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

Dr. Mohit Garg

Central Library

IIT Delhi

Week: 06

Lecture 23: Descriptive Analysis III

Dear Learners, welcome again to the lectures. So, in the last two lectures we have discussed a few of the concepts of descriptive analytics. So, first let us revise those concepts that we have discussed. and then we will go ahead about measures of dispersion. So, what we have discussed in the last two lectures. So, first we have imported the data and we have used the function read.csv. And if we have the data in other formats, we can use the other functions.

I will be sharing all those functions with the study materials. So, after importing the data, we have used the function View() to see what data is imported. So, what function have we used? We have used the function view and here the View() is in the capital letter. If you will do like this view(), all are in lowercase letters, it will not work.

So, I will request that you should try with both the function and what the output you get, please let us know. Then we have discussed the frequency, then we have discussed frequency and then in frequency we have discussed how we can count it. and then we discussed the cumulative frequency. And we imagine that we don't know the function to calculate the cumulative frequency. Then we seek the option of help in the RStudio itself. We went to the fourth quadrant of the RStudio and then we sought help and then we got the idea, we have to use the function Cumsum().

And after that we have discussed relative frequency. and how we will be calculating the relative frequency and how we have to use those arithmetic operations. Next lecture I will discuss this, how we can create our own functions. because throughout this course on R we have discussed a lot many functions. But we haven't discussed how we can create our own functions. So, we have created our own functions. So, if you remember we have created Mohit one function, then we have created Mukesh and we have also created Priyanka. And I requested that I always give a valid name related to the task. So that you can remember what exactly that function is. and how it is working.

And in functions we have discussed about the function which doesn't require arguments and some of the functions like addnum require argument. So, we have also discussed the functions which require one argument and then two arguments. Then we have discussed the mean, how we can calculate the mean using the existing function. and then we have created our own function. Then for the median we have also calculated step by step and also by using the inbuilt functions. And for mode there are like no inbuilt functions are there. So, we have created our own functions for mode. and we have also seen the whole analysis step by step in R.

So this is how we imported the data. This was our dataset. I have also discussed the path and why it's very important. and if your current directory and the data where data is existing are the same. then you don't have to give the whole path. But in case your current directory is different. But your data is in another directory. then you have to give the full path where your data is. Then we discussed this View() function.

If I like it, I already asked you to try this. like let's see if I do like this view. and now if I run this. it's saying that I could not find a function view. And if you remember I have discussed in very starting that what are the things we have created in R it's case sensitive. So, like VIEW() all uppercase, or View() with one letter as an uppercase and all others like view() word as all lowercase are three different words. So, you have to take care of this whenever you are pulling the function. it should be the right function.

Then we have imported the data with the location. Then we have counted the frequency and we have also counted this Table() function. So, Table() function is the important function which counts the occurrence of a particular value. So, from here we have counted that for like 2019 there is 1, then 2021 there is 1, then in 2022 there are 5 occurrences, and then in 2023 there is only 1 occurrence. So, after that we have saved the data. So, once we have done the analysis we have to save it as an external source. So that we can show it in our presentation. or maybe we can use that analysis in our word document. and in the same way the path is very important whenever you are saving the file.

So the same way the function for importing the data, data which is in the CSV file we have used read.csv() to save the data which is in the data-frame and we want to save it as a CSV we have to use this function write.csv(). So, if you see here that many of the functions are task oriented. Their names are based on the task related to that.

So if you see write.csv() it is showing that there is something. There is some data there you save as a CSV file. or if there is data you read it as a CSV file. So, you do not have to remember all the functions. Just you have to see I have to import the CSV data. So, import it into the read.csv function.

So you will be like when you will be doing a lot of practice, it will automatically help you in recalling those functions name. Then we discussed the cumulative frequency. and here I discussed about the help that we do not know what exactly function we have to use for calculating the cumulative frequency. And then we have moved to this fourth quadrant and we have seen after searching that we got that. Okay. cumulative frequency we will use cumsum(). Okay. After calculating the frequency and then the cumulative frequency we have made a single data-frame and then we have externally saved that particular data. Also we have changed the colnames() for our understanding.

