Predicting Chocolate Ratings: An R-Based Classification Approach

1. Introduction

This project aims to leverage data-driven techniques to predict whether a chocolate bar will receive a rating above 3 ("Satisfactory") or below 3 ("Unsatisfactory") on an expert rating scale. By framing this as a classification task, the goal is to provide actionable insights that can help manufacturers and stakeholders assess the likelihood of a chocolate bar's success based on various attributes.

The analysis primarily focuses on tree-based models, given their ability to handle complex relationships between features and their robustness to non-linearities in the data. The project approach included several key steps: data exploration and preprocessing, feature engineering to enhance predictive power, hyperparameter tuning for model optimization, and rigorous model selection using comprehensive evaluation metrics. The detailed methodology will be discussed in the following sections.

2. Data Description

a) Variable 'Rating'

Chocolate bar ratings range from 1 to 5, with approximately 69% values fall between the scores of 3 and 4 (Appendix 1). The average ratings across the board is 3.2.

b) Numerical Variables: REF, Review Date, and Cocoa.Percent (Appendix 2)

REF represents the recency of the review entry, the higher the value the more recent a review is. REF and Review Date display little to none correlation with ratings, which means they have limited relevance for predicting ratings. Cocoa.Percent has a weak negative relationship with ratings, suggesting very high cocoa content might slightly decrease consumer satisfaction.

c) Categorical Variables: (Appendix 3)

Company.Maker.if.known (Chocolate manufacturer): There are 416 distinct manufacturers in the dataset, the most popular ones are Soma, Bonnat, and Fresco.

Specific.Bean.Origin: The specific region the cocoa beans are sourced from, with more than 1000 unique regions included. The most popular location is Madagascar, followed by Peru and Ecuador.

Company.Location: Companies are based across 60 different countries, with the majority based in the U.S., followed by France, Canada, and the UK.

Broad.Bean.Origin: The broad region or the country the cocoa beans are sourced from. 100 regions are included, with the most popular regions being Venezuela, Ecuador, and Peru.

Bean.type: The breed of beans used for making the chocolate, if provided. There are 41 unique bean types, with the most popular one is a blend, followed by Trinitario.

d) Feature Importance in Predicting 'Rating' (Appendix 4)

Cocoa.Percent demonstrates significant importance, particularly in terms of its impact on prediction accuracy (%IncMSE), indicating that the cocoa content in chocolate strongly influences consumer ratings and satisfaction.

REF scores high in both %IncMSE and IncNodePurity but is less relevant to the business context as it does not directly relate to product characteristics.

Company.Location exhibits high importance, suggesting that the geographical origin of the manufacturer and the type of bean used significantly influence consumer perceptions.

Company.Maker.if.known. exhibits medium to high importance and also scores high on the Gini index. This suggests that the brand of the chocolate has moderate influence on consumer ratings.

Bean.Type exhibits moderate importance, suggesting that the bean used for chocolate production impacts ratings to some extent.

Broad.Bean.Origin and Specific.Bean.Origin exhibit low to moderate importance, reflecting its potentially less significant role in determining the ratings of the chocolate bars. One consideration is having two variables in the model might cause redundance, since Broad.Bean.Origin already captures a higher-level geographic influence.

3. Data Preprocessing, Model Selection, and Methodology

a) Target Variable: 'Rating Category'

The original variable 'Rating' was transformed into 'Rating Category,' where ratings below 3 are categorized as 'Unsatisfactory,' and ratings of 3 or above are labeled as 'Satisfactory.' This transformation shifts the focus from predicting precise numerical ratings to determining whether a product meets consumer satisfaction thresholds. By grouping ratings into broader categories, the model delivers more actionable business insights, such as identifying products likely to succeed or require intervention.

b) Numerical Variable: Cocoa.Percent

The feature's importance score of this predictor highlights its strong predictive power for determining whether a chocolate bar will be rated as satisfactory or not. Including this predictor ensures the model captures consumer preference, directly aligning with the task of classifying ratings based on product characteristics.

c) Categorical Variables: 'Company.Location', 'Bean.Type', 'Broad.Bean.Origin', and 'Company Rating Mean'.

High-Cardinality Variables: As discussed above, these categorical variables have a lot of of unique observations, making factoring these variables not a favorable choice since doing that can create a very sparse matrix and therefore increase model complexity. The following approaches were implemented to address this:

- Reclassification for low-frequency categories: For variables such as 'Company.Location', 'Bean.Type', and 'Broad.Bean.Origin', categories with frequency below a specific threshold were reclassified as 'Other'. (Refer to Appendix 5 for details on reclassification)
- Target encoding for 'Company Maker': Since the maker of the chocolate bar holds moderate importance in determining the ratings, this variable was target encoded using the company average rating with cross validation to avoid data leakage. The newly created variable is called 'Company Rating Mean'.

Feature Engineering: To improve the model's predictive power, additional variables were constructed and evaluated for their contribution to accuracy. After careful analysis, only one variable was included in the final model, 'Company_Popularity'. This feature captures the frequency of a company's occurrence in the dataset. The underlying assumption is that companies with higher frequencies tend to have more products available, indicating broader market exposure, while less frequent companies may cater to niche audiences. Companies were categorized into four tiers: Low, Low to Medium, Medium, and High popularity. Including this variable enables the model to assess whether a company's brand identity and presence influence consumer acceptance.

These chosen predictors capture important information that significantly enhance the model's predictive capabilities.

