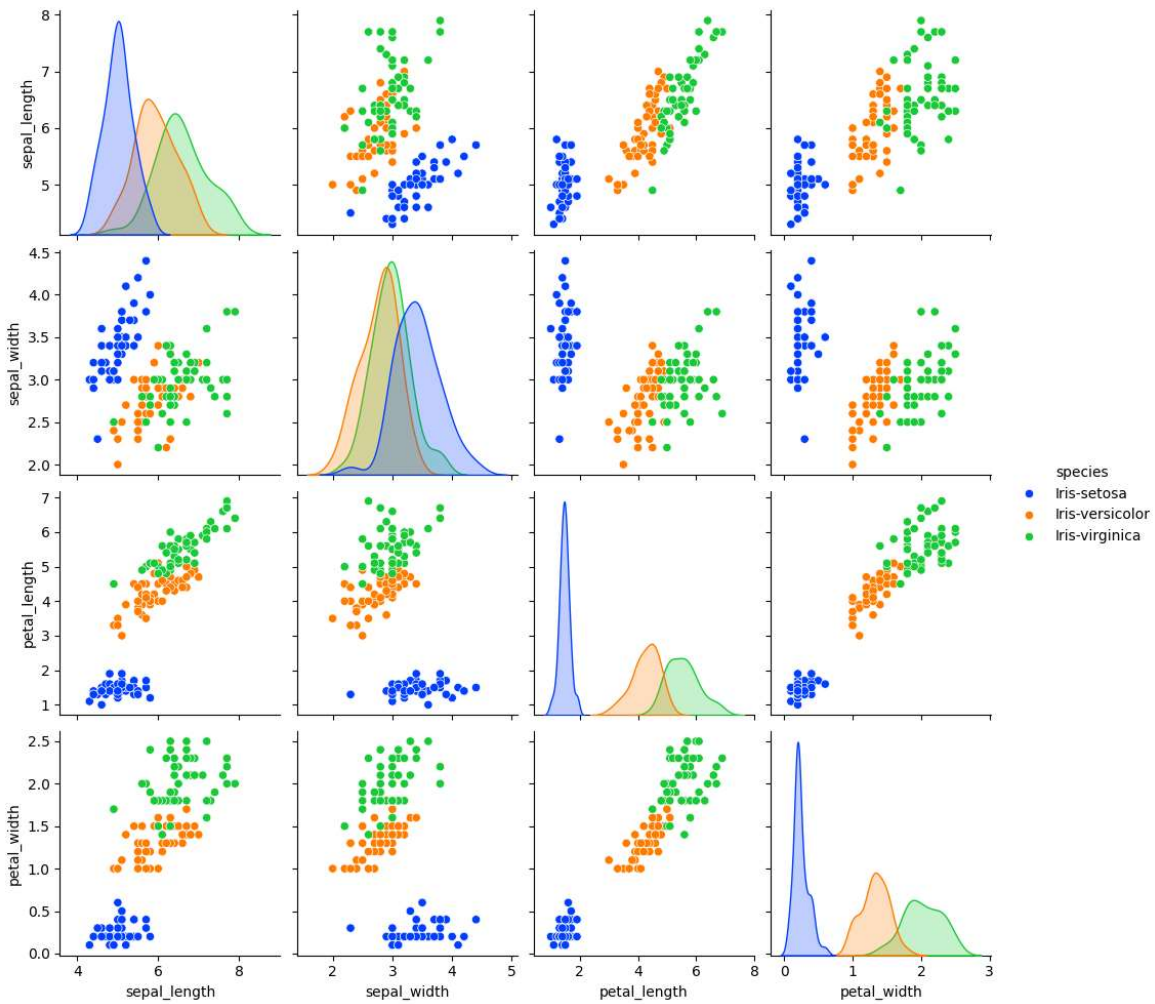
0.0.1 Scatter Plot

Problem 1: Scatter plot

Use different colors to visualize the relationship between a pair of variable in different scatterplots similar to the scatterplot below. (20 points)

