HW 4 & 5

STA 9797
Erik Carrion
Fall 2021

The Finch dataset contains 80 observations collected over 39 years for 2 classes of Finches – Fortis and Scandens. They are measured along three dimensions: Beak Length, Beak Width, and Beak Depth. The purpose of our analysis is to develop an accurate, reliable model capable of determining class membership. To achieve this end, we employ Linear Discriminant Function Analysis and Logistic Regression.

Our first step is to conduct an exploratory data analysis. We review the summary statistics for each group, the distributions of their attributes, as well as check for trends over time. Then we will move on to our formal analysis.
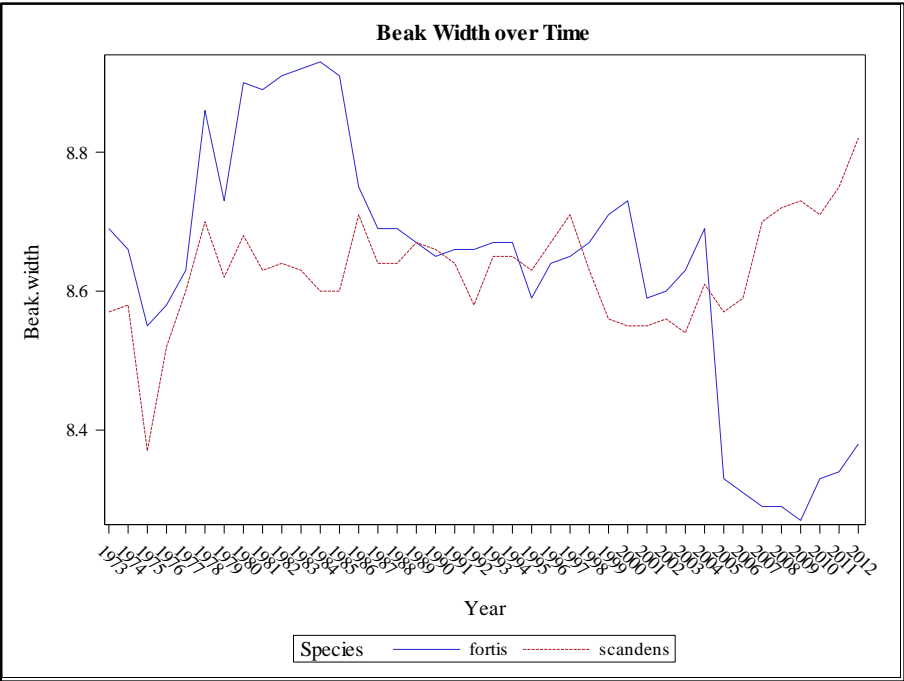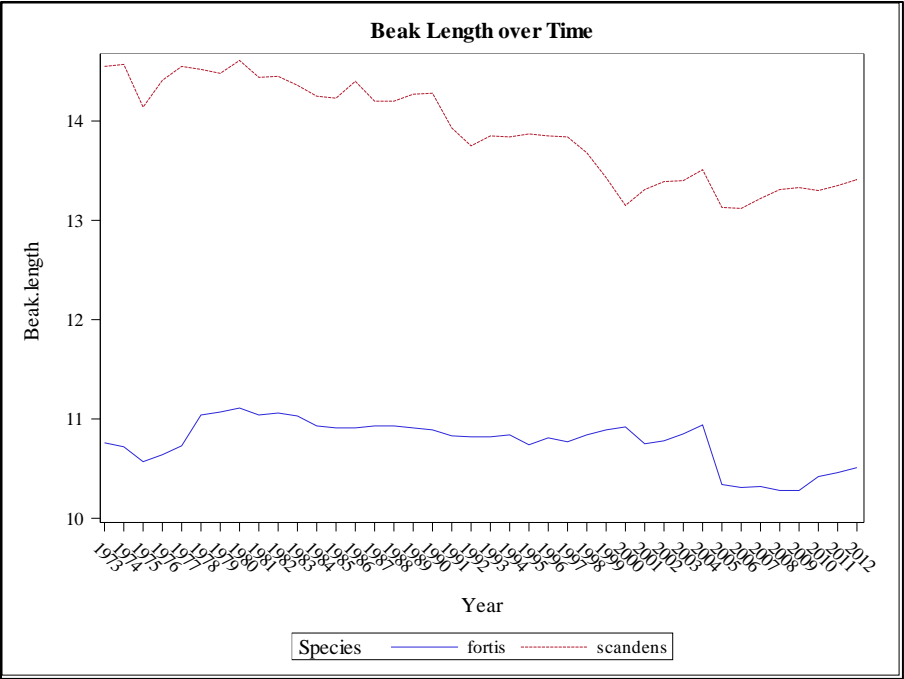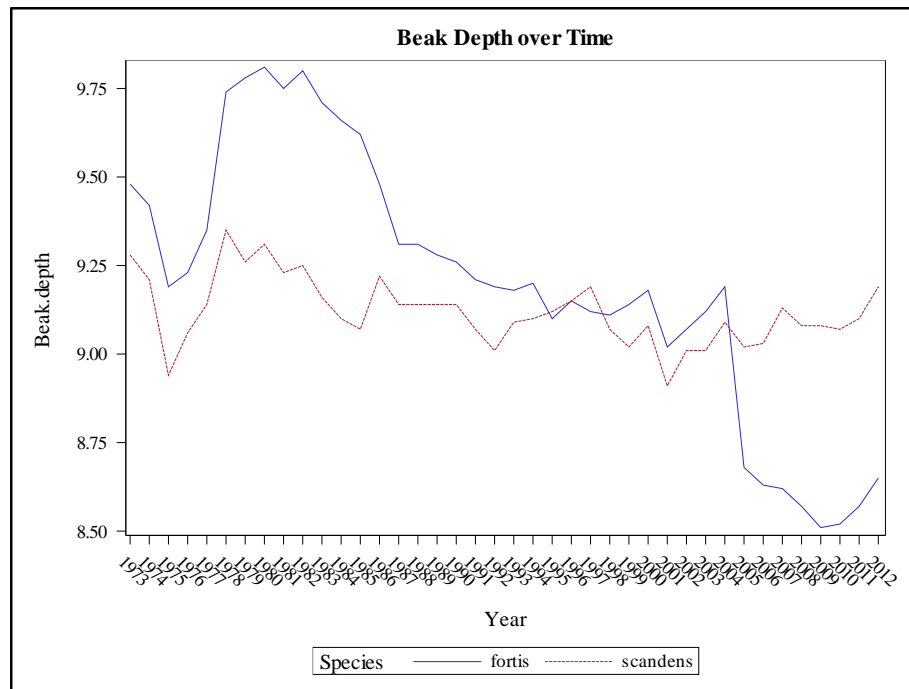
Exploratory Data Analysis

Species	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
fortis	40	Beak_length	Beak.length	40	10.7675000	0.2360438	10.2800000	11.1100000
		Beak_width	Beak.width	40	8.6327500	0.1889239	8.2700000	8.9300000
		Beak_depth	Beak.depth	40	9.1977500	0.3798548	8.5100000	9.8100000
scandens	40	Beak_length	Beak.length	40	13.8970000	0.4987790	13.1200000	14.6100000
		Beak_width	Beak.width	40	8.6295000	0.0772591	8.3700000	8.8200000
		Beak_depth	Beak.depth	40	9.1190000	0.0958043	8.9100000	9.3500000

Looking at the summary statistics, we see that the Means, Minimums and Maximums for Beak Width and Beak Depth are similar for both species. With respect to Width and Depth, the standard deviations, however, are quite different with the Fortis Species exhibiting significantly greater variability (.1889 & .3798) than the Scandens Species (.0772 & .0958).