For this classification task, two machine learning methods were evaluated: Random Forest and Gradient Boosting. Their performances were assessed using accuracy, balanced accuracy, and F1 scores as key evaluation metrics. Given the class imbalance in the dataset, emphasis was placed on the F1 score during hyperparameter tuning, as it provides a more comprehensive measure of model performance by balancing precision and recall.

4. Result Discussions

The best-performing model was selected based on its ability to generalize well across evaluation metrics, striking a balance between high overall accuracy and the capacity to handle imbalanced class distributions. This ensures the model performs well in real-world scenarios where class imbalance is common. The Random Forest model emerged as the best-performing one. For the comparison between Random Forest and Gradient Boosting models' performance, please refer to Appendix 6.

a) Evaluation Metrics:

For a detailed summary of the metrics, please refer to Appendix 7

- Accuracy (79.05%): Approximately 79% of predictions are correct, indicating that the model performs well overall. However, other metrics need to be examined to get a full picture of model performance.
- Balanced Accuracy (62.75%): This suggests that the model is imbalanced and favor the majority class (Satisfactory' chocolate bars) over the minority class ('Unsatisfactory' chocolate bars). Balanced accuracy is derived from the average of sensitivity and specificity:
    - Sensitivity (30.34%): The model identifies only 30% of unsatisfactory chocolate bars correctly, indicating a struggle to detect the minority class.
    - Specificity (95.17%): The model performs well at identify chocolate bars with satisfactory ratings, correctly classifying 95% of these cases.
- Positive Predicted Value (67.5%): When the model predicts a chocolate bar to be "Unsatisfactory," it is correct 67.5% of the time. While moderately good, this value is undermined by the low sensitivity.
- Negative Predicted Value (80.5%): When the model predicts a chocolate bar to be "Satisfactory," it is correct 80.5% of the time.

The model's strong performance in specificity, or identifying chocolate bars with ratings over 3, highlights its potential to predict successful product based on features like cocoa percentage, sourcing origin, and manufacturers. However, its limited ability to accurately identify unsatisfactory chocolate bars indicates opportunities for further improvement.

b) Feature Importance:

For a detailed summary of the feature importance, please refer to Appendix 8.

To assess the contribution of each predictor to the model's performance, feature importance was evaluated using Mean Decrease Accuracy and Mean Decrease Gini. These metrics reveal how each feature contributes towards making accurate predictions and better splits during model development:

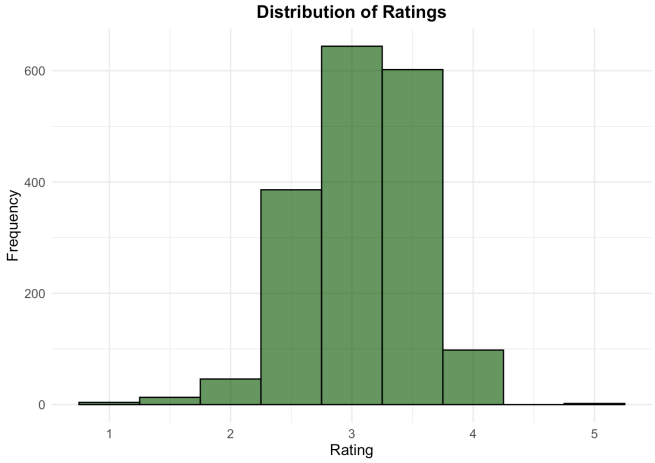
- Company Rating Mean, or a company's average rating, is the most important feature across all metrics, indicating that a company's historical performance strongly predict the acceptance of future products.
- Broad Bean Origin, or the country where the cocoa beans are sourced, ranked 1st in Gini Index, indicating that the origin of the cocoa beans is a key differentiator. Since different countries have different climates and quality standards, cocoa beans produced in certain region will have more attractive flavor profiles.
- Cocoa Percent moderately affects the model's accuracy and decision splits. This suggests that certain cocoa percentages resonate more positively with consumers.

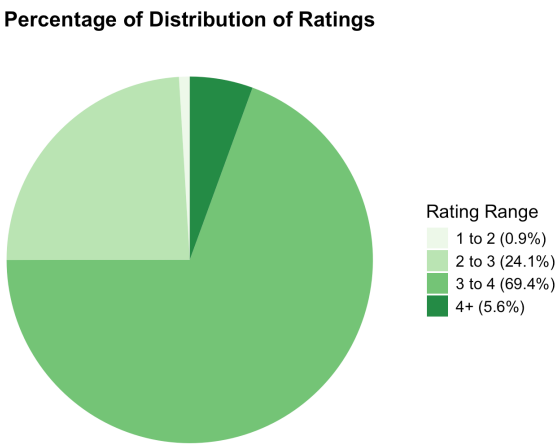
- Company Popularity ranks high in its predictive power. This means that the popularity of a company can impact how well received its products are.
- Bean Type is low in importance, suggesting minor correlation between the types of bean used and the reception of the chocolate bars
- Company Location has a moderate impact on deciding whether the chocolate bar will be rated satisfactory or not. This in understandable since different countries have access to different ingredients, follow different quality standards, and cater to regional preferences.

A strong correlation between the Company Rating Mean and predicted sentiment underscores the importance of maintaining consistent product quality. To achieve satisfactory ratings, companies must focus on delivering products that align with consumer expectations, by either strategically sourcing cocoa beans from specific regions or adjusting cocoa percentages to match consumer preferences, or both. This targeted approach can help businesses better cater to market demands and strengthen their brand reputation.

