Classification of bulls by breed

160175125

Introduction

The report contrasts different machine learning techniques with the aim to classify groups of bulls, in terms of breed, only using various measurements related to their dimension and body composition. The data set used for the report contains information on the height, weight, body fat, and breed of 76 bulls and is taken from the book "Applied Multivariate Statistical Analysis" by Johnson & Wichern [1] (see Appendix 1 for more information about the data set).

The first section of the report covers a Principle Components Analysis (PCA) of the data. The aim of the PCA was to identify which size characteristics are the most releveant for explaining the variation in the size of the bulls, and also to identify which of those size characteristics can be used to distinguish the three breeds of bull. The next section of the report covers building a logistic regression (LR) classifier. This was done by obtaining a set of LR models that were trained and tested to classify bulls as either Angus or Hereford bulls. The models were collectively tested for accuracy to approximate the mean and variance of a randomly selected models' accuracy. This information was then used to assess the reliability of any specific model, constructed in the same way, to accurately classsify new observational data. The last section of the report covers a Linear Discriminant Analysis (LDA) of the data. The aim was to transform the data into a form which tries to seperate the breeds as much as possible, and find hyperlanes which partition the data which can be used as a classification rule. The model created by the LDA was then tested for accuracy using the method of leave-one-out cross validation. After the accuracy of the model was quantified, it was used to classify new observational data.

PCA Analysis

The analysis began by identifying the number of principle components (PC's) necessary to explain the variation in the size of the bulls. The first 4 PC's are enough to explain approximately 95% of the variation in the data, hence the first 4 components were taken to interpret the plots in figure 1 (see Appendix 2 for the PC variances). Appendix 3 is a screeplot of the PC variances for illustration.

Figure 1 shows two of the six orientations of the data plotted in terms of the principle components (see Appendix 4 for the remaining plots). The gray axes are included to aid the explanation of the data interpretation. The only obvious candidates for potential outliers are the left-most observation, and possibly the bottom-most observation in plot A1, which belong to observations 51 and 16 of the data respectively. Table 1 shows the data for these two observations. The data shows that observation 51 is a particularly large and very lean bull, and observation 16 is a small, but very fat bull. There is no obvious error in the data entry and the numbers seem plausible, therefore these observations were considered influential and were not excluded from the analysis.

Plot A1 shows PC2 plotted against PC1. The two grey diagonal lines are the lines PC1 = PC2 (dashed), and PC1 = -PC2 (dotted). The plot shows a clear split between Simmental bulls, and the Angus & Hereford bulls along the line PC1 = PC2. Therefore using the PC1 = -PC2 line as an axis, the negative and positive scores on this axis identify the size characteristics that explain the difference between the Simmental and non-Simmental breeds. Negative scores on this axis correspond to large bulls with a low amount of body fat. Positive scores correspond to smaller bulls with a larger amount of bodyfat (see appendix 5 for derivation). This implies that Simmental bulls are larger and leaner than Angus and Hereford bulls, and that Angus and Hereford bulls are similar in size and the amount of bodyfat they carry. Note that Hereford bulls are generally only found on the positive side of the PC1=-PC2 axis, whille Angus bulls are distributed evenly over both the positive and negative side. The PC1=PC2 line can therfore be used to explain this difference.

Since it was shown Angus and Hereford bulls are similar in size and bodyfat, this implies they must differ in weight, and that Angus bulls must have the potential to carry more muscle mass than the Hereford bulls.

Plot A2 shows PC4 plotted against PC3. PC3 compares the size of the bull against the lean body weight of the bull. The variation in PC3 is due to the variation in the ratio of size to lean body weight; Hereford bulls are less varied in PC3 which implies a stronger correlation between size and lean body weight for Hereford bulls than the other breeds. Since the Simmental bulls are all large and lean, the variance in PC4 must be due to variation in weight which implies variation in muscularity. This agrees with the Simmental bulls occupying both sides of the PC1=-PC2 line like the Angus bulls in plot A1, which were shown to also vary in musclarity.

In conclusion the PCA revealed that Simmental bulls tend to be the largest and leanest breed of bull; all breeds generally do not vary in the proportion of bodyfat they carry; the Simmental and Angus breed vary in muscularity; the Hereford breed is generally constant in it's size characteristics. The only ways to distinguish the three breeds is by observing bulls that are a combination of both being very large and very lean (corresponding to occupying the left and top triangular quadrants in plot A1 which would distinguish them as Simmental bulls), or smaller and fatter but very muscular (corresponding to occupying the bottom triangular quadrant in plot A1, in which case it would likely be an Angus bull). There is no obvious way to distinguish a Hereford bull from an Angus bull by it's size characteristics.

