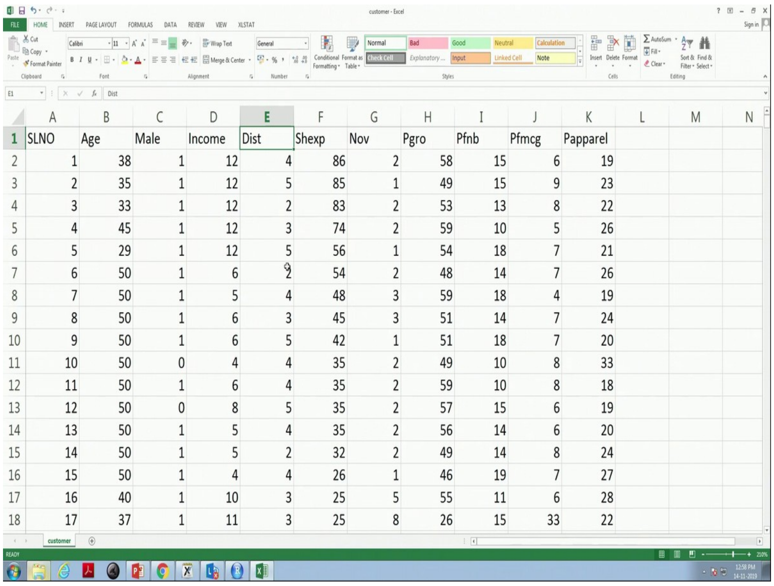
Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture 16
Segmentation Targeting and Positioning (Contd.)

Hello everybody, welcome to Marketing Analytics course, this is week 3, session 4 and I will be discussing about Segmentation Targeting and Positioning. This is Dr. Swagato Chatterjee from VGSOM, IIT Kharagpur who will be taking this course for you.

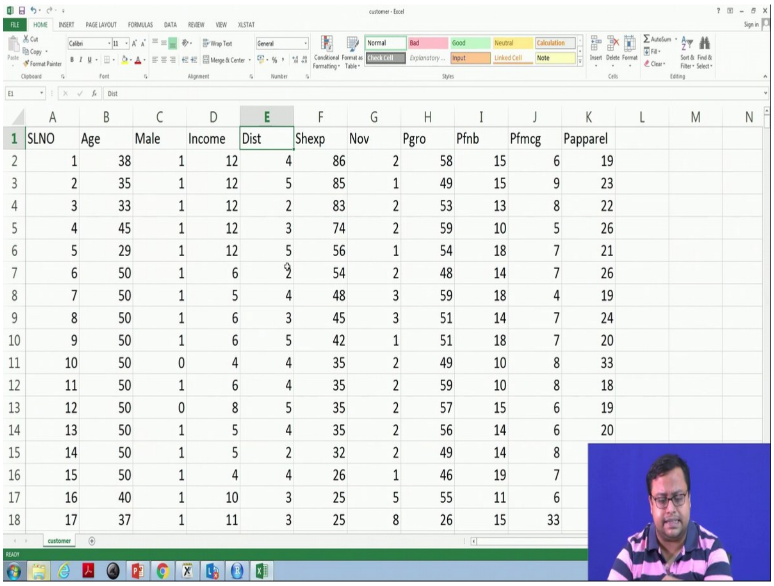
(Refer Slide Time: 0:35)

So, till the last presentation we have discussed about how to do segmentation targeting positioning using clustering methods. So, here we will actually work on a particular problem.

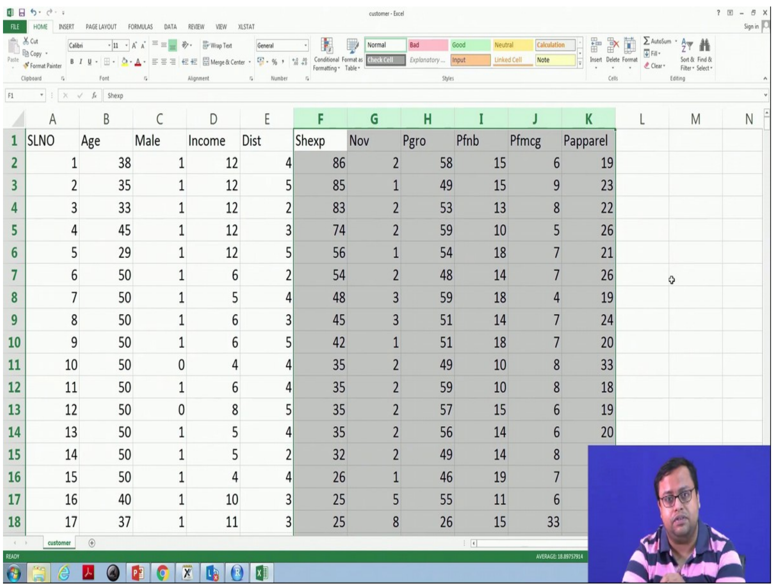
(Refer Slide Time: 0:42)



	A	B	C	D	E	F	G	H	I	J	K
1	SLNO	Age	Male	Income	Dist	Shexp	Nov	Pgro	Pfnb	Pfmcg	Papparel
2	1	38	1	12	4	86	2	58	15	6	19
3	2	35	1	12	5	85	1	49	15	9	23
4	3	33	1	12	2	83	2	53	13	8	22
5	4	45	1	12	3	74	2	59	10	5	26
6	5	29	1	12	5	56	1	54	18	7	21
7	6	50	1	6	2	54	2	48	14	7	26
8	7	50	1	5	4	48	3	59	18	4	19
9	8	50	1	6	3	45	3	51	14	7	24
10	9	50	1	6	5	42	1	51	18	7	20
11	10	50	0	4	4	35	2	49	10	8	33
12	11	50	1	6	4	35	2	59	10	8	18
13	12	50	0	8	5	35	2	57	15	6	19
14	13	50	1	5	4	35	2	56	14	6	20
15	14	50	1	5	2	32	2	49	14	8	24
16	15	50	1	4	4	26	1	46	19	7	27
17	16	40	1	10	3	25	5	55	11	6	28
18	17	37	1	11	3	25	8	26	15	33	22



	A	B	C	D	E	F	G	H	I	J	K
1	SLNO	Age	Male	Income	Dist	Shexp	Nov	Pgro	Pfnb	Pfmcg	Papparel
2	1	38	1	12	4	86	2	58	15	6	19
3	2	35	1	12	5	85	1	49	15	9	23
4	3	33	1	12	2	83	2	53	13	8	22
5	4	45	1	12	3	74	2	59	10	5	26
6	5	29	1	12	5	56	1	54	18	7	21
7	6	50	1	6	2	54	2	48	14	7	26
8	7	50	1	5	4	48	3	59	18	4	19
9	8	50	1	6	3	45	3	51	14	7	24
10	9	50	1	6	5	42	1	51	18	7	20
11	10	50	0	4	4	35	2	49	10	8	33
12	11	50	1	6	4	35	2	59	10	8	18
13	12	50	0	8	5	35	2	57	15	6	19
14	13	50	1	5	4	35	2	56	14	6	20
15	14	50	1	5	2	32	2	49	14	8	
16	15	50	1	4	4	26	1	46	19	7	
17	16	40	1	10	3	25	5	55	11	6	
18	17	37	1	11	3	25	8	26	15	33	



	A	B	C	D	E	F	G	H	I	J	K
1	SLNO	Age	Male	Income	Dist	Shexp	Nov	Pgro	Pfnb	Pfmcg	Papparel
2	1	38	1	12	4	86	2	58	15	6	19
3	2	35	1	12	5	85	1	49	15	9	23
4	3	33	1	12	2	83	2	53	13	8	22
5	4	45	1	12	3	74	2	59	10	5	26
6	5	29	1	12	5	56	1	54	18	7	21
7	6	50	1	6	2	54	2	48	14	7	26
8	7	50	1	5	4	48	3	59	18	4	19
9	8	50	1	6	3	45	3	51	14	7	24
10	9	50	1	6	5	42	1	51	18	7	20
11	10	50	0	4	4	35	2	49	10	8	33
12	11	50	1	6	4	35	2	59	10	8	18
13	12	50	0	8	5	35	2	57	15	6	19
14	13	50	1	5	4	35	2	56	14	6	20
15	14	50	1	5	2	32	2	49	14	8	
16	15	50	1	4	4	26	1	46	19	7	
17	16	40	1	10	3	25	5	55	11	6	
18	17	37	1	11	3	25	8	26	15	33	

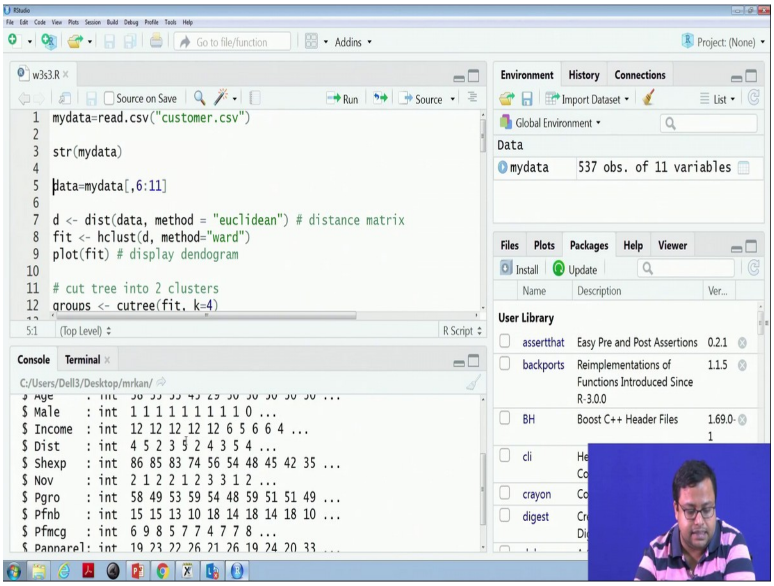
So, if you have seen the in your files there is a customer.csv data set and the data set looks like this. So, then the data set has serial number, age of a customer, male or female, male this, 1 means male and 0 means female it is a dummy variable. The income of the customers, the distance of the retail store, so it is a retail data let us say and the distance of the retail store from the address. So, how do I know the address of the customer? I actually come to know about the address of the customer when they fills it up.

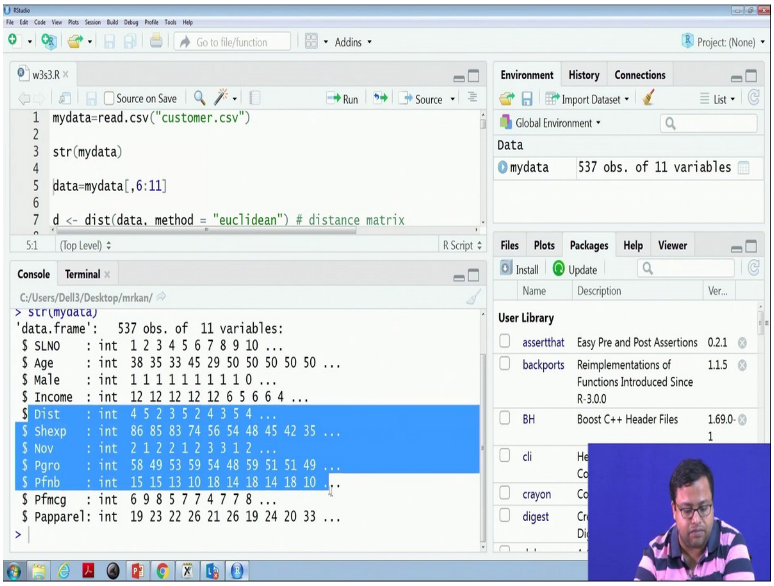
So, these these customer data has been tracked through the loyalty card they have. So, whenever you actually buy something to gain points, you swipe your loyalty card. And when you swipe your loyalty card, I come to know about your data. So, I, while you actually register for that loyalty card. What we had was your address and I know that the zip code of your address or not the zip code sometimes we know the Google location of the address as well, at least the lanes location. And from my retail store I can find out using Google Maps certain distance.

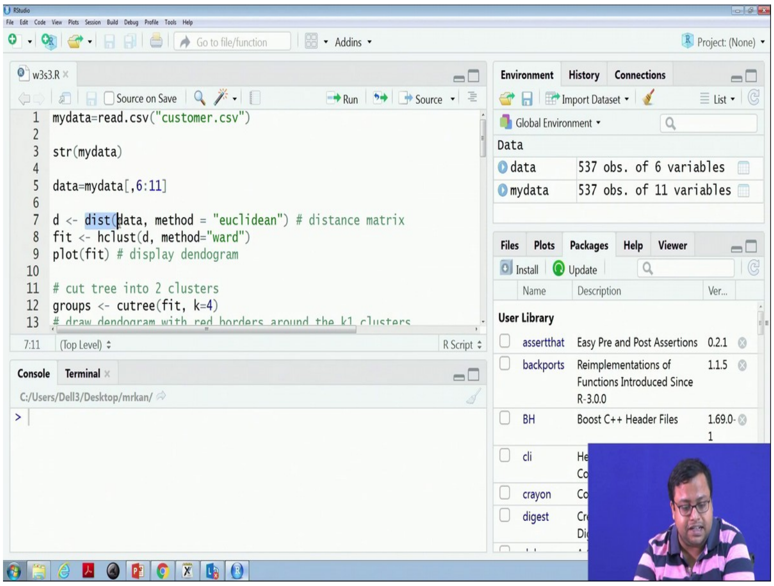
So, this is the part which will be done by a coder and this is not something that you assume probably have to do. You might have certain business analysts and etcetera. If you are doing for academic purposes, sometimes you have to do on your own, but some other expert can do it using Google Maps. So, that is number one, a distance. And then I have certain behavioral data. So, column B, C, D, E are all demographic data of the people, persons and I will not use that for my segmentation I will focus on the behavior of the people.

