What Make Coffee Popular?

Ken Trinh, Michelle Lee, Tanmay Mahapatra

03/30/2022

Contents

1 Introduction

Coffee is one of the most popular non-alcoholic beverages globally prized for its attributes. Coffee is brewed from roasted beans of the plant species Coffea, which is native to sub-Saharan Africa and individual islands in the Indian Ocean. Ever since its discovery, coffee has grown to more than 70 tropical countries, making up a multi-billion dollar market. Despite being consumed for over 120 years, there are still new trends entering the market every year, shifting the popularity of different coffee types. Some of these trends include different beans, caffeine concentration, and brewing techniques. Constant changes in coffee trends - everything from the variety of the plant, the chemistry of the soil, the weather, the amount of rainfall and sunshine, and even the precise altitude at which the coffee grows - can affect the popularity of the final product.

In an effort to understand what drives the popularity of coffee, our team proposed an investigation following the question:

How does the country of origin affect the popularity of a coffee?

Our research planned to tackle this question using the coffee data set gathered by tidytuesday[1]. We believe that answering this question would enable us to identify how different factors affect the rating of coffee. Ultimately, the result of our analysis would help ACME with identifying new coffee products to include in the store by prioritizing features as well as with improving our marketing strategy.

2 Operationalization

We conceptualized the popularity of coffee as a rating point given to a cup of coffee out of 100 points and country of origin as the country that the coffee bean came from. Some covariates that we plan to consider in the study are `flavor` and `aroma`. We conceptualized both of these two covariates as a numerical grade given to the coffee bean out of 10.

As mentioned, our analysis will use the 2020 coffee data set gathered by tidytuesday[2]. We planned to operationalize the observational data collected as follows:

- First, we use a column named `total_cup_points` that came with the data set as a measurement for our conceptualization for the popularity of coffee. This variable is already given as a numerical value out of 100 - the greater the value, the more popular the coffee is. Since this is a numerical value, we will apply the appropriate transformation to get the value normally distributed. We will also filter out extraneous variables such as infinity, NaN, NAs as needed in our modeling process.
- Second, we identified our main input variable of interest - `county_of_origin`. This variable represents the country that the coffee bean is from, such as Ethiopia, Brazil, and Peru.
- Third, we identified both the `flavor` and `aroma` covariates as additional variables that affect coffee bean rating. These two variables are recorded as numerical values ranging from 0 to 10, where 0 is the lowest grade given and 10 is the highest rate given. Since these two covariates are numeric, we will apply the appropriate transformation to get the value normally distributed.
- Finally, we plan to transform the `flavor` and `aroma` using multiplication in order to create an interaction term. Doing so, we will be able to extract some form of non-perfect collinearity information from these two variables. The results are then tested appropriately in this report.

[1]Rfordatascience. (2020). Tidytuesday/readme.md at master · rfordatascience/tidytuesday. GitHub. Retrieved March 23, 2022, from https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-07- 07/readme.md

[2]Rfordatascience. (2020). Tidytuesday/readme.md at master · rfordatascience/tidytuesday. GitHub. Retrieved March 23, 2022, from https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-07- 07/readme.md

3 Data Understanding

To answer our research question, we will be conducting an explanatory analysis using linear models to determine if there is a **causal** relationship between country of origin and total cup points.

3.1 Input Variable I: Country of Origin

First, we will look at how many unique countries of origin are included in the data along with how many data points we have for each country. Looking at the results below, we see that out of 37 unique countries of origin, 80% of the data comprises top 10 countries and 60% of top 5. In order to simplify the model, we will filter to the top 5 countries of origin for the analysis.

```
## [1] 37
```

country_of_origin	count	prop	runningprop
Mexico	236	17.63	17.63
Colombia	183	13.67	31.30
Guatemala	181	13.52	44.82
Brazil	132	9.86	54.68
Taiwan	75	5.60	60.28
United States (Hawaii)	73	5.45	65.73
Honduras	53	3.96	69.69
Costa Rica	51	3.81	73.50
Ethiopia	44	3.29	76.79
Tanzania, United Republic Of	40	2.99	79.78

3.1.1 Distribution of Total Cup Points by Top 5 Countries of Origin

The histogram in Figure 1 shows the distribution of `total_cup_point` with respect to countries of origin. For each of the country, the plot shows a slight left skew with the distribution looking more symmetric past 70 total cup points. The distribution of each country looks fairly similar to each other, with their median aligning around 82 points.

