«Study of the influence of the characteristics of Arabica coffee varieties on the cupping score»

Product Name:
«Study of the influence of the characteristics of Arabica coffee varieties on the cupping score»
Product Version: 1.0

Team name: "jvcr."
Team members: Buldin Dmitriy, Mityaev Alexei, Nikitin Maxim, Zamkovoy Ivan

Product Phase: Ready to use
Current Date: 2022-05-22

Table of Contents

TABLE OF CONTENTS

Abstract

Cupping is the practice of observing the tastes and aromas of brewed coffee. It is a professional practice but can be done informally by anyone or by professionals known as "**Q Graders**". A standard coffee cupping procedure involves deeply sniffing the coffee, then slurping the coffee from a spoon so it is aerated and spread across the tongue. The coffee taster attempts to measure aspects of the coffee's taste, specifically the body, mouthfeel, sweetness, acidity, flavour, and aftertaste.

This study will be devoted to the discovery of these characteristics.

Introduction

1. **Motivation**
   The impetus for the study was the desire to understand the processes of evaluating a good coffee, called specialty. Because specialty coffee shops often source their beans from local roasts, the difference between coffee shop flavors can vary greatly. Coffee differences are influenced by different characteristics that are determined by cupping, and this study was created to better understand this process.

2. **Brief Literature Review**
   Research on Arabica varieties has been carried out by various research groups and published in journals and websites of coffee service organizations or academia. The main source of information was the Coffee Quality Institute (CQI). Dataframes with

cupping data and general information about cupping processes were taken from the databases of this organization.

3. **Research Problem**
   This study is devoted to investigating the results of **cupping** Robusta and Arabica coffee beans. The work uses datasets processed from information obtained from the site of the organization "Coffee Quality Institute" (CQI). The main object of study of the results are the final points, from 0 to 100. If a lot of coffee gets more than 80 points, it receives the status of "specialty". In our study, this threshold is an additional marker in the study.

4. **Research Question**
   The main issue of the study is how much each of the characteristics separately and together affects the final points. Since the number of coffee positions is more than a thousand, and the variety of coffee, arabica, one, the differences in the results of the caping are the desired object of the study

5. **Hypotheses**
   The hypothesis lies in the fact that the effect of individual characteristics on the final points can differ greatly from the quality of coffee, and also differ in the degree of influence on the increase or decrease of the final score. Separately, the hypothesis of the study includes the assumption that more humid grain has higher qualities, since fryers of coffee often based on this characteristic when frying coffee.

6. **Main Results**
   The main results mostly correspond to the hypothesis. Most characteristics differ in the degree of influence on the formation of the final score, however, in different groups of grain quality, the influence of individual characteristics is weaker or stronger than the rest. The relationship was also discovered between opposite characteristics that make up the final assessment, for example, between flavor (sweet) and acidity, balance and aroma, clarified by conducting a correlation comparison. Also, the assumption about the dependence of the quality of the fried grain on its moisture was refuted - the grain was of the highest quality with a moisture content of 10-14%, but also with a near zero score of moisture.

Data and Methods

1. **Data sources**
   The data was obtained from Coffee Quality Institute (CQI) - a non-profit organization which operates internationally and helps coffee producers to improve the quality of

their products.
Link to original database:
https://github.com/jldbc/coffee-quality-database/tree/master/data

2. **Levels of analysis**
   The level of study is two-dimensional: first goes the state-level, which then includes different regions within these countries. Such analysis allows for a deeper comprehension of the importance of the geographic location not only around the world, but also around particular states.

3. **Scope**
   The size of the sample is more than 1200 lines. The number of data pieces is enough to observe common features and trend lines and to differentiate other characteristics within the largest country-producers of coffee. Beyond the scope of our analysis are small coffee producers which are located in states that are often not associated with the production of coffee beans.

4. **Time period of analysis**
   According to CQI, the data was being gathered for 8 years - from 2014 to 2022. This time frame is suitable for establishing the trend for either improving the quality of coffee or its declining.

