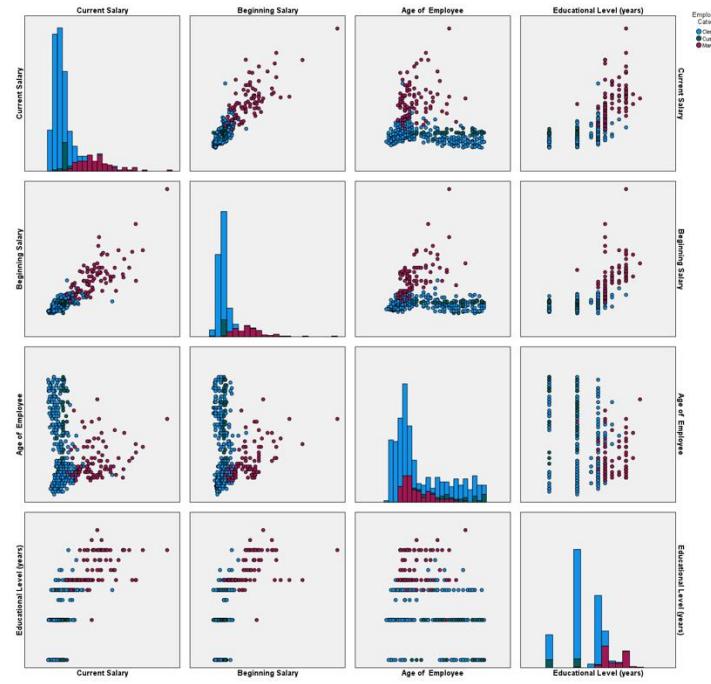


# Correlation analysis with SPSS Statistics

**Jarlath Quinn – Analytics Consultant**



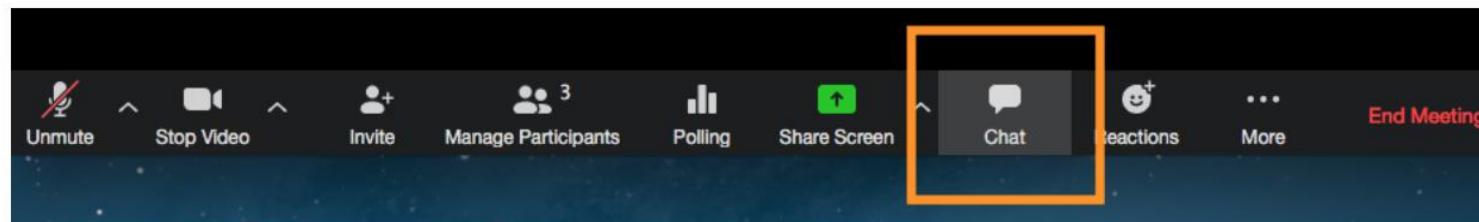
# Correlation analysis with SPSS Statistics

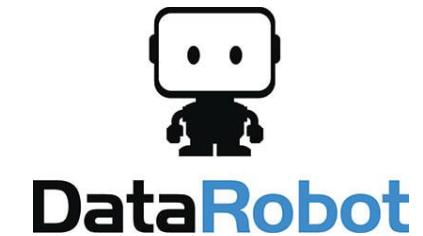
Just waiting for all attendees to join...

**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 chat panel – if we run out of time we will follow up with you.





- Gold 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



A SELECT INTERNATIONAL COMPANY

# Agenda

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



# Why are correlations useful?

**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).

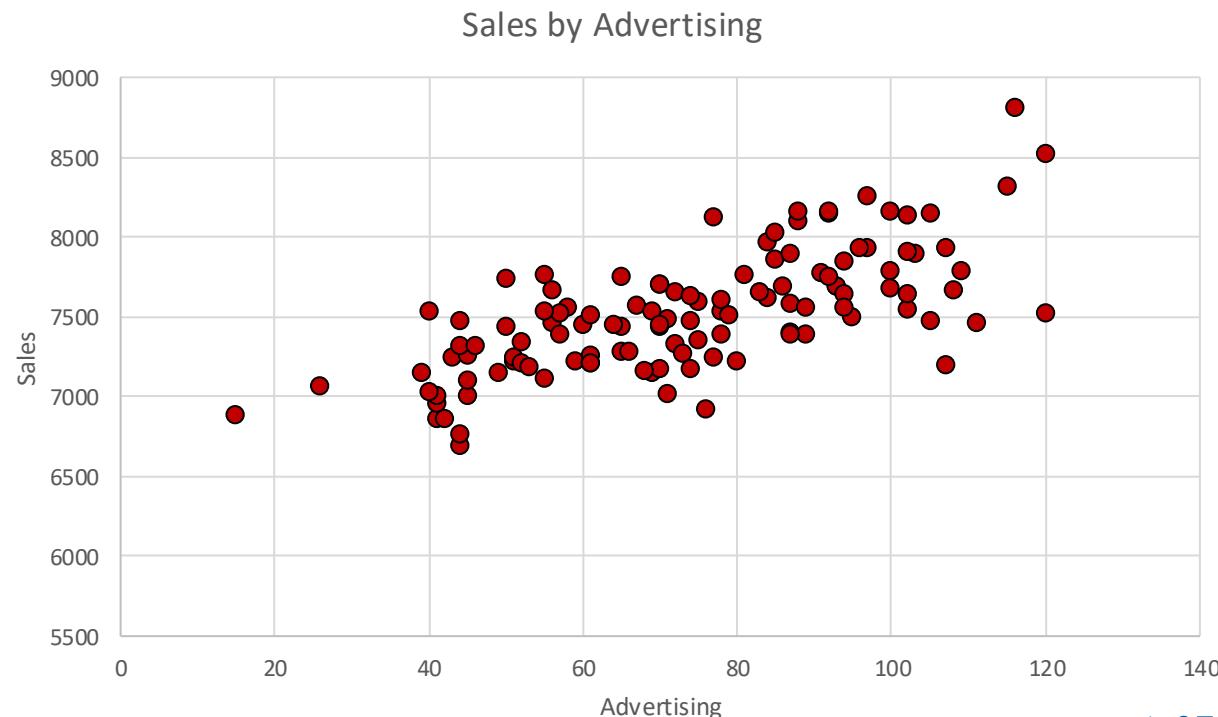
# Why use correlations?

- 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.





