

1. Đọc file `iris.csv` và lưu dữ liệu vào DataFrame với tên `iris`.
2. In 5 dòng đầu tiên của `iris`
3. Hiển thị thông tin tổng quan về bộ dữ liệu IRIS, bao gồm số dòng, số cột và thông tin về các cột.
4. In ra 5 dòng đầu tiên của bộ dữ liệu IRIS. In ra 5 dòng cuối cùng của bộ dữ liệu IRIS.
5. Hiển thị thông tin cơ bản về các thuộc tính (features) của dữ liệu IRIS, chẳng hạn như min, max, mean, median, std, và các percentiles.
6. In ra giá trị trung bình của từng cột trong bộ dữ liệu IRIS.
7. In ra giá trị trung bình của từng thuộc tính (features) cho từng loại hoa (species) trong IRIS.
8. In ra tất cả các loài hoa (species) có trong bộ dữ liệu IRIS.
9. Tạo một biểu đồ histogram cho một thuộc tính (ví dụ: `sepal_length`) của bộ dữ liệu IRIS.
10. Tạo một biểu đồ scatter plot cho thuộc tính (feature) `sepal_length` và `sepal_width` với mỗi điểm dữ liệu được màu sắc khác nhau theo loài hoa.
11. Vẽ biểu đồ boxplot so sánh các thuộc tính (features) của từng loài hoa.
12. Vẽ biểu đồ pairplot cho toàn bộ bộ dữ liệu IRIS để xem sự tương quan giữa các thuộc tính (features) và phân phối của chúng.
13. Tạo biểu đồ heatmap để hiển thị ma trận tương quan (correlation matrix) giữa các thuộc tính (features) trong IRIS.
14. Tính tỷ lệ số lượng mẫu (samples) của mỗi loài hoa trong bộ dữ liệu IRIS và vẽ biểu đồ cột (bar chart) thể hiện sự phân phối này.
15. Tạo biểu đồ pie chart để thể hiện tỷ lệ phần trăm của mỗi loài hoa trong bộ dữ liệu IRIS.
16. Vẽ boxplot so sánh từng thuộc tính của từng giống hoa trong `iris`
17. Vẽ một biểu đồ có 6 subplots, mỗi subplot là một scatter_plot thể hiện từng cặp thuộc tính của iris, phân biệt giống hoa bởi màu sắc
18. Tính và vẽ biểu đồ cột (bar chart) cho tỷ lệ giữa chiều dài cánh hoa (`petal_length`) và chiều rộng cánh hoa (`petal_width`) cho từng loài hoa.

```
#1. Đọc file iris.csv vào và lưu với tên iris.
import pandas as pd
```

```
iris_df = pd.read_csv("iris.csv")
```

```
#2. In 5 dòng đầu tiên của iris
iris_df.head()
```


	sepal.length	sepal.width	petal.length	petal.width	species
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

Các bước tiếp theo: [Tạo mã bằng iris_df](#) [Xem các đồ thị được đề xuất](#) [New interactive sheet](#)



```
#3. Hiển thị thông tin tổng quan về bộ dữ liệu IRIS, bao gồm số dòng, số cột và thông tin về các cột.
print(iris_df.shape)
print(iris_df.columns)
```

```
(150, 5)
Index(['sepal.length', 'sepal.width', 'petal.length', 'petal.width',
       'species'],
      dtype='object')
```


```
#4. In ra 5 dòng cuối cùng của bộ dữ liệu IRIS
iris_df.tail()
```





	sepal.length	sepal.width	petal.length	petal.width	species
145	6.7	3.0	5.2	2.3	Virginica
146	6.3	2.5	5.0	1.9	Virginica
147	6.5	3.0	5.2	2.0	Virginica
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3.0	5.1	1.8	Virginica


#5. Hiển thị thông tin cơ bản về các thuộc tính (features) của dữ liệu IRIS, chẳng hạn như min, max, mean, median, std, và các percentiles.
iris_df.describe()



	sepal.length	sepal.width	petal.length	petal.width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000





