MONASH University

Information Technology

FIT1006
Business Information Analysis

Lecture 7
Correlation

Topics covered:

- Bivariate data.
- The linear model.
- Calculating q and r by hand.
- Calculating r using Excel and SYSTAT.
- Interpreting q and r.
- Visual estimation of q and r.

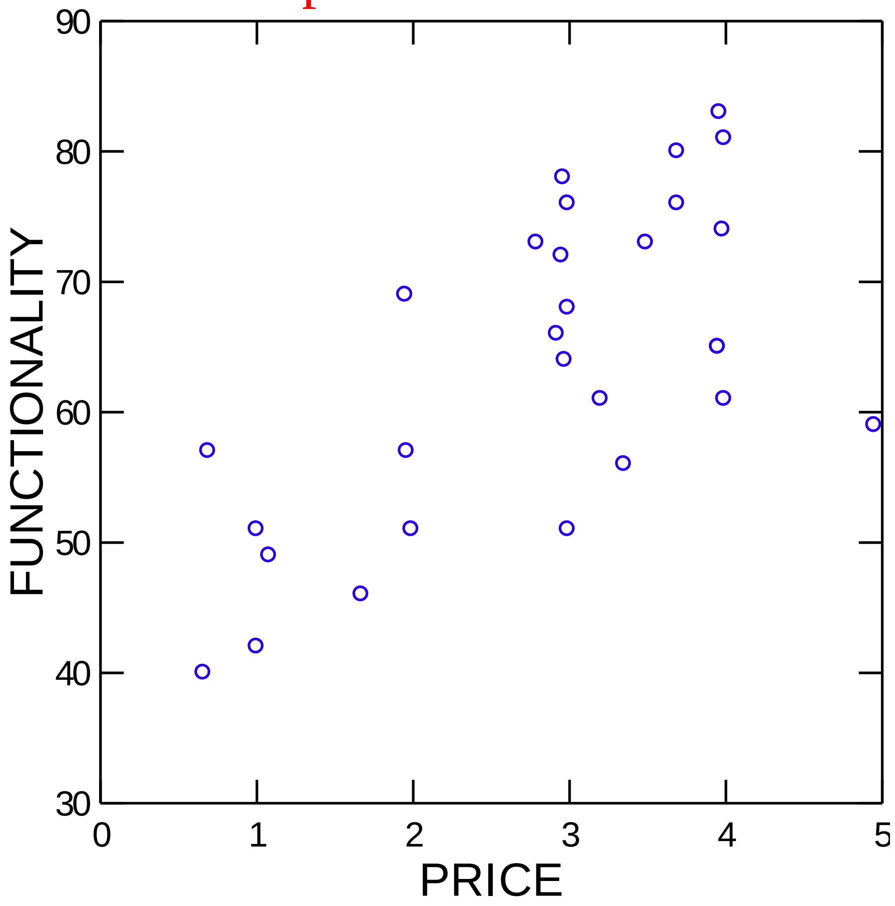
Motivating Question

- In 1998, *Choice* magazine tested 1500 toothbrushes.
- A summary of price and functionality score is on the right.
- Is the functionality of the toothbrush related to the price? (Selvanathan 4th Ed p 679)
- Answers later…

Price	Functionality
3.96	83
3.99	81
3.69	80
2.96	78
3.69	76
2.99	76
3.98	74
2.79	73
3.49	73
2.95	72
1.95	69
2.99	68
2.92	66
3.95	65
3.95	65
2.97	64
3.99	61
3.20	61
4.95	59
0.69	57
1.96	57
3.35	56
1.00	51
2.99	51
1.99	51
1.08	49
1.67	46
1.00	42
0.66	40

Motivating Question

Scatterplot

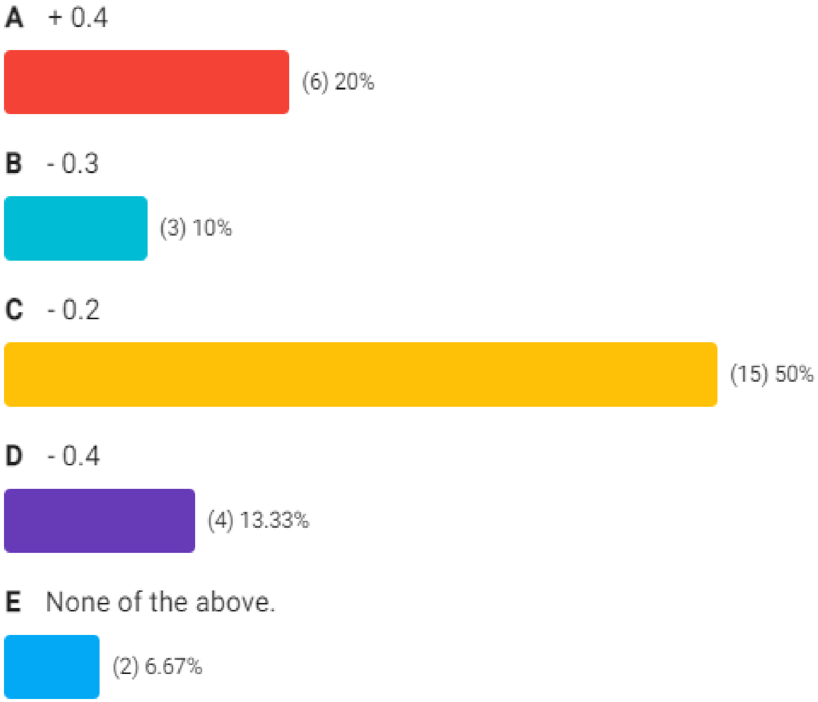


Price	Functionality
3.96	83
3.99	81
3.69	80
2.96	78
3.69	76
2.99	76
3.98	74
2.79	73
3.49	73
2.95	72
1.95	69
2.99	68
2.92	66
3.95	65
3.95	65
2.97	64
3.99	61
3.20	61
4.95	59
0.69	57
1.96	57
3.35	56
1.00	51
2.99	51
1.99	51
1.08	49
1.67	46
1.00	42
0.66	40

https://flux.qa (Feed code: SJ6KGV)

Question 1

From the scatterplot on the RHS below, the *q*-correlation coefficient is:



A + 0.4

(6) 20%

B - 0.3

(3) 10%

C - 0.2

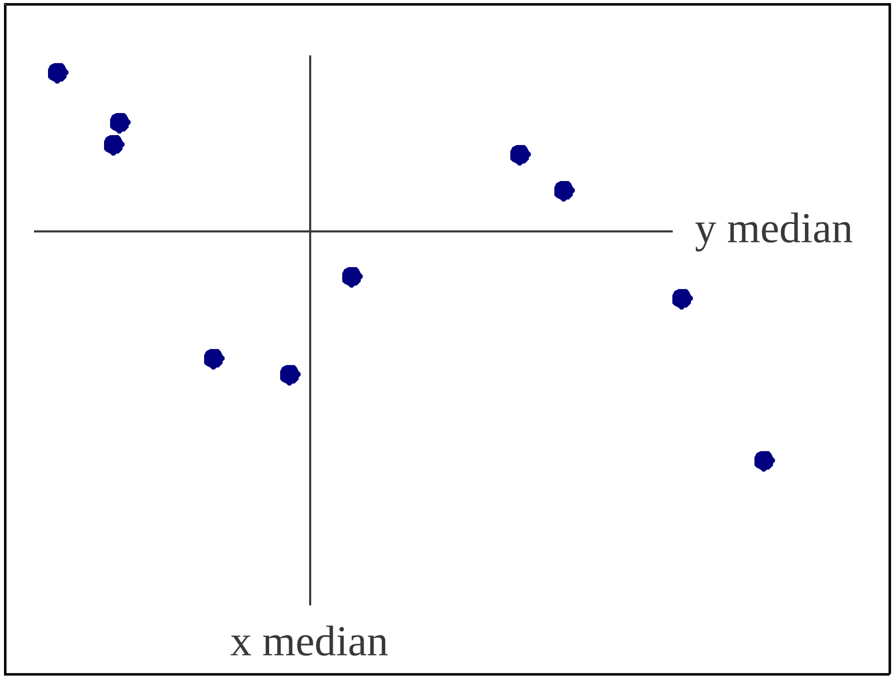
(15) 50%

D - 0.4

(4) 13.33%

E None of the above.