Analysis

1. **Descriptive Statistics**
   Our analysis begins with the fact that we cite the main coffee producing countries. The table shows 32 countries, country numbers (No: of Countries) and Quality Score. The next item we made is a scutterplot that shows the number of coffee producer (country) and the overall Quality Score in it. According to the scutterplot, we see that, for example, a country that has a number more than two hundred, and such, based on the table, is only one and that is Mexico, has a quality score above 87. From the graph and the table, it can be concluded that the best producers of Arabica varieties are Brazil, Mexico, Colombia and Guatemala with 90.58, 87.25, 89.00 and 88.25 quality points, respectively.
   We can also say that there are no producers with a quality index of less than 86, and most countries produce coffee with a quality score of about 87 points. The next step was to set the indicator "color". The table below has shown that there are three variants of Arabica variety colors: Blue-Green, Bluish-Green, Green. The Blue-Green group has the highest quality soon (90.58), followed by Bluish-Green (89.92) and Green (89.75).

	color	No: of Countries	quality_score
0	Blue-Green	82	90.58
1	Bluish-Green	112	89.92
2	Green	849	89.75

The table provides information and indicators for the best Arabica varieties. The table first shows us the number of quality points and the name of the variety. After that, the table represents the country of growth and the name of the variety on the farm. After that, there are statistical indicators: aroma, flavor, acidity, body and balance. The extreme indicators in the table are the cup scores, humidity and grain color. Thus, we conclude that the largest indicators of total points are Arabica beans from Ethiopia with the farm name Metad PLC, they are also the owners of the largest total cups of points. When building the table, we used probability tables.

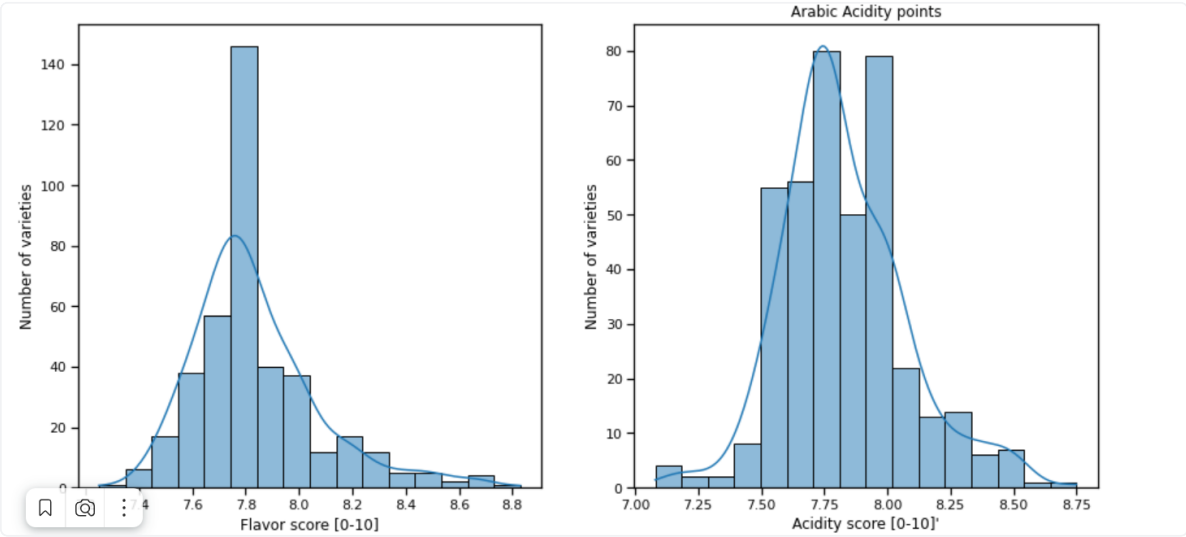
Out[62]:

	quality_score	species	country_of_origin	farm_name	altitude	region	harvest_year	aroma	flavor	aftertaste	acidity	body	balance	cupper_pc
0	90.58	Arabica	Ethiopia	METAD PLC	1950-2200	GUJI-HAMBELA/GOYO	2014	8.67	8.83	8.67	8.75	8.50	8.42	
1	89.92	Arabica	Ethiopia	METAD PLC	1950-2200	GUJI-HAMBELA/ALAKA	2014	8.75	8.67	8.50	8.58	8.42	8.42	
2	89.75	Arabica	Guatemala	San Marcos Barrancas "San Cristobal Cuch	1600 - 1800 m	NaN	NaN	8.42	8.50	8.42	8.42	8.33	8.42	
3	89.00	Arabica	Ethiopia	Yidnekachew Dabessa Coffee Plantation	1800-2200	Oromia	2014	8.17	8.58	8.42	8.42	8.50	8.25	
4	88.83	Arabica	Ethiopia	METAD PLC	1950-2200	GUJI-HAMBELA/BISHAN FUGU	2014	8.25	8.50	8.25	8.50	8.42	8.33	

