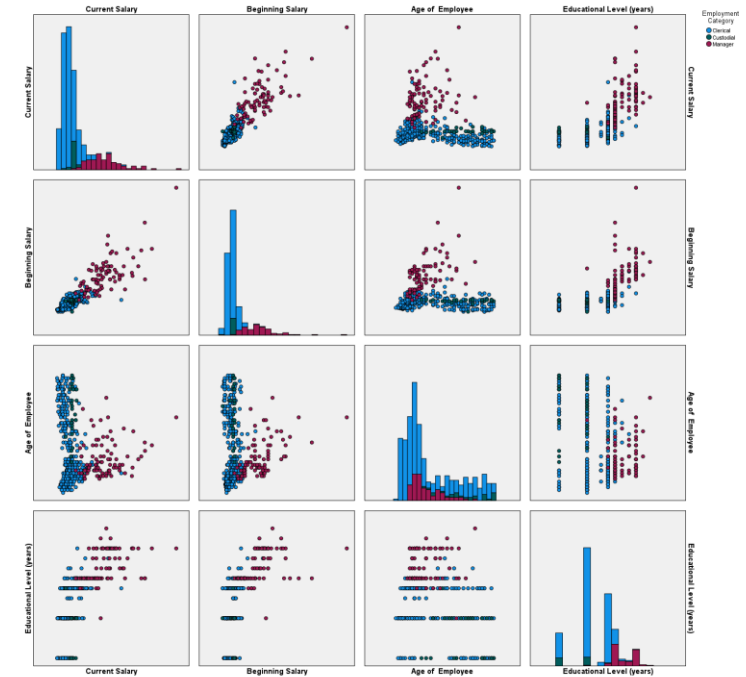


# Correlation analysis with SPSS

Jarlath Quinn – Analytics Consultant



Just waiting for all attendees to join...

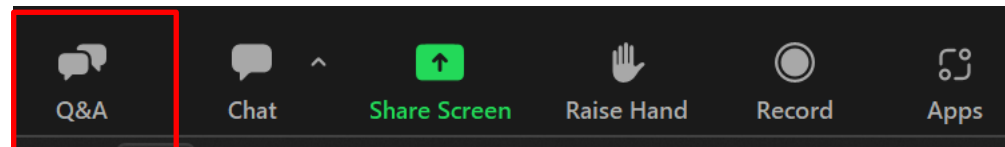


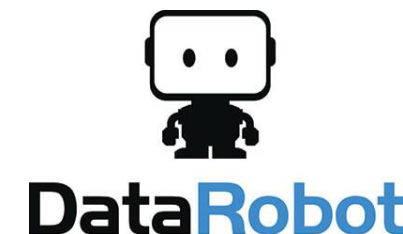
# Correlation analysis with SPSS

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
- Automatically highlighting strong correlations
- How correlations are calculated
- 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).

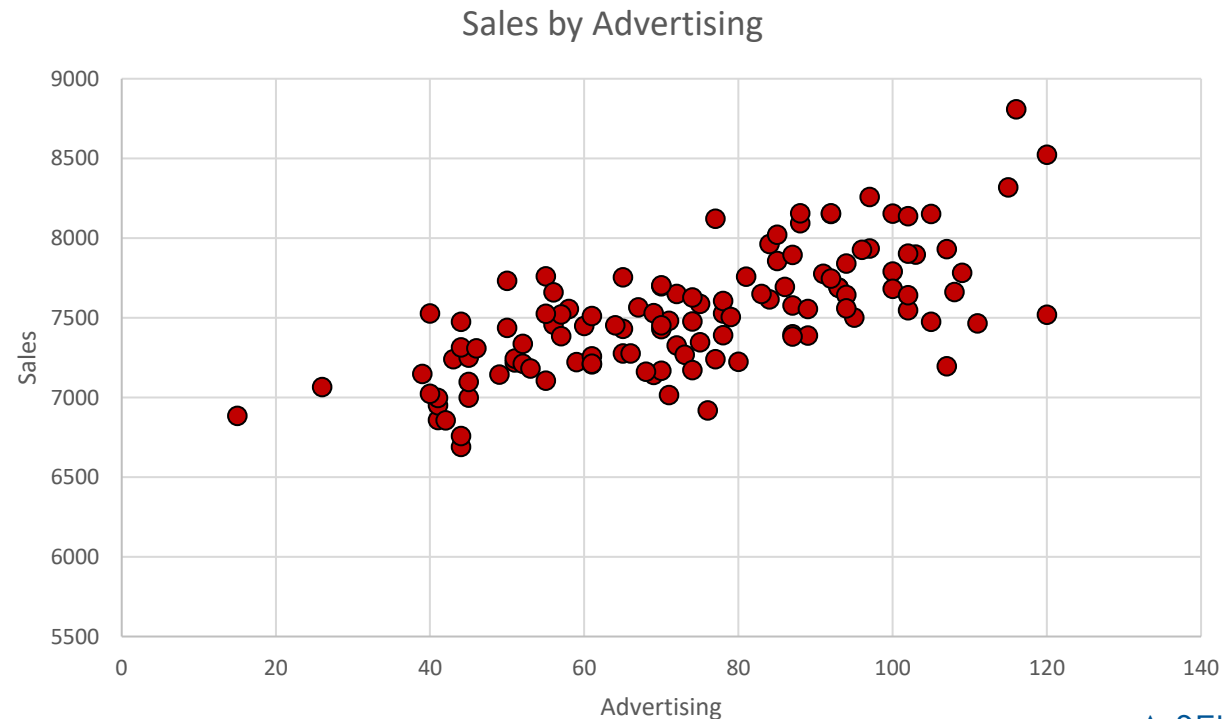
# Why are correlations useful?

