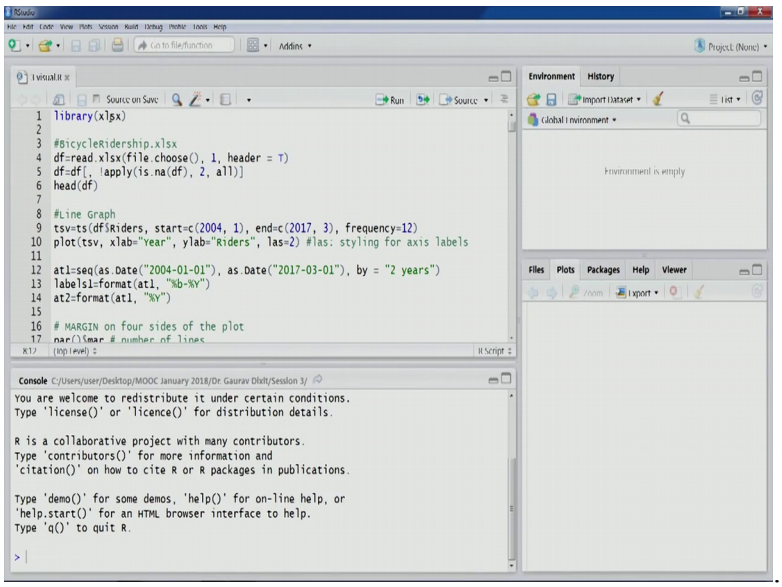
Business Analytics & Data Mining Modeling Using R

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

Lecture – 09
Visualization Techniques- Part III Heatmaps

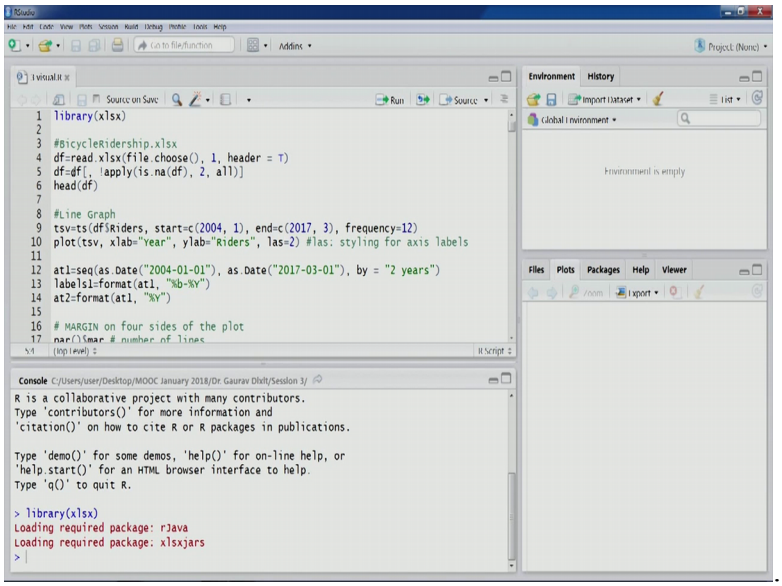
Welcome to the course business analytics and data mining modelling using R this is our 5th lecture and we were covering restarted visualization techniques in the previous lecture. So, let us start, we stopped at you know R studio where we were doing some of the examples. So, let us go back and complete some of them and then we will come back and start our discussion on our next particular plot that is on that is heat maps. So, let us go back.

(Refer Slide Time: 00:55)



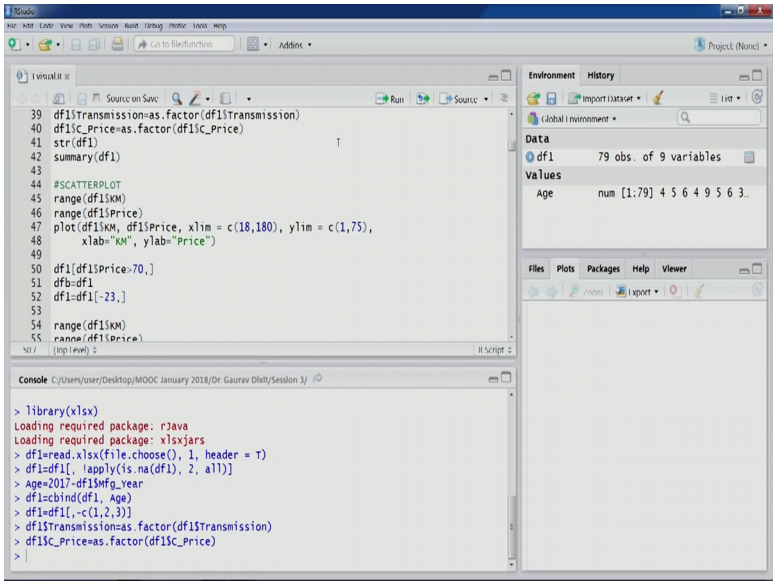
So, again we will have to do some of the loadings and importing data set we will have to reload the library and everything. So, let us load this particular library xlsx. So, once it is loaded.

(Refer Slide Time: 01:07)



So, mainly in this particular lecture we would be using used cars xlsx file. So, let us import that particular data set see here. So, let us import it.

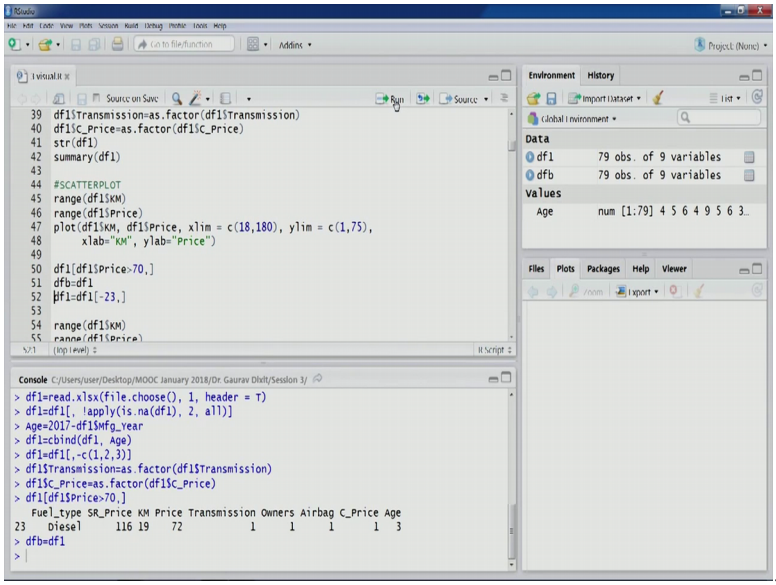
(Refer Slide Time: 01:40)



So, we will rerun the same lines that we did in the last in the previous lecture. So, we can see that there are 79 observation 11 variables in the environment section and then let us re create this age variable has discussed in the previous lecture let us append it to the data frame and let us subset the data frame right and let us also convert these variables. Now you might remember in the last session we had eliminated one of the observation which