3.2 Input Variable II: Flavor

Figure 2 shows a distribution of `flavor` grade across the top 5 selected countries. The histogram looks fairly symmetric and normal with no heavy skews or outliers. For this variable, it doesn't seem like any transformation would be needed.

3.3 Input Variable III: Aroma

Figure 3 shows a distribution of `aroma` grade across the top 5 selected countries. The histogram shows that `aroma` grade has an outlier near 5, but it seems fairly normal overall. We believe that no additional transformation is needed on this particular variable.

3.4 Outcome Variable: Total Cup Points

After filtering to the top 5 countries of origin, Figure 4 (right) shows that we still see a slight left skew. However, the `total_cup_point` distribution looks fairly symmetric past the 70 total cup points and no other transformation seems to be needed. Given that we have over 800 data points, we can apply CLT here and no additional transformation is needed.

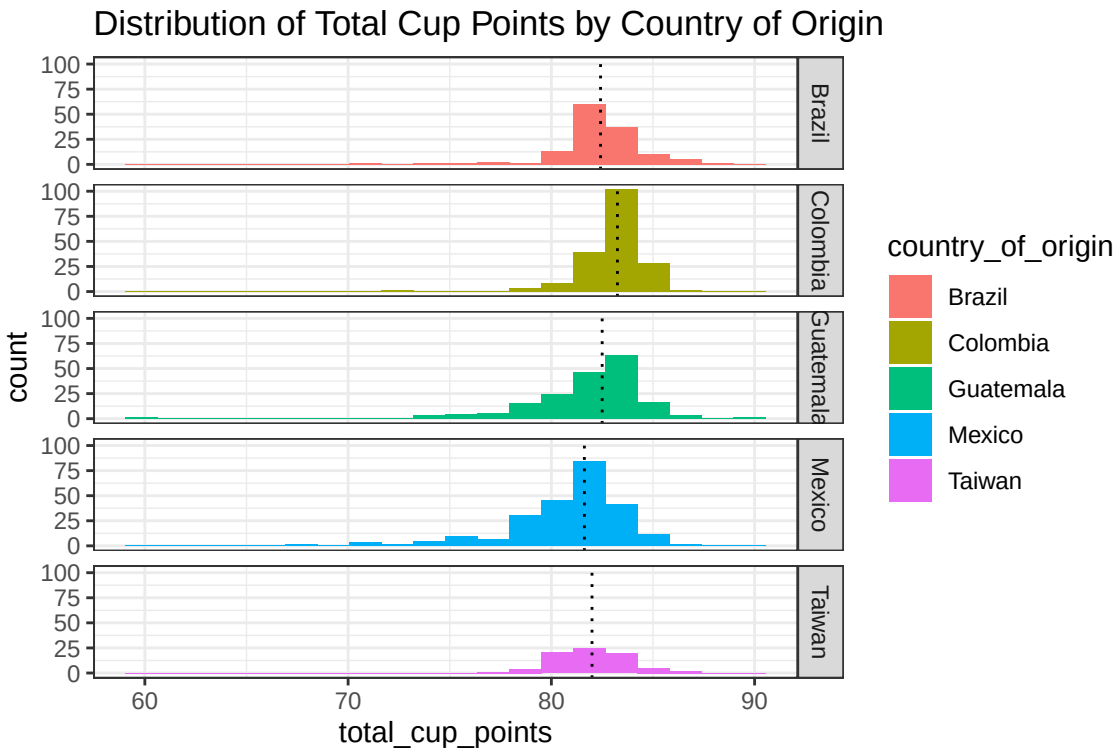

Figure 1: Coffee Distribtion Top 5 Countries

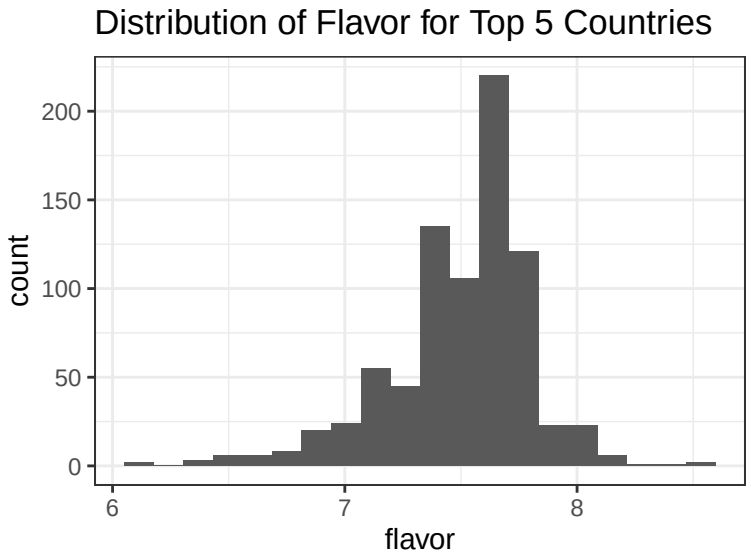

Figure 2: Distribution of Flavor for Top 5 Countries

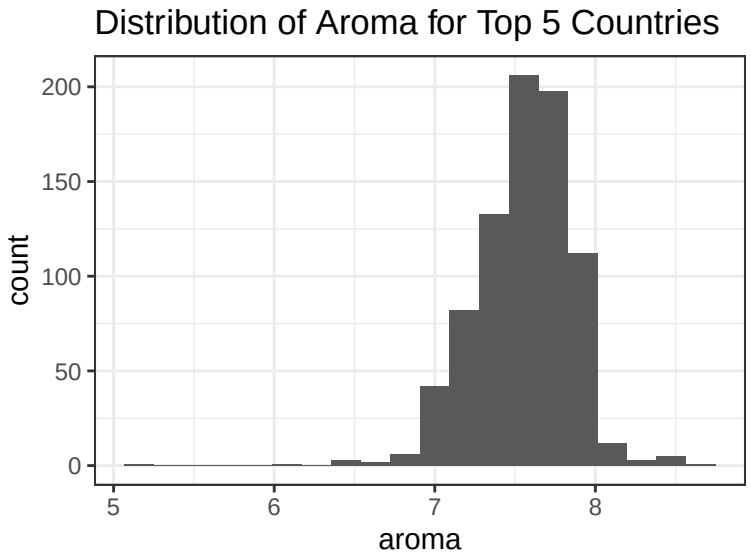

Figure 3: Distribution of Aroma for Top 5 Countries

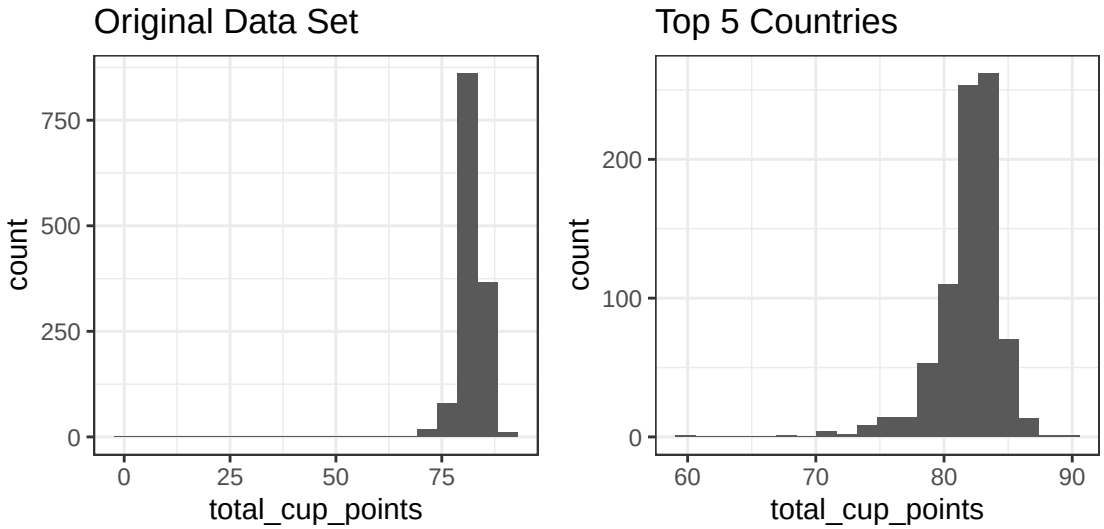

Figure 4: Total Distribution of Total Cup Points (Left) and Top 5 Countries Distribution of Total Cup Points

4 Model Specification & Building