#6. In ra giá trị trung bình của từng cột trong bộ dữ liệu IRIS.
for c in iris_df.columns[:-1]:
print(c)
print("Mean of {} = {:.4f}".format(c, iris_df.loc[:,c].mean()))



```
Mean of sepal.length = 5.8433
Mean of sepal.width = 3.0573
Mean of petal.length = 3.7580
Mean of petal.width = 1.1993
```

#7. In ra giá trị trung bình của từng thuộc tính (features) cho từng loại hoa (species) trong IRIS.
print(iris_df.groupby(by = "species").mean())



species	sepal.length	sepal.width	petal.length	petal.width
Setosa	5.006	3.428	1.462	0.246
Versicolor	5.936	2.770	4.260	1.326
Virginica	6.588	2.974	5.552	2.026

```
iris_df['species'] == 'Setosa'
```

```

↗
species
0      True
1      True
2      True
3      True
4      True
...     ...
145    False
146    False
147    False
148    False
149    False
150 rows × 1 columns

```

dtype: bool

#8. In ra tất cả các loài hoa (species) có trong bộ dữ liệu IRIS.
 print("Tất cả loài hoa: ", iris_df['species'].unique())

```

↗
Tất cả loài hoa:  ['Setosa' 'Versicolor' 'Virginica']

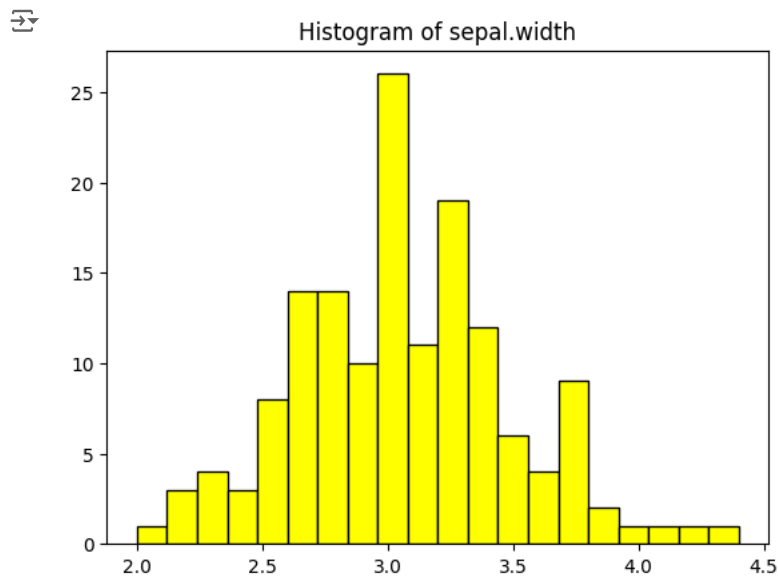
```

#9. Tạo một biểu đồ histogram cho một thuộc tính (ví dụ: sepal_length) của bộ dữ liệu IRIS.
 import matplotlib.pyplot as plt

```

plt.hist(iris_df.iloc[:,1], bins = 20, color="yellow", edgecolor="black" )
plt.title(f"Histogram of {iris_df.columns[1]}")
plt.show()

