Shuiming Chen
01/30/2025

Answer the following (20 points):

1. Explain your data collection process.

- 1.1 data comes from the flowers bought from hannaford.
- 1.2 Pick up those with enough leaves and flowers and google the species name.
- 1.3 using a scale with centimeters to measure the length and width and write it down.

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

- Scale bought from home depot with centimeter measurement.

3. Argue the accuracy and precision of your instrument.

- Very high accuracy and precision because it's a standard product.

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

- 34 lists in total, and each species has a different number, because the flowers were mixed by different species, and the leaves were limited.

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

- subset size are:
- Species(a) Alstroemeriaceae: 11 leaves
- Species(b) Orange-lily: 8 leaves
- Species(c) Cinnamomum-osmophloeum: 15 leaves
- full data set size is: N here is n(a) + n(b) + n(c) = 34

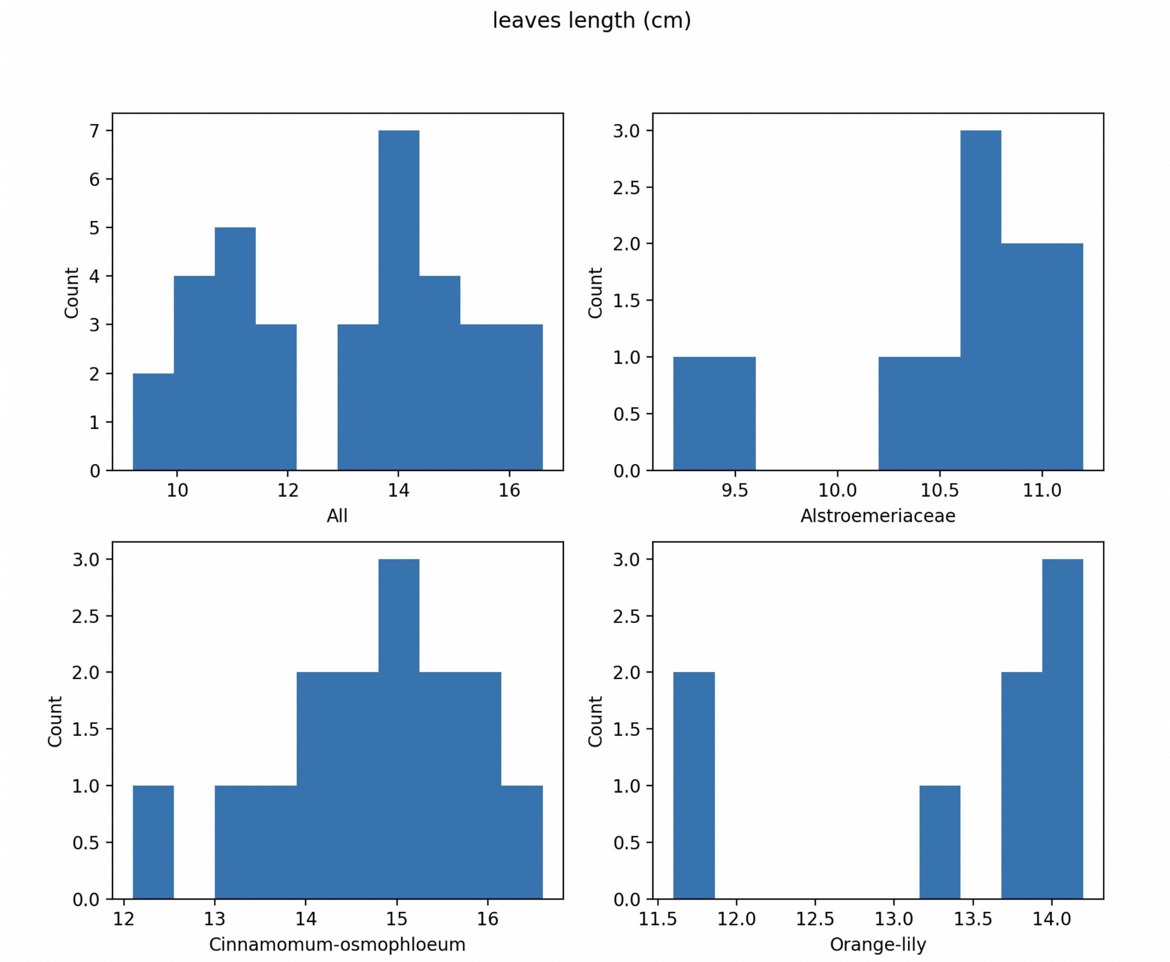
6. Explain any problems that you ran into during the data collection process.

