



Basic Data Cleaning in SPSS





Errors in Data

Errors and irrelevancies in data can occur due to:

- Errors in data input
- Errors in recording information between different data entry operators or due to changes over time.
- Information collected on non-applicable events or subjects.
- Mismatches between database tables.
- Difference in how various systems encode or represent data

Typical Tasks Encountered When Cleaning Data

- Identifying records or fields with missing values.
- Identifying records or fields with a degree of variability that is too high or too low. **For example, you have a variable with 5 levels but 99% of the data comes from only one level.**
- Values that are out of range. **For example, an entry for the age variable that is a negative value.**
- Duplicate records.
- Variables that are incorrectly formatted.
- Values in the variables that seem to contradict each other, or to imply errors in the data. **For example, if we are recording the ages of adult patients and we see an age of 11 years.**

Most data errors fall into two main classes:

1. Problems in the data storage and formatting in other data systems.
2. Problems with the data itself that were caused by human error or systematic issues.

Data Storage and Formatting Issues

- Issues with Variable Types - String, Numeric, etc.
- Issues with Variable Names.
- Issues with Variable Value Labels.
- Definition of Missing data values.
- Issues with how Data/Time data are recorded and stored.

Issues Stemming from human/system error:

- Irrelevant or non-applicable values or variables
- Duplicated data values
- Inconsistencies and illogical relationships
- Entry errors



Data Formatting Problems

Employee data file

1 - Formatting problem: variable name

	#id	gender	City	DOB	educ	var001	salary	salbegin	jobtime	prevep	minority	age
1	1 m	Chicago	03/FEB/1952	15	Manager		\$52,000	\$27,000	98	144	No	39
2	2 m	Chicago	23/MAY/1958	16	Clerical		\$40,200	\$18,750	x	36	No	33
3	3 f	Chicago	26/JUL/1929	12	Clerical		\$21,450	\$12,000	98	381	No	62
4	4 f	Chicago	15/APR/1947	8	Clerical		\$21,900	\$13,200	98	190	No	44
5	5 m	Chicago	09/FEB/1955	15	Clerical		\$45,000	\$21,000	98	138	No	36
6	6 m	Chicago	22/AUG/1958	15	Clerical		\$32,100	\$13,500	x	67	No	33
7	7 M	Chicago	26/APR/1956	15	Clerical		\$36,000	\$18,750	98	114	No	-9
8	8 f	Chicago	06/MAY/1966	12	Clerical		\$21,900	\$9,750	98	0	No	25
9	9 f	Chicago	23/JAN/1946	15	Clerical		\$27,900	\$12,750	98	115	No	45
10	10 f	Chicago	13/FEB/1946	12	Clerical		\$24,000	\$13,500	98	244	No	45
11	11 F	Chicago	07/FEB/1950	16	Clerical		\$30,300	\$16,500	98	143	No	41
12	12 m	Chicago	11/JAN/1966	8	Clerical		\$28,350	\$12,000	98	26	Yes	-9
13	13 m	Chicago	17/JUL/1960	15	Clerical		\$27,750	\$14,250	98	34	Yes	31
14	14 f	Chicago	26/FEB/1949	15	Clerical		\$35,100	\$16,800	98	137	Yes	42
15	15 m	Chicago	29/AUG/1962	12	Clerical		\$27,300	\$33,500	97	66	No	29
16	16 M	Chicago	17/NOV/1964	12	Clerical		\$40,800	\$15,000	97	24	No	27
17	17 m	Chicago	18/JUL/1962	15	Clerical		\$46,000	\$14,250	97	48	No	29
18	18 m	Chicago	20/MAR/1956	16	Manager		\$103,750	\$27,510	97	70	No	35
19	19 m	Chicago	19/AUG/1962	12	Clerical		\$42,300	\$14,250	97	103	No	29
20	20 F	Chicago	23/JAN/1940	12	Clerical		\$26,250	\$11,550	97	48	No	51
21	20 f	Chicago	23/JAN/1940	12	Clerical		\$26,250	\$11,550	97	48	No	51
22	21 f	Chicago	19/FEB/1963	16	Clerical		\$38,850	\$15,000	97	17	No	28
23	22 m	Chicann	24/SEP/1948	12	Clerical		\$21,750	\$12,750	97	114	Yes	41

