BT3041 – Analysis and Interpretation of Biological Data
Assignment 1

Atharv Karkar, ED21B014
IIT Madras

March 2, 2025

1 Introduction

This report presents an analysis of chlorophyll levels (chlorophyll a and chlorophyll b) in plant samples collected from Deciduous and Evergreen forests. The dataset consists of 100 samples, with 50 from each type of forest. The analysis follows the specific questions outlined in the assignment.

2 Analysis

2.1 Question 1: Histogram of Chlorophyll in Deciduous Forests

Task: Visualize the distribution of chlorophyll a and chlorophyll b values using histograms for Deciduous forests.

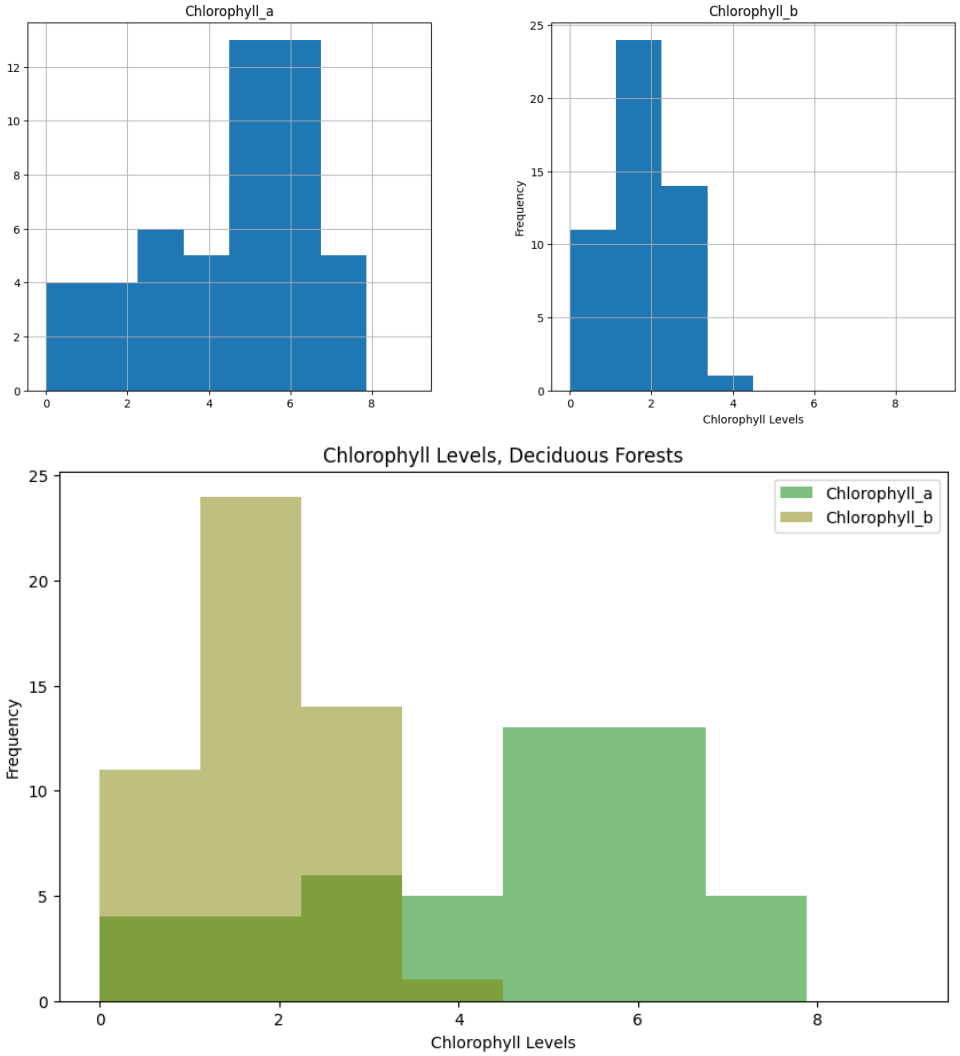


Figure 1: Histogram of chlorophyll a and chlorophyll b in Deciduous forests

2.2 Question 2: Histogram of Chlorophyll in Evergreen Forests

Task: Visualize the distribution of chlorophyll a and chlorophyll b values using histograms for Evergreen forests.

