Handling Large-Scale Unit Level Data Using STATA Professor Pratap C. Mohanty Department of Humanities and Social Sciences, Indian Institute of Technology, Roorkee Lecture 18 Tabulation and Creation of New Variables in Stata – II

Welcome friends once again to this module of NPTEL on Handling Large-Scale Data Using Stata. We have been so far trying our best to give you the hands on experience with using the core dataset which we usually require for our everyday research. We require maybe for our PhD work, maybe for our paper and over last three, four lectures we have been trying to clarify the very basic commands of Stata which is usually very important for all the researchers who uses Stata.

So this lecture is a continuation of the previous one, where we have already discussed the prerequisite of unit level data in Stata. Where we discussed tabulation and creation of new variable with the help of Stata. In the last class particularly, I have shown you how to derive result based on the core dataset, results like I have explained you how to get tab and table results or tabulate versus table. The first and the very clear difference between these are to identify frequency as compared to percentages.

In table command I have already shown you that at best we can derive the frequency, only the frequency of the variables, whereas in case of tabulate or tabulation, we can get many more information as well, like percentages, even cumulative percentages, which are very important for researchers to start filtering the variables for further research.

(Refer Slide Time: 02:53)



	EX *		a 1 a	5 . % . ^			+ BKSP			
al - 👔 - 🖬 🗎 🗉			2 3							
T 3 × ZOW	P	q w	e r	r y		P P	ľļľ	·	Variables	,
e 0	Caps	a s	d	f g h	j k		· -		A Filter variable	es here
.c Ty	pe Shift	z	x c	v b	n m <	· · · · · /	† Shift		Name	Label
taset_N owner	Ctrl 🕷	Alt	· · ·		Alt 🛛	Ctrl +	F - F		Sample	Sample
10w			(A. 1/A			<u> </u>			State Region	State-Region
	49.33	50.67	100.00						State	State
									State_District	State-District
	48.25	51.75	100.00						Stratum	Stratum Sub-Stratum-no
	1 .								SubRound	Sub-Round
	55.56	41.44	100.00						SubSample	Sub-Sample
	4 560	786	1.346						FODSubRegi	FOD-Sub-Regio
	41.60	58.40	100.00						Second_Strat_	Second-Stage-S
	5 441	497	938						ent_no	Sample enterpri
	47.01	52.99	100.00						Properties	
	6 454	81	535						8	
	84.86	15.14	100.00						* Variables	
	7 205	197	406						Label	Sector
	51.48	48.52	100.00						type	str3
	9 80	60	140							
	\$7.14	42.86	100.00						Notes	
	Total 35,766	36,762	72,528						4 Data	
	49.31	50.69	100.00							
									Notes	
									Vanables	42
Comm	and								Coservations	72,528 18,54M
									Memory	64M
									Sorted by	
tice dataset, NSS78, block2 d Graphics Statistics Use Li · M · C C C			2 3 e r	⁵ 4 [%] 5 ^	6 ^{&} 7 * 8 u i c		Bitsp		- 20 ()	
tice dataset, NSS73, block2 d Graphics Statistics Use al. X 2 3 3 Y 2 X rc 0	tu Z m roo S Fab Caps	• • • • • • • • • • • • • • • • • • •	2 3 e r d	⁵ 4 [%] 5 ^ t y f g h	6 ⁸ 7 [°] 8 U i q j k	⁶ 9	- + = 8ksp - - - 0 -		 Variables Fiter variable 	E N 412 P 8/17/2 = 0
the dealer MSTI Look 2 d	Cops Shift		2 3 e r d x c	§ 4 % 5 ^ t y f g h v b	6 [@] 7 [°] 8 u i c j k n m <		. [+ _ [8ksp]]]] ↓ [0 .]] .] 0 .]] 0		Variables	es here Label
Inc. denset. NSTI Dock Ja Gapha Santra Use Inc. denset. NSTI Dock Ja Santra Use Inc. denset. NSTI Dock Ja Inc. denset. NSTI Dock Ja Santra Use Inc. denset. NSTI Dock Ja Inc. denset. Ja	Ctri 2	AR	2 3 e r d x c	§ 4 % 5 ^ t y f g h v b	6 [@] 7 [°] 8 u i c j k n m <	9 0 - 9 1 0 - 1 - 7 / cor - 7 /			Variables Variables Criter variable Name Sector	Al 279
Inc. detaset, NSST book 26 Graphics: Statistics: Use	Caps Shift Ctrl 2	Q AR	2 3 e r d x c	∫ 4 % 5 ^ t y f g h v b	6 ⁰⁰ 7 ¹ 8 U i a j k n m ²	9 0 - P 7 1 - P 7 1 - P 7 1 - P 7 1 - P 7 1	- Fr		Variables Variables Titter sanable Name Sector State_Region	Al 279
Contract (NST) Look2 d Contract (NST) Contr	Caps Shift Ctrl 2 7 2 205	 1 Q W a z Att 	2 3 e r d x c	§4 %s ^ t y f g h v b	6 ⁰⁰ 7 ¹ 8 U i a j k n m ²	(9 0 0 0	. [* _]eksp]/]] \ D .]		Variables Variables Variables Variables Variables Sample Sector State, Reyon State Samp Annone State S	A 3279
tor denser (NST) (bool 2 Sophic Statistic Use du	Image: Comparison of the company of the com	Q W A a s 2 Att	2 3 e r d x c	§ 4 % 5 ^ t y f g h v b	6 & 7 & 8 u i a j k n m <	[9] 0 [-	- 1 - Bitsp 1 - 1 - D - D - D - D - D - D - D - D		Variables Variables Variables Variables Variables Variables Sample Sampl	All 27 M
Constraint NSST Local 2 Constraint Status U Constraint Status Cons	Image: Control of the second	Q W A S Z AR 1 1 G A Z AR 2 J 7 48.52	2 3 e r d x c	\$_4_ % 5_ ^ ty fg_h vb	6 ²⁰ 7 ⁸ 8 10 i 6 11 8 11 8 11 8	(° 9) 0 °	- F		Variables Variables Teter sample Sample Seter State, Region State State, State, State	All 27 M
rec.deaset.NSS71 Jon224 Gopher Statistic Use *********************************	Image: Control of the contro	4 157 4 157 4 4 .52 4 4 .52	2 3 e r d x c 406 100.00	^S 4 [®] s n ty fgnh vb	6 ⁽² 7 ² 8 u i c j k n m 2 At 7	⁽ 9) 0 − P] (1 ; ; 1 , 2 , 1 / , 2 , 2 , 1 / , 2 , 1 /	. r		Variables Variables Variables Priter sympth Name Sample Sector State State, Ontoict South SubSound	 All 20 M All 20 M<
Inc. Anser, 1007 Jood 2 Copies Stores Control of the Control of th	Image: Solution of the	4 .52 4 .52 4 .52 5 .762	2 3 e r d x c 406 100.00 140 100.00 72,528 100.00	⁵ 4 [∞] s ^ t y f g n v b	6 ⁸ 7 [°] 8 Ju i G Jj k n m r At m	[9] 0 .	. 1 = 10kop 		Variables Variables Ceter single Sate Sate Sate Datiet Solar	All 2 Ph
