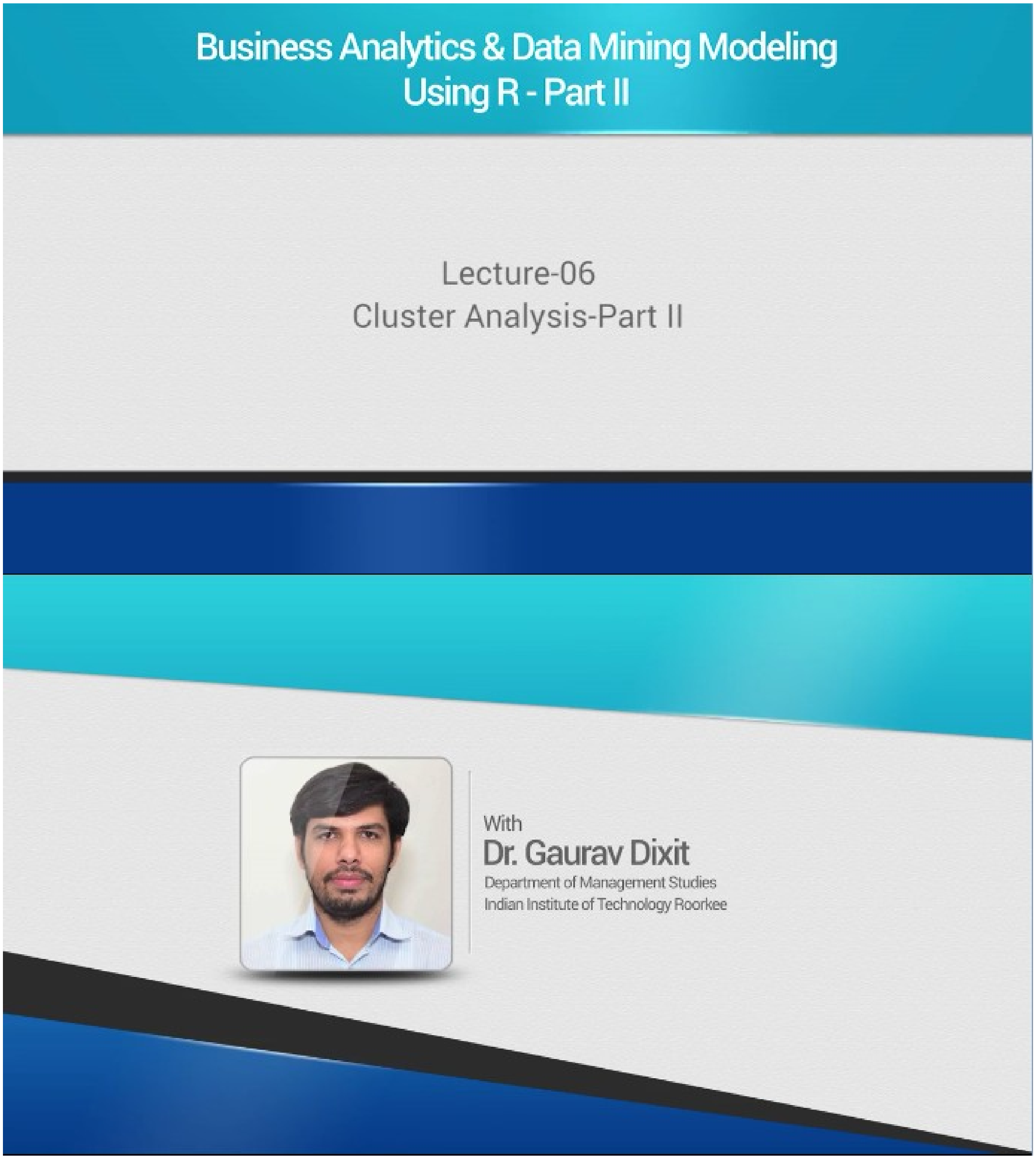
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
NPTEL
NPTEL ONLINE CERTIFICATION COURSE
Business Analytics & Data Mining Modeling
Using R – Part II
Lecture-06
ClusterAnalysis– Part II
With
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee



Welcome to the course Business Analytics and Data Mining Modeling Using R – Part 2, so in previous lecture we started our discussion on cluster analysis, so let's do a small review of what we discussed so far.

Cluster Analysis

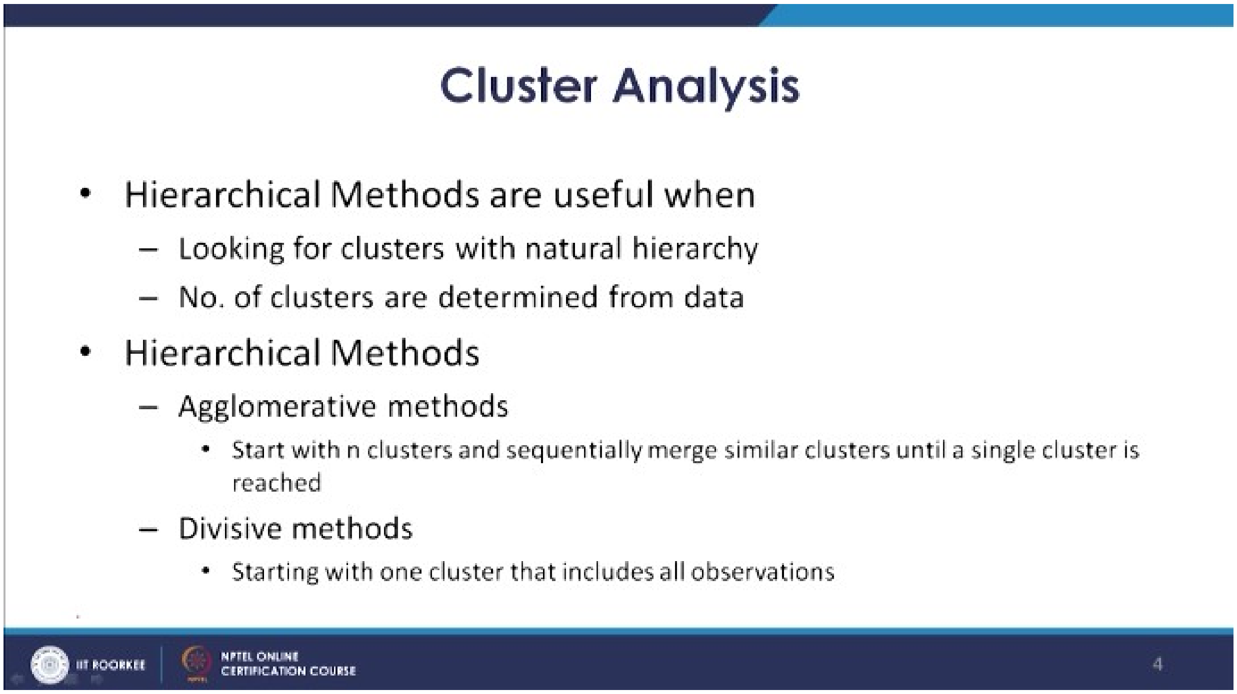
- Used
  - For unsupervised learning task of clustering
  - To form groups or clusters of similar records based on measurements taken on several variables for these records
- Main Idea is
  - To characterize the clusters in ways that would be useful for generating insight
- Applied in many domains
  - Customer segmentation, Market Structure analysis, Balanced Portfolios, Industry Analysis

So we talked about cluster analysis and you know the usefulness of this particular technique, the applications, the scenario, we also talked about the main idea we discussed that to characterize the clusters in ways that could be useful for the, for generating some insides, so that being the main idea.

