Arabica Coffee Reviews

MIDS W200 Fall 2022, December 5th, 2022

Elias Tavarez Robles, Michael Thottam, and Bailey Kuehl

Background and Motivation

The berries harvested from Coffea plants, better known as coffee, are used to make one of the world's most popular drinks: Coffee [1]. In fact, research shows that nearly 40% of the world's population drinks coffee everyday [2]. However, coffee drinkers are infamous for their particular tastes and preferences when it comes to consuming their beloved beverage. To better understand the relationship between coffee drinkers' perceived taste quality and external coffee production factors (such as harvesting methodology, geographic location, and farming conditions), we investigated coffee reviews from the Coffee Quality Institute's review pages [3].

Primary Research Questions

For this investigation, we focused our research on the following primary questions:

1. Are there correlations between how reviewers perceive coffee's various properties and how they score the coffee overall?
2. Does the geographic region (region, country, etc.) in which the coffee is grown and harvested affect the perceived quality of the coffee?
3. How does the farm's owner and processing methods affect the coffee's flavor?

Project Details

Project Repository

The following repository contains the code and reports used for our investigation.
https://github.com/UC-Berkeley-I-School/Project2_Tavarez_Thottam_Kuehl

Dataset

The dataset we analyzed for our investigation is contained in the following repository:
https://github.com/jldbc/coffee-quality-database/blob/master/data/arabica_data_cleaned.csv

Data Structure

Our dataset is structured as follows:

- **43 columns** - each column is either a review category, quality score, or metadata about the coffee reviewed
- **1311 rows** - each row represents an individual coffee review from one person
- The data was scraped in 2018 from the Coffee Quality Institute's website using a Selenium headless browser and Beautiful Soup
- The data is formatted as a CSV and has been minimally cleaned by github user *jldbc*.

Data Cleaning, Validating, and Exploring

Data Cleaning

To begin the investigation, we reviewed the dataset and determined which of the 43 variables were essential in order to answer our research questions. After an initial glance, we dropped seven columns that had redundant or irrelevant data for our analysis. For example, columns like the `species, owner_1` were not helpful since all species were Arabica and the primary `owner` column already existed. Additionally, columns like `ico.number, certification_address, certification_contact` provided no meaningful information for our analysis of the coffee itself. Lastly, we decided to drop `altitude_low_meters` and `altitude_high_meters`. We assumed that the remaining column, `altitude_mean_meters,` would suffice for our analysis, as it would be more informative than the range of altitude given by subtracting `altitude_high_meters` and `altitude_low_meters`.

Sanity Checks and Validation

Now that we decided which variables to analyze, we wanted to ensure that the remaining variables were formatted as expected and contained enough data to analyze.First, we checked the composition of non-null values for each variable. Given the initial number of rows of 1311, we wanted to ensure that we would have at least 33% of the data in each column to work with to get the best results. We found that `lot_number` had roughly only 20% of the values to be non-null, and thus we decided to drop this column as well.

Data Transformations

There were three main columns which required relatively extensive feature engineering: `harvest_year, coffee_weights,` and `altitude_mean_meters`. Beginning with `harvest_year`, we had an array of different formats (`March 2010, May-August, 2017, 08/09 crop`) which we needed to standardize. To fix this column, we took the following steps:

1. Drop all NaN values
2. Extract the numerical values from each observation
3. Drop all NaN values again (observations like May-August returned as NaN)
4. Apply custom function to convert any strange extractions (i.e. 08 to 2008)

This transformation led to a total of 57 observations removed. The next column we looked at was `bag_weights`, a column we expected to have solely integer values. However, we found that it was an object type composed of the bag weight and the unit (`60 kg, 15 lbs`). Thus, we applied the following transformations:

1. Split the column into two columns `weight` (as `int`) and `unit`
2. Apply custom function to convert any weights with unit in kg to lbs
3. Drop outliers with weights greater than 10,000 lbs

After these steps, a total of 12 observations were removed. Lastly, the final column we engineered was the `altitude_mean_meters` by taking the following steps:

1. Scale outliers with unrealistic altitudes
2. Apply function to convert meters to feet
3. Impute missing values by applying custom function based on `country_of_origin` mean altitude
4. Impute remaining missing values by applying custom function based on `region` mean altitude
5. Drop missing values

With this set of imputations, only a single observation was dropped. As a final feature engineering step, we converted the date columns (`grading_date` and `expiration`) to `datetime` objects.

