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.



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)

Min – max Normalization	
 This is simple normalization technique in which we fit the data, in a pre-defined interval. Generally interval define [0,1]. 	
Transform point = $V' = \frac{V - min}{max - min} * (newmax - newmin) + newmin$ $(V - min)$ $(V - min)$	
sanah (s	10

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)

🗆 Exa	mple – Data giv	en on age	e and mon	thly income	:	
In the age Min = 24 In the inc Min = 500	e column , max = 47 ome column 00, max = 10000	V' =	$\frac{V-min}{max-min} *$ New max	(<i>newmax —</i> = 1 , New mir	newmin) + 1 = 0	newmin
Age	Income		Normal			
25	10000		Age	Income		
36	5000		0.04			
45	8000		0.52	0		
47	7000		0.91	0.6		
24	6547		(0.4		
36	6874		0.52	0.37		
39	8700		0.65	0.74		

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 4th 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)

🗆 Unnor	Exa malized t	ample	of Nor	malizat	tion Ta	ble	
	Student	Advisor	Adv - Room	Class 1	Class 2	Class 3	
	1022	Ram	412	101-07	143-01	159-02	
	4123	Govind	216	201-01	211-02	214-01	
🎯 swayan 🔮							21

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)

First Normal Form : no repeating groups											
	Student	Advisor	Adv -room	Class							
	1022	Ram	412	101-07							
	1022	Ram	412	143-01							
	1022	Ram	412	159-02							
	4123	Govind	216	201-01							
	4123	Govind	216	211-02							
	4123	Govind	216	214-01							

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)

) Second	Exa I Normal F	mple c	of N	lormaliza ate redundan	tion Table	
St	tudent				Regist		
	Student	Advisor	Adv-		Student	Class	
			room		1022	101-07	
	1022	Ram	412		1022	143-01	
	4123	Govind	216		1022	159-01	
					4123	201-01	
					4123	211-02	
					4123	214-01	
0 5	wayam 🛞						23

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)

Example of Normalization Table Third Normal Form : eliminate data not dependent on key Student												
nt		Fa										
Advisor	Adv-	Name	Room	Depart								
	room	Ram	412	42								
Ram	412	Govind	216	42								
	Ram Govind	roomRam412Govind216	roomRamRam412Govind216	roomRam412Ram216Govind216	roomRam41242Ram412Govind21642							

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)

Aim of Database Normalization
Arranging data into logical groups such that each groups describes a small part of the whole.
Minimizing the amount of duplicated data stored in a database.
Building a database in which you can access and manipulate the data quickly and efficiently without compromising the integrity of the data storage.
Organising the data such that, when you modify it, you make changes in only one place.
Swayant 👲 25

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)

📑 Stata/SI 15.1			- 0 ×
File Edit Data Graphics Statistics	User Window Help		
©880≡·⊥·₫·₫0			
Review T & X	. epen mintotal = min(Medical_amp_total)	Variables	тях
🔨 Liter commands have 🛛 🔒 🔒		A liter variables h	
# Command _rc	· · · · · · · · · · · · · · · · · · ·	Name L	abel
1 use "CIUtien/Studio A/Des	· · · · · · · · · · · · · · · · · · ·	Destauro	terms to show.
2 tab Medical_exp_total			
3 summ Medical proj total	. gen med_total = (Medical_exp_total=mintotal)/(maxtotal= mintotal)		
4 egen mintatal + min/Medi	(4,520 missing values generated)		
5 egen-martetal = max(Med			
6 gen med_total = (Medical	br med tota)		
7 br med_total			
8 egen meaneup + mean(M			
 egen salesy + sal measure. man salesy + sale measure. 			
12 per covere i construcción	·		
12 pen medical decimal = Mr.			
13 Is medical decimal			
14 gen log, medical, total + lo	. egen si_exp = si(Medical_exp_total)		
15 br Medical, exp. tatal med			
15 clear			
	(4,520 missing values generated)		
	. br z score	Properties	a ×
		8.11	
		# Variables	
		Norse	
	.gen_medical_decimal = Medical_emp_total/1000	Label	
	(4,520 miesing values generated)	3,94	
		Volum School	
	br metical decimal	Notes	
		+ Data	
		Liename	
	. gen log_medical_tota) = log(Medical_exp_total)	Label	
	(6,32 missing values generated)	Notes	
		Vanables	
	. br Medical sup total med total a score medical decimal log medical total		
			5M
	. clear		
	Command		
	ume "C:\Deers\Studio A\Desktop\January 2022\Dr. Fratap Mohaniy\Lec 15UM55 75th block -09.dta"		
Collection of the device of 2010.	Amountation of the set		
Crower dates in being canally 2021bl.	reacy investory (search)		DH _
P Type here to search	🔿 😫 🍘 🛷 📙 Satal 3 🕐 🧶 MRIL Lecture 11, 🦻 Preservatives Side S 👯 Statal 3533 🔮 Do-Site Editor - nor 🛛 😡 2910, 🔿 📀 17, 1	4× d ⁴ 046	29-19-2021

