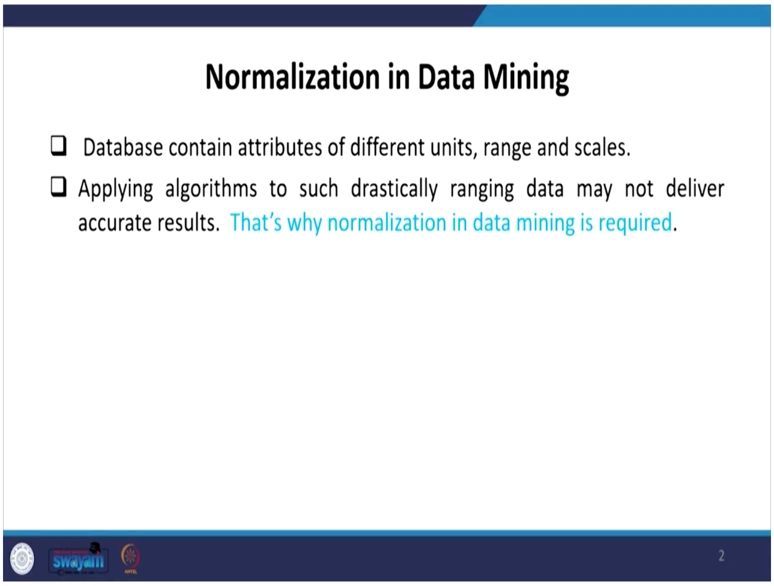
Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 13
Normalizing Data

Welcome friends once again to the NPTEL module on handling healthcare survey data. We are on the verge of 3rd week where we have been trying to understand, how to normalize health care data. Especially we wanted to address the specific practical sessions within the lecture where you guys will be quite handy in understanding or manipulating data as per the use.

My name is Dr. Pratap Mohanty I am attached with the Department of Humanities and Social Sciences at IIT Roorkee. This I have been handling for quite a couple of years. So, I can deal with all your doubts related to normalizing data. So, without further discussion let us start with the meaning of it.

(Refer Slide Time: 01:22)

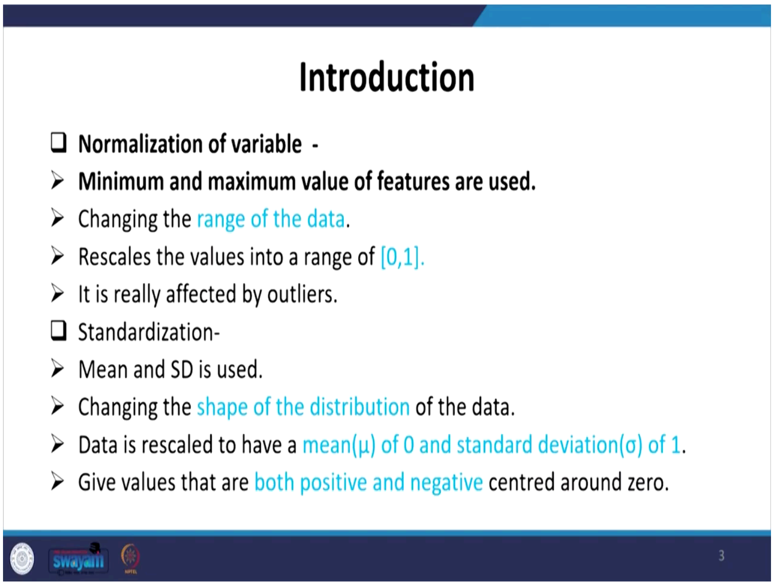


Regarding normalization in data mining, the first one we wanted to highlight is that it contains certain attributes such as different units, ranges, and scales. what does this mean? We often deal with various data sets in our day-to-day life. Where each of the variables is coming with its different units, their range is defined differently.

I have already clarified what you mean by range, scale, and measurement in my previous lectures. Now applying an algorithm to such drastically ranging data may not deliver accurate results, that is why normalization in data mining is required.

So, in the very introduction, we wanted to mention here that we will be discussing the normalization of variables. It is the maximum-minimum values of the features that are most often used like changing the range of the data and rescaling the values into a range of 0 and 1. It is really affected by outliers. This is related to the normalizing of the variable. Then the second aspect within the normalization is called standardization.

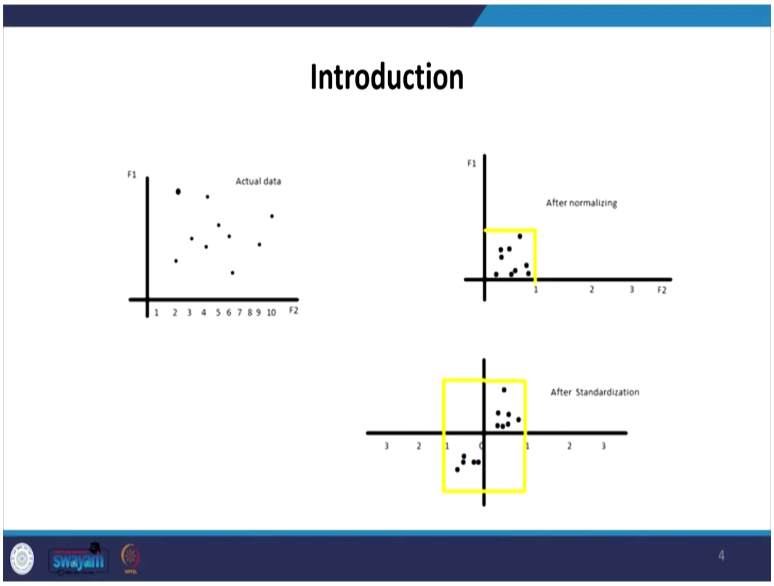
(Refer Slide Time: 02:58)



In standardization, we should use the mean and standard deviation. The standardization changes the shape of the distribution of the data and accordingly which mean or median value is used. Data is rescaled to have a mean value of 0 and a standard deviation of 1.

Usually, in that format, we standardize the data. Likewise, you might have earlier dealt with normal distribution kind of data. This gives values that are both positive and negative centered around 0. So, it is standardizing standard normal distribution.

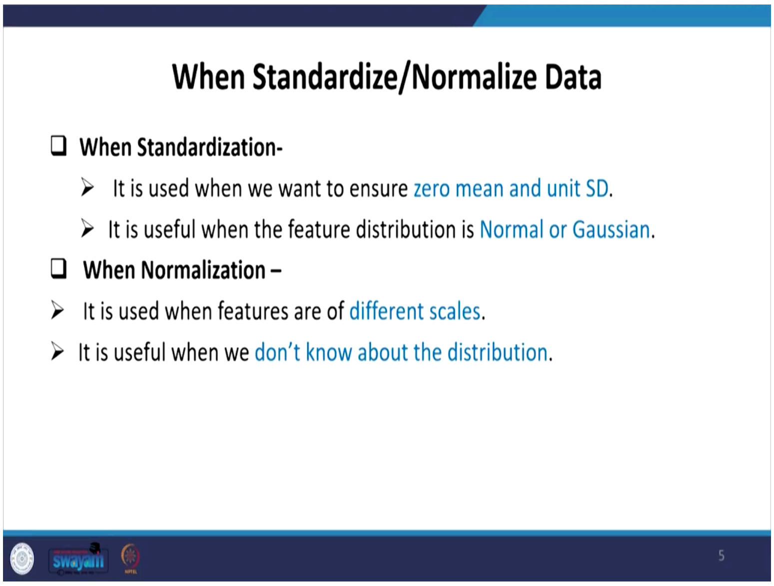
(Refer Slide Time: 04:05)



Here on the very first chart, we have actual data, then on the other chart, we have shown the normalized data. The actual data seem to be very scattered.

But after normalization, it is within a caveat, and it is within the range. We have seen that it should be within a range from minus value to plus value as per the standardization.

(Refer Slide Time: 05:01)

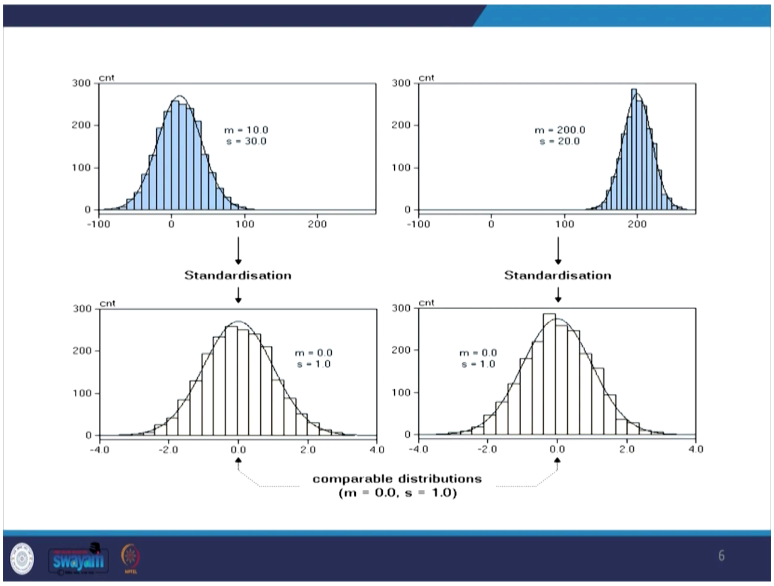


When to standardize or normalize the data? Standardization is used when we want to ensure 0 mean and unit standard deviation. It is useful when the feature distribution is normal or

Gaussian. So, Gaussian distribution is usually referred to in the context of the normal distribution of data.

Then in the case of normalization, it is used when features are of different scales. It is useful especially when we do not know about the distribution. So, when the distribution is skewed and is not well within the range in that case normalization helps.

