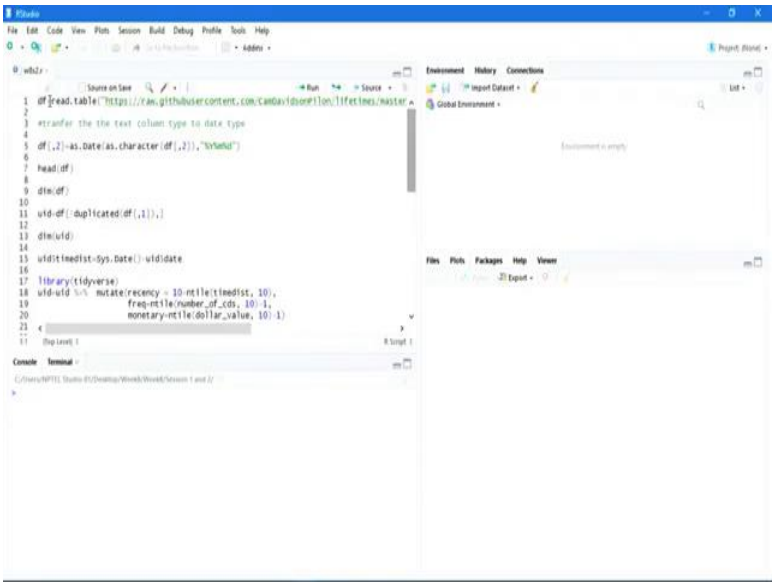
Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology Kharagpur
Lecture 41
RFM and Market Basket Analysis (Contd.)

Hello everybody, welcome to Marketing Analytics course, this is Doctor Swagato Chatterjee from VGSOM from IIT Kharagpur who is taking this course and we are in week 8, session 2. We are discussing RFM analysis, Recency Frequency Monetary Analysis and I told in the last video that this is a analysis technique which is used to do behavioral segmentation of customers based on their purchase data.

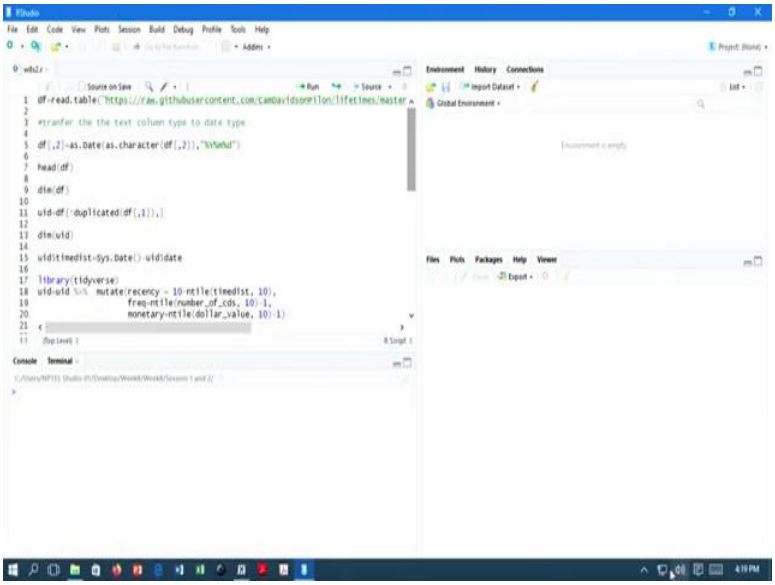
And R stands for recency which is, how close, how recently you have purchased and F stands for frequency which means that, how many times you have purchased in a certain period of time within a certain period of time and then monetary is how much money you have generated, how much expenditure you have done within a certain period time. So, based on that, we will do segmentation, RFM score high means is a profitable customer and RFM score low means this is a customer which you have to deactivate.

(Refer Slide Time: 1:14)+++



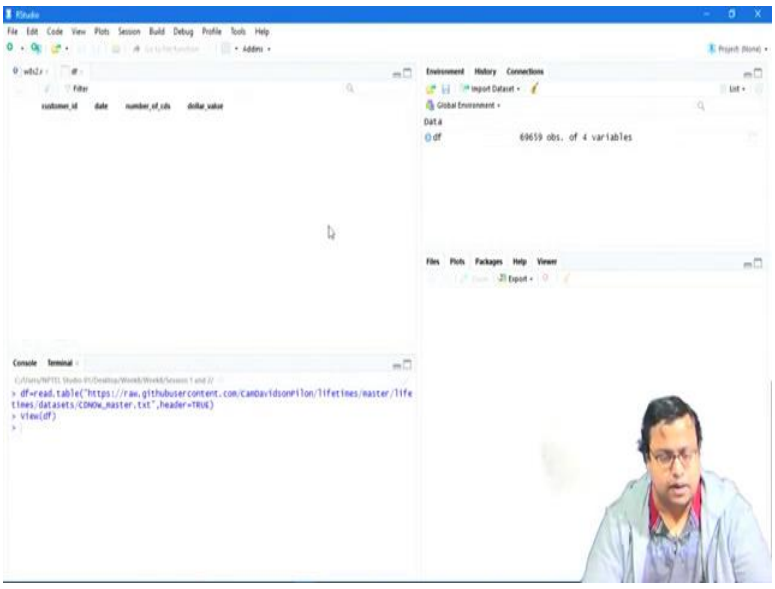
So, here in week 8 s to dot R file, if you open so, you will find out it is in the in your in your file link in the in the in the in NPTEL portal that you have and there if you just download it. Remember in this if you just carefully see the code here we have not used any data set data set is there in internet and I will right now show you how to connect with Internet in terms of the data creation.

(Refer Slide Time: 1:45)



So, you have to first check that the internet is connected which is there for me. And so, I will just run this line. So, read dot table and then the text. So, the when if if you can go and check this particular link and see that there will be a file there is nothing, one particular page where lots of data has been written which is like a notepad, but ever you write in the notepad similar thing is there. So, when you read that kind of a data, you use read dot table and if by chance the header names is at the top you should write header is equal to true and that is what you get.

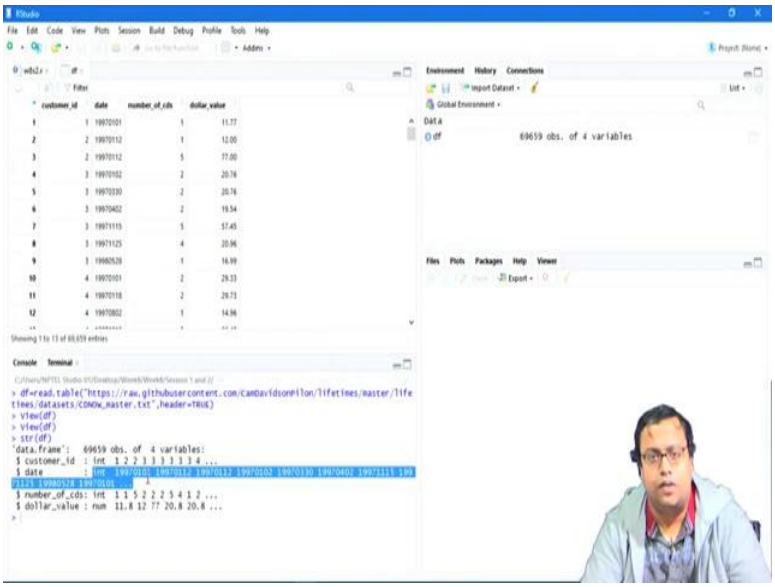
(Refer Slide Time: 2:23)



So, I have got customer ID, the date, the date is given as, this is what nine 9, 9, 97 01 01 means, first January 1997. Number of purchases and dollar value so, RFM data has been

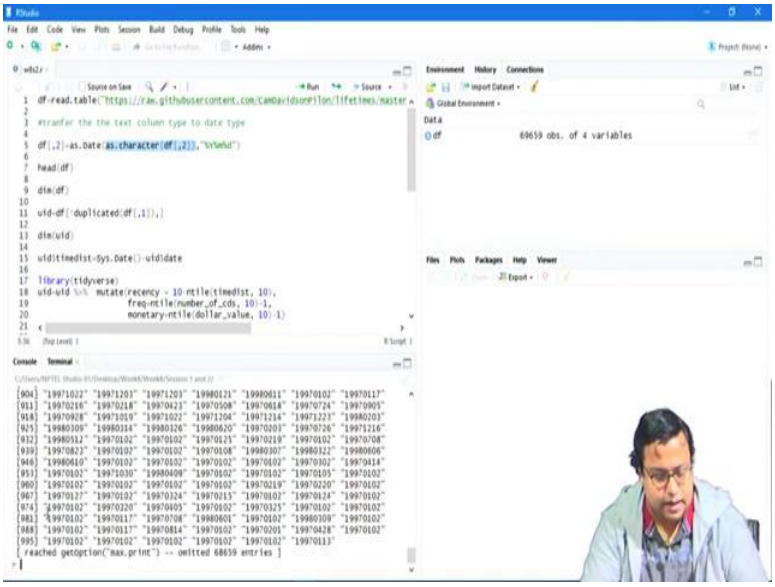
given to me. So, the first job that I have to do is to find out that how we recent this data date is? The last purchase date is this that, how recent it is? So, I have to convert this particular column which is the date column to our.

(Refer Slide Time: 2:57)



Right now, down for just check the structure of my data set, see the date is basically a integer, see it is an integer, so I have to somehow convert it to date. So, the first thing that I am doing here is changing it to character.

