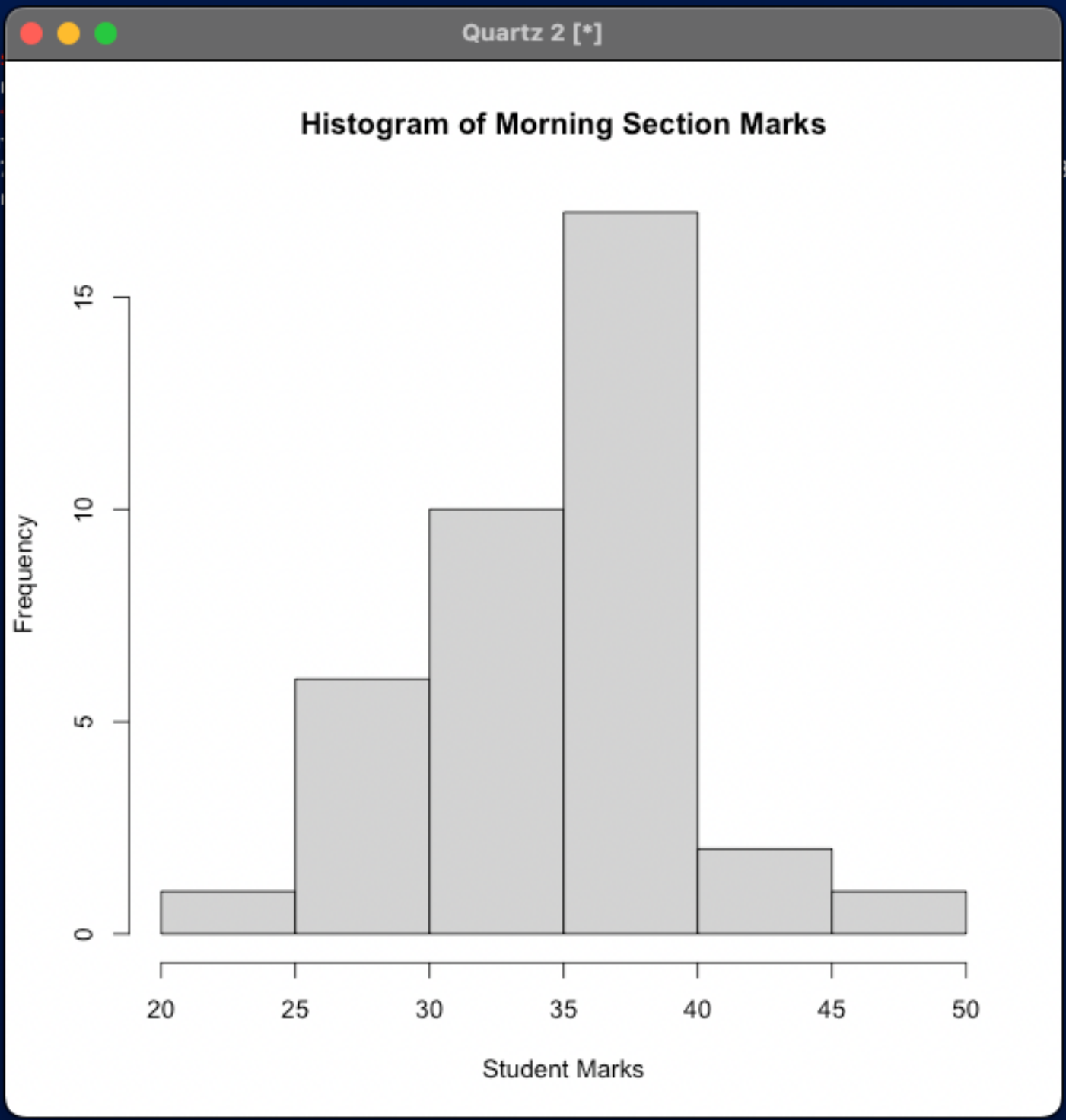
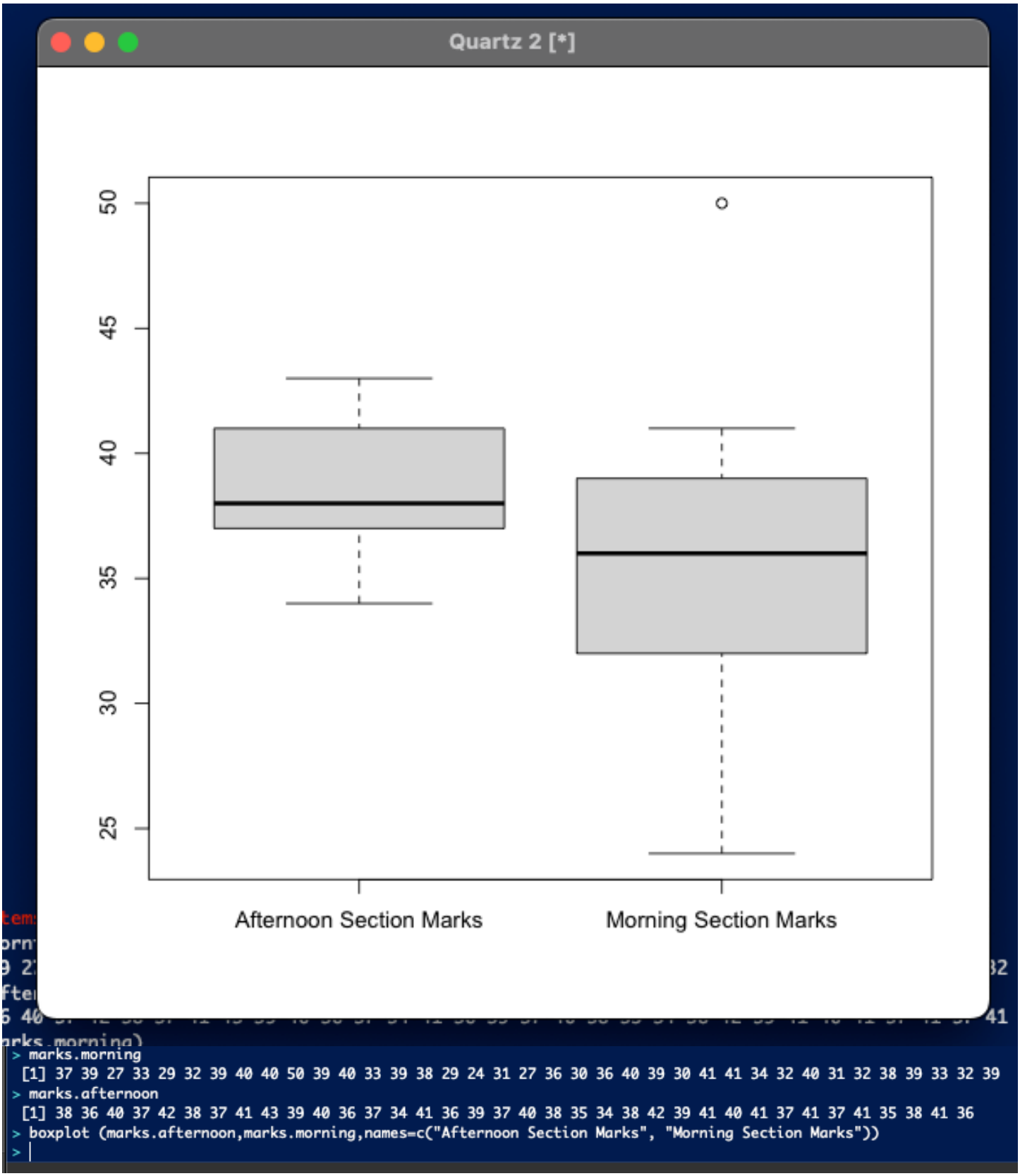
Q-1A

```
> marks.morning
 [1] 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
> marks.afternoon
 [1] 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 (marks.morning)
> hist (marks.morning, main="Histogram of Morning Section Marks", xlab="Student Marks")
>
```



