

# Classification of Cherry and Pear Leaves by Length and Width

*Kaisa Roggeveen, Scott Graham*

*March 22nd 2018*

Dr. Steven M. Vamosi  
Associate Dean, Diversity, Equity and Inclusion  
Professor, Population Biology  
2500 University Drive NW  
Department of Biological Sciences  
University of Calgary  
Calgary AB  
T2N 1N4 Canada

## Introduction

The intent of this paper is to develop a method for classifying leaves as either Cherry or Pear, based on their measured length and width. This method was developed for Dr. Steven Vamosi, a botanist from the University of Calgary.

The classification method used was Linear Discriminant Analysis (LDA), developed by R.A Fischer. In order to ensure models have strong predictive power, Leave One Out Cross Validation (LOOCV) was used for all the models described in this paper. In order to determine the accuracy and usefulness of our models we used LOOCV to compare a predicted classification vs. the actual classification in order to determine if there are any misclassifications.

Cherry and Pear leaves are both leaves from fruit trees. Cherry trees belong to the genus *Prunus* and Pear trees belong to the genus *Pyrus* [2],[3]. A common feature amongst the leaves is that they both have a midrib, which is the central vein of the leaf which extends along the leaf's center line.

## Data

### Measurement Process

The first step taken in the measurement of the leaves was to give each leaf an identification number based on the species. The method used to measure the dimensions was to create a box with the minimum length and width in which the entire leaf would be encompassed in the box.

To begin creating the sides of the box, a ruler was aligned parallel to the midrib, which is the central vein in the leaf and moved towards the left and the right of the picture until only one point on the leaf remained [1]. From the single point on the side of the leaf, a line was drawn parallel to the midrib of the leaf.

Next, the base and point of the leaf were measured, a ruler was placed perpendicular to the midrib and the ruler was moved towards to tip of the leaf until a single point remained, a line was draw perpendicular to the

midrib at this point. At the base of the leaves the length of the leaf was set as the point where the leaf ends and the stem begins, at this point a line was drawn perpendicular to the midrib.

After all the boxes were created, the width (lines parallel to midrib) and the length (lines perpendicular to midrib) were measured and the results were recorded in a spread sheet.

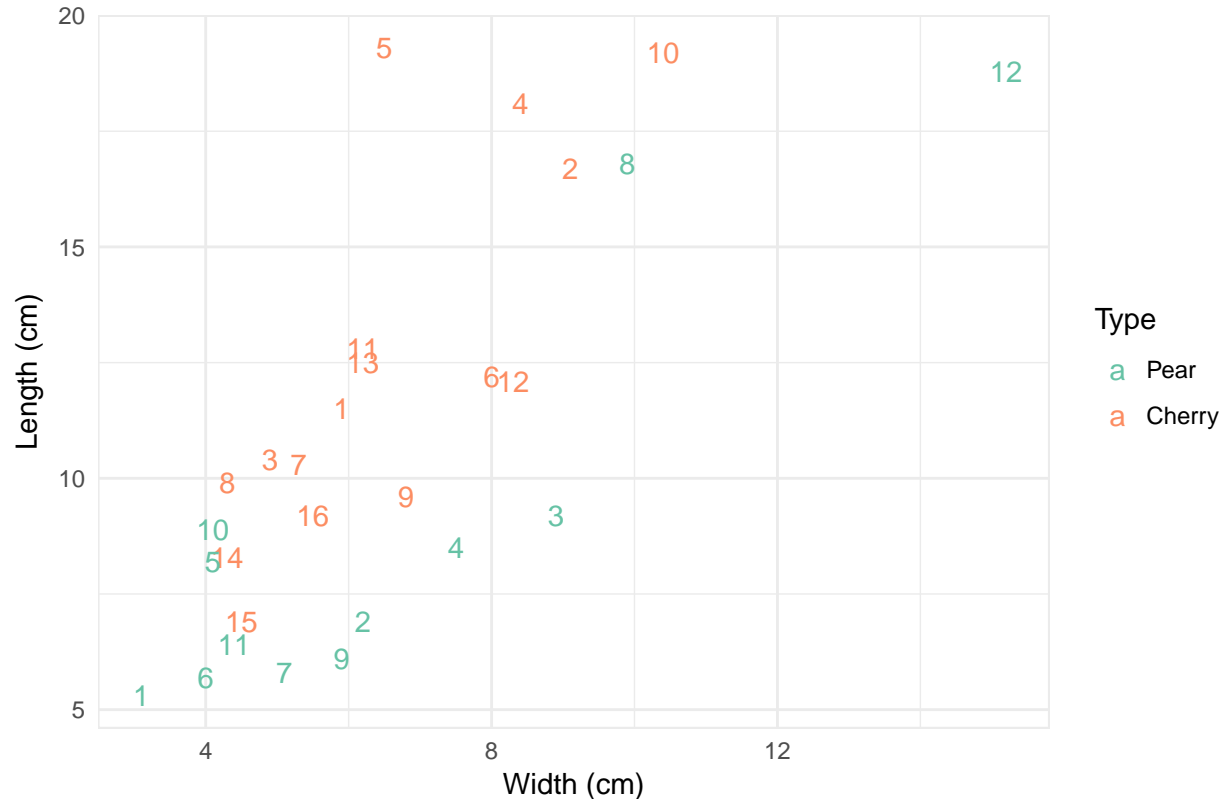
## Data Creation

Table 1: Data Summary

Type	Length	Width
Pear :12	Min. : 5.300	Min. : 3.100
Cherry:16	1st Qu.: 7.875	1st Qu.: 4.475
NA	Median : 9.750	Median : 6.050
NA	Mean :10.914	Mean : 6.536
NA	3rd Qu.:12.575	3rd Qu.: 8.075
NA	Max. :19.300	Max. :15.200

In this original data set there are a few issues that need to be acknowledged. The first issues that occurred during the data measurements was the result of the leaves that were distributed as the training sample were images, in which the images were not to scale. This resulted in a few outlines, which much larger lengths and widths compared to the other leaves in the set. These outlines included Pear#12, Cherry#10 and Cherry#5. However, based on the nature of this project in just observing the ratio between the length and width, this should not be affected by the size of the image, unless the image was stretched in either direction.

Figure 01: Length vs Width Scatter Plot



In Figure 01 there is a distinct separation in the data of the cherry and the pear. As mentioned above, the outlines are Cherry#10, Cherry#5 and Pear#12 these outlines appear to follow a similar grouping and therefore they were kept in the data set. The raw data is located in Appendix A.

## Classification Procedure (LDA)

### Training Data

Table 2: LDA Prior Probabilities

Type	Probability
Pear	0.4286
Cherry	0.5714

Table 3: LDA Group Means

	Length	Width
Pear	8.8833	6.5333
Cherry	12.4375	6.5375

Table 4: LDA Coefficients of Linear Discriminants

Dimension	Coefficient
Length	0.4194
Width	-0.5311

Tables 2-4 represent the output of a LDA done on the raw data. Table 2 represents the prior probabilities of falling in a particular type. The prior probability of being a Pear leaf is 0.4286. The prior probability of being a Cherry leaf is 0.5714.

The LDA coefficients act similar to those in a ordinary least squares model/regression model, where the sum product of those with a row of data gives us a score for that leaf. That score is then transformed into an estimated probability for both types of leaves, using Bayesian methodology. The LDA coefficient for Length is 0.4193533, and for Width it is -0.5310692.

