CORRELATION & REGRESSION Part 1

Question 1 (**)

The annual car sales of a small car manufacturer, c, and the annual advertising expenditure, £a, has product moment correlation coefficient r_{ac}.

The data is coded as

$$x = c - 7000 \quad \text{and} \quad y = \frac{a}{1000},$$

and the summary is shown in the table below.

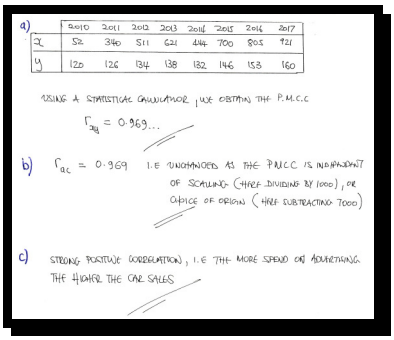
Year	2010	2011	2012	2013	2014	2015	2016	2017
x	52	340	511	621	444	700	805	921
y	120	126	134	138	132	146	153	160

a) Find, by a statistical calculator, the value of the product moment correlation coefficient between x and y, denoted by r_{xy}.

b) State with full justification the value of r_{ac}.

c) Interpret the value of r_{ac}.

$\boxed{}$, $\boxed{r_{xy} \approx 0.969}$, $\boxed{r_{ac} \approx 0.969}$



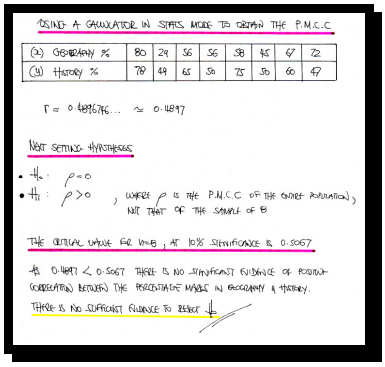
Question 2 (**)

The percentage mock exam marks, of a random sample of 8 G.C.S.E. students, in Geography and History are recorded in the table below.

Student	A	B	C	D	E	F	G	H
Geography	80	29	56	56	58	45	67	72
History	78	49	65	50	75	50	60	47

Test, at the 10% level of significance, whether there is evidence of positive correlation between the percentage mock exam marks in Geography and History.

[ ], [not significant evidence as $0.4897 < 0.5067$]


USING A CALCULATOR IN STATS MODE TO OBTAIN THE P.M.C.C

(x) GEOGRAPHY %	80	29	56	56	58	45	67	72
(y) HISTORY %	78	49	65	50	75	50	60	47

$r = 0.489576\ldots \simeq 0.4897$

NEXT SETTING HYPOTHESES

- $H_0: \rho = 0$
- $H_1: \rho > 0$, WHERE ρ IS THE P.M.C.C OF THE ENTIRE POPULATION, NOT THAT OF THE SAMPLE OF 8

THE CRITICAL VALUE FOR $n=8$, AT 10% SIGNIFICANCE IS 0.5067

AS $0.4897 < 0.5067$ THERE IS NO SIGNIFICANT EVIDENCE OF POSITIVE CORRELATION BETWEEN THE PERCENTAGE MARKS IN GEOGRAPHY & HISTORY.

THERE IS NO SUFFICIENT EVIDENCE TO REJECT H_0

Question 3 (**)

The table below shows the number of Maths teachers x, working in 8 different towns and the number of burglaries y, committed in a given month in the same 8 towns.

Town	A	B	C	D	E	F	G	H
x	35	42	21	55	33	29	39	40
y	30	28	21	38	35	27	30	k

a) Use a statistical calculator to find the product moment correlation coefficient between the number of maths teachers and the number of burglaries, for the towns A to G.

b) Interpret the value of the product moment correlation coefficient in the context of this question.

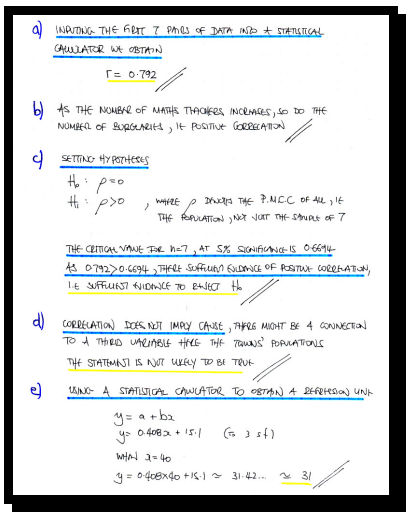
c) Test, at the 5% level of significance, whether there is evidence of positive correlation between the number of maths teachers and the number of burglaries, for the towns A to G.

d) Comment on the statement

"… the Maths teachers are likely to be responsible for the burglaries …"

e) Use linear regression to estimate the value of k, for town H.

$\boxed{}$, $\boxed{r \approx 0.792}$, $\boxed{\text{significant evidence as } 0.792 > 0.6694}$, $\boxed{k \approx 31}$



Question 4 (**)

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H
Test 1	28	39	18	30	42	43	33	10
Test 2	12	23	16	16	28	18	24	7

The first test is out of 50 marks while the second test is out of 30 marks.

Let x and y represent the marks obtained in Test 1 and Test 2, respectively.

a) Use a statistical calculator to find the value of the product moment correlation coefficient between x and y.

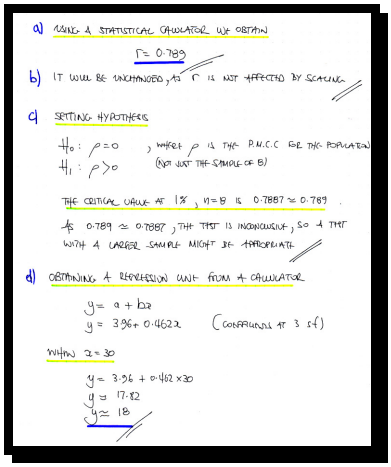
b) Explain how the value of the product moment correlation coefficient between x and y will be affected if the individual test marks were converted into percentage marks.

c) Test, at the 1% level of significance, whether there is evidence of positive correlation between x and y.

A student was absent from the second test but he obtained 30 marks in the first test.

d) Use linear regression to estimate this student's mark in the second test.

$\boxed{}$, $\boxed{r \approx 0.789}$, $\boxed{\text{unchanged}}$, $\boxed{\text{inconclusive test as } 0.789 \approx 0.7889}$, $\boxed{\approx 18}$



Question 5 (**)

The table below shows the maximum daytime temperature, in °C, at a certain city centre, and the amount of a certain pollutant in mg per litre.

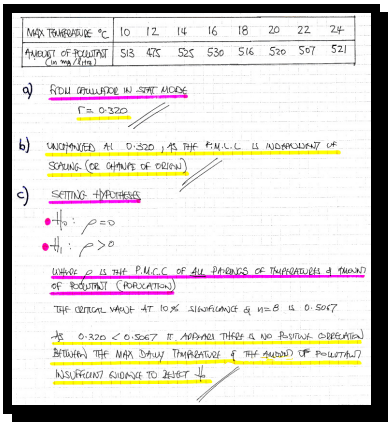
Maximum Temperature	10	12	14	16	18	20	22	24
Amount of Pollutant	513	475	525	530	516	520	507	521

a) Find, using a statistical calculator, the value of the product moment correlation coefficient for the above data.

b) State, with justification, the value of the product moment correlation coefficient, if the maximum daily temperatures were to be measured in degrees Fahrenheit.

c) Test, at the 10% level of significance, whether there is evidence of positive correlation in these bivariate data.

$\boxed{}$, $\boxed{r = 0.320}$, $\boxed{\text{unchanged}}$, $\boxed{\text{no significant evidence as } 0.320 < 0.5067}$



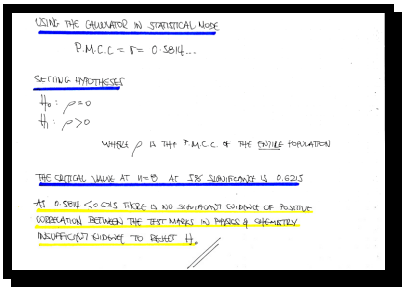
Question 6 (**)

The percentage test exam marks, of a random sample of 8 students, in Physics and Chemistry are recorded in the table below.

Student	A	B	C	D	E	F	G	H
Physics	70	36	56	56	58	45	67	72
Chemistry	78	49	55	50	75	50	60	57

Test, at the 5% level of significance, whether there is evidence of positive correlation between the percentage test marks in Physics and Chemistry.

[ ], [not significant evidence as $0.5814 < 0.6215$]



Question 7 (**)

The table below shows the daily number of shoplifting incidents in a shopping mall, for a given seven day week and the number of the security guards employed in each of these seven days.

Number of Shoplifting Incidents	17	20	23	11	35	32	21
Number of Security Guards Employed	6	6	5	7	4	3	5

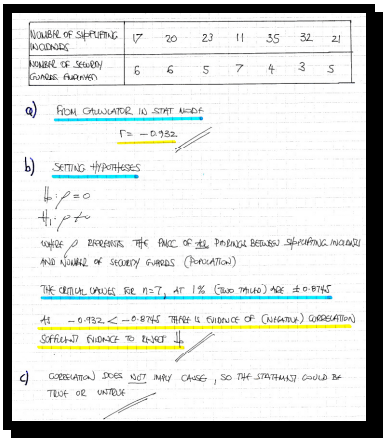
a) Find, using a statistical calculator, the value of the product moment correlation coefficient for these data.

b) Test, at the 1% level of significance, whether there is evidence of correlation in these bivariate data.

c) Briefly comment on the statement:

"… Increasing the number of security guards will result in a decrease in the shoplifting incidents …"

$\boxed{}$, $\boxed{r=-0.932}$, $\boxed{\text{significant evidence as } -0.932 < -0.8745}$



Question 89 (**)

An electrical appliances supplier wishes to investigate the impact of advertising on the sales of his washing machines.