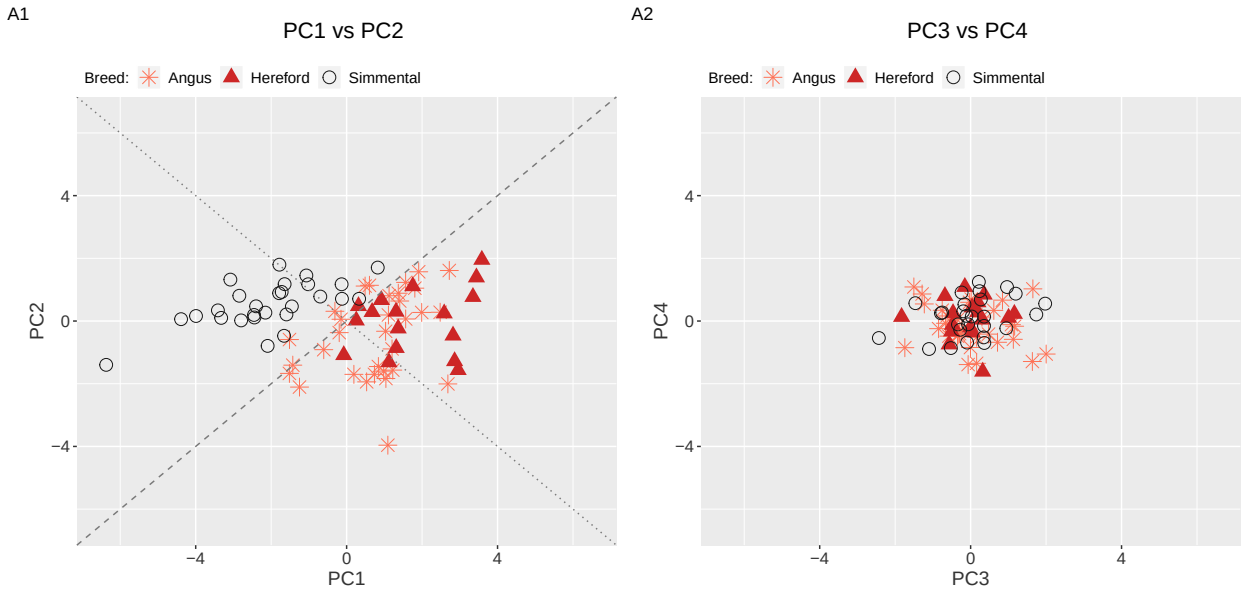

Figure 1: Principle Component Plots. Plot A1 shows PC2 plotted against PC1. The difference between the Simental breed, and the Angus and Hereford breeds is attributed to size and leanness; Simmental bulls are the largest and leanest of the three breeds, while Angus and Hereford are both smaller and fatter, but Angus bulls have the potential to be more muscular than Hereford bulls. Plot A2 shows PC4 plotted against PC3. The lack of variation in the Hereford bulls implies the Hereford bulls are less varied in size and weight than the other breeds.

Logistic regression

This section of the report assesses the ability of LR models to correctly identify which breed of bull new observational data belongs to. Since Angus and Hereford bulls were difficult to differentiate in the PCA, and logistic regression is a binary classification method, the logistic regression focuses only on classifying

Table 1: Outlying observations from plot A1. Both observations are outliers in terms of leanness and size. There is no obvious error in the data entries so the observations were included in the analysis.

	breed	price	frame	height	leanweight	pcntlean	backfat	finalht	finalwt
51	3	1450	8	53.3	1383	81.4	0.2	59.6	1904
16	1	2300	6	49.6	975	68.2	0.5	52.9	1842

Table 2: Averages of the performance of the LR models as a percentage. Contains the mean accuracy, mean sensitivity, and mean specificity. Also contains the absolute value of two sigma distance from the mean for each measurement. The two sigma distance is very large in each case and implies that logistic regression is unreliable for accurately classifying new data if the model is built only using the current data.