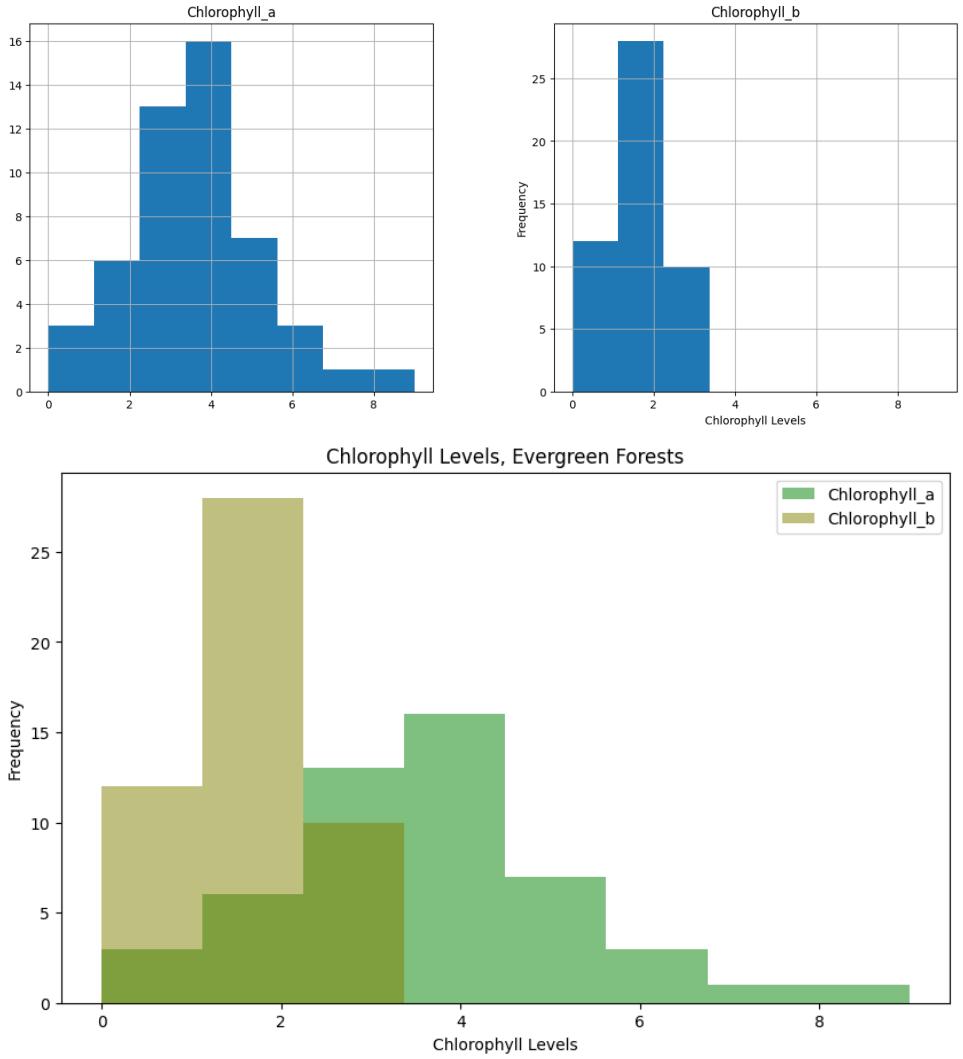


Figure 2: Histogram of chlorophyll a and chlorophyll b in Evergreen forests

2.3 Question 3: Histogram Without Separating Forest Types

Task: Visualize the distribution of chlorophyll a and chlorophyll b values using histogram without separating the measurements from each forests.

