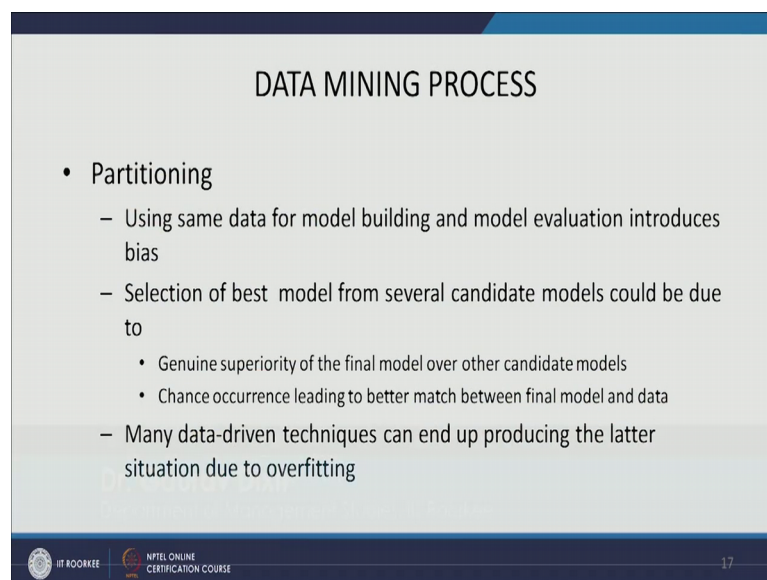


Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 06
Partitioning

Welcome to the course Business Analytics and Data Mining Modelling Using R. So, we are into second specific subject data mining process. So, last time we stopped at stopped our discussion on partitioning. So, let us pick up from there. So, in the data mining process another specific point is partitioning.

(Refer Slide Time: 00:45)



DATA MINING PROCESS

- Partitioning
 - Using same data for model building and model evaluation introduces bias
 - Selection of best model from several candidate models could be due to
 - Genuine superiority of the final model over other candidate models
 - Chance occurrence leading to better match between final model and data
 - Many data-driven techniques can end up producing the latter situation due to overfitting

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 17

So, as we discussed in this statistical modelling we generally use the same sample to build the model and then check and then perform and check it is validity again. In the data mining process we generally do partitioning wherein we split the data set into 2 or 3 partitions, 2 or 3 or even more partitions and then one of them the largest partition is generally used for model building and then other partitions are either used for fine tuning the models fine tuning the selected model or for model evaluation.

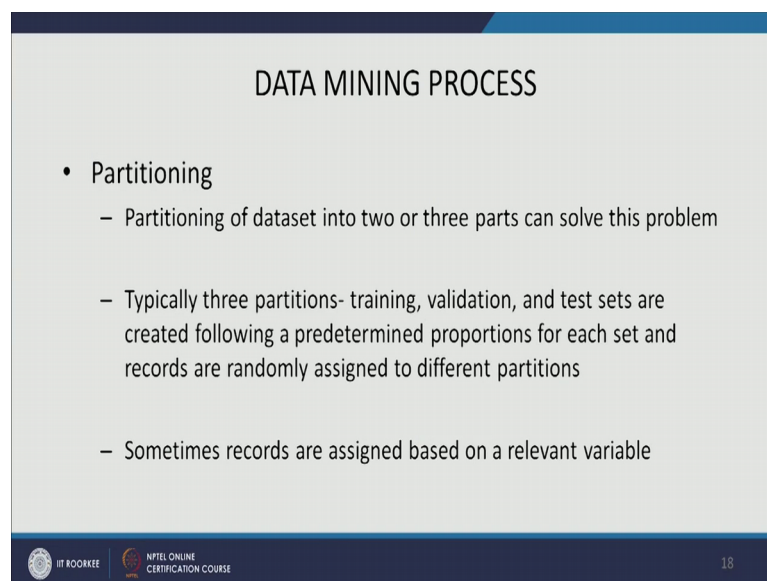
Now, another important point that we need to understand is that several candidate models how do we select the our best model. So, it could be due to a 2 main reasons. So, first one is the acceptable region the region that we want genuine superiority of the final model over other candidate model is. So, it might so happen that the a final model the

selected model is a giving superior performance genuinely in comparison to other candidate models.

The second is the problematic second reason is the problematic part that we want to minimise or remove chance occurrence leading to better a match between final model and data. So, it might. So, happen that you have 3 or 4 candidate models m_1 m_2 m_3 m_4 and it might due to a some chance occurrence that model number 3 that is m_3 is better matching with the a data and therefore, giving the superior performance. So, therefore, we need to minimise this particular situation we need to manage this particular situation.



So, partitioning is a one way to do that mainly data driven techniques they lack in structure. So, they do not impose any specific structure in on data during their modelling. So, therefore, they might end up a producing a this later situation chance occurrence because their data remain. So, their main focus is on data and that might lead to over fitting.

(Refer Slide Time: 02:57)



DATA MINING PROCESS

- Partitioning
 - Partitioning of dataset into two or three parts can solve this problem
 - Typically three partitions- training, validation, and test sets are created following a predetermined proportions for each set and records are randomly assigned to different partitions
 - Sometimes records are assigned based on a relevant variable

 IIT ROORKEE  NPTEL ONLINE CERTIFICATION COURSE 18

Now, as we said that partition of data set into 2 or 3 parts can actually solve this particular problem. So, a typically 3 partitions are created they are called training set or second one is validation set and the third one being test set.

So, again these partitions are created following a predetermined proportions. So, typically the partitions are created following 60 20 20 rules; that means, 60 a percentage

of the points 16 60 percentage of the point observation they going to training set and 20 percent go into a validation set and the remaining 20 percent go into test set. So, though the that is the typical proportion that is used.

So, but you can anyway change this so, but it has to be predetermined and this predetermined proportion is then used to create partitions; however, the records are randomly assigned to different partitions. So, the proportion is predetermined, but the records are randomly assigned to different partitions sometimes the situation might require that the records are assigned based on using some relevant variable. So, in those cases the variable decides which record will go into which particular partition.

(Refer Slide Time: 04:22)

DATA MINING PROCESS

- Partitioning
 - Training Partition
 - Usually largest
 - To build the candidate models
 - Validation Partition
 - To evaluate the candidate models
 - Or to fine-tune and improve the model
 - Test Partition
 - To evaluate the final model

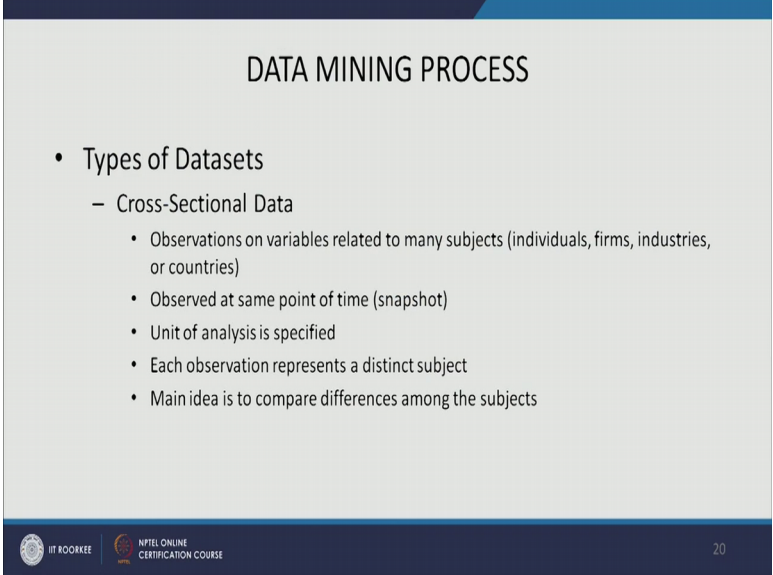
IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 19

Now, let us discuss the role of each of these partitions. So, first one being training partitions. So, usually this is the a largest partition and this is I used to the same partition the same sample exactly used to build the candidate models. So, different models that you can think of to a tackle your classification prediction task can be used then the second partition is the validation partition. So, in this particular partition is actually used to evaluate candidate models or sometimes we also use this particular partition to fine tune and improve our model.