Imputations

The final set of adjustments we made to our dataset was around data imputation. Columns like `farm_name, mill, producer, color, processing_method, company, variety,` and `region` all had a notable number of missing values. However, instead of dropping these sets of valuable observations, we imputed data for these columns using the following strategies:

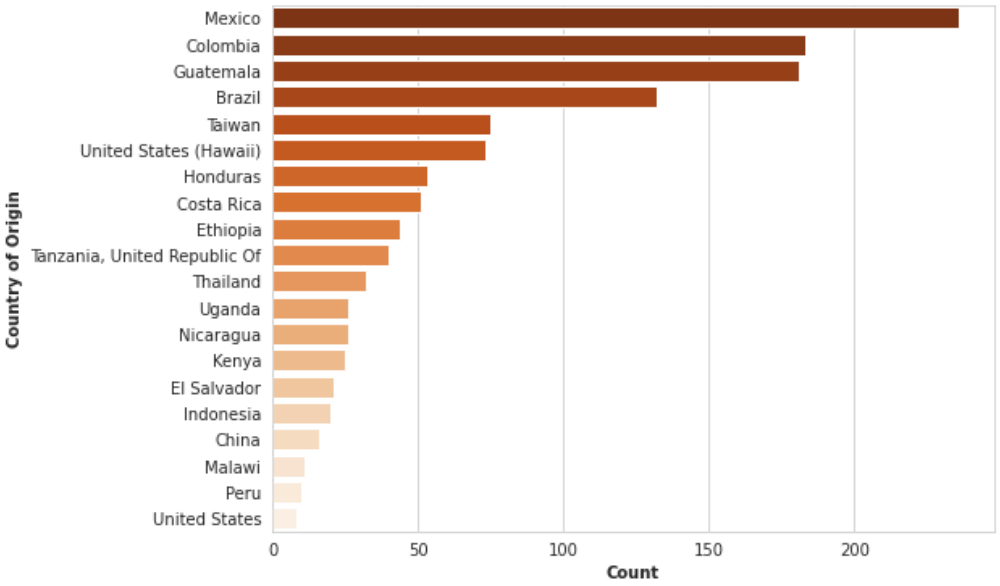
1. Impute `No <column> Specified` (i.e. `No Farm Name Specified`) for columns that had no sort of other/none (`farm_name, mill, producer, company, region`)
2. Impute `Other/None` for columns which already had those two options as part of the dataset (`variety, color, processing_method`)

In the end, we narrowed down the list of variables to be explored to 36 variables and 1241 observations (removed 70 observations, which was only 5.3% of the dataset). While this seems like a lot of remaining variables, it is important to note that 9 of them (25%) all fall within the category of coffee profile and will have similar analyses.

Initial Exploration

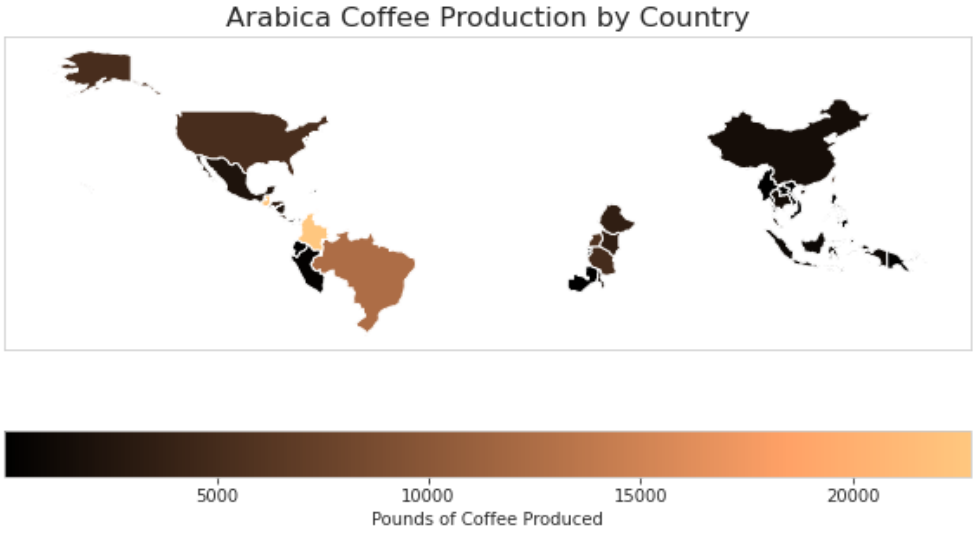
As an initial exploration of the dataset, we examined the variable ‘country_of_origin’ since it was one of the key variables involved in many of our primary research questions. From Figure 1, we found that some countries were underrepresented in the dataset. For instance, coffee coming from the United States and Peru have substantially fewer reviews than those of Mexico and Columbia. This distribution was an important factor to keep in mind during the analysis in order to be aware of easily skewed values for less represented countries.

