Correlation Co-efficient

r

At a junior tournament, a group of young athletes throw a discus. The *age* and *distance thrown* are recorded for each athlete.

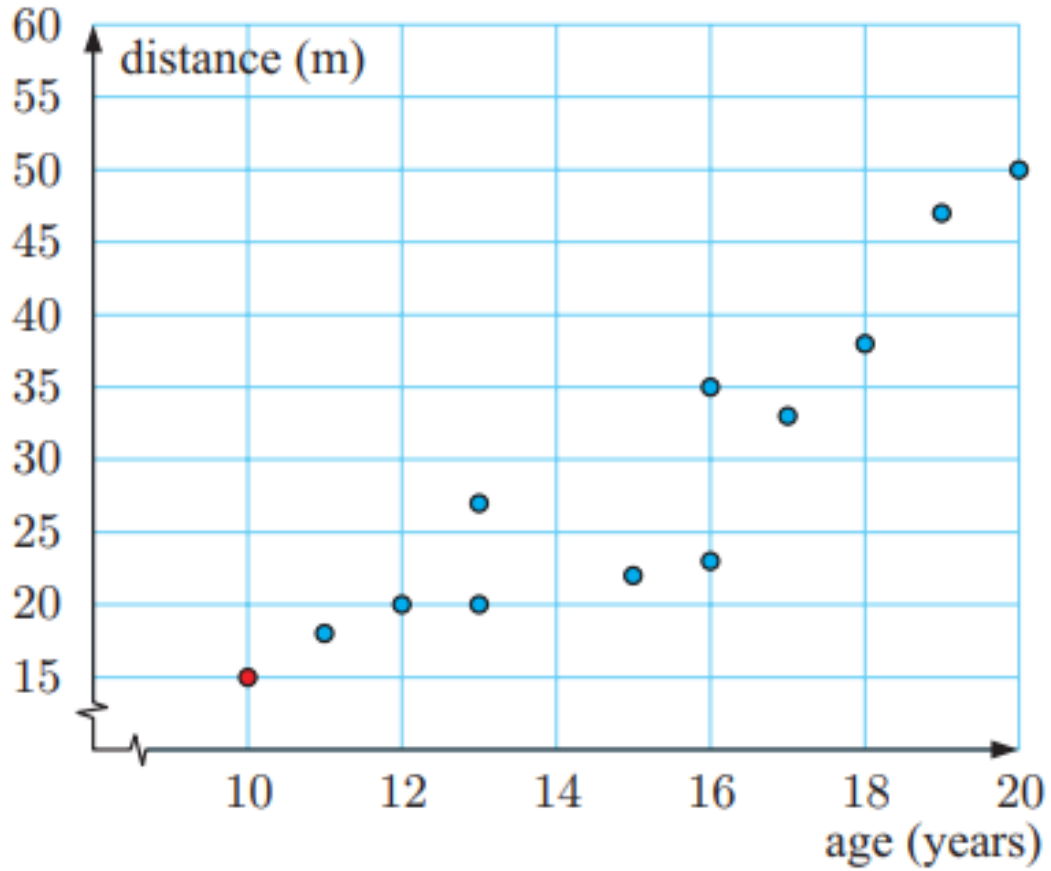
Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Age (years)	12	16	16	18	13	19	11	10	20	17	15	13
Distance thrown (m)	20	35	23	38	27	47	18	15	50	33	22	20

We can observe the relationship between the variables by plotting the data on a **scatter diagram**.

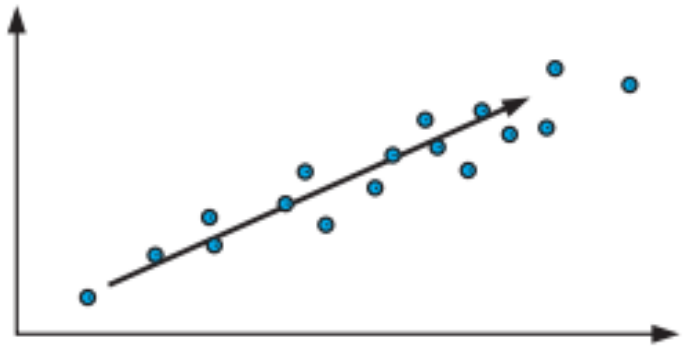
We place the **independent variable** *age* on the horizontal axis, and the **dependent variable** *distance* on the vertical axis.

We then plot each data value as a point on the scatter diagram. For example, the red point represents athlete H, who is 10 years old and threw the discus 15 metres.

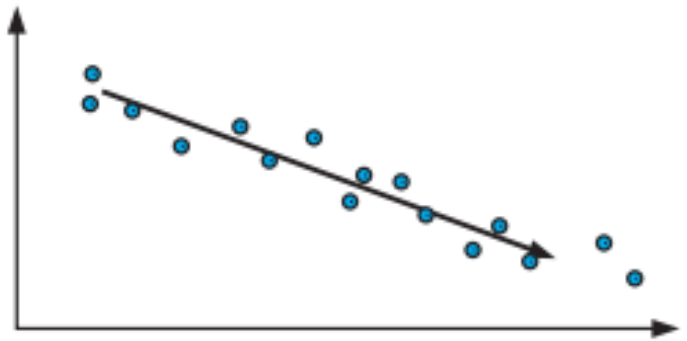
From the general shape formed by the dots, we can see that as the *age* increases, so does the *distance thrown*.



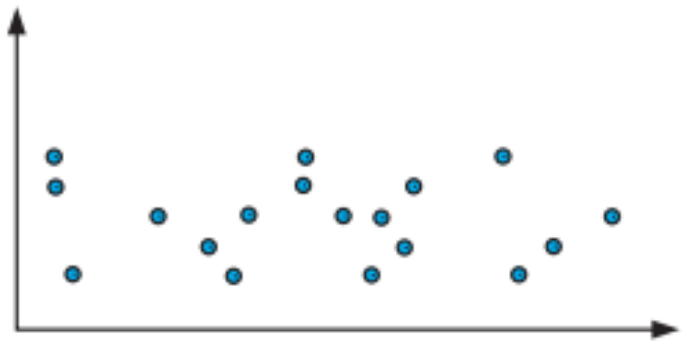
DIRECTION



For a generally *upward* trend, we say that the correlation is **positive**. An increase in the independent variable means that the dependent variable generally increases.



For a generally *downward* trend, we say that the correlation is **negative**. An increase in the independent variable means that the dependent variable generally decreases.

