NBA Dataset EDA

Introduction:

This report presents an analysis based on the visualizations and explorations of the NBA Dataset sourced from Kaggle. The focus is on uncovering trends, identifying relationships, and spotting anomalies in player statistics, country distributions, and performance indicators.

1. Dataset and Data Cleaning

- **Outliers**: Significant outliers were identified through box plots, revealing unusual data points in several key attributes. These were addressed with appropriate techniques tailored to the distribution of each variable.
- **Missing Data**: Approximately 15% of the dataset contained missing values, which were managed using an imputation method to ensure data completeness without introducing bias.
- **Categorical Typos**: Columns such as *College* and *Country* exhibited typographical errors and inconsistencies. These were standardized to improve data accuracy.

2. Univariate Analysis

2.1 Country Distribution:

- The pie chart analysis indicates that 83.8% of players hail from the USA, which reflects the country's long-standing dominance and influence in professional basketball.
- 16.2% of the players are international, representing the growing globalization of the NBA. This includes countries like Canada, Australia, and European nations contributing a smaller yet significant talent pool.
- **Insight**: While the NBA remains predominantly American, the increasing diversity highlights the league's effort to scout talent worldwide. Programs aimed at fostering international players appear to have had substantial success.

2.2 Points Per Game (PPG) Distribution

- Scoring Hierarchy in the NBA: The skewed distribution of points per game (PPG) illustrates a clear stratification among players. While the majority (scoring between 2 and 10 PPG) serve as role players or specialists, only a small elite group surpasses the 30+ PPG mark. These outliers signify the league's top-tier players who contribute disproportionately to their teams' scoring outputs and often define their eras.

- Scoring Evolution and League Dynamics: The right skew of the PPG data could also reflect changing offensive strategies in the NBA over time. For example, a heavier emphasis on team-based scoring in certain eras could explain the relatively low median scoring range, as opposed to the isolation-heavy playstyles that benefit star players.

- Opportunities for Role Players: The clustering of players in the 2-10 PPG range highlights a consistent demand for role players in the league. This group likely includes players specializing in defense, rebounding, or passing, rather than scoring, which reinforces the diverse skill sets required in professional basketball.

- **Insight:** While the data confirms the dominance of a few scoring superstars, it also underscores the importance of team dynamics and the varied roles players fulfill. Exploring the trends across eras or comparing PPG distributions between star players and role players could yield further insights into how team compositions and strategies evolve.

2.3 Player Performance and Attributes Distribution

2.3.1 Physical Attributes (Height, Weight, Age):

- Height: Most players fall within the 6'4" to 6'9" range, reflecting the positional demands of basketball. Outliers at both extremes signify specialized roles like guards (shorter) or centers (taller).

- Weight: Concentrated around 210–230 lbs, indicating a balance between agility and physicality necessary for the game.

- Age: Peaks between 22 and 30 years, declining sharply after 35. This aligns with the physically demanding nature of professional basketball.

- Insight: These distributions emphasize the specificity of physical attributes tied to success in basketball positions and the challenges of career longevity in a high-intensity sport.

2.3.2 Performance Metrics (PPG, REB, AST, NET_RATING):

- **Points Per Game (PPG)**: The majority of players average 5–15 PPG, with a small elite cohort exceeding 30 PPG, indicating the league's reliance on a few superstar scorers.

- **Rebounds (REB) & Assists (AST)**: Players generally average 2–8 rebounds and 1–5 assists, showing the dominance of specialized roles (e.g., big men for rebounds, guards for assists).

- **Net Rating (NET_RATING)**: A balanced distribution reflects varying levels of contribution to team performance, with outliers indicating standout or underperforming players.
- **Insight**: The dominance of role players, paired with a few exceptional performers, highlights the NBA's dependence on both teamwork and individual brilliance to succeed.

2.3.3 Advanced Metrics (USG_PCT, TS_PCT, AST_PCT):

- **Usage Percentage (USG_PCT)**: Most players are low-usage contributors, with a few dominating possessions—a testament to star-focused offensive systems.
- **True Shooting Percentage (TS_PCT)**: Concentrated around 50–60%, indicating the critical role of efficiency in scoring success.
- **Assist Percentage (AST_PCT)**: Low averages suggest concentrated playmaking among specific positions like guards.
- **Insight**: Advanced metrics reveal the nuances of team strategies, showcasing the importance of efficiency and specialized roles in overall performance.