Figure 1. Distribution of Countries in the Dataset



Additionally, we see that production of Arabica coffee (by weight) was quite different across countries. The lionshare of Arabica coffee in the dataset was produced in the Americas, with smaller productions in Asia and Oceania. Figure 2 visualizes the global distribution of production.

Figure 2. Distribution of Countries in the Dataset



Outliers

Some variables in the dataset did have outliers that were removed for the sake of analysis. Notably, there were observations with a `total_cup_points` of zero that were removed. We believed these scores to be erroneous and they had high leverage on some of our analyses when included.

There were also a handful of observations with extremely large `bag_weights` reported. In some cases, the data suggested that bags of coffee were over 100 pounds. We considered these

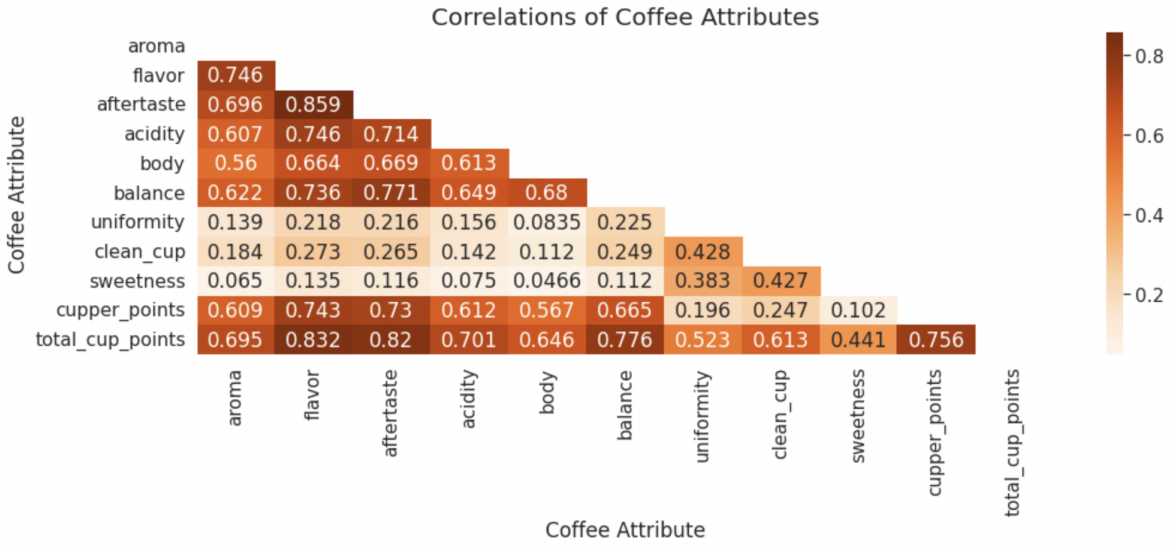
values to be due to entry error and were removed from the dataset (as described in the Data Transformations section).

Results

Question 1: Effects of Coffee Properties on Flavor, Total Cup Points

Our first research question related to coffee properties including acidity, aroma, sweetness, etc. We were curious to find out whether reviewers rankings of these properties on a scale of 1-10 had any correlations with each other. For example, we hypothesized that aroma scores would correlate with flavor scores given that smell and taste are closely connected. To start this exploration, we made a heatmap containing the correlations for each set of coffee property variables (see Figure 3).

