STAT 5650 Statistical Learning and Data Mining 1 Spring 2020

Name: Michael Huber

Submission Date: 01/27/2020

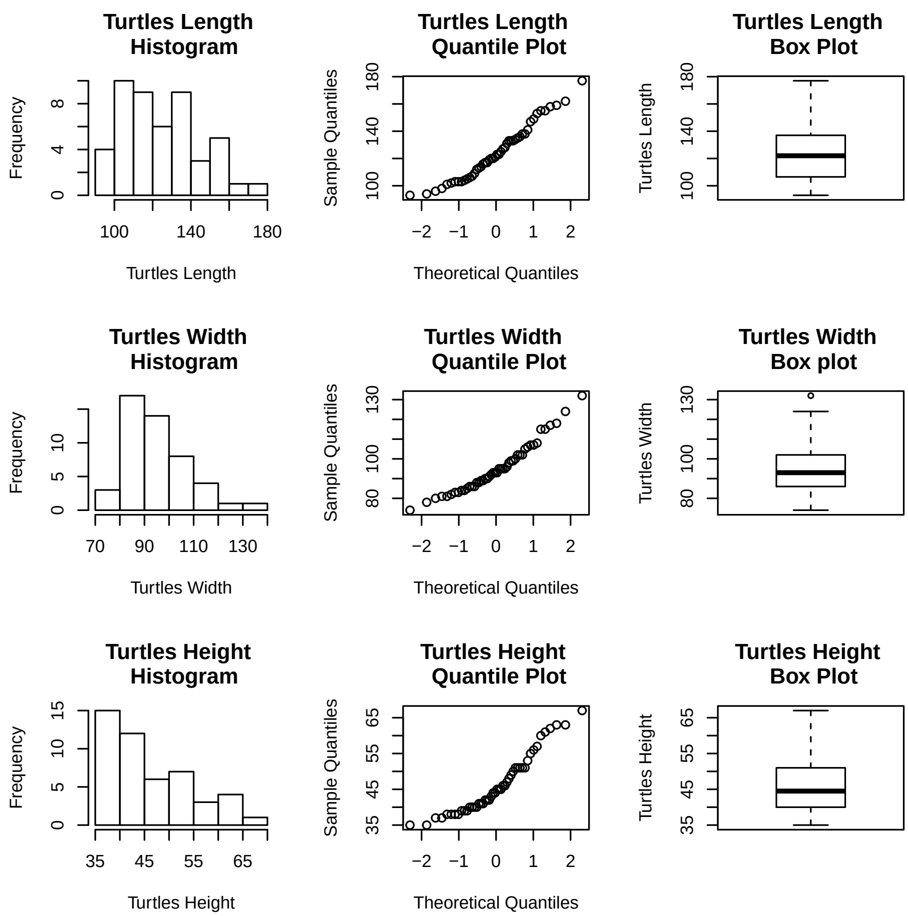
Homework 1

100 Points — Due Wednesday 01/29/2020 (via Canvas by 11:59pm)

The purpose of this homework is to get you comfortable using R to carry out basic matrix and vector calculations. The turtle data contains measurements (length, width, height) on male and female turtles of the same species. If you are familiar with SAS but not R, you can check many of the calculations in SAS but please do the calculations in R.

