Proposal

John-Tianyu-Fabian

2022-04-11

Notes:

Introduction

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.

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.

Data and Methodology

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.

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

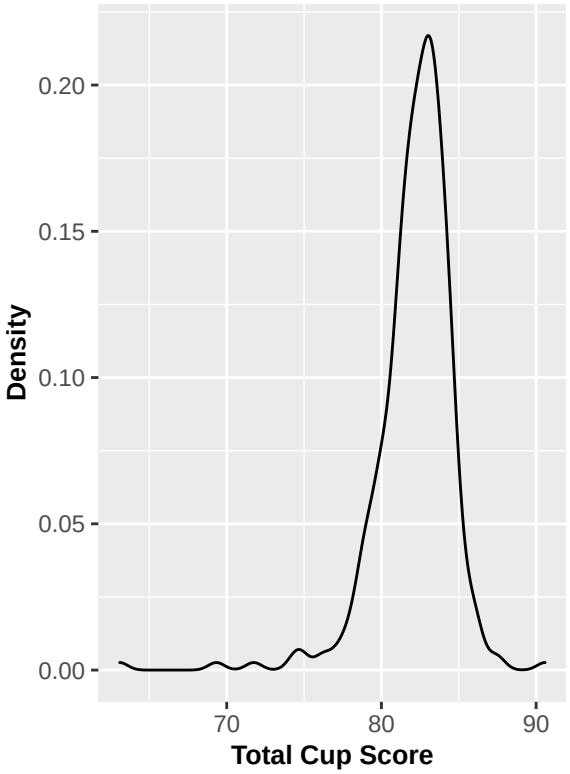
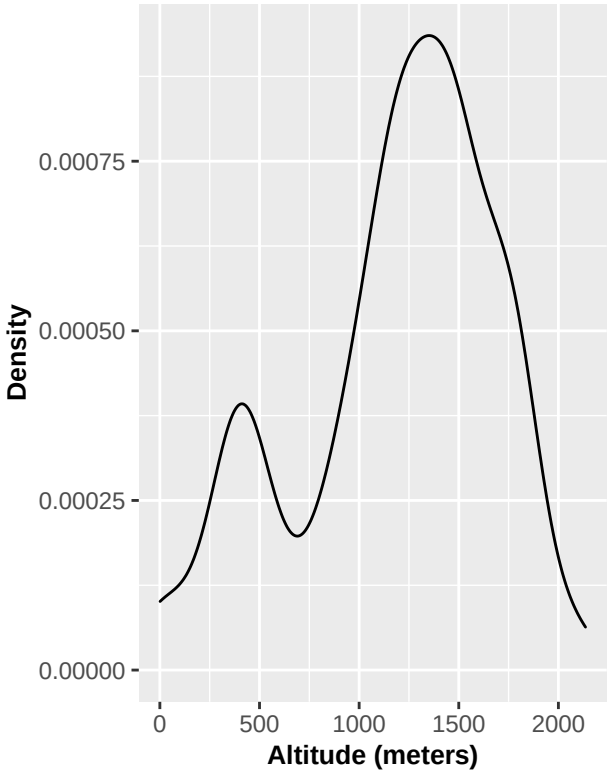