```



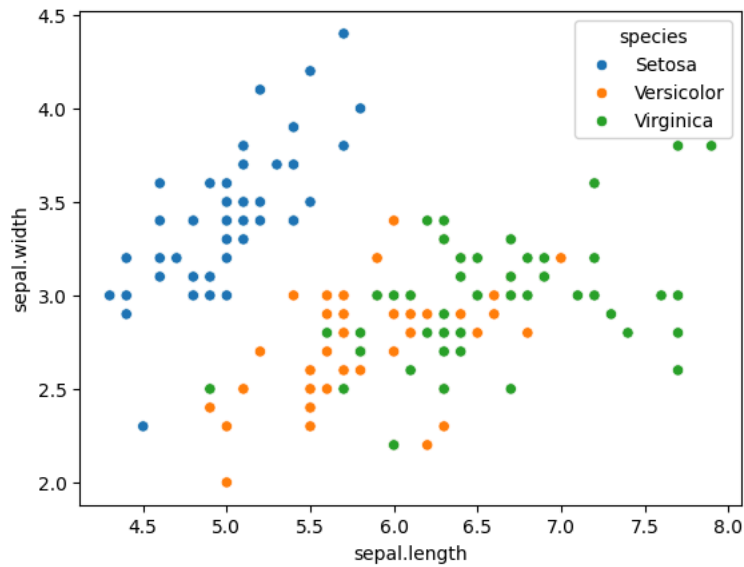
#10. Tạo một biểu đồ scatter plot cho thuộc tính (feature) sepal_length và sepal_width với mỗi điểm dữ liệu được màu sắc khác nhau theo loài
 import seaborn as sns

```

sns.scatterplot(data = iris_df, x ="sepal.length", y="sepal.width", hue="species")

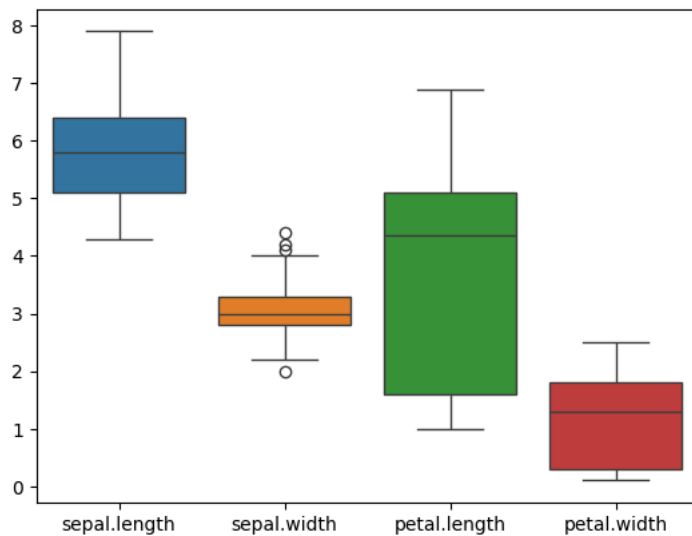
```

