Artificial iNtelligence
AI in Best Way

AN INTRODUCTION TO STATISTICS

LECTURE 5

BY/ ALY MAHER ABDELFATTAH

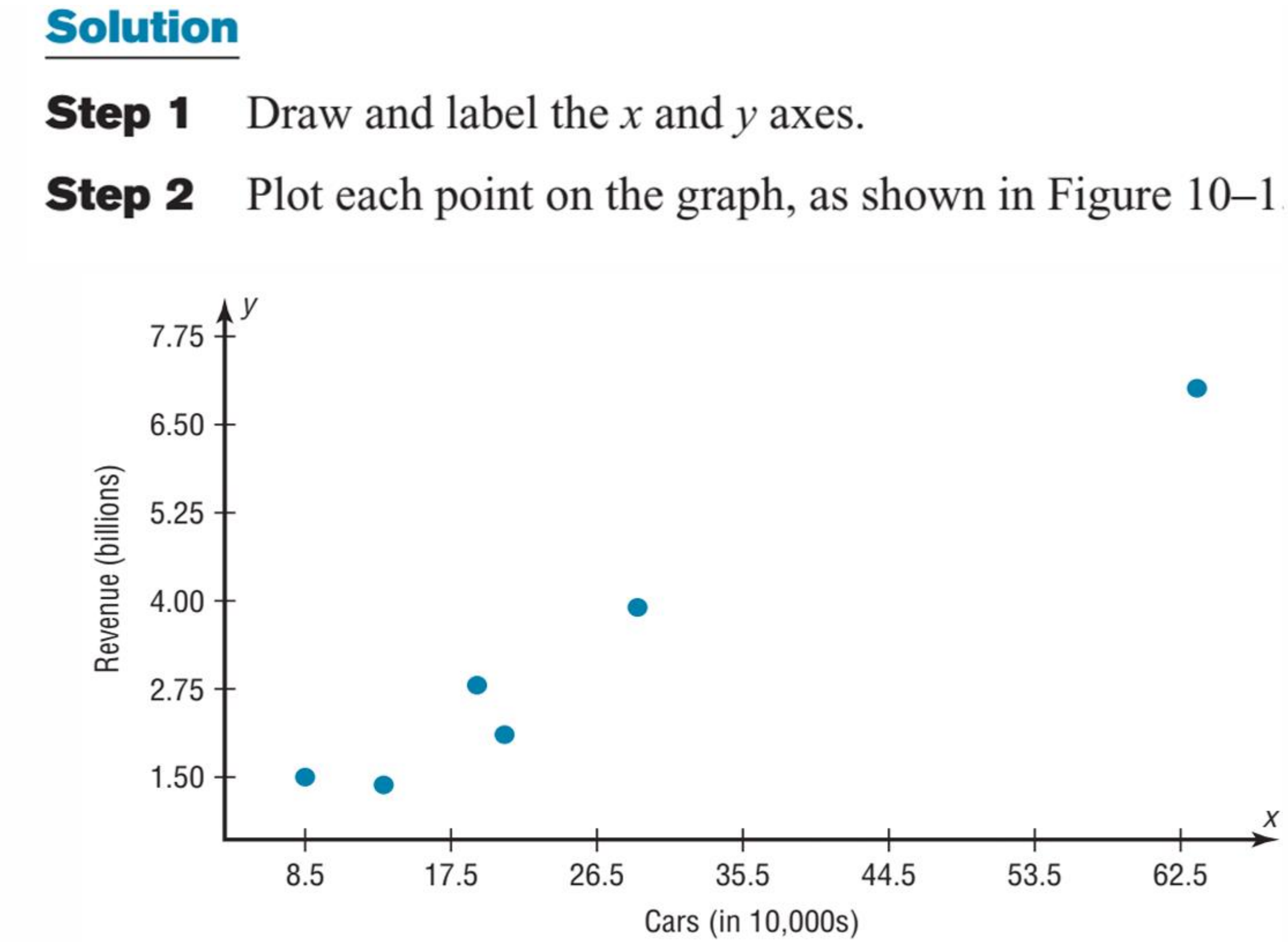
SCATTER PLOT

- A scatter plot is a graph of the ordered pairs (x , y) of numbers consisting of the independent variable x and the dependent variable y.

- Construct a scatter plot for the data shown for car rental companies in the united states for a recent year

Company	Cars (in 10 thousands)	Revenue (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

SCATTER PLOT



Solution

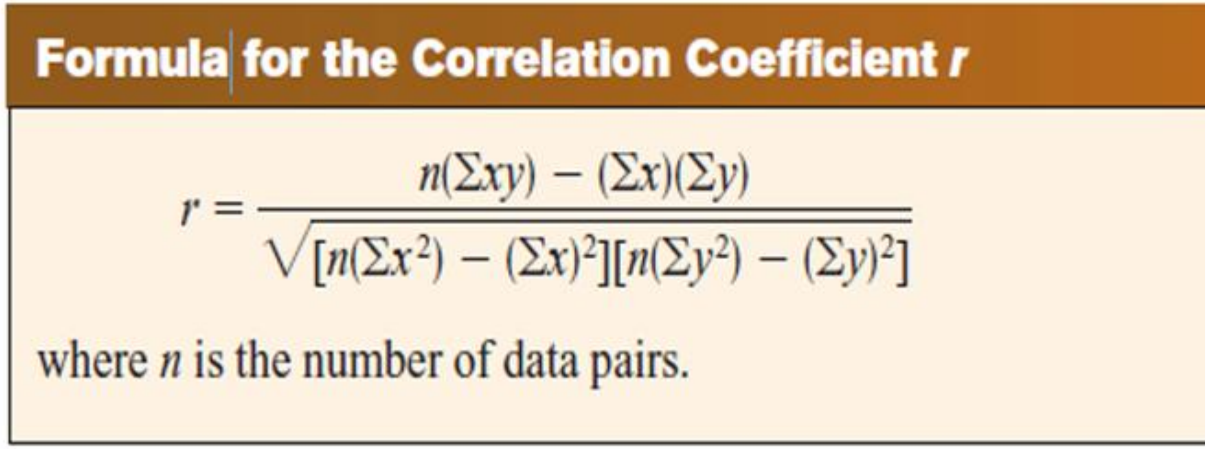
Step 1 Draw and label the *x* and *y* axes.