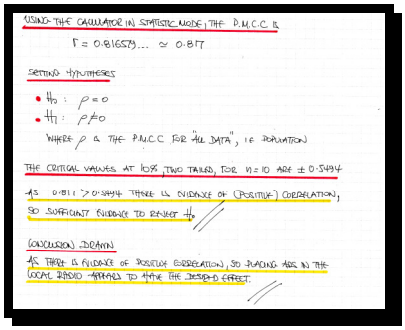
He records the number of monthly advertisements placed on the local radio station and the number of washing machines sold.

This is a table of his results.

Number of Advertisements (x)	52	37	66	45	77	27	80	19	47	40
Number of Washing Machines Sold (y)	180	115	171	166	177	99	174	100	143	164

Test, at the 10% level of significance, whether there is evidence of correlation between x and y, and explain what conclusions the electrical appliances supplier should make from this value.

[ ], [significant evidence as $0.817 > 0.5494$]



Question 9 (**)

The table below shows the number of Maths teachers x, working in 8 different schools and the number of students y, in each of these 8 schools.

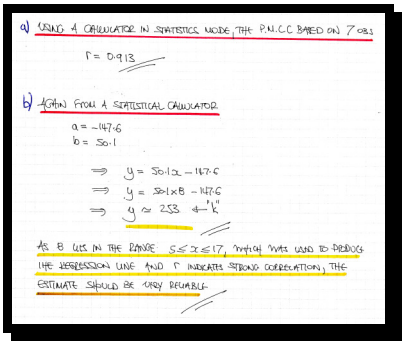
School	A	B	C	D	E	F	G	H
x	5	9	11	17	12	10	9	8
y	225	247	334	811	382	340	285	k

a) Use a statistical calculator to find the product moment correlation coefficient between the number of maths teachers and the number of students, for the schools A to G.

b) Use linear regression to estimate the value of k, for school H.

Justify the reliability of the estimate.

$\boxed{}$, $\boxed{r \approx 0.913}$, $\boxed{k \approx 252 - 253}$



Question 10 (**)

The table below shows the times obtained by a group of students, in two separate runs of a lap of the school's stadium.

Student	A	B	C	D	E	F	G	H
Run 1 (sec)	65	76	71	73	76	69	60	66
Run 2 (sec)	71	78	68	68	74	75	64	66

Let x and y represent the times obtained in Run 1 and Run 2, respectively.

a) Use a statistical calculator to find the value of the product moment correlation coefficient between x and y.

b) Explain how the value of the product moment correlation coefficient between x and y will be affected if the individual times were converted into minutes.

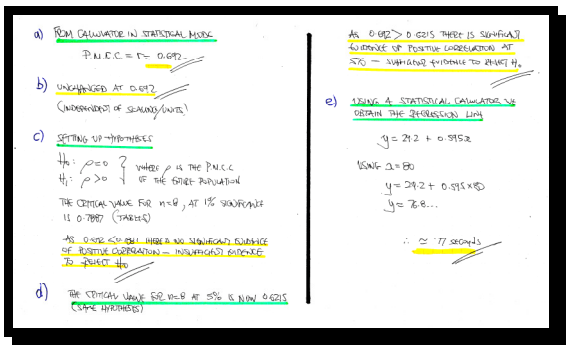
c) Test, at the 1% level of significance, whether there is evidence of positive correlation between x and y.

d) Repeat the test of part **(c)** at the 5% level of significance,

A student was absent from Run 1 but he ran Run 2 in 80 seconds.

e) Use linear regression to estimate this student's time in Run 2.

$\boxed{}$, $\boxed{r \approx 0.692}$, $\boxed{\text{unchanged}}$, $\boxed{\text{not significant at 1\% as } 0.692 < 0.789}$,
$\boxed{\text{significant at 5\% as } 0.692 > 0.6215}$, $\boxed{\approx 77}$



Question 11 (**)

The table below shows the number of priests x, working in 8 different towns and the number of shoplifting incidents y, committed in a given month in the same 8 towns.

Town	A	B	C	D	E	F	G	H
x	15	12	11	25	23	19	19	22
y	310	281	215	328	305	277	300	k

a) Use a statistical calculator to find the product moment correlation coefficient between the number of priests and the number of shoplifting incidents, for the towns A to G. **(1)**

b) Interpret the value of the product moment correlation coefficient in the context of this question. **(1)**

c) Explain how the value of the product moment correlation coefficient between x and y will be affected if the burglaries were recorded in **hundreds**. **(1)**

d) Test, at the 5% level of significance, whether there is evidence of positive correlation between the number of priests and the number of shoplifting incidents, for the towns A to G. **(4)**

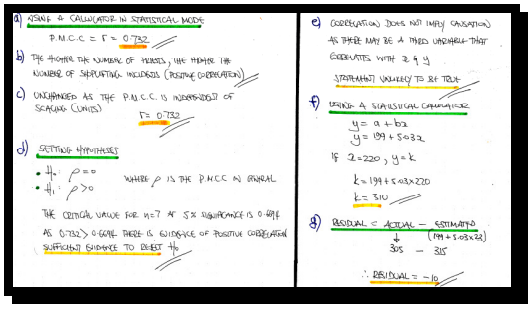
e) Comment on the statement

"… the priests are likely to be responsible for the shopliftings …" **(1)**

f) Use linear regression to estimate the value of k, for town H. **(3)**

g) Calculate the residual for town E. **(2)**

$\boxed{}$, $\boxed{r \approx 0.732}$, $\boxed{\text{significant evidence as } 0.732 > 0.6694}$, $\boxed{k \approx 310}$, $\boxed{\approx -10}$



Question 12 (***)

The table below shows the number of revolutions of a drill bit N, and the maximum temperature T °C reached by this drill bit after revolving for one minute.

N	600	700	900	1050	1300
T	49	47	52	52	53

a) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

b) Use a statistical calculator to determine …

i. … the value of the product moment correlation coefficient between N and T.

ii. … the equation of the regression line between T and N, giving the answer in the form

$$T = a + bN,$$

where a and b are constants.

c) Interpret in the context of this question the physical meaning of a and b.

Comment further on the likely value of a in a real life scenario.

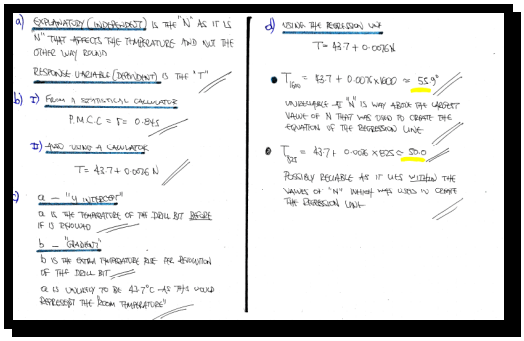
d) Use the equation of the regression line to estimate the value of T when …

i. … $N = 1600$.

ii. … $N = 825$.

Comment further on the reliability of each of these estimates.

$\boxed{}$, $\boxed{r = 0.845}$, $\boxed{T = 43.7 + 0.0076N}$, $\boxed{N_{1600} \approx 56}$, $\boxed{N_{825} \approx 50}$



Question 13 (***)

The table below shows the amount spent per month by a local radio on marketing M, in £1000, and the number of listeners L, in 1000, in that month.

M	6.5	8	11.5	14.25	15
L	87	139	119	127	147

a) Use a statistical calculator to find …

i. … the value of the product moment correlation coefficient between M and L.

ii. … the equation of the regression line between M and L, giving the answer in the form

$$L = a + bM \ ,$$

where a and b are constants.

b) Use the equation of the regression line to estimate the number of cars that are expected to be sold in a month where the amount spent on marketing and advertising is …

i. … £9,800.

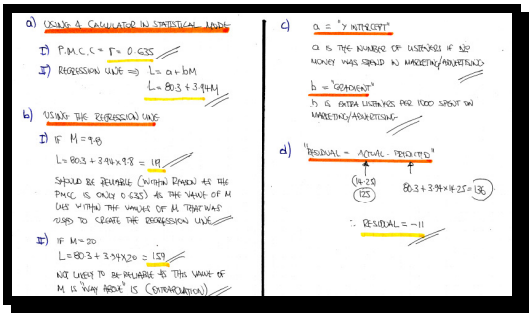
ii. … £20,000.

Comment further on the reliability of each of these two estimates.

c) Interpret in the context of this question the physical meaning of a and b.

d) Calculate the residual of the month where £14,250 was spent.

$\boxed{}$, $\boxed{r = 0.635}$, $\boxed{L = 80.3 + 3.94M}$, $\boxed{M_{9.8} \approx 119}$, $\boxed{M_{20} \approx 159}$, $\boxed{\approx -11}$



Question 14 **(***)**

The table below shows the amount spent per month by a car dealership on marketing and advertising m, in £1000, and the number of cars c sold that month.

m	6	7	8	9	10
c	8	13	11	12	14

a) Use a statistical calculator to find …

i. … the value of the product moment correlation coefficient between m and c.

ii. … the equation of the regression line between m and c, giving the answer in the form

$$c = a + bm,$$

where a and b are constants.

b) Use the equation of the regression line to estimate the number of cars that are expected to be sold in a month where the amount spent on marketing and advertising is …

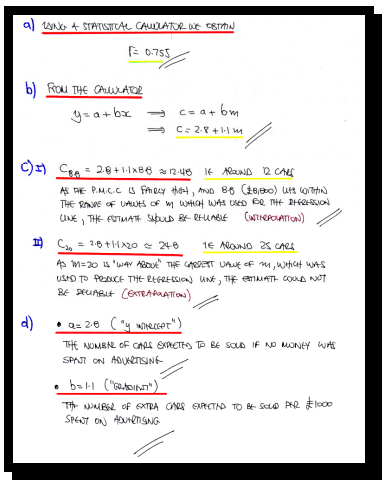
i. … £8,800.

ii. … £20,000.

