DATA 622 Assignment 2

CUNY: Spring 2021

Philip Tanofsky

19 March 2021

Introduction

The purpose of this project is to apply generative model approaches to the Palmer Penguin data set available at https://allisonhorst.github.io/palmerpenguins/articles/intro.html (Horst, Hill, and Gorman 2020). The first approach performs linear discriminant analysis (LDA) on the dataset in order to predict the species of the penguin subjects. The second approach performs a quadratic discriminant analysis in order to predict the species of penguin subjects. The final approach uses the Naive Bayes modeling approach to also predict the species of the penguin subjects. The exploratory data analysis informs the decisions of the predictor variables included for each generative model. A final comparison of the models are presented in order to compare the accuracy of each.

```
# Import required R libraries
library(palmerpenguins)
library(tidyverse)
library(caret)
library(MASS)
library(ggplot2)
library(mvtnorm)
library(e1071)
library(klaR)
library(pROC)
library(corrplot)
theme_set(theme_classic())
```

Initial Data Inspection

```
ds <- penguins

head(ds)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
```

```
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3              193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6              190        3650 male
## # ... with 1 more variable: year <int>
```

```
summary(ds)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
dim(ds)
```

```
## [1] 344   8
```

The palmer penguins dataset consists of 8 variables, 7 independent variables and 1 dependent variable (species).

Variables

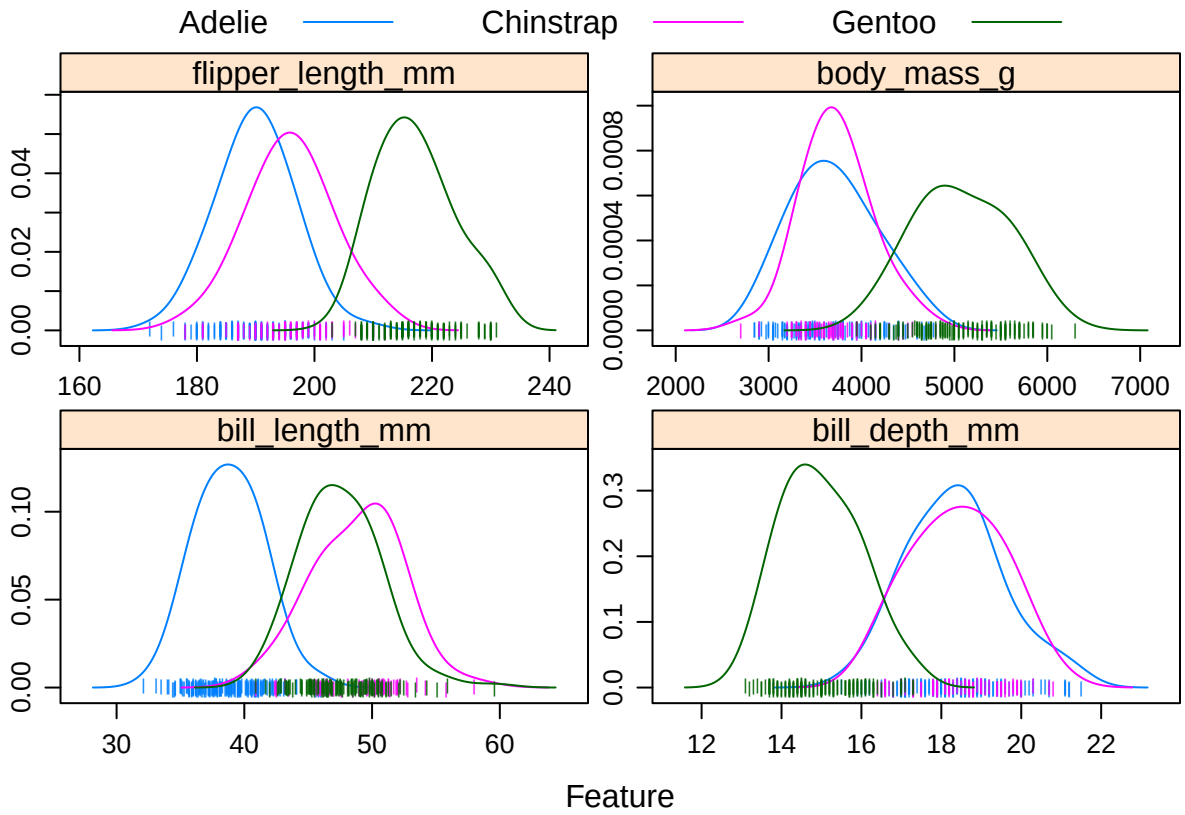
- species: species of the penguin observed (dependent variable)
- island: island of penguin's observation
- bill_length_mm: penguin bill length in millimeters
- bill_depth_mm: penguin bill depth in millimeters
- flipper_length_mm: penguin flipper length in millimeters
- body_mass_g: penguin body mass in grams
- sex: penguin sex
- year: year of observation

Exploratory Data Analysis

For linear discriminant analysis, two assumptions are made about the data. One, the predictors are normally distributed, which means the data follows a Gaussian distribution for each class. Two, the classes have class-specific means but also have equal variance and covariance.

First step, confirm multivariate normal distribution. A density plot of the four continuous variables by species indicates the normal distribution for the class-specific plots.

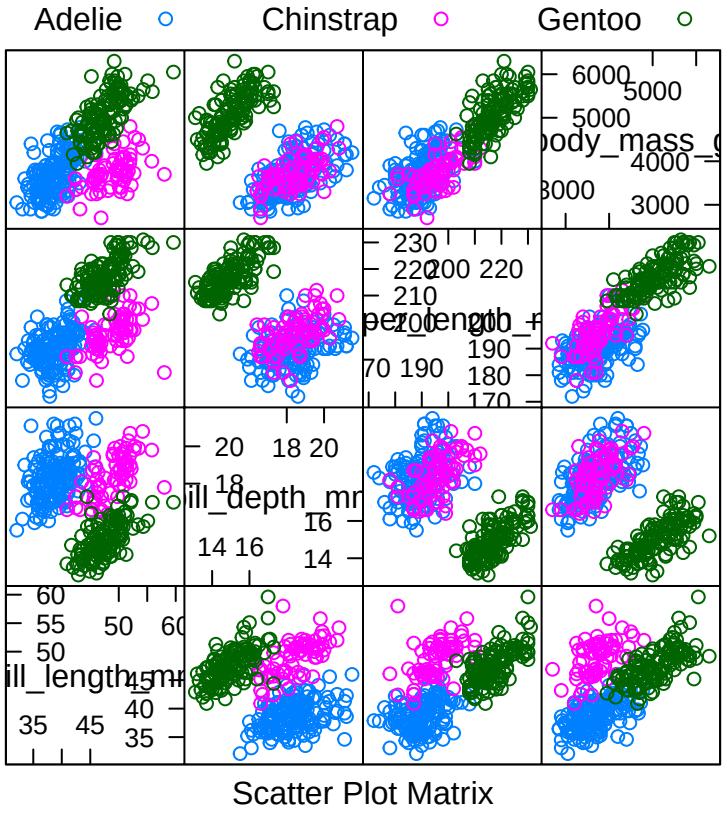
```
# Overlayed density plots
featurePlot(x = penguins[, 3:6],
            y = penguins$species,
            plot = "density",
            # Pass in options to xyplot() to
            # make it prettier
            scales = list(x = list(relation="free"),
                          y = list(relation="free")),
            adjust = 1.5,
            pch = "|",
            layout = c(2, 2),
            auto.key = list(columns = 3))
```



The variable $flipper_length_mm$ shows the clearest example of normal distribution by species class. The four density plots also indicate common evaluations between the Adelie and Chinstrap species for the independent variables flipper_length_mm, body_mass_g, and bill_depth_mm. The remaining continuous variable, bill_length_mm, shows an overlap between the Chinstrap and Gentoo species.

