Key Drivers for Quality Coffee

Bosniack, Lesley, Bowen, Janet, Higgins, Thomas, & Palacios, Gerardo
DSC 424 - Advanced Data Analysis
DePaul University

According to the Coffee Quality Institute (CQI), coffee quality is one of the most important variables that influence a coffee's value. One way that coffee quality can be measured is through a blind tasting, also known as cupping, by certified coffee analysts using the SCAA Cupping Protocol. This protocol gives guidelines for evaluation ranging from necessary equipment to preparation of the coffee. Ratings are given in various categories such as aroma, flavor, aftertaste, acidity, body, balance, sweetness, clarity, consistency, and overall impression. The final grade sums the ratings against a total of 100 points, similar to rating scales used for wine and similar goods. Anything rated over 80 points is considered a premium coffee. This method of evaluation provides a consistent and objective methodology for capturing some of the beans' sensory aspects and for evaluating quality.

The CQI's goal is to improve the quality of coffee and the lives of coffee producers. As a result, the CQI compiled a dataset of samples submitted for evaluation for coffees worldwide. The dataset provides a profile of coffee growers, coffee beans and the quality of the coffee grown, as measured according to the SCAA categories. Analysis of this dataset could result in valuable findings that would improve production practices. Our data was gathered from *https://www.kaggle.com/volpatto/coffee-quality-database-from-cqi*, which contained 1339 observations and 43 variables from the CQI database of coffee ratings from 2010 through 2018. Each observation represents a sample of coffee and 43 variables associated with it.

The main purpose of our analysis was to figure out if there were associations between the quality ratings of coffee and other information contained in the dataset, and if so, whether these had predictive power. We also investigated to see if there were other patterns in the data. The variables in the set generally fall into three categories that portray different characteristics for each sample of coffee (seen below). A more detailed listing of variables is provided in the appendix.

Quality Measures	
Aroma	Flavor
Aftertaste	Acidity
Body	Balance
Uniformity	Cup Cleanliness
Sweetness	Cupper Points

Bean Metadata
Processing Method
Color
Species
Moisture
Number of Defects

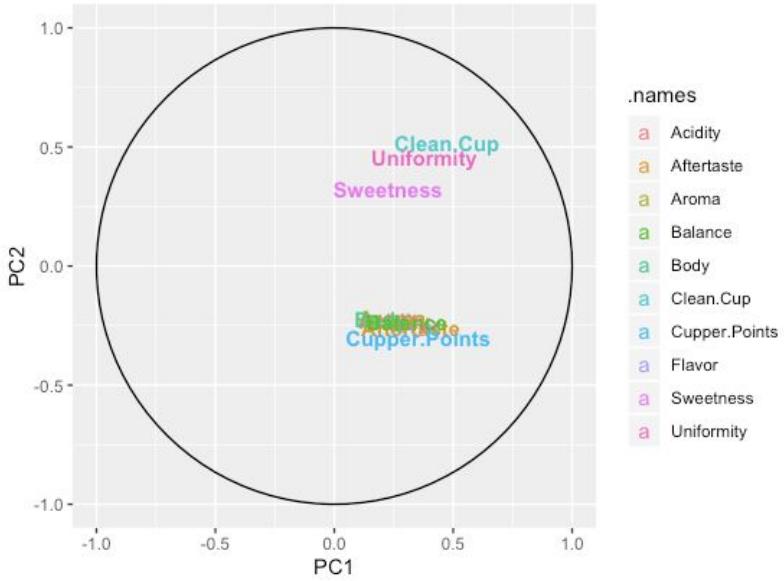
Farm Metadata	
Owner	Origin Country
Farm Name	Lot Number
Mill	Company
Altitude	Region

To carry this out, we used several methods of analysis to attempt to draw useful conclusions from the dataset; these are briefly detailed below, along with the results produced by each.

In one area of analysis, we used our data to predict the Cupper Points grade (that is, the coffee grader's subjective review). This process involves calculations that tweak our model weight variables differently (or even remove them); this is done to handle variables being redundant or not useful for accurately predicting the grade. This model was successfully built, and we found that we were able to account for about 80% of the variation in the cupper points grade, with Flavor, Aftertaste, Acidity, Body, Balance, and Altitude all helping to predict it.

We also analyzed the data to see if we could find different groupings of related variables. These groups can be used to better understand how variables are related to each other; they can also be used as inputs for other techniques, which can often result in a simpler and more accurate model. Using this method, we found that our data grouped well into several categories - two different sets of quality grades, a group of data related to quality control, and a group related to the amount of coffee produced. An example of two of the groupings can be found in the next chart. The chart on the next page shows the ten quality measures from the data set

and the two distinct groupings into which they separated. Looking at the top half of the chart, Clean Cup, Uniformity, and Sweetness have a relationship distinct from the other seven. As it turns out, these groupings show a divide in how these quality scores are measured.



For cluster analysis, we ran a series of calculations to discover groups of data entries that were similar to each other. These clusters were plotted on a visually simplified version of our dataset, and checked against various categorical labels (where the coffee was produced, etc.) to see what, if anything, they related to. In this case, though there were clusters (largely associated with quality control grades), we did not find any variables in our dataset that they were associated with.

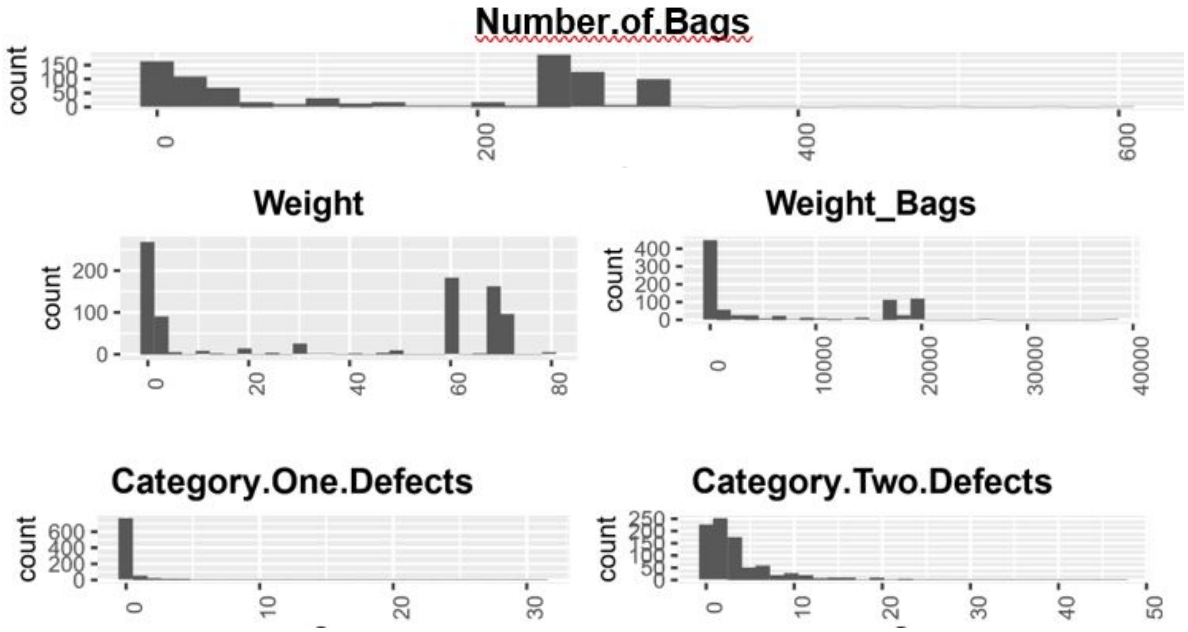
Finally, we ran calculations to see if we could use our data to separate the coffee by processing method. This was not successful; a 70% success rate may seem decent at first, but unfortunately this is roughly what would've happened if we had just guessed that every coffee was produced using the most common method.

In summary, we were able to discover relationships between the quality grades and other variables in our data. We were not successful at predicting the categories of coffees contained in the dataset, despite trying several methods - this could be because the coffee in the dataset is biased towards being high quality - although this is not something we can establish based on the information we have.

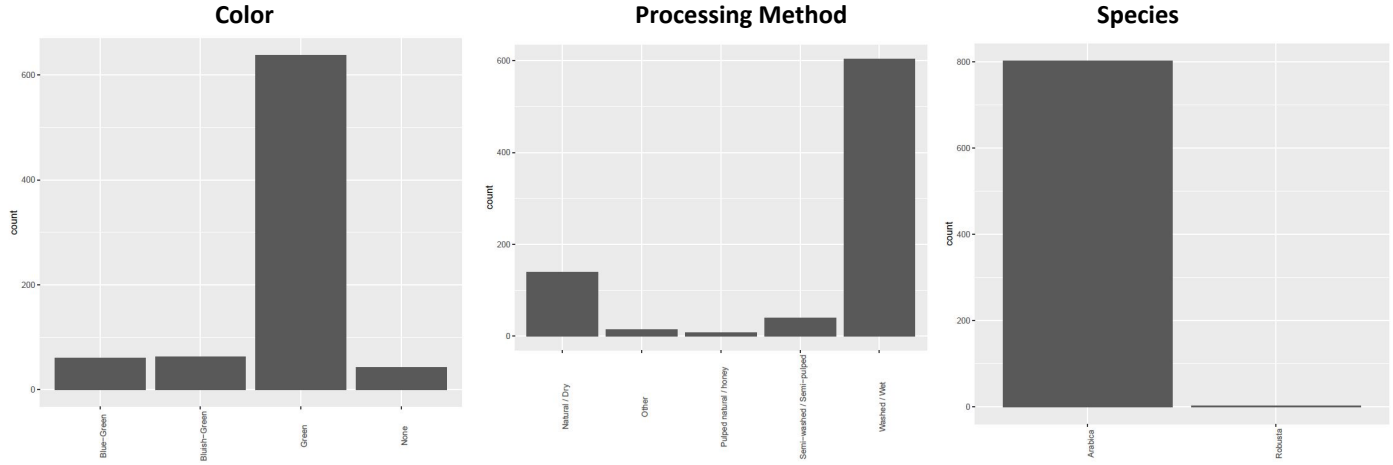
Technical Summary

Data Exploration

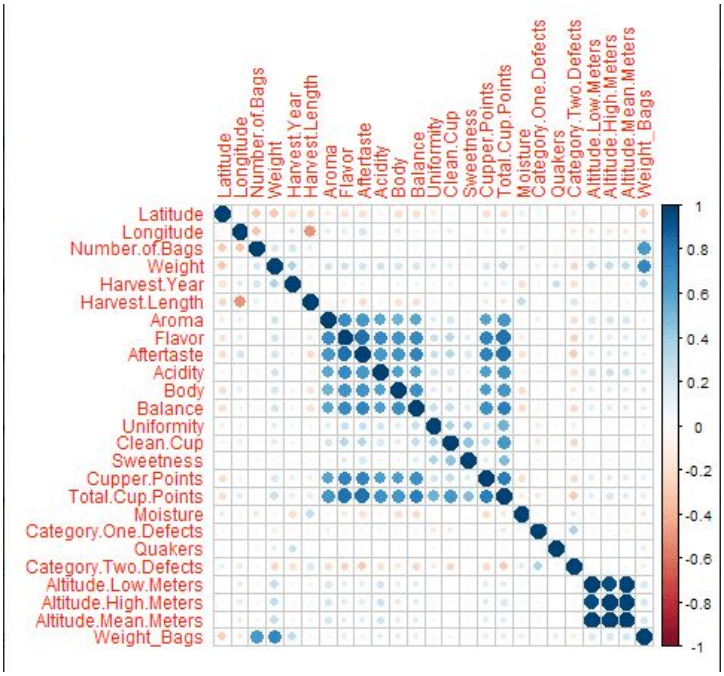
The majority of the variables were fairly normally distributed including the Altitude and Grader Score variables. Number of Bags has a highly skewed-right bi-modal distribution at zero and 250 bags. Weight and Total Weight (Number of Bags x Weight) are also highly skewed-right (that is, most have low weight bags). Transformations did not or nominally improve normality and in some cases created bimodal relationships where the original distribution did not exhibit as obvious a pattern. This was also the case for Category One and Two Defects with the majority of defects at or near zero.



Some of the categorical variables were highly disproportionate with the majority or most of the observations falling under one level. Color, Processing Method and Species are examples.



Correlations were reviewed across all variables. The next page contains a full correlation matrix of relevant variables. It should be noted that altitude variables all show perfect multicollinearity suggesting repetitive data. The Specialty Coffee Association of America ("SCAA") quality scores are all positively correlated with the lowest correlation of 0.51 between aroma and body. This suggests that if quality is high on some criteria, it is likely that quality will be high for other criteria and for the overall quality score, Total.Cup.Points. Similarly, where low quality scores exist on some criteria, low quality scores are expected on other criteria and on the combined score. Defects 1 and 2 and moisture levels exhibit mild negative correlations to the SCAA scores. Altitude does not appear to be correlated with SCAA scores.



Aroma	Flavor	Aftertaste	Acidity	Body	Balance
3.204241	6.980606	6.240338	3.460694	3.057331	3.805068

The above variable inflation factor numbers are taken from an exploratory regression model on the data; the numbers on Flavor and Aftertaste are high enough to indicate that this dataset could benefit from the use of techniques that handle multicollinearity.

As far as feature reduction is concerned, a number of the categorical variables, that had many levels or acted to some extent as primary keys, were removed. Those included Farm Name, Owner, and Mill. Low and high-altitude entries for many of the rows were the same and the same as the mean altitude. This reduced the value of the low and high, and as a result, these two variables were removed.

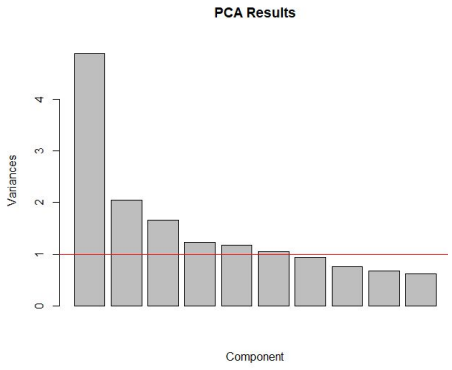
Multivariate Analysis

Principal Components Analysis ("PCA") and Least Squares Regression ("OLS")

To address the multicollinearity across many of the predictive variables, amongst coffee scores in particular, PCA was performed. Analysis was applied with three sets of variables: 1) all numeric variables, 2) coffee scores only, an 3) all numeric variables excluding Altitude, Moisture and Weight. Altitude, Moisture and Weight were not highly correlated with other variables and did not contribute to initial PCA results.

