Assignment 1

Question 0. You are provided with a dataset comprising measurements of chlorophyll levels (chlorophyll a and chlorophyll b) from 100 plant samples collected from two types of forests. Fifty samples were taken from deciduous forests, and the remaining fifty from evergreen forests. Perform the below mentioned analysis on the given data:

Solution. I will be using the Python programming language to code for each of the questions. Firstly, I have loaded the data for deciduous and evergreen, respectively. The initial part of the code is as follows:

```
# Importing the necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Loading the data
data = pd.read_csv("Data_for_assignment1.csv")
data_split_num = 50
deciduous = data[:data_split_num]
evergreen = data[data_split_num:]
evergreen = evergreen.reset_index(drop=True)
```

Listing 1. Loading the data

I begin with an elementary analysis of the data. The command **data.describe()** gives me the descriptive statistics of the data.

- It includes measures of central tendencies like the mean, median, minimum, and maximum.
- It includes measures of dispersion like quartiles (25%, 50%, and 75%) and standard deviation
- This statistic has been obtained for the entirety of the data. Similar statistics can be obtained separately for deciduous and evergreen forests as well.

	Chlorophyll_a	Chlorophyll_b
count	100.000000	100.000000
mean	4.095036	1.742074
std	1.862227	0.794192
min	0.071410	0.149904
25%	3.032940	1.147407
50%	4.098685	1.769393
75%	5.540312	2.240584
max	8.320829	3.422601

Figure 1. Statistics

Using the commands **data.isnull().sum()**, **data.isna().sum()**, I checked for the presence of any null or NaN cells in the above dataframe. I found that none of the cells matched the above two types, indicating a completely filled dataframe.
The dataframe has 4 columns:

- **Sample_id**: A unique string ID denoting a specific sample
- **Source**: Type of forest from which sample was collected
- **Chlorophyll_a**: Level of chlorophyll A in the sample
- **Chlorophyll_b**: Level of chlorophyll B in the sample

As a preliminary test, I wanted to test the relationship between Chlorophyll_a and Chlorophyll_b for the entire dataframe. I calculated the Pearson correlation coefficient between the two.

```
from scipy.stats import pearsonr
print(pearsonr(data.Chlorophyll_a,data.Chlorophyll_b))
# Result
PearsonRResult(statistic=-0.006806478054569109, pvalue=0.9464146782039555)
```

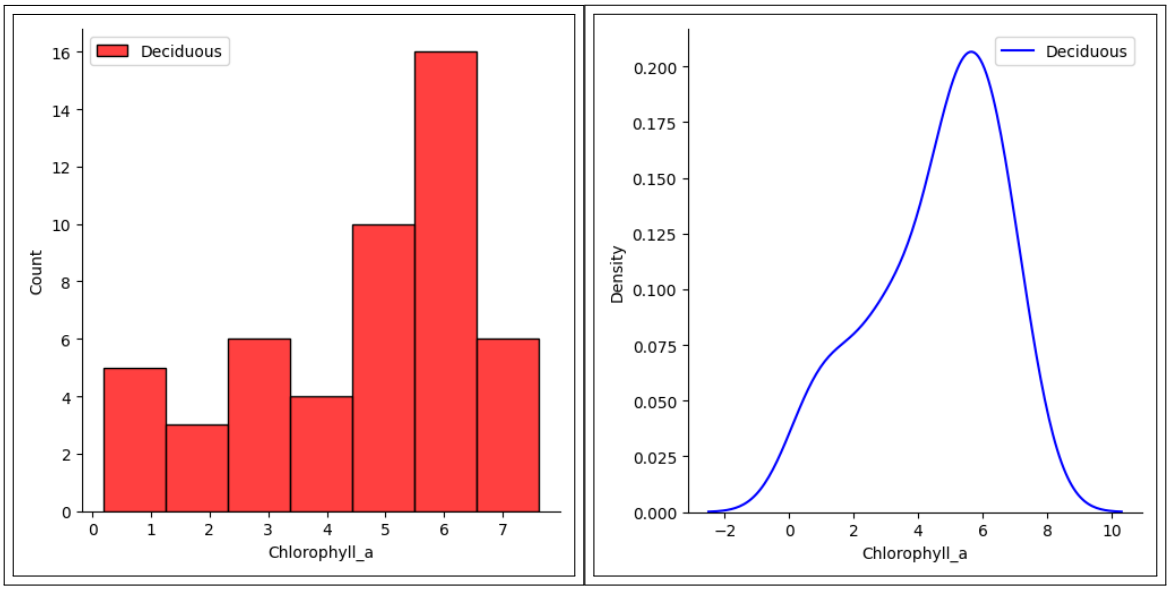
Listing 2. Pearson correlation coefficient

Question 1. Visualize the distribution of chlorophyll_a and chlorophyll_b values using histograms or density plots in deciduous forests.