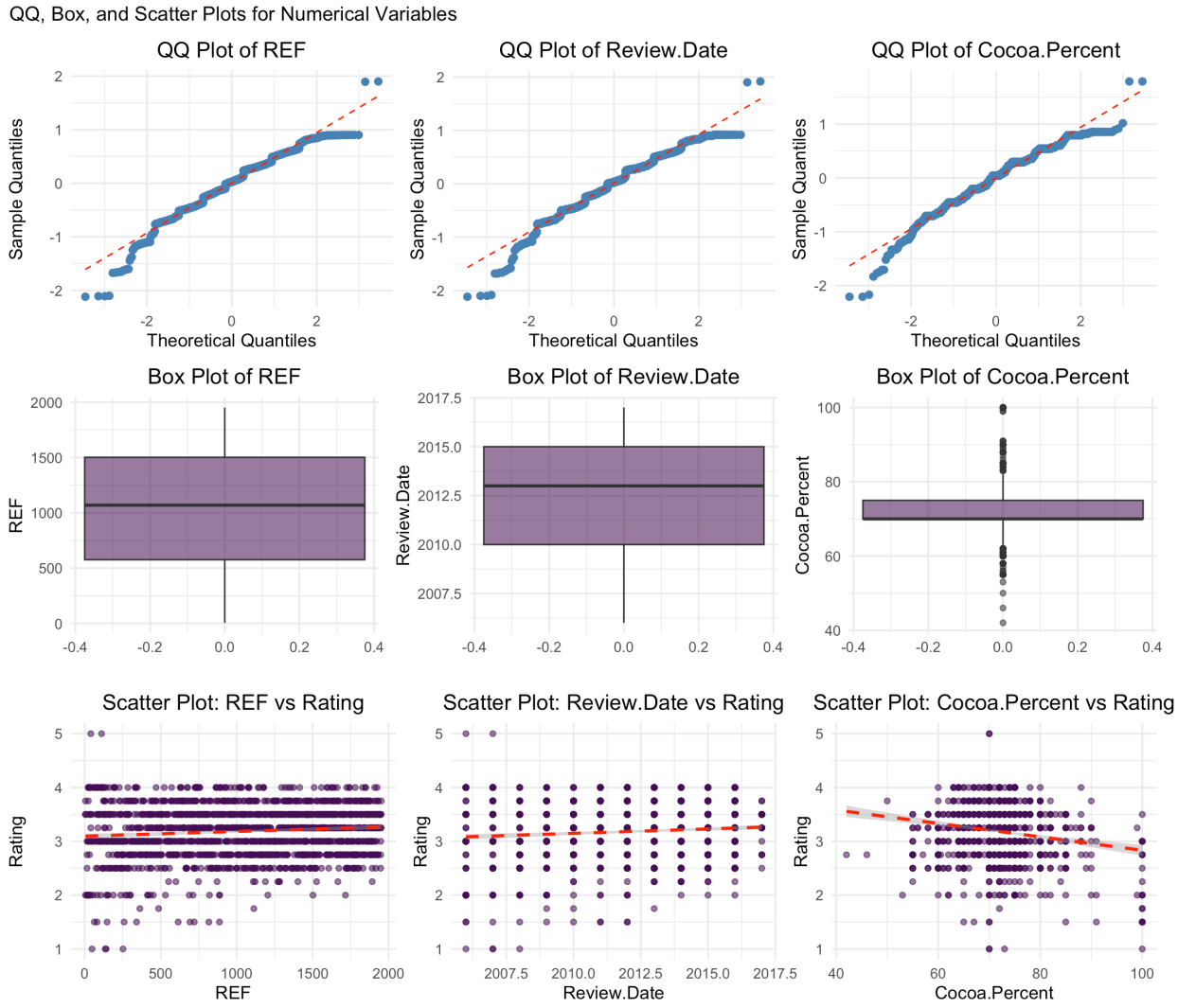
5. Appendices

Appendix 1: Rating Distributions



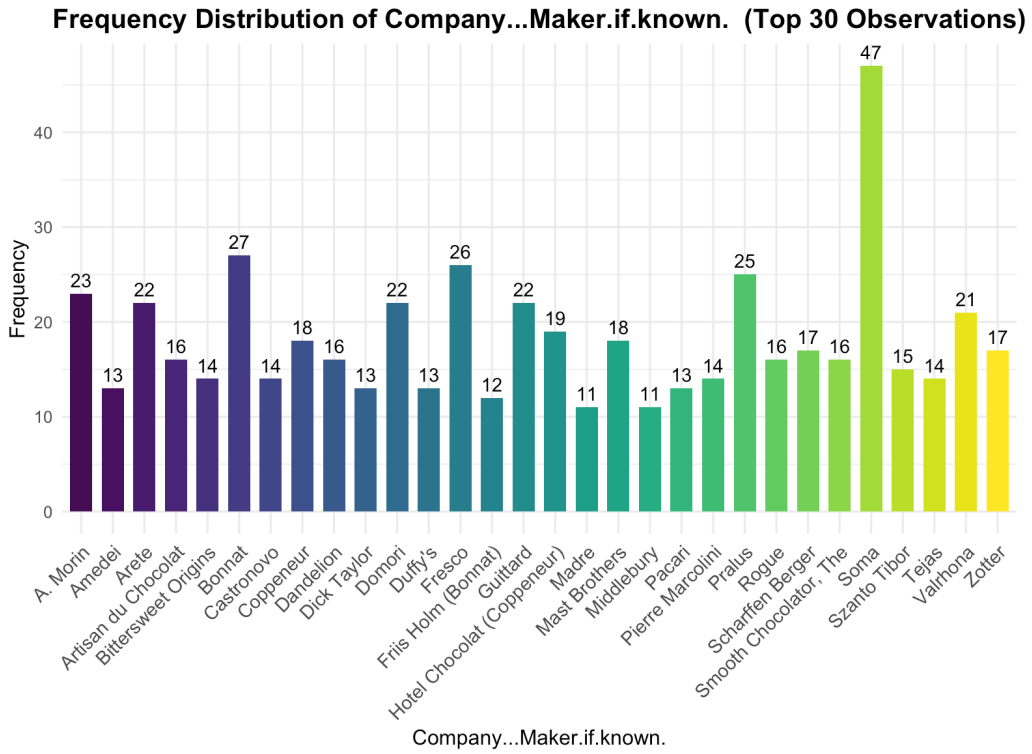


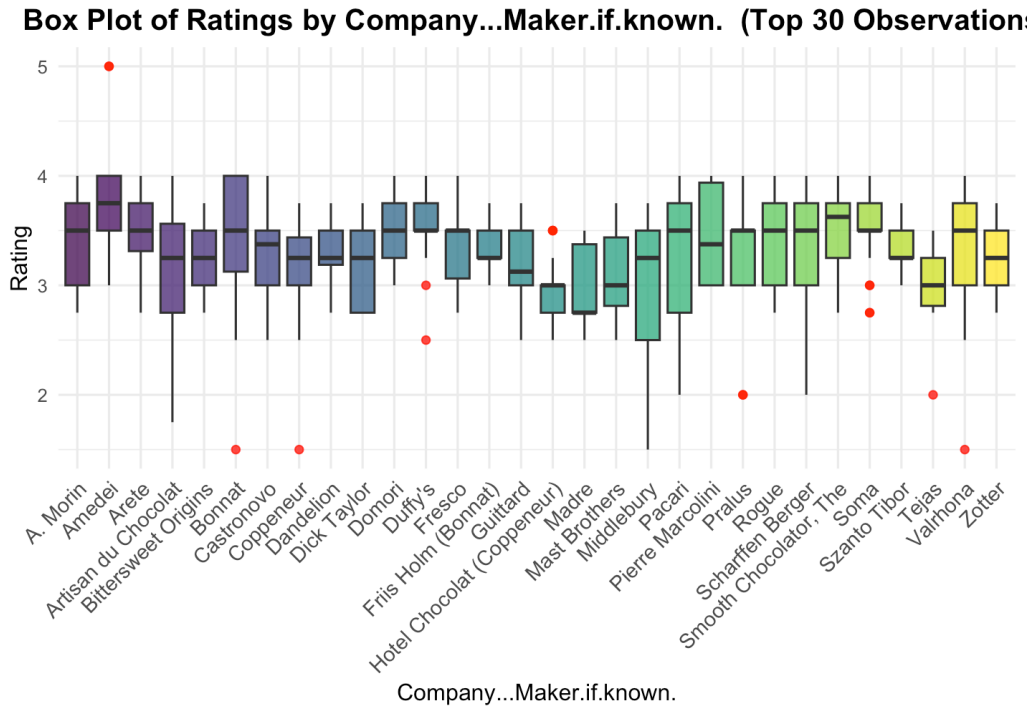
