Lecture 14 : Vector and Data Frame

Science Communication: Research Productivity and Data Analytics using Open Source Software

Mohit Garg

Central Library

IIT Delhi

Week: 04

Lecture: 14

Dear Learners, welcome again.

So, in the last lecture we discussed four of the value types that an object can take. That is one is numeric, then we have discussed integers, then we have discussed characters and finally, we have discussed logical. And we are left over 2 of the value type complex and row. Because that is not relevant to this course. Also I have asked you to create 2 objects in R. one is by your Name and another is by your pin area. Okay. Let us now first see how we can create those object Names and then we will go ahead. Let us say I am creating an object with the Name, myname and then I just assign myvalue. Okay. Now create another object with the object Name mypin and I assign the value 110016 and if I run this.

So, this object name is created and then if I run this. So, this object with mypin name will be created. Let us now see what the value of mypin is. So, the value of mypin is 110016.

So, this is the solution of the practice assignment I gave you in the last lecture. So, in this lecture I will be discussing the data structures in R. So, these are very helpful during the analysis. So, until now what we have discussed is this is an object 1, then this is another object. We have assigned a 1 value to this object. So, this is one object. then this is the second object and we have assigned this value.

What about if we have to assign many values to this object. Okay. If we have to add one more value here one more value here. say for example this is a name object and this is a pin object. Okay. Now in this we want to have another name like Mukesh, then we have Priyanka. and we want to give another pin. So, how can we do this kind of assignment to this particular object? So, these are possible with the help of data structure.

So, this one particular kind of this is a one kind of data structure. and then this is another kind of data structure. Okay. And what exactly this data structure is we will discuss now. In R there are six kinds of data structures. So, first is vector, second is data-frame, third is list, fourth is matrices, fifth is factor, and sixth is array. Okay. And the two of the data structures that we will be using very regularly is vector and data-frame.

So, first we will be discussing the vector and data-frame. and then we will go ahead to other data structures. Okay. So, let us first discuss what a vector is. Okay. So, vector is the most simplest form of data structure in R. Okay. So, if we see here in this particular two object type. So, this is a vector and this is also a vector. Okay.

So, in simpler terms you can say that a vector is just like a column. Okay. So, we can say that this is a column that stores a pin. Okay. So, this is another column where we store the name. Okay. Now, if we have to add more value into this column. Okay. most of you have created the column in a different spreadsheet. how you do it. So, you open the file and then you simply add the value 12, 13, 14, 15, or the name Mohit, Mukesh, Priyanka, Raman. Okay.

But in R we do this by using the function C. Okay. So, what C is? So, C means concatenation. So, we have to add more values to this particular object. So, what exactly we are doing here is we are directing R that you create an object which is a type of a vector and assign this 5, 6 value whatever we will be giving. Okay. So, we will see just examples in R how we can do that. Okay.

So, this is a function which we will be using for creating a vector. So, say for example I am creating an object with a name pubdoc and it can have 3 values. So, what do we have to do? So, first we have to name the object. So, this pubdoc is a name. then we have to use this C function. and then we have to give the value. Okay. So, it is 12, 14, 15. and if I run this, this is created. Okay.

And if I have to check the value what we have to do? We just put this and it if I run this. So, these 3 values are assigned and this is a kind of vector. What we have done here is we have created a vector pubdoc, which can have 3 values 12, 14, 15. Okay. So, this is the very first thing we have done. So, this is how our first vector is created.

Now, we will create another vector with names. So, how will we do that? So, now we are creating a vector with names. So, I am naming this vector as a researcher and the same function I am using is C and I am mentioning the 3 names. The only thing if you remember I have discussed in the last lecture is that whenever you have to define any character value any text value it should be in the inverted comma. So, the same way I am doing here. I am keeping all the characters in the inverted comma. and then putting the commas between them.

And if I run this, run this and if I have to check the value it is showing this particular value. The values are Raman, Mohit, Lokesh. So, I have created another vector and we will go back to the. So, this is another vector we have created, researcher Raman, Mohit, Lokesh. Okay. So, until now what we have done here is we have created 2 objects one is a pubdoc and another is a researcher.

Now if you check the class of this particular object and if you run this it is a character. So, this researcher is a character and if I check the class of our pubdoc it is numeric. Okay. So, we have discussed in the last lecture that if we do not explicitly mention that it is an integer it would not take it will be taken as numeric only. So, this is how you can create the vector in R. Okay. It is very simple now you have to remember a few things whenever you are creating the vector.