Solution. The skewness of the histograms has been calculated. The distribution of chlorophyll_a is heavily left-skewed. The distribution of chlorophyll_b is very slightly right-skewed.

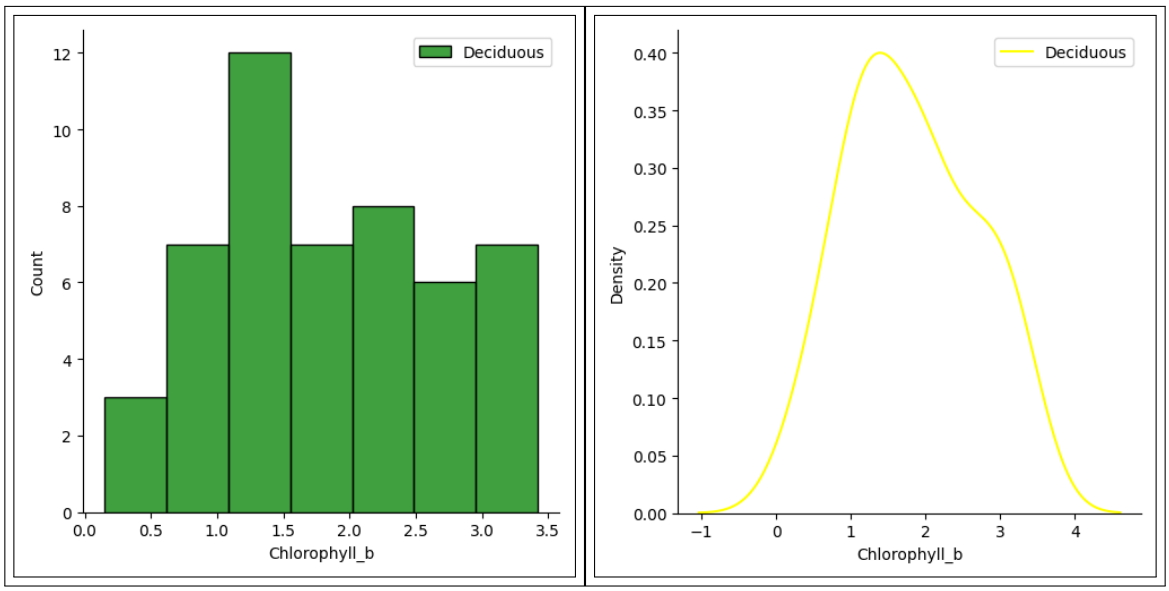
```
print("Skewness of Chlorophyll_a is", skew(deciduous.Chlorophyll_a))
print("Skewness of Chlorophyll_b is", skew(deciduous.Chlorophyll_b))
# Results
Skewness of Chlorophyll_a is -0.6337978111120774
Skewness of Chlorophyll_b is 0.16988631913025232
```

LISTING 3. Skewness of data



(A) Chlorophyll a in deciduous (Histogram) (B) Chlorophyll a in deciduous (Density Plot)

FIGURE 2. Chlorophyll a distribution



(A) Chlorophyll b in deciduous (Histogram) (B) Chlorophyll b in deciduous (Density Plot)