Appendix 2: Numerical Variables Distributions and Relationships with Ratings



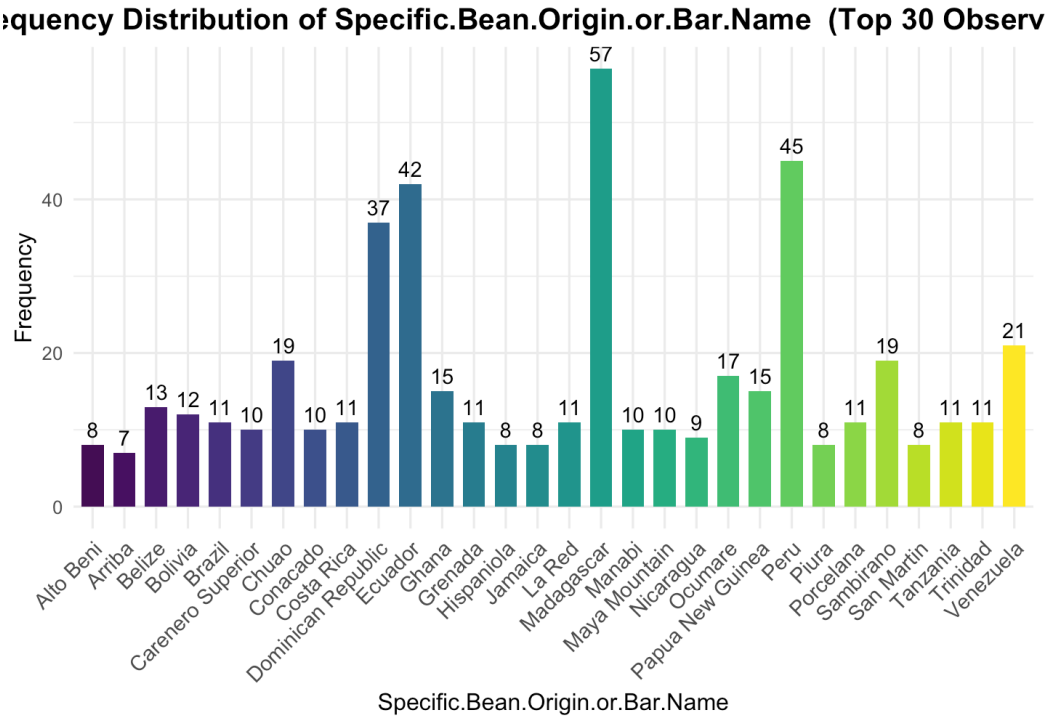
Appendix 3: Categorical Variables Distributions and Relationships with Ratings