# 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.







### 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

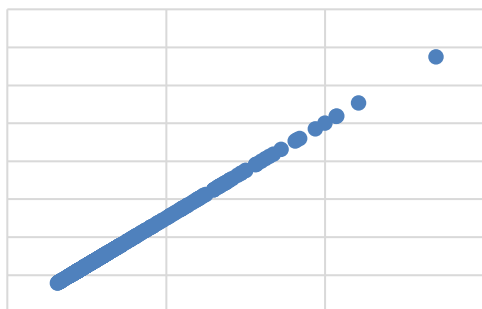
+0.5

0

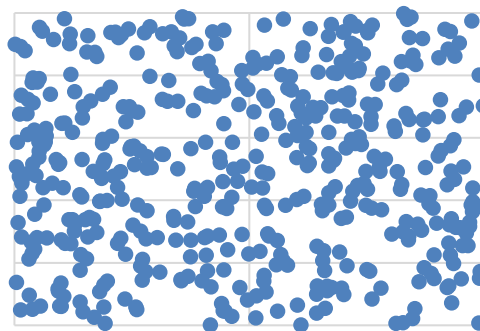
-0.5

-1

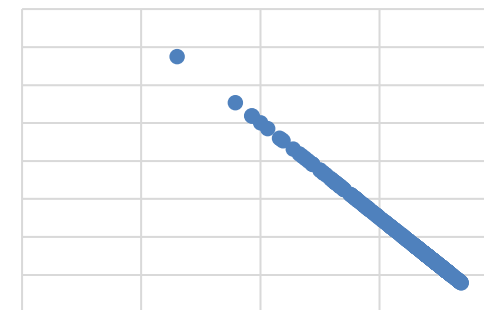
Perfect Positive Linear Relationship



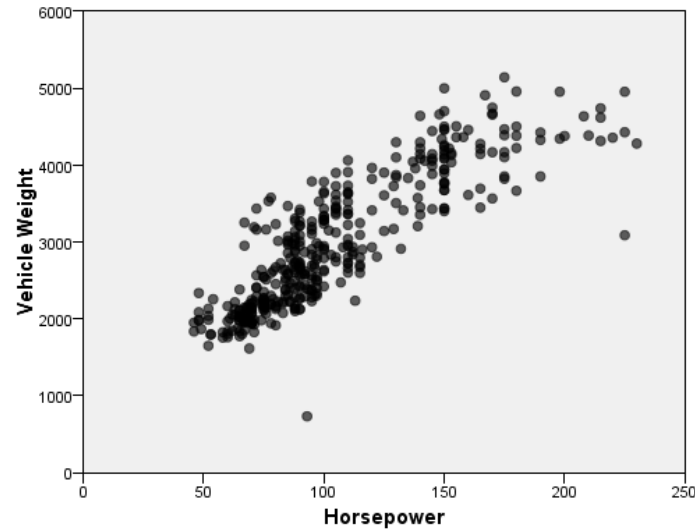
No Linear Relationship



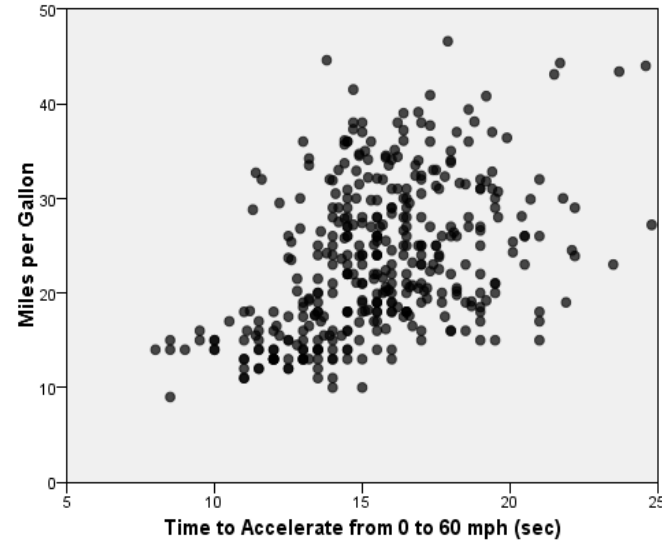
Perfect Negative Linear Relationship



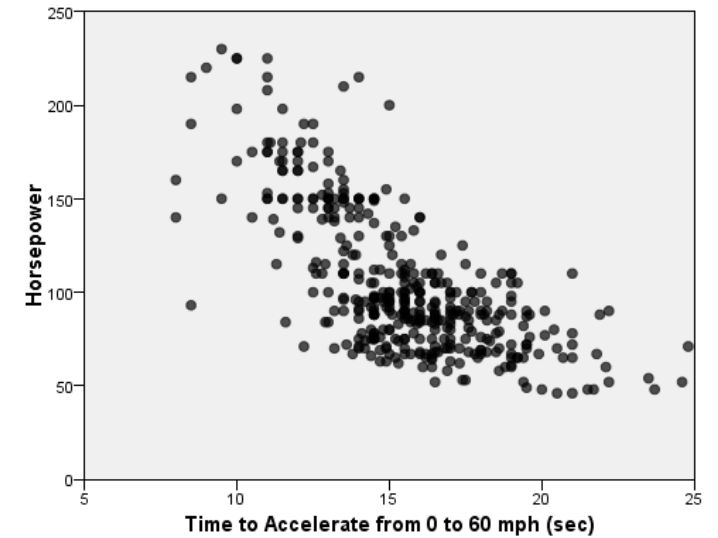
# Pearson's $r$ correlations



0.859



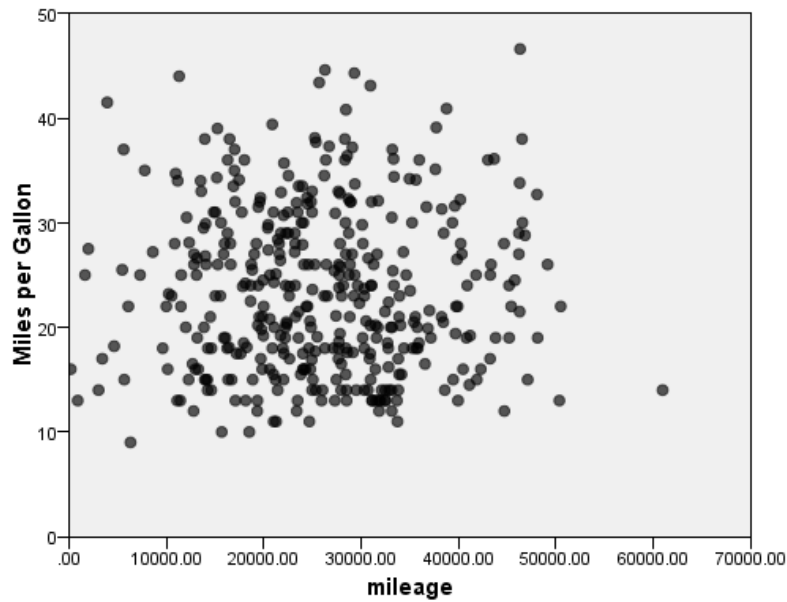
0.434



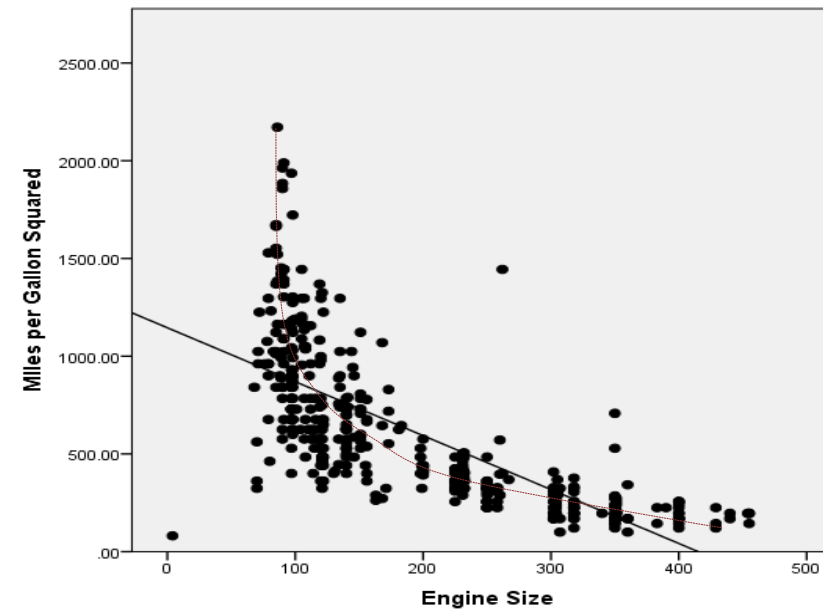
-.701

Pearson's  $r$  correlation coefficients

# Non-Linear Relationships



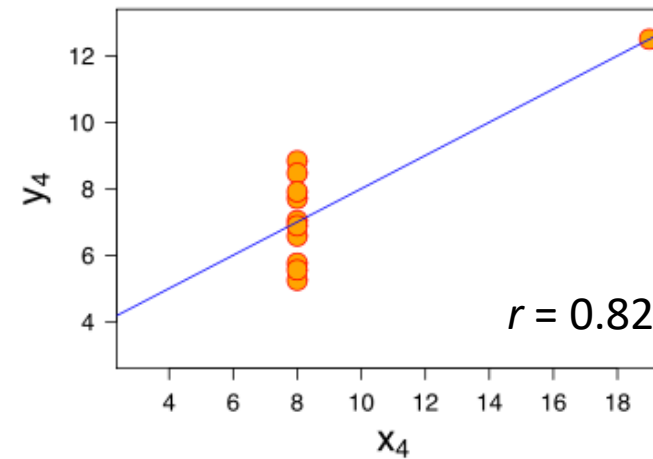
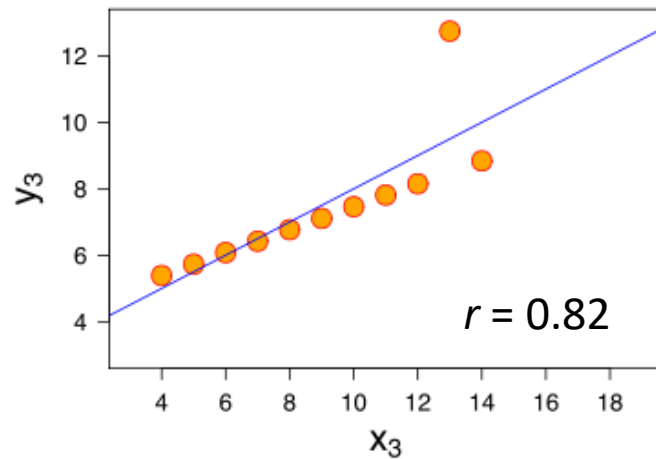
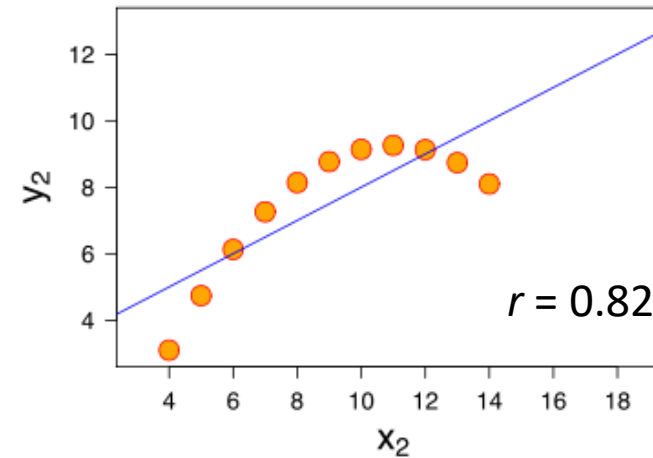
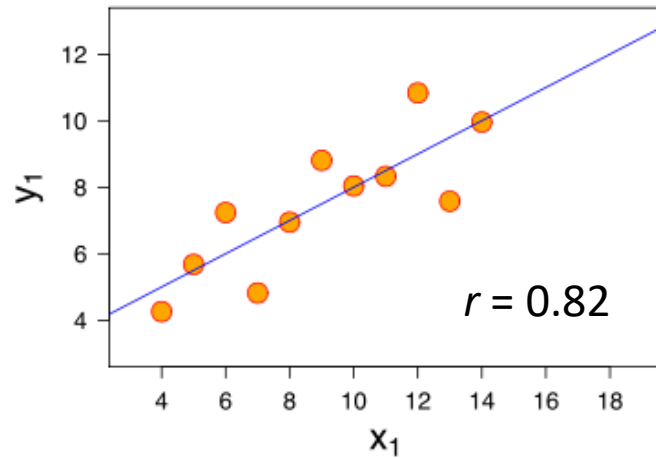
-0.005



