Statistical Learning and Data Mining 1 Spring 2020

Name: Michael Huber

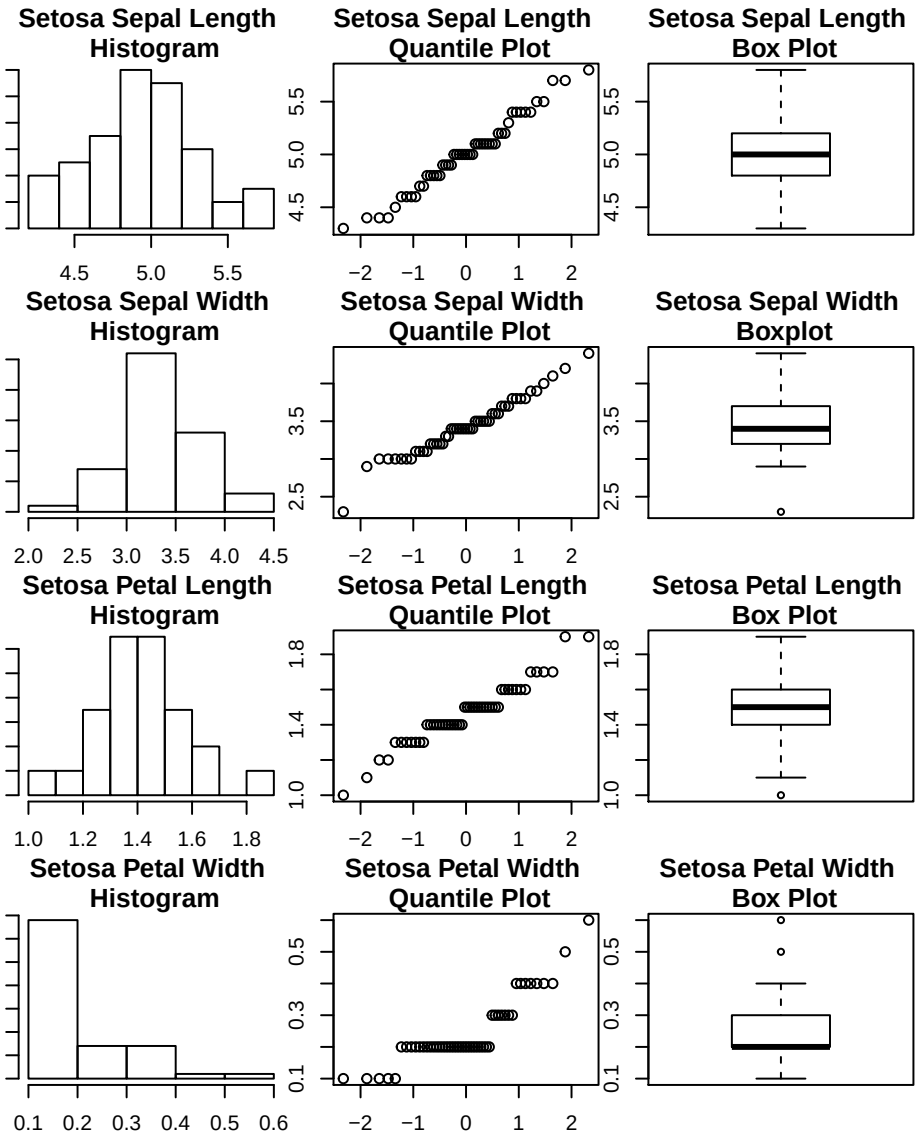
Submission Date: 02/17/2020

Homework 2 (01/31/2020)

100 Points — Due Monday 02/17/2020 (via Canvas by 11:59pm)

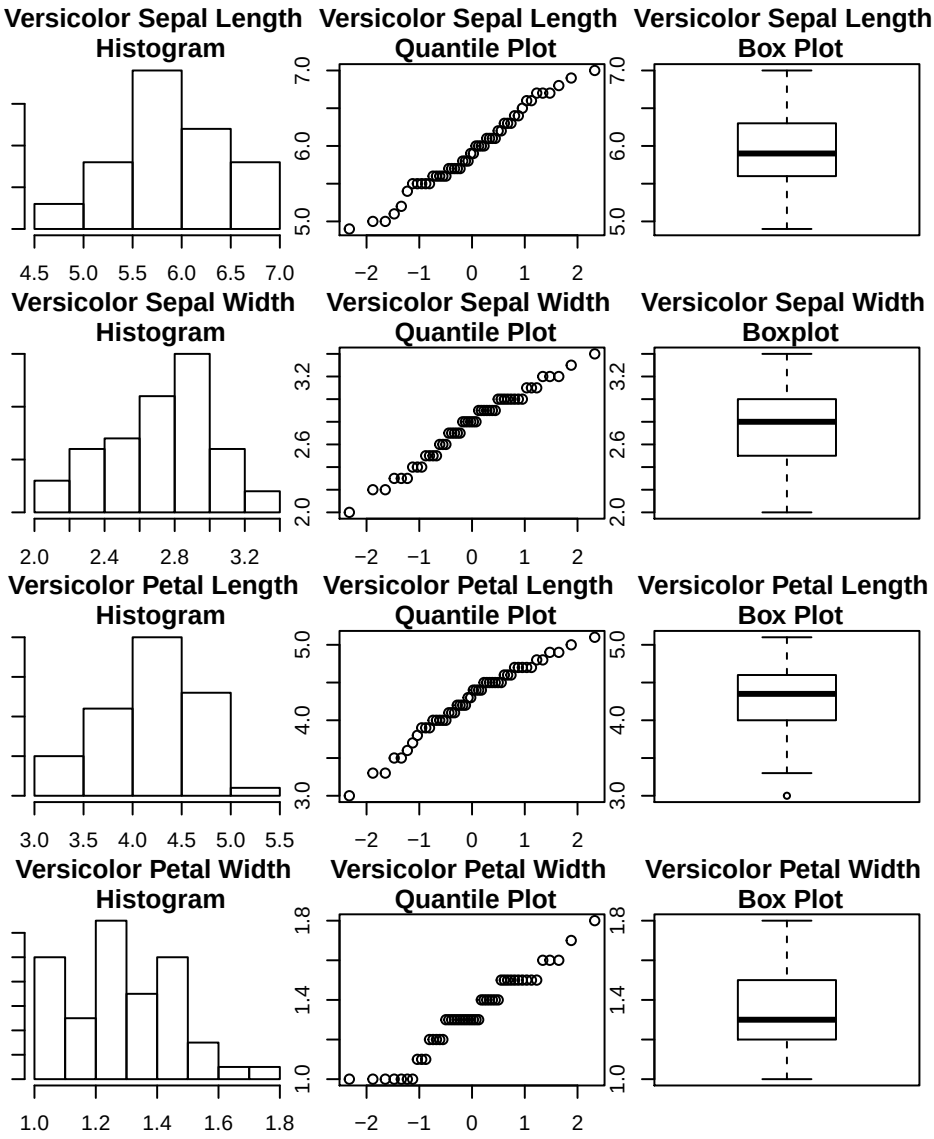
(i) **Question 1:** This question concerns Fisher's iris data, one of the most well-known and, perhaps, overused datasets. I have placed the data ("Iris.csv") on the Canvas site for the class. For this homework, I would like you to apply linear and quadratic discriminant analysis to see if the measured values or petal length and width and sepal length and width can be used to discriminate among the three species of iris.

