

SESSION-6: STATA DATA ANALYSIS AND VISUALIZATION

Course detail: <u>http://julhas.com/jsedutech/stata-level-one.html</u> Mentor: Julhas Sujan

Session Outline:

- Append, Merge, Substring, Concat commands
- Example: how to make age groups from age
- Basic and advance statistical graphs bar, column, histogram, line, box

Lession-1: Append, Merge, Substring and Concat commands

• Append: Appending two datasets require that both have variables with exactly the same name. **Step-1:** Copy our example dataset and keep the variables name as it is. Then add some data in the new dataset and save as Dataset-2.

int 🔻 🔞 Photo Print										
Name	Date modified	Туре	Size							
🛛 Dataset-1	4/11/2021 9:55 PM	Microsoft Excel W	12 KB							
📭 Dataset-2	4/11/2021 9:55 PM	Microsoft Excel W	12 KB							

We want to append Dataset-2 with Dataset-1. So, the columns/ variables name should be exactly same as Dataset-1.

Excel Dataset-1: (Data from 1-30)

	Δ	B	C	D	F F		G	н
1	s/N	Gender	Age	Case definition	Inicial We	Current V	Site of Disease	Outcome
2	1	Female	32	Clinically diagnosed case	58.0	58.0	Extrapulmonary	Successful
3	2	Female	19	Clinically diagnosed case	40.0	40.0	Pulmonary	Unsuccessful
4	3	Male	58	Clinically diagnosed case	64.0	64.0	Extrapulmonary	Successful
5	4	Female	28	Clinically diagnosed case	25.0	25.0	Extrapulmonary	Unsuccessful

Excel Dataset-2: (New data from 31-35)

	А	В	С	D	E	F	G	н
1	S/N	Gender	Age	Case definition	Inicial We	Current V	Site of Disease	Outcome
2	31	Female	25	Clinically diagnosed case	29.0	29.0	Pulmonary	Unsuccessful
3	32	Male	60	Clinically diagnosed case	56.0	56.0	Extrapulmonary	Unsuccessful
4	33	wale	57	Clinically diagnosed case	50.0	50.0	Extrapulmonary	Unsuccessful
5	34	Male	55	Clinically diagnosed case	60.0	60.0	Extrapulmonary	Unsuccessful
5	35	Female	30	Bacteriologically confirmed	44.0	44.0	Extrapulmonary	Successful
7								

Step-2: Now you need to convert the two datasets in Stata .dta format (you can find the data import process in our session-4).

Name	Date modified	Туре	Size
Dataset-1	4/17/2021 3:16 PM	Stata Dataset	8 KB
Dataset-2	4/20/2021 12:08 PM	Stata Dataset	6 KB

Step-3: Open the Dataset-1 and you can see the following data with 30 records

· · ·	. 📑 🖻		Q Y	-								
		sn[1]		1								
	sn	gender	age	casedefinition	inicialwei~t	currentwei~t	siteofdisease	outcome	^	Variables		
1	1	Female	32	Clinically diagnosed case	58	58	Extrapulmonary	Successful		K Filter variables	here	
2	2	Female	19	Clinically diagnosed case	40	40	Pulmonary	Unsuccessful		Name	Label	Type
3	3	Male	58	Clinically diagnosed case	64	64	Extrapulmonary	Successful			S/N	bute
4	4	Female	28	Clinically diagnosed case	25	25	Extrapulmonary	Unsuccessful		v gender	Gender	strfi
5	5	Male	23	Clinically diagnosed case	45	45	Extrapulmonary	Successful		Ø general Ø age	Age	byte
6	6	Female	35	Clinically diagnosed case	68	68	Extrapulmonary	Successful		✓ casedefinition	Case definition	str32
7	7	Male	78	Clinically diagnosed case	50	50	Pulmonary	Unsuccessful		✓ inicialweight	Inicial Weight	byte
8	8	Female	35	Bacteriologically confirmed case	65	65	Extrapulmonary	Unsuccessful		currentweight	Current Weight	byte
9	9	Male	17	Clinically diagnosed case	46	46	Extrapulmonary	Unsuccessful		☑ siteofdisease	Site of Disease	str14
8	10	Female	26	Clinically diagnosed case	53	53	Extrapulmonary	Successful		✓ outcome	Outcome	str12
1	11	Female	27	Clinically diagnosed case	44	44	Pulmonary	Successful				
2	12	Female	16	Clinically diagnosed case	54	54	Extrapulmonary	Successful		<		
3	13	Male	55	Clinically diagnosed case	50	50	Extrapulmonary	Unsuccessful		variables Snapsho	ots	
4	14	Male	55	Clinically diagnosed case	57	57	Extrapulmonary	Unsuccessful		Properties		
5	15	Female	17	Clinically diagnosed case	35	35	Extrapulmonary	Unsuccessful		▲ Variables		
6	16	Female	25	Clinically diagnosed case	45	45	Extrapulmonary	Unsuccessful		Name	sn	
7	17	Female	67	Clinically diagnosed case	45	45	Pulmonary	Successful		Label	S/N	
8	18	Male	48	Clinically diagnosed case	49	49	Extrapulmonary	Successful		Туре	byte	
9	19	Male	65	Clinically diagnosed case	40	40	Pulmonary	Unsuccessful		Format	%8.0g	
0	20	Female	65	Clinically diagnosed case	42	42	Extrapulmonary	Unsuccessful		Value label		
1	21	Male	36	Clinically diagnosed case	64	64	Pulmonary	Unsuccessful		A Data		
2	22	Male	4	Clinically diagnosed case	17	17	Extrapulmonary	Unsuccessful		Frame	default	
3	23	Female	60	Clinically diagnosed case	45	45	Extrapulmonary	Successful		Filename	Dataset-1.dt	8
4	24	Female	25	Clinically diagnosed case	29	29	Pulmonary	Unsuccessful		Label		
15	25	Male	68	Clinically diagnosed case	56	56	Extrapulmonary	Unsuccessful		Notes		
6	26	Male	57	Clinically diagnosed case	50	50	Extrapulmonary	Unsuccessful		Variables	8	

