In [ ]:

```
youtube link:- https://www.youtube.com/watch?v=U5oCv3JKWKA&list=RDCMUCCWi3hpnq_Pe03nGxuS7isg&index=1&ab_channel=Campus
```

In [1]:

```
import numpy as np
import pandas as pd
```

In [2]:

```
df=pd.read_csv("cars.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	brand	km_driven	fuel	owner	selling_price
0	Maruti	145500	Diesel	First Owner	450000
1	Skoda	120000	Diesel	Second Owner	370000
2	Honda	140000	Petrol	Third Owner	158000
3	Hyundai	127000	Diesel	First Owner	225000
4	Maruti	120000	Petrol	First Owner	130000

In [4]:

```
df['brand'].unique()
```

Out[4]:

```
array(['Maruti', 'Skoda', 'Honda', 'Hyundai', 'Toyota', 'Ford', 'Renault',
       'Mahindra', 'Tata', 'Chevrolet', 'Fiat', 'Datsun', 'Jeep',
       'Mercedes-Benz', 'Mitsubishi', 'Audi', 'Volkswagen', 'BMW',
       'Nissan', 'Lexus', 'Jaguar', 'Land', 'MG', 'Volvo', 'Daewoo',
       'Kia', 'Force', 'Ambassador', 'Ashok', 'Isuzu', 'Opel', 'Peugeot'],
      dtype=object)
```

In [5]:
```
df['brand'].value_counts()
```

Out[5]:
```
Maruti           2448
Hyundai          1415
Mahindra          772
Tata              734
Toyota            488
Honda             467
Ford              397
Chevrolet         230
Renault           228
Volkswagen        186
BMW               120
Skoda             105
Nissan             81
Jaguar             71
Volvo              67
Datsun             65
Mercedes-Benz      54
Fiat               47
Audi               40
Lexus              34
Jeep               31
Mitsubishi         14
Force               6
Land                6
Isuzu               5
Kia                 4
Ambassador          4
Daewoo              3
MG                  3
Ashok               1
Opel                1
Peugeot             1
Name: brand, dtype: int64
```

In [6]:
```
df['owner'].value_counts()
```

Out[6]:
```
First Owner             5289
Second Owner            2105
Third Owner              555
Fourth & Above Owner     174
Test Drive Car             5
Name: owner, dtype: int64
```

In [7]:
```
df['fuel'].value_counts()
```

Out[7]:
```
Diesel    4402
Petrol    3631
CNG         57
LPG         38
Name: fuel, dtype: int64
```

1. ONE HOT ENCODING

In [9]:
```
pd.get_dummies(df,columns=['fuel','owner'])
```

Out[9]:

	brand	km_driven	selling_price	fuel_CNG	fuel_Diesel	fuel_LPG	fuel_Petrol	owner_First Owner	owner_Fourth & Above Owner	owner_Second Owner	owner_Test Drive Car	owı
0	Maruti	145500	450000	0	1	0	0	1	0	0	0	
1	Skoda	120000	370000	0	1	0	0	0	0	1	0	
2	Honda	140000	158000	0	0	0	1	0	0	0	0	
3	Hyundai	127000	225000	0	1	0	0	1	0	0	0	
4	Maruti	120000	130000	0	0	0	1	1	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	
8123	Hyundai	110000	320000	0	0	0	1	1	0	0	0	
8124	Hyundai	119000	135000	0	1	0	0	0	1	0	0	
8125	Maruti	120000	382000	0	1	0	0	1	0	0	0	
8126	Tata	25000	290000	0	1	0	0	1	0	0	0	
8127	Tata	25000	290000	0	1	0	0	1	0	0	0	

8128 rows × 12 columns

2. (K-1) ONE HOT ENCODING

In [12]:
```
# K is the no. of categories (K minus 1)
```

In [ ]:
```
# Multicollearity
```

In [11]:

```
pd.get_dummies(df,columns=['fuel','owner'],drop_first=True)
```

Out[11]:

	brand	km_driven	selling_price	fuel_Diesel	fuel_LPG	fuel_Petrol	owner_Fourth & Above Owner	owner_Second Owner	owner_Test Drive Car	owner_Third Owner
0	Maruti	145500	450000	1	0	0	0	0	0	0
1	Skoda	120000	370000	1	0	0	0	1	0	0
2	Honda	140000	158000	0	0	1	0	0	0	1
3	Hyundai	127000	225000	1	0	0	0	0	0	0
4	Maruti	120000	130000	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
8123	Hyundai	110000	320000	0	0	1	0	0	0	0
8124	Hyundai	119000	135000	1	0	0	1	0	0	0
8125	Maruti	120000	382000	1	0	0	0	0	0	0
8126	Tata	25000	290000	1	0	0	0	0	0	0
8127	Tata	25000	290000	1	0	0	0	0	0	0

8128 rows × 10 columns

3. ONE HOT ECODING using SKlearn

In [13]:

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(df.iloc[:,0:4],df.iloc[:,-1],test_size=0.2,random_state=2)
```

In [14]:
```
X_train.head()
```

Out[14]:

	brand	km_driven	fuel	owner
5571	Hyundai	35000	Diesel	First Owner
2038	Jeep	60000	Diesel	First Owner
2957	Hyundai	25000	Petrol	First Owner
7618	Mahindra	130000	Diesel	Second Owner
6684	Hyundai	155000	Diesel	First Owner

In [15]:
```
from sklearn.preprocessing import OneHotEncoder
```

In [16]:
```
ohe = OneHotEncoder(drop='first',sparse=False,dtype=np.int32)
```

In [17]:
```
X_train_new = ohe.fit_transform(X_train[['fuel','owner']])
```

```
C:\Users\harsh\anaconda3\lib\site-packages\sklearn\preprocessing\_encoders.py:828: FutureWarning: `sparse` was renam
ed to `sparse_output` in version 1.2 and will be removed in 1.4. `sparse_output` is ignored unless you leave `sparse
` to its default value.
  warnings.warn(
```

In [18]:
```
X_test_new = ohe.transform(X_test[['fuel','owner']])
```

In [19]:
```
X_train_new.shape
```

Out[19]:
```
(6502, 7)
```

4. OneHotEncoding with Top Categories

In [20]:
```
counts = df['brand'].value_counts()
```

In [21]:
```
df['brand'].nunique()
threshold = 100
```

In [22]:
```
repl = counts[counts <= threshold].index
```

In [23]:
```
pd.get_dummies(df['brand'].replace(repl, 'uncommon')).sample(5)
```

Out[23]:

	BMW	Chevrolet	Ford	Honda	Hyundai	Mahindra	Maruti	Renault	Skoda	Tata	Toyota	Volkswagen	uncommon
5963	0	0	0	0	1	0	0	0	0	0	0	0	0
6846	0	0	0	0	0	0	1	0	0	0	0	0	0
4527	0	0	0	0	0	0	0	0	0	0	0	0	1
2588	0	0	0	0	0	0	1	0	0	0	0	0	0
4590	0	0	0	0	0	0	1	0	0	0	0	0	0

In [ ]:
```

```