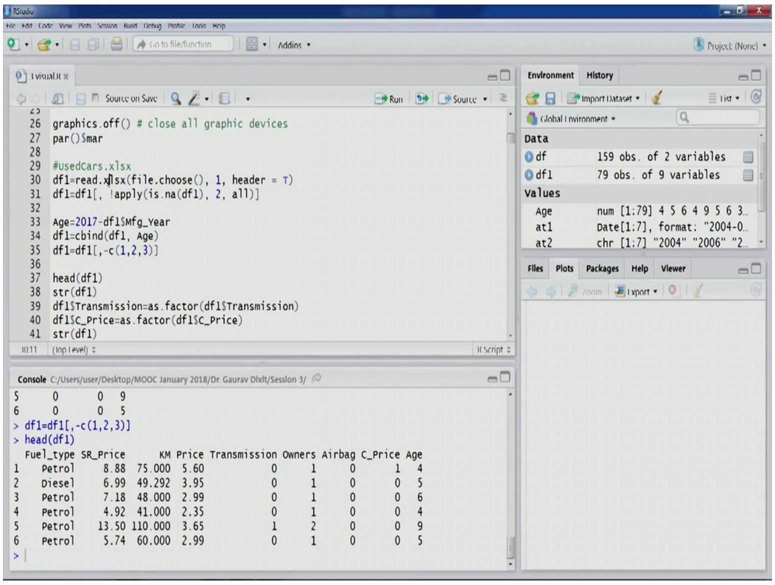
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 08
Visualization Techniques- Part II

Welcome to the course Business Analytics & Data Mining Modelling Using R. So, previous lecture we talked about visualization techniques. So, will continue our discussion from the same point where we left in this in that lecture. So, we were discussing bar plots and we were in this code we had imported this particular a data set used car. So, this has already remaining imported in data frame one.

(Refer Slide Time: 00:42)



```
graphics.off() # close all graphic devices
par()$mar

#UsedCars.xlsx
df1=read.xlsx(file.choose(), 1, header = T)
df1=df1[, !apply(is.na(df1), 2, all)]

Age=2017-df1$Mfg_Year
df1=cbind(df1, Age)
df1=df1[,-c(1,2,3)]

head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
5      0       0   9
6      0       0   5
> df1=df1[,-c(1,2,3)]
> head(df1)
  Fuel_type SR_Price      KM Price Transmission Owners Airbag C_Price Age
1    Petrol     8.88  75.000  5.60            0      1      0       1   4
2    Diesel     6.99  49.292  3.95            0      1      0       0   5
3    Petrol     7.18  48.000  2.99            0      1      0       0   6
4    Petrol     4.92  41.000  2.35            0      1      0       0   4
5    Petrol    13.50 110.000  3.65            1      2      0       0   9
6    Petrol     5.74  60.000  2.99            0      1      0       0   5
> 
```

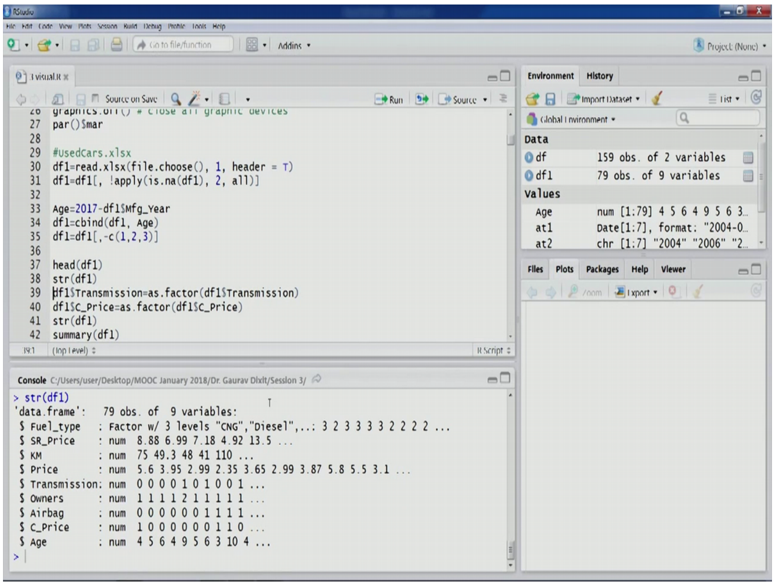
Data
df 159 obs. of 2 variables
df1 79 obs. of 9 variables
Values
Age num [1:79] 4 5 6 4 9 5 6 3...
at1 Date[1:7], format: "2004-0...
at2 chr [1:7] "2004" "2006" "2...

And we were also able to create this particular new variable age. So, this has been created using the a manufacturing year, wherever that was available in the data set then it was appended to the data frame. And then we had eliminated some of the variables which were not useful for our purpose.

So, after that we were left with this particular data frame. So, now, we have 9 variables you can see fuel; fuel type, then showroom price kilometres and price transmission, owners, air bags, see price and age.

So, then another command another useful command, another useful function that is available in r is S t r. So, we have discussed this a particular function before as well.

(Refer Slide Time: 01:37)



```
graphics.off() # close all graphic devices
par()$mar

#UsedCars.xlsx
df1=read.xlsx(file.choose(), 1, header = T)
df1=df1[, !apply(is.na(df1), 2, all)]

Age=2017-df1$Mfg_Year
df1=cbind(df1, Age)
df1=df1[,-c(1,2,3)]

head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
> str(df1)
'data.frame':   79 obs. of  9 variables:
 $ Fuel_type   : Factor w/ 3 levels "CNG","Diesel",..: 3 2 3 3 3 3 2 2 2 2 ...
 $ SR_Price    : num  8.88 6.99 7.18 4.92 13.5 ...
 $ KM          : num  75 49.3 48 41 110 ...
 $ Price       : num  5.6 3.95 2.99 2.35 3.65 2.99 3.87 5.8 5.5 3.1 ...
 $ Transmission: num  0 0 0 0 1 0 1 0 0 1 ...
 $ Owners      : num  1 1 1 1 2 1 1 1 1 1 ...
 $ Airbag      : num  0 0 0 0 0 0 1 1 1 1 ...
 $ C_Price     : num  1 0 0 0 0 0 0 1 1 0 ...
 $ Age         : num  4 5 6 4 9 5 6 3 10 4 ...
> 
```

```
Data
df      159 obs. of 2 variables
df1     79 obs. of 9 variables
Values
Age     num [1:79] 4 5 6 4 9 5 6 3…
at1     Date[1:7], format: "2004-0…
at2     chr [1:7] "2004" "2006" "2…
```

So, these actually help us understanding the structure of the data set different variables. So, you can see fuel price this is factor variable and with 3 labels being 3 label being CNG diesel and petrol.

Another important point that I would like to highlight here is the factor variable or categorical variable that are created in r**.** The labels would be in the alphabetical order. So, therefore, you would see that CNG has been displayed first. So, this has an impact on many functions that are available in r, wherein the CNG is taken as the a default category.

When we do we will start one of the formal analysis let say regression, then will come across some of these important peculiarities in r and different r function specifically with respect to factors.

Now, other variables you can see that showroom price, kilometre price, transmission owners, airbag all have been displayed as a miracle variable, but if you really look at the variable transmission and C_ Price.

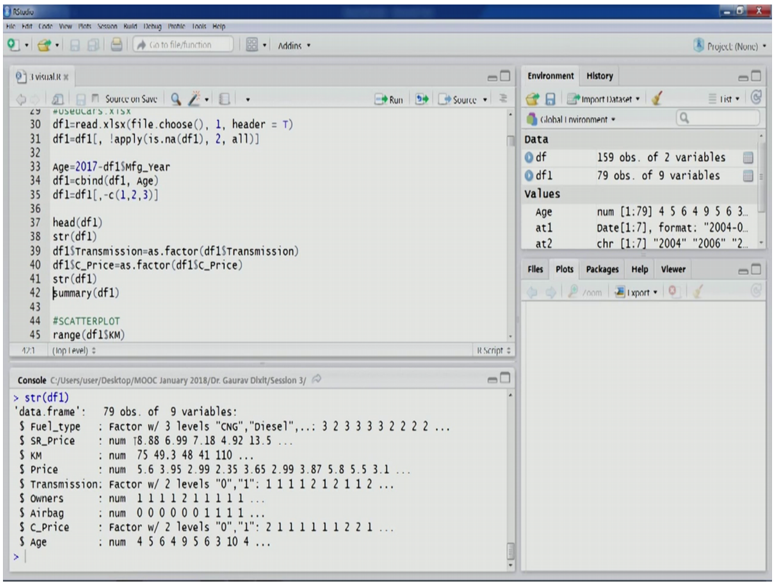
So, they are actually categorical in nature because transmission can have only 2 values that is 0 for 0 and 1 for automatic and manual. So, this is important. So, therefore, we need to convert this numeric variable into a factor variable.

Similarly a C a price that has been created by us manually as the as discussed in the previous lecture, were 1 was assigned for price amore than 4 lakhs equal to or more than 4 lakhs and for the cars having price as less than 4 lakh well assigned 0.

So, therefore, only 2 values are pass will 0 and 1. So, this variable also being categorical or factor variable therefore, we need to convert C-Price variable as well into factor variable. So, let us do that do this. So, as we have talked in a supplementary lectures hash dot factor is the command that can be used to course a numeric variable into factor variable or categorical variable. So, let us execute this particular line.

So, first transmission so, we run this line and then let us also run, C-Price.

