Beatriz Acosta

Task 6.1

Data Source

The dataset was sourced from Kaggle (Chocolate Bar Ratings), the ratings were compiled by Brady Brelinski, a founding member of the Manhattan Chocolate Society. Is publicly available under the CCO: Public Domain license, meaning it can be freely used for analysis.

Data source: Chocolate Bar Ratings

Data Overview: The data Includes 1,795 sensory reviews of 440 unique chocolate bars, including details such as:

Company: Name of manufacturer

Specific Bean Originor Bar name: The specific geo-region of origin for the bar.

REF: Value linked to when the review was entered in the database.

Review date: When the chocolate was reviewed

Cocoa Percent: the percentage of cocoa solids in chocolate

Company Location: The country where the manufacturer is based

Rating: Expert rating based on flavor, texture and aftermelt

Bean Type: The variety of cacao beans used.

Broad Bean Origin: The Broader geographical region of the bean source.

Why This data:

I chose the chocolate bar dataset because it meets the specific criteria outlined in the project brief while allowing me to develop essential analytical skills. This dataset is open-source, comes from an authentic source, and contains non-anonymized column names, making it ideal for structured analysis. Additionally, includes at least two continuous variables (such as cocoa percentage and ratings), two categorical variables (such as brand and country of origin), and exceeds 1,500 rows, ensuring a rich data sample. The dataset also includes a geographical component, allowing for meaningful geospatial analysis. Beyond fulfilling these technical requirements, I find working with chocolate data engaging and enjoyable. This dataset offers opportunities to apply exploratory,

geospatial, and predictive analytics techniques, refining my skills in Python, SQL, and visualization tools like Tableau. By choosing a dataset that excites me, I ensure an immersive and insightful learning experience.

Data Profile

Data Cleaning:

Standardize Column Names:

'Company (Maker-if Know)': 'Company',

'Specific Bean Origin or Bar Name': 'Bar_Name',

'Review Date': 'Review_Date',

'Cocoa Percent': 'Cocoa_Percent',

'Company Location': 'Location',

'Bean Type': 'Bean_Type',

'Broad Bean Origin': 'Bean_Origin'

Data Type conversion: Transformed cocoa_percent string to a numeric value (float) for example from 63% to 63.0.

Missing Values:

Bean Type: 1 missing value

Bean Origin: 1 missing value

Both missing values were replaced with the label " Unknown" to preserved the rows intact.

Duplicates values:

The dataset shows zero duplicates values.

Basic Descriptive Statistics:

Total Entries: 1,795 chocolate bar reviews.

Review Date Range: From 2006 to 2017, with an average review year of 2012.

Cocoa Percentages: Range from **42% to 100%**, with a median around **70-75%** most bars tend to be in the dark chocolate range.

Ratings: Range from **1 to 5**, with an average rating of **3.19**. The middle 50% of ratings are between **2.88 and 3.5**, meaning most chocolates are rated fairly well, but only a few reach the perfect score.

Data limitations

Temporal Scope: The data sets were last updated 8 years ago, meaning newer chocolate bars are not included.

Missing Values: Some entries lack information on bean type or specific origin.

Limited Scope: Focuses primarily on plain dark chocolate, excluding chocolate varieties.

Subjectivity in Ratings: The ratings are based on expert opinions, which may introduce bias.

Data Ethics

Attribution and transparency:

Since the data is from Kaggle and is community contributed, proper citation is required.

Be clear about how the data was collected and any limitations in its scope.

Privacy and Anonymity

The dataset doesn't include personal details; it's good practice to respect their anonymity.

Bias Awareness

The data was gathered from chocolate enthusiasts, meaning it might favor high end chocolate rather than the general market, it's good to be aware of potential bias in ratings.

Responsible Use

If any findings end up being used for commercial purposes, it's worth double checking licensing rights to make sure everything stays compliant.

Questions to explore:

- Cocoa Content & Quality
  How does cocoa percentage correlate with ratings?
- Origin & Bean Variety
  Do certain cocoa bean varieties receive higher ratings?
  Does country of origin influence chocolate quality perception?

- Global Trends in Chocolate Preferences
  Are there regional preferences for higher or lower cocoa content?
  Do ratings vary by geographic origin of the chocolate makers?