Table 5: LDA Misclassification Results

Predicted	Actual	Length	Width	Cherry Probability	Pear Probability	Correct Prediction
Pear	Cherry	9.6	6.8	0.3582	0.6418	FALSE
Pear	Cherry	12.1	8.3	0.4481	0.5519	FALSE
Pear	Cherry	6.9	4.5	0.3892	0.6108	FALSE
Cherry	Pear	8.2	4.1	0.6631	0.3369	FALSE
Cherry	Pear	16.8	9.9	0.8116	0.1884	FALSE
Cherry	Pear	8.9	4.1	0.7528	0.2472	FALSE

Table 6: LDA Confusion Matrix

Prediction	Pear	Cherry
Pear	9	3
Cherry	3	13

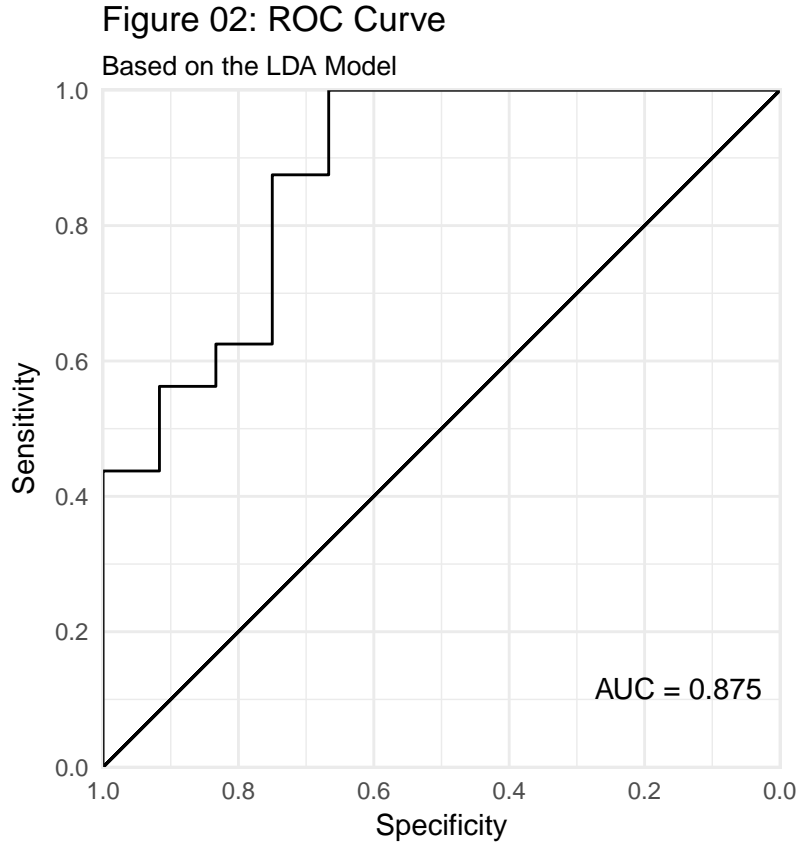
Table 7: LDA Confusion Matrix Stats

	x
Sensitivity	0.7500000
Specificity	0.8125000
Pos Pred Value	0.7500000
Neg Pred Value	0.8125000
Precision	0.7500000
Recall	0.7500000
F1	0.7500000
Prevalence	0.4285714
Detection Rate	0.3214286
Detection Prevalence	0.4285714
Balanced Accuracy	0.7812500

Tables 5-6 are the results from the LDA, in this model six leaves were misclassified, which included three pear and three cherry. Upon examining where these leaves are situated in the scatter plot these leaves are along the boundary lines.

In the confusion matrix, out of the 12 pear leaves 9 were classified correctly and 3 were misclassified and out of the 16 cherry leaves 13 were classified correctly and 3 were misclassified.

The sensitivity represents the proportion of predicted pear leaves that were actually pear leaves, which was 0.75. The specificity represents the proportion of predicted cherry leaves that were actually cherry leaves, which was 0.8125.



The Receiver Operating Characteristic (ROC) Curve represents the matched pairs of Specificity and Sensitivity at different threshold levels. What this means is that for a given data point we can assign a leaf type based on the estimated probability. To do this, we choose a threshold for this assignment, for example anything with a probability of being a cherry leaf of  $\geq 0.60$  we would assign a predicted type of cherry, with a threshold of 60%. To determine the best threshold to use in terms of maximizing both the specificity and sensitivity, we find the point on the ROC curve which is furthest away from the line splitting the graph along the diagonal. The best threshold to use as determined by the ROC Curve is 0.3468.

It is important to have an Area Under the Curve (AUC) that approaches one, as this ensures that the ROC curve approaches 1 for both specificity and sensitivity. For the LDA, we have a AUC of 0.875, which is pretty good.

## New Data

Table 8: LDA New Data Predictions

Predicted	Cherry	Pear	Number	Length	Width
Cherry	0.8003229	0.1996771	1	8.2	3.2
Pear	0.2772400	0.7227600	2	5.2	3.8
Cherry	0.5942515	0.4057485	3	7.6	4.0

Table 8 represents the predicted lead type based on data that was not originally included in the data set.

