Data Management in SPSS

Dr. Md. Atiqul Islam

M.Sc. (SUST, BD), M.Sc. (UHasselt, BE), PhD (RuG, NL)

Professor
Department of Statistics
Jagannath University, Dhaka-1100
E-mail: atique@stat.jnu.ac.bd

Outline

▷ Compute New Variables

▷ Recode Variables

▷ Merge Data Sets

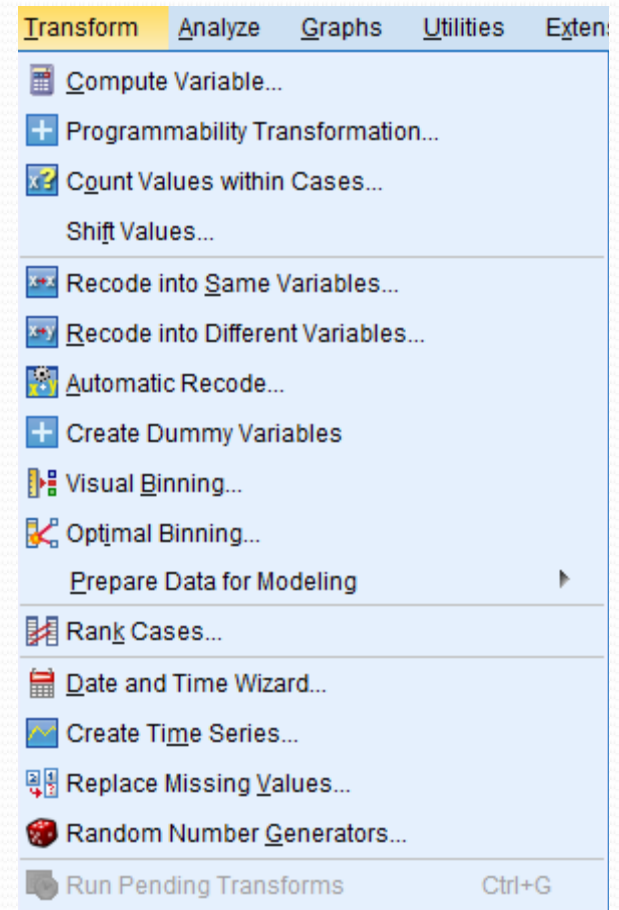▷ Compute Aggregate Statistics

▷ Some Descriptive Analysis

Datasets

▷ LIFESPAN.sav: Lifespan data of 195 Rates

▷ dads.sav: Three (3) dads' ID, name and their income

▷ faminc.sav: Three (3) dads' ID and three years income

▷ kids.sav: Information of kids belong to the three dads

▷ LIFESPAN+DIET.sav: Lifespan and diet data of 195 Rates

Data Management in SPSS

▷ DATA and TRANSFORM are the data management Menu Bar.

▷ **DATA**: It is used in manipulation of the format of the data spreadsheet to be used for analysis.

▷ **Transform**: It is used for generating new variables such as computing a new variable, recoding an existing variable.

TRANSFORM Menu Bar

▷ **Data Transformation**

▷ Most data transformations can be done with the **Compute** command.

▷ We want to use the data file "LIFESPAN+DIET.sav".

Transform	Analyze	Graphs	Utilities	Extens

- 🖩 Compute Variable...
- ➕ Programmability Transformation...
- 🔢 Count Values within Cases...
-    Shift Values...
- 🔤 Recode into Same Variables...
- 🔤 Recode into Different Variables...
- 🎲 Automatic Recode...
- ➕ Create Dummy Variables
- ▮ Visual Binning...
- 📈 Optimal Binning...
-    Prepare Data for Modeling ▶
- 📊 Rank Cases...
- 📅 Date and Time Wizard...
- 〰 Create Time Series...
- 🔢 Replace Missing Values...
- 🎲 Random Number Generators...
- 🔲 Run Pending Transforms   Ctrl+G

COMPUTE Function

▷ Suppose we want to calculate standardized value of LIFESPAN (X). The formula is,

$$Z = \frac{X - mean\ (X)}{standard\ deviation\ (X)}$$

▷ Suppose mean and standard deviation of LIFESPAN are 835 and 275.

▷ **Manually**

  ▷ Transform Compute Variable > Define target variable (**STD_LIFESPAN) > ((LIFESPAN-835)/275)**" > OK

▷ **Syntax**

  COMPUTE STD_LIFESPAN=(LIFESPAN-835)/275.

  EXECUTE.

Compute

Compute

Target variable

Desired formula

Target Variable:
STD_LIFESPAN = Numeric Expression:
(LIFESPAN-835)/275

Type & Label...

IDENTIFICATION N...
Type of Diet [DIET]
Length of Life [LIFE...
CATEGORICAL

Variable can be selected manually
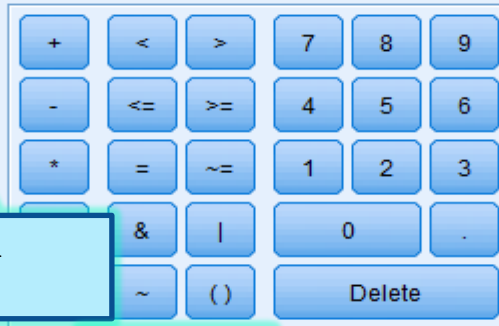
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation

+	<	>	7	8	9
-	<=	>=	4	5	6
*	=	~=	1	2	3
	&	|		0		.
~	( )	Delete			

Functions and Special Variables:

ID	DIET	LIFESPAN	CATEGORICAL	STD_LIFESPAN
86.00	Restricted Diet	60.50	<500	-2.82
1.00	Restricted Diet	105.00	<500	-2.65
9.00	Restricted Diet	193.00	<500	-2.33
16.00	Restricted Diet	211.00	<500	-2.27
23.00	Restricted Diet	236.00	<500	-2.18
30.00	Restricted Diet	302.00	<500	-1.94
37.00	Restricted Diet	363.00	<500	-1.72
44.00	Restricted Diet	389.00	<500	-1.62
51.00	Restricted Diet	390.00	<500	-1.62

Generated values

Reset Cancel Help

8

Recode into different variable

Class of LIFESPAN	Frequency	Percentage
<300	7	3.6
300-600	19	9.7
600-900	93	47.7
900-1200	55	28.2
≥1200	21	10.8

▷ The Target is to see the distribution of LIFESPAN

▷ "Recode into different variable" is used to create a new variable (CAT_LIFESPAN) from a source variable (LIFESPAN)

▷ **Syntax:**

RECODE LIFESPAN (Lowest thru 299=1) (300 thru 599=2) (600 thru 899=3) (900 thru 1199=4) (1200 thru Highest=5) INTO **CAT_LIFESPAN**.

VARIABLE LABELS CAT_LIFESPAN 'Distribution of Lifespan'.

VALUE LABELS CAT_LIFESPAN 1 "<300" 2 "300-600" 3 "600-900" 4 "900-1200" 5 "≥1200".

EXECUTE.

Recode into different variable

Recode into different variable

LIFESPAN <300

Recode into different variable

LIFESPAN 300-600

Recode into different variable

LIFESPAN ≥ 1200

Need to click on change

Recode into different variable

Before Recode

After Recode

14

Recode into same variable

Category of LIFESPAN	Frequency	Percentage
<300	7	3.6
300-600	19	9.7
600-900	93	47.7
900-1200	55	28.2
≥1200	21	10.8

> We wish to merge 1st two groups

▷ **Recode into same variable**

  ▷ Will overwrite the source variable

  ▷ Use when want to change the source variable permanently

  ▷ **Syntax:**

RECODE CAT_LIFESPAN (1=1) (2=1) (3=2) (4=3) (5=4).

