**Project Document**

[Link to the Dataset "Flavors of Cacao" on Kaggle](https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings?resource=download)

***Data Source***

**Data source**: External Data Source (<https://flavorsofcacao.com/>) . The data were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society.

**Data collection method**: Expert Ratings & Observational Study

**Data contents**: Expert ratings of over 1,700 individual chocolate bars, along with information on their regional origin, percentage of cocoa, the variety of chocolate bean used and where the beans were grown (manufacturer-provided details)

**Data limitations**: Manufacturer Data can be incorrect or incomplete– the cocoa percentage, bean variety, and origin might come from manufacturers, who may not always provide accurate or standardized information. Also, expert ratings are subjective and may reflect personal taste rather than universal quality.

I have chosen this dataset for analysis due to my strong interest in chocolate and a want to explore this topic deeper. By examining expert ratings, cocoa percentages, and bean origins, I aim to identify patterns that reveal where the highest-quality chocolate beans are grown and how different factors influence the taste and quality of chocolate. This analysis will provide valuable insights into regional variations in cocoa production and the characteristics that define premium chocolate.

**Variables**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Type |
| Company (Maker-if known) | Name of the company manufacturing the bar. | object |
| Specific Bean Origin or Bar Name | The specific geo-region of origin for the bar. | object |
| REF | A value linked to when the review was entered in the database (from 5 to 1952). Higher = more recent. | int64 |
| Review Date | Date of publication of the review. | int64 |
| Cocoa Percent | Cocoa percentage (darkness) of the chocolate bar being reviewed. | object |
| Company Location | Manufacturer base country | object |
| Rating | Expert rating for the bar, where:   * 5= Elite, * 4= Premium , * 3= Satisfactory * 2= Disappointing * 1= Unpleasant | float64 |
| Bean Type | The variety (breed) of bean used, if provided. | object |
| Broad Bean Origin | The broad geo-region of origin for the bean. | object |

***Data Profile***

**Cleaning/Wrangling Steps**

* Datatype of “Cacao Percentage” changed from “object” to “float64”
* The column “Bean type” deleted as it has 888 missing values and irrelevant for my analysis
* The row containing the missing value in the "Broad Bean Origin" column was deleted

***Ethical Considerations***

* Many cocoa beans come from regions with issues related to child labor, low wages, and poor working conditions (e.g., West Africa countries like Togo, Gabon).
* If the dataset does not account for ethical sourcing practices, it may unintentionally promote brands that use exploitative labor practices.
* Ethical certifications (e.g., Fair Trade, Rainforest Alliance) are not included in the dataset, so high-rated chocolates might not always be ethically sourced.
* “Eucador “ in “Company Location” changed to “Ecuador
* “Amsterdam” in “Company Location” changed to "Netherlands”
* “Niacragua” in “Company Location” changed to “Nicaragua”

**Questions**

* Where are the best cocoa beans grown?
* Which countries produce the highest-rated bars?
* What’s the relationship between cocoa solids percentage and rating?
* Which company have the most highly rated bars?
* Has the average chocolate rating changed over time?