Business Analytics & Data Mining Modeling Using R

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

Lecture – 13
Dimension Reduction Techniques- Part I

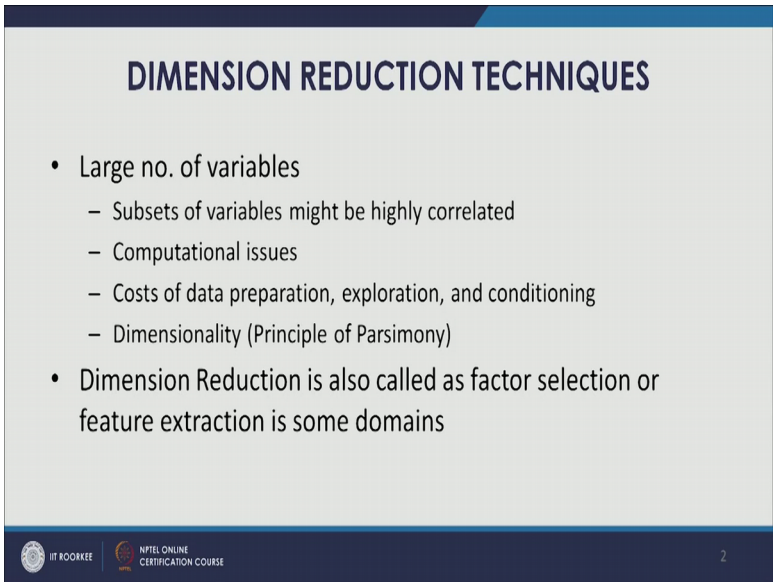
Welcome to the course business analytics and data mining modelling using R. So, we are going to start our discussion on next topic that is dimension reduction techniques. So, let us start. So, first, let us understand the context, so generally whatever discussion we have done till now, we have been talking about the different variables predictors that we generally require to build a model. So, sometimes if there are too many variables because we are also dealing with a large data set, when we talk about data mining modelling, we are generally dealing with large data sets. So, sometimes there are could be too many variables and subset of some of these variables might be highly correlated.

So, we are of course, looking for some amount of correlations, so that some link or association can be established, which could be used for prediction, but if the correlation is on the higher side then the information overlap could be problematic because few variables if they are highly correlated they can lead to spurious relationships because they are going to dominate the model.

So, their coefficient for example, if we are running regression, if we are doing regression analysis then they will dominate; their coefficient will dominate and the others, other variables coefficient might come down. So, therefore, they will dominate the influence on the outcome variable that is y. So, we do not want that, we do not want our model to be spurious or results to be relationships, results to be meaningless. So, we want to get rid of this highly correlated variable problem.

So, dimension reduction techniques are actually used for this purpose, then there are computational issues as well, for example, if we are dealing with too many variables, sometimes computation might take more time, especially if you are running different variants of your techniques and so many candidate models are there. So, it might be time consuming too much of time would be required to do your modelling exercise. So, for that also we would like to reduce the number of variables.

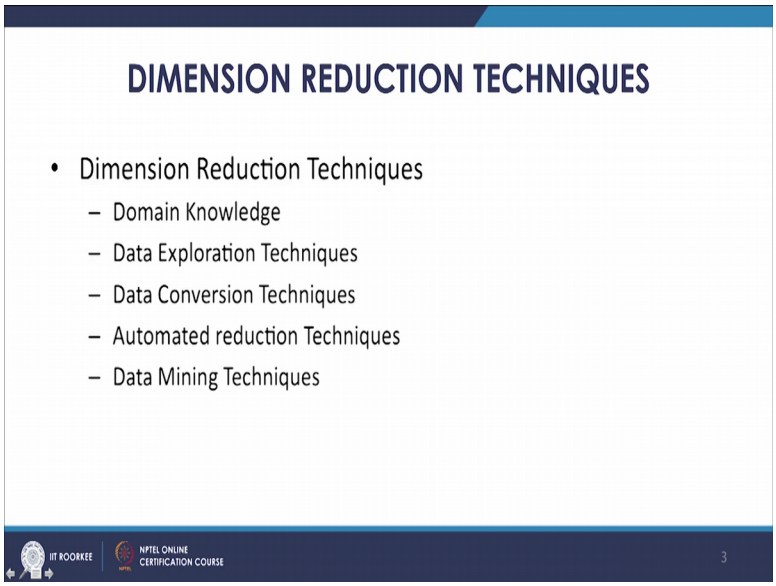
(Refer Slide Time: 02:35)



Now, another reason could be cost of data preparation exploration and conditioning. So, we talked about this space and the kind of things that we are required to do in data preparation exploration and conditioning. So, that might take more time, if there are more number of variables for example, visual techniques that we covered, that might be, that might require more time and so the cost of modelling would go up, so that could be another reason. So, dimensionality, so we talked about in previous lectures, we talked about principle of parsimony, were we said that we want a variable to have as few variables as possible and still we able to explain most of the variation in the outcome variable, so that is the desirable property.

So, therefore, we would ideally look to reduce the dimensionality of a model. So, therefore, large number of variables we always look to reduce the number of variables, but we want to retain the important variables in the model. So, we need to identify them and some of the problems for example, some variables been highly correlated, we want to get rid of that also. Dimension reduction is also called factor selection or feature selection in feature extraction in some domains, specially machine learning, artificial intelligence, so they are some other names are also being used.

(Refer Slide Time: 04:02)



So, what are the different dimension reduction techniques that can be used? So, these are the few that we are going to cover domain knowledge, data exploration techniques, data conversion techniques, automated reduction techniques and data mining techniques. So, domain knowledge is the one, that we generally that can be very handy, if you understand the phenomena, if you understand the underlying theories that are explain that particular phenomena, you understand the key variables construct and keeping in mind the kind of task that you have to perform, you can identify; easily identify key variables because you have been working on that area. So, because of the knowledge that you have throughout the phenomena about the variables construct and everything else, that can help you, that can put you in a better position in identifying key variables.

