EXAMINING YOUR DATA

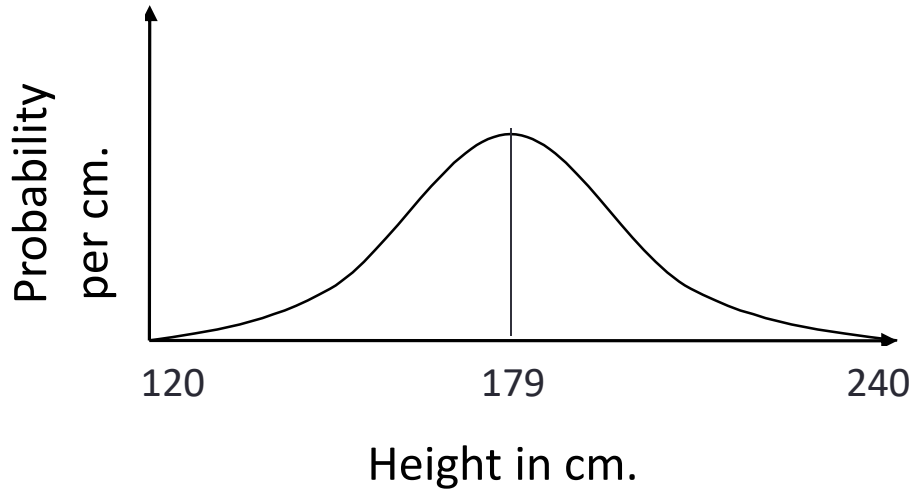
Chapter 2: Multivariate Insurance

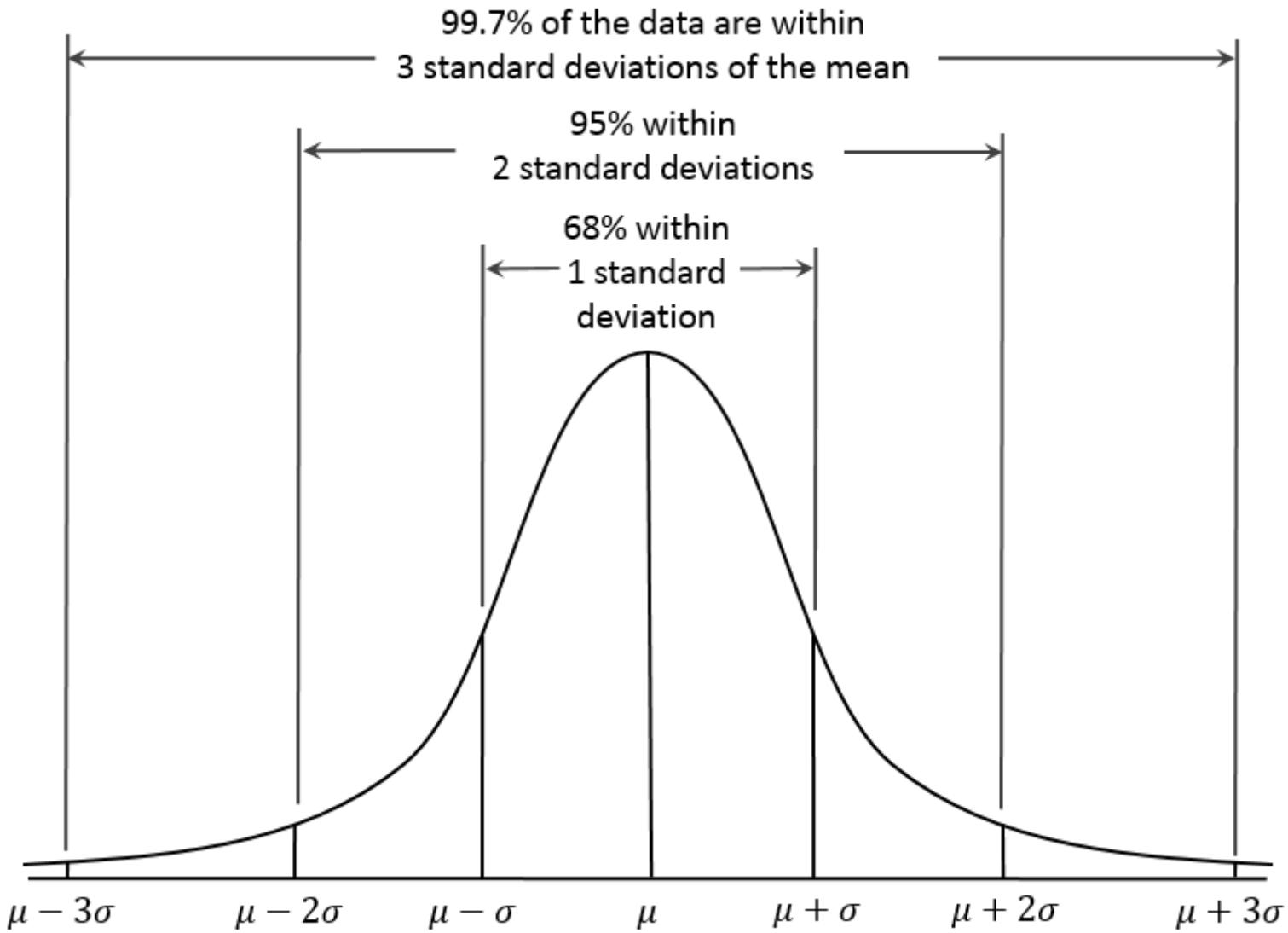
Topics

1. The normal distribution and normality.
2. Relationships between variables (correlation).
3. Outliers.
4. Missing data.

1. The Normal Distribution

A probability sample of male, 19 year old Swedes had their height measured. The majority were 179 cm tall, with some shorter and some taller.
The distribution is symmetric around the average (the mean $\bar{X}$).




99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Checking Normality

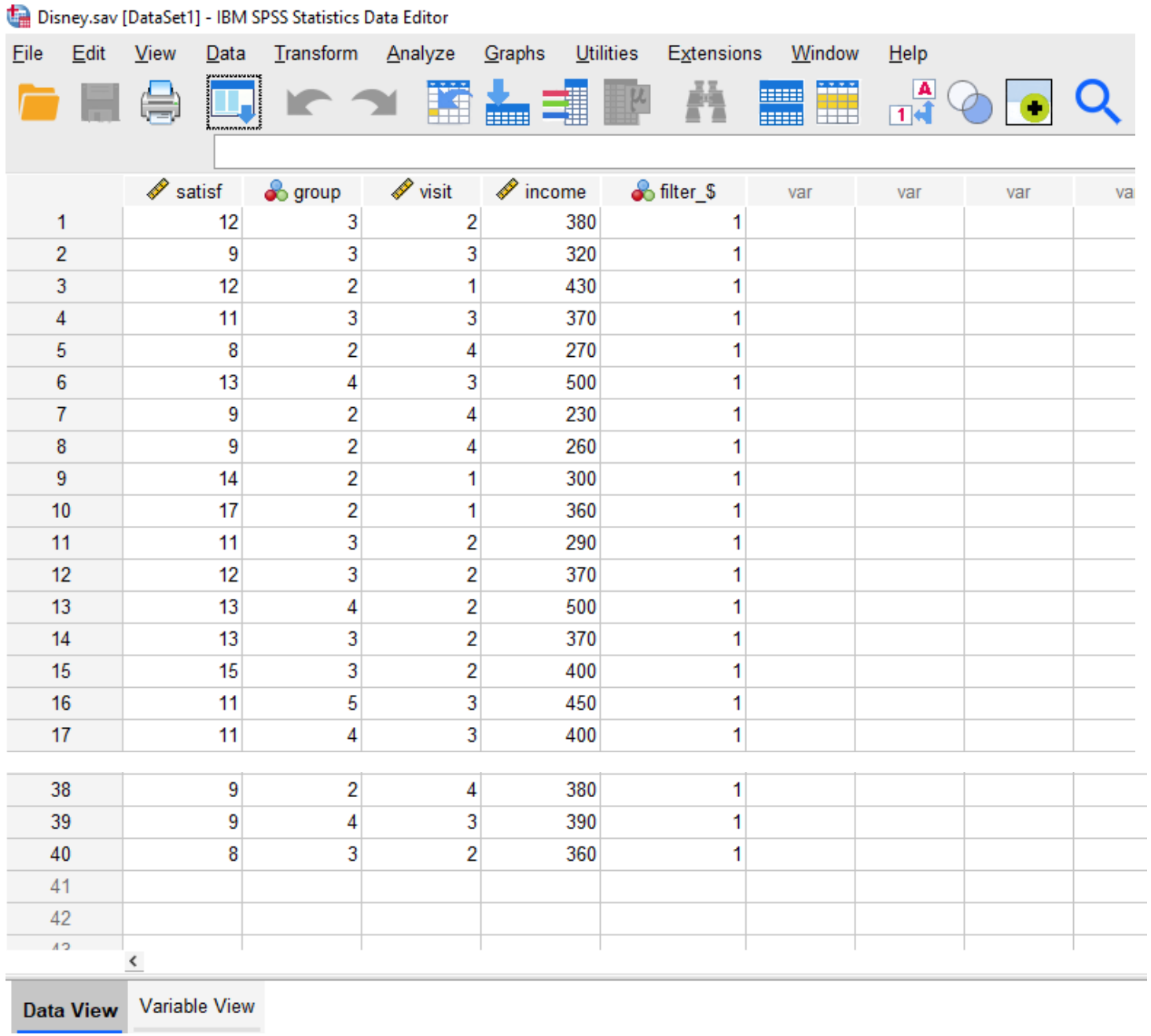
You can examine normality visually or statistically. There is no, ”one best way”, so it is often good to *consider several*.

