

# Data Analysis and Visualisation

Correlation Coefficients

# Introduction

- Today we'll look at
  - Correlation Coefficient
  - Pearson Correlation Coefficient
  - Interpreting Correlation Coefficients
  - Spearman Rank Correlation
- Until now we've been looking at data that can be described using a single variable
- Today we will start to look at the relationship between two variables

# Correlation Coefficients

- The correlation coefficient  $r_{xy}$   
shows the degree of association  
values from -1.0 to +1.0

When  $r_{xy} = 0$  no relationship exists between  $x$  and  $y$

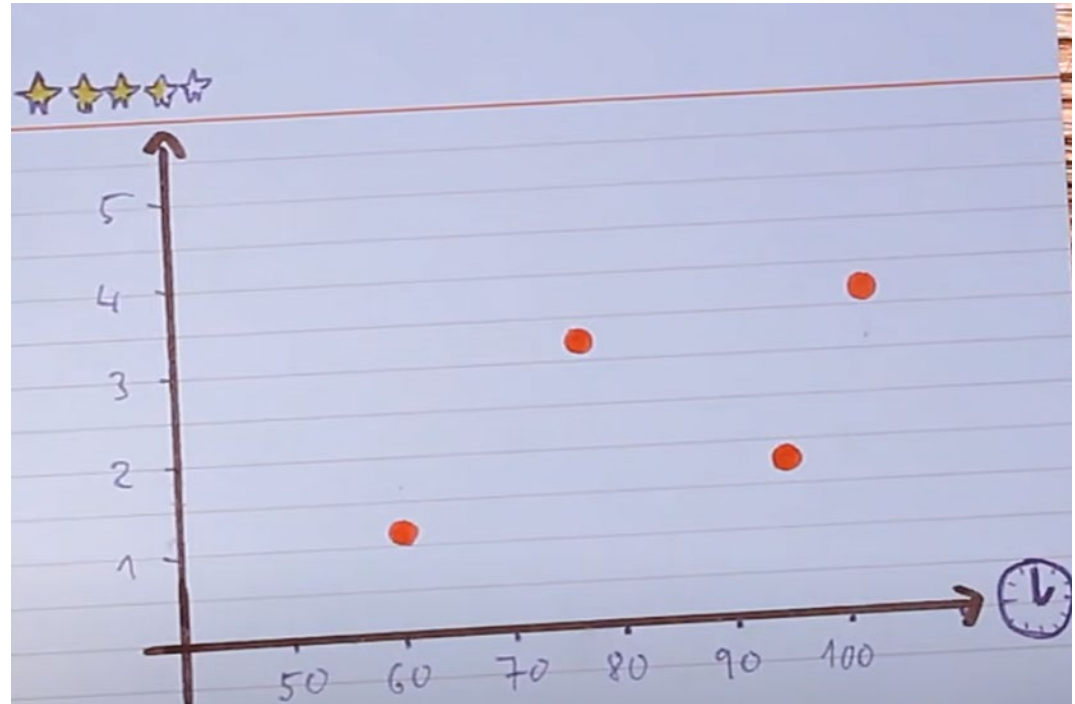
Strength of relationship  $\propto$  absolute value of  $r_{xy}$

# Correlation Coefficients

Consider the example:

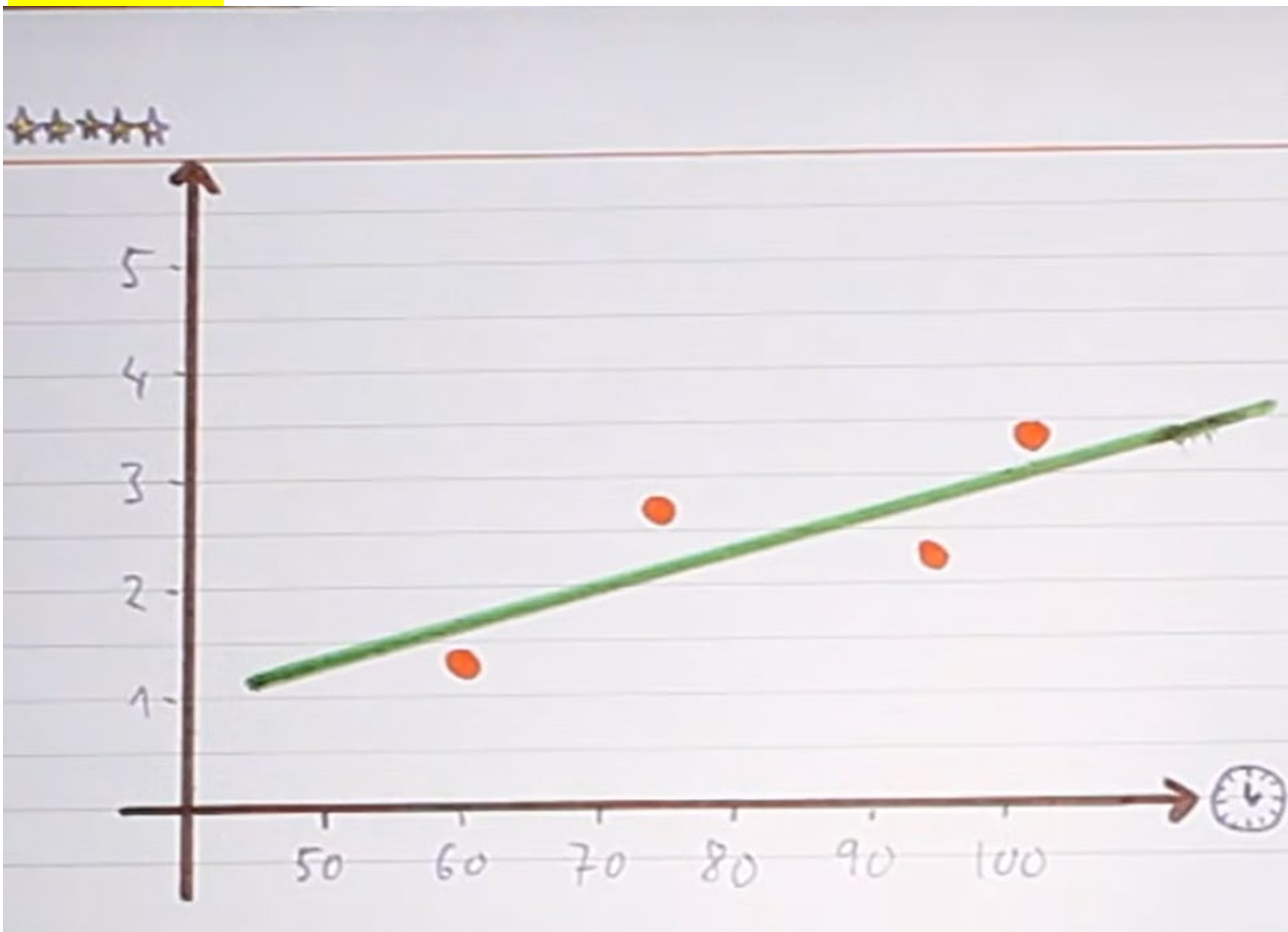
Movies length on the x-axis,

Rating for the movie between 0 and 5 stars on y axis



# Correlation Coefficient Steps

## 1. Fit Line

