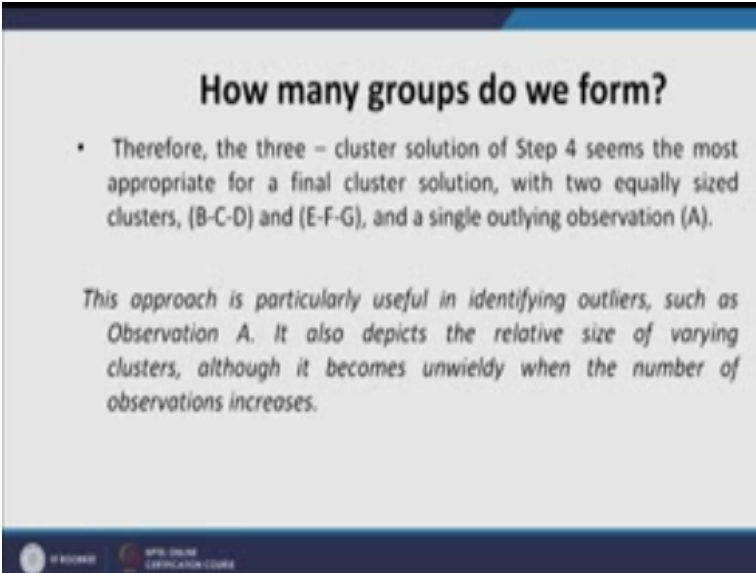
But now suppose I am saying that this difference is minimal is minimal right minimum so when the difference is minimum I am saying let us stop why because now the behavior between the you know that the clusters is getting more or less similar so there is hardly any difference to be found out so we cannot stay okay this cluster and that cluster are sufficiently different so there we stopped, so and we stay okay this three clusters are sufficiently large and sufficiently different from each other.



So we can say now how many clusters are there in the study now we can say three clusters are there so finally we say there are three clusters and now this is something very interesting here some researcher can argue that I will take only two clusters y3 right, well it is all up to you to decide how many clusters you want to take but there has to be a sufficient justification or you know backing behind it, okay why you are taking two why you are taking three why you want to take four anything.

That you want to take you to justify the reason behind it right obviously, if you take two many it is not wise if you take two less it is not serving the purpose okay, so how many groups so therefore three cluster.

(Refer Slide Time: 09:27)



**How many groups do we form?**

- Therefore, the three – cluster solution of Step 4 seems the most appropriate for a final cluster solution, with two equally sized clusters, (B-C-D) and (E-F-G), and a single outlying observation (A).

*This approach is particularly useful in identifying outliers, such as Observation A. It also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.*

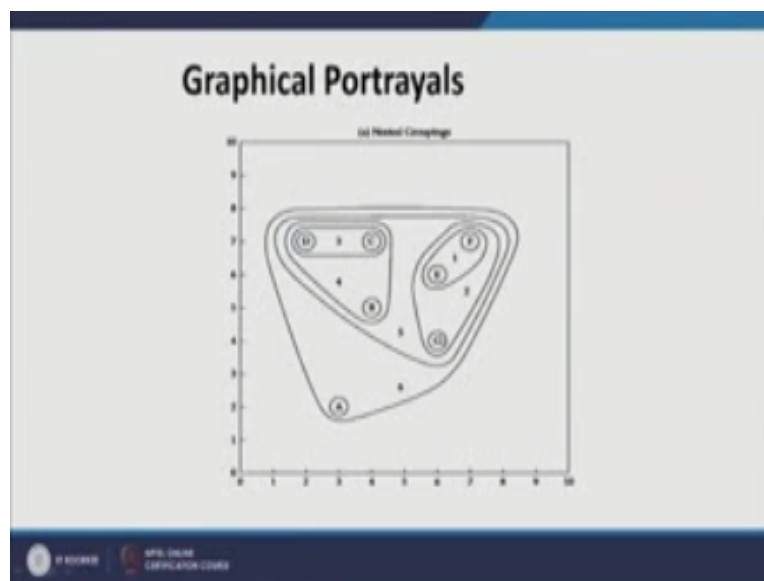
BY EDUCATION | NPTEL ONLINE CERTIFICATION COURSE

Solution of step force is most appropriate for a final cluster solution right so we had this three clusters B C D as one group E F G as another group and sorry A as the only signal one right this approach is particularly useful and also identifying outliers such as observations A through C now observation A was sufficiently different from the rest right why look at the value it has the lowest value right and it was individually unique in itself right.

It depth it is the relation size of the varying clusters although it becomes unwieldy the when the number of observation increases obviously. So when you are doing a cluster analysis unit to identify how many groups do we form and this is a very useful method, so hierarchal cluster in technique which is the one which is just did, there are two basically cluster in techniques.

