Basic Stats in PSPP

For NICAR 2016 Denver / Norm Lewis, University of Florida, nplewis@ufl.edu

PART 1: INTRODUCTION

What is PSPP?

PSPP is free software that emulates SPSS.

OK, so what is SPSS?

SPSS stands for Statistical Package for the Social Sciences. It was created in 1968 to allow social scientists (and others) to perform statistical tests. IBM purchased it in 2009 and renamed it IBM SPSS. It and SAS are used widely for data analysis.

Then why PSPP?

SPSS is excellent. It also costs more than $1,000. So the author of PSPP (the acronym does not stand for anything) created a free program that follows the nomenclature and style of SPSS. Unlike the statistics program R, which is also free, PSPP does not require that you learn a new language. And its graphical-user interface is an advantage for those less comfortable with a command-line environment.

PSPP is not as competent as its paid rival For example, PSPP misinterpreted a CSV file I created that SPSS handled just fine. And PSPP is missing some advanced statistical tests. However, PSPP performs enough statistics to be useful for journalists.

Why would journalists use statistics?

Life is random. What appears to be a pattern may just be chance. Statistics let us distinguish between a meaningful pattern and random chance. In turn, that lets journalists avoid mistaking coincidence for significance.

For example, does that cancer cluster have an environmental cause or is it just bad luck? Are improvements in school test scores the product of a new curriculum or good luck? Is a change in the local crime rate substantially different from the state or national average? In other words, are these differences newsworthy or the product of chance? Statistics can tell us.

PART 2: INSTALLATION

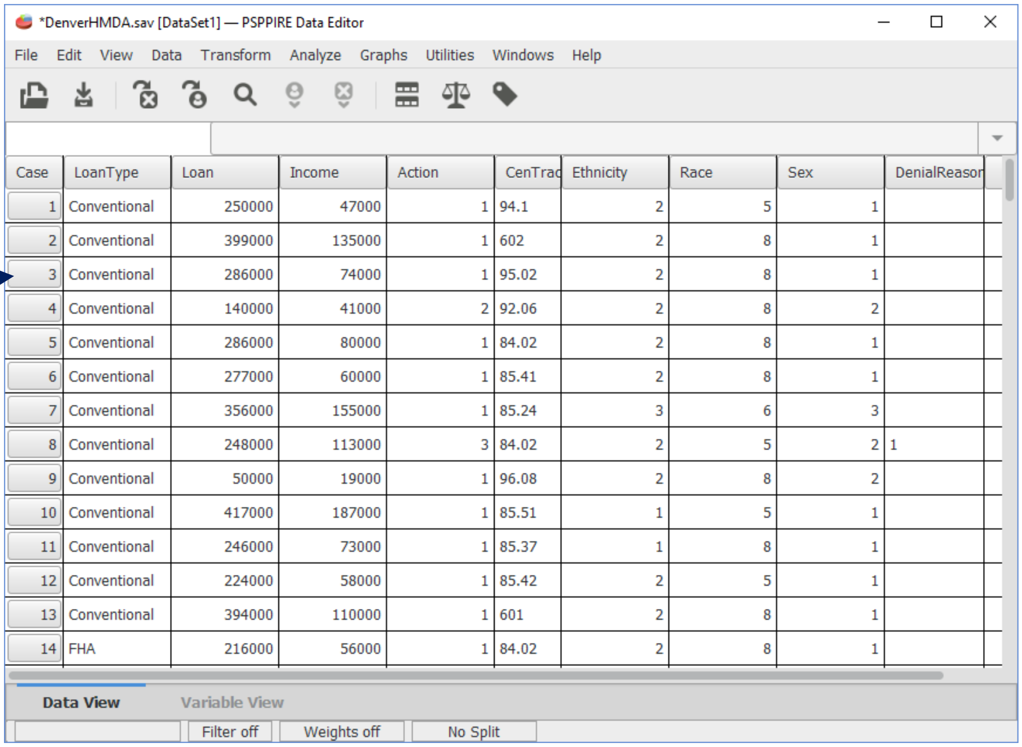
For Windows users, the easiest installation is to download a compiled PSPP from Sourceforge. The file is about 24 MB. (If you prefer binary packages, consult GNU.)

For Macintosh users, installation is more complicated, according to Softpedia, which has created an installation tutorial.

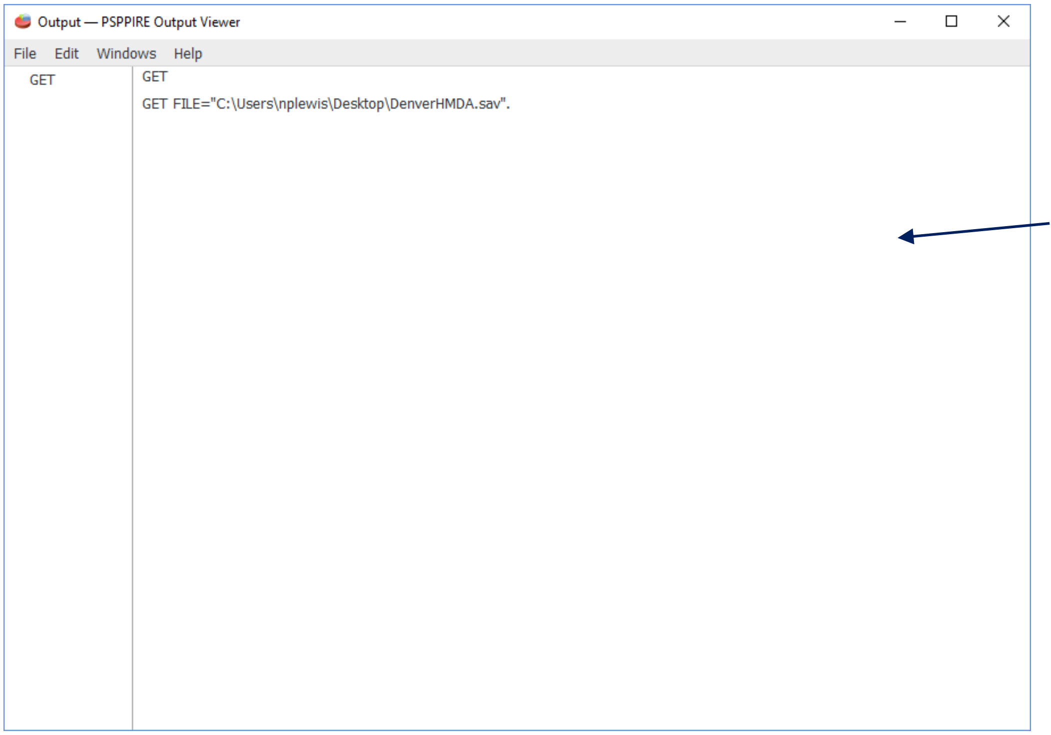
PART 3: USER INTERFACE

Open up the data file (with a .sav suffix, just like SPSS). You will see two windows.

This is the Data Editor Window, used to work with data and variables.



Case	LoanType	Loan	Income	Action	CenTrac	Ethnicity	Race	Sex	DenialReason
1	Conventional	250000	47000	1	94.1	2	5	1	
2	Conventional	399000	135000	1	602	2	8	1	
3	Conventional	286000	74000	1	95.02	2	8	1	
4	Conventional	140000	41000	2	92.06	2	8	2	
5	Conventional	286000	80000	1	84.02	2	8	1	
6	Conventional	277000	60000	1	85.41	2	8	1	
7	Conventional	356000	155000	1	85.24	3	6	3	
8	Conventional	248000	113000	3	84.02	2	5	2	1
9	Conventional	50000	19000	1	96.08	2	8	2	
10	Conventional	417000	187000	1	85.51	1	5	1	
11	Conventional	246000	73000	1	85.37	1	8	1	
12	Conventional	224000	58000	1	85.42	2	5	1	
13	Conventional	394000	110000	1	601	2	8	1	
14	FHA	216000	56000	1	84.02	2	8	1	

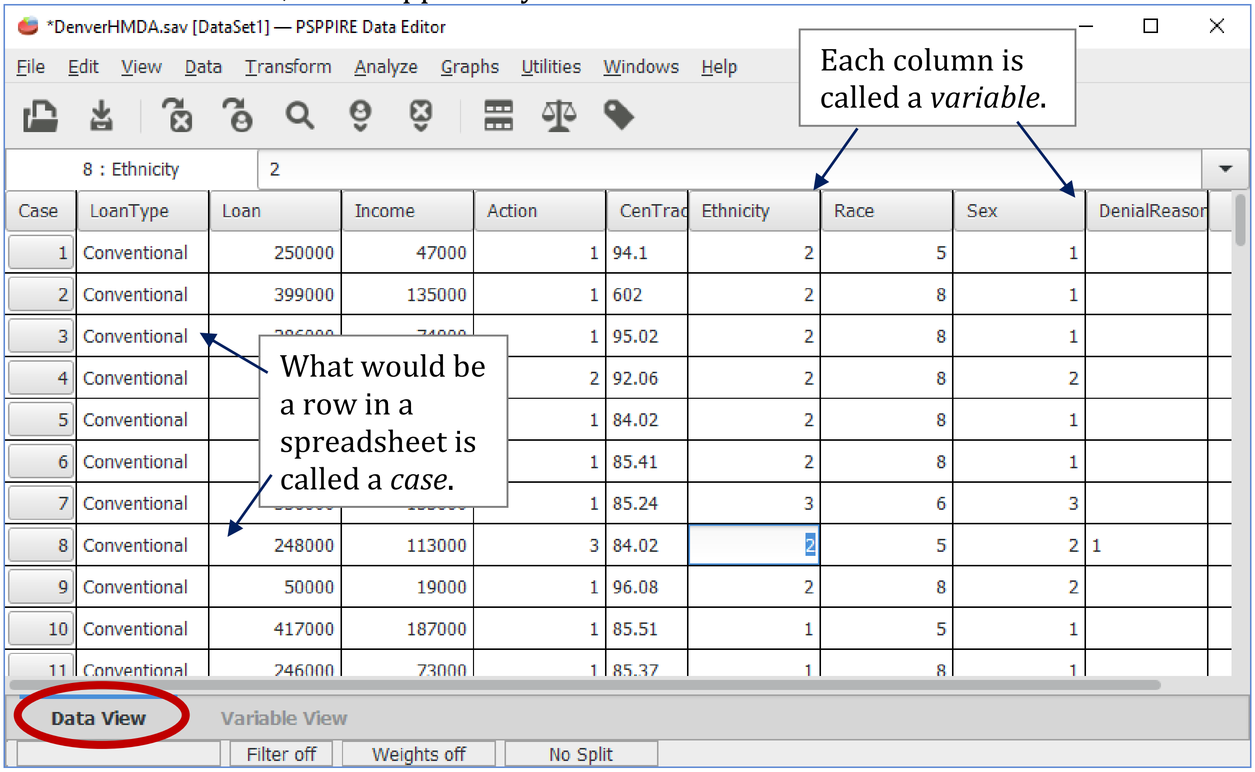


```
GET
GET FILE="C:\Users\nplewis\Desktop\DenverHMDA.sav".
```

This is the Viewer Window, used to output commands issued from the Data Editor such as statistical tests.

The Data Editor Window has two views.

This is the data view, which appears by default.

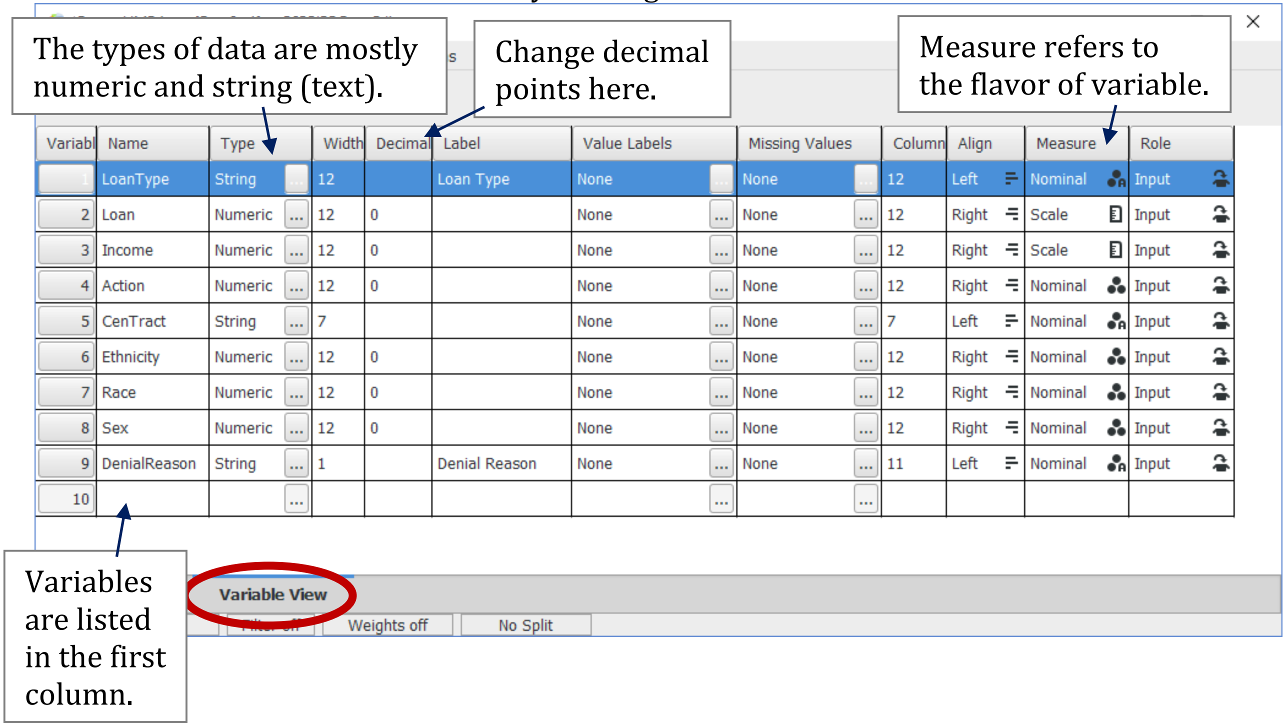

*DenverHMDA.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help
8 : Ethnicity
2
Data View
Variable View
Filter off
Weights off
No Split

Each column is called a *variable*.

What would be a row in a spreadsheet is called a *case*.

Case	LoanType	Loan	Income	Action	CenTrac	Ethnicity	Race	Sex	DenialReason
1	Conventional	250000	47000	1	94.1	2	5	1	
2	Conventional	399000	135000	1	602	2	8	1	
3	Conventional	[illegible]	[illegible]	1	95.02	2	8	1	
4	Conventional			2	92.06	2	8	2	
5	Conventional			1	84.02	2	8	1	
6	Conventional			1	85.41	2	8	1	
7	Conventional	[illegible]	[illegible]	1	85.24	3	6	3	
8	Conventional	248000	113000	3	84.02	2	5	2	1
9	Conventional	50000	19000	1	96.08	2	8	2	
10	Conventional	417000	187000	1	85.51	1	5	1	
11	Conventional	246000	73000	1	85.37	1	8	1	

This is the Variable View, obtained by clicking on those words at the bottom.



The types of data are mostly numeric and string (text).

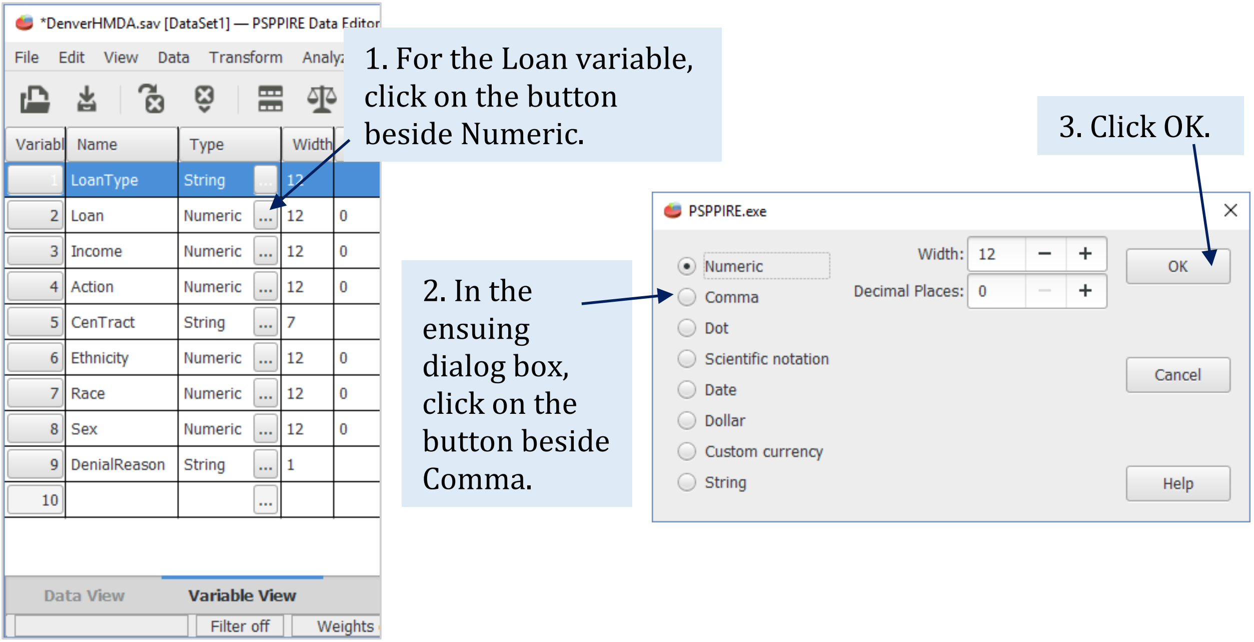
Change decimal points here.

