Coffee Quality Analysis

Pablo Aganza
Moez Hudda
Mohammed Elzubeir
Sonia Song
GitHub: https://github.com/UC-Berkeley-I-School/Project2_Team_7

History of Coffee

Coffee has long been part of civilization. The exact origin is unknown, but one of the earliest discoveries of coffee beans can be traced back to the 11th century in Ethiopia. Coffee became popularized in Europe in the 17th Century and began replacing common breakfast drinks such as beer and wine. From there, coffee was brought to New York and started changing household beverage preferences in the United States.

According to NOAA Climate.gov, coffee lovers consume more than 2.25 billion cups of coffee a day. It is among the most valuable tropical exports on the planet. Cool to warm tropical climates are ideal for growing coffee. Soil richness will also affect coffee quality. For this reason, the top coffee-growing countries tend to concentrate in countries along the Equator.

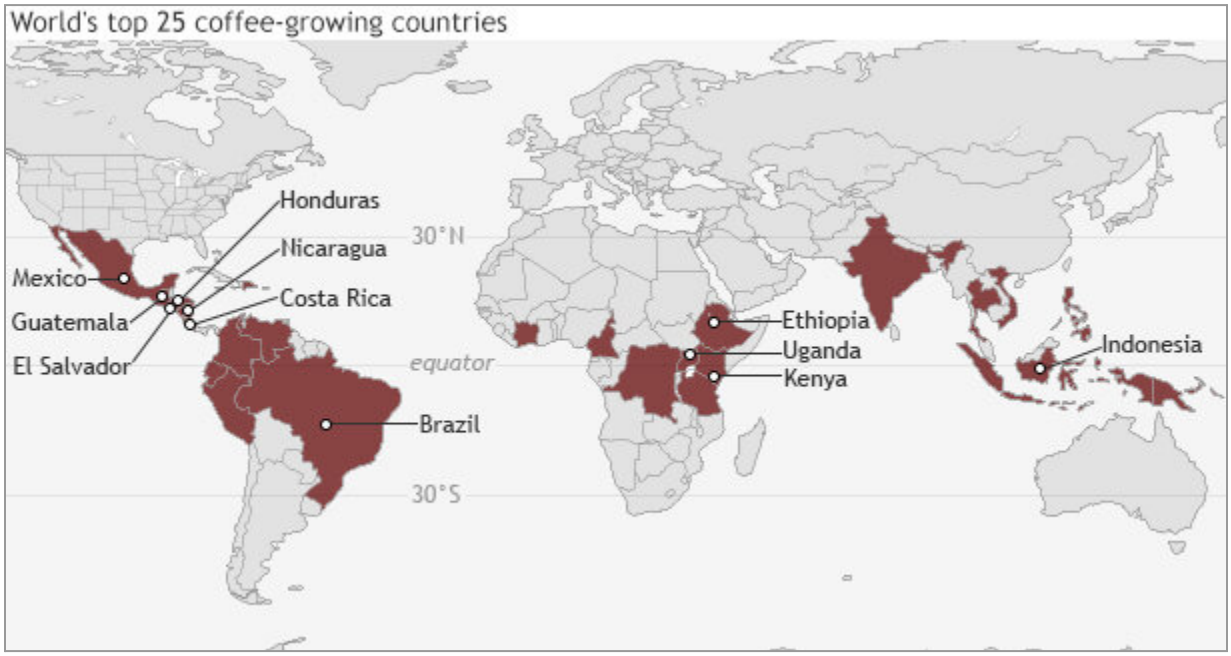


Image source: NOAA Climate.gov

The price of coffee has shifted throughout history. Variations in the supply chain, such as extreme weather events (e.g. severe drought and frost), shipping bottlenecks, and labor shortages have all contributed to sharp increase in coffee price historically. Today, the average daily coffee price is around $2 / pound.

Main Question
Which attributes of coffee influence the ranking of Coffee?

