DAV-3

HYPOTHESIS TESTING

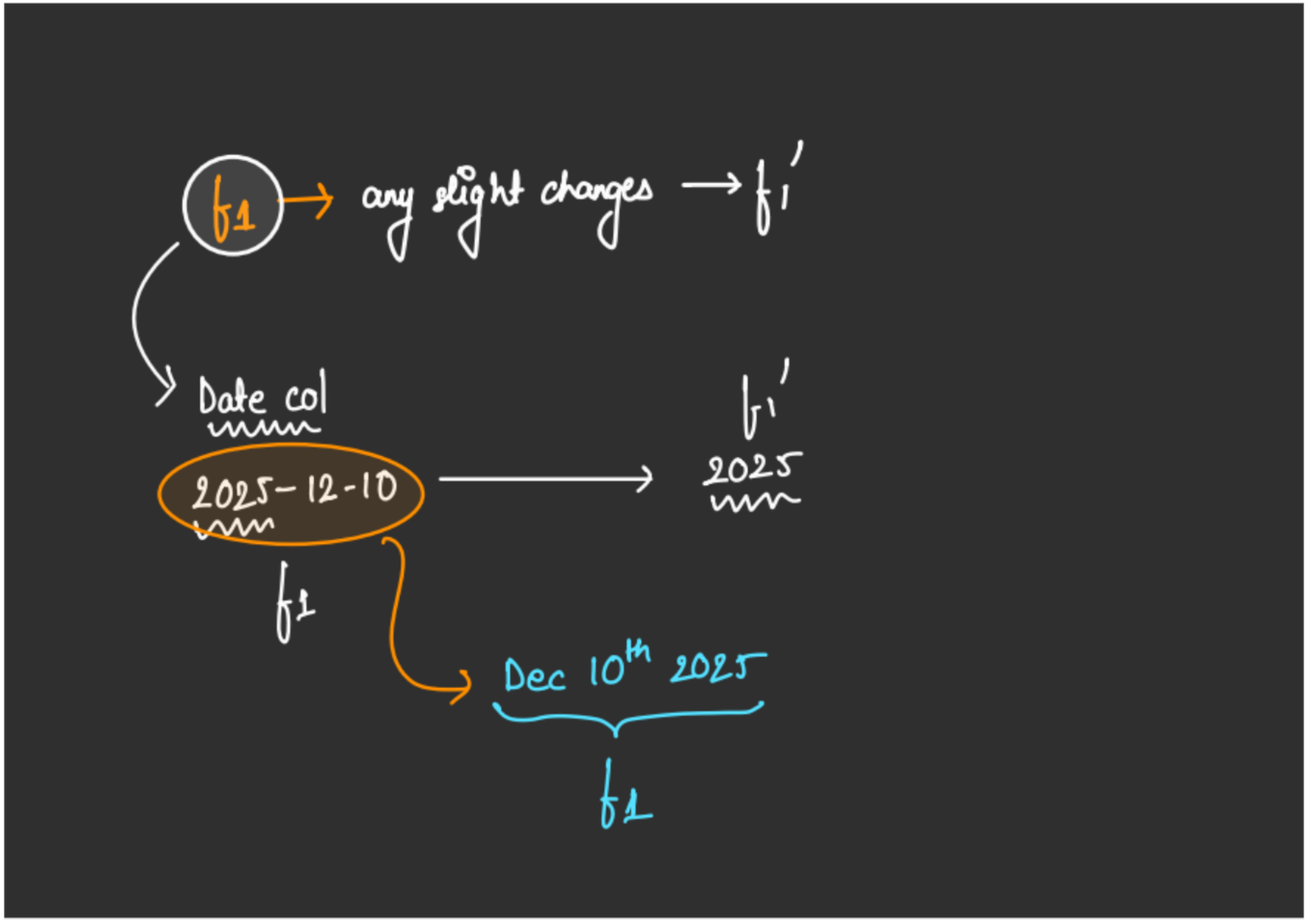
(Class starts @ 9:10 PM)

Lecture 10: Feature Engineering 2

#Agenda

1. Creating New Features
2. Missing Values
3. Outlier Treatment
4. Categorical Encoding & Types → TE left
5. Normalization & Standardization

↓ Next Class


f1
any slight changes
f1'
Date col
2025-12-10
f1
f1'
2025
Dec 10th 2025
f1

f1 wrt Target ⟶ univariate

f1, f2 wrt Target ⟶ Bivariate

Missing Values

① Identifying Null Values

→ { isnull().sum()
  { isna().sum()

② Why to fix NULL Values ?

