Identifying Determinants of Coffee Bean Quality

John-Tianyu-Fabian

2022-04-11

1 Introduction

1.1 Movtivation

The average American consumes ~2 cups of coffee per day, making coffee the preferred American drink (Holcomb, 2021). The coffee market is expected to grow 13.2% yearly over the next 4 years (Maida, 2022). Though some brew coffee at home to save money, it is generally known that the best coffee comes from a true coffee shop, where trained baristas used complex machinery to delicately convert specialty coffee beans into a delicious drink. There are many different ways to brew and consume coffee; some prefer espresso, some prefer pour-over, others prefer drip coffee. However, the brewing process only has a minor effect on the quality of the end product; instead, the quality of coffee is primarily determined by the quality of the bean sourced.

Before making it to the local coffee shop, the coffee bean goes through a lengthy cultivation, harvesting, and treatment process. The coffee bean is actually a seed extracted from a coffee cherry, the fruit of a coffee tree. Coffee trees are farmed under very specific conditions, primarily in warm climates around the world. Once a coffee tree is ready for harvest, the bean is extracted from the coffee cherry before being processed or washed and then shipped to different buyers around the world. Existing literature shows that the location of the farm is critical for the bean quality. Primarily, higher altitude has been shown to result in higher quality beans (Fleisher, 2017). This is largely because higher altitudes cause slower photosynthesis, which allows plants to metabolize nutrients more gradually and produce bigger and higher quality beans (Aprile 2017). Additionally, the distance from the equator has a significant impact; the coffee belt refers to the many farms within 30 miles from the equator (Clayton, 2021). Each step of the cultivation, harvesting, and washing process has a tremendous impact on the quality of the final product. Given this, coffee shops and distributors place great attention and emphasis on sourcing quality beans in order to improve their product.

To provide transparency to the coffee market, each bean is assessed by a set of trained judges every year. The Coffee Quality Institute (CQI) determines the "total cup score" of each bean through a rigorous, and consistent process to prevent bias from affecting the total cup scores. Total cup score is a metric first defined by the Specialty Coffee Association, and it is the result of rating 10 different attributes of the coffee on a scale 6-10 (such as aroma, acidity, flavor, balance, etc.) and then summing across dimensions for the total score. For each bean, the total cup score is determined by a "Q Grader", who first has to pass a series of certification tests by the CQI to qualify. The CQI also imposes rules on the Q Graders to ensure the integrity of bean ratings and avoid bias on database:

1. Q Graders must have no known ownership or interest in the coffees to be graded and will evaluate the coffee sample objectively

2. Q Graders must have the bean sample assigned to them on the CQI Database prior to grading the sample

3. Q Graders must be regularly rotated to provide equal opportunities and to ensure the transparency of the Q Coffee System
   This total cup score is commonly used to evaluate different coffees for importation and selections for specialty coffee shops.

This total cup score is commonly used to evaluate different coffees for importation and selections for specialty coffee shops.

1.2 Research Question

The goal of this analysis is to understand the key determinants of coffee bean quality (total cup score), to enable farmers to grow better beans and coffee shop owners to source better beans. In particular, this analysis will seek to understand the causal impact of altitude on bean quality through the use of linear models. The main research question is:

> How does the altitude of the bean cultivation affect the bean's total cup score?

Given the importance of sourcing quality beans and the expected growth of the coffee market, it is important for coffee shops to be diligent in the sourcing of their beans to provide a quality product that will retain existing and win new clientele. Additionally, coffee farmers must make better decisions throughout the cultivation process to improve the quality of their beans and compete in the growing market.

2 Data and Methodology

2.1 About the Data

Our analysis leverages a dataset produced by the Coffee Quality Institute (CQI) and hosted on Kaggle. After cleaning, the dataset has about 1,000 total records of Arabica coffee beans produced by different suppliers in a range of countries, spanning from 2009 through 2018. Each record includes the total cup score of the bean, as well as descriptive data points on the type of bean, the cultivation process, and the location of cultivation.

A second dataset, sourced from Kaggle was used to map the latitude of each country into the data.

2.2 Total Cup Score Characteristics

The total cup score variable is the outcome variable of interest. Each coffee bean in our dataset has a score between 60-100 since it is the sum of 10 coffee attributes, each rated on a scale of 6-10. According to the Specialty Coffee Association of America, the component attributes are:

Attribute	Description
Acidity	From bright (good) to sour (bad)
Flavor	Measurement of enjoyable flavor after sipping and before swallowing
Aroma	Dry fragrance and wet fragrance evaluated
Aftertaste	Measurement of enjoyable flavor after coffee is consumed
Uniformity	Consistency of flavor across samples
Balance	Balance of Body, Flavor, Acidity and Aftertaste
Body	The feeling of tongue and mouth after consumption
Sweetness	From sugary (good) to astringent flavors (bad)
Clean Cup	Lack of any negative attribute across consumption process
Cupper Points	General score from grader

The probability density function (pdf) of total cup score approximates a normal distribution with a left skew, as shown in Figure 1. This is likely a result of some beans being very low quality due to poor conditions or mistakes in the cultivation. The total cup score variable has a mean of 82 points and standard deviation of 3.2 points.