Then we have seen how we will be calculating the relative frequency. So, this is how we can create the relative frequency. So, this is like there is no inbuilt function. So, we have created our own function. So, this is our division operation and here one operation is this sum() and another is this is calling one by one those values. Okay.

After that we have like once our both the analysis is done we have saved this data as an external CSV file. and the file is stored to that particular location. But where the current directory is. Okay. And then we have also seen that whatever the things we have discussed throughout the lecture, how we can do that in just two or three commands. Okay. So, this is how we can do that. So, I request you again whenever you are starting. if you are a beginner with the R you always go with this long way because this is how you will get more understanding. and once you are comfortable in handling the functions and all, you go the shorter way. And most of the time like many of your bigger tasks will be solved in just one function. Okay. We will see all those examples.

In the second lecture I have discussed the functions. and how we can create the functions. So, to create a function we have to use the syntax like this. So, to define a function we have to give the name of the function. then assignment operation. then we have to call this function(). and then we have to under those parentheses (). we have to give exactly the task we aim to do in this function. Okay. And we have created a lot of functions. Okay. Then we have discussed the function based on a single argument. then based on the two arguments. After that we will have also discussed what output the function will give. Will it be like all those operations you have given, or on the last operation. So, we have found that the function will only return the last operation. and we have seen those examples here. and if you have to return the value after multiple operations, you have to do it like this. So, first you have to assign the thing that whatever the task is done. and then you have to call it as a last statement. Okay.

Whatever the operation will be in the last statement it will be the function that will give that output. Okay. So, we have calculated this mean. First we have like I did the all the frequency and then we have divided by the number of observations. and then we have also seen that a straight forward function is there. So, this is how we have calculated the mean. then for the median we just like to calculate the median we have ordered the data. So, it is up to you whether you are doing it in ascending order or descending order.

So the only thing you have to remember is that for an even number of values the median will be the average of two middle values. But for an odd number of observations the median will be the middle values. Okay. So, whatever the value is in the middle location that will be the median. Okay. Then we have seen the median by a shortcut way. So, there is an inbuilt function. There we do not have to do all those things. we have to just simply call the function median().

Then we have seen about this mode because there is no inbuilt function for it. First we have done manually then we have this we have created our own functions. and the last thing I have said in week 4 also that. There are like multiple packages which have a large number of functions to solve the various problems. Okay. So, instead of writing your own functions you can use the other functions which are already created by some other member of the community. So, we have called this package DescTools. So, DescTools has a function named mode() which gives you the value after calculating the mode of that particular dataset. Okay.

So in this lecture I will be discussing the measure of dispersion. and the similar way that I have discussed in the previous two lectures. I will be calculating this measure of dispersion first manually and then we will do it in R. Okay.

So if you remember our dataset was this. This was the year 2019, then 2021, 2022, 2023. This was frequency 1, 1, 5, 1. So, in measure of dispersion the first thing we will be calculating is the maximum value. Okay. So, if you see this data. we can easily see what the maximum value is. So, the maximum value is 5. So, here MaxValue is 5. If we see this particular data the minimum value is 1. So, the MinValue is 1. So, we will be calculating the maximum and minimum values in R. So, how will we do that? So, just think of what will be the function name for calculating the maximum and minimum value. Okay.

So we have to calculate the maximum. and then we have to calculate the minimum. So, as I already mentioned that the function names are related to the task. So, here if you see for maximum, the function name is Max(). and to calculate the minimum function it is Min(). Okay. So, there is no such function name as Mukesh which calculates the maximum value or Mohit which calculates the minimum value. Okay.

If you are totally new to R, just think of what the function name could be for maximum. Okay. This can also be a possibility, if a function can be a Maximum(), or maybe Max(), or maybe Maxi(). So, we can try all those things. and whatever the things, it will result we can use that function. Okay. So, let us assume that we don't know what function we have to use to calculate the maximum value in a dataset. Okay. So, we'll do like this. So, I just do maximum on our data is Max(DfdatFq$).

