Intro

Logistic Regression. The correlation between $V8$ and $V9$ is 0.357 ($3dp$), this is a weak positive correlation.

2. The pair of variables that have the greatest correlation are $V3$ and $V7$ with a positive correlation of 0.631 ($3dp$).

3. The trace of the variance matrix is 53.473 ($3dp$).

4. The data is fitted to the principal component model and the 11 observations are plotted according to the first and second PCs, including my registration number (red).

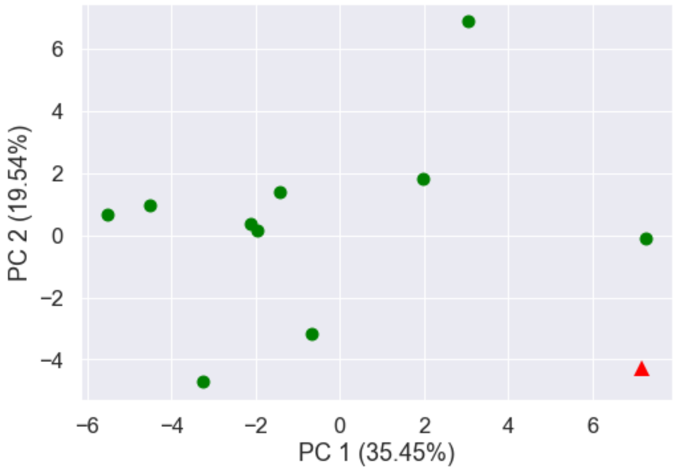

Figure 1. Scatter plot showing the first two principal components

5. The first 4 principal components make up 82.881% of the information in the data set.

6. If we suppose $V1$ is a class variable and let $Ob2, ..., Ob11$ be our training set, then after performing linear discriminant analysis we want to see whether $Ob1$ is predicted correctly. By fitting our new data set (excluding $V1$ and $Ob1$) and setting our target vector to be the $V1$ column of the training set, we see that the predicted class from our LDA is 2. The posterior probability of this being the case is 1.0. As wee see from the classes of the entire data set, $Ob1$ is in class 1 and therefore the LDA prediction is wrong. This is most likely due to the small number of observations.

Part Two

Rubies from three countries have their price evaluated depending on a variety of characteristics, specifically being the analysis of the variability of the rubies on the qualities ('color', 'diameter', 'thickness', 'angle', 'cut', 'clarity' and 'caratwt". Throughout this section, the plots referred to in the analysis will be coded as: 1.Burma (●), 2.Thailand (●) and 3.Cambodia (○), unless stated otherwise.

1. The quality of a ruby is based upon all variables other than location and price. From the cumulative variance graph [figure 2] it is shown that the first 3 principal components are responsible for 58.87%, 19.10% and 10.58% of the total variance respectively. The first two components account for 77.97% of the total variance, this suggests that the data is mostly spread out in two dimensions. However, we want more than 77.97% in order to obtain an appropriate model, so the by taking the first 3 components instead we can then account for 88.56% of the total variance of the data. This reduces our dimensionality from 7 to 3, allowing us to manipulate and visualise the data more easily and it can then also be stored in less space.

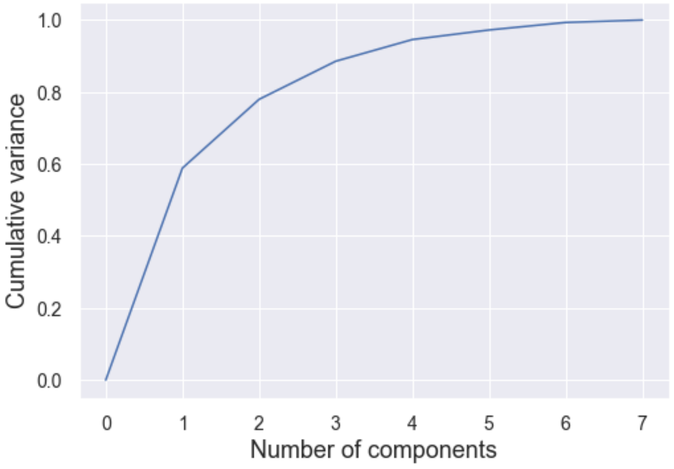

FIGURE 2. Scree plot of cumulative variance against the principal components

2. In order to describe the main sources of variation amongst the qualities of the rubies, we will look at the 7 principal components calculated from the variance matrix and look at their signs.

variable	$a1$	$a2$	$a3$	$a4$	$a5$	$a6$	$a7$
color	+		-	+		+	-
diameter	+		-				+
thickness	+		+	+	-		
angle	+	-	+	+	+		
cut	-	+		+		-	
clarity	+			-		-	-
caratwt	+	+	+	-	+	+	

From this we see that the first principal component ($a1$) is responsible for 58.8% of the total variance and is a made up of high values in all variables except 'cut'. This suggests that PC1 reflects the general shape and aesthetic qualities of the ruby. This would suggest that if a ruby had a high value for the first principal component, then it would generally be a gem of larger size and clearer red colour.

