Data Analysis using R

Veeral Patel

Regis University

March 26th, 2021

1. **Introduction**

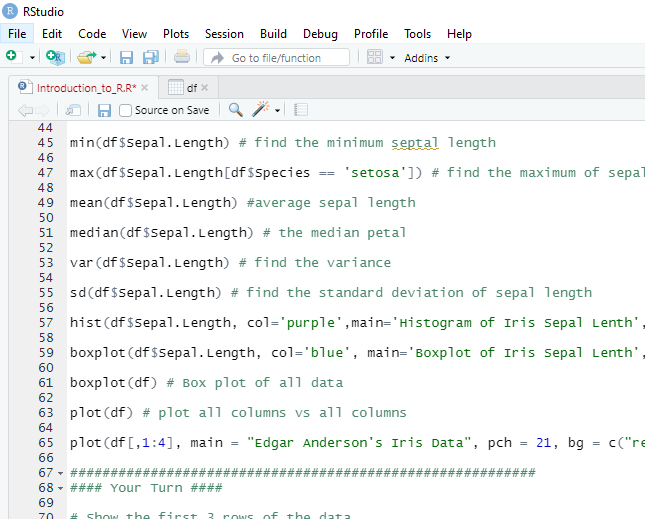
For this project, we were expected to download R and R Studio in order to be able to open data files, and use R for data analysis. We were using a dataset that was included in one of the R packages. This dataset included species of plants and their petal measurements along with other measurements of different species of plants. We were to use R in order to perform mathematical operations on that dataset which helps us gather more insight on the data.

1. **Methods/Code/Screenshots**

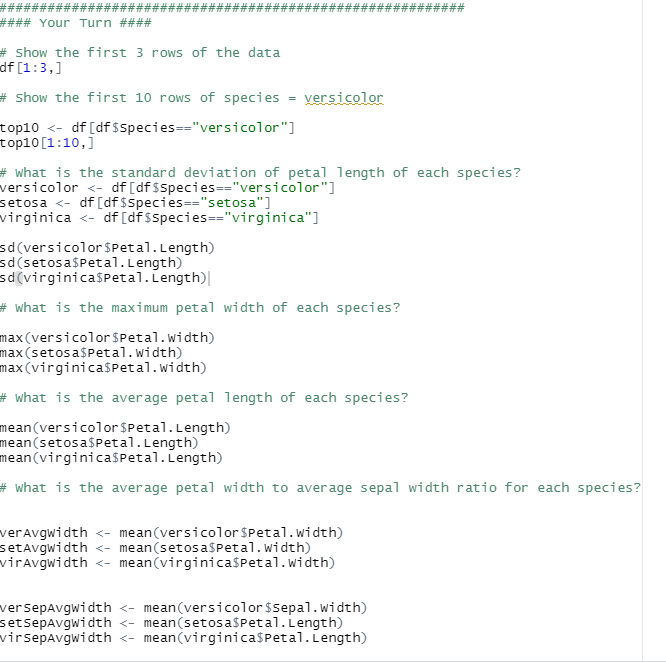
I used a windows computer and I had to install R, R Studio (IDE) and all of the dependency modules that were needed to perform the data analysis. I have worked with R before but it was very brief and it was in a science class. I had only used R before to calculate averages of large sets of data that was in an .rda format which is like a really compressed file. This was a new adventure for me to work with R Studio with more data analysis. I watched a few quick tutorials on YouTube in order to assist me with getting everything installed on my computer. Most of the code in this assignment was supposed by the professor. I was in charge of coming up with a few calculations that included standard deviation, mean, box plot, histogram, and ratios. I also found it generally helpful to actually understand the data by just opening it up in R studio in a data frame before getting started with the assignment. I also used the head() function in order to get a quick glimpse of the data within R Studio.

