# Univariate Imputation- Categorical Data

09 July 2023 16:45

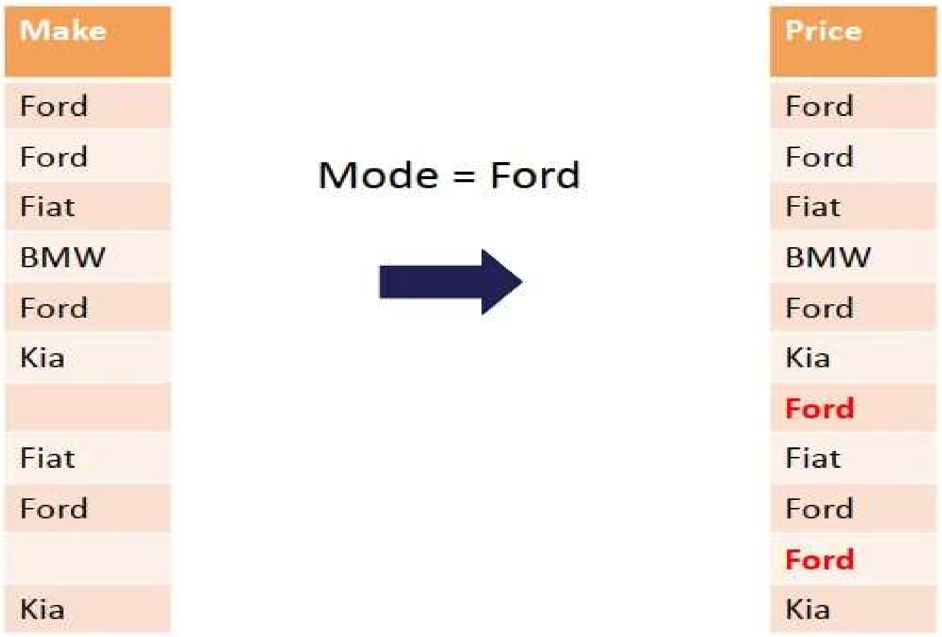
**Univariate Imputation - Categorical Feature:-**

a. **Mode Imputation** – most frequent value/category

b. **"Missing" Category Imputation**

## 1. Most Frequent Value or Mode Imputation:

*Mode imputation means replacing all the missing values within a feature by the mode of that feature, which in other words refers to the* ***most frequent value*** *or* ***most frequent category****.*



| Make |
|---|
| Ford |
| Ford |
| Fiat |
| BMW |
| Ford |
| Kia |
| |
| Fiat |
| Ford |
| |
| Kia |

Mode = Ford

| Price |
|---|
| Ford |
| Ford |
| Fiat |
| BMW |
| Ford |
| Kia |
| Ford |
| Fiat |
| Ford |
| Ford |
| Kia |

Mode Imputation can be applied for both numerical and categorical variables(columns). But Mean/Median gives best result for numerical variables so Mode is not preferred.

***When to use ?***

a. *When Data is* ***Missing Completely At Random****.*

   *A way to determine the type of missingness is by performing imputation methods and observing the impact on the distribution, correlation etc.*

b. *When Missing values are* ***less than 5%*** *of the total values in the variable.*

c. *The* ***most frequent category*** *should be present in* ***far greater number of rows*** *in comparison to other features.*

This approach is **easy** to implement but it **significantly changes the correlation of most frequent category with other features** in the data.

2. **"Missing" Category Imputation:**

   Here we add a **new category** in the feature by replacing all the missing values with the word "**Missing", "Not Defined", "NA"** etc. This is how we tell the model where the missing values are so that the model considers this too while training.

   ***When to use ?***

   a. When missing values  are more than 5%

   **Advantages:**

   a. Easy to implement

   **Disadvantages:**

   a. Introduces additional randomness in the data.
   b. Does not give good result as such.