-.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

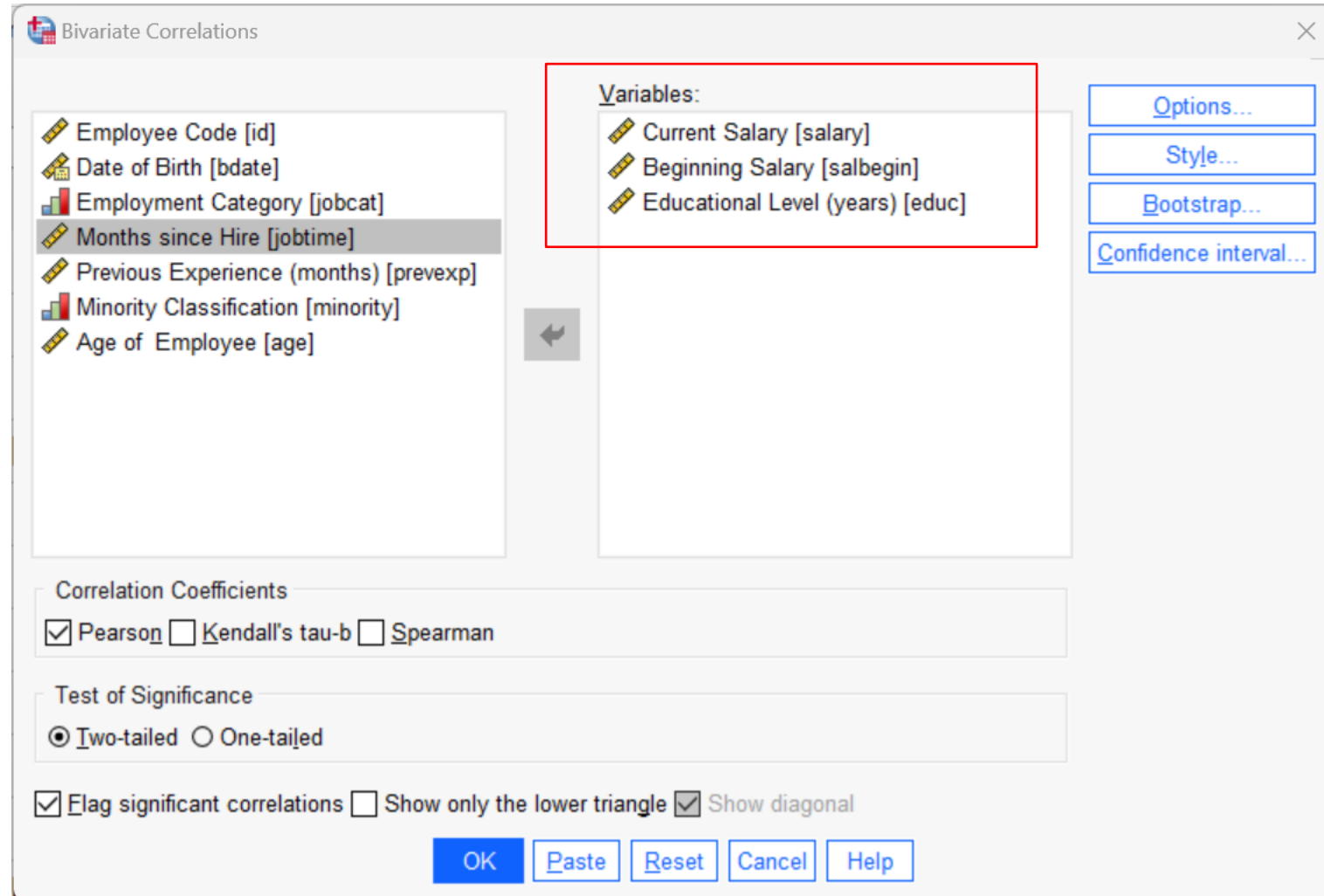
Power Analysis  
Meta Analysis  
Reports  
Descriptive Statistics  
Bayesian Statistics  
Tables  
Compare Means and Proportions  
General Linear Model  
Generalized Linear Models  
Mixed Models  
**Correlate**  
Regression  
Loglinear  
Neural Networks  
Classify  
Dimension Reduction  
Scale  
Nonparametric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

Search application

	id	age	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				
19	36	Female				
20	37	Male				
21	38	Male				
22	39	Male				
23	40	Male				
24	41	Male				
25	42	Male				
26	43	Male				
27	44	Male				
28	45	Male				
29	46	Male				
30	47	Male				
31	48	Male				
32	49	Male				
33	50	Male				
34	51	Male				
35	52	Male				
36	53	Male				
37	54	Male				
38	55	Male				
39	56	Male				
40	57	Male				
41	58	Male				
42	59	Male				
43	60	Male				
44	61	Male				
45	62	Male				
46	63	Male				
47	64	Male				
48	65	Male				
49	66	Male				
50	67	Male				
51	68	Male				
52	69	Male				
53	70	Male				
54	71	Male				
55	72	Male				
56	73	Male				
57	74	Male				
58	75	Male				
59	76	Male				
60	77	Male				
61	78	Male				
62	79	Male				
63	80	Male				
64	81	Male				
65	82	Male				
66	83	Male				
67	84	Male				
68	85	Male				
69	86	Male				
70	87	Male				
71	88	Male				
72	89	Male				
73	90	Male				
74	91	Male				
75	92	Male				
76	93	Male				
77	94	Male				
78	95	Male				
79	96	Male				
80	97	Male				
81	98	Male				
82	99	Male				
83	100	Male				
84	101	Male				
85	102	Male				
86	103	Male				
87	104	Male				
88	105	Male				
89	106	Male				
90	107	Male				
91	108	Male				
92	109	Male				
93	110	Male				
94	111	Male				
95	112	Male				
96	113	Male				
97	114	Male				
98	115	Male				
99	116	Male				
100	117	Male				
101	118	Male				
102	119	Male				
103	120	Male				
104	121	Male				
105	122	Male				
106	123	Male				
107	124	Male				
108	125	Male				
109	126	Male				
110	127	Male				
111	128	Male				
112	129	Male				
113	130	Male				
114	131	Male				
115	132	Male				
116	133	Male				
117	134	Male				
118	135	Male				
119	136	Male				
120	137	Male				
121	138	Male				
122	139	Male				
123	140	Male				
124	141	Male				
125	142	Male				
126	143	Male				
127	144	Male				
128	145	Male				
129	146	Male				
130	147	Male				
131	148	Male				
132	149	Male				
133	150	Male				
134	151	Male				
135	152	Male				
136	153	Male				
137	154	Male				
138	155	Male				
139	156	Male				
140	157	Male				
141	158	Male				
142	159	Male				
143	160	Male				
144	161	Male				
145	162	Male				
146	163	Male				
147	164	Male				
148	165	Male				
149	166	Male				
150	167	Male				
151	168	Male				
152	169	Male				
153	170	Male				
154	171	Male				
155	172	Male				
156	173	Male				
157	174	Male				
158	175	Male				
159	176	Male				
160	177	Male				
161	178	Male				
162	179	Male				
163	180	Male				
164	181	Male				
165	182	Male				
166	183	Male				
167	184	Male				
168	185	Male				
169	186	Male				
170	187	Male				
171	188	Male				
172	189	Male				
173	190	Male				
174	191	Male				
175	192	Male				
176	193	Male				
177	194	Male				
178	195	Male				
179	196	Male				
180	197	Male				
181	198	Male				
182	199	Male				
183	200	Male				
184	201	Male				
185	202	Male				
186	203	Male				
187	204	Male				
188	205	Male				
189	206	Male				
190	207	Male				
191	208	Male				
192	209	Male				
193	210	Male				
194	211	Male				
195	212	Male				
196	213	Male				
197	214	Male				
198	215	Male				
199	216	Male				
200	217	Male				
201	218	Male				
202	219	Male				
203	220	Male				
204	221	Male				
205	222	Male				
206	223	Male				
207	224	Male				
208	225	Male				
209	226	Male				
210	227	Male				
211	228	Male				
212	229	Male				
213	230	Male				
214	231	Male				
215	232	Male				
216	233	Male				
217	234	Male				
218	235	Male				
219	236	Male				
220	237	Male				
221	238	Male				
222	239	Male				
223	240	Male				
224	241	Male				
225	242	Male				
226	243	Male				
227	244	Male				
228	245	Male				
229	246	Male				
230	247	Male				
231	248	Male				
232	249	Male				
233	250	Male				
234	251	Male				
235	252	Male				
236	253	Male				
237	254	Male				
238	255	Male				
239	256	Male				
240	257	Male				
241	258	Male				
242	259	Male				
243	260	Male				
244	261	Male				
245	262	Male				
246	263	Male				
247	264	Male				
248	265	Male				
249	266	Male				
250	267	Male				
251	268	Male				
252	269	Male				
253	270	Male				
254	271	Male				
255	272	Male				
256	273	Male				
257	274	Male				
258	275	Male				
259	276	Male				
260	277	Male				
261	278	Male				
262	279	Male				
263	280	Male				
264	281	Male				
265	282	Male				
266	283	Male				
267	284	Male				
268	285	Male				
269	286	Male				
270	287	Male				
271	288	Male				
272	289	Male				
273	290	Male				
274	291	Male				
275	292	Male				
276	293	Male				
277	294	Male				
278	295	Male				
279	296	Male				
280	297	Male				
281	298	Male				
282	299	Male				
283	300	Male				
284	3					