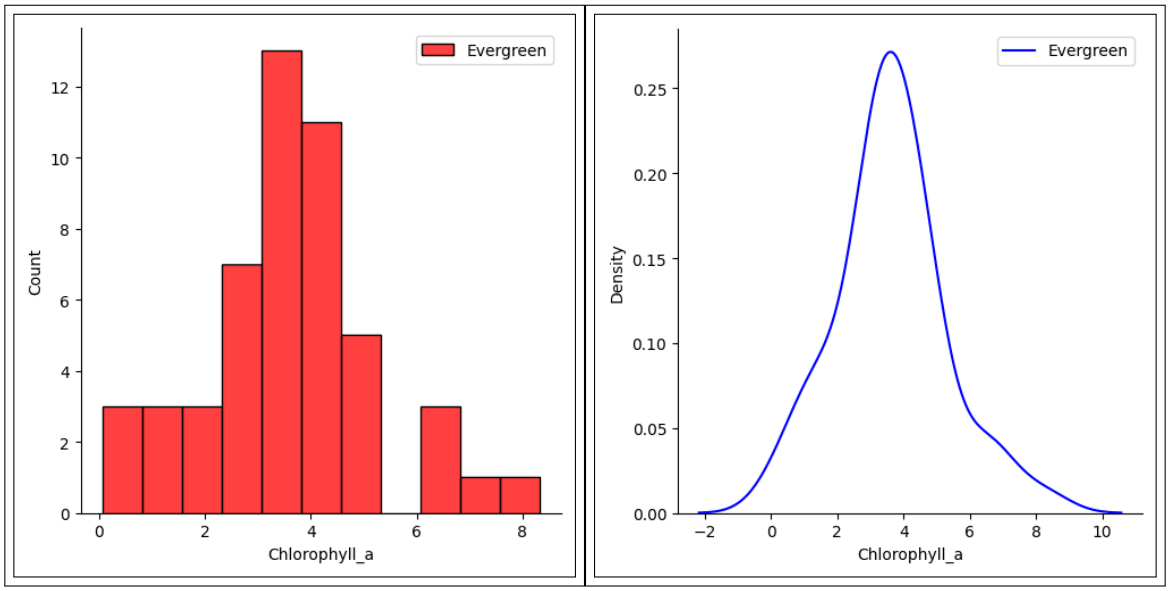
FIGURE 3. Chlorophyll b distribution

Question 2. Visualize the distribution of chlorophyll_a and chlorophyll_b values using histograms or density plots in evergreen forests.

Solution. The skewness of the histograms has been calculated. The distribution of chlorophyll_a is slightly right-skewed. The distribution of chlorophyll_b is slightly left-skewed.

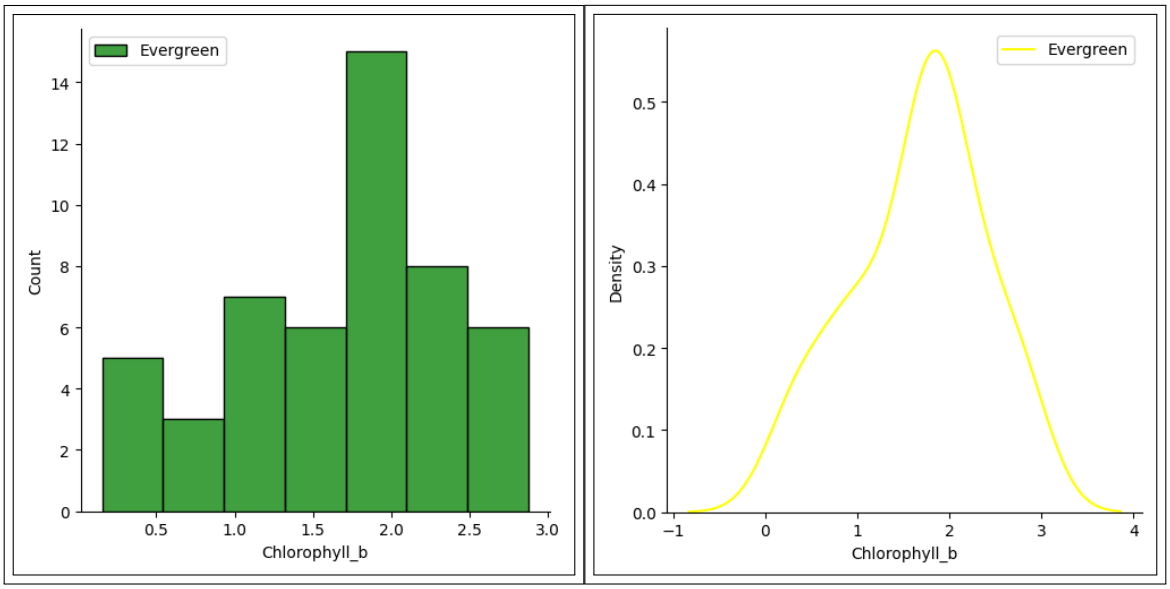
```
print("Skewness of Chlorophyll_a is", skew(evergreen.Chlorophyll_a))
print("Skewness of Chlorophyll_b is", skew(evergreen.Chlorophyll_b))
# Results
Skewness of Chlorophyll_a is 0.3597775204089479
Skewness of Chlorophyll_b is -0.31742988070810224
```

LISTING 4. Skewness of data



(A) Chlorophyll a in evergreen (Histogram) (B) Chlorophyll a in evergreen (Density Plot)

FIGURE 4. Chlorophyll a distribution



(A) Chlorophyll b in evergreen (Histogram) (B) Chlorophyll b in evergreen (Density Plot)

