Chapter 6 - Exercise 7: Area plot, Boxplot

Thực hành vẽ Area plot, Box plot trên 2 tập dữ liệu khác nhau.

Part 1: Area Plot

```
import pandas as pd
import matplotlib.pyplot as plt


# Cho Dữ liệu Số giờ nắng các tháng trong năm 2016, 2017 tại trạm quan trắc Vũng Tàu:
df = pd.DataFrame(
    {
        'Month': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12],
        'Hours_2017': [183.4, 211.8, 286.4, 287.5, 238.8, 200.3, 187.4, 233.8, 225.5, 149.
        'Hours_2016': [272.8, 254.0, 296.0, 298.0, 240.1, 197.8, 240.3, 219.5, 212.7, 134.
    }
)


# Hiển thị nội dung của df
df
```

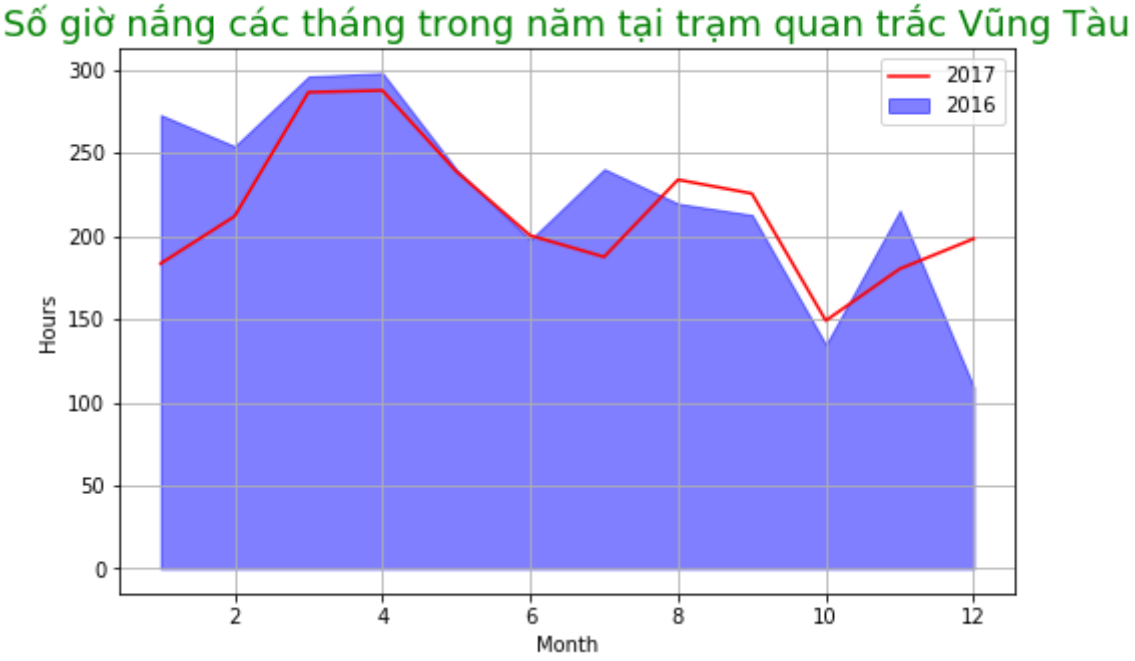
	Month	Hours_2017	Hours_2016
0	1	183.4	272.8
1	2	211.8	254.0
2	3	286.4	296.0
3	4	287.5	298.0
4	5	238.8	240.1
5	6	200.3	197.8
6	7	187.4	240.3
7	8	233.8	219.5
8	9	225.5	212.7
9	10	149.1	134.7
10	11	180.2	215.3
11	12	198.3	109.1

```
# Trên cùng một biểu đồ, hãy vẽ:
#        Area plot cho 12 tháng nắng trong năm 2016
#        Line plot cho 12 tháng nắng trong năm 2017
```

```
# Bạn nhận xét gì về biểu đồ vừa vẽ

plt.figure(figsize=(8,5))
plt.fill_between(df.Month, df.Hours_2016, color='blue', label = '2016', alpha=0.5)
plt.plot(df.Month, df.Hours_2017,color='red', label='2017')

plt.title("Số giờ nắng các tháng trong năm tại trạm quan trắc Vũng Tàu", fontsize=18, colo
plt.xlabel("Month")
plt.ylabel("Hours")
plt.legend()
plt.grid(True)
plt.show()
```