Measure refers to the flavor of variable.

Variabl	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Column	Align	Measure	Role
1	LoanType	String	12		Loan Type	None	None	12	Left	Nominal	Input
2	Loan	Numeric	12	0		None	None	12	Right	Scale	Input
3	Income	Numeric	12	0		None	None	12	Right	Scale	Input
4	Action	Numeric	12	0		None	None	12	Right	Nominal	Input
5	CenTract	String	7			None	None	7	Left	Nominal	Input
6	Ethnicity	Numeric	12	0		None	None	12	Right	Nominal	Input
7	Race	Numeric	12	0		None	None	12	Right	Nominal	Input
8	Sex	Numeric	12	0		None	None	12	Right	Nominal	Input
9	DenialReason	String	1		Denial Reason	None	None	11	Left	Nominal	Input
10											

Variables are listed in the first column.

Here is an example of how the Variable View controls the Data View.



Variabl	Name	Type	Width	
1	LoanType	String ...	12	
2	Loan	Numeric ...	12	0
3	Income	Numeric ...	12	0
4	Action	Numeric ...	12	0
5	CenTract	String ...	7	
6	Ethnicity	Numeric ...	12	0
7	Race	Numeric ...	12	0
8	Sex	Numeric ...	12	0
9	DenialReason	String ...	1	
10		...		

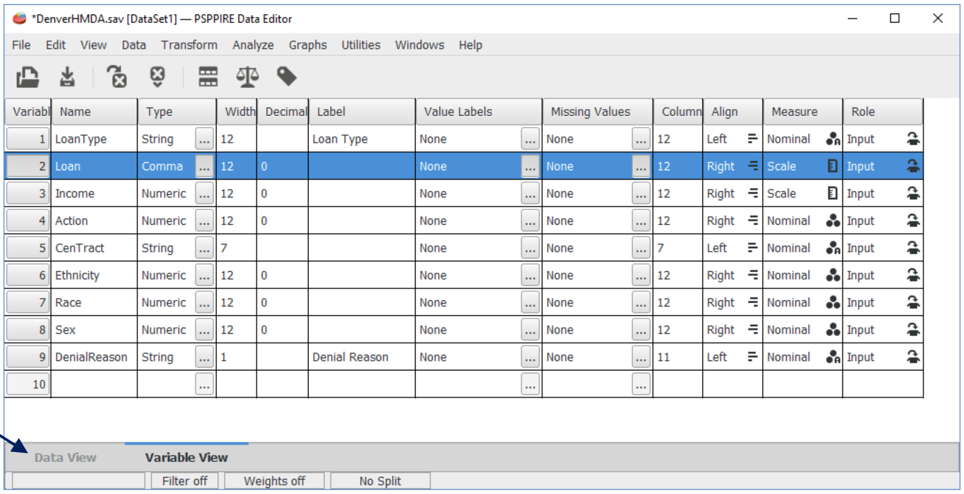
1. For the Loan variable, click on the button beside Numeric.

2. In the ensuing dialog box, click on the button beside Comma.

3. Click OK.

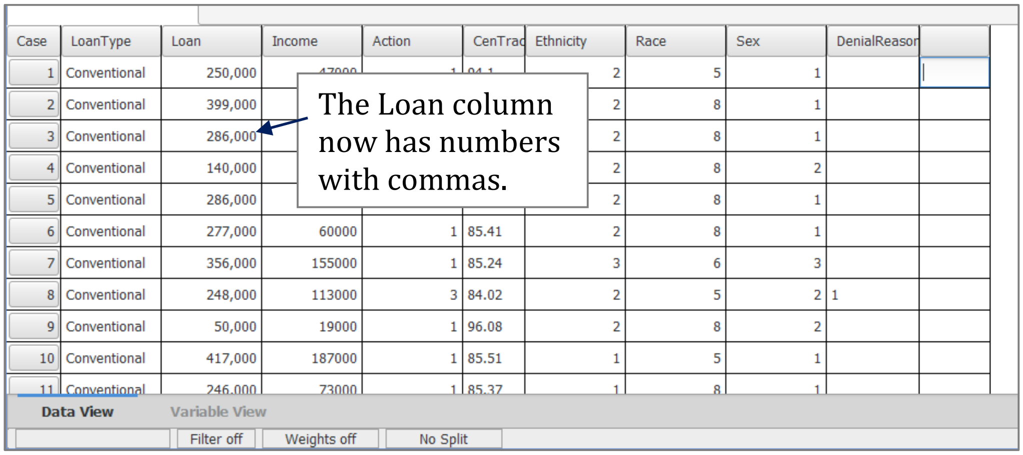
● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String

Width: 12 − +
Decimal Places: 0 − +



Variabl	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Column	Align	Measure	Role
1	LoanType	String ...	12		Loan Type	None ...	None ...	12	Left	Nominal	Input
2	Loan	Comma ...	12	0		None ...	None ...	12	Right	Scale	Input
3	Income	Numeric ...	12	0		None ...	None ...	12	Right	Scale	Input
4	Action	Numeric ...	12	0		None ...	None ...	12	Right	Nominal	Input
5	CenTract	String ...	7			None ...	None ...	7	Left	Nominal	Input
6	Ethnicity	Numeric ...	12	0		None ...	None ...	12	Right	Nominal	Input
7	Race	Numeric ...	12	0		None ...	None ...	12	Right	Nominal	Input
8	Sex	Numeric ...	12	0		None ...	None ...	12	Right	Nominal	Input
9	DenialReason	String ...	1		Denial Reason	None ...	None ...	11	Left	Nominal	Input
10		...				...	...				

4. Click on Data View.



Case	LoanType	Loan	Income	Action	CenTrac	Ethnicity	Race	Sex	DenialReason
1	Conventional	250,000	47000	1	94.1	2	5	1	
2	Conventional	399,000				2	8	1	
3	Conventional	286,000				2	8	1	
4	Conventional	140,000				2	8	2	
5	Conventional	286,000				2	8	1	
6	Conventional	277,000	60000	1	85.41	2	8	1	
7	Conventional	356,000	155000	1	85.24	3	6	3	
8	Conventional	248,000	113000	3	84.02	2	5	2	1
9	Conventional	50,000	19000	1	96.08	2	8	2	
10	Conventional	417,000	187000	1	85.51	1	5	1	
11	Conventional	246,000	73000	1	85.37	1	8	1	

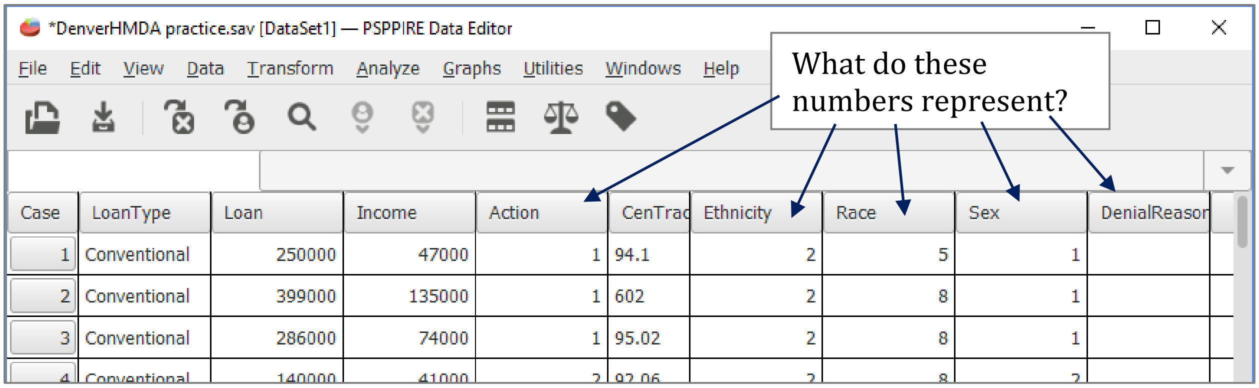
The Loan column now has numbers with commas.

PART 4: PREPARING THE DATA FOR ANALYSIS

This data comes from the Home Mortgage Disclosure Act, or HMDA ("hum-dah"), managed by the U.S. Consumer Financial Protection Bureau. Data are released in September for the previous calendar year.

This 2014 data is for the Denver metropolitan area. It is for mortgages to buy a single-family home to be occupied by the owner. It excludes loans for other dwellings or for refinancing or remodeling. Several variables were reduced to simply this tutorial.

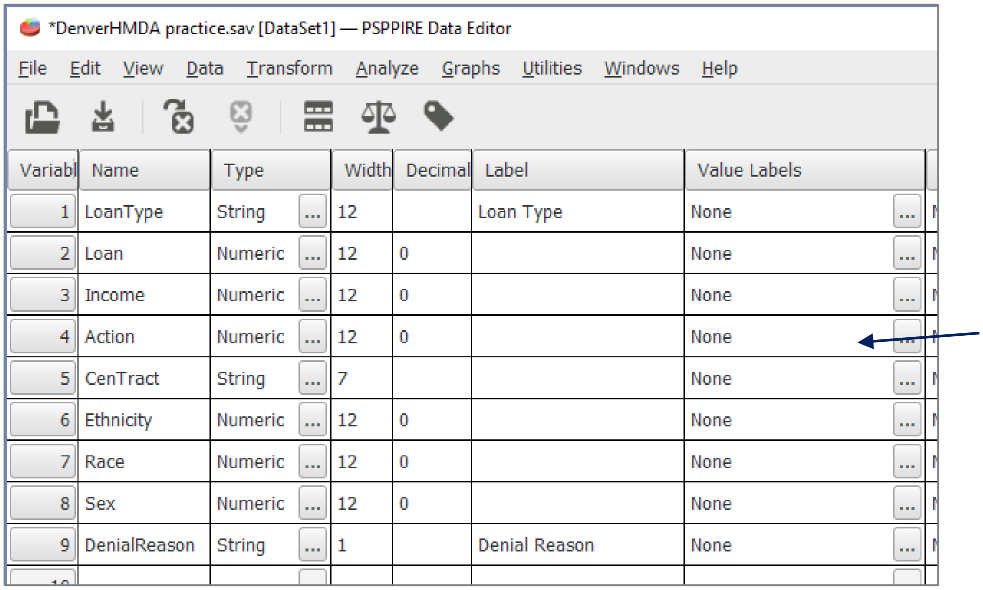
Loan and Income are numbers that reflect actual values. But variables like Action, Ethnicity, Race and Sex have no obvious meaning. What ethnicity is represented by a "2"?


*DenverHMDA practice.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help

What do these numbers represent?

Case	LoanType	Loan	Income	Action	CenTrac	Ethnicity	Race	Sex	DenialReason
1	Conventional	250000	47000	1	94.1	2	5	1	
2	Conventional	399000	135000	1	602	2	8	1	
3	Conventional	286000	74000	1	95.02	2	8	1	
4	Conventional	140000	41000	2	92.06	2	8	2	

We click on Variable View to see if these have variables have labels. None do.


*DenverHMDA practice.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help

Variabl	Name	Type	Width	Decimal	Label	Value Labels
1	LoanType	String ...	12		Loan Type	None ...
2	Loan	Numeric ...	12	0		None ...
3	Income	Numeric ...	12	0		None ...
4	Action	Numeric ...	12	0		None ...
5	CenTract	String ...	7			None ...
6	Ethnicity	Numeric ...	12	0		None ...
7	Race	Numeric ...	12	0		None ...
8	Sex	Numeric ...	12	0		None ...
9	DenialReason	String ...	1		Denial Reason	None ...

None of these have labels, so we will have to enter them.

Where do we get the labels? From something called a *code sheet*, as the next page shows.

The Code Sheet

To know what each number represents, we need the code sheet. It can be found on the original HMDA website and choosing the 2014 code sheet.

Action

1. Loan originated
2. Application approved but not accepted
3. Application denied by financial institution
4. Application withdrawn by applicant
5. File closed for incompleteness
6. Loan purchased by the institution
7. Preapproval request denied by financial institution
8. Preapproval request approved but not accepted (optional reporting)

Ethnicity

1. Hispanic or Latino
2. Not Hispanic or Latino
3. Information not provided by applicant in mail, Internet, or telephone application
4. Not applicable

Race

1. American Indian or Alaska Native
2. Asian
3. Black or African American
4. Native Hawaiian or Other Pacific Islander
5. White
6. Information not provided by applicant in mail, Internet, or telephone application
7. Not applicable
8. No co-applicant

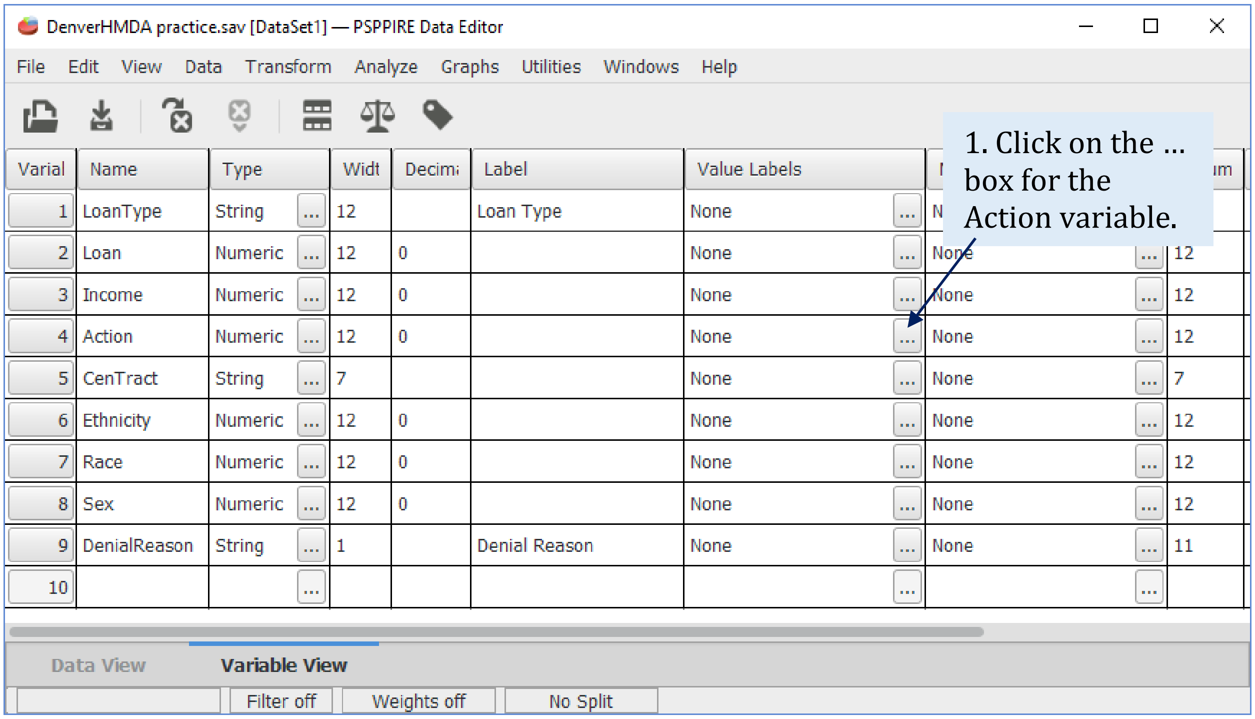
Sex

1. Male
2. Female
3. Information not provided by applicant in mail, Internet, or telephone application
4. Not applicable

DenialReason

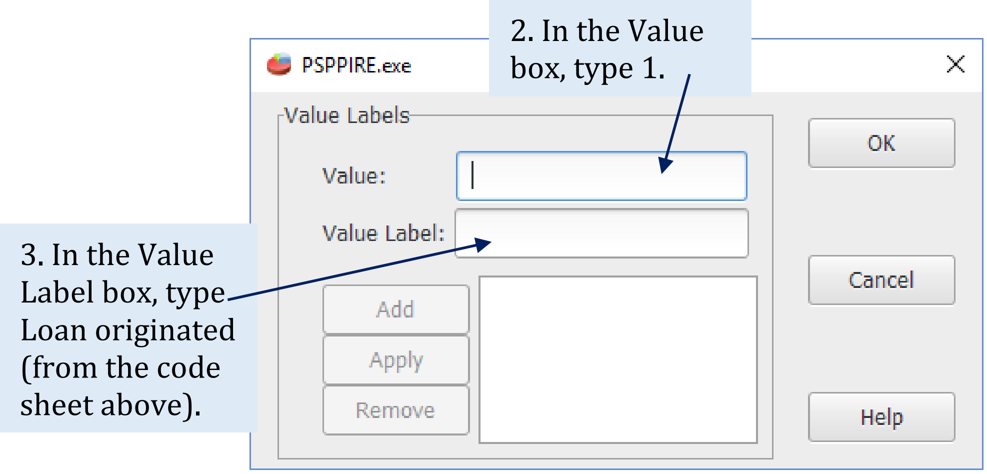
1. Debt-to-income ratio
2. Employment history
3. Credit history
4. Collateral
5. Insufficient cash (down payment, closing costs)
6. Unverifiable information
7. Credit application incomplete
8. Mortgage insurance denied
9. Other