Figure 3. Heatmap of Coffee Attribute Correlations

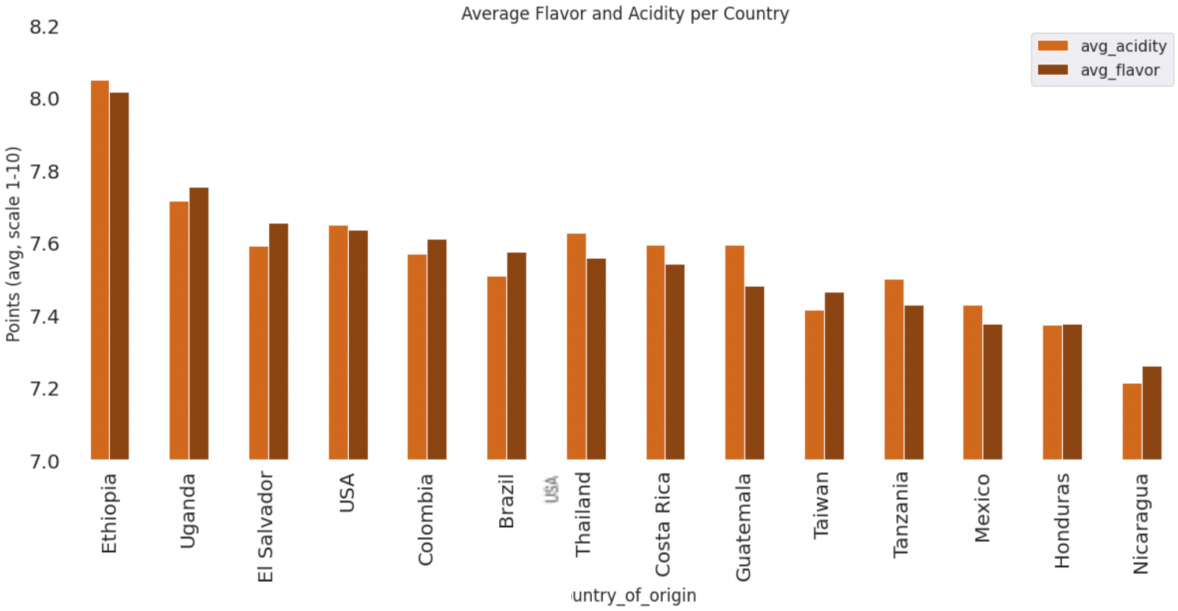


From Figure 3, we find that there is a strong correlation (0.859) between `aftertaste` and `flavor`. We also find strong correlations between `total_cup_points` and `flavor` (0.832), as well as `total_cup_points` and `aftertaste` (0.82). Though slightly weaker, we also find a fairly strong correlation between `aroma` and `flavor` (0.746) as expected based on the close relationship between taste and smell. There is also a strong correlation between `acidity` and `flavor` (0.746).

From this information, we draw the conclusion that the aftertaste, being the lasting impression of the coffee on the reviewer, likely has a high impact on the reviewer's perceived overall flavor. Based on these two ratings, the overall score would likely be subsequently affected, which is reflected by a correlation between these two attributes and `total_cup_points`. Based on these findings, we became more curious about some of these correlations and took a deeper dive into the data.

As we know that our dataset contains reviews of coffee from several countries, we were curious to investigate how flavor perception varied from country to country. Our findings are shown below in Figure 4.

Figure 4. Distributions of Flavor and Acidity Ratings by Country



As seen in Figure 4, we found that Ethiopia has the highest rated `flavor` with an average score of 8.01 across 40 different samples. It is also interesting to note that Ethiopia also rated highly for `acidity` (8.05). Similarly, many of the countries appear to have similar `acidity` and `flavor` scores. Generally, we see that as acidity decreases, flavor ranking also tends to decrease. Thus, this confirms the correlation demonstrated in Figure 3 between `flavor` and `acidity`.

Question 2: Effects of Geographical Conditions

To understand if flavor profiles differ by region, we evaluated flavor ratings at both the country- and continent-level. The overall finding was that flavor ratings were not well differentiated by country but there were some countries that had unique characteristics.

While the Americas produced more coffee in general some African countries (including Kenya, Ethiopia, and Rwanda) seemed to score slightly higher on flavor metrics such as Aroma, Acidity, and Balance). The entire profile of scores can be seen in Figure 9 (in the appendix, due to size).