So, the first thing is that a vector can have the value of the same type only. Okay. If you give the numbers and then you will give them some text in it. So, what will happen let us see now in R itself. If I do like this, say for example I am creating another vector by vector 1 which has 2 numeric values and 1 character value. If I do like this and if I check the class of this, first let us see the value and what value it has been assigned. Okay.

So, if you see the values here this 19, 20 are not numeric. These are in character. Okay. What happened here is. So, in this myvector1 what have we have created this 19, 20 and this A, B. So, these 19 and 20. These are not numeric. Okay. So, these are what exactly are things that. So, R thinks that it is a text only. you have to remember that whenever you are creating a vector it should be of the same type.

Now let us say for example if we check the class of this vector. Okay. So, it is a character and it is not numeric. Okay. Now if I create another vector with only character, and if I run this. So, the class of myvector1 and myvector2 is character. Okay. So, this 19 and 20 is not numeric; you cannot do any operation.

Let us see through R only that why we cannot do the operation on this 19, 20. Okay. So, say for example, if I multiply myvector1 with 2. Okay. So, if we see through our pen paper example if this is a vector and this is a number if we multiply this by 2. Okay. So, it will be like this. So, the answer should be 38 and 40. It should be like that.

But what should be the answer here we will see in R. So, it is showing the error that it is a non-numeric argument. Okay. So, this is you have to be very clear that whenever you are creating a vector it should be of the same type. Okay. If it is a character it will be a character. if it is a numeric it should be a numeric. if it is an integer it should be an integer. Okay. Whenever you will be doing multiple large scale analysis and these value types become very critical Okay.

So, you have to be very careful whenever you are creating the vector that should be of the same type. So, these are some examples of valid vectors. So, these are of the same type. and this is also the same kind of vector in its characters and this is numeric. What could be the possible invalid vector type is. Okay. So, if I do like this it is a invalid vector type. it will create the vector. that if you direct R that you create a vector with these 3 values it will create. But it would not be of the type which you want. Okay.

If you are doing this way you mean that 19, 20 are numeric and A, B is character. Okay. So, that you cannot do in vector. Okay. It will take 19, 20 as a character only. This you can identify from the output itself. Okay. If you see the output of myvector1 this is in inverted commas. Okay. That means 19, 20 is not numeric. It is a character. But if you see here if I run this here again if I run this. Okay. and now if I run this myvector1 this 19, 20 there is no inverted comma. Okay. So, you have to remember this whenever you are doing the analysis of how the output of that particular vector is.

So, now I request you all that you practice to create this particular vector in your RStudio and let me know what exactly the response you get. Okay. Let us now see what exactly these different kinds of vectors are. Okay. So, one is numeric and as the name says that a numeric vector is a vector which takes all the numeric values. then we have an integer vector which can have only the integer value. then we have a character vector, which has the character value. So, whatever that value type we have discussed in the last lecture we can create the vector of each value type in R. So, let us say for example I am creating a numeric vector. So, it is a numeric if I run this it is 12 if you want to use this 12 as an integer you have to use l we have discussed in the last lecture. So, this is an integer vector if I run this 12 there. if I have to create a character vector we can do like this. and this is how our other vector is created. and if I see the response of this. So, my character vector has two values Mohit and Mukesh. and if I check the value of int vector. and this. So, if you see here this 1, 2 and this 12 are different. because this 1 2 is integer. and this 1 2 is numeric. Okay. So, this you have to remember whenever you are creating the vector. that what value type you are using in it is. Okay.

Now, let us do some basic operations on these vectors. So, I am creating a vector here myvector 19, 20 21 and if I run this myvector is created and this is another vector. So, how the multiplication is done. So, the way we do the arithmetic operations on the columns. So, similarly we can do the same kind of operation on this vector also. before doing these operations in R. Let us first see how we used to do these column operations.

So, what we have done is we have created one vector here vector1 and we have created here vector2. Okay. 19, 20, 21, 2, 3, 4. Okay. So, this is our vector1 and this is our other vector. If I have to do addition, how will we do it? It is cell wise. So, 21+ 4, 25. 20+3, 23. 19+2, 21. So, this is an addition. if I have to do multiplication we have to do 21*4, 84.

20*3, 60. 19*2, 38. So, this is multiplication cell wise. then if I have to do subtraction we will do 21-4, 17; 20-3, 17; 19-2, 17.

So, this is subtraction and division 21/4, 20/3, 19/2. So, this will come in decimal. So, this is division. the same way we can do all these operations in R itself. So, we will see how we will do in R. So, addition is very simple. you have to use this '+' symbol we have discussed in the last lecture. and it is if I will just click here and I will run here. So, the answer is correct.

