Optimizing Coffee Selection: Data-Driven Insights for Coffee-Shop Owners to Maximize Price-to-Performance Across Regions

Project Report

For

DS 5110-Data Management and Processing

Team 12

Jiajun Fang: fang.jiaj@northeastern.edu

Henry Guo: guo.zeh@northeastern.edu

Xin Wang: wang.xin18@northeastern.edu

Table of Contents

1 Summary

Selecting the right coffee products that resonate with customers is a key challenge for coffee shop owners. Decisions regarding coffee selection are often influenced by subjective experiences or limited market data, considering factors such as origin, roast level, aroma, and price. This project aims to address these challenges by providing data-driven insights into the key coffee features that influence customer ratings. By understanding these relationships, coffee shop owners can make more informed decisions to optimize their offerings, boosting both customer satisfaction and sales.

To explore this, we collected data from the CoffeeReview website, scraping over 8,000 reviews that provided detailed information on coffee attributes like ratings, sensory features (aroma, acidity, body, flavor), roast levels, price, and origin. Our analysis uncovered key patterns, including linear modeling, clustering, and text analysis. The linear model showed that sensory attributes like flavor, acidity, and aroma are strong predictors of ratings, with the final model achieving an R^2 of 0.9945. Clustering revealed two main segments: a premium group with higher ratings and prices, and a diverse group with a broader range of roast levels and origins. Text analysis highlighted that positive sentiment and specific terms like "sweet" and "chocolate" were associated with higher ratings. These insights offer coffee shop owners valuable guidance in optimizing product offerings based on customer preferences.

2 Methods

2.1 Dataset Description

The coffee review dataset is sourced from CoffeeReview, a leading platform that provides detailed reviews and ratings for a wide variety of coffee products. It includes over 8,000 coffee reviews, capturing both sensory evaluations (such as aroma, acidity, body, and flavor) and descriptive text offering insights into the coffee's origin, roast level, and price. This dataset was collected through web scraping and serves as a valuable resource for understanding how different coffee features influence customer ratings. By analyzing the relationships between coffee characteristics and ratings, coffee businesses can optimize product offerings, while consumers can make informed decisions based on their personal preferences.

Variables of Interest:

- **Rating** - float - Overall rating of the coffee.
- **Coffee Origin** - str - Origin of the coffee beans.
- **Roaster** - str - Name of the roaster company.
- **Roast Level** - str - Type of the roast (Light, Medium-Light, Medium, Medium-Dark, Dark).
- **Agtron** - str - This variable contains 2 values: the left one indicates the Agtron reading for Whole Bean and the right one indicates the Agtron reading for Ground coffee. An Agtron machine reflects light on a sample of coffee to objectively and accurately assign a number to the beans' roast color. The smaller the number, the darker the roast. How to Interpret Agtron Values: Dark Roast: Agtron 0-35 (very dark brown to almost black); Medium Roast: Agtron 35-55 (medium brown); Light Roast: Agtron 55-100 (lighter shades of brown to cinnamon-like colors).
- **Aroma** - float - This value assesses how intense and pleasurable is the aroma when the nose first descends over the cup and is enveloped by fragrance.
- **Acidity** - float - Acidity is evaluated on a 1 to 10 scale, reflecting both intensity and quality. A higher score (e.g., 8-10) indicates a lively, bright acidity, often associated with fruity or citrusy notes, while a lower score (e.g., 4-6) suggests a smoother, more subdued acidity, common in darker roasts or naturally low-acid coffees.
- **Acidity/Structure** - float - In some reviews, Acidity/Structure is used instead of Acidity to highlight how acidity contributes to the overall balance of the coffee, including its relationship with body, flavor, and aftertaste. This provides a more comprehensive evaluation of the coffee's complexity. The choice between Acidity and Acidity/Structure may depend on the reviewer's preference or the coffee's characteristics.
- **Body** - float - Body refers to the coffee's texture and mouthfeel, rated on a 1-10 scale. Higher scores indicate a richer, fuller mouthfeel. Full-bodied coffee is rich and creamy, often found in dark roasts or high-oil coffees like Sumatran. Medium-bodied coffee has a balanced texture, common

in many Central American coffees, while light-bodied coffee is delicate and tea-like, often seen in high-altitude coffees like Ethiopian Yirgacheffe.
- **Flavor** - float - Flavor encompasses all sensory experiences not covered by acidity, aroma, and body, effectively serving as a synthesis of these elements.
- Aftertaste - float - In coffee tasting, aftertaste refers to the sensations that linger after the coffee has been swallowed (or spit out).
- **Est. Price** - str - Price of the coffee.
- **Review Date** - str - Date of the coffee review.
- **Blind Assessment** - str - The Blind Assessment section provides an unbiased sensory evaluation of the coffee, describing its aroma, flavor, acidity, body, and finish without knowing its origin or brand.
- **Notes** - str - The Notes section provides background information about the coffee, including its origin, variety, processing method, and roaster.
- **Who Should Drink It** - str - The Who Should Drink It section provides recommendations for the ideal audience based on the coffee's characteristics and appeal.
- **Bottom Line** - str - The Bottom Line section in Coffee Review provides a concise summary of the coffee's key characteristics and overall impression.