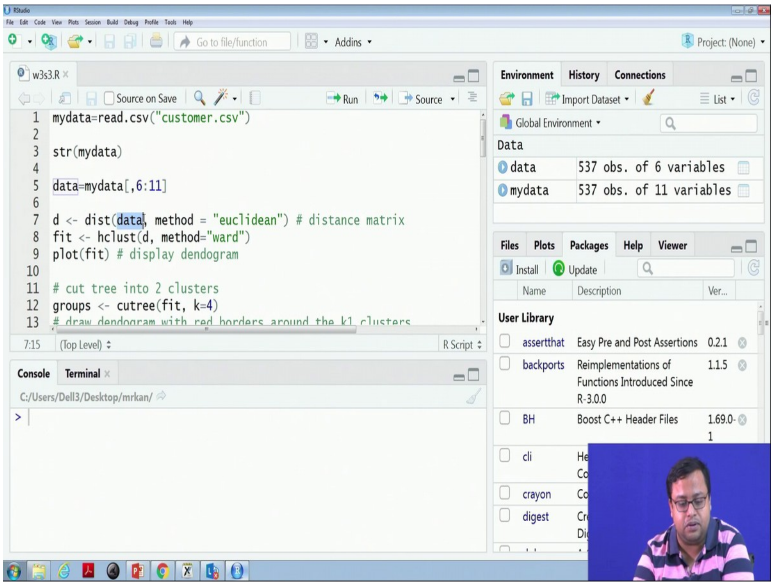
So, F column is shopping experience, G column is Nov that means number of visits, not shopping experience sorry shopping expenditure, F column was shopping expenditure in lakhs or in thousands per month. Nov was number of visits in units and then Pgro, Pgro is actually how many, so how what percentage of your purchase is in grocery? What percentage of your purchase is in these in food and beverage F and B? What percentages in FMCG? And what percentage is in apparel? And there can be another various kinds of observations that we can find out, behaviors you can find out, but I am focusing on these six. So based on these six, we will do our cluster analysis and then we will try to find out what kind of customers they are?

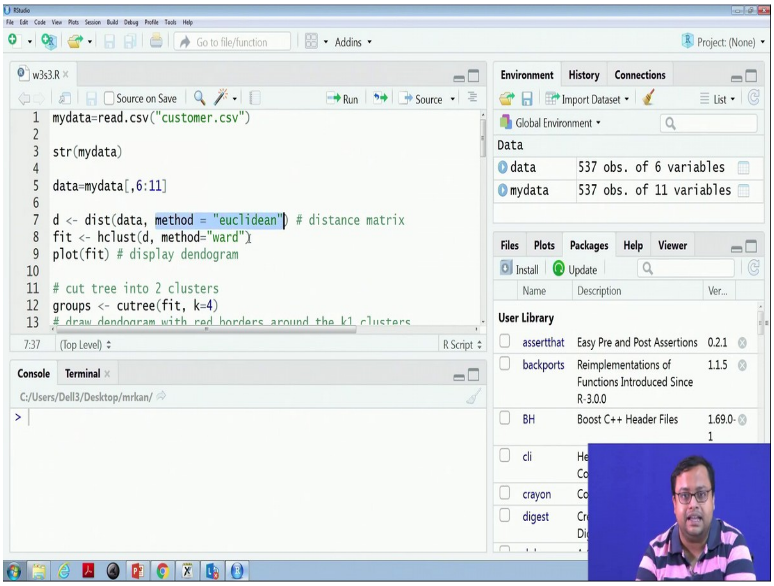
(Refer Slide Time: 3:28)

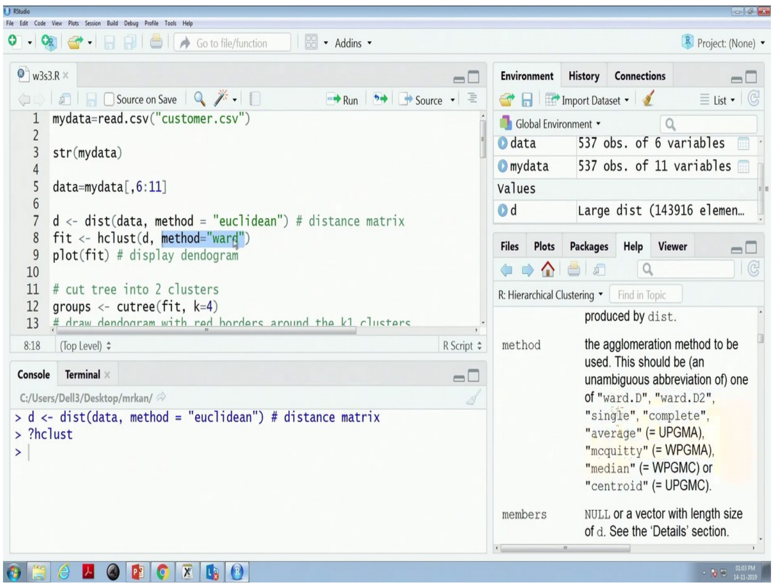


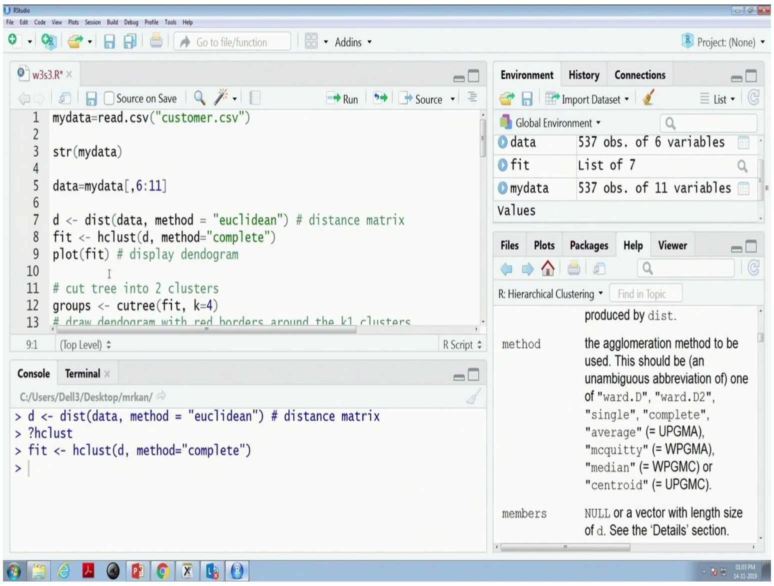


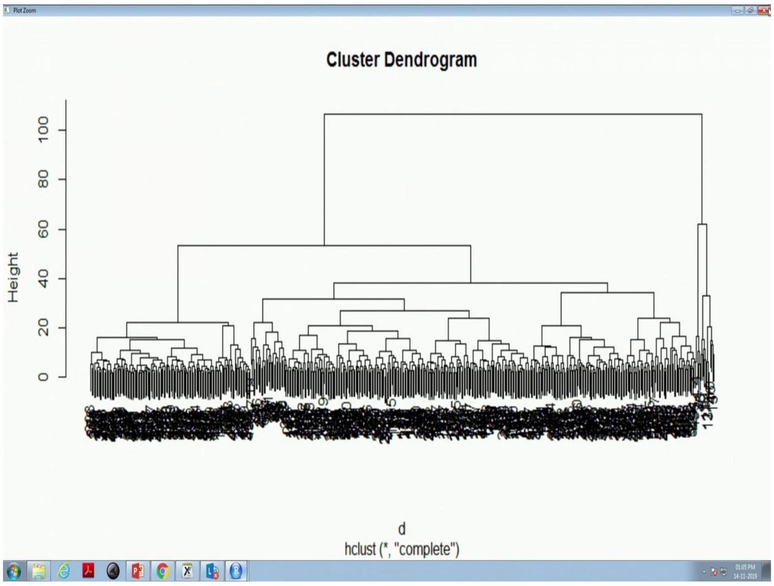



w3s3.R
mydata=read.csv("customer.csv")
str(mydata)
data=mydata[,6:11]
d <- dist(data, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram
# cut tree into 2 clusters
groups <- cutree(fit, k=4)
Environment History Connections
Global Environment
Data
data 537 obs. of 6 variables
mydata 537 obs. of 11 variables
Files Plots Packages Help Viewer
User Library
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.5
BH Boost C++ Header Files
cli
crayon
digest
Console Terminal
C:/Users/Dell3/Desktop/mrkan/


w3s3.R
mydata=read.csv("customer.csv")
str(mydata)
data=mydata[,6:11]
d <- dist(data, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram
# cut tree into 2 clusters
groups <- cutree(fit, k=4)
Environment History Connections
Global Environment
Data
data 537 obs. of 6 variables
mydata 537 obs. of 11 variables
Files Plots Packages Help Viewer
User Library
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.5
BH Boost C++ Header Files
cli
crayon
digest
Console Terminal
C:/Users/Dell3/Desktop/mrkan/


w3s3.R
mydata=read.csv("customer.csv")
str(mydata)
data=mydata[,6:11]
d <- dist(data, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram
# cut tree into 2 clusters
groups <- cutree(fit, k=4)
Environment History Connections
Global Environment
data 537 obs. of 6 variables
mydata 537 obs. of 11 variables
Values
d Large dist (143916 elemen...
Files Plots Packages Help Viewer
R: Hierarchical Clustering
produced by dist.
method the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of "ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) or "centroid" (= UPGMC).
members NULL or a vector with length size of d. See the 'Details' section.
Console Terminal
C:/Users/Dell3/Desktop/mrkan/
> d <- dist(data, method = "euclidean") # distance matrix
> ?hclust





So, the first thing that I will do is I will set marking directory to source file location. So, a w3s3.r actually s4 it should be s4.r that is the file that we are working on. And your global environment should be empty, your console should be empty. In that position, we are calling this particular data. So, I am calling this data, this data has 537 observations of 11 variables. And the structure of the data looks like this that all of the below ones are integer variables so they are numeric. So, I do not have any issue I do not care, I do not have to change anything.

So, then I get a subset of the data because I will work on the last six columns, I create the subset of the data and I create a data which is 537 observations of 6 variables, which are basically the behavioral variables and then the first thing that I do is I create a hierarchical cluster. So to do that, I create a distance matrix. So the function is dist, dist. And what is the input? Input is data and what is the method? Method this Euclidean, so it will create our distance metrics for every person to every person 537 to 537 it will create a distance matrix and that is what has been created here.

So d is equal to dist and then if I run this, I get a d which is a large distance metrics. Using this matrix which you can print here but 537 into 537 is very huge, is no point on printing. If you want to save it, you can save it by right dot csv and change this d to data frame and then save it. But I do not want to do that, I create a hierarchical clustering. Formula is simple, hclust is the code h class and then the input is d. and the method I am saying I am asking is use word's method you can use some other method as well.

So, if you just search for hclust, here they will say that okay, method is equal to complete or method is equal to single or method is equal to average, which one will you want you can choose, any one of these things you can choose.

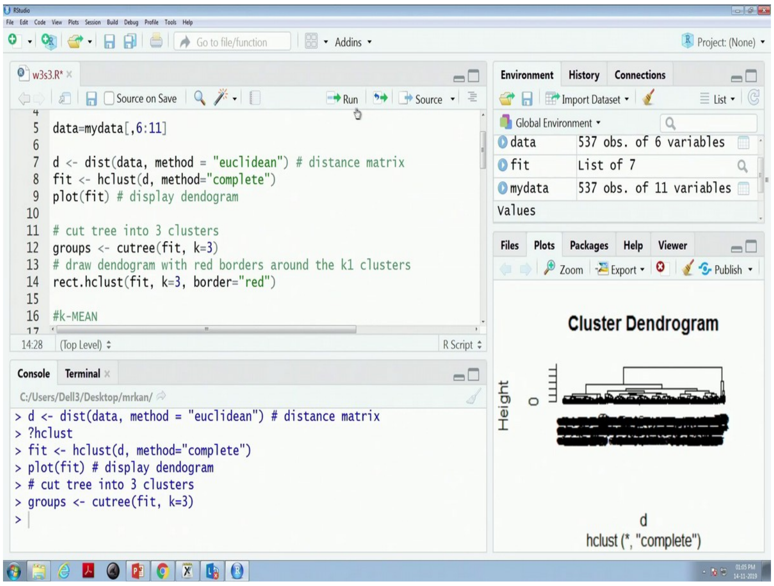
So, I am saying that let us say method is equal to instead of complete or whatever I want methods is equal to let us say centroid or let us say complete only, complete, method is equal to complete. So, complete means it will find out every distance. So, if I just run that, I got this. And then if I want to plot it, it gives me a plot and I want to show that. So, the below one is a Dendogram, study that the Dendogram carefully, so the all these small small small things I have 537 observations and they have come together, that is why it is so, so clumsy. But all of these things are each of these lines at the bottom is one single person.