So, what we can do here is that we can create a new vector which stores the additions of these two vectors. So, here we have created another vector addvector, and we have transferred the addition of these two vectors into this particular vector. So, if I run this and if I see the value of this addvector. So, this is the same answer. Now, if I have to do the subtraction. The same way I can do the subtraction and multiplication.

So, for multiplication we have to use '*', that I will do this. and for division if I do this. Now, if I see the output of these three vectors. So, subtraction all three values are 17. Now, multiplication 38, 60, 84. and then division if you see. So, it is in decimal.

So, this is how you can do the basic operations on vectors and this we have to use during our analysis. So, during the analysis many times we have to access the particular values of a vector. So, to access that particular value of that vector we must know where that value is located. So, in R the indexing system of a vector starts from 1. So, this particular value 19 is at location 1, then this 20 is the second location, and this is at the third location.

If we see here. So, our vector was like and this was our vector and I have to access the first value. So, it will be a first. If I have to access the third value, it will be on the third location. So, this is the third value. So, this is how you can access the particular value of that vector.

When you have the access of that particular value then you can do any operations also. So, you can do the operations of this particular row also. Say for example, you do not want to do the whole addition or the subtraction for the whole of the vector. But you want to do addition of this particular row. So, how will you do that? So, you will say that from this vector take the second value and from that this vector take the second value and multiply or add or subtract. Instead of doing for the whole vector you can do this for only one particular cell also.

So, let us now take this interesting example that you are creating two vectors and these two vectors are character vectors. So, I am taking these examples. So, you are creating these two vectors and they both are characters. So, I am running it. Now if you do the arithmetic operations on these vectors what will be the output? So, you do the practice

and let us know what is the response you get when you are running this particular command.

So, if you run this, what response you get you let us know through the Discussion forum. So, this is all about a vector.

Now we will discuss the data-frame. So, what data-frame is data-frame in R is nothing is a dataset. So, this one particular vector is there and this one vector is there. So, the data-frame is this whole of all the vectors is a data-frame. and whenever we will be doing the analysis we will be handling this particular data structure type during the course. So, a data-frame consists of vectors which are of different value types. We will see each other one by one. Let us now see how we can create a data-frame in R. So, the simplest way to create a data-frame in R is first you create this one vector and then you create this vector and then you combine these two vectors then you combine it.

So, the step one is you create individual vectors and then you combine it as a data-frame your data-frame will be created. So, we will see how we can create a data-frame in R. So, we have already created this vector. But I will run it again.

So, this is a vector. Now this is another vector. To create a vector we have used the function C. But to create a data-frame we have to use another function which is data.frame. So, how will we do that? So, we have to use this data.frame function. So, in the data.frame and here we have to use this researcher. So, if I run this data-frame is created and I want to supply this to another object which I named it as RESpub. If I do like this. So, now our object has this. So, we have created this particular data-frame and we have named it RES. So, this is how we can create a data-frame in R.

So, for this, this is the name of the object. So, these are the vector names or you can say it as a column name. So, this is a vector researcher which has three values Raman, Mohit, Lokesh. And this is another vector which has three values 12, 14, 15.

Now say for example, you have a data-frame and you want to add some values to it. So, the one way is. that you add values here like 16, 17 and then you add names here as Sumit and then Arvind. So, you create one vector here as a pubdoc and then another vector as researcher then you combine these two vectors and convert it to as a data-frame. So, how will we do in R? So, I have to add one more name here. So, I will just add one more name here. if I run this here, this is done. Now we will combine this as a data-frame.

So, this kind of error you also face. and this is an error of the Naming convention of what the name you have given to the object. So, myobject is a researcher. But I am using it here as RES is missing. If I do like this, it is okay. Now, one more value is added to this particular data-frame.

So, you do not have to worry much about these errors. During the analysis whenever you are using R you will face many of the errors. But those errors are straight forward. So, if you see here also the error of like. There is something missing. So, if you see here the error is that this one particular object is there. This is also one object that is there. But myobject is this one which has four values.

It is saying that there are different numbers of rows. So, you have to be very careful whenever you are creating a data-frame by using this approach. So, this is a lengthy approach and many times error comes because of wrong naming convention of the object. If you see here my data-frame is created with one more value and my earlier data-frame was this with only three names. Now, we have one more name. So, this is a lengthy way and this is like for this there is only one value. If we have to add 10 or 20 values to this particular data-frame it will take a huge time.

