Unwrapping Chocolate: Navigating the Chocolate Industry Through Culture and Taste

Introduction

"Candy is dandy, but chocolate is divine!" (Dahl, 1964). Every year, the world consumes over 7 million tons of chocolate, driving a global market worth more than $133 billion (Dame Cacao, 2023; Expert Market Research, 2023). However, behind this beloved sweet treat lies a complex, fragile and highly competitive industry. Cocoa farming, the backbone of chocolate production, is increasingly challenged by unpredictable weather, fluctuating prices, and supply chain disruptions, particularly in key growing regions like West Africa (Wired, 2023). Willy Wonka famously said, "A little nonsense now and then is relished by the wisest men" (Dahl, 1964), but in the world of chocolate, even a little disruption in the supply chain can ripple across the global market. These challenges highlight the urgent need for reliable, high-quality cocoa sources to sustain the industry and ensure its future. At the same time, consumer preferences are also evolving rapidly. Modern chocolate enthusiasts crave more than just traditional flavors—they are drawn to rare, seasonal, and premium offerings. In fact, 74% of consumers report excitement about trying unique and innovative chocolates(Callebaut,2023). Cocoa percentage in chocolate bars has also become a key differentiator, appealing to those seeking bold flavors or health-conscious options. Meanwhile, growing consumer awareness around sustainability has amplified the demand for ethically sourced cocoa, presenting brands with a dual challenge: meeting taste expectations while promoting responsible production practices(Confectionery Production, 2023).

This study explores some of the most pressing aspects of the chocolate industry to address these challenges. It focuses on identifying cocoa-growing regions that consistently deliver high quality and reliability, making them ideal partners for strategic sourcing. Additionally, it examines how regional taste preferences influence chocolate ratings, providing insights to tailor production and marketing strategies. The relationship between cocoa percentage and consumer satisfaction is analyzed to guide product development that balances bold flavors, sustainability, and consumer demand. Lastly, the study predicts which chocolate bars are most likely to resonate with consumers, enabling data-driven decisions to optimize product positioning and market success.

Data Description

Data Cleaning and Reprocessing

The dataset initially contained 1,759 reviews. To ensure the data's integrity, a rigorous cleaning process was implemented. Missing values, comprising less than 1% of the data (2 rows), were removed to maintain consistency without significantly reducing the sample size. Outliers in numerical variables like *Cocoa Percentage* and *Rating* were identified using z-scores with a ± 3 standard deviation threshold. This approach effectively balances the need to remove extreme values that could distort trends while retaining most of the data's natural variability. Specifically, 20 outliers in Cocoa Percentage and 8 overlapping outliers in Rating were flagged and removed, ensuring a more accurate analysis of relationships without undue influence from anomalies. The removal of these outliers is visualized in the *Cocoa Percent Histogram* and *Ratings Histogram* (Figure 5). Redundant columns like *REF* were dropped, and duplicate rows were removed to ensure that companies or bean types were not overrepresented. By addressing missing values, minimizing the impact of outliers, and streamlining (giving simpler names to the variables) the dataset, the risk of skewed or unreliable results was significantly reduced by testing for heterodeskadicity and others collinearity components. This clean, analysis-ready dataset provided a solid foundation for exploring consumer preferences and industry trends confidently.

Exploration of Numerical Variables

Ratings: The *Ratings* variable exhibits a near-normal distribution, with most values clustered between 3.0 and 4.0, indicating that the dataset predominantly represents chocolate products of average to above-average quality. This narrow range underscores the importance for producers to differentiate their offerings within this competitive quality spectrum. Producers should focus on strategies to push products toward the higher end of this range (above 3.5), as these are more likely to appeal to discerning consumers and generate favorable reviews. The distribution of ratings is visualized in the *Ratings Histogram* Figure 5.

Cocoa Percentage: *Cocoa Percentage* is heavily concentrated around 70%, reflecting the dominance of dark chocolate products. Histogram analysis reveals that percentages between 79% and 95% achieve the highest average ratings, peaking at approximately 3.21. In contrast, percentages exceeding 95% see a sharp decline to an average rating of 2.4, likely due to excessive bitterness. Similarly, lower cocoa percentages (below 63%) are associated with reduced ratings, suggesting a preference for balanced flavors over extremes. These findings highlight an "optimal cocoa range" of 70%-85%, where producers can maximize consumer satisfaction while maintaining product versatility. The distribution of cocoa percentages is illustrated in the *Cocoa Percent Histogram* Figure 5 and Figure 6.

Review Date: *Review dates* range from 2006 to 2017, with 68% of reviews conducted between 2010 and 2015. Temporal trends indicate a slight dip in ratings around 2010, potentially reflecting a temporary market disruption or heightened consumer scrutiny during that period. Ratings gradually improved in subsequent years, suggesting advancements in product quality and adaptation to evolving consumer preferences. The distribution of review dates is shown in the *Review Date Histogram* Figure 6.

The relationships between numerical variables were further explored using a *Correlation Heatmap* and *Residual Plots* to detect potential linear dependencies and heteroscedasticity. As shown in Figure 8, weak correlations were observed between key variables such as *Cocoa Percentage* and *Ratings*, indicating that non-linear modeling techniques may be more effective for uncovering trends. The residual plots also highlight non-linear patterns and variability, particularly in the relationships between *Review Date*, *Cocoa Percentage*, and *Rating* which will be explored later.

Exploration of Categorical Features

Bean Type: *Cocoa bean* types display notable differences in consumer ratings. Criollo and Trinitario beans consistently achieve higher average ratings (3.35 and 3.30, respectively) compared to Forastero beans (3.10). These patterns suggest that premium beans like Criollo, Trinitario or EET (which also rates higher) are perceived more favorably, while blends exhibit greater variability in ratings. This variability could reflect inconsistencies in product quality or diverse flavor profiles, as visualized in the *Average Ratings by Bean Type* plot Figure 9.

