Answer the following (20 points)

1. Explain your data collection process

I collected data by measuring LeafLength and LeafWidth of three tree species: Linden, Maple, and Oak. Since it's winter in Boston, collecting leaves was challenging as most trees had already shed their leaves. To overcome this, I focused on areas such as parks, large lawns, and botanical gardens, where leaves were more likely to remain. The leaves I collected were mostly from the ground, and I ensured they belonged to the intended species by identifying nearby trees based on their bark, branch structure, and tree markers when available. From each tree or its immediate surroundings, I selected the leaves, carefully avoiding those that were damaged or decayed, to ensure the dataset's quality.

2. What instrument did you use to collect data with?

I used a digital caliper to measure both LeafLength and LeafWidth. This instrument was ideal for the task due to its high precision, measuring up to 0.01 cm, which allowed me to capture even subtle differences between leaves. It was portable and easy to use in outdoor environments, making it suitable for fieldwork. Additionally, the caliper could be recalibrated easily, ensuring consistent accuracy during the data collection process. Its reliability and precision were crucial for obtaining a detailed and accurate dataset.

3. Argue the accuracy and precision of your instrument

The digital caliper I used was both accurate and precise for measuring leaf dimensions. It had an error margin of ±0.02 cm, which was negligible for the size differences I was studying, as most variations between species were larger than this margin. To ensure accuracy, I calibrated the caliper before and during data collection. Its resolution of 0.01 cm allowed me to detect very small differences, ensuring detailed and reliable measurements. In cases of suspected handling errors, I repeated measurements, and the caliper consistently produced accurate results.

4. How many data points did you collect? Why?

I collected a total of 150 data points, with 50 samples for each species (Linden, Maple, and Oak). This number was chosen to ensure a statistically significant dataset, enabling meaningful calculations of metrics such as variance, mean, median, and standard deviation. Equal representation of each species ensured a balanced dataset, which is essential for fair comparisons across groups. Despite the challenges of collecting data in winter, I achieved this number by focusing on multiple collection sites, making the dataset reliable and comprehensive.

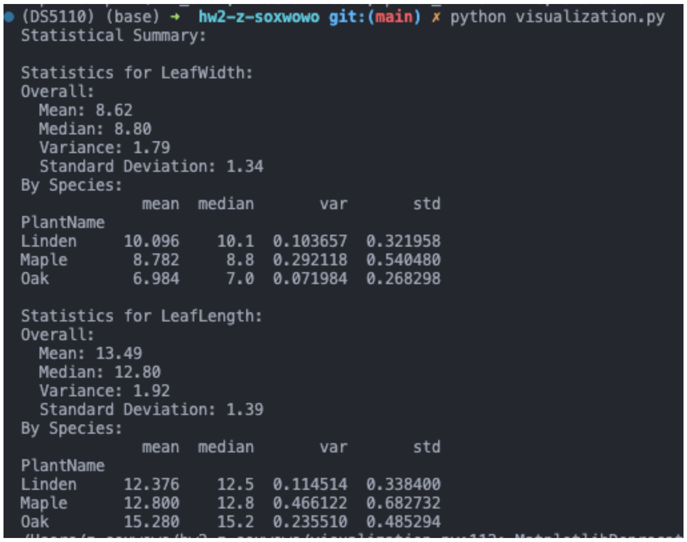
5. Define the size of your data in terms of both N (full data set size) and n (each subset size)

The dataset has a full size (N) of 150 samples, which includes all three species combined. Each subset (n) contains 50 samples for the three species: Linden, Maple, and Oak. This equal distribution ensures a balanced dataset, enabling fair and accurate comparisons between species without introducing bias. The balanced structure also enhances the statistical reliability of the dataset for analysis.

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

During the data collection process, I encountered several challenges. Since it was winter in Boston, most trees had already shed their leaves, and finding intact leaves was difficult. I collected leaves from the ground rather than directly from the trees, but many of these were damaged or decayed, reducing the usable sample pool. Identifying the species was also challenging, as there were no leaves directly on the trees; I relied on bark texture, branch patterns, and tree markers to confirm the species. Weather conditions, such as cold temperatures and occasional snowfall, complicated the process, as some leaves were damp or frozen, requiring extra care before measurement. Access to collection sites was another issue, as not all parks and lawns were accessible, which forced me to search extensively.