(2) 6.67%



Discussion in groups

A	B
C	D

$$q = \frac{N_B + N_C - \left(N_A + N_D\right)}{N_A + N_B + N_C + N_D}$$

https://flux.qa (Feed code: SJ6KGV)

Question 2

From the scatterplot on the RHS below, the *q*-correlation coefficient is:

A. + 0.4

B. – 0.3

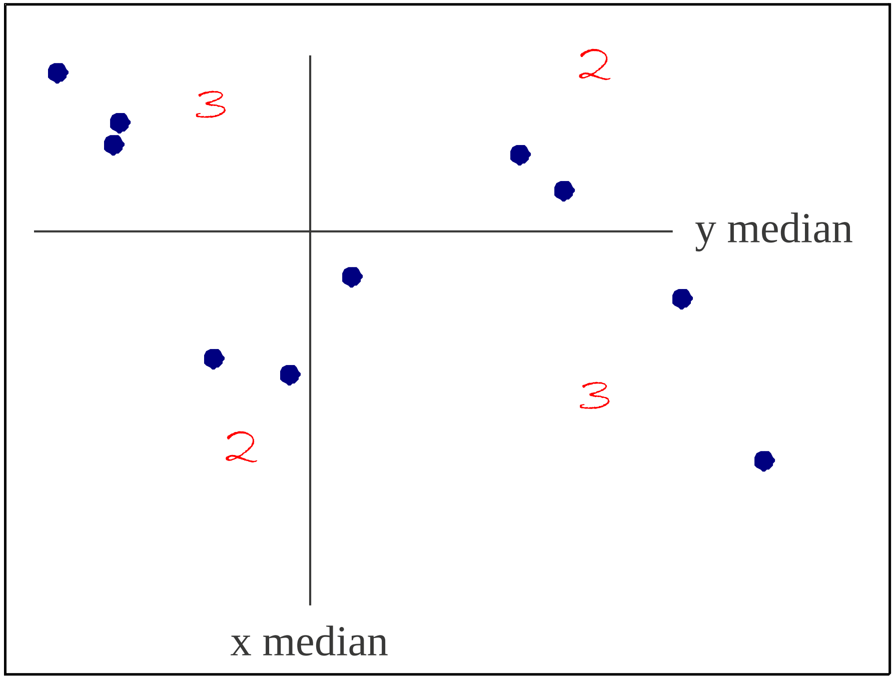
✓ C. – 0.2

D. – 0.4

E. None of the above.

$$q = \frac{(2+2) - (3+3)}{3+2+3+2}$$

$$\rightarrow q = -0.2$$

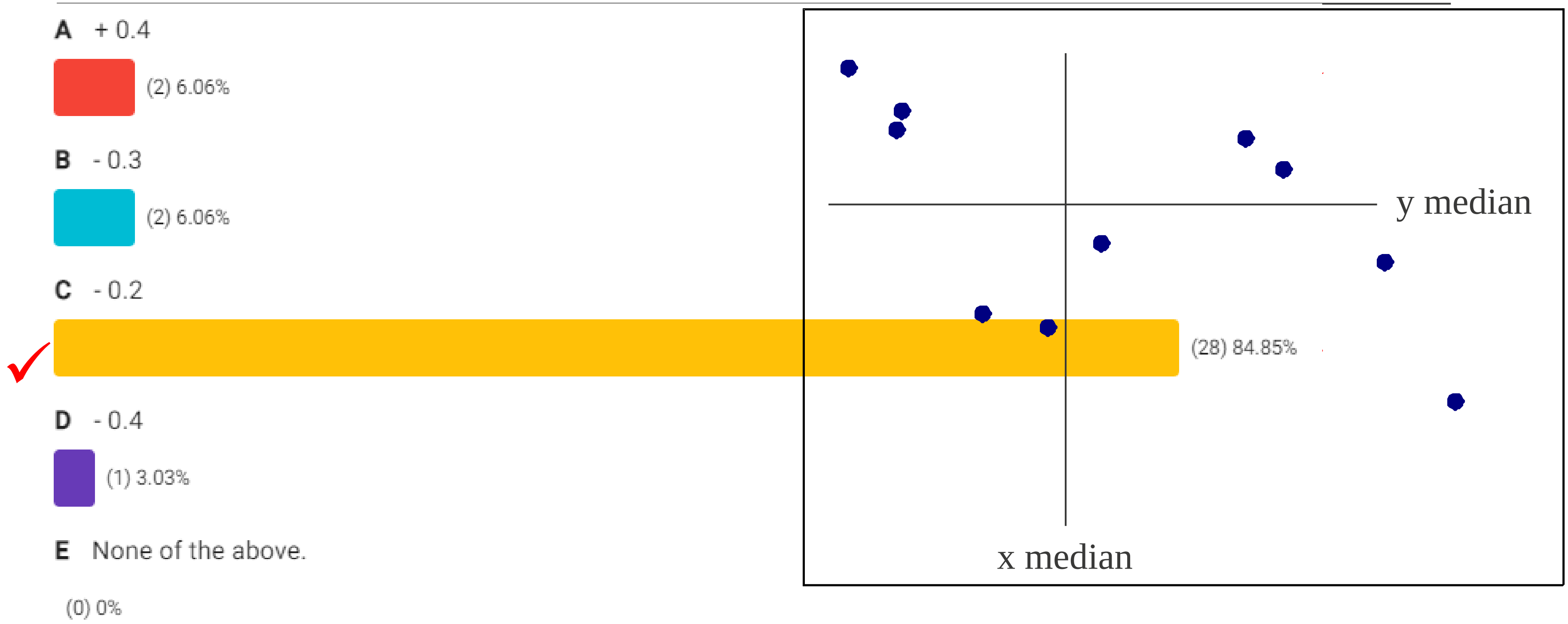


https://flux.qa (Feed code)

Peer-assisted learning – definitely improves results!

Question 2

From the scatterplot on the RHS below, the *q*-correlation coefficient is:



A + 0.4

(2) 6.06%

B - 0.3

(2) 6.06%

✓ C - 0.2

(28) 84.85%

D - 0.4

(1) 3.03%

E None of the above.