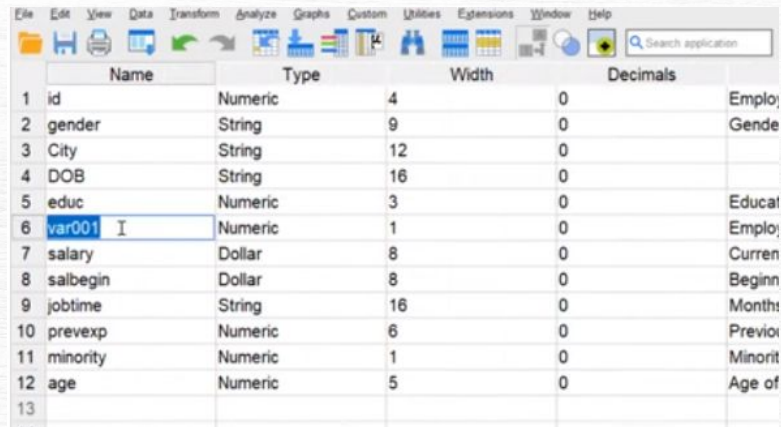
Column headings are variable names (not Labels), and spss found an issue in the naming of this variable in the system from which the file was imported, Excel, etc. In spss a name cannot begin with a number or any character other than a letter. So, if a variable is imported into spss with a name, let's say, beginning with a 5, then spss will toss out that name and give the variable a default name, var001, var002, etc.

How to fix this issue?

Double Click in the cell with “var001”



You will be taken to Variable View, where you can enter a proper name for the variable.



A screenshot of the Variable View window in SPSS. The window shows a list of variables with their names, types, widths, and decimals. The variable 'var001' is highlighted in blue.

	Name	Type	Width	Decimals	
1	id	Numeric	4	0	Emplo;
2	gender	String	9	0	Gende
3	City	String	12	0	
4	DOB	String	16	0	
5	educ	Numeric	3	0	Educational
6	var001	Numeric	1	0	Employment
7	salary	Dollar	8	0	Currency
8	salbegin	Dollar	8	0	Beginning
9	jobtime	String	16	0	Month
10	prevexp	Numeric	6	0	Previous
11	minority	Numeric	1	0	Minority
12	age	Numeric	5	0	Age of
13					

Employee data file

2 - Formatting problem: variable type

	id	gender	City	DOB	educ	var001	salary	salbegin	jobtime	prevexp	minority	age
1	1	m	Chicago	03/FEB/1952	15	Manager	\$57,000	\$27,000	98	144	No	39
2	2	m	Chicago	23/MAY/1958	16	Clerical	\$40,200	\$18,750	x	36	No	33
3	3	f	Chicago	26/JUL/1929	12	Clerical	\$21,450	\$12,000	98	381	No	62
4	4	f	Chicago	15/APR/1947	8	Clerical	\$21,900	\$13,200	98	190	No	44
5	5	m	Chicago	09/FEB/1955	15	Clerical	\$45,000	\$21,000	98	138	No	36
6	6	m	Chicago	22/AUG/1958	15	Clerical	\$32,100	\$13,500	x	67	No	33
7	7	M	Chicago	26/APR/1956	15	Clerical	\$36,000	\$18,750	98	114	No	-9
8	8	f	Chicago	06/MAY/1966	12	Clerical	\$21,900	\$9,750	98	0	No	25
9	9	f	Chicago	23/JAN/1946	15	Clerical	\$27,900	\$12,750	98	115	No	45
10	10	f	Chicago	13/FEB/1946	12	Clerical	\$24,000	\$13,500	98	244	No	45
11	11	F	Chicago	07/FEB/1950	16	Clerical	\$30,300	\$16,500	98	143	No	41
12	12	m	Chicago	11/JAN/1968	8	Clerical	\$28,350	\$12,000	98	26	Yes	-9
13	13	m	Chicago	17/JUL/1960	15	Clerical	\$27,750	\$14,250	98	34	Yes	31
14	14	f	Chicago	26/FEB/1949	15	Clerical	\$35,100	\$16,800	98	137	Yes	42
15	15	m	Chicago	29/AUG/1962	12	Clerical	\$27,300	\$33,500	97	66	No	29
16	16	M	Chicago	17/NOV/1964	12	Clerical	\$40,800	\$15,000	97	24	No	27
17	17	m	Chicago	18/JUL/1962	15	Clerical	\$46,000	\$14,250	97	48	No	29
18	18	m	Chicago	20/MAR/1956	16	Manager	\$103,750	\$27,510	97	70	No	35
19	19	m	Chicago	19/AUG/1962	12	Clerical	\$42,300	\$14,250	97	103	No	29
20	20	F	Chicago	23/JAN/1940	12	Clerical	\$26,250	\$11,550	97	48	No	51
21	21	F	Chicago	23/JAN/1940	12	Clerical	\$26,250	\$11,550	97	48	No	51
22	21	f	Chicago	18/FEB/1963	16	Clerical	\$38,850	\$15,000	97	17	No	28
23	22	m	Chicann	24/FEB/1948	12	Clerical	\$21,750	\$12,750	97	116	Var	61

When spss reads in a data variable, it tries to guess what type of variable it is. Sometimes it gets it wrong. If it guesses a variable to be ordinal, it places rectangles next to the variable name ( var001). If it guesses it to be nominal, it places 3 circles next to the variable name ( salary). The three circles at Salary” tells us that it guessed the salary variable to be nominal. However, we know that salary is a numeric or scale variable, so we need to correct it.