2.2 Data Tidying and Preprocessing

Our dataset was sourced from CoffeeReview and contains expert reviews of coffee products from 1997 to 2025. Each entry includes sensory ratings, metadata about the coffee and its roaster, price information, and review details. Below are the key data tidying steps we performed:

- **Roast Level:** The "Roast Level" variable was converted into an ordinal categorical variable, with ordered levels ranging from "Light" to "Very Dark," maintaining the natural progression of roast intensity for modeling purposes.
- **Agtron:** The original "Agtron" column contained two values, one for Whole Bean and the other for Ground coffee. We split this into two separate numeric columns—**Agtron_whole** and **Agtron_ground**. Additionally, obvious typos and extreme outliers in these columns were identified and corrected by cross-referencing review sources and replacing them with reasonable values based on similar entries.
- **Roaster Location:** The "Roaster Location" column exhibited varying formats (e.g., "City, State," "City, Country," or just "Country"). To handle this, we combined the "Roaster" and "Roaster Location" fields into a single identifier string and used the Google Maps API to extract latitude and longitude coordinates for future geographic analysis.
- **Estimated Price:** The "Est. Price" column featured diverse formats, including different currencies, units, and quantities. We used regular expression (regex) functions to extract the currency, numeric price, quantity, and unit from each entry. Prices were then converted to USD per 100 grams for fair comparison. Historical prices were adjusted for inflation using the U.S. CPI, with 2024 as the base year.
- **Review Date:** The "Review Date" was converted into a datetime format to facilitate time-based analysis and inflation adjustment.
- **Acidity:** The "Acidity" and "Acidity/Structure" columns, which were semantically similar and mutually exclusive, were merged into a single field for consistency.
- **Coffee Origin:** The "Coffee Origin" field was standardized to reflect each coffee's country name. For example, entries referencing specific regions (e.g., "Yirgacheffe") were categorized under their corresponding country (e.g., "Ethiopia").

2.3 Models

2.3.1 Linear Model

To predict coffee ratings, we explored multiple regression approaches grounded in the sensory and pricing attributes of each coffee entry. We began by conducting **exploratory data analysis (EDA)** to

assess feature distributions, relationships, and correlations. Key features such as Flavor, Acidity, Aroma, Aftertaste, and Body showed the strongest linear relationships with the target variable, Rating. We built an initial baseline model using simple linear regression with Flavor as the sole predictor, as it had the highest F-statistic during feature importance testing.

To improve predictive performance, we applied **stepwise feature selection** (forward selection using p-values) to identify the most impactful subset of features for multiple linear regression. This method iteratively selected features based on statistical significance and model improvement.
All models were trained and evaluated using a train-test split (e.g., 80/20 split). We evaluated model performance using R^2, RMSE, and p-values for coefficient significance. We also performed collinearity tests using VIF scores to examine if the features used in our model exhibits collinearity.

2.3.2 Clustering

The goal of the clustering analysis was to group coffee samples based on various sensory features, origin, pricing, and other characteristics in order to identify meaningful clusters. The features used for clustering included Roast Level, Agtron, Aroma, Acidity, Body, Flavor, and Aftertaste. Additionally, coffee origin, ratings, and prices were considered in the analysis to provide a more comprehensive view of the coffee clusters. For the cluster analysis part, the coffee origin was mapped with its corresponding continent.

2.3.2.1 Dimensionality Reduction

Due to the high correlation between certain features, dimensionality reduction was performed to improve the efficiency of clustering. Specifically, **Kernel Principal Component Analysis (Kernel PCA)** was applied to reduce redundancy while retaining key features. This step was necessary as Roast Level, Agtron, and Acidity exhibit strong relationships, with light roasts typically corresponding to higher Agtron values and higher Acidity. Although **PCA** was initially attempted, it resulted in a relatively low Silhouette Score, prompting the use of **Kernel PCA**, which is better suited for non-linear dimensionality reduction. A Silhouette Score of 0.4420 was calculated to assess the quality of the clusters.

2.3.2.2 Clustering Approach

The **K-Means** algorithm was used to partition the dataset into clusters based on the sensory features, pricing, and ratings. We performed K-Means clustering for different values of K, ultimately selecting **5** clusters based on the elbow method and some other clustering metrics like the largest average silhouette width.

Hierarchical clustering with Ward.D linkage rule was used as a complementary approach to gain a different perspective on the data. This method builds a dendrogram that represents how each data point is merged into clusters. The optimal number of clusters was determined by analyzing the dendrogram and identifying the largest height interval, which suggested **2** clusters as the best solution for our data.

2.3.2.3 Cluster Analysis

After clustering, the following analyses were conducted to evaluate and compare the clusters:

- **Coffee Origin:** Analyzed the distribution of coffee origins across the clusters to identify any patterns or dominance of particular regions.
- **Roast Level:** Examined the proportion of different roast levels within each cluster.
- **Acidity:** Analyzed the acidity levels in each cluster to determine whether specific clusters correspond to higher or lower acidity.
- **Ratings and Price:** Investigated how ratings and prices were distributed across the clusters to see if certain clusters were associated with higher-quality coffee or premium pricing.

2.3.3 Text Analysis

All analyses were performed in R (v4.2+) using the **tidyverse** and **tidytext** ecosystems.

In **TF-IDF** feature extraction, to quantify “distinctiveness” of vocabulary in each review, we counted word frequencies (n) per (review_id, document), then computed TF–IDF scores and aggregated them into an average TF–IDF per review. These avg_tf_idf values were joined back to the Rating column for visualization and correlation analysis.

In **Topic Modeling (LDA)** We constructed a document–term matrix (DTM) from the same tokens—treating each <document>_<review_id> pair as a unique “document”—and fit a Latent Dirichlet Allocation (LDA) model with topicmodels (k = 3,5,7): We extracted the top-10 words per topic (highest β), which characterize each latent theme, and examined their prevalence across comment columns.