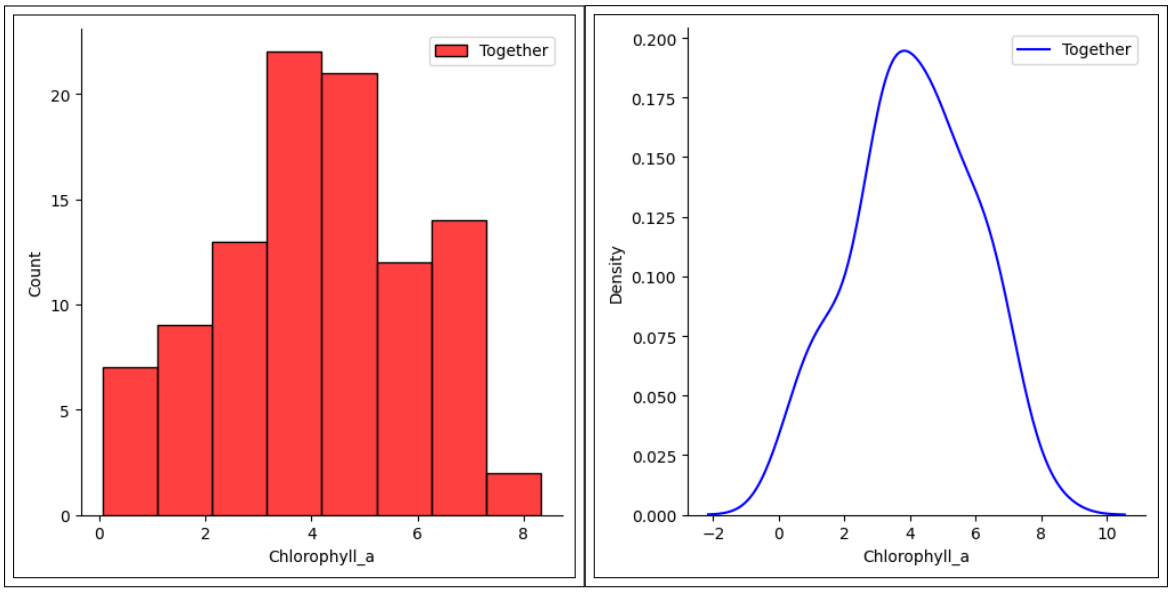
FIGURE 5. Chlorophyll b distribution

Question 3. Visualize the distribution of chlorophyll_a and chlorophyll_b values using histogram or density plots without separating the measurements from each forests.

Solution. The skewness of the histograms has been calculated. The distribution of chlorophyll_a is very slightly left-skewed. The distribution of chlorophyll_b is very slightly right-skewed.

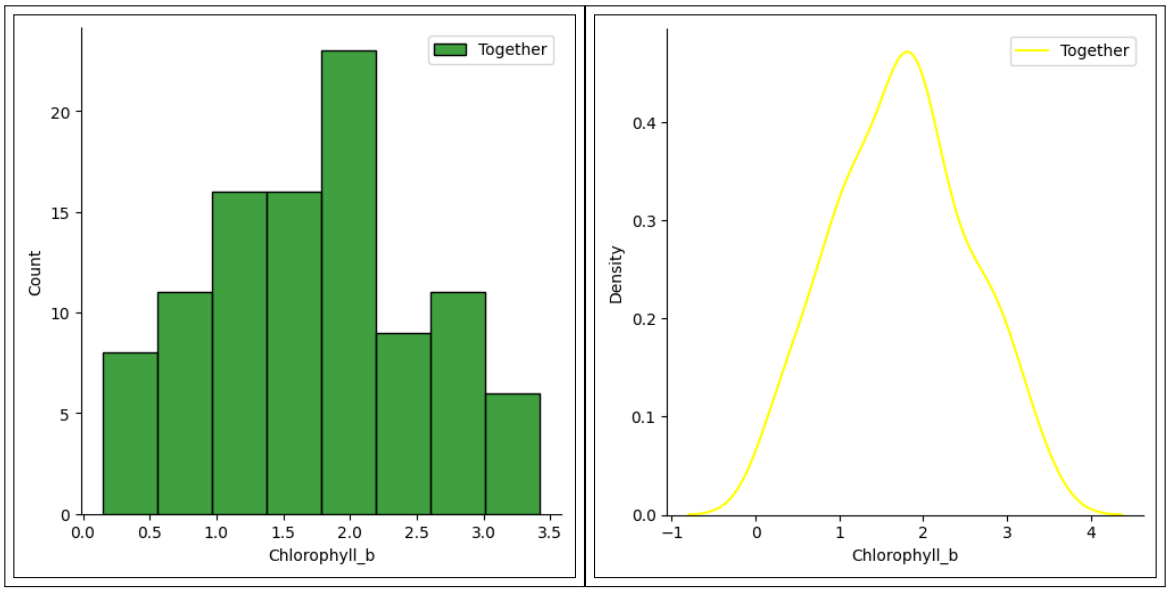
```
print("Skewness of Chlorophyll_a is", skew(data.Chlorophyll_a))
print("Skewness of Chlorophyll_b is", skew(data.Chlorophyll_b))
# Results
Skewness of Chlorophyll_a is -0.11888017040249085
Skewness of Chlorophyll_b is 0.04347253963995298
```

LISTING 5. Skewness of data



(A) Chlorophyll a in together (Histogram) (B) Chlorophyll a in together (Density Plot)