(a) Summarize the four measured variables for the three types of iris. Are the data approximately normal, and do they look like they have the same covariance matrix for all 3 species?

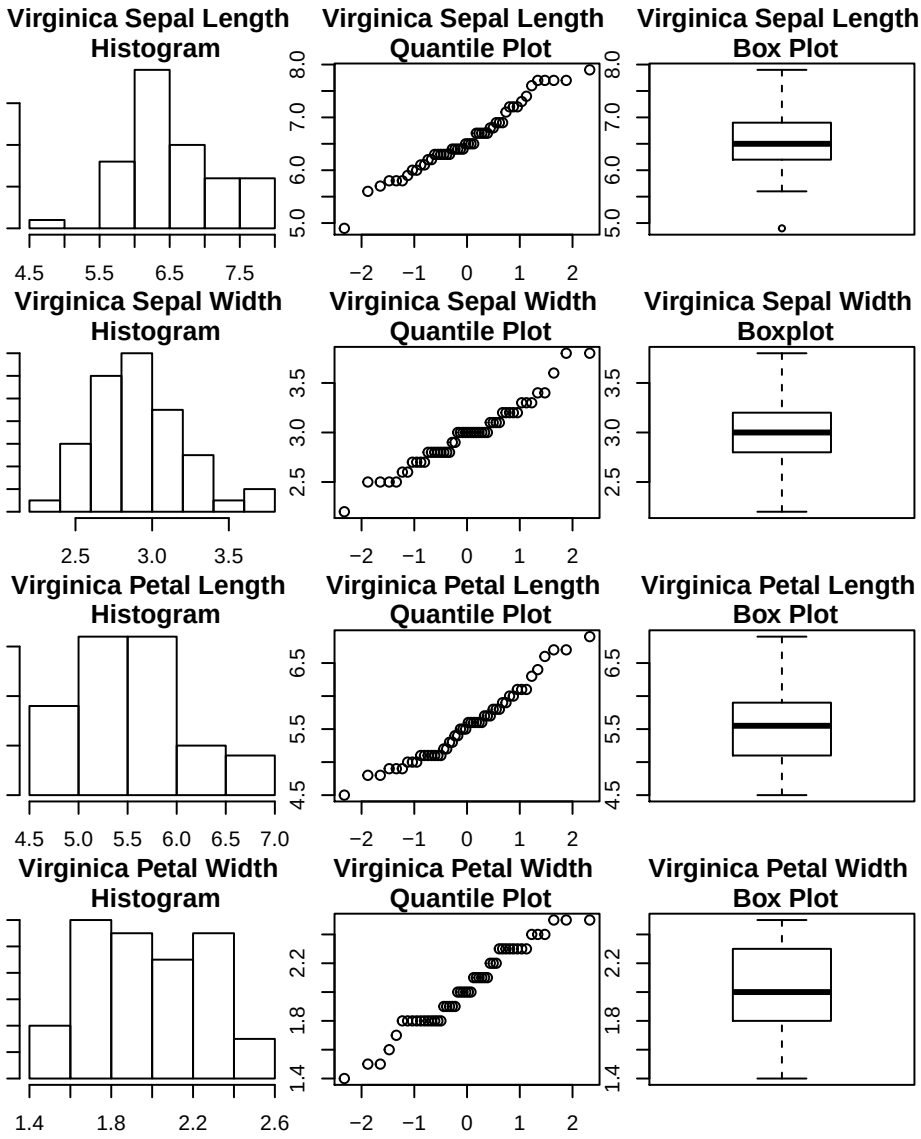


Setosa: This species is approximately normal for the variables Sepal Width,

Sepal Length, and Petal Length. In looking at the histogram for Petal Width Setosa appears to have a right skew to the data.



Versicolor: For this species of Iris the variables of Sepal Width, Sepal Length, and Petal Length are approximately normal. While the Petal Width appears to be closer to normal than the Setosa species, but there does still appear to be a large spike near the left side of the histogram.



Virginica: For this species of Iris the variables of Sepal Width, Sepal Length, and Petal Length are approximately normal. While the Petal Width appears to be almost a normal distribution but it appears to be slightly bimodal in its distribution.

```
[1] "Setosa Covariance Matrix"
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	0.12424898	0.099216327	0.016355102	0.010330612
SepalWidth	0.09921633	0.143689796	0.011697959	0.009297959
PetalLength	0.01635510	0.011697959	0.030159184	0.006069388
PetalWidth	0.01033061	0.009297959	0.006069388	0.011106122

```
[1] "Versicolor Covariance Matrix"
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	0.26643265	0.08518367	0.18289796	0.05577959
SepalWidth	0.08518367	0.09846939	0.08265306	0.04120408
PetalLength	0.18289796	0.08265306	0.22081633	0.07310204
PetalWidth	0.05577959	0.04120408	0.07310204	0.03910612

```
[1] "Virginica Covariance Matrix"
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	0.40434286	0.09376327	0.30328980	0.04909388
SepalWidth	0.09376327	0.10400408	0.07137959	0.04762857
PetalLength	0.30328980	0.07137959	0.30458776	0.04882449
PetalWidth	0.04909388	0.04762857	0.04882449	0.07543265

Summary of Covariance Matrices: The covariance matrices of the 3 different Iris species appear to all differ from one another. Some of the corresponding values across matrices appear to be close in value to one another but other parts of the matrices have very different values.

(b) Test to determine whether the covariance matrices for the three species may be pooled.

The hypotheses are defined as

H0: The Covariance matrices are homogeneous

H1: The Covariance matrices are not homogeneous

```
Summary for Box's M-test of Equality of Covariance Matrices

Chi-Sq:         140.943
df:         20
p-value: < 2.2e-16

log of Covariance determinants:
    setosa versicolor  virginica     pooled
-13.067360 -10.874325  -8.927058  -9.958539

Eigenvalues:
       setosa  versicolor  virginica     pooled
1 0.236455690 0.487873944 0.69525484 0.44356592
2 0.036918732 0.072384096 0.10655123 0.08618331
3 0.026796399 0.054776085 0.05229543 0.05535235
4 0.009033261 0.009790365 0.03426585 0.02236372

