Business Analytics & Data Mining Modeling Using R

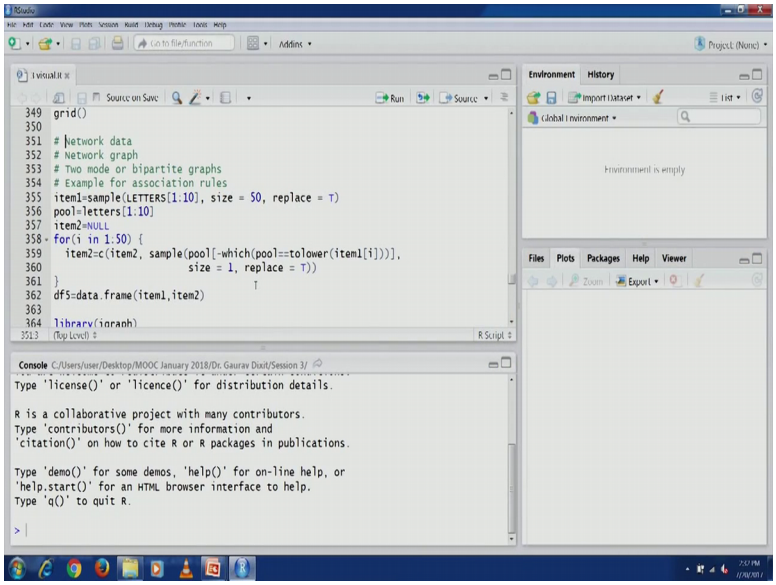
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 12
Specialised Visualization Techniques-Network Graph

Welcome to the course business analytics and data mining modelling using R. We have these 2 our last part of our visualization techniques. In the previous lecture we were we had we had started our discussion on specialised visualization we will continue from there. In the previous lecture, we started set size on network graphs, that is for network data. We will start from there and then we will cover 3 maps that is that is mainly for hierarchical data and then map charts that is for geographical data. Till now what we have we mainly dealing with before specialised visualization, was mainly cross-sectional data or time series data.

Now, these types of data set network data, hierarchical data and geographical data they are for different specialised visualization and they have different kind of data set as well. So we will discuss them in more detail as we go along.

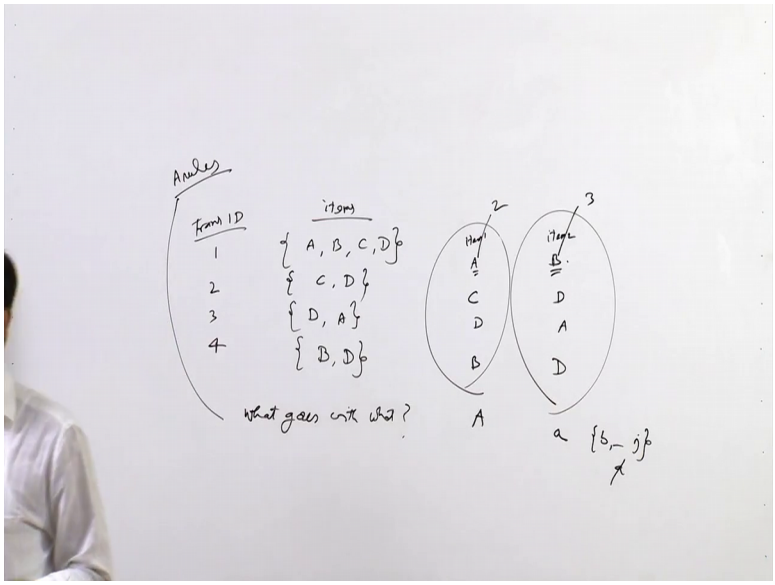
(Refer Slide Time: 01:26)



Let us open R studio, for network a graph we are going to use this network data which is mainly we also talked about a bit for this in previous a lecture that this is mainly in

association rules context. If we try to have we need to understand a bit more about association rules.

(Refer Slide Time: 01:53)



Now association rules mainly is generally done on transaction data basis.

Generally, transactions are in this format, transaction ID 1, 2, 3, 4 and then items and generally when a customer visits to a retail store; generally they purchase a few items. Let us say these are those items that were purchased by the first customer that is reflected in transaction ID 1. Another customer might purchase C, D; another customer might purchase D and A another customer might purchase B and D.

There are 4 items that are available you know that are mainly under discussion for these 4 transactions. They are being purchased by different customers and transactions are being recorded. Each in each transaction which items have been purchased, you can see it here. If we are interested in finding out you know generally, association rules mining is about, finding out what goes with what.

More detail more discussion on association rules we would be doing in a much later lecture. When will devote much more time on association rules right now, we can understand what association rules is about what goes with what. We try to identify which item is being bought along with which item. We are interested in finding out those items.

That we can plan our you know store layout and other things promotional offerings and acting offering we can bundle some of the items to boost our sales.

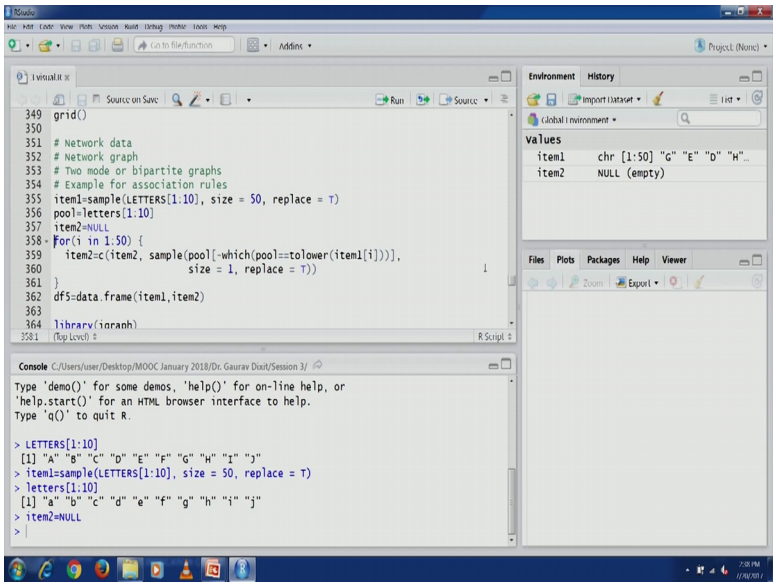
Those kind of things can be done, if we are able to identify. Now in this particular case when we are discussing network data or network graph we are interested in understanding transaction-based data you know database data set. In in a particular format were we are interested in finding which item was purchased first and then second another item would be purchased after that in the same transaction. Similarly, for this 1 CDD and a and B and D.

This particular information can also be depicted this particular information can also be depicted in through a network graph and we can understand from there we can make some sense about if we are dealing with large amount of data then in that larger amount of data visual analysis can be really helpful.

There we can now in this example that we are just doing we have just 4 observation 4 items 4 transactions 4 items, but if we are dealing with large amount of data visual analysis visualizing the data in a network data format in a network graph would be much more helpful for us to understand what is happening. Which items to identify to be able to identify some patterns and then that can be really helpful for us in our association rules mining; we will just do that.

What we are going to do is we are going to create bipartite graph that is 2 more graph. In this case as I talked about item 1 in a particular transaction and the item 2 you know the second item that is being bought that is being purchased in the same transaction. There are these 2 types that we are going to create. First, let us create the network data now this is hypothetical data we are going to create. Let us say we have letters 1 to 10. That is let us execute this particular code hitter letters A to letters from A to J.

(Refer Slide Time: 06:03)



These letters are representing different item. Let us execute this code let us create this pool item 1. These are you know items that are being purchased in different transactions.

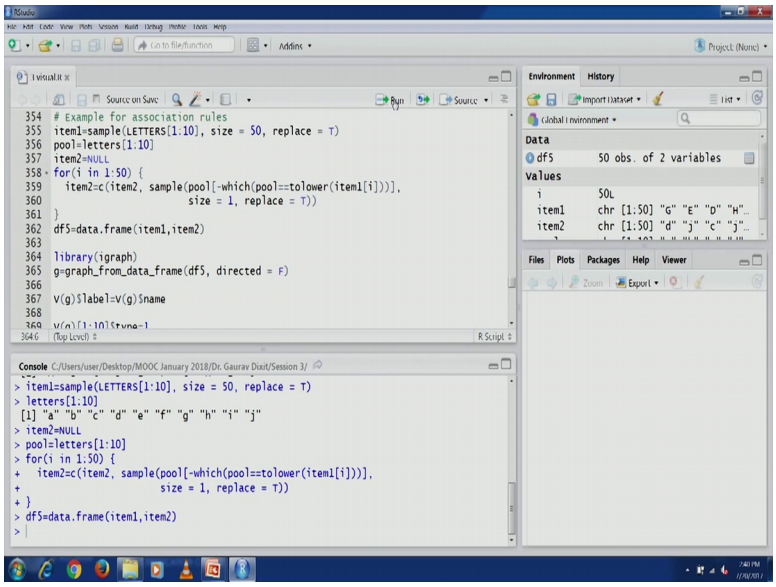
Now, second one is for example because the number of items are same. We are using the same items, but in lower cases this is mainly for the coating purposes. Otherwise, whether it is whether the item is represented in upper case or in lower case they both mean the same item, but just that we want differentiate between what was purchased first and what was purchased a along with that first item.