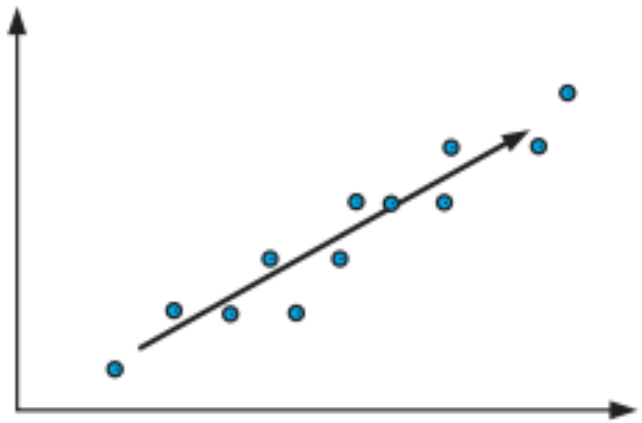


For *randomly scattered* points, with no upward or downward trend, we say there is **no correlation**.

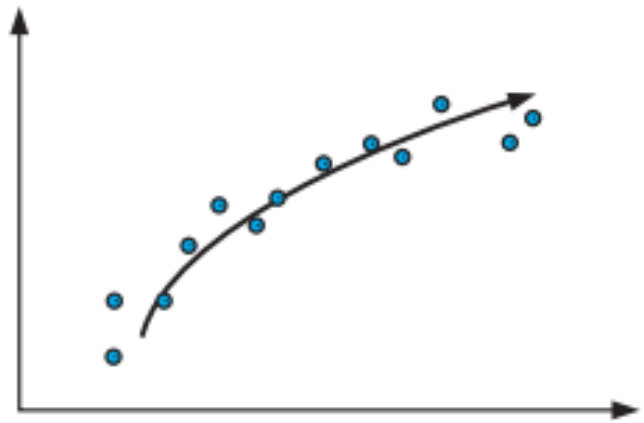
LINEARITY

We determine whether the points follow a **linear** trend, or in other words approximately form a straight line.

These points are roughly linear.

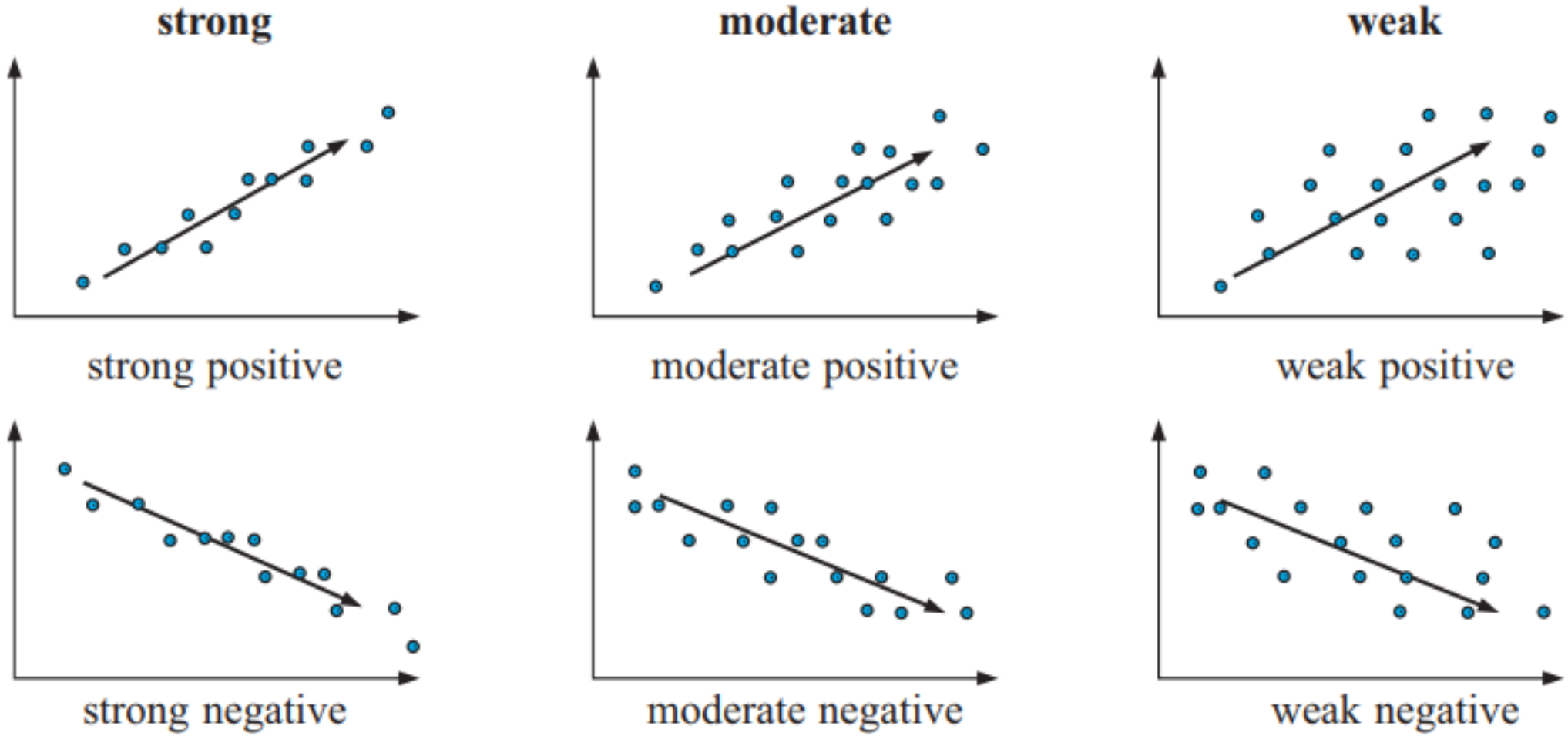


These points do not follow a linear trend.



STRENGTH

We want to know how closely the data follows a pattern or trend. The strength of correlation is usually described as either strong, moderate, or weak.

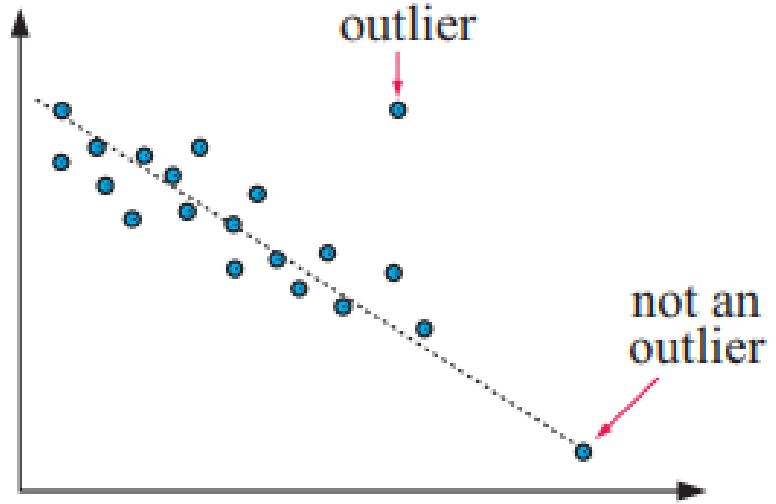


OUTLIERS

We observe and investigate any **outliers**, or isolated points which do not follow the trend formed by the main body of data.

