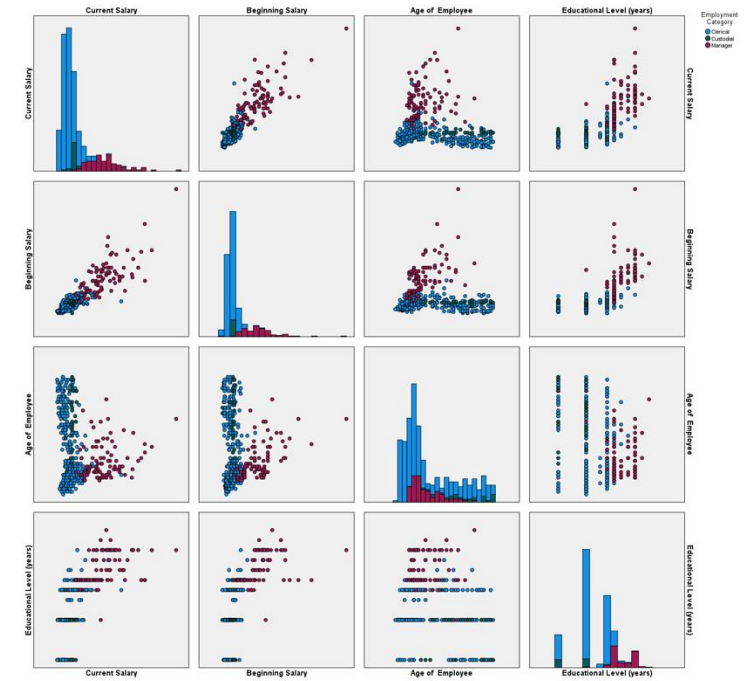
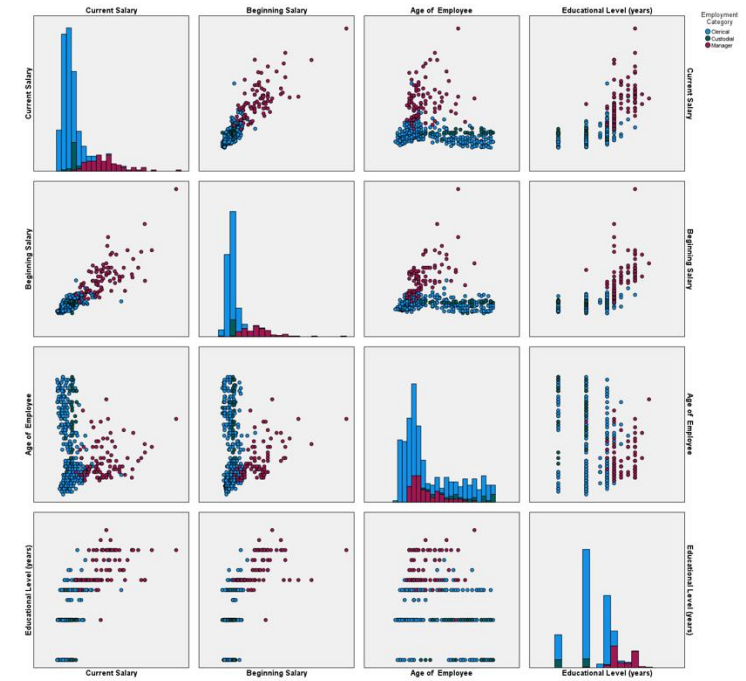




Correlation analysis with SPSS Statistics

Jarlath Quinn

www.sv-europe.com



Correlation analysis with SPSS Statistics

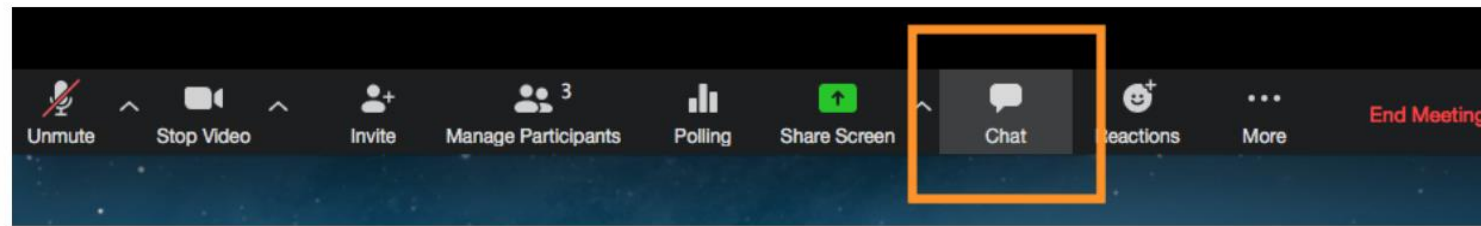
Just waiting for all attendees to join...

Jarlath Quinn

www.sv-europe.com

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 EDB specialising in advanced analytics & AI technologies
- Work with open-source technologies (R, Python 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
 - Healthcare/Pharma
 - Finance/Insurance
 - Media/Telecoms
 - Utilities
 - FMCG
 - Charity/Housing/Government



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



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).

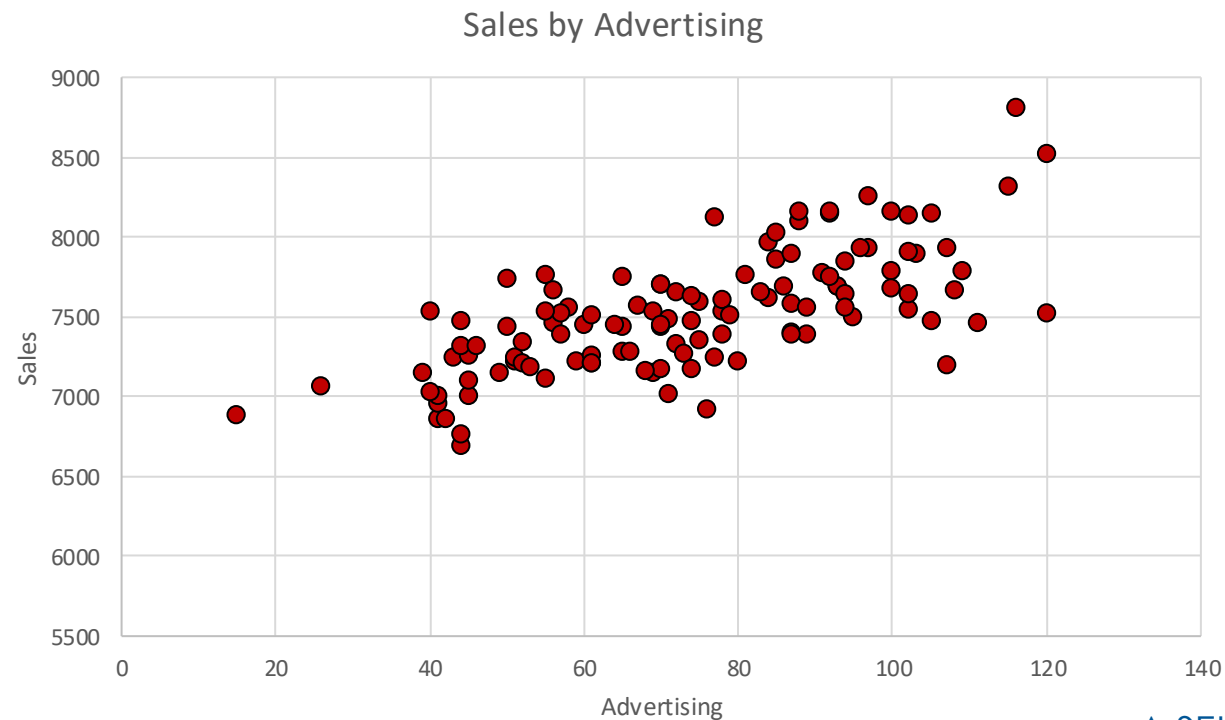
Why are correlations useful?

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.