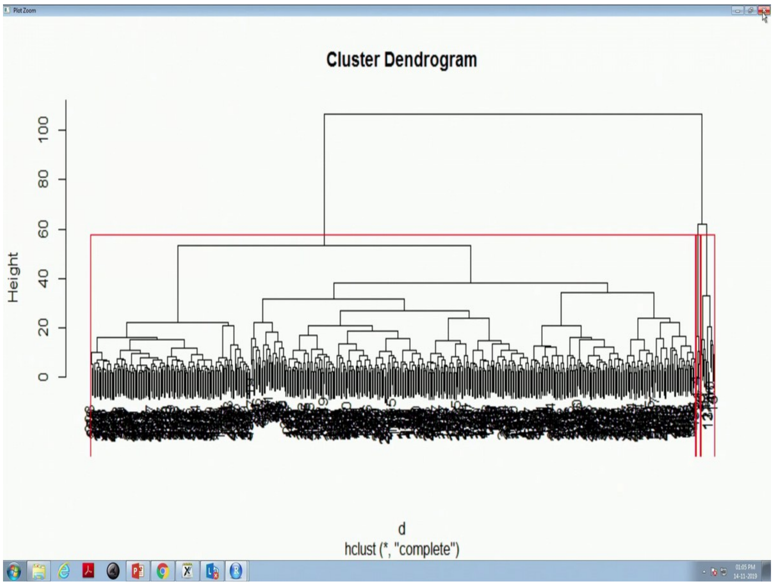
And then they join two persons, they join two persons in one group and when they join these two persons in one group, they actually achieve some distance. So, either you can start looking from the top and from the bottom. So, at the bottom, when everybody is in one segment you will see that these segments are so close to each other that you cannot even think that they are very different from each other. So, this does not make any sense to you.

So, why does it make sense? So, you will carefully see that these are the guys, sorry, these are the guys who are in one segment and all these guys are in another segment. So, there are two trees that is coming up at the top, one is this people and another is these set of people. So, that means this small group are very different from the other group. Now, remember this is a data that has been created for to have this kind of a data set. Now then, out of this huge group, the next term comes here that these people, these people are very different from this set of people.

So there are, I can see three segments, one is this, another is probably from here to here this segment and then a small segment at the back. So, there are three segments that I can see which is properly visible. So, when there are three segments, which are there, we can actually run for the rest of the thing. So, what we can do?

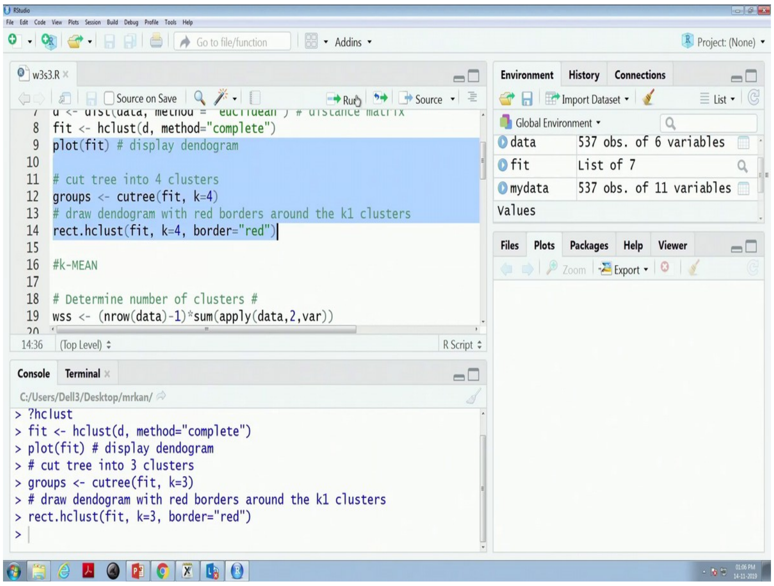
(Refer Slide Time: 8:17)

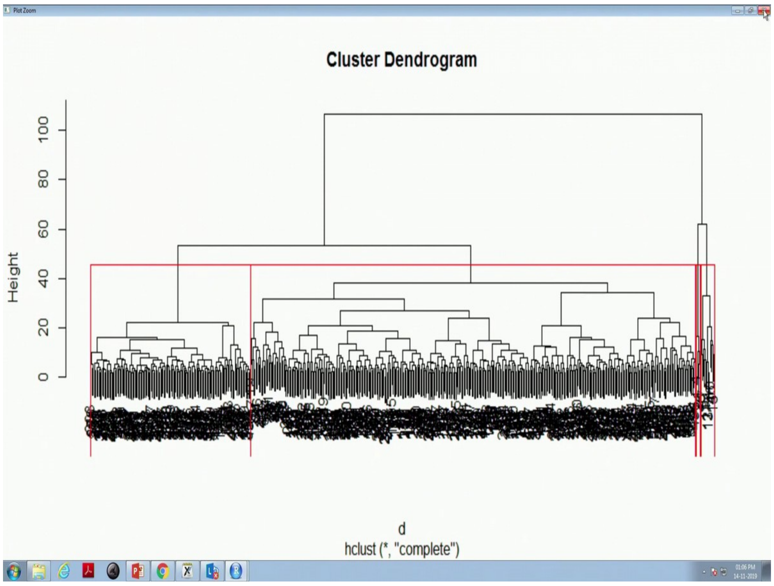


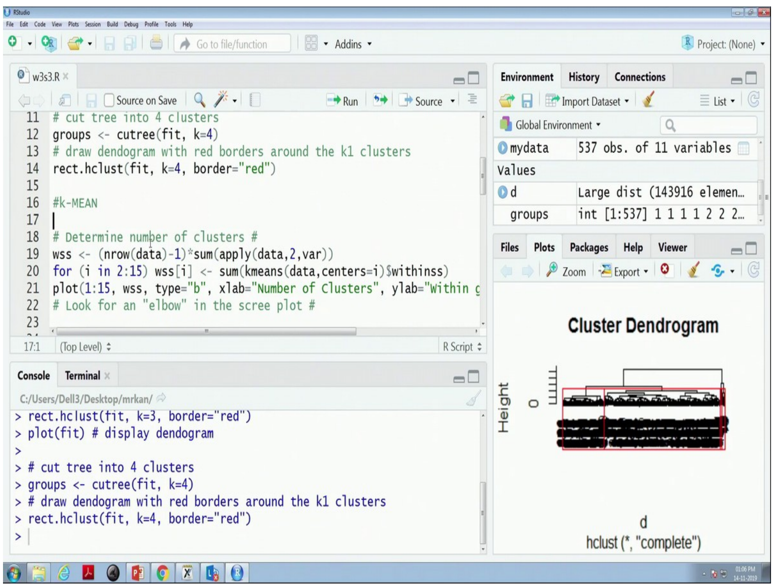


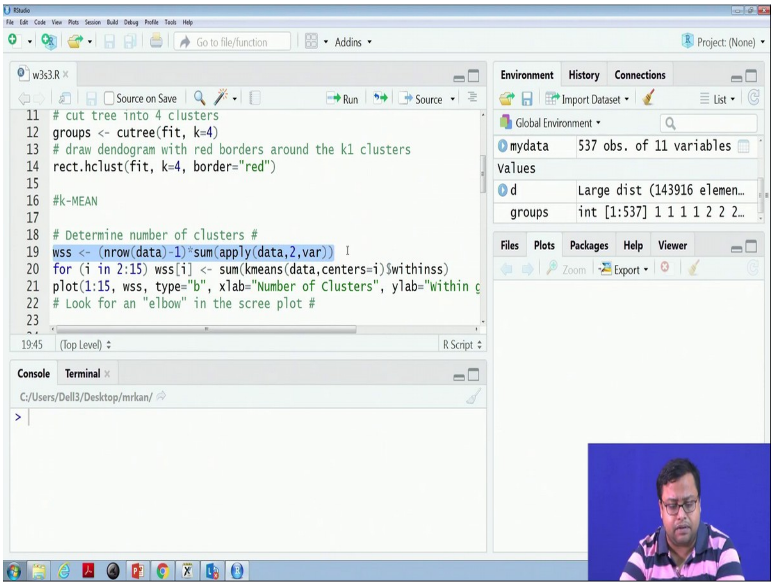
So, I will say that cut the tree in three segments. So, I will say groups cut 3*3, run and then border them and you will get the borders properly. So, here, so here there is one more segment that is coming up here very strictly. And I will talk about that segment probably.

(Refer Slide Time: 8:41)





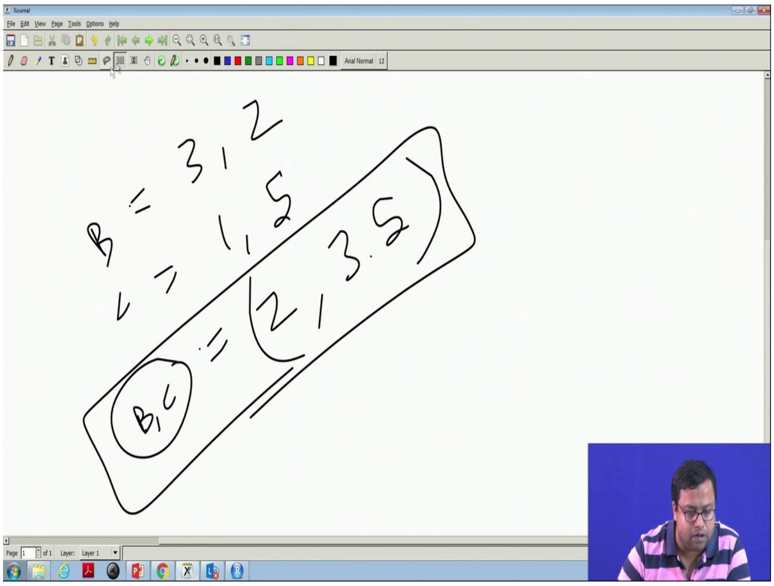


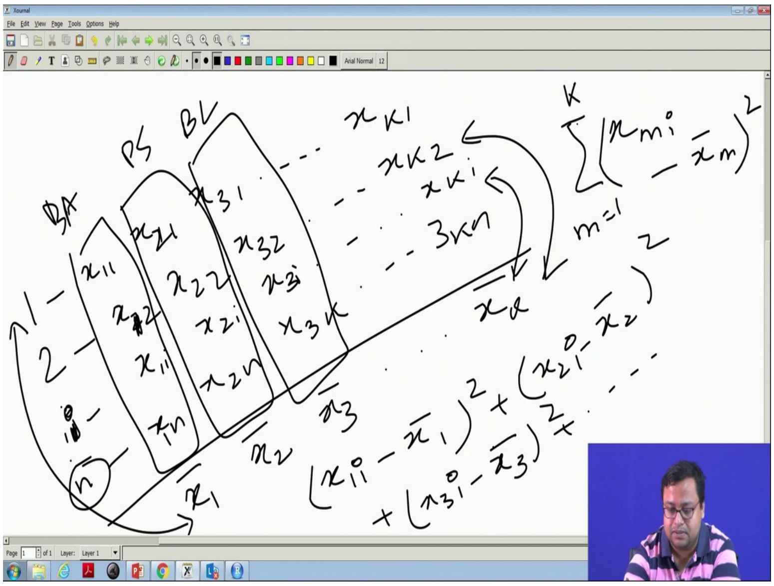


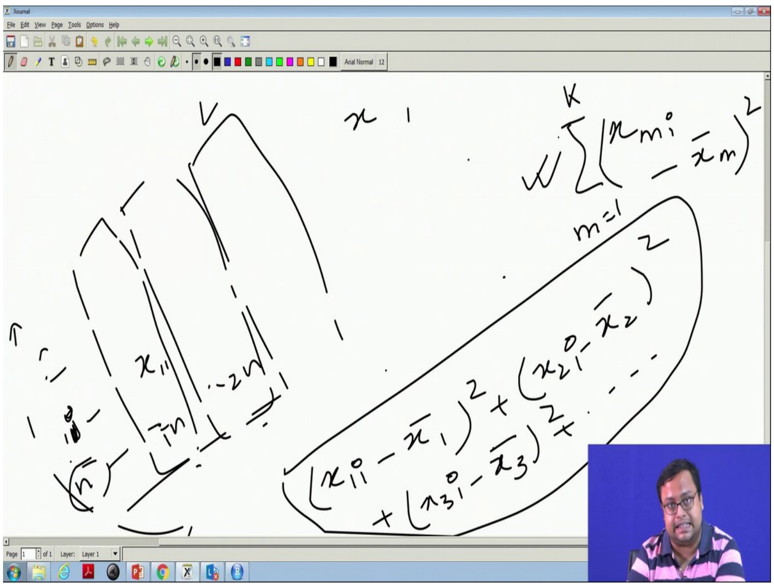
So, if I just clear this up and break it into four segments instead of three segments. If I run it into four segments, then I probably will have a better view. So, I plot this once more and then I break it into four segments and now carefully see that each segment has some meaning this is one segment, this is segment number two, which is a big one. And segment three and segment four are very small, but they are have some meaning. That is why they are coming up like that segments and we will discuss about that. The next step is k MEAN, so to k MEAN, I told you that you have to decide how many segments do you want.

So, one thing is I told that okay, you can have four segments using this other thing, other thing I will actually run that. So, what I am doing? I am writing. So, when there is no model, when there is absolutely no clusters, everybody is in the same thing. Then what is the net I would say, what is the segment I would say?