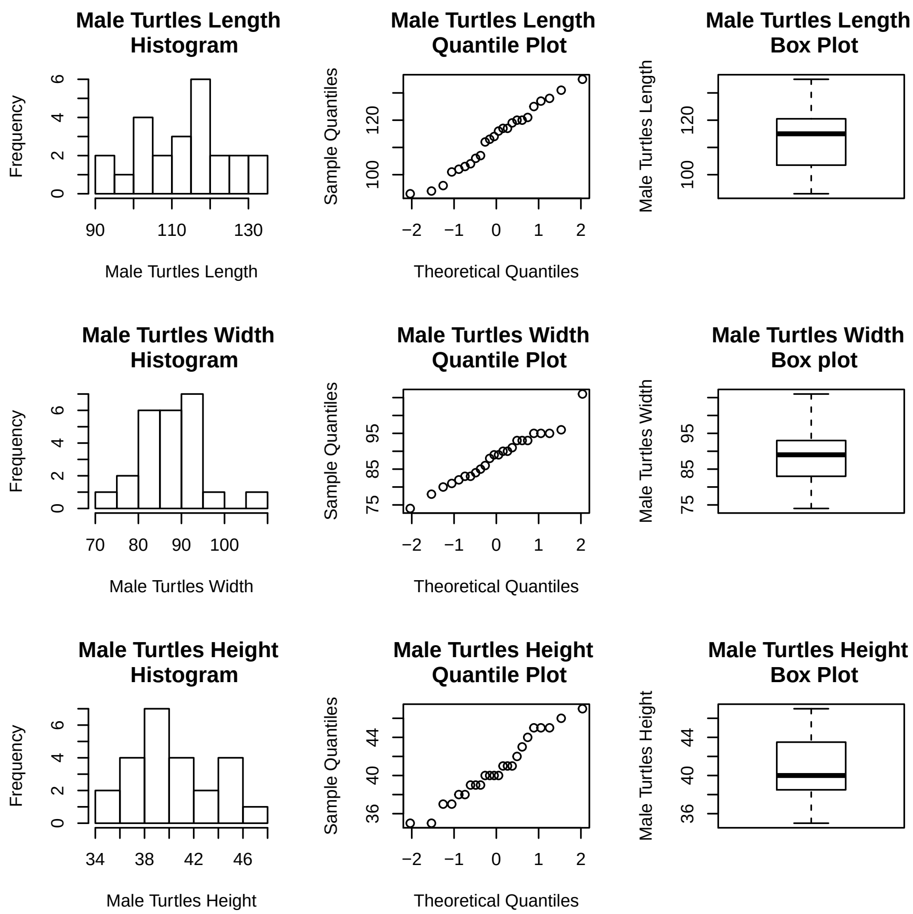
(i) **Question 1:** Graphically summarize the distributions of the three variables using boxplots, histograms, and normal quantile plots. Do the summaries for the combined data and for each gender of turtle.

(a) Combined Data Summary Plots



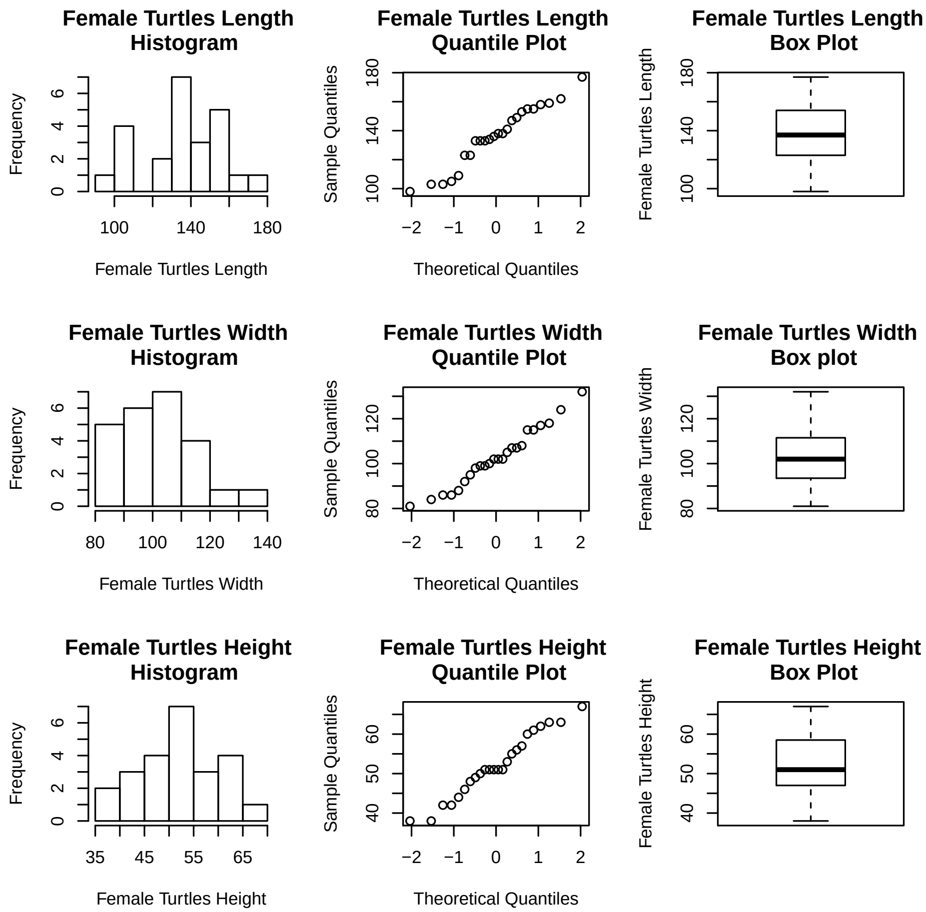
Summary: The combined turtle data is close to a normal distribution. We can see some right skewness in the histogram for the height of the turtle. We can also see that the quantile plot dips down slightly and the Box plot is also pushed down towards the bottom. Plots for length and width however, do appear to be more normally distributed. With only some slight variations.

