Key Drivers for Quality Coffee

Bosniack, Lesley, Bowen, Janet, Higgins, Thomas, & Palacios, Gerardo

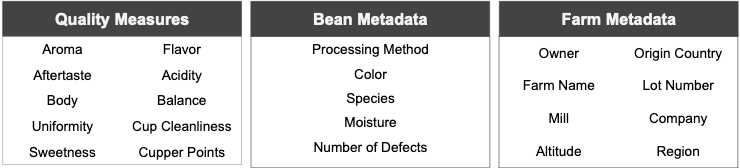
DSC 424 - Advanced Data Analysis

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*](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.

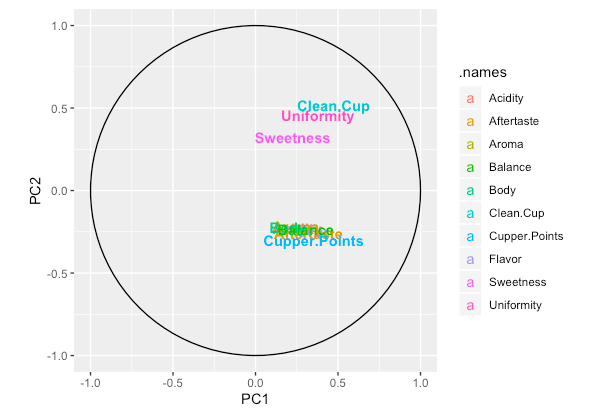
The main purpose of our analysis was to figure out if there were associations between the quality ratings of coffee and other information contained in the dataset, and if so, whether these had predictive power. We also investigated to see if there were other patterns in the data. The variables in the 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.



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.

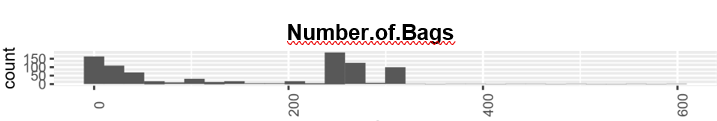
Finally, we ran calculations to see if we could use our data to separate the coffee by processing method. This was not successful; a 70% success rate may seem decent at first, but unfortunately this is roughly what would've happened if we had just guessed that every coffee was produced using the most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

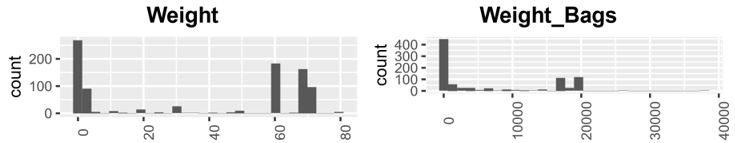
In summary, we were able to discover relationships between the quality grades and other variables in our data. We were not successful at predicting the categories of coffees contained in the dataset, despite trying several methods - this could be because the coffee in the dataset is biased towards being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

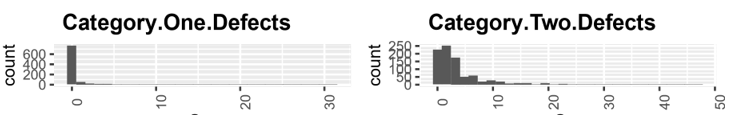
**Technical Summary**

*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.

