STAT 260 - R Assignment 1

Parker DeBruyne - V00837207

18/01/2023

Part 1

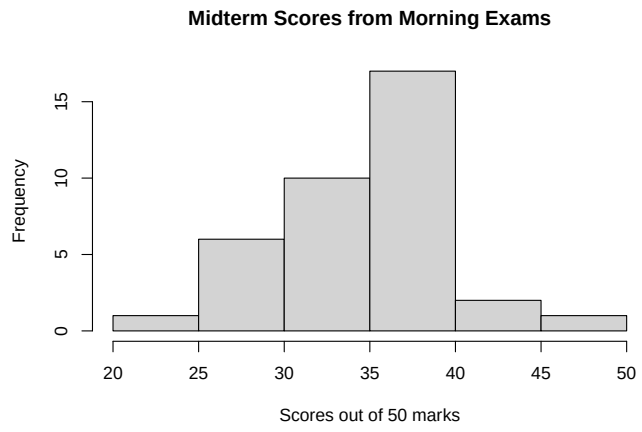
We wish to compare the midterm scores of two sections of a particular stats class. Both sections wrote exams out of 50 possible points. Collections of students from each of the two sections were randomly sampled and their exam scores are summarized below.

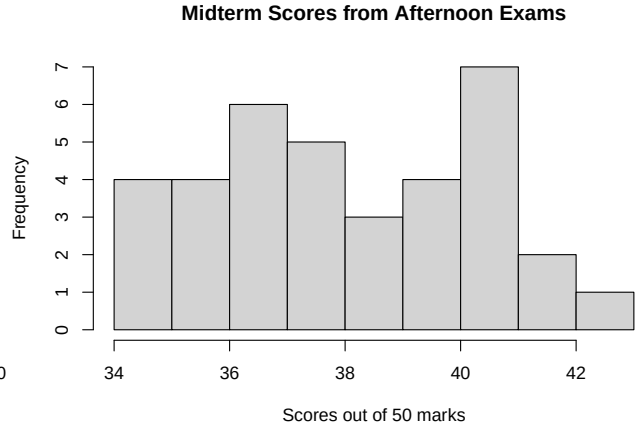
Morning Section Marks:
37 39 27 33 29 32 39 40 40 50 39 40 33 39 38 29 24 31 27 36 30 36 40 39 30 41 41 34 32 40 31 32 38 39 33 32 39

Afternoon Section Marks: 38 36 40 37 42 38 37 41 43 39 40 36 37 34 41 36 39 37 40 38 35 34 38 42 39 41 40 41 37 41 37 41 35 38 41 36

(a) [2 marks] Create a histogram of the marks from the morning section of the class. The title and x-axis for the histogram should have appropriate labels. Copy and paste the relevant command from the R Console Window and the resulting histogram into your Word document. (You do not need to include the code for how you stored the data into R, just include the one line of code for the histogram.)