→ Data loss

→ ML model don't understand Null values

C_1	C_2
1	x
2	y
Null	z
3	a
4	b

③ How to fix them ?

→ Statistical way → Mean, Median Mode, Max, Min

→ Library Based

Column

Numerical

- Mean
- Median
- Mode
- Constant Value

Categorical

- Mode
- Constant Value

Num & has outliers → Median

Num & has almost No outliers → Mean

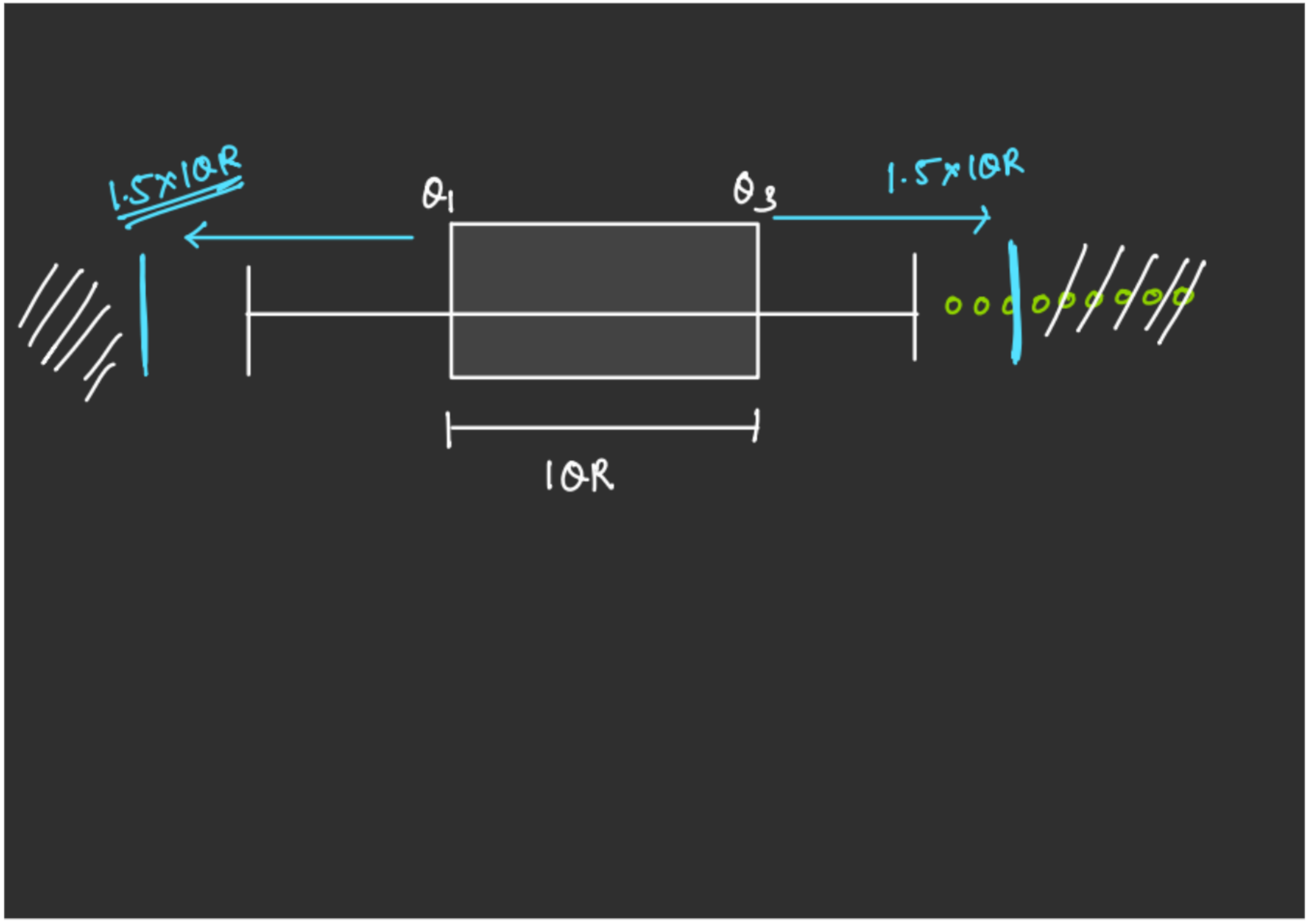
Cat → Mode
→ Const Value

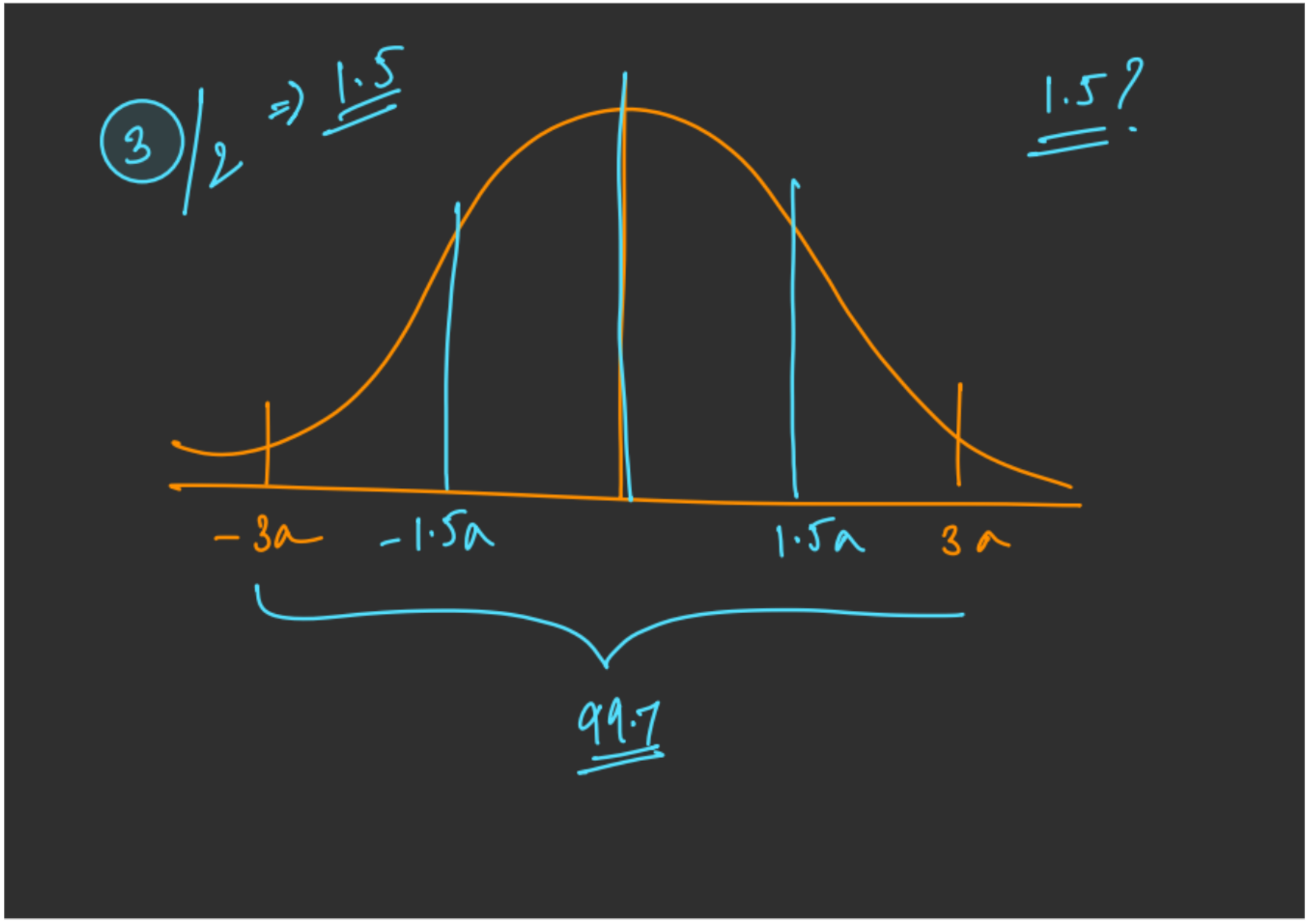
Num & you want to fill with your own value → Const Value

#Outlier Treatment

Q) How to Identify outlier? (Check the distribution → Boxplot
→ Histogram)

Q) What are some ways the remove outliers? (IQR, Z Score)


1.5×IQR
Q1
Q3
1.5×IQR
IQR


3 1/2 => 1.5
1.5?
-3a
-1.5a
1.5a
3a
99.7

#Categorical Encoding

- What is categorical Encoding ? [Convert Cat to Num]
- Why do we need this? [ML model don't understand Cat Variable]
- What are the Types ? [There are 3 Types]

Types of Encoding

1. One hot Encoding (OHE)
2. Label Encoding
3. Target Encoding.

① One Hot Encoding (OHE)