In those situations when we use validation partition to fine tune or improve our model validation part partition also becomes part of model building



So, therefore, it might create a bias in the model evaluation if the this particular partition is used for the a evaluation purposes therefore, in those cases test partition becomes mandatory to evaluate the final model and that is the role of test partition to evaluate the final model. Now at this point we need to discuss different types of data sets.

(Refer Slide Time: 05:34)



DATA MINING PROCESS

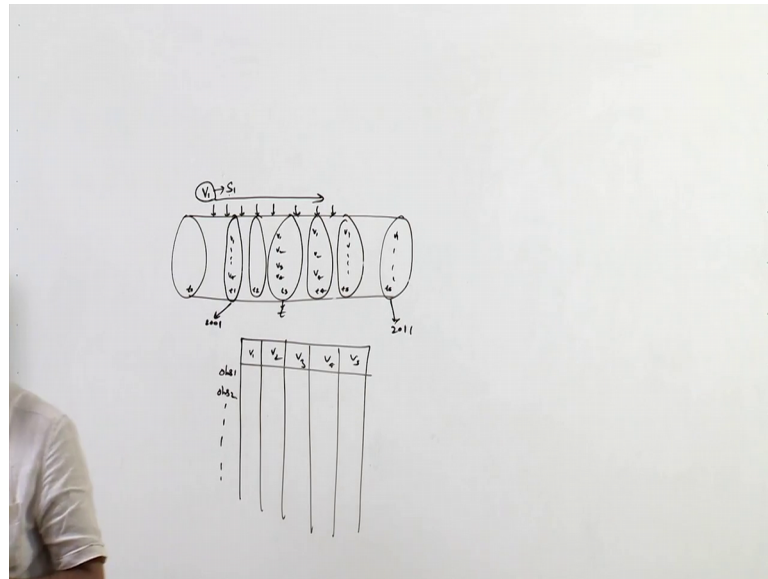
- Types of Datasets
 - Cross-Sectional Data
 - Observations on variables related to many subjects (individuals, firms, industries, or countries)
 - Observed at same point of time (snapshot)
 - Unit of analysis is specified
 - Each observation represents a distinct subject
 - Main idea is to compare differences among the subjects

 IIT ROORKEE  NPTEL ONLINE CERTIFICATION COURSE 20

So, till now the partition partitioning related discussion that we just did it is mainly applicable to cross sectional data now what different type's data sets are generally used in the statistical modelling or data mining modelling. So, let us discuss so first one be cross sectional data.

So, cross sectional data are observations on variables related to many subjects. So, variables could be relate to related to individuals they could be related to firms industries or countries regions and there are many variables and they are they are could be a many subjects. Now they are observed at same point of time so it is snapshot kind of a kind of a snapshot is taken.

(Refer Slide Time: 06:22)



So, let us assume our data set to be a cylindrical pipe. So, our variables observations on variables are related to many subjects they are taken at a cross section. So, let us see this is the point. So, all the observations on different variables v_1 v_2 v_3 v_4 for different subjects they are taken at same point.

So, this is called cross sectional data. Now generally when we are doing when we do a cross sectional analysis generally unit of analysis is also is specified though the variables might be on different subjects individual firms, but there has to be unit of analysis because the different observation that we are going to recall they are going to represent a distinct subject. For example, if we in our sample if we have. So, let say we have this sample and we have different variables in the column side and in this side we have different observations. So, each observation each observation represents a distinct subject for example, if the unit of analysis is individual. So, therefore, each observation will represent an individual.

If unit of analysis is firm then each observation will represent a firm even though the variables v_1 and v_2 v_3 v_4 could be on you know different subjects now the main idea when we do cross sectional analysis to compare differences among the subjects. So, whenever you need of analysis is individual we are trying to compare some differences that are arising out of differences among those individuals or if our unit of analysis is

firm then we are trying to steady some you know and compare differences data their among firms.

Now, second type of data is time series data, now in time series data observations on a variable. So, related to one subject. So, in the in time series data we do not deal with many subject there is just a one subject and observation on the on a variable related to that subjects are actually taken. Now observation they are the this particular variable is observed over a successive equally spaced points in time. So, each observation represents a distinct time period. So, in time series we have the same subject let us say this is the variable related to the subject one and at equally spaced times the observation would actually be made.

So, observation or on the a same subject one subject and observed over a successively equally spaced points and time each observation representing a distinct time period. So, now, again here also that this unit of time could be different it could be day's weeks or years or month. So, based on that equally spaced point in time the observations are recorded.

Now, the main idea in time series analysis is to examine changes in subject over time. So, in the this subject changes are examined over time. Now another type of data set that you may come across is the panel data sometime it is also called longitudinal data. So, panel data any way takes different features of cross sectional data and time series data. So, observations on variables related to same subjects over a successive equally spaced a point's in time are taken.

Now, the main idea in panel a panel data analysis is to compare a differences among the subjects and to examine changes in the subjects over time. So, in a way panel data can also be understood as cross sections with time order. So, we go back to our this cylindrical tube for a data set that is a this is another cross section this is another cross section.

So, all these cross sections with different equally spaced successive equally spaced time can actually be used for panel data analysis. So, you have many variables here could be v_1, v_2, v_4 in all cross sections and they are they have a definite time order.

So, we are trying to study the same subjects. So, therefore, same variables and we are taking different cross sections. Now another type of data set is called pooled cross sectional data. So, observation on a variables related to subjects at different time periods. So, you take observations, but these observations on subject, but these subjects need not to be same. So, they could be different, but the observation are made at different time periods. So, what is the main idea? So, main idea is to examine the impact of on subject due to you know environmental changes caused by some falsely intervention or some event.



So, for example, population senses that is one example of pooled cross sectional data were in India generally the population senses happens in in a in 10 years. So, let say 2001 population senses and 2011 senses and the senses is done the subjects might change, but the senses is happening at different time periods and the we are looking at due to passage of time populations we are looking at different characteristic related to population different features or variables on population.

So, again pooled cross sectional data can be understood as independent cross sections from different time periods. So, let say this is for pooled cross section data. So, this could be 2001 related data on subjects and this could be 2011 data on subjects. So, this subject need not be same and to independent cross section in the panel data the cross sections had a time order in the pooled cross sectional data they do not have a time order. So, these cross sections are independent let us discuss our next phase of data mining process that is model building.

(Refer Slide Time: 13:47)

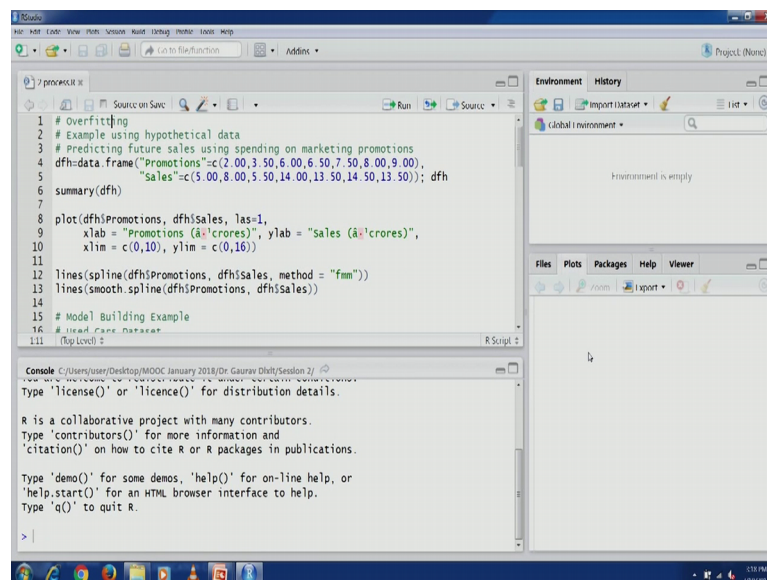
DATA MINING PROCESS

- Model Building
 - An example with Linear Regression
 - Open RStudio

24

So, we will go through this particular phase using an example with linear regression. So, let us open R studio.