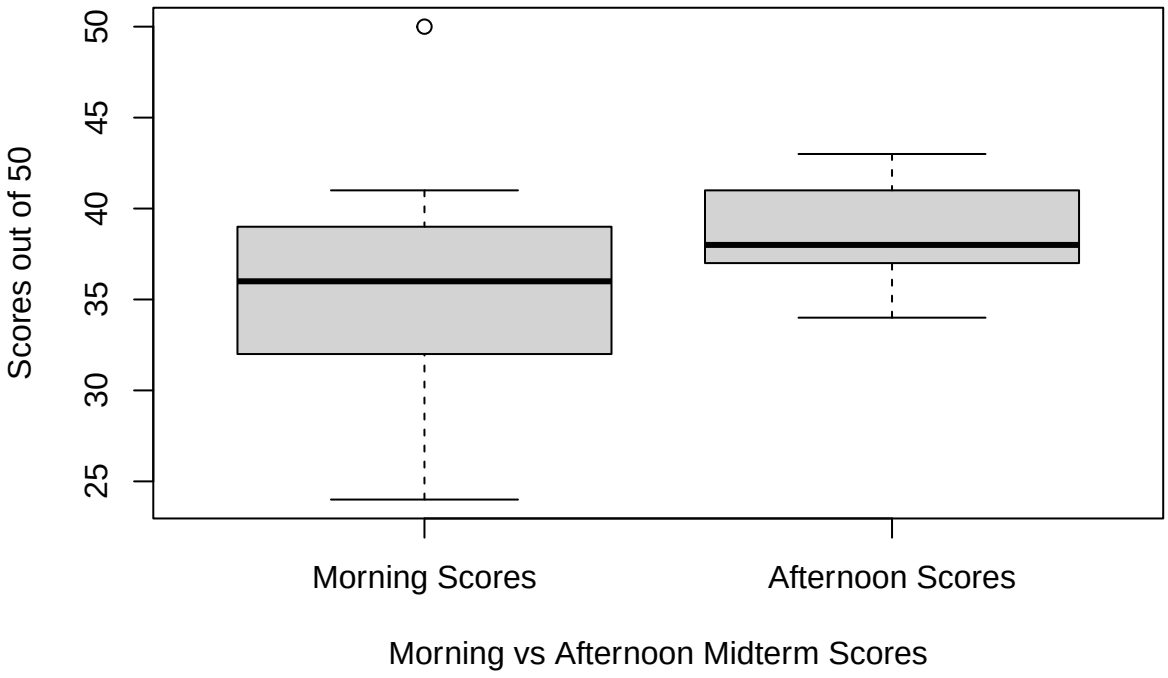
```
morning = c(37,39,27,33,29,32,39,40,40,50,39,40,33,39,38,29,24,31,
            27,36,30,36,40,39,30,41,41,34,32,40,31,32,38,39,33,32,39)

afternoon = c(38,36,40,37,42,38,37,41,43,39,40,36,37,34,41,36,39,37,
              40,38,35,34,38,42,39,41,40,41,37,41,37,41,35,38,41,36)

hist(morning, main="Midterm Scores from Morning Exams", xlab="Scores out of 50 marks")



hist(afternoon, main="Midterm Scores from Afternoon Exams", xlab="Scores out of 50 marks")
```





(b) [2 marks] Create one side-by-side boxplot of the two sets of marks (i.e. both boxplots on the same axes). The picture should have an appropriate title, the x- and y-axes for the boxplot should have appropriate labels, and the two groups should be labelled. (The code to change the title and axes for a boxplot is the same as how to change the title and axes for a histogram.) Copy and paste this boxplot and the line of code used to create it into your Word document. The boxplots themselves may be either horizontal or vertical (your choice).

```
boxplot(morning, afternoon,
        names=c("Morning Scores", "Afternoon Scores"),
        ylab="Scores out of 50",
        xlab="Morning vs Afternoon Midterm Scores")
```



(c) [2 marks] Use R to calculate the mean and standard deviation of the marks for both the morning and afternoon sections. Copy and paste the relevant commands and output from the R Console Window into your document. Write a short statement summarizing the values of your R output.

```
morning.mean = mean(morning)
morning.sd = sd(morning)

afternoon.mean = mean(afternoon)
afternoon.sd = sd(afternoon)

output = paste("MorningMmean: ", morning.mean, "\n",
               "Morning Standard Deviation: ", morning.sd, "\n",
               "Afternoon Mean: ", afternoon.mean, "\n",
               "Afternoon Standard Deviation: ", afternoon.sd, "\n")

writeLines(output)
## MorningMmean:  35.3783783783784
##  Morning Standard Deviation:  5.28284094104757
##  Afternoon Mean:  38.4722222222222
##  Afternoon Standard Deviation:  2.40815413734386
```

(d) [1 mark] Answer the following question: Which class appears to have performed better on the test? Write a few sentences explaining your opinion. You should make reference to the some of the relevant features of the two data sets (e.g. the mean or median, the spread of the data, minimum/maximum values, etc.). Use your results from both parts (b) and (c) to support your statement. You may wish to run the summary command in R for each class section to gather information about the maximum and minimum values.

```
# The Afternoon class had a higher mean than the morning class, 38.5 vs 35.4.
# The Afternoon class had a narrower spread as shown by the standard deviation,
# 2.4 vs 5.3.
# The min value of the Afternoon class was higher, 34 vs 24.
# The maximum value (excluding outlires) was higher in the Afternoon class, 43 vs 41.
# The morning class did have an outlier however with a maximum of 50.
# Overall, given the mean, minumum value, and spread of the data we can conclude that
# The Afternoon class did much better.
```