(0) 0%

q-Correlation

- To calculate *q*, find the horizontal and vertical medians and divide the data into four quadrants.
- Count the number of observations in each quadrant. Do not count any observations lying on the median lines.
- Calculate the *q*-correlation as follows:

A	B
C	D

$$q = \frac{N_B + N_C - \left(N_A + N_D\right)}{N_A + N_B + N_C + N_D}$$

- Note that *q* is robust to outliers.

https://flux.qa (Feed code: SJ6KGV)

Question 3

From the scatterplot on the RHS below, the *q*-correlation coefficient is:

A. + 0.7

[✓ B. + 1.0]

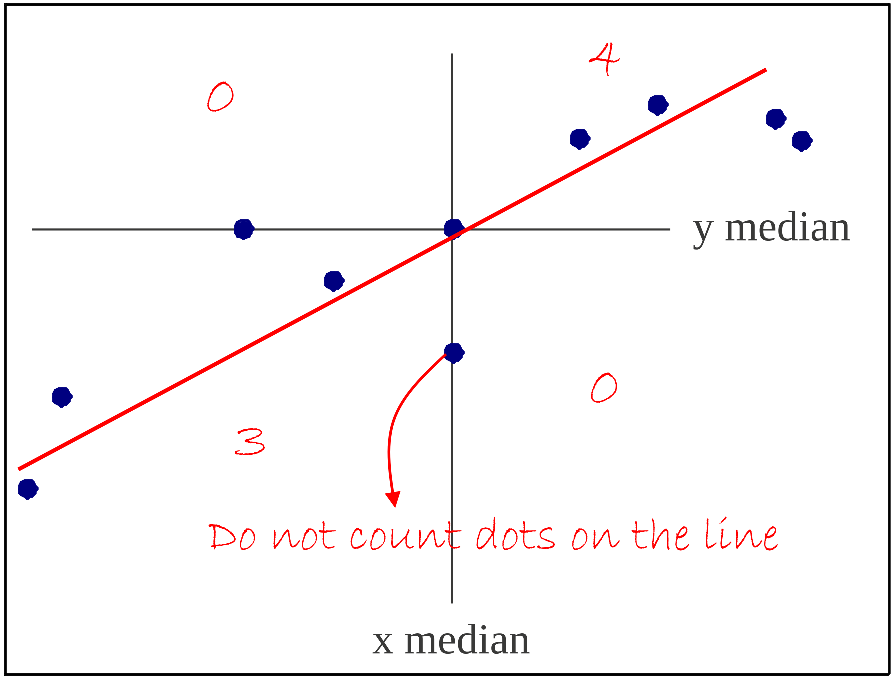
C. – 0.1

D. + 0.1

E. None of the above.

$$q = \frac{(4+3) - (0+0)}{0+4+3+0}$$

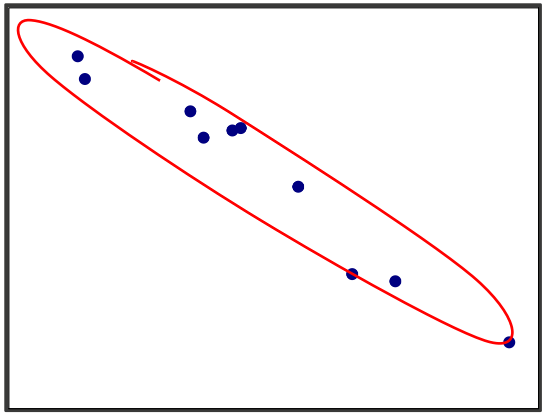
$$\rightarrow q = 1$$



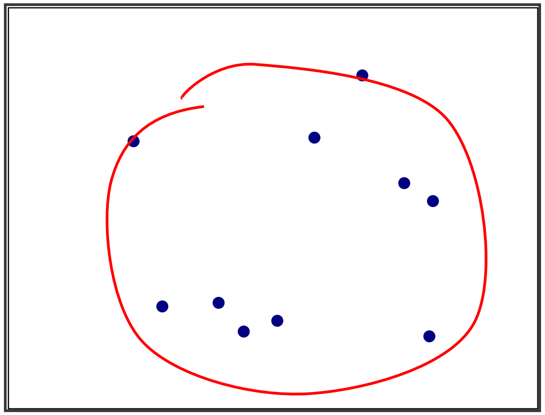
Question 4 (Feed code: SJ6KGV)

Which plot has a *q*-correlation closest to 0?

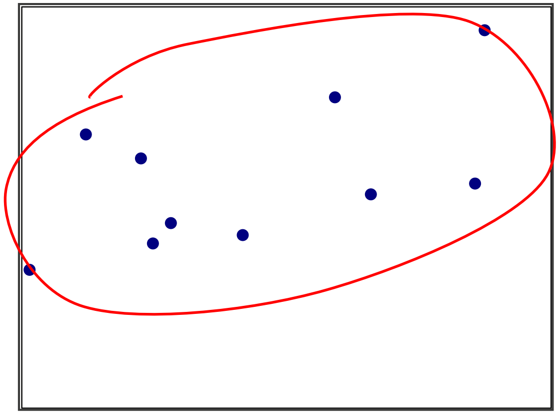
A.



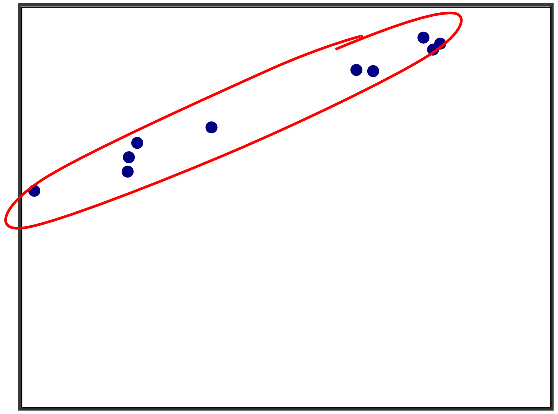
✓ B.



✓ C.



D.

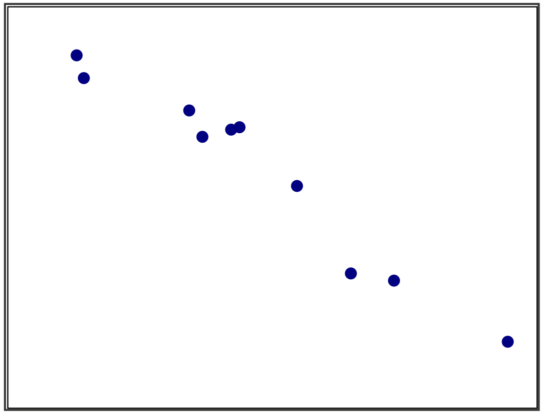


MONASH University

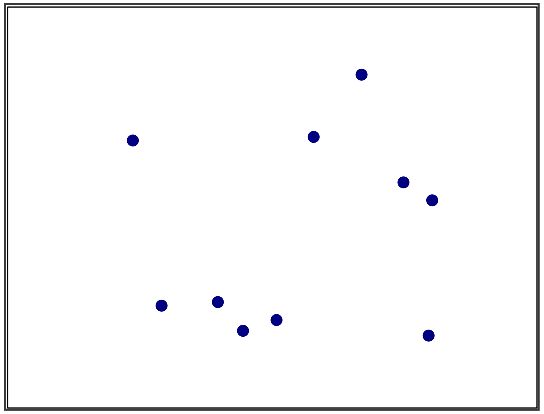
Question 5 (Feed code: SJ6KGV)

Which plot has a *q*-correlation closest to – 1?

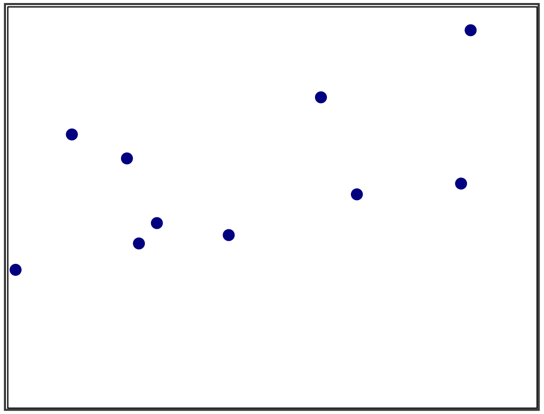
✓ A.



