flow to perform EDA process

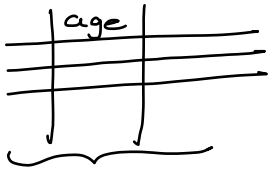
Why we perform EDA ? → Solve a business problem

Clear ✓ Business understanding : Correlate the insights from the data with your business understanding

Credit card : Data

↓

{ ← [age ≥ 18]

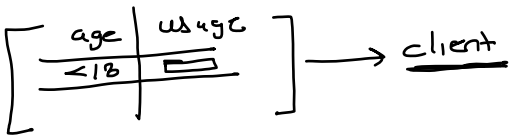


20,000

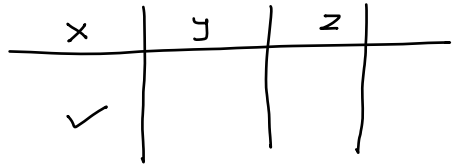
↓

[age < 18] ←

↓



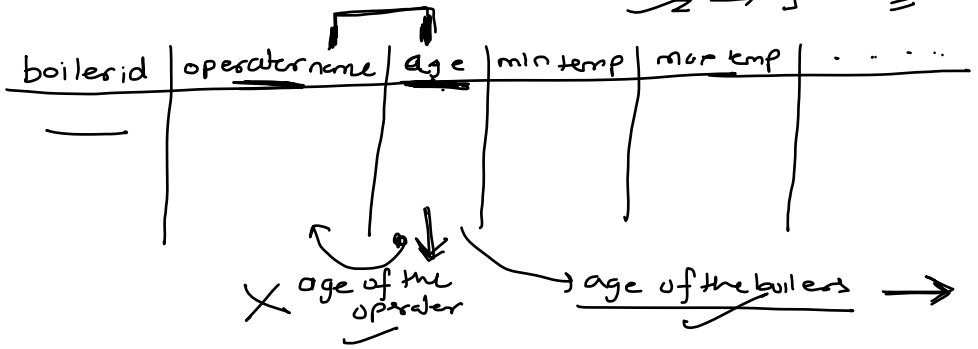
ii) Data understanding :



+ Data dictionary

x →
y →
z →
] very Imp

oil & gas



- Data cleaning → Check for duplicated rows
  - → Data type correction
  - → category check (check the list of unique categories)

Complete duplicated row : keep one drop other

Duplicates in a certain coloumn → Investigate → There will always be a reason behind it → according take next step

c_1	c_2	c_3	c_4	c_5
a_1	b_1	c_1	d_1	e_1
a_1	b_1	c_1	d_1	e_1

→ duplicated rows

Ecomm

city
- New Delhi
- Delhi
- Deli
- South Delhi
- NCR

→ (New Delhi)

Python → Pandas

- → Integer → int
- → decimal → float
- → String / text / alphanumeric → object
- → Datetime → Datetime

c.id o.id p.id v.id

c_1	c_2	c_3	c_4
a_1	b_1	c_1	d_1
a_1	b_2	c_2	d_2

→ Here the data is duplicated because there are two orders placed by one customer

customer count	order date
1.0	01/01/2011
2.0	01/02/2011
3.0	
7.0	
9.0	

→ object → Pandas Datetime format

int

→ Since this is count (wrong data type)

$\frac{50+29+28+21}{4}$ $\frac{29+25+28+21}{4}$ =

avg(Sal) = 32k
avg(age) = 2575

Age	Gender	Sal
29	M	50k
25	F	NA
28	M	29k
NA	M	28k
21	F	21k

25.75

32

Garbage value
NOISE

Analysis ✓

actual insights ?

Wrong
Biased
Not True

- objective → EDA → we don't perform missing value imputation

- objective → Machine Learning modelling → various ML-Algo. (They can't operate on a data with missing values)

Necessity to perform missing value imputation ✓

df.duplicated() ⟶

	c_1	c_2	c_3	c_4	
0	a_1	b_1	c_1	d_1	⟶
1	a_1	b_1	c_1	d_1	⟶
2	a_2	b_2	c_2	d_2	
3	a_1	b_1	c_1	d_1	⟶

