Estimating Coffee Ratings - Dataset from Coffee Quality Institute

Abraham Yang, Erica Nakabayashi, Melissa Olivera

2023-04-21

Introduction

The global specialty coffee market is expected to reach USD 71 billion by 2028 and accounts for about 20% of coffee exports worldwide. Specialty coffee is graded on a 100-point scale by a certified coffee taster (SCAA) or by a licensed Q Grader(CQI) where a score of 80 is the minimum qualification for a specialty coffee classification, 80-84.99 is considered "very good", 85-89.99 is "excellent", and 90-100 is "outstanding". Many factors impact the final taste of coffee, and these include the variety of the plant, the soil, the weather, the altitude where the coffee plant is grown, the processing of the coffee beans, and brewing practices.

In this report, we attempt to answer: "how do geographical features such as altitude, climate, and regional location affect coffee cup scores?" Altitude refers to the location of the farm growing the coffee plant, climate refers to whether the farm is in a fully tropical, partially tropical, or non-tropical climate.

Region refers to continental sub-regions such as "Central America," "South Central Asia." Cup score ranges from 0 to 100, with 80 being the minimum score for specialty coffee rating.

The goal of our project is to inform anyone intending to secure real estate and pursue specialty coffee market. Therefore, factors impacting the cultivation of specialty coffee unrelated to geography were lumped into our epsilon of unmeasured factors.

Description of the Data

The primary dataset we used was gathered from Kaggle, originally from Coffee Quality Database from the Coffee Quality Institute (CQI), courtesy of Buzzfeed data scientist James LeDoux.CQI is a worldwide, non-profit organization involved in coffee quality.

It created the Q Programme which is a system engaging both producers and buyers by certifying coffee grade, and offering opportunities for producers to access premium prices, and promote competitiveness.

This dataset includes the producers growing the coffee plants, the species of coffee, dates of harvest, beans processing methods, coffee beans qualities and defects, CQI grades, certification body, and altitude.

Each row of data represents one producer submission for grading. We also obtained a supplementary dataset from World Population Review to categorize countries as tropical, partially tropical, and non-tropical. Both datasets are observational.

Key Concepts are Operationalized

From our datasets, we picked as our key variables: "Altitude", "Subregion", and "Tropical". We performed independence analysis using a single-outcome causal graph. We determined that the causal pathways were strictly one-way from each key variable to the outcome "Total Cup Score". We also determined that there is no common ancestors between each key variable to the outcome. Alternative variables impacting "Total Cup Score" were put in our Epsilon. They included "Country" and "Bean Defects".

Explanation of Key Modeling Decisions

CQI data set presented a few relevant inconsistencies, specially on altitude our main interest, as the data appear to be manual. Incomplete and inconsistent observations were excluded. In order to assess validity, our team cross checked if the altitude data was reasonable within country and world level elevation measurements. After wrangling, over 300 observations were dropped, and the total sampling size is composed of 1034 different scores. Those were divided into train and test set using a stratification method in order to assure representatives of each sub region to the training set.
The main interest of this research is to understand how in the observational gathered data altitude, geographical sub region and climate (being tropical, not tropical or partially tropical). Because of that, variables not related to these topics, or that showed common ancestors (such as country), were not included on the modeling section.
A glimpse into altitude and the CQI scores showed a positive correlation of 0.13. Some specific regions are known for being coffee producers, as well as being from a warmer climate is also excepcted to be positive correlated with score, since it's a better growing condition for the coffee plant.