1B




```
> marks.morning
 [1] 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
> marks.afternoon
 [1] 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
> boxplot (marks.afternoon,marks.morning,names=c("Afternoon Section Marks", "Morning Section Marks"))
>
```

1C

The mean for the morning marks would be 35.37838
The mean for the afternoon marks would be 38.47222
The standard deviation for the morning section marks is 5.282841. It means all the marks are on average 5.28 points far from the mean. The standard deviation for the afternoon section marks is 2.408154. It means all the marks in the afternoon section are on average 2.40 points far from the mean.

The calculation from the R is given as follows:

```
> marks.morning
 [1] 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
> marks.afternoon
 [1] 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
> mean (marks.morning)
[1] 35.37838
> mean (marks.afternoon)
[1] 38.47222
> sd (marks.morning)
[1] 5.282841
> sd (marks.afternoon)
[1] 2.408154
>
```

1D

The afternoon section/class seems to have better performed in the test than the morning counterpart.

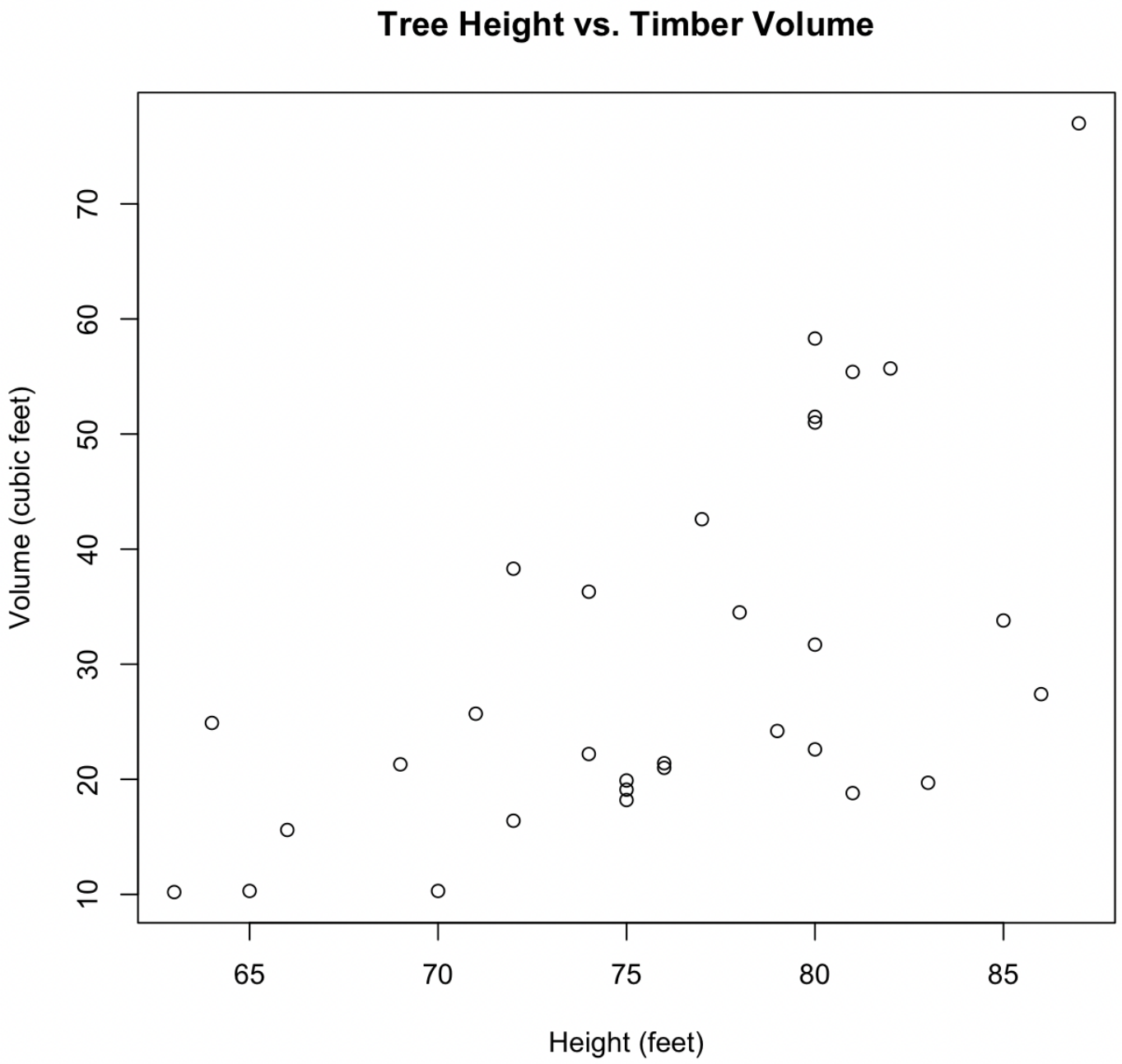
The results from the morning section are more spread out and have a highest mark which is more than the highest mark from the afternoon section. But if we look at the mean of both sections, the afternoon section has a higher mean than morning section. More students seem to get 38 in the afternoon section than 36 in morning section, which also indicates more students getting a higher grade in the afternoon section (although this might not indicate the better performance of a section, but still a good thing to note). Lastly, looking at the first and last quartile (first 25% and last 25%), we can see that more students got grades lower than 32 in the morning section, which is lower than the minimum in the afternoon section. For the last quartile, more students got marks in the 39-50 points than afternoon section (41-43). Even though more people got higher grades on average in the last quartile, that's not close to how the morning section performed for the first quartile, 25% students got lower than the minimum in the afternoon section. TLDR: Afternoon section performed better even though the morning section has a higher maximum grade.