Because, here we are trying to create a data set were if in item has been purchased first it should not be you know because there could be multiple quantities a particular customer might be buying more than one item of this or more than one item of this as well. We are not interested in the quantity, we are interested in finding out if a is bought what is the second item that is also being bought or purchased along with this.

That is why you know we need to exclude this in our coding. Let us create this variable item 2 and you would see I am running this loop for from one to 50. 50 observations are going to be created, a just like item 1 right, but because the same item cannot be depicted now. It cannot be a and then second item can all cannot be because we are trying to get hypothetical data. We do not want this kind of situation we want it to be something other than. It could be from b to j and it should not be a. That is what we are trying to do in our coding here. Would see this item 2 lower item 1 I and it has been converted into the

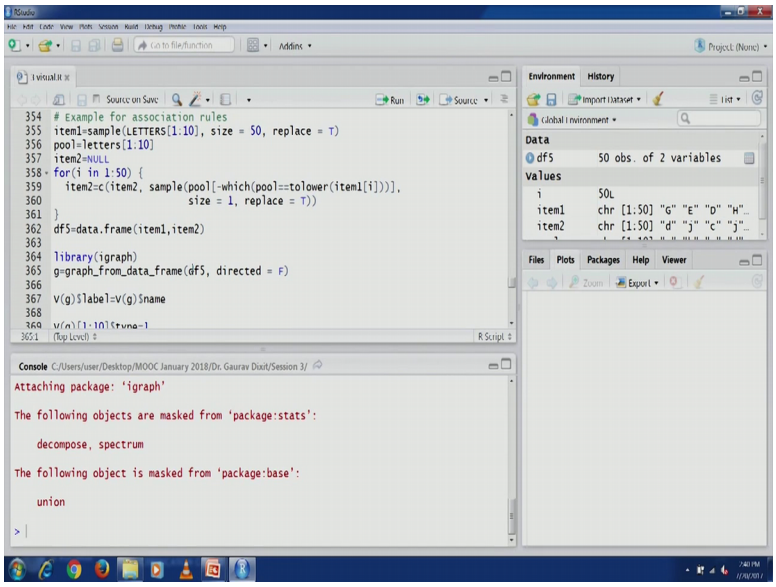
lower case and then we are comparing it with the pool items pool items are these. Let us create pool items. These are the pool items a to j smaller cases and then we need to eliminate this when we do sampling for item 2. That we get a proper item 2 vector.

(Refer Slide Time: 08:45)



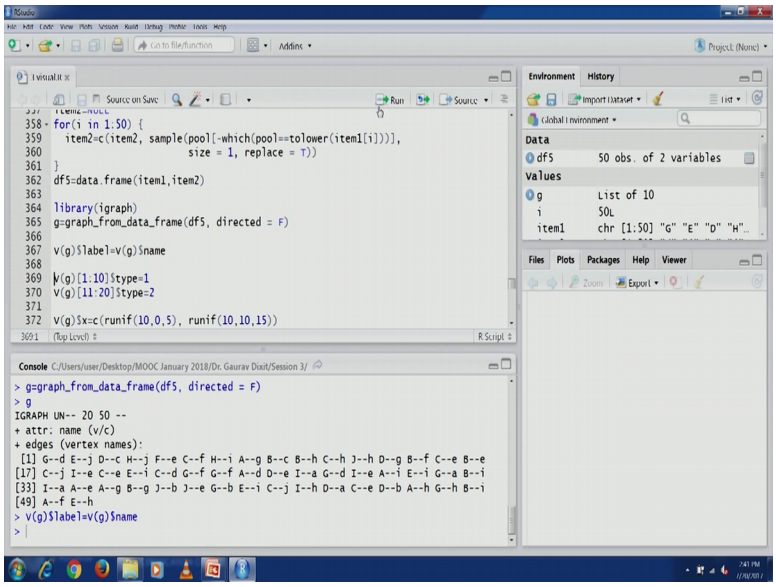
Now, let us crate the data frame of these 2 variables. You would see data frame has been created 50 observation of 2 variables now I graph is the library that is generally used for network data and network graphs. Let us load this library.

(Refer Slide Time: 09:05)



Now, from this data frame we are trying to create the a graph. This is the function that it is being used graph from data frame and because this is not a directed graph. That we want to create. Therefore, directed has been assigned as false. Let us execute this.

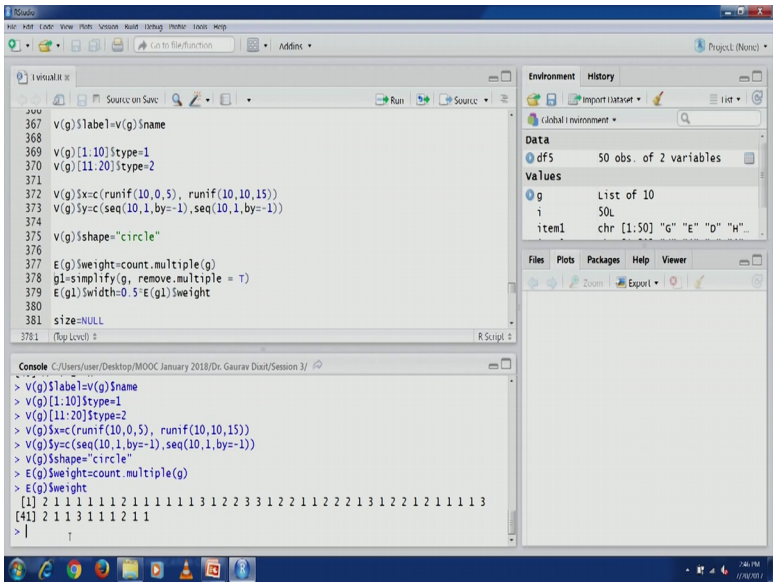
(Refer Slide Time: 09:27)



It you would see that a g has been created in the environment section and the this is undirected graph now a labels of these vertices if we at this a particular time if we look at the details of our graph you can see g and you would see that there are you know a 20 vertices and in 50 observations are there, then the 50 edges are there, you can actually see the edges. These are different items g and d was also bought along with this then a e and j then d and c and then h and j. These kind of this kind of list edges you would have now let us rename these vertices. They can be renamed after using their name itself. Let us execute this line. This is done.

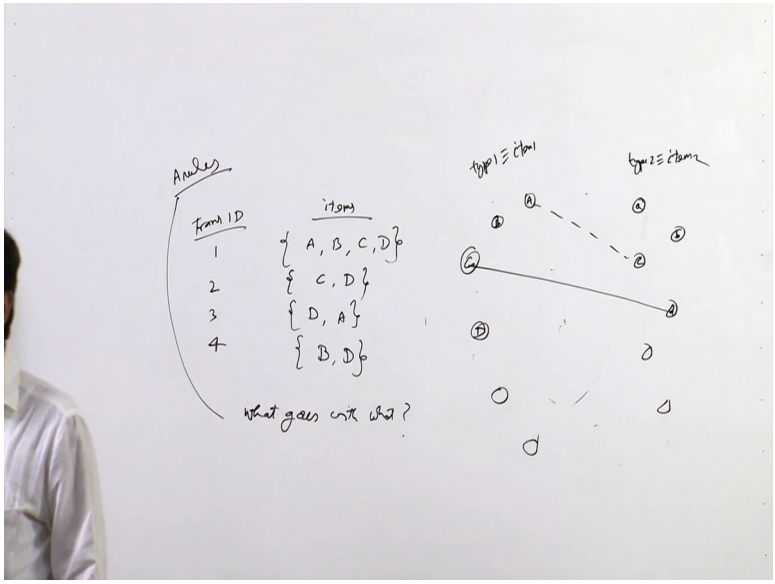
Now, as we said that we are trying to create 2 groups because we want to create bipartite graphs. Therefore, we need to have 2 groups. That is being done through this type a variable. V g and the type is type one is actually for item 1 and the type 2 is actually for item 2. We want to create 2 different groups. That we can we are able to create our network graph later on. Let us do this.

(Refer Slide Time: 10:53)



Now, we are trying to create a network layout. How the graphs are going to be depicted. We are we will be doing that. We want to be we want to create our network lay layout.

(Refer Slide Time: 11:17)



All these items are going to be represented by some vertices may be in circles and this is like type 1 and type 2 type 1 this is actually for item 1 and type 2 this is for item 2. These are like these elements. Represented by these elements ABCD in this fashion. Similarly, ABCD A could be with C, C could be with D depending on the transactions that we have. This kind of network graph we want to create. That we are able to visualize the data.