- 6.1 manual measurement can lead to some tiny inaccuracies, maybe up to 0.1 to 0.2 centimeter inaccuracies.
- 6.2 data sets are not big enough to get the pattern of leaves size range.

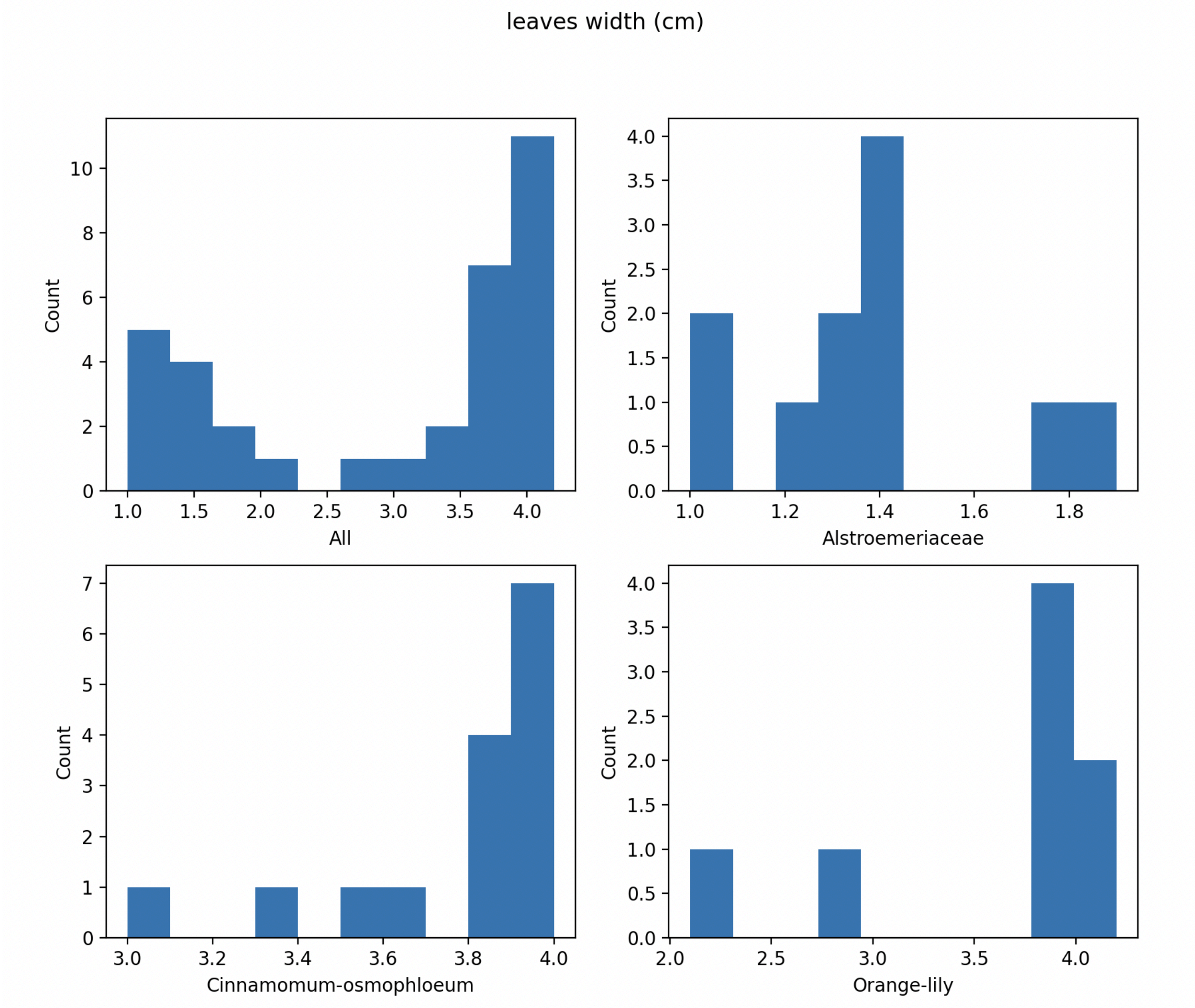
Analysis/Visualization - (50 points):