FIGURE 6. Chlorophyll a distribution



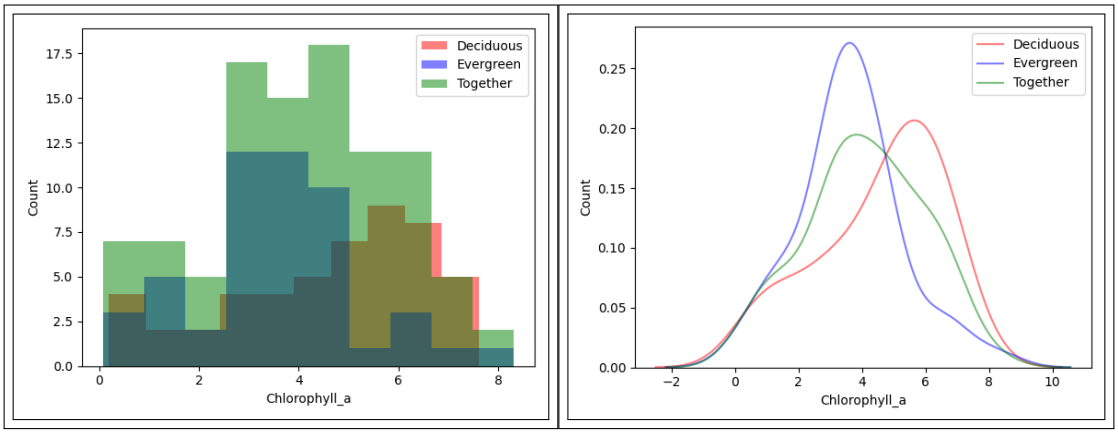
(A) Chlorophyll b in together (Histogram) (B) Chlorophyll b in together (Density Plot)

FIGURE 7. Chlorophyll b distribution

Question 4. Plot 1, 2, and 3 in the same plot together and explore how they change.

Solution. All 3 plots have been plotted together in a single plot.
For chlorophyll_a:

- In the case of all forest samples together, the data is almost symmetrical with a very slight left skew. The case of deciduous forests is heavily left-skewed, and the case of evergreen forests is quite right-skewed.
- In the case of evergreen forests, the concentration of values is higher in the center, indicating a positive kurtosis, while in the other cases of deciduous forests and all forest samples together, the concentration of values is higher towards the tails, indicating a negative kurtosis.

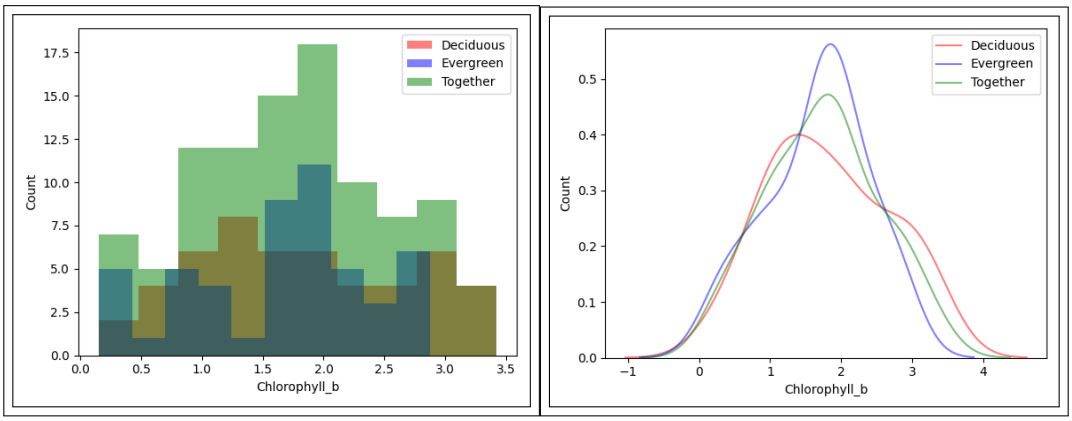


(A) Chlorophyll a variation (Histogram) (B) Chlorophyll a variation (Density Plot)

FIGURE 8. Chlorophyll a distribution among all 3 variations

For chlorophyll_b:

- In the case of all forest samples together, the data is almost symmetrical with a very minute right skew. The case of deciduous forests is slightly right-skewed, and the case of evergreen forests is quite left-skewed.
- In the case of deciduous forests, the concentration of values is higher in the tails, indicating a highly negative kurtosis. In the other cases of evergreen forests and all forest samples together, the concentration of values is higher towards the tails but lesser than in deciduous forests, indicating a negative kurtosis with a magnitude lesser than in deciduous forests.