This can be more explained though the intuitive summary function from R:

```
> summary (marks.morning)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.00   32.00   36.00   35.38   39.00   50.00
> summary (marks.afternoon)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.00   37.00   38.00   38.47   41.00   43.00
>
```

End of Question 1

Q-2A (R Functions on the next page)




2B

Linear relationships show a straight-line pattern in the scatterplot, while non-linear relationships show a curved pattern. Positive relationships show that as one variable increases, the other variable also increases. Negative relationships show that as one variable increases, the other decreases. The strength of the relationship is determined by how closely the points in the scatterplot follow the pattern of the relationship. A strong relationship means that the points are close to a straight line or a curved line, while a weak relationship means that the points are widely scattered. [1]

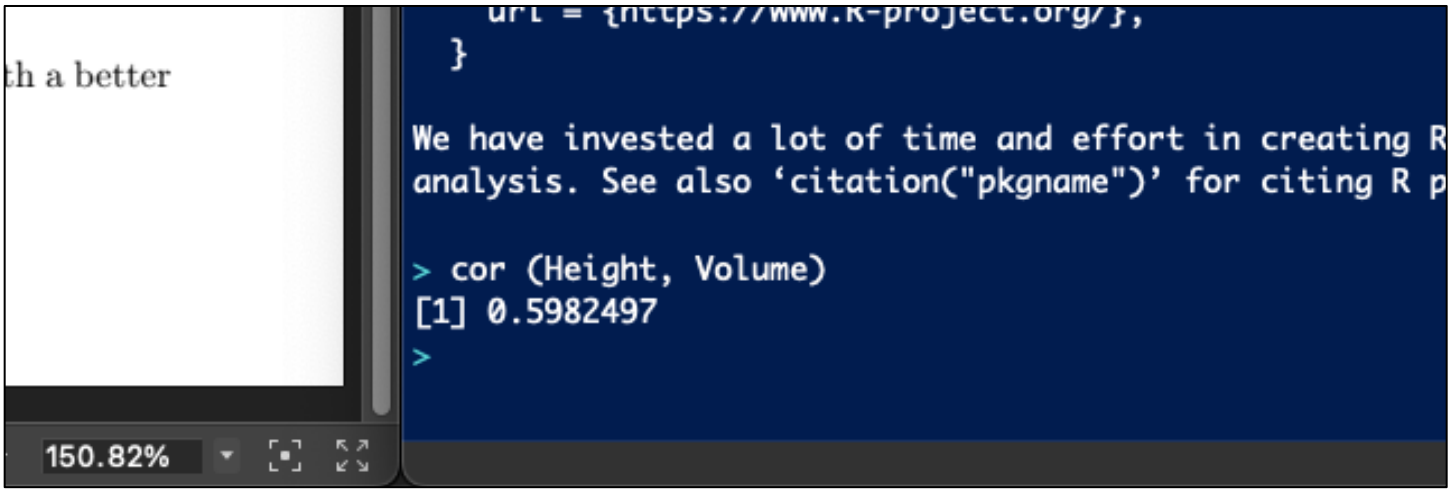
1. **Non-linear**: The relationship between the two variables shows a curved pattern in the scatterplot, which suggests a non-linear relationship.
2. **Positive:** As the tree height increases, the timber volume also increases, indicating a positive relationship between the two variables.
3. **Strong:** The points in the scatterplot follow a pattern that is close to a curved line, which suggests a strong relationship between the two variables.