	Accuracy	Sensitivity	Specificity
Mean	65.65000	44.83356	77.30591
Two sigma	28.06952	56.31677	33.93984

Angus and Hereford bulls. The data used was identical to the data set in the PCA, except the data for the Simmental breed is excluded. The data was randomly partioned into a training set and a test set where the training set contained around 80% of the observations and the test set contained the remaining 20%. The training set contained the measurements of 39 randomly selected bulls from the measurements dataset; the test set contained the remaining 10 bulls that were not selected for the training set. The LR model was then trained and tested on the data, and the sensitivity, specificity, and accuracy were recorded for the model. This process was iterated 1000 times to provide a good estimate of the mean and variance of the accuracy to determine how reliable a randomly chosen model from the set is.

Table 2 shows the average performance of a set of 1000 regression models, trained and tested on the randomly partitioned data, in successfully classifying new data as either Angus or Hereford bulls. 95% of the time, the real accuracy of a randomly chosen model will be in the interval [37.5% ,93.7%]. The mean accuracy suggests a specific model will likely correctly classify bulls with at least a low amount of accuracy, however the wide interval implies that any particular model from the set cannot reliably classify new data correctly. The sensitivity measures the proportion of Hereford bulls correctly classified as Hereford bulls, and likewise the specificity measures the proportion of Angus bulls correctly classified as Angus bulls. The table shows that the LR models are, on average, able to correctly classify the Angus bulls much better than the Hereford bulls; in fact the models tend to have a bias towards wrongly classifying Hereford bulls as Angus bulls. This imples the model is defaulting to classifying new data as Angus bulls when it is unsure. This bias explains the high success rate of classification for the Angus bulls.

The conclusion of assessment of the LR models is that while it the LR models have the potential to make accurate predictions, they are far too unreliable and biased to be useful machines for classifying new observational data in this case.

Table 3: Confusion table of the LDA model cross validation accuracy. The columns indicate which breed the model classifies the data to be, The rows indicate which breed the data actually belongs to.

	Angus actual	Hereford actual	Simmental actual
Angus prediction	26	5	1
Hereford prediction	11	6	0
Simmental prediction	2	1	24

Discriminant Analysis

This last section covers a Linear Discriminant Analysis (LDA) of the data in an attempt to classify the bulls with a greater reliability than the LR models. The discriminant analysis uses the same data as used in the PCA. The LDA produced a classification model, the accuracy of which was tested using the leave-one-out cross validation technique. This helped to determine how likely a new observation is to be correctly classified.

Table 3 shows the distribution of the predictions of the LDA model in terms of correct, and incorrect predictions. The predictions were obtained after the model had been trained with cross validation. The rows are the distribution of the model predictions for each breed. The number of correct predictions are the diagnoal entries, while the number of incorrect predictions are the off-diagonal entries. For example, the first entry of row 1 shows the model correctly classified 26 observations as Angus bulls, incorrectly classified 5 Angus bulls as Hereford bulls, and incorrectly classified 1 Angus bull as a Simental bull.

The accuracy of the model was estimated by dividing the total number of correct classifications by the total number of classifications and gives a value of `73.7`% for the approximate accuracy of the model which seems noticably better than the LR models. In a similar fashion, the accuracy of the model in classifying each breed was also calculated. The model has a success rate of `81.2`%, `35.3`%, and `88.9`% of correctly classifying Angus, Hereford, and Simmental bulls respectively. The model has a very high success rate at correctly classifying Simental bulls; this is due to the clear seperation between the Simmental breed and the other breeds (see Appendix 6). The model however suffers from the same bias as seen in the LR model, in which the model seems to default to classifying observations as Angus bulls if it is unsure.

A new observation was given to the LDA model of a bull with measurements: frame = 7 units, height at shoulder after one year = 50 inches, fat free body weight = 1000 pounds, percent fat free body weight = 73%, back fat = 0.17 inches, height at shoulder at sale = 54 inches, and weight at sale = 1525 pounds. The model predicts the new data comes from a bull which belongs to the Angus breed, with probability `0.86` (See appendix 7 for the distribution of the observation over the breeds. See appendix 6 for the plot of the data with the new observation). Despite the model bias, the location of the data point on the plot suggests the model prediction is likely correct in this case.

The conclusion of the LDA is that while the LDA model performs better than the LR models, it still suffers from the same bias as the LR models in that it often has difficulty distinguishing between Angus and Hereford bulls. The LDA model however performs extremely well at correctly classifying whether a bull is a Simmental bull, or not a Simmental bull. Visual interpretation of the location of the observations can help informally determine whether the classifications are likely to be correct or not.