(Refer Slide Time: 9:51)







So, if you remember the formula, the formula was like this, the formula of each, when everybody were in different segments. The formula was like this that so what I will do, I will say that x person, so ith persons jth x value, so its ith person is let us say x_{11}, x_{12} up to x_{1k}. So, ith persons, jth characteristics, ith persons, jth characteristics minus the mean of jth characteristics, square them up, square them up, that is what? That is the mean of the distance of ith person from the mean and then sum that up.

So, what will I do? Once more, once more, carefully see what I am writing. So, let us say, let us say there are our first person whose observations are x_{11}, x_{21}, x_{31}, x_{k1}. The second person which is x_{22} sorry x_{12}, x_{22}, x_{32}, x_{k2}. Then there will be n number of persons which is x_{1n}, x_{2n}, x_{3n}, x_{kn}. And there will be in between a jth person, ith person who is x_{1i}, x_{2i}, x_{3i}, up to x_{ki}. So, total number of people is n and when I denote any one person I denote it with i. Now, if I take a mean what is a mean observation, the mean observation is actually the mean of this, then the mean of this, then the mean of this. So, each one will have one mean $x_1^/$, $x_2^/$, $x_3^/$ and $x_k^/$ everyone will have every parameter.

So, this is brand awareness, this is price sensitivity, this is brand loyalty, this is something else, everything will have a mean I tried when they are all in different segments. So, sorry probably, when they are all in one segment or in that case, how much is the error when when am I cannot explain anything of the model.

So, I will find out how 1 is distance from this, how 2 is distance from this, how three is distance from this and so on. So, I will find out all of that thing. So, what is that? So, how 1 is distant from this mean? That is $(x_{11} - x_1^/)^2 + (x_{21} - x_2^/)^2 + (x_{31} - x_3^/)^2 + \ldots$ this is the distance of one, guy number 1 with the mean.

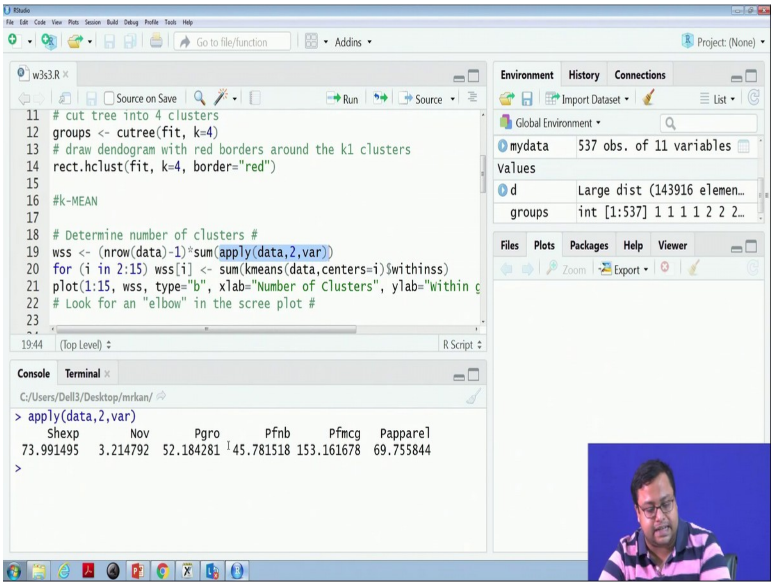
So, what is the distance of ith person with the mean? x_{1i}, x_{2i}, x_{3i} and so on. So, can I write it like this

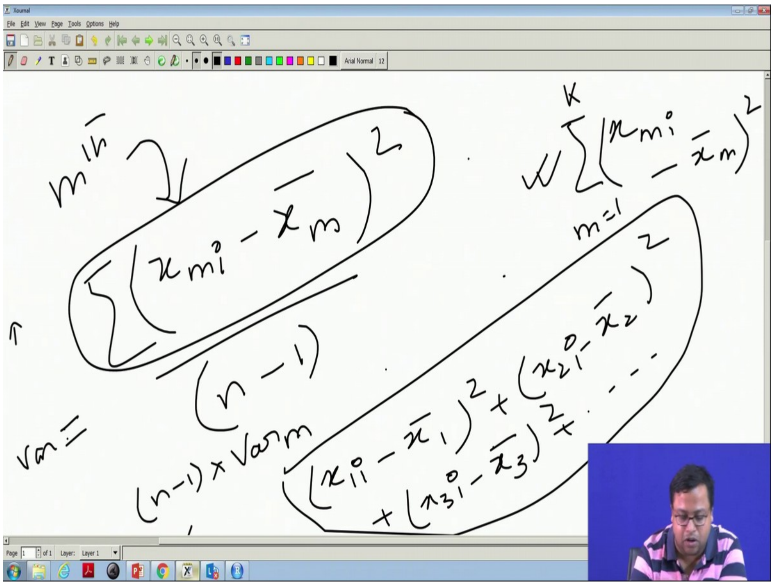
$(x_{1i} - x_1^/)^2 + (x_{2i} - x_2^/)^2 + (x_{3i} - x_3^/)^2 + \ldots$

i.e. $(x_{mi} - x_m^/)^2$ [for m=1 to m=k]

that x m i minus x dash m square summation m varies from 1 to k, I can write that. So, I can write this part carefully see, I can write so, I will just rub this off so that we can understand it properly. So, this is the part that I am focusing on. So, I can write the equation that has been written below this equation as this, I can write it like this. So, that is what I am writing and that is the distance.

(Refer Slide Time: 14:27)





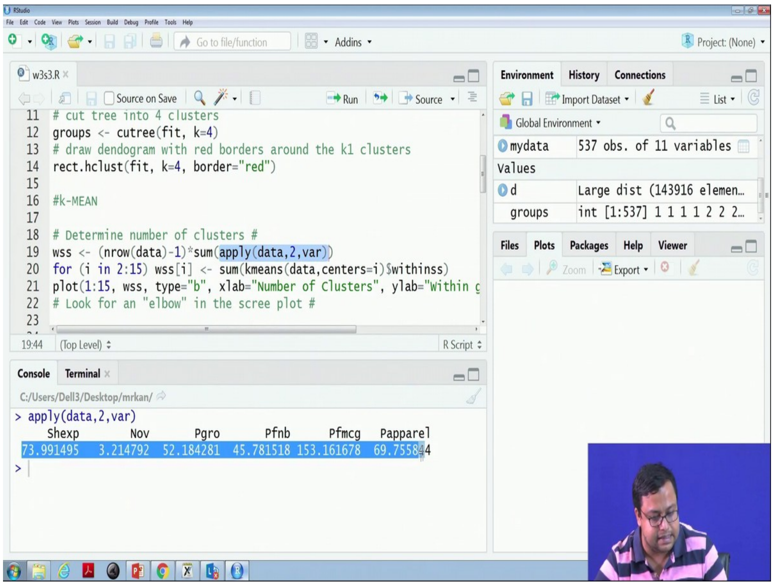
And then the so that is why what I am doing here in this code? In this code is I am writing apply data 2 var. What does this do? (apply(data,2,var)), if I just run this much, it will actually give the variance of each of the column one at a time, each of the columns hold data sets variance. Now, what is variance? Variance of a single column can be written like this. Just check, variants of a single column can be written like this, variance of a single column can it not be written like this.

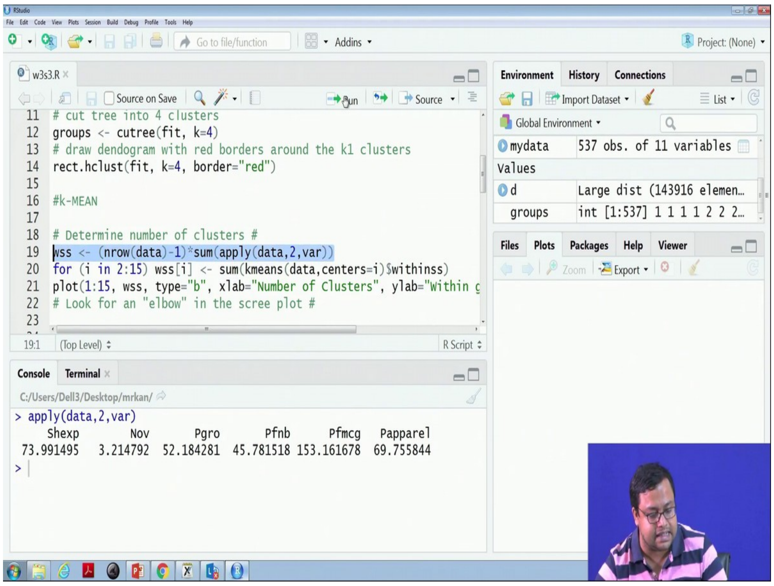
So, if it is mth column, then correspondence this thing is individual observations minus the mean of that particular guy, summation of that by n minus 1 number of observations minus 1 that is variance. The root over of this is, the root over of this is standard deviation. So, if that is variance, then can I not write that this one is nothing but n minus one into variants of my mth column.

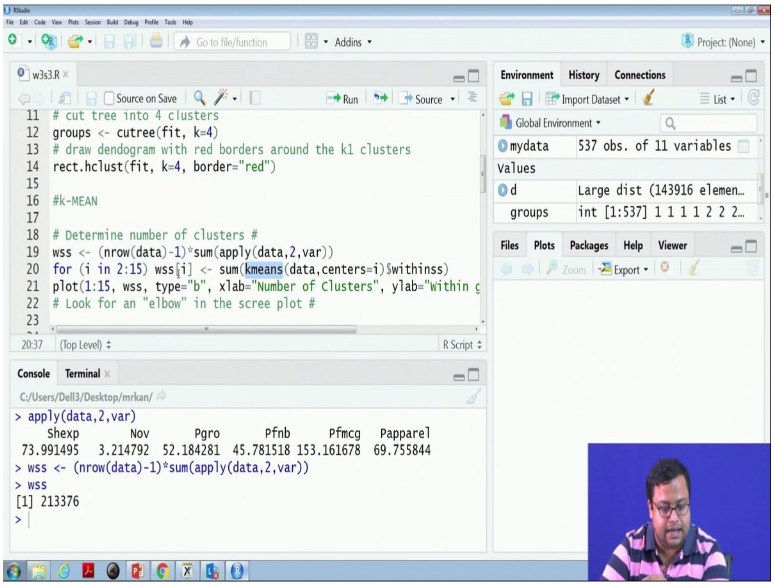
Variance= $\Sigma (x_{mi} - x_m^{/})^2 / (n-1)$

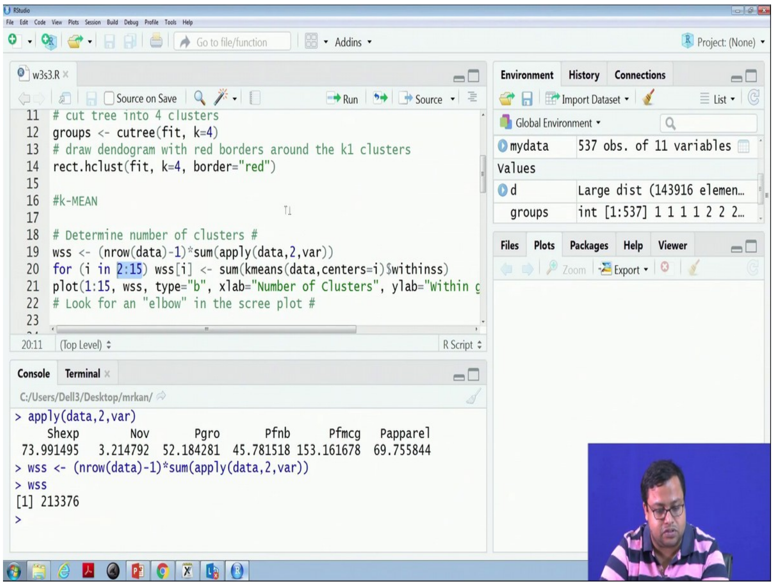
Standard deviation= $\sqrt{\Sigma (x_{mi} - x_m^{/})^2 / (n-1)}$

(Refer Slide Time: 15:51)