## Observation Space

Figure 03a: Length vs Width Scatter Plot

Overlaid with the Convex Hull Based on the LDA Predicted Type

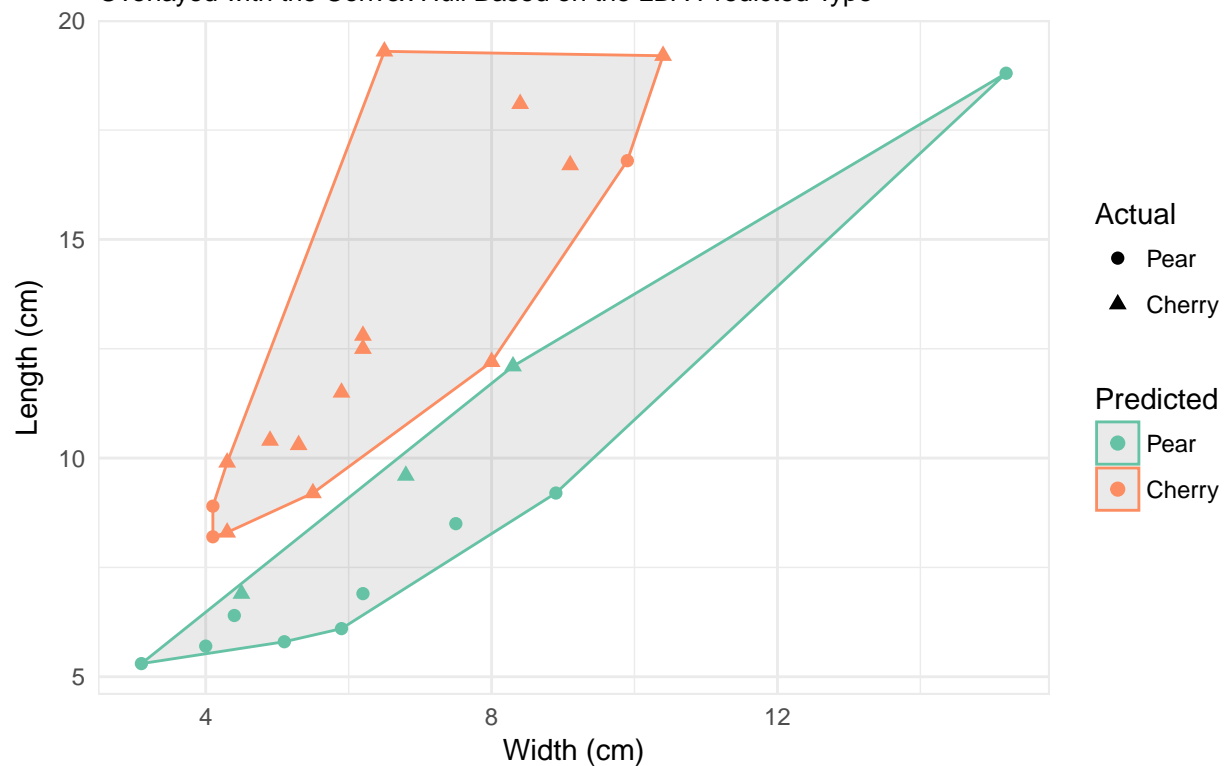
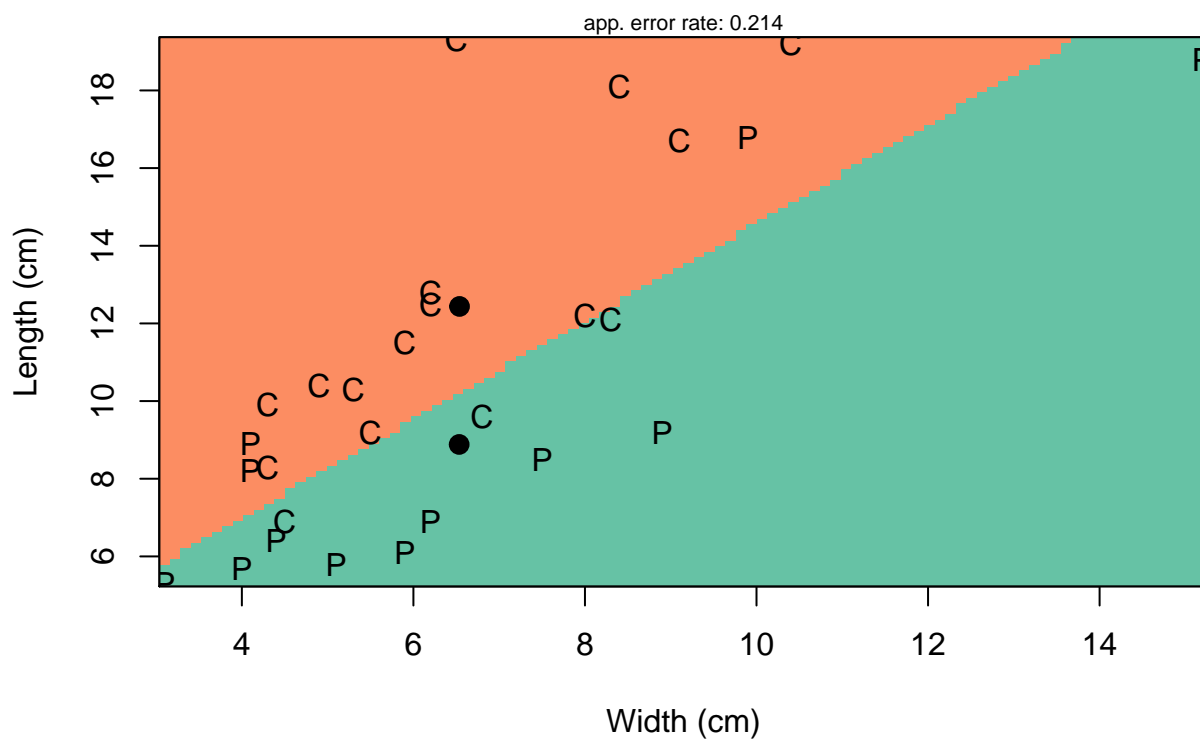


Figure 03b: LDA Partition Plot

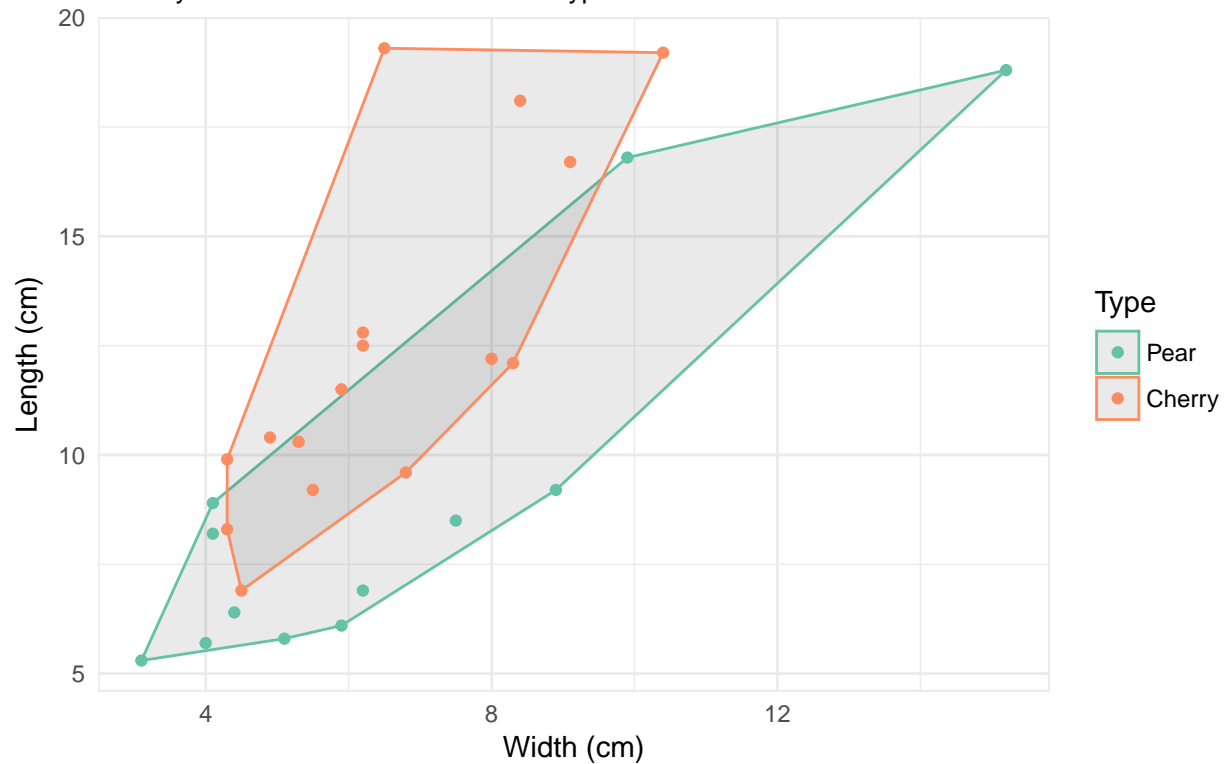


The convex hull in Figure 03a represents the region that captures all the points of a given leaf type, and is convex in nature. The convexity ensures that any linear combination of points in the set is still in the set. We can evaluate this on the predicted types given by the LDA to see the separating hyper plane between the two convex sets, as this gives a good approximation of the line used to differentiate between the types by the LDA. As well, by looking at the combination of shape and colour, we can see which points were misclassified as per Table 3.