Conclusion

Overall, the machine learning techniques had limited success in being able to correctly classify bulls in terms of breed. While the techniques were generally very successful in distinguishing Simmental bulls from non-Simmental bulls, they generally could not distinguish between Angus and Hereford bulls; this is due to the data for those two breeds sharing the same space, i.e Angus and Hereford bulls are very similar in size characteristics on average. This implies that having more observations likely would not improve the accuracy of the models, and that new types of measurements, such as width for example, would need to be introduced to seperate the data in another dimension.

References

[1] R.A.Johnson & D.W.Wichern, "Applied Multivariate Statistical Analysis", Pearson, 6th edition (2015). Acccessed November 2018.

Table 4: Sample of the Bulls data set.

index	breed	price	frame	height	leanweight	pcntlean	backfat	finalht	finalwt	colour	shape
1	1	2200	7	51.0	1128	70.9	0.25	54.8	1720	coral1	8
2	1	2250	7	51.9	1108	72.1	0.25	55.3	1575	coral1	8
3	1	1625	6	49.9	1011	71.6	0.15	53.1	1410	coral1	8
4	1	4600	8	53.1	993	68.9	0.35	56.4	1595	coral1	8
5	1	2150	7	51.2	996	68.6	0.25	55.0	1488	coral1	8
6	1	1225	6	49.2	985	71.4	0.15	51.4	1500	coral1	8

Table 5: Sample of the Bulls data set.

	breed	price	frame	height	leanweight	pcntlean	backfat	finalht	finalwt
	Min. :1.000	Min. : 975	Min. :5.000	Min. :47.20	Min. : 841.0	Min. :64.90	Min. :0.1000	Min. :49.40	Min. :1285
	1st Qu.:1.000	1st Qu.:1375	1st Qu.:6.000	1st Qu.:49.17	1st Qu.: 932.5	1st Qu.:68.60	1st Qu.:0.1500	1st Qu.:52.83	1st Qu.:1474
	Median :2.000	Median :1550	Median :6.000	Median :50.35	Median : 990.5	Median :70.85	Median :0.1500	Median :54.30	Median :1538
	Mean :1.934	Mean :1742	Mean :6.316	Mean :50.52	Mean : 995.9	Mean :70.88	Mean :0.1967	Mean :54.13	Mean :1555
	3rd Qu.:3.000	3rd Qu.:1900	3rd Qu.:7.000	3rd Qu.:51.73	3rd Qu.:1039.2	3rd Qu.:72.25	3rd Qu.:0.2500	3rd Qu.:55.50	3rd Qu.:1648
	Max. :3.000	Max. :4600	Max. :8.000	Max. :54.80	Max. :1383.0	Max. :81.40	Max. :0.5000	Max. :59.60	Max. :1904

Appendix

Appendix 1: The data set.

The Bulls data set was used for the report and contains data of 9 variables for 76 bulls. The variables are the following:

- `breed` 1 Angus; 5 Hereford; 8 Simmental
- `price`: price at sale
- `frame`: scale from 1 [small] to 8 [large]
- `height`: height at shoulder after one year (inches)
- `leanweight`: fat free body weight (pounds)
- `pcntlean`: percent fat free body weight
- `backfat`: back fat (inches)
- `finalht`: height at shoulder at sale (inches)
- `finalwt`: weight at sale (pounds)

For the variables: frame, leanweight, pcntlean, and backfat, the time the measurements were taken is not specified; it is therefore assumed these measurements were taken at the time of sale. Table 4 shows a sample of the dataset. Table 5 shows a summary of the data set.

Appendix 2: The PCA summary.