Step 2 Plot each point on the graph, as shown in Figure 10–1.

CORRELATION COEFFICIENT

- The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two quantitative variables. The symbol for the sample correlation coefficient is r.
- The range of the correlation coefficient is from -1 to +1. if there is a strong positive linear relationship between the variables, the value of r will be close to +1. if there is a strong negative linear relationship between the variables, the value of r will be close to -1. when there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.

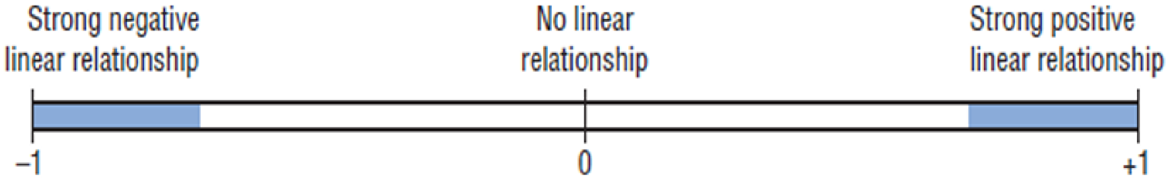
CORRELATION COEFFICIENT

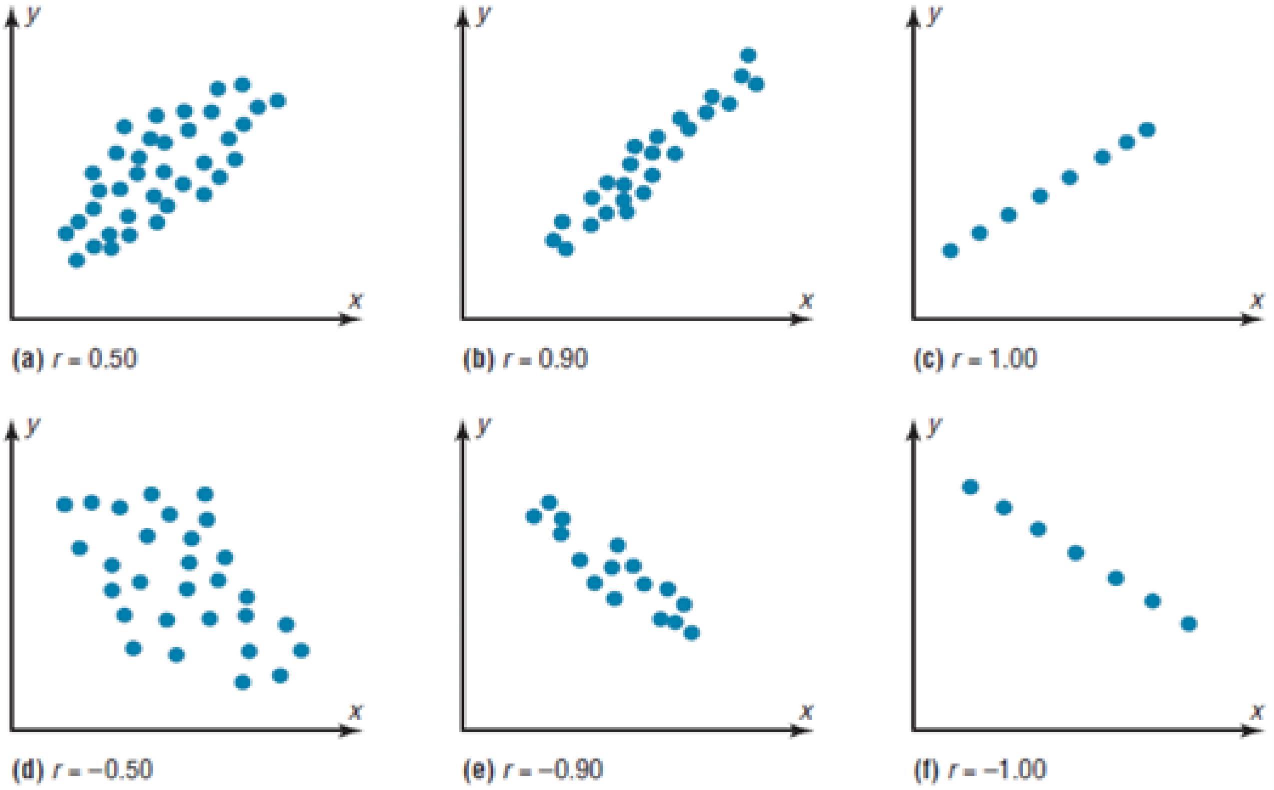


Formula for the Correlation Coefficient *r*

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

where n is the number of data pairs.


