**1. Data Source – Summary**

The chosen external dataset is the “Chocolate Bar Ratings” dataset sourced from Kaggle. This dataset includes sensory reviews of over 1700 unique chocolate bars, along with information on their regional origin, percentage of cacao, bean variety and where the beans were grown. These ratings were scored by cacao tasting experts. The ratings were compiled by Brady Brelinski, who is the Founding Member of the Manhattan Chocolate Society. More information on this dataset and background information about the ratings can be found on his website, listed below.

[Kaggle Dataset Link](https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings?select=flavors_of_cacao.csv)

[Flavors of Cacao Link](https://flavorsofcacao.com/index.html)

There were a few reasons why I chose this specific dataset for the final achievement of this project. The primary reason is that it met the all the criteria and specifications to complete the achievement. The dataset contains geographical components (country of bean origin, and company location), while also containing numerical components to do further statistical analysis on (chocolate ratings, review date, and cocoa percentage). For more in depth data visualizations, I can combine these components with categorical factors such as bean type and company to create filters or drill down views. All this makes for a comprehensive, yet simple, dataset that works perfect for this project. Lastly, I’ve always had a sweet tooth and affinity for chocolate, so I believe I will be able to present all my findings and analysis on the topic in a more passionate way, making a more compelling presentation in the end.

**2. Data Profile**

**Data Cleaning**

* After an initial glance of the data table, the **column names** we’re confusing and contained abnormal formatting. Each column name was changed to a more standardized format, words separated by an underscore.
* The **data types** were accurate except for 3 columns. The ‘Cocoa\_Percentage’ column was changed from a string to a float (numeric), and the ‘%’ symbol was removed for readability in the table. After a check for mixed-type columns, both ‘Bean\_Type’ and ‘Bean\_Origin’ contained mixed types. Both these columns were changed to string, as they didn’t contain numerical data.
* There were 887 **missing values** in the ‘Bean\_Type’ column, and 73 in the ‘Bean\_Origin’ column. These “Null” values were displayed as ‘\xa0’ but were blank in the table. All instances of these null values for both columns were replaced with ‘Unknown’ values, since contact couldn’t be made with the owner of the dataset about what they should be. All other columns did not have missing values
* There was no instance found for **duplicated** data in the table

**Understand the Data**

Below are the key attributes (columns) in the dataset, followed by a brief description of what reach represents:

* **Company**: This is the company name of the chocolate maker. There are over 400 unique chocolate makers (Categorical component)
* **Bar\_Name:** This refers to a specific bean origin, or bar variety. The data contains over 1000 unique bar varieties (Categorical component)
* **Reference\_Number**: This reference number refers to the unique chocolate bar
* **Review\_Date**: This column shows the year of publication of the review. The years the reviews span are from 2006 to 2017. (Trending/Numeric component)
* **Cocoa\_Percentage**: This represents the percentage of cocao (darkness) in the bar. (Numerical component)
* **Company\_Location**: This shows the manufacturer base country. There are about 60 countries. (Geographic component)
* **Rating**: This is the expert ratings of the specific bar. The rating system scale is from 1.0 to 5.0, and the descriptions are as follows for each rating: (Numerical component)
  + 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)
* **Bean\_Type:** The variety of bean used for the specific chocolate bar. There are over 40 bean types (Categorical component)
* **Bean\_Origin**: The country in which the cocoa bean originated. There are 100 distinct countries. (Geographic component)

Below is a screenshot of the descriptive statistics of the numerical categories (Review\_Date, Cocao\_Percentage, Rating). There are no serious outliers, and all the numbers seem to match up in accordance with what the dataset describes. There are 1795 rows of data and 9 columns.

A screenshot of a data

AI-generated content may be incorrect.

**Limitations & Ethics:**

1. There is a **temporal limitation** with this dataset, as the reviews only span from the years 2006-2017. While not a huge factor considering that these are ratings for chocolate bars, more up to date data and new reviews from experts could change the trends in which the data is displayed form this dataset.
2. Unfortunately, there was a decent chunk of **missing data**, as there were 887 missing values for Bean Type, and 73 missing values for Bean Origin. It would be unrealistic to track down the accurate information for these blank values, so for now we placed the ‘Unknown’ placeholder in these value slots. While not a huge issue for analysis, the data could be skewed a certain way considering the large chunk of missing data.
3. **Bias** is another factor for this dataset, specifically in the form of sampling and measurement bias. We don’t know the full details on all the “tasting experts”, but the flavors of cacao and founding members of the Manhattan Chocolate Society seem to be reputable. Since they are from Manhattan, the sampling bias could mean these “experts” are only American, and not evenly distributed throughout the globe (after all, chocolate is enjoyed in almost every country). While the scale for ratings is very precise when scoring, taste will always be a subjective quality. One expert’s taste may not be as “refined” as another’s, so measurement’s must be taken with some grain of salt. Even with this in consideration, the data seems to be trustworthy and efficient in relaying the message of how chocolate should be scored based on taste.
4. There are not too many ethical dilemmas in regards to privacy and security of the data, since it is public and can be used on almost any forum/platform for personal use. There is no PII data we’re dealing with in the dataset, so sensitive information is not a cause for concern. From an ethical standpoint, this dataset is okay to use for this project, and should not require any added permissions from the owner/license holder, so long as we don’t plan to profit off the findings and results.

**3. Questions to Explore**

* What countries grow the best cocoa beans?
* What countries produce the highest rated bars?
* Is there a correlation or relationship between cocoa percentage and rating?
* Is there a specific trend in time of rating chocolate bars, or the percentage of chocolate involved?
* Is there a way to predict a chocolate bar’s rating based on its other attributes (cocoa percentage, company, bean origin)?
* Are ratings better when bean origin and company location are closer together (proximity leads to better chocolate)?
* Is there a specific region of the world that produces the best ratings per chocolate bar (Americas vs. Europe vs. Asia)?
* Which bean types have the best ratings, and what correlation is there between bean percentage?
* Year by year display of count of bean type, company, bean origin, and cocoa percentages