Foundations of R Software

Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

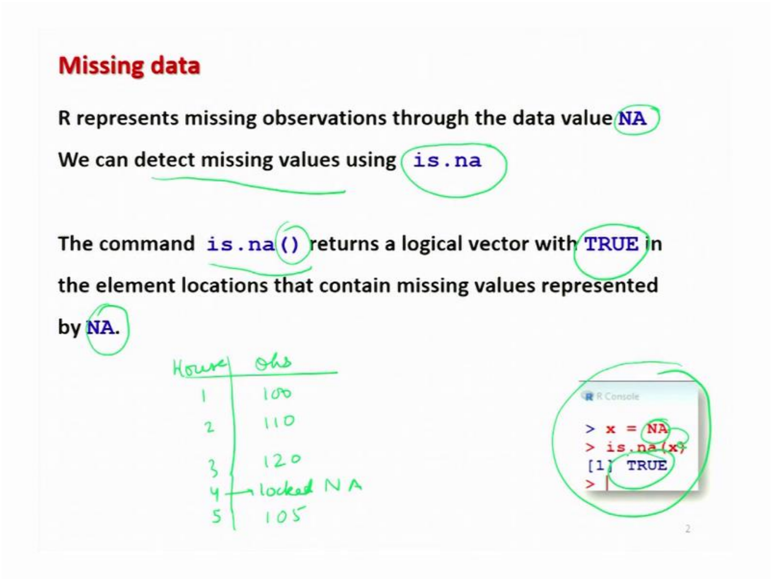
Basics of Calculations
Lecture - 18
Missing Data Handling

Hello friend, welcome to the course Foundations of R Software. In this lecture, we are going to begin with the new topic; this is about Missing Data Handling. What is this missing data handling? So, let me try to take a simple example to explain you; suppose you are trying to collect some data and suppose you have to go to five houses and you have to ask some value and then you have to enter it. Suppose you go to the fourth house and you find that the house is locked.

So now, you cannot obtain the value of the observation from the house which is locked. So, now in this case what we have to do and how we have to enter the data, so that it is handled in the R software? One thing I can inform you that in statistics, we have a full area, whole area which is called as missing data models; that we try to impute the value of the missing data by using the data that is available to us and we try to repair the data. So, you try to find out the value of the missing data and replace it there and then you try to conduct your statistical analysis.

So, there is a whole area, but we are not going into that; but I just wanted to convince you that how this missing data is handled in statistics, but here in this course we have a very different approach. The first question is this, in case, if this data is missing, how this data has to be entered or what is the rule by which R will come to know that this data is missing? And then once this data is missing, how to know and how to handle different types of operations? So, these are couple of questions, which we are trying to answer in this lecture.

(Refer Slide Time: 02:36)



Missing data

R represents missing observations through the data value NA

We can detect missing values using `is.na`

The command `is.na()` returns a logical vector with TRUE in the element locations that contain missing values represented by NA.

House	obs
1	100
2	110
3	120
4	locked NA
5	105

```
> x = NA
> is.na(x)
[1] TRUE
>
```

So, let us begin this lecture and try to understand how to handle the missing data values, ok. So, the first question comes that, how do you know that if there is any value which is missing? So, for that there is a rule in the R software that whenever we try to enter the data and if I know that the data has to be read in the R software; then the rule is that missing data is indicated by capital N, capital A. So, this is also a reserved word.

The example which I took here, suppose if I have here house number say 1, 2, 3 and here 4 and 5. Now, that fellow goes and it tries to collect the observation. So, suppose from house number 1, the value is 100; from the house number 2, the value is 110; from house number 3, the value is 120; from house 4, this house is locked.

So, the data will be recorded as here say NA and then after this the house number 5 has suppose some value 105 and that is all. So, this is how the data is recorded. So, as soon as you write NA, the R software will understand well this data is missing, that is the first rule. Second rule is this when you are trying to do any analysis in the R software, then the data is coming from some external sources and that data is in the form of some file etc.

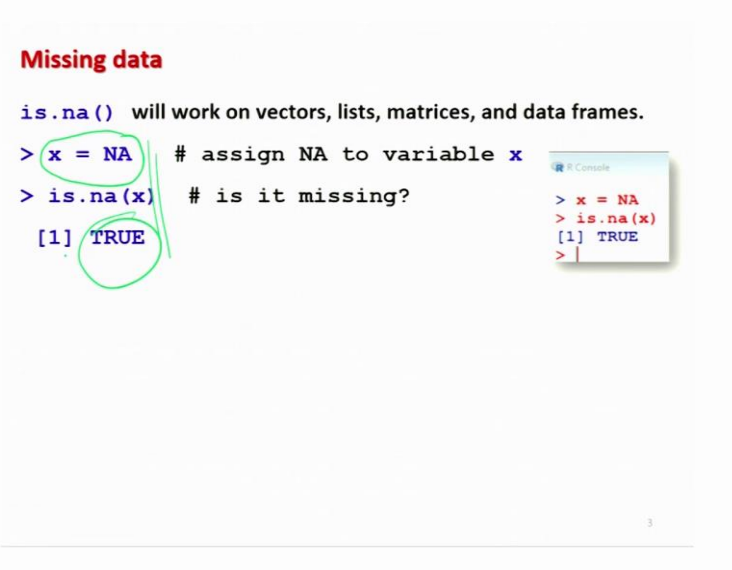
And that file may contain the data, which is say 1000 million etc.; there may be 1000 value, million values, etc. So, it is not really possible for you to scroll through the whole file and try to find that is there any value which is missing. So, you would like to have a command, so that you can detect the missing value from the file or from the variable.