The second component ($a2$) makes up 19.1% of the total variation and consists of the carat weight and the quality of the cut of the ruby. Therefore for a ruby to have a high PC2 value, it would need high values in both of these variables with a low value in the sharpness measurement.

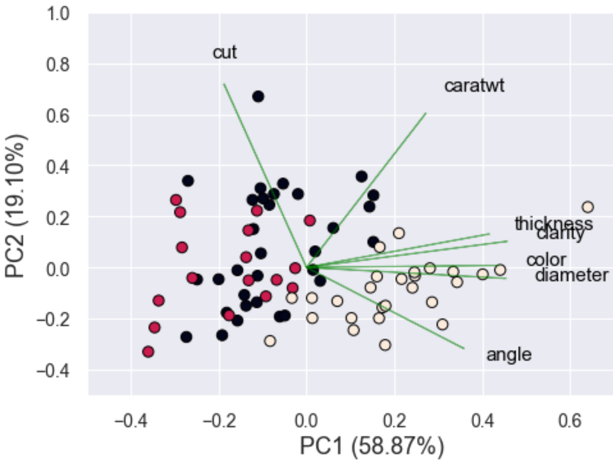

FIGURE 3. Scatter plot of the first two principal components

Another 10.59% of the variation is from the third principal component ($a3$). In order for a ruby to have a high value, it would need high values in variables thickness, angle and carat and low values in the variables color and diameter. This would suggest that these rubies are heavy, short and wide gems with a light red color.

3. Our data is multivariate and therefore the use of a pairplot allows us to view multiple bivariate distributions in the dataset and the patterns in the relationships between the variables. After looking at the full pairplot, the variables 'price', 'angle' and 'caratwt' were taken out as these gave no useful relationships.

In order to visualise the characteristics of each country's rubies, box plots can be used to gain information about the mean and variance of the variables in each country. Figures 5 and 6 show the variables 'diameter' and 'color' and these tell us that rubies from Cambodia (3) have the highest average ’diameter’ and ’color’ score. From similar box plots not in this document, it is also found that Cambodia (3) has the highest average ’thickness’, ’cut score’ and ’clarity’. This suggests that rubies from Cambodia (3) are generally larger and more aesthetically pleasing, but are a lower quality gem. Applying the same process to the rubies from Burma(1), figure 4 shows that the clustering of ’cut’ scores are higher than that of the other countries. Suggesting that rubies from Burma (1) generally have a higher quality of cut. The rubies from Thailand (2) are shown to have the lowest average ’diameter’, ’thickness’ and ’clarity’ score, suggesting that they are smaller and of a lighter shade of red than the other countries. Figure 4 shows that comparing these characteristics can distinguish the ruby’s country of origin. The pairplot is very difficult to read from when the number of variables is high and so the principal component analysis in figure 3 is a much better way to visualise trends and clusters within the data. From this, it is shown that there is a distinct separation between classes and it is clear that the characteristics mentioned above are accurate.

