Machine Learning Report

Ryan Pollard

07/12/2020

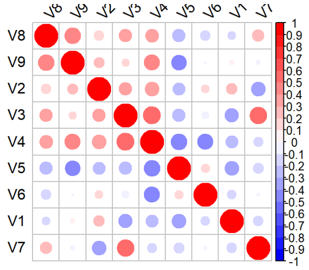
NB: All answers outside R code are given to 2dp. Some outputs and charts have been omitted due to page limit. The sizes of the charts are not ideal but I did not want to omit some things I put time into producing.

PART 1

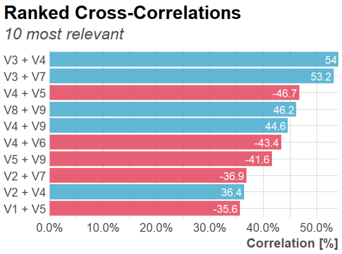
1. Give the correlation between variables V8 and V9.

$\rho = 0.46$

2. Which pair of variables has the greatest correlation?



Heatmap of correlated variables (v3, v4) and (v4, v7) look to be most correlated



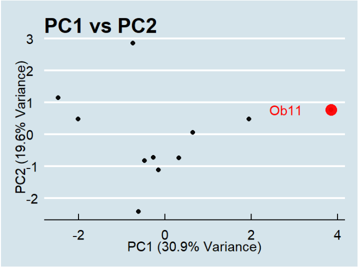
Ordered list of highest correlated pairs of variables

Answer: (v3, v4) are the most correlated with pearson correlation value of 0.54

3. What is the trace of the variance matrix of your data set?

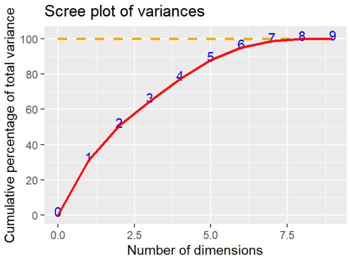
$trace = 55.93$

4. Plot the points on the first two principal components



Scatter plot of PCA1 and PCA2

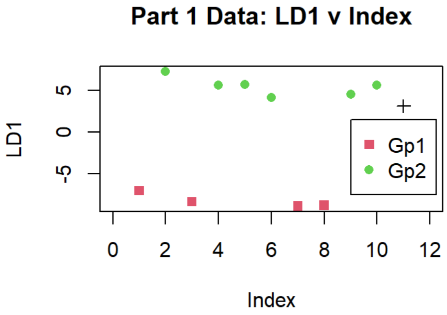
5. What proportion of the information in the data set is given by the first 4 principal components?



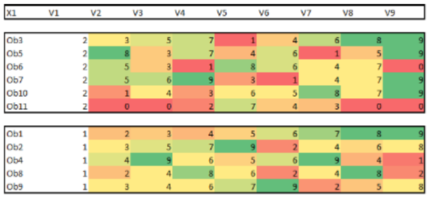
Scree plot shows first 4 components account for ~77% of variance

Answer: 0.77 (taken from summary of pca)

6. LDA



Plot of LDA against index of observations (LD2 not present). Cross-hair point at index = 11 shows the test data predicted as being, wrongly, in group 2, but it belongs to group 1



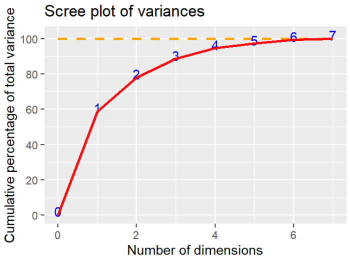
X1	V1	V2	V3	V4	V5	V6	V7	V8	V9
Ob3	2	3	5	7	1	4	6	8	9
Ob5	2	8	3	7	4	6	1	5	9
Ob6	2	5	3	1	8	6	4	7	0
Ob7	2	5	6	9	3	1	4	7	9
Ob10	2	1	4	3	6	5	8	7	9
Ob11	2	0	0	2	7	4	3	0	0
Ob1	1	2	3	4	5	6	7	8	9
Ob2	1	3	5	7	9	2	4	6	8
Ob4	1	4	9	6	5	6	9	4	1
Ob8	1	2	4	8	6	2	4	8	2
Ob9	1	3	4	6	7	9	2	5	8

Answer: The observation 1 is in group 1, but it is predicted to be in group 2, hence the value of 1 in the 2 column.