After the data is collected and processed, we plan to tackle this problem by building three models to better understand what makes a highly rated cup of coffee. In each of these models, we will utilize the data outlined in Section 2 of this report.

4.1 Model 1

The first model will capture the effect of country of origin on coffee ratings. Based on the result of our exploratory data analysis, we identified that there are 37 unique countries. For simplification, we will use only the top 5 countries with the most data recorded. The five countries that we include in our model are Brazil, Columbia, Guatemala, Mexico and Taiwan. Here our model will drop one country as a dummy variable to avoid linear dependence. The following equation describes the first model:

$$\begin{aligned} total\_cup\_points = \beta_0 + \beta_1 Columbia + \beta_2 Mexico \\ + \beta_3 Guatemala + \beta_4 Taiwan \end{aligned}$$

4.2 Model 2

Because there could be other confounding variables, our second model will include both `flavor` and `aroma` as additional covariates that could affect the rating of a cup of coffee. We will add the two new covariates as a continuation of `model_1`. The following equation is a representation of the second model:

$$\begin{aligned} total\_cup\_points = \beta_0 + \beta_1 Columbia + \beta_2 Mexico \\ + \beta_3 Guatemala + \beta_4 Taiwan \\ + \beta_5 Flavor + \beta_6 Aroma \end{aligned}$$

4.3 Model 3

