Marketing Analytics

Professor Swagato Chatterjee

Vinod Gupta School of Management

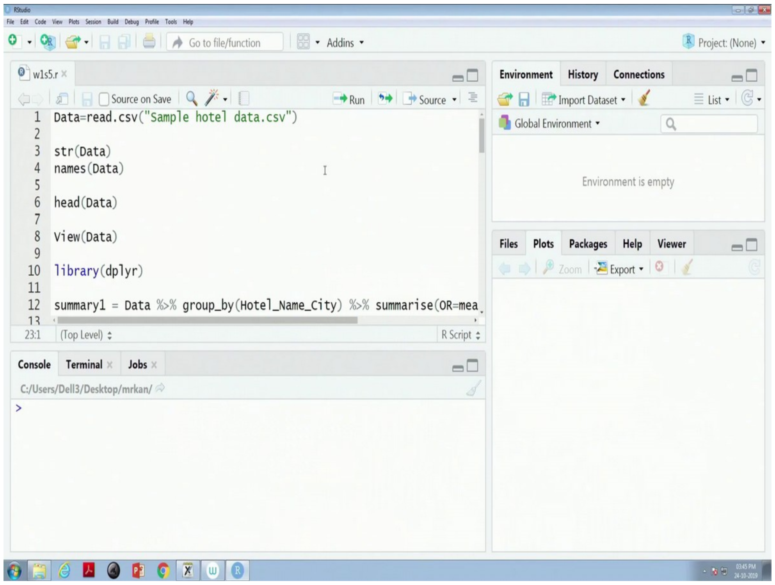
Indian Institute of Technology, Kharagpur

Lecture 06

Introduction to R Programming (Contd.)

Hello everybody. Welcome to Marketing Analytics course. This is Professor Swagato Chatterjee from VGSOM IIT, Kharagpur who will be taking this course for you. So, in the last video we were dealing with a hotel review data set and what did we do till now?

(Refer Slide Time: 0:35)



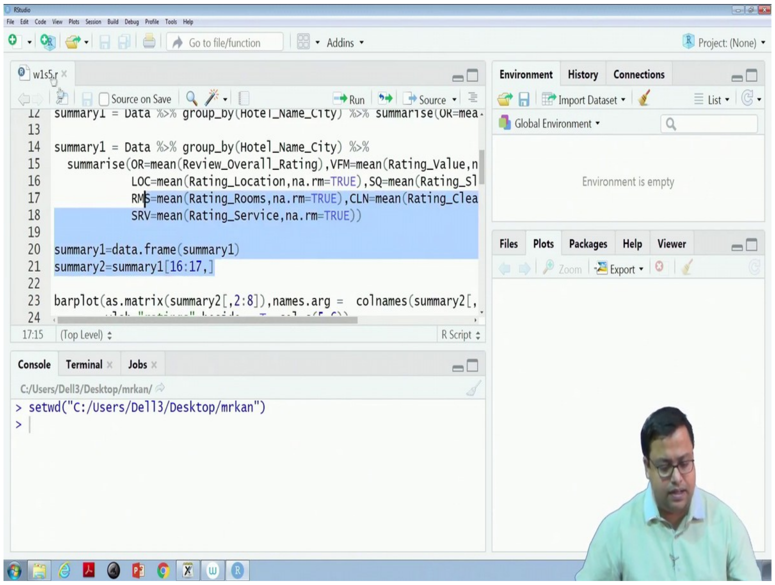
```
Data=read.csv("Sample hotel data.csv")

str(Data)
names(Data)

head(Data)

View(Data)

library(dplyr)

summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mea
```

```
C:/Users/Dell3/Desktop/mrkan/
>
```

So, I have done so if you see this. I have read the data, get a little bit of idea of the data, and then summarized the data. And when summarize the data, we use dplyr library. So, if you do not have that, you have to install that. And then I created the summary of the dataset for the whole hotel, all 23 hotels. And that summary had the overall rating, and that summary also had the mean of the location, service, and various other aspects, 6 other aspects.

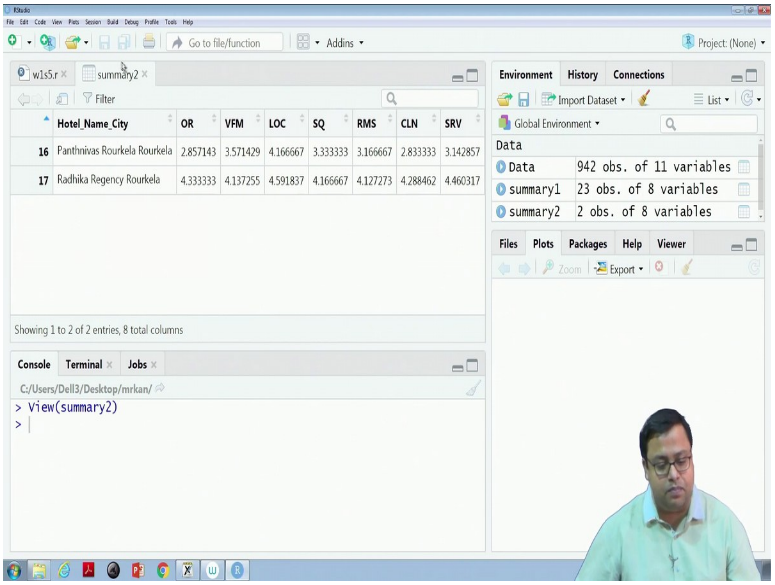
And then, we also found out how I can create, how I can actually compare my hotel with my competitor’s hotel. So, we selected two hotels from Rourkela by chance, and then we created summary 2. And we were about to create a bar plot of the performance of these two hotels. So, that is where we have stopped.

(Refer Slide Time: 1:53)



```
summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mea

summary1 = Data %>% group_by(Hotel_Name_City) %>%
  summarise(OR=mean(Review_Overall_Rating),VFM=mean(Rating_Value,n
            LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
            RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
            SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,
```

```
C:/Users/Dell3/Desktop/mrkan/
> setwd("C:/Users/Dell3/Desktop/mrkan")
>
```



	Hotel_Name_City	OR	VFM	LOC	SQ	RMS	CLN	SRV
16	Panthnivas Rourkela Rourkela	2.857143	3.571429	4.166667	3.333333	3.166667	2.833333	3.142857
17	Radhika Regency Rourkela	4.333333	4.137255	4.591837	4.166667	4.127273	4.288462	4.460317

Showing 1 to 2 of 2 entries, 8 total columns

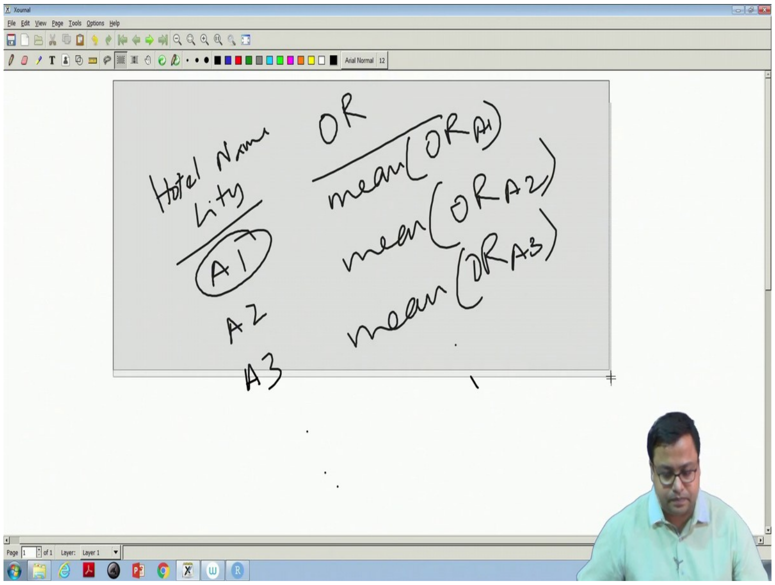
```
Data
Data        942 obs. of 11 variables
summary1    23 obs. of 8 variables
summary2    2 obs. of 8 variables
```

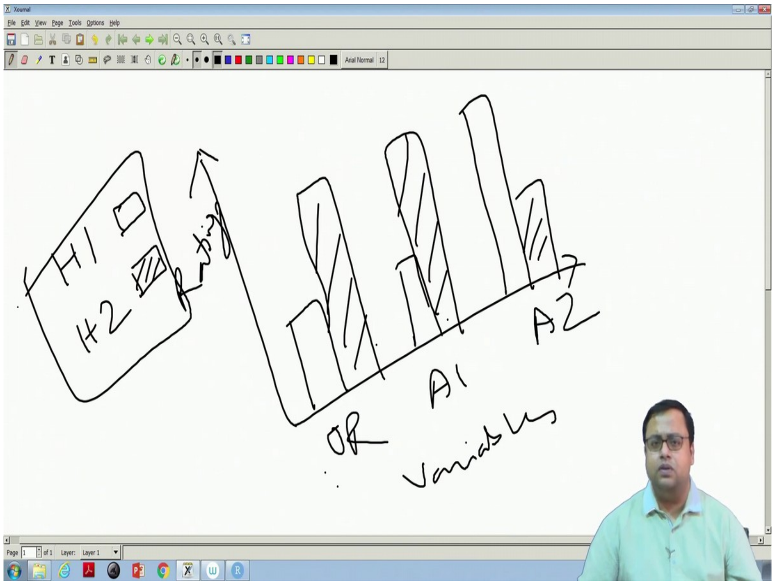
```
C:/Users/Dell3/Desktop/mrkan/
> View(summary2)
>
```

So, when I started this video, I have cleaned my console, cleaned my global environment. So, if you are continuing from the last video just now, then you do not have to do all this stuff. But all of those guys who are starting this video and seeing this video separately then the probably the session 5 video, you have to do the same thing that I am doing right now. So, clean your global environment, clean your console, set your working directory to source file location again. And then from line 1 to line 21, select the whole stuff, line 1 to line 21, and press run so that you and I remain on the same page. That is very important as we go ahead.

So, now that we are on the same page and I have the summary 2, which looks like this but there are 2 hotels and their overall rating, and their performance of each of the attributes has been given. I will actually try to do a little bit more work.

(Refer Slide Time: 2:38)

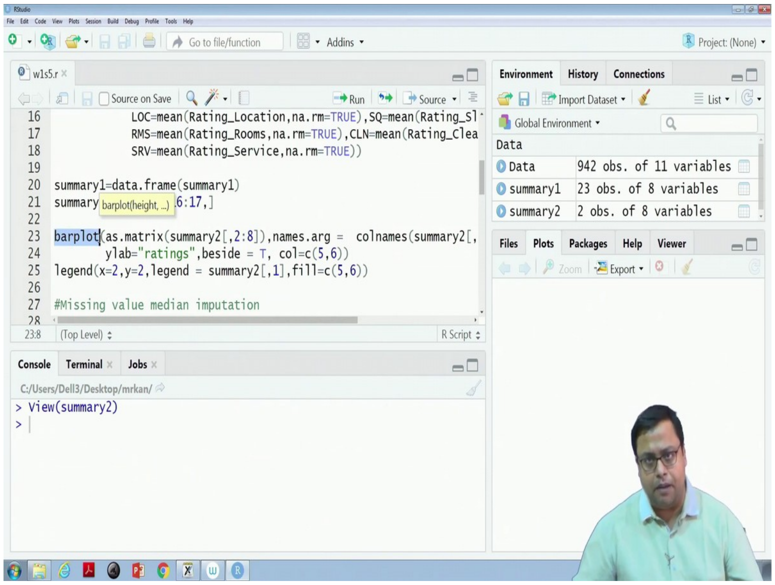




So the first thing that I will try to do is create a bar plot, bar plot looks like look something like this, I will clean this first. And so let us say. I am planning to create something like this, where, so for hotel 1 it will look like this, hotel 2 it will look like this. And then there will be a curve. And the second column will be like this always. So, this is let us say overall rating. This is aspect 1, this is aspect 2. And this is the ratings, and this is the variables probably. And this is hotel 1, and this is hotel 2; this is the legend, so something like this is what I will be trying to create.

Why have I been trying to create this? Because this gives a fairly measurable idea that Overall rating wise hotel 1 is worse than hotel 2. And that is majorly contributed by A1, though A2 wise hotel 1 is much better than hotel 2. It might not matter probably that is why this kind of result is coming. So, this is something that we are going to create now.

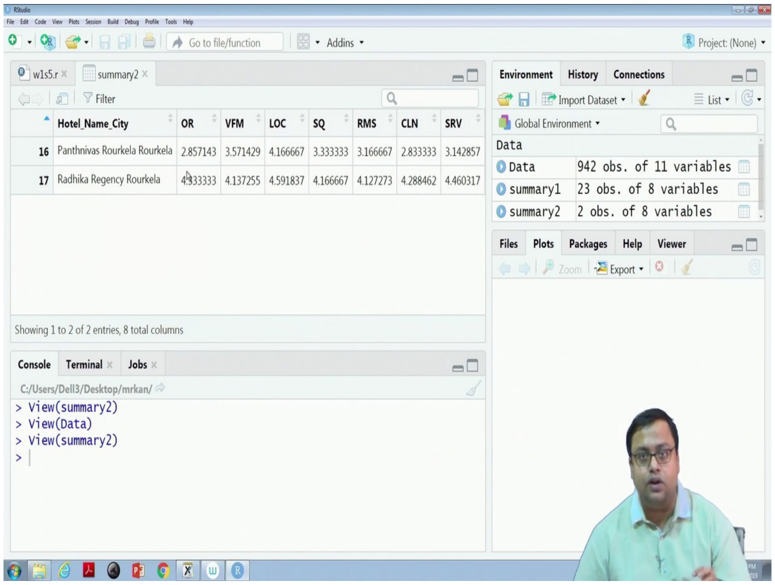
(Refer Slide Time: 3:50)



```
               LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
               RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
               SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary                 6:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,
          ylab="ratings",beside = T, col=c(5,6))
legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))

#Missing value median imputation
```

```
C:/Users/Dell3/Desktop/mrkan/
> View(summary2)
> 
```

Data	
Data	942 obs. of 11 variables
summary1	23 obs. of 8 variables
summary2	2 obs. of 8 variables

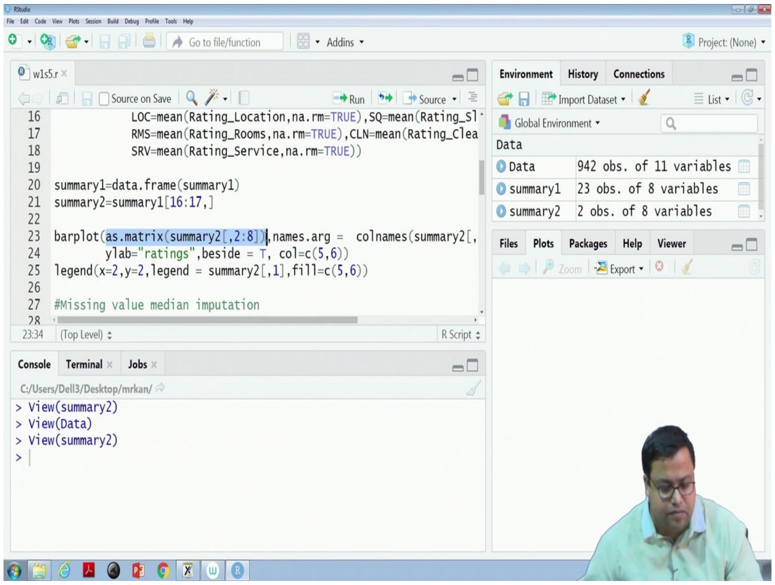


	Hotel_Name_City	OR	VFM	LOC	SQ	RMS	CLN	SRV
16	Panthnivas Rourkela Rourkela	2.857143	3.571429	4.166667	3.333333	3.166667	2.833333	3.142857
17	Radhika Regency Rourkela	4.333333	4.137255	4.591837	4.166667	4.127273	4.288462	4.460317

Showing 1 to 2 of 2 entries, 8 total columns

```
C:/Users/Dell3/Desktop/mrkan/
> View(summary2)
> View(Data)
> View(summary2)
> 
```

Data	
Data	942 obs. of 11 variables
summary1	23 obs. of 8 variables
summary2	2 obs. of 8 variables



```
           LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
           RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
           SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,
        ylab="ratings",beside = T, col=c(5,6))
legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))

#Missing value median imputation
```

Data

Data	942 obs. of 11 variables
summary1	23 obs. of 8 variables
summary2	2 obs. of 8 variables

```
C:/Users/Dell3/Desktop/mrkan/
> View(summary2)
> View(Data)
> View(summary2)
> 
```

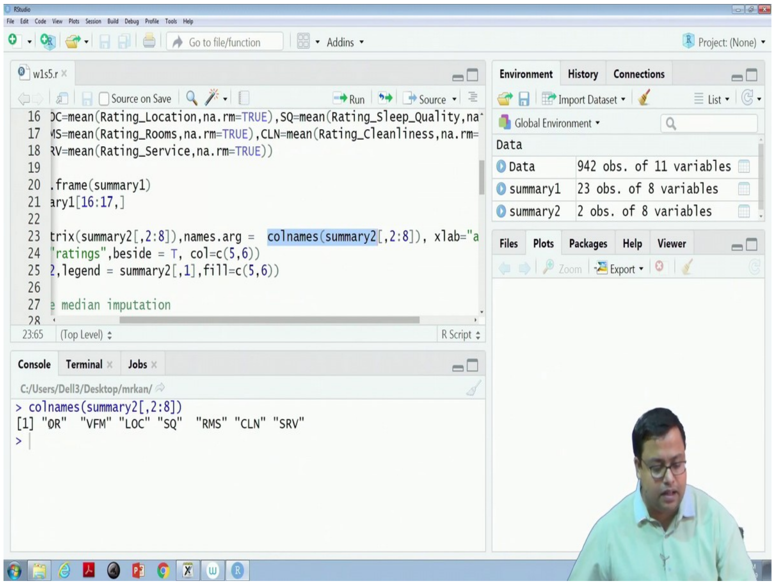
So to that, the code is bar plot. The function is bar plot, so you can go and find out the help of bar plot; how to find out the help of bar plot? You have to write a question mark. And then bar plot and press enter, or you can write help within bracket bar plot. So, any of these two options you can do.

So, bar plot will use a matrix in it. So, what I am trying to plot here, I am trying to plot here in the data set, not here, in this dataset summary 2 whole this. All this from here up to SRV, everything I am trying to plot, but I cannot give the input. This is something that has been written by somebody. And he has written in that way that the input has to be either a vector or a matrix.

So here, because we are creating a bar plot of multiple parameters, it is not a single parameter, it is multiple parameters. So, the input will be a matrix. So, I have given the input as a "matrix.Summary2," . So, you will see summary 2 is the dataset, summary 2 datasets. If I just see this much, summary 2 datasets all the row because nothing written before comma, and 2 to 8. That means leave aside the first column, which is the hotel name.

And other than that, use all other columns that mean overall rating to service all other columns you use and then change it to its matrix form as dot metrics. So, if I just click it and print it here, it will be just this, this is a matrix format. This is the input in the bar plot. This particular matrix is the input in the bar plot.