From the excel table above (a comparison of the data between V1 values), it looks as though that V9 is having a big sway on this, those in group 2 only have values of 0 and 9 for this variable, as does the first observation, therefore it's wrongly placed into group 2. The group 1s have values in between 1 and 8 inclusive.

PART 2

1. Determine an appropriate number of principal components

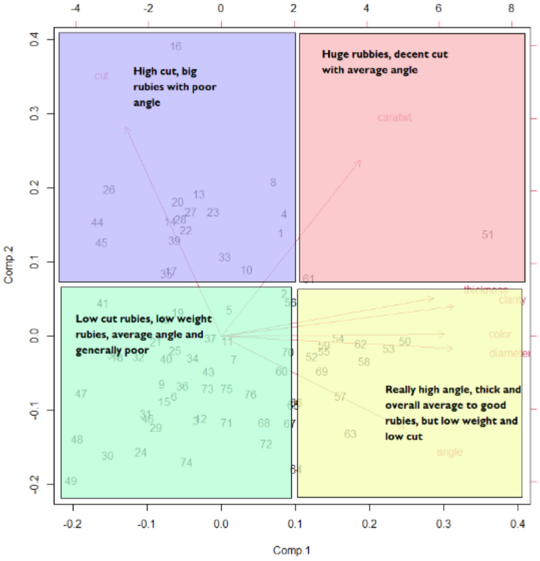


Scree plot showing a plateau at 4 PCAs

Answer: The jump from 2 to 3 PCAs explains ~10% more variance, as does 3 to 4 PCAs, so no strong justification to drop the 4th PCA. The jump from the 4th PCA to the 5th only provides an extra ~3%, hence we see a pleatau in the figure and discard the PCAs from 5 onwards

2. Give a description of the main sources of variation of the quality of the rubies

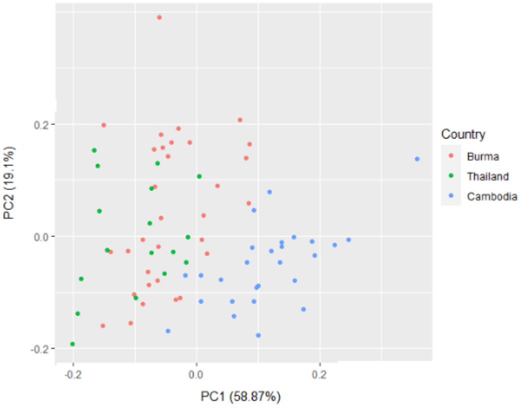
```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## color      0.434         0.452  0.243  0.143  0.434  0.582
## diameter   0.450         0.416  0.113                -0.775
## thickness  0.412  0.130 -0.450  0.247 -0.719  0.177
## angle      0.356 -0.316 -0.568  0.315  0.579 -0.128
## cut       -0.187  0.715         0.618  0.160 -0.208
## clarity    0.453  0.101  0.177 -0.216 -0.110 -0.799  0.237
## caratwt    0.270  0.601 -0.253 -0.582  0.291  0.277
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```



Answer: The first two PCAs account for ~78% of the variance. Between the 2 PCAs, the cut accounts for a lot of the variance, as does the angle, qs the difference between the two loadings of these viarables is the largest. Using these PCAS, the data can be split up into 4 groups and these properties account for the variance of quality. 1. High cut, big rubies with poor angle 2. Huge rubies, lesser but still good cut, with average angle 3. Low weight, low cut rubies, average angle and generally poor quality. 4. Really high angle, thick and average to good rubies but offset with a poor cut and lower weight rubies.

3. What are the characteristics that distinguish the three countries of origin?

This chart highlights where the rubies from different countries sit in the comparison of PC1 and PC2.



Burma : Generally the rubies sit in the bottom left and top left. A portion of Burmese rubies are poor, however a large portion of rubies are also large with high cut, with average angle

Thailand : Generally low quality rubies with an average angle, however some are bigger rubies with a high cut and poor angle. Similar to Burmese rubies but poorer quality.

Cambodia: These rubies are generally good quality, scoring high in many attributes with a high angle, and generally good shape. However tend to be smaller and with a poor cut.

4. Outliers

Using the identify function, we see 2 outliers in row 16 and row 51.

Observation 16 : The Burmese ruby scores a little below average for the attributes except it has a maximum cut of 10 and one of the heighest weights of 1.22, therefore we see it sit in the top left. It's the ruby with the highest cut of 10, and the next highest is a distant 8, which makes it outlie so much. It's the 2nd biggest ruby, 2nd to the next observation.

Observation 51 : The Cambodian ruby scores really well in all the attributes, except cut (4/10), and see its point on the far right of the scatter plot. It is the thickest and best angled, most clear and biggest rubies in the data. It's an outlier due to having the best scores in 5 of the 7 attributes.

5 and 6. Do a logistic regression and discuss how successful is the classification?

For consistency, this analysis will not include price to serve as a comparison with the PCA.