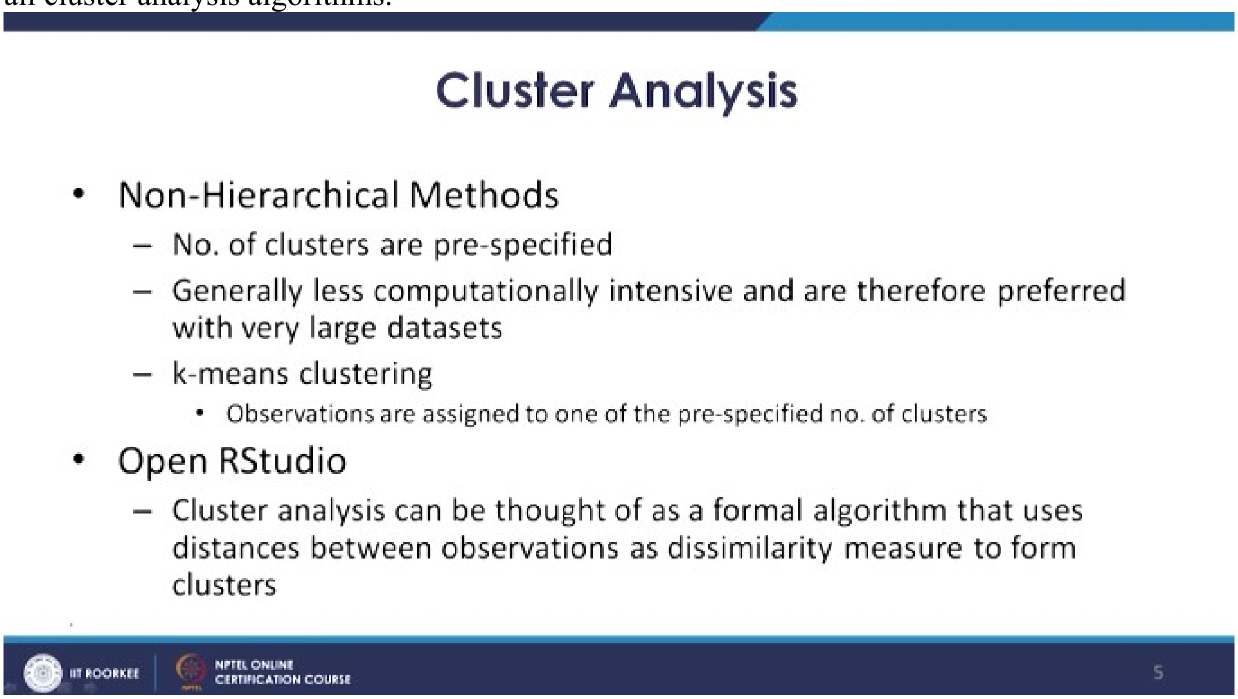
Cluster Analysis

- Example: Breakfast Cereals
  - Records to be clustered are cereals
  - Clustering would be based on eight measurements on each cereal
- Open RStudio
- Types of Clustering algorithms
  - Hierarchical Methods
  - Non-Hierarchical Methods

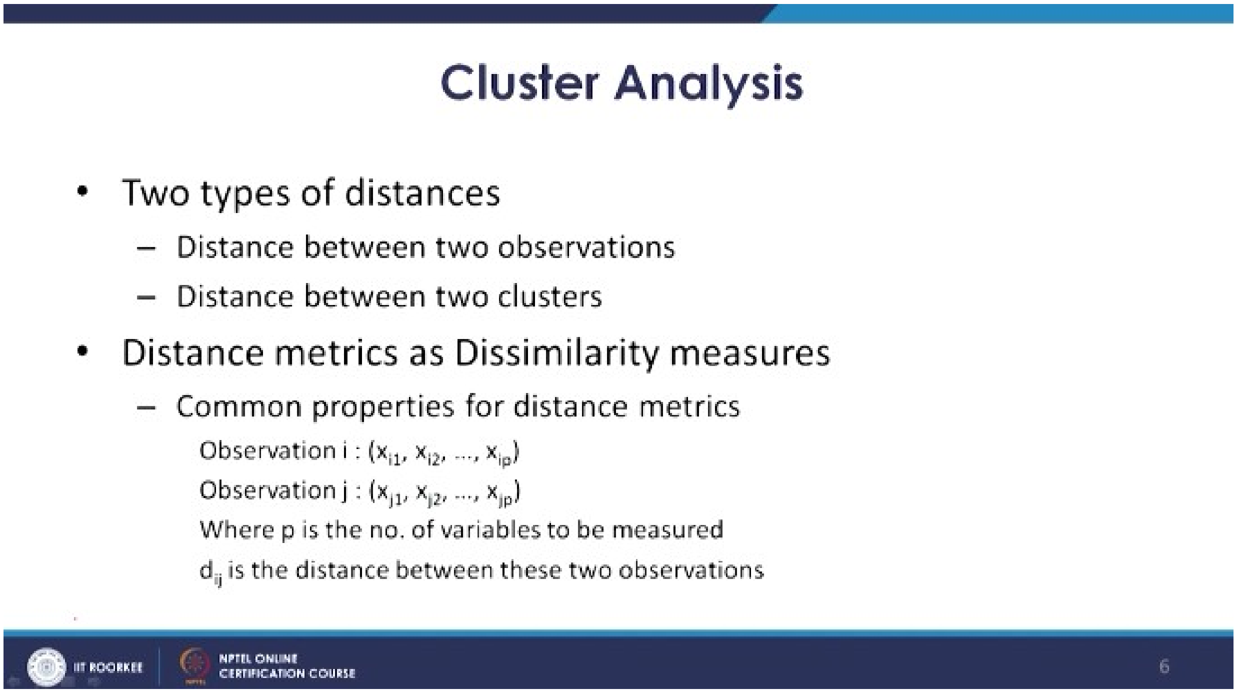
So then we also discussed this particular dataset, breakfast cereals and we did small exercise in R studio as well, then we talked about two types of clustering algorithm, hierarchical methods and non-hierarchical methods. We understood when these methods could be useful then within hierarchical methods we talked about agglomerative methods and divisive methods.



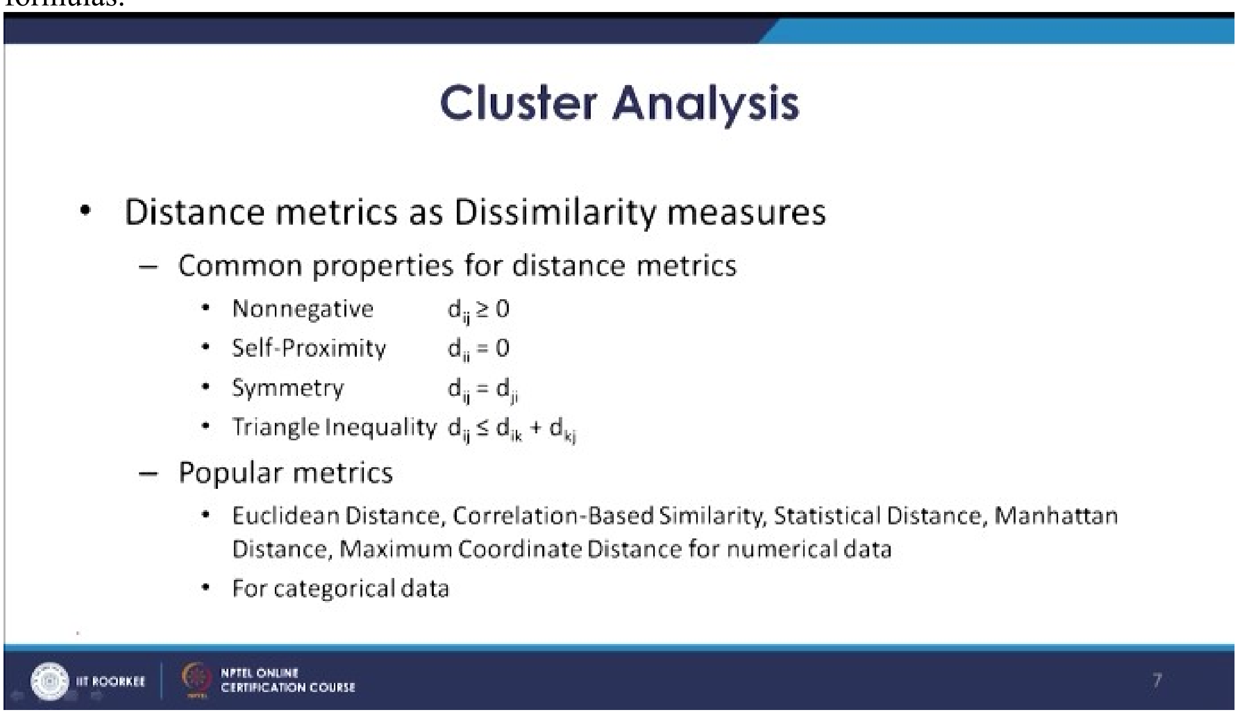
We also touched upon non-hierarchical methods and through an exercise in R studio we understood the main, how these methods cluster analysis method are typically implemented, we understood that distances between observations, so those become the major part of you know all cluster analysis algorithms.