Step-04: Go to Stata >> Data >> Combine Datasets >> Append Datasets

Stata/MP 16.0 - F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\Dataset-2.dta

File Edit	Data Graphics Statist	tics Use	r Window Help
ビ 🗄 🖨 📔	Describe data	•	0 - 0
History Filter cor	Data Editor Create or change dat) a ⊧	800-STATA-PC http://www.stata.com 979-696-4600 stata@stata.com
# Commar 1 append u 2 save "F:\S	Variables Manager Frames Manager Data utilities Sort	Þ	979-696-4601 (Tax) user 2-core Stata network license expires 20 Aug 2022: Serial number: 501609213901 Licensed to: www.carrotchou.blog
	Combine datasets Matrices, Mata langu Matrices, ado langua ICD codes Other utilities	⊧ age ¢ }	Merge two datasets Form all pairwise combinations within groups Append datasets Form every pairwise combination of two datasets d using "F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\Dataset-1.xlsx"

Step-5: Click on Append datasets and you can see the following screen and browse the second dataset and click on Ok button

Append datasets on disk to current d	ataset		
Filename of dataset on disk:			
'elopments\Stata Course\All Session	ns\Level-1\Sess on-6\ <mark>D</mark> a	taset-2.dta	Browse
Select an additional file			
List of files on disks:			
Options			
Name of new variable to mark result	s:		
Keep all variables from appending	g dataset		
○ Variables to be kept from appendi	ing dataset:		
			~
Do not copy value label definition	is from dataset on disk		
Do not convinctes from dataset o	n disk		
_ bo not copy notes norm dataset o	II UISK		

Step-06: Results

File	Edie M	our Data	Teels											
The second														
	i 🗸 📄 🗖		BUT	•										
		sn[1]		1										
	sn	gender	age	casedefinition	inicialwei~t	ei~t currentwei~t siteofdisease outcome					 Variables 			ņ
16	16	Female	25	Clinically diagnosed case	45	45	Extrapulmonary	Unsuccessful			K Filter variables	here		_
17	17	Female	67	Clinically diagnosed case	45	45	Pulmonary	Successful				Itabat	17.000	
18	18	Male	48	Clinically diagnosed case	49	49	Extrapulmonary	Successful			V Name	C/N	lype	
19	19	Male	65	Clinically diagnosed case	40	40	Pulmonary	Unsuccessful			le sn ⊡ son	S/IN Condex	byte	70
20	20	Female	65	Clinically diagnosed case	42	42	Extrapulmonary	Unsuccessful			iv gender	Gender	stro	70
21	21	Male	36	Clinically diagnosed case	64	64	Pulmonary	Unsuccessful			⊠ age	Age	byte	76
22	22	Male	4	Clinically diagnosed case	17	17	Extranulmonary	Unsuccessful			Casedefinition	Case definition	str32	76
22	22	Camala		Clinically diagnosed case	45	17	Extrapulmonary	Successful		_	Inicialweight	Inicial Weight	byte	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
2.5	2.5	Femble.	00	clinically diagnosed case	47	40	Exci opulationary	Juccessius		_	✓ currentweight	Current Weight	byte	%
24	24	Fendle	25	clinically diagnosed case	29	29	Pulmonary	Unsuccessful		_	✓ siteofdisease	Site of Disease	str14	%
25	25	Mate	68	Clinically diagnosed case	56	56	Extrapulmonary	Unsuccesstul		_	✓ outcome	Outcome	str12	%
26	26	Male	57	Clinically diagnosed case	50	50	Extrapulmonary	Unsuccessful						
27	27	Male	55	Clinically diagnosed case	60	60	Extrapulmonary	Unsuccessful			K Mariahlan Casashi			,
28	28	Female	30	Bacteriologically confirmed case	44	44	Extrapulmonary	Successful			variables shapshi			
29	29	Female	14	Clinically diagnosed case	46	46	Extrapulmonary	Successful			Properties			ц.
30	30	Female	18	Clinically diagnosed case	48	48	Extrapulmonary	Successful			▲ Variables			^
31	31	Female	25	Clinically diagnosed case	29	29	Pulmonary	Unsuccessful			Name			
32	32	Male	60	Clinically diagnosed case	56	56	Extrapulmonary	Unsuccessful			Label			
33	33	Male	57	Clinically diagnosed case	50	50	Extrapulmonary	Unsuccessful			Туре			
34	34	Male	55	Clinically diagnosed case	68	68	Extrapulmonary	Unsuccessful			Format			
35	35	Female	30	Bacteriologically confirmed case	44	44	Extrapulmonary	Successful			Value label			
											Notes			
											4 Date			