VALUE LABELS CAT_LIFESPAN 1 '<600' 2 "600-900" 3 "900-1200" 4 "≥1200".

EXECUTE.

Recode into same variable

Recode into same variable

Category of LIFESPAN	Frequency	Percentage
<600	26	13.3
600-900	93	47.7
900-1200	55	28.2
≥1200	21	10.8

"DATA" Menu Bar

▷ The **Data** drop-down menu provides procedures for reorganizing the structure of a data file.

▷ Procedures shown in the figure allow to change the format of a data file.

▷ We will discuss

   ▷ Sort Cases

   ▷ Merge Files
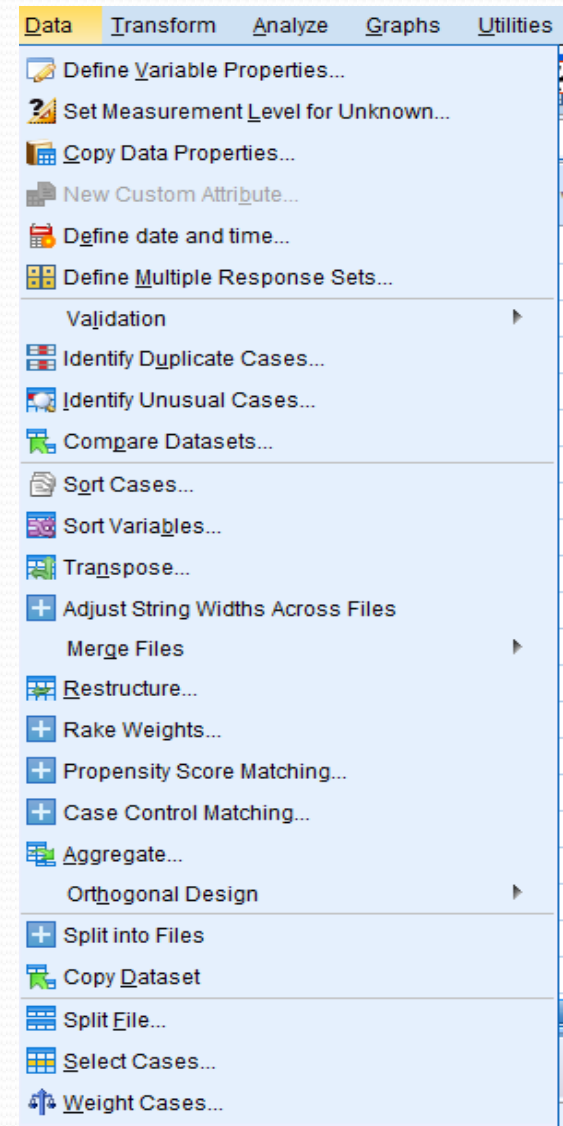
Sort Cases

▷ **Sort Cases** opens a dialogue box that allows sorting of cases (rows) in the spreadsheet according to the values of one or more variables.

▷ Sort the data set by age, sex, religion

▷ **Manual:** *Data> Sort Cases> Select one or more Variables>Ok*

▷ **Syntax:**

SORT CASES BY AGE(A)

➢ **A** indicates Ascending

➢ **D** indicates Descending
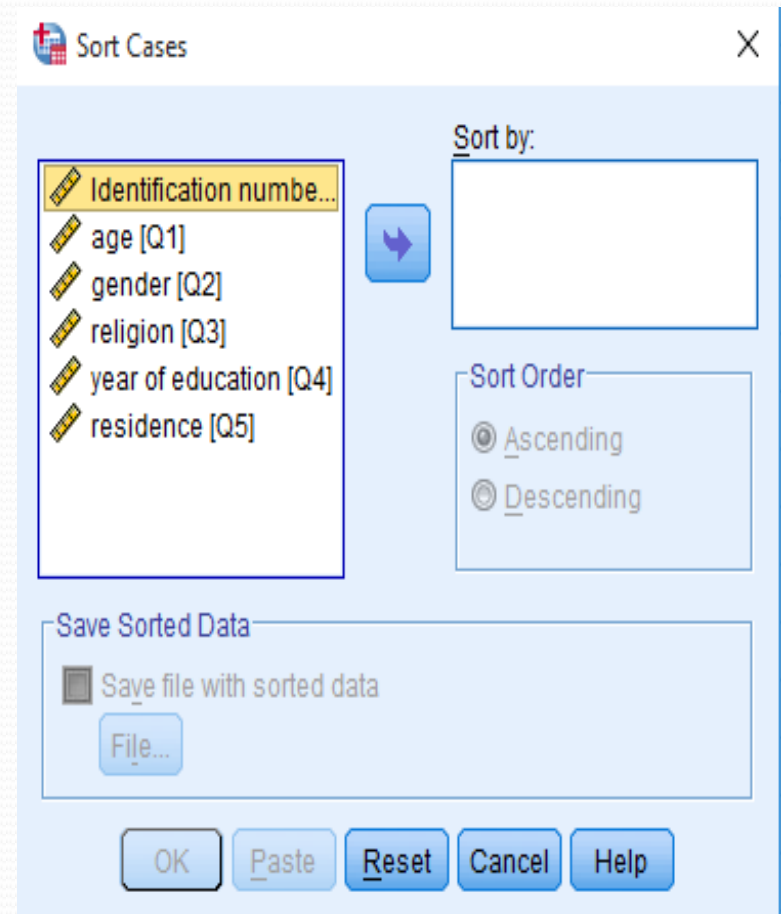
Sort Cases

Before sorting

	Id	Q1	Q2	Q3	Q4	Q5	var
1	1.00	45.00	male	muslim	16.00	rural	
2	2.00	35.00	female	hindu	12.00	rural	
3	3.00	33.00	female	others	14.00	urban	
4	4.00	23.00	male	hindu	12.00	rural	
5	5.00	18.00	male	muslim	15.00	urban	
6							

After sorting by Age

	Id	Q1	Q2	Q3	Q4	Q5	var
1	5.00	18.00	male	muslim	15.00	urban	
2	4.00	23.00	male	hindu	12.00	rural	
3	3.00	33.00	female	others	14.00	urban	
4	2.00	35.00	female	hindu	12.00	rural	
5	1.00	45.00	male	muslim	16.00	rural	
6							

Merge Two Datasets

▷ **Merge files** allows

   ▷ **Add cases:** When two data sets have same variables and combined them into one file.

   ▷ **Add variables:** One-to-one Merge when both data sets have unique cases.

   ▷ **Add variables:** One-to-many Merge when in one data set has duplicate cases.

Merge files: Add Cases

▷ **Add Cases** (new values of the existing variables)

    ▷ **Data1** - "data1.sav": 100 cases with two variables ID & DIET

    ▷ **Data2** - "data2.sav": 95 cases with two variables ID & DIET

▷ What do we want?

    ▷ Task 1: Merge **Data1** and **Data2** by ID – "data3.sav"

Merge files: Add Cases

▷ We want to add cases of "data2.sav" into "data1.sav" and save it as "data3.sav".