(Refer Slide Time: 04:01)



```
#usedcars.xlsx
df1=read.xlsx(file.choose(), 1, header = T)
df1=df1[, !apply(is.na(df1), 2, all)]

Age=2017-df1$Mfg_Year
df1=cbind(df1, Age)
df1=df1[,-c(1,2,3)]

head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)

#SCATTERPLOT
range(df1$KM)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

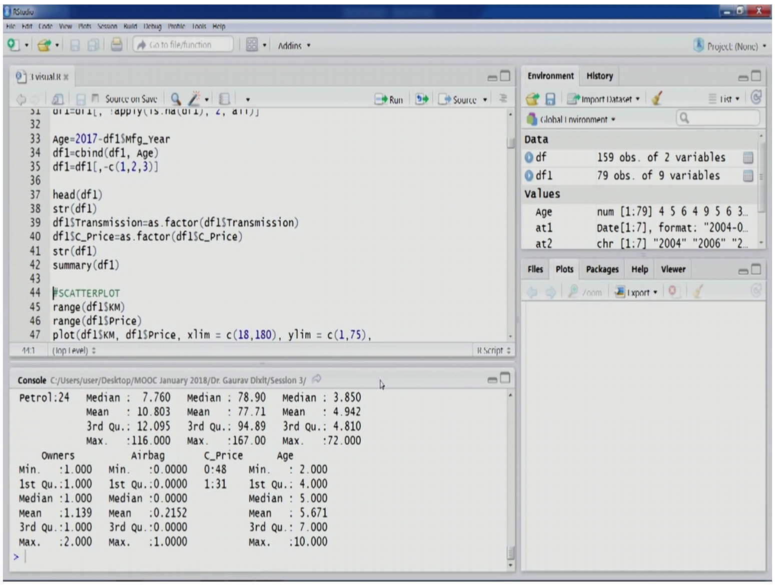
```
> str(df1)
'data.frame':   79 obs. of  9 variables:
 $ Fuel_type   : Factor w/ 3 levels "CNG","Diesel",..: 3 2 3 3 3 3 2 2 2 2 ...
 $ SR_Price    : num 18.88 6.99 7.18 4.92 13.5 ...
 $ KM          : num  75 49.3 48 41 110 ...
 $ Price       : num  5.6 3.95 2.99 2.35 3.65 2.99 3.87 5.8 5.5 3.1 ...
 $ Transmission: Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 1 2 ...
 $ Owners      : num  1 1 1 1 2 1 1 1 1 1 ...
 $ Airbag      : num  0 0 0 0 0 0 1 1 1 1 ...
 $ C_Price     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 2 2 1 ...
 $ Age         : num  4 5 6 4 9 5 6 3 10 4 ...
> 
```

```
Data
df       159 obs. of 2 variables
df1      79 obs. of 9 variables
Values
Age      num [1:79] 4 5 6 4 9 5 6 3…
at1      Date[1:7], format: "2004-0…
at2      chr [1:7] "2004" "2006" "2…
```

So, these 2 variables have been converted into factor variable. Let us look at the structure command as structure function again. Now you would see in the transmission you would see factor with 2 labels 0 and 1. They have been created. So, the variable converted into a factor variable, you would also see that C-Price has also been converted into a factor variable with 2 levels 0 and 1.

So, with this most of the variables in our imported data set used cars, most of the variables they are they are being you know presented in they are they are being used stored in their suitable a variable type.

(Refer Slide Time: 04:47)



```
df1=df1[, !apply(is.na(df1), 2, all)]

Age=2017-df1$Mfg_Year
df1=cbind(df1, Age)
df1=df1[,-c(1,2,3)]

head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)

#SCATTERPLOT
range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
```

Data
df 159 obs. of 2 variables
df1 79 obs. of 9 variables
Values
Age num [1:79] 4 5 6 4 9 5 6 3…
at1 Date[1:7], format: "2004-0…
at2 chr [1:7] "2004" "2006" "2…

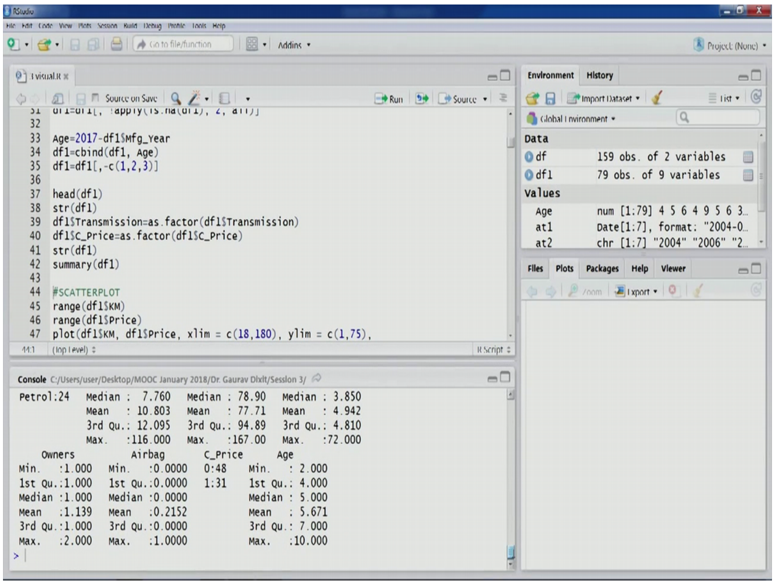
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
Petrol:24   Median :  7.760   Median : 78.90   Median : 3.850
            Mean   : 10.803   Mean   : 77.71   Mean   : 4.942
            3rd Qu.: 12.095   3rd Qu.: 94.89   3rd Qu.: 4.810
            Max.   :116.000   Max.   :167.00   Max.   :72.000
    Owners          Airbag        C_Price      Age
Min.   :1.000   Min.   :0.0000   0:48     Min.   : 2.000
1st Qu.:1.000   1st Qu.:0.0000   1:31     1st Qu.: 4.000
Median :1.000   Median :0.0000            Median : 5.000
Mean   :1.139   Mean   :0.2152            Mean   : 5.671
3rd Qu.:1.000   3rd Qu.:0.0000            3rd Qu.: 7.000
Max.   :2.000   Max.   :1.0000            Max.   :10.000
> 
```

So, now let us look at the summary results. So, summary results you can find out that that the categorical variable for example, fuel type you can see counts have been displayed for different categories for example, there are 3 only 3 records having CNG as fuel type 52 records or observation having diesel as fuel price and 24 records having petrol as the fuel type.

Similarly you can look at the transmission a sixty 3 records having a transmission 0 that is manual and 16 records having transmission as automatic that is 1 represented by 1.

(Refer Slide Time: 05:29)



```
df1=df1[, !apply(is.na(df1), 2, all)]

Age=2017-df1$Mfg_Year
df1=cbind(df1, Age)
df1=df1[,-c(1,2,3)]

head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)

#SCATTERPLOT
range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
```

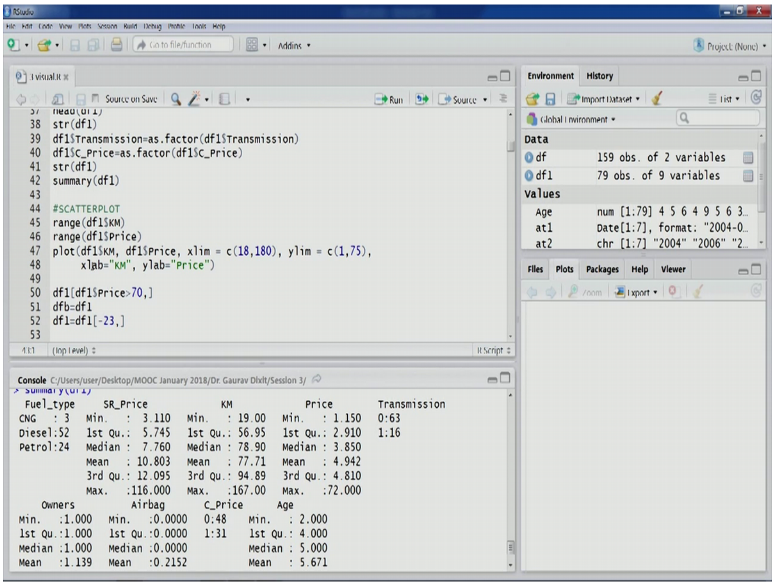
```
Petrol:24   Median :  7.760   Median : 78.90   Median : 3.850
            Mean   : 10.803   Mean   : 77.71   Mean   : 4.942
            3rd Qu.: 12.095   3rd Qu.: 94.89   3rd Qu.: 4.810
            Max.   :116.000   Max.   :167.00   Max.   :72.000
    Owners          Airbag        C_Price      Age
Min.   :1.000   Min.   :0.0000   0:48    Min.   : 2.000
1st Qu.:1.000   1st Qu.:0.0000   1:31    1st Qu.: 4.000
Median :1.000   Median :0.0000           Median : 5.000
Mean   :1.139   Mean   :0.2152           Mean   : 5.671
3rd Qu.:1.000   3rd Qu.:0.0000           3rd Qu.: 7.000
Max.   :2.000   Max.   :1.0000           Max.   :10.000
>
```

```
Data
df      159 obs. of 2 variables
df1     79 obs. of 9 variables
Values
Age     num [1:79] 4 5 6 4 9 5 6 3...
at1     Date[1:7], format: "2004-0...
at2     chr [1:7] "2004" "2006" "2...
```

Similarly, C-Price for label that also now being a factor variable or categorical variable 0 0 there are 48 cars, which are actually having a price value less than 4 lakhs rupees. And 31 car used cars are there, which are having price value more than equal to or more than 4 lakh rupees 4 lakhs rupees.

So, other variables are numerical in nature therefore, many descriptive statistics have been displayed for example, mean median a maths that we already understand. So, let us move to some of the basic plots.