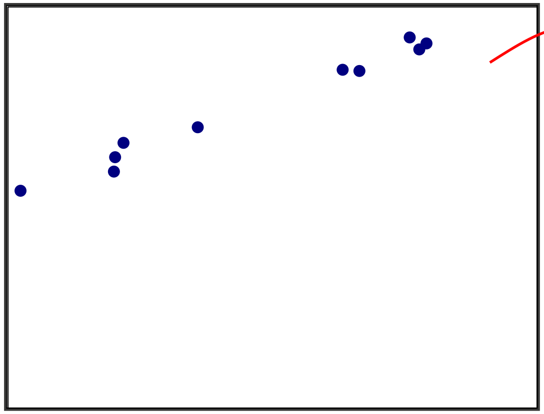
B.



C.



D.

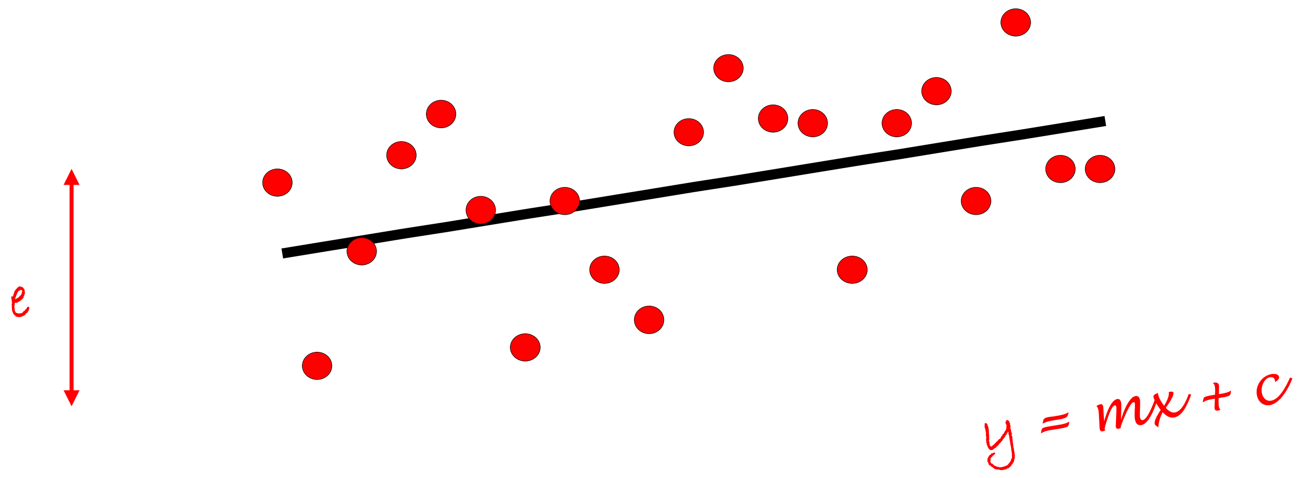


+1

Linear relationship

- When we determine the degree of correlation between variables we are assuming that the variables have a linear relationship.
- For two variables *x*, and *y*, we say that $y = ax + b + e$, where *e* are random, normally distributed errors.

error term



Pearson's *r*

- Pearson's *r* is the most commonly used measure of correlation. S_{xy} is the covariance of *x* and *y*.
- You should be able to calculate *r* if given the sum terms: Σx, Σy, Σx^2, Σy^2, Σxy, and *n*.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}\sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}}$$

std dev x

Calculating *r*

- Pearson’s *r* is built into Excel, SYSTAT and probably your calculator.
- In EXCEL use = CORREL(RANGE1, RANGE2) or draw a scatter plot and fit linear model.
- In SYSTAT use the menu:
  - Graph > Plots > Scatterplot
  - Statistics > Correlations > Simple
- For multivariate data use:
  - Graph > Multivariate Displays > Scatterplot Matrix (SPLOM)

Question 6 (Feed code: SJ6KGV)

Pearson's *r is* an appropriate correlation measure for

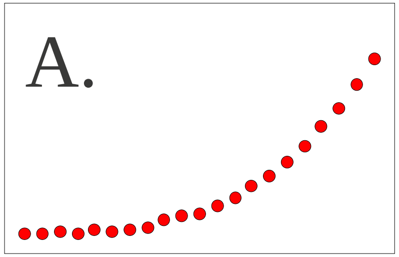
A. A – F.

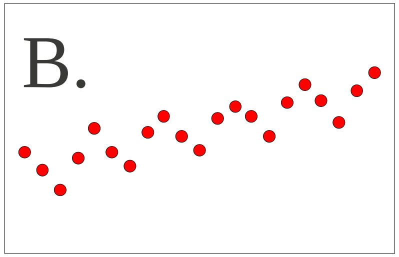
B. B, C, D, F.

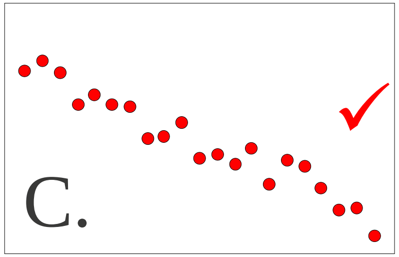
[✓ C. C, D, F.]

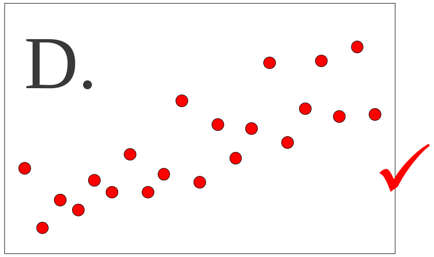
D. C, B, D.

E. C, D.

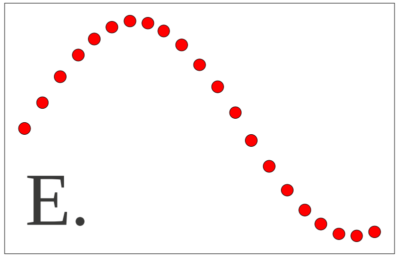


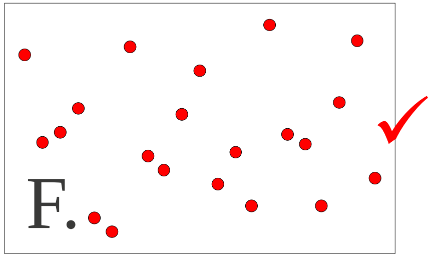






✓



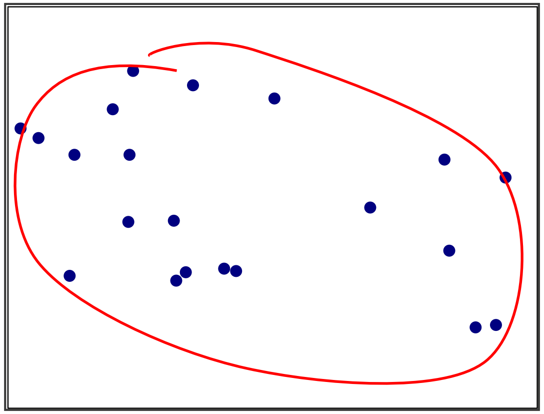


✓

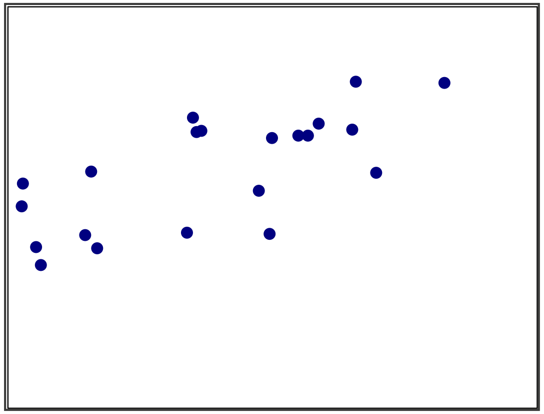
Question 7 (Feed code: SJ6KGV)

For which plot is *r* closest to 0?