<Axes: xlabel='sepal.length', ylabel='sepal.width'>



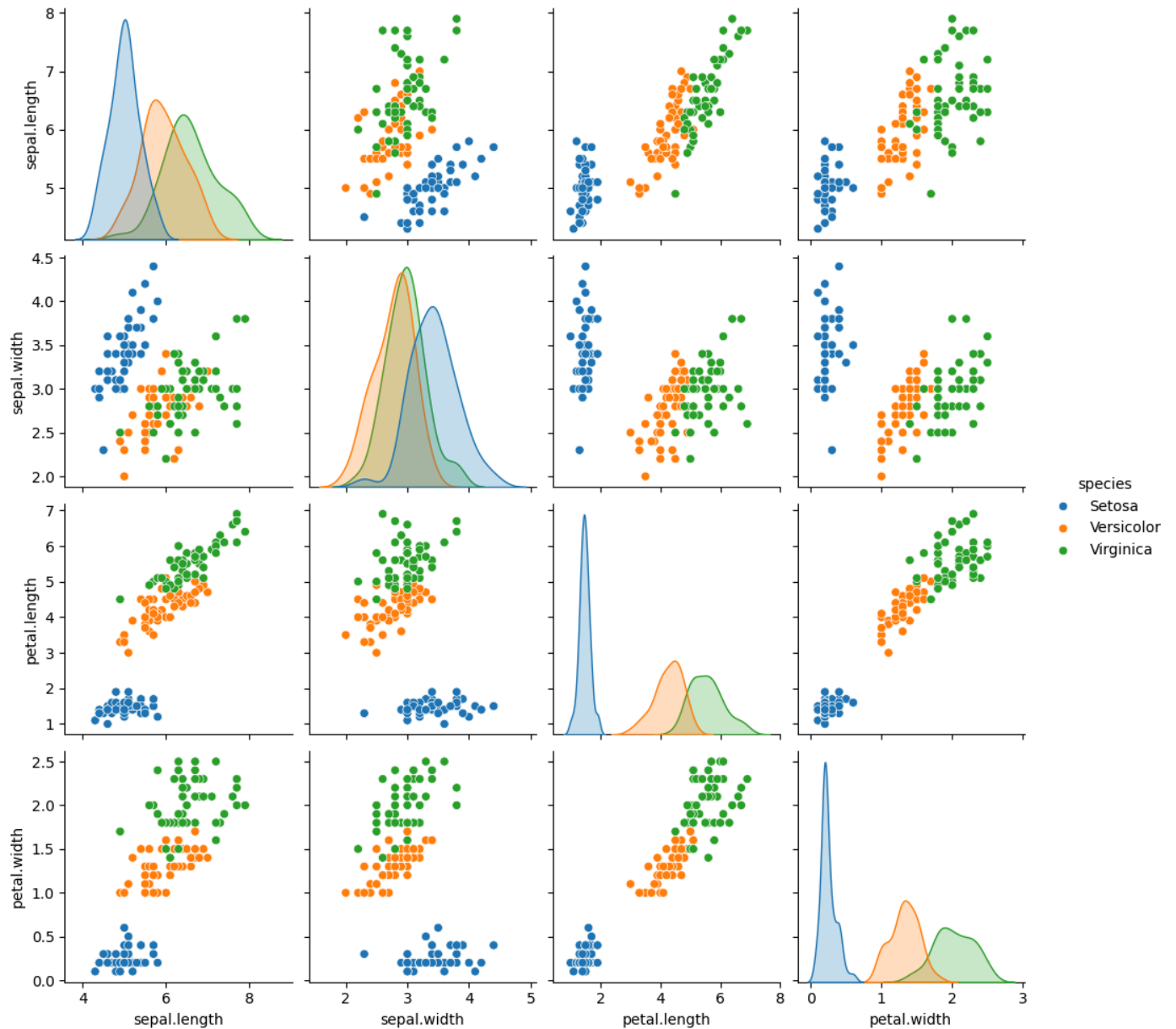
#11. Vẽ biểu đồ boxplot so sánh các thuộc tính (features) của từng loài hoa.
 sns.boxplot(data = iris_df)
 # sns.boxplot(data = iris_df, x = "species", y = "sepal.length")

<Axes: >



#12. Vẽ biểu đồ pairplot cho toàn bộ dữ liệu IRIS để xem sự tương quan giữa các thuộc tính (features) và phân phối của chúng.
 sns.pairplot(data = iris_df, hue = "species")

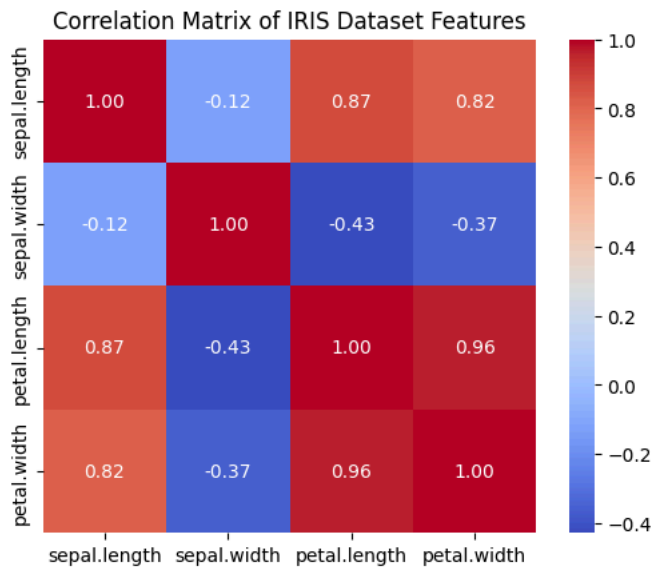
<seaborn.axisgrid.PairGrid at 0x7c1916dfa4d0>