Furthermore, we wanted to better understand whether or not there is a non-perfect collinearity relationship between `flavor` and `aroma` that could affect the model. For this, we applied a multiplication technique to create an interaction term between `flavor` and `aroma`. The following equation is a representation of the third model:

$$\begin{aligned} total\_cup\_points = \beta_0 + \beta_1 Columbia + \beta_2 Mexico \\ + \beta_3 Guatemala + \beta_4 Taiwan \\ + \beta_5 Flavor + \beta_6 Aroma \\ + \beta_7 Flavor\_Aroma\_Interaction \end{aligned}$$

5 Results

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sat, Apr 09, 2022 - 11:00:32 PM % Requires LaTeX packages: dcolumn

Table 1: Factors Affecting Coffee Ratings

	Dependent variable:		
	Total Cup Points		
	(1)	(2)	(3)
Columbia	0.701***	0.472***	0.451
	(−0.035)	(−0.045)	(0.595)
Guatemala	−0.559***	−0.084***	−0.081
	(−0.035)	(0.002)	(−1.047)
Mexico	−1.516***	−0.347***	−0.355
	(−0.035)	(−0.084)	(1.143)
Taiwan	−0.405***	0.222*	0.215
	(−0.035)	(−0.125)	(0.331)
Flavor		5.525***	8.924
		(−0.628)	(−116.425)
Aroma		0.989***	4.487
		(0.081)	(−130.265)
Flavor-Aroma Interaction			−0.462
			(16.378)
Constant	82.406***	33.091***	7.399
	(0.035)	(4.197)	(928.025)
Observations	807	807	807
R^2	0.108	0.676	0.677
Adjusted R^2	0.104	0.674	0.674
Residual Std. Error	2.358 (df = 802)	1.422 (df = 800)	1.421 (df = 799)
F Statistic	24.264*** (df = 4; 802)	278.662*** (df = 6; 800)	239.297*** (df = 7; 799)
Note:			*p<0.1; **p<0.05; ***p<0.01