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

(Refer Slide Time: 26:38)

the conversate here 0 Connexed R ref Cluberator Marken Marken are Cluberator Marken Marken (4) comm Medical gray stat gen metal transfording gen metal transfording gen metal transfording gen metal transfording gen metal transfording gen metal transfording	apan markold = mark Markol, ang julai gan markold = "Markol yang julai anatolai//markoli-anatolai) 200 markoly molem pennandi	Filter vanables Nawe FSU Segment SSS Household	Label
Connand In cer COMMANDadde ADes tab Medical cop Stat anom Medical cop Stat anom Medical cop Stat anom Medical cop Stat prometty and medical prometty and extended anomaly anomaly and anomaly prometty and anomaly	ge met stal = America ang usal-mentani/America). mentani) 22 minung wina generateki 24 met_taka	Name FSU Segment SSS Househeld	Label
ee "ChloentGoudo Albes. tab Misical ya Stati warm Medical ya Stati egen metatal = molNeds. egen metatal = molNeds. egen metatal = molNeds. egen metatal = mer(ML) egen maarog = mere(ML) egen station = nill Midical.	pe met sisa - Medinal pe sisa-asisa)/Awasisa: Asisa) 2010 Alexain sisa perintah Ar met sisa	FSU Segment SSS Household	
tab Medical, exp [34] [4] umm Medical, exp (stat) exprementati = mm/Medi exprementati = (Medical,) lar med (stat) exprementation = mean(M) expression = mit Medical expression = mit Medical	30 minist volume previous) In and (sold)	Segment SSS Household	
summ Medical pop total oppen material = min(Medi	ir sel_tstal	555 Household	
open mintatal = mintMedi open martetal = man(Medi pen med sotal = (Medical to med sotal open meanop = mean(M open size = sid(Medical	er med_tota)	Household	
open mastatal = mark (Med pen med tatal = (Medical to med total open maanoup = mean(M open siz oup = sid (Medical open sizene = 104Medical	br met_total		
gen med total = (Medical) In med total egen meaneug = mean(M) egen id_eug = id(Medical) en i unen = (Medical)		SI_ro_of_spe	
br meditatal open meaneup = mean(M open sd_exp = sd(Medical one s scene = (Medical es		Silverment.	
open meaneup + mean(M		Age	
egen sdjeup + sd(Medical		Whethmedic	
and a second a shador of the	egen meanexp = mean(Medical_exp_total)	Surgery	
per contra a contra		Medicine_HC	
en medical decimal + Me.	egen af exp = ad(Medical exp total)	York FCG #	
to medical decimal		Other dags.	
gen log, medical, total + lo		Doctors surg.	
br Medical, exp. total med	gen s_core = (Medical_exp_total = meanerp)/ sd_exp	Medicines, et	
dear	,520 Blashing Values generated)	Med.orp.Ot.	
use "Cillbien/Studio A/Des		Diegonistic_t	
	br z score	Other, medic	
		Medical_0.p.,	
		Properties	
		8	
	gen medical decimal = Medical way total/1000	 Variables 	
(4	520 missing values generated)	Norre	
		Label	
		1094 Locat	100
	e. Beardy _ Sectary	Value label	
		Notes	
	gen log_medical_total = log(Medical_exp_total)	+ Data	
(6	,122 missing values generated)	Liename	NSS 79th block -
		Label	
i	he Matical any total mad total x anora madical desimal log madical total		
	a manadationa motora dona manadona atalanadona		30
	olear		5.554
		Memory	64M
	sse "C'lüsersistadio Albesktopläensary 2022/Dr. Fratap Mohantyllao 131855 75th block -09.dta"	Sorted by	
Ľ.	· · · · · · · · · · · · · · · · · · ·		
Con	•		

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

(Refer Slide Time: 26:51)

the filter - unrelated products - 0 × to the the fight him - 0 × to the fight him - 0 ×	The number hard Total Lotal Total Lotal State State State
The The Topologic Section 2015 and 2015	View provide the order View provide the order View of the order of the
int the Paper Man Marchine Paper Man Marchine Paper Marchine Marchine Paper Marchine Marchine Paper Marchine M	Name (104) Figure Figur
2 U U B A () D P U T T Z S (). match	199 Segment 201 Rec. and Annu- K. yang aman K. Yang aman Madanang m. Madanang m. Madanang m. Madanang m. Madanang m. Madanang m. Madanang m.
<pre>mediatory.prof } instruction { instruction { restruction { rest</pre>	segmen 35 Noundrahd S., yun, et., yun, ag Wathmedic. Samp Madione, etc. Materice, etc. Data (Col.L. Otto, dapa. Data (se, etc. Dagang, etc.) Dagang, etc.
<pre>to Restar proj (net) was National (net)</pre>	s Annualed Series of Series Series press. Annuale Weithweite. Samp Medicine, etc Medicine, etc District, Series District, Series Medicine, etc Medicine, etc District, Series Medicine, etc Dispensiol, J.
<pre>sem Musici, Aug. Long.: per Musici, Aug. Long.: per Musici, Aug. Long.: per Musici, Aug. Aug. Aug. Aug. Aug. Aug. Aug. Aug.</pre>	Sing of year. Sing years. Sing years. Sing years. Materials. Materials. Other space. Destroy years. Materials. Destroy years. Materials. Destroy years.
<pre>que maneil * "confinitai" (grafitation) que maneil * confinitation (grafitation) te maneil * confinitation te</pre>	A y and a second degramment. Surger Medicanter. Neg (CGL). One degram. Dealer, yang. Medicanter. Medicanter. Dealer, yang. Medicanter.
<pre>equation = main [main] [mu] [mu] equation = main [mu] [mu] [mu] equation = main[mu] equation = main[mu] equation = main[mu] [mu] equation = main[mu] equation = main[mu] equation = main[mu] equation = main[mu] equation = main[mu] equation = main[mu] [mu] equation = main[mu] [mu] equation = main[mu] [mu] [mu] equation = main[mu] [mu] equation = main[mu] [mu] [mu] equation = main[mu] [mu] equation = main[mu] equation = m</pre>	kge Wertworker, Sangers Madione, pr Nad, yee, etc. Otwars, yee, Otwars, yee, Madione, a. Doorter, yee, Madione, a. Doorter, yee, Madione, a.
<pre>if up (inc) = (in</pre>	Whetheredic Sargery Medicine, str Medicine, str Na, ICO, BL Other, Angan Other, Sargan Medicine, str Damantic, L. Damantic, L.
<pre>even manage = mana finalises page (mana) management = management = management</pre>	Sargan Malakina, se Majura, etc Van, 160, B.L. Obara, Jung Malakina, a Malakina, a Daganak, J
ere mensal erem " men frank "proving (not) n mensal erem " men frank "proving (not) n mensal erem " not provide the second of the second n mensal erem " not provide the second of the second of the second n mensal erem " not provide the second of the	Medicine, etc Met, etc., etc. Kong KCG, EL. Obten, Songa, Medicine, etc Medicine, etc Medicine, etc Disponsitio, t
<pre>train_first_rest memory_up_first</pre>	Med.yes, etc. 799, E.G. (EL. Ober, degen. Dector, jurg., Medicene, ex., Med.yeg, Ob. Degenerate, f.,
n Tanu n Tanu n mura fara and an tanu tanu mura faran para faran n mura fara ang ang ang ang ang ang ang ang ang an	Yoy, ECG, EL Oliver, diagn Doctore, jung Medicines, en Medicines, en Diagonistic, f
er mennal senait - mennal per jenaitson: er mennal senait - mennal per jenaitson: er lagemenal senait - sign findat per senait tr mennal gen genait - mennal per senaitson jenaits ingeniens jenai	Other, degn Doctors, surg Medicines, ts Medicesp. Ot Diagonistic, t
na matrix factoria mel terri a factoria materix factoria (al materix factoria) na materix factoria mel terri a factoria materix factoria (al materix factoria) na materix factoria mel terri a factoria materix factoria (al materix factoria)	Doctors, surg Medicines, ex Medi _c orp.Ot Diagonistic, t
on for polici, test - top fortical of result. In Solical op ford and test i police material defined top material total	Medicine_es Medice.p.Ot Diagonistic_t
be Menical_exp_inted ann_totel s_bonic minical_decimal log_menical_totel	Med_exp_Ot Diagonistic_t
	Diagonistic_t
	Other, medic
	Medical_exp
	Properties
	B +
	 Variables
	Name PSU
	lan ad
	Format Sale
	Value Tabel
	Notes
	* Data
	Filename NSS 725 bloc
	Label
	Notes
	Variables 30 Observations 41 MD
	Gran S.NM
	Memory bild
, , , , , , , , , , , , , , , , , , , ,	÷
Line 16, Col 73 (JJP MAM (VIR	
	*
	Sector 1 control

