



Identifying Patterns

LECTURER: MARK GILLAN

Correlations

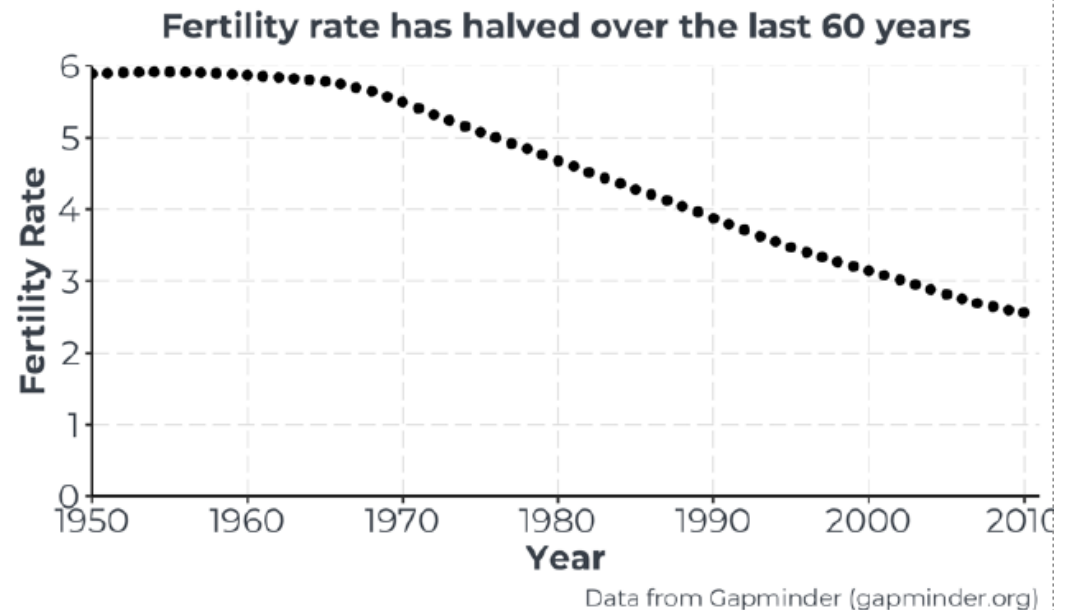
INCREASING, DECREASING, CYCLICAL

Once data has been manipulated and visualised, a pattern may be immediately obvious. This could be increasing, or decreasing, or cyclical.

The pattern could be over time, or by age band, or another independent analysis variable.

For example, this graph from the Gapminder (dataed.in/gapdata) dataset shows a very clear decreasing pattern in fertility rate in India over the last half century.

