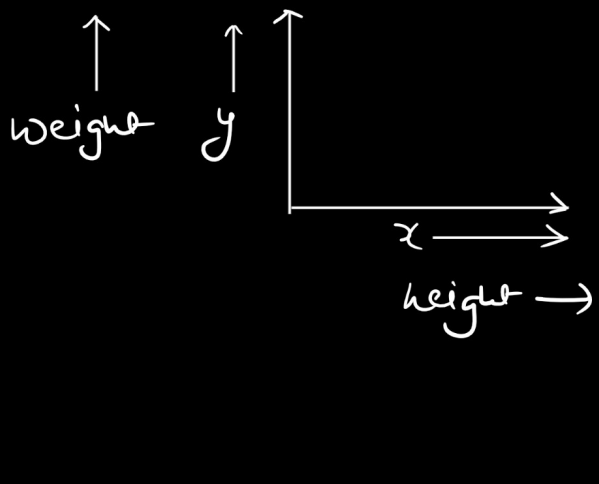
Dimensionality Reduction

1 Dimension → features / Columns



100, 1000, 10000 features / dimensions

① feature selection

20 features → ① feature importance {DT, Random forest}

10 features → highest feature importance value (feature)

② Lasso Regularization technique

③ Correlation heatmap

② Feature Extraction / Projection

or

Dimensionality Reduction

Redesign new features

↳ 10 → 90%
95%
80%.