Beak Length seems to be a defining feature for either of the two Species, with Scandens Finches having much longer beaks than Fortis Finches. The Fortis Finch has an average Beak Length of 10.77 while the Scandens Finch has a mean Length of 13.897. Furthermore, we see that the minimum and maximums for the two species display no overlap.

Given the data were generated over 39 years from 1973 to 2012, we wondered if there was an observable trend over time. We plotted measurements for each species and the following graphs were generated. We see an overall downward trend for Length and Depth for the two species while Beak Width for Scandens Finches saw an increase over time. Compared to the changes in Beak Length, the changes in Width and Depth display a more volatile progression. Overall, this suggests an environmental change which has selected for generally smaller beaks.

Beak Length over Time

Beak.length vs Year, Species: fortis, scandens

Beak Width over Time

Beak.width vs Year, Species: fortis, scandens

Next, we look at the distributions of the three variables and their association with each other to determine if the data exhibit characteristics which would threaten the internal validity of our results. Specifically, we check if the data are approximately normal, don't have too many outliers, and the variables aren't too correlated.

Below we can see that for the Fortis Finches, the distributions of the three variables seem approximately normal but the variables are highly correlated with each other. Given the near perfect pairwise linear relationship among the variables, for Fortis Finches we would consider eliminating Width and Depth in further analyses as they don't seem to contribute additional information beyond what Beak Length already provides.

The story is a bit different for Scandens, though. As we can see in the graph below, the variables are also approximately normally distributed but unlike Fortis, the pairwise scatterplots show very little linear relationship among the variables.

To identify outliers and get a different perspective on the distribution of the measured attributes, we turn to the boxplots. The presence of outliers, more so than deviations from Normality, threaten the ability of our Discriminant Functions to properly classify observations. If there are serious outliers, we would not be confident in the resulting function's ability to properly discriminate.

