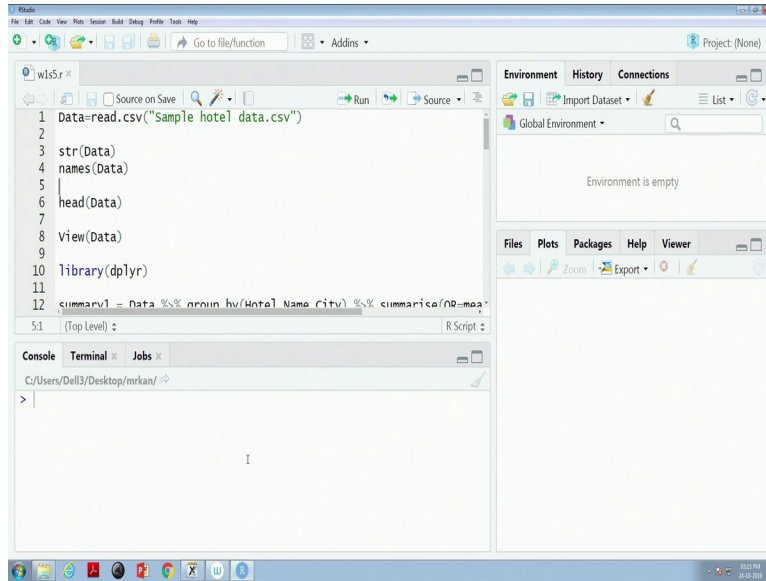


Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology Kharagpur
Lecture 05
Introduction to R Programming (Contd.)

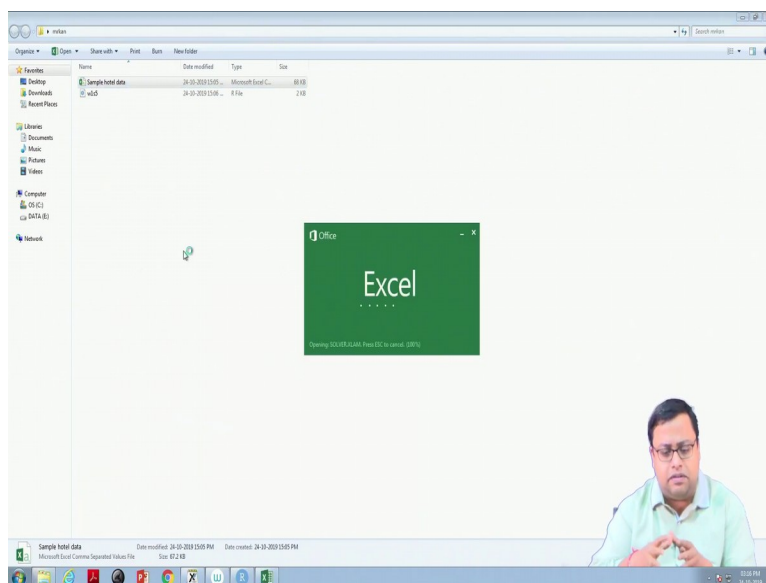
Hello, welcome to Marketing Analytics course. This is doctor from Swagato Chatterjee from VGSOM, IIT Kharagpur who is taking this particular course for you. And today this is week 1, session 5. So till now we have discussed about a little bit of R programming. We have started with vector and then we have also dealt with numeric and categorical vector, factor vectors. We have also talked about matrices and dataset. So till now we have not come into the marketing analytics per say.

We are still dealing with basic data analytics. So whenever you have some data, how to clean them, how to deal with them and etc. So today, I will actually bring in a little bit of flavor of marketing, but still today the major focus will be how to handle it. So, in your files you will see that there is a W1S5 dot R file. So, I have opened that particular file which is the R file and you will see that there is also a data file that has been given to you and that file is called sample hotel data dot csv.

(Refer Slide Time: 1:26)



So, we will today deal with these 2 files; one is this WIS5 that is week 1 session 5 dot R file. So, I would suggest again keep your console clean, keep your global environment clean, leave whatever is there in the fourth quadrant. And in your second quadrant where the editor is open, only W1 S5 dot 4 dot R should be open. Any other files should not be open at that particular part. So when we are at this particular stage where everything is clean, my console is clean, my global environment is clean, and the only this file is open, we can further work on it. So, let us start. (Refer Slide Time: 2:16)



Hotel_Name	Review_0 date_of_r	Month_of_Review_T	Rating_Va	Rating_Lo	Rating_Sle	Rating_Ro	Rating_Cle	Rating_Service
Bombay H	3 8/26/2014	Aug-14 with famil	3	4	NA	NA	NA	3
Bombay H	1 #####	Jun-13 NA	1	2	NA	1	1	3
Coffea Arc	2 10/20/201	Oct-14 with frien	2	NA	NA	2	NA	1
Coffea Arc	3 9/29/2014	Sep-14 with frien	NA	4	NA	NA	3	3
Coffea Arc	3 #####	Jul-14 as a coupli	NA	NA	3	2	NA	5
Coffea Arc	2 #####	Jul-14 with famil	2	3	NA	NA	NA	2
Coffea Arc	3 7/22/2014	Jul-14 as a coupli	NA	NA	NA	NA	NA	NA
Coffea Arc	4 #####	Jun-14 with famil	4	5	5	3	3	3
Coffea Arc	5 #####	May-14 as a coupli	5	5	5	4	4	5
Coffea Arc	4 4/30/2014	Apr-14 with famil	4	4	3	3	3	5
Coffea Arc	4 4/21/2014	Apr-14 as a coupli	4	5	4	4	5	
Coffea Arc	5 3/18/2014	Mar-14 as a coupli	5	5	5	5	5	
Coffea Arc	4 3/15/2014	Feb-14 with frien	5	3	3	4	4	
Coffea Arc	4 #####	Apr-13 with famil	5	5	3	3	3	
Coffea Arc	4 #####	Jan-14 with frien	4	4	3	3	3	

So, as you know that the file that we are working on with this sample hotel data dot csv. This is actually more related to the kind of research work I do, and I will try to show you that what this particular, this is a subset of the data that I have collected for one of my research work. So, this data looks like this, and I will let a little bit make it a little bit bigger so, that it can be seen.

So, now, if you can see this, you will see that there is a csv file. And there are lots of data rows that are available, and probably a few columns are also available. And before I jump into the data analytics of this particular thing, I would want to explain to you what this data has. So, this data has been collected from some websites, which is a review website where people actually post their online reviews.

So it is a very popular review website and from where we have collected, so there are around 40 randomly selected 40 hotels and for each of the hotels, we have selected some of their reviews. Probably almost all of their reviews till at a certain point of time, and it was collected long back around if I am not wrong around probably 5 years back.

So we have this data which was collected, and then lots of research paper has come up on this data. But I will focus on a simple problem here. So when you post a review, if you know that

even if you have ever posted a review or if you have ever seen a review that has been posted in an online, you will see that, that people actually post lots of stuff, they give a overall heading of the review and then the content of that review.

Now here we are not doing any kind of text mining, at least in this class. So I will not go into the table heading content or the or that actual text content of the review. But along with that, you also give from 1 star to 5 stars one kind some quantitative rating. And along with that, you also give along with the quantitative rating. Sometimes in some of the review websites, people actually ask for other attribute wise ratings as well. So by attribute wise rating I mean to say let us say value for money or location or quality of the food, quality of the rooms, and so on.

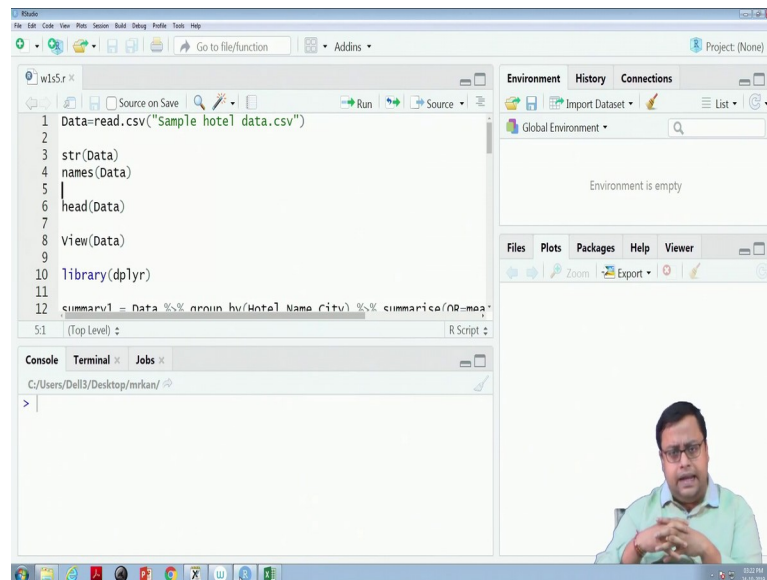
So, this data set also has those kinds of stuff. So the first column is the hotel name. The second column is the overall review, review the overall rating. So this talks about the overall review, the in a 1 to 5 point scale whatever you have given. Then the date in which the review has been posted, the month of visit in which month he has visited, the reviewer type whether he has visited as a family or let us say with their business colleagues or with his spouse or solo even solo alone.