If an outlier is the result of a recording or graphing error, it should be discarded. However, if the outlier proves to be a genuine piece of data, it should be kept.

For the scatter diagram for the data in the **Opening Problem**, we can say that there is a strong positive correlation between *age* and *distance thrown*. The relationship appears to be linear, with no outliers.



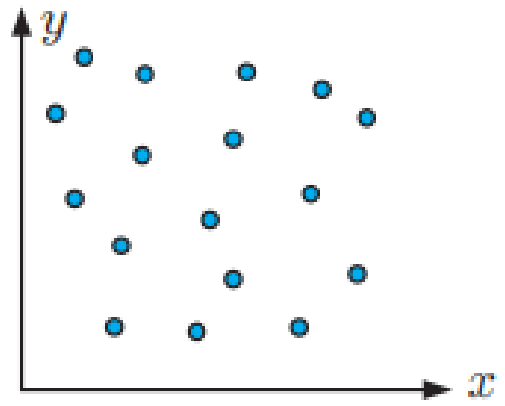
2 For the following scatter diagrams, comment on:

i the existence of any *pattern* (positive, negative or no correlation)

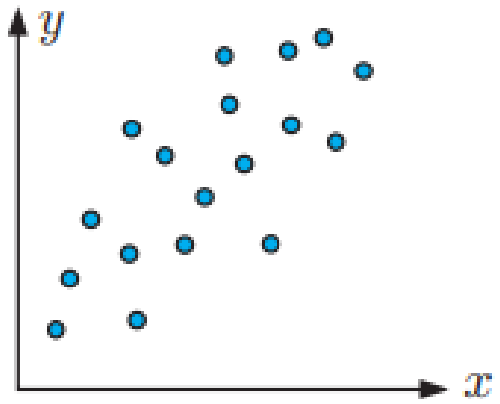
ii the relationship *strength* (zero, weak, moderate or strong)

iii whether the relationship is linear **iv** whether there are any outliers.