Sub Questions:
Which Countries produce the highest ranked coffee?
Which Countries produce the lowest ranked coffee?
Are there patterns we see in characteristics about coffee which reinforce why a country may be ranked higher or lower?

Data
We used a main data set and several supplemental data sets to come to our conclusions. As we dove further into the initial main dataset we selected within our proposal, we found that the inferences pulled weren't sufficient but it would make for a great supplemental data set due to having data on: Robusta Production, Arabica Production and Consumption by Country. After exploring other supplemental datasets, we found that the main data set had the attributes necessary to answer our main and sub questions. Due to curiosity, we also built our own dataset from scraping weatherandclimate.com to pull temperature averages for the highest ranked coffees. We didn't use this in our final analysis. One note to consider is that the values for the country United States are exclusively in Hawaii.

From our main data set, we used the following variables:

- Acidity: Acidity in coffee refers to the brightness or liveliness of the taste.
- Altitude: Height of the coffee farm within the country
- Balance: Balance refers to how well the different flavor components of the coffee work together.
- Moisture: The amount of water within the coffee beans
- Sweetness: It can be described as caramel-like, fruity, or floral, and is a desirable quality in coffee.
- Uniformity: Uniformity refers to the consistency of the coffee from cup to cup.
- Ranking: Score given on the coffee from the data set.

Main Data Set:
Coffee Quality Data (CQI) https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi

Supplemental Data Sets
Coffee Distribution Across 94 Countries
https://www.kaggle.com/datasets/parasrupani/coffee-distribution-across-94-counties
https://weatherandclimate.com/

Wrangling of data set process:
In regards to wrangling our data set, we dropped columns we didn't see value in (there were over 40 columns in some data sets!) and used only columns which would help us in answering our question. We also built additional tables that had the means of each particular characteristic listed above.

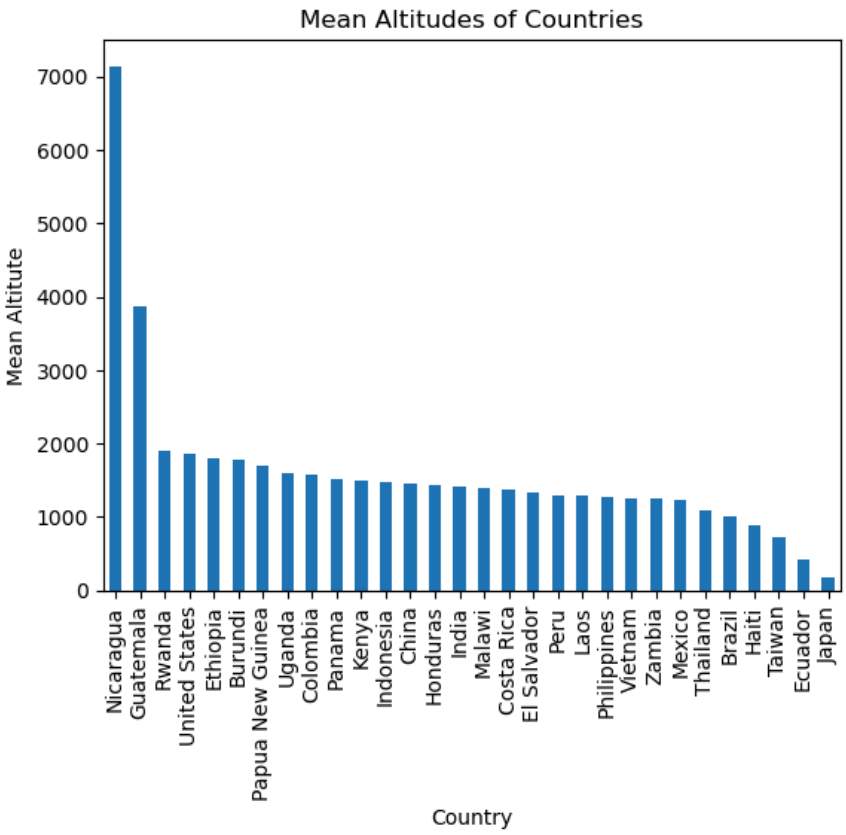
Initially, we merged our main data set with the coffee distribution supplemental data set and dropped tables we didn't see value in. Next, we renamed columns to make them more relevant and uniform across the merged table. From there we began to have fun with them. We first wanted to see which countries had the best coffee, so we grouped by coffee, took the mean of the ranking value and sorted them. Next, we took the means for each characteristic that was relevant to telling a story about coffee. Acidity, Body, Balance, Uniformity, Moisture and Sweetness. We even created a mean table and merged all the mean data together to get a broader view.