So we also started our discussion on the types of distances that we need to compute, in cluster analysis distances between two observations being the first one, and the distance between two

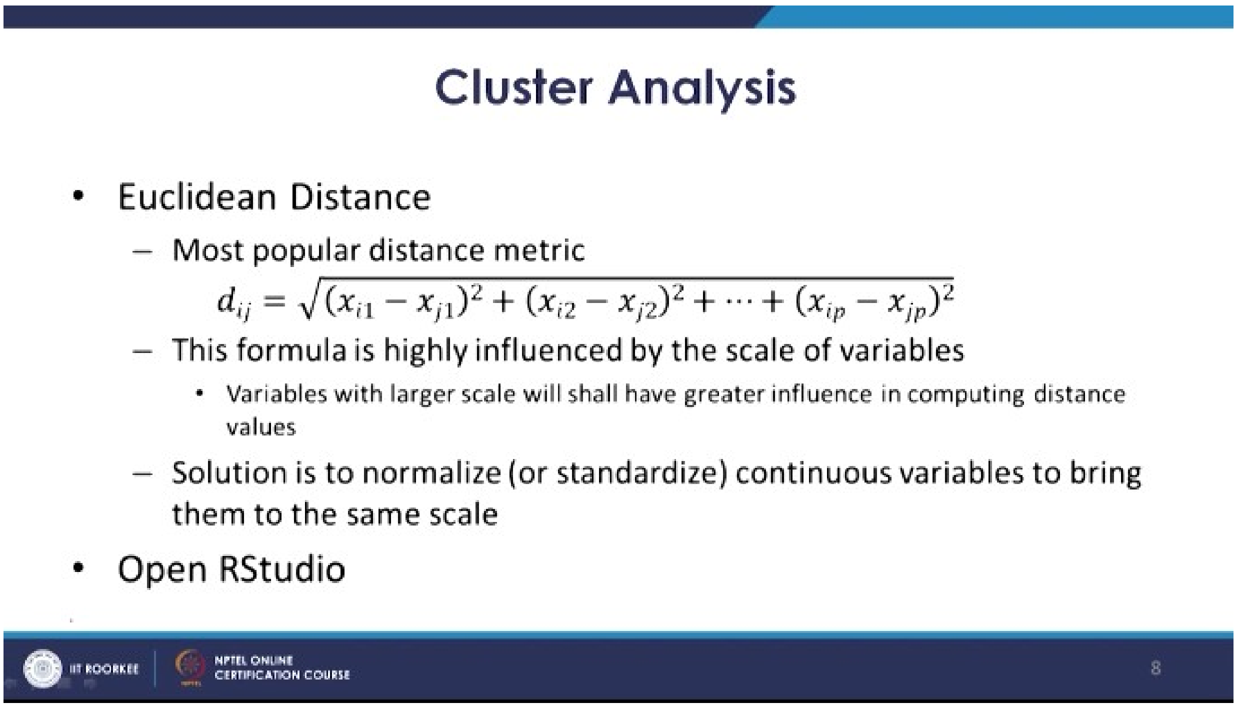


clusters being the another one, so let's start from this point, so we talked about how distance matrix can be also be conceptualize or thought of as the dissimilarity measures, so many of the you know popular similarity measures can be used because they can easily be converted into dissimilarity measures, and therefore can be used as distance matrix. So the kind of matrix that we are going to use as distance you know to measure distance for our cluster analysis would be the directly based on you know distance kind of formulas or would be based on similarity formulas.

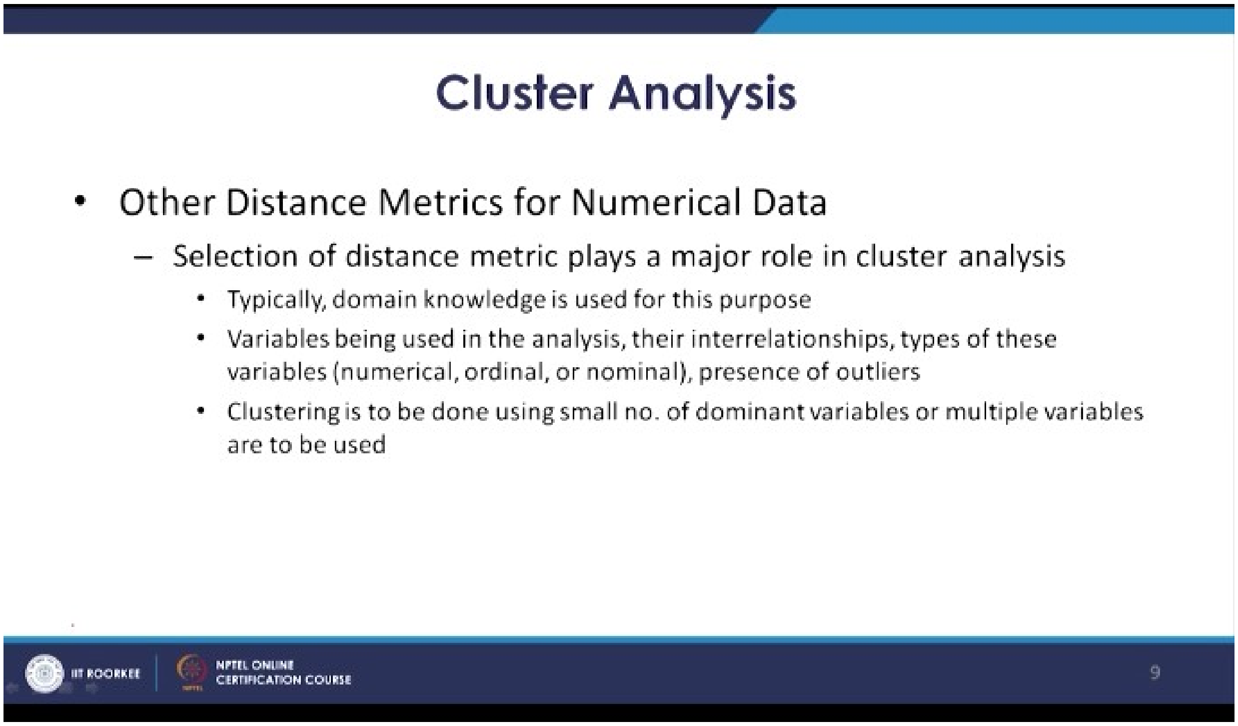


So we started our discussion with this you know two observation, we can think about two observation, first one observation I having coordinate as XI1, XI2 up to XIP, then observation J having coordinates XJ1, XJ2, up to XJP where P is the number of variables to be measured. So P is acting to you know what we use to you know have in our supervise learning techniques, the number of predictors, so the P essentially is the same, however we don't have any you know dependent variable as such in unsupervised learning methods. So we denote, we use this notation DIJ to indicate the distance between two observations.

We also talked about some common properties that has to be you know satisfied for a metric to be defined as to be accepted as distance metric, so these properties are nonnegative, self-proximity, symmetry and triangle inequality, so these properties as we discuss in the previous lecture have to be satisfied.