Statistics based on eigenvalues:
                setosa   versicolor    virginica       pooled
product   2.113088e-06 1.893828e-05 0.0001327479 4.732183e-05
sum       3.092041e-01 6.248245e-01 0.8883673469 6.074653e-01
precision 5.576122e-03 7.338788e-03 0.0169121236 1.304819e-02
max       2.364557e-01 4.878739e-01 0.6952548382 4.435659e-01
```

Conclusion: Where the p-value for the chi-squared test is so small. We would reject the null hypothesis that the covariance matrices are homogenous which means the covariance matrices for the 3 species cannot be pooled. Instead we would use the within covariance matrices in the discriminant function. This means that mathematically we would choose QDA for our analysis since LDA pools the values. But we will test LDA below to see how it performs.

(c) Apply both LDA or QDA. Obtain the cross-validated confusion matrices and accuracy or error rates (by species and overall).

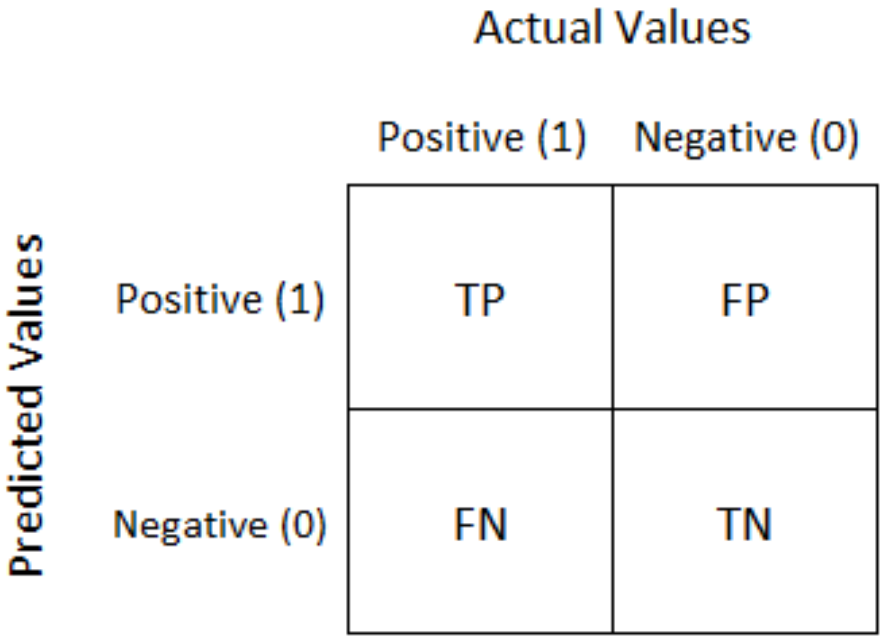


Figure 1: All confusion matrices included in this document will be formatted in the following way unless stated otherwise.

i. **LDA**

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         1
  virginica       0          2        49

Overall Statistics

               Accuracy : 0.98
```

```
                  95% CI : (0.9427, 0.9959)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.97

 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: setosa	Class: versicolor	Class: virginica
Sensitivity	1.0000	0.9600	0.9800
Specificity	1.0000	0.9900	0.9800
Pos Pred Value	1.0000	0.9796	0.9608
Neg Pred Value	1.0000	0.9802	0.9899
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3200	0.3267
Detection Prevalence	0.3333	0.3267	0.3400
Balanced Accuracy	1.0000	0.9750	0.9800

LDA Accuracy by Species and Overall:

Setosa: 1.0

Versicolor: 0.98

Verginica: 0.9750

Overall: 0.98

ii. QDA

Confusion Matrix and Statistics

	actual		
predicted	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	1
virginica	0	3	49

Overall Statistics

```
               Accuracy : 0.9733
                 95% CI : (0.9331, 0.9927)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.96

 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: setosa	Class: versicolor	Class: virginica
Sensitivity	1.0000	0.9400	0.9800
Specificity	1.0000	0.9900	0.9700
Pos Pred Value	1.0000	0.9792	0.9423
Neg Pred Value	1.0000	0.9706	0.9898
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3133	0.3267
Detection Prevalence	0.3333	0.3200	0.3467
Balanced Accuracy	1.0000	0.9650	0.9750

QDA Accuracy by Species and Overall:

Setosa: 1.0

Versicolor: 0.9650

Verginica: 0.9750

Overall: 0.9733

Summary of LDA and QDA: Even though the test from b. would have us conclude that the data should not be pooled and that we would do better using QDA over LDA. We can see from the results of actually running the test that LDA performs slightly better than QDA when we comapare their accuracy scores.

(d) Determine whether some of the measured variables are redundant and can be removed.

i. **Less Petal Width QDA**

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         46         4
  virginica       0          4        46

Overall Statistics

               Accuracy : 0.9467
                 95% CI : (0.8976, 0.9767)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.92

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9200           0.9200
Specificity                 1.0000            0.9600           0.9600
Pos Pred Value              1.0000            0.9200           0.9200
Neg Pred Value              1.0000            0.9600           0.9600
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3067           0.3067
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.9400           0.9400
```

<u>Conclusion:</u> When I remove the variable PetalWidth the overall accuracy decreases to 94.67%. The sensitivity and specificity for Versicolor and Virginica are also decreased. But they stay the same for Setosa.

ii. Less PetalLength QDA

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         45         3
  virginica       0          5        47

Overall Statistics

               Accuracy : 0.9467
                 95% CI : (0.8976, 0.9767)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.92

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9000           0.9400
Specificity                 1.0000            0.9700           0.9500
Pos Pred Value              1.0000            0.9375           0.9038
Neg Pred Value              1.0000            0.9510           0.9694
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3000           0.3133
Detection Prevalence        0.3333            0.3200           0.3467
Balanced Accuracy           1.0000            0.9350           0.9450
```

<u>Conclusion:</u> When I remove the variable PetalLength the overall accuracy decreases to 94.67% as well. The sensitivity and specificity for Versicolor and Virginica are also decreased. But they stay the same for Setosa.

iii. Less Sepal Width QDA

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          2        48

Overall Statistics

               Accuracy : 0.9733
                 95% CI : (0.9331, 0.9927)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.96

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9600           0.9600
Specificity                 1.0000            0.9800           0.9800
Pos Pred Value              1.0000            0.9600           0.9600
Neg Pred Value              1.0000            0.9800           0.9800
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3200           0.3200
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.9700           0.9700
```

<u>Conclusion:</u> When I remove SepalWidth the overall accuracy is the same value as when all 4 variables are present in QDA. It also looks like it increases the accuracy of predicting the Vir-

ginica up to 97% but it lowers the prediction accuracy of the Versicolor from 97.49% down to 97% accuracy. The specificity for Virginica goes up 1% but they drop in all other places except for Setosa stays the same.

iv. Less Sepal Length QDA

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         46         1
  virginica       0          4        49

Overall Statistics

               Accuracy : 0.9667
                 95% CI : (0.9239, 0.9891)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.95

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9200           0.9800
Specificity                 1.0000            0.9900           0.9600
Pos Pred Value              1.0000            0.9787           0.9245
Neg Pred Value              1.0000            0.9612           0.9897
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3067           0.3267
Detection Prevalence        0.3333            0.3133           0.3533
```

```
Balanced Accuracy          1.0000           0.9550          0.9700
```

Conclusion: **Removing the variable SepalLength lowers the overall accuracy slightly and it lowers the accuracy of predicting the Versicolor and Virginica species. Overall it looks like you could remove the SepalWidth Variable and still have near to the same prediction results as you do with all four variables using QDA.**

v. Less Petal Width LDA

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         4
  virginica       0          2        46

Overall Statistics

               Accuracy : 0.96
                 95% CI : (0.915, 0.9852)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.94

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000           0.9600          0.9200
Specificity                 1.0000           0.9600          0.9800
Pos Pred Value              1.0000           0.9231          0.9583
Neg Pred Value              1.0000           0.9796          0.9608
Prevalence                  0.3333           0.3333          0.3333
```

Detection Rate	0.3333	0.3200	0.3067
Detection Prevalence	0.3333	0.3467	0.3200
Balanced Accuracy	1.0000	0.9600	0.9500

<u>Conclusion:</u> When I remove the variable PetalWidth from the LDA model the overall accuracy decreases to 96%.

vi. Less PetalLength LDA

Confusion Matrix and Statistics

predicted \ actual	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	5
virginica	0	3	45

Overall Statistics

Accuracy	0.9467
95% CI	(0.8976, 0.9767)
No Information Rate	0.3333
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.92
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: setosa	Class: versicolor	Class: virginica
Sensitivity	1.0000	0.9400	0.9000
Specificity	1.0000	0.9500	0.9700
Pos Pred Value	1.0000	0.9038	0.9375
Neg Pred Value	1.0000	0.9694	0.9510
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3133	0.3000