Inputting Value Labels


DenverHMDA practice.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help
Data View
Variable View
Filter off
Weights off
No Split

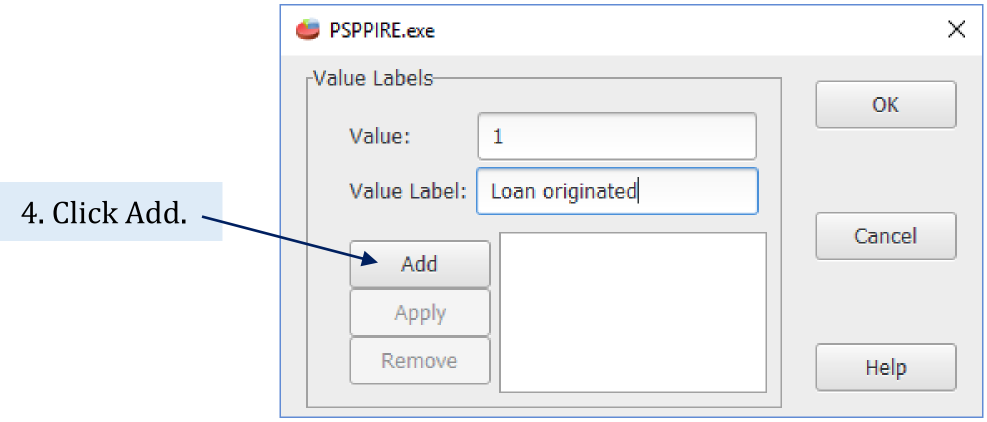
Varial	Name	Type	Widt	Decimi	Label	Value Labels	I	um
1	LoanType	String	12		Loan Type	None	N	
2	Loan	Numeric	12	0		None	None	12
3	Income	Numeric	12	0		None	None	12
4	Action	Numeric	12	0		None	None	12
5	CenTract	String	7			None	None	7
6	Ethnicity	Numeric	12	0		None	None	12
7	Race	Numeric	12	0		None	None	12
8	Sex	Numeric	12	0		None	None	12
9	DenialReason	String	1		Denial Reason	None	None	11
10								

1. Click on the ... box for the Action variable.



2. In the Value box, type 1.

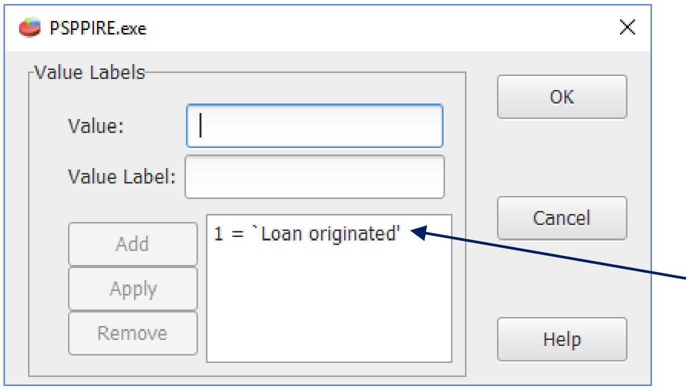
3. In the Value Label box, type Loan originated (from the code sheet above).



Value: 1

Value Label: Loan originated

4. Click Add.



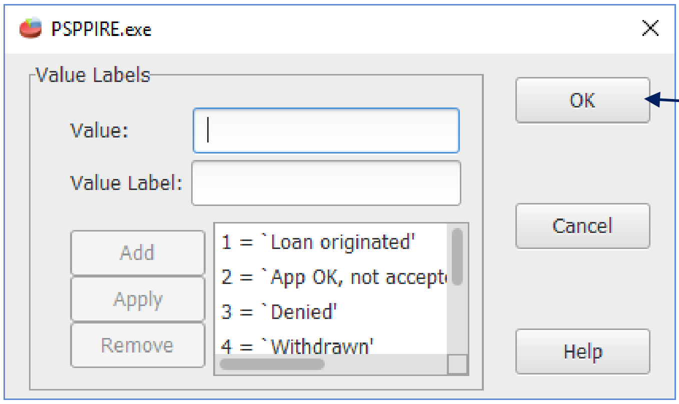
Value:

Value Label:

1 = `Loan originated'

Note the label is added to the box.

The process is repeated until all the labels are added. Shortened labels will be useful later. Then click OK and move to the next variable that needs labels.



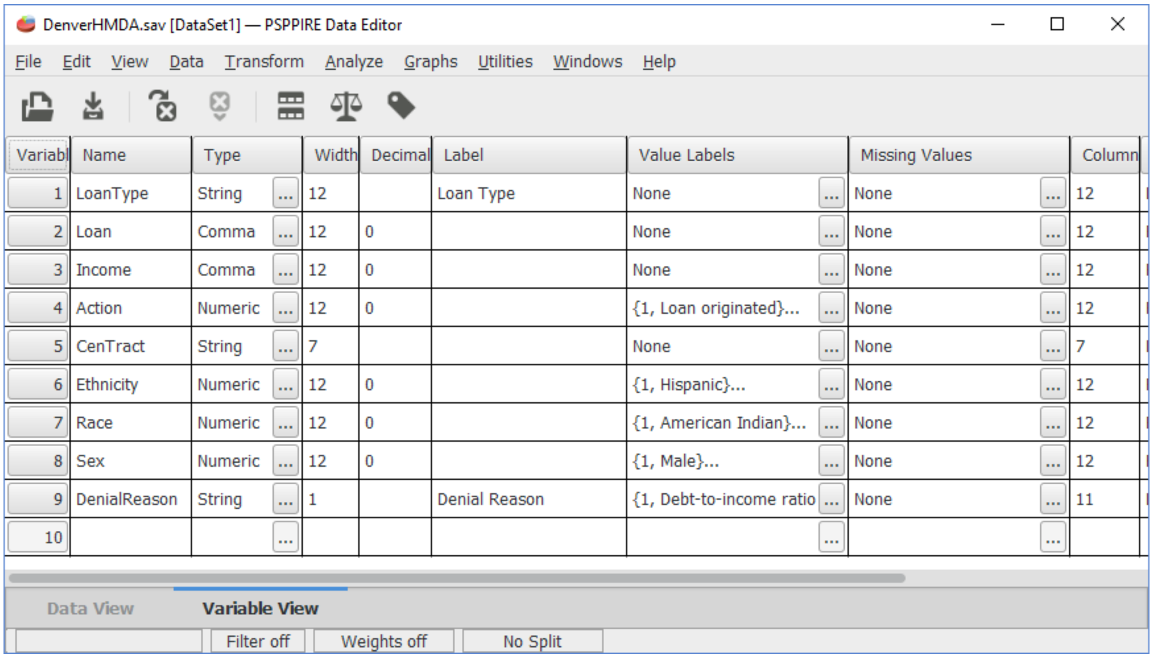
Value:

Value Label:

1 = `Loan originated'
2 = `App OK, not accept
3 = `Denied'
4 = `Withdrawn'

When all labels are added, click OK.

When completed, the pertinent fields have labels.


DenverHMDA.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help
Data View
Variable View
Filter off
Weights off
No Split

Variabl	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Column
1	LoanType	String ...	12		Loan Type	None ...	None ...	12
2	Loan	Comma ...	12	0		None ...	None ...	12
3	Income	Comma ...	12	0		None ...	None ...	12
4	Action	Numeric ...	12	0		{1, Loan originated}... ...	None ...	12
5	CenTract	String ...	7			None ...	None ...	7
6	Ethnicity	Numeric ...	12	0		{1, Hispanic}... ...	None ...	12
7	Race	Numeric ...	12	0		{1, American Indian}... ...	None ...	12
8	Sex	Numeric ...	12	0		{1, Male}... ...	None ...	12
9	DenialReason	String ...	1		Denial Reason	{1, Debt-to-income ratio ...	None ...	11
10		...				...	...	

Transforming a Variable

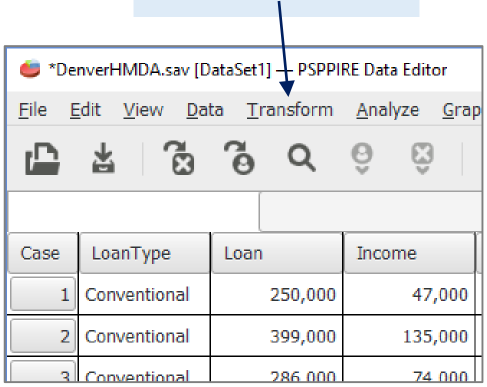
That leaves just one problem: the LoanType field. (Confession: The LoanType field comes with a number. I changed it to words to show how to convert string fields into numbers.)

HMDA offers four types of loans:

1.	Conventional	Any loan other that the three below
2.	FHA	Federal Housing Administration
3.	VA	Veterans Administration
4.	FSA/RHS	Farm Service Agency or Rural Housing Service

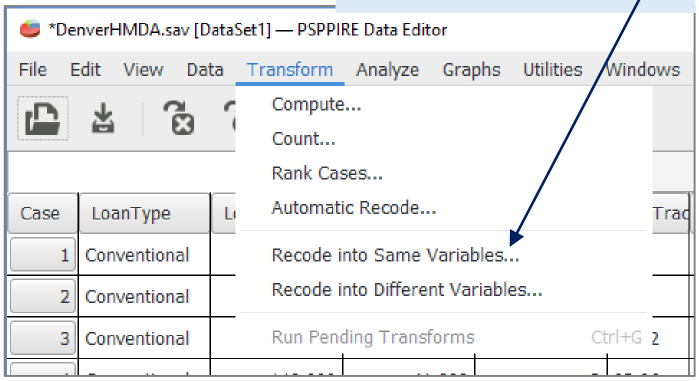
We will now convert those words to numbers.

1. Click on the Transform menu.

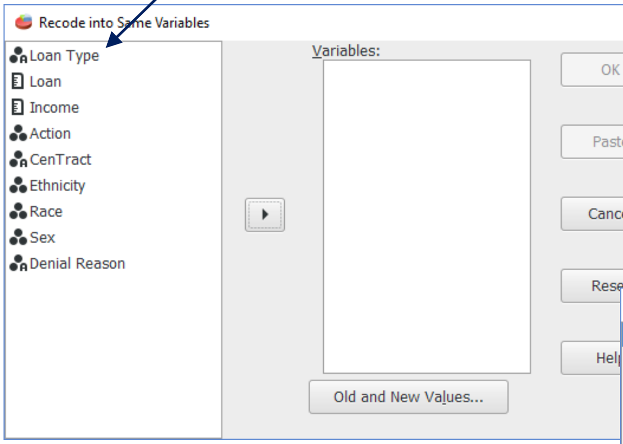


Case	LoanType	Loan	Income
1	Conventional	250,000	47,000
2	Conventional	399,000	135,000
3	Conventional	286,000	74,000

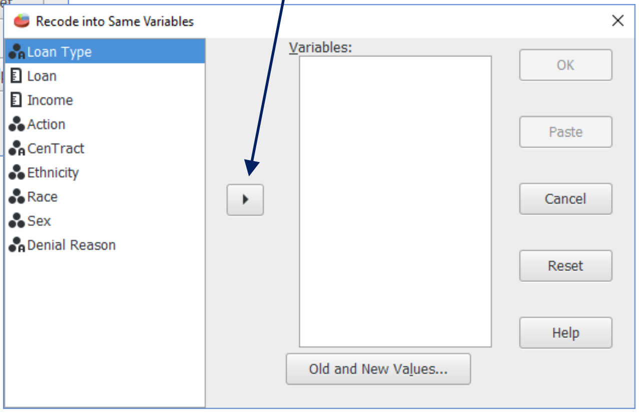
2. Choose Recode into Same Variables ...



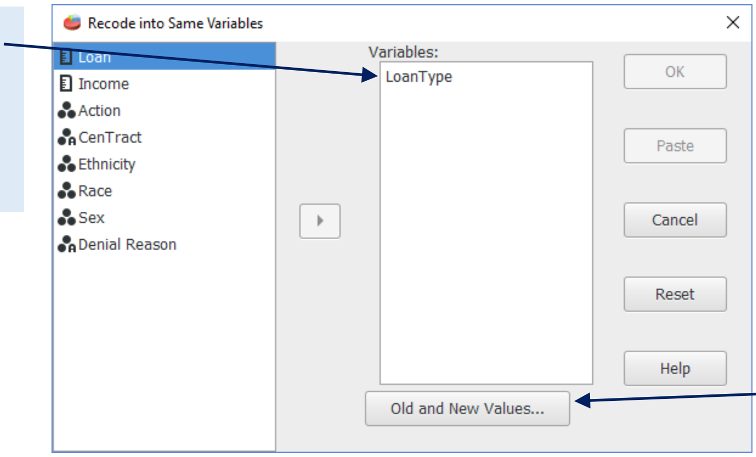
3. Click on LoanType.



4. Click on the forward button to move LoanType into the Variables box.



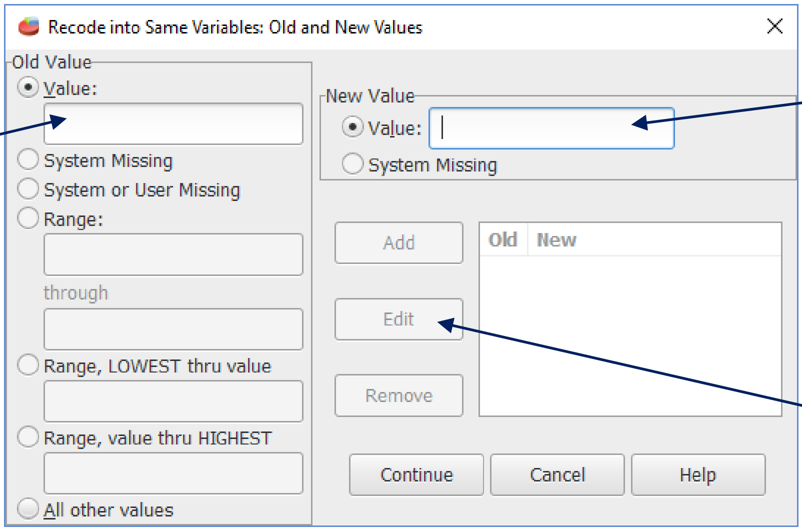
5. Click on LoanType in the Variables box.



Variables: LoanType

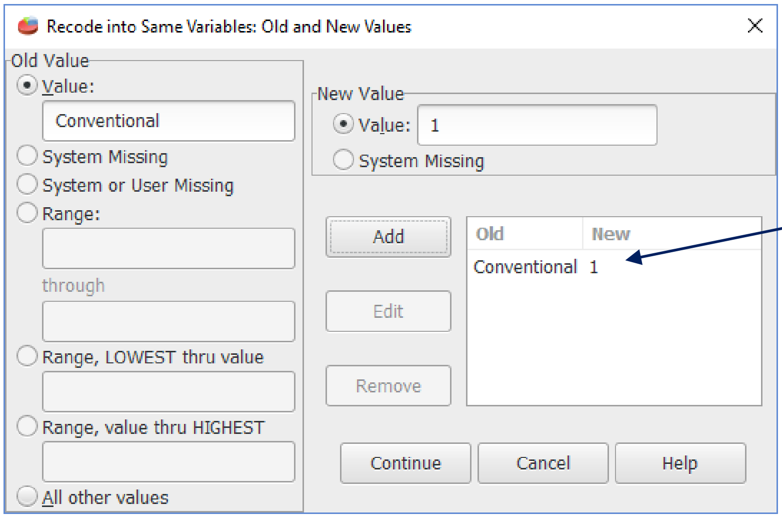
6. Click on Old and New Values ...

7. Under the Old Value section, in the Value box, type Conventional (no spaces).



8. Under the New Value section, in the Value box, type 1.

9. Click Add.



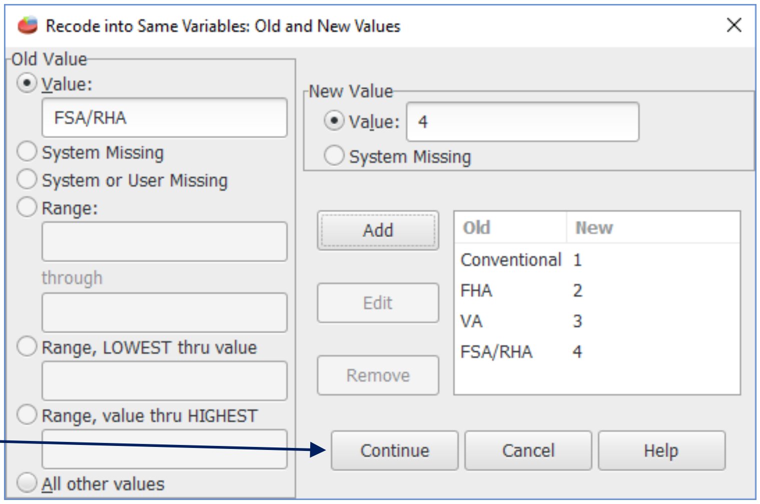
Old Value
● Value: Conventional

New Value
● Value: 1

Old	New
Conventional	1

Old and New values appear.