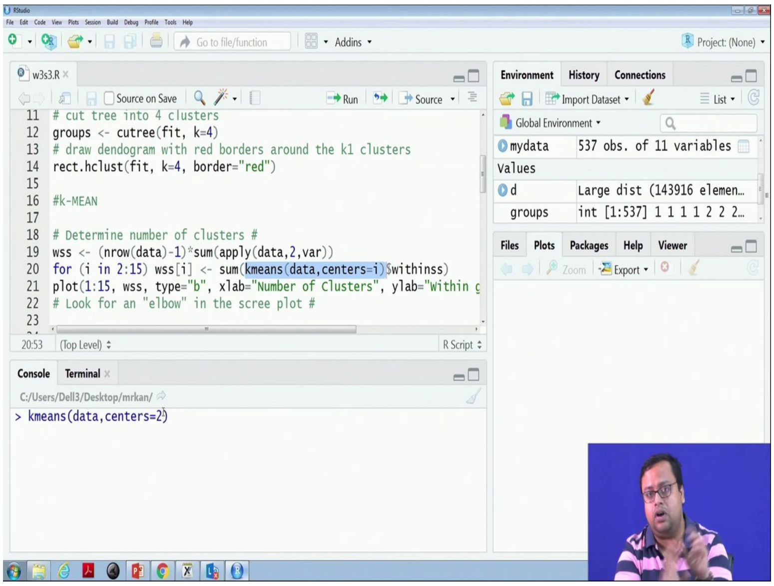


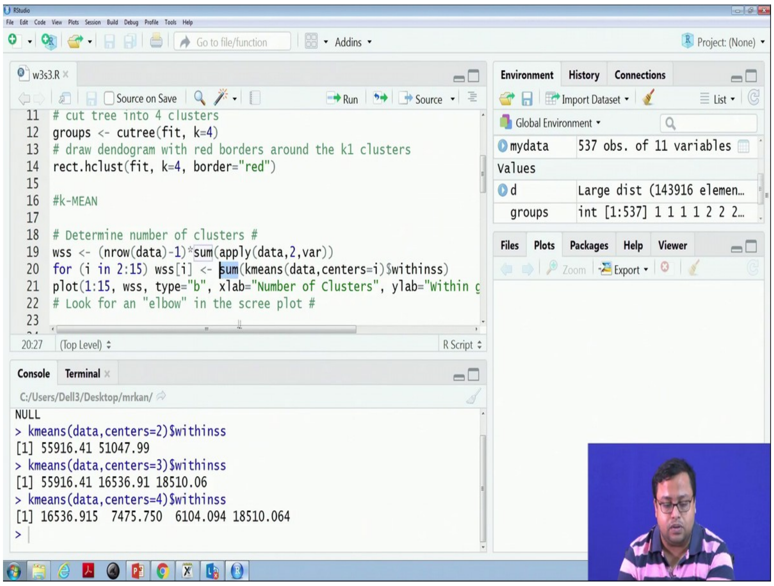


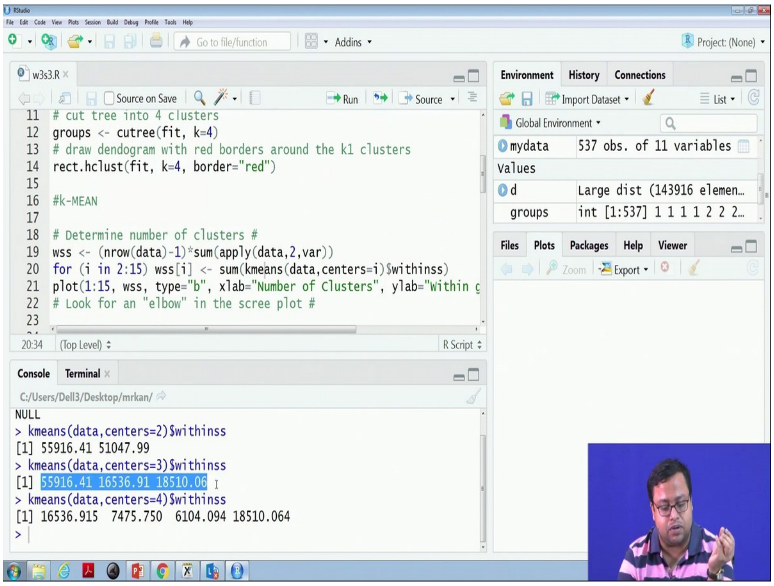
If I can write that properly, then you see that this is what I am writing. So, apply data to variance. So, these are the variances, then multiply with n row of data and n row data means 537 minus 1. So, multiply with it n minus 1 and then add them up, sum it up. So, this is the distance within sum squares when there is no clusters. So, when there is only one segment, if not more than one segment.

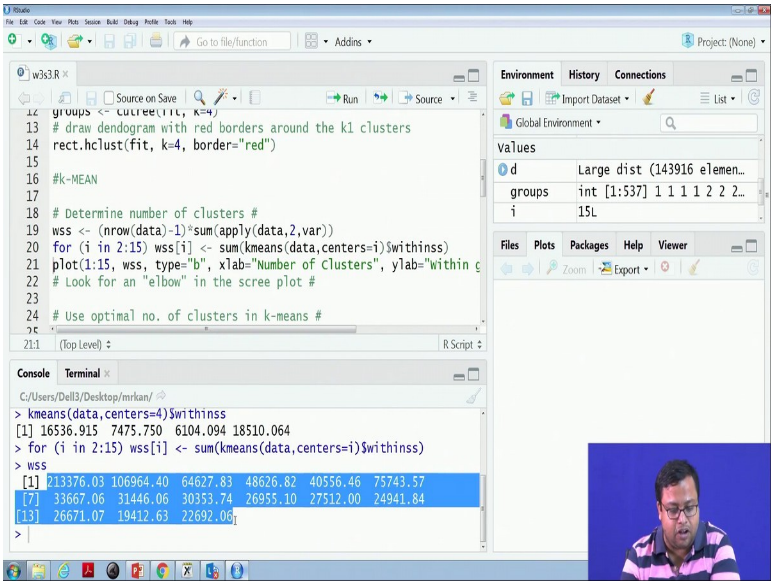
Now, when they are the more than one segments, I will actually find down that within some of square using a function called k means, k means and the i will vary from 2 to 15 means, I am varying the number of clusters from 2 to 15. So, when cluster number is 2, let someone cluster number is 2.

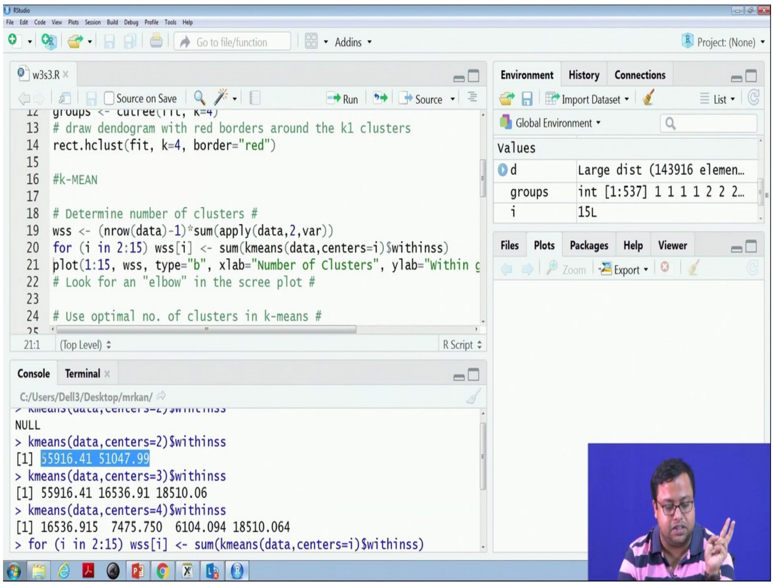
(Refer Slide Time: 16:53)











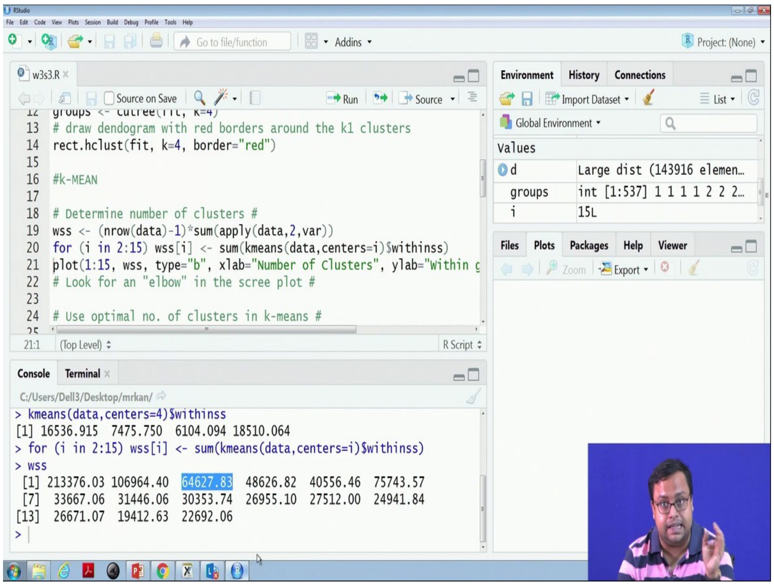
What does kmean will give? kmean will say that this is k means, what does k means say? k means will say okay centers is equal to 2. That means it will do read the data and create segments with two segments. Now, I do not need the segments right now, because I do not know how many segments I have. But I want to know when there are two segments what is there within ss, within sum of squares. So, if I break the data set into two segments and for each segment I find out the sum of square, how much how much total I get and that value is when centers is equal to 2 okay so, sum of that, sorry.

So, okay so, withinss this I have not written correctly withinss with. So, this is the 2 sum of squares of 2 segments. If there are 3, then each of the guy will have 3 different sum of squares. Sum of squares means the distance between each of the observations from the centroid of that particular segment from the mean of that particular segment. So, I find out the distance. So, the, these are the three distance when there are three segments.

When there are four segments, these are the distance and then add them up I joined them and add them up and that is why a sum sign. So, then I put those summation, so summation of these two and say that when segment is two this much will be that total withinss, when when within ss that means within sum of squares. When there is 3 segments if I add these 3 guys, this one will be that total within sum of squares. If I add these four guys up, this one will be the total within sum of squares. I save that in this wss thing.

So, if I run this I get a wss which is nothing but basically 15 observation. This is the observation when there was only one segment, this is the observation which is there were there are two segments which is nothing but summation of these two.

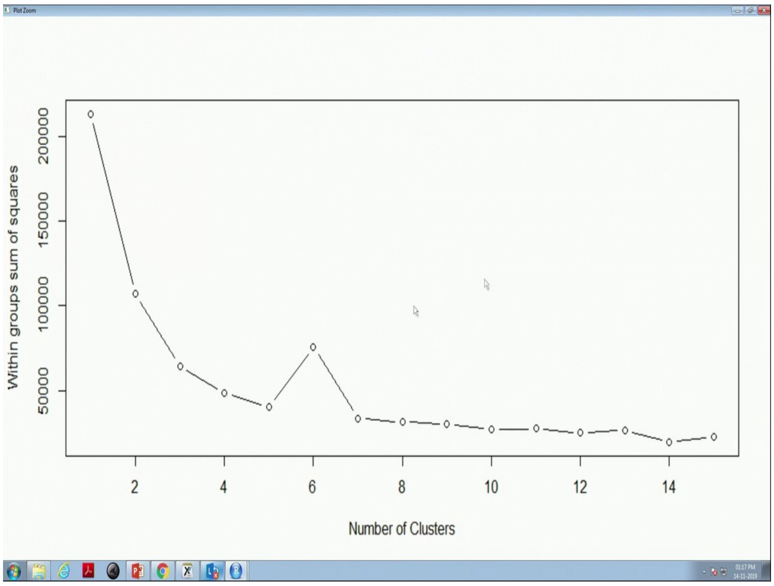
(Refer Slide Time: 19:00)

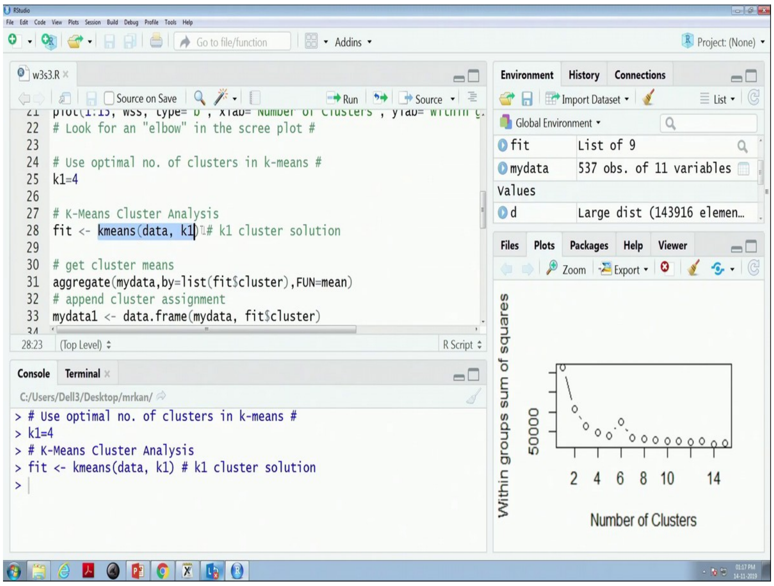


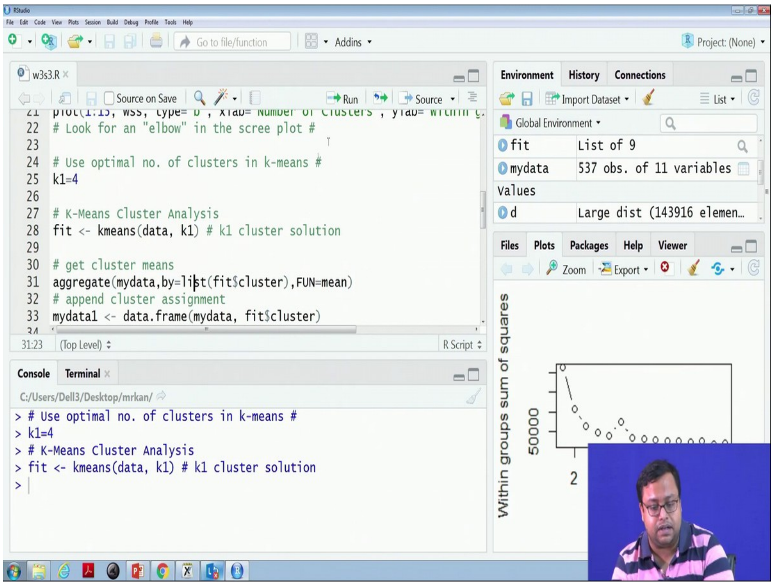
This is when there is three segments, so C, slowly the within sum of square is going down. When there is 1 segment you put bananas and oranges and apples everybody in one group. That is why the distance between the group mean from the group mean individual guys distance is very high. Now, when you have two groups, the bananas come in one group and apples and oranges comes in the second group. But because bananas are there in one group, the distance of individual bananas with the group mean of banana is 0 or very low. So that is why the distance comes goes down.

