Final Project Report

Coffee Ratings

Caitlin Kostna
7856868

04.13.2023
STAT 3690

1 Abstract

Coffee beans of different species are grown and harvested in various countries around the world. Each type has a distinct portfolio with various features of taste. People often have different opinions of the strength of these features in their coffee. The purpose of the following research is to see the effect of different features on the quality of the coffee, and whether being from a specific place, species, or processing method has an impact on the taste.

2 Introduction

2.1 Data Description

The datasets were gathered from the Coffee Quality Institute (CQI) in January 2018. They contain reviews from specialized reviewers for two types of coffee. There are three CSV files:

- An Arabica coffee pre-cleaned dataset
- A Robusta coffee pre-cleaned dataset
- A dataset constructed by merging the above datasets

Each dataset contains a variety of features that fall into three categories: quality measures, bean metadata, and farm metadata.

The following analysis uses the merged dataset, including some further cleaning that:

- only includes relevant columns
- removes rows that were missing a *Processing Method*
- adjusts column names where necessary
- simplifies some of the country names (i.e. change “United States (Hawaii)” to “Hawaii”)
- simplifies the names of the *Processing Methods* (i.e. change “Natural / Dry” to “Dry”)

2.2 Data Source

There is a GitHub repository containing the data, but it is also found on *Kaggle* at this link: Coffee Quality Data. The data was scraped directly from the Coffee Quality Institute website.

2.3 Exploratory Analysis

The following data visualizations allow us to get an idea of what values our coffee ratings take on, and if there are any interactions between our numerical and categorical variables.

It is important to note that the ratings for *Aroma*, *Flavour*, *Aftertaste*, *Acidity*, *Body*, *Balance*, *Uniformity*, *Clean Cup*, *Sweetness*, and *Overall* are summed to get the *Total Cup Points* for each observation. We will refer to these ratings as the set of **quality measures** for the remainder of the analysis.