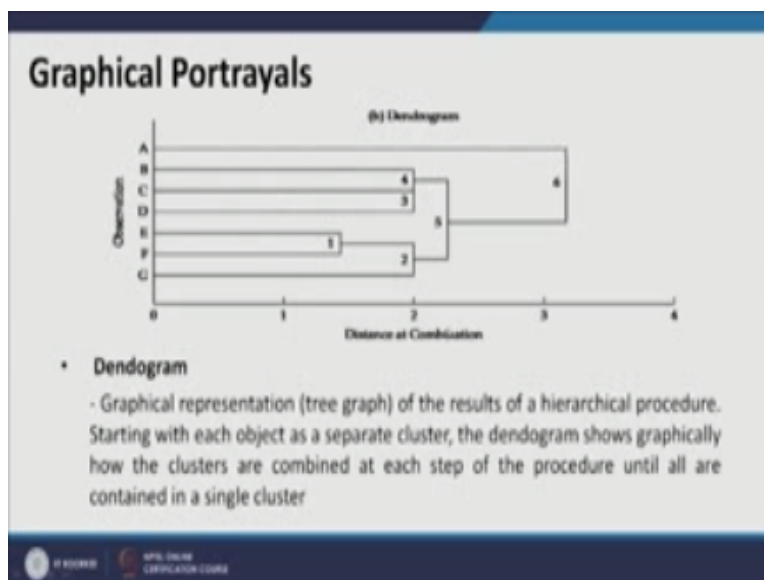
One call the hierarchical clustering the other is called the non-hierarchical clustering but one of them is very popular among it which called a K means and we will see how that is also used, right. So graphical portrayals if you can see this is how it looks, so A is here B,C,D is here another.

(Refer Slide Time: 10:43)



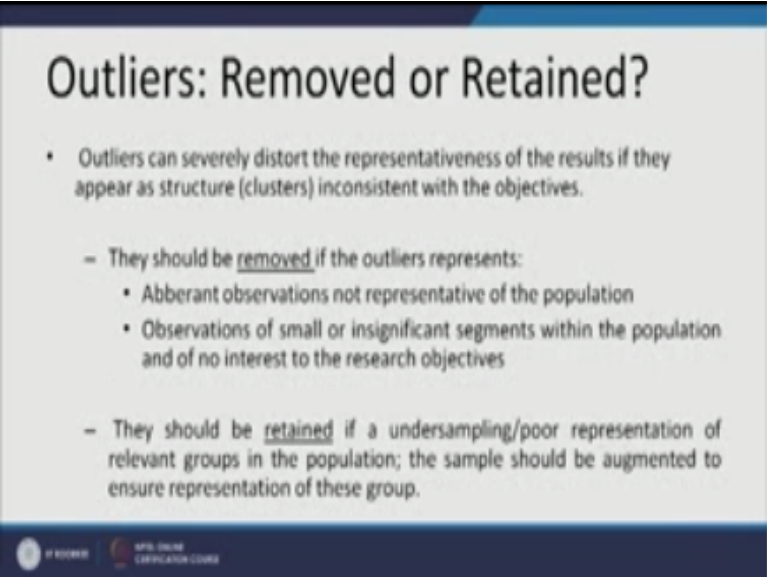
C, E, F is here if you take totally this becoming a single cluster, right so this is how it would look like.

(Refer Slide Time: 10:49)



This is something called a dendrogram, a dendrogram is a tree like structure it is a tree graph it is called tree graph if you can see here, tree graph right. So it does is basically it just does a graphical representation of the adding up of the clusters together to form a single final cluster, right, so this is the basic purpose of a dendrogram.

(Refer Slide Time: 11:11)



The slide is titled "Outliers: Removed or Retained?". It contains a bulleted list of guidelines for handling outliers. The first bullet point states that outliers can distort results if they appear as structure inconsistent with objectives. The second bullet point, marked with a minus sign, discusses when outliers should be removed or retained. The "removed" section lists aberrant observations and insignificant segments. The "retained" section discusses undersampling and the need to augment the sample.

## Outliers: Removed or Retained?

- Outliers can severely distort the representativeness of the results if they appear as structure (clusters) inconsistent with the objectives.
- They should be removed if the outliers represents:
  - Abberant observations not representative of the population
  - Observations of small or insignificant segments within the population and of no interest to the research objectives
- They should be retained if a undersampling/poor representation of relevant groups in the population; the sample should be augmented to ensure representation of these group.

ST RICHMOND    NPTEL ONLINE CERTIFICATION COURSE

Okay, outliers now in the last session also I have discussed with you about outliers now why are outliers important, outliers are important because any outlier can completely distort the formation of clusters it can disturb the formation of clusters right, so outliers should be always removed right, if they are present right, so but all though I have written it can be retained also in the last if you can see they should be retained if only there is a under sampling now if you are sample is too less right, then you may retain but let me tell you please if there is an outlier you need to remove it the best thing is that.

Because yes, sometimes if your samples are too less anything it would look like an outlier a small even change could look like an outlier but if you are, but mostly in all most all the cases outliers are critically problems which can distort all the values, right. So you should avoid outliers, you should remove the outliers the best thing, right.

(Refer Slide Time: 12:26)

## Detecting Outliers

- Outliers can be identified based on the similarity measure by:
  - Finding observations with large distances from all other observations.
  - Graphic profile diagrams highlighting outlying cases.
  - Their appearance in cluster solutions as single – member or small clusters.

So you can you know identifying outliers you can do to several methods the similarity method is the graphical method right, you can just find out you know the frequency of the values also and seek whether some value is there which is abnormally higher, abnormally low right, so that also can be a measure the several ways.

(Refer Slide Time: 12:46)

## Sample Size

- The researcher should ensure that the sample size is large enough to provide sufficient representation of all relevant groups of the population
- The researcher must therefore be confident that the obtained sample is representative of the population.