Comment further on the reliability of each of these two estimates.

c) Interpret in the context of this question the physical meaning of a and b.

$\boxed{}$, $\boxed{r = 0.755}$, $\boxed{c = 2.8 + 1.1m}$, $\boxed{c_{8.8} \approx 12}$, $\boxed{c_{20} \approx 25}$



Question 15 (***)

The table below shows the maximum temperature T °C on five different days and the corresponding ice cream sales, N, of a certain shop on those days.

T	15	20	25	30	35
N	69	165	172	200	232

a) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

b) Use a statistical calculator to determine …

i. … the value of the product moment correlation coefficient between T and N.

ii. … the equation of the regression line between N and T, giving the answer in the form

$$N = a + bT ,$$

where a and b are constants.

c) Interpret in the context of this question the physical meaning of a and b.

d) Use the equation of the regression line to estimate the value of N when …

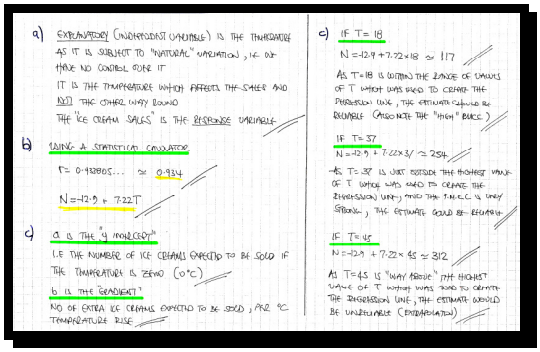
i. … $T = 18°\text{C}$.

ii. … $T = 37°\text{C}$.

iii. … $T = 45°\text{C}$

Comment further on the reliability of each of these estimates.

$\boxed{}$, $\boxed{r = 0.934}$, $\boxed{N = 7.22T - 12.9}$, $\boxed{N_{18} \approx 117}$, $\boxed{N_{37} \approx 254}$, $\boxed{T_{45} \approx 312}$



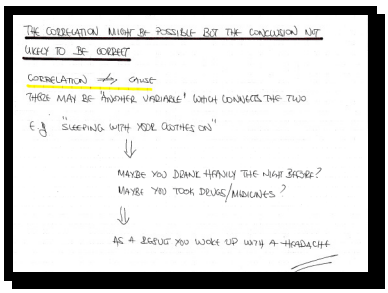
Question 16 (***)

It is an **actual fact** that "sleeping with your clothes and shoes on is strongly correlated with waking up with a headache".

Evidently the conclusion is that "sleeping with your clothes and shoes on causes a headache".

Discuss the validity of the above conclusion indicating how a strong correlation is possible in the above scenario.

[ ], [explanation as appropriate]



CORRELATION & REGRESSION Part 2

Question 1 (**)

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H
Test 1	27	38	17	29	41	42	32	9
Test 2	13	24	17	17	29	19	25	8

The first test is out of 50 marks while the second test is out of 30 marks.

Let x and y represent the marks obtained in Test 1 and Test 2, respectively.

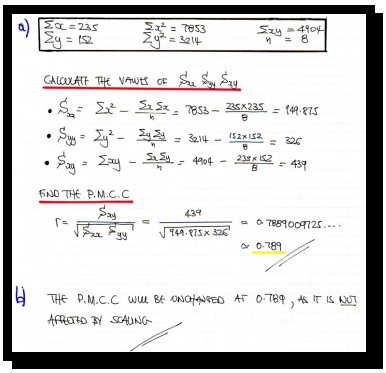
The following summary statistics are given.

$$\sum x = 235\,,\ \sum x^2 = 7853\,,\ \sum y = 152\,,\ \sum y^2 = 3214\,,\ \sum xy = 4904\,.$$

a) Find the value of the product moment correlation coefficient between x and y.

b) Explain how the value of the product moment correlation coefficient between x and y will be affected if the individual test marks were converted into percentage marks.

$\boxed{}$, $\boxed{r \approx 0.789}$, $\boxed{\text{unchanged}}$



Question 2 (**)

The table below shows the number of Maths teachers x, working in 8 different towns and the number of burglaries y, committed in a given month in the same 8 towns.

Town	A	B	C	D	E	F	G	H
x	37	40	21	50	32	27	39	40
y	30	28	20	35	34	27	31	26

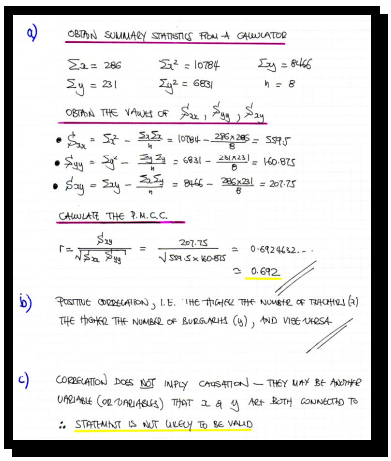
a) Calculate the product moment correlation coefficient between the number of maths teachers and the number of burglaries.

b) Interpret the value of the product moment correlation coefficient in the context of this question.

c) Comment on the statement

"… the Maths teachers are likely to be responsible for the burglaries …"

$\boxed{}$, $\boxed{r \approx 0.692}$



Question 3 (**)

An electrical appliances supplier wishes to investigate the impact of advertising on the sales of his washing machines.

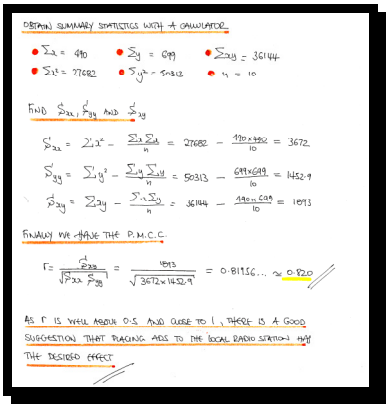
He records the number of monthly advertisements placed on the local radio station and the number of washing machines sold.

This is a table of his results.

Number of Advertisements (x)	52	37	66	45	77	27	80	19	47	40
Number of Washing Machines Sold (y)	80	75	81	76	77	49	84	50	63	64

Find, by detailed calculations, the value of the product moment correlation coefficient between x and y, and explain what conclusions the electrical appliances supplier should make from this value.

$\boxed{}$, $\boxed{r = 0.820}$



Question 4 (+)**

An electrical tester wishes to test the accuracy of a voltmeter used in a lab.

He uses a carefully calibrated voltage source and takes readings with the voltmeter he wishes to be tested.

This is a table of his results. x

Actual Voltage (x)	10	20	30	40	50	60	70	80	90	100
Voltmeter Reading (y)	9	19	34	39	54	61	68	80	92	99

a) Show, by detailed calculations, that the product moment correlation coefficient between x and y is approximately 1.

b) Determine the equation of the regression line between x and y, giving the answer in the form

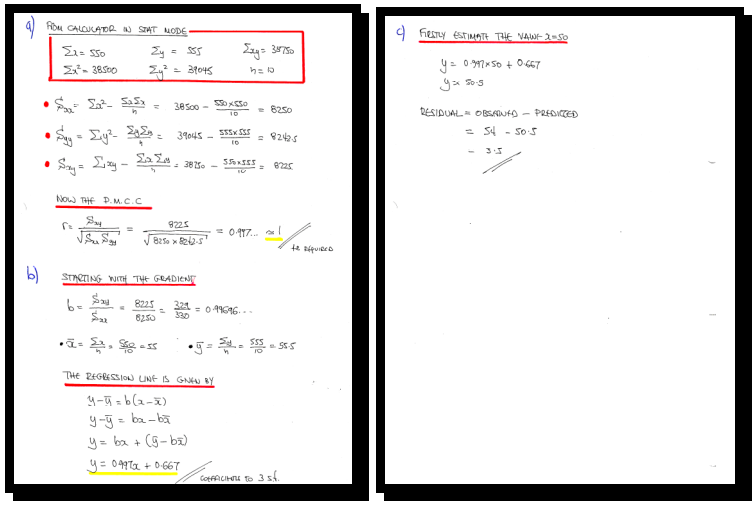
$$y = a + bx,$$

where a and b are constants.

Full workings must be shown for this part of the question.

c) Calculate the residual for $x = 50$.

[ ], $\boxed{y \approx 0.667 + 0.997x}$



a) FROM CALCULATOR IN STAT MODE

$\Sigma x = 550$ $\quad \Sigma y = 555$ $\quad \Sigma xy = 38750$

$\Sigma x^2 = 38500$ $\quad \Sigma y^2 = 39045$ $\quad n = 10$

$S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 38500 - \frac{550 \times 550}{10} = 8250$

$S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 39045 - \frac{555 \times 555}{10} = 8248.5$

$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 38750 - \frac{550 \times 555}{10} = 8225$

NOW THE P.M.C.C

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{8225}{\sqrt{8250 \times 8248.5}} = 0.997\ldots \approx 1$ AS REQUIRED

b) STARTING WITH THE GRADIENT

$b = \frac{S_{xy}}{S_{xx}} = \frac{8225}{8250} = \frac{329}{330} = 0.99696\ldots$

$\bar{x} = \frac{\Sigma x}{n} = \frac{550}{10} = 55$ $\quad \bar{y} = \frac{\Sigma y}{n} = \frac{555}{10} = 55.5$

THE REGRESSION LINE IS GIVEN BY

$y - \bar{y} = b(x - \bar{x})$

$y - \bar{y} = bx - b\bar{x}$

$y = bx + (\bar{y} - b\bar{x})$

$y = 0.997x + 0.667$ COEFFICIENTS TO 3 S.F.

c) FIRSTLY ESTIMATE THE VALUE $x = 50$

$y = 0.997 \times 50 + 0.667$

$y = 50.5$

RESIDUAL = OBSERVED − PREDICTED

$= 54 - 50.5$

$= 3.5$

Question 5 **(**+)**

The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H	I	J
Test 1	17	11	16	9	12	12	11	4	7	15
Test 2	24	21	24	20	22	18	18	9	15	21

Let x and y represent the marks obtained in Test 1 and Test 2, respectively.

d) Find the value of S_{xx}, S_{yy} and S_{xy}.