10. Repeat steps 7-8-9 with FHA, VA and FSA/RHA.



Recode into Same Variables: Old and New Values

Old Value

● Value: FSA/RHA

○ System Missing

○ System or User Missing

○ Range:

through

○ Range, LOWEST thru value

○ Range, value thru HIGHEST

○ All other values

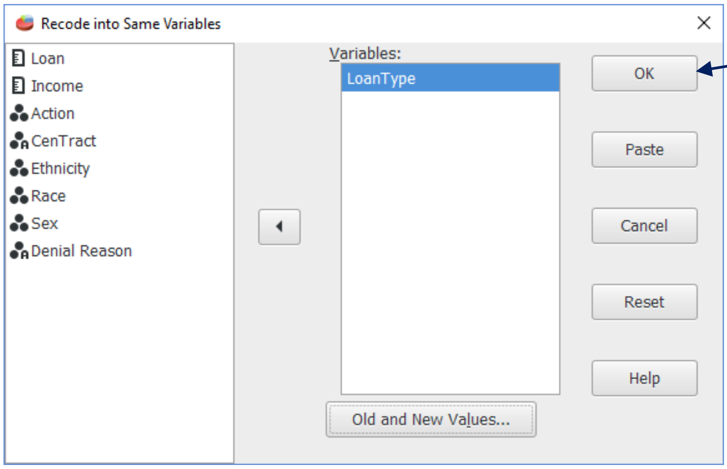
New Value

● Value: 4

○ System Missing

Old	New
Conventional	1
FHA	2
VA	3
FSA/RHA	4

11. Click Continue.



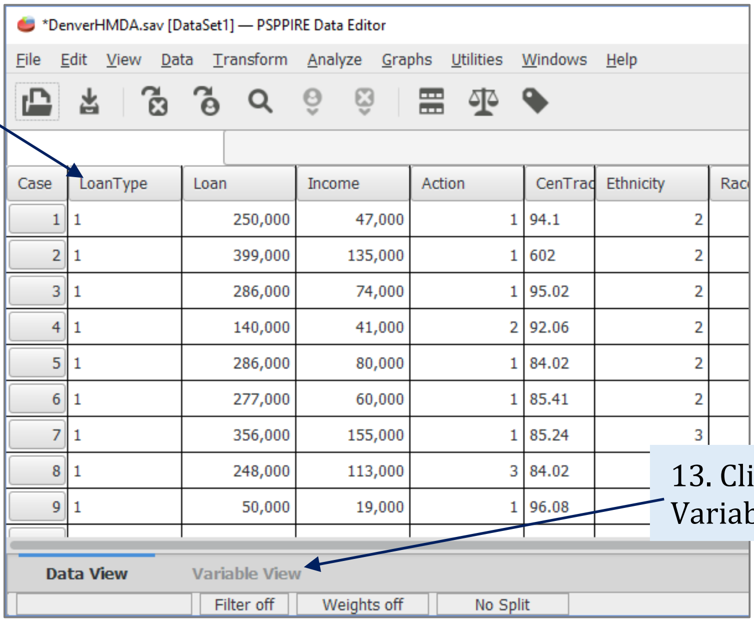
Recode into Same Variables

Loan
Income
Action
CenTract
Ethnicity
Race
Sex
Denial Reason

Variables: LoanType

12. Click OK.

LoanType variable now has numbers instead of words.



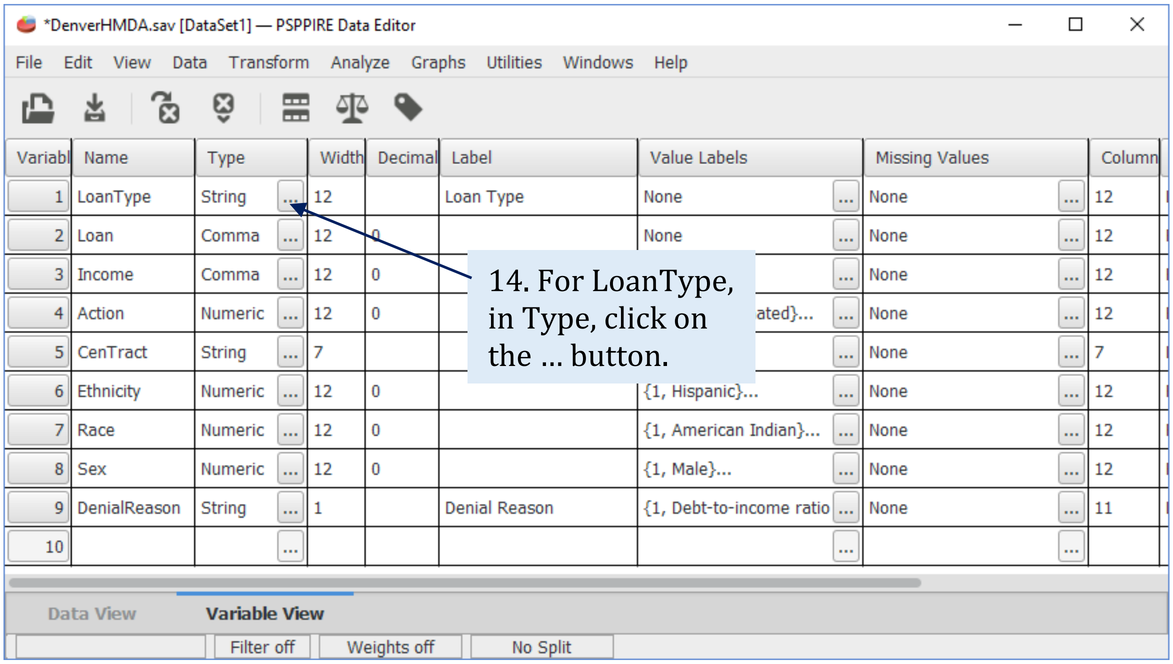
*DenverHMDA.sav [DataSet1] — PSPPIRE Data Editor

File Edit View Data Transform Analyze Graphs Utilities Windows Help

Case	LoanType	Loan	Income	Action	CenTrac	Ethnicity	Rac
1	1	250,000	47,000	1	94.1	2	
2	1	399,000	135,000	1	602	2	
3	1	286,000	74,000	1	95.02	2	
4	1	140,000	41,000	2	92.06	2	
5	1	286,000	80,000	1	84.02	2	
6	1	277,000	60,000	1	85.41	2	
7	1	356,000	155,000	1	85.24	3	
8	1	248,000	113,000	3	84.02		
9	1	50,000	19,000	1	96.08		

Data View
Variable View
Filter off
Weights off
No Split

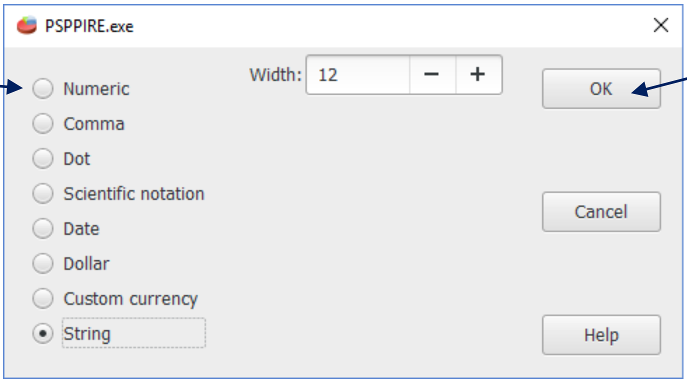
13. Click on Variable View.


*DenverHMDA.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help
Data View
Variable View
Filter off
Weights off
No Split

Variabl	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Column
1	LoanType	String	12		Loan Type	None	None	12
2	Loan	Comma	12	0		None	None	12
3	Income	Comma	12	0			None	12
4	Action	Numeric	12	0		ated}...	None	12
5	CenTract	String	7				None	7
6	Ethnicity	Numeric	12	0		{1, Hispanic}...	None	12
7	Race	Numeric	12	0		{1, American Indian}...	None	12
8	Sex	Numeric	12	0		{1, Male}...	None	12
9	DenialReason	String	1		Denial Reason	{1, Debt-to-income ratio	None	11
10								

14. For LoanType, in Type, click on the ... button.

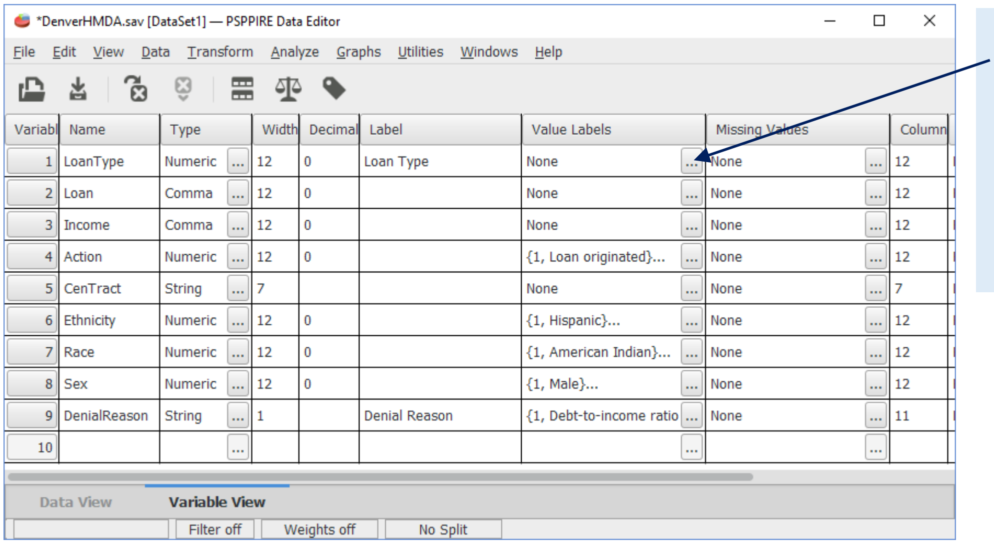
15. Click on Numeric button.



Width: 12

- ○ Numeric
- ○ Comma
- ○ Dot
- ○ Scientific notation
- ○ Date
- ○ Dollar
- ○ Custom currency
- ● String

16. Click OK.


*DenverHMDA.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows Help
Data View
Variable View
Filter off
Weights off
No Split

Variabl	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Column
1	LoanType	Numeric	12	0	Loan Type	None	None	12
2	Loan	Comma	12	0		None	None	12
3	Income	Comma	12	0		None	None	12
4	Action	Numeric	12	0		{1, Loan originated}...	None	12
5	CenTract	String	7			None	None	7
6	Ethnicity	Numeric	12	0		{1, Hispanic}...	None	12
7	Race	Numeric	12	0		{1, American Indian}...	None	12
8	Sex	Numeric	12	0		{1, Male}...	None	12
9	DenialReason	String	1		Denial Reason	{1, Debt-to-income ratio	None	11
10								

17. Click on the Value Labels ... button and assign words (Convention, FHA, etc.) to the numbers.

PART 5: SCALES OF MEASUREMENT

Before we can analyze the data, we have to understand what kind it is. This involves *scales of measurement*, which come in four types.

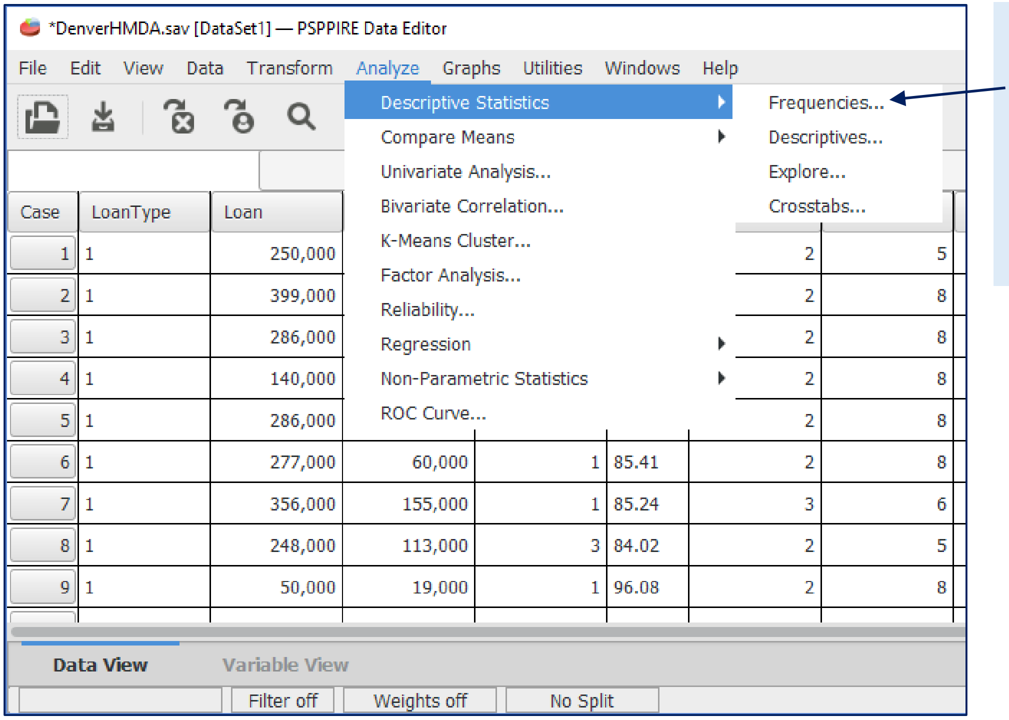
	Scale	Description	Examples
⛪	Nominal	Categorizes data without ranking; group 1 is not "better" than group 2	Religion, political party, sex, color, national origin
🏃	Ordinal	Categorizes and ranks; placing first is better than placing second but intervals between places are unequal	Finish order, grades, rank in class, Likert scale
🌡	Interval	Categorizes and ranks with equal distance but without a true zero; 0 degrees Celsius is not zero energy	Fahrenheit and Celsius temperature scales, time of day
⚖	Ratio	Categorizes and ranks with equal distance, and has a true zero	Kelvin temperature scale, age, weight, distance, money

For now, we can group these four into two: *categorical* and *continuous*. Then we can parse the HMDA variables into these two groups and consider potential stats.

	Scale	Group	DenverHMDA variables	Potential stats
⛪	Nominal	Categorical	LoanType Action CenTract Ethnicity Race Sex DenialReason	Sum, percent, mode, chi-square
🏃	Ordinal			
🌡	Interval	Continuous	Loan Income	Sum, percent, mean, median, standard deviation, t-test, ANOVA, regression
⚖	Ratio			

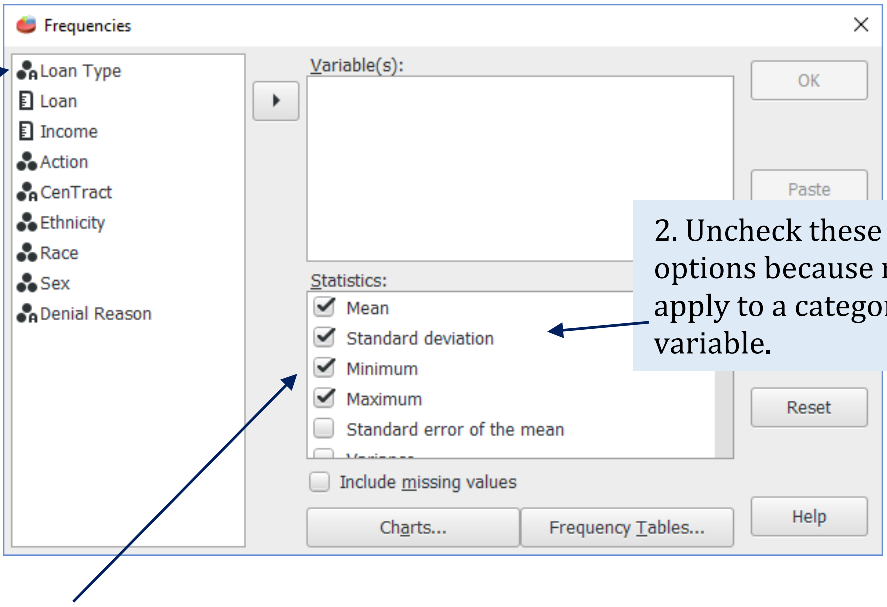
PART 6: FREQUENCIES

Now we can begin to interview the data.