So, in that fashion you can reduce the dimension, you might not use other variable state and your modelling and that can is that is going to help. Removing redundant variable, some variables through that same process if you understand your area, you will immediately understand that these variables are not useful, so you can get rid of those variables.

Sometimes, if you have been involved in data preparation exercise and the collection exercise you would also know some of the variables which might be having a redundant values because of the way the data might have been collected, it might not you know, those variables might not be suitable because some errors will come into your model and

the results that you actually might expect might not come and your own arguments and theories might not be justified. So, therefore, we can also identify these redundance variables and if we can handle them, we should go for that, otherwise we can get rid of these variables as well.

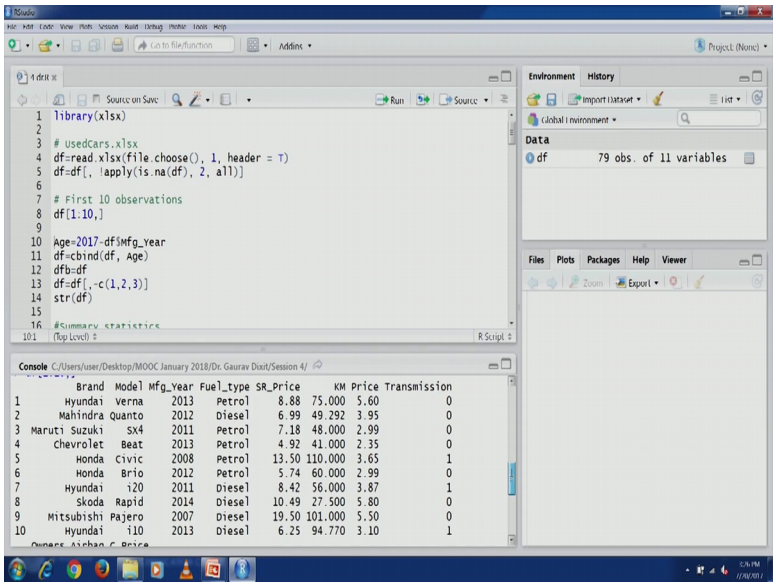
Measurement issues for variables, so sometimes we also have to see, if a variable if there is an we need to repeat the exercise can the variable be measured again, is it easy to measure that particular variable, so those kind of things also. So, that is from the project point of view, if you are running a project for your business organisation and there is some business roles and the relevant business analytics exercise that you have to perform.

So, you will also have to see, whether a particular variable can be again it can be data on that variable can be again measured and collected. So, that is also you have to see whether the exercise can be repeated or not, so those kind of issues could also be there. Next is data exploration techniques, so some of them, some of these things we have already discussed in previous lectures for example, descriptive statistics, that can also help us in reducing the dimension for example, summary statistics.

Sometimes it might be easily able to see, the you know, from the summary itself, the we can identify important variables and we can also identify the variables which we can get rid of, pivot tables in excel that can also be used to, again look at the summary level data and make further decisions about your modelling exercise. Correlation analysis can also help you identifying the variable which are correlated and should be included in the model, highly correlated variable can also be identified there that can again help you in determining which variable you need to, you have to perform closer inspection and so that is also, so correlation analysis is also going to be quite helpful in reducing the dimension.

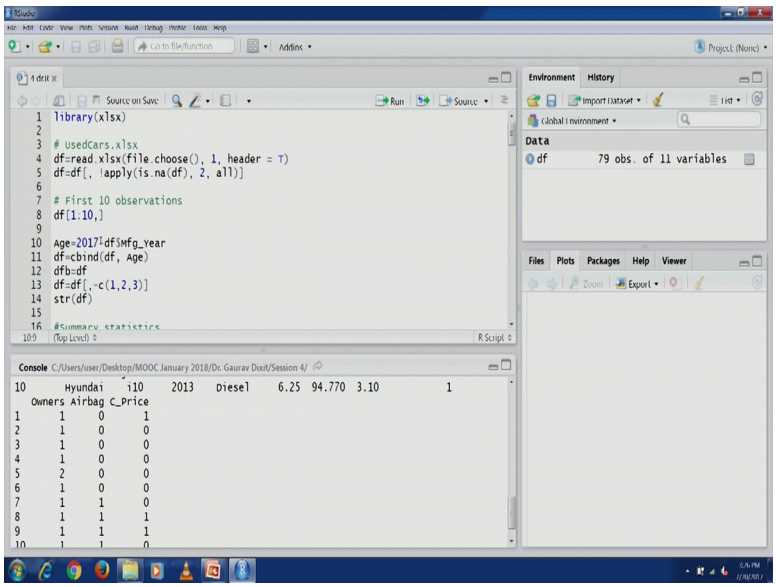
Then we have visualization techniques that we have covered in the in previous lectures. So, let us go through some exercises for these techniques. So, let us open R studio.

(Refer Slide Time: 08:16)



Let us, load this particular library that we have been using in every exercise. So, used cars data set this is the 1 that we are going to use this for this exercise. Let us import this data set again, now first an observation we have already seen this data in previous lecture, but let us have a look again. So, these are the variables, as you can see you are by now yon should be familiar with this particular data set, brand model, manufacturing year, fuel type, showroom price, kilometre price, transmission, owner's airbag and C price. So, these are the variables.