Strong negative linear relationship
No linear relationship
Strong positive linear relationship
−1
0
+1

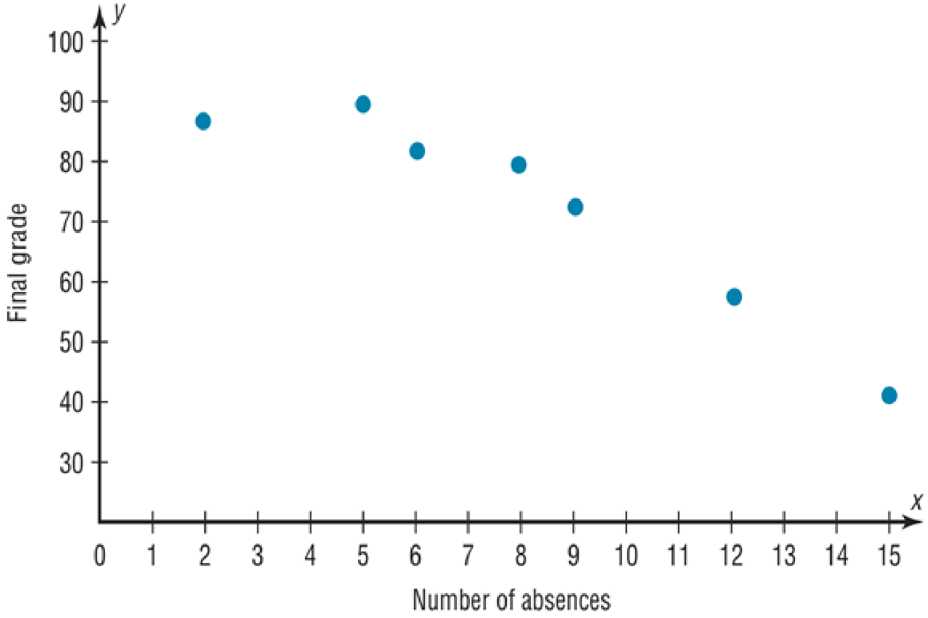


(a) $r = 0.50$ (b) $r = 0.90$ (c) $r = 1.00$

(d) $r = -0.50$ (e) $r = -0.90$ (f) $r = -1.00$

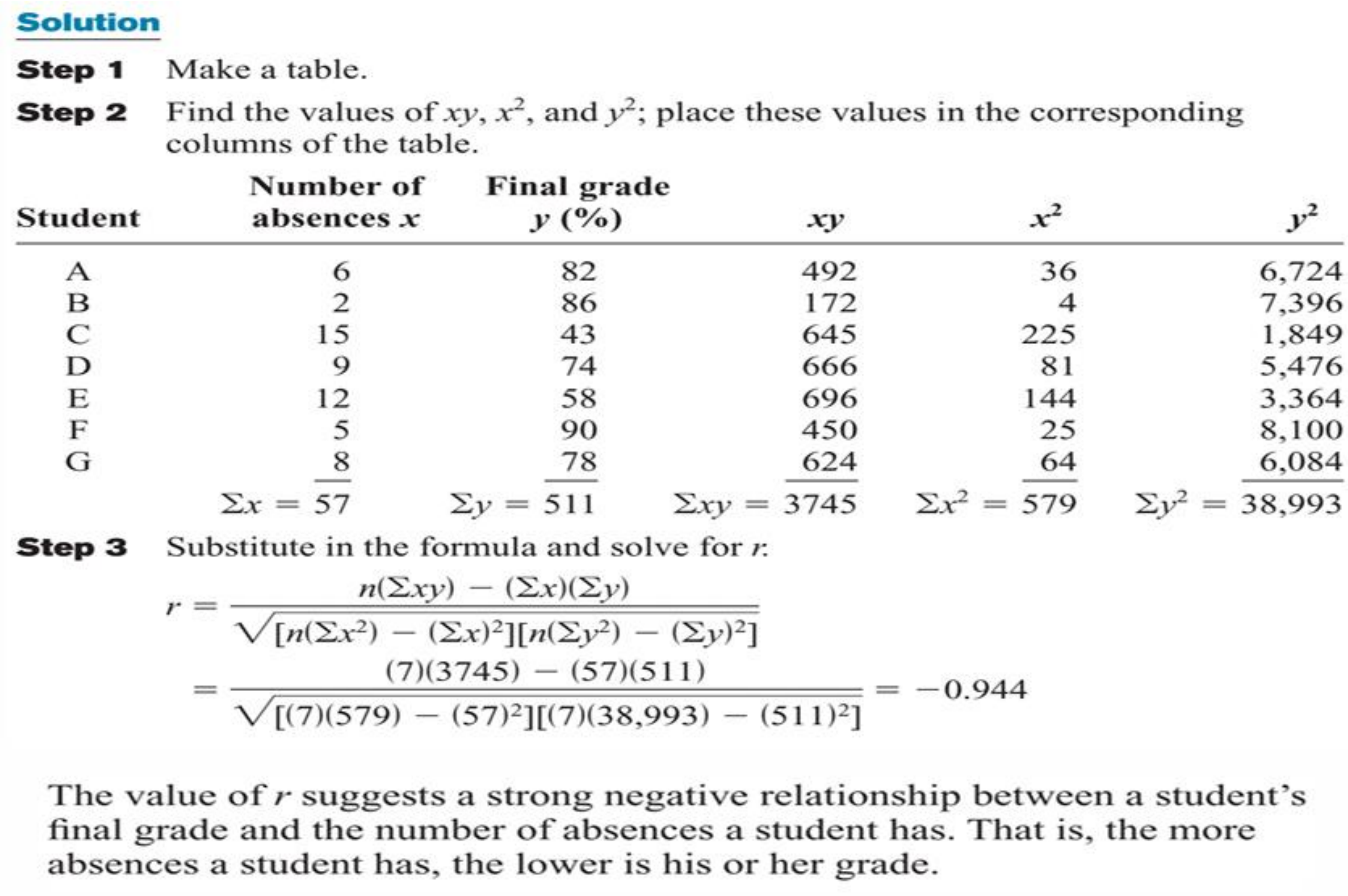
SCATTER PLOT & CORRELATION COEFFICIENT (EXAMPLE)

- Construct a scatter plot and evaluate Pearson's correlation coefficient for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.