The output shows the importance of the PC's in explaining the variation in the data. The first 4 components explain almost 95% of the variation in the data.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation     2.0299502 1.1563431 0.8610357 0.6491727 0.4310521
## Proportion of Variance 0.5886711 0.1910185 0.1059118 0.0602036 0.0265437
## Cumulative Proportion  0.5886711 0.7796896 0.8856014 0.9458050 0.9723487
##                            Comp.6      Comp.7
## Standard deviation     0.38275628 0.216925572
## Proportion of Variance 0.02092891 0.006722386
## Cumulative Proportion  0.99327761 1.000000000
```

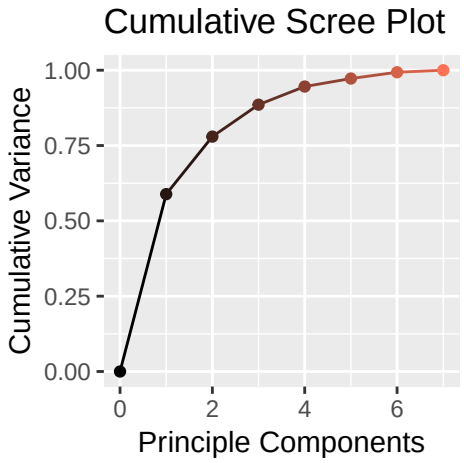

Figure 2: Cumulative Scree Plot. Shows the amount of variation explained in the data as more PC's are taken.

Appendix 3: Cumulative Scree Plot.

See figure 3.

Appendix 4: Remaining orientations of the PCA transoformed data.

See figure 2. None of the plots B1, B2, B3, and B4 revealed any new information about the size characteristics of the bulls, therefore were omitted from the main analysis.

Appendix 5: Derivation of the PC's for plot A1 and A2.

See output 1 and output 2.

Output 2 gives the unit vector parallel to the PC1 = -PC2 axis. Variation on this axis is mainly explained by size and leanness. Larger, leaner bulls will score more negatively.

Output 1 gives the unit vector parallel to the PC1 = PC2 axis. Variation on this axis is mainly explained by variation in weight. Heavier bulls will score more negatively.

Table 6 gives the PC'S in terms of linear combinations of the original variables.

```
# Output 1
loadings(bulls.pca)[,1] + loadings(bulls.pca)[,2]
```

```
##       frame      height  leanweight    pcntlean     backfat     finalht
## -0.44168511 -0.40714111 -0.54216213 -0.04005403 -0.52801456 -0.55416890
##     finalwt
## -0.87046179
```