Then, bigrams and trigrams are used to capture common phrases. Top-15 n-grams per column with faceted bar charts are visualized. To explore word associations, we built a co-occurrence network of words appearing in the same review (using widyr’s pairwise_count()) and plotted it with ggraph—filtering to pairs with ≥ 7 co-occurrences and repelling node labels for clarity.

In sentiment analysis, we compared three lexicons: Bing, AFINN, and NRC. For each lexicon,

1. Joined tokens to the lexicon, keeping only relevant categories.
2. Computed net sentiment (positive–negative word counts for Bing/NRC or summed value for AFINN).
3. Averaged net sentiment per review per column.
4. Plotted net sentiment vs. Rating with faceted scatterplots + linear fits.

We quantified the linear relationship between average TF-IDF and Rating for each comment column by merging avg_tf_idf with Rating, we computed Pearson’s correlation coefficient by document. Lastly, we compared text sentiment against time to identify any temporal connections.

3 Results

3.1 Linear Model

The baseline model using only the Flavor feature achieved the following performance:

- **Train R^2**: 0.660
- **Test R^2**: 0.658
- **Test RMSE**: 1.256

After applying stepwise feature selection, the final multiple linear regression model included: Flavor, Acidity, Aroma, Aftertaste, and Body.
This model yielded substantially better results:

- **Train R^2**: 0.993
- **Test R^2**: 0.9945
- **Test RMSE**: 0.158

These results indicate that sensory attributes have strong predictive power for expert-rated coffee quality, with all included variables showing positive and significant contributions to the rating.

3.2 Clustering

[Detailed **cluster analysis figures** are included in the **Appendix A** section.]

3.2.1 K-Means Clustering Analysis

The **K-Means clustering** algorithm was initially applied to group the coffee samples into five clusters. Below is the detailed breakdown of the clustering results:

- **Cluster 0**: Predominantly medium-light roasts, with a high proportion of Ethiopian coffees and a significant presence of African coffees. Coffees in this cluster exhibit higher acidity (8.0-9.0) and are priced moderately, showing higher ratings compared to other clusters.

- **Cluster 1**: This cluster is dominated by medium roasts and contains a diverse set of origins, including North American and Asian coffees. It has a medium acidity range (7.0-8.0) and generally lower ratings and more moderate prices compared to Cluster 0.
- **Cluster 2**: Light roasts are predominant in this cluster, with a strong representation of Ethiopian coffee. This cluster is notable for its higher ratings and higher price compared to other clusters, particularly in the upper quartile. The acidity is also higher (8.0-10.0), emphasizing the bright, lively profile typical of light roasts.
- **Cluster 3**: This cluster has a more diverse roast profile, including medium-dark and very dark roasts. It features a high proportion of Indonesian coffee and has lower acidity (3.0-8.0), reflecting a more subdued coffee profile. The ratings are generally lower, and the **prices** are also on the more affordable side.
- **Cluster 4**: Similar to Cluster 0, this cluster is primarily composed of medium-light roasts, with a notable representation of Ethiopian coffee. The coffees in this cluster are characterized by higher acidity (scores ranging from 8.0 to 9.0, with a dominant score of 8) compared to Clusters 1 and 3. Additionally, this cluster exhibits higher ratings than Cluster 1, while maintaining moderate pricing.

In summary, the K-Means clustering results led to the identification of five distinct groups. However, based on the features analyzed, it became apparent that these groups could be simplified into two main segments:

- **Premium Group** (Clusters 0, 2, and 4): Representing higher ratings, higher acidity, and a predominance of Ethiopian and African coffees.
- **Diverse Group** (Clusters 1 and 3)**:** This group includes a higher proportion of North American and Asian coffees, with a mix of medium-dark to dark roasts, a wider range of acidity, and generally lower prices.

3.2.2 Hierarchical Clustering Analysis

This method provides a dendrogram that suggests the best number of clusters based on the largest height interval. The analysis suggested dividing the dataset into **two clusters**, reinforcing the division observed in the K-Means clustering.

- **Cluster 0** (Hierarchical): This cluster contains a balanced mix of South American, North American, and Asian coffees, with a higher proportion of Brazilian and Indonesian coffees. The roast level is primarily medium roast, and the acidity is in the 7.0-8.0 range. The coffees in this cluster have slightly lower ratings (median around 90 points) and moderate pricing (up to 40 USD/100g).
- **Cluster 1** (Hierarchical): Cluster 1 is dominated by Ethiopian coffees, with a strong representation of African coffees overall (around 42%). The roast level is mostly medium-light, and acidity scores are high (8.0-9.0). This cluster is associated with higher ratings (median around 93 points) and higher prices, reflecting the premium quality of the coffees.

3.2.3 Comparison and Insights

The K-Means analysis initially identified five clusters, but further examination revealed that these could be grouped into two main segments: the **Premium Group** and the **Diverse Group**. Hierarchical clustering confirmed this division, with Cluster 0 representing a diverse mix of coffees with darker roasts, moderate prices, and ratings, while Cluster 1 was more homogeneous, dominated by Ethiopian and African coffees, with higher acidity, ratings, and prices. Both clustering methods highlighted two key segments: premium coffees and a diverse, lower-priced group, offering valuable insights for

coffee-shop owners to optimize product selection and pricing based on customer preferences for origin, roast level, and sensory attributes.

3.3 Text Analysis

The text analysis revealed several key factors influencing coffee ratings. Terms such as "Sweet," "Chocolate," and "Fruit" were strongly associated with a net positive sentiment and higher ratings, as evidenced by raw word frequency, bigrams, and Latent Dirichlet Allocation (LDA) across multiple topic models.