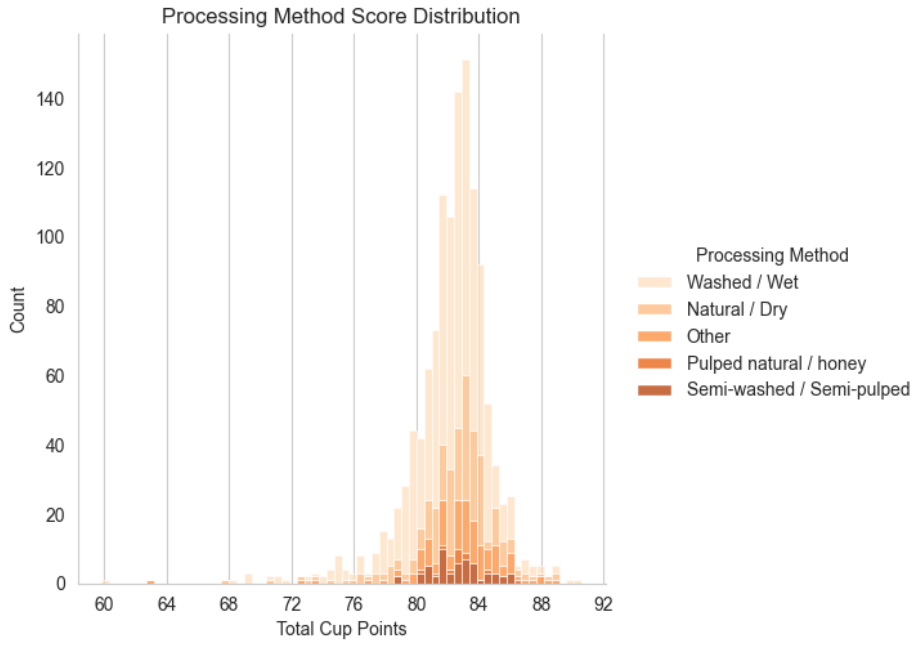
Note that Papua New Guinea also had very high scores, driving overall high performance for the Oceania continent. But it is worth mentioning that their production comprised less than 1% of global production.

We hypothesized that the altitude of a coffee farm might have an impact on the overall flavor score [6]. The correlation between altitude and total score was only 0.18, signaling that there might be other geographical features that are more predictive of flavor than altitude.

Question 3: Effects of Farming Processes

In our third question, we sought out to answer how the different processing methods affected the coffee's quality. Generally speaking, there are three main processing methods: washed/wet, natural/dry, and pulped natural/honey [4] . In our dataset, we had three all three groups and two additional groups: semi-washed (also referred to as wet-hulled processing [5]) and "other". To determine whether there was a method which tended to score higher, we generated the following plot:

Figure 5. Processing Method Effect on Total Cup Points



As seen from Figure 5, while there was certainly a varying number of observations for each group, they all shared a similar, normal distribution centered around a score of 82. This is no surprise since there really isn't a "best" method, but each method does draw out certain characteristics from the coffee.

For example, we observe the following attributes for each method:

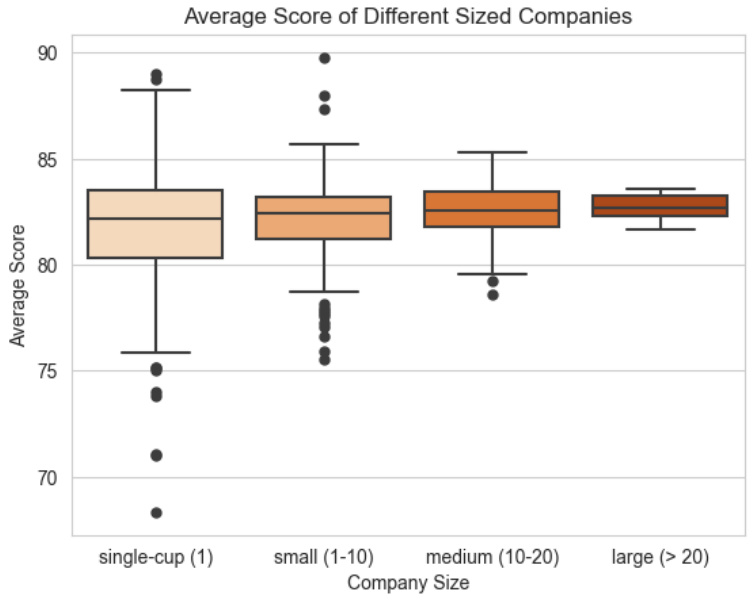
Table 1

Method	Washed/Wet	Natural/Dry	Pulped Natural/Honey	Semi-Washed
Attribute	Adds bright, aci flavors	Adds sweetness/ fruitiness	Has best attributes of b washed/wet and natural/dry flavors.	Has best attributes of both washed and wet and natural/dry flavors as it's a hyb of the two

Aggregating the data by the processing groups maximum and average values, we do find conformance with the washed/wet method as it had the coffee with the highest acidity. However, the rest of the characteristics for each group were not as definitive (see appendix figure X).