▷ **It is noted that variable names should be same in both data sets.**

▷ **Syntax:**

Get FILE='G:\Data File\data1.sav'.

***ADD FILES** /FILE=**

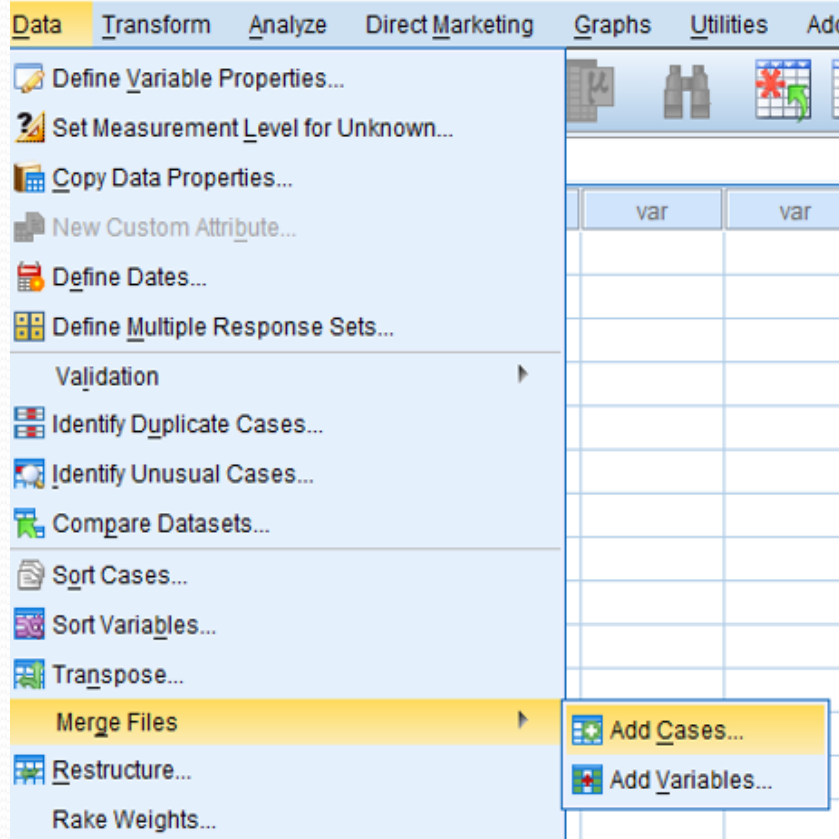 */FILE='G:\Data File\data2.sav'.*

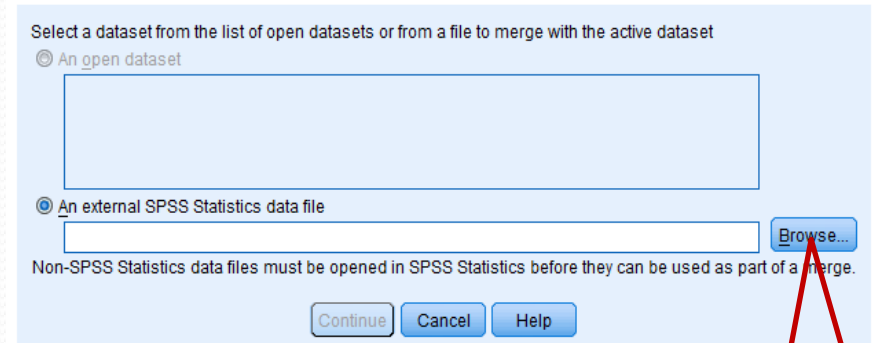*SAVE OUTFILE 'G:\Data File\data3.sav'.*

EXECUTE.

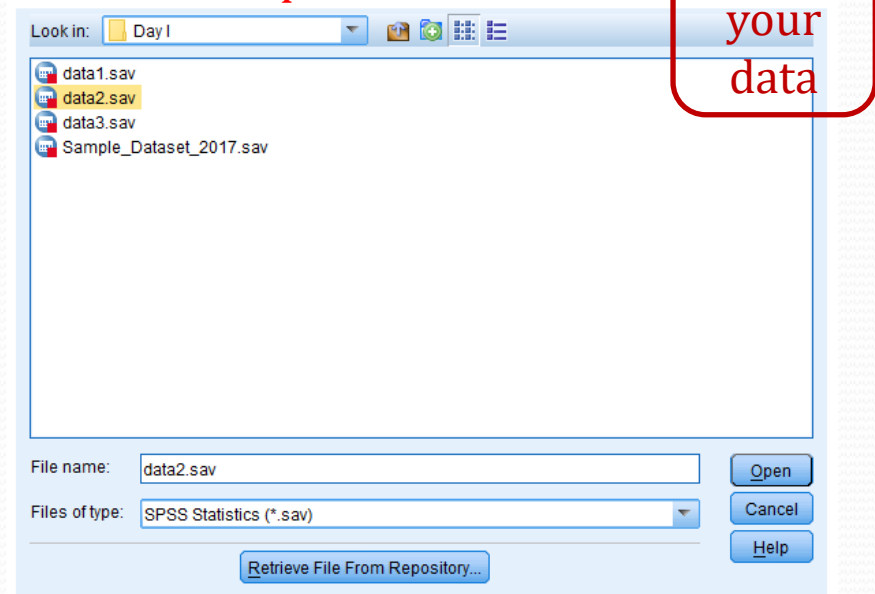**Lets Go to SPSS Window**

Merge files: Add Cases (Manually)

Step 1

Step 2

Step 3

Locate your data

Merge files: Add Cases

data1.sav

File Edit View Data Transform Analyze Direct Marketing Graph Utilities Add-ons Window Help

Visible: 2 of 2 Variables

	ID	DIET	var	var	va
95	95.00	Restricted Diet			
96	96.00	Restricted Diet			
97	97.00	Restricted Diet			
98	98.00	Restricted Diet			
99	99.00	Restricted Diet			
100	100.00	Restricted Diet			
101					

data2.sav

data2.sav [DataSet4] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graph Utilities Add-ons Window Help

Visible: 2 of 2 Variables

	ID	DIET	var	var	va
1	101.00	Restricted Diet			
2	102.00	Restricted Diet			
3	103.00	Restricted Diet			
4	104.00	Restricted Diet			
5	105.00	Restricted Diet			
6	106.00	Restricted Diet			
7	107.00	Free Diet			

Saving the new File

Step 6

Look in: Day I

- data1.sav
- data2.sav
- data3.sav
- Sample_Dataset_2017.sav

Keeping 2 of 2 variables.

File name: data4.sav

Save as type: SPSS Statistics (*.sav)

Encoding:

- ☑ Write variable names to spreadsheet
- ☐ Save value labels where defined instead of data values
- ☐ Save value labels into a .sas file
- ☐ Encrypt file with password

[Variables...] [Save] [Paste] [Cancel] [Help]

Store File To Repository...

data3.sav

data4.sav [DataSet20] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct M

	ID	DIET	var
94	94.00	Restricted Diet	
95	95.00	Restricted Diet	
96	96.00	Restricted Diet	
97	97.00	Restricted Diet	
98	98.00	Restricted Diet	
99	99.00	Restricted Diet	
100	100.00	Restricted Diet	
101	101.00	Restricted Diet	
102	102.00	Restricted Diet	
103	103.00	Restricted Diet	
104	104.00	Restricted Diet	
105	105.00	Restricted Diet	

Merge files: Add Variables

▷ Suppose we have two data sets "dads.sav" and "faminc.sav"

▷ We want add variables of "faminc.sav" into "dads.sav" file

dads.sav

famid	name	inc
2.00	Art	22000.00
1.00	Bill	30000.00
3.00	Paul	25000.00

faminc.sav

famid	faminc96	faminc97	faminc98
3.00	75000.00	76000.00	77000.00
1.00	40000.00	40500.00	41000.00
2.00	45000.00	45400.00	45800.00

Merge files: Add Variables

▷ This is a one to one merge because there is a one to one correspondence between the dads and faminc records.