(Refer Slide Time: 13:58)



```
1 # Overfitting
2 # Example using hypothetical data
3 # Predicting future sales using spending on marketing promotions
4 dfh=data.frame("Promotions"=c(2.00,3.50,6.00,6.50,7.50,8.00,9.00),
5               "Sales"=c(5.00,8.00,5.50,14.00,13.50,14.50,13.50)); dfh
6 summary(dfh)
7
8 plot(dfh$Promotions, dfh$Sales, las=1,
9      xlab = "Promotions (â'crores)", ylab = "Sales (â'crores)",
10     xlim = c(0,10), ylim = c(0,16))
11
12 lines(spline(dfh$Promotions, dfh$Sales, method = "fmm"))
13 lines(smooth.spline(dfh$Promotions, dfh$Sales))
14
15 # Model Building Example
16 # Read Cars dataset
17 (Top Level)
```

Console: C:\Users\user\Desktop\MOOC January 2018\On Gaurav Dhill\Section 2\

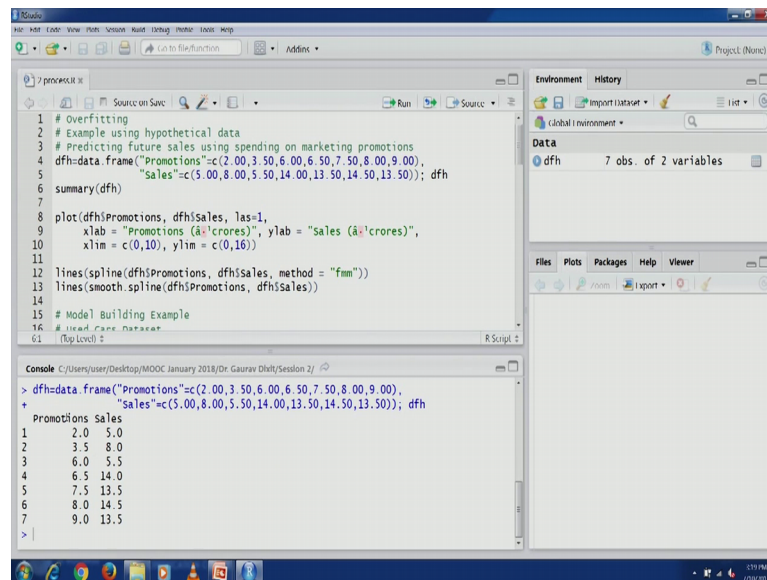
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

In the previous lecture we talked about over fitting. So, let us revisit the same concept through an example. So, this is again an hypothetical data this is about predicting future sales using a spending on marketing promotions. So, I have created I am going to create some hypothetical data. So, the you can see this data frame this code is about creating this data frame promotions is 1 variable you can see different numbers are there.

(Refer Slide Time: 14:44)



The screenshot shows the RStudio IDE. The script editor contains the following R code:

```
1 # Overfitting
2 # Example using hypothetical data
3 # Predicting future sales using spending on marketing promotions
4 dfh=data.frame("Promotions"=c(2.00,3.50,6.00,6.50,7.50,8.00,9.00),
5               "Sales"=c(5.00,8.00,5.50,14.00,13.50,14.50,13.50)); dfh
6 summary(dfh)
7
8 plot(dfh$Promotions, dfh$Sales, las=1,
9      xlab = "Promotions (â'crores)", ylab = "Sales (â'crores)",
10     xlim = c(0,10), ylim = c(0,16))
11
12 lines(spline(dfh$Promotions, dfh$Sales, method = "fmm"))
13 lines(smooth.spline(dfh$Promotions, dfh$Sales))
14
15 # Model Building Example
16 # Read Data Dataset
17 (Top Level) >
```

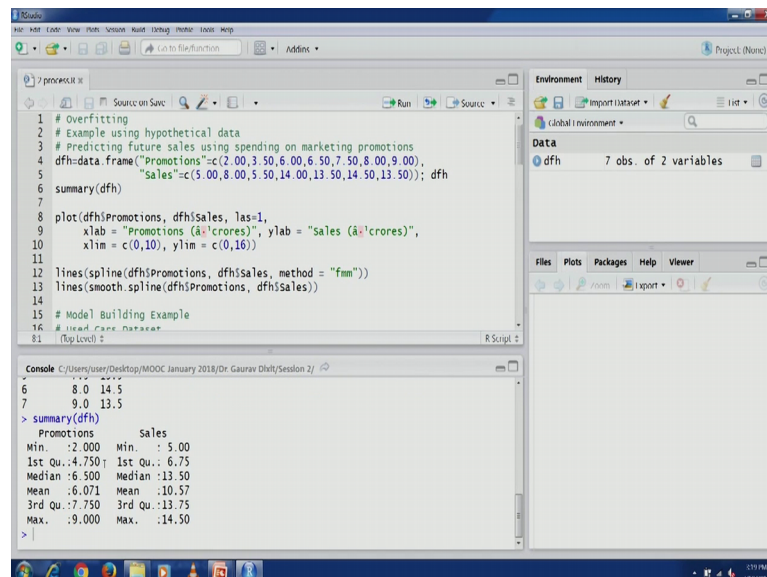
The console shows the output of the data frame creation and summary:

```
> dfh=data.frame("Promotions"=c(2.00,3.50,6.00,6.50,7.50,8.00,9.00),
+               "Sales"=c(5.00,8.00,5.50,14.00,13.50,14.50,13.50)); dfh
Promotions Sales
1          2.0    5.0
2          3.5    8.0
3          6.0    5.5
4          6.5   14.0
5          7.5   13.5
6          8.0   14.5
7          9.0   13.5
```

The Environment pane on the right shows the data frame 'dfh' with 7 observations and 2 variables.

So, these numbers are suppose are in rupees [FL] and then you have a sales. So, these numbers are again in corrodes. So, we are going to create this particular hypothetical data. So, let us create it can see promotion and sales number. So, we have same in observations on these 2 variables let us look at the summary.

(Refer Slide Time: 14:54)



The screenshot shows the RStudio IDE. The script editor contains the same R code as the previous slide. The console shows the output of the summary function:

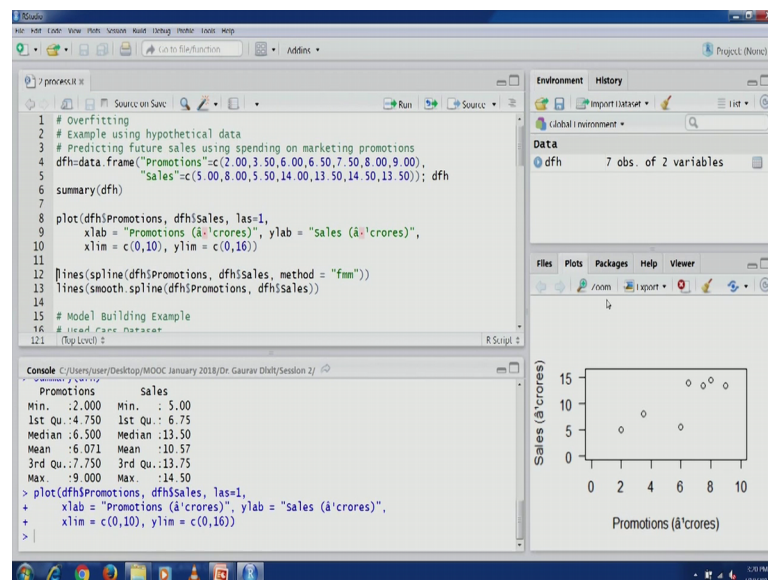
```
> summary(dfh)
Promotions Sales
Min.       :2.000   Min.    : 5.00
1st Qu.: 4.750   1st Qu.: 6.75
Median : 6.500   Median :13.50
Mean   : 6.071   Mean   :10.57
3rd Qu.: 7.750   3rd Qu.:13.75
Max.   : 9.000   Max.   :14.50
```

The Environment pane on the right shows the data frame 'dfh' with 7 observations and 2 variables.

So, these are the some of the statistics on these 2 variables promotions and sales. So, let us plot. So, you can see I am going to plot promotions on x axis and sales on y axis and you can see that x axis label and y axis label is given there and then limit on x axis and y

axis. So, I we discussed in the previous lecture limit can be given their using the results from summary command you can see the most of the value some minimum in promotion is 2 and maximum is 9. So, therefore, most of the value will lie in the range of 0 10s that is why I (Refer Time: 15:34) this assessment as 0 to 10 similarly for y limit.