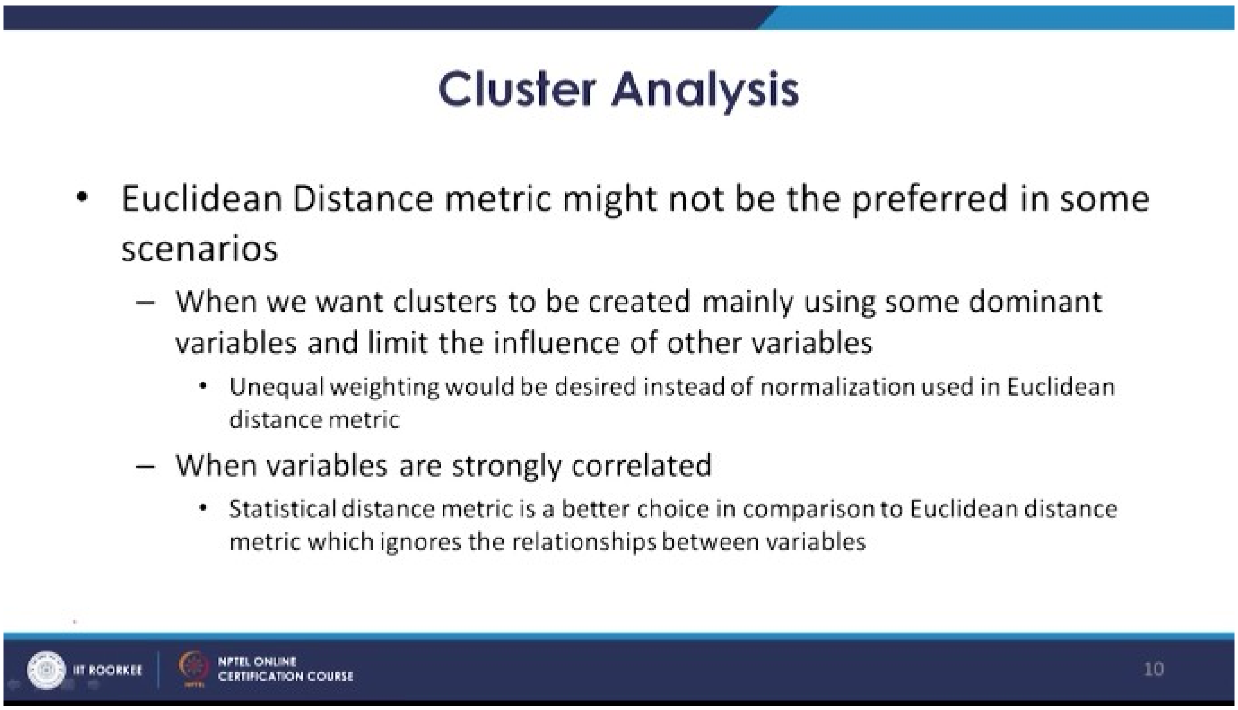


Then these are some of the popular metrics, Euclidean distance metric, correlation based similarity, statistical distance, Manhattan distance, maximum coordinate distance, so these are some of the matrix that we are going to cover in today's this particular lectures, this particular lecture we also, we'll also start our discussion on you know few metrics about categorical data, so let's start with the Euclidean distance, so we were able to discuss to some extent this particular metric in previous lecture as well, so Euclidean distance as we know this is the formula, this being the most a popular distance metric formula as we can see in the slide, this is as we talked about this is very scale dependent formula, so any particular variable which is having or carrying larger scale will obviously dominate the distance competition, will obviously dominate the distance value, therefore we need to you know overcome this particular issue, so one solution is to normalize or standardize the, you know standardize all continuous variable, right, so when we talked about standardize and latest computing statistic code and which is nothing but you know subtracting the actual value by mean value and dividing by standard deviation. So in that fashion we can bring all the numerical variables into similar kind of a scale.



We did a small exercise in R studio for this as well, so let's discuss some other metrics, so what are the other distance metrics for numerical data? So first before we start with them few more points, selection of distance metric why we need? So you know some people might think that we have Euclidean distance metric why not use it all the time, so there are few reasons for this why sometimes we might require to use some other metrics, so few points have been noted here, so selection of distance metrics plays an major role in cluster analysis, typically you know which particular metric is going to be suitable for which kind of problem, which kind of data set and situation you know for that you know domain knowledge becomes important part, so depending on the domain knowledge different distance metrics might be suitable for clustering task as such, so variables being used in this analysis they're entered relationships, types of these variables whether the variables are numerical in nature or categorical in nature, you know whether they are ordinary, nominal or numerical, so whether the out large, whether they are any out large and the data set, so all these things have to be kept in mind, have to be analyzed while deciding which particular distance metric to be used for clustering task.

Clustering you know we also have to think about you know sometime whether the clustering is to be done using a small number of dominant variables or multiple variables are to be used, so this kind of scenario will also decide, will also play a role in terms of what distance metric we would like to use.



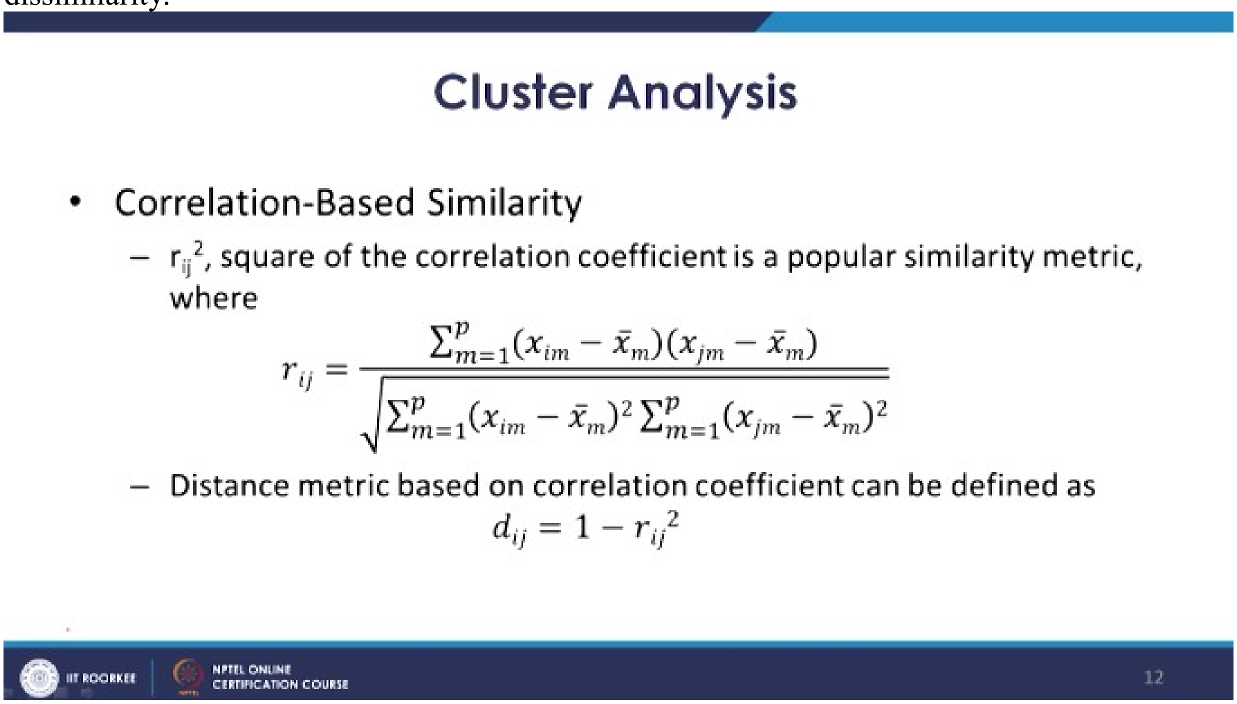
So let's move forward, so some of the scenarios that we talked about, some of the reasons that we talked about we'll be able to understand them better after these few points, so these few points are related to Euclidean distance and how it might not be the preferred metric in some scenarios, so what are these scenarios? So when we want clusters to be created, mainly using some dominant variables and limit the influence of other variables, so this one scenario where Euclidean distance metric might not be useful, because we would have to do unequal weighting and instead of normalization, so for you know Euclidean distance metric to remove it's scale you know dependency we normalize, however so that you know essentially weighting you know kind of equal weighting kind of scenario in Euclidean distance metric, however in some situations, Because we would like to you know account for some dominant variables a bit higher than some other you know, other variables, so in that, in those situation unequal weighting you know becomes important, so therefore Euclidean distance metric might not be the preferred distance metric, because after normalization it would not be able to you know do that.