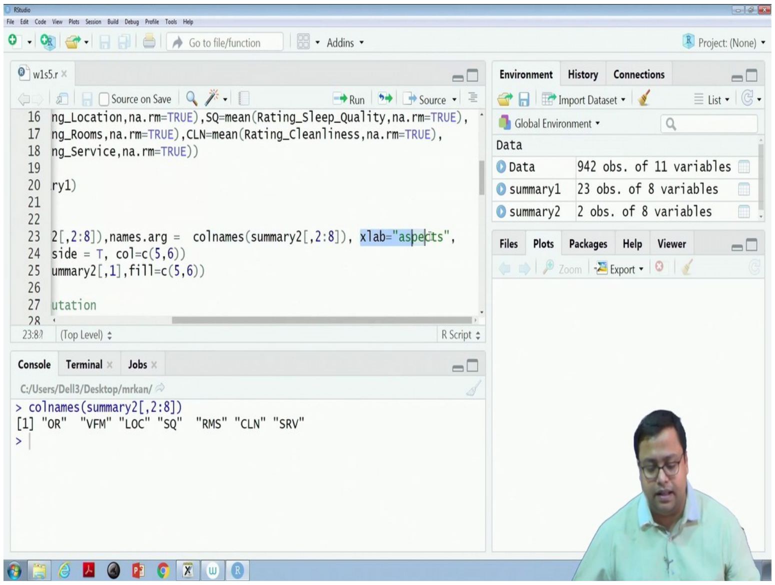
(Refer Slide Time: 5:50)



```
)C=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sleep_Quality,na
MS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Cleanliness,na.rm=
RV=mean(Rating_Service,na.rm=TRUE))

.frame(summary1)
ary1[16:17,]

trix(summary2[,2:8]),names.arg =  colnames(summary2[,2:8]), xlab="a
"ratings",beside = T, col=c(5,6))
2,legend = summary2[,1],fill=c(5,6))

e median imputation
```

```
> colnames(summary2[,2:8])
[1] "OR"  "VFM" "LOC" "SQ"  "RMS" "CLN" "SRV"
>
```

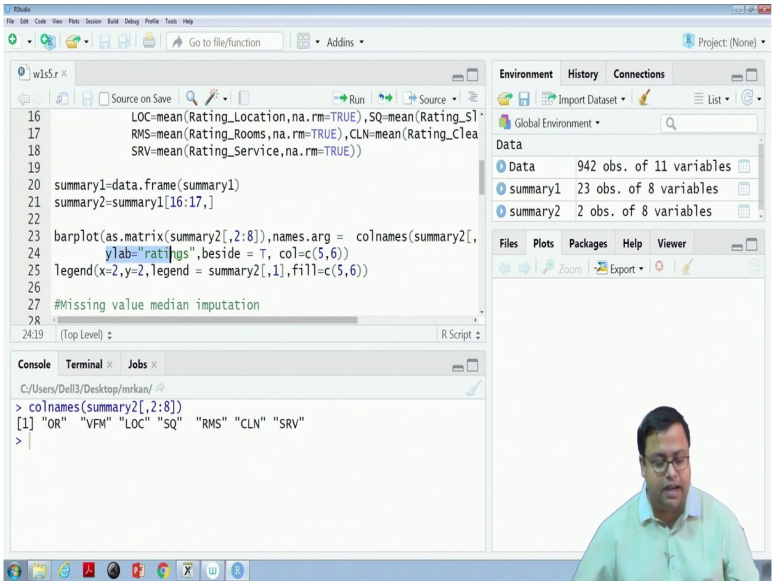
Data
Data 942 obs. of 11 variables
summary1 23 obs. of 8 variables
summary2 2 obs. of 8 variables



```
ng_Location,na.rm=TRUE),SQ=mean(Rating_Sleep_Quality,na.rm=TRUE),
ng_Rooms,na.rm=TRUE),CLN=mean(Rating_Cleanliness,na.rm=TRUE),
ng_Service,na.rm=TRUE))

ry1)


2[,2:8]),names.arg =  colnames(summary2[,2:8]), xlab="aspects",
side = T, col=c(5,6))
ummary2[,1],fill=c(5,6))

utation
```

```
> colnames(summary2[,2:8])
[1] "OR"  "VFM" "LOC" "SQ"  "RMS" "CLN" "SRV"
>
```

Data
Data 942 obs. of 11 variables
summary1 23 obs. of 8 variables
summary2 2 obs. of 8 variables



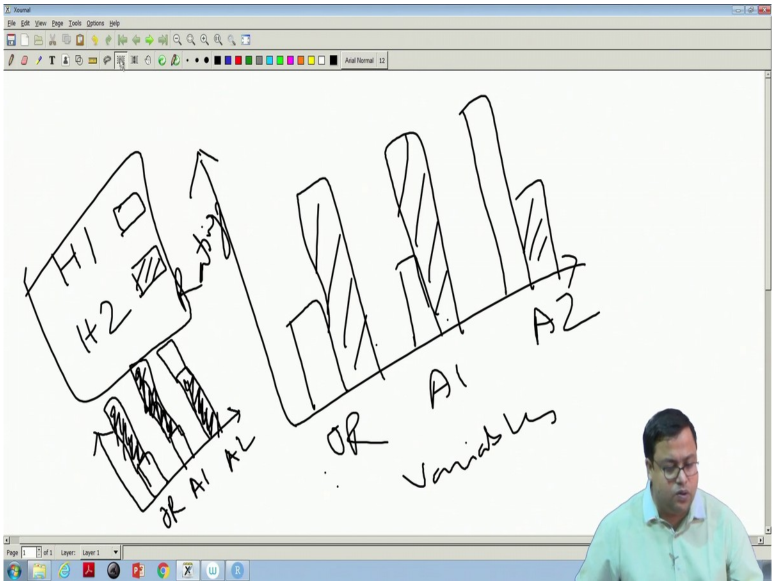
```r
            LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
            RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
            SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,
        ylab="ratings",beside = T, col=c(5,6))
legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))

#Missing value median imputation
```

```
> colnames(summary2[,2:8])
[1] "OR"  "VFM" "LOC" "SQ"  "RMS" "CLN" "SRV"
>
```

Global Environment

Data

Data	942 obs. of 11 variables
summary1	23 obs. of 8 variables
summary2	2 obs. of 8 variables



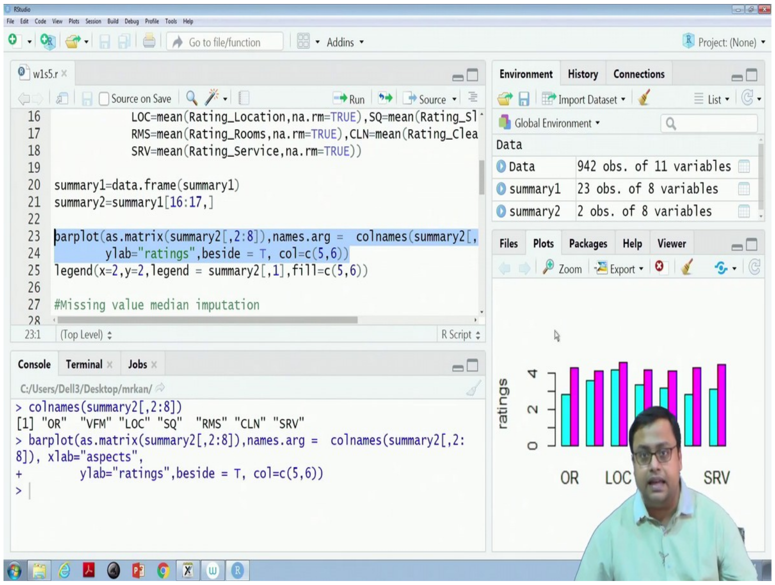
And what else? It is asking me that when I put the bar plot, what I will write here? So, here I have written OR, A1, A2. So, what will I write in my case? So, that is the names argument; names.arg, names argument is what? The column names of Summary 2. What are the column names of Summary 2? If I write column names of Summary 2. And then 2 to 8. So, that will give me this column names. Because the second column to the eighth column. I am creating a subset of the data set, and then I am finding out the column names.

So, these are the aspects, and the overall rating comes at the bottom of the chart. So, that is how I have created. Then I have also asked what my x label is? So X-axis, what will be the label? The label is aspects. Y -axis what is the label? Ratings. You can give any other name does not matter. Now, beside is equal to true. Beside is equal to true means. I have 2 hotels, so one way of representing it is this.

Another way of representing is somebody can create something like this. So it is hotel 1. This is Hotel 2. This is Hotel 1. This is Hotel 2. And this is hotel 1. And this is Hotel 2. Overall A1, A2. So, basically stacked, this one stacked view. So, something like this. Somebody can create this stack view. So, when you try to create a stack view, you will write.

Beside is equal to false, when you try to create, do not want to create a stack view, you will write beside is equal to true. So, I have written beside is equal to true and color you can choose anything. I have written 5 comma 6. You have to choose 2 colours. There are lots of colours available. You can choose 1 comma 2. It will come black and probably red. And then if you, so RBG and so on.

(Refer Slide Time: 8:04)



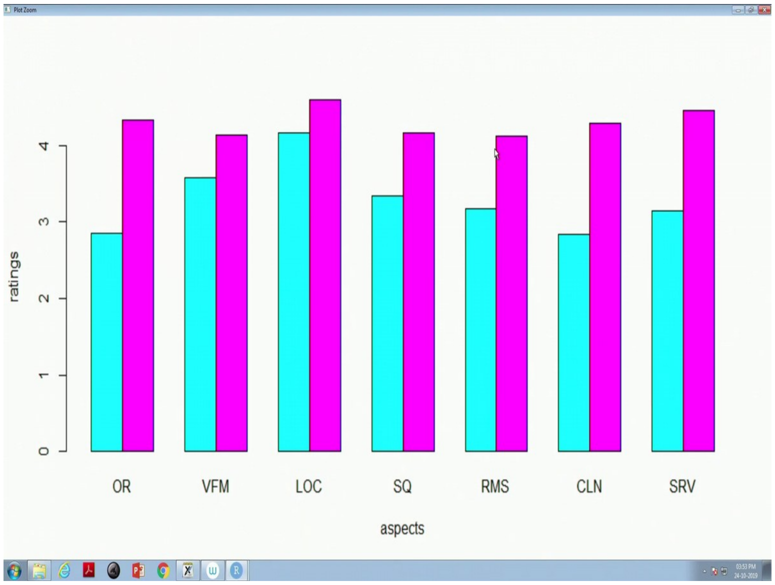
```r
              LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
              RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
              SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,
         ylab="ratings",beside = T, col=c(5,6))
legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))

#Missing value median imputation
```

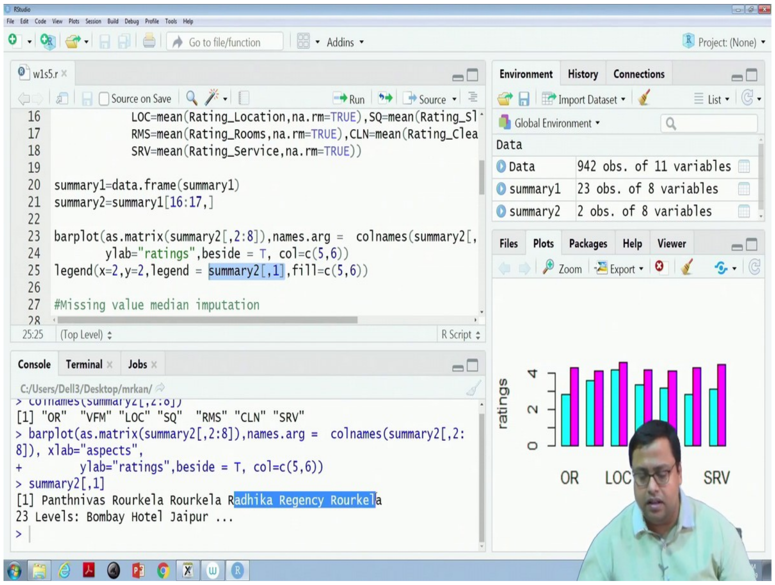
```
C:/Users/Dell3/Desktop/mrkan/
> colnames(summary2[,2:8])
[1] "OR"  "VFM" "LOC" "SQ"  "RMS" "CLN" "SRV"
> barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,2:
8]), xlab="aspects",
+         ylab="ratings",beside = T, col=c(5,6))
> |
```

Global Environment

Data

Data	942 obs. of 11 variables
summary1	23 obs. of 8 variables
summary2	2 obs. of 8 variables





```r
            LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
            RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
            SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,
        ylab="ratings",beside = T, col=c(5,6))
legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))

#Missing value median imputation
```

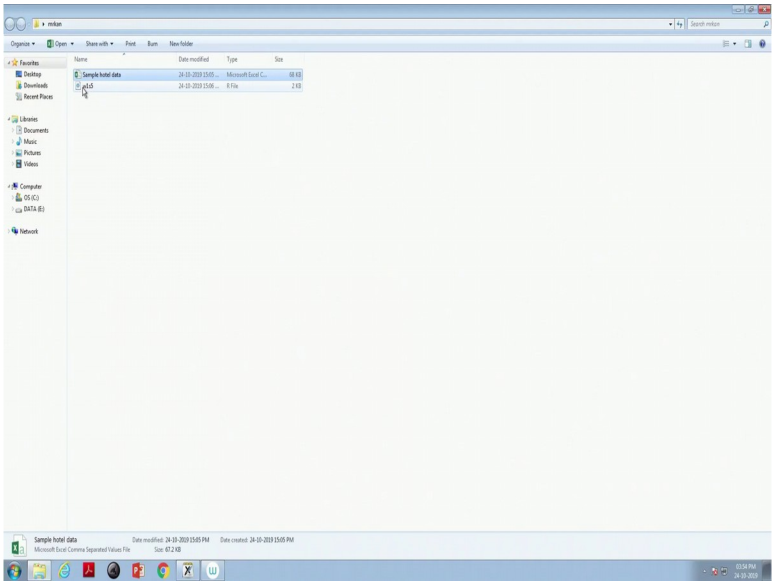
```
C:/Users/Dell3/Desktop/mrkan/
> colnames(summary2[,2:8])
[1] "OR"  "VFM" "LOC" "SQ"  "RMS" "CLN" "SRV"
> barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,2:
8]), xlab="aspects",
+         ylab="ratings",beside = T, col=c(5,6))
> summary2[,1]
[1] Panthnivas Rourkela Rourkela Radhika Regency Rourkela
23 Levels: Bombay Hotel Jaipur ...
> 
```

Data
Data	942 obs. of 11 variables
summary1	23 obs. of 8 variables
summary2	2 obs. of 8 variables

So, if I select these two lines the bar plot line and run this one, I get here in the plotting area, I get a plot like this, and I will actually make it a little bit broad. And you will see that there are overall ratings, value for money, locations, sleep quality, and so on. And the ratings are coming up, and I can fairly see by seeing this. And if you can probably also want to add up the legend on this.

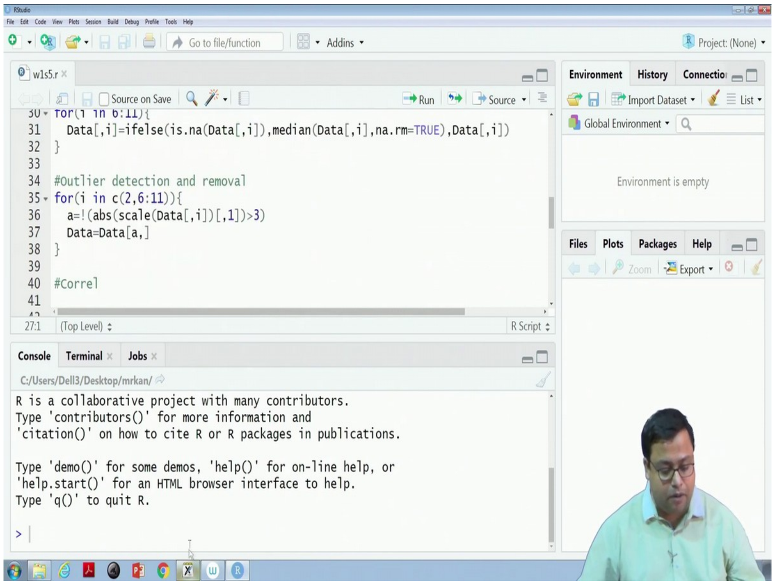
So, the legend is x is equal to 2, y is equal to 2 is actually the location of the legend via X axis and in Y-axis. At what level it will be printed. And what is the legend? Legend=summary2,1. If I just print out this much. This is nothing but see two levels. One is panthnivas rourkela and other is radhika regency rourkela. So, these two names get printed, and the colour is the same colour that you have chosen 5 comma 6.

(Refer Slide Time: 9:02)



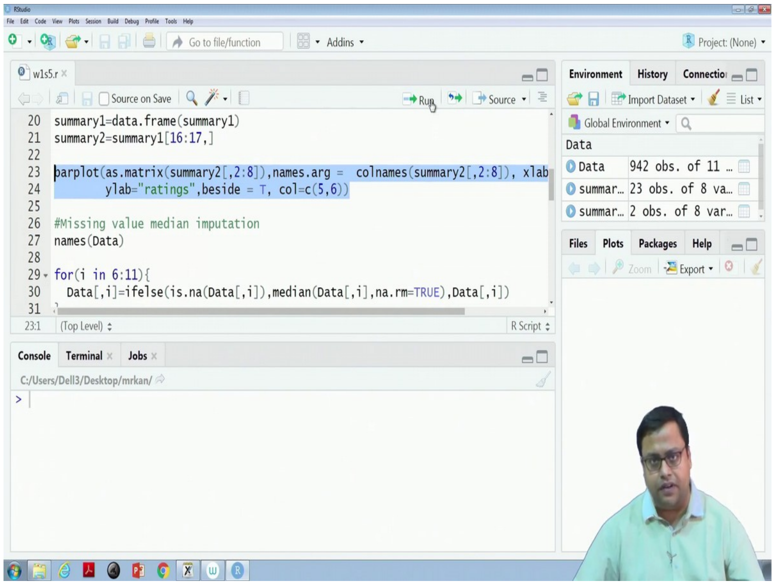
Name	Date modified	Type	Size
Sample hotel data	24-10-2019 15:05 ...	Microsoft Excel C...	68 KB
w1s5	24-10-2019 15:06 ...	R File	2 KB

Sample hotel data
Microsoft Excel Comma Separated Values File
Date modified: 24-10-2019 15:05 PM
Size: 67.2 KB
Date created: 24-10-2019 15:05 PM



```r
for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=!(abs(scale(Data[,i])[,1])>3)
  Data=Data[a,]
}

#Correl

```

```
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.

>
```



```r
summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,2:8]), xlab
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)

for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}
```

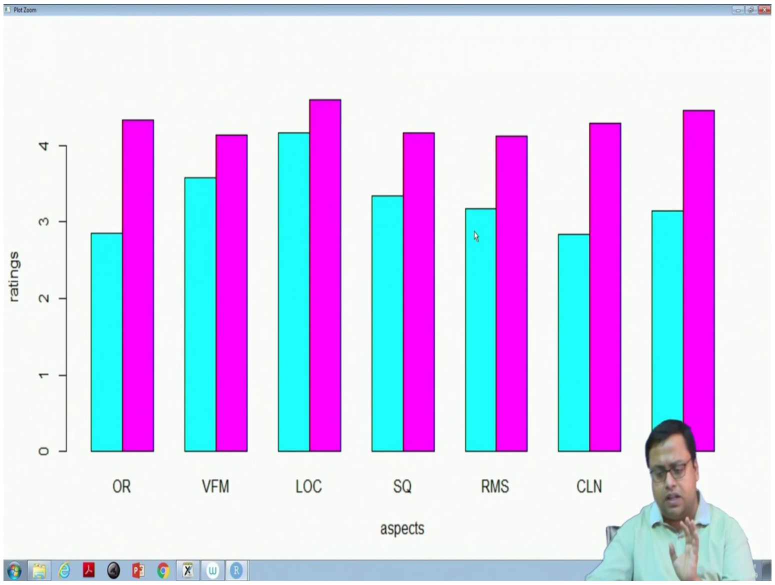
23:1

C:/Users/Dell3/Desktop/mrkan/

```
>
```

Data

Data	942 obs. of 11 …
summar…	23 obs. of 8 va…
summar…	2 obs. of 8 var…



So, if I run this, the legends come here. So, this is a very common thing that happens in R that sometimes it hangs so it is. That is why it is a very good practice that you keep on saving whatever files that you are creating so that you do not lose stuff. Yes. So what we will do now is that we will create the bar plot, and the bar plot will look like this.