Our sentiment analysis further confirmed that a positive sentiment correlates with higher ratings across three different sentiment lexicons. Additionally, a comparison between ratings and the average TF-IDF score per column showed that coffee reviewers tend to assign lower ratings when using more distinctive or specific comments.

Correlation analysis indicated that coffee ratings are most significantly influenced by the blind assessment, followed by the notes, who the coffee is recommended for, and finally, the bottom line. Finally, our temporal trend analysis demonstrated that real-world events, such as the financial crisis and the COVID-19 pandemic, had a significant impact on coffee ratings, reflecting broader societal influences on consumer preferences.

4 Discussion

4.1 Meaning & Impact of the Results

Sensory Attributes Are Paramount. The linear regression confirmed that *Flavor* and *Acidity* together explain the vast majority of rating variance (adjusted $R^2 > 0.99$). This quantifies what tasters have long known: bright, fruit-like acidity balanced by sweetness is the hallmark of a top-scoring coffee.

Distinct Market Segments Emerge. Both K-Means and hierarchical clustering consistently split our samples into a Premium Group—high-acidity, predominantly Ethiopian/East African coffees with higher prices and scores—and a Diverse Group of darker roasts, broader origins, and more moderate ratings and pricing. This segmentation provides a data-driven foundation for differentiated product lines.

Language Mirrors Quality. Text mining of four review sections showed that positive descriptors—especially “sweet,” “fruit,” and “chocolate”—cluster in the highest-rated coffees. Sentiment analysis (Bing, AFINN, NRC) confirmed that more positive language predicts higher scores, most strongly in the **Blind Assessment** section. Conversely, higher average TF-IDF (more specialized or defect-focused terms) correlates with lower ratings.

4.2 Who Benefits?

Roasters & Green-bean Buyers can prioritize origins and processing methods that maximize sweetness and fruit complexity.

Café Owners & Baristas can tailor roast profiles, brewing parameters, and menu descriptions to highlight the same sensory highlights that experts reward.

4.3 Turning Insights into Decisions

Sourcing & Roasting: Use the linear model coefficients to forecast rating impacts of marginal changes in acidity or flavor scores, guiding green-bean procurement and roast curve adjustments.

Menu Design: Emphasize tasting notes that align with our high-rating word clusters in menu descriptions and tasting cards.

Quality Control: Monitor reviewer language (e.g., spikes in defect-related TF-IDF terms) as an early warning signal for off-profile lots.

4.4 Limitations & Future Work

- **Domain-Specific Sentiment:** Our use of generic lexicons sometimes mislabels coffee-specific jargon. Developing a custom, coffee-focused sentiment dictionary would sharpen analysis.
- **Consumer vs. Expert Language:** We analyzed expert reviews—future comparisons to consumer feedback (e.g., e-commerce reviews) could validate whether the same linguistic signals hold in the mass market.
- **Supervised Topic Models:** Applying Structural Topic Models with rating as a covariate would directly reveal which topics drive score variation.
- **Multimodal Features:** Integrating chemical assays (e.g., Agtron and cupping scores) or roast-profile metadata alongside text features may improve predictive power and deepen sensory-textual links.

4.5 Takeaways

By combining rigorous statistical modeling, unsupervised clustering, and natural-language techniques, we've created a reproducible, end-to-end framework that ties specific coffee attributes—and specific words—to expert scores. Coffee professionals can leverage these insights to make more informed sourcing, roasting, and marketing decisions, ultimately aligning product offerings with the sensory qualities that matter most.

5 Statement of Contributions

Jiajun Fang

1. Processed columns: Est. Price, Acidity/Acidity Structure
2. Initial EDA of the dataset and linear regression models
3. Wrote data tidying and methods + results section for linear regression model

Xin Wang

1. Scrapped the data from the coffee review website and parsed the text data
2. Processed columns: Agtron, Roaster Location
3. Dimensionality reduction and clustering models

Henry Guo

1. Processed columns: Coffee origin, Roast Level, Review Date
2. Text mining on comment columns: Review Date, Blind Assessment, Notes, Who Should Drink It, Bottom Line
3. Analysis on mined texts