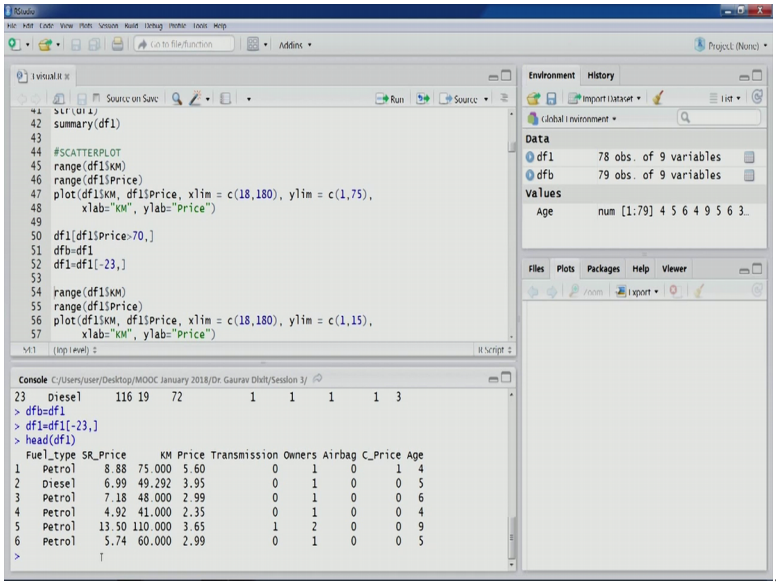
was which was actually outlier. So, let us perform the same operation again. So, this was the observation let us take back up and then again eliminate the observation.

(Refer Slide Time: 02:28)



Now, we want to have a look at the data set this is the data set you can see now d f1 78 observation and 9 variable.

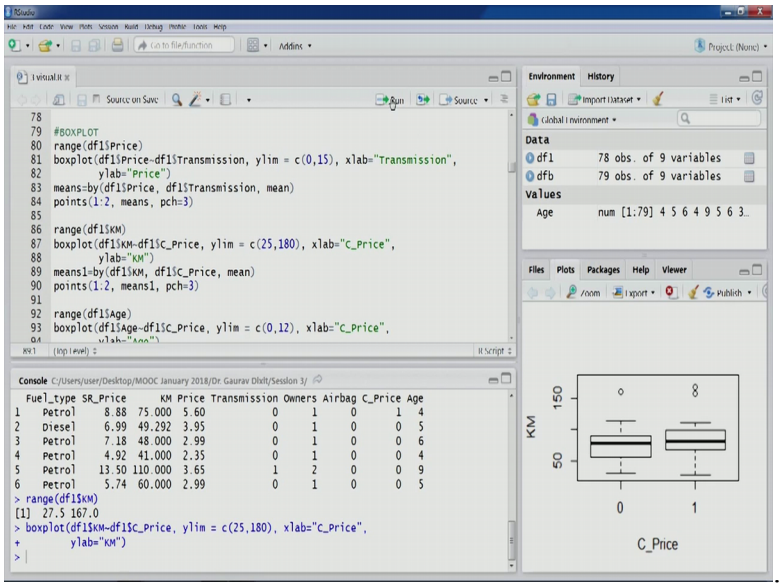
(Refer Slide Time: 02:40)



So, let us go back to the point where we stopped in the previous lecture. So, we were going through some of the examples of box plots. So, I think what I remember is we

completed a one box plot. So, let us discuss another one, this particular box plot is between kilometre and the categorical price. So, let us look at the range of kilometre because we would have to specify that in the y limit because this particular variable is going to be on the y axis. So, let us do that you can see the range and you can see the y limit that we have specified in this particular line is actually covering this range for kilometre. So, now, let us create the box plot.

(Refer Slide Time: 30:39)



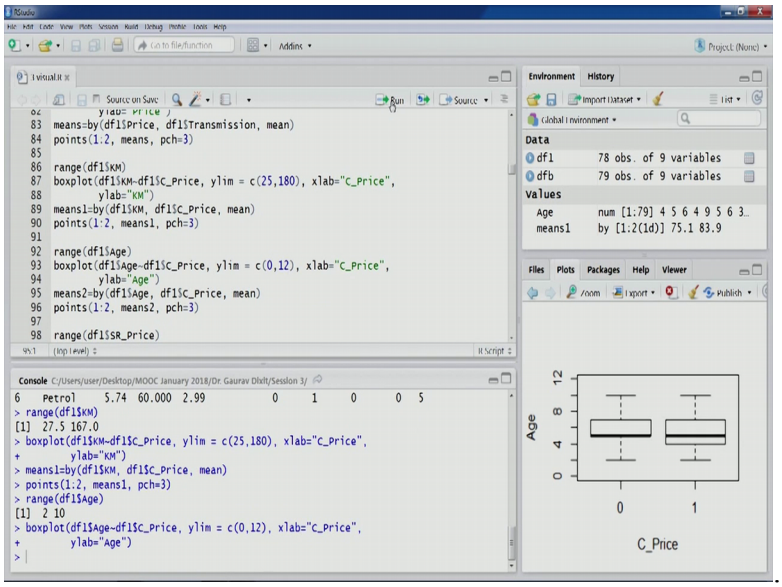
You can see this like the last like the last session last lecture this is the box plot that we have. So, other understanding of the box plot remains same like the last session if you want to display mean as well and that can also be done, but for that we will have to compute the mean first. So, this is the code that we discussed in the previous section as well. So, means are computed you can see means variable have been a created has been created means one variable has been created and 2 values are there.

Now, let us plot these 2 points and you can see the plot now here also if you want to compare how the kilometre variable km variable is actually distributed for 2 groups 0 and 1 group 0 and group 1 you can see. So, in comparison to our the previous example that we saw there is the both these boxes are closer to each other, but group 0 is slightly on the lower side the distribution is slightly on the lower side and there is some difference between box 0 and box 1. So, this kind of box plots as we discussed can also help us in understand the difference between groups and we can also help you know

decide whether we need to include any interaction variable because of you know if we see a significant difference in the distribution of data in 2 groups. So, we will have more discussion on interaction and on other related concepts in coming lectures.

So, let us plot another one. So, this one is between age and the categorical price that we have these 2 variables. So, let us look at the range because again this is going to be plotted on the y axis we can see the range is 2 to 10.

(Refer Slide Time: 05:50)

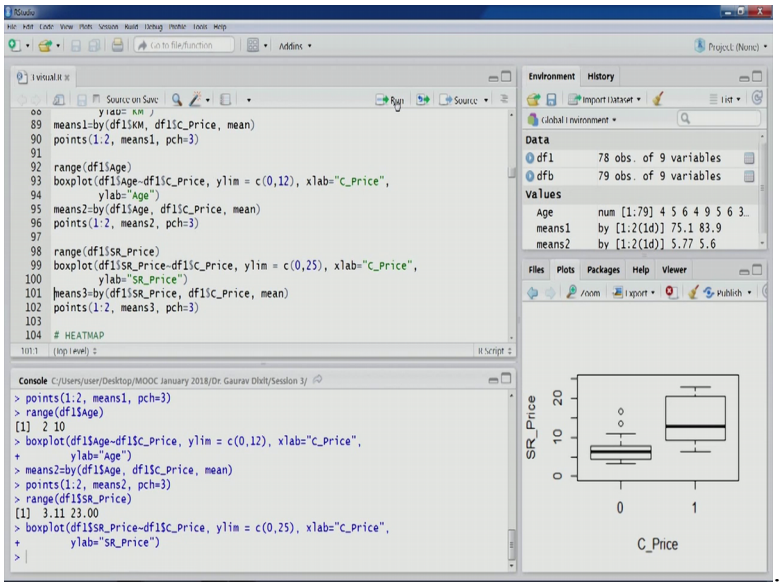


So, you can see the limit is also specified appropriately and other things remaining similar. So, can see these 2 plots this one plot and we can also calculate means and plot these 2 means for these 2 box plots let us look at the graphic.