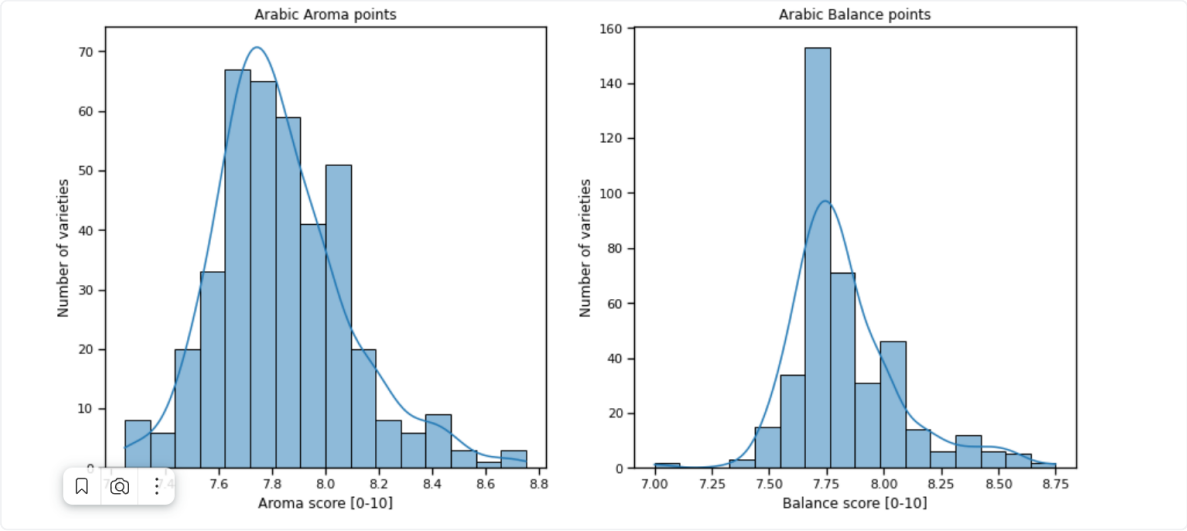
2. Exploratory Data Analysis

This map was created using the Folium library to display the coffee producing countries. We've tried our best, take a look at it.

Here are two linear graphs that represent the taste points for grain and acidity points. The first graph shows the taste points, which range from 0 to 10, also on the graph you can see the number of varieties that are shown on the vertical part of the graph. Their number reaches 400 species. After studying the line graph, we can conclude that the maximum number of varieties have about 7.8 taste points. The minimum number of points (7.3) has only one variety of grains, as well as the maximum number of points (8.8), has only one variety of grains. The second table shows us the number of acidity points. On the horizontal part we see glasses that range from 0 to 10, and on the vertical part the number of varieties of grains. The maximum number of varieties of seeds have 7.75 points, and the minimum number of seeds have 8.60 and 8.70 points. We can also add that in the second graph, the results range from 7.5 to 8, whereas in the first graph there is a dominant indicator - 7.8 To study these graphs, we used frequency analysis. Also, we used Subplot for visualisation.

