Here I'm going to give examples of good and bad homework assignments. I will ask very similar questions for the iris data as for the cars data to illustrate.

The idea for the homework is to write a report that answers the questions. The report should use complete sentences and be written to be understood by someone who does not know R. This is not like a math homework where you circle your answer and don't use complete sentences. The goal is to communicate the results as much as to find the results.

Imagine that this is an assignment at work and your boss has asked you to write a report answering these questions. Your boss thinks of you as a data analyst but your boss does not know R. You can assume that your boss is familiar with the statistical concepts used in the assignment, such as standard deviations, but has never used R.

It is still a good idea to put some of your code in an appendix. One reason for this are that if you have an incorrect answer, the grader can look at your code and see how you got the answer. Sometimes a minor typo in the code leads to a big error in the results, so having the code in the appendix can lead to better partial credit.

Also, **you should each turn in separate assignments where each individual students run the code themselves**. As discussed in class, R code looks different depending on your personal style, including things naming of variables, use of white space, order of commands and so on. **If your homework looks identical to someone else (especially in the errors), the code in the appendix can help clarify that you didn't reuse someone else's code. Also, since English writing is unique to each person, your homework should not have identical sentences to other students in class**.

The next pages give example homework solutions to a slightly different homework set.

Example homework

Use the iris data, which is a built in data set in R, to answer the following questions.

- “iris" is a built in dataset in R with 50 observations on two variables:
- “Petal.Length" gives the length of individual petals
- “Petal.dist" gives the width of individual petals
- “Species” gives the names of the three species
- All measurements are in cm

Please refer to columns in the data using the “$’’ sign, if we use , d1<-cars, then the species column is d1$Species, and the petal length column is d1$Petal.Length.

a. Plot petal length against petal width. Do you see a pattern?

b. Compute the mean, median, standard deviation, and interquartile range for the petal length data.

c. Make a stem-and-leaf display, histogram, and boxplot for the petal length data. Is there much difference between the mean and median? Discuss, briefly, whether the size and the direction of the difference is sensible, given the graphical summaries.

d. Using the graphical summaries, describe the shape of the distribution. Discuss modality, presence/absence of outliers, whether skewness is present, and if so, in what direction, and whether it would be reasonable to assume that the distribution is normal?

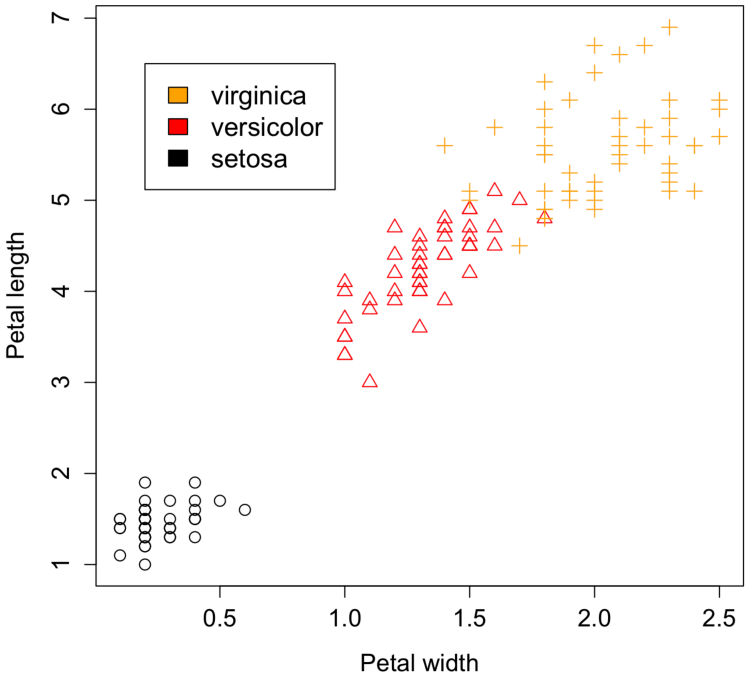
e. Repeat d restricting the analysis to the species versicolor.

Solutions are on the next page.

