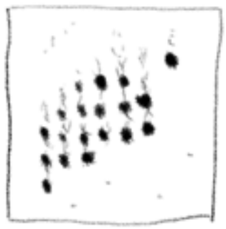
- Two variables are related if one is implicit of information about the other.

SCATTER PLOTS



"Unjittered"

"Jittered"

Jittering removes the effect of rounding, and can make the relationships clearer.

How-to jitter — add random noise to reverse roundings.

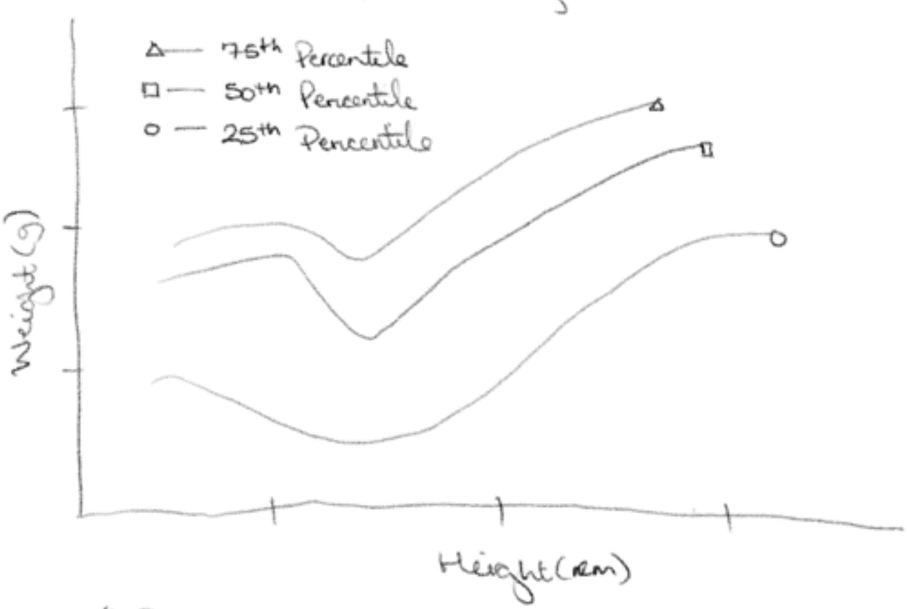
Saturation — disproportionate emphasis to outliers, obfuscates density — or overlapp.

- Can use transparency to reflect density.
- Can "bin" data to better depict density.

CHARACTERIZING RELATIONSHIP

1) Bin to Percentile Plot

Bin one var plot percentiles of the other var.



"Percentiles of weight for a range of height bins".

CORRELATION

Statistic which quantifies the strength of a relationship between two variables.

In order to compute this quantity, the relationships between the variables must be comparable. The values have to be in the same units and share a frame of reference so their comparisons mean anything.

Methods:

I. Standard Score – number of standard deviations from mean. The "Pearson Product-Moment Correlation Coefficient". (PMCC

II. Rank – an index in a sorted list of values. The "Spearman Rank Correlation Coefficient".

Standard Z-Score Calculation

$$\underbrace{Z_i = \frac{(x_i - \mu)}{\sigma}}$$

x_i – is datapoint
μ – is average
σ – is std.

Dividing the deviation standardizes the magnitudes of difference.

Dimensionless (no units).

Distribution of Z-var has mean of (0) and Variance $(\sigma^2) = 1$.

- The distribution of Z-scores is the same as the underlying distribution X-var.
  - ↳ When Z is not of a normal distribution

it is better to translate that distribution to rank order. Rank-order acts as a global shared frame of reference. Vitally rank-order is uniformly distributed.

COVARIANCE

The degree to which two variables similarily vary.

$dx_i = x_i - \bar{x}$
$dy_i = y_i - \bar{y}$

If variables vary similarly — then their deviations will share sign.

$Cov(X, Y) = \frac{1}{n} \sum dx_i dy_i$, where n is the length of both series. $\|X\| = \|Y\|$

PEARSON'S CORRELATION (PMCC)

Cov is only useful in some calculations. Not great summary statistic.

↳ combines units in dot product.

* Equal to the dot-product. $\vec{dx} \cdot \vec{dy}$

+ = similar
0 = orthogonal / neutral
− = dissimilar / opposite.

