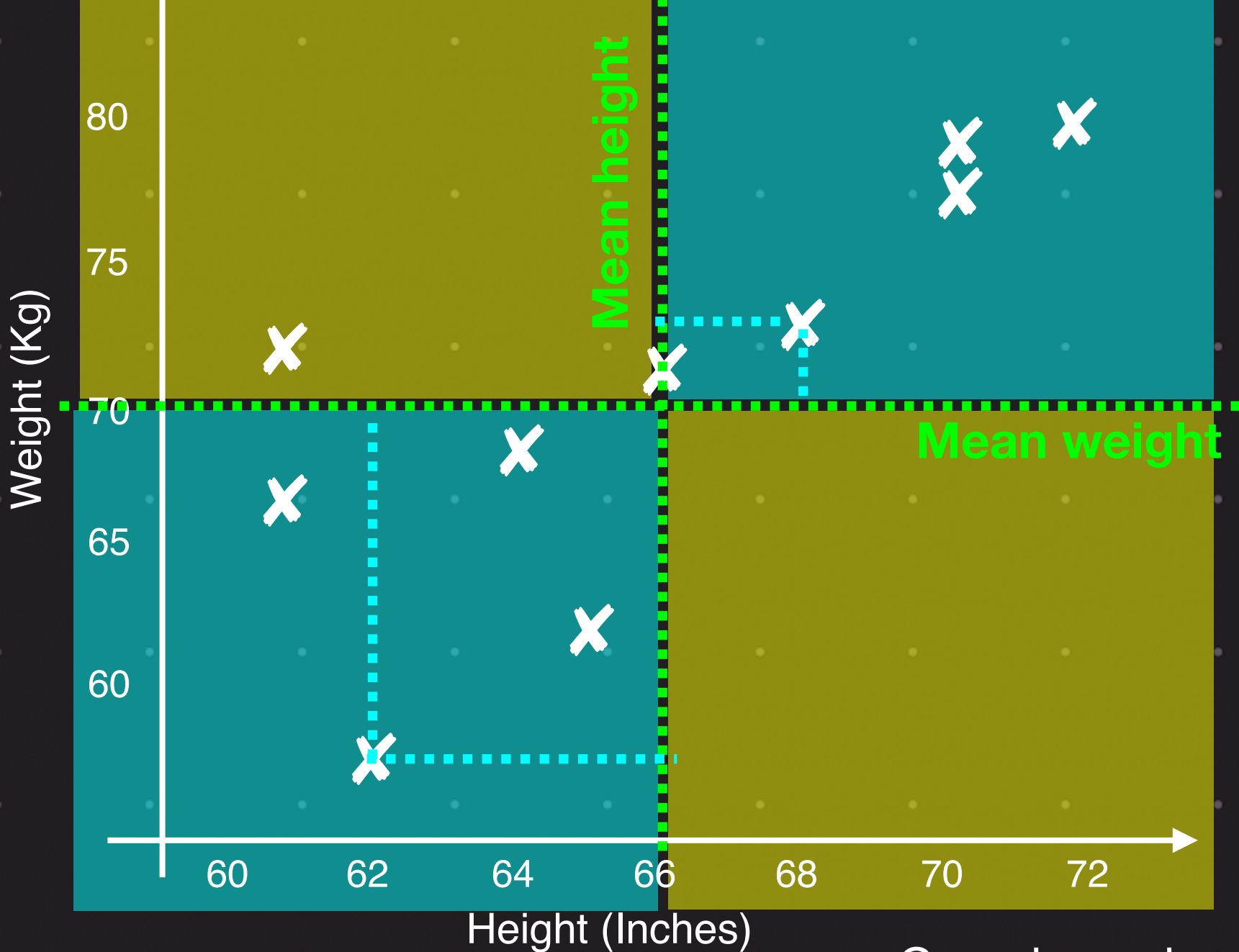


Height (inches)	Weight (kg)
68	72
62	58
64	67
61	72
70	79
66	61
61	68
65	64
71	80
72	79
$\bar{h} = 66$	$\bar{w} = 70$