The scatterplot matrix of the continuous variables indicates the relationships between the independent variables for each of the three penguin species.

```
# Use featurePlot
# https://topepo.github.io/caret/visualizations.html

# Scatterplot
featurePlot(x = penguins[, 3:6],
            y = penguins$species,
            plot = "pairs",
            # Add a key at the top
            auto.key = list(columns = 3))
```



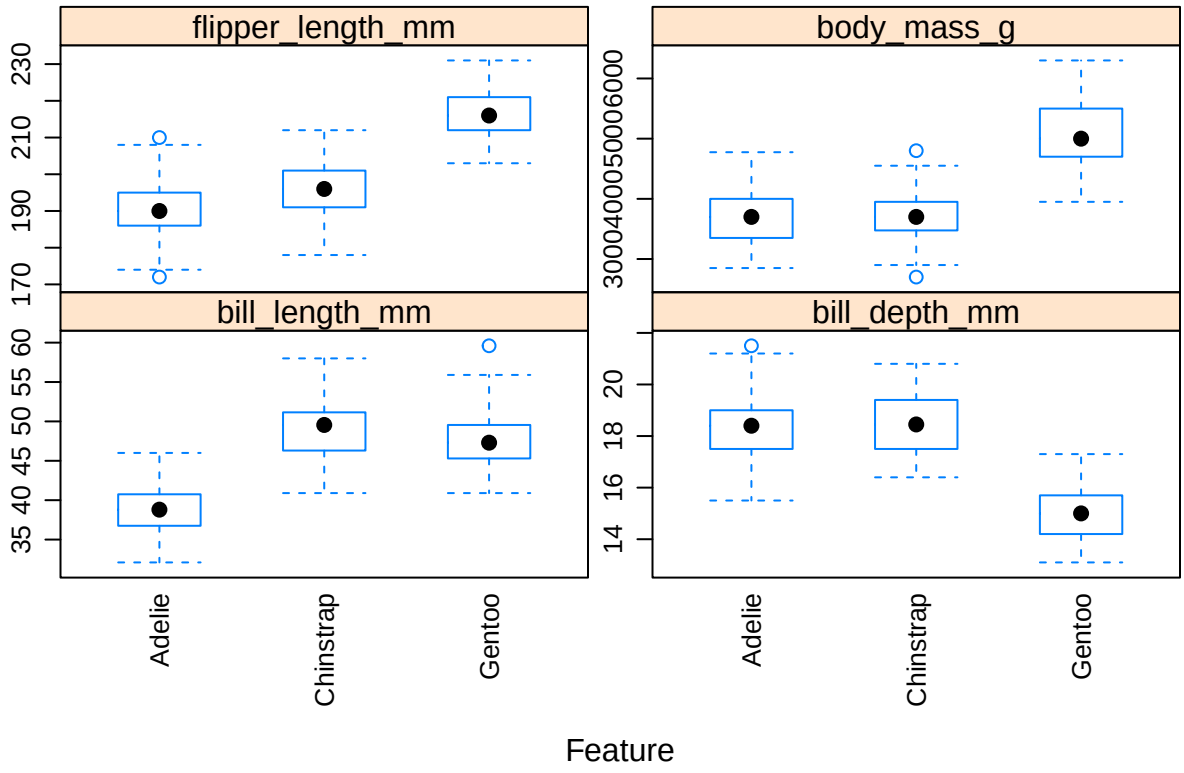
The independent variable $bill_length_mm$ stands out by providing a clear distinction between the three penguin species when compared with any of the other three independent variables. As seen in the above scatterplot, when using any two of the other three independent variables, the plot results in a clear overlap of the Adelie and Chinstrap species, an expected result given the prior density plots.

So far, the independent variable $bill_length_mm$ is a prime candidate to be included in the generative models. The key will be identifying which of the other variables will provide additional significance to the model.

The following boxplots provide another perspective of the continuous variable relationships across the three species. As already noted, Adelie and Chinstrap show similar means and distributions for three of the four variables. The boxplot does indicate a few outliers across the variable values but nothing egregious.

```
featurePlot(x = penguins[, 3:6],
            y = penguins$species,
```

```
            plot = "box",
            ## Pass in options to bwplot()
            scales = list(y = list(relation="free"),
                          x = list(rot = 90)),
            layout = c(2,2),
            auto.key = list(columns = 2))
```



With the first assumption confirmed, the second assumption is to confirm the similar covariance across the species classes. The following covariance checks further identify the relationship between the variables.

```
# Covariance matrix
p_g <- penguins %>% filter(species == 'Gentoo')
cov_mat <- cov(p_g[,3:6],  use = "complete.obs")
round(cov_mat, 2)
```

```
##                    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## bill_length_mm               9.50          1.95             13.21     1039.63
## bill_depth_mm                1.95          0.96              4.50      355.69
## flipper_length_mm           13.21          4.50             42.05     2297.14
## body_mass_g               1039.63        355.69           2297.14   254133.18
```

```
p_a <- penguins %>% filter(species == 'Adelie')
cov_mat <- cov(p_a[,3:6],  use = "complete.obs")
round(cov_mat, 2)
```

```
##                    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## bill_length_mm               7.09          1.27              5.67      670.36
## bill_depth_mm                1.27          1.48              2.45      321.44
## flipper_length_mm            5.67          2.45             42.76     1404.03
## body_mass_g                670.36        321.44           1404.03   210282.89
```

```
p_c <- penguins %>% filter(species == 'Chinstrap')
cov_mat <- cov(p_c[,3:6],  use = "complete.obs")
round(cov_mat, 2)
```

```
##                    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## bill_length_mm              11.15          2.48             11.23      659.20
## bill_depth_mm                2.48          1.29              4.70      263.79
## flipper_length_mm           11.23          4.70             50.86     1758.54
## body_mass_g                659.20        263.79           1758.54   147713.45
```

Comparing the variance of each continuous independent variable across the three species classes, *body_mass_g* shows the divergence in the variance evaluation, particularly for the Gentoo species. Then again, this variable is a weight measured in grams, so perhaps dividing each value by 1000 to measure in kilograms would make the variance seems less divergent. The prior density plot does show the wider distribution among Gentoo species for *body_mass_g*.

The covariance, in which a low number indicates a weak relationship, *bill_length_mm* and *bill_depth_mm*. Two variables, *bill_depth_mm* and *flipper_length_mm*, also show a comparatively low covariance. *bill_length_mm* and *flipper_length_mm* indicates a covariance higher than the aforementioned pairs, but the resulting values across the species classes are much better than the not listed variable pairs.

Assessing the mean of each continuous independent variable for each species class, the variable *bill_depth_mm* highlights the least difference in mean values of four variables. In fact, the difference in mean for Adelie and Chinstrap is one tenth of the measurement.

```
# Means
p_a[,3:6] %>% summarise_each(funs( mean( .,na.rm = TRUE)))
```

```
## # A tibble: 1 x 4
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##            <dbl>         <dbl>             <dbl>       <dbl>
## 1           38.8          18.3               190.       3701.
```

```
p_g[,3:6] %>% summarise_each(funs( mean( .,na.rm = TRUE)))
```

```
## # A tibble: 1 x 4
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##            <dbl>         <dbl>             <dbl>       <dbl>
## 1           47.5          15.0               217.       5076.
```

```
p_c[,3:6] %>% summarise_each(funs( mean( .,na.rm = TRUE)))
```

```
## # A tibble: 1 x 4
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##            <dbl>         <dbl>             <dbl>       <dbl>
## 1           48.8          18.4               196.       3733.
```