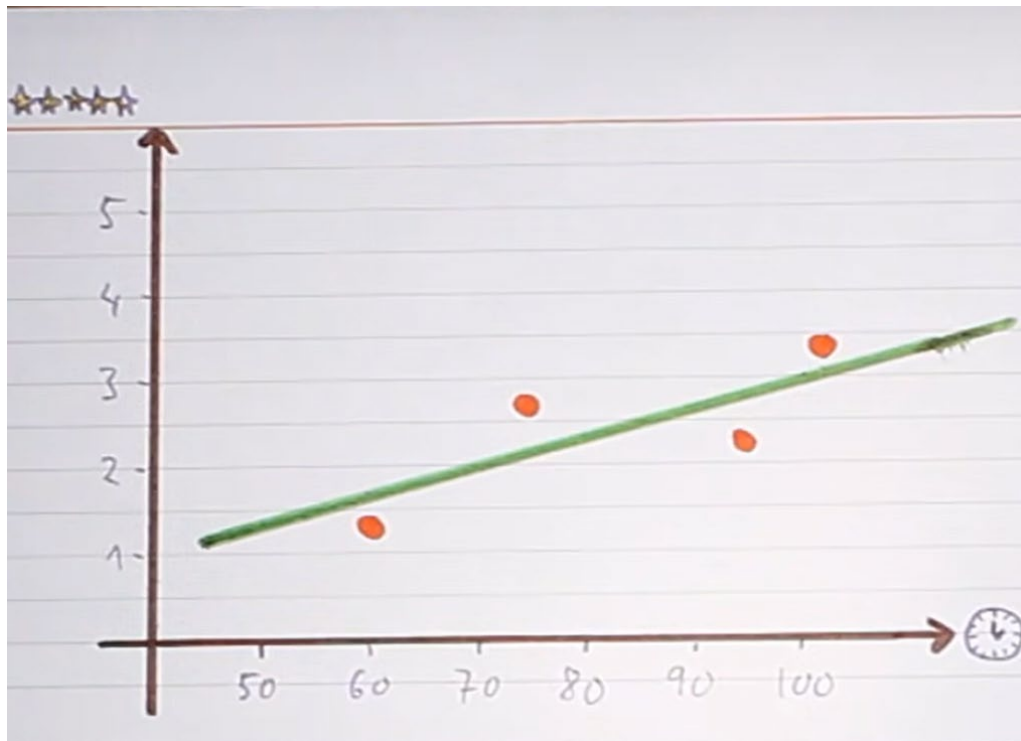


# Correlation Coefficient Steps

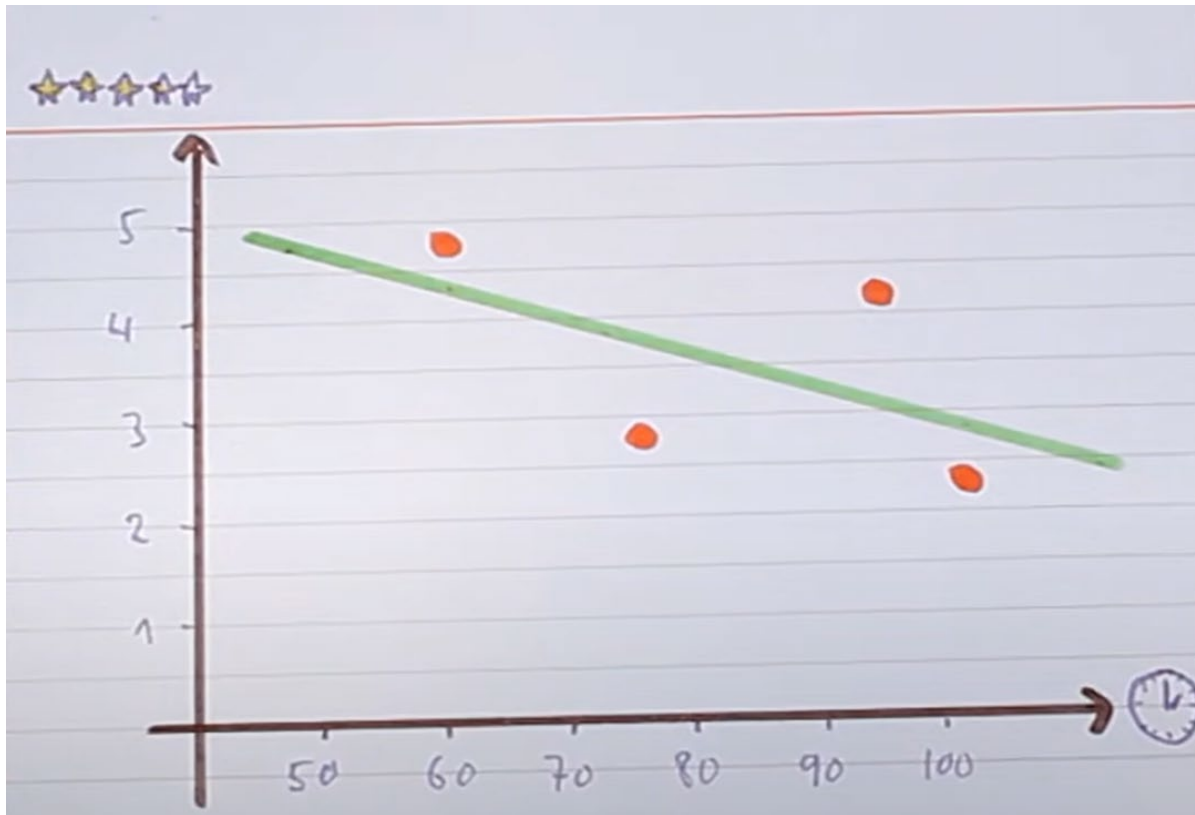
1. Fit Line
2. Direction of Slope (upwards/downwards)