# Interpreting 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).

# Linear Correlation Scale

+1

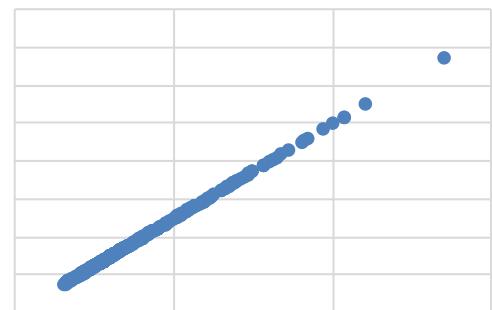
+0.5

0

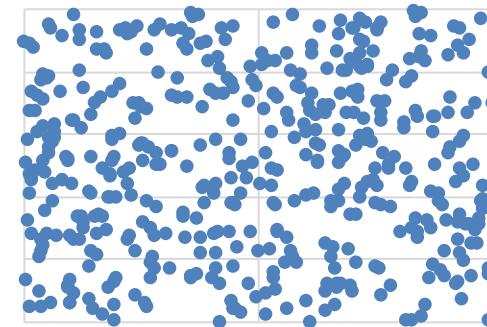
-0.5

-1

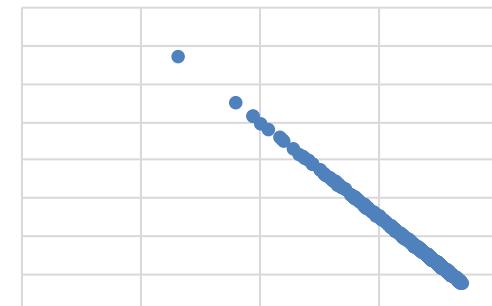
Perfect Positive Linear  
Relationship



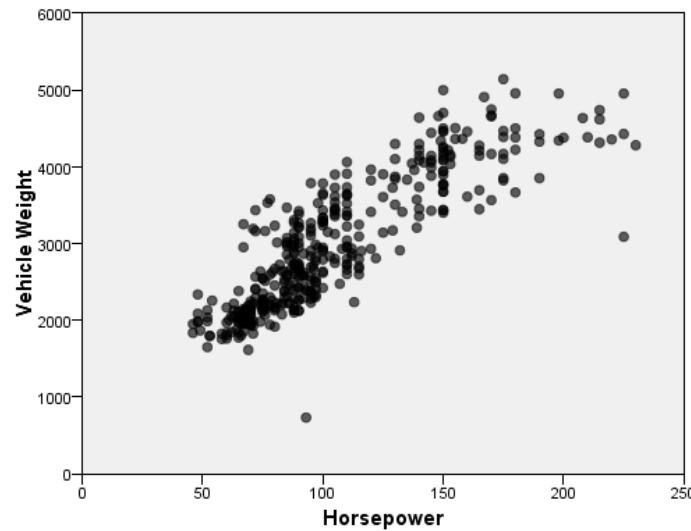
No Linear Relationship



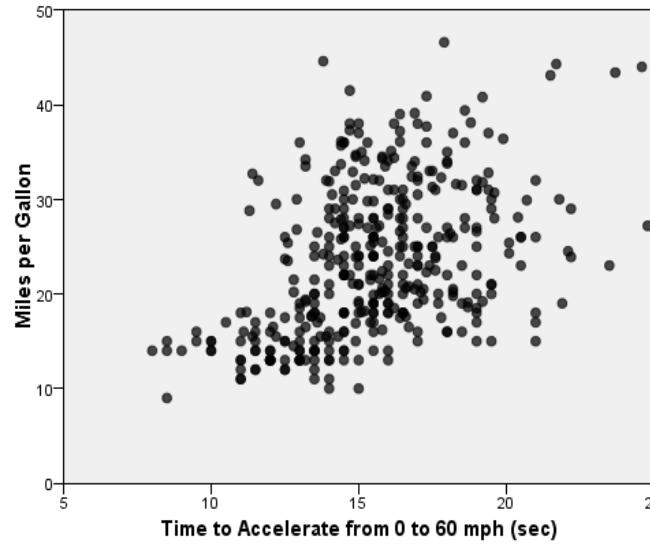
Perfect Negative  
Linear Relationship



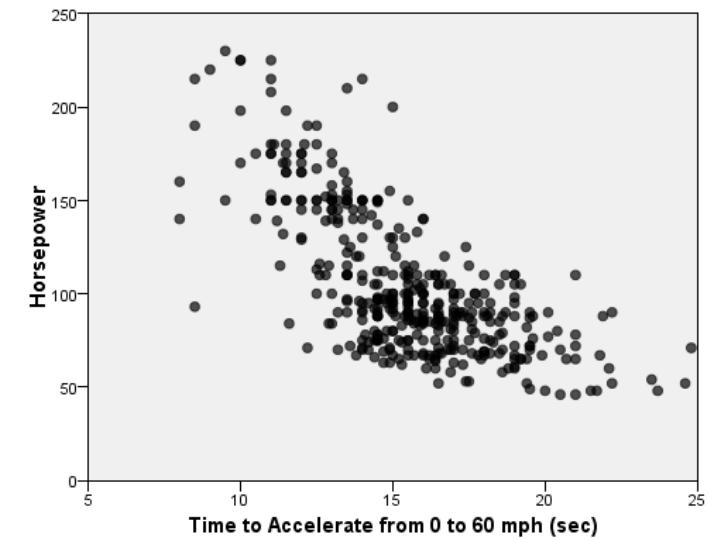
# Pearson's $r$ correlations



0.859



0.434

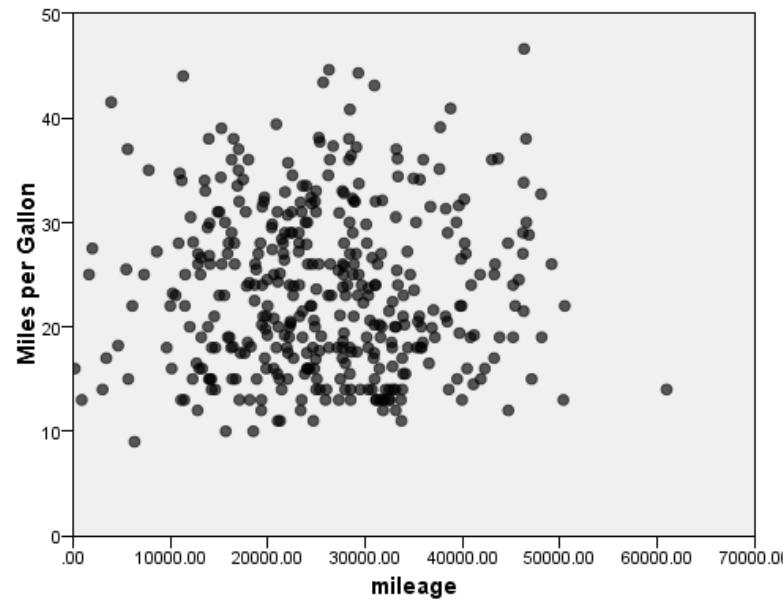


-.701

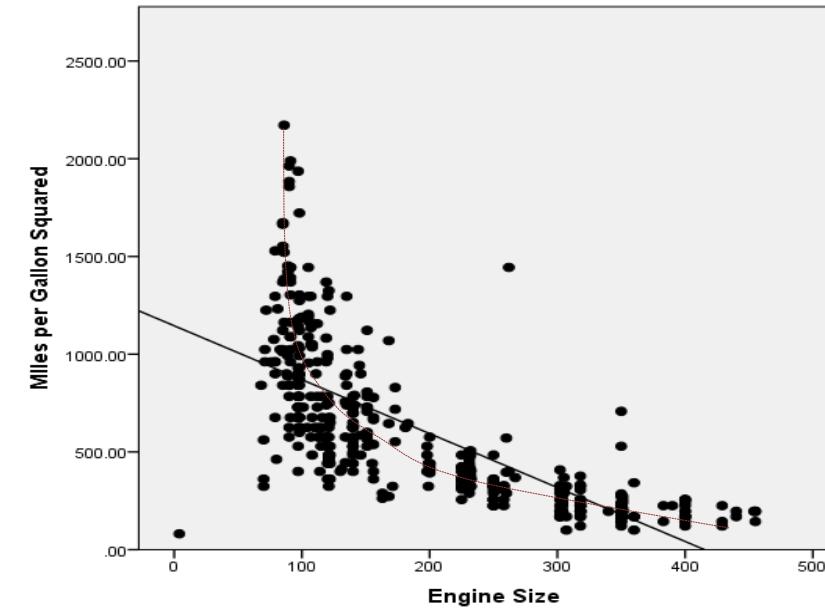
Pearson's  $r$  correlation coefficients



# Non-Linear Relationships



-0.005



-.671

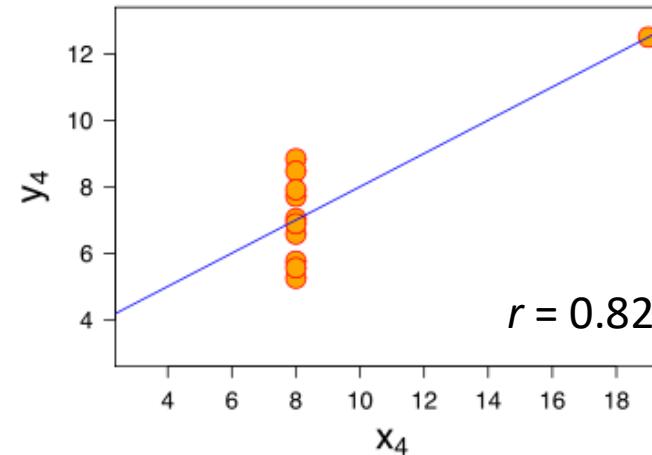
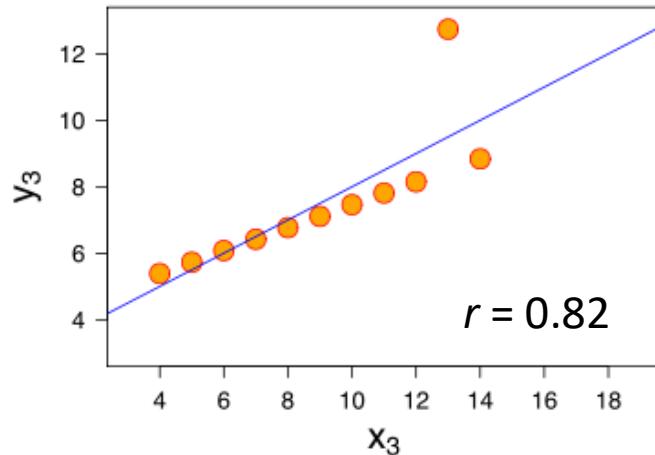
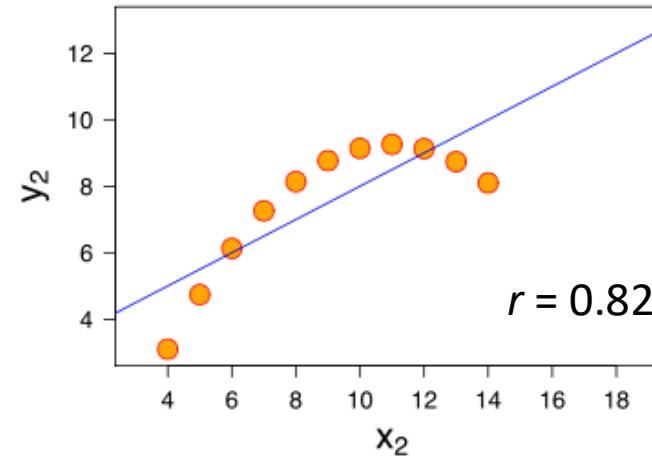
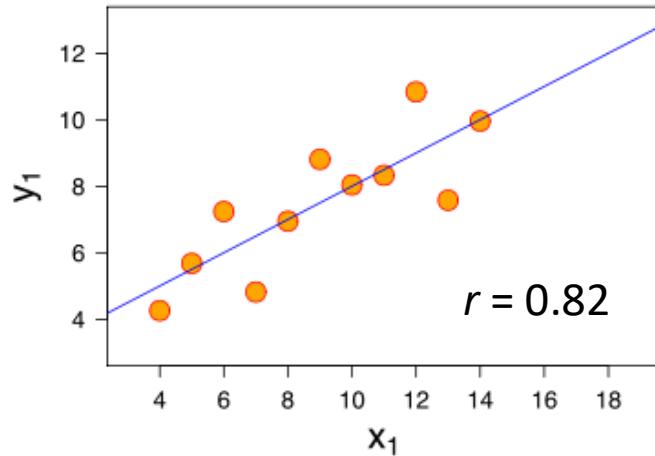
Pearson's  $r$  Correlations



