# Deal With Categorical Variable

# **What is a Variable?**

A variable is any characteristic, number, or quantity that can be measured or counted. We call them ****variables**** because the values they take may vary, and usually do.

The following are examples of variables:

* Age (23, 52, 60, …).
* Gender (male, female)
* Country of birth (Algeria, USA, Japan, …)
* Eye color (green, brown, blue, and purple maybe)
* Vehicle Brand (Buggati, Ferrari)

We classify variables in a dataset into one of these major types:

* ****Numerical variables****
* ****Categorical variables****
* ****Datetime variables****
* ****Mixed variables****

## Numerical Variables

The values of numerical variables are (predictably) numbers. For example, total rainfall measured in inches, heart rate, the number of cheeseburgers consumed in an hour—all numerical values.

We can further classify them into:

* Continuous variables
* Discrete variables

### @ Continuous Variable

A ****continuous variable**** is one that can take on an uncountable set of values. It may contain any value within a given range.

* For example****,**** the total amount paid by a customer in a supermarket is continuous. The customer can pay $20, $16.50, $150, and so on.

**To visualize such a variable we have a range of options, including:**

* Density plot
* Histogram
* Box plot
* Scatter plot.

### @Discrete Variable

A discrete variable is a variable that can only take on a finite number of values, and these values are integers ( — which means numbers that are not a fraction), they are counts.

* For example, the number of things bought by a customer in a supermarket is discrete. The customer can buy 2, 20, or 150 things, but not 10.4 items. It is always a round number.

**To visualize such discrete variables, you can use the following type of plots:**

* Count plot
* Pie chart

## Categorical Variables

The values of a categorical variable are selected from a group of ****categories****, also called ****labels****. Examples are gender (male or female) and marital status (never married, married, divorced, or widowed).

Other examples of categorical variables include:

* Gender (male, female)
* Mobile network provider (Mobilis, Vodafone, Orange)
* City name (Tiaret, Algiers, Texas, Dubai)

## Dates and Times

A particular type of categorical variable are those that take dates or time as values.

For example:

* Date of birth (16–04–1997, 12–01–2012)
* Date of application (2020-Jan, 2022-Feb)

## Mixed Variables

Finally, mixed variables are those whose values can contain both numbers and labels.

For a variety of reasons, mixing variables can occur in a given dataset, especially when filling its values.

# Imputing Missing Values.

## Data Imputation

Data imputation is the act of replacing missing data with statistical estimates of the missing values.

### Missing Data Imputation Techniques

We’re going to dive into techniques that apply to numerical and categorical variables, and also some methods that apply to both:

### Numerical Variables

* Mean or median imputation
* Arbitrary value imputation
* End of tail imputation

### Categorical Variables

* Frequent category imputation
* Add a ****missing**** category

# **Mean or Median Imputation**

Mean or median imputation consists of replacing all occurrences of missing values (NA) within a variable with the ****mean or median**** of that variable.

*This method is suitable for numerical variables.*

Here are some points to consider when using this method:

* If the variable follows a normal distribution, the mean and median are approximately the same.
* If the variable has a skewed distribution, then the median is a better representation.

# **Missing Category Imputation**

This method consists of treating missing data as an ****additional****label or category of the variable. Thus, we create a new label or category by filling the missing observations with a ****Missing****category.

CODE for IMPUTER of MISSING VALUES

|  |
| --- |
|  |
|  | from sklearn.impute import SimpleImputer |
|  | # create the imputer, the strategy can be mean and median. |
|  | imputer = SimpleImputer(missing\_values=np.nan, strategy='mean') |
|  |  |
|  | # fit the imputer to the train data |
|  | imputer.fit(train) |
|  |  |
|  | # apply the transformation to the train and test |
|  | train = imputer.transform(train) |
|  | test = imputer.transform(test) |
|  |  |

**# create the imputer, the strategy can be mean and median.**

**imputer = SimpleImputer(missing\_values=np.nan, strategy='mean')**

**# create the imputer, with fill value 999 as the arbitraty value**

**imputer = SimpleImputer(missing\_values=np.nan, strategy='constant', fill\_value=999)**

**# create the imputer (End Tail Impouter)**

**imputer = EndTailImputer(distribution='gaussian', tail='right')**

**# create the imputer, with most frequent as strategy to fill missing value.**

**imputer = SimpleImputer(missing\_values=np.nan, strategy='most\_frequent')**

**# create the imputer, with most frequent as strategy to fill missing value.**

**imputer = SimpleImputer(missing\_values=np.nan, strategy='constant', fill\_value="Missing")**

# **Arbitrary Value Imputation ( Constant Imputation)**

Arbitrary value imputation consists of replacing all occurrences of missing values (NA) within a variable with an arbitrary value. The arbitrary value should be different from the mean or median and not within the****normal values of the variable.****

We can use arbitrary values such as 0, 999, -999 (or other combinations of 9s) or -1 (if the distribution is positive).

# **End of Tail Imputation**

End of tail imputation is roughly equivalent to arbitrary value imputation, but it automatically selects the arbitrary values at the end of the variable distributions.

*This method is suitable for numerical variables.*

Here are ways to select arbitrary values:

* If the variable follows a normal distribution, we can use the mean plus or minus 3 times the standard deviation.
* If the variable is skewed, we can use the IQR proximity rule.

*The values to replace missing data should be calculated only on the train set.*

# **Frequent Category Imputation**

Frequent category imputation—or ****mode imputation****—consists of replacing all occurrences of missing values (NA) within a variable with the mode, or the most frequent value.

*This method is suitable for numerical and categorical variables, but in practice, we use this technique with categorical variables.*