Part 2: Boxplot

```
# Cho dữ liệu baseball.csv. Đọc dữ liệu từ baseball.csv và lưu vào biến data, hiển thị 10
data = pd.read_csv(r'data\baseball.csv', index_col=0)
data.head(10)
```

	height	weight
0	1.8796	81.646560

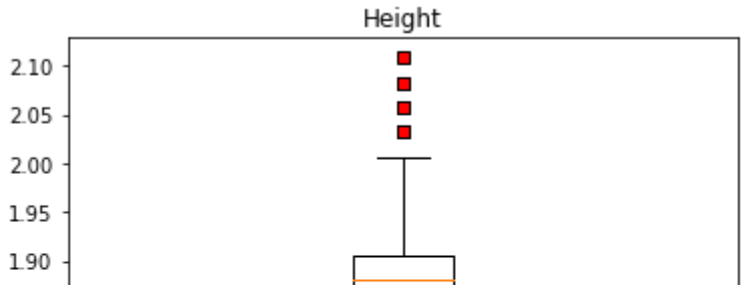
```
# Cho biết thông tin thống kê chung của data
data.describe()
```

	height	weight
count	1015.000000	1015.000000
mean	1.871717	91.330191
std	0.058774	9.445198
min	1.701800	68.038800
25%	1.828800	84.368112
50%	1.879600	90.718400
75%	1.905000	97.522280
max	2.108200	131.541680

```
# Vẽ boxplot cho dữ liệu height và weight
red_square = dict(markerfacecolor = 'r', marker = 's')

height = plt.boxplot(data.height, flierprops=red_square)
plt.title("Height")
plt.show()

weight = plt.boxplot(data.weight, flierprops=red_square)
plt.title('Weight')
plt.show()
```



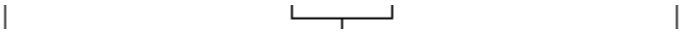
Kiểm tra xem dữ liệu có outliers hay không? Nếu có thì loại bỏ các outliers. Vẽ lại boxplot


```
# Tìm, đếm các outliers
```



```
# height
Q1_H = data.height.quantile(0.25)
Q3_H = data.height.quantile(0.75)
IQR_H = Q3_H - Q1_H
IQR_H
```

```
0.07620000000000005
```



```
H_lower_bound = Q1_H - (1.5 * IQR_H)
H_lower_bound
```

```
1.7145
```

```
count_H_lower_ouliters =  data.height[data.height < H_lower_bound].count()
count_H_lower_ouliters
```

```
2
```

```
H_upper_bound = Q3_H + (1.5 * IQR_H)
H_upper_bound
```

```
2.0193000000000003
```

```
count_H_upper_ouliters = data.height[data.height > H_upper_bound].count()
count_H_upper_ouliters
```

```
10
```

```
# weight
Q1_W = data.weight.quantile(0.25)
Q3_W = data.weight.quantile(0.75)
IQR_W = Q3_W - Q1_W
print(IQR_W)
```

```
13.154168000000013
```

```
W_lower_bound = Q1_W - (1.5 * IQR_W)
W_lower_bound
```

```
64.63685999999998
```

```
count_W_lower_ouliters =  data.weight[data.weight < W_lower_bound].count()
count_W_lower_ouliters
```

```
0
```

```
W_upper_bound = Q3_W + (1.5 * IQR_W)
W_upper_bound
```

```
117.25353200000004
```

```
count_W_upper_ouliters = data.weight[data.weight > W_upper_bound].count()
count_W_upper_ouliters
```

```
7
```

```
# Loại bỏ các outliers
result = data
result.head(10)
```

	height	weight
0	1.8796	81.646560
1	1.8796	97.522280
2	1.8288	95.254320
3	1.8288	95.254320
4	1.8542	85.275296
5	1.7526	79.832192
6	1.7526	94.800728
7	1.8034	90.718400
8	1.9304	104.779752
9	1.8034	81.646560

```
result = result.drop(result[result.height < H_lower_bound].index)
result = result.drop(result[result.height > H_upper_bound].index)
result = result.drop(result[result.weight < W_lower_bound].index)
result = result.drop(result[result.weight > W_upper_bound].index)
```

```
data.shape
```

```
(1015, 2)
```

```
result.shape
```

```
(998, 2)
```

```
# Vẽ lại box plot
height = plt.boxplot(result.height, flierprops=red_square)
plt.title("Height")
plt.show()

weight = plt.boxplot(result.weight, flierprops=red_square)
plt.title('Weight')
plt.show()
```