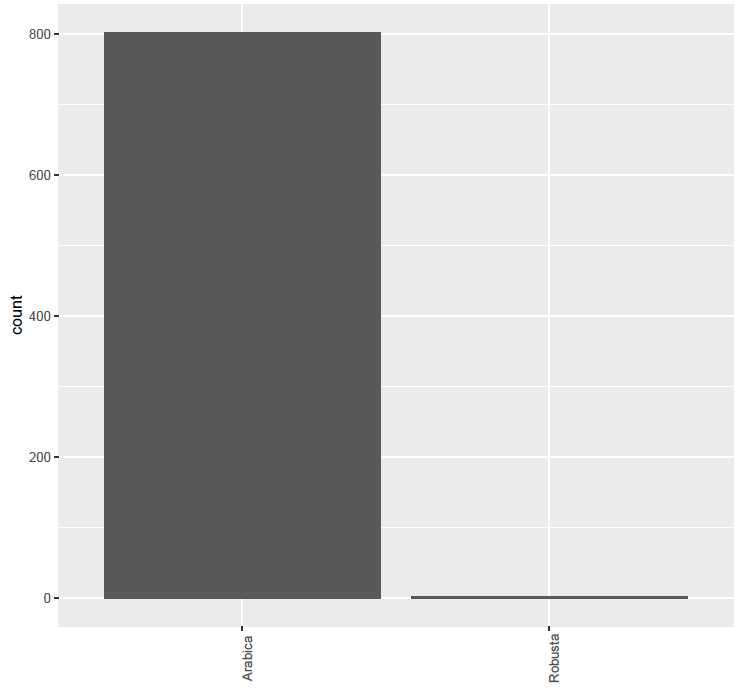
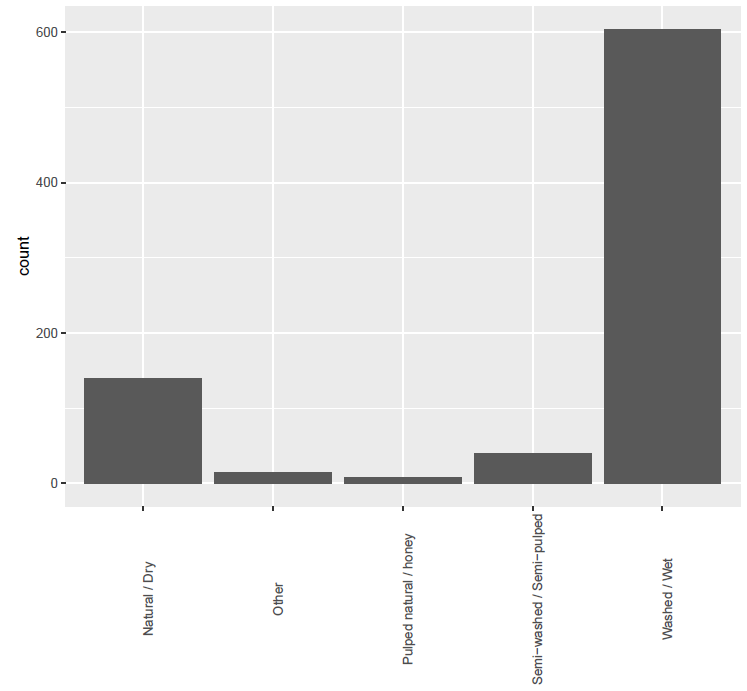
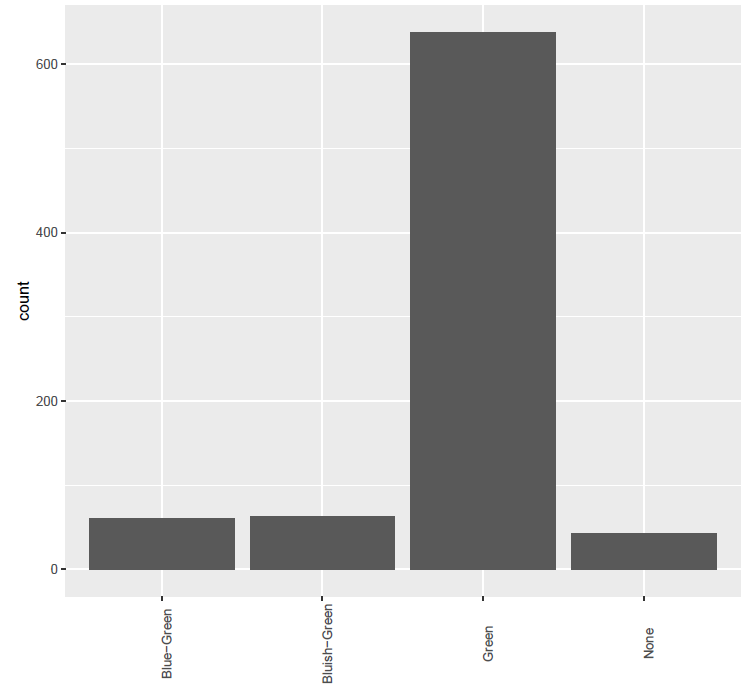




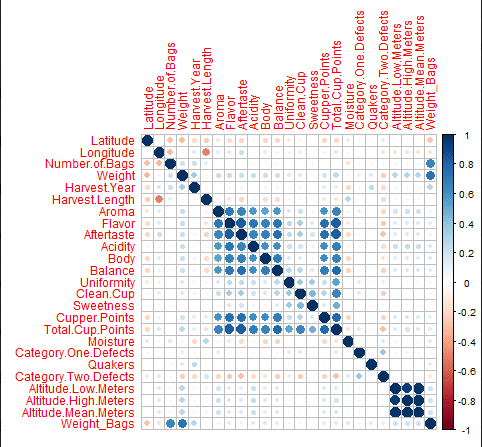


Some of the categorical variables were highly disproportionate with the majority or most of the observations falling under one level. Color, Processing Method and Species are examples.

**Color Processing Method Species**



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.



