Final Project: Coffee Ratings

Avery Klauke, Justin Hartenstein, Cecilia Diaz

2023-12-04

Introduction

With coffee being consumed by 30-40% of the world's population daily and over 70 countries participating in its production, we aim to conduct an analysis of the factors contributing to the overall appeal of coffee. In this report, we introduce a linear regression model that depicts the relationships between the overall score of a cup of coffee and its distinct characteristics, as measured by experts in the field. This statistical approach is crucial for the understanding of key aspects that influence on coffee ratings. Moreover, it serves as a foundation to give recommendations on the attributes that contribute to a good cup of coffee. In the next sections we will expand on the methodology used to build the regression model as well as the discussion of our findings.

Data Description

The data comes from TidyTuesday, who previously collected it from the Coffee Quality Database. The specific dataset used was provided by the Coffee Quality Institute in January 2018. It contains data on two types of coffee beans: Robusta and Arabica. Both of them are rated by professionals on several aspects like sweetness, acidity, fragrance and color. It includes 1,339 observations and 23 variables, four of which are categorical. The following table details the variables included in the dataset:

Variable	Class	Description
total_cup_points	double	Total rating/points (0 - 100 scale)
species	character	Species of coffee bean (arabica or robusta)
number_of_bags	double	Number of bags tested
harvest_year	character	When the beans were harvested (year)
processing_method	character	Method for processing
aroma	double	Aroma grade
flavor	double	Flavor grade
aftertaste	double	Aftertaste grade
acidity	double	Acidity grade
body	double	Body grade
balance	double	Balance grade
uniformity	double	Uniformity grade
clean_cup	double	Clean cup grade
sweetness	double	Sweetness grade
cupper_points	double	Cupper Points
moisture	double	Moisture Grade
category_one_defects	double	Category one defects (count)
quakers	double	Quakers
color	character	Color of bean
category_two_defects	double	Category two defects (count)
altitude_low_meters	double	Altitude low meters
altitude_high_meters	double	Altitude high meters

Variable	Class	Description
altitude_mean_meters	double	Altitude mean meters

Explanatory Data Analysis

We first began by removing all of the descriptive information related to manufacturing partners, lot numbers, etc. We also removed the altitude_high and altitude_low variables because of the similarities and correlation to altitude_mean variable. We can notice that the variables "harvest_year", "quakers", processing_method", "color", and "altitude_mean_meters" have missing values. To deal with this will omit all rows with missing values due the number of them being small across the dataset.

We will start by examining the summary statistics:

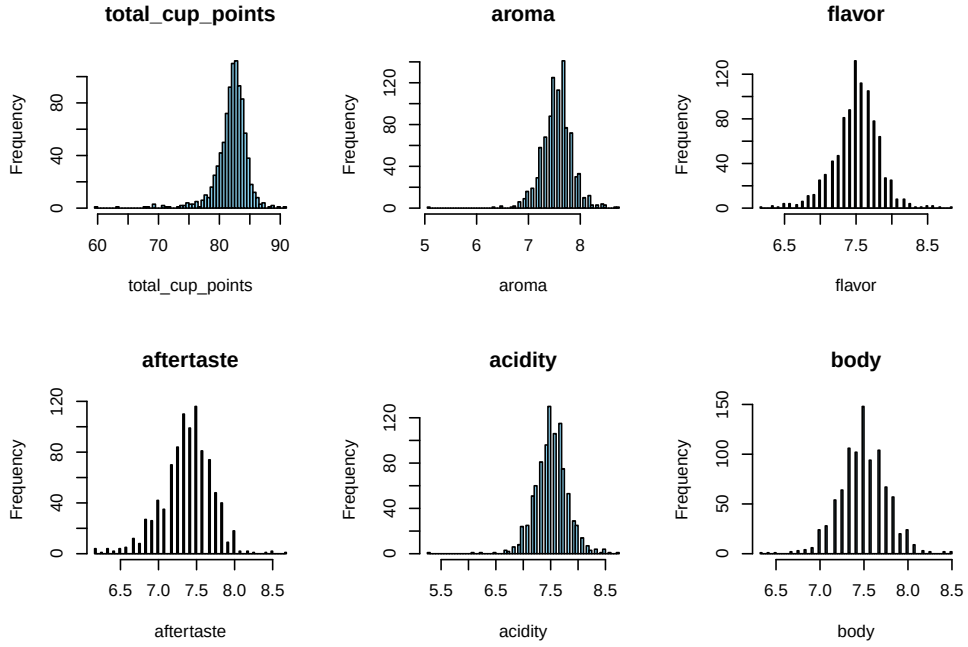
Variable	N	Mean	SD	Min	Max	Range	SE
total_cup_points	1339	82.09	3.50	0.00	90.58	90.58	0.10
species	1339			Inf	-Inf	-Inf	
harvest_year	1292			Inf	-Inf	-Inf	
processing_method	1169			Inf	-Inf	-Inf	
aroma	1339	7.57	0.38	0.00	8.75	8.75	0.01
flavor	1339	7.52	0.40	0.00	8.83	8.83	0.01
aftertaste	1339	7.40	0.40	0.00	8.67	8.67	0.01
acidity	1339	7.54	0.38	0.00	8.75	8.75	0.01
body	1339	7.52	0.37	0.00	8.58	8.58	0.01
balance	1339	7.52	0.41	0.00	8.75	8.75	0.01
uniformity	1339	9.83	0.55	0.00	10.00	10.00	0.02
clean_cup	1339	9.84	0.76	0.00	10.00	10.00	0.02
sweetness	1339	9.86	0.62	0.00	10.00	10.00	0.02
cupper_points	1339	7.50	0.47	0.00	10.00	10.00	0.01
moisture	1339	0.09	0.05	0.00	0.28	0.28	0.00
category_one_defects	1339	0.48	2.55	0.00	63.00	63.00	0.07
quakers	1338	0.17	0.83	0.00	11.00	11.00	0.02
color	1121			Inf	-Inf	-Inf	
category_two_defects	1339	3.56	5.31	0.00	55.00	55.00	0.15
altitude_mean_meters	1109	1775.03	8668.63	1.00	190164.00	190163.00	260.31