[1] J. Frieman, D. A. Saucier, and S. S. Miller, Principles & methods of statistical analysis. Los Angeles, CA: SAGE Publications, Inc., 2018.

```
> attach(trees)
The following objects are masked from trees (pos = 3):

    Girth, Height, Volume

The following objects are masked from trees (pos = 4):

    Girth, Height, Volume

> 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 81 82 80 80 80 87
> Weight
Error: object 'Weight' not found
> 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 22.2 33.8 27.4 25.7 24.9 34.5 31.7
[23] 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
> plot(Height,Volume,main="Tree Height vs. Timber Volume",xlab="Height (feet)",ylab="Volume (cubic feet)")
```

Figure: R Functions of 2A

2C



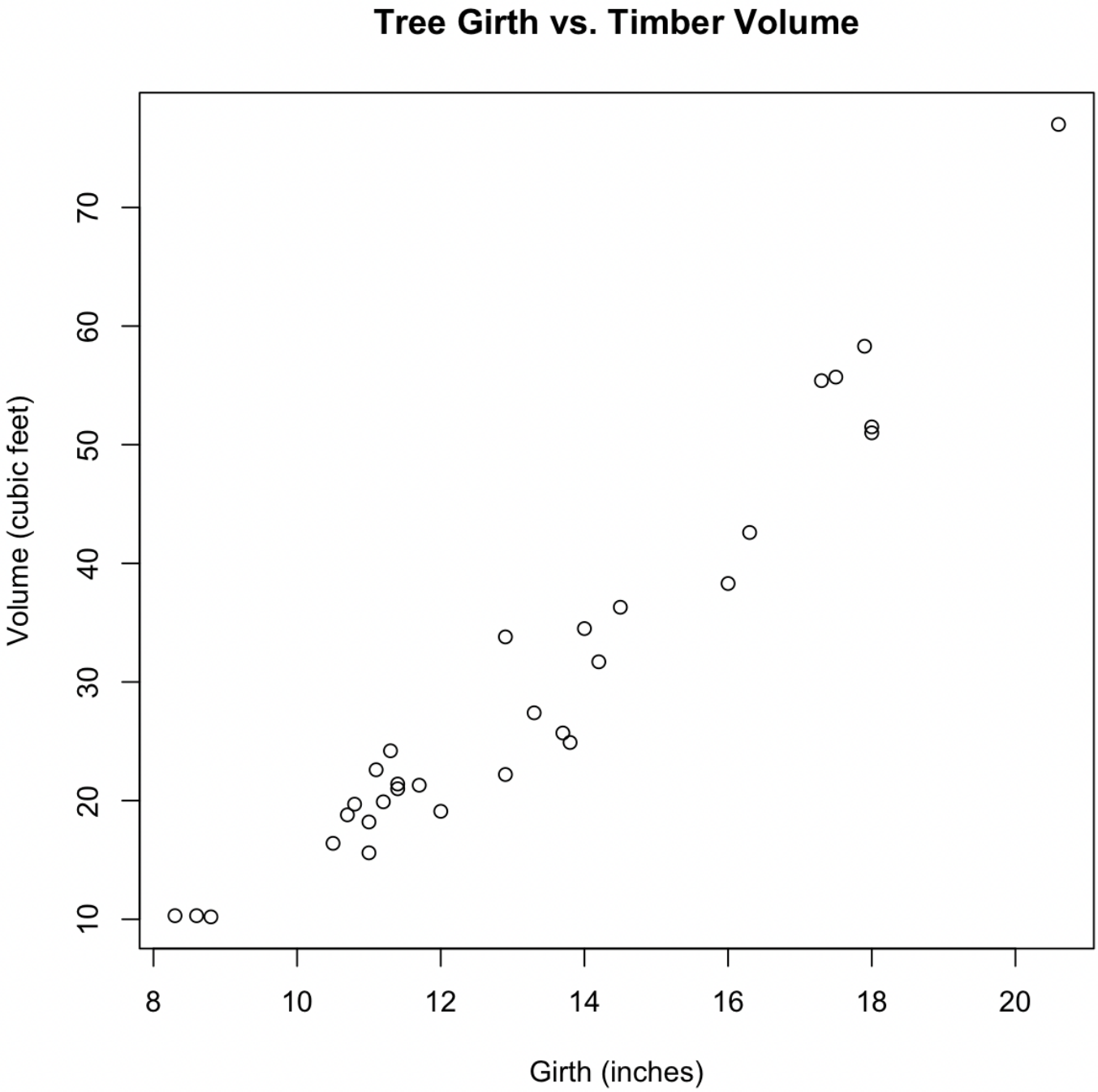

The correlation coefficient between tree height and timbre volume is 0. 5982497, which is a moderate strong relationship between two variables. This means that as timber volume increased, the tree height also increased, but not in a straight line (hence the correlation is not 1). It means the correlation coefficient is suggesting that the relationship is not a weaker one but is also not a really stronger one either. A stronger relationship is one with the correlation coefficient of 1, which doesn't happen in our case. Since it's close to 1, it's a moderate stronger relationship.

2D

```
>
>
> 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 22.2 33.8 27.4 25.7 24.9 34.5 31.7
[23] 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