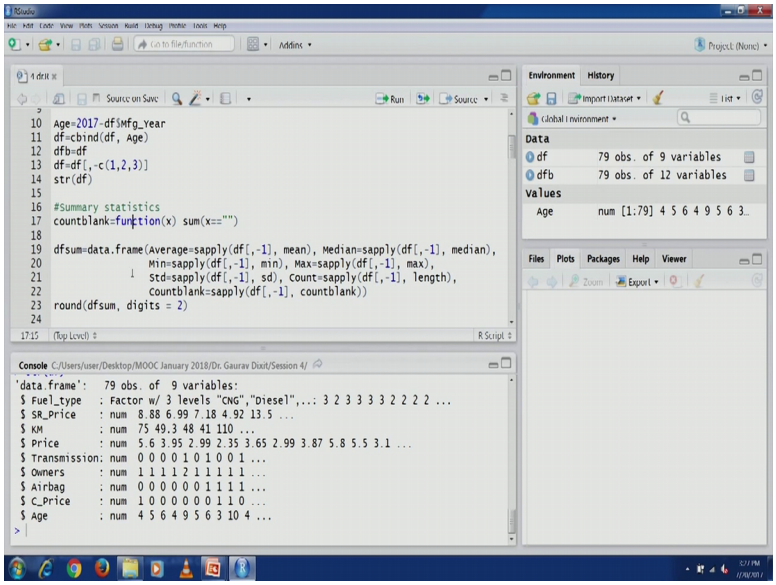
(Refer Slide Time: 09:07)



Now, another variable of interest that we have been using in other lectures as well is age. So, we are going to compute age like we did in previous lectures from manufacturing year, let us append this to our data frame, let us take a back up and first 3 variables are we were not interested.

So, let us drop them and now let us have a look at the structure. So, these are the variables of interest, 9 variables, you can see 1 being the categorical, others being numerical. Now, let us look at the summary statistics.

(Refer Slide Time: 09:44)

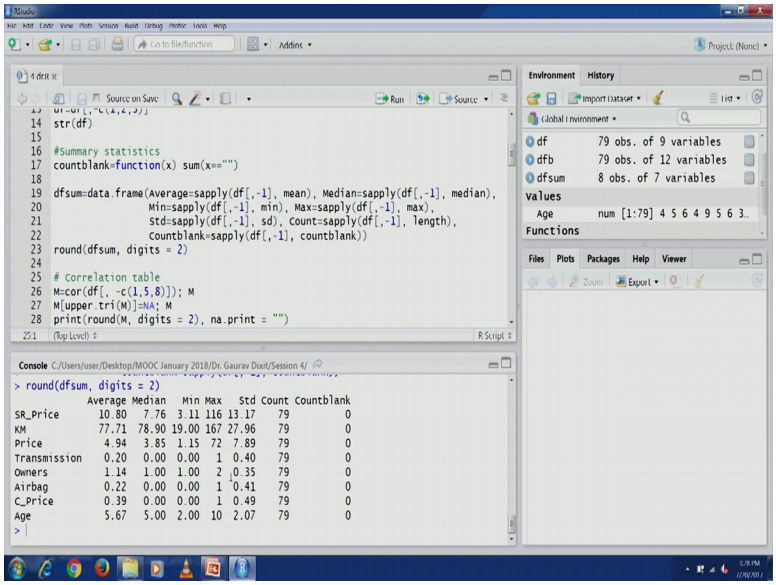


Now, this is 1 function that is also part of summary statistics that we are going to create a count blank. So, this is going to count the number of a blank cells that are there in a particular variable, in a particular column. So, that we can identify there are any missing values that are represented through blanks. So, if they are represented through NA or some other format, so we can use that and create our; write our own function and find out, if there are any such values.

So, let us create this function, so we can see a function has been created; count blank function has been created. Now, this is the data frame that we are going to create that is based on summary statistics that we are going to compute. So, mean is 1, then we are going to compute median, you would see that first column we have eliminated because this being fuel type being categorical. So, there is summary statistics are mainly applicable to numerical variables, mean, median then a minimum value and the

maximum value. So, that will tell us the range, then the standard deviation, then the length for that particular column, how many observations are there, then the count blank also. So, that we can easily spot, which variables are having all the values and if there are any missing values at all. So, let us execute this line and find out more about this variables, let us look at the results.

(Refer Slide Time: 11:18)



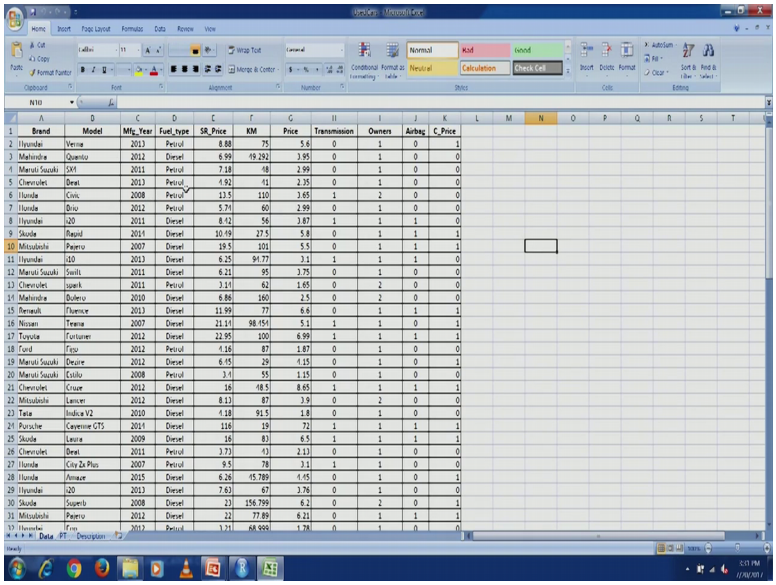
Now, you can see all these variables are having 79 as count. So, therefore, 79 observations are there and no variable is having any missing values, count blank numbers or 0 for each variables. So, we are on the safer side on these 2 counts.

Now, let us look at the range minimum and maximum value, you can see the range for different variables SR price and price, kilometres, so different range you can see, you can also look at the median values and the mean values. So, these values can also help you understand, average values are going to give you the central sense of data centralise, a numbers of the data, a minimum and maximum value will give you the range of the data.