WU ONLINE  
LERNPLATZ ONLINE

Sample size should be large enough, yes when you are doing a cluster analysis you should ensure, the researcher should ensure the sample size is large enough to provide sufficient representation of all the relevant groups of the population that means you are not missing anything right, so it is a representative of the population, the researcher must be confident that the sample the obtain sample is representative of the population, okay.

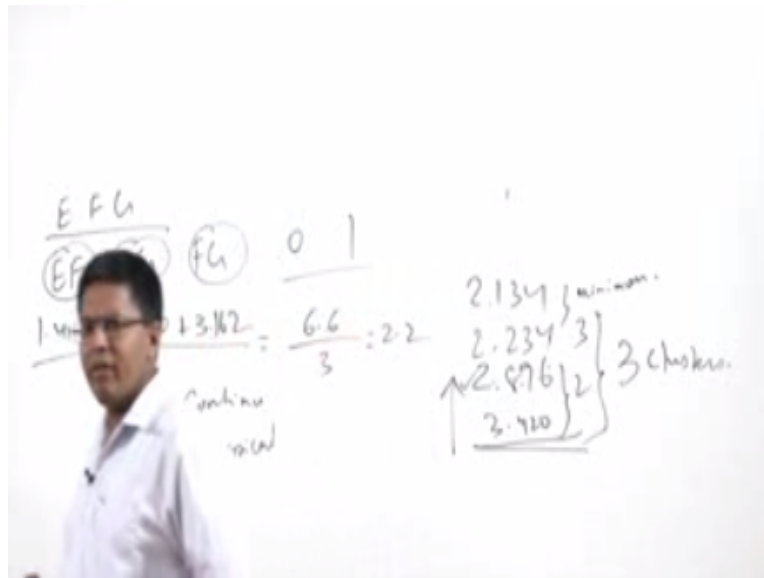
(Refer Slide Time: 13:15)

## Standardizing the Data

- Clustering variables that have scales using widely differing numbers of scale points or that exhibit large differences in standard deviations should be standardized.
  - The most common standardization conversion is Z score (with mean equals to 0 and standard deviation of 1).

Now question is, when you are having attributes as I said your attributes could be measured in several scales some could be continuous, some could be non-continuous also categorical also, right.

(Refer Slide Time: 13:25)



So if I have a condition where my variables are both categorical and continuous then in that case can I use both the attributes for analyzing or interpreting or making a cluster analysis yes, you can do it, the best way to do it is to standardize the data. Now when you standardize the data that means what they become more or less are represented on a single platform right, so the most common standardization is the conversion into the Z score, Zee score, Z score my pronunciation always is you know changes from Z and Zee please do not confuse.

So I sometimes says Z sometimes Zee, so the Z score where the mean is equal to 0 and the maximum standardization of 1 right, so when you standardize the data, so all the data will lie between 0 and 1, so when then it becomes easier maybe your income was in lakhs, cores whatever it is millions and you are let us say gender was only in male and female but still when you are bringing all them into standardizing they are comparable now right, so that is one the biggest benefits.

(Refer Slide Time: 14:46)



**Deriving Clusters**

- There are number of different methods that can be used to carry out a cluster analysis; these methods can be classified as follows:

- ❖ **Hierarchical Cluster Analysis**
- ❖ **Nonhierarchical Cluster Analysis**
- ❖ **Combination of Both Methods**

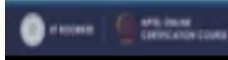
WU BUSINESS WU ONLINE CERTIFICATION COURSE

So deriving clusters so when you how to derive the clusters, how does the researcher derive the cluster is a very important thing, so we saw through the hierarchical clustering method through the algometric schedule that you are looking at the change in the value and the point at which you find the maximum change you retain that as a number of clusters right, so as I said hierarchical clustering analysis.

(Refer Slide Time: 15:07)

## Hierarchical Cluster Analysis

- The stepwise procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics using an algorithm either agglomerative or divisive, resulting to a construction of a hierarchy or treelike structure (dendrogram) depicting the formation of clusters. This is one of the most straightforward method.
- HCA are preferred when:
  - The sample size is moderate (under 300 – 400, not exceeding 1000).

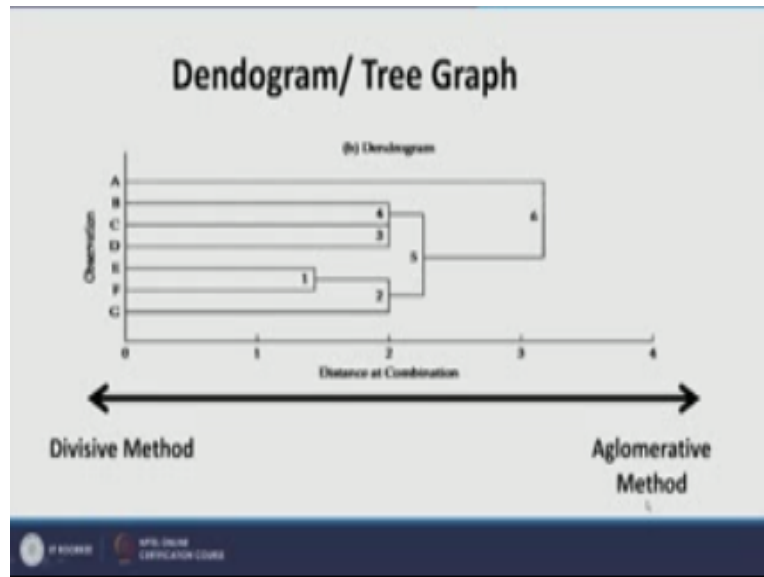


We just discussed about it so this is prepared when the sample size is moderate around 300 to 400 and not exceeding 1000, now why is it not preferable for a very large sample size the question is that the number of hydrations that the you know you have to do or the software has to do which you are using it would be very large suppose you have only 400 sample the number of combinations or the hydrations that we will do or the multiple combination that will happen may be would be around 80000 right.

