DS5110

Homework 1 - Data Collection

Due 1/24/2025

Introduction

There is only so much I can talk about the data collection process in class. This homework is designed so that you can get hands on experience collecting, processing, building an understanding, and visualizing data.

Problems

Data Collection (30 points)

You are tasked with collecting and creating your own data set like the Iris data set that we use in class. You are tasked with collecting data from 3 different types of plants around campus (or close to where you are currently located). Data must be saved in a .csv format and consist of leaf-width, leaf-length, and name of the Plant (You need to figure that out).

Answer the following (20 points):

1. Explain your data collection process.
2. What instrument did you use to collect data with?
3. Argue the accuracy and precision of your instrument.
4. How many data points did you collect? Why?
5. Define the size of your data in terms of both N (full data set size) and n (each subset size).
6. Explain any problems that you ran into during the data collection process.

Analysis/Visualization - (50 points)

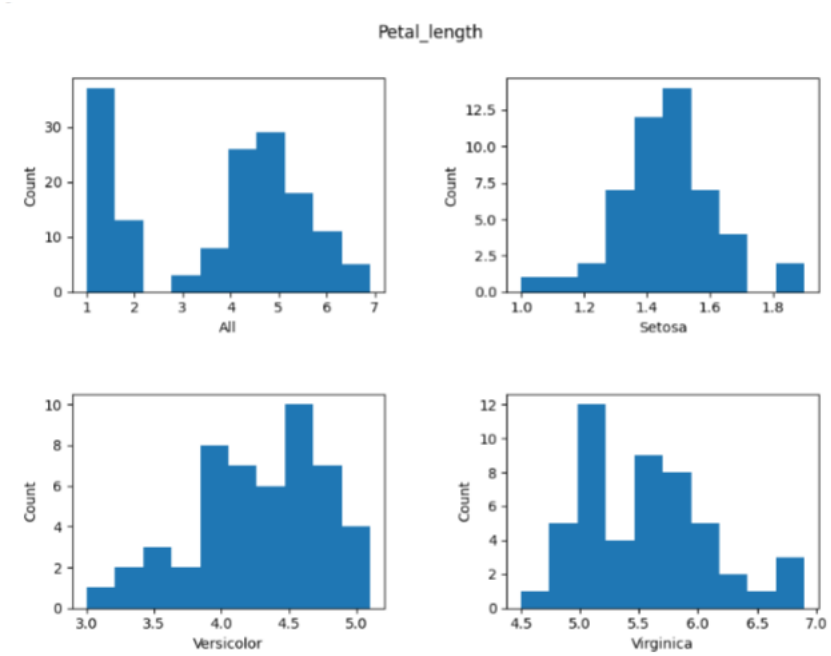
Now that you have collected the data you will now need to analyze and visualize the data.