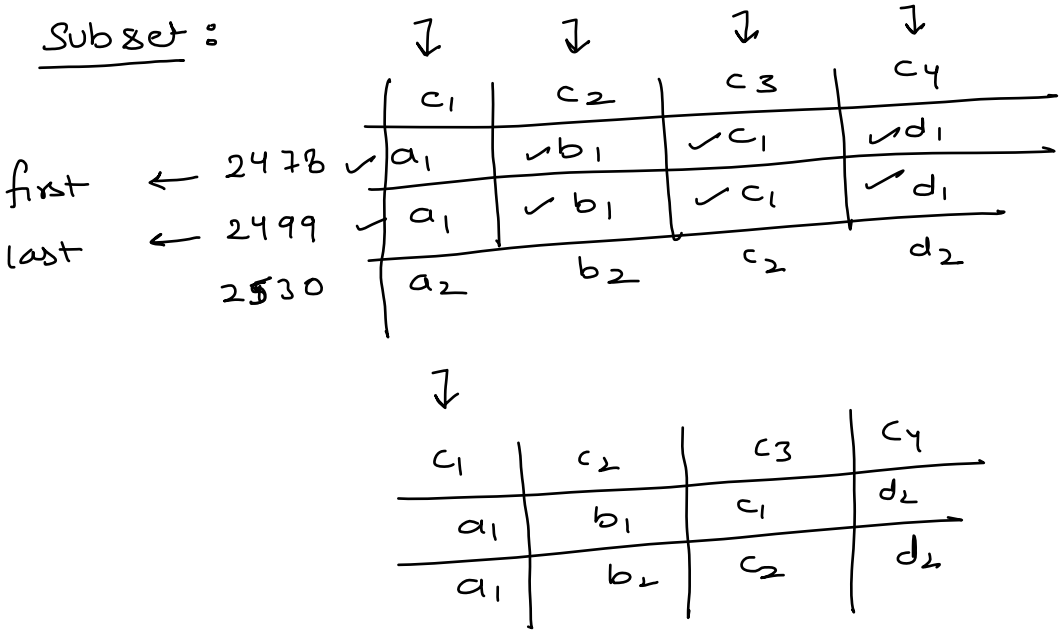
0	False ⟶
1	True ⟶
2	False
3	True ⟶

0
1
0
1
⟶ (2)
= Count of True

True ⟶ 1
False ⟶ 0

df.drop_duplicates()

Subset :

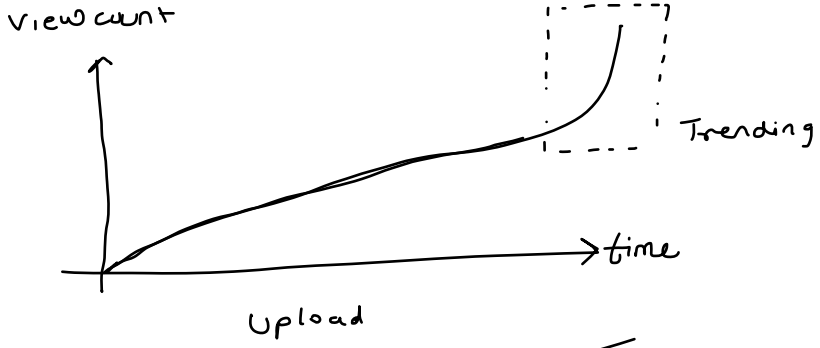


complete duplicated rows

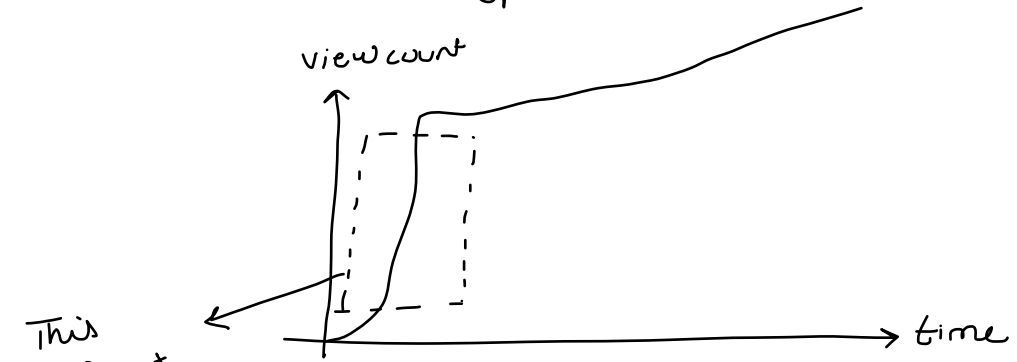
We are looking at the data by considering all the columns ?

Subset → column name on which you want to check for duplicated entries

by default it will check all the columns (complete duplicated rows)



Upload



This significant increase in the no. of viewership makes your vide become trending

Live stream