ASSIGNMENT 1

ADAM H WISE

Part 1

1. The correlation between variables V8 and V9 is -0.0064457; This is small enough to suggest there is nothing relating the two variables.

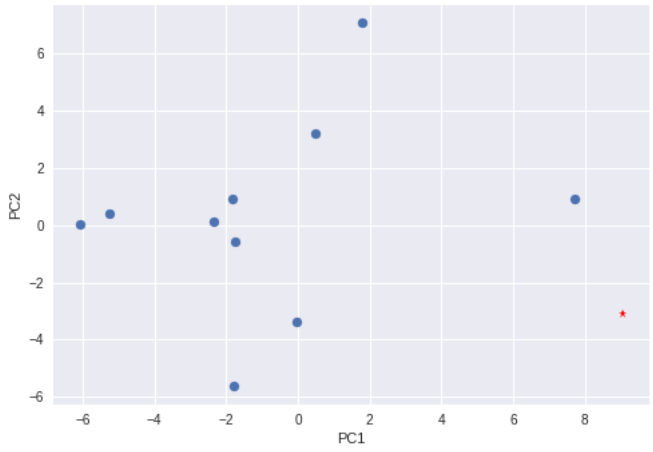
2. Variables V3 and V4 have the greatest positive correlation with a value of 0.579867. However the largest absolute correlation is between V4 and V6 with a value of -0.622541.

3. Using the pandas package allows the generation of a covariance matrix.

	V1	V2	V3	V4	V5	V6	V7	V8	V9
V1	0.272727	0.2	0.009091	0.018182	-0.618182	-0.3	-0.063636	0.081818	0.645455
V2	0.200000	5.4	-1.900000	-1.100000	0.200000	1.7	-3.400000	-0.600000	-2.500000
V3	0.009091	-1.9	4.963636	3.527273	-1.927273	-2.2	2.954545	-1.727273	0.618182
V4	0.018182	-1.1	3.527273	7.454545	-3.354545	-4.3	-0.890909	-0.954545	4.836364
V5	-0.618182	0.2	-1.927273	-3.354545	5.654545	1.6	-1.309091	-0.345455	-4.236364
V6	-0.300000	1.7	-2.200000	-4.300000	1.600000	6.4	-0.800000	-1.000000	-2.100000
V7	-0.063636	-3.4	2.954545	-0.890909	-1.309091	-0.8	6.218182	0.190909	-0.527273
V8	0.081818	-0.6	-1.727273	-0.954545	-0.345455	-1.0	0.190909	2.054545	-0.036364
V9	0.645455	-2.5	0.618182	4.836364	-4.236364	-2.1	-0.527273	-0.036364	15.490909

By either converting this to a numpy array or by summing the main diagonal, the trace of this matrix can be found to be 53.909091.

4. The covariance matrix above seems to suggest that we shouldn't need to normalise our data before we find the principle components, due to the small relative difference in size between the variables. Plotting the first two principle components against each other gives the following graph:



Observation 11 is denoted by the red star at the rightmost extreme of the graph.

5. 41.62% of the information in the data set is given by the first principle component, 63.01% is given by the first two principle components, 78.27% is given by the first three, and 87.956% is given by the first four principle components.

Part 2

The data will need some editing before it can be meaningfully interpreted. Firstly, the columns for breed, price and the arbitrary numbering of each bull must be removed and then the data will need to be normalised as the size between weight and height for instance is great enough to misrepresent the data.

1. By computing the cumulative sum of the PCA's explained variance ratio we can observe that 5 principle components are sufficient to explain 95.3% of the data. For our purposes this seems sufficient to be able to make meaningful predictions about the data at a later time.

2.

	PC1	PC2	PC3	PC4	PC5
frame	0.433957	0.007728	-0.452345	0.242818	0.142995
heightyear	0.449931	-0.042790	-0.415709	0.113356	0.065871
fatreebody	0.412326	0.129837	0.450292	0.247479	-0.719343
pcntfree	0.355562	-0.315508	0.568273	0.314787	0.579367
backfat	-0.186705	0.714719	-0.038732	0.618117	0.160238
finalht	0.452854	0.101315	-0.176650	-0.215769	-0.109535
finalwt	0.269947	0.600515	0.253312	-0.582433	0.290547

From the loadings data-frame, we can see that the first principle component is mostly determined by the bull's frame, height at one year, fat free body weight, percentage of fat free body weight, and it's height at the time of the sale. These all have coefficients of approximately 0.4, with final height having the largest coefficient of 0.453 and final weight appearing to be the least significant with a value of 0.270. The only negative value in the loading for the first principle component is the amount of back fat of the bull.