(Refer Slide Time: 05:50)

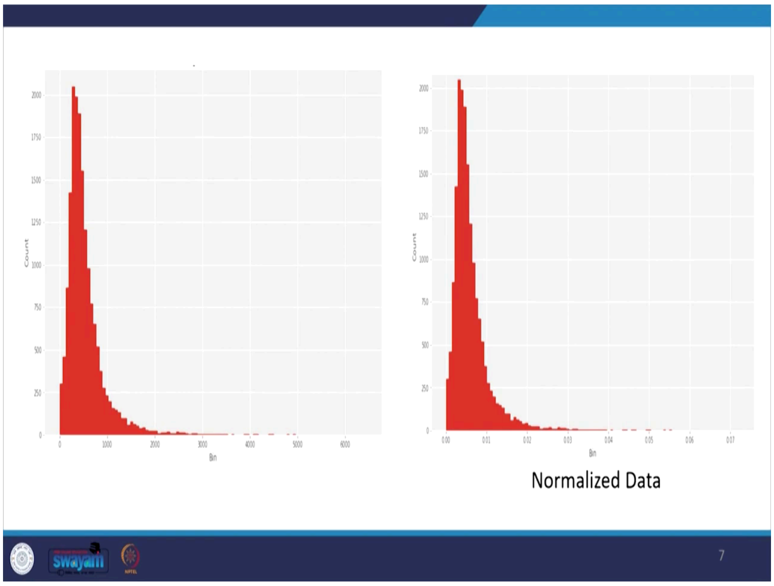


Here are some of the charts that are guiding you, and you can compare the distribution. So, in the first block of distribution, we are discussing standardization. You can just see how it looks.

So, on the first one, it seems the data is more or less confined near 0 and it is normally distributed data on the first scale.

The other chart again is normalized to 0 and it is with a range standard deviation of 1, but initially, it has a different standard deviation and mean value. The mean value initially was 10, in the latter case it was 200 and on the redefine and distribution we get our value min as 0 and standard deviation as 1.

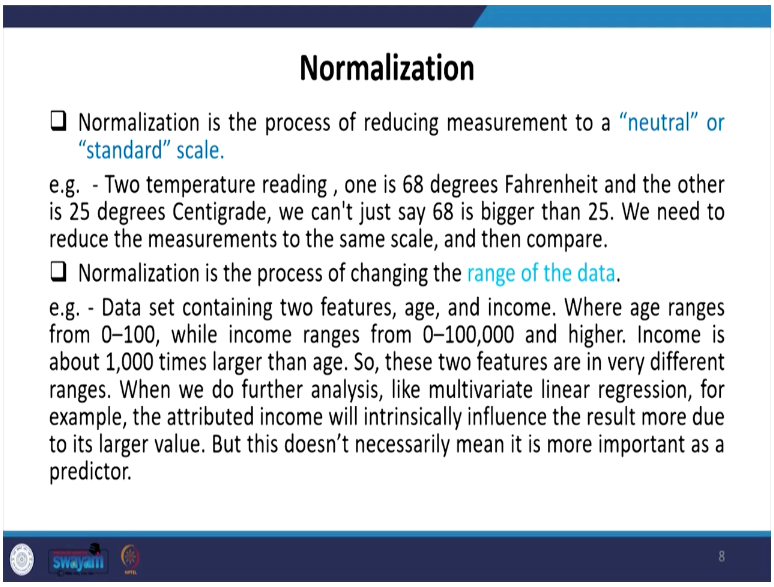
(Refer Slide Time: 07:14)



So, far as normalized data is concerned when it is quite skewed like the range on the first chart it is from 0 to around 5000.

For example, more frequencies are around let it be around 500 but after normalizing the data we can scale down to 0 at maximum till 0.05 or 0.06. But in the first chart, you could easily see that the variation is too huge and in the second one, the variation is very less. Though the distribution may not be properly standardized.

(Refer Slide Time: 08:23)



Normalization is the process of reducing measurement to a neutral or standard scale. For example, two temperature readings, one is 68 degrees Fahrenheit and the other one is 25 degrees centigrade. We cannot just say 68 is bigger than 25 because they are on a different scale. We need to reduce the measurement to the same scale.

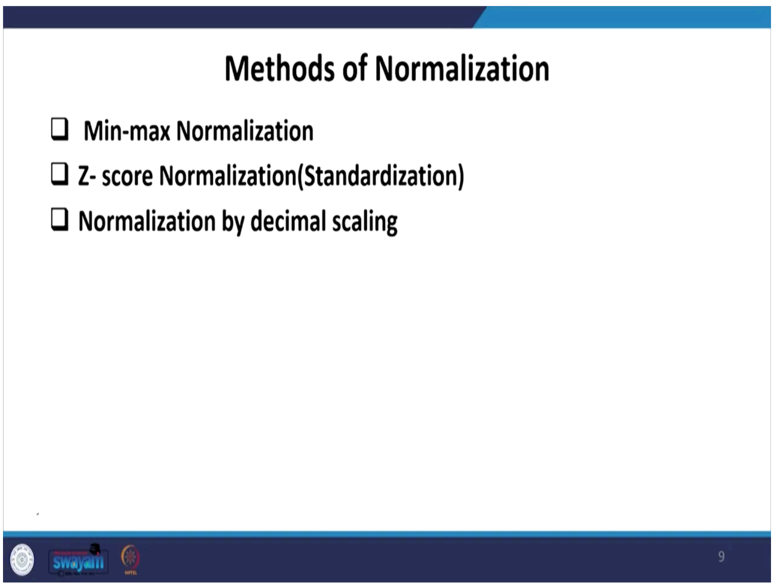
Through the normalization process, we can avoid Fahrenheit or the centigrade scale. It only converses the number and those numbers are comparable. Normalization is the process of changing the range of the data.

For example, the data set contains two features age and income. Where age ranges from 0 to 100 while income ranges from 0 to 100000 and higher. Income is about 1000 times larger than age. So, these two features are in very different ranges.

When we do further analysis like multivariate linear regression models we have to scale down. Otherwise, the implications of one variable may have higher loading on the final regression coefficient. When we do further analysis as I already say multivariate regression the attributed income will intrinsically influence the result more due to its large value. But this does not necessarily mean it is more important as a predictor.

So, as a researcher, I will strictly suggest that you start with normalizing your data. Even if they are on the same scale, it is free from their unit of a scale.

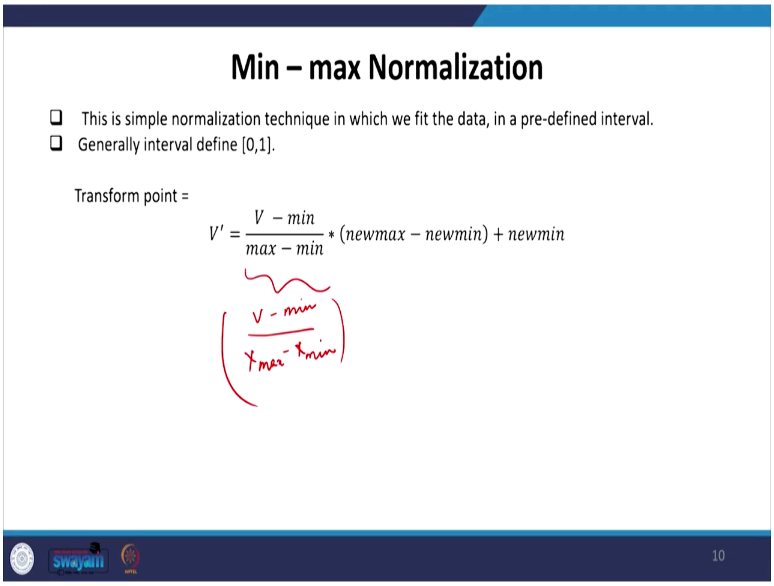
(Refer Slide Time: 10:40)



We usually consider three important methods of normalization one is min-max normalization second one is Z-score normalization usually called standardization and the third one is called normalization by decimal scaling. We have all those practical sessions at the end. I am now clarifying all the concepts to you and once you understand and set the tuning of these details then I am sure you will go into the depth of this with the practical sessions at the end of this particular lecture.

Starting with min and max normalization. This is a simple technique of data mining or data filtration, or data simplifications. This is a simple normalization technique in which we feed the data in a predefined interval.