▷ Syntax:

```
GET FILE="D:\spss\dads.sav".
SORT CASES BY famid.
SAVE OUTFILE="D:\spss\dads2.sav".

GET FILE="D:\spss\faminc.sav".
SORT CASES BY famid.
SAVE OUTFILE="D:\spss\faminc2.sav".

MATCH FILES FILE="D:\spss\dads2.sav"
 /FILE="D:\spss\faminc2.sav"
 /BY famid.
SAVE OUTFILE="D:\spss\OnetoOneMerge.sav".
```
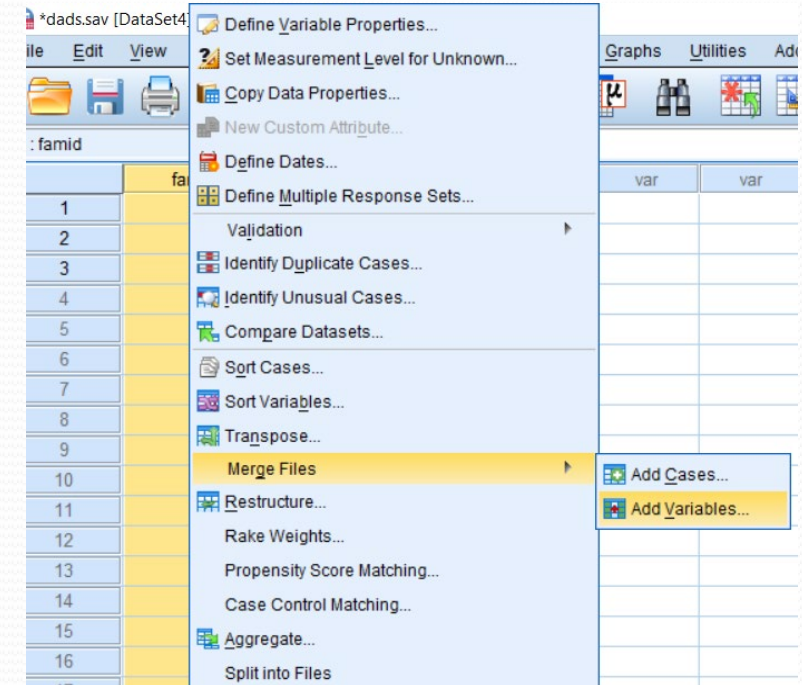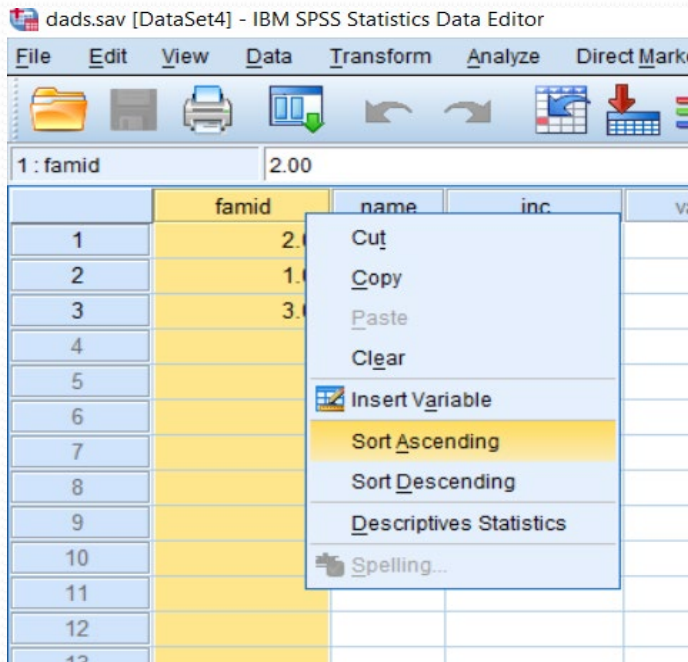
Merge files: Add Variables

Step-1: Both files need to have a common indexing key, so make sure both files are sorted in ascending order before trying to merge.
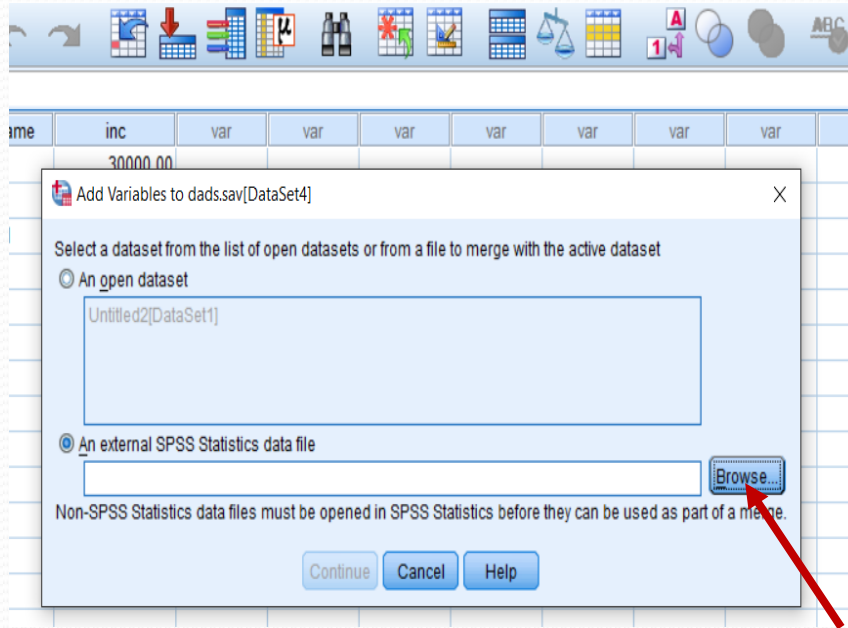
Step-2: Open the first file that you wish to merge. Next