3. Bivariate Analysis

3.1 Distribution of Players Across Seasons Played

- **Observation**: The majority of players have careers spanning 1 to 10 seasons, with 10 seasons being the most common (1,040 players). Beyond this, the number of players decreases significantly, with only a few sustaining careers beyond 15 seasons, and just 21 players playing 21 or more seasons.
- **Insight**:
  - The steep drop-off after 10 seasons highlights the competitive and physically demanding nature of the NBA, where longevity is reserved for exceptional talents who can adapt and consistently perform.
  - Players who sustain careers of 15+ seasons are likely those with a combination of exceptional skill, versatility, and adaptability to evolving playstyles or team needs.

3.2 Trends in Maximum Points Per Game (PPG)

- **Observation**: The line graph shows fluctuations in the maximum PPG recorded each NBA season from 1996-97 to 2022-23. Notable peaks include 35.4 PPG (2005-06) and 36.1 PPG (2018-19). The most recent season, 2022-23, recorded a maximum of 33.1 PPG.
- **Insight**:
  - The peaks correspond to seasons dominated by superstars known for their high-scoring prowess, such as Kobe Bryant (2005-06) or James Harden (2018-19). These values may also signify periods where offensive playstyles were prioritized across the league.

    - The decline in the most recent season (33.1 PPG) suggests either a shift back to more team-oriented scoring or a lack of players hitting previous superstar scoring thresholds.

3.3 Top Impactful Players: Games Played vs. Net Rating

Observations:

- Players with the highest Net Ratings include Bruce Bowen (~27.5) and Max Strus (~25), despite having relatively low games played (~35 each). This indicates significant impact in limited appearances, likely as role players in highly efficient systems.

- Sekou Doumbouya (~17.5) also exhibits strong efficiency with a similar number of games played, reflecting potential within specific contexts or rotations.

- Veteran players such as Tim Duncan, John Stockton, Manu Ginobili, and David Robinson demonstrate consistent moderate Net Ratings (~10–12.5) over a larger number of games (60–75), emphasizing longevity combined with sustained impact.

- Robert Parish (~15) stands out for maintaining a strong Net Rating with slightly fewer games (~45), highlighting his effectiveness within that duration.

Insights:

- High Net Ratings in Limited Games: Players like Bowen and Strus show that strategic deployment in specific roles or systems can significantly amplify efficiency, even with fewer minutes or games. These players may excel in team-oriented roles such as three-and-D specialists or situational contributors.

- Sustained Efficiency in Long Careers: The consistent performance of veterans like Duncan, Stockton, and Ginobili underscores their adaptability and ability to maintain high-level contributions despite aging and increased demands over extended careers.

- Variation Across Roles: The players' distribution suggests diverse roles—some excel as system-driven contributors with fewer games, while others display longevity-driven sustained excellence. Teams leveraging both types of players likely balance short-term results with long-term consistency.

3.4 Correlations Among Numerical Variables

The heatmap offers a clear window into how player physical attributes and on-court performance metrics interact:

- Physical Coupling vs. Functional Specialization
  - **Strong Positive Relationship:** A robust correlation between player height and weight (≈0.82) confirms that body size is fundamentally interlinked. This reinforces traditional positional roles, where size is a primary criterion for evaluating defensive and rebounding responsibilities.
  - **Inverse Association:** The significant negative correlation (≈–0.63) between height and assist percentage suggests that taller players are generally less involved in creating scoring opportunities. This supports a strategic disconnect, as playmaking is typically reserved for more agile, shorter players.
- **Interplay of Efficiency and Usage**: While not detailed numerically here, moderate correlations among usage percentage, true shooting percentage, and points hint that high-volume players are not automatically the most efficient scorers. Instead, the balance between effort (usage) and efficiency illustrates that overall impact is nuanced—star roles thrive not only on volume but also on scoring efficiency, a relationship subtly captured by net rating.
- **Insight:** The heatmap doesn't just confirm expectations; it emphasizes a layered portrait of basketball performance. Physical traits dictate certain positional predispositions—taller players deliver through size and strength, while shorter players excel in facilitation and dynamic playmaking. This balance, now quantified through these correlations, provides deeper context for how teams might evolve roles and scouting strategies in an era of versatile playstyles.

4. Multivariate Analysis:

4.1 Best Scoring Seasons of All Time

This scatter plot weaves together multiple dimensions—season (year), individual players, and their points per game (PPG)—revealing nuanced narratives behind NBA scoring excellence.