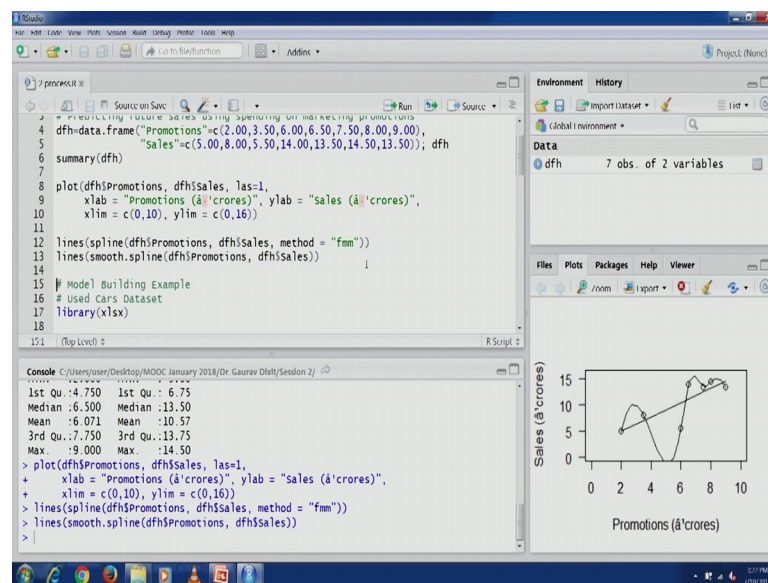
(Refer Slide Time: 15:40)



So, we can plot this this particular in the plot this particular symbol that you might see this was actually supposed to this was supposed to be a symbol for rupees may be in this system that particular symbol is not supported. So, therefore, it is coming as garbage in this case, but we look at the data. So, this is sales verses promotions you can see.

So, looking at this data we can try to affect different models which can actually help us understand the relationship between sales and promotions, but because we are doing a business analytics course. So, our idea is not just to understand the relationship, but to understand relationship and use them in a fashion. So, that we can improve our prediction we can do some predictions.

(Refer Slide Time: 16:41)

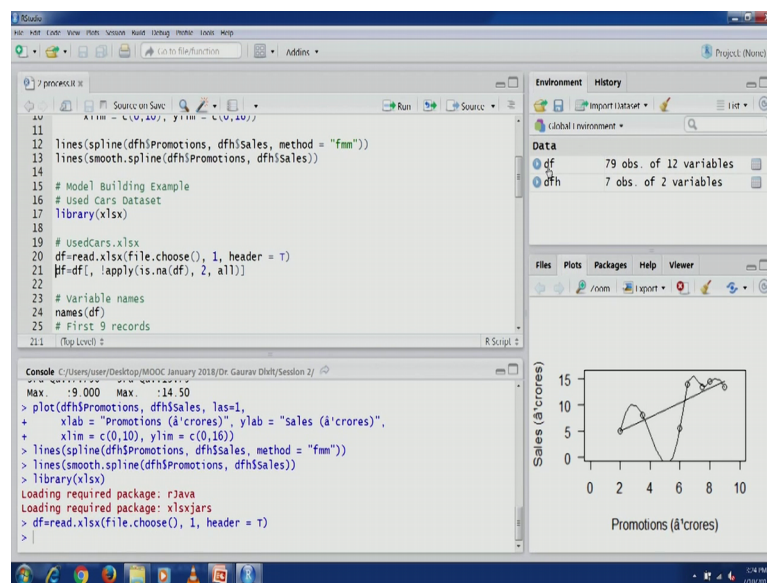


So, therefore, if we try to fit some complex model let say let us try to fit a cubic curve you can see that in the plot and this is the cubic curve that we have tried to fit here. So, we if you look at this curve so supposing this being a complex function which we are trying to fit over data and this is leading to perfect 100 percent fitting 100 percent match.

So, therefore, this is in a way causing over fitting of in the model you can see this hard to imagine, when you move when you know increase your promotions spending on promotions from 4 corrodes to 5 corrodes and the sales is actually dropping. So, that kind of relationship is difficult to imagine with this is just an example, how a complex model if it is fitted on data it can lead to over fitting a better model could be this particular line. You would see most of the observation they all closed to this line and this line could be the better model for this observations and this sample.

Now, let us go through our model building example. So, we are going to use this particular hypothetical data said used car data set it is actually based on the many post related to a used car sales that are made online, but mainly it is a hypothetical data. So, let us load this particular file this particular library.

(Refer Slide Time: 18:45)

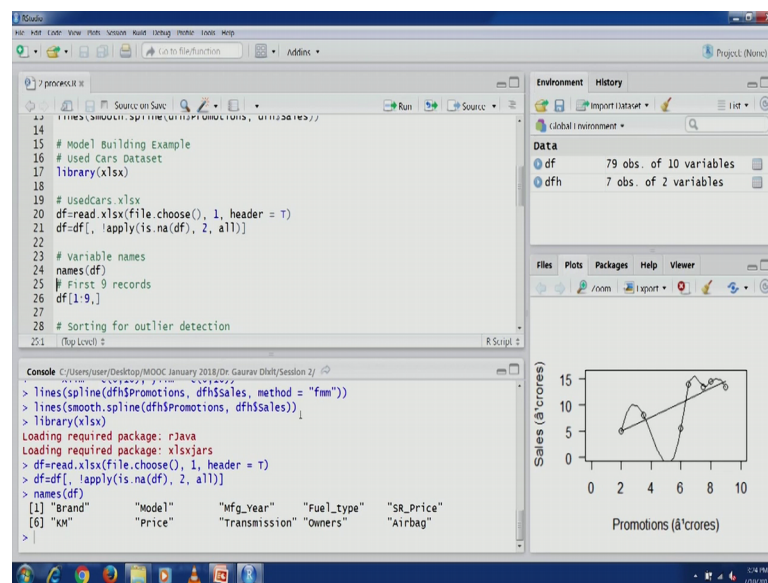


And this used car used car excel file you would see that this 1 d f has been created with 79 observations and right now it is showing as 12 variables.

But it is actually some 3 type deleted columns in excel that are that have also been picked up by R as variables even though there is no data those columns have been deleted. So, for that we have this particular code which will actually help us remove these deleted columns from excel.

So, that is why we have this particular line. So, it will actually remove those 2 variables you can see now and the environment section we have 79 observation with 10 variables those 2 variables our actually deleted columns in excel file.

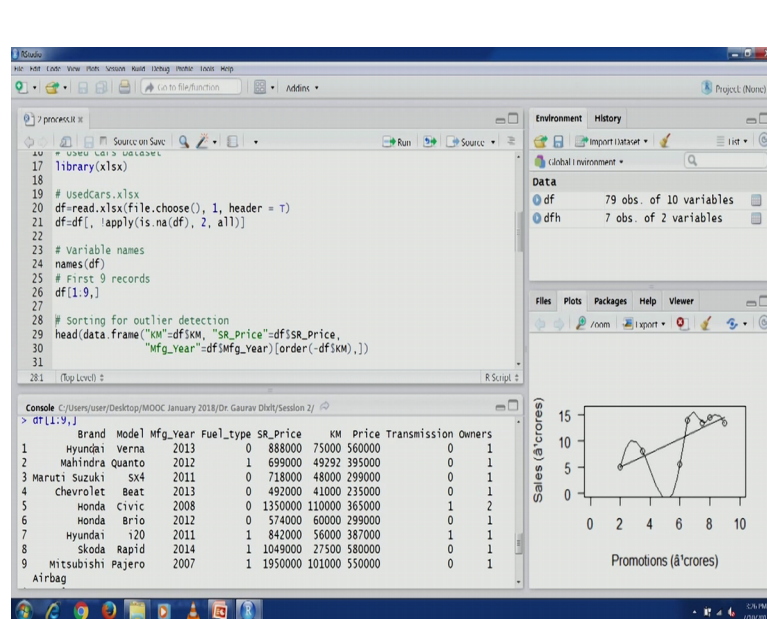
(Refer Slide Time: 19:41)



Let us look at the variable names. So, these are the variables we have brand of the car, model of the car and we have manufacturing year of the car then we have fuel type whether it is petrol diesel or c n g then we have S R price which is actually showroom price for the car then we have kilometres a accumulated. So, this is the related to another related to car then the price the offer price for the used car, then the we have another variable on transmission whether it is automatic or manual then we have another variable on owners. So, the number of owners who have actually had the ownership of the car and it is life time and then air bag number of air bags that are there in the car.

