EDA (Exploratory Data Analysis)

Thực hành trên bộ dữ liệu về xe hơi

Import thư viện

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

1. Hiểu tổng quan về bộ dữ liệu

1.1. Đọc dữ liệu từ file csv và hiển thị 5 dòng dữ liệu đầu tiên

In [2]:
```
# Đọc dữ liệu từ file csv
df = pd.read_csv('../data/cars_data.csv', header=0, delimiter=',', encoding='utf-8')
```

In [3]:
```
# Hiển thị 5 dòng dữ liệu đầu tiên
df.head(5)
```

Out[3]:

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Ma
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Tun
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxu
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxu
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	

1.2. Hiển thị 5 dòng dữ liệu cuối cùng

In [4]:
```
# Hiển thị 5 dòng dữ liệu cuối cùng
df.tail(5)
```

Out[4]:

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Nu l
11909	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	
11910	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	
11911	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	
11912	Acura	ZDX	2013	premium unleaded (recommended)	300.0	6.0	AUTOMATIC	all wheel drive	
11913	Lincoln	Zephyr	2006	regular unleaded	221.0	6.0	AUTOMATIC	front wheel drive	

1.3. Hiển thị kiểu dữ liệu của từng cột

In [5]:
```
df.dtypes
```

Out[5]:
```
Make                object
Model               object
Year                int64
Engine Fuel Type    object
Engine HP           float64
Engine Cylinders    float64
Transmission Type   object
Driven_Wheels       object
Number of Doors     float64
Market Category     object
Vehicle Size        object
Vehicle Style       object
highway MPG         int64
city mpg            int64
Popularity          int64
MSRP                int64
dtype: object
```

2. Tiền xử lý dữ liệu

2.1. Xóa đi các cột không quan tâm

Hãy xóa các cột: 'Engine Fuel Type', 'Market Category', 'Vehicle Style', 'Popularity', 'Number of Doors', 'Vehicle Size' Sau đó, hiển thị lại 5 dòng dữ liệu đầu tiên

In [6]:
```
df.drop(['Engine Fuel Type', 'Market Category', 'Vehicle Style', 'Popularity', 'Number of Doors', 'Vehicle Size'
```

Out[6]:

	Make	Model	Year	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	highway MPG	city mpg	M
0	BMW	1 Series M	2011	335.0	6.0	MANUAL	rear wheel drive	26	19	46
1	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	19	40
2	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	20	36
3	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	29
4	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	34
...	...	...	...	...	...	...	...	...	...	
11909	Acura	ZDX	2012	300.0	6.0	AUTOMATIC	all wheel drive	23	16	46
11910	Acura	ZDX	2012	300.0	6.0	AUTOMATIC	all wheel drive	23	16	56
11911	Acura	ZDX	2012	300.0	6.0	AUTOMATIC	all wheel drive	23	16	50
11912	Acura	ZDX	2013	300.0	6.0	AUTOMATIC	all wheel drive	23	16	50
11913	Lincoln	Zephyr	2006	221.0	6.0	AUTOMATIC	front wheel drive	26	17	28

11914 rows × 10 columns

In [7]:
```
df.head(5)
```

Out[7]:

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Ma
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Tun
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxu
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxu
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	

2.2. Đổi tên các cột để ngắn gọn hơn

Đổi tên các cột như sau:

"Engine HP": "HP",

"Engine Cylinders": "Cylinders",

"Transmission Type": "Transmission",

"Driven_Wheels": "Drive Mode",

"highway MPG": "MPG-H",

"city mpg": "MPG-C",

"MSRP": "Price"

Sau đó, hiển thị lại 5 dòng dữ liệu đầu tiên

In [8]:
```
df = df.rename(columns={
    "Engine HP": "HP",
    "Engine Cylinders": "Cylinders",
    "Transmission Type": "Transmission",
    "Driven_Wheels": "Drive Mode",
    "highway MPG": "MPG-H",
    "city mpg": "MPG-C",
    "MSRP": "Price"
})
```

In [9]:
```
df.head(5)
```

Out[9]:

	Make	Model	Year	Engine Fuel Type	HP	Cylinders	Transmission	Drive Mode	Number of Doors	Market Categ
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Fact Tuner,Luxury,Hi Performar
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performar
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Hi Performar
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performar
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Lux

2.3. Xử lý dữ liệu trùng lặp

Hiển thị số dòng và số cột của bảng dữ liệu

In [10]:
```
df.shape
```

Out[10]: (11914, 16)

Hiển thị các dòng dữ liệu bị trùng

In [11]:
```
duplicates = df.duplicated()
print(df[duplicates])
```

```
       Make     Model  Year             Engine Fuel Type     HP  Cylinders  \
14      BMW  1 Series  2013  premium unleaded (required)  230.0        6.0
18     Audi       100  1992             regular unleaded  172.0        6.0
20     Audi       100  1992             regular unleaded  172.0        6.0
24     Audi       100  1993             regular unleaded  172.0        6.0
25     Audi       100  1993             regular unleaded  172.0        6.0
...     ...       ...   ...                          ...    ...        ...
11481  Suzuki     X-90  1998             regular unleaded   95.0        4.0
11603   Volvo     XC60  2017             regular unleaded  302.0        4.0
11604   Volvo     XC60  2017             regular unleaded  240.0        4.0
11708  Suzuki      XL7  2008             regular unleaded  252.0        6.0
11717  Suzuki      XL7  2008             regular unleaded  252.0        6.0

     Transmission        Drive Mode  Number of Doors  \
14         MANUAL   rear wheel drive              2.0
18         MANUAL  front wheel drive              4.0
20         MANUAL  front wheel drive              4.0
24         MANUAL  front wheel drive              4.0
25         MANUAL  front wheel drive              4.0
...           ...               ...              ...
11481      MANUAL   four wheel drive              2.0
11603   AUTOMATIC    all wheel drive              4.0
11604   AUTOMATIC  front wheel drive              4.0
11708   AUTOMATIC    all wheel drive              4.0
11717   AUTOMATIC  front wheel drive              4.0

                Market Category Vehicle Size Vehicle Style  MPG-H  MPG-C  \
14            Luxury,Performance      Compact        Coupe     28     19
18                       Luxury      Midsize        Sedan     24     17
20                       Luxury      Midsize        Sedan     24     17
24                       Luxury      Midsize        Sedan     24     17
25                       Luxury      Midsize        Sedan     24     17
...                         ...          ...          ...    ...    ...
11481                       NaN      Compact      2dr SUV     26     22
11603  Crossover,Luxury,Performance      Midsize      4dr SUV     29     20
11604             Crossover,Luxury      Midsize      4dr SUV     30     23
11708                    Crossover      Midsize      4dr SUV     22     15
11717                    Crossover      Midsize      4dr SUV     22     16

      Popularity  Price
14          3916  31500
18          3105   2000
20          3105   2000
24          3105   2000
25          3105   2000
...          ...    ...
11481        481   2000
11603        870  46350
11604        870  40950
11708        481  29149
11717        481  27499

[715 rows x 16 columns]
```

