Anirudh Rao
BE21B004

BT3041 – Analysis and Interpretation of Biological Data

Assignment 1

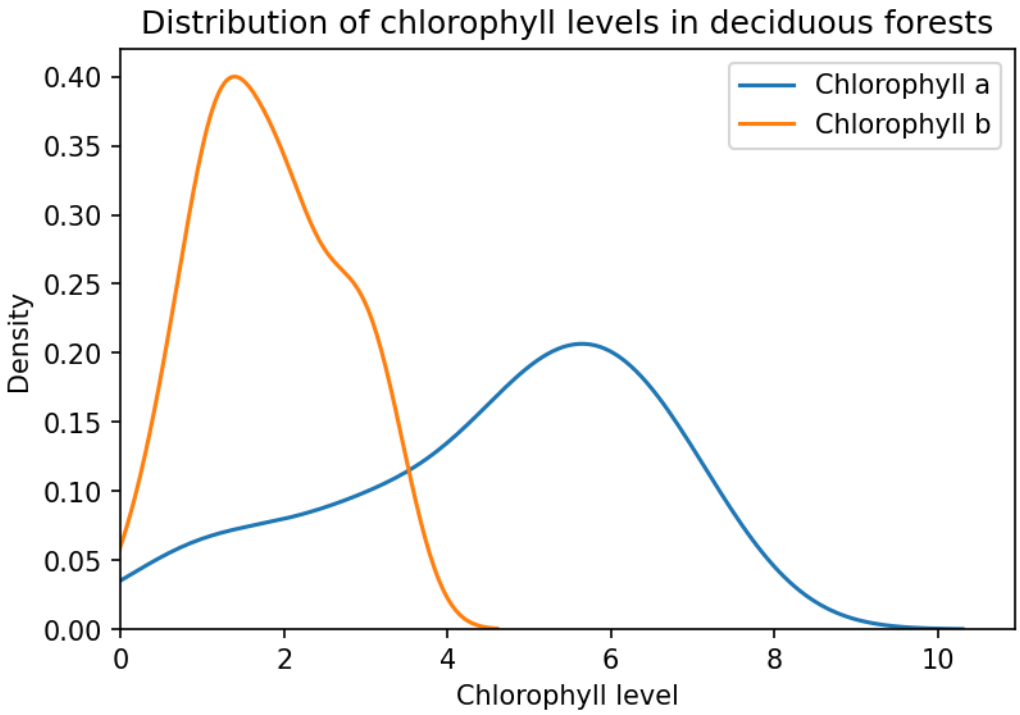
Report

The link to the Python notebook used to generate the visualisations and perform the statistical analysis of the given dataset is given below:

https://drive.google.com/file/d/1VG17owDqvToKZg5HLO373n8NHS75h2vh/view?usp=sharing