Student	Number of absence x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

SCATTER PLOT & CORRELATION COEFFICIENT (EXAMPLE)



Solution

Step 1 Make a table.

Step 2 Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the table.

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

Step 3 Substitute in the formula and solve for r:

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

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38{,}993) - (511)^2]}} = -0.944$$

The value of r suggests a strong negative relationship between a student's final grade and the number of absences a student has. That is, the more absences a student has, the lower is his or her grade.

LINEAR REGRESSION EQUATION

Linear Regression Equation:

$$Y = a + bX$$

Where;

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

And

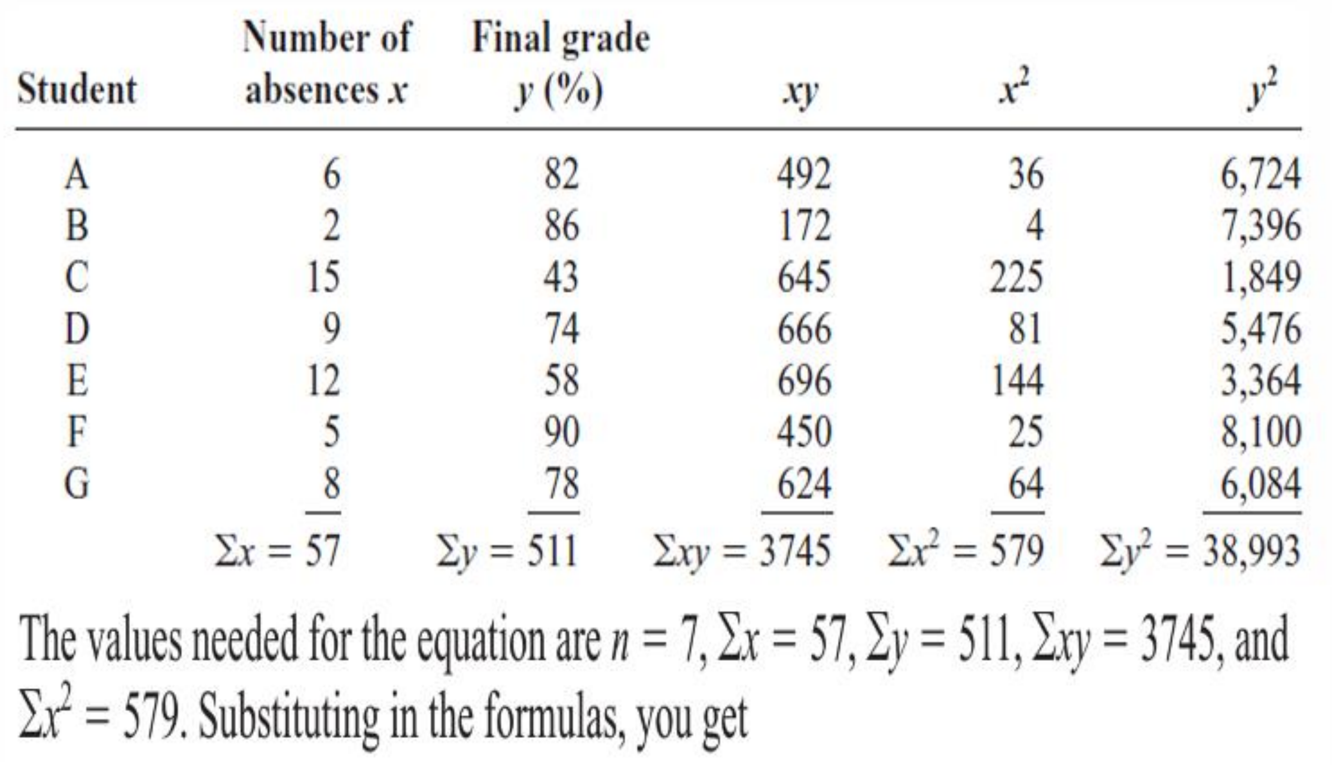
$$a = \left(\frac{\sum Y}{n}\right) - b\left(\frac{\sum X}{n}\right)$$

LINEAR REGRESSION EQUATION (EXAMPLE)

- Find the regression line for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

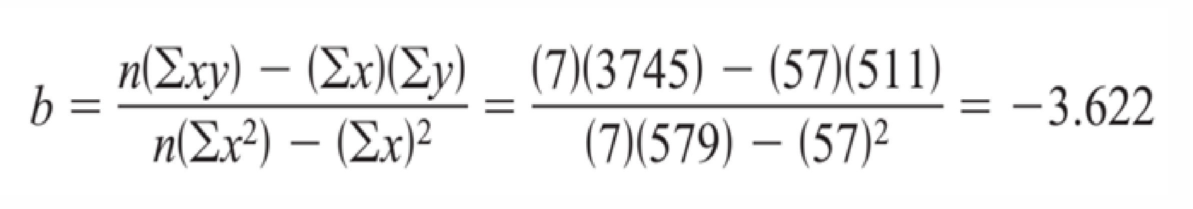
Student	Number of absence x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