Now, in this case this is in this case you would see the boxes are you know very you know in the same range those these 2 boxes they are in the same range, but you would see the median is coinciding with the first quartile in the box 0 and the box 1 it is separated you can also see the which are means are also looking at the same value. So, therefore, very close very little difference between these 2 distribution for these 2 groups and 0 with respect to age.

Now let us do another example this is bit being showroom price and categorical price. So, let us go through this range you can see appropriately specified in this particular line box plot code.

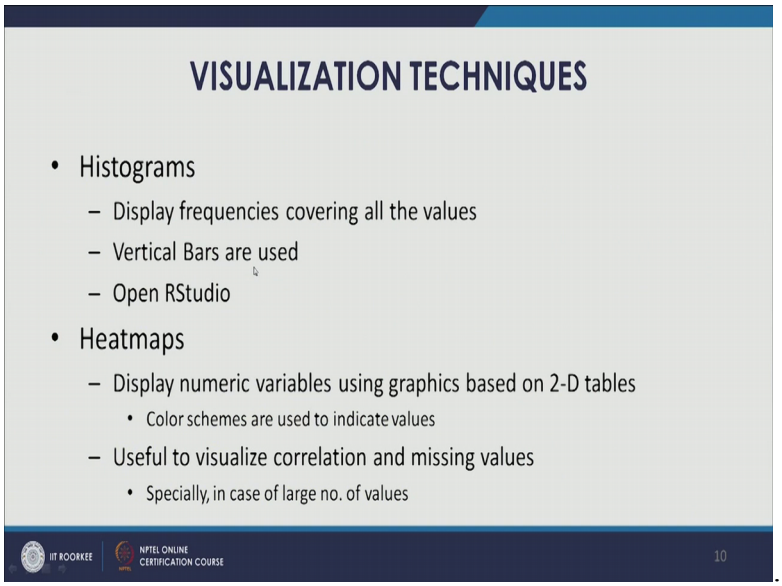
(Refer Slide Time: 07:02)



And we can plot the box plots and then means let us plot them now let us look at this particular graph this looks much interesting you can see these 2 boxes a much bigger difference you can see for group 0 you would see that price is showroom price was you know showroom price distribution is on the a lower side and for group 1 the showroom price is on the higher side. So, that is nothing unusual this is actually because of the way categorical price has been created the showroom prices actually following that you know indicating the same difference at because both are related to pricing of the cars. So, therefore, this difference this separation is very clearly visible or depicted in the box plot because both these variables are related to price.

Now, let us come back to another plot let us come back to our slides.

(Refer Slide Time: 08:09)



So, heat maps is the our next discussion, heat maps again they are another you know they can be combined with the basic plots and distribution plots they display numeric variables using some graphics based on 2 D tables will see how that is possible. Then we can also use some of the colour schemes that could be use to indicate values. So, different colours and different shades of colours could actually be used to indicate a different range of value let say if a value lying between 0 and 0.1.

So, one particular shade could be used if the value is lying between 0.1 to 0.2 a darker shade could be used if the value is lying between 0.3 and 0.4 a little bit more darker colour could be used. So, therefore, in that in that fashion a colour ordering you know we can use the colour scheme and that can the intensity of the shade that can help us indicate whether the value is on the higher side all or on the lower side. So, 2 D tables any kind of data that we can have that we can actually have in 2 D you know table format a table format. So, therefore, that can be displayed using heat maps and the colour coding can actually help us in understanding the data and some relevant insights developing some relevant insight.

Now, as we talked about in the previous lecture as well that our human brains are capable of doing you know much higher degree of visual processing. So, therefore, heat maps can really be helpful especially when we are dealing with large amount of data. So, we have large number of value values it might be difficult for us to find out different things

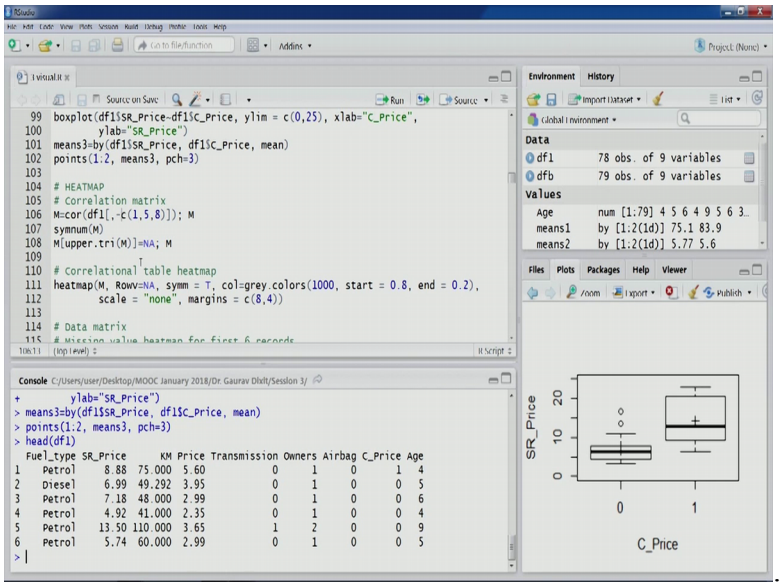
different insights about the data therefore, heat map this colour things can actually help us in building our visual perception, now those visual perception can be carry forward for subsequent analysis and used for then later on used for formal analysis.

Now as you can see in the slide second point about heat maps is useful to visualize correlation and missing value. So, as we talked about because different colour shades are going to be used therefore, in the correlation metrics if there is a higher value if there is a high degree of correlation between 2 particular numerical variables. So, that can be shown with the darker shade and if there is low degree of correlation, low value of correlation coefficient then a lighter state could actually be selected. So, therefore, the different shades in density of these colour shades that can all actually help us in finding out which variables are highly correlated and which variables are have or having low correlation values right.

Similarly, missing values can also be spotted. So, if we have the data generally as we talked about in the starting lectures that generally data is displayed in metrics format or in the tabular format. So, therefore, that data can actually be displayed and if there are any missing values. So, they can be represented using you know whiter white colour and the cells where the values are there can be represented as the darker colour or the black colour. So, it would be easier to specify missing value we can also. So, heat maps can also be used to help us understand the missing ness in a particular data set if there are too many missing values right that can also be spotted if there are duplicate rows and columns probably because of the colour shades if the colour shade is a very similar for 2 particular rows or 2 particular columns or multiple rows or multiple columns. So, we can again do a manual check to find out whether the values are whether it is a duplicate row or column. So, heat maps can help us finding these problems.

