**Data Source**

**Project overview**

Chocolate stands as the most favored sweet indulgence globally, celebrated for its exceptional flavor with roots tracing back to ancient times. The Maya revered chocolate, particularly a cocoa beverage prepared with hot water, deeming it the "Food of the Gods." In the United States, it holds a substantial share, with a consumption rate of 59%, surpassing various other confectioneries. Nevertheless, the quality and origin of chocolate can vary significantly. The dataset encompasses expert assessments of over 1,700 distinct chocolate bars, detailing their regional origins, cocoa percentages, chocolate bean varieties, and the specific geographic locations where the beans were cultivated.

I was intrigued to explore deeper into the origins of chocolate, its creation locations, the predominant production percentages, and demand trends. My goal was to uncover fresh insights into chocolate data, which is why I opted for this particular dataset.

Flavor of Cocoa rating system.

* 5= Elite (Transcending beyond the ordinary limits)
* 4= Premium (Superior flavor development, character and style)
* 3= Satisfactory (3.0) to praiseworthy (3.75) (well made with special qualities)
* 2= Disappointing (Passable but contains at least one significant flaw)
* 1= Unpleasant (mostly unpalatable)

Each chocolate is evaluated from a combination of both objective qualities and subjective interpretation. A rating here only represents an experience with one bar from one batch. Batch numbers, vintages and review dates are included in the database when known.

The database is narrowly focused on plain dark chocolate with an aim of appreciating the flavors of the cacao when made into chocolate. The ratings do not reflect health benefits, social missions, or organic status.

**Data Profile**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Time-variant / invariant** | **Structured / Unstructured** | **Qualitative / Quantitative** | **Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous** |
| Company \n(Maker-if known) | invariant | Structured | Qualitative | Nominal |
| Specific Bean Origin\nor Bar Name | invariant | Structured | Qualitative | Nominal |
| REF | invariant | Structured | Qualitative | Ordinal |
| Review\nDate | Time-variant | Structured | Quantitative | Continuous |
| Cocoa\nPercent | Time-variant | Structured | Quantitative | Continuous |
| Company\nLocation | invariant | Structured | qualitative | Nominal |
| Rating | Time-variant | Structured | Quantitative | Continuous |
| Bean\nType | invariant | Structured | Qualitative | Nominal |
| Broad Bean\nOrigin | invariant | Structured | Qualitative | Nominal |

**Data cleaning:**

# identifying the columns and rows.

df.shape result

|  |
| --- |
| (1795, 9) |

Missing data:

|  |
| --- |
| df.isnull().sum() |

|  |  |
| --- | --- |
| Column name | Missing count |
| Company \n(Maker-if known) | 0 |
| Specific Bean Origin\nor Bar Name | 0 |
| REF | 0 |
| Review\nDate | 0 |
| Cocoa\nPercent | 0 |
| Company\nLocation | 0 |
| Rating | 0 |
| Bean\nType | 1 |
| Broad Bean\nOrigin | 1 |

Looking at the result I could either delete the row, but the question is, is it going to impact further in analysis. And since it is less than 5 % of the I can remove the 2 rows from the data set.

When apply this code it only showed me Nan value but not empty, looking in to the column there were empty spaces, just to see how many where empty spaces

|  |
| --- |
| empty\_space\_count = df\_clean1['Bean\nType'].astype(str).str.isspace().sum()    if empty\_space\_count > 0:          print(f"Column 'Bean\\nType' has {empty\_space\_count} empty spaces.")  else:          print("Column 'Bean\\nType' has no empty spaces.")  Result : Column 'Bean\nType' has 887 empty spaces. |

Since there was a large number of empty spaces I would not delete and row since it can alter my result significantly and lose a large amount of the data.

Mix data time: None missing value

|  |
| --- |
| #checking for mix data types  for col in df\_clean1.columns.tolist():    weird = (df\_clean1[[col]].applymap(type) != df\_clean1[[col]].iloc[0].apply(type)).any(axis = 1)    if len (df\_clean1[weird]) > 0:      print (col) |

Duplicates: None duplicates were identified.

|  |
| --- |
| #creating new subset that contains only duplicated  df\_clean1\_dup = df\_clean1[df\_clean1.duplicated()]  df\_clean1\_dup |

|  |  |  |
| --- | --- | --- |
| Column Name | Datatypes | Changes |
| Company \n(Maker-if known) | object | none |
| Specific Bean Origin\nor Bar Name | Object | none |
| REF | int64 | Name to Reference Number |
| Review\nDate | int64 | Inter 32 to save space |
| Cocoa\nPercent | object | none |
| Company\nLocation | object | none |
| Rating | float64 | Float 16 to save space |
| Bean\nType | object | none |
| Broad Bean\nOrigin | Object | none |

**Data Limitations:**

-So many factors are taken in place to determine the flavor of the chocolate, putting a rating on one simple number may not be the best way to evaluate the chocolate flavor. This is one of the limitations of the flavor. Along with that there might be a lack of expertise that might not have the same knowledge as every other expert who is rating the chocolate flavors.

-The data entry is manually inputted, illustrating an example of limitations in the dataset due to the manual entry process.

-It appears that a significant portion of the chocolate samples in the dataset is derived from the USA, indicating a potential bias in the samples. To mitigate this bias and ensure a more representative dataset, it would be beneficial to obtain a comparable number of samples from various other countries. This approach aims to provide a more balanced and comprehensive representation of chocolate production and characteristics globally, allowing for a more accurate analysis and interpretation of the data.

-Time frame is from 2006 to 2017 which does not give us a recent data set to work with.

- There could be constraints related to the scope and specificity of the data, where it may be narrowly concentrated or lack the necessary depth for a thorough analysis. This limitation hinders its applicability for broader and more comprehensive insights.

Questions to answer

1. Which country created the highest rating chocolate bar?
2. Which country makes the highest percentage of cocoa?
3. What country products are being evaluated the most?
4. What is the time frame of the review
5. Relationship between chocolate percentage and rating?
6. How does the distribution of cocoa percentages look across different chocolate types?

Tableau Link

[Chocolate Rating | Tableau Public](https://public.tableau.com/app/profile/bandana.sharma6456/viz/ChocolateRating_17084661541070/ChocolateRating?publish=yes)

Link to git hub [Bandana2024/Chocolate-rating (github.com)](https://github.com/Bandana2024/Chocolate-rating)