Hiển thị số lượng giá trị phân biệt (khác nhau đôi một) của từng cột dữ liệu

In [14]:
```
df.count()
```

Out[14]:
```
Make               11914
Model              11914
Year               11914
Engine Fuel Type   11911
HP                 11845
Cylinders          11884
Transmission       11914
Drive Mode         11914
Number of Doors    11908
Market Category     8172
Vehicle Size       11914
Vehicle Style      11914
MPG-H              11914
MPG-C              11914
Popularity         11914
Price              11914
dtype: int64
```

Xóa đi những dữ liệu bị trùng lặp

In [12]:
```
df.drop_duplicates(inplace=True)
```

Hiển thị lại số lượng giá trị phân biệt (khác nhau đôi một) của từng cột dữ liệu

In [14]:
```
df.nunique().to_frame()
```

Out[14]:

	0
Make	48
Model	915
Year	28
Engine Fuel Type	10
HP	356
Cylinders	9
Transmission	5
Drive Mode	4
Number of Doors	3
Market Category	71
Vehicle Size	3
Vehicle Style	16
MPG-H	59
MPG-C	69
Popularity	48
Price	6049

In [15]:
```
df.count()
```

Out[15]:
```
Make               11199
Model              11199
Year               11199
Engine Fuel Type   11196
HP                 11130
Cylinders          11169
Transmission       11199
Drive Mode         11199
Number of Doors    11193
Market Category     7823
Vehicle Size       11199
Vehicle Style      11199
MPG-H              11199
MPG-C              11199
Popularity         11199
Price              11199
dtype: int64
```

2.4. Xử lý dữ liệu rỗng

Hiển thị số lượng giá trị rỗng của từng cột dữ liệu

các cột có giá trị false tức là không rỗng, còn có giá trị true tức là giá trị đó rỗng

In [19]:
```
df.isnull()
```

Out[19]:

	Make	Model	Year	Engine Fuel Type	HP	Cylinders	Transmission	Drive Mode	Number of Doors	Market Category	V
0	False	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	
11909	False	False	False	False	False	False	False	False	False	False	
11910	False	False	False	False	False	False	False	False	False	False	
11911	False	False	False	False	False	False	False	False	False	False	
11912	False	False	False	False	False	False	False	False	False	False	
11913	False	False	False	False	False	False	False	False	False	False	

11914 rows × 16 columns

Xóa đi các giá trị rỗng, sau đó hiển thị lại số lượng giá trị phân biệt của từng cột.

In [16]:
```
df.dropna(how='any', inplace=True)
```

In [17]:
```
df
```

Out[17]:

	Make	Model	Year	Engine Fuel Type	HP	Cylinders	Transmission	Drive Mode	Number of Doors	
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Fa
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	I
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	
...	...	...	...	...	...	...	...	...	...	
11909	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11910	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11911	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11912	Acura	ZDX	2013	premium unleaded (recommended)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11913	Lincoln	Zephyr	2006	regular unleaded	221.0	6.0	AUTOMATIC	front wheel drive	4.0	

7735 rows × 16 columns

Hiển thị lại số lượng giá trị rỗng của từng cột dữ liệu

In [18]:
```
df.isna().sum()
```

Out[18]:
```
Make                0
Model               0
Year                0
Engine Fuel Type    0
HP                  0
Cylinders           0
Transmission        0
Drive Mode          0
Number of Doors     0
Market Category     0
Vehicle Size        0
Vehicle Style       0
MPG-H               0
MPG-C               0
Popularity          0
Price               0
dtype: int64
```

