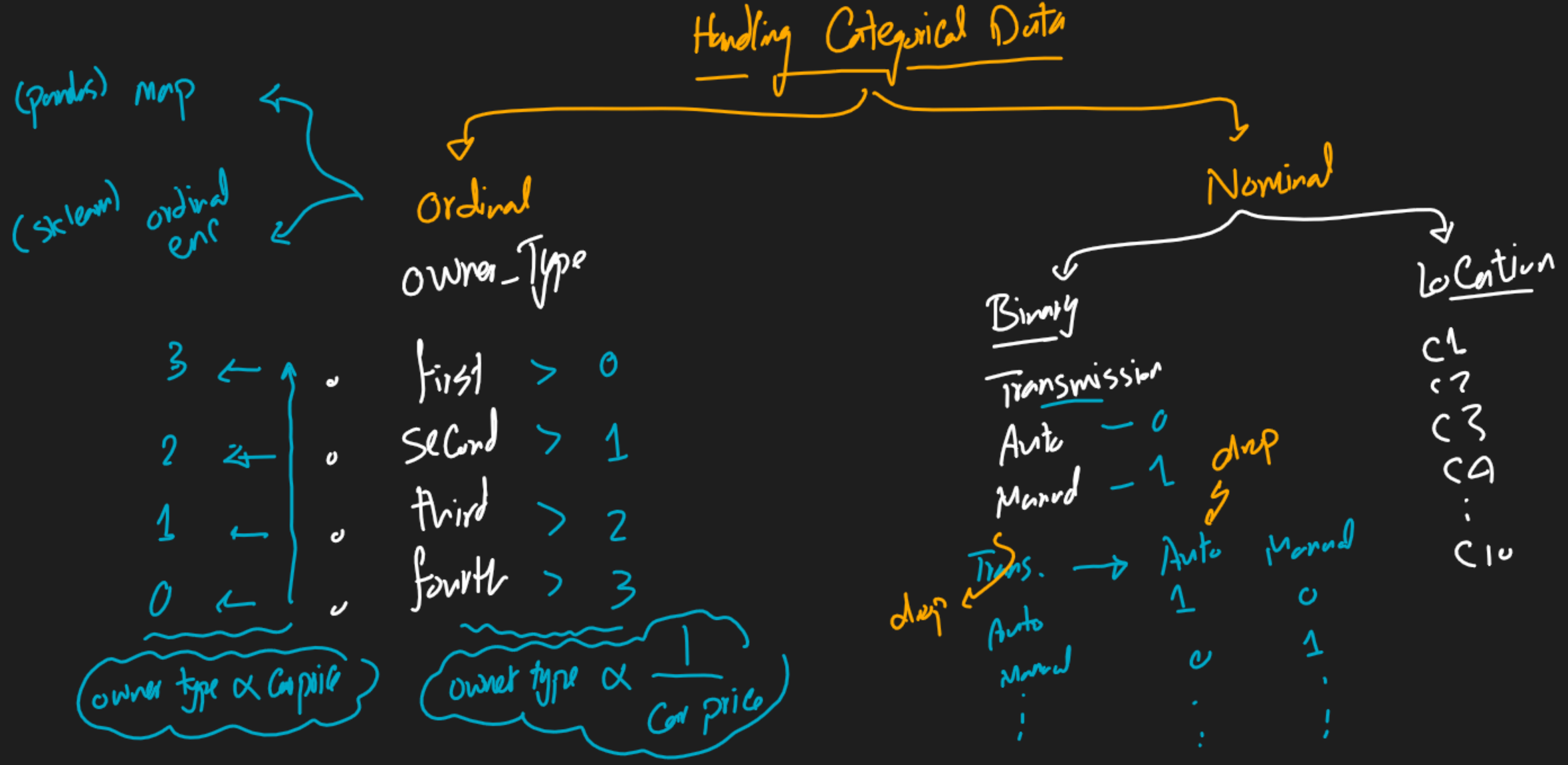
