* Data Preprocessing - Interview Question

completed ① Handling missing value - Measure of central Tendency

Encoding concept

② Handling | Char | object | variable ⟸ Label Encoder, One-hot-encoder, dummy variable

③ Handling outlier - Transformation approach

log, Sqrt, cuberoot etc.

Removal / Trimmer

Capping (very popular)

IQR, Z-Score

④ Feature scaling ⟸ Normalization, Standardization → by default

⑤ Imbalance Treatment → This is only applicable with Classification problem

(2) Encoding concept — variable - Char/Object

Label encoder
"Tree based model"

One-hot-encoder + dummy variable
"Linear family"
KNN

Maths/Algorithm — Number

COVID-19 — "Location" — Object

Significant → Inferential Stats
Hypothesis Testing

Feature importance

$p \leq 0.05$
CL $\geq 95\%$

Location	Label Encoder
Delhi	3
Mumbai	5
Pune	6
Bangalore	0
Chennai	2
Gujrat	4
U.P.	7
Bihar	1

Dummy variable

$= n - 1$

↓

(Total No. of column - 1)

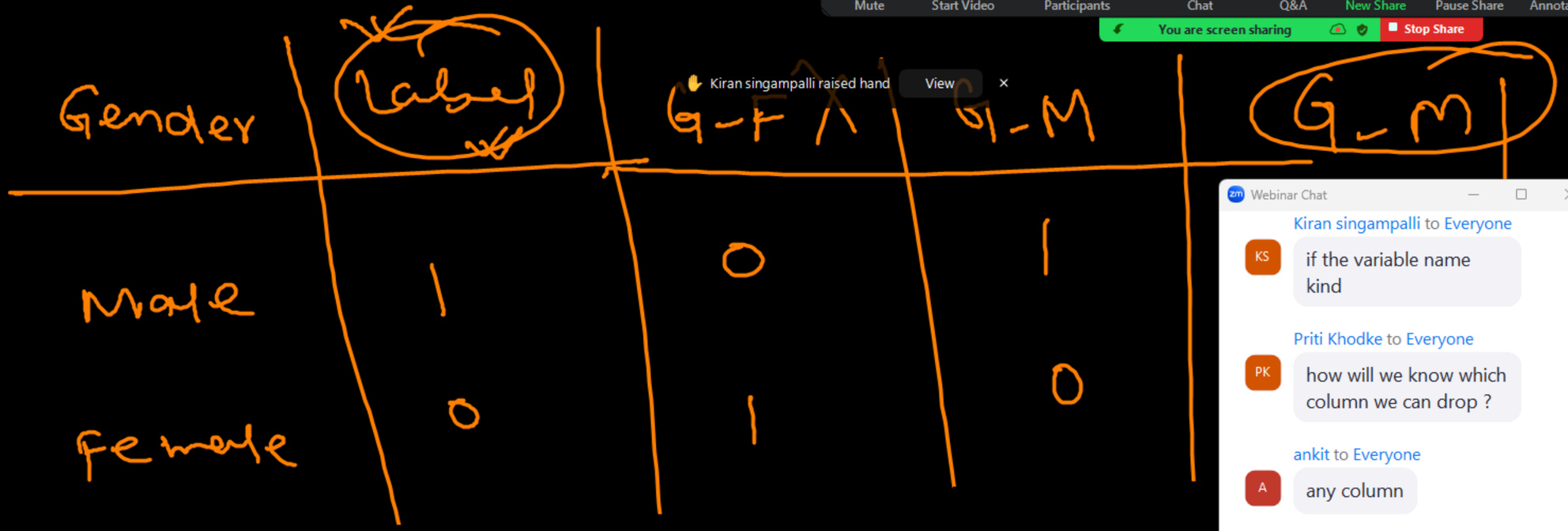
benefit :- "Multicorrelearity"

Linear Regression

One-hot-encoder

Dummy variable

Gender	Label Encoder	G_Female	G_Male	G_Trans	G_M	G_T
Male	1	0	1	0	1	0
Female	0	1	0	0	0	0
Female	0	1	0	0	0	0
Male	1	0	1	0	1	0
Transgender	2	0	0	1	0	1
M	1	0	1	0	1	0
F	0	1	0	0	0	0
T	2	0	0	1	0	1



Webinar Chat

Kiran singampalli to Everyone

KS: if the variable name kind

Priti Khodke to Everyone

PK: how will we know which column we can drop ?

ankit to Everyone

A: any column

milind pawar to Everyone

MP: whenever we have multiple columns then we can go for one hot encoding

Who can see your messages? Recording On

To: Hosts and panelists

Type message here...

Feature Scaling

Age	salary
20	50K - 50000
25	60K
30	80K
35	110K
45	150K
55	200K
60	250K - 250,000

Linear Regression (DV dependent)

$$Y = m_1 x_1 + m_2 x_2 + C$$

Dep slope slope Intercept

$$Y = m_1 * Age + m_2 * Salary$$

m_1 & m_2 = 1

$$Y = Age + Salary * 2$$

*2

Feature Scaling

where variable is known (Range)

Normalization

- Always give the value between 0 and 1
- NO outliers

$$Nor = \frac{x_i - min\_value}{max\_value - min\_val}$$

where variable is unknown (Range)

Standardization

- give both the +ve & -ve value between -3 and +3
- there is an outlier

$$Std = \frac{x_i - mean(\mu)}{std(\sigma)}$$