In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import LabelEncoder
```

## Reading the data.

In [4]:
```
import os
os.chdir(r'C:\Users\acer\Desktop\PythonProgramming')
data_pre = pd.read_csv('Data.csv', na_values = np.nan)
```

## Imputing with mean in columns = ['Age', 'Salary']

In [5]:
```
data_pre['Age'].fillna(data_pre['Age'].mean(), inplace = True)
data_pre['Salary'].fillna(data_pre['Salary'].mean(), inplace = True)
```

In [6]:
```
data_pre
```

| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 0 | France | 44.000000 | 72000.000000 | No |
| 1 | Spain | 27.000000 | 48000.000000 | Yes |
| 2 | Germany | 30.000000 | 54000.000000 | No |
| 3 | Spain | 38.000000 | 61000.000000 | No |
| 4 | Germany | 40.000000 | 63777.777778 | Yes |
| 5 | France | 35.000000 | 58000.000000 | Yes |
| 6 | Spain | 38.777778 | 52000.000000 | No |
| 7 | France | 48.000000 | 79000.000000 | Yes |
| 8 | Germany | 50.000000 | 83000.000000 | No |
| 9 | France | 37.000000 | 67000.000000 | Yes |

## Copying the dataframe.

In [8]:
```
from copy import deepcopy
```

In [9]:
```
data_mod = deepcopy(data_pre)
```

## One Hot Encoder Object of the class OneHotEncoder.

In [10]:
```
ohe = OneHotEncoder()
```

## Target column is Country.

In [11]:
```
tt = data_mod.iloc[:, 0].values.reshape(-1,1)  # Reshaping is for making it a 2D array.
tt
```

```
array([['France'],
       ['Spain'],
       ['Germany'],
       ['Spain'],
       ['Germany'],
       ['France'],
       ['Spain'],
       ['France'],
       ['Germany'],
       ['France']], dtype=object)
```

## Fitting into the object.

In [21]:
```
ohe_array = ohe.fit_transform(tt).toarray()
ohe_array = ohe_array[:,1:]    # Removed the france column for multiple co - linearity.
ohe_array
```

```
array([[0., 0.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [1., 0.],
       [0., 0.],
       [0., 1.],
       [0., 0.],
       [1., 0.],
       [0., 0.]])
```

## Reminder!!

In [17]:
```
uu = {'France': 0, 'Germany': 1, 'Spain': 2}
```

## Combining the newly formed - encoded array and the original dataframe.

- First we need to convert the series in to DataFrame(or series) using pd.DataFrame(ohe_array).

In [26]:
```
new_df = pd.concat([pd.DataFrame(ohe_array), data_mod], axis = 'columns')
new_df
```

| | 0 | 1 | Country | Age | Salary | Purchased |
|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | France | 44.000000 | 72000.000000 | No |
| 1 | 0.0 | 1.0 | Spain | 27.000000 | 48000.000000 | Yes |
| 2 | 1.0 | 0.0 | Germany | 30.000000 | 54000.000000 | No |
| 3 | 0.0 | 1.0 | Spain | 38.000000 | 61000.000000 | No |
| 4 | 1.0 | 0.0 | Germany | 40.000000 | 63777.777778 | Yes |
| 5 | 0.0 | 0.0 | France | 35.000000 | 58000.000000 | Yes |
| 6 | 0.0 | 1.0 | Spain | 38.777778 | 52000.000000 | No |
| 7 | 0.0 | 0.0 | France | 48.000000 | 79000.000000 | Yes |
| 8 | 1.0 | 0.0 | Germany | 50.000000 | 83000.000000 | No |
| 9 | 0.0 | 0.0 | France | 37.000000 | 67000.000000 | Yes |

## Dropping the Country column as it is not of much significance.

In [31]:
```
new_df = new_df.drop(['Country'],axis = 1)
```

## Encoded DataFrame.

In [32]:
```
new_df
```

| | 0 | 1 | Age | Salary | Purchased |
|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 44.000000 | 72000.000000 | No |
| 1 | 0.0 | 1.0 | 27.000000 | 48000.000000 | Yes |
| 2 | 1.0 | 0.0 | 30.000000 | 54000.000000 | No |
| 3 | 0.0 | 1.0 | 38.000000 | 61000.000000 | No |
| 4 | 1.0 | 0.0 | 40.000000 | 63777.777778 | Yes |
| 5 | 0.0 | 0.0 | 35.000000 | 58000.000000 | Yes |
| 6 | 0.0 | 1.0 | 38.777778 | 52000.000000 | No |
| 7 | 0.0 | 0.0 | 48.000000 | 79000.000000 | Yes |
| 8 | 1.0 | 0.0 | 50.000000 | 83000.000000 | No |
| 9 | 0.0 | 0.0 | 37.000000 | 67000.000000 | Yes |

# End.