# 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

<sup>\*\*</sup>. 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

<sup>\*\*</sup>. 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

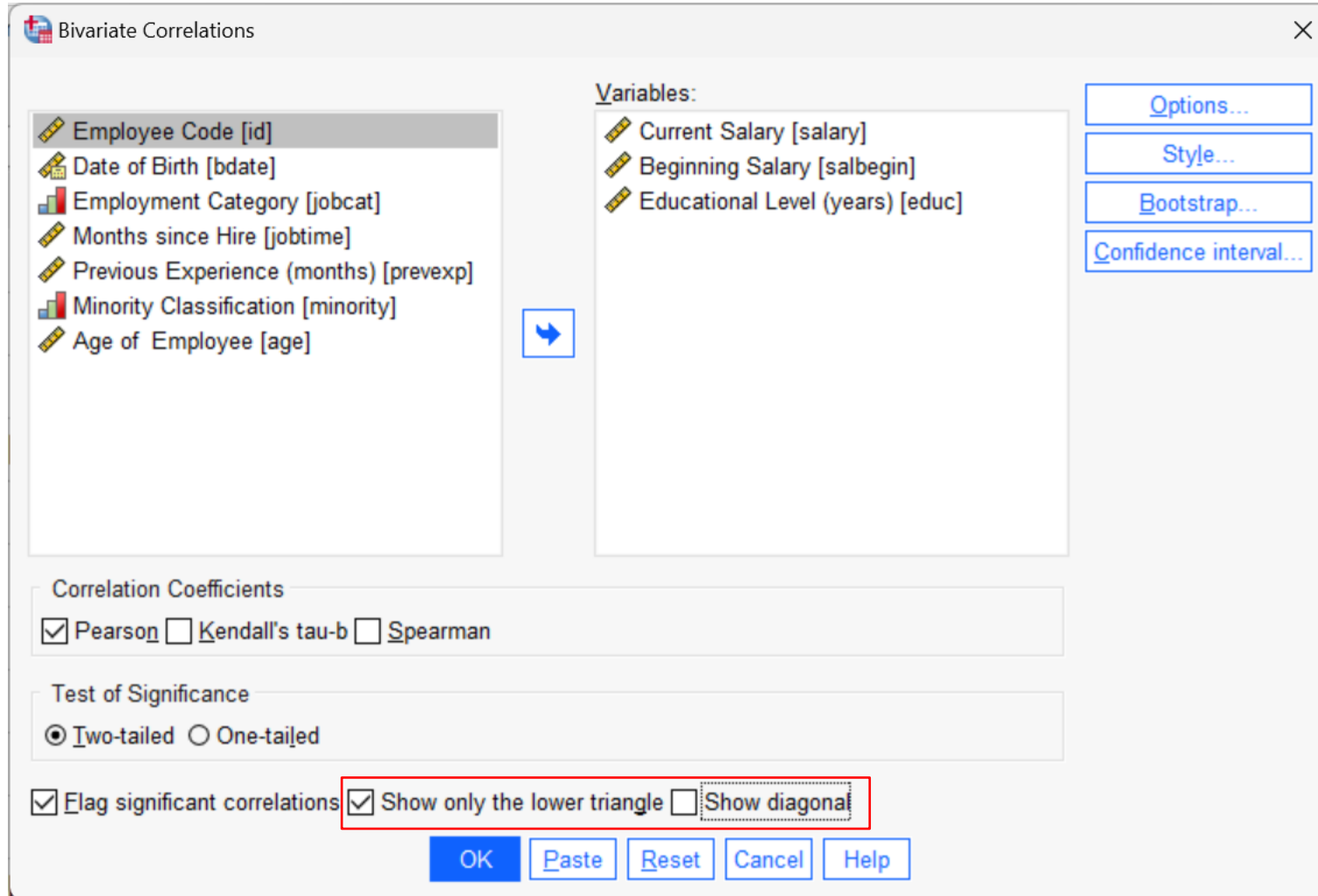
**Correlations**

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880 <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

- 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, with 'Kendall's tau-b' and 'Spearman' unchecked. The 'Test of Significance' section has 'Two-tailed' selected, with 'One-tailed' unselected. At the bottom, 'Flag significant correlations' is checked, and 'Show only the lower triangle' is also checked (highlighted with a red box), while 'Show diagonal' is unchecked. On the far right, there are buttons for 'Options...', 'Style...', 'Bootstrap...', and 'Confidence interval...'. At the bottom center are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'.

# 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

<sup>\*\*</sup>. 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

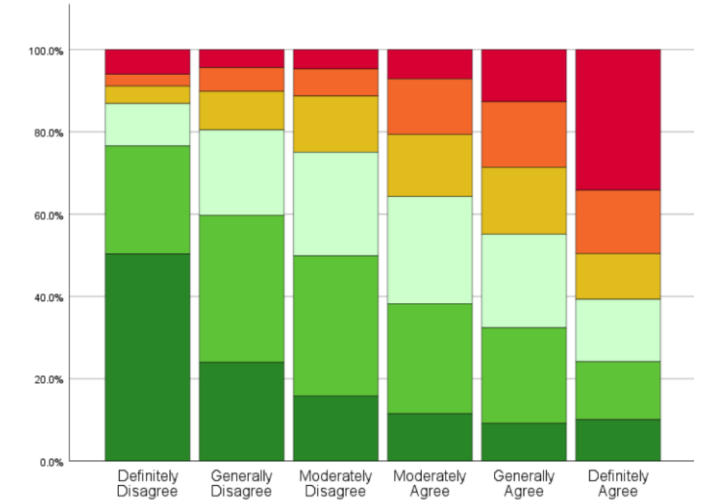
$x_i$  = values of the x-variable in a sample

