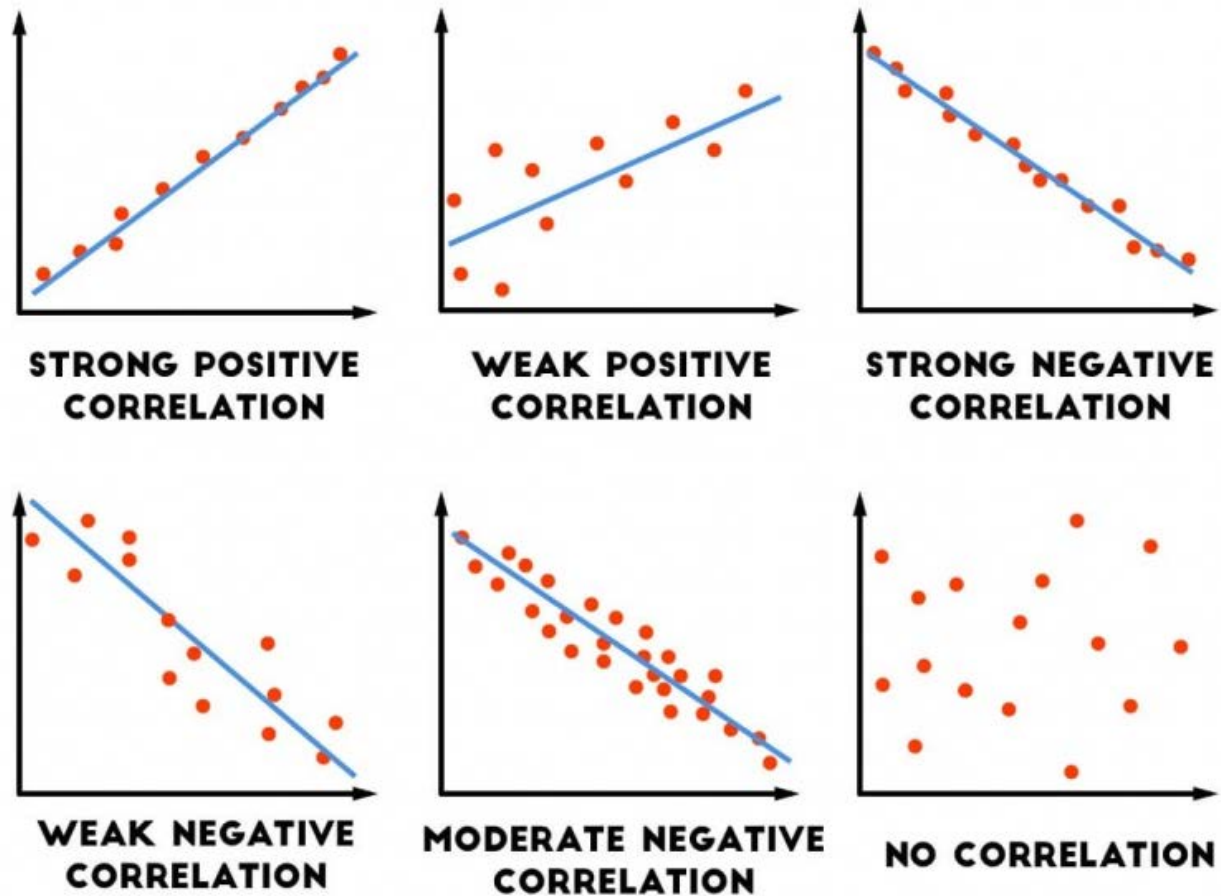# FEATURE SELECTION

Themistoklis Diamantopoulos

Correlation

- Similar information

Correlation Statistics

- Variables X, Y
- Variance

$$\text{var}(X) = E\left[(X - E[X])^2\right]$$

- Covariance

$$\text{cov}(X,Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

- Correlation

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

Values from [-1, +1]
+: Positive correlation
-: Negative correlation
0: No correlation

Source: https://en.wikipedia.org/wiki/Covariance

Mutual Information

- Variables X, Y
- Measures the dependence between random variables

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

- Higher dependence → Higher mutual information
- If the variables are independent then mutual information is 0

Chi-Squared Statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$

O_i: actual count
E_i: expected count

	Play Chess	Don't Play Chess	Sum
Like Science Fiction	250 90	200 360	450
Don't Like Science Fiction	50 210	1000 840	1050
Sum	300	1200	1500

- Example:

$$\chi^2 = \frac{\left(250 - 90\right)^2}{90} + \frac{\left(50 - 210\right)^2}{210} + \frac{\left(200 - 360\right)^2}{360} + \frac{\left(1000 - 840\right)^2}{840} = 507.93$$