Encoding Categorical Data

Types of Categorical Data

- Nominal Data
- Ordinal Data

Nominal data is a type of data that categorizes or labels items or individuals without any order or ranking.

For e.g

- Person 1: Blue
- Person 2: Red
- Person 3: Green
- Person 4: Blue
- Person 5: Yellow

Ordinal data is similar to nominal data, but with an added level of hierarchy or ranking. In ordinal data, categories or labels have a specific order or rank associated with them.

For e.g

- Very Dissatisfied
- Dissatisfied
- Neutral
- Satisfied
- Very Satisfied

Label Encoder:

A Label Encoder is a tool used in machine learning and data preprocessing to convert categorical data into numerical values. In simpler terms, it's a way to transform text labels into numbers so that machine learning algorithms can process them.

Label Encoder is always used on Y variables, but If you have an Ordinal category in X variables you will apply their Ordinal Encoding.

One Hot Encoding:

One-hot encoding is another technique used in machine learning and data preprocessing, particularly for handling categorical variables. It's used to convert categorical data into a binary format, where each category is represented as a binary vector.

Transformation: Each categorical value is replaced with a binary vector indicating which category it belongs to. For example:

- "Red" might be represented as [1, 0, 0]
- "Blue" might be represented as [0, 1, 0]
- "Green" might be represented as [0, 0, 1]

Dummy Variable Trap

Multicollinearity: Multicollinearity occurs when two or more independent variables in a regression model are highly correlated. When the dummy variable trap is present, the dummy variables are perfectly correlated, leading to multicollinearity. Multicollinearity makes it challenging to assess the individual effects of each variable on the dependent variable, inflates standard errors, and can lead to unstable coefficient estimates.

One Hot Encoding using Most Frequent Variables

We keep the topmost frequent categories and transfer all remaining categories into a completely new category, let's say using others named columns.