**SMART VISION**  
Europe

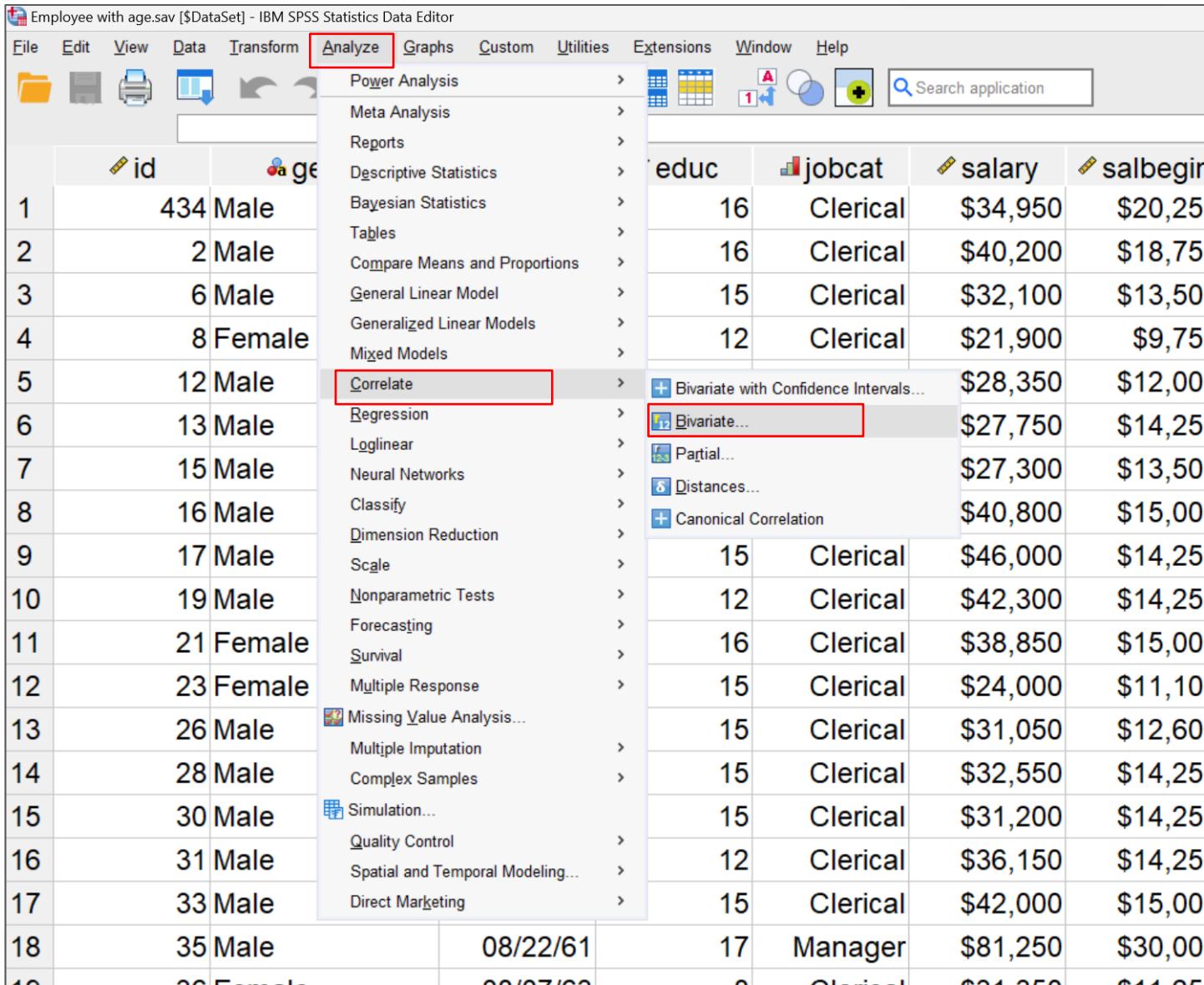
A SELECT INTERNATIONAL COMPANY

## *A word of warning: always investigate the relationship*



# Example SPSS Correlations

- Analyze
  - Correlate
    - Bivariate



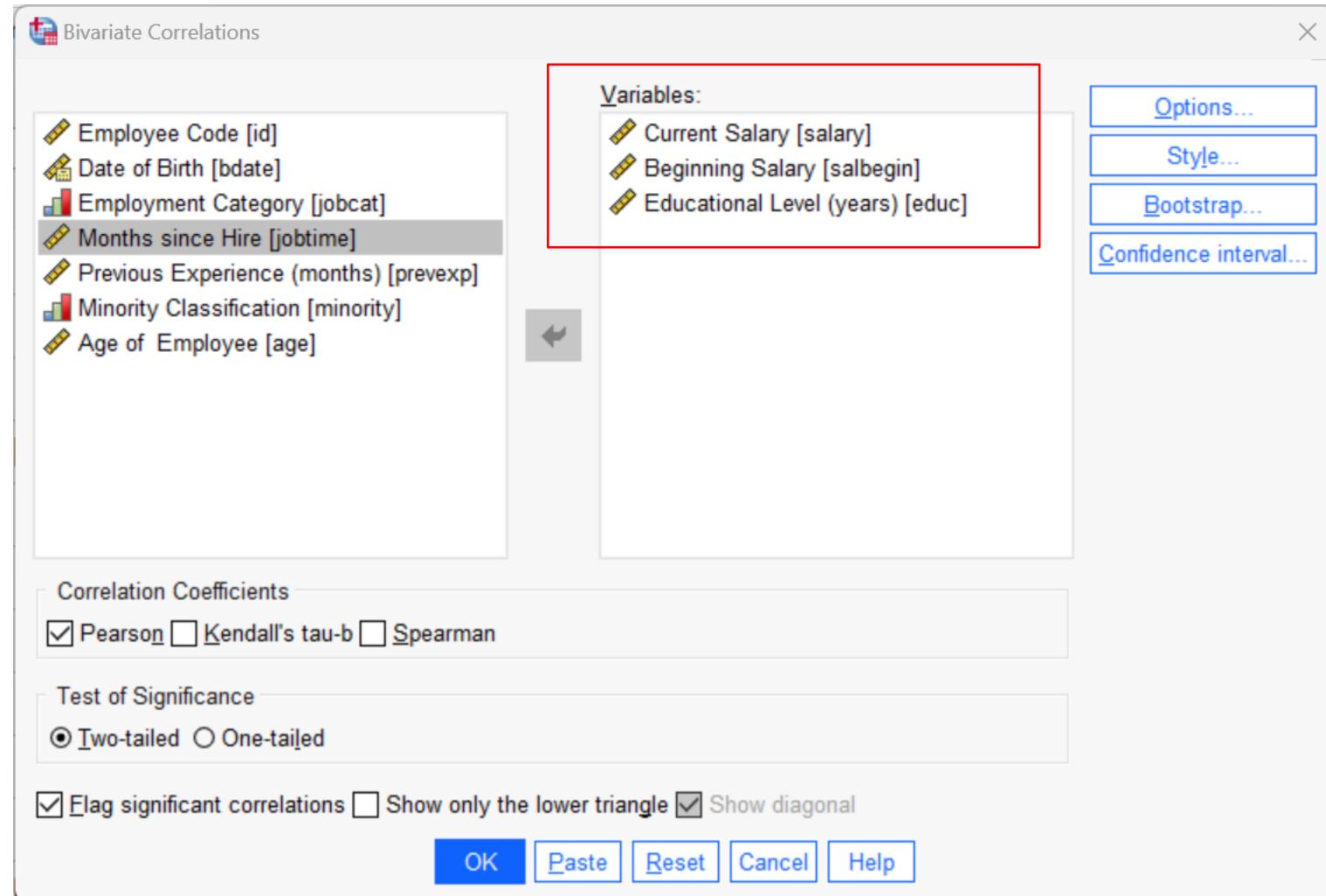
The screenshot shows the IBM SPSS Statistics Data Editor interface. The menu bar is visible with 'Analyze' highlighted in red. The 'Analyze' menu is open, showing various statistical options. The 'Correlate' option is also highlighted in red. Below the menu, a data table is displayed with columns for 'id', 'age', 'jobcat', 'salary', and 'salbegin'. The data shows 18 rows of employee information, including their ID, age, job category, salary, and start date.

	id	age	jobcat	salary	salbegin
1	434	Male			
2	2	Male			
3	6	Male			
4	8	Female			
5	12	Male			
6	13	Male			
7	15	Male			
8	16	Male			
9	17	Male			
10	19	Male			
11	21	Female			
12	23	Female			
13	26	Male			
14	28	Male			
15	30	Male			
16	31	Male			
17	33	Male			
18	35	Male	08/22/61		
19	36	Female	08/27/61		



