19120688

June 27, 2022

- MSSV: 19120688
- Họ và tên: Đỗ Nhật Toàn

1 Import các thư viện

[1]:
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
data = pd.read_csv('./ToyotaCorolla.csv', header=0)
data = data.dropna()
```

[2]:
```
data.shape
```

[2]:
```
(1436, 12)
```

[3]:
```
data.head()
```

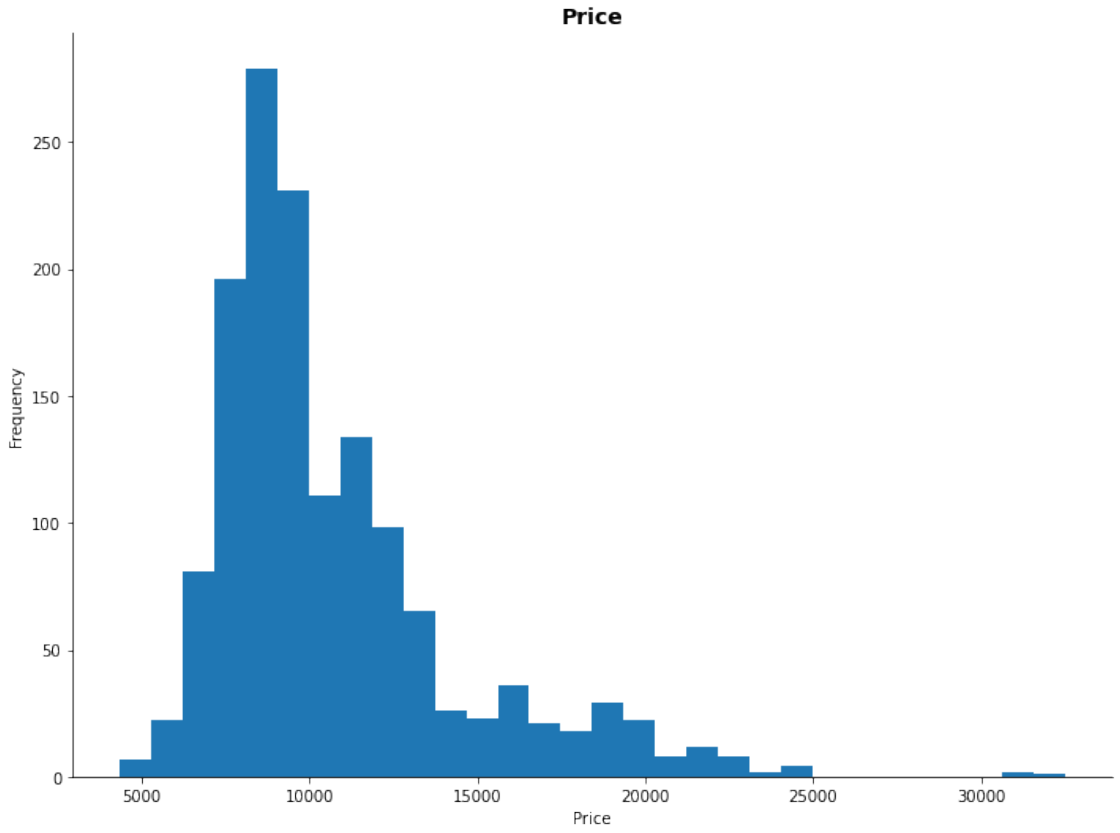
[3]:

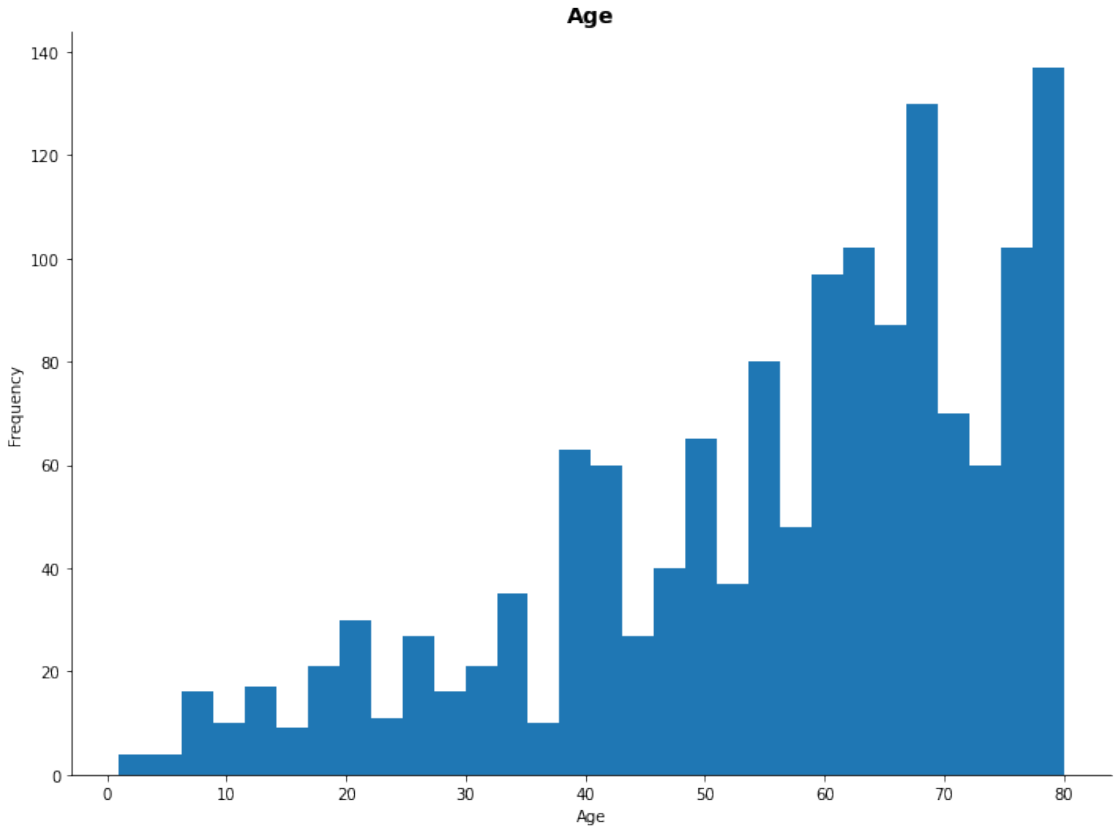
	Price	Age	Kilometers	Fuel_Type	HP	Metallic	Color	Automatic	CC	Doors	Quarterly_Tax	Weight
0	13500	23	46986	Diesel	90	1	Blue	0	2000	3	210	1165
1	13750	23	72937	Diesel	90	1	Silver	0	2000	3	210	1165
2	13950	24	41711	Diesel	90	1	Blue	0	2000	3	210	1165