(A) Chlorophyll b variation (Histogram) (B) Chlorophyll b variation (Density Plot)

FIGURE 9. Chlorophyll b distribution among all 3 variations

The skewness and kurtosis values for the above data have been calculated.

```
# Results of skewness
Skewness of Chlorophyll_a in deciduous forests is -0.6337978111120774
Skewness of Chlorophyll_a in evergreen forests is 0.3597775204089479
Skewness of Chlorophyll_a in all forests is -0.11888017040249085
Skewness of Chlorophyll_b in deciduous forests is 0.16988631913025232
Skewness of Chlorophyll_b in evergreen forests is -0.31742988070810224
Skewness of Chlorophyll_b in all forests is 0.04347253963995298

# Results of kurtosis
Kurtosis of Chlorophyll_a in deciduous forests is -0.6230270157101536
Kurtosis of Chlorophyll_a in evergreen forests is 0.6547041688013007
Kurtosis of Chlorophyll_a in all forests is -0.60405787117746
Kurtosis of Chlorophyll_b in deciduous forests is -0.935898963831411
Kurtosis of Chlorophyll_b in evergreen forests is -0.6086918652072542
Kurtosis of Chlorophyll_b in all forests is -0.6588517829131963
```

Listing 6. Skewness and Kurtosis of data

Question 5. Calculate summary statistics (mean, median, mode, and standard deviation) of chlorophyll a and chlorophyll b measurements from deciduous forests separately, evergreen forests separately, and both forests together.

Solution. The code to compute the summary statistics (mean, median, mode, and standard deviation) is given below:

```
d_c_a = deciduous.Chlorophyll_a
deciduous_mode_a = "None" # Each value occurs only once
print(d_c_a.mean(), d_c_a.median(), deciduous_mode_a, d_c_a.std())

e_c_a = evergreen.Chlorophyll_a
evergreen_mode_a = "None" # Each value occurs only once
print(e_c_a.mean(), e_c_a.median(), evergreen_mode_a, e_c_a.std())

all_c_a = data.Chlorophyll_a
all_mode_a = "None" # Each value occurs only once
print(all_c_a.mean(), all_c_a.median(), all_mode_a, all_c_a.std())
```

Listing 7. Translation

The mode in all the cases should be **None**, because each measurement of chlorophyll level is unique (as verified by performing **data.Chlorophyll_a.value_counts()**). This holds true for chlorophyll_a. Hence, the mode in all cases is **None**.
For calculating the other summary statistics, we have direct functions in Python: **.mean()**, **.median()**, and **.std()**.
The results for **chlorophyll_a** as obtained from the above code are given below.

Forest	Mean	Median	Mode	Std_dev
Deciduous	4.5643655123	4.9747212349	None	1.9643988111
Evergreen	3.6257064717	3.5314637960	None	1.6426079864
Together	4.0950359920	4.0986846120	None	1.8622273524

Forest	Mean	Median	Mode	Std_dev
Deciduous	1.8178106263	1.7635951109	None	0.8637127649
Evergreen	1.6663382311	1.7813223946	None	0.7187833082
Together	1.7420744287	1.7693933056	None	0.7941923488

Similarly, summary statistics can be obtained for **chlorophyll_b** as well. The mode in all the cases should be **None**, because each measurement of chlorophyll level is unique (as verified by performing **data.Chlorophyll_b.value_counts()**). This holds true for chlorophyll_b as well. Hence, the mode in all cases is **None**.
For calculating the other summary statistics, we have direct functions in Python: **.mean()**, **.median()**, and **.std()**.

Question 6. In the same boxplot, compare how the distribution of chlorophyll a and chlorophyll b values compares in Deciduous forests and Evergreen forests

