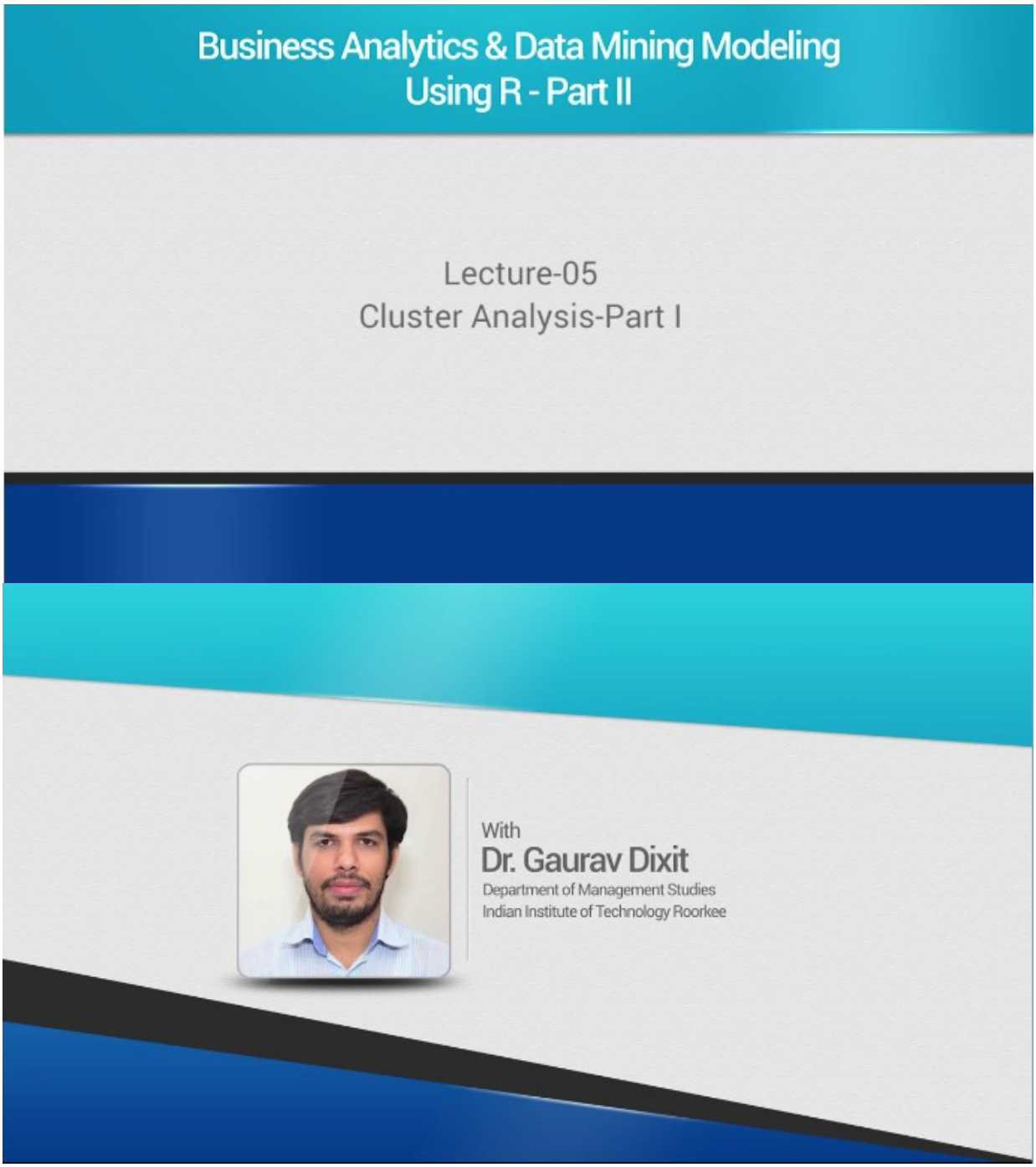
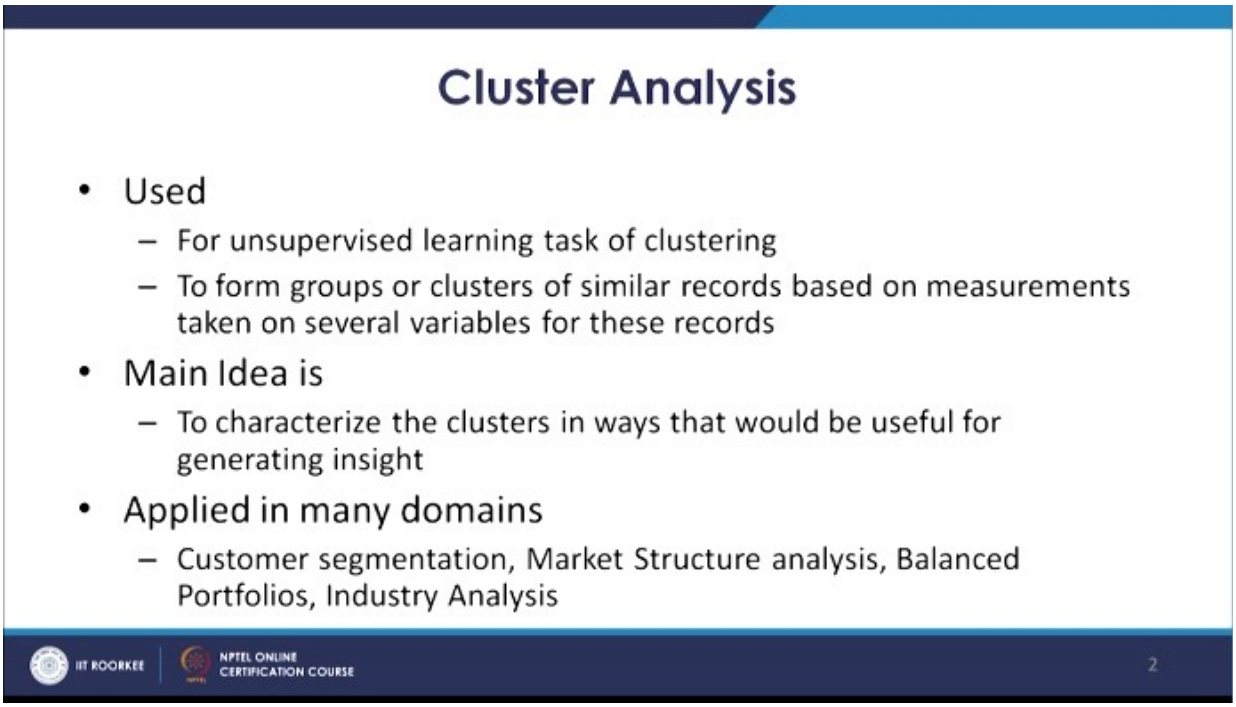
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
NPTEL
NPTEL ONLINE CERTIFICATION COURSE
Business Analytics & Data Mining Modeling
Using R – Part II
Lecture-05
ClusterAnalysis– Part I
With
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee



Welcome to the course Business Analytics and Data Mining Modeling Using R Part 2, so in this particular lecture we are going to start our discussion on cluster analysis, so let's start. So cluster analysis is used for unsupervised learning task of clustering, so as we have discussed in previous course that in supervise learning method we typically have two types of task, one is prediction, the other one is classification.



In this particular course we have already discussed association rules and this is the second technique that we have started our discussion on cluster analysis, so clustering and association rule mining are the two unsupervised learning task, so this particular technique, this particular method cluster analysis is used for unsupervised learning task of clustering, so used to form groups or clusters of similar records based on measurements taken on several variables for these records, so the data for this particular method technique is the tabular format of data that we typically use for most of the data mining techniques, and the variables are going to be on the column sides, and the observations would be on the row side, and the measurements on these variables are going to be used to find similarity among observations, and then based on that similarity we will try and create groups or clusters of those observations, so whole idea of cluster analysis is to identify and form these clusters based on some similarity you know the similarity can be you know computed using the different you know variable measurements.
So as you can see in the next point so the main idea behind cluster analysis to characterize the clusters in ways that would be useful for generating insides, so why we want to create these cluster or groups, because we are interested in some specific characteristics of those groups or clusters and these characteristics can throw some insides, some light for further research, so cluster analysis in a sense as we have talked about in previous course as well that typically unsupervised learning methods, unsupervised learning techniques they are used for the exploratory researched, so first you know research, you know very first research the studies in a new field, typically used you know these techniques unsupervised learning techniques, and once some idea about that particular new field or domain is established it is only then supervise learning techniques you know become useful, so main idea is when we are learning about, when we are venturing into new field and we are trying to learn something about you know different concepts, theories, models, variables that are involved, you know, these unsupervised learning techniques provide us the tool to go ahead, so one of them could be characterization of different things, finding out natural hierarchies you know among different things that are there in that field, so cluster analysis is one way to characterize you know to form these groups or clusters or natural hierarchies and characterize them, so using different variables, using

information that is there containing those variables, we can always find similarities and the similarities and form clusters and this can later on be used to characterize the same clusters. Application of cluster analysis is in many domains, so you know if we look at just the management area and customer segmentation, you know, market structure analysis, balanced portfolios, industry analysis, so all these are some of the examples where cluster analysis has been used and you know remains very popular. For example industry analysis, so how do you, you know, club or you know form you know, you know forms in unorganized sectors and clubbed them and find clusters, find similar forms working on similar kind of product process or similar kind of you know, practices that they might have, and some sort of variables to be used and create clusters and to identify forms in different groups, different segregated groups, so for that cluster analysis can be really useful, so industry analysis can be you know, cluster analysis can be useful in industry analysis, for example performance you know so, whether the firms, which firms are highly performing, which firms are you know not doing so well, and what we're depending on the different variables that could be used, we can create those clusters, high performing you know firms, you know average firms, and then poorly performing firms, so all that kind of industry analysis can also be done.