Company: The dataset included 393 unique companies, although a small number stood out for their dominance in reviews and performance. Domori and A. Morin emerged as top performers, with median ratings of 3.85 and 3.70, respectively, and consistently narrow interquartile ranges reflecting stable product quality. By comparison, companies like Hotel Chocolat displayed lower rating from 2.8 to 4.5 and some outliers. This variability is visible in the *Distribution of Ratings for Companies with Most Reviews* Figure 10. These observations could hint at potential quality control issues or differences in consumer expectations, especially for brands with broader product portfolios.

Company Location: Producers from 62 locations were represented in the dataset, with Switzerland and the U.S. accounting for 26% and 10% of reviews, respectively. Switzerland stood out for its stable ratings, achieving a median of 3.5. Madagascar, however, exhibited greater variability in ratings, ranging from 2.8 to 4.4, portraying a mix of diverse production standards and consumer perceptions. The *Distribution of Ratings by Continent* Figure 11) and *Average Ratings by Continent* Figure 12highlight regional trends and variability. While geographic origin can shape reputation to some extent, company-specific factors—such as production techniques (cocoa percentage) and branding—were found to have a stronger influence on ratings. This was supported by ANOVA tests, which showed that company effects significantly outweighed location-based influences, highlighting the importance of brand-driven strategies in shaping consumer evaluations.

Model Selection, Methodology and Results

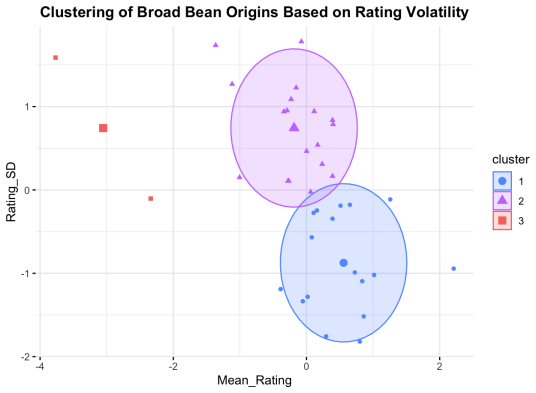
Identifying Stable Cocoa Bean Origins

To identify cocoa bean origins that exhibit consistent ratings over time, **K-means clustering** was applied. This method grouped broad bean origins based on two key metrics: **mean ratings** (representing quality) and **standard deviation** (capturing quality inconsistency). The optimal number of clusters (k) was determined to be 3 using the elbow method, which minimizes within-cluster variance while ensuring clear distinctions between clusters. This approach allowed for the identification of three distinct groups of cocoa bean origins, each characterized by their quality and stability. The clustering results are summarized in Table 1a and Table 2, and the clustering patterns are visualized in Figure 1.

K-means clustering grouped 37 broad bean origins into three distinct clusters based on their mean ratings and standard deviations. Cluster 1, which includes Ghana and Mexico, exhibited moderate quality with low quality inconsistency, making these regions ideal for long-term partnerships focused on consistent supply. To further enhance value, manufacturers could invest in targeted training and resources for farmers in these regions, ensuring not only stable quality but also opportunities to differentiate their beans through certifications or specialty products. Cluster 2, represented solely by São Tomé & Príncipe, demonstrated high quality inconsistency despite optimisitc mean ratings, as highlighted in Figure 1 and Figure 13.This shows the need for better quality control to tackle the wide range of ratings and help ensure more consistent chocolate quality. Cluster 3, encompassing Belize and Brazil, combined high ratings with low quality inconsistency, positioning these origins as ideal for premium chocolate production.

Cluster	Mean	SD	Count
1	2.93	0.54	8
2	3.18	1.02	1
3	3.23	0.42	28

(a) K-means Clustering Results for Cocoa Bean Origins



(b) Clustering of Broad Bean Origins Based on Rating Volatility

Figure 1: K-means clustering results alongside visualization of broad bean origins.

Exploring Regional Taste Preferences

To explore regional taste preferences and uncover potential cultural biases, **Linear Discriminant Analysis (LDA)** was initially employed. LDA is a supervised learning method that projects data onto a lower-dimensional space, maximizing the separation between predefined classes. The formula used for the first LDA model is as follows:

$$\text{lda}(Continent \sim Cocoa\_Percent + Rating)$$

where:

- **Continent:** The categorical dependent variable representing continents.
- **Cocoa_Percent:** The percentage of cocoa in the chocolate, a numerical predictor.
- **Rating:** The average rating of the chocolate, another numerical predictor.

The first LDA model classified data at the **continent level** using predictors such as **Cocoa_Percent** and **Rating**, achieving moderate accuracy (0.61). However, the second LDA model, which aimed to classify individual **South American countries**, struggled due to imbalanced class sizes and small sample sizes (accuracy 0.3). These challenges necessitated the adoption of **Hierarchical Clustering**, an unsupervised method that iteratively merges clusters to reveal nuanced relationships without requiring predefined class labels. Hierarchical clustering was performed using Ward's method, which minimizes within-cluster variance at each step. This method successfully grouped countries based on standardized **Cocoa_Percent** and **Rating**, revealing patterns missed by LDA. The clustering results are summarized in Table 3, and visualized using the hierarchical clustering scatter plot Figure 2 .

Regional taste preferences were examined using LDA and hierarchical clustering. European and Oceanian consumers displayed a preference for chocolates with higher cocoa content, reflected in their average ratings of 3.31 and 3.21, respectively. In contrast, South American markets exhibited diverse preferences, with distinct clusters emerging for countries like Ecuador and Peru. These clusters features are explained in Table 3 and Figure 2, which group South American countries based on their standardized cocoa percentages and ratings.

Cluster 1 prominently includes Brazil, Colombia and Venezuela, representing countries with consistent cocoa quality and ratings. Cluster 2 is smaller, focusing on Bolivia and Peru, which show moderate distinctions in cocoa characteristics. Cluster 3 captures countries like Argentina and larger representations for Ecuador, indicating diversity in cocoa quality and rating patterns within the cluster.

