## MGSC662 Final Project: Decoding the Art of Chocolate Ratings

Maggie Huang - 261229497

### Introduction

Chocolate, a universally cherished indulgence, offers a spectrum of flavors and qualities that captivate the senses. The "Chocolate Bar Ratings" dataset, curated by Brady Brelinski of the Manhattan Chocolate Society, compiles expert evaluations of over 1,700 chocolate bars, detailing attributes such as company, cocoa percentage, and origin.

This project aims to uncover relationships and patterns within the chocolate bar ratings dataset and offer insights into the attributes that contribute to higher quality assessments. Specifically, I will harness the power of random forest algorithm to predict chocolate bar ratings. Random forests, a powerful ensemble learning method, construct multiple decision trees during training and output the average prediction for regression tasks. It excels at capturing non-linear relationships and handling mixed datatypes. Additionally, they provide valuable insights by ranking feature importance. Based on my model results, the top three factors influencing chocolate bar ratings are the cocoa percentage, the specific origin of the beans, and the company that crafted the chocolate bar.

In addition to random forest, I will apply the k-prototypes clustering algorithm to address the dataset's mixed data types. K-prototypes extends the k-means algorithm to accommodate mixed data types by combining the Euclidean distance metric for numerical data with a dissimilarity measure for categorical data. This method enables the grouping of chocolate bars into clusters that share similar characteristics across both numerical and categorical attributes. However, interpreting clusters, particularly in high-dimensional spaces, can be challenging. As a result, the insights derived from clustering models are limited.

### Data Cleaning Overview

The dataset was cleaned and prepared to ensure it was accurate, consistent, and ready for analysis. The dataset was checked for duplicates and missing values. ‘broad\_bean\_origin’ has a few missing values, which were all removed. However, ‘bean\_type’ column exhibited nearly 50% missing values. This high level of missingness limited the extent to which analyses involving bean\_type could be conducted reliably.

For the exploratory data alaysis (EDA) and random forest, the missing values in ‘bean\_type’ were retained. Random forest is able to handle missing values inherently. Conversely, k-prototype clustering requires complete datasets for distance calculations; therefore, missing values in bean\_type were imputed using the K-Nearest Neighbors (KNN) imputation method with K = 5 by default.

The `broad\_bean\_origin` column, which contained various inconsistencies such as abbreviations, misspellings, and multiple entries, was thoroughly cleaned and standardized. Entries with multiple countries were split into separate rows, and inconsistent names were replaced with standardized ones.

### Exploratory Data Analysis (EDA)

#### Geographical Distribution of Chocolate Companies and Cocoa Beans (Graphs 1 & 2):

Most chocolate production companies are concentrated in the U.S., Canada, France, Italy, Belgium, Australia, and Ecuador. The highest concentration is in South America, particularly countries like Venezuela, Ecuador, Peru, etc. This aligns with the fact that South America is a major producer of cocoa due to its favorable climate and history in cocoa cultivation.

#### IMPORTANT - Ratings and Sample Size Considerations (Graphs 3–5):

Average ratings must be considered in context with sample size. The circle size on graphs 3 – 5, represents sample size, which impacts the reliability of the average rating. Larger circles indicate more samples, making their average ratings more statistically robust. Smaller circles reflect fewer samples, meaning their higher ratings may be less robust and subject to variability. For example, although bean type "Criollo (Ocumare 67)" has the highest average rating, there is only one review for it. Hence, we cannot confidently say that it is the best type of bean.

#### Bean Origin and Ratings (Graph 3):

Bean origins such as Java, Haiti, and Solomon Islands appear at the top with average ratings around 3.5. Puerto Rico has the lowest average rating of 2.5.

#### Company and Ratings (Graphs 4.1–4.2):

Top-rated companies include Tobago Estate (Pralus) at 4.0 and Heirloom Cacao Preservation (Zorzal) and Ocelot at 3.875. Lower-rated companies, such as Callebaut and Machu Picchu Trading Co., average 1.75.

