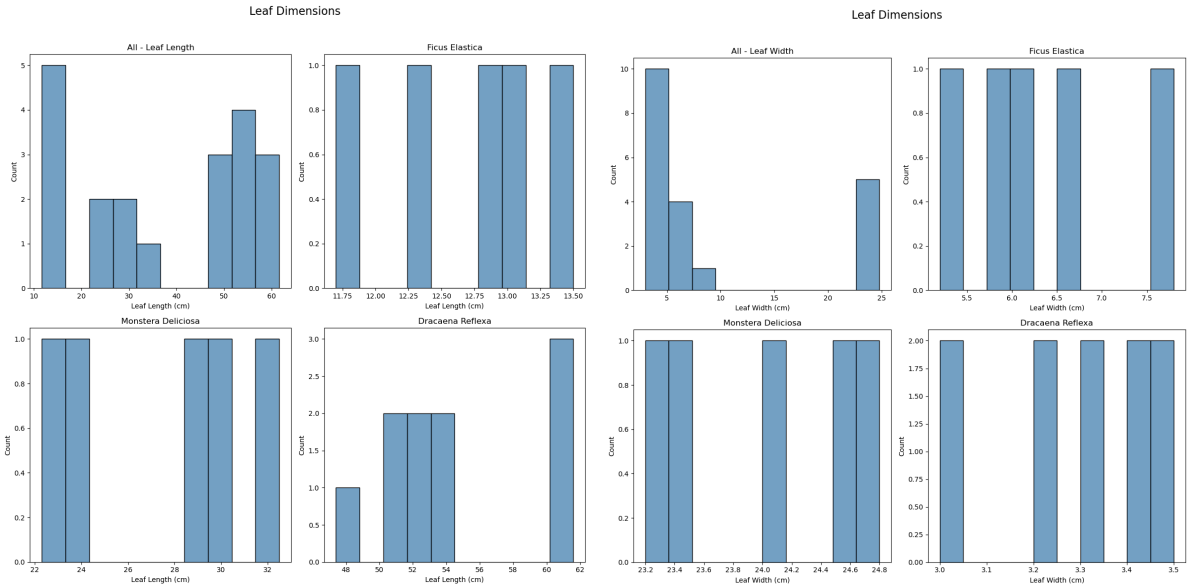
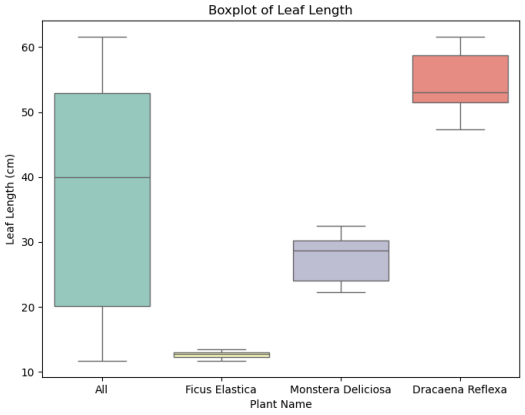
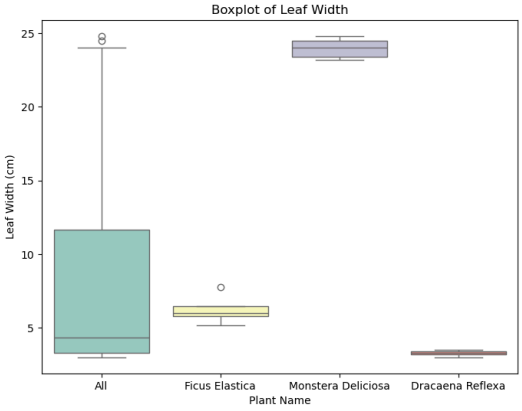
Graphs:

Histograms:

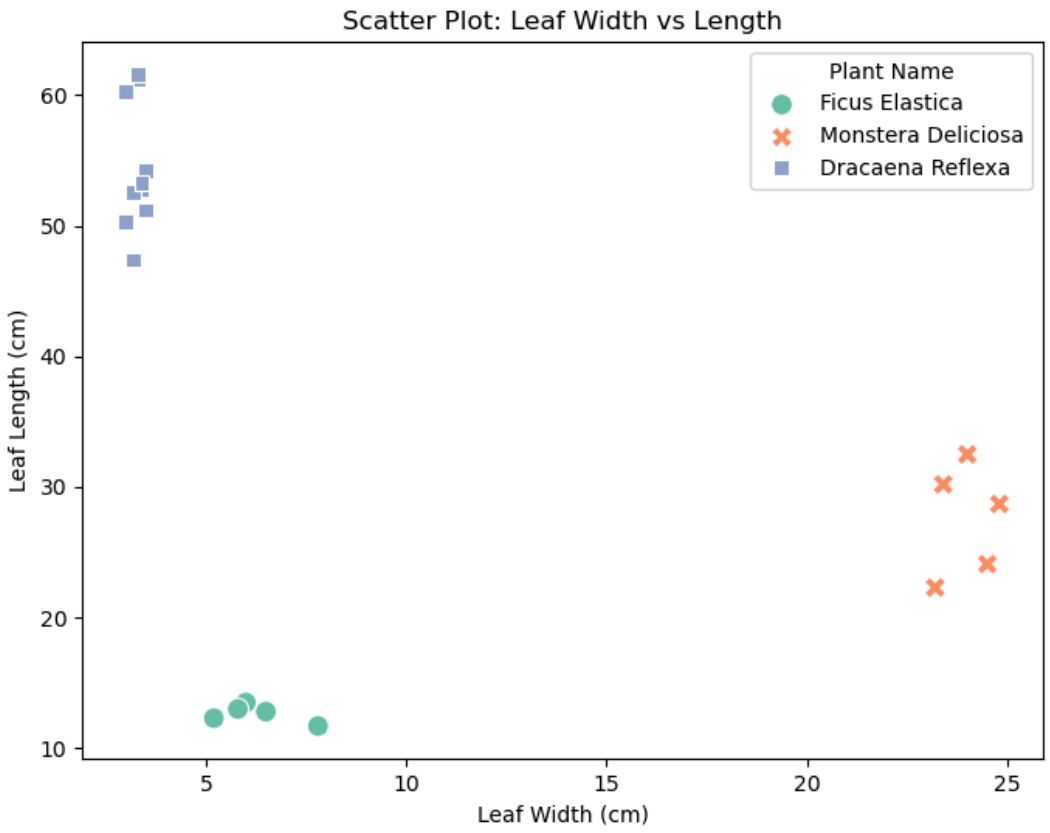


Boxplots:





Scatter Plot:



Explain your data collection process.

- To collect the data, I began by using the PlantNet application to identify the plant species. The three plants I identified were **Ficus Elastica**, **Monstera Deliciosa**, and **Dracaena Reflexa**.
- For each plant, I selected multiple leaves to measure, ensuring that I captured variation both within a single plant and across different plants of the same species. The measurements recorded included the **leaf-width** (measured at the widest point of the leaf) and **leaf-length** (measured from the base of the leaf to its tip). I aimed to create a dataset that reflected the natural variation in leaf size for each species, providing a diverse and representative sample.

What instrument did you use to collect data with?

- I used the following instruments:
    - **PlantNet application**: To accurately identify the species of each plant.

    - **Ruler and measurement app (Measure App in iPhone)**: To measure the leaf-width and leaf-length precisely.
- Measurements were taken manually, and data was recorded directly into a google spreadsheet.

Argue the accuracy and precision of your instrument.

- **Accuracy**:
    - The ruler and Measure I used is accurate to millimeters, which is sufficient for measuring leaf sizes in the range of a few centimeters. This level of accuracy ensures the data reliably reflects the actual dimensions of the leaves.
- **Precision**:
    - The precision of the instrument is high, as the ruler and app both offer clear, consistent measurements. The main reason I prepare to measure tools is for cross validation. Precision is limited only by potential human error, such as misaligning the ruler or measuring a bent leaf which can be reduced due to validation.
- **Potential Errors**:
    - Human error in positioning the ruler or app could introduce small inconsistencies, typically within ±0.1 cm.
    - Environmental factors, such as difficulty accessing certain leaves, may have also affected measurements slightly. However, these factors were mitigated by careful handling during data collection.

How many data points did you collect? Why?

- **Total data points (N)**: I collected **20 data points**.
- **Data points per plant species (n)**:
    - **Ficus Elastica**: 5 data points
    - **Monstera Deliciosa**: 5 data points
    - **Dracaena Reflexa**: 10 data points
- **Reasoning**:
    - I collected at least 5 samples for **Ficus Elastica** and **Monstera Deliciosa** because these plants had larger leaves, which required more careful selection.
    - I collected 10 samples for **Dracaena Reflexa** because its smaller leaves were easier to measure and more samples were readily available.
    - This approach ensured that each species was adequately represented in the dataset while balancing the time and effort required for measurements.

Define the size of your data in terms of both N (full data set size) and n (each subset size).

- **N (full dataset size)**: 20 total data points.
- **n (subset sizes for each plant species)**:

- **Ficus Elastica**: 5 data points collected from two individual plants of the same species for comparison.
- **Monstera Deliciosa**: 5 data points collected from two individual plants of the same species for comparison.
- **Dracaena Reflexa**: 10 data points collected from three individual plants of the same species for comparison.

Explain any problems that you ran into during the data collection process.

- While using the PlantNet application, some plants were challenging to identify. This required cross-referencing with other resources or making educated guesses based on the most probable match.
- Leaves were not always flat, which made measuring the full width or length more difficult. For curved leaves, it was challenging to position the ruler accurately.
- We recently suffered a snowstorm, so it is difficult to find outdoor plants for cross comparison.

Explain each graph in terms of variance, mean, median, and standard deviation.