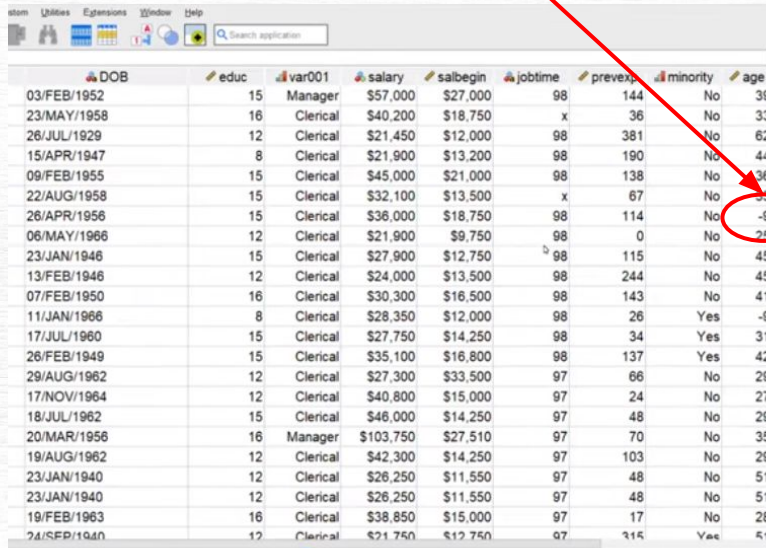
How to fix this issue?

To fix this problem, we double click on the variable name cell. We will be taken to Variable view and the Measure column, where we enter the correct measure. In this case, we will enter it as “**scale**”.

	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
7	0	Current Salary	None	\$0	8	Right	Nominal	Input
8	0	Beginning Salary	None	\$0	8	Right	Scale	Input
9	0	Months since Hire	None	None	8	Right	Nominal	Input
10	0	Previous Experienc...	None	None	8	Right	Scale	Input
11	0	Minority Classification (0, No)...	9	9	8	Right	Ordinal	Input
12	0	Age of Employee	None	None	5	Right	Scale	Input
13								

Employee data file

3 - Formatting Problem: Missing values

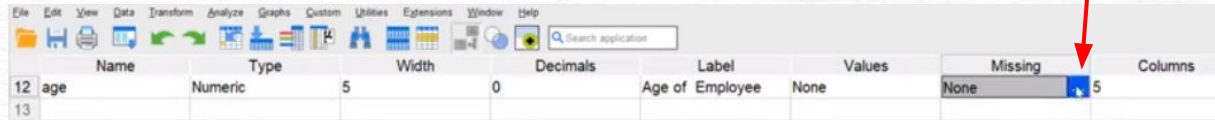


DOB	educ	var001	salary	salbegin	jobtime	prevexp	minority	age
03/FEB/1952	15	Manager	\$57,000	\$27,000	98	144	No	39
23/MAY/1958	16	Clerical	\$40,200	\$18,750	x	36	No	33
26/JUL/1929	12	Clerical	\$21,450	\$12,000	98	381	No	62
15/APR/1947	8	Clerical	\$21,900	\$13,200	98	190	No	44
09/FEB/1955	15	Clerical	\$45,000	\$21,000	98	138	No	36
22/AUG/1958	15	Clerical	\$32,100	\$13,500	x	67	No	33
26/APR/1956	15	Clerical	\$36,000	\$18,750	98	114	No	-9
06/MAY/1966	12	Clerical	\$21,900	\$9,750	98	0	No	25
23/JAN/1946	15	Clerical	\$27,900	\$12,750	98	115	No	45
13/FEB/1946	12	Clerical	\$24,000	\$13,500	98	244	No	45
07/FEB/1950	16	Clerical	\$30,300	\$16,500	98	143	No	41
11/JAN/1966	8	Clerical	\$28,350	\$12,000	98	26	Yes	-9
17/JUL/1960	15	Clerical	\$27,750	\$14,250	98	34	Yes	31
26/FEB/1949	15	Clerical	\$35,100	\$16,800	98	137	Yes	42
29/AUG/1962	12	Clerical	\$27,300	\$33,500	97	66	No	29
17/NOV/1964	12	Clerical	\$40,800	\$15,000	97	24	No	27
18/JUL/1962	15	Clerical	\$46,000	\$14,250	97	48	No	29
20/MAR/1956	16	Manager	\$103,750	\$27,510	97	70	No	35
19/AUG/1962	12	Clerical	\$42,300	\$14,250	97	103	No	29
23/JAN/1940	12	Clerical	\$26,250	\$11,550	97	48	No	51
23/JAN/1940	12	Clerical	\$26,250	\$11,550	97	48	No	51
19/FEB/1963	16	Clerical	\$38,850	\$15,000	97	17	No	28
24/SEP/1940	12	Clerical	\$21,750	\$12,750	97	115	Yes	51

