**Chocolate Bar Ratings Data Set**

**(Plain Dark Chocolate)**

**Data Source:**

|  |  |
| --- | --- |
| **Chocolate Ratings** | |
| Source | The source of this data set is from Kaggle and is called [Chocolate Ratings](https://www.kaggle.com/datasets/andrewmvd/chocolate-ratings). This data set was taken from a site called [Flavors of Cacao](http://flavorsofcacao.com/chocolate_database.html). It is an external source. |
| Collection | Chocolate Ratings was collected by Brady Brelinski (Co-Founder of Manhattan Chocolate Society) from the year 2006 to 2021. And was collected by reviews/surveys. |
| Contents | This data set contains **reviews on more than 2400 different chocolate bars** from 2006 – 2021 along with metadata and information on US and Canadian based producers.  Data set chocolate\_ratings contains 10 columns with 1 that I personally imputed (so 11 columns all together). And 2531 rows of data including the column title. So basically 2530 rows of data.  Data set chocolate\_makers contains 5 columns and 317 rows of data including the column title. So basically 316 rows of data.  Chocolate Ratings Data Set:   * REF * Company (Manufacturer) * Company Location * Review Date * Country of Bean Origin * Specific Bean Origin or Bar Name * Cocoa Percent * Ingredients * Most Memorable Characteristics * Rating * Rating Scale |
| Relevancy |  |

**Data Profile:**

Consistency checks: [6.1 Consistency Checks](6.1%20Task.ipynb.zip)

Changes made:

* I boldened the column titles and fill colored the title cells and added filters.
* Added the rating scale that was provided with the data sets.
* There were some right aligned company names, so I had to left align them. There were 7 cells of data.
* 87 cells of blanks found in the ingredient’s column. Imputing Missing Ingredients in those columns.
* Dropped 'Country of Bean Origin', 'Specific Bean Origin or Bar Name', 'Most Memorable Characteristics' columns because they were unnecessary for this project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Quant/Qual | Structured/  Unstructured | Time-Variant/  Invariant | *Nominal/ordinal/*  *Discrete/continuous* |
| REF | Quantitative | Structured | Invariant | Discrete |
| Company (Manufacturer) | Qualitative | Structured | Invariant | Nominal |
| Review Date | Quantitative | Structured | Invariant | Discrete |
| Country of Bean Origin | Qualitative | Structured | Invariant | Nominal |
| Specific Bean Origin or Bar Name | Qualitative | Structured | Invariant | Nominal |
| Cocoa Percent | Quantitative | Structured | Variant | Continuous |
| Ingredients | Qualitative | Unstructured | Variant | Ordinal |
| Most Memorable Characteristics | Qualitative | Unstructured | Variant | Nominal |
| Rating | Quantitative | Unstructured | Variant | Ordinal |
| Rating Scale | Quantitative | Structured | Invariant | Ordinal |

Limitations and Ethics:

Limitations – Because the data only starts from the year 2006, we can’t know how each company’s chocolate bars has produced over time. All we can predict is from the years 2006 to 2021, when the production of chocolate bars has started since the year 1847.

Ethics – There may be some collection bias, and I say some because the reviewees can review the chocolate bars with their own description and what they think of the bar but only from a list of descriptive words that are given. And that each person has their own opinions and likes of chocolate bars. While one may taste bitterness in the bar another might not. That is why I say some collection bias.

Update: I recently got new information that it was not a group of reviewees, it was solely just one person that reviewed the chocolate bars. It was Brady Brelinski. So, there is still some collection bias. Because it is only reviewed by one person, but he is a chocolate maker for over 6 years and is co-founder of Manhattan Chocolate Society.

Questions:

* How do the ratings differ from each country from the different years?
* Does the percentage of cocoa used change the ratings of the chocolate bar?
* What are the ratings for each country? Like are the ratings mostly 3.50 above or less than 3.50?
* Where and what companies are the ones that have a 4.0 rating and above?
* What percentage of cocoa has the highest ratings?
* What origin of cocoa bean has the highest rating?
* What are the ratings for cocoa percent of 70%?
* What are the ratings for cocoa percent of 80%?
* What are the ratings for cocoa percent of 90%?
* What are the ratings for cocoa percent of 100%?