1. Graph histograms of your data with appropriate labels.

- Leaves length histogram visualization

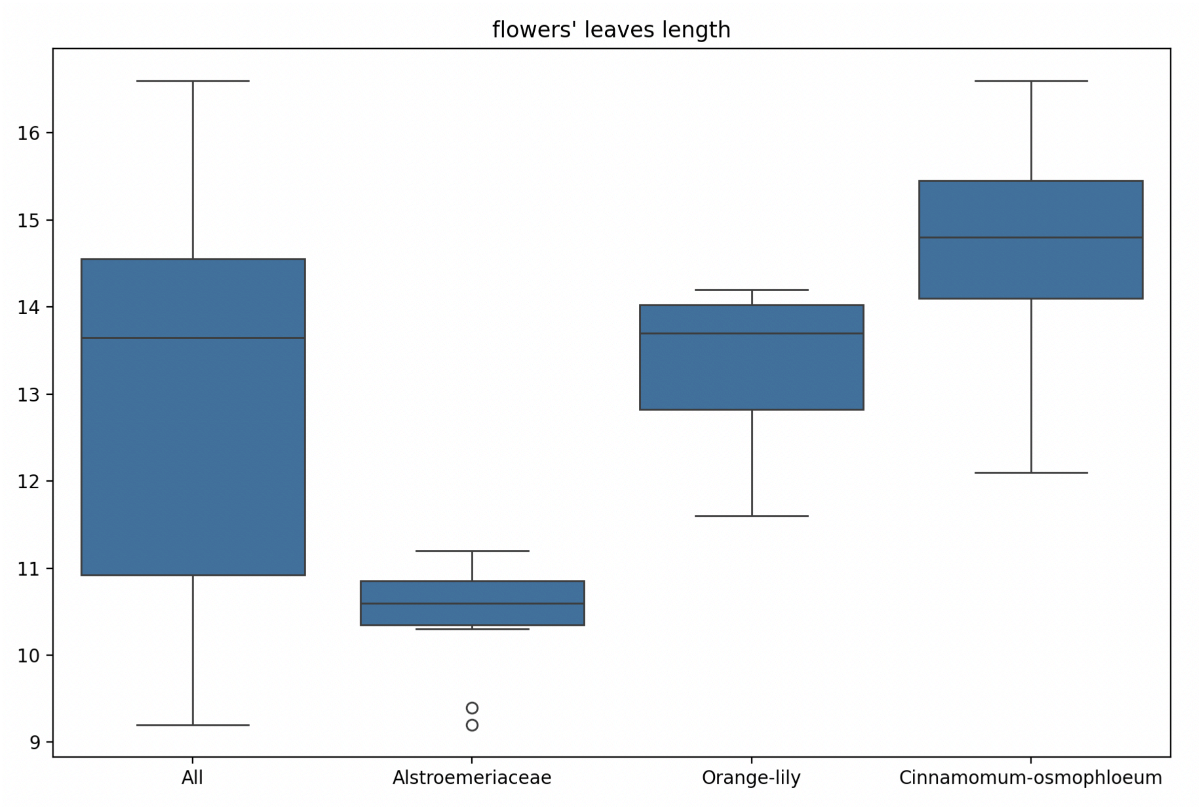


- Leaves width histogram visualization

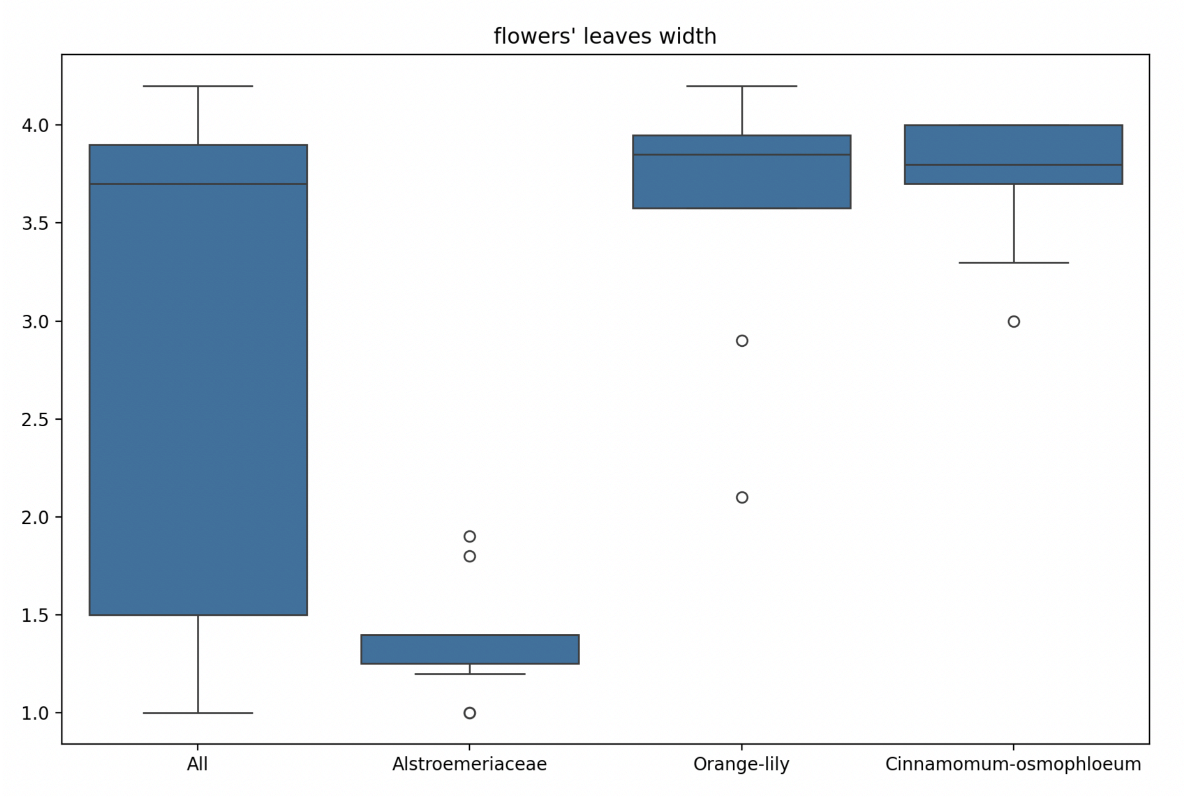


2. Graph boxplots of your data with appropriate labels.

- Leaves length boxplot visualization

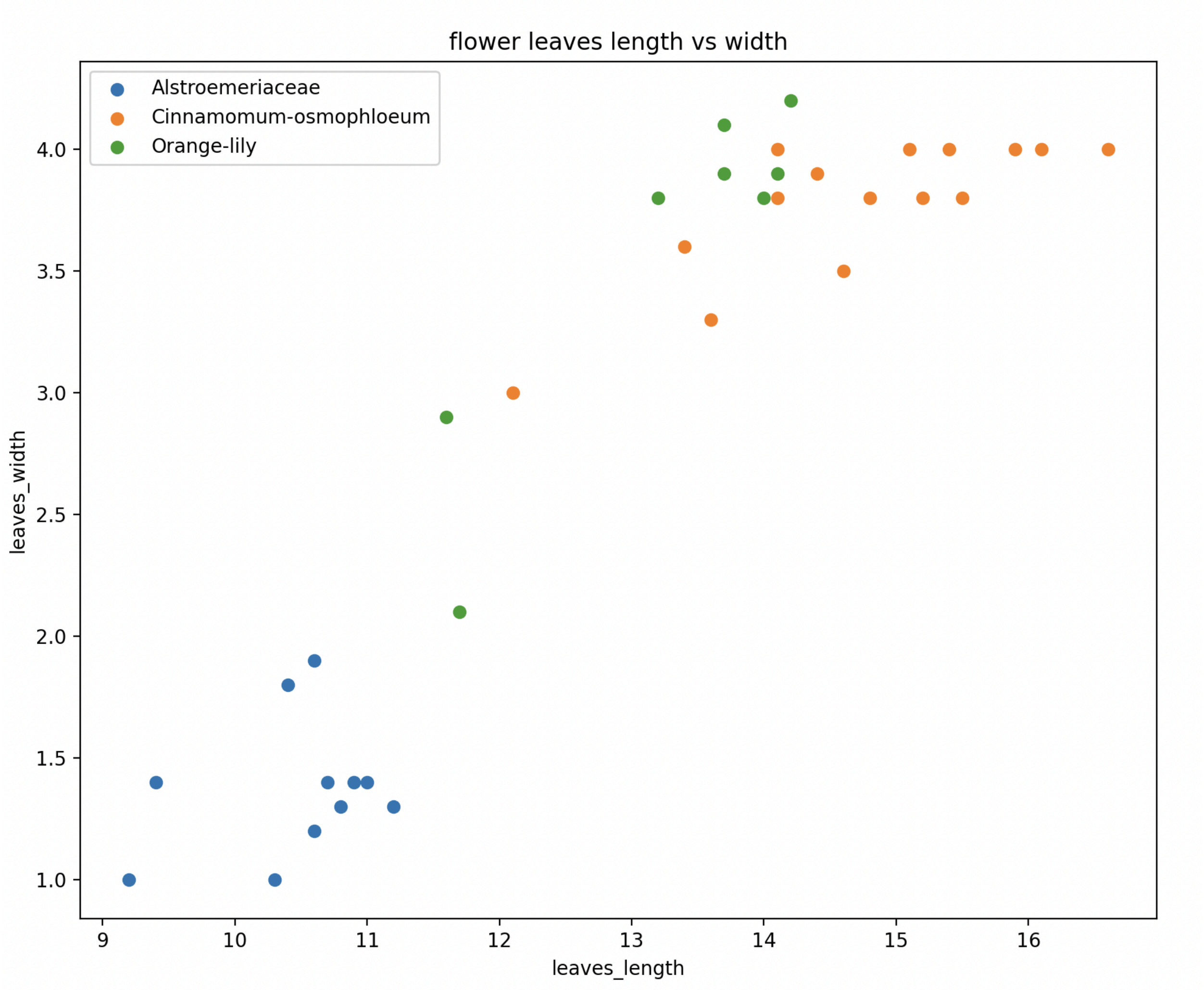


- Leaves width boxplot visualization



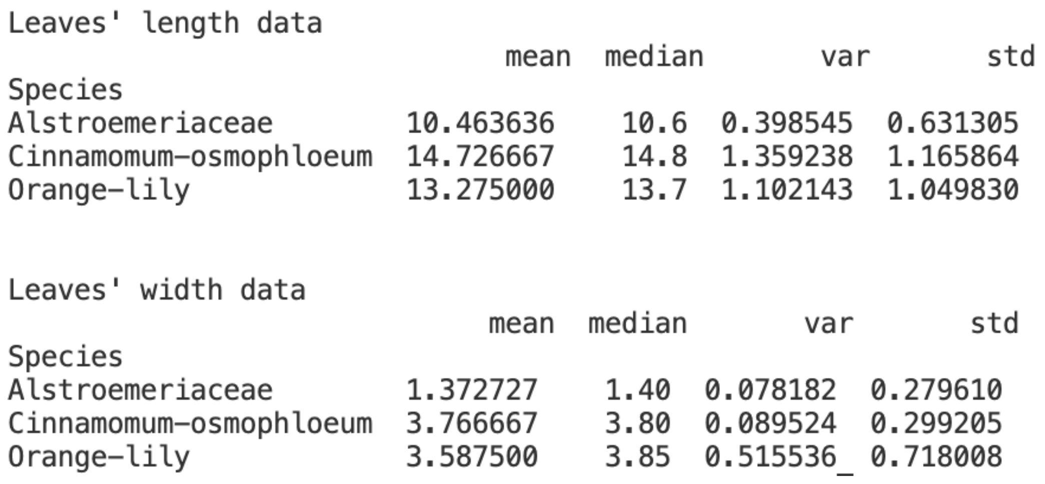
3. Graph a scatter plot of your entire data set with each subset different color and a ledger.

- scatter visualization



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

Leaves' length and width statistical data are calculated and shown below: see cal_statistics.py file



```
Leaves' length data
                            mean  median       var       std
Species
Alstroemeriaceae       10.463636    10.6  0.398545  0.631305
Cinnamomum-osmophloeum 14.726667    14.8  1.359238  1.165864