```
Detection Prevalence        0.3333            0.3467           0.3200
Balanced Accuracy           1.0000            0.9450           0.9350
```

<u>Conclusion:</u> When I remove the variable PetalLength from the LDA model the overall accuracy decreases to 94.67% which is the same as the QDA model.

vii. Less Sepal Width LDA

```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          2        48

Overall Statistics

               Accuracy : 0.9733
                 95% CI : (0.9331, 0.9927)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.96

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9600           0.9600
Specificity                 1.0000            0.9800           0.9800
Pos Pred Value              1.0000            0.9600           0.9600
Neg Pred Value              1.0000            0.9800           0.9800
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3200           0.3200
```

Detection Prevalence	0.3333	0.3333	0.3333
Balanced Accuracy	1.0000	0.9700	0.9700

<u>Conclusion:</u> When I remove SepalWidth from the LDA model the overall accuracy is the same value as when all 4 variables are present in QDA.

viii. Less Sepal Length LDA

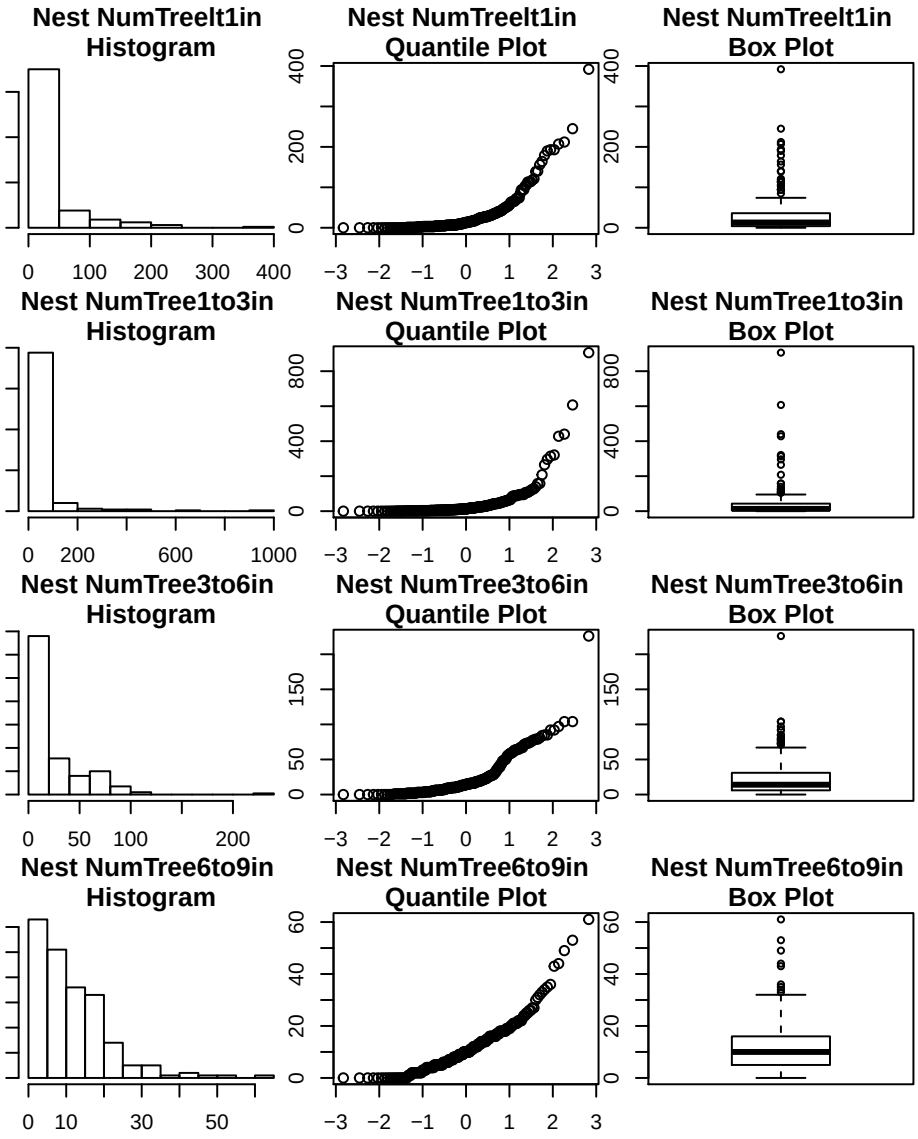
```
Confusion Matrix and Statistics

            actual
predicted    setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         3
  virginica       0          2        47

Overall Statistics

               Accuracy : 0.9667
                 95% CI : (0.9239, 0.9891)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.95

