Research for Marketing Decisions

Vaibhav Chawla

Department of Management Studies

Indian Institute of Technology Madras

Week - 07

Lecture - 32

Data Analysis using SPSS: Data Cleaning and Descriptive Statistics

Second one is, find the minimum and maximum value for each variable in this study let's do that, so we have this, how to do the second operation, second exercise. Find the minimum and maximum value, go to analyze analyze left click you have descriptive statistics click on descriptives, and for all the variables whether they are nominal, ordinal, or whatever, except the respondent number you can.... so this is the box operations box where statistical techniques will be the variables on which statistical techniques have to be applied will be in will be entered so we are moving the variables on which we want to do the range the test of identifying minimum and maximum value, we are moving them into variable box. Go to the options and what we want minimum and maximum. We at this point of time mean and standard deviation are also ticked automatically default, let them be there, and default settings don't touch click continue and click okay, you will get minimum and maximum in the output screen. Are you getting? analyze descriptive statistics descriptives. Let's do it again. Reset, I did it, and then all the variables you can you need not do it one by one, you can click on one of the variables then press shift and then put the down arrow so you would be able to choose everything at once, move them to variable box. Go to options, you have minimum maximum already ticked, click continue, and click ok. So this is the first output what is the output? what is the minimum maximum for each variable tell me quickly?

Is there any problem here? Where is the problem? For the nominal What is the problem? Yes, for nominal does not make any sense.

Correct. Any other problem? Is that a problem here? What is the problem? Next problem.

See, if you look at the familiarity with the internet, minimum is 2, maximum is 9, whereas we measured it on 7 point scale. So we identified that there is some problem in this variable in data entry or data collection. Familiarity with the internet how can, on a 7 point scale, somebody could give the score of 9. So So, we know now there is, and for likewise, for each variable for nominal, it does not make sense, but you know that minimum maximum ok 1 and 2 there is nothing else than that.

It does not make sense, but it has to be done to know whether there is some problem in any of the variable whether there is some missed entry of data somewhere. So our second exercise is also, second point is also complete. Find the minimum and maximum value for each variable. And now we know there is a problem. Third one is fix the missing values.

There are missing values also, right? When you did the data entry, did you find missing values? Can we see missing values? There is one here, and one here, but let's first go to the second point itself, because one of the point familiarity with the internet somewhere there was a problem because maximum score of 9 where is it is where is it? respondent number 10, we have identified because it 30 respondent, suppose 30,000 what you will do?

This is 30,000 what you will do? The SPSS can take 55,000 data points. So this is 30 respondents we have identified. 30,000 it will take entire day and you will go mad. This data will be corrected or not there will be some problem.

You have to somehow tell them to highlight those points which are possible. There is no voice bot. We can't tell them. We can't tell the software but it is a good yes if somebody can do that. Understand our commands. So here then what you have to do is, we will do one operation with the help of which we will get to know, go to transform, compute, and you will get a box like this. so we know that familiarity with the internet there was some problem because the maximum value was nine out of a scale point of seven, now how to find out where it is? we will create

another variable let's say, familiarity with the internet underscore one, and we will in this numeric expression box we will try to transform the familiarity with the internet variable in such a way that we know which respondent is the problem, so what is the maximum value scale point, seven, seven minus familiarity with the internet, which means for every respondent we are creating a variable familiarity with the internet underscore one, in such a way that it it it gives us seven minus the value that it actually has, so what will happen is shall I do it? redo it, okay, so transform this is for correcting the data entry problem, where the minimum and maximum there is some problem in one of the variables, we have to go to transform compute variables, suppose the data points are 30000, then we have to do this go to or 1000, 2000, 3000, so which is the variable familiarity with the internet, let's create a new variable familiarity with the internet underscore 1, and in the numeric expression box all the operations have to be done inside the brackets, so first we will introduce the brackets and then what is the maximum scale point 7, Why we are doing putting 7? Because 7 scale point and we will subtract all the values for we will subtract the value of the respondents and we will create a new variable familiarity with the internet underscore 1.