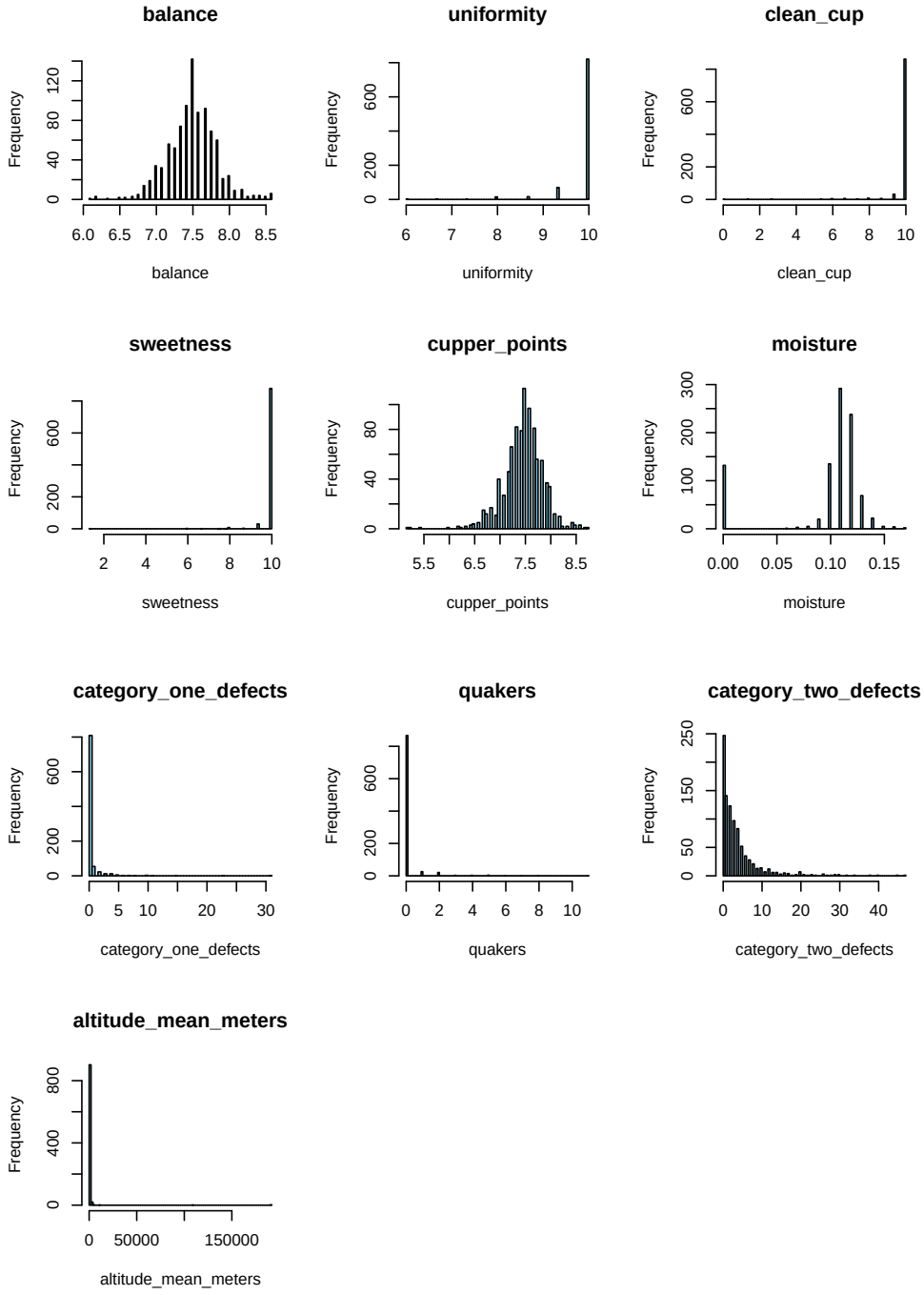
The new summary statistics are as follows:

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Length
total_cup_points	59.8	81.2	82.4	82.1	83.5	90.6	928
species							928
harvest_year							928
processing_method							928
aroma	5.08	7.42	7.58	7.56	7.75	8.75	928
flavor	6.17	7.33	7.50	7.51	7.67	8.83	928
aftertaste	6.17	7.17	7.42	7.38	7.58	8.67	928
acidity	5.25	7.33	7.50	7.52	7.67	8.75	928
body	6.33	7.33	7.50	7.50	7.67	8.50	928
balance	6.08	7.33	7.50	7.49	7.67	8.58	928
uniformity	6.00	10.00	10.00	9.87	10.00	10.00	928
clean_cup	0.00	10.00	10.00	9.85	10.00	10.00	928
sweetness	1.33	10.00	10.00	9.92	10.00	10.00	928

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Length
cupper_points	5.17	7.25	7.50	7.47	7.67	8.75	928
moisture	0.0000	0.1000	0.1100	0.0974	0.1200	0.1700	928
category_one_defects	0.00	0.00	0.00	0.41	0.00	31.00	928
quakers	0.00	0.00	0.00	0.15	0.00	11.00	928
color							928
category_two_defects	0.00	0.00	2.00	3.8	5.00	47.00	928
altitude_mean_meters	1	1100	1311	1868	1600	190164	928

The new summary statistics might suggest skewness in some of the variables. To visualize this we plot the distribution of the individual variables and view the skewness ratio:

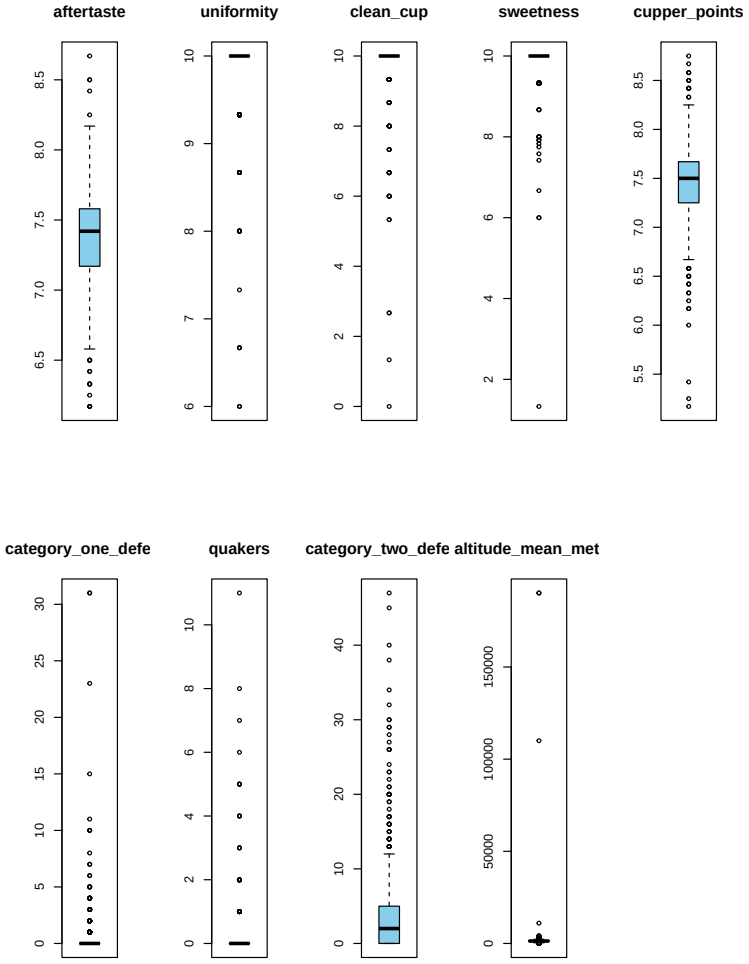




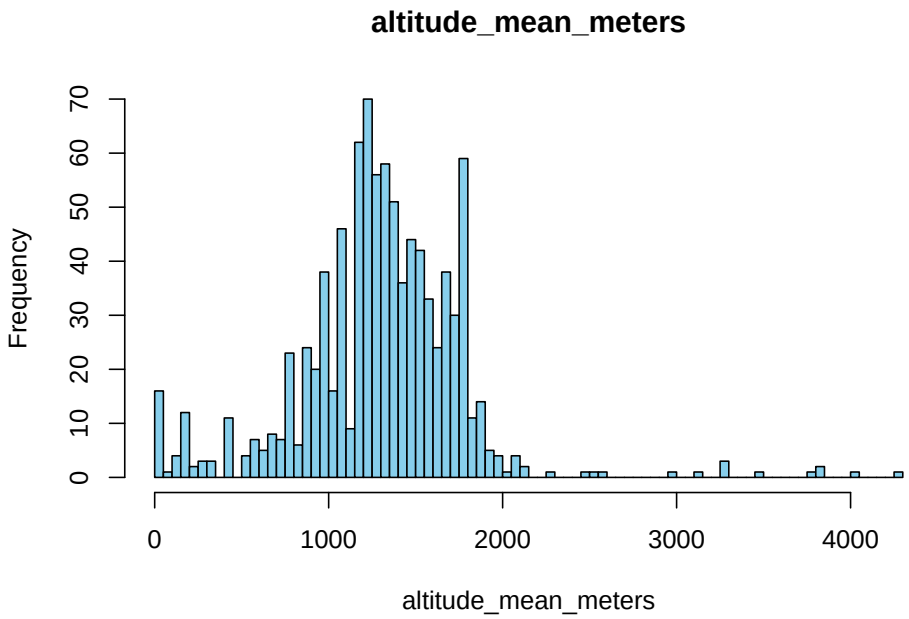
Variable	Skewness_Ratio
total_cup_points	0.86400
aroma	1.06250
flavor	1.00000
aftertaste	0.64000
acidity	1.00000
body	1.00000
balance	1.00000

Variable	Skewness_Ratio
uniformity	NA
clean_cup	NA
sweetness	NA
cupper_points	0.68000
moisture	1.00000
category_one_defects	NA
quakers	NA
category_two_defects	1.50000
altitude_mean_meters	1.373718

We notice that some of the variables have normal distributions, while others seem to appear skewed. This could lead to a poor fit of the linear model if not treated. Another aspect of the linear regression we want to investigate is the presence of outliers in the numerical values.

