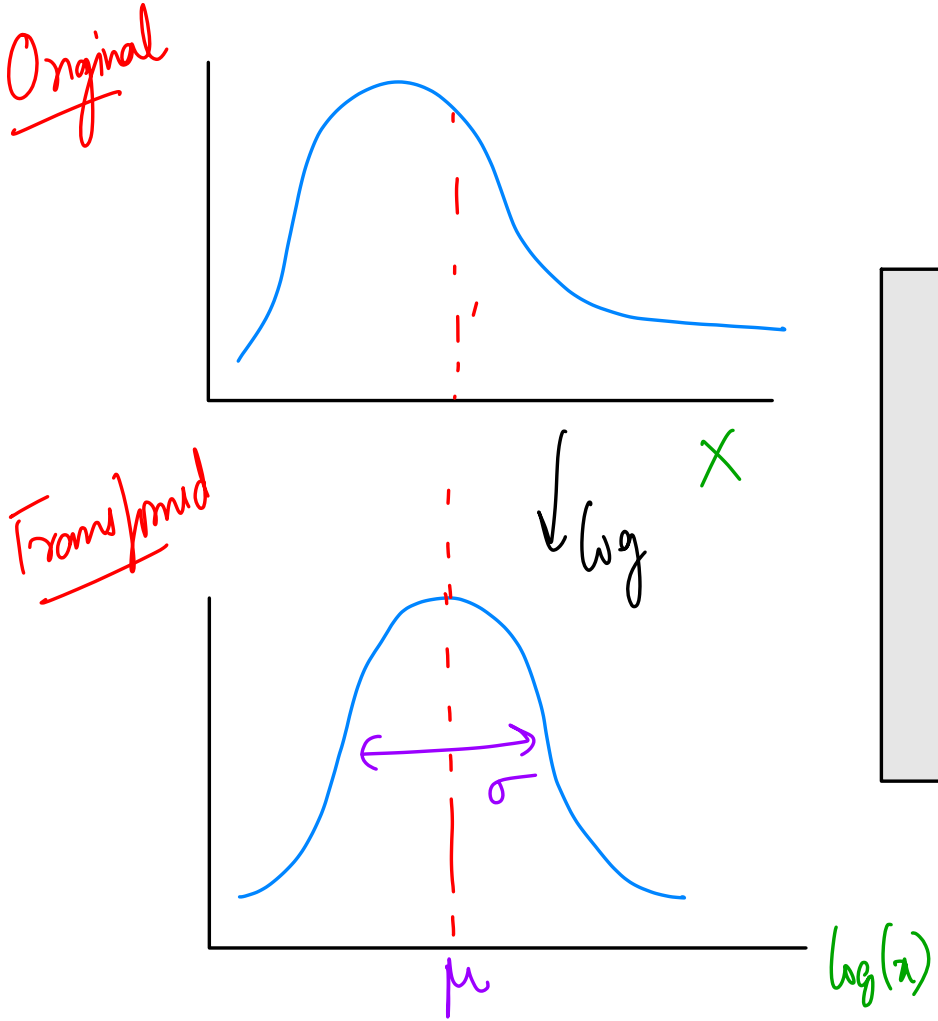
FEATURE ENGINEERING - 1

LOG NORMAL DISTRIBUTED



$$E(X) = \mu_X = e^{\left(\mu + \frac{\sigma^2}{2}\right)}$$

$$Var(X) = \left(e^{\sigma^2} - 1\right)\left(e^{2\mu + \sigma^2}\right)$$

Feature Engineering

Aerofit

Features / attributes

Target

Datapoint / Records

	Age	Income	Fitness	Miles	Gender	Education	Product
X_1							KP281
X_2							KP481

input → Model → prediction

Classification ←

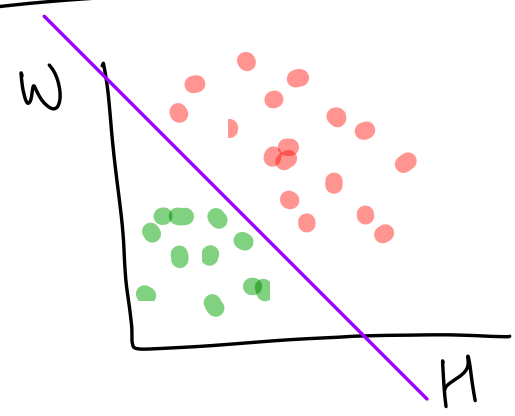
Regression

Clustering

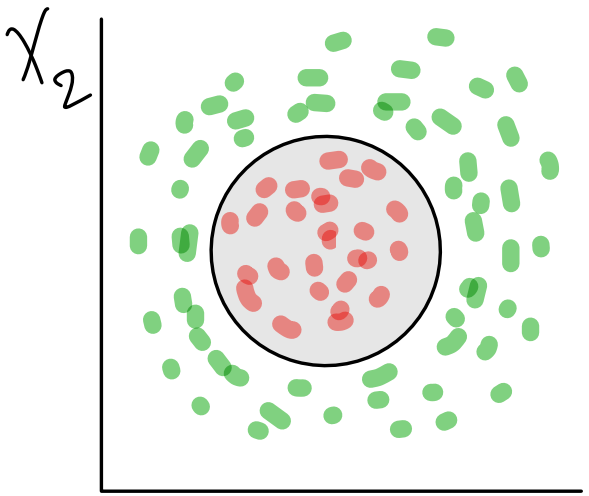
mathematical function

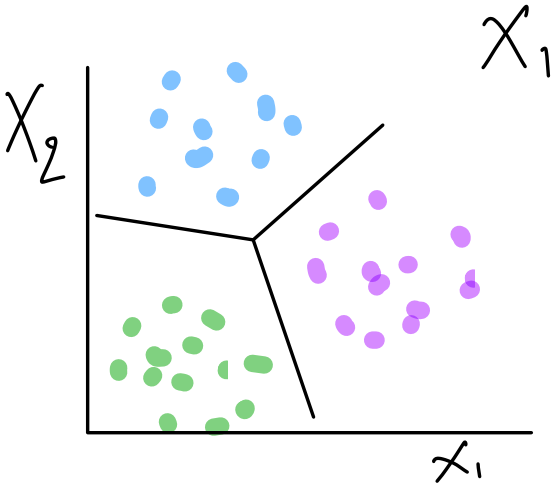
$y = mx + c$

Complex complicated



$$x^2 + y^2 + 2gx + 2fy + c = 0$$


X_2
X_1


X_2
X_1

Height	Weights	BMI $\frac{H}{W^2}$	Fitness

* Z test → μ, σ ✓ (Numerical v/s Categorical)

* T test (ttest_1samp, ttest_rel, ttest_ind, ttest_ind_formstats)
  σ ✗ → $n<30$, (Numerical v/s Categorical)

* KS Test → $\mu_1 \simeq \mu_2$, distributions are diff.

* χ^2 → Goodness of fit (CATEGORICAL)
  → Test of Independance (CAT v/s CAT)

* ANOVA → More than 2 grps (Numerical)

* KRUSKAL → when assump^n of ANOVA FAIL

* CORRELATION → (NUMERICAL v/s NUMERICAL)