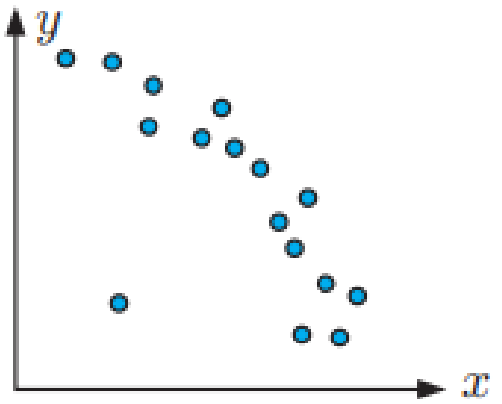
a



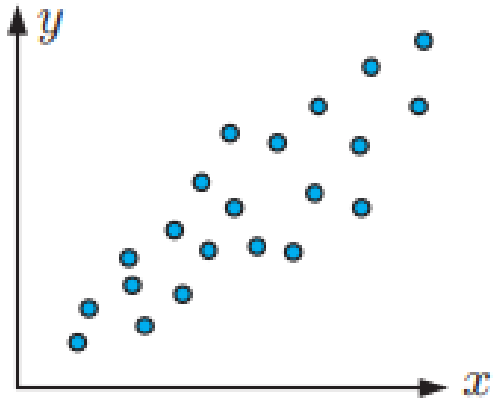
b



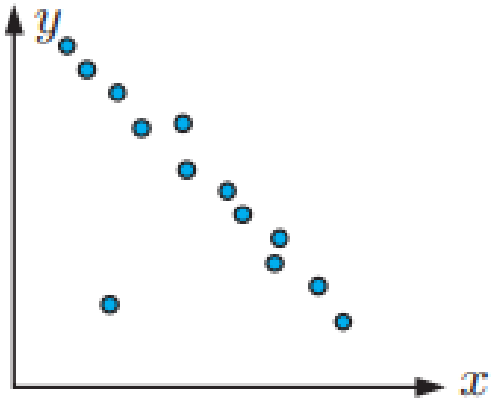
c



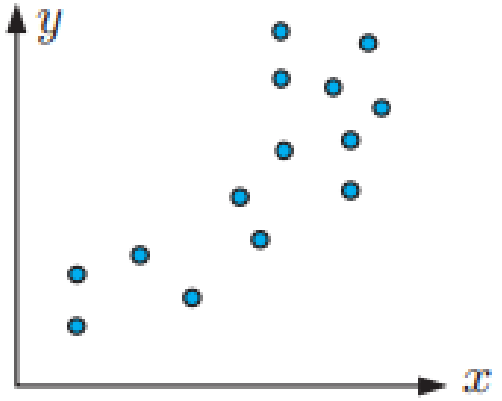
d



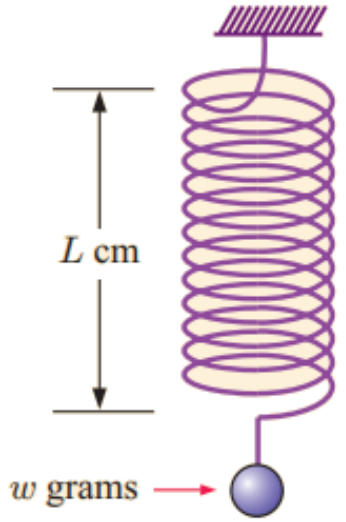
e



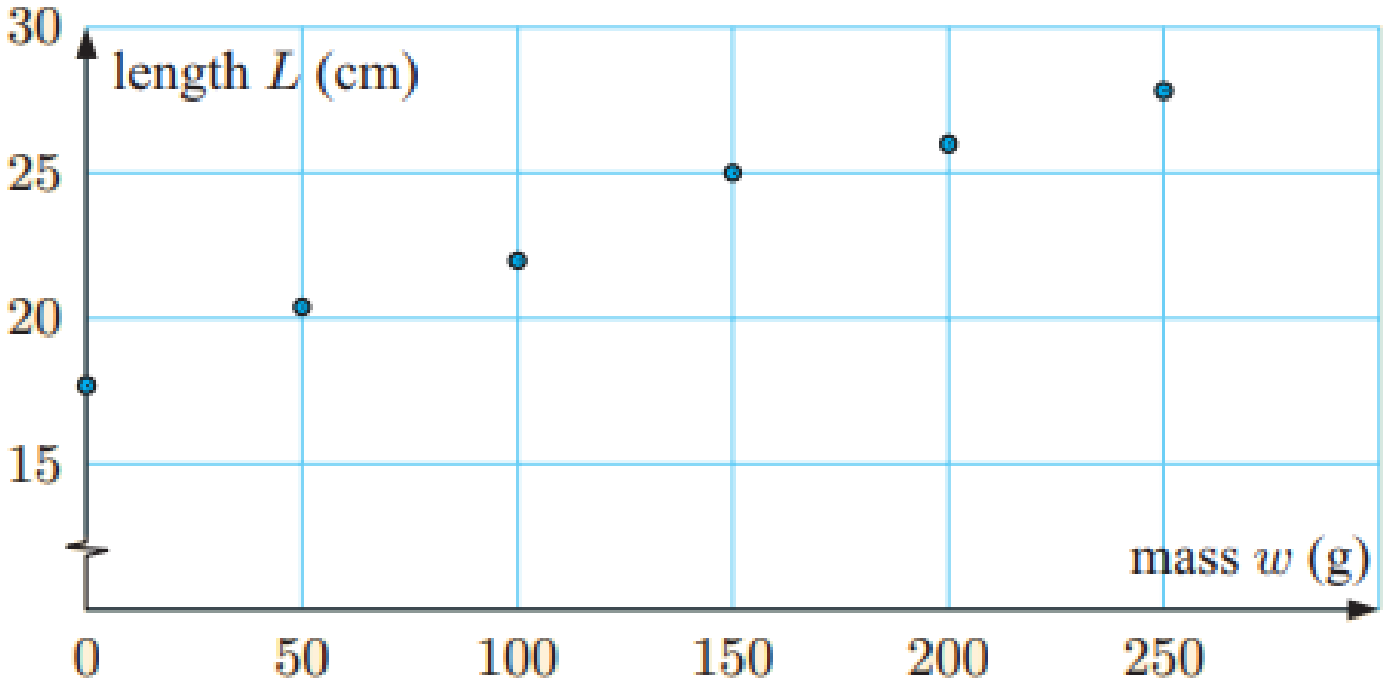
f



Suppose we wish to examine the relationship between the *length* of a helical spring and the *mass* that is hung from the spring.

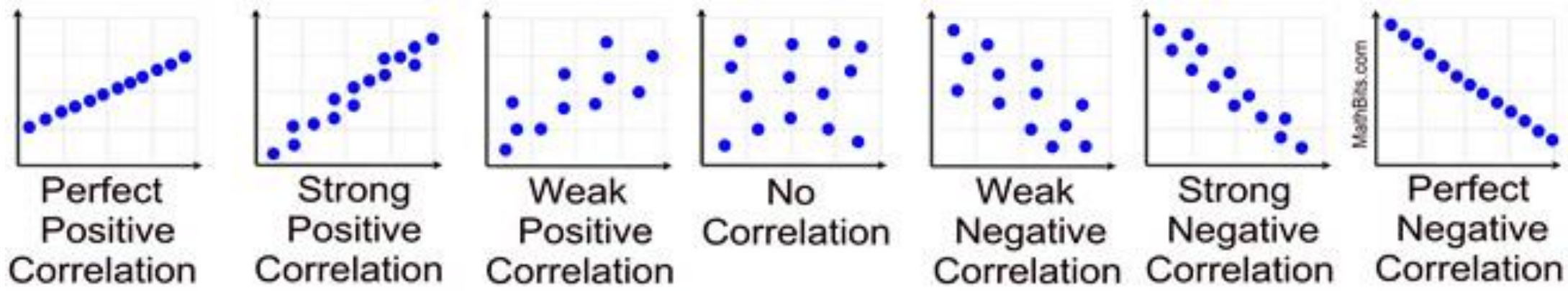


Mass w (grams)	0	50	100	150	200	250
Length L (cm)	17.7	20.4	22.0	25.0	26.0	27.8



Relationship Strength

How closely the points in the scatterplot fit in the straight line.



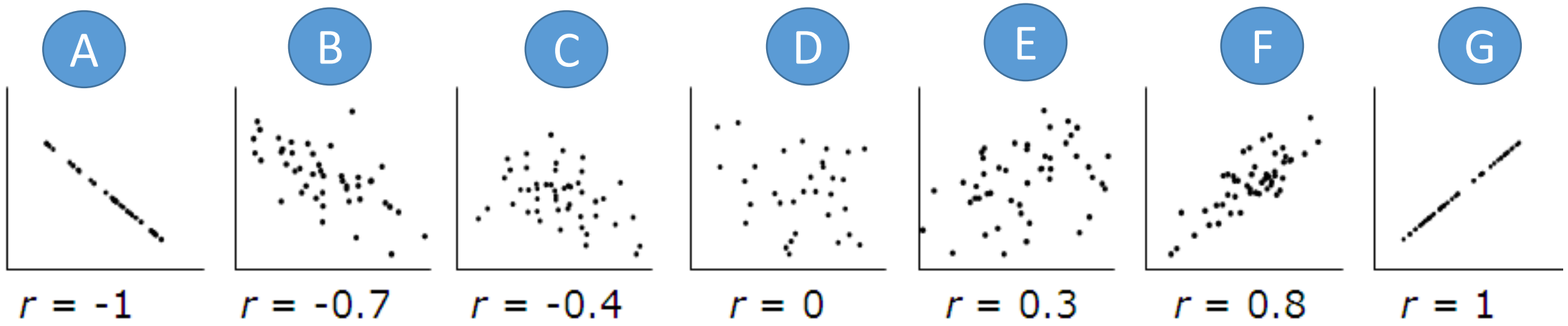
Pearson Correlation Co-efficient

r

A mathematical measure of the strength of a linear relationship.

If r is positive, the relationship is positive.
If r is negative, the relationship is negative.

If |r| is >= 0.75, the relationship is strong.
Else If |r| is < 0.75 and >= 0.5, the relationship is moderate.
Else If |r| is < 0.5 and >= 0.25, the relationship is weak.
Otherwise, there is no correlation.



If r is positive, the relationship is positive.
If r is negative, the relationship is negative.

If |r| is >= 0.75, the relationship is strong.
Else If |r| is < 0.75 and >= 0.5, the relationship is moderate.
Else If |r| is < 0.5 and >= 0.25, the relationship is weak.
Otherwise, there is no correlation.

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Yuck.