So, you would see that some of the some many of these variables are directly regiment in a sense to predict the to predict the offered price of a used car. So, what we are trying to the task that we are trying to perform here is prediction of offered price for used cars using these variables you would see accumulated kilometres and the age of the car which can be competed using the manufacturing year the transmission time number of owners. So, these are some of the variables which could be relevant in terms of making our prediction task related to offering offered price.

(Refer Slide Time: 21:25)



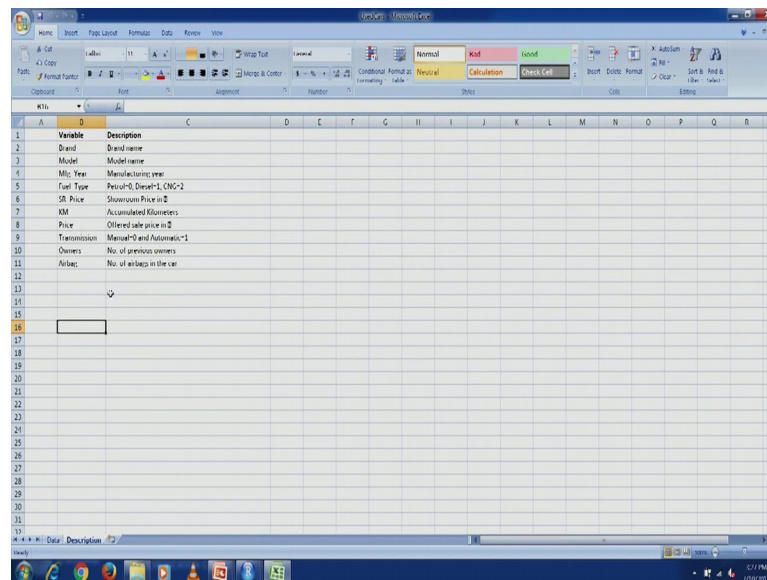
So, let us look at the first 9 records you can see here. So, different you know different used cars you know from different models Hondai, Mahindra, Maruthi Suzuki honda. So, these all are the then the model name model names are also available then the manufacturing here then the fuel type is there we go back to our excel data set.

(Refer Slide Time: 21:55)

	Brand	Model	Mfg_Year	Fuel_Type	SR_Price	KM	Price	Transmission	Owners	Airbag
1	Hyundai	Verna	2013	0	888000	75000	560000	0	1	0
2	Mahindra	Quantro	2012	1	699000	49292	395000	0	1	0
3	Maruti Suzuki	Sx4	2011	0	718000	48000	299000	0	1	0
4	Chevrolet	Beat	2013	0	492000	41000	235000	0	1	0
5	Honda	Civic	2008	0	1350000	110000	365000	1	2	0
6	Honda	Brio	2012	0	574000	60000	299000	0	1	0
7	Hyundai	i20	2011	1	842000	56000	387000	1	1	1
8	Skoda	Rapid	2014	1	1049000	27500	580000	0	1	1
9	Mitsubishi	Pajero	2007	1	1950000	101000	550000	0	1	1
10	Hyundai	i20	2013	1	625000	91770	310000	1	1	1
11	Maruti Suzuki	Swift	2011	1	621000	95000	375000	0	1	0
12	Chevrolet	Spark	2011	0	321000	62000	165000	0	2	0
13	Mahindra	Bolero	2009	1	688000	160000	250000	0	2	0
14	Maruti	Fluence	2013	1	1199000	72000	660000	0	1	1
15	Nissan	Teana	2007	1	2111000	98154	510000	1	1	0
16	Hyundai	Elantra	2012	1	2295000	200000	699000	1	1	1
17	Tata	Indica V2	2008	1	118000	87000	187000	0	1	0
18	Maruti Suzuki	Swift	2012	1	647000	38000	415000	0	1	0
19	Maruti Suzuki	Celerio	2008	0	340000	55000	115000	0	1	0
20	Chevrolet	Cruze	2012	1	3600000	48500	885000	1	1	1
21	Mitsubishi	Lancer	2012	1	813000	87000	390000	0	2	0
22	Tata	Indica V2	2008	1	128000	71500	180000	0	1	0
23	Porsche	Cayenne GT5	2014	1	13600000	18000	7200000	1	1	1
24	Skoda	Laurel	2009	1	1600000	83000	650000	1	1	1
25	Chevrolet	Beat	2011	0	372000	43000	212000	0	1	0
26	Hyundai	City Zeta Plus	2007	0	950000	78000	310000	1	1	0
27	Hyundai	Elantra	2005	1	638000	15789	449000	0	1	0
28	Hyundai	i20	2013	1	763000	67000	376000	0	1	0
29	Skoda	Superb	2008	1	2300000	156799	620000	0	2	0
30	Mitsubishi	Pajero	2012	1	2200000	77890	621000	0	1	1
31	Hyundai	Elantra	2012	0	1111000	48950	178000	0	1	0

And let us look at some of the dummy coding that has been done here.

(Refer Slide Time: 22:00)



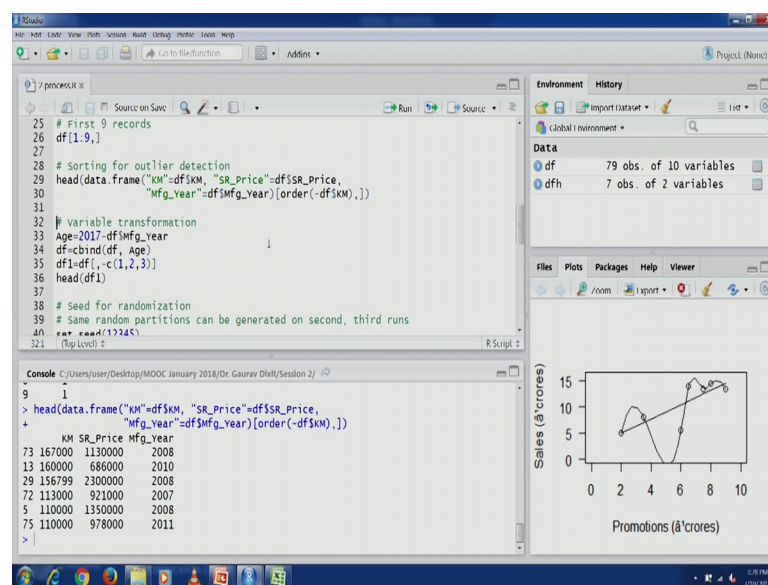
Variable	Description
Brand	Brand name
Model	Model name
Mfg. Year	Manufacturing year
Fuel Type	Petrol=0, Diesel=1, CNG=2
SR Price	Showroom Price in ₹
KM	Accumulated Kilometers
Price	Offered sale price in ₹
Transmission	Manual=0 and Automatic=1
Owners	No. of previous owners
Airbag	No. of airbags in the car

You would see that fuel type it has been indicated as petrol for petrol it is 0 then for diesel it is 1 and for c n g it is 2 then you have, another per transmission 0 means manual and 1 means automatic let us go back.

So, again for our layer during our discussion on outliers we talked about that sorting can also help's in terms of outlier ejection if there is some value related to some variable some measurement which is which looks out of place which does not seem to be real then that can that can be found out using sorting. So, let us do that. So, we have picked up these 3 important variables kilometres showroom price and the manufacturing year.

So, in this case everything looks, but sometimes some value might look like out of place it could be due to typing error or something else.