So an error popped up saying that there is no such function. Now if we go for our Max() function Maxi(). and I'll search here Df$. And if I run this here again for maxi, there is no such function there. or let us take one more example. So, I am assuming that there is a function called M which calculates the maximum value. But if there is a function named M(). So, everyone will be confused. So, M() is used for maximum or minimum. Okay. That is why the maximum of the function names are in such a way that they are related to one specific task. Okay.

So we'll see M also. Okay. So, df$ if I run this could not find function M(). Okay. Now we'll go to our next function named max(). So, if I see here and the output is 5. Okay. So that is true. So, whatever the value we have calculated manually. So, it's the same result.

So max() is a function to calculate the maximum value in a dataset. We have seen all those examples of possible function names. Similarly we will see the minimum also. So, for minimum value let us assume that there is a function called minimum. and now if I say. Okay. So, there is no function either starting with minimum() or min().

Okay. I have already mentioned that the MINIMUM() in upper case, and lower case will be different. So, there could be a possibility that there are some function names that are in upper case with the minimum also. So, we'll see that also. So, yes. there is no such function whether it's in upper case or lower case. Let us try all function names in upper case only. Okay.

So we'll try like MIN() then. Okay. So, there is no such function with the MIN name which calculates the minimum value in a dataset. Okay. So, MIN() in upper case there is no function. But min() in lower case is the function which we will be using to calculate the minimum value. Let us check now how we can do that. Okay. So, if I run this I got the right value 1. that 1 is the minimum value in the dataset. So, max() and min() are the two functions which are used to calculate the maximum value in a dataset then a minimum value.

Then in the next concept in the measure of dispersion is range(). So, what range is range() is the difference of maximum value and the minimum value. Okay. So, range() is maximum value-minimum. So, we have already calculated this maxvalue() by minValue(). So, the range will be 5-1 = 4. Okay. So, first we will not think of any inbuilt function. because now we have to just simply use the arithmetic operation. We have already calculated the maximum value. we have already the minimum value and we have to subtract that and we will get the range. Okay. So, this was our maxvalue and this was our minvalue. If I do this operation I get this and the range is 4. Okay. Or what you can do also is that you can just assign this maximum value to here maxi(). Okay. and this is maxi value and then also what you can do here you can do mini(). Okay. If I run this. Now if I have to calculate the range of what I will do I will call maxi()-mini(). and if I run this and it is saying that object maxi is not found. Okay. What is the error? Can anybody guess? So, if anybody can guess, just post the video here. and let us know in the discussion forum why this error maxi object not found came. Okay. So, now if you see here that we just simply wrote the object name here. But we have not run this object. Okay. So that is why the object was not created. So, we have to do first this maxi and then this mini. Okay.

Now if I do like this it is again showing that maxi() is not created. So, it is again showing maxi not found why? Because there is no such function called maximum() and I have assigned the object name maxi() to a wrong function name. Okay. So, what I have to do here is I have to just simply copy or maybe we will cut and then this is our maxi(). Okay. Now if I run this now if I again calculate the range, you will get the right answer this time. I hope so. Yes, I got the right answer of value 4. Okay. So, whatever the value that I have got earlier. So, whenever you will be doing in R there is a possibility you will be finding some of the errors. Okay. during your calling the functions or something is missed. Okay.

So you don't have to worry about it every time it is showing errors. Whenever you are running errors come. Okay. This is how we will learn that. Okay. if an error is coming something is missed. Okay. So, if you see in this whole example just what the example. While I was discussing it, I faced two errors. Okay. First error was the maxi object name to that value but I haven't created it. Okay. and the second error was that I have assigned the maxi object name to the wrong function name. Okay. and you see this error also. They are very straightforward. one can easily understand that you have created a maxi object to a wrong function where a function doesn't exist. Okay. So, I got the idea that where that error is there and if you see that to resolve these errors I didn't go to the internet or any kind of documentation. Now whatever the error came I just read the error. Okay. So, this is the error: could not find the function maxi. earlier I got here that object maxi was not found. These errors are self declarative that this error is because of this possible reason and you resolve this and this problem will be resolved. Okay. So, now we will go to the next concept is variance. So, what variance is we will calculate here variance. So, the formula of variance is var = $\sum(x_i-\bar{x})^2/n$.