- Skewness & Kurtosis
- Histograms: not so good with small samples.
- Statistical tests like Kolmogorov-Smirnov or Shapiro-Wilk (n<50).
- Box plots: Handy for identifying outliers.

How normal does the data need to be?

- It depends on the multivariate technique.
- e.g. Linear regression is quite robust against violations of normality.

Now we get SPSS going for everyone in the class



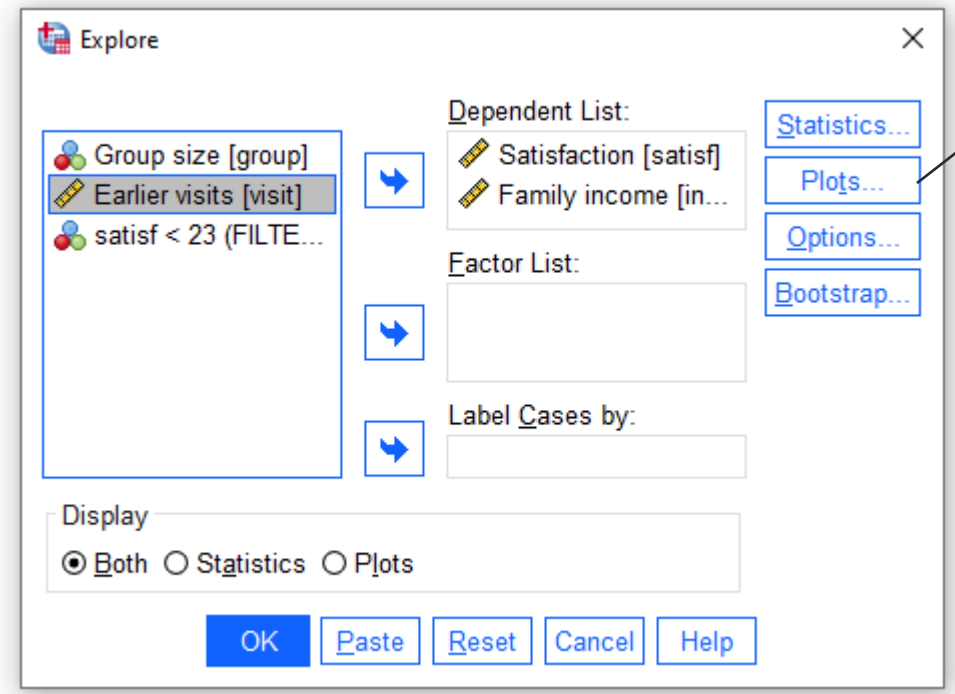
	satisf	group	visit	income	filter_$	var	var	var	va
1	12	3	2	380	1				
2	9	3	3	320	1				
3	12	2	1	430	1				
4	11	3	3	370	1				
5	8	2	4	270	1				
6	13	4	3	500	1				
7	9	2	4	230	1				
8	9	2	4	260	1				
9	14	2	1	300	1				
10	17	2	1	360	1				
11	11	3	2	290	1				
12	12	3	2	370	1				
13	13	4	2	500	1				
14	13	3	2	370	1				
15	15	3	2	400	1				
16	11	5	3	450	1				
17	11	4	3	400	1				
38	9	2	4	380	1				
39	9	4	3	390	1				
40	8	3	2	360	1				
41									
42									

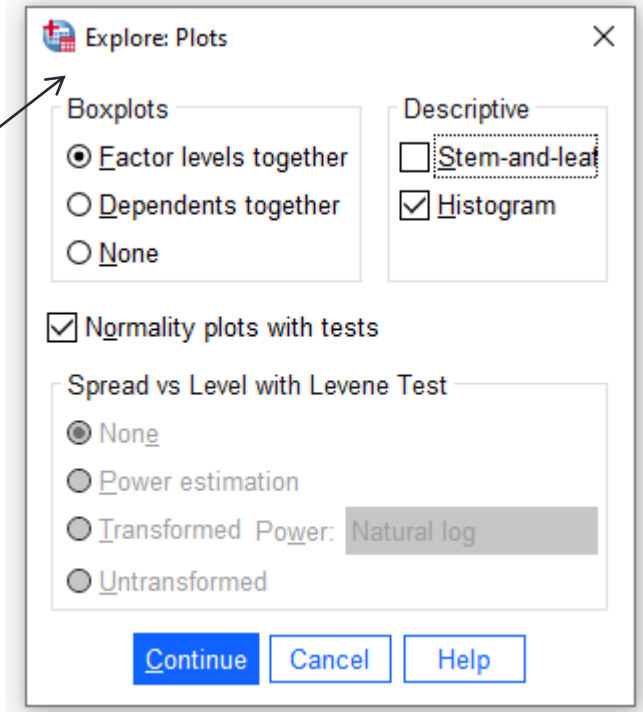
Disney.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	satisf	Numeric	8	0	Satisfaction	{1, not satis...	None	8	Right	Scale	Input
2	group	Numeric	8	0	Group size	None	None	8	Right	Nominal	Input
3	visit	Numeric	8	0	Earlier visits	None	None	8	Right	Scale	Input
4	income	Numeric	8	0	Family income	None	-99, 99	8	Right	Scale	Input
5	filter_$	Numeric	1	0	satisf < 23 (FIL...	{0, Not Sele...	None	10	Right	Nominal	Input
6											
7											
43											
44											

Data View **Variable View**


Explore
Group size [group]
Earlier visits [visit]
satisf < 23 (FILTE...
Dependent List:
Satisfaction [satisf]
Family income [in...
Factor List:
Label Cases by:
Statistics...
Plots...
Options...
Bootstrap...
Display
Both
Statistics
Plots
OK
Paste
Reset
Cancel
Help


Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Normality plots with tests
Spread vs Level with Levene Test
None
Power estimation
Transformed
Power:
Natural log
Untransformed
Continue
Cancel
Help

Skewness & Kurtosis

Descriptives

			Statistic	Std. Error
Satisfaction	Mean		11,33	,535
	95% Confidence Interval for Mean	Lower Bound	10,24	
		Upper Bound	12,42	
	Skewness		1,188	,374
	Kurtosis		2,865	,733
Family income	Mean		369,75	16,316
	95% Confidence Interval for Mean	Lower Bound	336,75	
		Upper Bound	402,75	
	5% Trimmed Mean		365,00	
	Median		370,00	
	Variance		10648,654	
	Std. Deviation		103,192	
	Minimum		200	
	Maximum		700	
	Range		500	
	Interquartile Range		148	
	Skewness		,624	,374
	Kurtosis		1,359	,733

Rule of Thumb:
Less than 1 (absolute value) is normal

Side note: If you ask a statistician who calculates skewness by hand, s/he will say the cutoff is +/- 3. most software standardizes both measures to +/- 1.

Statistical Tests

Rule of Thumb:
Big (> 0.05) is good (normal).

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Satisfaction	,146	40	,032	,917	40	,006
Family income	,087	40	,200*	,956	40	,122

*. This is a lower bound of the true significance.