$\bar{x}$  = 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

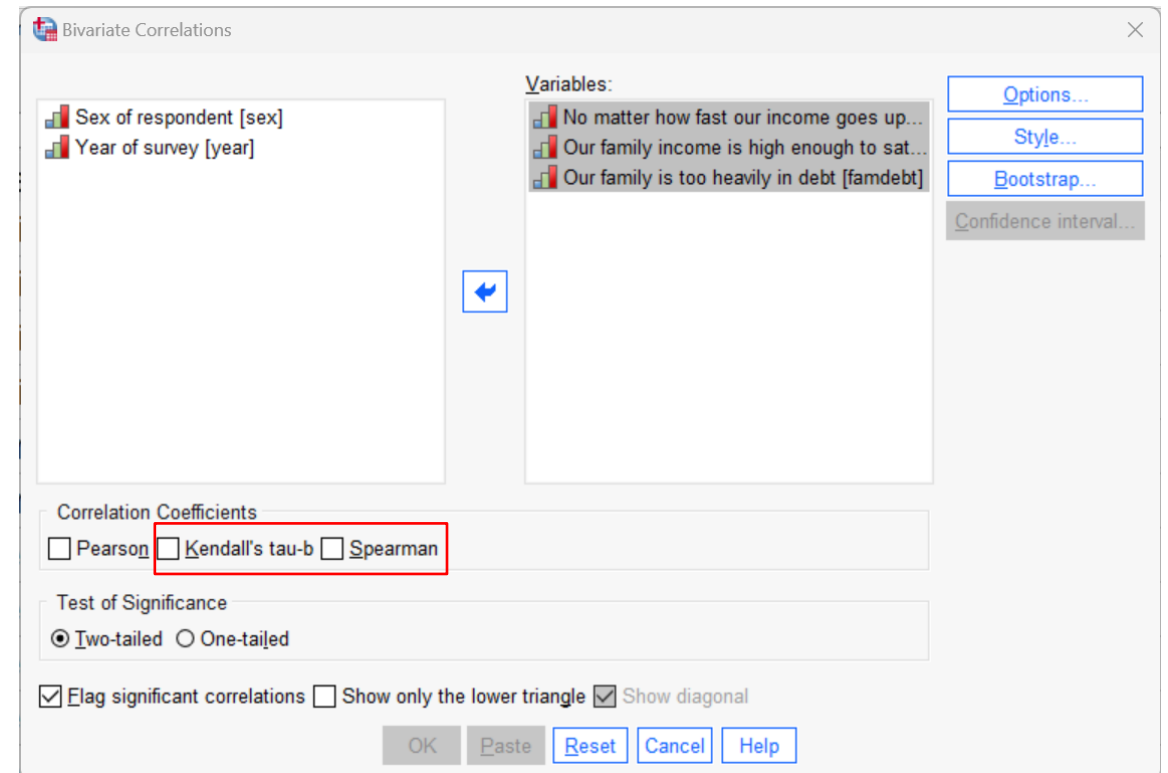
**Let's see an example of calculating  
Pearson's R**



# 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

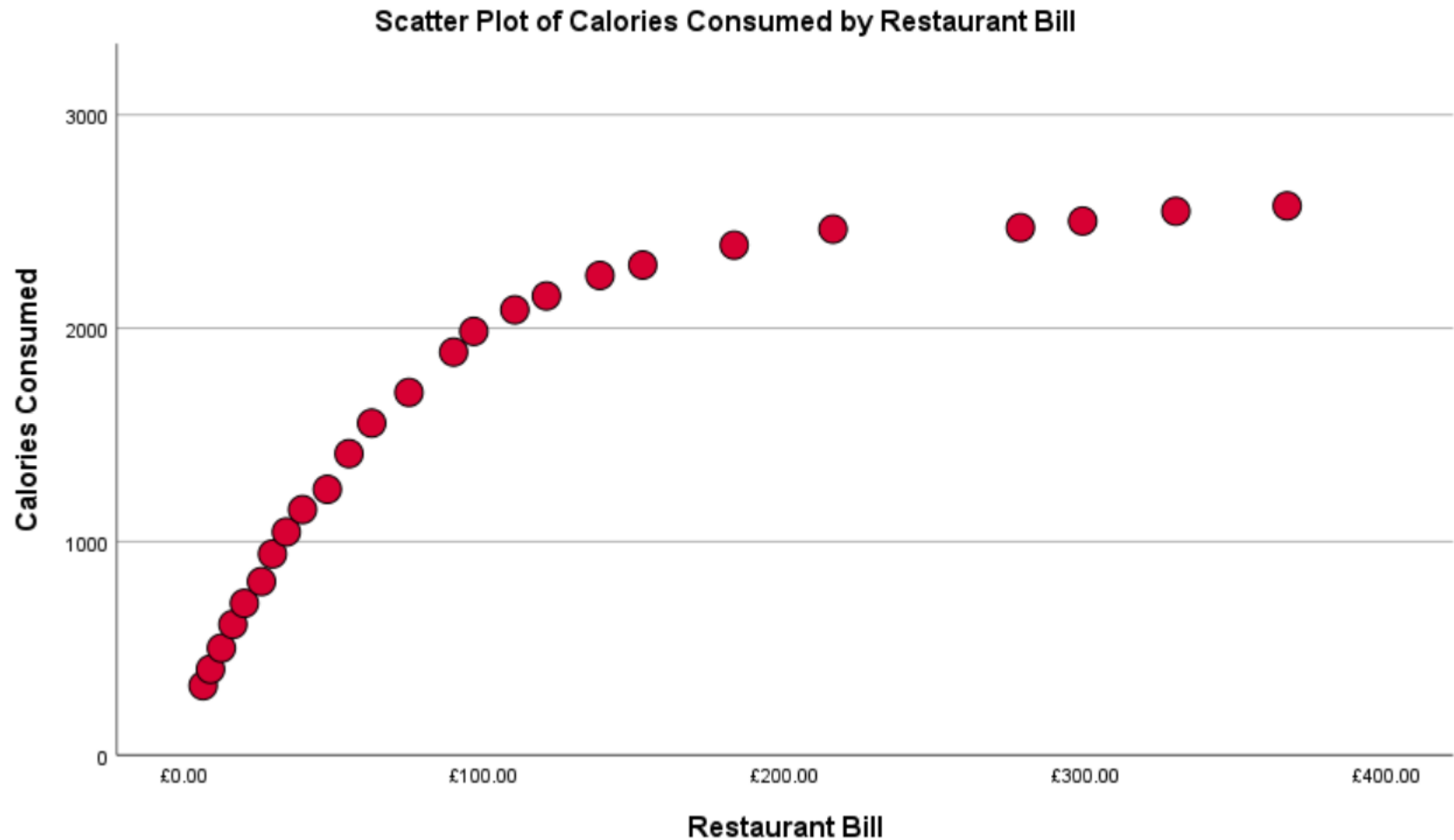


# Spearman's Rho

- **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
- 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....