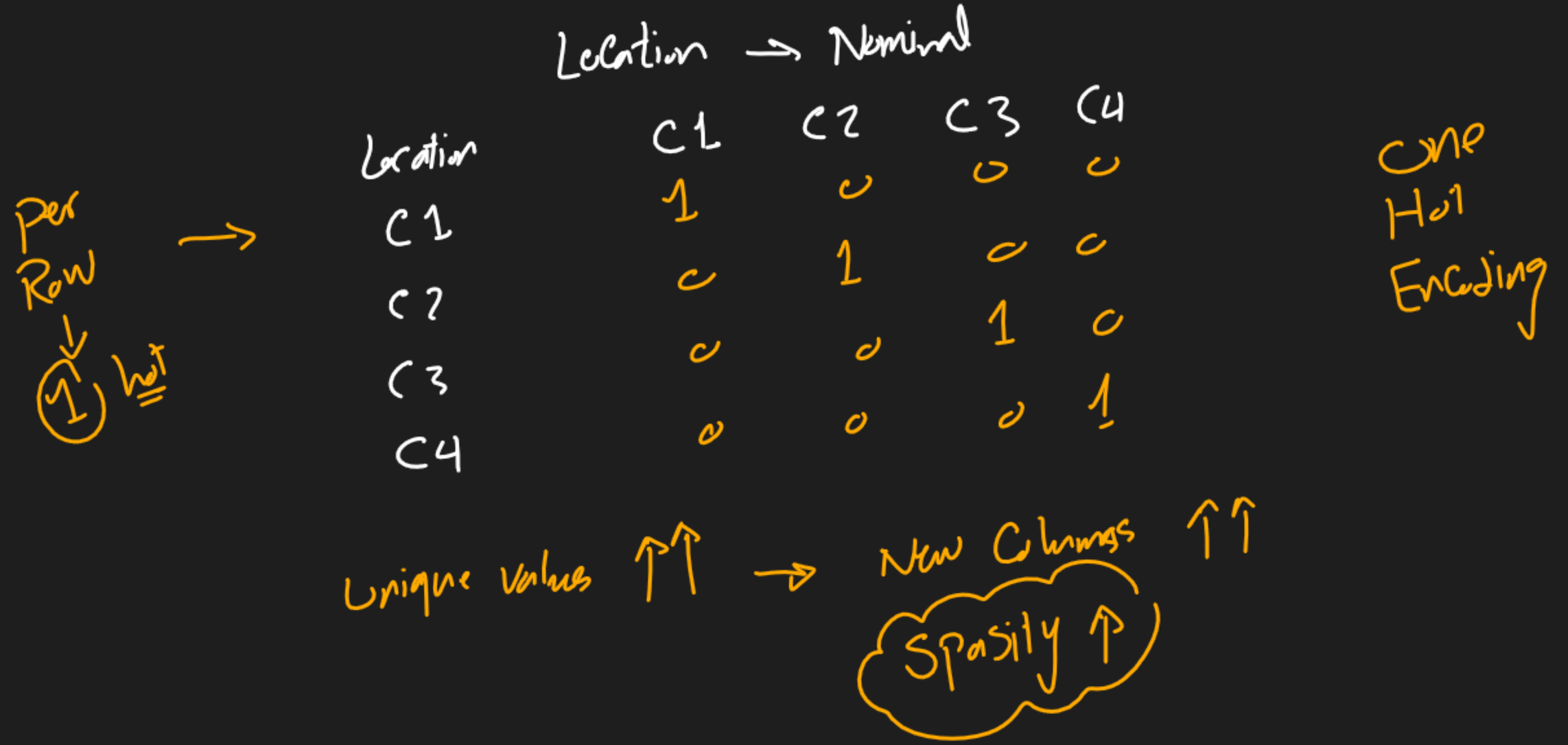