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)

vev Tax	7600	2	0.01	99.12					A 1	ariables -	τ.,
Liber commands have	7750	4	0.01	99.13					1	Litervasables	
Command or	7750		0.01	99.34						Neme	Label
contration (Alberta State Alberta	7100	10	0.03	99.17						601	case
tab Medical eue total	7910	í	0.00	99.38						Generated	
samm Medical and total	7950	1	0.00	99.38						111	
egen mintatal + min/Medi-	8300	25	0.05	99.25						Household	
egen mantatal = max(Med	8350	1	0.00	99.25						Sil no of spe	
pen med total + (Medical	8110	1	0.00	99.25						Sil no mem.	
by med, total	8200		0.01	99.26						Age	
egen meanerig + mean(M.,	4100		0.00	99.26						Whethmedic	
egen sd_exp + sd(Medical	1350	í	0.00	99.27						Surgery	
gen zussere = (Medical, ex	8460	2	0.01	99.28						Medicine, Ht	
br 2, score	8500	,	0.02	99.30						Med, rece, et	
gen medical_decimal + Mc	8504	1	0.00	99.31						Xray_ECG_EE	
br medical, decimal	1460	1	0.00	99.31						Other, dags	
gen log, medical, total + lo	8450	3	0.00	99.31						Dectors_surg	
br Medical, exp. total med,	4100		0.00	99.31						Medicines_et	
clear	8735		0.00	\$9.32						Med.orp.Ot	
use "C10sen/Studio A/Des	9750	1	0.00	99.32						Diagonistic_t	
tab Medical_exp_total	8800	2	0.01	99.33						Other_medic	
	9950	3	0.00	99.33						Medical_exp,	
	8950	2	0.01	99.33					P	operties	
	9950	1	0.00	99.34					1		
	9100		0.00	59.30						Variables	
	9070		0.00	99.39						Norre	FSU
	9100	2	0.01	99.39						Label	
	9150	1	0.00	99.39						7,91	and .
	9200	1	0.00	99.40						Format	5.91
	9250		0.01	99.40	1	5				Value label	
	9360	1	0.00	99.41		*				Notes	
	9350	1	0.00	99.43					· · · · ·	Data	
	9800		0.01	53.42						Land	No ODDER -
	9100	2	0.01	99.42						Label	
	9700	2	0.01	99.43							20
	9765	1	0.00	99.43							41,240
	9100	2	0.01	99.43					- 11		5.20M
	9950	1	0.00	99.44						Memory	64M
	10000	24	0.06	99.50						Sorted by	
	10070	1	0.00	+9.30							
	14143		0.00						¥		
	Command										

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)