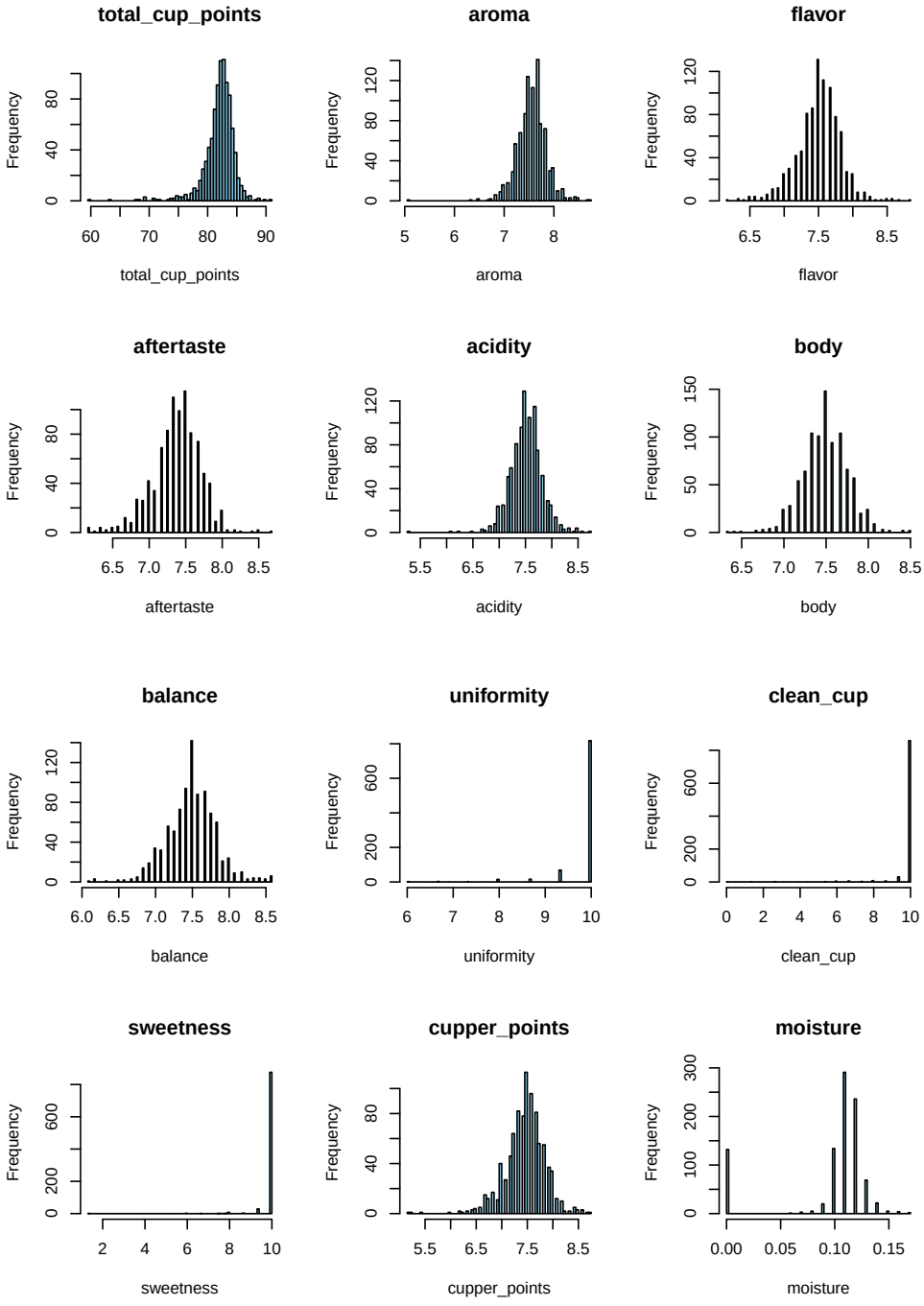


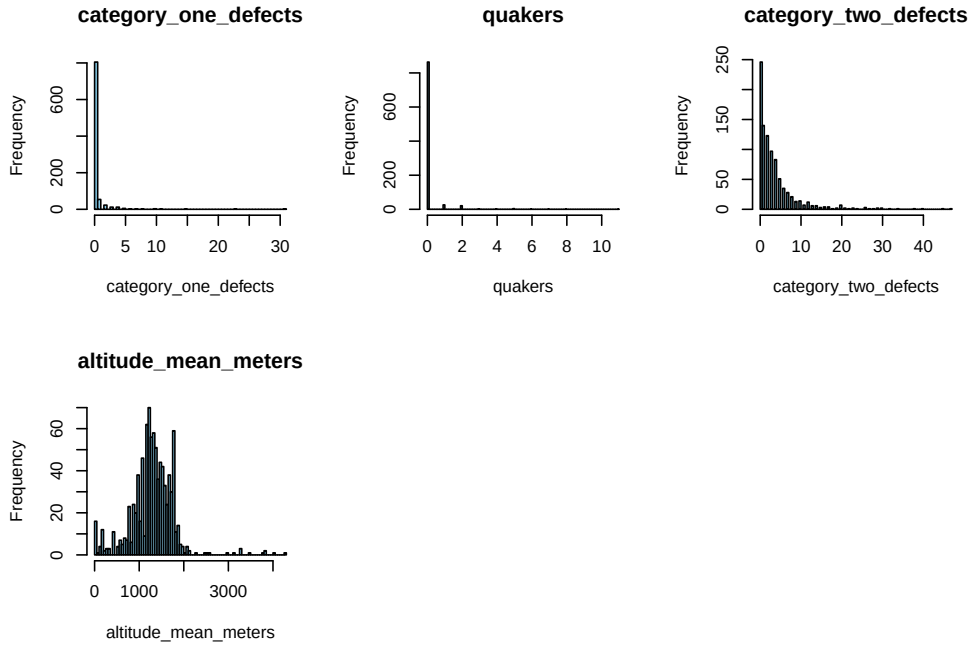
For example, for the variable related to altitude, we notice values surpassing 10,000 meters, which is impossible considering that the highest altitude above sea level is 8,800 meters. Therefore, we understand this to be an error and filter these values out of the dataframe. This results in the new distribution below:



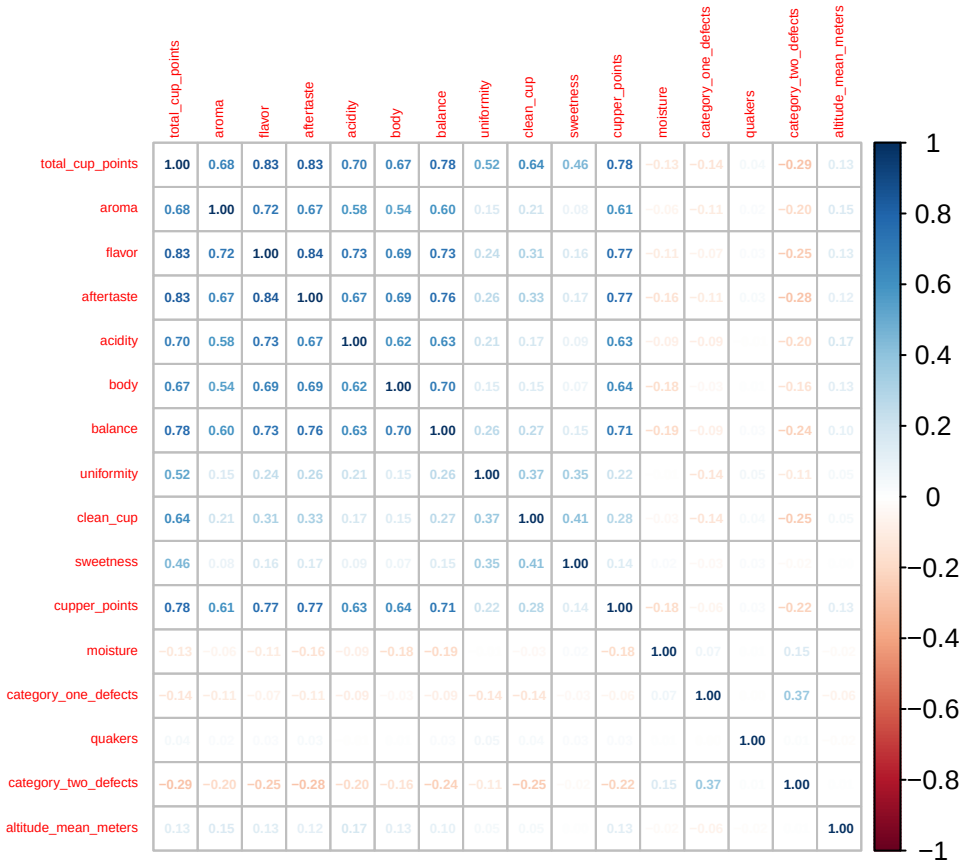
Now we transform the other columns using a Box-Cox transformation. This results in a new skewness-ratio below:

Variable	Skewness_Ratio
total_cup_points	0.8763059
aroma	1.0858750
flavor	1.0229265
aftertaste	0.6579849
acidity	1.0229265
body	1.0136937
balance	1.0136937
uniformity	NA
clean_cup	NA
sweetness	NA
cupper_points	0.6993627
moisture	1.0090498
category_one_defects	NA
quakers	NA
category_two_defects	0.5267934
altitude_mean_meters	1.3481246


total_cup_points
Frequency
0
40
80
60
70
80
90
total_cup_points
aroma
Frequency
0
40
80
120
5
6
7
8
aroma
flavor
Frequency
0
40
80
120
6.5
7.5
8.5
flavor
aftertaste
Frequency
0
40
80
6.5
7.0
7.5
8.0
8.5
aftertaste
acidity
Frequency
0
40
80
120
5.5
6.5
7.5
8.5
acidity
body
Frequency
0
50
100
150
6.5
7.0
7.5
8.0
8.5
body
balance
Frequency
0
40
80
120
6.0
6.5
7.0
7.5
8.0
8.5
balance
uniformity
Frequency
0
200
600
6
7
8
9
10
uniformity
clean_cup
Frequency
0
200
600
0
2
4
6
8
10
clean_cup
sweetness
Frequency
0
200
600
2
4
6
8
10
sweetness
cupper_points
Frequency
0
40
80
5.5
6.5
7.5
8.5
cupper_points
moisture
Frequency
0
100
200
300
0.00
0.05
0.10
0.15
moisture