Positive correlation

- Top right
- Bottom left

Negative correlation

- Top left
- Bottom right

$$\text{cov}(h, w) = \frac{1}{n} \sum_i (h_i - \bar{h})(w_i - \bar{w})$$

$$\begin{aligned} (68 - 66)(72 - 70) &= 2 * 2 = 4 \\ (62 - 66)(58 - 70) &= (-4) * (-12) = 48 \\ (64 - 66)(67 - 70) &= (-2) * (-3) = 6 \\ (61 - 66)(72 - 70) &= (-1)(2) = -2 \\ &\vdots \\ (72 - 66)(80 - 70) &= (6)(10) = 60 \end{aligned}$$

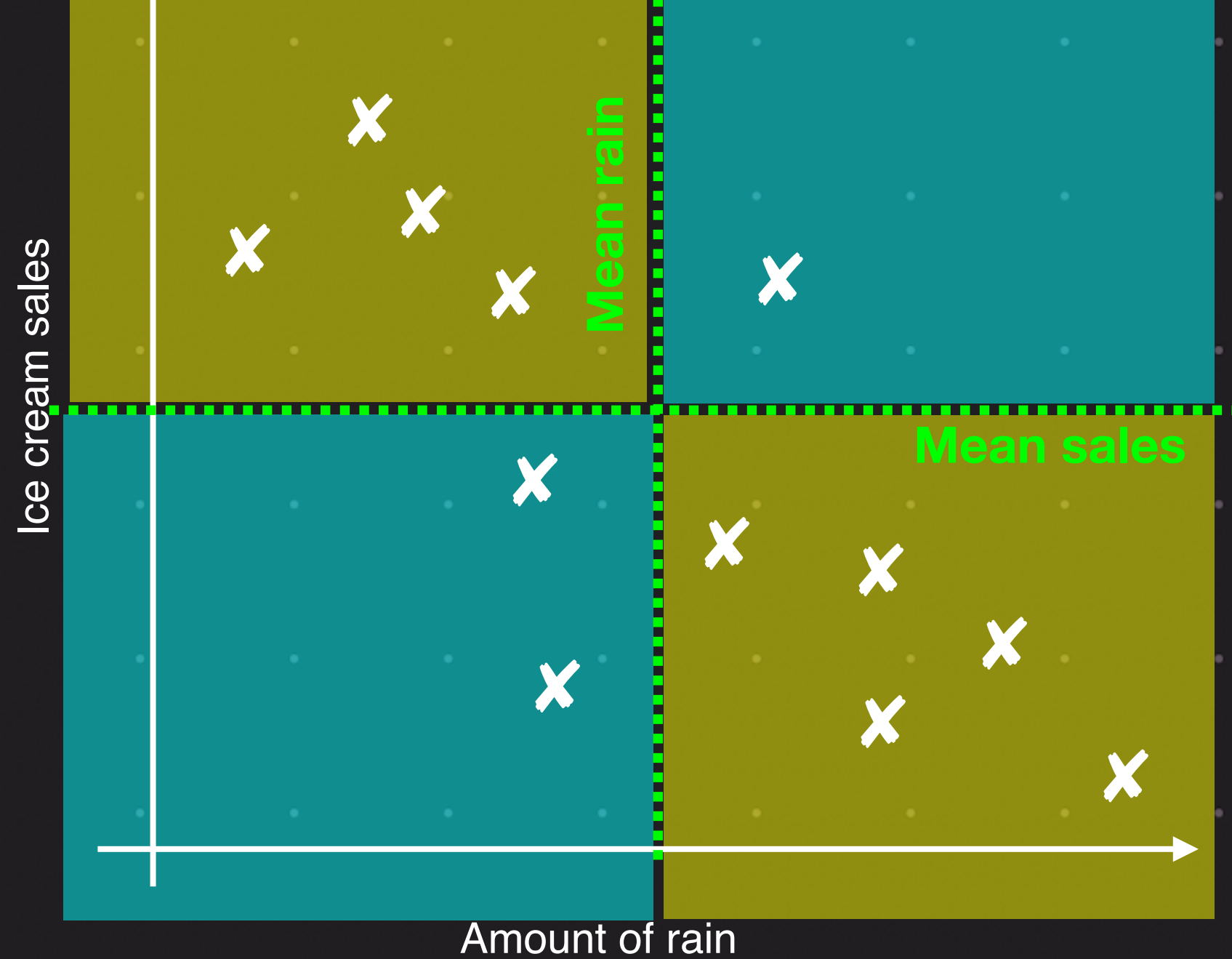
Covariance is the average of all these numbers