**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

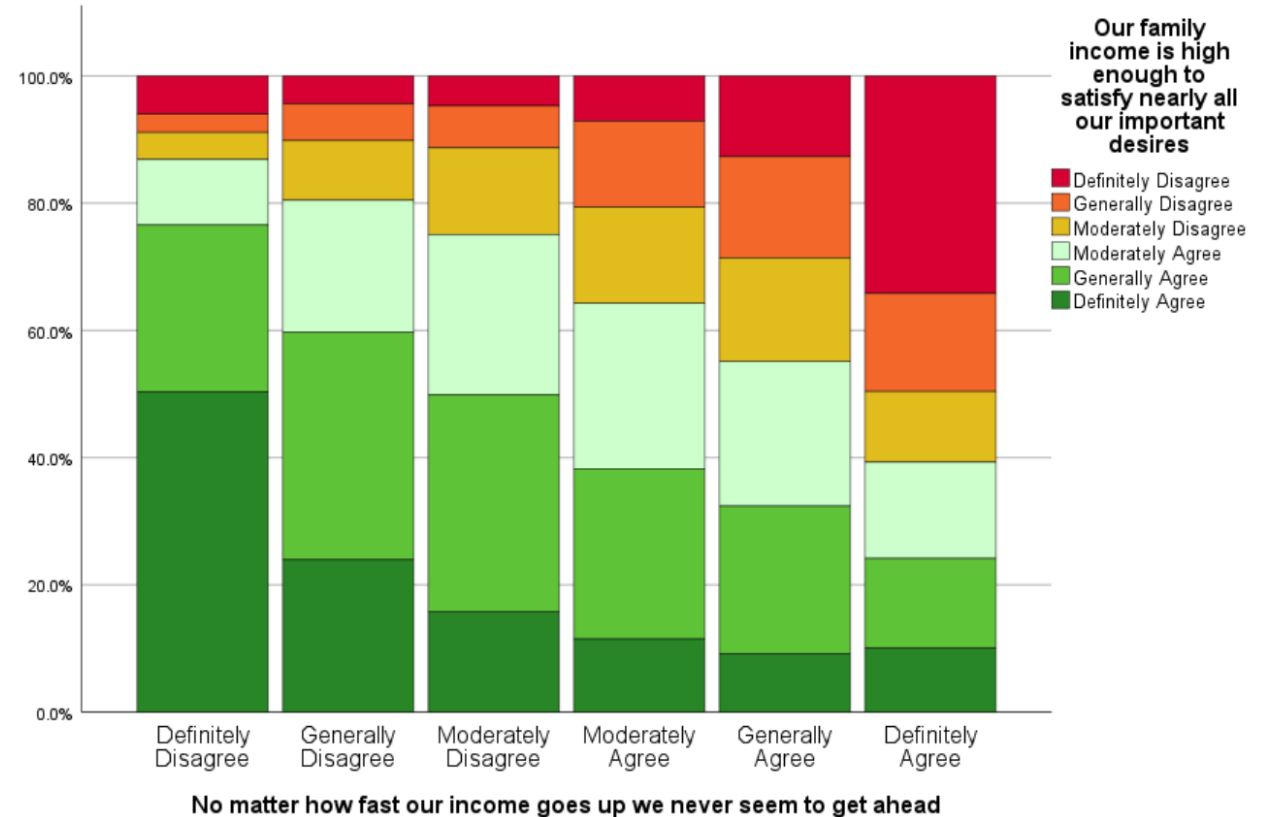
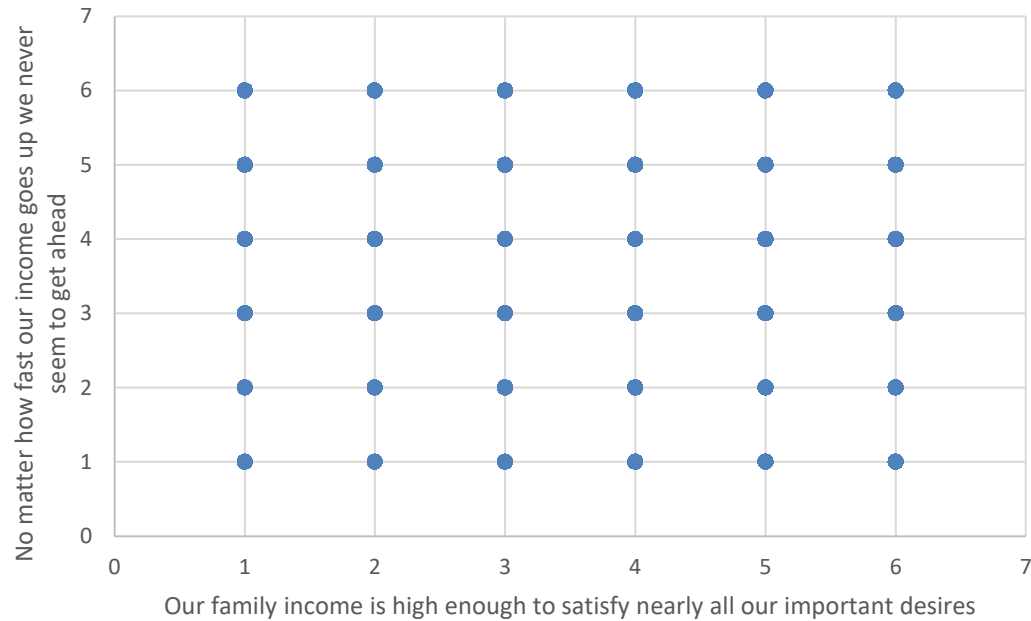
# Kendall's Tau

- 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

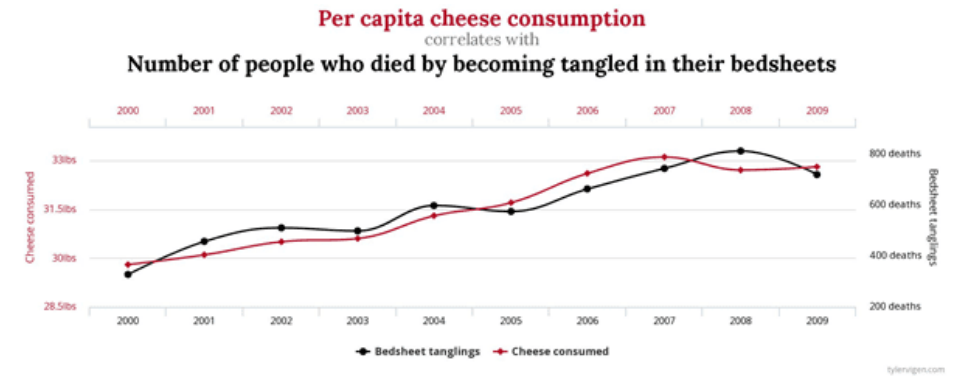
	🎬 Movie_Title	📊 IMDb	📊 Rotten_Tomatoes
1	Uncorked	1	6
2	Spenser Confidential	2	1
3	The Willoughbys	3	5
4	Tigertail	4	3
5	Extraction	5	2
6	The Half of It	6	8
7	To All the Boys I've Loved Before	7	9
8	LA Originals	8	4
9	Miss Americana	9	7
10	Crip Camp: A Disability Revolution	10	10