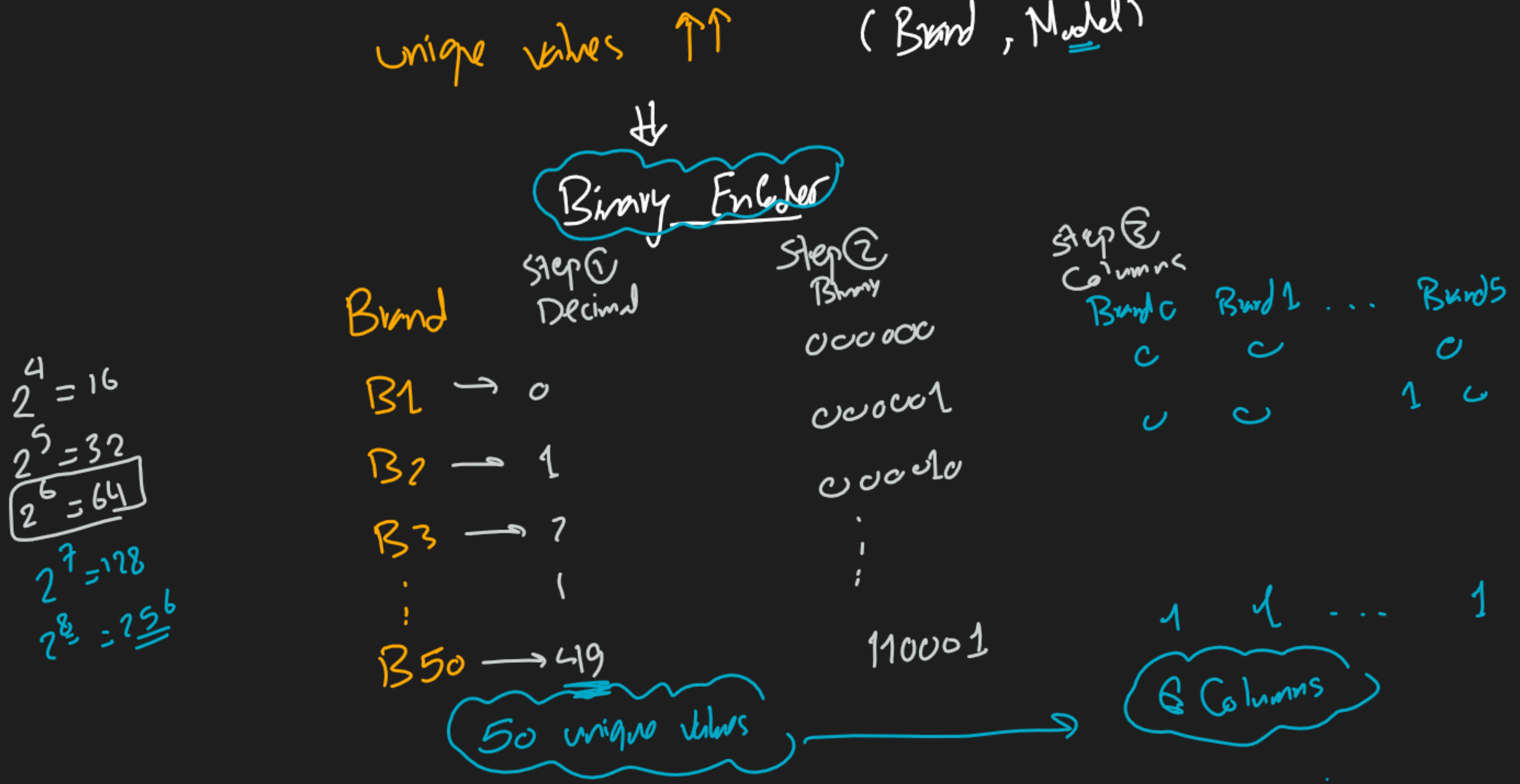