(Refer Slide Time: 06:16)



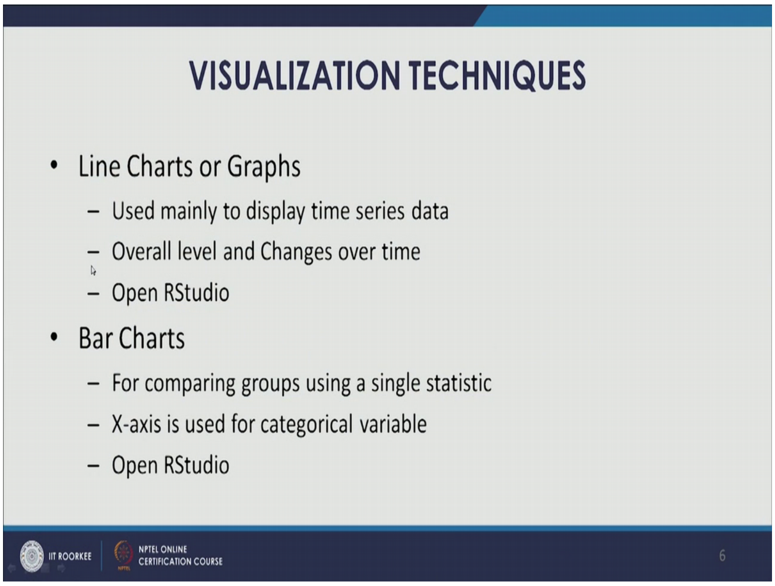
```
head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)

#SCATTERPLOT
range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
     xlab="KM", ylab="Price")

df1[df1$Price>70,]
dfb=df1
df1=df1[-23,]

```

```
> summary(df1)
  Fuel_type     SR_Price            KM             Price         Transmission
CNG   : 3   Min.   :  3.110   Min.   : 19.00   Min.   : 1.150   0:63
Diesel:52   1st Qu.:  5.745   1st Qu.: 56.95   1st Qu.: 2.910   1:16
Petrol:24   Median :  7.760   Median : 78.90   Median : 3.850
            Mean   : 10.803   Mean   : 77.71   Mean   : 4.942
            3rd Qu.: 12.095   3rd Qu.: 94.89   3rd Qu.: 4.810
            Max.   :116.000   Max.   :167.00   Max.   :72.000
    Owners          Airbag        C_Price      Age
Min.   :1.000   Min.   :0.0000   0:48    Min.   : 2.000
1st Qu.:1.000   1st Qu.:0.0000   1:31    1st Qu.: 4.000
Median :1.000   Median :0.0000           Median : 5.000
Mean   :1.139   Mean   :0.2152           Mean   : 5.671
```

```
Data
df      159 obs. of 2 variables
df1     79 obs. of 9 variables
Values
Age     num [1:79] 4 5 6 4 9 5 6 3...
at1     Date[1:7], format: "2004-0...
at2     chr [1:7] "2004" "2006" "2...
```

So, 2 basic plots that we want to cover that bar chart and this scatter plot. So, let us first start with scatter plot. So, let us go back to our slides and let us understand some of the key things about scatter plots.

(Refer Slide Time: 06:27)



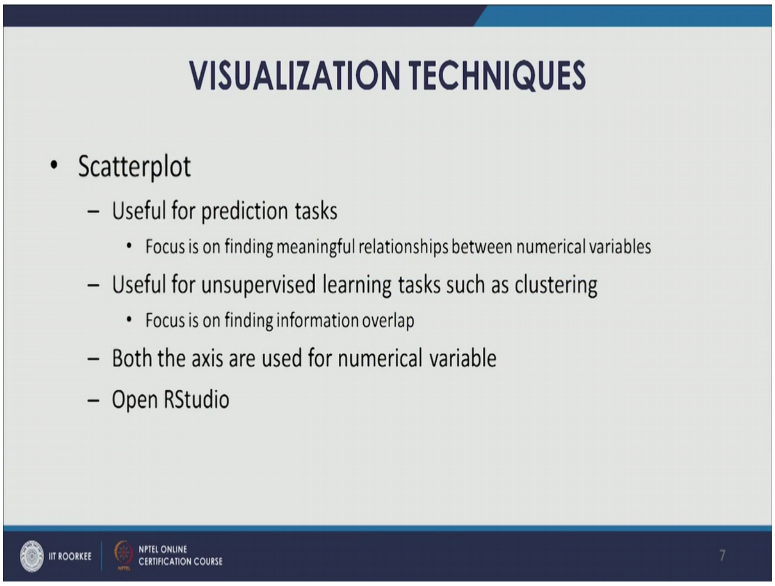
VISUALIZATION TECHNIQUES

- Line Charts or Graphs
  - Used mainly to display time series data
  - Overall level and Changes over time
  - Open RStudio
- Bar Charts
  - For comparing groups using a single statistic
  - X-axis is used for categorical variable
  - Open RStudio

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
6

So, generally scatter plots they are mainly useful for prediction task. So, the focus when we say that prediction task and how is scatter plot could be useful with respect to prediction task, focusing on finding meaningful relationship between numerical variables.

(Refer Slide Time: 06:41)



VISUALIZATION TECHNIQUES

- Scatterplot
  - Useful for prediction tasks
    - Focus is on finding meaningful relationships between numerical variables
  - Useful for unsupervised learning tasks such as clustering
    - Focus is on finding information overlap
  - Both the axis are used for numerical variable
  - Open RStudio

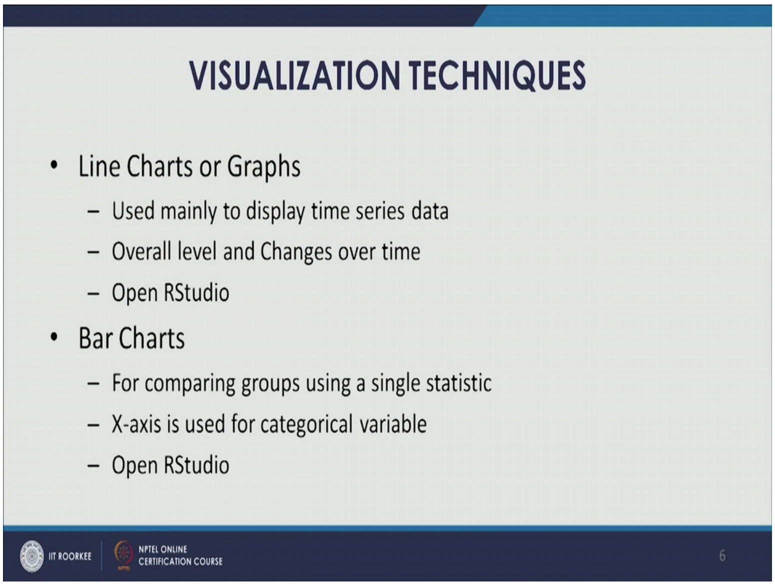
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
7

So, a scatter plot is mainly for numerical variable both the axis they are used for numerical variable both x axis and y axis, and for prediction task focus being identifying some meaningful relationship by from the plot. Now for unsupervised learning task is as clustering focus is on finding information overlap.

So, why this is useful that when we coming to our unsupervised learning lectures and start clustering our discussion on clustering will learn more about this, but focus being finding information overlap between different variables for unsupervised learning task and scatter plot can be useful in that.

Now both the axis as said are used for numeric variables in the bar charts x axis is used for categorical variable.

(Refer Slide Time: 07:36)



VISUALIZATION TECHNIQUES

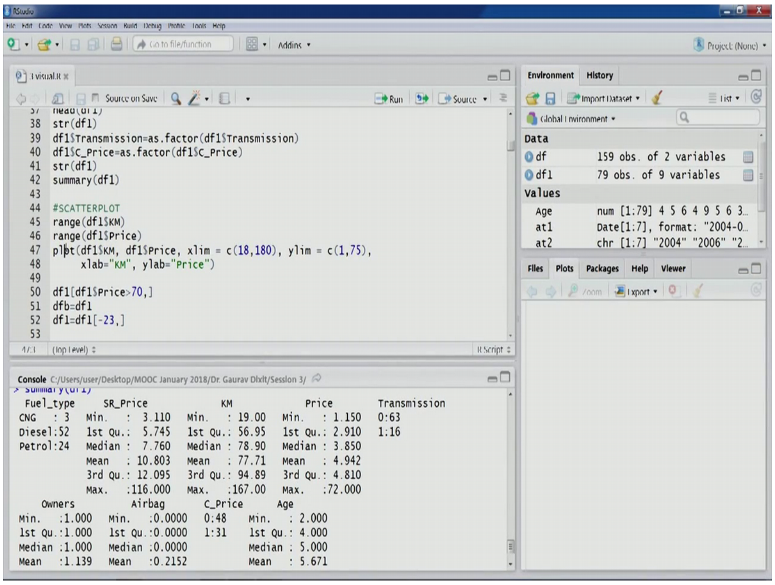
- Line Charts or Graphs
  - Used mainly to display time series data
  - Overall level and Changes over time
  - Open RStudio
- Bar Charts
  - For comparing groups using a single statistic
  - X-axis is used for categorical variable
  - Open RStudio

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
6

So, this is reserved for categorical variable and different groups because this being a categorical variable the variable on x axis is being categorical variable we can create different groups. So, there are going to be different categories and they can be different groups can be there and statistics can be displayed on y axis and that can help us understanding the differences between groups and compare them.

So, let us go back to our studio and let us start with scatter plot.

(Refer Slide Time: 08:10)



```
head(df1)
str(df1)
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)

#SCATTERPLOT
range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
     xlab="KM", ylab="Price")

df1[df1$Price>70,]
dfb=df1
df1=df1[-23,]

```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

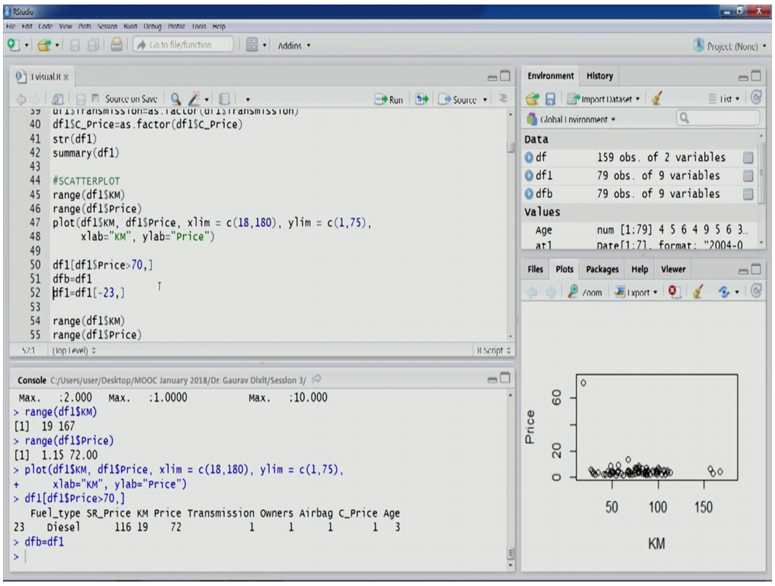
```
> summary(df1)
  Fuel_type     SR_Price           KM            Price         Transmission
 CNG   : 3   Min.   :  3.110   Min.   : 19.00   Min.   : 1.150   0:63
 Diesel:52   1st Qu.:  5.745   1st Qu.: 56.95   1st Qu.: 2.910   1:16
 Petrol:24   Median :  7.760   Median : 78.90   Median : 3.850
             Mean   : 10.803   Mean   : 77.71   Mean   : 4.942
             3rd Qu.: 12.095   3rd Qu.: 94.89   3rd Qu.: 4.810
             Max.   :116.000   Max.   :167.00   Max.   :72.000
     Owners          Airbag        C_Price      Age
 Min.   :1.000   Min.   :0.0000   0:48    Min.   : 2.000
 1st Qu.:1.000   1st Qu.:0.0000   1:31    1st Qu.: 4.000
 Median :1.000   Median :0.0000           Median : 5.000
 Mean   :1.139   Mean   :0.2152           Mean   : 5.671
```

Environment

```
Data
df          159 obs. of 2 variables
df1         79 obs. of 9 variables
Values
Age         num [1:79] 4 5 6 4 9 5 6 3...
at1         Date[1:7], format: "2004-0...
at2         chr [1:7] "2004" "2006" "2...
```

So, first plot that we are going to generate is between kilometre K M and price. So, before plotting because we need to specify limits on x axis and y axis so, that our plot looks much clear and we get more clear picture on the data. So, let us first understand the range of this variable. So, this range values of these kilometre and price is going to help us in the remaining the limits.