If the, you can also compare average and median value, so if the average is higher than the median value then probably the data is rightly right skewed, if the it is average mean value is less than the median value then the probably the data is left skewed. So, that kind of thing can also be understood from these statistics. Now, this brings us to our next part that is pivot tables. So, pivot tables can also be used, so we can if there is some level of interactivity that can be achieved using excel pivot tables and some kind of combining

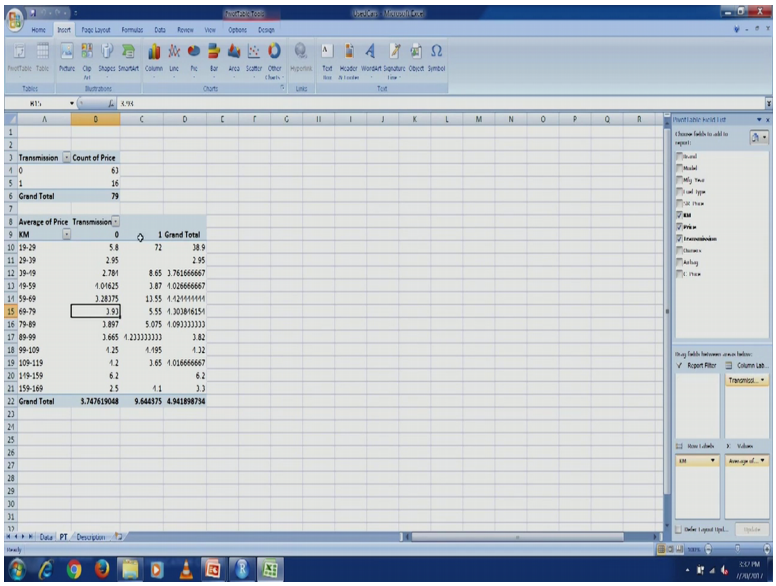
category is an understanding the relationship between variables can be understood using pivot tables we will see how. So, let us open the excel file.

(Refer Slide Time: 12:59)



So, data, so you will have to select the data and once you have selected all the data. So, let us select it, select all these variables and all the observation, you can go into the insert tab and there is pivot table there, you can create the pivot table; we have already created 1 for you. So, this p v, pivot table is there.

(Refer Slide Time: 13:19)



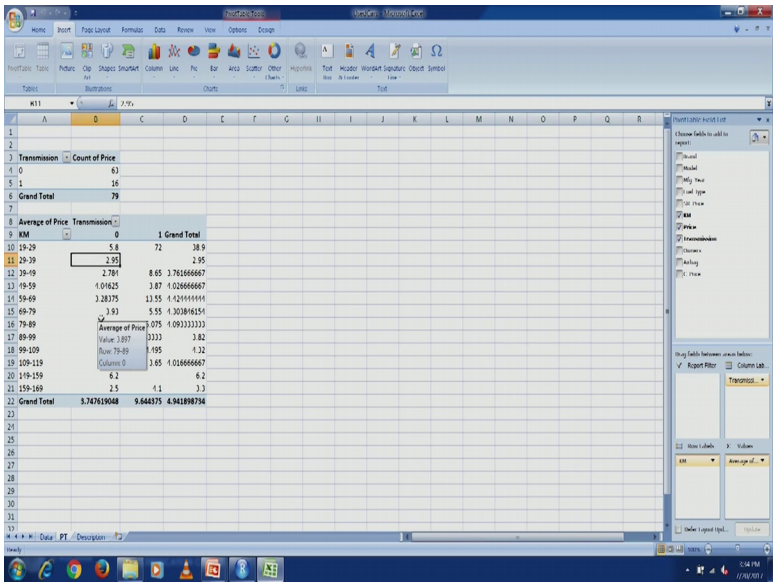
If you look at the first pivot table, so once you create the pivot table, you would have this kind of a view. So, where all the variables would be listed on this particular right top of your excel file and all the variables you can see.

Now, you would have report filter column, label, row labels and a values. So, in this particular example we have taken transmission variable in the row side and the in the values section we have taken the price and for the price also it is the counting of those values that has been done. So, if we look at this particular table, you would see transmission 0, 63 such cars are there and for transmission 1 that is for automatic 16 cars are there.

If you look at the another pivot table that is in more detail you would see we have now 2 labelling, 1 for kilometre is on the row side and the transmission is on the column side. So, transmission 0 and 1 and then different ranges have been created. Once you create a pivot table you will have to group them. So, you will have group them to have this kind of result that we have, if you are in your own time you can learn more about the pivot tables.

So, different binning's of a kilometres have been binned into these groups and you can see the values for these bins, across these 2 categories of transmission and if you look at again the price, average of price values have been taken, so the values that you see in these cells, these are actually the average price you can see average of price value 5.8 and row is this 1929. So, these kind of values they can easily have here. From there, from this, these kind of values you can easily spot, if some of the categories can be combined, for example, if KM was a variable which we wanted to convert into a categorical variable, therefore, how binning's can be done and the kind of sizes of those bins and number of such bins that those kind of decision can be performed. For example, if you look at the 79 to 89 range and 89 to 99 range, if you look at the average pricing for these 2 range this is quite close. So, probably we could have clubbed these 2 bins.