#### Bean Types and Ratings (Graph 5):

Criollo beans (e.g., "Criollo (Ocumare 67)") appear to consistently receive higher average ratings compared to other types. Forastero beans (e.g., "Forastero (Arriba)") and their blends tend to cluster in the lower-rated portion of the graph. This could be due to their more common use in mass-market chocolate, as they are less expensive and have less complexity in flavor.

#### Cocoa Percentage and Ratings (Graphs 6 & 7):

Most chocolates fall in the 60%–80% cocoa range, with average ratings around 3.0–4.0. Bars at 100% cocoa often score lower (~2.0) due to bitterness. The highest-rated bars generally contain about 70% cocoa.

#### Ratings Over Time (Graph 8):

Ratings appear relatively stable over the years, with median ratings consistently hovering around 3.0–3.5. There is no strong evidence of either significant improvement or decline in chocolate quality over time based on the ratings.

#### Review Counts (Graphs 9 & 10):

The most reviewed chocolates cluster around a cocoa percentage of 70%, with a clear peak. The majority of reviews fall within the 60%-80% cocoa range, indicating that chocolates in this range are likely the most popular in the market. Soma dominates with the highest number of reviews of over 60. Companies with higher review counts likely have more exposure, possibly through effective marketing or reputation.

### Model Methodologies – Random Forest

To develop and evaluate my predictive model, I employed a Random Forest algorithm from the ranger package and a systematic approach to hyperparameter tuning using the caret package. The overarching goal was to identify the combination of model parameters that yielded the best predictive performance, as measured by the Root Mean Squared Error (RMSE).

#### Data Splitting:

I began by partitioning the cleaned dataset (data\_cleaned) into three subsets: training (60%), validation (20%), and testing (20%). The training set was used to fit the model, the validation set guided hyperparameter tuning and model selection, and the test set served as a final, independent evaluation of model performance.

#### Hyperparameter Tuning:

I focused on two critical hyperparameters for the Random Forest model:

* mtry: The number of variables to consider at each split.
* min.node.size: The minimum number of observations in each terminal node.

I created a grid of candidate values for these parameters and iteratively trained a series of Random Forest models on the training set. For each candidate combination, we evaluated the model’s performance on the validation set.

#### Best Hyperparameters:

The Random Forest model achieved optimal performance with an mtry value of 4, meaning it considered 4 randomly selected predictor variables at each split, and a min.node.size of 3, requiring each terminal node to have at least 3 data points. These hyperparameters strike a balance in the bias-variance tradeoff. A smaller mtry would introduce higher bias by limiting flexibility in splits, while a larger mtry could lead to overfitting by reducing generalizability. Similarly, a smaller min.node.size might overfit the model by creating overly complex trees, whereas a larger value could oversimplify and miss fine-grained data patterns.

The validation RMSE of 0.4273 indicates the model's average error in predicting rating values, with a lower RMSE reflecting better accuracy. This result suggests that the model performs reasonably well on the validation dataset and is reliable for rating predictions.

#### Final Model Training:

Once I identified the best hyperparameters (mtry = 4, min.node.size = 3), we retrained the Random Forest model using the combined training and validation data to maximize the amount of data available for learning.

#### Model Evaluation:

I assessed the final model’s performance on the held-out test set, which played no role in model building or selection. The test RMSE is 0.4358859, which is very close to the validation RMSE. This means that our model performs relatively well on new, unseen data.

#### Feature Importance:

To gain insights into the factors most influencing the model’s predictions, I examined variable importance scores (graph 11). For the interpretations of feature importance, refer to the Models Interpretations section.

### Model Methodologies – K-Prototype Clustering

#### Handling Missing Data via kNN Imputation

One critical step in preparing my dataset for clustering was dealing with missing values. Most clustering algorithms cannot handle incomplete observations, so I needed an effective imputation strategy. I chose k-Nearest Neighbors (kNN) imputation, which leverages the similarity of data points to infer missing values.

The kNN imputation process works as follows: For each data point with a missing value, I identified the k nearest neighbors in the feature space, using Euclidean distance for numeric variables and appropriate similarity measures for categorical features. If I do not explicitly set k, the algorithm defaults to k=5, meaning it uses the five most similar data points as a reference. The missing values are then replaced with a suitable summary statistic—such as the mean, median, or mode—calculated from these neighbors. This approach helped me maintain the dataset’s inherent structure while ensuring that I did not lose valuable information.