Here we see that an Age entry says “-9”. Now, this can be classified as a systematic or human error. However, it becomes a missing value problem since we need to remove the ‘-9’ and we do not know the true age of this employee.

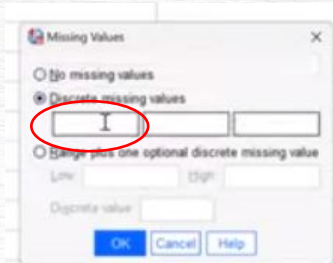
How to fix this issue?

To fix this, we double-click on the heading cell with the Age variable. We will be taken to Variable view. In the Missing Column, click on the right (blue) side of the cell.

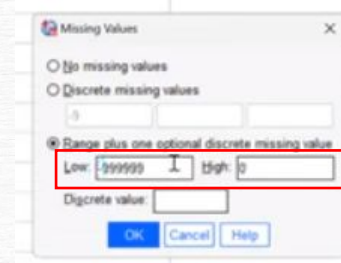


The following dialog box will come up. And we can do one of two things.

In the “Discrete missing value” box, enter **-9**. Click OK. SPSS will remove the negative number.



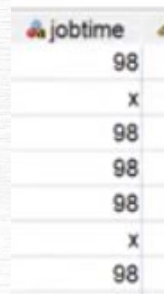
Or you could give spss a range of values which, if it encounters any one number in this range, it will consider it to be a missing value. In this case, we could give it the range of -99999 to 0, as shown below.



More on Missing Values

If a value is missing in a numeric variable field, spss places a period in the cell of the missing value. This says that the missing value is a “system missing.”

If the value is missing in a nominal field, it is simply left blank. For example, in the variable “jobtime” we have several cells with “x” entered.



jobtime
98
x
98
98
98
x
98

Upon closer inspection, we find that “jobtime” which is a numeric variable, has 3 circles next to its name. SPSS has guessed it to be nominal. So we double-click and change its Type to “numeric” and its measure to “scale”. After doing this, we will see that the “x” entries are now “.”



jobtime
98
.
98
98
98
.
98

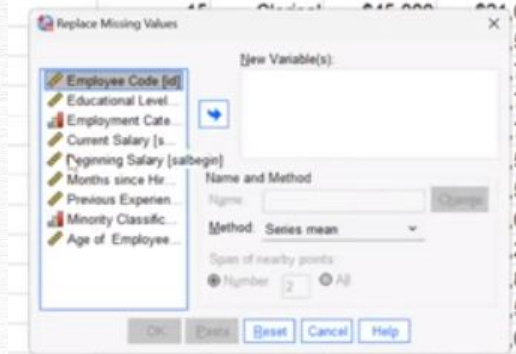
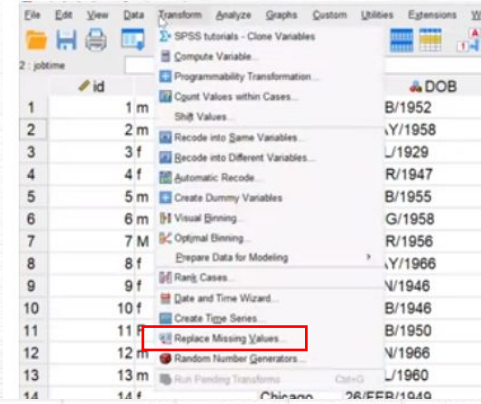
What do we do with Missing Values?

We can either ignore the missing value, in which case spss will do its calculations of descriptive statistics of the variable without this value, or we could have spss impute some value in its place.

For numeric data, the more commonly used practice is to replace the value with the series mean. SSPSS will replace the value with the mean for that variable.

Highlight the “.” that indicates a missing value. Go to the Transform menu. Click on “**Replace Missing Values**”.