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

# Visualising Correlations for Ordinal Variables

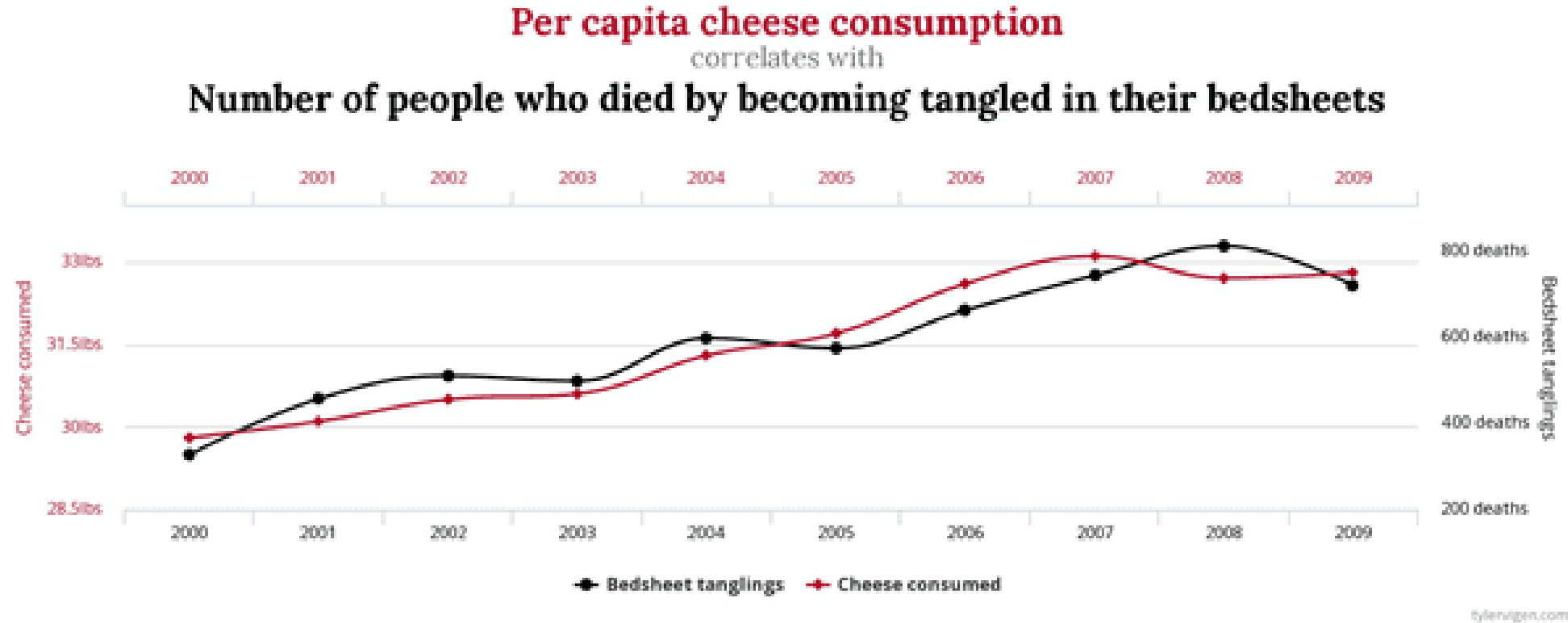


May require a different approach than scatterplots



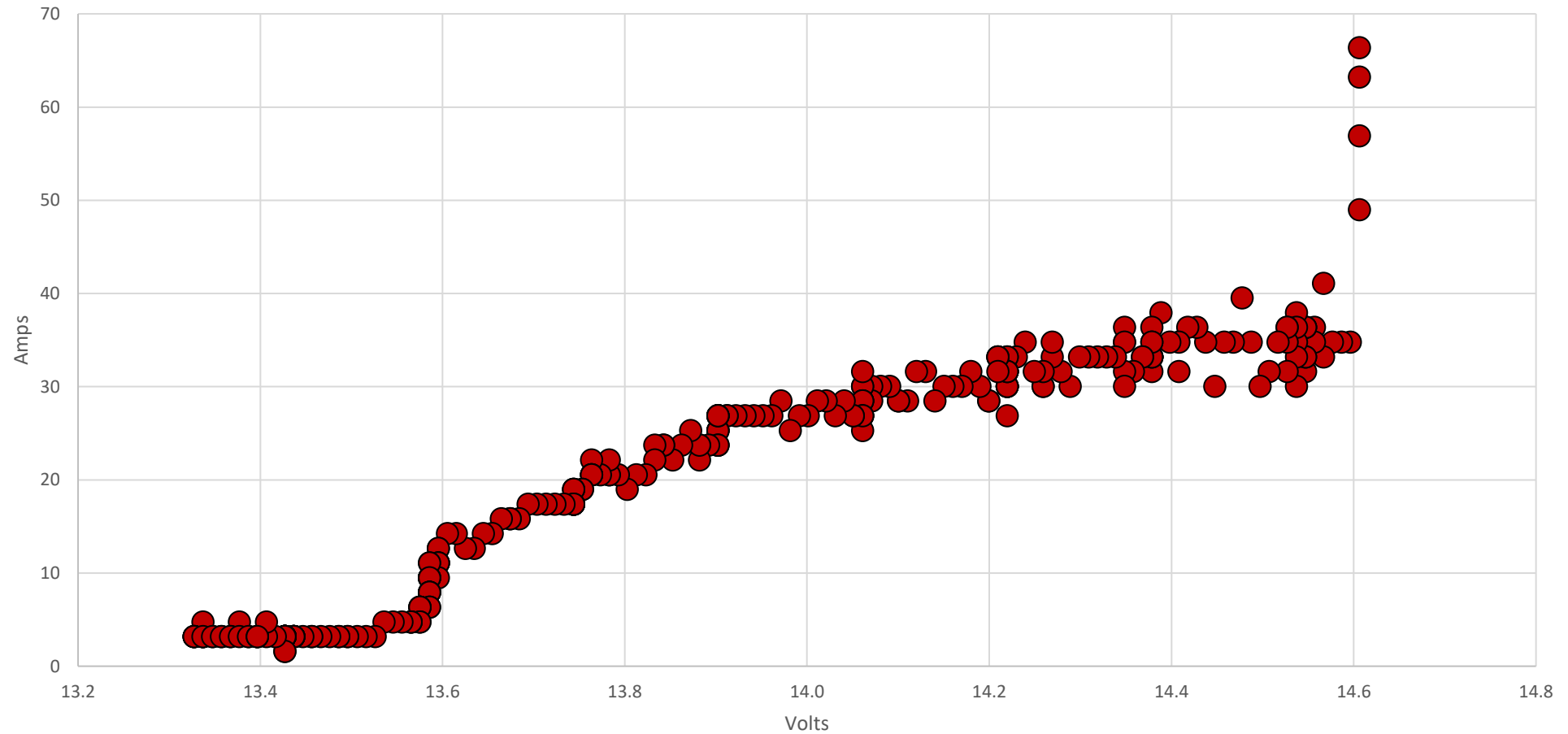
# The Limitations of Correlations

# Correlation does not indicate causation



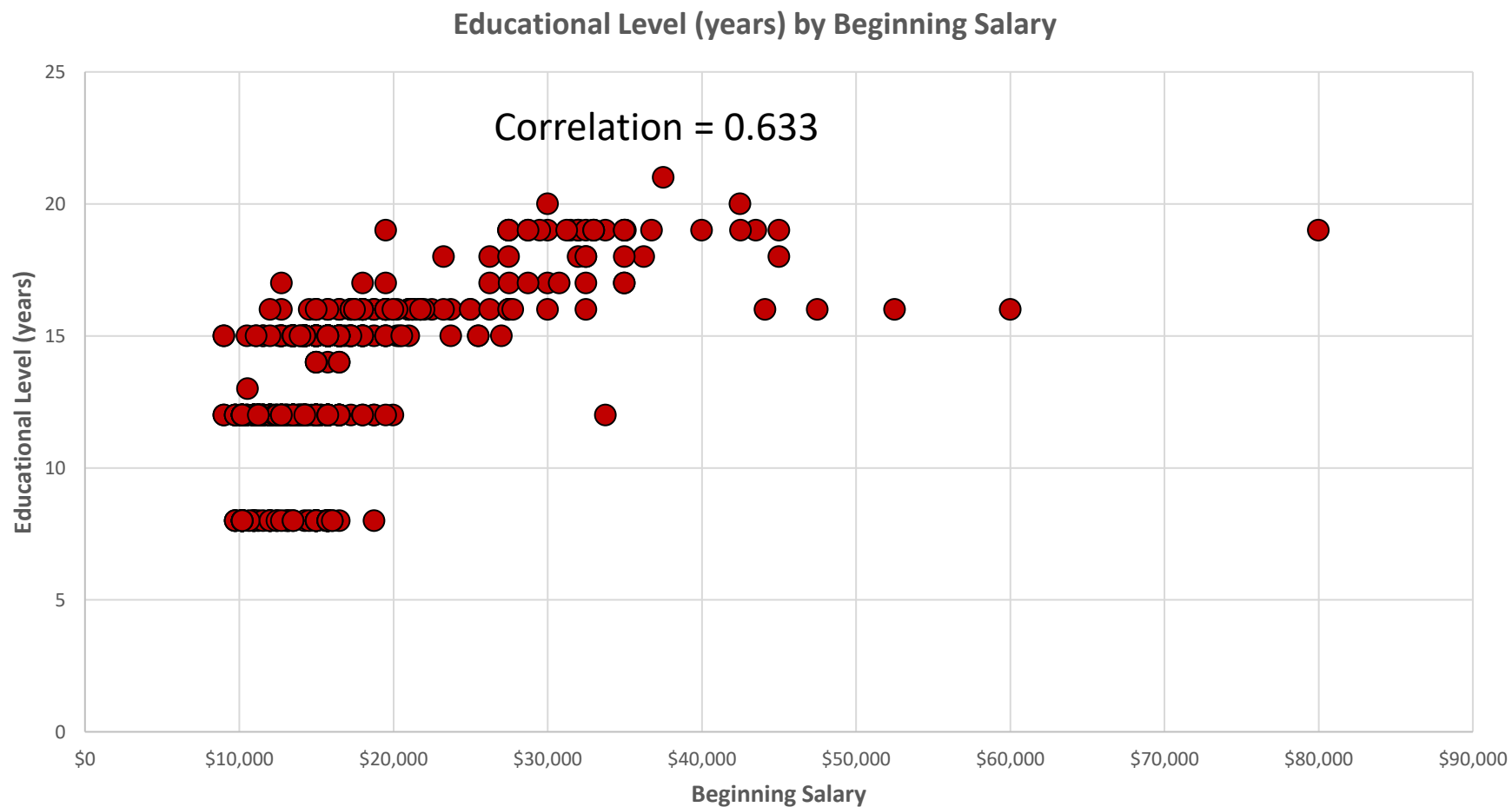
<https://www.productleadership.com/does-causation-imply-correlation/>