(Refer Slide Time: 08:40)



```
df1$Transmission=as.factor(df1$Transmission)
df1$C_Price=as.factor(df1$C_Price)
str(df1)
summary(df1)

#SCATTERPLOT
range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
     xlab="KM", ylab="Price")

df1[df1$Price>70,]
dfb=df1
df1=df1[-23,]

range(df1$KM)
range(df1$Price)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
 Max.   :2.000   Max.   :1.0000           Max.   :10.000
> range(df1$KM)
[1]  19 167
> range(df1$Price)
[1]  1.15 72.00
> plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
+      xlab="KM", ylab="Price")
> df1[df1$Price>70,]
   Fuel_type SR_Price KM Price Transmission Owners Airbag C_Price Age
23    Diesel      116 19    72            1      1      1       1   3
> dfb=df1
> 
```

Environment

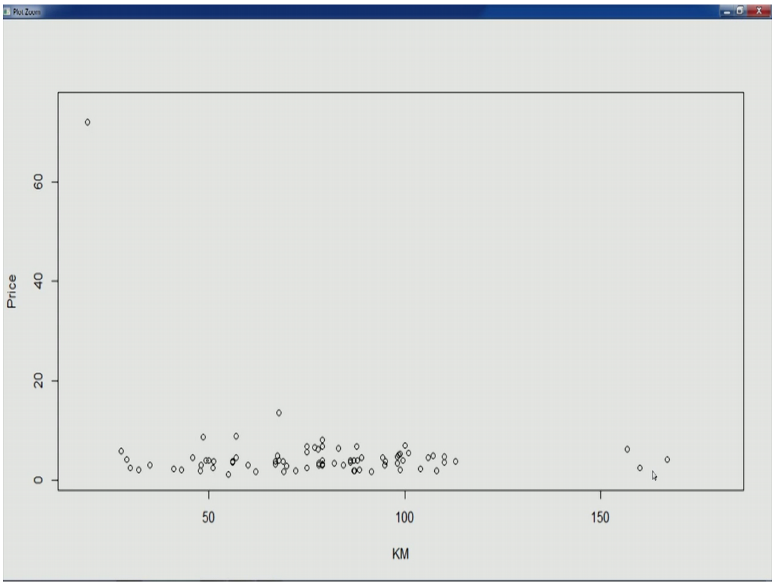
```
Data
df          159 obs. of 2 variables
df1         79 obs. of 9 variables
dfb         79 obs. of 9 variables
Values
Age         num [1:79] 4 5 6 4 9 5 6 3...
at1         Date[1:7], format: "2004-0...
```

So, you can see in the plot function you would see that for kilometre the range is between 19 and 167 and you can see the x limit that I have specified there is 18 and 180.

So, all the values are going to lie within this particular range similarly on the y axis the y limit that I have specified is 1 and 75 and you can look at the a price values. So, they are 1.15 and 72. So, therefore, these values are also going to be lie with within this range.

So, x axis is a kilometre is on the x axis and price is on the y axis you would had discussed before price is the outcome variable of interest in this particular data set and therefore, price has is being displayed on y axis. Label for x axis and y axis have been given appropriately using x lab and y lab arguments. So, we can run this particular code and you would see graph has been depicted.

(Refer Slide Time: 09:38)



Now, if we zoom into this particular graph you would see that there is 1 extreme out layer there. This particular where you kilometres accumulated they are far less they are far less than 25, 000 kilometres, but you would see that price that is offered it is a much higher. So, it is more than it is more than 70 in this case. So, the price is much much higher more than 70 lakhs that is.

So, but if you look at the other values of the majority of values are lying within 0 to 20 lakhs range, but this is the only out layer. So, from this we can understand that a most of the used cars that are they are in the data set they are in a smaller range.

Therefore, it would not be appropriate for us to study this stream out layer along with the these points. So, you know we have to restrict our analysis to this range as well to be it is it would be. So, we can eliminate this particular point and focus on this major chunk of points mainly lying between 0 to 20 in terms of price.

You would also see that in this plot some of the some of the points which all lying closer to x axis, but for you know far away from the 0 value and the majority of the values. So, they have more kilometres accumulated, but price offered for them is also in the same range between 0 to 20.

So, let us go back and here you would see first we are trying to identify we are trying to identify that particular out layer point, you would see I have given this price value greater than 70, you would see in the graph that this looks more than 70 price looks more than 70 lakhs. So, therefore, a let us run this particular code.

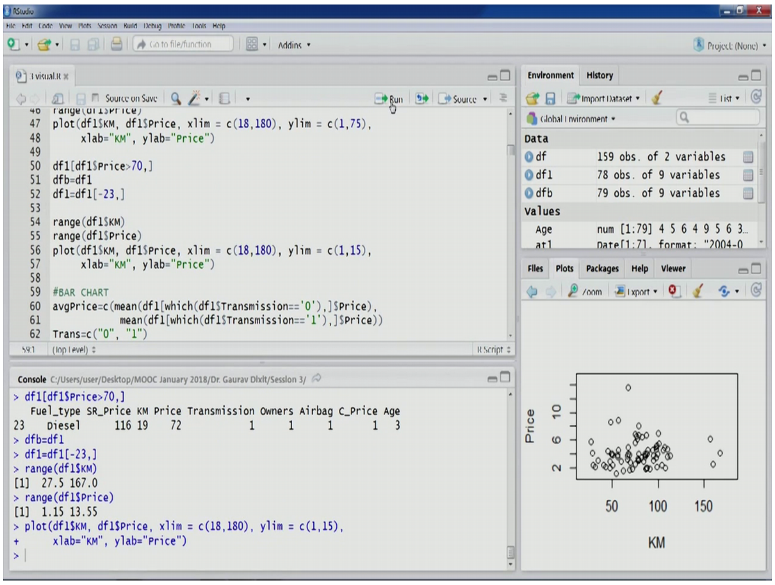
So, you would see this is point number 23. So, this is observation number 23 having fuel type a diesel and showroom price 100 and 16 lakhs. And then you would see that showroom price is 100 16 lakhs and the offered price is 72. So, these are high numbers in comparison to the majority of other observation.

So, we can get rid of this particular observation because this seems to be this can be considered a very distinct group. So, let us take a backup of previous data frame and now eliminate this particular point. Now this is how we can eliminate this point in the data frame we can use these brackets and the point number that index, we can specify as minus 23 minus will give this instruction that this point is to be removed from the data, frame and the data frame would be again stored in the same.

So, let us execute this or this particular data frame is gone. Now let us again lettle let us again have a look at the range values max and min values for both these variables kilometre and price and let us again re plot the graphic.

So, you would see that kilometres there is not much change price then here you would see some change for example, earlier values was were ranging from 1.15 to 72.

(Refer Slide Time: 13:26)



```
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,75),
     xlab="KM", ylab="Price")

df1[df1$Price>70,]
dfb=df1
df1=df1[-23,]

range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,15),
     xlab="KM", ylab="Price")

#BAR CHART
avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
           mean(df1[which(df1$Transmission=='1'),]$Price))
Trans=c("0", "1")
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

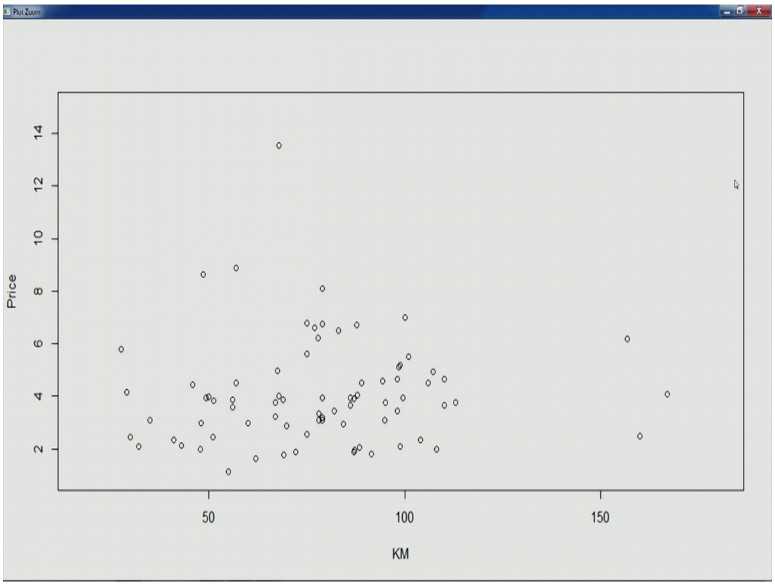
```
> df1[df1$Price>70,]
  Fuel_type SR_Price KM Price Transmission Owners Airbag C_Price Age
23    Diesel      116 19    72            1      1      1       1   3
> dfb=df1
> df1=df1[-23,]
> range(df1$KM)
[1]  27.5 167.0
> range(df1$Price)
[1]  1.15 13.55
> plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,15),
+      xlab="KM", ylab="Price")
> |
```