And then there are 6 attribute wise rating that means the different aspects of the hotel you have given rating. So value, location, probably sleep quality and then if you go ahead rooms and then cleanliness and service. So, these are the 6 aspects western, which people have given this rating, and we have this data set.

So, again I will ask you at this point to please pause the video, and probably before you go ahead in the video, please pause the video. And think probably for 2-3 minutes that if this kind of a dataset comes up to me, what will I do with this data set? But I elect I am one of the hotel owners. So out of these 900 something reviews that are there, probably around 100 reviews are for me, for my hotel.

So then if that is the case, if let us say on around 50 reviews are for my hotels, and that is the reviews that I have got. Then what will I do with this data set? So you can think a little bit.

(Refer Slide Time: 6:15)



What could you do? Some of the possible options when I asked the same thing in my class, some of the earlier answers that I have got from the students is that one hotel can know that what is his overall performance. So what is overall performance? He can get a probably all the so if there are 50 reviews, and there is overall rating; he can get a mean value of that rating.

So that will give him an idea of how much is the average and mean, and standard deviation might also be there. So that will give you an idea that what is the average satisfaction score of the people regarding that particular hotel and they will also know how that varies a little bit. So then somebody might want to ask that okay. So there every time when we try to create a marketing story, there is a what, there is a why and there is a how. So the what part is given that okay, I am doing good or bad than the average.

So, I can get my hotel's mean value. I can get my competitor's hotels mean value as well, or even if, if I do not get only the competitors, I can get the overall industry average as well.

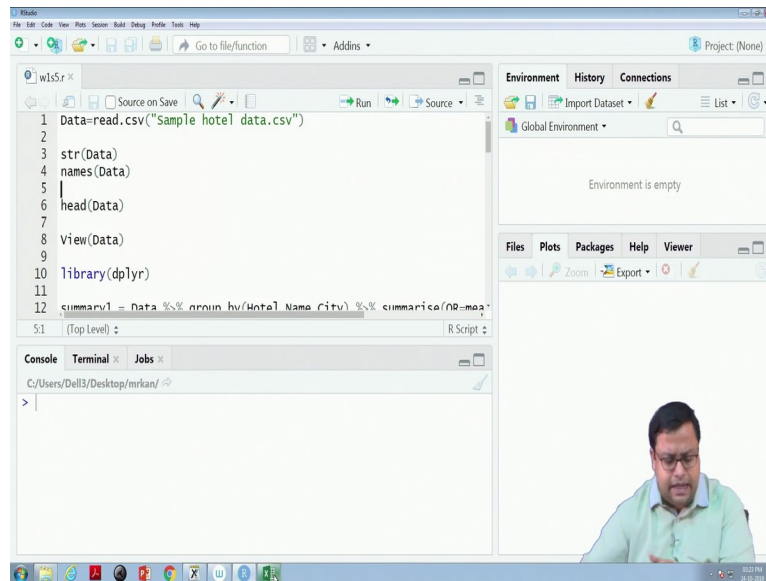
And then, I can compare my overall performance with the industry average that is one state.

But that is a primary step that you have to do.

Now, let us say once I find out that and I found out that okay, I am not doing very good, I am doing a little bit lower than the average. Then comes that, okay, so if I am doing a little bit lower than the average, why? The next question is, why? Why am I doing a little bit lower than the average? Obviously, the second question. So, then you might be actually finding out all possible options which can actually be leading to this thing. So, there is a vice course available. How? So, ideally, we also have 6 aspects if you have seen.

(Refer Slide Time: 8:19)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Hotel_Nar	Review_	O date_of_ r	Month_of Review_ T		Rating_Va	Rating_Lo	Rating_Sle	Rating_Ro	Rating_Cle	Rating_Ser	vice
2	Bombay H	3	8/26/2014	Aug-14	with famil	3	4	NA	NA	NA	3	
3	Bombay H	1	#####	Jun-13	NA	1	2	NA	1	1	3	
4	Coffea Arc	2	10/20/201	Oct-14	with frien	2	NA	NA	2	NA	1	
5	Coffea Arc	3	9/29/2014	Sep-14	with frien	NA	4	NA	NA	3	3	
6	Coffea Arc	3	#####	Jul-14	as a coupl	NA	NA	3	2	NA	5	
7	Coffea Arc	2	#####	Jul-14	with famil	2	3	NA	NA	NA	2	
8	Coffea Arc	3	7/22/2014	Jul-14	as a coupl	NA	NA	NA	NA	NA		
9	Coffea Arc	4	#####	Jun-14	with famil	4	5	5	3	3	3	
10	Coffea Arc	5	#####	May-14	as a coupl	5	5	5	4	4	5	
11	Coffea Arc	4	4/30/2014	Apr-14	with famil	4	4	3	3	3	5	
12	Coffea Arc	4	4/21/2014	Apr-14	as a coupl	4	5	4	4	5		
13	Coffea Arc	5	3/18/2014	Mar-14	as a coupl	5	5	5	5	5		
14	Coffea Arc	4	3/15/2014	Feb-14	with frien	5	3	3	4	4		
15	Coffea Arc	4	#####	Apr-13	with famil	5	5	3	3	3		
16	Coffea Arc	4	#####	Jan-14	with frien	4	4	3	3	3		

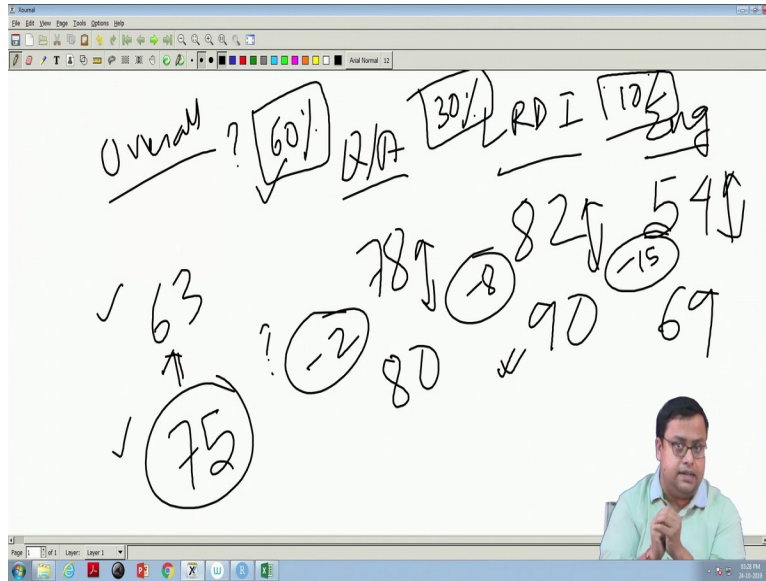


These are the 6 aspects based on which people also give reviews. So, either you are good in your 6 aspects or bad in those six aspects in comparison to your industry average or comparison to the competitors. So, then the question comes in that which of the aspects out of the 6 aspects, I am good and which of them I am bad? So, that might probably give you a little bit of hint about why you are performing badly. If you find out that there are 2 aspects of 3 aspects out of those 6 where you are much lower than the industry average, then they are the possible reasons.

Now, next, comes how? How means should I improve if I aspect 1, aspect 2, and aspect 3? If these 3 aspects are doing bad, which one will I improve, and how will I improve? So this is a very classic problem. So this comes under the overall problem of resource allocation, which will do into a later class, a little bit of service improvement, which also comes under the purview of operations and marketing together.

