**Design Document for Cricket Dataset Analysis**

**1. Introduction**

This notebook analyzes a dataset of cricket deliveries to derive insights into player performance, specifically focusing on batsman statistics. The dataset includes information such as match details, batting and bowling teams, runs scored, and wicket information. This analysis computes key statistics like batting average, strike rate, and other player metrics such as hundreds and fifties.

**2. Libraries Used**

**2.1 Pandas**

* **Purpose**: Used for data manipulation and analysis. It is employed for reading the CSV file, handling missing values, grouping data, and calculating aggregates.
* **Functions**: pd.read\_csv(), df.groupby(), df.agg(), df.fillna()

**2.2 Matplotlib**

* **Purpose**: Used to create visualizations such as histograms and bar charts to help interpret the dataset visually.
* **Functions**: plt.hist(), plt.figure(), plt.bar(), plt.title(), plt.xlabel(), plt.ylabel()

**3. Loading the Dataset**

* **Description**: The pd.read\_csv() function reads the cricket delivery dataset from a specified file path and loads it into a DataFrame df. This DataFrame will be the primary object for data manipulation throughout the analysis.

**4. Data Overview**

* **Description**:
  + df.head() displays the first five rows of the dataset, providing an overview of the available columns such as match\_id, batting\_team, batter, batsman\_runs, etc.
  + df.info() gives the structure of the DataFrame, including the number of non-null entries in each column, the data types, and memory usage. This helps identify columns with missing values.

**5. Statistical Analysis of Runs and Wickets**

**5.1 Runs**

* **Description**: This section calculates basic statistics for batsman\_runs:
  + **Mean**: Average runs scored per delivery.
  + **Median**: Middle value when the runs are sorted.
  + **Mode**: Most frequent run value.

**5.2 Wickets**

* **Description**: This section calculates statistics for the is\_wicket column, which indicates whether a wicket was taken (1 for yes, 0 for no):
  + **Mean**: Average number of wickets per delivery.
  + **Median**: Middle value for wickets.
  + **Mode**: Most common value (either 0 or 1).

**6. Visualizations**

**6.1 Distribution of Runs**

* **Description**: This histogram shows the distribution of runs scored per delivery. It helps understand the frequency of each run value (e.g., how often a batsman scores 0, 1, 2, etc.).

**6.2 Wickets Distribution**

* **Description**: A bar chart that shows the number of deliveries with and without wickets. The color scheme differentiates between deliveries where a wicket was taken (green) and where no wicket was taken (orange).

**7. Batsman Performance Calculation**

**7.1 Grouping Data by Batsman**

* **Description**:
  + The data is grouped by the batter column to aggregate batsman-specific statistics:
    - **total\_runs**: Total runs scored by each batsman.
    - **balls\_faced**: Total number of balls faced.
    - **dismissals**: Count of times the batsman was dismissed.
    - **hundreds**: Count of hundreds (runs >= 100).
    - **fifties**: Count of fifties (runs >= 50 and < 100).

**7.2 Batting Average and Strike Rate**

* **Description**:
  + **Batting Average**: Total runs divided by the number of dismissals.
  + **Strike Rate**: Total runs divided by balls faced, multiplied by 100.

**7.3 Normalizing Metrics**

* **Description**: This section normalizes each statistic to a scale of 0-100, where the maximum value for each column is transformed to 100.

**8. Scoring Batsman Performance**

**8.1 Assigning Weights and Calculating Overall Score**

* **Description**:
  + Weights are assigned to each metric to reflect their importance (e.g., strike rate is weighted more heavily than fifties).
  + The overall\_score is computed as a weighted sum of the normalized metrics.

**9. Final Output: Top Batsmen**

* **Description**:
  + The batsmen are sorted by the overall\_score in descending order, and the top 10 batsmen are displayed based on their overall performance score.

**10. Conclusion**

This analysis provides insights into individual batsman performance in the dataset, including key metrics such as batting average, strike rate, hundreds, and fifties. The overall score combines these metrics to rank batsmen. The visualizations further enhance the understanding of the distribution of runs and wickets across the dataset.