```
# Output 2
loadings(bulls.pca)[,1] - loadings(bulls.pca)[,2]
```

```
##      frame     height leanweight   pcntlean    backfat    finalht
## -0.4262287 -0.4927215 -0.2824890 -0.6710696  0.9014242 -0.3515387
##    finalwt
##  0.3305679
```

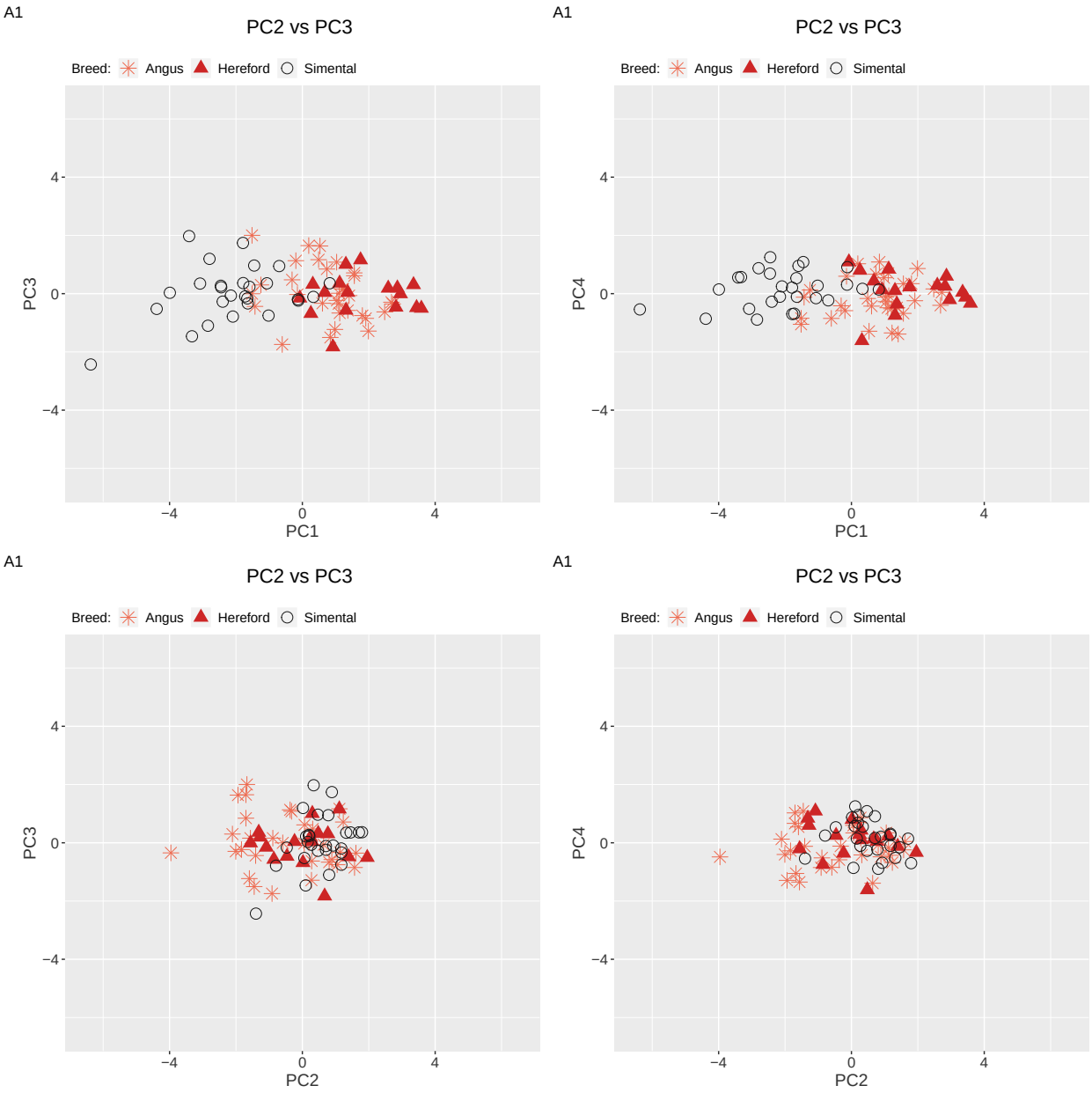

Figure 3: Principle Component Plot

Table 6: PC's as a combination of the original variables.

	Comp.1	Comp.2	Comp.3	Comp.4
frame	-0.4339569	-0.0077282	0.4523450	-0.2428179
height	-0.4499313	0.0427902	0.4157089	-0.1133565
leanweight	-0.4123256	-0.1298365	-0.4502924	-0.2474787
pcntlean	-0.3555618	0.3155078	-0.5682731	-0.3147874
backfat	0.1867048	-0.7147194	0.0387320	-0.6181171
finalht	-0.4528538	-0.1013151	0.1766504	0.2157694
finalwt	-0.2699470	-0.6005148	-0.2533119	0.5824327

Appendix 6: LDA data plot with new observation classification.

See figure 4.

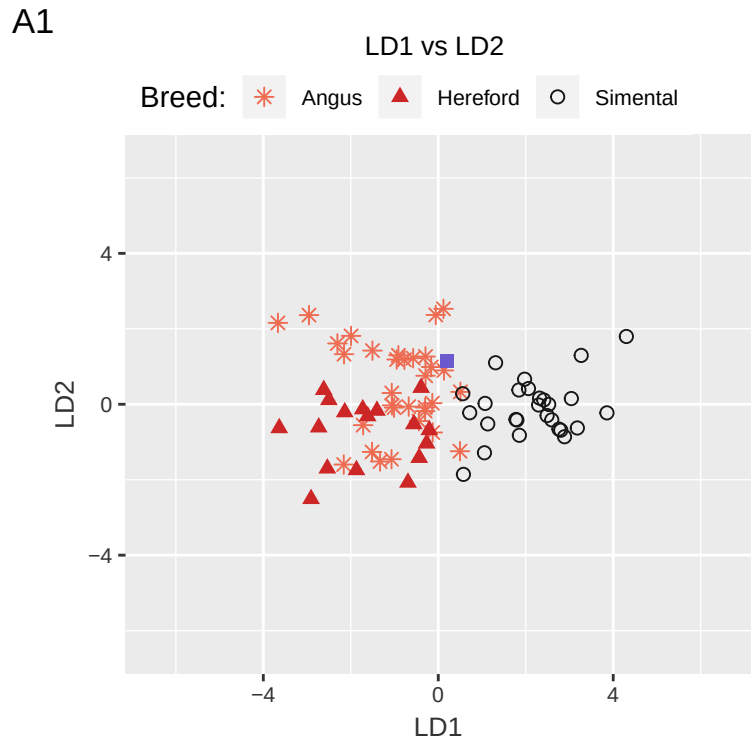

Figure 4: Linear Discriminant Plot. The plot shows LD2 plotted against LD1. The plot shows the transformation of the data where the breeds are most split apart from each other. The blue dot is a new observation of a bull of unknown breed. Observing the location suggests it is likely to be an Angus bull, which is in agreeance with the posterior probabilities shown in Appendix 7.

Appendix 7: Probability distribution of the new observation.

See output 3.

The posterior probabilities show the certainty in which the new datapoint is classified. The model thinks the new observation is likely to be an Angus bull.

```
#Output 3
new.pred$posterior
```

```
##              1          2         3
## [1,] 0.8550122 0.03096714 0.1140206
```