

Lecture 3: Data cleaning

BIOSTAT 212, Summer 2017



Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

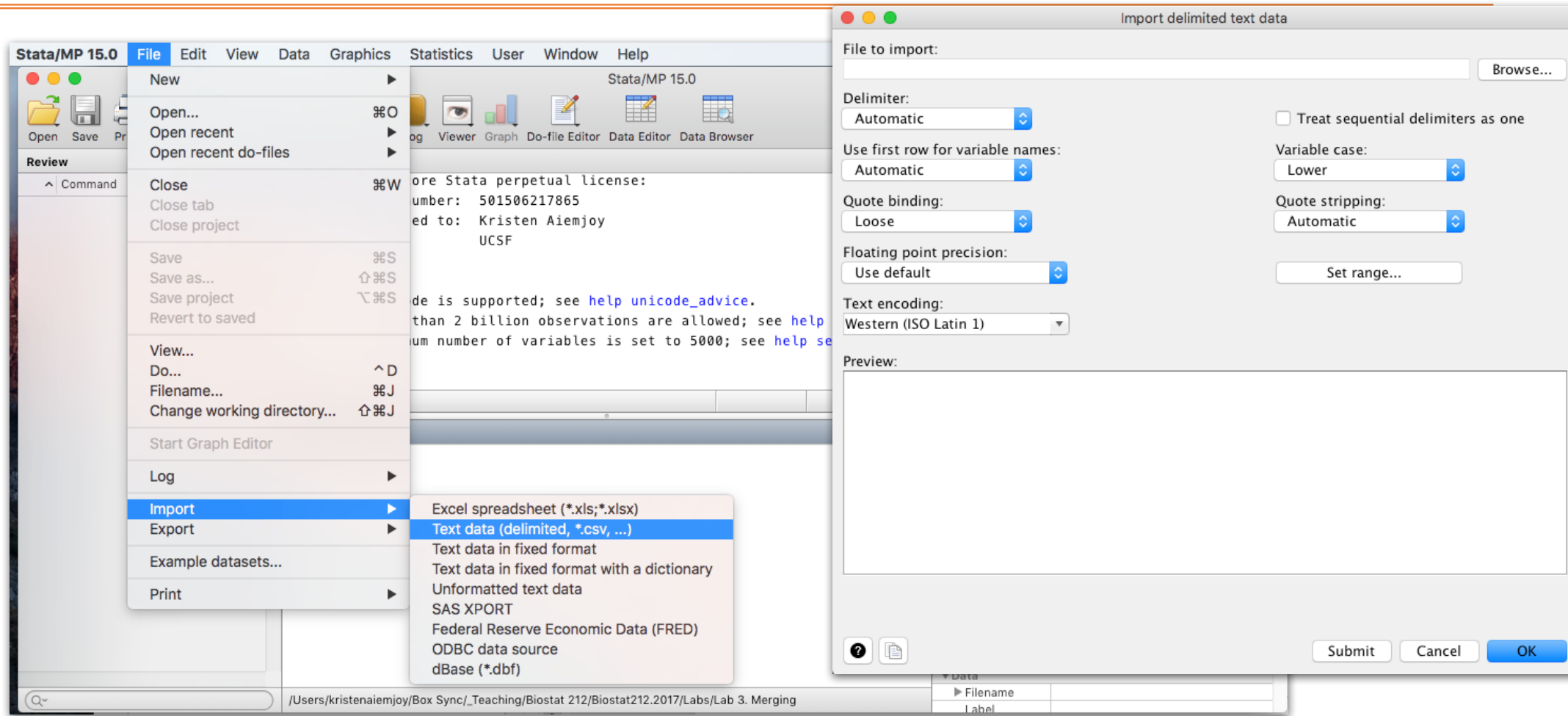
Review: tab vs sum

- ◇ tabulate categorical variables
 - *race, sex, heartattack*
- ◇ summarize continuous variables
 - *BloodPressure, Age, Weight*

Review: importing files

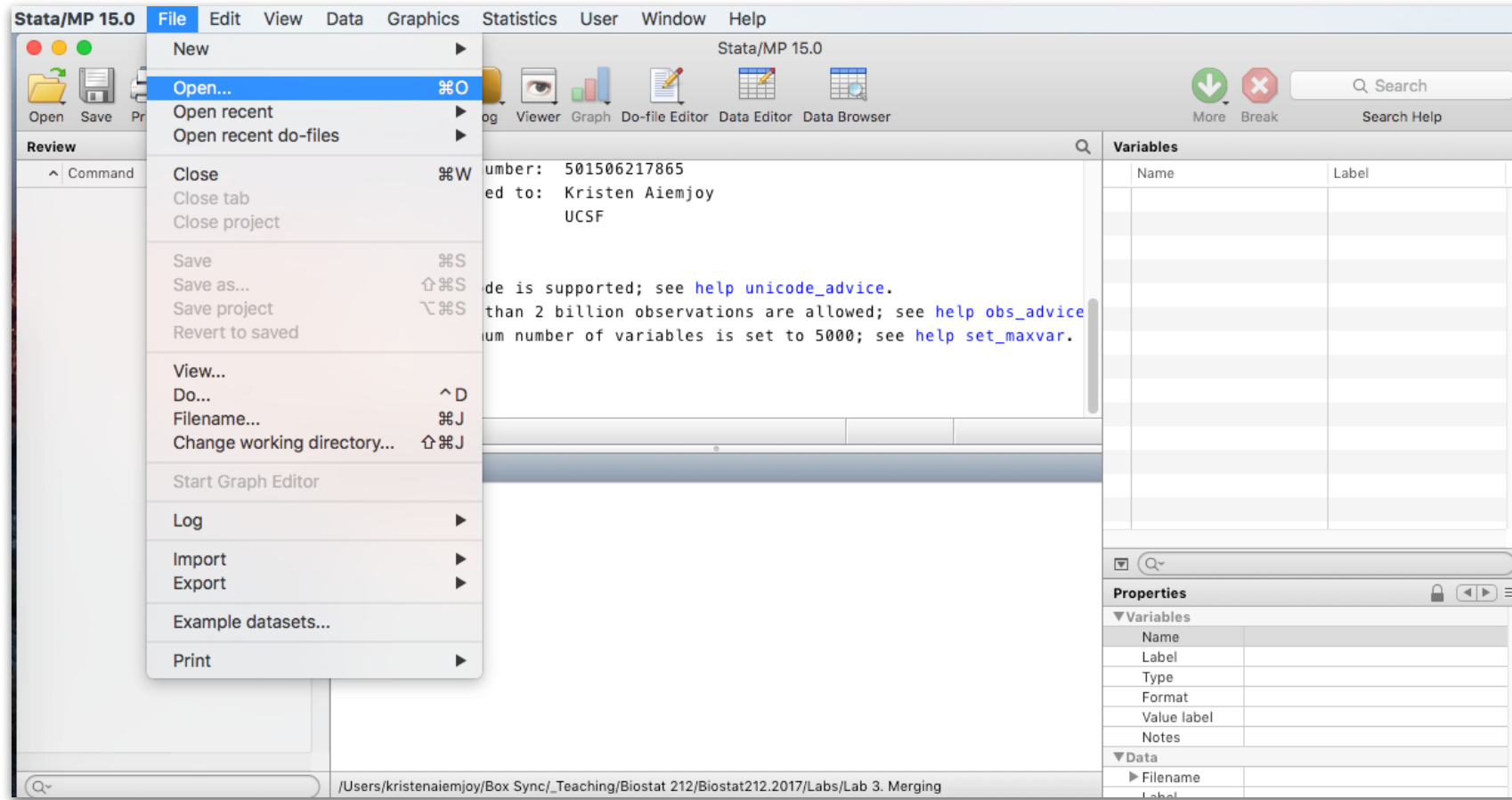
Source program	Export as	Stata command
Excel	.csv	import delimited filename.csv
Excel	.xls	import excel filename.xls
SPSS	.dta	use filename.dta
SAS	.xpf	import sasexport filename.xpf
SQL		odbc "help odcb" for more information