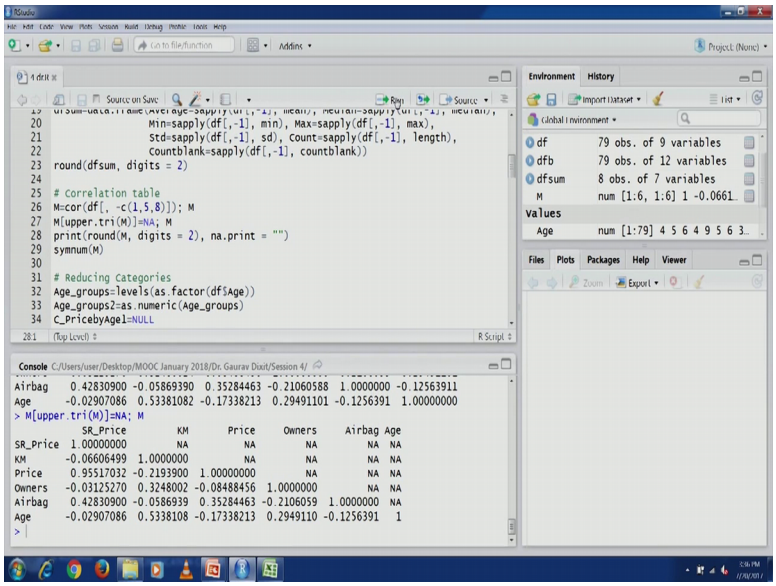
(Refer Slide Time: 15:59)



Similarly, if you look at 29 to 39 and we look at 39 to 49 the average pricing is quite close. So, therefore, we could have combined these 2 bins. So, this kind of combining can be done because as you know, when you have categorical variable and when you do your modelling, you will have to create dummy variables depending on the number of categories that are there in the variable. So, that would actually increase the dimensionality of your model and you are always looking to reduce the dimensionality, if there are few categories that is better for the model.

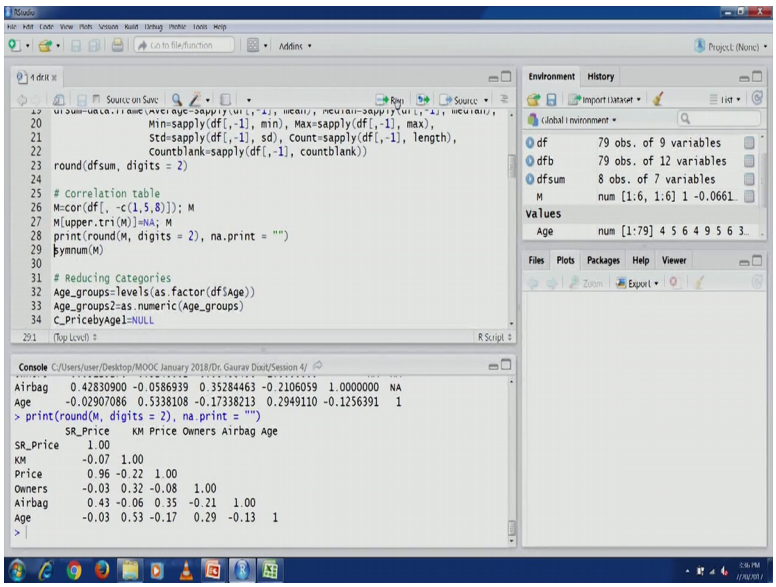
So, we are always looking to reduce the number of categories for a variable. So, we are always looking to combine some of the categories or drop some of the categories if required, so that kind of thing can be easily done using pivot tables. Now, next thing that we can do is correlation analysis. So, as we talked about correlation analysis can be helpful in identifying the highly correlated variables. So, let us open R studio and will understand through an exercise. So, again you would see the same data frame, the used cars data set we are going to use, you would see the categorical variables we are not going to consider for correlation tables because the variables have to be numerical. So, let us, so car car is the function. So, this is something that basic stat is something that we have covered in our supplementary lectures. So, you are advised to actually go through those videos, those lectures. So, correlation let us compute between the variables.

(Refer Slide Time: 17:48)



You would see SR price, so this matrix you would see in the row and column we have the same names, all the numerical variables have been selected SR price, KM price, owners airbag and age and you would see. Now, this particular metrics is symmetrical from the diagonal all the diagonal values are 1 because the variable is going to be 100 percent correlated with itself, therefore, all these values are 1.

(Refer Slide Time: 18:28)



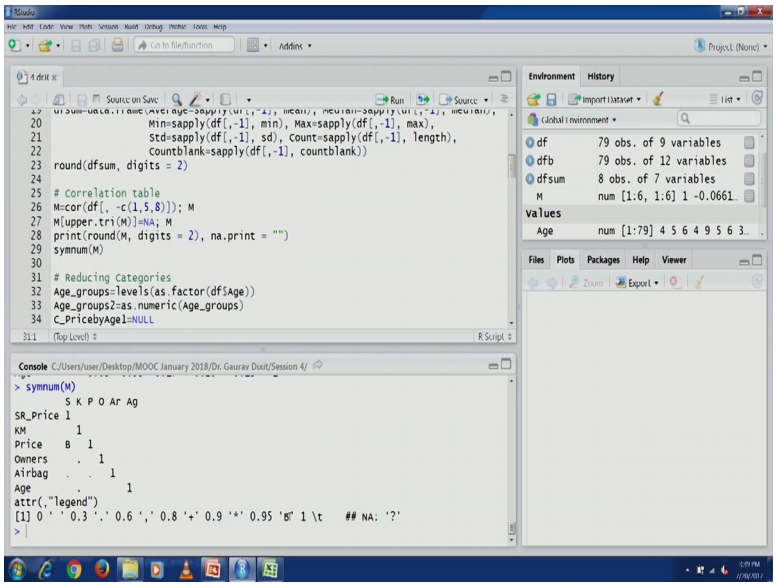
Now, we want to get rid of the upper triangular, so this we can do through this particular function upper try, we will set it to NA and then later on will get rid of this, print is the

function that can be used, so we can leave NA as blank and will have this particular correlation table, which is now showing just a 1 half lower triangular matrix and we can see the numbers. If we look at the numbers, we can see price and showroom price they are very highly correlated 96 percent correlation .96 is the correlation coefficient here.