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*

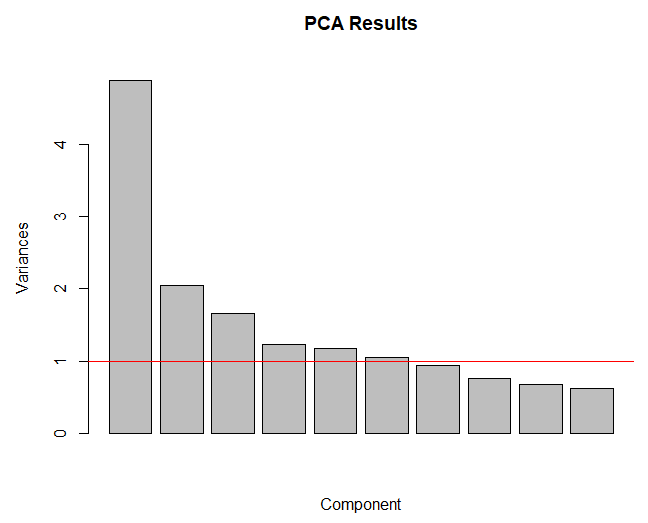
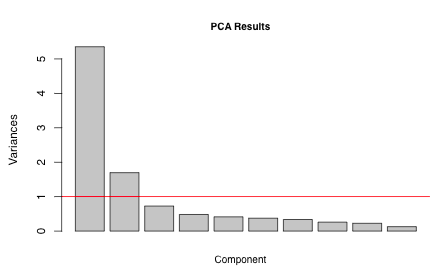
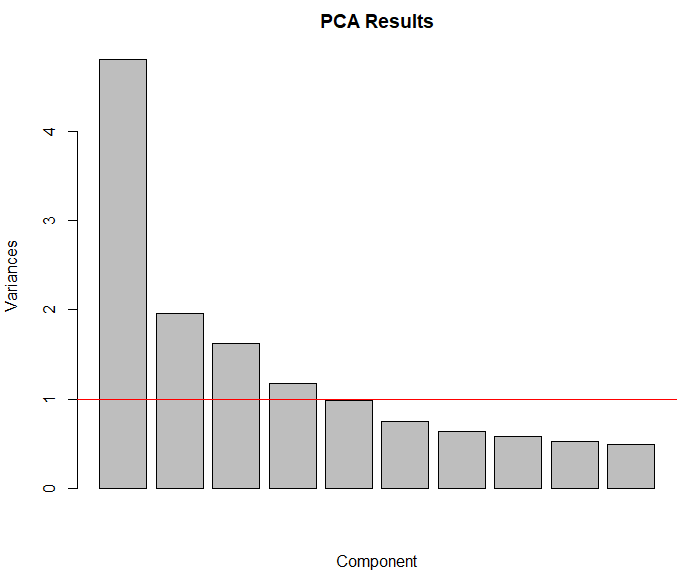
*Principal Components Analysis (“PCA”) and Least Squares Regression (“OLS”)*

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 Coffee Scores Only Excluding Altitude,**

**Moisture and Weight**

|  |  |  |
| --- | --- | --- |
| Note: 5 or even 6 components above Var>1 capture 61-67% of the variance. Elbow is between the 4th and 5th components. | Note: 2 components above Var>1 capture 71% of the variance. Elbow is between the 2nd and 3rd components. | Note: 4 or 5 components above Var>1capture 60-73% of the variance. Elbow is between the 4th 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.

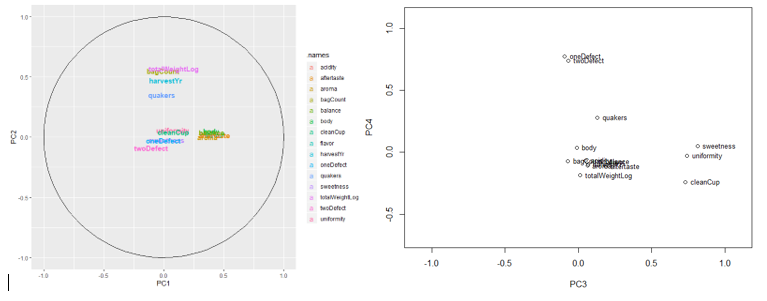
Outliers were identified based on observed values of variables and through OLS analysis. Nine data points were removed. The nine outliers exhibited very low or very high Cupper Points whereas the values for other coffee scores were not at extreme values. Given that initial OLS identified coffee scores 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.

