```
import numpy as np
import pandas as pd
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
df = pd.read_csv('/content/drive/MyDrive/python-Saylani/cars.csv')
```

```
df.head()
```

	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

```
df['owner'].value_counts()
```

owner	count
First Owner	5289
Second Owner	2105
Third Owner	555
Fourth & Above Owner	174
Test Drive Car	5

dtype: int64

```
x=df.drop(columns='selling_price')
y=df['selling_price']
```

```
from sklearn.model_selection import train_test_split
```

```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2)
```

```
x
```

	brand	km_driven	fuel	owner
0	Maruti	145500	Diesel	First Owner
1	Skoda	120000	Diesel	Second Owner
2	Honda	140000	Petrol	Third Owner
3	Hyundai	127000	Diesel	First Owner
4	Maruti	120000	Petrol	First Owner
...	...	...	...	...
8123	Hyundai	110000	Petrol	First Owner
8124	Hyundai	119000	Diesel	Fourth & Above Owner
8125	Maruti	120000	Diesel	First Owner
8126	Tata	25000	Diesel	First Owner
8127	Tata	25000	Diesel	First Owner

8128 rows × 4 columns

```
y
```

	selling_price
0	450000
1	370000
2	158000
3	225000
4	130000
...	...
8123	320000
8124	135000
8125	382000
8126	290000
8127	290000

8128 rows × 1 columns

dtype: int64

3. OneHotEncoding

```
from sklearn.preprocessing import OneHotEncoder
```

```
ohe = OneHotEncoder(drop='first', sparse_output=False, dtype=np.int32)
```

```
x_train_new = ohe.fit_transform(x_train[['fuel','owner']])
```

```
x_train_new
```

```
array([[0, 0, 1, ..., 0, 0, 0],
       [1, 0, 0, ..., 1, 0, 0],
       [0, 0, 1, ..., 0, 0, 0],
       ...,
       [1, 0, 0, ..., 0, 0, 0],
       [0, 0, 1, ..., 1, 0, 0],
       [0, 0, 1, ..., 0, 0, 0]], dtype=int32)
```

```
x_test_new = ohe.transform(x_test[['fuel','owner']])
```

```
x_train_new.shape
```

```
(6502, 7)
```

```
np.hstack((x_train[['brand','km_driven']].values,x_train_new))
```

```
array([['Hyundai', 60000, 0, ..., 0, 0, 0],
       ['Nissan', 82000, 1, ..., 1, 0, 0],
       ['Tata', 20000, 0, ..., 0, 0, 0],
       ...,
       ['Mahindra', 75000, 1, ..., 0, 0, 0],
       ['Maruti', 26000, 0, ..., 1, 0, 0],
       ['Ford', 9500, 0, ..., 0, 0, 0]], dtype=object)
```

Most Frequent Categories

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

```
counts
```

brand	count
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
Land	6
Force	6
Isuzu	5
Ambassador	4
Kia	4
MG	3
Daewoo	3
Ashok	1

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

```
repl = counts[counts >= threshold].index
```

```
df = df[df["brand"].isin(repl)]
```

```
df.head()
```

	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

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

brand	count
Maruti	2448
Hyundai	1415
Mahindra	772
Tata	734
Toyota	488
Honda	467
Ford	397
Chevrolet	230
Renault	228
Volkswagen	186
BMW	120
Skoda	105

dtype: int64

```
Start coding or generate with AI.
```