So if you around a 500 sample size around more then 100000, so such high number of combination or hydrations that happen is one important thing that is which is you know the reason why we used hierarchical clustering analysis in sample size which is moderate or not so large right if it is a large one then it becomes really hectic and very difficult okay. So there two methods in hierarchical clustering as you can see here the agglomerative or divisive right so either you take one by one add them club it as we did and bring it to t single cluster or you can do it divisive method where you first form the single cluster and then you start the one which not related and then go o dividing it till you find all the variables separately right.

So did you understand what it is a bottom of approach, now either you come this way or you go from the bottom so I have first one cluster then I have two cluster then I have three this is a divisive method okay. And in the other method we have seven first then 6 then 5 then 4 till we reached one okay so the same thing right. So divisive also I explain so this is how the dendrogram look like right.

(Refer Slide Time: 17:10)




So this is the divisive method so you are and this is the agglomerative method.

(Refer Slide Time: 17:17)

## Agglomerative Algorithms

- Among numerous approaches, the five most popular agglomerative algorithms are:
  - Single – Linkage
  - Complete – Linkage
  - Average – Linkage
  - Centroid Method
  - Ward's Method
  - Mahalanobis Distance

BYJU'S ONLINE CERTIFICATION COURSE

So in one place I have shown now how is the distance to how what are the approaches to you know join the clusters or create the combination of the clusters. So there are several methods the single linkage method the complete linkage method let us see all these methods.

(Refer Slide Time: 17:34)

## Agglomerative Algorithms

- *Centroid Method*

- *Cluster Centroids*

- are the mean values of the observation on the variables of the cluster.

- The distance between the two clusters equals the distance between the two centroids.



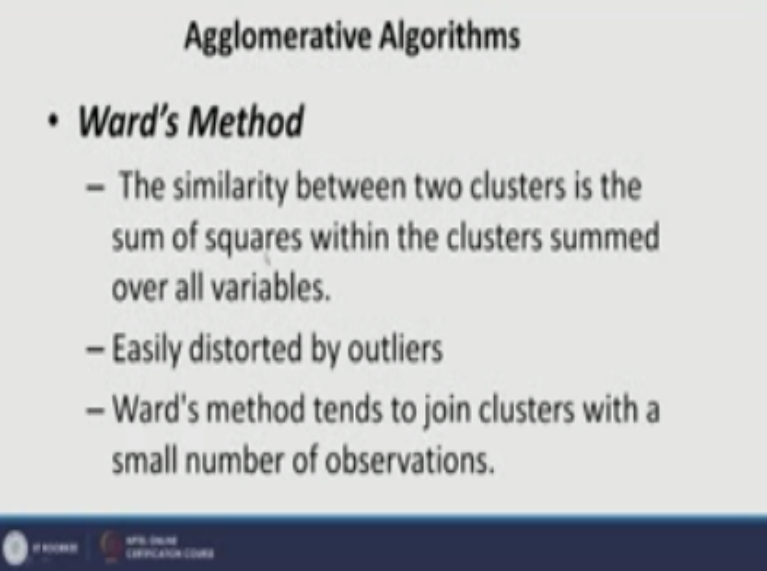
The first method of combining the clusters is called the single linkage method where what is happening is it takes the two different points from two different clusters which are nearest to each other right so they define the similarity between clusters as the shortest distance from any object in one cluster to another object in the other cluster that means suppose we will find out the distance is among all the clusters right all the variable and the one the minimum we will take the minimum right and suppose there are three clusters let us say a b And c for example now all the data's are taken all the variables distances are measured right.

The one which is the minimum that those two clusters will first to be group together right second is the complete linkage where you take the furthest points right so you take the furthest points maximum distance between any two members in the two clusters okay. average linkage is another method where two clusters the distance is define as the average distance between all the pairs here we are taking only the single the extreme or the closest in the single linkage and the completer but here we are taking the average distance between all the pairs of two clusters members right all the members and their average okay.

So again the minimum will be taken centered method is a method which is also very nice what happens here what is the centered first it is the mean value of the observations on the variable on the clusters, so you find a mean value of all the variables the values of the variables and then the this mean of one cluster and the mean of another cluster is the distance is find out right and then you find out the minimum one and then you club them start clubbing together.

So you take the minimum value as the nodule point here okay so the distance between the two cluster equal the distance between the two centuries' okay.

(Refer slide time: 19:43)



The slide is titled "Agglomerative Algorithms" in a bold, black font. Below the title, there is a bullet point for "Ward's Method". Under this bullet point, there are three sub-points, each preceded by a minus sign. The first sub-point states that the similarity between two clusters is the sum of squares within the clusters summed over all variables. The second sub-point states that the method is easily distorted by outliers. The third sub-point states that the method tends to join clusters with a small number of observations. At the bottom of the slide, there is a dark blue footer bar containing two logos: a circular logo on the left and a rectangular logo on the right.