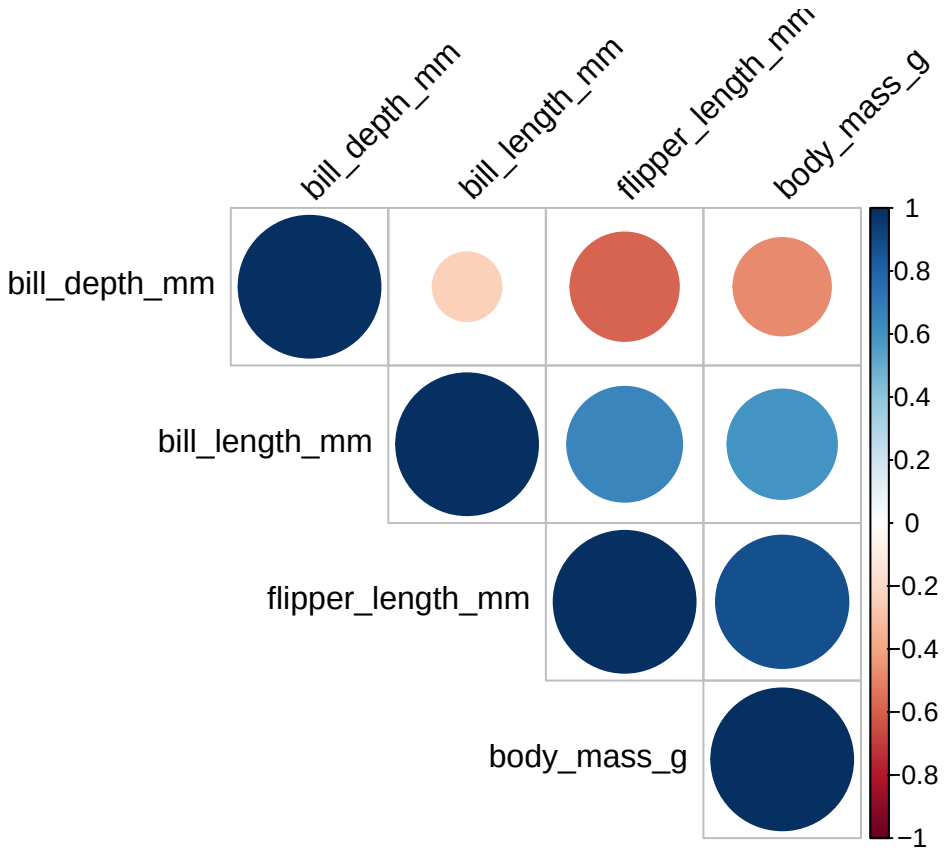
Correlation matrix

Based on the covariance matrix, the correlation matrix should confirm the variable relationships as previously noted. Again, *bill_length_mm* and *bill_depth_mm* show the least correlation. *body_mass_g* and *flipper_length_mm* have a high correlation.

```
# Compute correlation matrix
cor_mat <- cor(penguins[,3:6],  use = "complete.obs")
round(cor_mat, 2)
```

```
##                    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## bill_length_mm               1.00         -0.24              0.66        0.60
## bill_depth_mm               -0.24          1.00             -0.58       -0.47
## flipper_length_mm            0.66         -0.58              1.00        0.87
## body_mass_g                  0.60         -0.47              0.87        1.00
```

```
corrplot(cor_mat, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

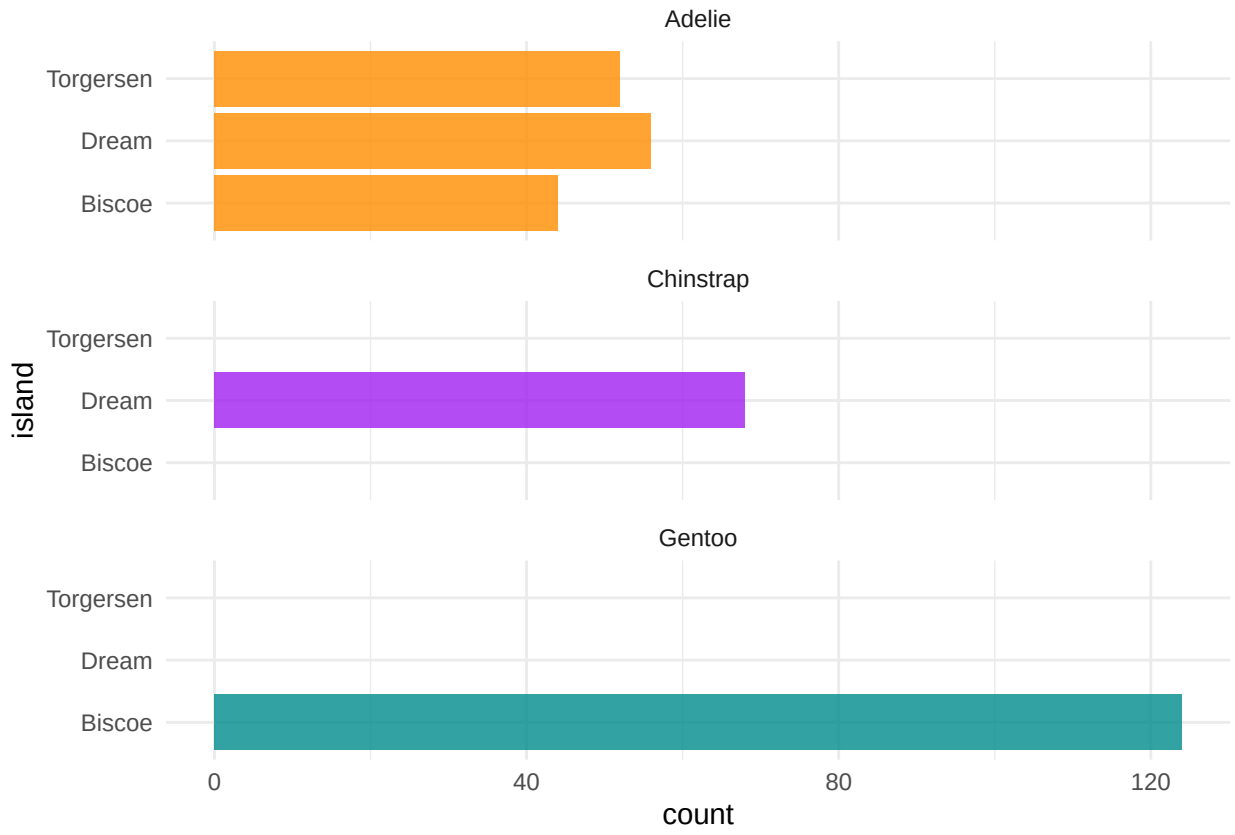


Categorical Variables

For comprehensive exploratory data analysis, the categorical independent variables are plotted and assessed.

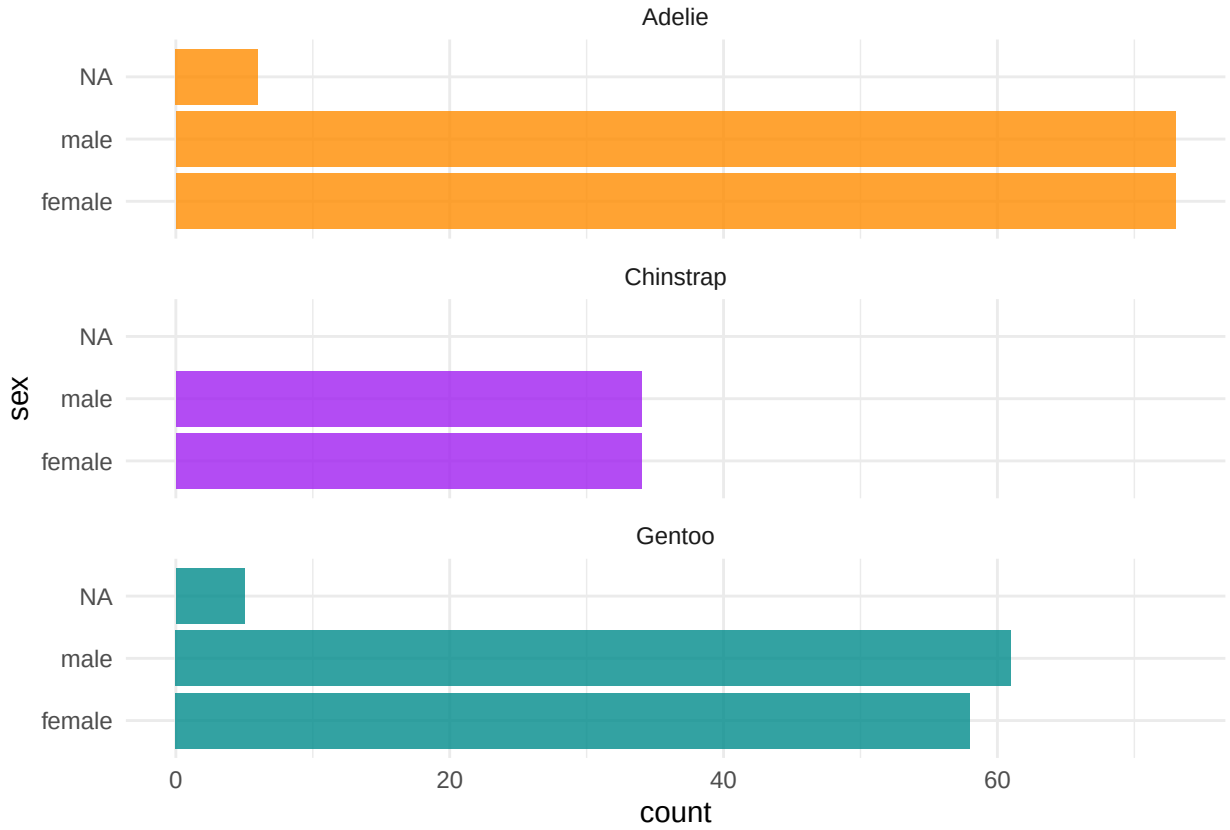
The breakdown of penguin species by island indicates an uneven distribution by island. Gentoo only appear on Biscoe island, Chinstrap only appear on Dream, and Adelie appears on all three island under consideration, Torgensen, Dream and Biscoe.