Figure 1: Distribution of Total Cup Point

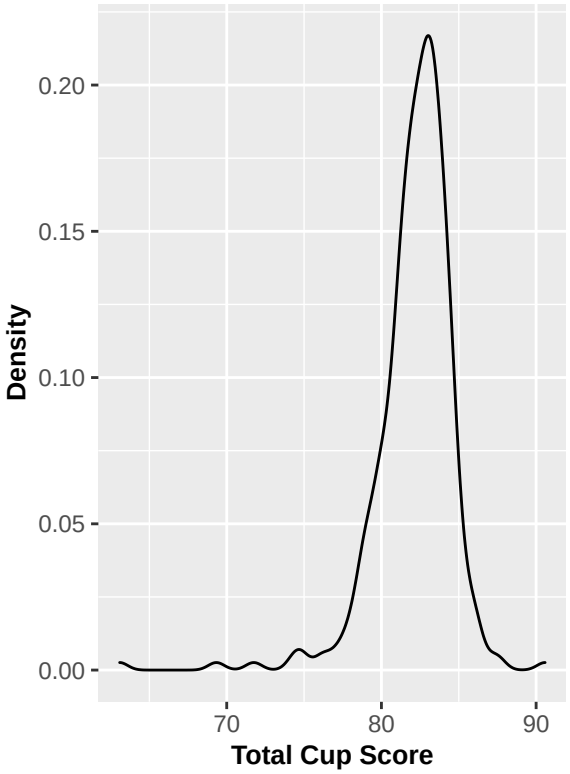
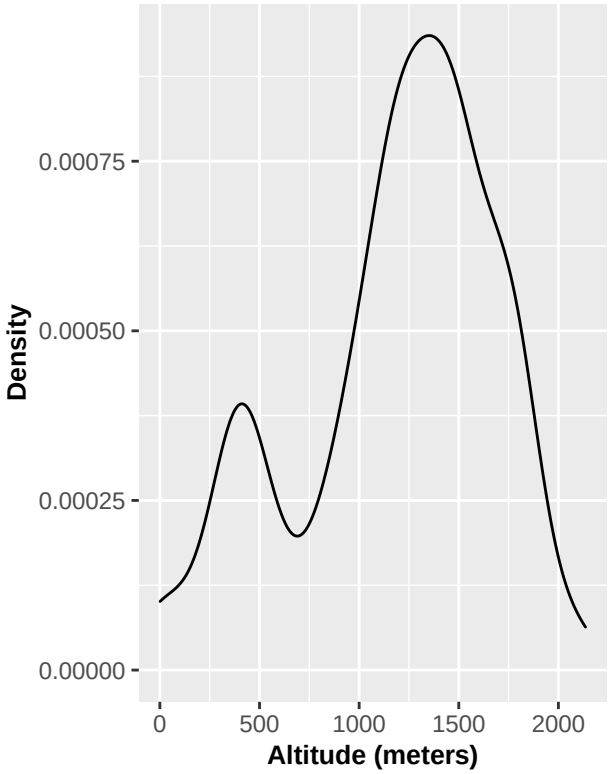