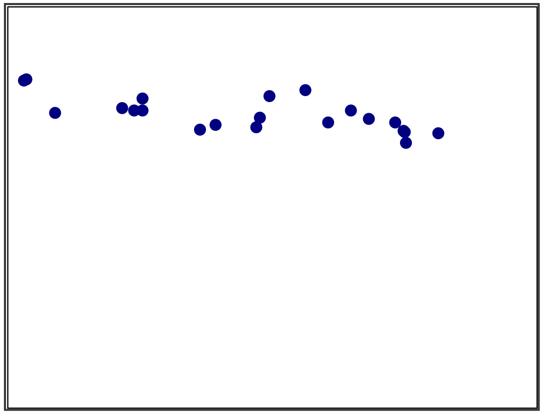
[✓ A.]



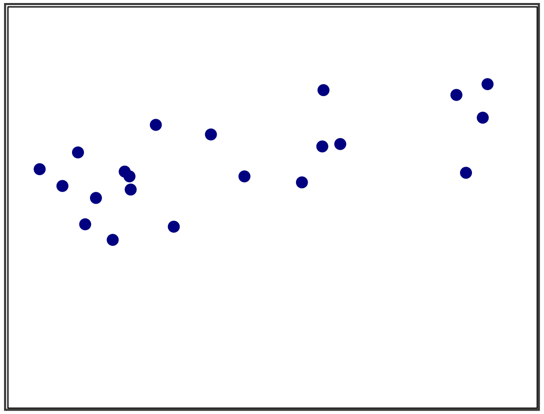
B.



C.



D.



Question 8 (Feed code: SJ6KGV)

If a data point moves as shown. Which of the following is true?

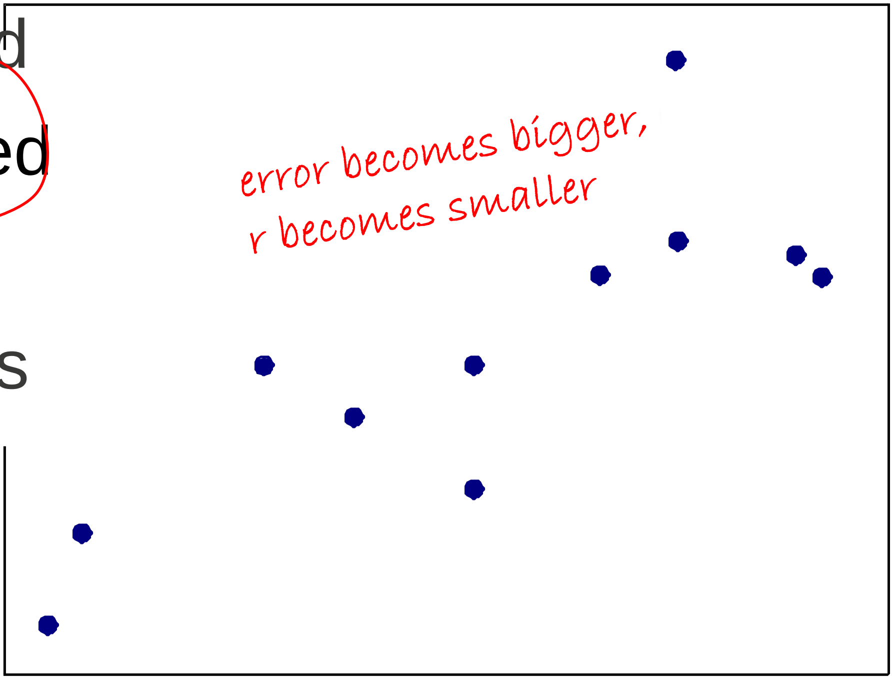
A. *r* increases, *q* unchanged

[✓ B. *r* decreases, *q* unchanged]

C. *r* increases, *q* increases

D. *r* decreases, *q* decreases

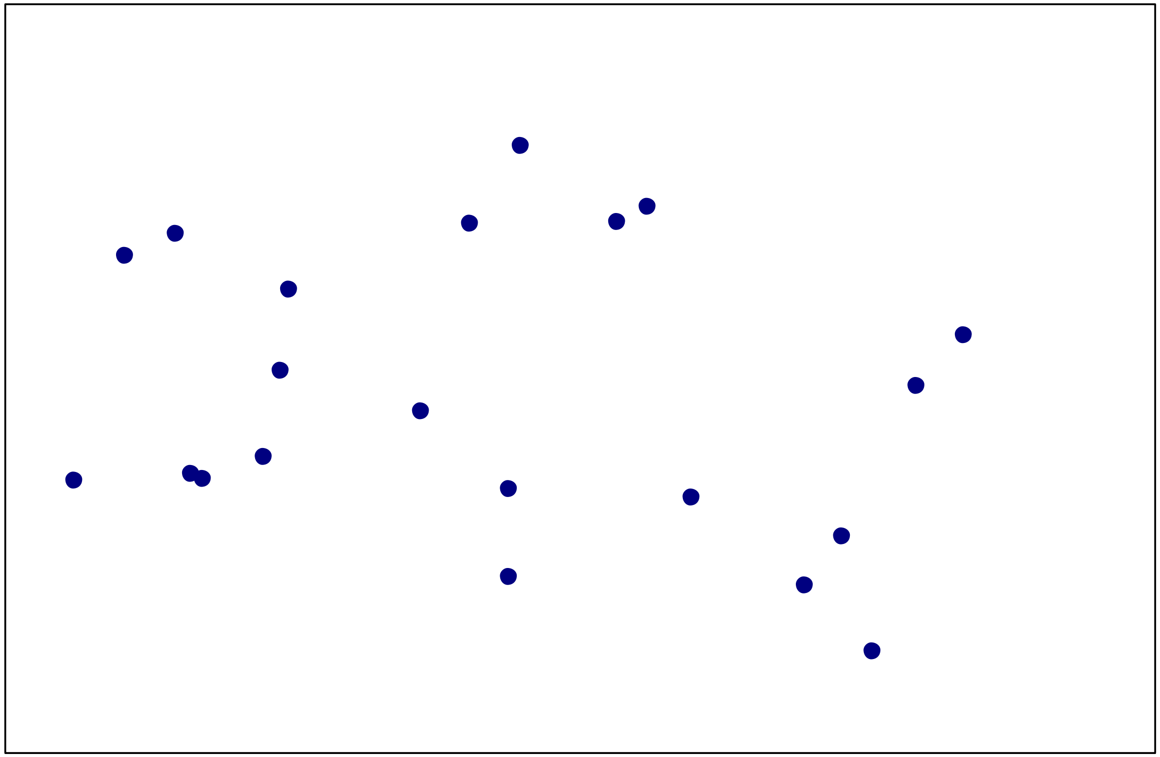
E. None of the above.