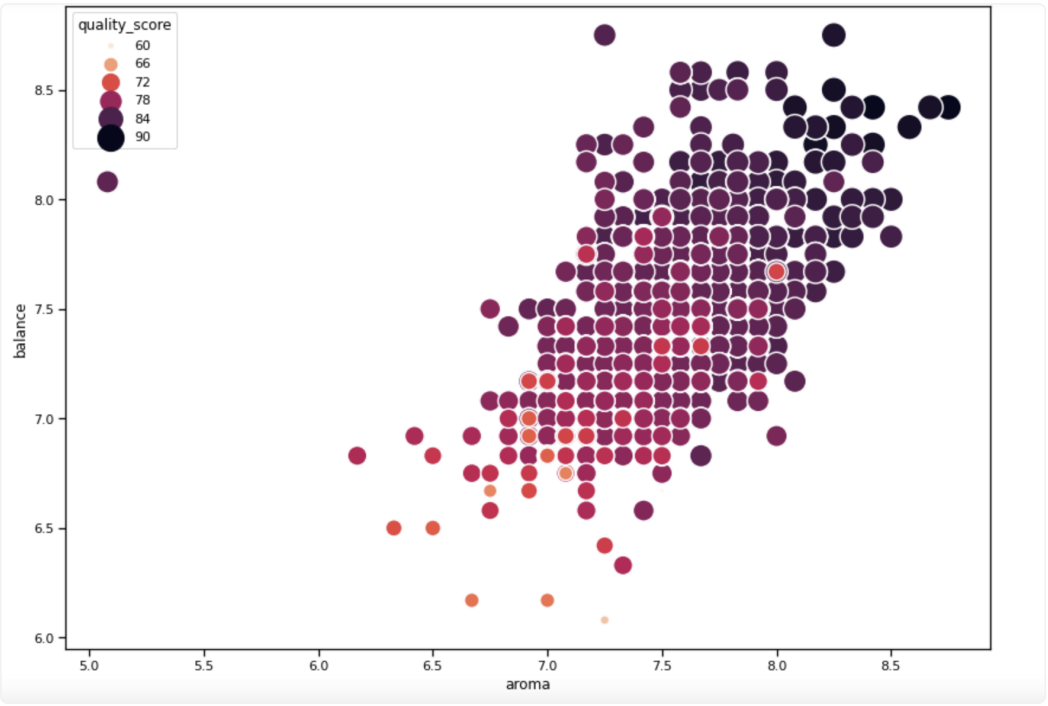


To consider this graph, we used frequency analysis, and to visualize a Subplot. On the left graph we can see aromatics points for different varieties of grains. The largest number of countries have about 67 points. The minimum number of points is approximately 8 countries, and the maximum is 4 countries. While the indicators on the chart can vary from 0 to 10. The right graph shows us the balance of speed for different types of coffee beans. We can understand that the points, as well as on the left graph, vary from 0 to 10. The maximum value of balance points is approximately 3 countries, and the minimum is 2 countries. The maximum number of points on the chart are 155 points. To describe the graphs, we used frequency analysis, and to visualize a Subplot. On the second chart, the results range from 7 to 8.75, whereas on the first chart, the indicator 7.75 dominates.

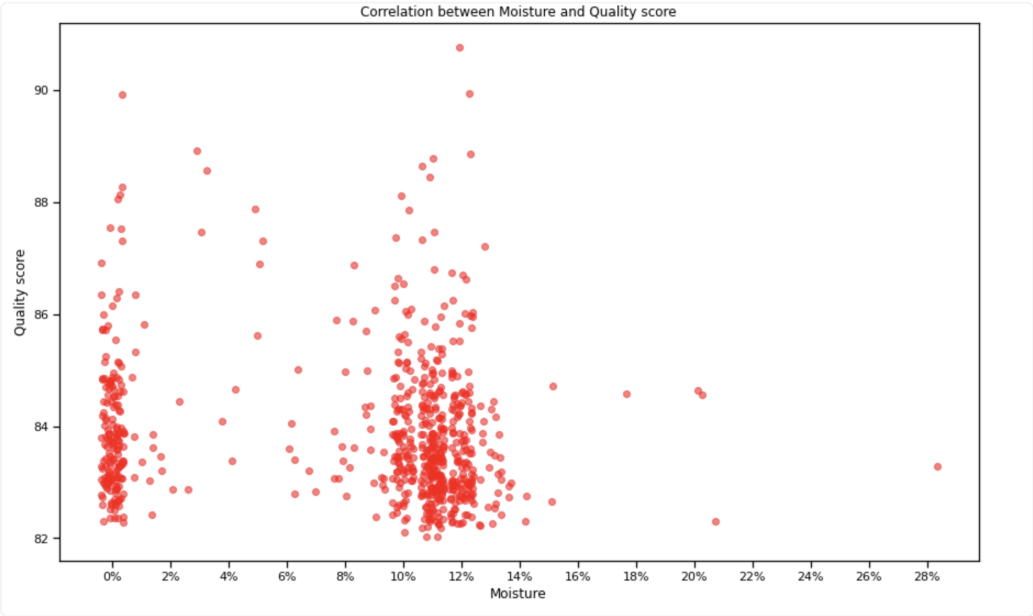


3. Advanced plots

This advanced scatterplot was created to combine the characteristics of balance and aroma in relation to the overall quality score. To fit three indicators in two planes, we created a gradient palette: the darker the point, the higher the speed quality of the corresponding characteristics. Moreover, the more quality speed, the larger the scatter dot. Thus, this chart combines a number of illustrative introductions for greater clarity.