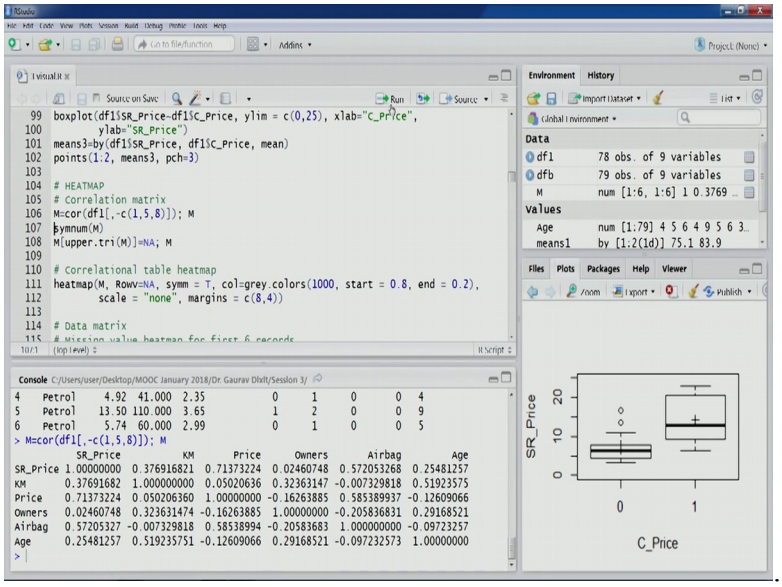
So, let us go back to R studio, first will cover the correlation matrix. So, heat map can be used for you know creating a correlation table heat map. So, first we need to compute correlation. So, in this case you would see that in the data frame that we have data set that we have let us have a relook.

(Refer Slide Time: 13:08)



So, you can see that column 1 5 and 8, 1 and then 5 and then 8 having left out in the correlation function reason being obvious that these are factor variable or categorical variables. So, that the correlation values it requires numerical variables. So, let us compute the correlation values among the remaining numerical variables this.

(Refer Slide Time: 13:36)

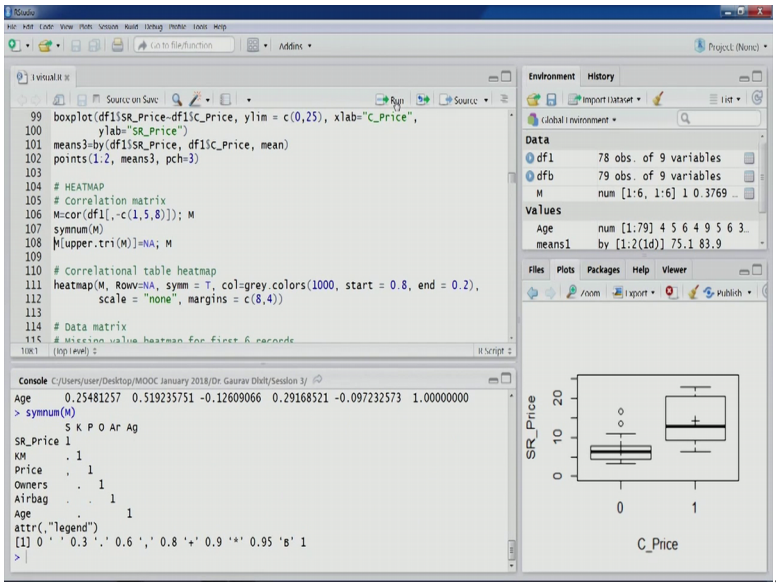


You can see a correlation matrix has been displayed there this metrics is symmetrical. So, upper half is symmetrical to the lower half you know this diagonal is there all the values are 1. So, this value 1 is between the same variables. So, the variable is going to be

hundred percent correlated with itself therefore, these values are one other values are showing the particular correlation coefficient.

Now we can have a different kind of a table for the same data.

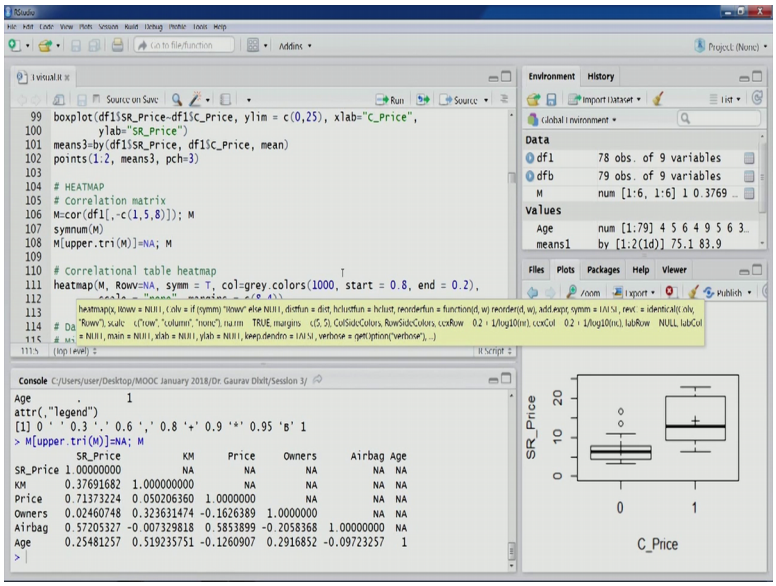
(Refer Slide Time: 14:13)



This is the function same num. So, you can see different kind of depiction here. So, you can see the variable names here in the rows and in the column also variable names are there. So, the 1 representing the a 100 percent correlation and you would see the some notations are given in the at the bottom of this particular output, see this single code is used for values lying between 0 and 0.3 dot is being used for values are lying between 0.3 and 0.6 comma is being used for values lying between 0.6 and 0.8 plus is being used for values lying between 0.8 and 0.9.

Hash trick is being used for values between 0.9 and 0.9 5 and the b is being used for values between 0.9 5 and 1. So, you can see there is 1 we can see 1 comma here and then several dots. So, this comma value might be somewhere between 0.6 and 0.8 and dots could be somewhere between 0.3 and 0.6. So, this is quite similar to what we were talking about the heat maps heat map will so the same thing using colours. So, in this case this particular function sym num is displaying different different value using different symbols. So, now, because the there is symmetry in the matrix.

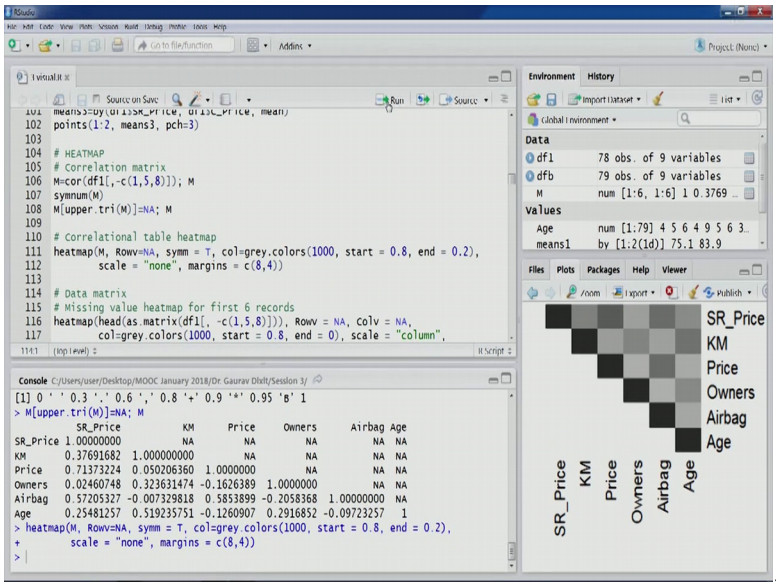
(Refer Slide Time: 15:51)



So, let us get rid of the one of the triangle. So, in this case we just want to keep the lower triangle. So, upper triangular values have been assigned as NA, now let us create the correlation table heat map. So, this is the code you can see the first argument in this particular function heat map is the metrics itself then there are some other arguments symmetry is in this case is 2 colour we have specified as grey colour. So, we want to have now we want to use the grey pallet we will understand more about colour schemes in R later in this lecture so, this particular function grey dot colours can be used to create a number of grey a number of grey shades.