Case	LoanType	Loan					
1	1	250,000				2	5
2	1	399,000				2	8
3	1	286,000				2	8
4	1	140,000				2	8
5	1	286,000				2	8
6	1	277,000	60,000	1	85.41	2	8
7	1	356,000	155,000	1	85.24	3	6
8	1	248,000	113,000	3	84.02	2	5
9	1	50,000	19,000	1	96.08	2	8

1. Click on the Analyze menu, click on Descriptive Statistics, and select Frequencies.



Statistics:

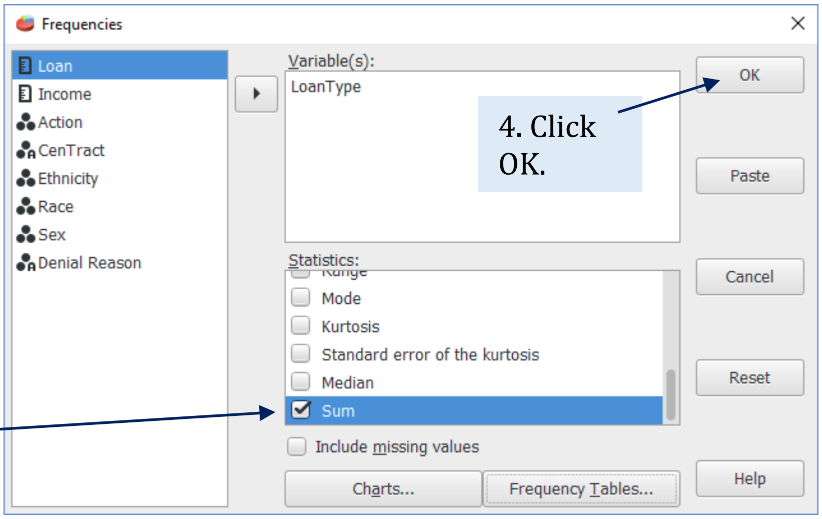
- ☑ Mean
- ☑ Standard deviation
- ☑ Minimum
- ☑ Maximum
- □ Standard error of the mean
- □ Variance

□ Include missing values

1. Click on LoanType to move it into the Variable(s) view.

2. Uncheck these four options because none apply to a categorical variable.

PSSS checks these four statistics by default. But LoanType is a categorical variable for which a mean (and thus a standard deviation) would hold no meaning. So let’s customize this list to match the variable type.



Variable(s):
LoanType

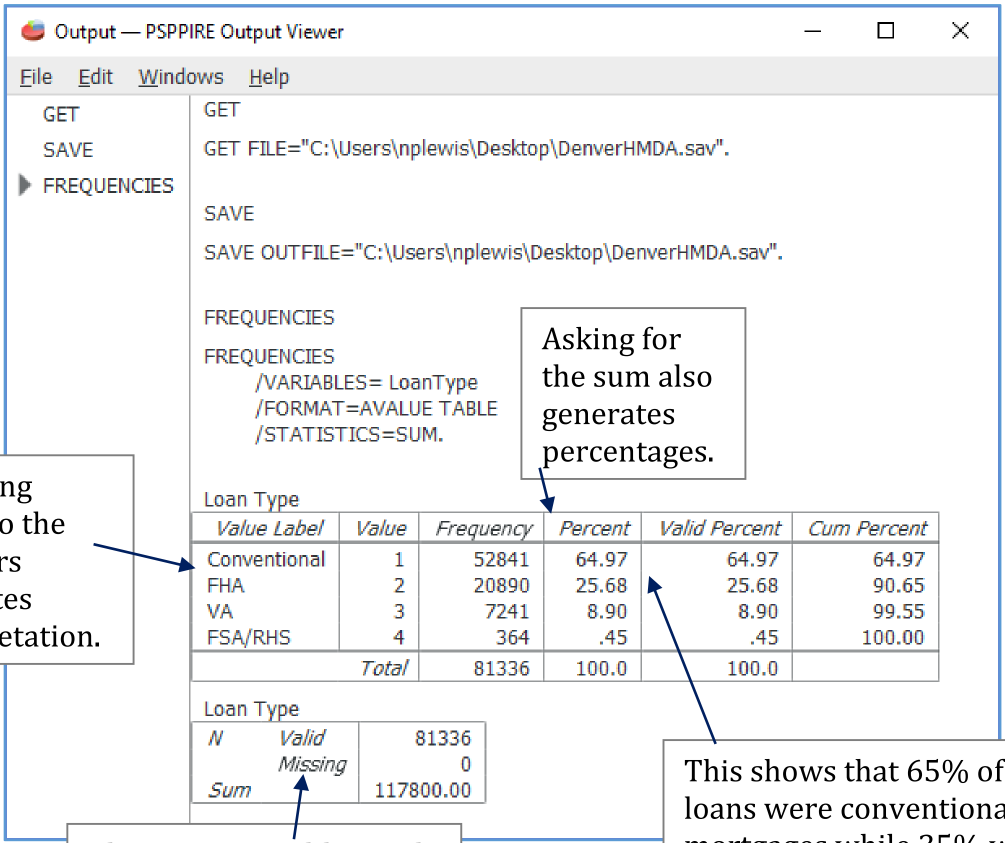
Statistics:
- □ Mode
- □ Kurtosis
- □ Standard error of the kurtosis
- □ Median
- ☑ Sum

□ Include missing values

3. Scroll through the list of suitable frequencies options, which for a categorical variable is only sum.

4. Click OK.

Results from the statistical procedure appear in the Output Viewer window.



```
GET

GET FILE="C:\Users\nplewis\Desktop\DenverHMDA.sav".


SAVE

SAVE OUTFILE="C:\Users\nplewis\Desktop\DenverHMDA.sav".


FREQUENCIES

FREQUENCIES
        /VARIABLES= LoanType
        /FORMAT=AVALUE TABLE
        /STATISTICS=SUM.
```

Loan Type

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum Percent*
Conventional	1	52841	64.97	64.97	64.97
FHA	2	20890	25.68	25.68	90.65
VA	3	7241	8.90	8.90	99.55
FSA/RHS	4	364	.45	.45	100.00
	Total	81336	100.0	100.0	

Loan Type

N	*Valid*	81336
	Missing	0
Sum		117800.00

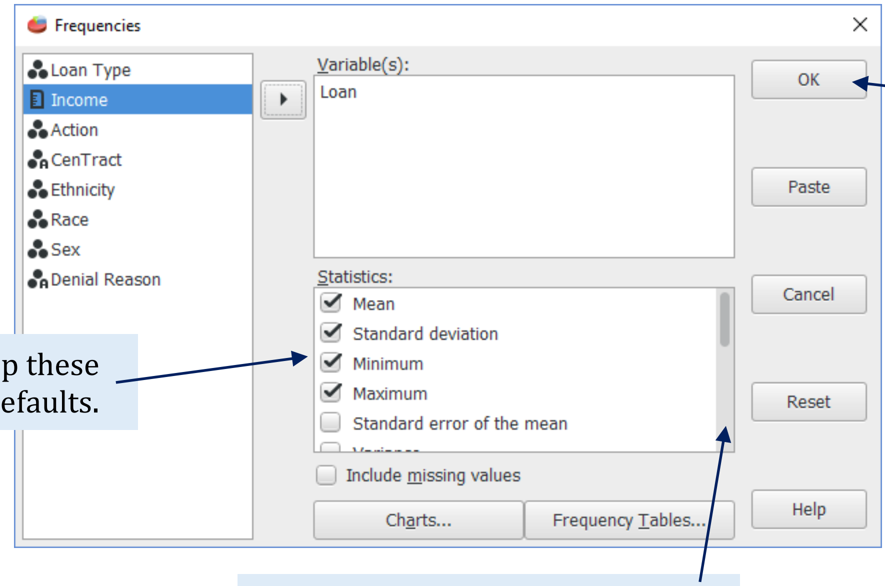
Asking for the sum also generates percentages.

Assigning labels to the numbers facilitates interpretation.

This shows that 65% of these loans were conventional mortgages while 35% were from government programs.

The output would reveal if any cases were missing data for LoanType.

Now let’s look at frequency statistics available for a continuous variable: loan.



Variable(s): Loan

Statistics:
☑ Mean
☑ Standard deviation
☑ Minimum
☑ Maximum
□ Standard error of the mean

□ Include missing values

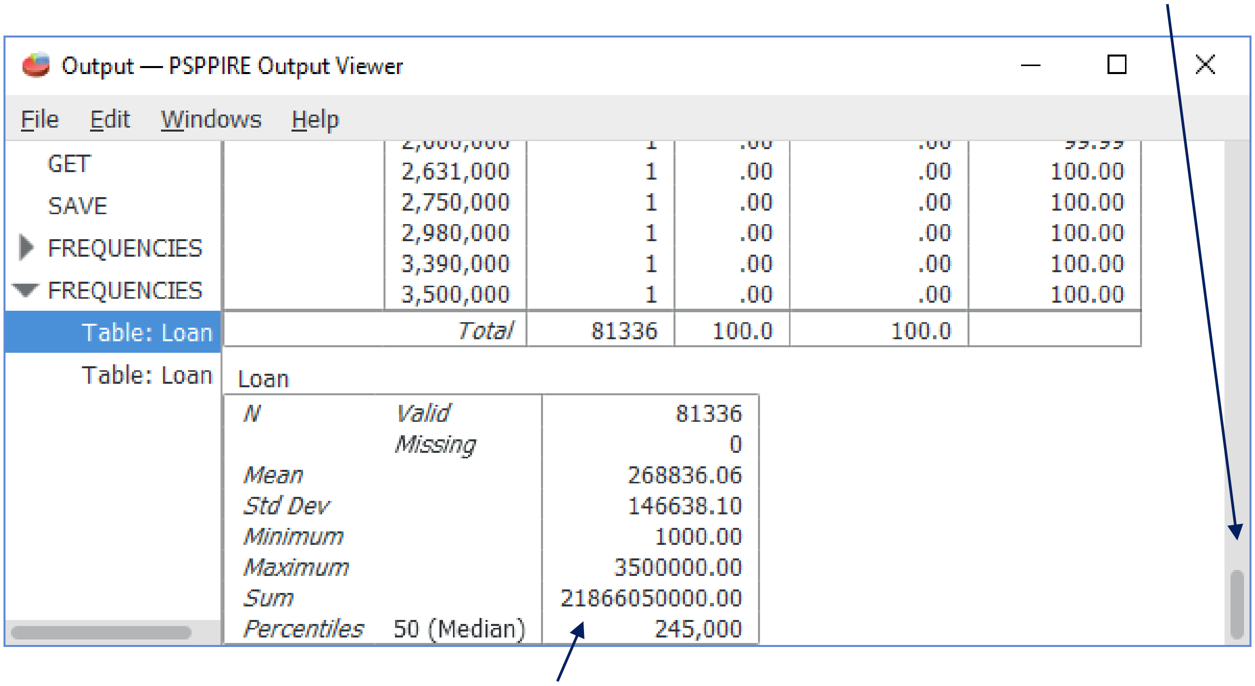
1. Keep these four defaults.

2. Scroll through the list to add Median (usually a better average for money than mean) and Sum.

3. Click OK.

(continued on next page)

Scroll through the Output Viewer to get to the good stuff at the bottom.



2,631,000	1	.00	.00	100.00
2,750,000	1	.00	.00	100.00
2,980,000	1	.00	.00	100.00
3,390,000	1	.00	.00	100.00
3,500,000	1	.00	.00	100.00
Total	81336	100.0	100.0	

Loan

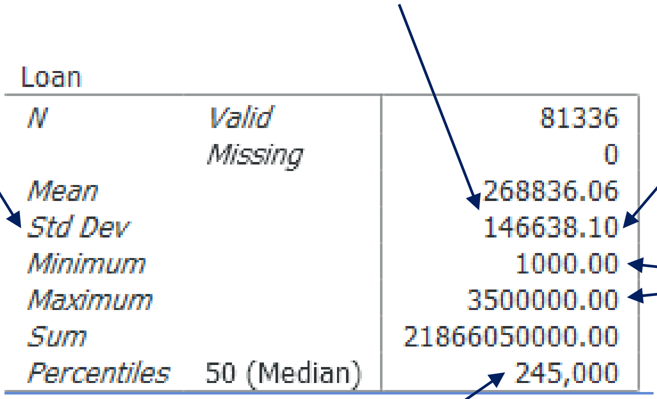
N	*Valid*	81336
	Missing	0
Mean		268836.06
Std Dev		146638.10
Minimum		1000.00
Maximum		3500000.00
Sum		21866050000.00
Percentiles	50 (Median)	245,000

Let's explore these numbers in more detail:

Std Dev is standard deviation. It is a measure of dispersion. When added and subtracted to the mean, it marks the boundaries that will include 68% of all loans in this set.

Std Dev and Mean combine to reveal that 68% of loans are between $415,474 and $122,198.

The standard deviation is 55% of the mean, which is relatively large and an indication that the loan data are widely distributed. But that's typical with home loans, so not newsworthy.



Loan

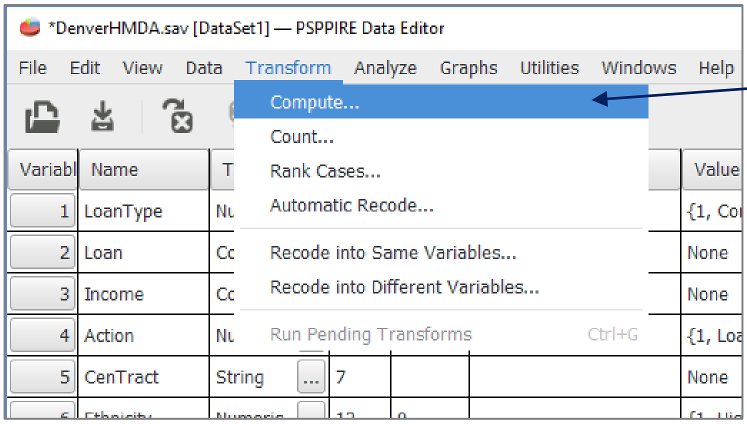
N	*Valid*	81336
	Missing	0
Mean		268836.06
Std Dev		146638.10
Minimum		1000.00
Maximum		3500000.00
Sum		21866050000.00
Percentiles	50 (Median)	245,000

Loans ranged from $1,000 to $3.5 million.

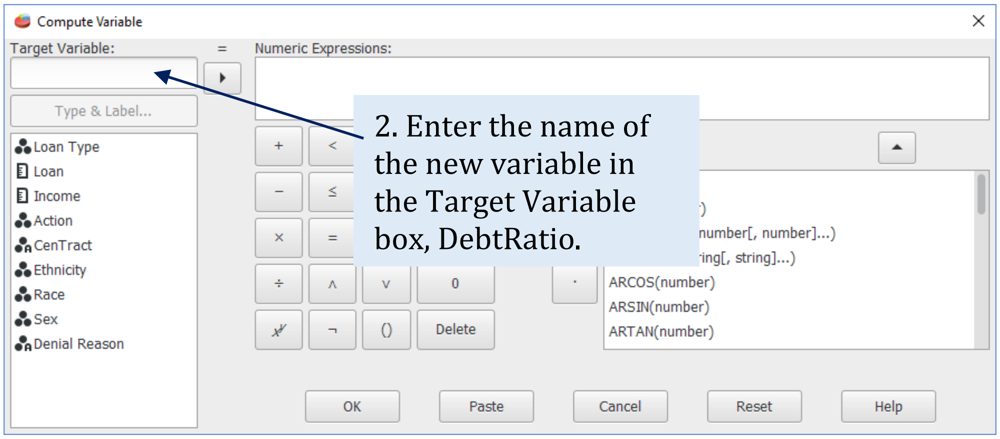
Median (the midpoint) is a better average than mean when numbers vary so widely, as is typical in home loans.

PART 7: CALCULATING A NEW VARIABLE

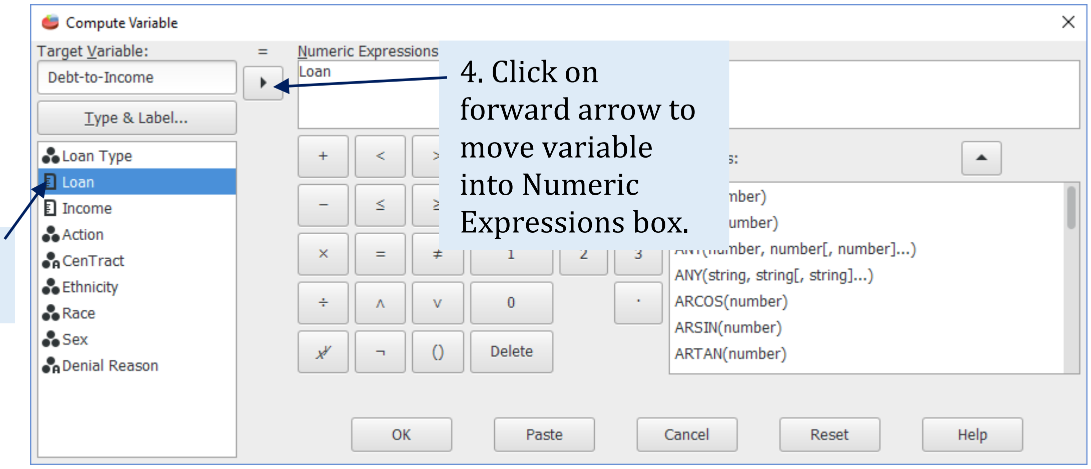
The frequencies for DenialReason showed that debt-to-income ratio was the leading reason for denial of a mortgage. This data set does not have debt-to-income ratio as a variable. But it has both the debt (loan) and the income data. So we can create a variable that calculates the debt-to-income ratio.