error becomes bigger,
r becomes smaller

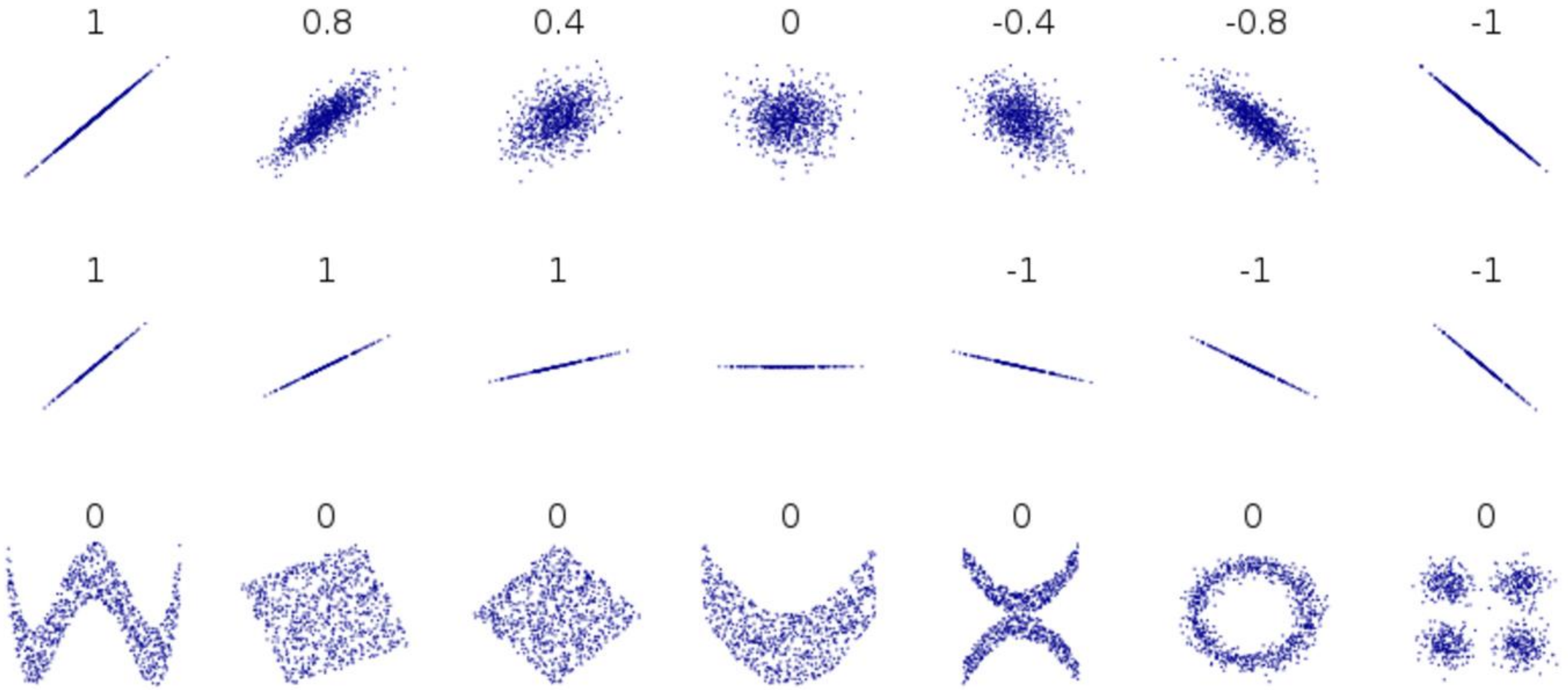
Estimating *r* and *q* by eye

- Practice using the ‘Correlation’ worksheet.



Estimating correlation

- From: https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient



Question 9 (Feed code: SJ6KGV)

For the motivating problem, *r* is closest to:

A. 0.1

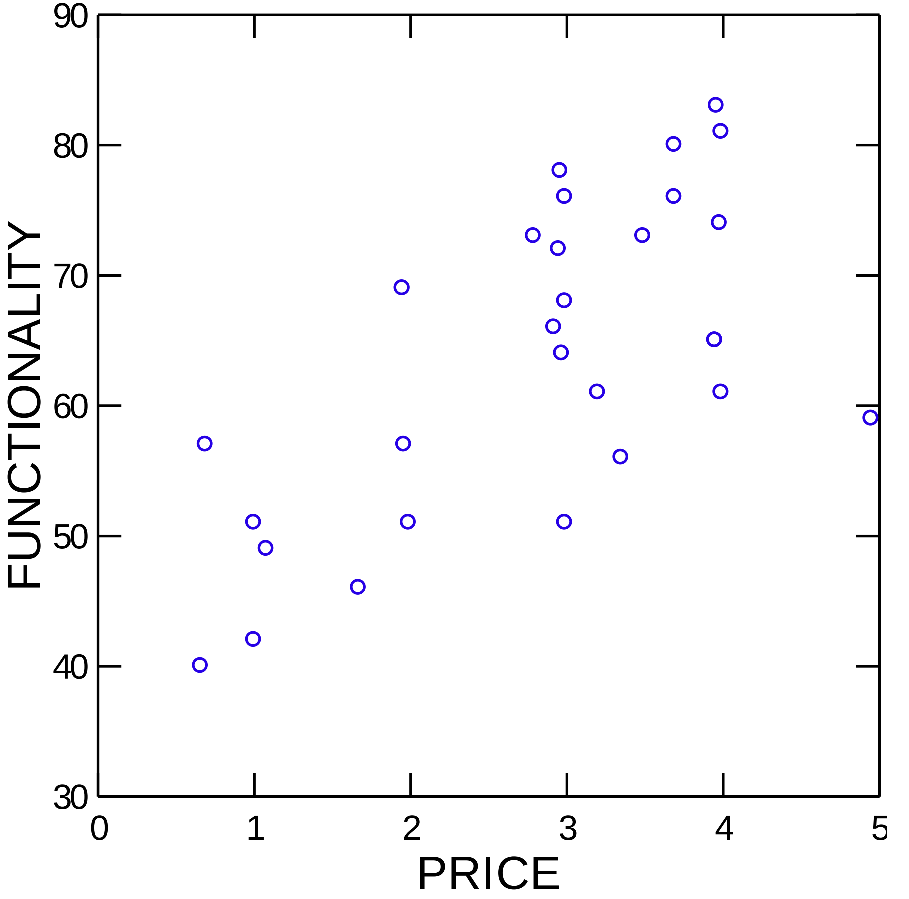
B. 0.3

C. 0.5

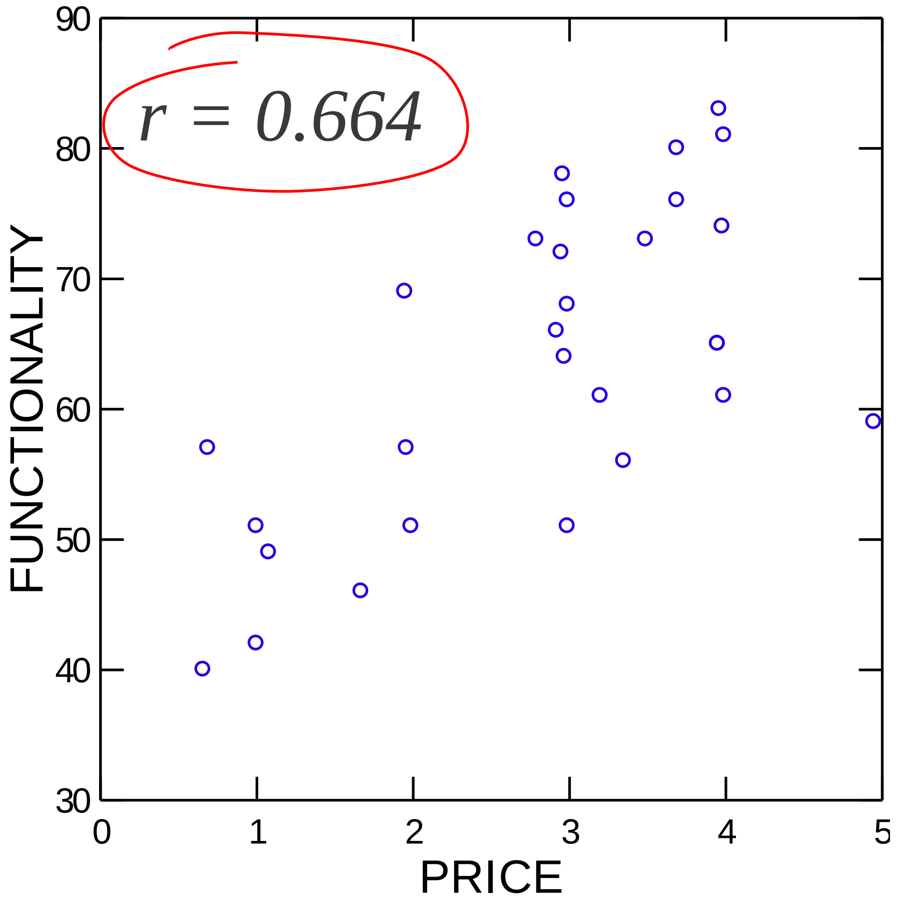
✓ D. 0.7

E. 0.9

See next slide for answer from Systat



Motivating Question



Price	Functionality
3.96	83
3.99	81
3.69	80
2.96	78
3.69	76
2.99	76
3.98	74
2.79	73
3.49	73
2.95	72
1.95	69
2.99	68
2.92	66
3.95	65
3.95	65
2.97	64
3.99	61
3.20	61
4.95	59
0.69	57
1.96	57
3.35	56
1.00	51
2.99	51
1.99	51
1.08	49
1.67	46
1.00	42
0.66	40

