# DS 5110 - HW 2

**Data Collection:**

1. To create the dataset, I chose three types of plants commonly found around near 144 State Street Portland, Maine : Eastern Hemlock (Tsuga canadensis), Northern White Cedar (Thuja occidentalis), and Inkberry Holly (Ilex glabra). For each plant, I measured the width and length of leaves, recorded the values in a notebook, and later compiled the data into a .csv file.

2. I used a ruler marked in millimeters for precise measurements of leaf dimensions. The ruler was easy to handle and allowed for accurate readings.

3. A standard ruler is generally accurate for small-scale measurements like leaf dimensions. However, minor deviations could occur if the ruler was not perfectly aligned with the leaf surface. The millimeter scale ensured a consistent level of precision across measurements. The ruler's simple design minimized variability. While the ruler provides accurate linear measurements, slight curvature in some leaves could have led to underestimating the actual dimensions.

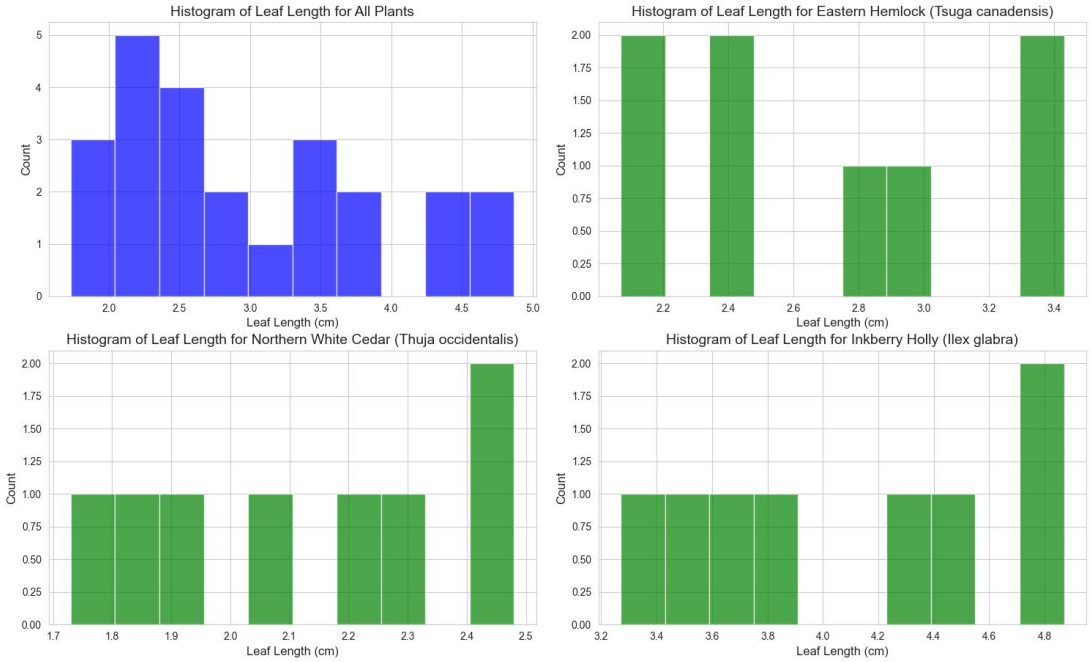
4. In all, 24 data points were collected, measuring each plant type 8 times. The number was chosen to provide enough variety in the dataset while considering the time and resources available. More than 8 measurements for each plant would be better to make the dataset robust, but it had to be balanced with the scale due to time and work limitations.

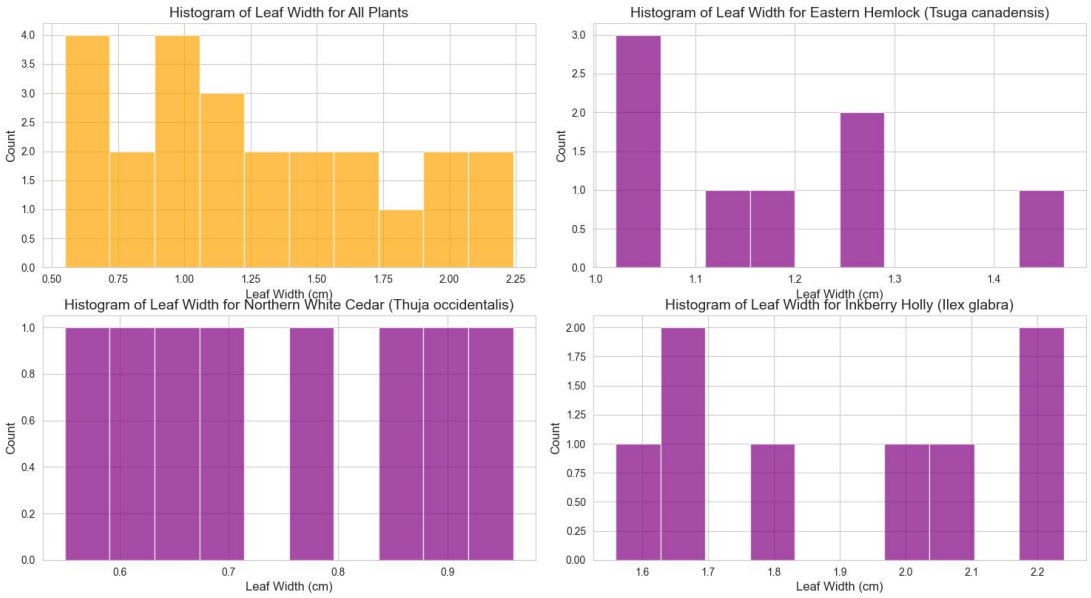
5.N (full dataset size): 24 data points 8 data per plant type  multiplied by 3 plant types, n (subset size): Each plant type had 8 data points.