{ f1, f2, f3, f4, f5, ---------- , f100 }

Computationally expensive

10 feature → selected

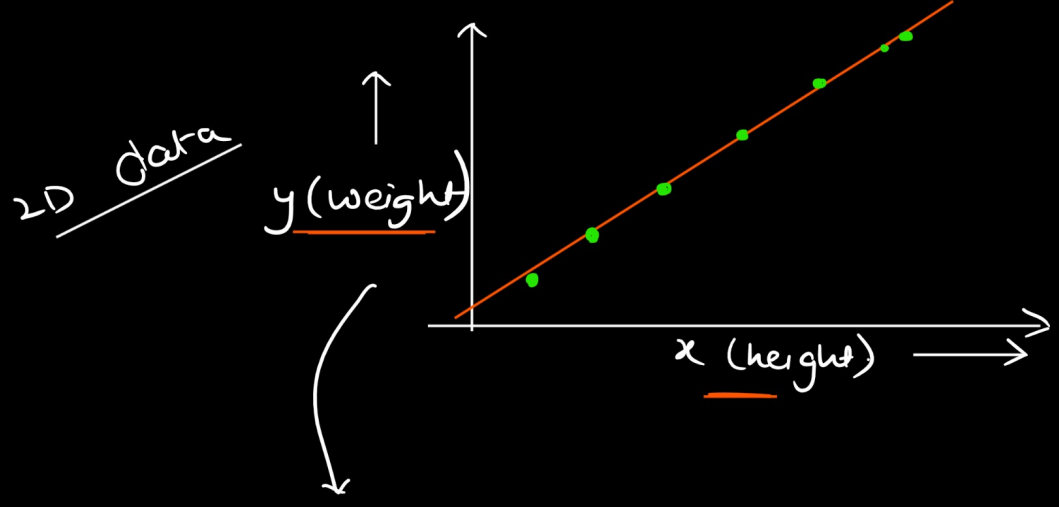
Huge information Loss

EDA → Data Preprocessing

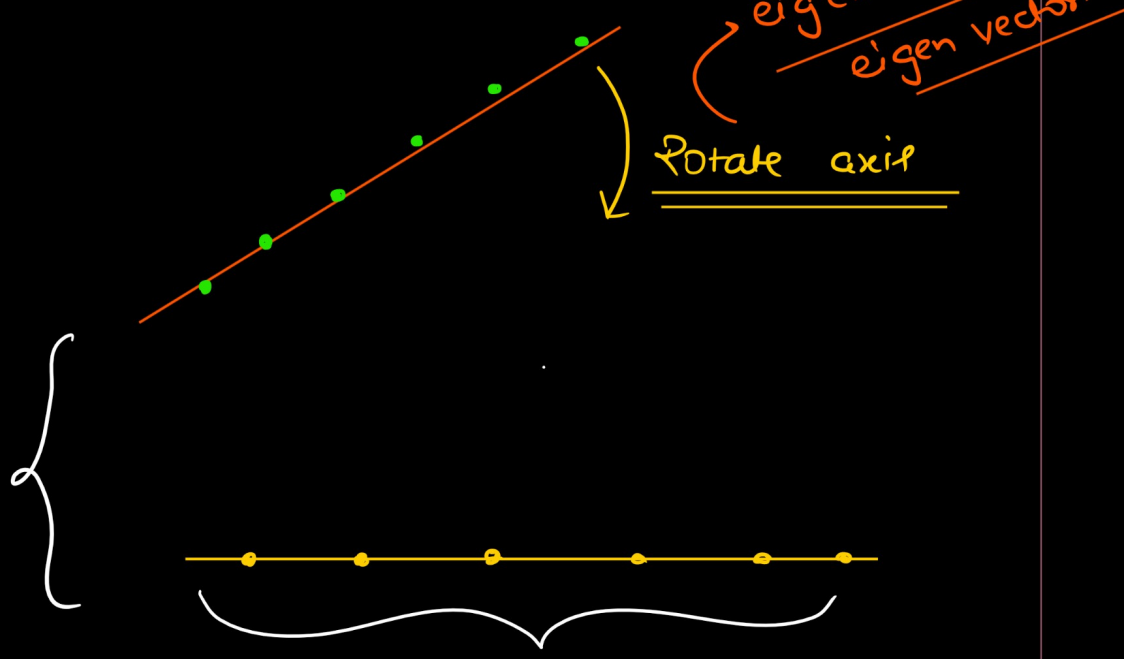
Unsupervised ML Technique

PCA → Principal Component Analysis

Eigen values & Eigen vectors


2D data
y (weight)
x (height)

1D data


eigen values &
eigen vector
Rotate axis

Step-by-step Maths of PCA

to get/extract new features

from the original features

sklearn.preprocessing

1. Standardise the data → to maintain the range of all the features

Age	Standardise → $z = \frac{x_i - \mu}{\sigma}$
23	
25	
47	
48	
72	
64	

2. Covariance matrix ←

A covariance matrix is a square, symmetric matrix in statistics that describes the variance and covariance between different variables in a dataset. The diagonal elements of the matrix contain the variances of the individual variables, indicating their individual spread or variability, while the off-diagonal elements represent the covariances between pairs of variables, showing how they vary together.

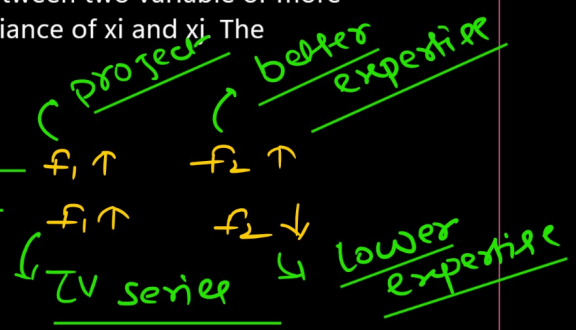
Covariance matrix ⟹ $\frac{\text{matrix}^T \text{ @ matrix}}{\text{num of records} - 1}$

(matrix multiplication)

Covariance provides the measure of strength of correlation between two variable or more set of variables. The covariance matrix element Cij is the covariance of xi and xj. The element Cii is the variance of xi.

-∞ to +∞

If COV(xi, xj) = 0 then variables are uncorrelated
If COV(xi, xj) > 0 then variables positively correlated
If COV(xi, xj) < 0 then variables negatively correlated



difference b/w correlation coefficient vs Covariance ??

Covariance	Correlation
Covariance is a measure of how much two random variables vary together	Correlation is a statistical measure that indicates how strongly two variables are related.
Involves the relationship between two variables or data sets	Involves the relationship between multiple variables as well
Lie between -infinity and +infinity	Lie between -1 and +1
Measure of correlation	Scaled version of covariance
Provides direction of relationship	Provides direction and strength of relationship
Dependent on scale of variable	Independent on scale of variable
Have dimensions	Dimensionless

They key difference is that Covariance shows the direction of the relationship between variables, while

(3) Eigen Values & Eigen vectors using Covariance matrix

2 features → 2 eigen values

↓

2 eigen vectors

Power of information

④ To evaluate the new principal components

PC1 > PC2 > PC3 _ _ _ ——

2 features ⟶ 2 Principal components

Transform your data

(dataframe @ eigen vector)

Properties of Principal components

① Each principal component is uncorrelated orthogonal to all others.

② PC1 captures the most variance/information, Subsequent components will have lesser

Real time examples of higher dimensional data

1. Healthcare data
2. NLP (Natural Language Processing ↳ Textual data)
3. Image & video data
4. financial data