Though oranges and apples are still which were in the same segment but they are still different. But the overall distance has come down. Now, when I you further break, you break apples and oranges also into two different groups the within sum of squares further comes down, so slowly it comes down if you see.

(Refer Slide Time: 20:00)



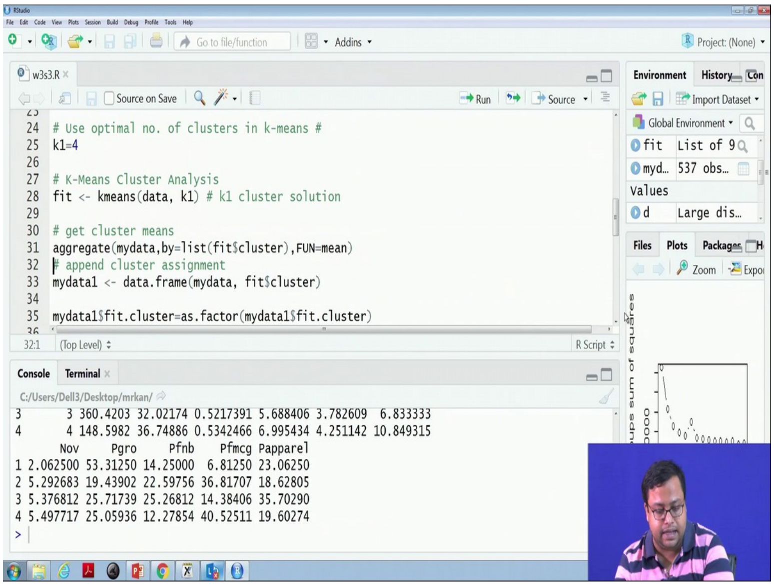


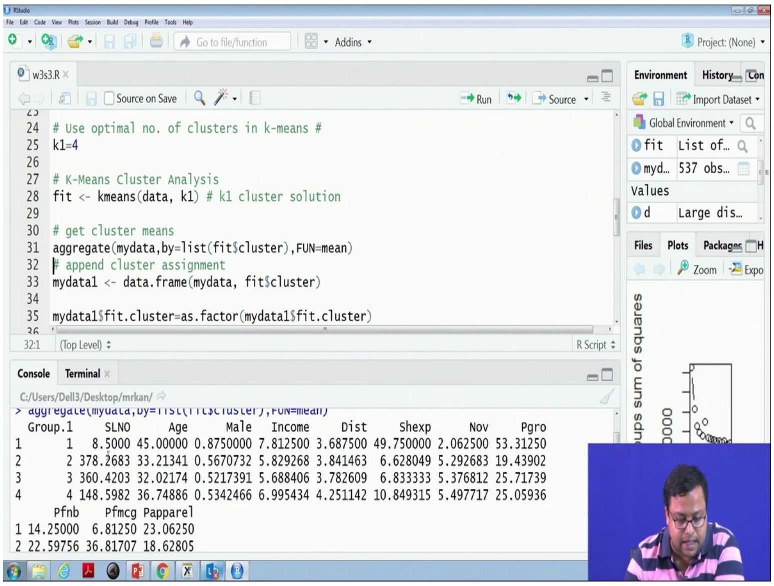


Now, I plot it and I want to see that how that plot comes down. So here, there is a germ but I can say that okay up to 4 or 5, I will take as a 4 or 3, it is a call, you can take the number of segments. So, let us k 1 is equal to 4 I decided. So, using this method I decided that k 1 I will take as 4. Now, if the number of k 1 means the number of means, number of segments I have take, we will take as 4.

So, then what will be the observations? I will just run this line which will break the data set into four clusters. And if I want to see how these clusters are, I am done and I have already done the work. Now, I am doing the third part, if you remember the third part is creating the segments, parameters the identification of the segment.

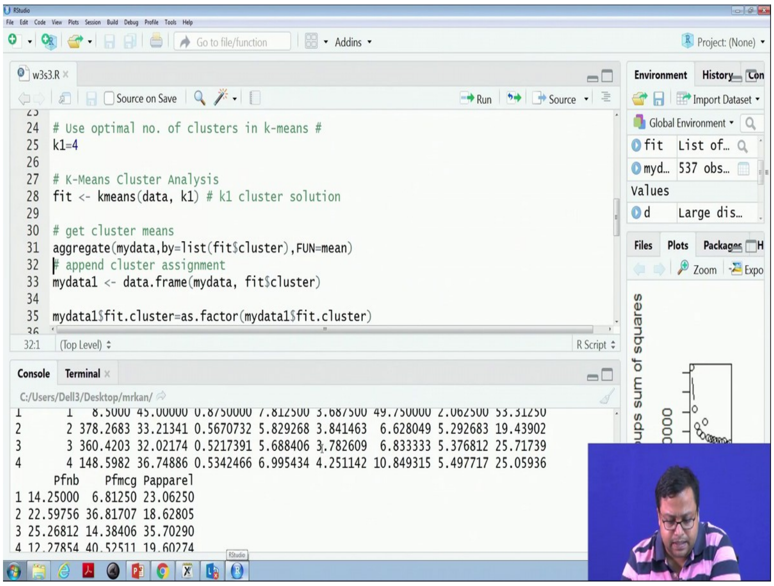
(Refer Slide Time: 20:50)





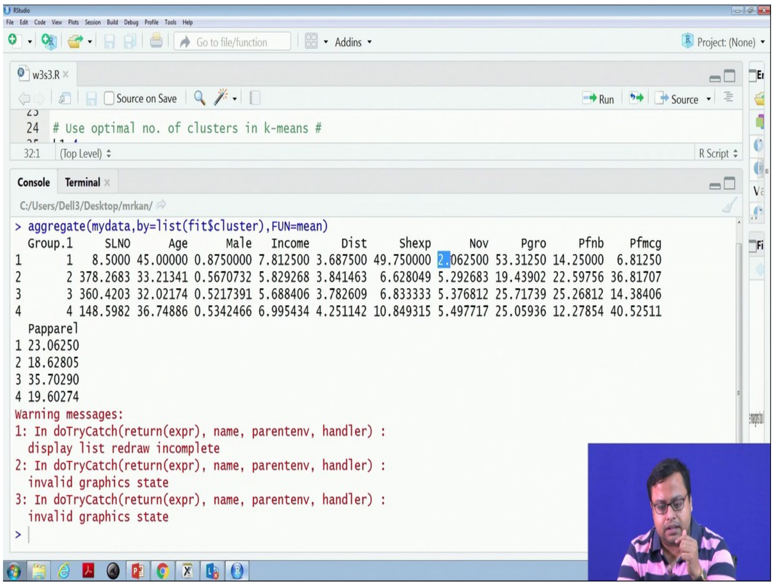
The identification of the segment will look like this. So, I have done the aggregation. So, I will just plot once more, the aggregate is like this, check it carefully. Serial number was, so the first, serial number does not make any meaning.

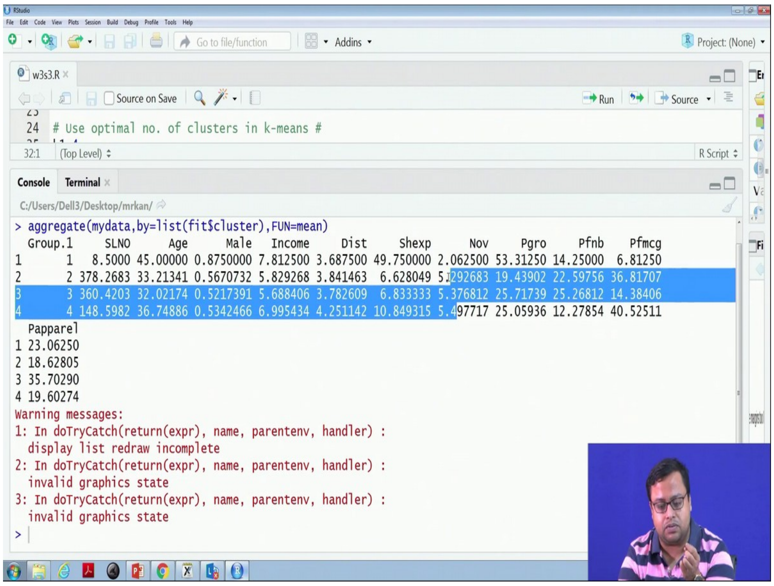
(Refer Slide Time: 21:06)



Segment 1 in terms of behavior if I try to only focus on behavior and nothing else and I will probably run once more, sorry. So, if I want to plot them once more in terms of the behavior, you will see that segment 1 has 49 parts 49,000 per month shopping experience, shopping expenditure. So their expenditure is 49,000 per month in one retail store. On the other hand, other guys are 6000, 6000 or (seven thou) 6.6, 6.8 and 10. So, I can say six point half and then seven and 11 something like that is the shopping expenditure. So, group one is high expenditure guys, that is number 1.

(Refer Slide Time: 22:01)





But group 5 visits very low amount of time only twice in a month and these guys average is 5, almost 5. So, group 1 is totally different from other guys, they visit for limited number of times, two times a month and purchase a lot. Their major purchase is grocery and apparel also, grocery and apparel is the major purchase, FMCG and food and beverage is not something that they purchase a lot.

What is the distance? The distance is not very different from each other, distance for all the groups are same. These guys income is a little bit higher than the other group. They are predominantly male 0.875 is male and middle aged. These guys are further younger 33, 32 and 36, this guy is 45. So, middle aged men, so middle aged men. So, if they are middle aged men, who are these persons. Can you, can you imagine?

So, there is a middle aged men comes twice in your shop only twice not more than that probably very low amount of time they come, but they bulk in, they purchase in bulk 49,000 average shopping expenditure means or probably total shopping expenditure. That means 25,000 in one shopping expenditure means they purchase in bulk and what do they buy? They buy grocery and apparel items.

Probably there, so, you do not buy 25,000 worth grocery or 25 percent worth apparel every month. So, a normal customer cannot buy and if that is not the case then probably these guys are B to B customers. That means they have small retail stores, they come and buy from these big retail store and sell it in the small retail store. They can be resellers, they can be small kirana stores or something like that. So, that is what a segment which is coming up, which is very prominent.

What are the other segments? The other three segments are, their shopping expenditure is 6, 6 and 10. So, they are not B to B, they are B to C. Now, though all of them visits around 5 times, so they are weekly visitors so, that is okay they are not very different about that. But segment 2 if you carefully see or probably segment 4 if you carefully see the FMCG is 40 percent of their purchase which is majorly they come to buy FMCG.

So, who are these people? They are a 50, 50 male-female, more female than, some more male than female. And their age is a little bit higher than the other group. And their income is also little bit higher than the other group. So, these guys who are a little bit higher income will focus more on FMCG and they are ready to come from a long distance also. So, they are coming 4.25 kilometers away from there also.

On the other hand, there is another group whose major purchase is apparel if you see 35 and these guys are youngest 32. And they are also more or less more famle they are than the other groups and then their distance is the shortest 3.78. So, see the closest guys are obviously the B to B guys and then these guys who come to buy apparel, they can, they also want to want to have a lower distance and then group 2 majorly buys FMCG 36 percent also.

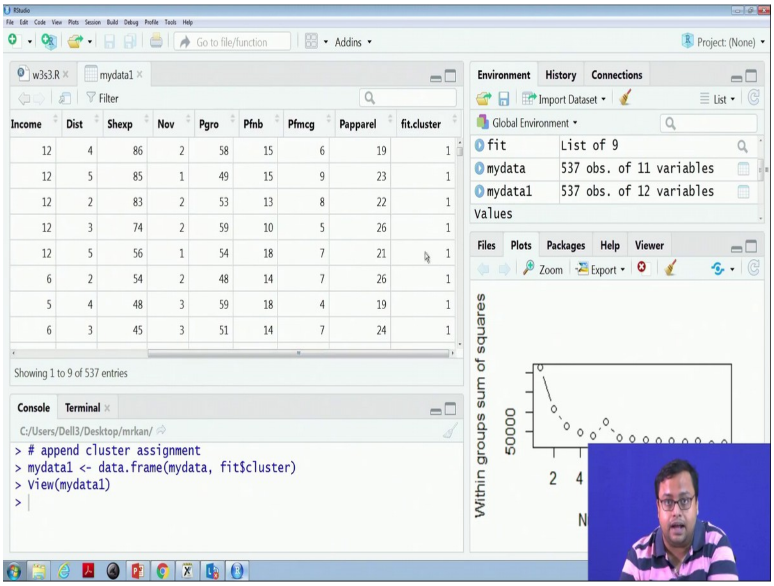
So, okay so we have to check now that how group 2 and group 4 are different. Group 2 also buys 36 percent FMCG, group 4 also buys 40 percent FMCG. Group 2 also makes five visits, group 4 also makes around 5 visits. Group 2 shopping expenditure is 6, but these guys shopping expenditure is 11. So there is some difference in terms of shopping expenditure and what does that come from? That come from probably from income, this guy's income is 5.8 it is 7.