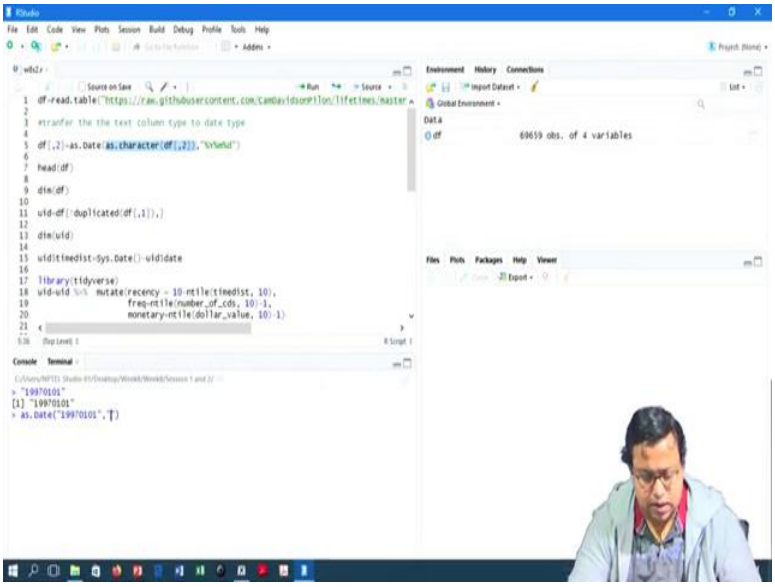
(Refer Slide Time: 3:12)



So, if I just write as character of D F 2 what will I get? Just copy this and paste it here, and press an enter. See each of them has been converted to a character form. Now, only a

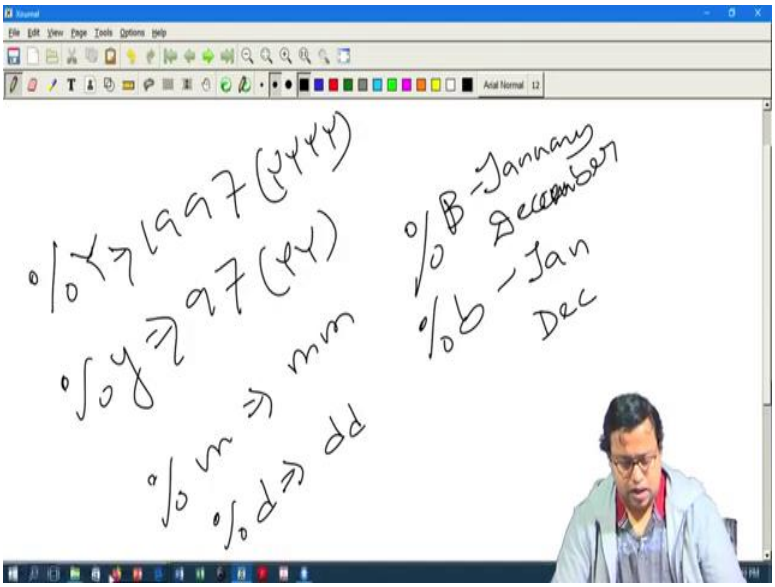
character form written in a specific format can be changed to a date form using a function called as dot date.

(Refer Slide Time: 3:34)



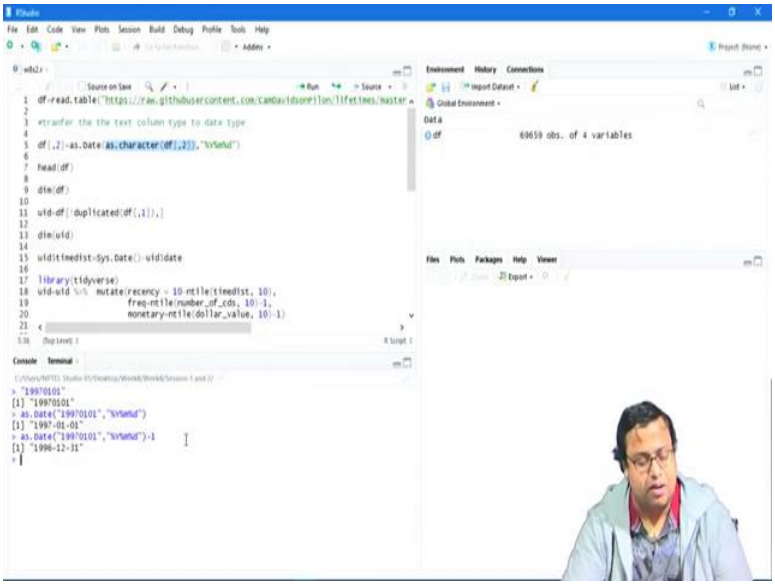
So, as dot date watch like this, let us say I am talking about this one 9997 01 01 that is the thing that I am asking. So, as dot date how will it work? as dot date, 1997 01 01. So, you I am saying that this particular text, you change it to a date form, and how will you read the text exactly whatever is written. So, for 1997 when the full year is written, you have to write percentage capital Y.

(Refer Slide Time: 4:10)



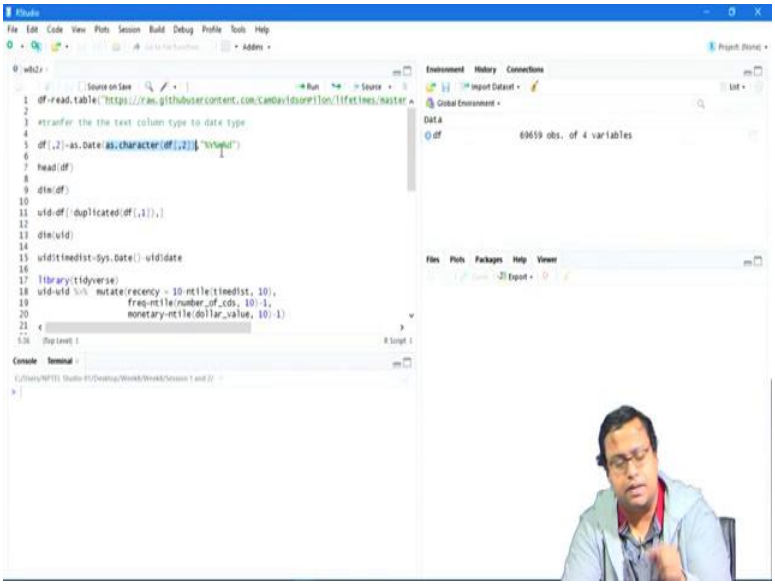
So, these are some of the details that I have probably discussed before I forgot, so I will just write percentage capital Y is for 1997, percentage small y is for only 97, so that kind of a. so, if it is y y y y then it is like that, if it is only y y, (())(4:27) font is in small y small. Percentage m is for m m, month, percentage d is for date d d and percentage B and percentage small b if I am not wrong just check. B is for full month, let us say January or let us say December, when it is written in a full form you use this and when it is written in small form so, Jan, Dec you will use small b and the rest of the will be exactly the same format in which the text is written. So, I will follow that.

(Refer Slide Time: 5:08)



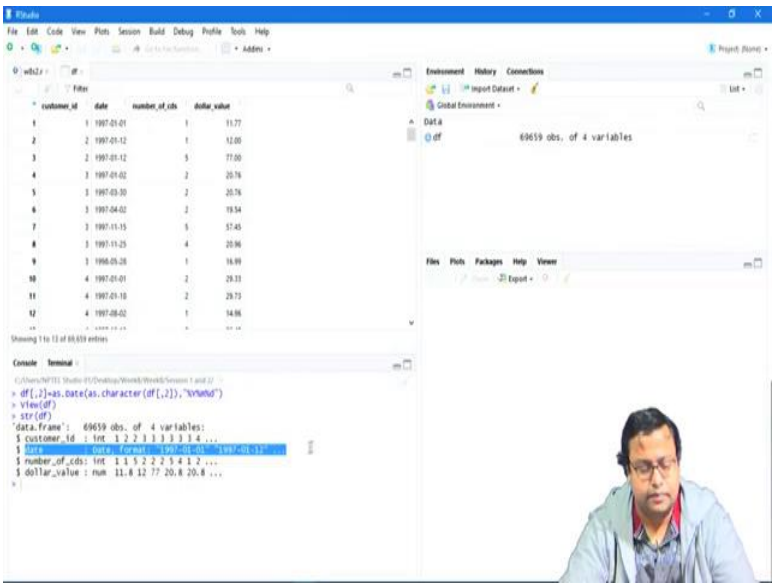
So, here what is written? Percentage. Why? Then percentage m and percentage d. So, if I just write then now it is reading it as a date, how will I know this is a date? Just do some date operation, so, I am saying this minus 1. So, that will give me, see 31st December 1996. So, that means one date before, so, now I know that this is how I have to write date. So, that is I have written it like this.

(Refer Slide Time: 5:37)



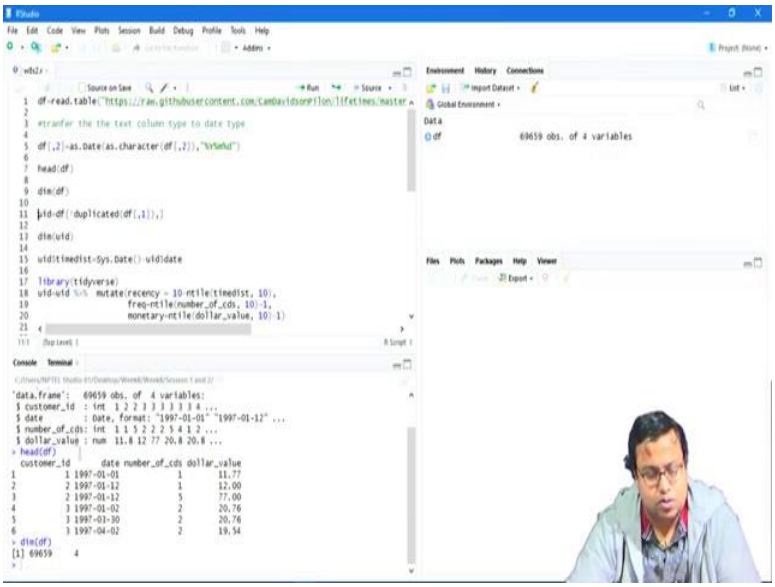
That as corrected d f comma 2 that means d f data sets, second column, this one you convert it to character and then change it to date using this format and then save it where? This date again to d f's second column. So, carefully understand once more. So, d f second column that means the date column, d f dollar date you could have written, d f dollar date you change it to character using as character, then you change use that character form as the input in the as date as dot date function and you also have the formatting which is percentage capital M, percentage y, percentage t.

