**NCAA EDA**

the Exploratory Data Analysis (EDA) provided is well-suited to support answering the primary research question:

"Can we predict a player's NCAA division (Division I, II, or III) based on their demographic and physical attributes (e.g., height, year/class, state or country of origin, position)?"

Here’s a detailed assessment of why this EDA aligns effectively with my research question:

**How the EDA supports my research question:**

**1. Instant Reports (Summary Statistics)**

* Quickly identify distribution patterns, ensuring my data is balanced and representative across NCAA divisions, player attributes (height, class, position), and geography.

**2. Descriptive Analysis**

* Clarifies the distributions of categorical features (e.g., position, state, year/class) across NCAA divisions.
* Detects possible skewness or imbalance that could influence the predictive modeling process.
* Reveals important insights such as whether certain positions or geographic origins are more prevalent in specific divisions.

**3. Correlation and Relationship Analysis**

* Helps detect if numeric attributes like total\_inches (height) have strong correlations or associations with the NCAA division, thus indicating their importance as predictive features.
* Provides insights into potential feature importance and multicollinearity, guiding the selection of meaningful predictors for my machine learning model.

**How to interpret EDA to effectively address my question:**

* **Feature importance:**  
  Identify attributes that clearly differ among divisions. For instance, if player height or positions vary notably between divisions, these become valuable predictive indicators.
* **Class distribution (balance of data):**  
  Confirm that you have sufficient representation from each NCAA division. If not, consider techniques like oversampling or undersampling when modeling.
* **Multicollinearity and redundancy:**  
  Identify highly correlated or redundant features through the correlation matrix. This aids in selecting or eliminating redundant predictors, improving model performance.