So, the command here is, is dot na; all this is dot na they are all written in the lower case alphabets. And then in case if you want to know that is there any value missing in some variables; so you try to write down is dot n a and inside the parenthesis you try to write down the variable name. And this is going to return you a value which is a logical value; that will be TRUE or FALSE.

So, in case if the value comes out to be here TRUE, the outcome comes out to be here TRUE; that means the value is missing. And if the value comes out to be FALSE; that means that there is no value which is missing. And in all those location where the value is missing, the value is represented by capital N, capital A; that is the rule what we try to do in the R software, right.

So, if I try to show you here a screenshot here, I will try to show you on the R software also that I try to take here a value itself as capital N, capital A and I try to say here is dot n a and inside the parenthesis x, it gives me answer here TRUE, right. So, that is the same thing which I am trying to show you here, right.

(Refer Slide Time: 05:41)



Missing data

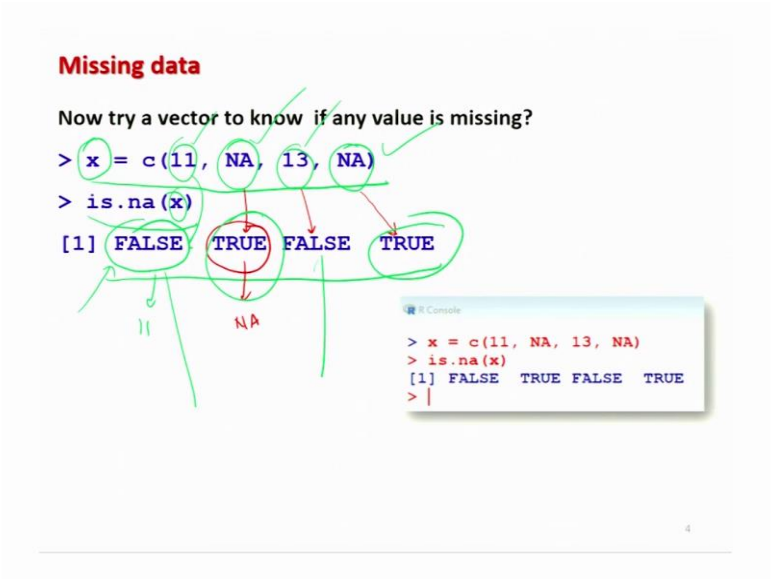
`is.na()` will work on vectors, lists, matrices, and data frames.

```
> x = NA       # assign NA to variable x
> is.na(x)      # is it missing?
 [1] TRUE
```

```
> x = NA
> is.na(x)
[1] TRUE
> |
```

So, when you yourself is trying to give a value NA; that means you are writing that the value is definitely missing. So, you are assigning the value NA to the variable and then you are asking is it missing. So, one you have assigned the value as NA; then when you are asking that is there any value which is missing, definitely the value is missing and you can see here that the answer comes out to be here TRUE.

(Refer Slide Time: 06:10)



Missing data

Now try a vector to know if any value is missing?

```
> x = c(11, NA, 13, NA)
> is.na(x)
[1] FALSE TRUE FALSE TRUE
```

```
> x = c(11, NA, 13, NA)
> is.na(x)
[1] FALSE  TRUE FALSE  TRUE
> |
```

Now, suppose I get a data and the data is a data vector. So, suppose the data has four values and out of four, two values are missing. So, let me try to write down this data here as say x which has four values; 11 and 13 they are known, they are some numerical values, but the values at second and fourth position they are missing.

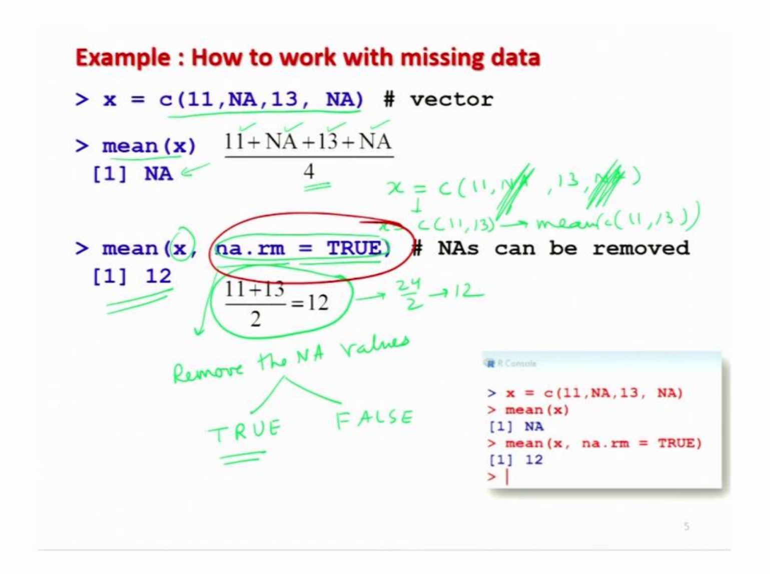
So, now I want to know in this data vector, where the values at first and third positions are available and the values at second and fourth positions are missing. How to know this thing? So, all this value have been stored in the variable x; so I try to execute here the command is dot n a and then I write down here the name of the variable x.