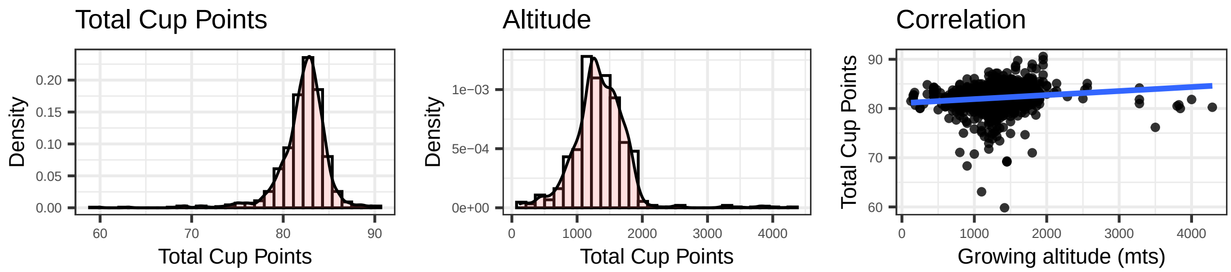

Figure 1: Histogram and Correlation on Cup Points and Altitude

As the altitude variable has an skewed distribution (skewness = 1.34), log transformation was considered to our analysis in order to reduce skewness on altitude and error term. This transformation was discarded mainly due to not achieving the expected effect, increasing model complexity, and reducing explain ability. For the last model, an interaction term was also added, and later on discarded due to perfect collinearity in some of the terms and lack of significance of the remaining variables.

Regression Models

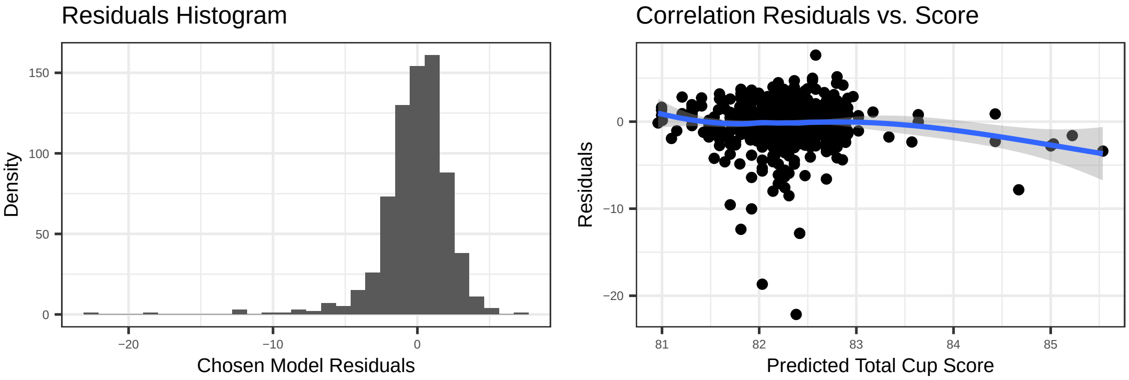

Figure 2: Residual Analysis - Model (5)

Model number (5) was selected as Top Performing, even though as shown on the regression table below, it has a slightly lower R-squared, because it both has more significant variables and is easier to interpret.

That said, as pictured on figure 2, model (5) still holds evidence for a non-normal, heavy tailed, and skewed errors as shown on the plots below. Shapiro Wilk and Jarque Bera tests also provided evidence in line by not rejecting the $H0$(Shapiro = c(W = 0.83792416333261), 1.63724294137714e-26, Shapiro-Wilk normality test, test$resids5, Jarque = c(`X - squared` = 8680.34929428485), c(df = 2), 0, JarqueBeraTest, test$resids5).