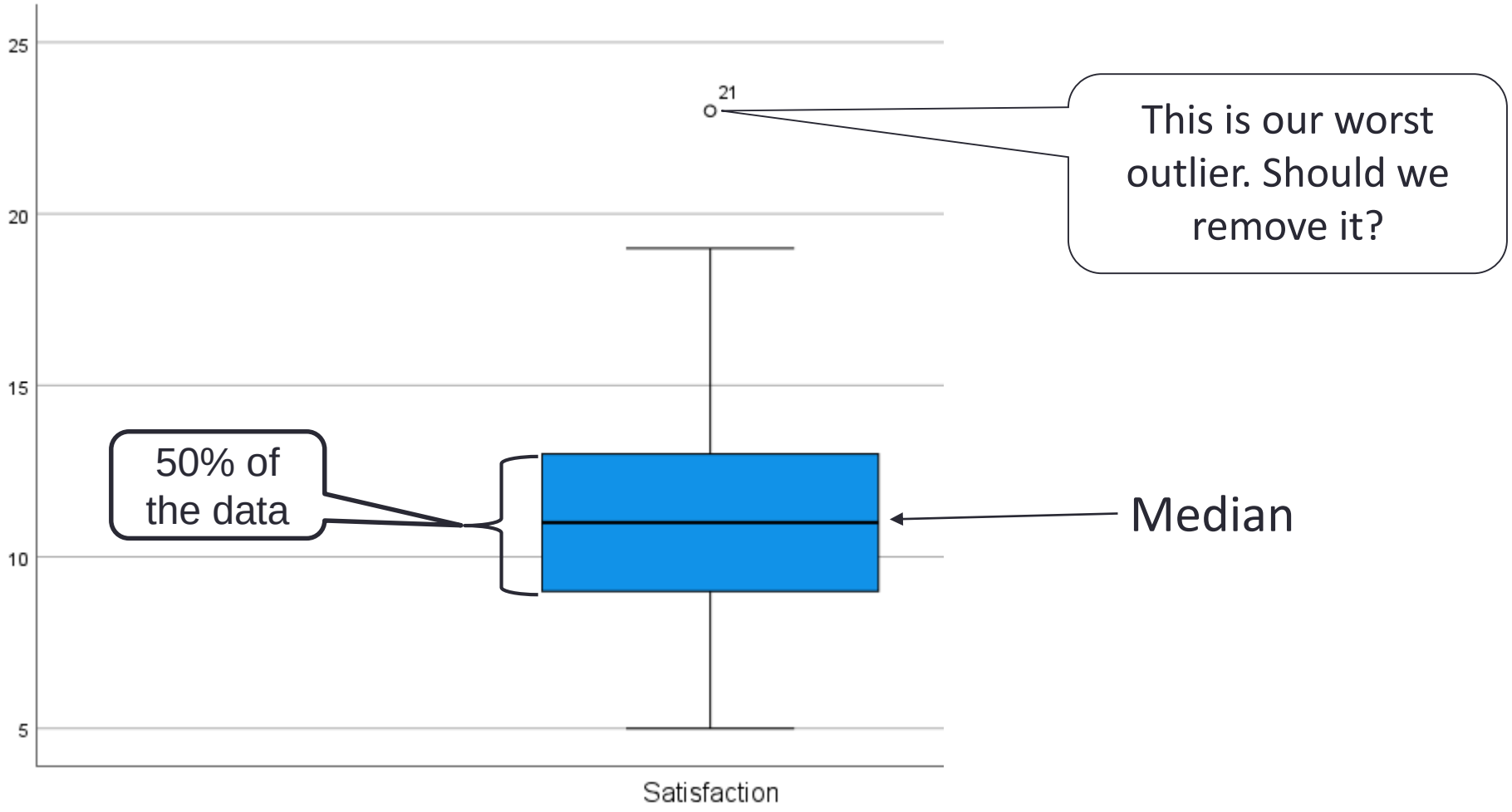
a. Lilliefors Significance Correction

Abnormal

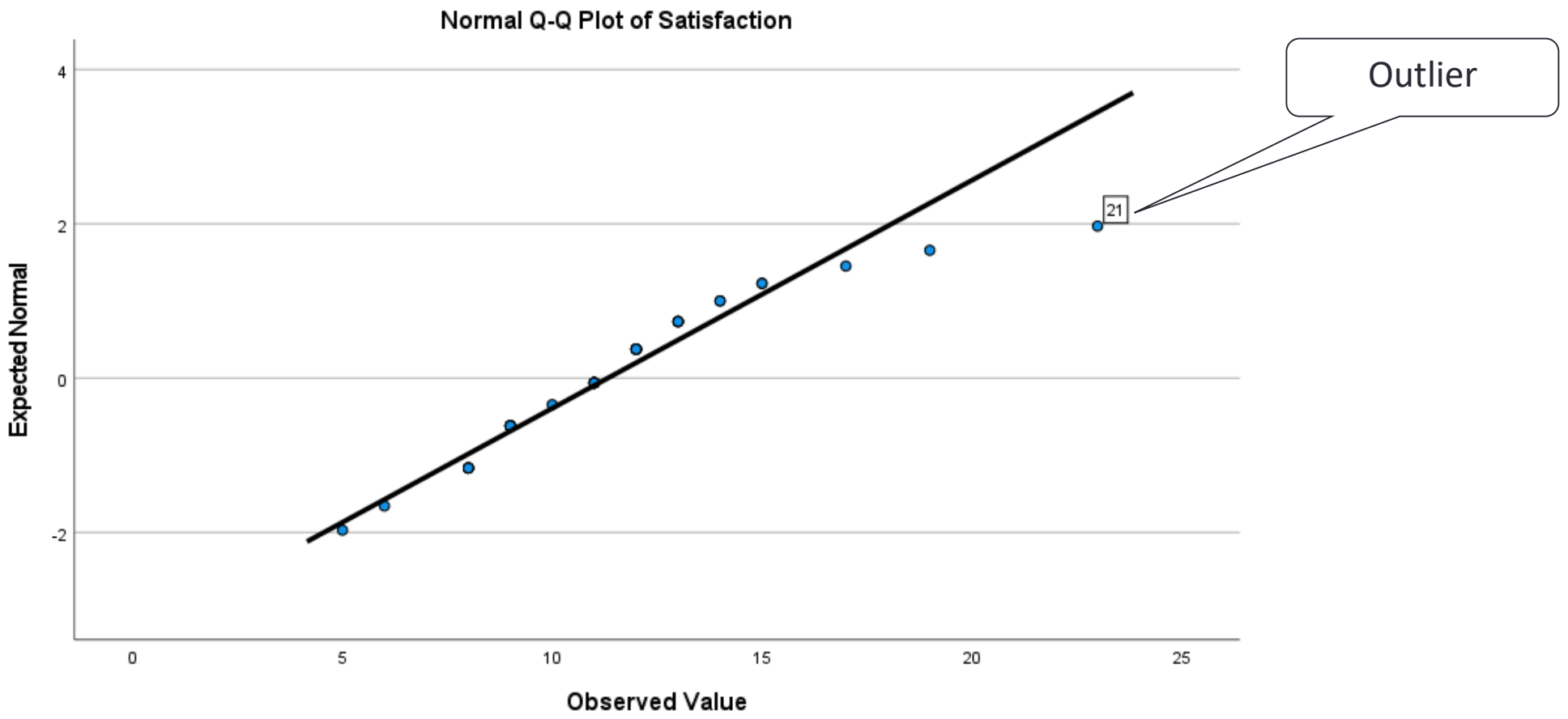
Normal

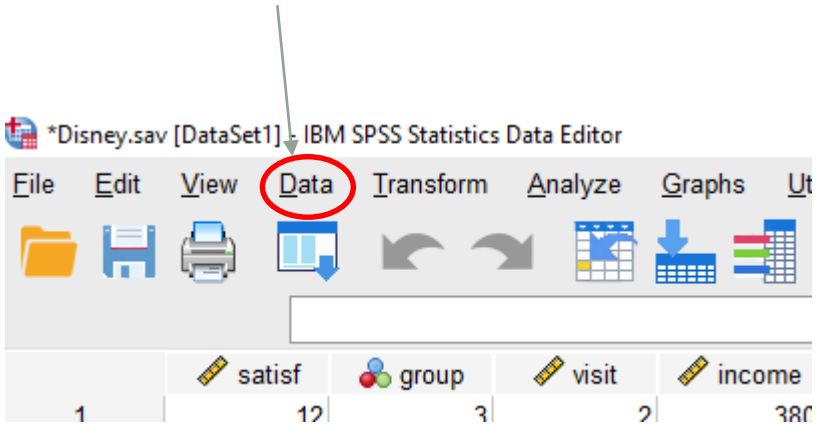
We look here because n < 50.