Data>Merge Files>Add Variables

Merge files: Add Variables

Step-3: Select the file you wish to merge

Step-4:
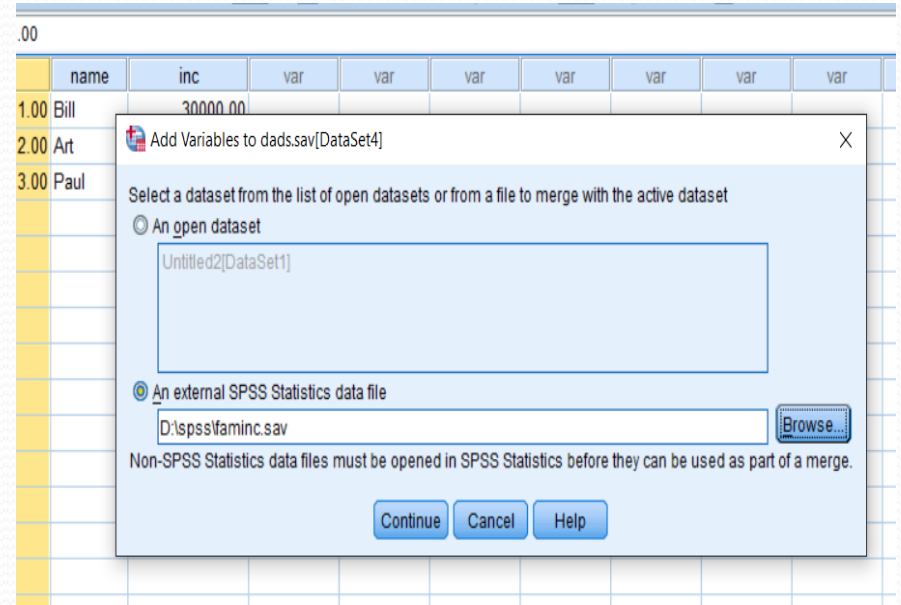
Merge files: Add Variables

Step-5: If both keys have the same name one will show up in the 'Excluded Variables' list. Select the key variable from the 'Excluded Variables' list. Check the 'Match cases on key variables in sorted files' checkbox. Select the 'Both files provide cases' radio button. Now press the right arrow button that is next to the 'Key Variables:' list box.
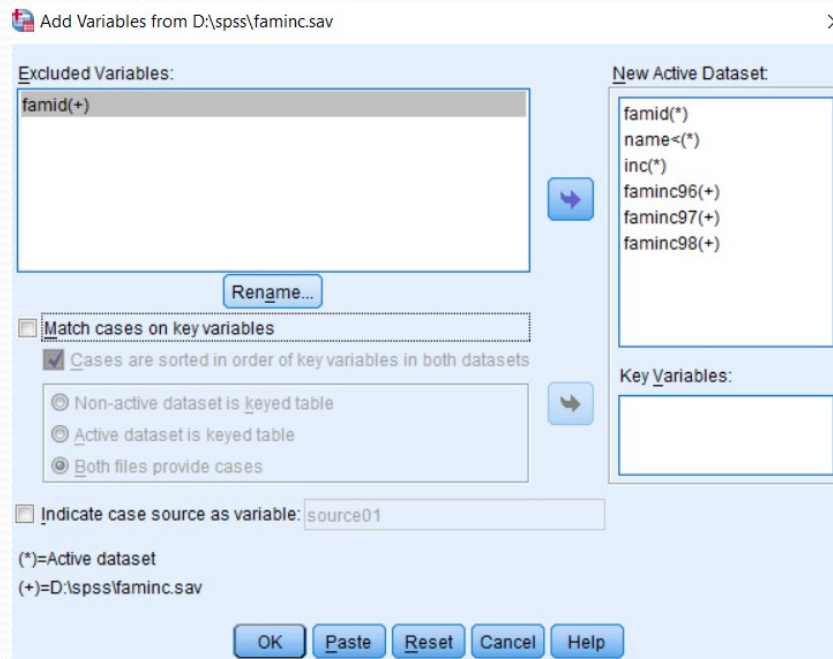
Merge files: Add Variables

Step-6: SPSS will give you a warning regarding sorted key variables. Make sure both files were sorted in ascending order before trying to do a file merge.

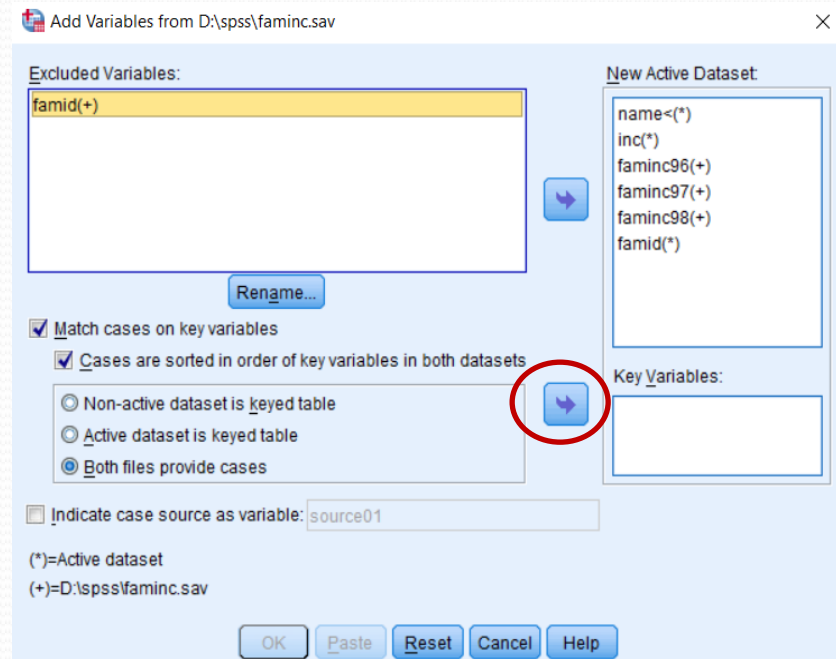Step-7: All variables from both files will be merged in a new datafile. Cases were matched by the key variable (famid in this example)
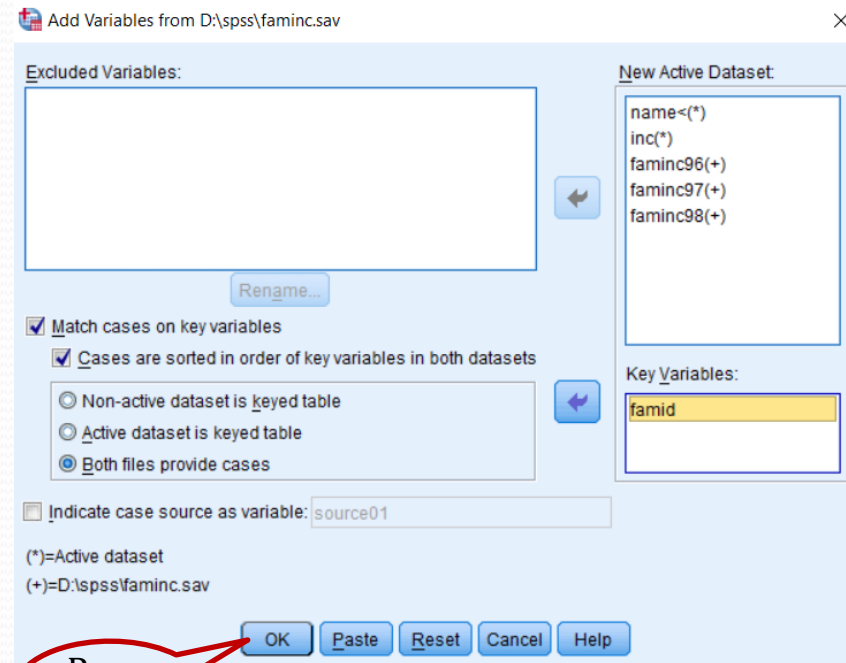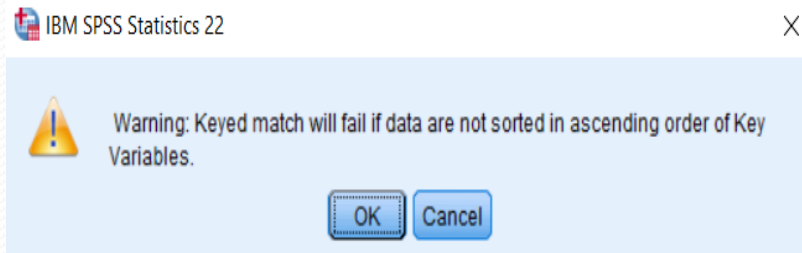
Merge files: One to Many Merge

▷ We have a file named **dads.sav** & a file with **kids.sav** where a dad could have more than one kid.