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import os
import zipfile


zip_path = 'C:\\Users\\Bilal\OneDrive - McGill University\\Git Hub\\Fall 2023\\Prob Stats\\Assignment 2\\iris.zip'
extract_folder = 'C:\\Users\\Bilal\OneDrive - McGill University\\Git Hub\\Fall 2023\\Prob Stats\\Assignment 2\\iris'

with zipfile.ZipFile(zip_path, 'r') as zip_ref:
    zip_ref.extractall(extract_folder)

file_path = os.path.join(extract_folder, 'iris.data')

iris_data = pd.read_csv(file_path, header=None)
iris_data.columns = ['sepal_length', 'sepal_width', 'petal_length',
'petal_width', 'species']

sns.pairplot(iris_data, hue='species', palette='bright')
plt.show()
```

[1] ✓ 8.7s

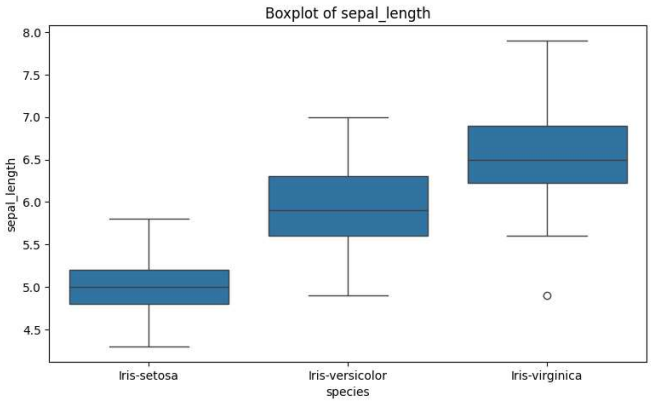


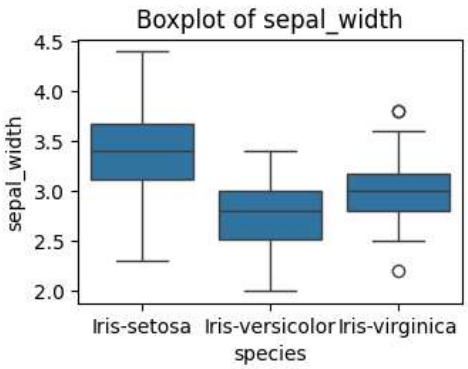
Problem 2: Box plot

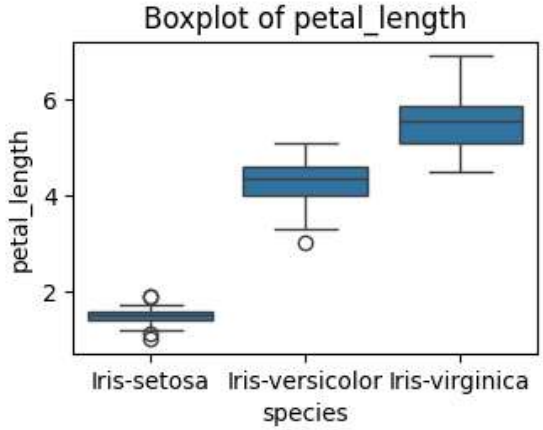
Generate the boxplot for each variable in each category and explain what you can say about the distribution of each variable in each class. (20 points)

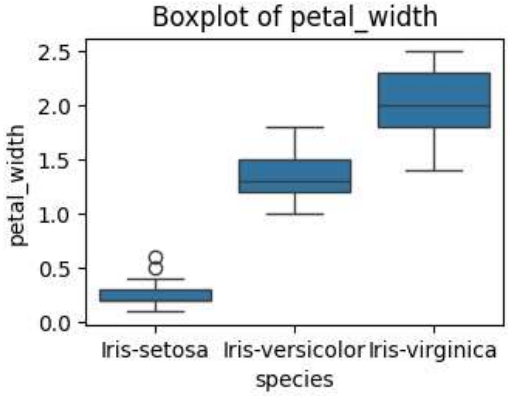
```
plt.figure(figsize=(15, 9))