Figure 03b shows the classification line used by the LDA. As expected it is a linear classification rule, which splits the observation space into two half spaces.

**Figure 04: Length vs Width Scatter Plot**

Overlaid with the Convex Hull of that Type



This is the convex hull of the raw data, and as you can see there is an overlap which indicates that there is no strict separation in the raw data and therefore it was necessary to conduct the LDA.

# Probability Distributions

## Contour

Figure 05: Length vs Width Scatter Plot

Overlaid with the Contour Plot



Contour plots show the clustering of data for pear and cherry trees. Each contour line represents the same density anywhere along that line. As the contour line density increases so does the steepness of the graph and the probability that a given leaf will have those characteristics.

In Figure 05, there are two distinct shapes for the contour plot of pear and cherry leaves, this is due to the different covariance matrices of the leaves as seen in Table 10 and 11. The pear leaves tended to have a more similar length and width whereas, the cherry leaves tended to have a longer length and a skinnier width.



Figure 06: Length vs Width Scatter Plot

Overlaid with a Contour Plot of that Type

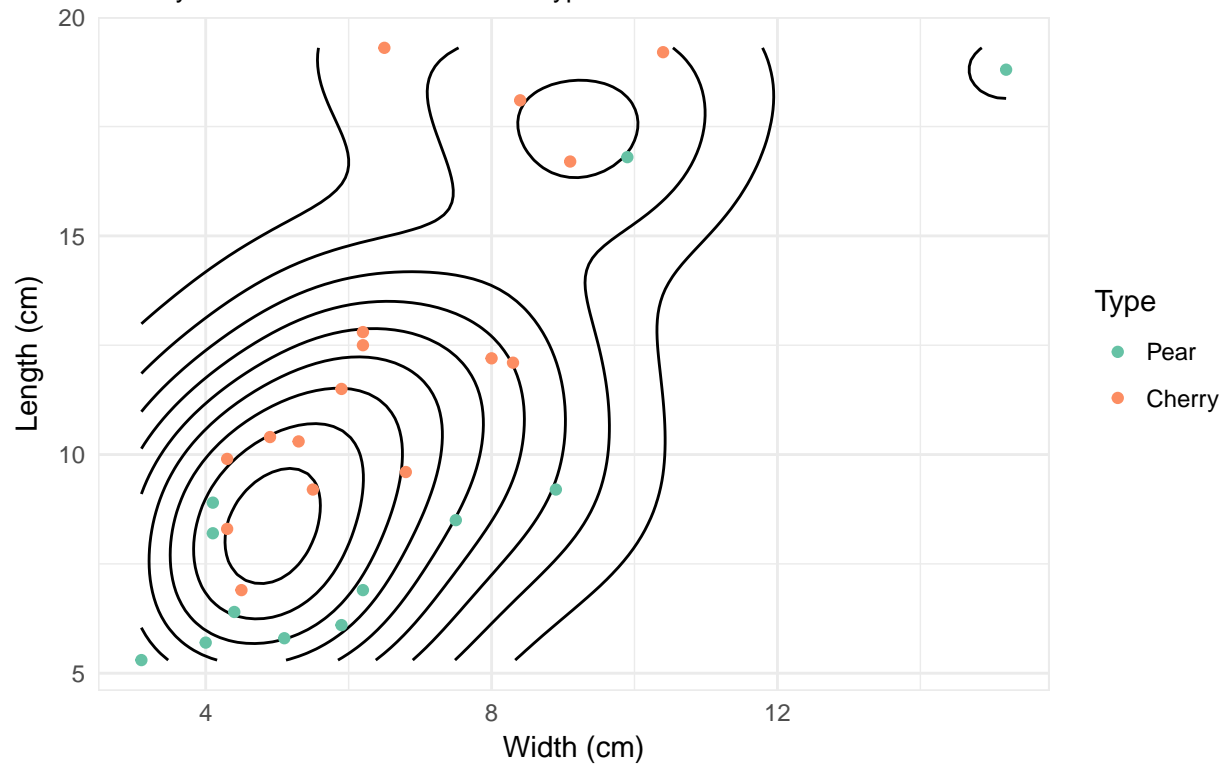


Figure 6 is the combined contour plot of the raw data. In this contour plot there is potential bimodality, as seen with the two peaks in the contour plot. However, these peaks fall along the diagonal and not along the vertical or horizontal axes, which indicates that the bimodality is shared between the length and the width.