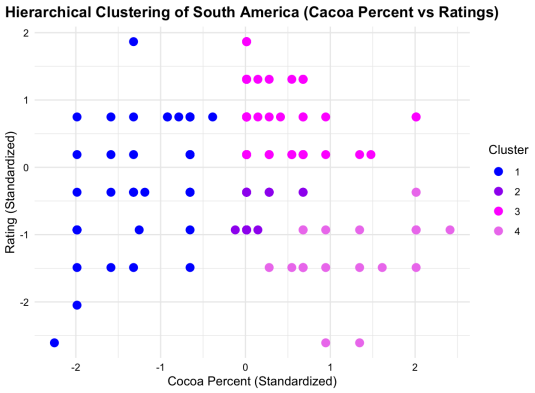


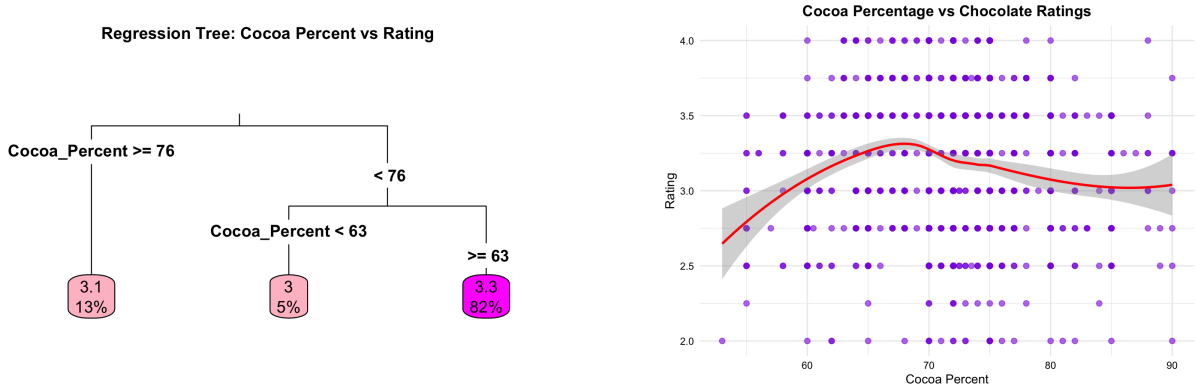
Figure 2: Hierarchical Clustering of South America: Rating vs. Cocoa Percent

Relationship Between Cocoa Percentage and Ratings

To analyze the relationship between cocoa percentage and ratings, a **Regression Tree** model was implemented. This decision-tree-based approach was chosen for its ability to uncover non-linear relationships and generate interpretable decision rules. The model used *Cocoa Percent* as the sole input variable and was evaluated using **Root Mean Squared Error (RMSE)** and has a **RMSE** of 0.471083 . The regression tree complexity parameters and performance are summarized in Table 4, while the structure of the regression tree is visualized in Figure 3a.

Regression Trees revealed a non-linear relationship between cocoa percentage and ratings. Ratings peaked within the 65%-80% range, averaging 3.21 which confirmed the results observed in the EDA. However, ratings declined for cocoa content exceeding 95%, averaging 2.4, likely due to excessive bitterness. Similarly, lower cocoa percentages (below 63%) were associated with reduced ratings, reflecting a preference for balanced flavors. The relationship between predicted and actual ratings, depicted in Figure 14, further emphasizes these trends and highlights systematic biases in the mid-range of ratings.

These insights are crucial for guiding product development. By optimizing cocoa content within the 65%-80% range, manufacturers can align their offerings with consumer preferences while promoting the sustainability of cocoa farming. Educational campaigns emphasizing the flavor complexities of this range could further enhance consumer appreciation.



(a) Regression Tree: Cocoa Percent vs Rating

(b) Cocoa Percentage vs Chocolate Ratings

Figure 3: Regression tree and cocoa ratings side-by-side comparison.

Predicting High-Rated Chocolate Bars

To predict whether a chocolate bar would achieve a high rating (greater than 3.5), three models were tested: Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression was initially chosen for its simplicity and interpretability in binary classification problems. However, the presence of potential non-linear relationships between predictors and outcomes led to explore other machine learning methods that might be better suited to handle non-linear relationships. Gradient Boosting ultimately outperformed the other models, making it the model of choice for this task.The predictors used were **Cocoa_Percent**, **Broad_Bean_Origin**, and **Company_Location**. Model performance was evaluated using metrics such as **Accuracy** and **Area Under the Curve (AUC)**.

The Gradient Boosting model was as followed:

$$\text{High\_Rating} = f(\text{Cocoa\_Percent}, \text{Broad\_Bean\_Origin}, \text{Company\_Location})$$

Where:	With the following parameters:
Cocoa_Percent: Percentage of cocoa in the chocolate.	**Number of Trees:** 1000
Broad_Bean_Origin: Factor representing the bean's origin.	**Tree Depth:** 3
Company_Location: Factor representing the company's location.	**Learning Rate:** 0.01
	Cross-Validation: 5

Gradient Boosting achieved an accuracy of 79.18% and an AUC of 0.64 (Figure 15, Table 7, Table 8 and Table 9). Its ability to capture complex, non-linear relationships makes it a valuable tool for understanding nuanced consumer preferences. The training progress of the Gradient Boosting model, including improvements in deviance, is summarized in Table 7 and through the ROC curve Figure 15.

The Gradient Boosting model identifies the Venezuela-Italy pairing as the highest-performing category, with a predicted probability of 0.51. This shows the significance of Venezuelan cocoa, particularly the rare and highly prized Criollo variety, known for its delicate and nuanced flavor profile (Núñez, 2023). Italian chocolate craftsmanship, exemplified by renowned brands like Venchi, emphasizes quality and traditional methods, resulting in rich, decadent flavors (Hill Country Chocolate, 2023). Other high-performing pairings, such as Venezuela-Switzerland and Madagascar-Italy, also present strategic opportunities. Swiss chocolate is celebrated for its precision and premium quality, while Madagascar's cocoa is prized for its unique flavor profile, characterized by bright, fruity notes and a hint of nuttiness (CocoTerra Company, 2022). These combinations likely resonate with consumers seeking exceptional taste experiences, offering a balance of authenticity, innovation, and superior quality.