8	lata/SE 15.1 - C1User/Shudio N/Desitoj	pilanuary 2022/Dr. Pr	utap Mohanty(Lec 1)	UNSS 79th block -	09. <i>ma</i>											- 0 ×
F6	e Edit Data Graphics Statistics	ther Window	Help													
2	N & U B - T - Q - Q (3000														
Rev	iev T # X	29400	1	0.00	99.91									~ V#	iables	τů×
٩	l iter commands have	30000	2	0.01	99.92										Liter variables P	
	Command rc	31400		0.00	\$9.92										Name	abel A
1	use "Cill/sen/Studio A/Des	31500	1	0.00	99.93										FSU	
2	tab Medical_exp_total	35000	4	0.01	99.94										Segment	
3	summ Medical_exp_total	35040	1	0.00	99.94										\$55	
4	egen mintatal + min/Medi	35200	1	0.00	99.94										Household	
5	egen mantotal = max(Med	34000	1	0.00	99.94										Sil, no, of, spe	
6	gen med_total + (Medical	45000		0.00	99.95										Siljeojmen	
7	br med_total	45200	í.	0.00	\$9.96										Age	
8	egen meaner#n mean(M.,	47700	1	0.00	99.96										Whethmedic	
9	egen sd_exp + sd(Medical	40200	3	0.00	99.96										Surgery	
10	gen zuscere + (Medicalues	50000	1	0.00	99.96										Мебсіне, нс	
11	ps t"recost	50150	1	0.00	99.97										Med_rece_ot	
12	gen medical_decimal + Me	52075		0.00	99.97										Xray_ECG_EE	
13	br medical_decimal	57450		0.00	99.97										Other_dags	
14	gen log_medical_tatal + lo	65000	i	0.00	\$9.90										Doctors_surg	
15	br Medical_exp_total med	71000	1	0.00	99.90										Medicines_ex	
15	Clear .	75000	3	0.00	59.90										Med_or_OL.	
17	use "Cilbien/Studio A/Des	92500	1	0.00	99.90										Dagonsbc_t	
18	tab Medical_exp_total	94000	1	0.00	99.99										Other medic	
	symm Medical (org. fotal	117700		0.00	59.99										Mebcal, exp	
- 20	eges minestal i ministed	155000		0.00	53.35									Pre	iperties	ų ×
21	egen martatal n max(Med	400000		0.00	100.00											
"	fer menintrar a (menina)	\$00000	1	0.00	100.00										Variables	
		Total	30,720	100.00											Label	
															194	100
		. sum Medica	il_exp_total													
		Variable	0.4	Nean	Std. Dev.	Min	Max									
															Data	
		Medical_ex-1	30,720	042.0752	3907.799	0	500000								Gename	N55 795 Block -014
															Label	
		. egen mintot	al = min(Medic	al_exp_total	0										Notes	
															Variables	
		. egen maxtot	si - sext sedi	cal_aub_coci											Observations	41,240
		data and tot	al = (Medical	em total-m	intotall/inarts	tal. mintor	411									5.69M
		14,520 misesin	g values gener	ated)												SIM
														1.00		
														v		
		Command														
		Connecto														
CIL	hen/Studio A/Desktog/January 2022/Dr.	Pratap Mohanty/Stat	u15													CIP NUM OF
									A		-				4 100	12-22
1	ype nere to search		0.1	a 💔 .	Side 15		e 4	renti technell,	energiet St	and H and	A CONTRACTOR OF	CO RELADOR HAL	- arc - x 0	0.00	- 0° - 0%	29-10-2021

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)

T # X	30300	3	0.00	99.92								~	lariables	,
er commande hara	31400	1	0.00	99.92									A Litter constitu	
	32500	1	0.00	99.93										a second
ommana jit	35000	4	0.01	99.94									76877.8	Line
e Crosevande Adel	35160	1	0.00	99.94									150	
o medical jesp jonal	34000		0.00	53.54									segment	
ener regela jug juga	43100		0.00	99.95									303 Mountabald	
permission of mercial and	45000	,	0.01	99.95									Sil en el cen	
n med tatal = (Medical -	45200	1	0.00	99.96									Sil no men.	
aned total	47700	1	0.00	99.96									And a	
en manena + man/M	48280	1	0.00	59.96									Whethmedia	
en ul ma - ul Medical	50000	1	0.00	99.96									Summ	
n z scere + (Medical ex	52875		0.00	\$9.97									Medicine rec.	
1 1/10	56750	i	0.00	99.97									Med serve at	
n medical decimal + Me.	57490	1	0.00	99.97									Trav FCG RE	
medical decimal	65100	1	0.00	99.98									Other daan.	
n log medical tatal + lo	71000	1	0.00	99.90									Doctors surg	
Medical exp total med	75000	3	0.00	59.90									Melicines ex.	
ter .	92500	1	0.00	99.98									Med. exp. OL.	
e "C1Usen/Studio A/Des	117700		0.00	59.59									Diagonistic 4	
b Medical_exp_total	120000		0.00	99.99									Other, medic	
mm Medical, exp. total	155000	1	0.00	99.99									Medical, exp.	
pn mintatal v min(Medi	400500	1	0.00	100.00									Incention	
en mastatal + max(Med	500000	1	0.00	100.00										
n med_total = (Medical													4 Variables	
med_total	Total	38,720	100.00										Name	FSU
													Label	
	- Fille Heards	"out-cocar											Ter.	10.5
	Variable	0.4	Nean	Std. Dev.	Nin	Max							Format	5.01
													Value label	
	Medical_ex-1	30,720	042.0752	3907.799	0	500000							Notes	
													# Data	
	. egen mintot	al = min(Medic	al_exp_total	0									Liename	NSS 75th big
	an (¢					Label	
	. egen 1 -4 -4												Notes	
													Vanables	
	(4,520)													41,240
							0	9						2.0004
	. br me													
						Editing tip:	5 B +	-						
												v.		
	f													
	Command										 	•		
			¥.											

