Standardization (Z-score scaling)

Formula: $z = (x - \mu)/\sigma$

Let's say we have heights and weights of five people:

Person	Height (cm)	Weight (Kg)	Z Score (Standardized)
A	150	50	
B	160	60	
C	170	70	
D	180	80	
E	190	90	

Min–Max Scaling (Rescaling to [0, 1])

Formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Let's say we have heights and weight of five people:

Person	Height (cm)	Weight (Kg)(x)	Scaled (x')
A	150	50	
B	160	60	
C	170	70	
D	180	80	
E	190	90	

Robust Scaling (Outlier-Resistant)

Formula:

$$x' = \frac{x - \text{median}}{\text{IQR}}$$

where IQR=Q3−Q1

Let's say we have heights and weight of five people:

Person	Height (cm)	Weight (Kg)(x)	Scaled (x')
A	150	50	
B	160	60	
C	170	70	
D	180	80	
E	300	200	

Nominal vs Ordinal + One-Hot Encoding

Nominal: Categories with no order
Ordinal: Clear ranking

Tiny dataset

id	color	size	price
1	red	Small	10
2	blue	Medium	12
3	green	Large	15
4	red	Medium	11

One-hot the nominal feature by hand

Step 1: Creating New columns:

- Color_red, color_blue, color_green

Step 2: The transformed table:

id	price	color	Color_red	Color_blue	Color_green
1	10	red	1	0	0
2	12	blue	0	1	0
3	15	green	0	0	1
4	11	red	1	0	0

Step 3: Row by row visualization:

- id 1 red → [1,0,0]
- id 2 blue→ [0,1,0]
- id 3 green→[0,0,1]
- id 4 red → [1,0,0]

Ordinal Encoding

Ordinal: Clear ranking

Tiny dataset

id	color	size	price
1	red	Small	10
2	blue	Medium	12
3	green	Large	15
4	red	Medium	11

Ordinal encoding for ordinal feature: size

Step 1: Decide Order:

Small: 1, Medium: 2, Large: 3

Step 2: The transformed table:

id	color	size	price	Size_encode
1	red	Small	10	1
2	blue	Medium	12	2
3	green	Large	15	3
4	red	Medium	11	2

Euclidean distance & Manhattan distance

What these measure

- **Euclidean distance**: straight line closeness.
- **Manhattan distance**: city-block closeness.

Formulas

For p = [p1, p2, ..., pk], q = [q1, q2, ..., qk]

- Euclidean: $d2(p,q) = \sqrt{\Sigma (pi - qi)^2}$
- Manhattan: $d1(p,q) = \Sigma |pi - qi|$

Tiny dataset

S1 = [70, 80]
S2 = [60, 90]
S3 = [85, 60]
S4 = [78, 76]
S5 = [62, 65]
q = [75, 70] # query point

S1 vs q

- Diffs: $(70 - 75) = -5$, $(80 - 70) = 10$
- Euclidean: $\sqrt{25 + 100} = \sqrt{125} \approx 11.180$
- Manhattan: $|-5| + |10| = 5 + 10 = 15$