We see from the boxplots that Beak Length has very few outliers whereas for Width and Depth there are quite a few. Looking back to our summary statistics we can determine how far the minimum and maximum observations are from their means. For Fortis, all observations fall within about 2 standard deviations of the mean. For the Scandens species, though, there are observations which are as much as 3.36 standard deviations away. With that said, the standard deviations are very small and considering that the minimum and maximum observations are between 92.52% and 1.067% of their means, we don't expect our analysis to be seriously affected by these outliers.
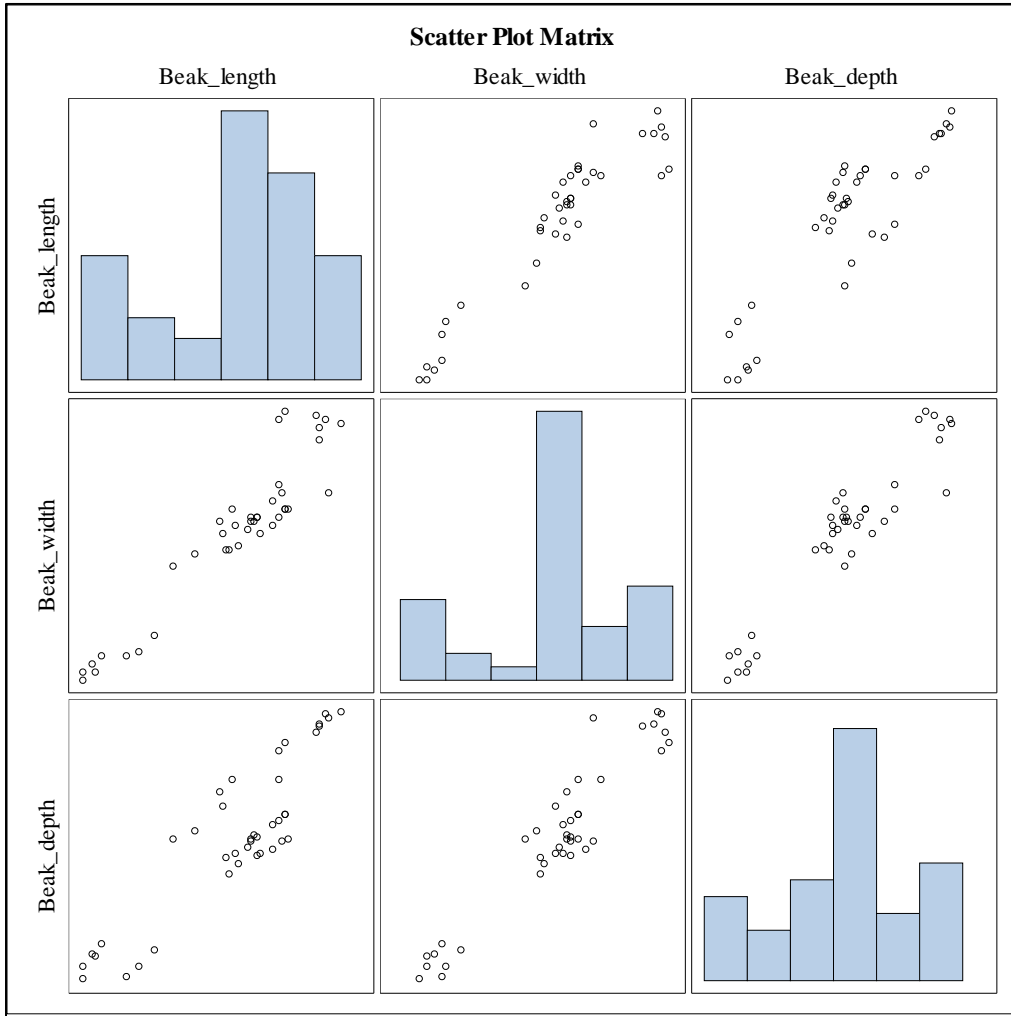
Figure 1 Species = Fortis
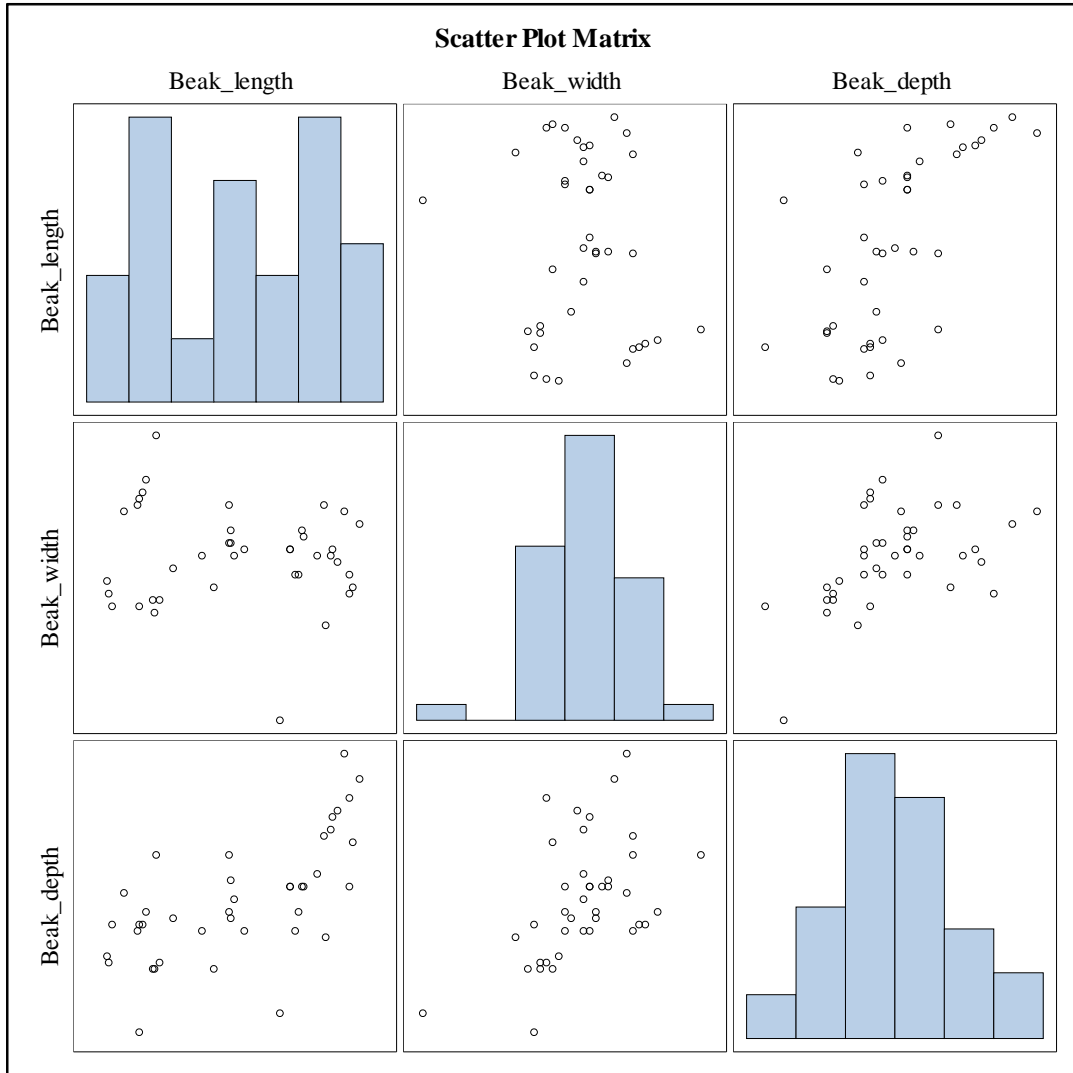
Figure 2 Species = Scandens
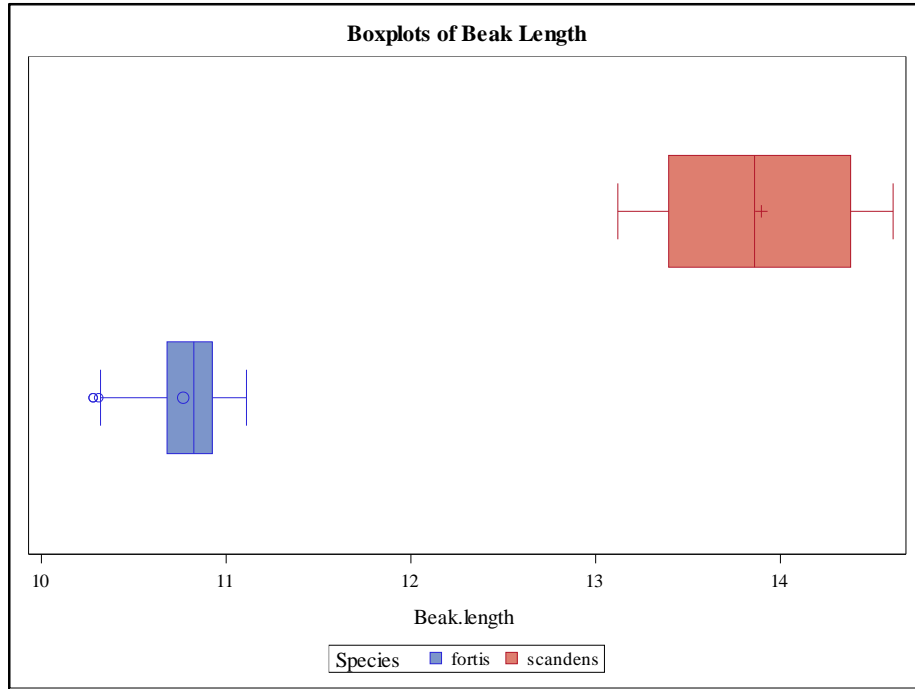