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

(Refer Slide Time: 32:06)

🖬 🕸 U 🗟 - 📖 - 🖻 - 🖉 🖬	0 0 0											
iew T&X	32500	1	0.00	99.93						~	Variables	1
Liter command chara 0	35000	4	0.01	99.94							A litervarable	
Comment	35060	1	0.00	99.94							Neme	Label
contrains (1)	35200	1	0.00	99.94							601	une
tab Maferi na tatal	43100		0.00	10.15							General	
same Medical ma total	45000	;	0.01	99.95							505	
egen mintatal + min/Medu.	45200	1	0.00	99.96							Household	
egen mantatal = man(Med	47700	1	0.00	99.96							Sil no of spe	
gen med total + (Medical	48280	1	0.00	99.96							Sil no men.	
by meditotal	50000	1	0.00	99.96							Age	
egen meanerig + mean(M	53475		0.00	99.97							Whethmedic	
egen sd_exp = sd(Medical	56750	í	0.00	99.97							Surgery	
gen 2, score + (Medical, ex	57490	i	0.00	99.97							Medicine, rec	
br 2, score	65000	1	0.00	99.90							Med, rece, et.,	
gen medical_decimal + Me	71000	1	0.00	99.90							Xyay_ECG_EE	
br medical, decimal	75000	1	0.00	99.90							Other, dags	
gen log_medical_total + lo	92500	1	0.00	99.90							Doctors_surg	
br Medical_exp_total med	117700		0.00	20.00							Medicines_ex	
clear	120100	í.	0.00	\$9.99							Med.orp.OL.	
use "Cillben/Studio A/Des	155000	i	0.00	99.99							Diagonistic_t	
tab Medical_exp_total	400000	1	0.00	100.00							Other, medic	
summ Medical_exp_total	\$00000	1	0.00	100.00							Medical_exp	
egen mintotal + min(Med											Properties	
egen martatal = max(Med	100.01	38,720	100.00								8	
gen med_total + (Medical,	sum Medica	em total									# Variables	
br med_total											Name	FSU
egen meanerp + mean(M	Variable	08.8	Nean	Std. Dev.	Nin	Маж					Label	
											1,94	10.5
	Medical_ex-1	38,720	042.0752	3907.799	0	500000					Lorent	5.9
		a an									Nature Labor	
	. egen annoc	er - annipaeore	at_exp_coca.								4.0.44	
	. egen maxtot	al = max(Med)	ical emp tota	11							Genane	NSS 799-March
											Label	
	. gen med_tot	al = (Medical	exp_total-m	intotal)/(BAX	total- minto	(tal)					Notec	
	(4,520 missio	ș values genes	rated)								Variables	34
											Obsenations	41,710
	. Br Bed_tota										518	5.00M
	egen meaner	p = mean(Med)	ical amp tota	11							Memory	64M
										v .		
	Command											
	open ad exp	pd(Medical_e	mp_total)									

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)

1 🛛 🕲 U 🖻 - 11 - 12 - 12 i	0 0 🗇										
niew TAX	35040	1	0.00	99.94						 Variables 	т и :
Liter commands have 0	35200	1	0.00	99.94						A filter variable	
Command rc	43000	1	0.00	99.94						Name	Label
use "Cillsien/Studio A/Des	45000		0.01	99.95						Other medic	
tab Medical_exp_total	45200	1	0.00	99.96						Medical_cop_	
summ Medical, erg. total	47700	1	0.00	99.96						Taroport, for	
egen mintatal + min/Medi	40200	1	0.00	99.96						Other, expension	
egen mantatal = max(Med	50000	1	0.00	99.96						Expenditure	
gen med_tatal + (Medical	52075		0.00	99.97						reimbursed	
br med_total	54750	i	0.00	99.97						Major_source.	
egen meanerp = mean(M	57490	1	0.00	99.97						Place_of_trea	
egen sd_exp + sd(Medical	65000	1	0.00	99.90						state_code	
genziscontre (Medical, ec	71000	1	0.00	99.98						Lass_ef_inco	
pi t'ircois	75100	1	0.00	99.98						N55	
gen medical_decimal + Me	94550		0.00	10.10						NGL	
to medical decimal	117700	1	0.00	99.99						MULT	
gening medical total in to	120000	1	0.00	99.99						Pharocean	
dea	155000	1	0.00	99.99						manotal and	
use 'C'illien/Statio A/Des	400000	1	0.00	100.00						0440510	
tab Medical mar total	500000	1	0.00	100.00						of me	
summ Medical ma total	Total	10.720	100.00								
egen mintotal + min/Medi										A	
egen mantatal = max(Med	. sum Medica	exp_total								Properties	
gen med_total = (Medical										4 Mariahier	
br med_total	Variable	C0.4	Nean	Std. Dev.	Nin	Маж				- variables	
egen meanerp + mean(M.,	Notice 1 and	10.110		1443 344						Label	
egen sid_exp = sid(Medical	WARTCHT KK. T	1 30,120	042.0752	3907.199	0	346004				5.94	first
	. egen mintot	al - min(Medi-	cal emp total)						Format	54.09
										Value label	
	. egen maxtot	al = max(Ned	ical_emp_tots	11						Notes	
										* Data	
	. gen med tot	AI = (Medical	exp_total-mi	ALACAL) / (BART	otal- Blatot	411					NSS 79th block -08
		1 calles dere								Label	
	. br med tota									Variables	
											41,240
	. open meaner	p = mean(Med	ical_exp_tots	11							6.02M
										Memory	6M
	. open sd_exp	- 65(NeSica	i_exp_total)							Sorted by	
	ľ.										
	Command									*	
	gen z_score =	(Medical_exp_	total - means	ab) / ad_exb							