* 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**

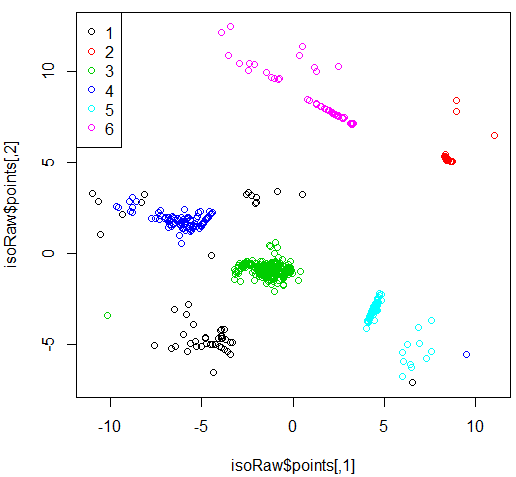
For this analysis, we wanted to see if the grades and quality characteristics of coffee visually cluster from this dataset. The quality scores and the defect measurements were used as inputs here; of these, the defect measurements were subject to a logarithmic transform. A final step of cleaning: in 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.

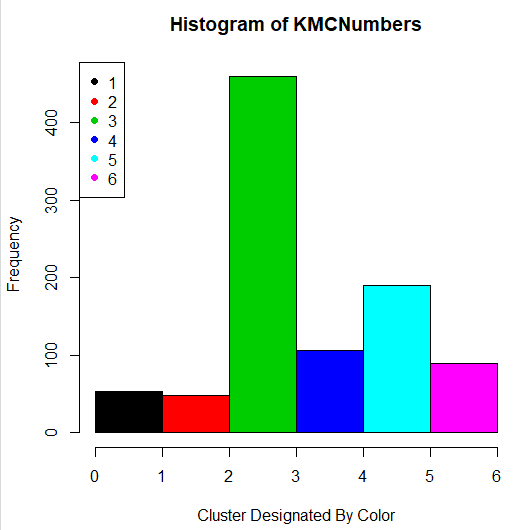
Having done that, we began by plotting both the classic solution for multidimensional scaling, and a converged isometric solution, both in 2 dimensions. The calculations for the isometric solution also gave us stress measurements for both, which could be used to tell us how distorted the visualization 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.

### 

### 

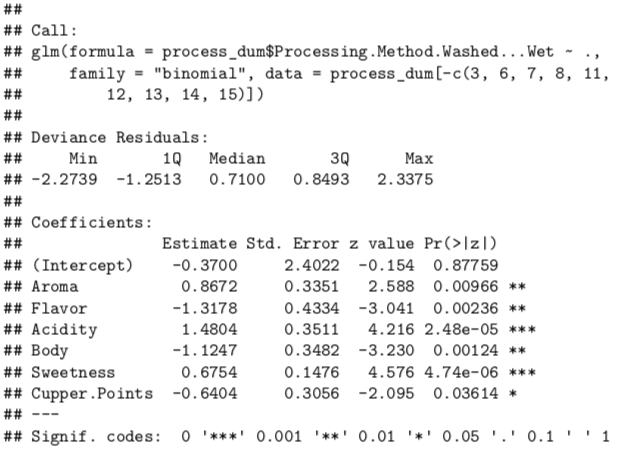
### Results

### 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.

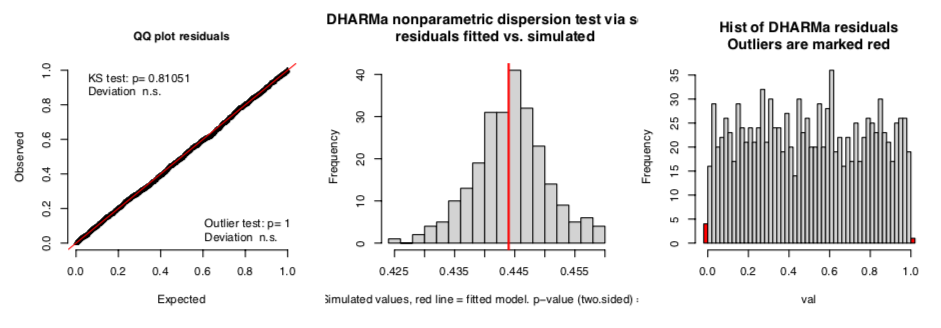
The beta coefficients suggest that the most important variables that can explain the variability in Cupper Points are **Flavor, Aftertaste, Acidity, Body and Balance and Altitude.** More importantly, it suggests that Flavor, Aftertaste and Balance are the strongest contributing indicators by beta weight 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.

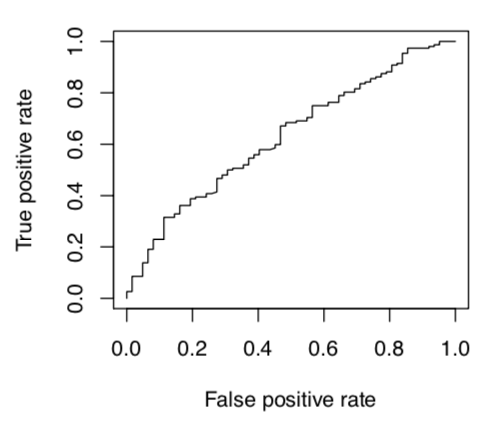
**Logistic Regression of Processing Method and LDA**

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.