3	14950	26	48000	Diesel	90	0	Black	0	2000	3	210	1165
4	13750	30	38500	Diesel	90	0	Black	0	2000	3	210	1170

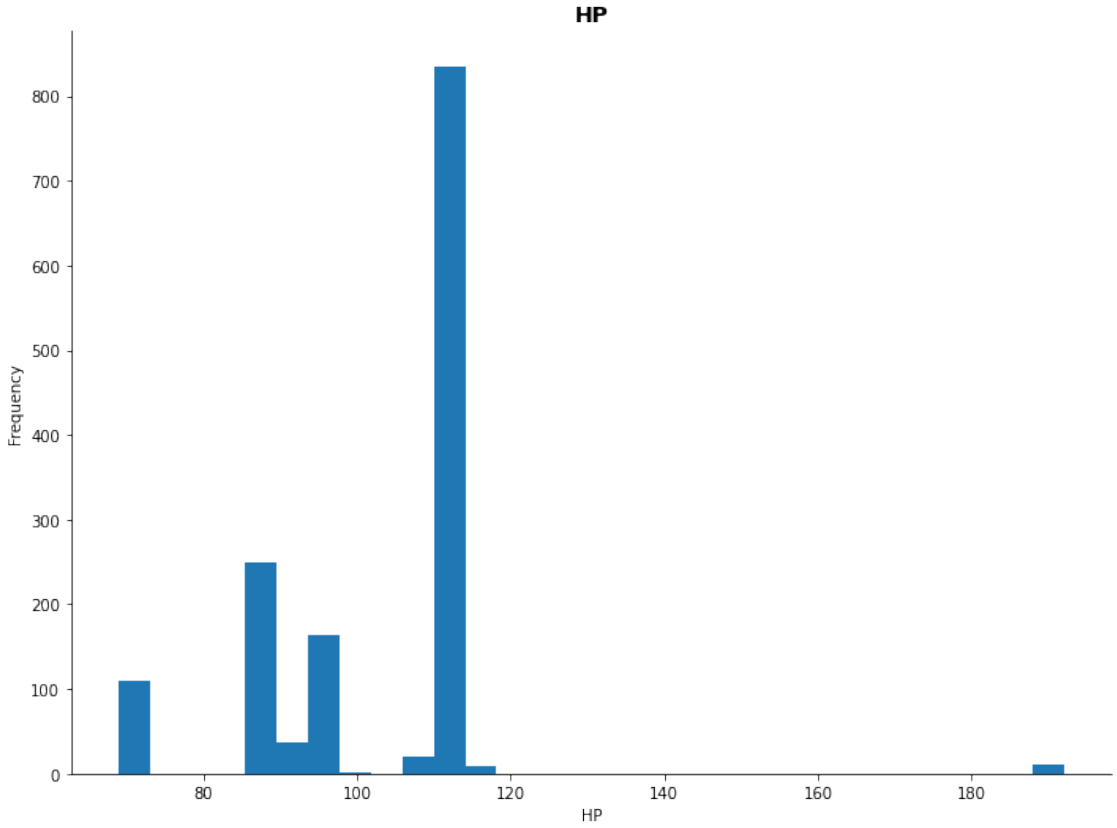
2 Hãy trực quan hóa các thông tin thống kê mô tả cho các biến

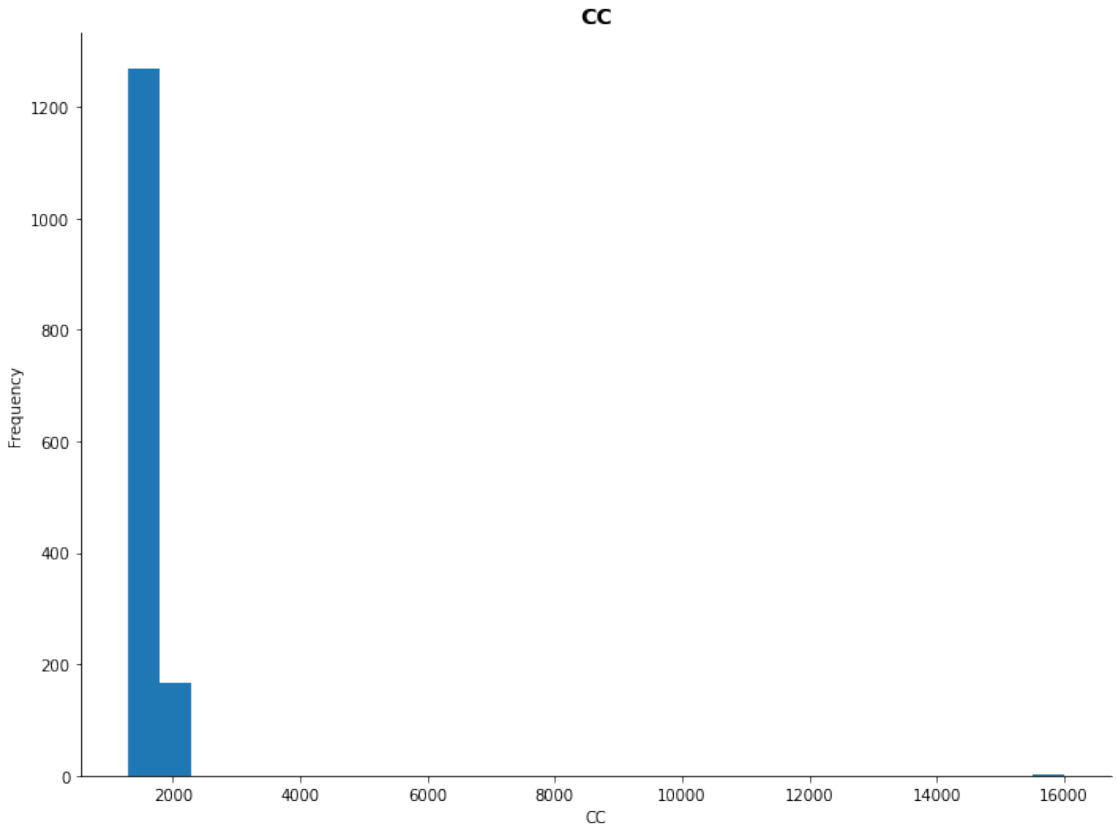
[4]:
```
atr_list = ['Price', 'Age', 'Kilometers', 'Fuel_Type', 'HP', 'Metallic',␣
 ↪'Color', 'Automatic', 'CC', 'Doors', 'Quarterly_Tax', 'Weight']
numeric_list = ['Price', 'Age', 'HP', 'CC', 'Doors', 'Quarterly_Tax', 'Weight']
categorical_list = ['Fuel_Type', 'Metallic', 'Automatic', 'Color']
```

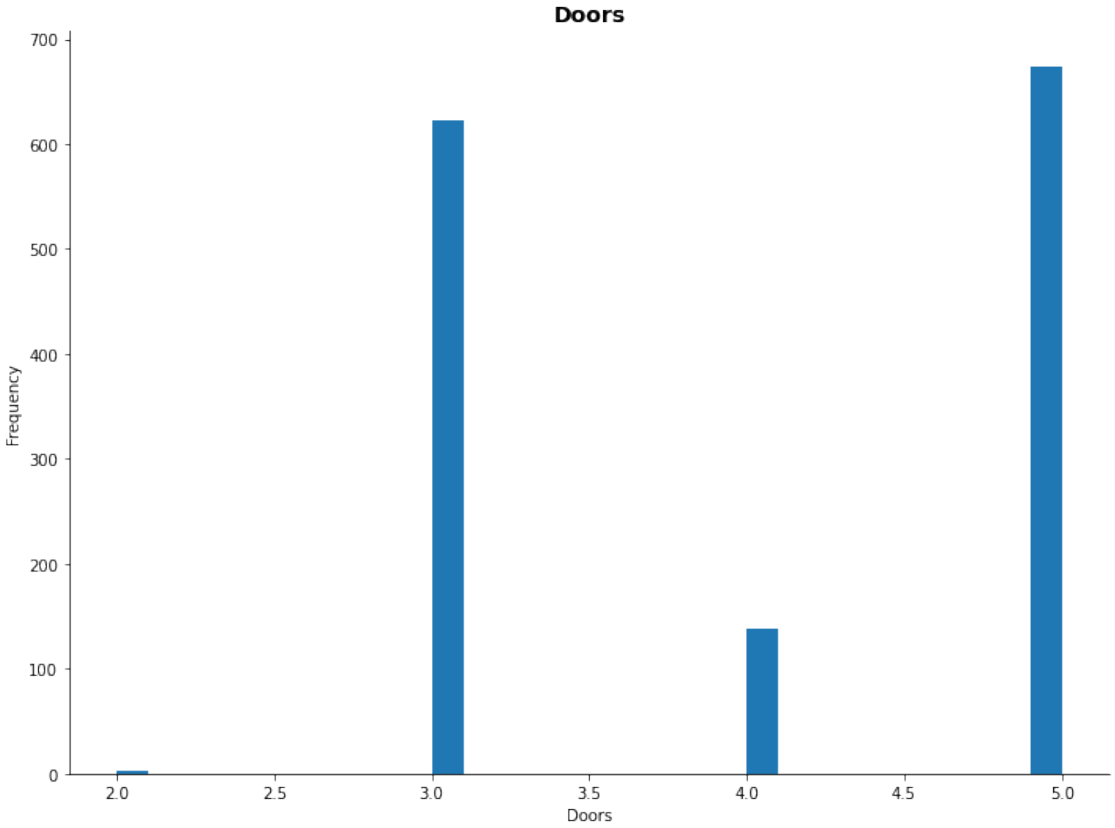
[5]:
```
for x in numeric_list:
    fig_obj = plt.figure(figsize=(10, 7.5))