# Correlation Coefficient Steps

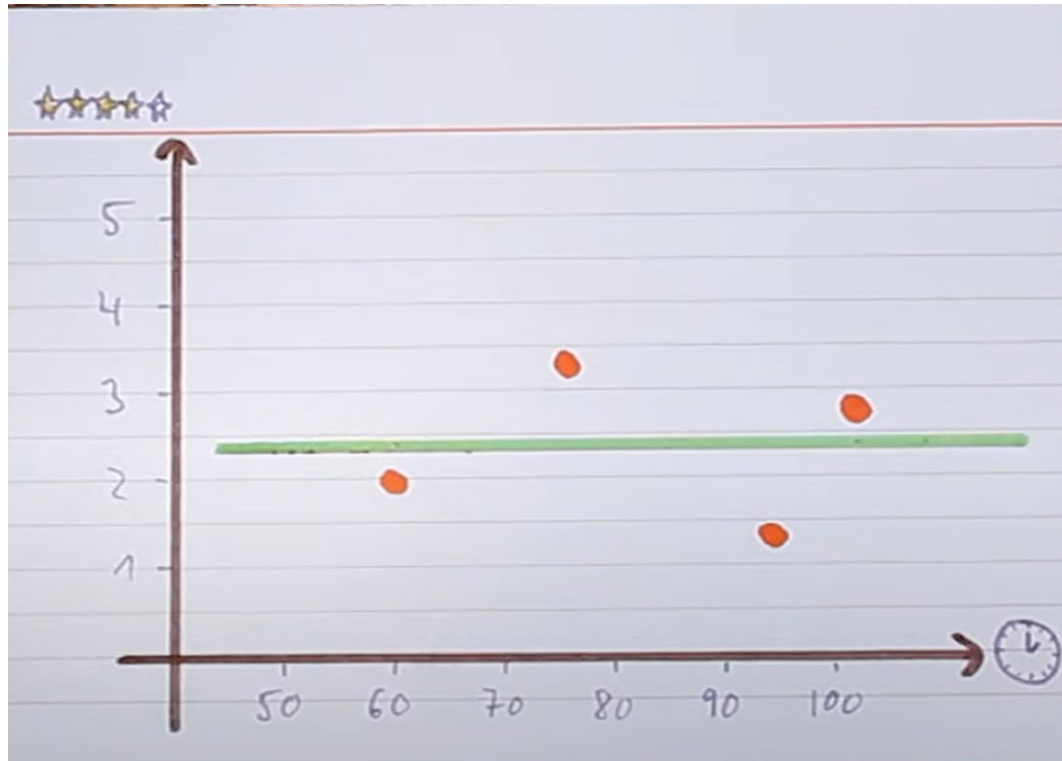
1. Fit Line
2. Direction of Slope



Correlation  
Coefficient  
= 0 and 1



Correlation Coefficient  
= 0 and -1

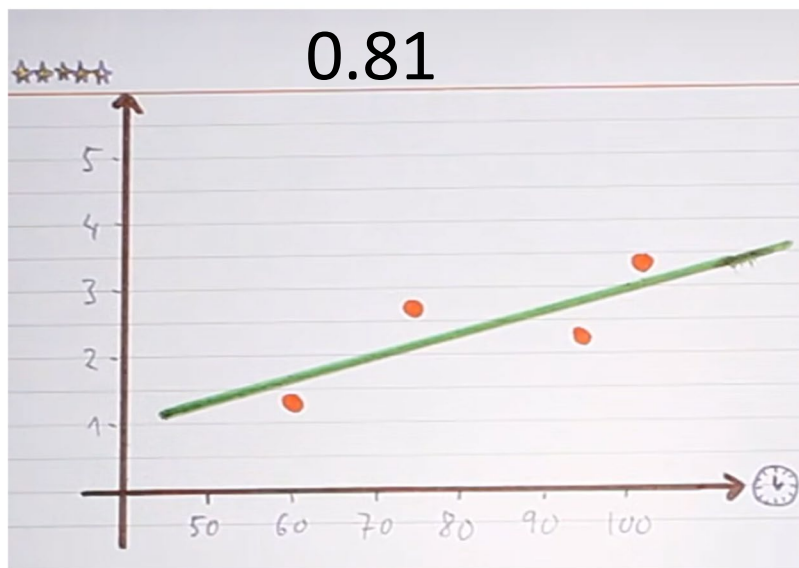


Correlation Coefficient = 0

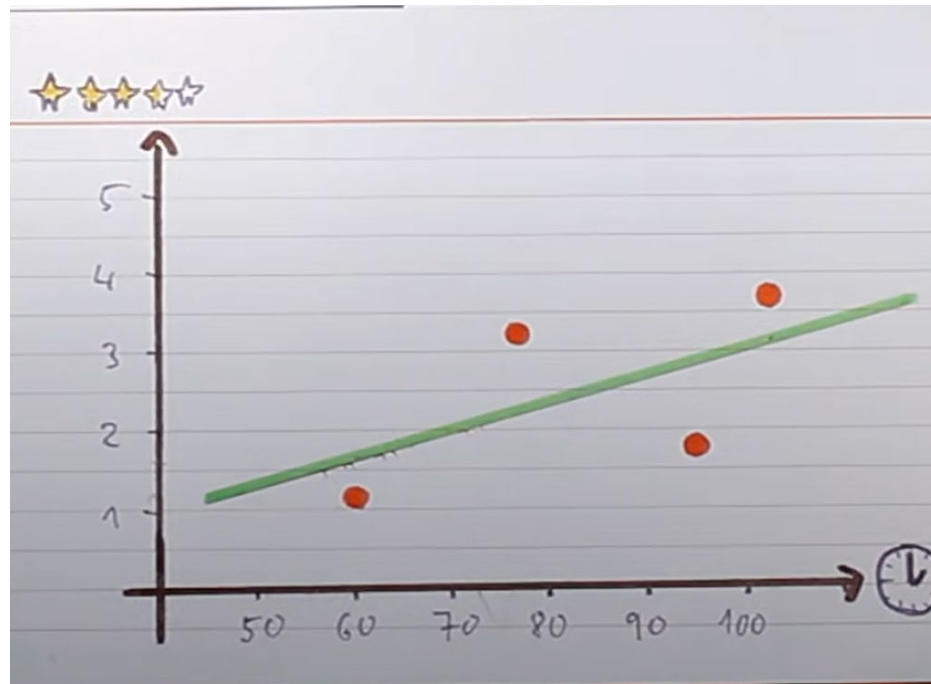
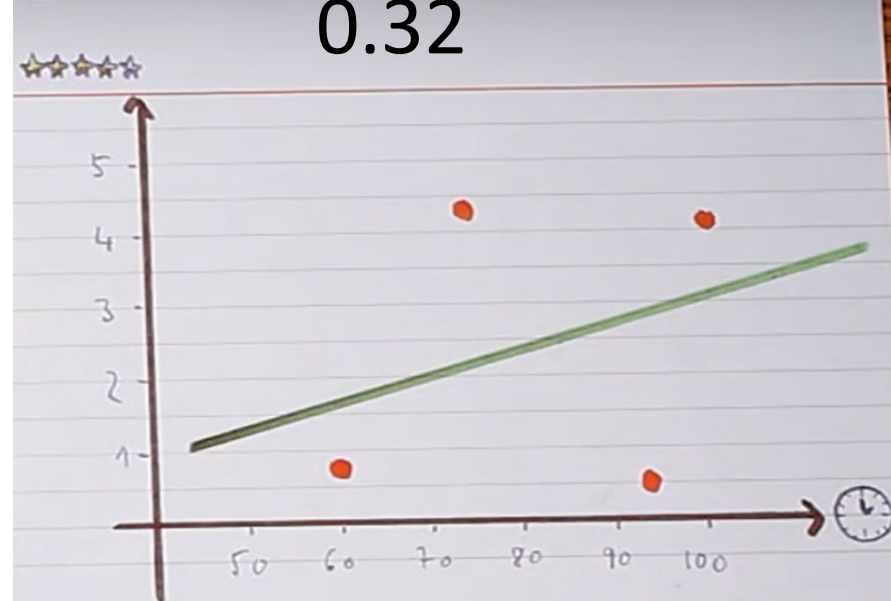
# Correlation Coefficient Steps

1. Fit Line
2. Direction of Slope
3. Quality of Fit

0.81



0.32



# Pearson Correlation Coefficient

- This describes the linear relationship between two variables  $x$  and  $y$

## Methods

- Deviation Score Formula
- Computing Formula

# Pearson Correlation Coefficient

- Deviation Score Formula

$$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y}$$

Where  $Cov_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$  or  $Cov_{xy} = \frac{\sum xy}{N - 1}$