So, you can see here you get this type of outcome. So, now, our job is that, we have to; see what is the meaning of this outcome? So, here the first value it is here FALSE; this FALSE is corresponding to this value 11. So, this value is not missing, this is available. So, your question is, is dot na is the value missing and the answer here is FALSE; that means the value is not missing.

Similarly, when you try to take here the second value, which is here missing NA and your answer here is TRUE. So, what is the meaning of this? You are asking the question is dot n a, the answer is yes the value is missing and this yes is indicated by this TRUE. And similarly when the value is available, this gives you here the value FALSE and when the value here is missing in the fourth place, this gives you the value TRUE.

So, you can see here that this FALSE and TRUE appear at those places, where the value is not missing and missing. So, you can see here at first and third position, the value is available; so you get here FALSE like this and the value at second and fourth position this is missing, so you get here the answer TRUE. So, that means when you are getting the answer TRUE; that means the value is missing.

(Refer Slide Time: 08:18)



Example : How to work with missing data

```
> x = c(11,NA,13, NA) # vector

> mean(x)
 [1] NA

> mean(x, na.rm = TRUE) # NAs can be removed
 [1] 12
```

$$\frac{11+NA+13+NA}{4}$$

$$\frac{11+13}{2}=12$$

R Console

```
> x = c(11,NA,13, NA)
> mean(x)
[1] NA
> mean(x, na.rm = TRUE)
[1] 12
> |
```

So, now the question comes here, what does it make when the value is missing and we want to use any mathematical function? So, for example, in case if I try to say this example that in this data set suppose I want to find out the arithmetic mean; arithmetic mean you know, because this is the sum of all the numbers present in the data vector divided by the total number of values.

So, in case if I try to say here mean of x. So, what will happen? All the values 11 plus NA plus 13 plus NA will be divided by 4, there are four values. So, and the, but the answer comes out to be here NA, this does not help us. What we actually wanted? We wanted that out of this x vector there are two values 11 and 13 at first and third position and there are two values at second and fourth positions which are missing.

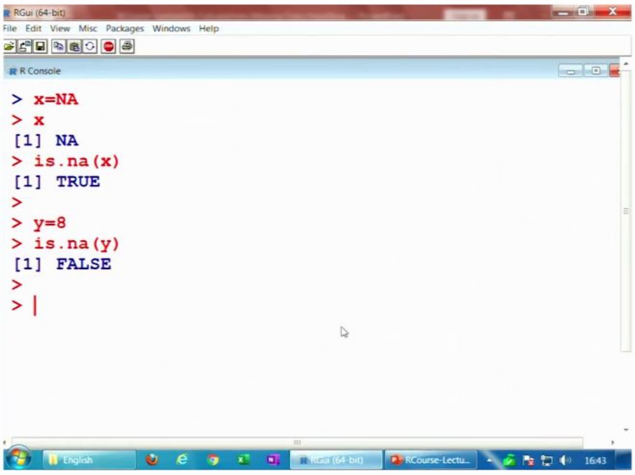
You wanted let this missing value are removed from this data vector x and then whatever is the data left in this data vector which is here 11 and 13, we try to find out the mean of this 11 and 13 like this. So, how to get it done that is the question. So, in case if you want that in any function, the missing values are removed and the function is operated over the available values; then you have to use an option which is n a dot r m is equal to TRUE, means n a dot r m means remove the NA values.

So, it has two possible outcomes; one solution is that you try to give here TRUE or you try to give here FALSE. So, when you are trying to give the answer as TRUE; that means it is asking or you are asking the R software that mister R please look into the values which are inside this data vector x and since I am asking that na dot r m is equal to TRUE, so please try to remove the missing value from the data and try to find out the arithmetic mean of the values of those values which are available inside the data after removing the missing values.

So, now if you try to see what will happen here, the answer comes out to be here 12. So, what is happening? That this NA values are removed and then you have here two values 11 and 13 and their arithmetic mean is coming out to be here 24 divided by 2 which is here 12. So, that is what happens when you are trying to deal with the missing values in the R software.

So, this is how you try to find out and this is how you try to operate them. Now, I try to first give you these operations inside the R software and then I will try to means come to remaining topics.

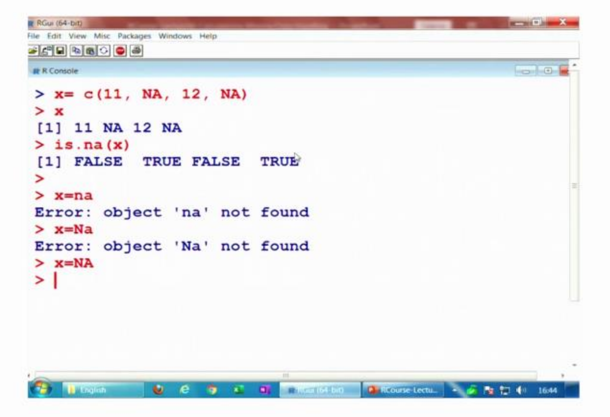
(Refer Slide Time: 11:17)



```
> x=NA
> x
[1] NA
> is.na(x)
[1] TRUE
>
> y=8
> is.na(y)
[1] FALSE
>
> |
```

So, let me try to take here this variable here let say x is equal to NA. So now, you yourself is giving the value x as NA. Now, if you try to type here is dot n a x, answer will be coming out to be here TRUE. On the other hand if you try to take here another value here y equal to 8 and if you try to find out here is dot n a y, it will say FALSE, right.