This regplot allows us to evaluate the effect of grain moisture on its quality. Thanks to the jitter tool, we can observe the distribution of scores and visually verify the distribution along the ascending column.



4. The method of the data collection

The data collection method was provided to us by an aggregated set of cupping data. This result was provided by the cupping forms created by the Coffee Quality Institute and by the graders who filled in the table data. Moreover, the compilation process was carried out by members of the github community for free access.

Discussion

1. **Substantial interpretation of the results**

   A more developed study of the results shows the following data. The leaders of the production of coffee are Brazil - 132 varieties, Colombia - 183 varieties, Guatemala - 181 variety, Mexico - 236 varieties. Despite the number of varieties of Arabica, a leader in the final one was Brazil. The grain in the study is divided into Blue-Green, Bluish-Green, Green. The most common grain is green. All three colors have an average of a friend who are close to a friend with average points, however, in a general series of comparisons, Blue-Green leads. Most of the grain from the study, about 80 percent, is specialty. Aftertaste and flavor, acidity and flavor have the highest correlation between a pair of characteristics to the maximum score, while acidity, body and aroma have the smallest correlation. The greatest amount of flavor has a score of 7.8, acidity 7.75 and 7.8, aroma 7.7 and balance 7.75. From this we can conclude that most of the grain has average characteristics below 80 specialty points, therefore, a number of prominent characteristics in the lot help it to become specialty. Also, the assumption about the dependence of the quality of the fried grain on its moisture was refuted - the grain was of the highest quality with a moisture content of 10-14%, but also with a near zero score of moisture.

2. **Limitations of the analysis**
   The limitations of our study were the difficulty of extracting and sorting the data. Initially, it was intended to use several coffee varieties for side-by-side and cross-sectional comparisons, but either there was not enough data, or they were not aggregated, and their processing would take a lot of time and our resources. Another limitation is the lack of a bean height study - the Q-graders did not adhere to a specific format when filling in the tables, and the preparation of data on coffee bean heights cannot be aggregated, since there are more than 1400 rows. Further, the true difference between cupping characteristics is difficult to trace, since most of the lots studied are specialty and have very close scores to each other. That is why the research can only be useful for researching specialty coffees in the high – high plus segment.

3. **Contribution to the literature**
   A significant addition to the literature used could be a study on the origin of coffee beans, carried out using ML technology. A group of researchers from Brazil conducted a country-specific study to track the evolution and development of coffee species. The origin of the grain would help us cross-study with other species, find differences or similarities in characteristics, draw parallels with natural mutations and their impact on the score. That is why the traceability of Arabica coffee by country of origin is feasible using quality attributes, machine learning, robust multivariate and univariate data science. A cause the number of records in the dataset was insufficient to classify the country of origin with high accuracy, and should be about 25 times higher.

   Link to the study:
   Machine learning to support geographical origin traceability of Coffea Arabica

4. **Perspectives for the future research**
   This study can be used in specialized establishments such as roasteries, training rooms and coffee shops as a guide to primary classifiers of grain quality and flavor profile. What's more, the graphics and illustrations can be used in training classes for initial learning material for barista trainees. The project can be further developed to include datasets with other coffee methods subjected to the same process of analysis and comparison. Moreover, the study can be included in a number of papers on the properties and characteristics of coffee beans or serve as the basis for a methodological manual on cupping.

References

1. Lingle, T. R., & Menon, S. N. (2017) Cupping and grading—Discovering character and quality. In *The craft and science of coffee* (pp. 181-203). Academic Press.

2. Caten, C. S. 2019 Sensory profile of fermented arabica coffee in the perception of American cupping tasters. In *Agricultural Sciences*, *10* (3), 321-329.
3. Nadhiroh, H. (2018) Effect of different post-harvest processing on the sensory profile of Java Arabica coffee. *Advances in Food Science, Sustainable Agriculture and Agroindustrial Engineering (AFSSAAE)*, *1* (1), 9-14.
4. Revi, I. 2019 Coffee Cupping: Evaluation of Green Coffee Quality. In *Coffee* (pp. 335-360).
5. Morjaria, A., & Sprott, M. (2018) Ugandan Arabica coffee value chain opportunities. In *International Growth Centre, Policy Brief*.
6. Lambri, M. (2021) Sensory profile of Italian Espresso brewed Arabica Specialty Coffee under three roasting profiles with chemical and safety insight on roasted beans. *International Journal of Food Science & Technology*, *56* (12), 6765-6776.