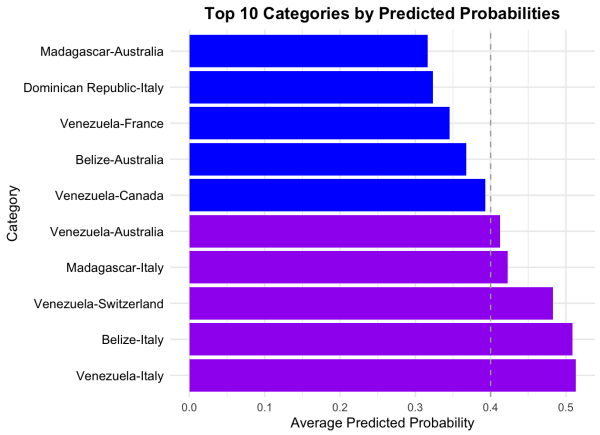


Figure 4: Top 10 Categories by Predicted Probabilities
The table show the top 10 category pairs, combining the source of the bean and the company location of the brand

Classification/Prediction and Recommendations

Regions such as Belize and Brazil emerge as top candidates for long-term partnerships due to their high-quality cocoa and low inconsistency, making them ideal for premium product lines with higher quality cocoa beans. Incorporating these origins into initiatives like a "Sustainable Heritage Cocoa Collection" could secure supply chain stability and align with growing consumer demand for ethically sourced products. In contrast, investing in Ghana and Mexico offers opportunities to transform mid-tier suppliers into reliable sources through targeted quality improvement programs and sustainable farming initiatives. Venezuela, with its strong positive

association with high ratings, is well-positioned for exclusive product launches, leveraging its reputation for producing rare and highly sought-after cocoa varieties.

To capitalize further on the unique strengths of regions like Venezuela and Belize, manufacturers can explore regional collaborations that extend beyond traditional marketing. For example, partnering with local artisans to co-create limited-edition chocolate collections can highlight the cultural heritage of these regions. These collaborations could be supported by pop-up events in urban markets, where consumers can interact with chocolatiers, sample exclusive offerings, and learn about the craftsmanship involved and then be expanded internationally in strategic location based company location and popular markets. Such initiatives will tap into consumer interest in authenticity and sustainability, reinforcing the brand's position in the premium market segment.

Data analysis reveals that cocoa content between 65% and 80% achieves peak consumer ratings, averaging 3.21. This range caters to both health-conscious consumers and premium segments (more attracted by the quality of the bean). Products in this range should be marketed as indulgent yet sustainable, using wellness-oriented messaging. Chocolates exceeding 95% cocoa can target connoisseurs through exclusive campaigns, such as collaborations with sommeliers or coffee brands for unique pairings (as even if the market it niche there is a a demand for it). For chocolates between 63% and 76%, approachable flavor profiles and family-focused messaging can broaden appeal. Educational initiatives, such as "The Authentic Sweet Spot," could reshape consumer perceptions of high-cocoa chocolates while increasing demand among health-conscious markets.

European consumers, who favor high-cocoa-content chocolates, present an opportunity for artisan-crafted offerings that emphasize rare beans like Criollo. Campaigns such as "From Criollo to Craft" could highlight culture and exclusivity, resonating with refined tastes. Regional sponsorships, such as supporting European craft fairs or South American cocoa festivals, can further enhance the brand's association with authenticity and culture, strengthening its connection to both local and global markets. Packaging could also portray the region's sustainable farming practices (implemented by the previous strategies and reflected through the analysis), reinforcing ethical branding. In South America, a dual strategy is recommended: premium high-cocoa options for connoisseurs and approachable blends for casual consumers. Limited-edition packaging highlighting unique regional flavor profiles, such as Ecuador's floral notes or Peru's fruity undertones, could enhance cultural resonance and foster stronger consumer connections and loyalty.

Pairings like Venezuela-Italy and Belize-Italy represent high-priority opportunities, combining premium cocoa origins with strategic company locations. Venezuela-Italy, with its strong association with quality, can anchor luxury-focused campaigns targeting consumers attached to high quality cocoa, while Belize-Italy offers an opportunity to explore emerging markets with high growth potential. These insights are summarized in Table 1, which highlights high-potential pairings and strategic recommendations.

Category	Avg. Prob.	Priority	Recommendation
Venezuela-Italy	0.513	High	Target premium segment
Belize-Italy	0.508	Medium	Emerging market
Venezuela-Switzerland	0.483	High	Focus on partnerships
Madagascar-Italy	0.423	Medium	Explore niche markets
Venezuela-Australia	0.412	Medium	Build regional campaigns
Venezuela-Canada	0.393	Medium	Expand presence in North America
Belize-Australia	0.368	Low	Test new product lines
Venezuela-France	0.346	Medium	Strengthen European foothold
Dominican R -Italy	0.323	Low	Trial limited editions
Madagascar-Australia	0.317	Low	Explore consumer interest

Table 1: Market-Oriented Prioritization Matrix for High-Potential Categories
The table show high-potential category pairs, combining the source of the bean and the company location of the brand

To maintain a competitive edge, manufacturers should leverage predictive analytics to prioritize regions and product categories with the highest consumer engagement potential. Storytelling campaigns, such as "Heritage Origins or Culture Origins," could spotlight premium pairings like Venezuela and Italy, using immersive QR codes to connect consumers with cocoa origins. Meanwhile, partnerships with regions like Switzerland can facilitate expansion into niche luxury markets, and testing markets through partnerships with medium-priority regions like Madagascar-Italy unlocking new opportunities.

All of these initiatives will allow manufacturers to optimize resource allocation, diversify supply chains, and enhance brand authenticity.

