DS5110 Homework 2 - Data Collection - Yueheng Yuan

1. Explain the data collection process

The dataset was collected by measuring two key attributes: leaf length and leaf width from three types of plants (ZZ Plant, Snake Plant, and Dumb Cane Plant). All measured plants grow close to each other, which shares similar conditions such as sunshine and moisture. Each measurement was recorded based on the longest dimension of the leaf (leaf length) and the widest part of the leaf (leaf width) in centimeters.

2. Instrument Used for Data Collection

A calibrated ruler was used to measure the leaf length and width, ensuring the precision of measurement results in centimeters.

3. Accuracy and Precision of the Instrument

Accuracy: the ruler is calibrated at the industrial level with the accuracy of ±0.02 mm, which ensures minimal measurement error for various leaves.

Precision: using the same instrument to measure the same plant in the same day should yield similar results.

4. Number of Data Points Collected and Why

The dataset contains 60 data points (N = 60) across three species, where each species has 20 sample measurements (n = 20).

The sample size is determined to contain enough data points for analysis while balancing statistical robustness with practical feasibility.

5. Define the Dataset Size in Terms of N (full data set size) and n (each subset size)

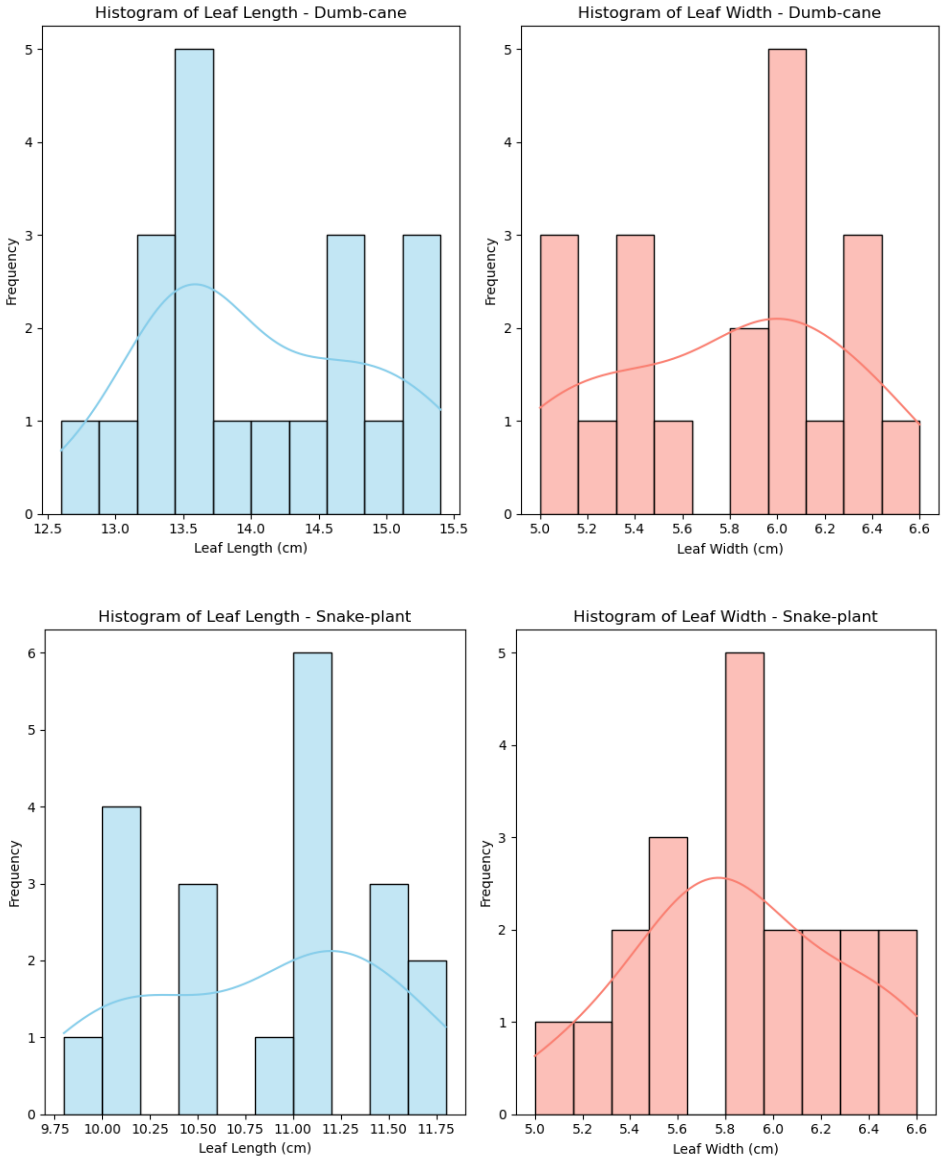
N = 60, n = 20 for ZZ Plant, Snake Plant, and Dumb Cane Plant each