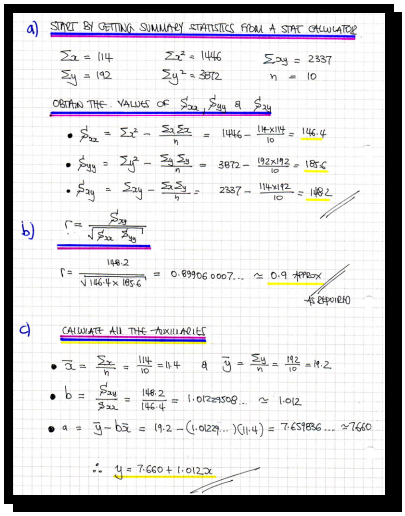
e) Show that the product moment correlation coefficient between x and y is approximately 0.9.

f) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{xx} = 146.4}$, $\boxed{S_{yy} = 185.6}$, $\boxed{S_{xy} = 148.2}$, $\boxed{y = 7.660 + 1.012x}$



Question 6 **(**+)**

The table below shows 10 pairs of bivariate data.

x	10	30	50	60	70	80	90	100	110	140
y	15	8	3	6	11	8	6	2	3	1

a) Determine the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the value of the product moment correlation coefficient between x and y.

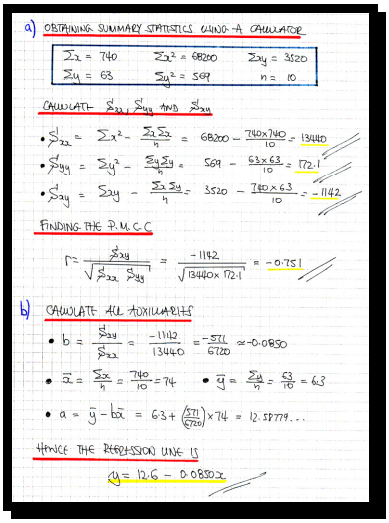
b) Find the equation of the least squares regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{xx} = 13440}$, $\boxed{S_{yy} = 172.1}$, $\boxed{S_{xy} = -1142}$, $\boxed{r = -0.751}$,

$\boxed{y = 12.6 - 0.0850x}$



Question 7 (+)**

The table below shows the heights and weights of a random sample of 10 pupils, where the heights are given to the nearest cm and the weights to the nearest $5\,\text{kg}$.

Pupil	A	B	C	D	E	F	G	H	I	J
Height (cm)	148	164	156	172	147	184	162	155	182	165
Weight (kg)	40	60	55	75	40	80	65	50	80	70

Let x and y represent the respective heights and weights of these pupils and r the product moment correlation coefficient between x and y.

a) Determine the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the value of r, correct to three decimal places.

b) Interpret in context the value of r.

c) State the value of r if the heights were measured in metres instead of cm.

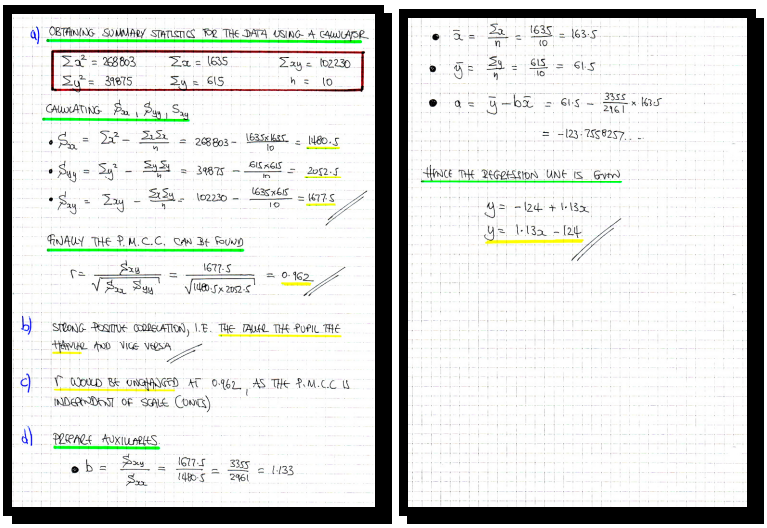
d) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = bx + a,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{xx} = 1480.5}$, $\boxed{S_{yy} = 2052.5}$, $\boxed{S_{xy} = 1677.5}$, $\boxed{r \approx 0.962}$

$\boxed{r \approx 0.962 \text{ regardless of units}}$, $\boxed{y = 1.13x - 124}$



Question 8 **(**+)**

The table below shows the midday daily temperature x, in °C, and the number of cups of tea y, sold in a small café.

x	20	25	26	27	29	29	32	36
y	100	80	72	74	65	69	63	60

a) Find the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the product moment correlation coefficient between x and y.

b) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

c) Use the equation of the regression line to estimate the value of y when …

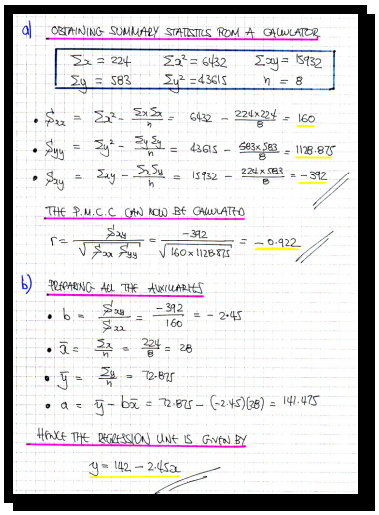
i. … $x = 40$.

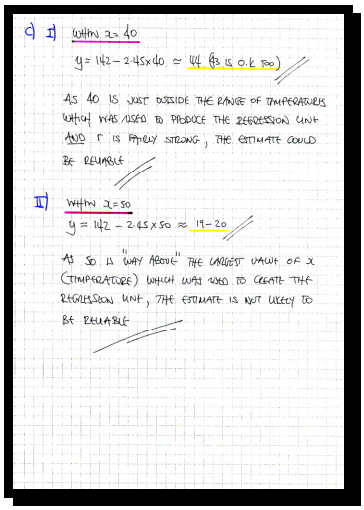
ii. … $x = 50$.

Comment further on the reliability of these two estimates.

$\boxed{}$, $\boxed{S_{xx} = 160}$, $\boxed{S_{yy} = 1128.875}$, $\boxed{S_{xy} = -392}$, $\boxed{r = -0.922}$,

$\boxed{y = 141.475 - 2.45x}$, $\boxed{y_{40} \approx 43}$, $\boxed{y_{50} \approx 19}$





Question 9 (***)

The table below shows the average midday temperature x of a seaside town, in °C, and the number of people y, that used a certain restaurant in that town.

x	17	20	25	29	27	21	20	24
y	40	42	42	43	44	39	41	45

a) Find the value of S_{xx}, S_{yy} and S_{xy}, and hence calculate the product moment correlation coefficient between x and y.

b) State the value of the product moment correlation coefficient between x and y if the temperature was measured in degrees Fahrenheit instead of Centigrade.

c) Determine the equation of the regression line between x and y, giving the answer in the form

$$y = a + bx,$$

where a and b are constants.

d) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

e) Interpret in the context of this question the physical meaning of b.

f) Use the equation of the regression line to estimate the value of y when …

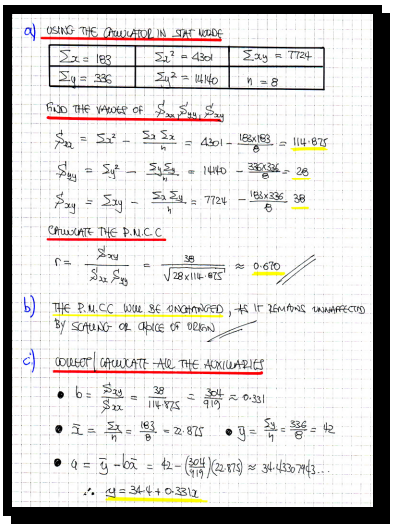
i. … $x = 16$.

ii. … $x = 35$.

Comment further on the reliability of each of these two estimates.

