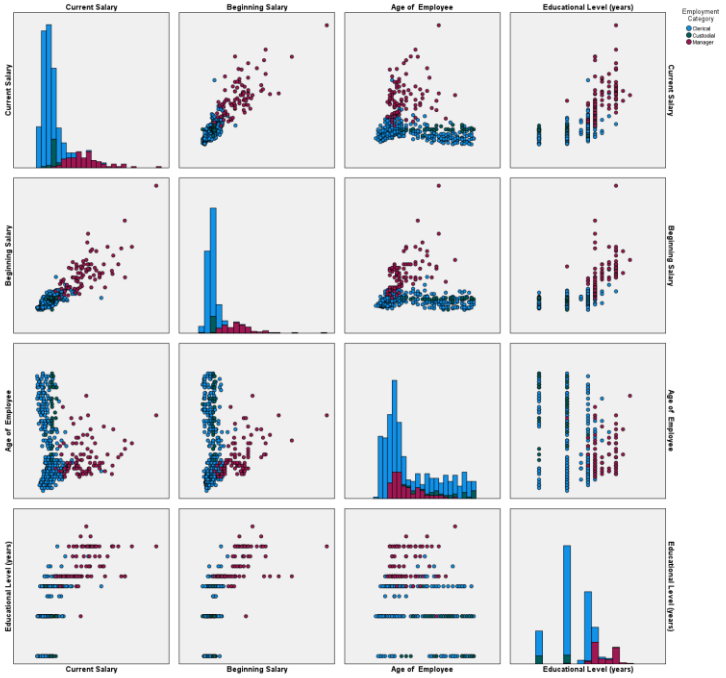
SMARTVISION Europe

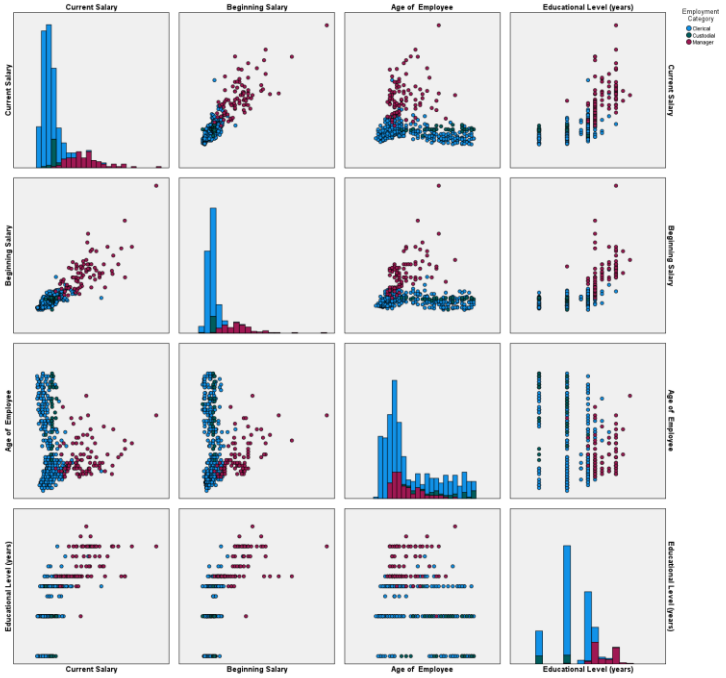


Correlation analysis with SPSS

Jarlath Quinn – Analytics Consultant

SMARTVISION Europe

Just waiting for all attendees to join…

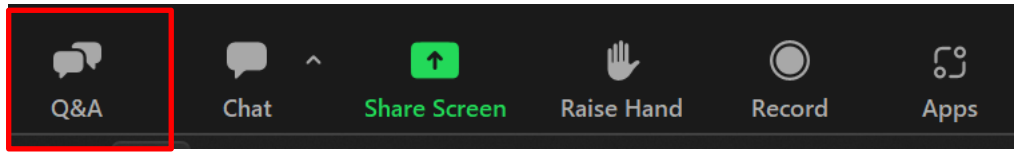


Correlation analysis with SPSS

Jarlath Quinn – Analytics Consultant

FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the Q&A panel – if we run out of time we will follow up with you.



- Premier accredited partner to IBM, Predictive Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry

- Deep experience of applied advanced analytics applications across sectors
  - Retail
  - Gaming
  - Utilities
  - Insurance
  - Telecommunications
  - Media
  - FMCG

Agenda

- Why are correlation values useful?
- Interpreting correlation coefficients
- Estimating correlation values with bootstrapping techniques
- How correlations are calculated
- Automatically highlighting strong correlations
- Linear vs non-linear relationships
- Non-parametric correlations
- The limitations of correlations

Correlations

		Current Salary
Educational Level (years)	Pearson Correlation	.661**
	Sig. (2-tailed)	<.001
	N	474

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

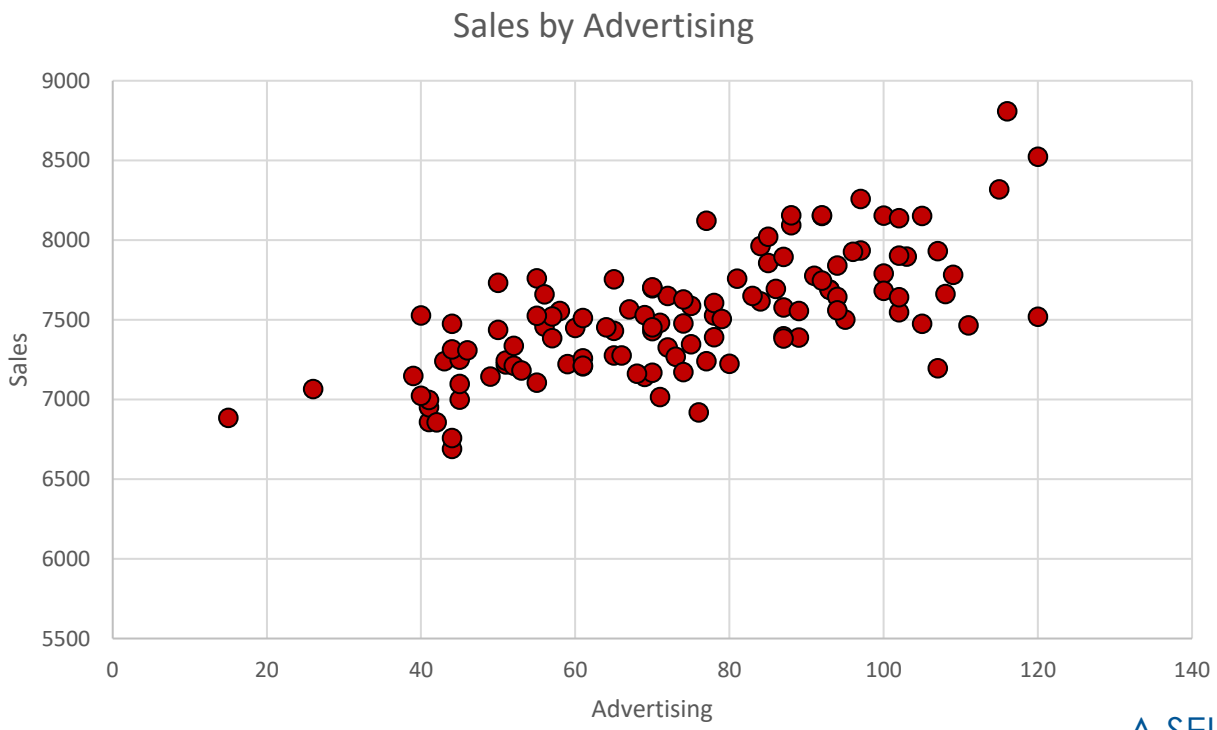
Why are correlations useful?

Why use correlations?

- Correlation is a term that we employ in everyday speech to denote things that *appear* to have some kind of relationship
- In analytics, correlations are specific values that are calculated in order quantify the relationships between variables
- This kind of analysis is powerful because, it allows us to *detect* and *measure* the strength of linear associations between an near infinite range of factors, such as:
  - Advertising spend and website hits
  - Product sales and competitor pricing
  - Vibration and component part failure
  - Rainfall and pollution
  - Study time and examination grade
  - Exercise and weight loss
  - Government spending and population health outcomes

The gateway to prediction

- Not only can we measure a linear relationship with correlation, but we can also use one variable to predict the other
- For example, if we know how much we're planning to increase our spend on advertising then we can use correlation to accurately predict what the increase in visitors to the website is likely to be.



SMARTVISION Europe

Correlations

		Current Salary
Educational Level (years)	Pearson Correlation	.661**
	Sig. (2-tailed)	<.001
	N	474

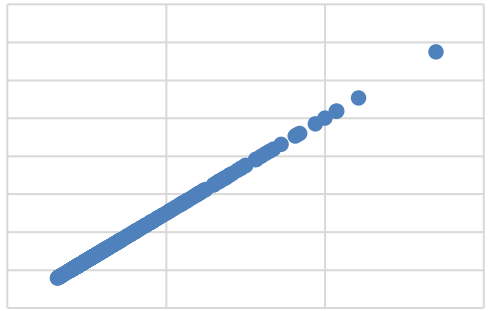
**. Correlation is significant at the 0.01 level (2-tailed).