$$\frac{1}{10}(4 + 48 + 6 - 2 + \dots + 60)$$

Which has more influence? Positive or negative
Positive has more influence

We say that these two features are positively correlated

Ice cream Vs Rain



Positive correlation

- Top right
- Bottom left

Negative correlation

- Top left
- Bottom right

$$\text{cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Which has more influence? **Positive** or **negative**

Negative has more influence

We say that these two features are positively correlated

Height Vs Rain



Positive correlation

- Top right
- Bottom left

Negative correlation

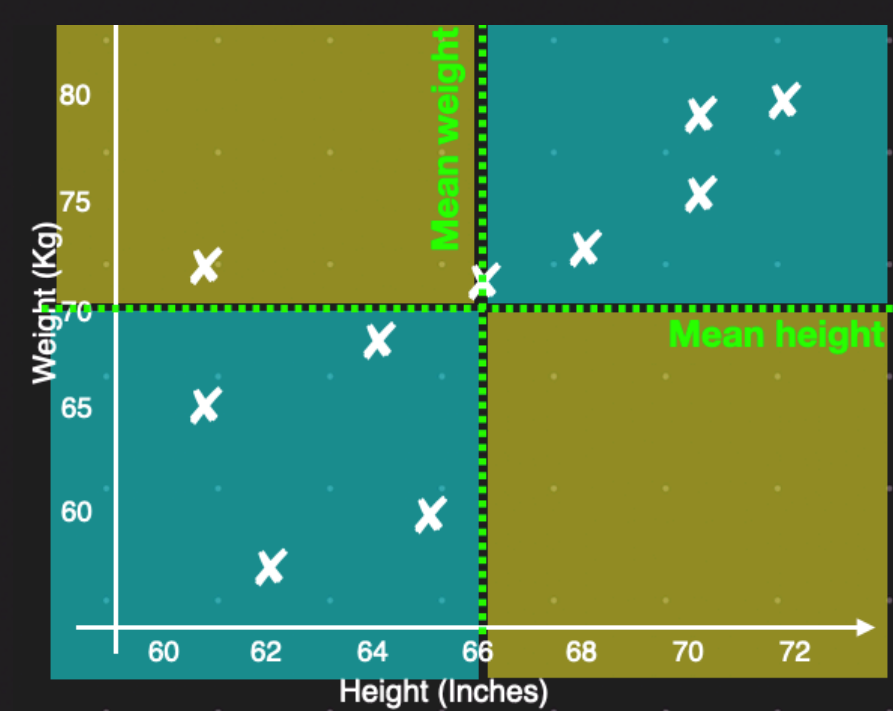
- Top left
- Bottom right

$$\text{cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Which has more influence? **Positive** or **negative**