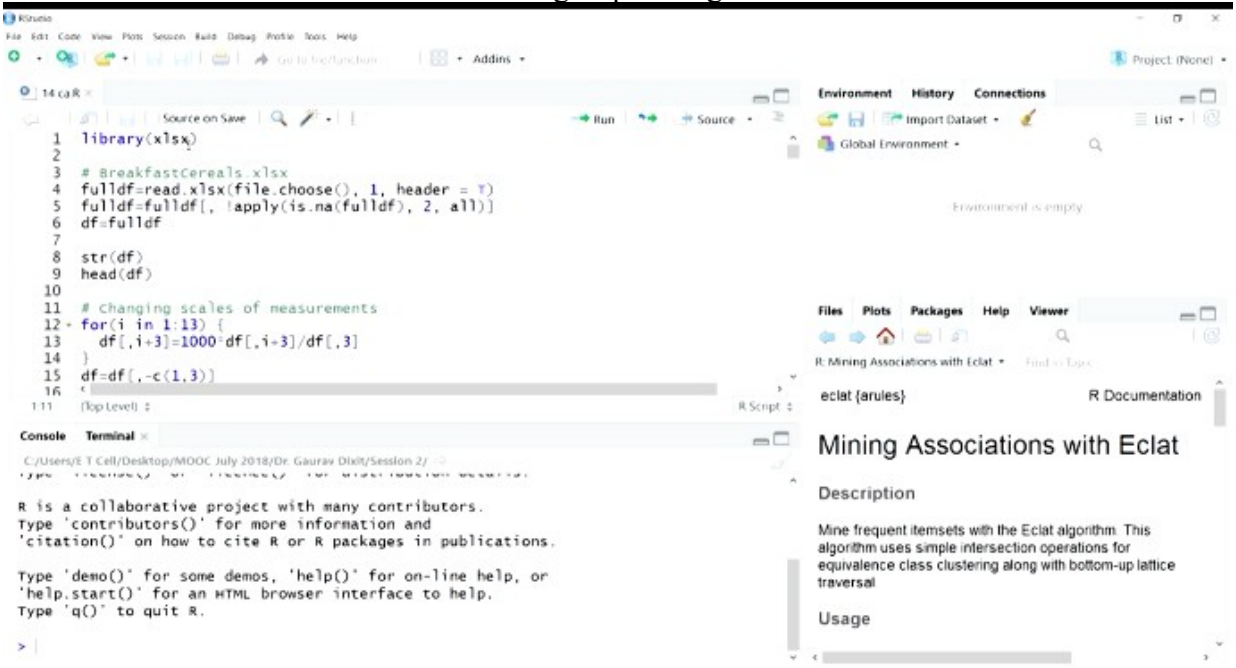
Balanced portfolio so as we all know that, we tried to create balanced portfolios to mainly you know cover the risk, so we would like to spread the risk among the various portfolio items that we you know investing, so how do we do that? So cluster analysis can really be useful to identify high risk you know, high risk items, low risk item and average risk items and then balanced out the investment portfolio in a sense, so cluster analysis can be useful there as well, market structure analysis so there are so many products that are available, that are being marketed and sold in different industry sectors, so how do we identify similar products or you know, if there are any dissimilarity between some products and senses, so that kind of grouping or clustering can also be done using clustering analysis.

Similarly customer segmentation based on some demographic information or personal characteristics of those customers, now we can apply cluster analysis and find out the similarity and dissimilarity and then create segments, and then develop our marketing and promotional strategies, differently for each of those segments, right, so these kind of you know applications, cluster analysis has been this kind of areas are there, where cluster analysis has been applied very successfully.

Cluster Analysis

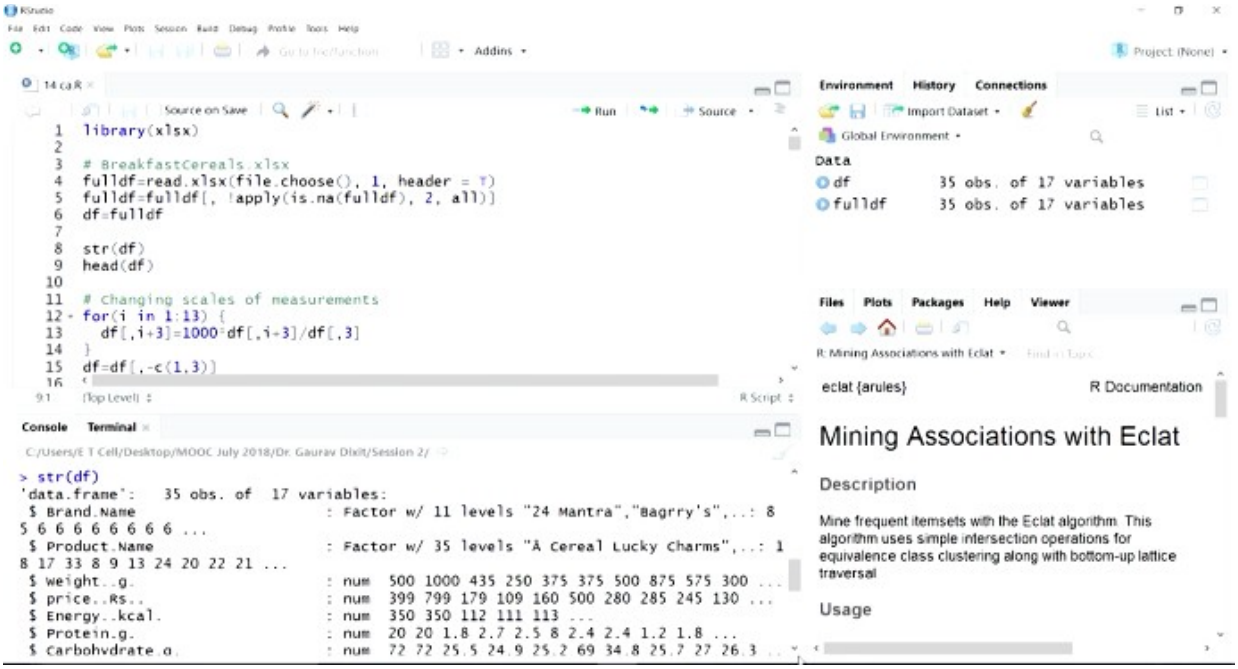
- Example: Breakfast Cereals
  - Records to be clustered are cereals
  - Clustering would be based on eight measurements on each cereal
- Open RStudio
- Types of Clustering algorithms
  - Hierarchical Methods
  - Non-Hierarchical Methods

So these are some of the examples in the management or business domain, how about cluster analysis remains popular in other science and engineering disciplines, so during the our discussion on cluster analysis will, you know, will have this, will discuss this example as well, breakfast cereals examples, so we have a number of you know cereals in this particular dataset, and information on you know different nutritional ingredients and other things for example rice and wheat and other things that is available in this particular dataset, and we would be using this dataset to find out different clusters of group among these cereals.