Conclusion

This report highlights how data-driven insights can help the chocolate industry adapt to evolving consumer preferences while addressing operational and sustainability challenges. By identifying stable, high-quality cocoa sources like Belize and Brazil, manufacturers can strengthen supply chains and enhance product quality to stand out in competitive markets. The analysis demonstrates that success requires more than sourcing premium beans or optimizing cocoa content. A well-rounded strategy that incorporates regional customization, ethical sourcing, and consumer education is essential. Tailoring products to specific market tastes—such as high-cocoa chocolates for Europe and diverse offerings for South America—can deepen customer connections and expand market reach and that even with the data some assumptions and qualitative research is still required to market chocolate and reach consumers. Looking ahead, the chocolate industry has an opportunity to combine innovation with responsibility. By fostering transparency in sourcing and emphasizing sustainable practices, manufacturers can meet growing consumer expectations for ethical products while maintaining profitability. With these strategies, chocolate makers can balance tradition and modern demands, securing a sustainable and competitive future.

Appendix

Numerical Variables

Ratings Histogram and Cocoa Percent Histogram

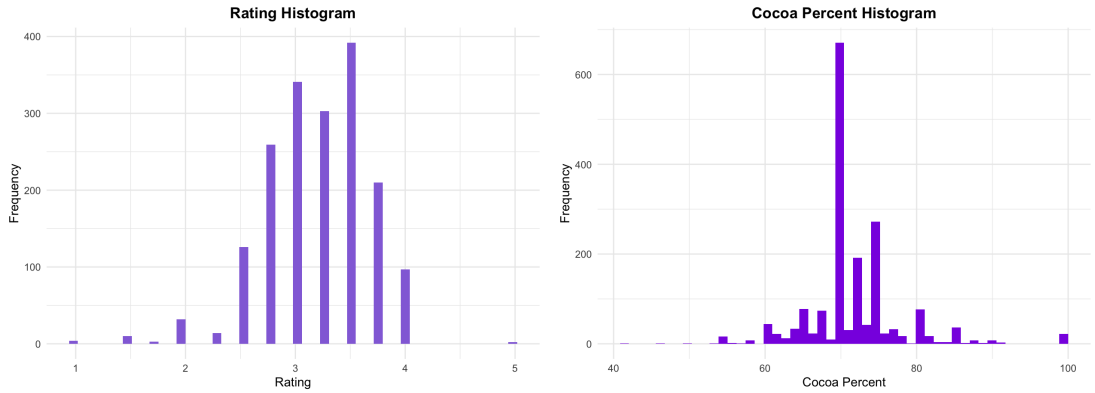


Figure 5: Ratings Histogram (left) and Cocoa Percent Histogram (right).

The *Ratings Histogram* shows that most ratings are clustered between 3.0 and 4.0, indicating that the majority of chocolates are rated as above average. This suggests a preference for moderately high-quality chocolate among consumers. Similarly, the *Cocoa Percent Histogram* highlights a sharp peak around 70%, indicating that chocolates with cocoa percentages around this range are the most common and likely the most preferred by consumers.

Review Date Histogram and Ratings vs Cocoa Percent

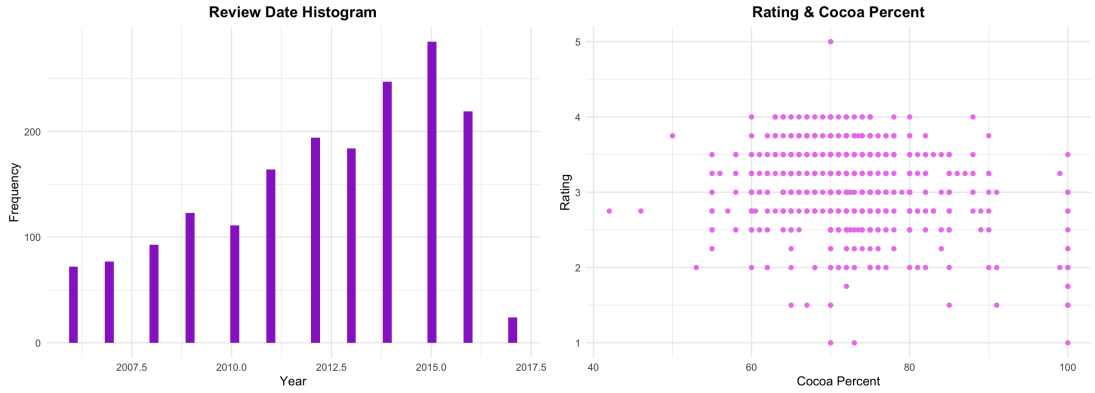


Figure 6: Review Date Histogram (left) and Ratings vs Cocoa Percent Scatter Plot (right).

The *Review Date Histogram* shows a steady increase in reviews over time, with a notable peak in 2015. This may reflect growing consumer interest in chocolate products and increased market engagement during this period. On the other hand, the *Ratings vs Cocoa Percent Scatter Plot* highlights that ratings tend to peak between 70% and 85% cocoa content. Ratings decrease for chocolates with either very low or very high cocoa percentages, suggesting that consumers prefer a balanced flavor profile rather than extremes.

Boxplots for Ratings and Cocoa Percentages

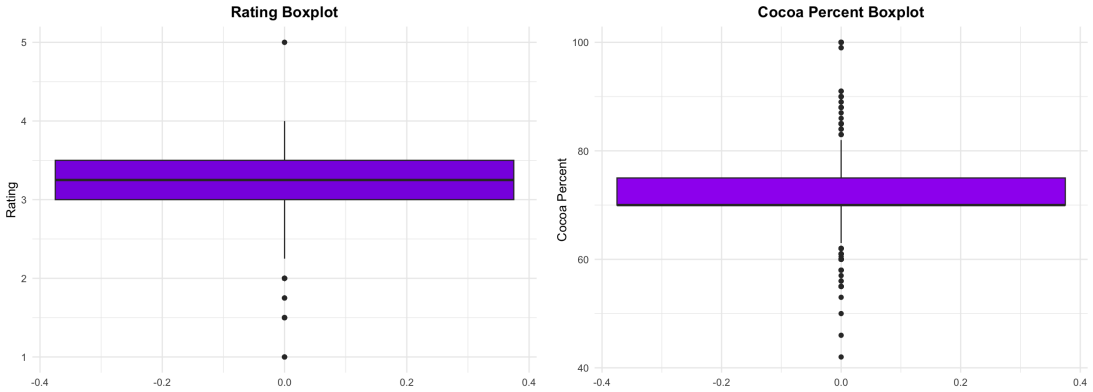


Figure 7: Ratings Boxplot (left) and Cocoa Percentages Boxplot (right).