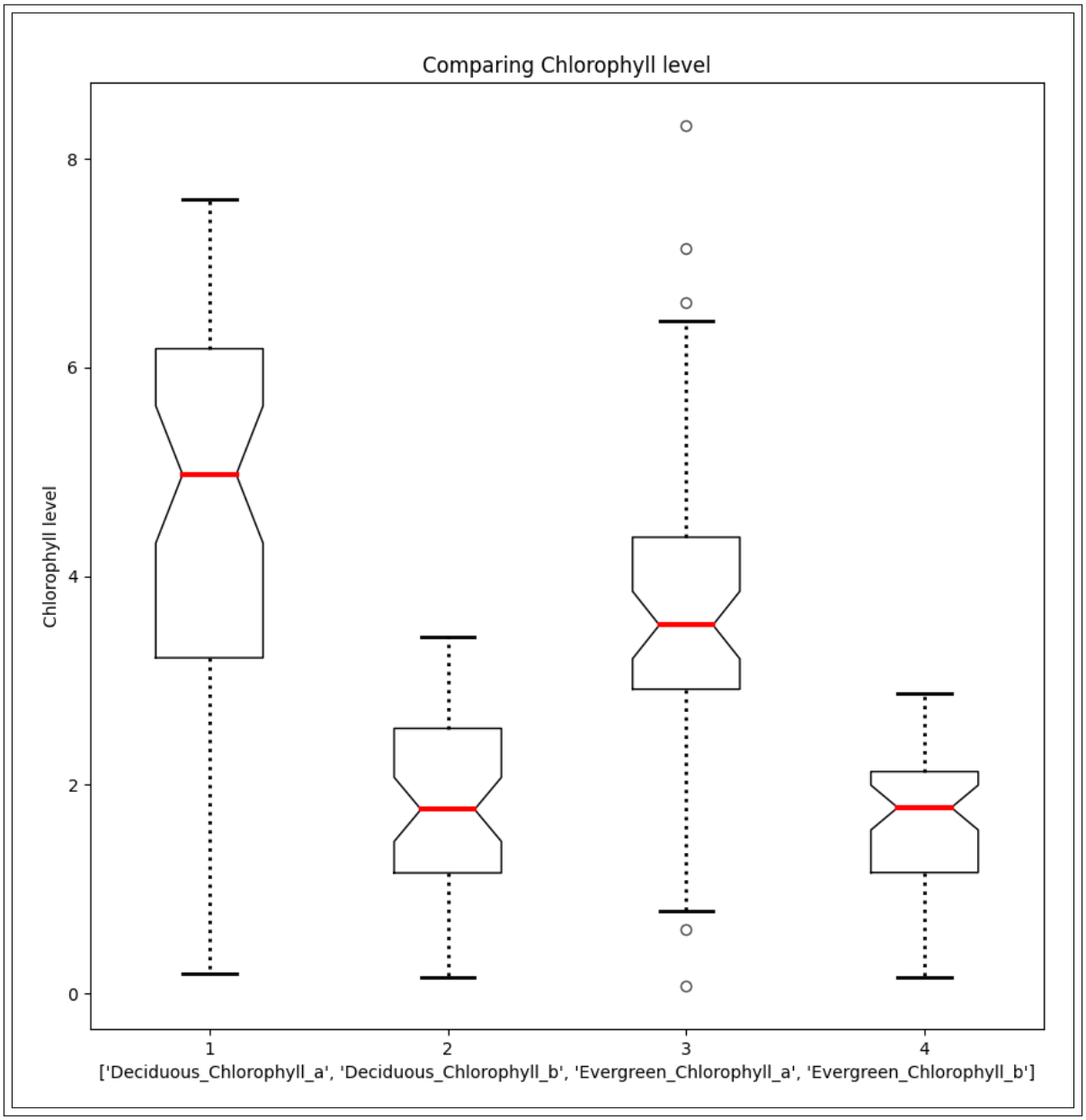


FIGURE 10. Box-plot for comparison

Solution.

- The **interquartile region** of deciduous_chlorophyll_a is located quite high up (in terms of chlorophyll level) when compared to deciduous_chlorophyll_b. This indicates that the **chlorophyll_a level is much higher than the chlorophyll_b level** for **deciduous forests**.
- The **interquartile range** of deciduous_chlorophyll_a is **much larger** when compared to deciduous_chlorophyll_b. The interquartile range is a measure of statistical dispersion. A greater interquartile range implies **greater variability** in the data. **This implies that there is greater variation in deciduous_chlorophyll_a than deciduous_chlorophyll_b**.
- From the boxplot, I observe that the deciduous_chlorophyll_a is **heavily left-skewed** (larger part of the IQR lies below the median), while the deciduous_chlorophyll_b is **slightly right-skewed**.
- The **interquartile region** of evergreen_chlorophyll_a is located quite high up (in terms of chlorophyll level) when compared to evergreen_chlorophyll_b. This indicates that the **chlorophyll_a level is much higher than the chlorophyll_b level** for evergreen forests.
- The **interquartile range** of evergreen_chlorophyll_a is **almost similar** to that of evergreen_chlorophyll_b. The interquartile range is a measure of statistical dispersion. A greater interquartile range implies greater variability in the data. This implies that the **variation in evergreen_chlorophyll_a and evergreen_chlorophyll_b is almost equivalent**.
- From the boxplot, I observe that the evergreen_chlorophyll_a is **slightly right-skewed** (larger part of the IQR lies above the median), while the evergreen_chlorophyll_b is **slightly left-skewed**.

Question 7. Are the variances between chlorophyll a and chlorophyll b measurements different significantly? Perform appropriate statistical tests to support your claim. Compare the variances of chlorophyll content from deciduous forests separately, evergreen forests separately, and both forests together.