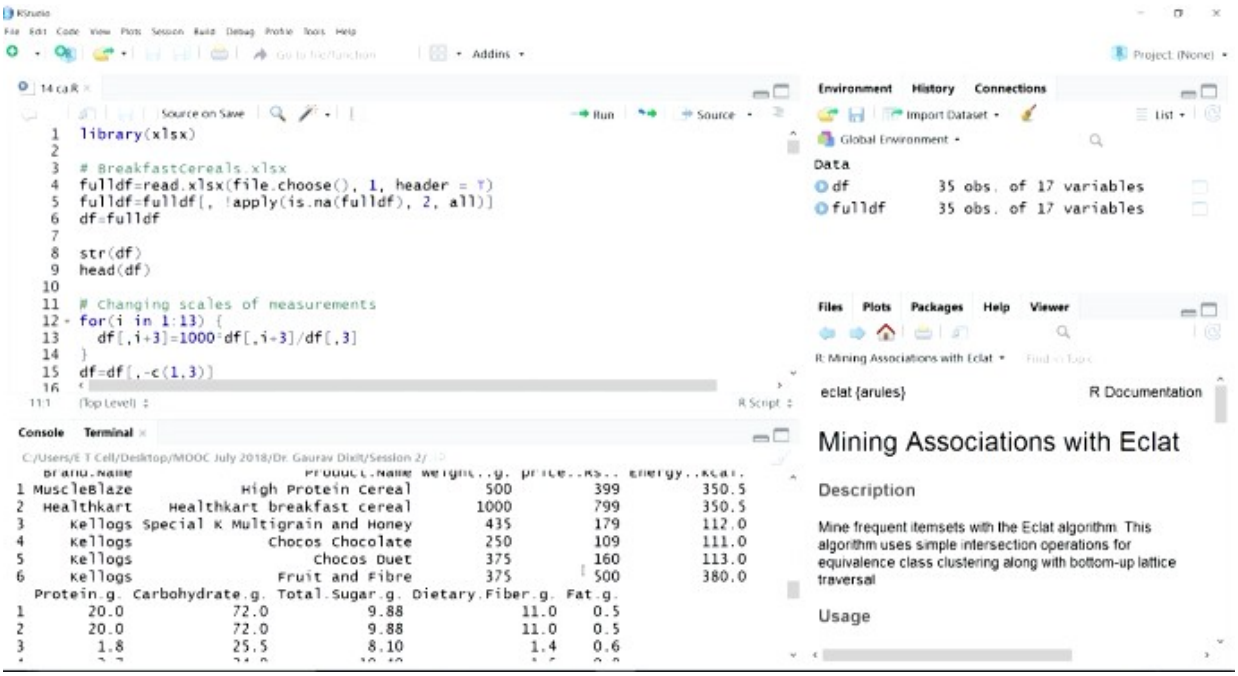


So as you can see here records to be clustered or cereals and clustering would be based on 8 measurements on each cereal, so we have data on 8 variables for these cereals, so let's look at this dataset because we would be referring to this particular example in our discussion on

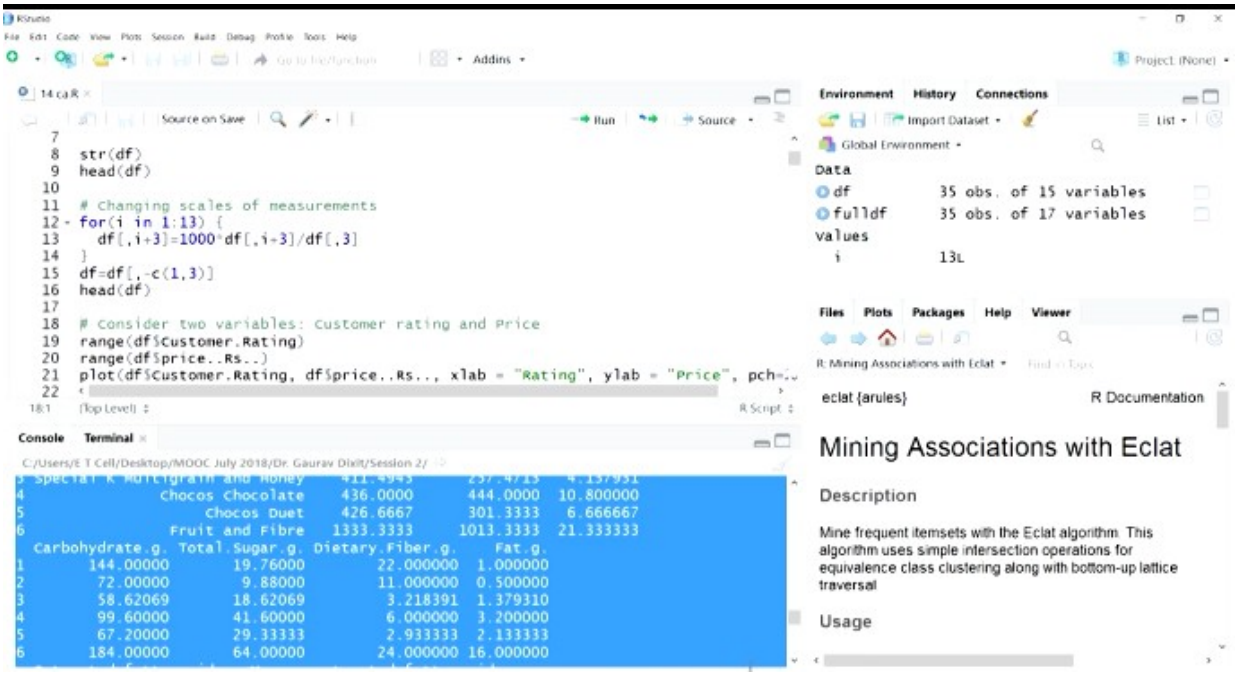
cluster analysis. So let's go back to R studio and let's have a look of this particular dataset, so let's load this library, library xlsx, so this would allow us to use functions which can be used to import XL data, so you can see the file that we are going to import here is the breakfast cereals slot xlsx so let's run this code.



So you would see this file is right here, it's imported you can see in the environment section here, 35 observations of 17 variables right now so we will be getting rid of some of these variables as we'll see, so total 35 cereals data on 35 cereals we have, and let's remove the redundant columns here and let's take a new copy of this dataset, let's look at the structure, so now we can see the kind of you know one measurements that we have took here on the cereals, so you can see 35 observations, 17 variables so first one is the brand name, second one is the product name then we have weight of those offerings, those market offerings products and then we have price of those packages, and then the energy, and the units are also mentioned as you can see, energy, protein contain carbohydrate contain the total sugar that is there, dietary fiber, fat, saturated fatty acid, monounsaturated fatty acid, polyunsaturated fatty acids, so different natural and other ingredients and other information about the cereals is available, cholesterol, sodium content, iron content and then we also have the customer ratings, how customers have rated these cereals, so all this information is available with us, so we would be using this particular dataset in our discussion of cluster analysis.



Let's look at the first 6 observation of this particular dataset, so let's scroll for first few columns, so as we discuss brand name, then product name, then weight, so then weight and pricing, so here it is important to notice here that the weights for these you know these cereals are different and therefore you know other variables, for example price and other ingredients, energy, you know other information like energy and other ingredients, protein, carbohydrate, they would also be you know they would not be in the same kind of a scale, so because the weights for this factories are different, so we need to do certain computation, certain transformation for this variables, so we would like to change the scale of these variables, so as you can see here in this particular line of code what I have done, I have tried to bring all these value to the same scale, so now you can see here 1000 is being used, so all these price and other information energy, protein, carbohydrate, etcetera, now they would be for 1000 gram of that particular cereal, so let's run this code and change the scale.



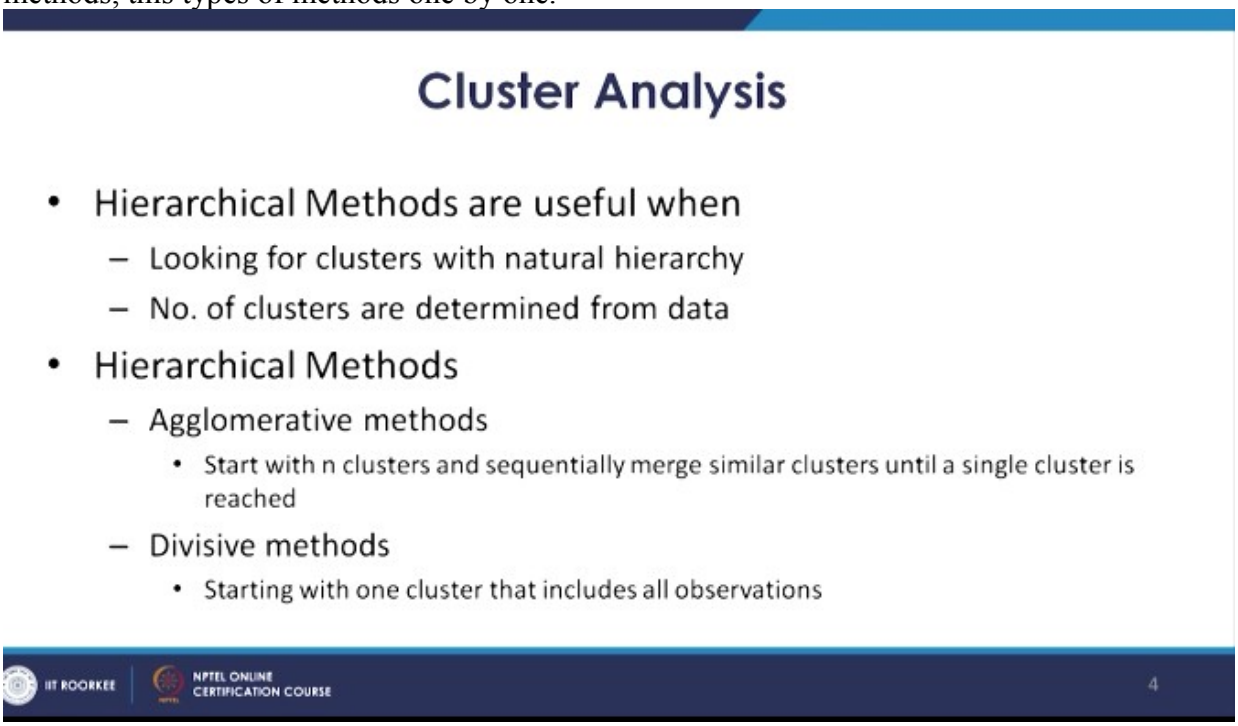
Now after this we would be getting rid of two columns, first one is about the brand name and another one is about the weight, because now weight is already incorporated and these scales have been changed, so we will just use the remaining variables for further analysis.