Figure 2: Distribution of Altitude

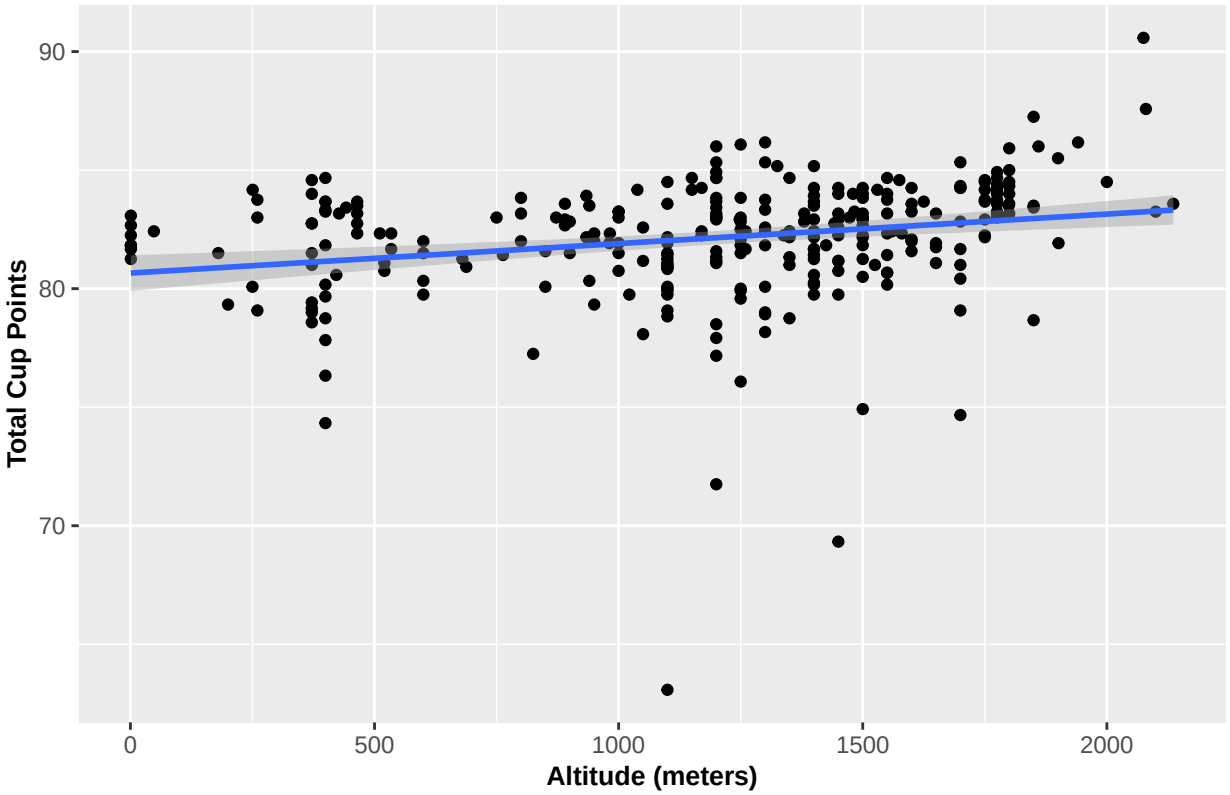


2.3 Coffee Bean Makeup and Cultivation Characteristics

The altitude variable is our main right hand side (RHS) variable. Altitude is measured in both meters and feet in the initial dataset, and we converted the feet measures into meters to standardize. It has a mean of 1,252 meters and standard deviation of 538 meters. As shown in Figure 2, there are two local maxima in the pdf of altitude: one around 450m and the other around 1300m. This is likely because some bean cultivators grow cheaper beans at lower altitude for traditional blends and others focus on higher quality beans with more intense flavors (Stopsack, 2017). The initial dataset had a few records of beans with altitudes over ~3,000 meters (up to 6,000m), which is significantly higher than the majority of the distribution. These were filtered out as they were outliers and less than 1% of the sample.

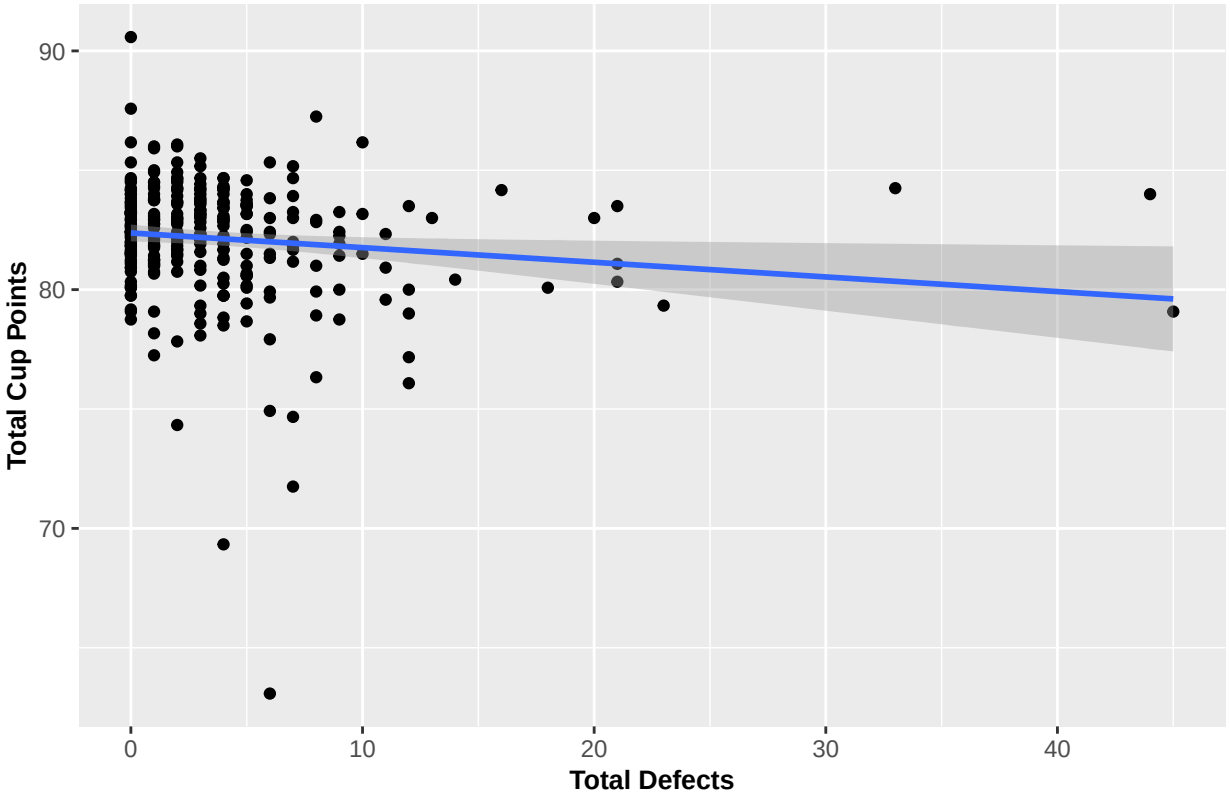
Preliminary data exploration (on a holdout set) shows a positive relationship between the altitude measure and the total cup score, as shown in Figure 3. The relationship is linear with small residual bounds across the range of altitude values. Intuitively, it seems likely that as altitude reaches extreme levels that bean production would not thrive given the cold conditions and difficulty of farming. Given this, we can expect that the altitude has a positive effect on bean quality up until a certain point, after which it decreases. It seems that the farmers of beans in this dataset have discovered this maximum altitude and largely avoided it, likely due to the cost of operating at higher altitudes. To account for this potential non-linear relationship, we will create an altitude-squared feature to include in one of our linear models. If our models validate that altitude has a causal effect on the total cup score, we can recommend that coffee shops source beans from the farms that are in the ideal altitude range. Additionally, we can recommend that farms search for land at higher altitudes when planting the beans.