The other scenario is when variables are strongly correlated, so as we know that, as we this particular aspect we have discussed in our previous course as well, so Euclidean distance metric cannot account for correlation you know between variables, so there is another metric called statistical distance metric or Mahalanobis you know distance which we have discussed in the previous course as well, so this metric is a better choice in comparison to Euclidean distance metric, because it can account for the correlations between variables. So we'll see how this particular metric is going to be useful in such situations.

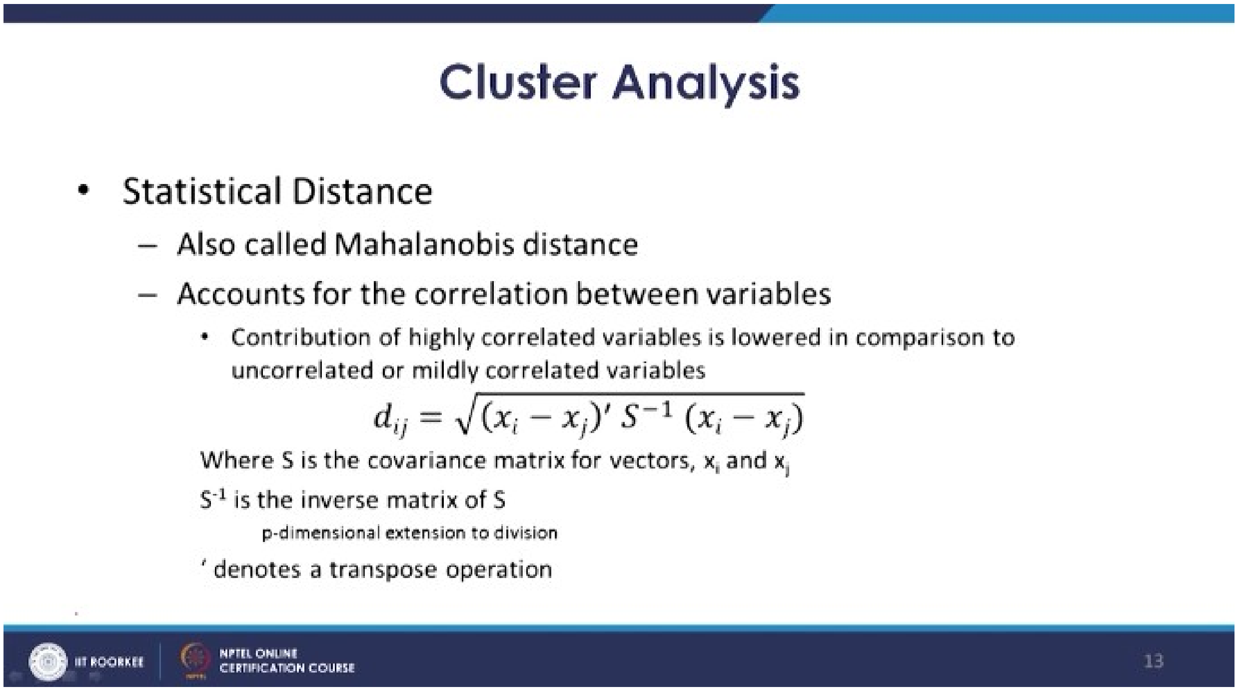
In other scenario where Euclidean distance metric might not be preferred is, when outliers are present and removal of these points is not desired, so as we can understand that Euclidean distance metric if the outliers are present then they can significantly influence the distance computations, and if we you know for some reasons because those points might be very important for the analysis as such, analysis as a whole so therefore we might not be interested in deleting or removing those points, so therefore Euclidean distance metric in such a situation might not be a preferred choice. So we can use you know because, we can use some other distance metric which are robust to these kind of situations, outliers, presence of outliers, so Manhattan distance is one which could be used, so what is Manhattan distance? We'll discuss that, that in the, this particular lecture itself.

So we talked about these few situation, so when we are trying to account for you know dominant variables, we are looking to give them more weightage and then other situation being when we are looking to incorporate the high correlation between variables, and the another situation being the presence of outliers, so these are some of the situations where Euclidean distance metrics might not remain the preferred choice.

So now let's start our discussion of these some other metrics which could be useful in you know different scenarios, so let's start our discussion with this metric, correlation with similarity, so we all understand the correlation coefficient, so this particular metric is actually based on that, so distance metrics measure dissimilarity, so we can also think about as I have talked about while discussion on cluster analysis that distance metrics just like you know similarity measures can be used to convert, can we convert it to measure dissimilarity and therefore can be you know used as distance metrics similarity, distance metrics can also be used to measure dissimilarity.



So similarity measures can be converted to measure dissimilarity and therefore can be read it as distance metrics, so when we think about the correlation coefficient, right, the formula is in this particular slide as you can see how you know correlation coefficient is computed, so when we think about the correlation coefficient it is essentially computing similarity between two variables, and the square of this correlation coefficient can be used as a similarity metric and this kind of formulation which we can see in the last point in the slide, DIJ is 1-RIJ square, so the square of this correlation coefficient now this is being converted into a dissimilarity metric and therefore can be used as a distance metric. So if we look at this particular you know metric so essentially we are trying to measure this similarity between variables and that similarity is being you know converted into dissimilarity, and therefore being used as a distance metric, so this is one metric that can be used to compute distances.



The second one, the more important, one of the more important metrics is this one, the statistical distance, so also called Mahalanobis distance, so this particular metric you know, can account for correlation between variables, in the sense the contribution of highly correlated variables can be lowered in comparison to the variables which are uncorrelated or mildly correlated, so this particular formula as you can see in the slide DIJ square root of you know this XI – XIJ transpose and S inverse, where S is the covariance matrix for vector XI and XJ, and then under square root itself the multiplication where XI – XJ, so you can see S inverse this, is the inverse matrix of S, so this is nothing, in a sense we can think about as S-1 as the P dimensional extension to division.

Then this particular donation, apostrophe donation denotes a transpose operation, so this formula is the statistical, this is how we can compute this statistical distance, and as you can see because the formula itself is accounting for, because of the presence of covariance matrix is, in a way accounting for the correlation among variables, and as I talked about contribution for highly correlated variable is lower, in comparison to the same of uncorrelated of mildly correlated variables.