Now let's look at the first 6 observations again, now you would see first column is now product name, then we have price, now you can see the price values have changed, so these values have been appropriately adjusted or computed for you know 1000 grams of the cereals, and you would see energy values and protein values all these values have been appropriately changed, now we have the information using same scale, so this is the dataset that we have, and this dataset is something that we are going to refer to during our discussion on cluster analysis.

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

Now let's come two types of clustering algorithms, so there are two general types of clustering algorithms, first one is called hierarchical methods, the second one is called non-hierarchical methods, so as I talked about that cluster analysis can also be used to understand the natural hierarchy you know among different things that we could be studying, so hierarchical methods are typically used for that, and then we have non-hierarchical methods, so let's discuss these methods, this types of methods one by one.



So as I said hierarchical methods are useful when we are looking for clusters with natural hierarchy, so a number of clusters are determined from data, because natural hierarchy we typically expect it to be in a certain form, so the number of clusters have to follow that you know that hierarchy, so typically useful, so this particular methods are useful when we are looking for this kind of natural hierarchy and number of clusters that we are interested in, they can always be determined from data, so within hierarchical methods we have two types, first one is called agglomerative methods, the second one is called divisive methods, so what's the difference between these two methods, so let's see.

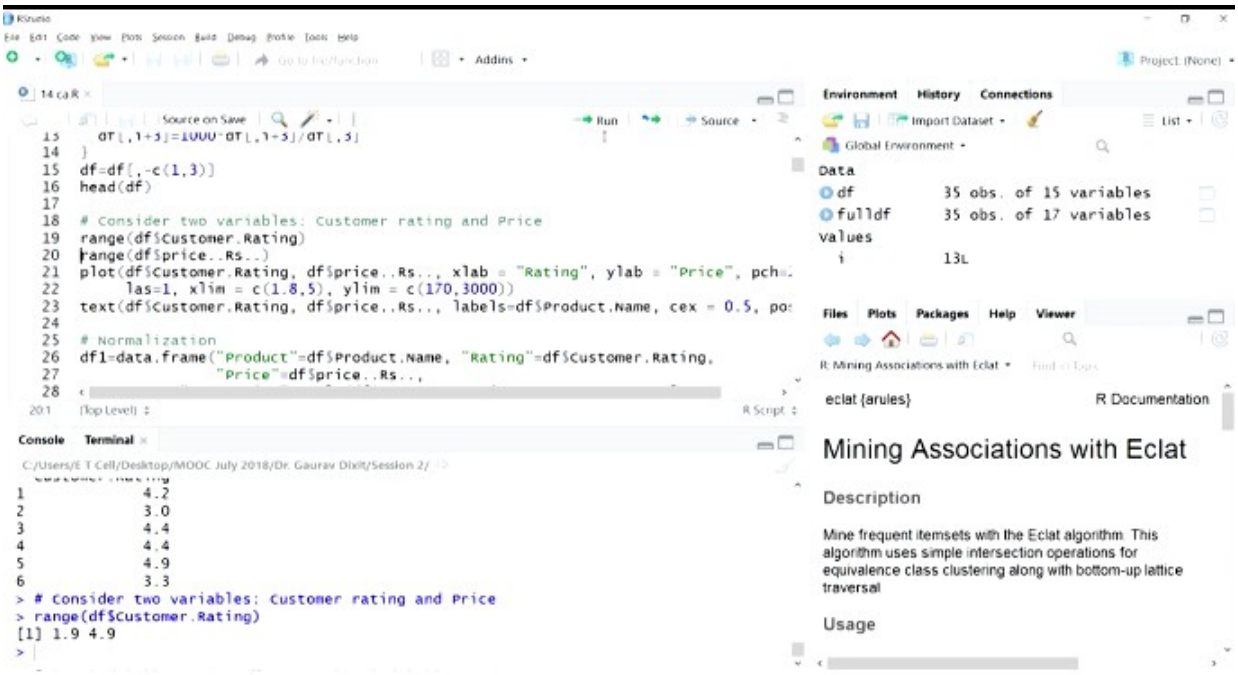
So in agglomerative methods we start with N clusters and sequentially merge similar clusters until a single cluster is reached, so that is why the name comes agglomerative, so we are creating a aggregation, so we start with N clusters, and the identified, we try to identify similar clusters and keep merging them until we reached to one particular clusters, so this is the typical process that adopted in agglomerative methods.

If we have to talk about the divisive methods, so it is just the opposite of the agglomerative methods, so here we start with one cluster which includes all the observations, and then we divide it into a number of clusters depending on the similarity or dissimilarity, so it is just the opposite process. So typically agglomerative methods are more popular, so in this particular you know discussion, in our discussion of cluster analysis, we would be using agglomerative methods.

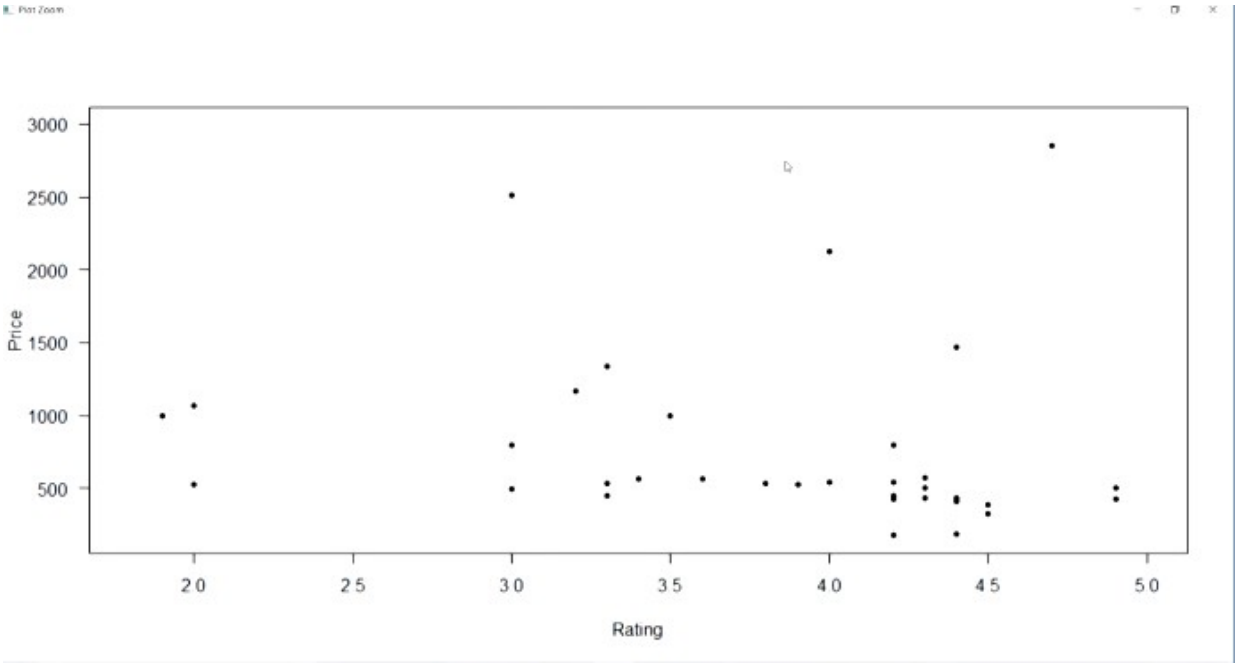
Cluster Analysis

- Non-Hierarchical Methods
  - No. of clusters are pre-specified
  - Generally less computationally intensive and are therefore preferred with very large datasets
  - k-means clustering
    - Observations are assigned to one of the pre-specified no. of clusters
- Open RStudio
  - Cluster analysis can be thought of as a formal algorithm that uses distances between observations as dissimilarity measure to form clusters