A A	Image: Control of the contro	a 5 44.55 44.55 46.55 42.46 5.55	2 3 e r d x c 100.00 140 100.00 72,528 100.00	⁵ 4 [®] 5 [°] t y f g h v b	6 ⁽⁰ 7 [°] 8 u i 0 j k n m < At 7		_ + Bkp] _ + + _ T _ Shit + + _ fr		Variables Variables The variables Name Same State, Region State, Region State, Ontoict State, Ontoict State, Ontoict State, Ontoict State, Ontoict State, Ontoict State, Ontoict State, Ontoict State, Ontoict State, Ontoict	A 422 Ph 8/37/20 shere Label Sample Sector State-District State State-District State
Image: A second control of the second contro of the second control of the second control of the second contro	Image: Control of the contro	2 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	2 3 e r d x c 100.00 140 100.00 72,528 100.00	⁵ 4 [®] 5 ⁶ t y f g V b	6 ⁰ 7 ⁸ 8 10 i 0 11 k 11 m ≪ 14 m				Vanables Vanables Citer vanable Name Sereter Sate	es here Label Sample Sector State St
tor dance (NDT1 Joba 22 Singha Status Unit 	Image: Constraint of the	a 5 7 40 52 40 52 40 52 50 65 50 65	2 3 e r d x c 406 100.00 100.00 72,528 100.00	⁵ 4 [™] 5 ∩ t y t g h v b	6 × 7 × 8 10 i 0 j k n m 2 At 8				Variables Variables Inter-sample Sample Sate Sample Sate Context Sate Sate Context Sate Sate Context Sate Sate Sate Sate Sate Sate Sate Sate Sate Sate Sate Sate Sate Sate Sate	es here Label Sample Sector State-Region State State-District Statum Sub-Sample Sub-Sample FOD-Sub-Region Sub-Sample FOD-Sub-Region Sub-Sample FOD-Sub-Region Sub-Sample FOD-Sub-Region
See Amount (NOT) Junck 2: See Amount (NOT) Junck 2: Copies Strengs Use 1: If I and I	Image: Section 1 Image: Section 1<	1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0	2 3 e r d x c 406 100.00 140 100.00 72,528 100.00	Sats ty fgh yb	6 ^{[0} 7 [*] 8 [0] j k [n] m [4 [44] n	Image: p			Variables Variables Variables Variables Variables Variables Samp	Algorithm Algorithm
Image: A second control of the second contro of the second control of the second control of the second contro	Image: Second	AR 36, 72 41, 52 42, 66 36, 72 50, 69 200 200 200 200 200 200 200 20	2 3 e r d x c 100.00 140 100.00 72,528 100.00	⁵ 4 [∞] 5 [°] t y f g h V b	6 [®] 7 [®] 8 u ji 0 ji k n m ∢ At π				Vanables Vanables Vanables Vane Sate	Algorithment A
	Image: Control of the contro	1 0 9 W 0 V 0 V 0 V 0 V 0 V 0 V 0 V 0 V	2 3 e r d x c 406 100.00 72.528 100.00 Total 100.00	⁵ 4 [™] 5 [°] t y f g h y b	6 × 7 × 8	Image: point of the second			Variables Variables Variables Variables Sector Sate, Paylon Sate, Paylon Sate, Ostocc Sate,	es here Label Sample Statum Sub-S
Image: Section of the sectio	Image: Control of the second	1 0 4 0 4 0 4 0 4 0 5 0 5 0 5 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6	2 3 3 e r d 406 100.00 140 100.00 72.528 100.00	Sats ty fgh vb	6 ⁸ 7 ⁶ 8 u i a j k n m - At P		L L L L L L L L L L L L L L L L L L L		Variables Variables Variables Varies State Colored State C	Annovement A
Image: Section 1000 12 000 12 000 12 000 12 000 12 000 12 000 12 000 12 000 100 1000 100 1000 1000 100 100 100 100 100 100 100 100 100 100 100 100 10	Image: Second	2 4 4 4 5 5 4 4 5 5 5 5 5 5 5 5 5 5 5 5	2 3 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5	⁵ 4 [∞] 5 [°] t y f g h ∀ b	6 27 8 10 11 0 17 14 17				Variables Variables Variables Same Jegon Same Jegon Sam	Anno and a second and a se
see anter (NOT) Social 2 see anter (NOT) Social 2 see anter (NOT) Social 2 see anter (NOT) see anter (NO	Image: Control of the second	2 2 4 4 5 5 4 5 5 5 5 5 5 5 5 5 5 5 5 5	2 3 3	⁵ 4 [™] 5 [°] t V f 9 h V b	6 ⁰ 7 ¹ 8 u i c j k n n r At n	Image: point of the second			Variables Variables Variables Sample Sampl	 In the second sec
Image: Section of the sectio	Image: Control of the second	2 2 4 4 4 4 4 4 4 4 4 4 4 4 4	2 3 4 4 4 100,00 100,00 100,00 100,00 100,00 100,00 100,00 100,00 100,00	Sats ty fgh yb	6 ⁸ 7 ⁸ 8 u i a j k n m - At P		Shift		Variables Variables	 Image: A second s
Image: Sector (NOT) 2002 2 Image:	Image: Second	Image: Control of the second	2 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	5 4 ∞ 5 ^ t y f g h ∀ b	6 27 8 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1				Vanables Vanables Vanables Vanable	In the second
incenter (NDT) local 2 incenter (NDT) local 2 incenter (NDT) local 2 incenter (NDT) incenter (NDT) incent	Image: Control of the contro	0 1 1 1 1 1 1 1 1 1 1 1 1 1	2 3 3 d d x 4 66 100.00 100.00 72.528 100.00 72.528 100.00 100.00 100.00 100.00 100.00 100.00	⁵ 4 [∞] 5 ² t V f g h V b	6 0 7 8	Image: point of the second	L L L L L L L L L L L L L L L L L L L		Vaniables Vaniables Cline or copy State St	Annual Control Co
to change (NOT) Junit2 (Capito Simons Union 1 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Image: Control of the contro	Image: Control of the second	2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	Satis ty fgh yb	6 ⁶ 7 ⁶ 8 u i a j k n m ² M P		V V V V V V V V V V V V V V V V V V V		Vanables Vanables Vanables Seren	Comparison of the second
Image: Sector (NOT) Joint 2: Image: Sector 2: Image	Image: Second	Image: Constraint of the second sec	2 4 3 4 7 4 7 4 8 4 8 4 8 4 8 4 8 4 8 4 8 4 8	∑ 4 ∞ 5 ° 1 y 1 9 h V b	6 27 8 1 1 0 1 K n m 1 At 7				Vanidas Vanidas Nore autorita Sart, Feyn Sart, Feyn Sar	Control Con
conserved (NOT) (Sec22 of a second se	Image: Section of the sectio	0 1 0 1 0 1 0 2 1 0 2 2 4 0 2 2 4 4 50 0 2 4 50 10 10 10 10 4 50 0 0 10 10 10	2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	⁵ 4 [∞] 5 ² t 7 f 9 h 7 b	6 0 7 8		Shift		Vaniales Vaniales Conservations State Stat	 Control Control C
to change (NOT Jacob 2 Copens Strong Use	Image: Control of the contro	Image: Control of the second	2 4 3 4 4 4 4 4 4 100 00 100 00 1	S 4 № 5 r t y f g h y b	6 8 7 8 8		Shift		Vaulas Vaulas Vaulas Vaulas Sere	 A second s