Gender	M	F
Male	1	0
female	0	1
Male	1	0
female	0	1
female	0	1

When cat=2

OHE is good

When cat = 5

Sparse data

Majority of your data has zeros

Travel	Low	Medium	High
Low	1	0	0
Medium	0	1	0
High	0	0	1

2 Label Encoding

Gender	New col
Male	1
Female	0
Male	1
Female	0

Categorical → Animal

Cat	dog	elephant	Rat	Fox
1	2	3	5	4

Nominal data

We use **Label Encoding** when we have **ordinal** data

- Ordinal → Data is ordered

Nominal Data

- No Order

Example

- High → 3
- Medium → 2
- Low → 1

③ Target Encoding

→ When cat = 2 then use OHE

→ When cat > 2 ⟶ Label Encoding (Ordinal data)

→ When cat > 2 ⟶ Target Encoding (Nominal data)

→ Loan Status (The column you are predicting)

Name	Target
Akash	1
Max	0
John	1
Max	1
John	1

⇒

Akash → 1 → 1/1 ⇒ 1 ✓

Max → 2 → (0+1)/2 ⇒ 0.5 ✓

John → 2 → (1+1)/2 ⇒ 2/2 = 1

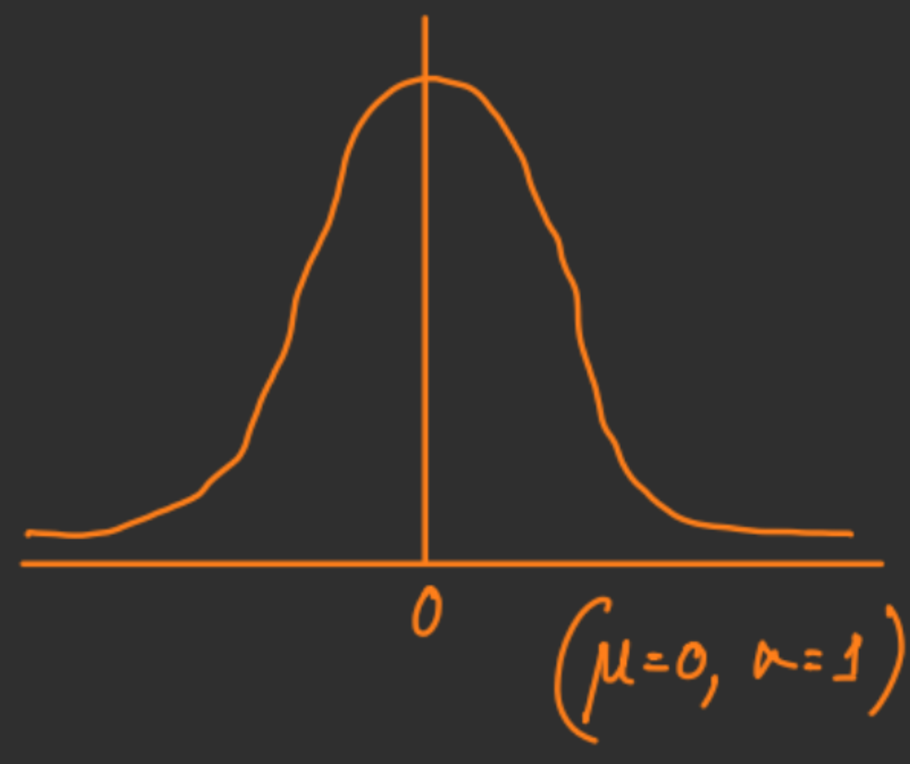
#Column Normalization & Standardization

1. Why do we need this? (To bring every feature into same scale)
2. What are Types?
   - Normalization & Standarization

#Standardization

"Standard Scaler" = $\left(\frac{x_i - x.mean()}{Std(x)} \right)$

→ Standardization library



#Normalization

"Min Max Scaler" = $\left(\frac{x - x_{min}}{x_{max} - x_{min}} \right)$

→ Normalization library

Range $\in [0, 1]$

Normalization and standardization are both used to rescale data, but they serve different purposes and are best suited for different situations:

1. **Normalization** (scaling to a range, typically 0 to 1) is ideal when:
   - You know the data has no extreme outliers, and values are already bounded or have a specific range.-
   - You're working with algorithms that assume data within a fixed scale (e.g., neural networks or k-nearest neighbors).

2. **Standardization** (scaling to a mean of 0 and standard deviation of 1) is ideal when:
   - Your data has a normal (bell-shaped) distribution or contains outliers.
   - You're using algorithms that assume data has a Gaussian distribution (e.g., linear regression or principal component analysis).

In short, normalize when you want all values on the same scale without outliers affecting the range, and standardize when you want to center and scale data, especially when the data might have extreme values.