1. Under the Transform menu, select Compute ...



2. Enter the name of the new variable in the Target Variable box, DebtRatio.

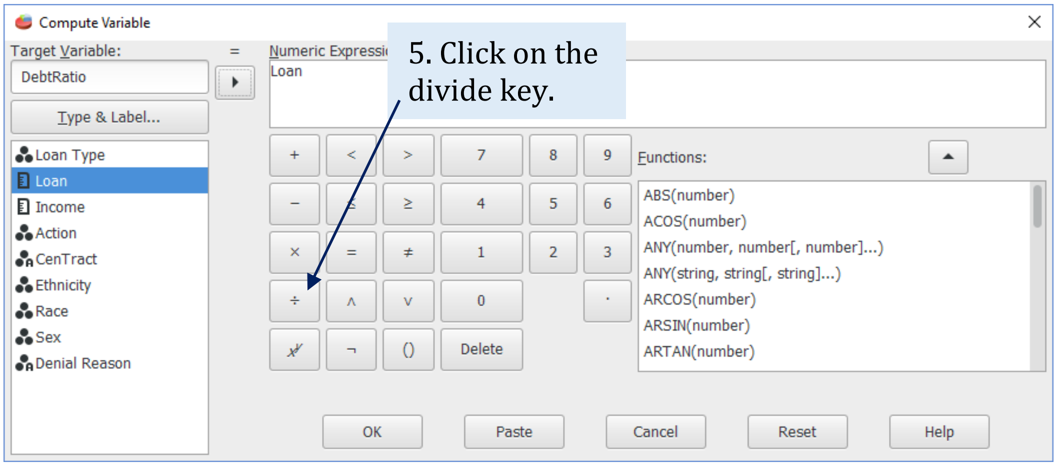


Target Variable: Debt-to-Income

Numeric Expressions: Loan

4. Click on forward arrow to move variable into Numeric Expressions box.

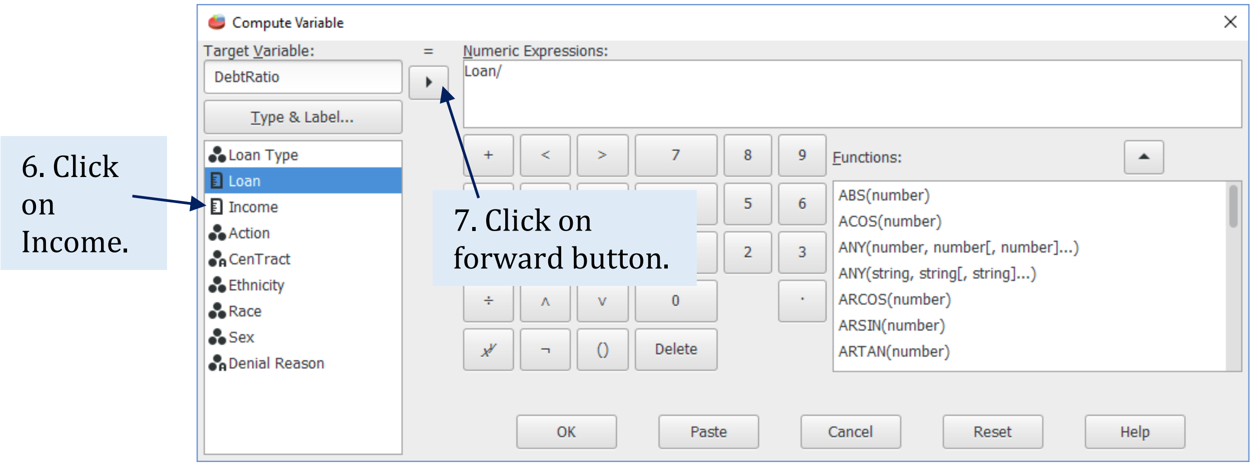
3. Click on Loan.



Target Variable: DebtRatio

Numeric Expression: Loan

5. Click on the divide key.

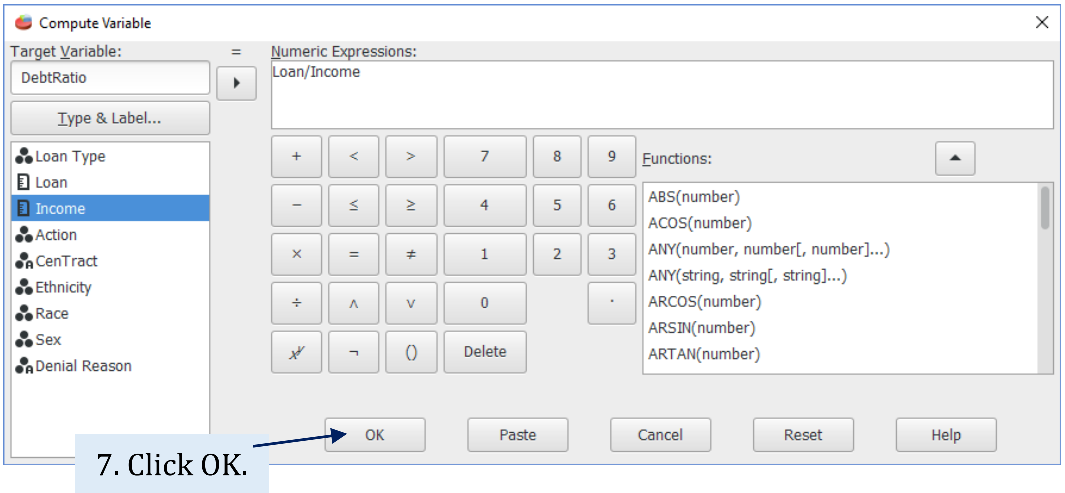


Target Variable: DebtRatio

Numeric Expressions: Loan/

6. Click on Income.

7. Click on forward button.

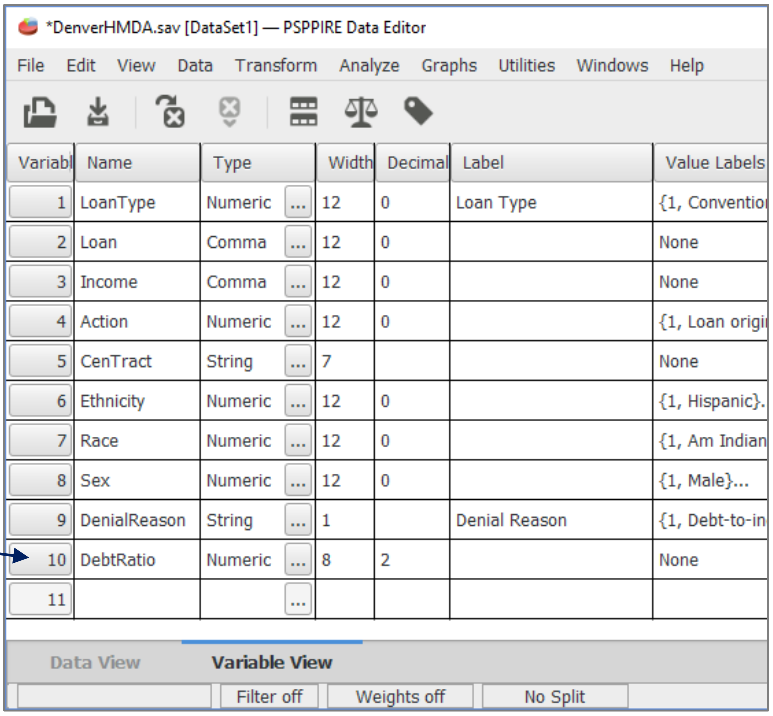


Target Variable: DebtRatio

Numeric Expressions: Loan/Income

7. Click OK.

Variable View shows the new variable DebtRatio.

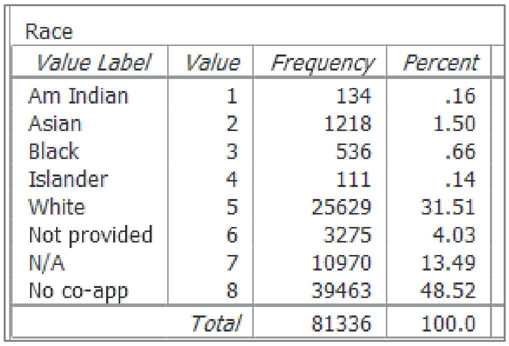


Variabl	Name	Type		Width	Decimal	Label	Value Labels
1	LoanType	Numeric	...	12	0	Loan Type	{1, Conventio
2	Loan	Comma	...	12	0		None
3	Income	Comma	...	12	0		None
4	Action	Numeric	...	12	0		{1, Loan origi
5	CenTract	String	...	7			None
6	Ethnicity	Numeric	...	12	0		{1, Hispanic}.
7	Race	Numeric	...	12	0		{1, Am Indian
8	Sex	Numeric	...	12	0		{1, Male}...
9	DenialReason	String	...	1		Denial Reason	{1, Debt-to-in
10	DebtRatio	Numeric	...	8	2		None
11			...				

PART 8: COLLAPSING DATA TO FACILITATE ANALYSIS

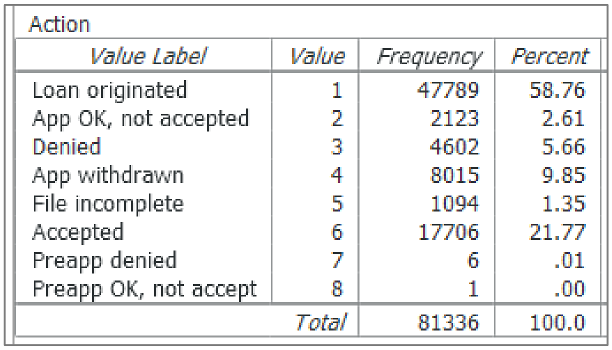
With home loan data like this, we want to know if the denial or acceptance rates differ by another factor, such as race.

Here are the descriptive statistics for the Race and Action variables:



Race

Value Label	Value	Frequency	Percent
Am Indian	1	134	.16
Asian	2	1218	1.50
Black	3	536	.66
Islander	4	111	.14
White	5	25629	31.51
Not provided	6	3275	4.03
N/A	7	10970	13.49
No co-app	8	39463	48.52
	Total	81336	100.0



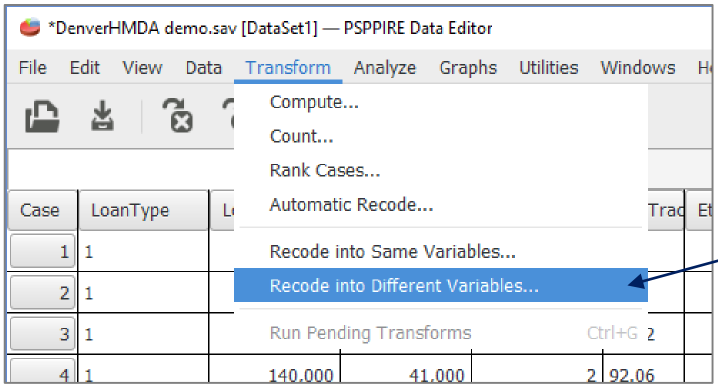
Action

Value Label	Value	Frequency	Percent
Loan originated	1	47789	58.76
App OK, not accepted	2	2123	2.61
Denied	3	4602	5.66
App withdrawn	4	8015	9.85
File incomplete	5	1094	1.35
Accepted	6	17706	21.77
Preapp denied	7	6	.01
Preapp OK, not accept	8	1	.00
	Total	81336	100.0

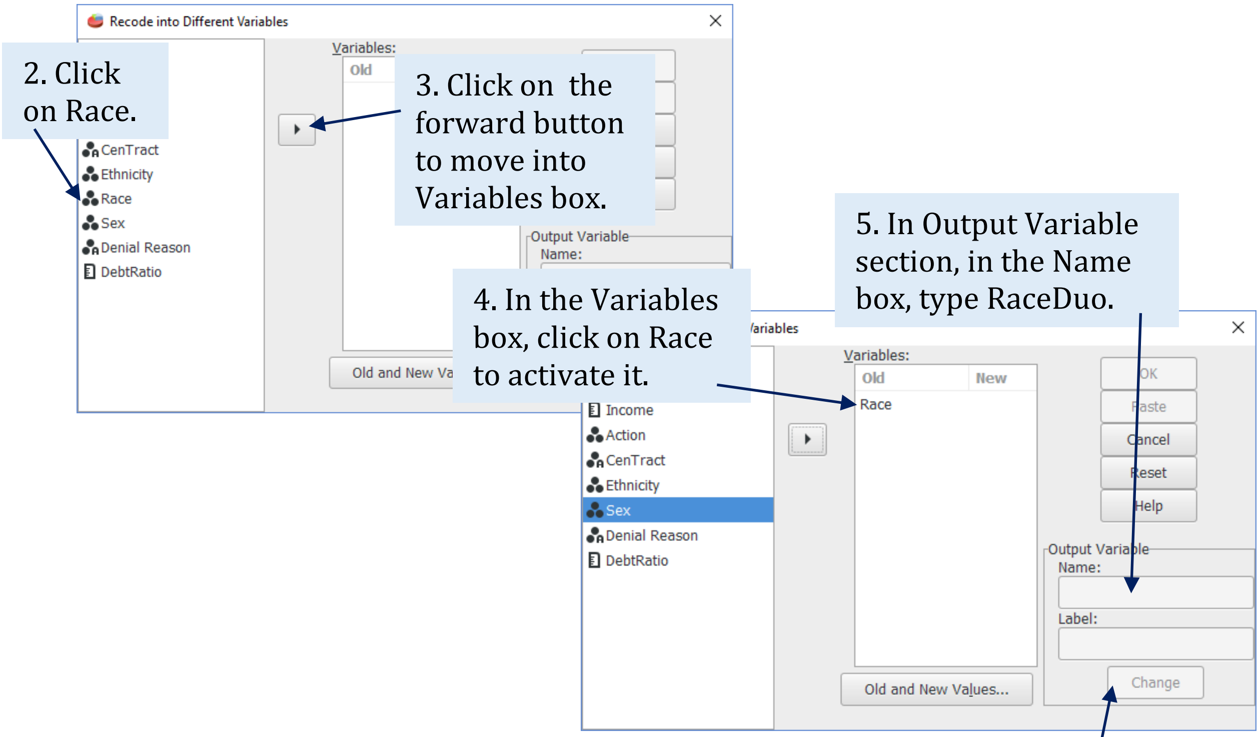
That’s messy. Each variable has too many options. And the results for race are greatly skewed by the “no co-applicant” option. So we need first to re-compute these into new variables before we can do meaningful analysis.

We will collapse Action into two groups: accept or reject. Because the numbers for race are so small, we also will collapse race into two groups: minority and white.

Collapsing data



1. On the Tranform menu, choose Recode into Different Variables.



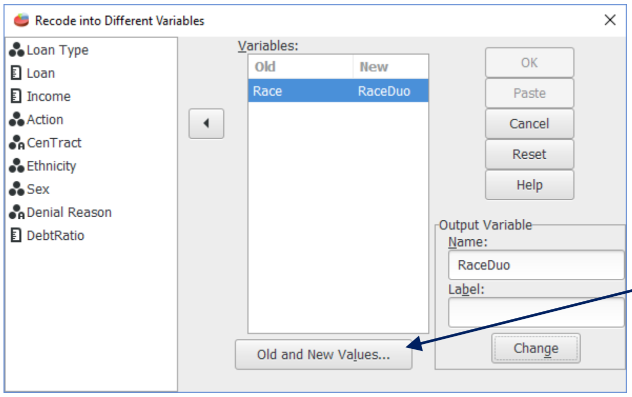
2. Click on Race.

3. Click on the forward button to move into Variables box.

4. In the Variables box, click on Race to activate it.

5. In Output Variable section, in the Name box, type RaceDuo.

6. Click on the Change button.

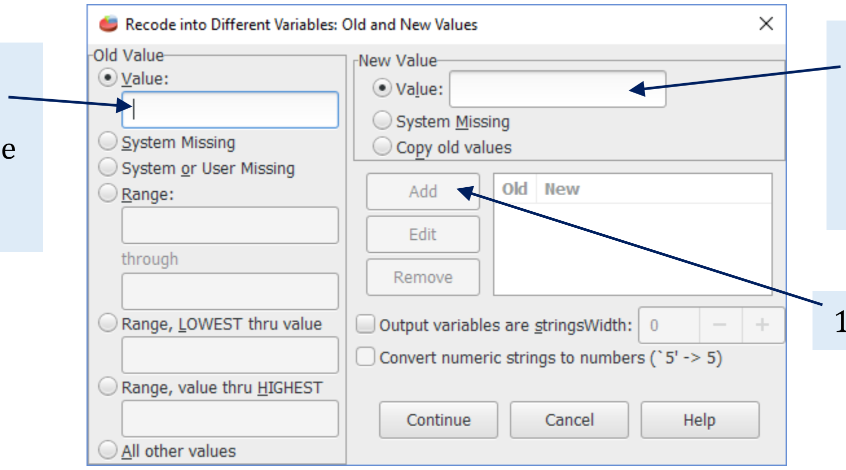


7. Click on the Old and New Values ... button.

These are the answer categories from the code book for race:

1. American Indian or Alaska Native
2. Asian
3. Black or African American
4. Native Hawaiian or Other Pacific Islander
5. White
6. Information not provided by applicant in mail, Internet, or telephone application
7. Not applicable
8. No co-applicant

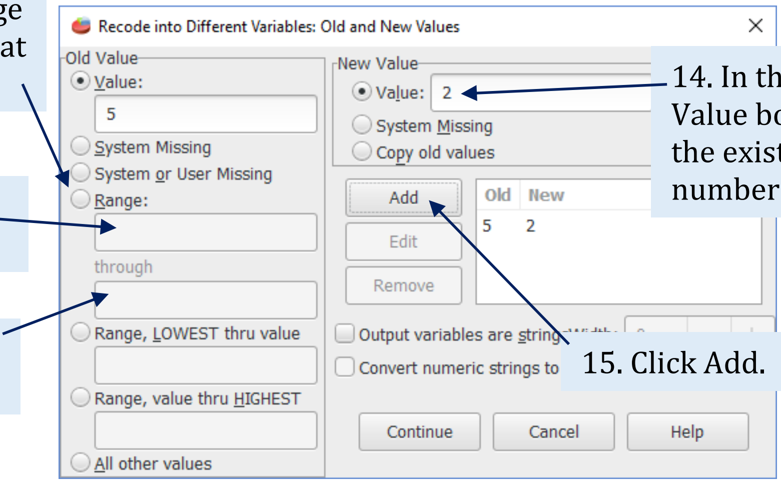
For RaceDuo, we will combine 1, 2, 3 and 4 into one Minority category (new value: 1) and renumber White from a 5 to 2. Ignore data from 6, 7 and 8.