	Dependent variable:					
	Total Cup Scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Altitude	0.001^{***}		0.001^{***}	0.001^{*}	0.001^{*}	0.001^{**}
	(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0004)
log(Altitude)		1.160^{***}				
		(0.425)				
Eastern Africa Sub-Saharan Africa			1.895^{***}		1.598^{***}	1.414^{***}
			(0.432)		(0.440)	(0.449)
Eastern Asia			1.218^{**}		1.688^{***}	1.959^{***}
			(0.546)		(0.565)	(0.582)
South America Latin America			1.498^{***}		1.495^{***}	0.957^{**}
			(0.339)		(0.335)	(0.432)
South-Eastern Asia			0.758		0.555	0.723
			(0.595)		(0.592)	(0.657)
Partially Tropical				-0.907^{***}	-0.914^{***}	-1.212^{***}
				(0.309)	(0.323)	(0.390)
E.Afr. x Partially Tropical						
E.Asia x Partially Tropical						
S.LAm x Partially Tropical						1.376^{*}
						(0.699)
SE.Asia x Partially Tropical						-1.439
						(1.581)
Constant	80.822^{***}	73.998^{***}	80.266^{***}	81.702^{***}	80.986^{***}	80.716^{***}
	(0.517)	(3.048)	(0.536)	(0.593)	(0.588)	(0.612)
Observations	309	309	309	309	309	309
R^2	0.028	0.024	0.122	0.055	0.145	0.159
Adjusted R^2	0.025	0.021	0.108	0.049	0.128	0.136
Residual Std. Error	2.496	2.502	2.388	2.466	2.361	2.350
F Statistic	8.944^{***}	7.453^{***}	8.430^{***}	8.878^{***}	8.523^{***}	7.084^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01		

Discussion of Results

As for Statistical significance, testing our chosen model, an evaluation in terms of correlation accuracy (corr = 0.3465156) did not show strong predictability, and also, analyzing predictors and predicted confusion matrix, we could not recommend the use of our findings with statistical certainty.In terms of Practical Significance, through our regression analysis, we discovered that the most important location features to consider when working to improve cup points are subregion and tropical category. If a coffee is from Eastern Africa, Eastern Asia, or South America, it is likely associated with a higher cup score of 1.41, 1.95 and 0.96 respectively. If a coffee is from a particular tropical subregion, this is likely associated with a lower cup score of 1.21 points. This information can be used to help new coffee producers decide where they would like to produce their specialty coffees. Given the low R-squared, however, we would encourage further research before making a location decision based on these results.

Limitations

To assess our model limitations, we considered both statistical and structural limitations. For the statistical limitations there are 2 large-sample statistical assumptions we evaluated. We first evaluated whether the data was IID. We concluded that this assumption was not met given that the data collected primarily focused on top coffees. >85% of the reviews had a total cup score >80 pts. We believe this is not a true representation of the population and geographical clustering is also a factor. The second statistical assumption we evaluated was unique BLP/no perfect collinearity. For this assumption, we evaluated the coefficients of each of our models and noticed that R did not drop any of our variables, indicating no perfect collinearity.
This assumption also includes the requirement that a unique BLP exists, however our distributions had heavy tails so this assumption was not met. For structural limitations, we identified Omitted Variable bias and Right Hand Side bias. For omitted variable bias, there were several variables we omitted that may bias our results such as moisture and variety due to too many missing values. These variables are positively correlated to total cup points, so we would expect a bias moving away from zero and thus making our hypothesis tests overconfident.
For RHS bias, we took this into consideration in our initial model development. One of the initial models we considered was using country of origin and altitude to predict total cup points. We soon realized that this would result in altitude as a RHS variable, as country can impact altitude and both altitude and country can impact total cup points. To overcome this, we instead used subregion in our models.

Conclusion

Tying this back to the original research question “How do geographical features such as altitude, climate and regional location affect coffee cup scores?”, we determined that our analysis was inconclusive and we are not confident that these features affect coffee cup scores. Our top performing model (using altitude, subregion and tropical to determine cup points) suggests that there is a positive relationship between cup points and altitude, a positive relationship between total cup points and subregions Eastern Africa, Eastern Asia, and South America, and a negative relationship between total cup points and being partially tropical.
However, even though this model had good statistical significance, we ultimately decided it is unreliable.
There are many other factors we did not consider in this OLS regression model largely due to missing or inconsistent data.
If we were to move forward with this research question, we would collect new data which accurately represents the coffee population and we would ensure data was complete and consistent across all reviewers.