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

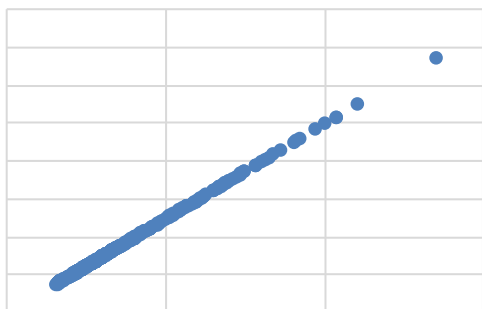
+0.5

0

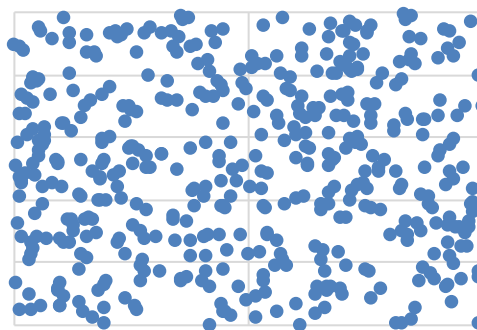
-0.5

-1

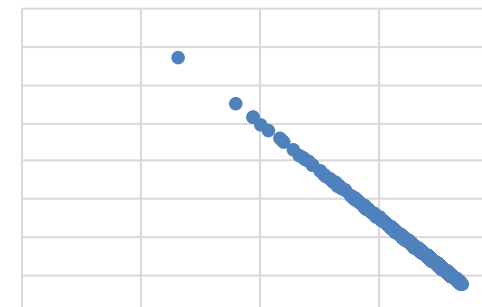
Perfect Positive Linear Relationship



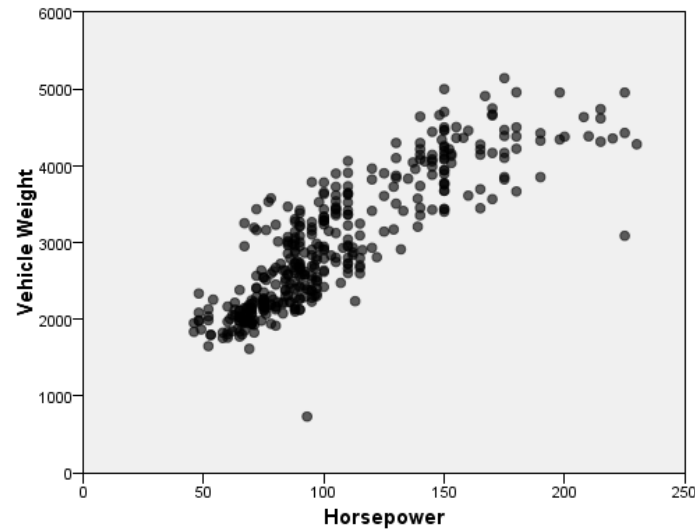
No Linear Relationship



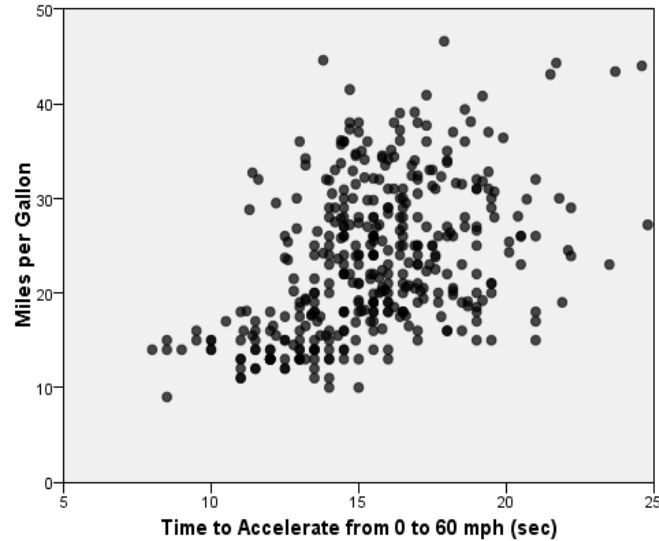
Perfect Negative Linear Relationship



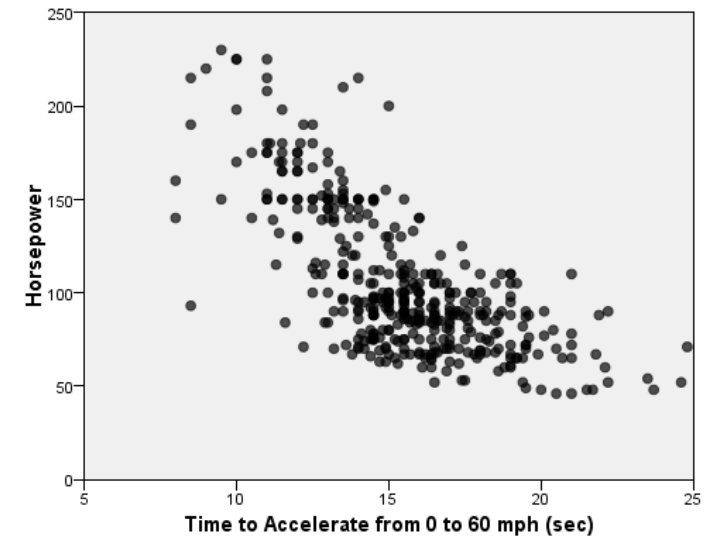
Pearson's r correlations



0.859



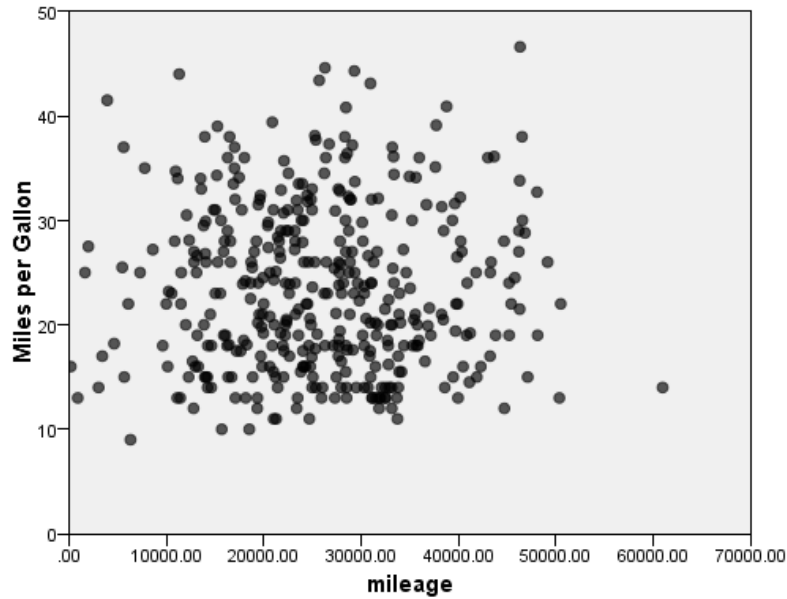
0.434



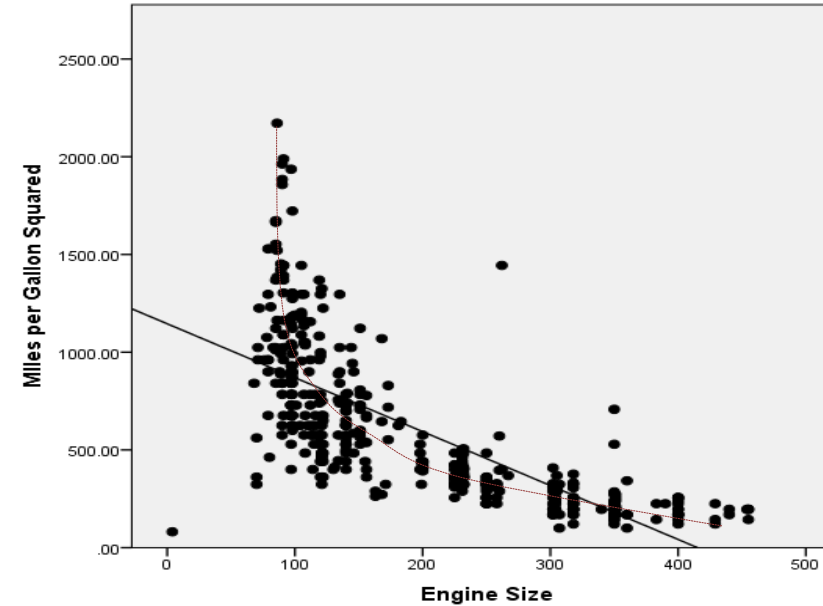
-.701

Pearson's r correlation coefficients

Non-Linear Relationships



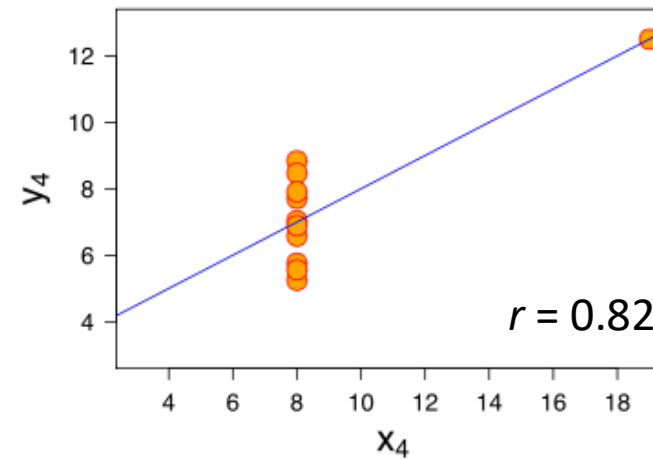
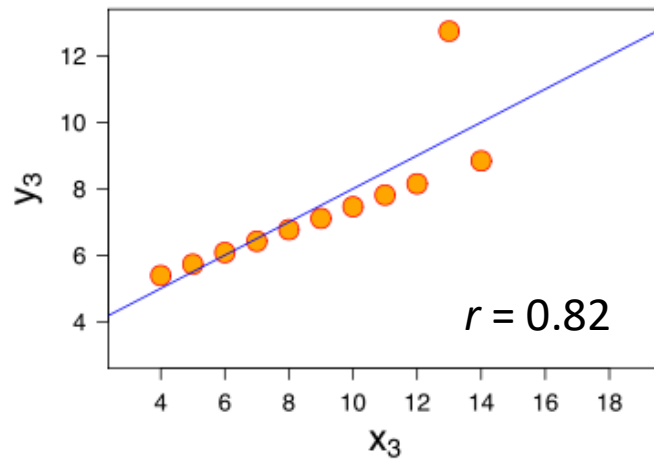
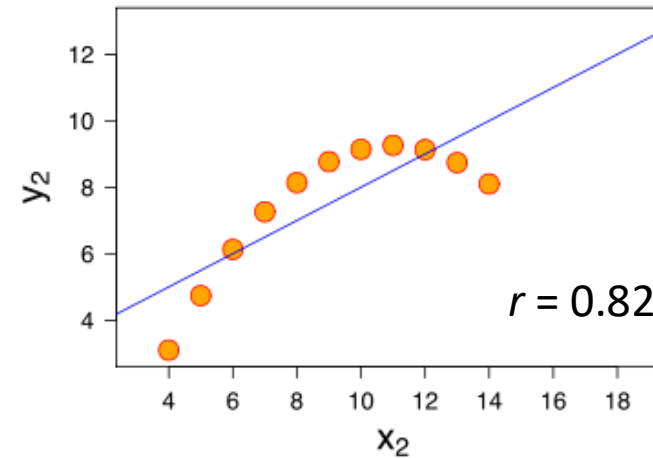
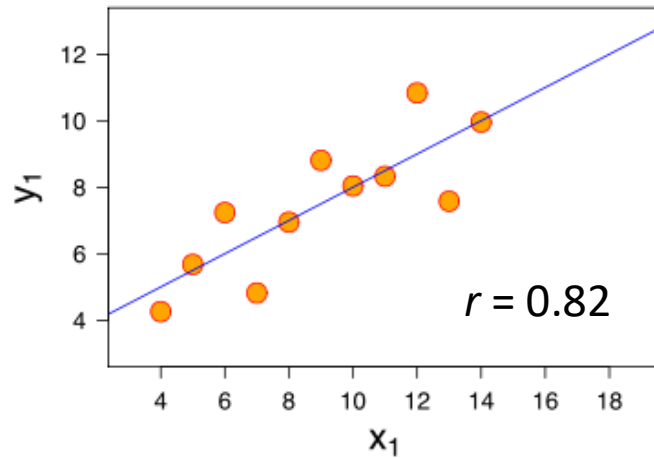
-0.005