Figure 07: Density Plot by Type

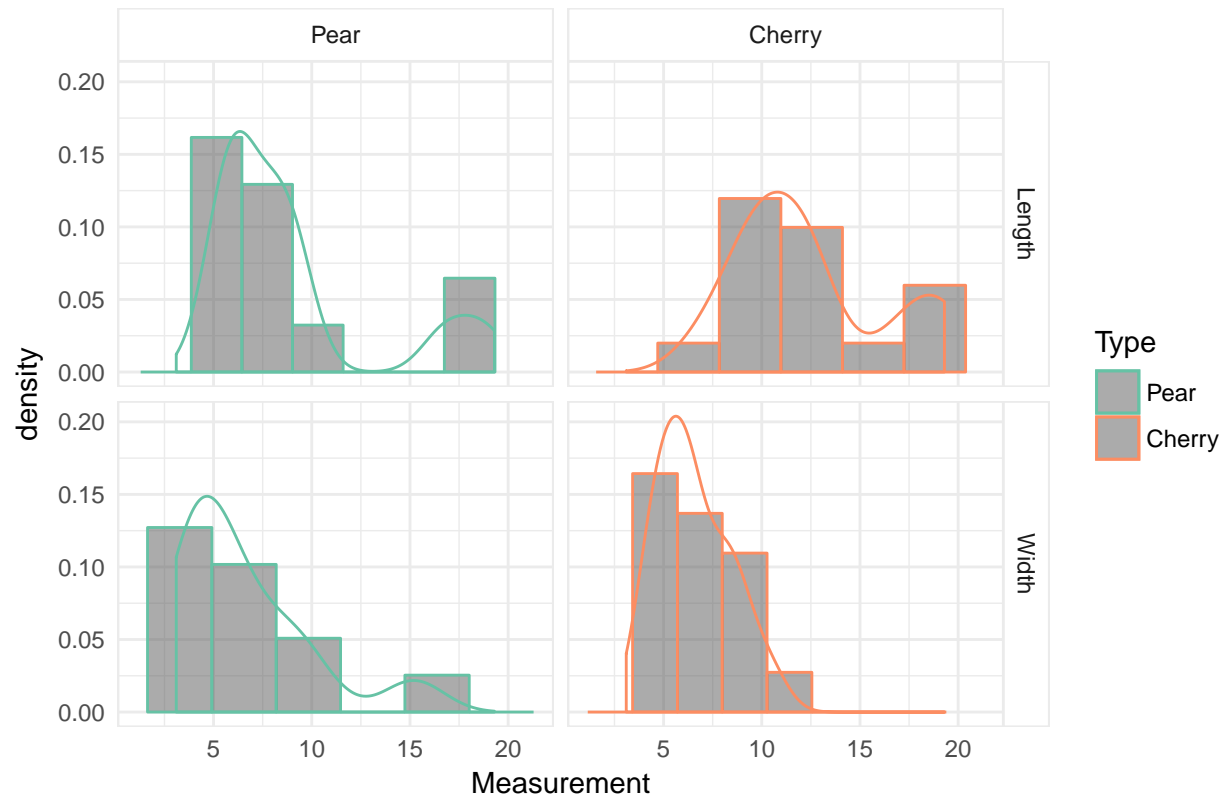


Figure 08: Density Plot

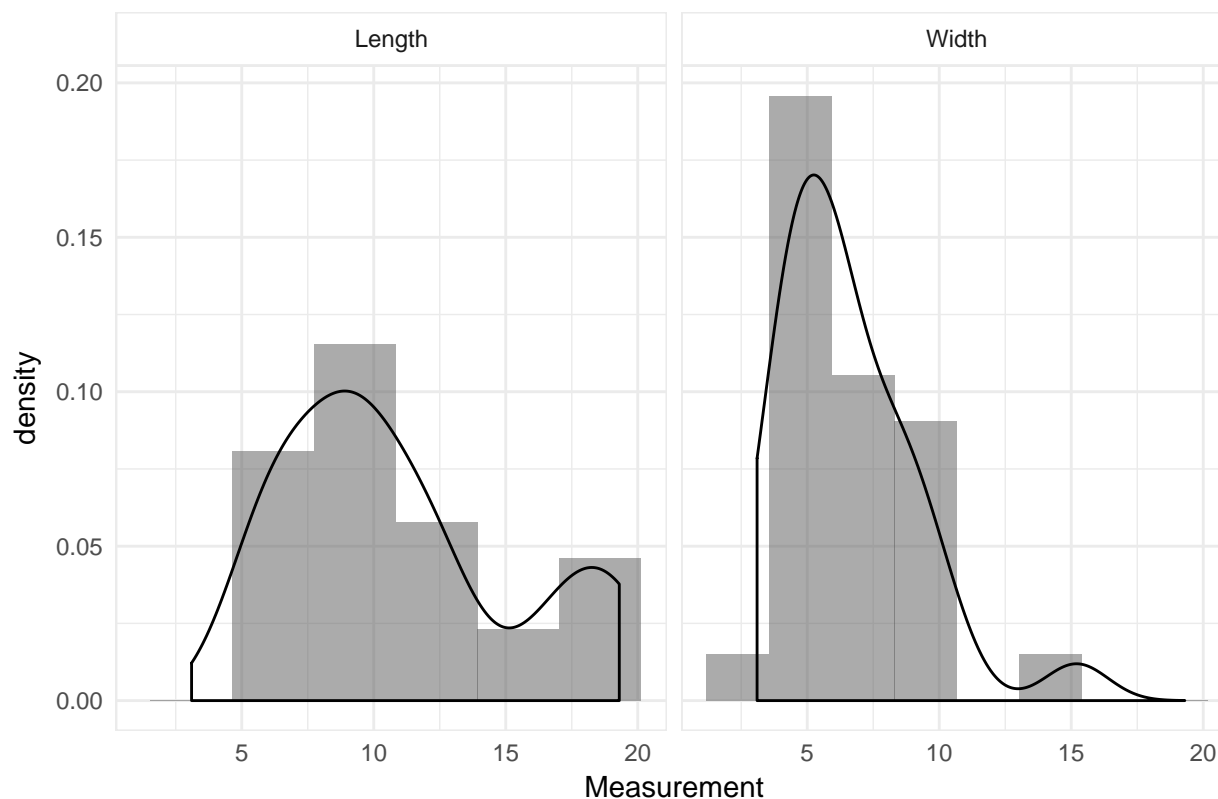


Figure 8 shows the bimodality of the data for both the length and the width. The bimodality of the data is more prominent in the length than in the width, which becomes especially apparent when you split it up by type.

## Covariance Matrix

Table 9: Shared Covariance Matrix

	Length	Width
Length	19.422011	8.476508
Width	8.476508	6.685344

Table 10: Cherry Covariance Matrix

	Length	Width
Length	15.047833	5.443833
Width	5.443833	3.357167

Table 11: Pear Covariance Matrix

	Length	Width
Length	19.27788	13.37333

	Length	Width
Width	13.37333	11.83152

In order to get an idea of how different the split by type covariance matrices are, we calculated the average distance between entries. This is effectively the  $\ell_2$ -norm divided by the number of entries in the covariance matrix.

$$\text{Average Distance} = \frac{\|\Sigma_{\text{Cherry}} - \Sigma_{\text{Pear}}\|_2}{r \times c} = \frac{\sqrt{\sum_{i=1}^2 \sum_{j=1}^2 (\sigma_{\text{Cherry},i,j} - \sigma_{\text{Pear},i,j})^2}}{4}$$

Which comes out to 3.6697. We can also compute this by comparing the shared and cherry covariance matrix, giving us 1.7429, and for the shared and pear covariance matrix, giving us 2.1573. As one would expect, the difference between the individual covariance matrices and the shared one is smaller than the difference between the two individual covariance matrices, due to the pooling property of the shared matrix. Because of this, it is safe to assume that the shared covariance matrix accurately estimates the individual covariance matrix for the two types of leaves.

## Classification Procedure (QDA)

The difference between LDA and Quadratic Discriminant Analysis (QDA) is that QDA doesn't rely on the assumption that both classes of data share a covariance matrix, which is a crucial assumption in LDA. This allows us to perform analysis on data where this assumption may not hold in exchange for an increased variance. As well, it doesn't require the classification rule to be linear, but instead can be a quadratic function.

## Training Data

Table 12: QDA Prior Probabilities

Type	Probability
Pear	0.4286
Cherry	0.5714

Table 13: QDA Group Means

	Length	Width
Pear	8.8833	6.5333
Cherry	12.4375	6.5375

The prior probability of being a Pear leaf is 0.4286. The prior probability of being a Cherry leaf is 0.5714.

Table 14: QDA Misclassification Results

Predicted	Actual	Length	Width	Cherry Probability	Pear Probability	Correct Prediction
Pear	Cherry	9.6	6.8	0.4686	0.5314	FALSE
Pear	Cherry	12.1	8.3	0.4363	0.5637	FALSE

Predicted	Actual	Length	Width	Cherry Probability	Pear Probability	Correct Prediction
Pear	Cherry	6.9	4.5	0.4710	0.5290	FALSE
Cherry	Pear	8.2	4.1	0.6431	0.3569	FALSE
Cherry	Pear	16.8	9.9	0.8095	0.1905	FALSE
Cherry	Pear	8.9	4.1	0.7287	0.2713	FALSE