Whatever be the case, it is a1, a2, a3 for an analytics point of view. This is a1, a2, and a3, and I want to know which one I should focus more? Think about a situation.

(Refer Slide Time: 9:46)



Let us say that you are a student and you got around 63 as your overall score, 63 overall and this overall score came up from let say quantitative ability and LRDI and English these are the three things that contributes to this 63, the overall average is 75 and you did lower than the average, let say you have done that.

Now quantitative ability, you got 78, LRDI got ,let us say 82 but English you got 54 something like that and the average, the in the average of other people in this group is let us say here it is 80. So you are a little bit lower than that. Here it is, let say 90, so you are still lower than that. But here it is let us say around; I do not know let us say 69. So, if I ask you now that which one you should focus on?

You see that here it is minus 2, here it is minus 8, and here it is almost minus 15. So our common sense is that okay, I will focus on English because English is something that I am doing most badly in comparison to the average. But then some, if I tell you that okay, the score is, this score is calculated in a different way. These when I calculated this score 63 and 75, I did not give equal weightage to all these 3 aspects.

I gave, let us say 60 percent to this somehow I do not know. Let us say 30 percent to this and 10 percent to this, this might not add up to 63, but I am just saying that if the weightage is

different, if the weightage is not the same, then this distance alone does not tell me anything. I have also to consider the weightages. So in some way or other, I have to find out these weightages. How much is the weight for the aspects that I have in my hand?

The story does not end here, as well. The story has another line. So let us say you find out that now, based on this, it is 60 percent, but you are almost closer to Q and A quantitative ability. So, though it is 60 percent, you will not focus on quantitative ability because the distance is very low. You might probably focus on LRDI because the distance is minus 8, and that has 30 percent important so, that has pretty good importance.

So, you focus on LRDI; you decided that I will focus LRDI. But then you came to know that till now we have not taken the help of anybody in the exam preparations. And you came to know that okay you have a person in your family let us say, your younger, your elder brother or elder sister or your father or mom, who is a professor of English, who can actually help you in English very or on any verbal ability very-very easily. He can actually contribute it very easily, and you have not taken his help till now.

On the other end, for LRDI you have to go to a coaching center, you have to actually pay a lot of money to improve your score. In other words, I am trying to say that let us say improving English is easy and improving LRDI is a little bit tough. If this information comes in your hand now, then again, the situation changes. You cannot only depend on this; you cannot only depend on this, then another variable comes into the picture, which is the cost. So, that is how I am trying to create the story. There are lots of aspects of the story.

Now coming back to our problem, let us say I want to find a hotel, and I want to find out that out of these 3 aspects, which one should I focus on? The first thing is to know which one I am doing worse, whether I am doing worse here or here or here, where the distance is higher? That is the first step; the second step is that I have to find out which of this distance is more

important than the other one? Where will I focus more, where the customers who are giving a review, if they are saying that I have got very bad reviews 3 or 2 out of 5, why it is that?

Is it, they are giving more focus to sleep quality, or they are giving more focus to let us say rooms or the location or the food which one? I have to find out that, how to find out that?

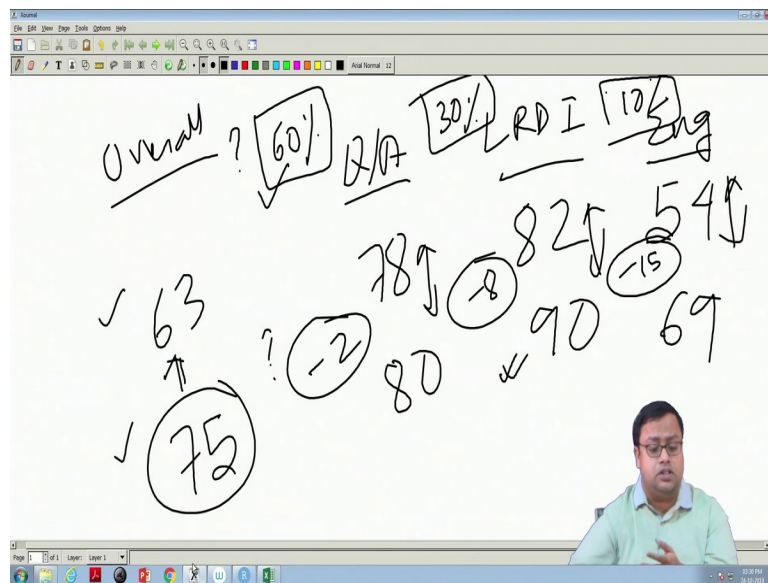
(Refer Slide Time: 14:27)

Hotel_Nar	Review_O date_of_r	Month_of_Review_Ty	Rating_Va	Rating_Lo	Rating_Sle	Rating_Ro	Rating_Cle	Rating_Service
Bombay H	3 8/26/2014	Aug-14 with famil	3	4	NA	NA	NA	3
Bombay H	1 #####	Jun-13 NA	1	2	NA	1	1	3
Coffea Arc	2 10/20/201	Oct-14 with frien	2	NA	NA	2	NA	1
Coffea Arc	3 9/29/2014	Sep-14 with frien	NA	4	NA	NA	3	3
Coffea Arc	3 #####	Jul-14 as a coupl	NA	3	2	NA	5	
Coffea Arc	2 #####	Jul-14 with famil	2	3	NA	NA	NA	2
Coffea Arc	3 7/22/2014	Jul-14 as a coupl	NA	NA	NA	NA	NA	
Coffea Arc	4 #####	Jun-14 with famil	4	5	5	3	3	3
Coffea Arc	5 #####	May-14 as a coupl	5	5	5	4	4	5
Coffea Arc	4 4/30/2014	Apr-14 with famil	4	4	3	3	3	5
Coffea Arc	4 4/21/2014	Apr-14 as a coupl	4	5	4	4	5	
Coffea Arc	5 3/18/2014	Mar-14 as a coupl	5	5	5	5	5	
Coffea Arc	4 3/15/2014	Feb-14 with frien	5	3	3	4	4	
Coffea Arc	4 #####	Apr-13 with famil	5	5	3	3	3	
Coffea Arc	4 #####	Jan-14 with frien	4	4	3	3		

If you think a little bit, you will understand that my B column is my y variable, and column F to column K is my x variables, and I can do a regression to find out that how these 6 aspects impact my Y variable. So, that is what we are at the end of the day going to do step by step with this data. So, in the first step, what we will do is we will collect this particular data read, this particular data.

And then we will try to find out what the various things that I can do with this data are? How can I find out the basic details about the data? And then after that, I will try to see how people are giving different relative importance to different aspects, aspect 1, aspect 2, aspect 3, and so on, how they are giving different importance to different aspects. Then, I might also want to know that, let us say some people let say in the same example that I was talking about.

(Refer Slide Time: 15:38)



In this particular example, were Q and A, LRDI, and engineer English is there. There will be some people who are from high con background. Let us assume that they have studied in engineering and probably the science subjects and etc. So they are a little bit more quantitative oriented, and there will be some people who will be not so much quantitative oriented.

So, the guys who will be quantitative oriented will might give more weightage to Q and A. And the guys who are not so much quantitative oriented might give more weightage to English. So if there are different people in the selection body who decides that how much weightage to be given to Q and A, LRDI, and verbal ability then the weightages given are also different for different kinds of people. So, that is something which also is an interesting story.