Interpreting correlations

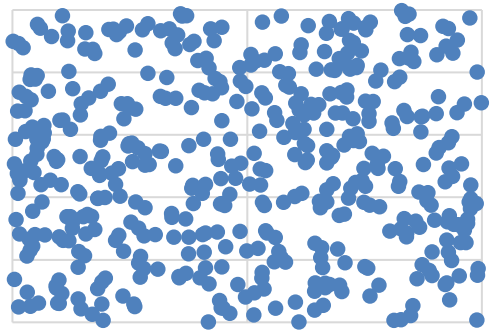
Linear Correlation Scale

+1 +0.5 0 -0.5 -1

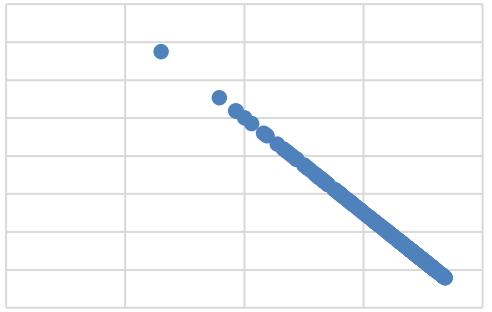
Perfect Positive Linear Relationship



No Linear Relationship



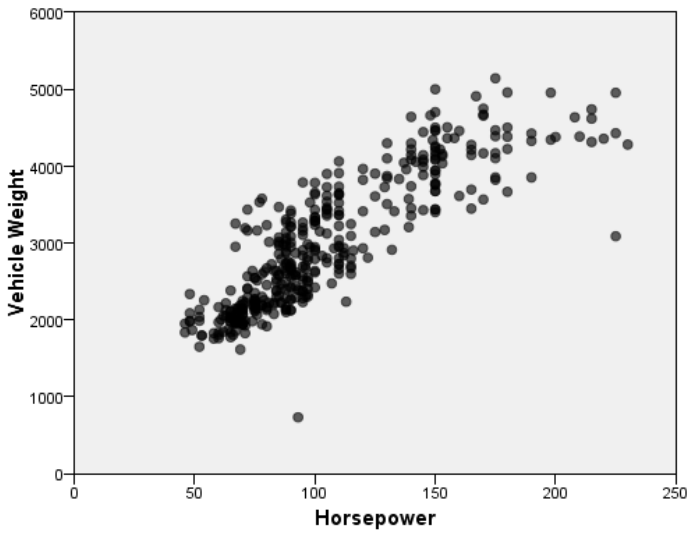
Perfect Negative Linear Relationship

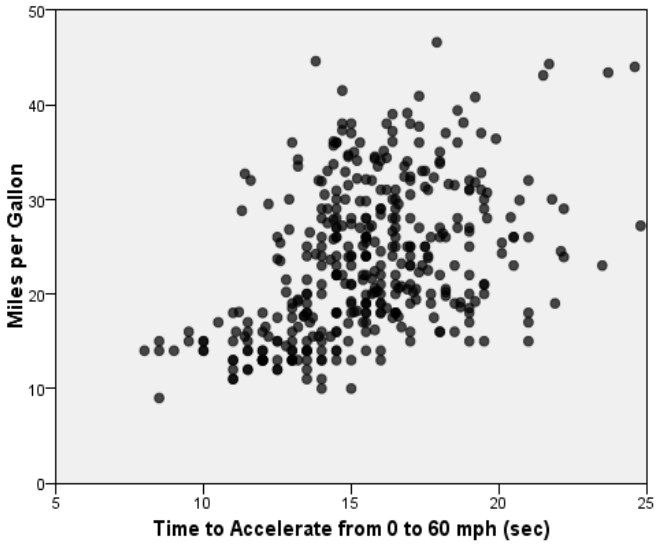


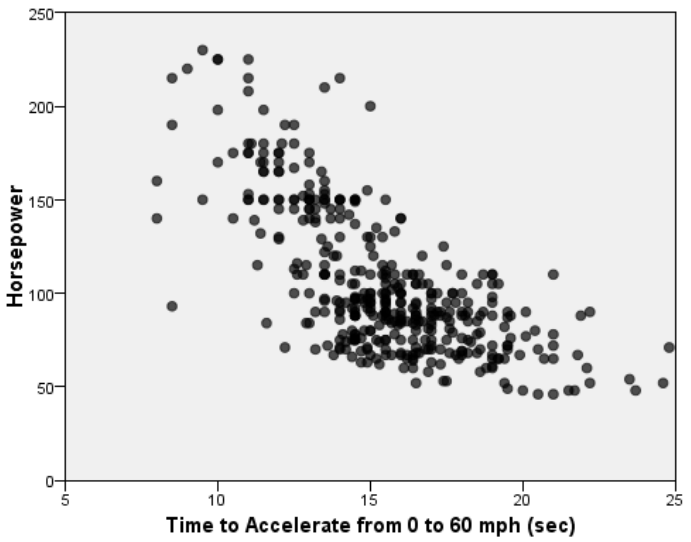
SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Pearson's *r* correlations