(Refer Slide Time: 6:26)



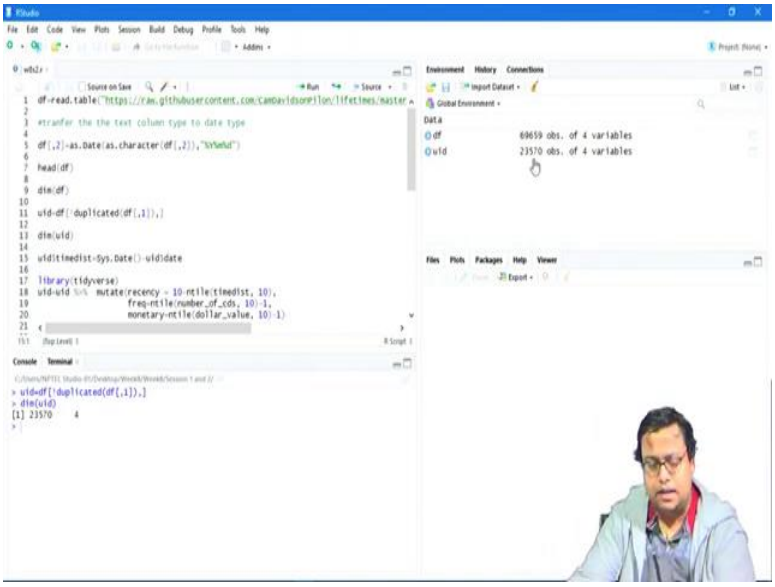
And when I run this, the whole thing has converted to dates. So, these are dates. So, if I just write str of d f now, they will find this is a date type of format, fair enough. So, if I can find out the date format now, how does my data set look like?

(Refer Slide Time: 6:48)



The data set looks like this. And what is the dimension of the data? the dimension of data looks like 69659 comma 4 that means 4 columns and this many rows. So, this is something that we have. Now, there might be certain duplicate data.

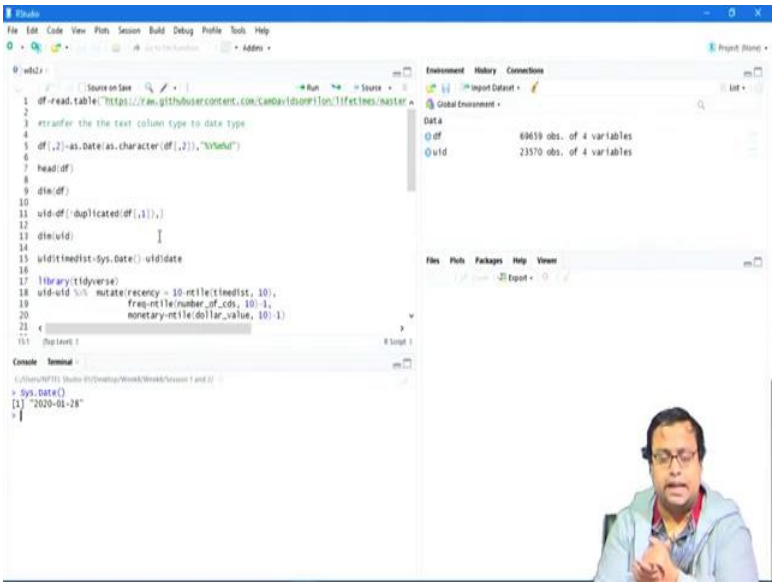
(Refer Slide Time: 7:07)



So, I will remove the duplicate data first using the duplicated function and then I will again check that dimension. Now, the dimension has stopped. So, there are lots of duplicate data.

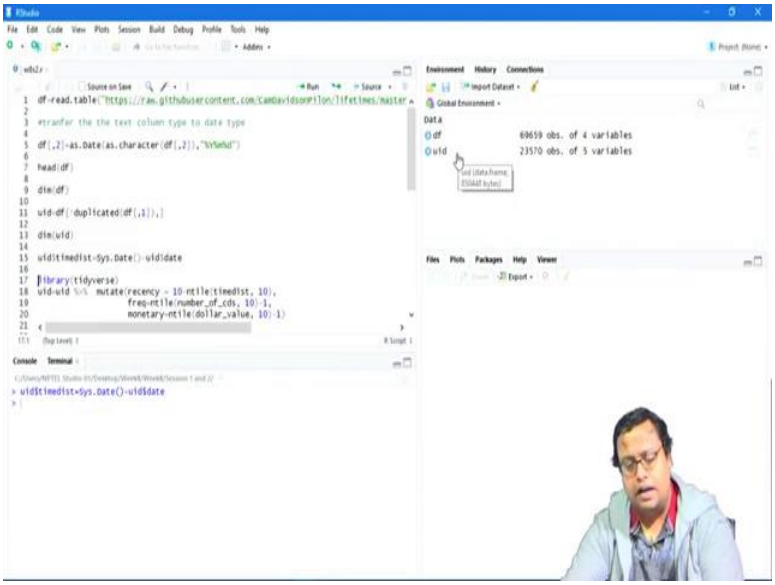
So, that is something that was a problem in this data set. So, now I have 23,570 observations of 4 variables, fair enough nothing. So, then what will I do? I have the date column, I have todays current date, how would I (())(7:36) for this todays current date, so, the today's current date can be found out by the system date.

(Refer Slide Time: 7:42)



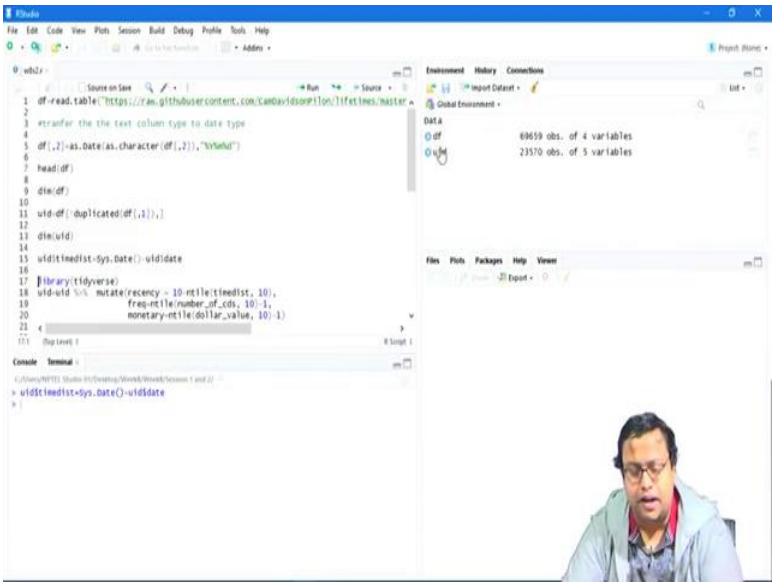
Now, here when I am actually recording this, today's date is basically, sys dot date, so, this is today's date, which is 28 January 2020 on the date of recording. So, you might get different value when you run this, when you see the values will be different obviously. So, you check your known system, whatever system date is coming that particular date should come now. Now, whatever be that case, a a certain data you have given and current today's date, that date and today's date difference is you are basically a measure, opposite of that is the measure of recency.

(Refer Slide Time: 8:21)



So, I will say that UID dollar time dist, time dist stands for time distance, is system date minus UID dollar date. So, today's system date minus that particular and you will get different distance also.

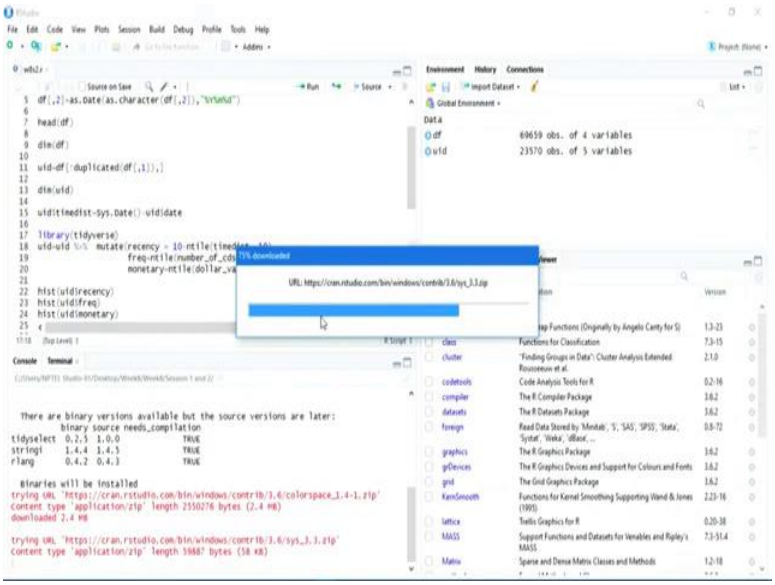
(Refer Slide Time: 8:37)



So, now I am getting 8427, 8416, your values will be different obviously, because you are running it in a different date. Now, you can say that, okay, for my this thing, I will fix it to a particular data from that date I will find out a distance, so that your dates and my dates are same does not matter, so this is something that you got us time distance.

Now, remember in this data said that time distances, higher is preferred or lower is preferred? Yes, lower is preferred and then dollar value higher is preferred or lower is preferred? Higher is preferred very good. And then number of (())(9:20) also? higher is preferred, very good. So, this is something that we have to understand. So, when I create the ranking between 9 to 0 or whatever, I have to keep that in mind, very well.