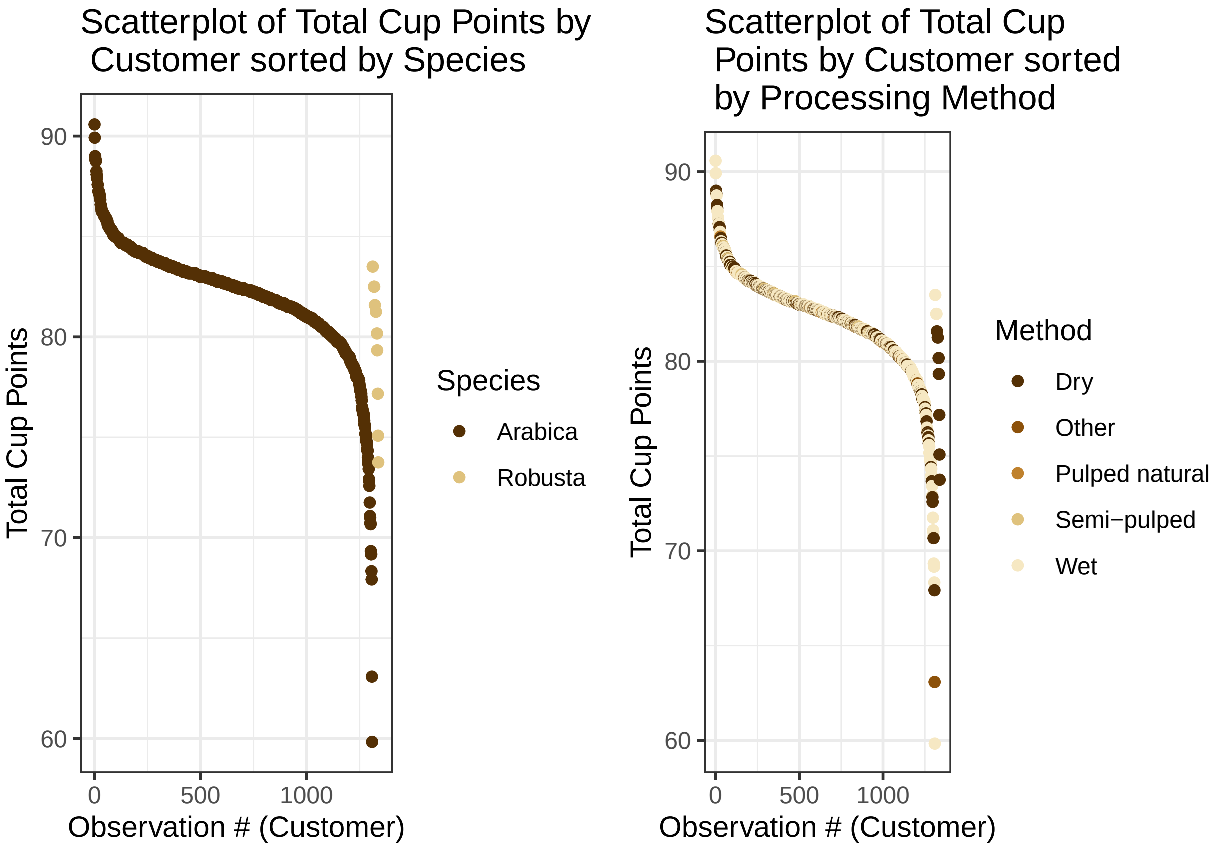

Figure 1: Dual Scatterplot of Total Cup Points

In figure 1, specifically in the plot on the left, we can see that there are far more ratings for Arabica coffee than Robusta coffee. While it is a bit difficult to identify each observation's colour in the plot on the right, the majority of the points correspond to the "Wet" *Processing Method*, and the next most being "Dry".