0.859

0.434

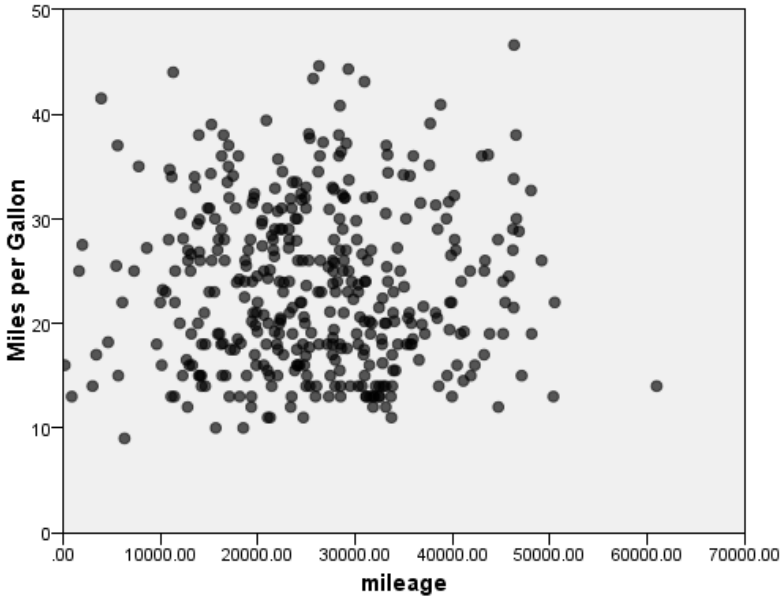
-.701

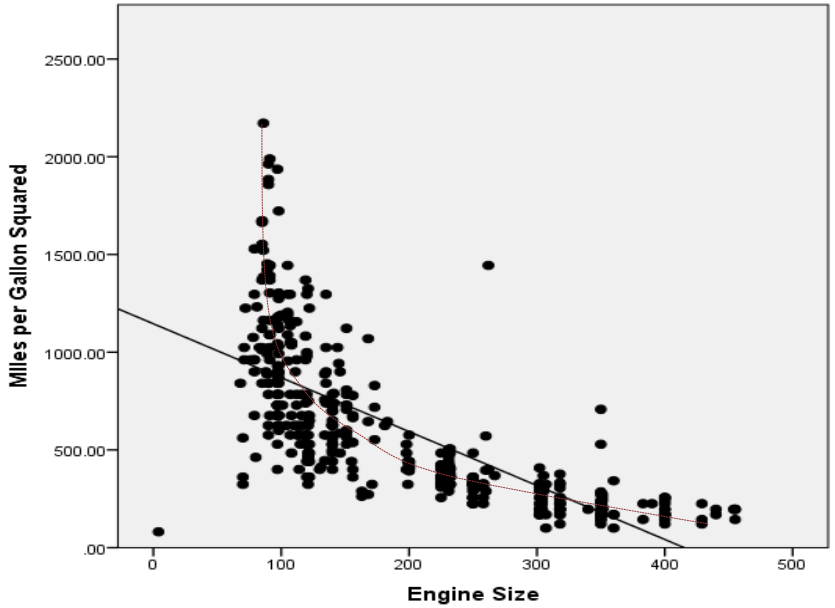
Pearson's *r* correlation coefficients

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Non-Linear Relationships





-0.005

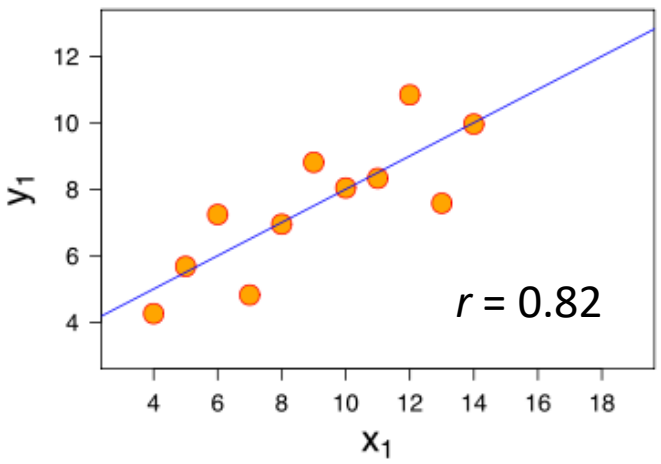
-.671

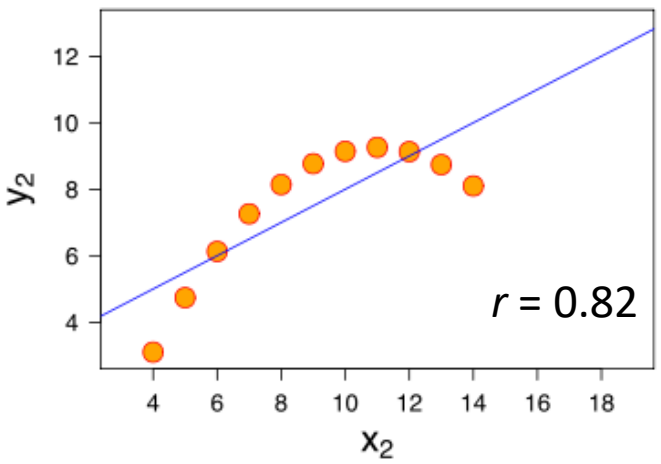
Pearson's *r* Correlations

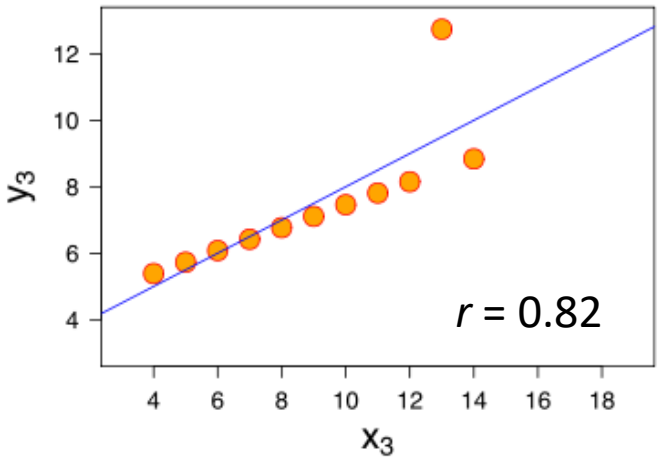
SMARTVISION Europe

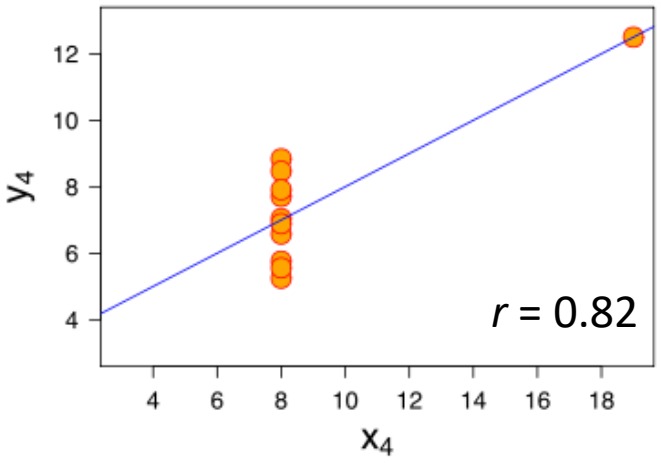
A SELECT INTERNATIONAL COMPANY

A word of warning: always investigate the relationship







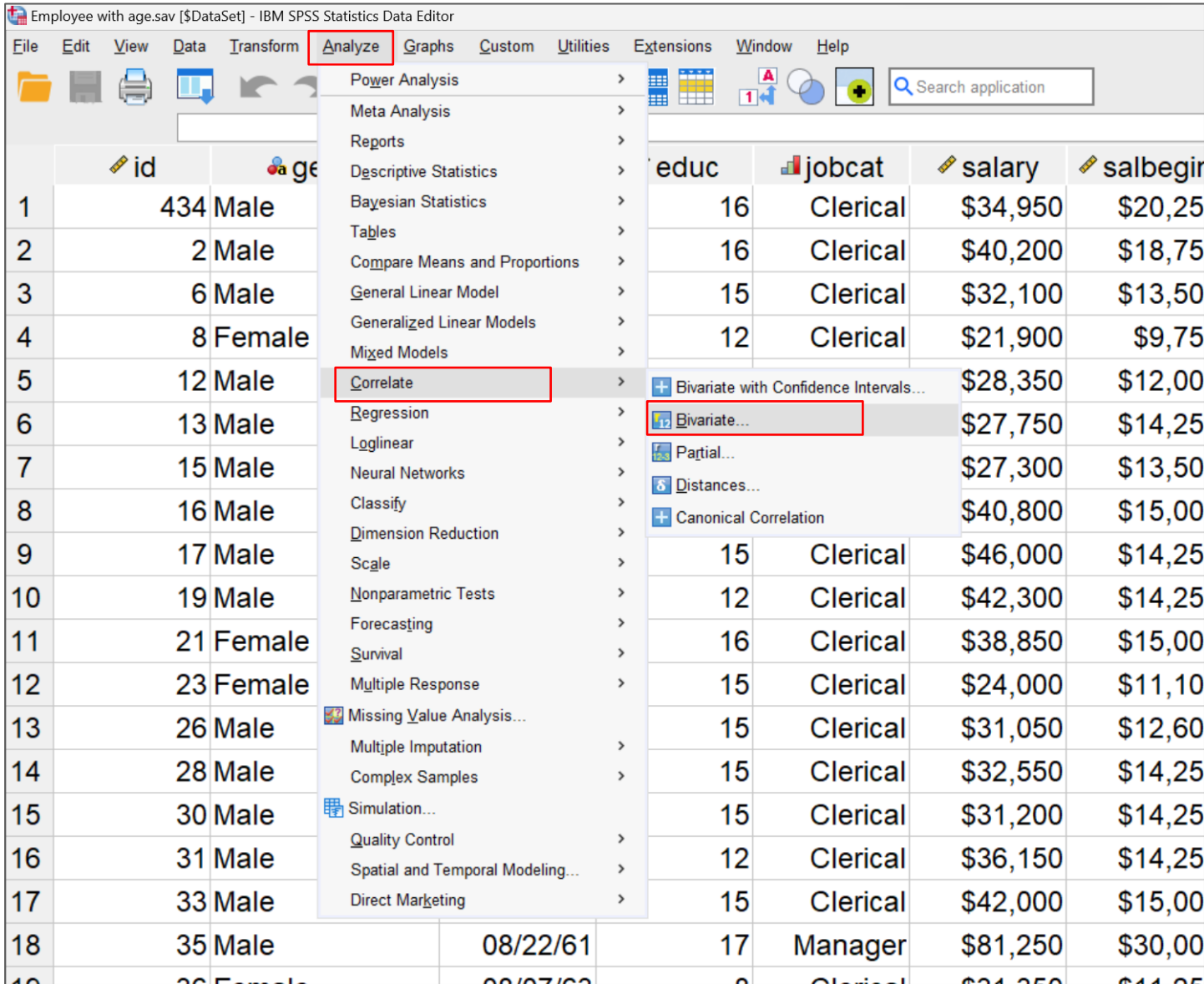


SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Example SPSS Correlations

- Analyze
  - Correlate
    - Bivariate

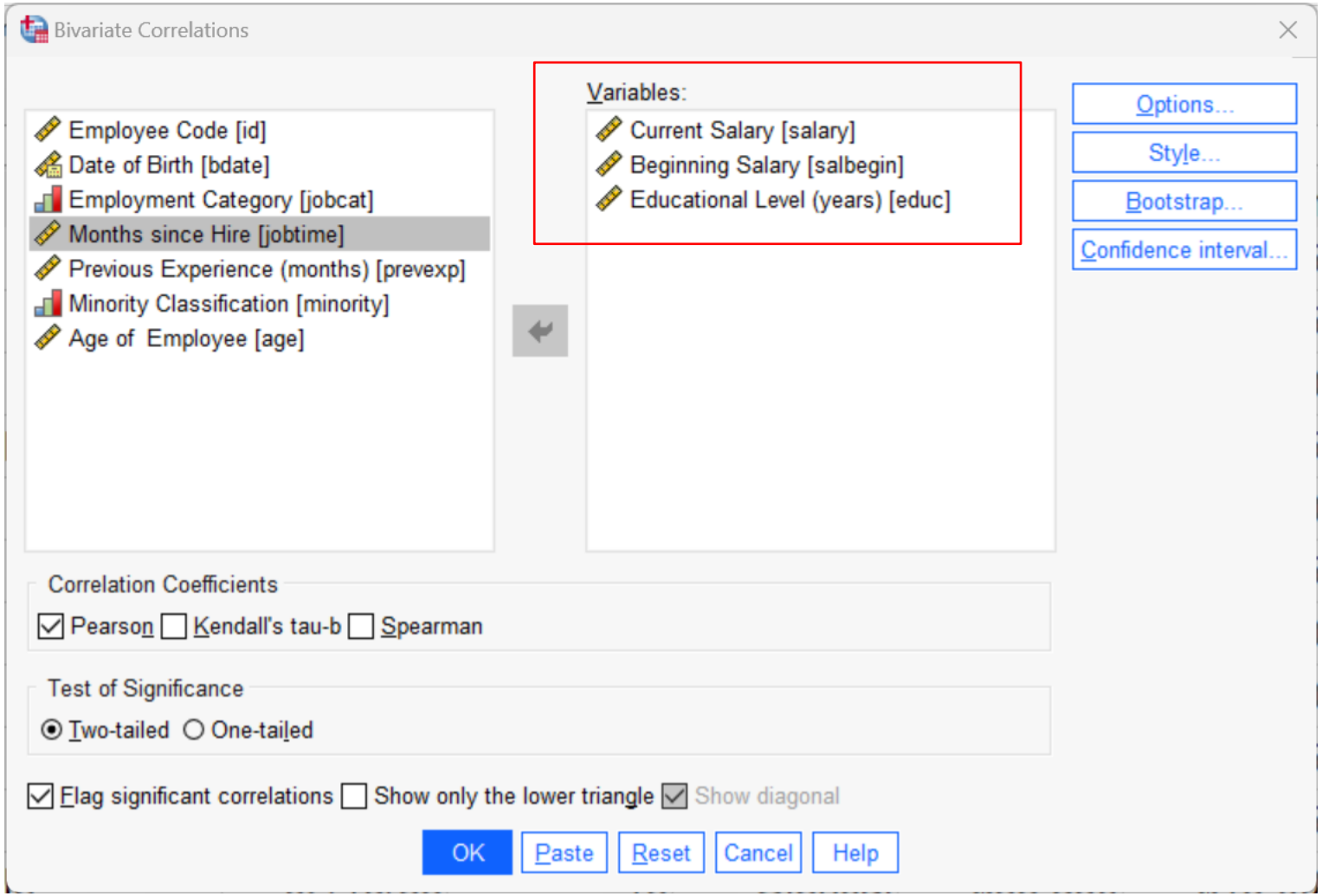


SMARTVISION Europe

Example SPSS Correlations

- Three variables chosen – so three pairs of correlations

1. Current Salary x Beginning Salary
2. Current Salary x Education Level
3. Beginning Salary x Education Level



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Example SPSS Correlations

Correlations

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880**	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