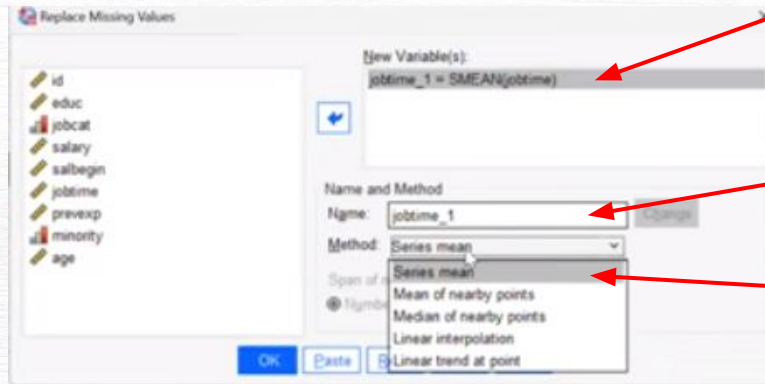
Highlight the variable from the column on the left (If you can't find your variable, then try its Label instead of Name). Click the middle arrow to move the variable to the right column.



What do we do with Missing Values?

This dialog box will arise.

Don't worry if spss changed the variable name here. You can always change it back in Variable view.



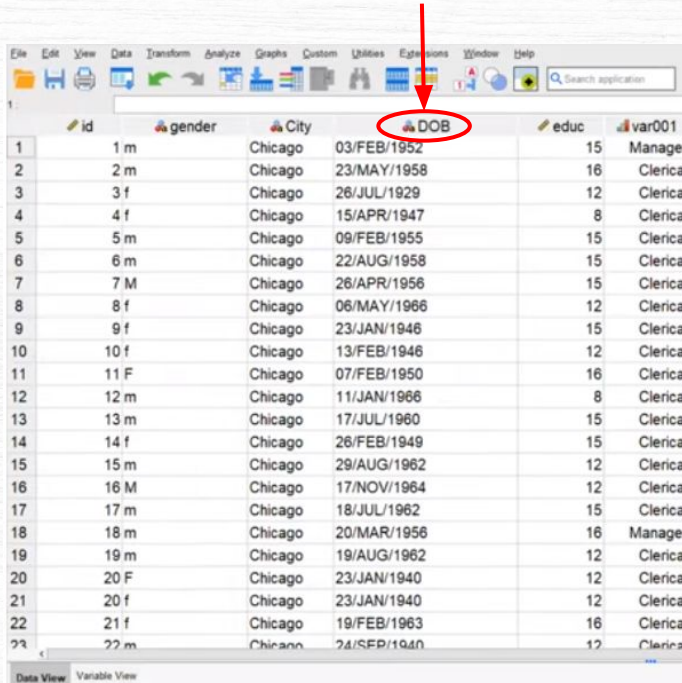
Spss will attach an underscore 1 (_1) to the variable name here. Just remove the underscore 1.

Under “Method” choose “series mean”

SPSS will now impute the series mean wherever it finds a missing value, or a “.”

Employee data file

4 - Formatting Problem: Date/Time



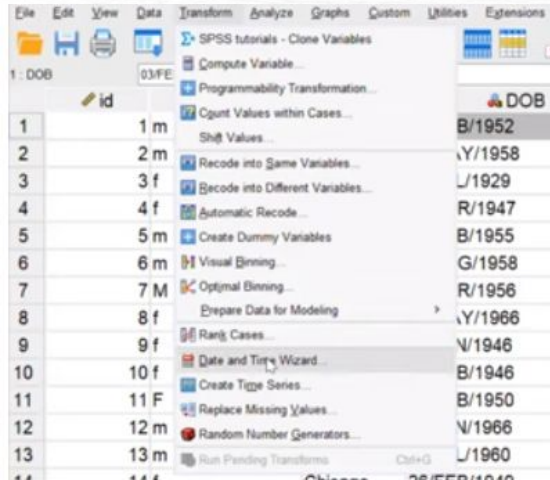
	id	gender	City	DOB	educ	var001
1	1 m		Chicago	03/FEB/1952	15	Manager
2	2 m		Chicago	23/MAY/1958	16	Clerical
3	3 f		Chicago	26/JUL/1929	12	Clerical
4	4 f		Chicago	15/APR/1947	8	Clerical
5	5 m		Chicago	09/FEB/1955	15	Clerical
6	6 m		Chicago	22/AUG/1958	15	Clerical
7	7 M		Chicago	26/APR/1956	15	Clerical
8	8 f		Chicago	06/MAY/1966	12	Clerical
9	9 f		Chicago	23/JAN/1946	15	Clerical
10	10 f		Chicago	13/FEB/1946	12	Clerical
11	11 F		Chicago	07/FEB/1950	16	Clerical
12	12 m		Chicago	11/JAN/1966	8	Clerical
13	13 m		Chicago	17/JUL/1960	15	Clerical
14	14 f		Chicago	26/FEB/1949	15	Clerical
15	15 m		Chicago	29/AUG/1962	12	Clerical
16	16 M		Chicago	17/NOV/1964	12	Clerical
17	17 m		Chicago	18/JUL/1962	15	Clerical
18	18 m		Chicago	20/MAR/1956	16	Manager
19	19 m		Chicago	19/AUG/1962	12	Clerical
20	20 F		Chicago	23/JAN/1940	12	Clerical
21	20 f		Chicago	23/JAN/1940	12	Clerical
22	21 f		Chicago	19/FEB/1963	16	Clerical
23	22 m		Chicago	24/SEP/1940	17	Clerical