Figure 3: Relationship Between Altitude and Total Cup Points



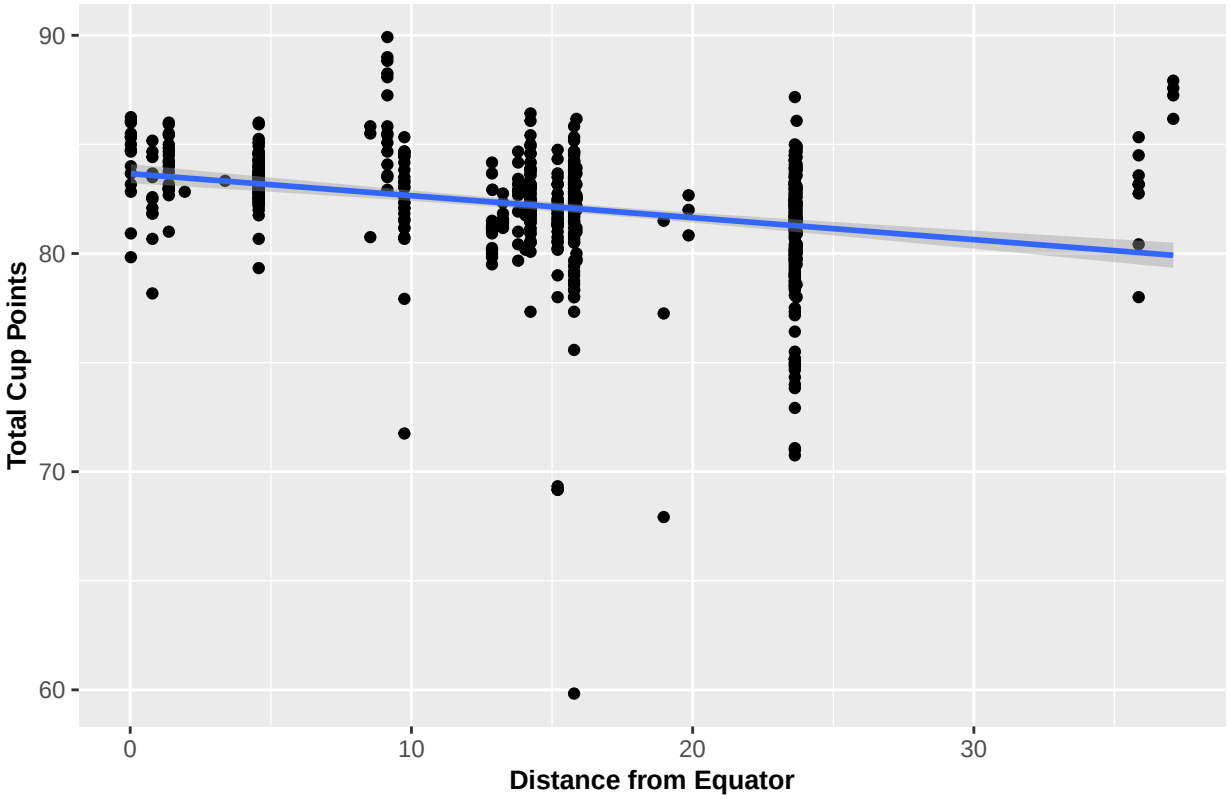
Our models will include additional covariates to provide more detailed insights into the many attributes that make up a quality coffee bean. If we encounter other variables that can also generate a causal effect on the total cup score we will be able to provide a better and more complete recommendation to local coffee shops. In our analysis, we also explored the relationship between the total defects variable and total cup points. The dataset contains counts of Category One and Category Two defects per bean, which are summed to derive the total defect. Category One defects occur when beans are full black or sour, and Category Two defects occur when beans are either chipped, partially black or sour, have insect damage, or have water damage. Figure 4 below shows a negative linear relationship between the number of defects and total cup points, which is what we expected based on intuition alone. A deeper analysis showed that the sum of defects had a stronger relationship with total cup score than either of the two defect categories alone.

Figure 4: Relationship between Total Defects & Total Cup Points



Another predictor of interest is the distance of bean cultivation from the equator. Research suggests that the best region of bean cultivation (called the "Bean Belt") is near the equator, between the tropic of Cancer and the tropic of Capricorn. The feature we used to account for this is the distance from the equator of the Country of cultivation. Our data shows a negative linear relationship between the distance from the equator and total cup points.

Figure 5: Relationship between Distance from Equator and Total Cup Point



In addition to the metric variables discussed above, our analysis explored the use of the bean variety and processing method categories as controls in our linear models.