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Correlation Coefficient Formula

Yuck.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Both x and y need to be numbers

Only for linear relationships

Correlation Coefficient Formula

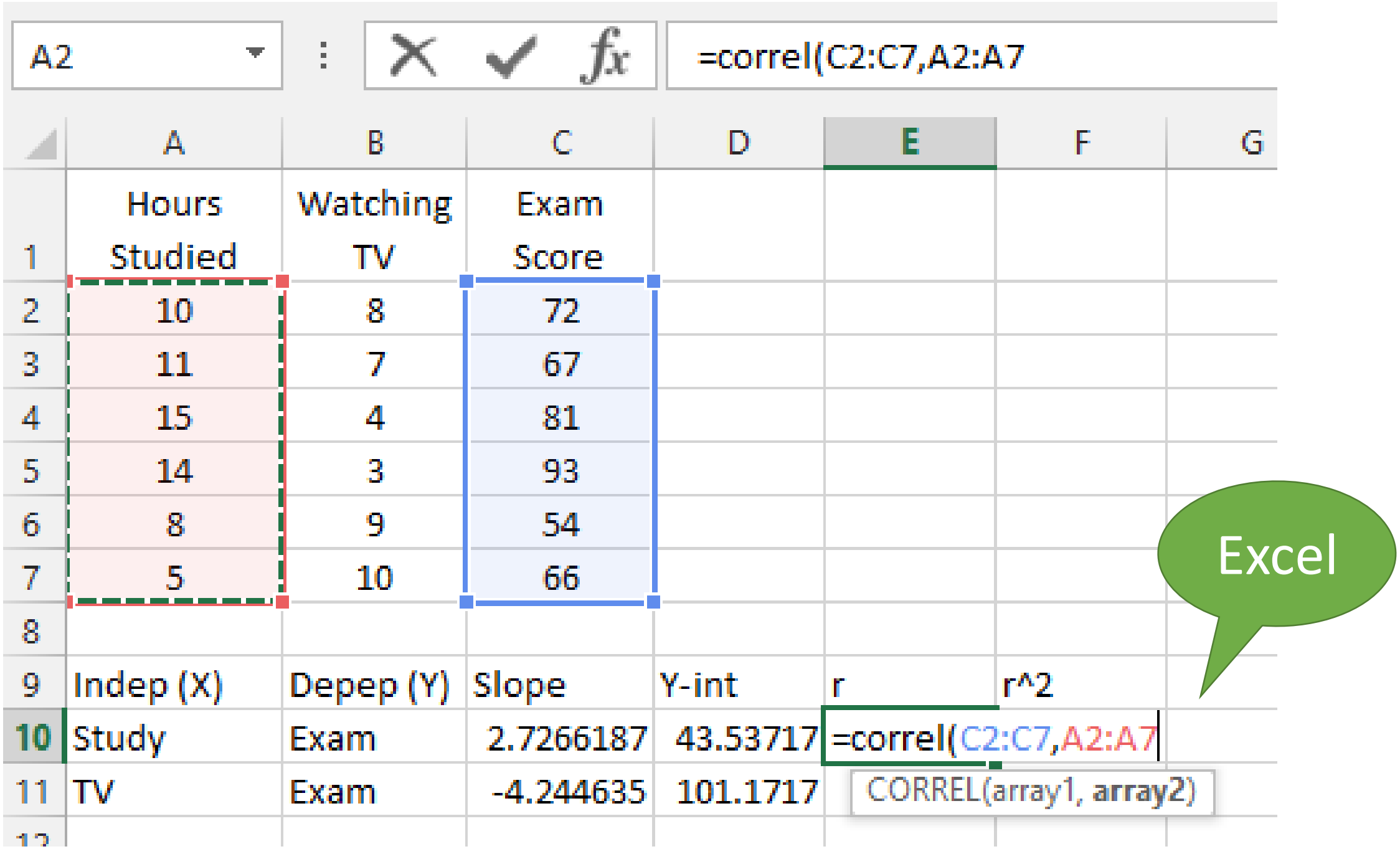
Yuck.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Both x and y need to be numbers

Only for linear relationships

Excel: =correl(y_values, x_values)


A2
=correl(C2:C7,A2:A7
Hours Studied
Watching TV
Exam Score
10
8
72
11
7
67
15
4
81
14
3
93
8
9
54
5
10
66
Indep (X)
Depep (Y)
Slope
Y-int
r
r^2
Study
Exam
2.7266187
43.53717
=correl(C2:C7,A2:A7
TV
Exam
-4.244635
101.1717
CORREL(array1, array2)
Excel

Pearson's Correlation Co-efficient (r)

$r = 1$	perfect positive correlation	$r = -1$	perfect negative correlation
$0.95 \leqslant r < 1$	very strong positive correlation	$-1 < r \leqslant -0.95$	very strong negative correlation
$0.87 \leqslant r < 0.95$	strong positive correlation	$-0.95 < r \leqslant -0.87$	strong negative correlation
$0.5 \leqslant r < 0.87$	moderate positive correlation	$-0.87 < r \leqslant -0.5$	moderate negative correlation
$0.1 \leqslant r < 0.5$	weak positive correlation	$-0.5 < r \leqslant -0.1$	weak negative correlation
$0 \leqslant r < 0.1$	no correlation	$-0.1 < r \leqslant 0$	no correlation