Both have (approximately) equal influence

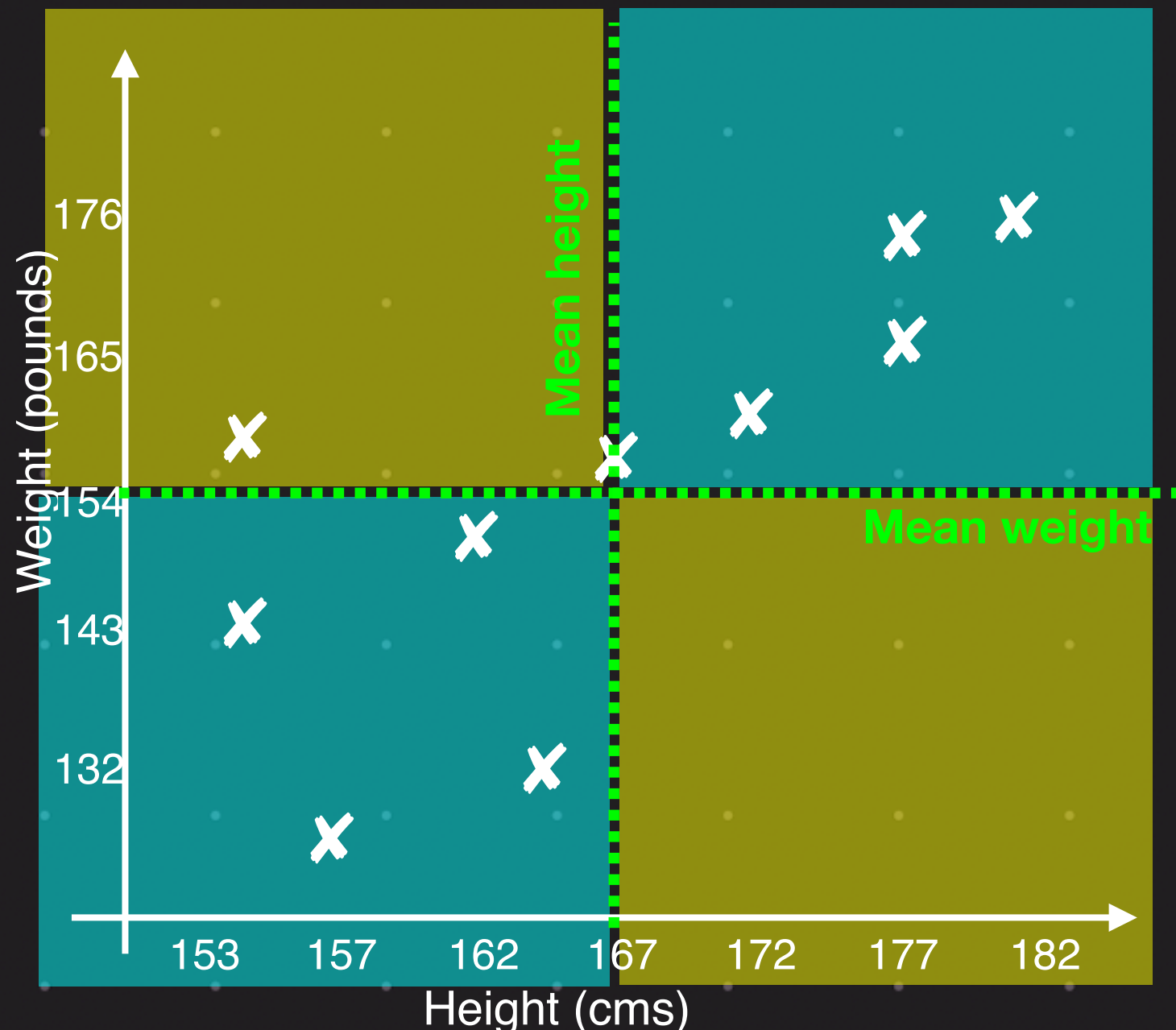
We say that these two features are uncorrelated



Suppose we express height in centimetres and weight in pounds

Simply stretching the axis should not have much influence on how we quantify correlation

The definition of “correlation” does a standardisation of “covariance”



If we apply the formula of correlation, we get the same number whether we use the inch/Kg axis or cms/pounds axis