Similarly, we can look at some other numbers for example, airbag and SR price 43 and then we can have look at other numbers like age and KM a 53, similarly airbag and price they are 35. So, there are some a highly correlated numbers, price and showroom being very highly correlated. So, these highly correlated numbers we can easily spot and then we can try to understand if some of these variables can be, some of the dimensionality can be reduced, but in this case SR price is an important variable, it being correlated with the price is because of the, that SR price is mainly determining the used price of a used car. So, a because of that this high correlation is there. So, we do not need to get rid of because this being a key variable and important variable and it is relationship is again with the outcome variable of interest that is price.

Now, there is another function that can be used to perform the same thing that is sym num, sym num is 1 function that again displays the, a correlation table, you can see here.

(Refer Slide Time: 19:56)

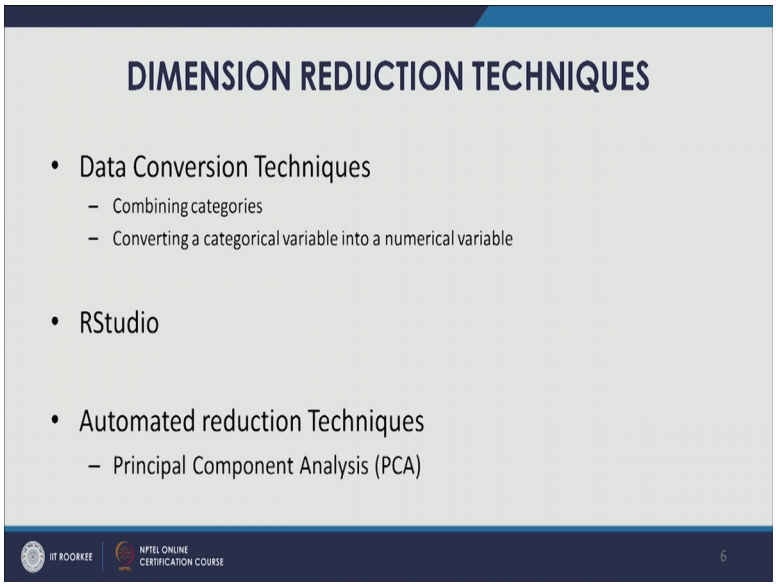


But the instead of the actual numbers they display the different symbols for different range. For example, any correlation value that is between 0 and 0.3 that is left blank, then any value that is between 0.3 and 0.6 they would displayed using dots, you would

see 4 dots there. So, 4 correlation values are dots. So, this is easier for because we are generally looking to identify the variables which are slightly on the higher side, so more than 0.3.

So, therefore, you would see any values, any correlation value which is more than 0.3, we are interested in knowing those 2 variables. So, 0.3 and 0.6 are being depicted by dot, there are 4 such values, then 0.6 to 0.8 they can be depicted by comma, we do not have any such value and 0.8 to 0.9 y plus then 0.9 to 0.95, that is why as trick. So, we do not have any such value, but 0.9521 that is depicted by B. So, that we have 1 value in this category that is price and showroom price. So, immediately we can spot through these symbols because we are interested in identifying the highly correlated values. So, this can actually help. So, this is another function. So, let us go back to slides.

(Refer Slide Time: 21:15)



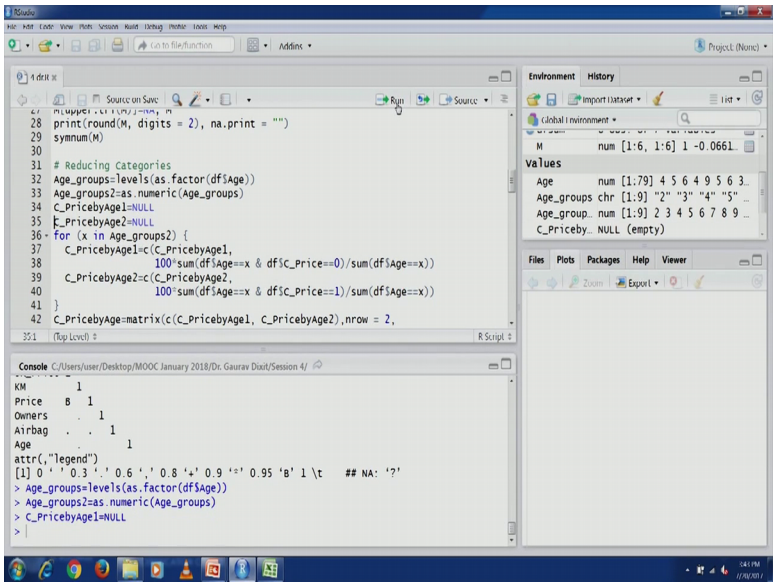
Now, another method for reducing dimension could be data conversion techniques. So, if we have a categorical variable and many categories are there and it will we can have a look at the data and identify the categories which can be combined. So, combining categories as we discussed that can reduce dimensions. Similarly, if we have a categorical variable and it is possible to convert it in to a numerical variable, so that can also help us in reducing the dimension. For example, suppose we have a collected data so on age, age groups.

So, we have not been able to collect the data on actual age of a individual, rather we have age group information. So, particular record belonging to a particular age group, so in that case depending on the number of age group that we have, so those many dummy variables have will have to be created for our modelling exercise and that will increase the dimension. So, we can reduce this scenario because age is essentially is a continuous variable. So, we can, if there are many more categories if there are many more categories for age groups we can treat it as a numerical variable. So, the middle value of a particular range can be assigned, that mid value of particular range can be assigned as the value for that particular, individual age value for that particular individual and we can convert this categorical variable into a numerical variable and thereby reducing the dimensions, so this can be done.

Similarly, for categorical variable as we talked about in the previous exercises as well, a for 2 groups if there are similar kind of numbers are there with respect to our outcome variable, then probably we can combine them. If there is 1 group which is having very few number of observation for that group, we can also drop it or we can combine it with the major category. So, that kind of combining can be performed using data conversion techniques. So, let us open R studio and let us do the same thing through an exercise.

