COFFEE QUALITY PREDICTION

GROUP 05

Descriptive Analysis

Kavindi Chamathka	- s16367
Nethmi Sansala	- s16252
Kavindu Weerasekara	- s16076

Abstract

Coffee is a valuable agricultural product, popular and appreciated worldwide due to its unique flavor and aroma. The search to improve coffee quality comes from many fronts, as do the many ways to measure quality and the factors that affect it. This report aims to present the findings of the exploratory data analysis conducted on **'Coffee Quality Data'** dataset obtained from Kaggle. The key objective of this analysis is to identify the factors that have a huge impact on the quality of coffee. The findings from this analysis can serve as a foundation for the consumers to get access to better tasting coffee by making informed choices.

Table of Contents

List of Figures

List of Tables

Introduction

Coffee is one of the most consumed beverages in the world and is crucial in the economy of many developing countries. Due to the economic impact of coffee around the world, research on coffee quality has become essential. Over decades of research, hundreds of studies have been concerned with investigating coffee quality. For consumers, the concept of product quality is relative and depends on their needs and interests. However, in the global coffee market, both producers and consumers are prioritizing quality. Hence determining the factors associated with the quality of coffee is important. Apart from consumers and producers, exporters, retailers, researchers and policymakers also will benefit throughout this analysis. This report focuses on the descriptive analysis of coffee quality prediction, exploring several key variables.

Description of the Question

In this descriptive analysis we focus on understanding the structure of the dataset, trends and relationships between variables. Through descriptive statistics and visualizations, this report aims to explore how different factors affect the quality of coffee, answering the following questions:

- What are the main factors associated with the quality of coffee?
- How well do descriptive statistical techniques help in understanding the distribution and relationships of those factors?
- Can the coffee samples be grouped into distinct clusters and how are these clusters related to quality grades?

Description of the Data Set

The **'Coffee Quality Data'** dataset obtained from Kaggle contains 207 observations with 41 features. A description of each variable can be found below.

Variable	Description
ID	
Index	
Country of Origin	The country where coffee was produced.
Farm Name	The specific farm where coffee was cultivated.
Lot Number	A unique identifier for the coffee batch.
Mill	The processing facility where the coffee was processed.
ICO Number	International Coffee Organization number.
Company	The company responsible for marketing or exporting the coffee.
Altitude	The elevation at which the coffee was grown.
Region	The specific geographical area within the country
Producer	The entity that cultivated the coffee beans
Number of Bags	The quantity of coffee bags
Bag Weight	Weights of the coffee bags
In Country Partner	The local organization
Harvest Year	Year when the coffee beans were harvested
Grading Date	Date when the coffee sample was officially graded
Owner	The individual or entity who owns the coffee farm
Variety	The specific botanical variety or cultivar of the coffee plant
Status	The current classification or state of the coffee batch
Processing Method	The technique used to process coffee cherries after harvest
Aroma	The fragrance perceived when the coffee is brewed.
Flavor	The overall taste profile
Aftertaste	The lingering taste left on the palate after swallowing
Acidity	The bright, tangy sensation that adds liveliness to the coffee.
Body	The weight or mouthfeel of the coffee
Balance	The harmony between acidity, body, and flavor.
Uniformity	Consistency of flavor across different cups from the same sample.
Clean Cup	The clarity and purity of the flavor
Sweetness	The pleasant, sugary notes that enhance the overall taste.
Overall	Rating assigned to the coffee sample after cupping
Defects	Total number of defects found in the coffee sample
Total Cup Points	The cumulative score assigned during cupping

Moisture Percentage	The percentage of moisture content in the coffee beans.
Category One Defects	Serious defects such as black beans, sour beans, or insect damage
Quakers	Immature beans that fail to roast properly
Color	The physical color of the coffee beans or roast
Category Two Defects	Less severe defects like broken beans or skins
Expiration	The date after which the coffee is no longer recommended
Certification Body	The organization that certifies the coffee's compliance
Certification Address	Contact or location details of the certification organization.
Certification Contact	Contact information for the certification body