```
# Count penguins for each species / island
ggplot(ds, aes(x = island, fill = species)) +
  geom_bar(alpha = 0.8) +
  scale_fill_manual(values = c("darkorange", "purple", "cyan4"),
                    guide = F) +
  theme_minimal() +
  facet_wrap(~species, ncol = 1) +
  coord_flip()
```



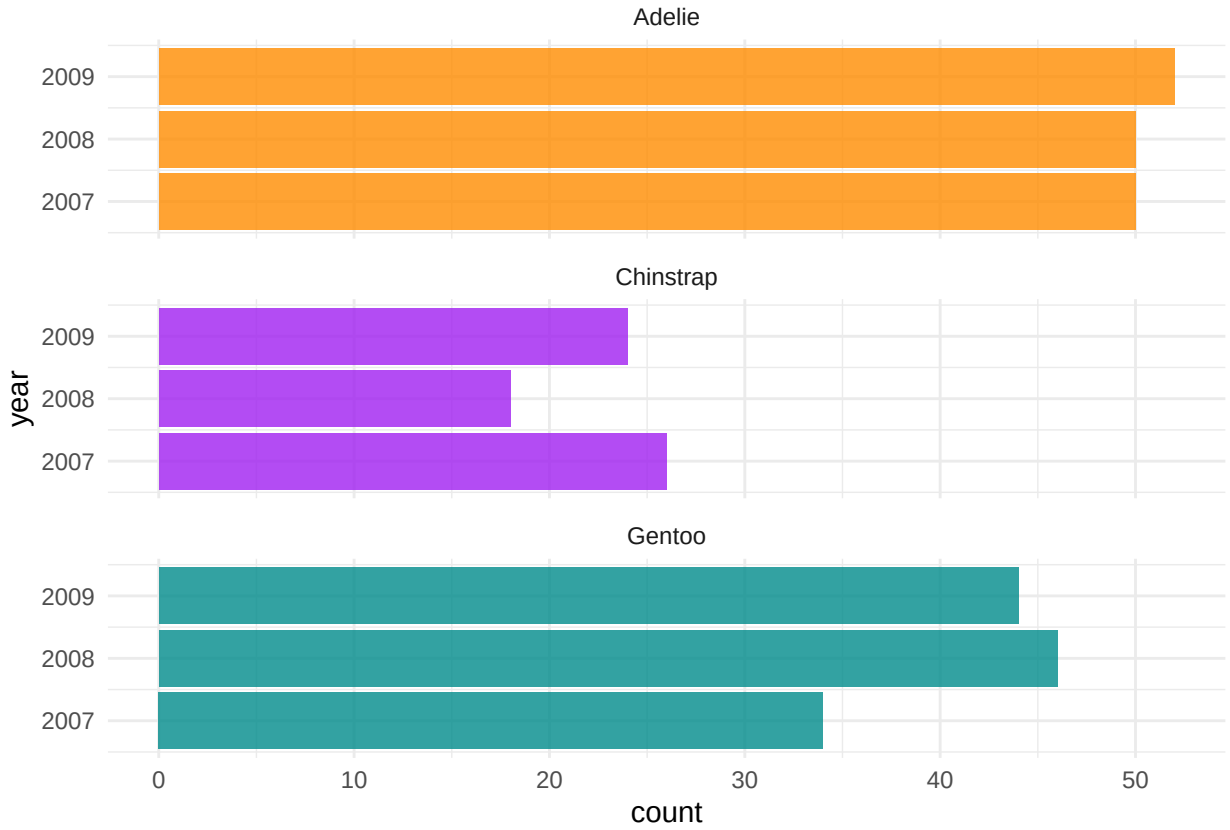
The penguin category for sex shows a clear even distribution by species with a few values missing for Adelie and Gentoo species.

```
# Count penguins for each species / sex
ggplot(ds, aes(x = sex, fill = species)) +
  geom_bar(alpha = 0.8) +
  scale_fill_manual(values = c("darkorange", "purple", "cyan4"),
                    guide = F) +
  theme_minimal() +
  facet_wrap(~species, ncol = 1) +
  coord_flip()
```



The penguin category for year also indicates a relatively even distribution for each species.

```
# Count penguins for each species / year
ggplot(ds, aes(x = year, fill = species)) +
  geom_bar(alpha = 0.8) +
  scale_fill_manual(values = c("darkorange", "purple", "cyan4"),
                    guide = F) +
  theme_minimal() +
  facet_wrap(~species, ncol = 1) +
  coord_flip()
```



As an additional evaluation of the independent variables, a MANOVA test is applies of the four continuous variables along with year, which is defined as numeric. The categorical variables are not applicable for the MANOVA test. The results show a high significance for the four continuous variables while also indicating the year variable is not significant. These results confirm previous understanding of the independent variables.

```
# MANOVA test
manova_res <- manova(cbind(bill_length_mm,bill_depth_mm,flipper_length_mm,body_mass_g,year) ~ species, d
summary.aov(manova_res)
```

```
##  Response bill_length_mm :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## species       2 7194.3  3597.2   410.6 < 2.2e-16 ***
## Residuals   339 2969.9     8.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response bill_depth_mm :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## species       2 903.97  451.98  359.79 < 2.2e-16 ***
## Residuals   339 425.87    1.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response flipper_length_mm :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## species       2  52473 26236.6   594.8 < 2.2e-16 ***
```

```
## Residuals   339  14953    44.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response body_mass_g :
##              Df    Sum Sq  Mean Sq F value    Pr(>F)
## species       2 146864214 73432107  343.63 < 2.2e-16 ***
## Residuals   339  72443483   213698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response year :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## species       2   0.485 0.24228  0.3615 0.6969
## Residuals   339 227.223 0.67027
##
## 2 observations deleted due to missingness
```

Linear Discriminant Analysis (LDA)

Linear discriminant analysis is a classification method that separates classes through linear directions, or linear discriminants. The linear directions are derived as linear combinations of the predictor variables. As outlined above, LDA assumes the feature variables come from multivariate normal distributions and each one continuous. The different classes should also have class-specific means and equal variance/covariance. LDA does not handle categorical data well.

To prepare the data for the LDA model, the observations with missing data are removed from the dataset. Next, the categorical variables sex and year are removed. Based on the exploratory data analysis, neither variable appears to distinguish one species from another. Despite LDA not handling categorical data well, I chose to leave the variable island in the dataset as a trial.

In preparation of evaluating the model's accuracy, the penguins dataset is split into training and set partitions. Finally, the training and test datasets are transformed through a preprocessing step to normalize the data. Given the different range and distribution of the independent variable values, particularly the *body_mass_g* variable, the normalization ensures each variable will be weighted based on predictive ability and not based on initial measurement value.

```
# Load the data
data("penguins")

# Only complete entries
penguins <- na.omit(penguins)

# Remove 'year' and 'sex' feature
# Apparently leaving 'island' in for LDA improves the model
drops <- c("year", "sex")
penguins <- penguins[ , !(names(penguins) %in% drops)]

#Split the data into training (75%) and test set (25%)
set.seed(123)
training.samples <- penguins$species %>%
  createDataPartition(p = 0.75, list=FALSE)
train.data <- penguins[training.samples, ]
test.data <- penguins[-training.samples, ]

#2. Normalize the data. Categorial variables are automatically ignored from normalizing
# Estimate preprocessing parameters
preproc.param <- train.data %>%
  preProcess(method = c("center", "scale"))

# Transform the data using the estimated parameters
train.transformed <- preproc.param %>% predict(train.data)
test.transformed <- preproc.param %>% predict(test.data)
```

Based on the exploratory data analysis, the three variables selected for the LDA model (and the two subsequent models) are *bill_length_mm*, *flipper_length_mm* and *bill_depth_mm*. As the *body_mass_g* variable was highly correlated with *bill_length_mm* and *flipper_length_mm*, the variable was omitted from the model. The categorical variables were also omitted, as LDA performs best on continuous variables.