Along with the processing methods, we were also curious about how the score of the coffee could be impacted by the size of the farm/company. In other words, if a company produces more than one coffee, is it likely that they will have a lower quality of coffee than say companies with fewer amounts of coffee since they should have greater control over their quality? To investigate this, we first grouped our data by the `company` column and aggregated it by the count of observations (in other words, the number of coffees a company had graded) and the mean `total_cup_points`. From that grouping, we then binned the companies into “single-cup”, “small”, “medium” and “large” companies and we generated the following plot:

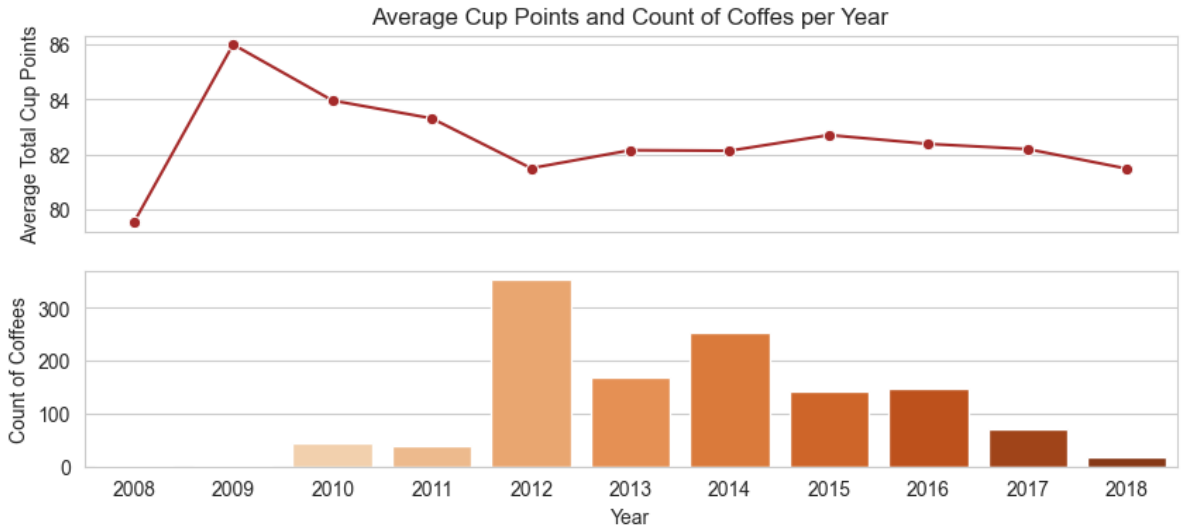
Figure 6. Company Sample Size Effects on Total Cup Points



From Figure 6, we can see that owners that produce several coffees tend to rank among the average ratings across all owners. As also noted from the graph, owners which received the highest scores (the outliers in this case) were owners that had fewer than ten cups of coffee, with most of them in the single-cup and small categories. Granted, those two categories also had the worst coffees, indicating that a company which has only a few coffees is not necessarily always better.

As the final piece of information we wanted to review regarding coffee farming conditions was the year of harvest. Was there a specific harvest year which outperformed the others? The answer to that question is what the following charts answered:

Figure 7. Effect of Coffee Harvest Year on Total Cup Points



With only the line plot one could say that coffee harvested in 2008 and 2009 was the worst and best respectively. However, by combining with the bar plot, we can see that those scores are composed of very few coffees (three coffees for 2008 and one single coffee for 2009). All other years had at least 19 or more coffees harvested and from that group of coffees, we see that 2010 had the best scoring coffee; albeit, there's not much separating this group of coffees year to year.

Concluding Thoughts

To summarize, we explored coffee reviews in order to gain insight into the tasting nuances hidden within a single cup of coffee, as well as understand how these perceived nuances may vary with respect to different harvesting and growing conditions.