So, in this particular case again same data set we are going to use. So, age is the variable that we have in our data set; age of used cars. So, let us convert it into age groups. So, d f age this was numerical variable we need to use the hash dot factor function to convert it into factor variable or categorical variable and now we are trying to extract the labels, so of this particular categorical variable once created.

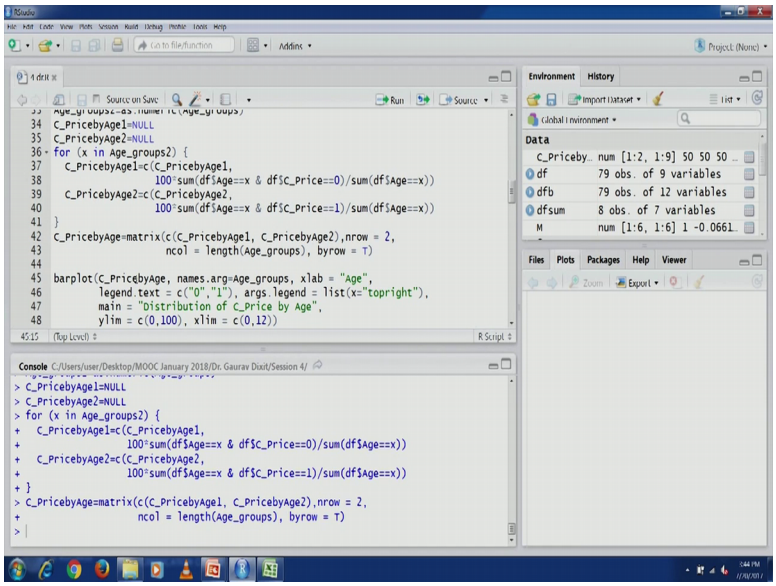
(Refer Slide Time: 23:59)



So, you would see age groups has been created and different labels can be seen here in the environment or data section 2, 3, 4, 5 and many others. Now, for further coding we would also require these groups in a numeric format. So, we are trying to create another variable to store the same information. Then, we want to a create a bar plot, which will see as later on. So, for that we need to create this C price that categorical pricing that we had created for this particular data set by age. So, let us do that.

So, there are 2 groups for categorical price 1 is 0 and 1 is for 1. So, used cars with less than 4 lakh value and then used car is greater than or equal to 4 lakh value. So, these 2 groups we have, one is, so 2 variables we need to create, for each of these 2 variables and for different age categories, we are trying to find out the this these values, we are trying to sum them up. So, you would see first C price age 1 is actually, we are trying to find out the percentage values, where age is a particular belong, belongs to a particular age is there and it belongs to particular age group and the C price is also 0. So, we want to find out the percentage of that, similarly for C price 1, we are going to perform the same thing.

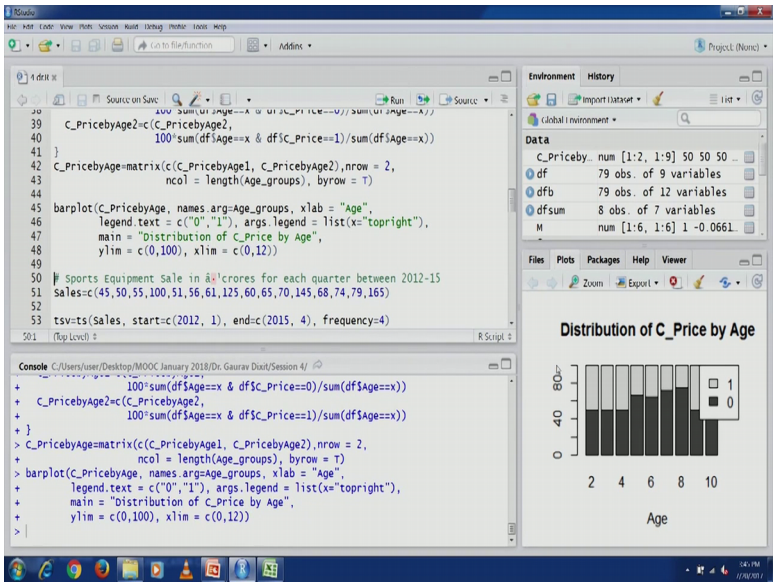
(Refer Slide Time: 25:50)



So, let us run this, you would see that different 2 different variables C price by age 1 and C price by age 2 have been created. So, 9 groups were there, you would see 9 groups were there and 9 values have been computed.

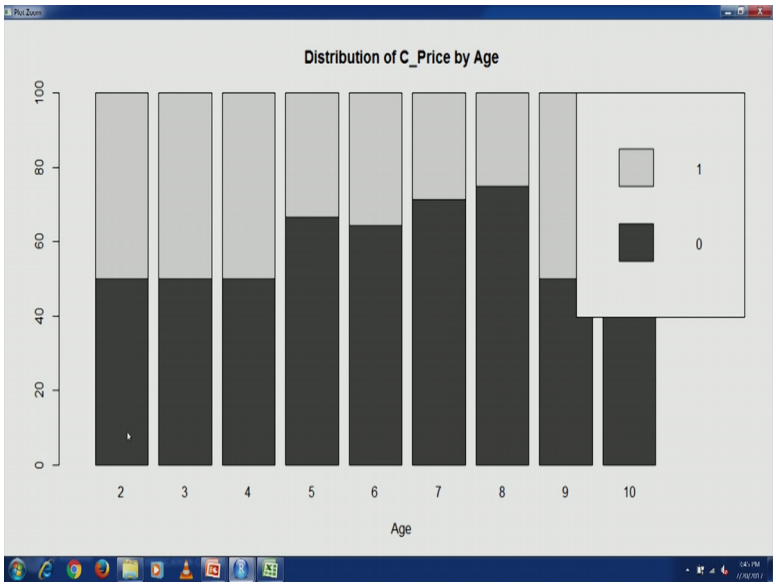
Now, let us create a matrix of these 2 variables, once this is done. Now, this particular information, this particular variable can be used to draw a bar plot, the x axis is again going to be categorical as we have been doing before. So, age groups; the labels from the age groups would be taken for this labelling, limits have been is specified every appropriately because this is going to be percentage number. So, 0 to 100 is the range for y limit and then because there are 9 groups, so this, they are going to be captured in this range for x axis.