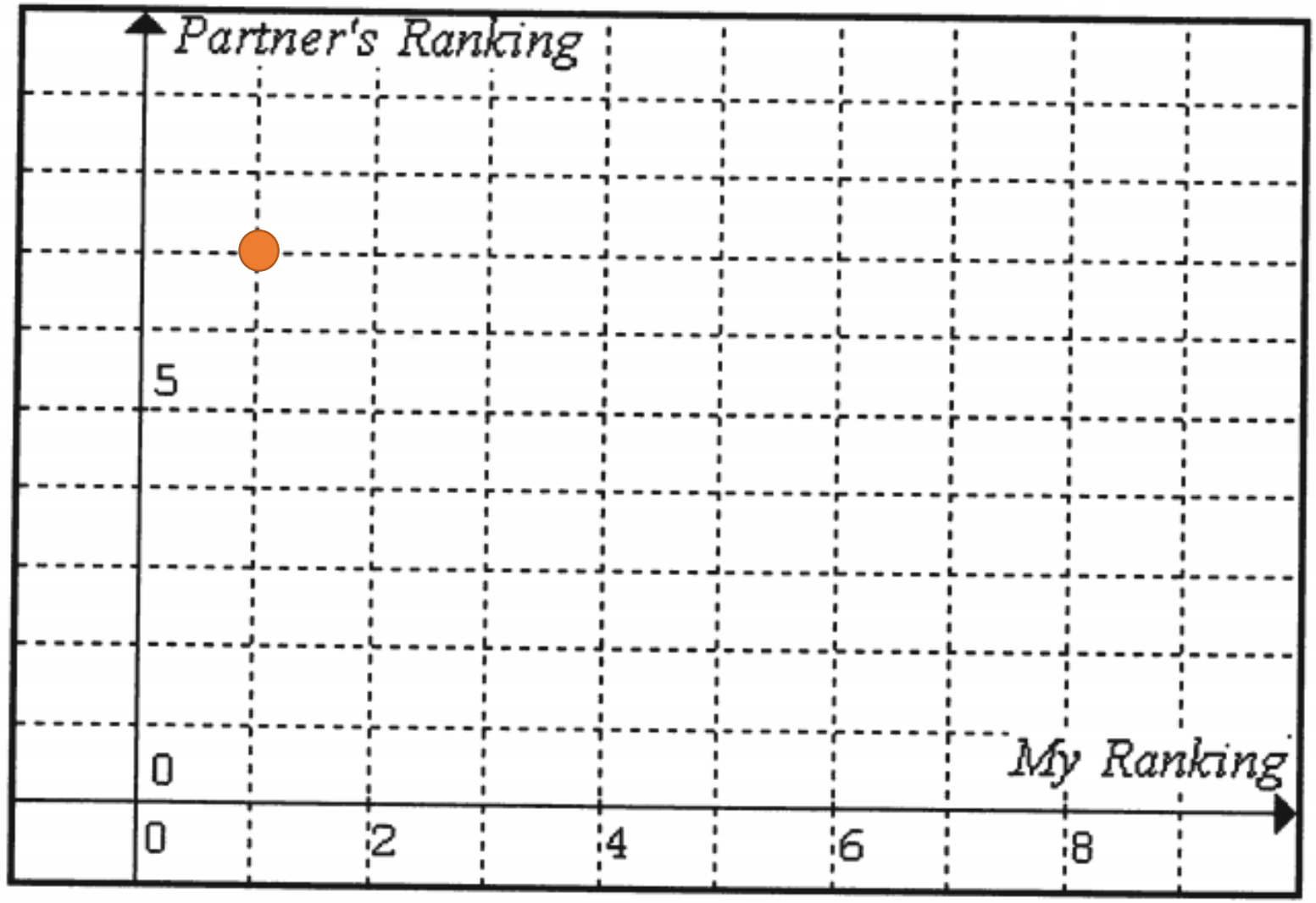
Movie Activity

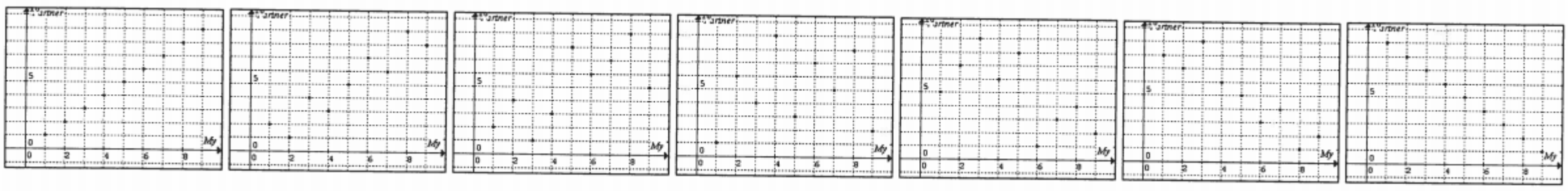
Rank from 1 to 10. No repeats.

10 = best
1 = worst

Your	Partner	Movie
		Joker - Arthur Fleck (Joaquin Phoenix) is a ay in Gotham's fractured society.
		Gemini Man - An action-thriller starring Will Sm assin, who is suddenly targeted and pursued by a mysterious young operative that seemingly can predict his every move.
		Addams Family - The first family of Halloween is back on the big s first animated comedy about the kookiest family on the block.
		Abominable - When teenage Yi encounters a young Yeti on building in Shanghai, she and her mischievous friends, embark on an epic quest
		Maleficent II - Maleficent and Aurora begin to question th s they are pulled in different directions by impending nuptials, unexpected alli
		Downton Abbey - A royal visit from the King and Queen of En andal, romance and intrigue that will leave the future of Downton hanging in the balanc
		IT: Chapter Two - Twenty-seven years after the Losers Club defeated Pennywise, he has returned to terrorize the town of Derry once more.
		Judy - Winter 1968 and showbiz legend Judy Garland arrives in Swinging London to perform a five-week sold-out run at The Talk of the Town.
		Jexi - When Phil (Adam Devine) is forced to upgrade his phone, the latest model comes with an unexpected feature, Jexi -- an A.I. life coach, virtual assistant and cheerleader.
		The Peanut Butter Falcon - Zak, a young man with Down syndrome, who runs away from a residential nursing home to follow his dream of attending the professional wrestling school of his idol.

Your	Partner	Movie
1	7	Joker - Arthur Fleck (Joaquin Phoenix) is a man struggling to find his way in Gotham's fractured society.
		Gemini Man - An action-thriller starring Will Smith as an elite assassin, who is suddenly targeted and pursued by a mysterious young operative that seemingly can predict his every move.
		Addams Family - The f about the kookiest fan
		Abominable - When te she and her mischievo
		Maleficent II - Malefice different directions by
		Downton Abbey - A ro intrigue that will leave
		IT: Chapter Two - Twe terrorize the town of D
		Judy - Winter 1968 an sold-out run at The Ta
		Jexi - When Phil (Adan unexpected feature, Je
		The Peanut Butter Fal nursing home to follov





r = 1	r = 0.8 (approx.)	r = 0.6 (approx.)	r = 0	r = – 0.6	r = – 0.8 (approx.)	r = – 1 (approx.)
Perfect line	Not so perfect	Even less perfect	Big blob	Not so perfect	More perfect	Perfect line
Positive slope	Positive slope	Positive slope	No slope	Negative slope	Negative slope	Negative slope
If you like it, so does your partner	If you like it, your partner probably does too.	If you like it, your partner might too	If you like it, you have no idea if your partner does	If you like it, your partner might not	If you like it, your partner probably does not	If you like it, your partner does not
Your value predicts your partners'	Your value sort of predicts your partners'	Your value rarely predicts your partners'	Your value has no relation to your partners'	The opposite of your value rarely predicts your partners'	The opposite of your value sort of predicts your partners'	The opposite of your value predicts your partners'

Co-efficient of Determination

r^2

r^2 is between 0 and 1.

It represents the proportion of the variation in one variable that can be explained by the other.

Only used in linear models.

If r^2 is 0.93, then 93% of the variation in Y is due to X.

X is a student's Science aptitude score.
Y is a student's Average.

r is calculated to be 0.8.
What is r^2?

What does the r^2 value mean?

X is a student's Science aptitude score.

Y is a student's Average.

r is calculated to be 0.8.

What is r^2?

$= 0.8 \times 0.8$

$= 0.64$

What does the r^2 value mean?

X is a student's Science aptitude score.
Y is a student's Average.

r is calculated to be 0.8.

What is r^2? $= 0.8 \times 0.8$

$= 0.64$

What does the r^2 value mean?