Solution. Here, in order to compare the variances of chlorophyll content for various forest cases, I employed the F-test. I first computed the f value, which is the ratio between the variances of chlorophyll_a and chlorophyll_b respectively, for each forest case. Then, I computed the **p_value** by subtracting the CDF of the F-distribution from 1. The code for computing the **p_value** is given below:

```
# Null Hypothesis: Ho = variane(c_a) = variance(c_b) for deciduous forests
# Alternate Hypothesis: H1 = variance(c_a) != variance(c_b) for deciduous
    forests

# n1 = 50; n2 = 50
# v1 = 49; v2 = 49
# alpha = 5% = 0.05

var_a = deciduous.Chlorophyll_a.var()
var_b = deciduous.Chlorophyll_b.var()
f_val_d = var_a / var_b
p_val_d = 1 - stats.f.cdf(f_val_d, 49, 49)
print("p value deciduous:",p_val_d)
# p value deciduous: 2.6136590558500927e-08

# Since the p value is less than alpha, I am rejecting null hypothesis
```

```
"""H1 = variance(c_a) != variance(c_b) for deciduous forests"""
```

LISTING 8. Computing p value via F-statistic

The above code computes the **p value** via the **F statistic method** only for the case of deciduous forests. Similarly, a p value can also be found for the other two cases of evergreen forests and all forests together.
My null hypothesis is that in each case, the variance of chlorophyll_a is equal to the variance of chlorophyll_b. My alpha for each case was 0.05. Based on the p_values, I obtained the following results:

- p value of **deciduous only**: 2.6136590558500927e-08. This value is very low when compared to the value of alpha. Hence, the **null hypothesis is rejected**.
- p value of **evergreen only**: 2.2224616258448293e-08. This value is very low when compared to the value of alpha. Hence, the **null hypothesis is rejected**.
- p value of **both forests together**: 5.551115123125783e-16. This value is very low when compared to the value of alpha. Hence, the **null hypothesis is rejected**.

From the above observations, one result is clear, which is, the **variance of chlorophyll_a is not equal to the variance of chlorophyll_b** in any case of forests. Therefore, **the variances of chlorophyll_a and chlorophyll_b are very different**. This claim is supported by the **F-test statistic** shown above.

Question 8. Finally, test whether the mean of chlorophyll a is greater than the mean of chlorophyll b using appropriate statistical test in all three combinations: deciduous forests separately, evergreen forests separately, and both forests together. Clearly state your null hypothesis, chosen significance criteria and the result of hypothesis testing.

Solution. In the previous question, I obtained the result that the variance of chlorophyll_a is not equal to the variance of chlorophyll_b. Hence, now, in order to compare the means, I have to use **Separate variance t-test**. I will employ the function **stats.ttest_ind()** to achieve this. I will explain the case of deciduous forests. The same can be extended for evergreen forests and both forests together.
My null hypothesis is that the **mean of chlorophyll_a is equal to the mean of chlorophyll_b**. I am considering the value of alpha (significance level) as 0.05. My alternative hypothesis is that the **mean of chlorophyll_a is greater than the mean of chlorophyll_b**. Also, the **variances of chlorophyll_a and chlorophyll_b are not equal**. The code for the above described test is below:

```
# Deciduous forests
stat_val_d, p_val_d = stats.ttest_ind(deciduous.Chlorophyll_a, deciduous.
    Chlorophyll_b, equal_var = False, alternative = "greater")
print("Statistic value of deciduous forests ttest:", stat_val_d)
print("p value of deciduous forests ttest:", p_val_d)

# Evergreen forests
stat_val_e, p_val_e = stats.ttest_ind(evergreen.Chlorophyll_a, evergreen.
    Chlorophyll_b, equal_var = False, alternative = "greater")
print("Statistic value of evergreen forests ttest:", stat_val_e)
print("p value of evergreen forests ttest:", p_val_e)

# Both forests together
stat_val_a, p_val_a = stats.ttest_ind(data.Chlorophyll_a, data.Chlorophyll_b,
     equal_var = False, alternative = "greater")
print("Statistic value of all forests ttest:", stat_val_a)
print("p value of all forests ttest:", p_val_a)
```

LISTING 9. Computing p value via separate variance t-test

These inputs are passed to the function **stats.ttest_ind()**. The statistic and p values are computed by the function. If the p value is less than alpha, the null hypothesis is rejected and the alternative hypothesis is accepted. On performing the tests on **stats.ttest_ind()**, the results are:

- p value of **deciduous only**: 1.5035020828357813e-13. This value is very low when compared to the value of alpha. Hence, the **null hypothesis is rejected** and **alternative hypothesis is accepted**.
- p value of **evergreen only**: 3.7014110909166746e-11. This value is very low when compared to the value of alpha. Hence, the **null hypothesis is rejected** and **alternative hypothesis is accepted**.
- p value of **both forests together**: 2.5301343434745824e-22. This value is very low when compared to the value of alpha. Hence, the **null hypothesis is rejected** and **alternative hypothesis is accepted**.

The alternative hypothesis states that the mean of chlorophyll_a is greater than the mean of chlorophyll_b. This point has been proven by the above **separate variance t-test**.

Here is the link to the Google Drive with the Jupyter ipynb notebook containing the codes for assignment 1.
Drive link to codes