Positive correlation

- Top right
- Bottom left

Negative correlation

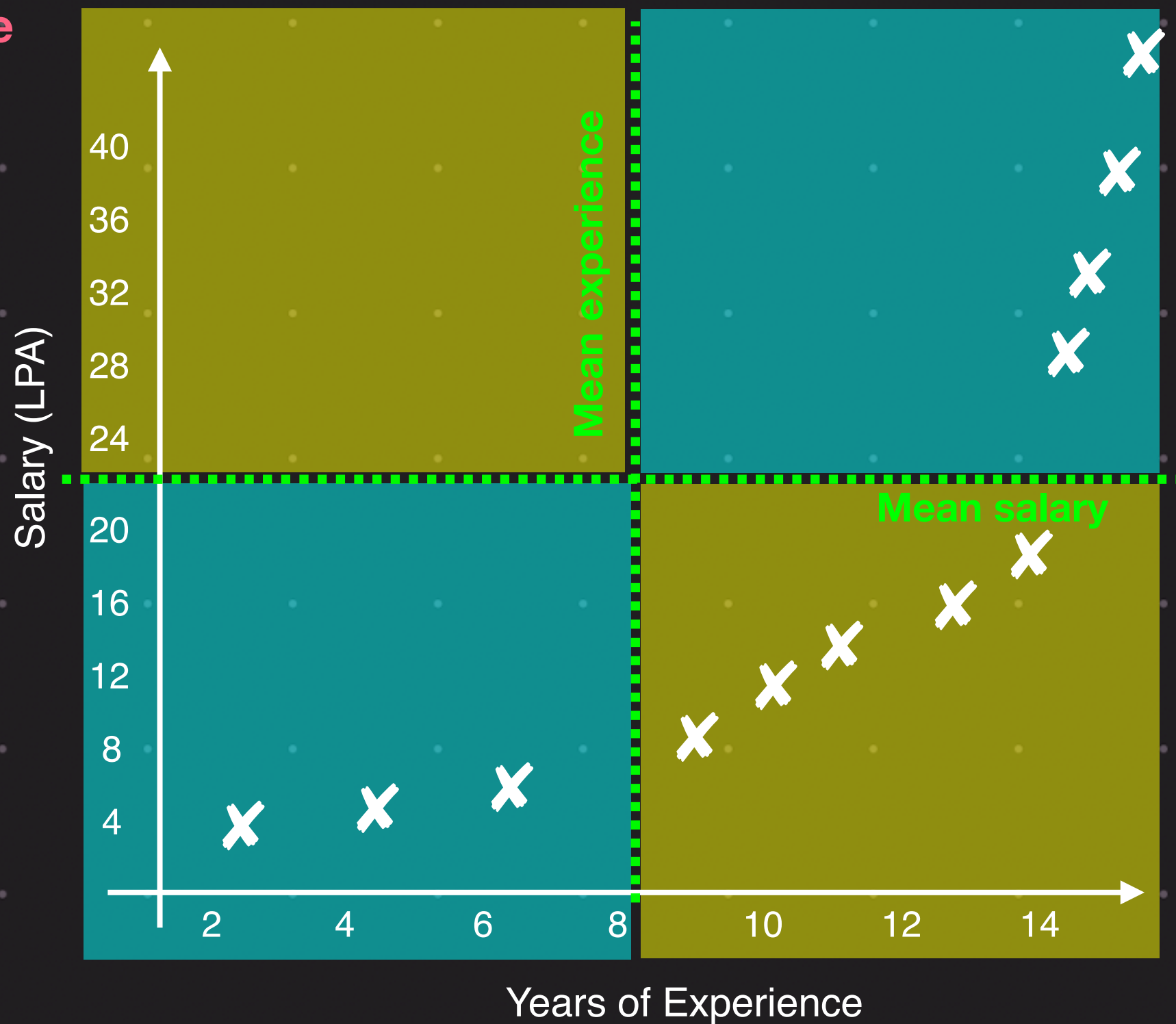
- Top left
- Bottom right

$$\text{cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho_{xy} \leq 1$$

Salary Vs Experience



Positive correlation

- Top right
- Bottom left

Negative correlation

- Top left
- Bottom right

$$\text{cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho_{xy} \leq 1$$



Strange phenomenon: Even though we know that the two features are related, the correlation turns out to be very low

Spearman to the rescue!!!

Rank along both the x and y-axis, then take the correlations of the ranks

Pearson Correlation

$$\text{cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{hw} = \frac{\text{cov}(h, w)}{\sigma_h \sigma_w}$$

Spearman Correlation Pearson correlation of rank(X) and rank(Y)