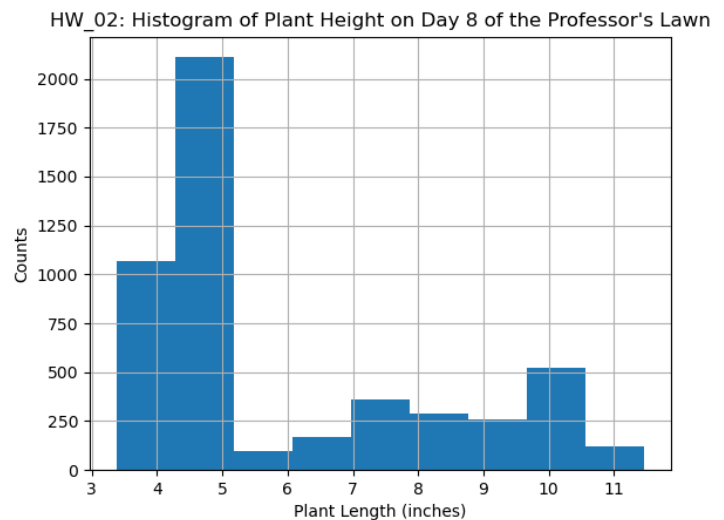
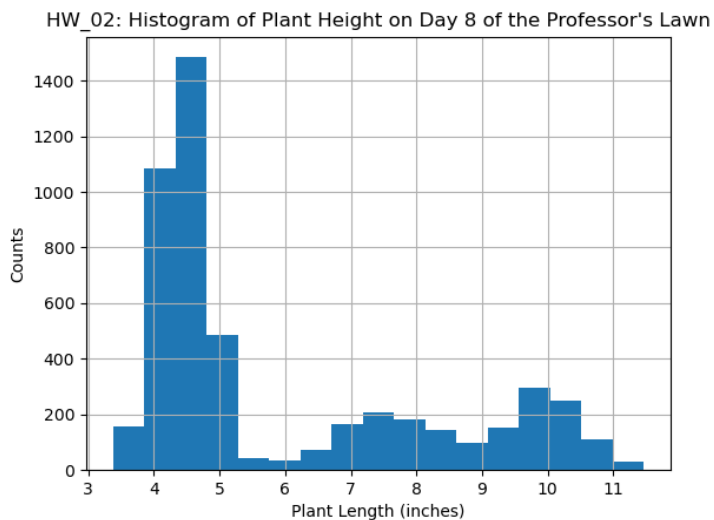
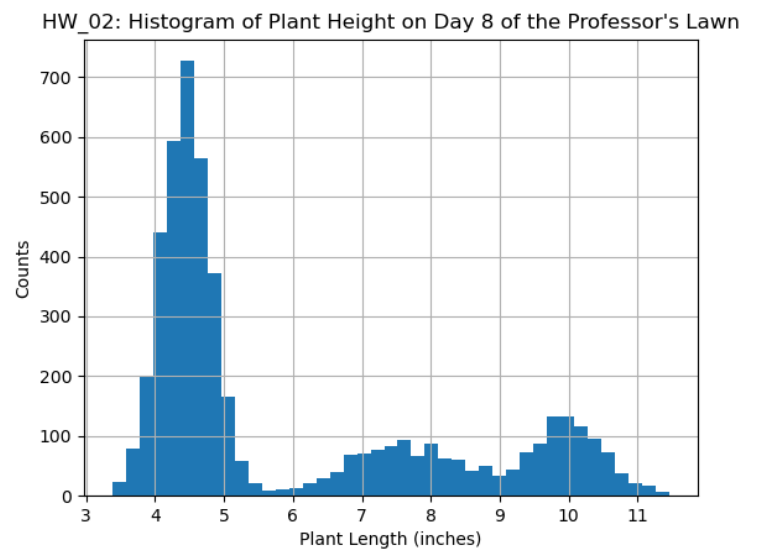
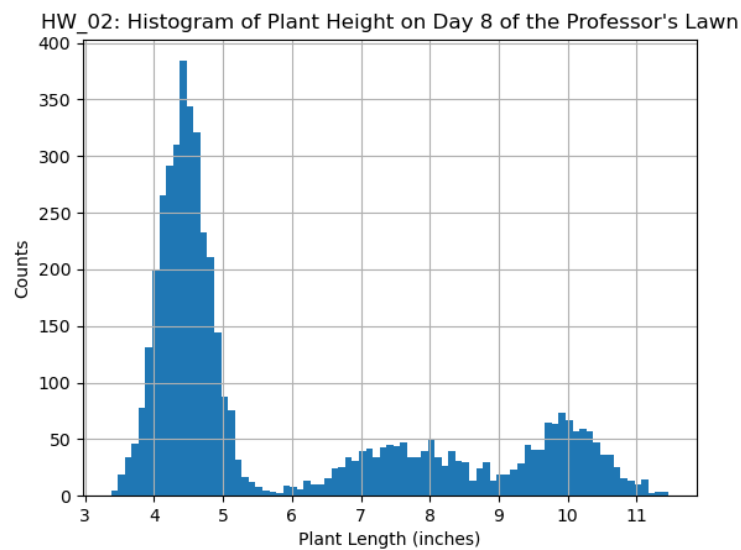
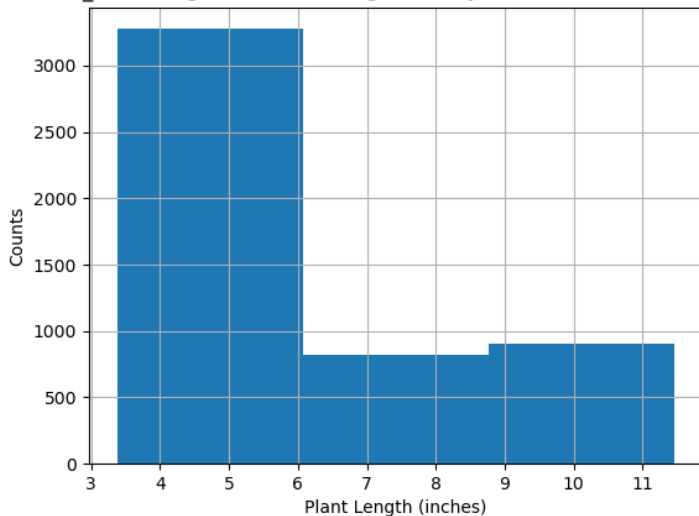