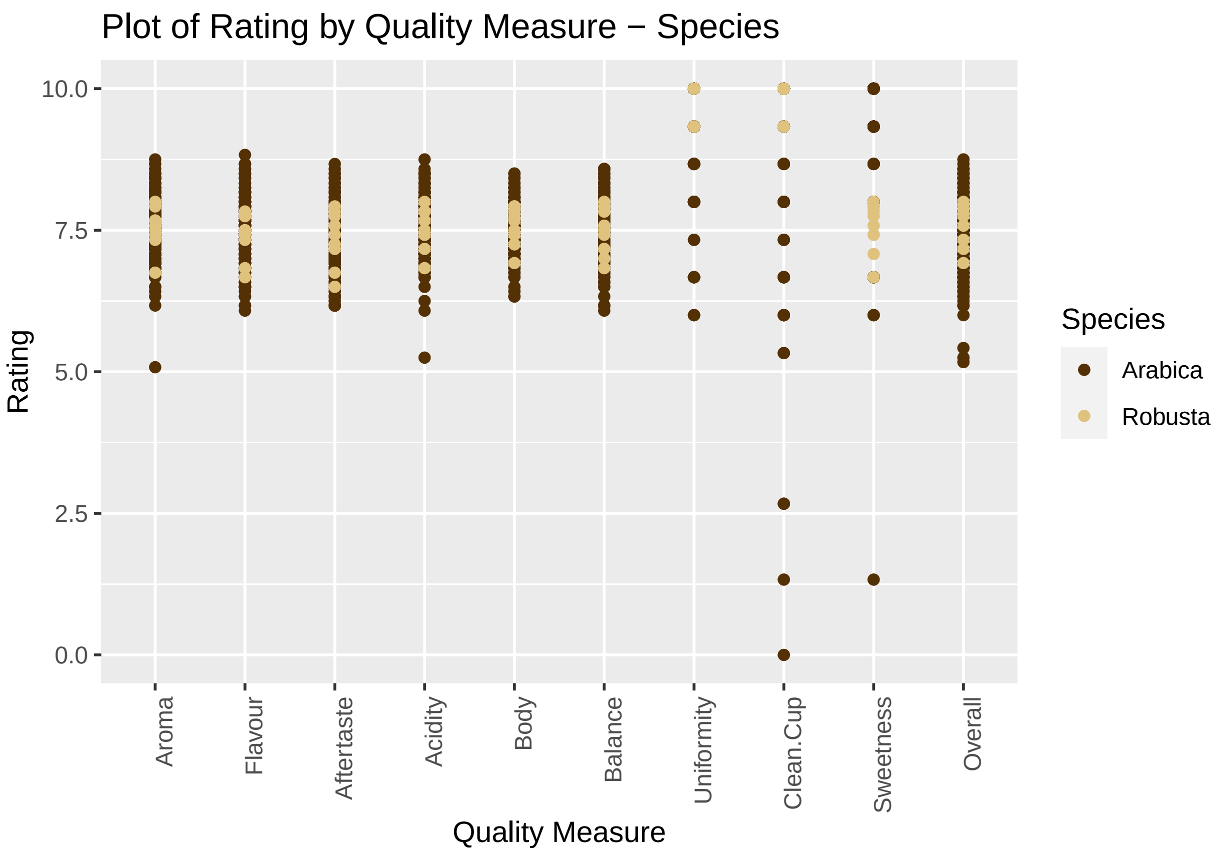

Figure 2: Scatterplot of quality measure ratings

Figure 2 breaks down the graph on the left hand side of figure 1 to see each quality measure. We can see that the ratings of the quality measures for the Robusta *Species* tend to be around the aspect's mean, while the ratings for the Arabica *Species* are more spread out.

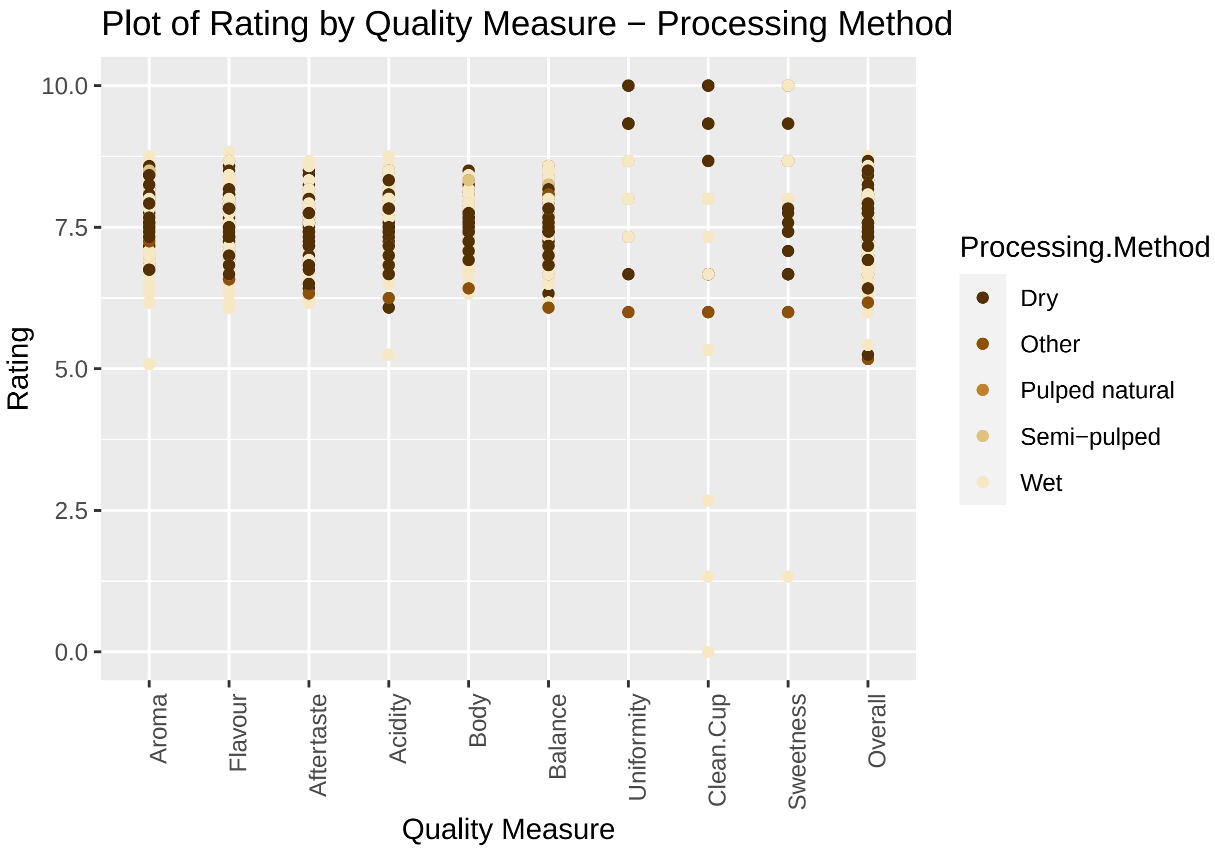

Figure 3: Scatterplot of Quality Measure ratings

Figure 3 breaks down the graph on the right hand side of figure 1 to see each quality measure. We can see that there is a decent distribution of colour for each quality measure. Since one row of our dataset has an entry for each quality measure, we know that there is at least one value for each *Processing Method* for each of our aspects. It is evident that some of the ratings for Wet could be considered outliers as they fall under five.