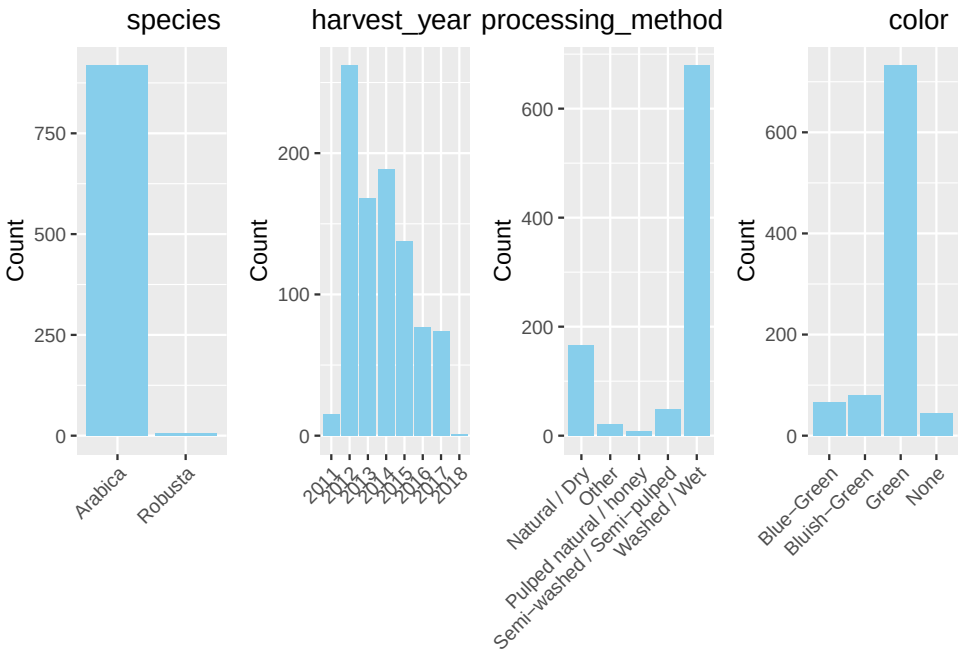


However, we decided to keep the "outlier" values for the remaining variables. For "uniformity", "clean_cup", "sweetness" and "quakers", we determined the outliers to be relevant because of the context behind the data. Some coffee beans were rated low on these categories, which directly influences the outcome of our predictor "total_cup_score". Also, the "category_one_defects" count and "category_two_defects" count are relevant because some coffee beans were found with several defects, which in turn also affects the score. To see this, we will look at the multivariate relationship of the data.



We notice a high correlation between the variable to predict "cupper_points" and the variables "aroma", "flavor,"aftertaste", "acidity", "body", "balance", "clean_cup". Also, the correlation seems high between the same explanatory variables mentioned. This will be something to address when running the linear regression, as collinearity between explanatory variables could introduce bias in their estimates and affect significance.

Now, we will view the categorical variables.



We notice that the sample contains data mostly on the Arabica species, the year 2012, the wet processing method, and the green-colored coffee beans.

Final regression model

Our final model is given by:

$$CupperPoints = \alpha + \beta_1 x_1 + \beta_2 Acidity + \beta_3 Moisture + \beta_4 Aroma + \beta_5 Body + \beta_6 CleanCup + \beta_7 ProcessingMethod_1 +$$

$$\beta_8 ProcessingMethod_2 + \beta_9 ProcessingMethod_3 + \beta_{10} ProcessingMethod_4 + \epsilon$$

, where

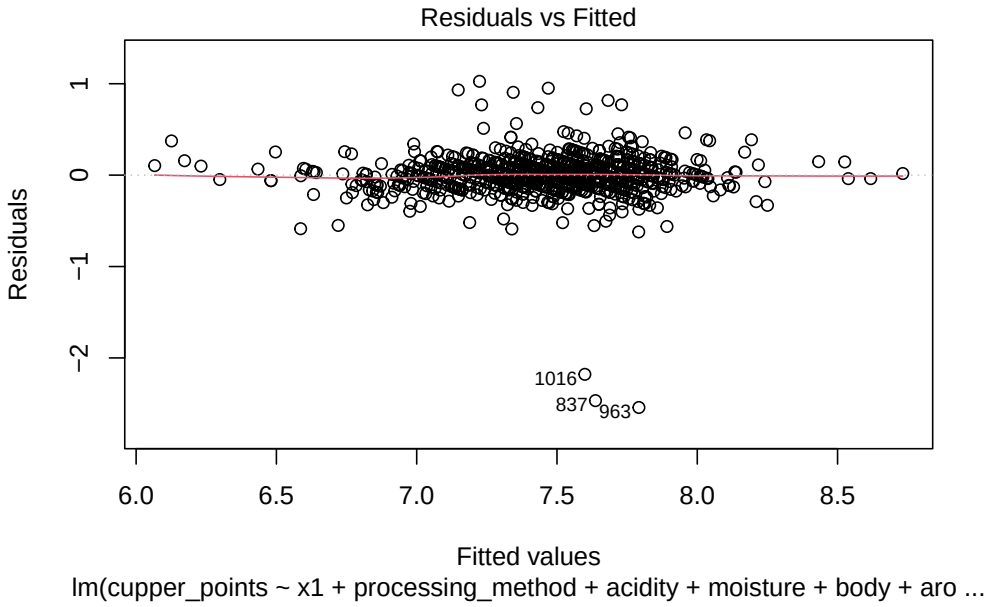
$$x1 = 0.93 \cdot flavor + 0.94 \cdot aftertaste + 0.90 \cdot balance$$