Review: Import .csv file



```
import delimited accupuncture_raw.csv, clear
```

Review: Open Stata data set



use acupuncture.dta, clear

Clean data

- ◇ The data sets we have worked with (nhanes2.dta and acupuncture_baseline.dta) were clean
 - The data set came as a .dta file
 - The variables were ready to analyze
 - Most variables were labeled

In the real world:

- ◇ Your data will arrive:
 - On paper forms
 - In a text file (comma or tab delimited)
 - In Excel
 - In Access
 - In another data format (SAS, etc)

Insider tips for data cleaning



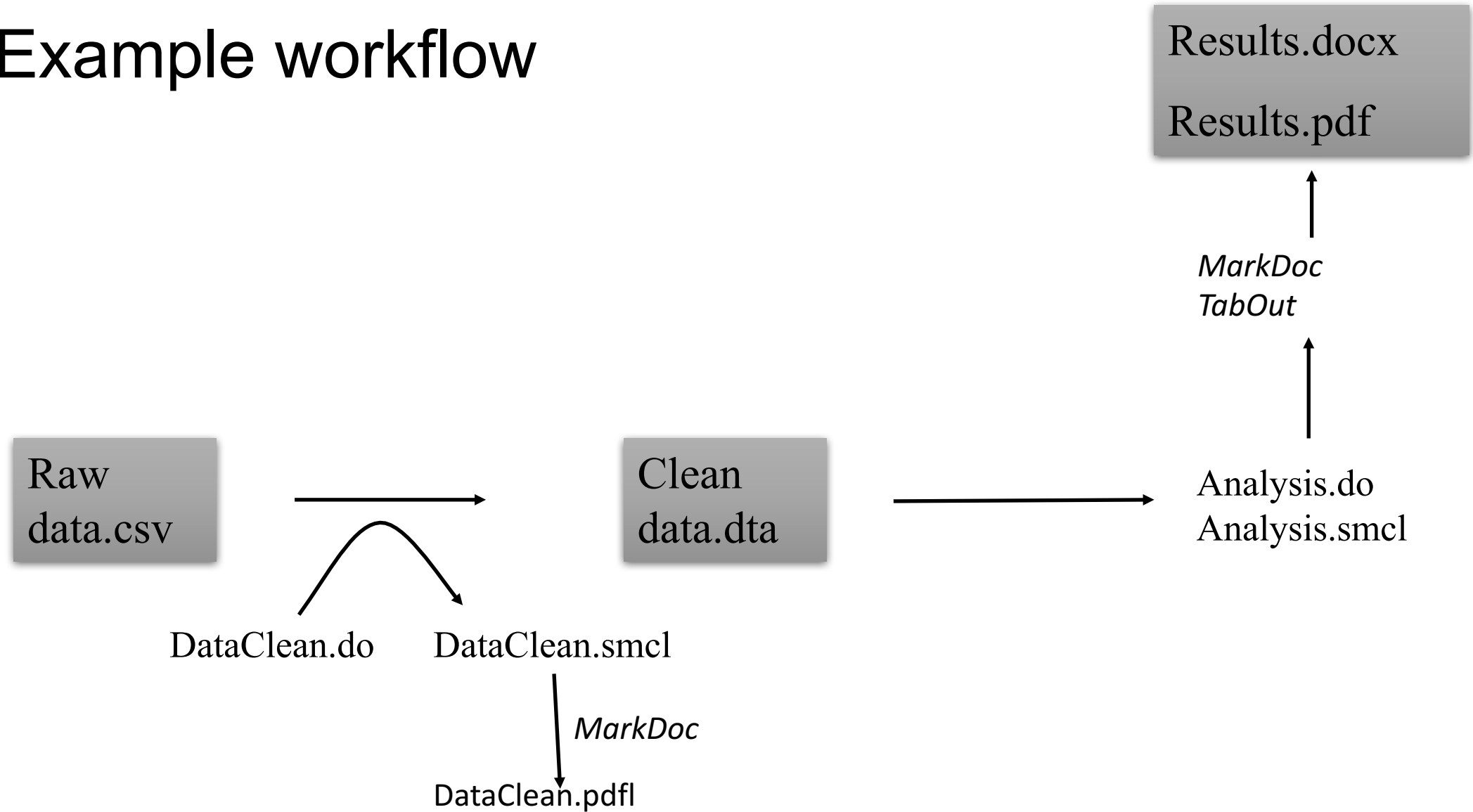
- We live in a dirty world
- Your data will be dirty also (no matter how well run the study is)
- Data cleaning can take up a huge amount of your Stata code and analysis time
- Keep a raw unedited version of the original dataset and do not ever touch it. Save revised data in a new file.

- All your data cleaning decisions should be fully documented in your do-file
- You will not remember what you did in 6 months time

```
/*Document everything and trust no  
one
```

```
...especially yourself */
```

Example workflow



Outline for today

- ◇ Name and label variables
- ◇ Checking ranges and removing outliers
- ◇ Missing data
- ◇ Generating variables
- ◇ Convert string variables to numeric
- ◇ Subset data
- ◇ MarkDoc

Load dataset for lecture today

◇ Download the “acupuncture_raw.csv” dataset from the course website

◇ Import the data set into Stata

```
import delimited acupuncture_raw.csv, clear
```

-or-

```
File<Import<Text data (*.csv)
```

Outline for today

- ◇ **Name and label variables**
- ◇ Checking ranges and removing outliers
- ◇ Missing data
- ◇ Generating variables
- ◇ Convert string variables to numeric
- ◇ Subset data
- ◇ MarkDoc

Name and label variables

- ◇ Do you need to rename the variable?
- ◇ Do you need to label the variable?
- ◇ Do you need to label the values of categorical variables?

Variable names: keep short and sweet!

Data Editor (Browse)

Edit Browse Filter Variables Properties Snapshots

id[1] 139

	chronicity	acupuncturist	group	basehs	posths	oneyrhs	f1
21	45	2	0	59	18.75	43.75	28
22	34	2	1	22.25	4.5	6.25	20
23	6	2	1	17	8.75	17.75	15
24	20	3	1	17.75	11	14.25	21
25	16	3	0	18.5	16.75	9.75	20
26	36	9	0	20	30.25	19.75	10
27	15	7	1	17.75	17.25	16.75	16
28	19	9	0	63.75	59	71.5	28
29	24	9	0	36.25	41.75	46.75	26
30	35	9	1	77	.	.	28
31	45	9	1	11	14.75	9.5	6
32	15	3	0	9.75	.	.	14
33	2	9	1	28.5	14	17.25	28
34	37	7	0	10.25	.	.	10
35	49	7	1	20.75	11	12.5	17
36	2	3	1	18.75	20.25	11	25
37	10	3	0	11.5	9	7.5	10
38	10	6	0	11.25	18.75	.	10
39	10	6	1	34	.	.	23
40	41	6	0	21.25	.	.	12
41	20	8	0	23.25	18	16	18
42	23	8	1	11.75	4.75	3.75	11
43	4	8	0	22.5	.	.	26

Variables

Name	Label
☒ age	
☒ sex	sex
☒ migraine	
☒ chronicity	
☒ acupuncturist	
☒ group	
☒ basehs	
☒ posths	
☒ oneyrhs	

Properties

Variables

Name	id
Label	
Type	int
Format	%8.0g
Value label	
Notes	

Data

Filename	
Label	
Notes	
Variables	78
Observations	401
Size	79.89K
Memory	64M

Vars: 78 Order: Dataset Obs: 401 Filter: Off

Change variable names

Variables Manager (78 variables)

Q Enter filter text here

Filter

Name	Label	Type	Format	Value label	Notes
id		int	%8.0g		
age		int	%8.0g		
sex	sex	str6	%9s		
migraine		byte	%8.0g		
chronicity		byte	%8.0g		
acupuncturist		byte	%8.0g		
group		byte	%8.0g		
basehs		float	%9.0g		
posths		float	%9.0g		
oneyrhs		float	%9.0g		
f1		float	%9.0g		
f2		float	%9.0g		
f5		float	%9.0g		
pf1		float	%9.0g		
rlp1		byte	%8.0g		
rl1		float	%9.0g		
ef1		float	%9.0g		

Variable properties

Name
group

Label

Type
byte

Format
%8.0g Create...