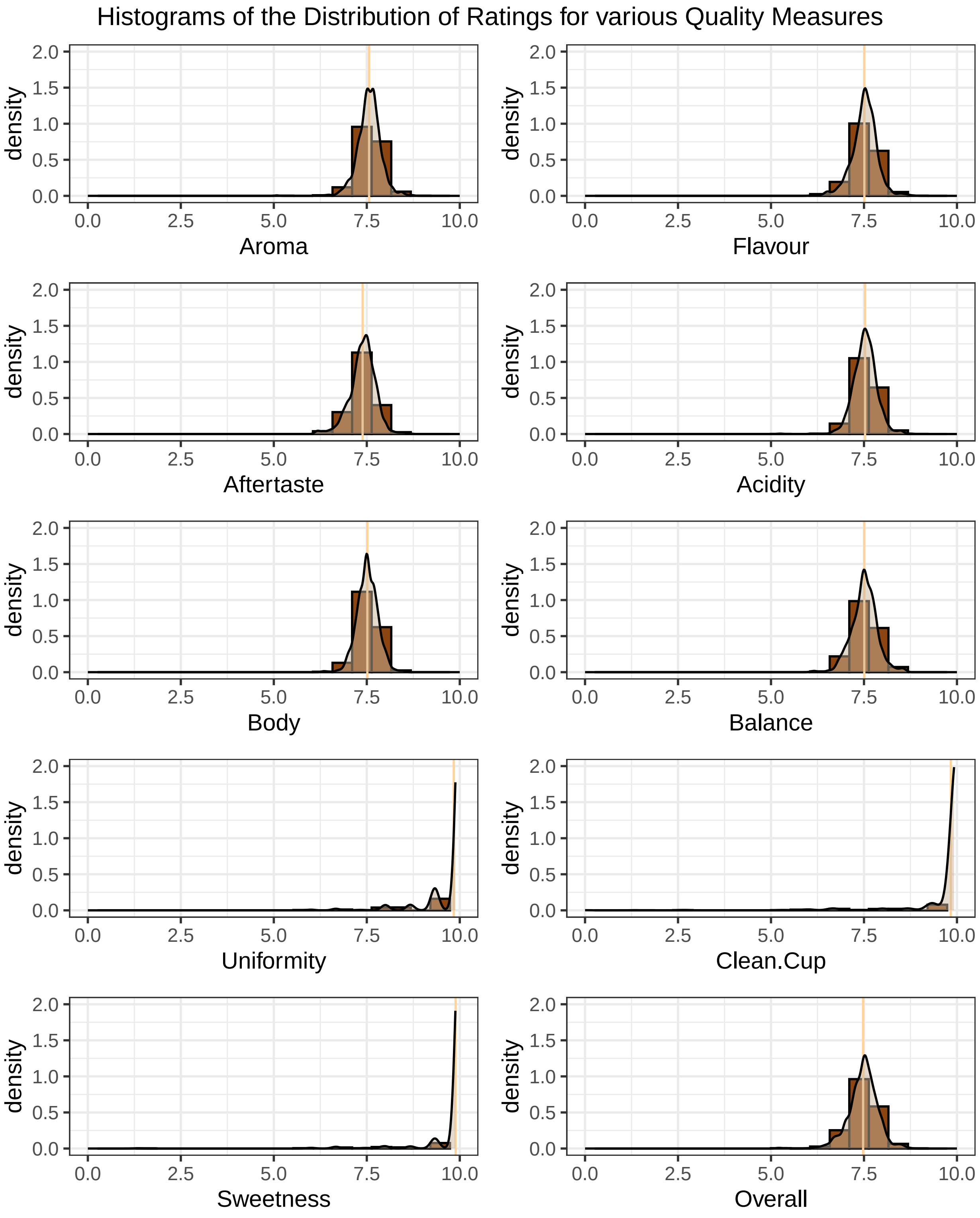

Figure 4: Histograms of Quality Measures

In figure 4, we can see the true distributions of the ratings for all quality measures. The vertical line in each histogram is the mean. The mean for *Aroma*, *Flavour*, *Aftertaste*, *Acidity*, *Body*, *Balance*, and *Overall* seem to be around 7.5, and their distributions appear to be normal. The mean for *Uniformity*, *Clean Cup*, and *Sweetness* seem very close to 10, and their distributions appear to be severely left skewed. It will be interesting to perform two tests to see if those two groups of aspects have similar means within them.