Following these different tables, we were ready to build our charts to visually describe what the data was telling us.

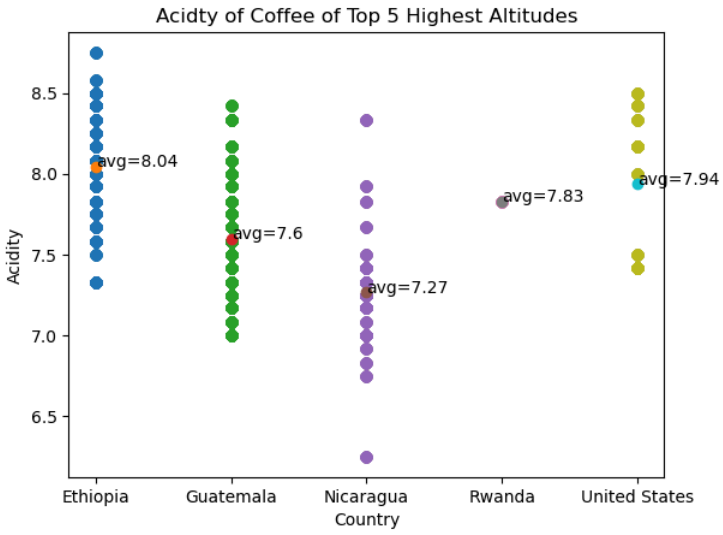
Hypotheses: Given origins in highlands and marketing (maybe) altitude has a positive relation to coffee quality
Null:

- Altitude has no effect on coffee quality
- Variables defined in data section have no effect on coffee quality

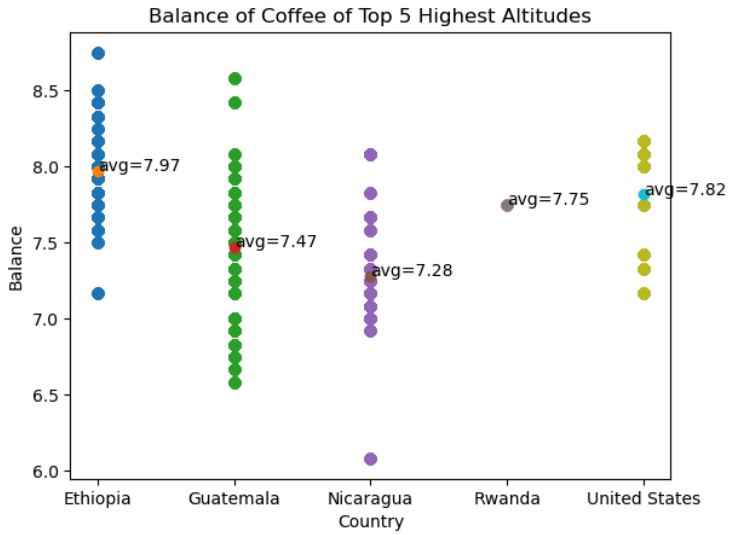
Graphs and discussion
From our hypothesis, we looked at the countries with the highest altitudes:



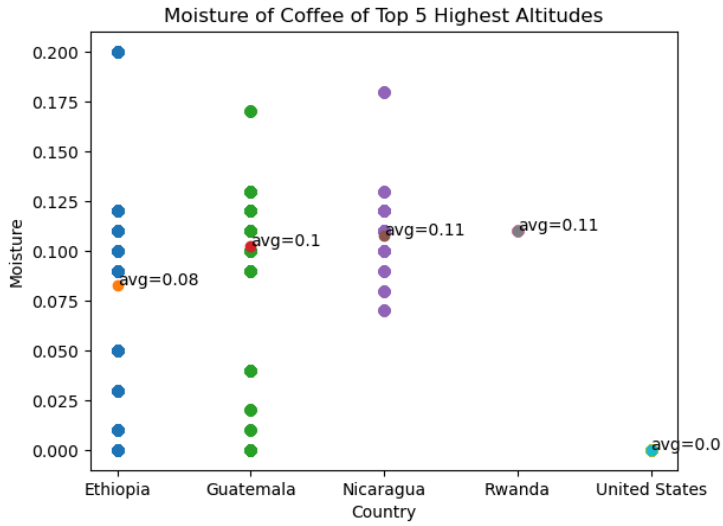
From this, we see the countries with the highest mean altitudes are Nicaragua, Guatemala, Rwanda, United States (Hawaii), and Ethiopia. The notable countries are Nicaragua and Guatemala as they are far above the average mean altitude of all countries. Next, we can look at the other six variables defined in the data section (Acidity, Balance, Moisture, Sweetness, Uniformity, and Ranking) for these five high altitude countries.



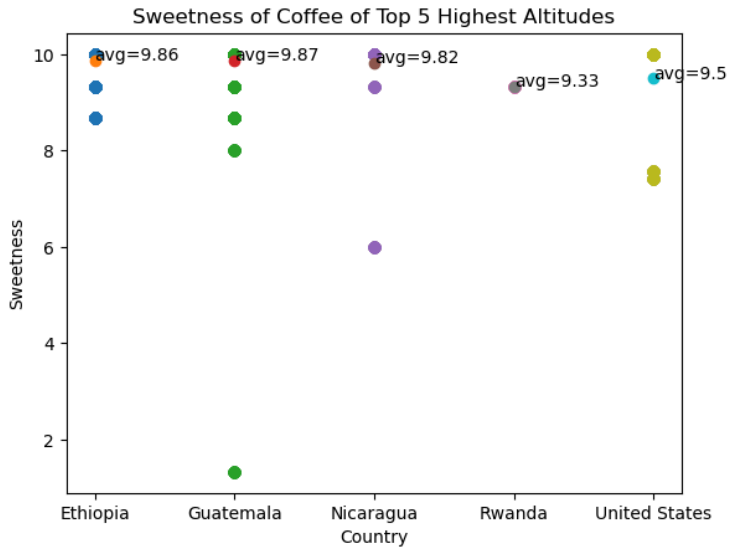
- Nicaragua and Guatemala have lower averages



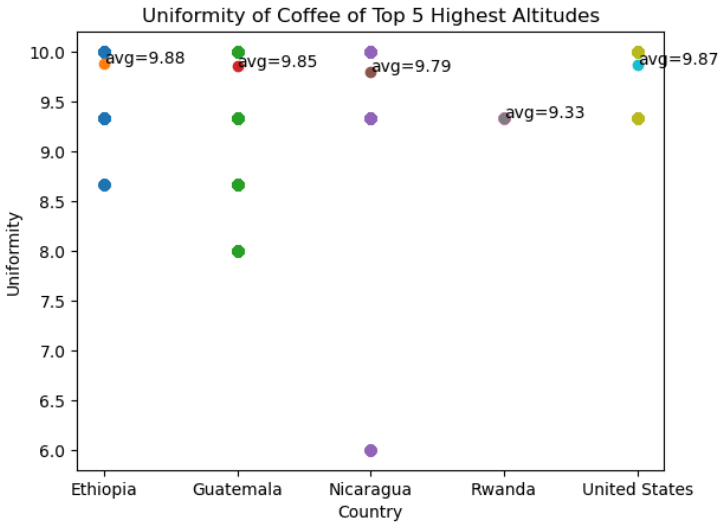
- Nicaragua and Guatemala have lower averages



