



What is Preprocessing?

w h name
✓ X BMW

Data missing:

Solution: Calc. avg. feature.

How to handle missing data?

Here height is numerical feature.

avg not possible ↓

F1 F2 F3 Label

value 1, 2, 3

avg: 2.23 X
not meaningful.

Since only 1, 2, 3 is applicable.

10 F1=1
30 F1=2
20 F1=3

choose this if data missing

calculate majority value

Example:

BMW

Toyota

10 F1=1
30 F1=2
20 F1=3

30 F1=1
20 F1=2
40 F1=3

if BMW data missing choose 30 F1=2

Or just choose highest in BMW and Toyota (40).

if data is not numerical?

w h → Convert to numerical.
Short = 1
medium = 2
long = 3