What exactly is the function you have to use? So, the function name is fixed. So, what we are doing here is we are fixing this particular data-frame. So, if we run this and this one data editor interface will come. Now, we can add more values here like Priyanka, and we can add like 4, 5, 8, 12, 13. So, if I close this and our new dataset is. So, these many values can be added by using this function. So, the beauty of this R is that it has many inbuilt functions that makes your task much easier during the analysis. and I will be sharing the names of all these functions that will make your task easier during the analysis. So, now say for example you want to add one more column to this dataset. So, this particular dataset has the name of the researcher then the number of publications.

Now, you want to add one more column that stores the number of citations. How will you do that? The same way the lengthy approach is that I will create another vector, like citation. Then I enter the values here like 2, 3, 4, 5 and this time what I will do here is that in one way I combine this vector and this data-frame. or I can combine all these 3 single vectors and then I combine it as a data-frame. So, this is a lengthy way to add one more vector to this particular data-frame. So, for example, I am adding more values here. So, this is another vector. if I run this. So, what I am doing here is attaching the column CitDoc to the data-frame. So, if I run this here our another column is attached to this particular data-frame. But this is a lengthy way to attach the one more column to a data-frame. Another way you can use this function is to fix. and if you do fix it here it will open here. and now you can name this as a CitDoc. It is numeric. Then you add the value here 1, 2. if I close this now this is the easiest way we can add the column to our data-frame.

If you want to go by this particular approach what will happen is that sometimes like say for example if you add here as these values. if I run this. now it will say that there are like 8 rows in that particular column. So, sometimes if you have a large dataset and if you do not have the knowledge about your dataset how many rows it has. So, it will show the

errors. So, the better way is to use this fix function and if you run this if you do not add this particular value if you remove this 8 here and it will automatically put NA and now if I like close this.

So, it will show the NA values. So, this is a better approach to attach a column to the dataset. Now our dataset is ready. So, this is our dataset. So, now this is our dataset. So, in a dataset the very first thing is that we want to know exactly the dimension of the dataset.

So, for that we have to use this function dim if I do this dim. So, it is 10, 3. So, what 10 and 3 means 10 means there are like 10 are the number of rows and 3 are the columns. for a small dataset. We can simply run the dataset and we can see here. and it is showing. But say for example you have 10,000 values. So, all those values will be shown here and some will be omitted. So, you would not be able to see the exact values. The better way is using dim function. So, if you do this you will get this output and this 10 and 3, 10 is number of rows and 3 is columns.

But you want to know only the number of rows and how many rows are there. So, for that the function is nrow and if you run this is 10. Here I request you to pause the video and just try to think what would be the possible command to know the number of columns in a dataset. So, to know the number of rows we have Nrow. to know the number of columns we will have ncol. So, if you thought of n col you are correct and if I do like this ncol number of columns are 3.

Now sometimes we want to know what exactly that structure of our dataset is. So, to do that we will use the str function it will give us more detail about this particular dataset. So, dimension and nrow and ncol function give us the just basic values that are the number of rows there are. These are the number of columns. But we want more details on how that particular data is. So, it is showing that there are 10 observations of 3 variables. and these 3 variables are one is character and other 2 are numeric and these are the values. So, this is how you can know about your dataset and then you can decide what analysis you will be doing on this particular dataset.

So, say for example you are calculating the average. and you put the average on this particular variable. So, it is a character variable how the average will be calculated. So, this is very important to know exactly the structure of your dataset. Then we have another function which summarizes the dataset that is exactly the minimum value in that particular dataset, what is the maximum value. So, for that we have this summary function and it summarizes the dataset. So, this is how it summarizes the dataset. that this particular dataset has 3 columns. One is researcher, another is pubdoc, and third is CitDoc. Then the length of the researcher is 10 and its class is character. Then pubdoc for pubdoc the minimum values are 4, and there is NA is also there, and maximum is 16.

Then similarly for CitDoc the minimum value is 2, and the maximum value is 12. So, say for example you have a dataset that has publication details of that particular institution.

So, if I take this as a dataset. So, we can say that the number of documents published by this author are 4. and there is one NA value of course there. So, we are not sure that this author has either published or not. So, we can say that at present we do not have the data of this particular author or how many documents that particular author has published. In A similar way we can say that the minimum number of citations of those documents, and then the maximum citations. So, another question for you is that if you notice here that these values are 4, 8, 12, 11, 14, 16 are coming. But these values are in decimals 2.0, 3.75, 5.50 why is it? So, please let us know in the discussion forum.