(b) Male Turtle Summary Plots



Summary: Looking at just the male turtle data. The plots appear to be close to a normal distribution across the length, width and height. They do differ slightly in the histogram for the height and length but they are close enough that I would still say they have a normal distribution.

(c) Female Turtle Summary Plots



summary: The female turtle data also has an approximately normal distribution. The female length and width do appear to vary a little in the histogram and the quantile plot. But the box plot for all variables appear to be normally distributed.

(ii) **Question 2:** Compute the covariance matrices and the correlation matrices for the male and female turtles, and visually compare them. (Later we will determine how to formally compare covariance matrices and mean vectors for different groups.)

(a) Male and Female Covariance Matrices

```
[1] "Male Covariance Matrix"

          Length    Width   Height
Length 138.76630 79.14674 37.37500
Width   79.14674 50.04167 21.65399
Height  37.37500 21.65399 11.25906

[1] "Female Covariance Matrix"

         Length    Width    Height
Length 451.5199 270.9746 165.95471
Width  270.9746 171.7319 101.84420
Height 165.9547 101.8442  64.73732
```

Comparison: Looking at the covariance matrices for male and female turtles. We can see that they all have an increasing linear relationship. The covariance values for female turtles is much greater than that for male turtles. But generally you will just use covariance to determine whether the variables have a positive linear relationship or a negative linear relationship.

(b) Male and Female Correlation Matrices

```
[1] "Male Correlation Matrix"

          Length     Width    Height
Length 1.0000000 0.9497846 0.9455580
Width  0.9497846 1.0000000 0.9122648
Height 0.9455580 0.9122648 1.0000000

[1] "Female Correlation Matrix"

          Length     Width    Height
Length 1.0000000 0.9731162 0.9706748
Width  0.9731162 1.0000000 0.9659029
Height 0.9706748 0.9659029 1.0000000
```

Comparison: The correlation matrices for females and males show that all of the variables are positively correlated and that they have very strong correlation between them. The correlation values do show greater strength among the female turtles. But the male turtles are still close behind them.

(iii) **Question 3:** For one of the genders, compute the matrix $T = \frac{1}{n-1}(\mathbf{Y}^T\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^T)$ and compare it with the covariance matrix you previously obtained.

```
[1] "Matrix Computed Manually"

          Length    Width   Height
Length 138.76630 79.14674 37.37500
Width   79.14674 50.04167 21.65399
Height  37.37500 21.65399 11.25906
```

```
[1] "Matrix Computed with cov() function"

          Length    Width   Height
Length 138.76630 79.14674 37.37500
Width   79.14674 50.04167 21.65399
Height  37.37500 21.65399 11.25906
```

Description: The first matrix I computed using the formula given in the question. The second matrix was the covariance matrix I obtained previously using the cov function. As you can see they produce the same answer. The equation is just the full equation for creating a covariance matrix. Which means the function cov is doing something similar when you use the given data as input values.