Good Solution

The data consists of 150 observations of iris flowers, with variables for sepal width, sepal length, petal width, and petal length. There are 50 observations from each of three different species: *setosa*, *versicolor*, and *virginica*.

a. Below is a plot of petal length against petal width. Species are coded by shape and color. In the plot, there is a linear relationship between petal length and width, and increasing width is associated with increasing length. The data points also appear to form clusters based on species, with petal lengths and widths *setosa* being especially well separated from *virginica* and *versicolor.*

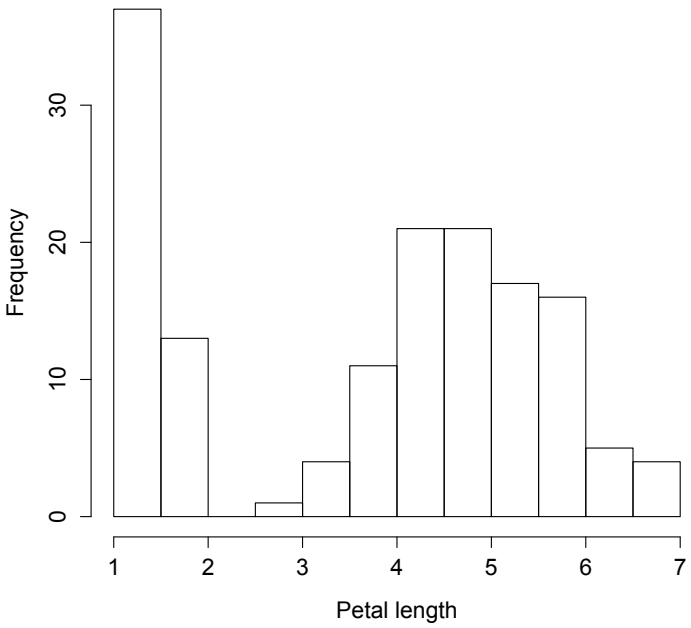


b. The petal length has a mean of 3.8cm and median of 4.35cm. The standard deviation and interquartile range are 1.8cm and 3.5 cm, respectively.

c. The stem-and-leaf plot is on the next page.

```
The decimal point is at the |

 1 | 01223333333444444444444
 1 | 55555555555556666666777799
 2 |
 2 |
 3 | 033
 3 | 55678999
 4 | 000001112222334444
 4 | 55555555666777778888999999
 5 | 000011111111223344
 5 | 55566666677788899
 6 | 0011134
 6 | 6779
```

A histogram for the petal lengths is given here.



A boxplot for the data is the following.

From part b, the mean is 3.8cm, and the median is 4.35cm. The median appears to be larger than the mean, which is more typical of left-skewed distributions. This is consistent with the boxplot, in which the median is offcenter towards larger values. However, examination of the histogram shows that the distribution appears to be bimodal, with a clear separation of lower values (less than 2cm) versus higher values (greater than 2.5cm). The shape of the histogram for higher values of petal length is roughly symmetrical, but there are many values below 2cm which might make the median and mean behave more like a left-skewed distribution.

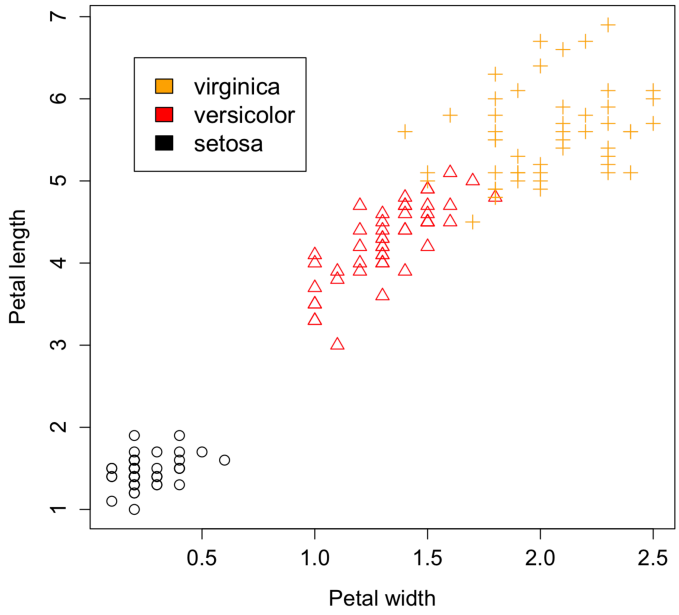
d. As discussed in part c, the distribution of petal lengths is bimodal. Based on the boxplot, as well as the histogram, there do not appear to be any outliers for petal length. The distribution does not appear to be symmetrical, and if anything is left-skewed. Based on these characteristics, the petal lengths do not appear to be normally distributed.

e. If we restrict the analysis to the species *versicolor*, then the data looks more symmetrical and closer to a normal distribution. The histogram looks like it might be slightly left-skewed. Consistent with this observation, the median is slightly larger than the mean (4.35cm versus 4.26cm). There still do not appear to be any outliers. This can be seen from a histogram of the *versicolor* petal lengths.