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

Example SPSS Correlations

Correlations

The table is a mirror image

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880**	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

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

Example SPSS Correlations

The diagonal values are all equal to one as they are the variables correlated against themselves

Correlations

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880**	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

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

Example SPSS Correlations

The Significance values show how likely one is to get a correlation like that assuming there's no relationship between the variables

Correlations

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880**	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

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

Example SPSS Correlations

Th N values show how many cases the correlation was based on

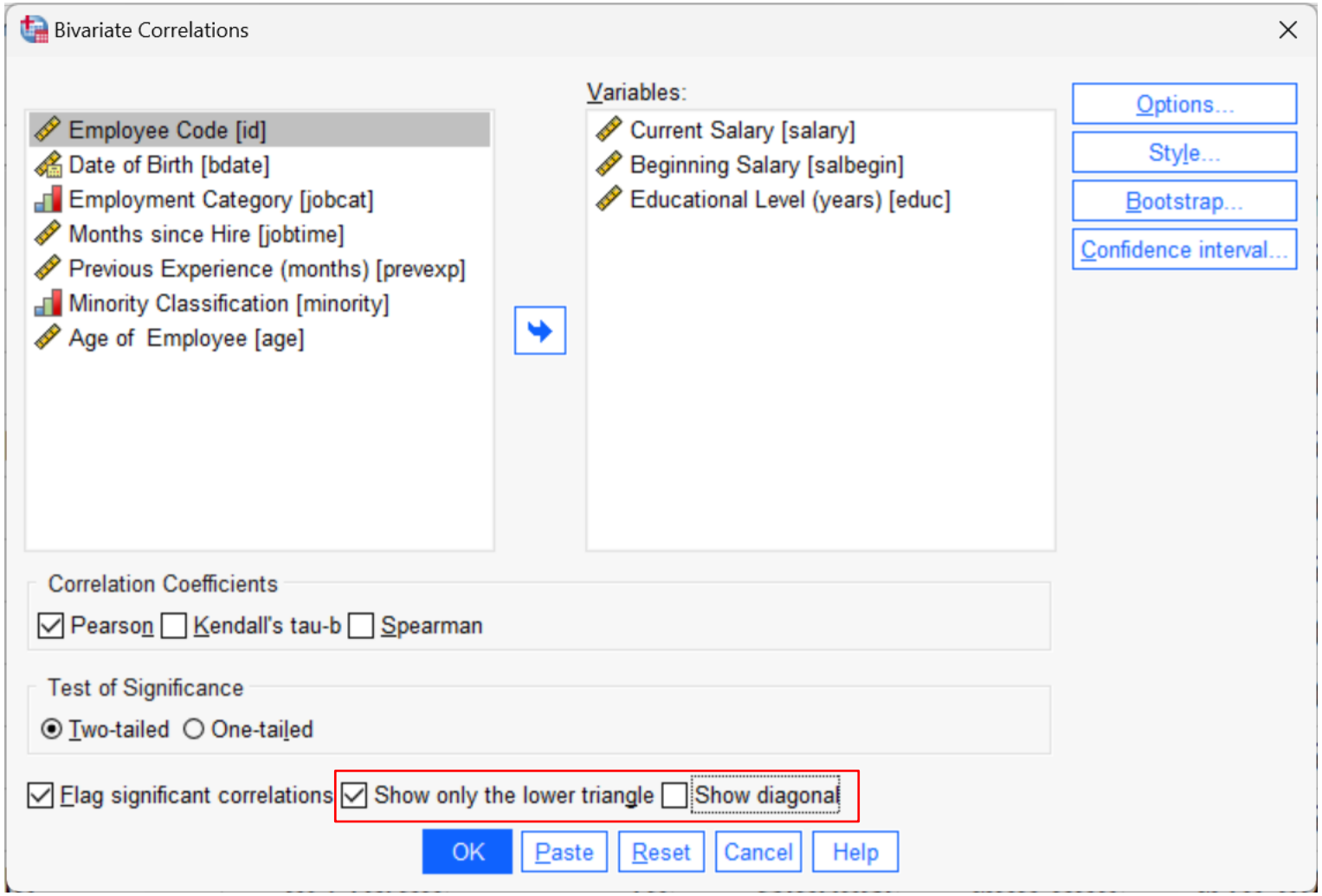
Correlations

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880**	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

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

Example SPSS Correlations

- We can re-run the analysis, but this time….
  - Show only the bottom half of the matrix
  - Don't show the correlations of each variable against itself



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Example SPSS Correlations

Correlations

		Current Salary	Beginning Salary
Beginning Salary	Pearson Correlation	.880**	
	Sig. (2-tailed)	<.001	
	N	474	
Educational Level (years)	Pearson Correlation	.661**	.633**
	Sig. (2-tailed)	<.001	<.001
	N	474	474

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

Let’s explore Pearson’s correlations in SPSS Statistics

SMARTVISION Europe

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

How is a Pearson's correlation calculated?

Pearson's *r* (the most well known correlation measure)

- In statistics, the **Pearson correlation coefficient** is also known as **Pearson's r** or the **Pearson product-moment** correlation coefficient
- Correlations describe data moving together
- This is a **parametric** procedure. That means it makes assumptions about the data. Strictly speaking Pearson's *r* assumes
  - The level of measurement of the variables are continuous/scale (i.e. interval or ratio)
  - There should be no extreme outliers in the correlated variables
  - The data are normally distributed - this is not needed for a reasonable sample size

Formula for Pearson's *r*

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

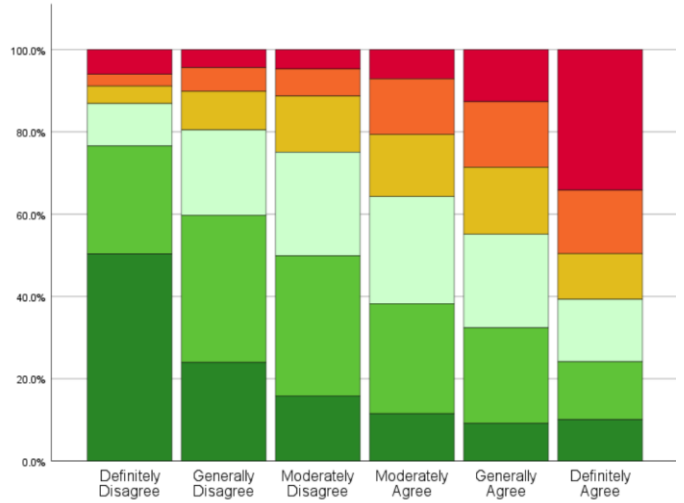
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

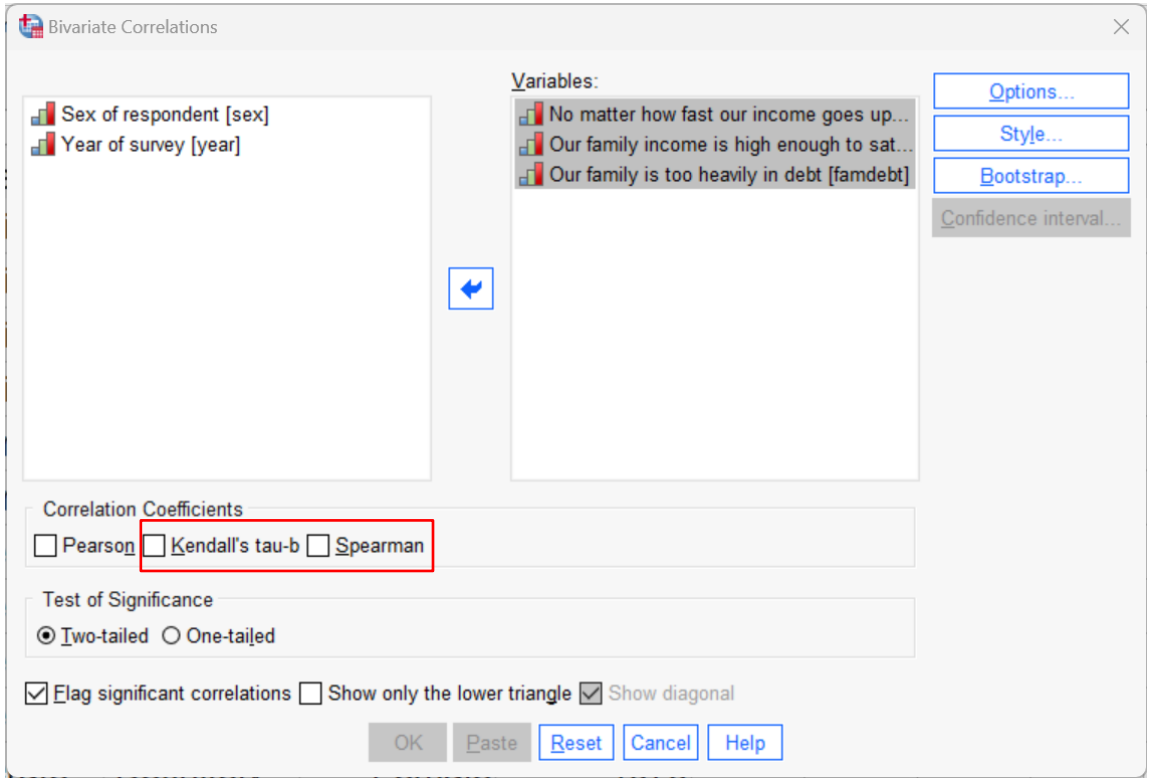
Let's see an example of calculating Pearson's R