▷ This is called a one to many merge since we are matching one dad observation to one or more (many) kids observations.

dads.sav

famid	name	inc
2.00	Art	22000.00
1.00	Bill	30000.00
3.00	Paul	25000.00

Kids.sav

famid	kidname	birth	age	wt	sex
1.00	Beth	1.00	9.00	60.00	f
1.00	Bob	2.00	6.00	40.00	m
1.00	Barb	3.00	3.00	20.00	f
2.00	Andy	1.00	8.00	80.00	m
2.00	Al	2.00	6.00	50.00	m
2.00	Ann	3.00	2.00	20.00	f
3.00	Pete	1.00	6.00	60.00	m
3.00	Pam	2.00	4.00	40.00	f
3.00	Phil	3.00	2.00	20.00	m

Merge files: One to Many Merge

▷ **Syntax**

```
GET FILE="D:\spss\dads.sav".
SORT CASES BY famid.
SAVE OUTFILE="D:\spss\dads2.sav".


GET FILE="D:\spss\kids.sav".
SORT CASES BY famid.
SAVE OUTFILE="D:\spss\kids2.sav".


MATCH FILES FILE="D:\spss\kids2.sav"
 /Table="D:\spss\dads2.sav"
 /BY famid.


SAVE OUTFILE="D:\spss\OnetoManyMerge.sav".
```
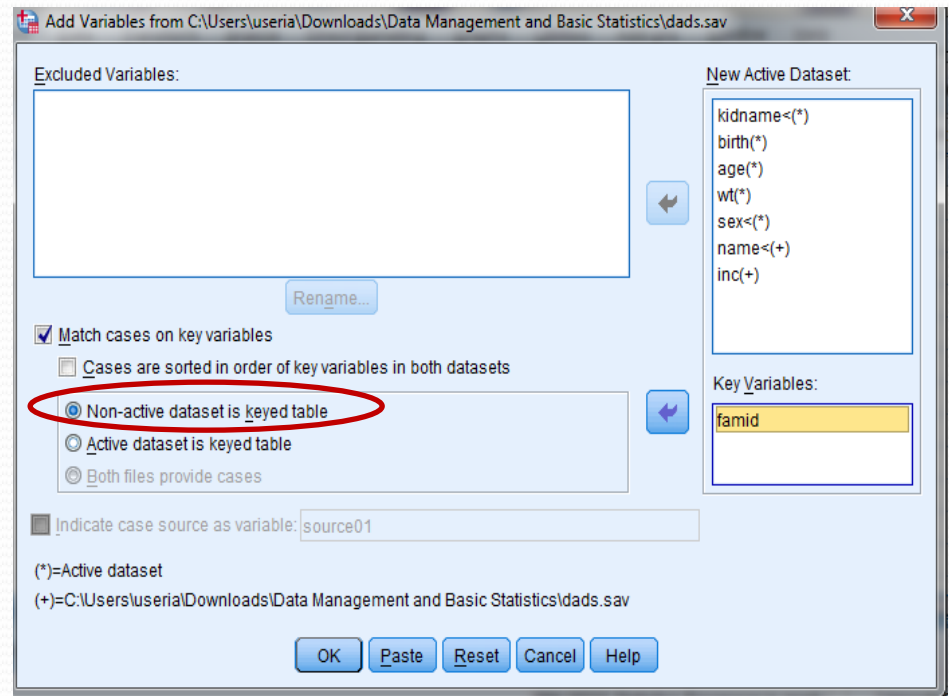
▷ **Manually:** Just like the one to one merge.

▷ **One change is required**

Add Variables from C:\Users\useria\Downloads\Data Management and Basic Statistics\dads.sav

Excluded Variables:

New Active Dataset:
kidname<(*)
birth(*)
age(*)
wt(*)
sex<(*)
name<(+)
inc(+)

Rename...

☑ Match cases on key variables
☐ Cases are sorted in order of key variables in both datasets
◉ Non-active dataset is keyed table
○ Active dataset is keyed table
○ Both files provide cases

Key Variables:
famid

☐ Indicate case source as variable: source01

(*)=Active dataset
(+)=C:\Users\useria\Downloads\Data Management and Basic Statistics\dads.sav

OK   Paste   Reset   Cancel   Help

34

Merge files: One to Many Merge

Kids.sav

famid	kidname	birth	age	wt	sex
1.00	Beth	1.00	9.00	60.00	f
1.00	Bob	2.00	6.00	40.00	m
1.00	Barb	3.00	3.00	20.00	f
2.00	Andy	1.00	8.00	80.00	m
2.00	Al	2.00	6.00	50.00	m
2.00	Ann	3.00	2.00	20.00	f
3.00	Pete	1.00	6.00	60.00	m
3.00	Pam	2.00	4.00	40.00	f
3.00	Phil	3.00	2.00	20.00	m

dads.sav

famid	name	inc
2.00	Art	22000.00
1.00	Bill	30000.00
3.00	Paul	25000.00

famid	kidname	birth	age	wt	sex	name	inc
1.00	Beth	1.00	9.00	60.00	f	Bill	30000.00
1.00	Bob	2.00	6.00	40.00	m	Bill	30000.00
1.00	Barb	3.00	3.00	20.00	f	Bill	30000.00
2.00	Andy	1.00	8.00	80.00	m	Art	22000.00
2.00	Al	2.00	6.00	50.00	m	Art	22000.00
2.00	Ann	3.00	2.00	20.00	f	Art	22000.00
3.00	Pete	1.00	6.00	60.00	m	Paul	25000.00
3.00	Pam	2.00	4.00	40.00	f	Paul	25000.00
3.00	Phil	3.00	2.00	20.00	m	Paul	25000.00