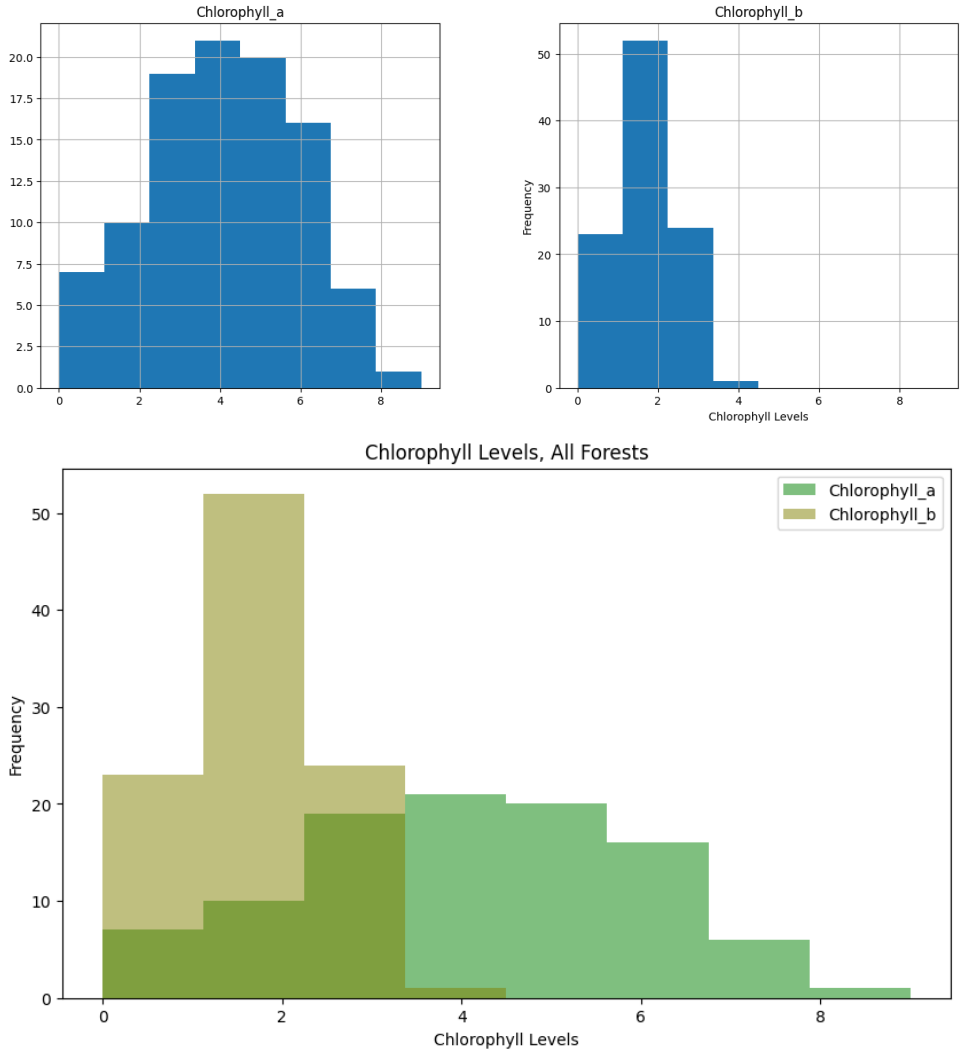


Figure 3: Histogram of chlorophyll a and chlorophyll b for all samples

2.4 Question 4: Density Plot Comparison

Task: Plot 1, 2 and 3 in same plot together as density plot and explore how it changes.