Correlations with rank order variables

Correlations for Ordinal Variables

- Rank order or ‘ordinal’ variables refer to variables such as rating scales
- These are not true numbers, but rather ranked ‘numerals’
- Two techniques exist in SPSS that deal with these kind of data
  - Spearman’s Rho
  - Kendall’s Tau-b
- Both are non-parametric methods



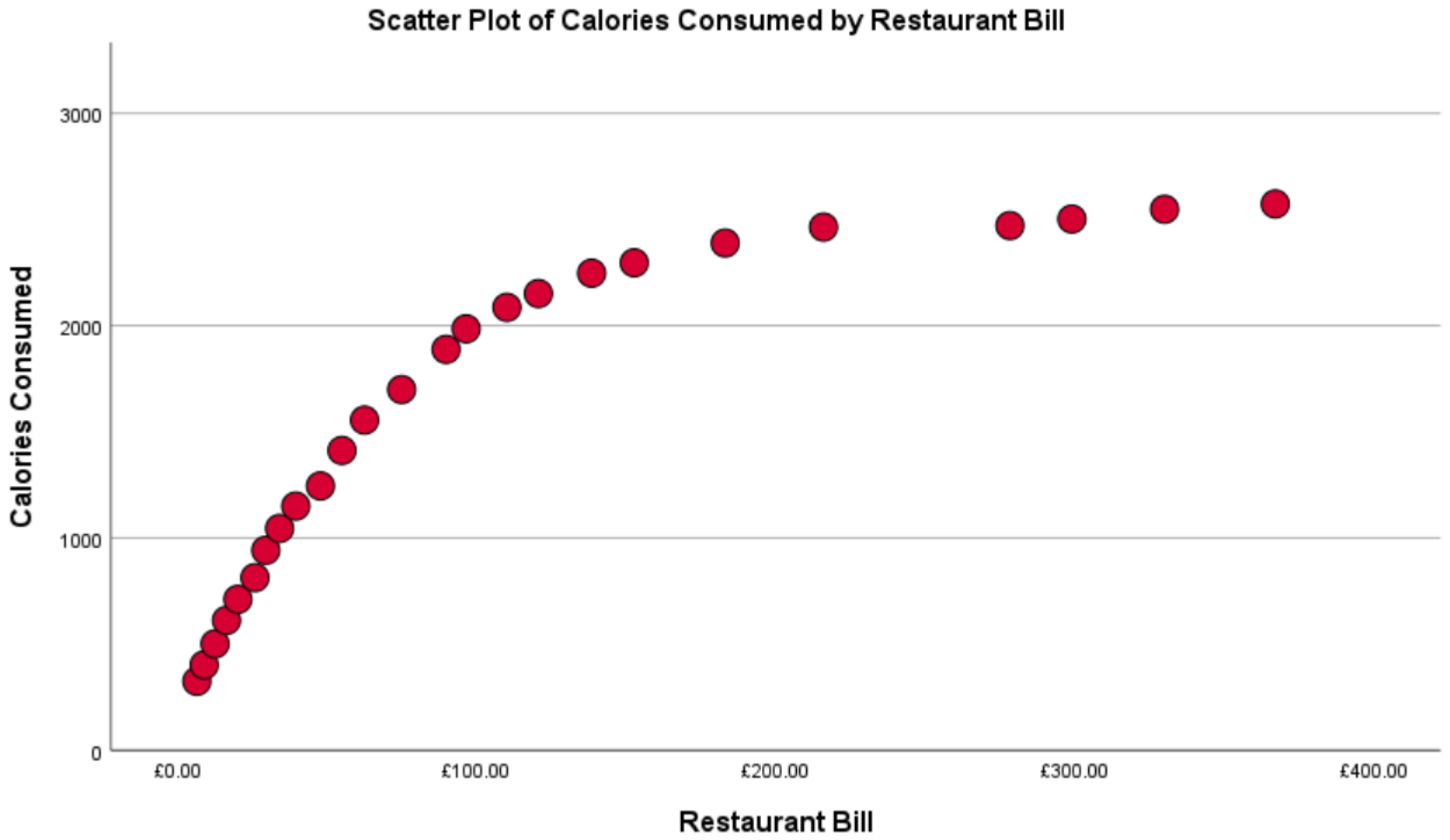
SMARTVISION Europe

Spearman's Rho

- **Spearman's Rho** works by ranking the original values from the lowest number to the highest
- For this reason, it's sometimes referred to as Spearman's Rank Correlation
- Unlike Pearson's Correlation here we don't know if the variables are linearly related as ranking the values hides this relationship
- Spearman's correlation detects *monotonic* relationships. A monotonic relationship is one in which, as the size of one variable increases, the other variables also increases, or where the as the size of one variable increases, as the other variable decreases.
- Spearman correlations are not affected by outliers but analysts should still consider whether extreme outliers are valid reflections of the population under consideration

Spearman's Rho

- Consider this non-linear relationship….



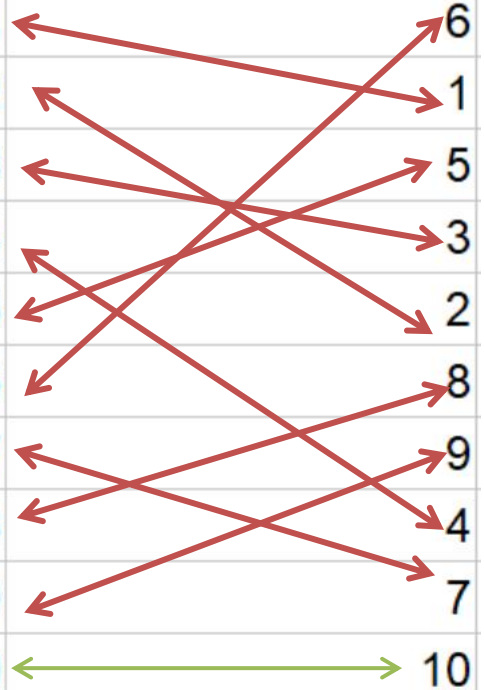
SMARTVISION Europe

Let’s explore how Spearman’s correlations work

Kendall's Tau

- **Kendall's Tau b** also works by ranking the original values from the lowest number to the highest
- However, this time the analysis focuses on *the degree of concordance and discordance* between two ranked columns of data

	Movie_Title	IMDb	Rotten_Tomatoes
1	Uncorked	1	6
2	Spenser Confidential	2	1
3	The Willoughbys	3	5
4	Tigertail	4	3
5	Extraction	5	2
6	The Half of It	6	8
7	To All the Boys I've Loved Before	7	9
8	LA Originals	8	4
9	Miss Americana	9	7
10	Crip Camp: A Disability Revolution	10	10

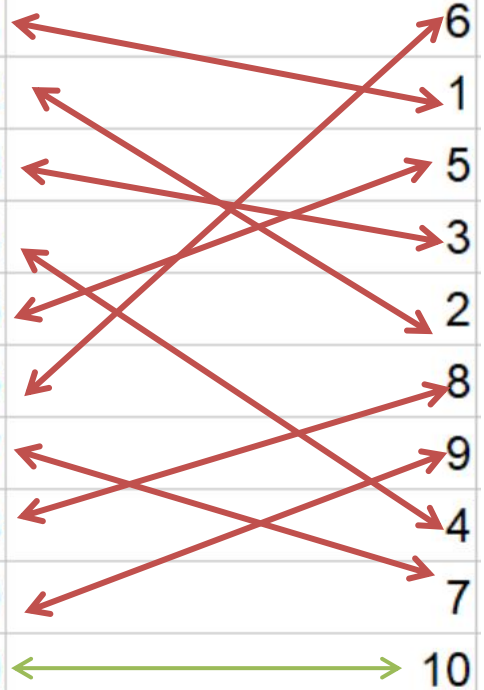


SMARTVISION Europe

Kendall's Tau

- Some argue that the significance estimates and confidence intervals for Kendall's Tau tend to be more reliable than for Spearman correlations.
- Kendall's Tau tend to give smaller correlation values than Spearman's and can also take much longer to calculate