(Refer Slide Time: 11:41)



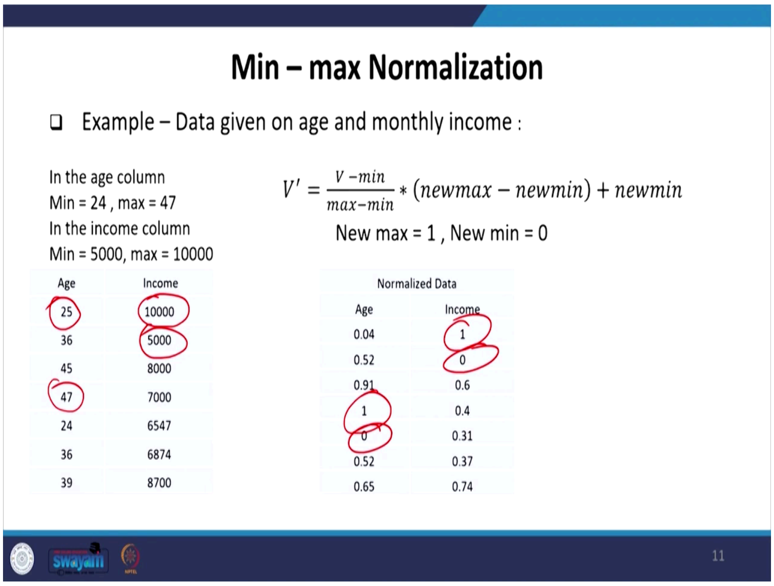
So, generally, the interval is defined with 0 and 1. So, we need to transform with this formula this is the standard formula for the min-max normalization. So, this one is basically:

$$V' = \frac{V \;- min}{max - min} * (newmax - newmin) + newmin$$

So, usually, some of the researchers only follow this part:

$$V' = \frac{V - min}{max - min} V' = \frac{V - min}{max - min}$$

(Refer Slide Time: 12:50)

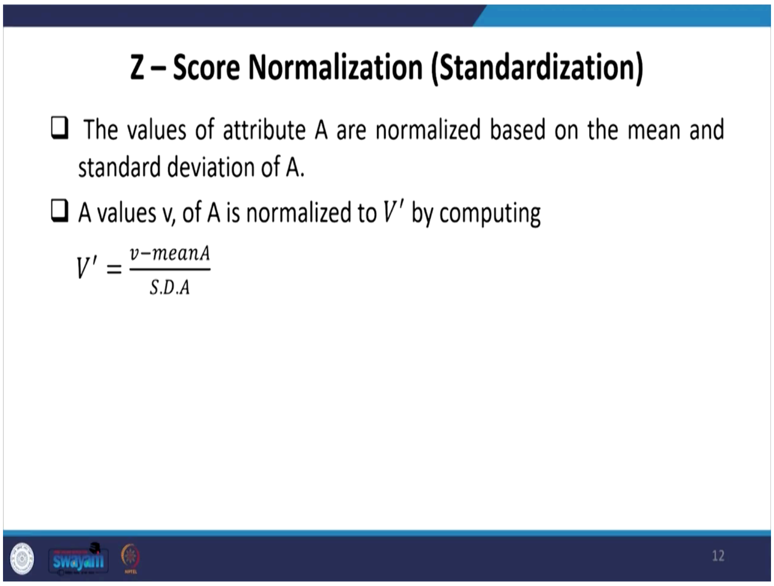


So, we will experiment and show it to you with the help of stata. The example of age and monthly income is taken. In the age, column minimum value is equal to 25 and the maximum value is 47 which you can easily see.

The second one is related to the income, and we have a maximum equal to 10000 and a minimum of 5000. So, after getting the value we can normalize the data with the help of this formula.

This gives us the new maximum value and new minimum value and with that new maximum value here again just with the first formula, we can get the new maximum and new minimum value. Here on the age scale, this is 0 and 1. So, the new maximum and minimum values can be derived based on the formula.

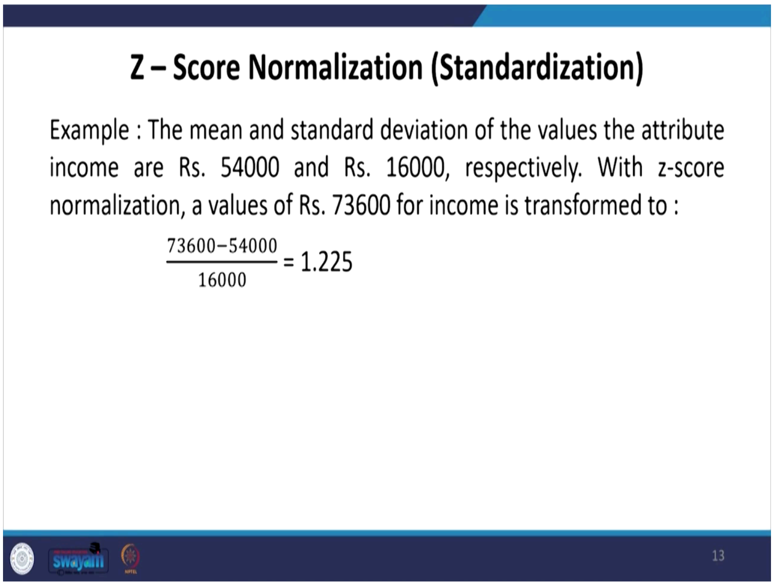
(Refer Slide Time: 14:24)



Now, coming to the Z score normalization or standardization. The values of attribute A are normalized based on the mean and standard deviation of A. So, values of V of A are normalized to V prime by computing.

$$V' = \frac{v-meanA}{S.D.A} V' = \frac{v-meanA}{S.D.A}$$

(Refer Slide Time: 15:00)



So, here we have cited one example for Z-score normalization is also called standard normal distribution or called standardization. So, you might be confused about standardized coefficient and non-standardized coefficient, which we will also discuss during our regression analysis during our multivariate analysis. The standardized β or non-standardized β is computing during regression analysis. Anyway, I will clarify those things later.

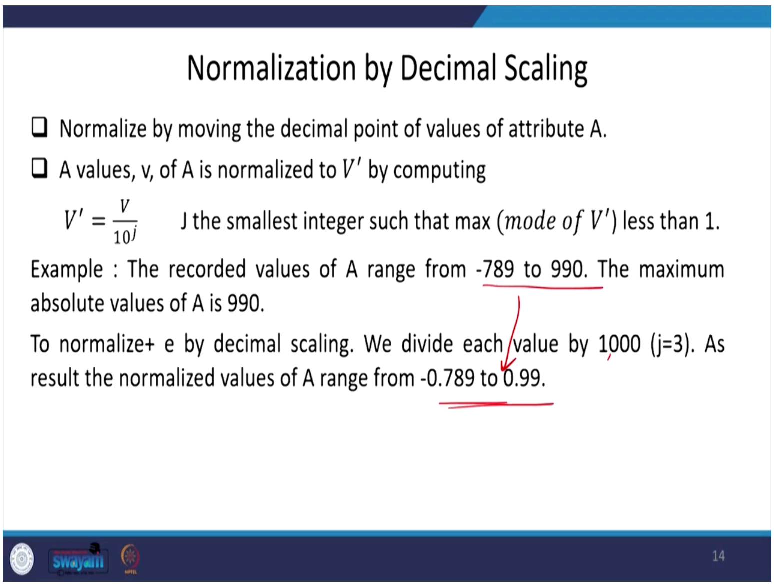
The mean and standard deviation of the values of the attributed income is rupees 54000 and 16000 respectively. So, that is the mean and standard deviation with Z-score we can just put this value. With Z-score normalization, the value of 73600 for income is transferred to 1.225.

So, the third approach we would be following is our normalized by moving the decimal point of values of attribute A. So, A is normalized to V prime by computing:

$$V' = \frac{V}{10^j}$$

So, j is the smallest integer such that the maximum less than 1 would be considered.

(Refer Slide Time: 17:19)

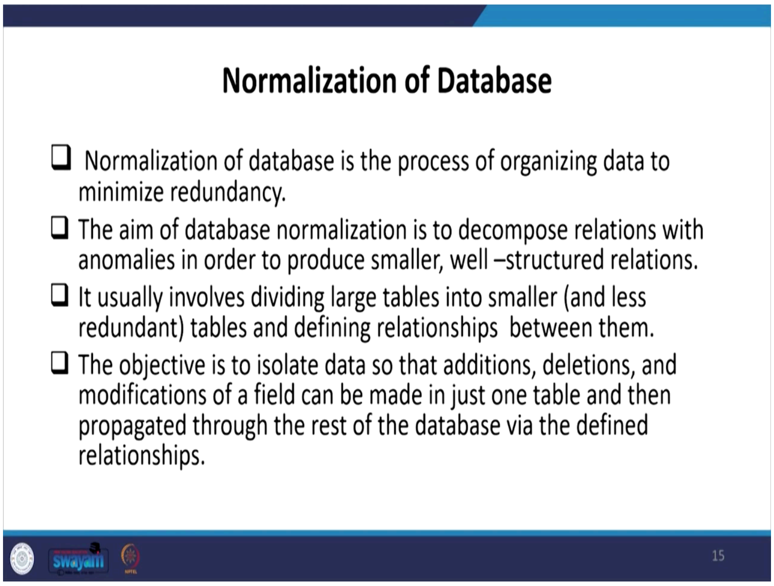


For example, the recorded values of A range from minus 789 to 990. The maximum absolute value of A is 990. To normalize by decimal scaling. We divide each value by 1000. So, since it is 10 to the power 3 times. As a result, the normalized value of A ranges from 0.789 to 0. 99.

So, it depends upon how you have normalized to the decimal scaling. I think if you remember I have discussed the weight of the variable in NFHS. We divided the weight variable and scaled it down to a normal range.

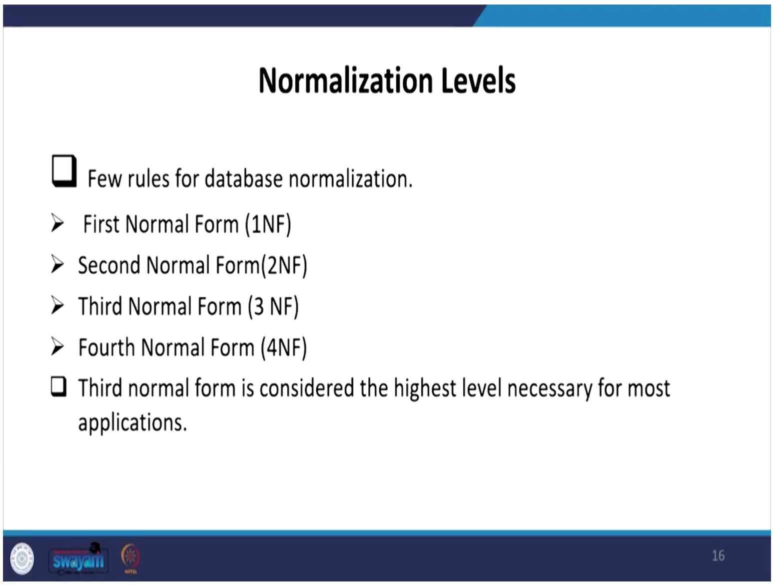
So, the weight in NSS is different than NFHS and it has to be on a ratio scale, rather than an absolute number. So, now I am discussing the normalization of the database. How can we go about normalizing the database? Normalization of a database is the process of organizing data to minimize redundancy. The aim of database normalization is to decompose relations with anomalies to produce smaller well structure relations.