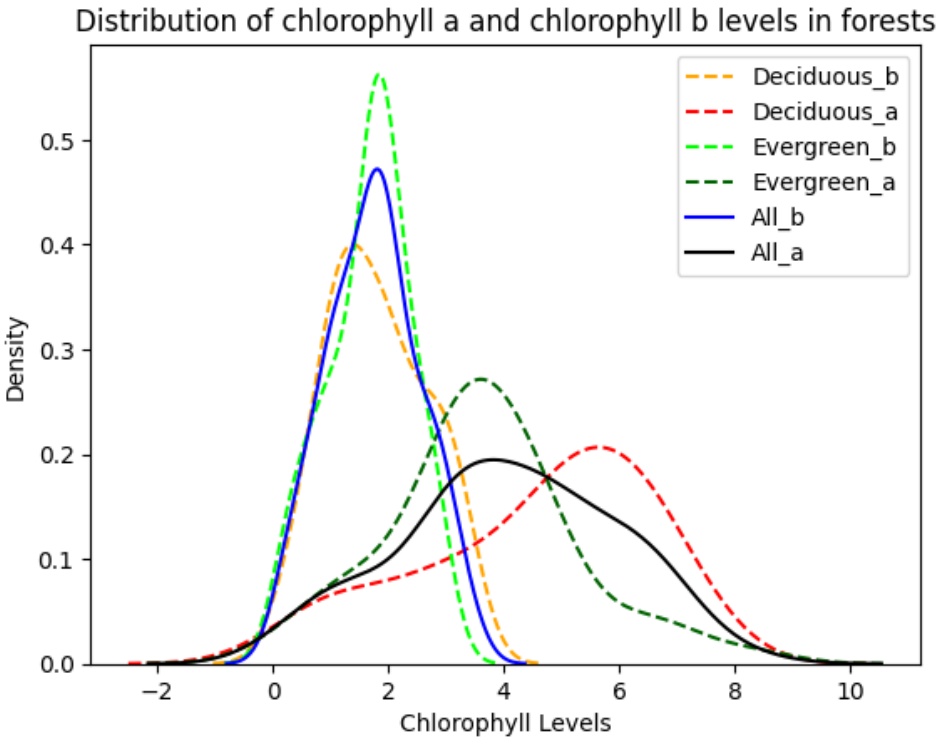


Figure 4: Density plot comparing chlorophyll a and b distributions

In this Sample dataset, regardless of forest type, it is observed that level of 'Chlorophyll a ' is greater than 'Chlorophyll b'. Though in all cases the curves look bell-curved and resemble the normal distribution, ' Chlorophyll a' curve looks more spread out , while 'Chlorophyll b' looks more centered and dense. This implies that 'Chlorophyll a' must have a higher variance than 'Chlorophyll b'. We will verify this further by summary statistics and Hypothesis testing.

2.5 Question 5: Summary Statistics

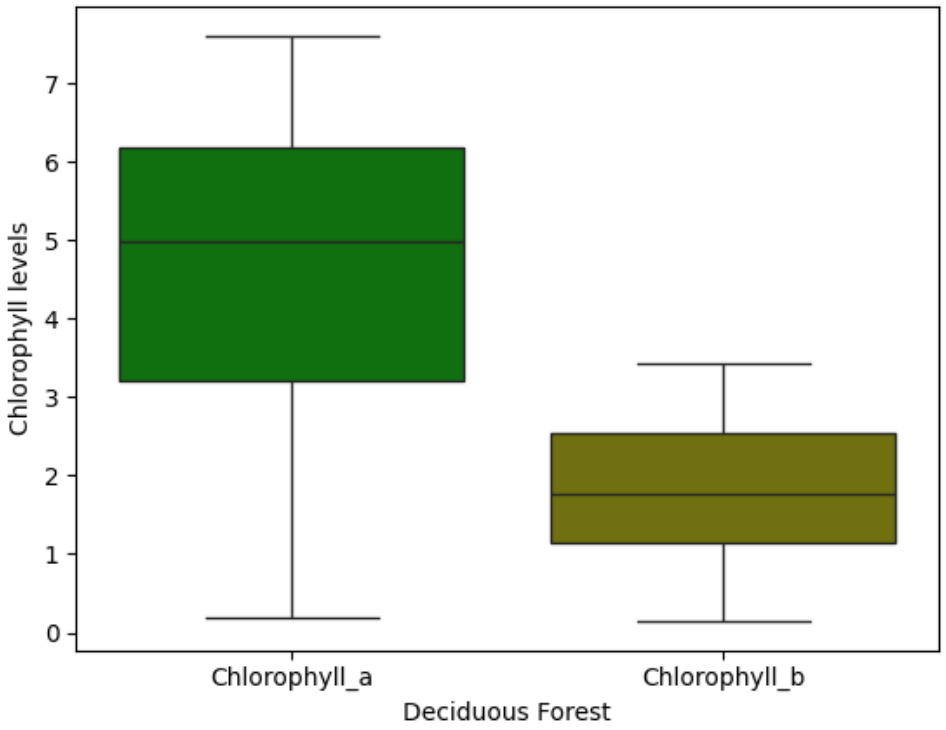
Task: Calculate summary statistics (mean, median, mode and standard deviation) of chlorophyll a and chlorophyll b measurements from Deciduous forests separately, Evergreen forests separately and both the forests together.