(Refer Slide Time: 9:36)



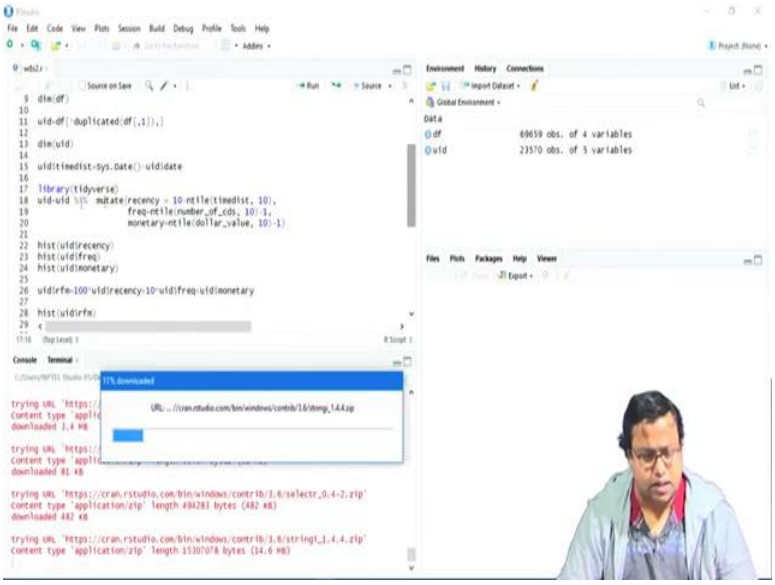
Now, what next? Next, I am calling a library called tidyverse, so tidyverse if by chance the package is not there, you have to install this. So, tidyverse, I will just call this library first here. Okay, I am not I am pretty sure that is not there in my system, so I will install it. So, if you have not installed it, you have to also install it so it will install all the supporting libraries as well, this kind of links will come this come, so, it will download certain libraries one by one, some are big, some are small it might take some time for some of these things, but otherwise it will be okay.

So, it will download lots of libraries. This tidyverse will be used because the data set is big and sometimes you might want to do some data processing with this data which is required and sudden DPLYR is a library that we have used if you remember at the early stage. So, here there is a library called tidyverse which has some some some features which can be used in creating. So, if you check the dependencies, it also has these DPLYR or DDPLY, DDPLY is the actually the function but the DPLYR or PLYR this kind of libraries will be there, they do basically, processing of columns.

So, in a huge data schedule sometimes, instead of doing row operations one by one, you do column operation, one column at a time you just change it to something or find out a mean or do certain operations, certain analysis with them. So, for those kind of piping and etc. kind of analysis, this tidyverse or DPLYR or PLYR packages are helpful.

Now I am installing it for the first time, so that is why it is taking quite a bit of time, but in your case, you should first see that what kind of libraries is required for for your analysis, and then you should install them and then you can go ahead for your analysis. So, it might take one more minute to analyze this to get get all the files ready.

(Refer Slide Time: 11:59)



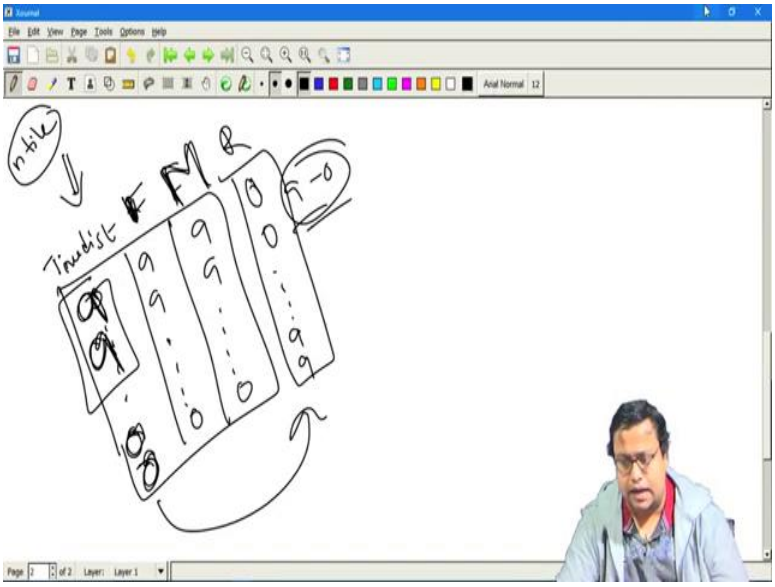
Once I a I got this ready, what I will do is, in the next step, if we just bring this I will just bring this below and then scroll this down you check what did I do in next, I use the mutate function, the same old DPLYR mutate function, to change what? To change to work on the data set called UID. I will mutate, mutate means it will change the data set and add certain columns on that.

What I am doing? I am saying that distance is equal to 10 minus n-tile of time distance comma 10. So, what is n-tile? n-tile is is the particularly the function for which we are using this tidyverse, n-tile is a function which breaks our numeric column into deciles or quartiles or n number of tiles basically. So, n-tile, so what it will do, it will actually sort it up from highest to lowest, and then break it into 10 group either either highest to lowest or lowest to highest basically and then it will highest to lowest in fact, and it will break in 10 groups, if I

have written n-tile time distance comma 10, if I have written 4, then it could have broken into 4 groups that is four percen-tiles.

Now, remember why did I write like 10 minus n-tile time distance 10. So, that is something that we have to understand.

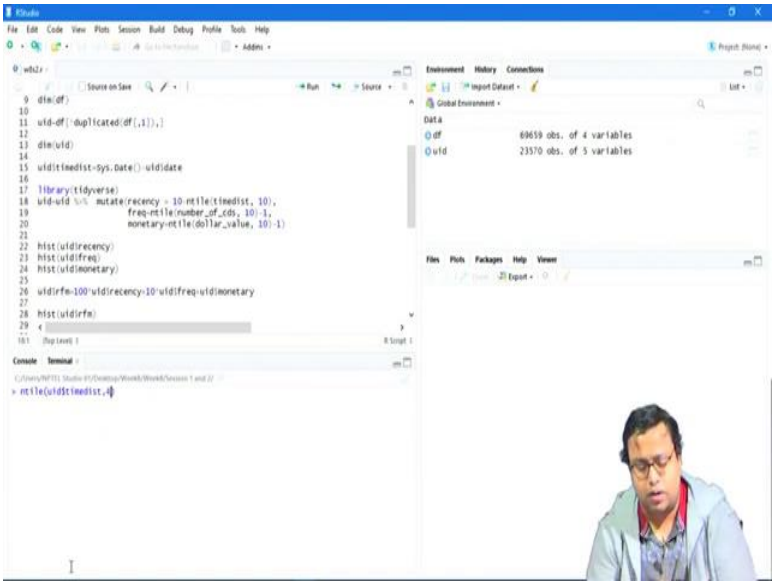
(Refer Slide Time: 13:20)



So, if I just now, let us say I had a data set, where I had time dist, I had recency and I had frequency, this is the data that I had. Now, I put n-tile in it. I put in tile in it., so this one is okay, this one the higher recency will get 9 and the lower recency will get 0. Here the higher frequencies will get 9, lower frequencies will get 0, but here in time distance, the lower frequencies will get 9 and the higher frequencies will get 0. So, sorry time distance higher will also get 9 and lower will be 0.

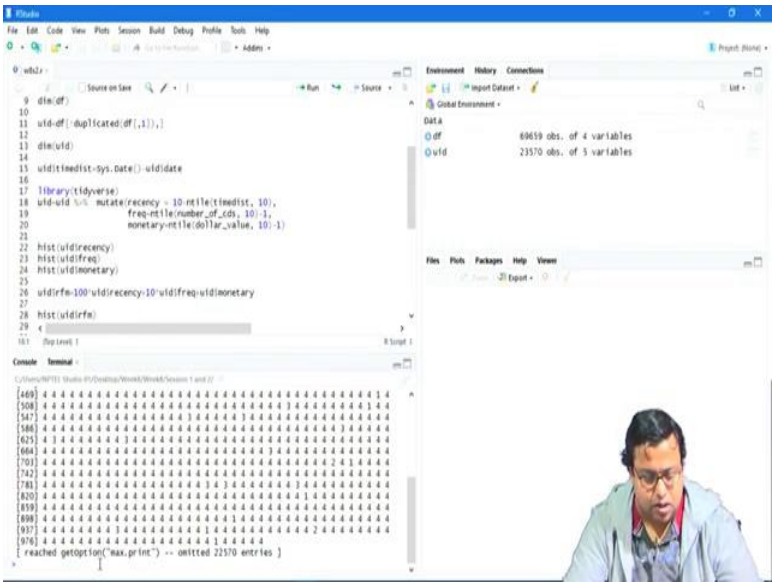
So, how to convert this thing this is 0. Now how to convert this thing this is frequency sorry and this is monetary I am very sorry. So, for frequency higher frequency will get 9 (monitor) lower frequency will get 0, for monetary higher monetary will get 9, lower manetary will get 0, for time distance higher time distance will get 9, lower time distance will get 0. Now, I have to convert it in some way recency column, but these guys will be 0 0 and these guys will be basically 9 9 or something like that and then practically not 9 and 0 will not also work. So, 0 to 9 is 10, so, that is fine, so 0 and 9. So, how will I get here, so, I might want to put 9 minus 0 or something like that. So, the moment I do 9 minus that will work for me. So, let us see how that is outcomes and based on that we will do something.