$x = (X - \bar{X})$ , the deviation scores for X

$y = (Y - \bar{Y})$ , the deviation scores for Y

$\sum xy$  = the sum of the products of the paired deviation scores

$N$  = the number of pairs of scores

$SD_x$  = the standard deviation of X

$SD_y$  = the standard deviation of Y

# Example

|           | $x = (X - \bar{X})$ |                  | $y = (Y - \bar{Y})$ |      |       |
|-----------|---------------------|------------------|---------------------|------|-------|
| Persons   | X                   | Y                | x                   | y    | xy    |
| Alice     | 37                  | 75               | -1.8                | -4.8 | 8.64  |
| Bob       | 41                  | 78               | 2.2                 | -1.8 | -3.96 |
| Charlotte | 48                  | 88               | 9.2                 | 8.2  | 75.44 |
| David     | 32                  | 80               | -6.8                | 0.2  | -1.36 |
| Eve       | 36                  | 78               | -2.8                | -1.8 | 5.04  |
| N=5       | $\sum X = 194$      | $\sum Y = 399$   | $\sum xy = 83.8$    |      |       |
|           | $\bar{X} = 38.8$    | $\bar{Y} = 79.8$ |                     |      |       |
|           | $SD_x = 6.1$        | $SD_y = 4.9$     |                     |      |       |

$$\bar{X} = \frac{\sum X}{N}$$

$$= 194/5$$

$$= 38.8$$

# Example

|           | $x = (X - \bar{X})$ |                  | $y = (Y - \bar{Y})$ |      |       |
|-----------|---------------------|------------------|---------------------|------|-------|
| Persons   | X                   | Y                | x                   | y    | xy    |
| Alice     | 37                  | 75               | -1.8                | -4.8 | 8.64  |
| Bob       | 41                  | 78               | 2.2                 | -1.8 | -3.96 |
| Charlotte | 48                  | 88               | 9.2                 | 8.2  | 75.44 |
| David     | 32                  | 80               | -6.8                | 0.2  | -1.36 |
| Eve       | 36                  | 78               | -2.8                | -1.8 | 5.04  |
| N=5       | $\sum X = 194$      | $\sum Y = 399$   | $\sum xy = 83.8$    |      |       |
|           | $\bar{X} = 38.8$    | $\bar{Y} = 79.8$ |                     |      |       |
|           | $SD_x = 6.1$        | $SD_y = 4.9$     |                     |      |       |

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{(-1.8)^2 + (2.2)^2 + (9.2)^2 + (-6.8)^2 + (-2.8)^2}{4}} = 6.05$$

# Example

| Persons   | X                | Y                | x    | y    | xy               |
|-----------|------------------|------------------|------|------|------------------|
| Alice     | 37               | 75               | -1.8 | -4.8 | 8.64             |
| Bob       | 41               | 78               | 2.2  | -1.8 | -3.96            |
| Charlotte | 48               | 88               | 9.2  | 8.2  | 75.44            |
| David     | 32               | 80               | -6.8 | 0.2  | -1.36            |
| Eve       | 36               | 78               | -2.8 | -1.8 | 5.04             |
| N=5       | $\sum X = 194$   | $\sum Y = 399$   |      |      | $\sum xy = 83.8$ |
|           | $\bar{X} = 38.8$ | $\bar{Y} = 79.8$ |      |      |                  |
|           | $SD_x = 6.1$     | $SD_y = 4.9$     |      |      |                  |

$$Cov_{xy} = \frac{\sum xy}{N-1} = \frac{83.8}{5-1} = 20.95$$

$$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y} = \frac{20.95}{(6.1)(4.9)} = \frac{20.95}{29.89} = +0.70$$

# Pearson Correlation Coefficient

- Computing Formula

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

| Persons | X         | Y         | X <sup>2</sup>          | Y <sup>2</sup>            | XY            |
|---------|-----------|-----------|-------------------------|---------------------------|---------------|
| Justin  | 37        | 75        | 1369                    | 5625                      | 2775          |
| Theresa | 41        | 78        | 1681                    | 6084                      | 3198          |
| Zachary | 48        | 88        | 2304                    | 7744                      | 4224          |
| Koko    | 32        | 80        | 1024                    | 6400                      | 2560          |
| David   | 36        | 78        | 1296                    | 6084                      | 2808          |
| <hr/>   |           |           |                         |                           |               |
| N = 5   | Σ X = 194 | Σ Y = 399 | Σ X <sup>2</sup> = 7674 | Σ Y <sup>2</sup> = 31,937 | Σ XY = 15,565 |

$$\begin{aligned} r_{xy} &= \frac{419}{596} \\ &= +0.70 \end{aligned}$$

- Pearson is a measure of the linear relationship between two variables
- Curvilinear fashion - not supported
- Dots in scatter plot- distributing along a curved line (curvilinear) rather than a straight line (linear).
- Other correlational techniques :
- Spearman correlation for ranked data.

# Spearman Rank Correlation Coefficient

- The Spearman rank correlation ( $r_s$ ) - *monotonically* related data
- $r_s$  used for ranked Data
  - First ranking of scores
  - Second using the formula:

$$r_s = \frac{Cov_{xy}}{SD_x SD_y}$$

Where x and y are ranks

# Example

| Persons   | X  | $X_r$ | Y   | $Y_r$ | x    | y    | xy    |
|-----------|----|-------|-----|-------|------|------|-------|
| Alice     | 30 | 1     | 100 | 6     | -2.5 | 2.5  | -6.25 |
| Bob       | 40 | 2.5   | 80  | 3     | -0.5 | -0.5 | 0.5   |
| Charlotte | 40 | 2.5   | 70  | 2     | -1   | -1.5 | 1.5   |
| David     | 50 | 4     | 90  | 5     | 0.5  | 1.5  | 0.75  |
| Eve       | 60 | 5     | 88  | 4     | 1.5  | 0.5  | 0.75  |
| Frank     | 70 | 6     | 60  | 1     | 2.5  | -2.5 | -6.25 |