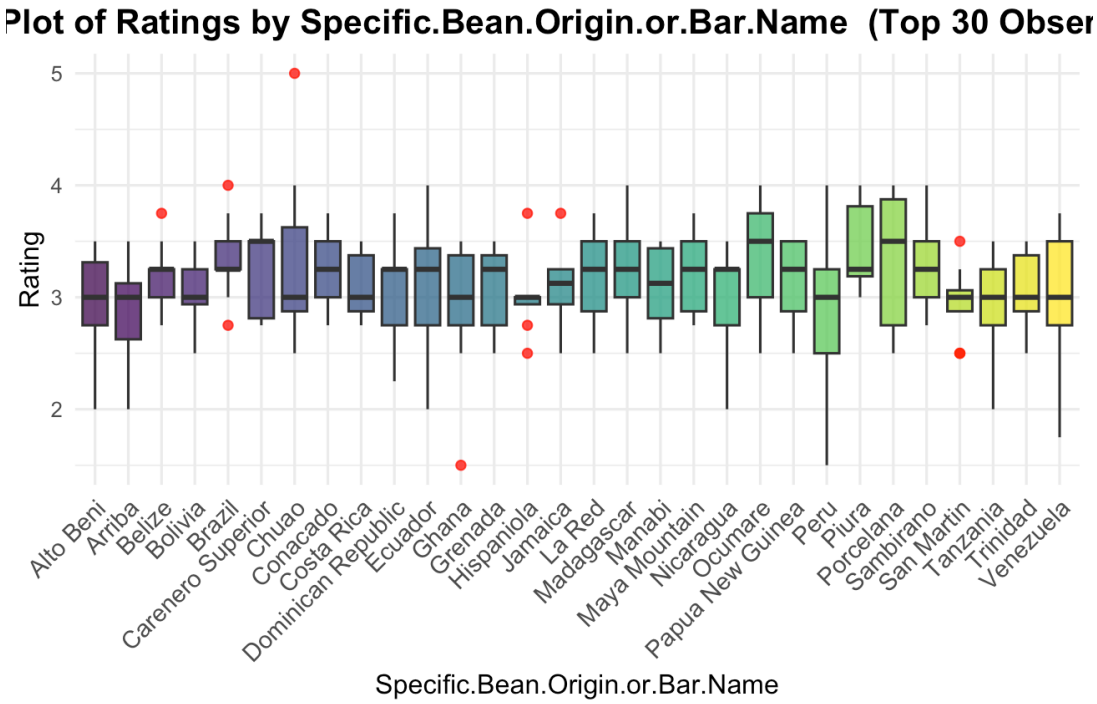
1. Appendix 3.1: Company Maker, or Manufacturer