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9600           0.9400
Specificity                 1.0000            0.9700           0.9800
Pos Pred Value              1.0000            0.9412           0.9592
Neg Pred Value              1.0000            0.9798           0.9703
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3200           0.3133
Detection Prevalence        0.3333            0.3400           0.3267
```

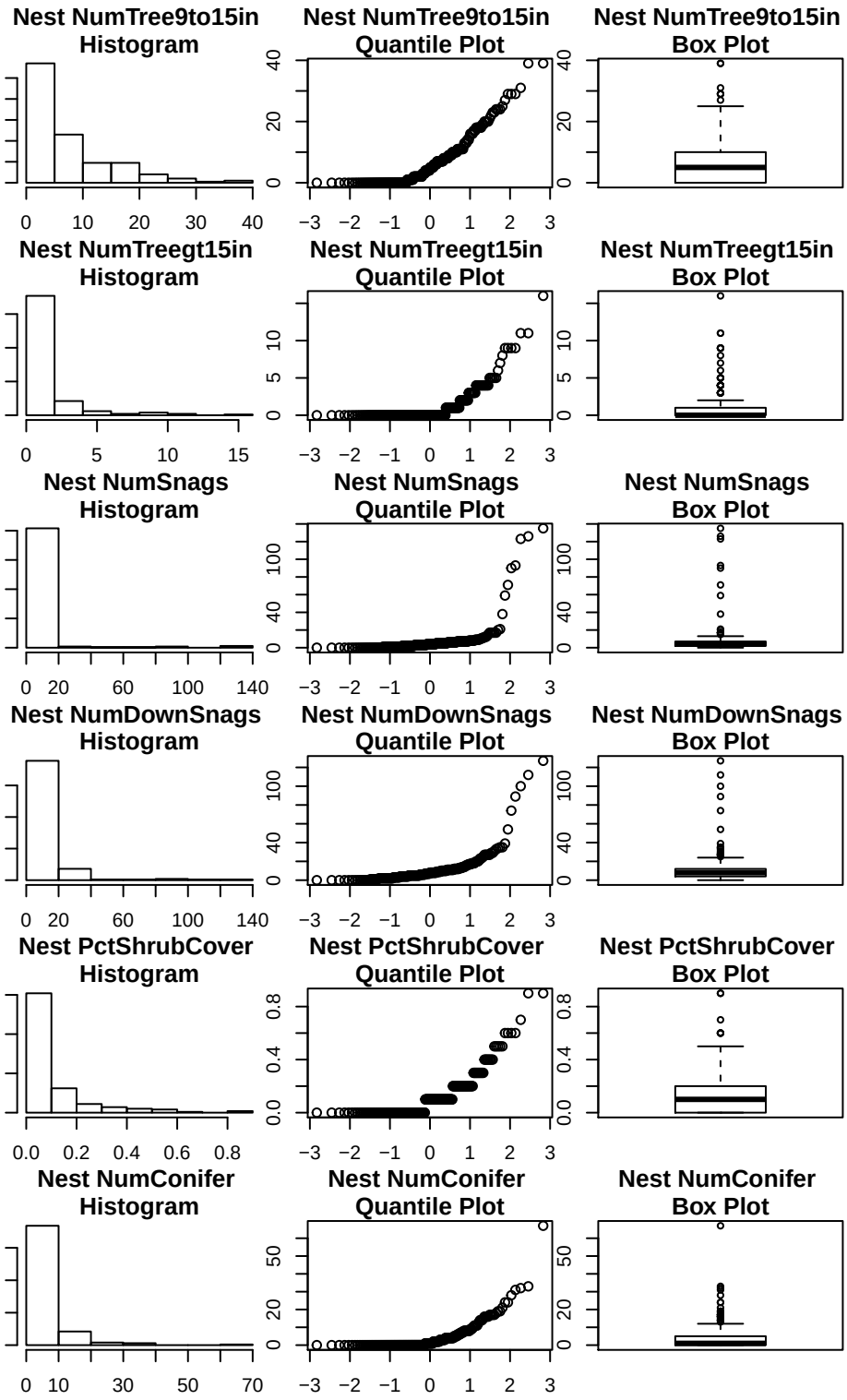
Balanced Accuracy	1.0000	0.9650	0.9600

<u>Conclusion:</u> Removing the variable SepalLength lowers the overall accuracy to 96.67%. So in the end it does not look like removing any of the variables outperforms the accuracy achieved in the LDA model when all 4 variables are present.

(ii) **Question 2:** I have placed a dataset called "Nest.csv" on the Canvas site for the class. This dataset contains data on nest sites for three bird species—(Northern) Flicker, (Mountain) Chickadee, and (Rednaped) Sapsucker—plus a bunch of sites at which none of these birds are nesting. The variable Nest is the response or grouping variable. Species indicates the species of nesting bird, and StandType is a dummy variable coded as 0 for pure aspen forest and 1 for mixed aspen and conifer.

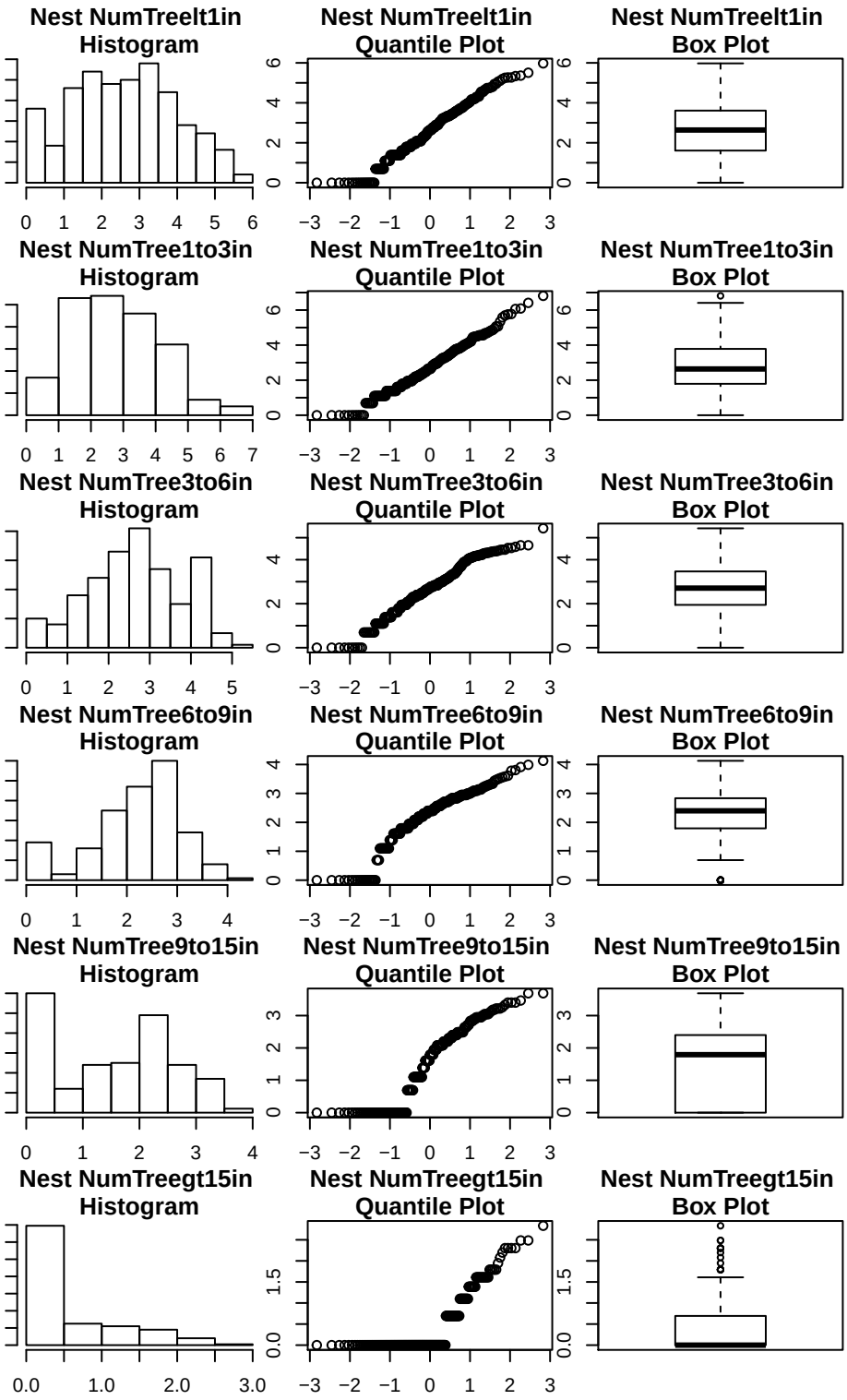
(a) Carry out numerical and graphical summaries of all the predictor variables except StandType. Are the variables approximately normal in distribution? If not, apply some transformation(s) to "improve" the distributions of these variables.

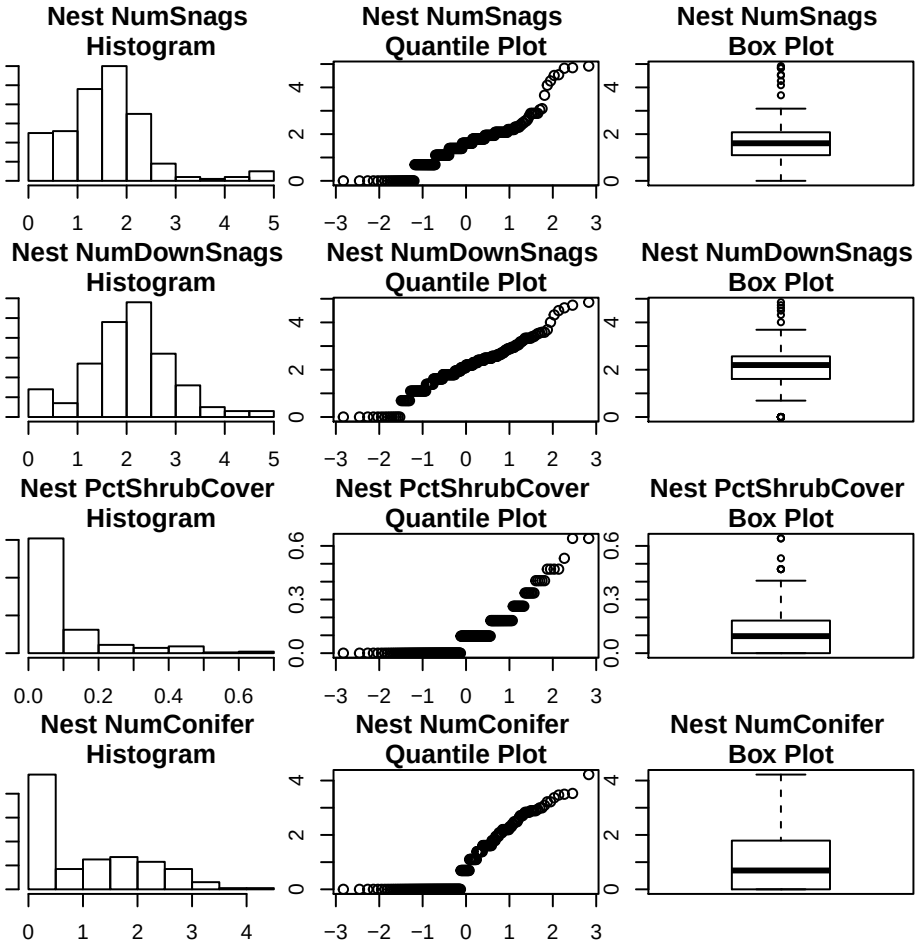



Nest NumTree9to15in Histogram
Nest NumTree9to15in Quantile Plot
Nest NumTree9to15in Box Plot
Nest NumTreegt15in Histogram
Nest NumTreegt15in Quantile Plot
Nest NumTreegt15in Box Plot
Nest NumSnags Histogram
Nest NumSnags Quantile Plot
Nest NumSnags Box Plot
Nest NumDownSnags Histogram
Nest NumDownSnags Quantile Plot
Nest NumDownSnags Box Plot
Nest PctShrubCover Histogram
Nest PctShrubCover Quantile Plot
Nest PctShrubCover Box Plot
Nest NumConifer Histogram
Nest NumConifer Quantile Plot
Nest NumConifer Box Plot