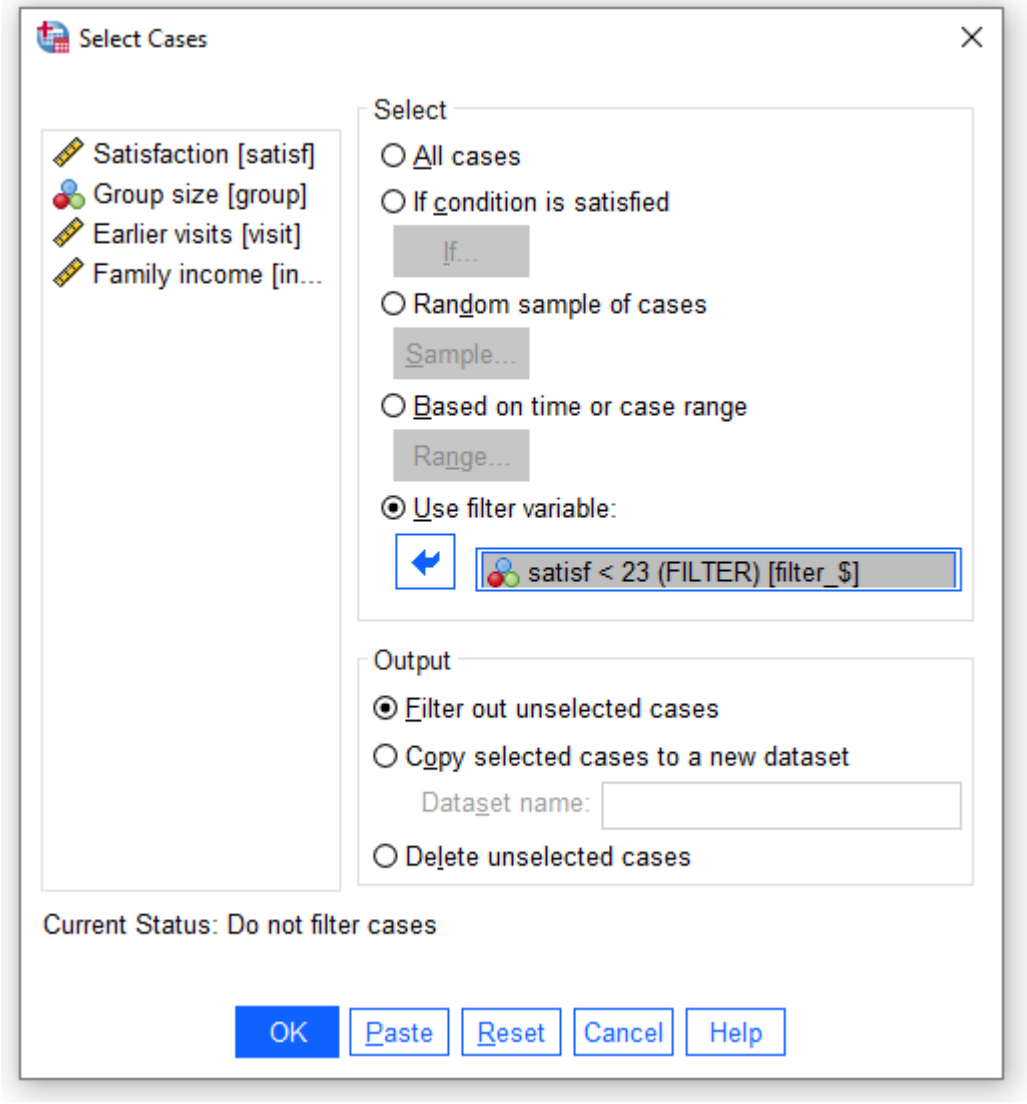
Boxplot


25
20
15
10
5
21
Satisfaction
This is our worst outlier. Should we remove it?
50% of the data
Median

Normal Q-Q Plot




*Disney.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
satisf
group
visit
income


Select Cases
Satisfaction [satisf]
Group size [group]
Earlier visits [visit]
Family income [in...
Select
All cases
If condition is satisfied
If...
Random sample of cases
Sample...
Based on time or case range
Range...
Use filter variable:
satisf < 23 (FILTER) [filter_$]
Output
Filter out unselected cases
Copy selected cases to a new dataset
Dataset name:
Delete unselected cases
Current Status: Do not filter cases
OK
Paste
Reset
Cancel
Help

Without the Outlier (21)

Now both are normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Satisfaction	,121	39	,158	,965	39	,267
Family income	,110	39	,200*	,972	39	,430

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

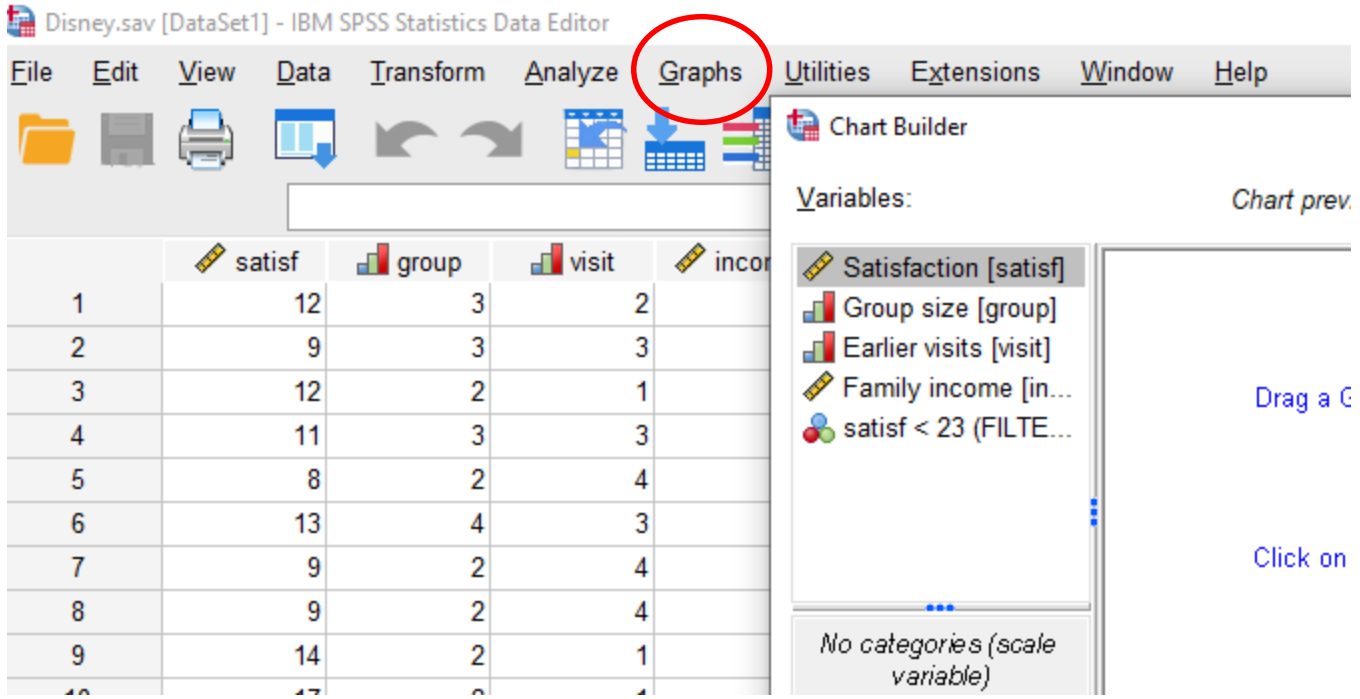
Rules of Thumb!

- Consider which technique you are using before worrying too much about normality.
- Remove outliers.
- Consider transforming the data.
- With larger sample sizes normality is less of an issue.

2. Relationships Between Variables

In, for example, regression we want strong linear relationships between the independent variables and the dependent variable, but weak linear relationships between independent variables.

- Examine scatterplots.
- Examine bivariate correlations.


Disney.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Extensions
Window
Help
Chart Builder
Variables:
Chart prev
Satisfaction [satisf]
Group size [group]
Earlier visits [visit]
Family income [in...
satisf < 23 (FILTE...
Drag a G
Click on
No categories (scale variable)
satisf
group
visit
1 12 3 2
2 9 3 3
3 12 2 1
4 11 3 3
5 8 2 4
6 13 4 3
7 9 2 4
8 9 2 4
9 14 2 1

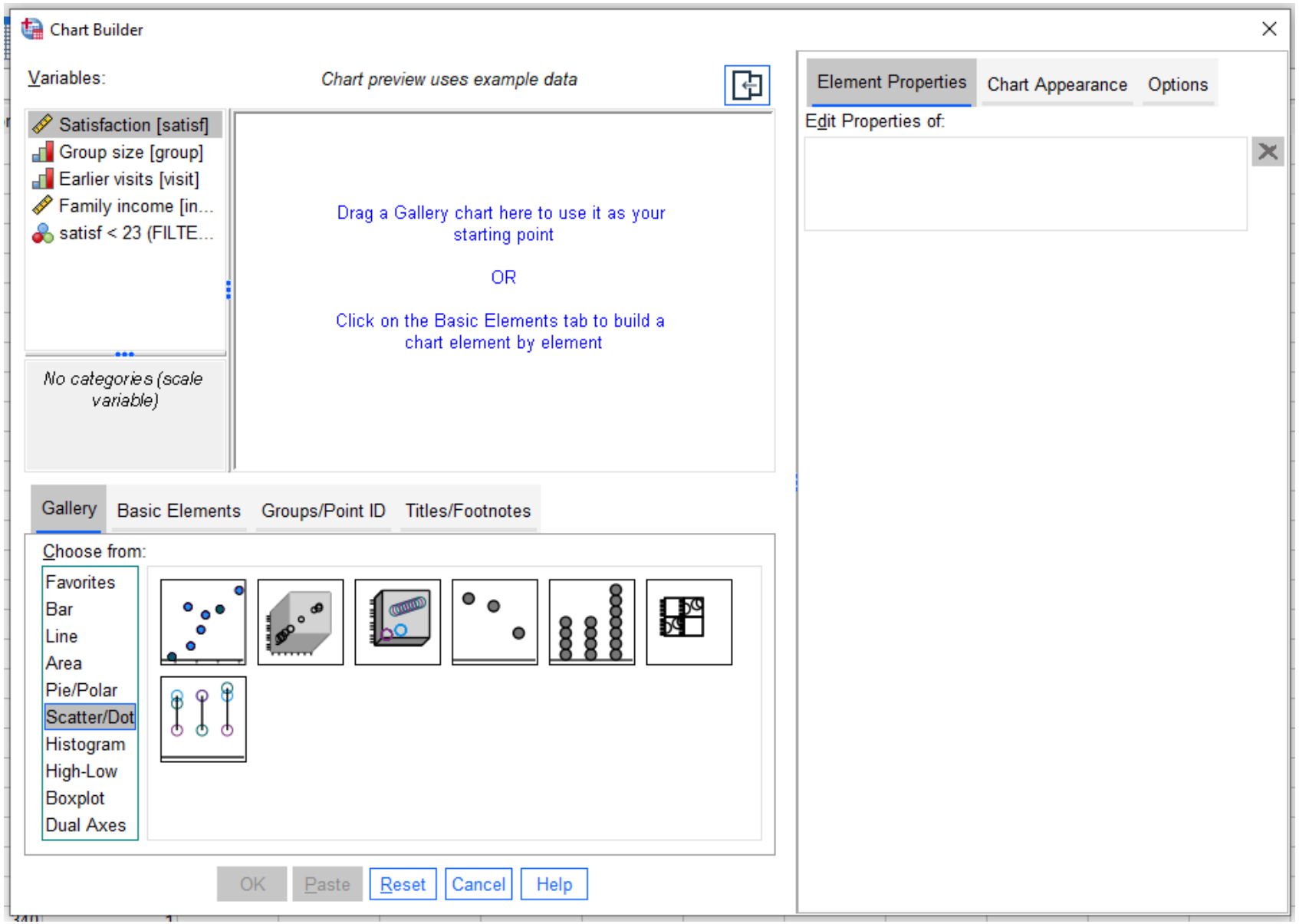


Chart Builder
Variables:
Chart preview uses example data
Satisfaction [satisf]
Group size [group]
Earlier visits [visit]
Family income [in...
satisf < 23 (FILTE...
Drag a Gallery chart here to use it as your starting point
OR
Click on the Basic Elements tab to build a chart element by element
No categories (scale variable)
Element Properties
Chart Appearance
Options
Edit Properties of:
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
OK
Paste
Reset
Cancel
Help