Overall, we found the following according to each research question:

- Question 1: Effects of Coffee Properties on Flavor, Total Cup Points
    - Reviewers tended to rate coffee flavor higher if it had a good aftertaste
    - Ethiopia has the most flavorful coffee, according to reviewers
- Question 2: Geographic impact of coffee quality
    - While the Americas produce the most Arabica coffee, it seems that the African countries seem to produce slightly higher quality coffee
    - Elevation had a relatively small impact on coffee grades
- Question 3:
    - There is no best processing method. In general all processing methods tend to have the same average score.
    - The number of coffee cups that a company makes does not affect the average score of their coffees. However, we found that both the best and worst scores came from companies which had less than 10 coffees graded.
    - The year in which the coffee was harvested did not have a notable impact on the coffee score. However, we did find that the majority of the graded coffees were harvested from 2012-2014.

Appendix

Table A.1 - Processing Method Aggregation by Average Coffee Characteristics

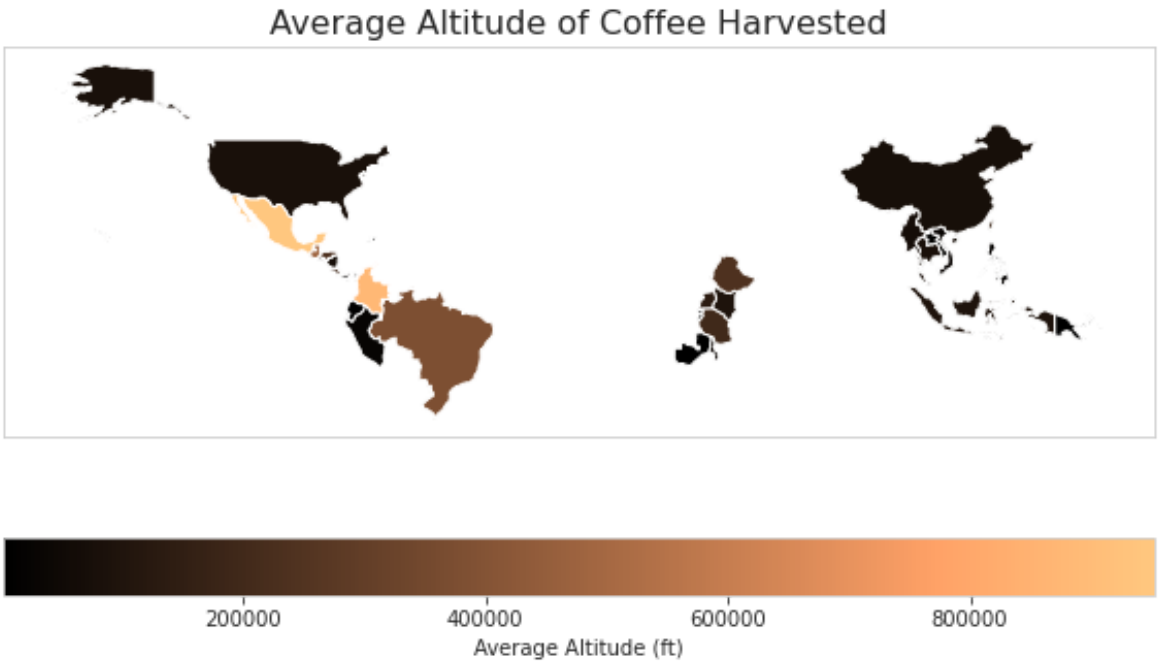
processing_method	aroma	flavor	aftertaste	acidity	body	balance	uniformity	clean_cup	sweetness	cupper_points
Natural / Dry	7.59	7.58	7.46	7.55	7.59	7.59	9.79	9.80	9.82	7.58
Other	7.61	7.61	7.49	7.58	7.55	7.58	9.88	9.88	9.92	7.66
Pulped natural / honey	7.54	7.51	7.53	7.55	7.62	7.52	10.00	10.00	10.00	7.54
Semi-washed / Semi-pulped	7.60	7.59	7.46	7.54	7.55	7.58	9.87	9.95	9.95	7.54
Washed / Wet	7.55	7.48	7.36	7.53	7.49	7.48	9.85	9.83	9.93	7.44