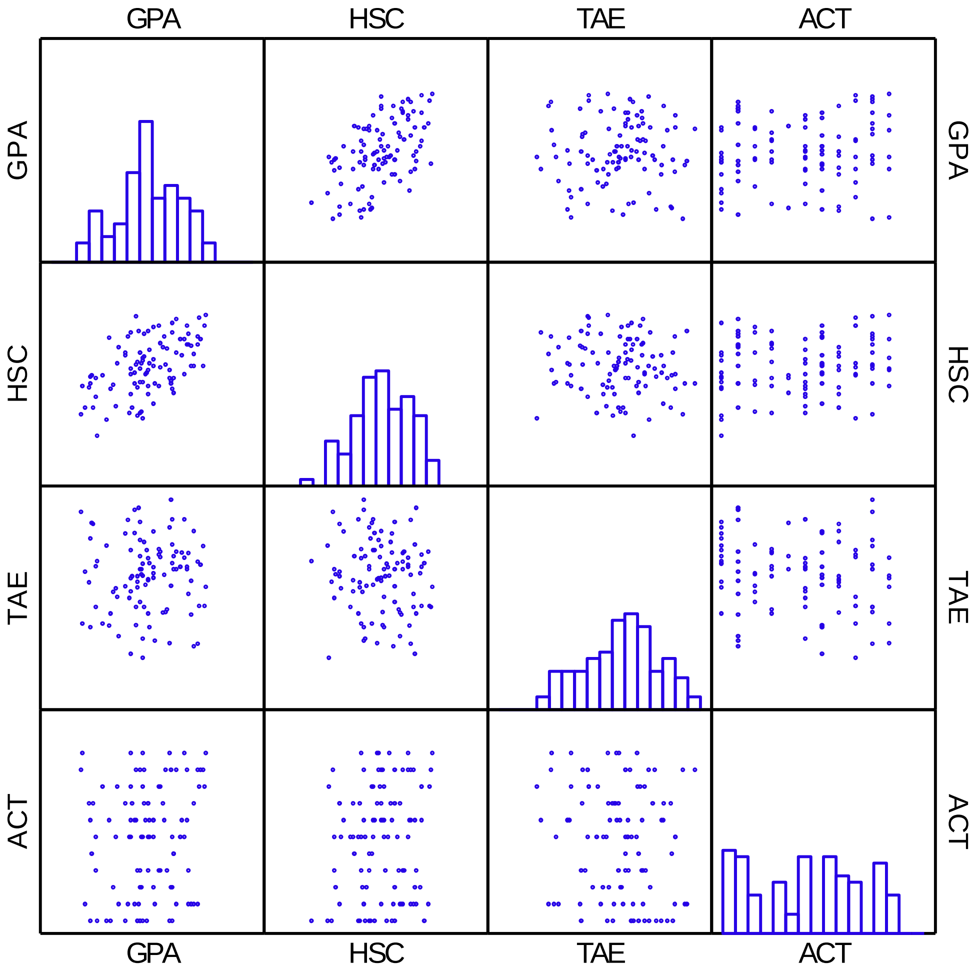
Interpreting correlation

- Some Cautions:
- non-linear relationships will have low correlation.
- Bivariate data are subject to outliers which tend to decrease the value of correlation coefficient.
- Correlation does not imply causation. Two variables may have a strong correlation but are not necessarily directly related. (They may be related by a third party)
- We tend to use correlation comparatively - that is one set of observations have a greater correlation than another set.

Discussion: Multiple Plots

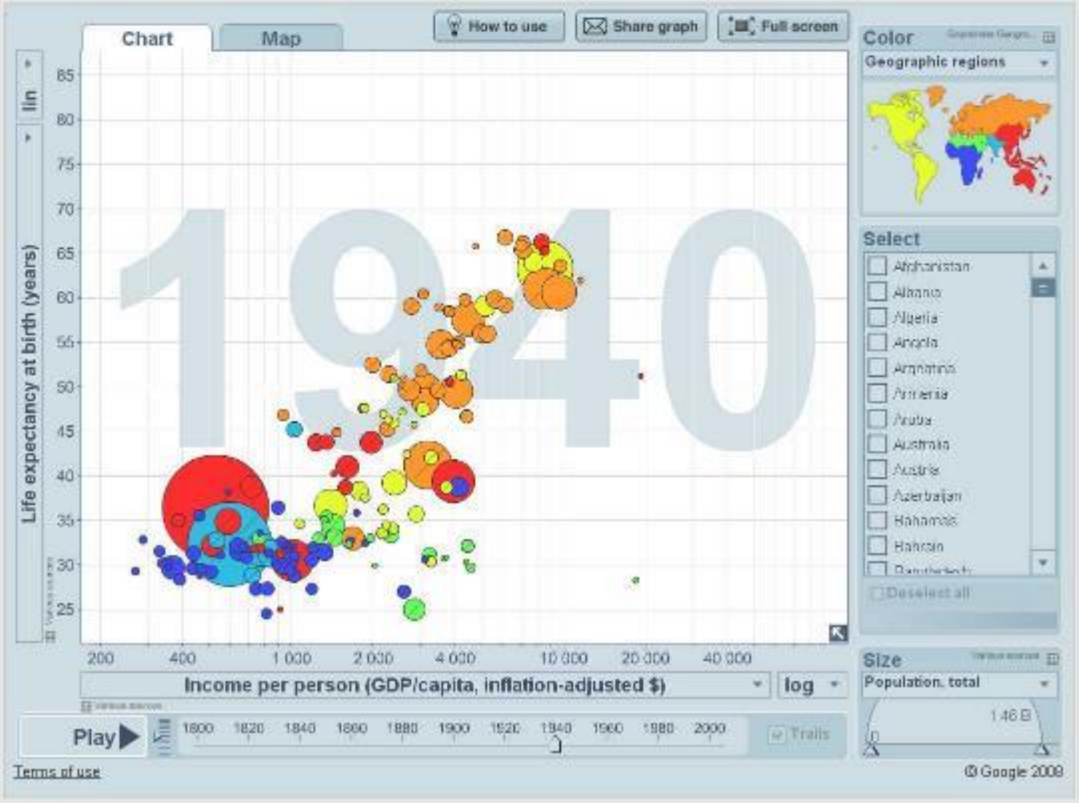
- Data XR 15-19 is admissions data looking success factors based on GPA over first 3 years at university.
- You have:
  - HSC grades
  - TAE (Tertiary admission score)
  - ACT – hour/week on extra curricular activities.
- Which is best predictor of GPA?