Recode into Different Variables: Old and New Values
Old Value
● Value:
○ System Missing
○ System or User Missing
○ Range:
through
○ Range, LOWEST thru value
○ Range, value thru HIGHEST
○ All other values
New Value
● Value:
○ System Missing
○ Copy old values
Add
Edit
Remove
Old
New
□ Output variables are strings Width: 0
□ Convert numeric strings to numbers ('5' -> 5)
Continue
Cancel
Help

8. Under the Old Value section, in the Value box, type 5.

9. Under the New Value section, in the Value box, type 2.

10. Click Add.



Old Value: 5

New Value: 2

Old	New
5	2

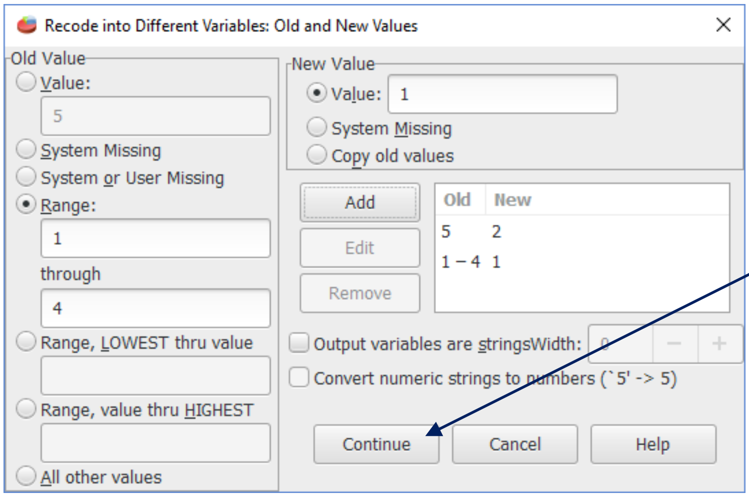
11. Click on the Range button to activate that section.

12. In the first Range box, type 1.

13. In the second Range box, type 4.

14. In the New Value box, replace the existing number with 1.

15. Click Add.



Old Value

○ Value: 5

○ System Missing

○ System or User Missing

● Range: 1

through 4

○ Range, LOWEST thru value

○ Range, value thru HIGHEST

○ All other values

New Value

● Value: 1

○ System Missing

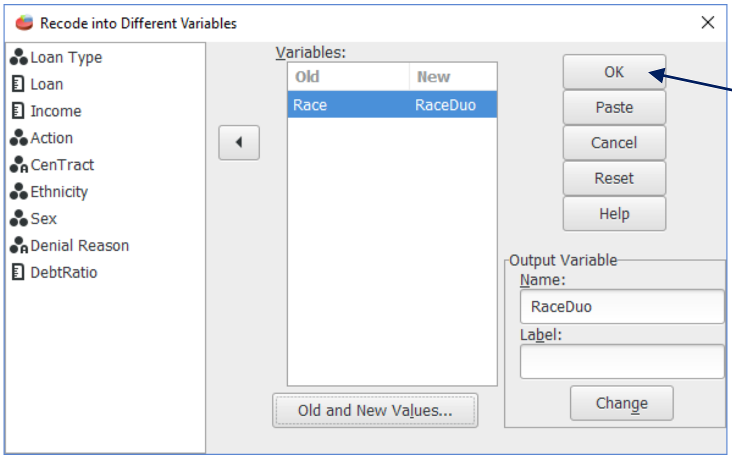
○ Copy old values

Old	New
5	2
1 – 4	1

□ Output variables are stringsWidth: 0

□ Convert numeric strings to numbers (`5' -> 5)

16. Click Continue.



Variables:

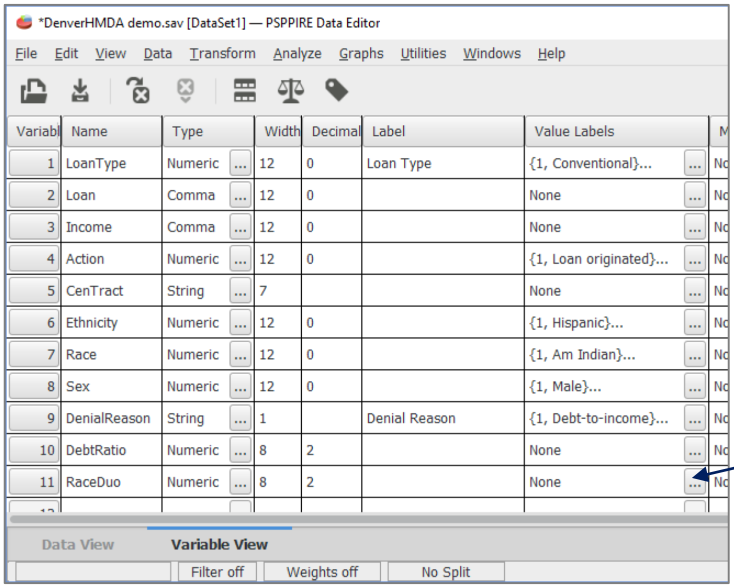
Old	New
Race	RaceDuo

Output Variable

Name: RaceDuo

Label:

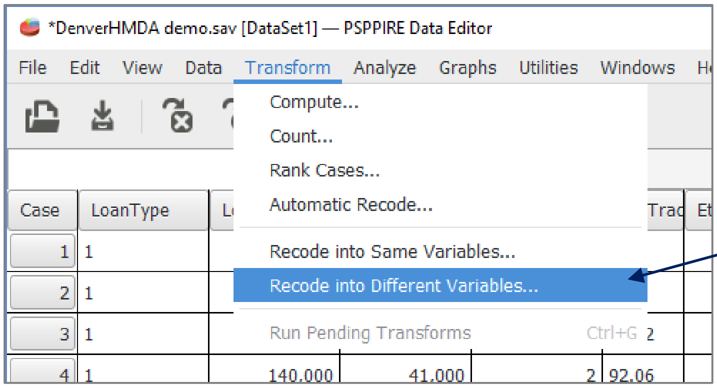
17. Click OK.



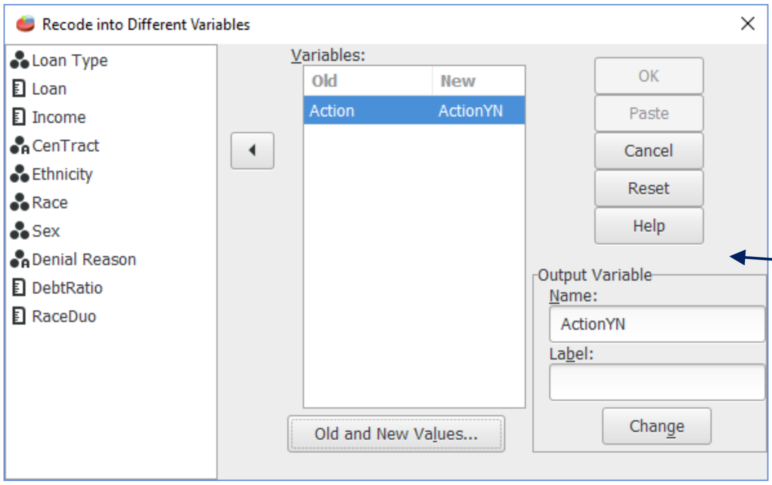
Variabl	Name	Type	Width	Decimal	Label	Value Labels	M
1	LoanType	Numeric ...	12	0	Loan Type	{1, Conventional}... ...	Nc
2	Loan	Comma ...	12	0		None ...	Nc
3	Income	Comma ...	12	0		None ...	Nc
4	Action	Numeric ...	12	0		{1, Loan originated}... ...	Nc
5	CenTract	String ...	7			None ...	Nc
6	Ethnicity	Numeric ...	12	0		{1, Hispanic}... ...	Nc
7	Race	Numeric ...	12	0		{1, Am Indian}... ...	Nc
8	Sex	Numeric ...	12	0		{1, Male}... ...	Nc
9	DenialReason	String ...	1		Denial Reason	{1, Debt-to-income}... ...	Nc
10	DebtRatio	Numeric ...	8	2		None ...	Nc
11	RaceDuo	Numeric ...	8	2		None ...	Nc

18. In the Data Editor window, for our new variable RaceDuo, click on the Value Labels ... button. Using the techniques learned earlier, assign a label of Minority to 1 and White to 2.

Repeat the process to create the AcceptYN variable.


*DenverHMDA demo.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Graphs Utilities Windows He
Compute...
Count...
Rank Cases...
Automatic Recode...
Recode into Same Variables...
Recode into Different Variables...
Run Pending Transforms Ctrl+G
Case LoanType
140,000 41,000 2 92.06

19. On the Tranform menu, again choose Recode into Different Variables.



Old	New
Action	ActionYN

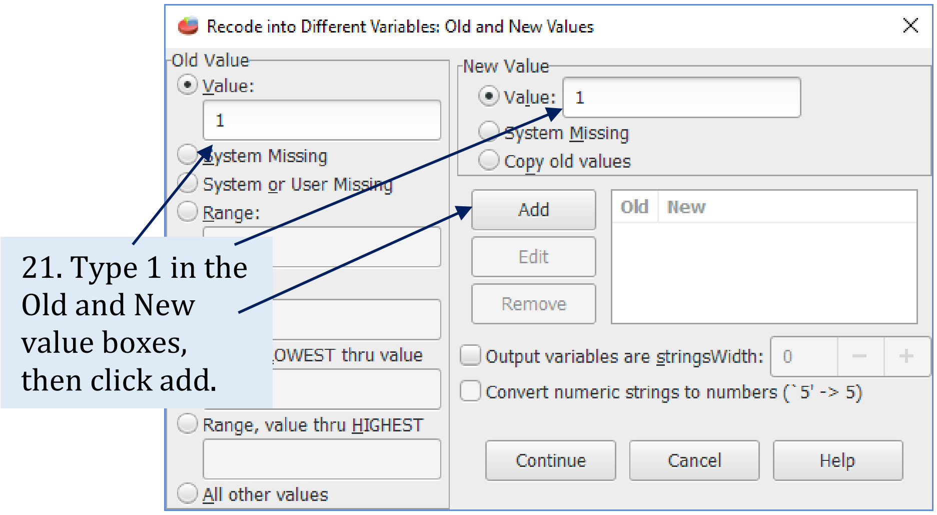
Output Variable
Name: ActionYN
Label:

20. Select the Action variable, move it into the Variables window and create the new Output Variable, ActionYN; click on Change and then on Old and New Values ...

The code book offered these values for the Action variable. The ones in bold (1, 2 and 6) indicate loan acceptance. No. 3 indicates the loan application was denied.

1. **Loan originated**
2. **Application approved but not accepted**
3. *Application denied by financial institution*
4. Application withdrawn by applicant
5. File closed for incompleteness
6. **Loan purchased by the institution**
7. Preapproval request denied by financial institution
8. Preapproval request approved but not accepted (optional reporting)

Thus, we will code 1, 2 and 6 as Accepted (new value: 1) and 3 as Denied (new value: 0) and ignore the rest.

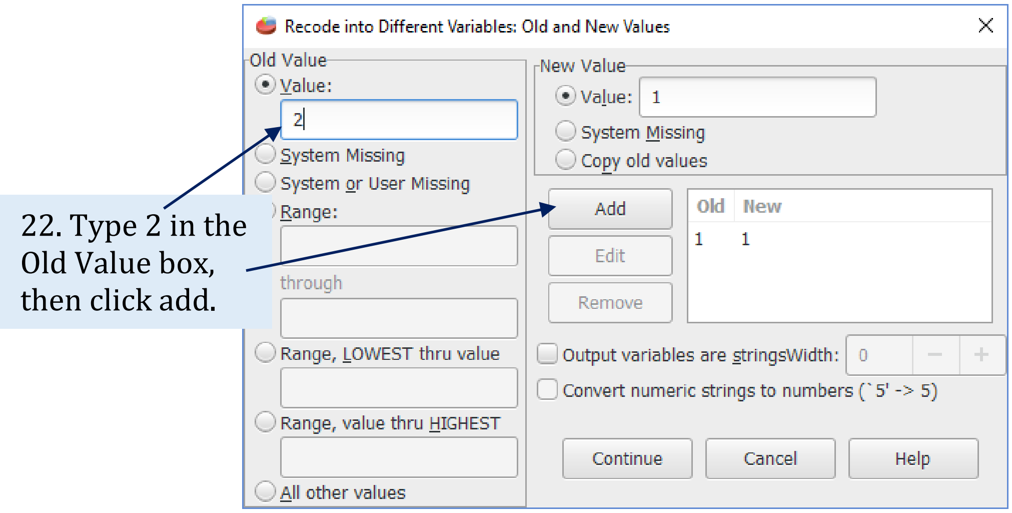


21. Type 1 in the Old and New value boxes, then click add.

Recode into Different Variables: Old and New Values

Old Value: ● Value: 1

New Value: ● Value: 1



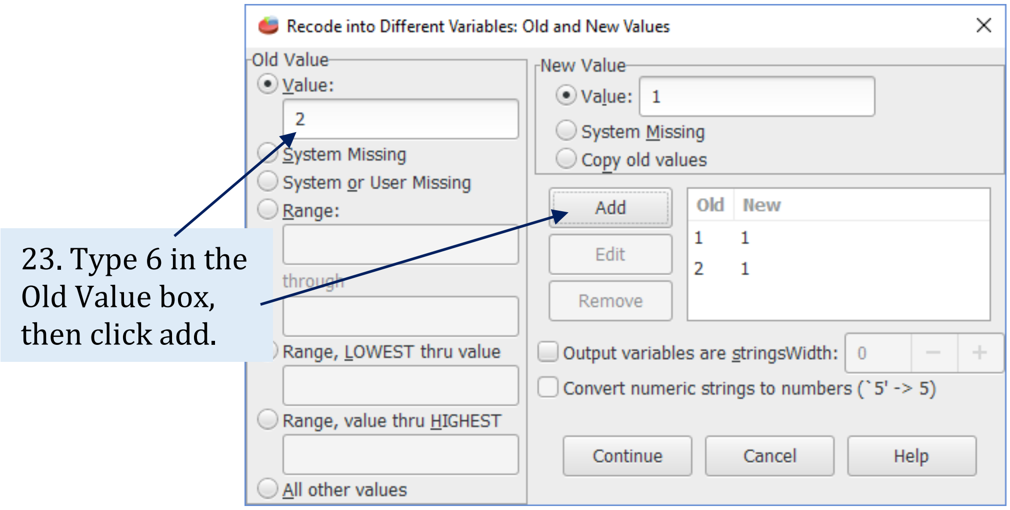
22. Type 2 in the Old Value box, then click add.

Recode into Different Variables: Old and New Values

Old Value: ● Value: 2

New Value: ● Value: 1

Old	New
1	1



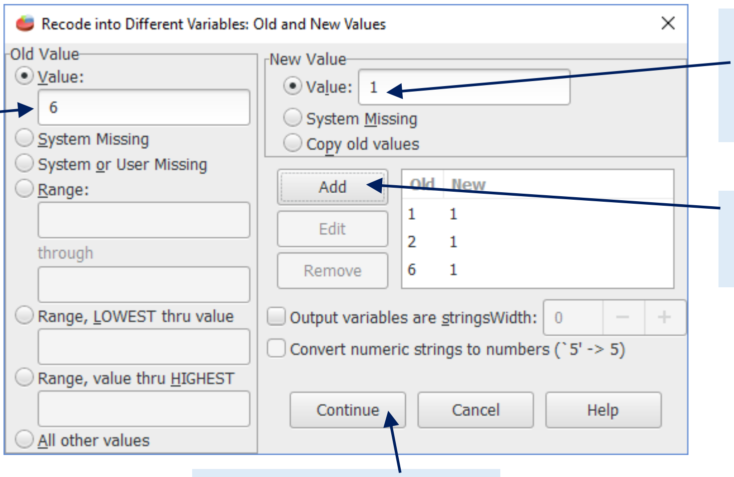
23. Type 6 in the Old Value box, then click add.