(Refer Slide Time: 15:10)



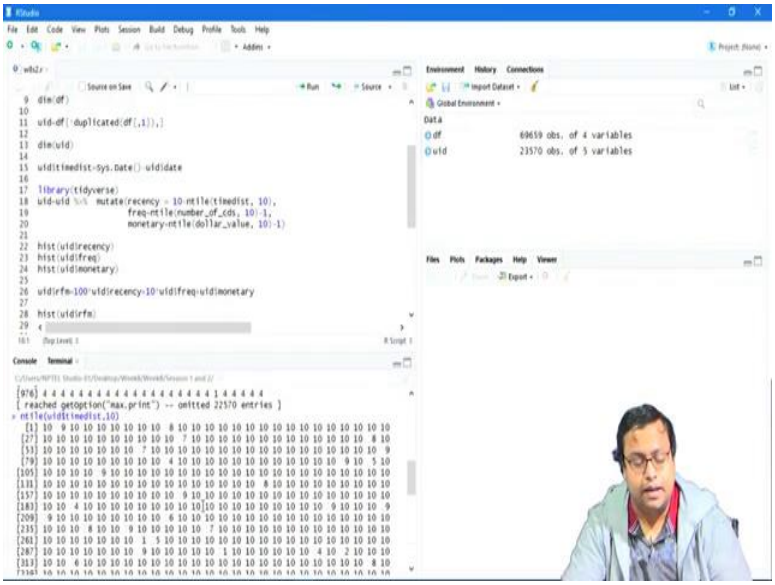
So here I have installed it I called the library. Okay, so, library has been called and then you just see n-tile if I just write n-tile n-tile let say I write time dist dollar, so d f dollar time dist, so UID dollar time dist comma 4, I will write 4, okay, so 4 categories it is breaking.

(Refer Slide Time: 15:46)



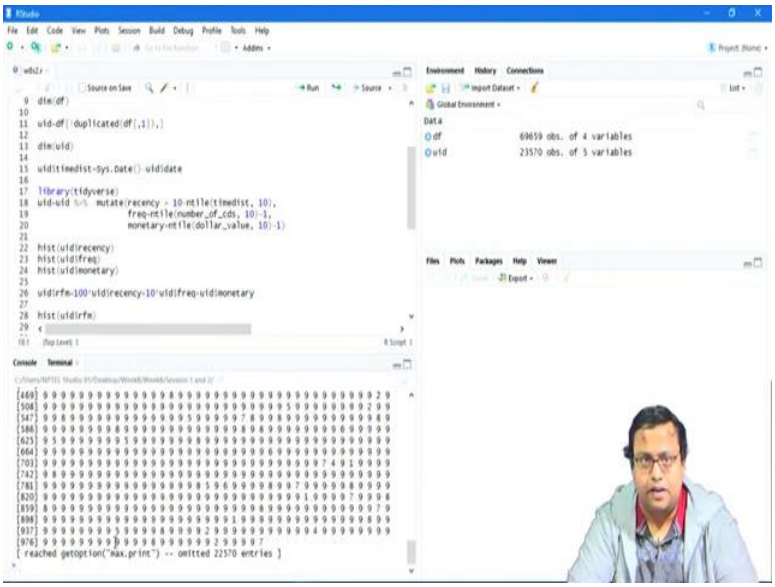
So, you say that the value that I am getting is 1, 2, 3 and 4, not actually 0 to 3.

(Refer Slide Time: 15:55)



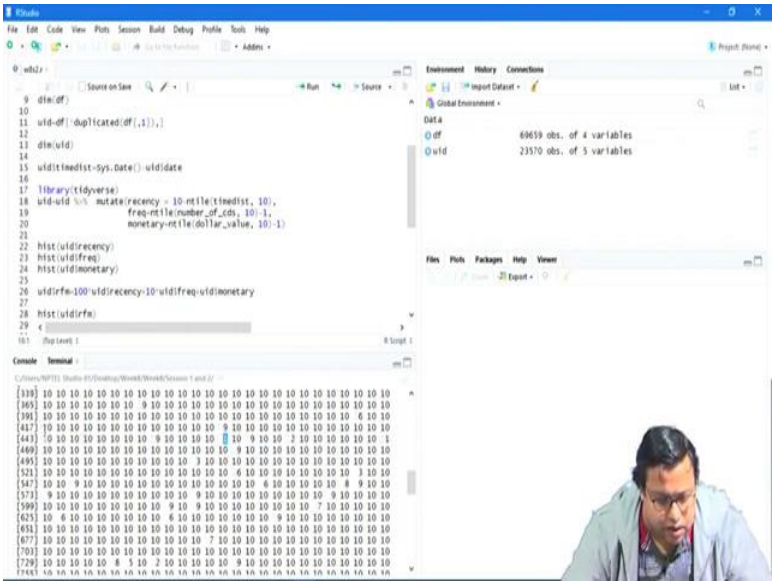
So, if I if I break it into time dist dollar 10, I will get 0 to 10 sorry 1 to 10, not 0 to 9, but I want 0 to 9, so what do I have to do?

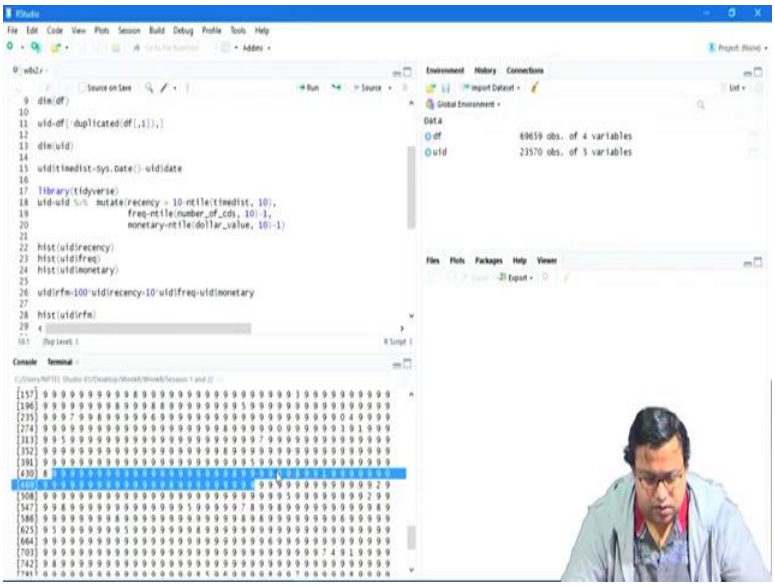
(Refer Slide Time: 16:05)



The first thing that I have to do is this minus 1 that will give me 0 to 9, fair enough. So, there see there will be some values, which you are not getting here n-values which as 0. So, because probably last values are very, very close to each other, you might get some values.

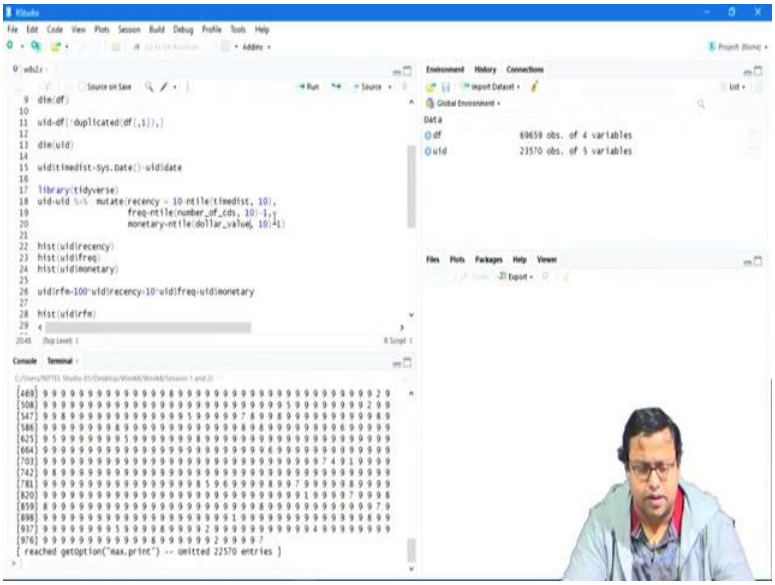
(Refer Slide Time: 16:24)





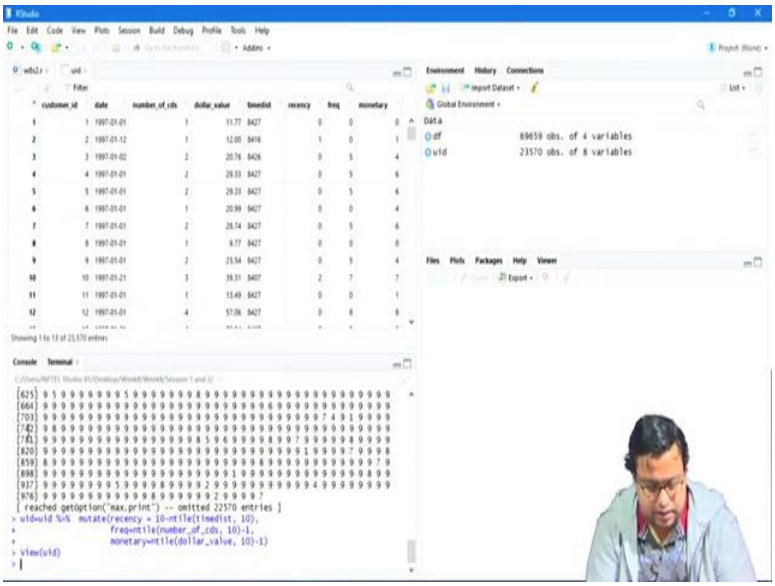
So, for this one, which is 443, and then another 10, let us say 453, I will get a 0. So, 4, yeah, so see there is a 0 here. So, some zero's will also be there, there is one 0 here, one 0 here and so on, so, some zero's will be there.