(Refer Slide Time: 11:44)

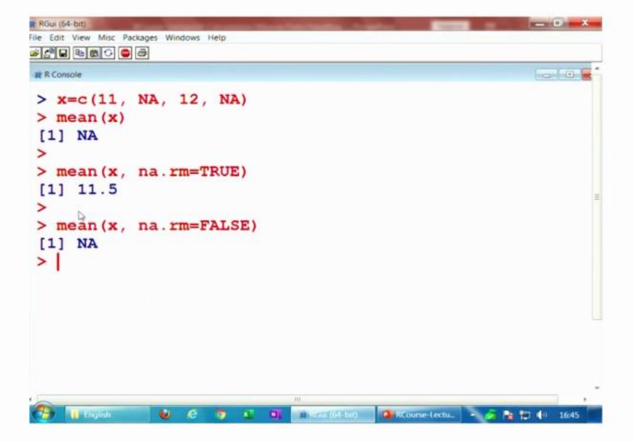


```
> x= c(11, NA, 12, NA)
> x
[1] 11 NA 12 NA
> is.na(x)
[1] FALSE  TRUE FALSE  TRUE
>
> x=na
Error: object 'na' not found
> x=Na
Error: object 'Na' not found
> x=NA
> |
```

So, similarly if you try to take here say vector here, say like as say 11, say NA, then 12 and then here NA. So, you can see here now this is your here x and if you try to find out here, if you try to give here is dot na x; you can see here that the values at first and third position they are available, so it is giving you answer FALSE and the values at second and fourth values are NA which are missing, so it is giving you the answer TRUE, right.

So, this is how it will work. Now, in case if you try to use suppose here small na; let us see what happened, this is not found. Even if you try to say small say capital N and small a, this is also not found; but if you try to see here NA, this is found. So, NA capital N, capital A this is a reserved vote, which you want to have.

(Refer Slide Time: 12:43)



```
> x=c(11, NA, 12, NA)
> mean(x)
[1] NA
>
> mean(x, na.rm=TRUE)
[1] 11.5
>
> mean(x, na.rm=FALSE)
[1] NA
> |
```

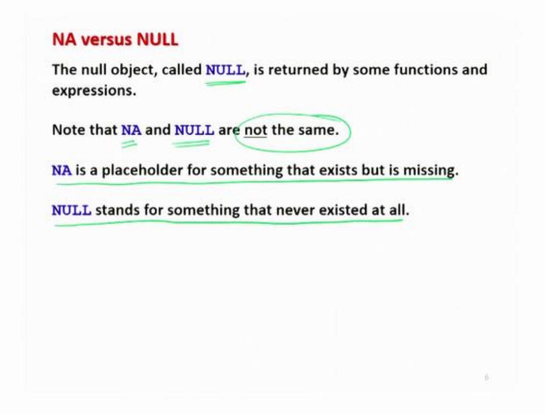
So, now in case if I try to show you here that is equal to if I try to take here two values here; say 11 and say here NA and then here say 12 and then here NA. And if I try to find out here the mean of x that are simply arithmetic mean. So, this comes out to be here NA; but in case if you try to add here an option that na dot r m is equal to TRUE; so you can see here now it is coming out to 11.5.

Why? Because 11 plus 12 that is 23; 23 divided by 2, this is 11.5. And in case if you try to make it here n a dot r m is equal to FALSE; then you will see that this is the default option when you are trying to use here, for example here mean of x. So, it will give you here NA and you are trying to tell well mister R please do not remove the NA values and try to find out the mean, ok.

So, you can see here it is not a very difficult operation, but surely I would like to address here one more concept. Here in this case you have seen that I have given you the concept of here na dot rm is equal to TRUE and n a dot r m is equal to FALSE and you have understood what really happens when this na dot rm is equal to TRUE and na dot rm is equal to FALSE.

Now, you will see that in forthcoming lectures, whenever we are trying to learn about any command; this option will always be there. Now, you have to take a call whether your data has missing values or not. In case if the values are missing, then you try to use this na dot rm is equal to TRUE always; whether you are trying to find out the mean or variance or median or anything else what we are going to do in the forthcoming lectures. So, that is another command which you can take that, I have explained you that how to remove the missing values and then how to compute the function.

(Refer Slide Time: 14:41)



NA versus NULL

The null object, called NULL, is returned by some functions and expressions.

Note that NA and NULL are not the same.

NA is a placeholder for something that exists but is missing.

NULL stands for something that never existed at all.

Now, just like NA, there is here another reserved word which is here NULL and you will see that sometime you are trying to execute a function and that function gives you an outcome as NULL. So now, let me try to explain you what is the difference between NA and N L and NULL. So, NA and NULL, N, U, double L which is in the all in capital letters. So, NA and NULL they are not the same; actually NA is a place holder for something that exist, but is missing and the NULL stands for something that never existed at all.

Actually what I can do that, I can take here a very simple example to explain you that what is the difference between NA and NULL. Suppose a school conducts an interest examination and suppose 100 students appear in the exam and suppose out of those 100 students, 70 students are admitted in their school.

Now, you have two classes of a student; a group of students who got admission in the school and another group of students who has not got the admission in the school. So, you know that in the school there is always an attendance every day, in which the class teacher marks absent or present.