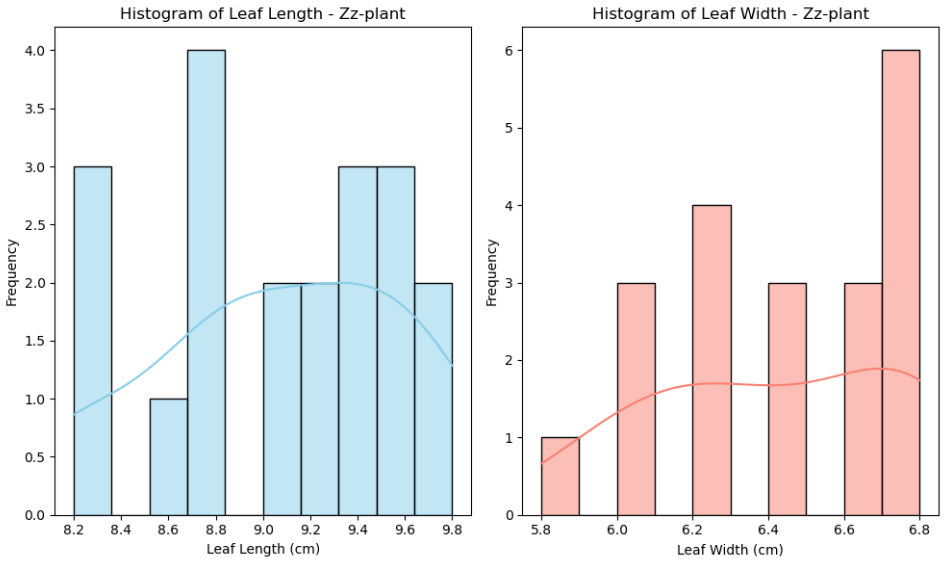
6. Explain Any Problems During Data Collection

Though environmental factors are mostly aligned during data collection, the natural curvature of leaves poses difficulty for width measurements and and may affect the consistency of measured results.