-.671

Pearson's r Correlations

A word of warning: always investigate the relationship



Example SPSS Correlations

- Analyze
 - Correlate
 - Bivariate

Employee with age.sav [DataSet] - IBM SPSS Statistics Data Editor

File Edit View Data Transform **Analyze** Graphs Custom Utilities Extensions Window Help

Power Analysis
Meta Analysis
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means and Proportions
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Spatial and Temporal Modeling...
Direct Marketing

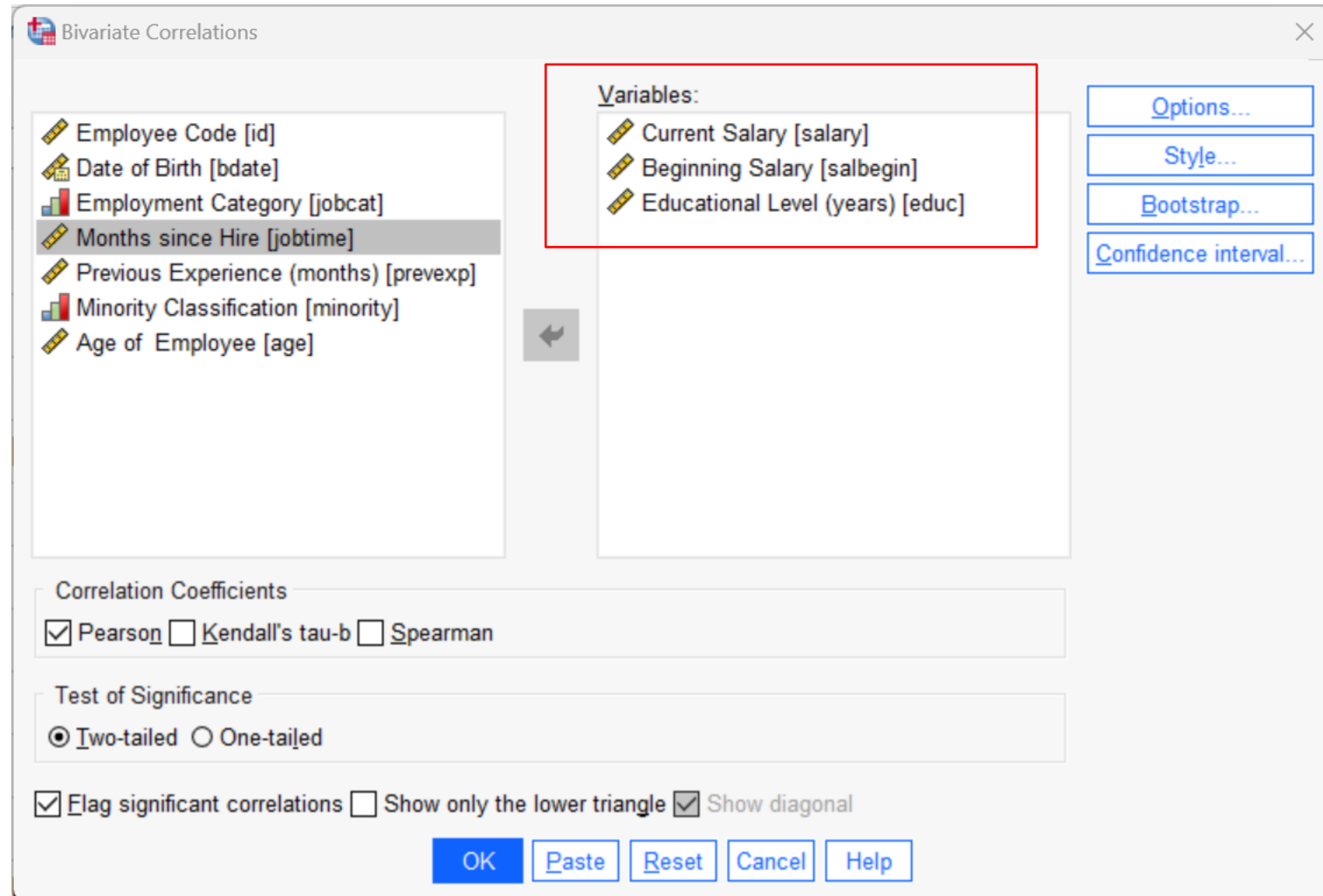
educ jobcat salary salbegin

1	434 Male	16	Clerical	\$34,950	\$20,25	
2	2 Male	16	Clerical	\$40,200	\$18,75	
3	6 Male	15	Clerical	\$32,100	\$13,50	
4	8 Female	12	Clerical	\$21,900	\$9,75	
5	12 Male			\$28,350	\$12,00	
6	13 Male			\$27,750	\$14,25	
7	15 Male			\$27,300	\$13,50	
8	16 Male			\$40,800	\$15,00	
9	17 Male	15	Clerical	\$46,000	\$14,25	
10	19 Male	12	Clerical	\$42,300	\$14,25	
11	21 Female	16	Clerical	\$38,850	\$15,00	
12	23 Female	15	Clerical	\$24,000	\$11,10	
13	26 Male	15	Clerical	\$31,050	\$12,60	
14	28 Male	15	Clerical	\$32,550	\$14,25	
15	30 Male	15	Clerical	\$31,200	\$14,25	
16	31 Male	12	Clerical	\$36,150	\$14,25	
17	33 Male	15	Clerical	\$42,000	\$15,00	
18	35 Male	08/22/61	17	Manager	\$81,250	\$30,00
19	36 Female	08/27/62	9	Clerical	\$34,350	\$14,25

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 ^{**}	.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