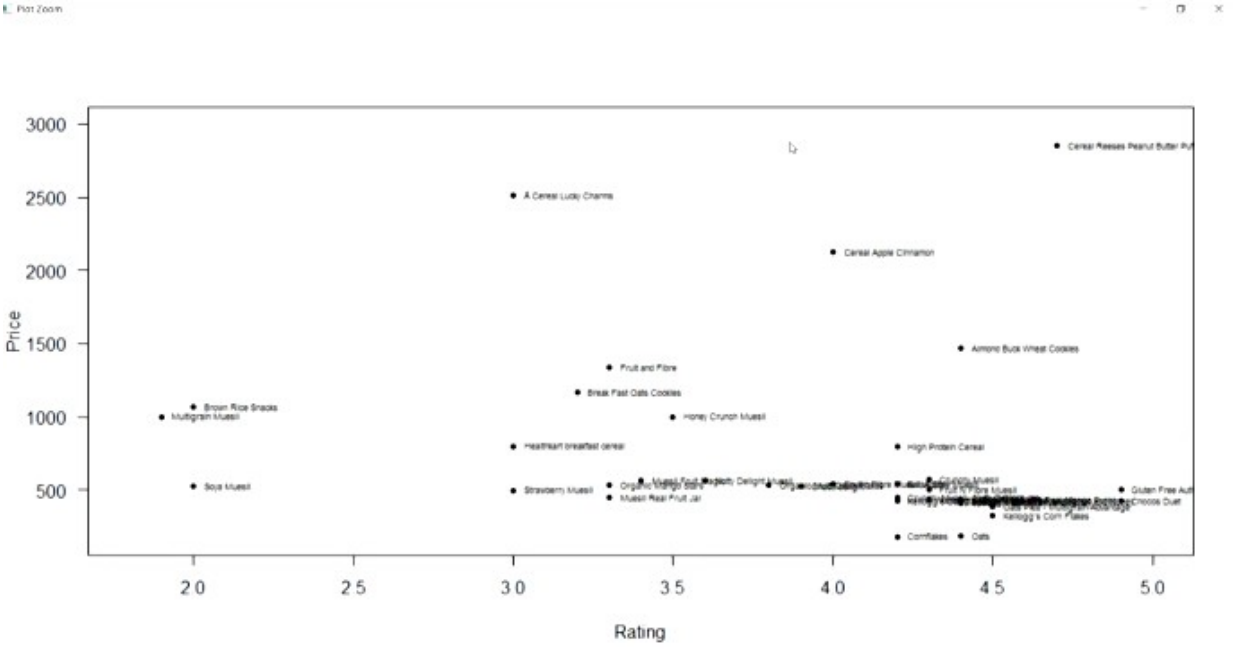
So let's move to next category of cluster analysis that is non-hierarchical methods, so in this case number of clusters are typically pre-specified, so we typically have you know we have you know have idea about due to about domain knowledge and other things that these would be the number of expected clusters and those clusters are typically pre-specified, so this particular technique is generally less computationally intensive, and preferred with large, very large datasets, so typically one of the technique that is quite popular is, K means clustering, so popularity is also because of the lower level of computational intensity, so in this particular technique K means clustering, observations are assigned to one of the pre-specified number of cluster, we can see the task is much you know simpler in this case, so we have the number risk specified number of clusters and the observations are assigned to one of them depending on the similarity or dissimilarity.



So let's look at, let's go back to R studio and go through our you know one more exercise to understand a bit more about cluster analysis, so let's consider two variables in our database of breakfast cereals, so these two variables are customer rating and price, so we'll consider these two variables and try to you know create a scat of lot using these two variables and understand about the groups or clusters that can be visually seen there, so let's look at the range of these variables, so customer rating, so you can see here it seems that customer rating you know is between lower range is 1.9, then 4.9, so it seems that you know rating options were 1 to 5, and then we have the price, let's look at the range, so it is from 180 to 2853 this value, so this is the range now this values have been used in this plot function to specify the limit so that we are able to use the most of the graphic space that is available with us, so let's plot this, we can see X axis you know rating is going to be plotted on X axis, and price is going to be plotted on Y

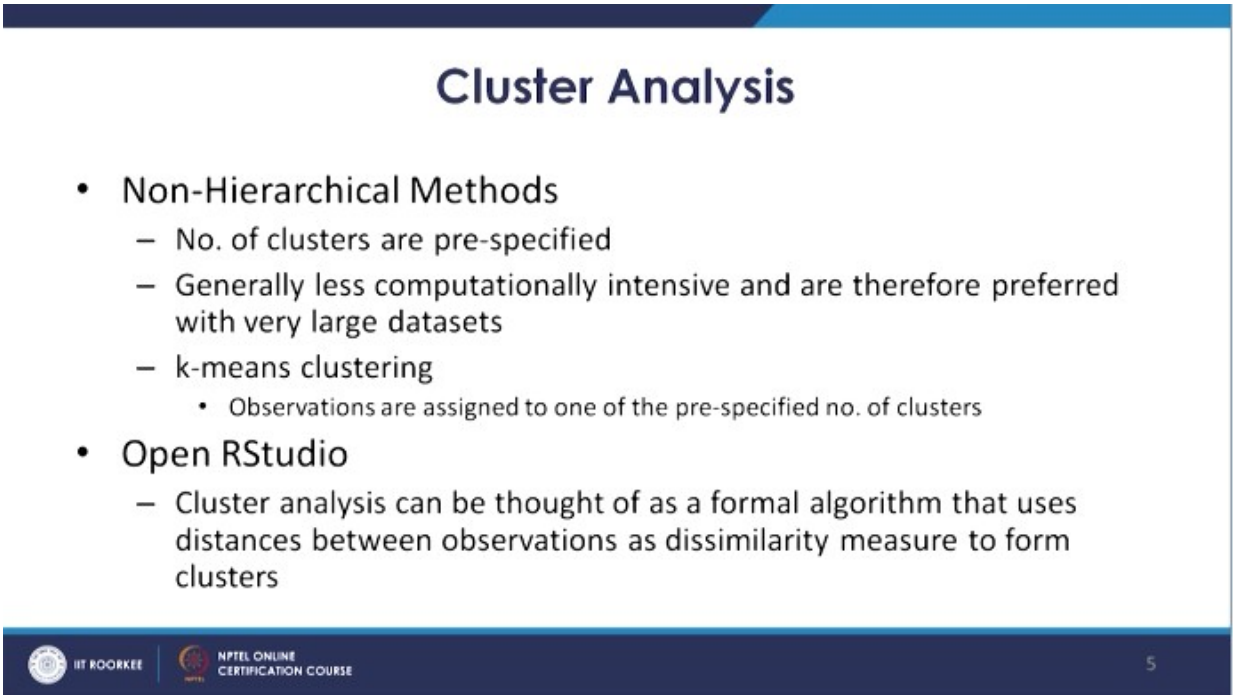


axis, so let's run this code, so you can see a plot has been generated, so if we zoom into this plot and so here if you look at this particular plot we can identify certain groups of you know observations, for example here these three observations right, example these three observations we can use here, so these three observations seem to be in one group, then if we look at here so these observations, this group of observations seems to be forming one cluster here, and then we have this group observation they're again seem to be forming one cluster here, and again these three points they're again can be thought of as forming a one particular cluster.



This one seems to be outline you know distant frame all of these clusters, so typically we can see four clusters here, so let's switch out, so now all these points, because all these breakfast cereals they have, they have you know product information we have, so that text can also be added on to this plot so let's run this code text function and you would see all the product

names, you see here or the product name, let's zoom in, you can see against all these you know breakfast cereals we also have attached the name of that products, so in this fashion we would be able to see the name on the cereals and get an idea about which products are being grouped with what other products, so if we look at the process that we have done manually here, we try to identified, we try to identify few groups from this database, so in a way in using this V, you know unknowingly we used this two dimensional distance here, and we try to identify these groups, so group one, group two, you know, cluster three group three, and four, so in a way using these two you know X axis and Y axis using these two axis we tried to you know use the two dimensional distance based on these two variables.