# 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



# Example SPSS Correlations

**Correlations**

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

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

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

# Example SPSS Correlations

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

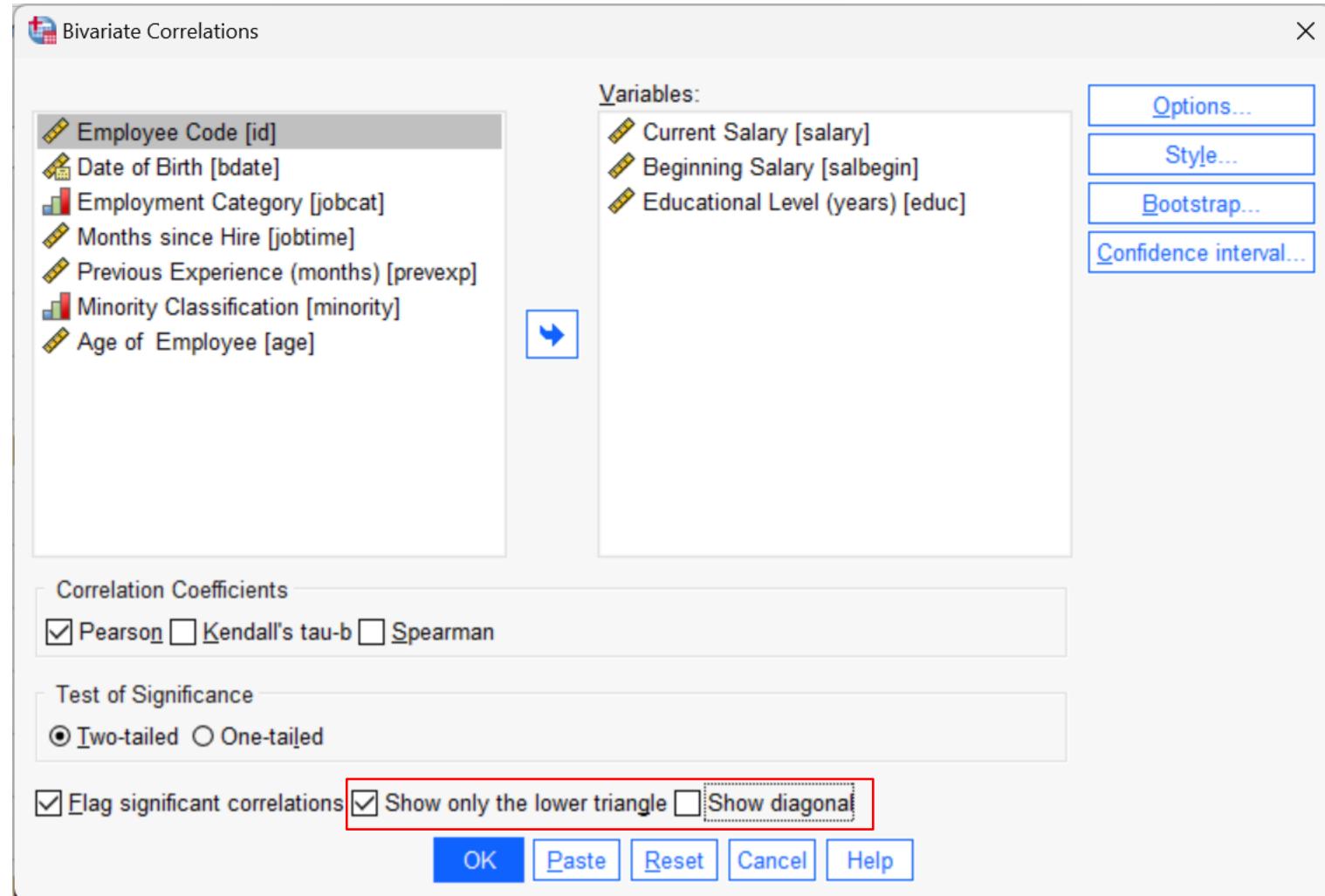
**Correlations**

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



# Example SPSS Correlations

**Correlations**

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



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

# 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 following:
  - 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

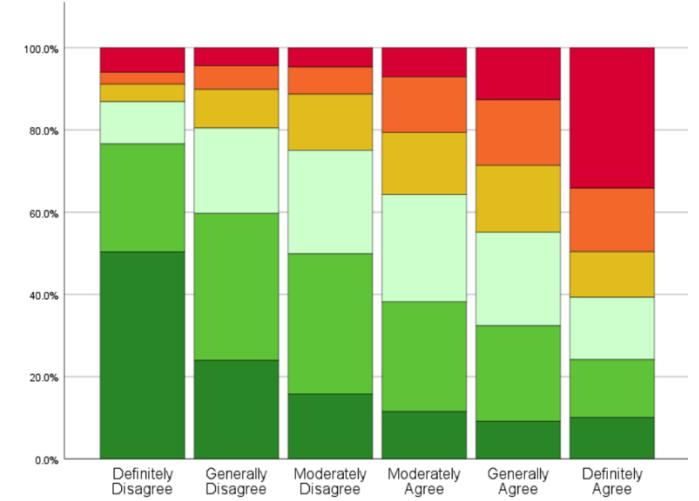
$\bar{x}$  = mean of the values of the x-variable

$y_i$  = values of the y-variable in a sample

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



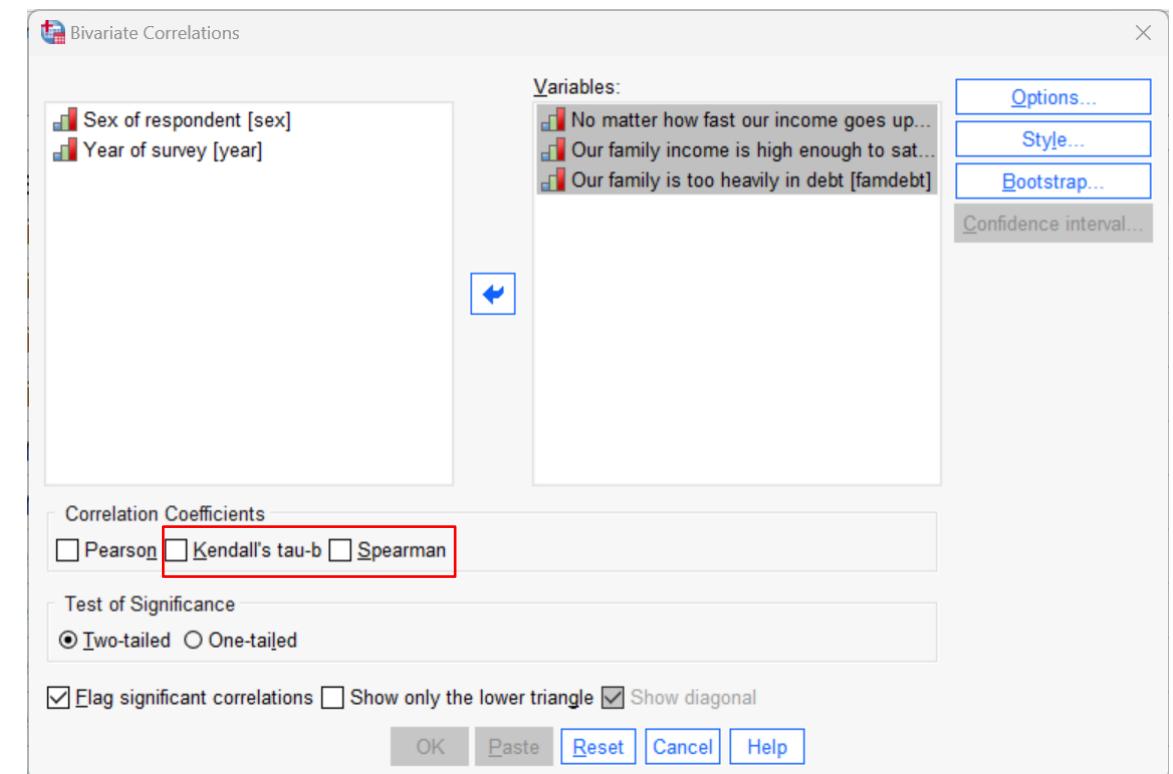
Let's see an example of calculating  
Pearson's R