(Refer Slide Time: 16:43)



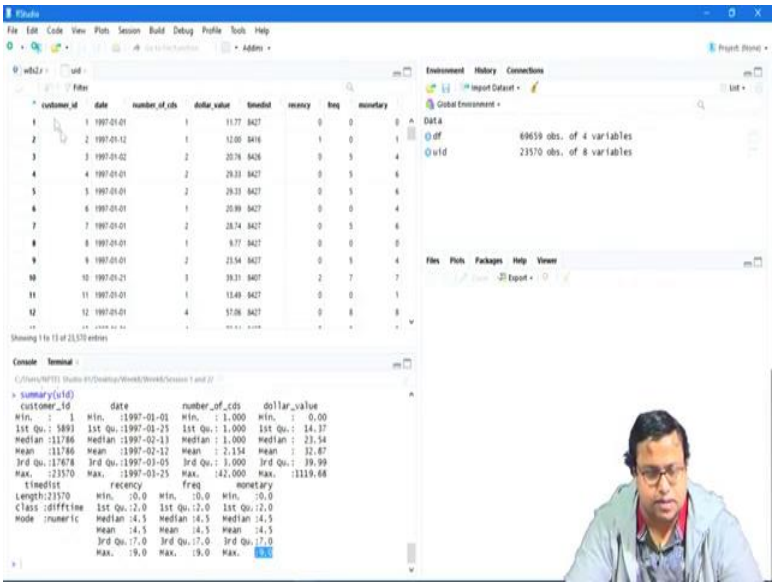
So, that is what I am doing exactly for frequency, I am doing number of cds, I am breaking it to 10 groups and then minus 1, for monetary also number of groups comma 10 minus 1, so dollar value comma 10 minus 1, minus 1 means instead of 1 to 10, 10 deciles it will give 0 to 9, 10 deciles, that is right. And here it is also gave him 1 to 10, so the moment I do 10 minus that it will it is automatically giving me 0 to 9, right? 1 to 10, 10 minus 10 becomes 0, and 10 minus 1 becomes 9. So, the moment I write 10 minus this it is giving me the highest values, the highest time distance values will get the recency score as 0 and the lowest time time distance values will get a recency score of 9, so that is what I am making sure.

(Refer Slide Time: 17:46)



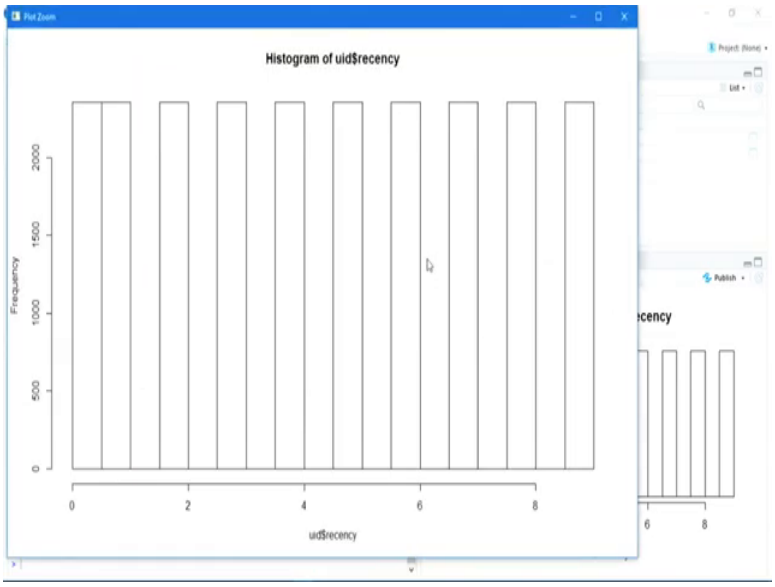
And mutate means, make these things these changes and save it in the original data set. So, if I just run this one, it will save it in the original data set, this is my original data set, frequency recency, frequency, monetary, three columns gets created. Each value will be between 0 to 1, so this is something.

(Refer Slide Time: 18:01)



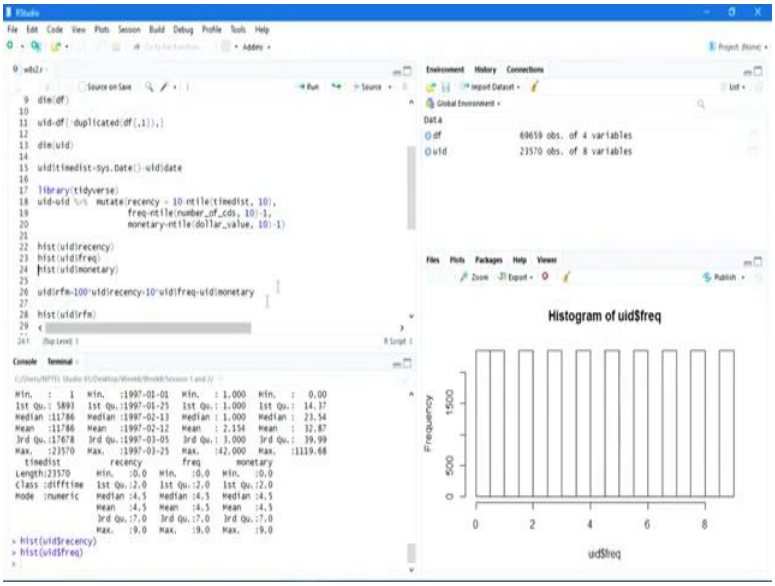
So, if I just find out the the probably the summary of my UID you will see for recency frequency monetary, the lowest value is 0, the highest value is 9. Here also the lowest value is 0, highest value is 9, here also the lowest value is 0, highest value is 9. Now, there is some problem when we do that, but anyway.

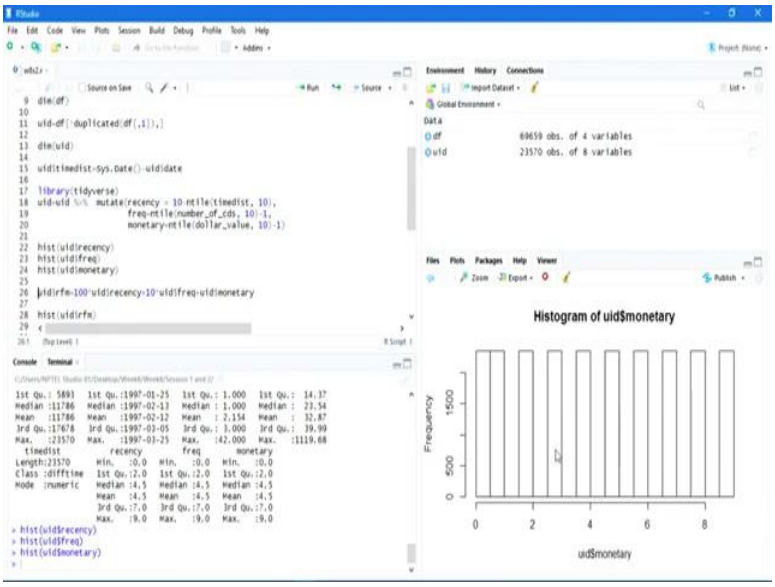
(Refer Slide Time: 18:26)



So, if I just draw the histogram of recency, you get the histogram as this, this is this is the current histogram.

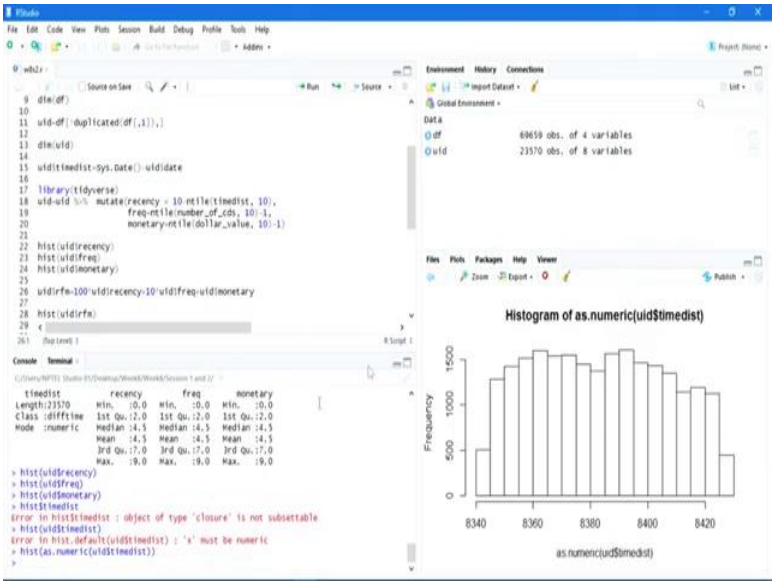
(Refer Slide Time: 18:40)





So, if I do the same thing for monetary let us say, for monetary or frequency, this is the history and monetary this is the histogram that is basically the problem that (())(18:52) is making making them into same number of people in each group, which might not be the case.

(Refer Slide Time: 18:58)



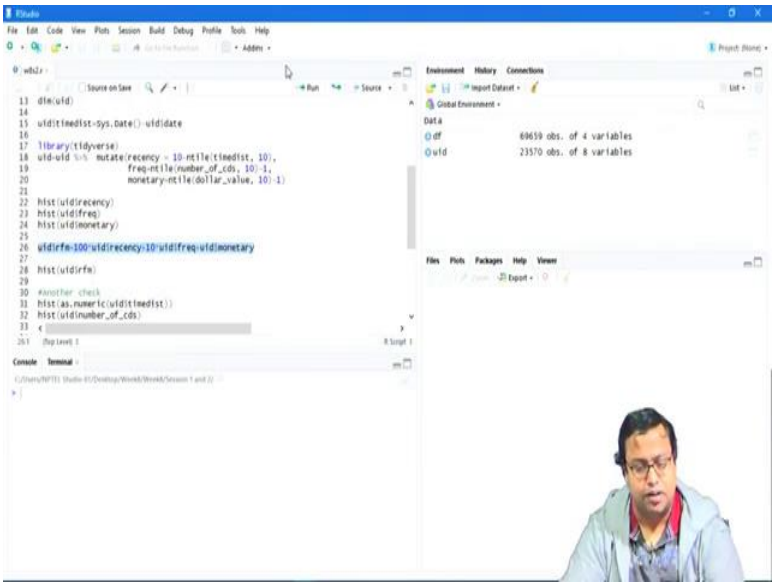
For example, let us say, hist dollar if I talk about time distance sorry, not hist or time distance, hist within bracket UID dollar time distance. Okay, time distance is not numeric, so just change it to numeric, yeah, so see the distribution looks like this, which is not exactly not exactly straight, so it is fairly distributed. In some months, there are probably more number of observations and some other months, so, this is something that we have to take keep into we should have a look on that.

