Data Processing and Feature Engineering: Encoding Categorical Variables

Categorical (also called discrete),
- Attribute can take a finite set of values

Examples
color ∈ (red, green, blue)
isFraud ∈ (true, false)

Pandas
supports special dtype = "category"

Ordinal - categories are ordered
e.g. size - small, medium, large

Nominal - categories are unordered
e.g. color

Issue: Categoricals are often represented as text but many algorithms require numericals as input
- we need special encoding to convert categoricals into numerical representation

loan_approval ~ encode Y or N as 1 or 0

garden_size - ordinal (S, M, L) use pandas map function
N - no garden 0, S - 5, M - 10, L - 20

use label_encoder from sklearn
wrong solution when no relationship between

If no ordering, or no relative categories sizing, conversion/mapping will not give us a good result.