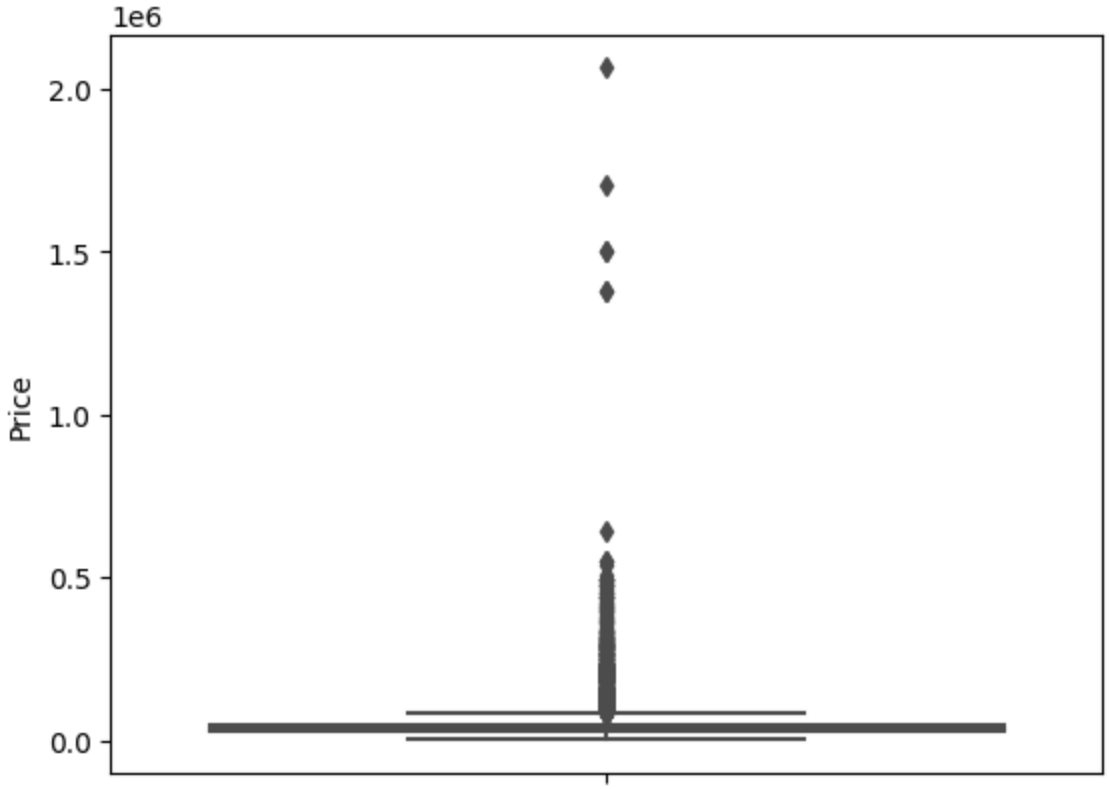
như vậy không còn giá trị rỗng trong từng cột

2.5. Xử lý outlier

Vẽ boxplot cho cột Price

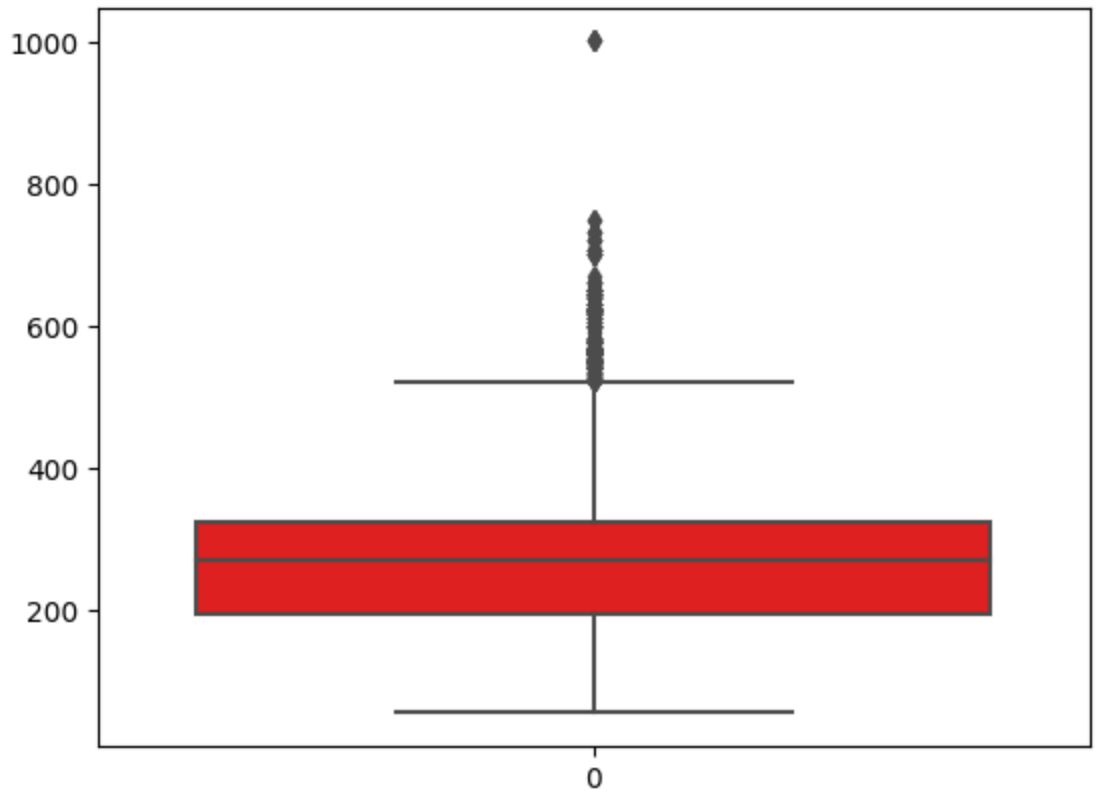
In [21]:
```
sns.boxplot(y=df['Price'], color='r')
plt.show()
```

```
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning:
is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, Categori
calDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```



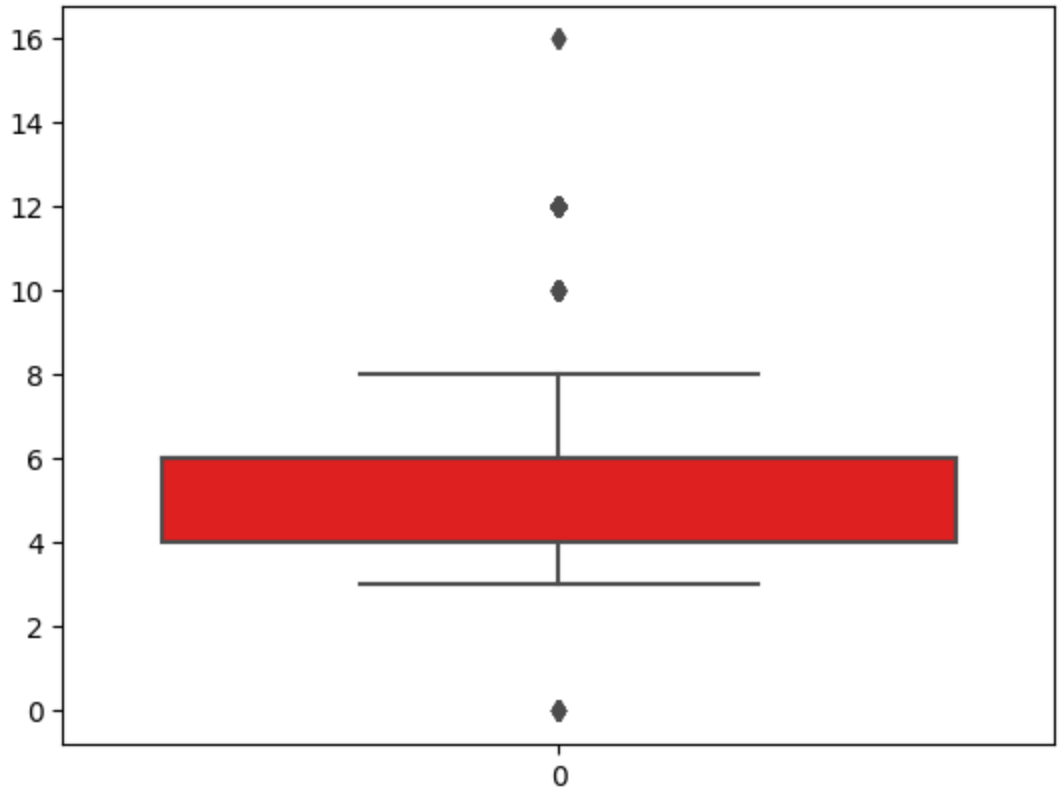
Vẽ boxplot cho cột HP

In [20]:
```
sns.boxplot(df['HP'], color='r')
plt.show()
```



Vẽ boxplot cho cột Cylinders

In [22]:
```
sns.boxplot(df['Cylinders'], color='r')
plt.show()
```



In [23]:
```
df
```

Out[23]:

	Make	Model	Year	Engine Fuel Type	HP	Cylinders	Transmission	Drive Mode	Number of Doors	
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Fa
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	l
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	
...	...	...	...	...	...	...	...	...	...	
11909	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11910	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11911	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11912	Acura	ZDX	2013	premium unleaded (recommended)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Cro
11913	Lincoln	Zephyr	2006	regular unleaded	221.0	6.0	AUTOMATIC	front wheel drive	4.0	

7735 rows × 16 columns

In [24]:
```
df.dtypes
```

Out[24]:
```
Make               object
Model              object
Year                int64
Engine Fuel Type   object
HP                float64
Cylinders         float64
Transmission       object
Drive Mode         object
Number of Doors   float64
Market Category    object
Vehicle Size       object
Vehicle Style      object
MPG-H               int64
MPG-C               int64
Popularity          int64
Price               int64
dtype: object
```

In [25]:
```
df['Year'].dtype
```

Out[25]:
```
dtype('int64')
```

Tìm độ trải giữa (IQR) của từng cột dữ liệu

In [26]:
```
for col in df.columns:
    if df[col].dtypes in ('int64','float64'):
        iqr = df[col].quantile(0.75) - df[col].quantile(0.25)
        print('Độ trải giữa của cột',col,'là',iqr)
```