(Refer Slide Time: 16:34)

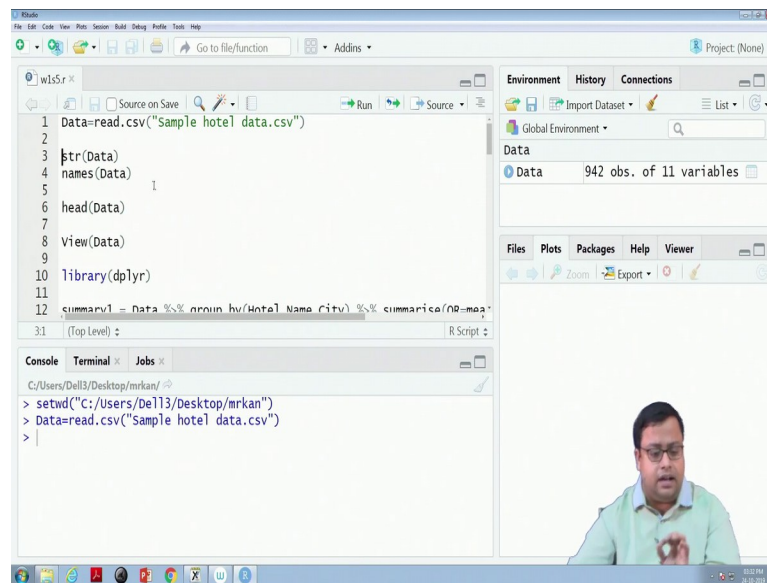
Hotel_Nar	Review_O date_of_r	Month_of Review_T	Rating_Va	Rating_Lo	Rating_Sle	Rating_Ro	Rating_Cle	Rating_Service
Bombay H	3 8/26/2014	Aug-14 with famil	3	4	NA	NA	NA	3
Bombay H	1 #####	Jun-13 NA	1	2	NA		1	3
Coffea Arc	2 10/20/201	Oct-14 with frien	2	NA	NA		2	NA
Coffea Arc	3 9/29/2014	Sep-14 with frien	NA	4	NA	NA	3	3
Coffea Arc	3 #####	Jul-14 as a coupl	NA	NA	3	2	NA	5
Coffea Arc	2 #####	Jul-14 with famil	2	3	NA	NA	NA	2
Coffea Arc	3 7/22/2014	Jul-14 as a coupl	NA	NA	NA	NA	NA	NA
Coffea Arc	4 #####	Jun-14 with famil	4	5	5	3	3	3
Coffea Arc	5 #####	May-14 as a coupl	5	5	5	4	4	5
Coffea Arc	4 4/30/2014	Apr-14 with famil	4	4	3	3	3	5
Coffea Arc	4 4/21/2014	Apr-14 as a coupl	4	5	4	4	5	5
Coffea Arc	5 3/18/2014	Mar-14 as a coupl	5	5	5	5	5	5
Coffea Arc	4 3/15/2014	Feb-14 with frien	5	3	3	4	4	5
Coffea Arc	4 #####	Apr-13 with famil	5	5	3	3		
Coffea Arc	4 #####	Jan-14 with frien	4	4	3	3		

So, for example, here, if I know that out of the 6 aspects, I know that okay service is more important. But if I, by chance, find out that for a customer who is traveling alone, service is not that important. Because he generally does not take all those services. He just probably goes into the restaurant, eats, goes to the bar, have some drinks, and then go back to his room that is all.

On the other hand, a person who is traveling with his friends or family will take lots of services. He will go, he will eat in the morning, probably go in lunch because his kids want to have a little bit of fun. His family members want to have a little bit of fun. So, all the various aspects of the hotel like the swimming pool and the restaurant and the free morning breakfast etc-etc everything he will take and that is why the service requirement is much more for him.

If that is something that we can find out from this kind of data, then that is also a new insight, and that is also something that will help in for a marketer, so given that as the background, I will start this particular coding.

(Refer Slide Time: 17:49)



```
1 Data=read.csv("Sample hotel data.csv")
2
3 str(Data)
4 names(Data)
5
6 head(Data)
7
8 View(Data)
9
10 library(dplyr)
11
12 summary1 = Data %>% group_by(Hotel Name, City) %>% summarise(n0=mea
3:1 (Top Level) R Script
```

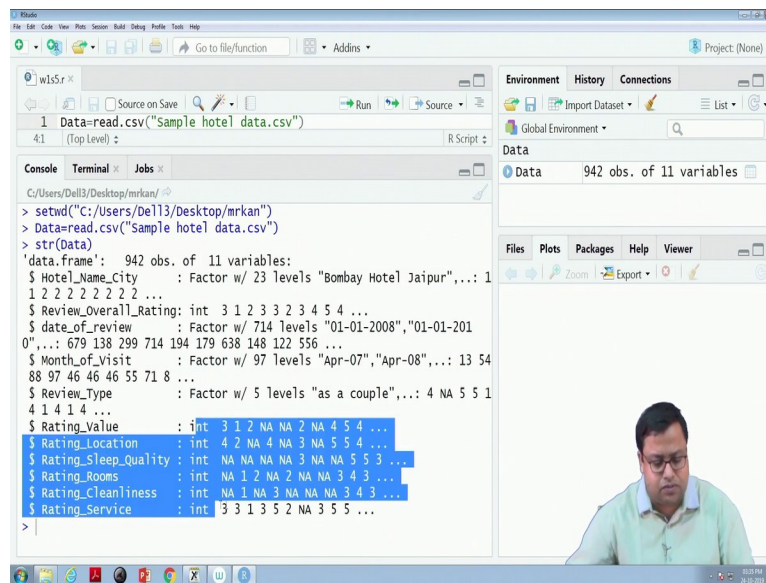
Environment History Connections
Global Environment
Data
Data 942 obs. of 11 variables

Console Terminal Jobs
C:/Users/Dell3/Desktop/mrkan/
> setwd("C:/Users/Dell3/Desktop/mrkan")
> Data=read.csv("Sample hotel data.csv")
>

So, the first thing I will do is I will read a dataset; I will set our working directory. So I have kept this W1 S5 dot r and my data set in the same folder, so I can just go to the session, working directory to the source file location, that sets my working directory. Then line number 1, the data set they mean sample hotel data dot csv. In the last video, I have shown you how to read and write a data set. So I am reading a data set.

And I am saving the data set in the name of Data, so that is the dataset. So, once you read the data set, the best thing to do is to see the structure of the data set. So how do you see the structure of the data set? The easy form function is str and then within bracket data.

(Refer Slide Time: 18:48)



```
1 Data=read.csv("Sample hotel data.csv")
4:1 (Top Level) : R Script :

Console Terminal x Jobs x
C:/Users/Dell3/Desktop/mrkan/ >
> setwd("C:/Users/Dell3/Desktop/mrkan")
> Data=read.csv("Sample hotel data.csv")
> str(Data)
'data.frame':  942 obs. of  11 variables:
 $ Hotel_Name_City   : Factor w/ 23 levels "Bombay Hotel Jaipur",...: 1
 1 2 2 2 2 2 2 2 ...
 $ Review_Overall_Rating: int  3 1 2 3 3 2 3 4 5 4 ...
 $ date_of_review     : Factor w/ 714 levels "01-01-2008", "01-01-201
0",...: 679 138 299 714 194 179 638 148 122 556 ...
 $ Month_of_Visit    : Factor w/ 97 levels "Apr-07", "Apr-08",...: 13 54
88 97 46 46 46 55 71 8 ...
 $ Review_Type       : Factor w/ 5 levels "as a couple",...: 4 NA 5 5 1
4 1 4 1 4 ...
 $ Rating_Value      : int  3 1 2 NA NA 2 NA 4 5 4 ...
 $ Rating_Location   : int  4 2 NA 4 NA 3 NA 5 5 4 ...
 $ Rating_Sleep_Quality: int  NA NA NA NA 3 NA NA 5 5 3 ...
 $ Rating_Rooms      : int  NA 1 2 NA 2 NA NA 3 4 3 ...
 $ Rating_Cleanliness: int  NA 1 NA 3 NA NA NA 3 4 3 ...
 $ Rating_Service    : int  3 3 1 3 5 2 NA 3 5 5 ...
```

So the moment I see that this is what comes up, the structure of the data set, the lots of things have come up, and I will try to explain one by one what these are. So in the first line, it is saying that it is a data frame. So, whatever he told, this guy is a data frame, which has 942 observations of 11 variables. So, this guy has 942 observations that mean 942 rows and 11 variables.

