Project 6: Cricket Records

You've just finished watching *Moneyball*, the iconic film about how data analytics transformed baseball by uncovering hidden patterns and undervalued talent. Inspired by the way numbers told the real story behind the game, you've decided to bring that same sharp-eyed curiosity to the sport of cricket.

Your goal? To build a clean, structured dataset of batting averages from past Cricket World Cups, using data from ESPN Cricinfo — cricket's biggest online archive. Unfortunately, this treasure trove of statistics is tucked away in web pages with no direct API access.

As a passionate data enthusiast, your task is to scrape these batting records, clean and organize them into a structured format, and lay the groundwork for deeper analysis. You might uncover which players truly shined under World Cup pressure, how performances evolved across eras, or which countries have produced the most consistent tournament batsmen.

For this project, you'll:

- Scrape detailed batting average data from World Cup records pages.
- Tidy and standardize the data to handle inconsistencies and missing values.
- Transform the dataset into a form ready for insightful, Moneyball-style analysis.
- Explore the data through visualizations or statistical summaries to spot trends and standout performances.

Part 1: Data Collection

In this stage, you'll scrape batting average data for all players who have participated in the Cricket World Cup, organized by country, from ESPN Cricinfo's records pages.

Step 1: Access the World Cup Records Page

- Visit the page: ESPN Cricinfo.
- Locate the **"Player Averages"** section.
- Within this section, find and select the **"Batting Averages"** tab.

Step 2: Collect Country Names and Their Links using Python

- On the **Batting Averages** page, you'll find a list of links, each corresponding to a different country that has participated in the World Cup.
- Extract the following information for each country:
  - **Country Name**
  - **URL Link** to that country's batting averages page.
- Store this information in a structured list or table for use in the next step.

Step 3: Scrape Batting Averages for Each Country

- Loop through the list of country links collected in Step 2.
- For each country's batting averages page:
  - Scrape the relevant data for each player listed.
  - For each player's record, attach the **Country Name** from the link it came from.

Step 4: Store the Collected Data

- Organize the scraped data into a structured format such as a CSV file or DataFrame.

Part 2: Data Cleaning & Preparation

Now that you've scraped the raw batting averages data, it's time to clean, standardize, and enhance your dataset to make it analysis-ready. Follow these structured steps:

Step 1: Standardizing Missing Data

- Review the dataset for any placeholder symbols (such as dashes "–" or "-") used to represent missing values.
- Replace these placeholders with appropriate values to ensure they are recognized as missing data.

Step 2: Renaming Ambiguous Columns

- Inspect the dataset for any column names that are abbreviated, unclear, or not self-explanatory.
- Rename these columns to more descriptive titles. For example:
  - Change **'no'** to **'Not Out'**
  - Change **'HS'** to **'Highest Ending Score'**
  - Change **'Sr'** to **'Batting Strike Rate'**
  - Change **'Mat'** to **'Matches'**

Step 3: Handling Missing Values

- Identify columns containing missing or null values.
- Investigate these columns to determine whether the missing values are acceptable (e.g., due to incomplete historical records) or if they need to be filled.
- Replace missing values in numerical columns (like **'Balls Faced'** or **'Batting Strike Rate'**) with zeros if it makes sense in context.
- Carefully document any assumptions made during this process.

Step 4: Removing Duplicate Records

- Check the dataset for duplicate player records.
- If duplicates are found, review them to confirm whether they are exact copies or if any details differ.
- Remove any fully duplicated rows, keeping only one occurrence of each player’s record.

Step 5: Splitting the ‘Span’ Column

- Locate the **‘Span’** column, which represents the range of years a player was active in the World Cup (e.g. "1987-2003").
- Split this into two separate columns:
  - **‘Rookie Year’** (the first year in the range)
  - **‘Final Year’** (the last year in the range)
- Remove the original **‘Span’** column after creating these new fields.

Step 6: Correcting Data Types

- Review the data types for each column to ensure they are appropriate for the values they hold (e.g. integers for matches played, floats for strike rates).
- Pay attention to columns that may contain symbols like asterisks (‘*’) or other non-numeric characters, which can prevent proper conversion.
- Clean such characters from the affected columns.
- Convert columns like **‘Matches’**, **‘Highest Ending Score’**, **‘Balls Faced’**, **‘Batting Strike Rate’**, **‘Rookie Year’**, and **‘Final Year’** into their correct numerical types.

Step 7: Debugging Problematic Data

- During type conversion, be vigilant for any persistent issues such as unexpected symbols or invalid entries.
- Investigate problematic rows, especially if conversion errors occur.
- Decide whether to clean or remove problematic records that disrupt data type conversions.
- Recheck for null values and data types after adjustments.

Step 8: Creating a ‘Career Length’ Feature

- Introduce a new column called **‘Career Length’**.
- Calculate each player's career length by subtracting their **‘Rookie Year’** from their **‘Final Year’**.
- This new feature will allow for deeper analysis, such as exploring the relationship between career length and batting performance.

Part 3: Data Analysis & Visualization Suggestions

Now that your dataset is clean, structured, and rich with player performance details, it’s time to explore it visually. The aim here is to uncover trends, patterns, and interesting stories through data visualization. Below are some suggested analyses you can perform:

- Distribution of Batting Averages: Plot a histogram or density plot of batting averages to see how player performances are distributed.
- Top 10 Highest Career Batting Averages: Create a bar chart of the top 10 players with the highest career batting averages in World Cup history.
- Average Batting Strike Rate by Country: Use a boxplot or bar chart to compare the distribution or average of batting strike rates across different countries.
- Rookie Year Trends: Use a line chart or bar plot to visualize the number of new players (rookie debuts) per World Cup year.