Value label
Manage...

Notes
No notes Manage...

< > Reset Apply

Data > Variable manager

Change variable names

```
rename old_varname new_varname
```

- changes the name of an existing variable
- the content of the variable is unchanged

Variable labels: help you remember for details

Data Editor (Browse)

Edit Browse Filter Variables Properties Snapshots

id[1] 139

	chronicity	acupuncturist	group	basehs	posths	oneyrhs	f1
21	45	2	0	59	18.75	43.75	28
22	34	2	1	22.25	4.5	6.25	20
23	6	2	1	17	8.75	17.75	15
24	20	3	1	17.75	11	14.25	21
25	16	3	0	18.5	16.75	9.75	20
26	36	9	0	20	30.25	19.75	10
27	15	7	1	17.75	17.25	16.75	16
28	19	9	0	63.75	59	71.5	28
29	24	9	0	36.25	41.75	46.75	26
30	35	9	1	77	.	.	28
31	45	9	1	11	14.75	9.5	6
32	15	3	0	9.75	.	.	14
33	2	9	1	28.5	14	17.25	28
34	37	7	0	10.25	.	.	10
35	49	7	1	20.75	11	12.5	17
36	2	3	1	18.75	20.25	11	25
37	10	3	0	11.5	9	7.5	10
38	10	6	0	11.25	18.75	.	10
39	10	6	1	34	.	.	23
40	41	6	0	21.25	.	.	12
41	20	8	0	23.25	18	16	18
42	23	8	1	11.75	4.75	3.75	11
43	4	8	0	22.5	.	.	26

Variables

Name	Label
☒ age	
☒ sex	sex
☒ migraine	
☒ chronicity	
☒ acupuncturist	
☒ group	
☒ basehs	
☒ posths	
☒ oneyrhs	

Properties

Variables

Name	id
Label	
Type	int
Format	%8.0g
Value label	
Notes	

Data

Filename	
Label	
Notes	
Variables	78
Observations	401
Size	79.89K
Memory	64M

Vars: 78 Order: Dataset Obs: 401 Filter: Off

Change variable label

Variables Manager (78 variables)

Q Enter filter text here

Filter

Name	Label	Type	Format	Value label	Notes
id		int	%8.0g		
age		int	%8.0g		
sex	sex	str6	%9s		
migraine		byte	%8.0g		
chronicity		byte	%8.0g		
acupuncturist		byte	%8.0g		
group		byte	%8.0g		
basehs		float	%9.0g		
posths		float	%9.0g		
oneyrhs		float	%9.0g		
f1		float	%9.0g		
f2		float	%9.0g		
f5		float	%9.0g		
pf1		float	%9.0g		
rlp1		byte	%8.0g		
rl1		float	%9.0g		
ef1		float	%9.0g		

Variable properties

Name
group

Label

Type
byte

Format
%8.0g

Value label

Notes
No notes

< >

Data > Variable manager

Change variable label

```
label variable varname "labelname (can have spaces)"
```

Practice

Variable f1 = headache frequency at baseline

1. Rename f1 → hfreq_bl
2. Variable f1 = headache frequency at baseline

```
rename f1 hfreq_bl
```

```
label variable hfreq_bl "headache frequency baseline"
```

Value labels

Data Editor (Browse)

Edit Browse Filter Variables Properties Snapshots

id[1] 139

	chronicity	acupuncturist	group	basehs	posths	oneyrhs	f1
21	45	2	0	59	18.75	43.75	28
22	34	2	1	22.25	4.5	6.25	20
23	6	2	1	17	8.75	17.75	15
24	20	3	1	17.75	11	14.25	21
25	16	3	0	18.5	16.75	9.75	20
26	36	9	0	20	30.25	19.75	10
27	15	7	1	17.75	17.25	16.75	16
28	19	9	0	63.75	59	71.5	28
29	24	9	0	36.25	41.75	46.75	26
30	35	9	1	77	.	.	28
31	45	9	1	11	14.75	9.5	6
32	15	3	0	9.75	.	.	14
33	2	9	1	28.5	14	17.25	28
34	37	7	0	10.25	.	.	10
35	49	7	1	20.75	11	12.5	17
36	2	3	1	18.75	20.25	11	25
37	10	3	0	11.5	9	7.5	10
38	10	6	0	11.25	18.75	.	10
39	10	6	1	34	.	.	23
40	41	6	0	21.25	.	.	12
41	20	8	0	23.25	18	16	18
42	23	8	1	11.75	4.75	3.75	11
43	4	8	.	22.5	.	.	26

Vars: 78 Order: Dataset Obs: 401 Filter: Off

Variables

Name	Label
☒ age	
☒ sex	sex
☒ migraine	
☒ chronicity	
☒ acupuncturist	
☒ group	
☒ basehs	
☒ posths	
☒ oneyrhs	

