Task 6.1_Mallika Banerjee

Data Source and Collection

The **Chocolate Bar Ratings** dataset was obtained from **Kaggle (chocolate-bar ratings)**. It contains **1,795 sensory reviews** of **440 unique chocolate bars**, contributed by a community of chocolate enthusiasts between **2006 and 2017**.

Key Attributes:

- **Company:** Name of the chocolate maker (*416 unique makers*)
- **Bar Name:** Specific bean origin or bar variant (*1,039 unique names*)
- **REF:** Unique bar identifier
- **ReviewDate:** Year of the Review
- **CocoaPercent:** Cocoa content expressed as a percentage
- **Location:** Country of production (*60 distinct countries*)
- **Rating:** Sensory score on a scale from *1.0 to 5.0*
- **BeanType:** Cacao varietal or blend (*41 types; ~50% missing values*)
- **BroadOrigin:** Country of bean origin (*100 countries*)

The reviews were aggregated from public submissions on Kaggle. Contributors—both professional and amateur tasters—provided detailed evaluations and scores after sampling each chocolate bar. Supplementary metadata, including origin, company location, and cocoa percentage, were supplied by tasters or sourced from manufacturer specifications.

This **community-driven dataset** offers a rich collection of sensory and production information. However, as it is based on voluntary contributions rather than systematic sampling, it may reflect certain participation biases.

Data Limitations

1. **Temporal Scope:**
   Reviews conclude in 2017, offering no insight into newer products or evolving market trends.
2. **Sampling Bias:**
   The dataset primarily reflects inputs from specialty chocolate enthusiasts, limiting representativeness of general consumer preferences.
3. **Missing Varietal Data:**
   Approximately 50% of *BeanType* entries are blank, constraining varietal-level analysis.
4. **Community Licensing:**
   Data is openly available for educational purposes; however, commercial use may require rights verification.

Why this data?

I selected this dataset due to my strong interest in **specialty foods and flavor profiling**. Its **geographic depth**—linking cacao bean origins with manufacturer locations—enables meaningful **spatial analysis**. The inclusion of both **continuous variables** (cocoa percentage, ratings, review year) and **categorical variables** (company, origin, bean type) allows for a wide range of analytical approaches. Overall, this dataset offers a compelling foundation for exploring how **origin, formulation, and time** influence **chocolate quality**.

Ethical Considerations

1. **Attribution:**
   The dataset is community-contributed on Kaggle; proper citation and acknowledgment of the source are essential.
2. **Privacy:**
   Although no personally identifiable information is included, contributor anonymity must be maintained and respected.
3. **Bias Awareness:**
   As the data is based on voluntary submissions, it may disproportionately represent niche or premium chocolate products. Recognizing this bias is crucial for accurate interpretation.
4. **Interpretation:**
   Conclusions should be presented with caution, avoiding overgeneralization. Any gaps, limitations, or contextual constraints should be clearly acknowledged when communicating findings.

Questions to Explore

- How does CocoaPercent correlate with Rating?
- What is the trend in average ratings from 2006 to 2017?
- Which BroadOrigin countries produce the highest average ratings?
- Do company location differences (e.g., U.S.A. vs. Europe) influence ratings?
- How do different BeanType categories compare in terms of average rating?
- Can a bar's Rating be predicted from its CocoaPercent, Origin, and Company using regression analysis?
- What do time trends suggest about changing preferences for darker vs. lower-percentage chocolate?

Data Cleaning Summary

- **Column Name Standardization:**
  Removed newline characters and irregular whitespace from all column headers. Renamed columns to standardized identifiers—**Company, BarName, REF, ReviewDate, CocoaPercent, Location, Rating, BeanType,** and **BroadOrigin**—for clarity and consistency.
- **Data Type Conversion:**
  Converted **CocoaPercent** from a string format (e.g., "70%") to a numeric float (70.0) by removing the percentage symbol. Ensured **ReviewDate** is stored as an integer year and **Rating** as a numeric value.
- **Missing Value Treatment:**
  Replaced **887 missing or blank entries** in *BeanType* with the label **"Unknown"** to retain all records. Noted **one missing value** in *BroadOrigin* and retained it intentionally for transparency in subsequent analysis.
- **Duplicate Verification:**
  Detected **no exact duplicate rows**. Confirmed that multiple appearances of the same **REF (bar ID)** are expected, as they represent distinct reviews for the same chocolate bar.
- **Data Integrity Checks:**
  Verified that all records contain valid entries for key fields—**Company, BarName, CocoaPercent, Rating,** and **ReviewDate**. Inspected for out-of-range values (e.g., cocoa percentages <0 or >100) and found none.