So in a way the clusters that we create they incorporate the you know similarity, dissimilarity using the distance matrix, right, so that brings us to our discussion back on certain important points about cluster analysis, so as you can see here the last point in the slide cluster analysis can be thought of as a formal algorithm that uses a distances between observations as this similarity measure to form clusters, right, so in the clauses that we did using this cutter plot you know inherently we used the distances, so the cluster analysis as such can be thought of a formal way of doing it where the distances between observation are being used as this similarity, so higher the distance more the dissimilarity, right, so the distances between observation can be used as dissimilarity measure and the clusters or groups can be identified.

Cluster Analysis

- Two types of distances
  - Distance between two observations
  - Distance between two clusters
- Distance metrics as Dissimilarity measures
  - Common properties for distance metrics

    Observation i : $(x_{i1}, x_{i2}, \ldots, x_{ip})$

    Observation j : $(x_{j1}, x_{j2}, \ldots, x_{jp})$

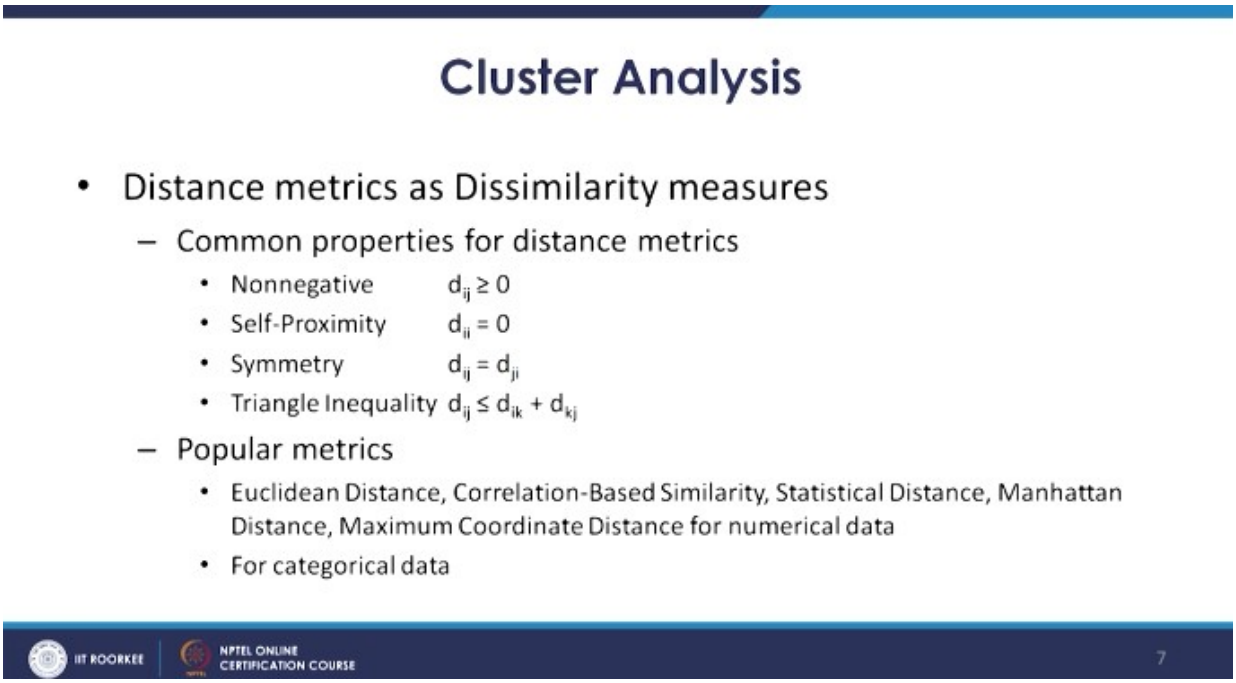
    Where p is the no. of variables to be measured

    d_{ij} is the distance between these two observations

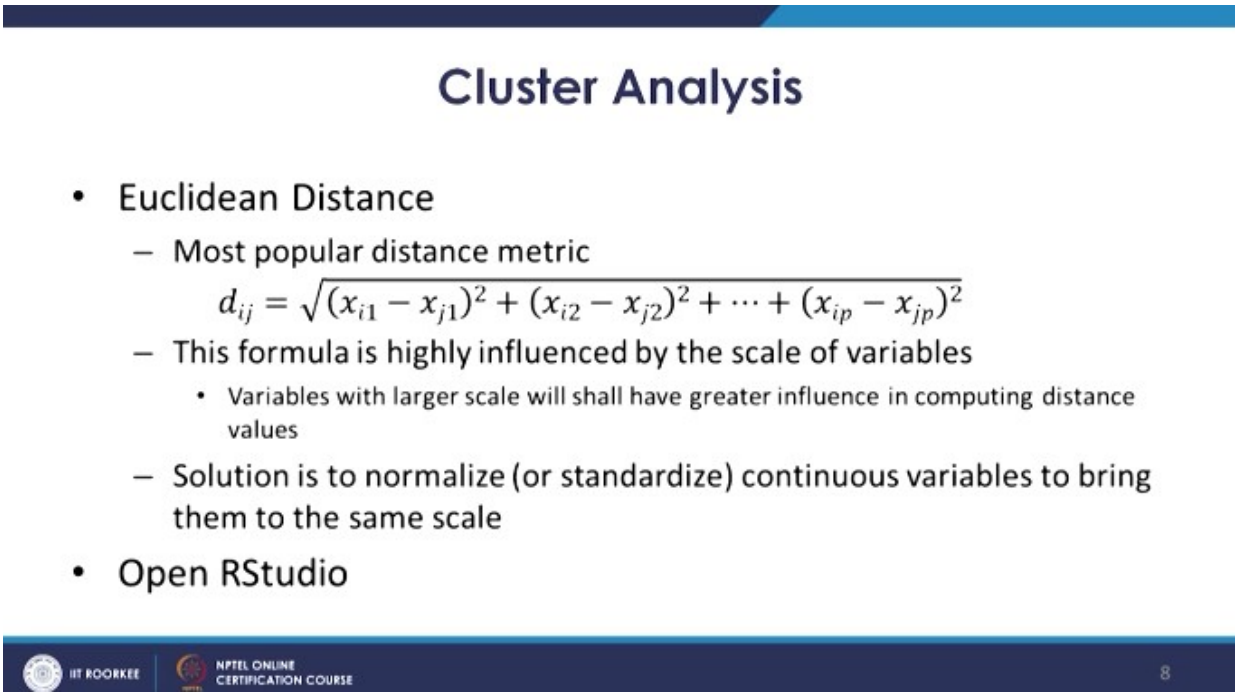
Now if distances are to be used then we need to see that there are two types of distances that we will have to compute here, so one would be distance between two observations, and the second one would be distance between two clusters, right, so once you know to identify clusters we need to compute the distances between observations then we would be able to identify the clusters or groups, and once that is done you know for any new observation that we want to you know assign to a particular clusters we need to know the distance between clusters you know for that you know that point you know and its distance from different clusters and all that, so these two times of distances that we can see, for see that have to be computed.

Now here as I talked about that distance matrix or being used as dissimilarity measures, so there are some common properties that are required for a metric to be defined as distance matrix, so these properties have to be follow, so here some notation has been mentioned which is to be used along in the discussion of cluster analysis and particularly the distance of distance metrics, so we are here assuming this observation I, XI1, XI2, up to XIP, so these are the coordinates for this particular observation, where P is the number of variables to be measured, so we have measurements on P variables and those P measurements are creating a coordinate for us for observation I, similarly for observation J we have XJ1, XJ2, up to XJP so these are the coordinates that we have for this observation.

Now the distance DIJ, the distance between observation I and J, so DIJ is the, you know notation that we are using for that distance.