Appendix for good solution.

```
> mycol <- c(rep(``black'',50),rep(``red'',50),
rep(``orange'',50))

> myplotchar <- c(rep(1,50),rep(2,50),rep(3,50))

> plot(iris$Petal.Width,iris$Petal.Length,xlab="Petal width",
ylab="Petal
length",cex=1.3,cex.lab=1.3,cex.axis=1.3,pch=myplotchar,col=mycol
)
> mean(iris$Petal.Length)
[1] 3.758
> median(iris$Petal.Length)
[1] 4.35
> sd(iris$Petal.Length)
[1] 1.765298
> fivenum(iris$Petal.Length)[4]-fivenum(iris$Petal.Length)[2]
[1] 3.5
> boxplot(iris$Petal.Length)

> stem(iris$.Petal)

>
hist(iris$Petal.Length[iris$Species=="versicolor"],breaks=10,cex=
1.3,cex.lab=1.3,cex.axis=1.3,main="",xlab="Petal lengths for
versicolor")

> boxplot(iris$Petal.Length,xlab="Petal
length",cex=1.3,cex.axis=1.3,cex.lab=1.3,main="")
```

Bad Solution



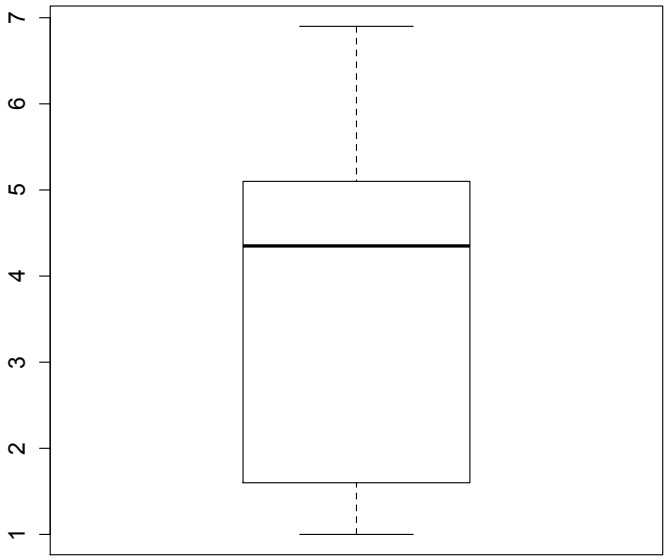
a.

b. The mean, median, standard deviation, and interquartile range can be seen in the R output below:

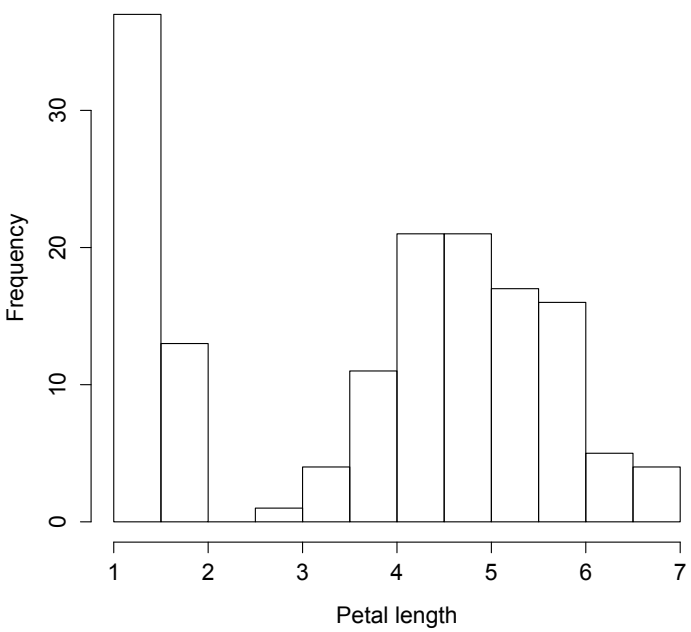
```
> mean(iris$Petal.Length)
[1] 3.758
> median(iris$Petal.Length)
[1] 4.35
> sd(iris$Petal.Length)
[1] 1.765298
> fivenum(iris$Petal.Length)[4]-fivenum(iris$Petal.Length)[2]
[1] 3.5
```

```
c.
stem(iris$Petal.length)

The decimal point is at the |

  1 | 012233333334444444444444
  1 | 55555555555556666666777799
  2 |
  2 |
  3 | 033
  3 | 55678999
  4 | 000001112222334444
```

```
4 | 555555556667777888899999
5 | 000011111111223344
5 | 55566666677788899
6 | 0011134
6 | 6779
```



Petal length



d. Based on plots, not normal.

e. For the new analysis, it looks better. More normal.