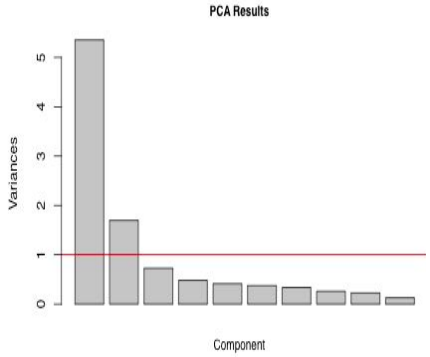
Initial PCA tested the number of components. Four components were selected for further PCA/CFA for the All Numeric Values and Excluding Altitude, Moisture and Weight datasets according to the elbow in each of the variance plots below. Two components were selected for Coffee Scores Only based on a variance threshold of 1.0.

All Numeric Values



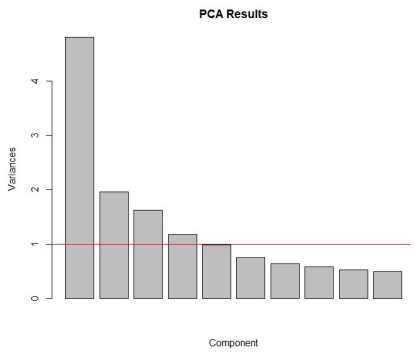
Note: 5 or even 6 components above Var>1 capture 61-67% of the variance. Elbow is between the 4th and 5th components.

Coffee Scores Only



Note: 2 components above Var>1 capture 71% of the variance. Elbow is between the 2nd and 3rd components.

Excluding Altitude, Moisture and Weight



Note: 4 or 5 components above Var>1capture 60-73% of the variance. Elbow is between the 4^{tl} and 5th components.

Varimax scaled rotation was employed due to the range of values across variables from Moisture as percentages to AvgAltitude in the thousands and due to its improved separation of variables into their key components with no change in chi-squared and RMSE results as seen in the chart to the right. All subsequent discussion focuses on Varimax scaled analysis.

Due to lack of normality of some variables, transformations were introduced. Log transformations were applied to totalWeight and Moisture. Both variables delivered distributions closer to normal. The table to the right shows the impact that Varimax has.

Variable	Rotation	
	Without	With
bagCount	0.12	0.00
harvestYr	0.19	0.04
aroma	0.75	0.77
flavor	0.90	0.90
aftertaste	0.89	0.88
acidity	0.78	0.81
body	0.76	0.82
balance	0.84	0.84
uniformity	0.38	0.15
cleanCup	0.43	0.18
sweetness	0.27	0.05
oneDefect	-0.20	-0.01
quakers	0.04	-0.03
twoDefect	-0.39	-0.21
totalWeightLog	0.34	0.18

Rotation comparison excludes modeling of Altitude, Moisture and Weight.

Outliers were identified based on observed values of variables and through OLS analysis. Nine data points were removed. The nine outliers exhibited very low or very high Cupper Points whereas the values for other coffee scores were not at extreme values. Given that initial OLS identified coffee scores as highly predictive of Cupper Points, the removal of outliers with dissimilar relationships was discouraged and as a result was very limited.

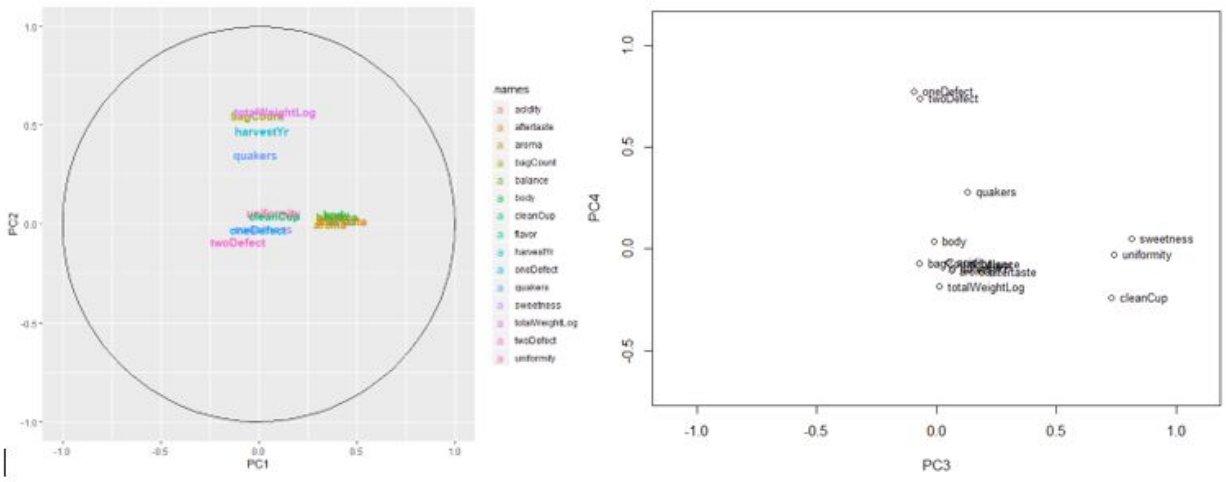
Validation with 60-40 training and test sets delivered reasonably comparable loadings between the two sets. Exceptions were the priority order of the components RC2 and RC3 and the size of the cleanCup load. PCA including all numeric variables delivered loadings, Chi-square test results and root mean squared errors similar to PCA excluding Altitude, Moisture and Weight. PCA results with the excluded variables are provided below for which 64% of the variance was captured, an improvement over PCA including all variables.

PCA Observations

Regardless of the set of variables (All numerics, Quality Scores only, Numerics excluding Altitude, Moisture and Weight) included in PCA, the message was clear as can be seen by the loadings to the right and PCA plots below.

```
                Principal Components Analysis
                    Loadings Cutoff of 0.4

Loadings:
                RC1    RC2    RC3    RC4
aroma          0.769
flavor         0.897
aftertaste     0.879
acidity        0.807
body           0.821
balance        0.838
bagCount              0.772
harvestYr             0.667
totalWeightLog        0.803
uniformity                   0.741
cleanCup                     0.731
sweetness                    0.813
oneDefect                           0.775
twoDefect                           0.742
quakers               0.493

                 RC1   RC2   RC3   RC4
SS loadings    4.332 1.970 1.880 1.384
Proportion Var 0.289 0.131 0.125 0.092
Cumulative Var 0.289 0.420 0.545 0.638

Test of the hypothesis that 4 components are sufficient.
The root mean square of the residuals (RMSR) is  0.07 with the
empirical chi square  1035.68  with prob <  1.4e-183
```

- Component 1 ("Ratings 1" in OLS modeling below) appears to represent a number of the scoring categories (Aroma, Flavor, Aftertaste, Acidity, Body and Balance).
- Component 3, an additional ratings component ("Ratings 2"), focuses on the remainder of the ratings (CleanCup, Sweetness and Uniformity) which are scored differently from those in Component 1 in the SCAA scoring process.
- Component 2 ("Quantity") contains quantity-oriented variables, Weight and BagCount, and also HarvestYr. HarvestYr may be connected to quantity in that some years produce more or less crop than others.
- Component 4 ("Defect") concentrates on defects in coffee beans. In the All Variables PCA, Moisture is included. This grouping may represent defects or possibly variables whose values are unfavorable the higher they are as opposed to the ratings variables which are more favorable the higher they are.



With multicollinearity addressed by PCA, in advance of full-blown regularized regression, OLS testing on Cupper Points targeting a significance of 0.01 was performed with the four components, numeric variables not included in the components, and the key categorical variables of Color and Processing Method identified as significant in regularized regression analysis. The F-test delivered solid overall model results. The amount of variability

(R-Squared) was at an acceptable level of 75%. All four components and Processing Methods were significant with the first ratings component contributing far more than the other four significant variables. The variance of the residuals was not consistent across the range of Cupper Points values. A binomial transformation was tested but did not deliver improved results. OLS results including residual plots appear in the OLS appendix.

Cluster Analysis

Multidimensional Scaling – cmdscale, isoMDS

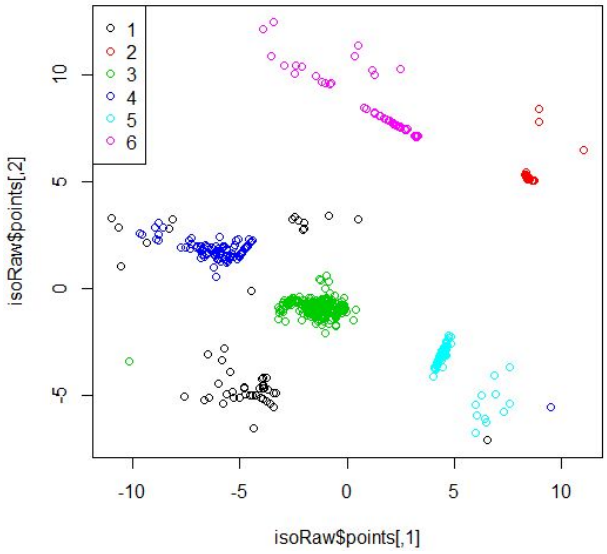
For this analysis, we wanted to see if the grades and quality characteristics of coffee visually cluster from this dataset. The quality scores and the defect measurements were used as inputs here; of these, the defect measurements were subject to a logarithmic transform. A final step of cleaning: in this subset of variables, entries 504 and 535 were now identical, resulting in a relative distance of zero, which isometric scaling does not allow; this was resolved by removing row 504.

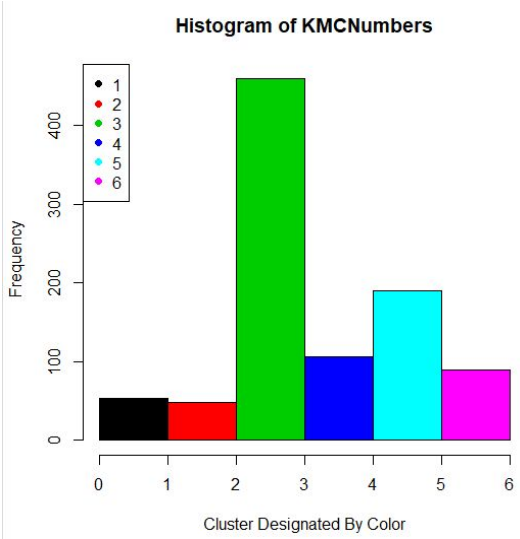
Having done that, we began by plotting both the classic solution for multidimensional scaling, and a converged isometric solution, both in 2 dimensions. The calculations for the isometric solution also gave us stress measurements for both, which could be used to tell us how distorted the visualization is – under 10% is ideal, under 20% is tolerable, and over 20% is probably not usable. The stress associated with the classic solution is 27.55% - this was well above 20%, so it couldn't be considered an accurate representation of the data. This improved to 12.88% for the isometric plot – not great (ideally, this would be below 10%), but good enough to work with. The uncolored charts of these were in the appendix; visually, it appeared obvious that there were clusters in this dataset, especially with the isometric plot.

Several clustering methods were tried (k-means, k-medoids, density); of these, k-means gave perhaps the most informative view, so we examined it here.

Cluster Analysis

K-means required specifying a number of clusters. The scree plot of this data (included in an appendix) bent sharply at 5 for minimizing mean squared error, so no fewer than 5 should be used. Visually, 6 appeared to be a strong possibility, and selecting 6 gave a visual output that mostly lined up with how we expected it to look, so the results here used 6.





The left plot is the isometric MDS,with cluster assignments by color; the right plot is a histogram of how many members each cluster has. For a measure of fit, we can use the between-cluster sum of squares divided by the total sum of squares, which is a measure of how much variance our clusters are explaining. For this dataset, we got around 89.3%, which is high. The visualization showed some points that were not classified as expected; k-means is susceptible to noise, and this superficially looked like a noise issue, but the same points behaved similarly with k-medoids. Our suspicion was that this was related to the stress in the visualization (recall that the stress was low enough to be usable, but still somewhat distorted), and that the points looked out of place but really weren't.

For all 6 clusters, the quality grades were similar – they weren't identical across clusters (Cluster 2 has higher grades than Cluster 4, for instance), but not all cluster pairs were significantly different in this regard. Even when they were significantly different, the differences were small (for example, with Welch's t-test the difference in balance between Clusters 2 and 4 was about 0.3 points in

favor of Cluster 2, with a p-value of 1.443e-07 - definitely different, but a small difference). Instead, these clusters were being separated by moisture and defects; in other words, quality control issues. The following table details the traits associated with each cluster, from interpretation of the cluster centers.

	Moisture	Quakers	Minor Defects	Major Defects
Cluster 1	High	Many	Many	Few
Cluster 2	Low	Few	Few	Few
Cluster 3	High	Few	Many	Few
Cluster 4	High	Few	Many	Many
Cluster 5	High	Few	Few	Few
Cluster 6	Low	Few	Many	Few

Looking at these, it made sense that cluster 4 had lower quality grades – it has the coffees with major defects. Cluster 2 has the coffees with few defects, so it also made sense that it had higher scores. Also, the most common quality control profiles involved few quakers and major defects, but several minor ones. It's worth noting that many combinations of these traits were not represented; more might be seen with a different number of clusters.

The kicker, unfortunately, is that none of these clusters corresponded to any of the categorical labels we have access to in the dataset. A few plots are included as an appendix to give an idea of this, but generally no label was strongly associated with any given cluster. Which is not to say that this was totally unproductive. The clusters gave some evidence that quality control problems were associated with lower quality grades, but the impact appeared limited.
Why don't the clusters mean anything? The basic answer is that the clusters were largely determined by defects, but our dataset doesn't appear to contain any categorical labels that were strongly associated with a coffee's defect profile. Country would've been an interesting one, to highlight differences in a national industry's quality standards, but that association didn't cleanly show up. Similarly, the association of defects with quality grades was weak – this could mean that they were of limited importance for coffee quality, but it could also mean that the dataset was biased towards high-quality coffee, and that coffees with defects were not submitted if they would not grade well; considering the business and reputation incentives at work, this had to be considered a strong possibility, but this wasn't something we could definitively determine here.