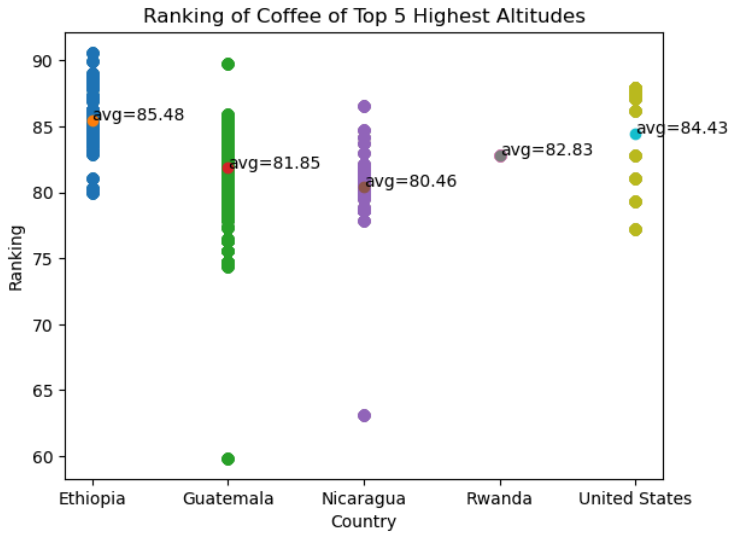
- Averages are about the same amongst all countries
- United States (Hawaii) has zero for Moisture, most likely indicating that this metric was not collected for these coffee beans



- Averages are about the same amongst all countries, with Rwanda being slightly lower



- Averages are all fairly close as well, excluding Rwanda which has a slighter lower average than the rest

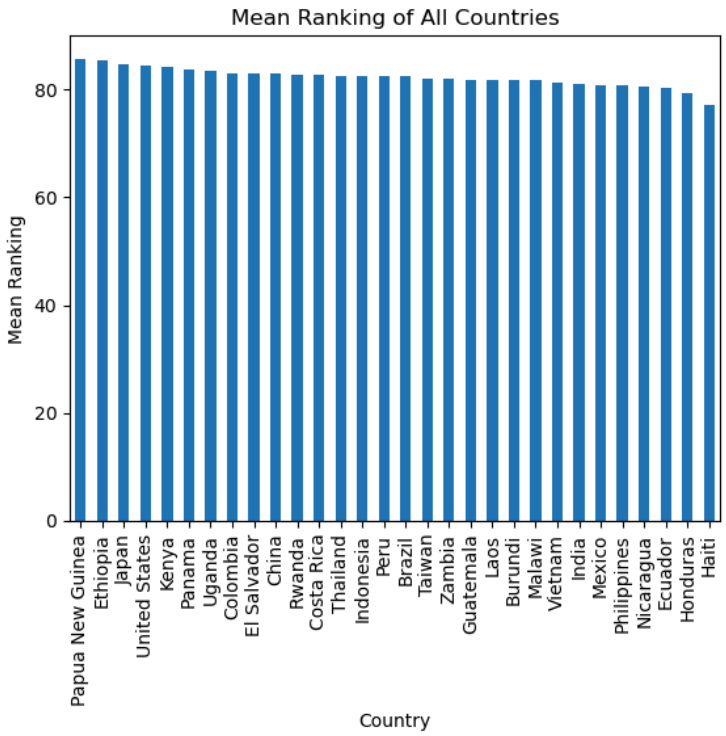


- Guatemala and Nicaragua have the lowest overall average ranking

Based on the values of these variables, we see the rankings of these countries resulted in averages of:

- Ethiopia - 85.48
- United States (Hawaii) - 84.43
- Rwanda - 82.83
- Guatemala - 81.85
- Nicaragua - 80.46

