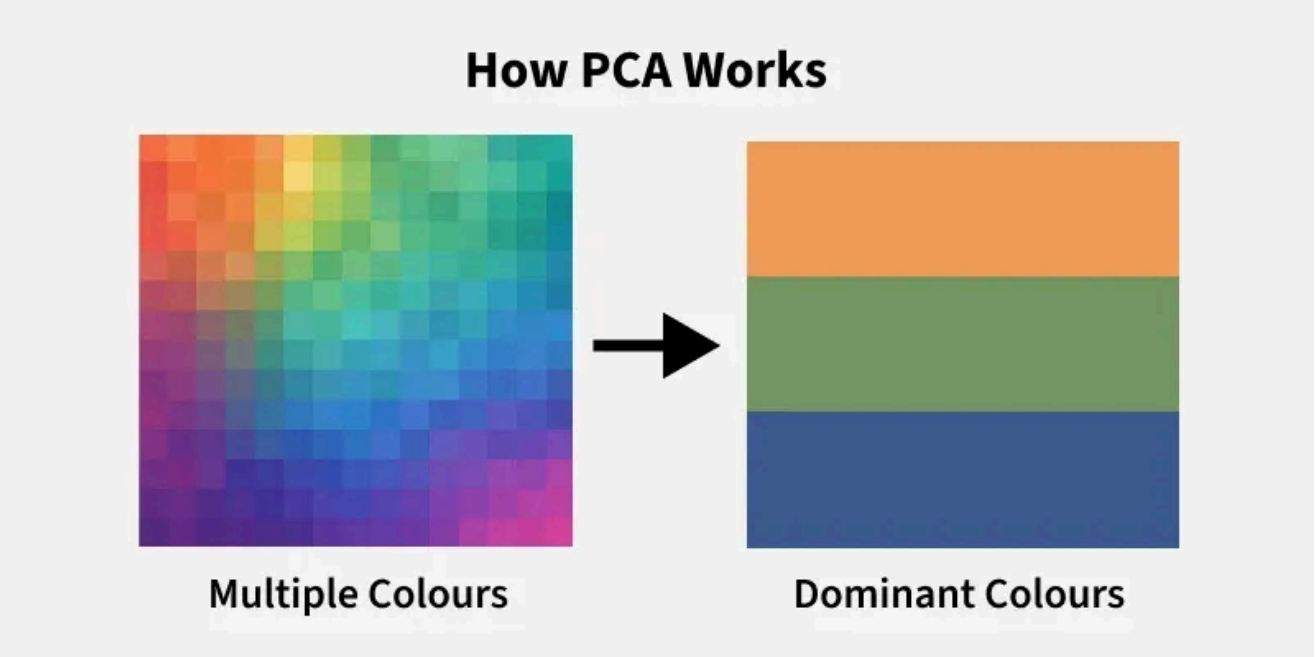
PCA (Principle Component Analysis)

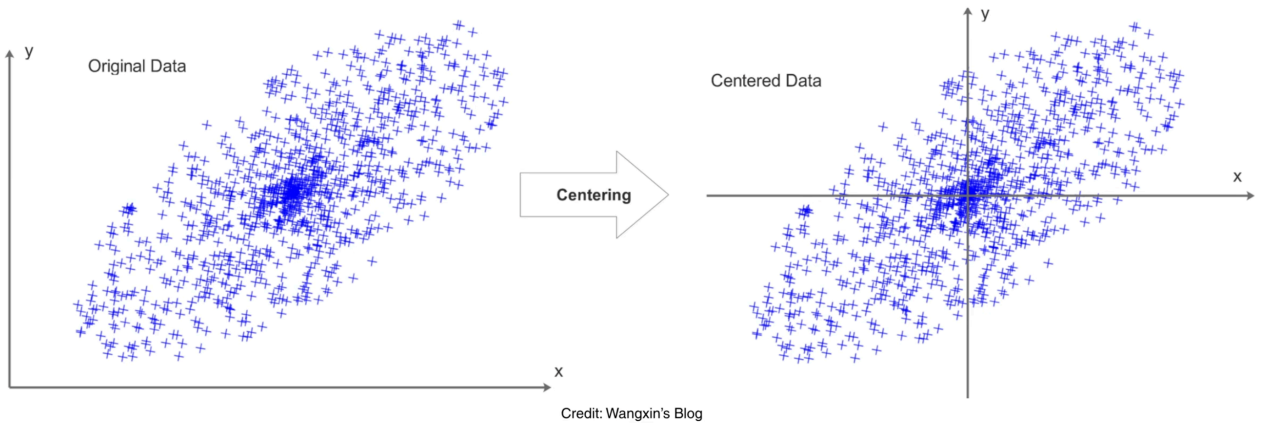
PCA (Principal Component Analysis) is a dimensionality reduction technique and helps us to reduce the number of features in a dataset while keeping the most important information. It changes complex datasets by transforming correlated features into a smaller set of uncorrelated components.



It uses only X not Y, that make it Unsupervised.
PCA finding a new axis for your dataset.
So this become the new features called Principle Component
and each of this component is simply a combination of your previous feature.

Principle Component Requirement:
-Capture as much variance between the data point as possible
-Each principle component is completely uncorrelated with all the other components
PC1: Capture distinct pattern of variation
PC2: Capture distinct pattern of another variation
--> No overlap

Variance = the squared distance of data points from the mean



--> Maximizing sum of squared projection = Maximizing Variance captured along u

Eigenvector = Principle Components
Eigenvalue = Explained variance of component

$$\mathbf{C} = \begin{bmatrix} \frac{8}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \end{bmatrix}$$

$$\lambda_1 \approx 2.54, \quad \lambda_2 \approx 0.45$$

$$\mathbf{u}_1 \approx \begin{bmatrix} 0.93 \\ 0.37 \end{bmatrix}, \quad \mathbf{u}_2 \approx \begin{bmatrix} -0.37 \\ 0.93 \end{bmatrix}$$

u1 = PC1
u2 = PC2

$$\mathbf{u}_1 \perp \mathbf{u}_2 \perp \mathbf{u}_3 \perp \mathbf{u}_4 \perp \ldots$$

as the vector is always symmetric, so the all the eigenvector are orthogonal to each other

$$\mathbf{u}_1 \approx \begin{bmatrix} 0.93 \\ 0.37 \end{bmatrix}$$

PC1 = 0.93 * Feature1 + 0.37 * Feature2

$$\mathbf{u}_2 \approx \begin{bmatrix} -0.37 \\ 0.93 \end{bmatrix}$$

PC2 = -0.37 * Feature1 + 0.93 * Feature2

combinations of the original features

All in all, the PCA essentially boils down to performing the Eigen decomposition of the covariance matrix of the original data.

PCA is not robust to outlier because it relies on the variance, which can be heavily influenced by the extreme values.