Data Summary: Flavors of Cacao Dataset

Data Source

The dataset was initially sourced from Kaggle, specifically from the dataset titled Chocolate Bar Ratings. It provides insights into chocolate bar reviews, including cocoa percentages, company origins, and ratings. The dataset was compiled by Brady Brelinski, a founding member of the Manhattan Chocolate Society. Additional content, including interviews with craft chocolate makers, can be found on his website: Flavors of Cacao.

After noticing a significant number of missing values in the Bean Type and Broad Bean Origin columns, I decided to verify the original data source. I discovered a more complete version of the dataset on the Flavors of Cacao Chocolate Database and opted to use this version to ensure a more comprehensive analysis. The updated version includes 2,693 individual chocolate bars and contains more detailed information.

Questions to Explore

- How does cocoa percentage influence chocolate ratings?
- Which companies consistently receive the highest ratings?
- Which countries produce the highest-rated chocolate bars?
- Do chocolates from certain regions or countries receive higher ratings on average?
- Is there a relationship between cocoa solids percentage and rating?
- How do different chocolate ingredients impact ratings?

Data Collection

This dataset contains expert ratings of over 2,693 individual chocolate bars. It includes information on their regional origin, percentage of cocoa, the variety of chocolate bean used, and where the beans were grown. The ratings are based on a combination of objective qualities and subjective interpretations, focusing primarily on flavor, texture, aftermelt, and overall opinion. The dataset is specifically centered on plain dark chocolate to highlight the flavors of cacao when processed into chocolate. The dataset was last updated on March 3, 2024.

Data Limitations

- Some missing values exist in the Ingredients column (87 missing entries).
- The Cocoa Percent column is stored as a string with a percentage sign, requiring conversion to a numeric format.
- The column names contain extra spaces and newlines, which should be cleaned for ease of analysis.
- The Country of Bean Origin field includes a mix of country names and specific regions or places within those countries, potentially causing inconsistencies in geographic analysis.
- The ratings represent experiences with specific chocolate bars from particular batches, meaning batch variations could impact results.
- The dataset focuses exclusively on plain dark chocolate, meaning other chocolate varieties (such as milk or white chocolate) are not represented.

Possible Ethical Considerations

1. Bias in Expert Ratings
   - Are the chocolate ratings subjective and influenced by personal preferences of reviewers?
   - Could batch variations affect consistency in ratings?
2. Data Representation & Completeness
   - The dataset focuses on dark chocolate—does this exclude a major portion of the industry?
   - Are certain regions underrepresented in ratings, leading to incomplete global insights?

Data Profile

Review Criteria: Flavors of Cacao Rating System

The dataset follows a structured rating system based on expert evaluation:

Rating Scale:

4.0 - 5.0 = Outstanding

3.5 - 3.9 = Highly Recommended

3.0 - 3.49 = Recommended

2.0 - 2.9 = Disappointing

1.0 - 1.9 = Unpleasant

Not all the bars in each range are considered equal, so to show variance from bars in the same range, a .25, .50, or .75 is assigned.

Each chocolate is evaluated based on both objective qualities and subjective interpretation. A rating represents an experience with one bar from a specific batch, considering factors like batch numbers, vintages, and review dates.

The database is narrowly focused on plain dark chocolate with an aim of appreciating the flavors of the cacao when made into chocolate.

Key Evaluation Factors

Flavor: The most important component, focusing on diversity, balance, intensity, and purity of flavors.

Texture: Plays a significant role in the overall experience, impacting flavor perception and craftsmanship.