    ax = plt.subplot(111)
    ax.spines["bottom"].set_visible(True)  # Set the spines, or box bounds␣
 ↪visibility
    ax.spines["left"].set_visible(True)
    ax.spines['right'].set_visible(False)
    ax.spines['top'].set_visible(False)
    ''' Plot the histogram of '''
    p = plt.hist(data[x], bins = 30)
    plt.title(x, fontsize=14, fontweight='bold')
    ''' Save figure '''
    plt.xlabel(x)
    plt.ylabel("Frequency")
    plt.tight_layout()
    plt.show()
```

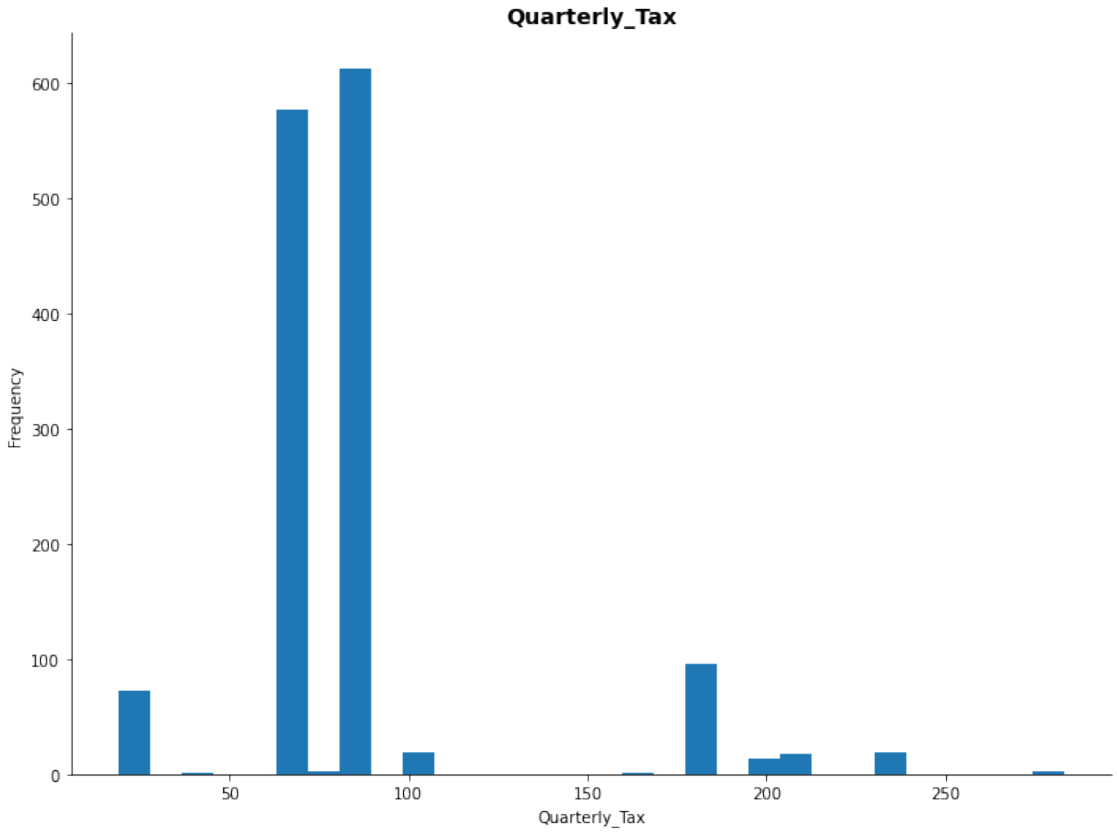


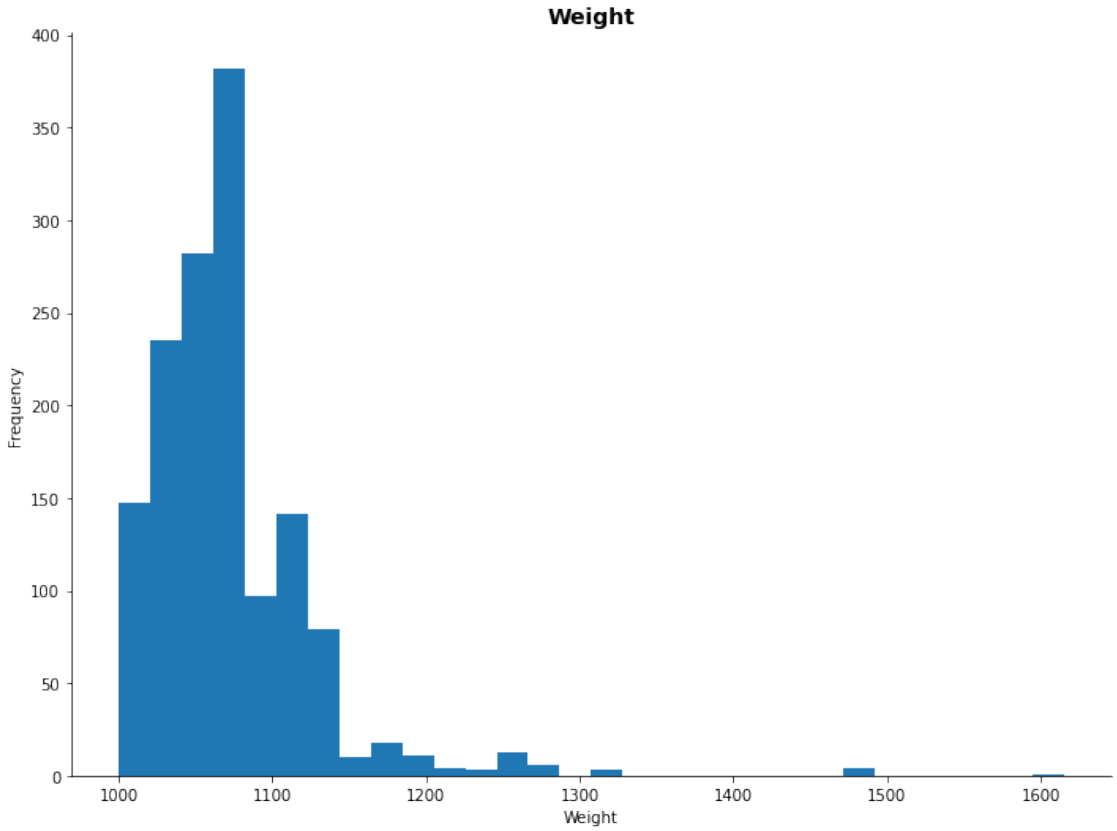

Age
Frequency
Age
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70
80


HP
Frequency
0
100
200
300
400
500