So n is your number of observations. So, what I will do is first calculate the $\bar{x}$. So, what $\bar{x}$ is. $\bar{x}$ is the mean of summation of this 5+1+1+1. So, we will be calculating the $(x_i-\bar{x})^2$. So, this will be $(1-2)^2$. this will be $(1-2)^2$. this will be $(5-2)^2$. and this will be $(1-2)^2$. So, this will be 1, 1. this will be 9, and then 1-2, 1. So, our $\sum(x_i-\bar{x})^2$ came 12. So, our variance will be 12/4. it will be 3. So, now we will calculate the variance in R. So, how we will do that exactly we will do this operation. So, let's do that. So, we have already calculated the mean(Fq). So, mean(Fq) is 2. So, now we will calculate first $(x_i-\bar{x})$. So, this is our $\bar{x}$ mean. and these values are there. So, if I do like this, we get this value. Okay. we will assign to this object. Now we have to square it this value. So, how will we do that?. So, we will use it this way. So, SmFq() was not found. why? you see here SmFq(). F is capital here f is small. So, I have to give this word. So, if I do like this I will assign this to SQ(Fq). So, SQ(Fq) is there. Now we will use the formula variance. So, this is our sum of SQ(Fq). So, this is 12. Now I will count the number of observations. So, it's N 4. So, it's 4 now and I divide those two. So, our variance is 3. I got it. So, whatever the output we are getting here it's 3. and here also we are good. So, we will assign this to well now if I see the output of this variant of Fq it's 3. Okay. So, whatever the value we have calculated manually is the same thing we have calculated in R also. So, we got our variance. Now think of what could be the possible name of the variance function. Okay. So, it can be either variance. either on the uppercase or maybe the lower case or like this. Okay. or maybe this it can be. Okay. you just give a try on all these possible function names the community can decide. and I just share with you the like what is the valid function name to calculate the variance this one. So, we will be using this particular function to calculate the variance. So, if I do this. Okay. So, our variance is 4. But what happened is that our variance we have calculated the value is 3. and now here the value of variance is coming 4. But why? So, R is wrong. or we are wrong. So, neither we are wrong nor R is wrong. So, what we have done is we have calculated the population variance when we were calculating the manually we were calculating the population variance but for the sample variance we have to use n-1. So, in R the function calculates the variance of a sample. Okay. So that is why I am taking you first to the manual analysis. and then I am doing it in R. Okay. because what function is given and behind that how that function is written you must know. Okay. then only you will have good results and then you can interpret the result. Okay. This is the sample variance. Okay. So, for example what you have to do is here we have to just simply give the length of frequency is n, n-1. So, if I do like this and if I now run it again. So, our value came 2, and if I like this but we haven't assigned no. first we will assign this to it's 2. and now if I do this it's 4. But why it's coming 2 values? Okay. But we are variance as per the R is 4. So, let us now see what will get the variance value when we will be calculating manually. Okay. So, what we are doing here is that our sample variance is $(x_i-\bar{x})^2/n-1$. Okay. So, $(x_i-\bar{x})^2$ is our value is 12. So, 12/(n-1). n is our 4. (4-1) that is equal to 3. So, 12/3 = 4. So, 4. So, now we are getting 4. So, now here we are getting the right answer. But R is calculating. So, we can say that R is calculating the right variance of a sample. So, what we can say here is that something is wrong in our writing that calculation. we will identify what that mistake is. So, here if you see this sum(sqFq)/length(DfdatFq$Frequency)-1. But how it is calculated. So, it is calculated this way. So, the way we have written those arithmetic operations. So, there is something wrong with that arithmetic operation. But that mistake is that. The mistake is this. So, the