Handling Categorical Data
(pandas) map
(sklearn) ordinal enc
Ordinal
Owner_Type
first > 0
second > 1
third > 2
fourth > 3
3
2
1
0
owner type ∝ Car price
owner type ∝ 1 / Car price
Nominal
Binary
Transmission
Auto – 0
Manual – 1
drop
Trans.
Auto
Manual
drop
Auto
Manual
1
0
0
1
Location
c1
c2
c3
c9
c10


Location → Nominal
Per Row → 1 Hot
Location
C1
C2
C3
C4
C1 C2 C3 C4
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1
One Hot Encoding
Unique values ↑↑ → New Columns ↑↑
Sparsity ↑


unique values ↑↑
(Brand, Model)
Binary Encoder
Brand
Step ① Decimal
Step ② Binary
Step ③ Columns
B1 → 0
B2 → 1
B3 → 2
B50 → 49
50 unique values
110001
6 Columns
Brand 0
Brand 1
Brand 5
$2^4 = 16$
$2^5 = 32$
$2^6 = 64$
$2^7 = 128$
$2^8 = 256$

Scaling

f1	f2	f3
1	1000	0.7
↓	↓	↓
10	5000	08

different scales

Standardization

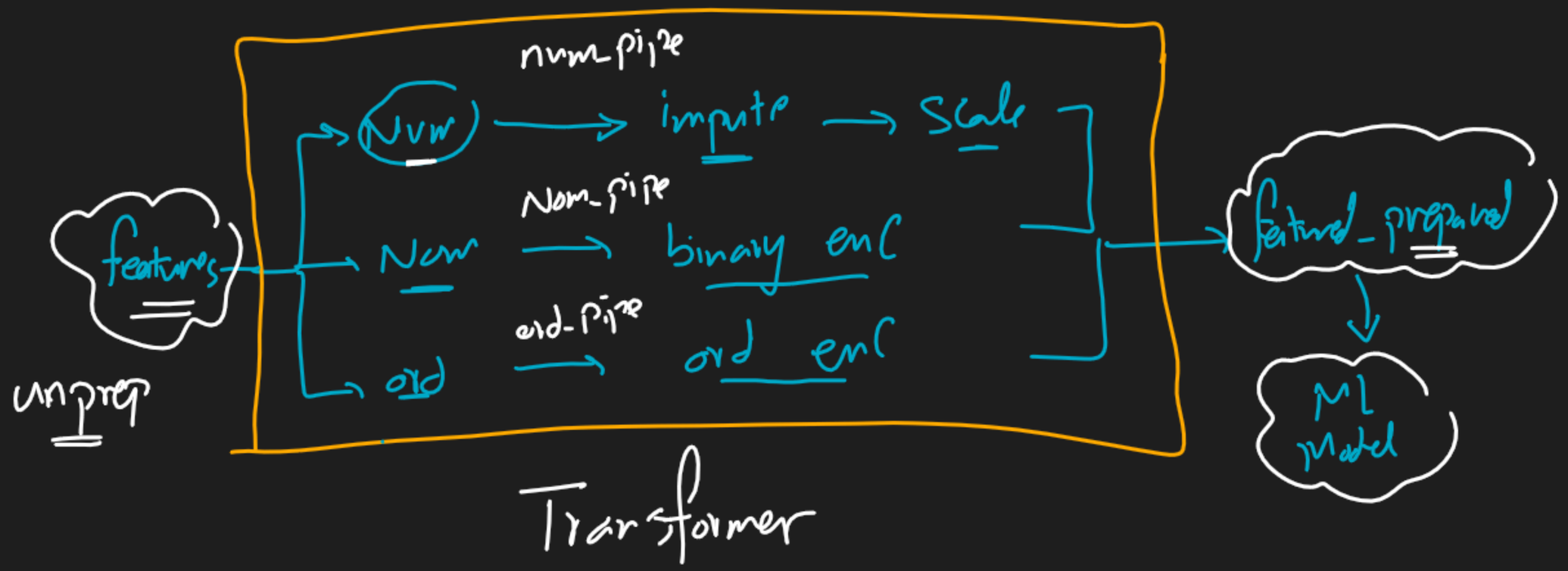
Standard Scale → $\frac{x - x1}{\sigma}$

Mean 0
Std 1

Normalization

MinMaxScaler → $\frac{x - x_{min}}{x_{max} - x_{min}}$

Min 0
Max 1


num_pipe
num
impute
scale
features
Nom_pipe
Nom
binary enc
ord_pipe
ord
ord enc
unprep
featured_prepared
ML
model
Transformer

Preprocessing

① Calculation → fit → train

use

② Conversion → transform → train, test

fillna("median")