> 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 12.9 12.9 13.3 13.7 13.8 14.0 14.2
[23] 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0 20.6
> plot(Girth,Volume,main="Tree Girth vs. Timber Volume",xlab="Girth (inches)",ylab="Volume (cubic feet)")
>
>
```

Figure: Plotted R Functions for Volume and Girth.

Looking at the scatterplot, we can see that the relationship is strong, positive, and linear. In comparison with the *Volume vs. Height*, *Volume vs. Girth* is more linear, stronger, and more positive.




```
>
> 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 22.2 33.8 27.4 25.7 24.9 34.5 31.7
[23] 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0 77.0
> 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 12.9 12.9 13.3 13.7 13.8 14.0 14.2
[23] 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0 20.6
> plot(Girth,Volume,main="Tree Girth vs. Timber Volume",xlab="Girth (inches)",ylab="Volume (cubic feet)")
> cor (Volume, Girth)
[1] 0.9671194
>
>
```

Figure: Correlation Coefficient shows a linear, strong, and positive relationship between Timber Volume and Tree Girth.

2E

Tree Girth is a better predictor of Timber Volume, since looking at the scatterplot, we can see that the relationship is strong, positive, and linear. In comparison with the *Volume vs. Height*, *Volume vs. Girth* is more linear, stronger, and more positive. Also, the Correlation Coefficient is closer to 1 in case of *Volume vs. Girth*, than *Volume vs. Height*, suggesting a stronger and positive linear relationship with Girth than Height.

End of Question 1