**Agglomerative Algorithms**

- **Ward's Method**
  - The similarity between two clusters is the sum of squares within the clusters summed over all variables.
  - Easily distorted by outliers
  - Ward's method tends to join clusters with a small number of observations.

Ward's method is also very popular method highly utilized what it does is basically it is nothing but the sum of squares within the clusters summed over all the variables now what it takes basically the similarity measures between two clusters. So that two clusters all the variables between the two clusters the distances are measured and the sum squared. So after doing this we take the minimum value obviously the Ward's method you can understand there is a lot of number of calculations are very high obviously right.

And more thing is this is susceptible or very highly you know influenced by outliers so if you have an outlier it is very dangerous because the simple reason is you are squaring it so if you have the suppose all the variables were between 1 to 5 and suddenly you have a let say income or let say to 5000 and suddenly there is one guy who has got 12000 and the square of 12000 then becomes 144 right.

(Refer Slide Time: 20:59)

## Hierarchical Cluster Analysis

- The Hierarchical Cluster Analysis provides an excellent framework with which to compare any set of cluster solutions.
- This method helps in judging how many clusters should be retained or considered.

BY EDUCARE MPIL ONLINE CERTIFICATION COURSE

So that becomes too much differences right so hierarchical cluster analysis we have explained right so how many cluster should be retained understand this.

(Refer Slide Time: 21:05)

## Non Hierarchical Cluster Analysis

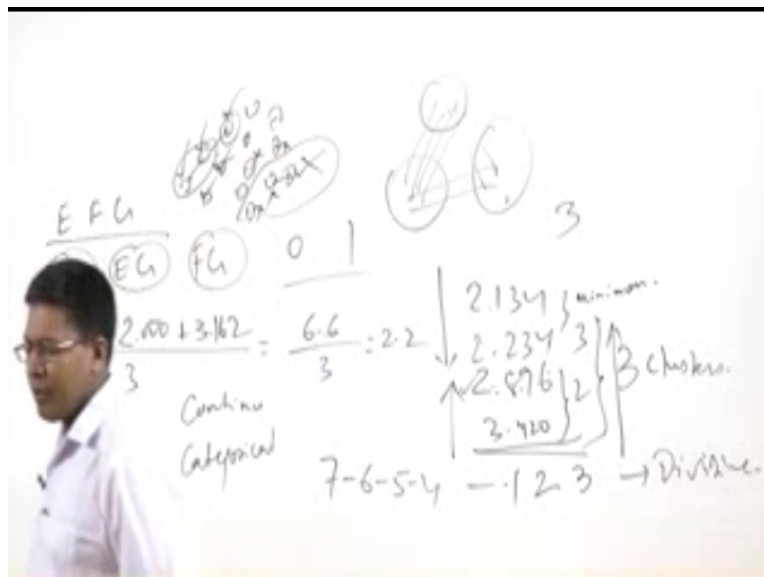
- In contrast to Hierarchical Method, the NCA do not involve the treelike construction process. Instead, they assign objects into clusters once the number of clusters is specified.
  - Two steps in Non HCA
    - 1.) **Specify Cluster Seed** – identify starting points
    - 2.) **Assignment** – assign each observation to one of the cluster seeds.

BY EDUCARE MPIL ONLINE CERTIFICATION COURSE

Then come to the case of non hierarchical so as I explained in the non hierarchical there is an advantage everything has a advantage and disadvantage and now what is the advantage with the non hierarchical in the non hierarchical the advantage is that you have a specified cluster seed or starting point right so once you have a starting point then you can find out easily which cluster is fall into which groups.

For example let say it does not involved at the stage like tree like construction process like you say in dendrogram hierarchical clustering instead they assign the objects into clusters once the number of cluster sir specified that means you take a point I will have three clusters right so if I have three clusters then accordingly what I do is I will see this clusters will fall how will the formation of this clusters is happen now for example so it says as each observation to one of the cluster seeds for example let say we have three clusters each clusters there is a starting point the starting point of the each clusters.

(Refer Slide Time: 22:45)



Now whatever value that cluster starting point is right from try to find out the variables or the distance or the score from that starting point now then you start minimum ones and start clubbing them together so what it results is automatically the variables for example suppose this



are the different clusters for example there is a variables let say so let say this is one of my starting point okay.

This is another starting point so closest one is this one to this so this is grouped them so this is all in the tick one right so this is the cross one so the cross one this is the cross one so this is the cross one this is the cross one so basically two clusters are forming them all this are they are adding up accordingly now for example this one will it come in tick or cross may be in the cross.

(Refer Slide Time: 23:29)

**Non Hierarchical Clustering Algorithm**

- Sequential Threshold Method
- Parallel Threshold Method
- Optimizing Procedures

• All of this belongs to a group of clustering algorithm known as **K - means**.

- **K - means Method**
  - This method aims to partition n observation into k clusters in which each observation belongs to the cluster with the nearest mean.
  - K - means is so commonly used that the term is used by some to refer to Nonhierarchical cluster analysis in general.