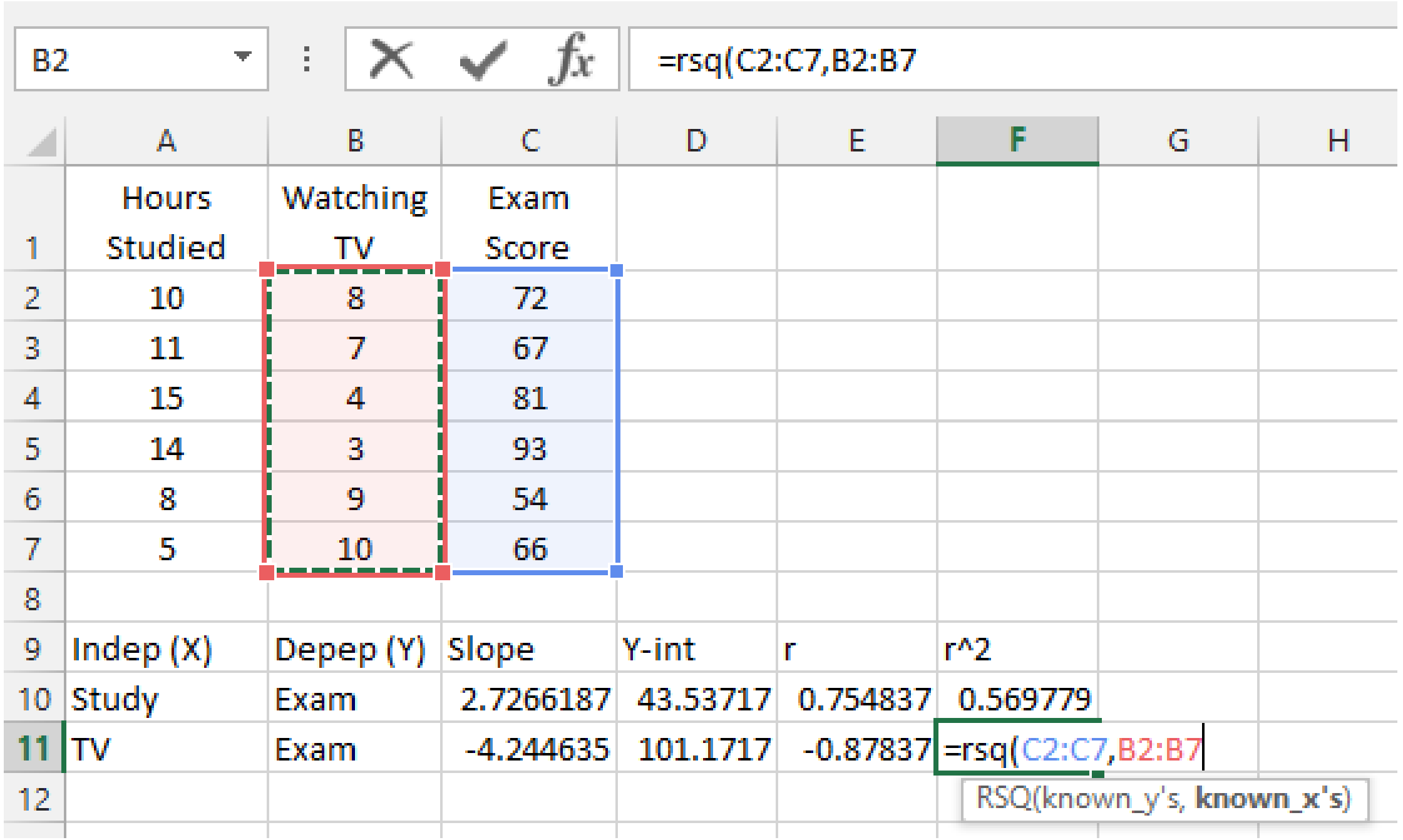
64% of the variation in your average is due to your science aptitude. Maybe you are taking two science courses.

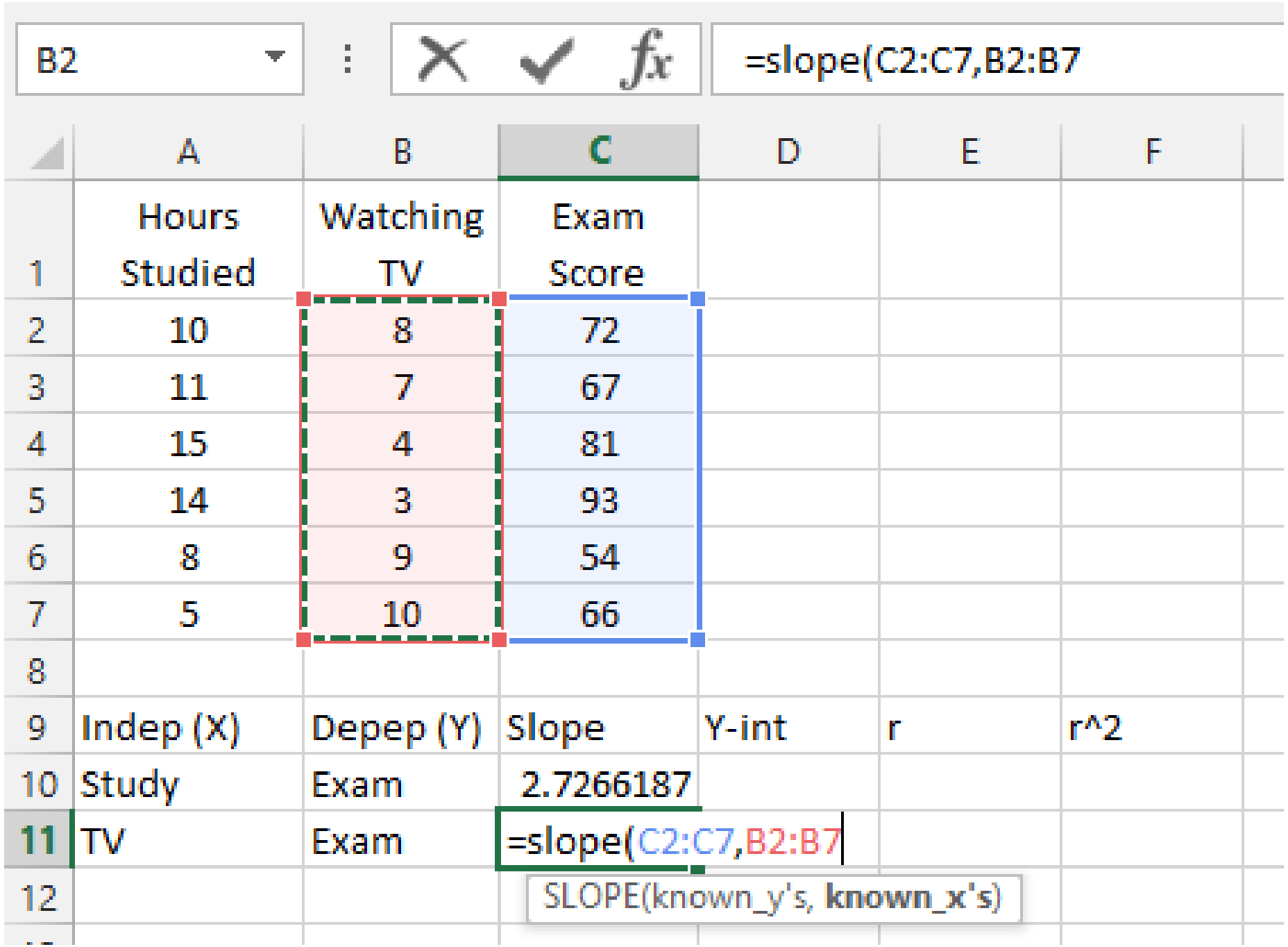
Suppose you are running an experiment to plot the likelihood of victory in a certain sport (y).