6 References

- Coffee reviews website
- Scrapped dataset
- Rates used to adjust prices

7 Appendix

GitHub Link

Appendix A: Clustering results

A.1 Cluster Analysis - KMeans

A.1.1 Coffee Origin

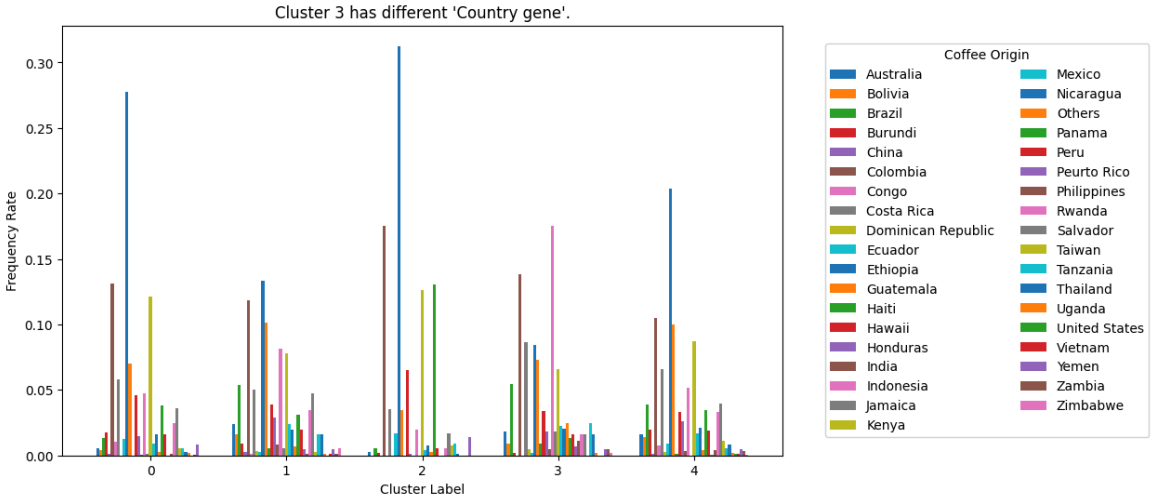

Figure A.1: The distribution of coffee origin (Country) across the clusters. Clusters 0, 2, and 4 have similar Country distributions: High level proportion of Ethiopian coffee. But Cluster 2 has more American coffee; Clusters 1 and 3 tend to have more diverse distributions, but Cluster 3 has the highest proportion of coffee from Indonesia.

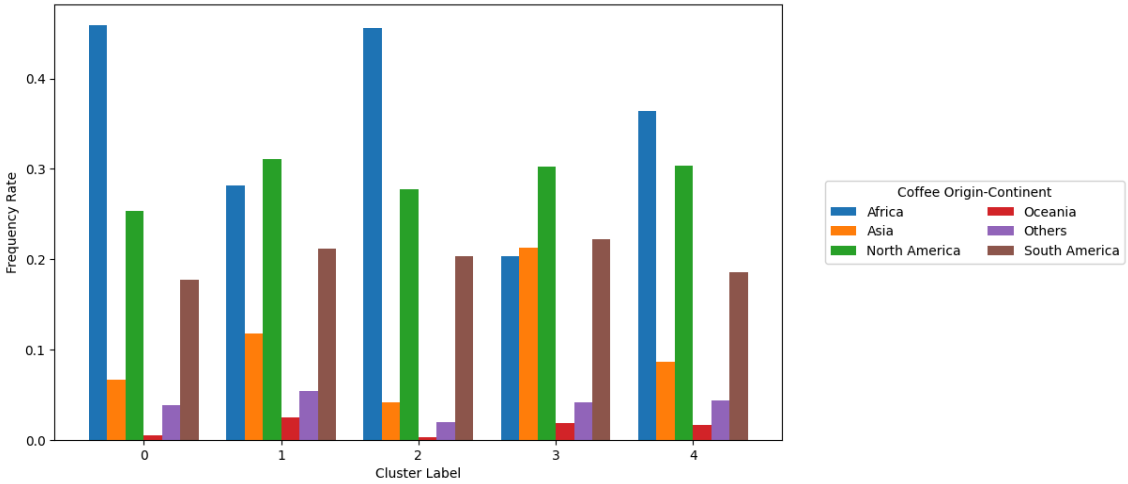

Figure A.2: The distribution of coffee origin (Continent) across the clusters. Clusters 0, 2, and 4 have similar Continent distributions with a high level proportion of African coffees; Clusters 1 and 3 tend to have more North American and Asian coffees.

A.1.2 Roast Level

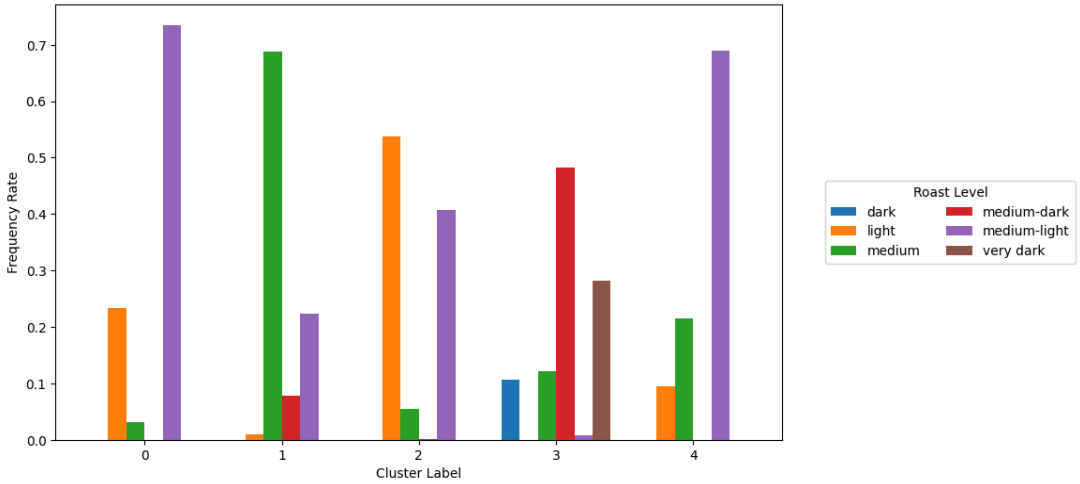

Figure A.3 illustrates the distribution of roast levels across the clusters. Clusters 0 and 4 are predominantly composed of medium-light roasts (purple), making up over 70% of each cluster. Cluster 1, on the other hand, is dominated by medium roasts (green), with approximately 70% of its coffees falling into this category. Cluster 2 is characterized by light roasts (orange), which represent more than 50% of the cluster. Cluster 3 has the most diverse roast profile, containing a mix of medium-dark, very dark, and dark roasts. The order of cluster labels based on the roast levels, from Light to Very Dark, is: 2, 0, 4, 1, 3.