(iv) <u>**Question 4:**</u> For one of the covariance matrices, compute the eigenvalues and eigenvectors of the inverseof the covariance matrix. What is the relationship between the eigenvalues and eigenvectors of a covariance matrix and its inverse?

(a) Eigenvalues and Eigenvectors of the inverse of the covariance matrix.

```
[1] 0.905934472 0.271108226 0.005120993

            [,1]        [,2]      [,3]
[1,]  0.23653541  0.48810477 0.8401219
[2,]  0.04687583 -0.86938426 0.4919082
[3,] -0.97049145  0.07697229 0.2285205
```

(b) Eigenvalues and Eigenvectors of the covariance matrix

```
[1] 195.274633   3.688564   1.103833

          [,1]        [,2]        [,3]
[1,] 0.8401219  0.48810477 -0.23653541
[2,] 0.4919082 -0.86938426 -0.04687583
[3,] 0.2285205  0.07697229  0.97049145
```

(c) <u>Comparison:</u> The eigenvalues of the inverse are just the reciprocals of the normal covariance matrix. So 1/195 should give you the eigenvalue of the inverse. The sign doesn't matter but the eigenvectors will be the same for the inverse just in different positions in the matrix. Since 195 is the largest eigenvalue of the normal matrix. 1/195 makes it the smallest eigenvalue of the inverse matrix.

(v) **Question 5:** Letting $\mathbf{u_1}$ denote the eigenvector corresponding to the largest eigenvalue, λ_1 . Verify that $\mathbf{Su_1} = \lambda_1 \mathbf{u}_1$ and $\mathbf{u}_1^T \mathbf{Su_1} = \lambda_1$.

(a) Verification that $\mathbf{Su_1} = \lambda_1 \mathbf{u}_1$

```
       Length    Width   Height
[1,] 164.0545 96.05719 44.62425

         [,1]     [,2]     [,3]
[1,] 164.0545 96.05719 44.62425
```

Description: The first output is the covariance matrix $\mathbf{S}$ times by the vector u_1. Below this is $\lambda_1 \mathbf{u_1}$. The outputs of the two equations are the same showing that they do in fact equal one another.

(b) Verification that $\mathbf{u}_1^T \mathbf{Su_1} = \lambda_1$.

```
         [,1]
[1,] 195.2746

[1] 195.2746
```

Description: The first output is $\mathbf{u}_1^T \mathbf{Su_1}$. The second output is the value of λ_1. This shows that the output of the first equation is the same as the second verifying that this is correct.

(vi) **Question 6:** The length, width, and height are all measured in the same units. Is there any reason we might prefer to use the correlation matrices over the covariance matrices if we were to carry out principal components analysis on this data?

Answer: Because the variances of the variables differ greatly from one another the correlatin matrix would be a good choice because it would standardize the values for the principal component analysis.

(vii) **Question 7:** The length, width, and height are all measured in the same units. Is there any reason we might prefer to use the correlation matrices over the covariance matrices if we were to carry out principal components analysis on this data?

(a) Covariance Matrix Eigenvalues and Eigenvectors.

```
[1] 678.365651   6.769697   2.853783

          [,1]       [,2]       [,3]
[1,] 0.8138808  0.5548963 -0.1723025
[2,] 0.4961059 -0.8180268 -0.2910518
[3,] 0.3024516 -0.1514012  0.9410636
```

(b) Correlation Matrix Eigenvalues and Eigenvectors.

```
[1] 2.93979909 0.03433769 0.02586323

           [,1]       [,2]       [,3]
[1,] -0.5781417 -0.1373949  0.8042853
[2,] -0.5771970 -0.6278484 -0.5221591
[3,] -0.5767112  0.7661129 -0.2836814
```

(c) Comparison: Based on the eigenvalues I would recommend retaining the first principal component in both instances. Then going in and comparing their eigenvectors with the covariance matrix. The Length of the turtle is weighted as having the biggest effect on the first principle component. In the correlation matrix all of the variables are weighted pretty evenly in the first principle component.