```
# Fit the model
model.lda <- lda(species~bill_length_mm + flipper_length_mm + bill_depth_mm, data = train.transformed)
```

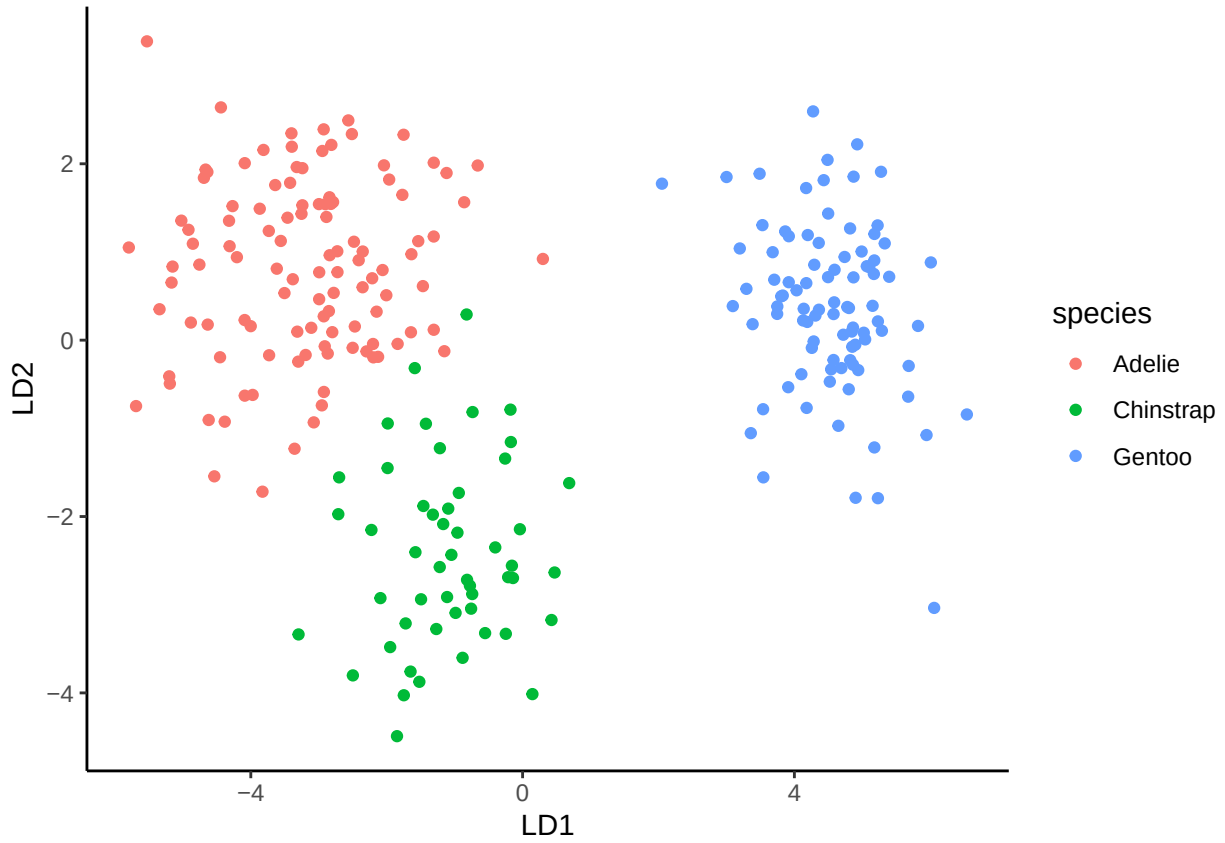
```r
# Output Model
model.lda
```

```
## Call:
## lda(species ~ bill_length_mm + flipper_length_mm + bill_depth_mm,
##     data = train.transformed)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4382470 0.2031873 0.3585657
##
## Group means:
##           bill_length_mm flipper_length_mm bill_depth_mm
## Adelie        -0.9366488        -0.7688194     0.5972070
## Chinstrap      0.8659650        -0.3862079     0.6503085
## Gentoo         0.6540795         1.1585193    -1.0984278
##
## Coefficients of linear discriminants:
##                          LD1         LD2
## bill_length_mm     0.8690417 -2.13504028
## flipper_length_mm  1.6558709  1.53996627
## bill_depth_mm     -1.8612474  0.01119463
##
## Proportion of trace:
##    LD1    LD2
## 0.8862 0.1138
```

When there are K classes, linear discriminant analysis can be viewed exactly in a K-1 dimensional plot. With three classes in the species dependent variables, two linear discriminants are generated.

The plot of the LD1 and LD2 values on the two-dimensional plot shows the classification of the species in the 2D subspace.

```r
# Plot
lda.data <- cbind(train.transformed, predict(model.lda)$x)
ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = species))
```



Now predict the species on the transformed test dataset.

```
# Make predictions
preds.lda <- model.lda %>% predict(test.transformed)

head(preds.lda$x, 3)
```

```
##         LD1        LD2
## 1 -4.540639 -0.2451809
## 2 -2.676708  0.1316585
## 3 -5.237623  0.6623137
```

```
# Confusion matrix
table(preds.lda$class, test.transformed$species)
```

```
##
##             Adelie Chinstrap Gentoo
##   Adelie        35         0      0
##   Chinstrap      1        17      0
##   Gentoo         0         0     29
```

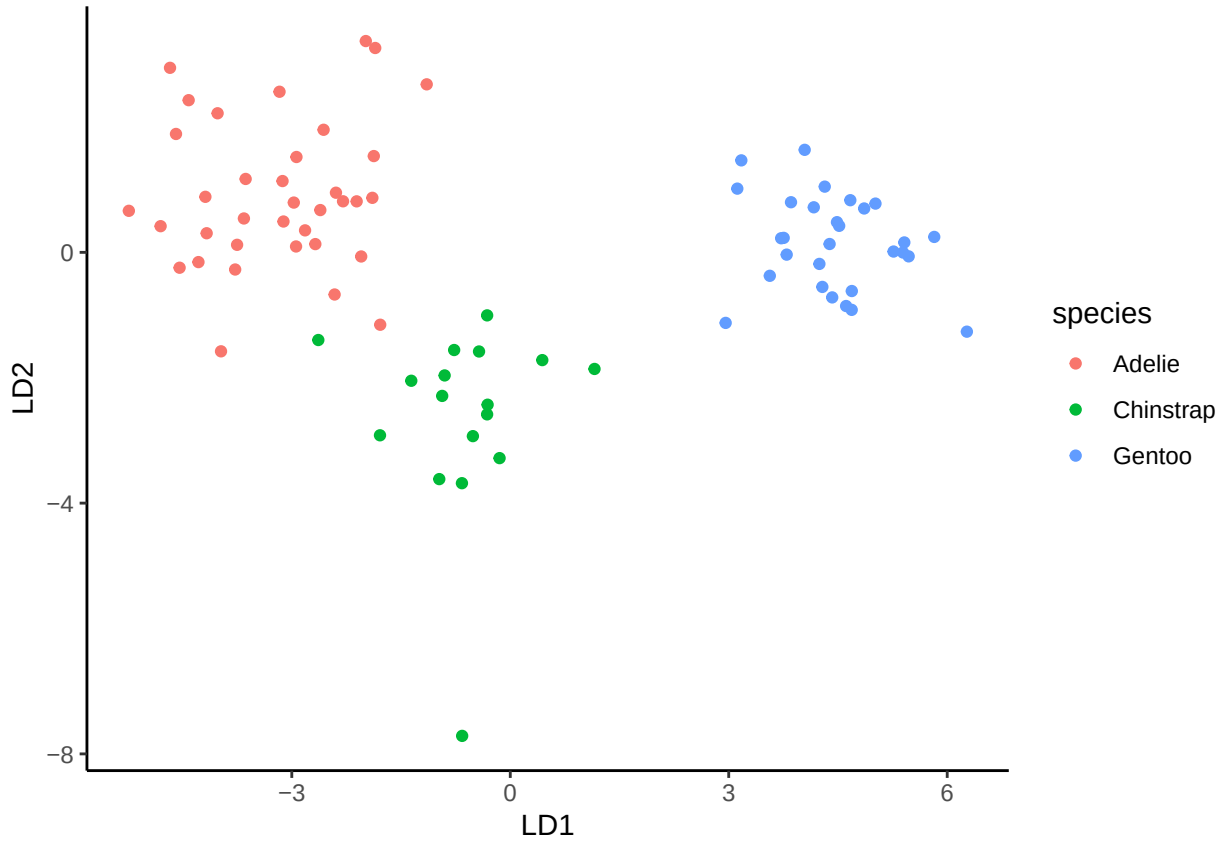
```
# Model accuracy
mean(preds.lda$class == test.transformed$species)
```