A.1.3 Acidity

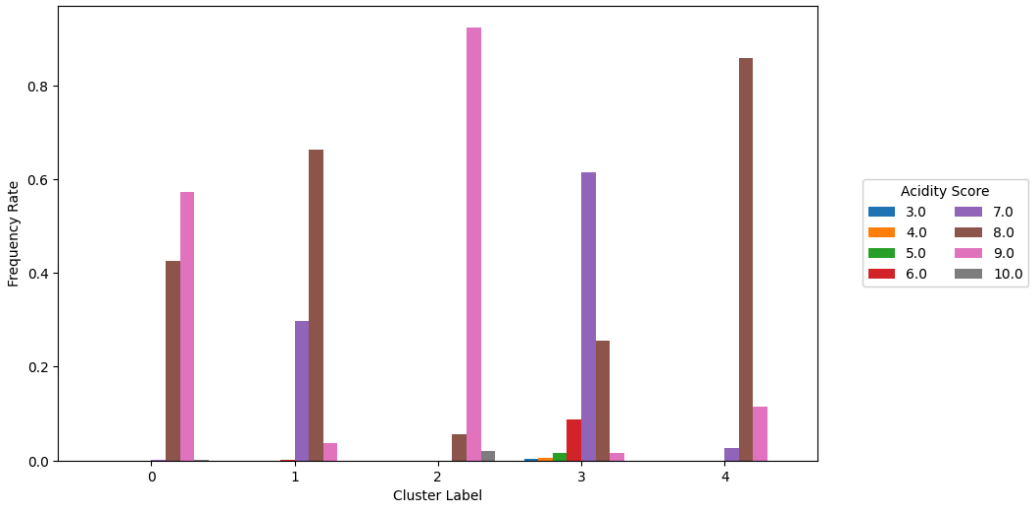

Figure A.4: Frequency distribution of acidity scores across different clusters. Cluster 0, Cluster 2, and Cluster 4 predominantly feature high acidity ratings, with scores of 8.0 and 9.0 being dominant. Cluster 1 exhibits acidity scores of 7.0 and 8.0, while Cluster 3 has a more diverse range of acidity scores, with values ranging from 3.0 to 8.0.

A.1.4 Rating

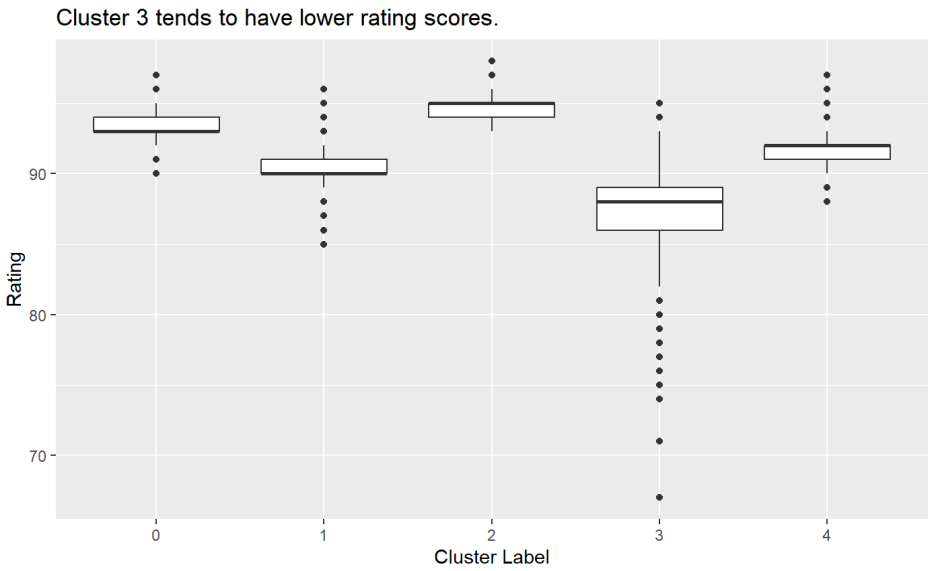

Figure A.5: Boxplot showing the distribution of ratings across different clusters. The plot reveals that Clusters 0 and 2 have similar, higher ratings compared to the other clusters. Cluster 1 and Cluster 4 also exhibit similar rating distributions, though they are slightly lower than Clusters 0 and 2. Cluster 3 tends to have the lowest ratings, with a more spread-out range and lower median compared to the other clusters.