```
Data
df      159 obs. of 2 variables
df1      78 obs. of 9 variables
dfb      79 obs. of 9 variables
Values
Age     num [1:79] 4 5 6 4 9 5 6 3…
at1     Date[1:71], format: "2004-0…
```

Now, this is ranging from 1.15 to 13.5 5. So, even less than 15.

So, the price is now less than 15 lakhs, now for a kilometres also now this new range is 27.5 to 167 earlier range was 19. So, you would see that kilometre range has also you know increased specifically for the minimum value. Now let us plot this now the x limit and y limit values have been appropriately changed modified would see 18 180 and 1 and 15 now instead of 75 earlier case, on it is plot this now this is the new plot that we have.

(Refer Slide Time: 14:09)



Now, you would see that the graphic is covering most of the points in a clear fashion. So, most of the points are covered in the graphic now. If we try to understand some of the relationship between these 2 a numerical variable kilometre and price you would see not much change, you we can have a constant line that is going somewhere at a price value of 4 lakhs and it can be a constant line.

So, it seems from this particular data points if we fit this particular these data points into a you know linear model then you would see that kilometres is not a important factor. So, the price is being offered irrespective of the kilometre. So, that is the kind of sense that we get from the data. So, this data can be restart by a horizontal line. So, therefore, kilometre is not a is not such a crucial variable in our analysis specifically focusing on predicts and task related to price.

So, these are some of the insights that we can get from these basic plots. For example, relationship between price and kilometres we can see that kilometres KM might not be such a useful indicator for offered price, respectfully this is what we get from this is, what we gather from this particular data set.

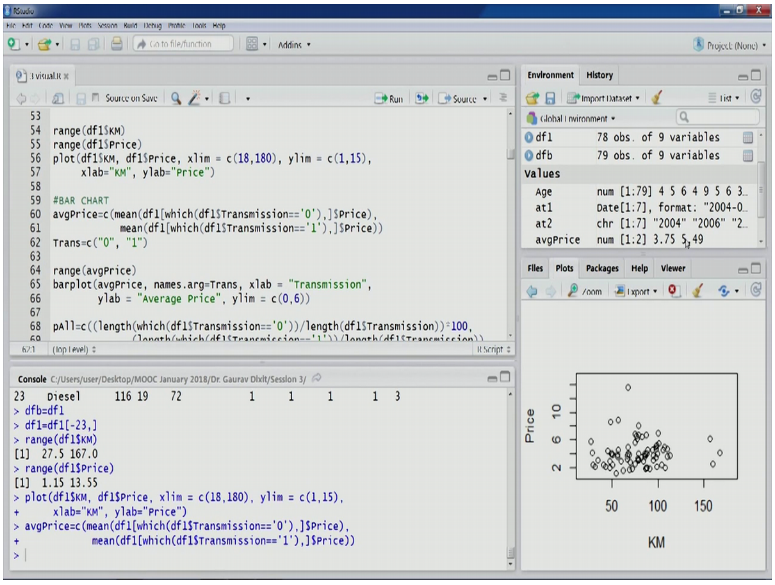
Now let us move to next basic chart that is bar chart. So, would see the bar chart we want to plot between price and transmission the transmission is the factor variable that we have already created and we can compute average price for different groups. So, we can get 2 groups based on transmission value. So, 1 is a 0 that is manual and the another transmission value could be 1 that is for automatic.

So, these 2 groups manual and automatic and average price value for these 2 groups can we can compute using this particular line up code, you can see I am trying to compute mean from the values. So, which is another function which can return more information on which you can find from the help section, but to give you a sense of this particular function, which can find out the indices of the observations were transmission value is transmission value is equal to 0.

So, those indices would be return and then the those particular observation would be retrieved or selected or subset or subset would be created for us to pass to the mean function and to calculate mean of that. And mean is again dollar notation indicating that mean is to be computed only for one variable that is price.

Similarly, for another group the same thing can be performed. So, let us compute average price you would see that average price a numerical variable has been created and there are just 2 values, these values corresponding to 2 different groups group, 0 that is go manual and group one that is for automatic car.

(Refer Slide Time: 17:15)



```
range(df1$KM)
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,15),
     xlab="KM", ylab="Price")

#BAR CHART
avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
           mean(df1[which(df1$Transmission=='1'),]$Price))
Trans=c("0", "1")

range(avgPrice)
barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
        ylab = "Average Price", ylim = c(0,6))

pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
       (length(which(df1$Transmission=='1'))/length(df1$Transmission))
```

```
23    Diesel      116 19    72              1       1       1       1   3
> dfb=df1
> df1=df1[-23,]
> range(df1$KM)
[1]  27.5 167.0
> range(df1$Price)
[1]  1.15 13.55
> plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,15),
+      xlab="KM", ylab="Price")
> avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
+            mean(df1[which(df1$Transmission=='1'),]$Price))
> |
```

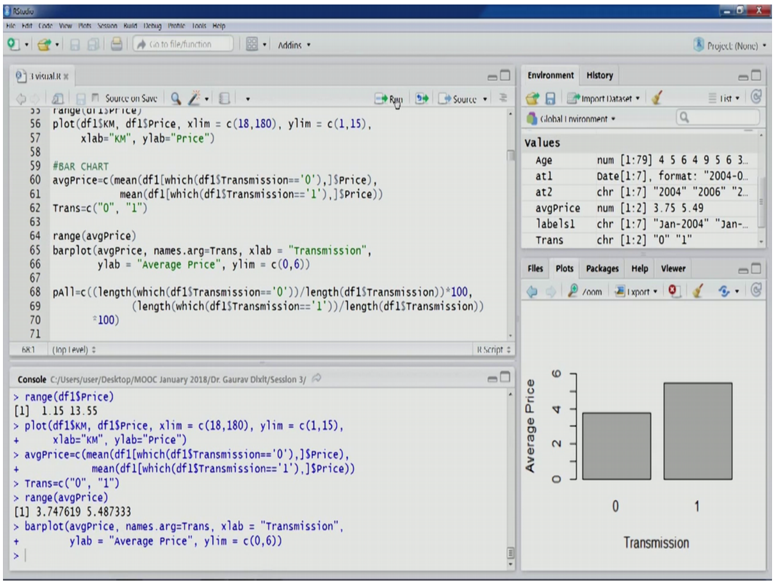
```
df1         78 obs. of 9 variables
dfb         79 obs. of 9 variables
Values
  Age       num [1:79] 4 5 6 4 9 5 6 3...
  at1       Date[1:7], format: "2004-0...
  at2       chr [1:7] "2004" "2006" "2...
  avgPrice  num [1:2] 3.75 5.49
```

So, these 2 mean values have been computed. Now another variable that we want to create before generating bar plot Trans is the name of the variable. So, this is just for the labelling purpose. So, the transmission the labels that we are going to use in our plot so 0 and 1, that are the names of the label for 2 different groups.

We could have we could have used a you know manual and automatic as well here. So, those could have also been the label names for our bar plot let us go with 0 and 1 right now. Now, because the this particular bar plot is between average price the variable we have just created and then a trans. So, let us look at the range; range is between 3.7 4 to 5.4 8. Now if you look at the plot bar plot function that I have written here the y limit is between 0 and 6. So, these in this particular range would be covered in this.

So, average price is the first variable. So, this will go into the y axis and the names dot r here this will go into the x axis and the labels for x axis that we have just created using transmission, x lab name is transmission and y lab y axis label is average price. So, let us execute this line.

(Refer Slide Time: 18:48)



```
range(df1$Price)
plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,15),
     xlab="KM", ylab="Price")

#BAR CHART
avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
             mean(df1[which(df1$Transmission=='1'),]$Price))
Trans=c("0", "1")

range(avgPrice)
barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
        ylab = "Average Price", ylim = c(0,6))

pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
               (length(which(df1$Transmission=='1'))/length(df1$Transmission))
        *100)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

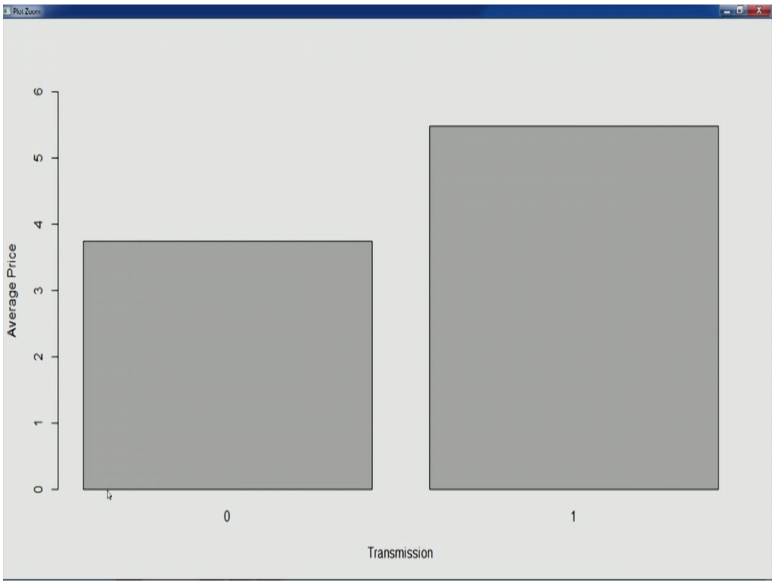
```
> range(df1$Price)
[1]  1.15 13.55
> plot(df1$KM, df1$Price, xlim = c(18,180), ylim = c(1,15),
+      xlab="KM", ylab="Price")
> avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
+              mean(df1[which(df1$Transmission=='1'),]$Price))
> Trans=c("0", "1")
> range(avgPrice)
[1] 3.747619 5.487333
> barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
+         ylab = "Average Price", ylim = c(0,6))
> |
```

Values

Age	num [1:79] 4 5 6 4 9 5 6 3…
at1	Date[1:7], format: "2004-0…
at2	chr [1:7] "2004" "2006" "2…
avgPrice	num [1:2] 3.75 5.49
labels1	chr [1:7] "Jan-2004" "Jan-…
Trans	chr [1:2] "0" "1"

And this particular bar plot has been created.

(Refer Slide Time: 18:53)



You can see this for group 0 that is the cars used cars you know manual with manual transmission, you can see their average price is somewhere between 3.5 and 4. And for group 1 that is the used cars with automatic transmission, their average price is ranging somewhere between 5 and 5.5.