```
## [1] 0.9878049
```

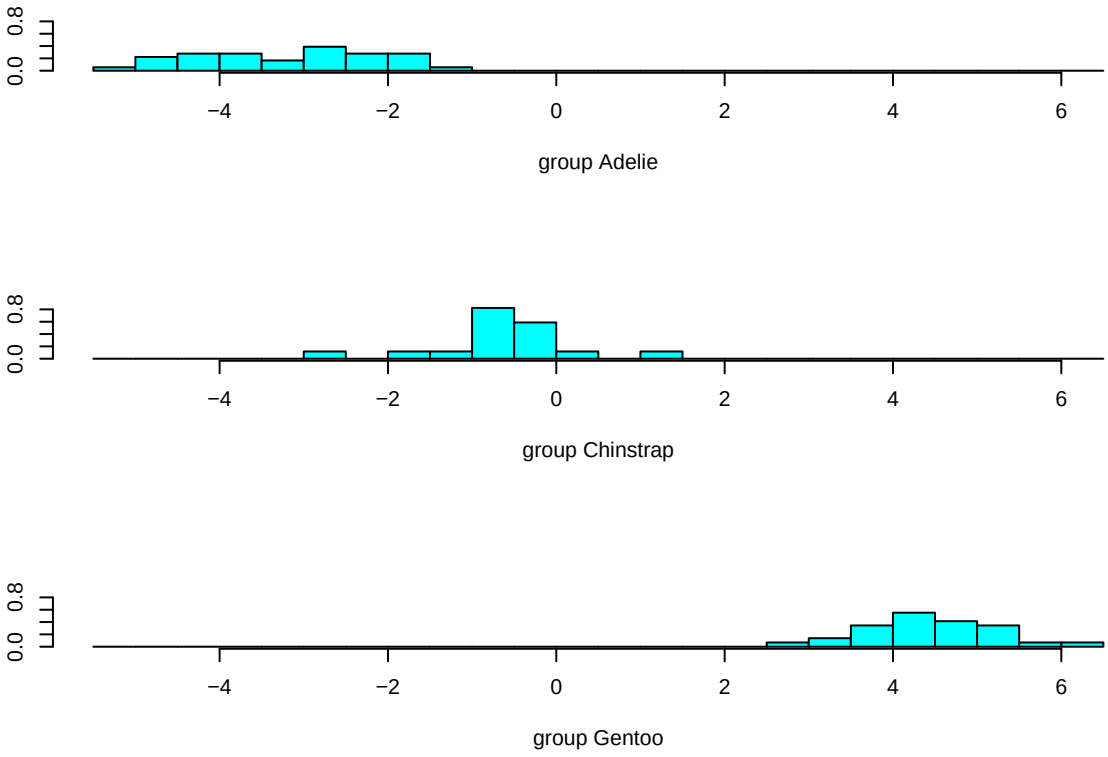
The confusion matrix indicates one incorrect prediction with an overall model accuracy of 98.78%.

The 2D subspace plot of LD1 and LD2 on the test dataset show a similar pattern to the scatterplot on the training data. The plot does show one outlier Chinstrap observation.

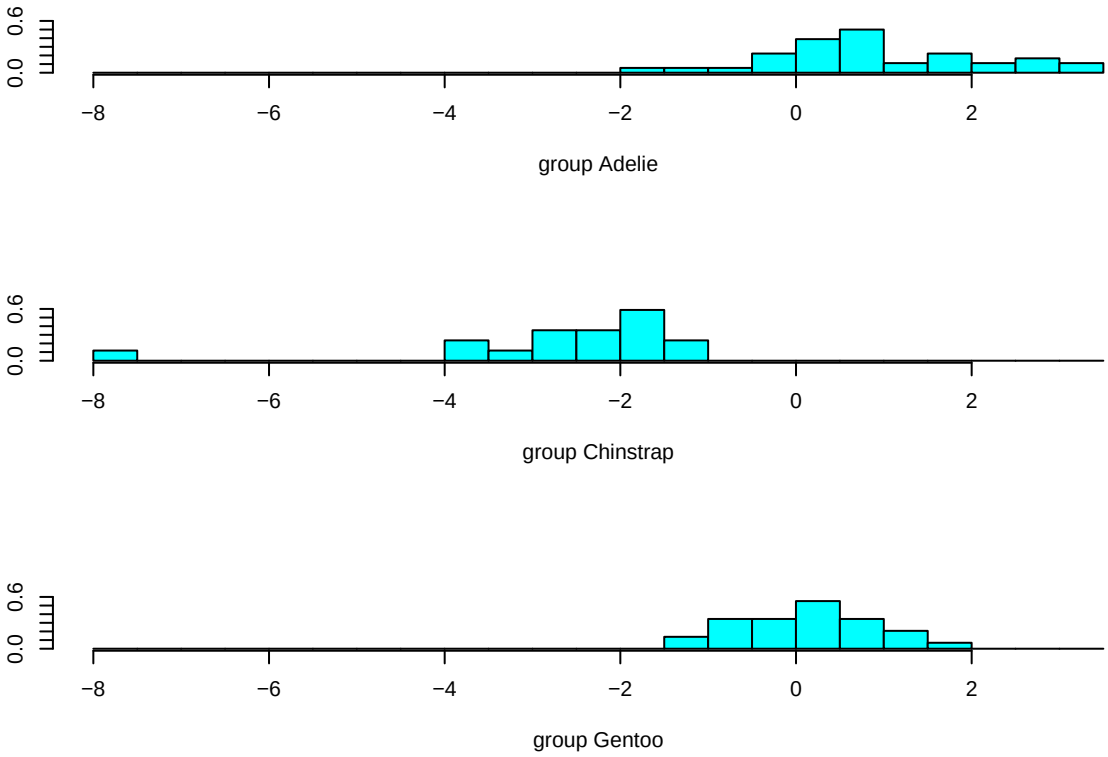
```
# Plot
lda.data <- cbind(test.transformed, preds.lda$x)
ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = species))
```




```
ldahist(data=preds.lda$x[,1], g=test.transformed$species)
```



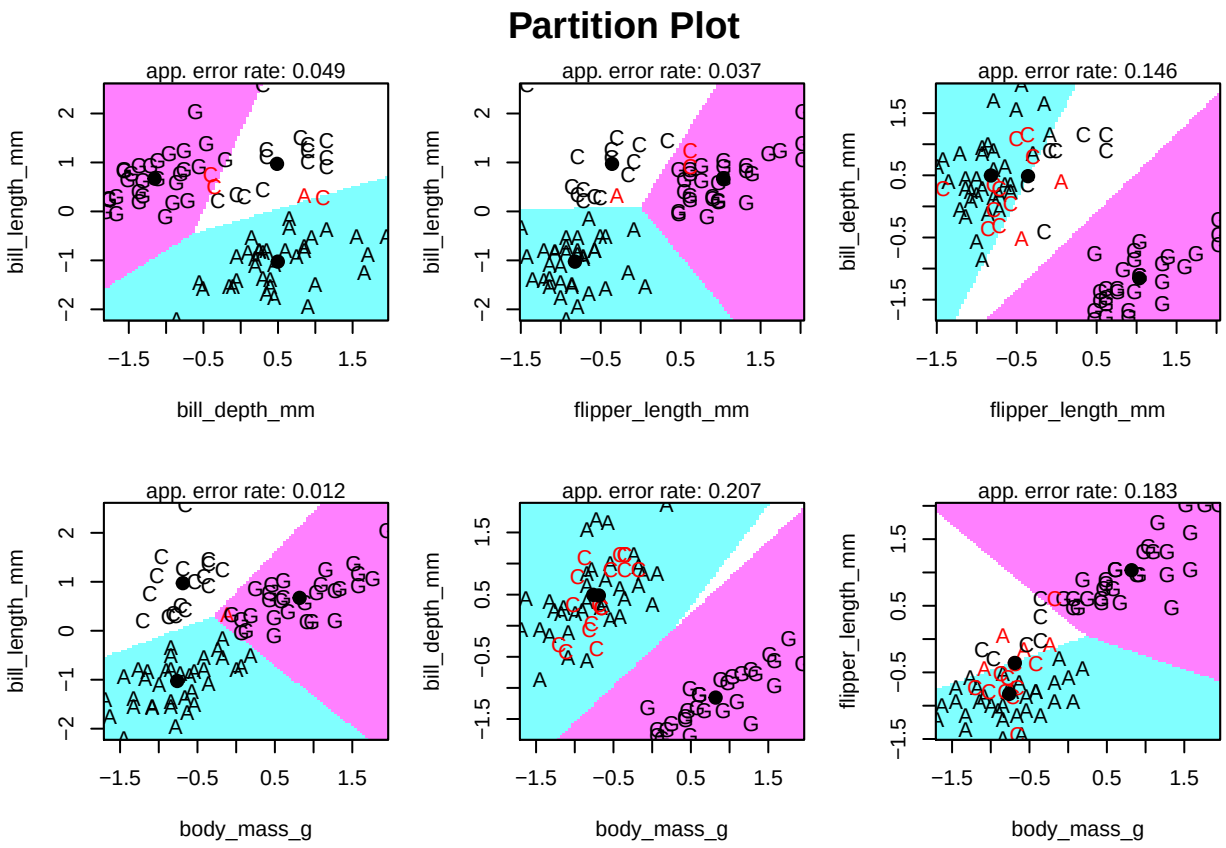