Now, using X and Y coordinates. We are trying to create we are trying to generate coordinates for all the vertices. Run if the if the command that we are going to use. You would see that we are trying to create a 10. Because, there are 10 items in in type one and similarly 10 items in type 2, you would see that run if 0 to 5 is range has been given and here 10 to 15 range has been given. The some space between these 2 groups because this is a bipartite graph some space is being left in the layout. Then we can easily visualize this vertices or nodes.

This is for x coordinate y coordinate is you can see that some spacing between different vertices or nodes is being run using sequence function. Once we have done this layout planning, now we can decide on the shape of nodes. In this particular case, we are going to select circle. Vertices they would be in this shape we can have squares and other shapes as well.

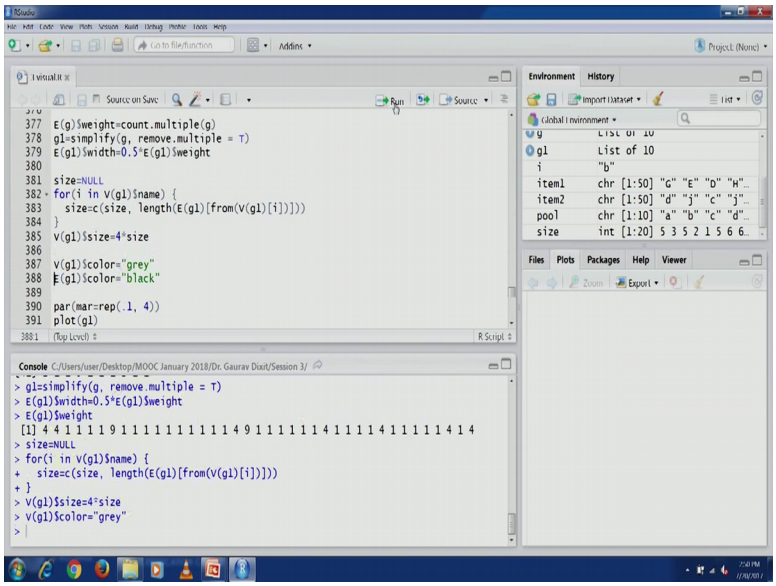
For this exercise we are selecting circle, then there are going to be there could be many edges between 2 particular nodes. If we are just trying to convert if we are just trying to depict transactions into a network graph. There could be a some customer who would be buying the same item 1 again B and D this kind of transaction would be there.

Therefore, between B and D there could be multiple edges, but in a network graph we do not want to show in this fashion. We want to show just one particular edge, but we would like to change the width of this edge. This would be a much a wider much wider edge representing more number of connections. This is how that is how we want to show that. We need to find out the multiple edges between 2 vertices.

This is what we are trying to do using count dot multiple function. In a graph, that has already been created with g in g. That we can find out on we are trying to. If we execute easy weight, let us see the values. You would see that for all these edges 350 edges that we have now. You would see some of the edges you can see that they are they are for example, this one is twice, some of the edges are thrice, in this fashion we can find out how many edges are how many edges are between the same vertices.

Now, we want to remove this multiple edges. This we can do using simplified function. Simplified function you would see that remove dot multiple argument is there. We are setting is it has 2. That is going to remove the multiple edges. Let us execute this 9 now in the next line you would see that.

(Refer Slide Time: 16:00)



Now we are trying to assign width that same way because we want edges to be of much wider width were the multiple edges were there. The same thing we are trying to you know you know assign it here. You would see Eg 1-dollar weight that has been assigned to Eg1 width and 0.5 is the kind of we are trying to control the scale of it. That the network graph it is visually more you know pretty or attractive. 0.5 is the reaction is scale that we have done from our side.

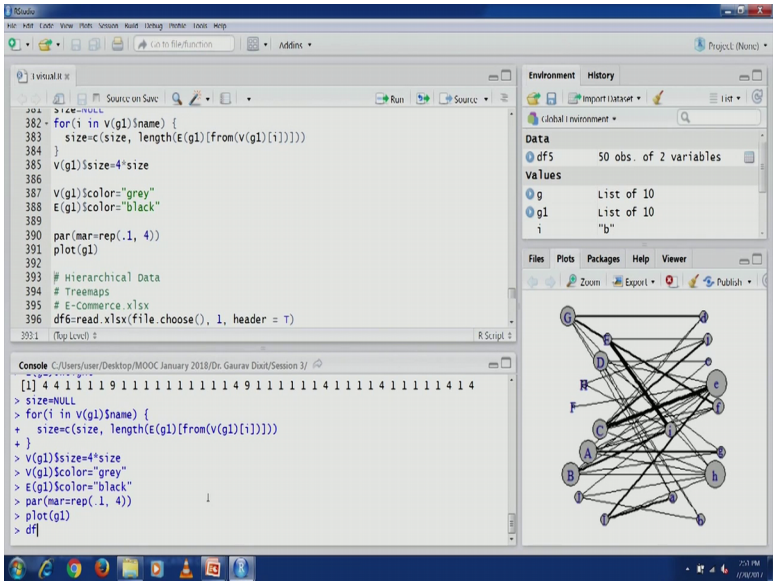
Depending on the way graph is being generated we can change this change this. Once Eg1 once e once simplify has been learnt if you are interested in finding out the you know this particular weight, you can see the same you can see there are many edge there are many edges for which the weights value have increased. 4 4 for some are 9 is also there. There many edges for which a weights value have increased that is because we have removed the you know a multiple edges and increase the width and through weight.

Now, the same exercise we can do for the vertices. Now the size of vertices that is also going to be based on the number of edges that is that are coming into a that are you know a coming into this vertices are going out of a particular node or vertice vertex. The same thing we are trying to do here. For every for every vertex we are trying to find out the length. Let us execute this code you would see that now this 4 is 4 here is again the boosting factor. Size. That again the vertices they are you know in a in a they have the particular shape which is reflecting the number of also the reflecting the number of edges

coming into it or going out of it and also the item that are being represented by a to j in capital letters or small letter they are also very well seen.

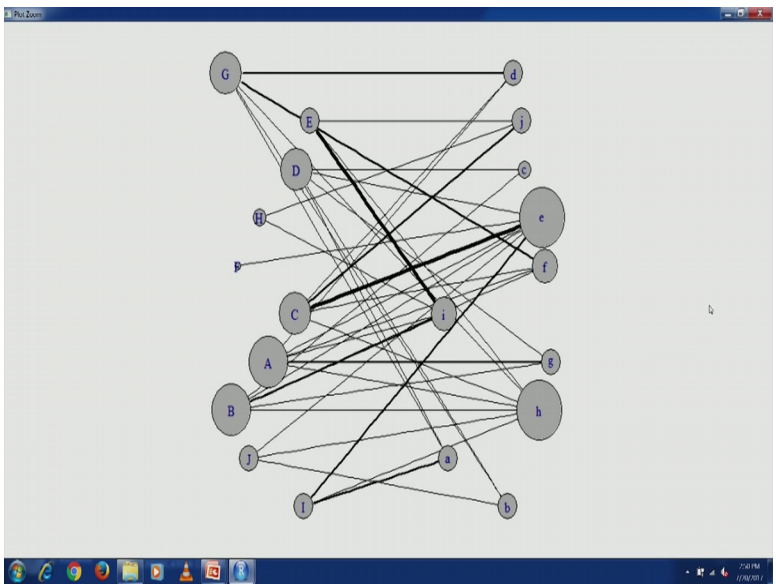
Now, let us come back to the colour 2 colours we are taking, per vertices are going to be generated going to be coloured with grey colour and edges are going to be coloured with black. Let us do this now let us set our parameter function margin 0.1.

(Refer Slide Time: 19:04)



On all you know sides and lest plot. Now, let us zoom out this particular now this is our graph.

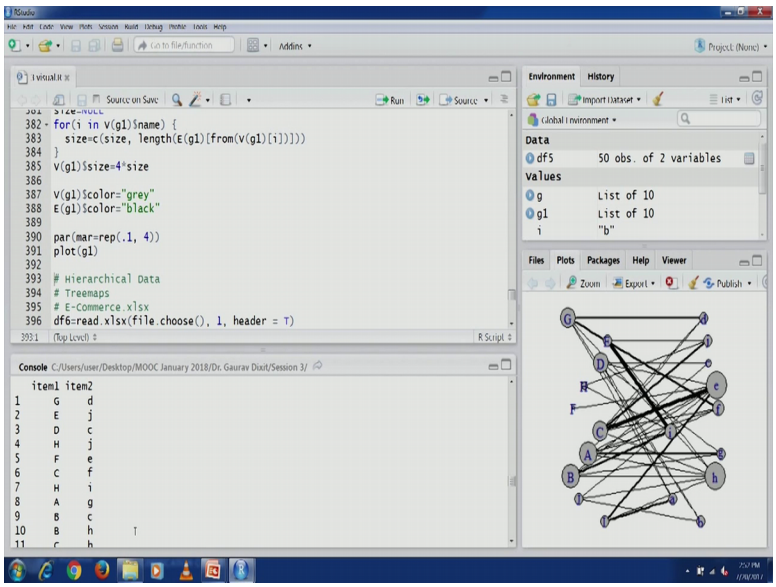
(Refer Slide Time: 19:16)



Now, from this graph you can see that items g d which are in in in which are having much greater size. Similarly, e and h there are many more edges coming into these vertices. You would see some of the edges their width is much higher. Because, there are more they are involved in more number of transactions. For example, e and I the width is quite high in comparison to other edges. Because, these 2 items have been brought more often. The same is reflected in the network graph.