Recode into Different Variables: Old and New Values

Old Value: ● Value: 2

New Value: ● Value: 1

Old	New
1	1
2	1



Recode into Different Variables: Old and New Values

Old Value

● Value: 6

○ System Missing

○ System or User Missing

○ Range:

through

○ Range, LOWEST thru value

○ Range, value thru HIGHEST

○ All other values

New Value

● Value: 1

○ System Missing

○ Copy old values

Old	New
1	1
2	1
6	1

☐ Output variables are strings Width: 0

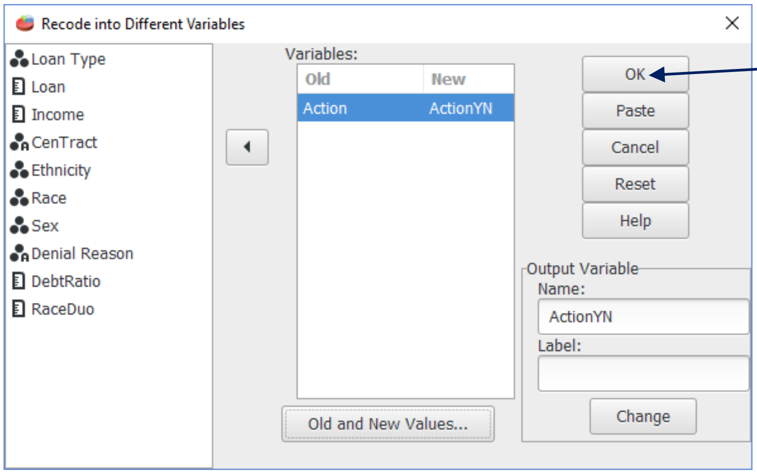
☐ Convert numeric strings to numbers (`5' -> 5)

24. Type 3 in the Old Value box.

25. Type 0 in the New Value box.

26. Click Add.

27. Click Continue.



Recode into Different Variables

- Loan Type
- Loan
- Income
- CenTract
- Ethnicity
- Race
- Sex
- Denial Reason
- DebtRatio
- RaceDuo

Variables:

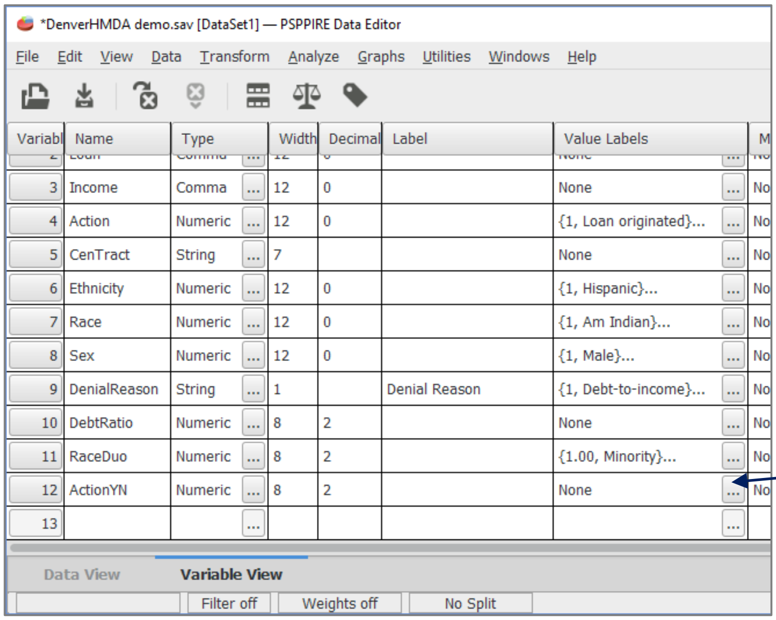
Old	New
Action	ActionYN

Output Variable

Name: ActionYN

Label:

28. Click OK.



*DenverHMDA demo.sav [DataSet1] — PSPPIRE Data Editor

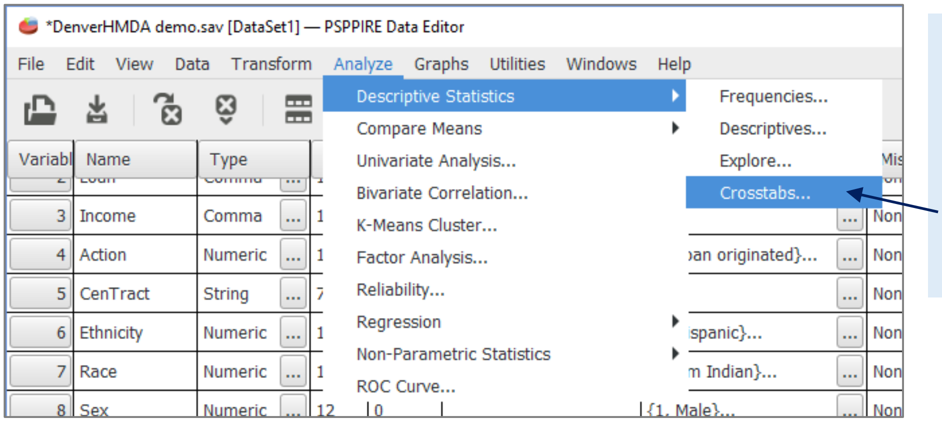
Variabl	Name	Type	Width	Decimal	Label	Value Labels	M
2	Loan	Comma	12	0		None	No
3	Income	Comma	12	0		None	No
4	Action	Numeric	12	0		{1, Loan originated}...	No
5	CenTract	String	7			None	No
6	Ethnicity	Numeric	12	0		{1, Hispanic}...	No
7	Race	Numeric	12	0		{1, Am Indian}...	No
8	Sex	Numeric	12	0		{1, Male}...	No
9	DenialReason	String	1		Denial Reason	{1, Debt-to-income}...	No
10	DebtRatio	Numeric	8	2		None	No
11	RaceDuo	Numeric	8	2		{1.00, Minority}...	No
12	ActionYN	Numeric	8	2		None	No
13							

28. In the Data Editor window, for our new variable ActionYN, click on the Value Labels ... button and assign a label of Denied to 0 and Accepted to 1.

PART 9: DATA ANALYSIS

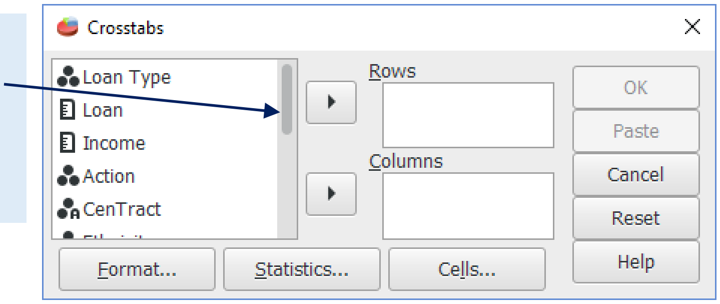
Now, let's see if there is a statistically significant relationship between action taken on loans and race. Keep in mind as we do that the numbers for each are small.

Because each of these is a categorical variable, few statistical tests are available. One that will work is a chi-square test available through a crosstab.



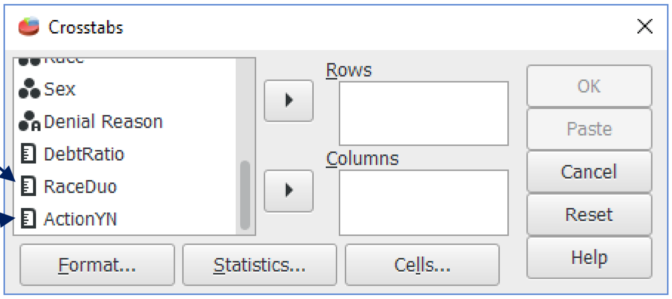
1. Under the Analyze menu, choose Descriptive Statistics and then Crosstabs ...

2. Scroll through the variables to get the two new ones at the bottom.

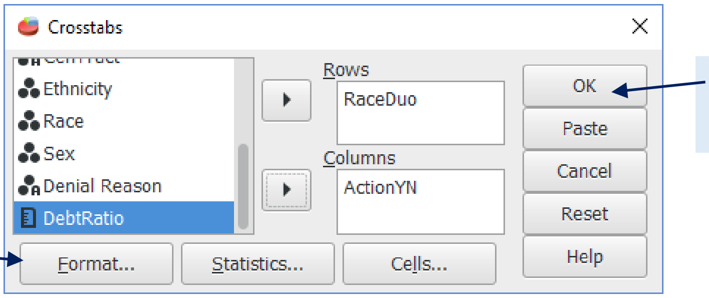


3. Select RaceDuo and move it into Rows.

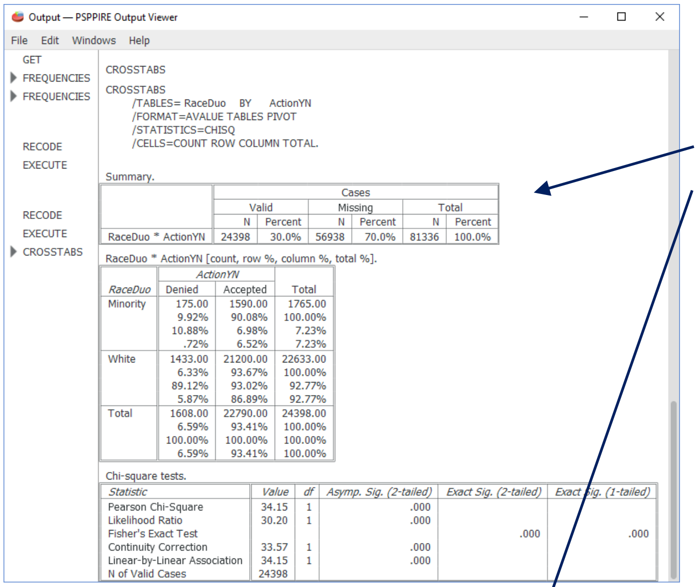
4. Select ActionYN and move it into Columns.



The Format, Statistics and Cells buttons offer many options, but the defaults will work fine for this test.



5. Click OK.



```
CROSSTABS

CROSSTABS
      /TABLES= RaceDuo    BY     ActionYN
      /FORMAT=AVALUE TABLES PIVOT
      /STATISTICS=CHISQ
      /CELLS=COUNT ROW COLUMN TOTAL.
```

Summary.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
RaceDuo * ActionYN	24398	30.0%	56938	70.0%	81336	100.0%

RaceDuo * ActionYN [count, row %, column %, total %].

	ActionYN		
RaceDuo	Denied	Accepted	Total
Minority	175.00 9.92% 10.88% .72%	1590.00 90.08% 6.98% 6.52%	1765.00 100.00% 7.23% 7.23%
White	1433.00 6.33% 89.12% 5.87%	21200.00 93.67% 93.02% 86.89%	22633.00 100.00% 92.77% 92.77%
Total	1608.00 6.59% 100.00% 6.59%	22790.00 93.41% 100.00% 93.41%	24398.00 100.00% 100.00% 100.00%

Chi-square tests.

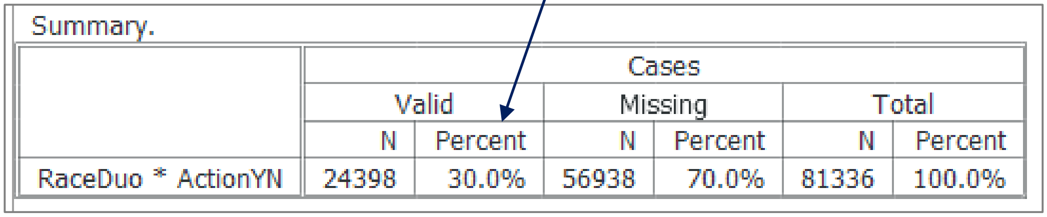
Statistic	*Value*	*df*	*Asymp. Sig. (2-tailed)*	*Exact Sig. (2-tailed)*	*Exact Sig. (1-tailed)*
Pearson Chi-Square	34.15	1	.000		
Likelihood Ratio	30.20	1	.000		
Fisher's Exact Test				.000	.000
Continuity Correction	33.57	1	.000		
Linear-by-Linear Association	34.15	1	.000		
N of Valid Cases	24398				

Results appear in the Output Viewer. Let's break these sections apart.

First, the Summary section reveals that just 30 percent of the 81,336 cases in the data set match both criteria in the variables ActionYN and RaceDuo:

- Applicant declares to be either white or a minority.
- Loan was either accepted or denied.

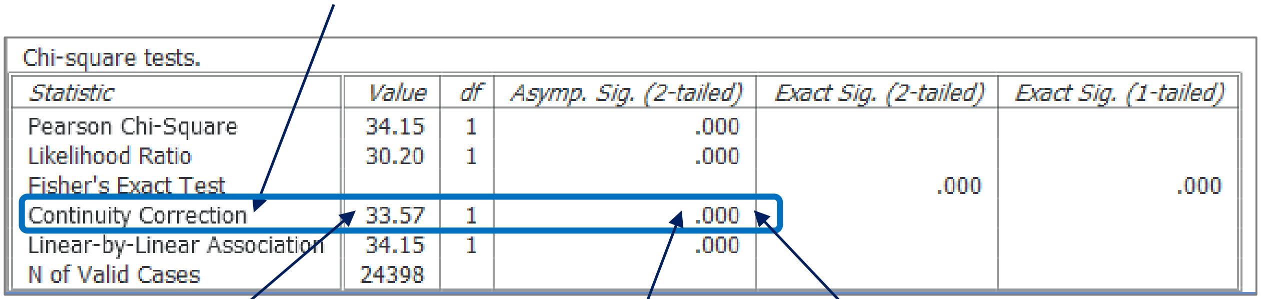
That 30% figure warns that we must be careful in evaluating the results. Our analysis is for just a subset.



Summary.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
RaceDuo * ActionYN	24398	30.0%	56938	70.0%	81336	100.0%

Next, skip to the bottom and the outcome of the chi-square test. Because this is a 2 X 2 test, we use the Continuity Correction line.



Chi-square tests.

Statistic	*Value*	*df*	*Asymp. Sig. (2-tailed)*	*Exact Sig. (2-tailed)*	*Exact Sig. (1-tailed)*
Pearson Chi-Square	34.15	1	.000		
Likelihood Ratio	30.20	1	.000		
Fisher's Exact Test				.000	.000
Continuity Correction	33.57	1	.000		
Linear-by-Linear Association	34.15	1	.000		
N of Valid Cases	24398				

The value reported, 33.57, is the size of the chi-square statistic. This is a relatively large value.

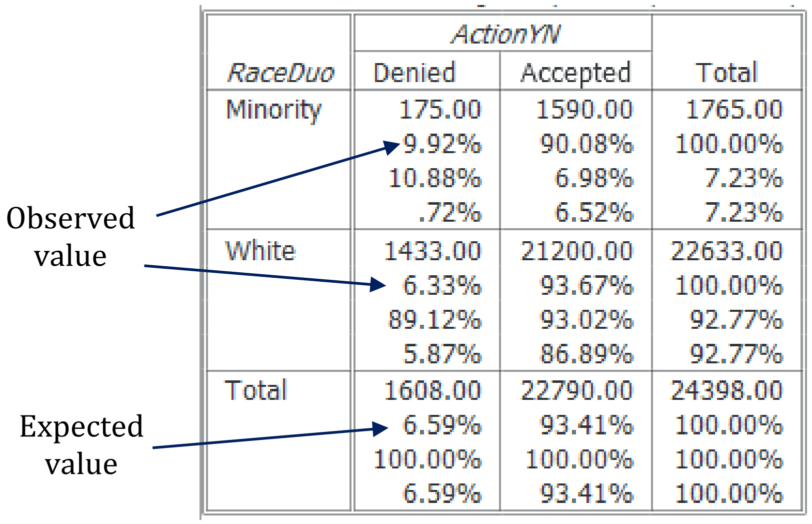
The .000 refer to probability. This is misleading, for a probability of 0 is impossible. PSPP is trying to tell us is that the probability is less than .001, or in the language of statistics: *p* < .001

For data involving people, a p-value of less than 5 percent (p < .05) is *statistically significant*. The *p* < .001 means that if there were no relationship between race and loan acceptance, the chance of getting these data would be less than 1 in 1,000. So something is going on here. We need to look at the middle table to figure out what that something is.

The chi-square statistic measures the difference between *expected* and *observed* values.

Expected value
In this data, 6.59 percent of all mortgage applications were denied. Thus, the *expected* value is that the 6.59 percent figure would be roughly similar for minorities and for whites.

Observed value
But the observed value is different. It shows that 9.92 percent of minority loans were denied while 6.33 percent of white loans were denied.



	ActionYN		
RaceDuo	Denied	Accepted	Total
Minority	175.00	1590.00	1765.00
	9.92%	90.08%	100.00%
	10.88%	6.98%	7.23%
	.72%	6.52%	7.23%
White	1433.00	21200.00	22633.00
	6.33%	93.67%	100.00%
	89.12%	93.02%	92.77%
	5.87%	86.89%	92.77%
Total	1608.00	22790.00	24398.00
	6.59%	93.41%	100.00%
	100.00%	100.00%	100.00%
	6.59%	93.41%	100.00%

Interpretation
PSPP has found a statistically significant relationship between race and denial of mortgage. Thus, we appear to have a news story here.

However, don't jump to the keyboard just yet. More analysis and evaluation are needed before you're ready to broadcast or publish.

Remember, this relationship involves only 30 percent of the mortgage applications. More choices are available than just accept or deny. Some withdraw an application or fail to complete it. And these mortgage applications are only for purchase of owner-occupied homes. They exclude other types of loans such as remodels or multi-family units.

Further, the denial portion is so small – 6.6 percent – that only a couple dozen cases in a year would change these results. Small numbers can yield unrepresentative results.

And we would want to look further at the loan-to-income ratio we computed to see if it is associated with race. It may be that the real difference is the debt ratio, not race.

But this exercise does show how journalists can find a story in data using a free statistical package. Happy sleuthing!