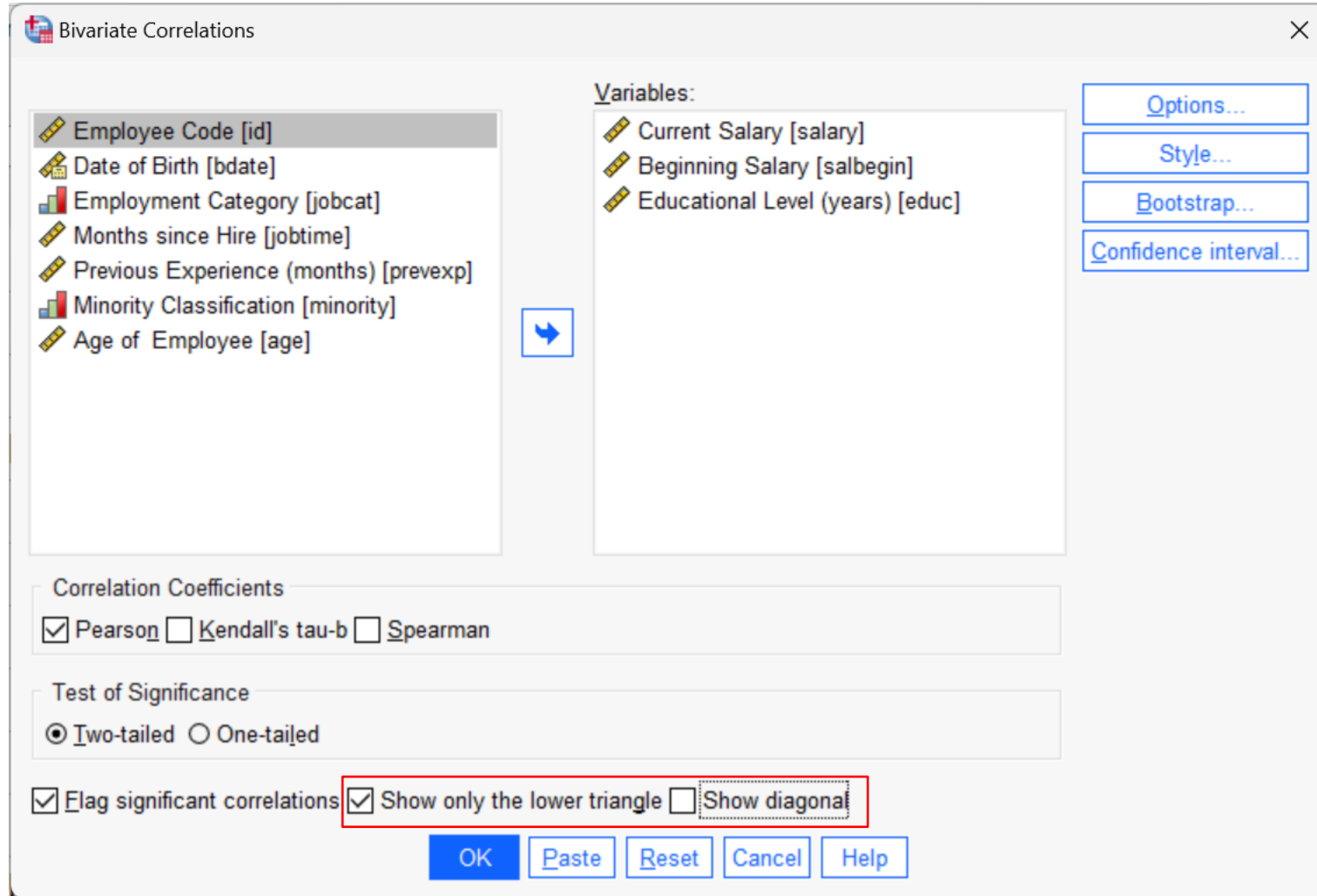
Correlations

		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



The image shows the 'Bivariate Correlations' dialog box in SPSS. On the left, a list of variables includes 'Employee Code [id]', 'Date of Birth [bdate]', 'Employment Category [jobcat]', 'Months since Hire [jobtime]', 'Previous Experience (months) [prevexp]', 'Minority Classification [minority]', and 'Age of Employee [age]'. On the right, under 'Variables:', are 'Current Salary [salary]', 'Beginning Salary [salbegin]', and 'Educational Level (years) [educ]'. A blue arrow button points from the left list to the right list. Below these lists, the 'Correlation Coefficients' section has 'Pearson' checked. The 'Test of Significance' section has 'Two-tailed' selected. At the bottom, 'Flag significant correlations' is checked, and 'Show only the lower triangle' is also checked and highlighted with a red rectangle. To the right of this are buttons for 'Options...', 'Style...', 'Bootstrap...', and 'Confidence interval...'. At the very bottom are 'OK', 'Paste', 'Reset', 'Cancel', and 'Help' buttons.

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



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

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

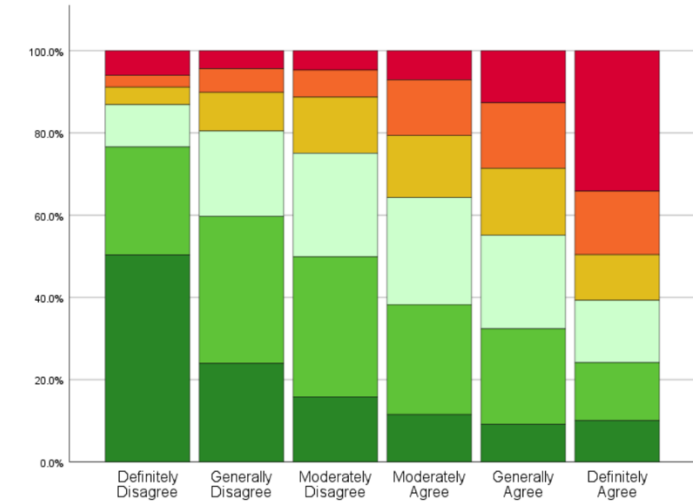
x_i = values of the x-variable in a sample

$\bar{x}$ = 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

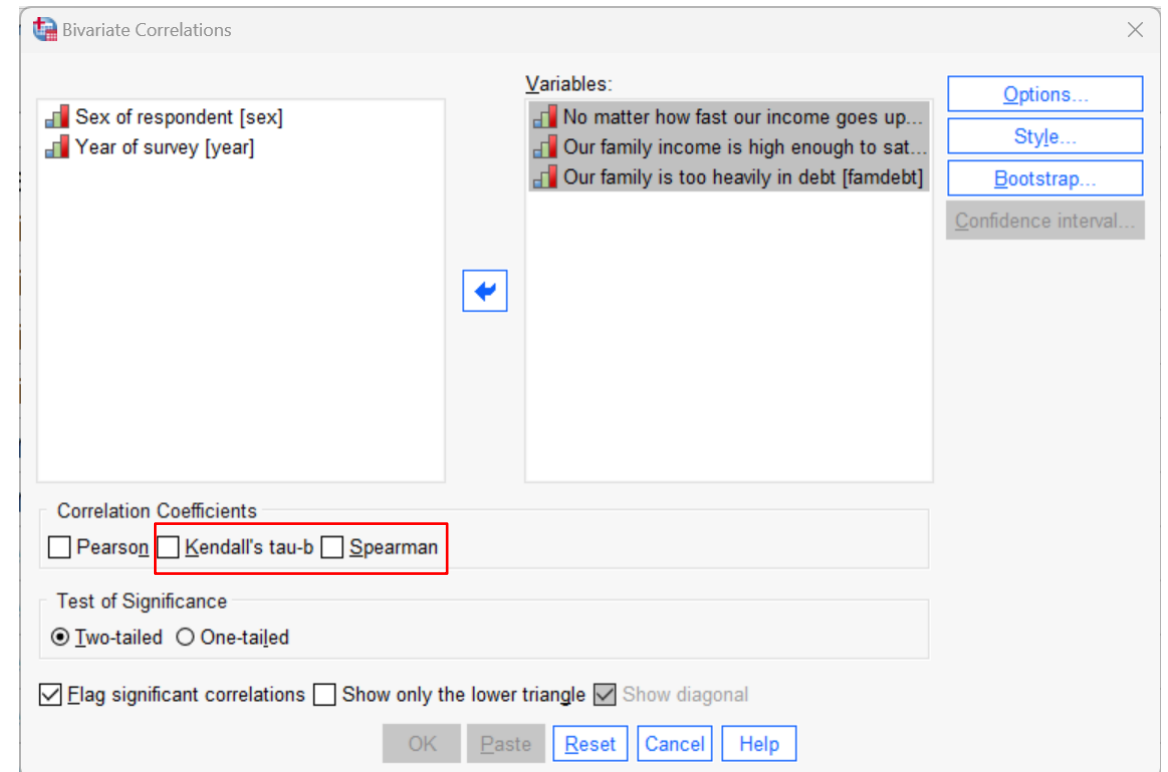
**Let's see an example of calculating
Pearson's R**