Scatter Plot Matrix

Figure 3 Boxplot of Beak Length

Figure 4 Boxplot of Beak Width
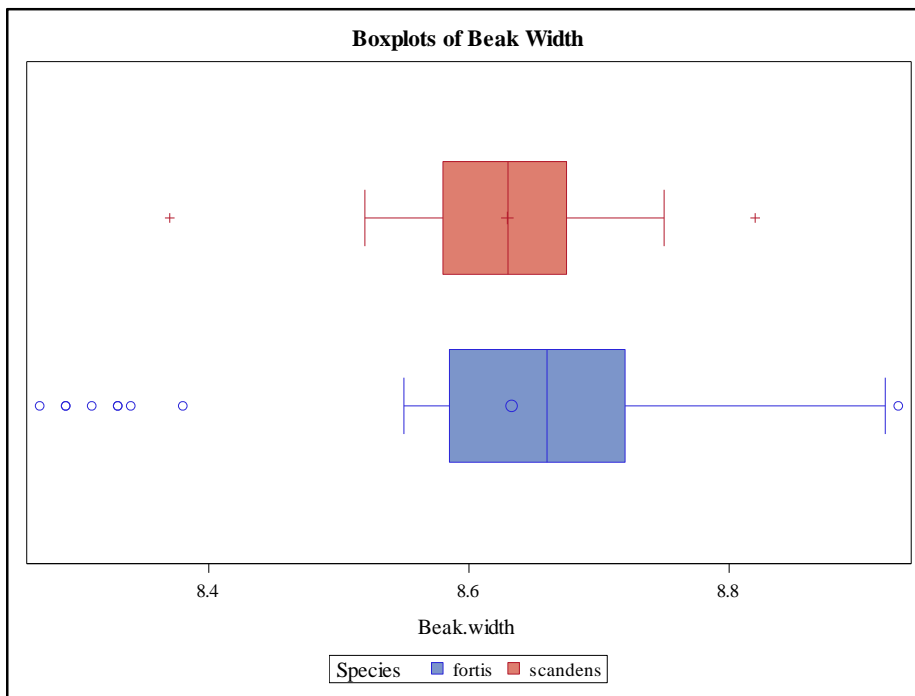
Figure 5 Boxplot of Beak Depth
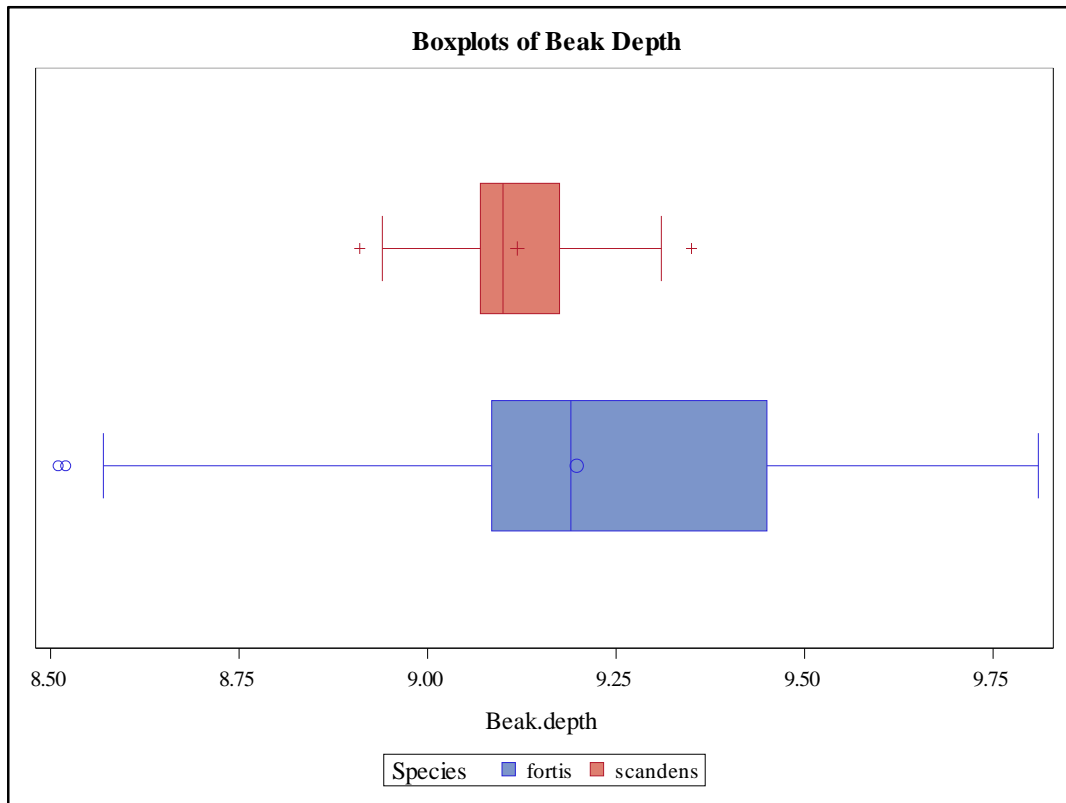
Number of Standard Deviations Away from the Mean

Species	Variable	Mean	SD	Min Value	Max Value
Fortis	Length	10.7675000	0.2360	**2.0653**	**1.4510**
	Width	8.6327500	0.1889	**1.9201**	**1.5734**
	Depth	9.1977500	0.3799	**1.8106**	**1.6118**
Scandens	Length	13.8970000	0.4988	**1.5578**	**1.4295**
	Width	8.6295000	0.0773	**3.3588**	**2.4657**
	Depth	9.1190000	0.0958	**2.1815**	**2.4112**

The goal of Linear Discriminant Function Analysis is to generate a discriminant function which can classify an observation as belonging to one group or another based on the association between the set of groups and the set of observed features. Since the analysis seeks to associate one set of variables with another set of variables, LDF shares characteristics with MANOVA. Unlike MANOVA, LDF does not place a stringent requirement on Multivariate Normality (MVN). It is sensitive to outliers, though. As concluded above, despite our data having some outliers, we don't deem them so abnormal as to threaten the validity of our analysis.

In SAS, we invoke the "pool = test" option to determine whether a Linear (LDF) or Quadratic Discriminant Function (QDF) is appropriate. The test for Homogeneity of Within Covariance Matrices yields a chi-square value with a p-value of <.0001 so we reject the null hypothesis that the Within Covariance Matrices are equal and deploy a QDF Analysis.

Chi-Square	DF	Pr > ChiSq
166.472325	6	<.0001

Below we have the test of the canonical correlation between the set of Species and set of observed Features with the correlation coming in at .981921 suggesting a very high association between the two sets, confirmed by the exact F-Statistic of 681.70 having a p-value of <.0001

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.981921	0.981682	0.004031	0.964170	26.9094		1.0000	1.0000

	Test of H0: The canonical correlations in the current row and all that follow are zero				
	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.03583023	681.70	3	76	<.0001

Looking to the Pooled Within Canonical Structure, which adjusts for group membership, we see the following output. Displayed are the simple correlations between the observed features and the discriminant function after controlling for group membership. As theorized above, Beak Length plays a large part in the overall canonical structure.

Pooled Within Canonical Structure		
Variable	**Label**	**Can1**
Beak_length	Beak.length	0.782914
Beak_depth	Beak.depth	-0.027751
Beak_width	Beak.width	-0.002198

To interpret each variable's relative contribution, we look to the standardized coefficients. Doing so reveals that Beak Length is now the 2nd most impactful variable with Beak Depth being the most impactful. As in the overall structure, Beak Width does not play as large a role as the other two variables.

Pooled Within-Class Standardized Canonical Coefficients		
Variable	**Label**	**Can1**
Beak_length	Beak.length	1.228318653
Beak_depth	Beak.depth	-1.451089463
Beak_width	Beak.width	0.880994008

The purpose of Linear Discriminant Function Analysis is to develop a model which can successfully differentiate between the different groups. To assess the model's ability to do so, we deploy it on the training data to see how well it classifies the individual observations. The results of the classification are below.

Number of Observations and Percent Classified into Species			
From Species	**fortis**	**scandens**	**Total**
fortis	40 100.00	0 0.00	40 100.00
scandens	0 0.00	40 100.00	40 100.00
Total	40 50.00	40 50.00	80 100.00
Priors	0.5	0.5	

Error Count Estimates for Species			
	fortis	**scandens**	**Total**
Rate	0.0000	0.0000	0.0000
Priors	0.5000	0.5000	

We see that the model was able to discriminate among the different types of Finches with 100% accuracy so we are confident this model will be able to categorize new observations well.

Our analysis continues with the use of logistic regression to predict class membership. Because we have three independent features our first step was to test if there were significant interactions among the features. We tested the three-way and two-way interactions and found each interaction was not significant, allowing us to focus on first order models.