600
700
800
80
100
120
140
160
180
HP

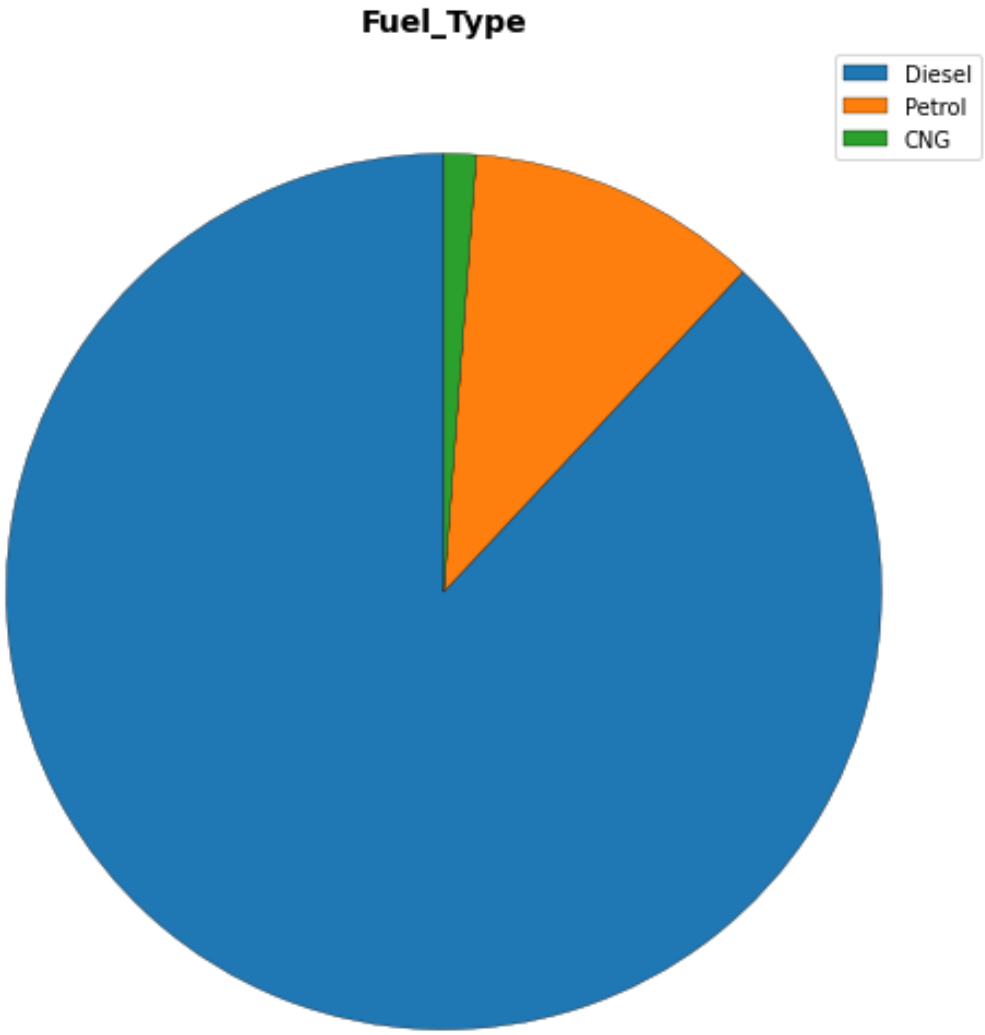

cc
Frequency
1200
1000
800
600
400
200
0
2000
4000
6000
8000
10000
12000
14000
16000
CC

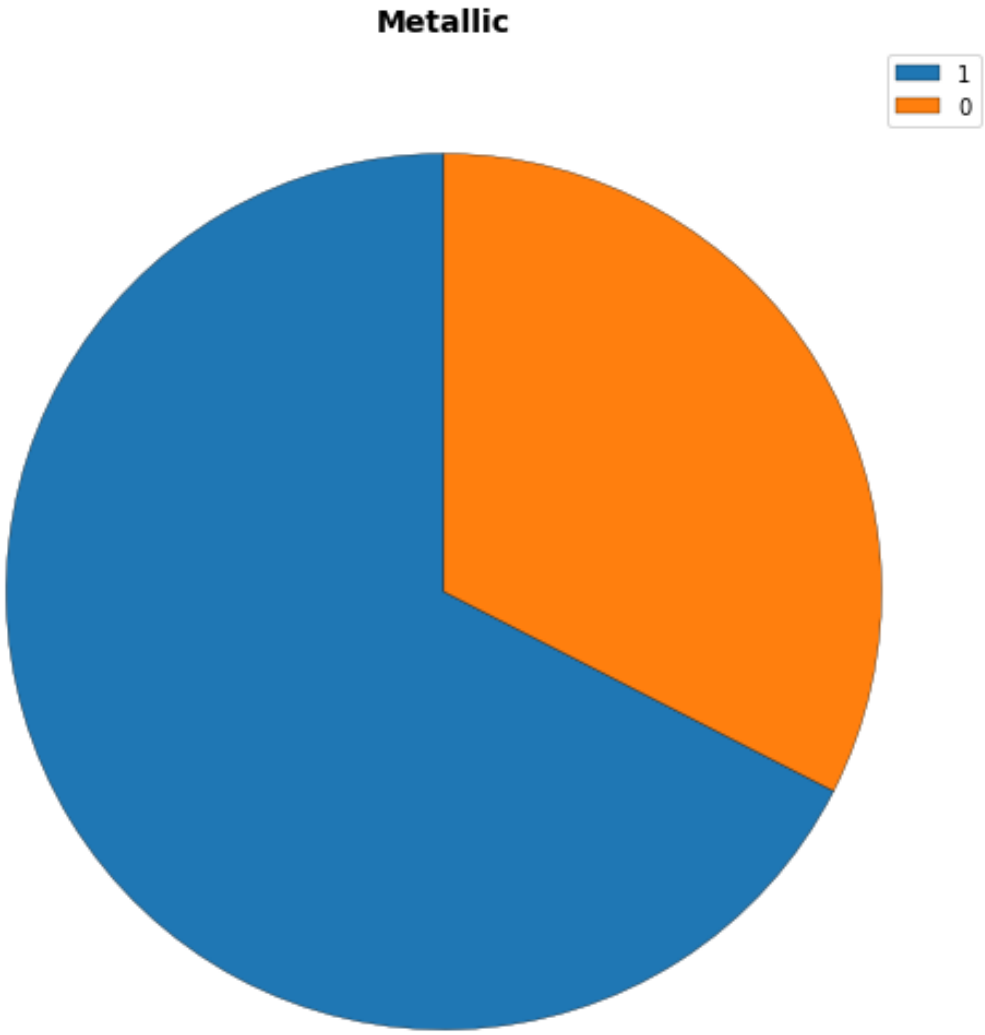

Doors
700
600
500
400
300
200
100
0
Frequency
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Doors

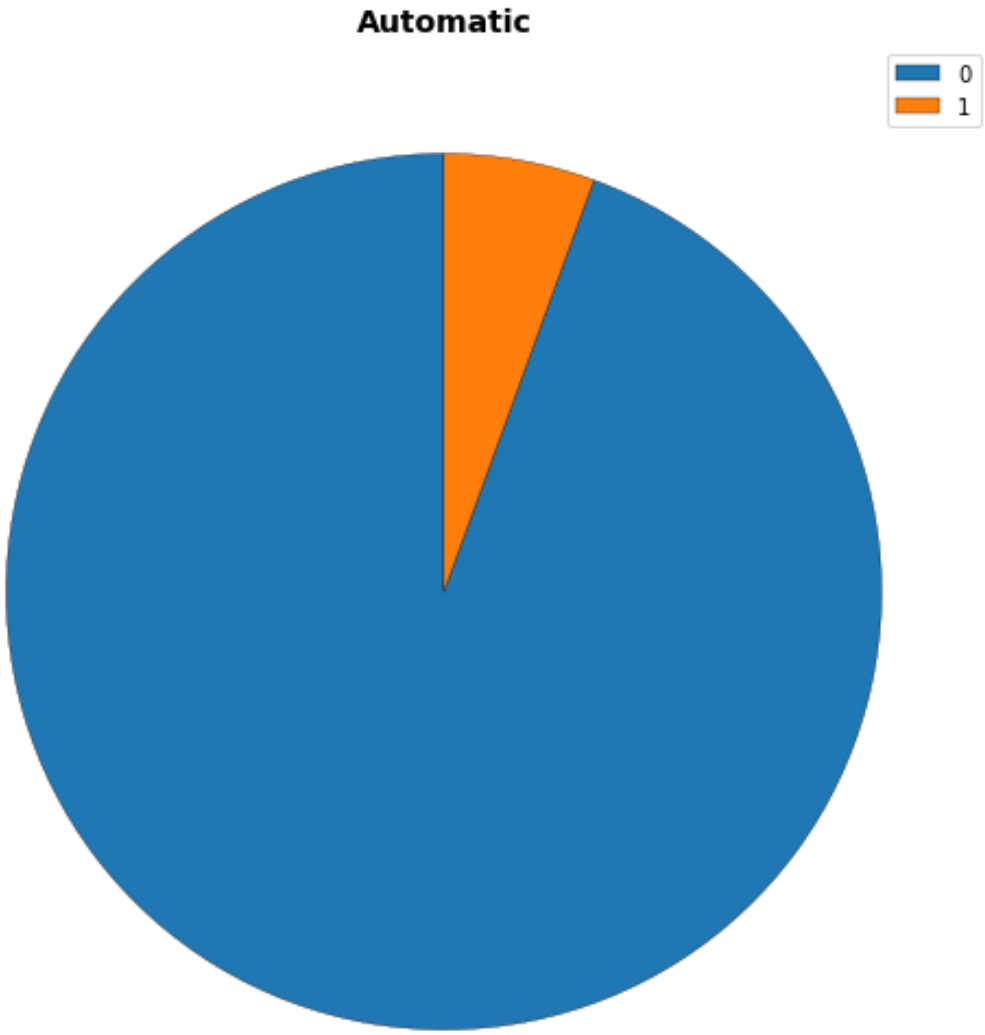

Quarterly_Tax
Frequency
600
500
400
300
200
100
0
50
100
150
200
250
Quarterly_Tax



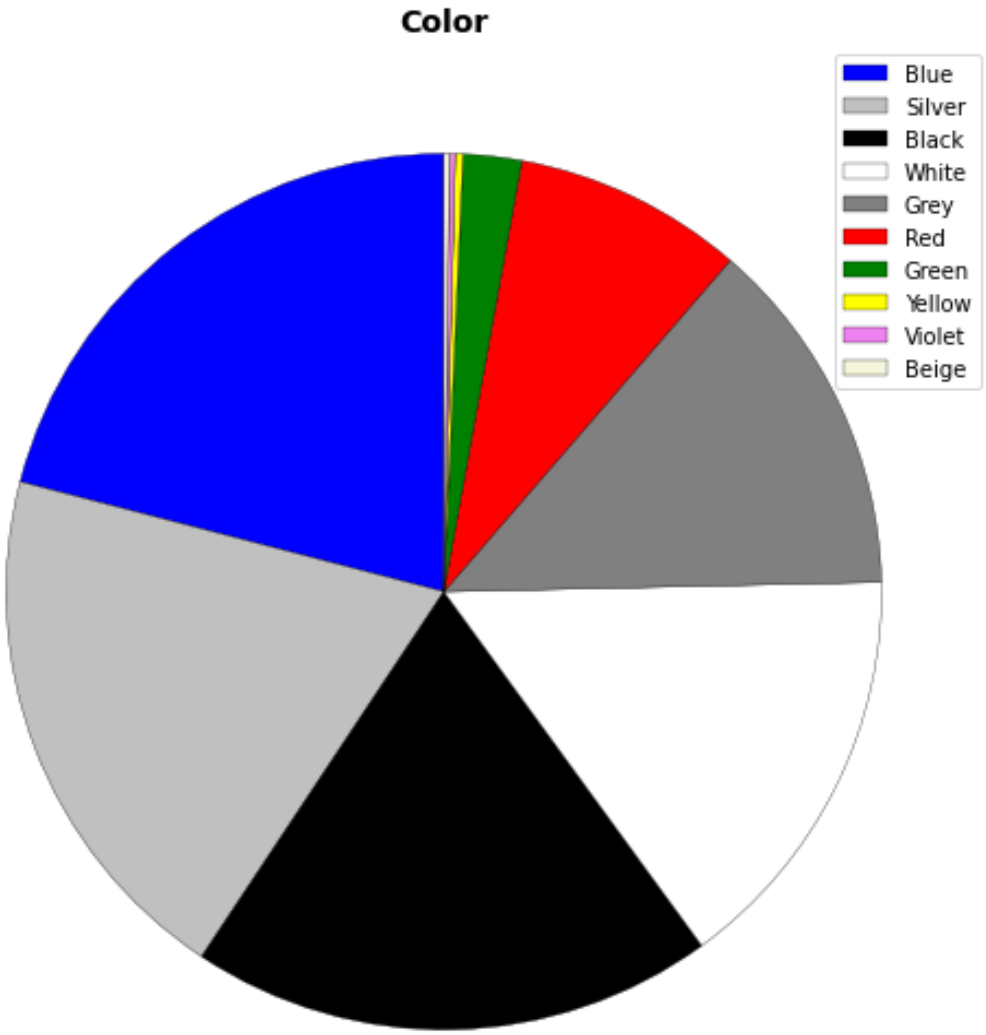
[6]:
```
for x in categorical_list[:-1]:
    fig_obj = plt.figure(figsize=(10, 7.5))
    fig_obj.set_facecolor('white')
    value = data[x].value_counts()
    label = data[x].unique().tolist()