Properties

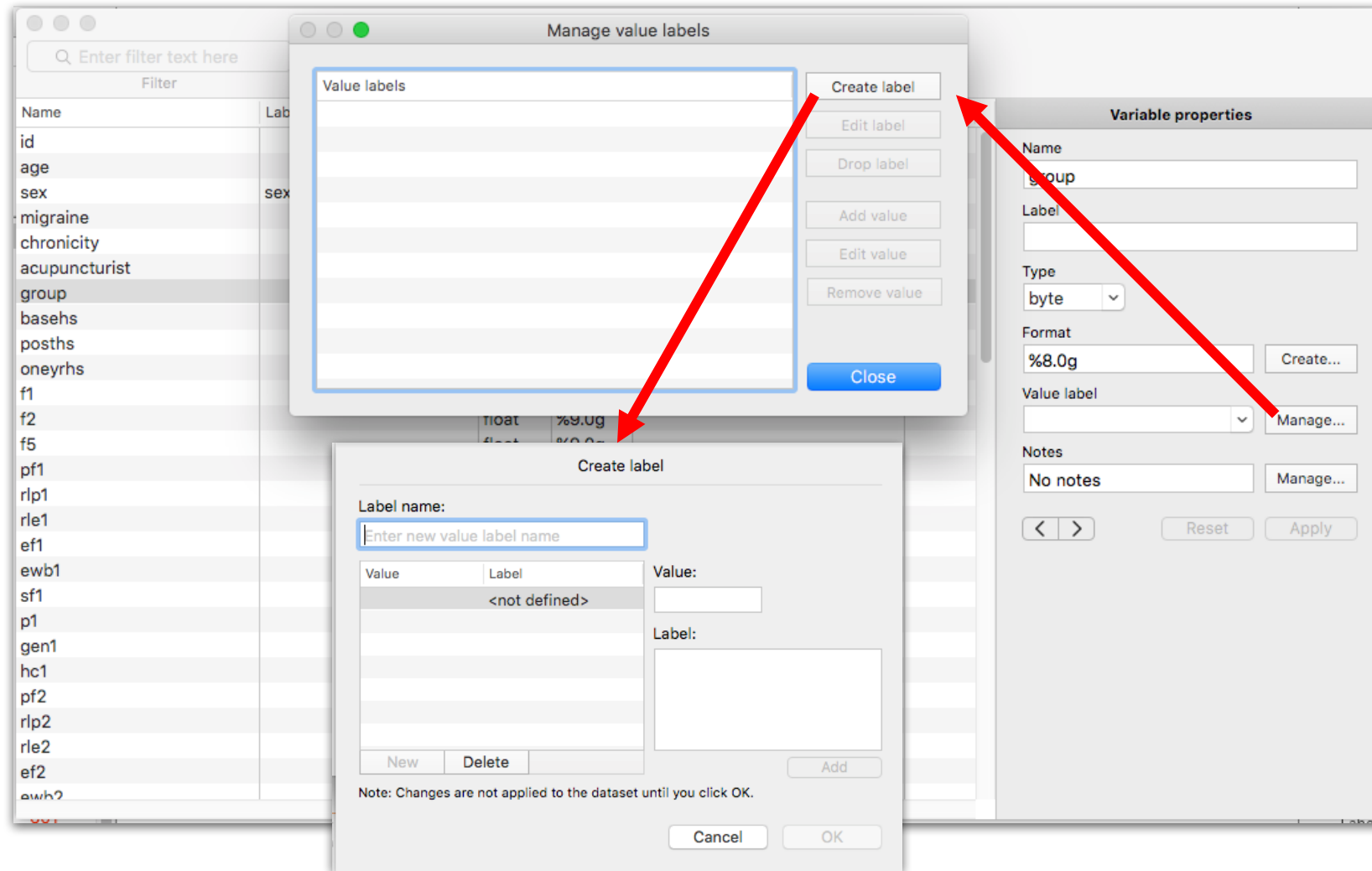
▼ Variables

Name	id
Label	
Type	int
Format	%8.0g
Value label	
Notes	

▼ Data

► Filename	
Label	
► Notes	
Variables	78
Observations	401
Size	79.89K
Memory	64M

Add value labels



Data >Variable manager

Add value labels

Label values of categorical variables

```
label define labelname # "label" # "label" ...
```

```
label values varname labelname
```

```
example label define yesno 0 "no" 1 "yes"
```

```
label values exercise yesno
```

Practice

- ◊ Label group (1=intervention) (2=control)

```
tab group
```

```
label define grouplabel 1 intervention 0 control
```

```
label values group grouplabel
```

```
tab group
```

Outline for today

- ◇ Name and label variables
- ◇ **Checking ranges and removing outliers**
- ◇ Missing data
- ◇ Generating variables
- ◇ Convert string variables to numeric
- ◇ Subset data
- ◇ MarkDoc

Range checks

◇ Are there any outliers? (unrealistic values?)

- `codebook varname`
- `inspect varname`
- `histogram varname`
- `summarize varname`
- `spikeplot varname`

Removing outliers: replace

```
replace varname = # if condition
```

– replace the value of a variable meeting a certain condition

Example

```
replace weight = 230 if id == "13des4"
```

```
replace age = . if age>110
```

Practice

- Are there any outliers for age?

```
spikeplot age
```

```
sum age
```

```
list if age >400
```

- **Replace outlier**

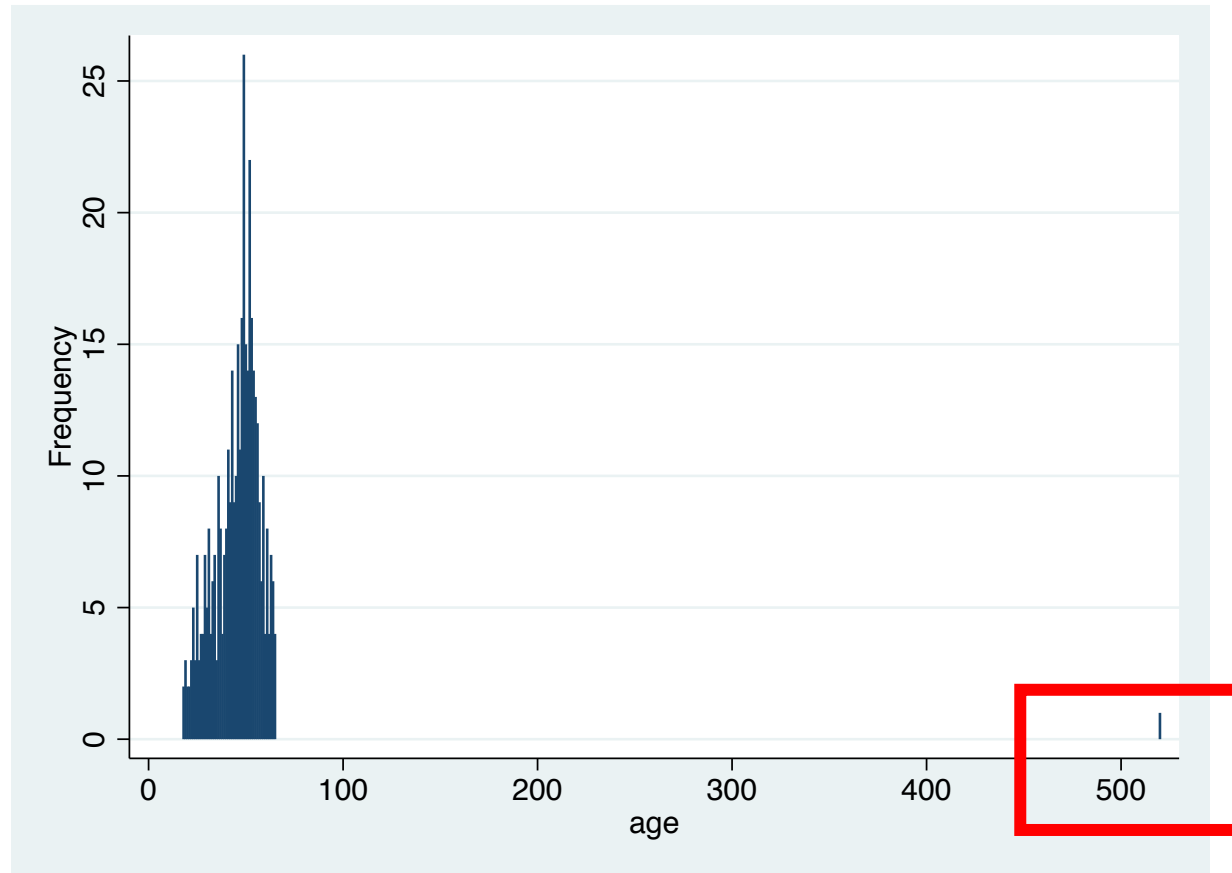
```
replace age = 52 if id==363      -or-
```

```
replace age = 52 if age>115
```

```
spikeplot age
```

Range checks

`spikeplot age`



Outline for today

- ◇ Name and label variables
- ◇ Checking ranges and removing outliers
- ◇ **Missing data**
- ◇ Generating variables
- ◇ Convert string variables to numeric
- ◇ Subset data
- ◇ MarkDoc

Missing data

- ◇ Is there missing data?
- ◇ How are missing values coded?
- ◇ Standardize all missing data to (.)

Missing data

- ◇ Stata designates missing data with a period (.)
- ◇ Important to distinguish between 'missing' and 'no'
- ◇ Why is it missing?
 - Loss to follow up
 - Death
 - Skip pattern in a survey
 - Certain questions can only be answered if respondent has children
 - Data collection errors
 - Respondent refusal/non-response