Now, for example, we want to create 1000 shades starting from value ranging from 0.8 and ending at 0.2. So, the values can be start between 0 to 1 range, but we are is a restricting ourselves to 0.8 to 0.2 scales we are not scaling the data that we have because this is already correlation value. So, they are already standardised margins we have specified.

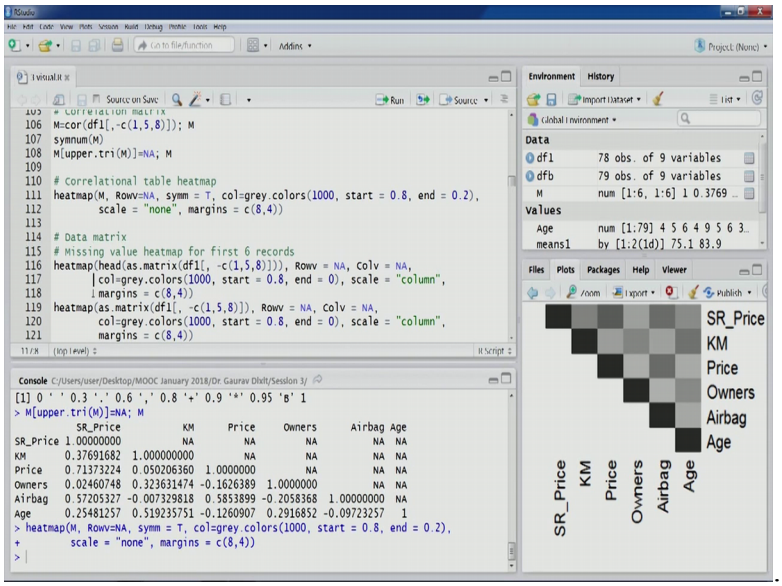
(Refer Slide Time: 17:18)



So, let us execute this particular code you can see the output this is the graphic that we have. So, in this graphic you can clearly see that a diagonal values they are in the darker shade because there is the each variable is going to be 100 percent correlated to itself therefore, these value are being repainted by the darker shade having perfect correlation other values you would see slightly a lighter a shades of grey have been used, but the intensity of the shade in indicate higher value and higher intensity indicates higher value and lower indicates lower intensity of the shade in indicates lower value. So, these a whitish kind of light grey kind of you know rectangles are squares shown here the correlation values for the corresponding variables pair of variables are in the lower side and the slightly darker for example, this particular one we mean price and SR price that is this we can understand that showroom price is going to be highly correlated with the price of the car.

So, therefore, the correlation is value correlation value is going to be on higher side and similarly the colour is on the you know this is higher intensity in higher shade of grey colour. So, this can actually help us visualizing in terms of finding out which particular periods of variable are highly correlated. So, therefore, we can say price and SR price we can also say SR price and airbag you can also say kilometres and age right similarly price and airbag. So, these a particular set of variables they seem to be highly correlated with SR price and price being very highly correlated.

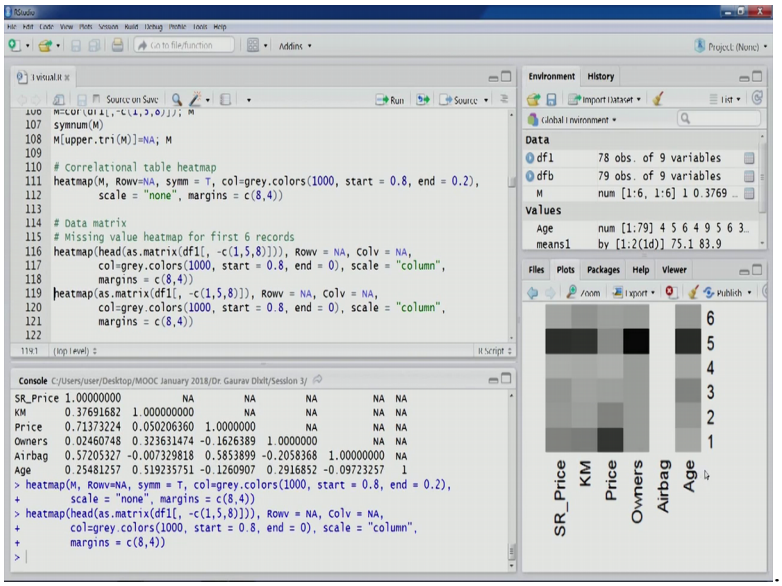
(Refer Slide Time: 19:18)



Similarly the data metrics or missing value heat map can be depicted using heat map function. So, what we are trying to do here is. So, generally missing value heat map is actually if the value is present it is generally shown in the darker shade, if the value is absent then it is shown in you know lighter shade. So, that generally you know the colour that is used is black for the value being present and white for value being absent so, but in this case we are not doing that because the data set that we have all the values are present. So, instead of that what we are trying to do here is to just show you to give you the feel of heat map in the data metrics case missing value heat map case for first 6 records and then later on for all the records we are depicting different shades of grey colour for different actual values. So, depending on the value different colour shade would actually be shown.

So, for first 6 records we are going to run this code you can see head is the function that has been used for to actually subset the a data frame for first 6 records you can see grey colours the scheme is this slightly different scale, now we want to standardise this scale because column wise. So, column wise scaling is going to be done margins are also mentioned there. So, let us run this code.

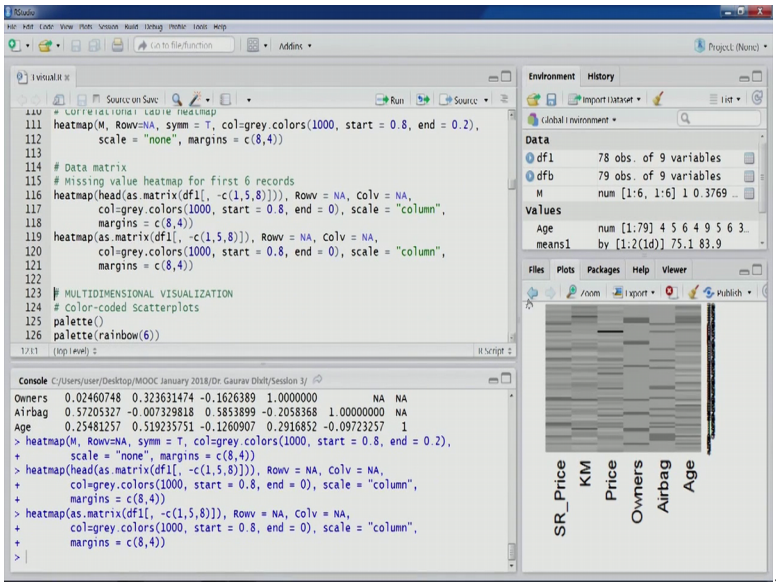
(Refer Slide Time: 20:46)



You can see for first 6 observations we can see for different columns different variables and the values. So, for example, you can see the airbag it looks white and the colour is predominantly white you can see most of the values in airbag they were actually 0. So, therefore, this whiter shade of grey colour has been used similarly you can see this 5th row this is mainly in the, you know many cells, many squares in this many cells in this particular row they are in the darker shade. So, higher values are there in this particular row.