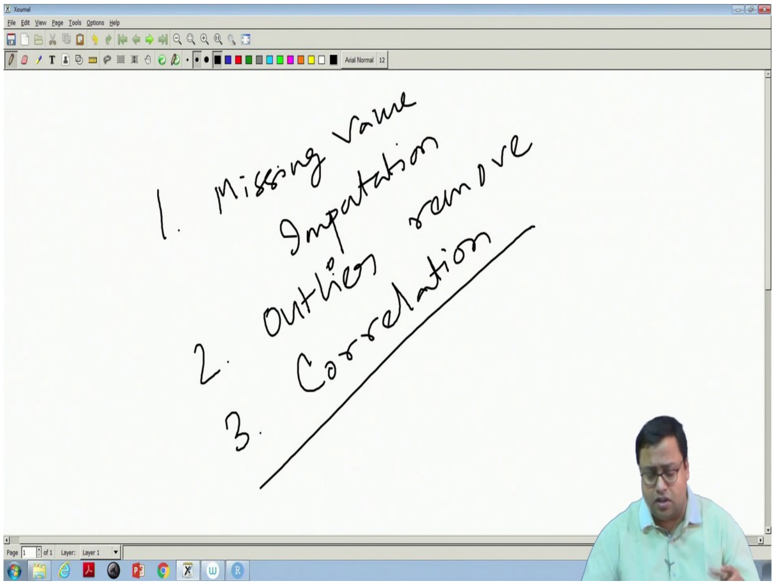
So, as I just told you that there is a matrix that we have to create. And then the names argument and x-label, y -label. And besides is equal to true. So, and colour is equal to 5, 6, so when I run this it looks like this. And if I just make it bigger, it will look like this. And I get an idea that, okay, so based on this thing, I can say that overall rating. So, overall rating, I can say that the blue one, I am not telling which one is which. Because the first one is basically blue and the second one is pink.

The first one, hotel 1 is not doing as good as hotel 2. Obviously, the overall rating is a little bit lower based on the data set that we have. And I can also say, by seeing this particular thing that VFM is more or less the same, location is also more or less the same. One of the major things which is not same is cleanliness.

So, cleanliness is much lower for the blue guy, and probably the SRV is also much lower. The service is also lower than, so I would strongly say that the last 4 aspects SQ, RMS, cleanliness and service. These are the four contributors. Now, it is very easy to say that cleanliness is something that you can improve right, very easily. And sometimes if you actually improve cleanliness, it might also have an effect on SQ that is service, sleep quality or if not sleep quality, at least in SRV the service it will have an impact on that. So, it might be a more economical choice that I will focus on cleanliness.

Now, other than that, I have to also focus on that which one is more important and which one is less important. This is something that matters. How? We have discussed in the previous video that I have to find out that using something called the regression equation where overall rating will be my y variable and x1 to x6, which is all the aspects will be my x variable.

(Refer Slide Time: 11:57)



1. Missing Value Imputation
2. Outliers remove
3. Correlation

Xournal
File Edit View Page Tools Options Help
Page 1 of 1 Layer: Layer 1

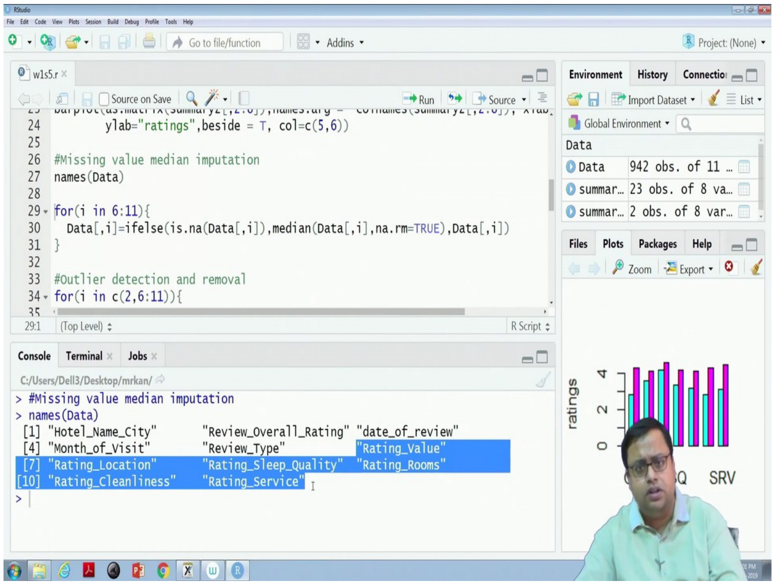
So, before we run regression, there are certain steps. What are those steps? The first step before you run a regression is that we have missing values, so missing value imputation is my first step. Then what is the step? So, I will deal with it. There are lots of ways of missing value imputation. Ideally, that should have been covered in a course called Introduction to Business Analytics.

Here, I will just follow something called a median imputation because you see all our observations you will see later point of time. All our observation is a little bit off. I would say bias towards the higher side. That means more numbers of 4's and 5's are there rather than 2's and 3's and 1's probably the review. So, it is not symmetric. And if it is not symmetric probably median is a better choice than mean to when I am doing an imputation because that will not change the distribution of the data. So, that is something that I am trying to do.

The second one is missing value imputation. After that, we have to also find out outliers because this is a linear regression. All of the variables are let us say I can assume them to be numerical and continuous and if by chance I am assuming them to be continuous, so we will see that whether the assumption works or not. But, if by chance that they are continuous, then there might be outliers, and I have to find out those outliers and remove them. That is step number 2.

And step number 3 is basically we have to see correlation. So, at least the x variables cannot be correlated with each other to run a regression. That is also an assumption that we have, and we have to check whether that is fulfilled or not. And I will stop here. There are other things that you have to check. We have to check for heteroscedasticity; you have to check for variance inflation factor and etcetera. But today's class is only about how to do it in Excel, sorry in R in an easy format, so I will not focus on that.

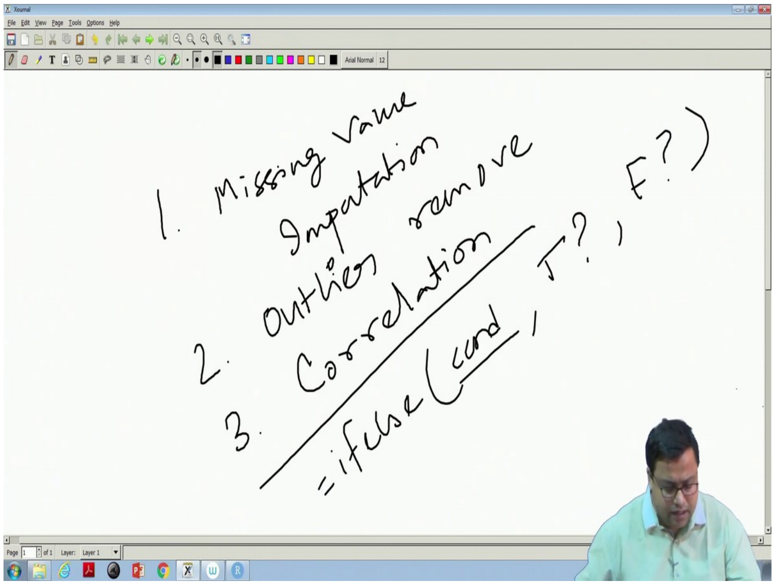
(Refer Slide Time: 13:56)



```
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)

for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}

#Outlier detection and removal
for(i in c(2,6:11)){
```

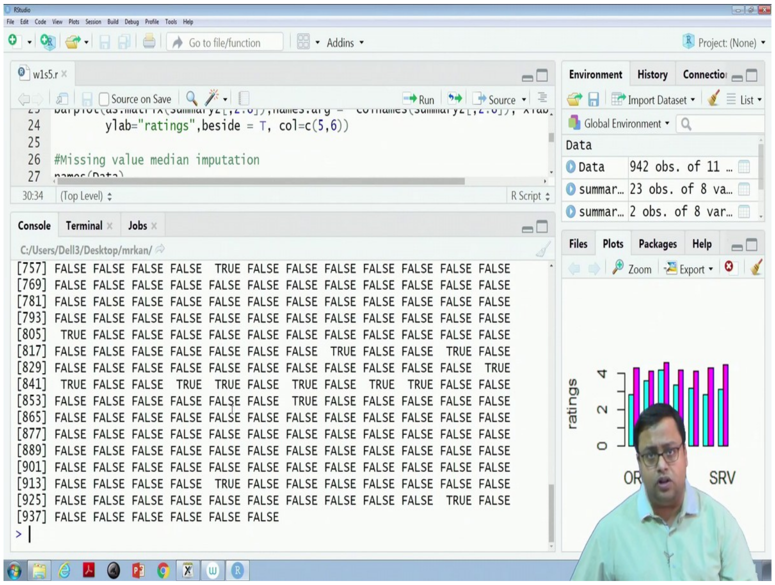
```
C:/Users/Dell3/Desktop/mrkan/
> #Missing value median imputation
> names(Data)
 [1] "Hotel_Name_City"       "Review_Overall_Rating" "date_of_review"
 [4] "Month_of_Visit"        "Review_Type"           "Rating_Value"
 [7] "Rating_Location"       "Rating_Sleep_Quality"  "Rating_Rooms"
[10] "Rating_Cleanliness"    "Rating_Service"
> 
```

Data	
Data	942 obs. of 11 ...
summar...	23 obs. of 8 va...
summar...	2 obs. of 8 var...



1. Missing value imputation
2. Outlier remove
3. Correlation

=ifelse(cond, T?, F?)



```
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
```

Data

Data	942 obs. of 11 ...
summar...	23 obs. of 8 va...
summar...	2 obs. of 8 var...

```
[757] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[769] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[781] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[793] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[805]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[817] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
[829] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[841]  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
[853] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[865] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[877] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[889] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[901] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[913] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
[937] FALSE FALSE FALSE FALSE FALSE FALSE
> |
```

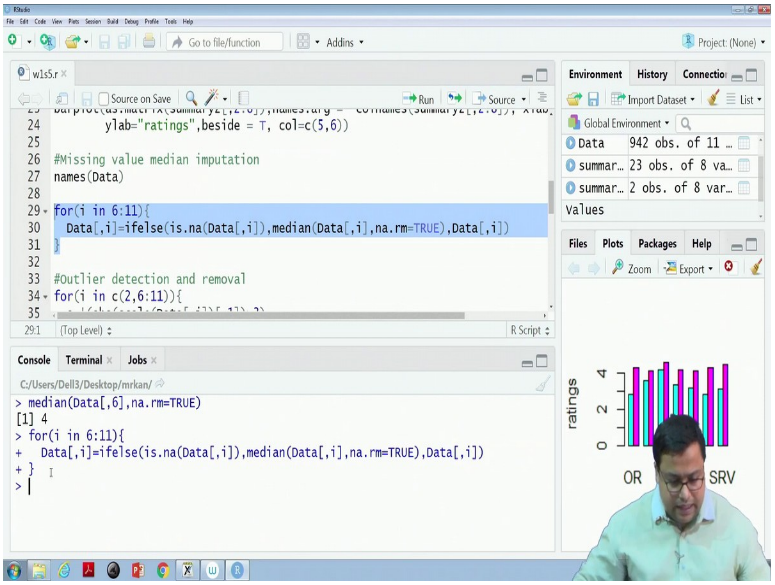
So, let us say the first one is a missing value. So, what I am writing here carefully see, for i is equal to 6 to 11. Now, how do I know 6 to 11? Because I have checked the names of data and the missing values are there in the service rating, the aspect wise rating part. So, those are 6 that is rating value to service, rating service. These are the 6 aspects where any value is there and which will be used as my x variable in the regression. So, I am focusing on them, 6 to 11.

So, for i is equal to 6 to 11. So, here I told, when these names within bracket data will come in handy. So, i in 6 to 11, what happens is this i keeps on changing, i takes the value of 6 and then 7, then 8. When i takes the value of 6, data, then comma 6 that means the 6th column will be replaced in such a way. Such that ifelse. Ifelse means if something happens, you do something. If that does not happen, you do something other. So, there will be ifelse if you remember there are 3 things in ifelse, in ifelse there were 3 things. Condition, if the condition is true, what happens? If the condition is false what happens?

So, these are the 3 entries that I have given in my ifelse function as well. So, ifelse is dot na data dollar i. So, let us say i is equal to 6. Let us assume i is equal to 6. Then what will this particular guy give is dot na, data dollar i. This guy will give me see lots of trues and false. There are some trues in between also and lots of false. So, it is actually checking out of my nine hundred observations.

Where the values are na , whether it is na like it is asking if it is yes it is na. Then it will come as true, if it is not na not missing value then it will come as false.

(Refer Slide Time: 16:08)



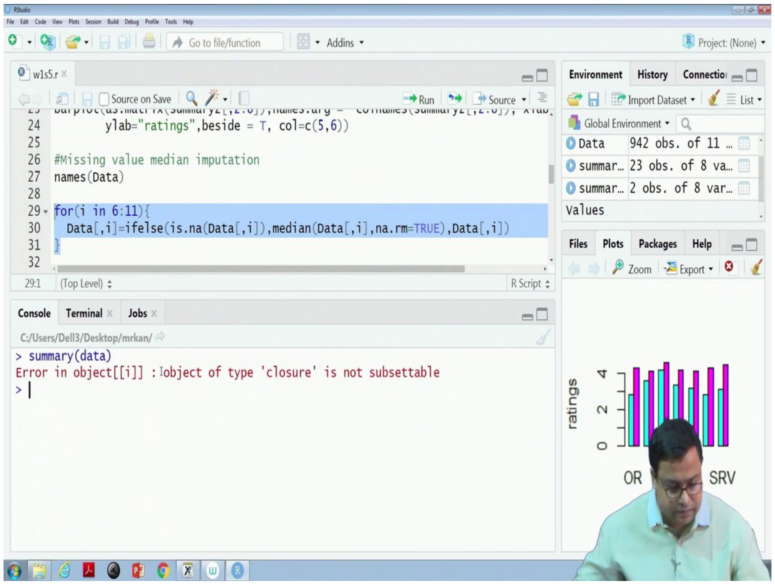
```r
barplot(as.matrix(summary2[,2:6]),names.arg = colnames(summary2[,2:6]), xlab,
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)

for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}

#Outlier detection and removal
for(i in c(2,6:11)){
```

```
C:/Users/Dell3/Desktop/mrkan/
> median(Data[,6],na.rm=TRUE)
[1] 4
> for(i in 6:11){
+   Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
+ }
> 
```

```
Data      942 obs. of 11 …
summar…   23 obs. of 8 va…
summar…   2 obs. of 8 var…
Values
```

So, whenever it is true, replace it with what? Median of my data comma 6. So, what is the median value of 6 when i is equal to 6? What is the median value? 4, and when that is not the case, then keep it as it is. When the is dot na is false we do no changes. So, this particular thing will actually change if I go one column at a time, find out what the na values are, replace it with that particular column median value. So, what I do that, I actually replace everything with the median value. So, how will I know that it has been replaced?
]

(Refer Slide Time: 16:53)

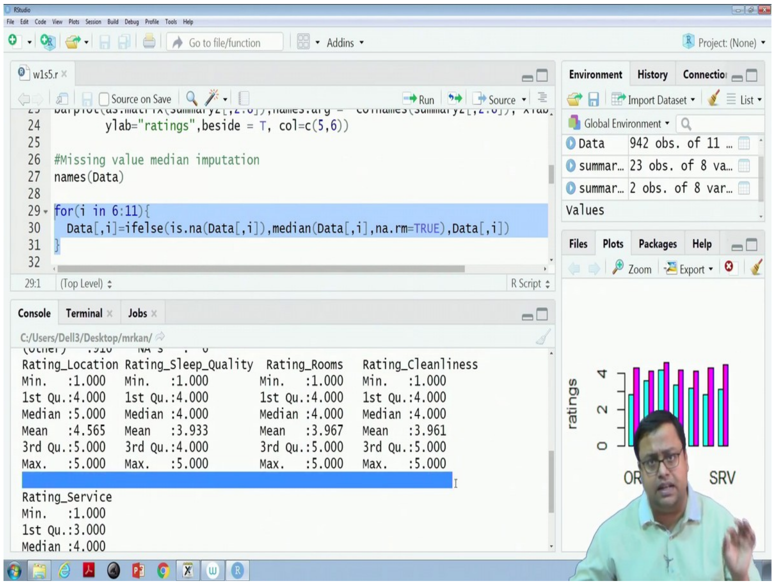


```
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)

for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}
```

```
Data       942 obs. of 11 ...
summar...  23 obs. of 8 va...
summar...  2 obs. of 8 var...
Values
```

C:/Users/Dell3/Desktop/mrkan/

```
> summary(data)
Error in object[[i]] : object of type 'closure' is not subsettable
> 
```

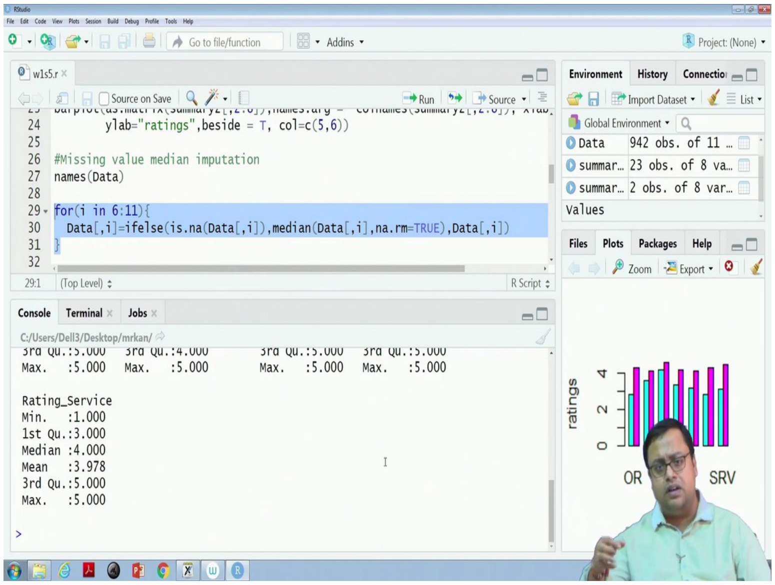


```
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)

for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}
```

```
Data       942 obs. of 11 ...
summar...  23 obs. of 8 va...
summar...  2 obs. of 8 var...
Values
```

C:/Users/Dell3/Desktop/mrkan/

```
 Rating_Location Rating_Sleep_Quality  Rating_Rooms   Rating_Cleanliness
 Min.   :1.000   Min.   :1.000        Min.   :1.000   Min.   :1.000
 1st Qu.:4.000   1st Qu.:4.000        1st Qu.:4.000   1st Qu.:4.000
 Median :5.000   Median :4.000        Median :4.000   Median :4.000
 Mean   :4.565   Mean   :3.933        Mean   :3.967   Mean   :3.961
 3rd Qu.:5.000   3rd Qu.:4.000        3rd Qu.:5.000   3rd Qu.:5.000
 Max.   :5.000   Max.   :5.000        Max.   :5.000   Max.   :5.000

 Rating_Service
 Min.   :1.000
 1st Qu.:3.000
 Median :4.000
```



```
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)

for(i in 6:11){
  Data[,i]=ifelse(is.na(Data[,i]),median(Data[,i],na.rm=TRUE),Data[,i])
}
```

Data	942 obs. of 11 ...
summar...	23 obs. of 8 va...
summar...	2 obs. of 8 var...

Values

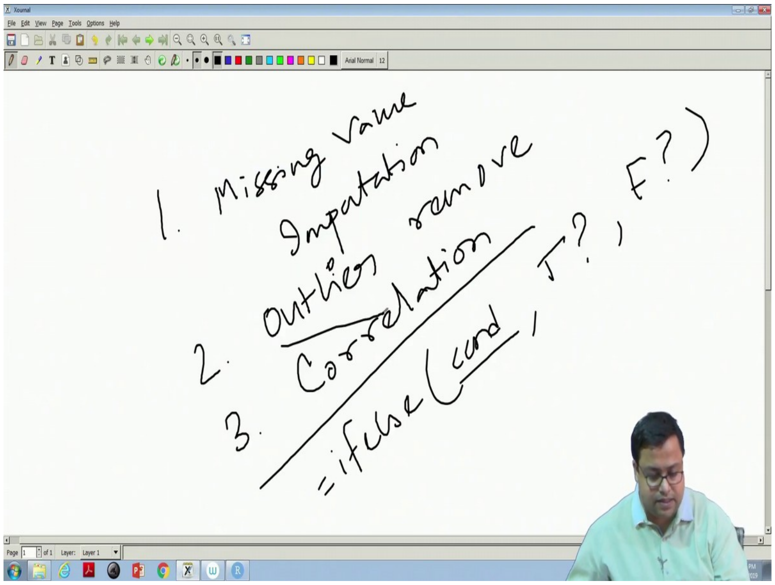
C:/Users/Dell3/Desktop/mrkan/

```
 Max.   :5.000   Max.   :5.000         Max.   :5.000   Max.   :5.000

 Rating_Service
 Min.   :1.000
 1st Qu.:3.000
 Median :4.000
 Mean   :3.978
 3rd Qu.:5.000
 Max.   :5.000

>
```

Now, if you do a summary of the data, you will see there is nothing. No na is in the last 6, here when there is a na. This is actually saying na, but this is review type we have not touched that, we have touched the last 6 attributes, there were some na values because you have seen that is dot na give me true in some of the cases, but right now I have removed them. Here there is no na coming in any of these things no na coming. So, that means that all the na values have been replaced.