Stata/SE 15.1 - G\practice_dataset_NSS7	block2.dts	Tools 9	•				- 0 - X
File Edit Data Graphics Statisti	cs User Esc	· •		2 3	\$ 4 % 5 6 % 7 8 9 0 - + = Bksp		
S 🛛 😂 🗊 🕞 - T - 🤾 - 🕄	Tal	, q	w	e r	It y u i o p []] , Del		
Review T # x	colum Ca	05		l d		Variables	Ťů×
 Filter commands here 		0	1, 1,	1. T		Titer variabl	es here
Command Jc Jc use 'G\practice_dataset_N	ovners		2	X C		Sample	Sample
2 ta b2_q204 Sector	Cr	"	AIT		Alt 10 Cm	Sector	Sector
3 ta b2_g204 Sector, row 4 ta b2_g204 Sector, row role	1	49.33	50.67	100.00		State_Region State	♦State-Region♦ State
5 ta b2_q204 Sector, row col		63.23	03.19	63.21		State_District	♦State-District●
	2	48.25	51.75 12.39	100.00		Stratum	Stratum
						SubStratumine	Sub-Stratum-no
	,	0.01	0.01	0.01		SubSample	Sub-Sample
	4	41.60	58.40	100.00		FODSubRegi.	FOD-Sub-Region Hamlet crown/Su
		1.57	2.14	1.86		Second_Strat.	Second-Stage-Str
	5	47.01	52.99	100.00		ent_no	Sample enterprise +
		1.23	1.35	1.29		Properties	φ×
	6	84.86	15.14	100.00		 Variables 	
		1.27	0.22	0.74		Name	Sector
	7	51.48 0.58	48.52	100.00		Label	Sector str3
						Format	\$31
	у	0.22	42.86	0.19		Value label	
	Total	49.31	50.69	100.00		4 Data	
		100.00	100.00	100.00		Filename	practice_dataset_NS!
						 Notes 	
						Variables	42
	Command					Size	72,528 1854M
						Memory	64M
8) (2 0) 🛷	0						₩ 8/17/2020
StatuSE 151 - Gifkample data setda Fie Edit Data Gaphics States	s User Wind	ow Help					8/17/2020
Stauffe 151- Gilampie dra getdu File fot Data Gapher Stanti- Bie 🖉 Die is - H - H - M Review 7 2 x	B. User Wind	енер 2.71	97.29	100.00		Variables	- 0° - X
StaarSE 151 - GManagie data gestuu Fare fon Data Gaphica Sansh Inde connauda here III	CS User Wind	0w Help 2.71 2	97.29 114	100.00		Variables Fiter variable	x 177/2020
StaarSt 151-GKBangie data gestuu File fon Data Gaphica Sansh Inder connand here Connand je	B User Wind	0 Help 2.71 2 1.72	97.29 9114 98.28	100.00 116 100.00		Variables Titles variable Name	
Claure 11 - Champe en uneren Re for Das Graphs Sonth Re for Das Graphs Sonth Champe - L - M - S For command terr (Command terr), Champe - Command terr (United terr), Champe - Command terr), Champe - Command terr (United terr), Champe - Command terr), Champe - Command terr, Champe -	CS User Wind	0w Help 2.71 2.72 764	97.29 97.29 114 98.28 35,002	100.00 116 100.00 35,766		e Variables Filter variable Name location ente. b0.2011	T 0 × 17/2000
Consult 131 - Of complete and particle Ine test Data Conjunct Status Total Conjunct Status Total Consult Test Consult Test Consult Test Consult Test L	C Uter Wind	0w Help 2.71 2 1.72 764 2.14	97.29 97.29 114 98.28 35,002 97.86	100.00 116 100.00 35,766 100.00		Variables Filter variable Name location, enter b2,2211 enterprise,ly-	T U X
Constitution of the second sec	C User Wind C User Wind Casual Total	ov Help 2.71 2.71 1.72 764 2.14	97.29 114 98.28 35,002 97.86	100.00 116 100.00 35,766 100.00		Variables Variables Variables Variables Variables Variables Variables variable variab	T Q X17/200 T Q X r here Label Location of the en. Enterprise type du. Enterprise type du. Enterprise type du. Netro of Increation
Clanced 131 Ottomore encouncer ine for Data Graphic Statistics in the Carl Data Graphic Statistics in the Carl Data Graphics Statistics in the Carl Data Graphics in t	D O O	Cov Help 2.71 2.71 2.11 2.14 2.14	97.29 114 98.28 35,002 97.86	100.00 116 100.00 35,766 100.00		Variables Filer sangle Name location enter b0 2011 enterprise ly- initial year nature openin nature constr. main	T Q X1/2000
Constant 132 Offensele and vertee The fiel Data Gaping South The fiel Data Gaping South The fiel Data Gaping South The constant field Constant (N) Constant (N) Constant (N) Use 200 Meditor are	D User Wind C User Wind C Casual Total -> Sector = Key	2.71 2.71 2.71 2.71 2.1.72 764 2.14	97.29 114 98.28 35,002 97.86	100.00 116 100.00 35,766 100.00		Variables Intervanado Intervanado Nane Iocation ente b2,2011 enteprise,ty- initial/year nature.oporta iocomputer ioennet	T & A17/200 T & X I &
Consect 131 Conserve and sectors Iner for Data Capitor Statement Conserve and Statement Conserve and Statement Sector Sector Sector Statement Se	C User Wind C User Wind C Casual Total -> Sector = Key frequence	2.71 2.71 2.1.72 764 2.14	97.29 114 99.28 35,002 97.86	100.00 116 100.00 35,766 100.00		Variables Variables A file and Rane Notion, ente b,2,011 enterpoint, initial year antra year antra year antra year antra year	AU1/200
Control 131 Collampie dan united For 160 Data Copies Sonio For 160 Data Copies Sonio For 160 Data Copies Sonio For 160 Data Copies Copies For 160 Data Copies	Uker Wind Uker Wind Casual Total Total Key frequent zow percei	ом Неір 2.71 2.73 1.72 764 2.14 2 гу тсяре	97.29 114 98.28 35,002 97.86	100.00 116 100.00 35,766 100.00		Vanishes Interview	I a stringe stringer stringer stringer light stringer
Common Control Co	C User Wind Casual Casual Total Very Frequence Compercent	w Help 2.71 2 1.72 764 2.14 2 y y	97.29 114 99.28 35.002 97.86	100.00 116 100.00 35,766 100.00		Variables Variables	O X 11/200 Support Section 2017 S
Canada 113 Offennerie and unter The feel Oau Gapina Santa Canada 119 - L	Casual Casual Total Casual Casual Total Casual Casual Compension Casual Compension Comp	00 Help 2.71 2 1.72 764 2.14 2 1 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2	97.29 114 99.28 35.002 97.86	100.00 116 100.00 35,766 100.00		Variables Variab	O I III/000 O III/000 O III/000 O III/000 O IIII/000 O IIIIIIIIIIIIIIIIIIIIIIIIIIIIII
Clause 131 Colonary en ancarce The fee Dua Coupe Stere The Stere	Casual Casual Total -> Sector = Key frequent row percent	v Help 2.71 2 1.72 764 2.14 2 v v tage Did the end seceiv assistance	97.29 114 99.28 35.002 97.86	100.00 116 100.00 35,766 100.00		Variables Variables Inter-analytic Inter-anaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa	C C C
