20-09-2025

Agenda:

EDA-I → movie sentiment } Textual
EDA-II → movie sentiment - II
→ wine quality } numerical
→ commodity data } date, numerical, missing value

EDA

Descriptive

Inferential

- can we claim this effect exist in the larger population, or is it just chance?

Tools:

→ Hypothesis testing
→ confidence interval

H0 : 1746 word | 1748 word

H1 : 1746 | 40

```
def (no_of_word):
    if no_of_w > x:
        Happy
    else:
        sad
```

EDA

→ grasp data / understand data
→ data looks / visualization
→ finding trends / pattern
→ To clean data
→ correlation
→ skewness
→ distribution

data → EDA → FE → ML

prob: → Todo list

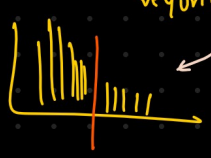
prob: separate young vs old people

data: age, income

① data → EDA → FE → ML

② data → EDA → think if ML required

form:



if p-age > x:
old
else
young

Hypo: must:
difference between -ve & +ve review length -

DS: Hypo → length

FE → ML →

Year

1 (DS) 2 → product

2 (DS) ↑ 2

3 → (DS) other

→ proceed with solving the prob

→ domain

4th → 1 → Real

→ 2 → ML

DS → 100%.

→

Python: (DS)

DS → Stats

→ CV ← master → libs

Agentic

→

→ CV ← research → bad

→ CV ← tve

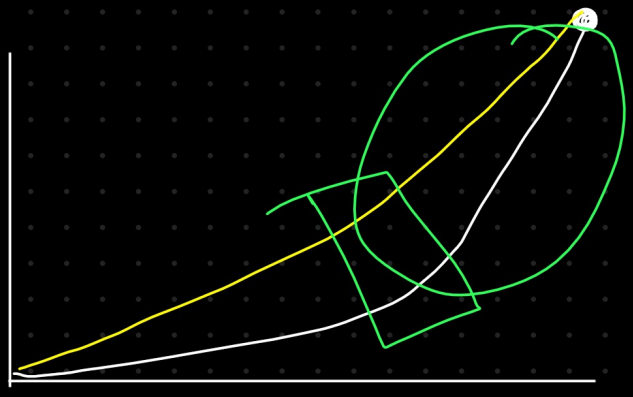
→ CV ←

→

Tech : CS → job
5%

:

Non-Tech : spec, ↑ → 30%



(X) (y)

$x_1 m_1 \quad x_2 m_2 \quad x_3 m_3$

↓

m_L

↓

Time-series data:

Data that contains date time

time-series → analysis

↳ regression → predict

↳ forecast

Mon	Tue	Wed	Thur	Friday
—	—	—	—	

Summary of ffill

ffill → forward fill

Data:

10
Nan
15
Nan
20

$\frac{10+15+20+40+50}{5}$

$\Rightarrow 27$

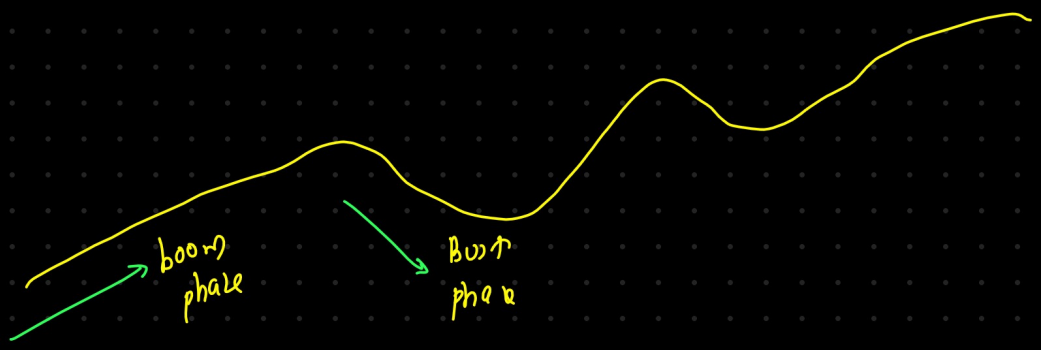
Time series

10
15
20
40
50

ffill → missing values would be replaced with the value row above it.

bfill → missing values would be replaced with the value below it.

Nan		Nan		Nan → 10
10	→ ffill	10	→ bfill →	10
20		20		20
30		30		30
40		40		40



pct_change

2025-01-01	100	NaN
2025-01-02	110	0.10
2025-01-03	105	-0.045
2025-01-04	126	0.20

→

(105 − 110) / 110

Goal → [.10, -1.5, -0.045, .20]

↓ std

std → mean → speed

A → 30 40 50 60 70

→ 10

B → 0 20 50 60 100

std → ↑