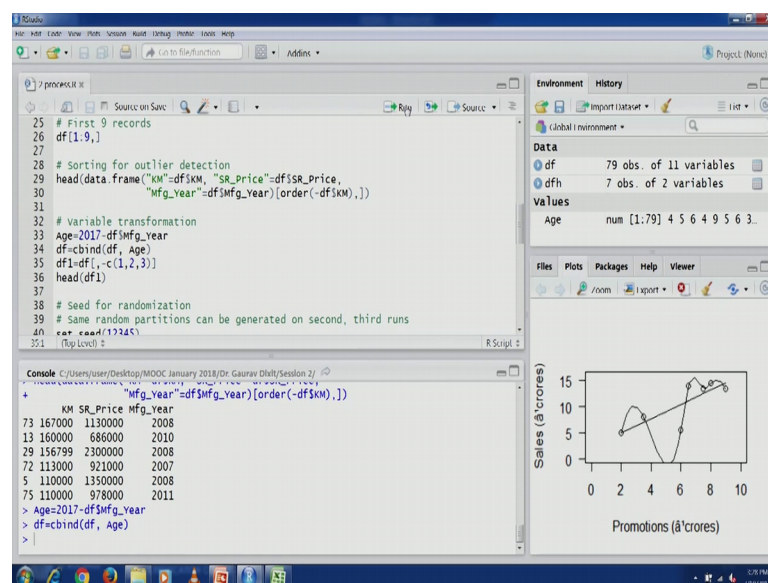
(Refer Slide Time: 23:16)



So, that can be easily identified and that value can then be handled then as we discussed age could be age could be an another important variable in terms of predicting the value of a used car.

So, let us compute this particular variable. So, you can see the current year if it is the current year is let say 2017 then it be subtracted by manufacturing year.

(Refer Slide Time: 23:41)



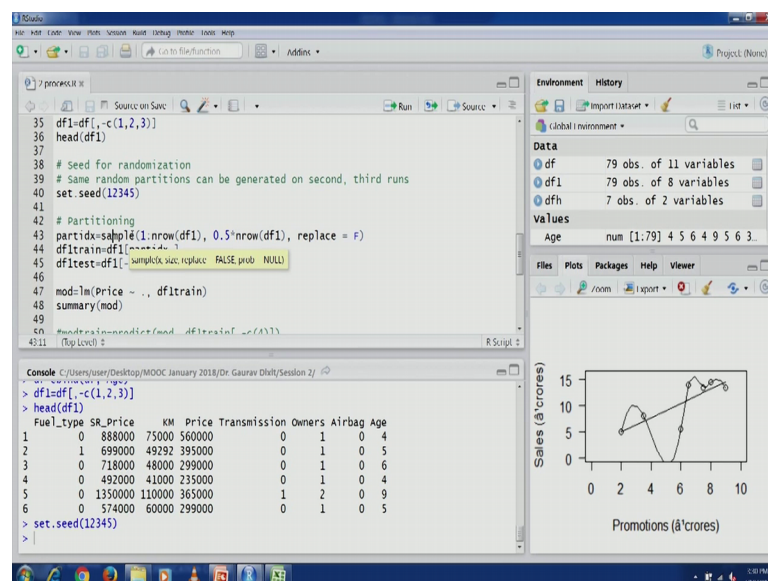
So, we could be able to a compute an age that is add it to the distinct data frame and since we are interested in since we are not interested in first 3 variables the brand name

the model name and the manufacturing year any more. Therefore, we can compute another data frame by the moving these 3 columns. So, these are the variables and of interest fuel type S R price showroom price kilometres price transmissions and transmission and owner's air bag and age.

Now, another important concept related to modelling is model building is seeding. So, generally many analysts prefer to do seeding actually help helps provide some you know flexibility in randomisation. When if you know if you want to use these same partitions in your second or third run seeding would actually help us duplicate the same random partitions. So, in R we have this function set dot seed which can be set up. So, you can further up in this case set dot seed we have given 1 2 3 4 5.

So, this this could be any number that you like and a seed would be created. So, next time when you want to create the same partitions this seed can actually helped you duplicate the same. So, therefore, let us execute this.

(Refer Slide Time: 35:20)



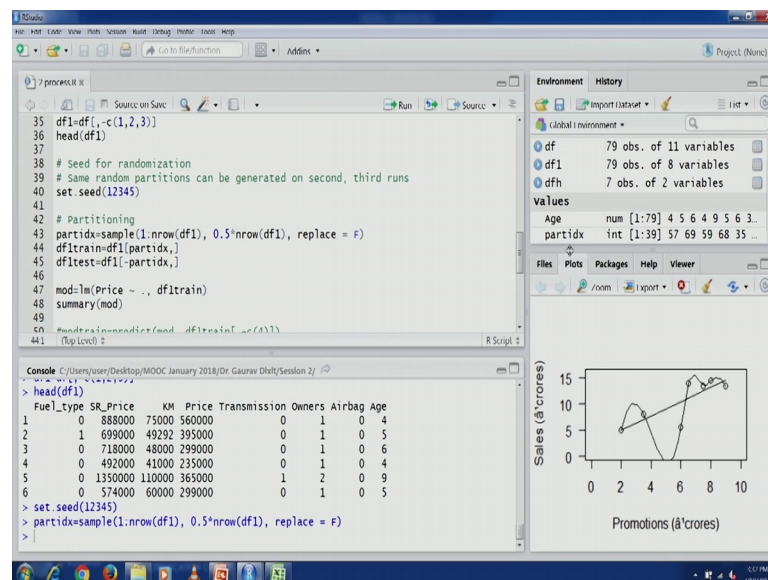
Now, let us move to partitioning. So, here you would see that you are using sample function that is available in r. So, this sample function can actually be used to randomly draw different observations and create an index which can be used to create an different partition for example, in this particular case we want to create just 2 partitions of 50 percent each. So, we just want to create 2 partition training and test.

So, in this case 50 50 percent you would see in the first argument we have given the range of values that we want in our sample. So, this is one form the sample size that is can be computed using n row command and passing the data frame as argument in R and you can see the second argument is 0.5 multiplied by again the size; that means, we want 50 percent of the observations in one sample and then same in the another second sample and you would see the replace third argument it is assigned as false.

So, therefore, this is without replacement. So, this sampling is without replacing because we are we want to create a you know we want to do partitioning. So, therefore, we do not want the same observation to again appear in validation partition or test partition. So, we want some of the observation randomly picked and being assigned to training partition and the other observation that randomly picked and being assigned to other partitions depending on the number of partitions.

So, let us execute this particular command. So, this will actually create a few are into the data section.

(Refer Slide Time: 27:06)



You would see the you would see that an integer Vectorize being created. So, these values are actually index 2 different observations. So, now, this index is indices can actually be used to assign different observation to different partition. So, for example, data frame one we can use this particular index to assign these observation you would see that the integer vector part ideas it actually has 39 observations. So, the data training

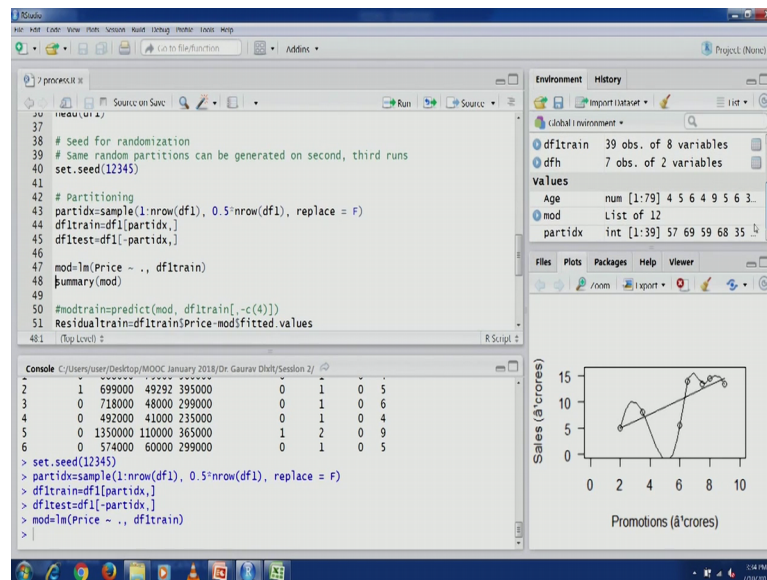
partition would actually end up with 39 observations you would see the `df1` train has been created 39 observation 8 variables. Now the remaining remaining observation will actually go to the test partition you would see forty observations have had been assigned to test partition.