Comment into Course connected for the Course Course Course for the Course Course Course for the Course Course	C User Wind C User Wind C Casual Total Total Total Exercise Key Exercise Nature of	ov Help 2.71 2 1.72 764 2.14 2 Uid the ent receive assistance querenent	97.29 114 99.20 35.002 97.86 100 the during	100.00 116 100.00 35,766 100.00		Vanables Vanables Information Informatio Informatio Informatio Informatio Inf	A 917/200 A 917/200 A 917/200 A 91/200
Common 111 of Changel even setter Ine for Data Gapha Sonth Ine for Data Gapha Sonth Command terr Command terr Command terr Ine Command terr Una Common Intern Una	Casual Casual Casual Casual Control Casual C	vv Help 2.71 2 1.72 764 2 vv itage bid the ent receive asistance V vv	97.29 114 99.28 35.002 97.86 37.86 4 doiing tree N	100.00 116 100.00 35,766 100.00		Vanishes	Strikes Image: Strike Strike Image: Strike Strike Strike Image: Strike Strike Strike Strike Image: Strike Strike Strike Strike Strike Image: Strike Strike Strike Strike Strike Strike Image: Strike Strik
Common 1131 of Chammele care set of the Congress State of the	User Wind Casual Total Total Sector = Key frequence zow percent Nature of operation perennial	Netp 2.71 2 1.72 764 2.34 2 Uid the ent receiv assistance government last Y 447	97.29 114 99.20 55.002 97.86 87.86 0 during httee N 36.017	100.00 116 100.00 35,766 100.00 Total		Variables variagramma description variagramma var	A 11/1/200 A 11/1/20
Control 11 Collampic data set of the for Out Coupes Some in for Out Coupes Some in the Out Coupes Some in the Control I I I I I I I I I I I I I I I I I I I	User Winds User Winds Casual Total Total Total Zector = Key Irequess Irequess percental	2.71 2.172 2.172 764 2.14 2 Did the end receiv assistance governmen last ti 1.23	97.29 99.29 99.28 99.28 99.28 99.28 99.28 99.28 99.28 99.28 99.29 90.29 90.20	100.00 115 100.00 33,765 100.00 Total 36,664 100.00		Vanishes	A 91/200 A 91/200 A 91/200 A 9 x
Common 111 of Charge conserver of the fee Out Couple Couples South Couple Couples South Couples C	a Une Wood Wood Wood Wood Wood Wood Wood Woo	<pre>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>></pre>	57.29 114 99.20 35.002 97.86 297.86 1000 1000 1000 1000 1000 1000 1000 10	100.00 116 100.00 35,766 100.00 701al 35,464 100.60		Variables Variable	A 917/200 A
Common Charles Conference under In the	a Lee Wede a Lee	View Prop 2.71 2.72 74 1.72 2.14 2.1	97.29 98.29 95.002 97.86 297.86 207.8	100.00 114 30,764 30,764 100.00 70041 30,664 36,664 36,664 32,25 100,00		Vanabes Vanabes Vanabes Vanabes Vanabes vana vanabes vana	AUX/200
Control 11 Control Con	a Low Word a Low Mord a Low Word a Low	2.71 2.71 2.72 1.72 2.74 2.14 2.14 2.14 2.14 1.15 1	97.29 97.29	100.00 116 10.00 55.766 100.00 701al 100.00 701al 100.00 701al		Vanishes Vanishe Va	A 401/200 A 401/200 A 3 x A 3 x A 3 x A 3 x A 4 x </th
Connect 111 of Connect connected for 101 of Connect connected for 101 of Connect connected for 101 of Connect connected for connected term for connected term for Connected term for Connected term the start of Connecte	a Loe Wood a Loe Wood a Doe	N M(2) 2 7 2 7 7 2 7 2 7 2 7 2 7 2 7 2 7 2 7 2 7 2 7 2 1 2 1 1.3 1 1.3 1 1.3	97.29 97.29 114 99.28 35,002 97.86 100 the t during htree N 98.67 71 98.67 71 98.67 71 98.71 99.67	100.00 116 100.00 35.766 100.00 701al 36.666 100.00 225 20.00 20 20 20 20 20 20 20 20 20		Vanishins Vanishins Vanishins Vanishins Vanishins Vanishins vacoro, yeek va	A 91/200 A 91/200 A 91/200 A 9 × A
Common 111 of Character and and the first of the Character and the Charac	a don Wood and a constant of a	2.71 2.72 7.72 7.74 7.72 7.74 7.74 7.74 7.74	97.29 114 99.29 35.002 97.86 297.86 297.86 100 100 100 100 100 100 100 100 100 10	100.00 114 100.00 100.00 100.00 100.00 100.00 205 100.00 225 100.00 225 100.00 225 100.00		Vanables	A 91/200
Constant 13 Conserve encourse Format 13 Conserve encourse Format 19 Conserve encourse Conservation Conse	a Los Wood a Los	With the set of the s	97.29 114 99.20 25.00 97.46 97.46 97.46 100 100 100 100 100 100 100 100 100 10	100 00 116 126 120,000 120,0000 120,0000 120,0000000000		Variables Variables Information Name Dependence Properties Properti	A 11/200 A 11/200 A 11/200 A 12/200
Comment 111 Containing encouncer of the feel Data Coupers South Terms of the Coupers of the Coupers South Terms of the Coupers South Terms of the Coupers of the Coupers of the Coupers of the Coupers South Terms of the Coupers of	a Une Weeder a Une Weeder a Status a Status	N NQ 2.73 2 2.73 2 74 2.14 2 2 7 74 2.14 2 7 2 7 1 1.20 1.30 1.33 1.39 4051 4051	97.29 114 99.20 35.002 97.86 97.86 97.86 97.86 8 98.87 98.87 98.87 98.87 98.87 94.71 94.71 94.71 94.71	100.00 1155 100.00 35,766 100.00 701.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00		Vanishes	
Common 11 of Champion construction In the Outer Couples South Common 1 of Couples Couples Couples Couples Couples Couples Cou	a too Wood casual totalt	2 71 2 72 744 2 74 2 74 1 72 744 2 74 1 72 1 72 1 74 1 75 1	97.29 114 99.25 97.86 97.86 97.86 97.86 97.86 97.86 98.07 98.07 98.07 98.07 98.07 98.07 98.07 98.07 98.07	100.00 114 35,744 100.00 700.00 205.0		Vanishes	
Stored State Concerning of the second state of	a Loo Word a Loo	7 ref 7 ref 7 ref 7 ref 7 ref 1	97.29 13.4 33.002 97.86 297.86 297.86 10.001 10.0001 10.0001 10.0001 10.00000000	100,00 116 100,00 33,766 100,00 36,664 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 72 255 100,00 725 100,000 725 100,0000 725 100,000 725 100,000 725 100,000 725 100,0000 725 100,000 725 100,000 725 100,0000 725 100,0000 725 100,0000 725 100,0000 725 1000000000000000000000000000000000000		Variables	A 11/200 A 11/200 A 11/200 A 11/200 A 14 12 A 14