Cluster Analysis

- Manhattan Distance (city block)
  - This distance metric uses absolute differences instead of squared differences between coordinates

$$d_{ij} = \sum_{m=1}^{p} |x_{im} - x_{jm}|$$

- Maximum Coordinate Distance
  - This distance metric uses the absolute difference between the coordinates which are farthest

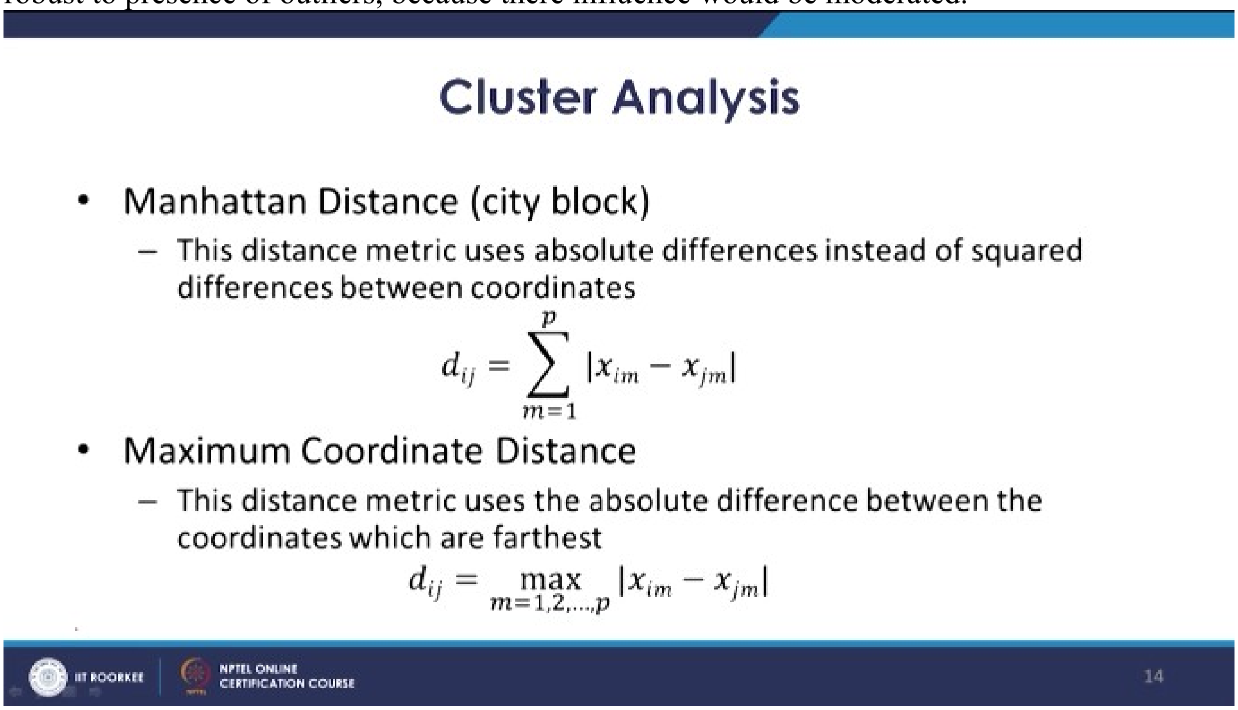
$$d_{ij} = \max_{m=1,2,\ldots,p} |x_{im} - x_{jm}|$$

Then there are few more matrix which can be used, for example Manhattan distance, so Manhattan distance this is also similar to a city block distances that are used, so this distance metric uses absolute differences instead of a square differences between coordinates, as you can see in the formula itself, so DIJ is being defined as summation of M 1 to P, and then absolute difference between XIM and XJM, so absolute of XIM – XJM, so here instead of looking, instead of incorporating the square differences in other formula, in other formulas that we have discussed so far, typically we are you know using the square differences, in this case Manhattan distance it is using the absolute differences between coordinates.

Cluster Analysis

- Euclidean Distance metric might not be the preferred in some scenarios
  - When outliers are present and removal of these points is not desired
    - Euclidean distance metric is sensitive to outliers
    - More robust distance metrics such as Manhattan distance could be used
- Correlation-Based Similarity
  - Distance metrics measure dissimilarity
  - Similarity measures can be converted to measure dissimilarity and therefore can be treated as distance metrics

If you go back there was one important point where we said that Manhattan distance could be useful, let's go back to the point and discuss that particular aspect, presence of you know presence of outliers, so when outliers are present we said that more robust distance metrics such as Manhattan distance could be used, so now you can see why this was mentioned, because the other metric, for example Euclidean distance metric itself is squared differences between coordinates of there in the formula, however if we look at the Manhattan distance which is the absolute you know difference between coordinates is taken there, so if the outliers are present the squareddifferences will actually lead to you know more you know variation, and more influence being accounted, and more influence of outliers being accounted, but if we use the Manhattan distance because it is just using the absolute difference between coordinates so the contribution of outliers is a bit lower, so comparatively Manhattan distance is going to be more robust to presence of outliers, because there influence would be moderated.

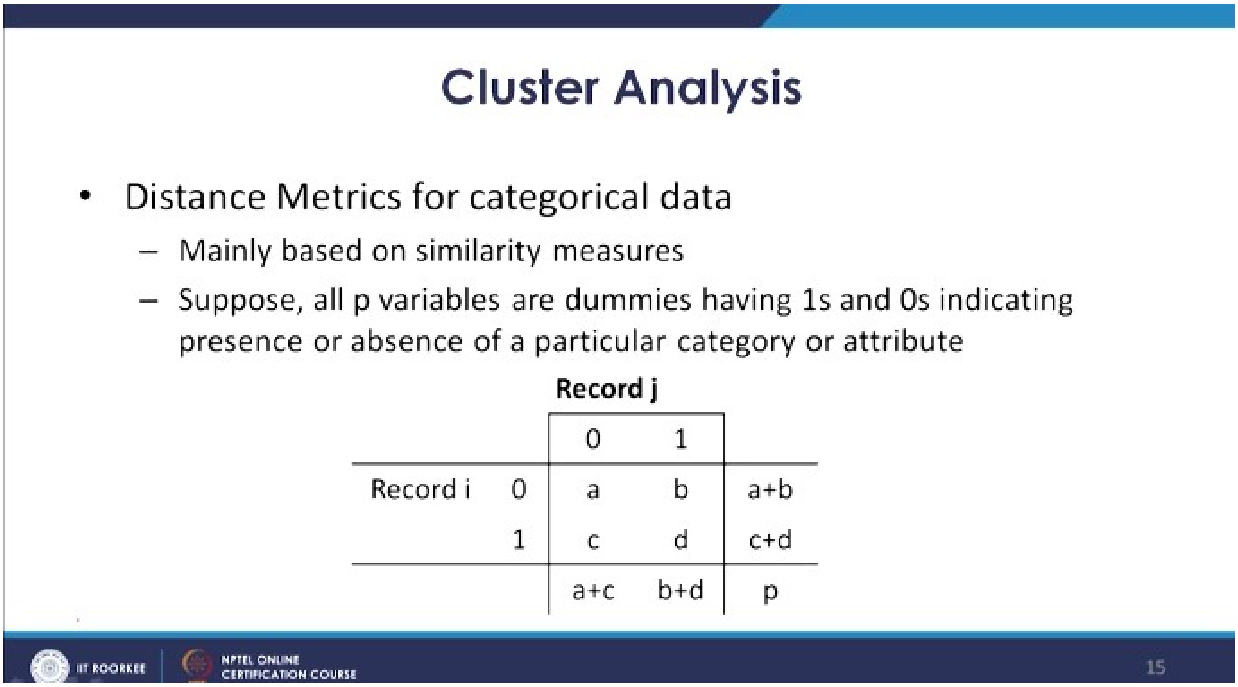


So let's go back and, so there is one more metric that could be used to compute distances for numerical data, so this is called maximum coordinate distance, so this particular distance metric is again going to use the absolute difference but this is you know just between the coordinates which are the farthest, so if we look at the you know other distance metric there more often they're not using summation, you know so the different you know coordinates we are taking differences or absolute differences or squared differences, and then we are summating all those values and using the final value as the distance you know, final distance value between two observations. However if we look at this particular metric maximum coordinate distance, we are just you know identifying you know two particular coordinates which are farthest in comparison to other pairs of points, other pair of coordinates, so the same is reflected in the formula you can see DIJ = maximum of M, 1 to P and then we are taking the absolute difference XIM – XJM, so we are just looking to you know, looking to identify the coordinates which have the maximum deviation, so that this particular difference is being taken as a distance, so that is why the name is also resting the same, maximum coordinate distance, so sometimes you know this particular distance metric can also be useful, so we'll see that, we'll see that while our discussion of cluster analysis in coming lectures as well.