Figure 2: Distribution of Altitude

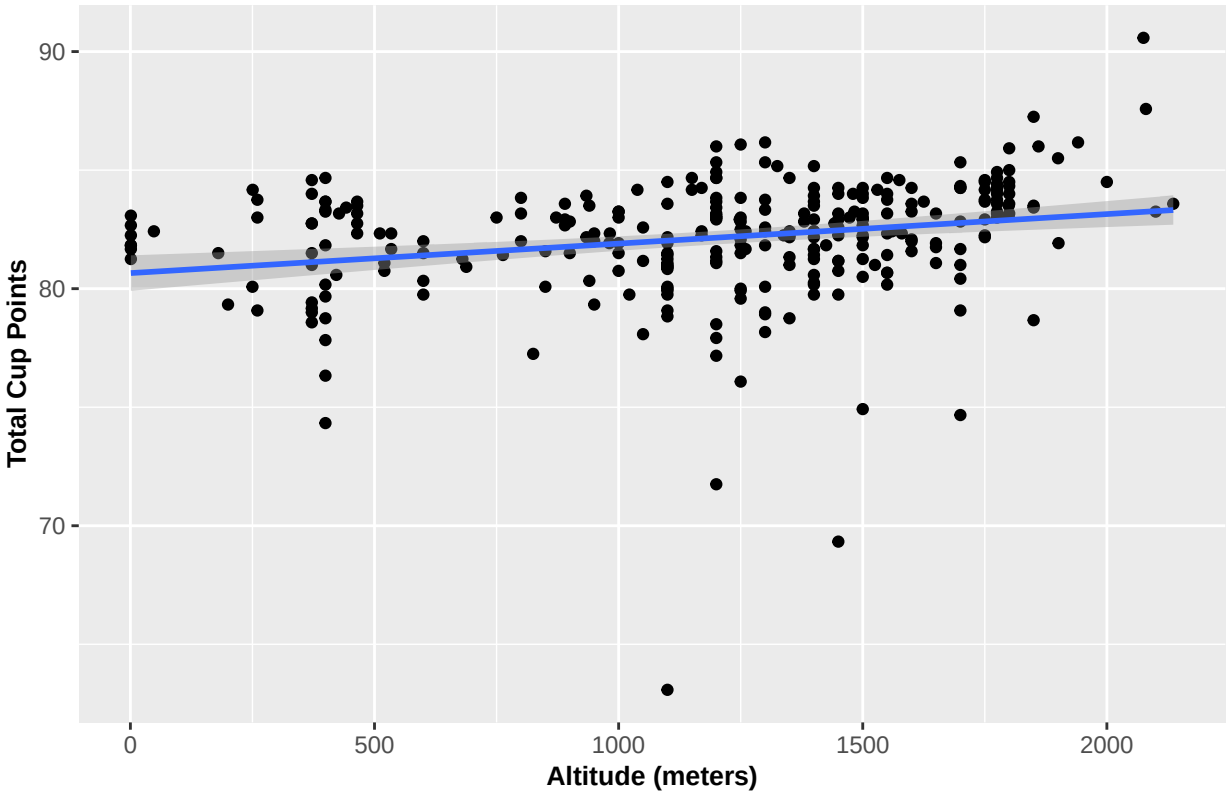


Coffee Bean Makeup and Cultivation Characteristics

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.

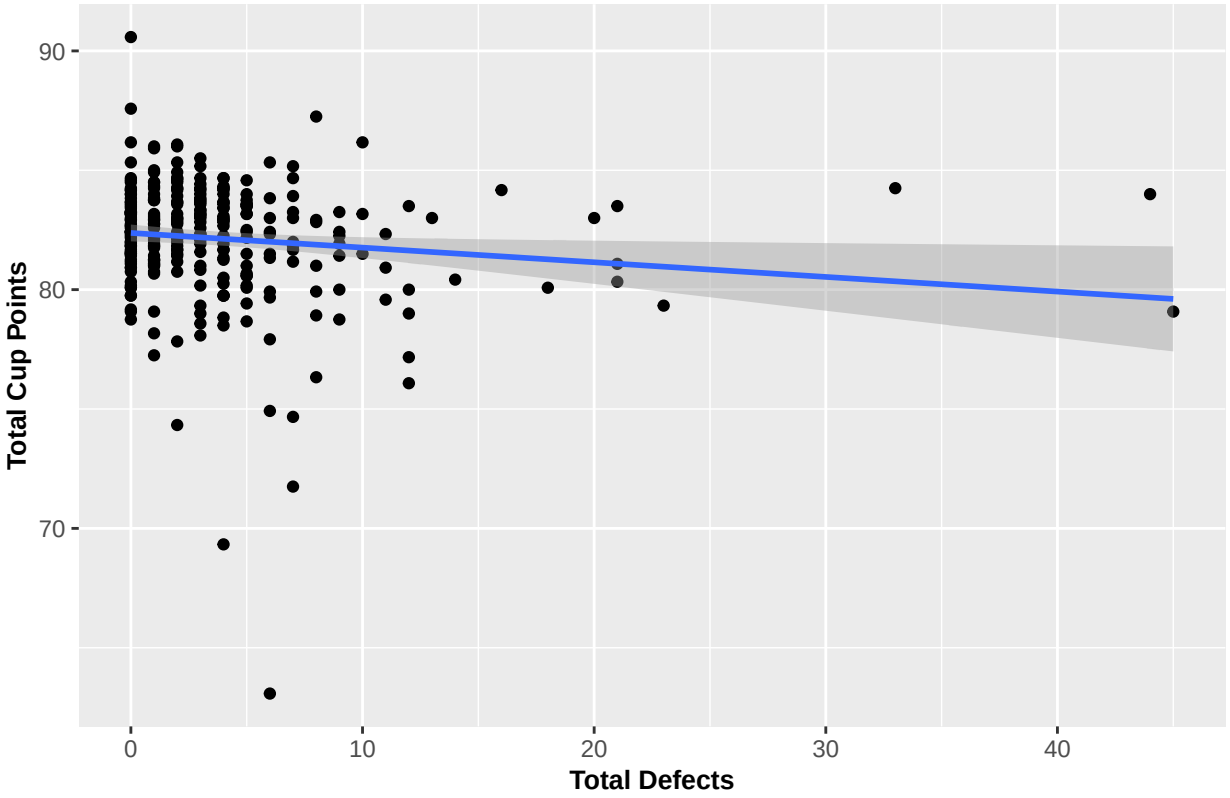
Preliminary data exploration (on a holdout set) shows a positive relationship between the altitude measure and the total cup score, as shown in Figure 3. The relationship is linear with small residual bounds across the range of altitude values. Intuitively, it seems likely that as altitude reaches extreme 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