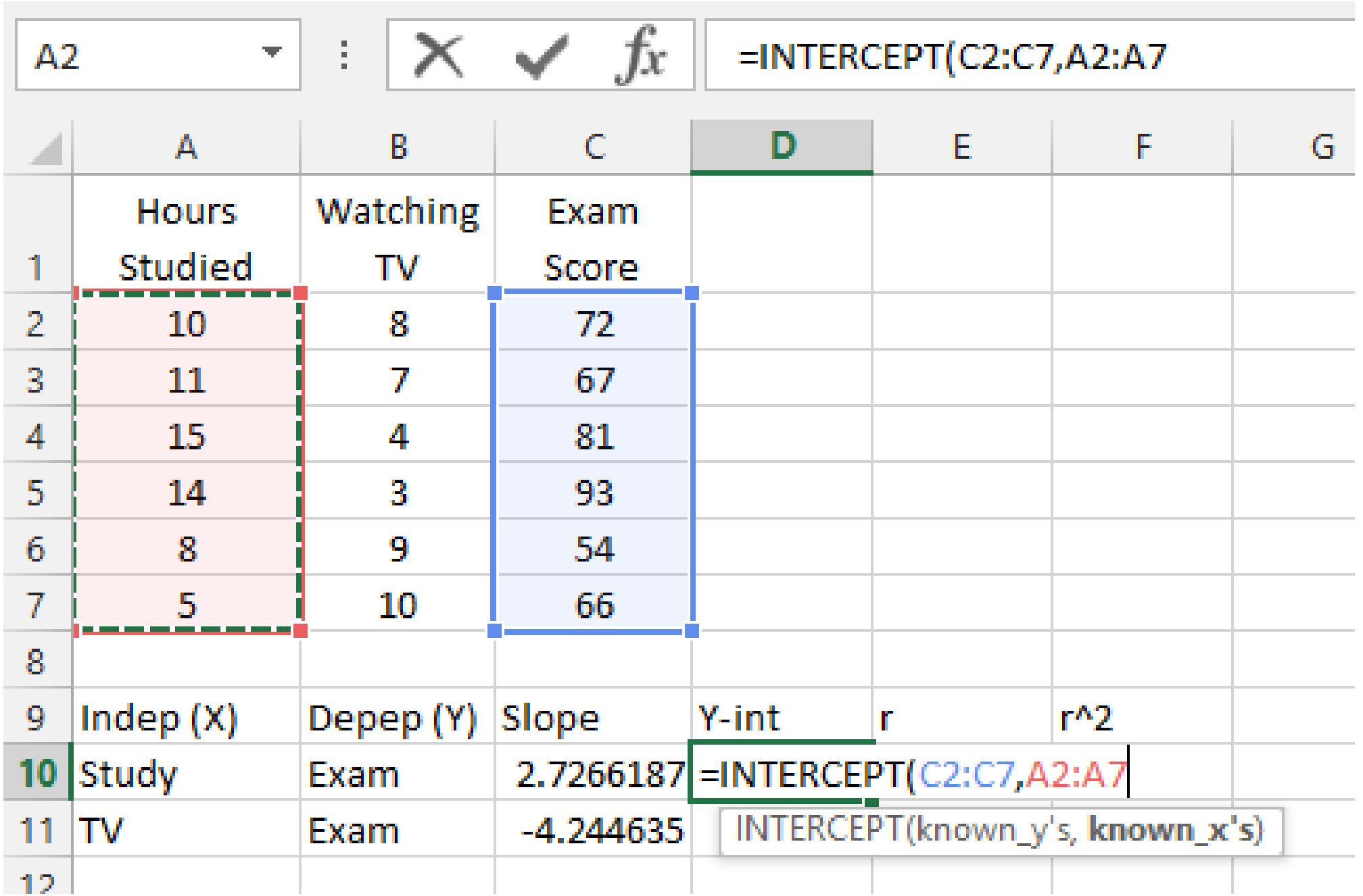
You plot each of these x values. What do the r2 values tell you about the importance of each factor in determining y?

Is this causation?

X – Dependent	r^2
Hours of Practice	0.2
Natural Ability Score	0.1
Opponent's Ranking	0.3
Attitude Score	0.1


B2
=rsq(C2:C7,B2:B7
A
B
C
D
E
F
G
H
Hours Studied
Watching TV
Exam Score
10
8
72
11
7
67
15
4
81
14
3
93
8
9
54
5
10
66
Indep (X)
Depep (Y)
Slope
Y-int
r
r^2
Study
Exam
2.7266187
43.53717
0.754837
0.569779
TV
Exam
-4.244635
101.1717
-0.87837
=rsq(C2:C7,B2:B7
RSQ(known_y's, known_x's)


B2
=slope(C2:C7,B2:B7
A
B
C
D
E
F
1
Hours Studied
Watching TV
Exam Score
2
10
8
72
3
11
7
67
4
15
4
81
5
14
3
93
6
8
9
54
7
5
10
66
8
9
Indep (X)
Depep (Y)
Slope
Y-int
r
r^2
10
Study
Exam
2.7266187
11
TV
Exam
=slope(C2:C7,B2:B7
12
SLOPE(known_y's, known_x's)


A2
=INTERCEPT(C2:C7,A2:A7
A
B
C
D
E
F
G
1
Hours Studied
Watching TV
Exam Score
2
10
8
72
3
11
7
67
4
15
4
81
5
14
3
93
6
8
9
54
7
5
10
66
8
9
Indep (X)
Depep (Y)
Slope
Y-int
r
r^2
10
Study
Exam
2.7266187
=INTERCEPT(C2:C7,A2:A7
11
TV
Exam
-4.244635
INTERCEPT(known_y's, known_x's)
12