5.1 Model 1 - Intepretation

Because countries of origin are given as categorical variables, one of the categories is dropped to ensure linear independence. Here, the first model shows that 5 out of 5 countries have statistically significant p-values, suggesting that country of origin does affect the rating given to a cup of coffee. At baseline, when Brazil is the country of origin, the total cup point is approximately 82.4. Interestingly, only Columbia is shown to have a positive coefficient, suggesting that coffee originating from Columbia added approximately 0.701 points to the total cup point when compared to coffee originating from Brazil. Both Guatemala and Mexico have negative coefficients, suggesting that there is a decrease in the total cup point by 0.559 points for coffee originating from Guatemala and 1.516 points for coffee originating from Mexico when compared to coffee originating from Brazil. Finally, coffee originating from Taiwan is shown to have a negative coefficient. We can interpret this in a similar fashion where coffee originating from Taiwan decreases the total cup point by 0.405 points.

5.2 Model 2 - Interpretation

As mentioned, the second model includes two covariates `flavors` and `aroma`. In our analysis, both covariates are statistically significant. For interpretation, we see that if the coffee originated from Brazil, an increase in 1 point of flavor grade, while holding aroma grade constant, added 5.525 points to the total cup point. Similarly, an increase in 1 point in aroma grade, holding flavor grade constant, added about 0.989 points to the total cup point. We can see a positive correlation between these two variables and total cup points.

5.3 Model 3 - Interpretation

Because flavor and aroma can have a non-perfect collinearity, we wanted to address how an interaction term between these two covariates would affect our model. Unfortunately, the coefficients obtained for this interaction term in our third model is not statistically significant so the analysis here might not be too insightful.

5.4 Covariates Enchancing Effect

Since we broken our modeling processes into three models, we can assess the significance of the added covariates. To do so, we conducted an F-test or our models. Here we assume the null hypothesis that `model_1`, using only `country_of_origin`, is sufficient in explaining the variance of `total_cup_point` population.. Our alternative hypothesis is that `model_2`, adding `flavor` and `aroma`, is more sufficient in explaining the variance of `total_cup_point` population. Our group decided against including `model_3` in the F-test because we did not introducing any additional confounding variables that `model_2` did not already capture.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
802	4457.450	NA	NA	NA	NA
800	1617.128	2	2840.321	702.5593	0

According to the F-test, adding `flavor` and `aroma` increase the degree of freedom by 2. The F statistics is computed to be 702.56 with the p-value being less than .001, suggesting a statistically significant result. Therefore we would reject the null hypothesis, thus adding in the two covariates is better in explaining the variance of `total_cup_point`. Essentially, this means that aside from country of origin, adding flavor and aroma of coffee beans also has an effect on the rating of coffee.

6 Model Assumptions and Limitations

We are looking at a sample drawn from a data set of over 800 data points which allows us to use the Central Limit Theorem (CLT) to approximate normality. Here, we will follow the Large Model Assumptions.

6.1 Large Model Assumption

1. IID

The data set is not I.I.D.; The coffee rating of the given data set might not be independent of each other because:

- Between countries, there is a relationship between geographic region and coffee species. In our data set, both Guatemala and Mexico are located in Central America, creating a geographic clustering. However, Figure 1 shows that the distributions are not too different from each other. So we can continue with our modeling approach.
- The same taster may have rated various other coffee beans so their preferred one would be rated higher than the others. However, Figure 4 shows that the distributions is normally distributed. So we can continue with our modeling approach.

The data is not identically distributed. Since our data set consists ratings for mostly Arabica species of coffee beans we believe that the ratings cannot be assumed applicable to all species of coffee beans.

2. Unique BLP exits

In order to determine if a unique BLP exists, we need to assess the data and ensure that there is no perfect collinearity between variables.

In the figures, we observe that the distribution does not exhibit large kurtosis which means that the it produces fewer and/or less extreme outliers and the covariance is measurable.