We see an issue with the DOB or date of birth variable. This is supposed to be a 'Date' type. However, we see the 3 circles next to its name, which means that spss has guessed it to be a string or nominal variable. We must fix this.

How to fix this issue?

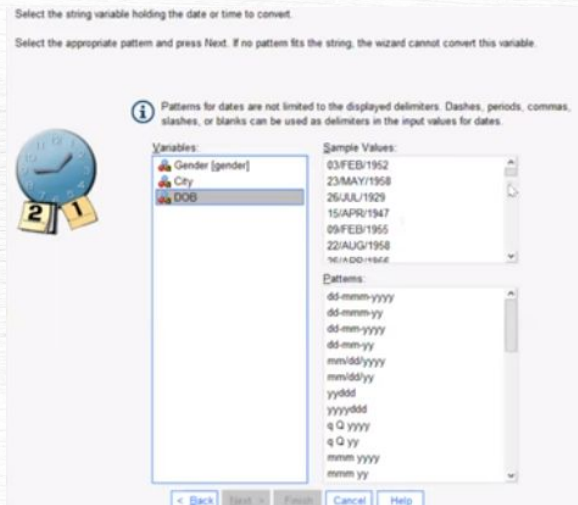
Transform → Date and Time Wizard

In the wizard, choose **“Create a date/time variable from a string containing a date or time”**.



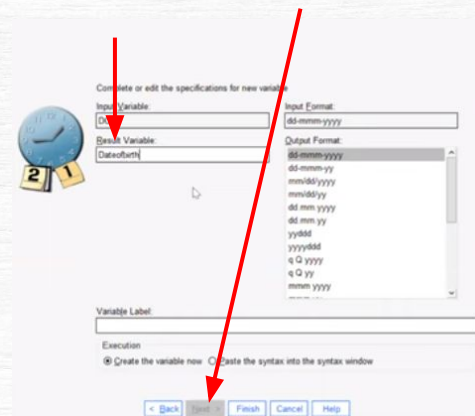
How to fix this issue?

Highlight your variable from the list of three string variables on the left column and choose the date/time format you want from the list on the right.



Next you must give the new variable a name. We call it **“dateofbirth”**. On the right will be an example of the output for this variable, or how the date values of the variable will look.

When satisfied, click **“Finish”**.



How to fix this issue?

City	DOB	Dateofbirth	educ
Chicago	03/FEB/1952	03-Feb-1952	15
Chicago	23/MAY/1958	23-May-1958	16
Chicago	26/JUL/1929	26-Jul-1929	12
Chicago	15/APR/1947	15-Apr-1947	8
Chicago	09/FEB/1955	09-Feb-1955	15
Chicago	22/AUG/1958	22-Aug-1958	15
Chicago	26/APR/1956	26-Apr-1956	15
Chicago	06/MAY/1966	06-May-1966	12
Chicago	23/JAN/1946	23-Jan-1946	15
Chicago	13/FEB/1946	13-Feb-1946	12
Chicago	07/FEB/1950	07-Feb-1950	16
Chicago	11/JAN/1966	11-Jan-1966	8
Chicago	17/JUL/1960	17-Jul-1960	15
Chicago	26/FEB/1949	26-Feb-1949	15

Old variable, which can now be deleted.

New Variable.



Problems with the Data Itself

Data Validation

The fundamental method of dealing with problems with the data itself is to validate the data.

Data validation is the process of verifying and validating data that is collected before it is used. Any type of data handling task, whether it is gathering data, analyzing it, or structuring it for presentation, must include data validation to ensure accurate results.