Discussion: SPLOM



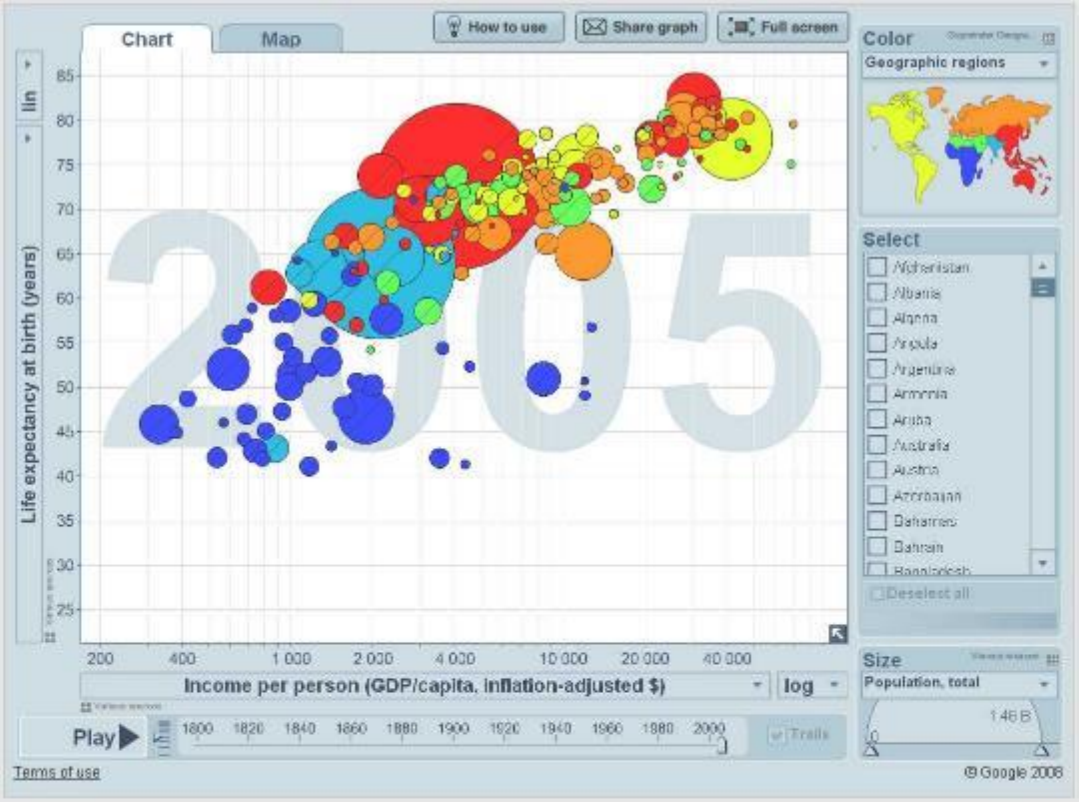
Scatterplots over multiple variables

For enrichment: go to http://www.gapminder.org/



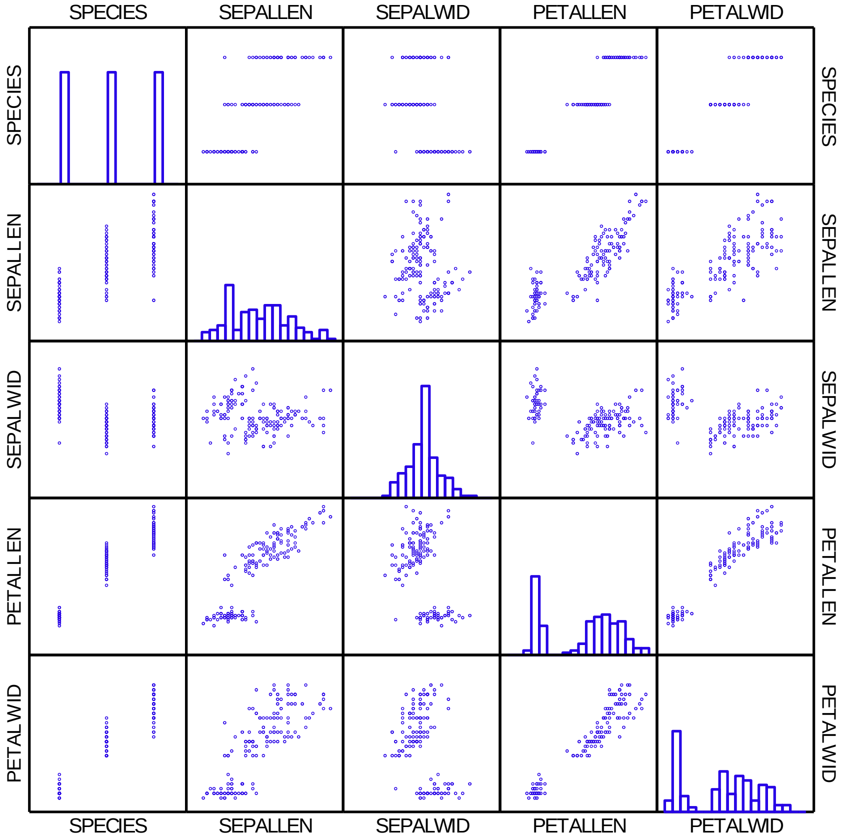
Scatterplots over multiple variables

Multivariate display shows: Income, Life expectancy, Geographic region and Population over time.



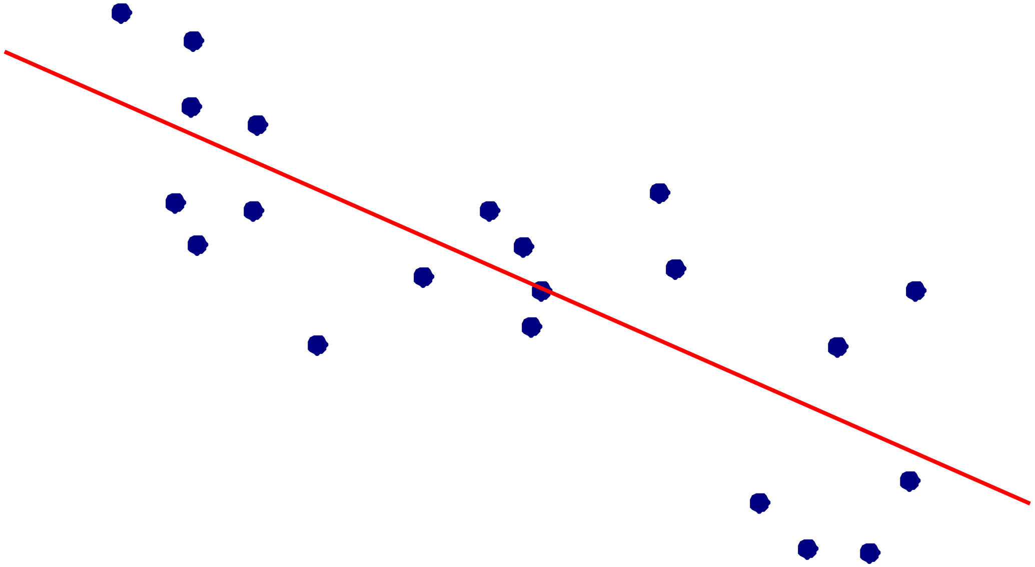
The Iris Data

A famous data set. See Wikipedia. Compares the sepal width & length and petal width & length for 3 species of iris.



Regression

The equation of the trend line is the other piece of important information we get from bivariate data. This is covered next lecture.



Reading/Questions (Selvanathan)

- Reading:
  - 7th Ed Sections 4.3, 5.5.
- Questions:
  - 7th Ed Questions 4.37, 4.38, 4.43, 4.44, 5.77, 5.81, 5.84, 5.85.