# 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

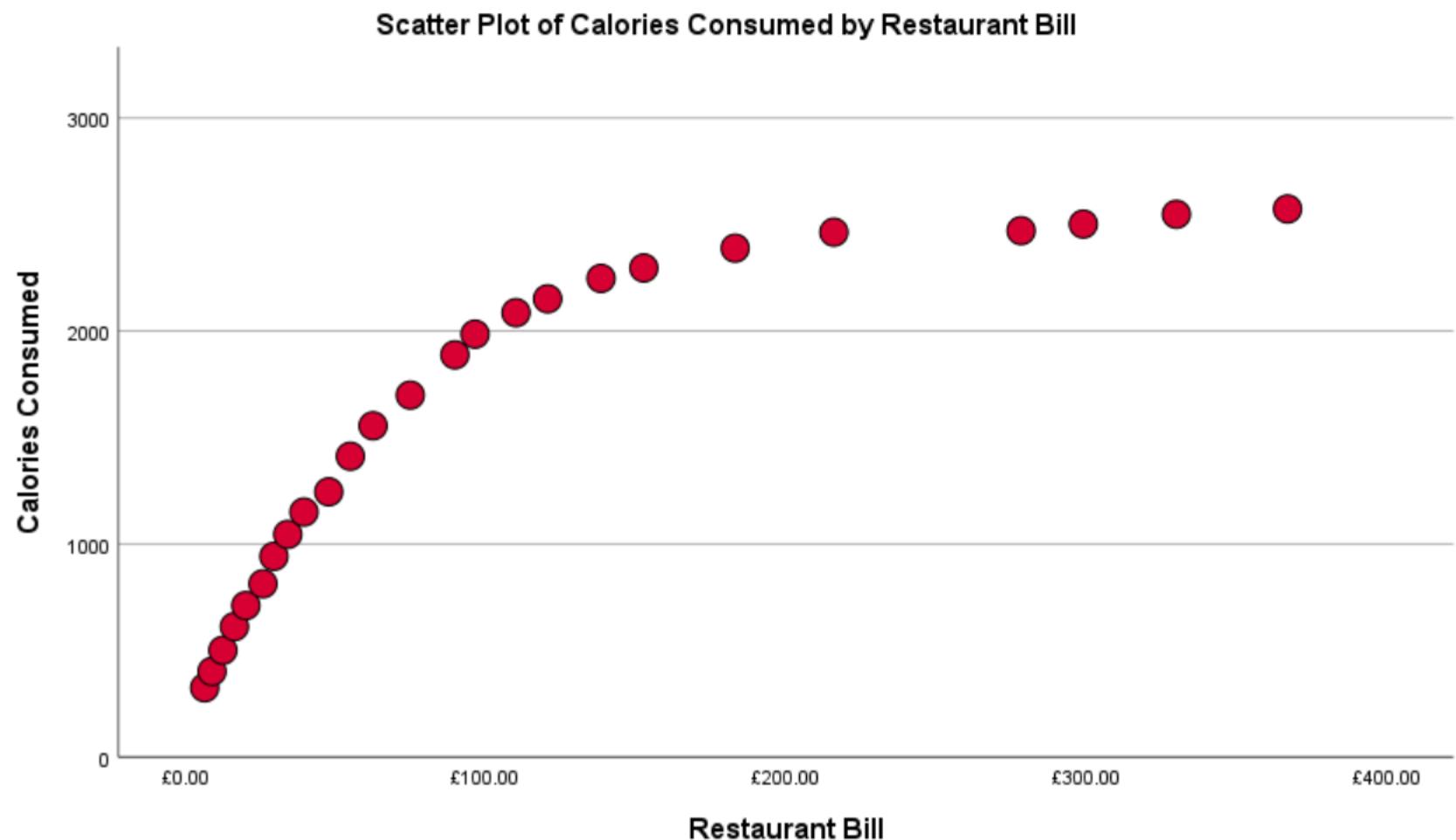


# Spearman's Rho

- **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
- 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....



Let's explore how Spearman's correlations work

# An alternative to Spearman's Correlation: Kendall's Tau b

# 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



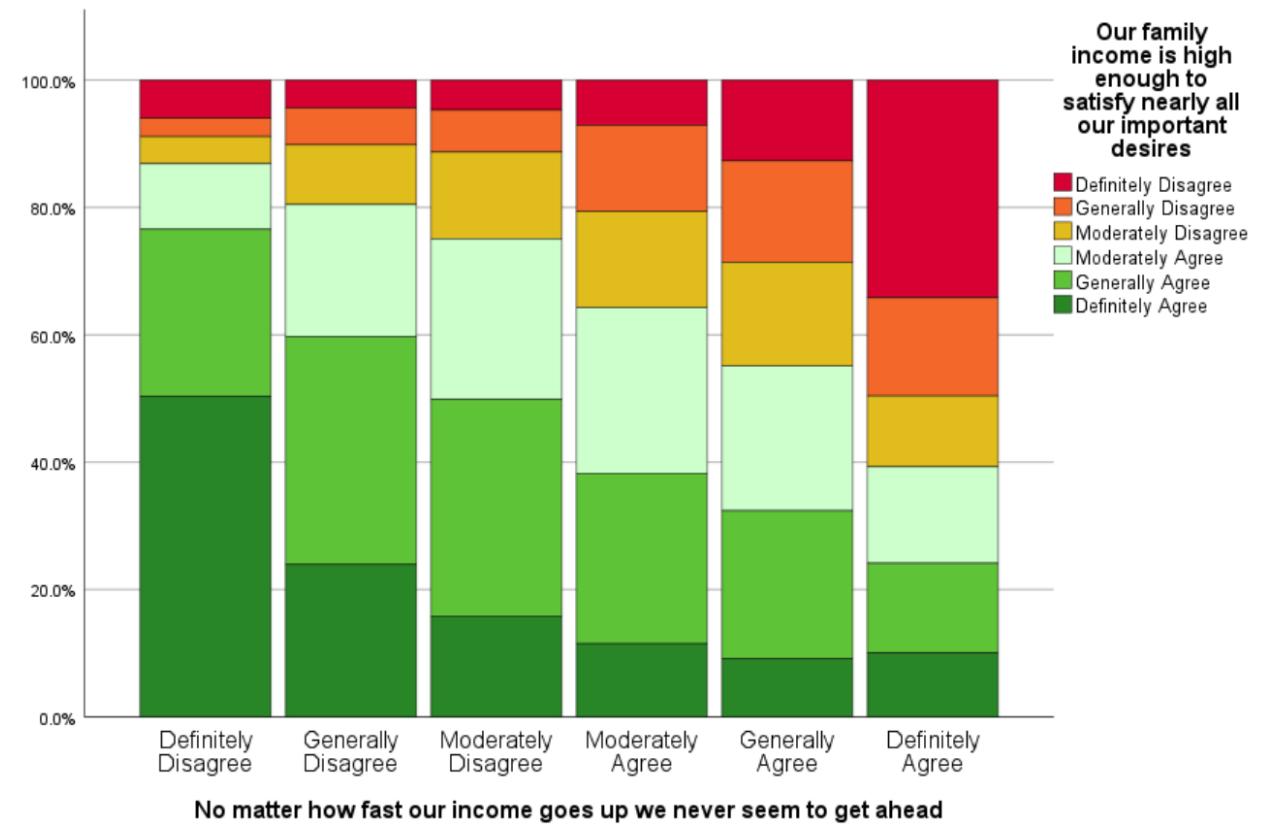
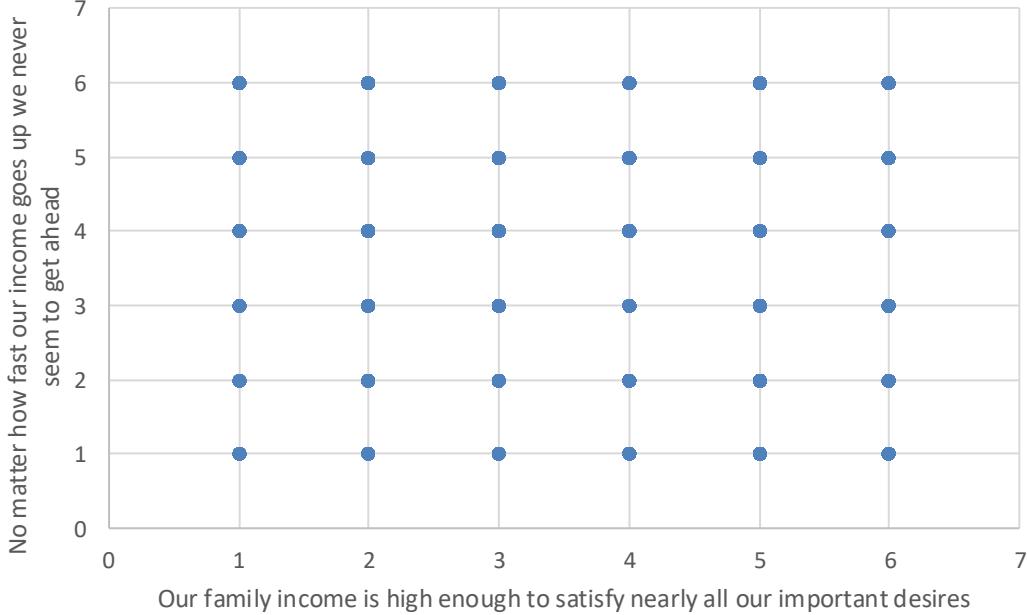
# Kendall's Tau

- 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



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

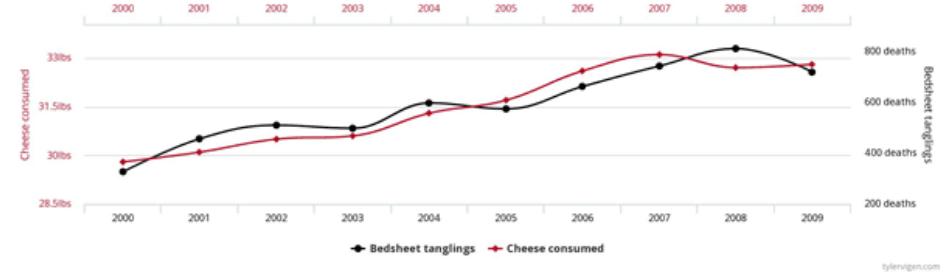
# Visualising Correlations for Ordinal Variables