1. Sort scores in ascending
2. Assign Ranks

# Example

| Persons   | X                      | $X_r$ | Y                      | $Y_r$ | x                | y    | xy    |
|-----------|------------------------|-------|------------------------|-------|------------------|------|-------|
| Alice     | 30                     | 1     | 100                    | 6     | -2.5             | 2.5  | -6.25 |
| Bob       | 40                     | 2.5   | 80                     | 3     | -0.5             | -0.5 | 0.5   |
| Charlotte | 40                     | 2.5   | 70                     | 2     | -1               | -1.5 | 1.5   |
| David     | 50                     | 4     | 90                     | 5     | 0.5              | 1.5  | 0.75  |
| Eve       | 60                     | 5     | 88                     | 4     | 1.5              | 0.5  | 0.75  |
| Frank     | 70                     | 6     | 60                     | 1     | 2.5              | -2.5 | -6.25 |
| N = 6     | $\sum X_r = 21$        |       | $\sum Y_r = 21$        |       | $\sum xy = -9.0$ |      |       |
|           | $\overline{X_r} = 3.5$ |       | $\overline{Y_r} = 3.5$ |       |                  |      |       |
|           | $SD_x = 1.8$           |       | $SD_y = 1.9$           |       |                  |      |       |

3. Find N,  $\overline{X_r}$ ,  $\overline{Y_r}$ ,  $SD_x$ ,  $SD_y$ , and  $\sum xy$  using the ranks

# Example

| Persons   | X                      | $X_r$ | Y                      | $Y_r$ | x                | y    | xy    |
|-----------|------------------------|-------|------------------------|-------|------------------|------|-------|
| Alice     | 30                     | 1     | 100                    | 6     | -2.5             | 2.5  | -6.25 |
| Bob       | 40                     | 2.5   | 80                     | 3     | -0.5             | -0.5 | 0.5   |
| Charlotte | 40                     | 2.5   | 70                     | 2     | -1               | -1.5 | 1.5   |
| David     | 50                     | 4     | 90                     | 5     | 0.5              | 1.5  | 0.75  |
| Eve       | 60                     | 5     | 88                     | 4     | 1.5              | 0.5  | 0.75  |
| Frank     | 70                     | 6     | 60                     | 1     | 2.5              | -2.5 | -6.25 |
| N = 6     | $\sum X_r = 21$        |       | $\sum Y_r = 21$        |       | $\sum xy = -9.0$ |      |       |
|           | $\overline{X_r} = 3.5$ |       | $\overline{Y_r} = 3.5$ |       |                  |      |       |
|           | $SD_{xr} = 1.8$        |       | $SD_{yr} = 1.9$        |       |                  |      |       |

5. Calculate the  $Cov_{xyr} = \frac{\sum xy}{N-1} = \frac{-9.0}{6-1} = -1.8$

6. Finally calculate the  $r_s = \frac{Cov_{xy}}{SD_x SD_y} = \frac{-1.8}{(1.8)(1.9)} = \frac{-1.8}{3.42} = -0.53$

# Correlation Coefficients Interpretation

- **Effect size** - magnitude and direction of the differences between 2 groups

Recall:

- perfect correlation between two variables would result in  $= +1.00$  or  $-1.00$
- No correlation between X and Y, then  $r_{xy} = 0$
- Example:  $r_{xy} = +0.70$  --→ Imply?

# Coefficient of Determination

$$r_{xy}^2$$

$$r_{xy} = 0.7$$

$$r_{xy}^2 = 0.49 = 49\%$$

Implies=> Almost half of the differences in Y are due to differences in X

# Coefficient of Determination

Suppose a correlation of 0.6 exists between the number of risk factors for skin cancer (X) a patient has and the likelihood of the patient suffering from skincancer (Y)

$$r_{xy} = 0.6$$

What does it imply?

# Coefficient of Determination

Suppose a correlation of 0.6 exists between the number of risk factors for skin cancer (X) a patient has and the likelihood of the patient suffering from skin cancer (Y)

$$r_{xy} = 0.6$$

What does it imply?

$$r_{xy}^2 = 0.36$$

36% of the risk is due to these factors, but 64% is due to other

unknown factors

- Large Correlation = 0.50
- Moderate Correlation = 0.30
- Small Correlation = 0.10

<https://twitter.com/bioluisinho/status/1224404802693668864>

Don't confuse correlation with causation.

# Exercise

Given the following data, what are the correlation and the coefficient of determination between the following?

(a) Scores on tests of intelligence (IQ) and emotional intelligence (EQ)