(Refer Slide Time: 19:47)

On the other hand, if I just do the same thing for let us say, dollar value, dollar value that is not absolutely highly scute, it is not at all distributed properly. So, often sometimes it might not be a good idea to break them into 10 equal deciles, but that is your subject to your call

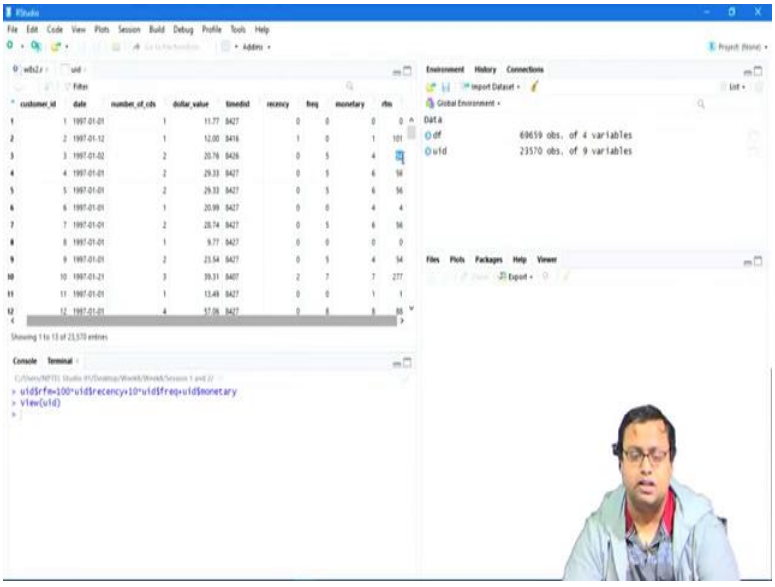
you have to decide that how much so, I if I were you I would have given that okay for this is 2, for this is 5 for this kind of data this is 100 probably the points that I am assigning and in that case the last one the (())(20:25) model based RFM analysis would have worked better, but here that will not work that much better, so, that is something that is also a very important decision model. So, if your variables are fairly distributed, you can do this kind of thing if it is not fairly distributed, you should do the other one.

(Refer Slide Time: 20:45)



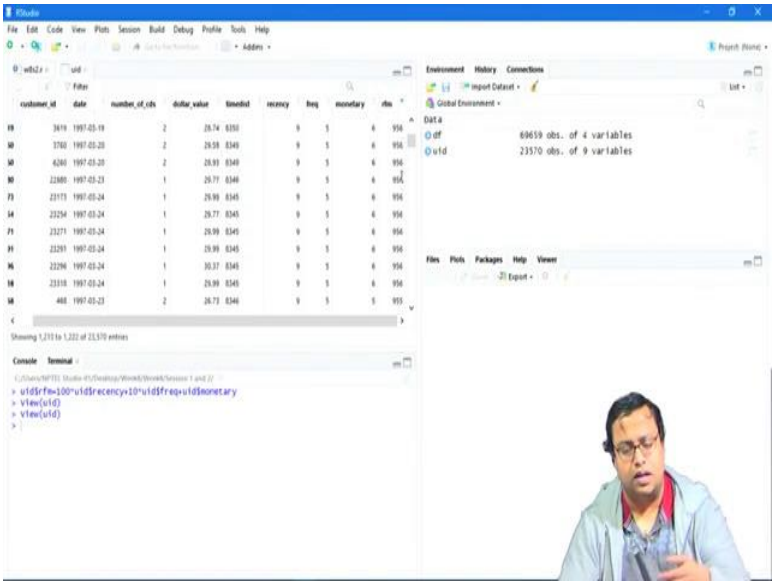
So, anyway, so mine is fairly distributed and then what do I do I create RFM score which is 100 into recency, 10 into frequency and then plus monetary, so that is the basic score of RFM that I am generating.

(Refer Slide Time: 21:00)



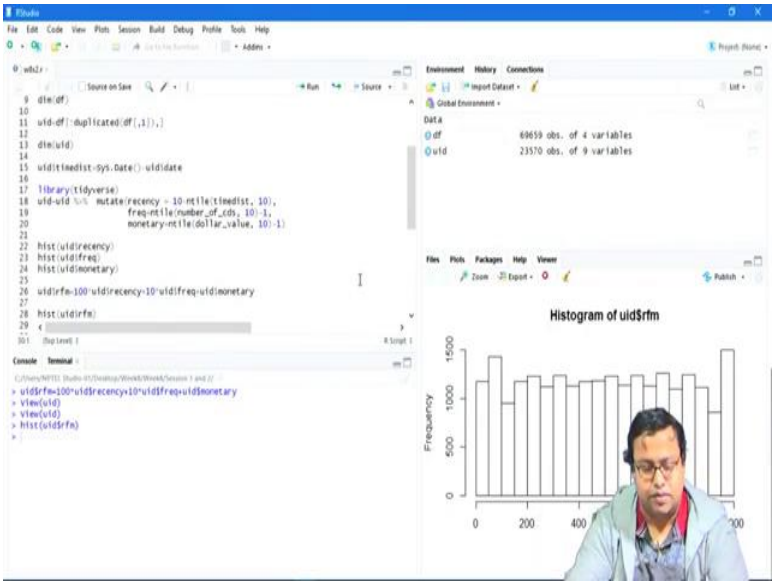
So, if I run this then I get another column in UID which is nothing but a combination of this course, so, 0 0 0 you are getting 0, 1 0 1 you are getting 101, 0 5 4 it , so the moment 54 is written, I know the recency score is 0, that is why it is coming 54, otherwise that come 154 to 54 or something. So, 54 means, recency score is 0, the (mon) frequency score is 5 and the monetary score is 4, so automatically I am getting RFM.

(Refer Slide Time: 21:37)



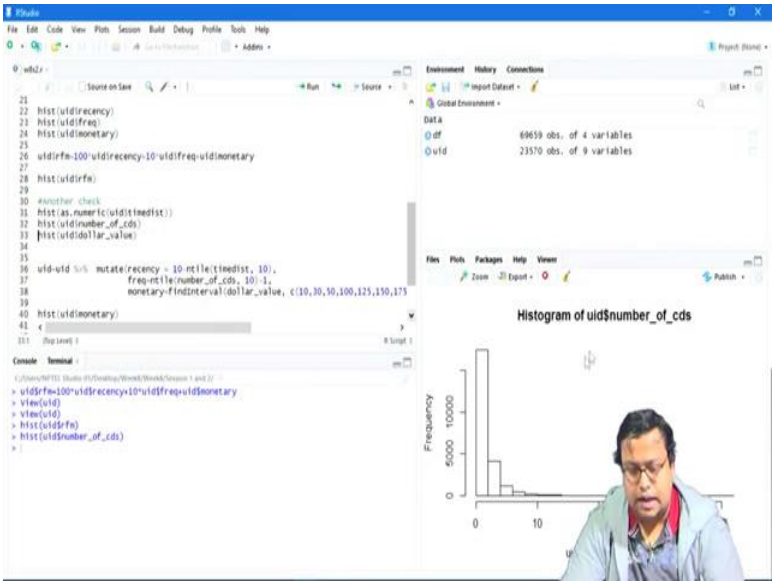
And if I want to, let us say if I want to find out the top, so these are my top guys, a 999 guys, and then comes the 998 guys, we scroll down, so quite a few 999 guys because many people might have. Okay, so I I okay after 999 there will be 953 and so on. So, there will be you can just scroll it down there will be so many people. So, first thing will be based on the last one and then the second one and the third one you will you can add on them.

(Refer Slide Time: 22:09)



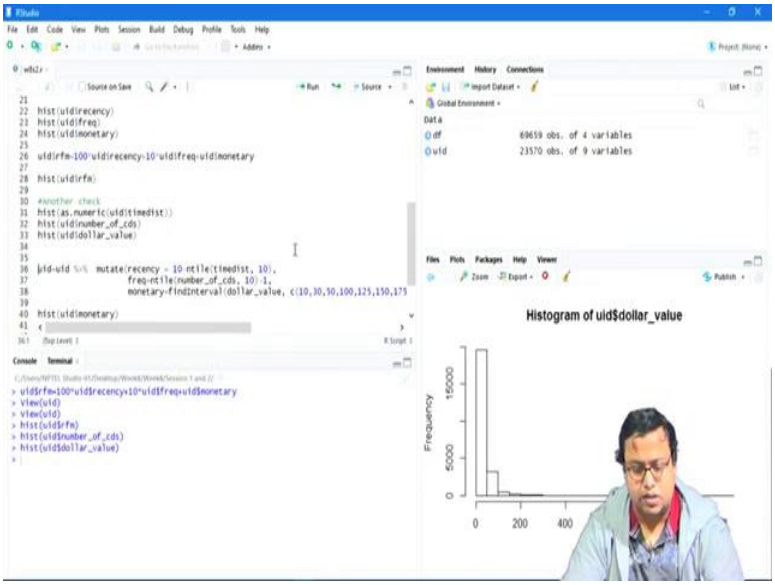
Where is the histogram? How does the histogram looks like it is similar. So, it is similar to the profitability that you can generate from them, so, which is good this is the histogram of RFM. And you should also check as I told that the RFM you should find out that how the histograms of the original variables are.