```
     Nest             Species     NumTreelt1in    NumTree1to3in
 Min.   :0.0000   Chickadee: 42   Min.   :  0.00   Min.   :  0.00
 1st Qu.:0.0000   Flicker  : 23   1st Qu.:  4.00   1st Qu.:  5.00
 Median :1.0000   Non-nest :106   Median : 13.00   Median : 13.00
 Mean   :0.5023   Sapsucker: 42   Mean   : 32.16   Mean   : 43.39
 3rd Qu.:1.0000                   3rd Qu.: 36.00   3rd Qu.: 43.00
 Max.   :1.0000                   Max.   :392.00   Max.   :906.00
 NumTree3to6in    NumTree6to9in  NumTree9to15in   NumTreegt15in
 Min.   :  0.00   Min.   : 0.0   Min.   : 0.000   Min.   : 0.000
 1st Qu.:  6.00   1st Qu.: 5.0   1st Qu.: 0.000   1st Qu.: 0.000
 Median : 14.00   Median :10.0   Median : 5.000   Median : 0.000
 Mean   : 24.81   Mean   :11.8   Mean   : 7.009   Mean   : 1.155
 3rd Qu.: 31.00   3rd Qu.:16.0   3rd Qu.:10.000   3rd Qu.: 1.000
 Max.   :226.00   Max.   :61.0   Max.   :39.000   Max.   :16.000
   NumSnags        NumDownSnags    PctShrubCover      NumConifer
 Min.   :  0.000   Min.   :  0.00   Min.   :0.0000   Min.   : 0.000
 1st Qu.:  2.000   1st Qu.:  4.00   1st Qu.:0.0000   1st Qu.: 0.000
 Median :  4.000   Median :  8.00   Median :0.1000   Median : 1.000
 Mean   :  7.671   Mean   : 11.41   Mean   :0.1188   Mean   : 4.366
 3rd Qu.:  7.000   3rd Qu.: 12.00   3rd Qu.:0.2000   3rd Qu.: 5.000
 Max.   :135.000   Max.   :127.00   Max.   :0.9000   Max.   :67.000
  StandType
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.4554
 3rd Qu.:1.0000
 Max.   :1.0000
```

Summary:All the data appears to be heavily skewed to the right. For this reason I will perform a log transformation and then see if the transformed data is closer to to a normal distribution.


Nest NumTreelt1in
Histogram
Nest NumTreelt1in
Quantile Plot
Nest NumTreelt1in
Box Plot
Nest NumTree1to3in
Histogram
Nest NumTree1to3in
Quantile Plot
Nest NumTree1to3in
Box Plot
Nest NumTree3to6in
Histogram
Nest NumTree3to6in
Quantile Plot
Nest NumTree3to6in
Box Plot
Nest NumTree6to9in
Histogram
Nest NumTree6to9in
Quantile Plot
Nest NumTree6to9in
Box Plot
Nest NumTree9to15in
Histogram
Nest NumTree9to15in
Quantile Plot
Nest NumTree9to15in
Box Plot
Nest NumTreegt15in
Histogram
Nest NumTreegt15in
Quantile Plot
Nest NumTreegt15in
Box Plot



Transformation Summary: After applying the transformations to my data it is still not perfectly normal but it is much closer to normal than it was before. But you can still see a grouping of data down near zero in many of the variables. PctShrubCover also appears to still be skewed to the right.

(b) Fit LDA and QDA to all the data (with the transformed predictor variables) treating the three birds as a single species. Compare the accuracies or error rates of your classifications using cross-validation.

i. **LDA Confusion Matrix**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 79 19
        1 27 88

               Accuracy : 0.784
                 95% CI : (0.7227, 0.8373)
    No Information Rate : 0.5023
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5679

 Mcnemar's Test P-Value : 0.302

            Sensitivity : 0.7453
            Specificity : 0.8224
         Pos Pred Value : 0.8061
         Neg Pred Value : 0.7652
             Prevalence : 0.4977
         Detection Rate : 0.3709
   Detection Prevalence : 0.4601
      Balanced Accuracy : 0.7839

       'Positive' Class : 0
```

ii. **QDA Confusion Matrix**

```
Confusion Matrix and Statistics
```

```
          actual
predicted  0  1
        0 85 19
        1 21 88

               Accuracy : 0.8122
                 95% CI : (0.7532, 0.8623)
    No Information Rate : 0.5023
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6244

 Mcnemar's Test P-Value : 0.8744

            Sensitivity : 0.8019
            Specificity : 0.8224
         Pos Pred Value : 0.8173
         Neg Pred Value : 0.8073
             Prevalence : 0.4977
         Detection Rate : 0.3991
   Detection Prevalence : 0.4883
      Balanced Accuracy : 0.8122

       'Positive' Class : 0
```

Comparison of LDA to QDA on Nest data: In this instance QDA and LDA perform very similar in their classifications. When I run them through different times the LDA and QDA actually change which one is performing better. The first time I ran the LDA it performed at about 80% While the QDA was at 79%. The Second time I ran the tests LDA dropped to 78.40% and QDA went up to 81.22%. So they appear to be very similar in classifying when all the species are grouped together. Another interesting thing we can see between the two models is they both have a Specificity of 82.24% but they have very different Sensitivity values. QDA comes in about 6% higher than the value for LDA.