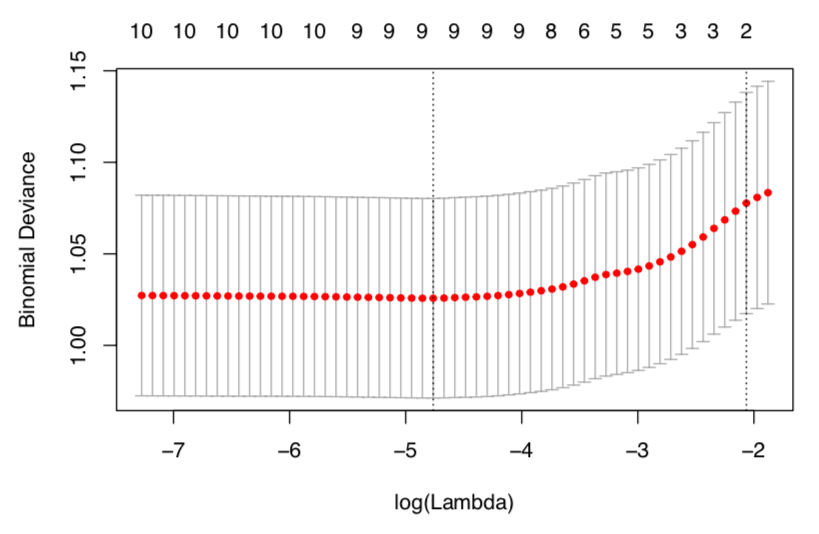
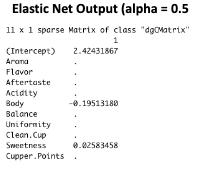


The data was split into a training (80%) and test (20%) split. Manual model building and stepwise regression all created the same models, with Aroma, Flavor, Acidity, Body, Sweetness, and Cupper Points noted as significant predictors. The DHARMa package in R has a testResidual function to handle residual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

****

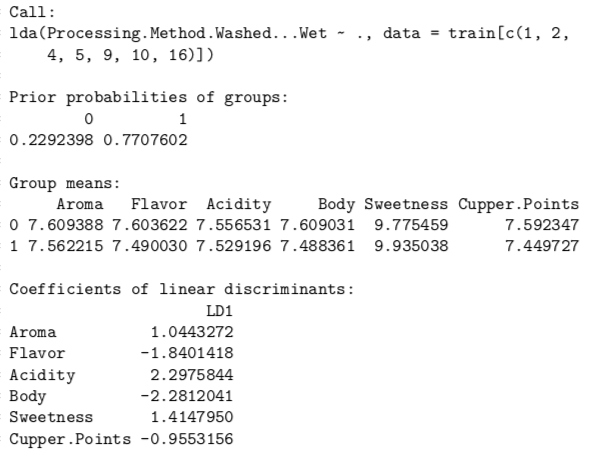
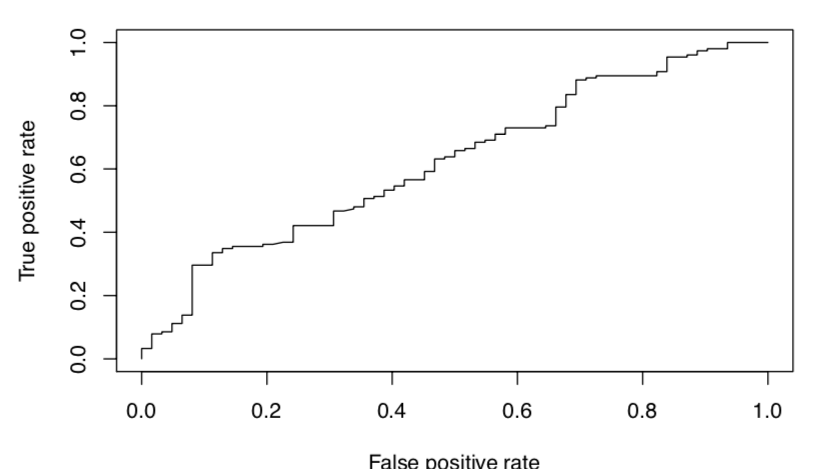
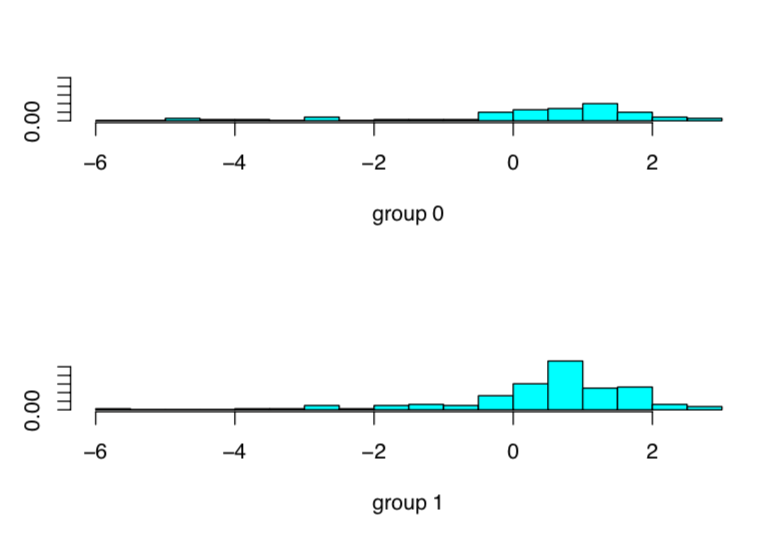