That being said, the clusters didn't appear forced by the algorithm – they were driven by an actual pattern in the data. If a category existed that was related to the defect scores, I would've expected it to show up in these clusters.

Regression Of Cupper Points and Processing Methods

Regularized Regression of Cupper Points

When modeling for Cupper Points, six iterations of model building were created. Three of the interactions used the original date set and the other three used the calculated PCA scores. Each iteration of model building follows a systematic approach. First, OLS models were created including forward and backward stepwise regression for comparisons. Then, Ridge, Lasso and elasticNet with varying levels of alphas were also created and done for both lambda.min and lambda.1se. For each iteration, the data was separated into training and test sets (80/20 split) in order to calculate the root mean squared error (RMSE). At each iteration, feature selection was performed by way of Lasso, elasticNet and p-value (OLS), and AIC stepwise selection methods. The most common variables between each feature selection method were then used to create the next iteration of models and repeat the process. In addition to feature selection, the standardized residuals for the best model in each iteration was calculated and any observations that exceeded plus or minus 3.5 standard deviations were removed from the dataset before running a new iteration.

Original Dataset Iterations

1. 26 predictors

species | country | region | bagCount | weight | harvestYr | gradeDate | variety | process | aroma | flavor | aftertaste | acidity | body | balance | uniformity | cleanCup | sweetness | moisture | oneDefect | quakers | color | twoDefect | expirDate | certBody | avgAltitude

Total number of observations is 956. Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.2.

2. 19 predictors, Less 8 residual outliers

species | harvestYr | process | aroma | flavor | aftertaste | acidity | body | balance | uniformity | cleanCup | sweetness | moisture | oneDefect | quakers | color | twoDefect | Bag_Weight | avgAltitude

Total number of observations is 948 Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.05. Many of the variables removed had too many levels without enough observations within each level. Also, time variables were removed because the date frames did not have consistent time frames.

3. 6 predictors, Less 6 residual outliers

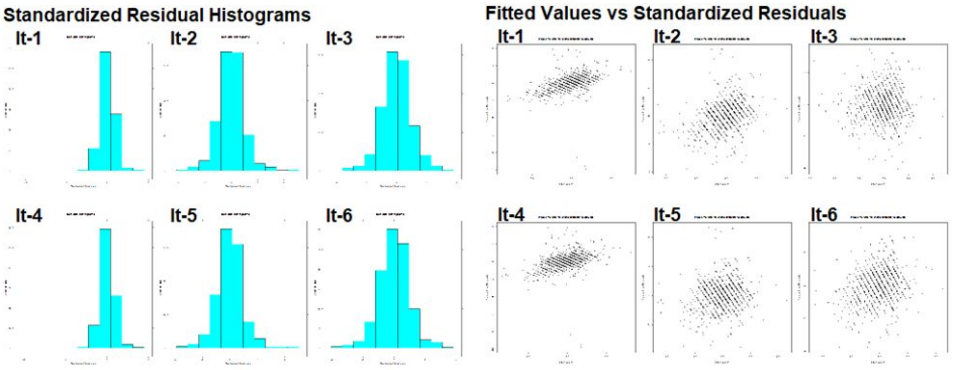
flavor | aftertaste | acidity | body | balance | avgAltitude

Total number of observations is 942 Overall, The best model was produced with elasticNet using Lambda Min at alpha = 0.35.

PCA Scores Iterations

1. All Observations
   Total number of observations is 956. Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.2.
2. Less 8 residual outliers
   Total number of observations is 948 Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.05.
3. Less 7 residual outliers
   Total number of observations is 941 Overall, The best model was also produced with elasticNet using Lambda 1SE at alpha = 0.05.

Residual Analysis



Residual analysis did not yield good results for the 1st or 4th iteration. The standardized residuals in the 1st and 4th iteration are highly skewed. This can also be noted by the high degree of heteroscedasticity when looking at the fitted values against the residuals. There are three obvious outliers in the bottom left hand corner that have standardized residuals exceeding -9. This is a very drastic deviation and may be strongly influencing the effects of the model. This may be an indication of some variables that are giving the model too much weight. For the final models, the residuals against the predictors and fitted residuals were homoscedastic.

Influential Outliers

From all the iterations, most of the residual outliers tended to be on the outer edges. As most of the observations below have absolute residual values between 3.5 and 4. The influential outliers can be argued to be in It-1. Where 3 of the residuals had absolute scores above 9. This is an indication of how the model is unable to represent those few points or something else may be going on.

	flavor	aftertaste	balance	sweetness	cupperPts	res	fitted
1	6.33	6.5	6.83	8.67	6	-3.63	6.92
2	7.58	7.5	7.75	10	5.42	-8.5	7.56
3	7.75	7.67	7.75	10	5.25	-9.57	7.67
4	7.83	7.58	7.83	10	5.17	-9.92	7.67
5	7.75	7.58	7.5	10	8.5	3.76	7.55
6	7.75	7.67	7.67	10	8.5	3.54	7.61
7	7	6.92	7.5	10	8.25	3.91	7.26
8	7.58	7.25	7.42	10	8.42	3.83	7.45

Best Model Training | Testing Results
Original Values

	alpha	RMSETest	R2Test	RMSETrain	R2Train
8	0.35	0.16	0.79	0.15	0.81

PCA Scores

	alpha	RMSETest	R2Test	RMSETrain	R2Train
5	0.2	0.17	0.79	0.16	0.78

Results

Overall, The best model was produced with elasticNet using lambda Min with an alpha of 0.35 with the original values and lambda at 1 SE with an alpha of 0.2. This produced the smaller RMSE error when comparing the training and test sets. The difference between the training and test RMSE are very small, but it does suggest that the model has, at most, a low degree of overfitting due to idiosyncrasies in the data. Both models have a strong goodness of fit capturing between 78% and 81% of the variance in Cupper Points.

Model Beta Coefficients

Original Values

	1
(Intercept)	-0.14147
flavor	0.27952
aftertaste	0.27918
acidity	0.1184
body	0.12038
balance	0.2205
avgAltitude	0

PCA Scores

	1
(Intercept)	7.46561
RC1	0.2878
RC2	0.00074
RC3	0.02493
RC4	-0.02467

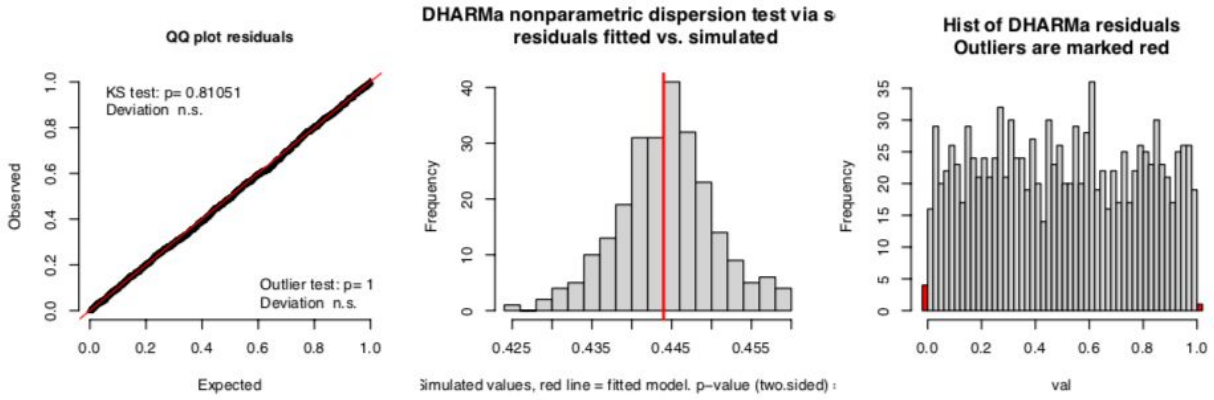
The beta coefficients suggest that the most important variables that can explain the variability in Cupper Points are **Flavor, Aftertaste, Acidity, Body and Balance and Altitude.** More importantly, it suggests that Flavor, Aftertaste and Balance are the strongest contributing indicators by beta weight that may influence the final score in Cupper Point ratings. Another important note is the beta coefficient of Altitude. Although below it is depicted as 0 it is in fact non-zero, just a very small beta weight. Without the variable included in the model the RMSE nearly doubles. For that reason, it remains in the model due to its significance and the reduction in RMSE.

Logistic Regression of Processing Method and LDA

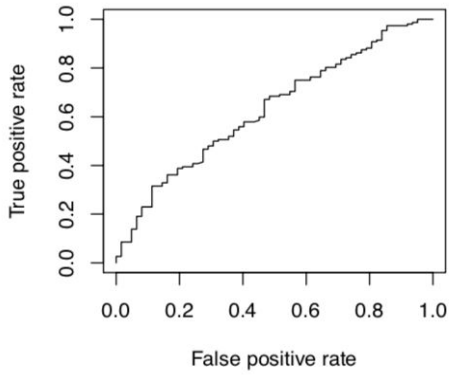
The response variable of the Processing Method naturally led to using logistic regression as the first step in the analysis, using both manual model building and stepwise modeling. Additionally, since a correlation plot of the variables showed high multicollinearity among many of the cupping scores, this called for regularized regression-- specifically Lasso or elasticNet.

The data was split into a training (80%) and test (20%) split. Manual model building and stepwise regression all created the same models, with Aroma, Flavor, Acidity, Body, Sweetness, and Cupper Points noted as significant predictors. The DHARMa package in R has a testResidual function to handle residual analysis on binomial glm regression. The analysis showed no lack of fit based on p values of the Q-Q plots and simulated residuals:

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.3700     2.4022  -0.154  0.87759
## Aroma           0.8672     0.3351   2.588  0.00966 **
## Flavor         -1.3178     0.4334  -3.041  0.00236 **
## Acidity         1.4804     0.3511   4.216 2.48e-05 ***
## Body           -1.1247     0.3482  -3.230  0.00124 **
## Sweetness       0.6754     0.1476   4.576 4.74e-06 ***
## Cupper.Points  -0.6404     0.3056  -2.095  0.03614 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



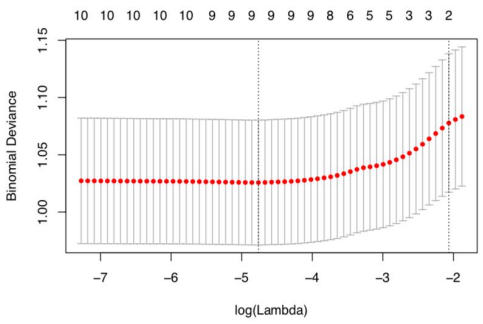
However, the ROC curve (right) determined that the model had little to no predictive value, which was not surprising given there was little difference between the null deviance and residual deviance numbers in the model summary. Prediction of the test data using the training model showed an accuracy of 73%.



For regularized regression, the data was analyzed using Lasso, due to the multicollinearity, as well as elasticNet. With Lasso, the result left only Body as a significant predictor variable using the lambda.1se criterion. Using elasticNet, the alpha value was determined to be 0.5 after testing the range from 0 to 1 in 0.1 increments. This model

Elastic Net Output (alpha = 0.5

```
11 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)   2.42431867
Aroma         .
Flavor        .
Aftertaste    .
Acidity       .
Body         -0.19513180
Balance       .
Uniformity    .
Clean.Cup     .
Sweetness     0.02583458
Cupper.Points .
```

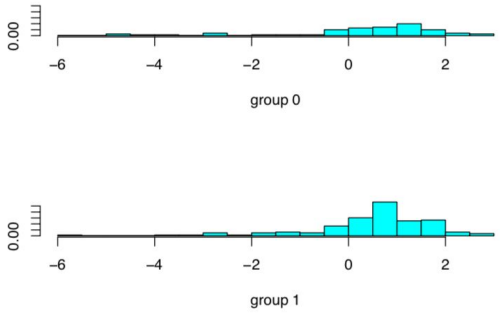


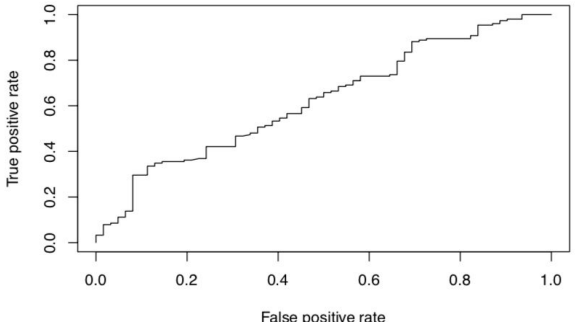
added Sweetness in addition to Body as a significant variable, though pulling in the opposite direction. Analysis of prediction results to the test set showed an accuracy rate of 71%, which was slightly worse than the manual model building, though a more parsimonious model. An analysis of the model plots, however, showed not much in terms of reducing error for either model.

Linear Discriminant Analysis

The final method of exploration was Linear Discriminant Analysis, to see if there was another way to separate the two Processing Method classes. Using lda from the MASS library on the training data produced a very similar model to the output of logistic regression: Aroma, Flavor, Acidity, Body, Sweetness, and Cupper Points were all significant predictor variables.. A plot of the groupings, however, did not show any separation and the ROC plot did not show improvement over logistic regression. A prediction of the training model on the test data had an accuracy rate of 71.5%, which was similar to other methods.

```
Coefficients of linear discriminants:
                    LD1
Aroma          1.0443272
Flavor        -1.8401418
Acidity        2.2975844
Body          -2.2812041
Sweetness      1.4147950
Cupper.Points -0.9553156
```





A confusion matrix comparison of all of the methods show that even though prediction accuracy rate is 70-71% for all, each model under-predicts dry processing (0 value) by quite a bit. In actuality, the model does not predict any better than if one were to guess wet processing for every observation (which is what elasticNet actually did).