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)

ie Lat Data Graphics Statistic I 🖬 🏶 🛄 🗷 - 🔟 - 🗭 - 🗹 -	s User Window	Help								
1880 B.L. 0. 01										
	🗈 🗂 O · O									
14.8	43000	1	0.00	99.95				^	Variables	та
Liter commands have	45000	3	0.01	99.95					A liter constitu	
Command	45200	1	0.00	99.96					Neme	Label
Contraction of the Allines	47700		0.00	99.96					Other made	case
tak Maileal and total	50000		0.00	33.36					Medical and	
same Medical and total	50150		0.00	99.97					Successfully	
even mintatal a miniMedi	52075	1	0.00	99.97					Other more	
epermental r methods.	56750	1	0.00	99.97					localities.	
oper manager of the product	57490	1	0.00	99.97					manhaned	
be and total	65000	1	0.00	99.90					Mains strate	
eren eren er eren M	71000	1	0.00	59.90					Nace of here	
equivalences of Medical	75000	3	0.00	99.90					state code	
ego record a Medical es	92500	1	0.00	59.50					Law of inco.	
he a second	1 117700		0.00	53.55					NCS NCS	
or sprom	120000		0.00	\$9.99					NSC	
b medical decimal	155000	- i	0.00	99.99					MUT	
contract conficul total a la	400000	1	0.00	100.00					- mintered	
he Medical and total and	\$00000	1	0.00	100.00					metotal	
clas									coal total	
use 'Cillion/Studio A/Des	Total	38,720	100.00						044000	
tab Medical martetal	and Made as								si ne	
comm Medical gas total	. Find Reales	r_eap_cocar							1 include	
even mintatal a miniMedi	Variable	0.4	Nean	Std. Dev.	Nin	Маж			1,000	
earn martial - mark Med.									Properties	
an real total + (Medical	Medical_ex-1	30,720	042.0752	3507.759	0	500000			8 **	
by med total									* Variables	
even meaning a mean/M.	. egen mintot	al = min(Medic	al_exp_total	13						
earn of eus - of Medical									1	10 cm
een z scere + (Medical ex	. egen maxtot	al - max(Medi	cal_emp_tots	-13					1994 Locat	
	and part tot	al = (Medical	ern total-mi	intotal)/imarto	al- mint.	1411				
	(4,520 missio	r values gener	ated)						Notes	
									+ Data	
	. br med_tota	1							Liename	NSS 795 block -0
	-								Label	
	. open meaner;	p = mean(Ned)	cal emp tota	411					Notes	
									Variables	35
	. egen sd_exp	- sd(Medical	exp_total)						Observations	41,240
		- Medical en							Sax	6.1944
	(4.520 miesto	values gener	ated)	enerth), auferth					Memory	64M
									Sorted by	
	Command		Δ.							
	hr s_score		"							
Usen/Studio A/Desktog/January 2022/Dr	r. Pratap Moharey/Stat	a15								CAP NUM C
D Tune here to count		0-	5 🙈	a 1 94-11			1971) Laterall 🔒 Provident Oda C 🔠 Bala (1951) (191) 📝 Da Ha (Alay ana 🤷 Mari	A 11 4	4 4 100	1225

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)

ew tak	X 45200	1 0.00 99.96			 Variables 	1
l iter commands have	47700	1 0.00 99.96			A Liter venables he	
Command IV	40200	1 0.00 99.96			Name La	abal
use 'Cillbert/Statio #/Des	50150	1 0.00 99.96			Other coeffic	
tab Medical exp total	52875	1 0.00 99.97			Medical exp	
summ Medical, exp. total	54750	1 0.00 99.97			Variabet. Nor	
egen mintatal + min/Medi	57490				Other expent.	
egen mantatal = man(Med	65000	[] Uata Editor (Brewse) - (NGS-79th block - 09.4ta)		- U 1	Expenditure	
gen med_tatal + (Medical	71000	File Edit View Data Tools		Close	reimbursed	
by med_total	75900	# # @ D D # # # T.			Major_source	
egen meaneup + mean(M.,	94000	t weekil (harden)			Place_of_trea	
egen sd_exp + sd(Medical	117700	Sisteral di Directeur			state_code	
gen zuscere + (Medicalues	120000	E_eccre	· · · · · · · · · · · · · · · · · · ·	Variables 🔍	Law, of jeco	
br z,score	155000	1 .0140245		A Filter variables here	N55	
gen medical_decimal + Me	400000	2 2043482		P Name Label A	NSC	
br medical_decimal	509900	31110576		R I SCOW	MURT	
gen log_medical_tatal = lo	Total	40603046		T FSU	mintotal	
br Medical_exp_total med		80000201		Segment	mantotal	
Clear	. sumn Medical_exp	61612116		555	med_total	
use "Cilbies/Studio AlDes		7 - 1737400		Household	meanerp	
tab Medical_exp_total	Variable	F - 1301301		St,re,et,spe-	rq"erb	
symm Medical erg total	No. 41 and 1			St.no,mem	1,0004	
egen minestal i minimete	Megrost "ex.7	10 - 1110030		C Apr	Properties	
egen manatar y man men.	egen mintotal - m			Whethmedic-	8 **	
key med total				Surgery	 Variables 	
energenergen v menni M.	. egen maxtotal = m	11		Medicine.rec. Y		
earn of roa - of Medical		13 - 1941341		Tasabes Stappeds	Table 1	
sen z scere + (Medical ex	. gen med_total = (141404723		Properties #	1994 Logend	52.04
br z score	14,520 milesing vers	15 - 1406723		+ Variables ^	Value John	
	br med total	14 - 1900253		Name Lycore	Notes	
		17 .0017044		Label	# Data	
	. open meanexp = me	102020416		Type Nout	Liename	NS5 799-106
		190905190		Format 59.5g	Label	
	. open sd_exp = sd	201737450		Vulue label	Notes	
		212113435		Notes	Variables	36
	(4.520 mission valu	222113435		Advanta BES Per Mark 7		41,240
		231235350		Low		6.1944
	br s_score	24 - 1204111		Notes	hierory formation	5411
		15		Variables 36	Sofer by	
		14 - 1411717		Observations 43,340	¥	
		Feedy	Vanc 1 of 16 Onder: Dataset Obc 43,740 Filt	e.Off ModelBrowse CD NUM		