Two-way table we already explained. To explain those things, I will show you once again how we operated the two-way one, then only we can discuss the three-way table. The two-way table, what does it mean? It is also called cross tabulation, two variables how they are related, maybe they are correlated, maybe one is dependent variable, one is independent variable and how you wanted to establish the relationship between two variables the very basic two-way table can able to clarify it.

So let me just open one dataset for you which we wanted to use, is the following. I have explained you for the NSS 73rd in the last class. Let me open for you once again. So the two

table we wanted to show and its relationship through the tabulation command is simply typing tab or only ta that I have shown you in the last class. I think in the last class I have shown you already how to go for it. And like this let me proceed. Here one variable I wanted to use is type of ownership b2_q204 with that by sector. Tab, so I will start with type of ownership here. Then I will also use another variable since I am interested now to discuss two-way table. So let it be sector here.

For a simple result without any percentages, we can simply enter. We will get the frequency with two variables and their cross relationship between these two variables. How each of the variable with their different codes are overlapping, like 1 here, then 1 there, and 1 stands for, I think already I clarified, 1 is rural and 2 for urban. Similarly, in the type of ownership, we have already discussed proprietary male, proprietary female, then proprietary maybe by joint venture, proprietary by SHG Group. These are very different code. You can find it from the original data from their layout also. We will also explain those things in a different format today. So this is one way we have discussed in the last class.

Just frequency and its number is cumbersome, in the sense that you may not get the right picture for analysis. You can at best say that 29,769 number are from rural areas and those were the male proprietor out of 72,528 number of unorganized enterprises as per the 73rd round of National Sample Survey, Government of India data. But this number probably you cannot compare with another number because these are not in percentages, these are in absolute number. You always require a percentage figure for better comparison.

So percentage you can derive, as I told you, by giving another addition to the command. Suppose you are interested for row wise percentages or column wise percentages; you need to add with another command to it that is row. So, row, so if you add it, it gives you the information like row wise percentages. Similarly, I have shown you all those things. So I am not going to spend much time today. You can have a revisit.

But one thing I just want to mention here, if you are not interested to come out with your result with the absolute frequencies, you only want to keep your result with their percentages, so in that case nofr is the right command. So you simply click there, no frequencies result, nofr. So row wise nofr or column wise nofr both. If you wanted to go by row wise, it will give you row wise or if you add both row and column with nofr. But you have to identify whether it is row wise or whether it is column wise. But if you want both, then you have to add row and column both. So if I add, it gives result like this, row wise as well as column wise.

Another aspect of looking at the data and its interpretation is through cell percentages. I think I have told you in the last class that these are the figures derived either by row total or by column total, but what about the absolute total, individual cell, single cell. I told you in the last class, single cell out of the total. So, in that case, you simply add cell command and that will give you the right result.

Let us come back to the three-way table. Two-way we discussed even some other additions to the two-way table and their interpretation we mentioned at large in the last class. Let's start with the three-way table. In the case of three-way table, why it is essential for a researcher. The three-way table gives some kind of comparison between two variables with respect to a third variable. The third variable when you are doing, it is always suggested to go by a sorting of the third variable. It sorts the third variable. If want to focus a particular variable, you start with that particular variable. It will sort with that particular variable in addition another two variables will be added.

In this case, three-way tables for categorical variable. One important aspect to be mention, when it is a scale variable or a continuous variable, this table can derive result, but it is very difficult to read, because scale, there are so many rows and columns will be derived in case of two-way or three-way comparison. So, Stata might tell you that there is too many information, so better to stick to categorical variable.

So, categorical variables limit your coverage of the data to particular group. So, the group comparison we have already shown it here, as given the variable type of ownership we have nine groups and regarding sector we have two groups. So, comparison is clearly visible and is understood better. But in case of continuous variable, it is too difficult to do it. And though we get the result, but interpretation is not going to be so apt.

So you cannot give tab a list of three variables to act on. Looking at three variables just tab three variables is not going to give you any result. So in that case, the right command is through by command. We would have to separate it through a command called by for each of the categorical variable. So we have given here with this right command, bysort of a particular variable, as I just

mentioned that you need to sort that variable and the right command is bysort of the particular variable then you add colon then tab, likewise we did for the cross tabulation.