The *Ratings Boxplot* demonstrates that the median rating is between 3.0 and 3.5, with a relatively narrow interquartile range, suggesting consistent quality among most chocolates. Few outliers with extreme ratings are present, which indicates that most chocolates meet consumer expectations. The *Cocoa Percent Boxplot* shows that cocoa percentages around 70% dominate the dataset, with a narrower spread compared to the Ratings Boxplot. This supports the idea that chocolates in this range are both more common and potentially more appealing.

Correlation Heatmap and Residual Plots

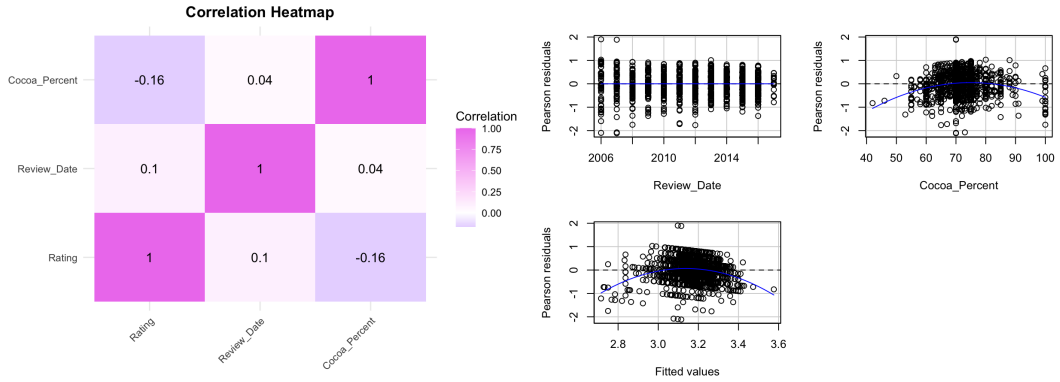


Figure 8: Correlation Heatmap (left) and Residual Plots (right).

The *Correlation Heatmap* reveals weak correlations among key variables, such as a -0.16 correlation between Cocoa Percent and Rating, indicating that these variables may not exhibit a strong linear relationship. This suggests that non-linear modeling techniques may be more effective for understanding these dynamics. Additionally, the *Residual Plots* demonstrate non-linear patterns and potential heteroscedasticity, particularly in the relationships between Review Date, Cocoa Percent, and Rating.

Categorical Variables

Analysis of Key Categorical Variables

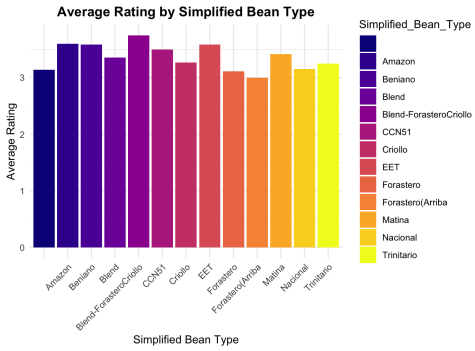


Figure 9: Average Ratings by Bean Type

Among the bean types, Criollo and blends containing Criollo consistently achieve the highest average ratings. These bean types are renowned for their rich, complex flavors and rarity, making them a preferred choice among chocolate enthusiasts. On the other hand, Forastero, known for its robustness and higher yield, tends to receive lower ratings, indicating a consumer preference for more refined and intricate flavor profiles.

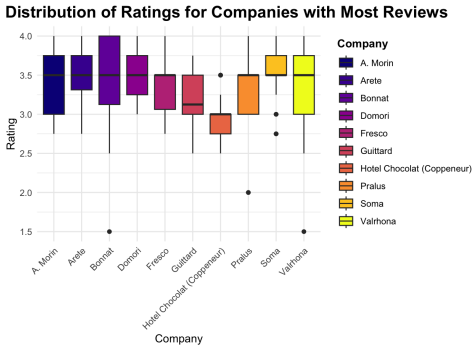


Figure 10: Distribution of Ratings for Companies with Most Reviews

The distribution of ratings for companies with the most reviews highlights variability in product quality among top producers. Companies such as Bonnat and Domori consistently deliver high ratings with minimal variation, reflecting their focus on quality and premium products. Conversely, companies like Hotel Chocolat show a broader range of ratings, indicating diverse product offerings or potential inconsistencies in quality.

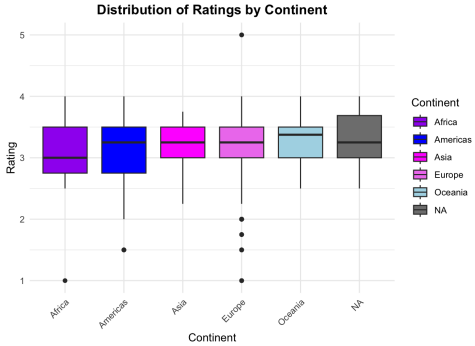


Figure 11: Distribution of Ratings by Continent

The distribution of ratings by continent reveals intriguing regional differences. Oceania and Asia exhibit the highest median ratings, likely driven by niche, high-quality chocolate producers catering to specific consumer preferences. Europe and the Americas display greater variability in ratings, suggesting a mix of both premium and mass-market chocolate products.

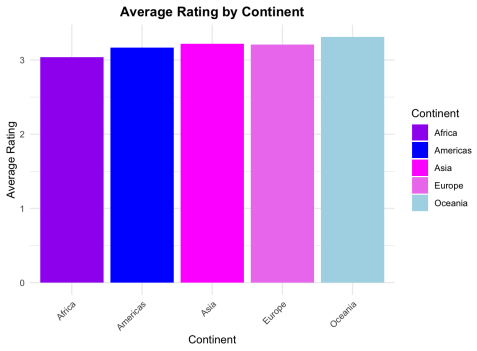


Figure 12: Average Ratings by Continent

When examining average ratings by continent, Oceania and Asia again emerge as leaders, reinforcing the dominance of these regions in producing highly-rated chocolates. Africa, while less prominent, shows consistent performance, possibly reflecting the quality of its raw cocoa production that contributes to the global chocolate supply chain.