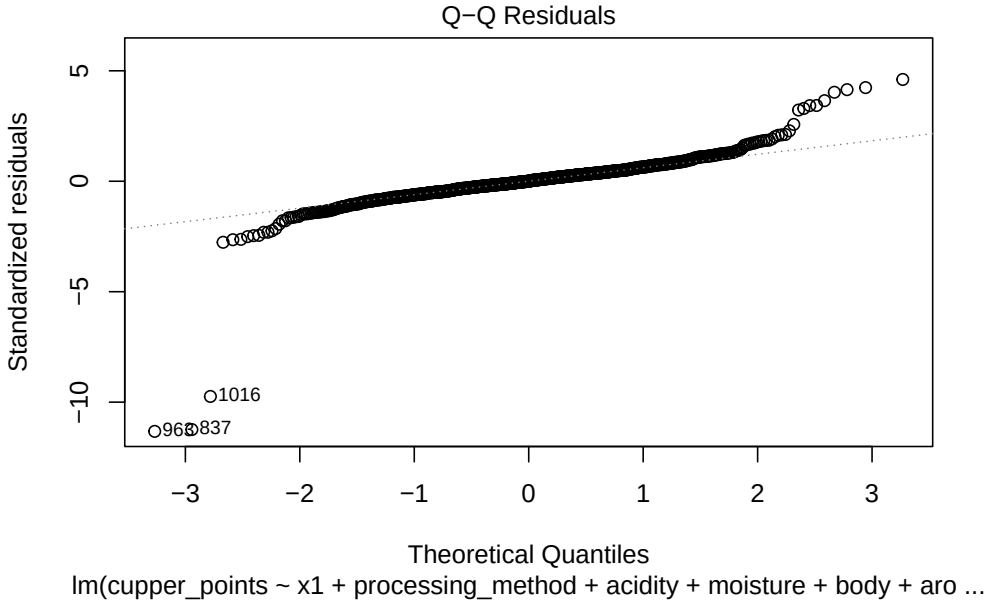
obtained by PCA.

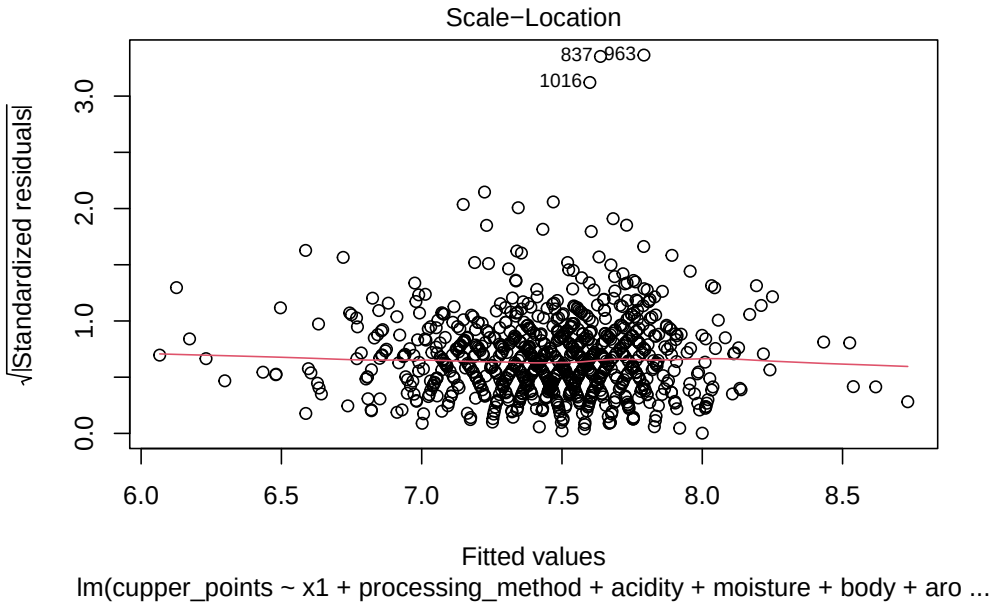
The estimates are given in the following table. We can see that x1 (flavor, balance, aftertaste), the intercept, and one processing method have significant effects on the coffee rating. We find no or only slight significant effects of the other variables in our sample. We also notice that some of the estimated coefficients, especially for the processing method, are very small in magnitude.

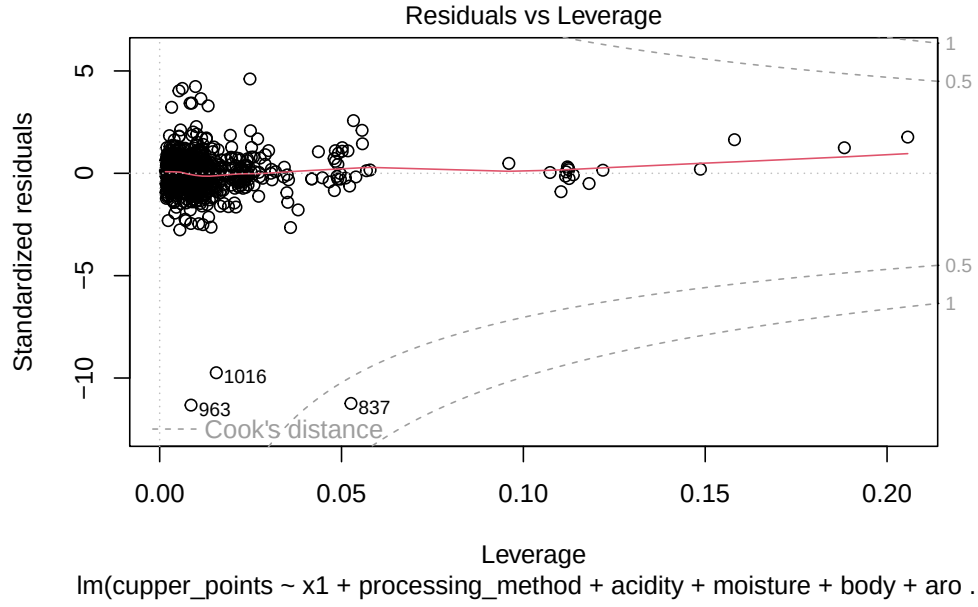
Coefficient	Estimate	Std. Error	t value	Pr(>
(Intercept)	5.723066	0.493290	11.602	< 2e-16 ***
x1	0.164624	0.009899	16.630	< 2e-16 ***
processing_methodOther	-0.174064	0.052559	-3.312	0.000963 ***
processing_methodPulped natural / honey	-0.052158	0.077211	-0.676	0.499515
processing_methodSemi-washed / Semi-pulped	-0.036994	0.036899	-1.003	0.316330
processing_methodWashed / Wet	-0.041842	0.019997	-2.092	0.036674 *
acidity	0.085767	0.037070	2.314	0.020907 *
moisture	-0.432694	0.184969	-2.339	0.019536 *
body	0.076217	0.042616	1.788	0.074036 .
aroma	0.058440	0.036204	1.614	0.106837
clean_cup	0.016566	0.010533	1.573	0.116117

Statistic	Value
Residual standard error	0.2255
Degrees of freedom	913
Multiple R-squared	0.6748
Adjusted R-squared	0.6713
F-statistic	189.5
F-statistic DF	10 and 913
F-statistic p-value	< 2.2e-16









In our original dataset, we noticed multicollinearity between the explanatory variables flavor, balance, and aftertaste (high VIF values). To tackle this issue, we applied principal component analysis and retained the first principal component, as it already explained over 85% of the variance of those three predictors. We used PCA on this subset of variables (as opposed to e.g. ridge regression) because the other explanatory variables showed no or only moderate collinearity. In our original linear regressions (see additional work) we noticed many insignificant predictors that were not able to explain much of the variation of our response variable. Since we wanted to research which factors were most important in determining coffee ratings (and not which factors are *not* important), we tried to find a smaller model that could answer this question better and be more interpretable. The model returned by the AIC criterion was able to select a subset of the original

predictors without losing explained variation ($R^2_{adj,full} = 0.6697$, $R^2_{adj,AIC} = 0.6713$).