A.1.5 Price

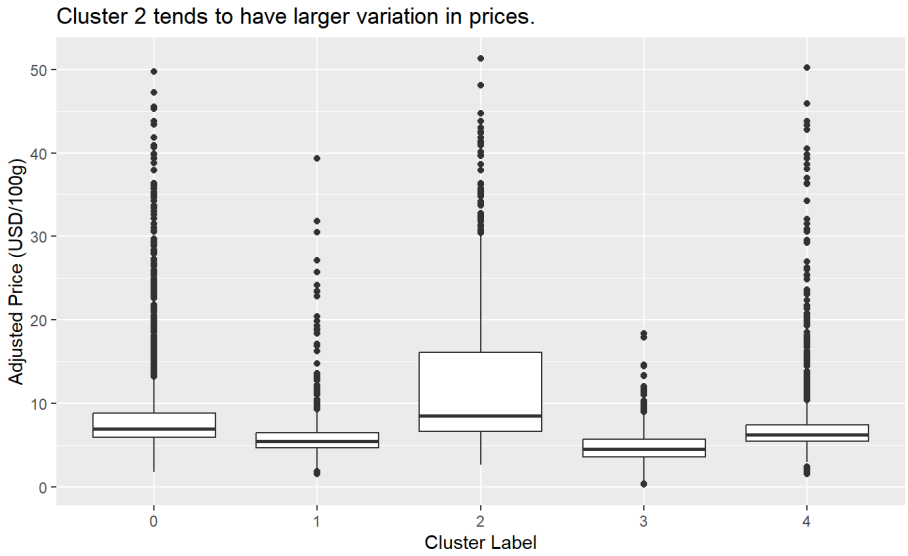

Figure A.6: Boxplot showing the distribution of adjusted prices (USD/100g) across the coffee clusters. Cluster 2 exhibits the largest price variation, with a higher median and upper quartile compared to the other clusters. Cluster 3 generally has lower prices, with minimal variation, indicating a more consistent and affordable pricing range. Clusters 0, 1, and 4 show moderate pricing, with Cluster 0 displaying a slightly wider distribution of prices compared to the others.

A.2 Cluster Analysis - Hierarchical clustering

A.2.1 Coffee Origin

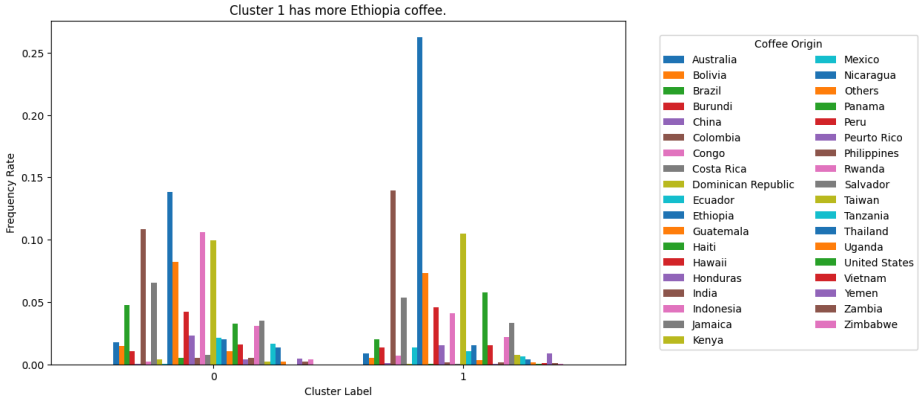

Figure A.7: The distribution of coffee origin (Country) across the clusters. Cluster 1 is predominantly composed of Ethiopian coffee, as indicated by the significant spike in frequency for Ethiopia in this cluster. In contrast, Cluster 0 contains a more balanced representation of coffee origins, with a broader distribution across countries such as Brazil, Colombia, Guatemala, and others.

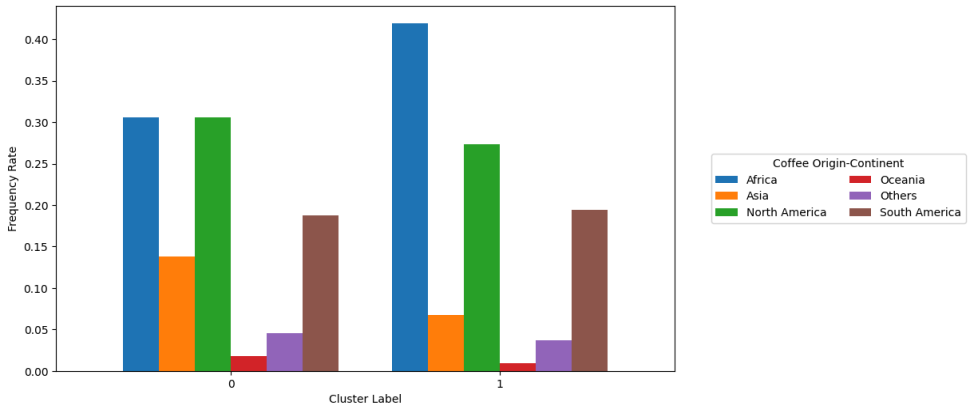

Figure A.8: The distribution of coffee origin (Continent) across the clusters. Cluster 1 has a higher proportion of African coffee, while Cluster 0 is more diverse in terms of coffee origin, with a greater presence from North America and Asia.

A.2.2 Roast Level

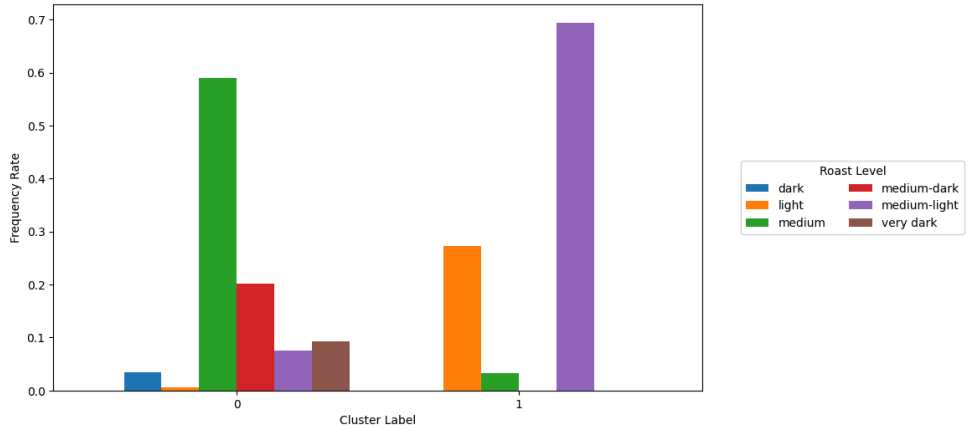

Figure A.9 illustrates the distribution of roast levels across the clusters. This figure highlights the roast level differences between the two clusters, with Cluster 1 showing a clear preference for medium-light roasts, while Cluster 0 features a wider mix of medium and medium-dark roasts.