So, summary is something that I wanted to show. So, that will actually remove all the. So, it does not remove; it actually shows the basic, I would say descriptive statistics of all the variables like what is the mean, what is the median, what is the first quantile, what is the third quantile, etcetera, etcetera. So, you can quickly see that in summary.

(Refer Slide Time: 18:03)



1. Missing value Imputation

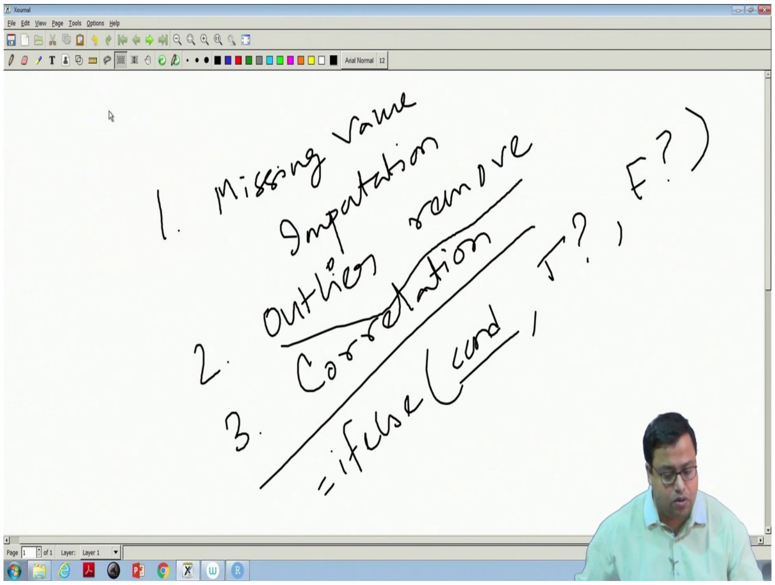
2. Outliers remove

3. Correlation

=ifelse(cond, T?, F?)

Now, we have done missing value. What is the next step? The next step I told, the next step is outlier removal. So, remove the outlier.

(Refer Slide Time: 18:09)

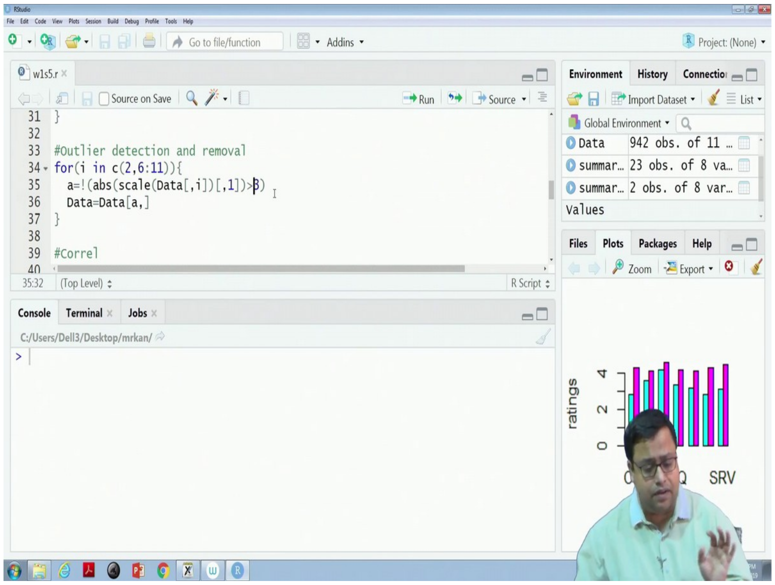


1. Missing value Imputation

2. Outliers remove

3. Correlation

=ifelse(cond, T?, F?)

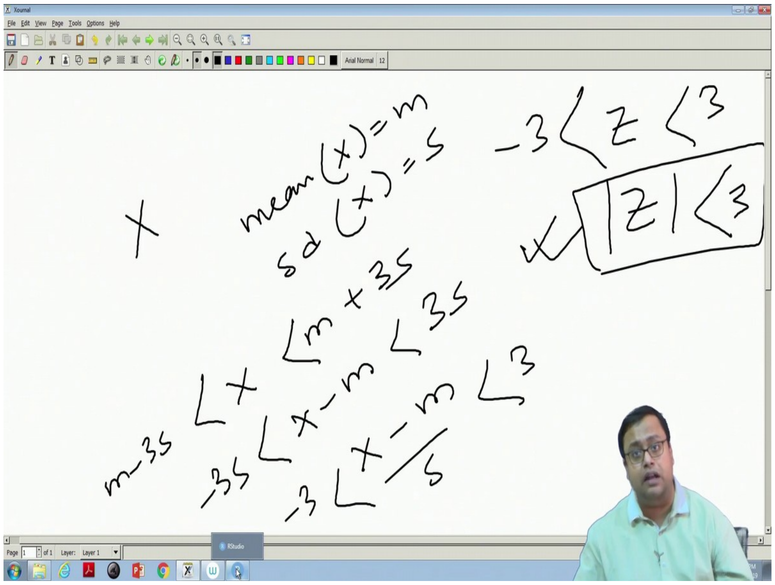


```
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=!(abs(scale(Data[,i])[,1])>3)
  Data=Data[a,]
}

#Correl
```

```
C:/Users/Dell3/Desktop/mrkan/
> 
```

So, how to remove the outlier? So, outlier will be there both in the y variable and the x variable. So, x variable was 6 to 11, but y variable was the second variable in my dataset 2. So, I have written when i varies from 2 and 6 to 11. So 2, 6, 7, 8, 9, 10, 11 these are the values that i will take. Then what I am doing? I have written something carefully, check what I have written. So, one by one, so when I try to do missing value I am trying to do outlier. What is an outlier?

(Refer Slide Time: 18:45)



X

$\text{mean}(x) = m$

$\text{sd}(x) = s$

$m - 3s < x < m + 3s$

$-3s < x - m < 3s$

$-3 < \frac{x - m}{s} < 3$

$-3 < z < 3$

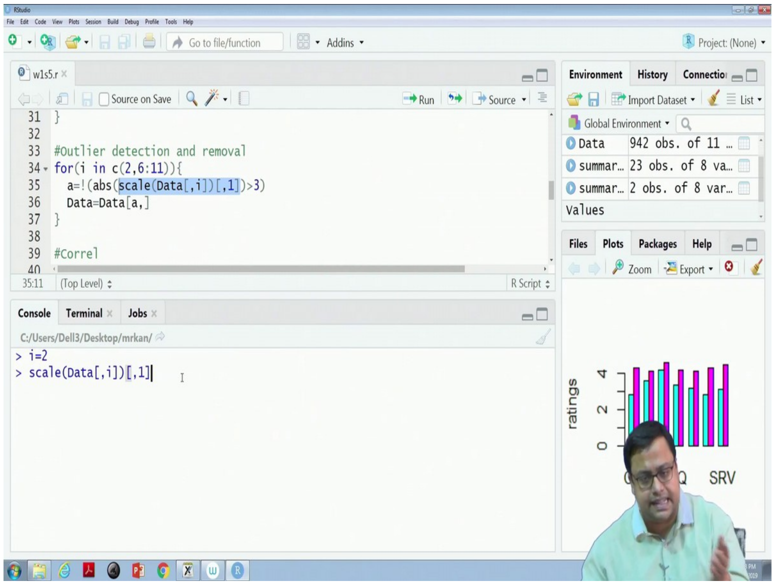
$\boxed{|z| < 3}$

So, there are various ways to define an outlier. Outlier is a value that is very away from the mean. And how do we decide? If there is an x, which is a long variable, we find out mean of x, which is 'm' and we find out the standard deviation of x which is 's' and anything, which is beyond let us say x is within m+3s and m-3s. This is not an outlier. 3 is something that I had taken, if you are a little bit more conservative, you can take 2 as per your wish.

So, x as long as it is within this limit it is not a outlier. So, I can also write (x-m) has to be within -3s to 3s or I can write (x-m)/s has to be within -3 and 3. That means I can also write x-m is what, actually z, z = (x-m)/s, so z has to be between -3 to 3.

That means the mod of z has to be lower than 3, when it is higher than 3 it is outlier, when this is lower than 3, I will keep it as simple as that. So, that is what I have done here. In my code I found out the z value for each of the columns, and I have checked that whether that z value is lower than 3 or not.

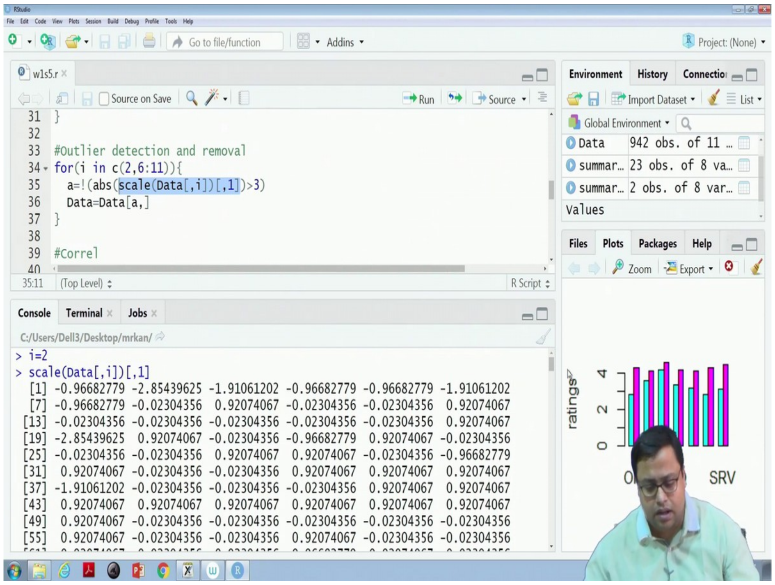
(Refer Slide Time: 20:27)



```
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=!(abs(scale(Data[,i])[,1])>3)
  Data=Data[a,]
}

#Correl
```

```
C:/Users/Dell3/Desktop/mrkan/
> i=2
> scale(Data[,i])[,1]
```

Data	942 obs. of 11 ...
summar...	23 obs. of 8 va...
summar...	2 obs. of 8 var...
Values	



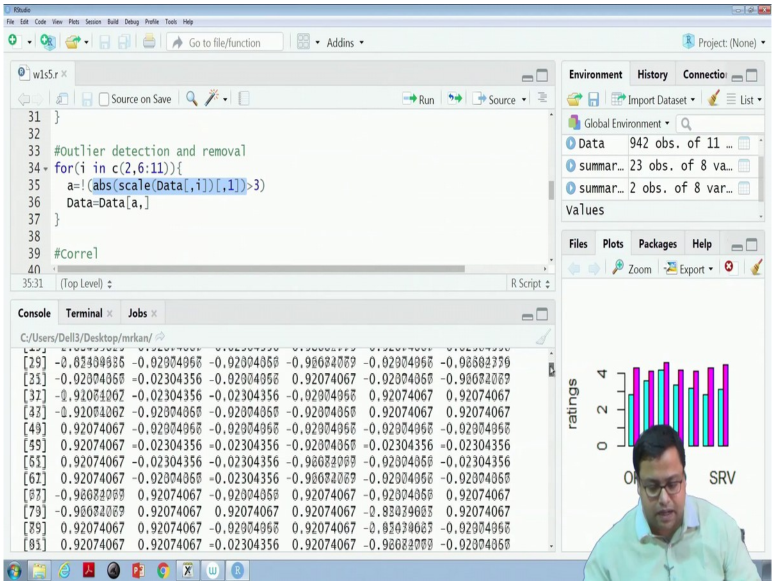
```
31  }
32
33  #Outlier detection and removal
34  for(i in c(2,6:11)){
35    a=!(abs(scale(Data[,i])[,1])>3)
36    Data=Data[a,]
37  }
38
39  #Correl
```

35:11 (Top Level) R Script

```
Data      942 obs. of 11 ...
summar... 23 obs. of 8 va...
summar... 2 obs. of 8 var...
Values
```

C:/Users/Dell3/Desktop/mrkan/

```
> i=2
> scale(Data[,i])[,1]
  [1] -0.96682779 -2.85439625 -1.91061202 -0.96682779 -0.96682779 -1.91061202
  [7] -0.96682779 -0.02304356  0.92074067 -0.02304356 -0.02304356  0.92074067
 [13] -0.02304356 -0.02304356 -0.02304356 -0.02304356 -0.02304356  0.92074067
 [19] -2.85439625  0.92074067 -0.02304356 -0.96682779  0.92074067 -0.02304356
 [25] -0.02304356 -0.02304356  0.92074067  0.92074067 -0.02304356 -0.96682779
 [31]  0.92074067 -0.02304356 -0.02304356  0.92074067  0.92074067  0.92074067
 [37] -1.91061202 -0.02304356 -0.02304356 -0.02304356  0.92074067  0.92074067
 [43]  0.92074067  0.92074067  0.92074067  0.92074067  0.92074067  0.92074067
 [49]  0.92074067 -0.02304356 -0.02304356 -0.02304356 -0.02304356 -0.02304356
 [55]  0.92074067 -0.02304356 -0.02304356  0.92074067 -0.02304356 -0.02304356
```



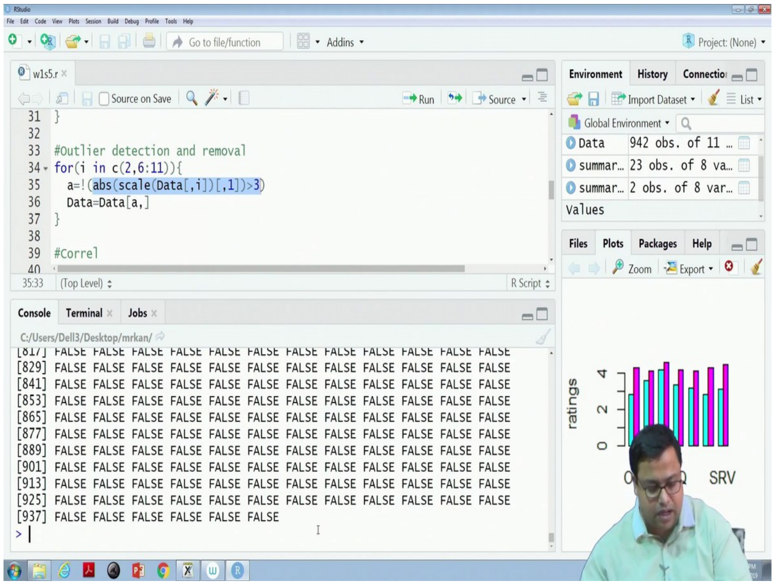
```
31  }
32
33  #Outlier detection and removal
34  for(i in c(2,6:11)){
35    a=!(abs(scale(Data[,i])[,1])>3)
36    Data=Data[a,]
37  }
38
39  #Correl
```

35:31 (Top Level) R Script

```
Data      942 obs. of 11 ...
summar... 23 obs. of 8 va...
summar... 2 obs. of 8 var...
Values
```

C:/Users/Dell3/Desktop/mrkan/

```
 [29] -0.02304356 -0.02304356 -0.02304356 -0.96682779 -0.02304356 -0.96682779
 [31] -0.02304356 -0.02304356 -0.02304356  0.92074067 -0.02304356 -0.96682779
 [37] -1.91061202 -0.02304356 -0.02304356 -0.02304356  0.92074067  0.92074067
 [43] -1.91061202 -0.02304356 -0.02304356 -0.02304356  0.92074067  0.92074067
 [49]  0.92074067 -0.02304356 -0.02304356 -0.02304356 -0.02304356 -0.02304356
 [55]  0.92074067 -0.02304356 -0.02304356 -0.02304356 -0.02304356 -0.02304356
 [61]  0.92074067 -0.02304356 -0.02304356 -0.96682779 -0.02304356 -0.02304356
 [67]  0.92074067 -0.02304356 -0.02304356 -0.96682779 -0.02304356 -0.02304356
 [73] -0.96682779  0.92074067 -0.02304356  0.92074067 -0.02304356  0.92074067
 [79] -0.96682779  0.92074067  0.92074067  0.92074067 -2.85439625  0.92074067
 [85]  0.92074067  0.92074067 -0.02304356  0.92074067 -2.85439625 -0.02304356
 [91]  0.92074067  0.92074067 -0.02304356  0.92074067 -0.96682779 -0.02304356
```



```
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=!(abs(scale(Data[,i])[,1])>3)
  Data=Data[a,]
}

#Correl
```

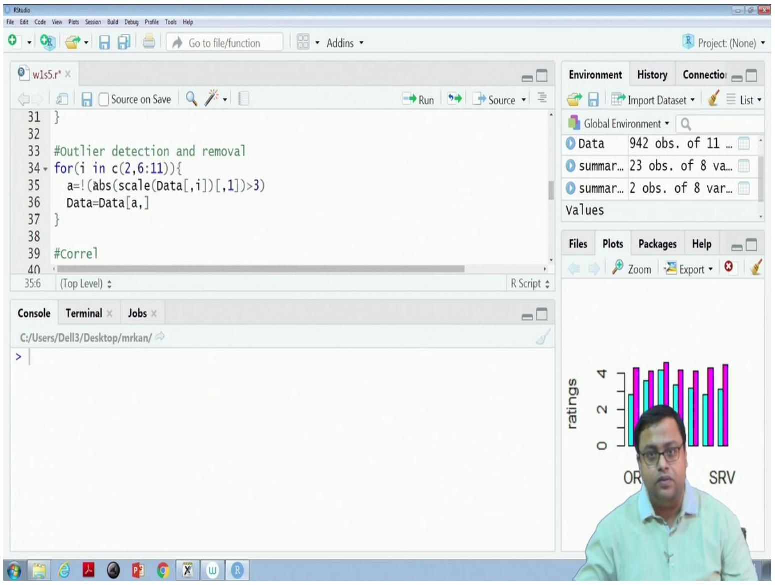
```
[817] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[829] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[841] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[853] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[865] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[877] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[889] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[901] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[913] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[937] FALSE FALSE FALSE FALSE FALSE FALSE
> 
```

So, what did I do in this part? So, let us say i is equal to 2. Then what does this do? Only this part does his work, it takes data's second column and scales it, scale it means it changes it to its z values. So, all the values that you are saying that is the second column z value; these are the z value of the second column, which is the overall rating. So, it has turned.

Now, I want the absolute value mod of z. So, abs is the function. So, abs will give me the same things, but in an absolute form, that means it will say that all the negatives will be positive and all the positive will be positive. So, see here are all the values coming are positive, no negatives, everything is positive.

Now, what am I doing? I am checking whether they are higher than 3 or not? If they are higher than 3 it will come as true, if it is lower than 3 it will come as false. So, all those guys are coming false. There might be some trues in between; I am not checking it right now. So, there will be trues and false.