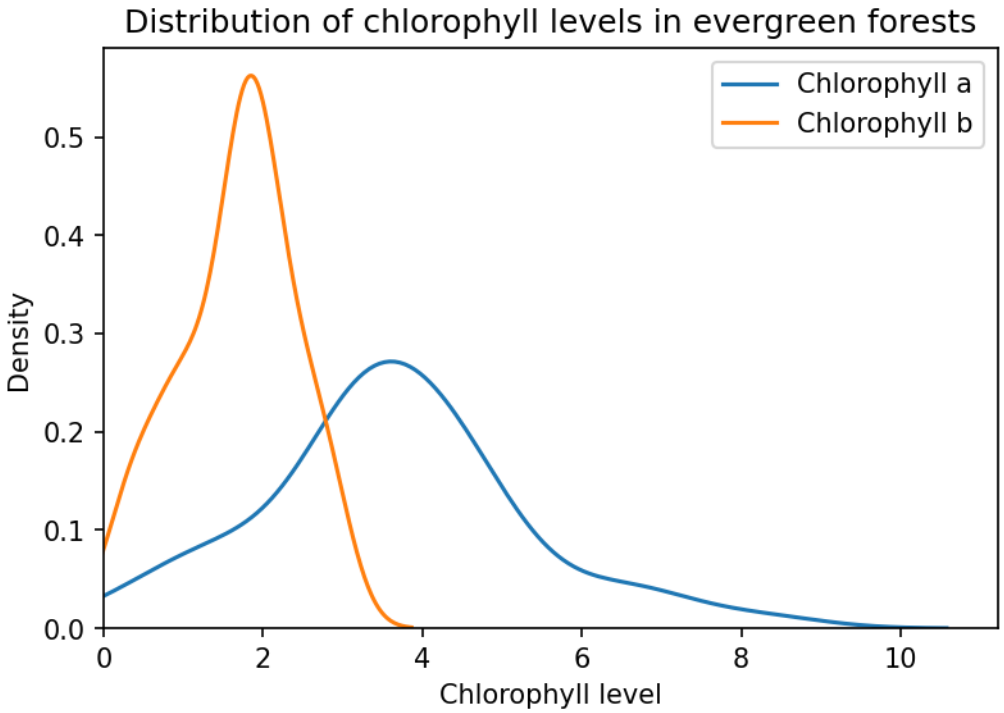
1

The distribution of chlorophyll a and chlorophyll b values in **deciduous** forests was visualised using a density plot. To prevent the density plot from extending to negative values, the plot was clipped on the left side at 0.



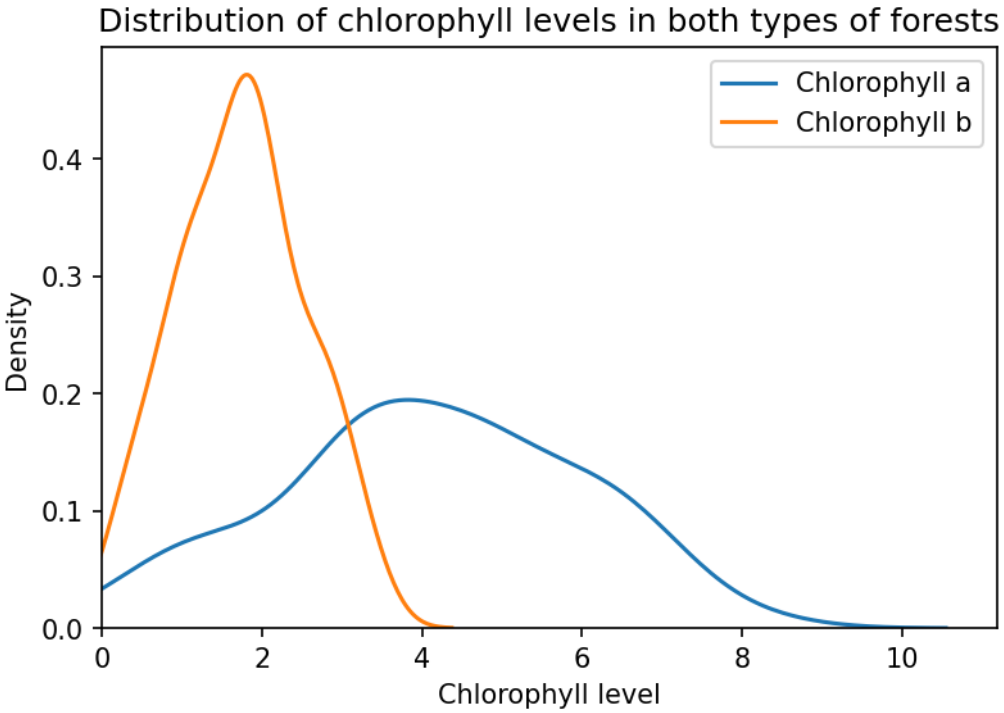
2

The distribution of chlorophyll a and chlorophyll b values in **evergreen** forests was visualised using a density plot. To prevent the density plot from extending to negative values, the plot was clipped on the left side at 0.