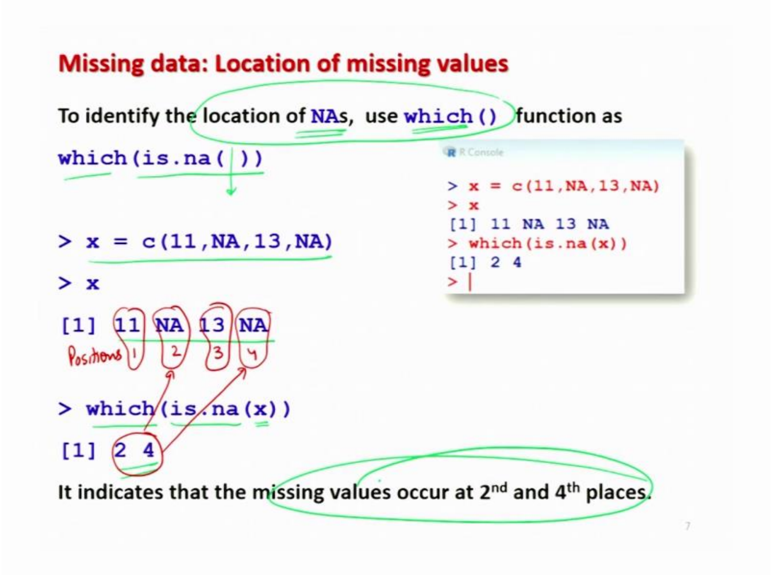
Suppose a student who has been admitted in the school is not there, the student is absent. So, the teacher will mark as absent. Now, try to consider one more situation, a student from the remaining group; that means the group of students who have not been given the admission, a student from that group is not coming to the school.

Do you think that the teacher is going to mark absent? No, why? Because this student was not admitted in the school; so teacher does not expect the student to be there. So, in the first case when a student who has got the admission is not coming to the school, the teacher will mark the absent; because teacher expected that the student will come to the school.

And since the student is missing today, so she will mark NA instead of absent, that is the same thing. And the student who was not admitted; in case if that student has to be marked in the school; whether that student is absent or present that has to be indicated, then that will be indicated by NULL. Why? Because that student did not got admission in the school; so nobody expect the student to be in the school, so that is why we use two options NA and NULL.

So, NA is trying to indicate that something was expected and the value is missing and NULL that is indicating that it was not expected, right. So, that is the difference between NA and NULL.

(Refer Slide Time: 18:41)



Missing data: Location of missing values

To identify the location of NAs, use which() function as

```
which(is.na( ))
```

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
```

```
> which(is.na(x))
[1] 2 4
```

It indicates that the missing values occur at 2nd and 4th places.

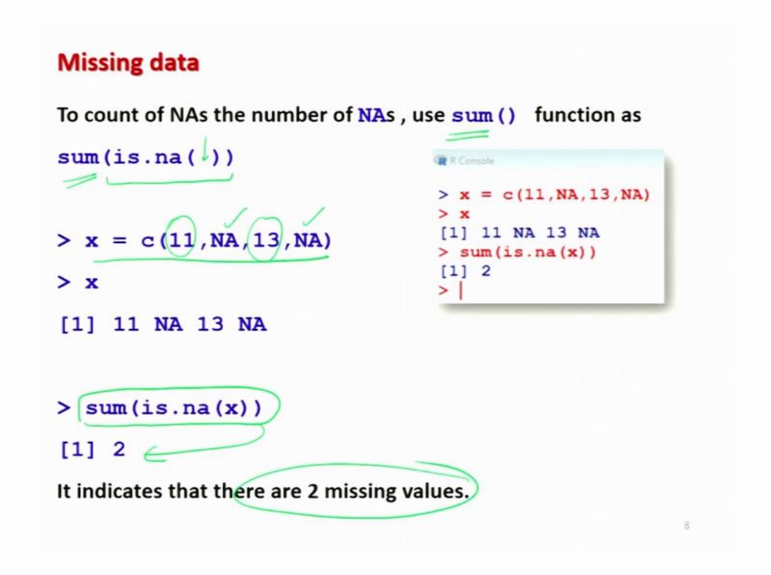
```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
> which(is.na(x))
[1] 2 4
> |
```

So, let us try to continue our lecture. So, now, we try to consider here some more basic operation about NA. Whenever some data vector has got some values which are missing and they are indicated by NA; we would like to know what are the values which are missing and for that our interest is that I would like to know the location of the values NA in the data vector. So, that will give us that the values at that location are missing.

So, now we are interested in finding out the location of the missing values in the data vector. So, for that we have a function here which; which all in lower case alphabets and we try to use here which and within the parenthesis which we write is dot na and then inside the parenthesis, we try to give the data vector. So, for example, we have taken here a data vector having four values x equal to 11, NA, 13 and NA, so which is here like this.

So, now, I try to write down here which is dot n a and inside the parenthesis x. So, it gives me here 2, 4. What does this mean 2, 4? So, if you try to see this data vector, there are four values and their positions are like 1, 2, 3 and 4. So, one is for 11, 2 is for NA, 13 is for has the position 3, and NA has got the position four. So, this 2 and 4 this is related to the positions of NA in the data vector. So, I can say here that, the values at second and fourth place in the data vector x they are missing, right.

(Refer Slide Time: 20:30)



Missing data

To count of NAs the number of NAs , use `sum()` function as

```
sum(is.na( ))
```

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA

> sum(is.na(x))
[1] 2
```