(b) IQ scores and exam scores

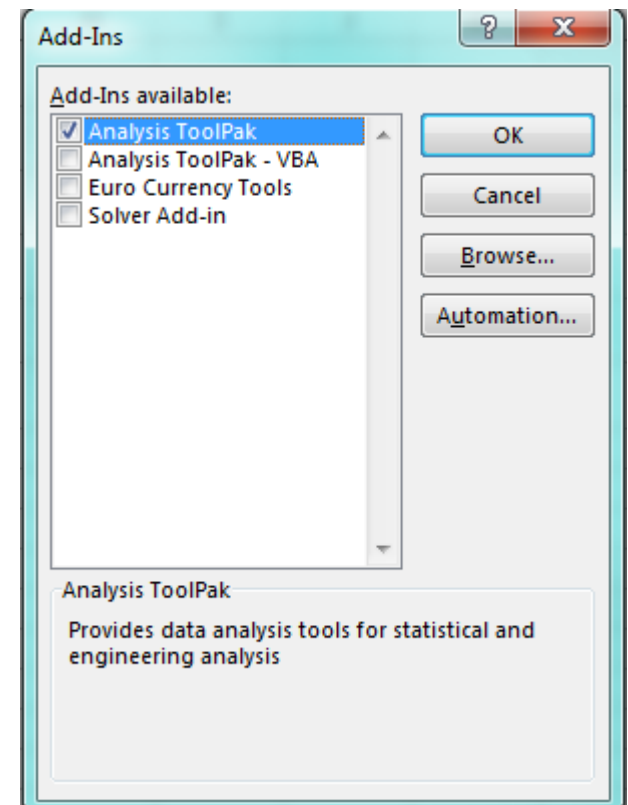
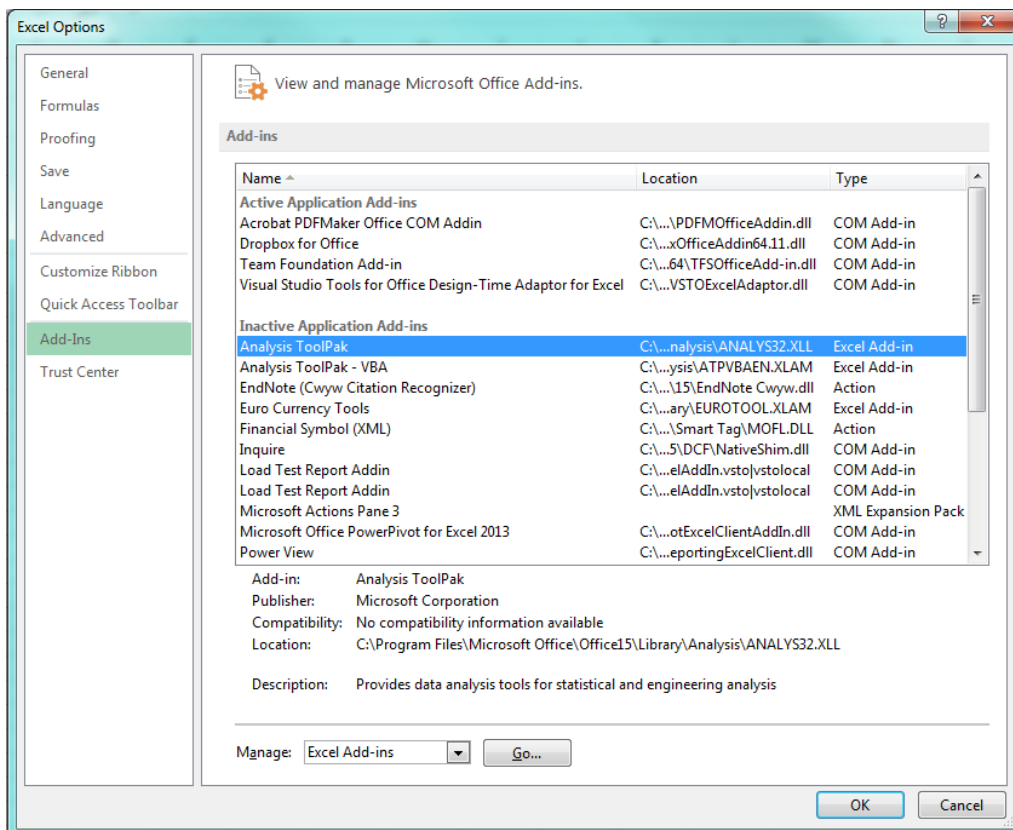
(c) EQ test scores and exam scores

In addition, how would you interpret the size of each of these correlations?

| Student | IQ (X) | EQ (Y) | Exam Scores (Z) |
|---------|--------|--------|-----------------|
| 1       | 140    | 14     | 42              |
| 2       | 130    | 20     | 44              |
| 3       | 120    | 29     | 35              |
| 4       | 119    | 6      | 30              |
| 5       | 115    | 20     | 23              |
| 6       | 114    | 27     | 27              |
| 7       | 114    | 29     | 25              |
| 8       | 113    | 30     | 20              |
| 9       | 112    | 35     | 16              |
| 10      | 111    | 40     | 12              |