LINEAR REGRESSION EQUATION (EXAMPLE)

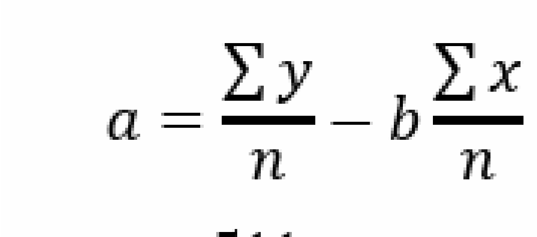


Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

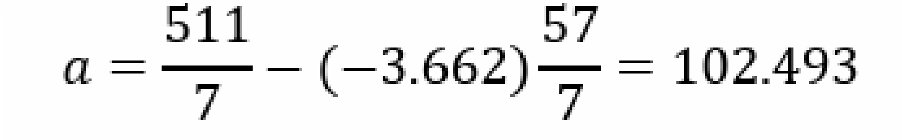
The values needed for the equation are $n = 7$, $\Sigma x = 57$, $\Sigma y = 511$, $\Sigma xy = 3745$, and $\Sigma x^2 = 579$. Substituting in the formulas, you get



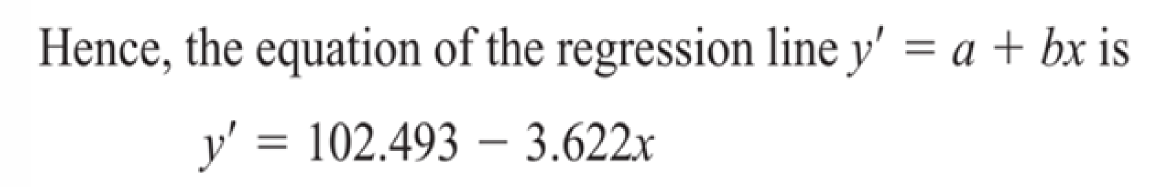
$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} = -3.622$$



$$a = \frac{\sum y}{n} - b\frac{\sum x}{n}$$



$$a = \frac{511}{7} - (-3.662)\frac{57}{7} = 102.493$$



Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 102.493 - 3.622x$$

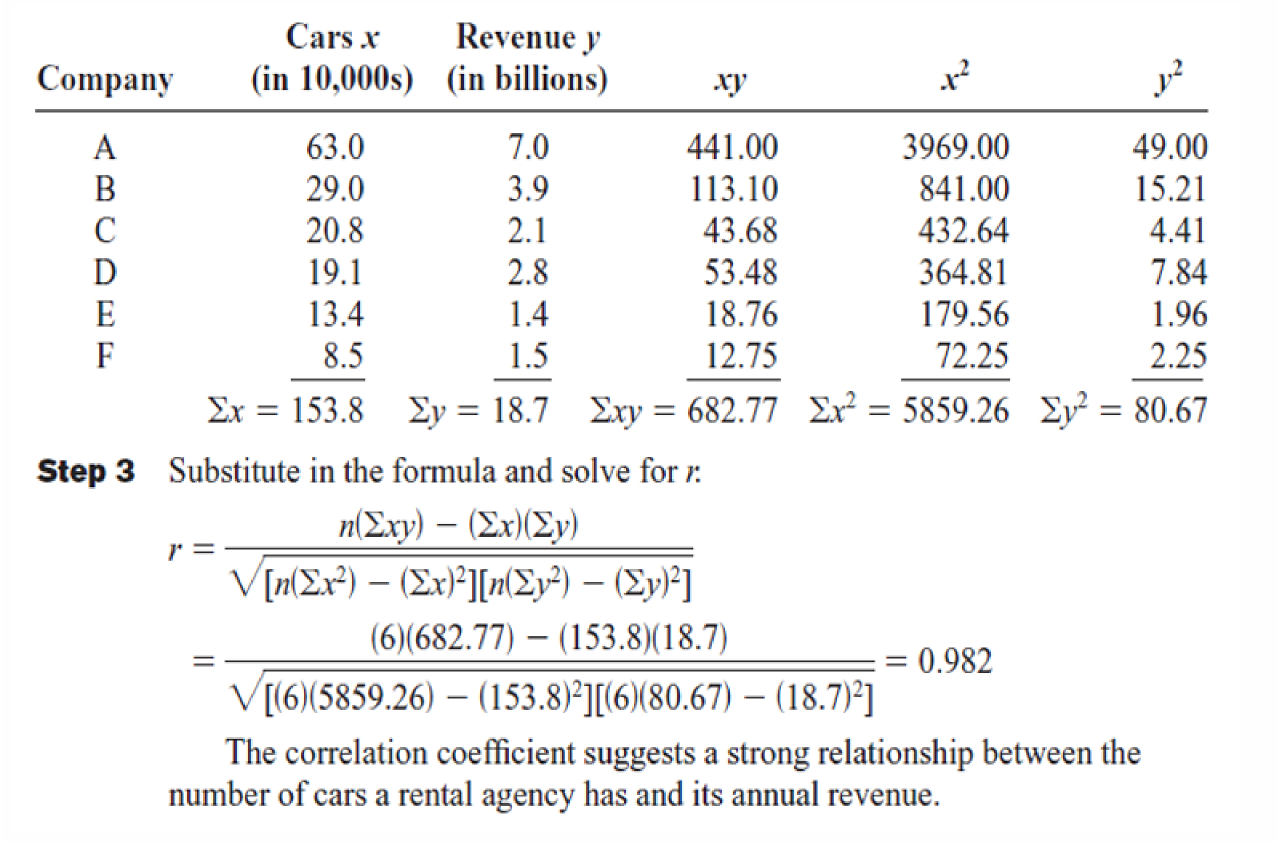
REGRESSION & CORRELATION COEFFICIENT (EXAMPLE)