In the coefficient table below, no variables were dropped during our modeling process. As such, we can conclude that a unique BLP exists for our linear model.

	coeficients
(Intercept)	7.3993831
I(country_of_origin)Colombia	0.4509475
I(country_of_origin)Guatemala	-0.0808141
I(country_of_origin)Mexico	-0.3545963
I(country_of_origin)Taiwan	0.2152821
flavor	8.9235392
aroma	4.4872686
I(flavor * aroma)	-0.4621801

6.2 Omitted Variable Bias

The factor being studied is in truth associated in the target population with few other factors that could influence the outcome of interest. We list the most important of them below:

1. Caffeine Content

Consumers are increasingly concerned about the caffeine levels of the coffee they drink. Caffeine has a flavor, plus it also reacts with other ingredients to produce new flavors, which can affect the flavor score. If the taster likes the flavor of a dark, bold roast, decaffeinated coffee probably won't taste as good to her so she is more likely to rate the former higher. On the other hand, if she likes a light roast, she might prefer the flavor of decaf and rate it higher. All things considered, we believe the caffeine content would have been an important factor to consider and it would be interesting to see how it affects the results.

In this case the direction of bias is toward zero in that perhaps the caffeine content creates an affinity that prompts the consumer to rate one type of coffee bean higher than the others.

2. Roasting Temperature

Many consumers have favored roast choices, be it temperature or time. For instance consumers assume that the strong, rich flavor of darker roasts indicates a higher level of caffeine and may rate the corresponding bean higher. The perfect roast is a personal choice that is also sometimes influenced by national preference or geographic location. We believe this factor could also provide some interesting insight.

In this case the direction of bias is yet again toward zero in that the consumers have a preference in taste that results from the roast (i.e. temperature and roasting time).

3. Type of Soil/Fertilizer

The differences between particular coffees are in the 'terroir'. Different environmental and regional factors influence that subtle and dramatic taste differences between coffees. Soil, altitude, sun, wind and rainfall all play a part in shaping the flavor of the coffee.

In this case perhaps the direction of bias is toward zero since the variation in the type of soil and the growing process across the countries contribute to the resulting attributes in the end product.

7 Conclusion

7.1 Overall Conclusion

In this research, we set out to investigate the effect of country of origin on the popularity of coffee. To provide insight into the popularity of coffee, we gathered data based on the bean's country of origin, the flavor and the aroma of the bean, and the total rating given to a cup of coffee brewed using the bean. Here, the popularity of coffee is defined as the `total_cup_point` variable in our data set. Since there were many countries producing coffee beans, we limited our study to the top five countries. The two additional covariates used in this study are `flavor` and `aroma` grade for the bean. We filtered out our data set based on outliers to bring each variable to normality. The final data set used in this study has 807 observations and 7 variables. From our analysis, we concluded that country of origin does have an effect on the popularity of coffee. However, the effect of each country is different. In our sample, only Columbian coffee beans show a positive effect on the total cup points when compared to Brazilian beans with an effect size of 0.701 units. Furthermore, our analysis shows that an increase of 1 point in `flavor` grade added 5.525 points to the total cup points and an increase of one point in `aroma` grade added 0.989 points to the total cup point. This result suggests that a high grade `flavor` or `aroma` statistically improved the popularity of a cup of coffee. Our analysis further shows that adding an interaction term between `flavor` and `aroma` is not statistically significant, suggesting that there is no collinearity between these two variables.

7.2 Business Strategy and Further Studies

While we believe that different sets of countries will show different results, we believe that it is fair to direct ACME's product research and development team to target a South American based coffee bean, particularly Brazil and Columbia, with high `flavor` and `aroma` grade as potential product to carry in our stores. Based on the result of our analysis, we believe that this strategy will boost the popularity of our coffee lineup, leading to more sales.

While this might be a good short term solution, it should be noted that the data set used in this study lacks information on the caffeine contents, roasting temperature, and the type of soil used. Due to this limitation, we wish to include these additional information, along with their potential interactions, in future studies. Our reasoning is that the inclusion of these variables would reduce some of our omitted variables bias, leading to better models and insights.