# Can't accurately measure curvilinear relationships

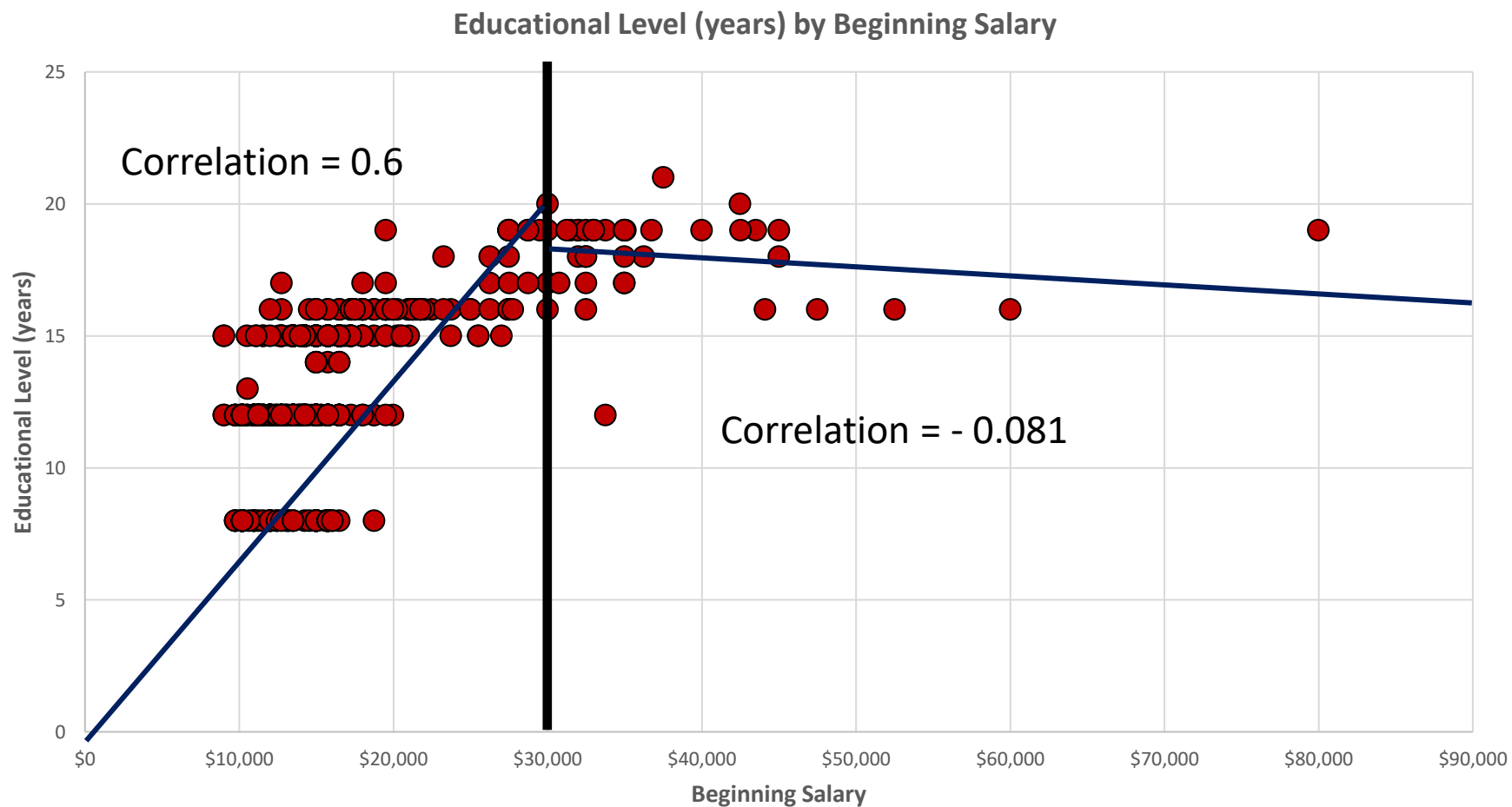




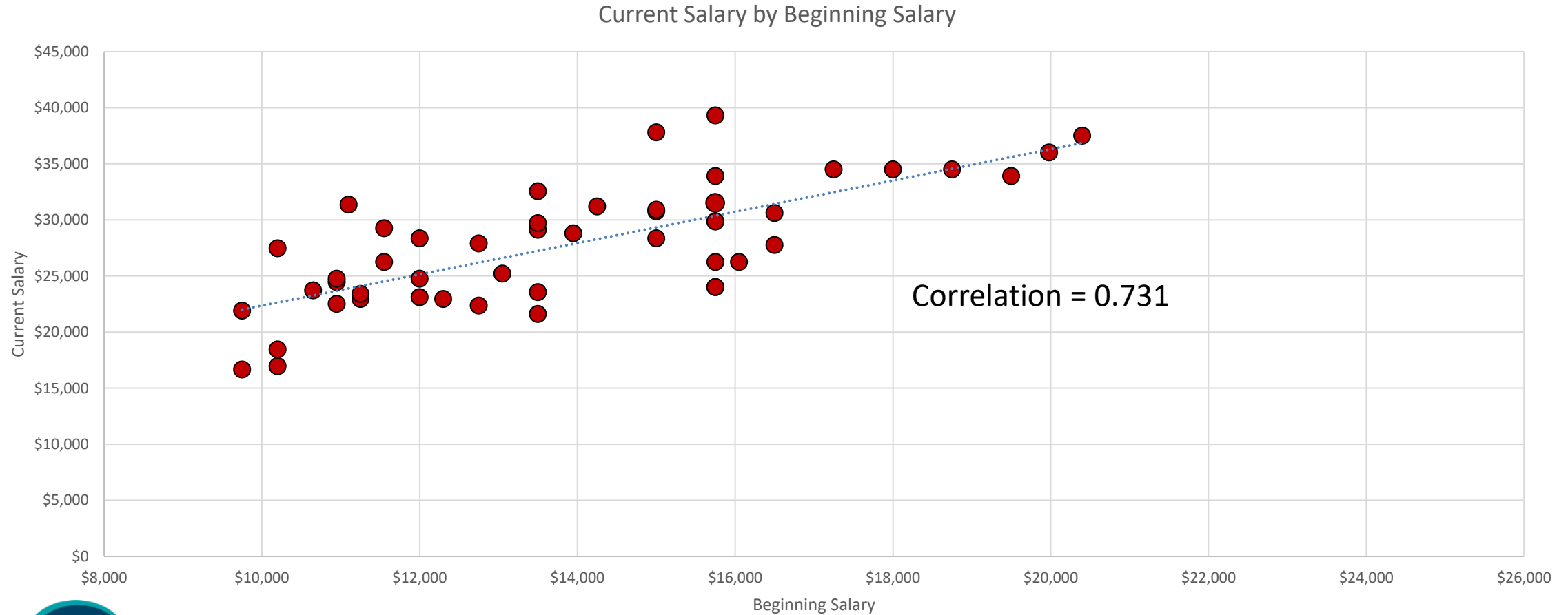
# Are influenced by the range of values in the sample



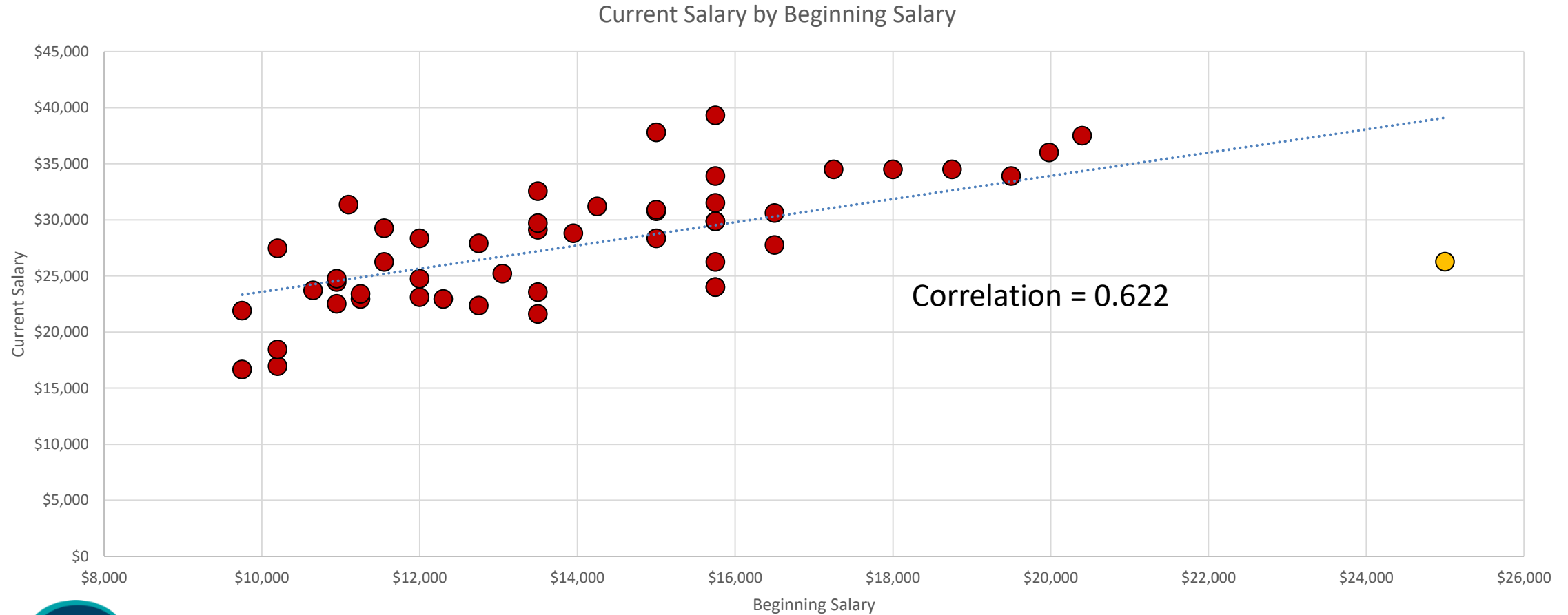
# Are influenced by the range of values in the sample



# Can be unduly affected by extreme/outlier values



# Can be unduly affected by extreme/outlier values



# 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