- Evaluate the Pearson’s correlation coefficient and regression line for the data shown for car rental companies in the United States for a recent y.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

REGRESSION & CORRELATION COEFFICIENT (EXAMPLE)

Pearson's correlation coefficient



Company	Cars x (in 10,000s)	Revenue y (in billions)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

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

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The correlation coefficient suggests a strong relationship between the number of cars a rental agency has and its annual revenue.

regression line

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

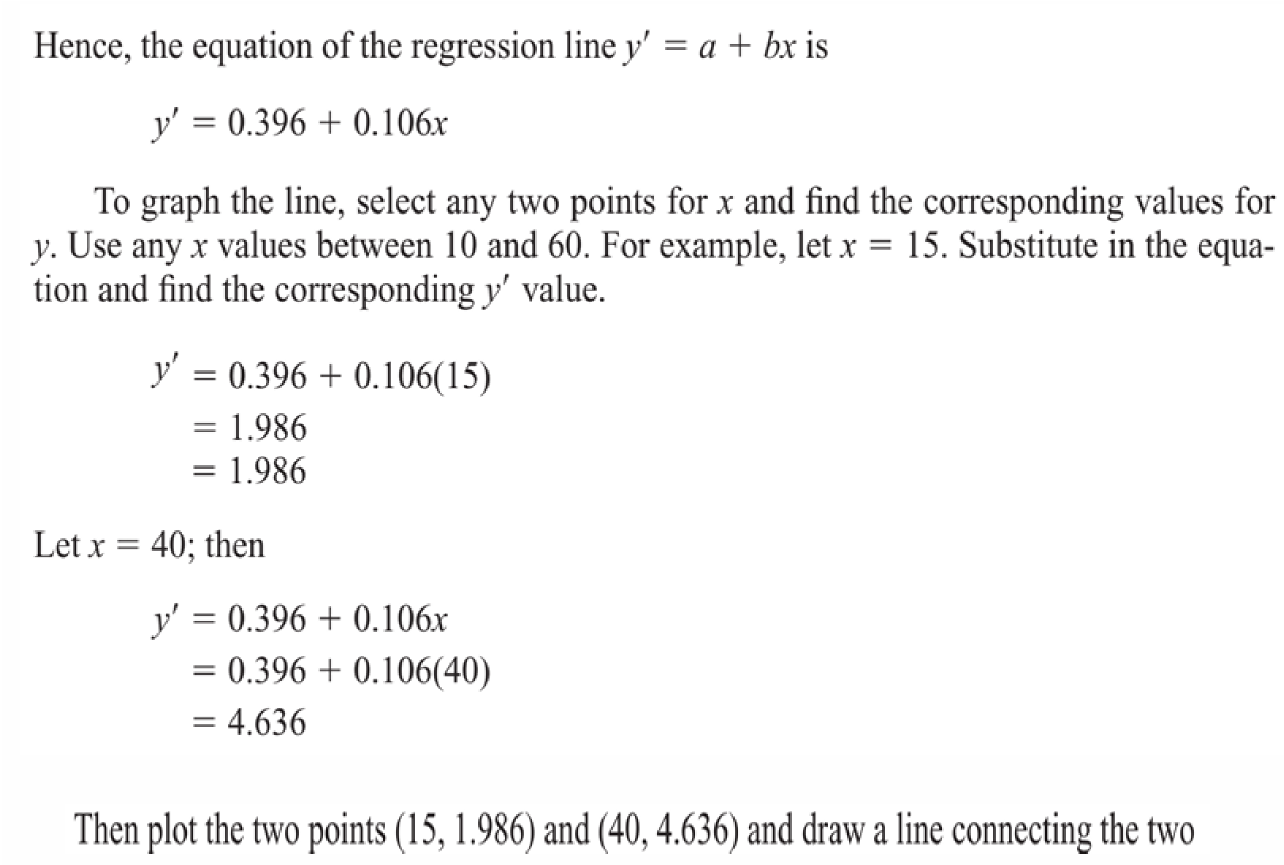
$$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n}$$

$$a = \frac{18.7}{6} - (0.106)\frac{153.8}{6} = 0.369$$

$$y' = a + bx$$

$$y' = 0.369 + 0.106x$$

REGRESSION LINE (EXAMPLE)



Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 0.396 + 0.106x$$

To graph the line, select any two points for x and find the corresponding values for y. Use any x values between 10 and 60. For example, let $x = 15$. Substitute in the equation and find the corresponding y' value.