```
ldahist(data=preds.lda$x[,2], g=test.transformed$species)
```



The bar plot for LD1 shows a better performance in classifying among the three species, while the LD2 bar plot shows more overlap among the three species. These plots align with the proportion of trace from the initial model output. LD1 accounts for over 88% of the trace, while LD2 accounts for less than 12%.

The partition plots below from the *partimat* function display the classification of each test observation for each combination of two independent variables. Turns out the plot with the lowest error rate is *bill_length_mm* and *body_mass_g*.

```
partimat(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g, data=test.transfor
```


Partition Plot
app. error rate: 0.049
bill_length_mm
bill_depth_mm
app. error rate: 0.037
bill_length_mm
flipper_length_mm
app. error rate: 0.146
bill_depth_mm
flipper_length_mm
app. error rate: 0.012
bill_length_mm
body_mass_g
app. error rate: 0.207
bill_depth_mm
body_mass_g
app. error rate: 0.183
flipper_length_mm
body_mass_g

Quadratic Discriminant Analysis (QDA)

Quadratic discriminant analysis, similar to linear discriminant analysis, also assumes a multivariate Gaussian distribution, but unlike LDA, QDA assumes each class has a unique covariance matrix. QDA draws the classification distinctions through quadratic decision boundaries, as the name implies, instead of the linear approach of LDA. To allow for even comparison among the three models, the same transformed training dataset is used for the QDA model.

```
# QDA
# Fit the model
model.qda <- qda(species~bill_length_mm + flipper_length_mm + bill_depth_mm, data = train.transformed)

# Output model results
model.qda
```

```
## Call:
## qda(species ~ bill_length_mm + flipper_length_mm + bill_depth_mm,
##     data = train.transformed)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4382470 0.2031873 0.3585657
##
## Group means:
##           bill_length_mm flipper_length_mm bill_depth_mm
## Adelie        -0.9366488        -0.7688194     0.5972070
## Chinstrap      0.8659650        -0.3862079     0.6503085
## Gentoo         0.6540795         1.1585193    -1.0984278
```

The model output for the QDA provides group means equal to the group means provided from the LDA model output. Thus, showing the similarity in behavior between the two model approaches. As the QDA model is quadratic, the model output does not contain coefficients for each independent variable as the LDA model does.

```
# Make predictions
preds.qda <- model.qda %>% predict(test.transformed)

# Confusion matrix
table(preds.qda$class, test.transformed$species)
```

```
##
##             Adelie Chinstrap Gentoo
##   Adelie        35         1      0
##   Chinstrap      1        16      0
##   Gentoo         0         0     29
```

```
# Model accuracy
mean(preds.qda$class == test.transformed$species)
```

```
## [1] 0.9756098
```

The confusion matrix indicates two incorrect predictions with an overall model accuracy of 97.56%. Based on the textbook description of QDA, LDA typically outperforms QDA when relatively fewer training observations are used. Also, as the covariance matrix outlined in the exploratory data analysis, the assumptions are met for the LDA model approach. That being said, the difference in accuracy is based on a single additional incorrect prediction.

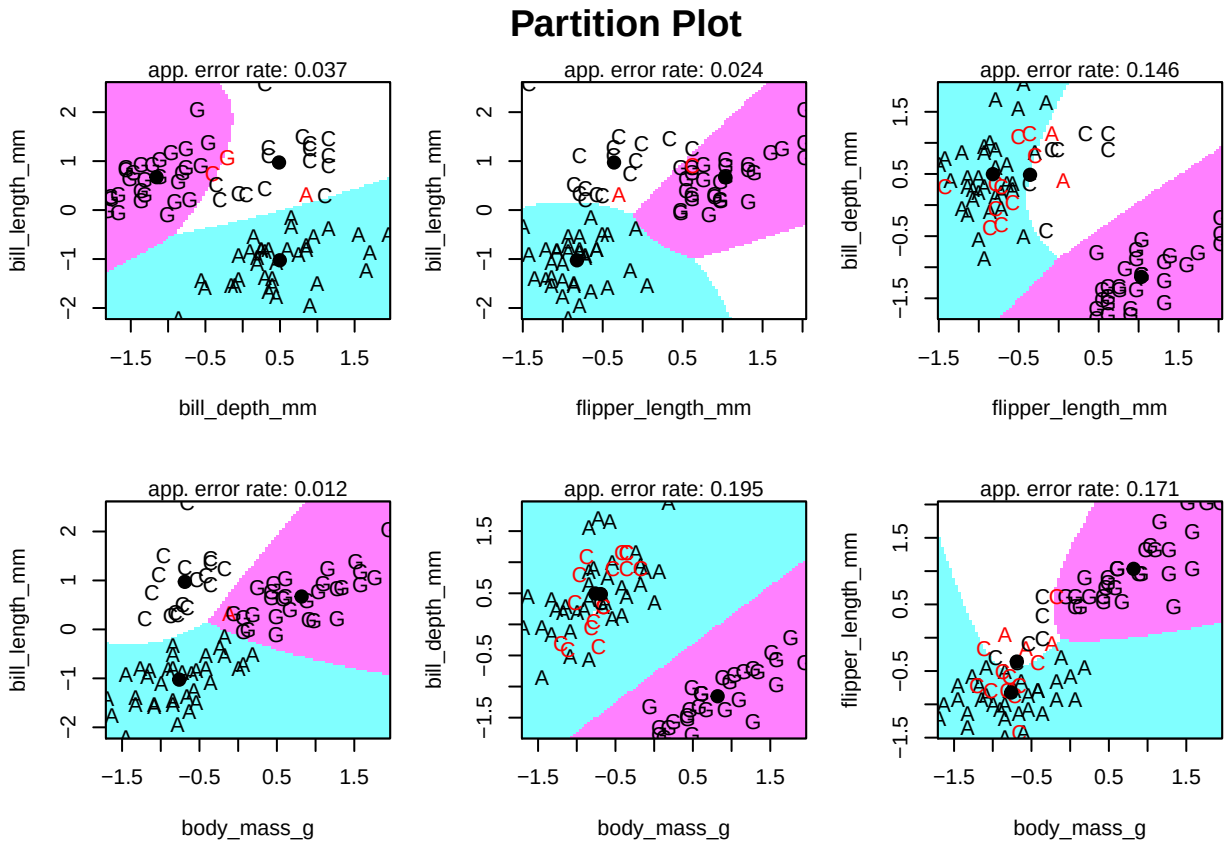
The below table indicates an error rate of 1.22% for LDA and an error rate of 2.44% for QDA.

```
test.transformed %>%
  summarise(lda.error = mean(preds.lda$class != test.transformed$species),
            qda.error = mean(preds.qda$class != test.transformed$species))
```

```
## # A tibble: 1 x 2
##   lda.error qda.error
##       <dbl>     <dbl>
## 1    0.0122    0.0244
```

The partition plots below capture the quadratic nature of the decision boundaries.