#### Choosing the Clustering Algorithm: k-Prototypes

My dataset comprised a mixture of categorical and numeric variables. Methods such as k-means are best suited for numeric data, while algorithms like k-modes handle categorical variables more effectively. To handle both simultaneously, I employed the k-prototypes algorithm. This method combines the strengths of k-means and k-modes, enabling me to consider both types of variables in a single clustering framework.

#### Determining the Optimal Number of Clusters

Selecting the right number of clusters (k) is crucial for producing meaningful groupings. To achieve this, I used the silhouette width as a guiding metric. The silhouette width measures how well each data point fits into its assigned cluster compared to other potential clusters. Higher values generally indicate clearer, more cohesive clustering solutions.

To find the optimal k, I:

* Ran the clustering algorithm for multiple values of k (from 2 to 10).
* Computed the silhouette width for each configuration.
* Identified the k value that maximized the average silhouette width.

In this case, the highest average silhouette width occurred around k=4, so I selected four clusters (graph 12).

#### Final Clustering

After determining the optimal number of clusters, I re-ran the k-prototypes algorithm with k=4. Cluster 1-4 has the size of 143, 476, 723, 444 observations respectively. Additionally, I summarized the mean values of the numeric variables (graphs 13 & 14) and examined the most common categories (modes) for the categorical variables within each cluster (graphs 15 - 20).

### Model Results

#### Random Forest Feature Importance

cocoa\_percent (20.75%) is the most important feature for predicting chocolate ratings. This makes sense, as the cocoa percentage directly influences the flavor intensity, bitterness, and richness of the chocolate. In the EDA, I found that ~70% cocoa has the highest average rating. This is likely due to that around 70% cocoa is often considered the "sweet spot" because it balances the rich, complex flavor of dark chocolate with enough sweetness to avoid being overly bitter. This percentage appeals to both dark chocolate enthusiasts and casual consumers.

Specific Bean Origin (18.32%) is the second most important feature. This is important as the specific region where the cocoa beans are grown determines the flavor profile due to differences in climate, soil, and farming practices (terroir). For example, the most highly rated bean origin is Java. This makes sense since Java's cocoa beans are renowned for their distinctive flavor profile, characterized by mild cocoa notes complemented by caramel and refreshing acidic hints of yellow fruits like banana and pineapple. This unique taste results from Java's rich volcanic soil and tropical climate, which contribute to the exceptional quality of its cocoa beans (1).

The company (18.27%) producing the chocolate ranks third in importance, almost tied with the bean origin. This reflects the influence of brand reputation, production methods, and consistency in quality. Companies with a history of high-quality production tend to receive better ratings. For example, the most highly rated company is Tobago Estate (Pralus). Tobago Estate uses high-quality Trinitario cocoa beans from their own estate in Tobago, known for rich and complex flavors with hints of tropical fruit and nuts. The company combines traditional cocoa cultivation methods with modern processing techniques, emphasizing sustainability and quality. Their chocolates have received international recognition, winning awards at various chocolate and fine food competition (2).

#### Clustering Patterns

There is no sharp distinction between the 4 clusters. For the numerical variables rating and cocoa\_percent, we can see on the boxplots (graphs 13 & 14) that their distributions across the 4 clusters are relatively uniform. For all the categorical variables (graphs 15 – 20), we see the most variations across clusters for company\_location. Specifically, cluster 1 is entirely made up of companies located in France. Cluster 2 and 4 are made up of mostly of companies located in the USA. However, interpreting clusters, particularly in high-dimensional spaces, can be challenging. It's not too clear what insights we could get in this case, since there are no sharply distinct clusters formed.

### Business Insights

Target the "Sweet Spot" of ~70% Cocoa

Chocolate producers could refine their product lines to highlight offerings within the 70% range, potentially expanding their 60%-80% cocoa portfolios to appeal to a broad segment of consumers who value complexity without overwhelming bitterness.