Table A.2 - Processing Method Aggregation by Average Coffee Characteristics

processing_method	aroma	flavor	aftertaste	acidity	body	balance	uniformity	clean_cup	sweetness	cupper_points
Natural / Dry	8.58	8.67	8.50	8.50	8.50	8.58	10.0	10.0	10.0	8.67
Other	8.67	8.67	8.58	8.42	8.58	8.75	10.0	10.0	10.0	10.00
Pulped natural / honey	8.00	8.00	8.00	8.25	8.00	8.17	10.0	10.0	10.0	8.17
Semi-washed / Semi-pulped	8.50	8.17	8.00	8.08	8.33	8.25	10.0	10.0	10.0	8.42
Washed / Wet	8.75	8.83	8.67	8.75	8.50	8.58	10.0	10.0	10.0	8.75

As seen in the two tables above, there is not much to distinguish the known characteristics of the different processing methods. The only notable exception is that the Washed/Wet did have the largest max `acidity` score, which matches processing expectations.

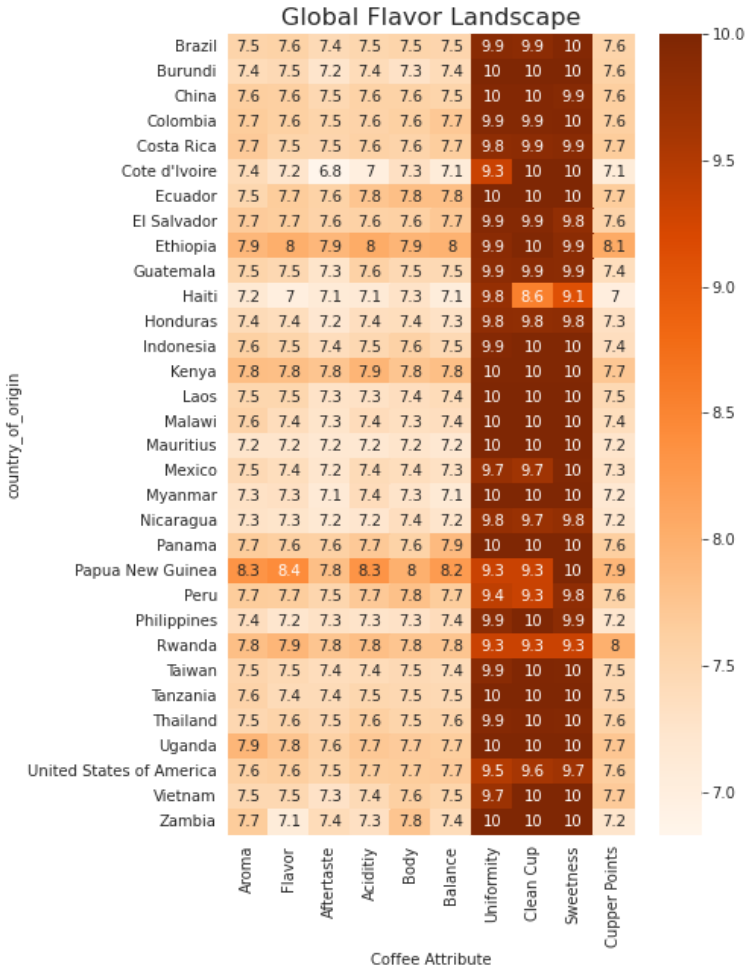
Figure A1. Average Elevation of Coffee Harvests by Country



We also evaluated the average elevation for each country's harvests. The figure above was compared against a country's production and average rating. The relationship between elevation and taste score was considered weak.

Figure A2. Flavor Ratings Aggregated by Country

The heatmap below shows the overall flavor ratings for coffees brewed in each country. We notice that generally there is a narrow range of scores for all coffee flavor ratings. Discussion of these results can be found in the “Effects of Geographical Conditions” section.



References

[1] What is coffee? https://www.ncausa.org/About-Coffee/What-is-Coffee
[2] Coffee Consumption Statistics: https://coffee-rank.com/world-coffee-consumption-statistics
[3] Coffee Quality Institute: https://www.coffeeinstitute.org/
[4] Barista Institute: https://www.baristainstitute.com/blog/jori-korhonen/january-2020/coffee-processing-methods-drying-washing-or-honey
[5] Wet Hulled Processing: https://www.trabocca.com/process/semi-washed/
[6] Impact of altitude on coffee flavor: https://caffenero.com/us/the-journal/how-altitude-affects-the-taste-of-coffee/