Now, once these partitions have been created we can do our modelling. So, now, `lm` is the function that is used for linear regression in R. So, first argument is actually the formula. So, in this case formula and how the formula is actually written here is actually the price you have to pick your outcome variable or your output variable the dependent variable which you want to predict. So, in this it is the price that is the offered price of the car and then we used these 2 this is the way we write the formula we used till date and then dot means all other variables that are present in the data frame they would be picked up as the input variables or the independent variables and they would be they would be used as predictors for model building.

Now I would see that generally a dollar notation is used with the data frame, but in this case because this is the way `lm` is implemented you can mention the name of the data frame in another argument and the name of variables in the formula. So, it will be taken other things will be taken care of within the `lm` function. So, let us execute this line and build our linear regression model.

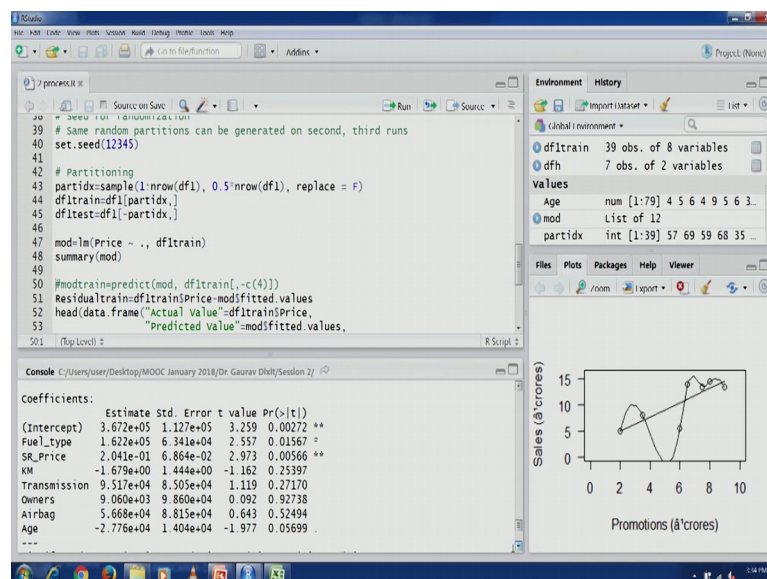
You would see again in the valued section and environment section and again values subtraction have been created you would see a `mod` has been created list of a 12.

(Refer Slide Time: 29:23)



Let us look at the summary in the summary you will get the results of your regression analysis, you can see the formula you can see the residuals is some statistics related to residuals the mean and max value median value and another things.

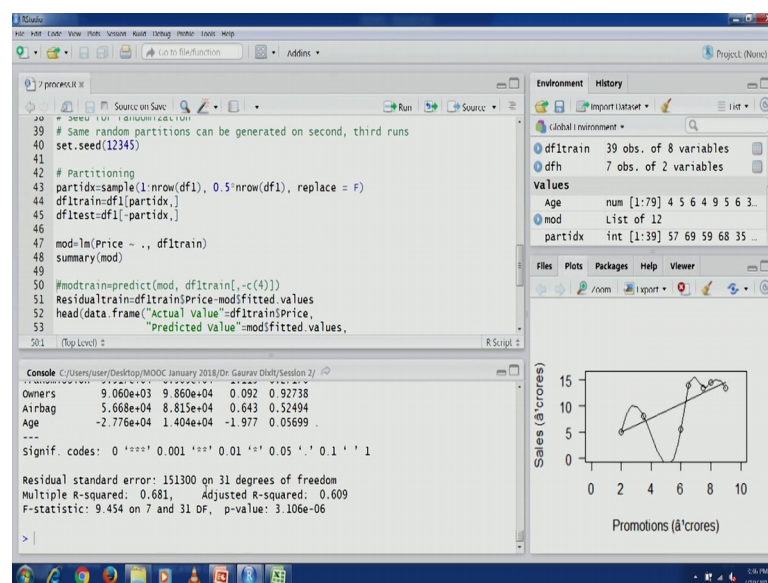
(Refer Slide Time: 29:48)



Now, let us focus on these coefficients parts you would see different predictors fuel type S R price K M transmission their estimates are given and you would also see that P values are given you would also see in the result phase that significance boards are also remained.

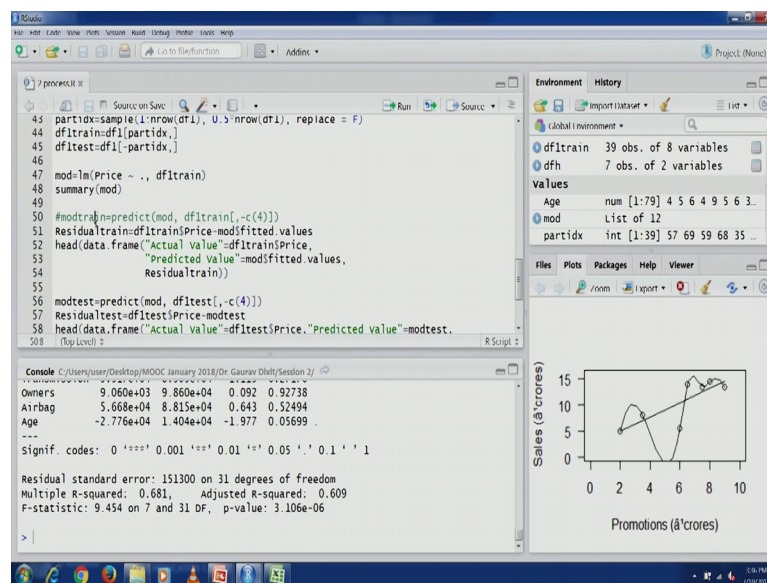
For example 0 for 100 percent kind of significance more than 99.9 percent significance 3 stars are used in 99 99.9 percent confidence interval 2 stars are used for then for 99 1 star and then for 95 dot is used. So, these are the notations in this case we c 3 less than say having 2 star and 1 star. So, would see the constant term and the fuel type and the showroom price you would see a dot in the age as well so these 4 variables. So, if we look at the main variables excluding and constant term fuel type S R price and the age of the variable, they are the main variable which are helping us in determining the offered price.

(Refer Slide Time: 31:04)



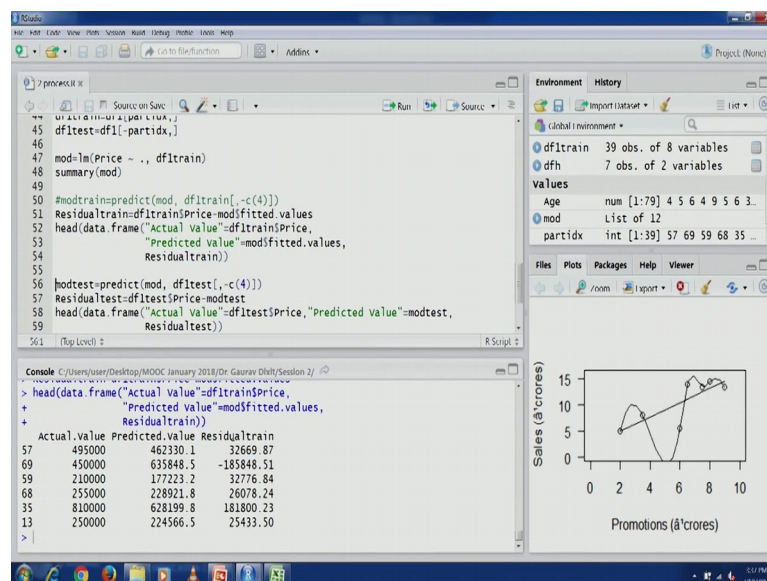
Other statistics related to this regression modelling are given for example, multiple R square and R square is given this seems to be 60 close to 61 percent, which is good enough. Now you would also see from the (Refer Time: 31:19) statistics that this particular model is significant. So, therefore, we can go ahead and interpret the results.