for i, feature in enumerate(iris_data.columns[:-1]):
    plt.subplot(2, 2, i+1)
    sns.boxplot(x='species', y=feature, data=iris_data)
    plt.title(f'Boxplot of {feature}')
    plt.tight_layout()
    plt.show()
```

[2] ✓ 1.6s









Trait	Setosa	Versicolor	Virginica
Sepal Length	Short, consistent size	Medium, moderate variety	Long, sometimes overlaps with Versicolor
Sepal Width	Wide, consistent size	Medium, often intersects	Medium, slightly less wide than Versicolor
Petal Length	Very short, distinct	Medium, some overlap	Long, similar to Versicolor but longer
Petal Width	Narrowest, minimal overlap	Medium, intersects	Widest, overlaps with Versicolor

Comparative Analysis:

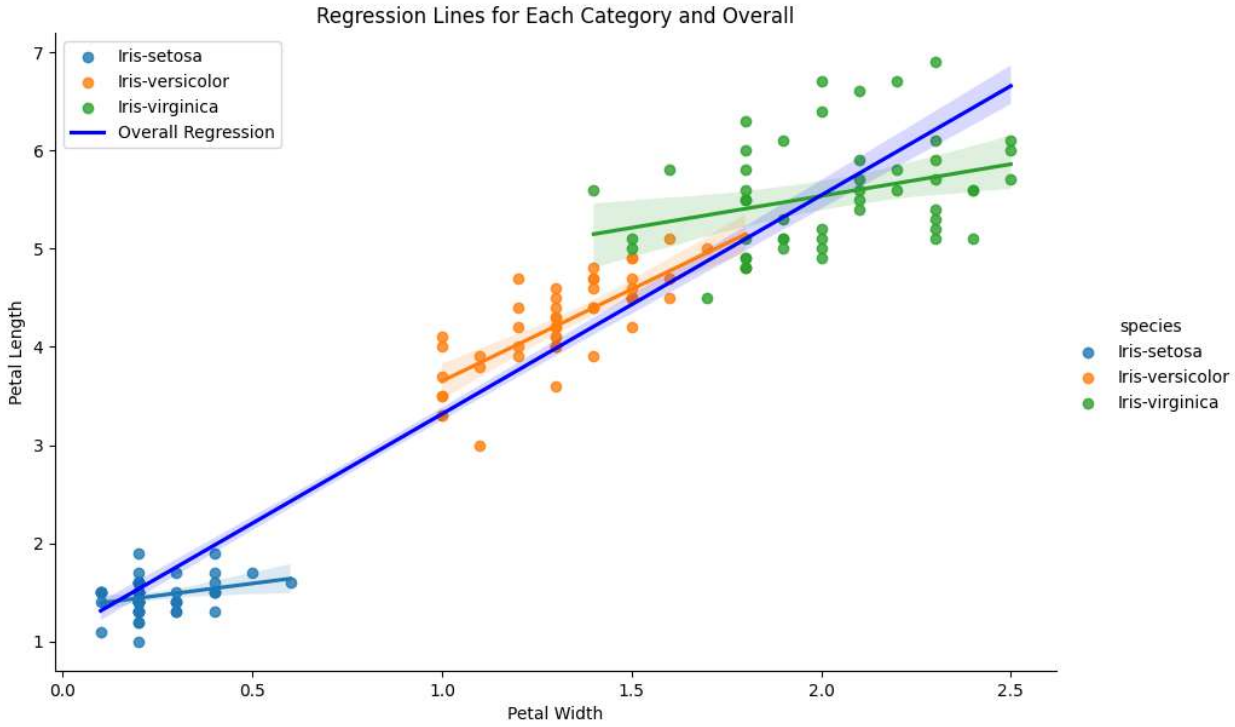
- **Setosa** is characterized by consistently shorter and wider sepals, and distinctly shorter and narrower petals.
- **Versicolor** presents with moderately long sepals and petals that have a tendency to overlap with the other two species, displaying a moderate variability in both traits.
- **Virginica** stands out with the longest sepals and widest petals, occasionally overlapping with Versicolor's measurements, suggesting a higher degree of variability.

Problem 3: Simple Regression

- Use the linear regression to predict the petal length using the petal width, i.e. to find β_0, β_1. Petal.Length = $\beta_0 + \beta_1$Petal.Width(20 points)

```
sns.lmplot(x='petal_width', y='petal_length', data=iris_data, hue='species', aspect=1.5, height=6)

sns.regplot(x='petal_width', y='petal_length', data=iris_data, scatter=False, color='blue', label='Overall Regression')

plt.legend()
plt.title('Regression Lines for Each Category and Overall')
plt.xlabel('Petal Width')
plt.ylabel('Petal Length')
plt.show()
```

[3] ✓ 1.7s



Problem 4: Descriptive Statistics

Generate the descriptive statistics that include min, max, mean, first quartile Q1, second quartile or median Q2, and the 3rd quartile Q3, for each of the continuous variables.(20 points)

```
descriptive_stats = iris_data.describe()
print(descriptive_stats)
```

[4] ✓ 0.0s

```
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.054000      3.758667     1.198667
std        0.828066     0.433594      1.764420     0.763161
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
```