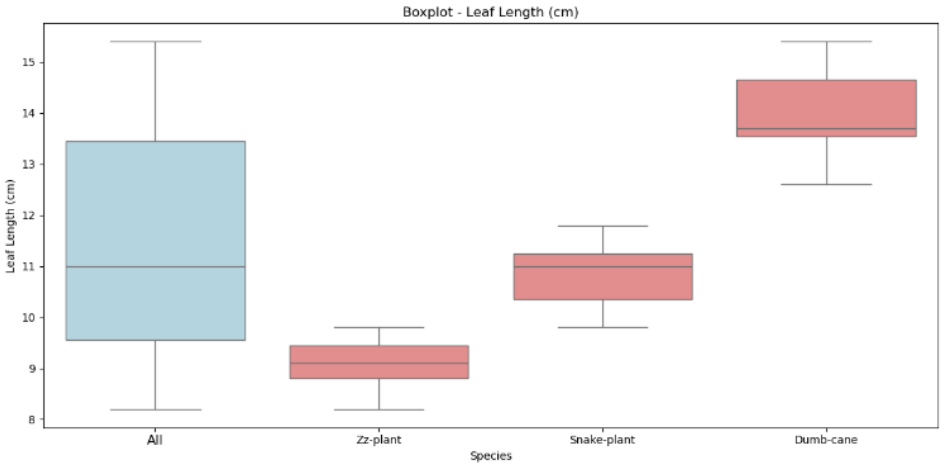
Analysis

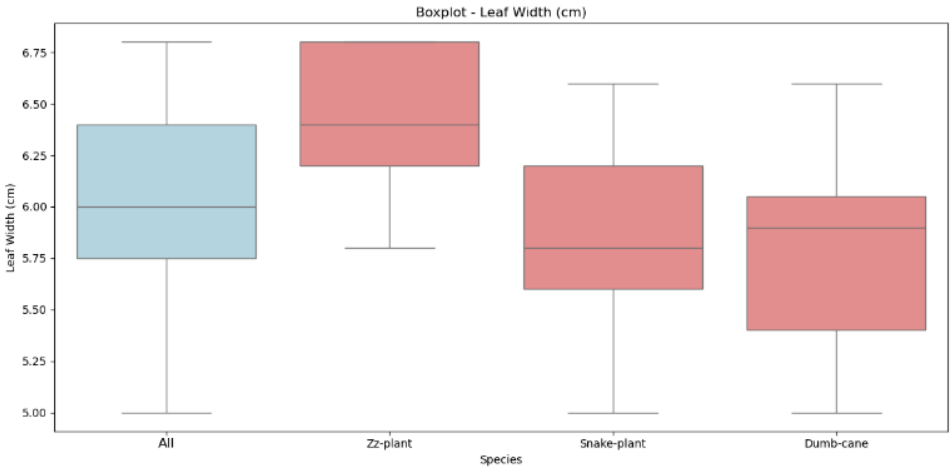
1. Histogram



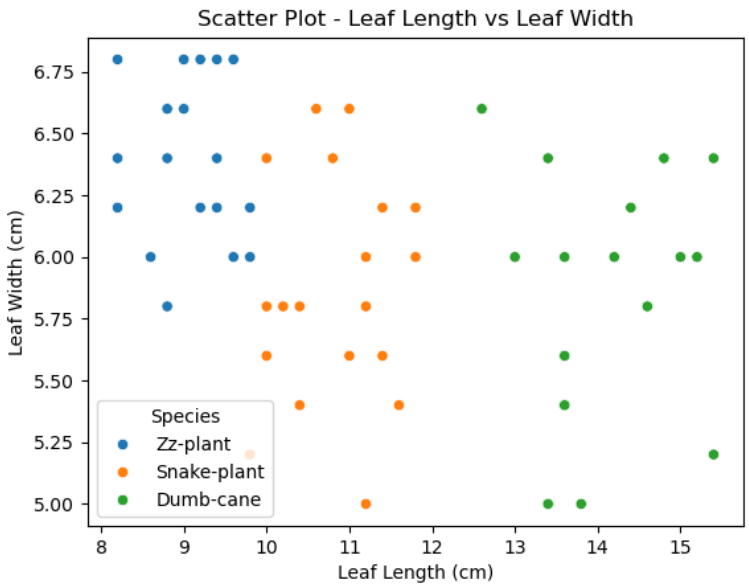

Histogram of Leaf Length - Zz-plant
Frequency
Leaf Length (cm)
Histogram of Leaf Width - Zz-plant
Frequency
Leaf Width (cm)

2. Boxplot


Boxplot - Leaf Length (cm)
Leaf Length (cm)
All
Zz-plant
Snake-plant
Dumb-cane
Species


Boxplot - Leaf Width (cm)
Leaf Width (cm)
All
Zz-plant
Snake-plant
Dumb-cane
Species

3. scatter plot



(Feel free to ignore those contents and jump to the next question)

For ZZ-plant,

Linear Regression: slope = -0.0213, intercept = 6.6135

The slope suggests a very small negative relationship between leaf length and width, and the intercept is 6.6135, which is not meaningful biologically.

