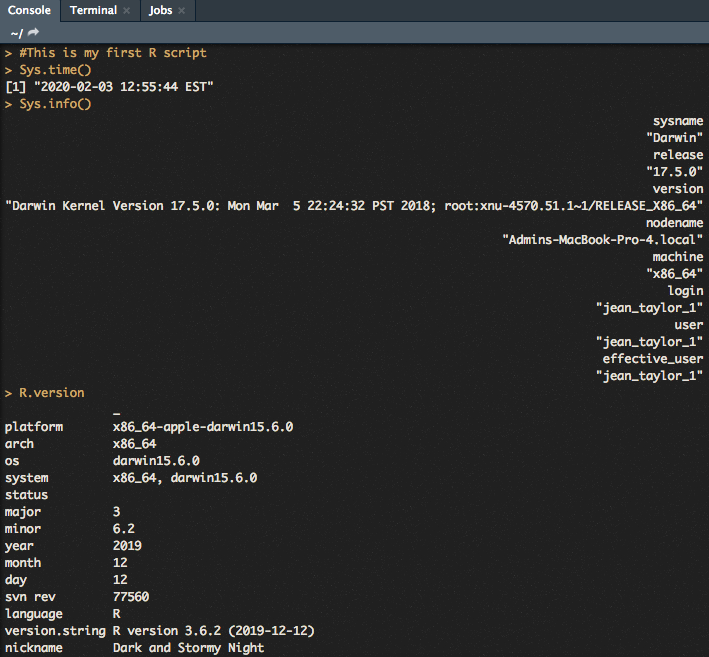
James Taylor

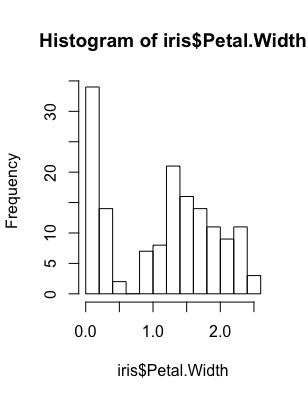
DATA 630

Exercise 1

**Part 2a.**







**Part 2b.**

These commands reveal the iris data has 5 variables. They are Sepal.Length, Sepal.Width, Petal.Length, Petal.Width and Species. All of them are numerical values except for the species. The maximum value of the Petal.Width variable in the dataset is 2.5. The summary statistics for Petal.Width show the distribution of the variable, which again has a max of 2.5. The histogram of Petal.Width has the distribution in more detail. The histogram output is above in Part 2a.

**Part 3.**

The histogram in line 10 reveals the distribution of the values of the Petal.Width variable. It shows many data points close to zero or at zero outliers (although the summary stats show no zero values). The distribution has another peak below 1.5. The distribution doesn’t show obvious. It doesn’t appear to be normally distributed. The max value is 2.5.

Visualization capabilities are important because it displays lots of detail quickly. Relationships can be uncovered fast. Those relationships could also go undetected if it weren’t for visualizing and only relying on summary statistics.

**Part 4.**

I wouldn’t say an open source tool is always better than a commercial one, but it does have great benefits. Firstly, they are free, so rich or poor can investigate data. Secondly, because R is free and popular, there are many people and resources on the internet for help and inspiration. There are tutorials everywhere and already-written scripts to help. The extensive modules that exist can dramatically increase someone’s speed and ability to complete a task. Commercial tools often have experts that can help, but the open-source tools have the ability to crowd-source.