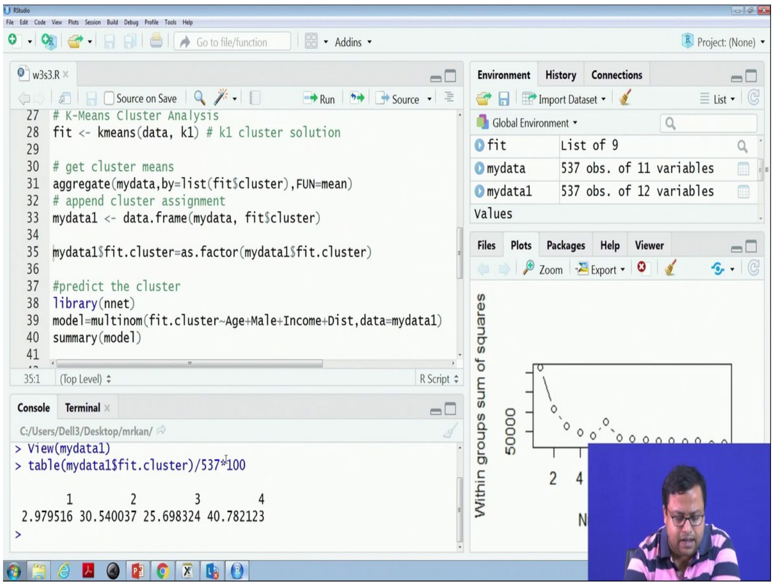
So, one lakh extra income that should not impact the shopping expenditure so much. There is any other difference, okay. These guys actually make expenditure on food and beverage also these guys have very low expenditure on food and beverage, only 12 percent versus 22 percent. So, does group 2 I can say that group 2 are mainly male, who are of age, average age of 33. And their income is around six lakhs per annum. They live close by within 3 kilometer or 4 kilometer distance from the from the retail store and their average shopping expenditure is around six and half thousand per visit or per month. And majorly they buy both FNB that means food and beverage and FMCG, FMCG is the primary but they also buy significant amount of food and beverage material.

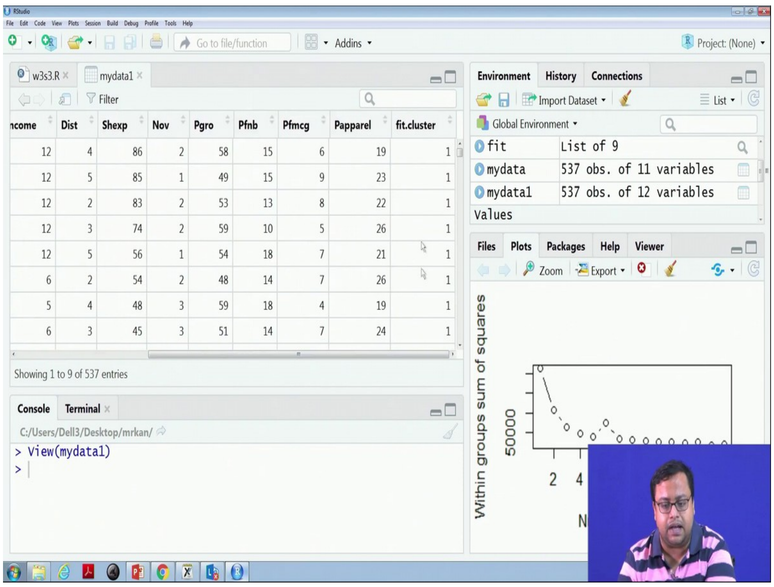
On the other hand group 4 have all of these things similar but their shopping expenditure is around 10,000, their income is around seven lakhs. And they only focus on FMCG, they do not focus on the rest of the thing. Probably they focus on groceries also FMCG and groceries. So, one group probably focus on food and beverage, which is packaged. Another group grocery means their food and beverage which is not packaged, which is, so the focus is different. The age is also a little bit higher for group 4.

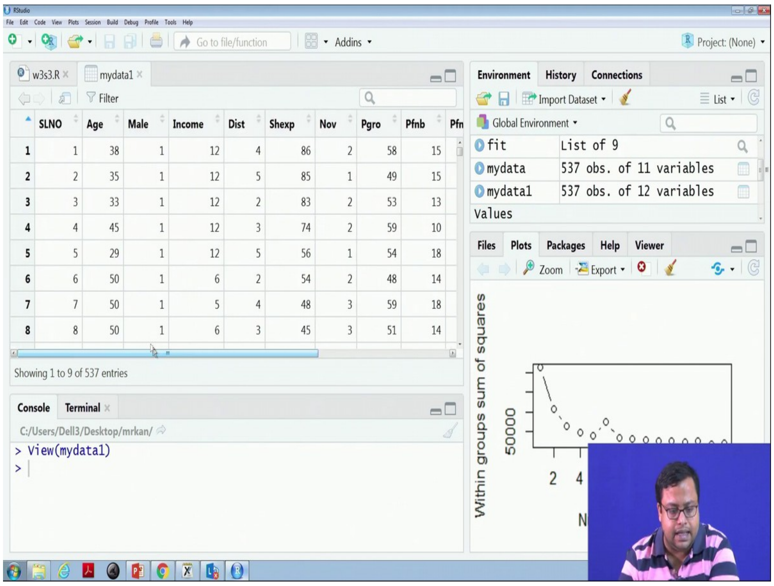
So, I can probably assume that these guys and age income is higher, distance they are coming from a little bit longer distance. So, I can imagine that this fourth group is a family person while the third group is not a family person or if fourth group might have a larger family, might have a car because he is coming larger distance and etcetera. So, these are giving my some idea about what the segment is, from the demographics and from the experience.

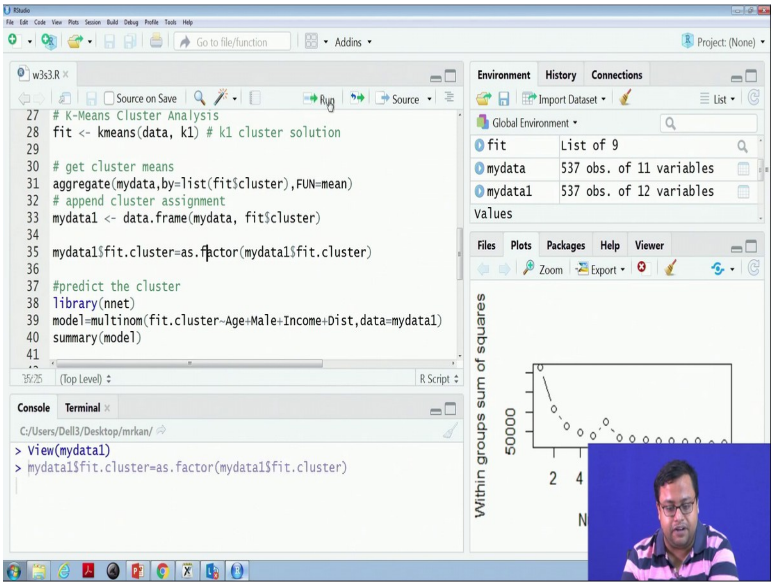
(Refer Slide Time: 28:18)









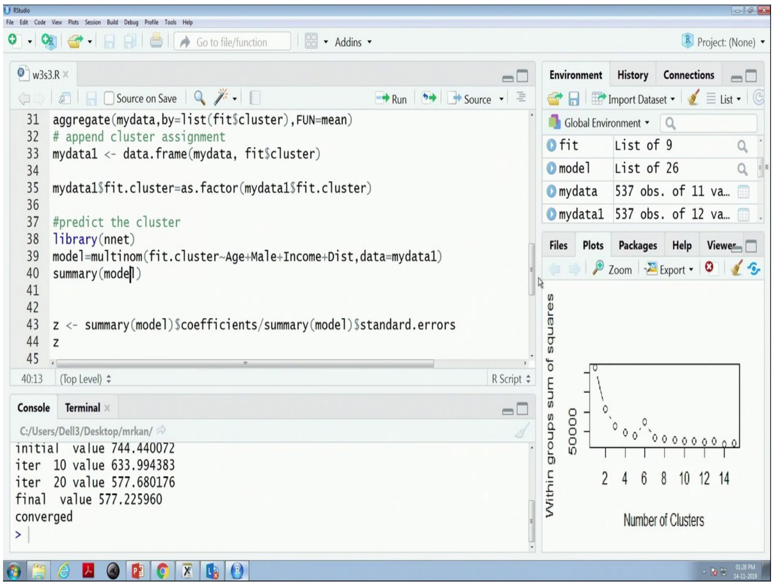


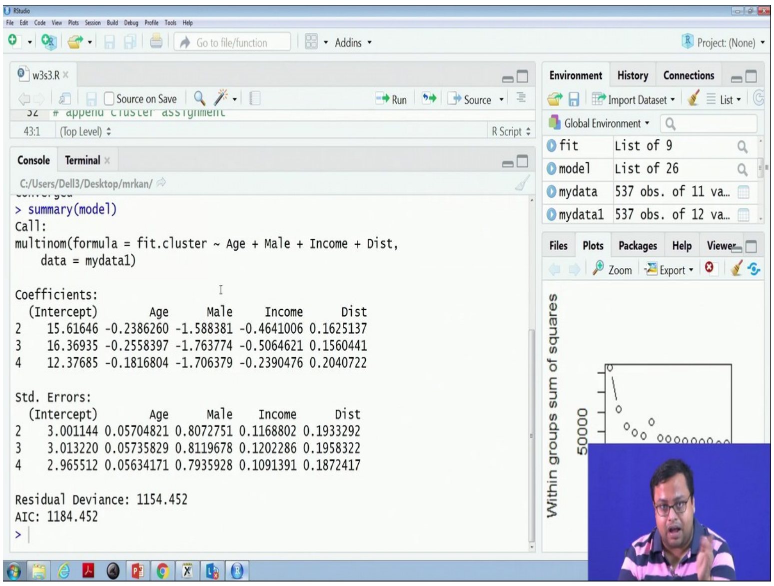
And now, at the last stage what I will do? I have to create a targeting mechanism. I have to find out that if a new customer comes up and registers with me, how will I know whether he is in segment 1 or segment 2 or segment 3 or segment 4? So, what I do is I put another data set, where I quit my data 1, where I actually have put the cluster numbers the who are in, which guy is in which cluster I have did that. So, if I just write my data 1 dollar fit cluster and then try to find out a table of that divided by 537 into 100. In percentage term I know in segment 1 has only 3 percent people.

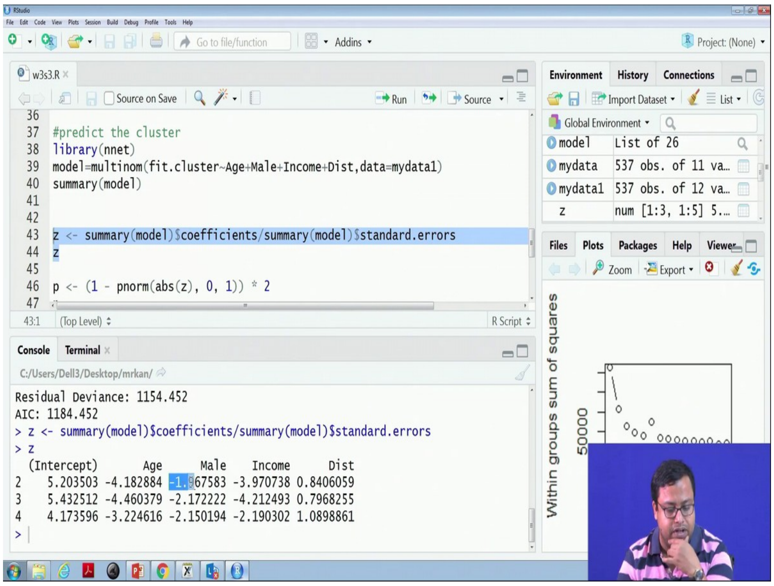
So, though they are very prominent, they are in number of, in numbers they are very small make sense because B to B buyers will be small. How many kirana stores will be there in the locality 20, 50? But in a, in, there will be probably 10,000 customers and 50 retail stores. So, B to B purchases will be less so that is why 3 percent. And the other one is 30 percent, 25 percent, 40 percent, so they are fairly well sized.

Now, I have in my my data if you remember, I have this as my positions of the persons and this as my age, male, income, distance these are the four demographic variables that I have. So, using this demographic variables I will try to predict whether he is. Now 1, 2, 3, 4 there are four categories, they are there. And this is not a linear regression because all these four categories are different. So, I change them to factor variable as factor.