```
Độ trải giữa của cột Year là 6.0
Độ trải giữa của cột HP là 131.0
Độ trải giữa của cột Cylinders là 2.0
Độ trải giữa của cột Number of Doors là 2.0
Độ trải giữa của cột MPG-H là 8.0
Độ trải giữa của cột MPG-C là 6.0
Độ trải giữa của cột Popularity là 1489.0
Độ trải giữa của cột Price là 23252.5
```

In [27]:
```
Q1, Q3 = df['Price'].quantile([0.25, 0.75])
IQR = Q3 - Q1
print("Độ trải giữa (IQR) của cột Price")
print(IQR)
```

```
Độ trải giữa (IQR) của cột Price
23252.5
```

In [28]:
```
Q1, Q3 = df['HP'].quantile([0.25, 0.75])
IQR = Q3 - Q1
print("Độ trải giữa (IQR) của cột HP")
print(IQR)
```

```
Độ trải giữa (IQR) của cột HP
131.0
```

In [29]:
```
Q1, Q3 = df['Cylinders'].quantile([0.25, 0.75])
IQR = Q3 - Q1
print("Độ trải giữa (IQR) của cột Cylinders")
print(IQR)
```

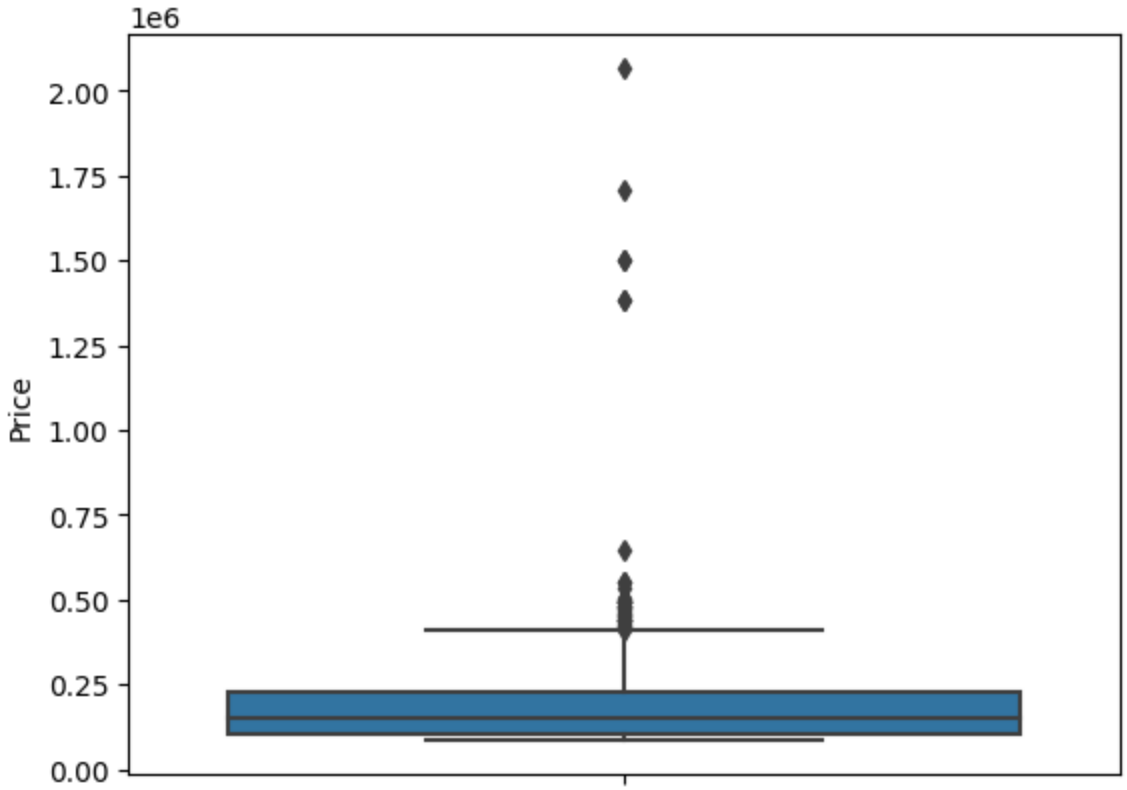
```
Độ trải giữa (IQR) của cột Cylinders
2.0
```

Loại bỏ đi outlier

In [30]:
```
# Xác định giá trị Q1 và Q3 của cột Price
Q1, Q3 = df['Price'].quantile([0.25, 0.75])
IQR = Q3 - Q1