So, when we do that what will happen is, if you see this variable is created. Now wherever you see minus sign which is easy to identify, because 7 is the maximum scale point, maximum answer can be 7, maximum here 7 minus whatever you do will be there cannot be a negative answer, because 7 is the highest scale point, so wherever you see negative sign you can highlight that saying that there is a problem. This is one way to identify by looking at negative sign which is little easily identifiable than the earlier. So, this is one of the ways, yeah, that is correct, but that this what I this is one of the ways find out other ways on your own. I have you know not worked with such big data.

So, that, but whatever techniques are simple and popular, I can tell you here this is one of the way where you can do this, transform the variable so that at least you are able to see the sign, if it is negative, where there is the problem right? filter might be there, one has to explore but yeah, some of the things it is left on students to explore and that's what what we will do next now, next is once you identify this is a problem go to the data sheet if it is physical, if it is internet survey, you internet survey you can download in SPSS format between the same data format or Excel format from Excel you can export or you can from Excel you can export it to SPSS you will get the same way so from SPSS data sheet you can convert it to excel from excel y,ou can convert it to SPSS there is an option to import the data in the excel sheet into SPSS that is also there so so once you have data from the internet web survey they'll such problems would not be there because they are automatic there cannot be an answer option more than seven points if it is seven-point scale but if it is of face to face survey and

Then they'll some some some this kind of problems would come. Now what to do in this case go to the data sheet survey sheet if it is a paper and pen survey and find out what is exactly the answer ticked. It is a more likely to be a problem of entering the data manually. So here once you go to that data sheet of the respondent number 10, you would be able to see where is a problem, and you would be able to correct that data point. Here let's say it is 7, so we can correct it to 7 which means we went back to the data sheet the survey sheet and we looked at this particular respondent sheet and his or her answer to the familiarity with internet we found it is 7 not 9, and then we can correct it, so once we correct it our job is done,

and we can delete this variable familiarity with the internet underscore 1 because this was created to identify the wrong data entry. Now there are missing values as well. There are missing values if you look at the third point fix the missing values. There are missing values again this is a small data set. If it is a large data set how to identify the missing value?

so how to do that is one way to identify the missing values. if you go to analyze and go to descriptive statistics go to frequencies, take them all to variable box and click ok, so you will be able to identify missing here that do you see this table? Can you find out, can you do this table on your screens as well? Go to analyze descriptives and frequencies and move all the variables into the operations box and you will be able to get this as an output and in that you will be able to see which has the missing values. One missing value is there in attitude towards the internet and one is there in gender. Now the exercise number three is fix the missing values, not just identify.

Identifying is the first step. Now we have to fix the missing value as well. Now this is 30, we can easily see what is the other way of identifying where is the missing value from the frequency table output, we will know which variable has how many missing values so attitude towards internet has one missing value gender has one missing value but we don't know which respondent suppose 3000 then what we will do is we will go to transform again go to Compute variable. Reset it.

Let's say attitude towards internet is one of the variable where missing values. Let's create another variable ATI from ATI underscore 1. Now what you have to do is go to the function group and click on all. This function group all is there. Click on all.

And then down there is one table. Here click on M where missing will come. And double left click. You will get something in the expression box. So let me redo it.

So you have to go to transform compute variable for ATI underscore ATI you want to identify missing values. Let's make a new variable ATI underscore one and first go to function group click on all and in the box below click on M so that you get the missing double click on that and you will get a numeric expression missing and in the brackets question mark, then move attitude towards internet in that question mark, and then click ok you will get a new variable ATI underscore one, and where the missing value wherever the missing value is you will get one otherwise zero. So you got it, which is the respondent number 17. Now one way to fix this missing value is, go to the survey and see whether the answer is field is it a data entry problem, if let's say, the answer is left blank, then how to fix this missing values? the most popular method is series mean mean of the all the respondents on that particular variable and substitute it there, so that how to do it, go to transform, replace missing values, for what you want to replace missing values? attitude towards internet, take it there and we want to take the series mean, which means mean of all the

all the respondent values will substitute there and click OK then. Name we can keep the name now ATA underscore 2. Then it is giving some value which doesn't makes sense that's what the original, just just a minute so ah... we want.... so actually it is creating a new variable with this name ATI underscore one, so once we know that ATI underscore one, which particular variable has missing we can actually delete this particular variable, or even you keep it.