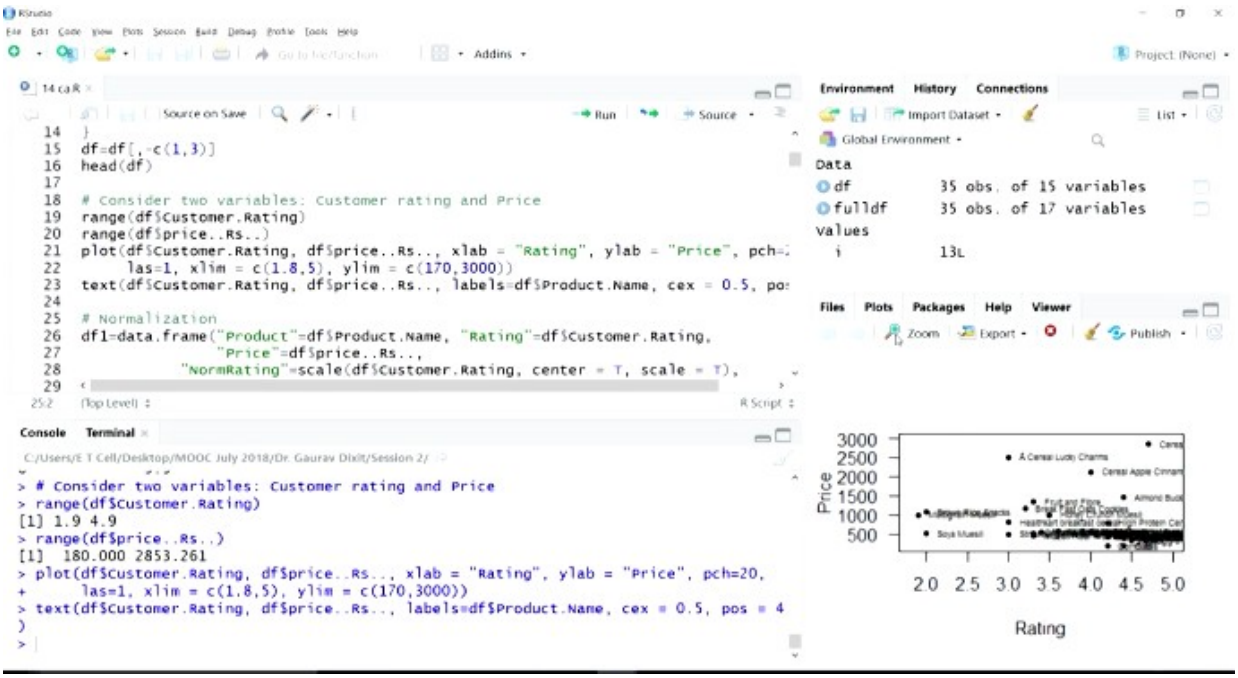


Now what are these common properties that we talked about? So these are the four property that has to be satisfied for a particular metric to be considered as a distance metric. So first thing is non-negative, so distance has to be a non-negative value so it would be either 0 or greater than 0, it has to be non-negative, then self-proximity, distance you know of a particular observation from itself has to be 0, right, so DII so distance of the observation from itself has to be 0, so that self-proximity rule has to be satisfied for any distance metric formula.
Symmetry, so distance between two observations I and J has to be same at the distance between observation J and I, so that symmetry has to be satisfied by the any distance in metric formula, triangle inequality, so distance between any two points that is I and J should be less than or equal to these summation of distance between two points I and K, and K and J so this triangle inequality is also to be satisfied, and it also to be satisfied for a distance, for a metric to be treated as you know, an appropriate distance metric.
Some of the popular matrix or Euclidean distance, correlation based similarity, statistical distance, Manhattan distance and maximum coordinate distance, so all these metrics that we just named here, they are typically for numerical data, so this metrics can be used to compute the distances between two observations for numerical data.

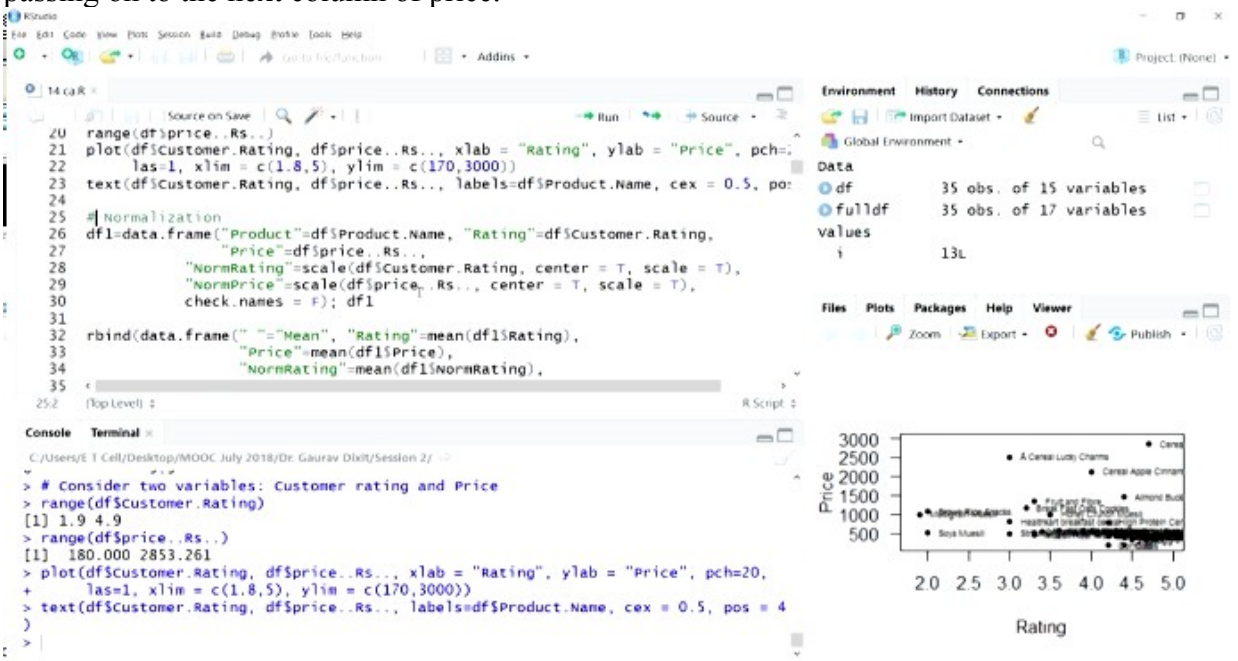


Similarly for categorical and data also there are few metrics that we would be covering in coming lectures, so let's move forward. So Euclidean distance is the, is the first metric that we are going to discuss, reason being it is the most popular distance metric, the formula for Euclidean distance we are already familiar with, so you can see here in the slide, DIJ distance between observation I and J can be defined using Euclidean distance formula as seen here, square root of XI1 – XJ1 square + XI2 – XJ2 square + up to XIP – XJP square, so this is the value as per the Euclidean formula, so far any two points for example, the observation, 35 observation that we had in the breakfast cereal dataset, so far between any two observation first and second observation we had the values for all the variables, so that can be used to compute the distance between those observations.

If you look at this formula, this formula is highly influenced by the scale of variables, so if the values, the actual values of the XI1 and XJ1 you know for any particular, you know for any particular variable if you know the values are on the higher side then that particular valuable might dominate the distance value, so this formula is a scale dependent, so variables with large scale will have a greater influence in computing distance values, so how do we overcome this problem? So the solution is to normalize or standardize continuous variables to bring them to same scale. So how this is done? Go back to R studio, and through an exercise will do this, so

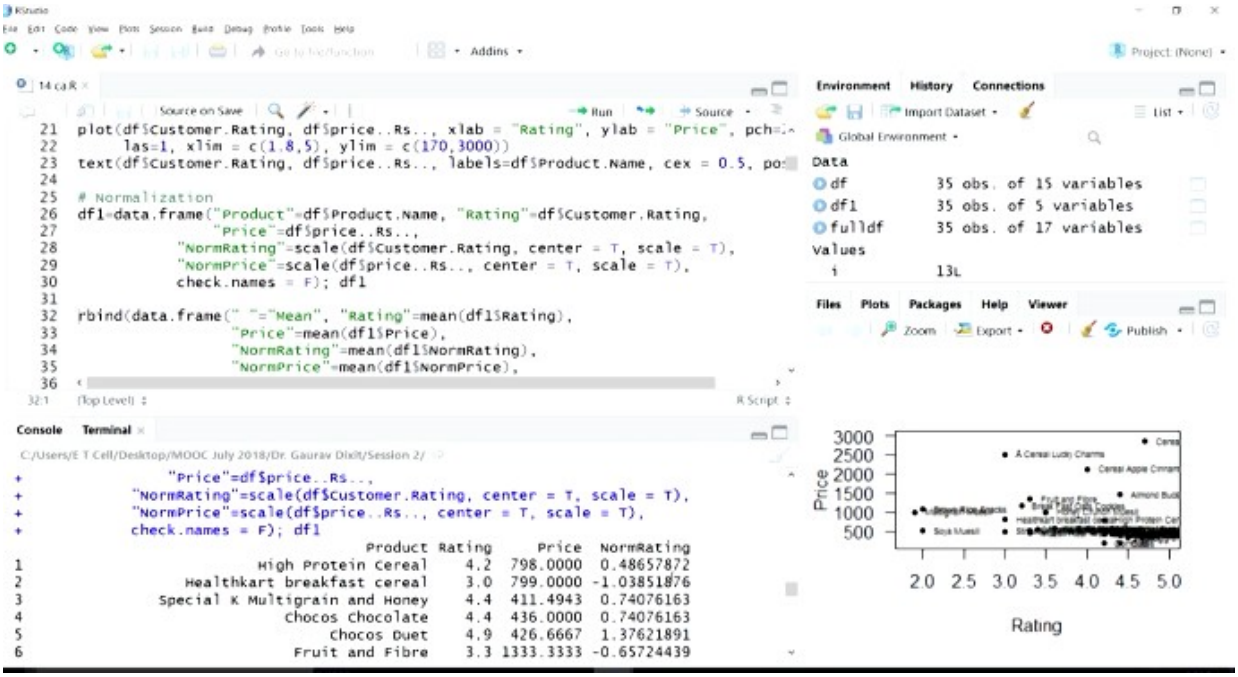


two variables that we have considered till now is customer rating and price in our dataset of breakfast cereals, so what we'll do is we'll do a small normalization exercise using these variables, so as you can see some code is written here, so we are calling this data.frame function which is going to create just a new data frame consisting of these columns, so first column is product, so nothing you know the previous data frame that we have already created DF, so we're passing on the product name of you know, from that data frame to this column, then the rating we are passing on the customer rating to this column, then we have price, then we are passing on to the next column of price.