1. **Results/Analysis**

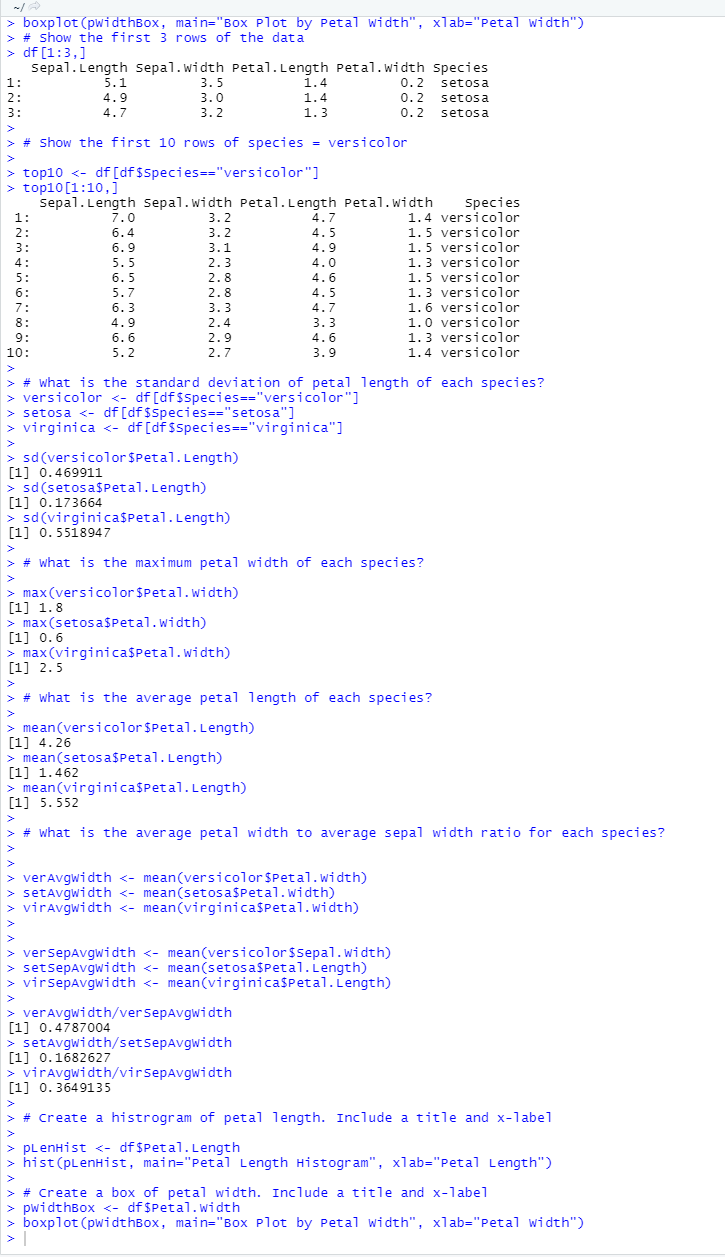
Installation of R Studio:



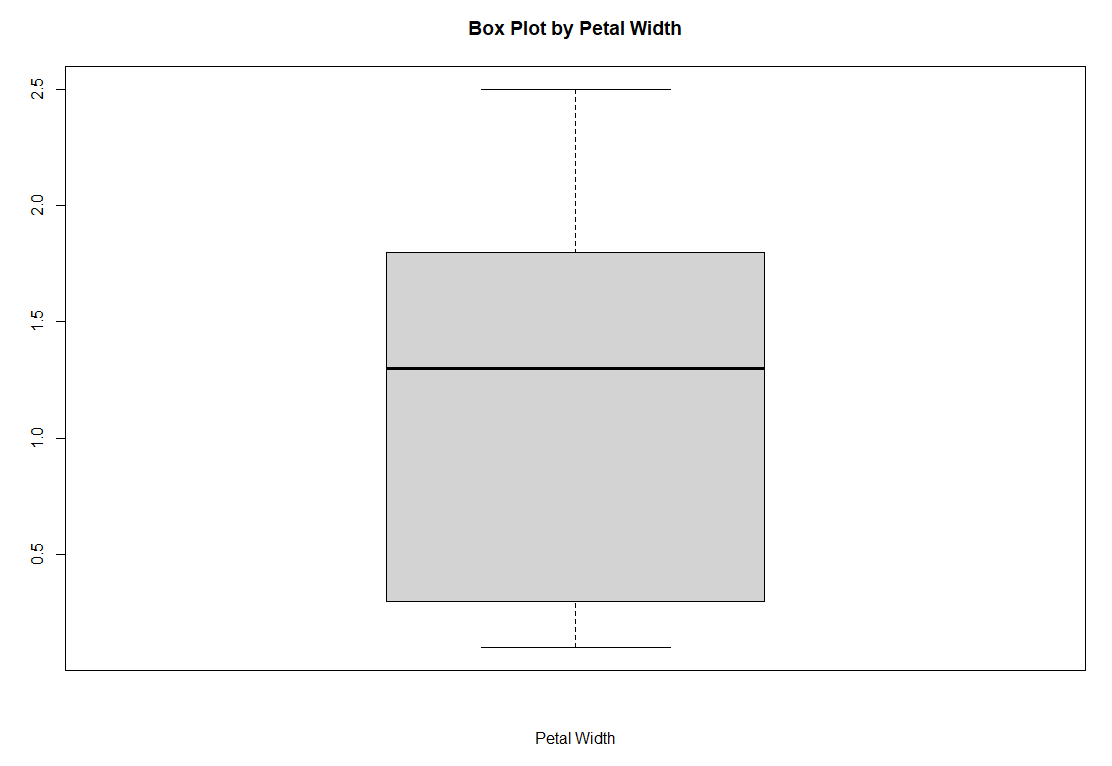
Code for the homework:



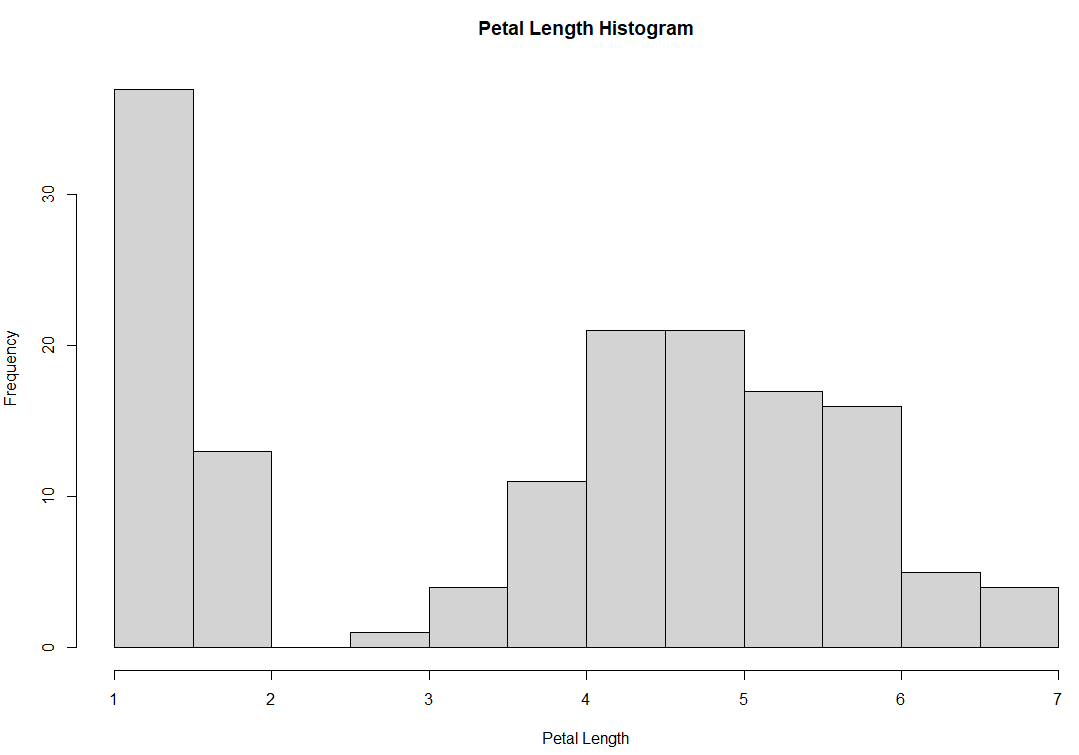
Homework Output:



Boxplot:



Histogram:



1. **Analysis of Results**

In order to analyze the results, we must understand what the dataset really is. Here, we have a dataset that contains of different petal sizes for different species of plants/flowers. There are 3 species that this dataset contains: versicolor, setosa, and virginica.

The standard deviation shows us how far apart the data is from the mean. We can see from the standard deviation that the setosa species has the closes petal length to the mean. We can also see that the largest petal width comes from the virginica species. Also, on average, the virginica has the largest average petal length as well. The viginica has the largest petal width to sepal width ratio out there. I thought that looking at the histogram and box plot was telling.

Looking at the histogram, it helps us to see where the data lies. For example, in the petal length histogram, we can see that the majority of the flowers, have a petal length of 1, this is data that we cannot see from taking the mean of the petal length. The boxplot shows us where the quartiles are and where the range of the petal width is. We can see that the mean is around 1.4, and that most of the range is between .4 to 1.9.

1. **Conclusion**

I thought that this was a really good example of being able to use simple arithmetic with a dataset in R to learn the basics of data science in R. I have actually worked with this dataset in python before and I thought that it was a really cool dataset and it was just simple enough to start learning the basics. The one thing that I did not like about this dataset was just my unfamiliarness with species of plants so the dataset itself didn’t mean that much to me since it was a subject that I was pretty unfamiliar with. This would be useful in the math and science fields since it is really useful to calculate this much in a few lines of code. I really think that for a lot of data science things, R seems to be less lines of code than python would have been. This was my first experience with R, the syntax looked a little bit weird to me, but after I got used to it, it was pretty simple.

1. **Reference:**

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O’Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 1-7.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".