There will be no problem. Just go to transform. Replace missing values. We want to replace for attitude towards internet. Create ATI underscore 2,

and if you click OK, just a minute.... Yeah, tell me. Yeah, that you can change actually.

That you can change. You can actually change it. It is scale only. So, when we are substituting it with series mean it creating a new variable ATI underscore 1. So, better to change the name here then better we will change the name here so that

let us write here it is ATI underscore 3, and then go to data view let us fix the missing value transform, replace, missing value for attitude toward internet, replace with the series mean, and the new variable will be created ATI underscore 1, which will have that missing value fixed and missing value replaced with the series mean, you click ok you will be able to get ATI underscore 1 new variable with the 5.17. 5.17 as the series mean. Now one way if you want to keep it as discrete like others 1, 2, 3, 4, 5 then round it off. Round it off to the nearest integer.

If it is 5.5 do it 6. If 5 less than 5.5 do it 5. You can keep it 5, and save it. So instead of doing it here, wherever missing value is there, you can keep it 5 here. And delete all other variables that we created.

So missing value is fixed here, right? Same way can we do it for gender? So we go to transform, replace missing values, reset for gender.

Gender 1, series mean, okay. So, we will get 1.52 which is greater than 1.5. We will make it 2. So, what is the mean here? So, mean of what?

Mean of the 1 and 2, the respondents values on gender. Is it right? Done? Done? Done?

Everybody? Done? Last? This was a mistake. It is a nominal variable.

Can we take the mean? Then why did you do that? If I am doing anything, that's what I told you to think and then do it. He thought about it. He told me, sir, we are taking mean for a nominal variable.

So here which means if it is a nominal variable we cannot take a mean and decide whether it will be male or female, right? So here we should not do that operation which means the one which we did G underscore 1 should not be done on nominal variables. So you should pay attention in the class. Think about whether what I am doing is right or wrong? I will do these things many times just to find out whether you are thinking.

Gender, how you will determine with their responses. What you are trying to do is whatever their responses are, you are trying to determine gender by taking nearby points. So here what you will do is, you will go back to the data sheet and find out whether the gender response was filled. If it is not filled, name is written. Can you make a guess from the name what will be the gender?

Sometime you can make, right? Or if the phone number is there, call back. Call back and just, when you call back, if there is a female there, just don't keep the call, assuming that female. Somebody else could be picking up call, ask, is it your name, is it the correct person? Okay, you gave the response, but one of the responses was unfilled.

Gender was there, so now we know the answer, thank you. But if the name is, you know, if just by listening to some male or female, by giving a, you know, what is that call? It is not a missed call. What is that call? Blank call.

Blank call. Do not assume that whoever is picking is the right person who filled the survey. Ask the name and then. So in this case here in this case we have kept equal actually 15 female 15 male.

So, whatever is missing is either 1 or 2. So, we will just look at the frequency that we frequency table for the gender yeah. So, male were 14 and female 15. So, male 1. So, we will substitute it 1.

So, that our data set is complete. So, we have now this is called data cleaning which is not very easy here I made it look easy. See once you these are generally less than 10 percent of the missing values In your data, total data, 10% of the missing values are acceptable. But you won't have that many.

You would have 1 to 2%. Then you will have to fix those. If you have a small sample, you cannot remove them. You have to fix them. And if you remove some of the respondents, you might create bias.

Because you are removing whatever sample size you calculated, you are removing them. If let's say, for a particular respondent many values are missing, then you can remove that particular respondent, if only one or two, try to substitute them by mean series mean, or if it is a multi-item scale, listen to me, if it is a multi-item scale, instead of taking series mean, take the mean of that respondent. Let's say five items are there.

In the third item, it is missing. What you can do is for the same person, you can take the mean of the answer on the four items and substitute it rather than taking series mean. That is better. If it is a multi-item scale, you are measuring, let's say, attitude towards internet on five statements. Answer for that particular respondent on four statements is given.

Fifth is not given, take the average of the four respondents score for that particular respondent and substitute it. Rounding it off to nearest integer. That is the way for multi-item scale.

Whereas for single item you can take the series mean. No. Once we are taking series mean. Yes multi item scale. You remember the item should be reliable,

consistent with each other, Correlated with each other, that is why we substitute it. So we will end the class here. And we will meet in the next class.