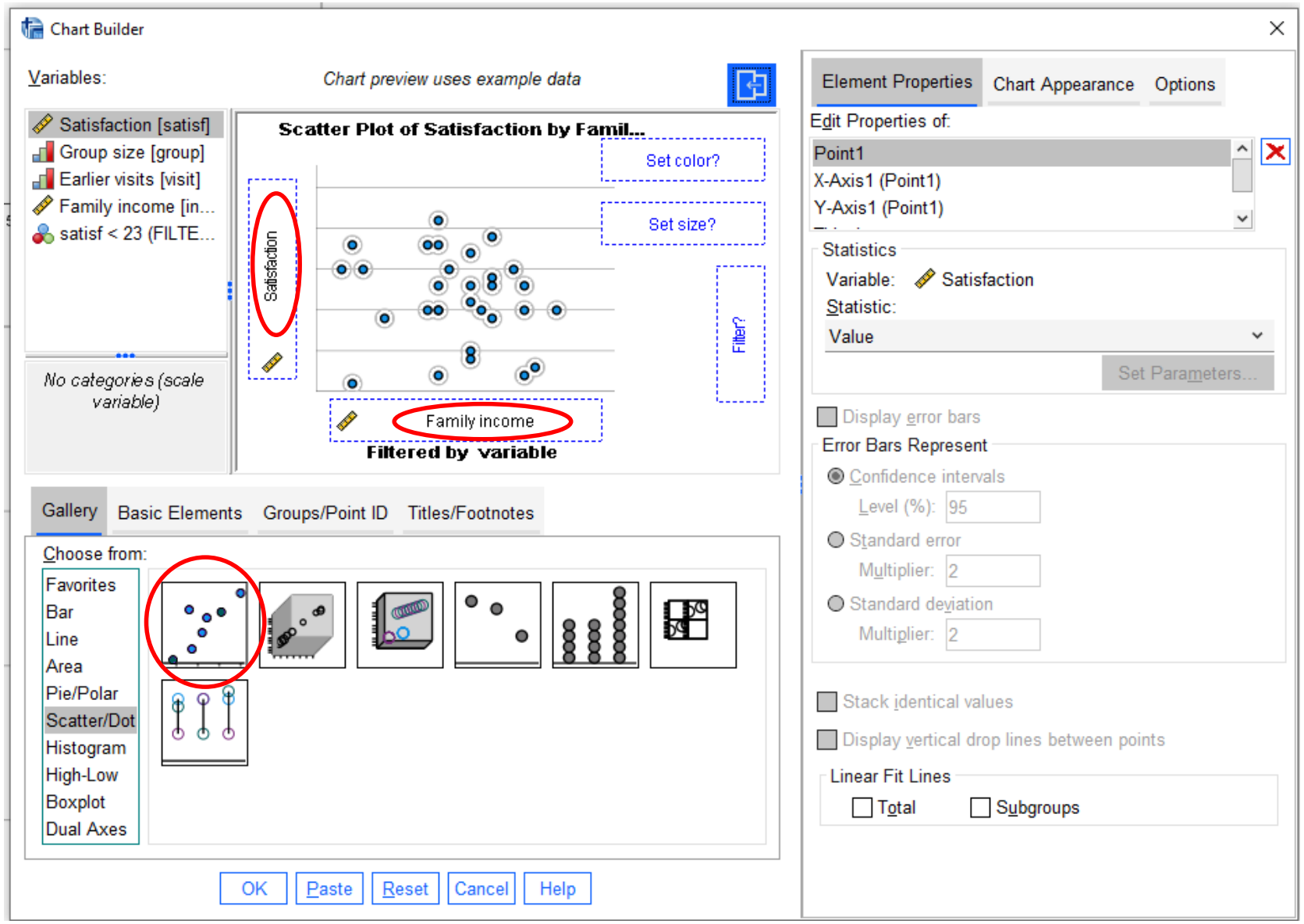


Chart Builder
Variables:
Chart preview uses example data
Satisfaction [satisf]
Group size [group]
Earlier visits [visit]
Family income [in...
satisf < 23 (FILTE...
No categories (scale variable)
Scatter Plot of Satisfaction by Famil...
Set color?
Set size?
Filter?
Satisfaction
Family income
Filtered by variable
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
OK
Paste
Reset
Cancel
Help
Element Properties
Chart Appearance
Options
Edit Properties of:
Point1
X-Axis1 (Point1)
Y-Axis1 (Point1)
Statistics
Variable: Satisfaction
Statistic:
Value
Set Parameters...
Display error bars
Error Bars Represent
Confidence intervals
Level (%): 95
Standard error
Multiplier: 2
Standard deviation
Multiplier: 2
Stack identical values
Display vertical drop lines between points
Linear Fit Lines
Total
Subgroups