```
##    column.pred
##      0  1
##   0 28  4
##   1  8  9
```

Here we have the output of the logistic regression and a table comparing the predictions of the data vs the real values. The table shows there's a good amount of bad predictions (non-column values). AIC, Null deviance and Residual deviance are relatively low which indicates good classification ability. The accuracy we see is 28 + 9/ 49 = 75% accuracy.

To further analyse how good it is at classifying new data, we split the data into train and test data. For each country, 70% of its data is taken for training, and 30% for testing.

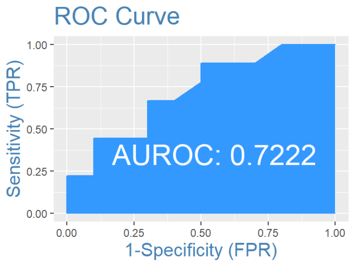
```
## 
## Call:
## glm(formula = where ~ color + diameter + thickness + angle + 
##     cut + clarity + caratwt, family = binomial, data = columntrain)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.3749  -0.6206  -0.3031   0.1841   1.8321  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept)   5.5536    16.7374   0.332   0.7400  
## color        -1.2773     1.8487  -0.691   0.4896  
## diameter      1.1111     1.2065   0.921   0.3571  
## thickness     1.1737     0.7739   1.517   0.1294  
## angle        -0.3437     0.1788  -1.923   0.0545 .
## cut          -0.3430     0.3634  -0.944   0.3453  
## clarity     -48.6226    27.6346  -1.759   0.0785 .
## caratwt      -1.1725     6.8787  -0.170   0.8647  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 34.795  on 29  degrees of freedom
## Residual deviance: 23.062  on 22  degrees of freedom
## AIC: 39.062
## 
## Number of Fisher Scoring iterations: 6
```

```
##    columntest.pred
##     0 1
##   0 8 2
##   1 5 4
```

```
## [1] "Sensitivity: 0.444444444444444"
```

```
## [1] "Specificity: 0.8"
```

```
## [1] "Accuracy: 0.631578947368421"
```



ROC plot

Accuracy of 63% doesn’t suggest it’s a good model for classifying rubies to countries, due to the low train size.

The AUC value of 0.72 suggests an acceptable ability to classify, however it is not good.

The confusion matrix shows a poor ability to correctly classify.

The specificity suggests the model is good at classifying Burmese rubies, but not the Thai rubies (low sensitivity).

Overall, it would be a mediocre model when it comes to predicting if a ruby is Burmese or Thai.

7a. Classifying new ruby

This prediction is based off the regression using all the data and not splitting it into train and test - in order to improve predictive power.

This ruby would be predicted to be a Burmese ruby as the probability is closer to 0 (2.3%).

7b. Predict price of new ruby

We would do this using an OLS with price as the outcome variable, and using the OLS formula, we would use the predict command to predict the new ruby price.

```
## 
## Call:
## lm(formula = price ~ color + diameter + thickness + angle + cut + 
##     clarity + caratwt, data = df_log)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -119.913  -50.778   -4.583   43.188  246.696 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -644.1215   424.1226  -1.519   0.1365  
## color         90.1636    41.7392   2.160   0.0367 *
## diameter      12.7569    28.5918   0.446   0.6578  
## thickness    -26.9892    12.8701  -2.097   0.0422 *
## angle          0.6582     2.5032   0.263   0.7939  
## cut            6.9462     9.0373   0.769   0.4465  
## clarity      821.8186   494.5303   1.662   0.1042  
## caratwt      252.8419   142.6789   1.772   0.0838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 85.32 on 41 degrees of freedom
## Multiple R-squared:  0.6881, Adjusted R-squared:  0.6348 
## F-statistic: 12.92 on 7 and 41 DF,  p-value: 1.234e-08
```

The new ruby price is predicted as 451.68. One of the more expensive rubies, in the top 25% of the data.