$\boxed{}$, $\boxed{r \approx 0.670 \text{ regardless of units}}$, $\boxed{y = 34.4 + 0.331x}$, $\boxed{y_{16} \approx 40}$, $\boxed{y_{35} \approx 46}$

[solution overleaf]



a) USING THE CALCULATOR IN STAT MODE

$\Sigma x = 183$	$\Sigma x^2 = 4301$	$\Sigma xy = 7724$
$\Sigma y = 336$	$\Sigma y^2 = 14140$	$n = 8$

FIND THE VALUES OF S_{xx}, S_{yy}, S_{xy}

$$S_{xx} = \Sigma x^2 - \frac{\Sigma x \Sigma x}{n} = 4301 - \frac{183 \times 183}{8} = 114.875$$

$$S_{yy} = \Sigma y^2 - \frac{\Sigma y \Sigma y}{n} = 14140 - \frac{336 \times 336}{8} = 28$$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 7724 - \frac{183 \times 336}{8} = 38$$

CALCULATE THE P.M.C.C

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{38}{\sqrt{28 \times 114.875}} \approx 0.670$$

b) THE P.M.C.C WILL BE UNCHANGED, AS IT REMAINS UNAFFECTED BY SCALING OR CHOICE OF ORIGIN

c) COLLECT / CALCULATE / ARE THE ACRONYMS

- $b = \frac{S_{xy}}{S_{xx}} = \frac{38}{114.875} = \frac{304}{919} \approx 0.331$
- $\bar{x} = \frac{\Sigma x}{n} = \frac{183}{8} = 22.875$ • $\bar{y} = \frac{\Sigma y}{n} = \frac{336}{8} = 42$
- $a = \bar{y} - b\bar{x} = 42 - \left(\frac{304}{919}\right)(22.875) \approx 34.4330793\ldots$

$\therefore$ $y = 34.4 + 0.331x$

d) THE EXPLANATORY VARIABLE (INDEPENDENT VARIABLE) IS THE TEMPERATURE x, AS IT IS THE TEMPERATURE THAT AFFECTS THE DINNERS y NOT THE OTHER WAY ROUND

y (NO OF DINNERS) IS CALLED THE RESPONSE VARIABLE

e) $b = 0.331$ IS THE GRADIENT OF LINE

IT REPRESENTS THE EXTRA NUMBER OF DINNERS PER DEGREE OF TEMPERATURE RISE

f) I) IF $x = 16$ $y = 34.4 + 0.331 \times 16 \simeq 40$

ALTHOUGH THE P.M.C.C IS NOT VERY STRONG, 16 IS ONLY 1 DEGREE LESS THAN THE SMALLEST VALUE OF x WHICH WAS USED TO CREATE THE REGRESSION LINE, SO IT COULD BE RELIABLE

II) IF $x = 35$ $y = 34.4 + 0.331 \times 35 \simeq 46$

GIVEN THAT THE P.M.C.C IS NOT STRONG, AND 35 IS "WAY ABOVE" THE HIGHEST TEMPERATURE WHICH WAS USED TO CREATE THE REGRESSION LINE, THE ESTIMATE WILL BE UNRELIABLE

Question 10 (***)

The table below shows the maximum temperature T °C on five different days and the corresponding ice cream sales, N, of a certain shop on those days.

T	15	20	25	30	35
N	79	145	182	255	302

a) Find the value of S_{TT}, S_{NN} and S_{TN}, and hence, determine the value of the product moment correlation coefficient between T and N.

b) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

c) Determine the equation of the regression line between N and T, giving the answer in the form

$$N = a + bT ,$$

where a and b are constants.

d) Interpret in the context of this question the physical meaning of b.

e) Use the equation of the regression line to estimate the value of N when …

i. … $T = 18°\text{C}$.

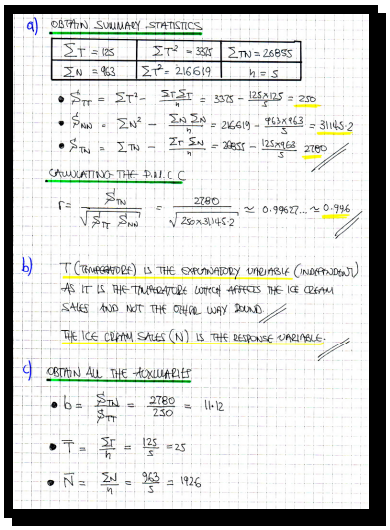
ii. … $T = 37°\text{C}$.

iii. … $T = 45°\text{C}$

Comment further on the reliability of each of these estimates.

$\boxed{}$, $\boxed{S_{TT} = 250}$, $\boxed{S_{NN} = 31145.2}$, $\boxed{S_{TN} = 2780}$, $\boxed{r = 0.996}$,
$\boxed{N = 11.12T - 85.4}$, $\boxed{N_{18} \approx 115}$, $\boxed{N_{37} \approx 326}$, $\boxed{T_{45} \approx 415}$

[solution overleaf]



a) OBTAIN SUMMARY STATISTICS

$\sum T = 125$	$\sum T^2 = 3375$	$\sum TN = 26855$
$\sum N = 963$	$\sum T^2 = 216619$	$n = 5$

- $S_{TT} = \sum T^2 - \frac{\sum T \sum T}{n} = 3375 - \frac{125 \times 125}{5} = 250$
- $S_{NN} = \sum N^2 - \frac{\sum N \sum N}{n} = 216619 - \frac{963 \times 963}{5} = 31145.2$
- $S_{TN} = \sum TN - \frac{\sum T \sum N}{n} = 26855 - \frac{125 \times 963}{5} = 2780$

CALCULATING THE P.M.C.C

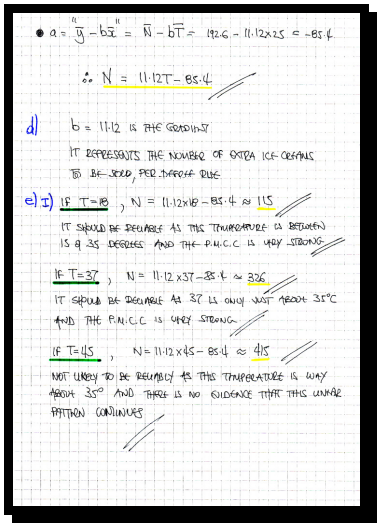
$$r = \frac{S_{TN}}{\sqrt{S_{TT} S_{NN}}} = \frac{2780}{\sqrt{250 \times 31145.2}} \approx 0.99627\ldots \approx 0.996$$

b) T (TEMPERATURE) IS THE EXPLANATORY VARIABLE (INDEPENDENT)
AS IT IS THE TEMPERATURE WHICH AFFECTS THE ICE CREAM SALES AND NOT THE OTHER WAY ROUND.

THE ICE CREAM SALES (N) IS THE RESPONSE VARIABLE.

c) OBTAIN ALL THE NECESSARIES

- $b = \frac{S_{TN}}{S_{TT}} = \frac{2780}{250} = 11.12$
- $\overline{T} = \frac{\sum T}{n} = \frac{125}{5} = 25$
- $\overline{N} = \frac{\sum N}{n} = \frac{963}{5} = 192.6$



- $a = \overline{y} - b\overline{x} = \overline{N} - b\overline{T} = 192.6 - 11.12 \times 25 = -85.4$

$\therefore N = 11.12T - 85.4$

d) $b = 11.12$ IS THE GRADIENT

IT REPRESENTS THE NUMBER OF EXTRA ICE CREAMS TO BE SOLD, PER DEGREE RISE

e) I) IF $T = 18$, $N = 11.12 \times 18 - 85.4 \approx 115$

IT SHOULD BE RELIABLE AS THE TEMPERATURE IS BETWEEN 15 & 35 DEGREES AND THE P.M.C.C IS VERY STRONG.

IF $T = 37$, $N = 11.12 \times 37 - 85.4 \approx 326$

IT SHOULD BE RELIABLE AS 37 IS ONLY JUST ABOVE 35°C AND THE P.M.C.C IS VERY STRONG.

IF $T = 45$, $N = 11.12 \times 45 - 85.4 \approx 415$

NOT LIKELY TO BE RELIABLY AS THE TEMPERATURE IS WAY ABOVE 35° AND THERE IS NO EVIDENCE THAT THIS LINEAR PATTERN CONTINUES

Question 11 (*+)**

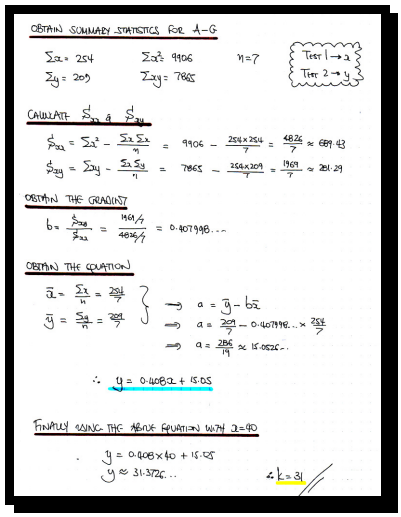
The table below shows the marks obtained by a group of students, in two separate tests.

Student	A	B	C	D	E	F	G	H
Test 1	35	42	21	55	33	29	39	40
Test 2	30	28	21	38	35	27	30	k

Use linear regression for the test marks of the students A – G , to estimate the value of k , for student H.

Detailed workings are expected.

$\boxed{}$, $\boxed{k \approx 31}$



Question 12 (*+)**

The table below shows the amount spent per month by a car dealership on marketing and advertising m, in £1000, and the number of cars c sold that month.

m	7	8	9	10	11
c	7	12	10	11	13

a) Find the value of the product moment correlation coefficient between m and c.

b) Determine the equation of the regression line between m and c, giving the answer in the form

$$c = a + bm,$$

where a and b are constants.

c) Use the equation of the regression line to estimate the number of cars that are expected to be sold in a month where the amount spent on marketing and advertising is …

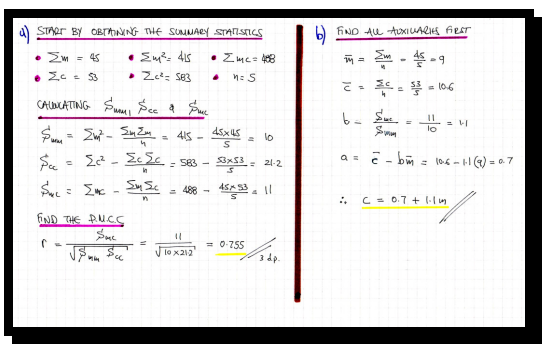
i. … £8,800.