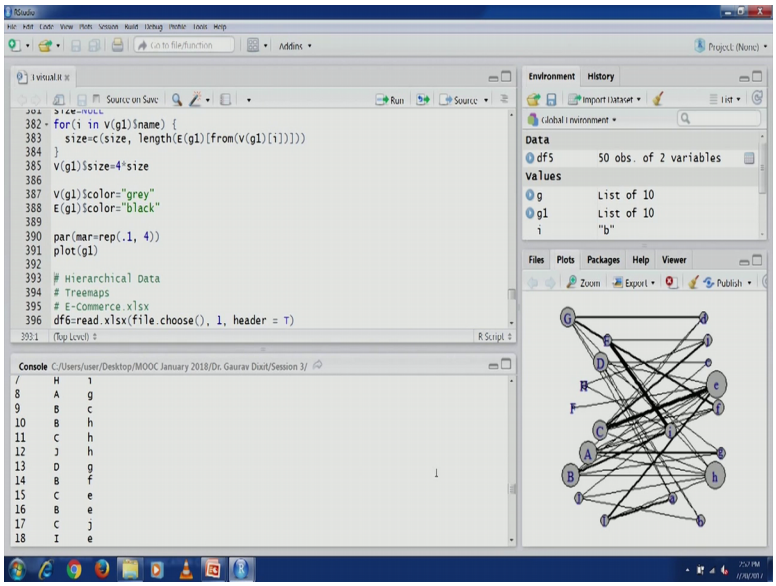
From here you can understand similarly c n e you would see that is smaller e this is also the size of this particular vertex is also on higher side. You would see e is anyway involved in more number of transactions in the hypothetical data set that we have generated and c and e they are being purchased together many times, that is reflected in the through the width of width between c and e. If you want you can have a look at the a data frame that we had created for this particular exercise or network graph, that is d f 5. The same would be same thing would be reflected in these transactions.

(Refer Slide Time: 20:47)



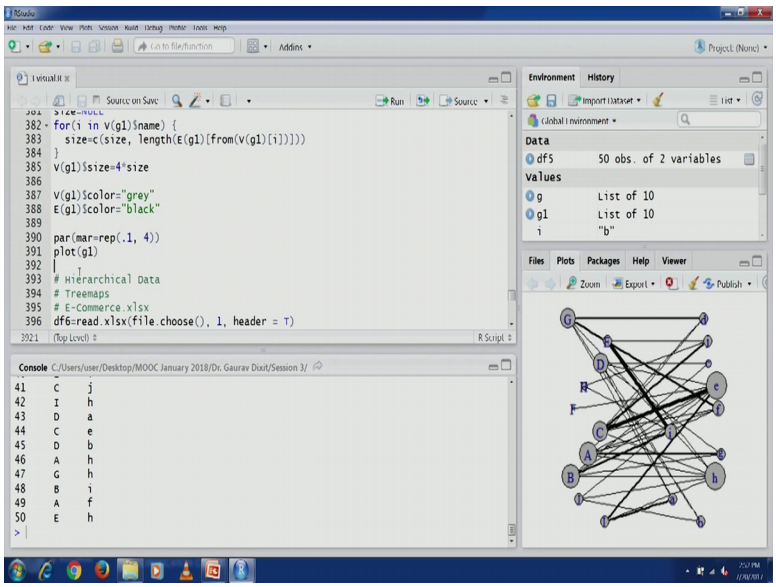
For example, in the graph as we can see c and e there seem to be more number of transactions and e there again being purchased many times, you might see that more often. You can see transaction 50 in this particular record c e is there. You can c in many transactions either in the item 1 or item category similar same case is there.

(Refer Slide Time: 21:11)



For e and c e can also be seen in many records. You can see this is again we can see c and e. The same thing we can verify here.

(Refer Slide Time: 21:44)

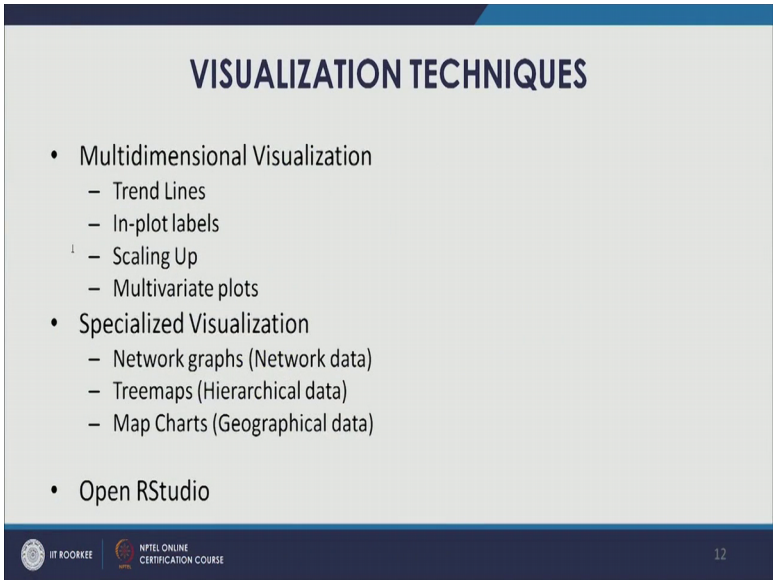


Now, let us move to the next specialised visualization, that is our about hierarchical data. Sometimes, we might have to deal with data that is more of a hierarchical more you know hierarchical in nature. They are going to be for example, in a university there are going to be departments and within departments there are going to be labs. Similarly, in

business organisation there are going to be organisation different verticals and then there are going to be different departments.

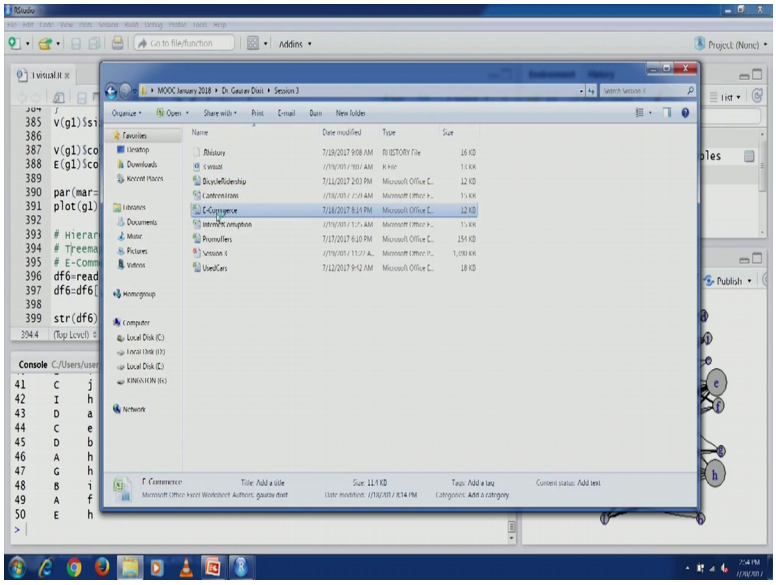
The way this particular information is going to presented it is more in a hierarchical format. How we can visualize hierarchical data? And how the kind of insights that we can generate through visual by applying visualization techniques? That we are going to cover now. Interesting plots that can be used to understand or visualize hierarchical data is tree maps. We are going to generate tree map through an exercise. We have this data set e commerce dot x lax.

(Refer Slide Time: 22:46)



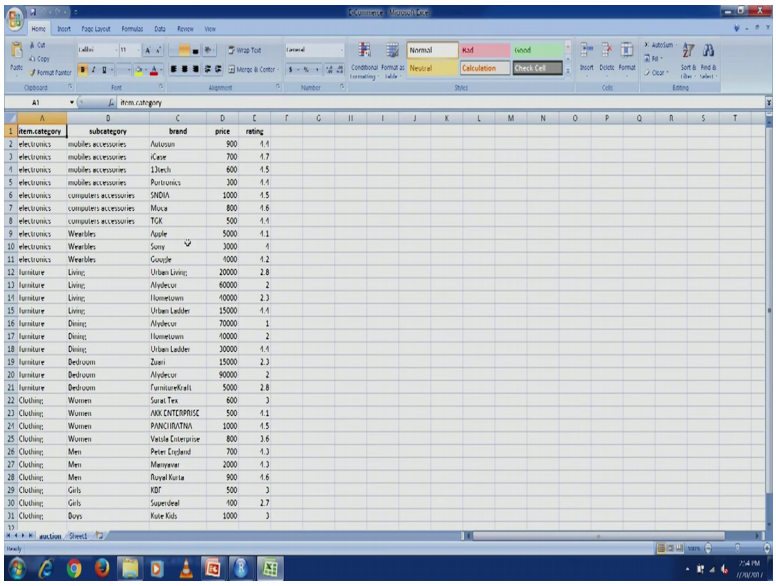
Let us look at the data first again this is also.