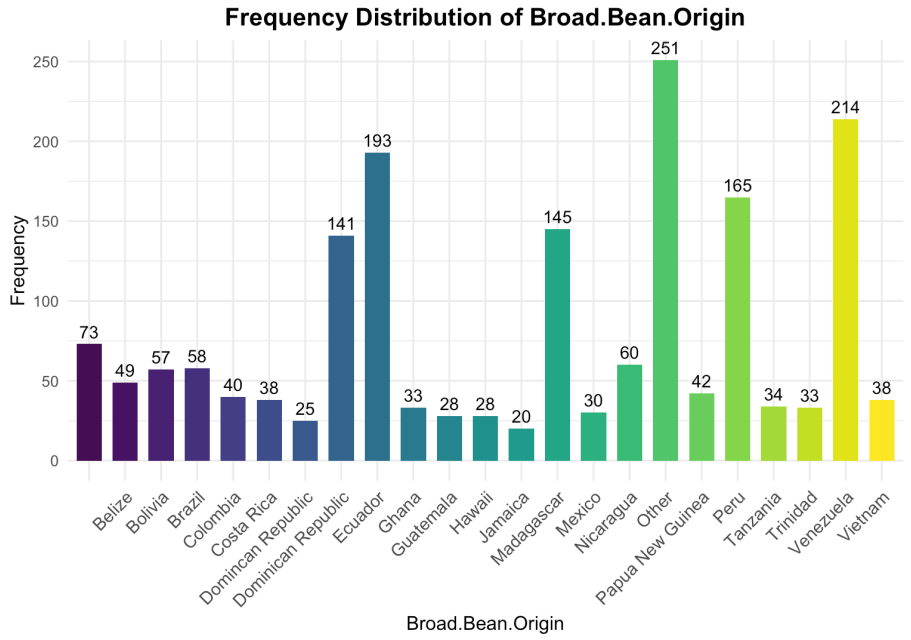


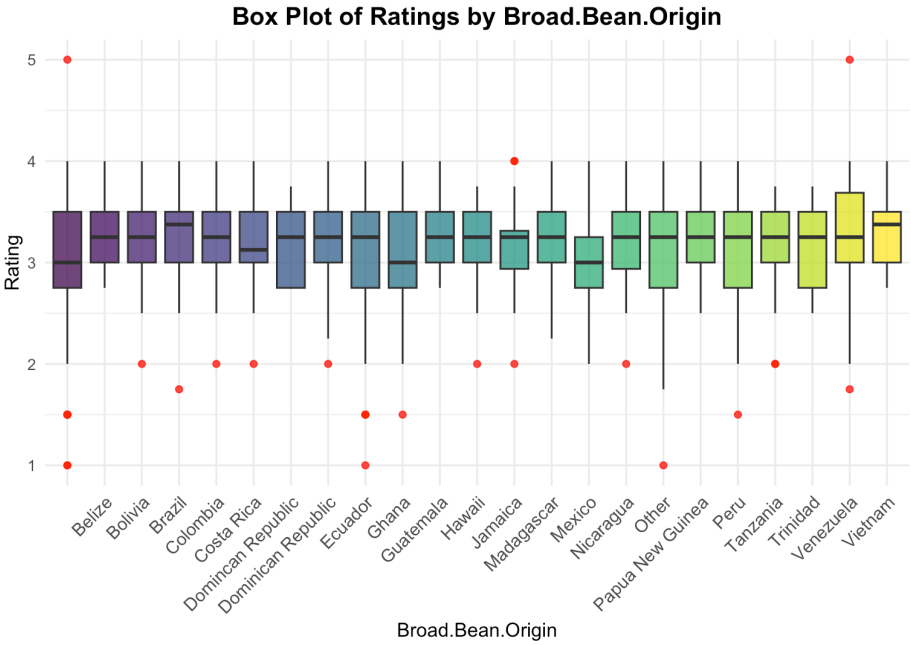
2. Appendix 3.2: Specific Bean Origin



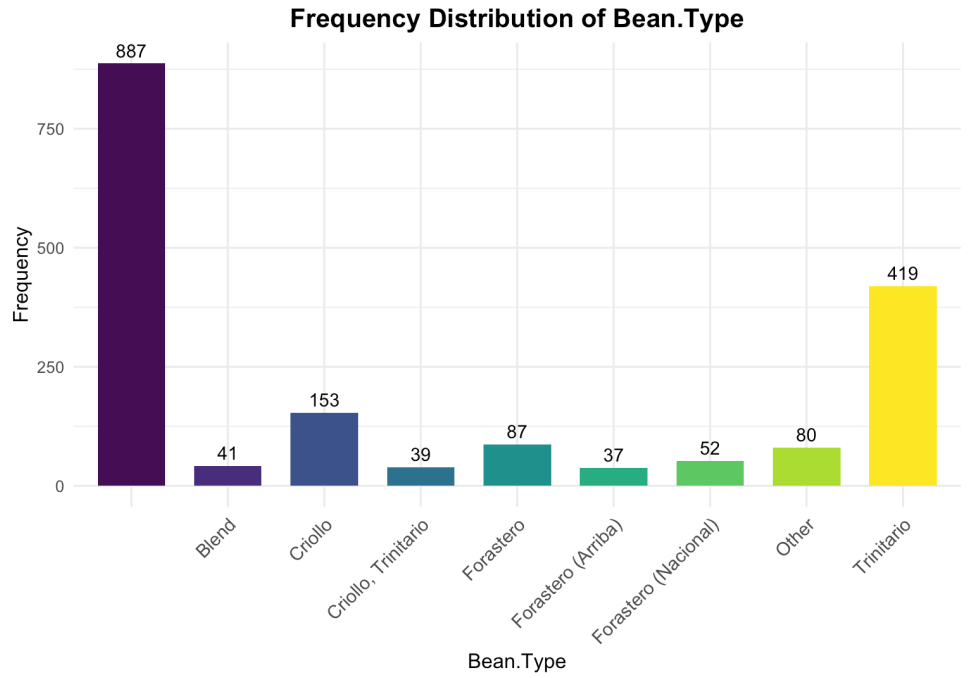


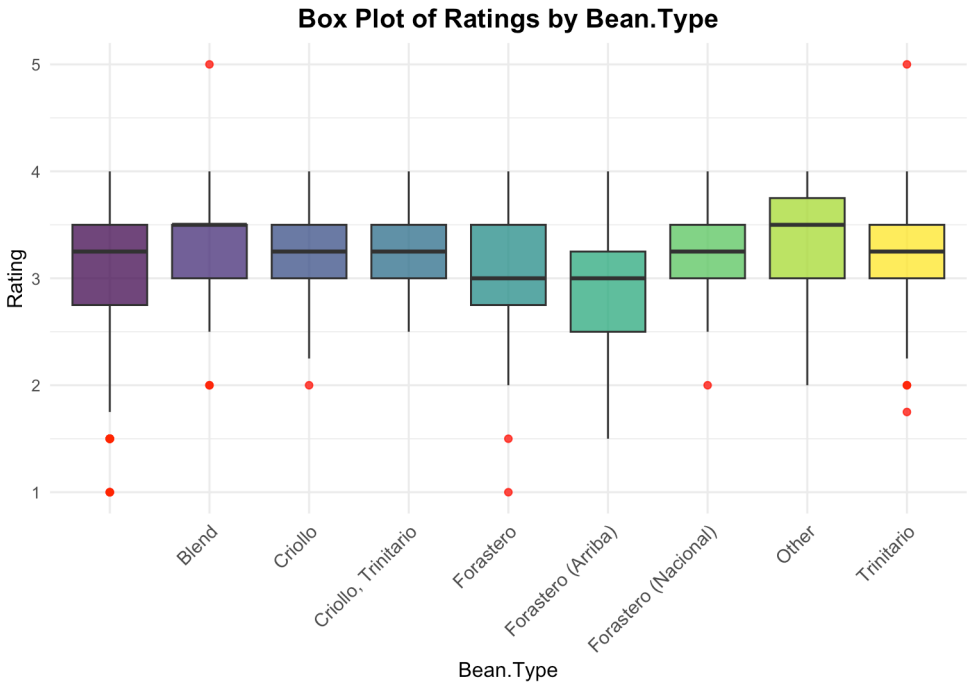
3. Appendix 3.3: Broad Bean Origin



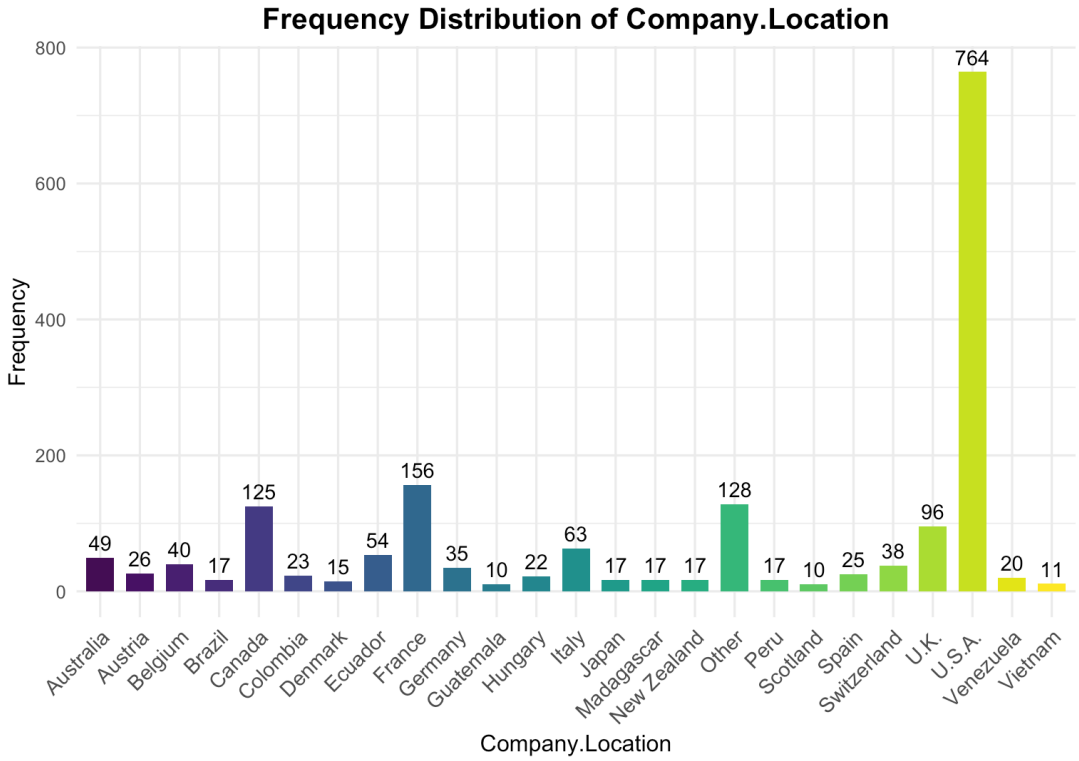


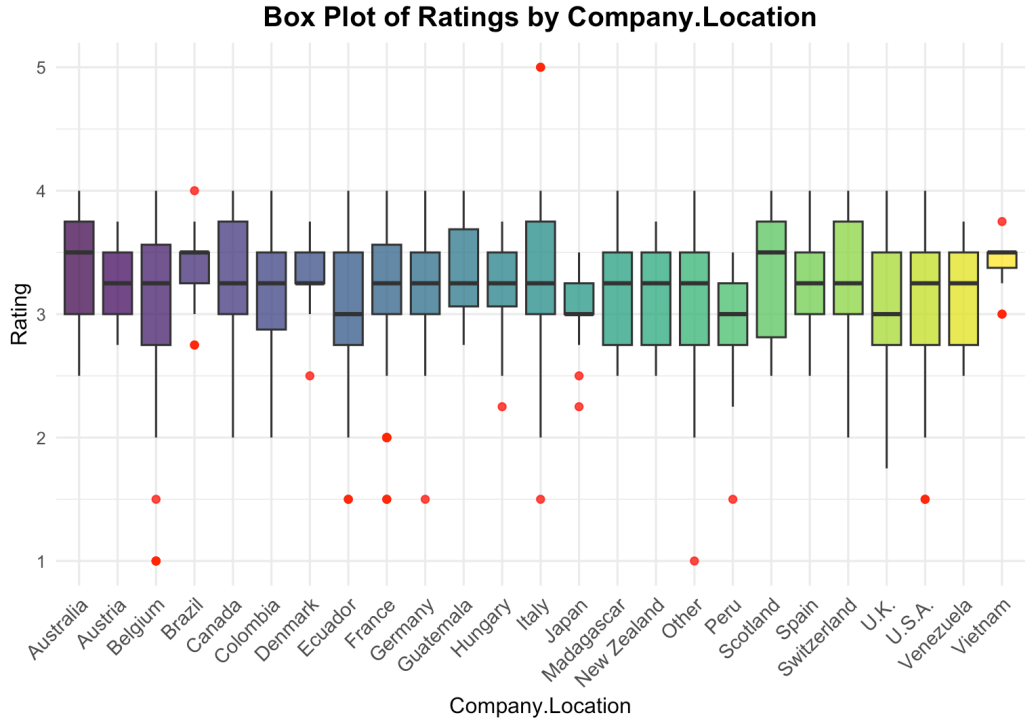
4. Appendix 3.4: Bean Type



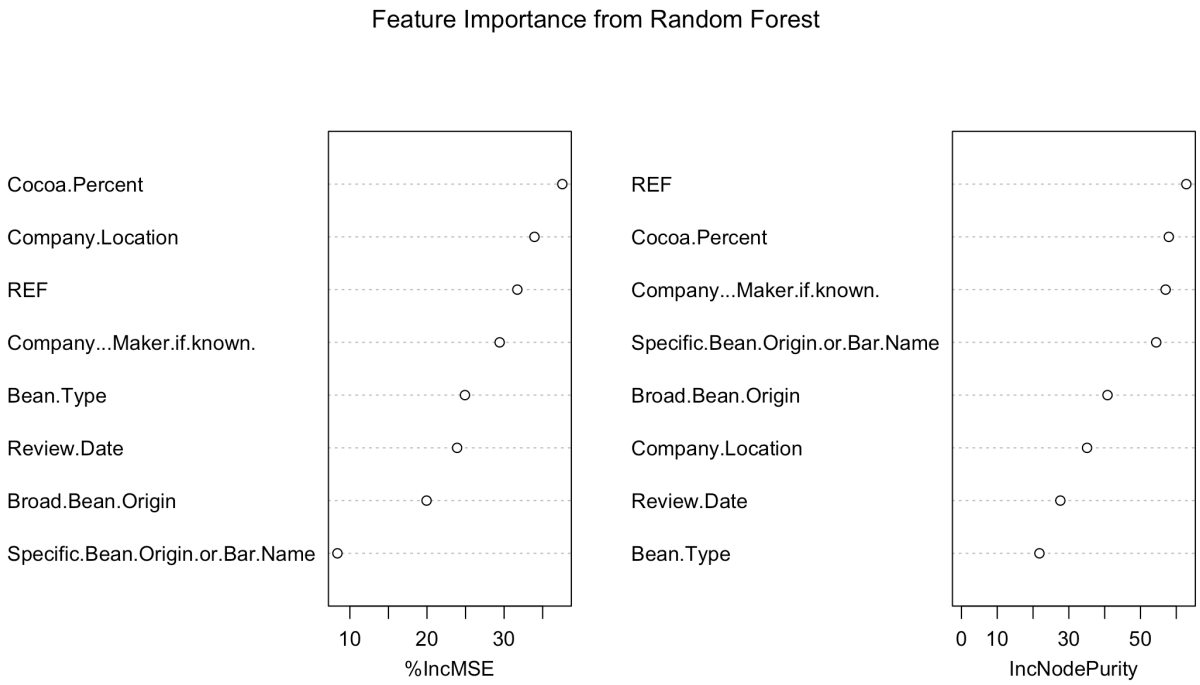


5. Appendix 3.5: Company Location





Appendix 4: Feature Importance Results from Random Forest



Appendix 5: Reclassification for Low-Frequency Categories

Variable	Threshold	Description
Company Location	5	If there are less than 5 unique companies in a location/country, the company location is classified as 'Other'.
