HANDLING CATEGORICAL VARIABLES

As machine learning model understanding numbering so it is necessary to change to numeric ones

town	area	price
monroe township	2600	550000
monroe township	3000	565000
monroe township	3200	610000
monroe township	3600	680000
monroe township	4000	725000
west windsor	2600	585000
west windsor	2800	615000
west windsor	3300	650000
west windsor	3600	710000
robbinsville	2600	575000
robbinsville	2900	600000
robbinsville	3100	620000

town	area	price
1	2600	550000
1	3000	565000
1	3200	610000
1	3600	680000
1	4000	725000
2	2600	585000
2	2800	615000
2	3300	650000
2	3600	710000
3	2600	575000
3	2900	600000
3	3100	6

onroe township = 1, West Windsor = 2, Robbinsville =

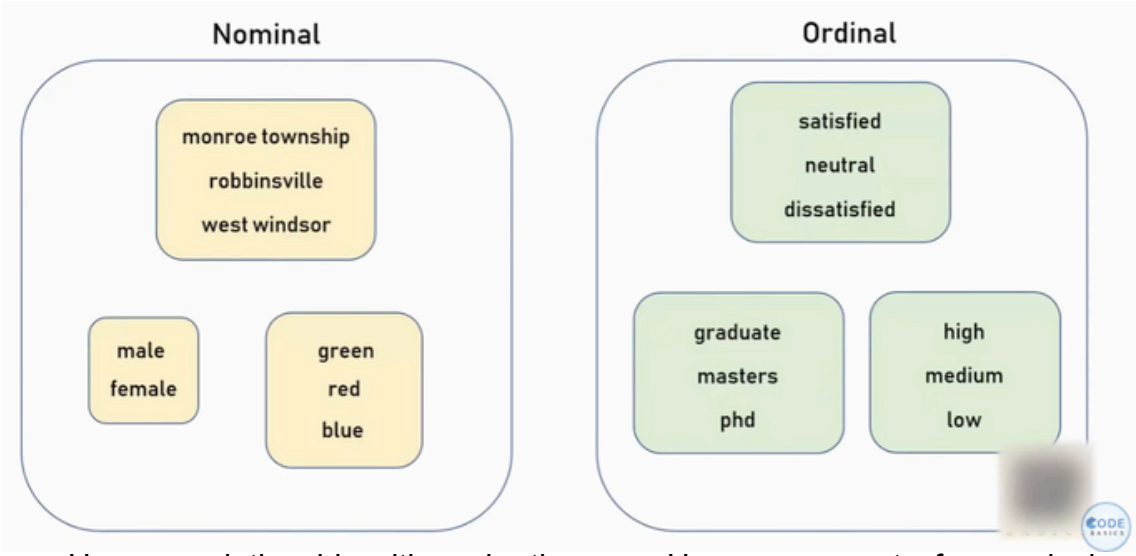
As given towns are nominal. So, when we assign like 1, 2, 3. It assume in order like below which might make no sense

Monroe township < West Windsor < Robbinsville ??

Or

Monroe township + West Windsor = Robbinsville ??

TYPES OF CATEGORICAL VARIABLES



Have no relationship with each other

Have some sort of numerical ordering/relation between them . For example we can say that high have low value than other .....

NOMINAL VARIABLES

Integer based encoding do not work for nomnal variables.So, we use one hot encoding(get_dummies)

GIVES RESULT LIKE THIS

town	area	price	monroe township	west windsor	robbinsville
monroe township	2600	550000	1	0	0
monroe township	3000	565000	1	0	0
monroe township	3200	610000	1	0	0
monroe township	3600	680000	1	0	0
monroe township	4000	725000	1	0	0
west windsor	2600	585000	0	1	0
west windsor	2800	615000	0	1	0
west windsor	3300	650000	0	1	0
west windsor	3600	710000	0	1	0
robbinsville	2600	575000	0	0	1
robbinsville	2900	600000	0	0	1
robbinsville	3100	620000	0	0	
robbinsville	3600	695000	0	0	

When to Use `LabelEncoder`:

- **Ordinal Data:** When there is a meaningful order to the categories. Examples include ratings (e.g., 'low', 'medium', 'high') or levels (e.g., 'beginner', 'intermediate', 'advanced').

When Not to Use `LabelEncoder`:

- **Nominal Data:** When there is no inherent order among the categories. Examples include types of fruit (e.g., 'apple', 'orange', 'banana') or city names.

- In label encoding ,each data value is assigned a **distinct number** instead of a **qualitative value**.
- In one-hot encoding, each unique value is transformed into a new binary (0/1) feature column.