Stata Commands:

use "F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\Dataset-1.dta" append using "F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\Dataset-2.dta"

Merge

Merging two datasets require that both have at least one variable in common (either string or numeric). If string make sure the categories have the same spelling (i.e. country names, etc.). The common variables must have the same name.

Example: We want to add 30 BMI results for each of the patients in our first Dataset. So we need to keep the one unique ID in both datasets.



Step-1: Keep the Dataset-1 as it is. Prepare sencond dataset as:



🔢 Stata/MP	16.0 - F:\Skill Development	s\Stata Course\All Sessions\Level-1\Session-6\Append-Dataset-1.dta		– 8 ×
File Edit D	ata Graphics Statistics	User Window Help		
ビ 🗄 🖶	Describe data			
History	Data Editor	>,	Variables	тџ×
S Filter cor	Create or change data	· (R)	Filter variables here	
# Comman	Variables Manager	// //	Name	Label
1 use "E:\Sk	Frames Manager	stics/Data Analysis StataCorp	sn	S/N
	Data utilities	luochenzhimu.com	gender	Gender
	Sort	- Parallel Edition College Station, Texas 77845 USA	age	Age
	Combine datasets	Merge two datasets Merge two datasets	casedefinition	Case definition
	Matrices, Mata language	e Form all pairwise combinations within groups ×)	inicialweight	Inicial Weight
	Matrices ado Janguage	Append datasets	currentweight	Current Weight
	ICD codes	g 2022:	siteofdisease	Site of Disease
	ico codes	Licensed to: www.carrotchou.blog	<u> </u>	Outcome
	Other utilities	China		
	No	otes: 1. Unicode is supported; see help unicode advice. 2. More than 2 billion observations are allowed; see help obs advice.		
		 Maximum number of variables is set to 5000; see help set_maxvar. set.iskill Developments\Stata Course\All Sessions\Level-3\Session-6\Append-Dataset-1.dta* 		

Step-3: Follow the below steps:

- Select One-one on key variables
- o Key variables as "sn"
- Browse the Dataset-2

😑 merge - Merge datasets	-		×						
Main Options Results									
Type of merge									
One-to-one on key variables									
Many-to-one on key variables (unique key for data on disk)									
One-to-many on key variables (unique key for data in memory)									
O Many-to-many on key variables									
One-to-one by observation									
Key variables: (match variables)			_						
sn			\sim						
Filename of dataset on disk:			-1						
F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\M	erge	Browse.							
			-						
? C D K Car	ncel	Sub	mit						

Click on the Ok button and you can see the following results

Stata/MP 16.0 - F:\Skill Developm	nents\Stata Course\All Sessions\Level-1\Session-6\Append-Dataset-1.dta			– 8 ×
File Edit Data Graphics Stati	istics User Window Help			
🍯 🔛 🖨 🚇 👁 × 🖬 × 📓 ×				
History T A X	979-696-4600 stata@stata.com 979-696-4601 (fax)	^	Variables	тů×
Commands inter Command Control of Control use FXSUIDevelopments merge th Sn using FXSUI	Single-user 2-core Stata network license expires 20 Aug 2022: Serial number: 50(400213001 Licensed to: waw.carrotoku.blog Olna Notes: 1. Unicode is supported; see help unicode_motice. 2. Nore than 2 billion observations are allowed; see help ubs_motice. 3. Nationm number of variables is set to 5000; see help ubs_motice. 3. Nationm number of variables is set to 5000; see help ubs_motice. 4. use "F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\Verge-Dataset-1.dta" merge 1:1 a using "F:\Skill Developments\Stata Course\All Sessions\Level-1\Session-6\Verge-Dataset-2.dta" Result <u># of obs.</u> <u>networks 0 (secrema)</u>		Name Name sn egender egge casedefinion inicialweight currentweight siteofdiases outcome built imerge	Label S/N Gender Age Case definition Inicial Weight Current Weight Site of Disease Outcome BMI
		~		
	Command	4		