LOG

	0	1
0	4	0
1	58	152

PCA

	0	1
0	8	4
1	54	148

Elastic Net

	0	1
1	62	152

LDA

	0	1
0	6	4
1	56	148

Concluding Remarks

According to the CQI, quality is one of the most important variables that influence a coffee's value. The goal of the CQI is to improve the quality of coffee and the lives of coffee producers by providing producers "access to the tools and support they need to understand the quality of their coffee, improve that quality, access markets that reward that quality, ultimately enabling them to make more informed business choices." [1] Knowledge ascertained from the CQI's dataset could provide support for sounder decision making by coffee producers. In our analysis of the dataset, we found that multicollinearity was apparent across all coffee ratings, suggesting that coffee quality across the rating criteria were associated. Our PCA analysis combined two rating components and quantity and bean quality components to predict Cupper Points with satisfactory predictability through least squares and elasticNet regression.. **Our analysis concluded that Flavor, Aftertaste and Balance are *some* of the strongest contributors towards high ratings. Suggesting that these may be some of the most prominent qualities that critics look into when rating the quality of coffee. Understanding the underlying drivers of the various rating criteria could improve the use of these observations by coffee producers. In turn, many producers could adjust their growing strategies or preparation strategies that may influence those variables or factors.** MDA and cluster analysis, although not highly conclusive, indicated to us the possibility of a bias towards high quality grades. **This may be due to the fact that coffee producers would only enter their coffee if they felt it was high quality already. Thus, it may be safe to assume that producers would not intentionally enter a poor quality coffee for quality rating.** And finally, logistic regression was not able to find much relationship between Processing Method and Acidity, Aroma, Body, Cupper Points, Flavor and Sweetness ratings.

[1] Coffee Quality Institute, URL: https://www.coffeeinstitute.org/our-work/

Appendix - Listing of Variables

Variable	Type	Description
Species	Categorical	
Number.of.Bags	Numeric	Number of Bags of Coffee in a Growing Season
Wt	Numeric	Average weight of bags in kilograms
Harvest.Year.Fixed	Date	Year the coffee was harvested
Grading.Date.Fixed	Date	Month, day and year coffee was graded
Variety	Categorical	Variety of coffee plant
Processing.Method	Categorical	Method of processing (wet/washed, dry, pulped and semi-washed)
Aroma	Numeric	SCAA aroma (dry fragrance) quality score ranging from 6-10
Flavor	Numeric	SCAA flavor quality score ranging from 6-10
Aftertaste	Numeric	SCAA aftertaste quality score ranging from 6-10
Acidity	Numeric	SCAA acidity quality score ranging from 6-10
Body	Numeric	SCAA body quality score ranging from 6-10
Balance	Numeric	SCAA balance quality score ranging from 6-10
Uniformity	Numeric	SCAA uniformity quality score ranging from 6-10
Clean.Cup	Numeric	SCAA clean cup quality score ranging from 6-10

Variable	Type	Description
Sweetness	Numeric	SCAA sweetness quality score ranging from 6-10
Cupper.Points	Numeric	Correction points added by grader for more appealing coffees not reflected in other scores
Total.Cup.Points	Numeric	Sum of Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup and Sweetness scores
Moisture	Numeric	Moisture percentage
Category.One.Defects	Numeric	Number of primary defects
Quakers	Numeric	Quantity of unripe or poorly roasted beans
Color	Categorical	Coffee color ranges from green to blue
Category.Two.Defects	Numeric	Number of secondary defect
Expiration.Fixed	Date	One year post grade date

Appendix - Summary Statistics of the Variables

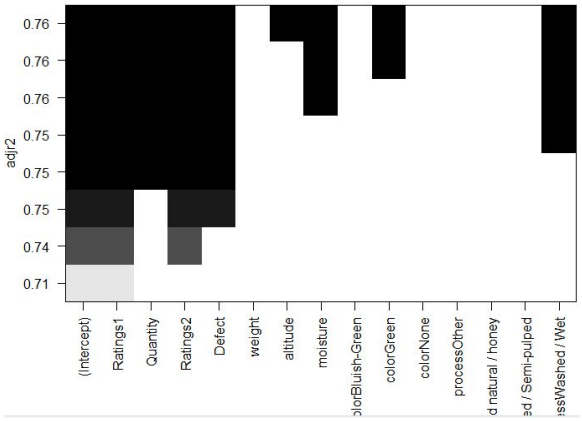
Statistic	Latitude	Longitude	Number.of.Bags	Weight	Harvest.Year	Harvest.Length	Aroma	Flavor	Aftertaste
1st Qu.	2.536	-91.47	20.0	1.0000	2012	92.0	7.420	7.330	7.170
3rd Qu.	17.059	-44.56	275.0	69.0000	2015	152.0	7.750	7.670	7.580
Max.	37.090	126.09	600.0	80.0000	2018	214.0	8.750	8.670	8.500
Mean	9.174	-44.41	159.2	36.1845	2014	133.3	7.561	7.505	7.376
Median	14.518	-84.04	200.0	60.0000	2014	151.0	7.580	7.500	7.420
Min.	-23.563	-118.28	1.0	0.4536	2011	31.0	5.080	6.170	6.170

Statistic	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points	Moisture
1st Qu.	7.330	7.330	7.330	10.00	10.000	10.000	7.250	81.17	0.10000
3rd Qu.	7.670	7.670	7.670	10.00	10.000	10.000	7.670	83.50	0.12000
Max.	8.580	8.420	8.580	10.00	10.000	10.000	8.580	89.92	0.17000
Mean	7.517	7.494	7.492	9.87	9.846	9.933	7.461	82.05	0.09807
Median	7.500	7.500	7.500	10.00	10.000	10.000	7.500	82.42	0.11000
Min.	5.250	6.330	6.080	6.00	0.000	1.330	5.170	59.83	0.00000

Statistic	Category.One.Defects	Quakers	Category.Two.Defects	Altitude.Mean.Meters	Weight_Bags
1st Qu.	0.0000	0.0000	0.000	1100	180
3rd Qu.	0.0000	0.0000	5.000	1550	17250
Max.	31.0000	11.0000	47.000	2560	37950
Mean	0.4263	0.1463	3.822	1293	6558
Median	0.0000	0.0000	2.000	1311	640
Min.	0.0000	0.0000	0.000	1	1

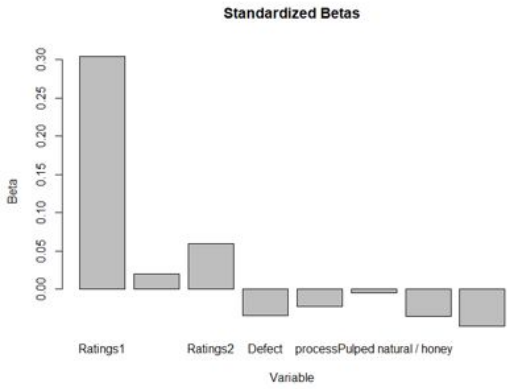
Appendix - OLS

Significant variables from All Subsets OLS regression is below.

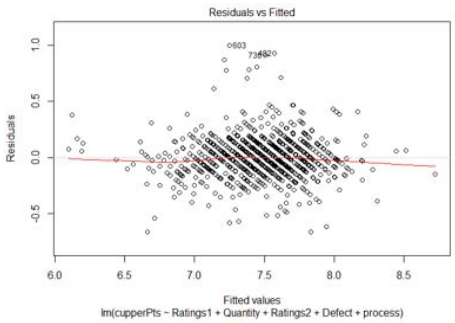


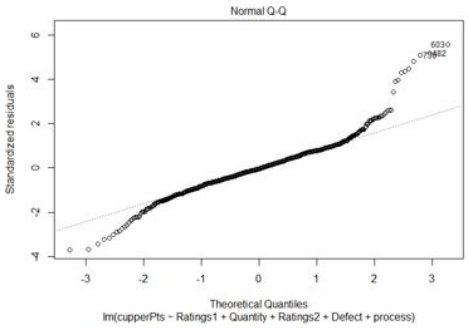
Below is the chosen OLS model and model results..

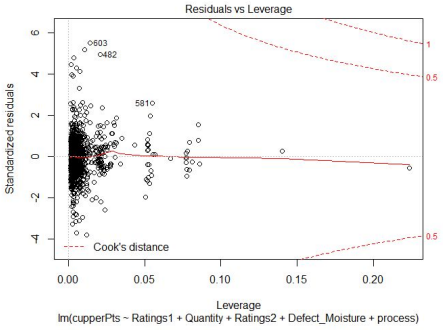
```
lm(formula = cupperPts ~ Ratings1 + Quantity + Ratings2 + Defect + process,
 data = coffeeScoreAlt)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66264 -0.10056 -0.00765  0.09429  1.00021

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     7.506554   0.013244 566.804  < 2e-16 ***
Ratings1                        0.304625   0.005896  51.670  < 2e-16 ***
Quantity                        0.019758   0.005960   3.315  0.00095 ***
Ratings2                        0.059369   0.005856  10.138  < 2e-16 ***
Defect                         -0.034844   0.005858  -5.948 3.83e-09 ***
processOther                   -0.022975   0.042498  -0.541  0.58889
processPulped natural / honey  -0.004754   0.052125  -0.091  0.92735
processSemi-washed / Semi-pulped -0.035587   0.028774  -1.237  0.21647
processWashed / Wet            -0.048479   0.014995  -3.233  0.00127 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.18 on 942 degrees of freedom
Multiple R-squared:  0.756,    Adjusted R-squared:  0.754
F-statistic: 364.9 on 8 and 942 DF,  p-value: < 2.2e-16
```



OLS diagnostic plots are below.



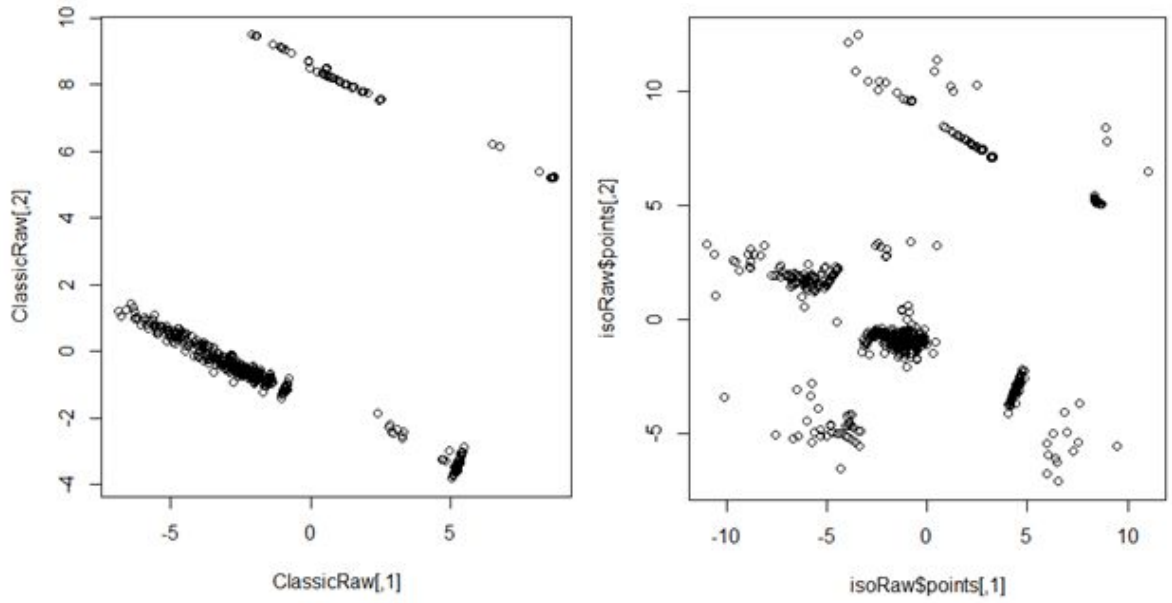




- ✓ F-test results excellent at 2.2e-16
- ✓ 75% of variability of Cupper Points explained
- ✓ t-test significance < 0.01
- ✓ No multicollinearity concerns
- ✓ Some heteroscedasticity and non-normal residuals in normal Q-Q at far left and right

Appendix - Additional Charts for MDS and Cluster Analysis

The multidimensional scaling plots are below. The left is the classic solution, the right is the isometric solution.

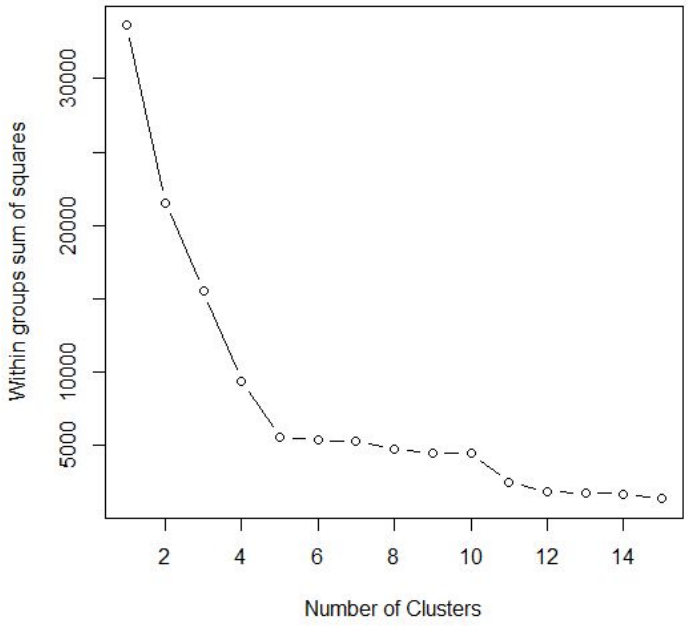


The cluster centers from the k-means clusters are below; from top to bottom, with rows denoted by cluster.