#13. Tạo biểu đồ heatmap để hiển thị ma trận tương quan (correlation matrix) giữa các thuộc tính (features) trong IRIS.

```
numeric_iris_df = iris_df.select_dtypes(include=['number'])
sns.heatmap(numeric_iris_df.corr(), annot=True, cmap="coolwarm", square=True,
            fmt=".2f").set_title("Correlation Matrix of IRIS Dataset Features")
```

Text(0.5, 1.0, 'Correlation Matrix of IRIS Dataset Features')

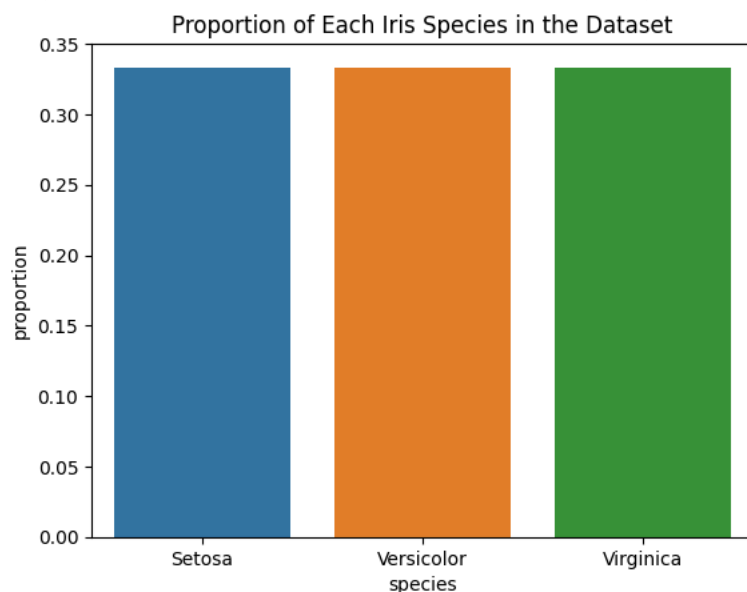


#14. Tính tỷ lệ số lượng mẫu (samples) của mỗi loài hoa trong bộ dữ liệu IRIS và # vẽ biểu đồ cột (bar chart) thể hiện sự phân phối này.

```
species_counts = iris_df['species'].value_counts(normalize=True).reset_index()
species_counts.columns = ['species', 'proportion']

sns.barplot(data=species_counts, x='species', y='proportion',
            hue = 'species').set_title("Proportion of Each Iris Species in the Dataset")
```

Text(0.5, 1.0, 'Proportion of Each Iris Species in the Dataset')



```
print("Tỷ lệ số lượng mẫu (samples) của mỗi loài hoa trong bộ dữ liệu IRIS:\n ){species_counts}")
```

Tỷ lệ số lượng mẫu (samples) của mỗi loài hoa trong bộ dữ liệu IRIS:

	species	proportion
0	Setosa	0.333333
1	Versicolor	0.333333
2	Virginica	0.333333

#15. Tạo biểu đồ pie chart để thể hiện tỷ lệ phần trăm của mỗi loài hoa trong bộ dữ liệu IRIS.