#### Highlight Premium Origins in Branding and Marketing

Specific bean origin significantly influences chocolate ratings, indicating that terroir and bean quality matter greatly. By sourcing beans from highly regarded origins—such as Java—companies can differentiate their products in a crowded market. Chocolatiers can emphasize the provenance on packaging, in advertising materials, and in tasting notes.

#### Build and Leverage Brand Reputation

The producing company’s identity and track record also heavily influence ratings. Consumers trust brands known for consistent quality, ethical sourcing, and innovative production methods. Businesses should invest in long-term brand development strategies.

### Conclusion

In conclusion, this analysis reveals that cocoa percentage, bean origin, and the producing company are the primary drivers of chocolate ratings. Bars around 70% cocoa strike the ideal balance, while premium bean sources and reputable chocolatiers earn consistently high scores. Although clustering did not yield highly interpretable segments, it confirms the complexity of this market. Overall, these findings can guide producers, marketers, and consumers toward refining product offerings, emphasizing quality origins, and building trusted brand reputations.

### Appendix

A map of the world with different colored countries/regions

Description automatically generated

Graph 1: Geographical Distribution of Chocolate Companies (‘company’)

A map of the world with different colored countries/regions

Description automatically generated

Graph 2: Geographical Distribution of Cocoa Beans (‘broad\_bean\_origin’)

A graph with colored dots

Description automatically generated

Graph 3: Average Rating by Broad Bean Origin

A graph with many colored dots and lines

Description automatically generated

Graph 4.1: Average Rating by Top 30 Companies (Descending)

A graph with dots and lines

Description automatically generated

Graph 4.1: Average Rating by Bottom 30 Companies (Ascending)

A graph with colored dots and lines

Description automatically generated

Graph 5: Average Rating by Bean Type

A screen shot of a graph

Description automatically generated

Graph 6: Rating by Cocoa Percentage and Broad Bean Origin

A graph showing a graph of a graph

Description automatically generated with medium confidence

Graph 7: Cocoa Percentage per Rating

A graph with different colored squares

Description automatically generated

Graph 8: Rating by Review Year

A graph with different colored lines

Description automatically generated

Graph 9: Number of Reviews by Cocao Percentage

A graph with blue and white bars

Description automatically generated

Graph 10: Number of Reviews by Company (Top 30)

A graph with blue bars

Description automatically generated

Graph 11: Feature Importance in the Random Forest Model

A graph with lines and numbers

Description automatically generated

Graph 12: Silhouette width for k = 2 – 10

A graph with multiple rows of rectangular objects

Description automatically generated with medium confidence

Graph 13: ‘rating’ Distribution by Cluster

A graph with a row of boxes

Description automatically generated with medium confidence

Graph 14: ‘cocoa\_percent’ Distribution by Cluster

A graph of a company

Description automatically generated with medium confidence

Graph 15: ‘company’ Distribution by Cluster (Top 10 + Other Categories)

A graph of a group of colored bars

Description automatically generated with medium confidence

Graph 16: ‘company\_location’ Distribution by Cluster (Top 10 + Other Categories)

A graph of different colored squares

Description automatically generated

Graph 17: ‘specific\_bean\_origin’ Distribution by Cluster (Top 10 + Other Categories)

A chart with different colors

Description automatically generated

Graph 18: ‘broad\_bean\_origin’ Distribution by Cluster (Top 10 + Other Categories)

A graph of a number of colored squares

Description automatically generated with medium confidence

Graph 19: ‘review\_year’ Distribution by Cluster (Top 10 + Other Categories)

A graph of different colored squares

Description automatically generated with medium confidence

Graph 20: ‘bean\_type’ Distribution by Cluster (Top 10 + Other Categories)

### References

Dataset: <https://www.kaggle.com/code/fangya/chocolate-bar-rating-eda-anova-svm#reference>

1. <https://www.callebaut.com/en-CA/chocolate-cocoa-nuts/JAVA-2B-U75/java>?
2. <https://www.tasteatlas.com/tobago-cocoa-estate?utm_source=chatgpt.com>