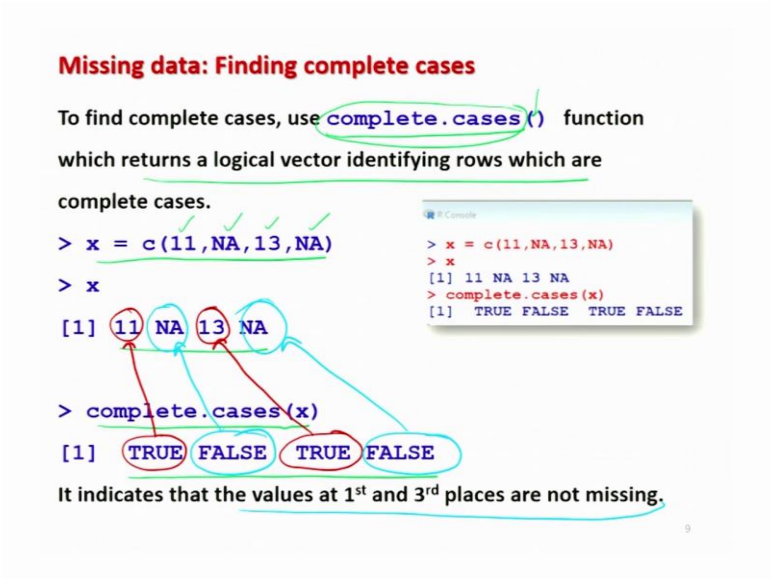
It indicates that there are 2 missing values.

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
> sum(is.na(x))
[1] 2
> |
```

On the other hand in case if we are interested in finding that how many values are missing in the data vector? So, in order to count the number of NAs, we use the function sum. And we use it like this sum inside the parenthesis I will write is dot na and then within the parenthesis I will write the data vector.

So, in case if I try to take the same example that x has got two missing values this NA and NA at second and fourth position and two values 11 and 13 they are available at the first and third position. Then if I try to operate here sum is dot na x; so you can see here the answer comes out to be here 2. So, that is indicating that there are two values in the data vector x which are missing. So, that is how we can know about this different type of information which we want to know from this data values.

(Refer Slide Time: 21:22)



Missing data: Finding complete cases

To find complete cases, use `complete.cases()` function which returns a logical vector identifying rows which are complete cases.

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA

> complete.cases(x)
[1]  TRUE FALSE  TRUE FALSE
```

It indicates that the values at 1st and 3rd places are not missing.

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
> complete.cases(x)
[1]  TRUE FALSE  TRUE FALSE
```

9

Now, another question, when you are trying to handle a data vector which has got missing values, you would like to identify that which of the data values are complete? So, we have here a function complete dot cases, c o m p l e t e complete dot c a s e s cases and this all is written in the lower case alphabets and inside the parenthesis, you have to give the name of the data vector.

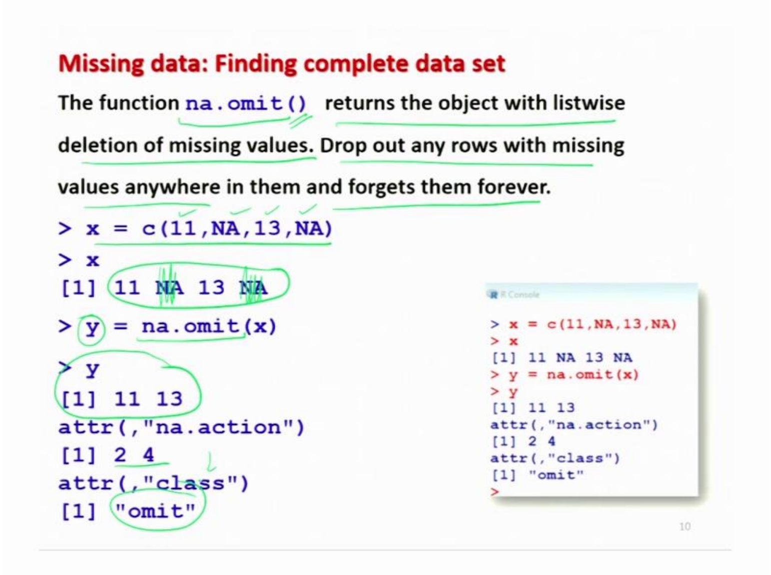
So, this is going to written as a logical vector and in which we have to identify that which are the complete cases. So, wherever the data is available, data is not missing, it will give the answer TRUE and wherever the data is missing, this operation will give the answer as FALSE. So, let us try to consider the same data vector which has 11 and 13 at the first and third position as available data and two values on the second and fourth position they are missing as NA.

So, this is my here vector x and I try to write down here complete dot cases and inside the parenthesis x, which gives me the operation; after this operation, I get the values here TRUE, FALSE, TRUE, FALSE. So, if you try to see here what is happening; this TRUE is corresponding to this 11, that means the data is available, this is the complete case. And similarly this another TRUE; this is at the third position, so this is corresponding to the third value in the data vector x which is 13 and that is indicating that the data is available and this is the complete case now.

Similarly if you try to look at the values which are FALSE; for example this here is FALSE, this is at the second place in the outcome. So, this is corresponding to the second value in the data vector x and here since the outcome here is FALSE. So, that is indicating that the value is missing and the case is incomplete. So, complete case is FALSE; that means the data is missing and similar is the story at the FALSE, which is occurring at the fourth position in the outcome, this is corresponding to the fourth value in the data vector x.

So, this is indicating that the values at the first and third position, they are not missing, but they are available and the values at the second and fourth position, where the answer is coming out to be FALSE, they are missing, right.

(Refer Slide Time: 23:59)



Missing data: Finding complete data set

The function `na.omit()` returns the object with listwise deletion of missing values. Drop out any rows with missing values anywhere in them and forgets them forever.

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
> y = na.omit(x)
> y
[1] 11 13
attr(,"na.action")
[1] 2 4
attr(,"class")
[1] "omit"
```

