**One-Hot Encoding vs. Label Encoding**

1. What is Categorical Encoding?
2. Different Approaches to Categorical Encoding

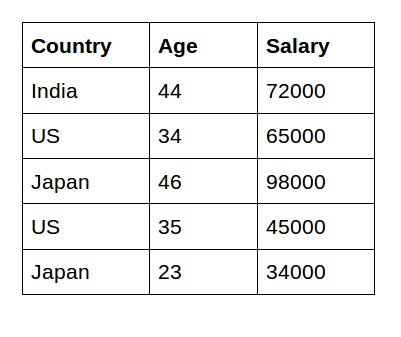
a. Label Encoding

b. One-Hot Encoding

1. When to use Label Encoding vs. One-Hot Encoding?

The reason we convert categorical columns to numerical columns is so that a machine learning algorithm understands it. This process is called **categorical encoding.**

A. Label Encoding



In the table above, we can see that ‘Country’ is a categorical feature. More specifically it is a Nominal feature since the Country names do not have an order or rank.

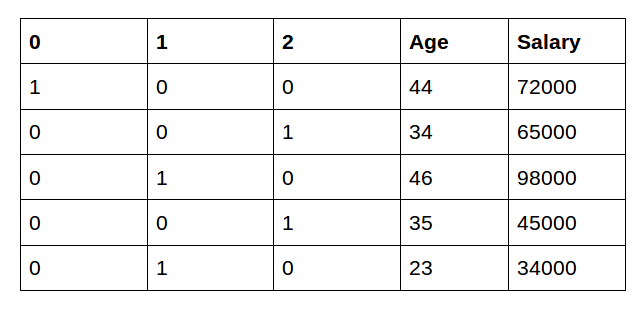
\*\* Now when we perform label encoding in **scikit learn**, the country names are labelled based on the order of the alphabets. i.e. India will be encoded with 0, the US with 2, and Japan with 1 giving a feel that there is a relationship between countries such as India < Japan < the US.

In **SPSS** we can use a reclassify node and manually label the countries as we wish.

B. One-Hot Encoding

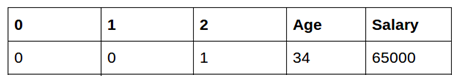
One-Hot Encoding is the process of creating dummy variables. In this encoding technique, each category is represented as a one vector.

India is represented by column 0, Japan by column 1, and US by column 2.



One-Hot Encoding results in a **Dummy Variable Trap** as the outcome of one variable can easily be predicted with the help of the remaining variables. i.e. in this case,

If the country is not India or Japan, it is definitely US. (given that it is a non nullable column)



Dummy Variable Trap is a scenario in which variables are highly correlated to each other. The Dummy Variable Trap leads to the problem known as **Multicollinearity**. Multicollinearity occurs where there is a dependency between the independent features. Multicollinearity is a serious issue in machine learning algorithms like Linear Regression and Logistic Regression.

So, in order to overcome the problem of multicollinearity, one of the dummy variables needs to be dropped. And therefore we can drop any 1 of the three columns. (which will be represented automatically by other 2 columns being 0).

**When to use a Label Encoding vs. One Hot Encoding**

We apply One-Hot Encoding when:

1. The categorical feature is **not ordinal** (like the countries above)
2. The number of categorical features is less so one-hot encoding can be effectively applied.

We apply Label Encoding when:

1. The categorical feature is **ordinal** (like Jr. kg, Sr. kg, Primary school, high school)
2. The number of categories is quite large as one-hot encoding can lead to many features/dimensions
3. When we have a large number of categories in a single feature, we can also do an aggregate to see the top 20-30 categories and label encode them, while the other sparse classes can be simply labeled as ‘others’