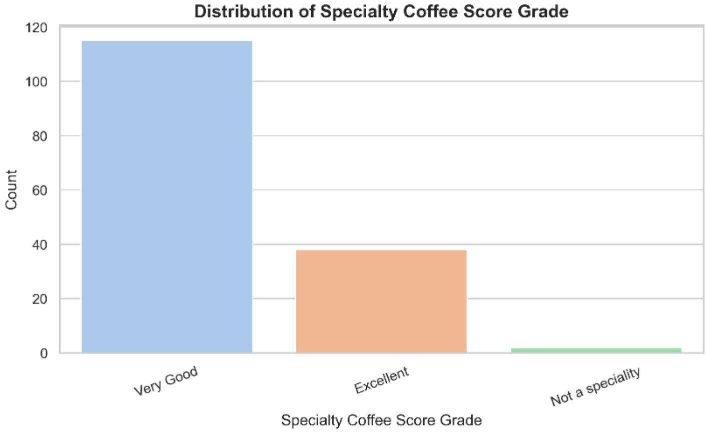
Table 1: Dataset Description

Data preprocessing

- Columns like *ID, Farm Name, Lot Number, Mill, ICO Number, Producer, Status, Defects, Certification Address, Certification Contact* were removed from the dataset due to lack of relevant information to Coffee quality or due to large number of missing values in columns.
- A new variable named Shelf_life_days was created by taking the time difference between Grading date and expiration date.
- In *Altitude* column some of the values were given as a range, therefore we took the average in such cases.
- A New response variable *Specialty coffee score grade* with 4 categories was created based on *Total cup points* variable
- Then 13 rows with missing values were removed.
- No duplicate records were found in the dataset.
- Checked for outliers. There were outliers but present in the numerical variables, except for clean cup, sweetness, and shelf-life days. However, we chose not to remove these outliers from the dataset, and proceeded with the analysis using the original data.
- Then the dataset was split into training and test sets such that the training set consisted of 155 observations.
- Afterward, the descriptive analysis was conducted using the training set.

Main Results of the Descriptive Analysis

Response variable - Specialty Coffee Score Grade



The response variable is a multiclass categorical variable with 4 categories, whether coffee quality is Outstanding, Excellent, Very Good or Not specialty (see appendix). In our dataset there were no records that belong to the Outstanding coffee quality and majority of records belong to Very good quality. Moreover, 2 records don't belong to any of the specialties.

Figure 1: Distribution of Response Variable

Distribution of each numerical variable by response variable (Specialty Coffee Score Grade)