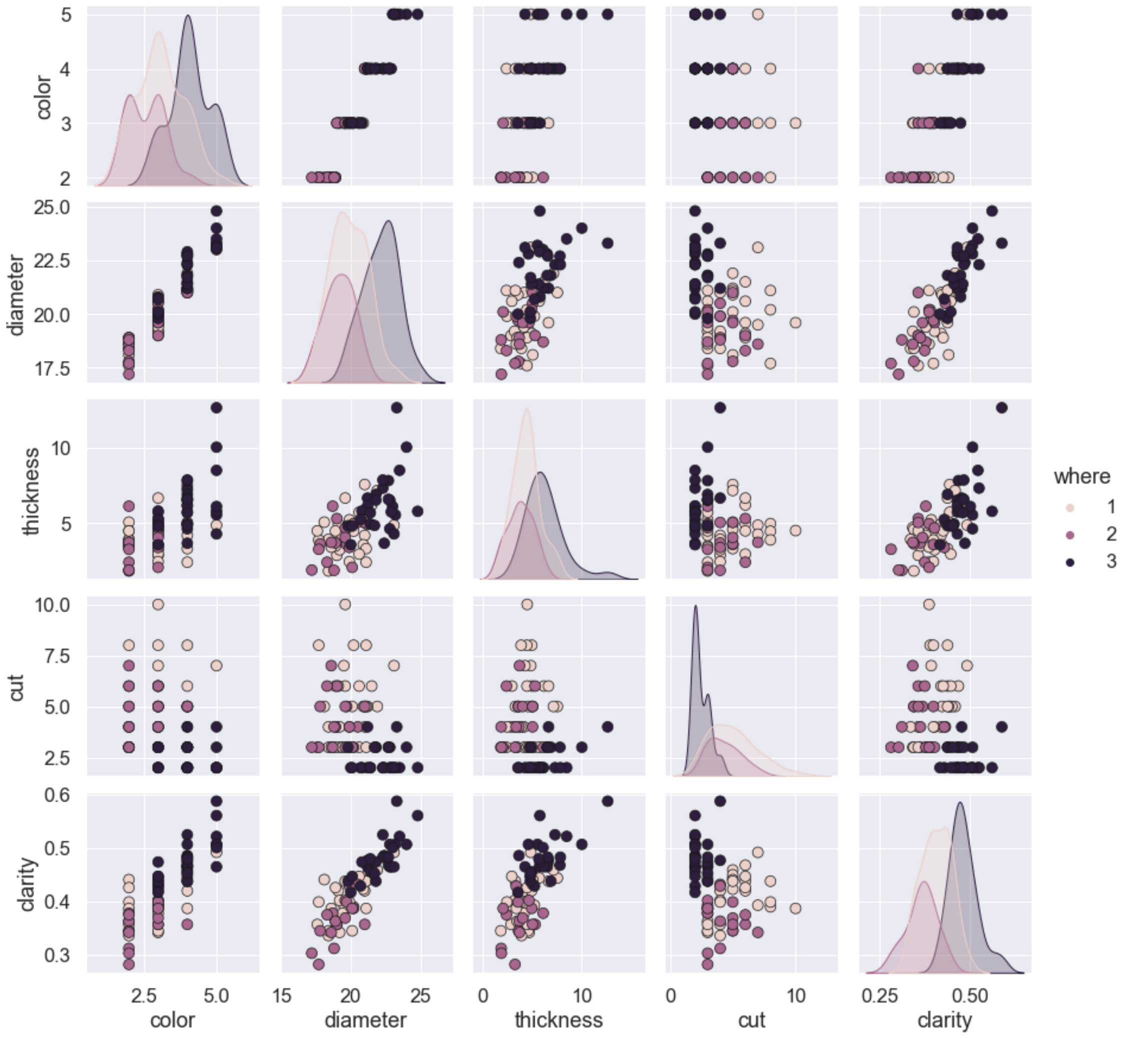

Figure 4

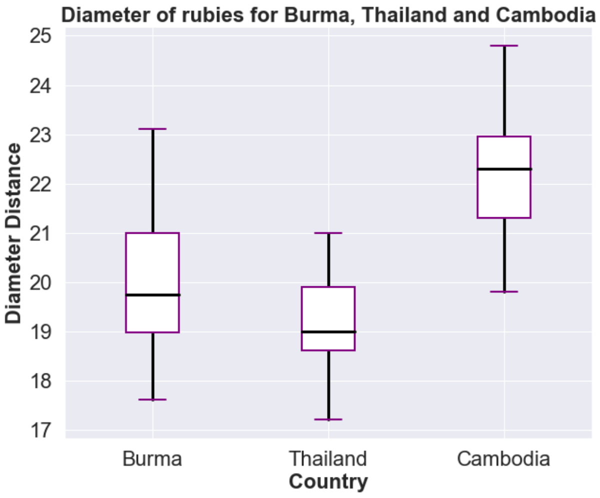

Figure 5

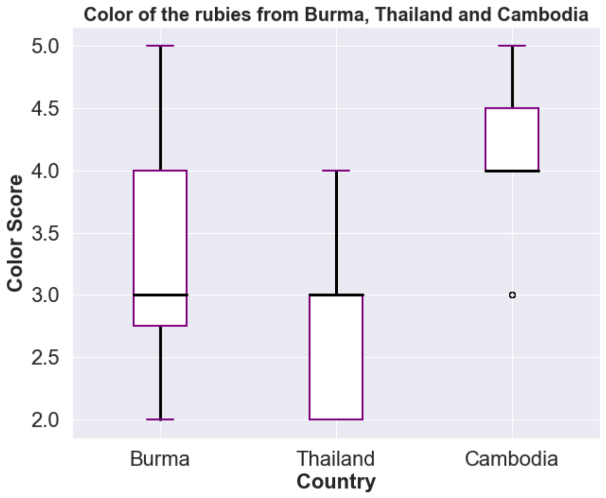

Figure 6

4. Figure 7 shows examples of possible outliers. This shows the 'thickness' of the rubies plotted against the country of origin. In this plot two values from Cambodia (3) lie on the far right hand side, which have been highlighted in green. They both have a thickness measurement that is higher than average, however, as our data set is not large for each country, it is not certain that these values are outliers. By removing these from the data, it is presumed that these values are unlikely to occur again in the 'thickness' variable of rubies from this country, even though they have occurred in this data set. There are a few extreme values in other variables but none that are significant enough to be disregarded from the data set.