Now all these metrics that we have discussed so far they involve you know, they involve numerical data so typically all these metrics that we have discussed maximum coordinate distance, Manhattan distance, statistical distance, you know correlation dissimilarity and the Euclidean distance all of these metrics there for numerical data, so what if you know, so what if the categorical data is present, so we would still like to perform the task of clustering even in the presence of categorical data, so are there any metrics which could be used for distance computation, because if you think about distance typically it is you know concept that we you know naturally associate with the numerical data, but you know distance can it be, can this concept can be extended to categorical data, can we do some sort of quantification and use that you know, use the same for our cluster analysis to perform a task of clustering.

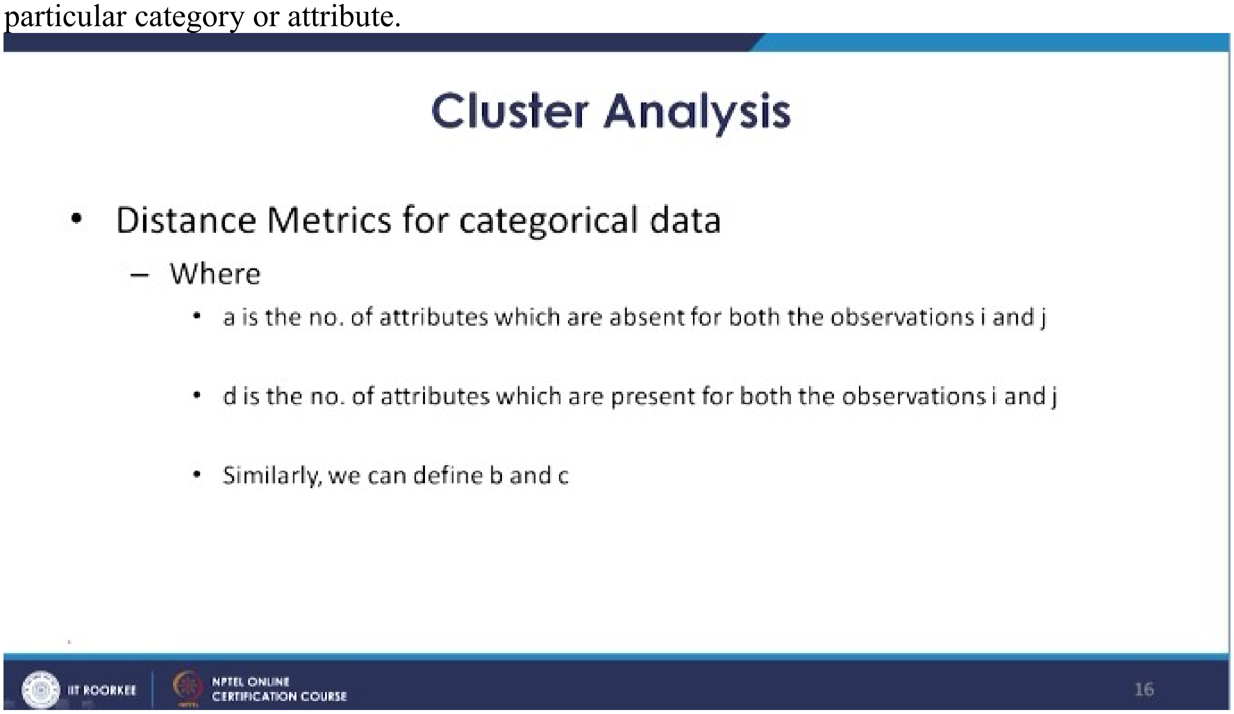
So yes, there are metrics which could be used for categorical data, so we are going to discuss few of them, so if we look at the distance metrics for categorical data they are mainly based on similarity measures which makes sense also because the distance computation as such is, because that is typically applicable for numerical data as we talked about, and we'll not make sense for categorical data, so essentially what we are looking for when we say distance metrics for categorical data we are essentially trying to measure the similarity or dissimilarity, so we can always talk about the you know similarity between you know two observations based on the attributes that they have for a particular categorical variable. So two observations might have the same, might have the same attribute for a categorical variable, for example if we are talking about the you know status where the particular you know particular record, particular customer is, the customer is a student or employed or retired, so in that case two observations, two of the customers might be having the same attribute, they both might be student, they both might be employed, or they both might be retired, or they might be different, one might be student, one might be employed or the other one might be retired, so we can always look at the similarity, so if both are student then we can take that as an evidence of similarity, if you know both are different, one is a student then other one is employed we can take that as evidence of dissimilarity.

And now you quantify this information and use this in a, you know, distance you know, like as a distance value for our cluster analysis. So the main idea is you know similar to this what I just discussed, so mainly based on similarity measures.



		0	1	
Record i	0	a	b	a+b
	1	c	d	c+d
		a+c	b+d	p

So now you know extend this thought to our you know P dimension, so suppose all P variables are dummies, so right now we are only thinking about the categorical variables, so you know the variables that we have in our final you know tabular format, right, so as we have discussed in our previous course that if the categorical variables are having more than two categories we can always create dummy variables and you know then those dummy variables are then used for you know formal analysis, right, so similarly we can think about all our P variables to be dummy variables and they are having ones and zeros indicating presence or absence of a particular category or attribute.

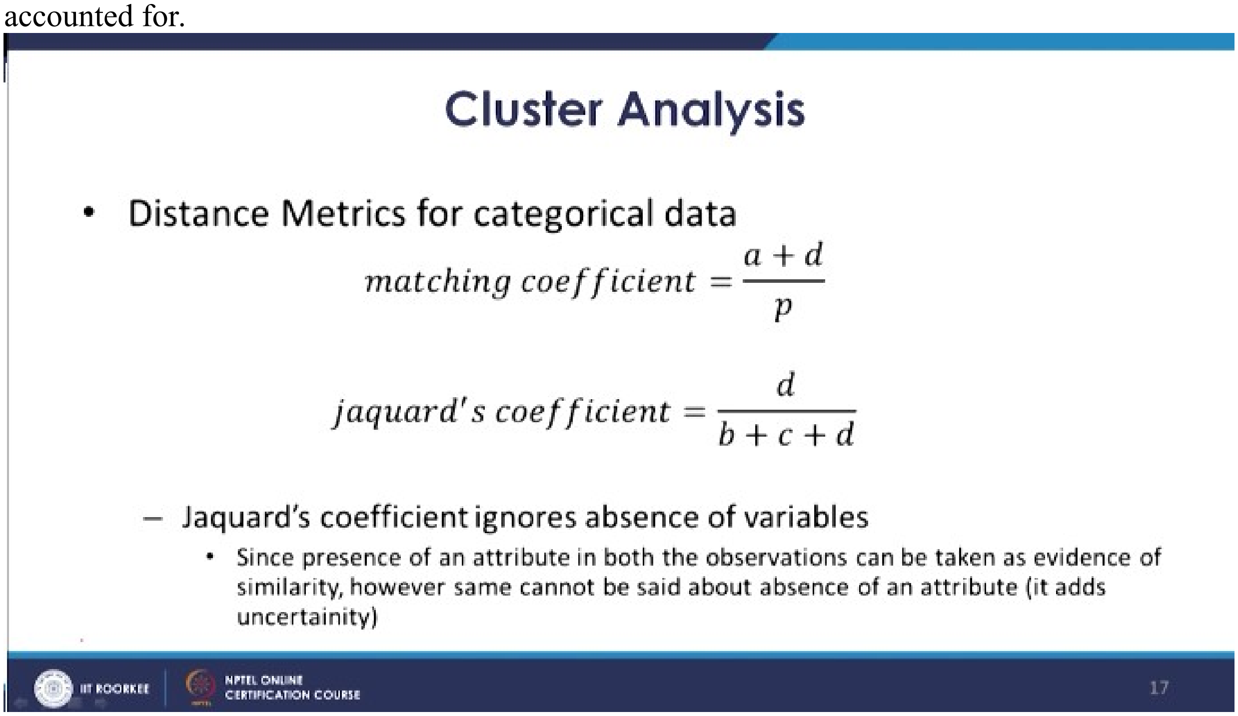


So as you can see here we have a small table here which has, which we are going to use to understand this particular concept, so we have record I and record J, and both these record, both these observations have you know, we have you know P variables and this particular table 0 and 1, and 0 and 1 as you can see is indicating the number of variables, so for example record I0, and record J0 so A is there, this A is actually you know indicating the number of attributes which are present, which are absent for both the observation I and J, 0 denoting the absence, and 1 denoting the presence, and these numbers are actually denoting the number of those attributes which are present or absent.