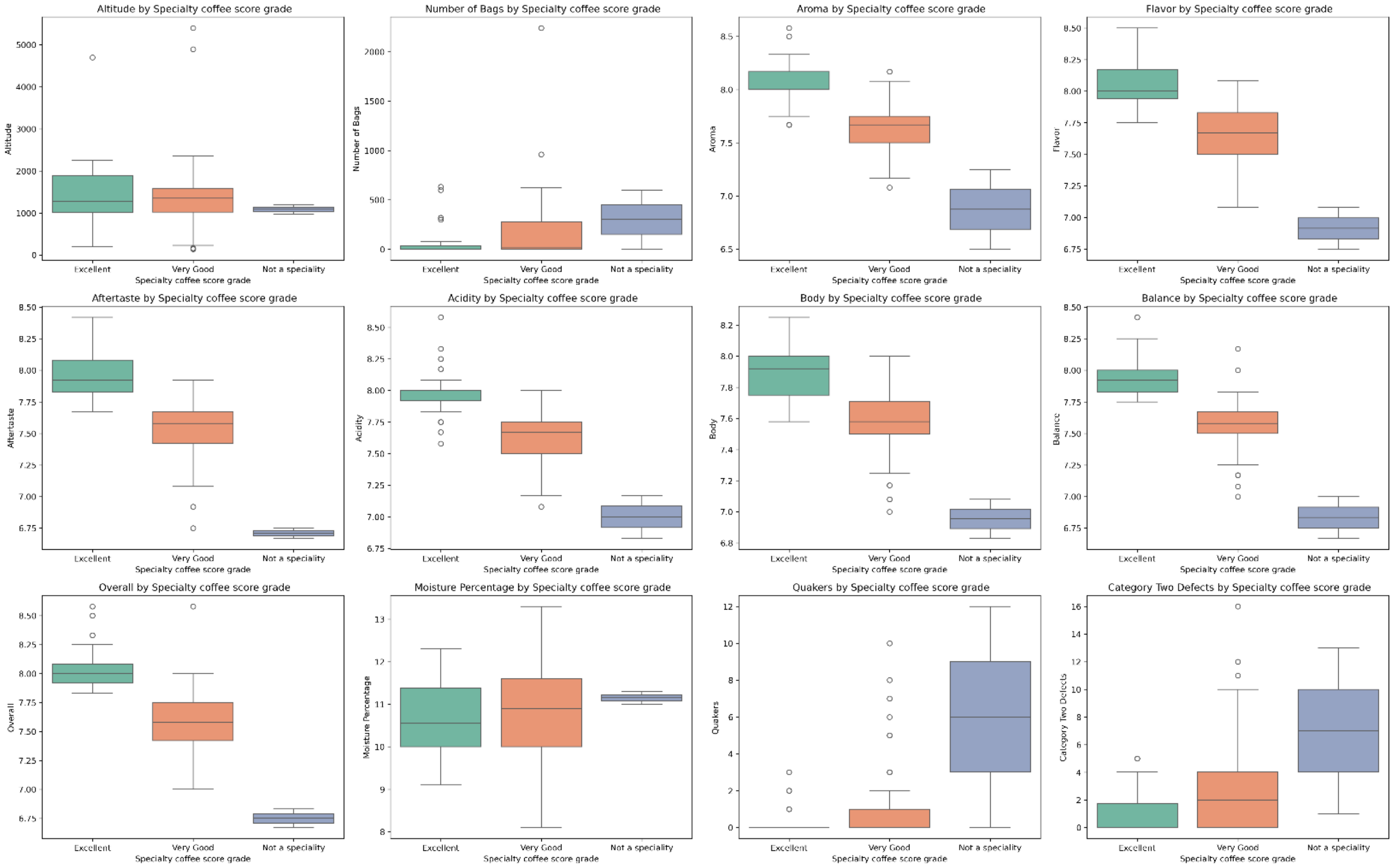


Figure 2: Distribution of Numerical Variables

The boxplots show how various characteristics of coffee differ across quality grades (Excellent, Very Good, and Not a specialty). Overall, higher-grade coffees have better scores for aroma, flavor, aftertaste, acidity, body, balance, and overall quality. Moisture content, on the other hand, shows more spread across samples, indicating that even coffees of the same grade can have quite different moisture levels. Meanwhile, the plots for defects clearly show that lower-grade coffees have more defects, which is expected since defects directly lower the quality grade. This visual pattern highlights how sensory quality improves and defects decrease as we move to higher specialty coffee grades.

Statistical Hypothesis Testing (Kruskal-Wallis test at 5% significance level)

To statistically support these observations, we performed hypothesis testing using ANOVA or Kruskal-Wallis tests (depending on normality within groups) to check whether there were significant differences among the three coffee quality grades for each variable.

Variable	p-value	Conclusion
Altitude	0.3102	Not Significant
Number of Bags	0.3440	Not Significant
Aroma	0.0000	Significant
Flavor	0.0000	Significant
Aftertaste	0.0000	Significant
Acidity	0.0000	Significant
Body	0.0000	Significant
Balance	0.0000	Significant
Overall	0.0000	Significant
Moisture Percentage	0.5445	Not Significant
Quakers	0.1993	Not Significant
Category Two Defects	0.0074	Significant

Table 2: Kruskal-Wallis Test Results

The results showed that variables such as aroma, flavor, aftertaste, acidity, body, balance, overall score, and category two defects have significant differences across the three classes. However, altitude, number of bags, moisture percentage, and quakers did not show statistically significant differences.

Correlation among numerical variables

The heatmap shows the presence of multicollinearity. It highlights that sensory quality attributes such as aroma, flavor, aftertaste, acidity, body, balance, and overall score are all strongly positively correlated, meaning coffees that score high in one of these tend to score high in the others too. In contrast, defects (category two defects and quakers) have moderate negative correlations with these quality measures, indicating that more defects generally lower the quality scores. Altitude shows a mild positive relationship with quality attributes, while moisture content and the number of bags have very weak or no clear correlations with quality scores, suggesting they do not directly influence sensory ratings in this dataset. Overall, the heatmap confirms that higher sensory scores tend to move together, and quality decreases as defects increase.

Note: In the above two analysis parts of numerical variables, we excluded some numerical variables like Clean Cup, Shelf-Life Days, Sweetness, Bag Weight, Uniformity, and Category One Defects, as they contained only a single value or very low variability and thus were not suitable for group comparison.