Orange-lily            13.275000    13.7  1.102143  1.049830


Leaves' width data
                           mean  median       var       std
Species
Alstroemeriaceae       1.372727    1.40  0.078182  0.279610
Cinnamomum-osmophloeum 3.766667    3.80  0.089524  0.299205
Orange-lily            3.587500    3.85  0.515536_ 0.718008
```

histograms

- variance: x label of the figure width which is the distributions indicate variance in leaf length, the narrower of the width, the smaller of the leaves variance, and vice versa.
- mean/median: The peak of the bin can help to estimate the mean/median for each species. If the dataset are normalized, the mean and median would be in the same spot, otherwise their results would be close to each other.
- standard deviation: the narrower of the bin means lower standard deviation, and vice versa.

boxplots

- Boxpots do not directly show the variance, standard deviation and mean. However, we can infer from the boxplot of median, which is the line inside the box.
- Boxplots has first quartile(Q1) which means the median of the upper half, and third quartile(Q3) which is the median of lower half.
- We can also infer from the boxplots of minimum and maximum values which are called whiskers, and outliers which are data points outside of whiskers.
- We can say that a wider box and longer whiskers can be used to imply higher standard deviation.

scatters

- variance: how spread out of the points are can be used to detect variance.
- mean/median: They can be inferred by looking for the central tendency of the points.
- standard deviation: The more spread out of the points from the central point (mean), the higher the standard deviation.

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

General Trends:

- Cinnamomum-osmophloeum: the width of the leaves are relatively close to 4cm, but the length is different.
- Orange-lily: except those tender leaves, all other ripe leaves are having the similar length and width.
- Alstroemeriaceae: the length is pretty stable but the width is more changeable.

Correlation Analysis:

- We can tell from the scatter figure that in general, the longer of the leaves, the wider the width.
- But not all longer length of leaves along with the wider width, it comes out species by species.

Species Comparison:

- Alstroemeriaceae has a relatively long length but pretty small size of the width of the leaves compared with the other two species.
- The distinct size patterns showed after grouping the data by species, and we can also see the species' leaf dimension.

Conclusion:

The leaf length and width reveals a general correlation but varies by specific species. If the data set is large enough, maybe we can use the pattern to study the plant's growing environment. Like indoor growing pattern and outdoor, or cold area vs warm area, etc.