Analysis/Visualization – (50 points)

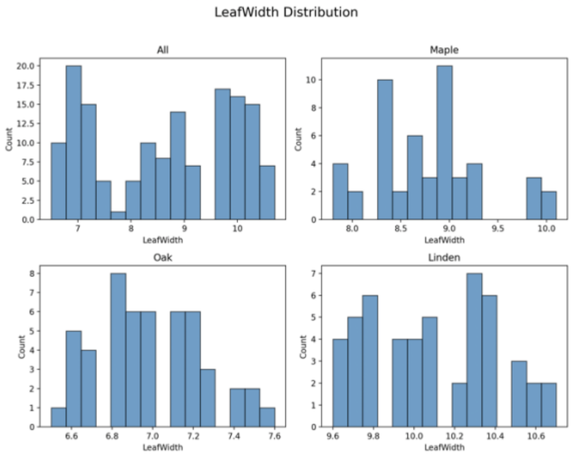


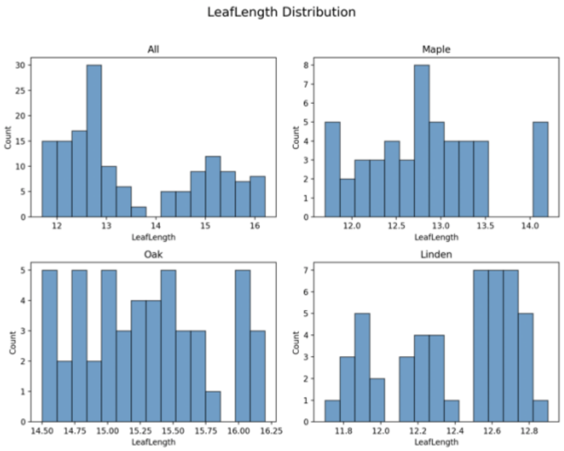
```
(DS5110) (base) → hw2-z-soxwowo git:(main) ✗ python visualization.py
Statistical Summary:

Statistics for LeafWidth:
Overall:
  Mean: 8.62
  Median: 8.80
  Variance: 1.79
  Standard Deviation: 1.34
By Species:
             mean  median       var       std
PlantName
Linden     10.096    10.1  0.103657  0.321958
Maple       8.782     8.8  0.292118  0.540480
Oak         6.984     7.0  0.071984  0.268298

Statistics for LeafLength:
Overall:
  Mean: 13.49
  Median: 12.80
  Variance: 1.92
  Standard Deviation: 1.39
By Species:
             mean  median       var       std
PlantName
Linden     12.376    12.5  0.114514  0.338400
Maple      12.800    12.8  0.466122  0.682732
Oak        15.280    15.2  0.235510  0.485294
```

4. Explain each graph in terms of variance, mean, median, and standard deviation. (Histograms)





1. LeafWidth Distribution

- **Overall**:
  - The mean of **Leaf**
  - **Width** is 8.62, which is slightly less than the median of 8.80, indicating a minor right-skewed distribution in the overall dataset.
  - The variance is 1.79, and the standard deviation is 1.34, reflecting a moderate spread in the data, as seen in the wide range from approximately 6.5 to 10.5 in the histogram.
- **By Species**:

- **Linden**:
    - The mean (10.096) and median (10.10) are identical, suggesting a perfectly symmetric distribution.
    - Variance (0.10) and standard deviation (0.32) are very low, as shown by the narrow spread of values concentrated between 9.8 and 10.4.
- **Maple**:
    - The mean (8.78) and median (8.80) are very close, indicating a slightly symmetric distribution with a moderate spread.
    - The variance (0.29) and standard deviation (0.54) are higher than Linden, consistent with the broader range of values between 8.0 and 10.0.
- **Oak**:
    - The mean (6.98) and median (7.00) are nearly identical, suggesting a symmetric distribution for Oak's **LeafWidth**.
    - The variance (0.07) and standard deviation (0.27) are the smallest among all species, as seen in the narrow spread of values around 6.8 to 7.4.

2. LeafLength Distribution

- **Overall**:
    - The mean of **LeafLength** is 13.49, while the median is 12.80, showing a right-skewed distribution due to higher values from Oak.
    - Variance (1.92) and standard deviation (1.39) indicate a moderate spread, as reflected by the two clusters of values: one around 12–13 and another around 15–16.
- **By Species**:
    - **Linden**:
        - The mean (12.38) and median (12.50) are very close, showing a symmetric and compact distribution.
        - The variance (0.11) and standard deviation (0.34) are small, indicating that the data points are tightly concentrated between 12.0 and 12.8.
    - **Maple**:
        - The mean (12.80) and median (12.80) are identical, indicating a balanced distribution.

        - Variance (0.47) and standard deviation (0.68) show moderate variability, as values are spread between 12.0 and 14.0.