Missing data

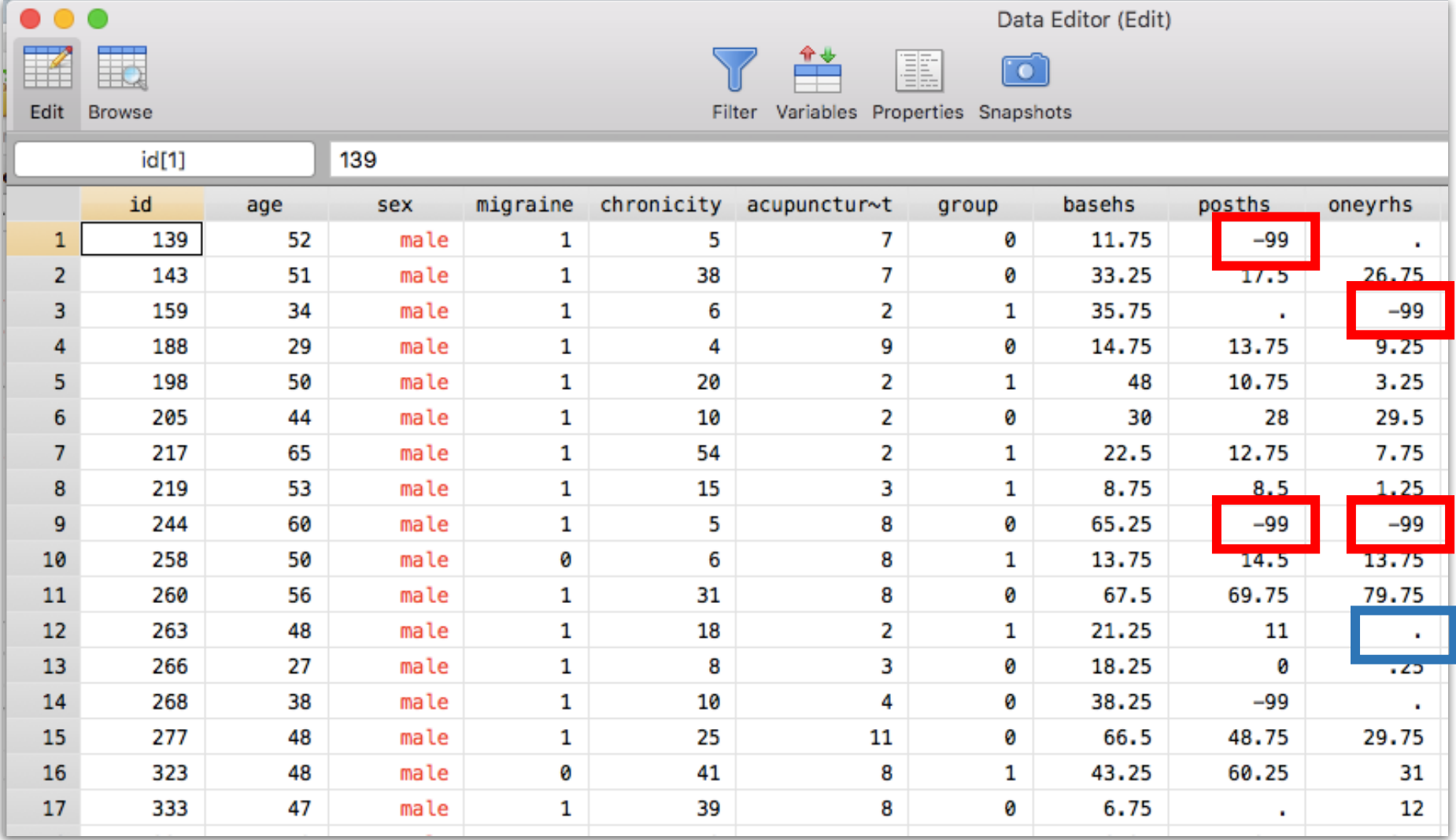
◇ Is there missing data?

```
tab varname, missing  
misstable sum
```

◇ How are missing values coded?

- Range checks (sum, codebook)
- tabulate varname, missing

Convert numbers to missing values

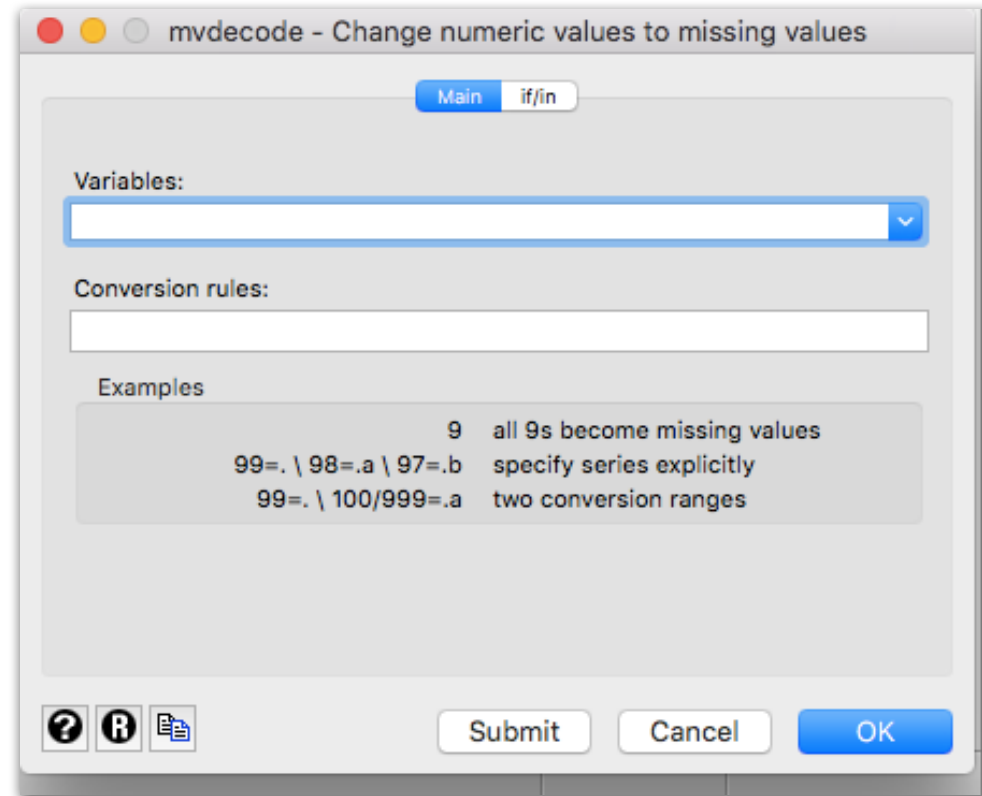


Data Editor (Edit)

	id	age	sex	migraine	chronicity	acupunctur~t	group	basehs	posths	oneyrhs
1	139	52	male	1	5	7	0	11.75	-99	.
2	143	51	male	1	38	7	0	33.25	17.5	26.75
3	159	34	male	1	6	2	1	35.75	.	-99
4	188	29	male	1	4	9	0	14.75	13.75	9.25
5	198	50	male	1	20	2	1	48	10.75	3.25
6	205	44	male	1	10	2	0	30	28	29.5
7	217	65	male	1	54	2	1	22.5	12.75	7.75
8	219	53	male	1	15	3	1	8.75	8.5	1.25
9	244	60	male	1	5	8	0	65.25	-99	-99
10	258	50	male	0	6	8	1	13.75	14.5	13.75
11	260	56	male	1	31	8	0	67.5	69.75	79.75
12	263	48	male	1	18	2	1	21.25	11	.
13	266	27	male	1	8	3	0	18.25	0	.25
14	268	38	male	1	10	4	0	38.25	-99	.
15	277	48	male	1	25	11	0	66.5	48.75	29.75
16	323	48	male	0	41	8	1	43.25	60.25	31
17	333	47	male	1	39	8	0	6.75	.	12

-99 → .

Convert numbers to missing values



Data > Create or change data > Other variable-transformation commands > Change numeric values to missing

Convert numbers to missing values

`mvdecode _all, mv(-99)` – recode all '-99' as '.'

`mvdecode varname, mv(-99)` – for specific `varname` recode '-99' as '.'

Practice

- ◊ Look for missing data

```
misstable sum
```

- ◊ Change all '-99' to (.)