In the diagnostics plots of our final model, we noticed no signs of heteroscedasticity or non-linearity. In the residual distribution, a slight tail was visible compared to a normal distributions. Cook's distance plot revealed three influential points. Upon investigation, we could find no evidence in the data that those points were errors or that they had underlying characteristics which would qualify them for elimination, which is why they were not removed from the model. It should be noted though that removing those outliers led to a much better estimation with more significant predictors and a better fit of $R^2_{adj,NoOutliers} = 0.7752$ (see additional work for details).

Discussion: limitations

One limitation of the model is human bias. For example, variables- such as flavor and bitterness- are continuous, yet some people will personally never select a non-integer rating. Human bias also affects the collinearity between aroma, flavor, aftertaste, acidity, body, balance, etc. Although these all refer to different aspects, many people blend the terminologies.

Another limitation is that our response variable, "cupper_points," is on a 1-10 scale. This means that our model could predict values above or below this specified range.

Conclusion

The final model depicts the main aspects that influence coffee ratings. The intercept suggests a baseline score of approximately 5.723 for coffee, and the variable x1 (flavor, balance, aftertaste), positively influences the score. On the other hand, the 'Other' processing method negatively affects the score, while 'Washed/Wet' also decreases it slightly but significantly. Higher acidity is beneficial for the score, whereas higher moisture content is not. The model explains about 67% of the variance in coffee scores, indicating a good fit. Variables such as body, aroma, and clean cup show positive associations with coffee scores but lack statistical significance at $\alpha = 0.05$.

An interesting observation from the analysis is the lack of influence from metadata attributes of the coffee beans, such as harvest year and color, on the ratings assigned by coffee experts. This suggests taste characteristics may hold a greater significance in the criteria employed by professionals. Alternatively, it is likely that the dataset used was not large enough to capture the nuances that these attributes may impart on expert evaluations. Further research with a larger and more varied sample could be useful to fully understand the impact of such metadata on coffee quality ratings.

Additional work

In addition to the final regression method above, we also tried basic linear regression, different model selection criteria, and ridge regression methods. In the initial linear regression model, we wanted to see the general relationship of all of the variables, with no collinearity or model selection considered. Running the full linear regression resulted in a residual plot that had a slight trend and a QQ- plot that still displayed some skewness.

The ridge regression model was tried as an attempt to decrease the multicollinearity issue we saw in the exploratory data analysis. When run, it also produced a "better" model compared to the full model. However, it had issues with the categorical variables and produced results that were more difficult to interpret than a simple regression.Therefore, we first tried model selection.

The model obtained by using BIC as the selection criterion was reduced to only one predictor while having a similar R^2. Since BIC tends too constrain the model heavily (worsens future prediction), we did not use this model as our final model.

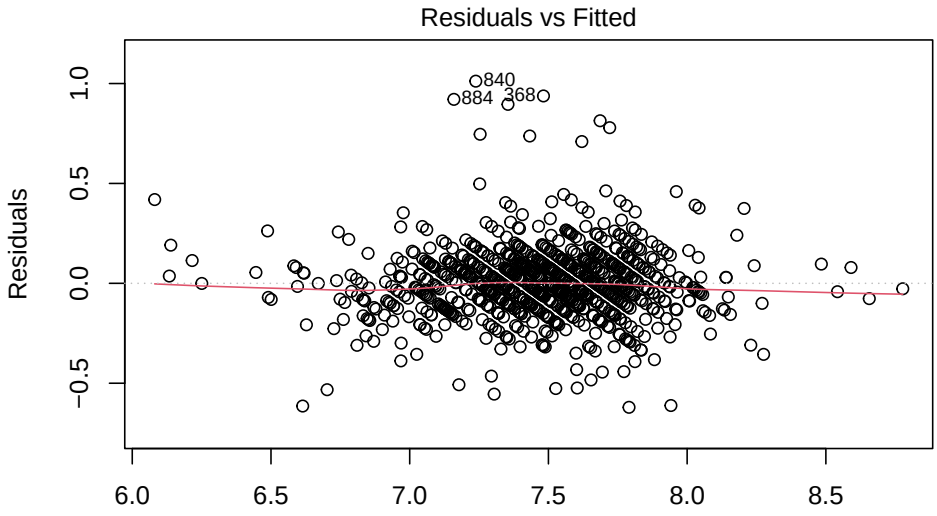
Coefficient	Estimate	Std. Error	t value	Pr(>
(Intercept)	7.465422	0.007519	992.85	<2e-16 ***
x1	0.200058	0.004702	42.55	<2e-16 ***

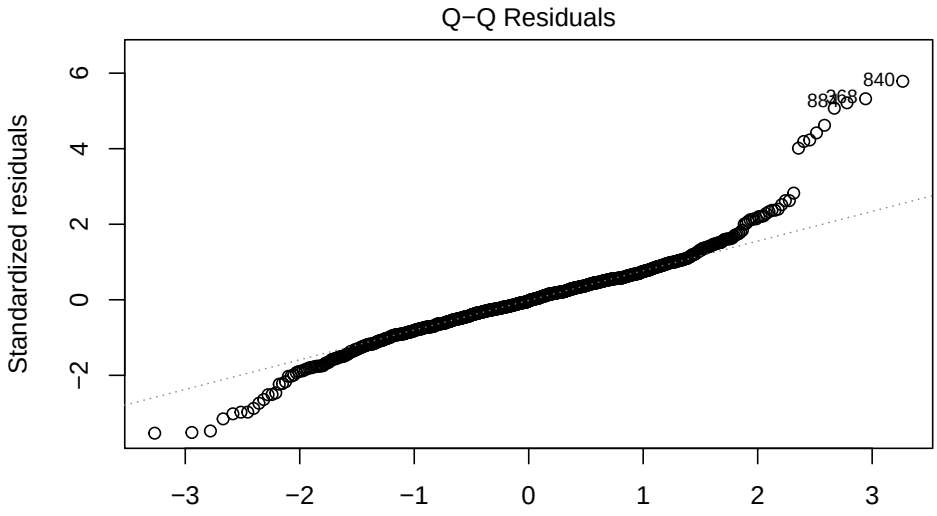
Statistic	Value
Residual standard error	0.2286
Degrees of freedom	922
Multiple R-squared	0.6626
Adjusted R-squared	0.6622
F-statistic	1811
F-statistic DF	1 and 922
F-statistic p-value	< 2.2e-16