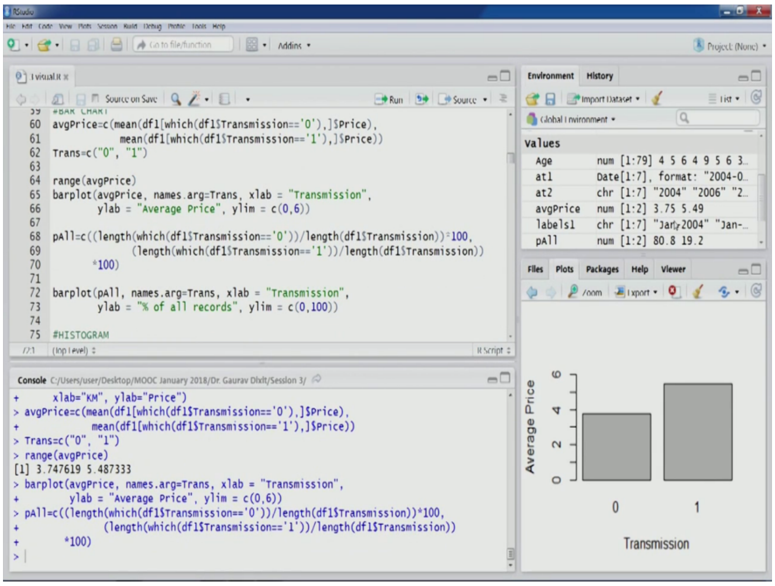
So, this kind of information, so, it seems that the cars automatic cars with auto automatic transmission they seem to be carrying more value. Now let us create another bar plot. So, this time this time we are going to use only the 1 variable. So, the this previous plot that we created. So, we had 1 numerical variable on y axis and 1 categorical variable on x x axis.

Now let us just focus on 1 variable and that has to be categorical. So, it is going to be on x axis again. So, this variable is again transmission. So, what we are trying to find out is the number of cars, which are manual and the number of cars, which are automatic the percentage. So, percentage of all records percentage of all records that we want to find out which are manual and which are how many are manual and how many are automatic.

So, this is the code that we can that we can actually use to compute this. So, can see length. So, I am trying to find out the length of a vector which is going to be determined by this rich command. So, rich command is going to return the indices for transmission were it is 0.

So, all those indices they would be counted using the length function and will get the number of records. And you would see this has been divided by divided by the all the all records in the vector transmission, that can be computed using again using the length function and passing on the argument transmission. So, this would be the ratio and then multiplied by 100. So, this will create a percentage number percentage value similarly for group 1 that is for automatic cars we can do this. So, let us execute this line.

(Refer Slide Time: 21:28)



```
#BAR CHART
avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
           mean(df1[which(df1$Transmission=='1'),]$Price))
Trans=c("0", "1")

range(avgPrice)
barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
        ylab = "Average Price", ylim = c(0,6))

pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
            (length(which(df1$Transmission=='1'))/length(df1$Transmission))
       *100)

barplot(pAll, names.arg=Trans, xlab = "Transmission",
        ylab = "% of all records", ylim = c(0,100))

#HISTOGRAM
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
+      xlab="KM", ylab="Price")
> avgPrice=c(mean(df1[which(df1$Transmission=='0'),]$Price),
+            mean(df1[which(df1$Transmission=='1'),]$Price))
> Trans=c("0", "1")
> range(avgPrice)
[1] 3.747619 5.487333
> barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
+         ylab = "Average Price", ylim = c(0,6))
> pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
+             (length(which(df1$Transmission=='1'))/length(df1$Transmission))
+       *100)
> 
```

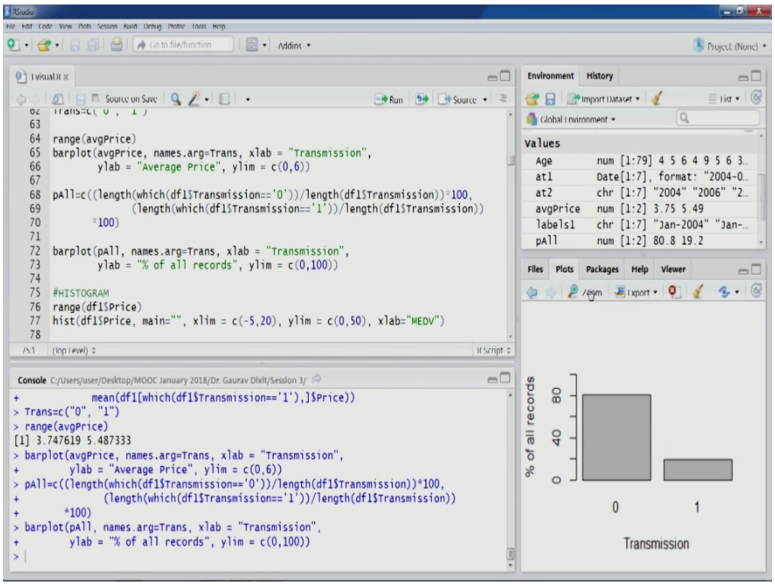
Values

```
Age        num [1:79] 4 5 6 4 9 5 6 3...
at1        Date[1:7], format: "2004-0...
at2        chr [1:7] "2004" "2006" "2...
avgPrice   num [1:2] 3.75 5.49
labels1    chr [1:7] "Jan-2004" "Jan-...
pAll       num [1:2] 80.8 19.2
```

You can see variable pAll has been created and there are 2 values 80.8. So, 80.8 percent is of records they actually belong to group 0 that is manual and 19.2 records they belong to group 1 that is for automatic cars.

So, let us generate a bar plot you can see pAll is the variable and the arguments Trans and limit. Now in this case is 0 100 because we are using percentage. So, that is the range standard range. So, let us create this plot.

(Refer Slide Time: 22:07)



```
Trans=c("0", "1")

range(avgPrice)
barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
        ylab = "Average Price", ylim = c(0,6))

pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
           (length(which(df1$Transmission=='1'))/length(df1$Transmission))
       *100)

barplot(pAll, names.arg=Trans, xlab = "Transmission",
        ylab = "% of all records", ylim = c(0,100))

#HISTOGRAM
range(df1$Price)
hist(df1$Price, main="", xlim = c(-5,20), ylim = c(0,50), xlab="MEDV")

```

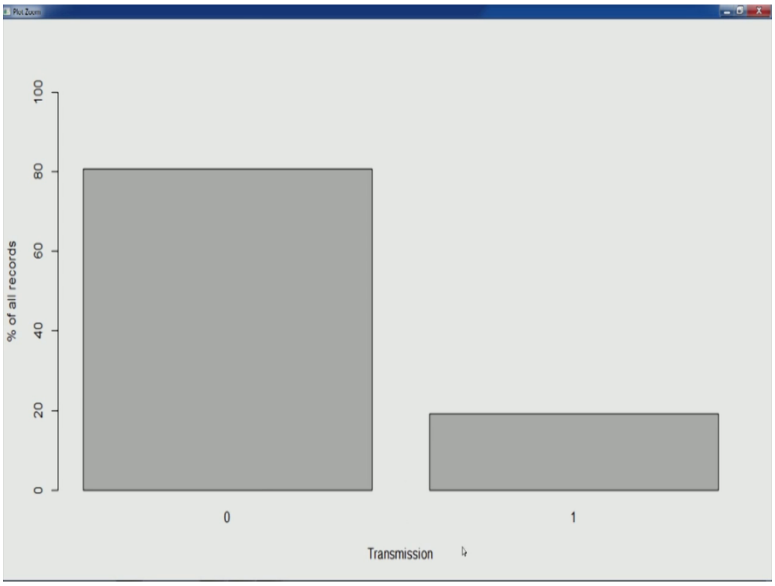
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
+             mean(df1[which(df1$Transmission=='1'),]$Price))
> Trans=c("0", "1")
> range(avgPrice)
[1] 3.747619 5.487333
> barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
+         ylab = "Average Price", ylim = c(0,6))
> pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
+            (length(which(df1$Transmission=='1'))/length(df1$Transmission))
+        *100)
> barplot(pAll, names.arg=Trans, xlab = "Transmission",
+         ylab = "% of all records", ylim = c(0,100))
> 
```