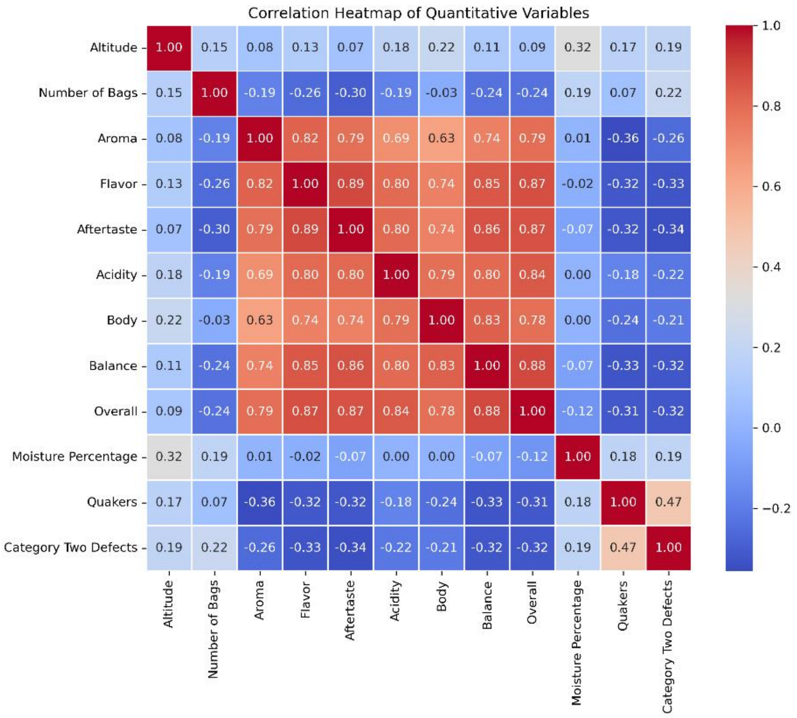


Figure 3: Correlation Heatmap

Chi-Square Test of Independence

The Chi-Square test of independence was conducted to examine the relationship between various categorical variables and the coffee quality grade (response variable). A significance level of 0.05 was used to determine statistical dependence.

Variable	Chi-Square	p-value	Conclusion
Country of Origin	57.136	0.0385	Not Independent
Company	135.914	0.0983	Independent
Region	274.002	0.0000	Not Independent
In-Country Partner	57.358	0.0227	Not Independent
Harvest Year	31.578	0.0000	Not Independent
Owner	147.241	0.0757	Independent
Variety	121.841	0.0011	Not Independent
Processing Method	91.036	0.0000	Not Independent
Color	17.025	0.7619	Independent
Certification Body	57.358	0.0227	Not Independent

Table 3: Chi-Square Test Results

The results indicate that Country of Origin, Region, In-Country Partner, Harvest Year, Variety, Processing Method, and Certification Body all have p-values below 0.05, suggesting that these variables are not independent of coffee quality grade. In other words, there is a statistically significant association between these factors and the quality classification, implying that the distribution of coffee quality grades varies across different countries, regions, harvest years, varieties, processing methods, and certification bodies.

Cluster Analysis

To explore whether any distinct groupings in the dataset containing both quantitative and qualitative variables, Factor Analysis of Mixed Data (FAMD) was applied. Since the variation explained by the first two factor components was very low (15.2%) K-Prototype clustering was used for mixed data. Some numerical variables like Clean Cup, Shelf-Life Days, Sweetness, Bag Weight, Uniformity and Category One Defects were not considered here due to lower variability which will add noise and reduce cluster quality.

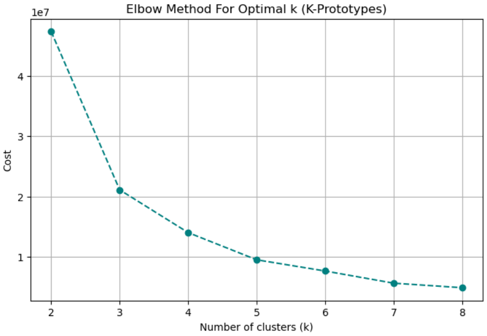


Figure 4: Elbow Method

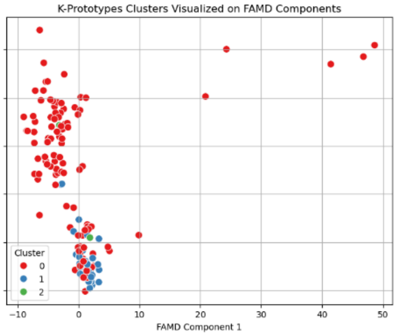
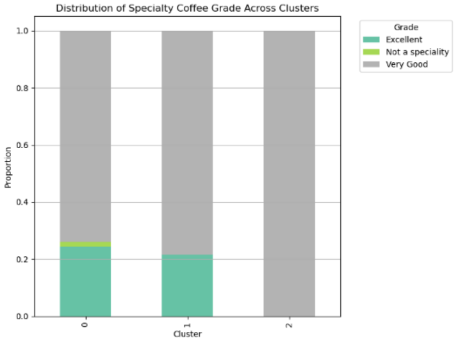


Figure 5: Cluster Visualization

Based on the Elbow Method applied to this data, k=3 is the point where the improvement in cost starts to level off, suggesting that using more than 3 clusters might lead to overfitting or unnecessary complexity. Figure clearly depicts the four clusters that were identified using the K-Prototype approach.