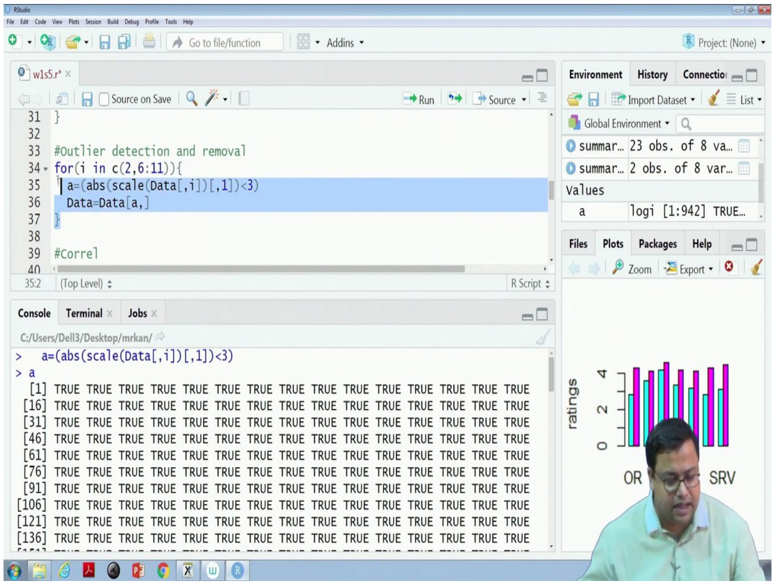
(Refer Slide Time: 21:45)



```
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=!(abs(scale(Data[,i])[,1])>3)
  Data=Data[a,]
}

#Correl
```

```
C:/Users/Dell3/Desktop/mrkan/
> 
```

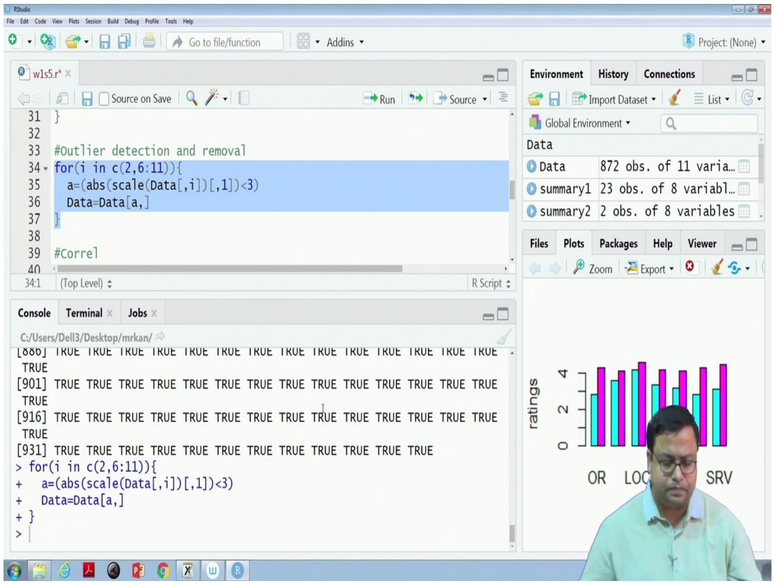
```
Data       942 obs. of 11 …
summar…    23 obs. of 8 va…
summar…    2 obs. of 8 var…
Values
```



```
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=(abs(scale(Data[,i])[,1])<3)
  Data=Data[a,]
}

#Correl
```

```
C:/Users/Dell3/Desktop/mrkan/
>   a=(abs(scale(Data[,i])[,1])<3)
> a
  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
summar…    23 obs. of 8 va…
summar…    2 obs. of 8 var…
Values
a          logi [1:942] TRUE…
```



```r
}

#Outlier detection and removal
for(i in c(2,6:11)){
  a=(abs(scale(Data[,i])[,1])<3)
  Data=Data[a,]
}

#Correl
```

C:/Users/Dell3/Desktop/mrkan/

```
[886] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[901] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[916] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[931] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> for(i in c(2,6:11)){
+   a=(abs(scale(Data[,i])[,1])<3)
+   Data=Data[a,]
+ }
> 
```

Data	
Data	872 obs. of 11 varia…
summary1	23 obs. of 8 variabl…
summary2	2 obs. of 8 variables

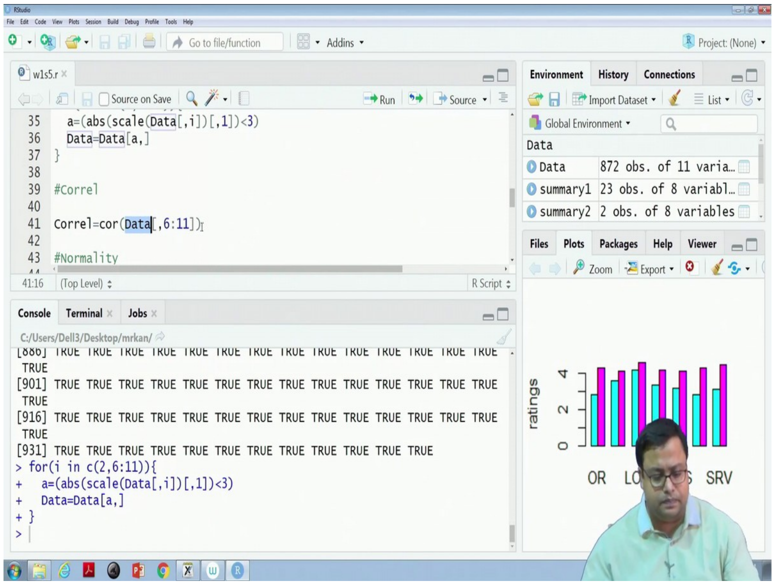
I want to keep the false ones. So, instead of doing this, let us make it simple, delete this one, just check what I am deleting. There was an exclamation sign which was not of whatever I was doing on the right side. So, I am removing this, and I am putting a smaller than sign. Smaller than sign means it will give me true values when guys are within 3, when the mod of z is lower than 3. It will give me true. And those are the values I want to keep in my data set because they are not outliers, so I am putting that. So, this is my 'a', this is my 'a,' for all the trues means these are not outliers.

So, if there is some by chance some falses in between then, that will be an outlier and take that subset of the data in such a way that I am taking only the trues and not the false. So, I am taking only such rows what this particular column is coming as to be true, rather than false, if it is by chance coming false, do not take the whole row, the whole row you just remove it.

So, this is the thing that I am running, so you stop the video again. Or go back to the few steps again in the video and listen and understand what I have done. The moment I do that you will see if I run this particular line. You will see that we have 952 or something like that from there I got

872, which is a huge drop. So, there will be around overall 80 odd outliers, and we have removed that.

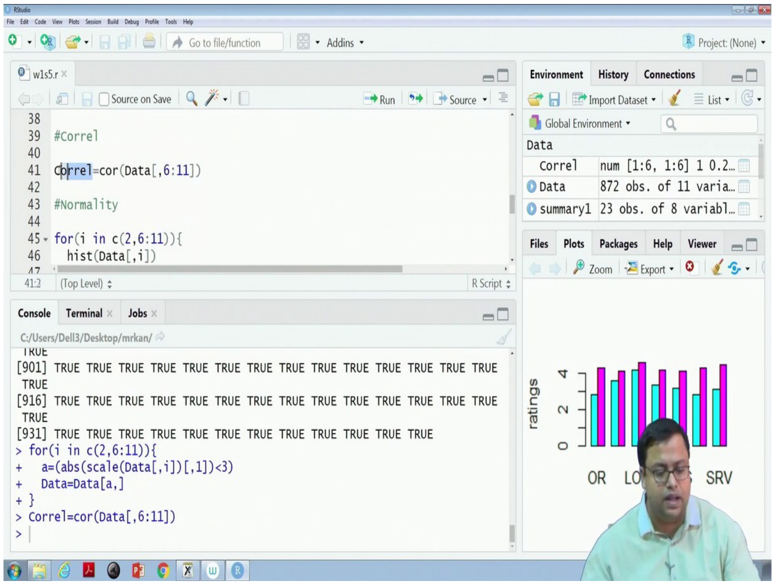
(Refer Slide Time: 23:29)



```
  a=(abs(scale(Data[,i])[,1])<3)
  Data=Data[a,]
}

#Correl

Correl=cor(Data[,6:11])

#Normality
```

```
C:/Users/Dell3/Desktop/mrkan/
[886] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[901] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[916] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[931] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> for(i in c(2,6:11)){
+   a=(abs(scale(Data[,i])[,1])<3)
+   Data=Data[a,]
+ }
> 
```

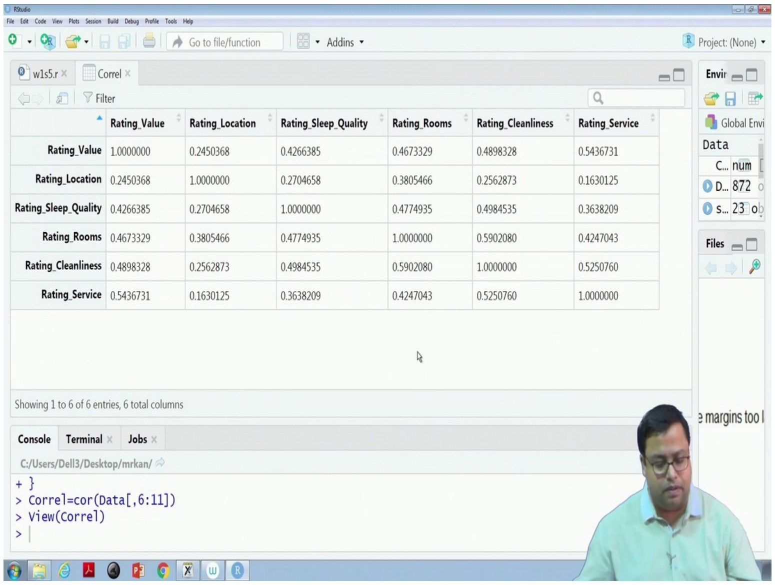
Data	
Data	872 obs. of 11 varia...
summary1	23 obs. of 8 variabl...
summary2	2 obs. of 8 variables



```

#Correl

Correl=cor(Data[,6:11])

#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
```

```
C:/Users/Dell3/Desktop/mrkan/
 TRUE
[901] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[916] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 TRUE
[931] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> for(i in c(2,6:11)){
+   a=(abs(scale(Data[,i])[,1])<3)
+   Data=Data[a,]
+ }
> Correl=cor(Data[,6:11])
> 
```

Data	
Correl	num [1:6, 1:6] 1 0.2...
Data	872 obs. of 11 varia...
summary1	23 obs. of 8 variabl...



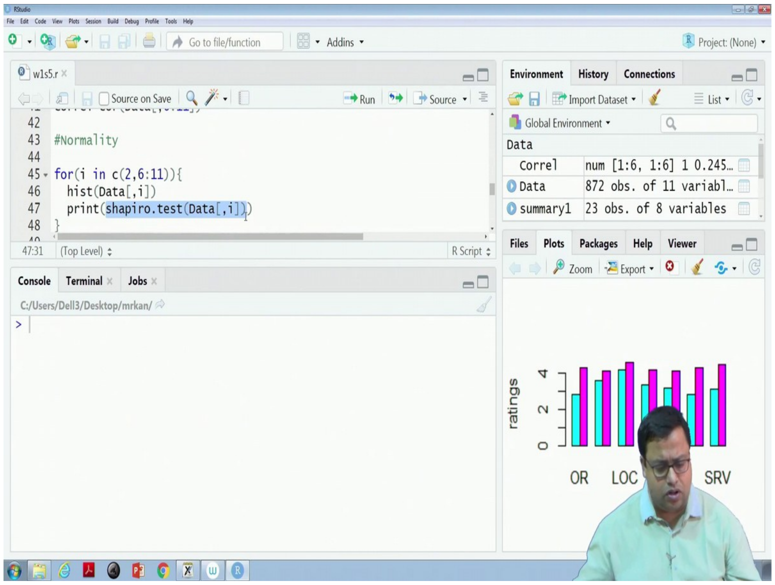
	Rating_Value	Rating_Location	Rating_Sleep_Quality	Rating_Rooms	Rating_Cleanliness	Rating_Service
Rating_Value	1.0000000	0.2450368	0.4266385	0.4673329	0.4898328	0.5436731
Rating_Location	0.2450368	1.0000000	0.2704658	0.3805466	0.2562873	0.1630125
Rating_Sleep_Quality	0.4266385	0.2704658	1.0000000	0.4774935	0.4984535	0.3638209
Rating_Rooms	0.4673329	0.3805466	0.4774935	1.0000000	0.5902080	0.4247043
Rating_Cleanliness	0.4898328	0.2562873	0.4984535	0.5902080	1.0000000	0.5250760
Rating_Service	0.5436731	0.1630125	0.3638209	0.4247043	0.5250760	1.0000000

Showing 1 to 6 of 6 entries, 6 total columns

```
+ }
> Correl=cor(Data[,6:11])
> View(Correl)
> 
```

The next step is correlation. So, I removed that, I only want to find out the x variables correlation 6 to 11. So, I am writing that correlation of what data is my data set. That data sets 6 to 11th column. So, if I run this that gets saved in something called correl. Correl is a matrix. I click on that and this gives me the matrix. And if you carefully see all the values in the matrix are lower than 0.6, I think so. Let us check, yes, all the values are lower than 0.6; it is coming lower than 0.6. So, there is no very hard and fast I would say multi collinearity issue or correlational problem right now, as of now.

(Refer Slide Time: 24:23)



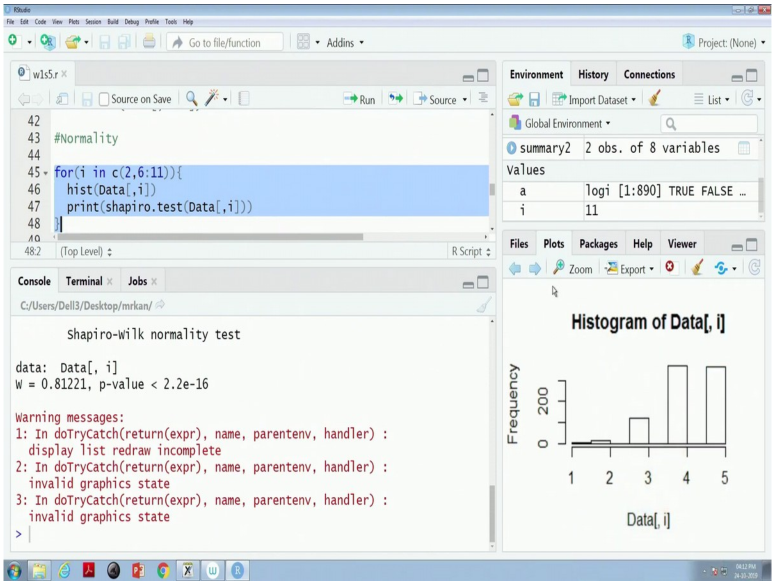
```
#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
  print(shapiro.test(Data[,i]))
}
```

C:/Users/Dell3/Desktop/mrkan/

```
>
```

Data

Correl	num [1:6, 1:6] 1 0.245...
Data	872 obs. of 11 variabl...
summary1	23 obs. of 8 variables



```
#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
  print(shapiro.test(Data[,i]))
}
```

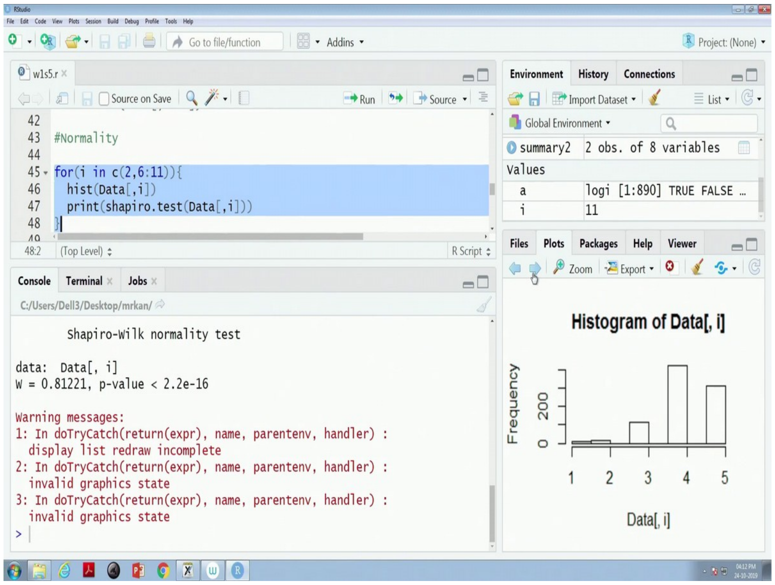
C:/Users/Dell3/Desktop/mrkan/

```

	Shapiro-Wilk normality test

data:  Data[, i]
W = 0.81221, p-value < 2.2e-16

Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
>
```

summary2	2 obs. of 8 variables

Values

a	logi [1:890] TRUE FALSE ...
i	11

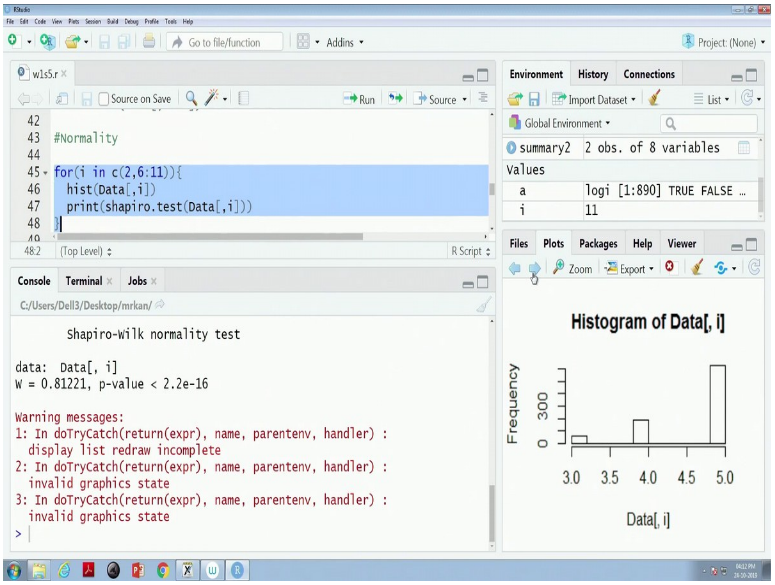


```r
#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
  print(shapiro.test(Data[,i]))
}
```

summary2	2 obs. of 8 variables
Values	
a	logi [1:890] TRUE FALSE …
i	11

C:/Users/Dell3/Desktop/mrkan/

```
        Shapiro-Wilk normality test

data:  Data[, i]
W = 0.81221, p-value < 2.2e-16

Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
>
```

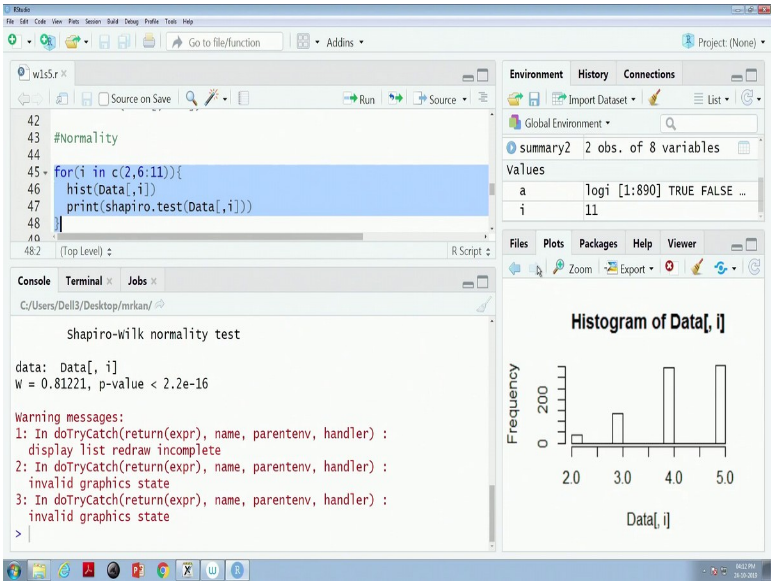


```r
#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
  print(shapiro.test(Data[,i]))
}
```

summary2	2 obs. of 8 variables
Values	
a	logi [1:890] TRUE FALSE …
i	11

C:/Users/Dell3/Desktop/mrkan/

```
        Shapiro-Wilk normality test

data:  Data[, i]
W = 0.81221, p-value < 2.2e-16

Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
>
```



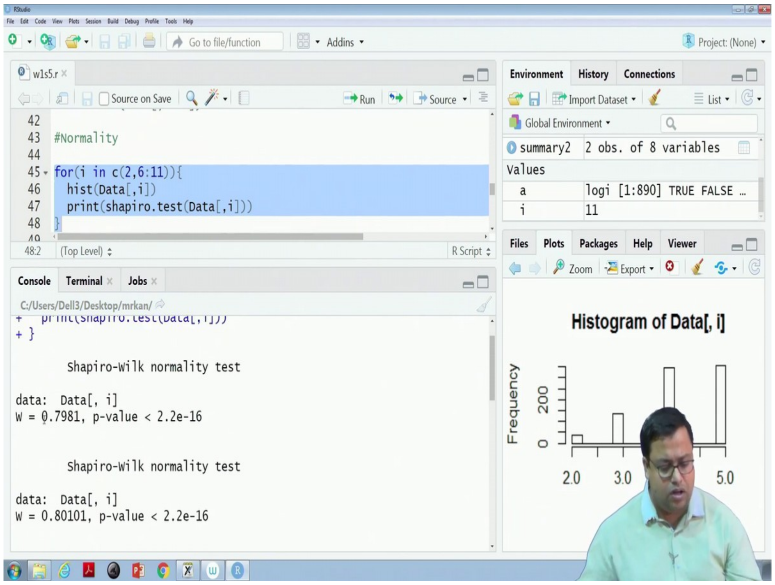
```
#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
  print(shapiro.test(Data[,i]))
}
```

```
        Shapiro-Wilk normality test

data:  Data[, i]
W = 0.81221, p-value < 2.2e-16

Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
>
```

And then we have to also check for normality. Another important thing is that all the x and y variables are normal. So, what we have to do is I have to find out the histogram of everything from the x variable, which is 2, and then 6 to 11 all the variables is a histogram. And there is something called Shapiro test which is a goodness of fit test for normality. So, it checks that whether it is normal or not, there is a w coefficient that calculates that value has to be higher than 1 if it is lower than 1 it is not normal.

