COFFEE QUALITY MEASURES PREDICTION

By :

1. Crysantha Monica Lim | 2602090076
2. Cherylene Callista Reksohartono | 2602087024
3. Gladys Lionardi | 2602073076

Problem Statement

Analyzing the "arabica_data_cleaned.csv" dataset to identify what affects the lowest quality measures of Arabica coffee in a specific country.

Data Source

The dataset downloaded from Kaggle contains 44 attributes and 1311 rows of data.

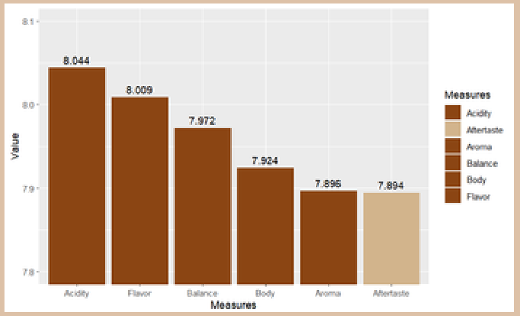
Analysis

I-Assessing the data
II-Cleaning the data
III-Training and building the model
IV-Evaluating the model's performance

Data & Method

1. Removing character variables
2. Imputation
3. Selecting atrributes that're hypotherized to have significant impact to the analysis

Clustering Result

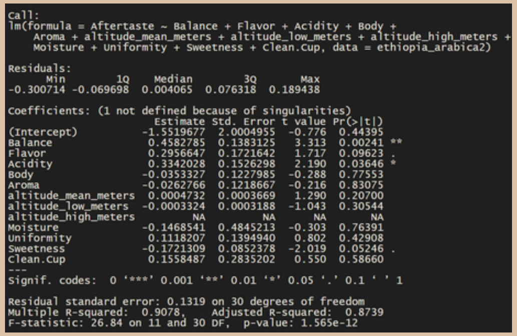


A graph that shows both speciality and weakness of Ethiopia's Arabica coffee.

It shows that Ethiopia 's Arabica Coffee is high in "acidity" while it is low in "aftertaste".

What affects the aftertaste of a coffee? How to maximize the quality of it?

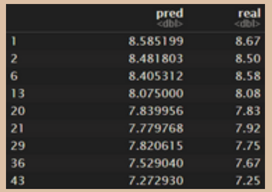
Model Report



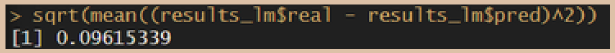
```
Call:
lm(formula = Aftertaste ~ Balance + Flavor + Acidity + Body +
    Aroma + altitude_mean_meters + altitude_low_meters + altitude_high_meters +
    Moisture + Uniformity + Sweetness + Clean.Cup, data = ethiopia_arabica2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.300714 -0.069698  0.004065  0.076318  0.189438

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.5519677  2.0004955  -0.776  0.44395
Balance              0.4582785  0.1383125   3.313  0.00241 **
Flavor               0.2956647  0.1721642   1.717  0.09623 .
Acidity              0.3342028  0.1526298   2.190  0.03646 *
Body                -0.0353327  0.1227985  -0.288  0.77553
Aroma               -0.0262766  0.1218667  -0.216  0.83075
altitude_mean_meters  0.0004732  0.0003669   1.290  0.20700
altitude_low_meters  -0.0003324  0.0003188  -1.043  0.30544
altitude_high_meters        NA         NA      NA       NA
Moisture            -0.1468541  0.4845213  -0.303  0.76391
Uniformity           0.1118207  0.1394940   0.802  0.42908
Sweetness           -0.1721309  0.0852378  -2.019  0.05246 .
Clean.Cup            0.1558487  0.2835202   0.550  0.58660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1319 on 30 degrees of freedom
Multiple R-squared:  0.9078,    Adjusted R-squared:  0.8739
F-statistic: 26.84 on 11 and 30 DF,  p-value: 1.565e-12
```

Prediction Model (Multi Linear Regression)

- Below shows the differences between the Aftertaste values that we predicted and the actual Aftertaste values from the dataset. The predicted values are collected based on the calculation of the other quality measures (Acidity, Balance, etc)
- Between the predicted and the dataset values, it has an error of 0.096, which is actually quite low. Therefore, we can conclude that our predictive model is quite accurate.

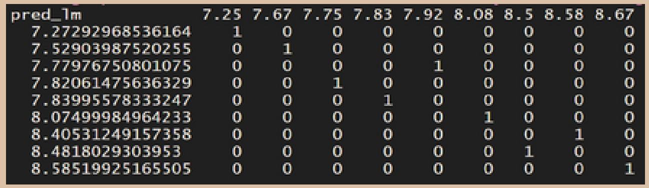


	pred	real
	<dbl>	<dbl>
1	8.585199	8.67
2	8.481803	8.50
6	8.405312	8.58
13	8.075000	8.08
20	7.839956	7.83
21	7.779768	7.92
29	7.820615	7.75
36	7.529040	7.67
43	7.272930	7.25



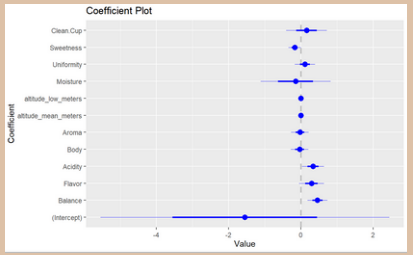
```
> sqrt(mean((results_lm$real - results_lm$pred)^2))
[1] 0.09615339
```

- The data explains how accurate our predicted values are. In the second visualization below, the dots represent the predicted values while the line represents the actual values. It can be seen that the dots tend to be close to the line, so we can say that our model is suitable to predict new values.