- Processing Method: Refers to the way in which a seed is removed from a coffee cherry after harvesting. Each method affects the body, sweetness, and acidity of the brewed coffee. The classes in this variable include: Natural/Dry, Washed/Wet, Semi-Washed, Honey, and Other processing methods. Figure X shows the distribution of processing methods:

Processing.Method	n
Pulped natural / honey	9
Other	25
Semi-washed / Semi-pulped	52
Natural / Dry	176
Washed / Wet	723

- Variety: Within the Arabica species, there are several varieties of coffee beans. The dataset includes 28 total varieties; however, we decided to select the top 5 categories (which account for 74% of the data) and assign each other variety to the "Other" category. The top 5 varieties are Caturra, Typica, Bourbon, Catuai, and Yellow Bourbon. Figure Z shows the distribution of bean varieties:

Variety	n
Arusha	1
Blue Mountain	1
Ethiopian Heirlooms	1
Marigojipe	1
Pache Comun	1

Variety	n
Peaberry	1
Sulawesi	1
Sumatra Lintong	1
Ethiopian Yirgacheffe	2
Java	2
Mandheling	2
Ruiru 11	2
Sumatra	2
Pacamara	7
SL34	7
Gesha	12
Pacas	13
SL28	13
SL14	16
Catimor	20
Mundo Novo	25
	26
Yellow Bourbon	30
Catuai	64
Other	96
Bourbon	201
Typica	206
Caturra	231

Regarding transformations, the altitude feature was divided by 100 to better interpret coefficients in 100m increments, rather than 1m increments. The categorical variables were converted into dummy variables for use in modeling. For the Processing Method dummy features, the "Natural/Dry" method was held out. For the Variety dummy features, the "Other" variety type was held out.

2.4 Research Design

In this analysis, we seek to understand the relationship between the altitude of coffee bean cultivation and the quality of the bean (total cup score) to answer our research question "How does the altitude of the bean cultivation affect the bean's total cup score?".

In this study, we will focus on the altitude variables as the main dependent variable of interest. We will conduct additional research to understand the effect that other variables related to bean cultivation and harvesting have on the resulting quality score. The data available is large enough to employ the large-sample linear model. The models in our analysis will use the total cup score (a measure of coffee bean quality) as the one outcome variable, and several other explanatory variables to identify the key determinants of bean quality. The total cup score variable is approximately normally distributed; however normality of variables or residuals are not required as only the large-sample model assumptions must be met in this study.

3 Modeling

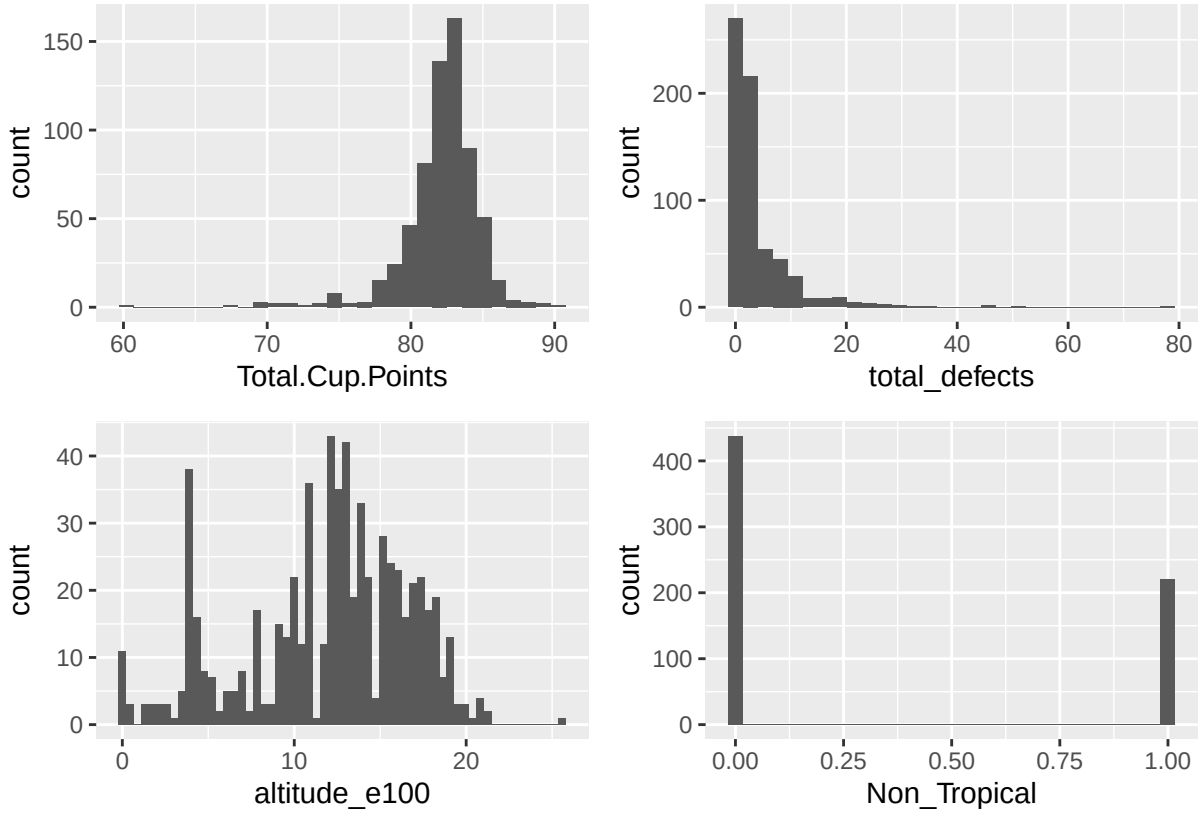
As mentioned from the previous section of exploratory data analysis, we found a positive relationship between the altitude measure and the total cup score. Therefore, altitude is being chosen as the key variable in our model. On top of the altitude, the climate and environment which coffee beans are growing in is considered as an additional factor which can impact the quality. Since coffee plants are well grown between the tropic of Cancer and the tropic of Capricorn, usually termed the bean belt or the coffee belt, we will mainly investigate the climate difference between tropical and non-tropical regions. In this study, the latitude information of each country is integrated from an extra data set which allows us to define the region of each country.