```
Values
  Age         num [1:79] 4 5 6 4 9 5 6 3...
  at1         Date[1:7], format: "2004-0...
  at2         chr [1:7] "2004" "2006" "2...
  avgPrice    num [1:2] 3.75 5.49
  labels1     chr [1:7] "Jan-2004" "Jan-...
  pAll        num [1:2] 80.8 19.2
```

And see the plot let us zoom in.

(Refer Slide Time: 22:13)



So, can see x axis transmission 2 groups 0 and 1 and in on, y axis percentage of all records and you can see 0 this is close to 80 and you can see 1, group 1, which is closer to 20. So, this kind of a using 1 categorical variable that is mainly a transmission we can also create these kind of plots.

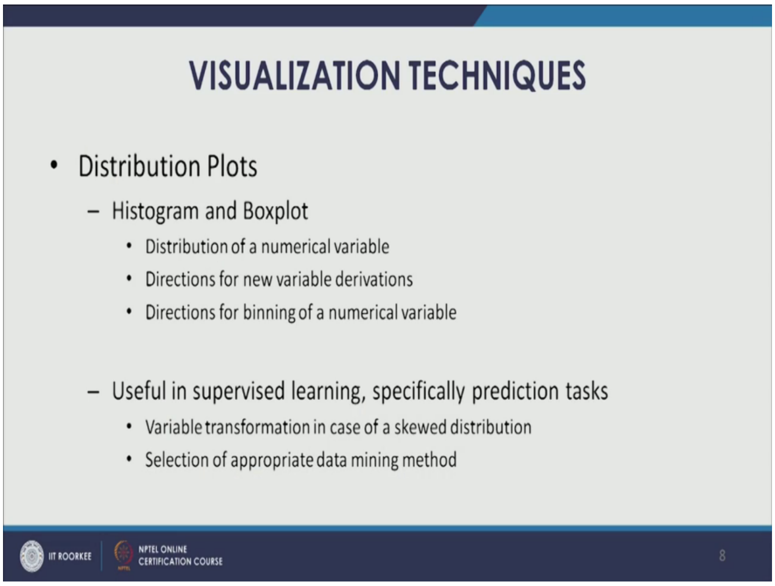
So, this again in a way help us in understanding the structure of the data from this we can understand that most majority of the cars, more than image ultimo it is around 80 percent most of the cars. They are manual and only 20 percent of the cars if smaller numbers of cars smaller percentage of cars are actually automatic.

So, therefore, this gives as a idea about the this particular structure for example, if it was less than 5 percent that then in that case it could have it could be you know defined as a rare category rare class you know the automatic.

So, therefore, this might have affected our formal analysis. So, these are kind of graphics an actually help us understanding some of the insights about the data and help us later on in formal analysis.

Now, let us go back to our slides let us discuss a next set of plots. So, next set of plots are actually distribution plots.

(Refer Slide Time: 23:37)



VISUALIZATION TECHNIQUES

- Distribution Plots
  - Histogram and Boxplot
    - Distribution of a numerical variable
    - Directions for new variable derivations
    - Directions for binning of a numerical variable
  - Useful in supervised learning, specifically prediction tasks
    - Variable transformation in case of a skewed distribution
    - Selection of appropriate data mining method

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 8

So, mainly 2, 2 distribution plots that we are going to cover in in this lecture in this course a mainly 2 being 1 first 1 is histogram the second 1 being box plot. So, as the name says these are distribution plots and help us understand the distribution of data. So, because this is distribution so, generally they are applicable to numerical variable.

So, we are interested we might be sometimes we might be interested in understanding the distribution of a numerical variable. This could be various reason for various reason that be will keep on learning as we go long. Once we understand the distribution of data we can we can some of the we can touch, the we can understand, we can verify some of the assumptions that are there mainly in statistical techniques.

So, distribution can help us for example, understanding whether the data is following normal distribution or not. So, therefore, if it is not following normal distribution what can be done?

So, these some of these plots are going to help us in the fashion. So, sometimes we might be required to transform those variables. So, that we are able to achieve the normal plot normal distribution. Sometimes if we want to convert a numeric variable into a categorical variable so, histogram and box plot the entire distribution that they display that can help us in terms of binning of those variables; how the bins are groups are to be created. So, those insights again help us in creating new variables.

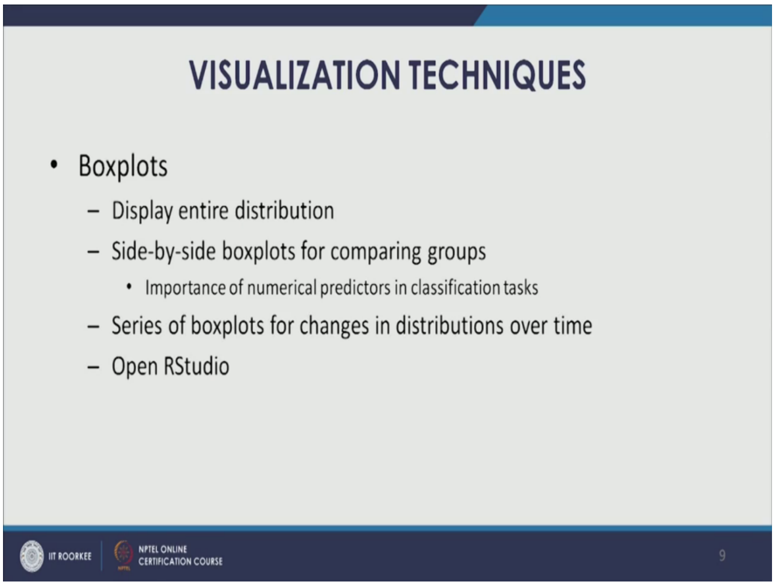
So, as shown in this slide histogram and box plot they are about distribution of a numerical variable, we get directions for new variable derivation as we discussed and we also get directions for binning of a numerical variable. Now useful in supervised learning is specifically in predict prediction task, because they are mainly applicable for numerical variable and therefore, prediction task doing the important type for this kind of plots where we can actually get some help for example, variable transformation in in case of a skewed distribution.

So, will learn more about the skew in coming lectures as well and in this lecture as well so, there could be a right skewed distribution or left skewed distribution. So, what are the transformation that can be done to achieve to actually reduce some of this skewness of the of the of the plot of the data. So, and so that can actually be find out that can be actually be shown in use in histogram and box plot.

Selection of appropriate data mining method for example, if the a data set is not able to follow or not able to meet some of the assumptions in a statistical technique. Then probably, we cannot apply them and probably, we have to go with some of the data remain technique some of the data mining technique. Because they are some of these youngsters are relaxed and those techniques can always be applied.

So, selection of appropriate method or technique can also be done using these plots. Now further discussion on box plot. So, box plot they display entire distribution.

(Refer Slide Time: 26:52)



VISUALIZATION TECHNIQUES

- Boxplots
  - Display entire distribution
  - Side-by-side boxplots for comparing groups
    - Importance of numerical predictors in classification tasks
  - Series of boxplots for changes in distributions over time
  - Open RStudio

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
9

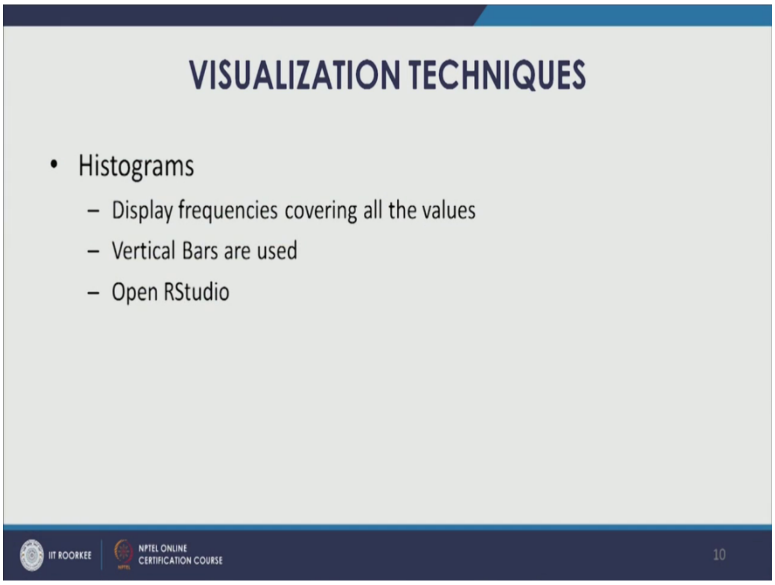
So, till now for example, whatever bar charts that we plotted they were focused on you know 1 or 2 groups or categories that were there in the categorical variable and the numerical values for the same were you know reflected in the y axis. So, that kind of information that that we could get from the bar plot, but in the box plot we get the whole entire distribution the whole range of values are covered. So, therefore, we can have a better look in the on the whole data the full data.

Now, there is another thing that can be done side by side box plots. So, we can create side by side box plot that can again help us in comparing and understanding the difference between groups, something that we did using bar plots that can also be done using box plot and in a more in a much better fashion.

So, this could be useful in classification task where we can understand the importance on numerical predictors. So, in classification task we are using some numerical predictors. So, this side by side box plot can actually help us in finding out how these numerical variables, numerical predictors, can be best utilised and their importance as well. Another usage of box plot could be in the you know time series kind of analysis where we can have series of box plot and we can look at changes in distribution over time.

So, that can also be done. So, let us open r studio and will go through an example go through examples for box plot and histogram. So, let us also cover histogram as well before we going to before will go through examples together for both of these kind of plots.

(Refer Slide Time: 28:47)



VISUALIZATION TECHNIQUES

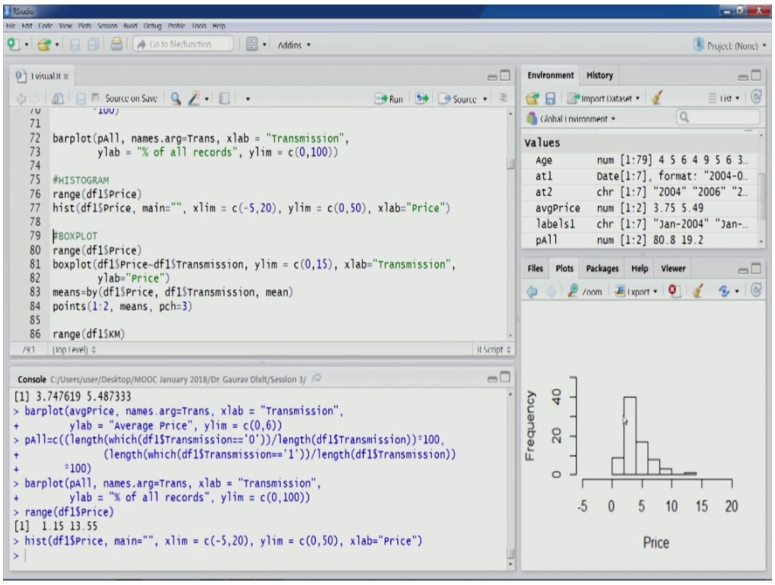
- Histograms
  - Display frequencies covering all the values
  - Vertical Bars are used
  - Open RStudio

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
10

So, histograms; histograms, they generally display a frequencies covering all the values. So, in the bar plots only few values are actually covered. So, in this case histograms we cover all the values and vertical bars are used more we will learn through r studio.

So, again we are going to use the same data used cars data. So, you can see that histories the function that can be used to plot a histogram, you can see because we are interested in this particular variable price that is our outcome variable. So, let us see the range. So, range is same as we saw here 1.1 5 to 13.5 5.

(Refer Slide Time: 29:27)



```
        *100)

barplot(pAll, names.arg=Trans, xlab = "Transmission",
        ylab = "% of all records", ylim = c(0,100))

#HISTOGRAM
range(df1$Price)
hist(df1$Price, main="", xlim = c(-5,20), ylim = c(0,50), xlab="Price")