The next thing is to know more about the dataset. So, how can you do that using this head function? So, head gives you the top few values of your dataset. Okay. So, if you run this. it gives the 6 top values. Okay. So, these are the 6 top values of this particular dataset and then if you want to see the last 6 values.

So, you can do this. you can use the tail function. So, it will show the 6 values from the bottom 10, 9, 8, 7, 6, 5. Now, if you do not want to see that top 6 or the last 6 values you want to see only 2 or 3 values. So, you can also do that. Okay. So, how you will do that for that you have to mention how many rows you want. So, you can just give the value here if you run this. So, these are the 3 values of this particular dataset and if you want to see the last 2 values you can do like this. Okay.

So, these 2 functions are basically used to know about the few values of the dataset. But you can use either of the functions for both the things also. Okay. So, this is how you can check the top few values of your dataset using these head on tail functions. So, now we have this as a dataset and we want to see what exactly the column name of that particular dataset is.

So, we can use this colname function and we can run it like this. It will give the output that these are the 3 column names of this particular dataset. Now, if you want to change the column name of this particular dataset. Okay. So, for that what we have to do is So, we have to use this particular function colname and we have to assign the 3 colnames. Okay. So, I do not want researcher, pubdoc, and Citdoc as a column name of this particular dataset. I want to use “sch”, “pub”, and “cit”. So, if I do this. But now if I run this see our column names have been changed. Okay. Another way that we fix once and we already have. So, we can use that fix also. like say for example, if I do this if I run it here and I want to change the colnames. Okay. So, this fix function is very helpful and always use this whenever you want to change any colnames or any value to the particular dataset.

So, let us now do some of the basic data manipulation on the data-frame. So, whenever we have a datasetș we use it to check if any missing values are there. nor are there any NAs there. So, we can use the function dot NA to know if there any NA values are there. So, here if you see there are like 4 Trues there. So, the presence of these 4 True are that these 4 values are NA. Other than that these all False are showing that there is no NA at this particular location.

So, we want to replace these NA values with some possible values. So, we can go ahead for the analysis. So, the easiest way we have already discussed many times is that we can use the fix function. So, we can use this way and we can put here 0 and if I close this. Now, if I run it again it will show only 3 True. That means, that only NA are available at only 3 locations.

So, we want to replace these NA with some possible value. So, this we can also fix also. But there is a special function in R which we can use to replace these values. That is the replace function and this replace function is very useful during our analysis. So, I request you to give a try using this replace function also. So, how you will do that if you do like this. And now if I check it here again is NA there is no False. That means, now there are no NA values if I see here.

So, here this 1, 2, 3 is not a numeric or not an integer. Here is this: 1, 2, 3 is a character. So, what you have to remember during this replace function is how you are replacing it. What values are you replacing it with? Now, say for example, you have a large data-frame which has like 50 columns and you want to access 1 particular column of that data-frame.

So, how do you do that? So, for that you have to use this $ sign. As soon as you put this $ sign after your data-frame name it will show all possible kinds of options. So, these are the columns if you click here and if you run here. it will show only the values of this particular column. and it will not show the whole data-frame. So, this is how you can do it if I go for another like for the public.

So, these are the details. Now say for example, you want to do the addition of all these publications. So, how can you do that? So, these are some of these values. So, this is how you can access that 1 particular column in a data-frame. Now, after this you want to access the particular value of a data-frame. How will you do that? The same way we do it for the vector kind of thing.

So, if we have to access the value we what we have to do? We have to Name the data-frame then which column. So, you see in col say for example and let us now take that we want to take the third value. So, if we run this here. So, a third value is location and similarly we can access the publication. So, say for example, you want to see that the number of publications of the third author and the number of publications of the fifth

author are equal or greater or less than how we will do that? So, for that we will be using it like this.

So, these are the number of publications of the third author. So, now we want to check that the number of documents published by the third authors are greater than the number of documents published by the fifth author. So, we will check. It is False. So, that means the number of documents published by the third author is greater than the number of publications published by the fifth author. So, I used this less than '<' symbol. But you can use the other symbol also greater than '>' if you do like this it will show True.

So, this is the True value. So, this is how you can create a data-frame and these are some of the basics operations of a data-frame. So, this is all about the vector and the data-frame. So, in the next lecture I will discuss four of the other data structures that we use in R. I just request you all to just give a try to create your own vector and data-frame in R. If you face any of the issues please let us know. Thank you. I will see you in the next lecture.