3 Way Interaction Length * Width * Depth

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	172.5	127.6	1.8281	0.1763
Beak_l*Beak_w*Beak_d	1	-0.1741	0.1286	1.8327	0.1758

2 Way Interaction Length * Width

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	93.5730	97.2189	0.9264	0.3358
Beak_leng*Beak_width	1	-0.8847	0.9169	0.9310	0.3346

2 Way Interaction Length *Depth

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	122.9	109.7	1.2563	0.2624
Beak_leng*Beak_depth	1	-1.0829	0.9647	1.2599	0.2617

2 Way Interaction Depth * Width

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8222	5.0534	0.9106	0.3400
Beak_dept*Beak_width	1	0.0610	0.0638	0.9125	0.3395

Given that the interactions were not significant we looked to a full first order model. Performing the analysis resulted in SAS being unable to find the maximum likelihood estimate for the parameters because it encountered complete separation of the data, resulting in parameter estimates which are unreliable and highly variable.

Warning: The maximum likelihood estimate does not exist.

Warning: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	51.1060	756.5	0.0046	0.9461
Beak_length	1	-6.4405	7.5505	0.7276	0.3937
Beak_width	1	-3.0277	161.1	0.0004	0.9850
Beak_depth	1	5.8247	92.4528	0.0040	0.9498

Complete separation occurs when the set of dependent variables can completely separate a set or subset of the features such that, given the feature or features in question, we can perfectly predict the group to which the observation belongs. When this occurs, our maximum likelihood estimates of the parameters do not converge and what estimates we do end up with have unacceptably high variability.

Above we saw there was zero overlap in the measurements for Beak Length between the two species. We must ask if these differences in Beak Length are endemic to the population of Fortis and Scandens Finches, or a peculiarity of the given sample. Below we reproduce the boxplot for Beak Length.
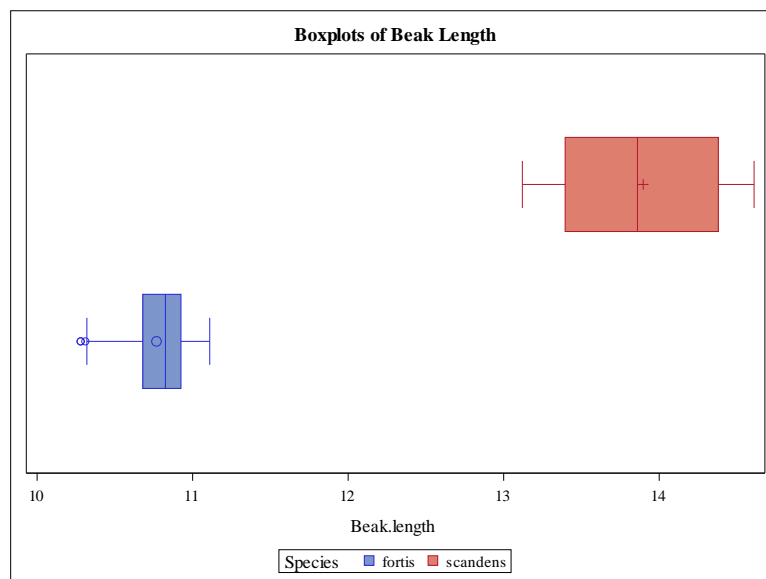
If the features which lead to complete separation are endemic, there is no need to generate a model because the decision function reduces to whether Beak Length falls within one of the measurement intervals or not. In other words, if a new finch has a Beak Length between 10.28 and 11.11 then it is a Fortis finch. If the measurement is between 13.12 and 14.16 then it is a Scandens finch.

If the features are not endemic to the population, we should adjust our model. There are several methods available to us, of which the following are three possible options.

1. We could remove the variable from the model completely. However, doing so means our conclusion that the complete separation is not endemic to the population has to be correct. Given there are 40 observations for each Species, we find this proposition untenable.
2. We could employ a regularization technique which would shrink coefficients towards zero and reduce the variability of our estimates. Doing so would retain the variable of interest but take in to account its effect and adjust accordingly.
3. We can take a Bayesian approach, as suggested by Gelman et al (2008), and allocate a prior distribution to our parameter estimates.

Of these three choices we will employ method 2 in our analysis and assess model performance via the AUC of the respective ROC curves.

Logistic Regression – Full Model – With Regularization

We first apply regularization to the full model and assess what effect, if any, it has on model performance. Applying regularization to the full model yields similar results as without regularization. Namely, the Maximum Likelihood Estimates do not converge. This time, however, we see the estimates and their standard errors have adjusted somewhat but the standard errors for width and depth are still very high. While the AUC of the ROC curve is 100% the model is not appropriate and we should consider a reduced model.

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-17.4300	54.3203	0.1030	0.7483
Beak_length	1	-2.8148	0.7480	14.1612	0.0002
Beak_width	1	8.1795	11.4621	0.5092	0.4755
Beak_depth	1	-2.1022	5.3261	0.1558	0.6931

Figure 6 ROC Curve Full Model

ROC Curve for Model
Area Under the Curve = 1.0000
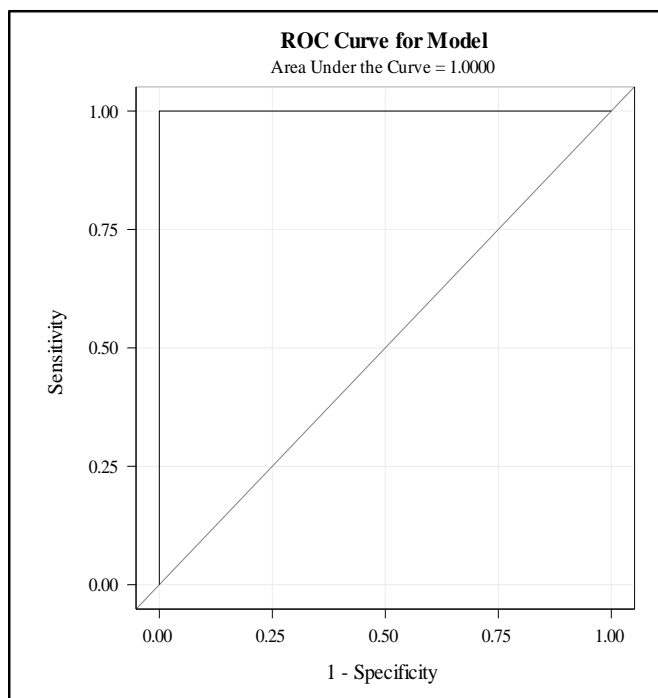
Logistic Regression – 2 Variable Model: Width & Depth

We evaluated a 2-variable model with just Beak Width and Beak Depth. The resulting analysis was able to converge on an MLE for the parameters and yielded the following output:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	28.6092	18.6930	2.3424	0.1259
Beak_width	1	-8.8559	4.0074	4.8836	0.0271
Beak_depth	1	5.2232	2.1134	6.1082	0.0135

The parameter estimates are significant at the .05 level but the standard errors are still relatively high and the AUC of the ROC curve below is .6469. After regularization we get the following parameter estimates which are not very different from the un-regularized model and, in addition, the ROC curves are identical, suggesting regularization is unnecessary for this model.

