Statistical Test

You work for a company that produces three types of energy drinks. The company wants to know if the average caffeine content differs significantly among these three types. You collect data on the caffeine content (in milligrams) from five samples of each drink:

Energy Drink A: 105 mg, 110 mg, 98 mg, 107 mg, 103 mg

Energy Drink B: 130 mg, 125 mg, 132 mg, 128 mg, 129 mg

Energy Drink C: 120 mg, 115 mg, 117 mg, 119 mg, 121 mg

Question: Is there a statistically significant difference in the average caffeine content between these three types of energy drinks?

group_A= [105, 110, 98, 107, 103]
group_B = [130, 125, 132, 128, 129]
group_C = [120, 115, 117, 119, 121]

$H_0 = \mu_A = \mu_B = \mu_C$

H_1 = Atleast one of the energy drinks have significant difference in the Caffeine Content

mean_A = 104.6
mean_B = 128.8
mean_C = 118.4

mean_ = 117.266

ssb_A = 802.13 → $5*(104.6-117.266)^2$
ssb_B = 665.16 → $5*(128.8-117.266)^2$
ssb_C = 6.42 → $5*(118.4-117.266)^2$

SSB= 1473.71

ssw_A = 81.2 → $(105-104.6)^2+(110-104.6)^2+(98-104.6)^2+(107-104.6)^2+(103-104.6)^2$
ssw_B = 26.8
ssw_C = 23.2

ssw = 131.2

MSB = 401.065 → $df_B = 2$
MSW = 10.933 → $df_W = 12$

F_value = 36.68 → MSB / MSW

P < 0.0001 → 0.0001

Alpha = 0.05
P < alpha Reject the hypothesis → atleast one of the energy drink have significant difference in the caffeine content

→ `sm.stats.anova_lm(model, typ=2)`

ANOVA → more than two mutual-fund scheme

T-test → compare the mean of two mutual fund schemes

Chi-Square Test → Categorical data

↓

Right-skewed distribution

2010 → weights of individuals into a small city & below are the numbers.

C1	C2	C3
< 50 kg	50 - 75 kg	> 75 kg
20%	30%	50%

Population data

2020 → $n = 500$ Sample size

$n = 500$

Expected data

	C1	C2	C3	< 50kg	50 - 75kg	> 75kg
	< 50 kg	50 - 75 kg	> 75 kg			
Observed data	140	160	200	100 (20% of 500)	150 (30% of 500)	250 (50% of 500)

Population diff of weights has changed in the last 10 years → True Statement

H_0 → observed = expected

H_1 → observed ≠ expected

$df = 3 - 1 = 2$

$$\chi^2 = \frac{\Sigma (\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$\chi^2 = 26.66$$

p_value = 0.0001

$\alpha = 0.05$

p_value $< \alpha$

↳ Reject H_0

↓

observed ≠ Expected in the last 10 years

Pencil Company → Product preference is evenly distributed among different age groups or not?

survey

Observed data

Product	Under 30	30-50	Over 50	Total
Electronics	50	30	20	100
Clothing	30	40	30	100
Home Goods	20	30	50	100
Total	100	100	100	300

500

Expected data

Product	Under 30	30-50	>50
Electronic	100/3	100/3	100/3
Clothing	100/3	100/3	100/3
	500/3	500/3	500/3
Home Goods	100/3	100/3	100/3

34, 33.33, 34

$$\chi^2 \text{ value} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$df = (r-1)(c-1)$$

r: # products, c: # categories

$$df = (3-1)(3-1)$$

$$df = 4$$

p-value α

A/B Testing

↳ Click-through rate or Conversion rate

C1	C2
Older	Newer

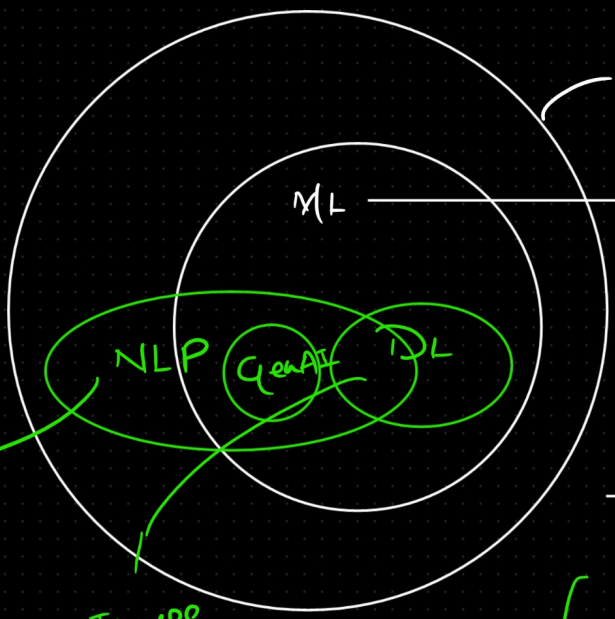
Fraud Detection

C1	C2
Corporate fraud	Retail fraud

Healthcare → Association b/w patient Characteristics & health outcomes

Self driving Car

Machine Learning (Reinforcement Learning)