meaning of this calculation is this. Okay. So, if you are running in this way, it will calculate a different formula. Okay. So, here you have to remember a very simple calculation BODMAS. Okay. So, what BODMAS is. BODMAS exactly defines the precedence of the operation. Okay. So, here if you see in this particular place the first division happened. Okay. and then subtraction happened. because in the BODMAS division comes first. So, division happened and then subtraction occurred. So that is why we are getting the wrong answer here. So, to correct this what we have to do here is that we have to do like this. now if I run this and if I check the value it's 4. and our sample variance formula is also. Okay. So, you have to be very careful whenever you are creating your own operations in R. Okay. because our earlier operation was like this sum/length-1. But what we want is a sum/length-1. Okay. So, there is a big difference in all these two operations. Okay. So, you have to remember this whenever you are creating your own calculation you just see how you are interpreting it. Okay. So that is why we are getting the wrong answers of variance. Okay. So, I hope you have got a clear idea of how you have to create your own operations in R. and what are inbuilt functions in R and what exactly that function is calculating. Okay. So, the variance the default function is calculating the sample variance. Okay. So, now we will calculate the standard deviation and we know the formula of standard deviation. It's a square root of variance. So, whatever the value we have, it's a square root of variance. So, our variance is 4. So, our standard deviation is 2. Okay. So, we will do the same in R also. So, how will we do that? So, we have to calculate the square root. Okay. So, there is a function in R for calculating the square root. what that function is that function is. So, this function is SqRt. So, SQRT is a function if you want to know about SQRT we will use the help. So, this is how this SQRT function works: it calculates the square root. So, as a beginner I highly recommend that whenever you are pulling any of the functions you must use this symbol ? question mark. Okay. So, this will take you to the help and give you the example of how exactly you have to call that function. What are some of the examples there ?. Okay. So, this variance Fq we have already calculated. Okay. So, what this was our 4 and now if I calculate the square root it's 2. and the standard deviation will assign to this. Okay. So, our standard deviation is 2. Okay. now again the same practice. if we have to calculate the standard deviation what function name it should be there. Okay. any guesses, if you can guess any of the names of the function post the video here and let us know in the discussion forum that what function name you thought of it will be the possible to for calculating the standard deviation. Okay. So, to calculate the standard deviation in R we have the function name Sd(). Okay. So, until now we could think of the function name as our maximum. We could think of maximum(), maxi(), or max(). But if you see that there are like operations is standard deviation. Okay. So that SD() means standard deviation. So, you have to think in that way also. Okay. it's not always that we can have something like StdDev(). Okay. we can think of that could be the function name but how we will write that particular function. Okay, whether we will like standard deviation is there if we do like. this standard deviation then we'll call. So, space cannot be there. So that is why you have to just give it a try. Okay. what could be the name of that function to calculate the standard deviation. So, if we run this too. So, our this and this value is equal. Okay. and also like you can use our logical operators. how we will do that simply. and if I run this it's going through that means the standard deviation but we have calculated without inbuilt function and the standard deviation but we have calculated using the inbuilt function is same that means our operations are. Okay. and the standard deviation the value of this particular data is 2.