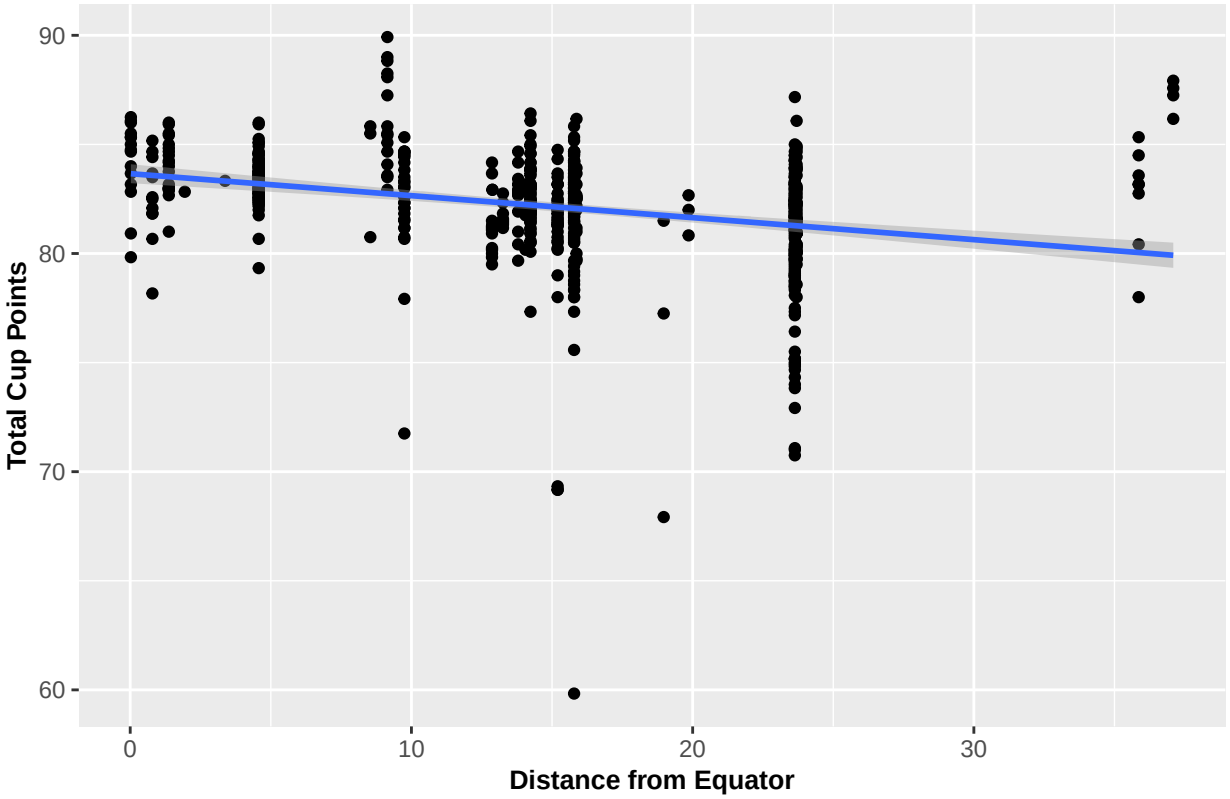
Our models will include additional covariates to provide more detailed insights into the many attributes that make up a quality coffee bean. If we encounter other variables that can also generate a causal effect on the total cup score we will be able to provide a better and more complete recommendation 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:
- 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:

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.

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.

#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.

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.

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 : NA, 3.5601354×10^{-9}

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 : NA, 7.284776×10^{-7}

Appendix