Objective 1

Clustering Summary

The clustering of cocoa bean origins is based on two variables:

- **Mean Rating (μ)**: The average rating of chocolate bars sourced from a particular region.
- **Rating Standard Deviation (σ)**: Measures the variability in ratings, indicating quality inconsistency.

The clusters are summarized as follows:

$$\text{Cluster Data: } \{\mu, \sigma\}$$

Cluster	Average Mean Rating (μ)	Average Rating SD (σ)	Number of Origins
1	3.28	0.372	17
2	3.16	0.467	18
3	2.70	0.467	2

Table 2: Clustering Results of Cocoa Bean Origins Based on Ratings

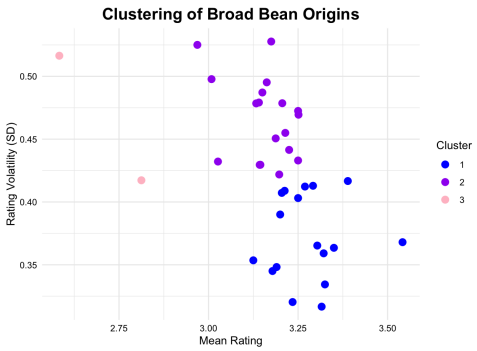


Figure 13: Clustering of Broad Bean Origins.

The scatter plot below illustrates the clustering of broad bean origins based on mean rating and standard deviation. The clustering analysis categorizes cocoa bean origins into three groups: high-rating and low-quality inconsistency regions ideal for partnerships, moderate-rating regions with higher variability, and low-rating, high-quality inconsistency regions that pose greater risks.

Objective 2

Cultural Influence	Cocoa Percent (%)	Rating
Africa	0.121	−0.468
Americas	0.016	−0.078
Asia	−0.043	0.084
Europe	0.045	0.043
Oceania	−0.226	0.157

Table 3: Cultural Influence on Cocoa Percentage and Ratings

Linear Discriminant Analysis (LDA)Table:

This table highlights the cultural influence of each continent on cocoa percent and rating. Africa shows the highest positive cocoa percentage but the lowest rating, suggesting potential quality inconsistency. Oceania exhibits the highest positive rating influence but a significant negative cocoa percentage. This divergence indicates regional differences in preferences and production quality.

Objective 3

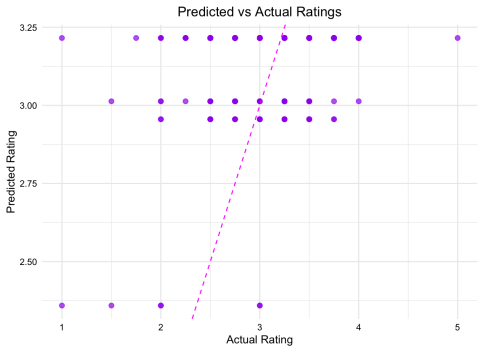


Figure 14: Predicted vs Actual Ratings

The scatter plot compares predicted ratings with actual ratings. The dashed line represents the ideal scenario where predicted ratings perfectly match the actual ratings. Most data points cluster around specific values, indicating some systematic bias in the prediction model, especially around the mid-range of ratings.

Regression Tree Model

The regression tree was fit to predict *Ratings* based on *Cocoa Percent*. The model uses an ANOVA method, with the complexity parameter (*cp*) set to 0.01. The summary of the regression tree is as follows:

$$\text{Rating} = f(\text{Cocoa\_Percent}) + \epsilon$$

CP	Splits	Rel. Error	CV Error	Std. Dev.
0.0387	0	1.0000	1.0015	0.0493
0.0138	1	0.9613	0.9834	0.0470
0.0100	3	0.9337	0.9624	0.0475

Table 4: Summary of Regression Tree Complexity Parameters and Performance

The table summarizes the model's performance at different levels of complexity. The relative error decreases as the number of splits increases, improving model fit.

Regression Tree Equation

The regression tree predicts the *Rating* (R) using a piecewise function derived from the splits in *Cocoa Percent*. The root node error is calculated as:

$$\text{Root Node Error} = \frac{\text{Sum of Squared Deviations}}{\text{Number of Observations}} = \frac{277.43}{1255} = 0.22106$$

This equation demonstrates the model's error at the root node before applying splits to improve prediction accuracy.

Objective 4

Logistic Regression Coefficients

The logistic regression model used to predict the probability of achieving a high rating ($Y = 1$) is as follows:

$$\text{Logit}(P(Y = 1)) = \beta_0 + \beta_1 \text{Cocoa Percent} + \beta_2 \text{Broad Bean Origin} + \beta_3 \text{Company Location}$$

where:

- $P(Y = 1)$: Probability of achieving a high rating.
- β_0: Intercept.
- $\beta_1, \beta_2, \beta_3$: Coefficients for Cocoa Percent, Broad Bean Origin, and Company Location respectively.

The following table summarizes the logistic regression coefficients and their standard errors:

Variable	**Estimate**	**Std. Error**
Intercept	-1.9402	0.62189
Cocoa Percent	-0.26082	0.08736
Broad Bean Origin (e.g., Venezuela)	1.2223	0.51607
Company Location (e.g., Canada)	0.19950	0.46198

Table 5: Logistic Regression Coefficients and Standard Errors

The negative coefficient for Cocoa Percent (-0.26082) indicates that as the cocoa percentage increases, the likelihood of receiving a high rating decreases, holding other factors constant. Positive coefficients for specific Broad Bean Origins (e.g., Venezuela) suggest that beans from these regions are associated with higher probabilities of high ratings. The effect of company location appears to vary, as seen in the small and mixed coefficient values.