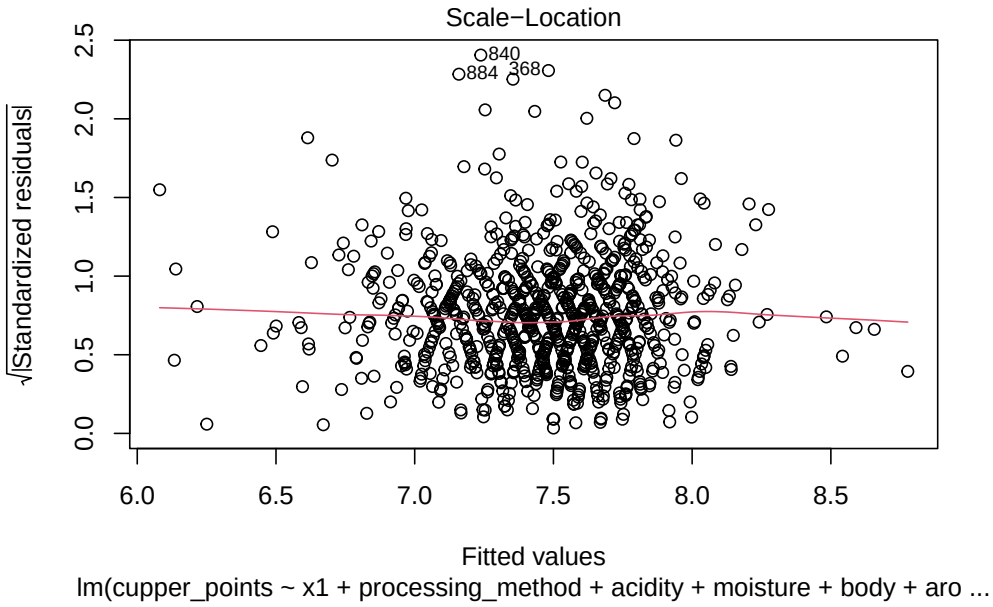
Finally, we ran our final model without the three influential points diagnosed with Cook's distance. This significantly improved the model.

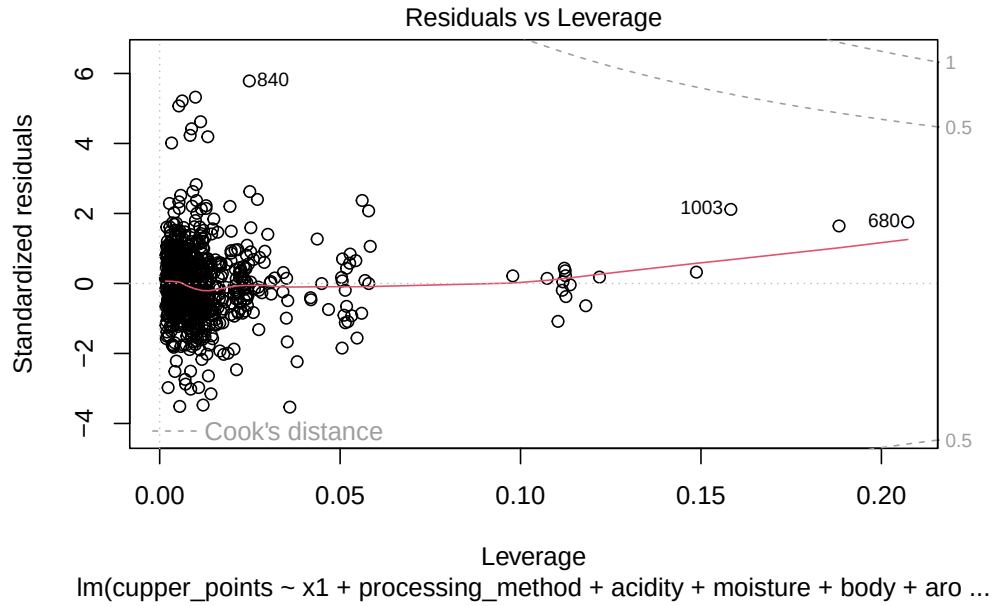
Coefficient	Estimate	Std. Error	t value	Pr(>
(Intercept)	5.301936	0.388104	13.661	< 2e-16 ***
x1	0.162241	0.007780	20.853	< 2e-16 ***
processing_methodOther	-0.057245	0.042251	-1.355	0.17579
processing_methodPulped natural / honey	-0.070329	0.060657	-1.159	0.24657
processing_methodSemi-washed / Semi-pulped	-0.052184	0.029006	-1.799	0.07234 .
processing_methodWashed / Wet	-0.050760	0.015753	-3.222	0.00132 **
acidity	0.087174	0.029135	2.992	0.00285 **
moisture	-0.282303	0.145432	-1.941	0.05255 .
body	0.149905	0.033653	4.454	9.45e-06 ***
aroma	0.037153	0.028531	1.302	0.19317
clean_cup	0.018332	0.008274	2.216	0.02697 *

Statistic	Value
Residual standard error	0.1771
Degrees of freedom	910
Multiple R-squared	0.7776
Adjusted R-squared	0.7752
F-statistic	318.2
F-statistic DF	10 and 910
F-statistic p-value	< 2.2e-16


Residuals vs Fitted
Residuals
Fitted values
lm(cupper_points ~ x1 + processing_method + acidity + moisture + body + aro ...
840
884
368
1.0
0.5
0.0
−0.5
6.0
6.5
7.0
7.5
8.0
8.5


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(cupper_points ~ x1 + processing_method + acidity + moisture + body + aro ...
840
6
4
2
0
−2
−3
−2
−1
0
1
2
3





References

Angela. "World Coffee Consumption Statistics." Coffee Rank, 30 Oct. 2022, coffee-rank.com/world-coffee-consumption-statistics/.

Mock, Tidy Tuesday- Coffee ratings, (2020), GitHub repository, https://github.com/rfordatascience/tidytuesday/tree/21674ea17c59e1bdcc4be7017a583d81ba41ae76/data/2020/2020-07-07

Link to Dataset: https://github.com/rfordatascience/tidytuesday/blob/21674ea17c59e1bdcc4be7017a583d81ba41ae76/data/2020/2020-07-07/coffee_ratings.csv