Table 15: QDA Confusion Matrix

Prediction	Pear	Cherry
Pear	9	3
Cherry	3	13

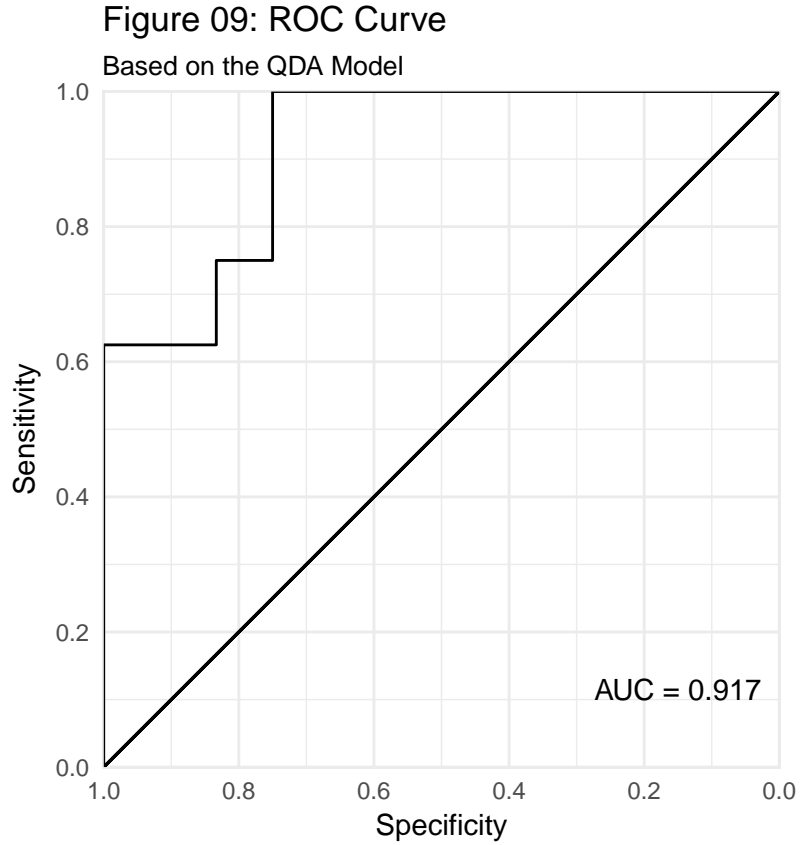
Table 16: QDA Confusion Matrix Stats

	x
Sensitivity	0.7500000
Specificity	0.8125000
Pos Pred Value	0.7500000
Neg Pred Value	0.8125000
Precision	0.7500000
Recall	0.7500000
F1	0.7500000
Prevalence	0.4285714
Detection Rate	0.3214286
Detection Prevalence	0.4285714
Balanced Accuracy	0.7812500

Tables 12-14 are the results from the QDA, in this model six leaves were misclassified, which included three pear and three cherry.

In the confusion matrix, out of the 12 pear leaves 9 were classified correctly and 3 were misclassified and out of the 16 cherry leaves 13 were classified correctly and 3 were misclassified.

The sensitivity represents the proportion of predicted pear leaves that were actually pear leaves, which was 0.75. The specificity represents the proportion of predicted cherry leaves that were actually cherry leaves, which was 0.8125.



For the QDA, we have a AUC of 0.9167, which is better than our LDA model. The best threshold to use as determined by the ROC Curve is 0.4306.

## New Data

Table 17: QDA New Data Predictions

Predicted	Cherry	Pear	Number	Length	Width
Cherry	0.6362745	0.3637255	1	8.2	3.2
Pear	0.3382849	0.6617151	2	5.2	3.8
Cherry	0.5712093	0.4287907	3	7.6	4.0

Table 17 represents the predicted leaf type based on data that was not originally included in the data set. These results are similar to what was seen in the LDA model.

Observation Space

Figure 10a: Length vs Width Scatter Plot

Overlaid with the Convex Hull Based on the QDA Predicted Type

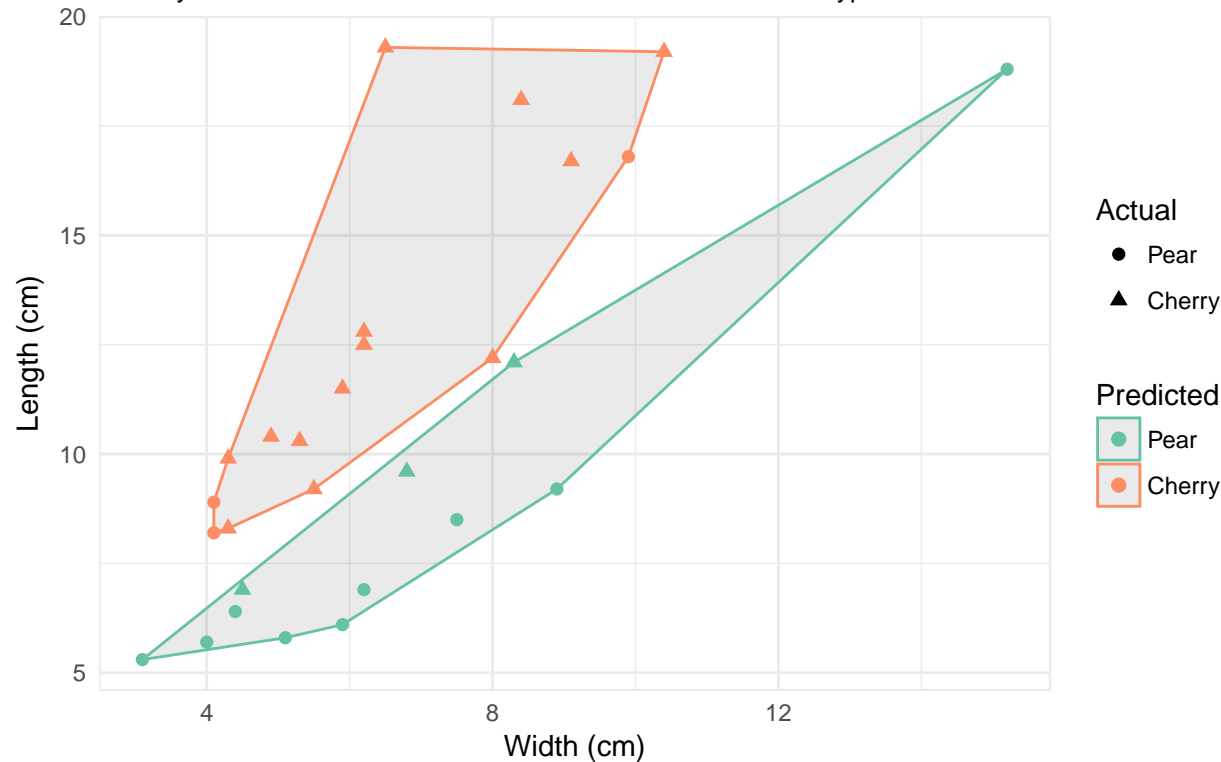
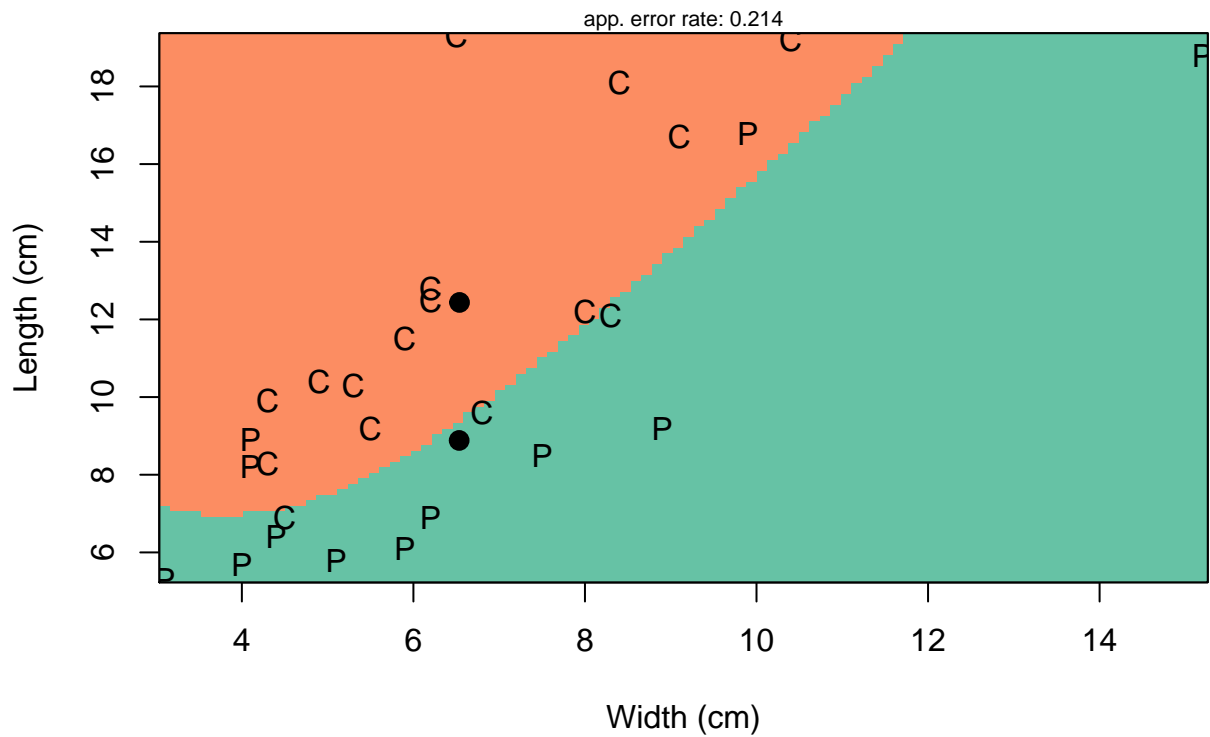