So in this case since 0 is denoting the absence for record I and record J, A number of attributes are absent you know for you know both these records, so the same thing is mentioned here in this slide, A is the number of attributes which are absent for both the observations I and J, if we look at the another you know value here in this table D, so D is the one indicating presence, so D essentially is indicating the number of attributes which are present in both the records I and J, both the observations, so this is what we have mentioned here, D is the number of attributes which are present for both the observations I and J.

Now there are going to be certain attributes which you know might be present for 1 and absent for the other one, right, so you can see here C, C is the number of attributes which are present for record I, which is against to 1 in record I, but against 0 in record J so we can see that C denotes the number of attributes which are present for observation I, but absent for observation J.

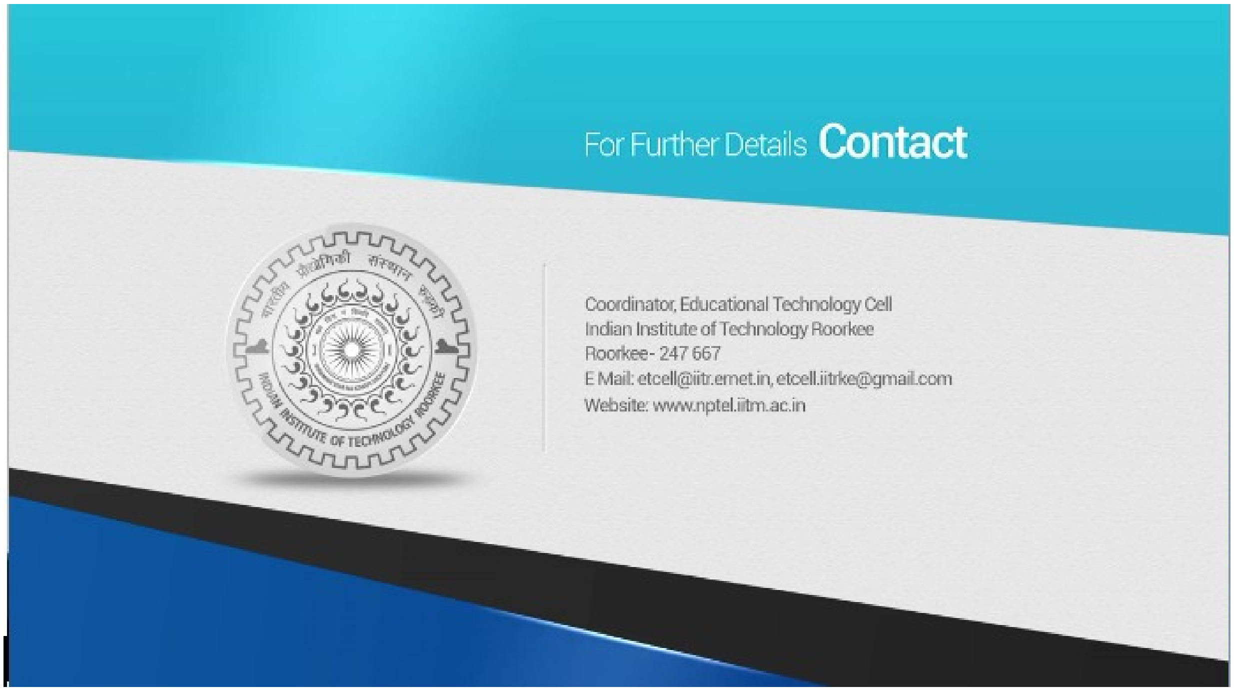
Similarly B is the number of you know attributes which are present in, present for observation J but absent for observation I, so these numbers tell us some story so if we look at this particular value D, so D indicating the number of attributes which are present in both the observation there by telling us you know giving us some sort of evidence of similarity, so D is essentially standing for the similarity between two observations, these two observation I and J. If we look at A which is again indicating the absence of something, so if we are okay with this idea that if something is absent, something is absent between two observation and taking it as a you know, you know taking it as a you know similarity between two observation this value A can also be accounted for.



However if we look at the B and C, so these two numbers you know definitely indicate the dissimilarity between these two observations I and J. So if we look at this information this particular information that we have just understood can be used to quantify the similarity and dissimilarity, so let's look at the metrics that actually can do this for us, so there are two metrics that we are going to talk about, one is the matching coefficient, the another one is Jaquard's coefficient, so if we look at the matching coefficient the value is, the way we define it is A + D divided by P, where P is the you know total number of attributes, so it is the summation of A + B + C + D, so if we look at the, we focus on the numerator part, so in the numerator we have A + D, so indicating the number of attributes which are absent for both the observations, and D indicating the number of attributes which are present for both the observation, so both A and D you know presence of number of attributes and absence of a number of attributes we are taking that as a you know similarity and we are you know using that in the this numerator of, numerator part of this particular coefficient, matching coefficient, and then dividing it by total number of attributes, so this, it gives us a ratio which we are taking as a you know this proportion in a way indicating the similarity between these two observations.

However some people might argue that absence of something, if something is absent you know in both the observation that would hardly indicate the similarity between two observation, rather it is you know at some uncertainty you know between two observation in terms of similarity or dissimilarity, so there is another coefficient which doesn't take the number of you know this value of A into account, so Jaquard's coefficient you can see in the numerator part and the denominator part, and both the parts the value of A has been taken off, so A is has been ignored because absence of you know particular attribute in both the observation is not being taken you know, as you know similarity or dissimilarity, it is being considered you know we are uncertain about it whether you know these you know attributes which are absent whether they would have indicate you know some similarity or dissimilarity, so A has been taken out of this formula so in the numerator we are just left with D, so D indicating the number of attributes which are present in both the observations, and in the you know denominator we have just B + C + D and therefore we get a particular ratio, a particular value, a particular ratio value which will indicate the similarity between two observation. If the same point is summation here in the slide also that Jaquard's coefficient ignores absence of variables and so you can see further that since presence of an attribute in both the observations can be taken as evidence of similarity, however same cannot be said about absence of an attribute you know since you know it acts, it brings uncertainty so therefore you know the value A has been ignored.

So you know with this we'll like to you know end this session here, and in the next session we'll start our discussion of this distance metrics for both numerical and categorical data, so we're more likely to encounter scenarios where in our data set both kind of variables, numerical and categorical both kind of variables are present, so how do we go about you know doing our, you know, task of clustering, so for that we need some similarity metric or you know distance metric which can incorporate, which can be used for both kind of data numerical and categorical data, so there is this one metric Gower's metric similarity metric, so we'll discuss about this one in the next lecture. Thank you.



For Further Details Contact
Coordinator Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee – 247 667
E Mail:-etcell@iitr.ernet.in, iitrke@gmail.com
Website: www.nptel.iitm.ac.in

Acknowledgement
Prof. Ajit Kumar Chaturvedi
Director, IIT Roorkee
NPTEL Coordinator
IIT Roorkee
Prof. B. K Gandhi
Subject Expert
Dr. Gaurav Dixit
Department of Management Studies
IIT Roorkee
Produced by
Mohan Raj.S
Graphics
Binoy V.P
Web Team
Dr. NibeditaBisoyi
Neetesh Kumar
Jitender Kumar
Vivek Kumar
Dharamveer Singh
Gaurav Kumar
An educational Technology cell