Forest Type	Chlorophyll Type	Mean	Median	Mode	Standard Deviation	Variance
Deciduous Forests	Chlorophyll_a	4.564366	4.974721	NaN	1.964399	3.858863
	Chlorophyll_b	1.817811	1.763595	NaN	0.863713	0.746000
Evergreen Forests	Chlorophyll_a	3.625706	3.531464	NaN	1.642608	2.698161
	Chlorophyll_b	1.666338	1.781322	NaN	0.718783	0.516649
All Forests	Chlorophyll_a	4.095036	4.098685	NaN	1.862227	3.467891
	Chlorophyll_b	1.742074	1.769393	NaN	0.794192	0.630741

Table 1: Summary statistics of chlorophyll measurements

2.6 Question 6: Boxplot Comparison

Task: In a same boxplot, compare how the distribution of chlorophyll a and chlorophyll b values compare in Deciduous forests and Evergreen forests.



In deciduous forests, over 75% of the 'chlorophyll a' values exceed those of 'chlorophyll b'. A comparison of their interquartile ranges (IQR) indicates that 'chlorophyll a' is more dispersed than 'chlorophyll b'. Additionally, 'chlorophyll a' exhibits slight negative skewness, whereas 'chlorophyll b' shows slight positive skewness.

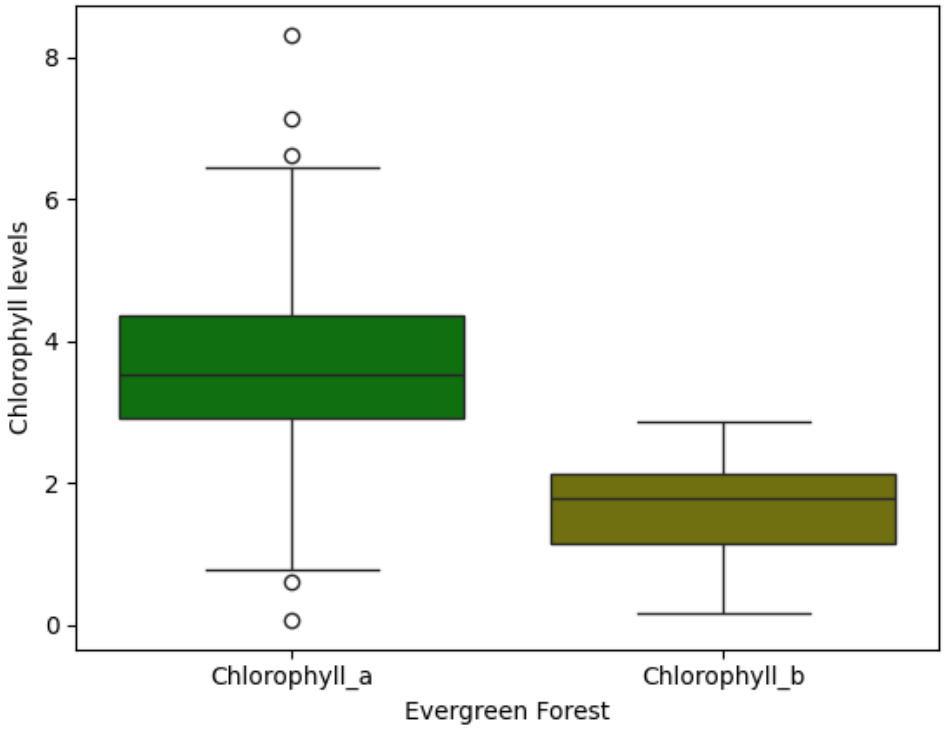


Figure 5: Boxplot comparing chlorophyll levels in Deciduous and Evergreen forests

Similarly, in evergreen forests, more than 75% of 'chlorophyll a' values are higher than 'chlorophyll b' values. The IQRs of both are comparable, but 'chlorophyll a' has a slight positive skew, while 'chlorophyll b' has a slight negative skew. Furthermore, the 'chlorophyll a' sample contains a few outliers.

2.7 Question 7: Variance Comparison

Task: Are the variances between chlorophyll a and chlorophyll b measurements differ significantly? Perform appropriate statistical tests to support your claim. Compare variances of chlorophyll content from Deciduous forests separately, Evergreen forests separately and both the forests together.