Stats:

1. In order to calculate stats data, we can use groupby and agg functions in Pandas. Benefit by Numpy, we can calculate mean, var(variance), and std (standard deviation).

```
grouped = df.groupby("Plant Name")


stats = grouped.agg(
   Mean_Width=("Leaf-Width (cm)", "mean"),
   Median_Width=("Leaf-Width (cm)", "median"),
   Variance_Width=("Leaf-Width (cm)", "var"),
   StdDev_Width=("Leaf-Width (cm)", "std"),
   Mean_Length=("Leaf-Length (cm)", "mean"),
   Median_Length=("Leaf-Length (cm)", "median"),
   Variance_Length=("Leaf-Length (cm)", "var"),
   StdDev_Length=("Leaf-Length (cm)", "std")
)


stats.to_csv(os.path.join(os.path.abspath('./stats/'), "stats.csv"), index=True)
```

We can get the data:

```
                Mean_Width  Median_Width  Variance_Width  StdDev_Width  Mean_Length  Median_Length  Variance_Length  StdDev_Length
Plant Name
Dracaena Reflexa        3.28           3.3        0.032889      0.181353        54.49          53.05        23.989889       4.897947
Ficus Elastica          6.26           6.0        0.958000      0.978775        12.66          12.80         0.473000       0.687750
Monstera Deliciosa     23.98          24.0        0.472000      0.687023        27.56          28.70        18.078000       4.251823
```

Histograms

1. **Leaf Length**:

- **Variance**:
  - *Dracaena Reflexa* has the largest spread in leaf length, ranging from ~47 to 61 cm, indicating high variance.
  - *Ficus Elastica* shows the least variance, with leaf lengths clustered tightly between ~12 and 13 cm.
  - In detail, *Dracaena Reflexa* has the variance of leaf length: 23.98. It is significantly greater than others. *Ficus Elastica* shows the least variance of 0.473. And the *Monstera Deliciosa* has the variance of 18.078
- **Mean & Median**:
  - *Dracaena Reflexa* has a mean (54.49) and median (53.05) leaf length in cm.
  - *Ficus Elastica* has a mean (12.66) and median (12.80) leaf length in cm.
  - *Monstera Deliciosa* has a mean (27.56) and median (28.70) leaf length in cm, indicating consistency.
- **Standard Deviation**:
  - *Dracaena Reflexa* has a standard deviation of 4.90 for leaf length in cm.
  - *Ficus Elastica* has a standard deviation of 0.69 for leaf length in cm.
  - *Monstera Deliciosa* has a standard deviation of 4.25 for leaf length in cm.
  - *Dracaena Reflexa* has the highest standard deviation, reflecting its large spread.

2. **Leaf Width**:
   - **Variance**:
     - *Dracaena Reflexa* has the variance: 0.032 in leaf width
     - *Ficus Elastica* has the variance: 0.96 in leaf width
     - *Monstera Deliciosa* has the variance: 0.47 in leaf width
     - *Monstera Deliciosa* has the largest variance in leaf width, while *Dracaena Reflexa* has minimal variation
   - **Mean & Median**:
     - *Dracaena Reflexa* has a mean (3.28) and median (3.3) leaf width in cm.
     - *Ficus Elastica* has a mean (6.26) and median (6.0) leaf width in cm.
     - *Monstera Deliciosa* has a mean (23.98) and median (24.0) leaf width in cm.
     - The mean width of *Ficus Elastica* (~6.2 cm) aligns closely with its median.
   - **Standard Deviation**:
     - *Dracaena Reflexa* has a standard deviation of 0.18 for leaf width in cm.
     - *Ficus Elastica* has a standard deviation of 0.99 for leaf width in cm.
     - *Monstera Deliciosa* has a standard deviation of 0.47 for leaf width in cm.
     - *Monstera Deliciosa* displays a higher standard deviation compared to the narrow distribution of *Dracaena Reflexa*.

Box Plots

1. **Leaf Length Boxplot**:
   - Shows the spread and central tendencies of leaf lengths.
   - *Dracaena Reflexa* has a significantly higher range.

    - *Ficus Elastica* exhibits the tightest interquartile range (IQR), reflecting minimal variance.
2. **Leaf Width Boxplot**:
    - *Monstera Deliciosa* has wider leaves with a moderate range.
    - *Dracaena Reflexa* maintains a narrow width range, as does *Ficus Elastica*, but with slightly higher variance.

Scatter Plot: Leaf Width vs. Leaf Length

- **Variance**:
    - Clear grouping of species shows that *Monstera Deliciosa* occupies the upper-right corner (broad, long leaves), while *Ficus Elastica* is in the bottom-left (short, narrow leaves).
    - *Dracaena Reflexa* shows long but narrow leaves, forming a vertical cluster.
- **Inferences**:
    - The relationships between width and length are species-specific, with minimal overlap.
    - Outliers are minimal, confirming the consistency of measurements within each species.

What can you infer with data and graphs that you have?

The data clearly separates the three species based on their leaf dimensions.

1. *Ficus Elastica*: Compact leaves with low variance in both length and width.
2. *Monstera Deliciosa*: Large leaves with higher variance in both dimensions.
3. *Dracaena Reflexa*: Long, narrow leaves with high variance in length but low variance in width.

These differences could be used to classify or identify plants based on their leaf dimensions. The scatter plot and boxplots, in particular, are effective for visualizing interspecies differences.