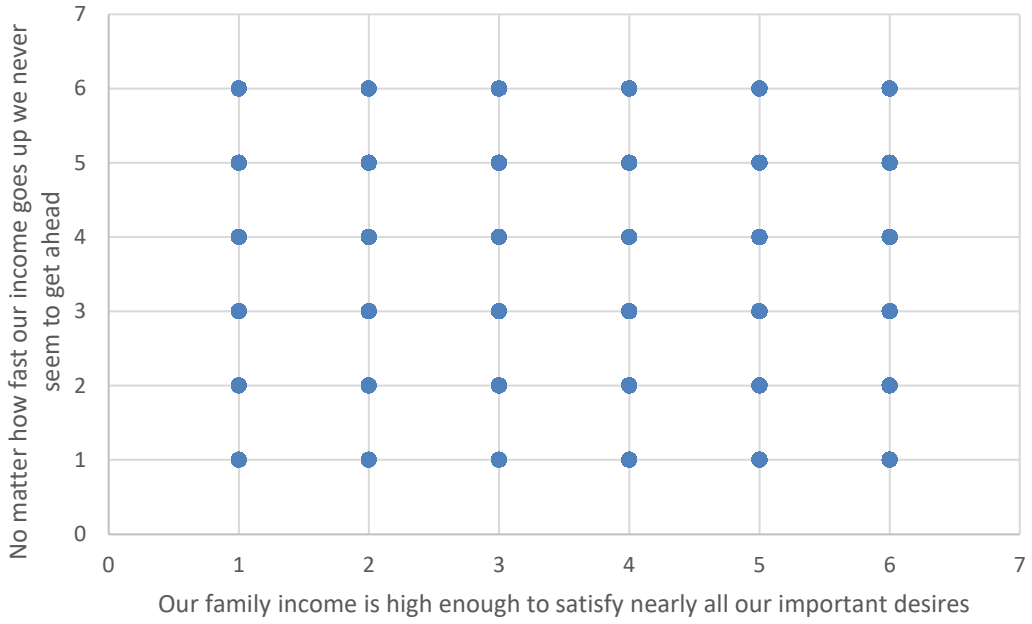
	Movie_Title	IMDb	Rotten_Tomatoes
1	Uncorked	1	6
2	Spenser Confidential	2	1
3	The Willoughbys	3	5
4	Tigertail	4	3
5	Extraction	5	2
6	The Half of It	6	8
7	To All the Boys I've Loved Before	7	9
8	LA Originals	8	4
9	Miss Americana	9	7
10	Crip Camp: A Disability Revolution	10	10