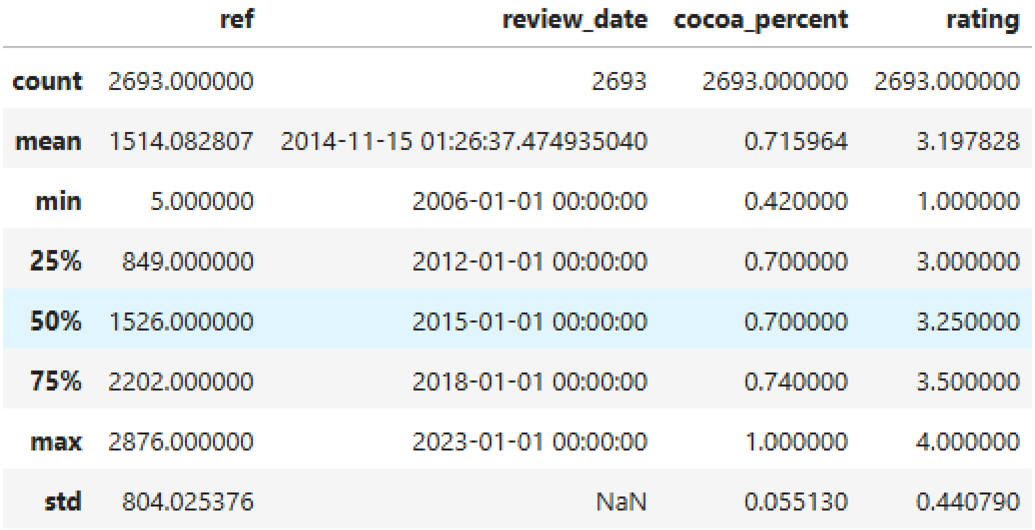
Aftermelt: The experience after the chocolate has melted, with higher-quality chocolate lingering longer and more enjoyably.

Overall Opinion: The final impression based on whether the components work together, summarizing key characteristics of the chocolate.