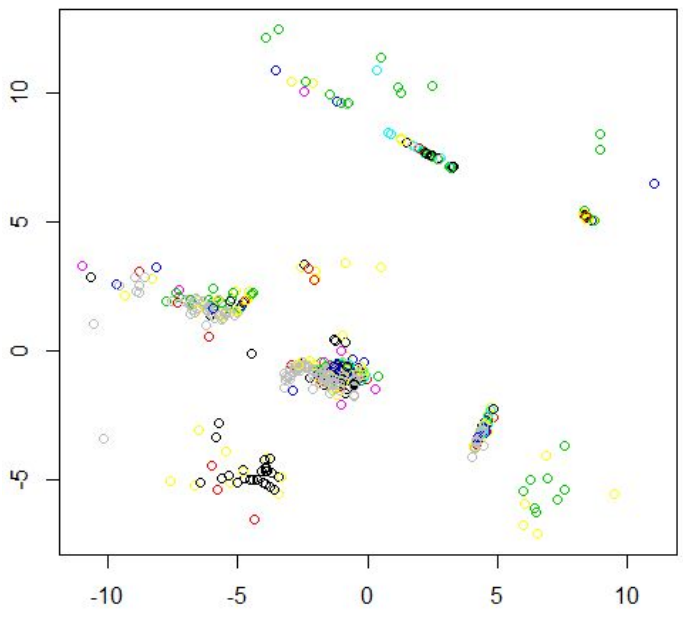
	aroma	flavor	aftertaste	acidity	body	balance	cupperPts	uniformityLog
1	7.530000	7.539245	7.390189	7.568113	7.496792	7.466981	7.504151	2.300068
2	7.655208	7.572500	7.490000	7.596042	7.591667	7.645417	7.646250	2.292489
3	7.561046	7.502614	7.366863	7.538105	7.493791	7.491852	7.460850	2.289968
4	7.439245	7.362736	7.202170	7.352358	7.377358	7.335849	7.272736	2.262731
5	7.602158	7.542684	7.434947	7.523684	7.503263	7.510211	7.490211	2.291704
6	7.553889	7.555333	7.473333	7.538000	7.607222	7.605778	7.552667	2.284496

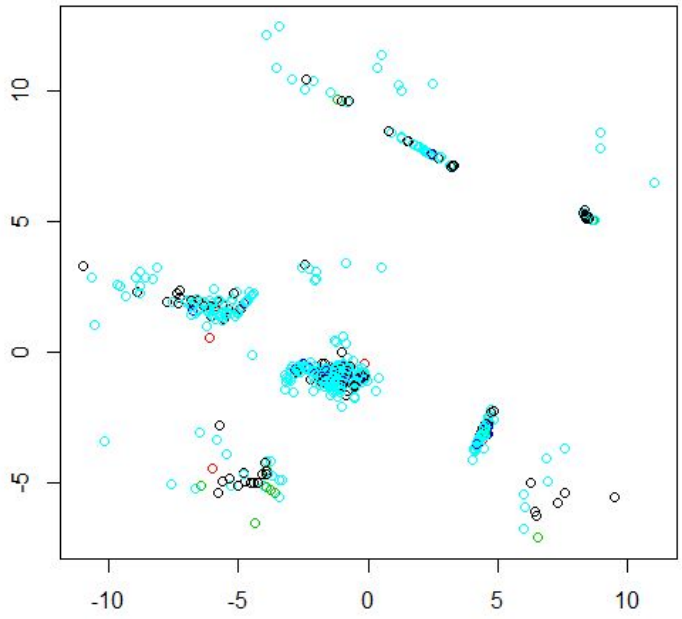
	cleanCupLog	sweetnessLog	moistureLog	oneDefectLog	quakersLog	twoDefectLog
1	2.300068	2.300068	-2.228282	-6.0787785	0.6435267	1.003852
2	2.298268	2.292579	-11.512925	-6.6054603	-6.7493927	-6.907755
3	2.262997	2.294214	-2.178691	-6.9077553	-6.9077553	1.107456
4	2.253338	2.280258	-2.141593	0.7361514	-6.8425783	1.764631
5	2.293407	2.295410	-2.201116	-6.5812920	-6.7950242	-6.907755
6	2.279505	2.292022	-11.512925	-5.6948907	-6.3197251	0.963791

Below is a scree plot, used as a diagnostic to pick the number of centers for k-means clustering. The sharpest bend happens at 5 clusters, although another drop happens at 11. Any number of clusters below 5 would be a bad choice - having a too high within groups sum of squares tends to indicate that some separate clusters are being grouped together as one.



Below are some example charts of the MDS plotted with categorical variables; the left is country, the right is production method. These plots are not very conclusive, and indicate that the clusters are likely not very associated with these variables.





Summary of Individual Analysis

Summary of Individual Analysis

Gerardo Palacios

Summary of Individual Analysis

Multivariate analysis has been an eye opening learning experience throughout the quarter. In fact, one of the most interesting aspects that has been an continuing learning theme is transforming a data set's weakness into strengths to extrapolate latent details in the data. A perfect example of this was learning to use PCA/CFA in combination with linear models in order to improve performance and gain further meaning. Our project is a great example of applying what I've learned about multivariate analysis. This project allowed me to combine multiple techniques that compliment each other. In addition, using PCA with modeling allowed me to have a better understanding how data points can relate to each other and finally being able to use geometric calculations to judge potential importance. As a result, my process can be broken down into three sections, initial analysis, individual analysis and comparisons.

Initial Analysis

The initial analysis involved merging and cleaning data, correlation visualizations and variable transformations. More cleaning was conducted by other members of the group while I performed transformations such as standardizing units (i.e. lbs to kg) and removing incomplete observations. After clean up, I created the correlation plots and stacked histogram visualizations on every variable showing the original, logged and squared root values in order to determining if any of the variables could be become more normal. Ultimately, none of the variables truly benefited from any of the transformations since many of the measurements were based on a predetermined scale. This meant that many of the variable transformations would not be appropriate if performed on the data set.

Individual Analysis

After the initial analysis we each conducted our own analysis on the data to compare methodologies (in terms of finding the best model). This involved PCA/CFA, OLS and regularization models. My modeling was based on sparser data set, I omitted all the blanks at read, while Les had more robust set where she used 3rd party data to fill it any potential blanks in the data. This was done in order to compare the sets to see if the extrapolated missing values would be drastically different from the sparser data set For the first portion of the analysis I ran PCA in 2 different iterations. Each iteration involved comparing the results with its scaled|unscaled and rotated | rotated counterparts.

1. No variables excluded

   Unscaled PCA held the entire variability in PC1, This would create uninterpretable results. Fortunately, using scaled PCA allowed the variability to become more evenly spread out. More

importantly, I was able to use the variance value of 1 as a breaking point for determining the number of components.

2. Excluding at CF = 0.90

   Using a correlation test with a confidence level of 90%, I was able to eliminate very high correlated variables. From the original 29 numerical variables reduced to only 4 variables.

As expected, using all the variables resulted in difficult interpretations. The scaled results, surprisingly are harder to interpret than that of the unscaled.

The regularization models were next. This was done in number of ways. First, OLS models were created this includes step wise AIC for modeling and feature selection. Then, was conducting the regularization models. Using cross validation and randomly splitting the data into training and test sets (80/20) I was able to compare the different models and judge for over fitting. The regularization models and outputs were stored in tables using a sequence of alpha from 0 to 1 by increments of 0.05 (21 models) and subsequently comparing the results using both calculated lambdas (lambda.min and lambda 1SE). This meant a total of 42 models for each iteration. The root mean squared errors were then calculated and the model with the smallest diagnostic was then used for residual analysis. This same process was repeated only using the PCA scores that were calculated earlier.

Residual analysis was conducted at 3 levels. At each level, a number of observations were removed that had calculated standardized residuals greater than 3.5. A total of 15 influential outlines were found using this methodology and 16 influential outliers were found when using the PCA scores in regularization. Without repeating outputs, the 15 residuals were did not seem to have any similarities except that their actual cupper points were rated low relative to other observations that had similar scores. The biggest ones were residuals over 10. Three observations (Kenya, Indonesia, and Taiwan) had high predictor scores, but had actual cupper points of less than 6. Compared to all of the models, each predicted these observations to be much higher. Ultimatly, it seems we are missing variables in the dataset that do seem to explain those observations very well.

Comparisons

Compared to Les's analysis there was not much difference in the models, ultimately deciding that using more observations was better since the 3rd party data did not seem to vary drastically. As a result, for the final report as well as the presentation we opted to use her PCA analysis coupled with my linear and regularization models. The results flowed very nicely together as the raw data and the PCA analysis complemented each others interpretation. For example, the first PCA component was composed of the same variables as my best linear model and regularization model using the raw data. This is evidence of the significance and potential importance for predicting the variability in cupper points.

Final Thoughts

I think the most important aspect that could be drawn from my contribution to the project would have to the residual analysis and visualizations. The models, when compared without removing the residuals outliers, had more than 10% less in calculated captured variance (R^2). Removing the outliers helped stabilize the model to reveal less inflated/deflated beta coefficients in both regularization and OLS.

Tom Higgins – Individual Report on MDS and Clustering

Summary of Work on Project Setup

In this group project, we tended to work by handling tasks as they came up, with multiple people pitching in for just about everything we did – the only formal division of roles was when we divided up the analysis, and even then we tended to consult with each other as we went. We worked on tasks as it made sense at the time; accordingly, I'll list a few tasks I was particularly involved in.

Along with everyone else, I did some research and information gathering so that we understood what was in our dataset, and which values we could and couldn't use. I wrote initial drafts of group milestones, although as submitted those had input from everyone. Related to that, I did some initial modeling on the dataset as a proof of concept to help demonstrate that regularized regression and factor analysis could be applied to the dataset. Analysis topics I worked on for the report and presentation included multidimensional scaling and cluster analysis. I also worked on common factor analysis, but that was cut for redundancy with principal component analysis.

Summary of Work on Cluster Analysis

For clustering inputs, I used aroma, flavor, aftertaste, acidity, body, balance, and cupperPts. Additionally, the following variables had a log transformation applied: uniformity, clean cup, sweetness, moisture, grade 1 defects, grade 2 defects, and quakers. The categorical labels I deemed usable for this analysis were production year, country, color, production process, and coffee variety.

This first needed to be scaled, so I ran both the classical and isometric MDS solutions available on R. The classical solution was very distorted; the isometric solution was still somewhat distorted, but the stress was about 12%, low enough to be usable.

Judging by how the scaled plot looked, I tried density-based, k-means, and k-medoids clustering methods, and settled on k-means giving the best view of the dataset (k-medoids tended to produce the same analysis, and density-based tended to either not classify enough of the dataset, or to group together clusters that seemed obviously different). I used a scree plot with total squared error to determine how many clusters I should use, combined with how the dataset appeared visually, and settled on 6.

The clusters produced turned out to be determined by quality control inputs, and using the between-cluster sum of squares divided by the total sum of squares as a measure of fit, they explained 89.3% of the variance, which is high. The cluster centers allowed for building a distinct quality control profile for each cluster. In some cases, the quality scores were also slightly different, particularly for clusters 2 and 4; however, the differences were slight (under half a point, in all cases) and not always statistically significant between a given pair of clusters, as judged by R's implementation of Welch's t-test.

Unfortunately, the clusters did not cleanly line up with any of the categorical labels I checked. Accordingly, the analysis was somewhat inconclusive; the clusters appear real and not forced by the

algorithm, but they don't line up with any of the categories we have access to. This probably means that none of those categories are strongly associated with what defects a coffee has.

Takeaways and Lessons

The implications of the defects apparently having only a minor effect on coffee grades aren't totally clear. It could be that they truly only make a minor difference, or it could be that coffee is only submitted for grading when it's expected to be high quality, and thus the dataset doesn't include defective coffees that have a low grade. Another method of analysis is likely more suited to examining this, and it would probably need more/different data.

The punch list entries changed this analysis a lot; the effect the transformations had on the clustering essentially forced me to redo the entire section of analysis after the presentation (including picking a different clustering method). What I learned from that was to consider whether some variables should be rescaled, etc. if an analysis is not giving results. Doing this to the quality grades didn't accomplish much, but the way the quality control/defect traits were distributed caused some differences to emerge that weren't particularly notable unscaled, after a log transform was applied. Funnily enough, the conclusion was still very similar - the clusters we found didn't line up well with any categories either way, but it's more plausible to me now that some category exists that lines up with them, and that we just didn't have it in our data.

Code Snippets

During my work on the report, I used the following snippets of code. This doesn't include every bit of cleaning, subsetting, etc. that was run for the analysis, and it doesn't include the many things that did not make the final report; it's just here to give an idea of what I did.

```
RawDist = dist(CoffeeRaw) # distance for MDS
ClassicRaw = cmdscale(RawDist, k = 2) # classic solution
isoRaw = isoMDS(RawDist, y = cmdscale(RawDist, 2), k = 2) # isometric, and
stress calculations, from MASS
set.seed(12) # consistency on k-means
wssplot(CoffeeRaw) #scree plot to determine number of centers.
# Not a base/library function – I used code from the following link:
#
https://stackoverflow.com/questions/33752645/scree-plot-for-determining-k-
in-k-means
set.seed(12) # more consistency
KMeans2 = kmeans(CoffeeRaw, 6, nstart = 10, iter.max = 100) # the
clustering
plot(isoRaw$points, col = KMeans2$cluster) # clustering visualization
plot(isoRaw$points, col = CoffeeTraits$variety) # label visualization;
CoffeeTraits is a subset of variables from the initial file
plot(isoRaw$points, col = CoffeeTraits$process) # another label
visualization.
t.test(Cluster2$aftertaste, cluster4$aftertaste) # To show the difference
in quality scores per cluster; not exhaustively run for every possibility
hist(KMeans2$cluster, col = 1:6, breaks = 0:6, xlab = "Cluster Designated By
Color") # histogram to show cluster population
```

```
legend("topleft", legend = 1:6, col = 1:6, pch = 19) # colors histogram, but
similar code also used for some other tables
```

Key Drivers for Quality Coffee

DSC424, Individual Report– Lesley Bosniack

This individual report focuses on my role in the development and execution of the analysis of coffee quality. Outlined are the steps I took to contribute to the overall conclusions drawn.