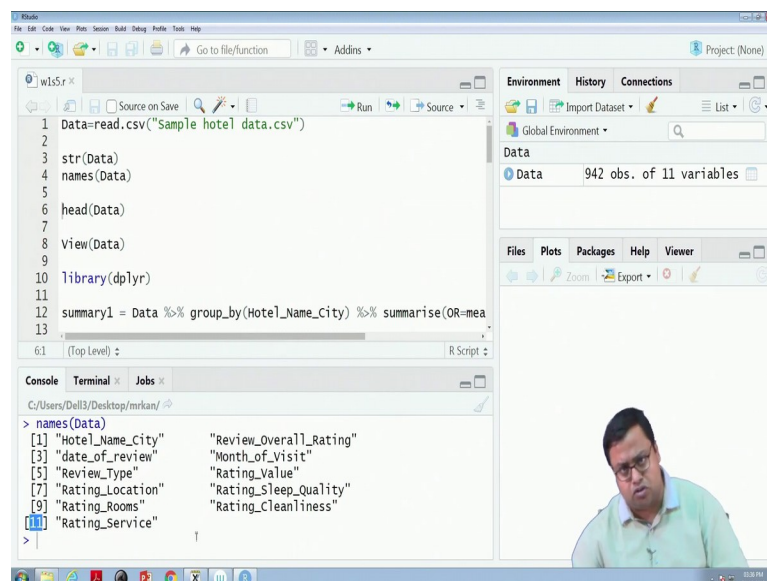
Now, it is also giving me listing down the variable names. So, the first one is the hotel name city, the second one is probably, review overall rating, and the third one is the date of the review. The fourth one is the month of visit and so on. And it is also telling me, what are the various types of these guys. So the first one hotel names it is a factor variable with 23 levels, 23 levels mean there are actually 23 unique hotel names are there.

Then comes the review overall rating that is in a 1 to 5 point scale, you are giving overall rating there, so that is an integer variable. The date of the review here it has been taken as a factor variable. So if I want to do anything with this, I have to change it to its date form. And in a later class, if it is required, we will show you how to change a factor variable to its date form; if it is recorded, it is a date actually.

So, there is a function called `as.date`, if you want to know now, you can go and search and find out how that works `as.date`. So it cannot change a factor variable directly; you have to change the factor variable to its character form, and from the character form, you can change it to date. So, you have to first change it to `as` dot character, and then you change it to character form and then `as` dot date, so I would suggest you to just Google it up. And then and in a different class, if it is required, we will also show you then review type.

Review type is like there are 5 levels whether you are a visited as a couple or visited as a businessman or whatever. And then these are the 6 aspects which are also there. They are all integers..

(Refer Slide Time: 21:01)



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

```
1 Data=read.csv("Sample hotel data.csv")
2
3 str(Data)
4 names(Data)
5
6 head(Data)
7
8 View(Data)
9
10 library(dplyr)
11
12 summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mea
13
61 (Top Level) R Script
```

The Environment pane shows a variable named 'Data' with 942 observations and 11 variables. The Console pane shows the output of the `names(Data)` command:

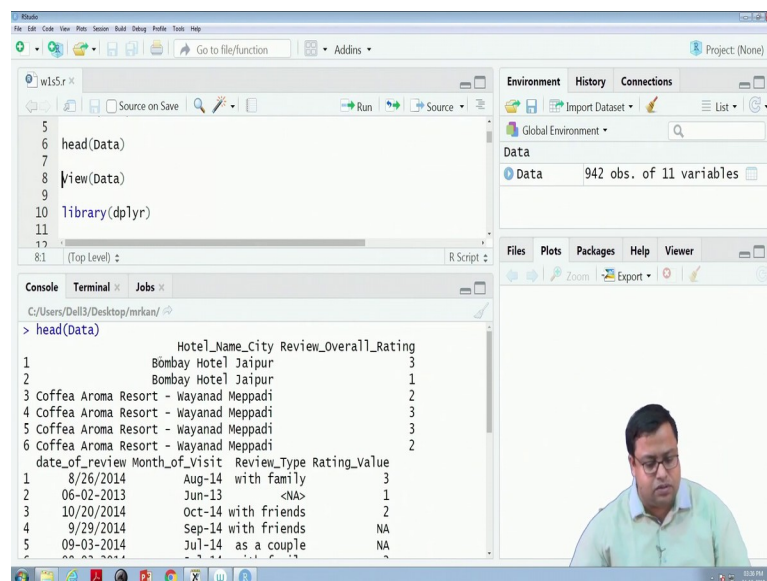
```
> names(Data)
[1] "Hotel_Name_City"      "Review_Overall_Rating"
[3] "date_of_review"      "Month_of_Visit"
[5] "Review_Type"         "Rating_Value"
[7] "Rating_Location"     "Rating_Sleep_Quality"
[9] "Rating_Rooms"        "Rating_Cleanliness"
[11] "Rating_Service"
```

So next, I have cleaned the console. Now, I also want to know the names of the column names of the data set. So if I just run this slide, it gives me the column names; these particular names within bracket data will come handy when you write the codes later. You see that there will be lots of codes that I will show you in the late, at a later point in time. You will see that the codes are probably 2-3 sentences long.

Now, I have not written those codes in one single line, one single go, when I was writing the code for this particular class, I was doing iterations, I was doing mistakes and whatever code you see is the final no mistake version of this particular thing. But ideally, it is an iterative process, and you slowly learn how to code, and you make mistakes, and when you make mistakes, these small-small things like names of data, an str of data. These small things come in handy to find out what this particular mistake is.

For example, these names within bracket data, which gives you a column names will come in handy to know that which column number is associated with which column name, for example, rating service is 11th, this information will be needed when I go ahead at a later point of time, and I will show you..

(Refer Slide Time: 22:21)



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

```
5  
6 head(Data)  
7  
8 view(Data)  
9  
10 library(dplyr)  
11  
12
```

The console shows the output of the `head(Data)` command:

```
> head(Data)  
  Hotel_Name_City Review_Overall_Rating  
1      Bombay Hotel Jaipur              3  
2      Bombay Hotel Jaipur              1  
3 Coffea Aroma Resort - Wayanad Meppadi  2  
4 Coffea Aroma Resort - Wayanad Meppadi  3  
5 Coffea Aroma Resort - Wayanad Meppadi  3  
6 Coffea Aroma Resort - Wayanad Meppadi  2  
date_of_review Month_of_Visit Review_Type Rating_Value  
1      8/26/2014      Aug-14 with family              3  
2      06-02-2013      Jun-13      <NA>              1  
3      10/20/2014      Oct-14 with friends              2  
4      9/29/2014       Sep-14 with friends              NA  
5      09-03-2014      Jul-14 as a couple              NA
```

The Environment pane on the right shows a data object named 'Data' with 942 observations and 11 variables.

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

```

5
6 head(Data)
7
8 View(Data)
9
10 library(dplyr)
11
12
13
14
15
16
17
18:1 (Top Level)

```

The console output for `head(Data)` is as follows:

	Hotel_Name	City	Review	Overall_Rating
1	Bombay Hotel	Jaipur		
2	Bombay Hotel	Jaipur		
3	Coffea Aroma Resort - Wayanad	Meppadi		
4	Coffea Aroma Resort - Wayanad	Meppadi		
5	Coffea Aroma Resort - Wayanad	Meppadi		
6	Coffea Aroma Resort - Wayanad	Meppadi		

The environment pane on the right shows 'Data' with 942 observations and 11 variables.

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

```

5
6 head(Data)
7
8 View(Data)
9
10 library(dplyr)
11
12
13
14
15
16
17
18:1 (Top Level)

```

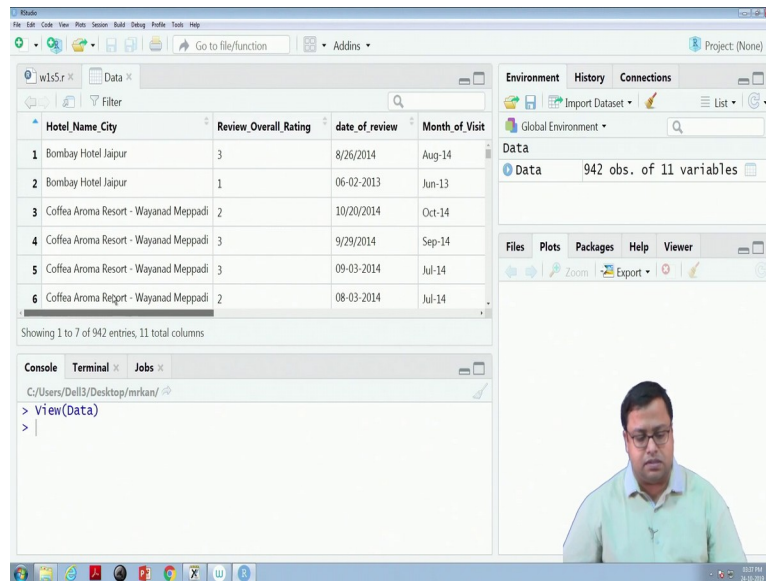
The console output for `View(Data)` is as follows:

	Rating_Cleanliness	Rating_Service
1	NA	3
2	1	3
3	NA	1
4	3	3
5	NA	5
6	NA	2

The environment pane on the right shows 'Data' with 942 observations and 11 variables.