2.4 Purpose of Further Statistical Analysis

The main goals of the following statistical analysis is to identify how each variable in our dataset is structured, and how variables are related to or influence one another. In figure 1, it was evident that the *Total Cup Points* had a bit of variation depending on the *Species* or *Processing Method*. Further analysis will be conducted to see if we can predict the *Total Cup Points* based on these two factors. In figures 2 and 3, it was evident that there is some clustering in *Species* or *Processing Method* for their ratings of each quality measure. Further analysis will be conducted to see if we can classify the *Species* or *Processing Method* based on the ratings for each quality measure. Lastly, an analysis will be conducted to see if the means for the various quality measures are equal, as the histograms in figure 4 suggest such a possibility.

3 Methods

For consistency, any tests or intervals will be constructed at a **95%** level of significance.

3.1 Check Normality

First, we check to see if the main numerical columns we will be working with approximately follow a normal distribution. This is to ensure that any further assumptions of normality for tests or models are fulfilled. This also determines if the mean is applicable as a representative value of our data or not.

3.2 Inference on μ

3.2.1 Tests of Means

As seen in figure 4, the quality measures could be split into two groups where they appeared to have similar medians within. We will perform a hypothesis test to determine whether or not there is a significant difference in their means. Our assumption is that each column of our dataset follows a multivariate normal distribution with different means and the same covariance matrix.

3.2.2 Simultaneous Confidence Intervals

Once performing tests of means for the ratings of quality measures, we will create multiple confidence intervals (original, Bonferroni, Scheffé) and compare which one is more precise. Our assumptions are the same as for the test of means.

3.3 Modeling

3.3.1 Univariate Linear Regression

Using two categorical variables (*Species* and *Processing Method*) coded as indicator variables, we will see if they have an effect on the *Total Cup Points*. Then, we will see which confidence intervals for the coefficients contain zero. Our assumption to create the confidence interval is that we have normality.

3.4 Classification

Lastly, we will implement linear discriminant analysis (LDA) twice - once using the *Processing Method* as the labels and once using the *Species* as the labels - with the various quality measures as our data. Our assumptions are that the sample measurements are independent, the distributions of the quality measures follow a normal distribution, and the covariance matrix is identical across each class (or type of *Processing Method* or *Species*).

4 Results

First, we check to see that our variables approximately follow a normal distribution using **qqPlots**.