(Refer Slide Time: 22:50)



A hypothetical data set.

(Refer Slide Time: 22:52)



You would see there are there is there are 5 columns in this particular excel sheet item category then sub category, brand, price and rating.

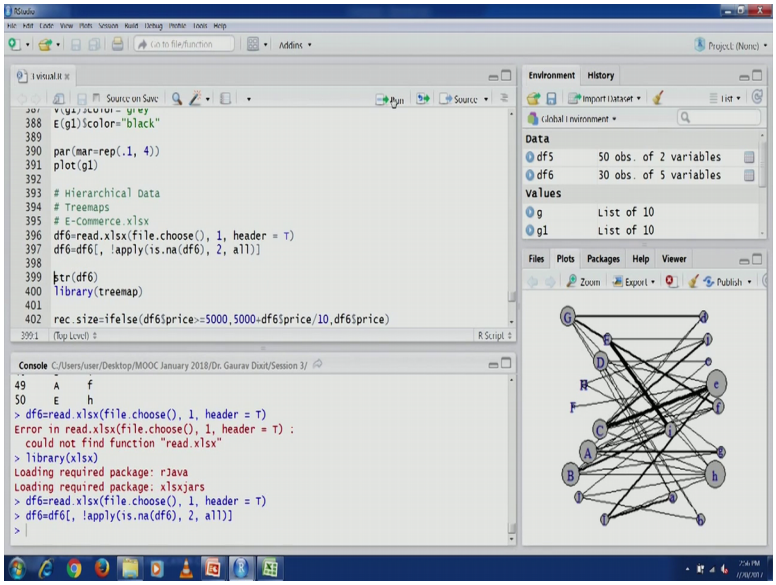
You would see that item category we have, electronics furniture and clothing and then sub category within electronics we have mobile accessories, computer accessories, variables in the furniture. We have sub category within you know living room furniture dining room furniture and bedroom furniture then clothing we have for clothing for

women men girls and boys. You can see for every category there is going to be a few sub categories.

A hierarchy easily a hierarchy can be seen in the data, then a brands are there. Of course, for every a sub category there are going to be multiple brands. And the another hierarchy and for each brands there are going to be you know different product. We are not covering the products here, but for each brands we have given some prices which are actually reflective of different products that are there. For each brand we have some prices and ratings for. These ratings or by customers for that is also available.

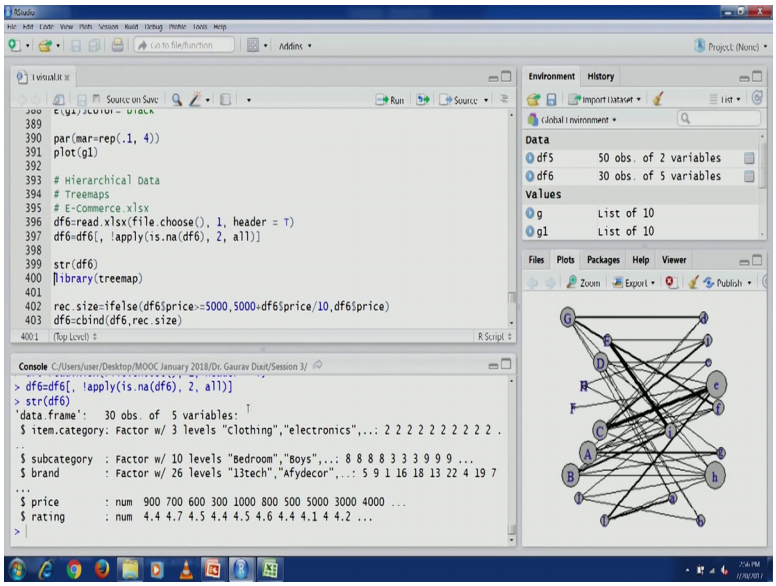
We are going to use this particular hierarchical data and we are going to plot tree maps and see what we can understand from those maps about specially about the data and how it can help in different task like prediction, classification and clustering. Let us load the import this particular data set.

(Refer Slide Time: 25:33)



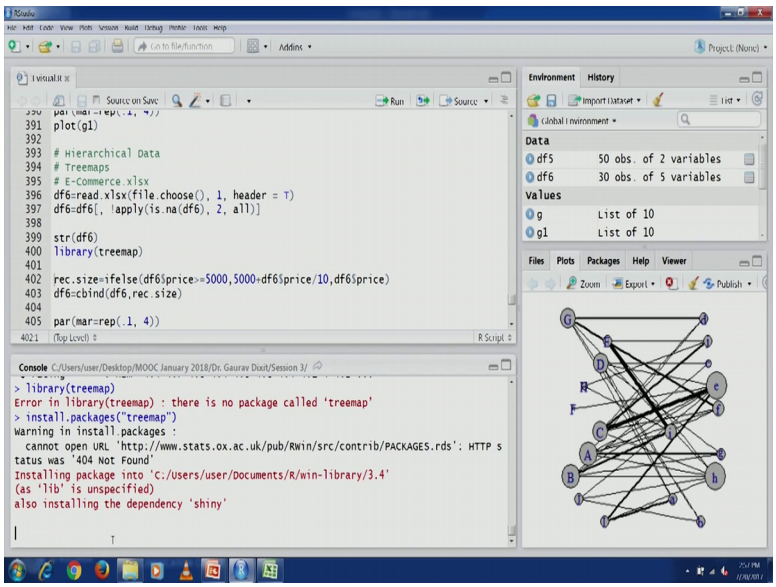
let us load this library first, mostly library has been loaded. Let us execute this line and import the data set you would see df6 has been imported 30 observation 5 variables the same excel file that we saw just now.

(Refer Slide Time: 25:03)

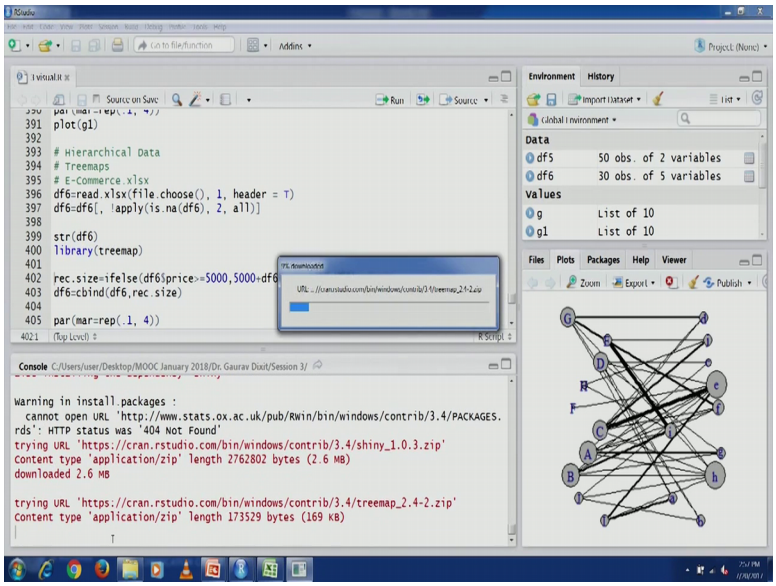


Let us execute this line as well let us look at the structure of the data set you can see item category factor fact sub categories also factor and the brand and then we have 2 numerical variable price and rating. The library that we require for tree maps is this tree map library. Let us reload this particular library.

(Refer Slide Time: 25:46)

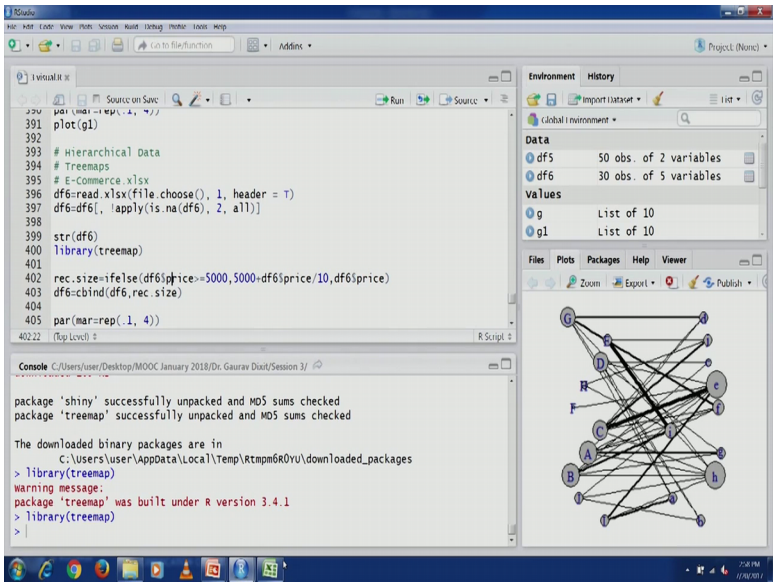


(Refer Slide Time: 26:00)