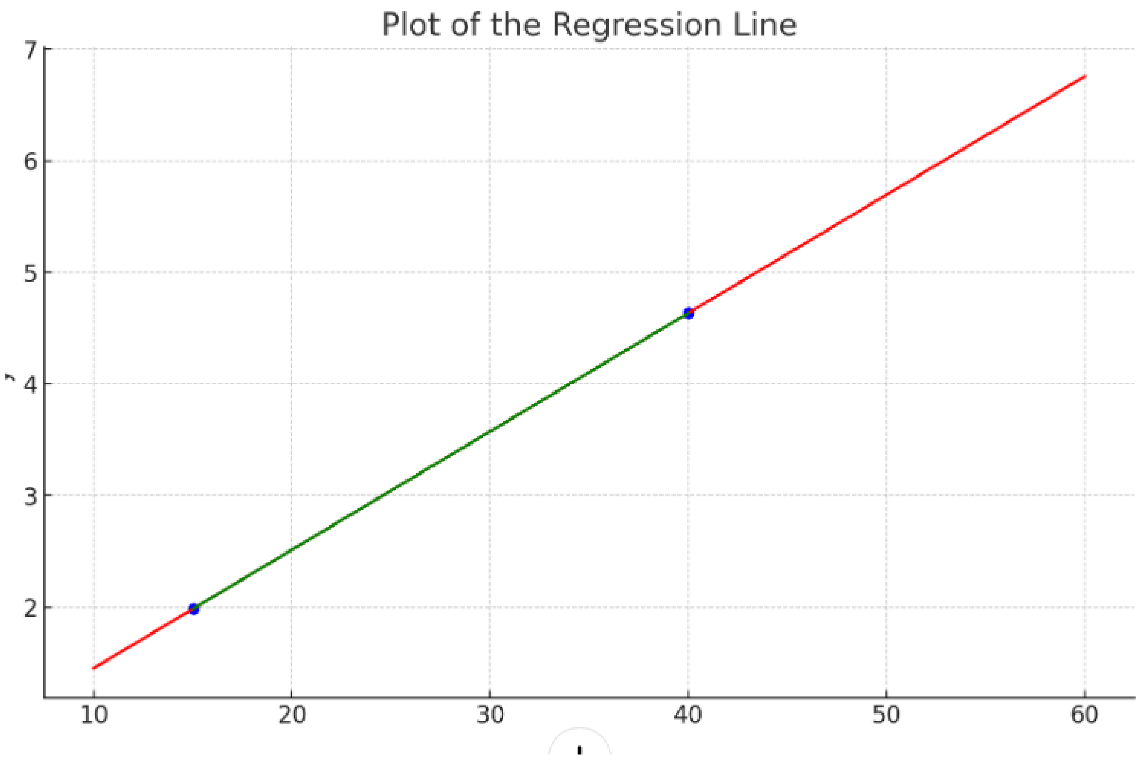
$$\begin{aligned} y' &= 0.396 + 0.106(15) \\ &= 1.986 \\ &= 1.986 \end{aligned}$$

Let $x = 40$; then

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(40) \\ &= 4.636 \end{aligned}$$

Then plot the two points (15, 1.986) and (40, 4.636) and draw a line connecting the two

Points



COEFFICIENT OF DETERMINATION

- The coefficient of determination is the ratio of the explained variation to the total variation and is denoted by r^2. That is,

r^2 = explained variation / total variation

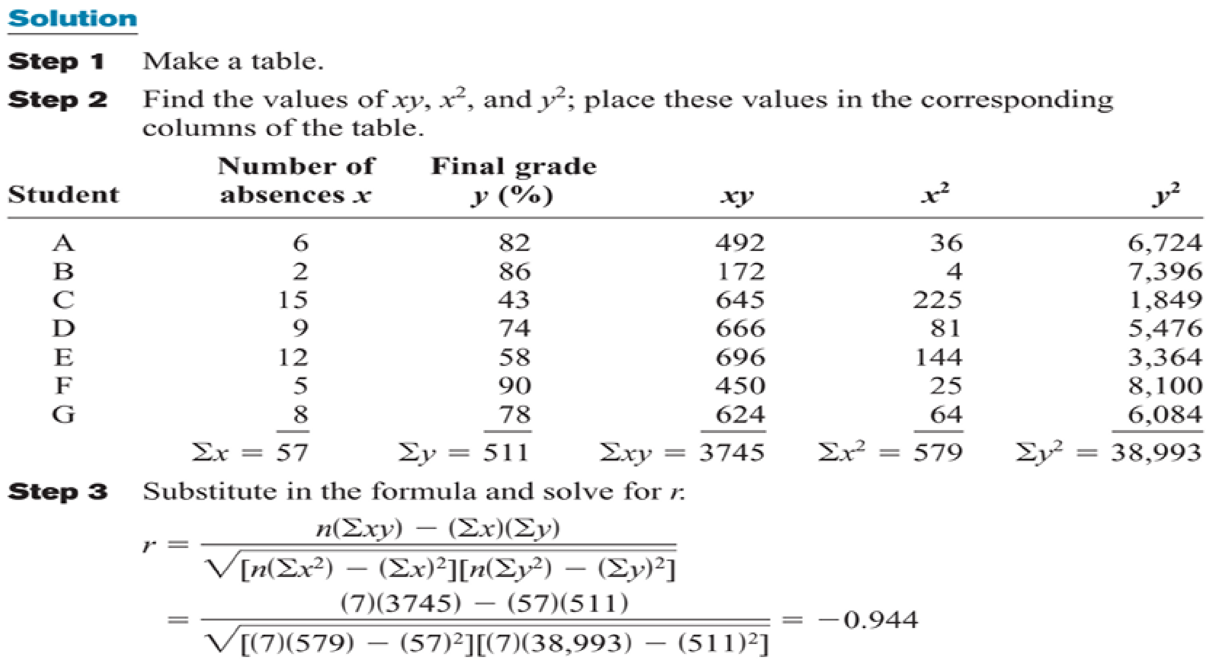
- **The coefficient of determination is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable. The symbol for the coefficient of determination is r^2.**

COEFFICIENT OF DETERMINATION (EXAMPLE)

- Find the coefficient of determination for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

Student	Number of absence x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

COEFFICIENT OF DETERMINATION (EXAMPLE)



Solution

Step 1 Make a table.