1. In addition to exploring various potential datasets, my high-level review of those found by classmates led me to do a deeper dive on the Coffee Quality dataset that Gerardo Palacios proposed. Collectively, my team explored additional weather sources to supplement the dataset.
2. We collectively drew up the dataset proposal for submission. My participation was more on the backend of the submission with some proposal writing. With the help of Gerardo's distributional and correlation plots, I reviewed and drew conclusions on distributions, possible transforms and areas where multicollinearity would need to be addressed. In addition, I participated in the build of the data dictionary which was not available to us through Kaggle or the CQI. The data exploration took place in this step, but I further explored in my individual analysis.
3. Decisions on the techniques used was also a collective effort. My area of focus along with Gerardo's was on PCA and regularized regression on Cupper Points. We independently explored both techniques for viability and compared our observations. This led to the proposal of the first technique which was then broken into two techniques (PCA & CFA / OLS vs Regularized Regression).
4. The development of PCA and CFA was the work that I performed. The analysis addressed multicollinearity across many of the factors and provided components that fed the cluster analysis and regularized regression techniques. As an extension to the component analysis, I wrapped up my work with ordinary least squares regression ("OLS") to test the predictiveness of the selected components as a preliminary indication of what to expect in the regularized regression explored by my colleagues.
5. With the lows and highs of our busy schedules, my colleagues and I each contributed sometimes more and other times less than others to the different collaborative milestones. The presentation and to some extent the final report is where I contributed more, in structuring and developing the "common" slides of the presentation. As in other aspects of the project, the presentation and the final report were shared equitably amongst all four of us.

Individual Analysis on PCA and CFA

Before PCA and CFA were performed, I did a deep dive on the dataset to develop a sense of the quality of the data and to further gain domain knowledge. Both fed the decision making in the modeling process of how much reliance to place on the various variables.

- The latitude, longitude, and weather API data collected by Gerardo, and the Harvest Dates collected by me to facilitate the merging of the coffee dataset with the weather data, came into question. Detailed location data at the appropriate altitude was necessary to include, and the Region variable did not provide this detail. In comparing the coordinates collected to the Country variables, I estimated that approximately 10% of the coordinate information was incorrect. This confirmed the choice to remove the coordinate and weather data so as to mitigate potential distortions to the analysis.
- Estimations on missing data were developed cautiously. For example, the relationship between Harvest Years and Expiration Date from other rows fed the estimation of missing Harvest Years and of incorrect the chronological ordering of the two variables. Altitude for all Hawaiian farms was researched and populated to preserve the use of a number of data points. Gglots were reviewed to dissect the data for patterns and identified the need to consider either Country or Certifying Body to supplement the Harvest Year estimation. Often enough, Variety, Processing Method, and Color were all empty for a number or rows. Patterns in enough cases were not clear enough to impute. Multiple imputation was tested but ultimately not utilized. The final dataset contained 956 rows and 28 variables.
- Nine rows were removed for outliers based on observed values of variables and identification through OLS modeling. These nine rows had very low or very high Cupper Points whereas other variables, particularly other coffee scores, were not at extreme levels. Because the other coffee scores were highly predictive of Cupper Points, removing rows exhibiting this pattern was done on a limited basis.
- Due to the lack of normality of some variables, transformations were considered. Most transformations did not deliver enough of an improvement to warrant. For example, cleanCup, Defect One, Defect Two, Quakers, Sweetness and

Uniformity, entries were highly concentrated on one value and often a value at one end of the range of values. Log transformations narrowed the spread of the values of the variables. The PCA results delivered three components, one less component relative to the PCA results provided in our final report. Two of the components reflected coffee ratings, similar to the final report, but the third component appeared to be a combination of somewhat unrelated variables that four components better segregated. Although PCA with transformed data provided less components, the explanatory nature of the results were the reason for discarding. The three component results were tested in the Tom Higgins' clustering analysis. Log transformations of Total Weight and Moisture delivered distributions closer to normal and were maintained in the analysis.

- Upon reviewing my OLS of Cupper Points, it became clear that the variance of the residuals was not consistent across Cupper Points. A binomial transformation was tested, did not deliver improved results, and therefore was not used. In comparing results with Gerardo, a solution to improve the OLS results was to remove more outliers. Given the concerns expressed above on outliers, I decided it was best not to remove more than the nine chosen.

The PCA results presented in the final report were a result of testing with 1) unscaled and scaled data, 2) including and excluding outlier data points, 3) with all numeric variables and with the removal of three variables with low correlations to other variables, and 4) with unrotated and varimax rotated components. The decision to include all variables in one version of the PCA analysis was because every variable was correlated to another variables with at least a 0.05 significance level. The variables removed were based on a combination of reviewing correlations of variables (untransformed and transformed), significance testing of correlations, initial PCA results, and domain knowledge of what variables might reasonably be combined under PCA. Well into PCA, it became obvious that some variables were less significant resulting in the alternative PCA that removed altitude, moisture, and weight. Along with PCA outputs on fit, mean squared error and the loadings themselves, the visualizations reviewed were variance bar plots to determine the number of components to further explore and the pairwise component plots of the first two and then the latter two components.

I performed CFA on each set of variables. Factor analysis delivered similar results to PCA, but PCA results were preferred due to its inclusion of the Defect component. Once CFA was set aside, my focus was on validating the PCA work with a 60-40 training and test set which proved successful for the most part.

Individual Analysis on OLS

Once I finalized the four components, I created a limited dataset with the response variable (Cupper Points), the four selected components, numeric variables not included in the components and key categorical variables of Color and Processing Method as identified in Gerardo Palacios' work. With additional correlation and also variance inflation factor calculation, I confirmed that multicollinearity had been sufficiently addressed. I performed all subsets OLS as a quick test targeting a t-test significance of 0.01. This was not intended to be a comprehensive analysis but a means of comparison to the regularized regression performed by Gerardo. OLS results were consistent regardless of what set of PCA components were included from the various datasets and the PCA assumptions outlined above. In addition to the review of the F-test, t-test and R-squared results, visualizations of residuals, normal Q-Q and Cook's distance plots were reviewed to determine the normality, heteroscedasticity and the effect of outliers. The all subsets OLS visualization assisted in identifying the variables most effective in predicting Cupper Points. Lastly, standardized beta bar plots identified the overwhelming contribution of the first component to the regression analysis relative to other variables.

Overall Conclusions

Through my analysis, I was able to confirm and address multicollinearity across the predictive variables in the dataset. This involved developing two meaningful principle components for coffee ratings and components that reflected quantity-oriented variables and the bean quality itself. These components along with Processing Methods were found to be significant in predicting Cupper Points whether through my OLS analysis or the regularized regression performed by Gerardo. The components with and without transformed variables also fed the cluster analysis performed by Tom Higgins.

Final Project Milestone 6: Summary of Individual Analysis

Janet Bowen

3/15/2020

Area of Exploration and Data Cleaning

Based on the avenues of research and techniques submitted in Milestone 3, each member in the group chose her or his area of interest. I chose to research processing method, specifically whether or not scores in each of the cupping categories could predict the processing method for each sample. Our research found that there are some flavor characteristics tied with coffee processing method: dry-process produces a coffee heavy in body, wet-process is better for acidic beans, etc. It would be interesting to see if that holds true for these samples.

Before breaking off for our individual analyses, we agreed on a starting point for the dataset so that everything would be consistent. I decided to further subset the data I needed–the ten individual scores and the processing method– and then began to clean and process the data:

- The categorical variable, Processing.Method, had 170 values with no type listed and were removed from the dataset.
- The score variables had a few zero ratings and three with a rating less than 3. Since the samples submitted to the CQI are supposed to be of higher quality, I chose to remove these as outliers as they don't seem to fit with the rest of the dataset. These were additionally confirmed outliers after running the analysis both removing zero scores only and scores less than three.
- I wanted to explore the difference specifically between wet and dry processing. Ninety-six of the observations were of other methods and were also removed.
- The dummy variables were created for the processing category: 1 for wet washed, 0 for dry processing. This was used as the response.

In total, 270 observations were removed from the dataset, leaving 1,069 observations. These observations were broken into an 80-20 split of training and test sets. A summary of the cleaned data can be found in Table 1 of the Appendix.

The Analysis

Logistic Regression

The nature of the response variable led to using logistic regression as the first step in the analysis, using both manual model building and stepwise modeling. Additionally, a corrplot of the variables (Appendix: Figure 1) showed high multicollinearity among many of the cupping scores, which called for regularized regression– specifically Lasso or Elastic Net.

Manual model building and stepwise regression all created the same models, with aroma, flavor, acidity, body, sweetness cupper points noted as significant predictors. Prediction on the test data set showed an accuracy of 73%. I used the testResidual function from the DHARMa package in R to handle residual analysis which showed no significance based on p values of the QQ plots and simulated residuals. However, the ROC curve determined that the model had little to no predictive value which was not surprising given there was little difference between the null deviance and residual deviance numbers in the model summary. The model summary, testResidual plots, and ROC curve can all be found in the Appendix: Table 2, Figure 2 and Figure 3, respectively.

For regularized regression, the data was analyzed using Lasso, due to the multicollinearity, as well as Elastic Net. The result left only body as a significant predictor variable using the lambda.1se criterion and Lasso. Using Elastic Net, the alpha value was determined to be 0.5 after testing the range from 0 to 1 in 0.1 increments. This model added sweetness to the Lasso model, though pulling in the opposite direction. Analysis of prediction results to the test set showed an accuracy rate of 71%, which was worse than the manual model building, though a more parsimonious model. However, an analysis of the model plots showed not much in terms of reducing error for either model. Figures 4 and 5 of the Appendix show the results of the Lasso and Elastic Net models.

Principal Component Analysis and Linear Discriminant Analysis

PCA and PFA

Since logistic regression was not a good fit for the data, principal components of the cupping scores were the next avenue of exploration. The scree plot on the scaled data suggested two components would be sufficient for the dataset using the cutoff at 1, which explained around 70-71% of the variance. This is similar to the results of the logistic regression models. An initial exploration of PCA showed some interesting things related to PC2, in that Sweetness, Uniformity, and Clean.Cup were separated from the remainder of the variables (Appendix: Figure 6).

Rotating the data using varimax and two factors to see showed the sane separation seen in the PCA plot. After some further research on cupping methodology, it is interesting to note that uniformity, clean cup, and sweetness variables are a consistency measure that aggregates scores across three cups of coffee. The remainder are a single score given to that sample. The different methods of scoring might be the reason for the separation of factors seen here. Looking at the cumulative variance explained, it is similar to everything see so far, explaining around 71% of the variance.

The transformed scores of PFA were plugged back into a logistic regression model to see if there was any improvement in the model. However performance remained the same, which can be seen in the ROC curve (Appendix: Figure 7).

Linear Discriminant Analysis

The final method of exploration was Linear Discriminant Analysis. Using lda from the MASS library on the training data to predict the test data showed a very similar accuracy rate to all other forms of analysis so far: 71.5%. Similarly, the ROC curve shows almost no predictive power for the model and histograms of the data show overlapping data across the categories. Results of the LDA can been seen in the Appendix, Figure 8.

Conclusion

Ultimately, the processing method in this data set did not do well as a response variable when using cupper scores as predictor variables. Despite using regression, PCA, PFA, and LDA, the prediction rate was ultimately not better than guessing wet washing for all observations. There could be a few reasons for this result. Due to the data being higher quality, it's possible that the graded samples are going to adjust other methods of processing (such as roasting) in order to make up for what this processing method changes about its bean. It's also possible that processing method does not create distinct enough differences to observe in cupping scores. Either way, the further analysis would likely need to be expanded to include a full range of coffee quality in order to know for sure.

Appendix

*Table 1. Summary of the Final Dataset

```
summary(process_dum2[-c(11:15)])
```

```
##      Aroma           Flavor       Aftertaste       Acidity
##  Min.   :5.080   Min.   :6.08   Min.   :6.170   Min.   :5.250
##  1st Qu.:7.420   1st Qu.:7.33   1st Qu.:7.170   1st Qu.:7.330
##  Median :7.580   Median :7.50   Median :7.420   Median :7.500
##  Mean   :7.564   Mean   :7.51   Mean   :7.387   Mean   :7.532
##  3rd Qu.:7.750   3rd Qu.:7.75   3rd Qu.:7.580   3rd Qu.:7.750
##  Max.   :8.750   Max.   :8.83   Max.   :8.670   Max.   :8.750
##       Body          Balance        Uniformity       Clean.Cup
##  Min.   :6.330   Min.   :6.170   Min.   : 6.000   Min.   : 5.330
##  1st Qu.:7.330   1st Qu.:7.330   1st Qu.:10.000   1st Qu.:10.000
##  Median :7.500   Median :7.500   Median :10.000   Median :10.000
##  Mean   :7.515   Mean   :7.508   Mean   : 9.838   Mean   : 9.858
##  3rd Qu.:7.670   3rd Qu.:7.750   3rd Qu.:10.000   3rd Qu.:10.000
##  Max.   :8.500   Max.   :8.580   Max.   :10.000   Max.   :10.000
##    Sweetness      Cupper.Points   Processing.Method.Washed...Wet
##  Min.   : 6.000   Min.   :5.250   Min.   :0.0000
##  1st Qu.:10.000   1st Qu.:7.250   1st Qu.:1.0000
##  Median :10.000   Median :7.500   Median :1.0000
##  Mean   : 9.894   Mean   :7.479   Mean   :0.7587
##  3rd Qu.:10.000   3rd Qu.:7.670   3rd Qu.:1.0000
##  Max.   :10.000   Max.   :8.750   Max.   :1.0000
```

```
dim(process_dum2)
```

```
## [1] 1069   16
```

*Table 2. Summary of Logit Regression Using Manual Model Building and Stepwise Regression

```
summary(model3)
```

```
##
## Call:
## glm(formula = process_dum$Processing.Method.Washed...Wet ~ .,
##     family = "binomial", data = process_dum[-c(3, 6, 7, 8, 11,
##         12, 13, 14, 15)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2739  -1.2513   0.7100   0.8493   2.3375
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.3700     2.4022  -0.154  0.87759
## Aroma           0.8672     0.3351   2.588  0.00966 **
## Flavor         -1.3178     0.4334  -3.041  0.00236 **
## Acidity         1.4804     0.3511   4.216 2.48e-05 ***
## Body           -1.1247     0.3482  -3.230  0.00124 **
## Sweetness       0.6754     0.1476   4.576 4.74e-06 ***
## Cupper.Points  -0.6404     0.3056  -2.095  0.03614 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1430.9  on 1164  degrees of freedom
## Residual deviance: 1354.1  on 1158  degrees of freedom
## AIC: 1368.1
##
## Number of Fisher Scoring iterations: 4
```