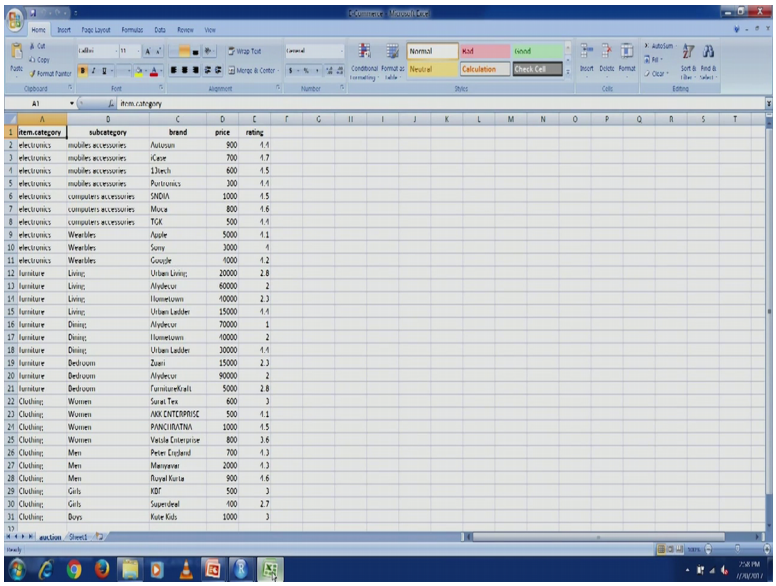
Let us reload this library tree map this one this has been reloaded.

(Refer Slide Time: 26:32)



Let us look at the let us look at the excel file again you would see that.

(Refer Slide Time: 26:55)



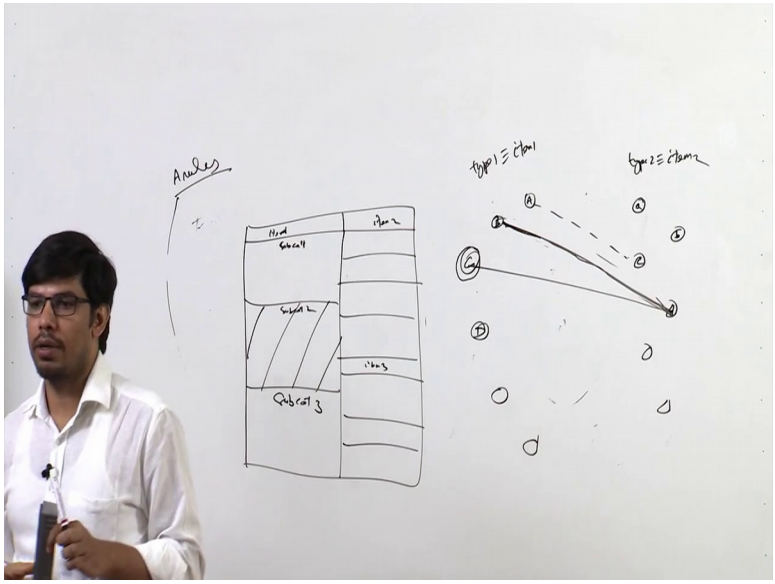
item.category	subcategory	brand	price	rating
electronics	mobiles accessories	Autosun	900	4.4
electronics	mobiles accessories	iKeye	700	4.7
electronics	mobiles accessories	1Tech	600	4.5
electronics	mobiles accessories	Portronics	300	4.1
electronics	computers accessories	SNDIA	1000	4.5
electronics	computers accessories	Moca	800	4.6
electronics	computers accessories	TGK	500	4.1
electronics	Wearables	Apple	5000	4.1
electronics	Wearables	Sony	3000	4
electronics	Wearables	Google	4000	4.2
furniture	Living;	Urban Living;	20000	2.8
furniture	Living;	Alydecor	60000	2
furniture	Living;	Hometown	40000	2.3
furniture	Living;	Urban Ladder	15000	4.1
furniture	Dining;	Alydecor	70000	1
furniture	Dining;	Hometown	40000	2
furniture	Dining;	Urban Ladder	30000	4.1
furniture	Bedroom	Zuari	15000	2.3
furniture	Bedroom	Alydecor	90000	2
furniture	Bedroom	FurnitureKraft	5000	2.8
Clothing;	Women	Surat Tex	600	3
Clothing;	Women	AKK ENTERPRISE	500	4.1
Clothing;	Women	PANCHRATNA	1000	4.5
Clothing;	Women	Vaisla Enterprise	800	3.6
Clothing;	Men	Peter England	700	4.3
Clothing;	Men	Manyavar	2000	4.3
Clothing;	Men	Nuvel Kurta	900	4.6
Clothing;	Girls	KDT	500	3
Clothing;	Girls	Superdeal	400	2.7
Clothing;	Boys	Kute Kids	1000	3

There are if you look at the price column. There are few items in this particular price column which are on which are up the very, very high value they for example, in the furniture the values are from 20000 to 90000 the other items other records that we see they are of lower value. See in here it is up to 5000 around 1000 or even less than 1000 other items.

There is quite a big gap between different price value and when we are going to create tree maps we would see that is going to a impact. The way that tree map is going to created. We want to reduce the scales of some of these values especially for create creating for the purpose of tree maps. In tree maps essentially, we are going to be creating different, different rectangles.

Let us see through an example.

(Refer Slide Time: 27:59)



T tree maps are generally in rectangular format. This kind of a rectangular plot is going to be created then you are going to have item categories item 1 item 2 and then you are going to have sub category sub category for these item sub category 1 sub category 2 3. Similarly, for item 2 there are going to be sub categories and then for item 3 and then sub categories.

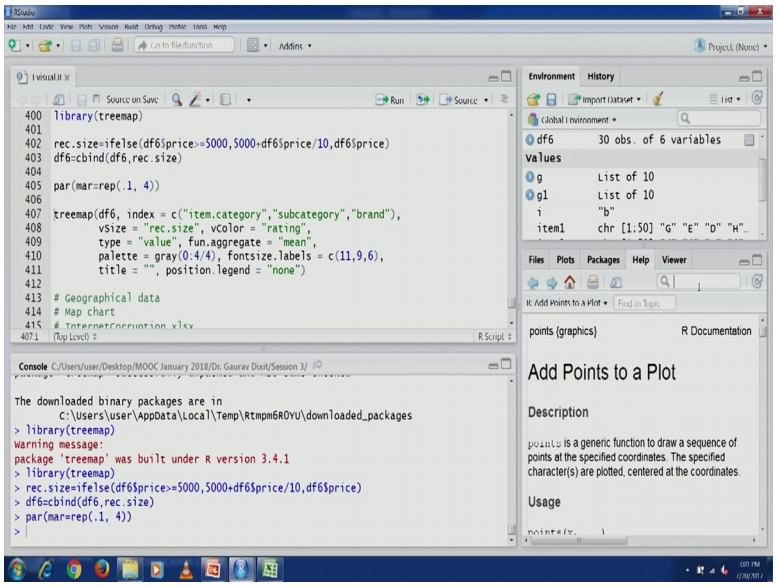
And then depending on the way you are trying to define the size of these rectangles rectangular zones, the way you want to colour them, the shade of the colour and the size of these rectangles that is going to convey some information that will see through an exercise but we want to we do not want some of the rectangular reasons to dominate and reduce these a size of other rectangle. That you know the visual that perception is lost.

We do not want to lose on our visual perception of some of the smaller rectangular. We need to control for that. If there is too much of you know too much of gap in the range for a particular variable which might be used for you know sizing of sizing of these rectangles, then that could be problematic. What we are trying to do in this particular line rectangular size for any price value. Price is going to be used to determine the size of these rectangular regions in tree maps.

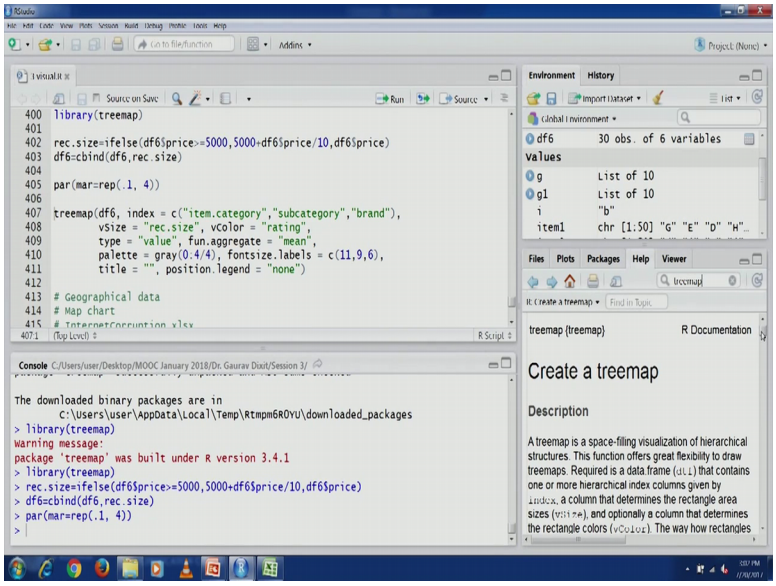
For any value which is greater than 5000. We are trying to reduce that particular value of the that value of that price. This is in this fashion 5000 and then we are dividing the value by 10. That is how the new value would be and that would determine the

rectangular size. If the value is less than 5000 then will keep the value as is. Let us create this particular variable. Let us also added into re data frame now let us again reset our par margin 0.1 on all 4 side. Now, tree map is the function that we are going to use. There are many there are many arguments that can be passed that can be used with this tree map function. If you are interested in more detail you can find out here in the help section.