Part 2

We are interested in analyzing the relationship between the height and the volume of timber produced by a felled black cherry tree. There is a built in data set in R that we will access for this purpose. Use the command attach(trees) to import the trees data set. This data set contains the height (measured in feet) and the volume of timber (measured in cubic feet) measurements for 31 trees (i.e. we have bivariate data here). The height measurements are already stored in a vector called Height and the volume measurements are already stored in a vector called Volume (i.e. attaching the trees dataset will also store the data lists of Height and Volume for you, so you can now use these two lists of data and call upon them by name). (Note also that vector names are case sensitive when calling upon the vectors of Height and Volume.)

```
attach(trees)

Girth
```

```
##  [1]  8.3  8.6  8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7 12.0
## [16] 12.9 12.9 13.3 13.7 13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0
## [31] 20.6
```

```
Height
```

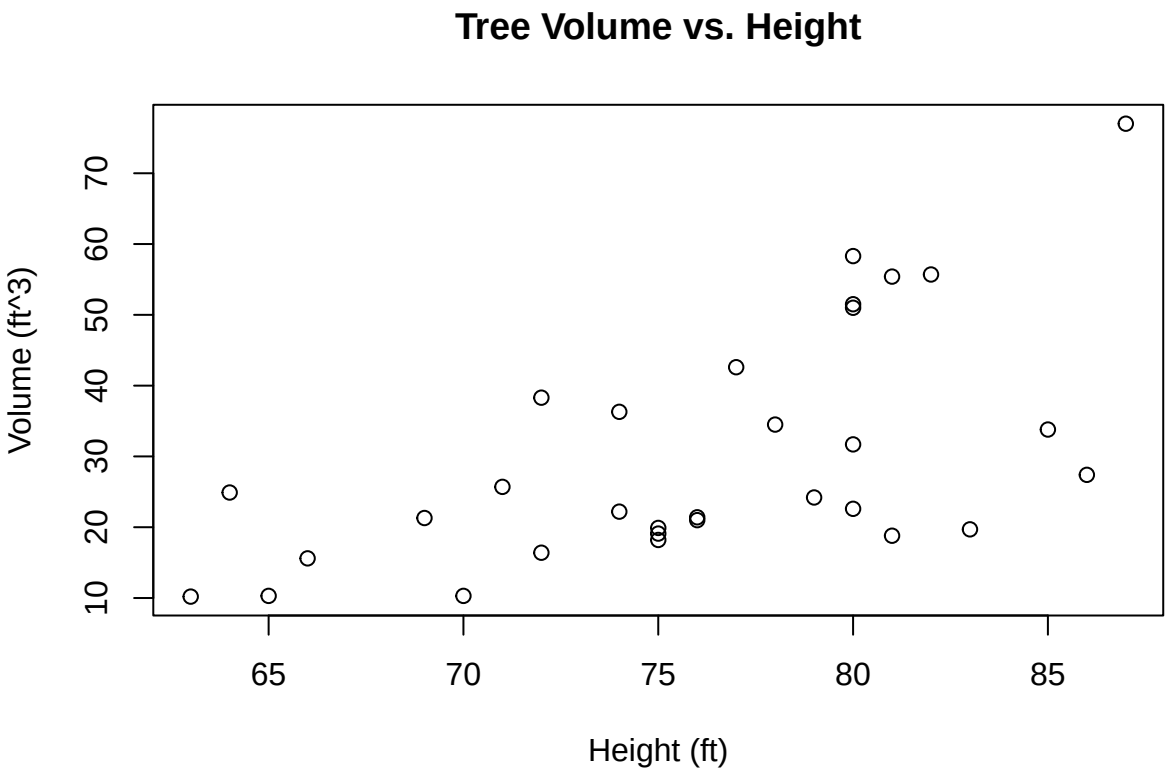
```
##  [1] 70 65 63 72 81 83 66 75 80 75 79 76 76 69 75 74 85 86 71 64 78 80 74 72 77
## [26] 81 82 80 80 80 87
```

```
Volume
```

```
##  [1] 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.1
## [16] 22.2 33.8 27.4 25.7 24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0
## [31] 77.0
```

(a) [2 marks] Create a scatterplot to compare the tree height to the volume of timber. (Hint: we potentially believe that tree height would influence volume of timber, so choose which data set will represent x and y wisely.) Your plot should have an appropriate title and the x- and y-axes should be labelled appropriately. Copy and paste the relevant commands and output from the R Console Window into your Word document.

```
plot(Height, Volume,
     main="Tree Volume vs. Height",
     xlab="Height (ft)",
     ylab="Volume (ft^3)")
```



(b) [1 mark] Describe the relationship (if any) between height and timber volume that this scatterplot shows. (e.g. Is it linear or not? Positive or negative? A strong or weak relationship?)

```
# It appears to have a positive linear relationship
# but because the points do not seem tightly packed,
# it appears to be weak.
```