(Refer Slide Time: 22:34)



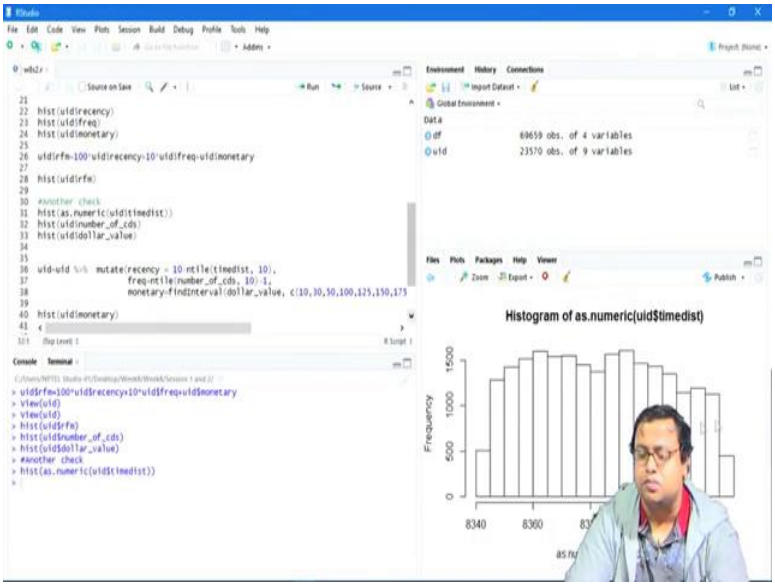
Now, the other thing that I am doing right now is that if the distributions are not very fairly distributed. For example, the histogram of cds, if I just check that is something very skewed.

(Refer Slide Time: 22:44)



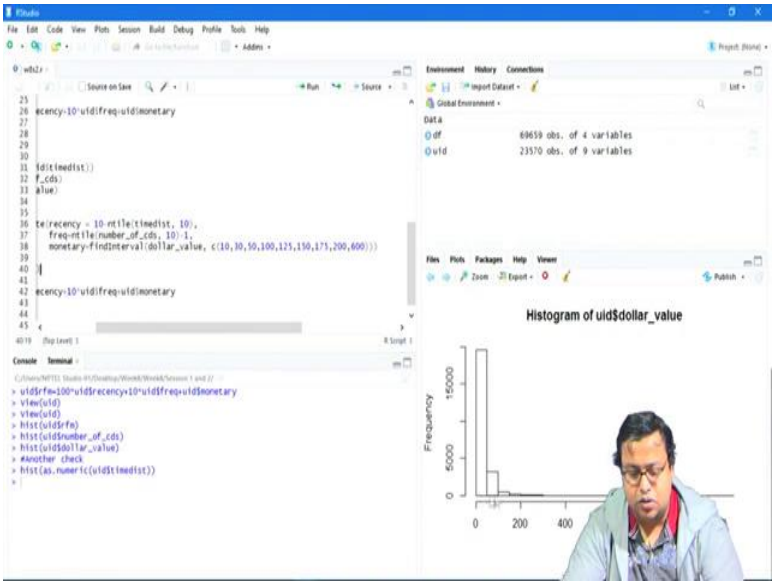
And then histogram of dollar value that is also very skewed.

(Refer Slide Time: 22:51)



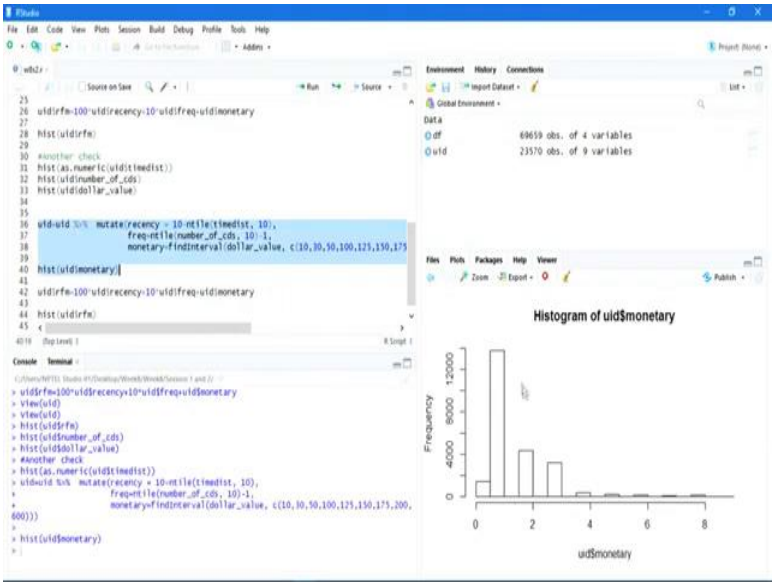
But histogram of time distance is more distributed than them. So, what I am doing here that instead of creating n-tile, what I am doing is I am fine breaking them into intervals. So, find interval dollar value, and what are the intervals? Below 10 it is 1, 10 to 30 it is 2, 30 to 50 it is 3, 50 to 100 is 4, 100 to 125 it is 5. So, I am breaking based on this picture that I got.

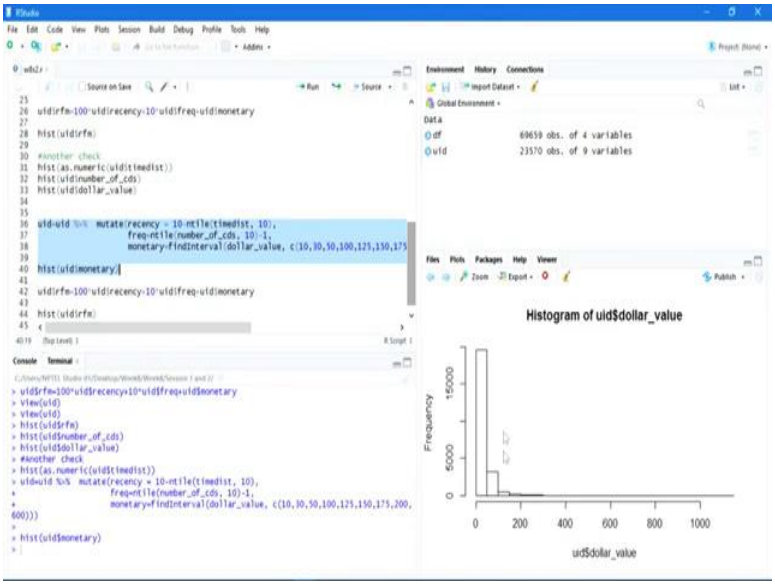
(Refer Slide Time: 23:18)



So, if you see this is my dollar value, 0 to 50, all this huge line is 1, sorry 0 to 10 the huge line is 1, 10 to 50 is 2. So, I am breaking here, these guys this big line and breaking into three groups, 1, 2 and 3. So, up to 50 there are serial number 1, 0 to 10, serial number 2, 10 to 30, and serial number 3, 30 to 50, so I am breaking this thing in three groups. And then breaking this one into two groups, 100 to 50 to 100, 100 to 125, 125 to 150, same groups again and so on. So, when I do that, I get a better distribution of monetary.

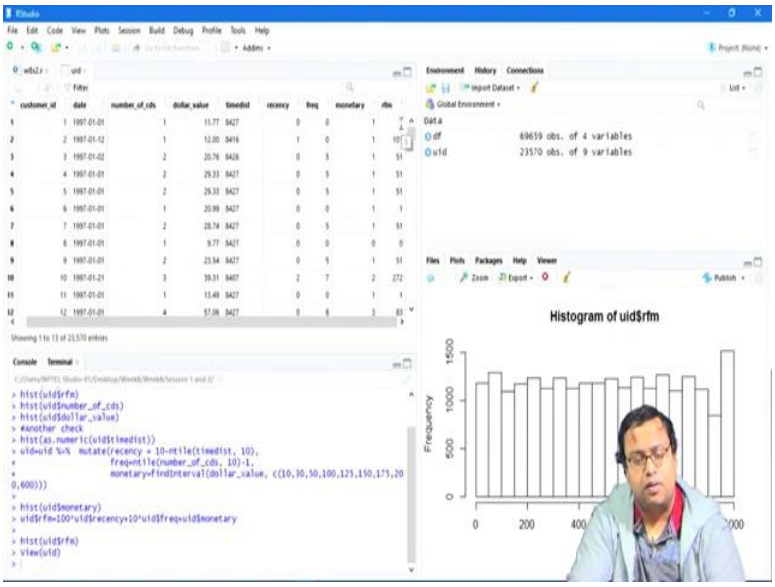
(Refer Slide Time: 23:55)





So, if I just run these now, you will see that monetary is small better distributed, this is a monetary score that I created now, and this monetary score is more distributed like the previous one. So, this distribution and this distribution, more or less looks similar, at least at least in terms of the shape they are similar. So, now you can use that kind of a score to create your RFM analysis, which is histogram.

(Refer Slide Time: 24:22)



And the RFM score, the new RFM scores are this one the similar kind of RFM that you got before similar RFM score, but remember this guy 51 before now I am getting 54 before now getting 51, this guy was 0 before not getting 1. So, I am changing this course somewhere other and I am getting certain values here. So, this is how we can also do RFM analysis which is more and more in line with the original data set that we have collected.

So, that is all about it is very easy to implement that is one of the major reasons it is used. Another major reason of using it is probably that you do not need a lot of and you can do it in Excel as well, because the data set is big, (())(25:07) because there is not unique data some data was duplicated and we are doing it in but in maximum cases it is not done in R or Python or something like that.

In maximum case the RFM analysis done by Excel only, so, they or broadly SQL if it is a database that they have certain values, they have certain certain macros and etc. already written, which will convert this time, date to time distance, time distance to monetary values, time distance to recency values and some other to monetary values these kind of changes are already written and the RFM score algorithm is already written and this is a very very, very I would say commonly used method in (con) in the context of segmentation through purchasing behaviour.

So, that is all for RFM analysis, we will continue with Market Basket Analysis in the next video and thank you very much I will come back with next video.