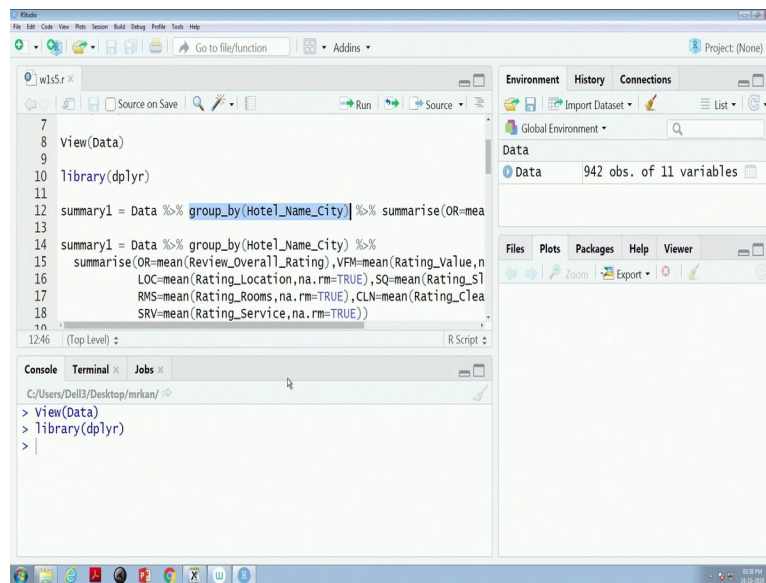
Then the head of data, so the head of data will actually print the first 6 observations, a top 6 observations. So this is how it is looked like the first is hotel name, city and like Bombay Hotel Jaipur and then Coffee Aroma Resort Wayanad. And this is the first 6 overall rating. This is the first 6 dates, first 6 months of visit, and so on. So oftentimes, we also print it to see the data set. We can also see the data set by running the line number 8 view dot data or I can click on here, click here.

(Refer Slide Time: 23:02)



Both will do the same job, which will give me a spreadsheet like view of the data set. So, this is how the data set looks like if you do not want to open in csv file, in your excel. If there are lots of rows, it is not suggested that you open in excel. In that case, it will open here using few data set.

(Refer Slide Time: 23:23)

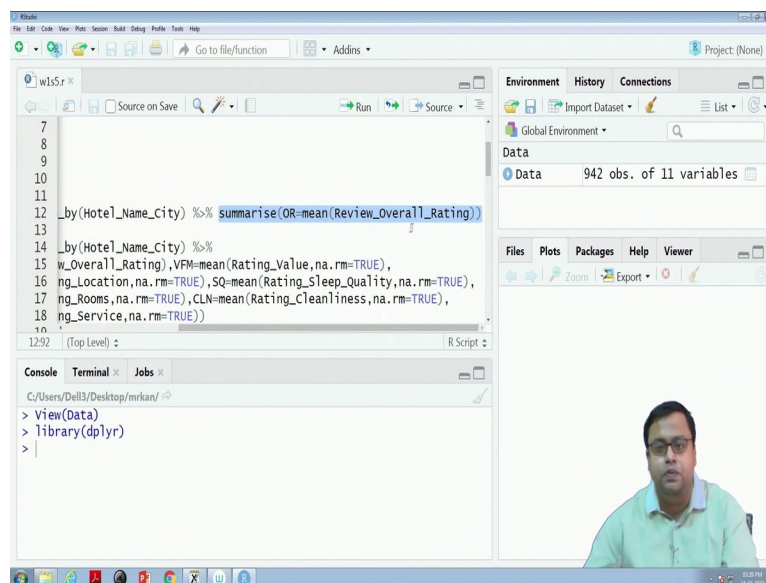


```
7
8 View(Data)
9
10 library(dplyr)
11 summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mea
12
13
14 summary1 = Data %>% group_by(Hotel_Name_City) %>%
15 summarise(OR=mean(Review_Overall_Rating),VFM=mean(Rating_Value,n
16 LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
17 RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_clea
18 SRV=mean(Rating_Service,na.rm=TRUE))
19
20 (Top Level) R Script
```

Environment History Connections
Global Environment
Data
Data 942 obs. of 11 variables

Files Plots Packages Help Viewer
Zoom Export

Console Terminal Jobs
C:/Users/Dell3/Desktop/mrkan/
> View(Data)
> library(dplyr)
> |




```
7
8
9
10
11
12 _by(Hotel_Name_City) %>% summarise(OR=mean(Review_Overall_Rating)
13
14 _by(Hotel_Name_City) %>%
15 w_Overall_Rating),VFM=mean(Rating_Value,na.rm=TRUE),
16 ng_Location,na.rm=TRUE),SQ=mean(Rating_Sleep_Quality,na.rm=TRUE),
17 ng_Rooms,na.rm=TRUE),CLN=mean(Rating_Cleanliness,na.rm=TRUE),
18 ng_Service,na.rm=TRUE))
19
20 (Top Level) R Script
```

Environment History Connections
Global Environment
Data
Data 942 obs. of 11 variables

Files Plots Packages Help Viewer
Zoom Export

Console Terminal Jobs
C:/Users/Dell3/Desktop/mrkan/
> View(Data)
> library(dplyr)
> |

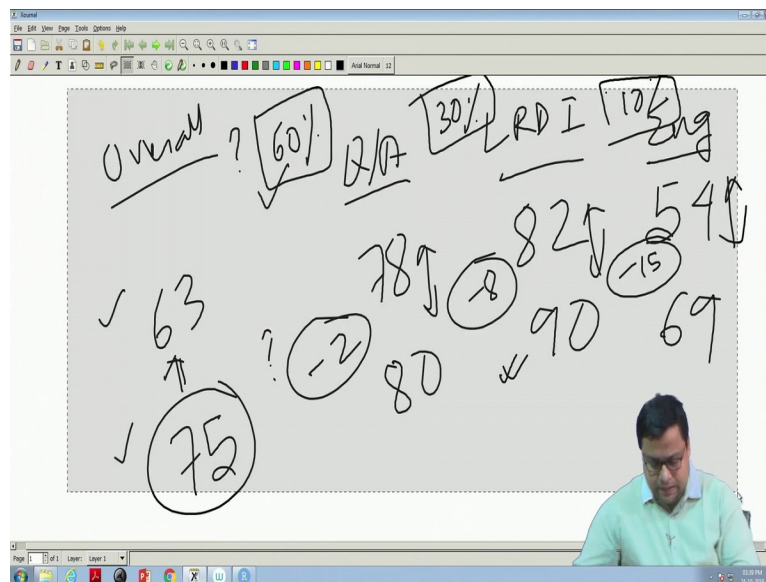


Now, the question that comes up first is whether I am doing good enough or not? That is the first question. So, to do that, what will I do? I will find out the average of my hotel and average of all other hotels, fair enough, an average of my hotel and average of all other hotels probably my competitor hotels, here the hotels were haphazardly corrected. So, I do not know who his competitor of whom.

So, I will randomly select around 40 hotels, so all the hotels. So, we have done this in the last class on how to summarize data. So, I see I am calling a library called dplyr. We have to install it if it is not there. And once you have actually installed it and you call this library dplyr, you have to go to summary one. Summary one is the name that I have given, summary one is equal to data; this is the original data set. And then this piping sign and then group by hotel name city.