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	26.7672	18.4881	2.0961	0.1477
Beak_width	1	-8.2716	3.9427	4.4014	0.0359
Beak_depth	1	4.8732	2.0750	5.5156	0.0188

Figure 7 Width & Depth - No Regularization   *Figure 8 Width & Depth - With Regularization*
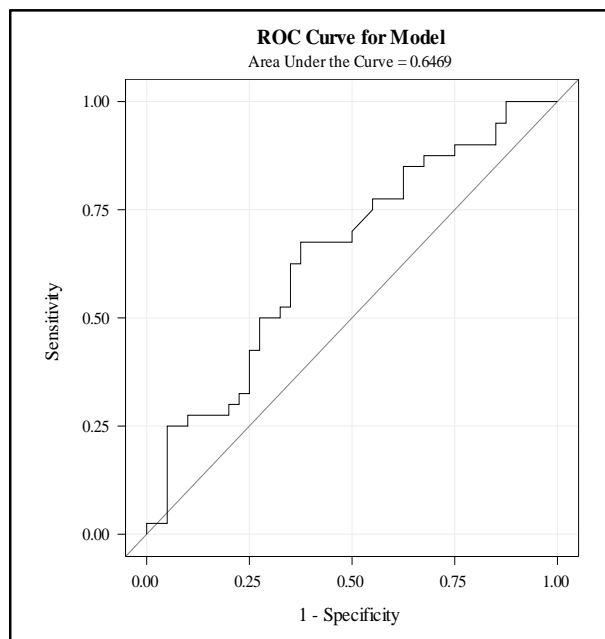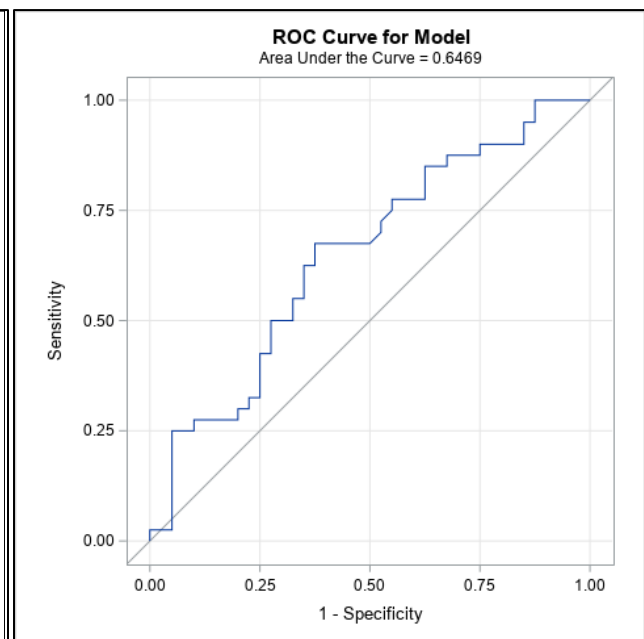
Logistic Regression – 2 Variable Models: Length & Depth, Length & Width

The 2-variable models which include length suffer from the same issues as the full model with respect to their inability to converge on MLE estimates. In both scenarios, Depth and Width are insignificant before and after regularization while Length is insignificant before regularization but becomes significant afterwards.

Length & Depth without Regularization

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	38.7064	397.3	0.0095	0.9224
Beak_length	1	-6.4810	7.3337	0.7810	0.3768
Beak_depth	1	4.3710	42.4753	0.0106	0.9180

Length & Depth with Regularization

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	10.9534	26.5460	0.1703	0.6799
Beak_length	1	-2.7447	0.6383	18.4930	<.0001
Beak_depth	1	2.4339	2.9040	0.7024	0.4020

Length & Width without Regularization

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	25.3330	667.1	0.0014	0.9697
Beak_length	1	-6.6852	7.3884	0.8187	0.3656
Beak_width	1	6.4494	76.9028	0.0070	0.9332

Length & Width with Regularization

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-14.9578	45.4394	0.1084	0.7420
Beak_length	1	-2.9115	0.7048	17.0658	<.0001
Beak_width	1	5.8124	5.6456	1.0600	0.3032

The ROC curves for each pairwise model including Length has an AUC of 100%. Given that beak length is the variable along which complete separation is possible and the fact that each regularized pairwise model with length included yielded parameter estimates for the covariate which were insignificant, we conclude that a model with just Length is the most appropriate. The un-regularized and regularized parameter estimates are below.

Without regularization, Beak Length is not significant and has a substantial standard error. While the boxplots suggest Beak length is a perfect predictor for Species, Logistic Regression is telling us that Beak Length is not a significant variable because of the high standard error of the parameter estimate. So, while Beak length yields very good results with respect to the ROC curves, the model estimates are so variable as to render it un-useable for predicting class membership on new data.

After regularization, Length is significant and has a much, much smaller standard error. Further, the AUC is still 100% so we conclude that a single variable regularized model with only Length as our predictor is most appropriate.

To further clarify why we believe a single variable model is best, we consider the vector plots of the variables. From the vector plots of Width vs. Depth we immediately see they each provide the exact same information. Since they are basically the same for each Species, including both in a model would be redundant, thus disqualifying the full first order model and the two-variable model with Width and Depth.

If we compare the Length vector against the Width and Depth vectors, we see the angles between the vectors are approximately equal so the two models will yield similar results which is confirmed by our analysis above. Given that the use of either Depth or Width in conjunction with Length renders those two variables insignificant, we are left with the single variable model with Length alone as our sole predictor.

As such, we conclude that a regularized single variable model using Length alone is most appropriate as it converges to a stable parameter estimate while yielding a perfect ROC curve.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	83.3915	91.6767	0.8274	0.3630
Beak_length	1	-6.8563	7.4946	0.8369	0.3603

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	38.8438	9.7893	15.7448	<.0001
Beak_length	1	-3.1820	0.8038	15.6705	<.0001