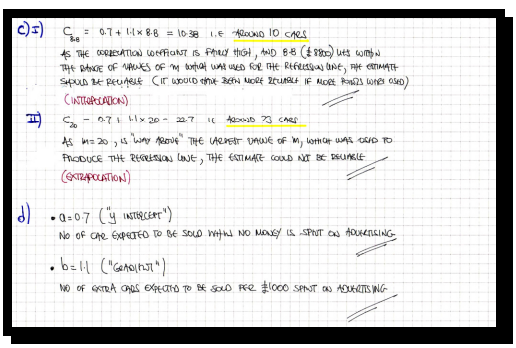
ii. … £20,000.

Comment further on the reliability of each of these two estimates.

d) Interpret in the context of this question the physical meaning of a and b.

$\boxed{}$, $\boxed{r = 0.755}$, $\boxed{c = 0.7 + 1.1m}$, $\boxed{c_{8.8} \approx 10}$, $\boxed{c_{20} \approx 23}$





Question 13 (*+)**

The table below shows the tomato yield obtained by a group of ten plants that were given different amounts of fertilizer and allowed to grow in otherwise identical conditions.

Plant	A	B	C	D	E	F	G	H	I	J
Amount of Fertilizer (grams)	0	10	20	30	40	50	60	70	80	90
Tomato Yield (kilograms)	1.2	1.9	2.1	2.4	2.5	2.7	3.0	k	3.2	3.1

a) Find an equation of the line of least squares using the plants A to G, I and J , and hence estimate the value of k , for the plant H.

Detailed workings are expected in this part

b) Interpret in context the gradient of the line of least squares.

c) Calculate the residual of plant J.

d) The residual of the plant A is -0.42. Find the …

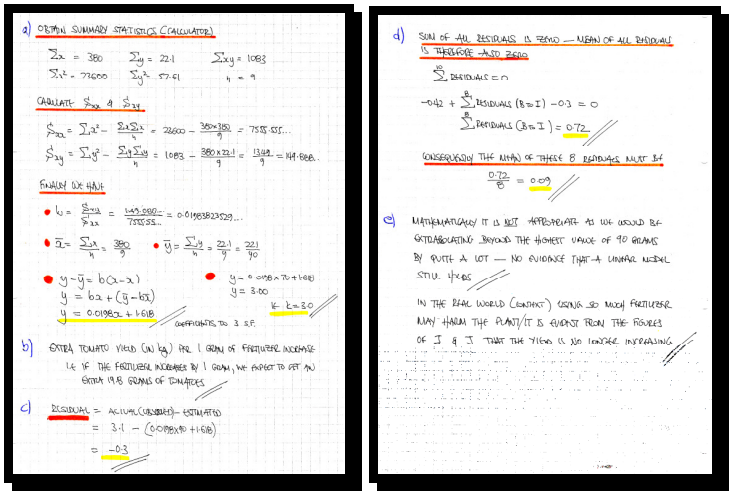
i. … sum of the residuals for the plants B to I.

ii. … mean of the residuals for the plants B to I.

Another plant N, not included in the table, was given 200 grams of fertilizer.

e) Discuss briefly, mathematically and in context, whether it is appropriate to use the line of least squares to predict its yield.

$\boxed{}$, $\boxed{Y=0.0198F+1.618}$, $\boxed{k=3.0}$, $\boxed{\mathrm{R_H}=-0.3}$, $\boxed{0.72}$, $\boxed{0.09}$



Question 14 (*+)**

The table below shows a set of bivariate data involving two variables x and y.

x	1003	1006	1012	1015	1021
y	0.0017	0.0027	0.0045	0.0056	0.0077

a) Use the coding equations

$$X=\frac{x-1012}{3} \qquad \text{and} \qquad Y=10000y-27$$

to find the value of S_{XX}, S_{YY} and S_{XY}.

b) Show that the product moment correlation coefficient between X and Y is approximately 0.9993.

c) State with justification the value of the product moment correlation coefficient between x and y.

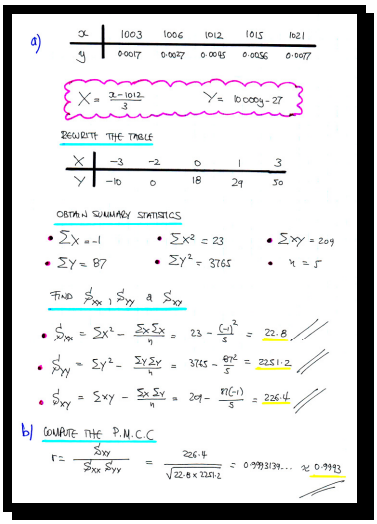
d) Determine the equation of the regression line between x and y, giving the answer in the form

$$y=a+bx,$$

where a and b are constants.

$\boxed{}$, $\boxed{S_{XX}=22.8}$, $\boxed{S_{YY}=2251.2}$, $\boxed{S_{XY}=226.4}$, $\boxed{r=0.9993}$,

$\boxed{y=0.00033x-0.33}$



Question 15 (*+)**

The table below shows a set of bivariate data involving two variables t and v.

t	151	154	157	163	169
v	8800	7800	7400	6500	3100

a) Use the coding equations

$$x=\frac{t-157}{3} \qquad \text{and} \qquad y=\frac{v}{100}$$

to find the value of S_{xx}, S_{yy} and S_{xy}.

b) Show that the product moment correlation coefficient between x and y is approximately -0.958.

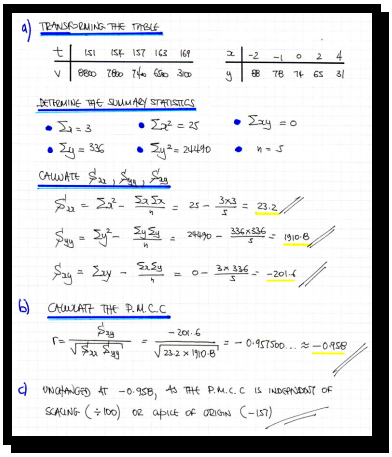
c) State with justification the value of the product moment correlation coefficient between t and v.

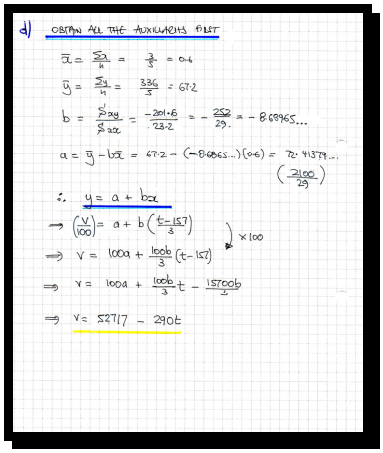
d) Determine the equation of the regression line between t and v, giving the answer in the form

$$v=A+Bt\,,$$

where A and B are constants.

$\boxed{}$, $\boxed{S_{xx}=23.2}$, $\boxed{S_{yy}=1910.8}$, $\boxed{S_{xy}=-201.6}$, $\boxed{r=-0.958}$, $\boxed{v=52717-290t}$





Question 16 **(***+)**

Clinical trials are carried out to determine the effect of a stimulant.

Ten volunteers were given different amounts of the stimulant, X milligrams, and the amount of their nightly sleep, Y hours, were recorded in the following night.

The following summary statistics were obtained.

$$\sum X = 900, \quad \sum Y = 78.4, \quad \sum X^2 = 114\,000, \quad \sum Y^2 = 616.18, \quad \sum XY = 6834$$

The following claims are made.

- Claim 1

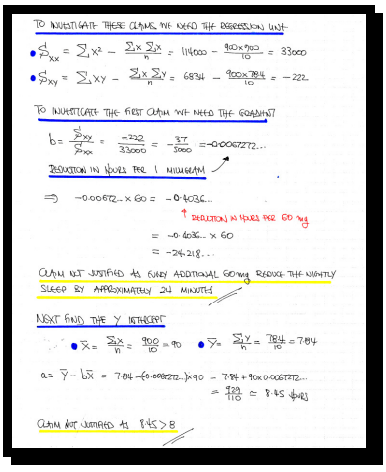
  For every additional 60 milligrams of the stimulant, the nightly sleep typically reduces by 40 minutes.

- Claim 2

  The expected nightly sleep would have been 8 hours if no stimulant was taken.

Comment briefly on these two claims, fully supported by appropriate calculations.