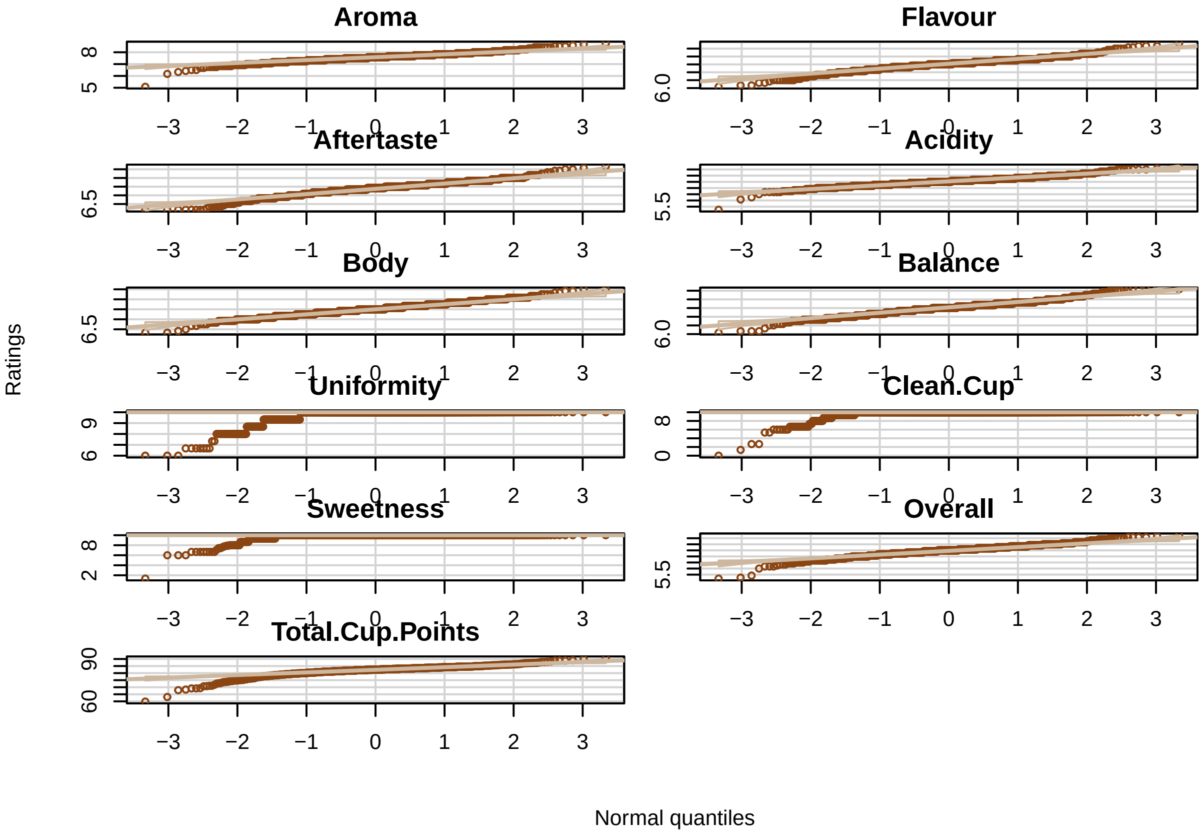

Figure 5: qqPlots to Check Normality

According to figure 5, the ratings for *Aroma*, *Flavour*, *Aftertaste*, *Acidity*, *Body*, *Balance*, *Overall*, and *Total Cup Points* all seem to fall within the shaded region along their respective lines, so we can assume they follow a normal distribution. *Uniformity*, *Clean Cup*, and *Sweetness* have a decent amount of ratings creating a left tail, so normality is a little questionable.

Our first test of means will be between *Aroma*, *Flavour*, *Aftertaste*, *Acidity*, *Body*, *Balance*, and *Overall* as they all seemed to have the same median in our boxplot. Let μ_1 be the mean rating for *Aroma*, μ_2 be the mean rating for *Flavour*, μ_3 be the mean rating for *Aftertaste*, μ_4 be the mean rating for *Acidity*, μ_5 be the mean rating for *Body*, μ_6 be the mean rating for *Balance*, and μ_7 be the mean rating for *Overall*.

Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$ vs H_1 : otherwise

Name of Approach: Likelihood Ratio Test (LRT)

Level of Significance: $\alpha = 0.05$

Test Statistic: 806.5755

Rejection Region: $[12.6925, \infty)$

P-value: ≈ 0

Conclusion: As our p-value $\approx 0 \leq 0.05 = \alpha$, we reject H_0. Conclude that there is sufficient evidence that the average rating across *Aroma*, *Flavour*, *Aftertaste*, *Acidity*, *Body*, *Balance*, and *Overall* does vary.

Below are simultaneous confidence intervals for the μ_1 to μ_7 as defined for the above test.

Table 1: Simultaneous Confidence Intervals for first test of means

	Orig L	Orig U	Bonf L	Bonf U	Sche L	Sche U
Aroma	7.545	7.580	7.538	7.587	7.528	7.597
Flavour	7.491	7.530	7.484	7.537	7.473	7.547
Aftertaste	7.369	7.408	7.361	7.415	7.351	7.426
Acidity	7.511	7.547	7.505	7.554	7.495	7.564
Body	7.499	7.531	7.493	7.537	7.484	7.545
Balance	7.489	7.528	7.482	7.536	7.471	7.546
Overall	7.454	7.500	7.446	7.508	7.433	7.520

By looking at table 1, we can see that the original confidence interval provides the smallest interval for each variable, and is thus the best confidence interval to use.

Our second test of means will be between *Uniformity*, *Clean Cup*, and *Sweetness* as they all seemed to have similar distributions in our boxplot. Let μ_1 be the mean rating for *Uniformity*, μ_2 be the mean rating for *Clean Cup*, and μ_3 be the mean rating for *Sweetness*.

Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$ vs H_1 : otherwise

Name of Approach: Likelihood Ratio Test (LRT)

Level of Significance: $\alpha = 0.05$

Test Statistic: 10.5921

Rejection Region: $[6.0120, \infty)$

P-value: 0.0052

Conclusion: As our p-value $= 0.0052 \leq 0.05 = \alpha$, we reject H_0. Conclude that there is sufficient evidence that the average rating across *Uniformity*, *Clean Cup*, and *Sweetness* does vary.

Below are simultaneous confidence intervals for the μ_1 to μ_3 as defined for the above test.

Table 2: Simultaneous Confidence Intervals for first test of means

	Orig L	Orig U	Bonf L	Bonf U	Sche L	Sche U
Uniformity	9.810	9.868	9.804	9.874	9.776	9.902
Clean Cup	9.793	9.879	9.784	9.889	9.743	9.930
Sweetness	9.860	9.919	9.853	9.926	9.824	9.955

By looking at table 2, we can see that once again, the original confidence interval gives the smallest interval for each variable, as is thus the best confidence interval to use.

Now, we will fit a linear regression model to predict the *Total Cup Points* using the *Species* and *Processing Method*. Since our model will use categorical variables, each level for each variable will be coded as indicator variables. Our reference group for *Species* is Arabica, and our reference group for *Processing Method* is Dry.

```
#
# Call:
# lm(formula = Total.Cup.Points ~ Species + Processing.Method,
#     data = coffee_rating)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -22.147  -0.978   0.436   1.523   8.603
#
# Coefficients:
#                                  Estimate Std. Error t value Pr(>|t|)
# (Intercept)                        82.314      0.169  485.79   <2e-16 ***
# SpeciesRobusta                    -2.530      0.861   -2.94   0.0034 **
# Processing.MethodOther            -1.035      0.555   -1.86   0.0625 .
# Processing.MethodPulped natural    0.494      0.740    0.67   0.5044
# Processing.MethodSemi-pulped       0.320      0.398    0.80   0.4218
# Processing.MethodWet              -0.337      0.194   -1.74   0.0824 .
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
#
# Residual standard error: 2.7 on 1163 degrees of freedom
# Multiple R-squared:  0.0141,  Adjusted R-squared:  0.00983
# F-statistic: 3.32 on 5 and 1163 DF,  p-value: 0.00556
```

For simplicity, the specific names for the types of *Species* and *Processing Methods* are abbreviated for our model. Our regression equation is:

$$\hat{y} = 82.314 - 2.530 * SR - 1.035 * PMO + 0.494 * PMPN + 0.320 * PMSP - 0.337 * PMW$$

Since our model has categorical variables (using indicator variables for the different levels), we can separate it into smaller models.

- Mean rating for Arabica species and Dry processing method: $\hat{y}_{AD} = 82.314$
- Mean rating for Arabica species and Other processing method: $\hat{y}_{AO} = 82.314 - 1.035 = 81.279$
- Mean rating for Arabica species and Pulped natural processing method: $\hat{y}_{APN} = 82.314 + 0.494 = 82.808$
- Mean rating for Arabica species and Semi-pulped processing method: $\hat{y}_{ASP} = 82.314 + 0.320 = 82.634$
- Mean rating for Arabica species and Wet processing method: $\hat{y}_{AW} = 82.314 - 0.337 = 81.977$
- Mean rating for Robusta species and Dry processing method: $\hat{y}_{RD} = 82.314 - 2.530 = 79.784$
- Mean rating for Robusta species and Other processing method: $\hat{y}_{RO} = 82.314 - 2.530 - 1.035 = 78.749$
- Mean rating for Robusta species and Pulped natural processing method: $\hat{y}_{RPN} = 82.314 - 2.530 + 0.494 = 80.278$
- Mean rating for Robusta species and Semi-pulped processing method: $\hat{y}_{RSP} = 82.314 - 2.530 + 0.320 = 80.104$
- Mean rating for Robusta species and Wet processing method: $\hat{y}_{RW} = 82.314 - 2.530 - 0.337 = 79.448$

If the *Species* is Robusta, it has a large negative impact on the mean rating. This might be because there are few ratings in the dataset that fall in this category. If the *Processing Method* is Other, it also has a decent negative impact on the mean rating. Two *Processing Methods* have a positive impact on the mean rating. The combination that gives the highest mean rating is a *Species* of Arabica and a *Processing Method* of Pulped natural.