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
> y = na.omit(x)
> y
[1] 11 13
attr(,"na.action")
[1] 2 4
attr(,"class")
[1] "omit"
>
```

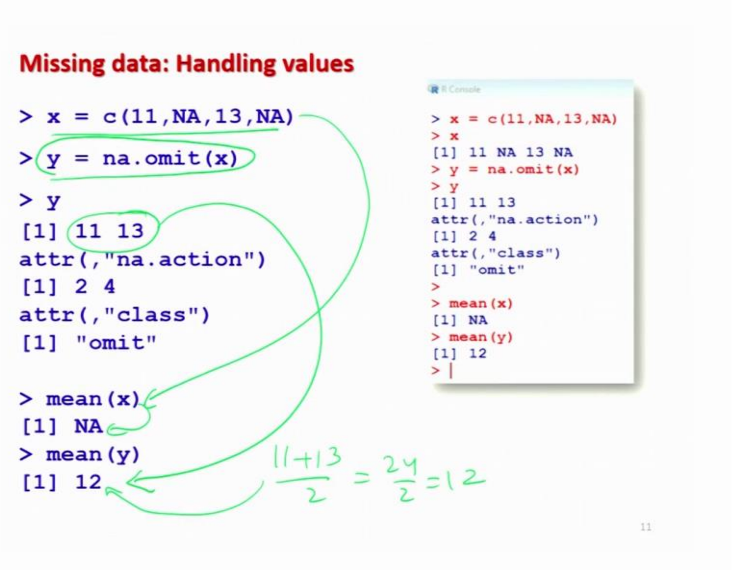
And after this I try to give you here one more operation, which is a very simple operation And in case if you want to create a data set after removing the missing values or after omitting the missing values; then how to get it done? Because all the time you are not interested only in the operations on the data values; but sometimes you also want to extract the data set which is completely, which has all the values, which is complete.

So, in order to do such operation, we have here a function na dot omit, na dot o m i t and within the parenthesis, we try to give the data value and this will return the object with list wise deletion of the missing values. So it, so what it does? It drop outs any row which has got the missing value anywhere in that data set and then forgets about it for always.

So, let me try to explain you this example; this concept through this example I try to take the same data set x is equal to c, 11, 13, NA, NA, in which first and third position are data is available and the data at second and fourth position is not available. So, I try to now create here another data vector y, which is obtained by using the command na dot omit and inside the parenthesis x. So, what I am trying to say, I have got here a data vector x like this; I am asking please omit NA and then please bring the data which is complete. So, that means please bring the data after omitting the NA values.

So, now you can see here this outcome comes out to be here like this 11, 13 and it has two attributes which are saying that it is 2, 4; that means out of 4 values 2 are missing, 2 have been removed, 2 are available and then this omit operation has been classified from the class this, because the class is omit.

(Refer Slide Time: 26:06)



Missing data: Handling values

```
> x = c(11,NA,13,NA)
> y = na.omit(x)
> y
[1] 11 13
attr(,"na.action")
[1] 2 4
attr(,"class")
[1] "omit"

> mean(x)
[1] NA
> mean(y)
[1] 12
```

$\frac{11+13}{2} = \frac{24}{2} = 12$

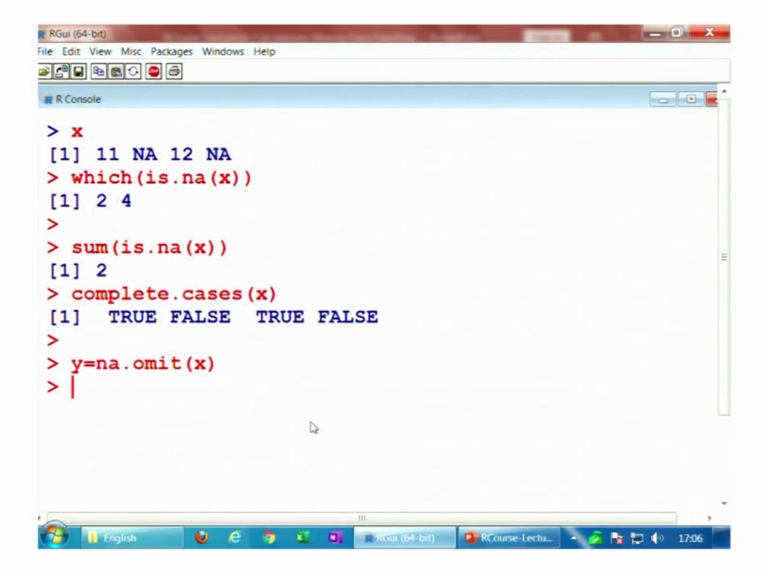
R Console