Similarly, you can see that KM column this is slightly on the darker side so; that means, higher values are there in the km column per KM variable. Similarly we can create the heat map for all the rows for all the records.

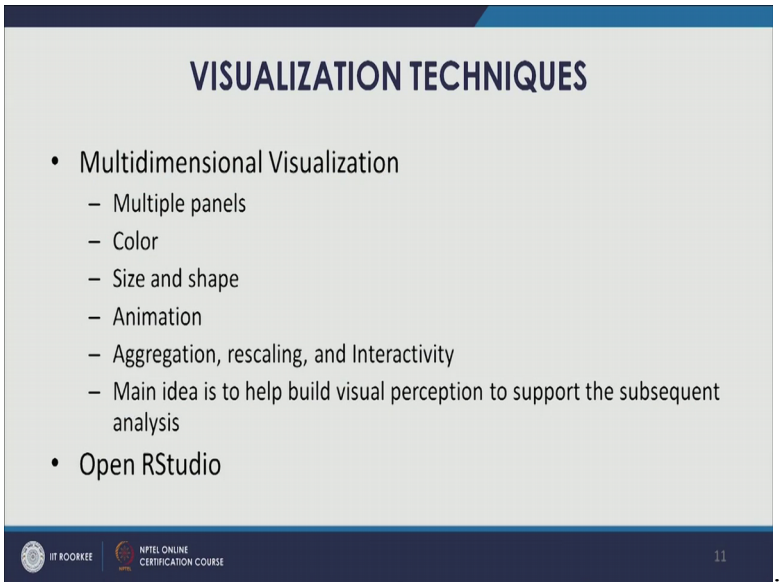
(Refer Slide Time: 21:49)



So, this is the heat map you can see now this is the number in indexing for the rows 1 to 79 because we had 79 observations you can see that. So, depending on the values itself the shade has been selected had it been for missing value actual missing value heat map. So, we would had we would see either black or white, white in the places where the value is absent and the black in the places where the value is present. So, we want to do a similar thing for in the for our data set whole table is going to look black because there is no missing value now that brings us to our next discussion.

So, let us come back to our slides, next discussion is on multidimensional visualization. So, most of the most of the visualization techniques or plots that we talked about they were mainly 2D 2 dimensional. Now, we can also have some features which can actually add to the 2D plots that we have a gone through till now and in a way they would be multidimensional because some of the features that are mentioned in this particular slide.

(Refer Slide Time: 23:21)



These features are going to give that multidimensional feel. So, our visual perception can be multidimensional you know using these features on 2D plots. So, you can see multiple panels if we can have multiple panels. So, if you just 1 scatter for example, if you use just 1 scatter plot only 2 variables can be selected, only 2 variables can be visualized, but if we have multiple panels right. So, we can have you know pair wise scatter plots for many variables and at in one go we can look at different variables and the relationships and the information overlap and many other things.

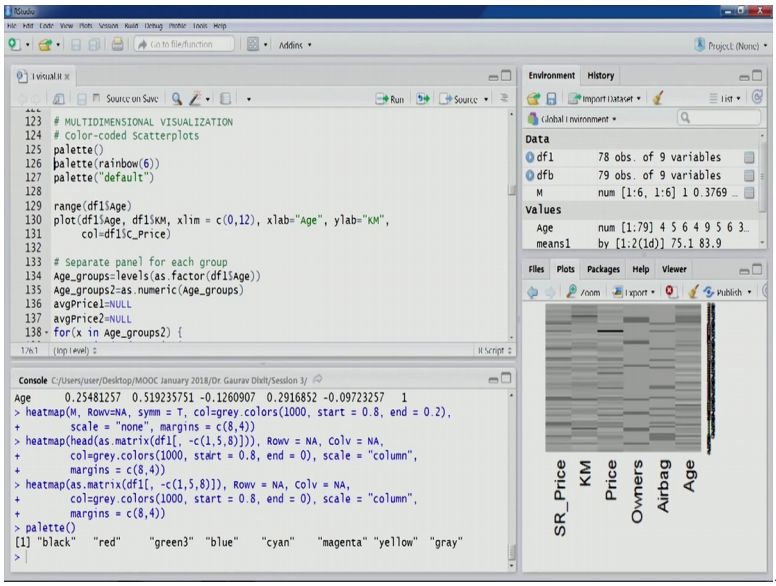
So, multiple panels can give us the multidimensional look using the 2D plots, similarly colour. So, colour coding can be done for different groups of a categorical variable. So, that will also give us the multidimensional and it will help us in building our visual perception size and shape. So, a different size and shape for points that are being depicted or shown forgettable graphics that is being generated different size and shapes can be can actually be used and therefore, that can help us give that multidimensional feel from the 2D plot animation can be done which can actually help us in visualizing a changes over time some operations like aggregation of data rescaling and interactive you know visualization that can also be done to have that multidimensional feel.

Now, when we create a real multidimensional visualization like 3D plots their visual perception is not that much clear it is difficult for us to a learn something from 3D plots. So, it is more easier for us to because of the way we having learning over the years our

learning with respect to 2D plots from 2D plots is much better than in higher dimensional done from higher dimensional plot. Now the main idea is again for these features and the operations that we talked about the main idea being help build visual perception that is going to help using support in the subsequent analysis.

So, let us go back to our studio and will go through some of the plots. So, first one being colour coded scatter plots. So, before we do some before we create some of the colour coded scatter plots let us understand the coloured schemes in R in R. So, there is this function pallet which can actually help us understand the default you know colour scheme for R in R.

(Refer Slide Time: 26:18)

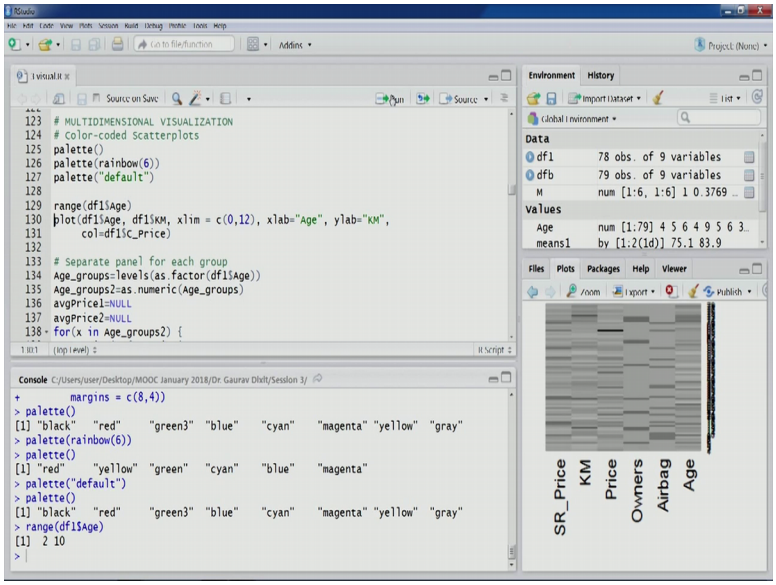


We run this particular function you can see different colours are depicted here black red green 3 green 3 blue 7 all these colours are depicted. So, therefore, for any whenever in any function we use the colour argument for example, in this plot this colour argument has been used. So, these particular colours would be picked up in that order. So, for you know first time black would be picked for different colouring red would be picked for the third different separate colouring green 3 would be picked. So, this particular colour scheme is going to be used to have different colours in your plots.