Figure 8 ROC Curve - Length Only with Regularization

Pairwise Vector Plots of Finch Features

Pairwise Vector Plots of Finch Features
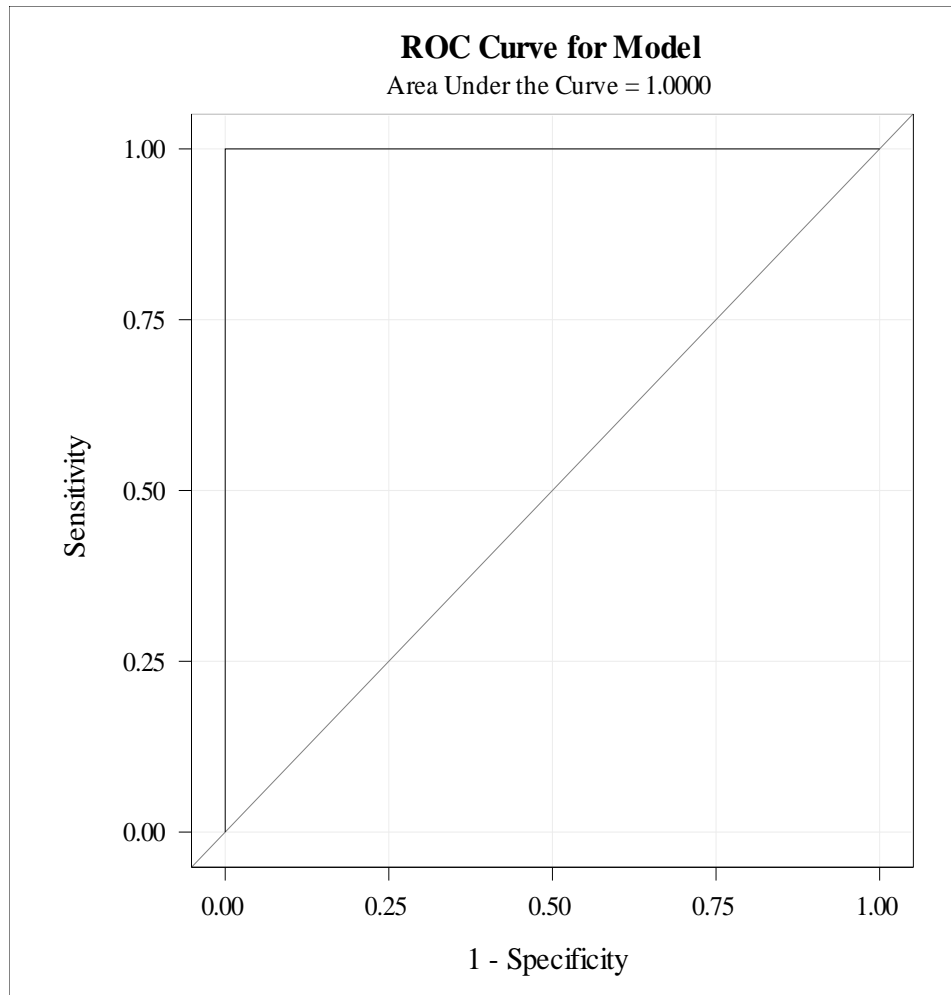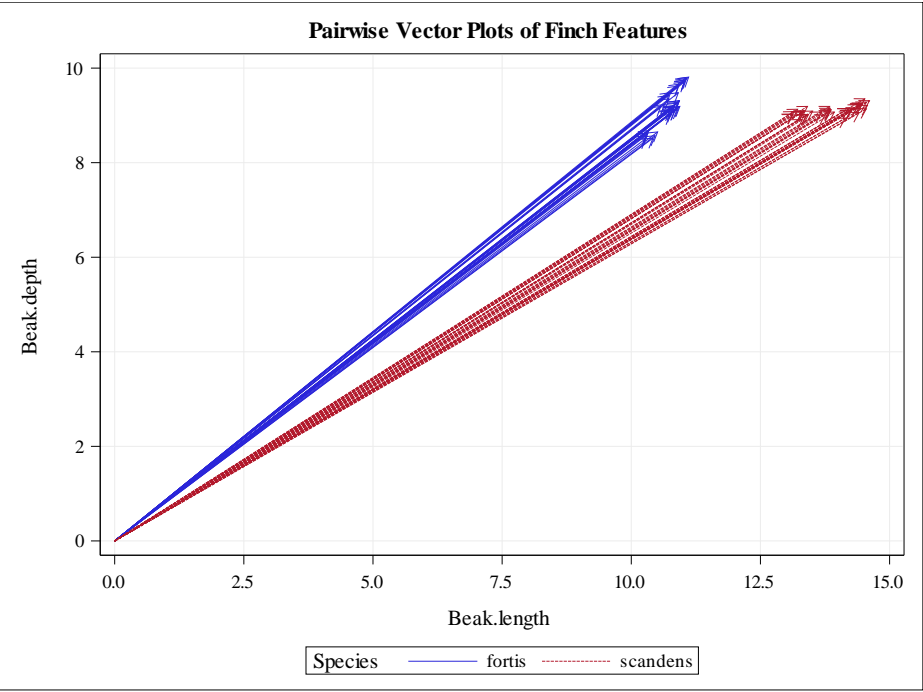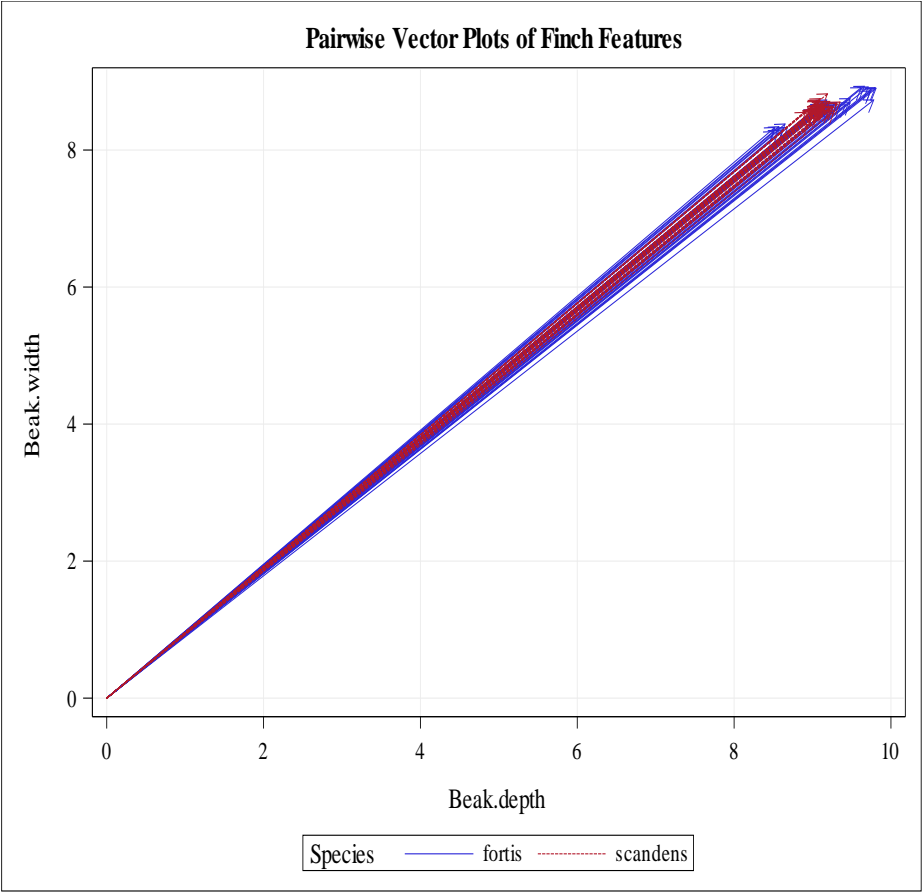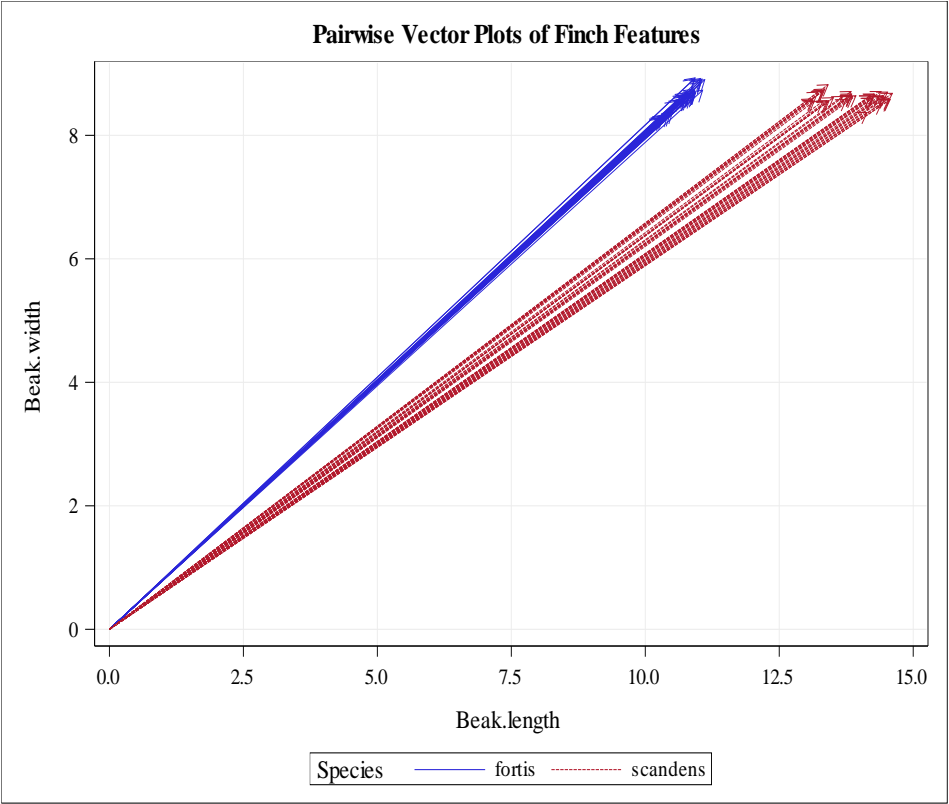
SAS CODE

```sas
proc import file = 'finches.xlsx' dbms = xlsx out = finches replace;
run;


/* Exploratory Data Analysis */

/* Summary Statistics */
proc means data = finches;
        class Species;
        var Beak_length Beak_width Beak_depth;
        title "Summary Statistics for Finches Dataset";
run;
/* Scatterplot matrix and Normality Assessment */
proc corr data = finches COV plots(maxpoints=NONE)=matrix(histogram);
        by Species;
        var Beak_length Beak_width Beak_depth;
        ods select MatrixPlot;
        title "Scatterplot Matrix - Assessing Multivariate Normality";
run;


/* Create Boxplots of the individual variables*/
/**** Beak Length ****/
proc sgplot data = finches;
        title "Boxplots of Beak Length";
        hbox Beak_length/group=Species;
run;
/**** Beak Width ****/
proc sgplot data = finches;
        title "Boxplots of Beak Width";
        hbox Beak_width/group=Species;
run;
/**** Beak Depth ****/
proc sgplot data = finches;
        title "Boxplots of Beak Depth";
        hbox Beak_depth/group=Species;
run;


/**** Is there a trend over time? ****/

/* Beak Length Over Time */
proc sgplot data = finches;
        title "Beak Length over Time";
        xaxis type = discrete;
        series x = Year y = Beak_length/ group = Species;
run;
/* Beak Width Over Time */
proc sgplot data = finches;
        title "Beak Width over Time";
        xaxis type = discrete;
        series x = Year y = Beak_width/ group = Species;
run;
/* Beak Depth Over Time */
proc sgplot data = finches;
        title "Beak Depth over Time";
        xaxis type = discrete;
        series x = Year y = Beak_Depth/ group = Species;
run;


/**** Linear Discriminant Analysis ****/
proc discrim data = finches pool = test method = normal list canonical anova bcorr bcov;
        class Species;
        var Beak_length Beak_depth Beak_width;
        title "Linear Discriminant Classification Method = Normal";
run;
```

```
/**** Logistic Regression ****/

/* Test the Interactions */
proc logistic data = finches;
        title "Finches LogReg - 3Way Interaction Test";
        class Species;
        model Species = Beak_length*Beak_width*Beak_depth/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;

/* Three-way interaction not significant */

/* Test the 2 Way Interactions */
proc logistic data = finches;
        title "Finches LogReg - 2 Way Interaction Length*Width";
        class Species;
        model Species = Beak_length*Beak_width/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;

proc logistic data = finches;
        title "Finches LogReg - 2 Way Interaction Length*Depth";
        class Species;
        model Species = Beak_length*Beak_depth/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;
/* Length*Depth not significant */

proc logistic data = finches;
        title "Finches LogReg - 2 Way Interaction - Width*Depth";
        class Species;
        model Species = Beak_width*Beak_depth/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;
/* Length*Width not significant */
/* All Interaction terms deemed not significant */
/* Now look at first order models*/

/* Full Model with and without regulariation */