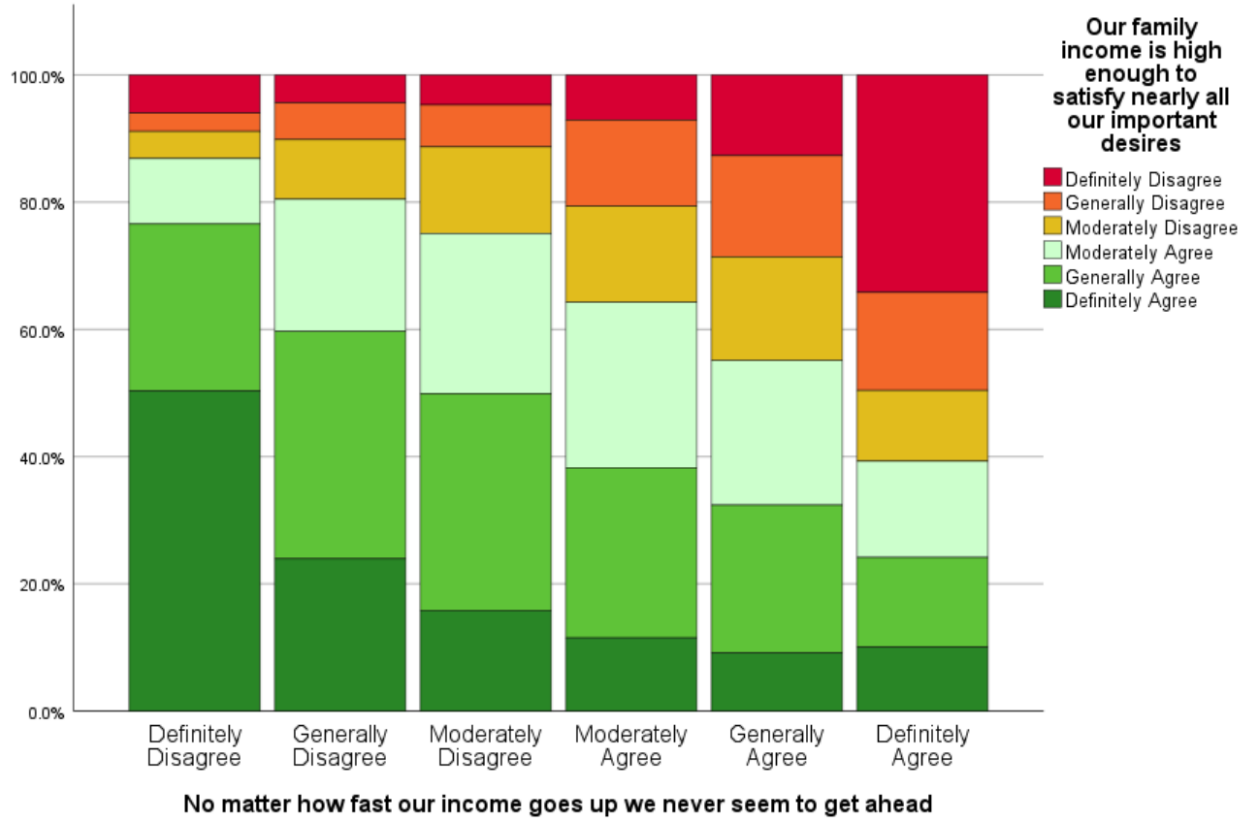


SMARTVISION Europe

Let's compare Kendall's Tau to Spearman's correlations

Visualising Correlations for Ordinal Variables



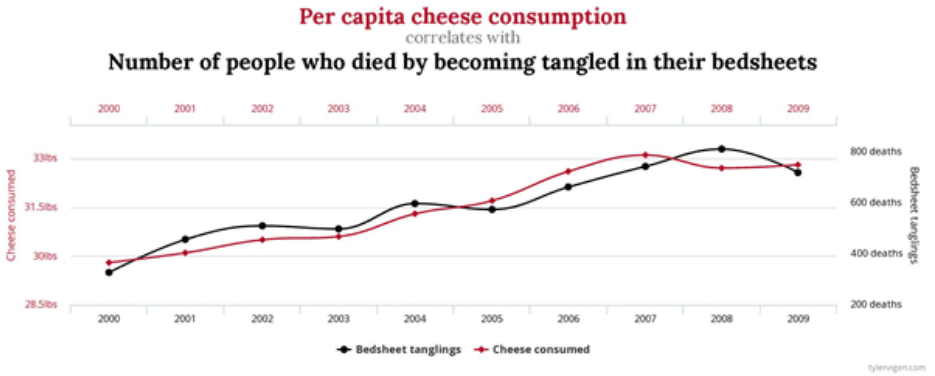


May require a different approach than scatterplots

SMARTVISION Europe

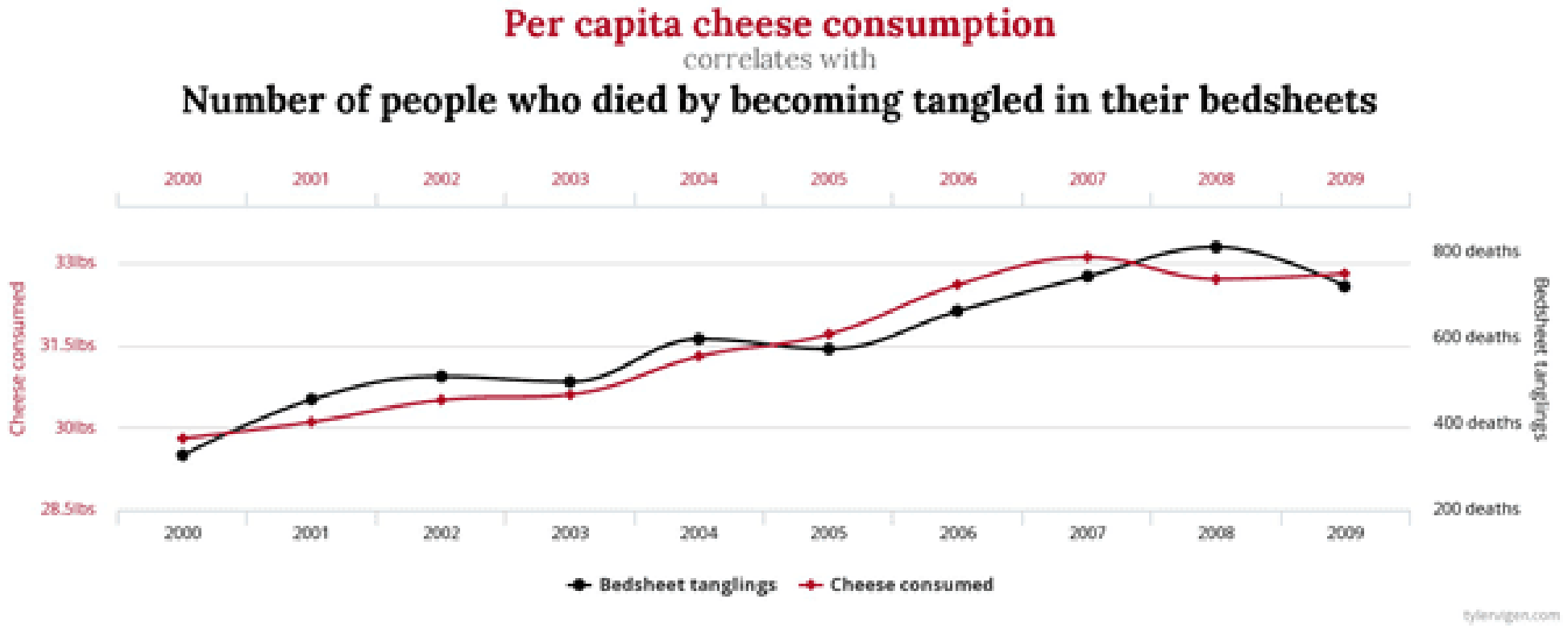
A SELECT INTERNATIONAL COMPANY

SMARTVISION Europe



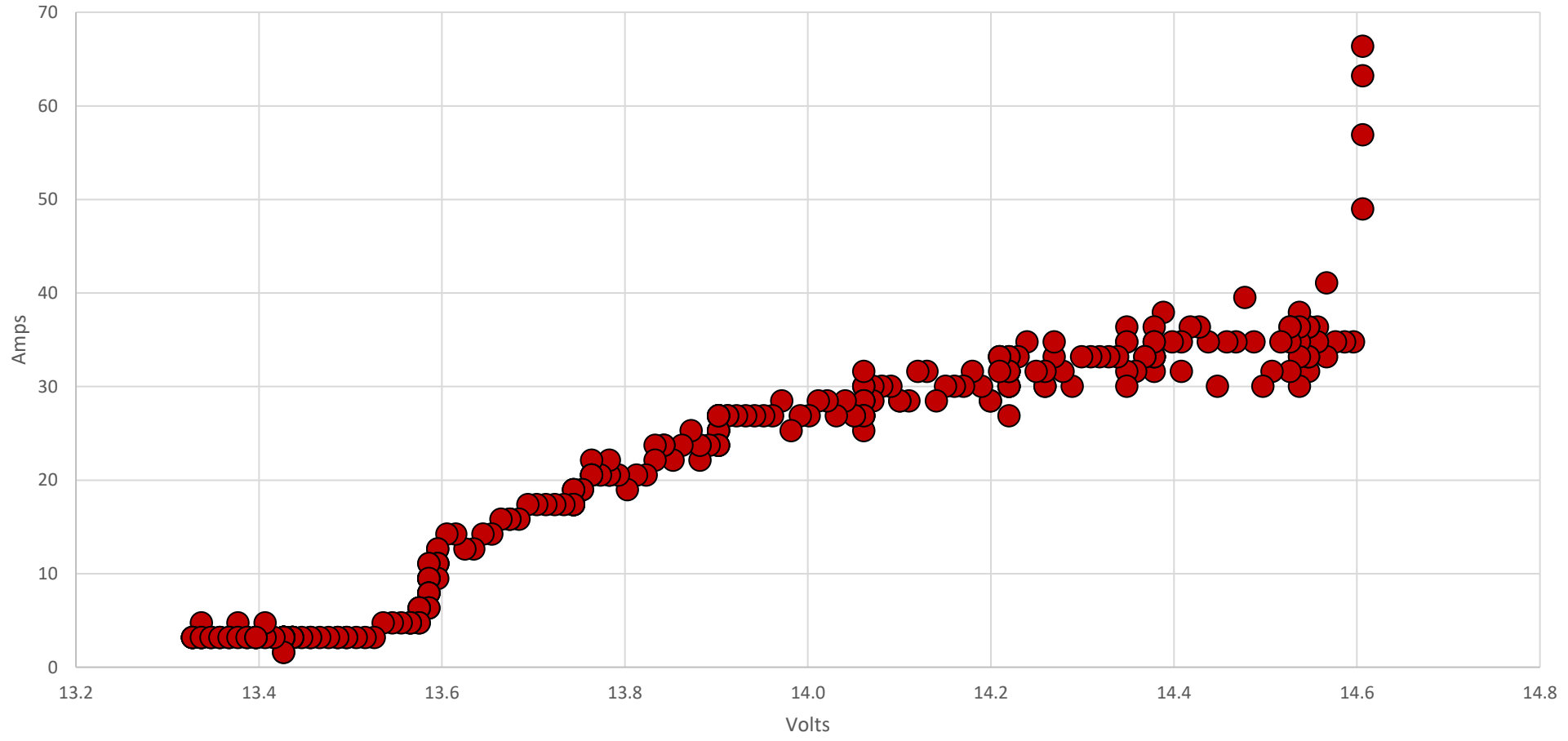
The Limitations of Correlations

Correlation does not indicate causation



https://www.productleadership.com/does-causation-imply-correlation/

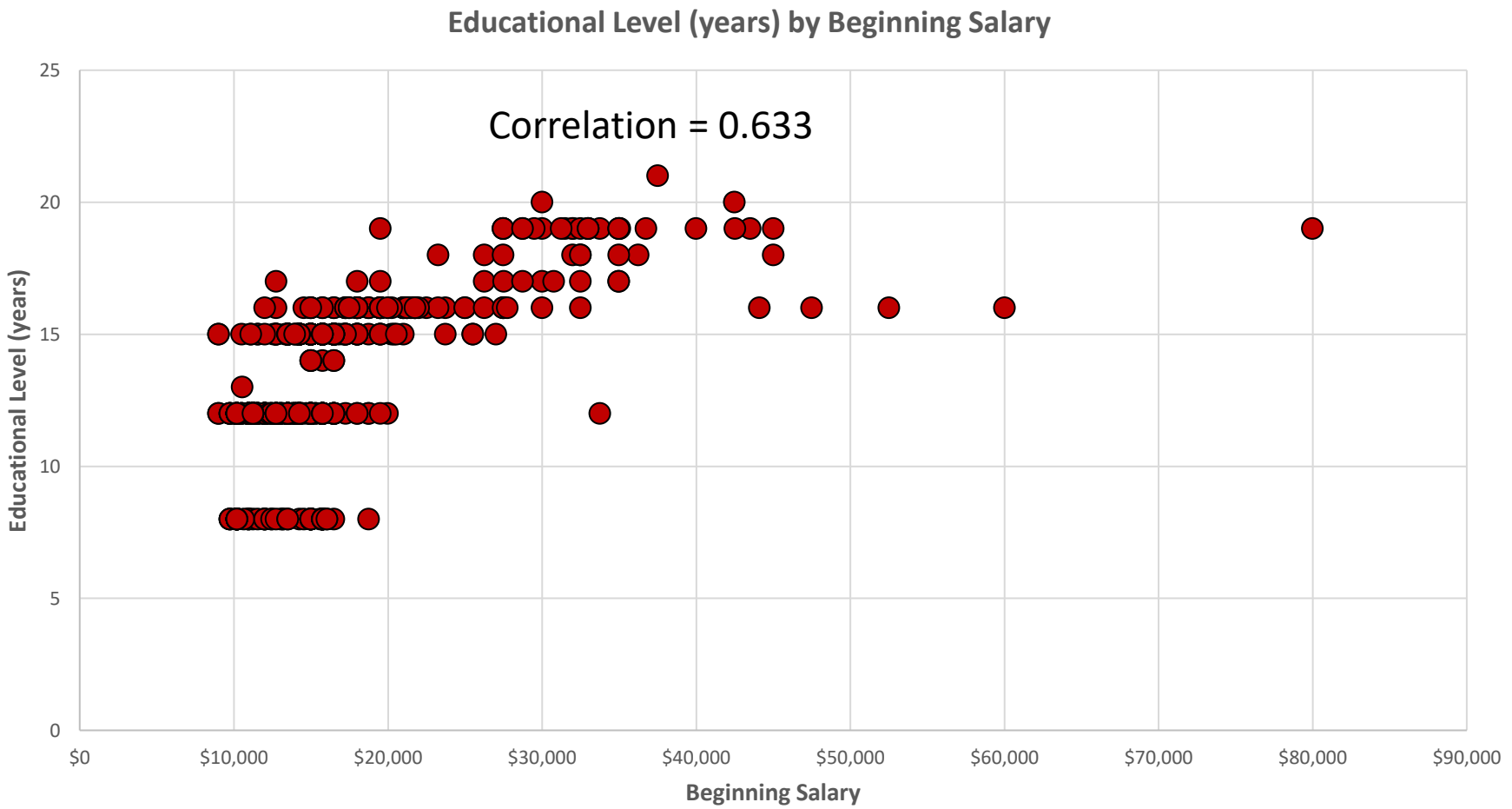
Can’t accurately measure curvilinear relationships