(Refer Slide Time: 30:10)







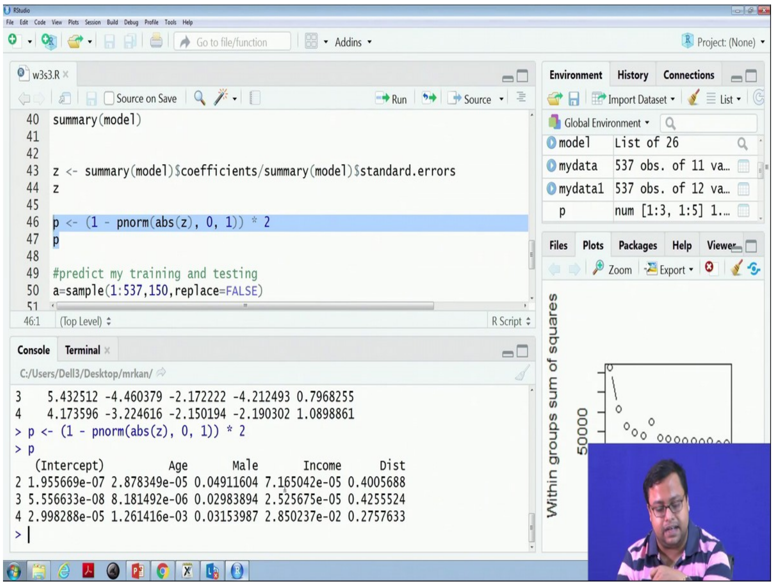
And then I run a multinomial logistic regression. So, for that I will require a library called nnet and then I will call this model. So, I will run multinom instead of lm I have written multinom. That is the only difference then fit dot cluster is mine, while variable age, male, income and distances my x variable and data is equal to my data 1 and if I just run this one and if I just see the summary of the model, I get this thing. So, in the summary of the model what it gives me? It takes 1 as, the observation 1 as the base point. So, observation 1 is 0. Observation 2 is the intercept is 15.

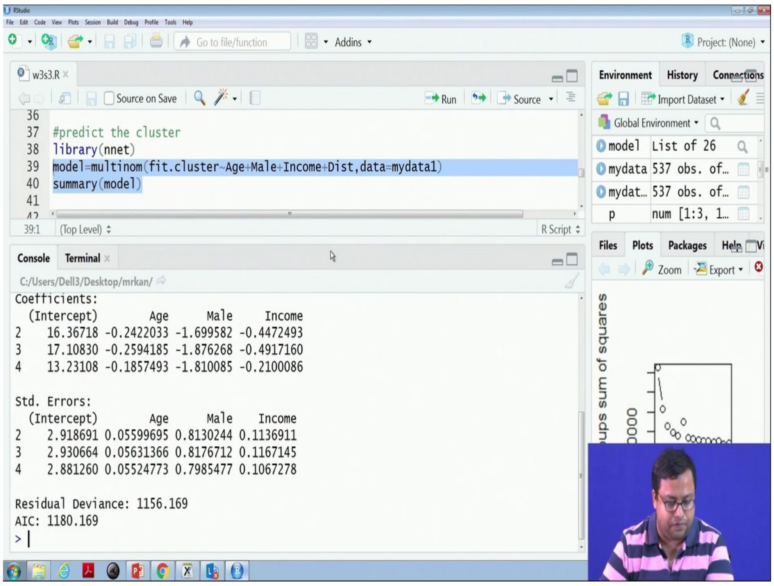
For each age increases, so, when age increases from 0 to 1 or for unit age is increase, the chances that you will be in group 2 is least or probably not group 3 is least. So, age increases your chances of being in group 1 increases. And if you are a male, then also the chances of in group 2, 3 and 4 decreases the highest chance is group A. So, these are all negative. Income, if your income increases your chances of being in group 2, 3, 4 is also lowered and probably lowest is in 3 and probably in comparison to that 4 is much better.

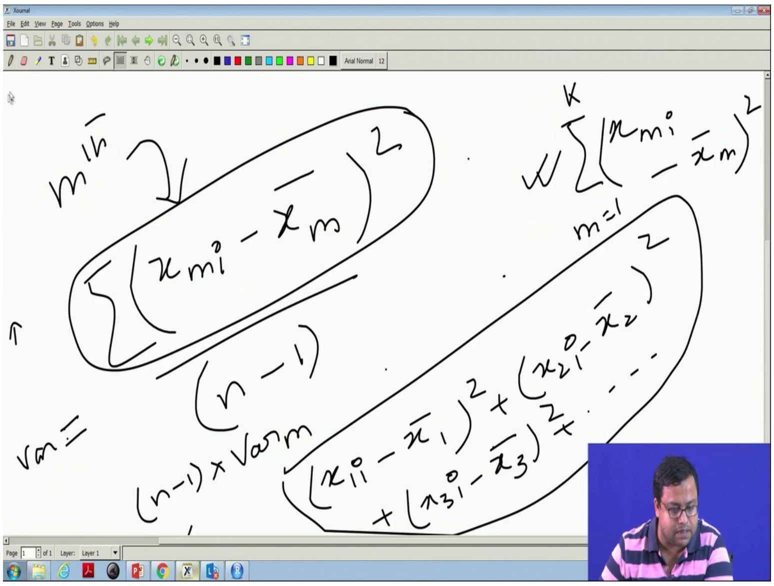
And as distance increases the chances of being in group 4 is highest. And these are the corresponding standard errors. So, what, how do we find out the how, whether they are significant or not? We do the coefficient by the standard error that will give us the T statistic or Z statistics.

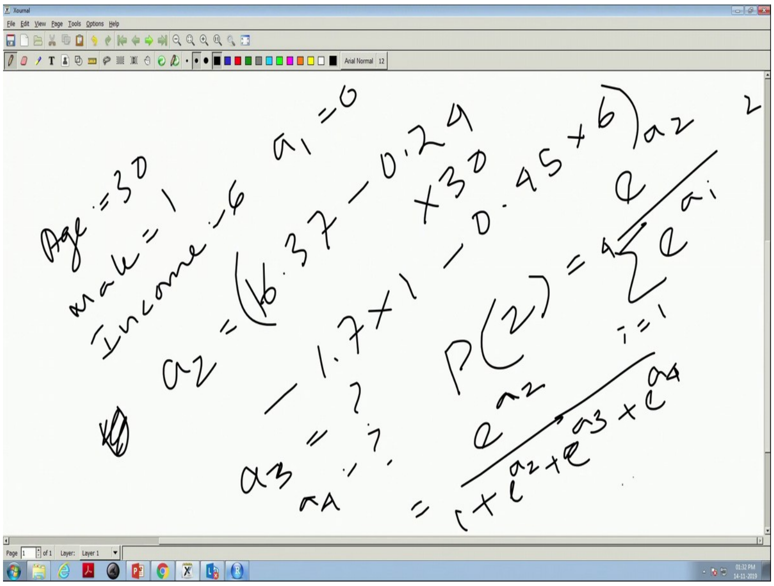
So, we find out that Z values, so Z values are for all of them probably this one is marginal and these guys, distance is not significant. The Z value is 0.84, 0.79, 1.08. So, remember it has to be higher than the mod value of the Z has should be higher than 1.96. So, but distance are not significant then, but the others are significant probably this one is also not significant.

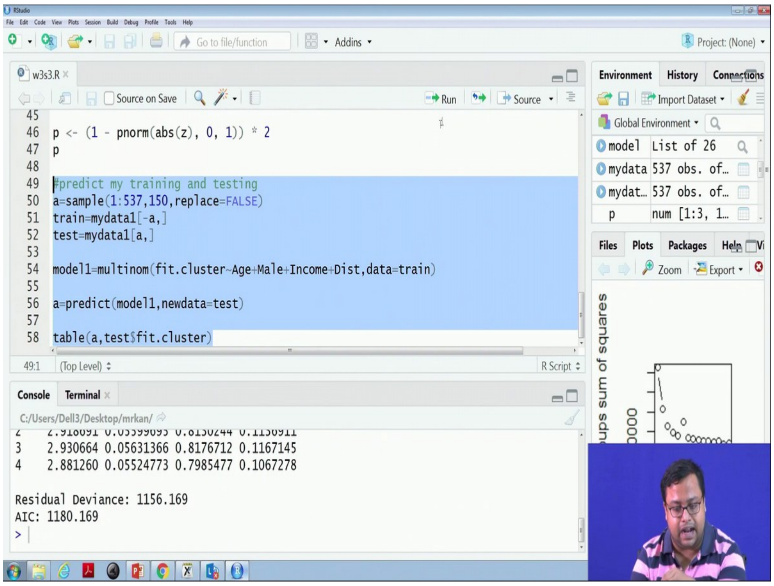
(Refer Slide Time: 32:27)











We can find out the probability, exact probability values and the p scores, okay. So, other than distance, which are all higher than point 0.5, the rest of the four things intercept, age, male and income. For all the observations, they are significant. So, that means that I can probably run this thing. I can probably run this thing using not using distance and then run and this is the right score. So, how to interpret it? If somebody comes as an age of, if somebody comes with this observation let us say, if some, somebody comes with a observation of let us say his age is of 30 years and he is a male. And his income is of six lakhs then what is the observation? The probability that he will be in group 2, first of all it is the U or whatever I do not know. So, let us say a of group 2 is basically 16.37 minus 0.24 into 30 minus 1.7 into male minus 0.45 minus 1.7 into male because 1 minus 0.45 into 6.

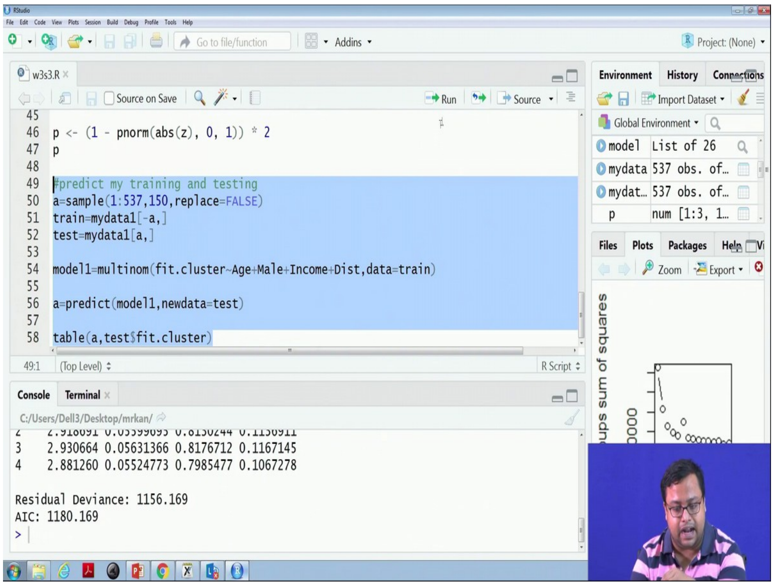
$a_2 = (16.37-0.24)*30-(1.7*1)-0.45*6$

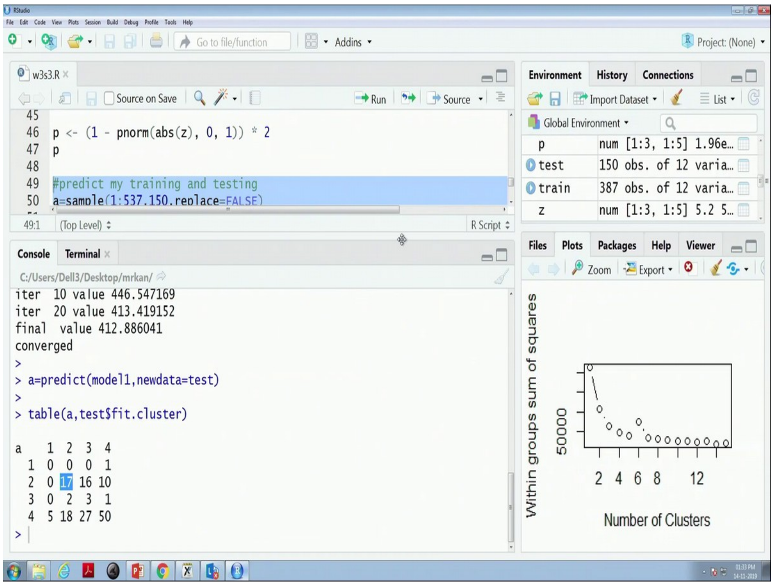
So, this is the probability that he will be in a 1 walk a 2. So, this is not probably, this is the measurement. And similarly I find out a 3, a 4 also, what is the values? And a 1 is equal to 0. So, probability that this guy will be in group 2 is basically e to the power a 2 by divided by e to the power a i's where i varies from 1 to 4.

In other words, e to the power a 2 by 1, 1 why? e to the power a 1 is e to the power 0 that means 1 plus e to the power a 2 plus e to the power a 3 plus e to the power a 4. So, something like that will actually give me the probability that this guy will be in group 2. And group 3 and group 4 and whenever, whichever segments probability is higher, I will put that person in that particular segment.

Similar thing we can do with LDA also. And I will not do that, we can also break the model and see that how is the predictive model. So, I have broken the data set in training and testing, created the model with the training data, predict it with the testing data and finding out the confusion matrix, it is a similar job that we have done for logistic regression you can try out that.

(Refer Slide Time: 35:38)





And here if I run these four lines together, this is the confusion matrix gets created. You can see that there are lots of off diagonal elements. Now, what are the accuracies? Basically 50 plus 353 and another 17, so around 70. So, 53 plus 17, 70 out of how many? Out of 150 or what, just one minute, so training data is after 150, 70 out of 150 is less than 50 percent which is bad. So, you have to find out some other demographic variables which explains the data set better and you have to try to improve your predictive modeling.

So, once you predict better predict the segmentation, which segment they will fall, the targeting becomes much easier. So, that is what we will have done about logistic regression, multiple logistic regression. We can discuss more about LDA in the next class with a small example on these data set itself. I will share 2, 3 lines of code and we will discuss that and we will go ahead with the next module from the next videos. Thank you very much for being with me. We will meet you in the next video once more. Thank you.