(Refer Slide Time: 19:32)



It usually involves dividing large tables into smaller (less redundant tables) and defining relationships between them. The objective is to isolate data so that additions, deletions, and modifications of a field can be made in just one table and then propagated through the rest of the database via the defined relationship.

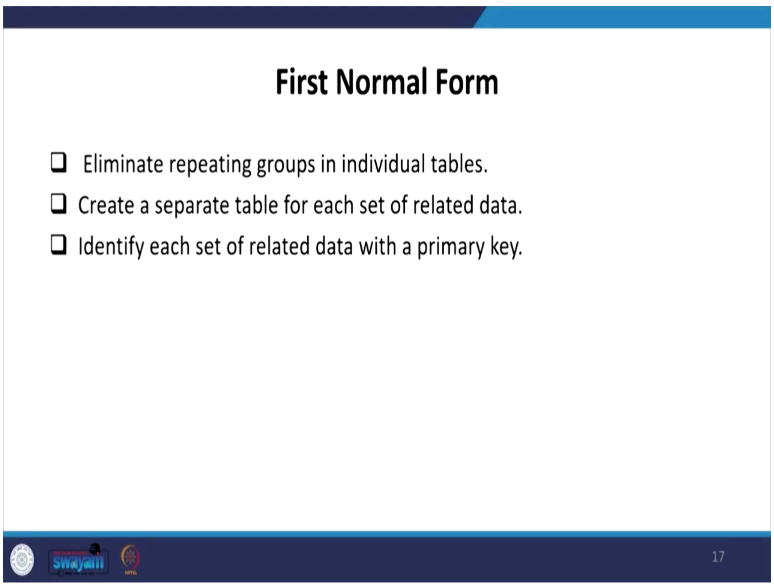
(Refer Slide Time: 20:03)



So, this is the meaning of normalization. So, there are different normalization levels of database normalization. The normalization forms like the first normalization form, second normalization form, third normalization form, and fourth normalization form. The third normal form is considered the highest level necessary for most applications.

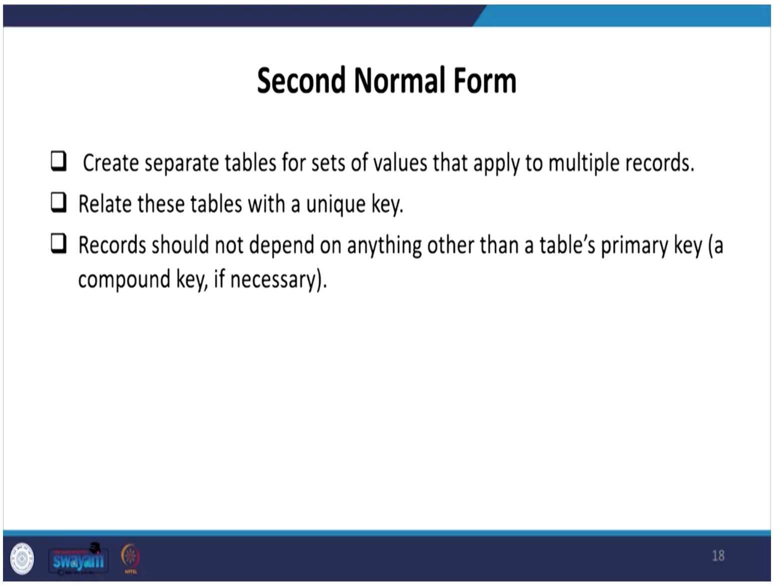
In the first normal form, we will eliminate repeating groups in individual tables. Any repeating groups are entered in individual tables, we will avoid or eliminate those. Create a separate table for each set of related data.

(Refer Slide Time: 20:47)



We identify each set of related data with a primary key, and with that, we can avoid repeated entries. So, though this is the first form and we can create separate tables for each set of filter data and accordingly identify which are repeated groups.

(Refer Slide Time: 21:11)

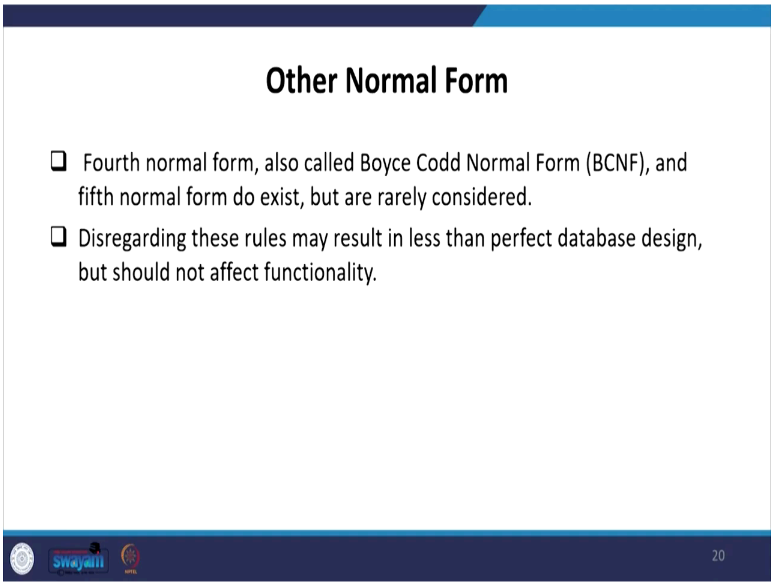


On the second normal form, we will create separate tables for sets of values that apply to multiple records. We will relate these tables with a unique key and with a unique key we can separate them. The record should not defend on anything other than a table's primary key.

The third form of normalization is the most useful form. We need to eliminate fields that do not depend on the key. If beyond the key some fields are not dependent on the key, then we need to eliminate those fields. The third normal form prohibits transitive dependencies, and a transitive dependency exists when any attribute in a table is dependent on any other non-key attribute in that table.

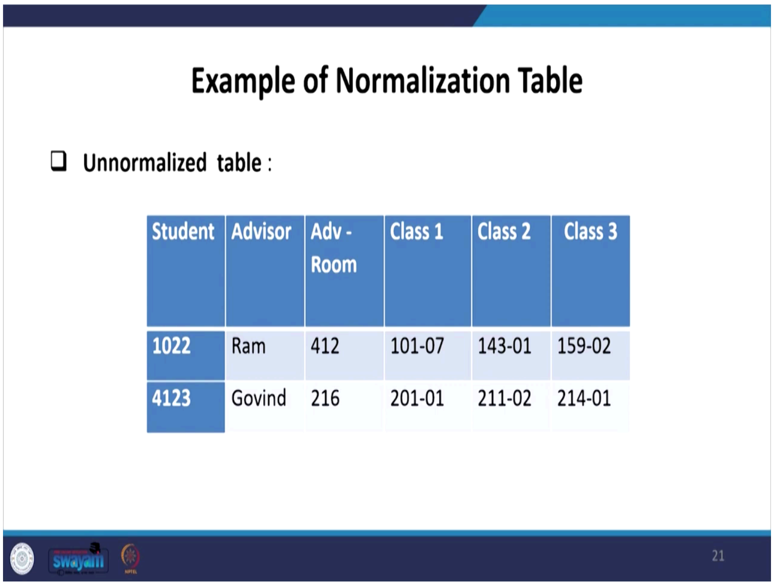
There might be collinearity issues in the data set and they should be avoided. The other normal form is the 4^{th} normal form, also called Boyce Codd normal form (BCNF) and the 5th normal form does exist but is rarely considered.

(Refer Slide Time: 23:01)



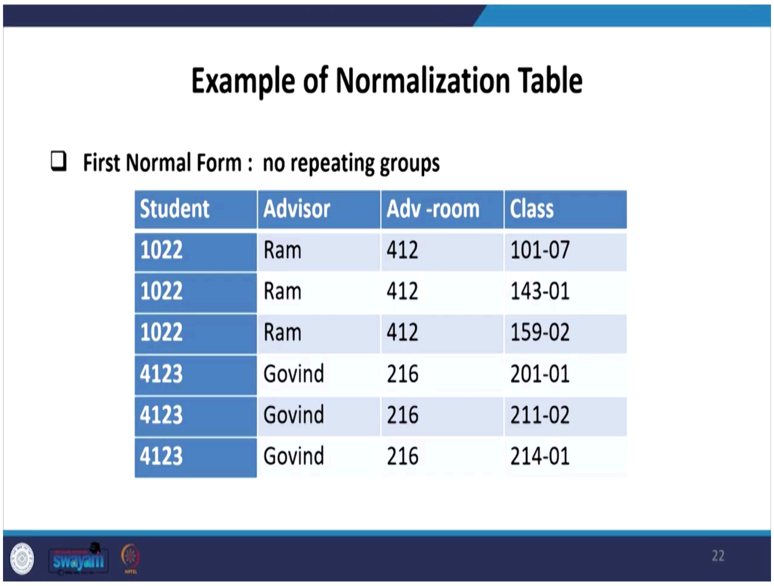
Disregarding these rules may result in less than perfect database design but should not affect functionality.