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

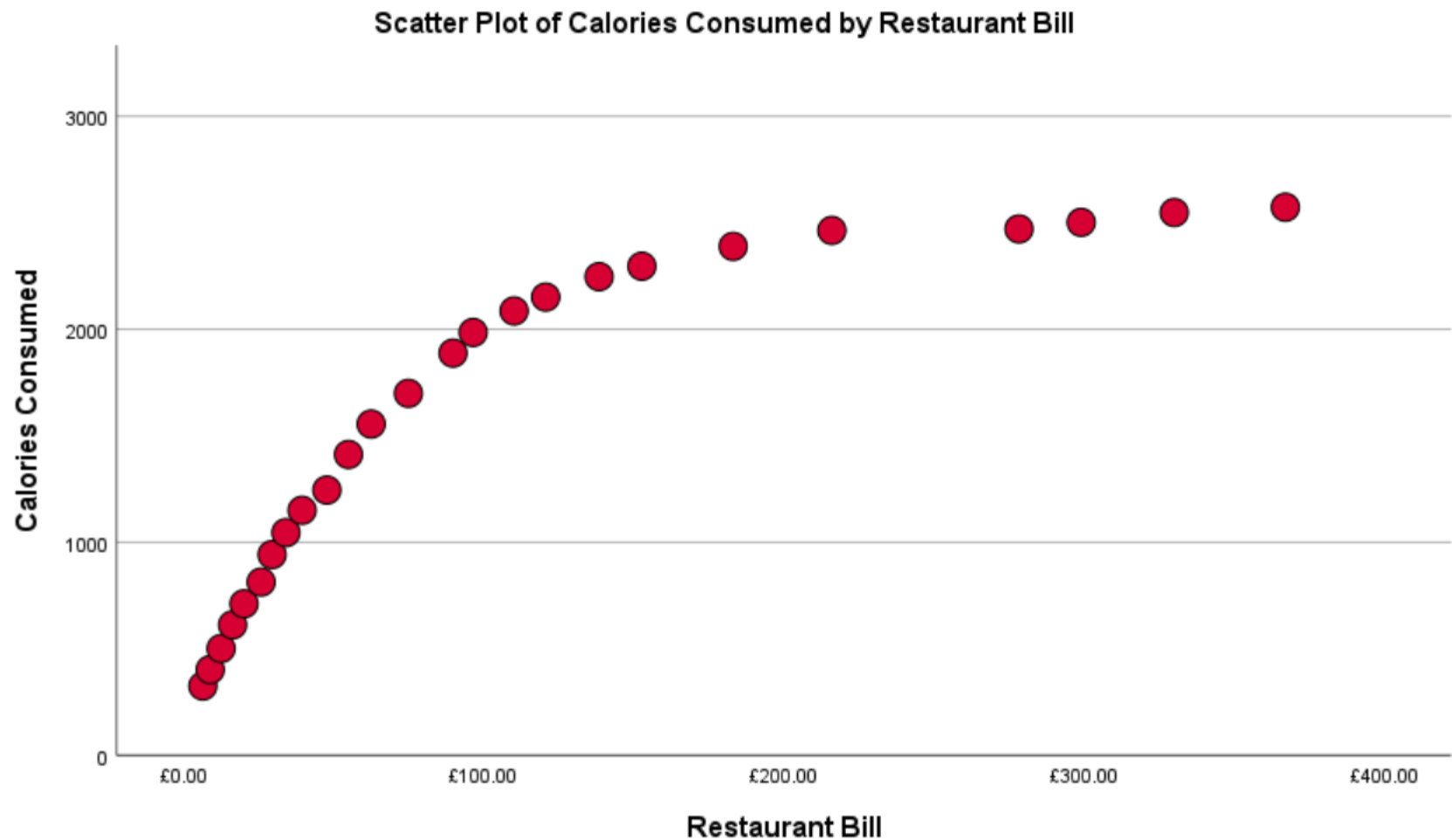


Spearman's Rho

- **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

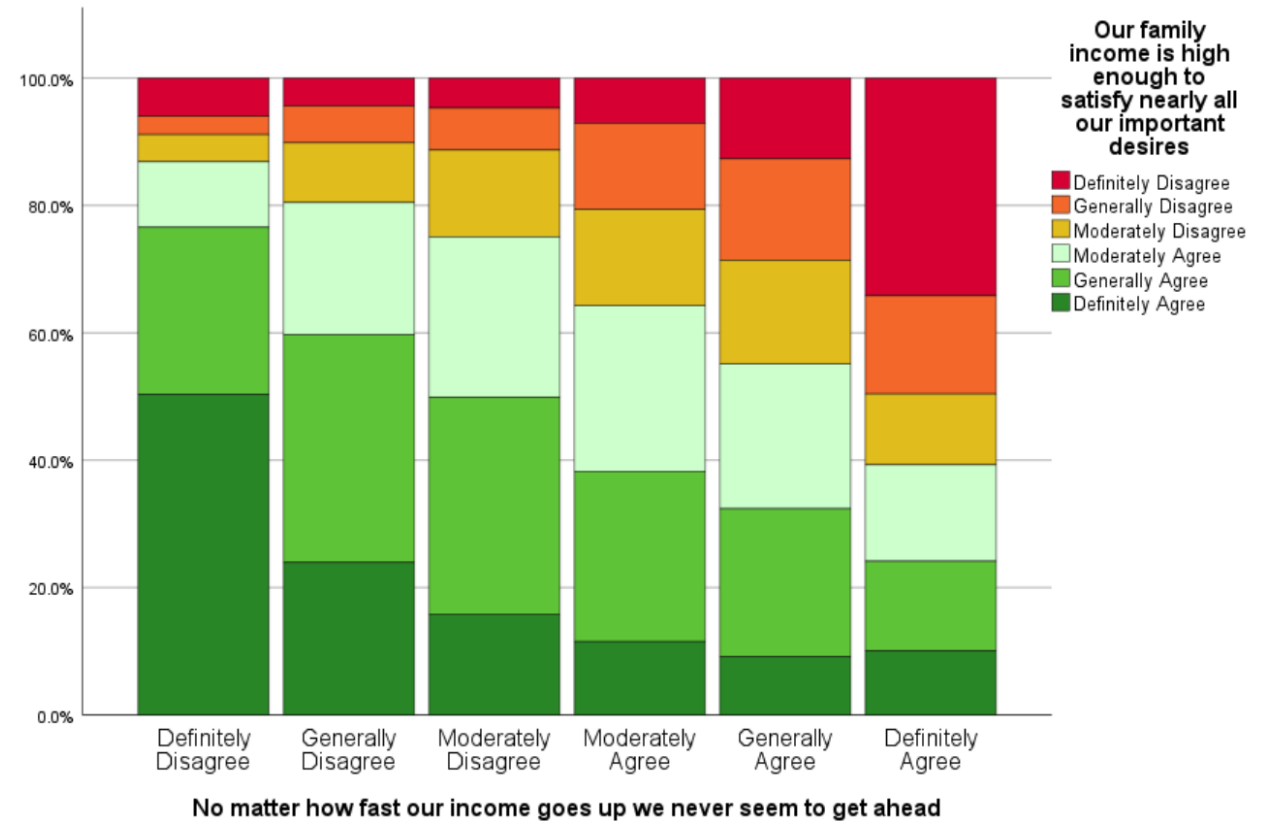
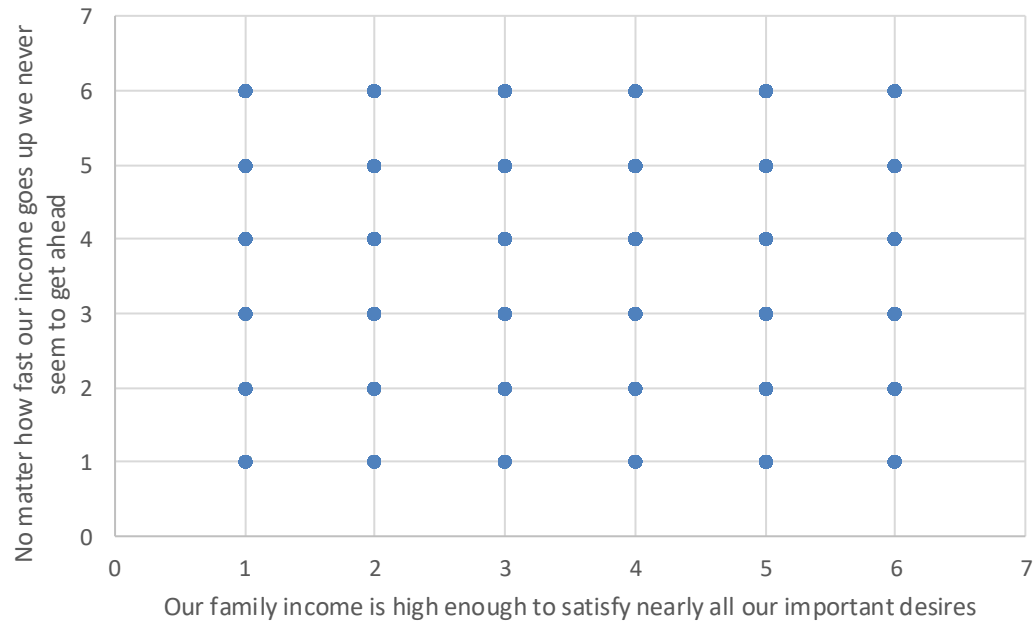
Kendall's Tau

- 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

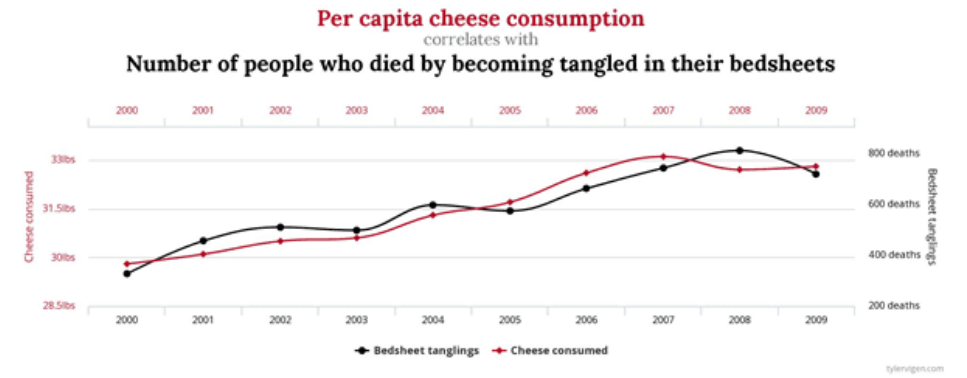
	🎬 Movie_Title	📊 IMDb	🍅 Rotten_Tomatoes
1	Uncorked	1	6
2	Spenser Confidential	2	1
3	The Willoughbys	3	5
4	Tigertail	4	3
5	Extraction	5	2
6	The Half of It	6	8
7	To All the Boys I've Loved Before	7	9
8	LA Originals	8	4
9	Miss Americana	9	7
10	Crip Camp: A Disability Revolution	10	10