So, if I just run this, you will see that there are so many diagrams that got created and I will show you one by one. So, the first one is overall rating. This is not normal. It cannot be normal because it is a categorical variable actually 1, 2, 3, 4, 5. Again this is not normal. This is not normal. And so on. So, these can be because they are categorical variables. They are not normal, but still, they are probably a little bit of up and down is being seen.

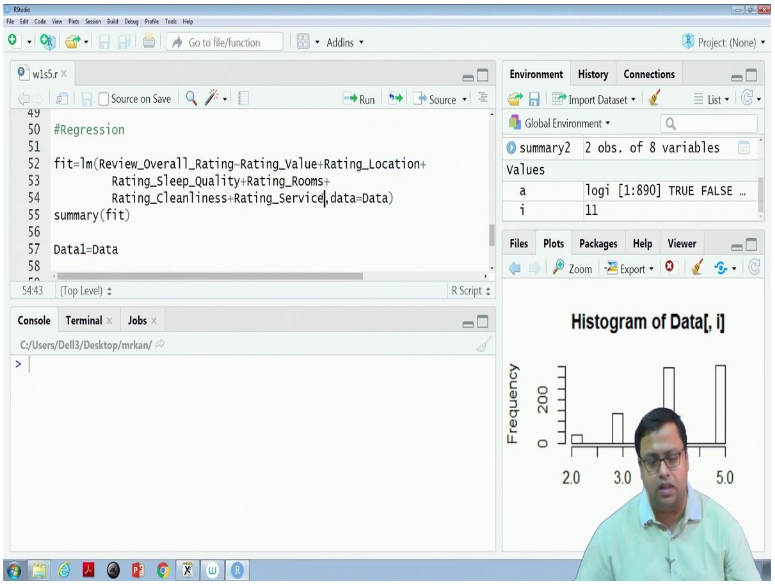
(Refer Slide Time: 25:32)



```
#Normality

for(i in c(2,6:11)){
  hist(Data[,i])
  print(shapiro.test(Data[,i]))
}
```

```
+   print(shapiro.test(Data[,i]))
+ }

        Shapiro-Wilk normality test

data:  Data[, i]
W = 0.7981, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  Data[, i]
W = 0.80101, p-value < 2.2e-16
```

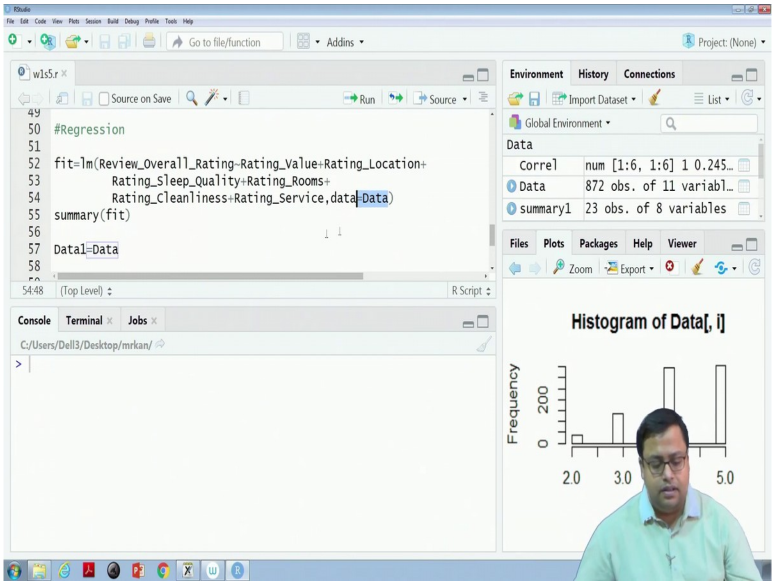
```
summary2    2 obs. of 8 variables
Values
a           logi [1:890] TRUE FALSE ...
i           11
```



```
#Regression

fit=lm(Review_Overall_Rating~Rating_Value+Rating_Location+
         Rating_Sleep_Quality+Rating_Rooms+
         Rating_Cleanliness+Rating_Service,data=Data)
summary(fit)

Data1=Data
```

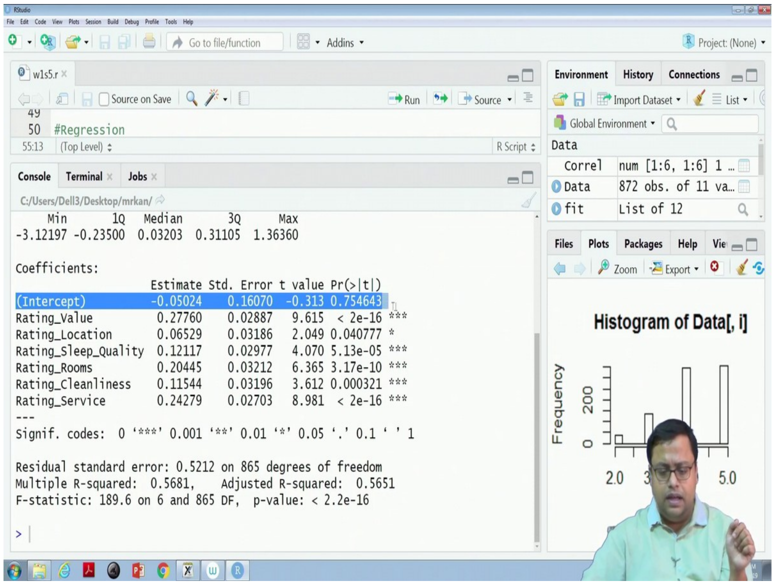
```
>
```

```
summary2    2 obs. of 8 variables
Values
a           logi [1:890] TRUE FALSE ...
i           11
```



```
#Regression

fit=lm(Review_Overall_Rating~Rating_Value+Rating_Location+
         Rating_Sleep_Quality+Rating_Rooms+
         Rating_Cleanliness+Rating_Service,data=Data)
summary(fit)

Data1=Data
```

```
Data
  Correl    num [1:6, 1:6] 1 0.245...
  Data      872 obs. of 11 variabl...
  summary1  23 obs. of 8 variables
```



```
     Min       1Q   Median       3Q      Max
-3.12197 -0.23500  0.03203  0.31105  1.36360

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.05024    0.16070  -0.313 0.754643
Rating_Value          0.27760    0.02887   9.615  < 2e-16 ***
Rating_Location       0.06529    0.03186   2.049 0.040777 *
Rating_Sleep_Quality  0.12117    0.02977   4.070 5.13e-05 ***
Rating_Rooms          0.20445    0.03212   6.365 3.17e-10 ***
Rating_Cleanliness    0.11544    0.03196   3.612 0.000321 ***
Rating_Service        0.24279    0.02703   8.981  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5212 on 865 degrees of freedom
Multiple R-squared:  0.5681,    Adjusted R-squared:  0.5651
F-statistic: 189.6 on 6 and 865 DF,  p-value: < 2.2e-16

>
```

```
Data
  Correl    num [1:6, 1:6] 1 ...
  Data      872 obs. of 11 va...
  fit       List of 12
```

Now, often times, we ignore these normalities. Because our major goal is not to focus on the rightness of econometrics techniques, our, sometimes our goal is to get an overall vague level idea that which one is more important, which one is less important. So, sometimes we ignore it in many many research papers of marketing people have ignored the normality part. But you probably should consider it that these are not normal. And if you have check this thing, you will see that all the w values, the, I have checked it for all this thing. All the w values coming lower than 1. So, they are none of them thing is normal.

So ideally, if it is a not normal you should not do linear regression but for current purpose I will do a linear regression, and then I will do something else also which will deal with this particular normality issue. But linear regression is something that most people know. Most people have an idea about those who are doing this particular course probably.

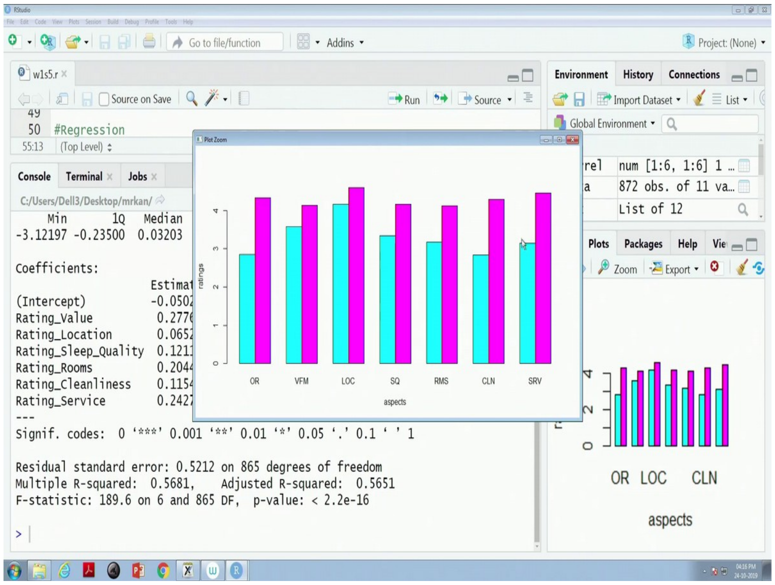
So, I will discuss about linear regression in further details in a different class, but here I am just showing you how to do linear regression. Which is, fit is the value where all the dataset gets saved, and all the results of the linear regression get saved and fit is equal to I have written in lm. lm stand for linear model and then within bracket.

My y variable, which is review overall rating and then tilde and then x1 plus x2 plus x3 up to services, so all the aspects I just write it down and data is equal to the data set name, which is capital DATA coming from here. So, once I run this and want to see the summary of my fit, the summary of the result I see something like this, so the call and then and this is the first thing that I will see that the F statistic and the corresponding p value. The p value is coming 2.2e^(-16) So, that means 0, 0, 0, 0, 0, 15 zeros after the decimal point and then 2 2. So, that means very small, that means the F statistic is significant. So, I can believe on this model.

Then the adjusted R square is 0.5651 it is good enough at least in the context of marketing it can be further improved, but as of now it is good enough, the intercept is also coming not significant that means see the how to know not significant the p value is higher than 0.05. So, I can say that okay, fairly good model. And I can find out all this value, location, etcetera has been measure in 1 to 5 point scale.

So, I can probably compare them; another way of comparing is before you do this regression you change all these 6 values to their scale form. So, you change to their standardized form; then you do the regression. So, that might be a better step, whatever, if you do the regression, I can see from the data set that service is the most important part. So, 0.24 is service and 0.11 is cleanliness, and 0.20 is rooms, and 0.12 is sleep quality. So, how that matters?

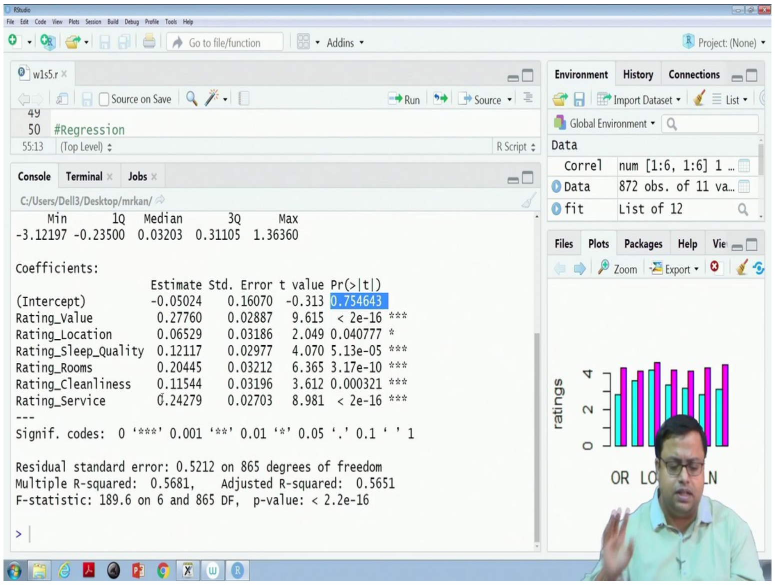
(Refer Slide Time: 29:07)



```
49
50 #Regression
```

```
C:/Users/Dell3/Desktop/mrkan/
     Min       1Q   Median
-3.12197 -0.23500  0.03203

Coefficients:
                    Estimat
(Intercept)         -0.0502
Rating_Value         0.2776
Rating_Location      0.0652
Rating_Sleep_Quality 0.1211
Rating_Rooms         0.2044
Rating_Cleanliness   0.1154
Rating_Service       0.2427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5212 on 865 degrees of freedom
Multiple R-squared:  0.5681,	Adjusted R-squared:  0.5651 
F-statistic: 189.6 on 6 and 865 DF,  p-value: < 2.2e-16

> |
```

```
rel   num [1:6, 1:6] 1 ...
a     872 obs. of 11 va...
      List of 12
```



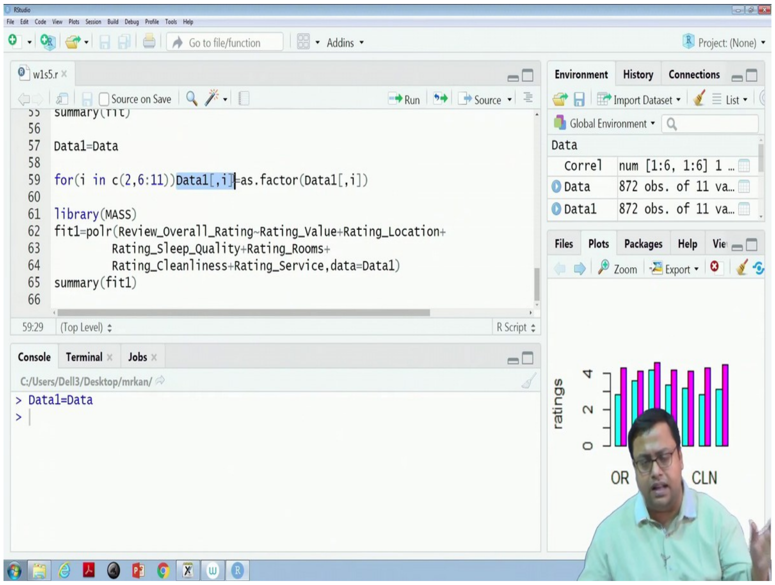
```
49
50 #Regression
```

```
C:/Users/Dell3/Desktop/mrkan/
     Min       1Q   Median       3Q      Max 
-3.12197 -0.23500  0.03203  0.31105  1.36360 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -0.05024    0.16070  -0.313 0.754643    
Rating_Value          0.27760    0.02887   9.615  < 2e-16 ***
Rating_Location       0.06529    0.03186   2.049 0.040777 *  
Rating_Sleep_Quality  0.12117    0.02977   4.070 5.13e-05 ***
Rating_Rooms          0.20445    0.03212   6.365 3.17e-10 ***
Rating_Cleanliness    0.11544    0.03196   3.612 0.000321 ***
Rating_Service        0.24279    0.02703   8.981  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5212 on 865 degrees of freedom
Multiple R-squared:  0.5681,	Adjusted R-squared:  0.5651 
F-statistic: 189.6 on 6 and 865 DF,  p-value: < 2.2e-16

> |
```

```
Data
  Correl   num [1:6, 1:6] 1 ...
  Data     872 obs. of 11 va...
  fit      List of 12
```

If I see now this particular, this particular one, so I was doing cleanliness was a bigger problem for me. So this jump is much bigger than this, but service is twice as important as cleanliness. So, then probably the weightage that the service gets is much higher than what the weightage the cleanliness gets. And you have to think about that. When you are dealing with this thing, and probably in comparison to the, all the aspects room and service are more or less similar, and sleep quality and cleanliness are more or less similar. And the location is the least important factor in this particular dataset that I have created; the location is the least important factor.

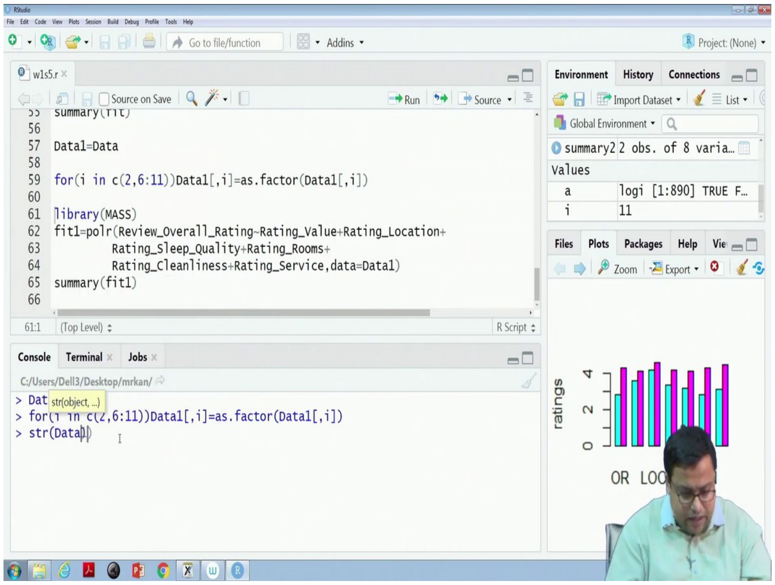
(Refer Slide Time: 29:51)



```
summary(fit)

Data1=Data

for(i in c(2,6:11))Data1[,i]=as.factor(Data1[,i])

library(MASS)
fit1=polr(Review_Overall_Rating~Rating_Value+Rating_Location+
          Rating_Sleep_Quality+Rating_Rooms+
          Rating_Cleanliness+Rating_Service,data=Data1)
summary(fit1)
```

```
> Data1=Data
> 
```

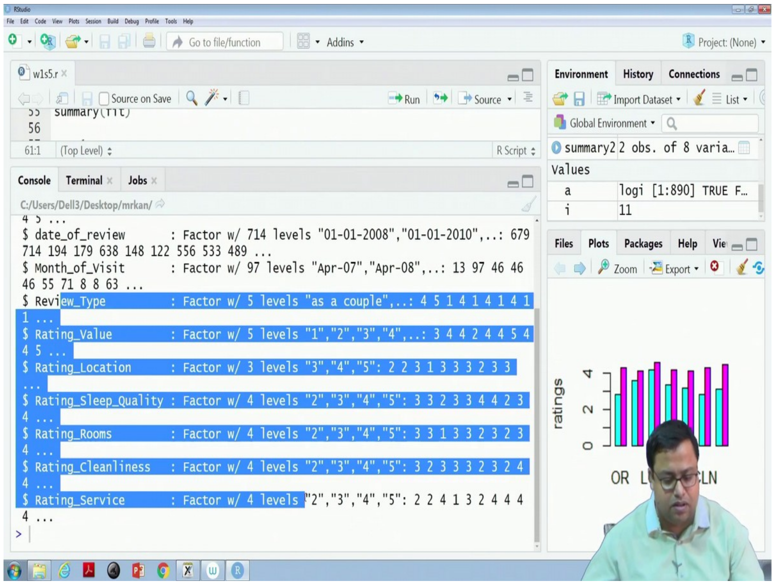
```
Data
  Correl   num [1:6, 1:6] 1 ...
  Data     872 obs. of 11 va...
  Data1    872 obs. of 11 va...
```



```
summary(fit)

Data1=Data

for(i in c(2,6:11))Data1[,i]=as.factor(Data1[,i])

library(MASS)
fit1=polr(Review_Overall_Rating~Rating_Value+Rating_Location+
          Rating_Sleep_Quality+Rating_Rooms+
          Rating_Cleanliness+Rating_Service,data=Data1)
summary(fit1)
```

```
> Dat str(object, ...)
> for(i in c(2,6:11))Data1[,i]=as.factor(Data1[,i])
> str(Data1)
```

```
  summary2 2 obs. of 8 varia...
Values
  a        logi [1:890] TRUE F...
  i        11
```



```
55  summary(fit)
56
```