(Refer Slide Time: 23:14)



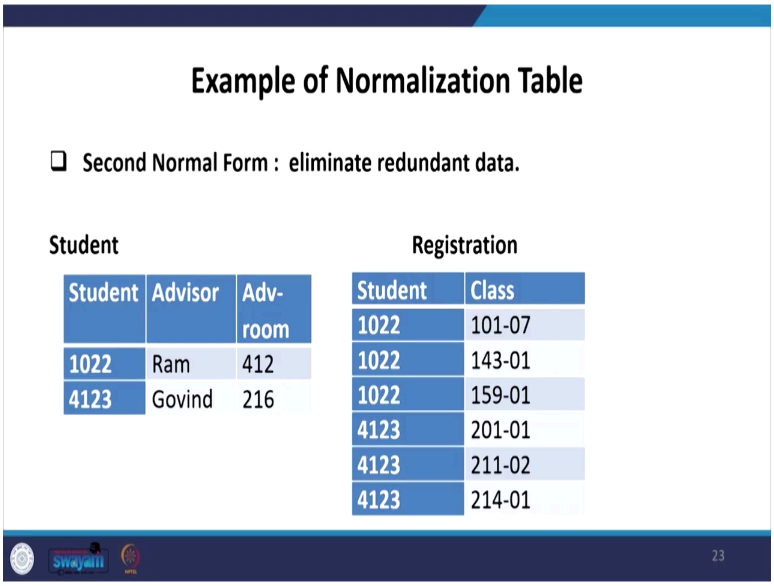
So, the example of the unnormalized table is given here. How it looks like you can just have a look. Class 1, class 2, class 3, and students' IDs, the name of the advisor, and their room numbers are given.

(Refer Slide Time: 23:38)



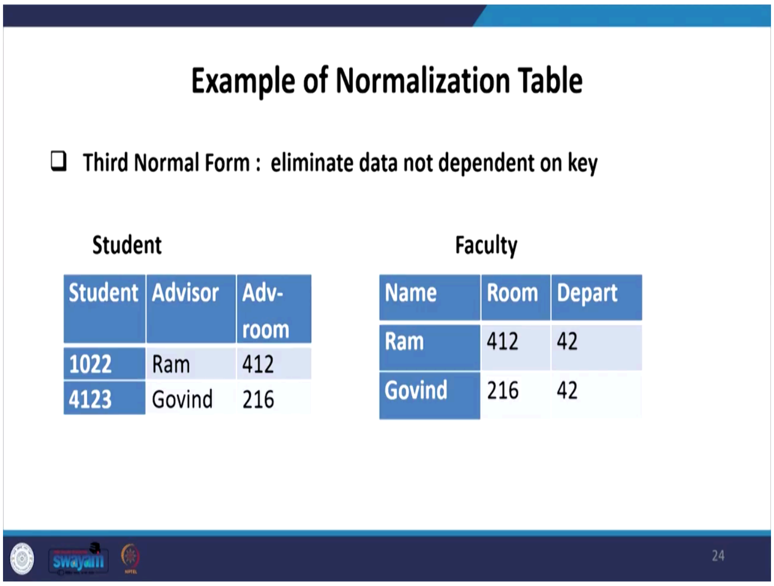
So, the first normal form with no repeating groups is visible here. So, the student details, and advisor details, we are taking this together instead of separately, and then we can have a comparison.

(Refer Slide Time: 24:04)



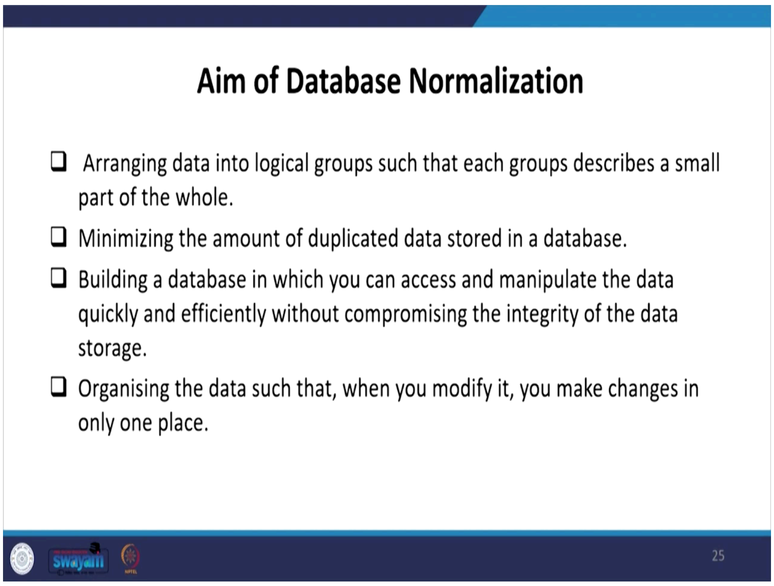
Then coming to the second normal form, we will eliminate the redundant data. Once we eliminate the redundant entries then we are left with a very specific need. Those are required for the analysis. Finally, we are left with our final form of the data since we have eliminated the redundant data.

(Refer Slide Time: 24:28)



The third normal form which we have been discussing eliminates data not dependent on the key. So, those that are not dependent on keys should also be avoided. Like here we have given student information and faculty information, I think here we have shown only data for advisor room class 1, class 2, class 3 but we have some non-dependent keys as well in this data that have also been avoided.

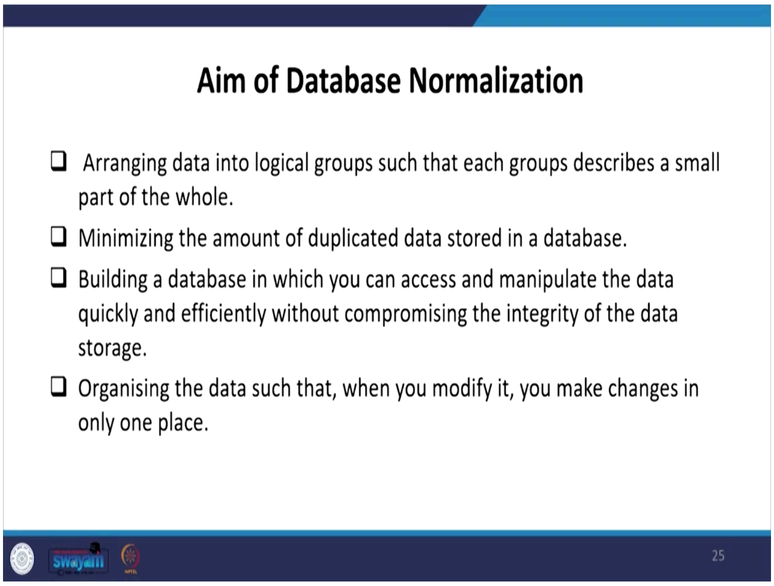
(Refer Slide Time: 25:07)



What are the important objectives of database normalization? Basically, to arrange data into logical groups such that each group describes a small part of the whole analysis. Minimize the amount of duplicated data stored in a database and build a database in which you can access and manipulate the data quickly and efficiently without compromising the integrity of the data storage.

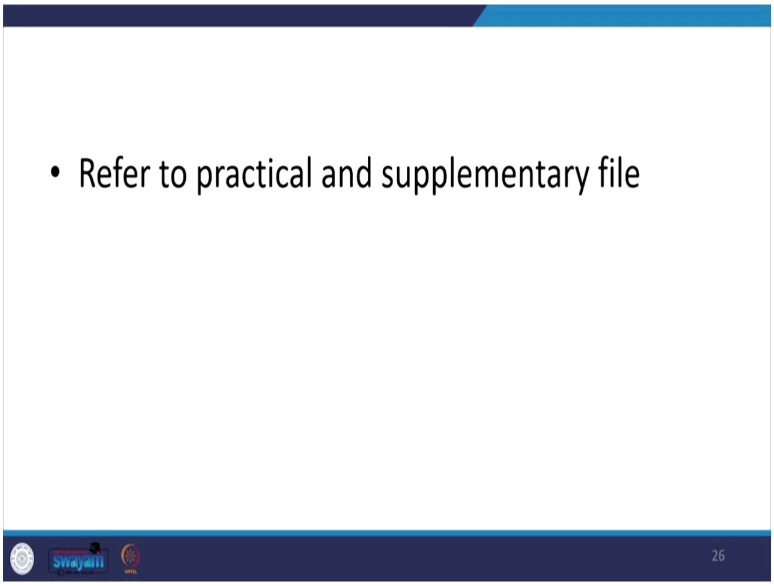
So, we said that first form, second form, and the third form we eliminated the repeated and redundant or non-dependent keys and then organized the data.

(Refer Slide Time: 25:55)



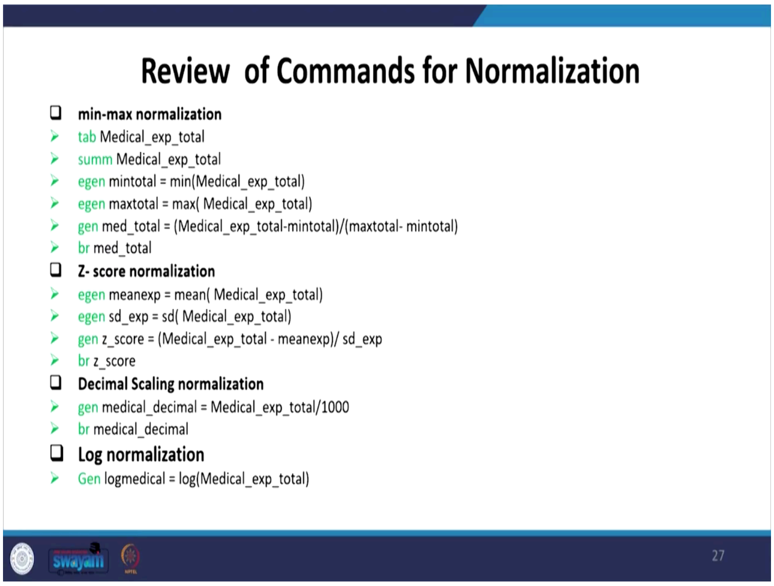
In such that when you modify it you make changes only in one place.