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

Visualising Correlations for Ordinal Variables

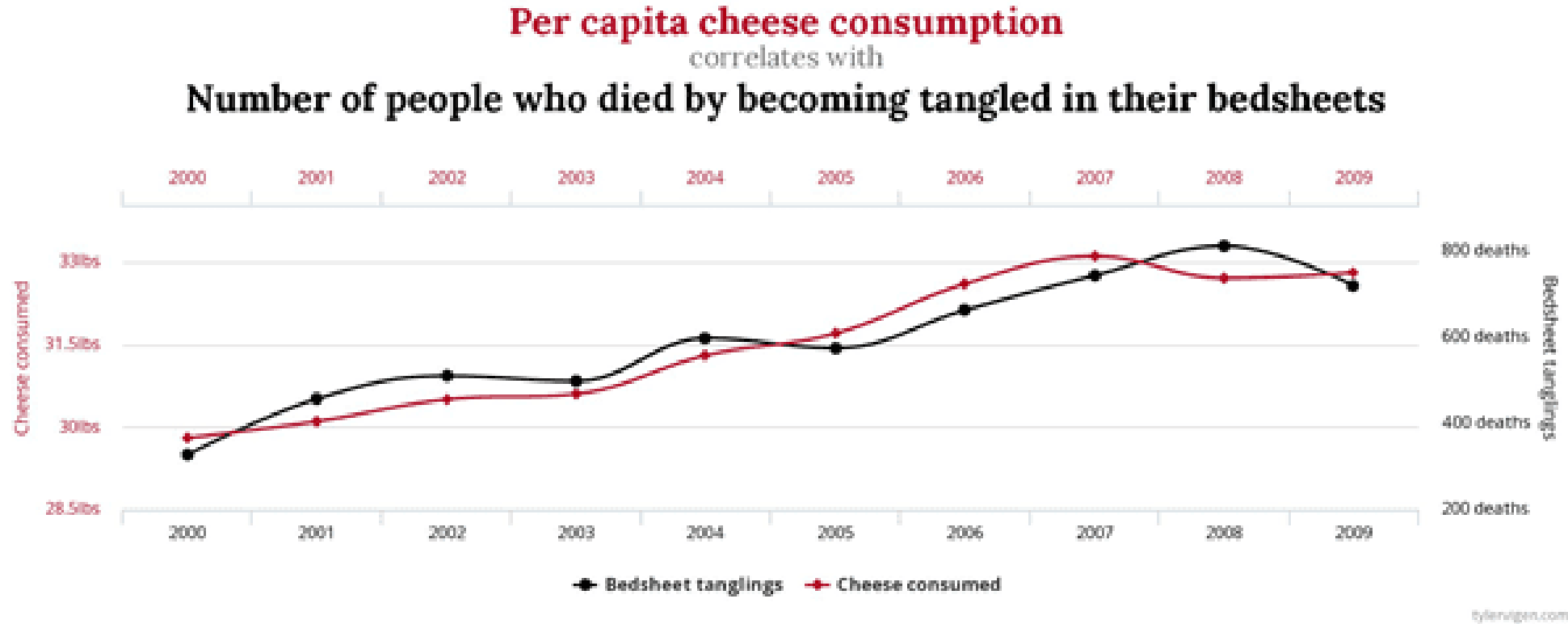


May require a different approach than scatterplots



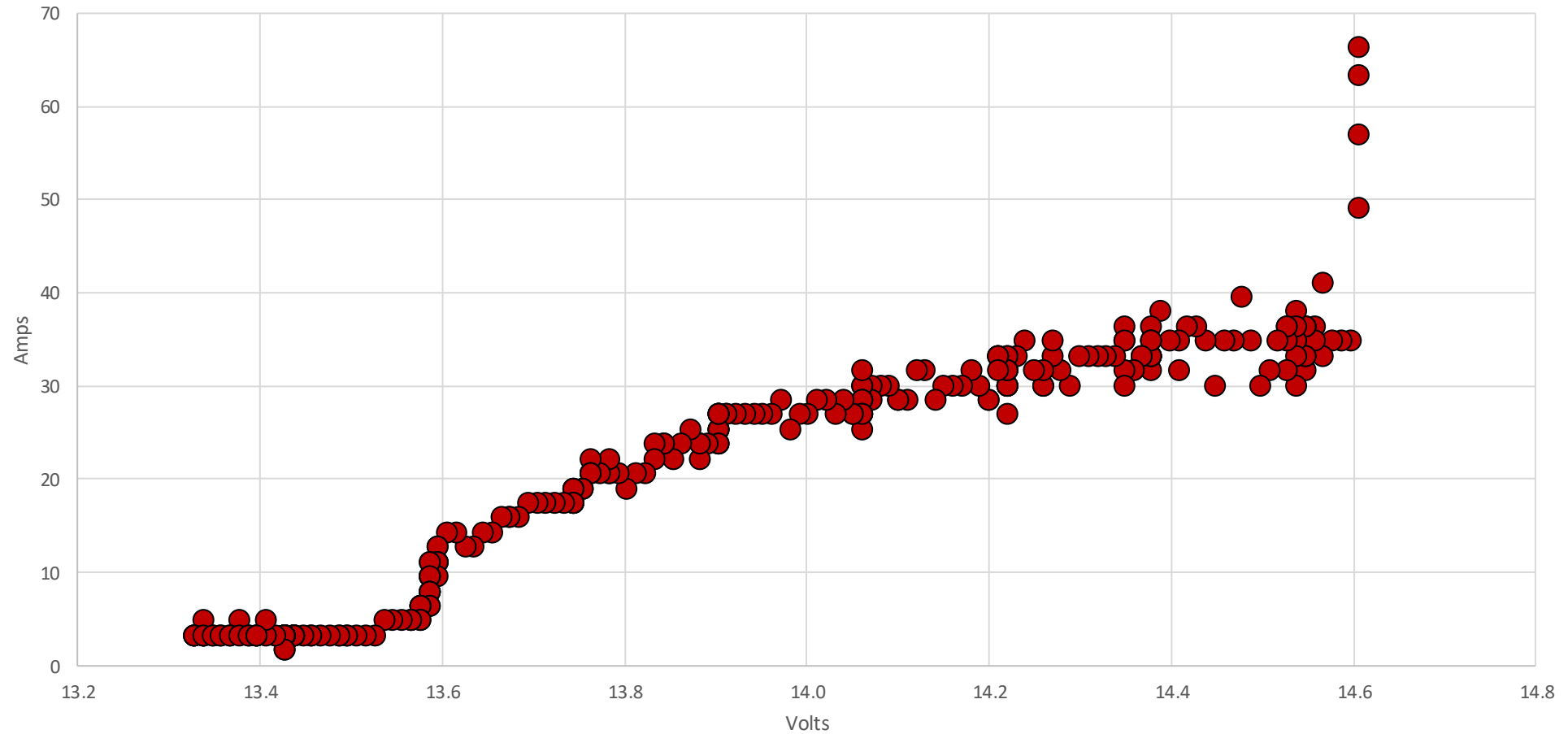
The Limitations of Correlations

Correlation does not indicate causation

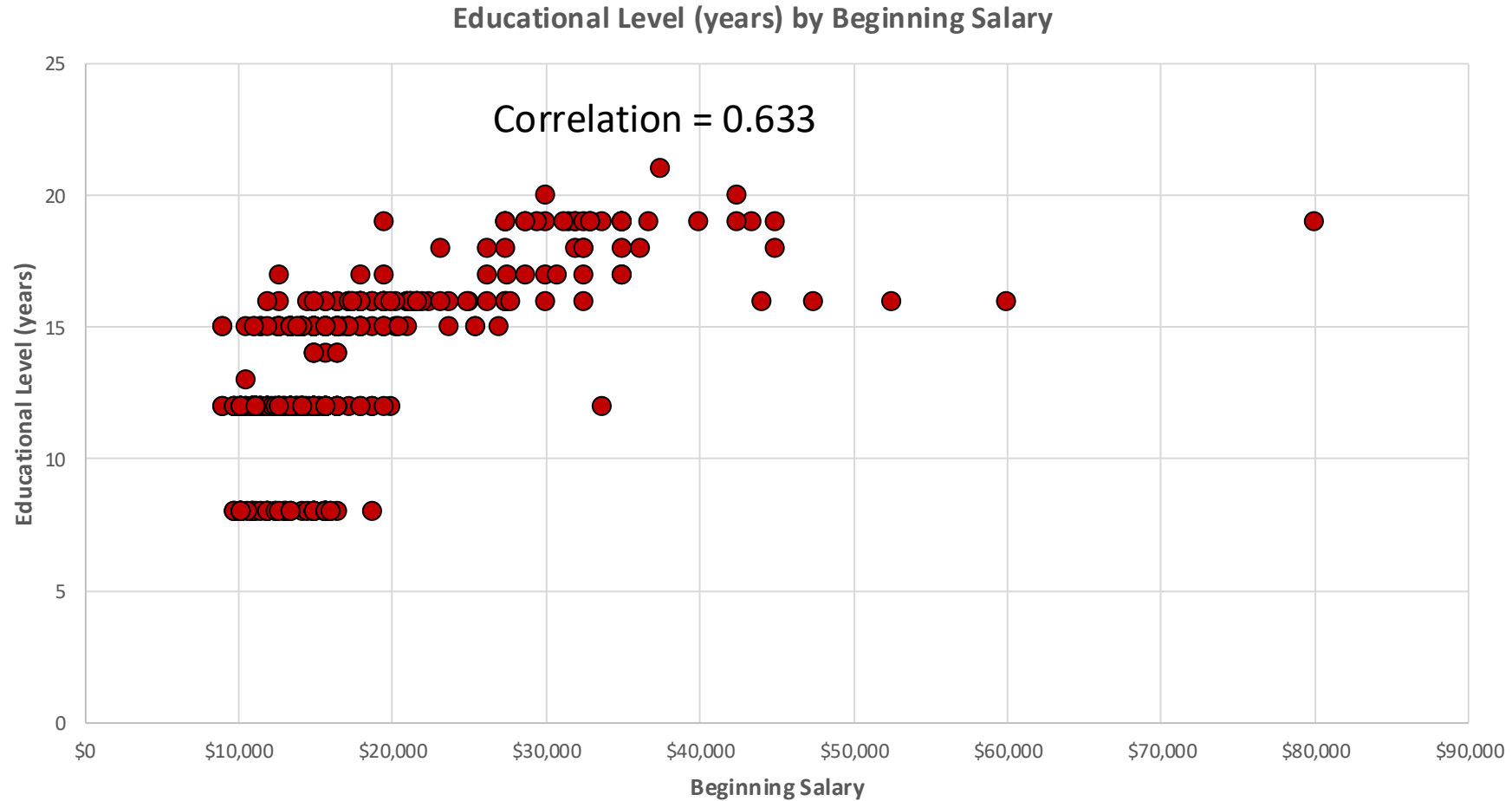