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

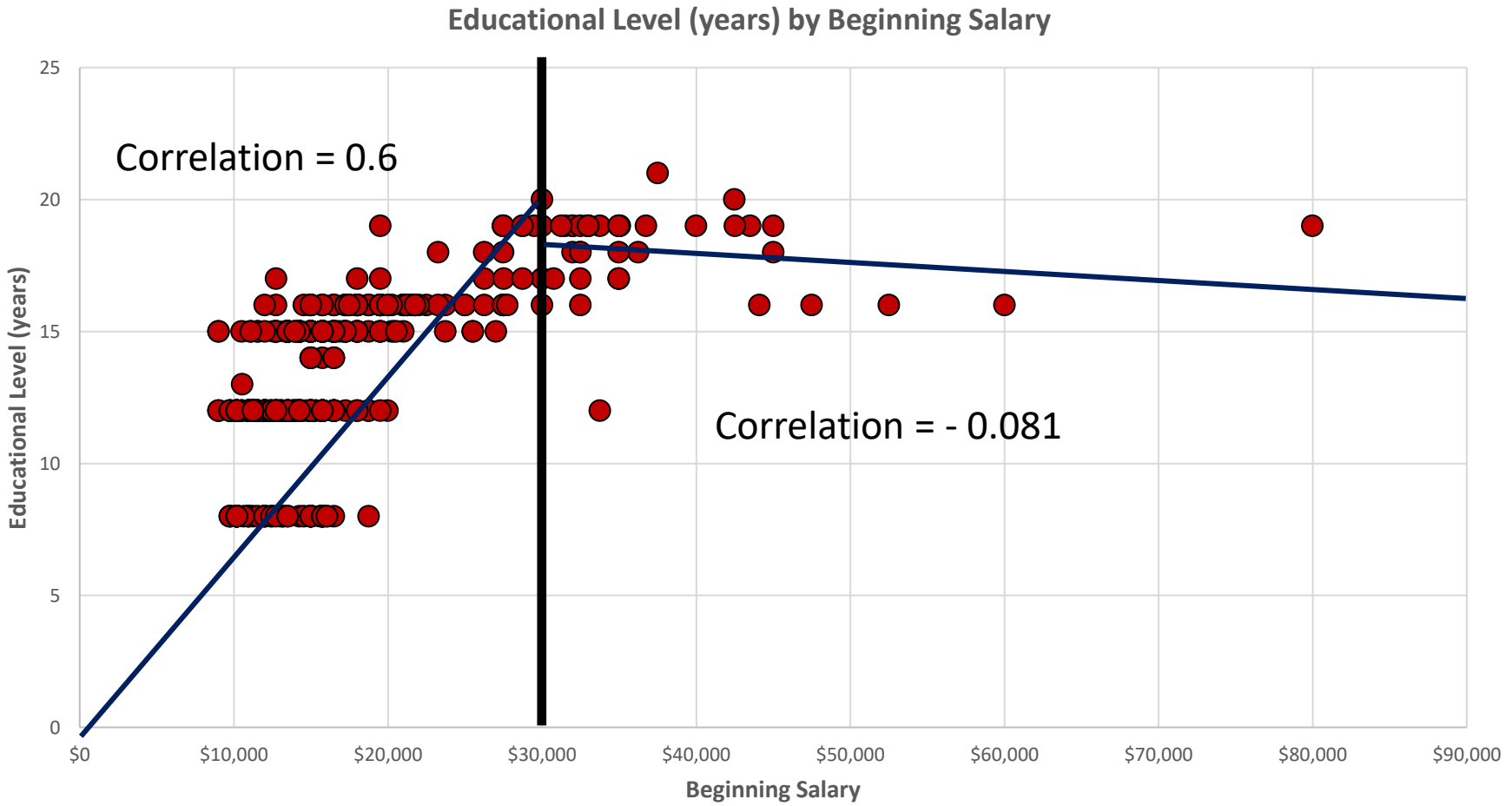
Are influenced by the range of values in the sample



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

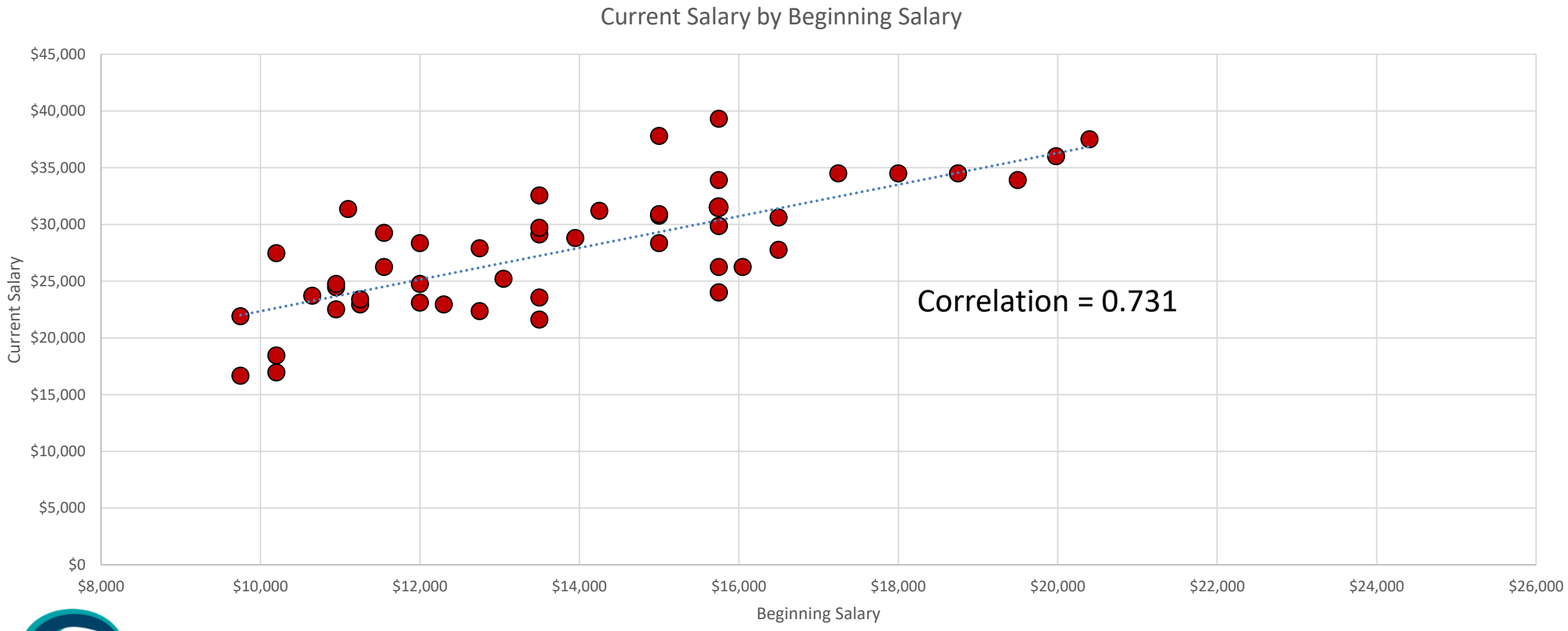
Are influenced by the range of values in the sample



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

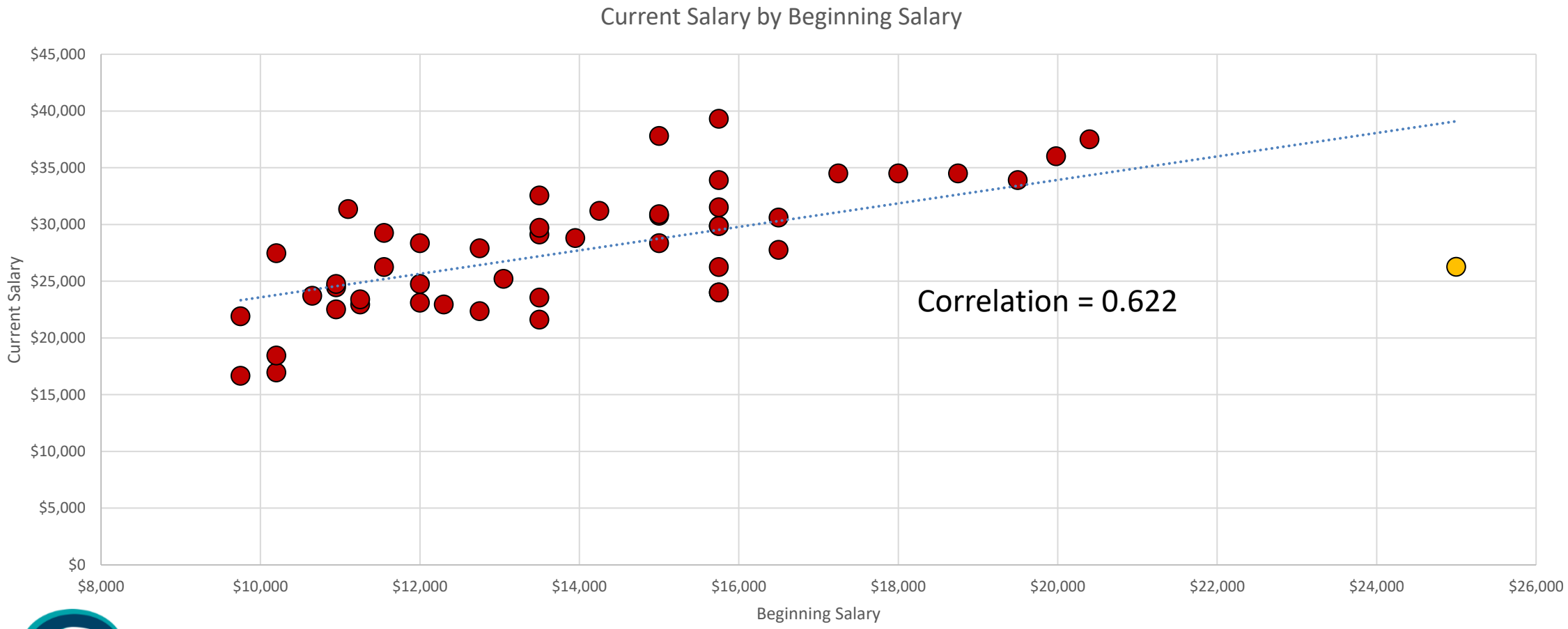
Can be unduly affected by extreme/outlier values



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Can be unduly affected by extreme/outlier values



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Working with Smart Vision Europe Ltd.

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - http://www.sv-europe.com/buy-spss-online/
- **Training and Consulting Services**
  - Guided consulting & training to develop in house skills
  - Delivery of classroom training courses / side by side training support
  - Identification & recruitment of analytical skills into your organisation
- **Advice and Support**
  - offer 'no strings attached' technical and business advice relating to analytical activities
  - Technical support services

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

SMARTVISION Europe

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In
Sign up for our Newsletter

Thank you