May require a different approach than scatterplots

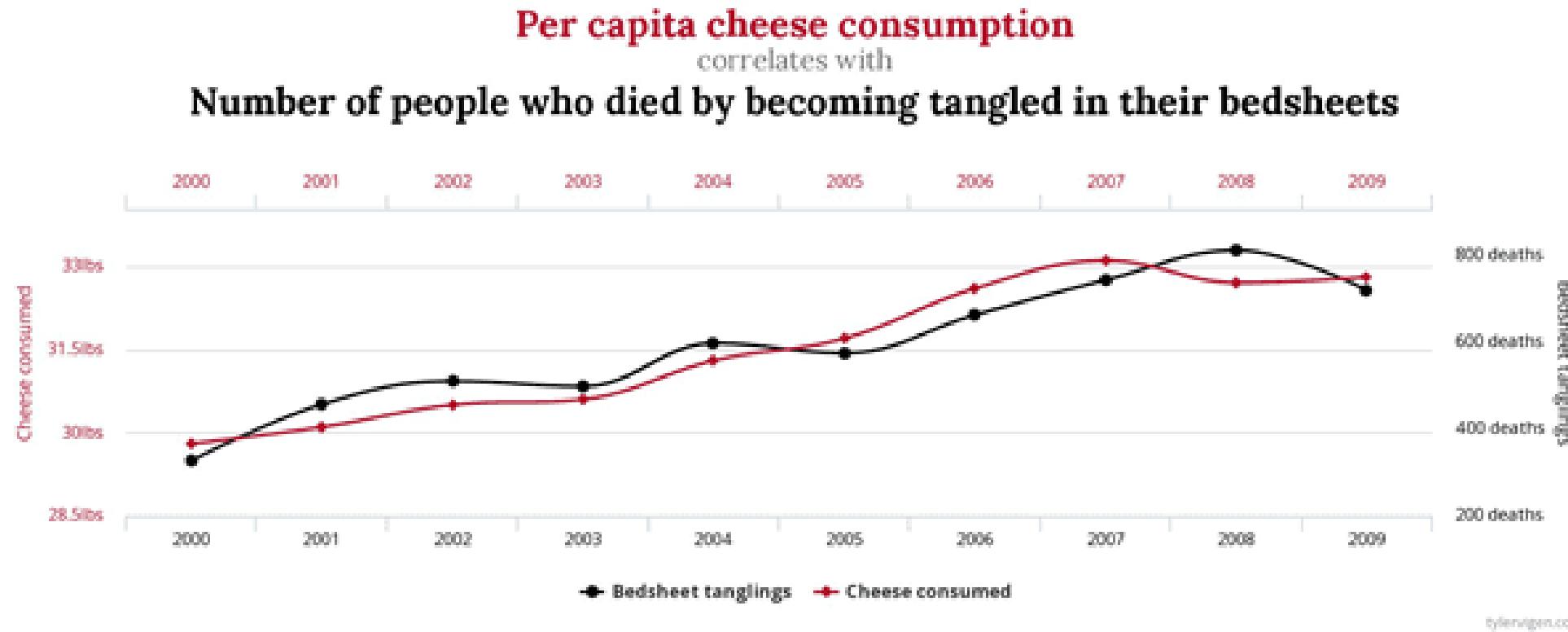


Per capita cheese consumption  
correlates with  
Number of people who died by becoming tangled in their bedsheets



# The Limitations of Correlations

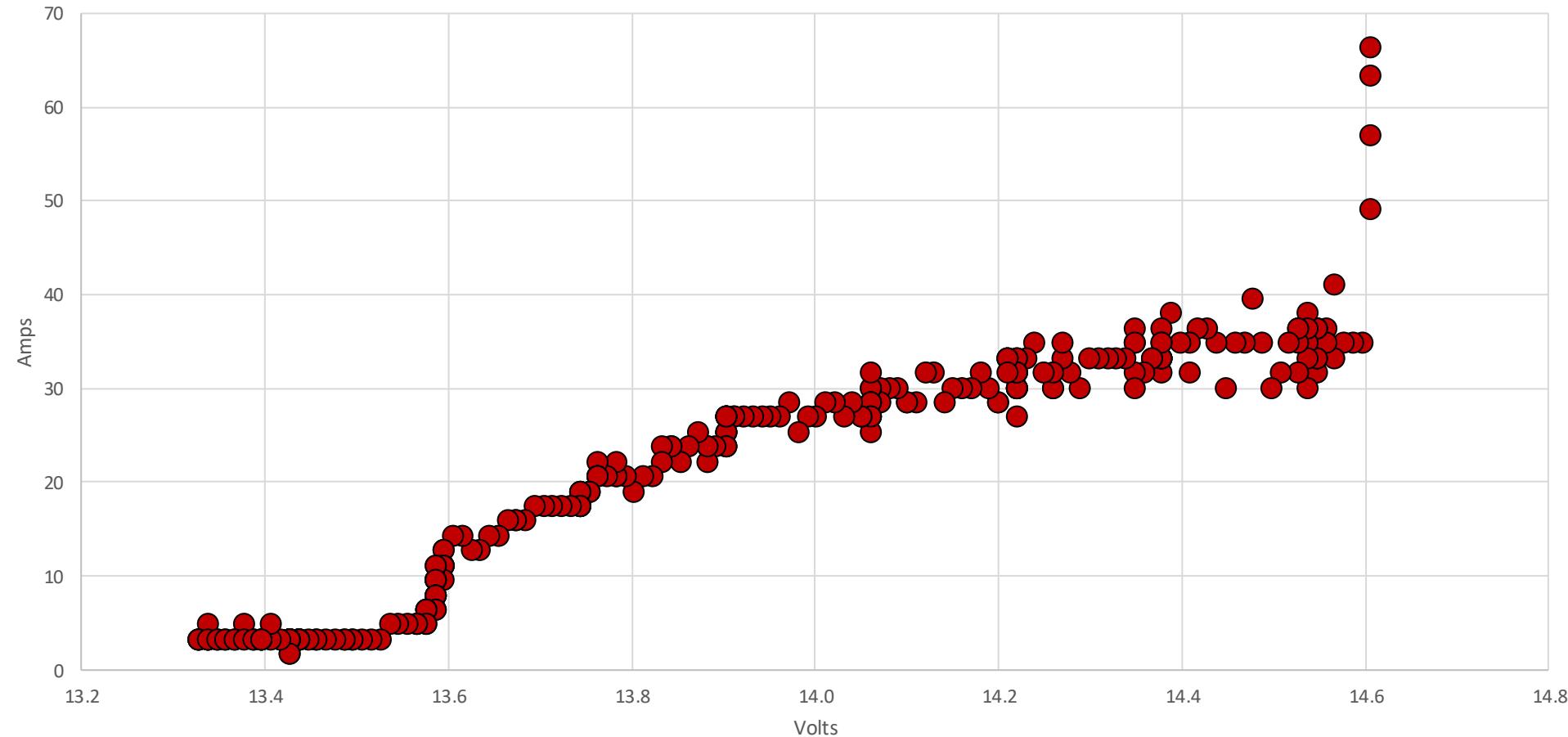
# Correlation does not indicate causation



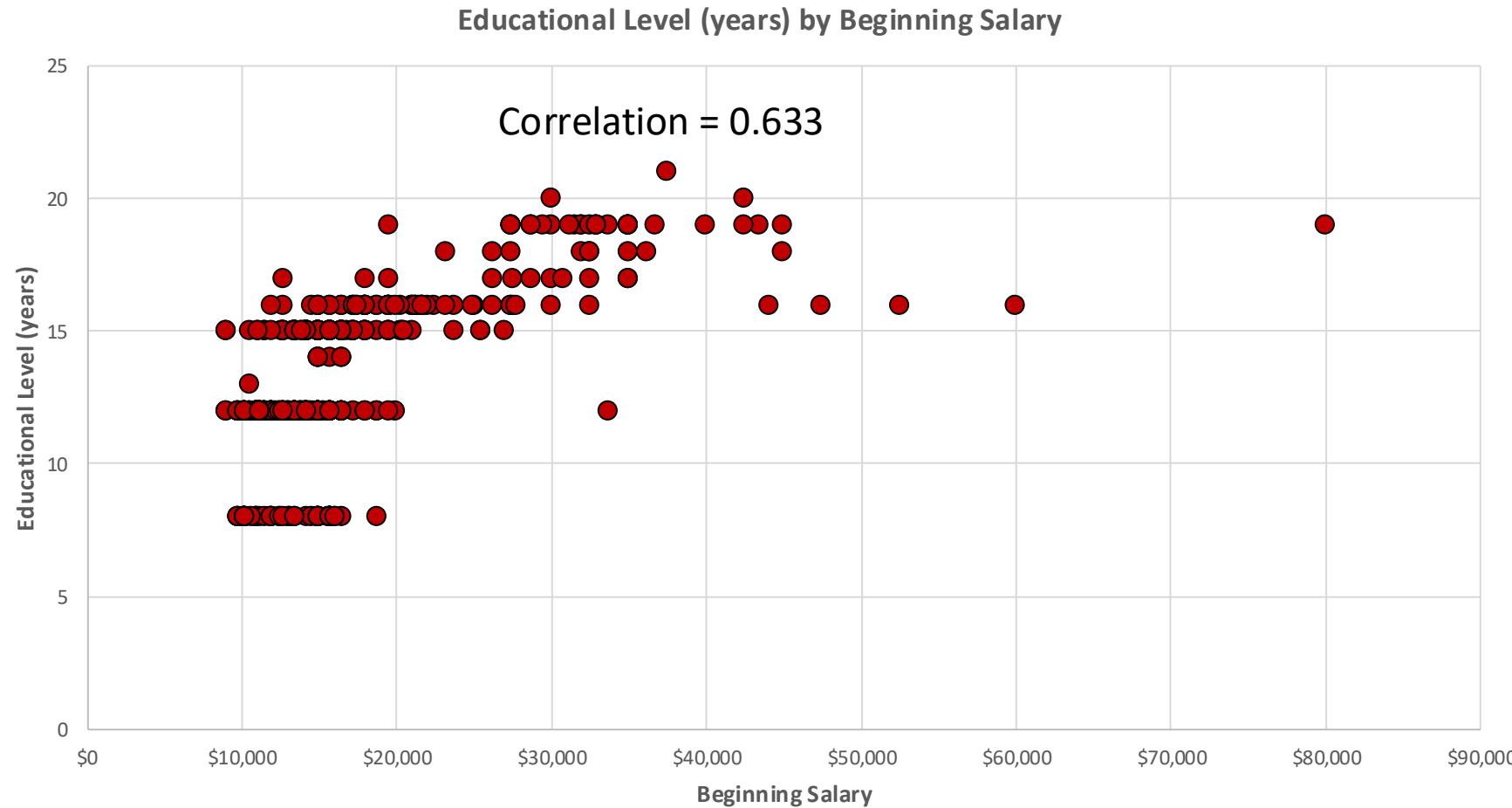
<https://www.productleadership.com/does-causation-imply-correlation/>