To remove the units can convert to standard z-scores.

$$p_i = \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x S_Y} \Rightarrow p = \frac{1}{n} \sum p_i \Rightarrow p = \frac{Cov(X, Y)}{S_X S_Y}$$

Named after Karl Pearson. Easy to use because its dimensionless.

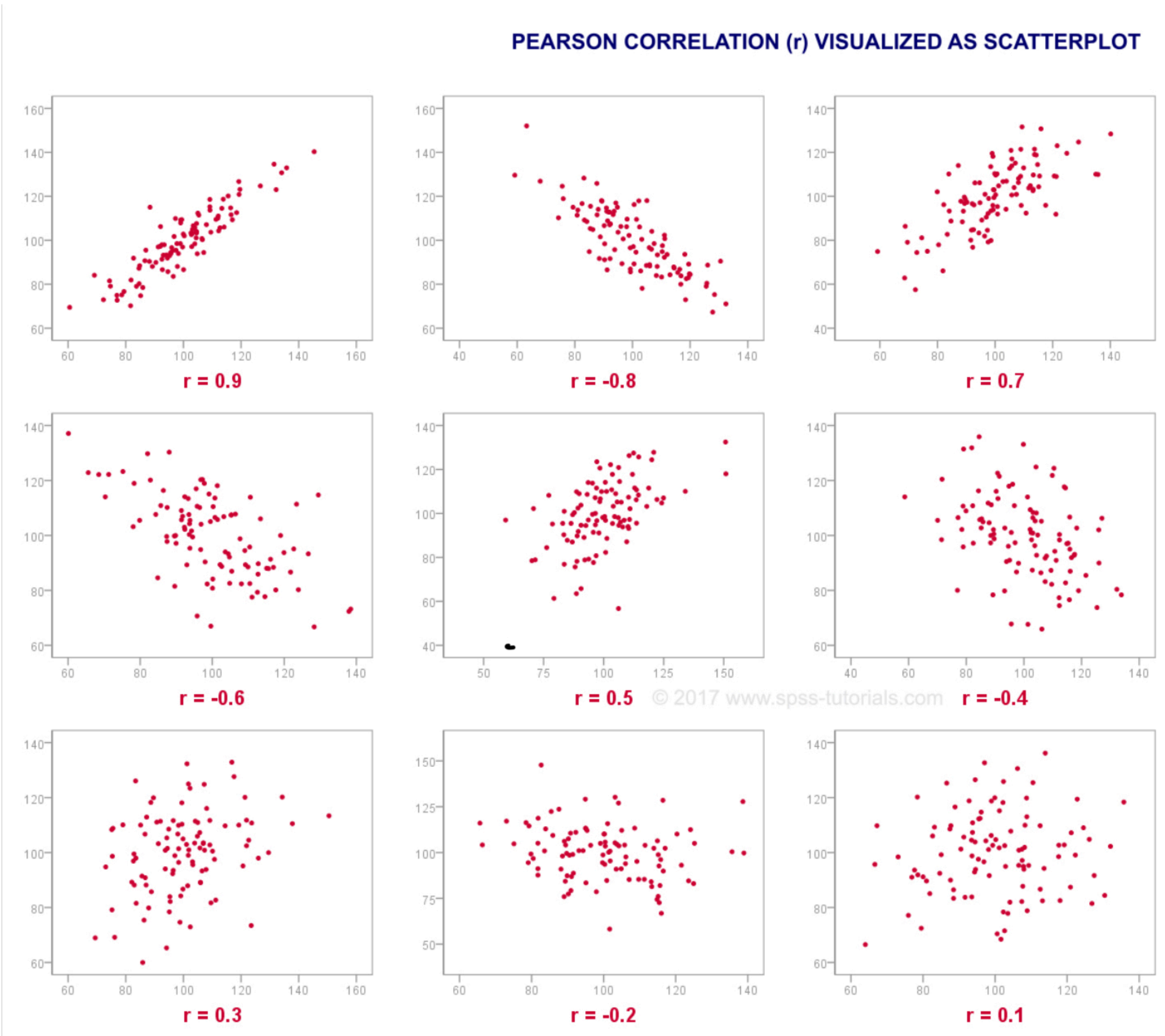
Meaning (from −1 to 1 inclusive)

(*) 1s are perfect correlation.

+1 — both high

-1 — one high other low

Should relate Pearson Correlation to other variables, with same Pearson Correlation applied.



