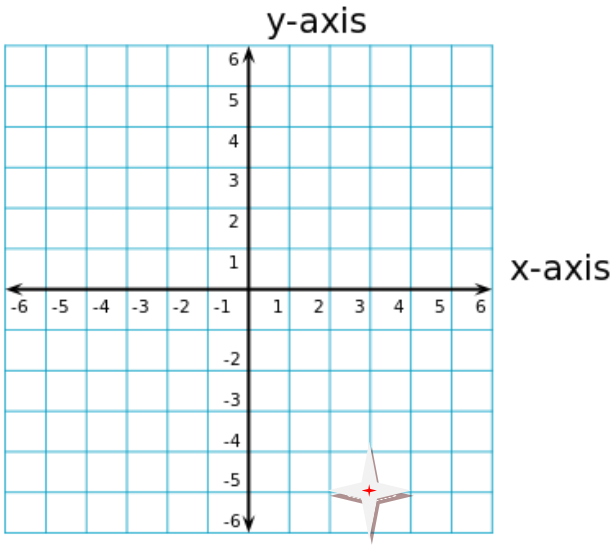
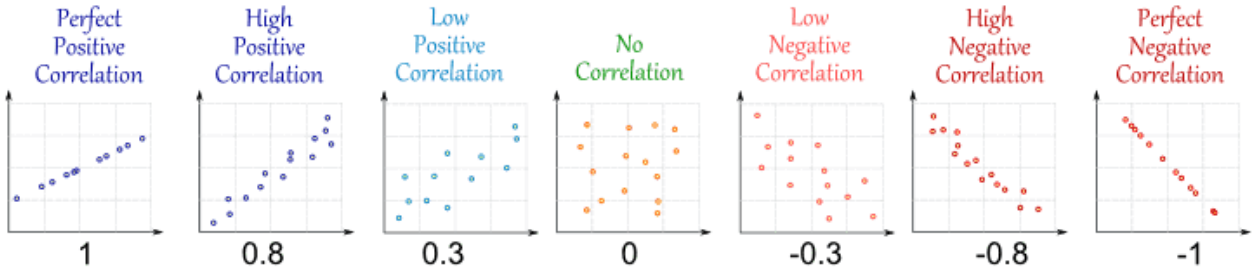
SCATTERPLOT REFERENCE SHEET

A **scatterplot** is a graph of x and y coordinates plotted to illustrate whether a relationship exists between two sets of data (the x and the y values).

The **Cartesian Plane** (conceived by René desCartes, a Swiss mathematician) is made up of an x-axis (or the horizontal axis) and the y-axis (or the vertical axis). A point, (x,y) can be located by moving x units along the x-axis followed by y units along the y-axis. For example, the point (x,y) = (3,-5) can be found by moving three units right (positive) along the x-axis and then one unit down (negative) along the y-axis.



A **correlation** defines the strength at which two sets of data are related.



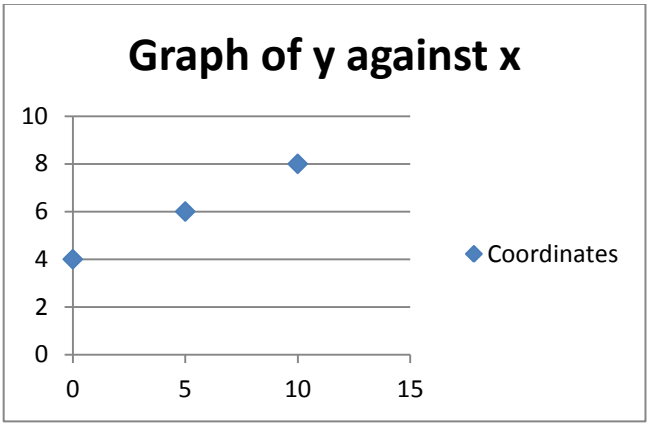
The scatterplot on the left is an example of a strong correlation because the data forms a near perfect line. When a line can be drawn across the data using a **line of best fit** we say that the data is **linear**. The The scatterplot in the centre is an example of a correlation where the two data sets are not related. In this case, we cannot draw a **line of best fit**.