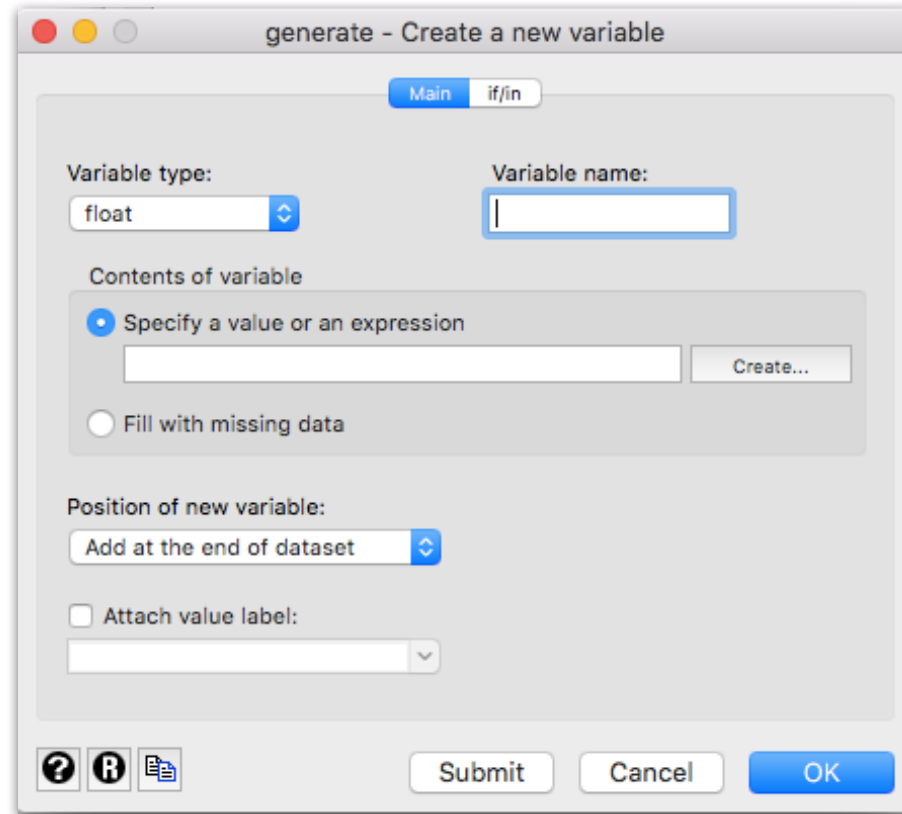
```
mvdecode _all, mv(-99)
```

```
misstable sum
```

Outline for today

- ◇ Name and label variables
- ◇ Checking ranges and removing outliers
- ◇ Missing data
- ◇ **Generating variables**
- ◇ Convert string variables to numeric
- ◇ Subset data
- ◇ MarkDoc

Create a new variable



Data > Create or change data > Create a new variable

Create a new variable

generate `new_varname = exp`

-generate a new variable

`exp` = 0, #, ".", _n, addition(+), subtraction(-)

example: `gen calsum = cal1 + cal2`

generate `new_varname=(condition)`

-generate a new dichotomous variable
where 1= condition_True &
0=condition_False

Options

`before(varname) / after(varname)`

–place new variable before/after
specified variable

Practice

- ◇ Generate a new running id variable and a blank variable

```
gen newid = _n  
gen blank = .
```

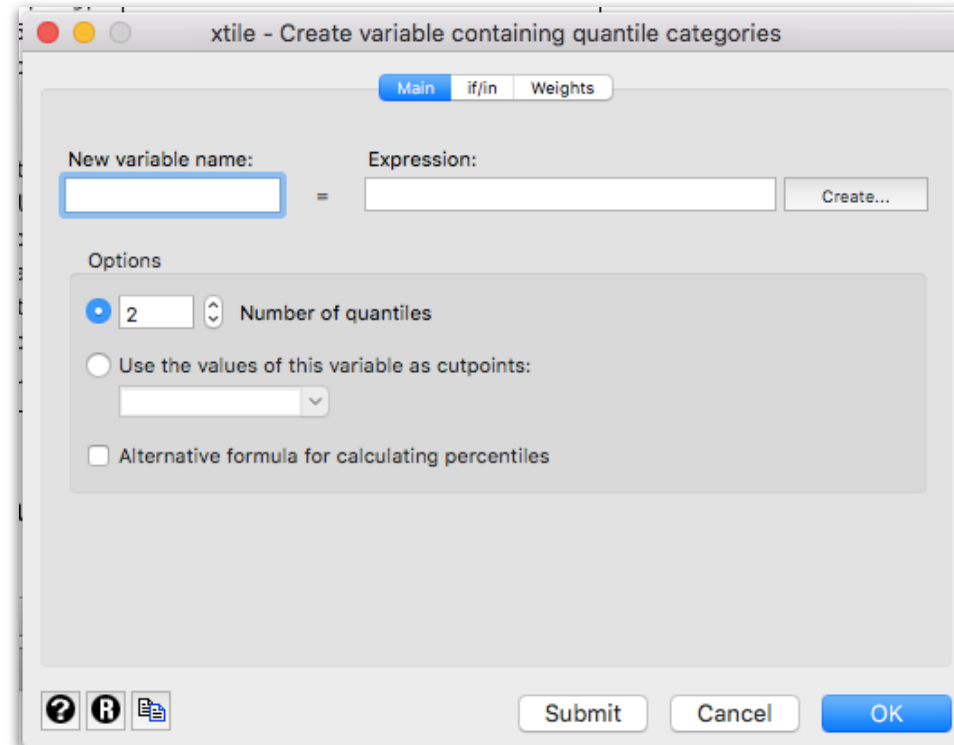
```
browse newid blank
```

- ◇ Generate a new variable for 'severe migraine' for a migraine + a headache score >50

```
gen severemigraine = (migraine==1 & basehs>50)
```

```
tab severemigraine
```

Create variable containing quantile categories



Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Create variable of quantiles

Create variable containing quantile categories

`xtile new_varname = varname` -create a new variable containing percentiles

Options

`if` –use logic operators

`nq (#)` –set number of percentiles (default is 2)

Practice

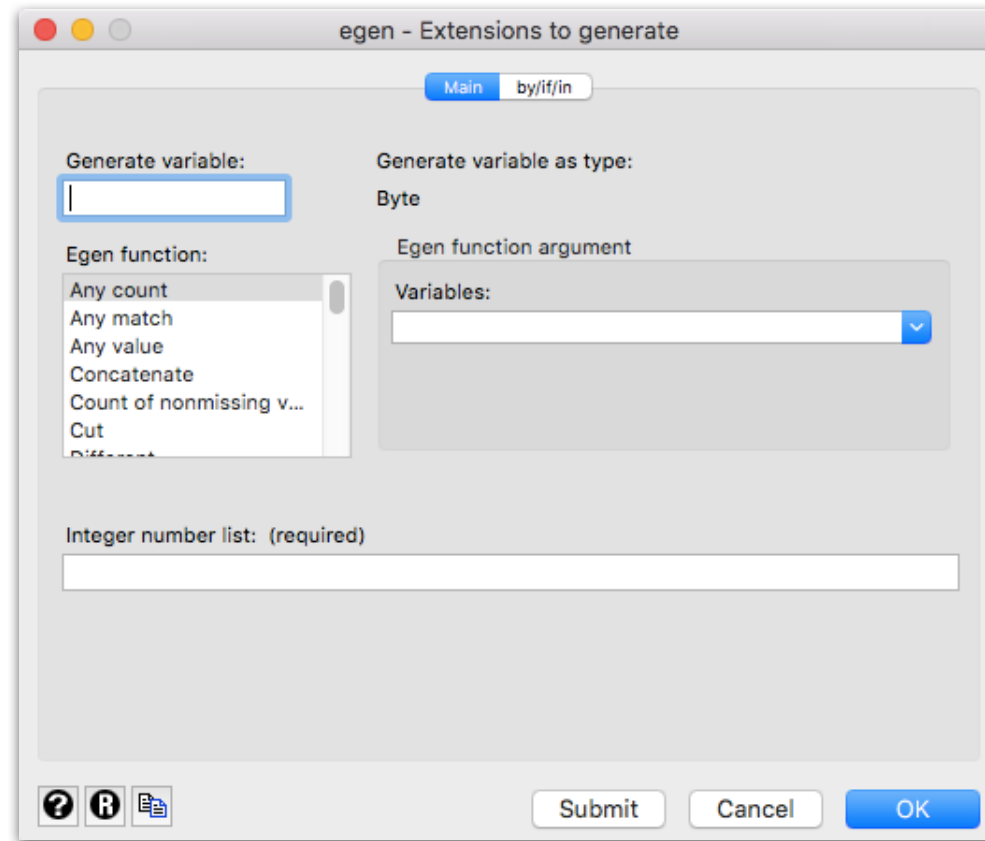
- ◇ Create a new categorical variable that is a quartile of baseline headache score.

```
xtile hs_quartile = basehs, nq(4)
```