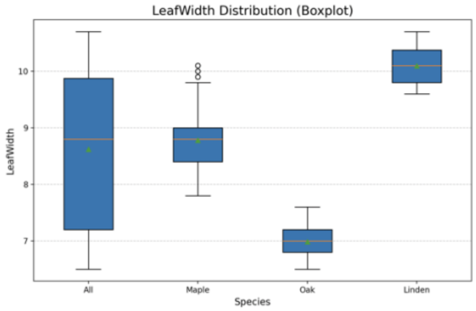
    - **Oak**:

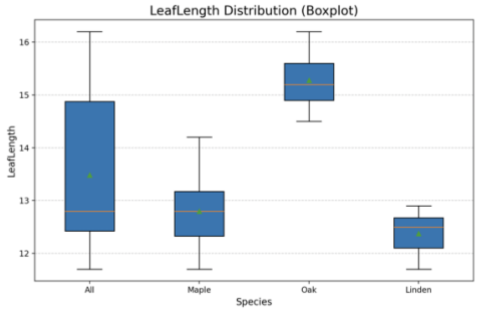
        - The mean (15.28) is slightly higher than the median (15.20), suggesting a slight right skew.

        - Variance (0.24) and standard deviation (0.49) indicate moderate spread, as values range between 14.5 and 16.0.

- **Linden** consistently exhibits the least variability for both **LeafWidth** and **LeafLength**, as seen by its low variance and compact histograms.
- **Maple** shows moderate variability in both features, with balanced distributions that are slightly more spread out compared to Linden.
- **Oak** has the most distinct characteristics, with the largest **LeafLength** values and the smallest **LeafWidth**, exhibiting a broader range for **LeafLength** and a tighter range for **LeafWidth**.

4. Explain each graph in terms of variance, mean, median, and standard deviation. (Boxplot)





1. LeafWidth Distribution

- **Overall**:

    - The mean for **LeafWidth** is 8.62 (green triangle), slightly below the median of 8.80 (red line), indicating a slight right-skewness.

    - The wide interquartile range and the presence of whiskers spanning from 7.0 to 10.0 reflect the high variance (1.79) and standard deviation (1.34) in the dataset.

- **By Species**:

    - **Linden**:

        - The mean (10.10) and median (10.10) are nearly identical, showing a symmetric distribution.
        - The small interquartile range and short whiskers confirm low variance (0.10) and standard deviation (0.32), meaning the data points are tightly clustered.
    - **Maple**:
        - The mean (8.78) is close to the median (8.80), indicating a balanced distribution.
        - The moderate interquartile range and presence of outliers highlight the higher variance (0.29) and standard deviation (0.54) compared to Linden.
    - **Oak**:
        - The mean (6.98) is nearly equal to the median (7.00), reflecting a symmetric distribution.
        - The smallest interquartile range and very short whiskers confirm the lowest variance (0.07) and standard deviation (0.27) among the species.

2. LeafLength Distribution

- **Overall**:
    - The mean (13.49) is higher than the median (12.80), indicating a slight right-skew caused by Oak's larger LeafLength values.
    - The large interquartile range and whiskers extending from 12.0 to 16.0 show high variance (1.92) and standard deviation (1.39).
- **By Species**:
    - **Linden**:
        - The mean (12.38) and median (12.50) are close, showing a symmetric and compact distribution.
        - The very small interquartile range and whiskers reflect low variance (0.11) and standard deviation (0.34).
    - **Maple**:
        - The mean (12.80) and median (12.80) are identical, indicating perfect symmetry.

    - The moderate interquartile range and slightly longer whiskers indicate moderate variance (0.47) and standard deviation (0.68).
  - **Oak**:
    - The mean (15.28) is slightly higher than the median (15.20), suggesting mild right skewness.
    - The wider interquartile range and longer whiskers show higher variance (0.24) and standard deviation (0.49) compared to Linden.

Linden:

- Consistently exhibits the least variability for both LeafWidth and LeafLength, as shown by its small interquartile range, short whiskers, and low variance.

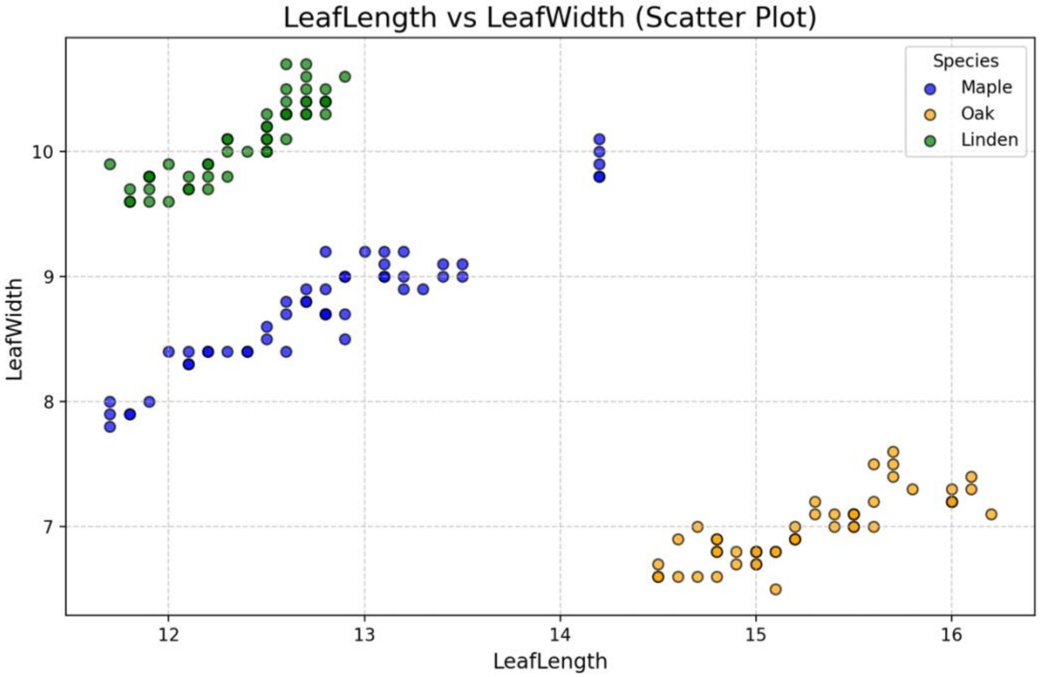
Maple:

- Displays moderate variability in both dimensions, with a symmetric distribution in LeafLength and a few outliers in LeafWidth.

Oak:

- Stands out with the largest mean and median LeafLength but the smallest LeafWidth. Its wider range in LeafLength reflects moderate variability.

4. Explain each graph in terms of variance, mean, median, and standard deviation. (Scatter Plot)



1. Linden

- **Mean and Median**:
  - The cluster for Linden is centered around a mean **LeafWidth** of 10.10 and mean **LeafLength** of 12.38.
  - The mean and median values for both dimensions are very close (median **LeafWidth** = 10.10, median **LeafLength** = 12.50), indicating a symmetric distribution.
- **Variance and Standard Deviation**:
  - The Linden cluster shows very little spread, with the lowest variance and standard deviation among all species for both dimensions:
    - Variance for **LeafWidth**: 0.10; **LeafLength**: 0.11.
    - Standard Deviation for **LeafWidth**: 0.32; **LeafLength**: 0.34.

  - This is visually represented by the compact and tightly packed points in the upper-left part of the scatter plot.

- **Linden** exhibits the least variability, and its consistent dimensions make it distinct.

2. Maple

- **Mean and Median**:
  - The Maple cluster is centered around a mean **LeafWidth** of 8.78 and a mean **LeafLength** of 12.80.
  - The mean and median are nearly identical (median **LeafWidth** = 8.80, median **LeafLength** = 12.80), suggesting a balanced and symmetric distribution.
- **Variance and Standard Deviation**:
  - The Maple cluster shows moderate spread compared to Linden:
    - Variance for **LeafWidth**: 0.29; **LeafLength**: 0.47.
    - Standard Deviation for **LeafWidth**: 0.54; **LeafLength**: 0.68.
  - This is visible in the moderate scatter of points along both dimensions, mostly clustered around the center but with some variation.
- Maple shows greater variability than Linden, but its cluster remains distinct and separate from the other species.

3. Oak

- **Mean and Median**:
  - The Oak cluster is centered around a mean **LeafWidth** of 6.98 and a mean **LeafLength** of 15.28.
  - The median values (**LeafWidth** = 7.00, **LeafLength** = 15.20) are close to the mean, indicating a symmetric distribution.
- **Variance and Standard Deviation**:
  - Oak has the smallest variance and standard deviation for **LeafWidth**:
    - Variance for **LeafWidth**: 0.07; Standard Deviation: 0.27.
  - However, for **LeafLength**, Oak has a larger variance (0.24) and standard deviation (0.49), leading to a more spread-out cluster along the x-axis.

- **Oak's** larger **LeafLength** values and smaller **LeafWidth** make it visually distinct from the other species. Its points form a clear, elongated cluster in the lower-right part of the plot.

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

- From the data and graphs, I can infer that LeafLength and LeafWidth are strong distinguishing features for identifying the three species: Linden, Maple, and Oak. The scatter plots and boxplots clearly show distinct clusters for each species, where Linden has the largest LeafWidth and moderate LeafLength, Maple exhibits moderate values for both features, and Oak stands out with the largest LeafLength and the smallest LeafWidth. The variability analysis reveals that Linden is the most consistent species with tightly packed data and low variance, Maple shows moderate variability with a balanced distribution, and Oak has higher variability in LeafLength but remains consistent in LeafWidth. Overall, the dataset reflects a moderate spread in both dimensions, with clear species-specific patterns that emphasize biological and ecological differences. These differences likely indicate adaptations to different environmental conditions, such as Linden's stability, Maple's flexibility, and Oak's adaptation for maximizing sunlight capture.