(Refer Slide Time: 30:41)



(Refer Slide Time: 30:48)



You can see tree map and you would see there are so many arguments that can actually be used to design your tree maps.

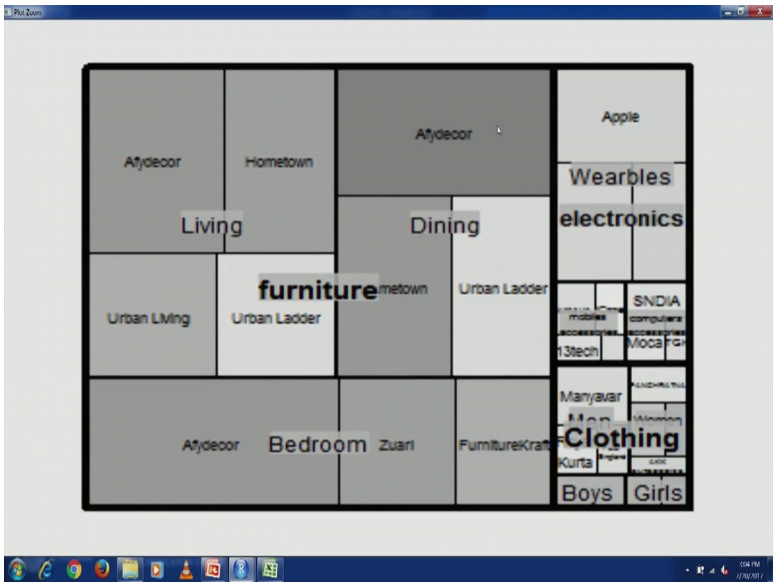
Some of them we are going to use in this particular exercise you would see that first argument is the data frame that we need to pass on to this function that is our data frame is df6 now the second is index. Indexing of a tree map is a dependent on the this hierarchy that we have. Our hierarchy is item brand item category then sub category and then the and then brand. That has to be passed on using a character vector. This is what we have done. Index takes character vectors.

We have created this then the argument that we are using in this exercise is v size that is size for the rectangular regions that are going to be created. We have already a created a variable rec dot size for the same. We have made sure that the sizes are in the appropriate sizes. That smaller rectangular they are not over you know overcrowded or reduced to a insignificant size because of bigger rectangles.

Now, the another argument that we have is v colour. For this colouring of different rectangular regions, we are using this rating column that we have. Ratings of these different, different products can be is going to be used to create different shades of a particular colour type, is another argument in this function value is going to be used here. Value of you would see value of price is going to be used to determine the this particular tree map. Then aggregate is using mean. Mean values they are going to be used for the aggregating aggregation.

Colour scheme is a grey. We are taking 4 labels of grey colours font size labels for different categories and sub categories depending on the hierarchy of the data this this we have given 11 for item names 9 for sub category names and then 6 for brand names this kind of font size labelling we are doing and the title we are not giving any title and legends. We do not want legend for our tree map. Let us execute this particular function. Let us see, this is the tree map that we have been able to generate.

(Refer Slide Time: 33:28)

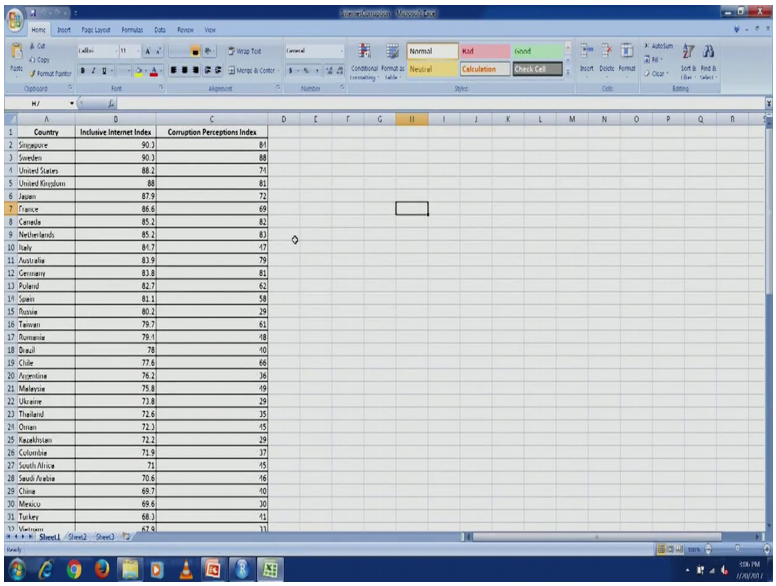


Let us look at. You can see that there is a one particular category furniture which is more dominant because of the higher value as we thought earlier. Living dining and bedroom are the sub category and then the brand names you can see rectangular regions, they are based on the average price the you would see that aggregation is based on the mean value or average value. That has been done. And so, that is there and the shading of this shading is based on the rating.

If a particular if a particular brand is a rated highly by the customers, the colour intensity is on the higher side. If it looks you know more of a grey higher intensity grey colour then therefore, customers have rated that particular thing highly more of you know light grey they have been poorly rated. The sellers of those items have been poorly rated. You can see this tree map now let us move to a next part that is geographical data.

To depict the a geo graph geographical data we generally use map chart. Again, we are going to use another data set this is a between this is this particular data set has information about internet inclusiveness and the corruption perception index and we are going to depict this information using geo using map chart.

(Refer Slide Time: 35:09)

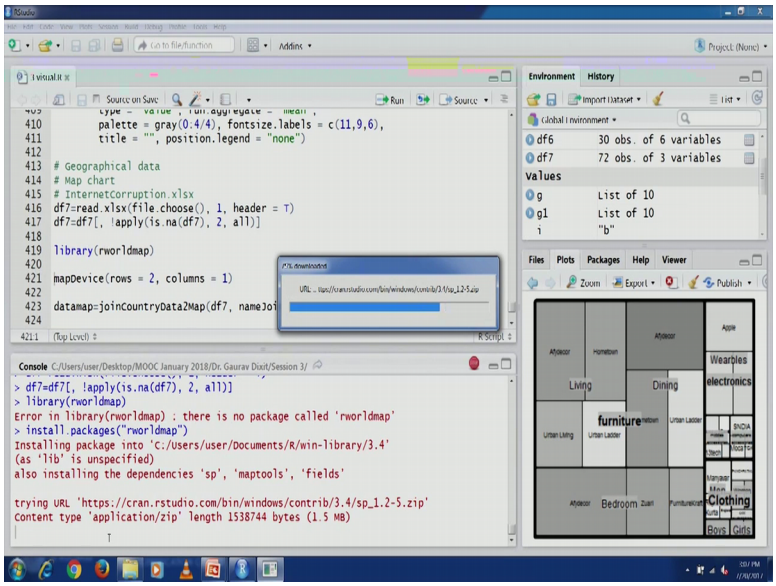


Country	Inclusive Internet Index	Corruption Perceptions Index
Singapore	90.3	84
Sweden	90.3	88
United States	88.2	71
United Kingdom	88	81
Japan	87.9	72
France	86.6	69
Canada	85.2	82
Netherlands	85.2	83
Italy	84.7	47
Australia	83.9	79
Germany	83.8	81
Poland	82.7	62
Spain	81.1	58
Russia	80.2	29
Taiwan	79.7	61
Romania	79.1	48
Brazil	78	40
Chile	77.6	66
Argentina	76.2	36
Malaysia	75.8	49
Ukraine	73.8	29
Thailand	72.6	35
Oman	72.3	45
Kazakhstan	72.2	29
Colombia	71.9	37
South Africa	71	45
Saudi Arabia	70.6	46
China	69.7	40
Mexico	69.6	30
Turkey	68.3	41

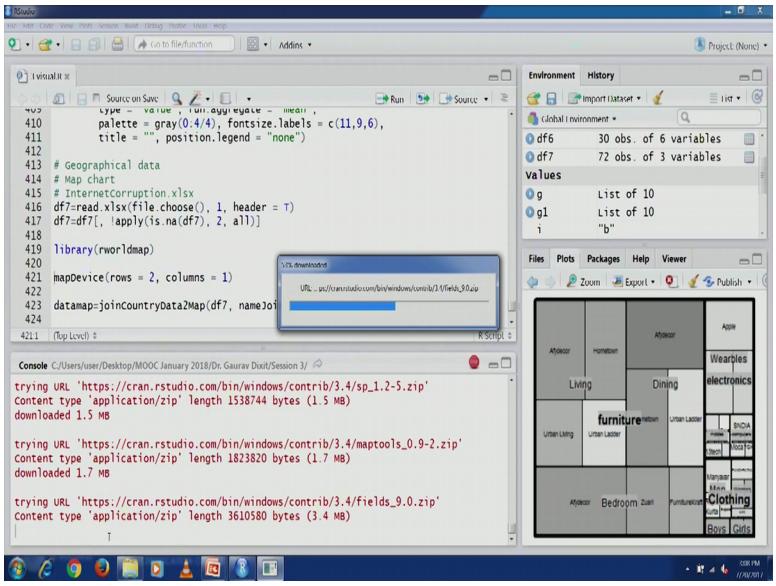
Let us look at the data set first, you would see we have 1 column first column is about the country. We have different country names here and then the their index about the kind of inclusive internet that they have. This index is reflecting that then the corruption perception index is also there.