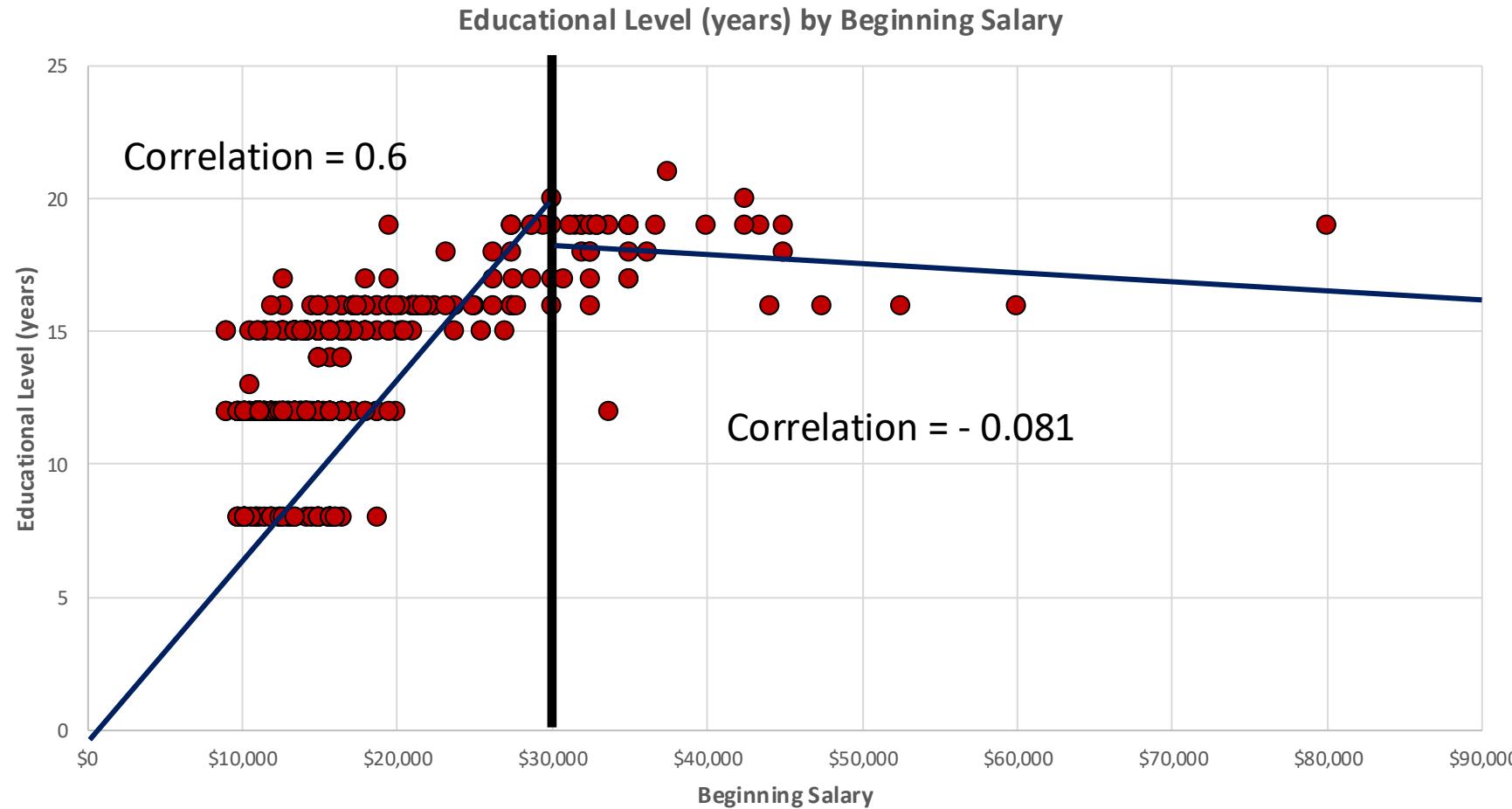
# Can't accurately measure curvilinear relationships



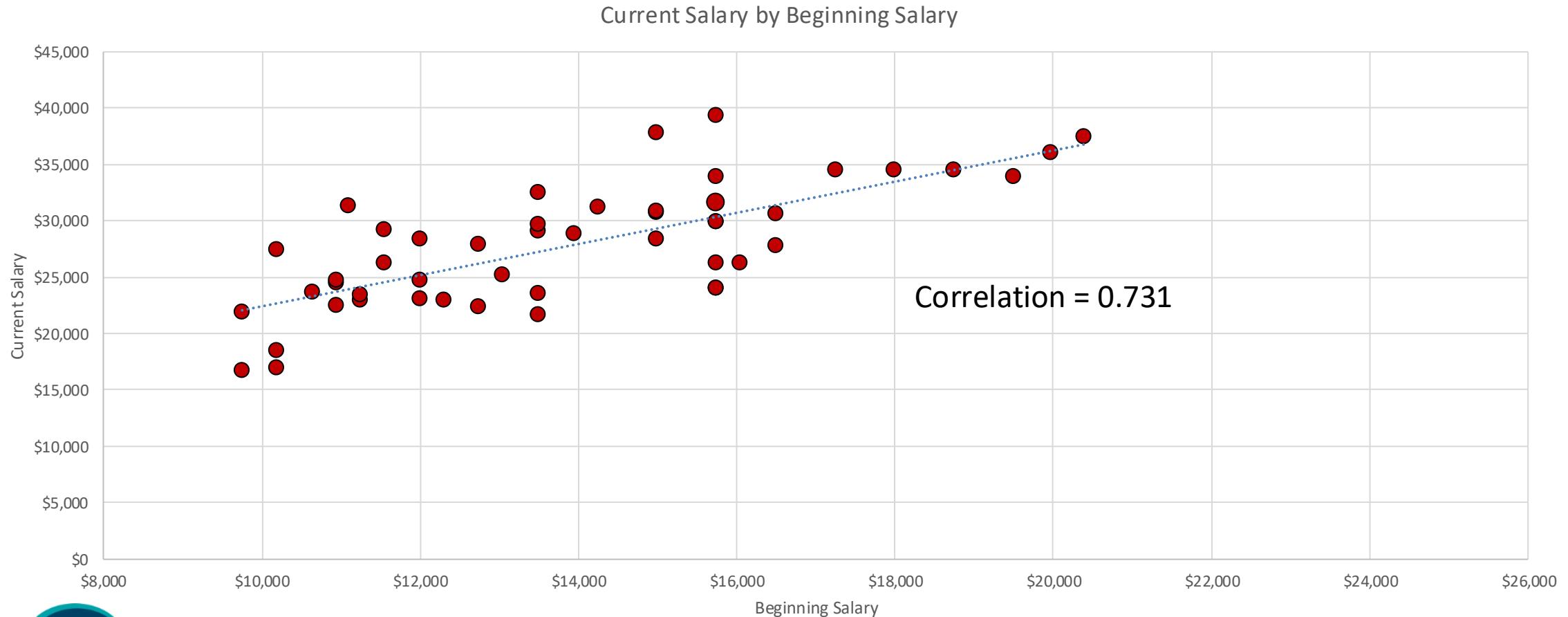
# Are influenced by the range of values in the sample