Figure 10b: QDA Partition Plot



The convex hulls in Figure 10a look identical to that described in the LDA section, due to the fact that the classification was identical between the two models. However when looking at Figure 10b, we can see that the model is using a curved or quadratic classification rule, which doesn't allow us to split the observation space into two half spaces, due to a lack of linearity.

## Classification Procedure (GLM)

The final classification method we looked at was using a cross-validated logistic regression model to determine the predicted type of leaf. This is based on a regression of the predicted log-odds of falling particular class on the predictors of length and width.

### Training Data

	Estimate	Std. Error	z value	Pr(> z )
<b>(Intercept)</b>	-1.74	1.519	-1.145	0.2522
<b>Length</b>	0.7764	0.2875	2.7	0.006931
<b>Width</b>	-0.9338	0.3815	-2.448	0.01438

(Dispersion parameter for binomial family taken to be 1 )

Null deviance:	38.24 on 27 degrees of freedom
Residual deviance:	24.26 on 25 degrees of freedom

Through the usage of a LOOCV Logistic regression, we derived the following model:

$$\text{Logit}(P(\text{Cherry}|L, W)) = -1.74 + 0.7764L - 0.9338W + \epsilon$$

Both of the slope variables were statistically significant at  $\alpha = 0.05$ . From this, the estimated odds in favor of being a cherry leaf vs. a pear leaf increases by a multiplicative factor of 2.1736 for every centimeter increase in length. The estimated odds in favor of being a cherry leaf vs. a pear leaf increases by a multiplicative factor of 0.393 for every centimeter increase in width.

Table 20: Logit Misclassification Results

Predicted	Actual	Length	Width	Cherry Probability	Pear Probability	Correct Prediction
Pear	Cherry	9.6	6.8	0.3460	0.6540	FALSE
Pear	Cherry	12.1	8.3	0.4759	0.5241	FALSE
Pear	Cherry	6.9	4.5	0.3578	0.6422	FALSE
Cherry	Pear	8.2	4.1	0.6895	0.3105	FALSE
Cherry	Pear	16.8	9.9	0.8868	0.1132	FALSE
Cherry	Pear	8.9	4.1	0.7927	0.2073	FALSE

Table 21: Logit Confusion Matrix

Prediction	Pear	Cherry
Pear	9	3



Prediction	Pear	Cherry
Cherry	3	13

Table 22: Logit Confusion Matrix Stats

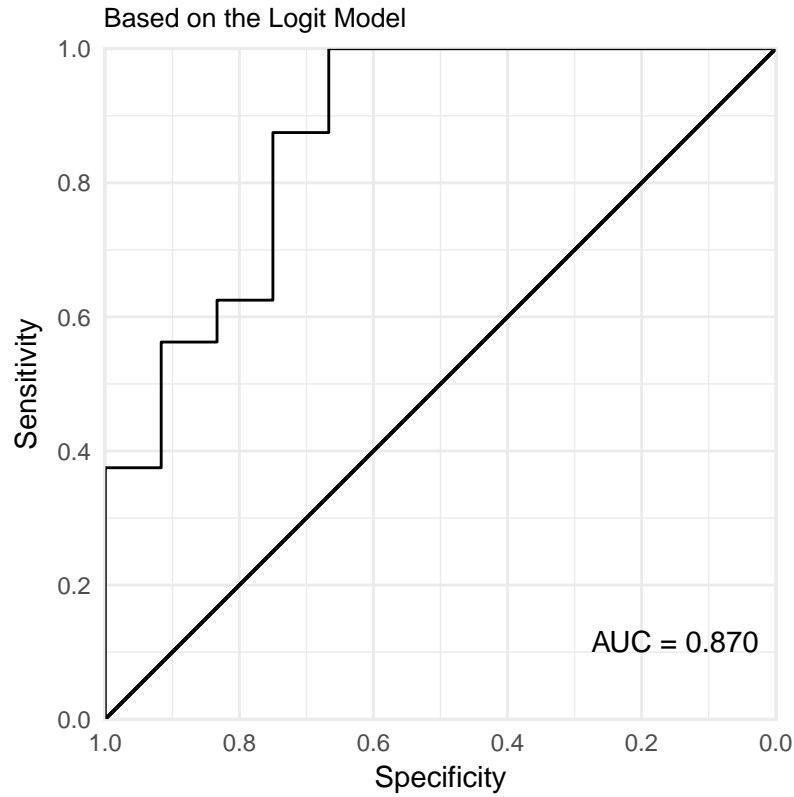
	x
Sensitivity	0.7500000
Specificity	0.8125000
Pos Pred Value	0.7500000
Neg Pred Value	0.8125000
Precision	0.7500000
Recall	0.7500000
F1	0.7500000
Prevalence	0.4285714
Detection Rate	0.3214286
Detection Prevalence	0.4285714
Balanced Accuracy	0.7812500

Tables 20-22 are the results from the logistic regression, in this model six leaves were misclassified, which included three pear and three cherry.

In the confusion matrix, out of the 12 pear leaves 9 were classified correctly and 3 were misclassified and out of the 16 cherry leaves 13 were classified correctly and 3 were misclassified.

The sensitivity represents the proportion of predicted pear leaves that were actually pear leaves, which was 0.75. The specificity represents the proportion of predicted cherry leaves that were actually cherry leaves, which was 0.8125.

Figure 11: ROC Curve



For the logistic regression, we have a AUC of 0.8698, which is worse than our LDA model. The best threshold to use as determined by the ROC Curve is 0.3196.