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)

eview T&X	45200	1	0.00	99.96			~	Variables	т и
titer corretands have 0	47700	1	0.00	99.96				A liter variable	
Command re	48280	1	0.00	99.96				Nava	Label
use "C\Ubien\Studio A/Des	50150		0.00	99.97				Other medic	
tab Medical_exp_total	52075	1	0.00	99.97				Medical_exp	
summ Medical, pag, total	54750	1	0.00	99.97				Tanpot, Inc.	
egen mintotal + min(Medi	57490	1	0.00	99.97				Other, expen.	
egen mantatal = max(Med	65000	1	0.00	99.90				Expenditure	
gen med_total + (Medical	71000		0.00	99.90				reimbursed	
by med_total	92500		0.00	\$9.90				Major_source	
egen meaneup + mean(M	94000		0.00	99.99				Pace_of_tota	
egen sd_exp + sd(Medical	117700	1	0.00	99.99				state_code	
0 gen zuicere + (Medicaluer	120000	1	0.00	99.99				Lass of jeco	
1 br zjacove	155000	1	0.00	99.99				NSS	
2 gen medical_decimal = Me	400000		0.00	100.00				NSC	
3 bi medical_decimal	507700		0.00	100.90				MULT	
4 gen log_medical_total + lo	Total	30,720	100.00					mintotal	
5 bi Medicaljevpjtatal medj								mentotal	
6 dear	. sum Medical	exp_total						med_total	
 Viel Cristel Studie Walds 								meanorp	
 tao mesika jergijota tomm Medical me total 	Variable	Cba	Nean	Std. Dev.	Nin	Max		10,010	
 some messagerg jorar some mintet die minibilitet 	Madical av-1	10.110	842 8752	1667 766		10000		1,0104	
1 egen martital = mar(Med.	monten_ex 1							Properties	
2 new resel total + (Medical -	. egen mintota	l = min(Medic	nal_exp_total	13				8 **	
3 by med total								* Variables	
4 egen meaners = mean(M.	. egen maxtote	1 = max(Ned)	ical_emp_tots	•11					
5 egen id. exp = id! Medical								Los	first.
gen zusare + (Medical, ex	. gen med tota	a = (Medical	exp_total-m	INTOCAL) / IBAXCO	CAI- BINC	otal)		Locat	34.04
br 2, score								Value label	
	. br med total							Notes	
	-							# Data	
	. egen Beanexy	- mean(Ned)	ical_exp_tots	413				Liename	NSS 735 block -0
								Label	
	- open ad_exp	- si Nedical	(_exp_total)					Notes	
	gen a score	- (Medical es	ep total - me	nanesp) / ed esp					36
	(4,520 818810)	values genes	rated)						41,240
									NM.
	. br s_score								
	·						×		
	Command								
	gen medical de	imal = Medic	al emp total	/1000					

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

(Refer Slide Time: 35:13)

ev T\$X	\$9900	1	0.00	99.96				Variables	T # 3
Liter commands have	50150	1	0.00	99.97				A Liter variable	
Command IV	52075	1	0.00	99.97				Name	Label .
use 'Cillion/Statio A/Des	56750		0.00	59.97				Other media	
tab Medical exp. total	65000		0.00	\$9.90				Medical exp	
summ Medical, exp. total	71000	1	0.00	99.90				Turopot, for	
egen mintatal + min(Medi	75000	1	0.00	99.98				Other, expen	
egen mantatal = man(Med	92500	1	0.00	99.98				Expenditure	
gen med_total = (Medical	117700		0.00	22.22				reinbursed	
br med_total	120000	i	0.00	\$9.99				Major_source	
egen meanerg + mean(M	155000	1	0.00	99.99				Place_of_trea	
egen sdjesp + sd(Medical	400000	1	0.00	100.00				state_code	
gen zuscere + (Medicalues	500000	1	0.00	100.00				Less_ef_ince	
to gate and find designed a Me	Total	10.720	100.00					NGS NGC	
by medical decimal from the								MUT	
orn inn medical total - in	. sum Medical	exp total						mintotal	
br Medical era total med								mentatal	
dear	Variable	C6.4	Nean	Std. Dev.	Nin	Ма		med_total	
use "Cillbien/Studio AlDes	Medical ex-1	30.720	842.8752	3807.759	0	10000	-	meantrip	
tab Medical_exp_total								sdjenp	
summ Medical_exp_total	. egen mintota	1 = min(Hedio	total_exp_total	3				2,50016	
egen mintatal + min(Med								Properties	a :
egen martatal = max(Med	. egen maxtota	al = max(Med)	ical_exp_tots	11				8. * *	
gen med_total = (Medical	. gen med tota	1 = (Medical	esp total-mi	ntotal)/(BARto	tal- minto	tal)		 Variables 	
ermenterer a manif M	(4,520 missing	values genes	(bsted)					Name	mototal
equivalence of Medical								Label	
pen z scere + (Medical ex	. br med_total							1/24 Locard	Tool .
br z score		- manual Medi	ical am tota					Value John	
gen medical decimal + Me	- open detailed	- ment men	and and and					Notes	
	. open ad exp	- ed(Medical	exp_total)					+ Data	
								Liename	NS5 799-block -05
	. gen s_score	- (Medical_es	ep_total - me	voesb) \ sq_est				Label	
	(4,520 M5481A)) values devel	(#543)					Notes	
	br z score								37
									41,740 6.15M
	. gen medical	decimal - Mec	fical_emp_tot	a1/1000				Memory	6M
	(4,520 mi##ing	y values genes	rated)						
	·							, ,	
	Command							*	
	br medical_deci	imal							

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)