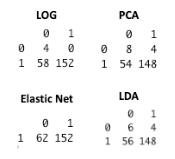
However, the ROC curve (right) determined that the model had little to no predictive value, which was not surprising given there was little difference between the null deviance and residual deviance numbers in the model summary. Prediction of the test data using the training model showed an accuracy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



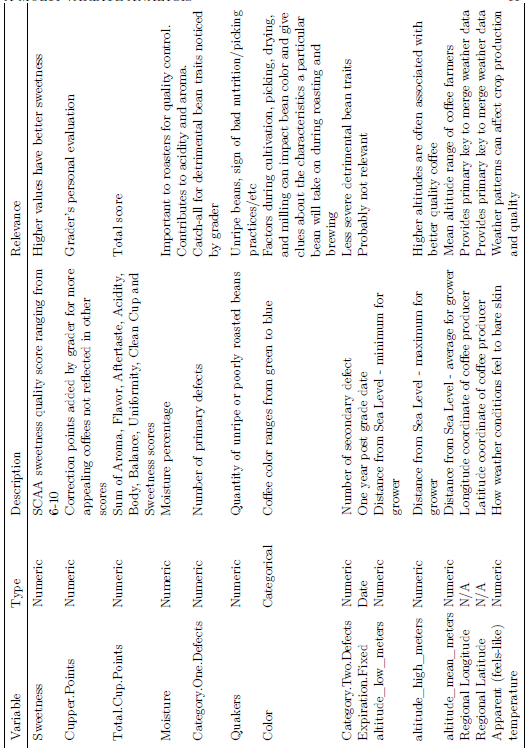
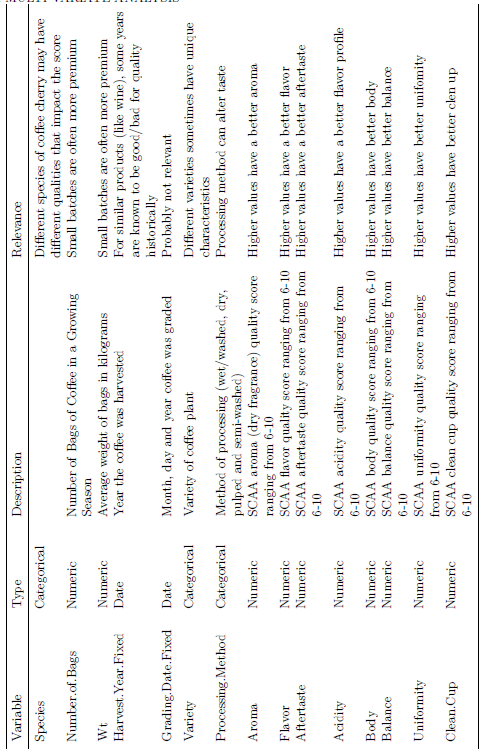
A confusion matrix comparison of all of the methods show that even though prediction accuracy rate is 70-71% for all, each model under-predicts dry processing (0 value) by quite a bit. In actuality, the model does not predict any better than if one were to guess wet processing for every observation (which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

**Concluding Remarks**

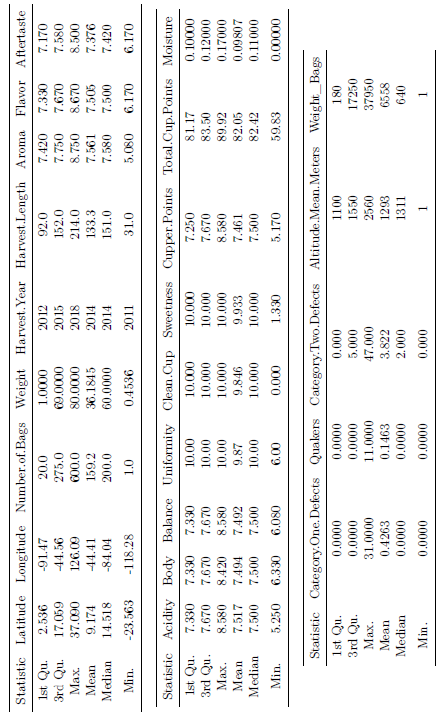
According to the CQI, quality is one of the most important variables that influence a coffee’s value. The goal of the CQI is to improve the quality of coffee and the lives of coffee producers by providing producers “access to the tools and support they need to understand the quality of their coffee, 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**