Beyond the two geographical factors mentioned above, we are considering to include the properties and processes that are associated with the beans as control variables in our model:

- Processing Method
- Variety
- Total Defects

While conducting the analysis and modeling, we found that the Total score increment described by the effect size will be too small if the variable unit increases by every meter. Thus, `altitude_e100` column is generated from the original altitude information divided by 100.

The analysis seeks to understand the causal impact of altitude on bean quality with a set of linear models: one model with a single explanatory variable (altitude), and several other model specifications that introduce additional covariates.



3.1 Base Model

In the base model, only the key explanatory variable altitude_e100 is included:

$$Total.Cup.Points = \beta_0 + \beta_1 \cdot altitude\_e100$$

In the base model with no covariates, the altitude variable is shown to be statistically significant, thus validating our hypothesis. The model's coefficient of determination(R^2) is 0.053, indicating that 5.3% of the variance in the total cup point is being described by the base model. The effect size of altitude per 100 meters is 0.128, indicating that a 100-meter increase of the altitude will increase the total cup score by 0.128 points.

3.2 Second Model

We introduces the quadratic term of altitude along with the linear term.

The second model assessed is thus as follows:

$$Total.Cup.Points = \beta_0 + \beta_1 \cdot altitude\_e100 + \beta_2 \cdot (altitude\_e100)^2$$

Model1 vs Model2 F test p value : $1.1752389 \times 10^{-10}$

3.3 Third Model

In the third model, we include a distance from the equator variable in the second model to explore another geographical element in the total cup point evaluation.

$$Total.Cup.Points = \beta_0 + \beta_1 \cdot altitude\_e100 + \beta_2 \cdot (altitude\_e100)^2 + \beta_3 \cdot dist\_equator$$

Model2 vs Model3 F test p value : 4.492368×10^{-6}

3.4 Fourth Model

In the fourth model, additional control variables are accounted for aspects of the coffee bean's properties and processing methods.
The third model assessed is thus as follows:

$$Total.Cup.Points = \beta_0 + \beta_1 \cdot altitude\_e100 + \beta_2 \cdot altitude\_e100^2$$
$$+\beta_3 \cdot dist\_equator + \beta_4 \cdot variety + \beta_5 \cdot Processing.Method + \beta_6 \cdot Total.Defect$$

Model3 vs Model4 F test p value : $4.2022955 \times 10^{-11}$

4 Results

The base model indicates an effect size of 0.128 with a coefficient of determination of 0.053. The effect size indicates that a 100-meter increase of the altitude increases the total cup score by 0.128 units.

The second model introduces the quadratic term of altitude along with the linear term. The effect size of altitude per 100 meters is -0.337, indicating that a 100-meter increase of the altitude will decrease the total cup score by 0.337 units. The effect size of squared altitude per 100 meter is 0.022, which means that the rate of change of total cup score increases by 0.044 for each 100-meter increase of the altitude. If we differentiate the model equation by altitude_e100, the change in total cup points per 100 meter change in altitude follows this equation:

$$d(Total.Cup.Points) = -0.337 + 2 \times 0.022 \times altitude\_e100$$

This means that at an altitude of 0, a 100-unit increase in altitude will decrease the total cup score by 0.337 points. However the rate of change of altitude is positive and increases at a rate of 0.044 for every additional 100 meters of altitude. Therefore, at an altitude of 1,000 meters (10 units of altitude_e100), a 100 meter increase in the altitude will increase the total cup score by 0.103 units ($2*0.022*10$ - 0.230). This means that the total cup score is minimized at an altitude of 766 meters, after which the total cup score increases for each increase in altitude.

The third model includes the distance from the equator variable, which is expected to explain meaningful aspects of the data. The effect size of the new variable is -0.068 with a coefficient of determination of 0.131. It indicates that a 1 degree increase in the latitude will decrease the total cup score by 0.068 units. With a F test p value of 4.492368×10^{-6}, the third model with additional distance from the equator variable fits better to the data than the base model.

Table 4:

	Dependent variable:			
	Total.Cup.Points			
	(1)	(2)	(3)	(4)
altitude_e100	0.128^{***}	-0.337^{***}	-0.250^{***}	-0.230^{***}
	(0.022)	(0.074)	(0.077)	(0.079)
I(altitude_e100^2)		0.022^{***}	0.016^{***}	0.016^{***}
		(0.003)	(0.004)	(0.004)
dist_equator			-0.068^{***}	-0.033^{**}
			(0.015)	(0.014)
total_defects				-0.105^{***}
				(0.019)
variety_adj_Bourbon				-0.182
				(0.261)
variety_adj_Catuai				-1.573^{**}
				(0.737)
variety_adj_Caturra				-0.445^{**}
				(0.210)
variety_adj_Typica				-0.890^{***}
				(0.309)
variety_adj_Yellow.Bourbon				-1.106^{***}
				(0.343)
Processing.Method_Washed...Wet				-0.967^{***}
				(0.229)
Processing.Method_Semi.washed...Semi.pulped				0.337
				(0.329)
Processing.Method_Pulped.natural...honey				-0.300
				(0.621)
Processing.Method_Other				-0.869^{*}
				(0.518)
Constant	80.563^{***}	82.485^{***}	83.421^{***}	84.315^{***}
	(0.298)	(0.407)	(0.438)	(0.513)
Observations	659	659	659	659
R^2	0.053	0.102	0.135	0.245
Adjusted R^2	0.051	0.099	0.131	0.229
Residual Std. Error	2.636 (df = 657)	2.568 (df = 656)	2.523 (df = 655)	2.375 (df = 645)
F Statistic	36.489^{***} (df = 1; 657)	37.122^{***} (df = 2; 656)	33.995^{***} (df = 3; 655)	16.063^{***} (df = 13; 645)
Note:				*p<0.1; **p<0.05; ***p<0.01