AI → Artificial Intelligence

ML — Machine Learning

Based on historical data

Train the model

Without any human intervention

models give the result

Textual Data

Image data

Numeric data

Banking Domain

① Supervised ML Models

(target label available)

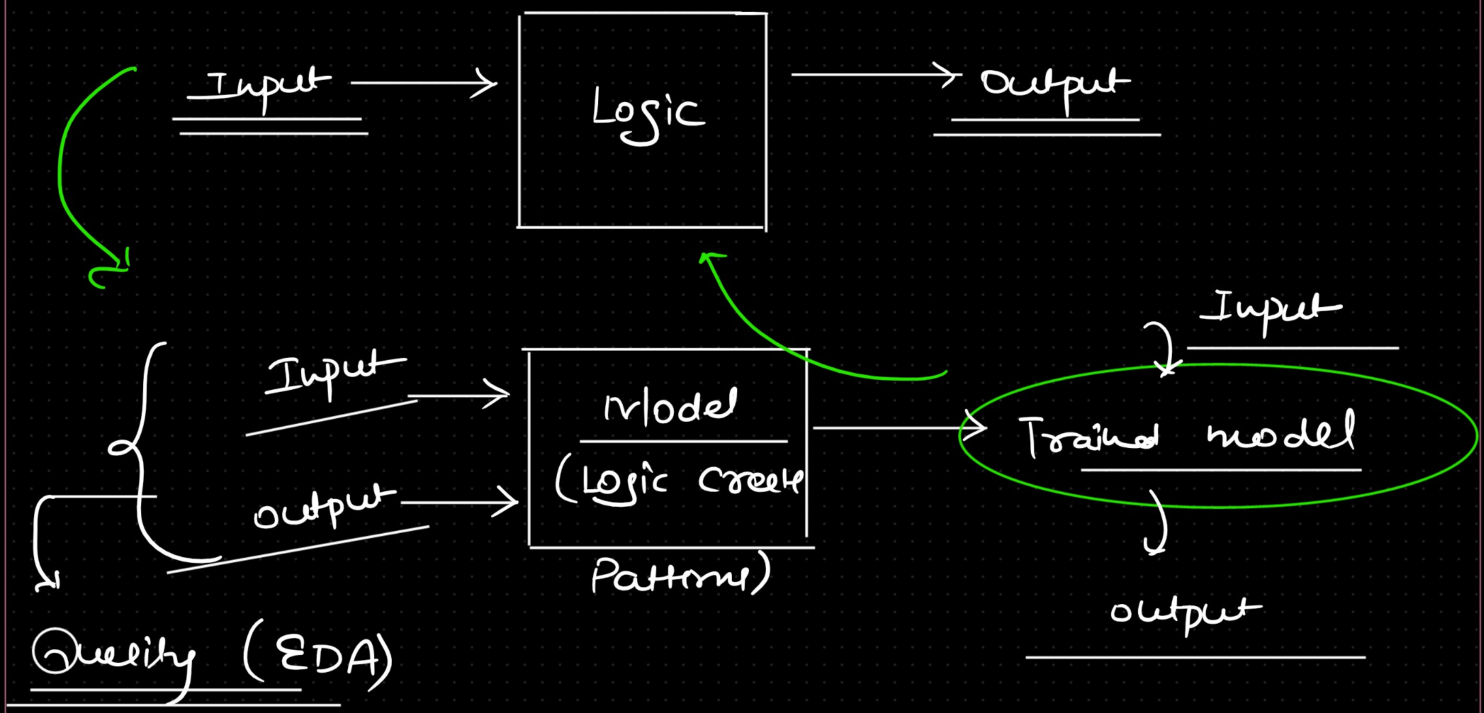
② Unsupervised ML Models

(target label unavailable)

* Mathematical Intuition

Models

Implementation


Input
Logic
Output
Input
output
Model
(Logic create
Patterns)
Trained model
Input
output
Quality (EDA)