(c) [1 mark] Use the cor function to compute the correlation coefficient for height and timber volume. (Copy and paste the relevant commands and output from the R Console Window into your Word document.) Does this value agree with your answer about the relationship between height and timber volume? Explain.

```
cor(Height, Volume)
```

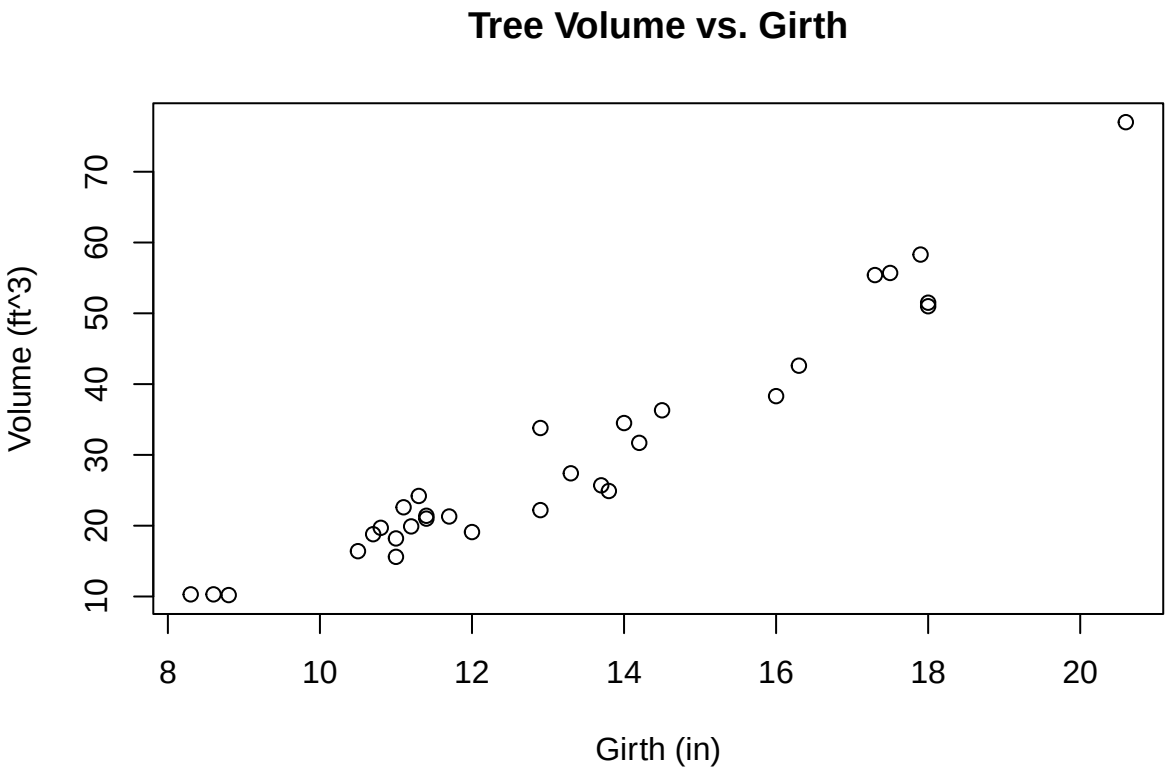
```
## [1] 0.5982497
```

```
# Because the correlation coefficient, ~0.60
# is greater than zero but less than 0.80,
# we say that Height (ft) and Volume (ft^3) has a weak,
# positive, linear relationship.
# This confirms our answer for 2.(b).
```

(d) [2 marks] The trees data set also contains information about the girth of the tree (measured in inches). This information is stored in the Girth vector. Create a scatterplot to compare girth to the volume

of timber, and then use the cor function to compute the correlation coefficient for girth and timber volume. (Copy and paste the R code and output for the scatterplot and the correlation coefficient calculation.)

```
plot(Girth, Volume,
     main="Tree Volume vs. Girth",
     xlab="Girth (in)",
     ylab="Volume (ft^3)")
```




```
cor(Girth, Volume)
```

```
## [1] 0.9671194
```

(e) [1 mark] According to your work in the previous parts of this question, is height or girth a better linear predictor of timber volume? Explain.

```
# Girth is a much better predictor of timber volume than height.
# It's correlation coefficient is ~0.97 which indicates a strong,
# positive, linear relationship.
# Girth's correlation coefficient of 0.97 is greater than
# Height's correlation coefficient of 0.60.
```