6. Initially, I had difficulty identifying the plant types. Using google lense helped me resolve this issue. Some leaves were curved, which made accurate measurements challenging. Flattening the leaves carefully before measuring helped reduce this error.
windy conditions occasionally caused leaves to move during measurement, adding slight variability in readings. I mitigated this by stabilizing the leaves with one hand while measuring. Cold temperature during winter also cause problem staying outside for too long. Selecting leaves of varied sizes within the same plant type required extra effort to ensure a diverse dataset.

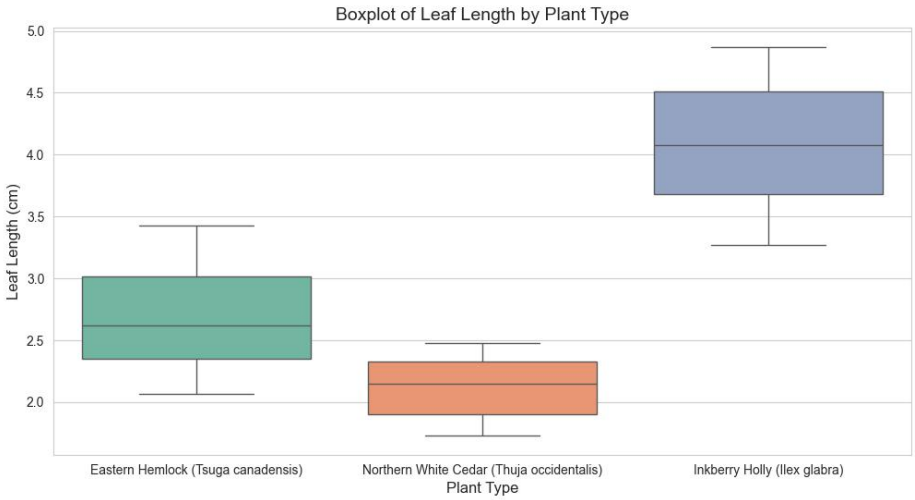
## Analysis/Visualization:

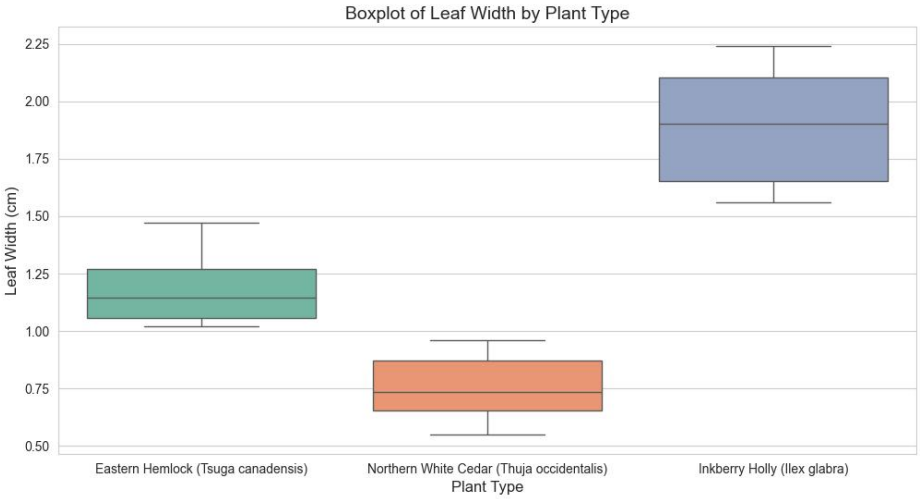
1.