# Tính toán giá trị lower bound và upper bound
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Loại bỏ các giá trị nằm ngoài lower bound và upper bound
df = df.loc[df['Price'] >= lower_bound]
df = df.loc[df['Price'] >= upper_bound]

# Vẽ biểu đồ boxplot
sns.boxplot(data=df, y='Price')
```

```
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```

Out[30]: <Axes: ylabel='Price'>

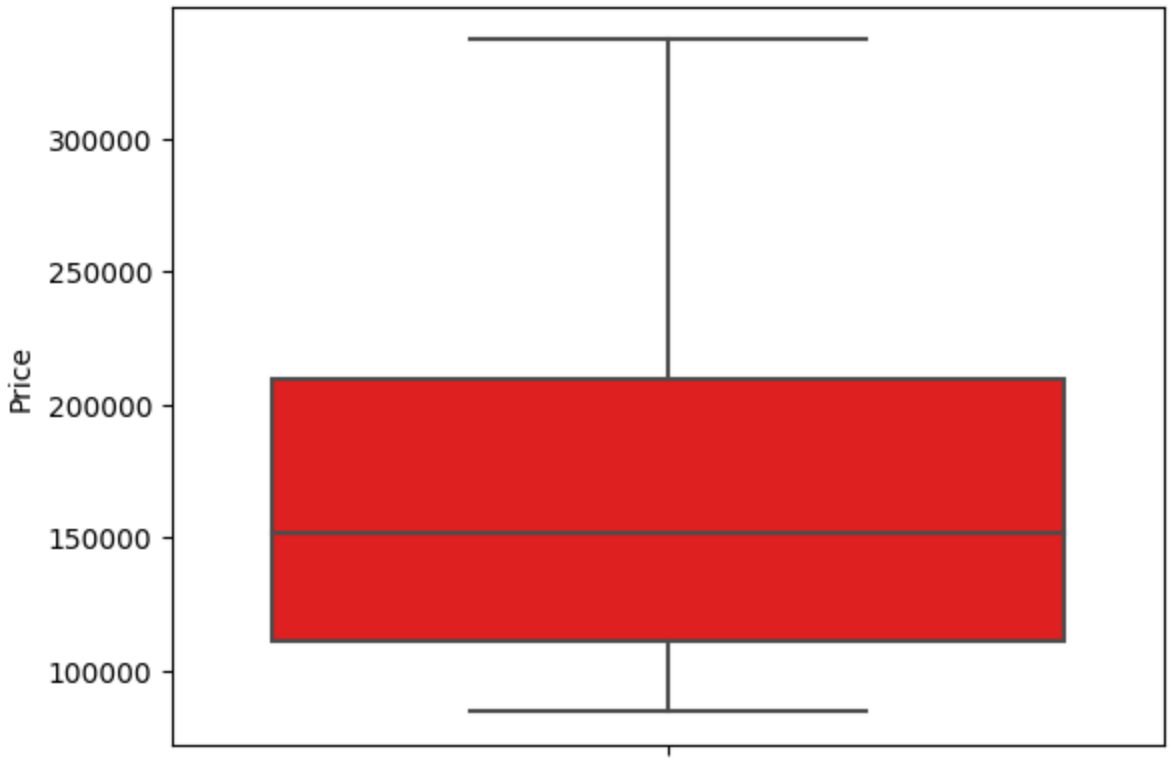


In [39]:
```
for col in df.columns:
    if df[col].dtypes in ('int64', 'float64'):
        q1 = df[col].quantile(0.25)
        q3 = df[col].quantile(0.75)
        iqr = q3-q1
        lower_bound = q1 - 1.5 * iqr
        upper_bound = q3 + 1.5 * iqr
        # print('Độ trải của cột',col,'là',iqr)
        # df = df[df[col].between(lower_bound,upper_bound)]
        df = df.loc[df[col] >= lower_bound]
        df = df.loc[df[col] <= upper_bound]
```