	2.5 %	97.5 %
(Intercept)	81.9812	82.6461
SpeciesRobusta	-4.2193	-0.8401
Processing.MethodOther	-2.1244	0.0540
Processing.MethodPulped natural	-0.9579	1.9463
Processing.MethodSemi-pulped	-0.4611	1.1009
Processing.MethodWet	-0.7164	0.0433

By looking at the above confidence intervals, four of them contain zero and it is all for the *Processing Methods*. Since those intervals contain zero, it is unclear as to whether or not there is a treatment effect. This lines up with the fact that the p-values for those coefficients (found in the linear regression output above) were not less than 0.05.

Our first classification task is to see if we can predict the *Processing Method* based on the ratings for the ten quality measures. From our linear discriminant analysis (LDA) model, we can see the proportion of each group in our training set.

```
#            Dry          Other Pulped natural    Semi-pulped            Wet
#        0.22564        0.02179        0.01410        0.04487        0.69359
```

The majority of ratings in our training set are of the Wet *Processing Method*. Then, we can look at the proportion of trace of our model, which is the variance explained by each linear discriminant function.

```
# [1] 0.69759 0.17149 0.10220 0.02871
```

The first linear discriminant function explains almost 70% of the variance, our second explains 17%, and the rest of the variance is between the last two discriminant functions. Lastly, we can use our model to perform some predictions on our test set, and calculate the accuracy.

```
# [1] 0.7095
```

Thus, this LDA model correctly identifies the *Processing Method* based on the quality measure ratings 70.95% of the time.

Our second classification task is to see if we can predict the *Species* based on the ratings for the ten quality measures. From our LDA model, we can see the proportion of each group in our training set.

```
#  Arabica  Robusta
# 0.991026 0.008974
```

The majority of ratings in our training set are of the Arabica *Species*. Then, we can look at the proportion of trace of our model, which is the variance explained by each linear discriminant function.

```
# [1] 1
```

The only linear discriminant function explains all of the variance. Lastly, we can use our model to perform some predictions on our test set, and calculate the accuracy.

```
# [1] 0.9871
```

Thus, this LDA model correctly identifies the *Species* based on the quality measure ratings 98.71% of the time.

5 Conclusion

While one person's opinion on coffee can differ from the next, it is interesting to see if we can find any patterns among ratings. Our exploratory analysis gave us some insight into the potential relationships between the different aspects of coffee, the type of coffee, and how the coffee was processed. We proceeded with utilizing various methods to take a deep dive in.

The majority of our quality measures follow a normal distribution, with only three of them having a left tail, leaving their normality questionable. For those that do follow a normal distribution, their mean can be used as a representative value.

Our tests of means revealed two rejections of our null hypotheses, indicating that at least one of the means in each group was not equal to the others. This was surprising as figure 4 had strong indications of such an effect. The tightest confidence intervals for both tests were the ones with no corrections.

Our linear regression analysis revealed that the Robusta *Species* and the Other *Processing Method* have a large negative impact on the mean rating. The coffee with the highest mean rating is an Arabica *Species* using a Pulped natural *Processing Method*. The coefficient confidence intervals revealed some uncertainty in treatment effectiveness.

Both of the linear discriminant analysis (LDA) models performed well. The model that used the *Processing Method* as the labels had an accuracy of 70%, and the model that used the *Species* as the labels had an accuracy of 98%.

As with any data analysis, there are limitations. One limitation is that there are unequal quantities of ratings when separated by *Species* and *Processing Method*. In particular, 99% of *Species* is Arabica, and 70% of *Processing Method* is Wet. This imbalance contributed to the performance of LDA, and led to the inability to implement quadratic discriminant analysis (QDA) as there wasn't enough of each type in the training and testing sets after splitting the data. A concern is that the dataset is rather small (~1200) and since the ratings are strictly opinion-based, it is unclear whether or not the same person rated multiple cups of coffee. If similar collection of this data is to be completed in the future, it would be beneficial to ensure that each type of *Species* and *Processing Method* are equally represented in the sample, as well as having the same people taste multiple kinds of coffee. This will provide more accurate data to use for modelling and classification.