*Figure 1. Correlation Table of Predictor Variables

```
corrplot(washcor, method = "circle", sig.level = 0.9, order="AOE", number.cex=0.75, tl.cex=1)
```

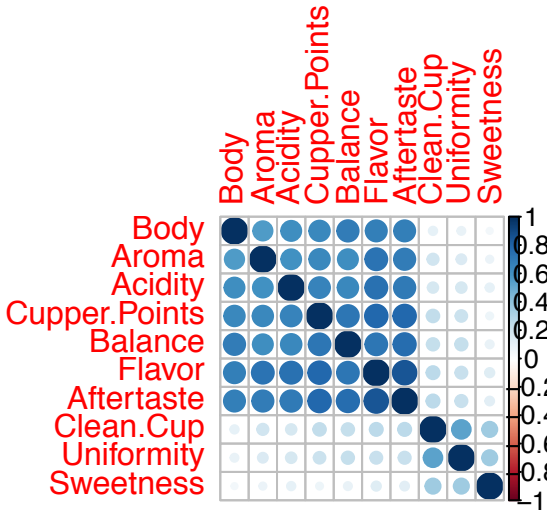
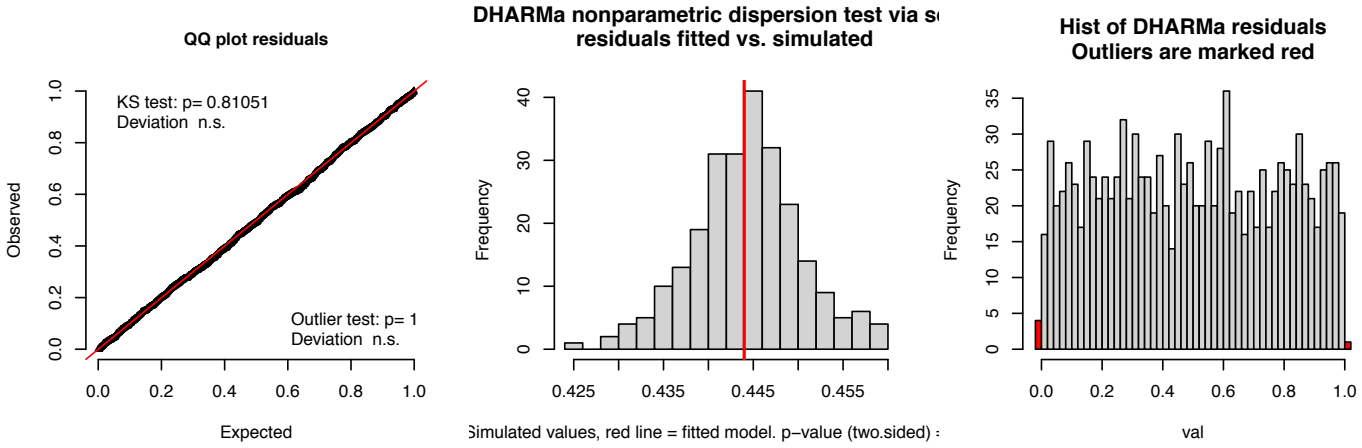


*Figure 2. DHARMa testResidual from Logistic Model

```
fit1 <- simulateResiduals(model3, seed=2020)
testResiduals(fit1)
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
```

```
## D = 0.018689, p-value = 0.8105
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted
##  vs. simulated
##
## data:  simulationOutput
## ratioObsSim = 0.99944, p-value = 0.92
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test
##
## data:  simulationOutput
## outLow = 4.0000e+00, outHigh = 1.0000e+00, nobs = 1.1650e+03,
## freqH0 = 3.9841e-03, p-value = 1
## alternative hypothesis: two.sided


## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.018689, p-value = 0.8105
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted
##  vs. simulated
##
## data:  simulationOutput
## ratioObsSim = 0.99944, p-value = 0.92
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test
##
## data:  simulationOutput
## outLow = 4.0000e+00, outHigh = 1.0000e+00, nobs = 1.1650e+03,
## freqH0 = 3.9841e-03, p-value = 1
## alternative hypothesis: two.sided
```

*Figure 3. ROC Plot from Logistic Model

```
glm_pdct <- ROCR::prediction(glm_test_results,test_y)
glm_prf <- ROCR::performance(glm_pdct, "tpr","fpr")
plot(glm_prf)
```

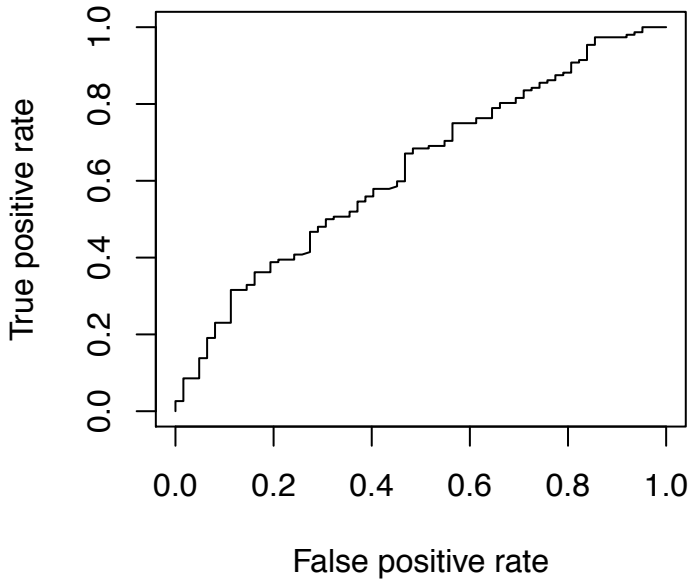
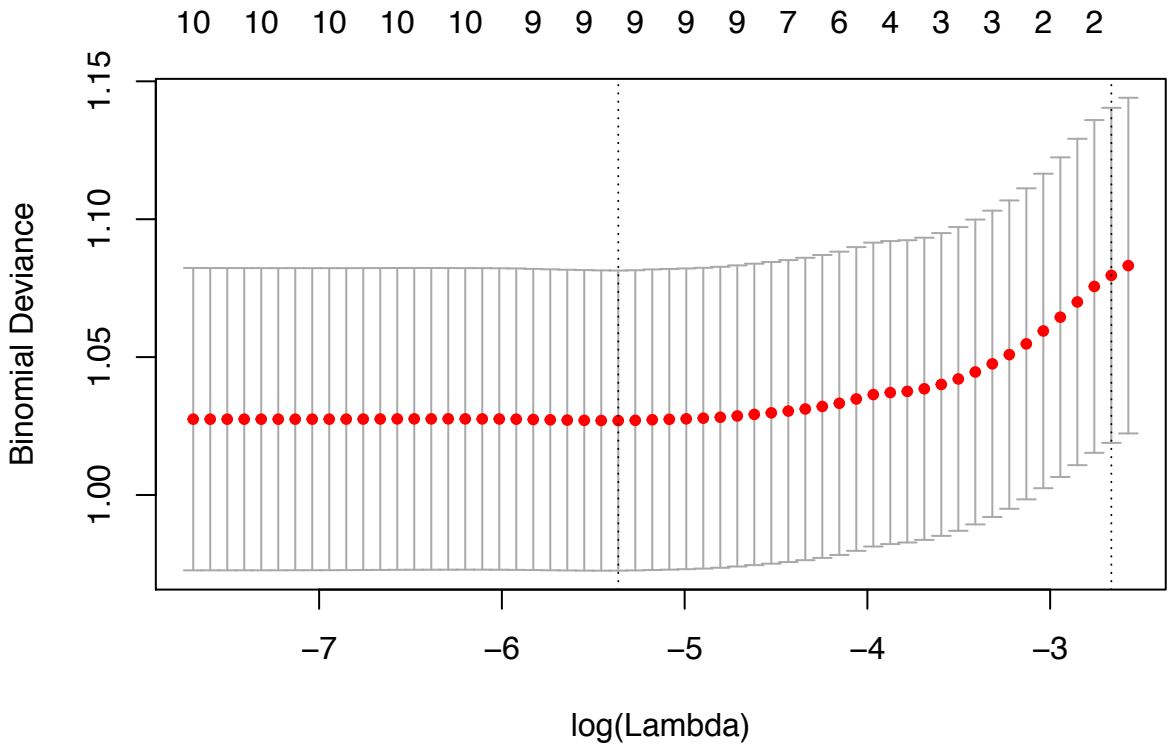


*Figure 4. Result of LASSO

```
plot(cv.lasso)
```

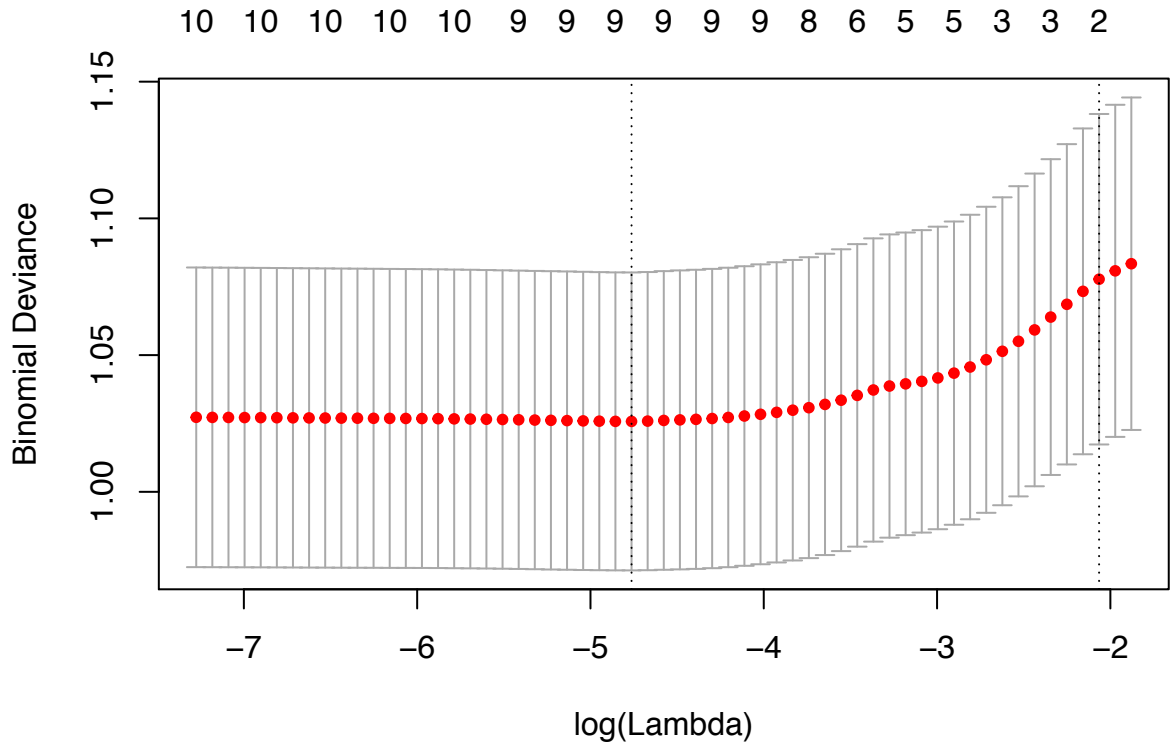


```
coef(cv.lasso, cv.lasso$lambda.1se)
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                             1
## (Intercept)    2.2471197
## Aroma          .
## Flavor         .
## Aftertaste     .
## Acidity        .
## Body          -0.1375876
## Balance        .
## Uniformity     .
## Clean.Cup      .
## Sweetness      .
## Cupper.Points  .
```

*Figure 5. Result of Elastic Net

```
plot(cv.elas)
```




```
coef(cv.elas, cv.elas$lambda.1se)
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                             1
## (Intercept)    2.42431867
## Aroma          .
## Flavor         .
## Aftertaste     .
## Acidity        .
```

```
## Body          -0.19513180
## Balance        .
## Uniformity     .
## Clean.Cup      .
## Sweetness      0.02583458
## Cupper.Points  .
```