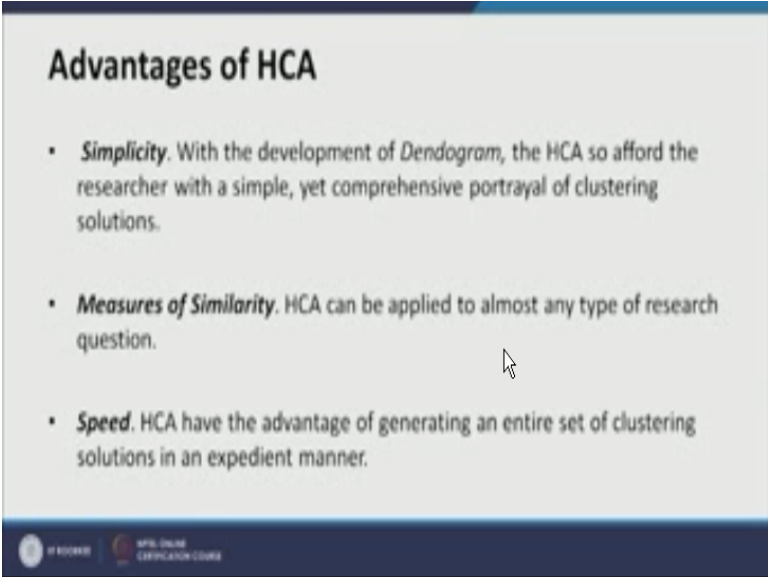
AT BOMBAY MPIL ONLINE EDUCATION COURSE

So what is doing is it ultimately tells you okay who are the variables which are following into which clusters okay so K-means method is this method aims to partition n observation into K clusters so let say 1000 observations or 500 observations into let say 4 or 5 clusters in which each observation belongs to the clusters with the nearest mean so it identifying point was there in the nearest mean.

And you started you connecting to that value and find out the closest ones K-means is so commonly used that this term is sometimes used to refer non hierarchical clustering analysis and

general right so the whole non hierarchical clustering which is if you can see algorithm as can be generally spoken as a K-means clustering technique okay.

(Refer Slide Time: 24:24)



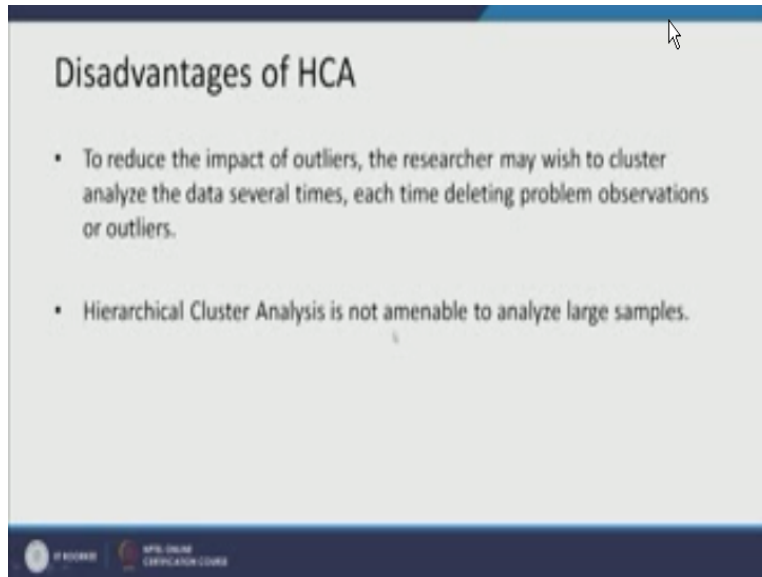
**Advantages of HCA**

- **Simplicity.** With the development of *Dendrogram*, the HCA so afford the researcher with a simple, yet comprehensive portrayal of clustering solutions.
- **Measures of Similarity.** HCA can be applied to almost any type of research question.
- **Speed.** HCA have the advantage of generating an entire set of clustering solutions in an expedient manner.

At the bottom of the slide, there are two logos: 'BY EDUCARE' and 'NPTEL ONLINE CERTIFICATION COURSE'.

So the advantages of hierarchical and disadvantages now some of the advantages of hierarchical clustering was it is simple right and it applies to any kind of research question where you want to find the similarity you know between the values or the attributes and finally it is it has an advantage because the entire set of clustering solutions is done in a faster method right this is the advantage.

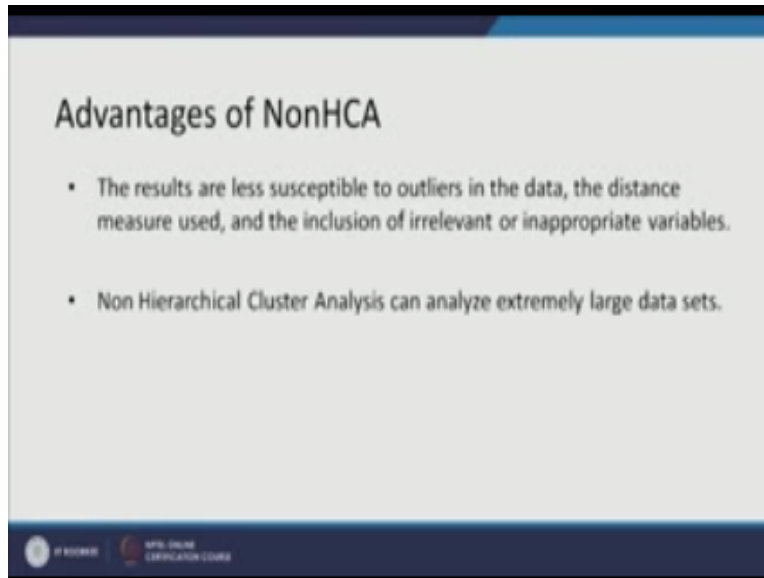
(Refer Slide Time: 24:49)



Now what is the disadvantages of hierarchical clustering the disadvantage is that if there is an out layer and you want to delete that each time deleting the problem observation of the out layers cravats a problem to the entire each time you have to run the thing again and fond out a new cluster right.