Forest Type	Chlorophyll Type	Variance
Deciduous Forests	Chlorophyll_a	3.858863
	Chlorophyll_b	0.746000
Evergreen Forests	Chlorophyll_a	2.698161
	Chlorophyll_b	0.516649
All Forests	Chlorophyll_a	3.467891
	Chlorophyll_b	0.630741

Figure 6: Variance of chlorophyll a and chlorophyll b for all samples

The variance differs significantly between 'chlorophyll a' and 'b' in all three cases. To check this we use the F-test for sample variances. We denote the variance of 'chlorophyll a' as σ_a^2 and 'chlorophyll b' as σ_b^2.

Deciduous Forests:

$$H_0 : \sigma_a^2 = \sigma_b^2, \quad H_A : \sigma_a^2 \neq \sigma_b^2$$
$$\sigma_a^2 = 3.858863, \quad ddof_1 = 49$$
$$\sigma_b^2 = 0.746, \quad ddof_2 = 49$$
$$F' = \frac{\sigma_a^2}{\sigma_b^2} = \frac{3.858863}{0.746} = 5.172739$$
$$F_{\alpha,v_1,v_2} = 1.607289$$

Since $F' > F_{\alpha,v_1,v_2}$, reject H_0 at 0.05 level of significance.

We conclude that the variances are not equal in Deciduous forests 'chlorophyll a' and 'b' values.

Evergreen Forests:

$$H_0 : \sigma_a^2 = \sigma_b^2, \quad H_A : \sigma_a^2 \neq \sigma_b^2$$
$$\sigma_a^2 = 2.698161, \quad ddof_1 = 49$$
$$\sigma_b^2 = 0.516649, \quad ddof_2 = 49$$
$$F' = \frac{\sigma_a^2}{\sigma_b^2} = \frac{2.698161}{0.516649} = 5.222425$$
$$F_{\alpha,v_1,v_2} = 1.607289$$

Since $F' > F_{\alpha,v_1,v_2}$, reject H_0 at 0.05 level of significance.

We conclude that the variances are not equal in Evergreen forests 'chlorophyll a' and 'b' values.

Both Forests Combined:

$$H_0 : \sigma_a^2 = \sigma_b^2, \quad H_A : \sigma_a^2 \neq \sigma_b^2$$
$$\sigma_a^2 = 3.467891, \quad ddof_1 = 99$$
$$\sigma_b^2 = 0.630741, \quad ddof_2 = 99$$
$$F' = \frac{\sigma_a^2}{\sigma_b^2} = \frac{3.467891}{0.630741} = 5.498122$$
$$F_{\alpha,v_1,v_2} = 1.394061$$

Since $F' > F_{\alpha,v_1,v_2}$, reject H_0 at 0.05 level of significance.

We conclude that the variances are not equal in Both forests 'chlorophyll a' and 'b' values.

2.8 Question 8: Hypothesis Testing for Mean Difference

Task: Test whether the mean of chlorophyll a is greater than the mean of chlorophyll b in all three cases.

From the data, we observe that the mean chlorophyll a levels are higher than chlorophyll b levels in all three cases. Since both measurements are taken from the same tree within a specific forest, their values are likely correlated. Therefore, the appropriate statistical test to use is the paired t-test.

We define:

- μ_a as the mean chlorophyll a level
- μ_b as the mean chlorophyll b level

Degrees of freedom:

- 49 for each forest type separately
- 99 when both forests are analyzed together

Significance level (α): 5

$$H_0 : \mu_a \leq \mu_b, \quad H_A : \mu_a > \mu_b$$

The following are the p-values for each case:

Forest Type	p-value	Conclusion
Deciduous	2.056196e-12	$p < 0.05$, reject H_0
Evergreen	6.614613e-10	$p < 0.05$, reject H_0
Both	1.904868e-20	$p < 0.05$, reject H_0

Table 2: Hypothesis testing results for mean comparison

Since the p-values in all three cases are much smaller than 0.05, we reject the null hypothesis at the 5% significance level. This confirms that, in the population, the mean chlorophyll a level is significantly higher than the mean chlorophyll b level.

3 Code and Data Access

The Python code used for analysis can be accessed at the following link: Google Drive Link.