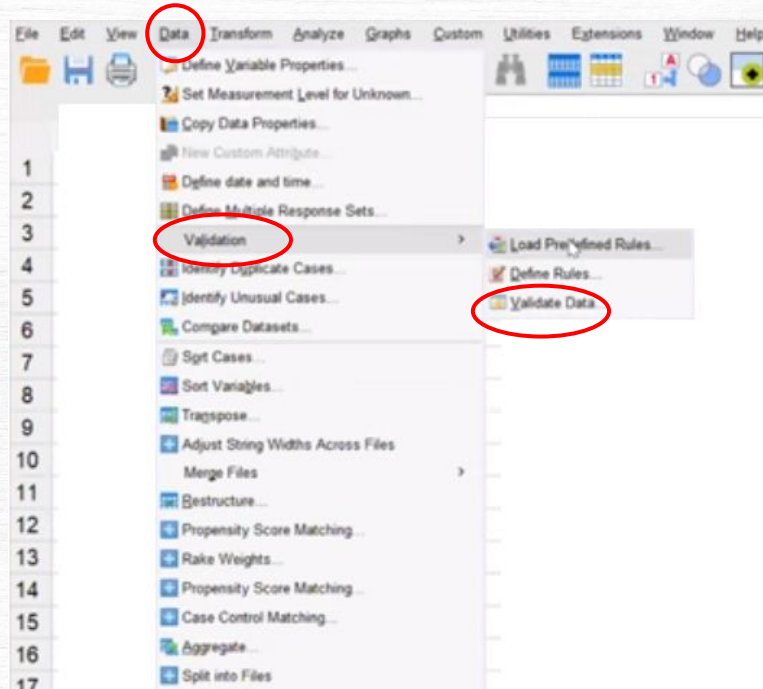
For example, in the gender field, we expect only M or F. Anything else entered into this field will be flagged as illegal or outside of the rules. As another example, there will be validation rules to ensure that the abbreviation for US states are appropriate. An entry “Ala” for Alaska will be flagged.

Several validation checks are built into the spss system which ensures the data being entered and stored has logical consistency

We can define our own rules for validation of the variables and entries in their fields, but the easiest procedure for data validation in spss is to use basic validation with predefined rules.

Data Validation

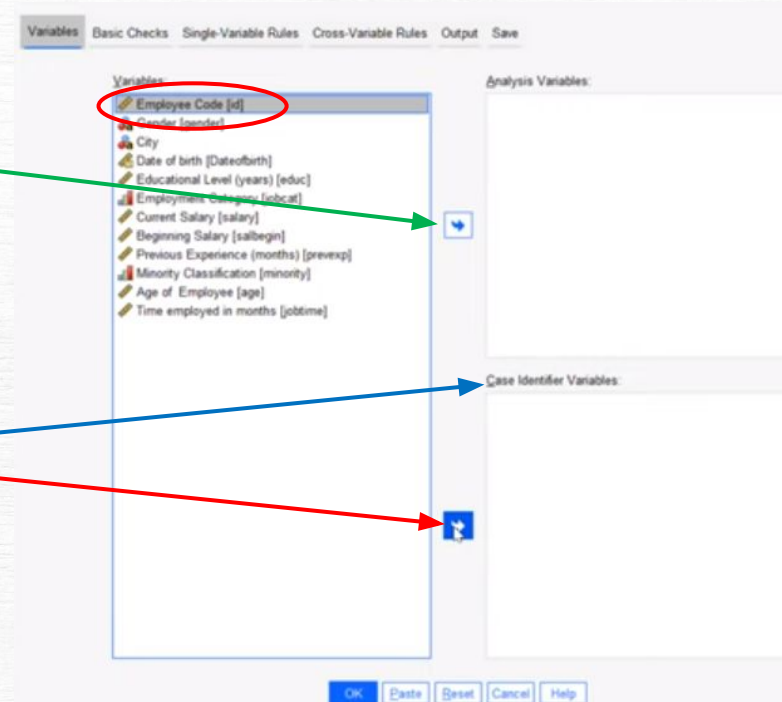
To do this, we go to the Data Menu → Validation → Validate Data.



Data Validation

Highlight all the other variables on the left and click on the upper arrow to send them all to the “**Analysis Variables**” box on the right.