Bysort variable, in this case let me clear this data and open another one for you. Now I am going to open the recoded data we have simplified already. I think this is the one. So recoded data where all the variables with their code, different names have been recoded and properly mentioned for our easy use. There are, like here, that time I could not able to read correctly. Later on I just read by this second labeling information, but the name was confusing.

Here I can read easily if they are registered or not, growth status or not, whether the particular enterprise has received assistance or not, whether the enterprise is using Internet or computer etc. nature of operation information also there. So it will be very easy to use. So let me use the command bysort. I need to look at the enterprise, I will sort it by sector but I will compare the status of the enterprise over last three years, where is that variable, nature of operation variable which just used.

So, nature of operation of the enterprise with respect to their assistance received. So I will compare whether nature of operation is having any relationship with assistance received, but I will sort it by sector. So I will start with bysort then I will start with that variable first by sector, isn't it? So I have added sector here. Then I will have to add a colon. This is a semi-colon, but I will have to add colon. Then the normal cross tabulation is going to be useful. Just ta is enough for us. What I wanted to look at how the nature of operation of the enterprises here is looked at through their assistance received. I think assistance received is given here. I will add, let it be by row wise percentage. So row wise percentage is here. After doing so, I derived the result like this.

So, it has divided the result into two parts. You just go little up. You will find. Here, it has divided into sector one. At the bottom it is sector two. At the top, it is sector one. So sector one in rural areas what is the relationship between their nature of operation, perennial, seasonal, casual. And in the urban areas how these are looked at through the context of whether the enterprise received assistance from the government during last three years. Sometimes we are interested to interpret through assistance received. If it is assistance received through government, we define them as formal assistance, formal credit.

So whether those enterprises who operate perennially or seasonally or casually, the total number of enterprises, whether they have received assistance or not. Look at the interpretation here. Out of 36,761 enterprises in urban areas, out of that we have divided into perennial, seasonal, we have derived row wise, first division is through assistance received or not received. 451 reported that they have received assistance, whereas 36,310 reported that they have not received any assistance.

Those operate perennially out of 36,461 in the urban area, only 1.23% reported having assistance from government sources, financial assistance from government sources. So is not it interesting for the researcher to note. Similarly, seasonally it is 1.33, casually that is 1.39. When you have already defined your variable very correctly, your interpretation is going to be very good. I am just showing you randomly how this is impacting our result and how you can derive the result.

So let us move on to our next set of explanation those will be going to be very interesting. So you can add as many variables as you need through by part of that particular command. However, the amount of output you will get can become cumbersome quickly if there are so many variables. So it is suggested that you stick to three variables. More variables generally there will be so many layers, sub-layer, then sub-layers, it will be too difficult to interpret.

(Refer Slide Time: 18:32)



So, three-way, four-way or five-way or more tables in Stata is always possible. The table command which I have already told you gives you the right result very quickly. Like table is the command with which you can produce such tables with the help of Stata. The simple command like table which simply give the variable's name, it gives all the results.

How we read those variables? The first variable is row variable; the second variable is column variable. Then the third variable whatever you have entered is called super column variable and the fourth one is called by variable, super row variable list. The variable you simply command, but the interpretation is generally by the way we interpret, it is by that command only you have to add by, there only. Without by it is not going to read that.

Then content with its frequency if you add, it gives the right result. One thing you need to note is that the third variable we use with table command is called super column variable. This, the list of variable attached with by options are called super row variable list, which is already underlined here or mentioned here. But it is always important to note that this gives frequency, not the percentages. So generally researchers use less table as the command.

(Refer Slide Time: 20:36)



Let us come to the interesting part of our lecture called creation of new variables. How to get new variables, create new variables with the help of Stata commands.

(Refer Slide Time: 20:49)



The original dataset variables may not necessarily give you what you need in the analysis. In this case, we often use or need variables that are transformations or combinations of other variables. So the original variable only gives the overview. For our analysis we need to transfer or combine the variable for our better interpretation. The most important Stata commands for those creation

of new variables are generate, replace, recode. We have already discussed replace and recode earlier, and we are discussing in the context of generating a variable. They are often used together, all those commands usually taken together for deriving a new variable.

(Refer Slide Time: 21:41)

Stata Variable Naming Rules-	
Begin variable name with a character.	
Use ONLY letters or numbers or the underscore (_).	
Do NOT use any other special characters in the variable name (such !@#\$).	n as
Variable names must be no longer than 32 characters.	
Use either camel case or snake case.	
Some suggestions-	
Use as short a name as possible; 8-10 characters or less is ideal.	
Choose a name that is as descriptive as possible.	
Avoid names that are the same as Stata names for commands functions.	or
🚳 - swayan 🧕	7

Some important rules behind creating new variables in Stata, so you begin the variable name with a character always and use only letters or numbers or the underscore. So, space in between two names or character is not going to be read by Stata. You have to either go by underscore or the snake based variable design you can also make it. You can start with a new name or new character with a capital letter. So that it can also read.

So another important rule is, do not use any other special characters in between or in the variable name such as exclamation mark or at the rate, hash or dollar. Those should be avoided, because Stata is not going to read these, not going to accept as the variable name. The variable names must be no longer 32 characters. It should not exceed 32 characters in the variable name.

So use either, as I just mentioned, snake based or camel based. Camel based as I told you, do not want to keep the space, you can start with a capital letter in the next separated name, but it has to be continuous with a higher capital letter. But in case of snake you need to give a space in between.

There are some suggestions. Use as short as possible for the name, usually 8 to 10 characters or less is the best way of naming the variable. Choose a name that is as descriptive as possible. Avoid names that are the same as Stata names for commands or functions.

(Refer Slide Time: 24:07)



So in order to understand the variable, we are supposed to go by generate and replace and there are specific commands for it and the primary commands for creating and changing variables are generate. Usually, we abbreviate it with gen or g with a comma replace. Replace is important because the earlier any kind of entries are there with the same name or even any database is there, it will not be confusing further. So which like other commands that can destroy the information and has no abbreviation like this.

So generate command creates a new variable using an expression that may combine constants, variables. We are going to discuss all those things, functions, arithmetical, logical operations. This is there in our slides we are going to show it. Replace can be used to create variables that are formed by performing arithmetic or logical operations on existing variables. We are going to show it.

(Refer Slide Time: 25:00)



It is very important to note that you should understand some of the meaning of equal sign and double equal sign, like the meaning of the single sign and double sign we are going to tell you. Use the single equal sign when you have to define variable value, whereas the double equal to sign, like here it is mentioned, the double equal to sign in the command if you are giving it while in the generate command when you are going to check a true or false function.