```
partimat(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g, data=test.transfor
```



Naive Bayes

The Naive Bayes classifier relies on Bayes theorem of probability. The naive Bayes classifer makes a simplifying assumpiotin which is all predictor variables are conditionally independent of each other in regards to the dependent variable. For numeric features, the Naive Bayes approach makes an assumption that the numerical variables are normally distributed. Due to the assumption of conditional independence, the NB classifier can lead to biased posterior probabilities.

```
# Modified for the penguin data

# Fitting the Naive Bayes model
model.nb <- naiveBayes(species~bill_length_mm + flipper_length_mm + bill_depth_mm, data=train.transforme

# Output the model
model.nb
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##    Adelie Chinstrap    Gentoo
## 0.4382470 0.2031873 0.3585657
##
## Conditional probabilities:
##            bill_length_mm
## Y                 [,1]      [,2]
##   Adelie    -0.9366488 0.4772224
##   Chinstrap  0.8659650 0.6246443
##   Gentoo     0.6540795 0.6034964
##
##            flipper_length_mm
## Y                 [,1]      [,2]
##   Adelie    -0.7688194 0.4873330
##   Chinstrap -0.3862079 0.4932397
##   Gentoo     1.1585193 0.4562644
##
##            bill_depth_mm
## Y                 [,1]      [,2]
##   Adelie     0.5972070 0.6099412
##   Chinstrap  0.6503085 0.5734907
##   Gentoo    -1.0984278 0.5151119
```

Interestlingly enough, the conditional probabilities of the NB model match the group means for the LDA and QDA models, a good sign for the upcoming predictions.

```
# Prediction on the dataset
preds.nb <- predict(model.nb, test.transformed)
# Confusion matrix to check accuracy
table(preds.nb, test.transformed$species)
```

```
##
## preds.nb    Adelie Chinstrap Gentoo
##   Adelie        35         0      0
##   Chinstrap      1        17      0
##   Gentoo         0         0     29
```

```
mean(preds.nb == test.transformed$species)
```

```
## [1] 0.9878049
```

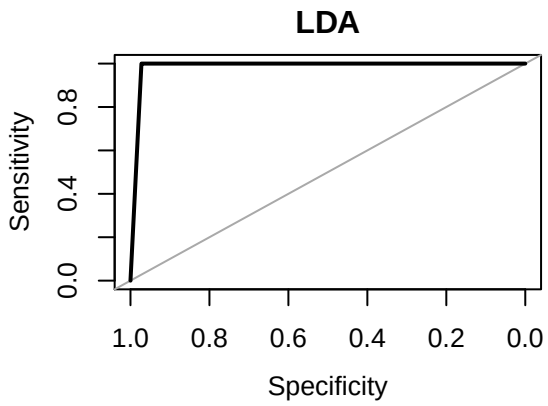
The confusion matrix indicates one incorrect prediction with an overall model accuracy of 98.78%.

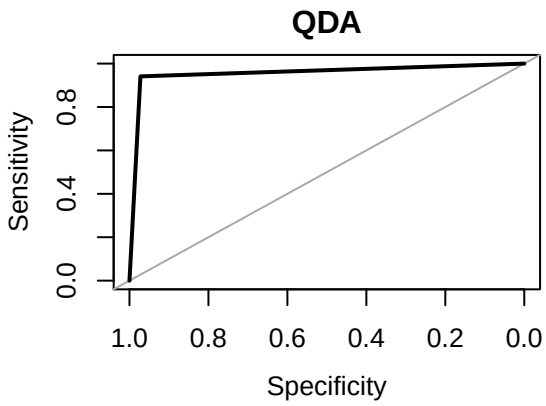
With the same conditional probabilities of the NB model as the group means of the LDA model, the predictions are the exact same. Each model contains one incorrect prediction of an Adelie penguin predicted as a Chinstrap.

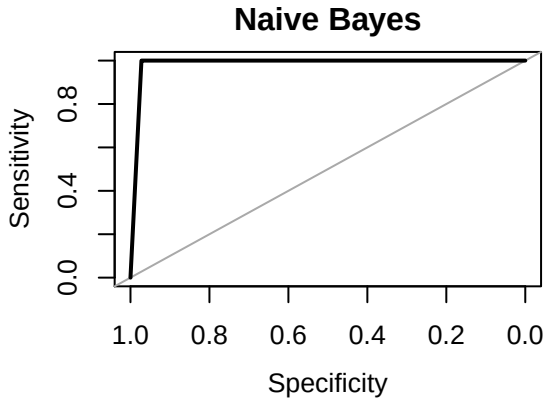
AUC Comparison

Give the three classifier models resulting near identicial group means and conditional probabilities, along with the near identical accuracy, the following area under the curve plots show results just short of 1.0.

```
# ROC curves

par(mfrow=c(2, 2))

preds.lda.num <- as.numeric(preds.lda$class)
roc.multi.lda <- multiclass.roc(test.transformed$species, preds.lda.num)
lda.auc <- auc(roc.multi.lda)
rs.lda <- roc.multi.lda[['rocs']]
plot.roc(rs.lda[[1]], main="LDA",
         xlim=c(1, 0), ylim=c(0, 1), add=FALSE, asp=NA)

preds.qda.num <- as.numeric(preds.qda$class)
roc.multi.qda <- multiclass.roc(test.transformed$species, preds.qda.num)
qda.auc <- auc(roc.multi.qda)
rs.qda <- roc.multi.qda[['rocs']]
plot.roc(rs.qda[[1]], main="QDA",
         xlim=c(1, 0), ylim=c(0, 1), add=FALSE, asp=NA)

preds.nb.num <- as.numeric(preds.nb)
roc.multi.nb <- multiclass.roc(test.transformed$species, preds.nb.num)
nb.auc <- auc(roc.multi.nb)
rs.nb <- roc.multi.nb[['rocs']]
plot.roc(rs.nb[[1]], main="Naive Bayes",
         xlim=c(1, 0), ylim=c(0, 1), add=FALSE, asp=NA)
```







Area under the curve results:

- LDA: 0.9953704
- QDA: 0.9855664
- Naive Bayes: 0.9953704

Conclusion

WORDS_HERE

$@Manual, title = palmerpenguins : PalmerArchipelago(Antarctica)penguindata, author = AllisonMarieHorstandAlison$

==== Prompt =====

```
Homework # 2 (Generative Models) (100 points) Due on March 12, 11:59pm EST
We will be working with the Penguin dataset again as we did for Homework #1. Please use "Species" as yo
Using the target variable, Species, please conduct:
a. LinearDiscriminantAnalysis(30points):
a. Youwanttoevaluateallthe'features'ordependentvariablesandsee what should be in your model. Please comm
```

```
b. Justasuggestion:YoumightwanttoconsiderexploringfeaturePlot on the caret package. Basically, you look
c. Fit your LDA model using whatever predictor variables you deem appropriate. Feel free to split the da
d. Lookatthefitstatistics/accuracyrates.
b. QuadraticDiscriminantAnalysis(30points)
a. Samestepsasabovetoconsider
c. Naive Bayes (30 points)
a. Samestepsasabovetoconsider
d. Commentonthemodelsfits/strength/weakness/accuracyforallthesethree models that you worked with. (10 po
```

http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/

Make sure each variable is normally distributed

https://web.stanford.edu/class/stats202/notes/Classification/LDA.html That is, within each class the features have multivariate normal distribution with center depending on the class and common covariance

In the covariance matrix in the output, the off-diagonal elements contain the covariances of each pair of variables. The diagonal elements of the covariance matrix contain the variances of each variable. The variance measures how much the data are scattered about the mean.

bill_depth, bill_length, body_mass

From: https://rpubs.com/Nolan/298913

plot(model, dimen = 1, type = "b")

https://www.geeksforgeeks.org/linear-discriminant-analysis-in-r-programming/

https://www.geeksforgeeks.org/naive-bayes-classifier-in-r-programming/

https://www.r-bloggers.com/2018/01/understanding-naive-bayes-classifier-using-r/

Focus on normal distributions

If n is small and the distribution of the predictors X is approximately normal in each of the classes, the LD model is more stable than logistic regression

Correlation will cause a line in the Gaussian density plot

References

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://allisonhorst.github.io/palmerpenguins/.