Q1 Data Encoding is an important pre-processing step in Machine Learning. It refers to the process of converting categorical or textual data into numerical format, so that it can be used as input for algorithms to process.

In computers, encoding is the process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage.

Q2 When we have a feature where variables are just names and there is no order or rank to this variable's feature. For example: City of person lives in, Gender of person, Marital Status, etc... In the above example, We do not have any order or rank, or sequence.

male/female (albeit somewhat outdated), hair color, nationalities, names of people, and so on

Q3 One-Hot encoding technique is used when the features are nominal(do not have any order). In one hot encoding, for every categorical feature, a new variable is created. Categorical features are mapped with a binary variable containing either 0 or 1. The number of categorical features is less so one-hot encoding can be effectively applied.

In [7]:
```
# example of a one hot encoding
from numpy import asarray
from sklearn.preprocessing import OneHotEncoder
# define data
data = asarray([['red'], ['green'], ['blue']])
print(data)
# define one hot encoding
encoder = OneHotEncoder(sparse=False)
# transform data
onehot = encoder.fit_transform(data)
print(onehot)
```

```
[['red']
 ['green']
 ['blue']]
[[0. 0. 1.]
 [0. 1. 0.]
 [1. 0. 0.]]
/opt/conda/lib/python3.10/site-packages/sklearn/preprocessing/_encoders.py:808: Fu
tureWarning: `sparse` was renamed to `sparse_output` in version 1.2 and will be re
moved in 1.4. `sparse_output` is ignored unless you leave `sparse` to its default
value.
  warnings.warn(
```

Q4 Encoding technique we used is one hot encoding: because in one hot encoding every categorical variables map with the binary variables.A common approach to encoding categorical features is to apply one-hot encoding. This method encodes categorical variables by adding one binary variable for each unique category.

Q5 While encoding Nominal data, we have to consider the presence or absence of a feature. In such a case, no notion of order is present. For example, the city a person lives in. For the data, it is important to retain where a person lives. Here, We do not have any order or sequence. It is equal if a person lives in Delhi or Bangalore.

In [8]:
```
import pandas as pd
import numpy as np

# Define the headers since the data does not have any
headers = ["symboling", "normalized_losses", "make", "fuel_type", "aspiration",
           "num_doors", "body_style", "drive_wheels", "engine_location",
           "wheel_base", "length", "width", "height", "curb_weight",
           "engine_type", "num_cylinders", "engine_size", "fuel_system",
           "bore", "stroke", "compression_ratio", "horsepower", "peak_rpm",
           "city_mpg", "highway_mpg", "price"]

# Read in the CSV file and convert "?" to NaN
df = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/autos/i
                 header=None, names=headers, na_values="?" )
df.head()
```

Out[8]:

	symboling	normalized_losses	make	fuel_type	aspiration	num_doors	body_style	drive_whe
0	3	NaN	alfa-romero	gas	std	two	convertible	r
1	3	NaN	alfa-romero	gas	std	two	convertible	r
2	1	NaN	alfa-romero	gas	std	two	hatchback	r
3	2	164.0	audi	gas	std	four	sedan	f
4	2	164.0	audi	gas	std	four	sedan	4

5 rows × 26 columns

Q6 a dataset containing information about different types of animals, including their species, habitat, and diet. We use Ordinal encoding to transform categorical data to a foramt suitable for machine learning algorithm.So this dataset is useful for every dataset.

Q7 We have a dataset with 5 features, including the customer's gender, age, contract type,monthly charges, and tenure.The encoding technique(s) will use to transform the categorical data is One-hot-encoding we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category.hese newly created binary features are known as Dummy variables. The number of dummy variables depends on the levels present in the categorical variable. This might sound complicated. Let us take an example to understand

this better. Suppose we have a dataset with a category like gender,age,monthly charges,tenure. Now we have to one-hot encode this data.