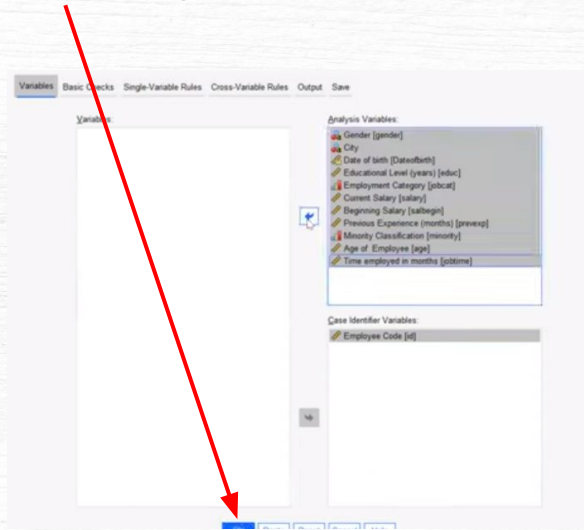
In the validation dialog box that comes up, we will use Employee Code ID as the identifier for all cases. So we will highlight this variable on the left and click **the bottom arrow** to send it to the “**Case Identifier Variables**” box on the right.



Data Validation

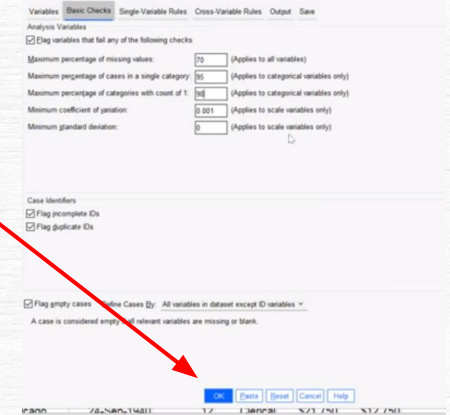
The dialog box should look like this when you are finished.

Click **OK** to begin validation.



By clicking on the tabs on the box you can see the types of rules that spss will use to flag data. Here you see the rules under Basic Checks. We are not going to get into writing your own rules via some of the other tabs like “Single Variable Rules” etc.

Click **OK** when finished looking at the tabs.



Data Validation

N	Valid	476
	Missing	3
Mean		12-OCT-1956
Median		31-JAN-1962
Std. Deviation		4305.14.32
Range		15340.00.0
Minimum		10-FEB-1929
Maximum		10-FEB-1971

Validate Data

Warnings

Some or all requested output is not displayed because all cases, variables, or data values passed the requested checks.

Variable Checks

Categorical Cases Constant > 95 City

Each variable is reported with every check it fails.

Identifier Checks

Incomplete Identifiers

Case	Identifier Employee Code
479	



Duplicate Identifiers

Duplicate Identifiers Group	Number of Duplicates	Cases with Duplicate Identifiers	Identifier Employee Code
1	2	20, 21	20



After clicking **OK**, you will get a validation report like the one shown below in which spss will report all flags.

Here the report will flag the variable City because rule that more than 95% of the data cannot belong to the same category. We can leave this one alone because all cases are from the city of Chicago, so there will be a high number of constant cases (100%). Alternatively, we could safely delete the field City because all cases will be from the same city. In other words, City is not a real variable.

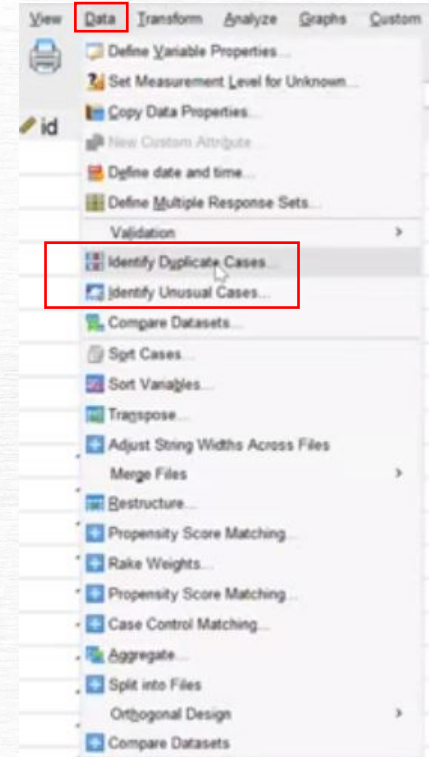
Here we are being told that case #479 (row #) has no entry in the Employee Code field. We go to case 479 and fix it.

Here we are being told that two duplicate cases have been found: Cases 20 and 21. We go there and check that they are actually duplicates. If they are, we can delete one of them.

Dedicated Functions

SPSS also has a dedicated function that can identify duplicate cases, another that can identify anomalies, that is, weird cases with entries totally different from the expected.

To find these functions, go to the Data menu and then to “**Identify Duplicate Cases,**” or ‘Identify Unusual Cases”



Although, for our purposes, we will not explore the other options in the validation function and other data cleaning methods available, there is a lot more to data cleaning in spss and I would encourage you to explore some of this on your own to see what's there.

I hope you found this presentation useful.

Thanks for stopping by.

See you in the next presentation.



**End of
Presentation**