Okay. Now we will discuss some of the other concepts in measure of dispersion like quartile, percentile, and decile. Okay. So, what quartile is. quartile is the 25th percent of the data. So, this is like Q1, Q2, Q3. Okay. So, this is at 50%. Okay. and if you see the median, the median is only Q2. Okay. So, Q1 is 25%. then Q2 is 50%, and Q3 is 75%. then we have to decide on the 1/10th. Okay. 10%, 20% like that. and then we have percentiles. Okay. 1/100, 1, 2, 3, 4. So, this is how we can also calculate this quartile, decile and percentile on our dataset. Our sample dataset is too small. So, we'll give a try based on what the inbuilt functions are there to calculate quartile, or decile, or percentile. Okay. So, again the same practice you do here is also. what should be the possible function name to calculate the quartile, decile, and percentile. Okay. So, if you have thought of the function name of Quartile() or maybe this Qua() or anything else. Okay. you are going in the right direction. At least you are thinking about what could be the possible function name. But in R the function name to calculate the quartile is Quantile(). Okay. and how you have to call that function quantile() is in this way. So, we have to call this quantile and then this is our dataset. and what is like how we want to define it like 0.25. This is the first quartile then this is the second quartile. and this is the third quartile. if I run this. So, 1,1,2. Okay. then if we have to only calculate the quartile one value Q1. So, we will be doing this. Okay. So, this 0.25 means 25% only. So, this is where we are calculating the quartile one and if we have to calculate the Q3 we will be using this. So, whatever the output we get here is the same way we get the output here. Okay. So, the value of Q1 is 1, Q3 is 2. Okay. the same as the answer to what we have got here. this quantile if we run here that 0.25, 0.50, and 0.75. So, what this means is we are calculating the Q1, Q2, and Q3. and if we run this the same value we have got. So, if we want to know that. Okay. what exactly this quantile does. Okay. So, we will again use this question mark sign. So, we will see what exactly it calculates. So, this function is a generic function. All details are given here. So, this is how you can calculate the quartile in a dataset. Now we will discuss decile(). So, decile() is basically dividing the dataset in 10 equal parts. So, here when we call the function quantile(). We have assigned the value to probes as 0.25, 0.50, and 0.75 for calculating the Q1, Q2, Q3. What will be the value of this when we will be calculating the design. Okay. So, this is given in that if you see the help it is given there all the details about this particular function but just you can think of then when we are using 0.25 0.50 0.75 for Q1 Q2 Q3 what could be for design because what exactly does it divides 10 equal parts. Okay. So, it will be 0.1, 0.2, 0.3. So, you can write in two ways in R. One way is a longer way. That is like the way we have written this. we will see another way using some functions that are there. So, we will see first by writing this from 0.1 to 0.9 .and then we will see the shortcut way of writing this. Okay. you. So, some errors came. what that error is. So, the error is this. they have put an extra comma. What error is there ? The error is that digits >= 1 is not true. Okay. What the issue is. we have to do like this because our syntax is this now if I run this. Okay. we got the results but see in this way there sometimes the problem like the way I got the error it can come. Okay. So, also try to use many of the functions that we have in R by default. So, to do this what exactly we are doing here is we are saying that this is a sequence. So, the sequence from 0.1 to 0.9. and there is a common difference is only 0.1. Okay. So, we see each number is increasing by 0.1. So that is what we are seeing: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%. Okay. So, in a shorter way, what we can do is we can use the function called seq(). So, this sequence function generates these numbers by defining what would be the common difference you want. Okay. So, here if we do like this. So, if I do this sequence, that is only. So, seq(0.1, 0.9, by = 0.1). What it says is that it creates a sequence from 0.1 to 0.9 with a common difference of 0.1. Okay. if I do here 0.2. it will create a sequence of. Let's see how it creates the sequence. So, it will be 20, 20, 20. Okay. So, now if I run this it's 10, 30, 50, 70, 90. Okay. It's not like 10, 20, 30. Okay. So, this is how you can play with the data. Okay, how do you want to calculate what exactly it is like whether you want to divide the dataset in 10 parts or like 100 parts or the 4 parts. Okay. So, in a similar way we will be calculating the percentile. So, we will be doing this in the same way. So, when we are calculating the percentile we are dividing the data in 100 parts. So, our first it will start from 0.01 and it will end at 0.99. So, we are saying that we create a sequence from 0.01 to 0.99 with a common difference of 0.01. Okay. So, if I do like this, my sequence is created. now if I run this my 100 percentile. This is how you can calculate either the quartile or the decile or the percentile. Okay. So, there is another thing that we used to calculate is that inter quartile range is exactly what inter quartile range is. So, the inter quartile range is the difference between quartile 3 and quartile 1. So, we will be subtracting the Q1 from the Q3. Okay. So, this is our Q1 and we will be subtracting into the Q3. Q1 we have already calculated. So, our Q1 is and and our Q3 is this. if I run this. now if I do 2, 3. So, the value is 1. Okay.