R-value = -0.0333, p-value = 0.8890, se = 0.1507

The R-value is very close to zero, implying no linear relationship between the two variables. This conclusion is also supported by the high p-value (>0.05 as the significance level), which fails to reject the null hypothesis of no significant linear relationship.

The standard error is relatively low, suggesting that the data points are relatively closer around the fitted regression line.

Pearson correlation coefficient = -0.0333

The negative Pearson correlation coefficient indicates a very weak negative correlation between leaf length and width, which also supports the results above.

For Snake-plant,

Linear Regression: slope = 0.0652, intercept = 5.1527

It may present a mild linear relationship based on the slope of the linear regression line.

R-value = 0.0930, p-value = 0.6965, se = 0.1645

Similarly, the R-value is relatively close to 0 with a high p-value, suggesting no significant statistical relationship between leaf length and width. The standard error indicates acceptable variation around the regression line.

Pearson correlation coefficient = 0.0930

The low Pearson correlation coefficient supports the result of p-value such that the linear correlation between leaf length and width is not statistically significant.

For Dumb-cane plant,

Linear Regression: slope = 0.0836, intercept = 4.6047

A mild positive linear relationship may be present based on the slope.

R-value = 0.1351, p-value = 0.5701, se = 0.1445

Similarly, as above, the R-value is close to 0, indicating a very weak correlation. The p-value is slightly above the significance level, which is not solid enough to reject the null hypothesis. The standard error indicates acceptable deviation of data points from the regression line.

Pearson correlation coefficient = 0.1351

Like above, the weak positive Pearson correlation coefficient supports the result of p-value such that there is no strong relationship between leaf length and width.

4. Explain in terms of variance / mean / median / SD

LeafLengthCm	Mean	Median	Variance	SD
ZZ plant	9.07	9.1	0.266421	0.516159859
Snake plant	10.85	11.0	0.403684	0.635361315
Dumb cane	14.06	13.7	0.669895	0.818471136

LeafWidthCm	Mean	Median	Variance	SD
ZZ plant	6.42	6.4	0.109053	0.330231737
Snake plant	5.86	5.8	0.198316	0.445326846
Dumb cane	5.78	5.9	0.256421	0.506380292

Generally, in terms of leaf length, the dumb cane plant appears to have the longest leaf and ZZ with the shortest leaf based on the mean and median value. For leaf width, ZZ plants tend to have slightly wider leaf than snake plants and dumb plants.

For all three plant species, the mean and median values are close to each other in terms of both leaf length and leaf width, potentially implying a fairly symmetric distribution without outliers. None of the variance or standard deviation appears to be significant, indicating that data collected is consistent among species.

For ZZ plants, the variance and SD for leaf length are moderate (Coefficient of variation: CV = 5.69%), indicating that the leaf lengths are somewhat spread out from the mean but not extremely varied. For leaf width, CV = 5.14%, meaning data points are tightly clustered and are consistent around the mean value.

For snake plants, it has slightly higher SD for both leaf length (CV = 5.85%) and width (CV = 8.76%) than ZZ plants, so data points are more spread out with greater variability.

For dumb cane plants, it has slightly higher SD for both leaf length (CV = 5.82%) and width (CV = 7.60%) among all recorded species, which means data points are more spread out around the mean.

5. What you can infer with data and graphs

In most cases, the dumb cane plant tends to have the longest leaf and ZZ with the shortest leaf. For leaf width, ZZ plants tend to have slightly wider leaf, while the leaf width for snake plants and dumb plants are close.

For all species, the leaf length and leaf width do not appear to have strong or statistically significant linear relationships (based on the R-value, p-value and Pearson correlation coefficient under the scatter plot).

For data collection, data points are generally tightly clustered around the mean value for each species, and no clear outlier observed.