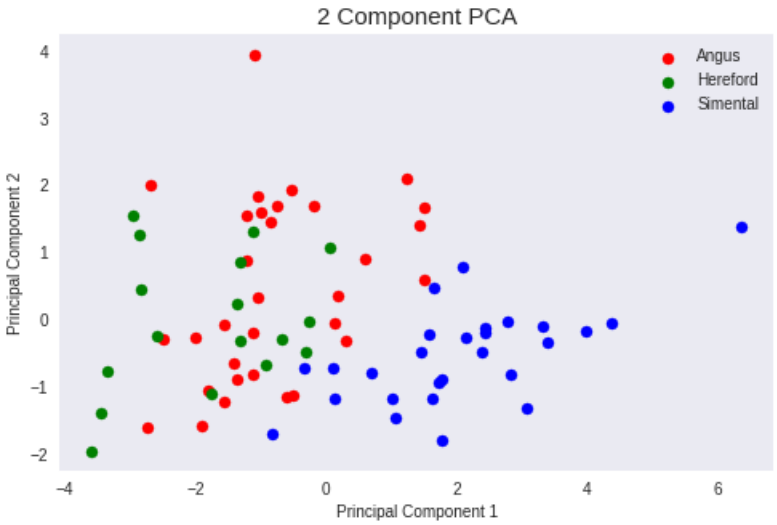
3. Firstly, we need to create a logistic regression model to understand which features of the bull affect its breed. We can view this model's coefficients to determine the effect that each feature has on determining the breed.

	coef
frame	-0.012986
heightyear	-0.373395
fatreebody	-0.300857
pcntfree	0.203534
backfat	1.134002
finalht	0.066536
finalwt	-0.014614

From this table, we can see that the each of the characteristics affects the linear model by a scale of its value. The size of the coefficient therefore determines how much each characteristic affects the model and therefore how much these characteristics affect the likelihood of a bull being a certain breed. We can interpret these by looking at the size of the coefficients.

The most significant factors appear to be: the bull's height after one year, the fat free body weight, the percentage of fat free body weight, and overwhelmingly the amount of back fat on the bulls which is over 3 times more impactful on the model than the other factors.

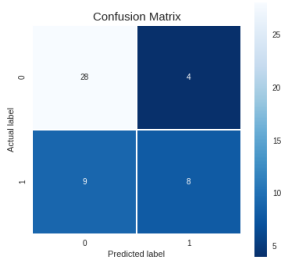
4. One way to search for outliers would be to use this linear model to predict which categories each bull would fit into and identify the bulls that are not a match with their proper breed. This method actually returns quite a few outliers, namely the bulls : 6, 11, 18, 25, 32, 33, 35, 36, 37, 39, 40, 42, 43, and 73. Much of these outliers were Hereford bulls being misclassified as Angus bulls suggesting a similarity between the two. A graph of the principle components may help to identify why this is the case.



While 2 principle components only hold about 78% of the information about the bulls, this graph is somewhat useful in explaining what's happening. Firstly, it shows a lot of overlap between the traits of Angus and Hereford bulls. This similarity seems to explain why the linear model misidentified many Hereford bulls as Angus bulls. Secondly, in the graph there are two obvious outliers: Bull number 15 with a second principle component equal to 3.96 and bull number 50 with principle component 6.37.

Bull 15 and Bull 50 appear to be outliers due to their very low fat free body weight: 893lbs and 844lbs respectively.

5. Begin by transforming the data in a similar way to before and eliminating the Simental bulls from the dataset. One way to test how successful the classification is, is to see how the model predicts the classes of the original dataset. Due to the similarity between the two breeds of bull that was discovered earlier, and the relative smallness of the dataset we expect this to be fairly poor, however. In fact, of the 50 bulls in this new dataset, 13 have been misclassified by this method. For further information, we study the confusion matrix.



The confusion matrix shows that while the probability of an Angus bull being identified as an Angus bull is quite likely (with 28 of the Angus bulls in the dataset projected to be correctly identified), this is far less accurate in the other situations. For instance, the confusion matrix shows that the model correctly identified 8 Hereford bulls as being Hereford bulls but identified 9 Hereford bulls as being Angus bulls.

This is therefore a fairly poor model, likely due to the small size of the dataset .

6. We can split the data into two sets, a training set and a test set, training the model on the training data to create a model that predicts the test set. After generating this we wish to see how accurate the model will be for new test data. The aim is to generate new models until one emerges that is sufficiently accurate. When this has been done the new test data will be entered into the model. Doing this returns that the model suspects that the new entry is a Simental bull. It is worth noting however that the model has a score of 0.803 and therefore may not be accurate.

This is confirmed as the proportion of correct predictions is only 0.735. Therefore while this is potentially a correct model a larger data set would be ideal.