So, now you think of any of the inbuilt functions which calculate the inter quartile range. Okay. So, earlier when we used to calculate either maximum value, minimum value or the sum we used to have this function max() then we had min(). we have seen min() also we have also seen SD() in small for standard deviation. So, these were like one word maximum, minimum, and for standard deviation we have sd() in lowercase. Okay. So, what could be the possible function name for inter quartile range inter quartile range. Okay. So, there is a common pattern where in the function name what we are seeing is that when there is a one word there we have the function name by the some of the letters of that word only like max mean sum or summing up the number. But when we have a standard deviation square root. We have the function name in lowercase characters. But one letter from each of the words like Sd(), then SqRt(). Okay. So, here there are two letters but if we think of it in a similar fashion we could think of one possible function name is iqr(). Okay. whole in small. Okay. So, let's see if there is any such function with iqr(). So, iqr() automatically shows that there is no such function with a smaller case as iqr(). So, this IQR() in uppercase is the right function name. Okay. So, then we don't have to calculate Q3, and Q1. if we are interested to know only inter quartile range. We can simply call this once an IQR() on our dataset. and it will give the value. Okay. So, the IQR of this is one and our earlier by when we will calculate by formula it was one. So, this is how we can calculate the inter quartile range. Whenever you are using different functions you have to take care of these uppercase and lowercase. Okay. So, not all functions are in lowercase, not all functions are in uppercase. So, mixed practice is there. Okay. there are functions there. So, you have to just take care of that. Okay. when you will be doing the practice will come to know that. Okay. This particular function is in uppercase for which we used to calculate the inter quartile range. Okay.