*Figure 6. PCA Plot on Cupping Score Variables

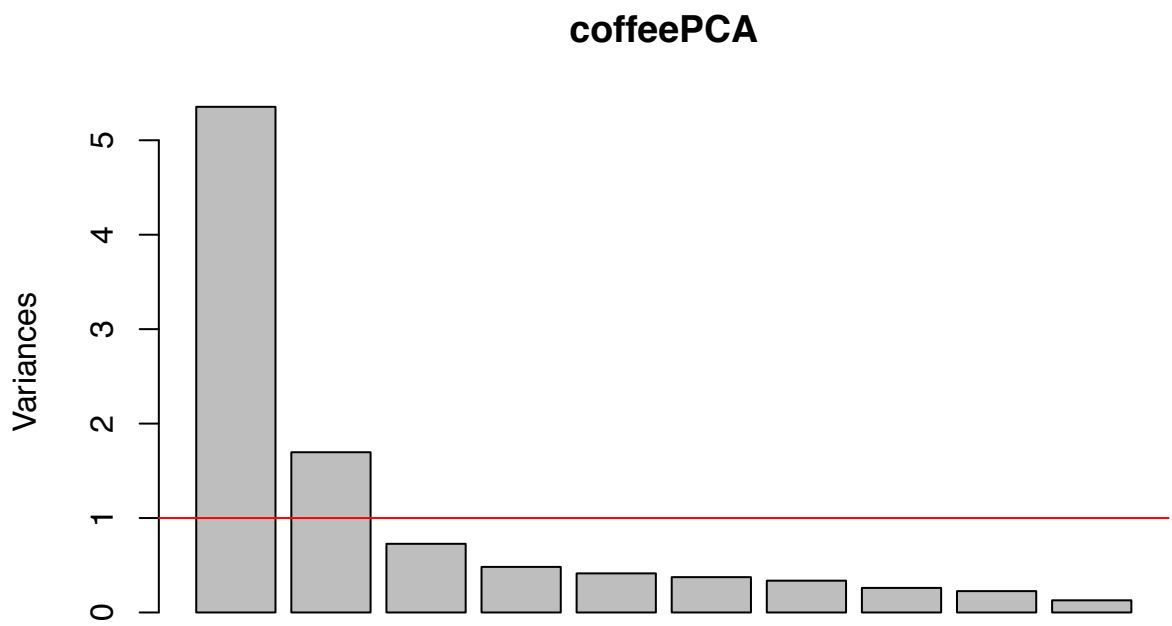
```
coffeePCA = prcomp(train_x, scale=TRUE)
summary(coffeePCA)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     2.3139 1.3025 0.8527 0.69470 0.64328 0.6116 0.58040
## Proportion of Variance 0.5354 0.1697 0.0727 0.04826 0.04138 0.0374 0.03369
## Cumulative Proportion  0.5354 0.7051 0.7778 0.82603 0.86741 0.9048 0.93850
##                           PC8     PC9    PC10
## Standard deviation     0.51006 0.47555 0.35868
## Proportion of Variance 0.02602 0.02261 0.01287
## Cumulative Proportion  0.96452 0.98713 1.00000
```

```
round(coffeePCA$rotation, 2)
```

```
##                PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
## Aroma         0.34  0.07 -0.04  0.70 -0.11 -0.49  0.27 -0.15 -0.18 -0.09
## Flavor        0.40  0.06 -0.02  0.17 -0.04  0.10 -0.14  0.24  0.41  0.74
## Aftertaste    0.39  0.06  0.02  0.04  0.04  0.04 -0.29  0.15  0.56 -0.64
## Acidity       0.35  0.11 -0.08 -0.01 -0.44  0.68  0.36 -0.16 -0.16 -0.11
## Body          0.34  0.19 -0.11 -0.47  0.33 -0.25  0.57  0.35 -0.05 -0.02
## Balance       0.37  0.07  0.06 -0.37  0.11 -0.19 -0.19 -0.79  0.02  0.12
## Uniformity    0.15 -0.59  0.32 -0.28 -0.58 -0.30  0.04  0.16  0.01  0.00
## Clean.Cup     0.16 -0.58  0.37  0.21  0.57  0.30  0.19 -0.09 -0.02  0.00
## Sweetness     0.09 -0.50 -0.85 -0.01  0.06  0.00 -0.08 -0.03 -0.03 -0.01
## Cupper.Points 0.38  0.06  0.07 -0.04  0.09  0.05 -0.53  0.31 -0.67 -0.02
```

```
plot(coffeePCA)
abline(h=1, col="red")
```




```
PCA_Plot(coffeePCA)
```

```
##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

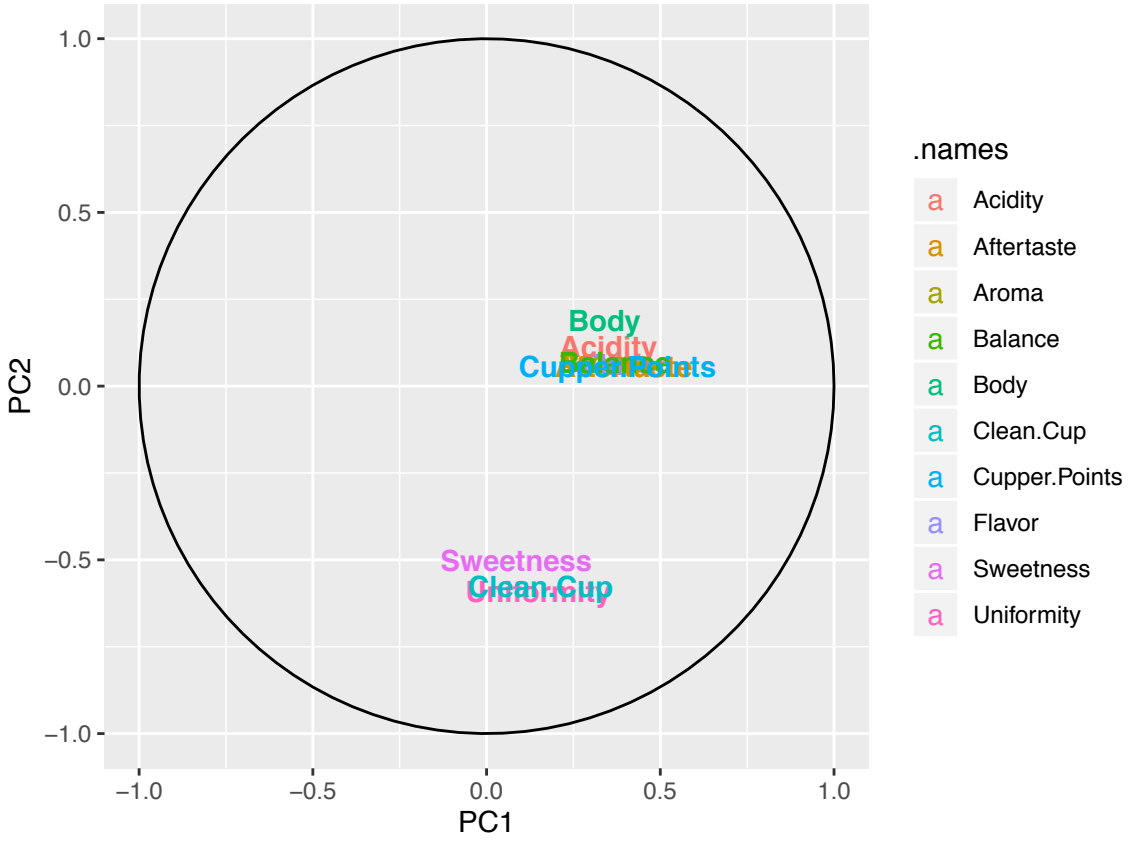


*Figure 7. ROC Plot of Principal Factors Used in Logistic Regression

```
plot(pf_prf)
```

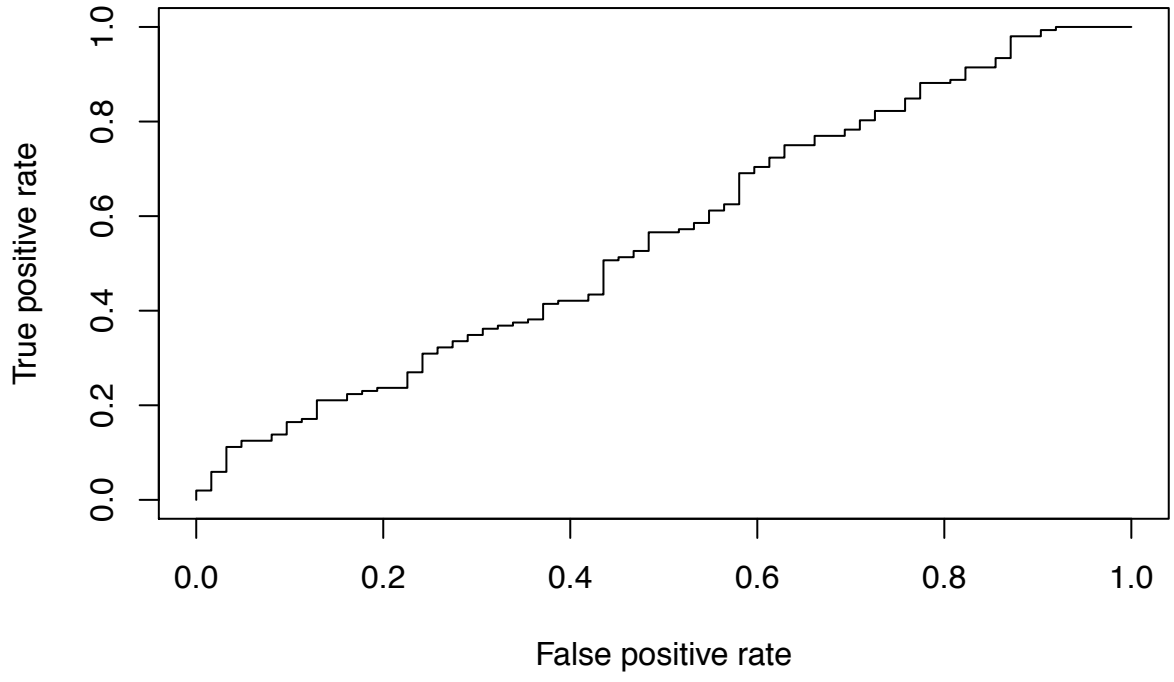


*Figure 8. Linear Discriminant Analysis

```
coffee.lda = lda(Processing.Method.Washed...Wet ~ ., data=train[c(1,2,4,5,9,10,16)])
print(coffee.lda)
```

```
## Call:
## lda(Processing.Method.Washed...Wet ~ ., data = train[c(1, 2,
##     4, 5, 9, 10, 16)])
##
## Prior probabilities of groups:
##         0         1
## 0.2292398 0.7707602
##
## Group means:
##      Aroma   Flavor  Acidity     Body Sweetness Cupper.Points
## 0 7.609388 7.603622 7.556531 7.609031  9.775459      7.592347
## 1 7.562215 7.490030 7.529196 7.488361  9.935038      7.449727
##
## Coefficients of linear discriminants:
##                      LD1
## Aroma          1.0443272
## Flavor        -1.8401418
## Acidity        2.2975844
## Body          -2.2812041
## Sweetness      1.4147950
## Cupper.Points -0.9553156
```

```
print(coffee.lda$scaling[order(coffee.lda$scaling[,1]),])
```

```
##          Body        Flavor Cupper.Points         Aroma     Sweetness
##    -2.2812041    -1.8401418    -0.9553156     1.0443272     1.4147950
##       Acidity
##     2.2975844
```

```
print("Conf Matrix - Train:")
```

```
## [1] "Conf Matrix - Train:"
```

```
table(train$Processing.Method.Washed...Wet, coffee.lda.values$class)
```

```
##
##       0   1
##   0  22 174
##   1  15 644
```

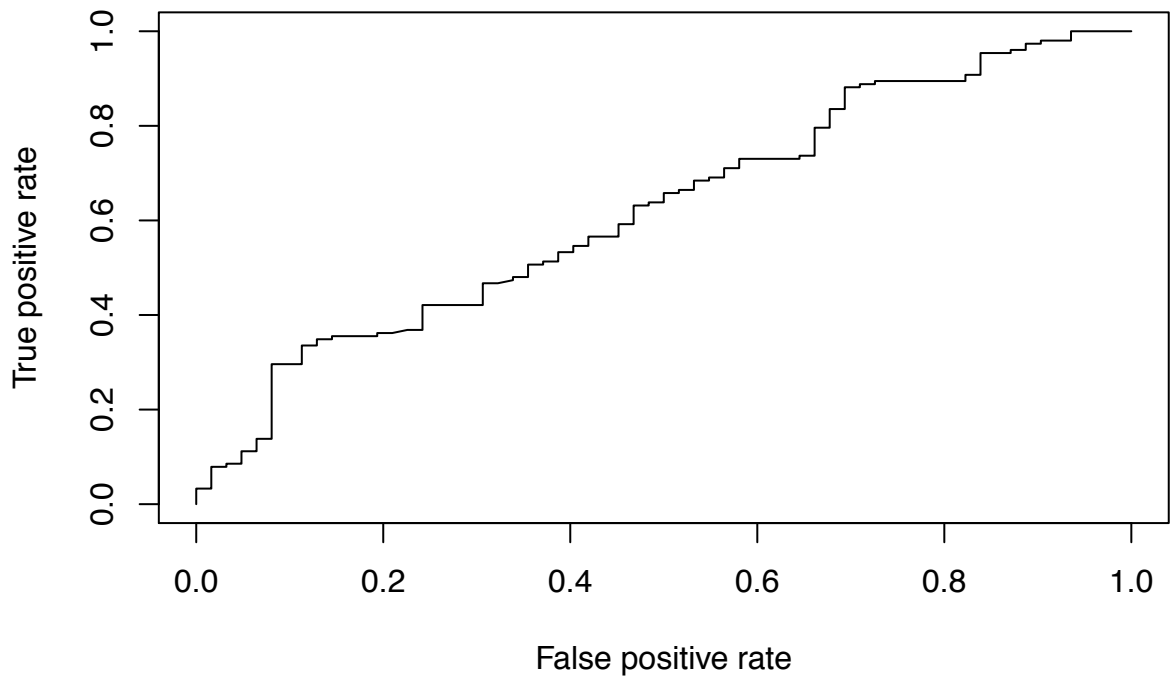
```
print("Conf Matrix - Test:")
```

```
## [1] "Conf Matrix - Test:"
```

```
table(test$Processing.Method.Washed...Wet, coffee.lda.values2$class)
```

```
##
##       0   1
##   0   6  56
##   1   4 148
```

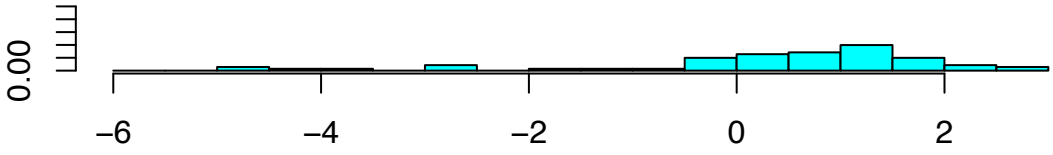
```
ldapred <- ROCR::prediction(coffee.lda.values2$posterior[,2], test_y)
ldaperf <- ROCR::performance(ldapred,"tpr","fpr")
plot(ldaperf)
```



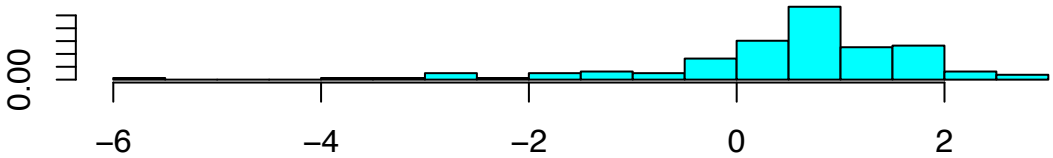

```
print(ROCR::performance(ldapred,"auc"))
```

```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
## [[1]]
## [1] 0.6229839
##
##
## Slot "alpha.values":
## list()
```

```
ldahist(data=coffee.lda.values$x[,1], g=as.matrix(test_y))
```



group 0



group 1