#BOXPLOT
range(df1$Price)
boxplot(df1$Price~df1$Transmission, ylim = c(0,15), xlab="Transmission",
        ylab="Price")
means=by(df1$Price, df1$Transmission, mean)
points(1:2, means, pch=3)

range(df1$KM)
```

Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

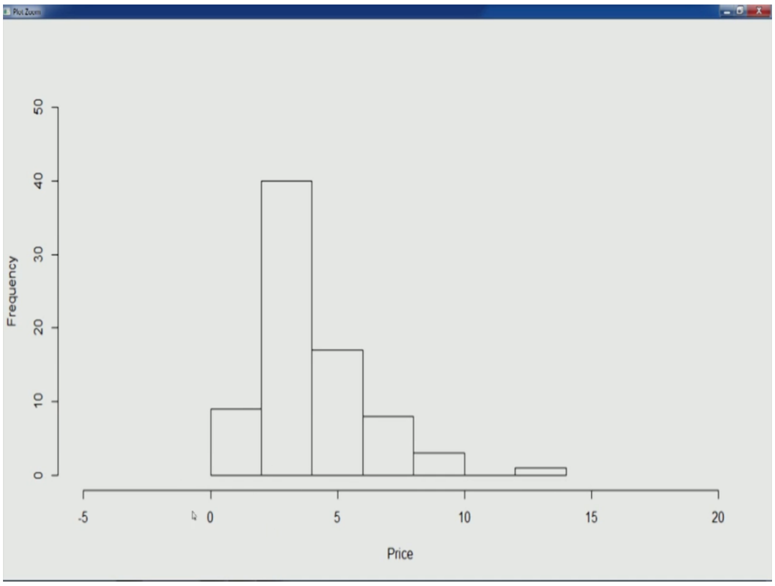
```
[1] 3.747619 5.487333
> barplot(avgPrice, names.arg=Trans, xlab = "Transmission",
+         ylab = "Average Price", ylim = c(0,6))
> pAll=c((length(which(df1$Transmission=='0'))/length(df1$Transmission))*100,
+              (length(which(df1$Transmission=='1'))/length(df1$Transmission))
+        *100)
> barplot(pAll, names.arg=Trans, xlab = "Transmission",
+         ylab = "% of all records", ylim = c(0,100))
> range(df1$Price)
[1]  1.15 13.55
> hist(df1$Price, main="", xlim = c(-5,20), ylim = c(0,50), xlab="Price")
> |
```

```
Values
  Age          num [1:79] 4 5 6 4 9 5 6 3…
  at1          Date[1:7], format: "2004-0…
  at2          chr [1:7] "2004" "2006" "2…
  avgPrice     num [1:2] 3.75 5.49
  labels1      chr [1:7] "Jan-2004" "Jan-…
  pAll         num [1:2] 80.8 19.2
```

So, let us so you can see that limit minus 5 to 20 has been used, why this slightly wider limit will see after the plot is created and y limit this is actually the frequency for different bins. So, will see once the once the plot is created. So, let us execute this line you can see the plot.

So, in this plot as you can see.

(Refer Slide Time: 30:18)

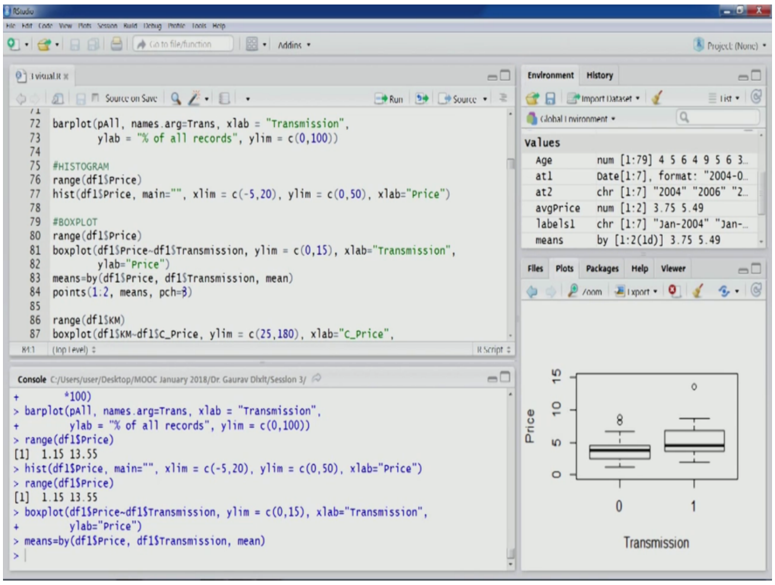


That the for better visibility of this histogram we have given this from minus 5 and this particular is also and on this x axis on this extreme a right extreme also some more range has been given. So, that we are able to visualize the whole histogram in 1 go in a much better fashions therefore, that is why this wider x limits were given and you can see the frequencies for different bins over there. So, this particular distribution because this histogram covers all the values of a numerical variable, we can get a sense whether this whether a particular distribution is following is normal distribution or not.

So, in this case in this case we can see that this seems to be a right skewed distribution right there is a slightly longest tale on the right side. So, this particular particular distribution does not seem to follow normal distribution. Therefore, it is going to be slightly difficult to apply some of this statistical technique for example, is linear regression etcetera. So, therefore, we would be required to do bit transformation to make it slightly more of a normal distribution.

Now, a let us move to the box plot. So, in again in the box for the box plot we are interested in these 2 variables price and transmission. So, let us look at the range which is again is going to be the same and box plot is the function that is used. So, in this case this is price verses transmission transmission is going to be on x axis and price is going to be on y axis. So, different groups in different groups different categories in transmission would be displayed on x axis and for each of those groups price distribution would be displayed a price values will be displayed on the y axis, limit for y axis as you can see 0 to 15 and the labelling is also you can understand.

(Refer Slide Time: 32:24)



```
barplot(pAll, names.arg=Trans, xlab = "Transmission",
        ylab = "% of all records", ylim = c(0,100))

#HISTOGRAM
range(df1$Price)
hist(df1$Price, main="", xlim = c(-5,20), ylim = c(0,50), xlab="Price")

#BOXPLOT
range(df1$Price)
boxplot(df1$Price~df1$Transmission, ylim = c(0,15), xlab="Transmission",
        ylab="Price")
means=by(df1$Price, df1$Transmission, mean)
points(1:2, means, pch=3)

range(df1$KM)
boxplot(df1$KM~df1$C_Price, ylim = c(25,180), xlab="C_Price",
```

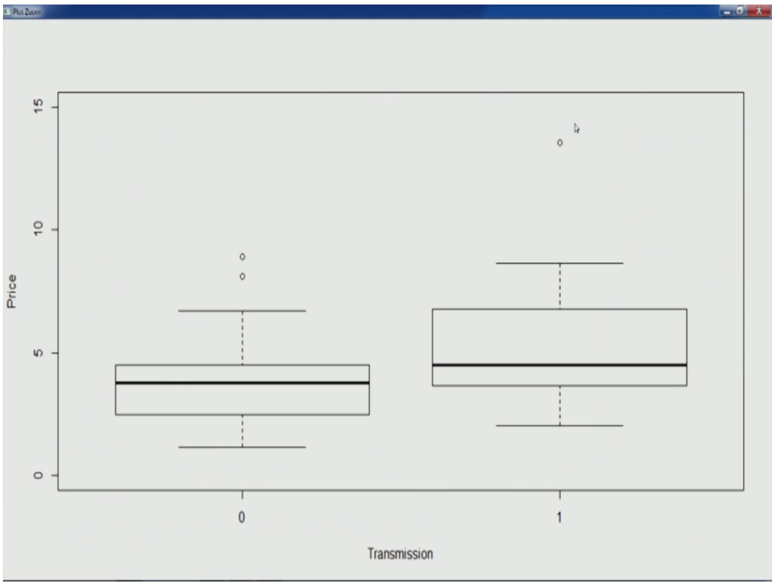
Console C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Session 3/

```
+        *100)
> barplot(pAll, names.arg=Trans, xlab = "Transmission",
+         ylab = "% of all records", ylim = c(0,100))
> range(df1$Price)
[1]  1.15 13.55
> hist(df1$Price, main="", xlim = c(-5,20), ylim = c(0,50), xlab="Price")
> range(df1$Price)
[1]  1.15 13.55
> boxplot(df1$Price~df1$Transmission, ylim = c(0,15), xlab="Transmission",
+         ylab="Price")
> means=by(df1$Price, df1$Transmission, mean)
> 
```

```
Values
  Age        num [1:79] 4 5 6 4 9 5 6 3…
  at1        Date[1:7], format: "2004-0…
  at2        chr [1:7] "2004" "2006" "2…
  avgPrice   num [1:2] 3.75 5.49
  labels1    chr [1:7] "Jan-2004" "Jan-…
  means      by [1:2(1d)] 3.75 5.49
```

So, this is the box plot let us have a look.

(Refer Slide Time: 32:28)



So, in the box plot you would see that 75 percent almost 50 percent of the values they are in the they are generally remain in the box. And the this black this this line this particular this particular line this is actually the median value and this is the starting point of the values and then you have in the box and you have first quartile and the this 1 is third quartile.

So, all the values in the box are between first quartile and third quartile. So, therefore, covering 50 percent of the values median is displayed. So, majority of the values are in this range between these 2 limits and some of the values that are displayed using a squares they are they can be called out layers.

Similarly, this is. So, this was for group 0 and for group 1 also again the same thing this this this line this particular line is median, then you have first quartile and third quartile and creating the box and other things remain same this being this value being the out layer. Now you can see between group 0 and group 1 the median value you can see this is much lower from group 1 median price value is much lower. So, the whole the whole box plot for group 1 is higher than much higher than the box plot for group 0.

So, there can be the clear separation between these 2 groups can be seen over there, to be want to look at the a mean value for a those 2 groups that can also be done. So, we will have to compute the means for those 2 groups. So, this can be done using the by command that is available in r. So, first is the you know first argument is the variable for which we have to generate the mean and the second argument is the categorical variable the now groups for which, we have to create a the mean and the mean function the because. So, let us execute this code.

So, means would be created and once means have been created we can plot them using the points command. So, points is the command which can again be used to plot the points on a particular graph. So, in this case p c h is the plotting character. So, in this case plotting character as understand as defined by value 3 is going to be displayed in the plot which is nothing, but plus sign.

So, let us execute this line more information on points command you can find out from the help section. So, you can see plus sign visible there let us look at the this plot, you can see plus sign exactly lying on the median value very close exactly matching or very close to median value for group 0. And you can see for group 1 plus sign is much higher than the median value. So, the skewness that we saw earlier may be that is coming because of the group 1 that is automatic cars.

So, let us stop here and we will create will continue from here. In the next lecture will create few more box plot and will go into. So, basic charts and these distribution plots they are mainly 2 d a graphics, we will go in more into you know multivariate or multidimensional graphics in the next lecture.

Thank you.