[ ], [claims not justified supported by the regression line equation]



Question 17 (*+)**

Dolphins are thought to communicate with each other by high pitch noises they produce. The frequency, v kHz, of the noise made by a dolphin is recorded at 15 different sea depths, d m. These data are summarized below.

$$\sum d = 385.5\,,\ \sum d^2 = 11543.25\,,\ \sum v = 22.5\,,\ \sum v^2 = 38.25\,,\ \sum dv = 650.25$$

a) State, with a reason, which is the explanatory variable in the above described scenario and state the statistical name of the other variable.

b) Find the value of S_{dd}, S_{vv} and S_{dv} for this data.

c) Calculate the product moment correlation coefficient between d and v.

d) Interpret the value of the product moment correlation coefficient in the context of this question.

e) Give a reason to support the fitting of a regression line of the form

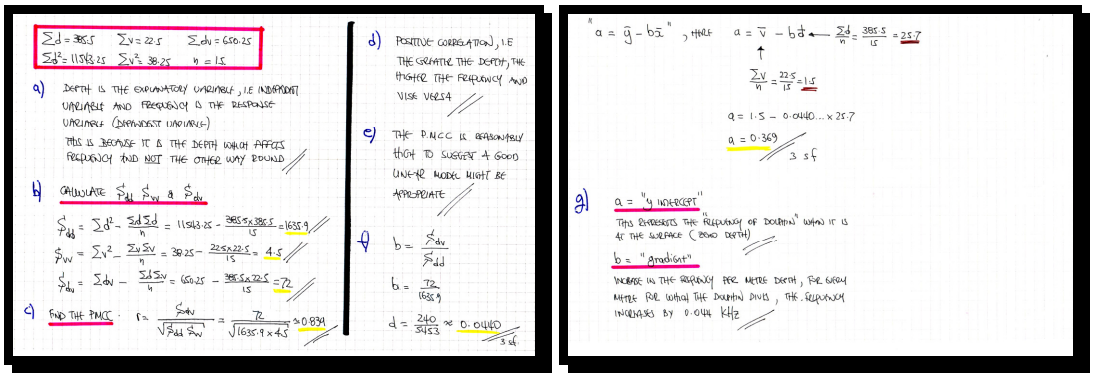
$$v = a + bd\,,$$

where a and b are constants.

f) Determine the value of a and b, correct to three significant figures.

g) Interpret in the context of this question the physical meaning of a and b.

$\boxed{}$, $\boxed{S_{dd} = 1635.9}$, $\boxed{S_{vv} = 4.5}$, $\boxed{S_{dv} = 72}$, $\boxed{r \approx 0.839}$, $\boxed{a \approx 0.369}$, $\boxed{b \approx 0.0440}$



Question 18 (****)

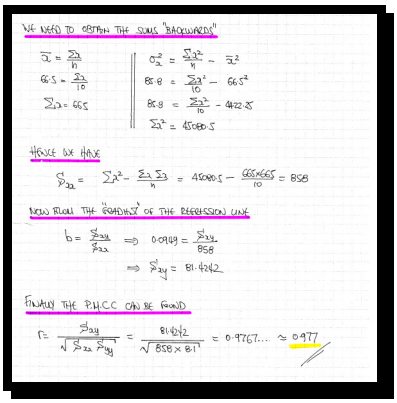
The mean and variance of 10 independent observations of a random variable x, are 66.5 and 85.8, respectively.

Based on of a random sample of 10 independent observations of another variable y, the regression line of y on x is

$$y = 0.0949x - 0.0130 .$$

Determine the product moment correlation coefficient between x and y. assuming further that $S_{yy} = 8.1$.

$\boxed{}$, $\boxed{r = 0.977}$



Question 19 (****)

A gym opened on the first day of January of a given year.

The months of that year were numbered as 1, 2, 3, … , 12.

The number of **new** members, N, at the end of each month m, was recorded for those 12 months.

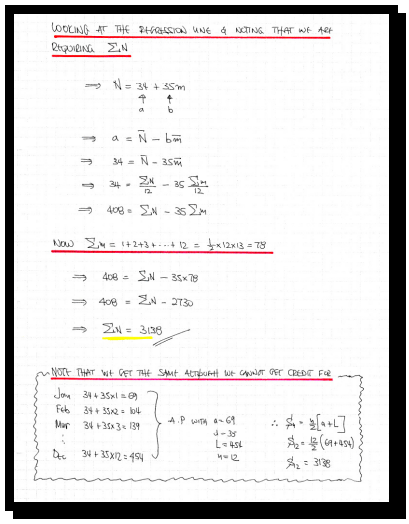
The regression line of N on m was found to be

$$N = 34 + 35m .$$

Use the regression line to find the total number of members which joined that gym during that year.

No credit will be given for adding 69, 104, 139, … , 419, 454.

$\boxed{}$, $\boxed{3138}$



Question 20 **(****+)**

Some summary statistics for a set of bivariate data, based two variables x and y, are given below.

$$n=10,\quad \overline{x}=15,\quad \overline{y}=48,\quad \sigma_x^2=186,\quad \sigma_y^2=172,\quad \sum xy=8850.$$

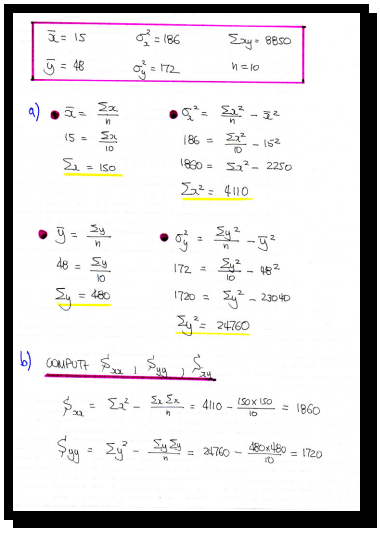
a) Find the value of each of the following sums.

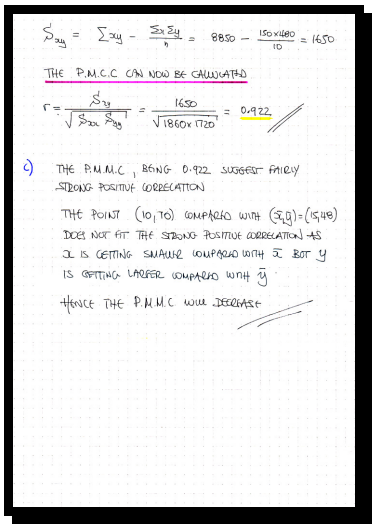
$$\sum x,\quad \sum y,\quad \sum x^2,\quad \sum y^2.$$

b) Calculate the product moment correlation coefficient between x and y.

c) Describe briefly the effect on the product moment correlation coefficient if another piece of data, $x=10$ with $y=70$, is added to the other 10 bivariate observations.

$\boxed{}$, $\boxed{\sum x=150}$, $\boxed{\sum y=480}$, $\boxed{\sum x^2=4110}$, $\boxed{\sum y^2=24760}$, $\boxed{r\approx 0.922}$





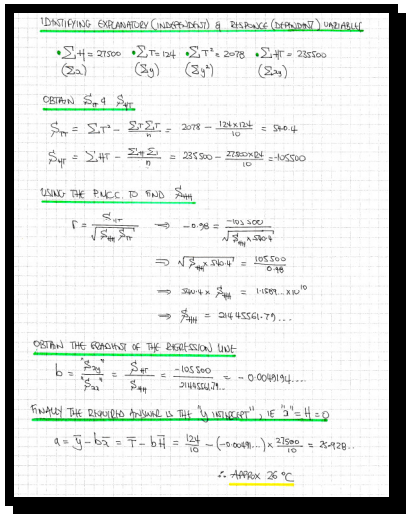
Question 21 (**+)**

On a certain mountain climb, a scientist recorded the temperature, T °C, at ten different heights, H m above sea level, and some of his results are summarized below.

$$\sum T = 124, \quad \sum T^2 = 2078, \quad \sum H = 27\,500, \quad \sum HT = 235\,500$$

If the product moment correlation coefficient for this data is -0.98, determine an estimate for the temperature at sea level on the day of the climb.

$\boxed{}$, $\boxed{\approx 25.9 \text{ °C}}$



Question 22 (**+)**

The number of letters x in people's first names and number of letters y in people's surnames is researched.

The summary data of the number of letters in the first names and the surnames of a random sample of 20 individuals is shown below.

$$\sum x = 125\,, \quad \sum x^2 = 796\,, \quad \sum y = 140\,, \quad \sum y^2 = 1032\,, \quad \sum xy = 882$$

a) Calculate the product moment correlation coefficient between x and y.

The name "Richard Edwards" is added to the sample, making the total number of people in the sample, 21.

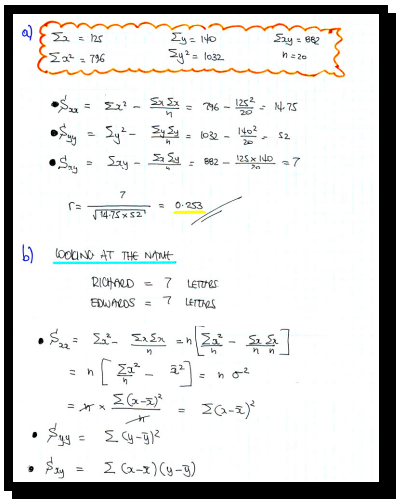
b) Without a direct recalculation , ...

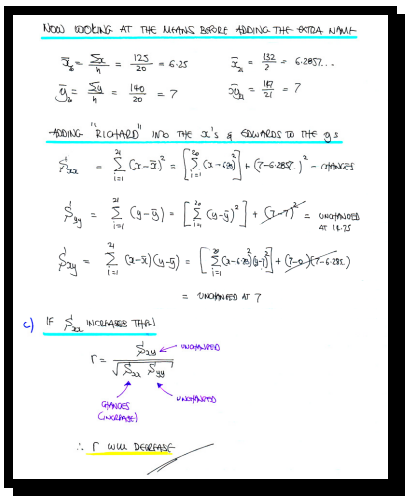
i. ... show that S_{xx} of the 21 first names is likely to have a different value to the original value of S_{xx} of the original 20 first names.

ii. ... determine the effect of adding "Richard Edwards" to S_{yy} and S_{xy}.

c) Given further that adding "Richard Edwards" increases the value of S_{xx} explain with justification whether the product moment correlation coefficient between x and y, increases or decreases.