    ax = plt.subplot(111)
    plt.title(x, fontsize=14, fontweight='bold')
    plt.pie(value,  startangle=90, wedgeprops = {"edgecolor" : "black",
                        'linewidth': 0.3,
                        'antialiased': True})
    plt.legend(label,loc=1)
    plt.tight_layout()
    plt.show()
```


Fuel_Type
Diesel
Petrol
CNG


Metallic
1
0



[7]:
```
fig_obj = plt.figure(figsize=(10, 7.5))
fig_obj.set_facecolor('white')
value = data['Color'].value_counts()
label = data['Color'].unique().tolist()
ax = plt.subplot(111)
plt.title('Color', fontsize=14, fontweight='bold')
plt.pie(value, colors=label, startangle=90, wedgeprops = {"edgecolor" : "black",
                      'linewidth': 0.3,
                      'antialiased': True})
plt.legend(label,loc=1)
plt.tight_layout()
plt.show()
```

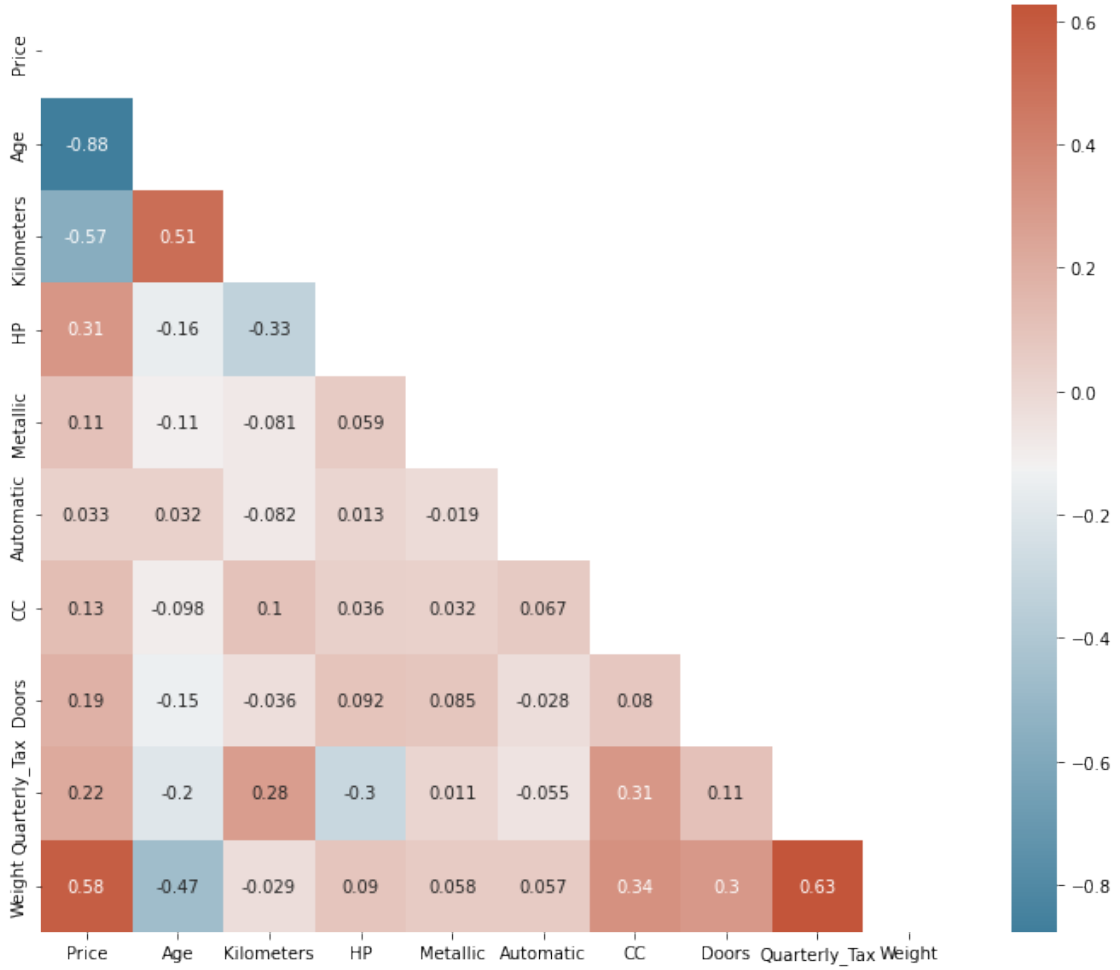


[8]:
```
# fig_obj = plt.figure(figsize=(10, 7.5))
    # ax = plt.subplot(111)
    # ax.spines["bottom"].set_visible(True)  # Set the spines, or box bounds␣
 ↪visibility
    # ax.spines["left"].set_visible(True)
    # ax.spines['right'].set_visible(False)
    # ax.spines['top'].set_visible(False)
    # ''' Plot the histogram of '''
    # p = plt.bar(data[x], bins = 40)
    # plt.title(x, fontsize=14, fontweight='bold')
    # ''' Save figure '''