(c) For each bird species separately, construct a dataset that comprises the data for all the nest sites for that species, and all the non-nest sites. Now, refit LDA and QDA, for each bird species separately and compare the results for the different methods.

i. **Chickadee LDA and QDA**

Chickadee LDA:

```
[1] "Chickadee & Non-Nest LDA"
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 91 14
        1 15 28

               Accuracy : 0.8041
                 95% CI : (0.7309, 0.8647)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.009482

                  Kappa : 0.5214

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.8585
            Specificity : 0.6667
         Pos Pred Value : 0.8667
         Neg Pred Value : 0.6512
             Prevalence : 0.7162
         Detection Rate : 0.6149
   Detection Prevalence : 0.7095
      Balanced Accuracy : 0.7626

       'Positive' Class : 0
```

Chickadee QDA:

```
[1] "Chickadee & Non-Nest QDA"
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 92 16
        1 14 26

               Accuracy : 0.7973
                 95% CI : (0.7234, 0.8589)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.01576

                  Kappa : 0.4941

 Mcnemar's Test P-Value : 0.85513

            Sensitivity : 0.8679
            Specificity : 0.6190
         Pos Pred Value : 0.8519
         Neg Pred Value : 0.6500
             Prevalence : 0.7162
         Detection Rate : 0.6216
   Detection Prevalence : 0.7297
      Balanced Accuracy : 0.7435

       'Positive' Class : 0
```

Chickadee Summary: Between the LDA and QDA being run on the Chickadee data they appear to be running at around the same accuracy. LDA performs with an accuracy of 80.41% while the QDA has an accuracy of 79.73%. Both LDA and QDA are very close in accuracy on the Chickadee data set but LDA did have a higher accuracy percentage on this data. We can also see that the Sensitivity values are about 1% apart but QDA has higher sensitivity at 86.79% which is interesting since LDA

outperforms QDA in accuracy. For the Specificity, both models are in the 60% area. But LDA outperforms QDA by about 3%.

ii. **Flicker LDA and QDA**

Flicker LDA:

```
[1] "Flicker & Non-Nest LDA"
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 99 12
        1  7 11

               Accuracy : 0.8527
                 95% CI : (0.7796, 0.9089)
    No Information Rate : 0.8217
    P-Value [Acc > NIR] : 0.2129

                  Kappa : 0.4506

 Mcnemar's Test P-Value : 0.3588

            Sensitivity : 0.9340
            Specificity : 0.4783
         Pos Pred Value : 0.8919
         Neg Pred Value : 0.6111
             Prevalence : 0.8217
         Detection Rate : 0.7674
   Detection Prevalence : 0.8605
      Balanced Accuracy : 0.7061

       'Positive' Class : 0
```

Flicker QDA:

```
[1] "Flicker & Non-Nest QDA"
Confusion Matrix and Statistics

          actual
predicted   0   1
        0 104  12
        1   2  11

               Accuracy : 0.8915
                 95% CI : (0.8246, 0.9394)
    No Information Rate : 0.8217
    P-Value [Acc > NIR] : 0.02057

                  Kappa : 0.5536

 Mcnemar's Test P-Value : 0.01616

            Sensitivity : 0.9811
            Specificity : 0.4783
         Pos Pred Value : 0.8966
         Neg Pred Value : 0.8462
             Prevalence : 0.8217
         Detection Rate : 0.8062
   Detection Prevalence : 0.8992
      Balanced Accuracy : 0.7297

       'Positive' Class : 0
```

Flicker Summary: For the Flicker dataset the QDA does appear to perform a littler better than the LDA. Here the LDA has an accuracy of 85.27% while QDA has an accuracy of 89.15%. So the QDA has accuracy in this instance that is around 4% better than the LDA. Another thing we can see about the models is they both have the same Specificity score at 47.83% which is not good at all. But they also both score high with Sensitivity. QDA comes in at 98.11% while LDA comes in at 93.40%.

iii. **Sapsucker LDA and QDA**

Sapsucker LDA:

```
[1] "Sapsucker & Non-Nest LDA"
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 91 12
        1 15 30

               Accuracy : 0.8176
                 95% CI : (0.7458, 0.8762)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.003047

                  Kappa : 0.5607

 Mcnemar's Test P-Value : 0.700311

            Sensitivity : 0.8585
            Specificity : 0.7143
         Pos Pred Value : 0.8835
         Neg Pred Value : 0.6667
             Prevalence : 0.7162
         Detection Rate : 0.6149
   Detection Prevalence : 0.6959
      Balanced Accuracy : 0.7864

       'Positive' Class : 0
```

Sapsucker QDA:

```
[1] "Sapsucker & Non-Nest QDA"
Confusion Matrix and Statistics
```

```
          actual
predicted  0  1
        0 92 15
        1 14 27

               Accuracy : 0.8041
                 95% CI : (0.7309, 0.8647)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.009482

                  Kappa : 0.5145

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.8679
            Specificity : 0.6429
         Pos Pred Value : 0.8598
         Neg Pred Value : 0.6585
             Prevalence : 0.7162
         Detection Rate : 0.6216
   Detection Prevalence : 0.7230
      Balanced Accuracy : 0.7554

       'Positive' Class : 0
```

Sapsucker Summary: Here again LDA and QDA are very close in comparison. The LDA for this iteration is slightly better at 81.76% while QDA is at 80.41% accuracy. But again they are so close to one another that they are very comparable in performance. But in this instance LDA did out perform QDA in Accuracy. When we look at sensitivity we can see that QDA does better than LDA at 86.79% coming in about 1% above LDA. LDA outperforms QDA though in Specificity, it comes in about 7% higher at 71.43%.

(iii) **Question 3:**

(a) Fit a logistic regression model with all the data. Compare the cross-validated accuracies or error rates with those for LDA and QDA that you obtained in the previous question.

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 81 19
        1 25 88

               Accuracy : 0.7934
                 95% CI : (0.7328, 0.8457)
    No Information Rate : 0.5023
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5867

 Mcnemar's Test P-Value : 0.451

            Sensitivity : 0.7642
            Specificity : 0.8224
         Pos Pred Value : 0.8100
         Neg Pred Value : 0.7788
             Prevalence : 0.4977
         Detection Rate : 0.3803
   Detection Prevalence : 0.4695
      Balanced Accuracy : 0.7933

       'Positive' Class : 0
```

Comparison between question 2 and 3 results: The accuracy rate for the cross validated logistic regression model on all the data is 79.34%. While that for LDA on all the data was 78.40% and QDA had an accuracy of 81.22%. So for the given iteration it appears that QDA performed the best

but again the accuracy between the three methods are very close together. Looking at the Sensitivity between the three models, LDA and the Logistic Regression perform the same at 74.53%, QDA does the best at 80.19%. Looking at the Specificity LDA and QDA tie for the best performance at 82.24%, Logistic Regression comes in at 81.31%

(b) Now apply some variable selection procedure (in logistic regression) and identify variables important to the classification. By how much did the cross-validated accuracies/error rates change?

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 79 20
        1 27 87

               Accuracy : 0.7793
                 95% CI : (0.7176, 0.8331)
    No Information Rate : 0.5023
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5585

 Mcnemar's Test P-Value : 0.3815

            Sensitivity : 0.7453
            Specificity : 0.8131
         Pos Pred Value : 0.7980
         Neg Pred Value : 0.7632
             Prevalence : 0.4977
         Detection Rate : 0.3709
   Detection Prevalence : 0.4648
      Balanced Accuracy : 0.7792

       'Positive' Class : 0
```