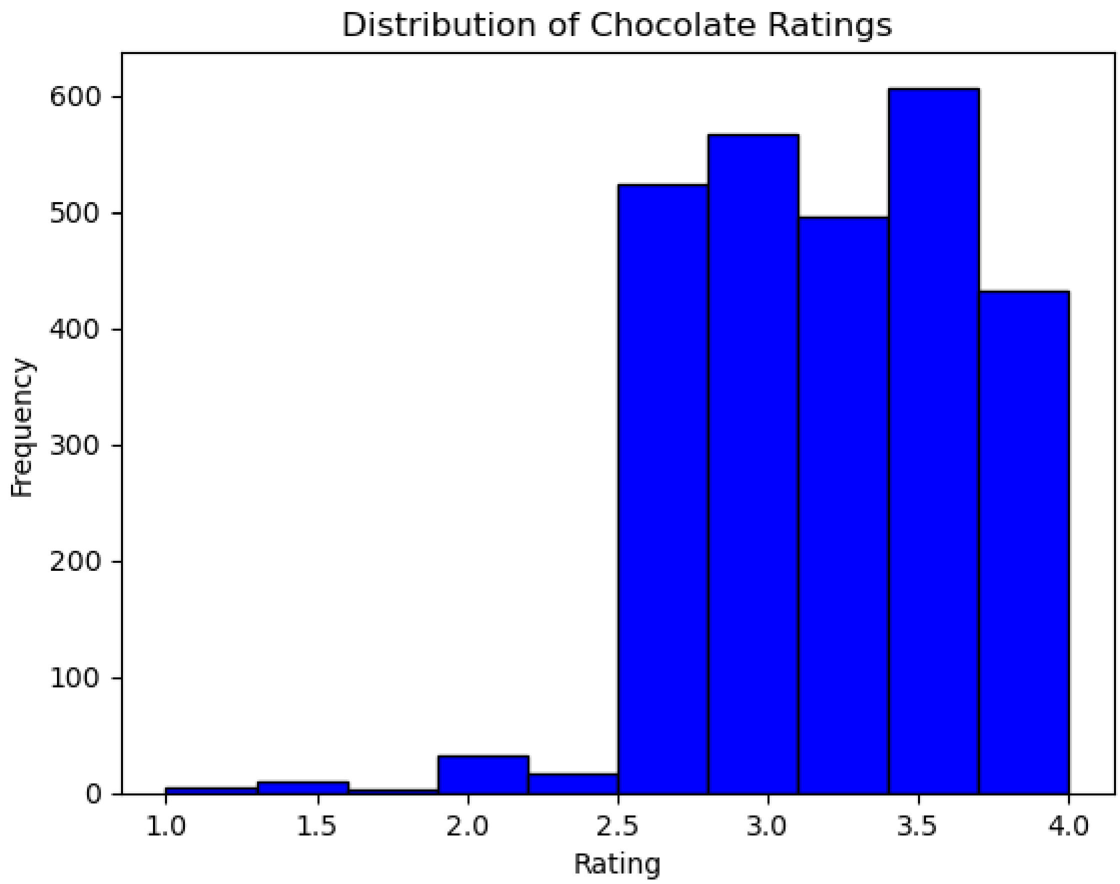
Data Cleaning Summary

1. Column names
   - Removing extra spaces and newline characters.
   - Converting all column names to lowercase and replacing spaces with underscores.
   - Removing brackets from column names.
   - Renaming columns for clarity and consistency:
     - company_manufacturer → company_name
     - specific_bean_origin_or_bar_name → bean_origin_or_bar
     - most_memorable_characteristics → memorable_characteristics
2. Handling Missing Values
   - Filling missing values in the *ingredients* column with "Unknown" to retain all data.
3. Ensuring Data Consistency
   - Removing periods from country names ("U.S.A." → "USA" and "U.K." → "UK").
   - Replacing non-country values with appropriate groupings or countries:
     - "Blend" → "Multiple Countries"
     - "Sulawesi", "Sumatra", "Bali" → "Indonesia"
     - "Principe" → "Sao Tome & Principe"
     - "Tobago" → "Trinidad & Tobago"
     - "France (Reunion)" → "Reunion"
4. Ensuring Data Types Are Correct:
   - Converting review_date from integer to datetime64 format to allow for time-based analysis.
5. Checking for Duplicates:
   - Verified that there are no duplicate rows in the dataset (df.duplicated().sum() == 0). Therefore, no further action was needed.

Descriptive Statistics



	ref	review_date	cocoa_percent	rating
count	2693.000000	2693	2693.000000	2693.000000
mean	1514.082807	2014-11-15 01:26:37.474935040	0.715964	3.197828
min	5.000000	2006-01-01 00:00:00	0.420000	1.000000
25%	849.000000	2012-01-01 00:00:00	0.700000	3.000000
50%	1526.000000	2015-01-01 00:00:00	0.700000	3.250000
75%	2202.000000	2018-01-01 00:00:00	0.740000	3.500000
max	2876.000000	2023-01-01 00:00:00	1.000000	4.000000
std	804.025376	NaN	0.055130	0.440790



Key Findings from Descriptive Statistics

1. Chocolate Ratings Distribution

- The majority of chocolates are rated between 3.0 and 3.5.
- The mean rating is 3.20, while the median rating is 3.25, indicating a slightly skewed distribution towards higher ratings.
- Very few chocolates score below 2.0, meaning most chocolates in the dataset are not considered bad.
- There are no ratings above 4.0, suggesting that 4.0 is the highest possible rating in this system.

2. Cocoa Percentage Trends

- The average cocoa percentage is 71.6%, with most chocolates having between 66% and 77% cocoa content.
- The median cocoa content is 70%, confirming that dark chocolate is the primary focus of this dataset.
- The range extends from 42% (low cocoa) to 100% (pure cocoa solids).

3. Insights from the Histogram

- The histogram confirms that ratings are concentrated between 3.0 and 3.5, with very few extreme values (either very low or very high).

- The absence of ratings between 4.0 and 5.0 indicates that the rating system might be capped at 4.0, meaning "Outstanding" chocolates cannot score above this threshold (potential expert review bias? The reviewers might not give perfect scores to maintain credibility)