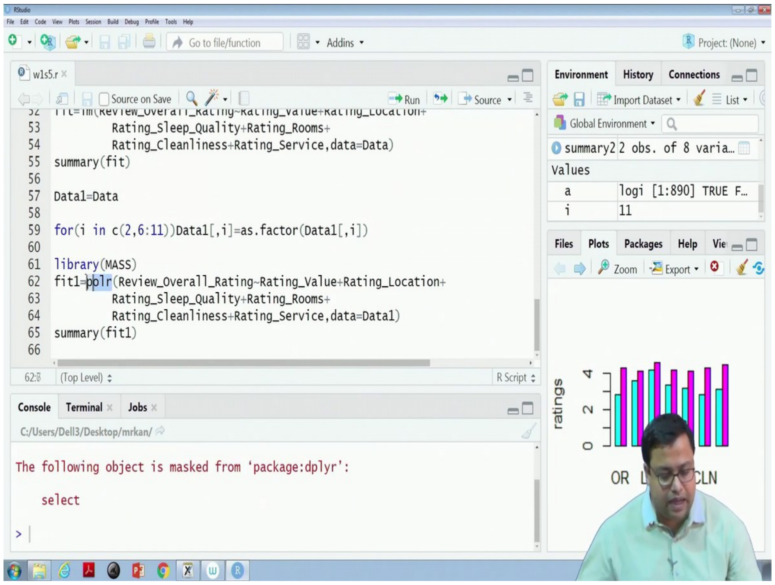
```
C:/Users/Dell3/Desktop/mrkan/
4 5 ...
$ date_of_review       : Factor w/ 714 levels "01-01-2008","01-01-2010",..: 679
714 194 179 638 148 122 556 533 489 ...
$ Month_of_Visit       : Factor w/ 97 levels "Apr-07","Apr-08",..: 13 97 46 46
46 55 71 8 8 63 ...
$ Review_Type          : Factor w/ 5 levels "as a couple",..: 4 5 1 4 1 4 1 4 1
1 ...
$ Rating_Value         : Factor w/ 5 levels "1","2","3","4",..: 3 4 4 2 4 4 5 4
4 5 ...
$ Rating_Location      : Factor w/ 3 levels "3","4","5": 2 2 3 1 3 3 3 2 3 3
...
$ Rating_Sleep_Quality : Factor w/ 4 levels "2","3","4","5": 3 3 2 3 3 4 4 2 3
4 ...
$ Rating_Rooms         : Factor w/ 4 levels "2","3","4","5": 3 3 1 3 3 2 3 2 3
4 ...
$ Rating_Cleanliness   : Factor w/ 4 levels "2","3","4","5": 3 2 3 3 3 2 3 2 4
4 ...
$ Rating_Service       : Factor w/ 4 levels "2","3","4","5": 2 2 4 1 3 2 4 4 4
4 ...
>
```

```
summary2  2 obs. of 8 varia…
Values
a   logi [1:890] TRUE F…
i   11
```

Now, the last step, and this is the step for those people who have a little bit of idea about econometrics and who thinks that linear regression is not the best choice here. What we have done is something called ordered logic. My y variable is 1, 2, 3, 4, 5. So, they are ordered, so ordered logistic regression is another way of dealing with this. So, what we did is we have actually converted all our x variable and y variable to their factor form. So, I took data 1 is equal to data another dataset, I have created another dataset.

And from 2 and 6 to 11 for all these columns. The y column and the x columns I have since written data 1$ i, i. That means that the data 1, the new dataset's ith column is the factor version of that particular variable. So, I have changed everything to factor. So, now if I write str data1, if we want to see the structure of data 1, you will see all these guys were integer now they have become factors. These guys have become factors; these guys have become factors. So, the structure says that the new dataset has all the x variables as factors.

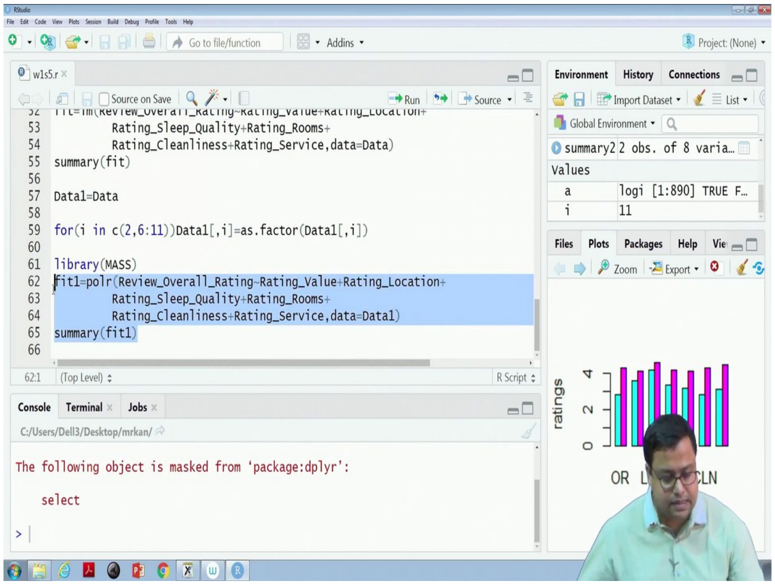
(Refer Slide Time: 31:04)



```
fit=lm(Review_Overall_Rating~Rating_Value+Rating_Location+
        Rating_Sleep_Quality+Rating_Rooms+
        Rating_Cleanliness+Rating_Service,data=Data)
summary(fit)

Data1=Data

for(i in c(2,6:11))Data1[,i]=as.factor(Data1[,i])

library(MASS)
fit1=polr(Review_Overall_Rating~Rating_Value+Rating_Location+
        Rating_Sleep_Quality+Rating_Rooms+
        Rating_Cleanliness+Rating_Service,data=Data1)
summary(fit1)
```

```
summary2  2 obs. of 8 varia…
Values
  a       logi [1:890] TRUE F…
  i       11
```

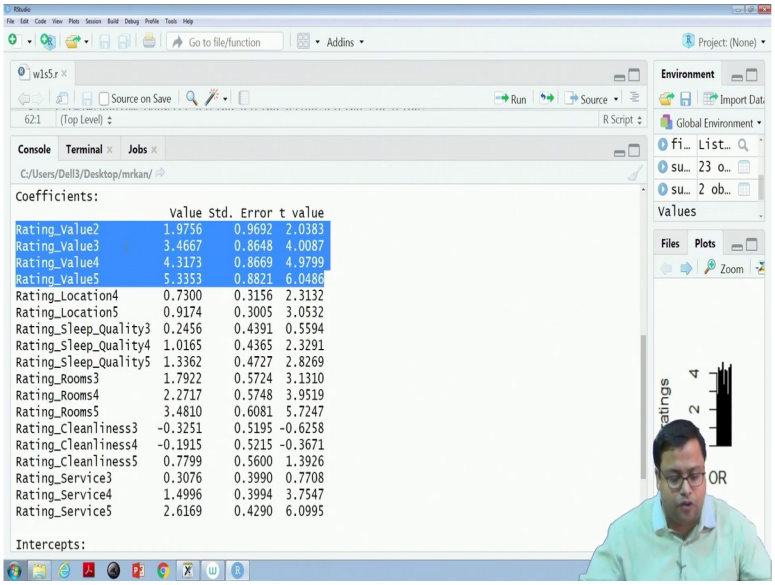
```
The following object is masked from 'package:dplyr':

    select

>
```



```
fit=lm(Review_Overall_Rating~Rating_Value+Rating_Location+
        Rating_Sleep_Quality+Rating_Rooms+
        Rating_Cleanliness+Rating_Service,data=Data)
summary(fit)

Data1=Data

for(i in c(2,6:11))Data1[,i]=as.factor(Data1[,i])

library(MASS)
fit1=polr(Review_Overall_Rating~Rating_Value+Rating_Location+
        Rating_Sleep_Quality+Rating_Rooms+
        Rating_Cleanliness+Rating_Service,data=Data1)
summary(fit1)
```

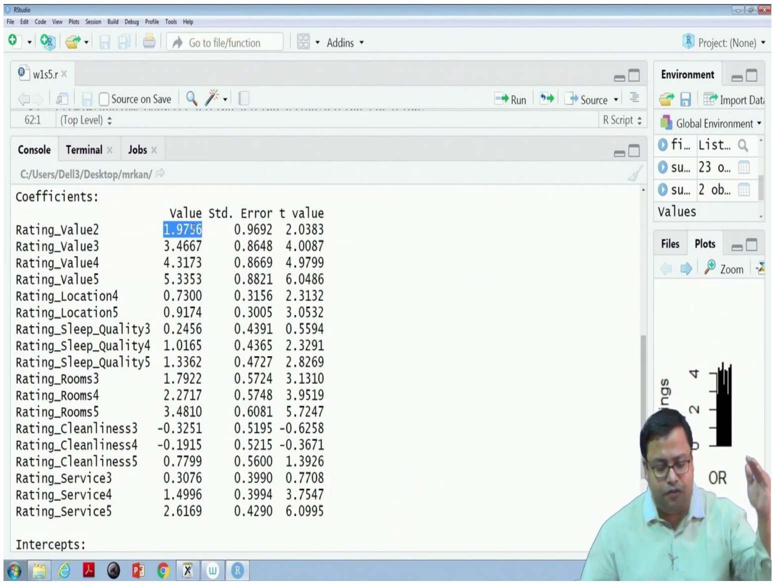
```
summary2  2 obs. of 8 varia…
Values
  a       logi [1:890] TRUE F…
  i       11
```

```
The following object is masked from 'package:dplyr':

    select

>
```



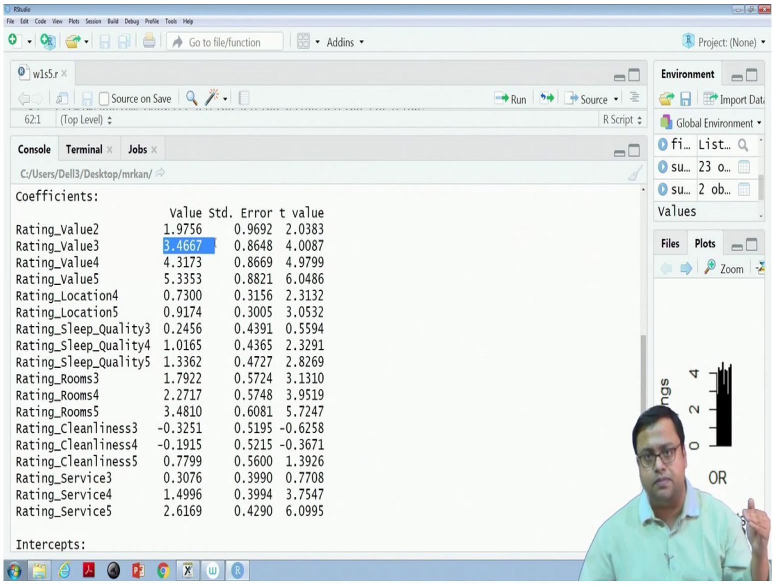
C:/Users/Dell3/Desktop/mrkan/

```
Coefficients:
                         Value Std. Error t value
Rating_Value2           1.9756     0.9692  2.0383
Rating_Value3           3.4667     0.8648  4.0087
Rating_Value4           4.3173     0.8669  4.9799
Rating_Value5           5.3353     0.8821  6.0486
Rating_Location4        0.7300     0.3156  2.3132
Rating_Location5        0.9174     0.3005  3.0532
Rating_Sleep_Quality3   0.2456     0.4391  0.5594
Rating_Sleep_Quality4   1.0165     0.4365  2.3291
Rating_Sleep_Quality5   1.3362     0.4727  2.8269
Rating_Rooms3           1.7922     0.5724  3.1310
Rating_Rooms4           2.2717     0.5748  3.9519
Rating_Rooms5           3.4810     0.6081  5.7247
Rating_Cleanliness3    -0.3251     0.5195 -0.6258
Rating_Cleanliness4    -0.1915     0.5215 -0.3671
Rating_Cleanliness5     0.7799     0.5600  1.3926
Rating_Service3         0.3076     0.3990  0.7708
Rating_Service4         1.4996     0.3994  3.7547
Rating_Service5         2.6169     0.4290  6.0995

Intercepts:
```

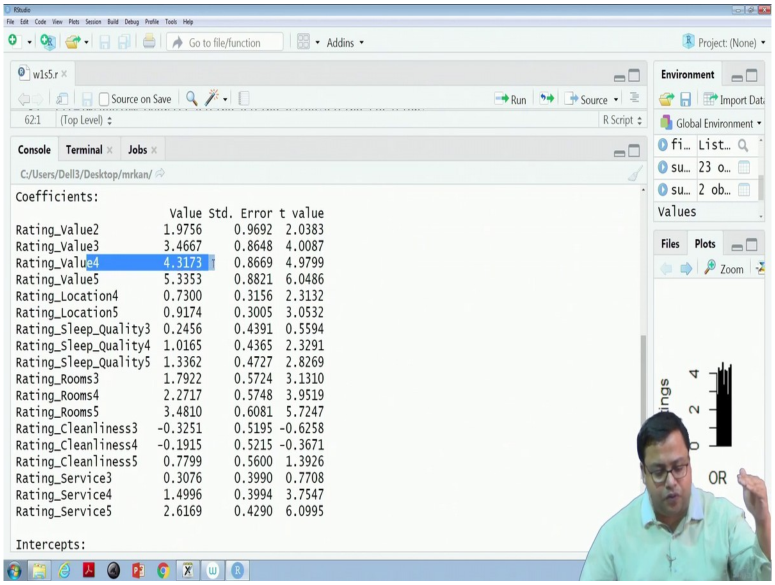


C:/Users/Dell3/Desktop/mrkan/

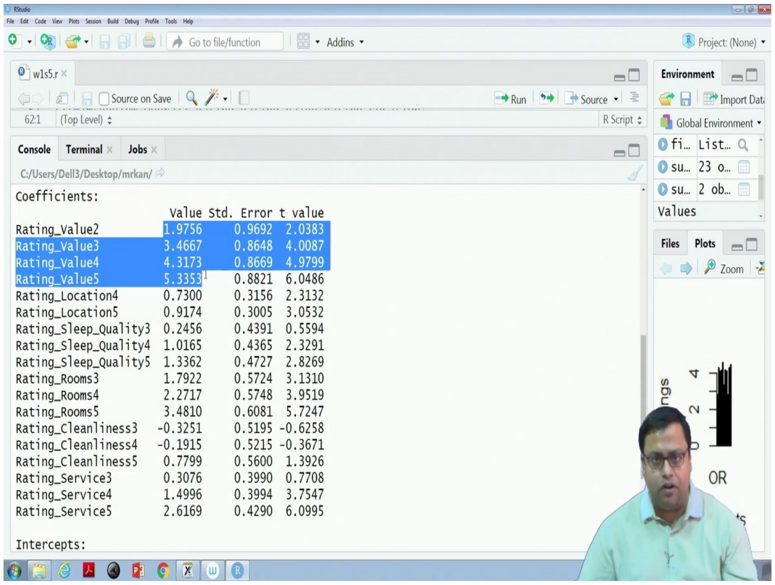
```
Coefficients:
                         Value Std. Error t value
Rating_Value2           1.9756     0.9692  2.0383
Rating_Value3           3.4667     0.8648  4.0087
Rating_Value4           4.3173     0.8669  4.9799
Rating_Value5           5.3353     0.8821  6.0486
Rating_Location4        0.7300     0.3156  2.3132
Rating_Location5        0.9174     0.3005  3.0532
Rating_Sleep_Quality3   0.2456     0.4391  0.5594
Rating_Sleep_Quality4   1.0165     0.4365  2.3291
Rating_Sleep_Quality5   1.3362     0.4727  2.8269
Rating_Rooms3           1.7922     0.5724  3.1310
Rating_Rooms4           2.2717     0.5748  3.9519
Rating_Rooms5           3.4810     0.6081  5.7247
Rating_Cleanliness3    -0.3251     0.5195 -0.6258
Rating_Cleanliness4    -0.1915     0.5215 -0.3671
Rating_Cleanliness5     0.7799     0.5600  1.3926
Rating_Service3         0.3076     0.3990  0.7708
Rating_Service4         1.4996     0.3994  3.7547
Rating_Service5         2.6169     0.4290  6.0995

Intercepts:
```



```
Coefficients:
                      Value Std. Error t value
Rating_Value2         1.9756     0.9692  2.0383
Rating_Value3         3.4667     0.8648  4.0087
Rating_Value4         4.3173     0.8669  4.9799
Rating_Value5         5.3353     0.8821  6.0486
Rating_Location4      0.7300     0.3156  2.3132
Rating_Location5      0.9174     0.3005  3.0532
Rating_Sleep_Quality3 0.2456     0.4391  0.5594
Rating_Sleep_Quality4 1.0165     0.4365  2.3291
Rating_Sleep_Quality5 1.3362     0.4727  2.8269
Rating_Rooms3         1.7922     0.5724  3.1310
Rating_Rooms4         2.2717     0.5748  3.9519
Rating_Rooms5         3.4810     0.6081  5.7247
Rating_Cleanliness3  -0.3251     0.5195 -0.6258
Rating_Cleanliness4  -0.1915     0.5215 -0.3671
Rating_Cleanliness5   0.7799     0.5600  1.3926
Rating_Service3       0.3076     0.3990  0.7708
Rating_Service4       1.4996     0.3994  3.7547
Rating_Service5       2.6169     0.4290  6.0995

Intercepts:
```



```
Coefficients:
                      Value Std. Error t value
Rating_Value2         1.9756     0.9692  2.0383
Rating_Value3         3.4667     0.8648  4.0087
Rating_Value4         4.3173     0.8669  4.9799
Rating_Value5         5.3353     0.8821  6.0486
Rating_Location4      0.7300     0.3156  2.3132
Rating_Location5      0.9174     0.3005  3.0532
Rating_Sleep_Quality3 0.2456     0.4391  0.5594
Rating_Sleep_Quality4 1.0165     0.4365  2.3291
Rating_Sleep_Quality5 1.3362     0.4727  2.8269
Rating_Rooms3         1.7922     0.5724  3.1310
Rating_Rooms4         2.2717     0.5748  3.9519
Rating_Rooms5         3.4810     0.6081  5.7247
Rating_Cleanliness3  -0.3251     0.5195 -0.6258
Rating_Cleanliness4  -0.1915     0.5215 -0.3671
Rating_Cleanliness5   0.7799     0.5600  1.3926
Rating_Service3       0.3076     0.3990  0.7708
Rating_Service4       1.4996     0.3994  3.7547
Rating_Service5       2.6169     0.4290  6.0995

Intercepts:
```



```
Coefficients:
                          Value Std. Error t value
Rating_Value2            1.9756     0.9692  2.0383
Rating_Value3            3.4667     0.8648  4.0087
Rating_Value4            4.3173     0.8669  4.9799
Rating_Value5            5.3353     0.8821  6.0486
Rating_Location4         0.7300     0.3156  2.3132
Rating_Location5         0.9174     0.3005  3.0532
Rating_Sleep_Quality3    0.2456     0.4391  0.5594
Rating_Sleep_Quality4    1.0165     0.4365  2.3291
Rating_Sleep_Quality5    1.3362     0.4727  2.8269
Rating_Rooms3            1.7922     0.5724  3.1310
Rating_Rooms4            2.2717     0.5748  3.9519
Rating_Rooms5            3.4810     0.6081  5.7247
Rating_Cleanliness3     -0.3251     0.5195 -0.6258
Rating_Cleanliness4     -0.1915     0.5215 -0.3671
Rating_Cleanliness5      0.7799     0.5600  1.3926
Rating_Service3          0.3076     0.3990  0.7708
Rating_Service4          1.4996     0.3994  3.7547
Rating_Service5          2.6169     0.4290  6.0995

Intercepts:
```

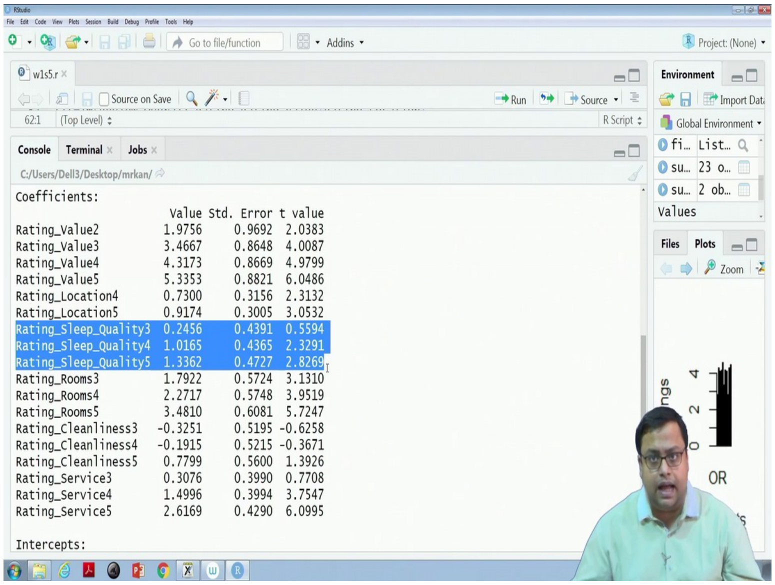
Now, there are library called mass, in which there is a polr function, which is actually a function which has been used for multinomial, actually ordinal logit, so, ordinal logit regression olr, so, fit 1 a new name I have given where all my details will be saved. Everything remains same, instead of LM, we write polr. That is all. So, polr make sure that this particular thing, the y variable is a categorical variable.