If you want to change this particular this particular colour scheme you can do that for example, rainbow 6 this is one function that can actually change your pallet scheme. So,

you can pass on this argument in the function you can again check the values that is a rerun pallet.

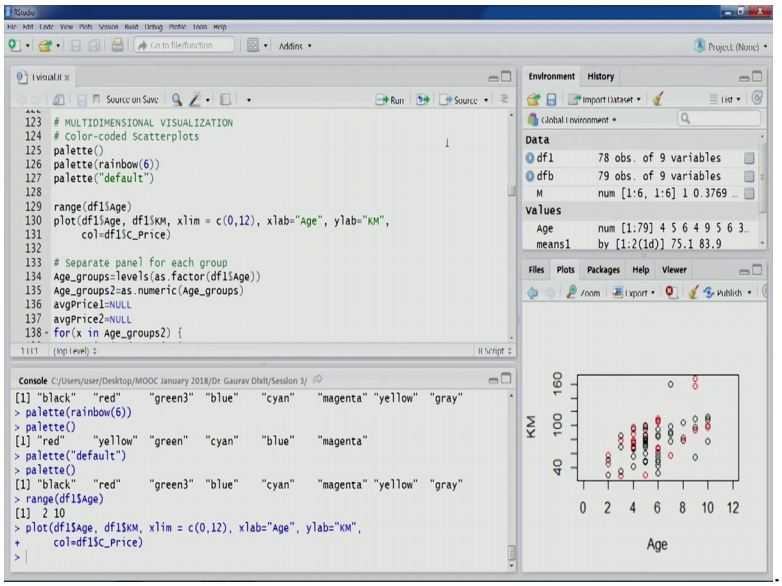
(Refer Slide Time: 27:19)



You can see now the, this colour scheme has changed to rainbow 6 red, yellow, green and these colours. So, right now will stick to the default scheme, let us reset it now you will see a default scheme is there you can recheck it you can see it is black, red, green 3. Now let us create a colour coded scatter plot. So, this is this plot we are going to create between the variable age and kilometres km. So, a colour is again colour is using. So, the colour feature that we are using this is for the categorical price variable. So, for different groups of this variable we have 2 groups in this variable in this categorical variable 0 and 1. So, for these those 2 different group different colours are going to be used, the points that are going to be plotted between age and km. So, for different groups different colours would be used.

So, let us a run these 2 lines. So, range is 2 10, appropriately specified in the plot function let us run you can see the plot.

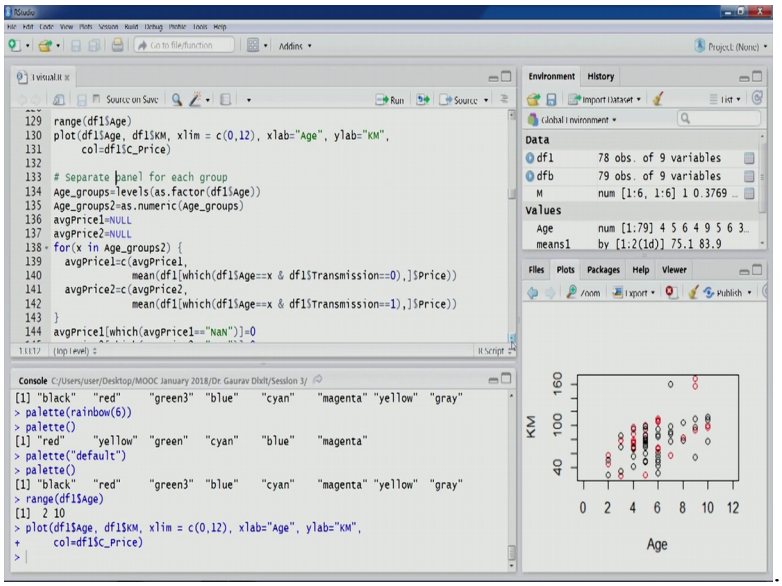
(Refer Slide Time: 28:30)



You can see 2 colours black and red as we have already seen that black and red are the first and 2 second option. So, black has been used for the group 0 and the red has been used for the group 1. So, you can see this particular plot. So, here we can have that 3 dimensional feel like we have 2 variables came in the y axis and age in the x axis and we can see the relationship between km and age in this particular this particular scatter plot as the age of a vehicle is more the number of kilometres accumulated or of course, going to be on the higher side, but you can also see that the red points or on the higher sides slightly on higher side there are few red points on the lower side. So, therefore, we can understand that categorical price were the you know which were assigned as one which means which were assigned as which were having value more than 4,00,000 equal to our more than 4,00,000 they have accumulated a more kilometres. So, those cars are being used more often. So, that 3 third dimension is being depicted using colour in this case.

Now, another kind of multidimensional visualization that we can create is multiple panel. So, we can create multiple panels each separate panel for each group. So, let us go through one example.

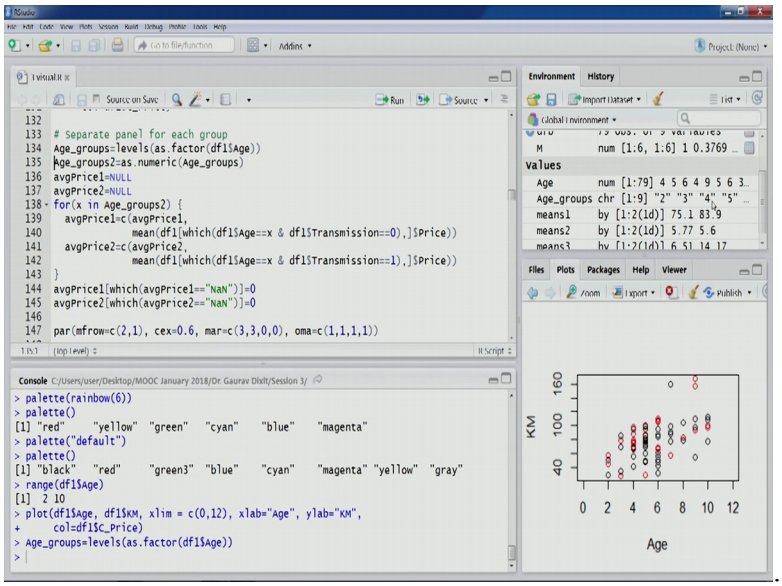
(Refer Slide Time: 30:19)



So, this particular example is being done using 3 variables. So, essentially we are trying to create bar plot and this bar plot is between this bar plot is mainly between price and age and then the different panels are going to be used depending on the transmission. So, for different transmission different panels are going to be used one panel per transmission 0 and one panel for another panel for transmission 1 and the main bar plot is between price and age were age is being used on the x axis. So, therefore, it has to be categorical.

So, therefore, we need to create a categorical we need to convert age into a categorical variable. So, let us start with that. So, age group is the variable is the categorical variable that we are going to create out of this age variable. So, you can see age dot factor and you can see age is already a factor, but to be saved this has this particular function has been used. So, we are extracting the labels with the function which can be used on a categorical variable to extract the labels different labels that are there in a particular variable. So, let us. So, age was a numerical variables, in this case s dot factor has been used to convert into a factor variable then it will have labels and labels function can be used to get to retrieve those labels.