In [40]:
```
sns.boxplot(y=df['Price'], color='r')
```

```
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```

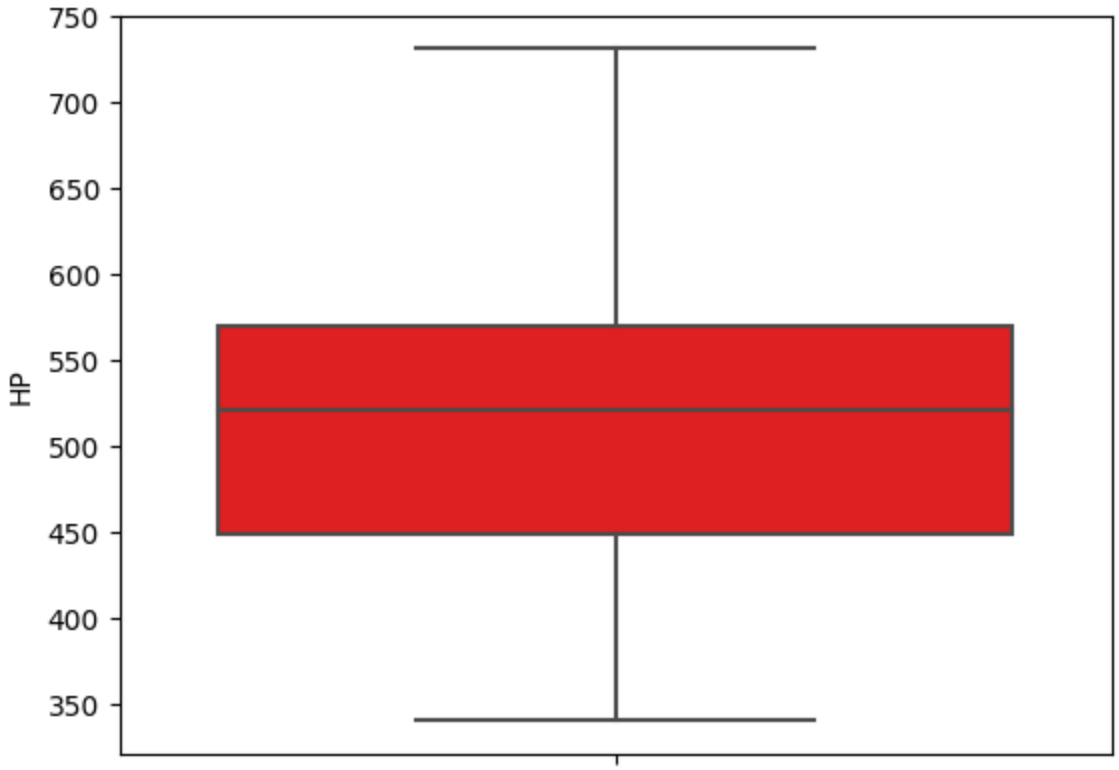
Out[40]: <Axes: ylabel='Price'>



In [41]:
```
sns.boxplot(y=df['HP'], color='r')
```