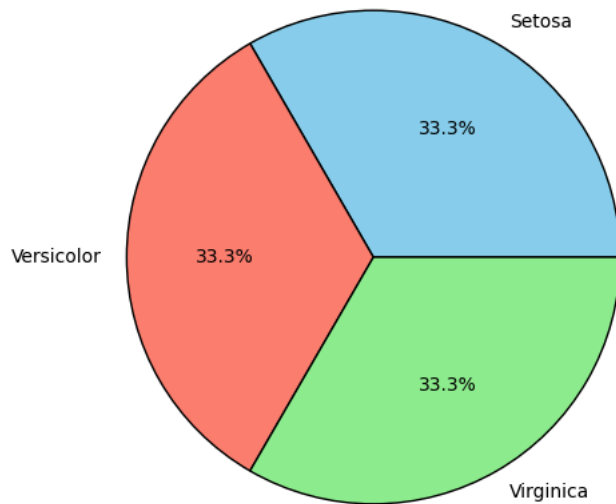
```
species_counts = iris_df['species'].value_counts(normalize=True)

plt.figure(figsize=(6,6))
species_counts.plot(kind='pie', autopct='%1.1f%%',
                    colors=['skyblue', 'salmon', 'lightgreen'],
```

```
wedgeprops={'edgecolor': 'black'})
plt.title('Proportion of Each Iris Species')
plt.gca().get_yaxis().set_visible(False)
# plt.text(0, 0, 'proportion', ha='center', va='center', fontsize=14, fontweight='bold')
plt.show()
```



Proportion of Each Iris Species



#16. Vẽ boxplot so sánh từng thuộc tính của từng giống hoa trong iris

```
plt.figure(figsize=(12, 8))

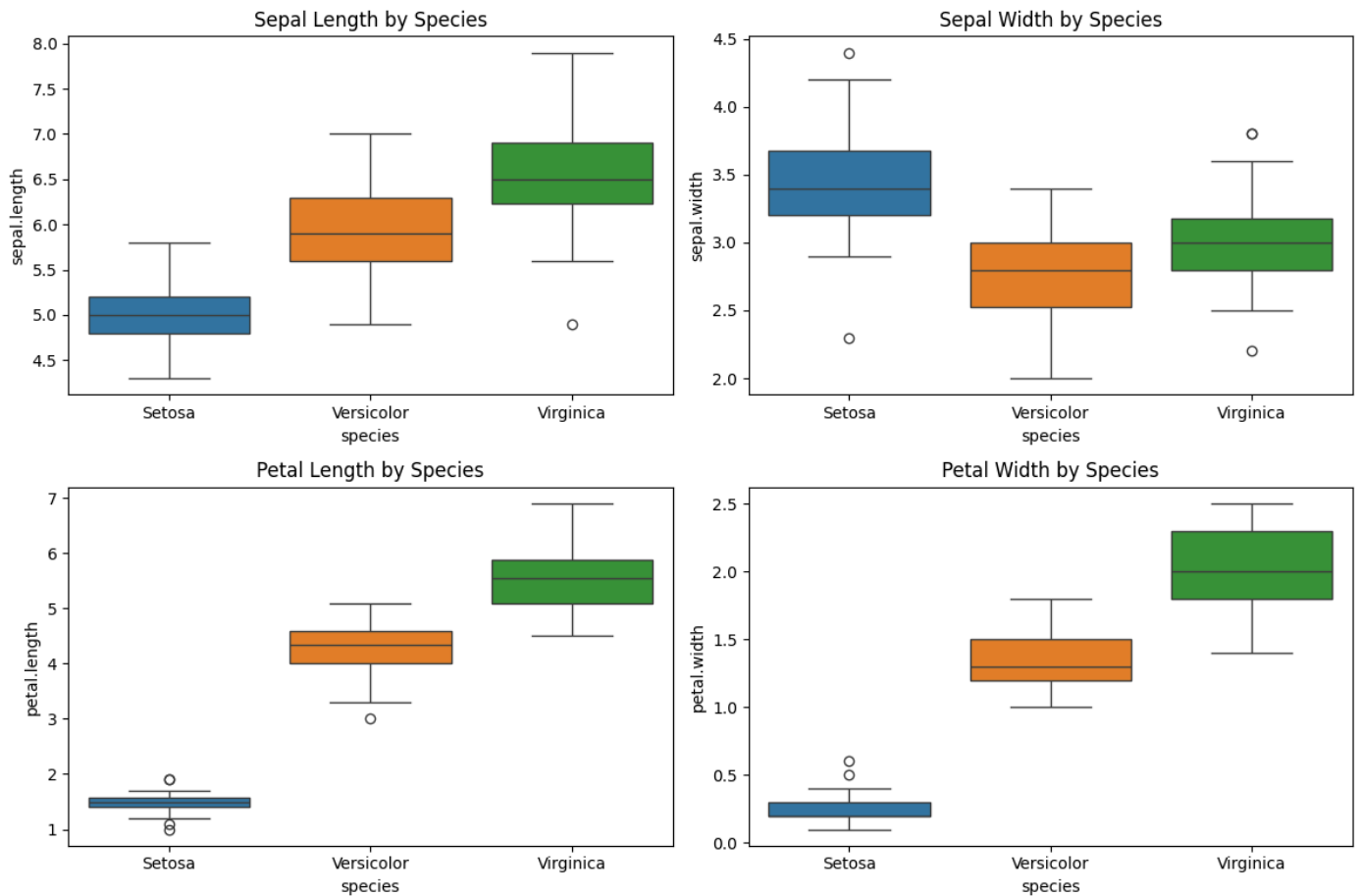
plt.subplot(2, 2, 1)
sns.boxplot(x='species', y='sepal.length', data=iris_df, hue = 'species')
plt.title('Sepal Length by Species')

# Boxplot cho sepal_width
plt.subplot(2, 2, 2)
sns.boxplot(x='species', y='sepal.width', data=iris_df, hue = 'species')
plt.title('Sepal Width by Species')

# Boxplot cho petal_length
plt.subplot(2, 2, 3)
sns.boxplot(x='species', y='petal.length', data=iris_df, hue = 'species')
plt.title('Petal Length by Species')

# Boxplot cho petal_width
plt.subplot(2, 2, 4)
sns.boxplot(x='species', y='petal.width', data=iris_df, hue = 'species')
plt.title('Petal Width by Species')

# Hiển thị các biểu đồ
plt.tight_layout()
plt.show()
```



