

# How can we handle mixed variables?

## What is Mixed Data?

In the real world, we often encounter different types of data in machine learning, and if your luck is bad, you may end up with mixed data. Mixed data refers to data where your columns contain a combination of numerical and categorical types of data. Handling such mixed data can be quite challenging because it's not always clear how to proceed. So, first, let me explain the various ways in which you can encounter this type of data.

C23 C25 C27	6
B57 B59 B63 B66	5
G6	5
F33	4
C22 C26	4
D	4
F4	4
F2	4
B96 B98	4
C78	4
A34	3
B58 B60	3
E34	3
C101	3
E101	3
B51 B53 B55	3
B69	2
B28	2

The first scenario is when you have a single column, such as the example of the Titanic dataset where there's a column named "Cabin" as mentioned earlier. In the "Cabin" column, you can see values like C23, C25, C27, and so on. Here, you can observe that these are not categories. If you were to consider them as categorical data, you would end up with numerous categories.

Now, the problem here is that information is hidden in two aspects. For example, C23 contains two pieces of information: first, it belongs to class C, and second, its cabin number is 23. So, C becomes our categorical data, while 23 becomes numerical.

The straightforward conclusion is that when you encounter such data, you need to split it into two parts or columns. One column where you store categorical information and another where you store numerical information. So, this is the first type where mixed variables can pose a challenge.

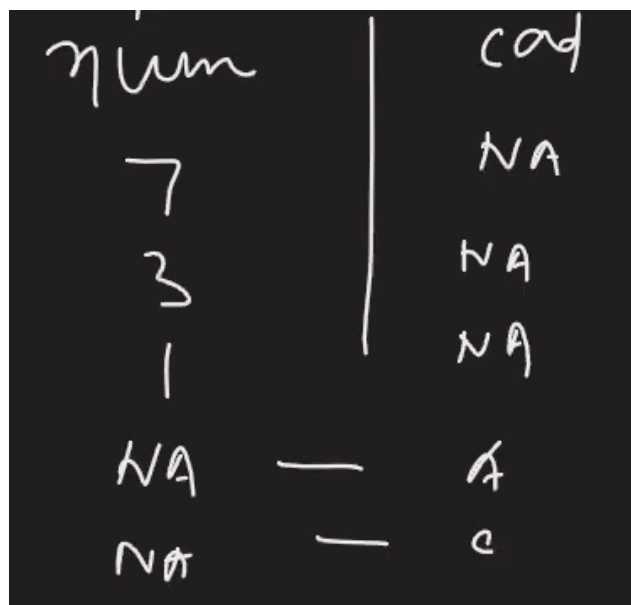
## 2nd Type of Handling of Mixed Variables

Another type of problem occurs when, for instance, you are given values like [7, 1, 4, 5, A, R, B, C] in a column. Here, the values are a mix of numeric and categorical data. This situation can be quite chaotic because some values are numeric while others are categorical. In this scenario as well, the best approach is to split the data into two separate columns. One column should be created for numerical values, and another for categorical values. So, the resulting columns might look something like this:

Numerical Column: [7, 1, 4, 5, \_, \_ , \_]

Categorical Column: [, \_, \_, \_, A, R, B, C]

Here, the underscore (\_) represents missing values or placeholders in the respective columns. By splitting the data in this way, you can handle the mixed types effectively, allowing for clearer analysis and processing.



A handwritten table on a black background with white text. The table has two columns. The first column is labeled 'num' and the second column is labeled 'cat'. The data rows are as follows:

num	cat
7	NA
3	NA
1	NA
NA	A
NA	C

Here, you can see that "NA" will appear in front of the ones opposite to the numerical ones and "NA" will appear in front of the ones opposite to the categorical ones, so the data will be distributed in two columns in this way.