```
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```

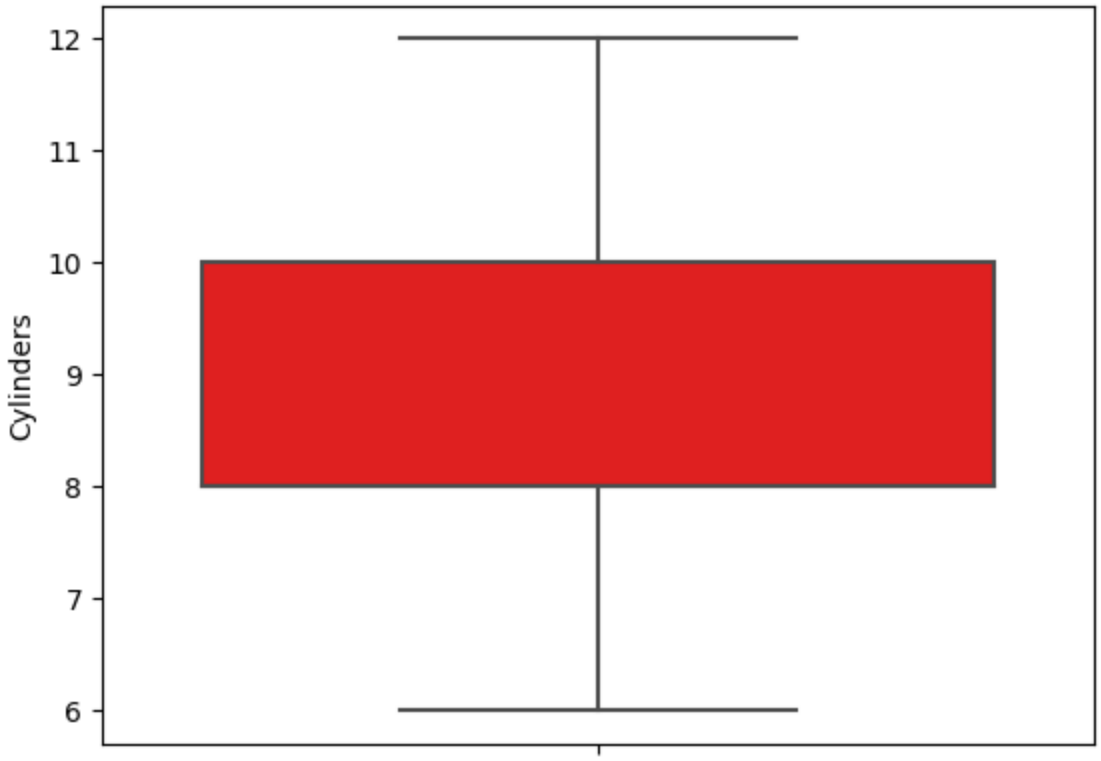
Out[41]: <Axes: ylabel='HP'>



In [42]:
```
sns.boxplot(y=df['Cylinders'], color='r')
```

```
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```

Out[42]: <Axes: ylabel='Cylinders'>

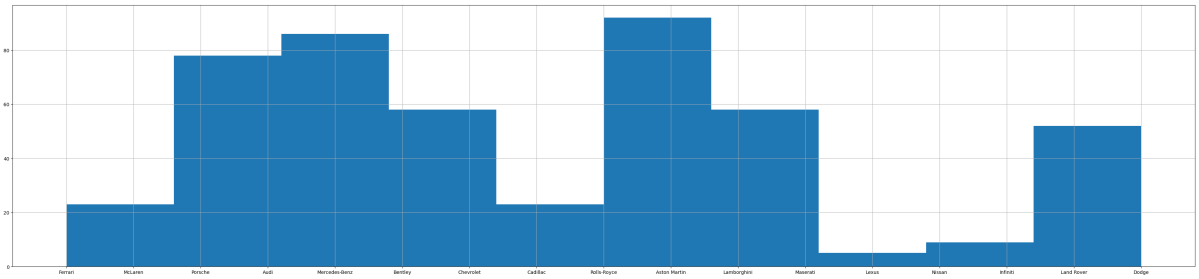


3. Trực quan hóa dữ liệu

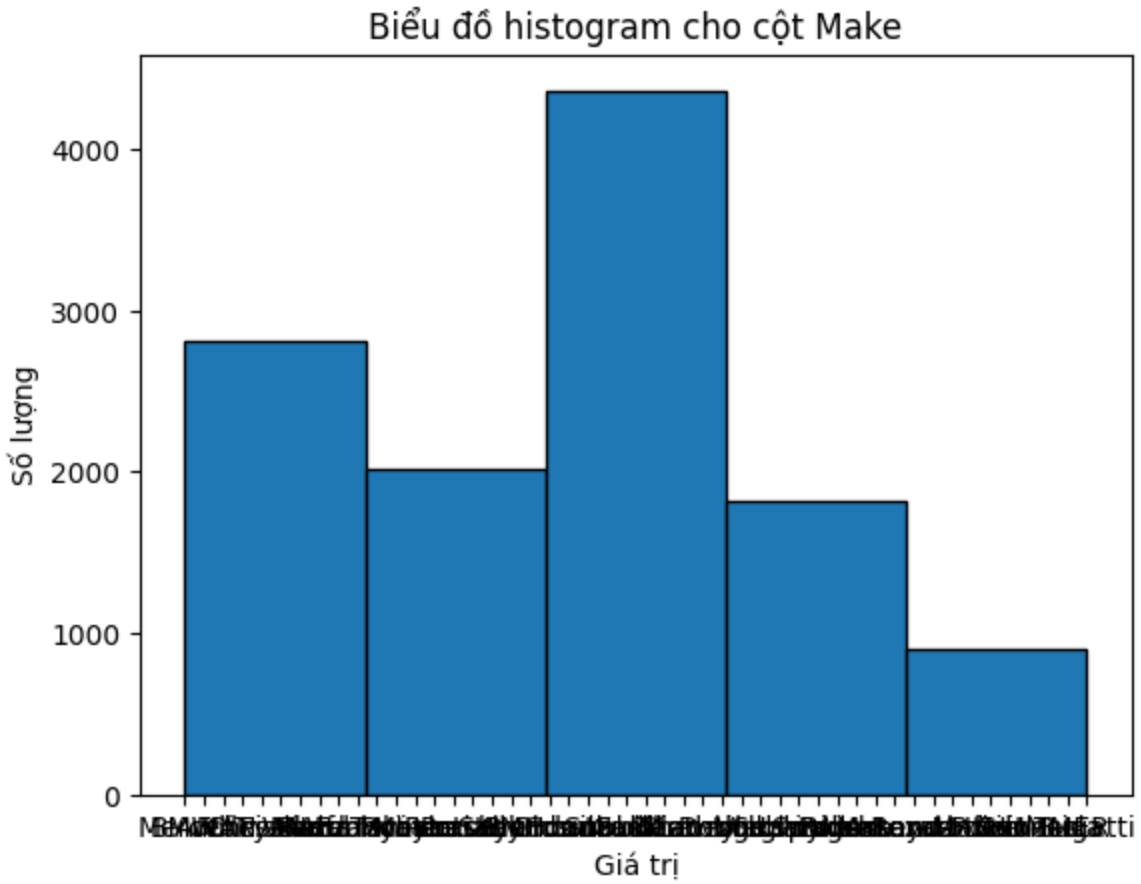
3.1. Vẽ histogram cho cột Make

In [43]:
```
plt.figure(figsize=(45, 10))
df['Make'].hist()
```

Out[43]: <Axes: >

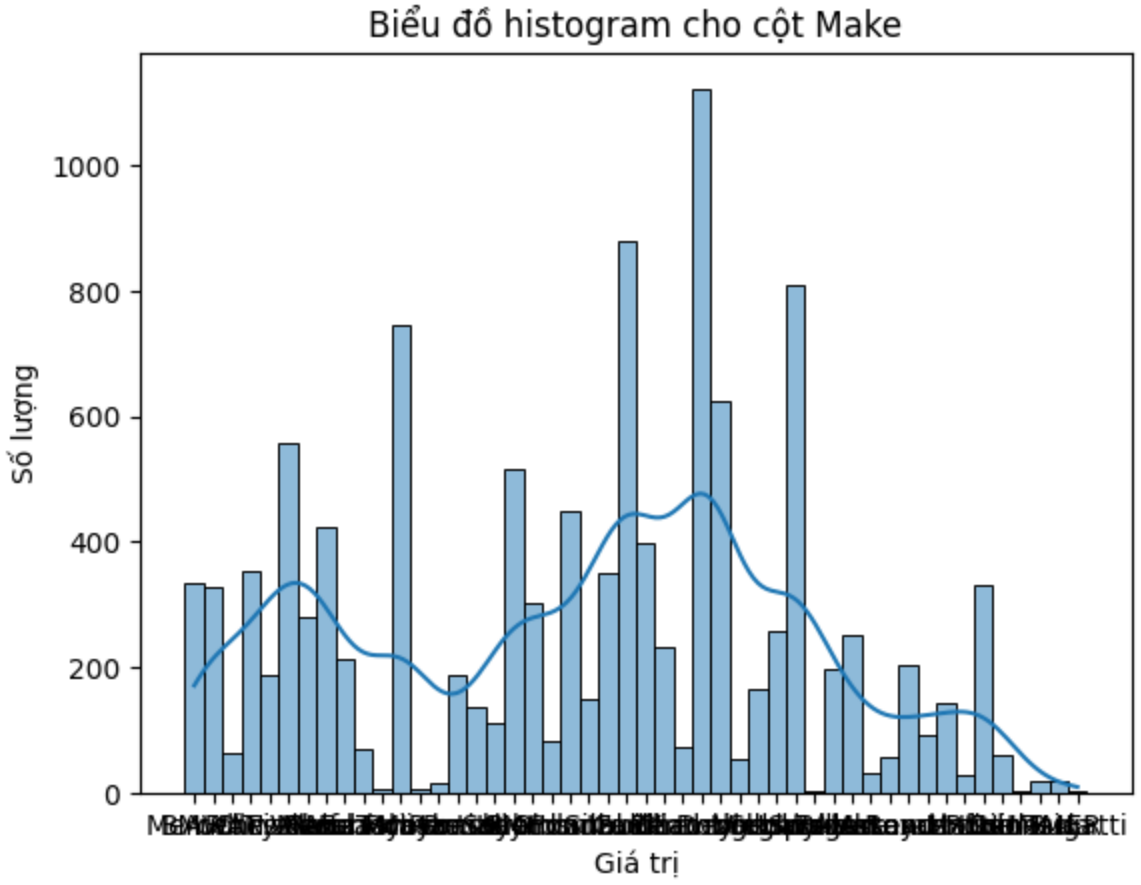


In [27]:
```
# Cách 1: Sử dụng Matplotlib