Now let us move to other two concepts like skewness and kurtosis in our dataset. So, basically how much our dataset is skewed or not. Okay. So, for that we used to calculate the skewness. So, to calculate the skewness and kurtosis in R we have to use some packages. So, for this kind of analysis we will call the package called moments. So, first if you have not installed that package. Okay. So, the first thing is that you install the package and we have already learned how we can install the package. So, there is either a call through command install.packages(). or we can go through tools to install the package. or we can also locally install the package. Okay. So, we will call this if you haven't installed this moment package. You just simply install this moment package by using this command. So, I'll just run this. now you can call this by this library. Okay. So, if I do like this. our package is done. Now this is our dataset. Okay. now if I have to check the skewness. Okay. So, if I do like this. Okay. So, this is the value of the skewness in the dataset. So, in a similar way I can calculate the kurtosis. So, this is the value of kurtosis. So, this is how you can calculate the skewness and kurtosis of your dataset. Okay.

So, until now whatever we have discussed is only related to one variable. Okay. So, what we have done until now is on the publication side. So, another variable can be the citation. Maybe we want to know how these number of publications and number of citations are related. Okay, is there any relationship between the number of publications there are and number of citations in a year. So, let's do that. Okay. So, this was our dataset. until now we have discussed this only single variable now this is our second variable citation. Okay. for. So, in 2023 we have got one paper which has got one citation. So, in 2022 we have five publications which have received citations 2, 3 or 2, 1 and one paper doesn't receive the citation. In 2021 we have received 2 citations to 1 paper. In 2019 there is one paper where we have received a single citation. Okay. So, now we will include this second variable also. Okay. So, our dataset was this. We have already imported and we have given it the name. Okay. So, this is the paper. Okay. and these are the years and these are the cited by now the column names. We will just see what title year, cited by. So, now what we will be doing is we want to know how much total number of citations in a year we got. Okay. like the way we have counted the number of publications in a year. Okay. for what function we have used we have used the table function. Okay. But for counting the citation we have to use this function called aggregate(). So, what aggregate does is aggregate if we do like this. So, in 2021 like there is in 2019 there was one publication it was cited by only one paper. and then in 2021 we have two citations. and in 2022 we though we have five publications. But we have these citations, some are 8. and in 2023 we have 1. Okay. So, this is our citation data. So, if I see here this and our publication data we already have frequency. So, this was our number of publications in the year 1151. So, we will add this extra column to our site where we have citations also. So, we will have the number of publications. So, we will do this. So, if I now check this. So, this is what we have the output is. what I have done is I have converted this dataset into a format like this, year 2019, 2021, 2022, 2023. So, the number of publications in 2019 is 1. in 2021 there is one. then in 2022, there are 5. then in 2023 there is 1. Okay. Now in 2019 there is only one citation. then in 2021, 2. then in 2022 we have this many 5 papers. So, 2+3, 5, 5+2, 7 and 8. and then in 2023 we have 1. Okay. So, these are. So, this is our dataset. and Now we will be doing the analysis. and this I have done in R. Okay. So, our dataset is this. Okay. So, when we were doing analysis on a single variable we used to do it for only this $\bar{x}$ and then we put xi-$\bar{x}$ and then we did all the analysis. So, here what we have to do is first we have to calculate the $\bar{x}$ of this then $\bar{y}$ of this then x_i-$\bar{x}$ and then y_i-$\bar{y}$. Okay. because we have to do the analysis on two variables. So, how will we do that? we will calculate the x_i-$\bar{x}$ of one variable. and in the similar way you can calculate that thing and I will request if you are a beginner you must first calculate manually on your paper. and then you go to the R. Okay. So, we have the mean() of the publications. So, this is two we have already calculated. So, this is the mean of our publications. now mean() of citations. So, we will be like this. So, if we put the $, it automatically shows me this. Okay. So, I will calculate this. So, the mean of the citations are 3. So, every citation we have received 3. Okay. I will assign this to an object mc now that is x_i-$\bar{x}$. Okay. So, I will calculate this. So, this is our mp1. I will just write it down as a comment. So, this is our x_i-$\bar{x}$. Okay. So, this small x is our value and then this $\bar{x}$ mean(). Okay. So, this will be helpful for you. I will be sharing this whole code in the github repository. So, you can also do the analysis. then this is our y_i-$\bar{y}$. Okay. then we will calculate for this. Okay. then we also have the number of observations. So, a number of observations will see how it is like for. So, we will assign to another object mn(). So, here you have to be very careful. Okay. for this bracket and all. Okay. So, what we are doing we are calculating this for covariance. We have a function called cov(). So, if I do like this. So, it's 6.66. this value is there. I'll assign to this and if I now check the value. So, again if you see. If I do like this. either I will put first frequency and then citations. or maybe first citation and then publication frequency. Okay. So, the covariance is the same. Okay. So, the covariance between x and y, or y and x will be the same. Okay. So, this is a conceptual thing. I think you might know those things. But we want to know how much the correlation is there between that. Okay. We have the functions to calculate the correlation also. So, first we will be calculating manually. and in R we will not be using the inbuilt function. So, we will be calculating the standard deviation of citation. So, we will be using this function. and this is our standard deviation of citations. Now this is our SdPc(). what SdPc is, SdPc() is covariance. Okay. This is our SdFq. We already have standard deviation of our frequency and sdc() is just now we have calculated standard deviation of citation. So, I run this. So, this is our correlation. now as we have the function for every of the operations. We have a function called cor() for calculating the correlation. So, if we do like this correlation, cor(). So, we got the same result. Okay. So, this is all about descriptive analytics.

So, in the last three lectures I have discussed how we can do the descriptive analysis in R by creating our own functions or by step by step operations or by using the inbuilt functions. Okay. We have also seen that we can use the packages also. There are some packages that are there for specific tasks. So, like what we have called the package called moments for calculating the skewness and kurtosis. So, now I request you all to as a practice you try on this practice dataset, which I will be providing on the Github repository you download this data, and do all the analysis whatever we have discussed in these three lectures. Okay. So, you just calculate the frequency of publications in each year, which year has the maximum number of publications or the citations. Then you also calculate the mean, median, variance of publications and citations. Whether there is any positive correlation there, or negative correlation there. or there is no correlation between publications and citations. Please let us know through a discussion forum on this practice dataset. So, in the next lecture I will be discussing how we can use R for others and like theoretical law like Bradford’s law. Okay. So, I will discuss some sample dataset and I will also share that dataset with you for your practice. So, see you in the next lecture. Thank you.