<https://www.productleadership.com/does-causation-imply-correlation/>

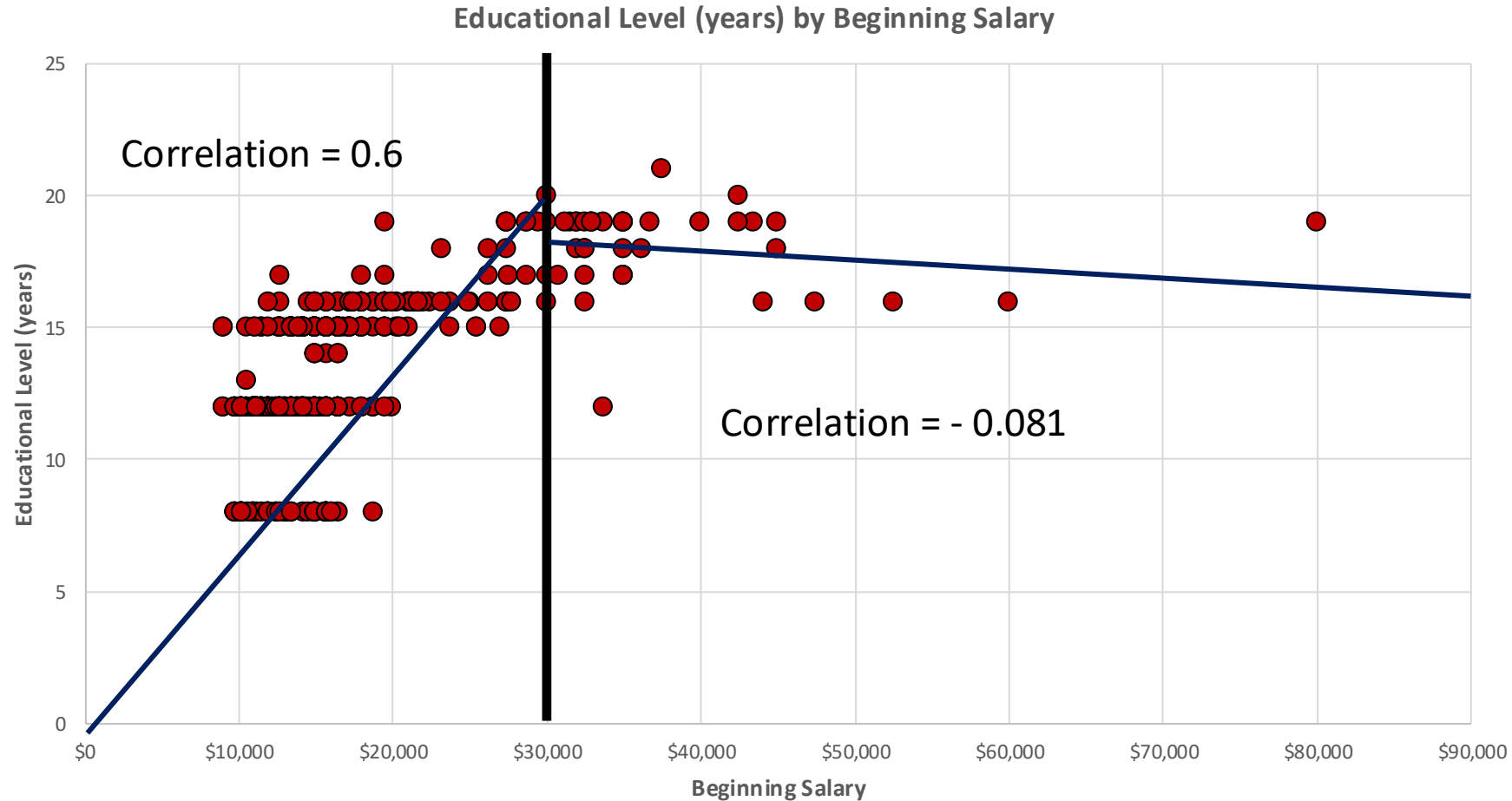
Can't accurately measure curvilinear relationships



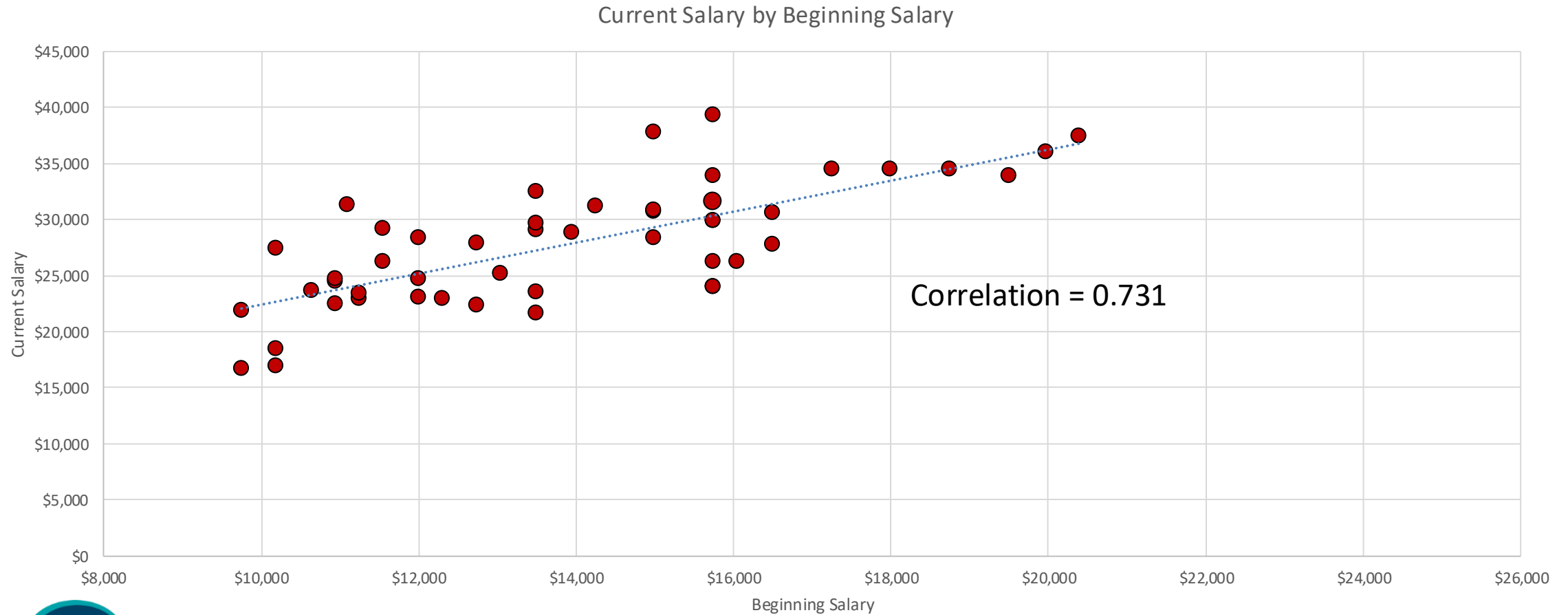
Are influenced by the range of values in the sample



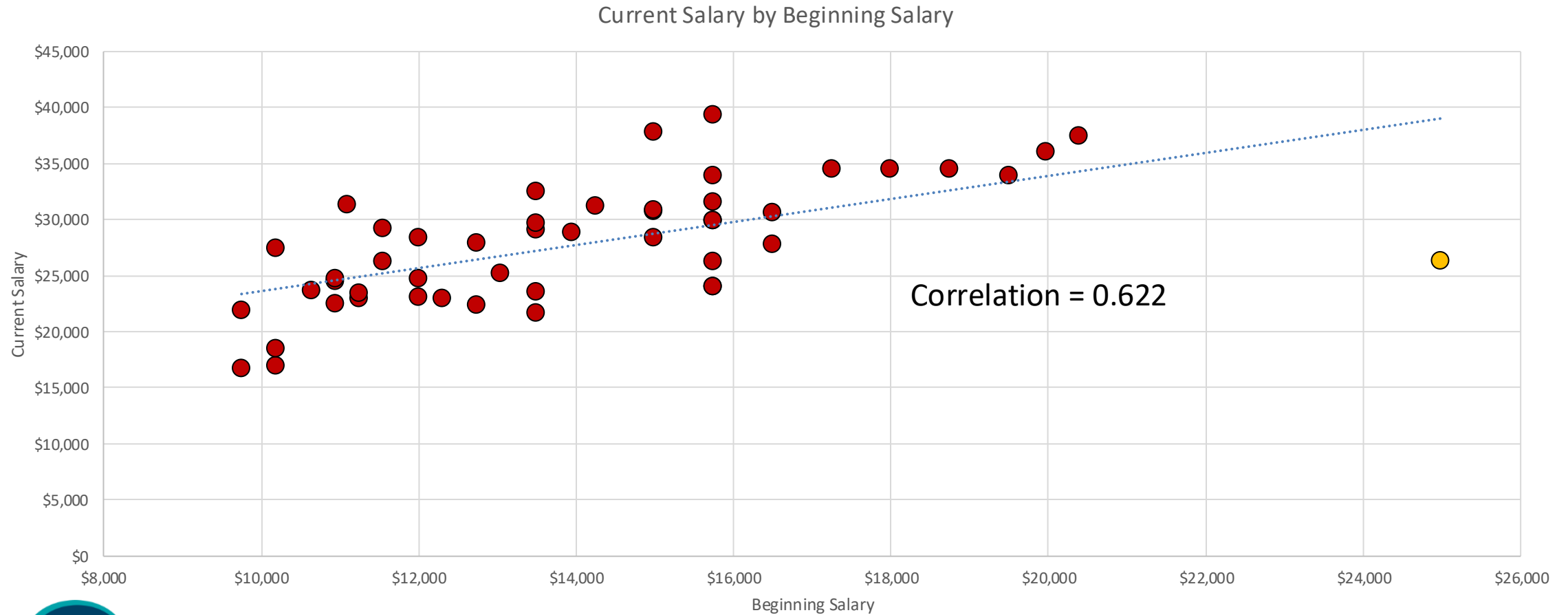
Are influenced by the range of values in the sample