Summary Variable Selection: Using the variable selection the accuracy of the model increased from 78.4% up to 79.34%. So the accuracy did increase but it did not increase significantly. The variables used in the final model were, "NumTree9to15in", "NumTree6to9in", "NumConifer", "NumDownSnags", "NumTreelt1in" and "NumTree3to6in"

(c) Repeat part a. using the datasets for the individual bird species.

i. **Chickadee Cross Validated Logistic Regression**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 93 17
        1 13 25

               Accuracy : 0.7973
                 95% CI : (0.7234, 0.8589)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.01576

                  Kappa : 0.4866

 Mcnemar's Test P-Value : 0.58388

            Sensitivity : 0.8774
            Specificity : 0.5952
         Pos Pred Value : 0.8455
         Neg Pred Value : 0.6579
             Prevalence : 0.7162
         Detection Rate : 0.6284
   Detection Prevalence : 0.7432
      Balanced Accuracy : 0.7363

       'Positive' Class : 0
```

Comparison of Chickadee Logsitic Regression vs. LDA and QDA:
Logistic Regression has an accuracy of 79.06% on the Chickadee data set, while on the same data set LDA got an accuracy of 80.41% and QDA got an accuracy of 79.73%. So in this case LDA and QDA both outperformed logistic regression, with LDA performing the best out of all the models. Looking at the Sensitivity Logistic Regression scored the best at 87.74% Then QDA got a score of 86.79% and QDA got a score of 85.85%. All the models performed poorly in the Specificity score. LDA was the highest at 66.67% then QDA and logistic regression tied at 61.90%.

ii. **Flicker Cross Validated Logistic Regression**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 99 15
        1  7  8

               Accuracy : 0.8295
                 95% CI : (0.7533, 0.8899)
    No Information Rate : 0.8217
    P-Value [Acc > NIR] : 0.4640

                  Kappa : 0.3262

 Mcnemar's Test P-Value : 0.1356

            Sensitivity : 0.9340
            Specificity : 0.3478
         Pos Pred Value : 0.8684
         Neg Pred Value : 0.5333
             Prevalence : 0.8217
         Detection Rate : 0.7674
   Detection Prevalence : 0.8837
      Balanced Accuracy : 0.6409
```

```
       'Positive' Class : 0
```

Comparison of Flicker Logsitic Regression vs. LDA and QDA: Flicker LDA had an accuracy of 85.27% while QDA had an accuracy of 89.15% and logistic regression had an accuracy of 82.17%. So again both LDA and QDA outperform logistic regression with their accuracy percentage. But in this instance QDA outperformed the other two models. Looking at the sensitivy QDA also scored the best coming in 98.11% while LDA came in at 93.40% and Logistic Regession came in at 92.45%. For specificity LDA and QDA tied at 47.83% both outperforming logistic regression which came in at 30.43%.

iii. **Sapsucker Cross Validated Logistic Regression**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 90 13
        1 16 29

               Accuracy : 0.8041
                 95% CI : (0.7309, 0.8647)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.009482

                  Kappa : 0.5281

 Mcnemar's Test P-Value : 0.710347

            Sensitivity : 0.8491
            Specificity : 0.6905
         Pos Pred Value : 0.8738
         Neg Pred Value : 0.6444
             Prevalence : 0.7162
         Detection Rate : 0.6081
```

```
   Detection Prevalence : 0.6959
      Balanced Accuracy : 0.7698

       'Positive' Class : 0
```

Comparison of Sapsucker Logsitic Regression vs. LDA and QDA:

Sapsucker LDA had an accuracy of 81.76% while QDA had an accuracy of 80.41% and logistic regression got an accuracy percentage of 80.41%. In this instance logistic regression performed just as well as QDA. In the end LDA was the best performing model with an accuracy of 81.76%. But all of these models accuracy scores are very close together. Looking at Sensitivity Logistic Regression tied with QDA for best performance at 86.79% LDA got a score of 85.85%. For Specificity Logistic Regression tied with LDA this time for the highest score at 71.43% while QDA got a score of 64.29%.

(d) Repeat part b. using the datasets for the individual bird species. Are there variables that are in the models for 2 or 3 of the species?

i. **Chickadee Variable Selection and Confusion Matrix:**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 91 20
        1 15 22

               Accuracy : 0.7635
                 95% CI : (0.6868, 0.8294)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.1168

                  Kappa : 0.3966

 Mcnemar's Test P-Value : 0.4990
```

```
            Sensitivity : 0.8585
            Specificity : 0.5238
         Pos Pred Value : 0.8198
         Neg Pred Value : 0.5946
             Prevalence : 0.7162
         Detection Rate : 0.6149
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.6912

       'Positive' Class : 0
```

ii. **Flicker Variable Selection and Confusion Matrix:**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 94 17
        1 12 25

               Accuracy : 0.8041
                 95% CI : (0.7309, 0.8647)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.009482

                  Kappa : 0.5

 Mcnemar's Test P-Value : 0.457614

            Sensitivity : 0.8868
            Specificity : 0.5952
         Pos Pred Value : 0.8468
         Neg Pred Value : 0.6757
             Prevalence : 0.7162
         Detection Rate : 0.6351
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.7410
```

```
       'Positive' Class : 0
```

iii. **Sapsucker Variable Selection and Confusion Matrix:**

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 94 12
        1 12 30

               Accuracy : 0.8378
                 95% CI : (0.7684, 0.8933)
    No Information Rate : 0.7162
    P-Value [Acc > NIR] : 0.0004038

                  Kappa : 0.6011

 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 0.8868
            Specificity : 0.7143
         Pos Pred Value : 0.8868
         Neg Pred Value : 0.7143
             Prevalence : 0.7162
         Detection Rate : 0.6351
   Detection Prevalence : 0.7162
      Balanced Accuracy : 0.8005

       'Positive' Class : 0
```

Final Comparison of the Species Data sets:

- Chickadee Logistic Regression Accuracy = 80.41%; Chickadee Logistic Regression Accuracy with Variable selection = 79.05%; In this instance the variable selection decreased the accuracy score from

80.41% down to 79.05%. An decrease of 1.34%. The sensitivity stayed the same at 89.62%. While the specificity decreased from 57.14% down to 52.38%.

- Flicker Logistic Regression Accuracy = 81.4%; Flicker Logistic Regression Accuracy with variable selection = 76.35% ; In this instance the variable selection decreased the accuracy of the model from 81.4% down to 76.35%. Sensitivity decreased from 92.45% down to 86.79%. While Specificity increased from 30.43% up to 50%.
- Sapsucker Logistic Regression Accuracy = 82.43%; Sapsucker Logistic Regression Accuracy with variable selection = 84.46% In this instance the variable selection increased the accuracy of the model about 2%. The sensitivity also increased going from 86.79% up to 88.68%. Specificity also increased from 71.43% up to 73.81%.
- "NumTree9to15in", "NumConifer" appears in both the Flicker data set and the Chickadee data set. "NumTreelt1in", "NumTree3to6in" appears in all 3 of the data sets. So we can assume that these are the most important variables that need to be used in the prediction of the models.