Broad Bean Origin	20	If the origin of the bean is not as popular (frequency is less than 20), the origin is classified as 'Other'.
Bean Type	11	If the bean is a rare type (frequency is less than 11), the bean is classified as 'Other'.

Appendix 6: Comparison between RF and GB model performance

Model	Accuracy	Balanced Accuracy	F1 Score	Parameters
GBM (baseline)	0.743	0.632	0.432	Default parameters
GBM (tuned)	0.737	0.622	0.427	Number of trees = 2000 Interaction depth = 5 Shrinkage = 0.1 Bag fraction = 0.5
Random Forest (baseline)	**0.791**	**0.628**	**0.417**	**Default parameters**
Random Forest (tuned)	0.76	0.569	0.283	Mtry = 4 Number of trees = 500 Node size = 10 Max nodes = 30

Appendix 7: Model Performance Results

Model Performance Metrics

Metric	Value
Accuracy	0.7905
95% CI	(0.7446, 0.8315)
No Information Rate	0.7514
P-Value [Acc > NIR]	0.0474
Kappa	0.3126
Mcnemar's Test P-Value	2.98e-08
Sensitivity	0.3033
Specificity	0.9517
Positive Predictive Value	0.675
Negative Predictive Value	0.8050
Prevalence	0.2486
Detection Rate	0.0754
Detection Prevalence	0.1117
Balanced Accuracy	0.6275

Appendix 8: Model Feature Importance

Model Feature Importance Table

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Company.Popularity	13.00	9.35	16.04	20.10
Company.Rating.Mean	26.18	20.34	31.85	79.14
Cocoa.Percent	12.87	7.60	13.43	50.97
Broad.Bean.Origin	12.89	6.60	12.52	80.16
Bean.Type	9.71	3.56	8.96	31.86
Company.Location	16.89	3.70	13.26	59.37