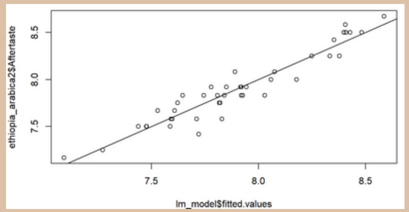


```
pred_lm            7.25 7.67 7.75 7.83 7.92 8.08 8.5 8.58 8.67
  7.27292968536164    1    0    0    0    0    0   0    0    0
  7.52903987520255    0    1    0    0    0    0   0    0    0
  7.77976750801075    0    0    0    0    1    0   0    0    0
  7.82061475636329    0    0    1    0    0    0   0    0    0
  7.83995578333247    0    0    0    1    0    0   0    0    0
  8.07499984964233    0    0    0    0    0    1   0    0    0
  8.40531249157358    0    0    0    0    0    0   0    1    0
  8.4818029303953     0    0    0    0    0    0   1    0    0
  8.58519925165505    0    0    0    0    0    0   0    0    1
```

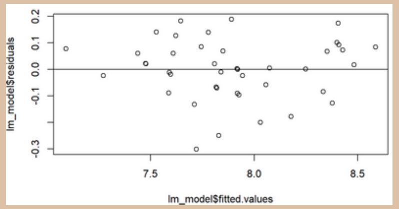
Visualizations



Plotting the coefficient



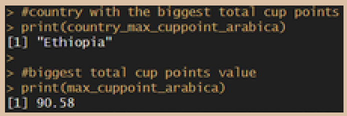
Plotting the Actual Vs. Predicted Values



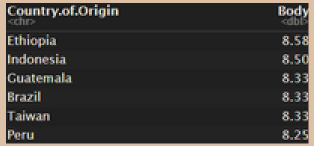
Plotting the residuals

Feature Selection

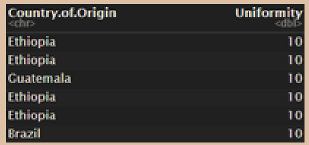
STRATIFIED RANDOM SAMPLING : (based on the highest total cups point and all the other quality measures factors)



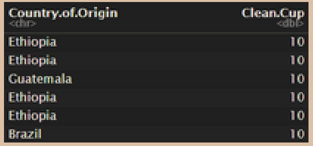
```
> #country with the biggest total cup points
> print(country_max_cuppoint_arabica)
[1] "Ethiopia"
>
> #biggest total cup points value
> print(max_cuppoint_arabica)
[1] 90.58
```



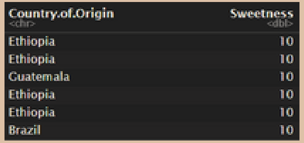
Country.of.Origin	Body
<chr>	<dbl>
Ethiopia	8.58
Indonesia	8.50
Guatemala	8.33
Brazil	8.33
Taiwan	8.33
Peru	8.25



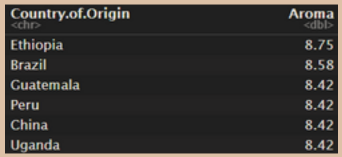
Country.of.Origin	Uniformity
<chr>	<dbl>
Ethiopia	10
Ethiopia	10
Guatemala	10
Ethiopia	10
Ethiopia	10
Brazil	10



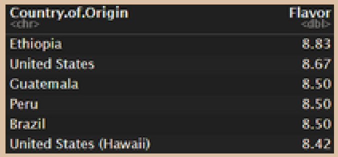
Country.of.Origin	Clean.Cup
<chr>	<dbl>
Ethiopia	10
Ethiopia	10
Guatemala	10
Ethiopia	10
Ethiopia	10
Brazil	10



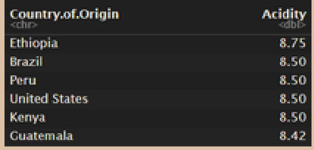
Country.of.Origin	Sweetness
<chr>	<dbl>
Ethiopia	10
Ethiopia	10
Guatemala	10
Ethiopia	10
Ethiopia	10
Brazil	10



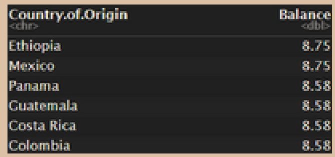
Country.of.Origin	Aroma
<chr>	<dbl>
Ethiopia	8.75
Brazil	8.58
Guatemala	8.42
Peru	8.42
China	8.42
Uganda	8.42



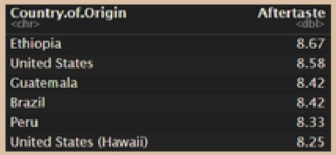
Country.of.Origin	Flavor
<chr>	<dbl>
Ethiopia	8.83
United States	8.67
Guatemala	8.50
Peru	8.50
Brazil	8.50
United States (Hawaii)	8.42



Country.of.Origin	Acidity
<chr>	<dbl>
Ethiopia	8.75
Brazil	8.50
Peru	8.50
United States	8.50
Kenya	8.50
Guatemala	8.42



Country.of.Origin	Balance
<chr>	<dbl>
Ethiopia	8.75
Mexico	8.75
Panama	8.58
Guatemala	8.58
Costa Rica	8.58
Colombia	8.58



Country.of.Origin	Aftertaste
<chr>	<dbl>
Ethiopia	8.67
United States	8.58
Guatemala	8.42
Brazil	8.42
Peru	8.33
United States (Hawaii)	8.25

Conclusions

- Balance, flavor, and acidity have a significant positive effect on aftertaste.
- Sweetness has a significant negative effect on aftertaste.
- Body, aroma, altitude_mean_meters, altitude_low_meters, altitude_high_meters, moisture, uniformity, and clean cup do not have a significant effect on aftertaste.

Reference

https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi?select=arabica_data_cleaned.csv