#17. Vẽ một biểu đồ có 6 subplots, mỗi subplot là một scatter_plot thể hiện từng # cặp thuộc tính của iris, phân biệt giống hoa bởi màu sắc

```
plt.figure(figsize=(15, 10))
```

```
# 1. Scatter plot giữa sepal_length và sepal_width
```

```
plt.subplot(2, 3, 1)
```

```
sns.scatterplot(x='sepal.length', y='sepal.width', hue='species', data=iris_df)
```

```
plt.title('Sepal Length vs Sepal Width')
```

```
# 2. Scatter plot giữa sepal_length và petal_length
```

```
plt.subplot(2, 3, 2)
```

```
sns.scatterplot(x='sepal.length', y='petal.length', hue='species', data=iris_df)
```

```
plt.title('Sepal Length vs Petal Length')
```

```
# 3. Scatter plot giữa sepal_length và petal_width
```

```
plt.subplot(2, 3, 3)
```

```
sns.scatterplot(x='sepal.length', y='petal.width', hue='species', data=iris_df)
```

```
plt.title('Sepal Length vs Petal Width')
```

```
# 4. Scatter plot giữa sepal_width và petal_length
```

```
plt.subplot(2, 3, 4)
```

```
sns.scatterplot(x='sepal.width', y='petal.length', hue='species', data=iris_df)
```

```
plt.title('Sepal Width vs Petal Length')
```

```
# 5. Scatter plot giữa sepal_width và petal_width
```

```
plt.subplot(2, 3, 5)
```

```
sns.scatterplot(x='sepal.width', y='petal.width', hue='species', data=iris_df)
```

```
plt.title('Sepal Width vs Petal Width')
```

```
# 6. Scatter plot giữa petal_length và petal_width
```

```
plt.subplot(2, 3, 6)
```

```
sns.scatterplot(x='petal.length', y='petal.width', hue='species', data=iris_df)
```



```
plt.title('Petal Length vs Petal Width')
```

```
plt.tight_layout()
```