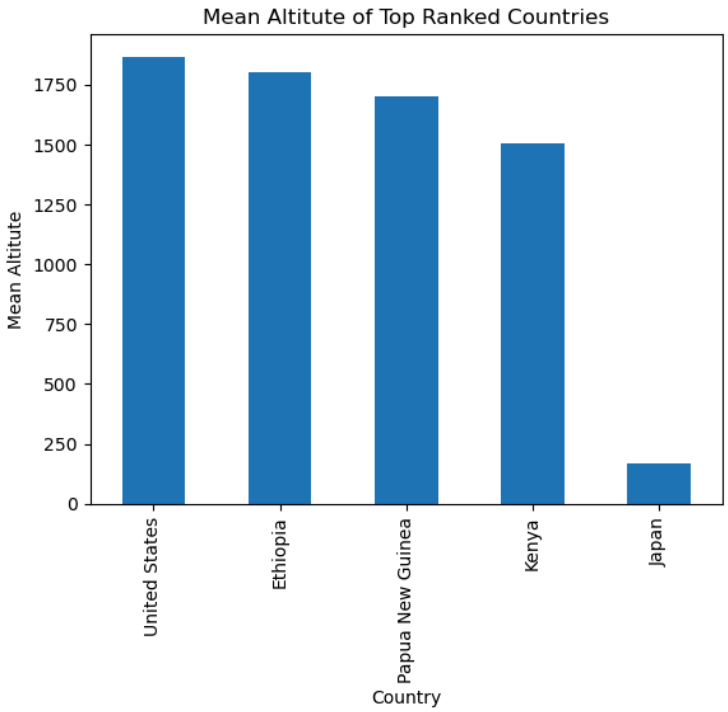
Based on this, we see that Nicaragua, being the country with the highest altitude by far of over 7000 meters, has the lowest ranking of these five countries. Now, we want to look at the average ranking of all countries.



From here, we see that the top five ranked countries are Papua New Guinea, Ethiopia, Japan, United States (Hawaii), and Kenya. Comparing this with the highest altitude countries, we see that rankings out of 30 positions are:

- Nicaragua - 27th
- Guatemala - 19th
- Rwanda - 11th
- United States (Hawaii) - 4th
- Ethiopia - 2nd

Looking at the altitude of these top ranked countries in the chart below, we see that they all have mean altitudes around the 1500-1750 meter range (excluding Japan), while the highest mean altitude was over 7000 meters in Nicaragua.



From this, we see our hypothesis of altitude having a direct positive correlation to the coffee quality is in question. We can infer that while altitude may not have a direct positive correlation, it may be that an average altitude in the range of 1500-1750 meters is most ideal for growing quality coffee, and having too high of an altitude can result in lesser quality coffee, as both Nicaragua and Guatemala are in the bottom half of the rankings and have altitudes of over 4000 meters. Looking back at the graphs for each variable of the highest altitude countries, we see that both Nicaragua and Guatemala have on average lower acidity and balance, but have similar sweetness, uniformity, and moisture averages compared to the other countries. These low averages in acidity and balance is most likely what caused their overall rankings to be low. From this, we can surmise that growing coffee at too high of an altitude results in unfavorable acidity and balance in coffee. Based on this, we can conclude that the variables acidity and balance may have the most influence on the coffee quality score.

Analysis

To understand which factors most influence coffee quality we opted to carry out an Ordinary Least Squares (OLS) regression, regressing the Coffee Quality Indicator represented by Total Cup Points on Sweetness, Moisture, Balance, Acidity, Altitude, Body, and Uniformity.

The regression is specified Below:

$$CQI = \beta_0 + \beta_1 \cdot Sweetness + \beta_2 \cdot Moisture + \beta_3 \cdot Balance + \beta_4 \cdot Acidity + \beta_5 \cdot Altitude + \beta_6 \cdot Body + \beta_7 \cdot Uniformity + \epsilon$$

Where:

CQI denotes the coffee quality indicator (Total Cup Points)
β_0 is the regression intercept
$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$
ϵ is the error term of the model, encapsulating the influence of variables not specified in the model.