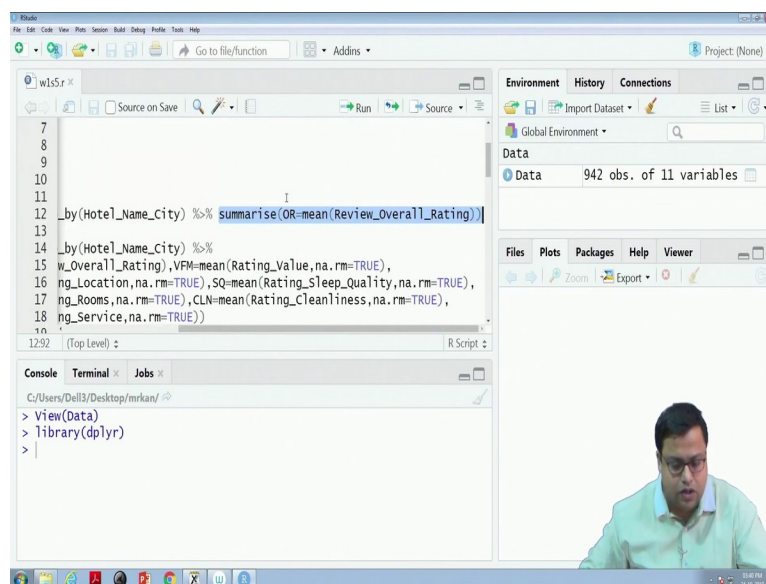
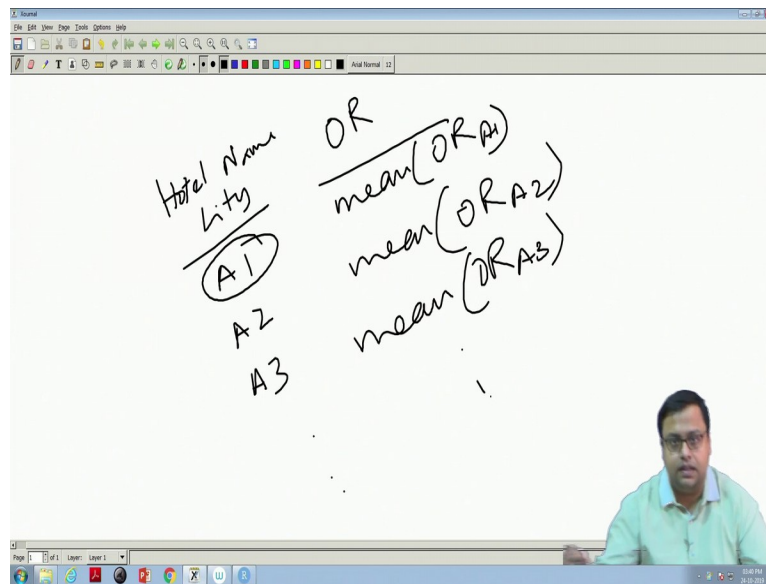
So, the hotel name city is the first column of your data-data set. So, you are breaking the whole data set based on the hotel name city. So, based on hotel name, you are actually breaking the whole data set. And then, in the next step, you summarize the data. You summarize the data in such a way that word is equal to, summarize means a new data set will be created. The previous data set will not be appended with anything. A new data set will be created and where OR is equal to mean of review overall rating, what will it give?

(Refer Slide Time: 25:16)



So what will it give? It will give something like this. So let me just clean it up. So what will it give?

(Refer Slide Time: 25:28)



It will give something like this; it will first create a hotel Name City. And A 1, A 2, A 3 are the hotel names let us say. And then it will create an overall rating, so a different this thing, which is OR is equal to mean of overall rating. So OR and if the A 1 hotel has 10 reviews, so those 10 reviews mean will come here. So mean of the overall rating of A 1s and then mean of the overall rating of A 2 and mean of the overall rating of A 3 and so on. That will come up here. So, that is something that I am expecting.

(Refer Slide Time: 26:18)

The screenshot shows the RStudio interface. The main window displays a data table with the following columns: **Hotel_Name_City** and **OR**. The data is as follows:

Hotel_Name_City	OR
1 Bombay Hotel Jaipur	2.000000
2 Coffea Aroma Resort - Wayanad Meppadi	4.129630
3 Costa Malabari Kannur	4.428571
4 Duke Palace Mathura	2.000000
5 Gmwn Tourist Rest House, Syalsaur Rudra, Prayag	4.200000
6 Hotel Classic Chandigarh Chandigarh	3.700000

The terminal window shows the following R code:

```
> View(Data)
> library(dplyr)
> summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mean(Review_Overall_Rating))
> View(summary1)
>
```

The Environment pane on the right shows the **Data** environment with 942 observations and 11 variables, and the **summary1** environment with 23 observations and 2 variables.

The screenshot shows the RStudio interface. The main window displays a data table with the following columns: **Hotel_Name_City** and **OR**. The data is as follows:

Hotel_Name_City	OR
19 RTDC Hotel Vinayak Ranthambore National Park	3.654762
20 Sneha Hotel & Resort Dhenkanal	4.600000
21 Sri Leela Bhavan Hotel Hyderabad	4.000000
22 Trishul Hotel Tiruvannamalai	4.000000
23 Unumbi Hill Palace Plantation Resort Idukki	4.666667

The terminal window shows the following R code:

```
> View(Data)
> library(dplyr)
> summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mean(Review_Overall_Rating))
> View(summary1)
>
```

The Environment pane on the right shows the **Data** environment with 942 observations and 11 variables, and the **summary1** environment with 23 observations and 2 variables.

So, if I just run this line you will get summary one which is exactly what I told. So, each of these hotels there are 23 hotels and corresponding overall ratings.

(Refer Slide Time: 26:34)

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

```
7  
8 View(Data)  
9  
10 library(dplyr)  
11  
12 summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mea  
13  
14 summary1 = Data %>% group_by(Hotel_Name_City) %>%  
15 summarise(OR=mean(Review_Overall_Rating), VFM=mean(Rating_Value, n  
16 LOC=mean(Rating_Location, na.rm=TRUE), SQ=mean(Rating_Sl  
17 RMS=mean(Rating_Rooms, na.rm=TRUE), CLN=mean(Rating_Clea  
18 SRV=mean(Rating_Service, na.rm=TRUE))  
19  
20 (Top Level) : R Script :
```

The Environment pane on the right shows the following data objects:

Object	Variables
Data	942 obs. of 11 variables
summary1	23 obs. of 2 variables

The Console pane shows the execution of the code:

```
> View(Data)  
> library(dplyr)  
> summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mean(Rev  
iew_Overall_Rating))  
> View(summary1)  
> |
```

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

```
7  
8  
9  
10  
11  
12 % group_by(Hotel_Name_City) %>% summarise(OR=mean(Review_Overall_R  
13  
14 % group_by(Hotel_Name_City) %>%  
15 in(Review_Overall_Rating), VFM=mean(Rating_Value, na.rm=TRUE),  
16 an(Rating_Location, na.rm=TRUE), SQ=mean(Rating_Sleep_Quality, na.rm=  
17 an(Rating_Rooms, na.rm=TRUE), CLN=mean(Rating_Cleanliness, na.rm=TRUE  
18 an(Rating_Service, na.rm=TRUE))  
19  
20 (Top Level) : R Script :
```

The Environment pane on the right shows the following data objects:

Object	Variables
Data	942 obs. of 11 variables
summary1	23 obs. of 2 variables

The Console pane shows the execution of the code:

```
> View(Data)  
> library(dplyr)  
> summary1 = Data %>% group_by(Hotel_Name_City) %>% summarise(OR=mean(Rev  
iew_Overall_Rating))  
> View(summary1)  
> |
```

Now, let us say I also told you that I also want to know where I am doing badly? It is the overall rating I am doing bad, I understand. My hotel is not up to the mark, I understand. But why, the question is, why? So, then I have to, so in the previous lines, I have added this part only this part is what I have added, so do not get scared there are 3-4 lines. But do not get scared of what do I have do?

Over and above, whatever I have done this OR is equal to mean of overall rating, then I wrote comma then VFM, value for money. This is a new column again I am adding on the summary data set. So, VFM is equal to mean of rating value, and why did I put comma na.rm=true, can you remember? That is because that there are 'na' values, so be if this particular column has 'na' values then you cannot do mean because mean will give you not available. There are some missing values it will give you. I do not know what to do with these missing values?

So, no result. So, you have to write na.rm=true. We did it in the last class, so that means we have to remove the 'na' value before I do the mean. So here, the first one is vfm, which is the mean of rating value, which is value for money. Then LOC, LOC stands for location, which is mean of rating location again na.rm=true. And then service quality, rma stands for I think rooms, CLN stands for cleanliness and SRV stands for service.

So all the 6 aspects I am finding out the mean. And if I just find out the mean if I run this line number 1 to line number 19 together, I have to run because all of these things are in the same code.