Scatterplot

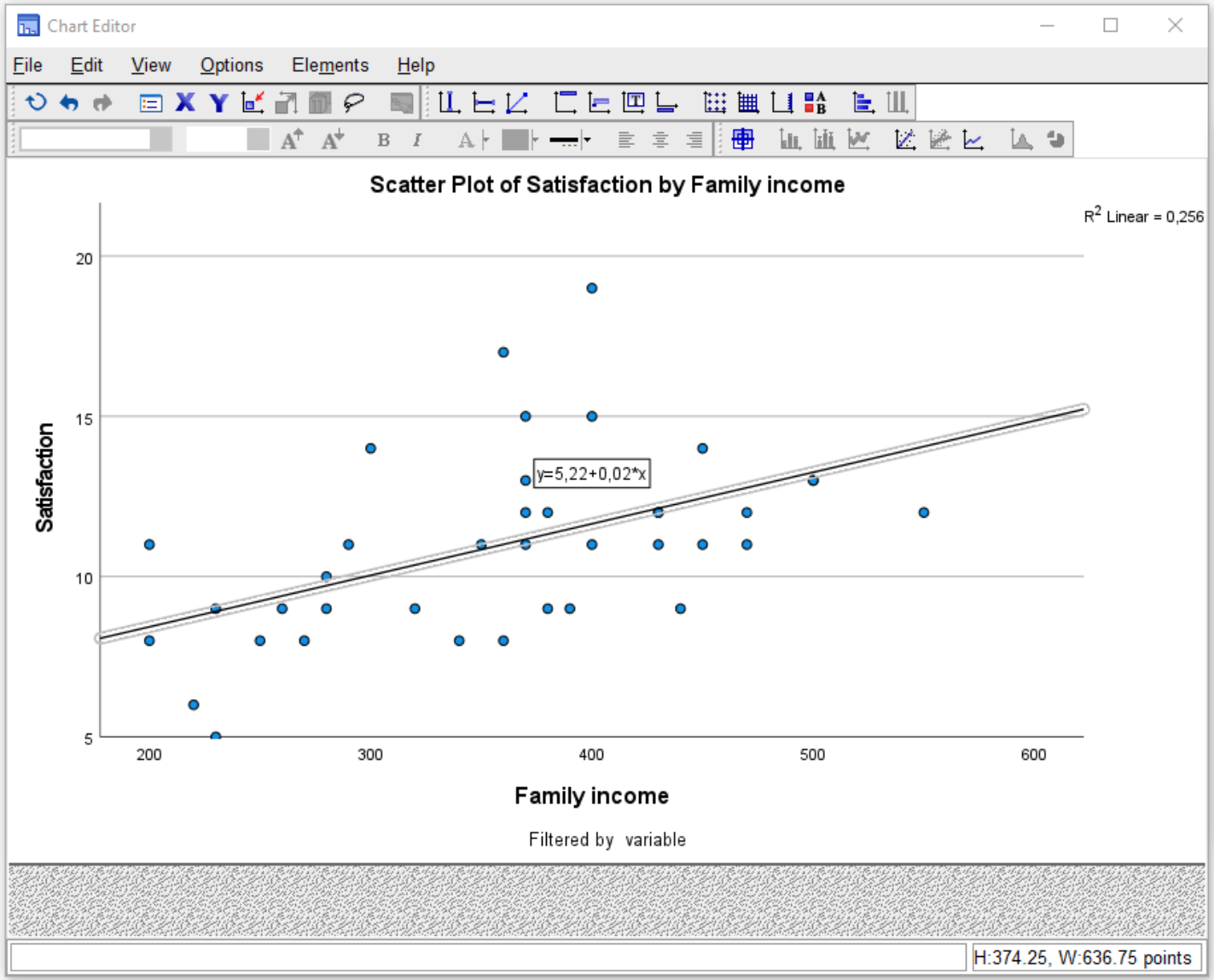
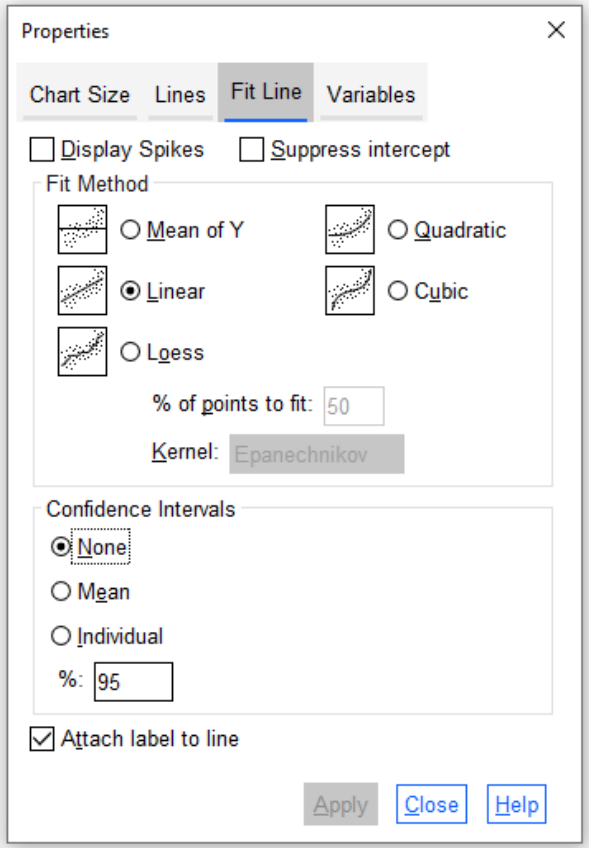
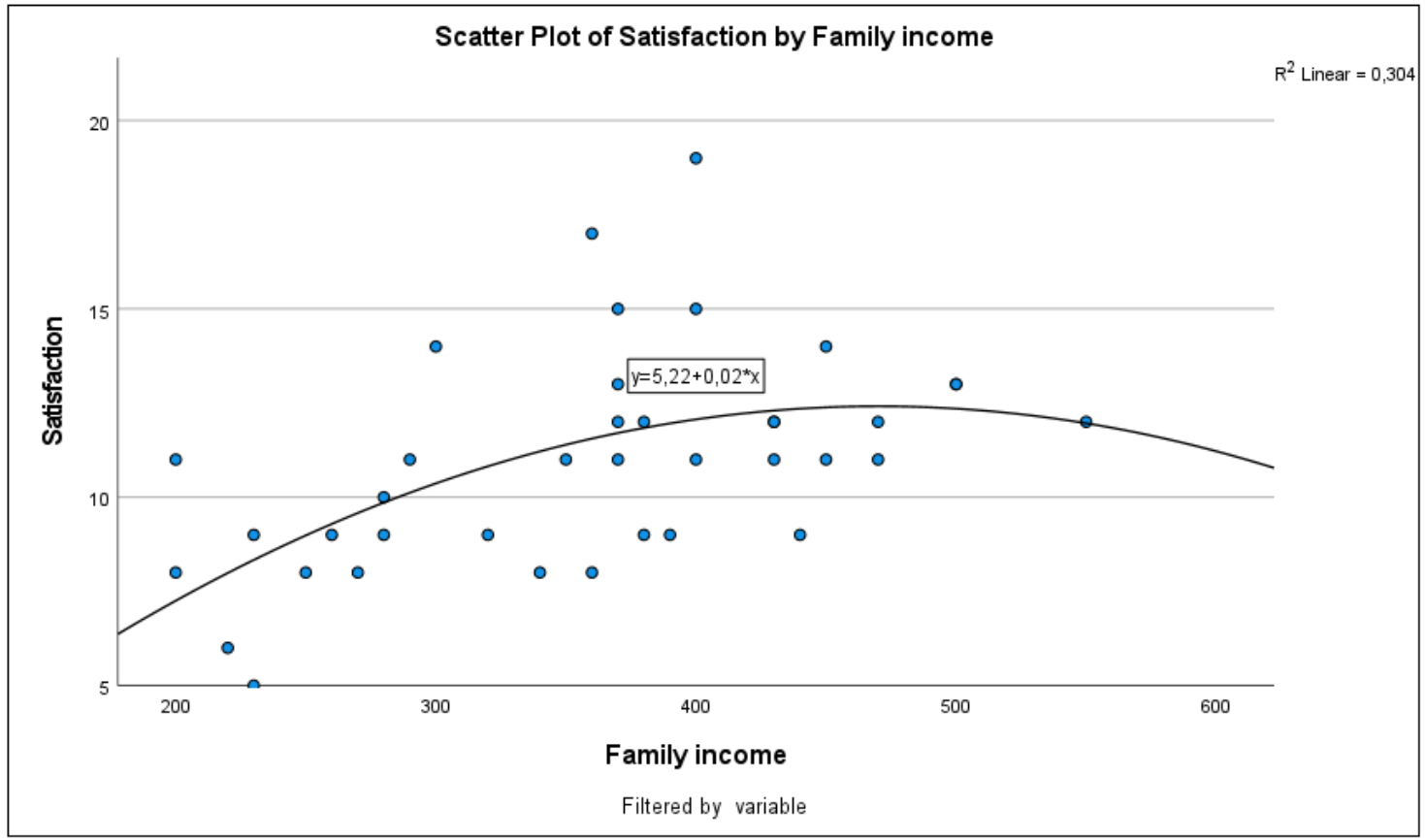
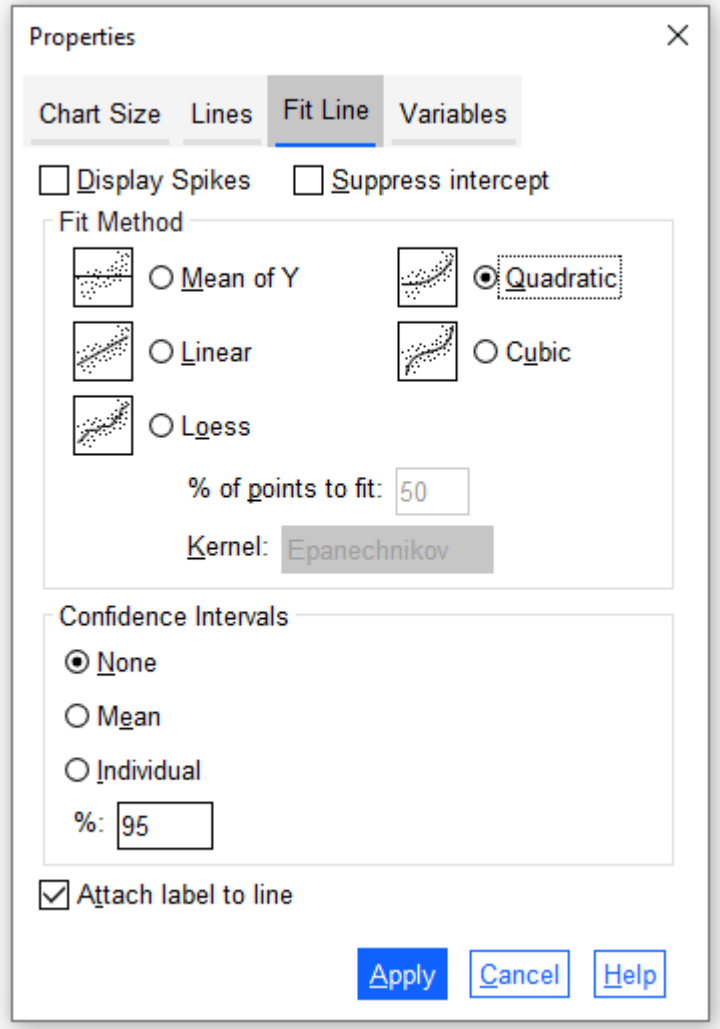


Chart Editor
File Edit View Options Elements Help
Scatter Plot of Satisfaction by Family income
R^2 Linear = 0,256
y=5,22+0,02*x
Satisfaction
Family income
Filtered by variable
H:374.25, W:636.75 points


Properties
Chart Size
Lines
Fit Line
Variables
Display Spikes
Suppress intercept
Fit Method
Mean of Y
Quadratic
Linear
Cubic
Loess
% of points to fit: 50
Kernel: Epanechnikov
Confidence Intervals
None
Mean
Individual
%: 95
Attach label to line
Apply
Close
Help


Scatter Plot of Satisfaction by Family income
R² Linear = 0,304
Satisfaction
20
15
10
5
y=5,22+0,02*x
200
300
400
500
600
Family income
Filtered by variable


Properties
Chart Size
Lines
Fit Line
Variables
Display Spikes
Suppress intercept
Fit Method
Mean of Y
Quadratic
Linear
Cubic
Loess
% of points to fit: 50
Kernel: Epanechnikov
Confidence Intervals
None
Mean
Individual
%: 95
Attach label to line
Apply
Cancel
Help

The thick line represents the range of the true data. The thin line shows values beyond the data.