And this is not applicable to large samples because I explained you because every time the iteration is done from the beginning and it takes to much iteration okay.

(Refer Slide Time: 25:27)



Advantages of non hierarchical or k means, so the advantages are here they are less susceptible to outliers in the data because we are not squaring it right, the distance and the inclusion of irrelevant or in appropriate variables. Why it is happening is that? In a non hierarchical it is not the tree like structure right, so there is no step wise approach, rather we have to identify some seed points as the identification at the basic starting point right.

So one way clustering helps you to once we have identified the number then it helps you to identify which variable will fall into or which responded will fall into which cluster right. non hierarchical cluster analysis can analysis the extremely large data sets that is the beautiful thing, so if you have only let say 1000 samplers then okay fine, you are still working with the hierarchical but suppose you have got to many, if you 10000 can you do the cluster, yes you can do the cluster.

And the best way is you were near to somewhere the starting point, you have to say that how many clusters that I need and one of clusters helps you in that okay.

(Refer Slide Time: 26:36)

## Disadvantages of NonHCA

- Even a nonrandom starting solution does not guarantee an optimal clustering of observations. In fact, in many instances, the researcher will get a different final solution for each set of specified seed points. How is the researcher to select the optimum answer? Only by analysis and validation can the researcher select what is considered the best representation of structure, realizing that many alternatives may be acceptable.
- Nonhierarchical methods are also not so efficient when a large number of potential cluster solutions. Each cluster solution is a separate analysis, in contrast to the hierarchical techniques that generate all possible cluster solutions in a single analysis. Thus, nonhierarchical techniques are not as well suited to exploring a wide range of solutions based on varying elements such as similarity measures, observations included, and potential seed points.

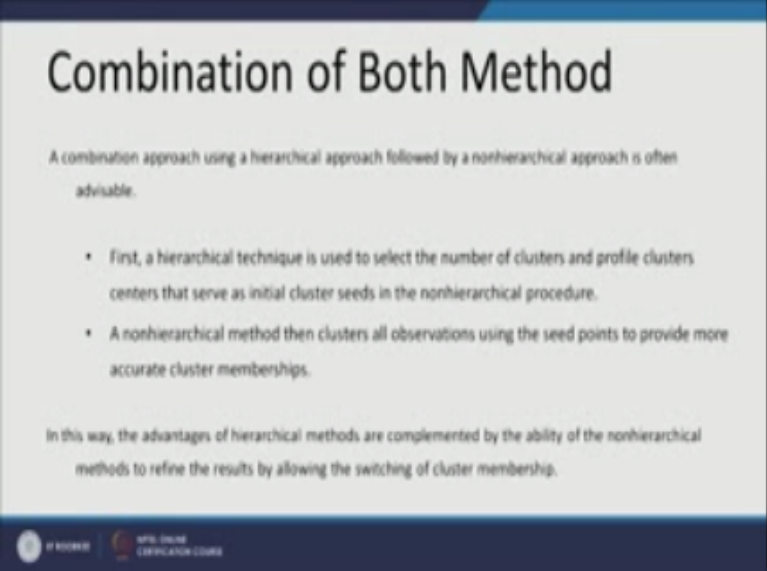
Either there is a disadvantage, yes, even in the nonrandom starting solution does not guarantee an optimal clustering of observations, so since you have done a sometimes the randomly you have done it, so it creates the problem right. so it is does not guarantee you a very effective solution, in many instances the researcher will get the different final solution, for each set of specified seed points. We can see that right, so how the researcher is going to find optimum answer is the question that is highly effective.

Which is very difficult to answer in case of the non hierarchical clustering, the other thing is other methods also not so efficient a large number of potential cluster solutions increases. That means if your cluster solutions are large in number then also the identifying the optimum solution there is the difficult thing. Because each cluster solution is separate analysis in contrast to techniques that generate all possible cluster solutions.

So what it does basically non hierarchical versus hierarchical the difference is hierarchical has got some advantage that is effective it is speedier, it is faster and you can get the clear pattern. But it has only one problem, that if you take into outliers and does not work well with the outliers, and does not work with the large data. On the other hand the non hierarchical is good but the problem is to identifying the seeds points.

So these are the basic advantages and disadvantages, so what is the best method, the best method is to had a club of both right, had club hierarchical cluster analysis and non hierarchical cluster analysis and do the final cluster analysis right.

(Refer Slide Time: 28:37)



## Combination of Both Method

A combination approach using a hierarchical approach followed by a nonhierarchical approach is often advisable.