(Refer Slide Time: 28:31)

The screenshot shows the RStudio interface. The top pane displays a data table with columns: name, City, OR, VFM, LOC, SQ, RMS, CLN, and SRV. The bottom pane shows the R console with the following code:

```

summary1 <- c(mean(Review_Overall_Rating), VFM=mean(Rating_Value, na.rm=
=TRUE),
+           LOC=mean(Rating_Location, na.rm=TRUE), SQ=mean(Rating_Sleep_
Quality, na.rm=TRUE),
+           RMS=mean(Rating_Rooms, na.rm=TRUE), CLN=mean(Rating_Cleanlin
ess, na.rm=TRUE),
+           SRV=mean(Rating_Service, na.rm=TRUE))
> View(summary1)

```

name	City	OR	VFM	LOC	SQ	RMS	CLN	SRV
veer	Regency Chitradurga	3.200000	3.000000	3.375000	3.142857	3.428571	3.142857	3.500000
ya	International Srimangala	3.125000	3.714286	3.857143	3.400000	3.166667	3.714286	3.142857
as	Rourkela Rourkela	2.857143	3.571429	4.166667	3.333333	3.166667	2.833333	3.142857
ko	Hotel Bodhgaya Bodh, Gaya	2.800000	2.428571	4.166667	2.333333	2.714286	2.500000	3.285714
Hotel	Jajpur	2.000000	2.000000	3.000000	NaN	1.000000	1.000000	3.000000

The screenshot shows the RStudio interface. The main window displays a data table with the following columns: name, City, OR, VFM, LOC, SQ, RMS, CLN, SRV. The 'OR' column is highlighted in blue for the row 'as Rourkela Rourkela' with a value of 2.857143. The console shows the following R code:

```

summary1 <- c(OR=mean(Rating_Overall, na.rm=TRUE),
              VFM=mean(Rating_Value, na.rm=TRUE),
              LOC=mean(Rating_Location, na.rm=TRUE),
              SQ=mean(Rating_Sleep_Quality, na.rm=TRUE),
              RMS=mean(Rating_Rooms, na.rm=TRUE),
              CLN=mean(Rating_Cleanliness, na.rm=TRUE),
              SRV=mean(Rating_Service, na.rm=TRUE))
View(summary1)

```

If I just run this, I get a better version of the summary. So which has the hotel name and then there is overall rating OR and then there is all these ratings. So here I can probably sort them up also. So here if I have sorted up based on the overall rating, I know that let us say this guy, scroll down version they say this guy 2.85 is not doing good. But 2.85 is not doing good because of what?

Not because of location because the location is pretty good enough. Probably because of cleanliness, so something like that a basic idea I sometimes came.

(Refer Slide Time: 29:05)

The screenshot shows the RStudio interface with the following R code in the editor:

```

14 summary1 = data %>% group_by(hotel_name_city) %>%
15 summarise(OR=mean(Review_Overall_Rating),VFM=mean(Rating_Value,n
16 LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
17 RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
18 SRV=mean(Rating_Service,na.rm=TRUE))
19
20 summary1=data.frame(summary1)
21 summary2=summary1[16:17,]
22
23 barplot(as.matrix(summary2[,2:8]),names.arg = colnames(summary2[,
24 ylab="ratings",beside = T, col=c(5,6)
25 legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))
26

```

The Environment pane shows:

- Data: 942 obs. of 11 variables
- summary1: 23 obs. of 8 variables

The Console shows the execution of the code, resulting in the creation of summary1 and summary2 data frames.

The screenshot shows the RStudio interface with the data table view of summary1. The table has 23 rows and 8 columns:

Hotel_Name_City	OR	VFM	LOC	SQ	RMS	CLN
13 Lotus Nikko Hotel Bodhgaya Bodhi_Gaya	2.800000	2.428571	4.166667	2.333333	2.714286	2.500000
14 Mahua Bagh Resort Kashi	3.968750	3.827586	4.137931	3.888889	3.793103	3.928571
15 Midtown Hotel Handwar	3.782609	3.900000	4.100000	3.526316	3.761905	3.816260
16 Parthivas Rourkela Rourkela	2.857143	3.571429	4.166667	3.333333	3.166667	2.833333
17 Radhika Regency Rourkela	4.333333	4.137255	4.591837	4.166667	4.127273	4.285714
18 Ravine Hotel Satara	3.855030	3.702422	4.652330	3.862903	3.974910	3.800000

The Console shows the execution of the code, including the creation of summary1 and summary2, and the use of View() to inspect the data frames.

The screenshot shows the RStudio interface with the following R code in the editor:

```

14 summary1 = data %>% group_by(hotel_name_city) %>%
15 summarise(OR=mean(Review_Overall_Rating),VFM=mean(Rating_Value,n
16 LOC=mean(Rating_Location,na.rm=TRUE),SQ=mean(Rating_Sl
17 RMS=mean(Rating_Rooms,na.rm=TRUE),CLN=mean(Rating_Clea
18 SRV=mean(Rating_Service,na.rm=TRUE))
19
20 summary1=data.frame(summary1)
21 summary2=summary1[16:17,]
22
23 barplot(as.matrix(summary2[,2:8]),names.arg = colnames(summary2[,
24 ylab="ratings",beside = T, col=c(5,6)
25 legend(x=2,y=2,legend = summary2[,1],fill=c(5,6))
26

```

The Environment pane shows:

- Data: 942 obs. of 11 variables
- summary1: 23 obs. of 8 variables

The Console shows the execution of the code, resulting in the creation of summary1 and summary2 data frames.

So, what I will do in the next probably 2 minutes is that I will change this particular data to a data frame form; this is still data frame form, but I am ensuring it than it is in Data Frame. And then, I am taking the 16th and 17th row of dataframe.

So if I see the summary, see somehow I find out while I was creating this that 16 and 17 both are from Rourkela. So by chance that 2 hotels that have been taken are from the same city, so I am taking these two and they probably they are not there each other's competitor, I do not know. So it is just a dummy data set. I am not trying to say anything about the hotels, but probably they are competitors. So I will just collect this 16 and 17, and then so I am taking 16 and 17 as the row number, nothing after the comma means all the columns I am subsetting this data set.

(Refer Slide Time: 30:12)

The screenshot shows the RStudio interface. The main window displays a data table with the following columns: Hotel_Name_City, OR, VFM, LOC, SQ, RMS, CLN, and SRV. The data is as follows:

Hotel_Name_City	OR	VFM	LOC	SQ	RMS	CLN	SRV
16 Panthivas 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

The console window shows the following R code being executed:

```
C:/Users/Dell3/Desktop/mrkan/ >  
+ SRV=mean(Rating_Score, na.rm=TRUE), CLN=mean(Rating_Cleanliness, na.rm=TRUE),  
+ SRV=mean(Rating_Service, na.rm=TRUE))  
> View(summary1)  
> summary1=data.frame(summary1)  
> View(summary1)  
> summary2=summary1[16:17,]  
> View(summary2)  
>
```

The screenshot shows the RStudio interface with the following R code in the script editor:

```
18 SRV=mean(Rating_Service, na.rm=TRUE)  
19  
20 summary1=data.frame(summary1)  
21 summary2=summary1[16:17,]  
22  
23 parplot(as.matrix(summary2[,2:8]), names.arg = colnames(summary2[,  
24 ylab="ratings", beside = T, col=c(5,6)  
25 legend(x=2,y=2, legend = summary2[,1], fill=c(5,6))  
26  
27 #Missing value median imputation  
28 names(Data)  
29  
30
```

The console window shows the same R code being executed as in the previous screenshot:

```
C:/Users/Dell3/Desktop/mrkan/ >  
+ SRV=mean(Rating_Score, na.rm=TRUE), CLN=mean(Rating_Cleanliness, na.rm=TRUE),  
+ SRV=mean(Rating_Service, na.rm=TRUE))  
> View(summary1)  
> summary1=data.frame(summary1)  
> View(summary1)  
> summary2=summary1[16:17,]  
> View(summary2)  
>
```

So, now, as I have substituted this data set and I have only summary 2 where only 2 hotels are there, you will see the 2 hotel data is there. So, I can know that okay, which hotel is doing good which hotel is doing badly? So, as I was telling you that if you want to compare with your competitors, you can do that. And you can also create a bar plot to do that. You can give a graphical view of that by doing a bar plot.

And that is the part that will do in the next video. Thank you for being with me in this particular video, and I will come back in a few minutes. Thank you.