We are we are going to create a map chart depending on these index values and will try to compare how the internet index is there and internet. How the internet index are there for different countries? And the a label of a corruption that are there whether there is any link between these 2 that we are going to do through a map charts. Let us import this data set. R1 map is the library that we would be requiring to create these map charts. Let us load this.

(Refer Slide Time: 36:38)

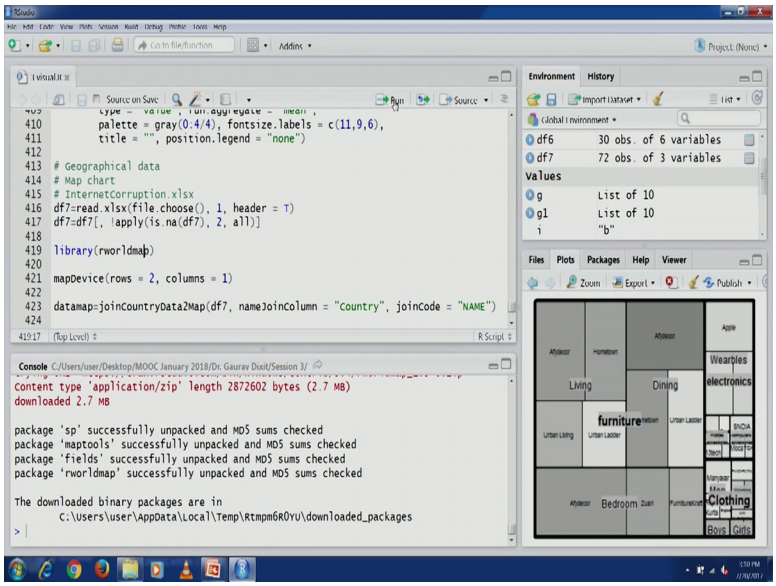


(Refer Slide Time: 37:00)



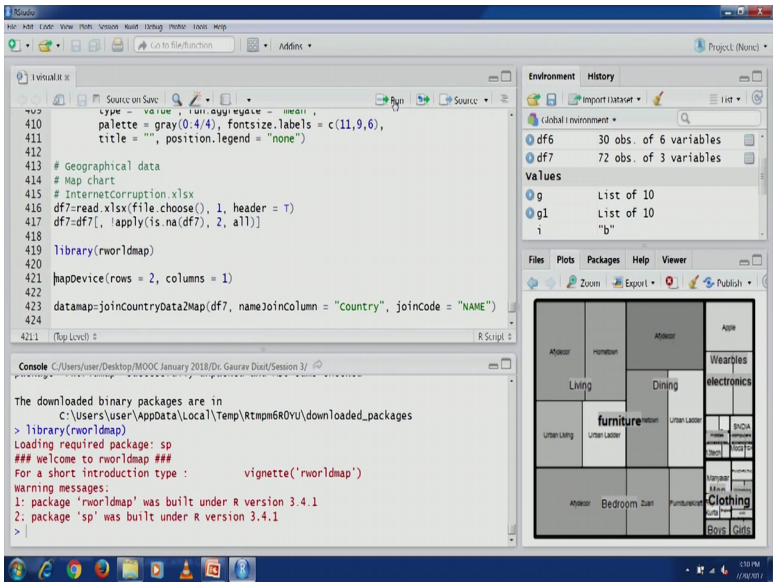
Let us reload this library.

(Refer Slide Time: 38:50)



R 1 map once it is loaded.

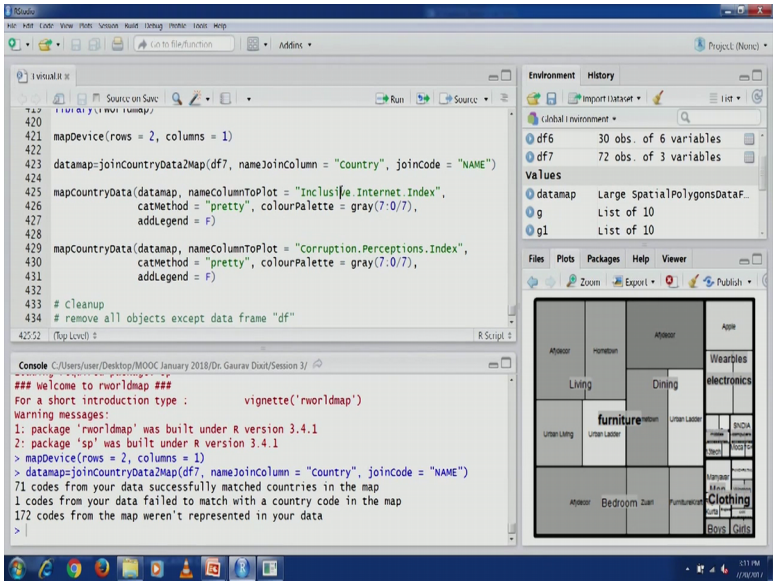
(Refer Slide Time: 38:54)



We need to create a device which is suitable for a generating map. We need we are going to create 2 maps. Therefore, 2 rows and 1 columns. Let us create this device. You can see a one device is active. Now, we are going to use this particular function join country data to map. The country data different in indexes that we had in the data set we are going to join them. Name join column that is country the column that we had in our data set join code is name that is why the using name of the country.

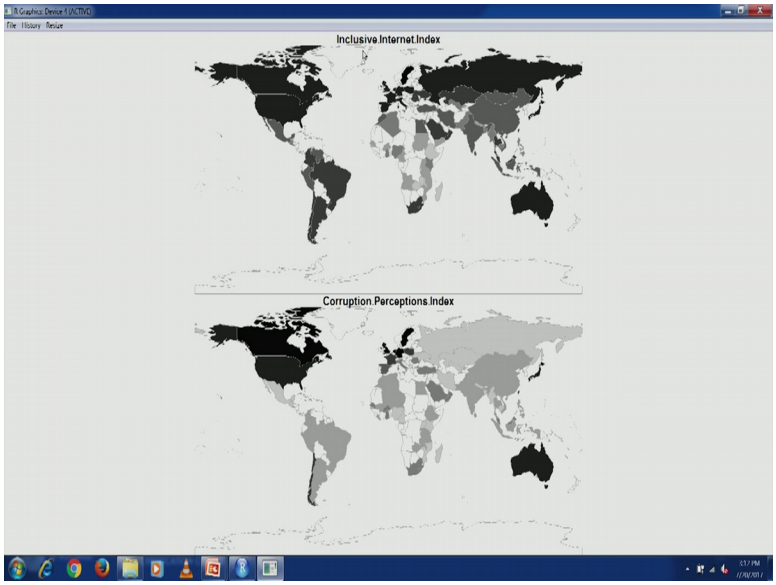
A data is available in in this particular data frame d f 7. Let us create this now you would see 71 codes have been. 1 code there was some a mismatch. That failed. Let us move on. There is another function map country data. This particular function is going to create the a map. Data map is going to be passed in then the appropriate the respective index that we want to plot in the map that we need to create.

(Refer Slide Time: 40:01)



The cat method is pretty the way the colours are going to be this selected colour pallet is given 7 to 0 then legend we do not want legend. Let us execute this code.

(Refer Slide Time: 40:26)



And you would see a map has been created in this particular device let us execute the another one this is representing the corruption index. You would see another map has been created now you can compare these 2 maps. The first one the is internet you would see the u s and Canada these 2 countries they are in the higher intensity a colour. They have higher inclusive internet and they also have a low levels of corruption. Higher intensity of is reflecting low intensity low level of corruption. You can see that. In this way you can actually visualize.

For example, inclusive internet inclusive internet index for India you would see you can see the shade of this colour grey and if you look at the corruption. This is in the lighter shade. Inclusive internet is there, but the corruption levels are not that much are not at that level. You can see Russia inclusive internet is much higher index, but if you look at the corruption there are much more corruption perception in Russia.

This kind of thing; we cannot, from this we can actually in general we can see that if the internet inclusiveness is on the higher side, we can also see that corruption perception is also a corruption is on the lower side in those countries. Some exceptions are there for example, Russia. We will stop here and in the next lecture will start our discussion on dimension reduction.

Thank you.