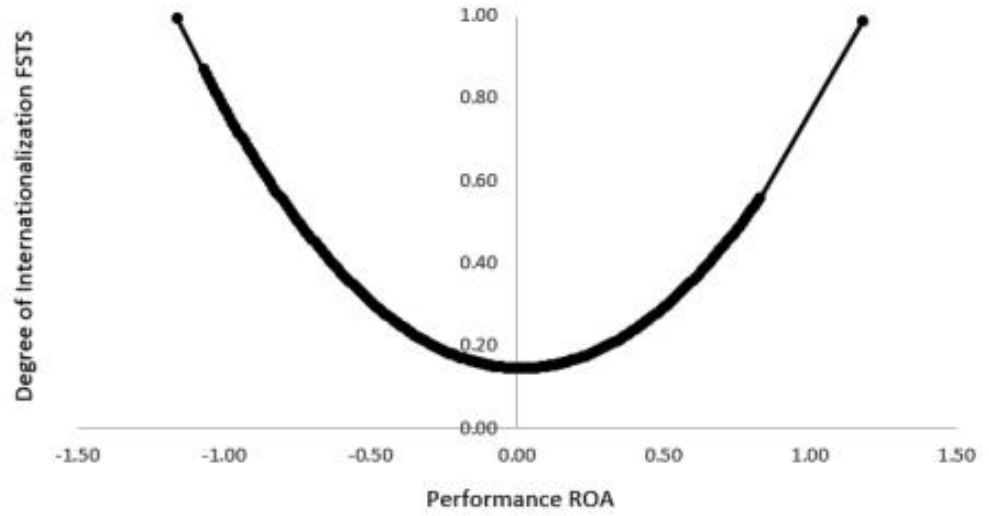
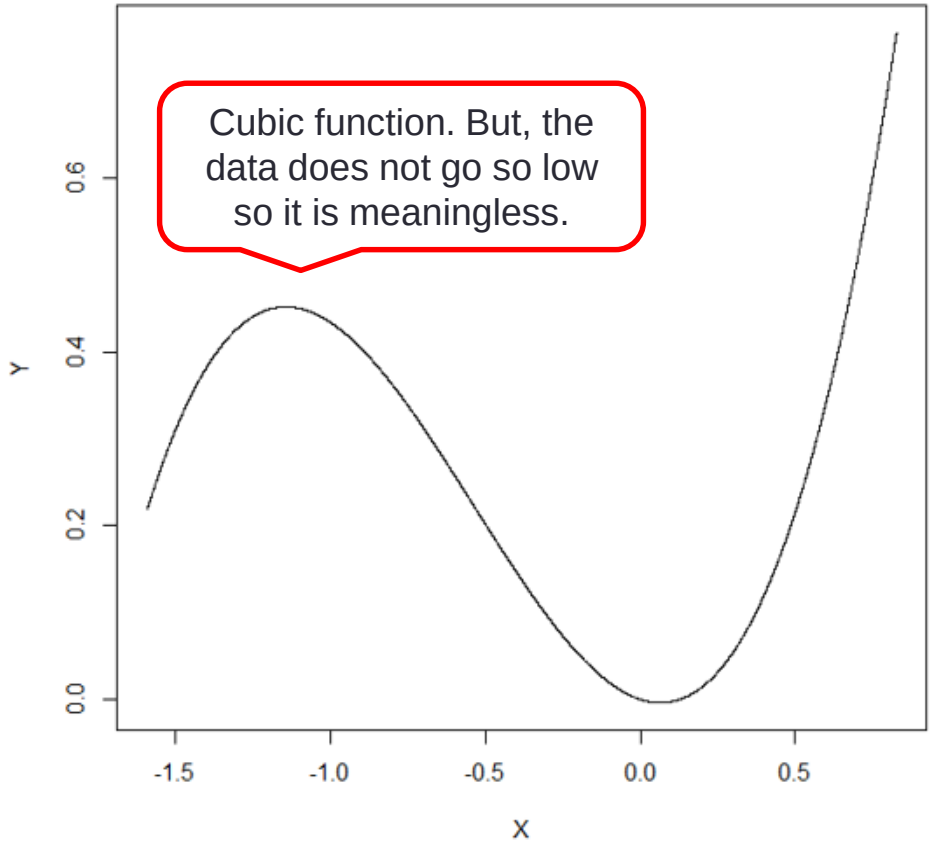
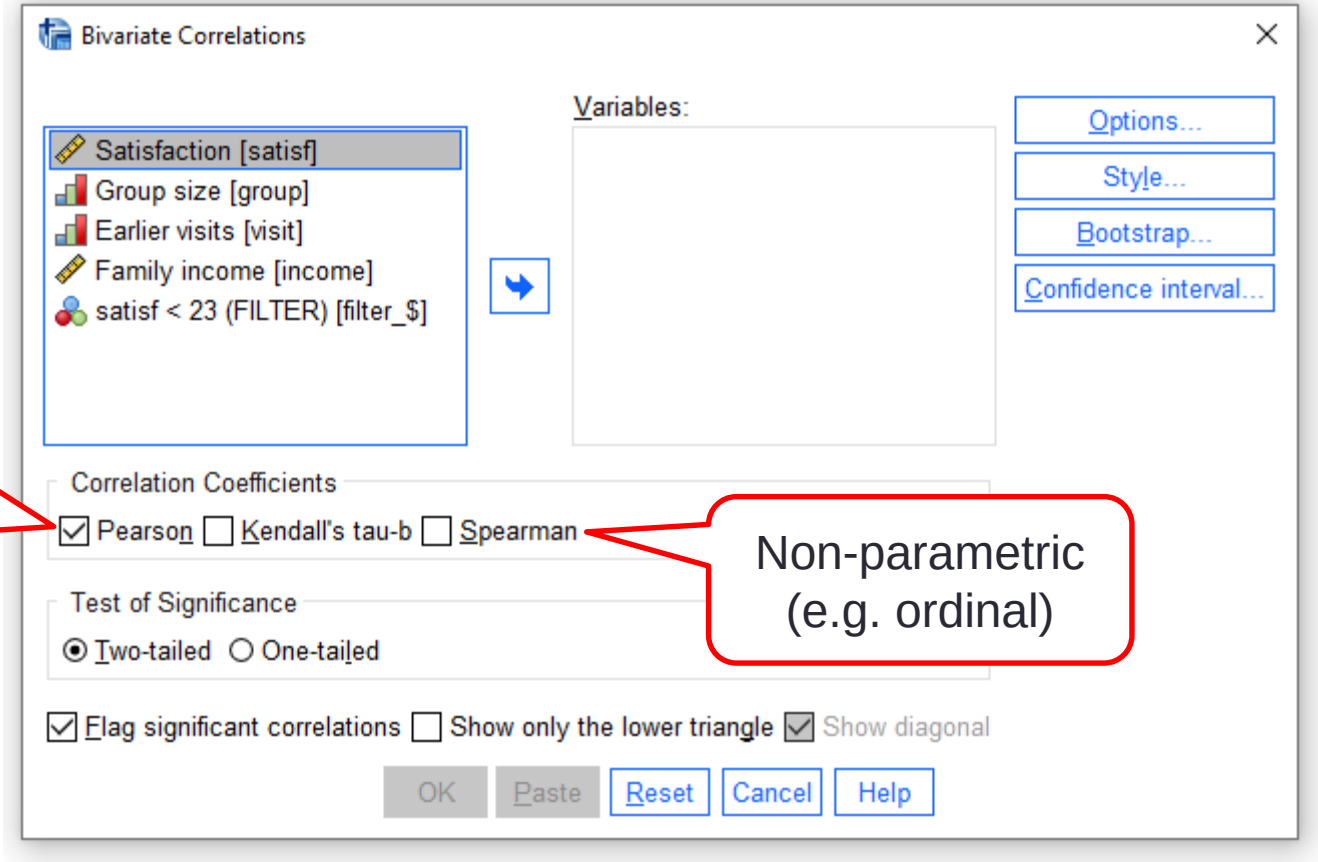


Fig. 1 Regression plot for predicted degree of internationalization

Cubic function. But, the data does not go so low so it is meaningless.



Parametric:
When people just say, “correlation”, they mean Pearson correlation



Non-parametric (e.g. ordinal)

Correlations

Correlations

		Satisfaction	Group size	Earlier visits	Family income
Satisfaction	Pearson Correlation	1	,282	-,573**	,506**
	Sig. (2-tailed)		,081	<,001	,001
	N	39	39	39	39
Group size	Pearson Correlation	,282	1	-,132	,764**
	Sig. (2-tailed)	,081		,423	<,001
	N	39	39	39	39
Earlier visits	Pearson Correlation	-,573**	-,132	1	-,345*
	Sig. (2-tailed)	<,001	,423		,031
	N	39	39	39	39
Family income	Pearson Correlation	,506**	,764**	-,345*	1
	Sig. (2-tailed)	,001	<,001	,031	
	N	39	39	39	39

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations – when considering regression

Correlations

		Y Satisfaction	Group size	# of earlier visits	Family income
Satisfaction	Pearson Correlation				
	Sig. (2-tailed)				
	N	39			
Group size X_1	Pearson Correlation	.282			
	Sig. (2-tailed)	.081			
	N	39			
# of earlier visits X_2	Pearson Correlation	-.573 **	-.132		
	Sig. (2-tailed)	.000	.423		
	N	39	39		
Family income X_3	Pearson Correlation	.506 **	.764 **	-.345 *	
	Sig. (2-tailed)	.001	.000	.031	
	N	39	39	39	

Want high correlations

Want low correlations (under I. 9I)

Rule of Thumb!