    # plt.tight_layout()
```

3 Tìm và trực quan mối quan hệ tương quan giữa các cặp biến (nếu có)

[9]:
```
corr = data.corr(method="pearson")
f, ax = plt.subplots(figsize=(12, 10))
mask = np.triu(np.ones_like(corr, dtype=bool))
cmap = sns.diverging_palette(230, 20, as_cmap=True)
sns.heatmap(corr, annot=True, mask = mask, cmap=cmap)
```

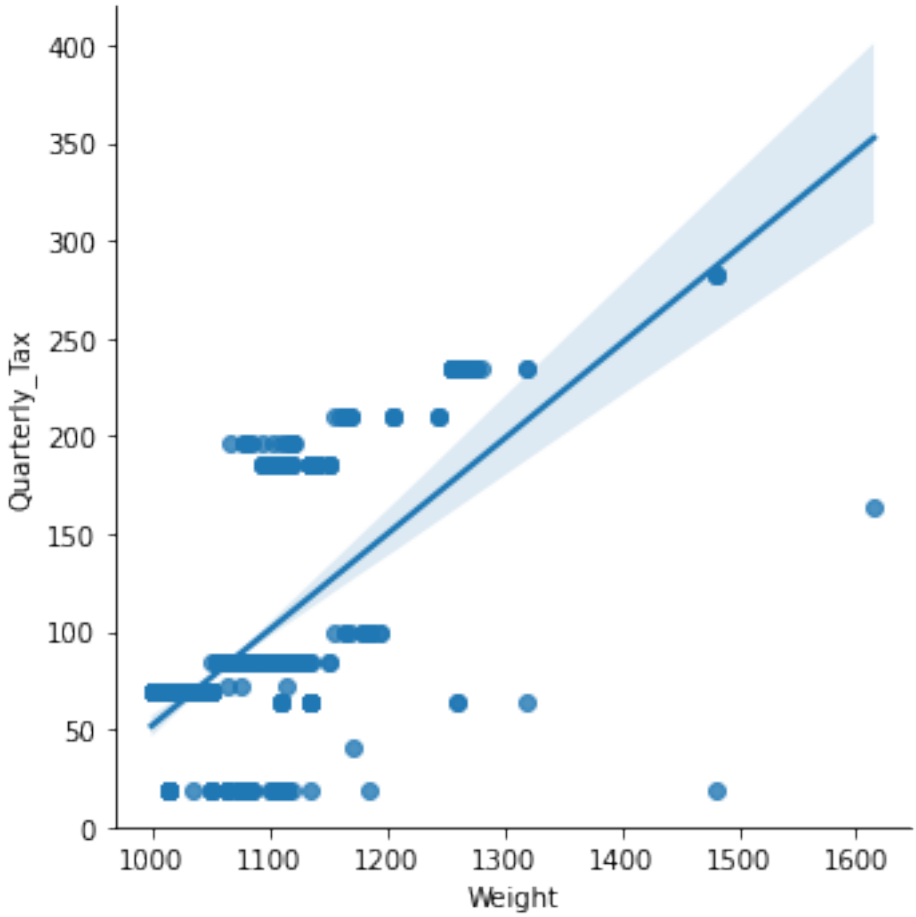
[9]: <AxesSubplot:>



3.1 Weight vs. Quarterly_Tax

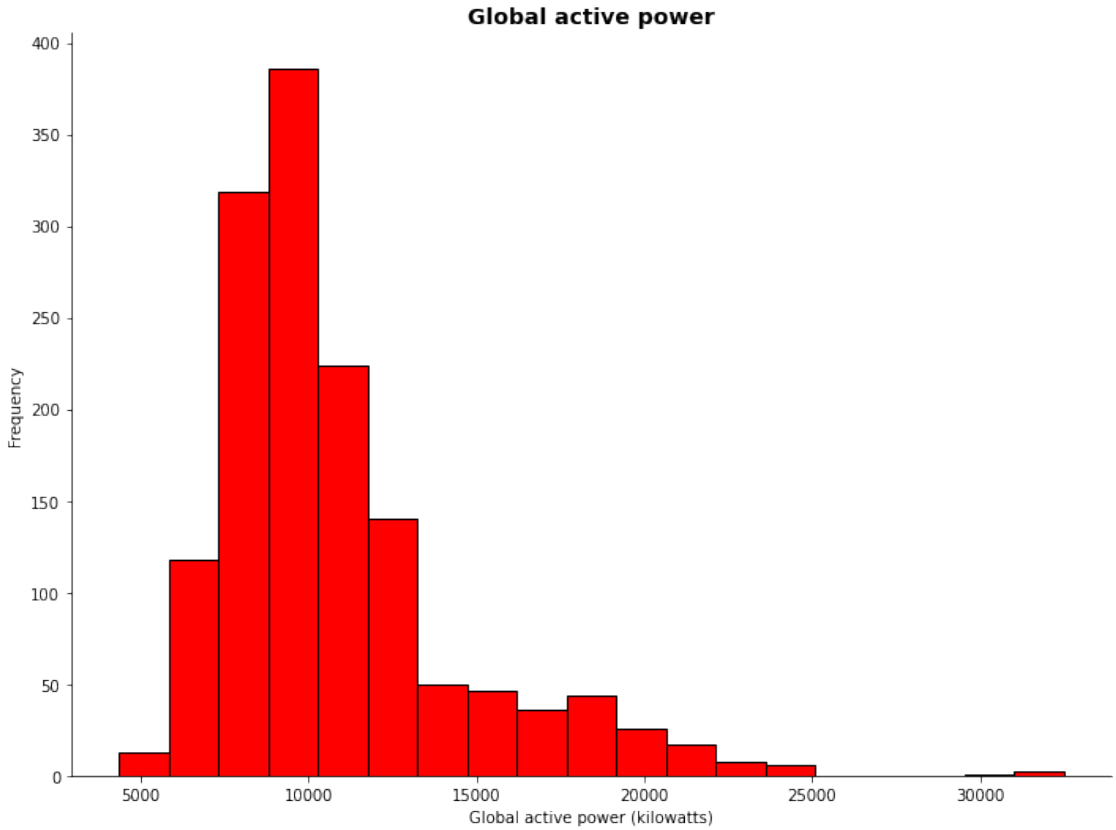
[10]:
```
sns.lmplot(x="Weight",y="Quarterly_Tax", data=data, order=1)
```

[10]: <seaborn.axisgrid.FacetGrid at 0x19b1c1af6d0>



4 Hãy trực quan hóa biểu đồ histogram cho Price theo từng biến biến theo Fuel_type và Color

[11]:
```
fig_obj = plt.figure(figsize=(10, 7.5))
ax = plt.subplot(111)
ax.spines["bottom"].set_visible(True)  # Set the spines, or box bounds␣
 ↪visibility
ax.spines["left"].set_visible(True)
ax.spines['right'].set_visible(False)
ax.spines['top'].set_visible(False)
p = plt.hist(data.Price, bins = 19, color="red", edgecolor='black')
plt.title("Global active power", fontsize=14, fontweight='bold')
plt.xlabel("Global active power (kilowatts)")
plt.ylabel("Frequency")
plt.tight_layout()
```


Global active power
Frequency
400
350
300
250
200
150
100
50
0
5000
10000
15000
20000
25000
30000
Global active power (kilowatts)