So, once I run this. I get a certain result, and I want to show you the result. So, here the p values are not coming, but it does not matter if the t is higher than 2, I can say that it is significant. So, the first thing is to see the rating value 1 got dropped; this categorical variable is a dummy variable. So, one of the dummy variables will get dropped, and the other will remain.

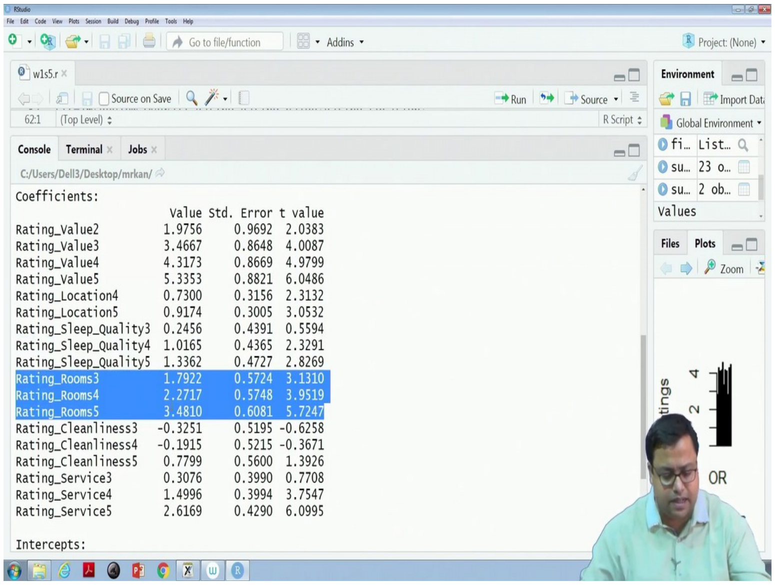
So here, for the rating value, 1 got dropped, and the other 4 are there, and you will see that you will carefully see that in comparison to rating value 1 that means value for money whoever is giving 1. If you by chance improve your value for money from 1 to 2. The overall thing increases by 1.9. If you further increase from 1 to 3, it increases by 3.4. If you increase by 1 to 4, it increases by 4.3. It is not a linear increase. It is logistic regression. So, logistic increase. But you will see that as you go ahead as you keep on increasing your rating in value for money the overall rating goes up, and all of these things are significant.

Similarly, here, 3 got dropped, for location 3 got dropped, and 4 and 5. So, anything lower than 3 is not even accepted. There is no review in my dataset where the location is lower than 3. Now, for those were the 3. If you go from 3 to 5, it will increase. But you see that jump is not so much 0.73, 0.9 in comparison to 5, 4, etcetera. So, then this jump is not so much. So, I can say that location has less impact.

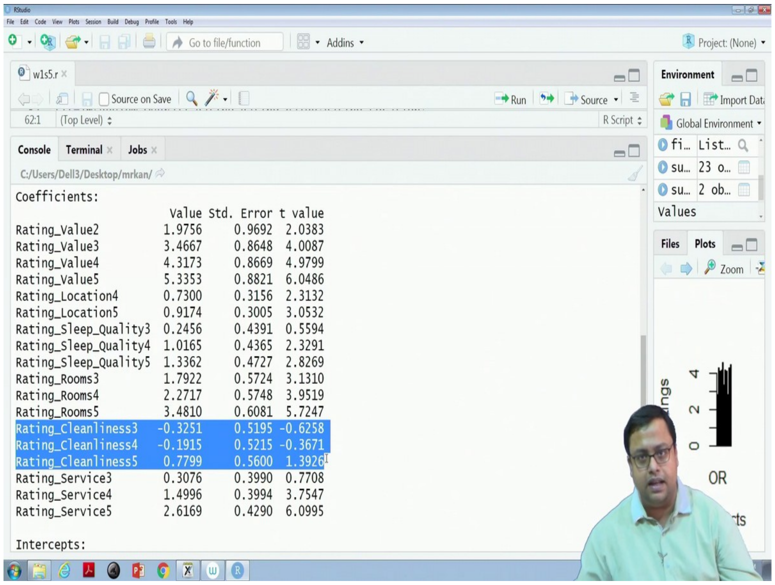
(Refer Slide Time: 33:28)



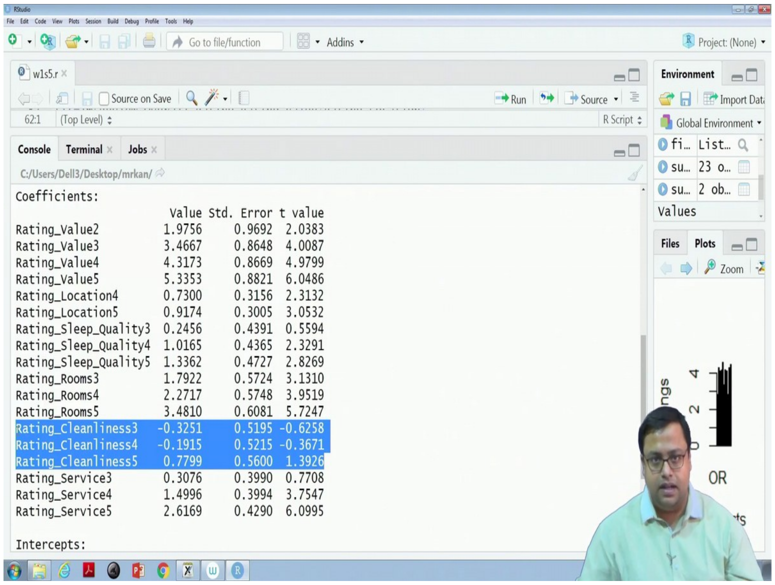
```
Coefficients:
                         Value Std. Error t value
Rating_Value2           1.9756     0.9692  2.0383
Rating_Value3           3.4667     0.8648  4.0087
Rating_Value4           4.3173     0.8669  4.9799
Rating_Value5           5.3353     0.8821  6.0486
Rating_Location4        0.7300     0.3156  2.3132
Rating_Location5        0.9174     0.3005  3.0532
Rating_Sleep_Quality3   0.2456     0.4391  0.5594
Rating_Sleep_Quality4   1.0165     0.4365  2.3291
Rating_Sleep_Quality5   1.3362     0.4727  2.8269
Rating_Rooms3           1.7922     0.5724  3.1310
Rating_Rooms4           2.2717     0.5748  3.9519
Rating_Rooms5           3.4810     0.6081  5.7247
Rating_Cleanliness3    -0.3251     0.5195 -0.6258
Rating_Cleanliness4    -0.1915     0.5215 -0.3671
Rating_Cleanliness5     0.7799     0.5600  1.3926
Rating_Service3         0.3076     0.3990  0.7708
Rating_Service4         1.4996     0.3994  3.7547
Rating_Service5         2.6169     0.4290  6.0995

Intercepts:
```



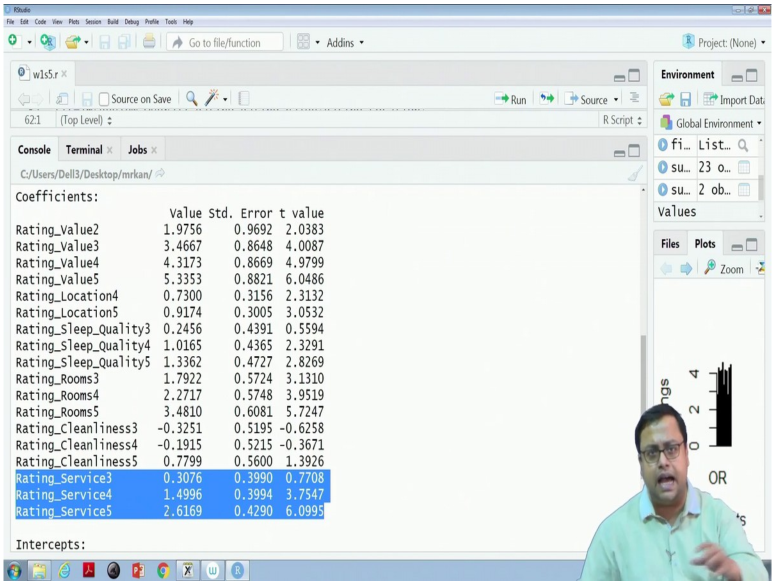
```
Coefficients:
                        Value Std. Error t value
Rating_Value2          1.9756     0.9692  2.0383
Rating_Value3          3.4667     0.8648  4.0087
Rating_Value4          4.3173     0.8669  4.9799
Rating_Value5          5.3353     0.8821  6.0486
Rating_Location4       0.7300     0.3156  2.3132
Rating_Location5       0.9174     0.3005  3.0532
Rating_Sleep_Quality3  0.2456     0.4391  0.5594
Rating_Sleep_Quality4  1.0165     0.4365  2.3291
Rating_Sleep_Quality5  1.3362     0.4727  2.8269
Rating_Rooms3          1.7922     0.5724  3.1310
Rating_Rooms4          2.2717     0.5748  3.9519
Rating_Rooms5          3.4810     0.6081  5.7247
Rating_Cleanliness3   -0.3251     0.5195 -0.6258
Rating_Cleanliness4   -0.1915     0.5215 -0.3671
Rating_Cleanliness5    0.7799     0.5600  1.3926
Rating_Service3        0.3076     0.3990  0.7708
Rating_Service4        1.4996     0.3994  3.7547
Rating_Service5        2.6169     0.4290  6.0995

Intercepts:
```



```
Coefficients:
                        Value Std. Error t value
Rating_Value2          1.9756     0.9692  2.0383
Rating_Value3          3.4667     0.8648  4.0087
Rating_Value4          4.3173     0.8669  4.9799
Rating_Value5          5.3353     0.8821  6.0486
Rating_Location4       0.7300     0.3156  2.3132
Rating_Location5       0.9174     0.3005  3.0532
Rating_Sleep_Quality3  0.2456     0.4391  0.5594
Rating_Sleep_Quality4  1.0165     0.4365  2.3291
Rating_Sleep_Quality5  1.3362     0.4727  2.8269
Rating_Rooms3          1.7922     0.5724  3.1310
Rating_Rooms4          2.2717     0.5748  3.9519
Rating_Rooms5          3.4810     0.6081  5.7247
Rating_Cleanliness3   -0.3251     0.5195 -0.6258
Rating_Cleanliness4   -0.1915     0.5215 -0.3671
Rating_Cleanliness5    0.7799     0.5600  1.3926
Rating_Service3        0.3076     0.3990  0.7708
Rating_Service4        1.4996     0.3994  3.7547
Rating_Service5        2.6169     0.4290  6.0995

Intercepts:
```



```
Coefficients:
                      Value Std. Error t value
Rating_Value2         1.9756     0.9692  2.0383
Rating_Value3         3.4667     0.8648  4.0087
Rating_Value4         4.3173     0.8669  4.9799
Rating_Value5         5.3353     0.8821  6.0486
Rating_Location4      0.7300     0.3156  2.3132
Rating_Location5      0.9174     0.3005  3.0532
Rating_Sleep_Quality3 0.2456     0.4391  0.5594
Rating_Sleep_Quality4 1.0165     0.4365  2.3291
Rating_Sleep_Quality5 1.3362     0.4727  2.8269
Rating_Rooms3         1.7922     0.5724  3.1310
Rating_Rooms4         2.2717     0.5748  3.9519
Rating_Rooms5         3.4810     0.6081  5.7247
Rating_Cleanliness3  -0.3251     0.5195 -0.6258
Rating_Cleanliness4  -0.1915     0.5215 -0.3671
Rating_Cleanliness5   0.7799     0.5600  1.3926
Rating_Service3       0.3076     0.3990  0.7708
Rating_Service4       1.4996     0.3994  3.7547
Rating_Service5       2.6169     0.4290  6.0995

Intercepts:
```



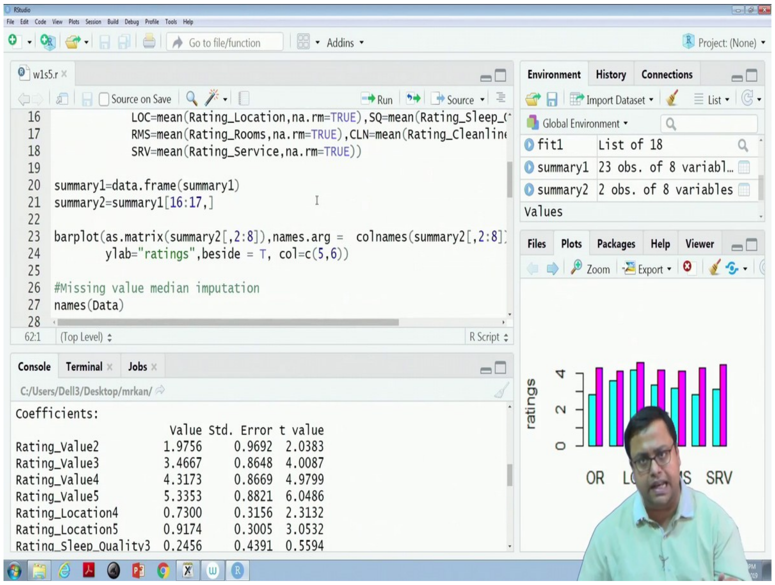
```
Coefficients:
                      Value Std. Error t value
Rating_Value2         1.9756     0.9692  2.0383
Rating_Value3         3.4667     0.8648  4.0087
Rating_Value4         4.3173     0.8669  4.9799
Rating_Value5         5.3353     0.8821  6.0486
Rating_Location4      0.7300     0.3156  2.3132
Rating_Location5      0.9174     0.3005  3.0532
Rating_Sleep_Quality3 0.2456     0.4391  0.5594
Rating_Sleep_Quality4 1.0165     0.4365  2.3291
Rating_Sleep_Quality5 1.3362     0.4727  2.8269
Rating_Rooms3         1.7922     0.5724  3.1310
Rating_Rooms4         2.2717     0.5748  3.9519
Rating_Rooms5         3.4810     0.6081  5.7247
Rating_Cleanliness3  -0.3251     0.5195 -0.6258
Rating_Cleanliness4  -0.1915     0.5215 -0.3671
Rating_Cleanliness5   0.7799     0.5600  1.3926
Rating_Service3       0.3076     0.3990  0.7708
Rating_Service4       1.4996     0.3994  3.7547
Rating_Service5       2.6169     0.4290  6.0995

Intercepts:
```

Similarly, sleep quality, you will see sleep quality has a lesser impact. So, first of all, from 2 to 3 it is insignificant, t value is lesser than even 1. So it is insignificant. But sleep quality in comparison 2 when you go to 4 and 5 you get, you make a jump from 2 to 4 or 2 to 5 it increases, but the increase is still at the range of 1, not 9, not 5, not 4 in the range of 1. Then if I go ahead, this is probably a little bit more important because the jump is at the range of 3.4.

Here, something is very interesting, if you increase cleanliness from 2 to 3 and 2 to 4. It is showing that your overall rating will drop. Actually, your overall rating does not drop because your t values are very low that so this results the first two results are not meaningful. These two results are not meaningful, but only when you jump, make a jump from 2 to 5. So, 2 to 3 does not happen, 2 to 4 does not mean anything. 2 to 5 there will be some change, which is positive still it is not significant.

So cleanliness has no meaning as per this particular result. It has no meaning. You do not even focus on cleanliness; whatever happens in cleanliness you do not care. On the other hand, services, it has some meaning. It can be seen when you jump from 2 to 5; the increase is 2.6. So, this guy gives you an idea that if I jump from 2 to 4 or 2 to 3 or 2 to 5 in terms of some improvement in the aspect ratings, how that will impact the overall rating.

(Refer Slide Time: 35:16)



```
            LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sleep_(
            RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Cleanline
            SRV=mean(Rating_Service,na.rm=TRUE))

summary1=data.frame(summary1)
summary2=summary1[16:17,]

barplot(as.matrix(summary2[,2:8]),names.arg =  colnames(summary2[,2:8]
        ylab="ratings",beside = T, col=c(5,6))

#Missing value median imputation
names(Data)
```

```
Coefficients:
                      Value Std. Error t value
Rating_Value2        1.9756     0.9692  2.0383
Rating_Value3        3.4667     0.8648  4.0087
Rating_Value4        4.3173     0.8669  4.9799
Rating_Value5        5.3353     0.8821  6.0486
Rating_Location4     0.7300     0.3156  2.3132
Rating_Location5     0.9174     0.3005  3.0532
Rating_Sleep_Quality3 0.2456    0.4391  0.5594
```

fit1	List of 18
summary1	23 obs. of 8 variabl…
summary2	2 obs. of 8 variables

So, in this particular video, what we did thoroughly is we have read a data set. We have grouped a data set and get certain insights. We have created a bar plot next. Then we have done a regression, what are the steps we find first, remove the missing values.

We then found out that whether there is any outlier or not, we remove the outliers. And then we actually found out that whether normality test are done for the x variables and the y variables, and we did not find them normal, but still, we actually tried for linear regression. We got a good result, good insight. And we also did ordinal logit, and we have got some bit of idea now.

Till now, I just wanted you to show the coding, not linear regression, not ordered logit. We will actually focus on each of them specifically in a later context. Here, I wanted you to actually get a little bit handy with the coding and that was our purpose. This is all that was there in week 1, week 2 onwards we will actually come into the marketing problems and will use various tools.

Again, I would suggest at this point in time, if you have studied statistics for management, the basic statistics, you should go and revise that. That is number 1. If you have studied marketing management, basic marketing management, you should go and revise that. And if you have studied, let us say, introduction to business analytics, some of the concepts of regression, if a little bit of machine learning techniques.

If you have ever heard about them or if you have ever studied about them, it is a good time to revise that because in our classes we will actually combine all of this stuff in solving marketing problems. And it is a prerequisite. It is an expectation that you might have a little bit of idea about marketing management, probably statistics, and interaction with business analytics.

So, this is the right time to go back and revise all of these things, from the next session, which will be week 2, we will actually focus on all of this stuff to solve marketing problems. Thank you. Thank you for being in this particular video. Thank you for being patient enough. And we will see you in the next video.