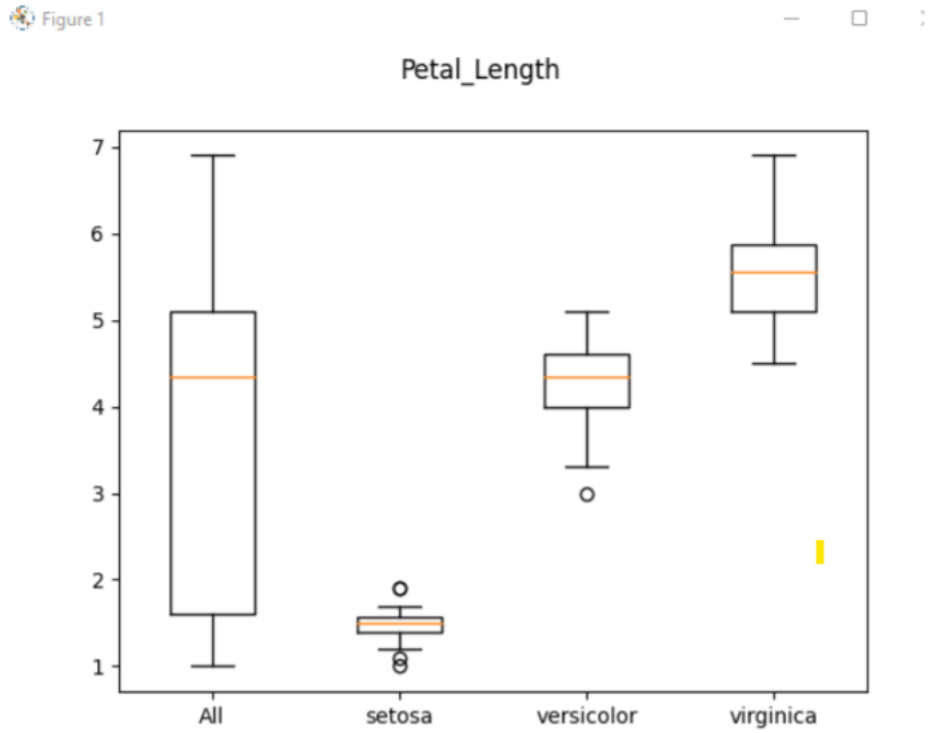
Complete the following:

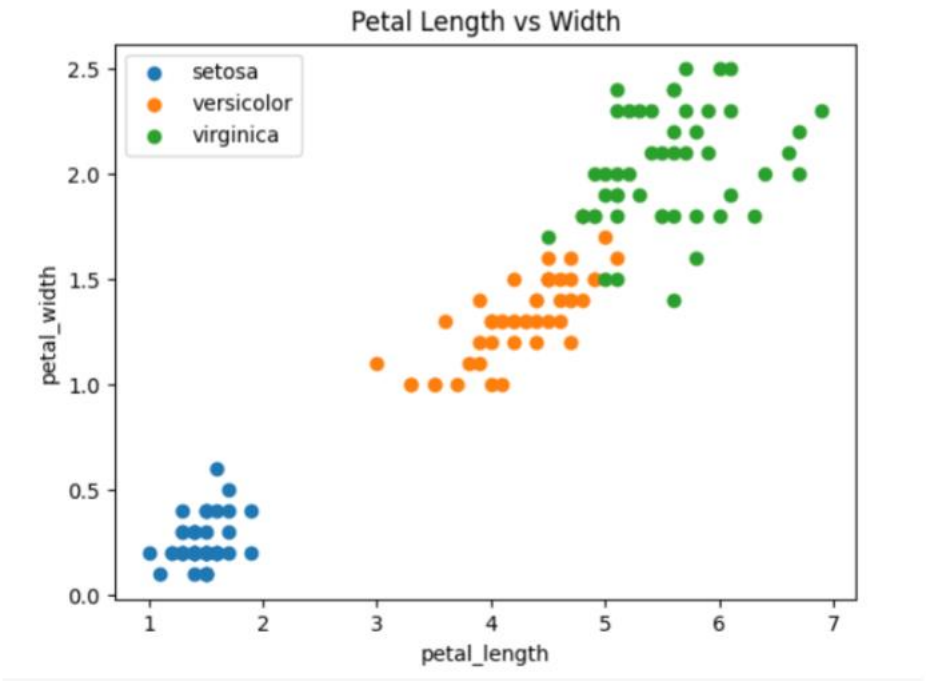
1. Graph histograms of your data with appropriate labels.
2. Graph boxplots of your data with appropriate labels.
3. Graph a scatter plot of your entire data set with each subset different color and a ledger.
4. Explain each graph in terms of variance, mean, median, and standard deviation.
5. What can you infer with data and graphs that you have?

Sample outputs

Here are how some of your graphs should look based on the Iris data set.







How to turn in

Answers to each of the written questions should be written either in your Readme.md or a separate .pdf file. The code that you wrote should be in a python (.py) file, and your data should be in a .csv file that is inside a data folder. Like in HW1.

Push all files to HW2 on GitHub.