(Refer Slide Time: 26:03)



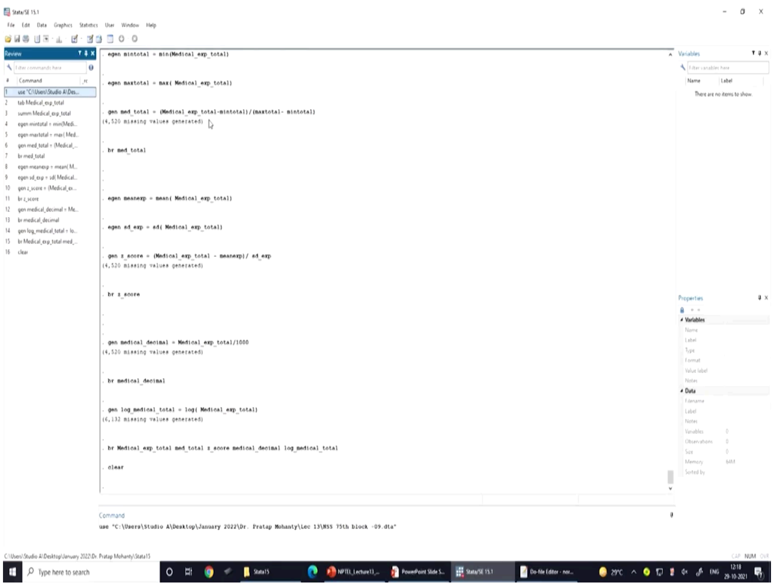
Now, we are referring to our practical and supplementary data.

(Refer Slide Time: 26:08)



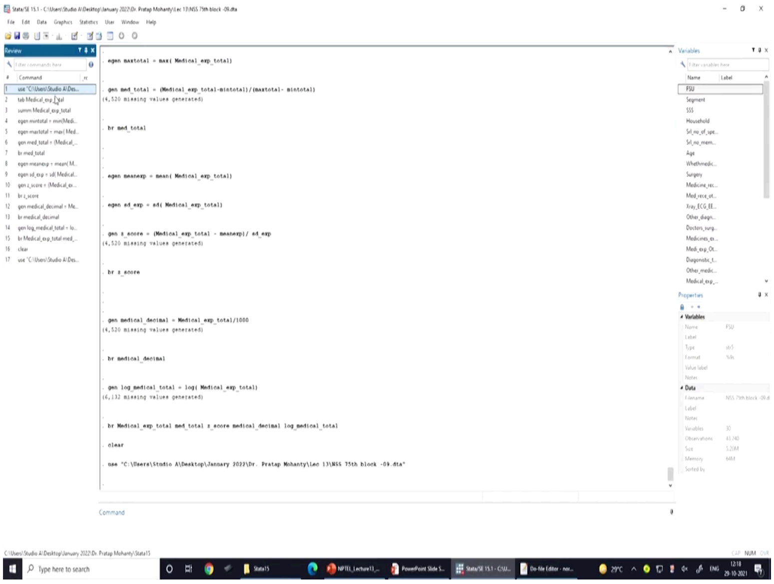
I am now going to experiment with these and then we will review them. We will review these once again with these four approaches, especially we are going to guide you through min-max, Z-score, decimal scaling, and log normalization. So, now let us go to that and clarify.

(Refer Slide Time: 26:31)



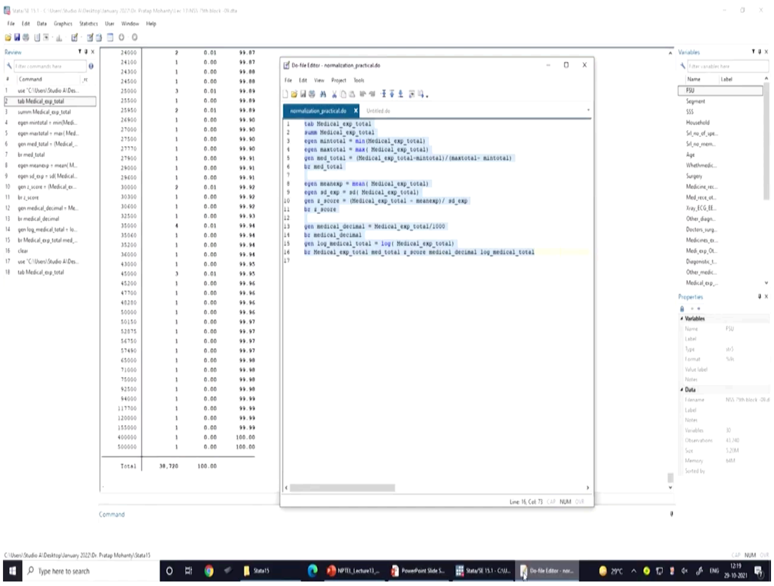
So, here it is on the screen first we will use the data.

(Refer Slide Time: 26:38)



So, now, the data is here we have opened and now we wanted to go with it.

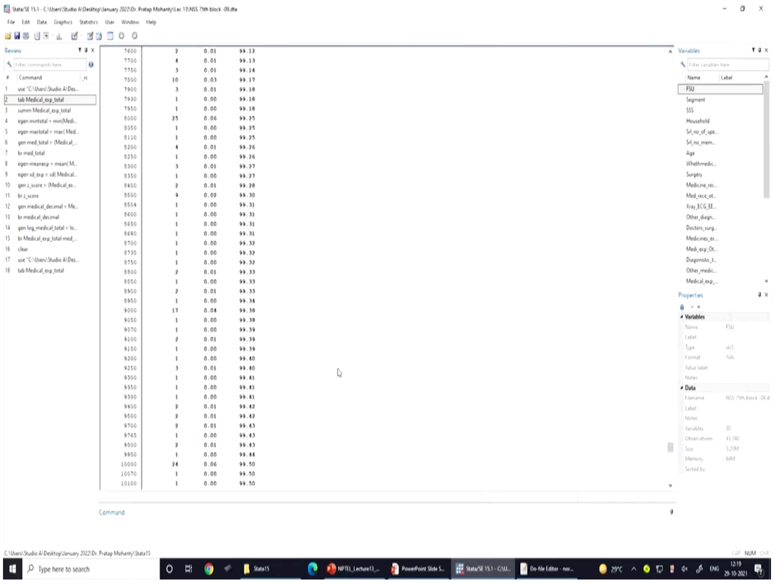
(Refer Slide Time: 26:51)



I can show you the min-max strategy. We are trying to understand all about the medical expenditure in our data set.

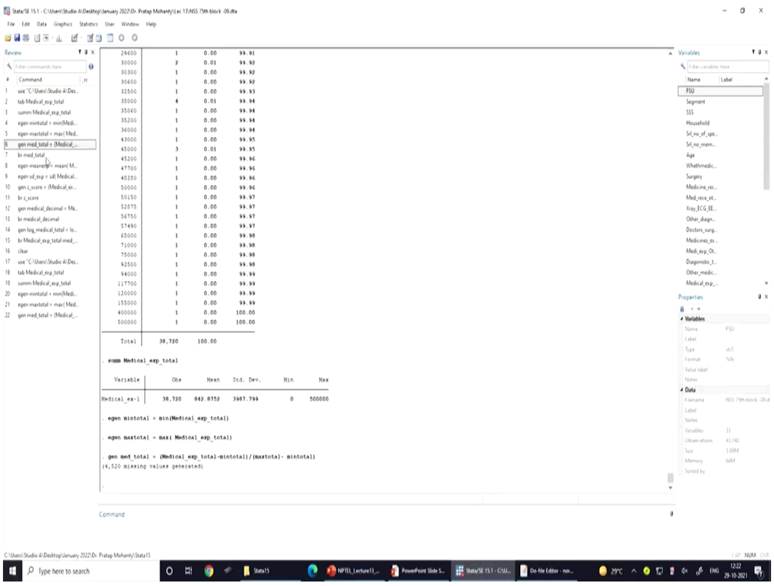
Then we will also analyze with summary statistics like mean, standard deviation, and min-max values as well. But we are now concerned with the min value and the max value. So, from the data now you can easily see.

(Refer Slide Time: 27:41)



Just a minute I am just going through the data first. So, many entries, but from the data, you cannot understand them, if there are lakhs of entries it is very difficult to read. So, the next command we will give here is to understand the summary statistics. So, it is here.

(Refer Slide Time: 27:57)



As we all know that medical expenditure data is usually skewed and usually overestimated whereas, the income data is underestimated. So, when it is skewed it is better to normalize it.