The fourth model, which includes the explanatory variables from the third model while also adding in control variables, shows an increase in explanatory power (Adjusted R^2 0.229 vs. 0.131). The addition of these control variables results in a reduction of effect size on an absolute value basis across all of the explanatory variables. For the variety variable, it shows that Catuai, Caturra, Typica, and Yellow Bourbon would have a statistically significant effect on the total cup score with the effect size of -1.573,-0.445,-0.890, and -1.106 respectively. Also, the fourth model indicates an effect size of -0.967 and -0.869 for the wet wash and other processing method respectively, which means that the wet wash method results in a 0.931 lower cup score compared to the dry processing method, and the other wash method results in a 0.869 lower cup score compared to the dry processing method. Again, with a F-test p value of $4.2022955 \times 10^{-11}$, the fourth model which includes these control variables should be considered as the most representative one for our data among all four models.

The final model selected is thus as follows:

$$Total.Cup.Points = \beta_0 + \beta_1 \cdot altitude\_e100 + \beta_2 \cdot altitude\_e100^2$$
$$+\beta_3 \cdot dist\_equator + \beta_4 \cdot variety + \beta_5 \cdot Processing.Method + \beta_6 \cdot Total.Defect$$

The total cup score of coffee beans is impacted by its growing altitude. At the beginning of 0 meter altitude, an increase in the altitude by 100 meter decreases the total cup score by 0.281 units, and is highly statistically significant. However, the rate of change increases by 0.044 for every 100 meters of altitude increment. The score will hit the bottom at about 766 meter altitude, after which, the score will go up again. As a guidance for coffee investors, without considering altitude above 3000m, it seems better to grow coffee either on a lower altitude area close to 0 meter or on a higher altitude area close to 3000 meter.

5 Limitations

5.1 Testing OLS model Assumptions

As we have completed the model analysis, we want to evaluate the assumptions used on the regression model. As a primary factor, we need to evaluate the amount of data that we have for our regression model. The assumptions for a model with a big amount of data are less restrictive than small ones used as we can rely on asymptotic properties like the Central Limit Theorem giving as statistics guarantees. This analysis was conducted using a dataset of ~1,000 records after cleaning the data, 30% of which were used for an exploratory data analysis and 70% used for modeling. Given this large sample size, we will need to evaluate that the data points are I.I.D ,the BLP exists and is unique.

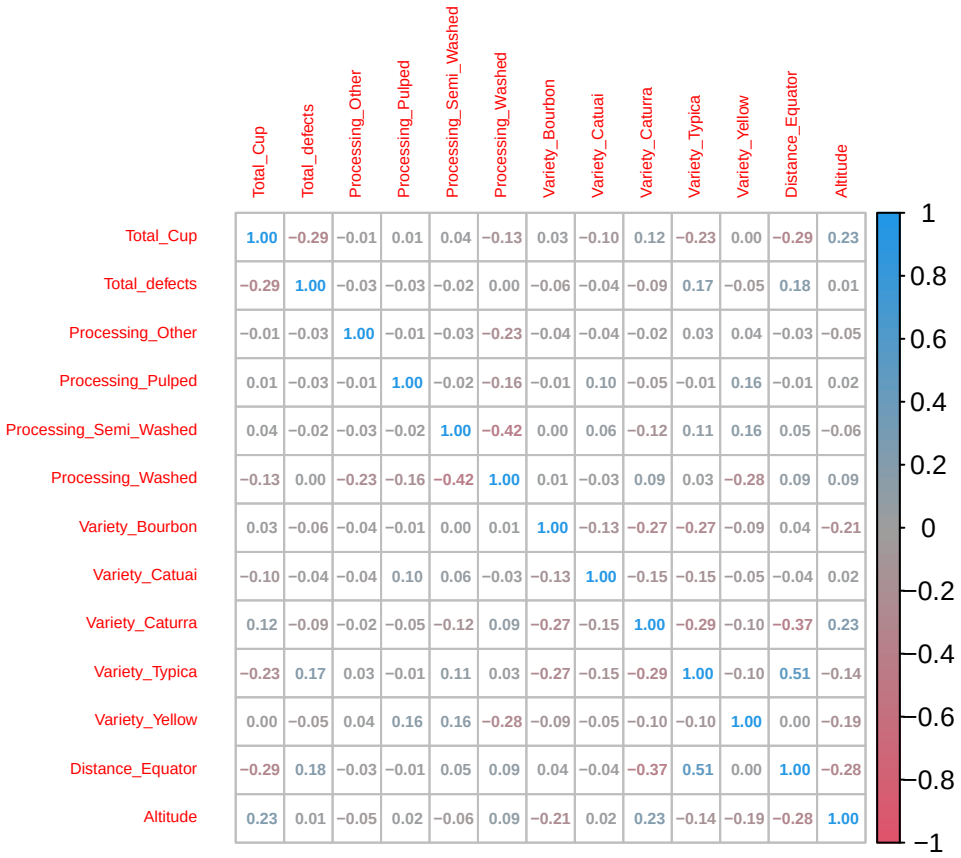
5.1.1 Assumption 1: Independent and Identically Distributed (I.I.D)