Aggregate Statistics

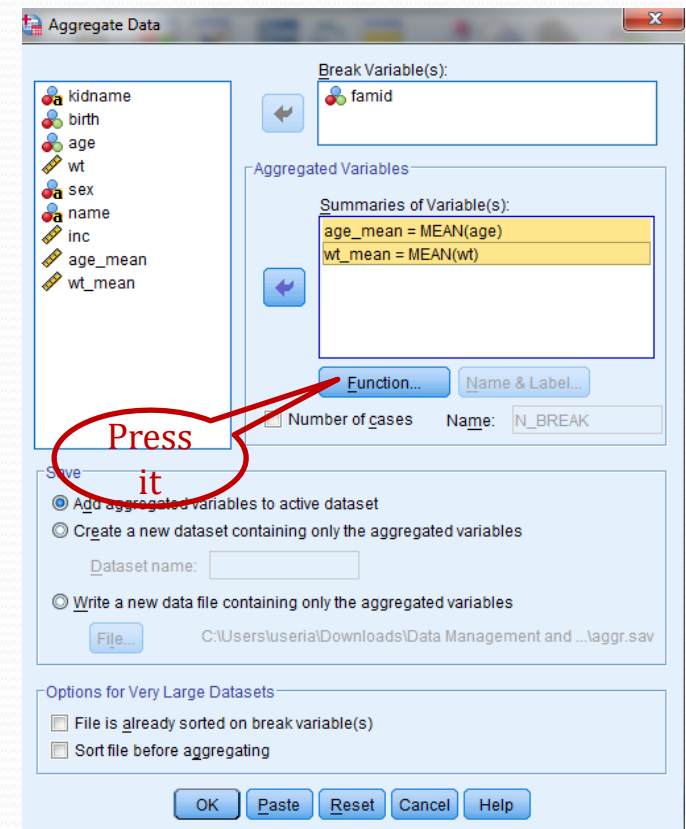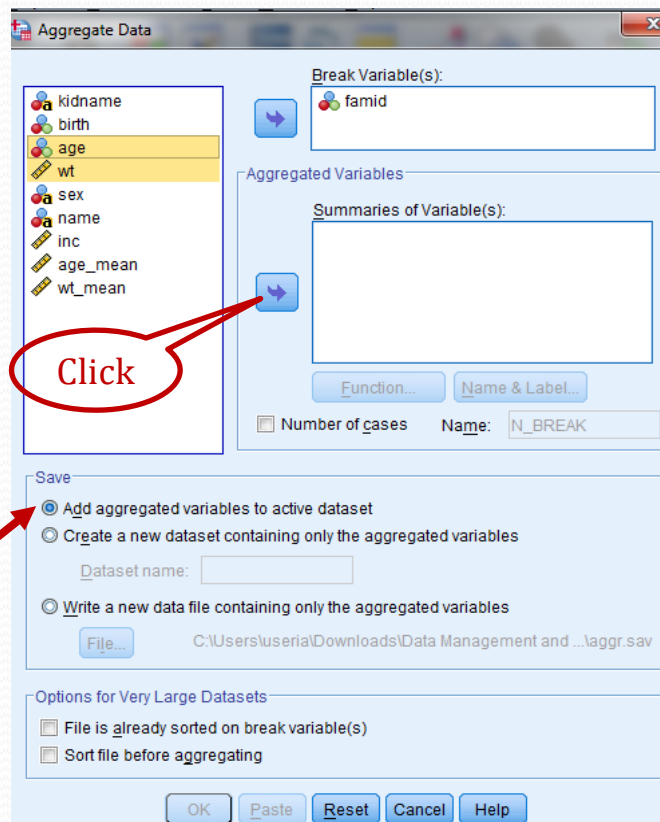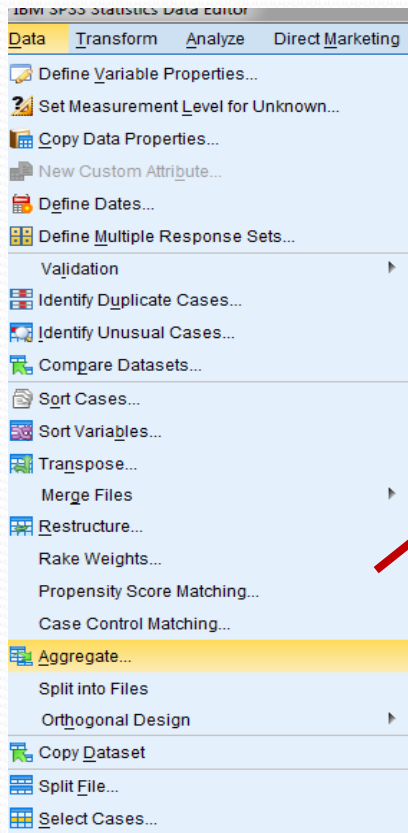▷ Find mean age and weight of children for each family.

Syntax:
AGGREGATE
 /OUTFILE=* MODE=ADDVARIABLES
 /BREAK=famid
 /age_mean_1=MEAN(age)
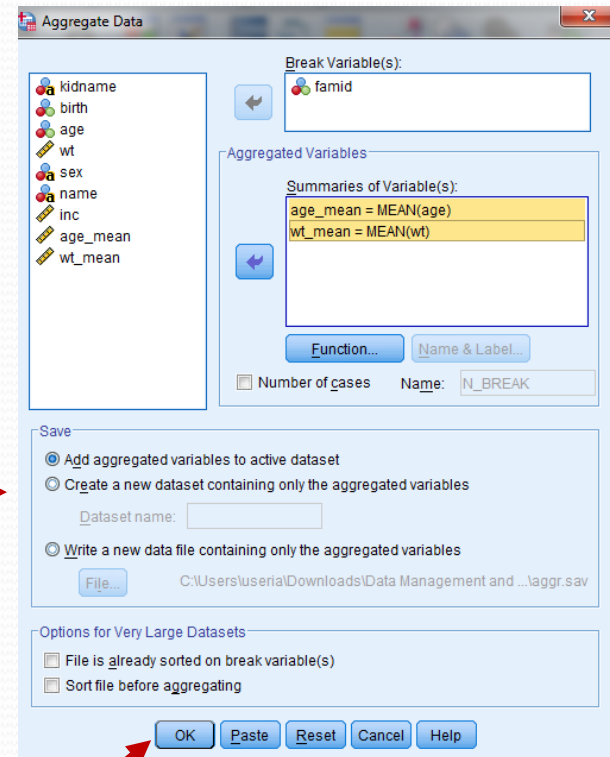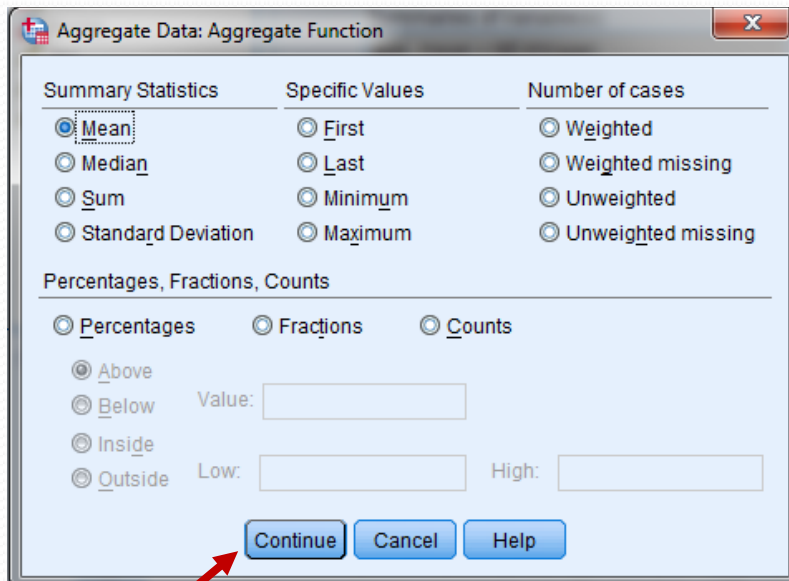 /wt_mean_1=MEAN(wt).