- ◇ What is the mean baseline headache score in the first quartile?

```
bysort hs_quartile: sum basehs
```

Create new variable (extended)



Data > Create or change data > Create new variable (extended)

Create new variable (extended)

```
egen new_varname = function(varname)
```

-create a new variable that is a function of a different variable(s)

Functions

mean	egen new_varname = mean(varname)
row mean	egen new_varname = rowmean(varlist)
row total	egen new_varname = rowtotal(varlist)
concatenate	egen new_varname = concat(varlist)

Options: by, if

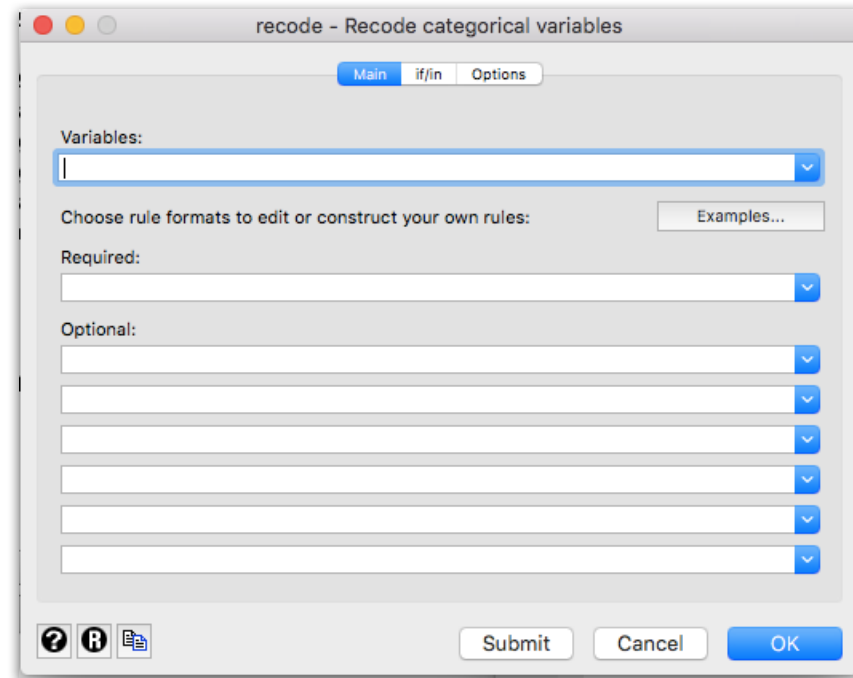
*many more functions → look in help file

Practice

- ◇ Create a variable for average headache score (basehs, posths, oneyrhs)

```
egen meanhs = rowmean(basehs posths oneyrhs)
```

Create categorical variable from continuous variable



Data > Create or change data > Other variable-transformation commands > Recode categorical variable

Create categorical variable from continuous variable

```
recode old_varname (min/# = 0 "value label") (#/# = 1 "value  
label") (#/max = 2 "value label"), gen(new_varname)
```

Practice

- ◇ Create a dichotomous variable for age (>50)

```
recode age (min/49 = 1 "less than 50") (50/max = 2 "50  
and older"), gen(age2)
```

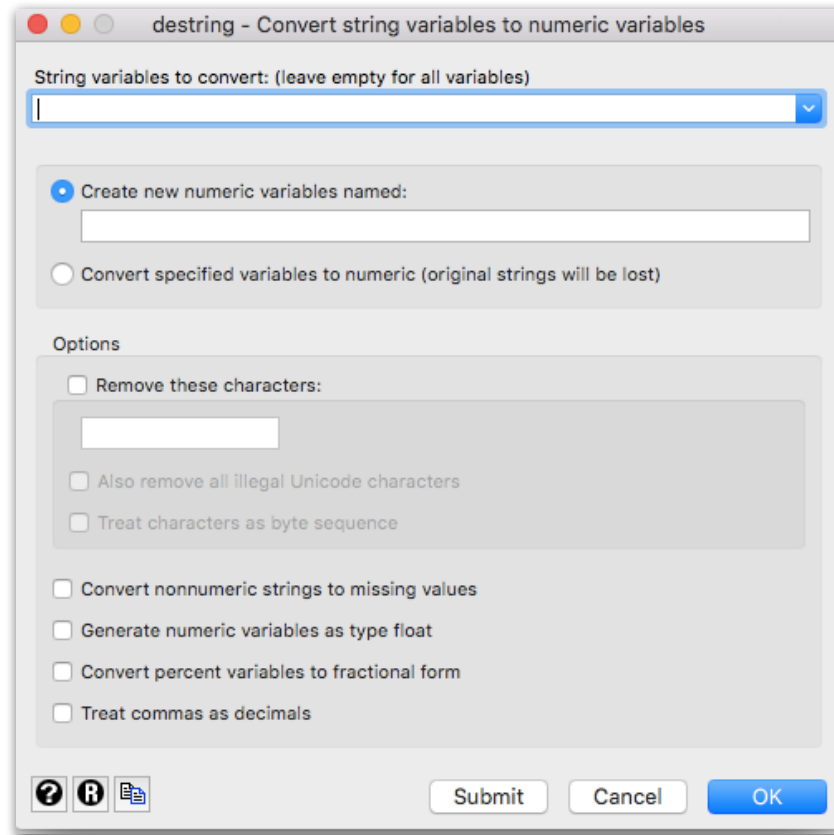
- ◇ How many study participants are 50 years old or older?

```
tab age2
```

Outline for today

- ◇ Name and label variables
- ◇ Checking ranges and removing outliers
- ◇ Missing data
- ◇ Generating variables
- ◇ **Convert string variables to numeric**
- ◇ Subset data
- ◇ MarkDoc

Convert string variables to numeric



Data > Create or change data > Other variable-transformation commands > Convert variables from string to numeric

Convert string variables to numeric

destring `varname`, generate(`new_varname`)

-convert string into numeric variables

encode `varname`, generate(`new_varname`)

-Encode string into numeric variable with labels

destring or encode?

encode and deststring both convert string variables to numeric variables. What's the difference?

◇ Use deststring to:

- convert intrinsically numeric variables such as **age** that do not have labels but happen to be stored as strings.

◇ Use encode to:

- convert categorical variables that will be included in statistical models but have been stored as strings. **encode** can convert the categories to value labels for the new numeric variable.

Practice

- ◇ Change `withdrawal_reason` from string to numeric/categorical:

```
browse withdrawal_reason
```

```
encode withdrawal_reason, gen(dropout)
```

```
browse withdrawal_reason
```

- ◇ What percent of individuals who dropped out withdrew consent?

```
tab dropout
```

Outline for today

- ◇ Name and label variables
- ◇ Checking ranges and removing outliers
- ◇ Missing data
- ◇ Generating variables
- ◇ Convert string variables to numeric
- ◇ **Subset data**
- ◇ MarkDoc

Subset data: drop

`drop if` -drops subset of data given condition

example `drop if age>100`

`drop in` -drops subset of observations

example `drop in 1/4`

`drop varlist` -drop one or more variables

****careful! you cant undo**

Subset data: keep

keep if -keep observations meeting condition, all others are dropped!

example keep if age<100

keep in -keep subset of observations

example keep in 1/30

keep varlist -keep one or more variables, all others dropped!