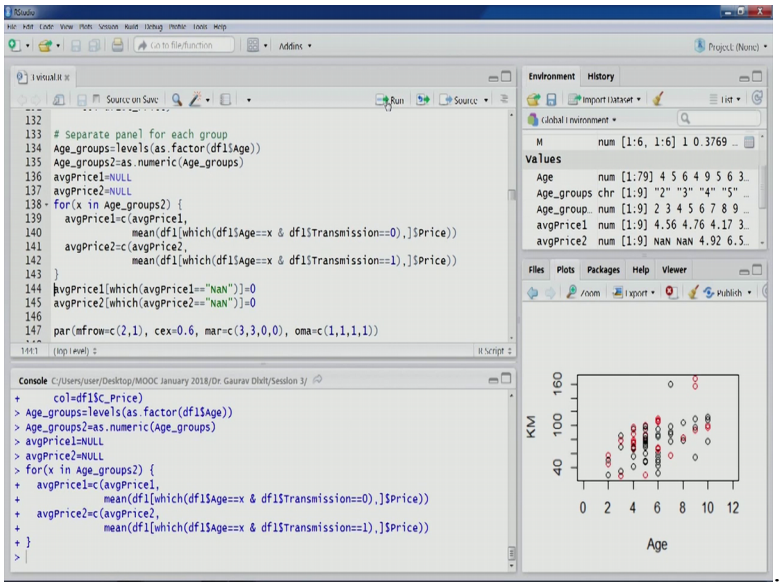
(Refer Slide Time: 31:55)



So, let us do that, age groups has been created you can see different labels 2 3 4 5, different cars with different ages. So, have now have been clubbed into different groups. So, again this is for coming you know further computation that we need to create this particular variable. So, we need to have you know we need to run a loop later you can see for loop is there. So, for that we need this age groups to which can help us in running through all the different age groups.

So, let us run then we are going to create a average price for each transmission group transmission 0 and transmission 1. So, for that we have created these 2 variable average price 1 and average price 2. So, let us initiate them once initialization is being run, for each age group we are going to run this particular loop and we are going to create these 2 variables we are going to fill feed data into these 2 variable average price 1 and average price 2 that is depending on that for transmission 0 and for you know all the groups for the transmission 0 and all the groups and for each group we are going to create an average price similarly for transmission 1 and for each age group we are going to compute the average price, let us run this particular loop.

(Refer Slide Time: 33:21)

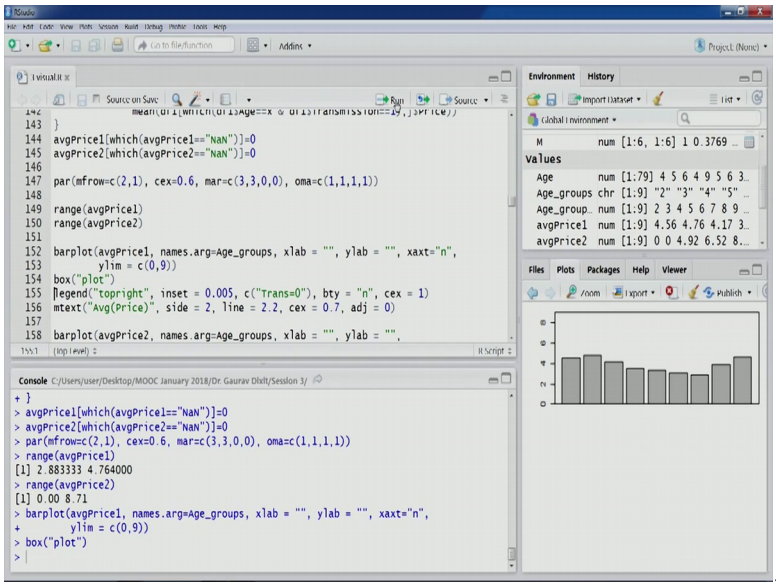


So, once this loop is done now there could be some groups where average price cannot be computed because for you know that combination did not match for transmission 1 some of the age groups there were no data, no record similarly for different transmission group transmission 0 there might be some age groups were there were no records. So, in those cases this N A N would be automatically you know assigned in R. So, therefore, we need to convert them to 0.

So, once this is done now we can now because we want to create a different 2 panels. So, in this case par is the command that can be used I mean you can see mf row is the argument. So, we want to create a 2 rows and 1 column right. So, we want to create 2 2 panels and the x axis is going to be the same. So, therefore, 1 column and there are going to be 2 panels and on y axis so, 2 and see x is again 0.6 this is actually for the labelling and this is actually for all the numbers that are going to be depicted. So, default is 1 and 0.6 is. So, we are scaling down the sizes of all the points all the numbers and text that is there margin you already know this is outer margin this is also specified here. So, let us run this command.

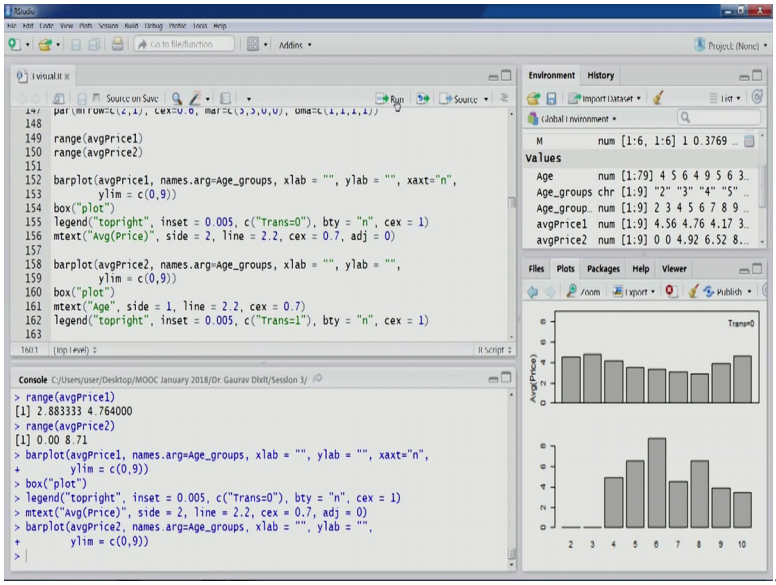
Let us have a look at the range because we are going to require that in bar plot. So, what is the range is you can see from these 2 we can see that, between 0 to 9, if we have a 0 to 9 limit it would be covered.

(Refer Slide Time: 35:15)



So, let us plot this let us create a box legend trans 0 and then another the name of the y axis.

(Refer Slide Time: 35:29)



And second plot walks name of the x axis and legend now you can see this plot has been created. So, these are the 2 panels you can see the scale for x axis is same because the same variable is being used on the x axis, but in the y axis the variable is same, but the average price for different groups could be different. So, therefore,, but is still we have used the same range. So, therefore, these 2 panels can actually be compared value by

value. So, therefore, you can see that most of the you know vehicles in different age group right and having a transmission 0 they are around 4,00,000 average price for some age group it is slightly slower as we move further this average price goes down till this particular age group age group 8 and then again for age group 9 and 10 it is increasing may be the cars over of the higher showroom price if you look at the transmission 1. So, these are automatic cars. So, the average price for these cars is a slightly on the higher side right more than 5 or closer to 6 right.

So, the automatic cars of course, they are going to be big costlier. So, therefore, used cars also are also going to be on the higher side they are also going to be costly that is reflected in this particular graphic and, but as the age increases you can see there is you know slight you know decrease as the age of a car is increasing some of course, some apprehension are there, but that is the general sense. So, will today we will stop here and will continue our discussion on some more visualization techniques in the next section.

Thank you.