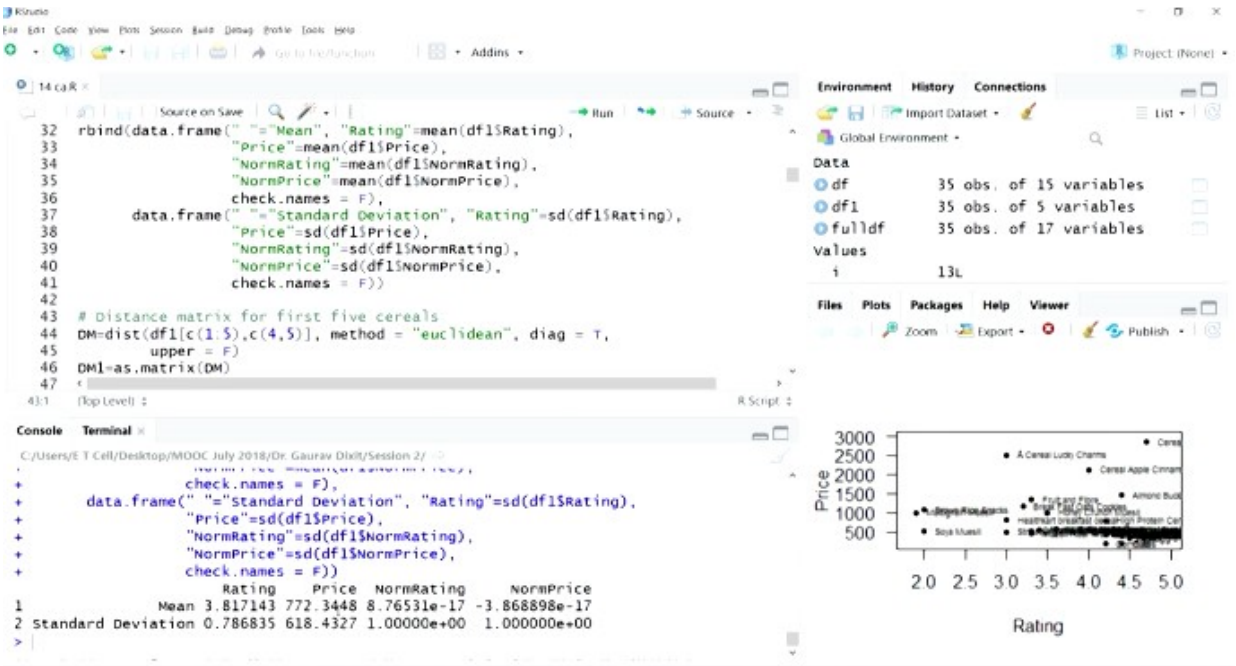


Then we have non-rating which is the normalized rating, so we would be doing normalization of rating, and the normalized price, norm price so we would be doing the normalization of

price, so the function that is being used here for this normalization process is scale, so you can see a scale function, more details on X-scale function you can always find from the help section, or you can also always refer back to the video lectures of previous course.
So first argument is the variable itself, DF$ customer rating, then we are center and a scale, so this is nothing but when we say center is true and a scale is true, so we are essentially going to compute Z scores, so typically when we talk about normalization typically we do, what we do is standardization which is nothing but computation of Z score, so what we do within this is that we subtract values is applied mean from the values and then divide by standard deviation, so the same process is going to be done using this particular formula, so rating and price, so both so we are considering rating here is also to be a numerical data, and so the rating of those cereals we are assuming that it has been quantified and numerically represented, so these variables so using this formula these variables are to be normalized, so let's run this code. Let's scroll and we will see that the first column is product, then we have rating and price, so these are actual

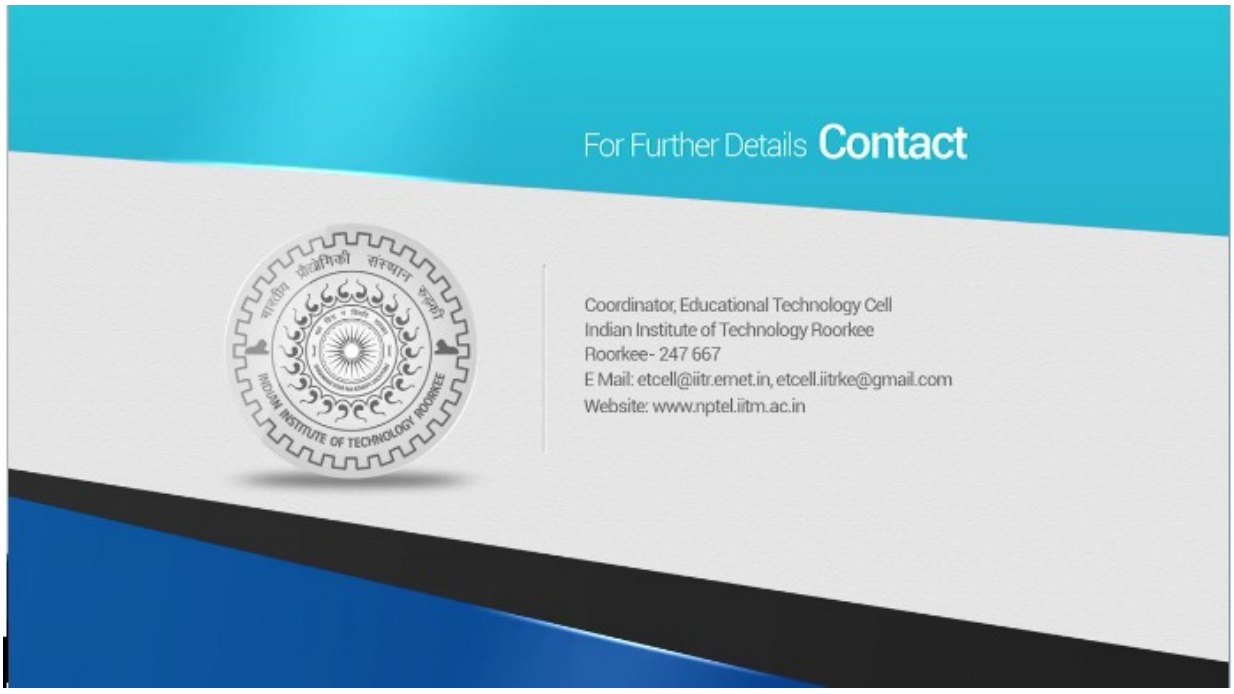


value, then we have norm rating and norm price, so you can see the values I have changed, so these values have been standardized now, so now these values now because of the scale that we have, these values can now be used, so no particular variable will be able to influence the distance computation, since standardize values would be in the same scale.
So to look more about the difference between the actual scale and standardize scale, some code has been written which is doing nothing special but we are just computing the mean values for the you know actual variables, and then the mean values for the normalized variables, and then the standard deviation for the actual variables and also same for the you know normalized variables, so let's run this code.



And you can see here, average rating was 3.8, and the average price was 772, if we look at the normalized rating so you can see both have come down to the same scales, right, then this values are almost close to 0, so in the normalized scale the values for both the variables are almost close to 0, so both have been brought down to the same scale, similar kind of values, but if we look at the actual values rating is about 4, and the price is about 7 you know it's about 800, so there was huge difference, so price would have dominated the distance values, right, but now using the normalize scales it won't happen.

Similarly standard deviation has also you know changed, you can see standard deviation is much lower for rating and much higher for price. And now if we look at the normalized scale so it has come down to the same value, right, so you can see using this kind of normalization and standardization we can get rid of this scale, you know, dependence problem of Euclidean distance metric. Okay, so till now so what we have done we've computed the normalized value for those two variables, and we could see how the, we can get rid of the scale dependence of Euclidean distance metric, so we'll stop at this point, and we'll continue our discussion on cluster analysis 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