Can be unduly affected by extreme/outlier values



Can be unduly affected by extreme/outlier values



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



Factor and Cluster Analysis with
IBM SPSS Statistics

£75.00
Jarlath Quinn



Introduction to Time Series
Forecasting with IBM SPSS
Statistics

£75.00
Jarlath Quinn



Understanding and applying
logistic regression techniques in
SPSS Statistics

£75.00
Jarlath Quinn



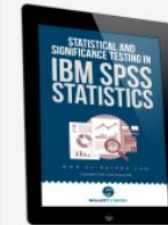
Understanding and Applying
Linear Regression Techniques in
SPSS Statistics

£75.00
Jarlath Quinn



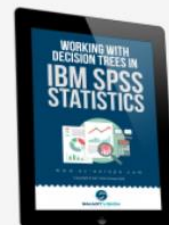
Building predictive models in
SPSS Modeler

£75.00
Jarlath Quinn

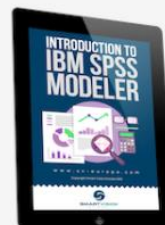


Statistical and significance
testing in SPSS Statistics

£75.00
Jarlath Quinn



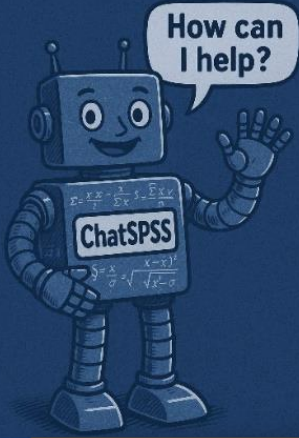
Working with decision trees in
SPSS Statistics



Introduction to SPSS Modeler
course



Introduction to IBM SPSS
Statistics course



ChatSPSS is a free generative AI chatbot developed by Smart Vision Europe Ltd based on official IBM documentation of SPSS procedures

A screenshot of the ChatSPSS web interface. The browser address bar shows "https://chatssp.com". The page has a dark blue header with "NEW CONVERSATION", "EXPORT PDF", "LOG OUT", "PRIVACY", and "SMARTVISION". The main content area shows a conversation. On the left, under "CONVERSATIONS", there are two messages: "How can I recode a string variable called region to create a numeric equivalent?" and "How can I recode a string variable called region to create a numeric equivalent?". On the right, a blue bubble contains the question "How can I recode a string variable called region to create a numeric equivalent?". Below this, the response says: "To recode a string variable called region to create a numeric equivalent, you can use the RECODE command with the INTO keyword." followed by "Steps to Recode the Variable" and a list of steps: "Use the RECODE command to specify the original values and the new values.", "Use the INTO keyword to specify the target numeric variable.", and "Ensure that the target variable is declared before using it." Below this is "Example Syntax" with the code:

```
RECODE region ('North'=1) ('South'=2) ('East'=3) ('West'=4) INTO region_num.
```

 and a list of explanations: "In this example, the string variable region is recoded into the numeric variable region_num." and "The original values 'North', 'South', 'East', and 'West' are replaced with the numeric values 1, 2, 3, and 4, respectively." Below this is "Alternative Method" with a text input field containing "Alternatively, you" and a "variable." label. Below the input field is the code:

```
AUTORECODE VAI  
/INTO region_num  
/PRINT.
```

 and a blue circular button with an upward arrow.

• NEW: Text Analytics with AI

- A new downloadable tool from Smart Vision Europe
- Powered by IBM Watsonx.ai
- Works seamlessly as part of IBM SPSS Statistics

hotel_reviews.sav [DataSet] - IBM SPSS Statistics

File Edit View Data Transform Power Analysis Meta Analysis Reports Extensions Window Help

Descriptive Statistics

	Guest_No	Gender
1	ID121012051...	Female
2	ID121014749...	Female
3	ID121032255...	Female
4	ID121036484...	Female
5	ID121038641...	Female
6	ID121046246...	Female
7	ID121052299...	Female
8	ID121053288...	Female
9	ID121054017...	Female
10	ID121062212...	Female
11	ID121062243...	Female
12	ID121073191...	Male
13	ID121083012...	Male
14	ID121094368...	Female
15	ID121104061...	Male
16	ID121108430...	Male
17	ID121121398...	Female
18	ID121131058...	Female
19	ID121138746...	Female
20	ID121153743...	Male

Watsonx.ai Text Categoriser: Demo

Main Files and fields Analytics Optional controls

Variables:

Sort: None

Target Text Variable: Guest_Reason_For_This_Score

Target flag variable for contribution analysis (optional):

Rows To Process

☒ First N rows

☐ All rows

Number of rows: 25

Classification mode

☐ Topic only

☐ Topic + sentiment

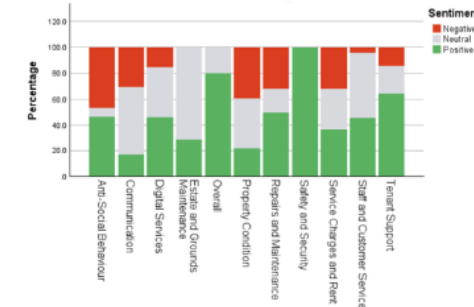
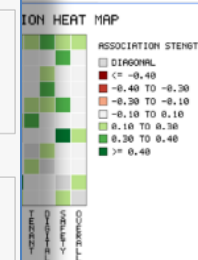
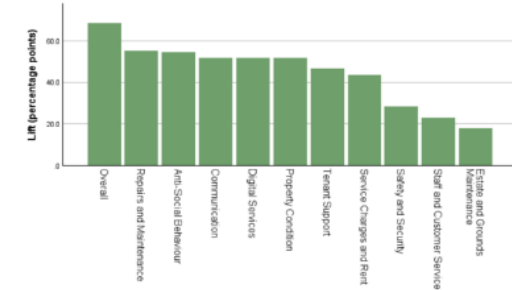
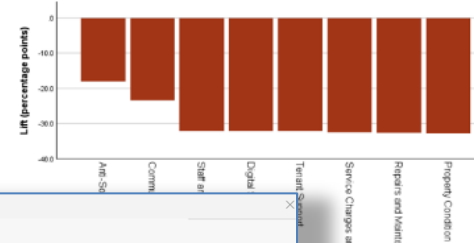
☒ Discovery

Maximum N topics to auto discover: 10

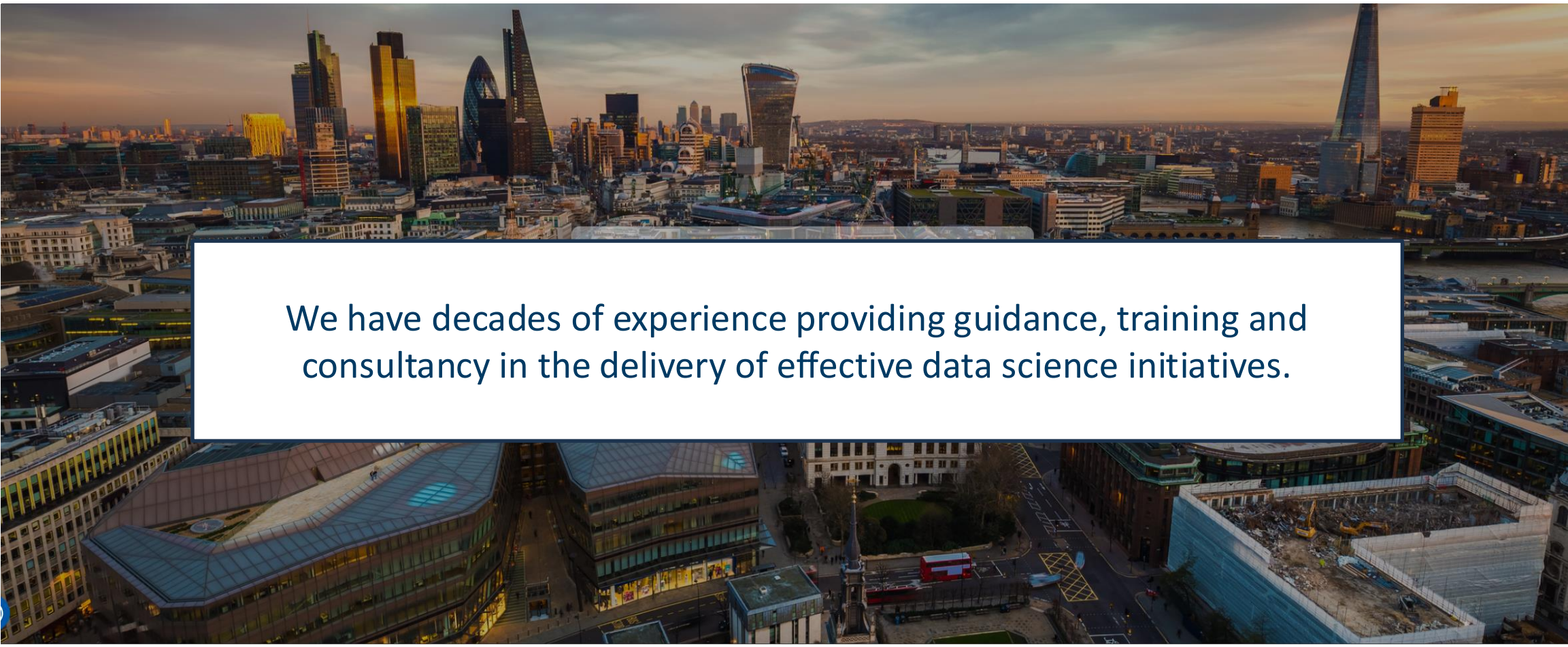
☐ Randomly select Discovery cases

☐ Write the Topic_map.csv file to the working folder

OK Paste Reset Cancel Help



Smart Vision Europe: Services and Expertise



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



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