proc logistic data = finches plots=roc;
        title "Finches LogReg - Full Model w/o Regularization";
        class Species;
        model Species = Beak_length Beak_width Beak_depth/pprob=(0.05 to .50 by .05) ctable lackfit;
run;
proc logistic data = finches plots=roc;
    title "Finches LogReg - Full Model w/ Regularization";
    class Species;
    model Species = Beak_length Beak_width Beak_depth/ firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;

/* Pairwise Models with and without regularization */

/* Width & Depth */
proc logistic data = finches plots=roc;
        title "Finches LogReg Width and Depth no Regularization";
        class Species;
        model Species = Beak_width Beak_depth/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;
proc logistic data = finches plots=roc;
        title "Finches LogReg Width and Depth w/ Regularization";
        class Species;
        model Species = Beak_width Beak_depth/firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;
/* Length & Depth */
proc logistic data = finches plots=roc;
        title "Finches LogReg - Length & Depth no regularization";
        class Species;
        model Species = Beak_length Beak_depth/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;

proc logistic data = finches plots=roc;
```

```sas
        title "Finches LogReg - Length & Depth w/ regularization";
        class Species;
        model Species = Beak_length Beak_depth/firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;


/* Length and Width */
proc logistic data = finches plots=roc;
        title "Finches LogReg Length and Width no regularization";
        class Species;
        model Species = Beak_length Beak_width/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;
proc logistic data = finches plots=roc;
        title "Finches LogReg Length and Width w/ regularization";
        class Species;
        model Species = Beak_length Beak_width/firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;


/* Single Variable Models */

/* Length Only */
proc logistic data = finches plots=roc;
        title "Finches LogReg - Length Only no regularization";
        class Species;
        model Species = Beak_length/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;

proc logistic data = finches plots=roc;
        title "Finches LogReg - Length Only w/ regularization";
        class Species;
        model Species = Beak_length/ firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;

/* Width Only */
proc logistic data = finches plots=roc;
        title "Finches LogReg - Width Only no regularization";
        class Species;
        model Species = Beak_width/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;

proc logistic data = finches plots=roc;
        title "Finches LogReg - Width Only w/ regularization";
        class Species;
        model Species = Beak_width/firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;

/* Depth only */
proc logistic data = finches plots=roc;
        title "Finches LogReg - Depth Only no regularization";
        class Species;
        model Species = Beak_depth/ pprob=(0.05 to .50 by .05) ctable lackfit;
run;
proc logistic data = finches plots=roc;
        title "Finches LogReg - Depth Only w/ regularization";
        class Species;
        model Species = Beak_depth/ firth pprob=(0.05 to .50 by .05) ctable lackfit;
run;

/* Using only Width or Depth yield unacceptable results */
/* As such, will not be considered in the analysis */
```