(Refer Slide Time: 25:45)





Filter commands here	seasonal	3	222 98.67	225 100.00	g	h	k	4	μ.	Variables	T B
Command Jrc use 'G\practice_dataset_N.	casuall	1.39	71 98.61	× 72 100.00						Name	Label Region
ta b2_q204 Sector ta b2_q204 Sector, row ta b2_q204 Sector, row nofr ta b2_q204 Sector, row nofr	Total	451 1.23	36,310 98.77	36,761						country popgrowth lexp gnppc	Country Avg. annual % gro Life expectancy at . GNP per capita
clear use 'G\Example data_set.d bysort Sector: ta nature_opr ta growth_status	. ta growth S enterpr	status tatus of the ise over the last 3 years	Freq	g. Percent	Cun.					safewater avg.gnpop InGNP	
clear sysue lifexp gen avg.gnpop +(popgrow gen InGNP+In(gnppc)		expanding stagnant contracting 9	24,32 28,88 6,55 9,37	14 35.18 13 41.77 18 9.48 18 13.56	35.18 76.95 86.44 100.00						
		Total	69,14	3 100.00						Properties	
										 Variables 	
	. sysuse 11: (Life expect . gen avg_g (5 missing	feexp tancy, 1998) mpop=(popgro values genera	(th+ gnppc)	/2						Name Label Type Format Value label Notes	gnopc GNP per capita Tioat %9.0g
	can locale	ala (mana)								4 Data	
	(5 missing	values genera	(ed)							Filename Label • Notes	lifeexp dta Life expectancy,
										Variables	8
	Command									Observations Size	68 117X
										244	
										Memory	64M

Like you can usually do that. I will also show you. Like creating a variable which is average of two variables. You wanted to get the average of two variables and you wanted to create another variable with average of two variables. So simply gen you can write and take a name, but better to combine with average underscore you can also take the name of that or simply name, this stands for a name, better to give a right name, because you are going to operate with average. So click the variable one or type the variable one plus variable two, close the bracket and divided by two. Slash two takes as divide. So it will show you.

Let me read out all those things and I will operate. Let us go by one example and in this particular data like you wanted to take the average of two variables, you have a continuous variable. If it is numeric one it will be good to average, but having categorical information averaging is not suggested unless the reason of averaging and your interpretation for it. So, better to stick to a continuous variable, for example, growth status. What is there in the growth status let me first check for you. So I simply wanted to check, if I just type like this growth status, it gives me information like it is also categorical: expanding, stagnant or contracting.

Look at, there is a nine entries. I have already discussed that, if any other entries are there, nine, triple nine, these as per the NSS data, this suggest that you please read those as missing. We have specific commands. Like number, what is this, number of first stage units. Let me just operate any two variables and its average. You can go by your own way. Otherwise, I will simply go by

the Stata directory. The directory it has better information. So let me clear this first, then I will start with the system use data.

Sysuse, then as I told you, lifeexp data we have. here population growth is there, then life expectancy is also there, you wanted to generate a variable. You simply write down gen you should take the name of that particular variable, let me take the name as avg, average underscore gnp and population growth. gnp per capita and population growth, just I wanted to take the average. How these continuous variables go by average number. So let it be average gnp, I am just using gnpop is equal to then bracket start, shift, then two variables, I have already said this plus this then bracket close then divided by two. So if you enter it, it defined you already, another variable with average genpop and it shows that there are five missing values generated. We will also discuss why missing values are generated in a short while.

Similarly, if you wanted to discuss about converting the variable into log of it, like usually the consumption expenditure or the income or the wealth type of variables are required to be converted because usually those data are not normal in nature. Those are usually skewed. Income of the population in the country is skewed. It skewed towards only 1% of or only 10% of the population occupies 90% of the total income in our country as per some of the reports. That means it is skewed.

But when you have used the skewed variable for your interpretation, it always gives biased result, because you have not scale down the variable, you have not made the variable normal. It will again bias your variable and its interpretation, other variable interpretation. So in that case you need to scale down the variable with any form of scaling. One important scaling down of that particular variable is through logarithmic transformation.

So generate, here we have used log of it, log of that particular within bracket of that variable, but I have used the gen, like in this case gnp per capita income is a continuous variable and it seems that it is not a normal variable, you need to scale it down. Then go by gen, you have create a variable, but how do you identify this is different than that of the original variable you have to attach with I somewhere, I stands for logarithm. You can say ln if it is natural logarithmic, ln you can underscore to it or you can simply write down capital GNP then just simply type equal to then ln, since we are taking the natural logarithmic function here, ln within bracket this variable then you close that bracket.

This variable is going to be converted into a proper scaled variable and it can be interpreted and used for further analysis. This is created. Here the variable has already been created as log gnp data. Similarly, for if you wanted to square a variable or cube a variable you can use the power, to the power two. Then that will give you a variable for sure. So I am not going to explain all those things. You can experiment. If there are any possible doubts, please note it down and ask to us. We will clarify for sure.

(Refer Slide Time: 33:54)

The replace command modifies existing variables in exactly the same way as *generate* creates new variables: gen *l* variable = \ln (variable) >replace I_variable = In (1) if I_variable == . /replace yr=yr-100 if yr >= 100- meaning it replaces 0,1,etc instead of 100,101 etc. 🎯 swayan 🧕

So, similarly, the replace command modifies existing variables in exactly the same way as generate creates new variables. Basically, here we are going to replace the variable with another way. Like generate variable we did log of that variable and log of this. Here, similarly, replace you take a name of it then log one. Basically why replace is important, I have already clarified in a point in the previous page in our slide that if one variable is by any reason turned out to be dot, like we converted some variable, but those are like I have taken log of it, but that has been converted as dot. How you can replace that to another one, because if you have missing values, it might be reducing your number of cases for the answer, for the interpretation or for the result. That may weaken your result.

Or like some variable for logarithmic transformation you have already an entry with dot, a lot of dot is not going to give you any result. So in that case try to change that to one, simple one. So lot of one is equal to zero. So it will convert a variable and at least the logarithmic function will be operated. So in that case you simply replace that variable with one, log of one. It will change that particular dot to one and it will operate your function.

Similarly, replace is useful. Look at the difference between replace and generate. Replace year but it has to be if it is till 100, if year is greater than equal to 100 that means it replaces zero one etc. instead of 100, 101 or any other value. So like if you are very clear with what you want 100 to be replaced and it replaces only zero and one, not above to that of 100, 101 accordingly. You can just go through and you will find out for sure. It is not that difficult and you can easily able to do it.