- Correlations between independent variables should not exceed 0.9. If they do you most likely have problems with multicolliniarity.

- Consider the substantive meaning of correlations. Small correlations, even if they are significant, are meaningless. By small I mean below I .3 I.

- Cohen (1988) established practical guidelines for interpreting statistically significant correlations.

Table 8.14 Correlation strength

Correlation strength	Absolute value of correlation coefficient
Small	0.10–0.29
Medium	0.30–0.49
Large	0.50–1.00

3. Outliers

- There is no rule of thumb. You simply have to look at them and decide if they are an aberration or if they are a natural part of the data.

- Examine how they influence your results.

Article

Organizational Research Methods
16(2) 270-301
© The Author(s) 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1094428112470848
orm.sagepub.com

SAGE

Best-Practice Recommendations for Defining, Identifying, and Handling Outliers

Herman Aguinis[1], Ryan K. Gottfredson[1], and Harry Joo[1]

4. Missing Data

Two questions:

1. Why is it missing?
   - You need to determine if the missing data is missing at random, or if there is some sort of pattern as to why it is missing.

2. How much is missing?
   - You have to consider how much data is missing. If not so much, just run your analyses with the default ”listwise deletion”.
   - Pallant often suggests pairwise deletion – this way you don’t lose so much data.
   - If a lot is missing, you may opt to replace missing values.
   - You may drop variables with a lot of missing values.

Four Steps

1. Type of missing data: ignorable – not ignorable?
   - Ignorable if substantively missing at random.
2. Extent of missing data: If not ignorable, is there so much missing as to bias results?
   - 10% cutoff for no problem.
3. Diagnose randomness (many months & large brain).
4. What to do?
   - Listwise deletion (complete case approach).
   - Pairwise deletion.
   - Imputation.

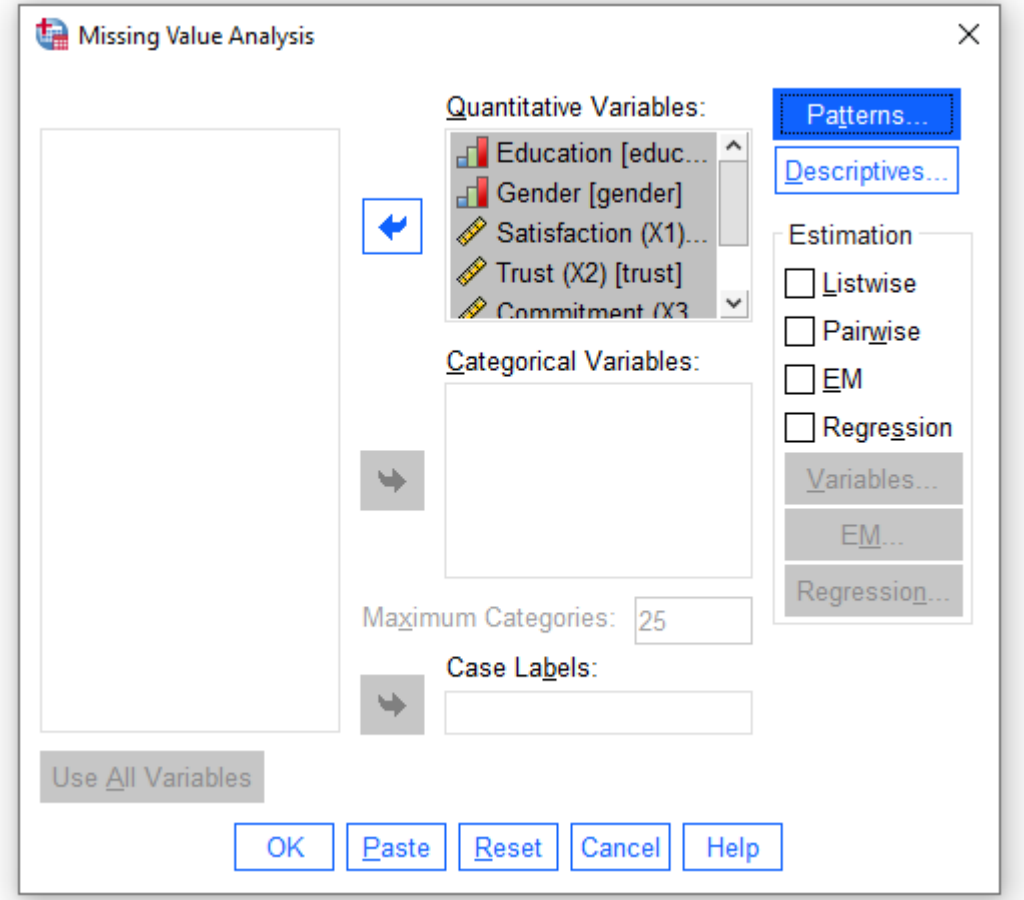
Rule of Thumb!

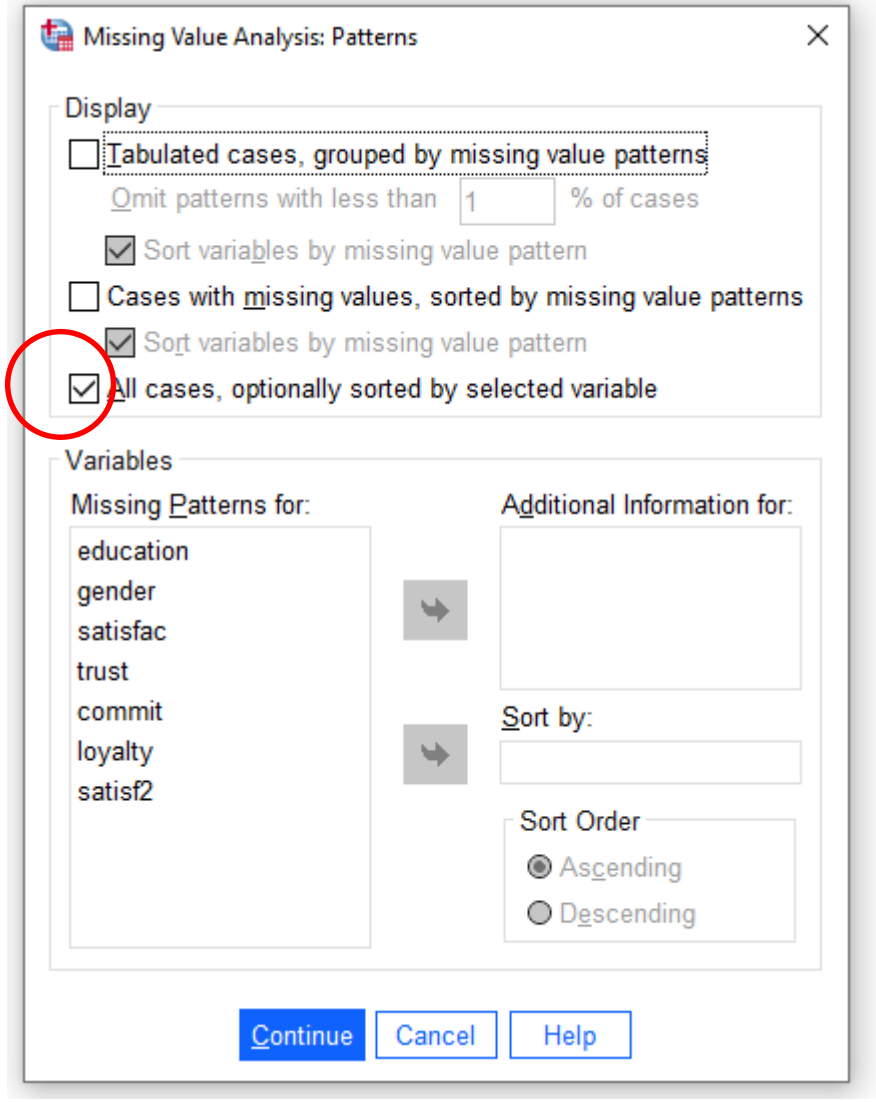
In SPSS you can test for differences between missing and non-missing data across all variables.

If there are significant patterns – hmm.

- Try to avoid missing data.
- If the % of missing values is low, ignore the patterns.
  - Around 10%.
- Are they variables you really need?
- Are the missing values concentrated in a few cases? Consider deleting them.

Use the “Banking” data





Missing data

Delete a bad variable?

Delete a bad case?

Data Patterns (all cases)

Case	# Missing	% Missing	Missing and Extreme Value Patterns: education	gender	satisfac	trust	commit	loyalty	satisf2
1	0	,0							
2	0	,0							
3	0	,0							
4	0	,0							
5	0	,0							
226	1	14,3						A	
227	0	,0							
228	3	42,9				A	A	A	
229	0	,0							
230	0	,0							
231	1	14,3				A			
232	0	,0							
233	0	,0							
234	5	71,4			A	A	A	A	A
235	1	14,3						A	