## New Data

Table 23: Logit New Data Predictions

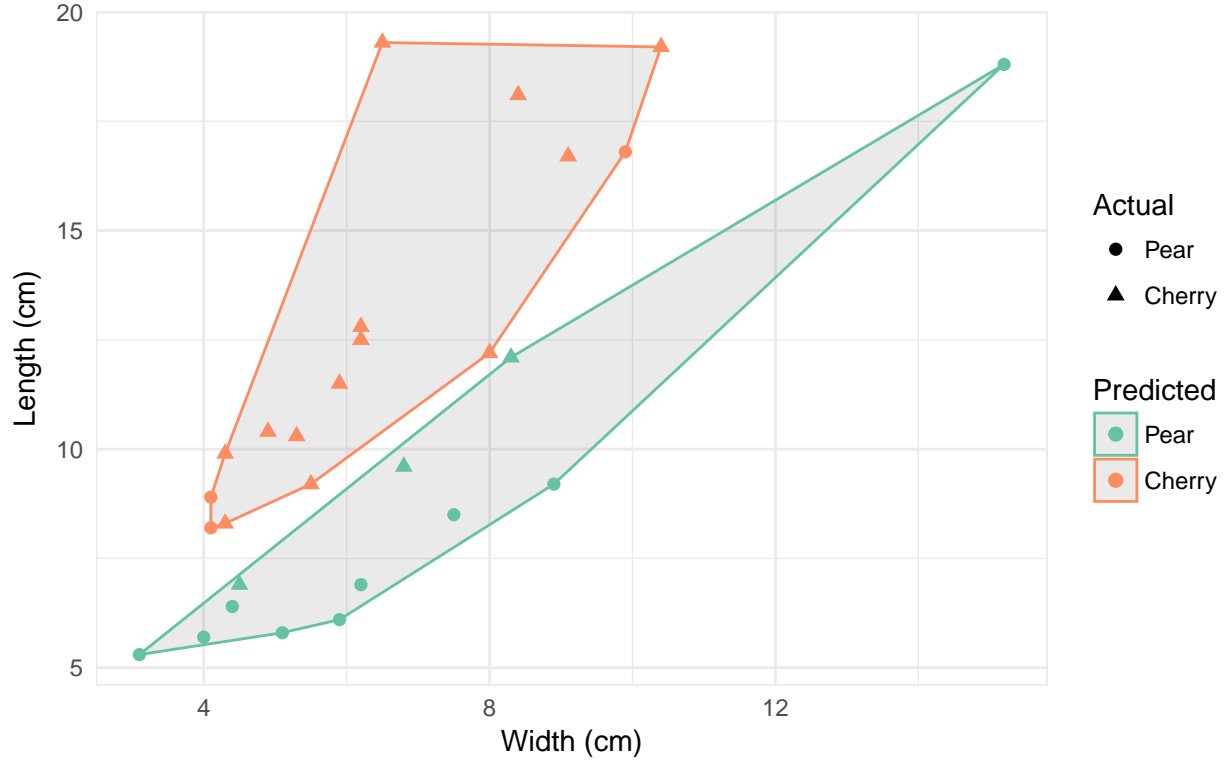
Predicted	Cherry Probability	Pear Probability	Number	Length	Width
Cherry	0.8373176	0.1626824	1	8.2	3.2
Pear	0.2225230	0.7774770	2	5.2	3.8
Cherry	0.6047999	0.3952001	3	7.6	4.0

Table 23 represents the predicted leaf type based on data that was not originally included in the data set. These results are similar to what was seen in both the LDA, and QDA models.

## Observation Space

Figure 12: Length vs Width Scatter Plot

Overlaid with the Convex Hull Based on the Logit Predicted Type



The convex hulls in Figure 12 look identical to the ones presented in the LDA and QDA methods. This is because the logistic regression came up with the same classifications as the other two models.

## Conclusion

Based on these models, the best model is the LDA, which offers the simplest model for ease of interpretation. All the model perform at a similar level of accuracy, in terms of their precision, specificity, and sensitivity, and AUC. The LDA model is a relatively simple model to interpret as it has a intuitive geometric interpretation that is accessible to those with a limited statistics background. As such it is our recommended model to use for this task.

```
## Warning: attributes are not identical across measure variables;  
## they will be dropped
```

Number	Length	Width	Model	Prediction
1	8.2	3.2	LDA Predicted	Cherry
2	5.2	3.8	LDA Predicted	Pear
3	7.6	4.0	LDA Predicted	Cherry
1	8.2	3.2	LDA Cherry Probability	0.8003
2	5.2	3.8	LDA Cherry Probability	0.2772
3	7.6	4.0	LDA Cherry Probability	0.5943
1	8.2	3.2	LDA Pear Probability	0.1997

Number	Length	Width	Model	Prediction
2	5.2	3.8	LDA Pear Probability	0.7228
3	7.6	4.0	LDA Pear Probability	0.4057
1	8.2	3.2	QDA Predicted	Cherry
2	5.2	3.8	QDA Predicted	Pear
3	7.6	4.0	QDA Predicted	Cherry
1	8.2	3.2	QDA Cherry Probability	0.6363
2	5.2	3.8	QDA Cherry Probability	0.3383
3	7.6	4.0	QDA Cherry Probability	0.5712
1	8.2	3.2	QDA Pear Probability	0.3637
2	5.2	3.8	QDA Pear Probability	0.6617
3	7.6	4.0	QDA Pear Probability	0.4288
1	8.2	3.2	Logit Predicted	Cherry
2	5.2	3.8	Logit Predicted	Pear
3	7.6	4.0	Logit Predicted	Cherry
1	8.2	3.2	Logit Cherry Probability	0.8373
2	5.2	3.8	Logit Cherry Probability	0.2225
3	7.6	4.0	Logit Cherry Probability	0.6048
1	8.2	3.2	Logit Pear Probability	0.1627
2	5.2	3.8	Logit Pear Probability	0.7775
3	7.6	4.0	Logit Pear Probability	0.3952

For the new data, Table 24 shows the results across all 3 models. The logistic regression and LDA were quite similar in terms of predicted probabilities. The QDA model’s probabilities were closer to the prior probabilities described in the raw data.

## Appendix

### Appendix A

Table 25: Data

Number	By Type	Type	Length	Width
1	Cherry	11.5	5.9	
2	Cherry	16.7	9.1	
3	Cherry	10.4	4.9	
4	Cherry	18.1	8.4	
5	Cherry	19.3	6.5	
6	Cherry	12.2	8.0	
7	Cherry	10.3	5.3	
8	Cherry	9.9	4.3	
9	Cherry	9.6	6.8	
10	Cherry	19.2	10.4	
11	Cherry	12.8	6.2	
12	Cherry	12.1	8.3	
13	Cherry	12.5	6.2	
14	Cherry	8.3	4.3	
15	Cherry	6.9	4.5	
16	Cherry	9.2	5.5	
1	Pear	5.3	3.1	

Number	By Type	Type	Length	Width
2		Pear	6.9	6.2
3		Pear	9.2	8.9
4		Pear	8.5	7.5
5		Pear	8.2	4.1
6		Pear	5.7	4.0
7		Pear	5.8	5.1
8		Pear	16.8	9.9
9		Pear	6.1	5.9
10		Pear	8.9	4.1
11		Pear	6.4	4.4
12		Pear	18.8	15.2

## References

- [1] The Parts of a Leaf. (17, October 30). Retrieved March 20, 18, from <http://www.robinsonlibrary.com/science/botany/anatomy/leafparts.htm>
- [2] Britannica, T. E. (2016, November 11). Cherry. Retrieved March 20, 2018, from <https://www.britannica.com/plant/cherry>
- [3] Britannica, T. E. (2015, May 13). Pear. Retrieved March 20, 2018, from <https://www.britannica.com/plant/pear>