- **Era-Spanning Scoring Peaks**: While the data spans more than two decades (2002–03 to 2022–23), the elite scoring performances are not confined to one specific period. Legends like Kobe Bryant and Allen Iverson in the mid-2000s, alongside modern stars such as James Harden (with multiple top-scoring seasons), Joel Embiid, and Luka Doncic in recent years, indicate that exceptional scoring output is both timeless and adaptable. This reveals how evolving styles—whether it's the isolation play of previous decades or today's pace-and-space strategy—still yield historical milestones when a player is at their peak.
- **Sustained Excellence Versus Burst Performances**: Players like James Harden, who appear more than once with different high-scoring seasons, hint at a sustained level of offensive prowess rather than a one-off anomaly. In contrast, a couple of single-season performances (e.g., Kobe's 35.4 PPG) suggest seasons where conditions dovetailed perfectly with individual talent. This dynamic underscores the interplay between personal skill, team strategy, and even rule evolutions that create an environment ripe for record-setting figures.
- **Contextual Clues Beyond Raw Numbers:** The elevated PPG values seen in these seasons are not merely isolated statistics but are products of multiple factors—team usage, offensive schemes tailored around star players, and possibly favorable rule changes. For example, the modern clustering of top scorers in recent years may reflect analytics-driven offense and a league-wide embrace of high-impact, three-point shooting strategies.
- **Insight:** This multivariate view not only celebrates individual scoring brilliance but also illustrates how systemic trends and strategic evolutions across eras amplify a player's ability to eclipse conventional benchmarks. The scatter plot invites further exploration into how

factors like playing time, team dynamics, and even defensive adjustments contribute to these record-breaking seasons—a fertile ground for deeper, contextual analysis.

4.2 Pair plot analysis

The pair plot integrating age, player height, points (pts), and games played (gp) reveals subtle yet profound insights into player performance and career dynamics:

- **Age as a Performance Pivot**: The histograms show a concentration of players in their mid-to-late 20s—likely the peak performance window. The scatter plot between age and points hints at a non-linear relationship: players tend to reach optimal scoring during these years before tapering off, suggesting that both rookie development and veteran adjustments shape scoring output.

- **Height and Role Differentiation**: While player height is clearly distributed across a broad spectrum, its relationship with points is diffuse. This implies that physical stature predominantly determines defensive or positional roles rather than directly driving scoring efficiency. Taller players, for example, may focus on interior defense and rebounds, while scoring is increasingly reliant on skill and role specialization.

- **Games Played and Impact**: The association between games played and overall points reaffirms that consistent on-court exposure generally translates to higher scoring totals. Yet, the scatter reveals clusters where high-scoring efficiency is achieved despite fewer game appearances. This suggests that some players excel as situational impact players, perhaps utilized strategically to maximize scoring potential in limited minutes.

- **Insight**: This multivariate view underscores that athletic success in the NBA is a multifaceted construct. Peak performance emerges not from a single attribute but from the interplay of age, physicality, and situational game experience—a dynamic balance that shapes both individual careers and team strategies.

4.3 Box plots across Features

- **By Teams & Draft Year:**
    - Teams: Box plots show that some franchises maintain a narrow age range—reflecting a focus on youth and cohesive development—while others exhibit broader distributions, indicating transitional phases with mixed veteran and rookie presences.
    - Draft Year: Older draft classes display a wider age spread, whereas recent drafts feature more uniform, younger profiles, reflecting modern scouting and developmental practices.
- **By Draft Rounds & Seasons:**
    - Draft Rounds: Early-round selections are consistently younger with tight age spreads, underscoring teams' prioritization of long-term potential. Later rounds show greater variability, often including older players from non-traditional paths.
    - Seasons: Seasonal trends (from 1996–97 to 2022–23) illustrate a gradual tightening of age distributions, mirroring changes in league recruitment, eligibility rules, and the rise of new development avenues like the G League.
- **Insight:** The combined insights reveal that NBA teams strategically shape their rosters through both draft practices and evolving recruitment trends, with an increasing emphasis on youth and long-term development. This reflects a broader evolution in talent acquisition that balances immediate performance with sustainable growth.

5. Conclusion

The analysis reveals that the NBA's on-court dynamics are driven by a blend of individual talent and strategic roster management:

- Data Quality: Rigorous cleaning ensured reliable insights, addressing outliers and missing values across 13k entries.
- Player Roles & Performance: Univariate and bivariate insights confirm a clear scoring hierarchy—where a few elite performers drive points, supported by specialized role players whose physical attributes (height, weight) directly influence their responsibilities.
- Evolving Talent Dynamics: Multivariate examinations show a strategic shift in roster composition. Teams and modern draft practices now emphasize a uniform, youthful cohort, reflecting consistent scouting and long-term development. In contrast, older draft classes exhibit wider age variability, highlighting past diversity in player entry.