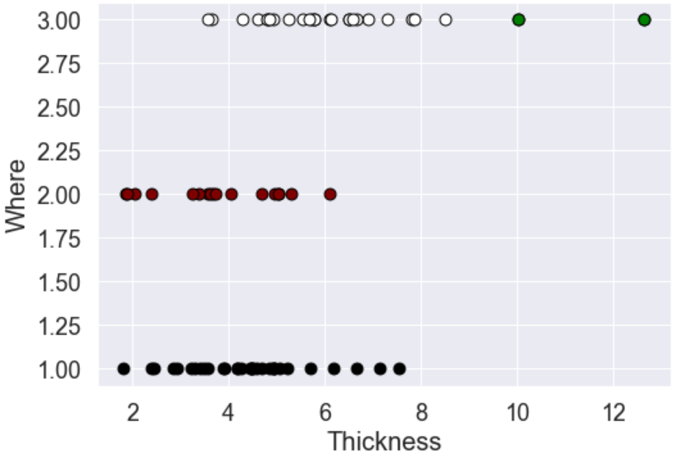

FIGURE 7. Scatter plot of ruby thickness against country of origin

5. Firstly, the logistic regression procedure is fitted to the data of Burma and Thailand. Then a prediction is made about the classes of our set to attain the posterior probabilities of assigning each ruby to a given class. By creating the confusion matrix the model's performance is summarised. Showing 41.2% (7) False Negatives, meaning that these are rubies from Thailand but whose class is predicted as being from Burma and 12.5% (4) False Positives, meaning that these are rubies which originating from Burma, but whose predicted class is Thailand. This model is predicts the country of origin of rubies that originate from Burma more successfully. The overall accuracy of this prediction is calculated at 0.776, therefore this classification is relatively successful. In order to provide insight into this score the roles of the training and test data are swapped and the accuracy score is assessed to see if it has a relatively similar value.

6. Changing from 2 classes to 3, away from the binary classification, linear discriminant analysis is used to form a classification rule for determining the country of origin of a ruby. After calculating the confusion matrix it shows that this classification rule works relatively well for rubies originating in Burma and Cambodia, with only 18.75% and 3.7% of predictions being wrong, respectively. However, the predictions for Thailand were wrong in 52.94% of calculations, suggesting that this classification is unable to properly distinguish between the characteristics of Thailand's rubies and that of the rubies from other countries. Therefore I estimate that this rule would not be very successful.

7. A new ruby was found with the range of characteristics,

```
color 4, diameter 20, thickness 5.00, angle 32.5, cut 3, clarity 0.42
and caratwt 0.900.
```

The country of origin can be predicted using linear discriminant analysis. The output gives a 77.56% probability of being in class 1, a 3.47% probability of being from class 2 and a 18.98% of being from class 3. Therefore it predicts that the new ruby originates from Burma. This is shown in figure 8, where this new ruby has been marked in orange. The LDA separates the 3 classes more efficiently than PCA.

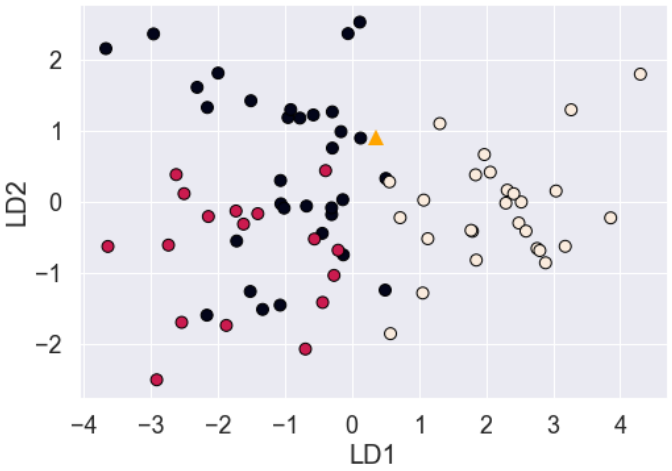

Figure 8. LDA plot of the data and new ruby

In order to predict a guide price, the original data set could be split into a training set and a test set. Using 90% of the data set is a reasonable proportion for the training set. A model is fit onto this set. The k nearest neighbours algorithm could then be used to predict the price of the rubies in the test data set. The training and test data are both mapped into a set of vectors. The algorithm then locates the closest k values of the training set that have the features that are most similar to those of the test set. The distances between the closest values and the test set can be computed. The inverse proportions of the distances from the k nearest neighbours to the test ruby are taken, and this gives the proportion of the price of each neighbour. Taking the sum and average gives a guide price for the test rubies. Once this is done this model can be evaluated on how it performs. If this model predicts the prices of the test set well, lets say within 25 price units, then this model can be used for the classification of any new rubies from any of the three countries.