3

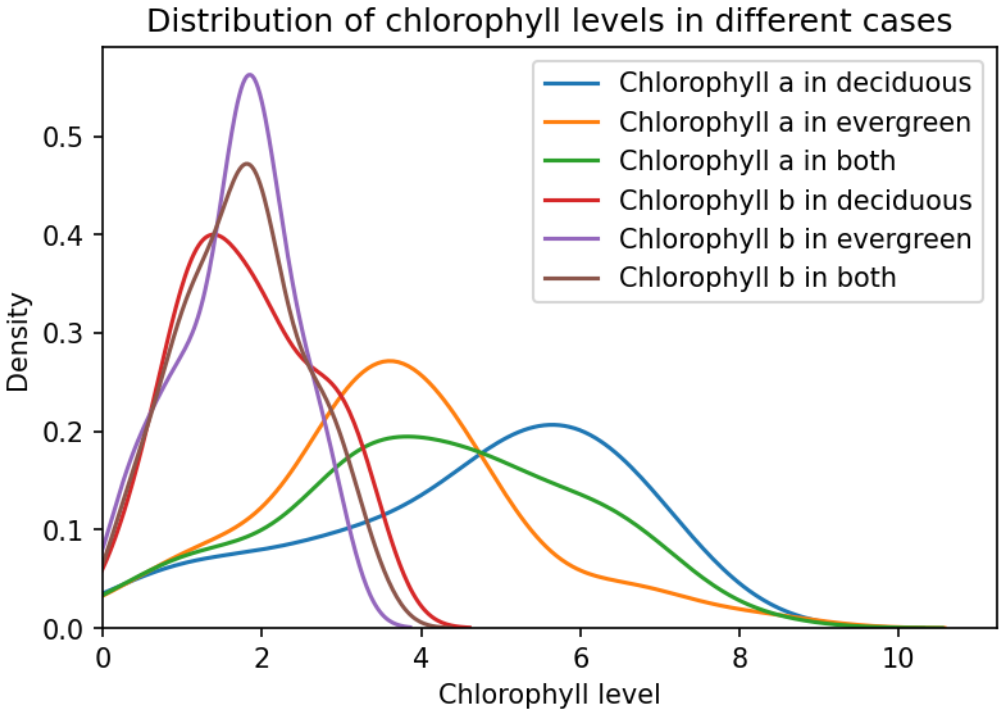
The distribution of chlorophyll a and chlorophyll b values in **both types** of forests was visualised using a density plot. To prevent the density plot from extending to negative values, the plot was clipped on the left side at 0.



From these three plots, we can observe that chlorophyll a levels tend to be higher than chlorophyll b levels.

4

The three plots were then plotted together to observe the differences in the distributions.



The following observations can be made:

- Chlorophyll a levels tend to be higher than chlorophyll b levels
- Chlorophyll a levels in deciduous forests tend to be higher than those in evergreen forests.
- Chlorophyll b levels in deciduous forests tend to be lower than those in evergreen forests.

5

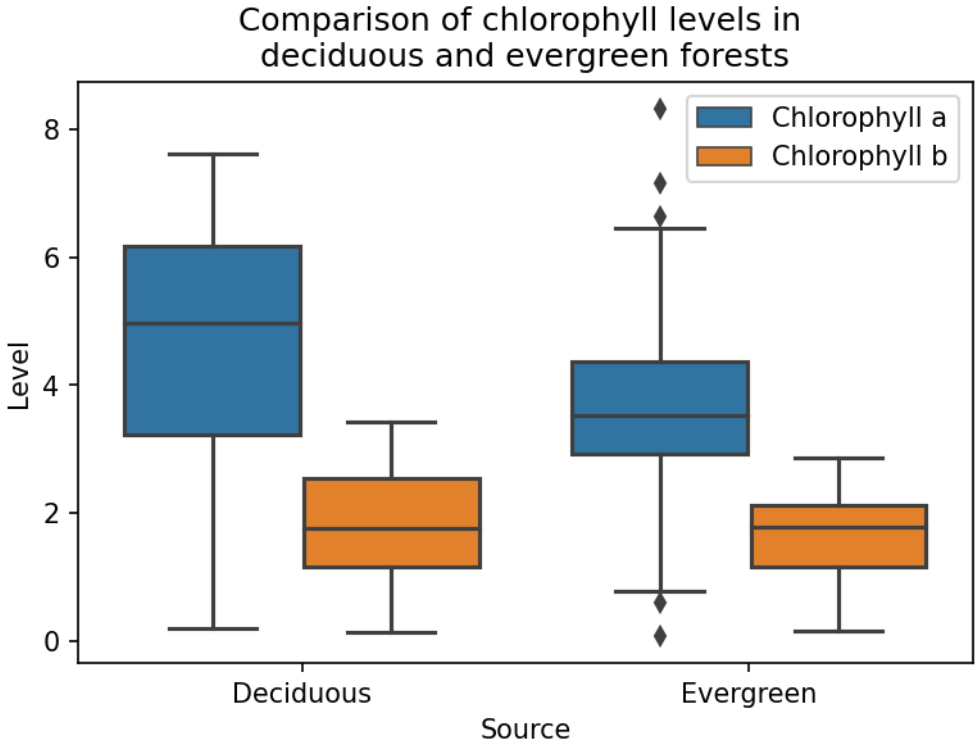
Summary statistics of chlorophyll levels were computed and are shown below:

		Deciduous	Evergreen	Both
Mean	Chl. a	4.56437	3.62571	4.09504
	Chl. b	1.81781	1.66634	1.74207
Median	Chl. a	4.97472	3.53146	4.09868
	Chl. b	1.76360	1.78132	1.76939
Mode	Chl. a	None	None	None
	Chl. b	None	None	None
Standard Deviation	Chl. a	1.96440	1.64261	1.86223
	Chl. b	0.86371	0.71878	0.79419

No modes exist as all values occur exactly once. The SD computed is the sample SD.

6

The distributions of chlorophyll a and chlorophyll b levels in deciduous and evergreen forests were compared using a boxplot, which is shown below:



- As seen earlier, chlorophyll b levels tend to be lower than chlorophyll a levels in both types of forests.
- Chlorophyll a levels in deciduous forests tend to be higher than those in evergreen forests. However, there are a few outliers in evergreen forests when it comes to chlorophyll a levels.
- Chlorophyll b levels in both types of forests have similar distributions.

7

- To separately compare the variances of the chlorophyll a and chlorophyll b levels in deciduous, evergreen, and both types of forests, we make use of the ***F*-test**, where the null hypothesis is that the variances are the same.
- The F-statistic in each case was computed as $F = \frac{s_a^2}{s_b^2}$, where s_a is the standard deviation in chlorophyll a levels and s_b is the standard deviation in chlorophyll b levels.
- This F-statistic was compared to a critical F-value obtained from the F-distribution corresponding to $\alpha = 0.05$, $v_1 = n_a - 1$, $v_2 = n_b - 1$, where n_a and n_b are the number of samples of chlorophyll a levels and chlorophyll b levels respectively.
- If the F-statistic is greater than the critical F-value, we conclude that the variances of the chlorophyll a and b levels are significantly different. Else, they are similar.

The following table summarises the result of the statistical tests performed:

	Deciduous	Evergreen	Both
***F*-statistic**	5.17274	5.22242	5.49812
Critical *F*-value	1.60729	1.60729	1.39406
Conclusion	Reject null hypothesis Variances are different	Reject null hypothesis Variances are different	Reject null hypothesis Variances are different
***p*-value**	2.61366e-08	2.22246e-08	5.55111e-16

Thus, we conclude that in all three cases, the variances of chlorophyll a levels and chlorophyll b levels are significantly different.

8

Since every sample in the dataset gives rise to a pair of measurements – one chlorophyll a level and one chlorophyll b level – it is appropriate to use a **paired *t*-test** while comparing the means of these two levels in the three cases (deciduous, evergreen, both types).

In all three cases, the null hypothesis we choose to adopt is that the mean of chlorophyll a levels is lesser than or equal to the mean of chlorophyll b levels.

$H_0: \mu_a \leq \mu_b$

$H_1: \mu_a > \mu_b$

We choose a significance level $\alpha = 0.05$ for our statistical tests. This is a one-tailed test.

If the p-value of the t-statistic computed is lesser than α, we reject the null hypothesis.

The following table summarises the result of the statistical tests performed:

	Deciduous	Evergreen	Both
***t*-statistic**	9.10513	7.45290	11.59395
Critical *t*-value	1.66055	1.66055	1.65259
***p*-value**	2.05619e-12	6.61461e-10	1.90487e-20
Conclusion	Reject null hypothesis	Reject null hypothesis	Reject null hypothesis

Thus, we conclude that in all three cases, the mean of chlorophyll a levels is significantly greater than the mean of chlorophyll b levels. This is in line with the observations we made initially with our visualisations.