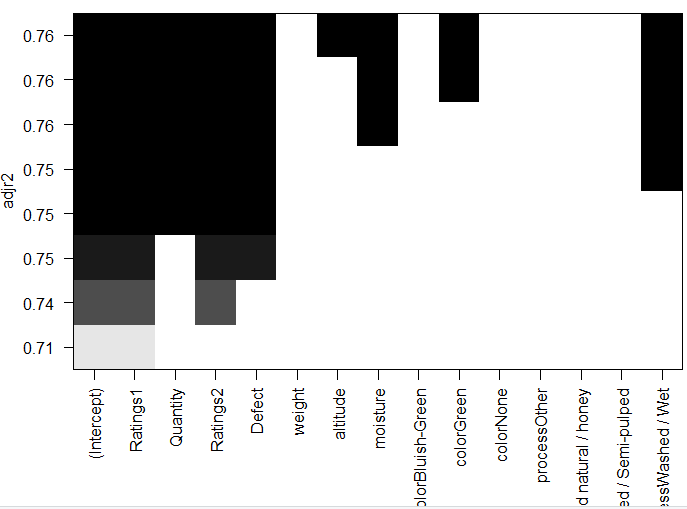


**Appendix - Summary Statistics of the Variables**

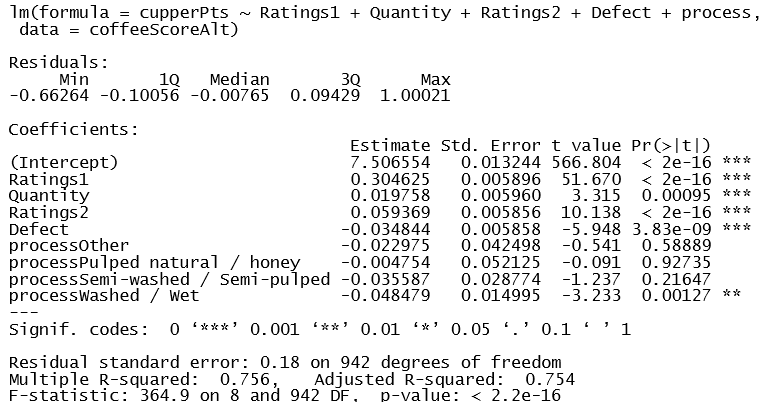
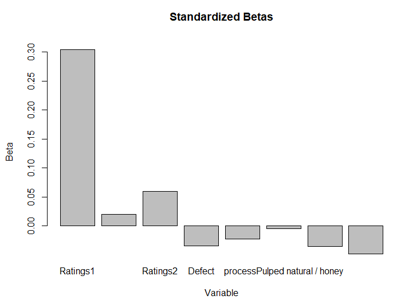
****

**Appendix - OLS**

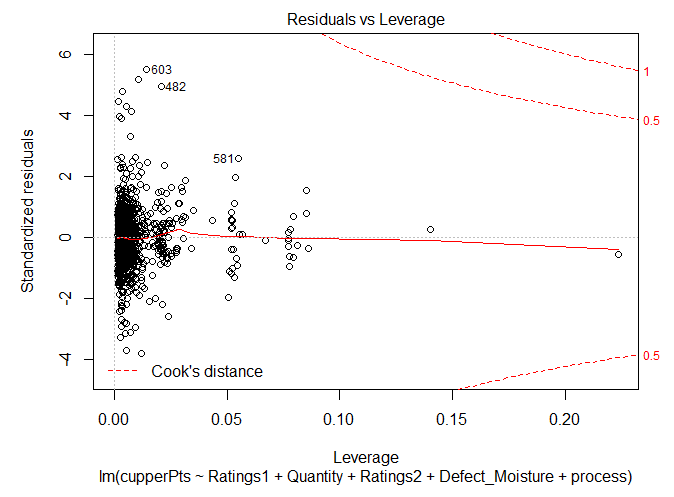
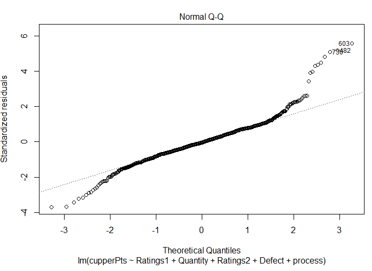
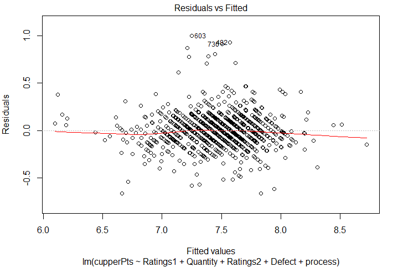
**Significant variables from All Subsets OLS regression is below.**