****careful! you cant undo**

Outline for today

- ◇ Name and label variables
- ◇ Checking ranges and removing outliers
- ◇ Missing data
- ◇ Generating variables
- ◇ Convert string variables to numeric
- ◇ Subset data
- ◇ **MarkDoc**

MarkDoc

- ◇ A package that transforms .smcl log files into PDFs and Word documents
- ◇ Creates “dynamic” documents
 - Fill tables/texts with numbers from the data, updates every time you run even if your data has changed
- ◇ Understands different markup languages including:
 - html, LaTeX, Markdown
 - Markdown is the simplest mark up language, simple syntax to create basic documents

```

1
2
3 qui log using example, replace
4
5
6 /**
7 ##Load auto dataset
8 ***/
9 sysuse auto
10
11 //OFF
12 foreach var of varlist weight price mpg {
13     summarize `var'
14     local `var'_mean : display %9.2f r(mean)
15     local `var'_sd : display %9.2f r(sd)
16 }
17
18
19
20 //ON
21 tbl ("Variable Name", "Mean", "SD" \           ///
22     "_Weight_", `weight_mean', `weight_sd' \    ///
23     "_Price_", `price_mean', `price_sd' \       ///
24     "_MPG_", `mpg_mean', `mpg_sd')             ///
25     , title("Table 1. Summary of _Weight_, _Price_, and _MPG_ variables")
26
27
28
29 qui log c
30
31 markdoc example, replace export(pdf) title("MarkDoc Package Example")
32
33

```



MarkDoc Package Example

Load auto dataset

```
. sysuse auto
(1978 Automobile Data)
```

Table 1. Summary of *Weight*, *Price*, and *MPG* variables

Variable Name	Mean	SD
<i>Weight</i>	3019.46	777.19
<i>Price</i>	6165.26	2949.50
<i>MPG</i>	21.30	5.79

```
9
10 qui log using example1.smcl, replace
11
12 /***
13
14 #Heading 1
15 ##Heading 2
16 ###Heading 3
17 ####Heading 4
18
19 _Italic text_
20
21 __Bold text__
22
23 **/
24
25
26 markdoc example1, replace install export(pdf)
27
28
29
```



Heading 1

Heading 2

Heading 3

Heading 4

Italic text

Bold text

```
9
10 qui log using example1.smcl, replace
11
12 /***
13
14 #Heading 1
15 ##Heading 2
16 ###Heading 3
17 ####Heading 4
18
19 _Italic text_
20
21 __Bold text__
22
23 ---
24
25
26 ***/
27
28 markdoc example1, replace install export(pdf) style(stata)
29
30
31
32
33
34
```



Heading 1

Heading 2

Heading 3

Heading 4

Italic text

Bold text

```
9
10 qui log using example1.smcl, replace
11
12 /***
13
14 #Heading 1
15 ##Heading 2
16 ###Heading 3
17 ####Heading 4
18
19
20 _Italic text_
21
22 __Bold text__
23
24 ----
25
26 ***/
27
28 markdoc example1, replace install export(pdf) style(formal)
29
30
31
```



Heading 1

Heading 2

Heading 3

Heading 4

Italic text

Bold text

Markdown notation

<code>/***</code>	start text segment
<code>*** /</code>	end text segment
<code>#text</code>	Heading 1
<code>##text</code>	Heading 2
<code>###text</code>	Heading 3
<code>####text</code>	Heading 4
<code>_text_</code>	Italic text
<code>__text__</code>	Bold text
<code>img</code>	Save image in dynamic document
<code>tbl</code>	Create table in dynamic document

```

10 qui log using example1.smcl, replace
11
12 /***
13
14 #Heading 1
15 ##Heading 2
16 ###Heading 3
17 ####Heading 4
18
19
20 _Italic text_
21
22 __Bold text__
23
24 ---
25
26 ***/
27
28 markdoc example1, replace install export(pdf) style(formal)
29

```

Make sure these names match. You create your markdoc from your .smcl log file so the names have to be the same

```
10 qui log using example1.smcl, replace
```

```
11
```

```
12 /***
```

```
13
```

```
14 #Heading 1
```

```
15 ##Heading 2
```

```
16 ###Heading 3
```

```
17 ####Heading 4
```

```
18
```

```
19
```

```
20 _Italic text_
```

```
21
```

```
22 __Bold text__
```

```
23
```

```
24 ---
```

```
25
```

```
26 ***/
```

```
27
```

```
28 markdoc example1, replace install export(pdf) style(formal)
```

```
29
```

Export as pdf, docx, html or tex

Assignment 2

- ◇ Answer questions [1-10]
- ◇ Run .do file and turn in the PDF created by MarkDoc

```
1  **Setup
2
3  cd "/Users/kristenaiejmjoy/Box Sync/_Teaching/Biostat 212/Biostat212.2017/Assignments/Assignment 2 cl
4
5  capture log close //close any open log files
6
7  set linesize 120 //increase the linesize (width of the page) helps for formatting the Markdown docum
8
9  quietly log using biostat212_assignment2_name.smcl, replace //using "quietly" will stop print out in
10
11
12
13  /**
14  #Biostat 212: 2017
15  #Assignment 2: Data cleaning
16  ## Your Name
17
18  ---
19
20  _Instructions_
21  In this assignment we will be cleaning a subset of the NHANES data set "nhanes_raw.csv"
22  We will import the data from a .csv file and export a clean data set.
23  Answer questions 1-10 in the .do file.
24  The commands you will need to answer these questions are already included in the
25  .do file. You should be able to run the entire .do file at once and create a PDF
26  with you answers and the code using the MarkDoc package. The syntax for calling the
27  MarkDoc package is already included in this .do file.
28  To get the marcdoc command to work, you need to install 3 packages "marcdoc" "weaver" and "statax".
29  To install these packages open the "MarkDoc_Installation.do" file on the course website and run all
30  three commands.
31
32  Turn in the PDF you create using MarkDoc, which should include your name and answers
33  to questions 1-10
34
35
36  The full .do file may take some time to run.
37
38
```



Biostat 212: 2017

Assignment 2: Data cleaning

Your Name

Instructions

In this assignment we will be cleaning a subset of the NHANES data set "nhanes_raw.csv". We will import the data from a .csv file and export a clean data set. Answer questions 1-10 in the .do file. The commands you will need to answer these questions are already included in the .do file. You should be able to run the entire .do file at once and create a PDF with you answers and the code using the MarkDoc package. The syntax for calling the MarkDoc package is already included in this .do file. To get the marcdoc command to work, you need to install 3 packages "marcdoc" "weaver" and "statax". To install these packages open the "MarkDoc_Installation.do" file on the course website and run all three commands.

Turn in the PDF you create using MarkDoc, which should include your name and answers to questions 1-10

The full .do file may take some time to run.[]

IMPORT

Import the raw .CSV file[]

```
. import delimited "nhanes_raw.csv", clear
(18 vars, 1,269 obs)
```

Browse data to make sure it worked. Use the exploring commands we learned last week to answer the following questions

1. How many observations are there? [Answer here]

To install `markdoc`

- ◇ Download “MarkDoc_Installation.do” from the CLE and follow the instructions

```
1
2  ***MarcDoc Installation Instructions
3  **8/7/2017
4
5
6
7  ***Step 1: Run the following chunk of commands (lines 8-14)
8  *it will take a few minutes to install, watch the progress bar at the bottom of your stata window
9  net install markdoc, replace from("https://raw.githubusercontent.com/haghigh/markdoc/master/")
10 net install weaver, replace from("https://raw.githubusercontent.com/haghigh/weaver/master/")
11 net install statax, replace from("https://raw.githubusercontent.com/haghigh/statax/master/")
12 net install github, replace from("https://haghigh.github.io/github/")
13 markdocpandoc //installs pandoc software
14 markdocwkhtmltopdf //installs html to pdf software
15
16
17 ***Step 2: run the following command to test that installation worked
18 *click on the "example.pdf" blue link in your results window after the command runs
19 markdoc, test
20
21
```