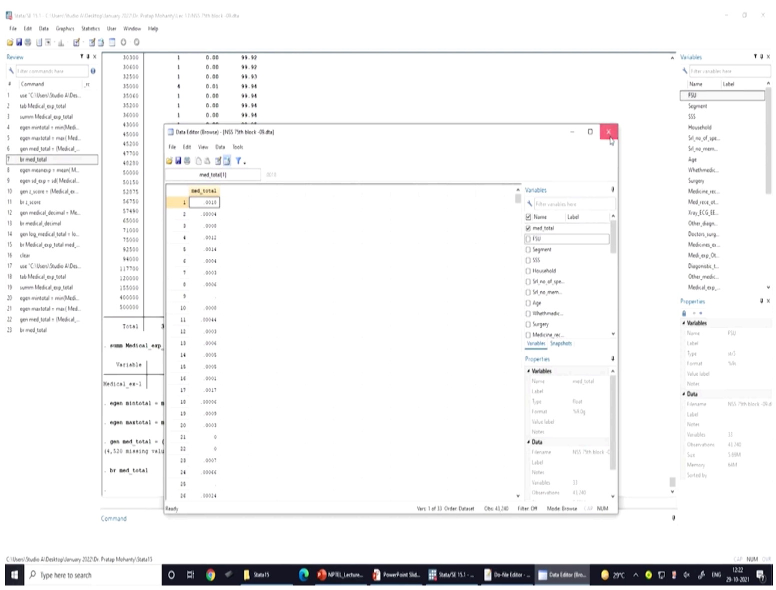
So, hereafter having the sum command, we get the summary of medical expenditure in total. There is a total of 38720 observations, and out of that the min value is 842.8 and the standard deviation was 3988 and the minimum is 0 and the maximum is 500000.

We have such a huge variation. Now, we will generate a min total and the max total. So, the next attempt is to go for the command egen min total. So, here it is. So, we have generated the variable here and similarly, we will also generate the minimum value of medical expenditure and that is generated and the second one is egen max total. We will get the normalized data with this command.

So, generate a medical total here with a min-max strategy. So, what is the medical expenditure? This is the variable I have been showing here minus the min total divided by max total minus mean total in the denominator within the bracket.

The min total, max total everything all those details we have already shown to you. So, then we have generated the min total value variable and the max total. Now, this is our normalized variable which is med_total. Let us see how it differs from the original one. So, we are going to browse this variable.

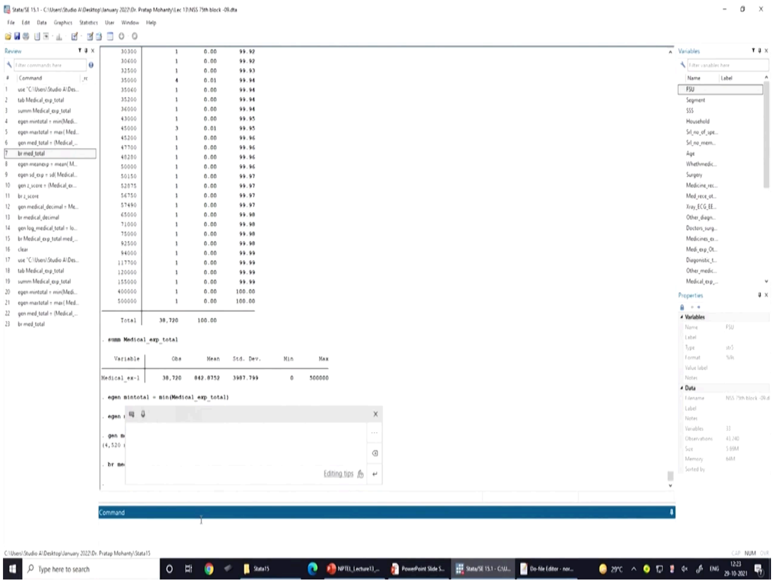
(Refer Slide Time: 31:00)



We will see this is how it looks like. Initially, there are absolute values now we have got them within a normalized scale is 0 and 1 values. So, let us close this and check the next one. Now we are going to operate with the next strategy which is to normalize the data through a standard normal approach or the standardization approach.

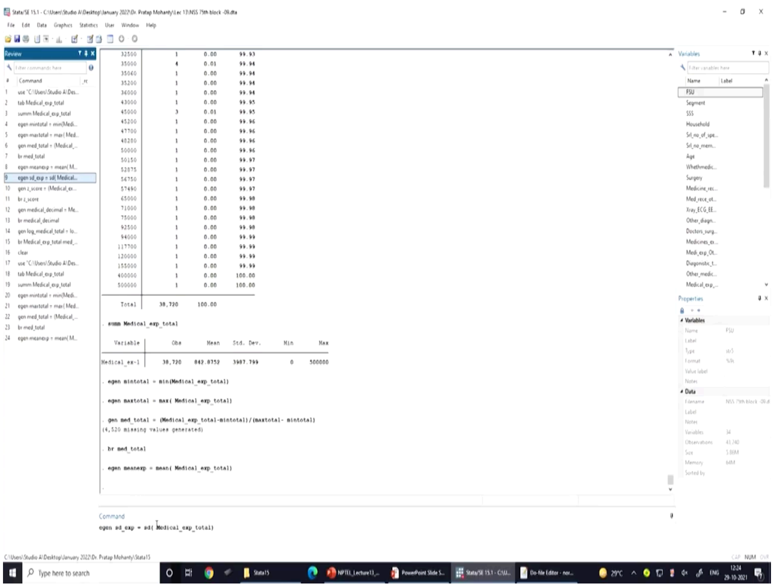
Here is what we require? We require the mean values and the standard deviation. We have to generate the mean values first then a standard deviation of that variable. Then we will compute the Z-score.

(Refer Slide Time: 31:58)



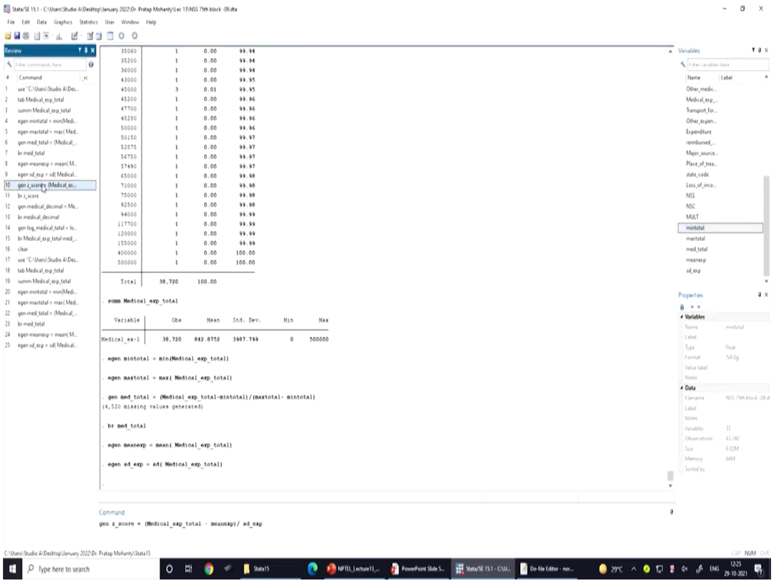
So, this is all presented systematically let us operate it and I am just closing it here.

(Refer Slide Time: 32:06)



Next, we are generating this egen mean expenditure is equal to the mean within the bracket of that variable. So, the mean expenditure value is derived. Next, we are generating the standard deviation. We have taken the name as egen sd_exp = sd(Medical_exp_total).

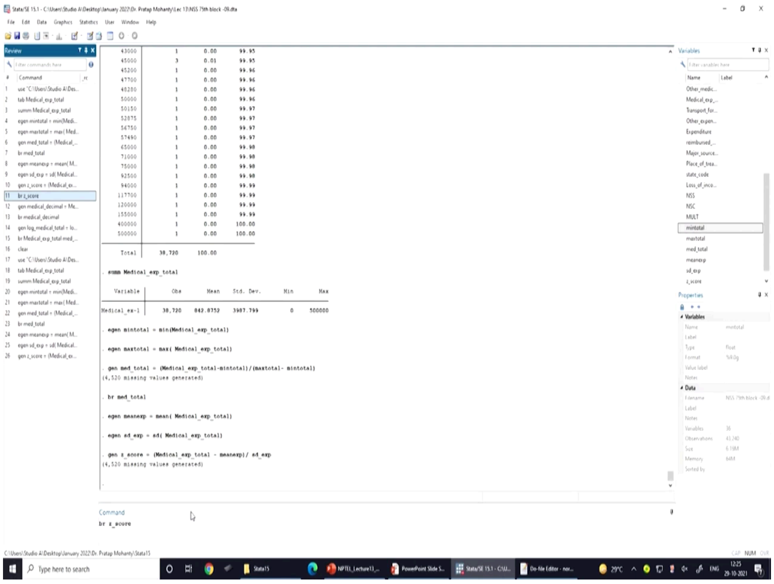
(Refer Slide Time: 32:39)



So, now with the enter, we will get that variable. You can also check in the variables window. All those are created one by one min total, max total, then normalized variable which you have created medical underscore total.

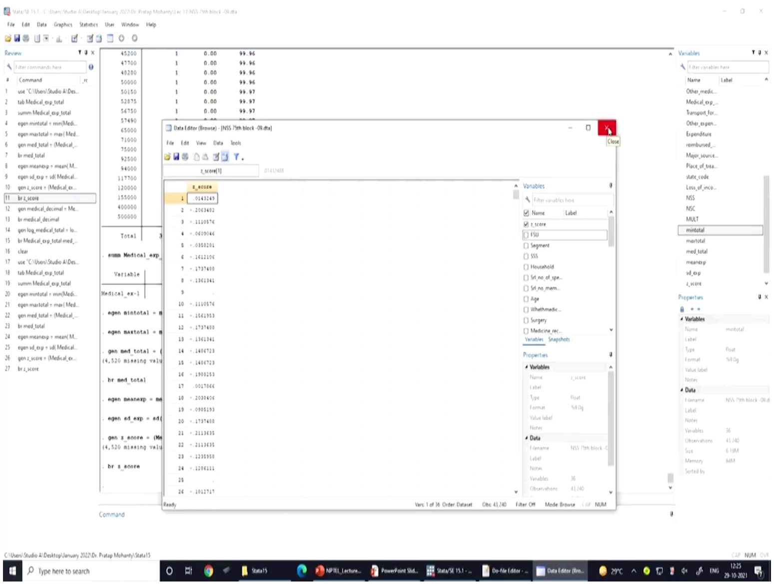
So, the next step is to generate that Z-score since we have already defined the mean value and the standard deviation. So, gen z_score = (Medical_exp_total - meanexp)/ sd_exp

(Refer Slide Time: 33:33)



Now, this is going to give us the Z-score value we can browse and check that as well. So, with the enter, we will get this.