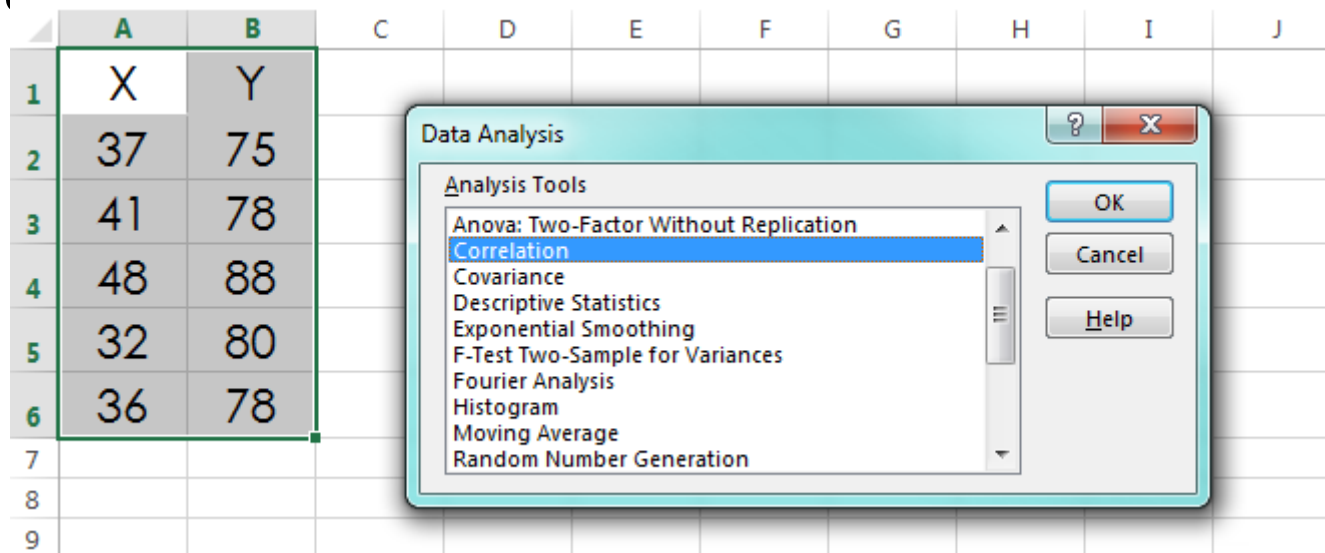
# Correlations in Excel

- To do any complex calculations in Excel requires the Analysis ToolPak
  - This is not part of the standard Excel build and has to be added in through the Excel 'Options' dialogue



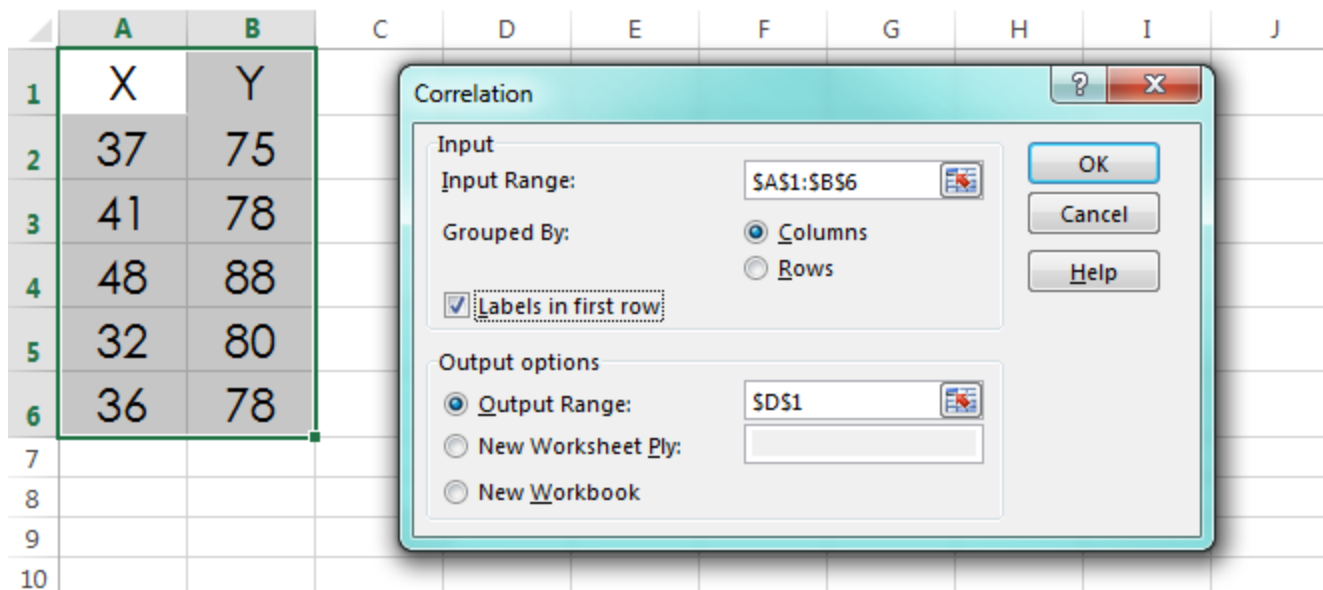
# Correlations in Excel

- Once the Analysis ToolPak has been added it will appear at the end of the Data tab of the ribbon
- Select 'Correlation' from the data analysis tool



# Correlations in Excel

- Input the data range for analysis, input a location for the data to be output, and tick the labels box if appropriate



# Correlations in excel

|   | A  | B  | C | D | E        | F | G |
|---|----|----|---|---|----------|---|---|
| 1 | X  | Y  |   |   | X        | Y |   |
| 2 | 37 | 75 |   | X | 1        |   |   |
| 3 | 41 | 78 |   | Y | 0.702981 | 1 |   |
| 4 | 48 | 88 |   |   |          |   |   |
| 5 | 32 | 80 |   |   |          |   |   |
| 6 | 36 | 78 |   |   |          |   |   |

- On completion, Excel will calculate the Pearson's correlation for the data
- Spearman's correlation is calculated in the same manner, only the rank data is used rather than the score data
  - N.B. Most of the time Spearman's correlation and Pearson's correlation do not give the same results (Spearman's rho for this dataset is 0.20)
  - Why might this be the case?

# Exercise/Homework

- Exercise 3: Repeat exercise 1 but calculate Spearman's rho rather than Pearson's rho
- Perform exercises 1, 2 and 3 using Excel

# Summary

- Topics covered
  - Correlation Coefficient
  - Pearson Correlation Coefficient
  - Interpreting Correlation Coefficients
  - Spearman Rank Correlation
- Next week we will cover
  - Regression Equations
  - Standard Errors of the Equations
- Regression allows us to start making predictions with statistics rather than simply describing a dataset

# Coming soon

- **Next session:** Assignment workshop
- **Following session:** Linear Regression
- **Homework:** Read chapter 8 of *Statistics for the Terrified*.