Aggregate Statistics

▷ Mean age and weight (Manually)

Aggregate Statistics

▷ Mean age and weight (Manually)

Store Aggregate Statistics

▷ Mean age and weight (Manually)

famid	kidname	birth	age	wt	sex	name	inc	age_mean	wt_mean
1.00	Beth	1.00	9.00	60.00	f	Bill	30000.00	6.00	40.00
1.00	Bob	2.00	6.00	40.00	m	Bill	30000.00	6.00	40.00
1.00	Barb	3.00	3.00	20.00	f	Bill	30000.00	6.00	40.00
2.00	Andy	1.00	8.00	80.00	m	Art	22000.00	5.33	50.00
2.00	Al	2.00	6.00	50.00	m	Art	22000.00	5.33	50.00
2.00	Ann	3.00	2.00	20.00	f	Art	22000.00	5.33	50.00
3.00	Pete	1.00	6.00	60.00	m	Paul	25000.00	4.00	40.00
3.00	Pam	2.00	4.00	40.00	f	Paul	25000.00	4.00	40.00
3.00	Phil	3.00	2.00	20.00	m	Paul	25000.00	4.00	40.00

Aggregate Statistics

▷ Find the number of children for each family.

Syntax:
AGGREGATE
/OUTFILE=* MODE=ADDVARIABLES
/BREAK=famid
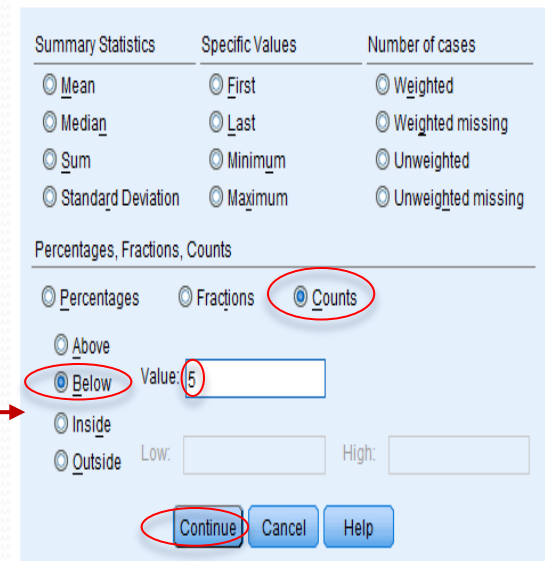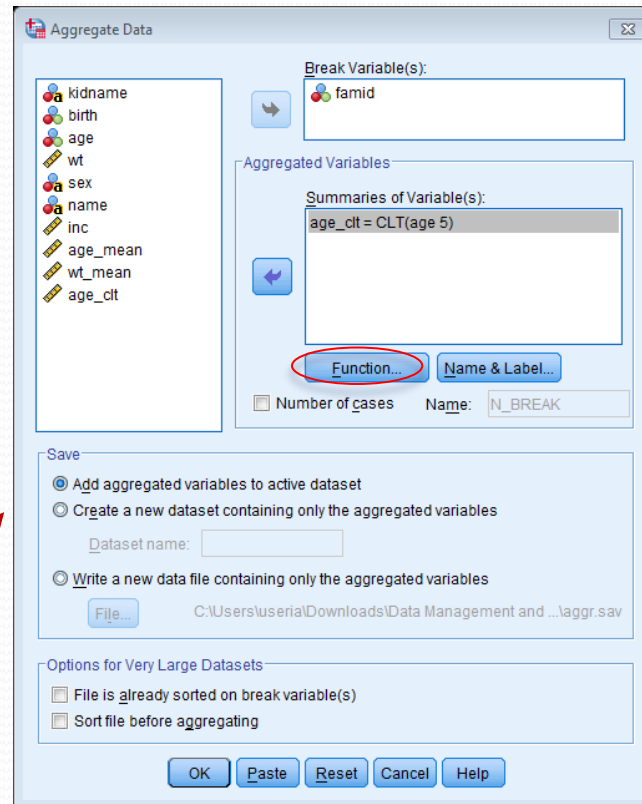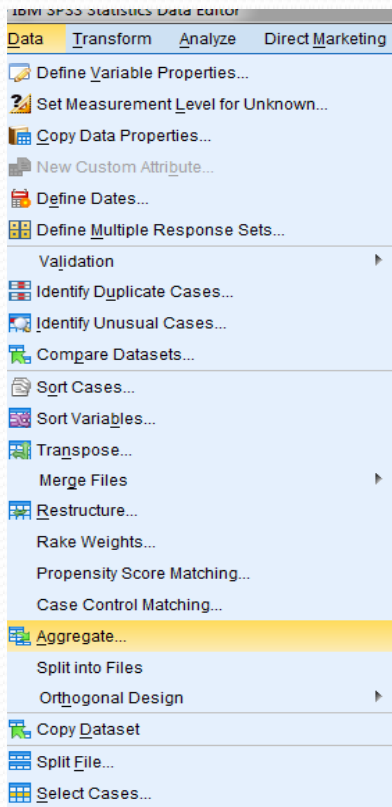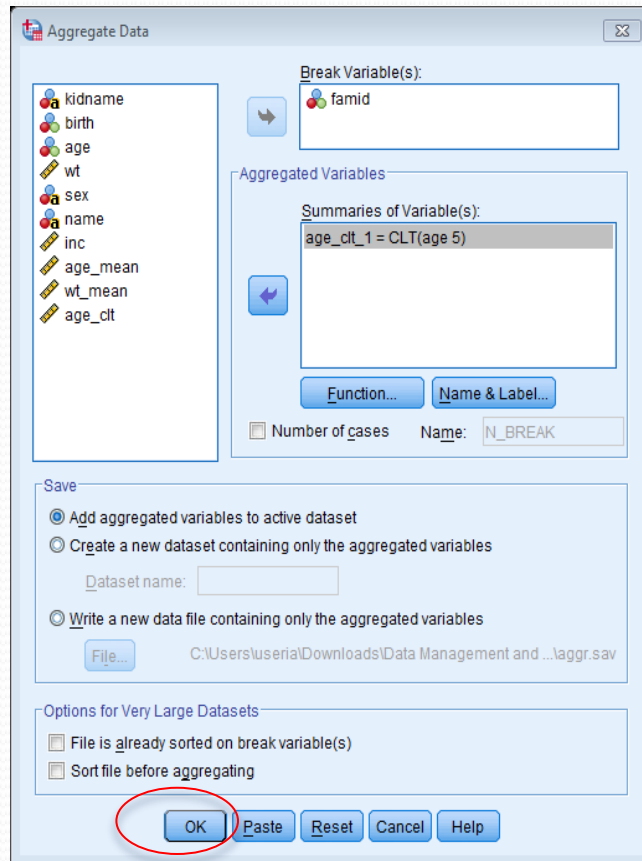/kidname_nu=NU(kidname).

Aggregate Statistics (Count)

▷ Find the number of children for each family.

Aggregate Statistics (Count)

▷ Find the number of children for each family.

Exercise

▷ Suppose we have two data set. How can we solve the problems in SPSS given next page?

dat1

dat2

HH ID	MAI	MAE
1	5000	4500
2	3500	4000
3	6000	5000

Child ID	1	5	2	4	3	6	7	10	9	8
HH ID	1	2	1	3	1	2	2	1	3	2
Sex of child	m	m	f	f	m	F	m	f	m	m
Age	10	15	8	5	5	10	4	1	2	1

▷ Combine information of dat2 with those of dat1 using household ID?

▷ Calculate mean (Mean_Age) and standard deviation (SD_Age) of children age by household ID.

▷ Can you calculate number of children (Children_number) per household? If yes, how?

▷ Add Mean_Age, SDS_Age, and Children_number in the dat1 file.

Thank You!