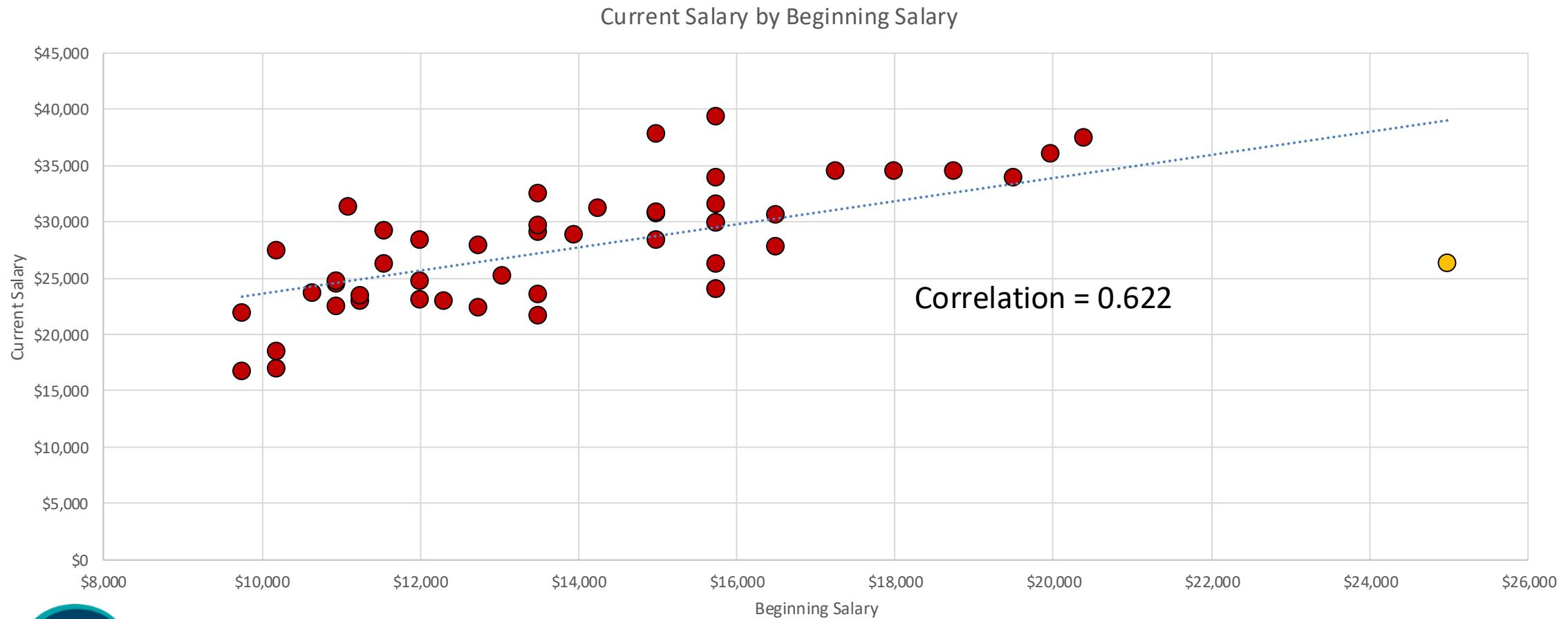
# Are influenced by the range of values in the sample



# Can be unduly affected by extreme/outlier values



# Can be unduly affected by extreme/outlier values



https://chatspss.com

NEW CONVERSATION

CONVERSATIONS

How can I remove unwanted LOG output?

How can I remove unwanted LOG output

Whats the difference between ANOVA and ANCOVA

To remove unwanted LOG output in IBM SPSS Statistics, you can hide the log messages in the Viewer. Here's how to do it:

1. From the menus, choose: **Edit > Options...**
2. Click the **Viewer** tab.
3. Double-click the **Log** icon to hide log messages in the output.
4. Click **OK** to save your changes and close the dialog box.

This will hide the log output from the Viewer, helping to declutter your output display.

Whats the difference between ANOVA and ANCOVA?

## Difference Between ANOVA and ANCOVA

ANOVA (Analysis of Variance) and ANCOVA (Analysis of Covariance) are both statistical methods used to compare the means of different groups. However, they differ in their approach and application:

- **ANOVA:** This method is used to compare the means of three or more groups to see if there are any statistically significant differences between them. It does not take into account any other variables that might influence the dependent variable.
- **ANCOVA:** This method is an extension of ANOVA that includes one or more continuous independent variables, known as covariates. ANCOVA adjusts the means of the groups based on the covariates, which can help reduce the error variance and increase the power of the test.

In summary, ANCOVA is used when you want to control for the effects of one or more continuous variables (covariates) while comparing the means of different groups, whereas ANOVA is used when you only want to compare the means without considering any other variables.

For more detailed information, please consult external sources or the official SPSS documentation.

### Training

- Training notes
- In person training
- Online self-paced training courses

### Consulting

- SPSS Boost
- One-to-one guided consultancy

### Support

- Technical support
- Frequently asked questions
- ChatSPSS – AI-Powered SPSS Tutor

## Meet ChatSPSS – Your AI-Powered SPSS Tutor

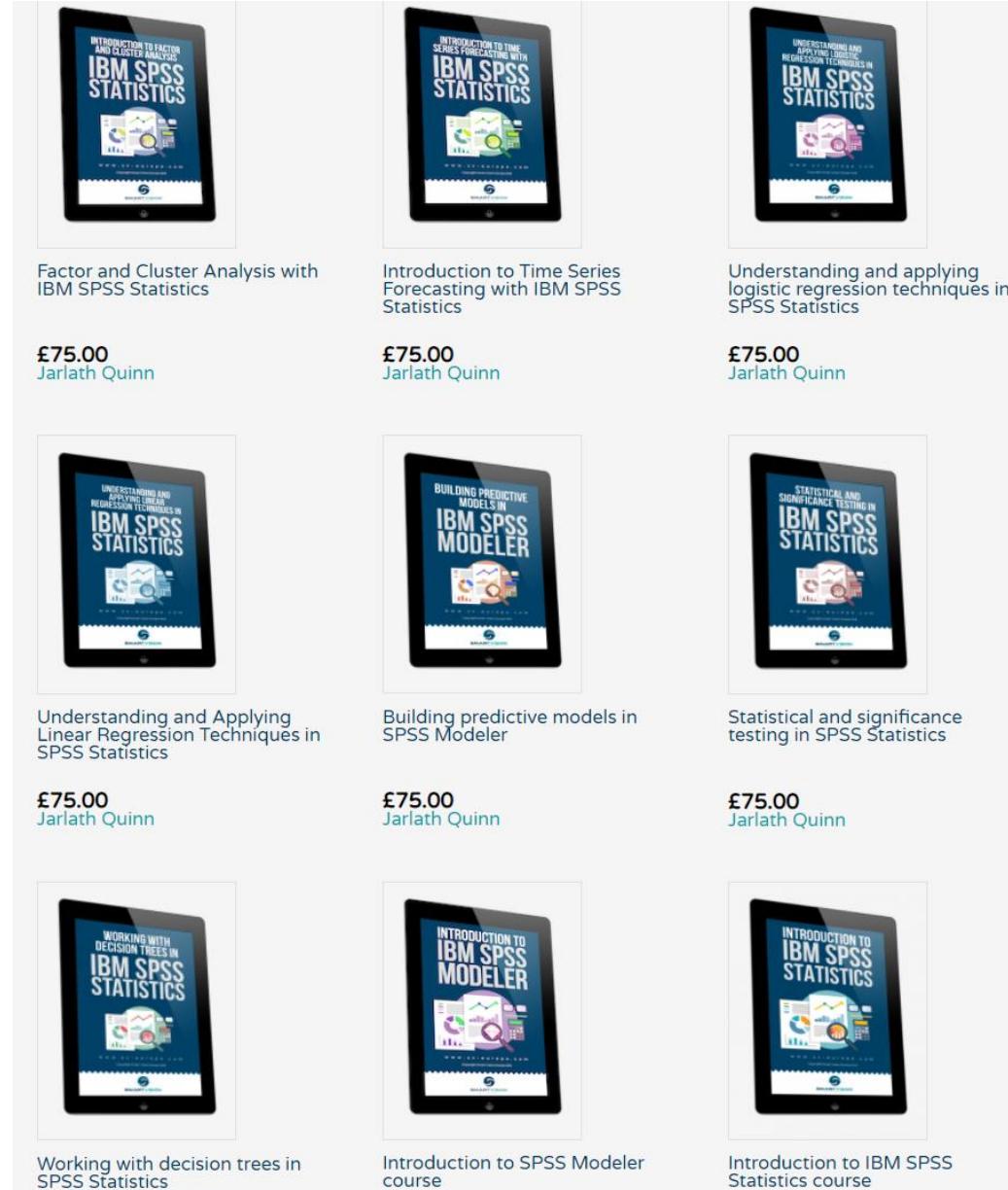
Get instant, accurate answers to your SPSS and statistics questions – anytime you need them.

Smart Vision Europe is proud to introduce **ChatSPSS**, the first AI-enabled tutor designed specifically for SPSS users. Whether you're new to SPSS or an experienced analyst, ChatSPSS helps you get the very most out of your software, saving you time and boosting your confidence with data analysis.

### What can ChatSPSS do for you?

- Guide you through loading and preparing data in SPSS
- Show you how to summarise results and run statistical tests
- Explain outputs from regression models and other advanced techniques
- Answer almost any SPSS or statistics-related question, on demand

Online training materials  
free to Smart Vision  
customers or available for  
purchase



# Smart Vision Europe: Services and Expertise



We have decades of experience providing guidance, training and consultancy in the delivery of effective data science initiatives.



**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





Contact us:

+44 (0)207 786 3568

[info@sv-europe.com](mailto:info@sv-europe.com)

Twitter: @sveurope



[Follow us on Linked In](#)



[Sign up for our Newsletter](#)

# Thank you