Nonlinear Relationships

- PMCC good for metrics ratios + intervals
- Not good for ordinal / nominal.

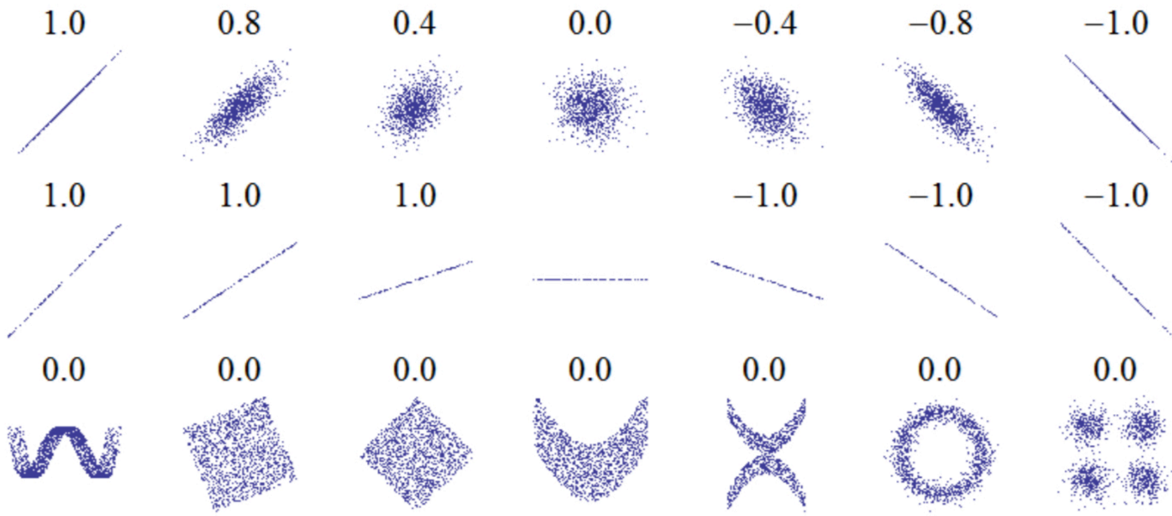


Figure 7.4: Examples of datasets with a range of correlations.

Pearson's Correlation only measures linear relationships.

- Perfect correlations are slope agnostic.
- Non-linear relationships can have a correlation coefficient of 0.
- Not good w-outliers
- Good for roughly normal distrib.

⇒ Look at the scatterplot before computing correlation coefficients.

CORRELATION & CAUSATION.

SPEARMAN'S RANK CORRELATION

- Robust against outliers and skewed distributions.
- Good for ordinal data.

1. Compute the rank of the deviations
2. Compute Pearson's Correlation with those ranks.

Alternatively, can remove skewness with logarithms. Then applying Pearson's Correlation.

The reason for applying logarithms to a skewed dataset is to elucidate a non-linear logarithmic relationship in the original scale.

Spearman's without tied ranks.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$, where d_i is the difference in paired ranks.

$n(n-1)$ in paired ...

Spearman's with tied ranks,

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

, where i is the paired score.

Spearman measures the strength and direction of two variables $\boxed{\text{monotonicity.}}$

Rank 1.	Rank 2.	d_i	d_i^2
9	4	5	25
3	2	1	1
10	10	0	0
4	7	3	9
6	5	1	1

$$\sum d_i^2 = 25 + 1 + 0 + 9 + 1$$

$$\rho = 1 - \frac{6(36)}{5(5^2-1)}$$

$$= 1 - \frac{216}{120}$$

, $\rho = r_s$ - analgous.

Correlation vs Causation

Correlation implies one of three cases,

$\left.\begin{array}{l}\text{1. A causes B.}\\ \text{2. B causes A.}\end{array}\right\}$ Causal Relationships

3. Some factors cause A & B.

Evidence of Causation.

- Use of time – the order of events can indicate the directionality of causation. It does not rule out external factors causing both A and B.

- Use Randomness – division of sample into multiple groups and then computing the means of variables should yield little difference. Can remove spurious relationships.
  ↳ should median be used as mean is sensitive to outliers.

This is the motivation for randomized control trials; and the introduction of change into one group.

In some cases it is possible to refer causation from regression.

Randomized Control Trial (RCT)

- aimed at removing selection bias.
  - randomization guards against the ignorance of unknown prognostic factors.