A.2.3 Acidity

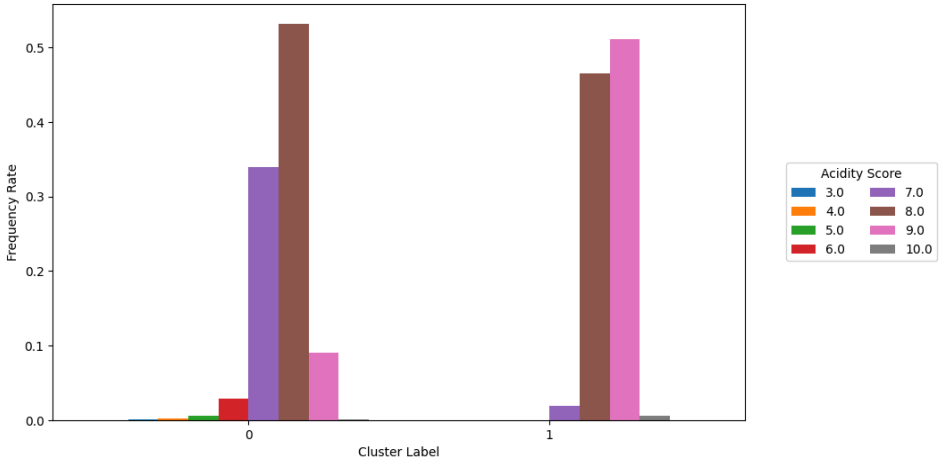

Figure A.10: Frequency distribution of acidity scores across different clusters. Cluster 1 features coffees with higher acidity compared to the more subdued acidity of Cluster 0.

A.2.4 Rating

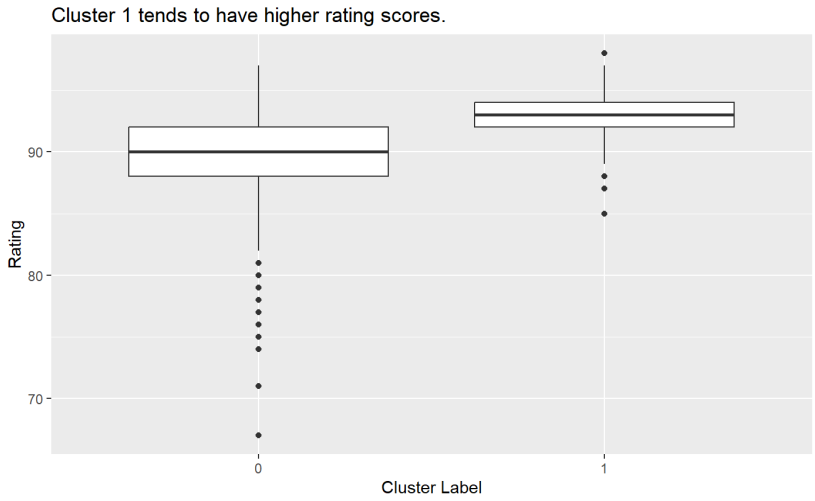

Figure A.11: Boxplot showing the distribution of **ratings** across different **clusters**. Cluster 0 tends to have slightly lower ratings, with the median around 90 points, and a wider distribution that includes more low-scoring outliers. In contrast, Cluster 1 has higher average ratings (median around 93 points) and a tighter distribution, indicating more consistency in ratings.

A.2.5 Price

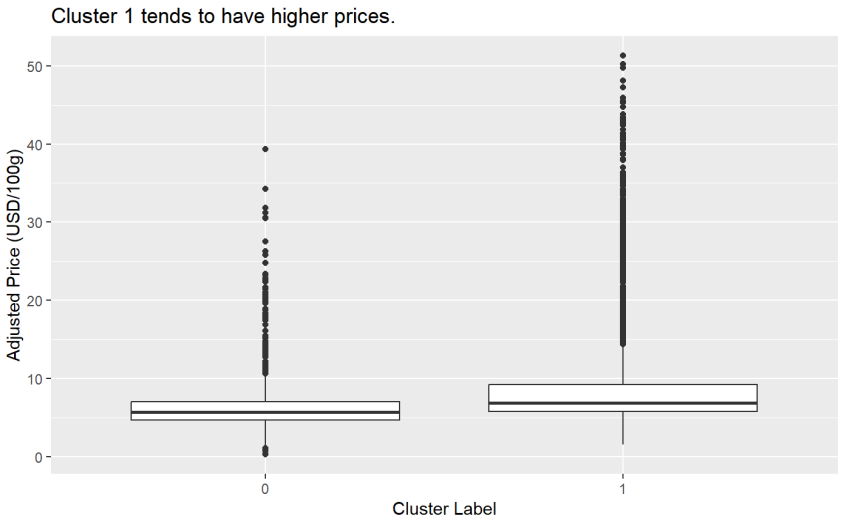

Figure A.12: Boxplot showing the distribution of adjusted prices (USD/100g) across the coffee clusters. Cluster 0 generally has lower prices, with a ceiling around 40 USD/100g and a relatively consistent distribution of prices. In contrast, Cluster 1 exhibits higher average prices and a higher price ceiling, reaching up to 50 USD/100g, reflecting the premium nature of the coffees in this cluster.