Step 2 Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the table.

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

Step 3 Substitute in the formula and solve for r.

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

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38{,}993) - (511)^2]}} = -0.944$$

$$r^2 = 0.8911$$

- This result means that 89% of the variation in the dependent variable (Final grade y) is accounted for by the variations in the independent variable (number of absence x). The rest of the variation, 0.11, or 11%, is unexplained.

SPEARMAN RANK CORRELATION

- Formula for Computing the Spearman Rank Correlation Coefficient

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

where

- ***d* = difference in ranks**
- ***n* = number of data pairs**

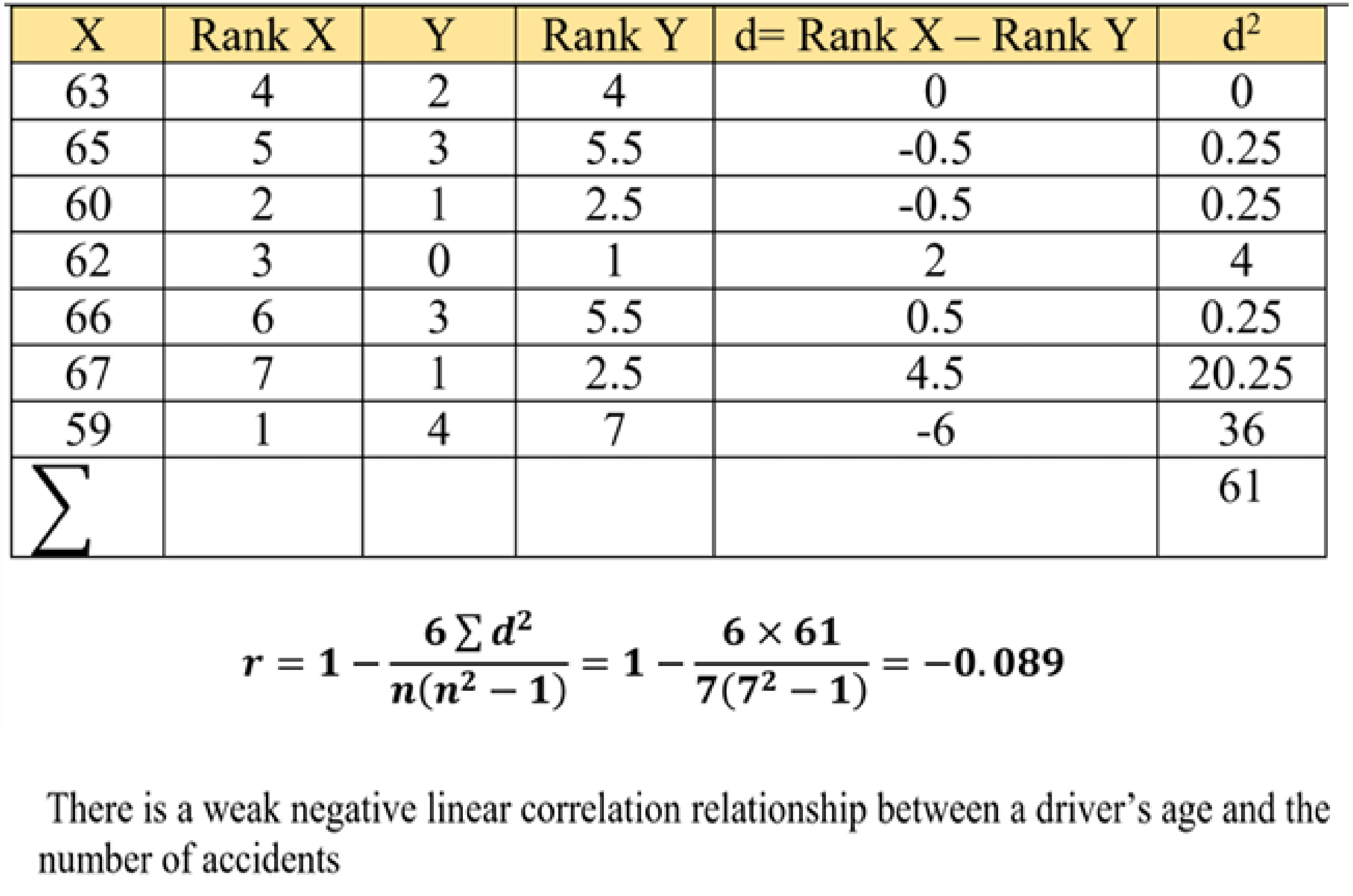
SPEARMAN RANK CORRELATION (EXAMPLE)

- A study is conducted to determine the relationship between a driver’s age and the number of accidents he or she has over a 1-year period. The data are shown here.

Driver’s age x	63	65	60	62	66	67	59
No. of accidents y	2	3	1	0	3	1	4

Calculate the Spearman’s correlation (Rank coefficient) coefficient.

SPEARMAN RANK CORRELATION (EXAMPLE)



X	Rank X	Y	Rank Y	d= Rank X – Rank Y	d^2
63	4	2	4	0	0
65	5	3	5.5	-0.5	0.25
60	2	1	2.5	-0.5	0.25
62	3	0	1	2	4
66	6	3	5.5	0.5	0.25
67	7	1	2.5	4.5	20.25
59	1	4	7	-6	36
$\sum$					61

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 61}{7(7^2-1)} = -0.089$$

There is a weak negative linear correlation relationship between a driver's age and the number of accidents