$\boxed{}$, $\boxed{r \approx 0.253}$





Question 23 (**+)**

Two variables, x and y, have the following regression equations, based on 5 observations.

$$y \text{ on } x : y = 18.5 + 0.1x$$

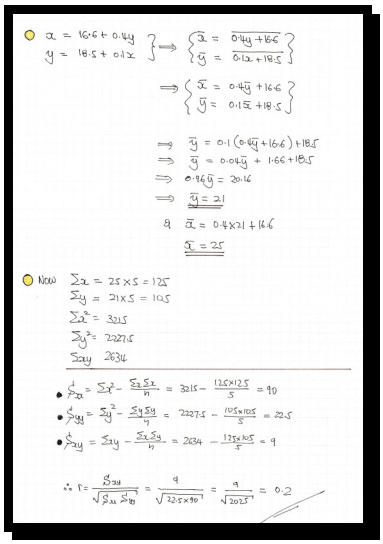
$$x \text{ on } y : x = 16.6 + 0.4y$$

The following summary statistics are also given.

$$\sum x^2 = 3215, \quad \sum y^2 = 2227.5, \quad \sum xy = 2634$$

Show that the product moment correlation coefficient between x and y is 0.2.

proof



SPEARMAN'S RANK

Question 1 (**)

Nine gymnasts performed in a gymnastics competition.

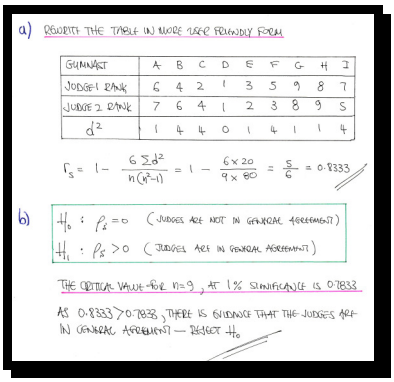
Their names were Arnold (A), Brian (B), Christian (C), Damon (D), Eli (E), Fabian (F), Gordon (G), Harry (H) and Ian (I).

Rank	1	2	3	4	5	6	7	8	9
Judge 1	D	C	E	B	F	A	I	H	G
Judge 2	D	E	F	C	I	B	A	G	H

a) Calculate Spearman's rank correlation coefficient for this data.

b) Test whether or not the judges are generally in agreement, at the 1% level of significance, stating your hypotheses clearly.

$\boxed{}$, $\boxed{r_s = \frac{5}{6} \approx 0.833}$, $\boxed{\text{evidence of agreement, } 0.8333 > 0.7833}$



Question 2 (**)

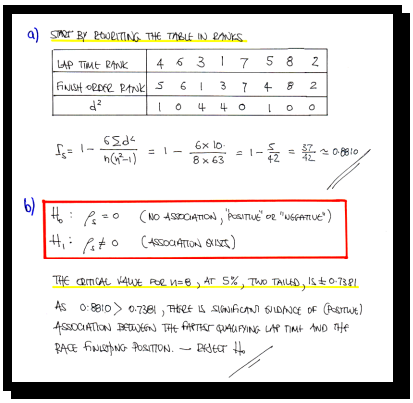
The data in the table below shows the time, in seconds, for the fastest qualifying lap for 8 different Formula One racing drivers, and their finishing order in the actual race.

Fastest Qualifying Lap	49.12	49.34	49.07	48.55	49.40	49.27	49.77	48.87
Finishing Position	5	6	1	3	7	4	8	2

a) Calculate Spearman's rank correlation coefficient for this data.

b) Test whether or not there is any association between the fastest qualifying lap time and the finishing position for Formula One racing drivers, at the 5% level of significance, stating your hypotheses clearly.

$\boxed{}$, $\boxed{r_s = \frac{37}{42} \approx 0.8810}$, $\boxed{\text{evidence of association, } 0.8810 > 0.7381}$



a) START BY RECREATING THE TABLE IN RANKS

LAP TIME RANK	4	6	3	1	7	5	8	2
FINISH ORDER RANK	5	6	1	3	7	4	8	2
d^2	1	0	4	4	0	1	0	0

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 10}{8\times 63} = 1 - \frac{5}{42} = \frac{37}{42} \simeq 0.8810$

b) $H_0: \rho_s = 0$ (NO ASSOCIATION, "POSITIVE" OR "NEGATIVE")
$H_1: \rho_s \neq 0$ (ASSOCIATION EXISTS)

THE CRITICAL VALUE FOR $n=8$, AT 5%, TWO TAILED, IS ± 0.7381

AS $0.8810 > 0.7381$, THERE IS SIGNIFICANT EVIDENCE OF (POSITIVE) ASSOCIATION BETWEEN THE FASTEST QUALIFYING LAP TIME AND THE RACE FINISHING POSITION. $\rightarrow$ REJECT H_0

Question 3 (**)

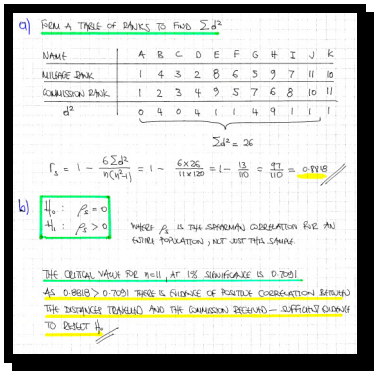
The table below shows the mileages travelled by eleven salesmen and the commission they got paid during a given month.

Name	Monthly mileage	Monthly commission
Alan	734	£800
Brian	650	£660
Christian	668	£620
Dominic	709	£610
Ethan	437	£450
Finlay	551	£560
Graham	580	£510
Hamish	387	£520
Ian	450	£460
James	298	£430
Kevin	325	£390

c) Calculate Spearman's rank correlation coefficient for this data.

d) Test whether or not there is evidence of positive correlation between the mileages travelled and the amount of commission received, at the 1% level of significance, stating your hypotheses clearly.

$\boxed{}$, $\boxed{r_s = \frac{97}{110} \approx 0.8818}$, $\boxed{\text{positive correlation, } 0.8818 > 0.7091}$



Question 4 (**)

The actual ages, in complete years, of seven cats is shown below.

Cat Name	Riri	Loulou	Ginge	Puss	Ollie	Rex	Mog
Age in years	3	4	18	21	5	11	9

These seven cats were seen by a vet, during a day's surgery, and the vet was asked to order them according to their age by examination only.

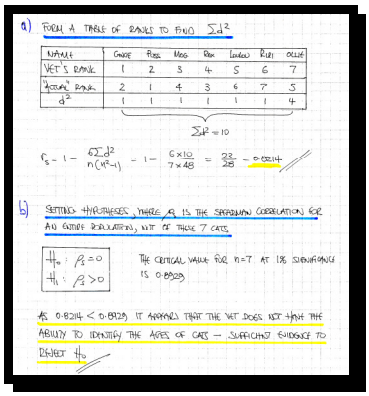
He ordered the cats' ages, older first, as follows.

Ginge, Puss, Mog, Rex, Loulou, Riri, Ollie.

c) Calculate Spearman's rank correlation coefficient between the actual age of the cats and the vet's order.

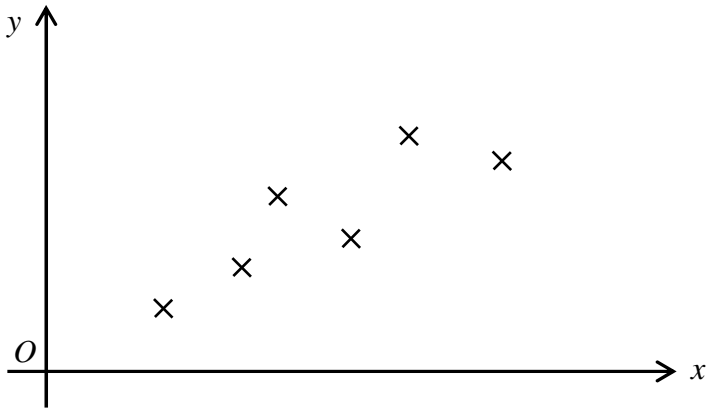
d) Test whether or not the vet has the ability to identify the age of cats, at the 1% level of significance, stating your hypotheses clearly.

$\boxed{}$, $\boxed{r_s = \frac{23}{28} \approx 0.8214}$, $\boxed{\text{no evidence of association, } 0.8214 < 0.8929}$



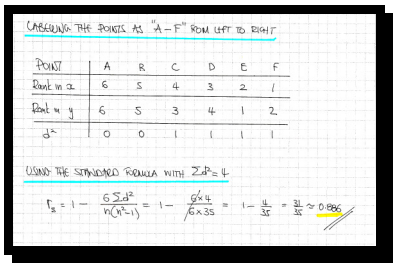
Question 5 (+)**

Six ordered pairs (x, y), of bivariate data, are shown in the following set of axes.



Determine the Spearman's rank correlation coefficient for this data.

$\square$, $r_s = \frac{31}{35} \approx 0.886$



Question 6 **(**+)**

The table below shows, for a group of students in a recent mock exam, the number of marks lost, y, and the corresponding number of papers, x, they practiced leading up to that exam.

Student	A	B	C	D	E	F	G	H	I	J
Number of Papers (x)	17	39	24	26	11	22	25	10	8	6
Number of Marks Lost (y)	12	5	11	14	10	9	8	15	19	17

a) Find the value of S_{xx}, S_{yy} and S_{xy}, and hence determine the value of the product moment correlation coefficient between x and y.

b) Comment briefly on the result of part **(a)**.

c) Obtain the Spearman's rank correlation coefficient between x and y.

d) Test, at the 1% level of significance, whether there is evidence of negative association between the ranks of x and y.

$\boxed{}$, $\boxed{S_{xx}=957.6}$, $\boxed{S_{yy}=166}$, $\boxed{S_{xy}=-317}$, $\boxed{r=-0.795}$, $\boxed{r_s=-0.745}$