Overview of clusters



Cluster 2 is the most homogenous, indicating consistent quality within this group. It represents a more stable but mid-level quality segment. Cluster 1 also shows a high proportion of Very Good coffee with a moderate share of Excellent samples and no lower grades. However, Cluster 0 is the most diverse, comprising all three categories. This suggests that Cluster 0 captures more variability in quality.

Figure 4: Distribution of response across clusters

The heatmap based on numerical variables significantly associated with specialty score reveals that Cluster 2 has the highest average number of Category Two Defects, indicating potential quality concerns. Also, Cluster 0 has got the highest average sensory scores indicating that it likely represents the highest quality segment of coffee samples in terms of sensory excellence.

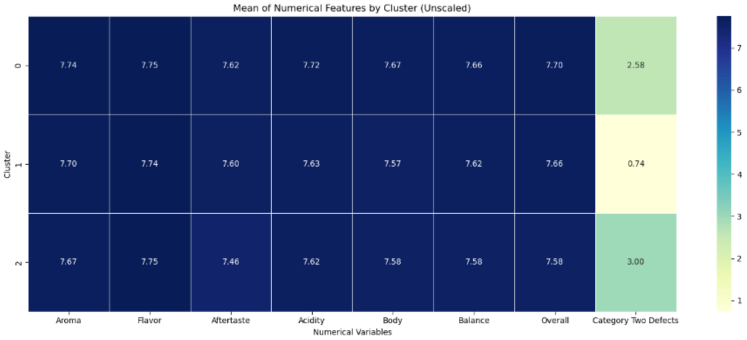


Figure 5: Heatmap of numerical variables across clusters

Conclusion obtained through cluster analysis

- *Cluster 0* - The most diverse group, with global origins and experimental processing, yielding the highest average sensory scores.
- *Cluster 1* - Consists stable-quality coffees with low defects from Taiwan, Thailand, Brazil, Vietnam and Hawaii and has a simpler taste profile, possibly suitable for everyday use.
- *Cluster 2* - Includes mid-tier coffees from Guatemala and Vietnam with moderate quality and traditional processing with a focus on varieties Bourbon and Catimor.

Conclusion

In conclusion, this descriptive analysis of the Coffee Quality dataset has highlighted the key sensory and physical attributes that influence coffee quality. Variables such as aroma, flavor, aftertaste, acidity, body, balance, and overall cup score emerged as significant differentiators among quality grades, while defects were found to negatively correlate with these attributes. Additionally, categorical factors like region, variety, processing method, and country of origin showed strong associations with coffee quality. The clustering analysis further revealed distinct groups characterized by sensory profiles, defect levels, and geographic origins, underscoring the multifaceted nature of coffee quality. These insights provide a foundational understanding for consumers, producers, and researchers, and can guide future advanced predictive modeling to support quality improvement and informed decision-making in the coffee industry.

Suggestion for a Quality Advanced Analysis

In advanced analysis machine learning models can be used, moving beyond traditional statistics. Given the small training dataset with only 155 observations, techniques like bootstrapping are crucial for robust model generalization. Since outliers were not removed, exploring their impact and employing outlier-insensitive models like tree-based models or SVM could be beneficial. Moreover, to address the class imbalance in coffee quality grades, using methods such as oversampling will be beneficial. Future models could also incorporate regularization techniques (Ridge, Lasso) during model training to further reduce dimensionality, prevent overfitting and handling multicollinearity as well.

Appendix

- Link for the Data set: https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi
- Python Codes

- **Creating specialty coffee score grade variable**

```
    data['Total Cup Points'] >= 90.00,
    (data['Total Cup Points'] >= 85.00) & (data['Total Cup Points'] <=
89.99),
    (data['Total Cup Points'] >= 80.00) & (data['Total Cup Points'] <=
84.99),
    data['Total Cup Points'] < 80.00
```

References

1. https://colipsecoffee.com/blogs/coffee/specialty
2. https://www.danioprea.com/projects/coffee-quality-analysis