```
> x = c(11,NA,13,NA)
> x
[1] 11 NA 13 NA
> y = na.omit(x)
> y
[1] 11 13
attr(,"na.action")
[1] 2 4
attr(,"class")
[1] "omit"
>
> mean(x)
[1] NA
> mean(y)
[1] 12
> |
```

And now if you try to see what will happen? From the mathematical point of view, we have just learnt that how to find out the mean after removing the missing values using the operator na dot rm. Now, I am giving you an alternative way, I am trying to consider here x like this and then I am trying to create another data vector y, which is obtained after removing the missing values which has values 11 and 13.

Now, in case if you try to find out the mean of this x, this will come out to be here NA; whereas if you try to find out the mean of this y, this will come out to be 11 plus 13 divided by 2, which is 24 divided by 2 is equal to 12. So, the same operation which we had obtained earlier by using the command mean and inside the parenthesis x comma na dot rm is equal to TRUE; the same command can be used here or a similar type of command can be used here to get the numerical values. So, now, let me try to show you here these operations on the R console also, so that you get here more confident about them.

(Refer Slide Time: 27:24)



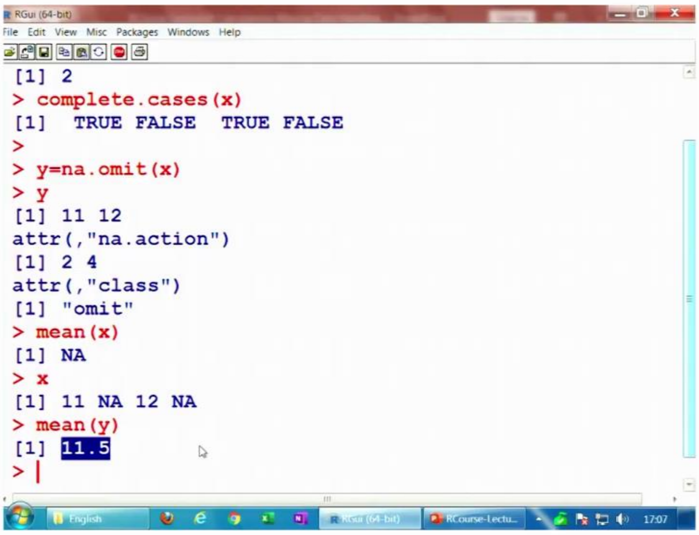
```
> x
[1] 11 NA 12 NA
> which(is.na(x))
[1] 2 4
>
> sum(is.na(x))
[1] 2
> complete.cases(x)
[1]  TRUE FALSE  TRUE FALSE
>
> y=na.omit(x)
> |
```

So, I try to first operate this which command and for that suppose I try to take here this x data vector, you can see here. So, in this data vector you can see there are four values and the values at first and third place are 11 and 12 respectively which are available and the values on second and fourth position they are missing. So, now, I want to use here this command which so, I try to say here which of the values are missing. So, it is giving me an answer 2, 4. So, the values at second and fourth position, they are missing.

So, after this I try to find out here their sum that, how many values are missing; so I try to use here the command here sum, which is like this. And you can see here that there are two values which are missing, ok. And in case after that if you want to see here that what are the complete cases in this case; if you try to see the command here, it is giving you here TRUE, FALSE, TRUE, FALSE.

So, these TRUEs, these two on the first and third position they are corresponding to 11 and 12 and these FALSE on the second and fourth position they are corresponding to these two NA, which are on the second and fourth position in the data vector x. And similarly, in case if you want to obtain here the data vector see here y, after omitting the values from x, so n dot omit and then from this here x. So, you can see here now y here is like this, ok.

(Refer Slide Time: 28:57)



```
[1] 2
> complete.cases(x)
[1]  TRUE FALSE  TRUE FALSE
>
> y=na.omit(x)
> y
[1] 11 12
attr(,"na.action")
[1] 2 4
attr(,"class")
[1] "omit"
> mean(x)
[1] NA
> x
[1] 11 NA 12 NA
> mean(y)
[1] 11.5
> |
```

So, now, if you try to find out here the mean of here x, so you can see here this is coming out to be NA why; because x has these missing values. And then if you try to find out here the mean of here y; this will come out to be 11.5, which is the mean of 11 plus 12 divided by 2, 23 divided by 2 which is 11.5.

So, this is how you can see that we can very easily handle these missing values in the R software; you just have to keep in mind the symbol and notation for indicating the missing value in the R software is NA. Actually if you try to go to different software, they have different ways of handling the missing data and usually in every software; there is a special symbol by which they try to indicate the missing value.

So, similar is the case in the R software also and R indicates it by NA; that is what you have to keep in mind and do not get confused with NULL, this is different thing that I explain you. So, why do not you try to create an artificial data set of couple of values and try to operate with this options; try to use some more functions which you know like a sum etc. and then try to see that how you can handle the missing values. For example, how would you like to find out the sum using the command na dot rm and then na dot omit and try to see are you getting the same value.

And similarly when we are moving now more into the R course, you will be coming to know you will come to know about different types of functions; try to operate them with and without missing values both. Actually that should be the default rule for the learning of the R software that whenever you are trying to learn a new function, always try to see how are you going to learn about the, how are you going to handle the missing values.

So, if you try to learn say for example, some commands to find out the median, variance etcetera in the further classes; also try to look yourself that how you can handle the missing values and how you can find the same median or same variance when the data is missing. So, that will give you a good practice. So, you try to practice and I will see you in the next lecture with more commands, more function, more details; till then goodbye.