- First, a hierarchical technique is used to select the number of clusters and profile clusters centers that serve as initial cluster seeds in the nonhierarchical procedure.
- A nonhierarchical method then clusters all observations using the seed points to provide more accurate cluster memberships.

In this way, the advantages of hierarchical methods are complemented by the ability of the nonhierarchical methods to refine the results by allowing the switching of cluster membership.

BY EDUCARE NPTEL ONLINE CERTIFICATION COURSE

The combination of the methods is first you do the hierarchical cluster, so when you want to do the clustering analysis, first you identify the number of clusters through the hierarchical clustering right, and after that use that number of clusters by defining into the k means or the non hierarchical right, so if you have 3 clusters or 4 clusters that starting point you can use it for understanding the behavior and trade of despondence and the booting them together. So what is the interpretation?

(Refer Slide Time: 29:17)

**Interpretation of Clusters**

- The cluster centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage:
  - Interpretation involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.
  - Cluster solutions failing to show substantial variation indicate other cluster solutions should be examined.
  - The cluster centroid should also be assessed for correspondence with the researcher's prior expectations based on theory or practical experience.

ST BUCKINGHAM  
MPS ONLINE  
LABORATORY COURSE

The cluster, the interpretation is that it involves the examining the distinguished characteristics of the clusters' profile right and identify the substantial differences between each clusters, so cluster 1 is different from cluster 2. Now it is different on what basis it is different that help you can through cluster analysis right. Cluster solution failing to show the substantial variation indicates the other cluster solutions should be examined okay.

Now suppose you find the two clusters are very similar in nature but they are still showing two different clusters in that case you need to be careful that you may need to work on it and find some other cluster solutions okay. So these are some of the you know things which are important yes one more thing is cluster centroid the mean value should be assessed for correspondence with the researchers prior expectations.

That means if you cluster whatever you are getting you should that should be you know that should be coming true with your earlier experience or your practical knowledge suppose you are getting a cluster and which you are getting very stage result and then there is something seriously wrong and you need to think about it so there has to be some kind of a logical conclusion that can be derived okay how do you validate the cluster now.  
(Refer Slide Time: 30:40)

## Validation of Clusters

- Validation is essential in cluster analysis because the clusters are descriptive of structure and require additional support for their relevance:
  - Cross – validation empirically validates a cluster solution by creating two subsamples (randomly splitting the sample) and then comparing the two cluster solutions for consistency with respect to the number of clusters and the clusters profiles.
  - Criterion/predictive validity: Validation is also achieved by examining differences on variables not included in the cluster analysis but for which a theoretical and relevant reason enables the expectation of variation across the clusters.

This is the last question to validate the cluster what you can do is you can split the entire data set into two half right so once you have two half and then you randomly after splitting the file you run this cluster solutions and check whether there is any difference coming or there is a similarity if there is a similarity then we would say that means there is it is valid okay so we can compare the two cluster solutions for consistency with respect to the number of cluster and the cluster profiles that is one thing the second thing is called the criterion or the predictive validity.

Now take variable which you have not used for the study or one attribute which have not used for the study right keep that has the basics right and then what you can do is but one thing suppose you have used this one variable which you have not used in the study but you know how this variable should be effecting the others the other variables used in the study so that is called a criterion and predictive validity it is achieved by examining differences on variables not included in the cluster analysis but for which a theoretical.

And relevant reason is there in my mind as a researcher that means suppose I have got certain clusters all the coefficient of the clusters are there with me so I know which variable is impacting how each cluster then suppose there is third variable which is not used in this cluster analysis but I know how that cluster that thing should be could affect the entire relation the GDP of the country and our study was to find the clusters on the bases of how people would respond to a new car right now the GDP is a variable which can effect I know how it can effect suppose I have a theoretical then that I should see well if suppose the income group is there then GDP



would affect the income group how GDP would affect the other the age the different age groups alright so by doing that also we can have predictive or a criterion validity.

So this is all we there are something more which also we can discuss may be some other time how it can you can run it in the spaces also but that is not may be the part of here the basic understanding you should know is that cluster analysis is very powerful technique which is used to create clusters and then identify that trades even what you can do is you can identify each and every individual member and say he falls into which cluster.

And when you know the trade of the cluster that means the marketer for example has an advantage of trying to motivate or try to change the mind of the consumer or the you know the respondent by rightly understanding to which cluster he belongs to and what is the behavior of the cluster right so these are some of the important utilizes of cluster analysis I think we have done with it today so thank you very much for this session.

#### **For Further Details Contact**

**Coordinator, Educational Technology Cell  
Indian Institute of Technology Roorkee  
Roorkee 247 667**

**E-Mail [Etcellitrke@gmail.com](mailto:Etcellitrke@gmail.com) [etcell@itr.ernet.in](mailto:etcell@itr.ernet.in)  
Website: [www.itr.ac.in/centers/ETC](http://www.itr.ac.in/centers/ETC). [www.nptel.ac.in](http://www.nptel.ac.in)**

**Production Team  
Sarath Koovery  
Mohan Raj. S  
Jithin. K  
Pankaj saini  
Graphics  
Binoy. V.P**

**Camera  
Arun. S  
  
Online Editing  
Arun.S  
Video Editing  
Arun.S**

**NPTEL Cooridinator  
Prof B.K Gandhi**

**An Educational Technology Cell  
IIT Roorkee Production**