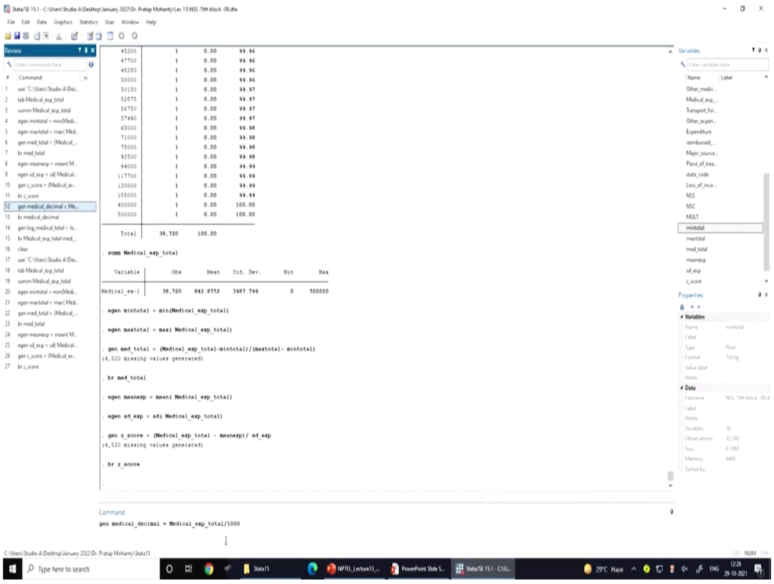
(Refer Slide Time: 34:04)



Z-score varies from minus 1 to 1. So, that is the reason why we have got minus entries as well. It has been standardized. So, we can also check it in summarized form, but it is not required you guys can easily check it.

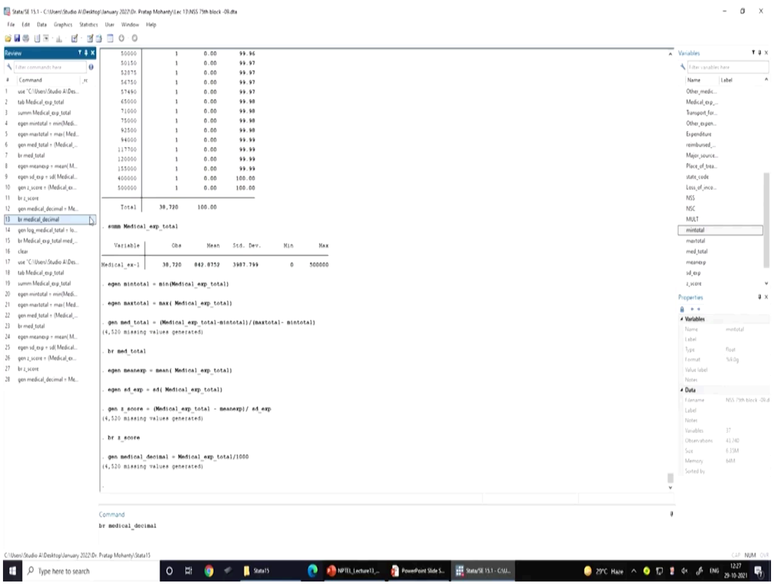
Now we are going to scale down the data with a decimal scaling approach. Then just for your example, we have divided by 1000 here.

(Refer Slide Time: 35:05)



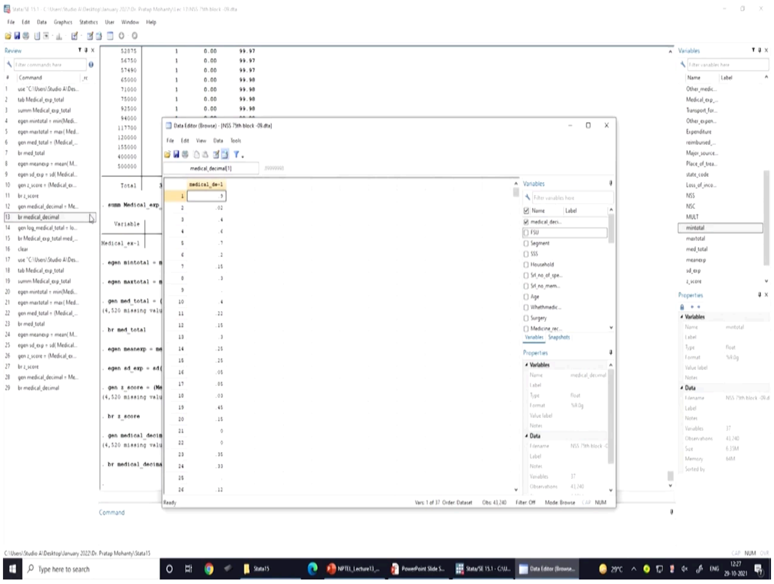
So, here the first approach is to generate a variable like gen medical_decimal = Medical_exp_total/1000.

(Refer Slide Time: 35:13)



You can also note the missing values are generated. So, these are also important in our research. Then you can check the new variable with its decimal scaling, and now you can check this in the browsing window. So, br the variable which you have created.

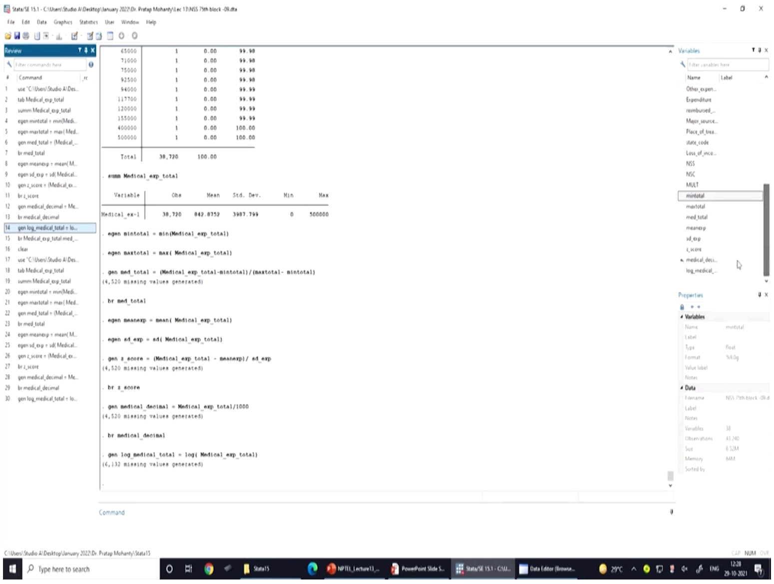
(Refer Slide Time: 35:47)



So, now everything is converted to a decimal range. Z-score normalization we have also done. Now we have clarified these. So, you have to check between these commands we have also given for your practice.

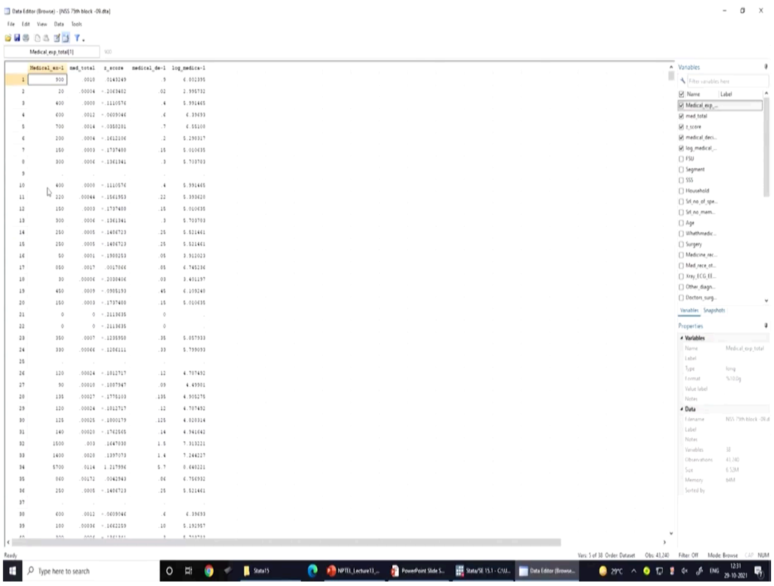
The last one is to get a log transformation of the variable. So, in that case, simply you generate with a variable name but it is better to take the name with log medical expenditure is equal to the log of within bracket of that variable.

(Refer Slide Time: 37:09)



So, we have generated that variable here gen log underscore medical total is equal to the log of this aspect. So, you can also check in our browse window and once we enter this variable will be generated. Now, we can check that in a browsing window.

(Refer Slide Time: 37:40)



Now which one should be used this may be a better call for further research. This is depending upon the data and how you are taking it like in the logarithmic transformation as per our last attempt.

But if there are more entries in 0 or in negative or even 1 in that case log value is not defined. Now another one is called decimal scaling and it reduces the volume of the data to a range. It is not exactly normalized the skewness of the data, but the other two Z-score and min-max strategies are used.

So, this is all our guidance for your better understanding I hope you must have enjoyed these operations and I suggest that you should operate them. You will learn lots of things from it and you can apply some approaches for index formation through the normalization scaling.

So, anyway those details I will guide you later. So, with this information, I think it is time to close if you have any difficulties and doubts do not hesitate and raise this in your live lecture or your chat box.

Thank you.