(Refer Slide Time: 26:57)



So, let us execute this.

(Refer Slide Time: 27:04)

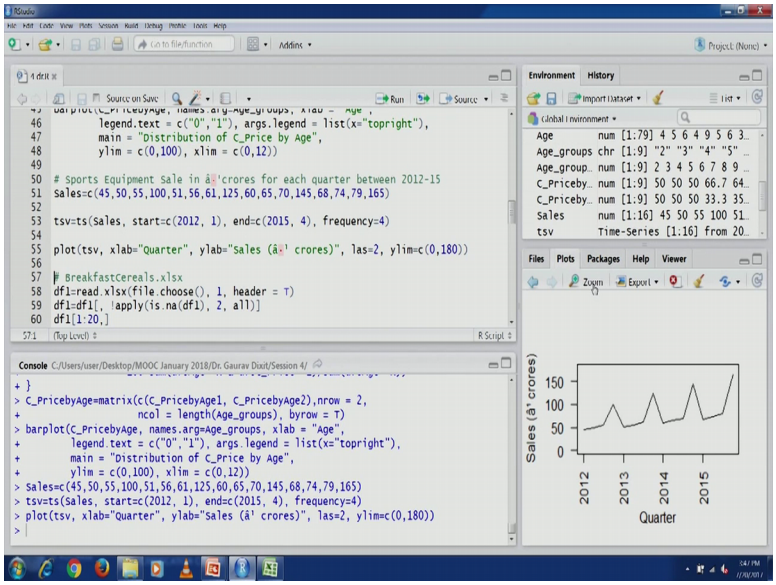


You would see that a bar plot has been generated, now you can see that 0 is represented by this higher intensity shade of grey colour and then 1 is represented by light grey colour. So, you can see for age group 2, 3 and 4 you would see that proportion of these values, this is number of cars for these is same and count of these cars for other groups this is slightly different, if you look at these 2, 3 and 4 from here, if we can see that probably age group 2, 3 and 4 they can be combined together and similarly age group 9

and 10 they can also be combined, if you look at age groups 5 6, they can be combined similarly, age group 7 and 8 can also be combined, that is mainly because the proportion does not change, so therefore, we can combine these categories.

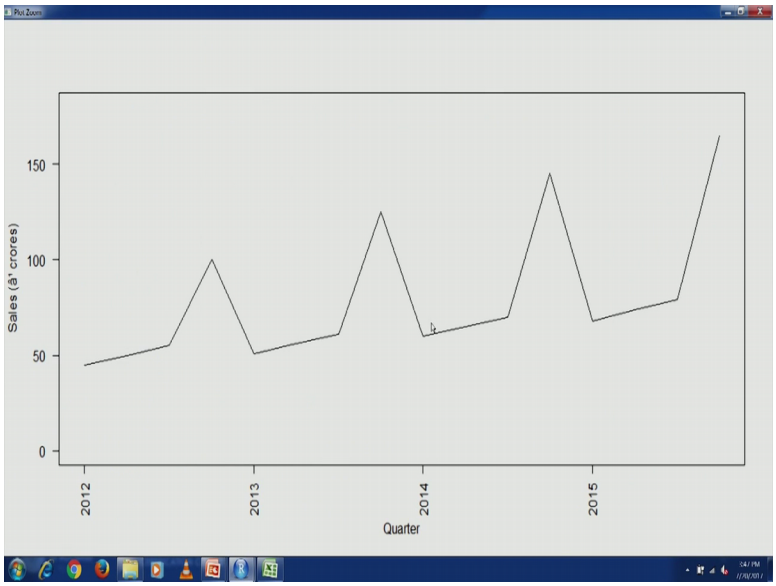
So effectively, we can reduce the number of categories from, right now they are 9 to 1, 2, 3 and 4. So, we can reduce from 9 categories to 4 categories, so thereby reducing a dimension by 5. So, this kind of combining categories can always be done as we talked about. Now, let us do 1 more exercise for time series data, so this hypothetical data we are going to create this is a time series data, so this is sales, a for force equipment sales in rupees crores for each quarter between 2012 to 2015 and for each quarters these numbers are given, let us create this variable sales and a time series is being created for the same.

(Refer Slide Time: 28:53)



You can see the frequency is 4 because this is quarterly data and the sales data is being used from 2012 onwards to 2015 and this time series has been created. Now, let us plot this.

(Refer Slide Time: 29:15)



Now, let us have a look at this particular plot. Now, you would see that let us a look at the 2012 data. So, there are going to be 3 points as we saw in the data itself, 1, 2nd point somewhere here and 3rd point here and then 4th point. So, it is always at the 4th point, where some peak is happening. So, it is at the in the 4th quarter were then there is huge spike in sales, but if we look at the a 1st and 2nd and 3rd quarter the numbers are quite close to each other. So, we are looking to combine categories, so we are looking at we have sales data for different quarters and we are interested in analysing this the same quarter wise performance.

So, probably we can combine the data for quarter 1, quarter 2 and quarter 3 because the kind of numbers that we have for this time series you would see that, their performance is quite similar. So, we can combine group them into 1 and the peak 1 could be the another group. So, this is how, combining can be done even for time series data. So, will stop here and in the next lecture will start our discussion on automated reduction techniques and the 1 that we are going to cover in this particular course is principle component analysis, so will start from there.

Thank you.