Through this homework we have been tasked with reading in csv file data (The length of 5000 plants measured over 8 days), and creating a histogram using one of the columns 'D8' (Day 8), which was done using primarily the pandas library, and the matplotlib library. The format of said histogram must be taken through arguments parsed from the command line, which was done using the argparse library.

The goal was to both learn how to better use the pandas for data analysis and argparse library as a formatted and simple way to parse data from the command line, and to better understand the importance of “how” you decide to represent the same data, and how that can affect what the data seems to show. In this case, using a quant argument parsed from the command line, we were expected to experiment with “playing around with it”, and seem how that affects the final histogram being shown. 5 examples of the histograms created (using a quant value of 0.1, 0.2, 0.5 and 1)



HW_02: Histogram of Plant Height on Day 8 of the Professor's Lawn



As can be seen in the smaller examples, we can begin to see that there seem to be 3 modes of heights for this data, which might suggest 3 different types of plants found here, with differing growth rates. However as we continue to increase the size of the bins, that fact becomes less and less obvious.

Through simply changing the size of each bin, we can begin to “hide” information that the data provides, all the while never representing false data. This is a perfect example of how one can lie with statistics. And this is with all else being constant. One of the biggest factors when it comes to generating histograms is where the first bin starts (which

in this case was set to always start at the smallest possible value), as it can completely change where the data seems to peak, and is one of the biggest ways that one can create misleading data representation and analysis, all the while using what is completely real data. This demonstrates the importance of the “how” you represent data. Being able to find the latent data within our collected data requires us to be able to understand this concept, and this homework demonstrates it quite succinctly.