plt.hist(df['Make'], bins=5, edgecolor='k')
plt.title('Biểu đồ histogram cho cột Make')
plt.xlabel('Giá trị')
plt.ylabel('Số lượng')
plt.show()
```



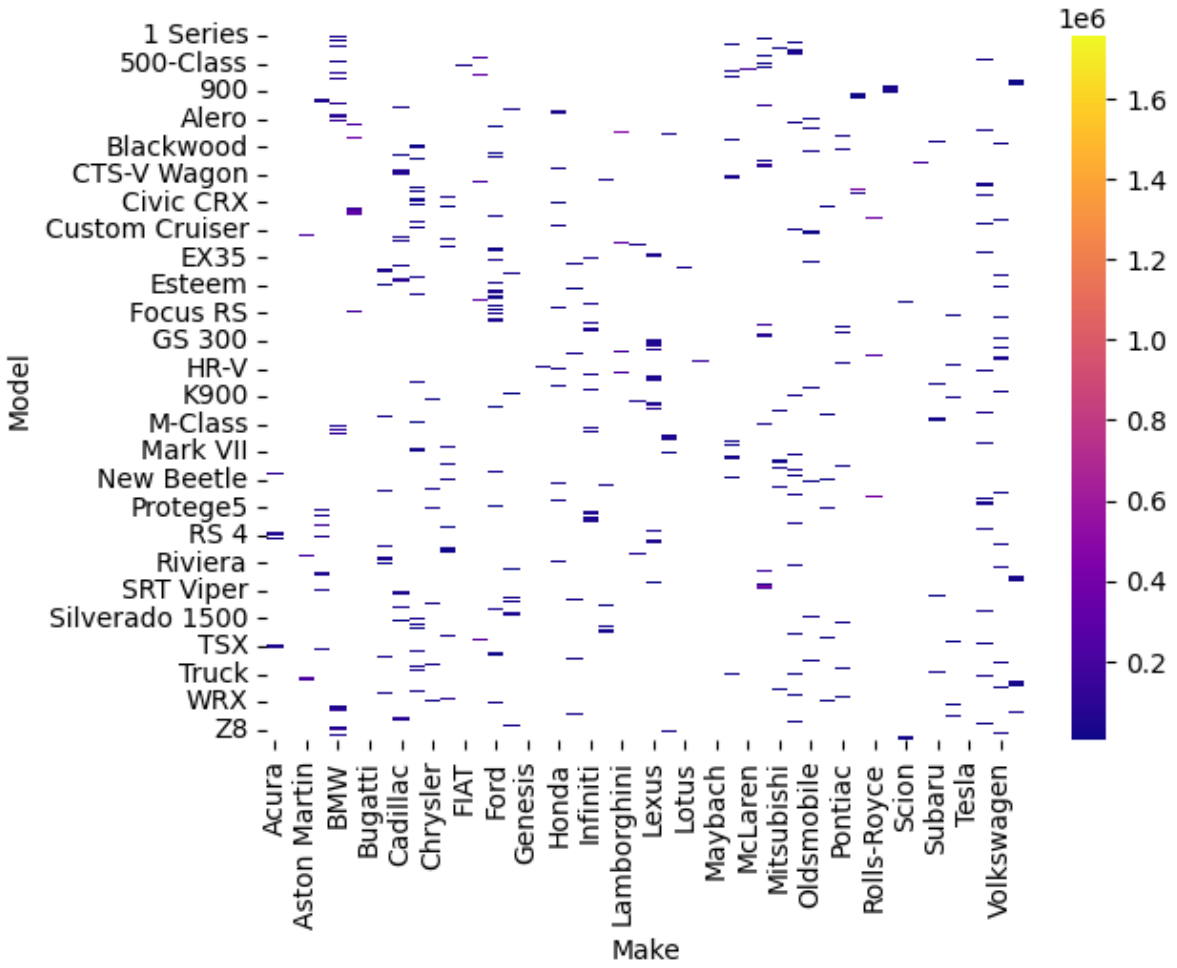
In [5]:
```
# Cách 2: Sử dụng Seaborn

sns.histplot(data=df['Make'], bins=5, kde=True)
plt.title('Biểu đồ histogram cho cột Make')
plt.xlabel('Giá trị')
plt.ylabel('Số lượng')
plt.show()
```

```
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
e:\LAP TRINH PYTHON\LT PTDL 1\venv\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



3.2. Vẽ biểu đồ nhiệt

In [13]:
```
sns.heatmap(df.pivot_table(values='Price', index='Model', columns='Make'), cmap="plasma")
plt.show()
```



3.3. Vẽ biểu đồ scatter cho 2 cột Price và HP

In [45]:
```
plt.scatter(df['Price'], df['HP'])
plt.show()
```