Interpretation

Each coefficient β_i reflects the expected change in the Coffee Quality Indicator (CQI) in response to a one-unit increase in the corresponding variable, keeping all other variables constant. The intercept β_0 indicates the value of the CQI when all predictors are zero, assuming such a scenario falls within the scope of the observed data.

Regression Summary

Dep. Variable:	Total Cup Points	**R-squared:**	0.947
Model:	OLS	**Adj. R-squared:**	0.930
Method:	Least Squares	**F-statistic:**	55.63
Date:	Sat, 13 Apr 2024	**Prob (F-statistic):**	1.57e-12
No. Observations:	30	**Log-Likelihood:**	-15.533
Df Residuals:	22	**AIC:**	47.07
Df Model:	7	**BIC:**	58.28
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	-1.2929	7.456	-0.173	0.864	-16.756	14.170
Sweetness	0.9060	0.198	4.581	0.000	0.496	1.316
Moisture	-0.6192	3.027	-0.205	0.840	-6.896	5.657
Balance	3.5673	1.156	3.087	0.005	1.171	5.964
Acidity	2.9133	0.818	3.563	0.002	1.218	4.609
Altitude (mean meters)	-3.648e-05	7.55e-05	-0.483	0.634	-0.000	0.000
Body	0.2879	0.582	0.494	0.626	-0.920	1.495
Uniformity	2.4100	0.572	4.215	0.000	1.224	3.596

Omnibus:	7.791	**Durbin-Watson:**	2.444
Prob(Omnibus):	0.020	**Jarque-Bera (JB):**	8.528
Skew:	-0.556	**Prob (JB):**	0.0141
Kurtosis:	5.363	**Cond. No.**	1.74e+05

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 1.74e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Interpretation of Regression Results

Our initial assessment of the results was received jubilantly by the team. The R-Squared of 0.947 implied that our predictors explained ~95% of the variance in the Coffee Quality Indicator.

P-values well below 1% significance for Sweetness, Balance, Acidity, and Uniformity highlight these attributes as statistically significant contributors to coffee quality. However, while these results were initially encouraging, we were puzzled by the note on the large condition number suggesting strong multicollinearity.

After re-reading the documentation on the dataset we discovered that in fact Total Cup Points was merely the numerical sum of the other factors, thus the exact function underlying the CQI variable is already known. This would be akin to predicting a person's weight using Body Mass Index (BMI) which is defined as weight/height^2.

Future analysis would entail possibly finding an alternative Coffee Quality Indicator (CQI) for the regression e.g the average scores given by experts in a blind test.

With a P-value of 0.634, altitude does not appear to have a statistically significant effect in determining coffee quality within the context of our current model. The altitude variable is not collinear with CQI as it isn't an input factor in the construction of the CQI, instead, it stands alone as an independent measure. Isolating it in future analysis could potentially allow us to explore altitude's influence more freely without the direct interference of other variables that are intricately linked to the CQI. However, the current findings suggest that altitude, in the confines of our model and data, does not significantly contribute to the variations in coffee quality.

Sources/Bibliography

A Brief History Of The Price of Coffee. *Sprudge Special Projects Desk*. https://specialprojects.sprudge.com/?p=353.

Climate zone finder. Global Historical Weather and Climate Data | Weather and Climate. 2024. https://weatherandclimate.com/.

Coffee Prices - 45 Year Historical Chart. Macrotrends. April 15, 2024. https://www.macrotrends.net/2535/coffee-prices-historical-chart-data.

Fatih B. *Coffee Quality Data (CQI May-2023)*. Kaggle. https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi.

Scott, Michon. *Climate & Coffee*. June 19, 2015. https://www.climate.gov/news-features/climate-and/climate-coffee#:~:text=Optimal%20coffee%2Dgrowing%20conditions%20include,the%20Middle%20East%3B%20and%20Asia.

The History of Coffee. *NCA*. https://www.ncausa.org/about-coffee/history-of-coffee#:~:text=An%20Ethiopian%20Legend&text=There%2C%20legend%20says%20the%20goat,want%20to%20sleep%20at%20night.