(Refer Slide Time: 31:28)



So, now let us look at how this model is going to perform over other partitions. So, let us first see. So, this mod fitted values this there are some values that are written from the lm function. So, one of them is being fitted values. So, we can compute the residuals using these fitted values more discussion on regression analysis we will do in a later lecture in this in this particular example we are just going through the data mining modelling process. So, let us compute the residuals let us look at the actual value predicted value and the error part.

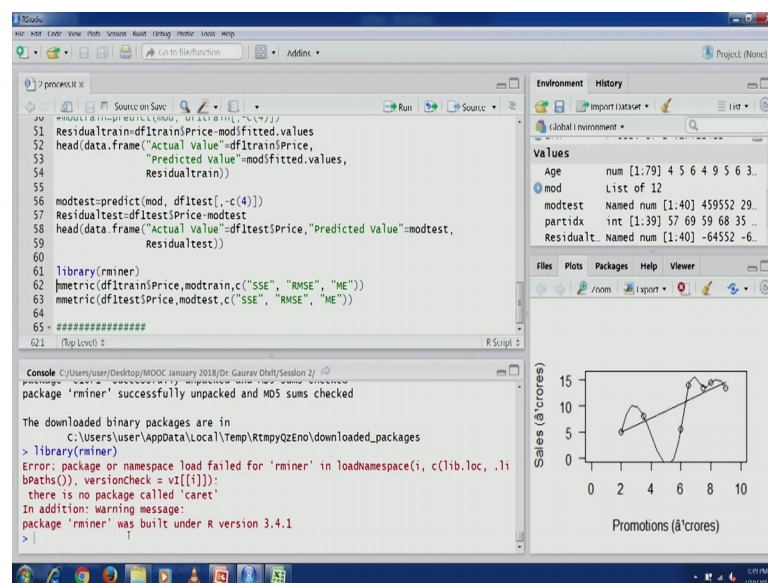
(Refer Slide Time: 32:14)



So, these numbers you can see now what was the actual value and what was the predicted value and what was the error these difference of these 2.

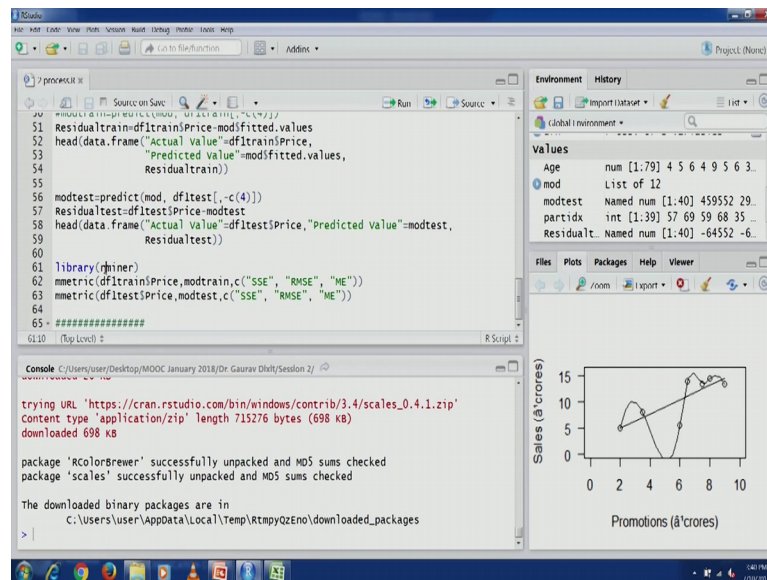
Now, similar thing can be done on the test partition those. So, we have predict function that can actually help us in scoring the test partition. So, in the predictive function first we need to pass on the model that is mod in this case and then the test partition is the second argument which has to be passed on. So, that we can find scoring do the scoring of test partition. So, let us execute this line.

(Refer Slide Time: 32:48)



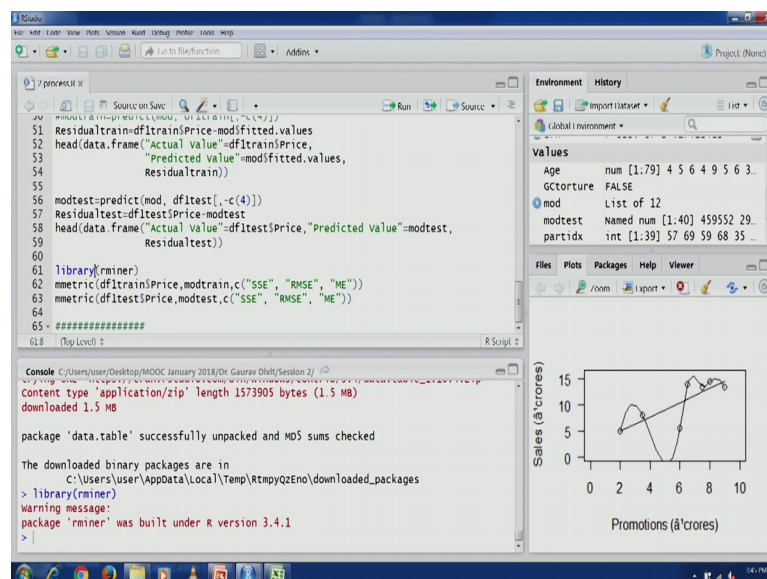
We would get the numbers let us again compute the residuals for test partition let us look at the numbers again the similar kind of output. Now we have another library that we need to load to see some of the metrics for evaluating the performance of the model. So, r miner is 1 particular r miner is 1 particular library that we need to load. So, let us install. So, we need to load this particular library r miner to be able to use some of the metrics.

(Refer Slide Time: 34:43)



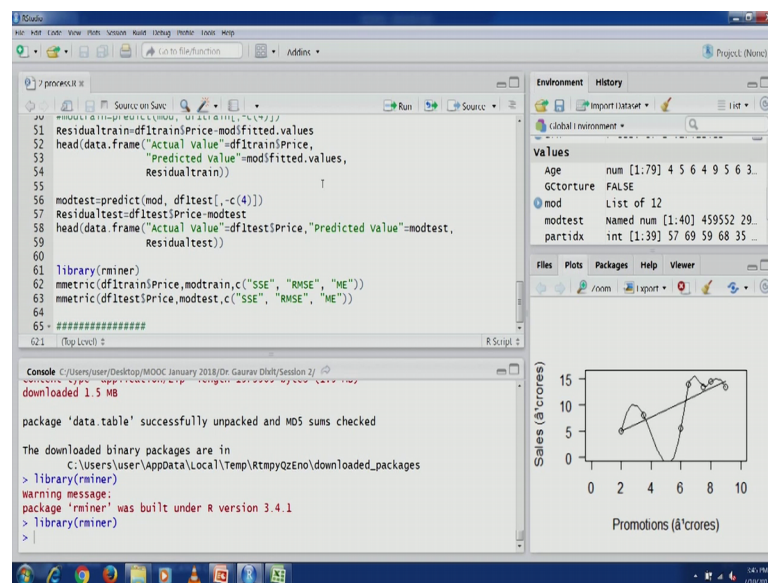
So, let us load this particular library r miner. So, that we are able to a able to compute some of the metrics. So, let us load this particular library r miner to be able to use some of the metrics for our model evaluation.

(Refer Slide Time: 36:04)



So, let us load this particular library R miner. So, that we have access to some of the metrics for our performance evaluation. So, let us load this.

(Refer Slide Time: 36:16)



Now, this is the function m metrics and we have first argument that we are passing is the price that is the actual value and then the a mod train the residual value and then we are going to compute these 2 these 3 metrics SSE RMSE and ME more discussion on metrics.

We will do in a later lecture. So, let us first compute this. So, we have this m metrics function in r miner that can be used to compute the metrics SSE RMSE and M E more discussion on this metrics we will do in a later lecture first argument is the price the actual value is the second argument is the fitted values. So, let us execute this particular code. So, these are the numbers we can see SSE RMSE and ME values there similarly we can compute for the a test partition again you would see the numbers there.

So, this is how the numbers of training partition and test partition then they can be compared and the performance of the model can then be assessed how well it is doing. So, we will do more discussion on this when we come to our regression analysis lecture.

Thank you.