Stata/SE 15.1 - C//User/Studio N/Dealt	oplianuus 2022/0: Partag Michaely Lee (19455-29th block-OLUTA		- 0 ×
File Edit Data Graphics Statistic	s User Window Help		
S ■ S U R · ⊥ · Z · Z	3 0 0		
Review T & X	45500 1 0.00 99.98	 Variables 	т # ×
1 iter commands have 0	71000 1 0.00 99.98	A filter variables	
A Commond or	75000 1 0.00 99.99	Neme	Label 0
1 con Villen State Uter	\$2500 1 0.00 99.99	Other succes	
1 tob Medical are total	71000 1 0.00 97.77 11770 1 0.00 98.88	formiture	
1 comm Medical and total	120100 1 0.00 99.99	mahanad	
4 mm mintatal a miniMedi	155000 3 0.00 99.99	Main source	
5 earn-marketel - mark Med	400000 1 0.00 100.00	Pace of two	
6 ann meil tatal + (Medical	505000 1 0.00 180.00	state code	
7 by med total		Lass of inco	
8 egen meanerig 1 mean(M.	Total 38,720 100.00	NSS	
9 egen (d. exp. + (d) Medical	emme Madical ave total	NSC	
10 gen zuscere + (Medical.ex	· · · · · · · · · · · · · · · · · · ·	MULT	
11 br z, score	Variable Obs Hean Std. Dev. Min Has	mintotal	
12 gen medical_decimal + Me		mantotal	
13 be medical, decimal	Medical_ex-1 30,720 042:0752 3507.759 0 50000	med_total	
14 gen log, medical, total + lo		meancrip	
15 br Medical, exp. total med	. epen ministal = min(Medical_exp_total)	sd_mp	
15 clear	anso martotal = mari Medical are totali	2,00016	
17 use "C10ses/Studio A/Des	, dan merone - met enner "ed-"ener)	n. medical_deci	- N
18 tab Medical_exp_total	. gen med_total = (Medical_emp_total-mintotal)/(maxtotal- mintotal)	log_medical	
19 summ Medical_exp_total	(4,520 missing values generated)		v
20 egen mintatal + min(Med		Properties	a x
21 egen martotal = max(Med	. br med_total	8 **	
22 gen med_total = (Medical		# Variables	
23 bi med_total	- shee meeter = meet meeter = she corri		mintotal
24 egen meanerig + mean(M.,	, epen as exp = as(Medica) exp tota);	Label	
25 egen sid, exp. + sid(Medical		5.94	Rut
26 gen zjiscere ir (Medicaljes	. gen I_score = (Medical_exp_total - meacemp) / sd_exp	Format	5.8.09
27 br zjscore	(4,520 missing values generated)	Value label	
28 gen medical_decimal + Mc			
29 to medical_decimal	. br i_score	* Data	
30 gen log_medical_total + lo	. men medical decimal = Medical ave total/1000	Linname	NSS CREATER - 24-3
	(4.12) missing values generated)	Labo	
		Variables	11
	- br medical_decimal		41,240
			6.53M
	. gen log_medical_total = log(Medical_exp_total)	Memory	64M
	[6,132 Biseldg Velues generated]	Sorted by	
		v	
	Command		
C1Dies/Studio #/Desktog/January 2022/0	Protag Misharehy Salaka S		CAP NUM OIR
.P Type here to search	🔿 🛱 🌍 🛷 📙 Sata 3 💽 🧶 👰 NVIII. Lecture 11, 👔 Recentioner State 5 🔡 Sata 55 (15) - C.O 📄 Data Later (Boosec 🔒 2010 - A 💿 💭	🚦 de 🎜 DNG	29-10-2021

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)

Med	cal,exp.total	[1]												
Neda	cal_es-1	med_total	1,10010	medical_de-1	log_medica-1								 Variables	
1	900	.0018	.0143249	.9	6.002395								 A File variables	
2	20	.00014	2063402	.02	2.996732								2 Name	Label
3	600	.0008	1110576		5.991465								Medical, exp.,	
4	600	.0012	0009046	. 6	6.27693								@ med,total	
\$	500	.0014	0350201	.7	6.55200								S 2,5COTE	
6	200	.0014	1412104	.2	5.290317								Ø metical, deci	
7	150	.0003	1737400	.15	5.010635								Ø log,medical,	
0	300	.0014	1361041	.1	5.703703								0 150	
5													C) Segment	
10	e00	.0008	1110576	.4	5.991465								C Hundeld	
1	IQ 220	.00044	1541953	.22	5.093420								C filment and	
2	150	.0003	1.1737400	.15	5.010635								O Stire men.	
3	300	.0004	1161041	.1	5.703700								0.44	
4	250	.0015	1414723	.25	5.521461								O Whithmade	
4	250	.0015	1416723	.25	5.521461								C) Surgery	
4	50	.0001	1900253	.45	3.912023								Medicine,rec	
7	850	.0017	.0017844	.45	6.745236								Med,rece,et	
	20	.00014	2030406	.03	3.401197								() Xay_ECG_EL	
1.5	450	.0019	0905190	-11	6.109240								[] Oher,dags	
10	150	.0013	1737400	.15	5.010435								C Dectors_surg	
11	0												Verables Snapsho	ets (
12	0		2113435										Properties	
10	150	.0007	1215950	.15	5.057333								+ Variables	
	322	.00044	- 1204111	.11	5.722022									Medical, pro
													Label	
2	120	00014	- 1412212		4 202422								1ar	long
		00018	- 1007947		4 43331								Lormat	51009
	135	00037	- 1776103	115	4 905275								Value label	
	120		- 1813313		4. 202423								Notes .	
	124		- 1800174	115	4 010314								Limane	NSS 798-10
	141		- 11424-4		4 141642								Label	
	140		- 1.42949										Notes	
	.500	.003											Variables	32
	1400	.0028	. 1.97079		7.244227								Observations	41,285
	*700	.0114	4.417994		*.***21								Sat	6 S3M
	160	.00173	.0042943	.14	6.756932									54M
6	250	.0005	1416723	.15	5.521461								Sorted by	
7														
	600	.0012	0609046	.4	6.19693									
9	100	.00034	1662259	. 10	5.192957									
											_			

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.