Regarding the independence assumption, there are a few ways in which this assumption may be infringed. Given the beans come from certain regions, mostly close to the Earth's equator, there is likely an influence of geographical clustering in the dataset. Additionally, some beans are planted by the same farm organization or owner, or sourced by the same supplier. Finally, there may be instances of competition in which one farmer mimics the decisions of another in planting beans. Thus, we can not assume there is perfect independence in the sample. Regarding the identically distributed assumption, the beans all belong to the same population distribution of total cup score. The Coffee Quality Institute (CQI) determines the total cup score of each bean through a rigorous, and consistent process to prevent bias from affecting the total cup scores. Given that each bean comes from the same population and the CQI ensures that ratings are as objective as possible, the identically distributed assumption is met.

5.1.2 Assumption 2: A Unique BLP Exists

Our analysis shows that the covariates used in our linear model are not perfectly collinear. There is likely some level of collinearity in our model with the full set of covariates given that coefficient values decreased compared to the simpler models (Model 1 to Model 3). Additionally, the matrix below shows some variables

are correlated with each other. However, there are no coefficients that are set to zero in the model fitting process, which means that there is no perfect collinearity. We can thus infer that a unique best linear predictor exists.



5.2 Structural Limitations

Up to this point we have evaluated our model based on the assumption that the causal theory of our model is correct - an increase in altitude causes an increase in total cup score. We are assuming that the altitude of the farm, the processing method to extract the bean, the coffee variety, the distance of the farm from the equator, and the total defects of the beans has a direct causal relation on the quality of the bean, and that there are no other significant variable that affect the total cup score. Additionally, we assume that there is not an inverse relationship between the coffee quality and the variables selected on this model. To assess these assumptions, we evaluate potential omitted variables and the effect each may have on the estimates if our causal theory is incorrect.

- Temperature: Extreme temperatures (low and high) can affect the harvest of any plant. For coffee beans, mild, consistent temperatures are generally preferred. This could explain in part why most coffee beans are produced in the warm regions near the Earth's equator. Temperature is likely to be correlated with the distance from the equator variable in our models as tropical areas are warmer than the non-tropical areas. Temperature also has a relationship with the quality of the coffee harvested. Research shows that beans grow better in mild temperatures than in hot or cool temperatures, which means that there is a non-linear relationship between temperature and bean rating. Additionally, changes in temperature due to global warming likely have an effect on the quality of the coffee bean. The omitted variable temperature has multiple effects on the coefficient and affects multiple variables simultaneously, therefore, we can't accurately estimate its effect on the coefficients.

- Soil: The health of a coffee tree depends on the quality of the nutrients that are provided in the soil. Ideally, our dataset would include an expert rating on the quality of the soil, which would likely explain

some additional variance in the total cup score distribution. The quality of soil is likely to be negatively correlated with the number of defects variable in our models. The omission of this variable is pushing the coefficient of total defects to be higher than its true value, thus the bias is positive and away from zero.

- Water: The right amount of watering is another essential variable in the development of the coffee bean. Similar to temperature, there is a sweet spot for each plant; the absence of water can cause the coffee bean not to develop to a correct size and flavor, and too much water can end up killing the tree roots. This variable does not have a linear relationship with total cup score and therefore its effect is very complex to describe, therefore, we can't accurately estimate its effect on the coefficients.
- Sunshine: Traditionally, quality Arabica coffee beans are grown under a canopy of trees to generate shade (Byloos, 2017). Farmers of Arabica beans want to provide shade throughout the day, and only allow trace amounts of sunshine to reach the beans. The omitted variable sunshine is thus negatively correlated with coffee quality but doesn't have a direct relation to the independent variables. There is no bias on the covariate coefficients but, the lack of this variable of the model reduces the power of the model to explain the variance in the total cup point variable.

After evaluating the potential omitted values, we believe that there are some important variables omitted in our models, but the omitted variable bias is not significant enough to discredit the results of the model. These omitted variables need to be taken into consideration in future research in order to improve the model and its results.

6 Conclusion

The goal of this study was to understand the relationship between altitude and coffee bean ratings. We primarily used a dataset produced by the Coffee Quality Institute, where each record includes the total cup score of the bean, as well as descriptive data points on the type of bean, the cultivation process, and the location of cultivation. Coffee bean quality rating was defined by the total cup score, which is a criteria for qualified coffee reviewers to measure. We used the altitude of the bean farm, and introduced several other metric and categorical variables that capture information about the bean makeup and cultivation methods.

Based on the results of this analysis we conclude that the altitude where the beans are being farmed has a significant effect on the quality of the coffee. In the base model, the total cup score increases by 0.128 for each 100 meters of increment of altitude. Though the effect is practically small, an increase by just a few total cup points could make a large difference in the competitive advantage of a coffee distributor. The findings of this study should be of interest to coffee shops, who need to source quality beans from the right location, and coffee farmers, who need to make real estate and cultivation decisions to grow quality beans. As the coffee industry continues to expand, it is of primary interest for these stakeholders to improve the quality of their products and expand clientele along with the market.