A correlation is said to be positive when **as the x-values rise so do the y-values**. In other words, if when the x-values **increase** so too do the y-values **increase**, then we have a **positive** correlation. If however, as the x-values **increase** but the y-values **decrease**, then we have a **negative** correlation.

A **line of best fit** is a line that the author must draw using their judgement in an attempt to quantify the linear relation. Correlations are deemed to be strong (very easy to draw the line), moderate (easy to draw but some judgement required and/or the data strays from the line), weak (difficult with a lot of judgement required and/or much data strays from the line) or none (it cannot be drawn despite judgement or it is not linear).

A **variable** is simply a representation of some concept's unknown quantity. The **independent variable** is located on the x-axis (time for example is **ALWAYS** on the x-axis) and is listed first on any table. This data is independent of the other variable. The **dependent variable** is located on the y-axis and is listed second on the table. This data is dependent on the independent variable.

Independent Variable x	Dependent Variable y
0	4
5	6
10	8



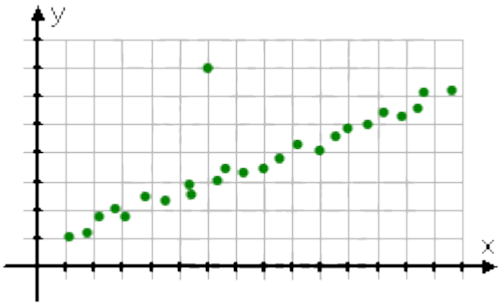
In statistics, we **interpolate** data when the author must make a prediction using a scatterplot when the result is **within** the data set. An author **extrapolates** data when a prediction is made using a scatterplot where the result is **outside** the data set. For example, if given data for every 10 years between the years 1950 to 2010, we would interpolate the data for 1975 but we would extrapolate the data for 1945 or 2014.

A **trend** is a description of how two data sets move in relation to each other. For example, as a child's age increases, so does the child's height.

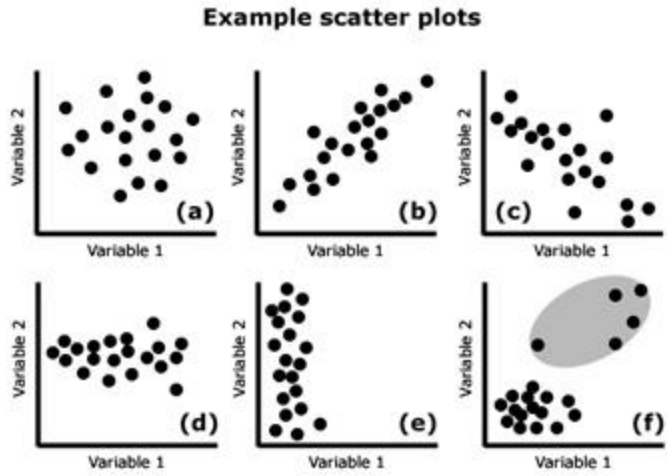
A **continuous** set of data is a description of data that exists for all values. For example, as time passes (increases) the distance that a jogger (either decreases, increases or remains). At any point in time, I can locate the jogger.

A **discrete** set of data is a description of data that only exists for specific values. For example, as time passes (increases) the number of holes that I dig (you can only have 0, 1, 2, …). In other words, you cannot dig half a hole.

An **outlier** is an anomaly point. This is a point that, for some reason, completely strays from the tendency of the correlation. In essence, we ignore this point when drawing the line of best fit and making our trend statement. In the case below, we would state that this scatterplot has a strong, positive correlation. We could further state, that an outlier exists at (6,7).



Examples:



In ALL examples, variable 1 is the independent variable because it lies on the x-axis, therefore, variable 2 is the dependent variable since it is on the y-axis.

In example (a), there is no correlation as the data is too dispersed to make trend statement.

In example (b), there is a strong correlation as it would be easy to draw a line of best fit. The correlation is positive since the trend is that "as variable 1 increases, variable 2 also increases".

In example (c), there is a moderate correlation as it would be easy to draw a line of best fit, but the data does drift away from the line. The correlation is negative since the trend is that "as variable 1 increases, variable 2 decreases".