Random Forest Model

The Random Forest model was trained with the following parameters:

- **Number of Trees:** 500
- **Number of Features per Split:** 2
- **Importance Calculation:** Enabled

The feature importance, as determined by the Random Forest model, is shown in Table 6:

Feature	**Importance Score**
$CompanyLocation$	40.23
$BroadBeanOrigin$	37.46
$CocoaPercent$	22.31

Table 6: Feature Importance in Random Forest Model

Gradient Boosting Model

The Gradient Boosting model was trained with the following parameters:

- **Number of Trees:** 1000
- **Tree Depth:** 3
- **Learning Rate:** 0.01
- **Cross-Validation:** 5

The training progress of the model is shown in the following table:

Iteration	Train Deviance	Step Size	Improvement
1	0.8761	0.0100	0.0004
2	0.8754	0.0100	0.0002
⋮	⋮	⋮	⋮
1000	0.7540	0.0100	−0.0002

Table 7: Training Progress for Gradient Boosting Model

Feature importance, as determined by the Gradient Boosting model:

Feature	Relative Influence (%)
CompanyLocation	40.99
BroadBeanOrigin	37.69
CocoaPercent	21.32

Table 8: Feature Importance in Gradient Boosting Model

Comparison of Model Performance

The performance metrics of all three models are summarized below:

Model	Accuracy	Precision	AUC
Logistic Regression	79.37%	0.00	0.583
Random Forest	76.39%	25.00%	0.537
Gradient Boosting	79.18%	20.00%	0.635

Table 9: Comparison of Model Performance Metrics

ROC Curve

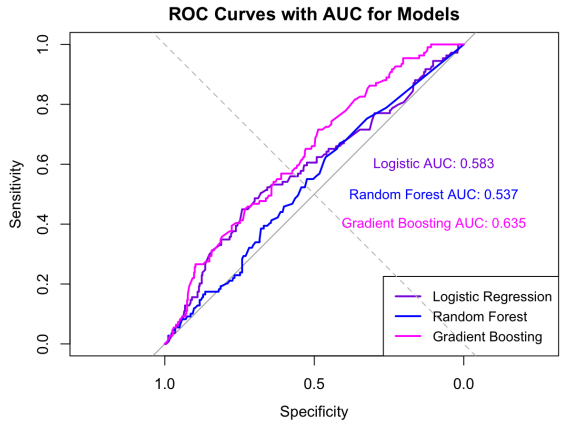


Figure 15: ROC Curve for Logistic Regression Model

The ROC curve provides an evaluation of the models' performance by plotting sensitivity against specificity. A model performing better than random guessing will have a curve above the diagonal line. The Logistic Regression model achieves an AUC of 0.583, indicating moderate discrimination between high and low ratings. The Random Forest model has a slightly lower AUC of 0.537, showing weaker performance in distinguishing between the two classes. In contrast, the Gradient Boosting model performs the best with an AUC of 0.635, suggesting improved predictive capability over the other two models.

Cocoa Percent vs Predicted Probability

The scatter plot below shows the relationship between standardized cocoa percent and the predicted probability of achieving a high rating, color-coded by the actual high rating (1 for high, 0 for low). The dashed line represents a fitted trend.

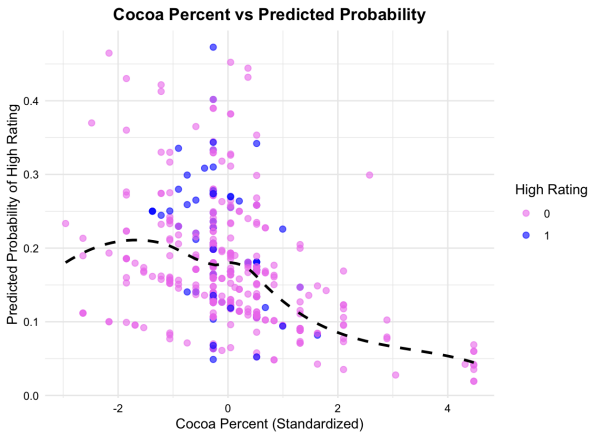


Figure 16: Cocoa Percent vs Predicted Probability of High Rating

The plot indicates a nonlinear relationship between cocoa percent and the probability of receiving a high rating. Higher cocoa percentages generally correspond to lower predicted probabilities of high ratings, with some variability based on actual outcomes.

References

[1] Dahl, R. (1964). *Charlie and the chocolate factory*. Alfred A. Knopf.

[2] Dame Cacao. (2023). *Global chocolate statistics*. Retrieved from https://damecacao.com/chocolate-statistics

[3] Expert Market Research. (2023). *Chocolate market size, trends, and growth analysis*. Retrieved from https://www.expertmarketresearch.com/reports/chocolate-market

[4] BakeMag. (2023). Consumer preferences in chocolate: New trends. Retrieved from https://www.bakemag.com/articles/18554-consumer-preferences-in-chocolate

[5] Wired. (2023). *Chocolate has a sustainability problem. Science thinks it's found the answer.* Retrieved from https://www.wired.com/story/chocolate-has-a-sustainability-problem-science-thinks-its-found-the-answer

[6] Confectionery Production. (2023). Barry Callebaut unveils major global chocolate trends for 2024. Retrieved from https://www.confectioneryproduction.com/news/46288/barry-callebaut-unveils-major-global-chocolate-trends-for-2024/

[7] CocoTerra Company. (2022, April 1). *10 distinct varieties of cacao: Ultimate cacao guide*. Retrieved from https://www.cocoterra.com/10-distinct-varieties-of-cacao/

[8] Hill Country Chocolate. (2023, June 15). *Exploring the rich history of Venchi chocolate*. Retrieved from https://www.hillcountrychocolate.com/blogs/chocolate-and-confections-1/venchi-chocolate

[9] Núñez, J. (2023, March 10). *Savoring the essence of Venezuelan cocoa: A guide to tasting notes and varieties*. Mesa Trading Group. Retrieved from https://mesatradinggroup.com/blog/savoring-the-essence-of-venezuelan-cocoa-a-guide-to-tasting-notes-and-varieties