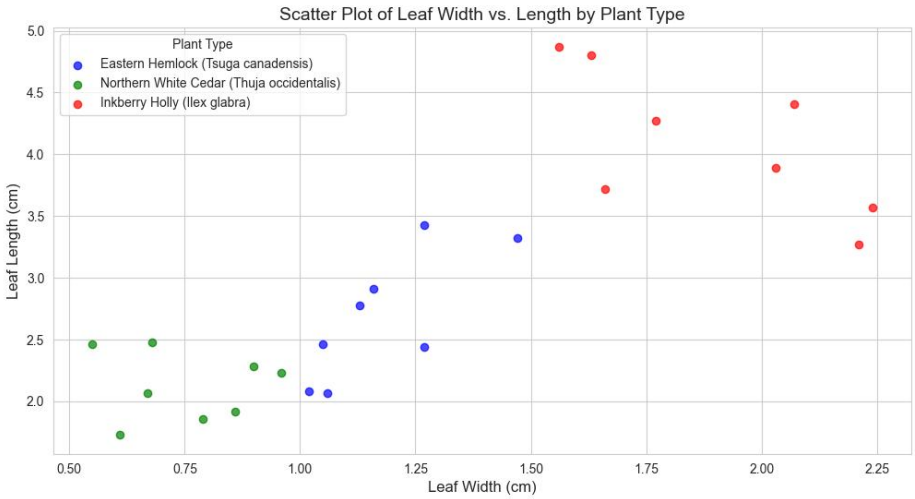



2.


Boxplot of Leaf Length by Plant Type
Leaf Length (cm)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
Eastern Hemlock (Tsuga canadensis)
Northern White Cedar (Thuja occidentalis)
Inkberry Holly (Ilex glabra)
Plant Type


Boxplot of Leaf Width by Plant Type
Leaf Width (cm)
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
Eastern Hemlock (Tsuga canadensis)
Northern White Cedar (Thuja occidentalis)
Inkberry Holly (Ilex glabra)
Plant Type

3.




4.

**Histograms for Leaf Length and Width (All Plants and Individual Plants)**

Mean: The mean is the central value of the distribution, which can be observed as the approximate center of the histogram's "bulk." The tallest bars generally cluster around the mean.

Median: The median is the midpoint of the data distribution, where half the data points are below it and half are above. In a symmetric histogram, the mean and median are close.

Variance: The variance measures the spread of the data. A wider spread (more bars with significant heights away from the center) indicates a higher variance.

Standard Deviation: The standard deviation is the square root of the variance and reflects how much the data deviates from the mean. A histogram with taller bars closer to the center has a smaller standard deviation.

Example Observations:

In the combined histograms, the spread across all plants is larger, as it includes variation across species.
In the individual plant histograms, you can observe tighter distributions (lower variance and standard deviation) specific to each species.

**Boxplots for Leaf Length and Width (Grouped by Plant Type)**

Mean: While boxplots do not explicitly show the mean, the mean often lies close to the center of the box if the data is symmetric.

Median: The line inside the box represents the median. It divides the box into two parts, with 50% of the data on each side.

Variance: The variance can be inferred from the length of the whiskers and the interquartile range (IQR). Longer whiskers or boxes suggest higher variance.
Standard Deviation: Similar to variance, the spread of the data can give an idea of the standard deviation. Outliers (dots outside the whiskers) indicate potential extremes in the data.

Example Observations:

Boxplots can reveal how leaf measurements vary across plant types. For example, Eastern Hemlock may have narrower distributions (lower variance) compared to Inkberry Holly.

The IQR (length of the box) shows the range where the central 50% of data lies. A larger box means a larger spread in the central data.

**Scatter Plot for Leaf Width vs. Length (Colored by Plant Type)**

Mean: The center of the cluster for each plant type represents the mean of the width and length for that plant.

Median: Similar to the mean, the approximate center of each cluster corresponds to the median.

Variance: The spread of points within a cluster reflects the variance. A tighter cluster indicates lower variance and standard deviation.

Standard Deviation: Similar to variance, clusters with tightly packed points have lower standard deviations.

Example Observations:

The scatter plot shows distinct clusters for each plant type, indicating that leaf width and length vary consistently within species but differ across species.

Overlapping clusters may indicate similar measurements for certain plant types.

5.

The data and graphs show clear differences in leaf measurements across the three plant types:

**Leaf Size Variability:**

Eastern Hemlock has the smallest and most consistent leaves.
Northern White Cedar shows moderate size and variation.
Inkberry Holly has the largest and most variable leaves.

**Plant Differentiation:**

Distinct patterns in histograms, boxplots, and scatter plots make it possible to differentiate plant species based on their leaf width and length.

**Key Insights:**

The variability in Inkberry Holly suggests adaptability or genetic diversity, while Eastern Hemlock is more uniform.
These patterns can support plant classification and ecological studies.