INDIAN FERTILITY RATE CHANGE



Correlation

RELATIONSHIP
BETWEEN TWO
VARIABLES



Correlation in detail

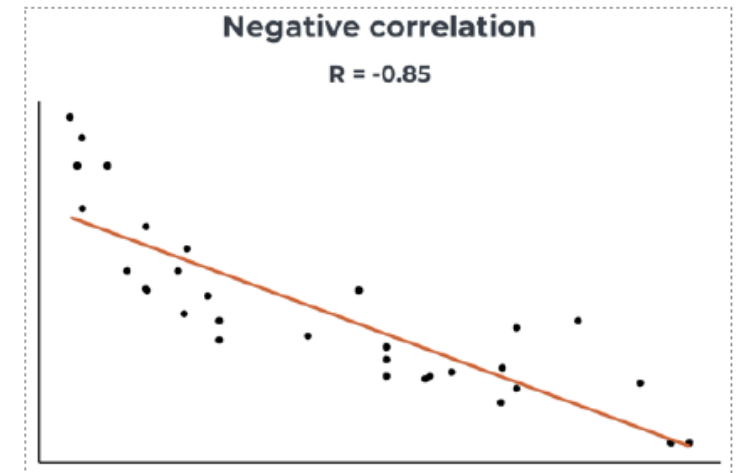
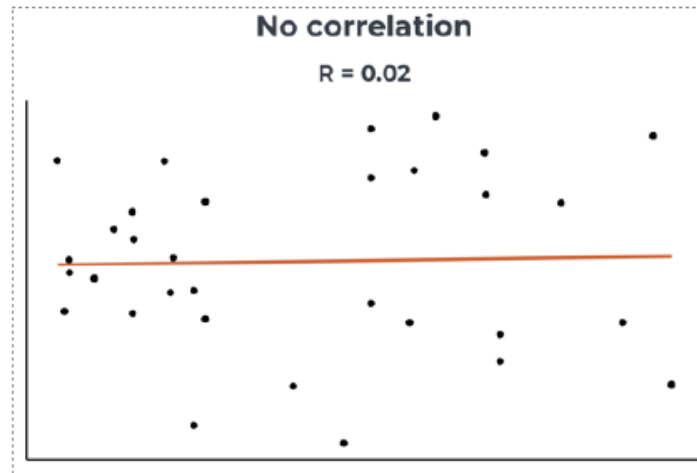
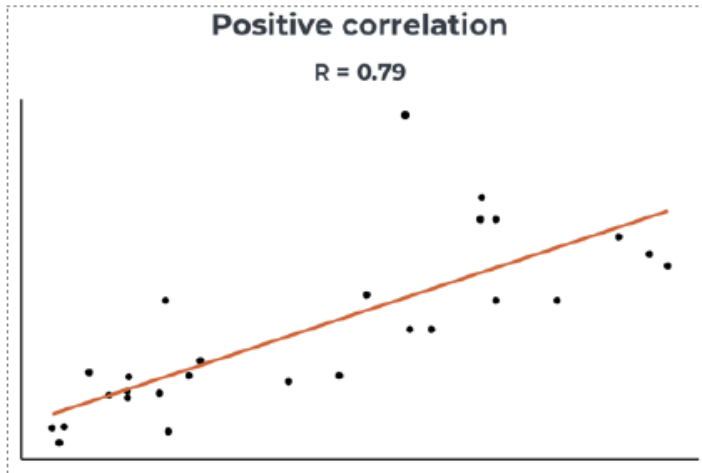
RELATIONSHIP BETWEEN TWO VARIABLES

SOMETIMES THE PATTERNS ARE LESS OBVIOUS, AND IT BECOMES NECESSARY TO STATISTICALLY TEST FOR THE PRESENCE OF A RELATIONSHIP. IT IS THEN USEFUL TO CALCULATE A CORRELATION COEFFICIENT.

THERE ARE A NUMBER OF DIFFERENT CORRELATION COEFFICIENTS, BUT THE MOST COMMONLY USED IS PEARSON'S (DATAED.IN/CORR) .

THIS IS A NUMBER THAT SUMMARISES BOTH THE STRENGTH AND DIRECTION OF THE RELATIONSHIP.

Correlation between -1 and 1



Correlation relationships

The correlation coefficient (R) takes a value between -1 and 1, with **-1 being a perfect negative correlation** and **+1 being a perfect positive correlation**.

A value of around 0 implies no correlation or relationship between the variables.

The assumption here is that the relationships are linear, they form a straight line.

Relationships can also be non-linear, and it may be necessary to fit a differently shaped function to the data to test for a relationship.

It is important to remember that “correlation does not imply causation”.

What this means is the change in one variable does not necessarily cause the other variable to change.

There may often be a **hidden confounding factor**, which is a variable that, related to both the variables being studied, which when changes, causes a change in both the analysis variables.

Extract Insights



Having identified a pattern in the data, this needs to be turned into an **actionable insight**. This ensures that the knowledge of the data showing a pattern provides information that can be used to solve a problem.



Before sharing this knowledge, the insight will need to be tested to ensure it holds true for further data sources, or over a different time period.



It is also important to review the analysis process carefully at this stage to ensure no mistakes have been made in the manipulation.



Finally, it is necessary to ensure the pattern isn't just a consequence of an **underlying bias** in the original data. Only at this point can the insight be used to make a decision, drive a new process, or prompt a different action.

Correlation:

What is the definition of correlation?

What does -1 mean?

What does +1 mean?