(Refer Slide Time: 36:34)



Let me further clarify what are other aspects of operations and expressions we need to do it. There are broadly arithmetic operations, logical operations and relational operations. So arithmetic operation add string, concatenation, subtract, multiply, divide, like power operations, the exclamation mark, not equal to or, and in logical operations, similarly, relational double equal to or exclamation with equal to that does mean not equal, please go through this you will clarify and try to operate the way we have entered logarithmic transformation or some greater than transformation. You try to operate and use these for your explanation. You will certainly have some doubts and we will be happy to clarity those later.

(Refer Slide Time: 37:46)



Let us come to creating a variable with if condition. If a generate command attach with an if condition or if a gen command has an if condition, the resulting variable will still exist for all observations. However, it will be assigned a missing value for observations where the if condition does not follow. You have an if condition and we have an example here. We will clarify accordingly.

Similarly, if a replace command has an if condition, observations where the if condition is not true will be left unchanged. One difference between gen and replace is that, in gen you are converting that particular variable. You are creating a variable with a new one. But in that case if your if condition is not right, that will generate a missing value. But in case of replace one, that will be left unchanged. That is very very important to be noted. This allows you to set variables to different values for different observations. We have example here. Let me go by that and clarify you.

(Refer Slide Time: 38:54)

In our example dataset, we have a variable called type of ownership of an enterprise which has 8 categories. If we want to create a new variable with only male entrepreneurs:
 gen maleEntrepreneur = 1 if ownership_type < 2 or == 1
 This command generates variable with values 1 for ownership type less than 2 or equal to 1 or the other gets missing.
 replace maleEntrepreneur = 0 if ownership_type > 1
 this command changes those missing to zeros.

In our example dataset, we have a variable called type of ownership, which I have already shown you, which has eight categories. If we want to create a new variable with only male entrepreneurs, so in that case generate male entrepreneurs, any name we will give, equal to if ownership type, that is the variable name given in the data, less than two or double equal to one. So ownership less than two or double equal to one, what does it mean. You wanted to create a male entrepreneur as one. These commands generate a variable with value one for ownership type less than two or equal to one or the other gets missing. Basically it is less than or two since we have limited already with one or less than two, if there are any other entries in the ownership type, there are other entries already I have shown you, so those will be converted into missing dot. Then later on we have to give another command for filling those dots.

But the replace command, if are going by the replace command with zero, like replace male entrepreneur equal to zero if ownership type is greater than one. Like all ownership type which has the value greater than one, it will be replaced with zero. So, these commands changes those missing to zero. Even if missing is there that will be also converting to zero. But generate is not going to do that. You can just practice, you will certainly get it. I have many aspects to deal with. So I am emphasizing other details as well. (Refer Slide Time: 40:57)



So another, I will just clarify, let me come to it. I will clarify this recode and then after recode we will take another lecture for another one. So, for recoding, variables are often not coded the way we want, often with too many categories or with values out of order. So with recode, like if it is out of order, the way you want to recode it is not there, so we can handle all recodes for a variable in a single step, simply recoding. We are going to show it. The core of the recode command is a list of the rules, each in parenthesis that tell it how a variable is to be recoded. They take the form like input value to output value. We will generate with a output with that particular recode. I am going to tell you the exact example.

(Refer Slide Time: 42:06)





The input value can be a single number. A single number, a list of numbers separated by spaces 1, 2, 3 is equal to 1, like there are codes given already 1, 2, 3, we want to recode it to 1 or a range of numbers specified with start to end. Like if we have a continuous series and codes are given like maybe 1 to 8 codes, in our ownership case, 8 codes are there, 1 to 8. You wanted to only code with converting those to 1 to 5 as 1, for example. So instead of writing 1, 2, 3 equal to 1 you simply write down 1/3 or in our case 1/5 equal to 1 it will recode. So output value will always be a single number that is 1, isn't it?

So in our example you can operate accordingly. So recode with that location enterprise that is, recode location underscore enterprise within bracket 1 is equal to 1. Basically you have kept 1 as 1, but others as 2. 2 to 6 as 2, you have given a code like 2/6 is equal to 2 then you can generate that variable, suppose you are recoding location, recoded you can also take a name redefined or location, another name you can define according to your own choice, we have given the name here as location.

I can recode for here like we have now, I have to go by that example the enterprise example. I can also recode that country, coding, I have to check that. It is probably problematic. Let us go by the variable. Let me clear first. Then I will open another one. I will open our original data. Then I will operate. So it is there.

If it is there, even from this also, I have already shown you, nature of operation is here, enterprise type, location, nature of operation, so here we wanted to show you location of that enterprise. So

let me check what type of codes are there in the location of that enterprise. I wanted to, there are six codes in the location of the enterprise given. I wanted to keep 1 as the code remain intact. For others I wanted to code it another one as 2.

So what I will do, I will go by the command recode then I have to discuss this location of the enterprise. I have to get this variable here. Then within bracket, what I will do, I will define as, there are 1 to 6 code, 1 as 1 that is basically within the household operation and outside the household operation. 1 stands for within the household and rest broadly categorize as outside household operations so far as enterprise operations is concerned.

1 equal to 1, isn't it, then bracket close, then within bracket another one, then either I will do one thing, I will do separate two, there are so many ways of doing. 2 is equal to 2, 3 is equal to 2, 4 is equal to 2, 5 is equal to 2, 6, there are so many ways of handling or in simple term 2, 3, 4, 5, 6 equal to 2 the way we have already shown you that we can do it or in simple way we will simply do 2/6 is equal to 2.

If I just close that and since I have recoded it I will have to generate a variable, otherwise it will be merging on the same variable which might be problematic because the same variable we may use it differently. So, better to generate a new variable with the new code. So you just generate as location, simple location which we, or you can simply type loc bracket close. So recode only allows numeric variables.

So this is already in, we will guide it, this data is in string format. I have already told you earlier that the string data does not read the numeric operation. You need to destring it or we will open. We have already destringed it. I will tell how, what do we mean by destringing. And now we have again retain the command recode with the same variables, the same operation, if you enter it gives the result.

You just look at and mark the changes carefully. Loc is not defined with the recode. It is interesting to note that when you have recoded as per our NSS data 1 stands for within household operation, 2 stands for outside, isn't it interesting to label those variables. So you have to label it. We have other slides to be discussed that how to label, 1 stands for within and 2 stands for outside. We are going to tell you in a short while. But I will finish the lecture with recoding today. From the next class we will discuss, but let me define the value labeling from the next

class. Egen then we have other things, but let me stop with this recoding till this and you may practice in between and raise your doubts if any. We will continue from the next class.

With this, thank you all.