****

**Below is the chosen OLS model and model results.**.

****

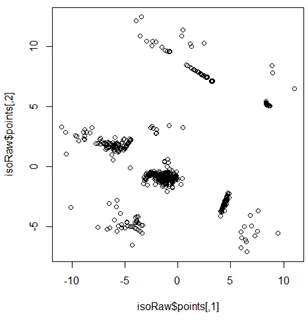
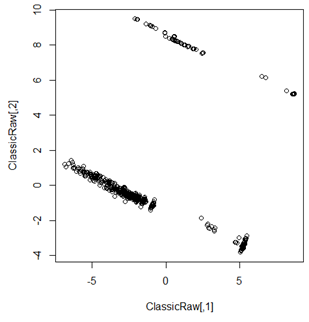
**OLS diagnostic plots are below.**



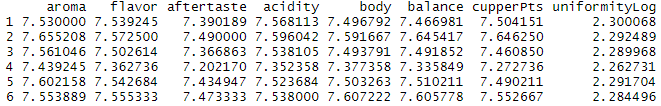


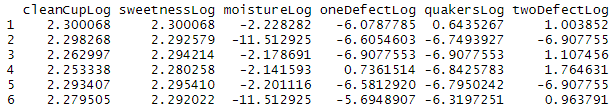
**Appendix - Additional Charts for MDS and Cluster Analysis**

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

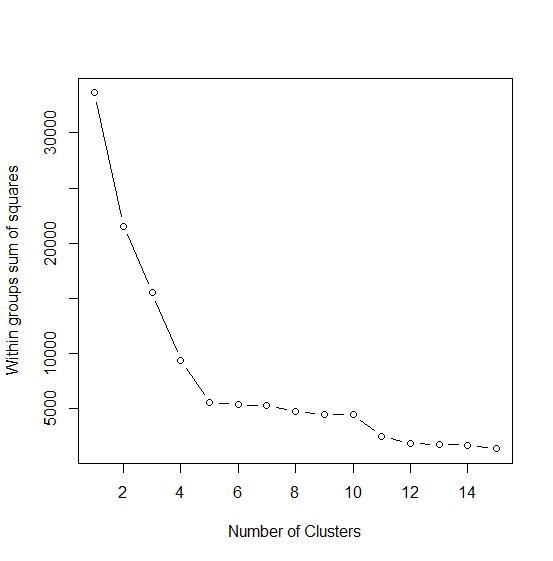


The cluster centers from the k-means clusters are below; from top to bottom, with rows denoted by cluster.

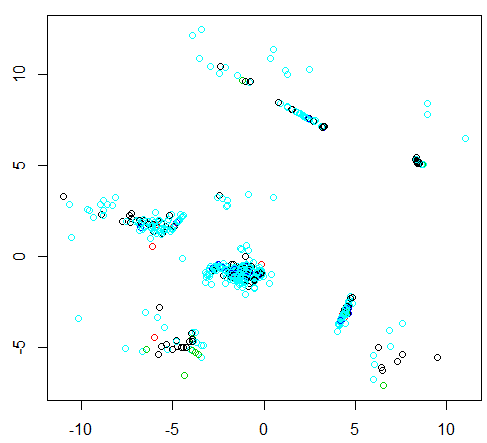
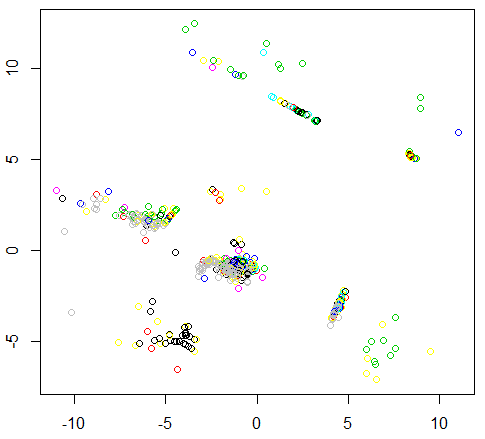




Below is a scree plot, used as a diagnostic to pick the number of centers for k-means clustering. The sharpest bend happens at 5 clusters, although another drop happens at 11. Any number of clusters below 5 would be a bad choice - having a too high within groups sum of squares tends to indicate that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