Step-04: results

File	Edit	View Data	Tools						
1	3 🗸 📑	🗄 🖨 🖣	ĩ6 🔍 🍸	-					
		sn[1]		1					
					1				
	sn	gender	age	casedefinition	inicialwei~t	currentwei~t	siteofdisease	outcome	bmi
1	1	1 Female	32	Clinically diagnosed case	58	58	Extrapulmonary	Successful	19.4
2	:	2 Female	19	Clinically diagnosed case	40	40	Pulmonary	Unsuccessful	29.1
з	:	B Male	58	Clinically diagnosed case	64	64	Extrapulmonary	Successful	30.1
4	4	4 Female	28	Clinically diagnosed case	25	25	Extrapulmonary	Unsuccessful	25.6
5		5 Male	23	Clinically diagnosed case	45	45	Extrapulmonary	Successful	22
6		5 Female	35	Clinically diagnosed case	68	68	Extrapulmonary	Successful	23
7		7 Male	70	Clinically diagnosed case	50	50	Pulmonary	Unsuccessful	21
8	1	B Female	35	Bacteriologically confirmed case	65	65	Extrapulmonary	Unsuccessful	18
9	9	9 Male	17	Clinically diagnosed case	46	46	Extrapulmonary	Unsuccessful	25.5
10	10	9 Female	26	Clinically diagnosed case	53	53	Extrapulmonary	Successful	24.4
11	1:	1 Female	27	Clinically diagnosed case	44	44	Pulmonary	Successful	19.4
12	13	2 Female	16	Clinically diagnosed case	54	54	Extrapulmonary	Successful	29.1
13	1	3 Male	55	Clinically diagnosed case	50	50	Extrapulmonary	Unsuccessful	30.1
14	14	4 Male	55	Clinically diagnosed case	57	57	Extrapulmonary	Unsuccessful	25.6
15	1	5 Female	17	Clinically diagnosed case	35	35	Extrapulmonary	Unsuccessful	22
16	10	5 Female	25	Clinically diagnosed case	45	45	Extrapulmonary	Unsuccessful	23
17	17	7 Female	67	Clinically diagnosed case	45	45	Pulmonary	Successful	21
18	1	3 Male	48	Clinically diagnosed case	49	49	Extrapulmonary	Successful	18
19	19	9 Male	65	Clinically diagnosed case	40	40	Pulmonary	Unsuccessful	25.5
20	20	9 Female	65	Clinically diagnosed case	42	42	Extrapulmonary	Unsuccessful	24.4
21	2:	1 Male	36	Clinically diagnosed case	64	64	Pulmonary	Unsuccessful	23
22	23	2 Male	4	Clinically diagnosed case	17	17	Extrapulmonary	Unsuccessful	21
23	2	B Female	60	Clinically diagnosed case	45	45	Extrapulmonary	Successful	18
24	24	4 Female	25	Clinically diagnosed case	29	29	Pulmonary	Unsuccessful	25.5
25	25	5 Male	60	Clinically diagnosed case	56	56	Extrapulmonary	Unsuccessful	24.4
26	20	5 Male	57	Clinically diagnosed case	50	50	Extrapulmonary	Unsuccessful	19.4

Lession-2: Variable groups - age

Example:

gen ageGroup = age recode ageGroup (0/17 = 0) (18/34 = 1) (35/51 = 2) (52/64 = 3) (65/140 = 4) label define ageGroup 0 "0-17 Years" 1 "18-34 Years" 2 "35-51 Years" 3 "52-64 Years" 4 ">= 65 Years" label values ageGroup ageGroup tab ageGroup

Lession-3: Statistical graph

- **Bar:** graph bar (mean) ageGroup
- **Column:** graph hbar ageGroup
- **Pie**: graph pie, over(gender)



• Dot: graph dot (mean) age (mean) bmi



- Line: twoway (line age bmi), line bmi ageGroup
- Histogram: histogram age



Next Session:

- Article writing style and Stata result input to your article
- Descriptive data analysis

Good Luck!

If you need any support, don't heasitate to let me know.