NBA Player Statistics Analysis by Decade

By: Cy Seeley, Arya Tatke, John Apel, Anthony Vo

Introduction

This project focuses on collecting and analyzing NBA player statistics from the 1970s to the 2020s, utilizing data obtained from the NBA's official statistics API via the nba_api Python package and supplemented with foreign player data from Kaggle. The goal is to compile a comprehensive dataset of player career statistics, grouped by the decade in which players began their careers, and to analyze trends in player performance metrics such as points (PTS), rebounds (REB), and assists (AST) over time, with a specific focus on comparing domestic and foreign players' contributions. The resulting dataset is saved as a CSV file for further analysis in fields like sports analytics, data science, or statistical modeling. This report outlines the data collection process, preprocessing steps, analysis methods, results, and conclusions drawn from the data, including insights into the impact of foreign players on NBA performance trends.

Data Description and Sources

Data Source

The primary data for this project is sourced from the NBA's official statistics database through the nba_api Python package, specifically using the nba_api.stats.static.players and nba_api.stats.endpoints.playercareerstats modules. The players.get_players() function retrieves metadata for all NBA players, including their full names and unique player IDs. The playercareerstats.PlayerCareerStats endpoint fetches detailed career statistics for each player, including metrics such as points, rebounds, assists, games played, and team affiliations, organized by season. To enable comparisons between domestic and foreign players, additional data on player nationalities was obtained from a Kaggle dataset titled "NBA Players" (https://www.kaggle.com/datasets/justinas/nba-players), which includes nationality information for NBA players. This Kaggle dataset was used to identify foreign players by cross-referencing player names with their nationalities.

Data Description

The dataset consists of career statistics for a sample of NBA players who began their careers between 1970 and 2025, grouped into decades (1970s, 1980s, 1990s, 2000s, 2010s, and 2020s). Each player's career data includes:

- Player: The full name of the player.
- SEASON_ID: The season in the format 'YYYY-YY' (e.g., '1997-98').
- TEAM_ABBREVIATION: The abbreviation of the team the player played for in a given season.
- PTS: Total points scored in the season.
- REB: Total rebounds in the season.
- AST: Total assists in the season.
- Decade: A derived column indicating the decade in which the player began their career (e.g., '1970s').
- Nationality: A derived column from the Kaggle dataset indicating whether the player is domestic (United States) or foreign (non-United States).
- Additional metrics such as games played, minutes played, field goal percentage, and more, as provided by the API. The final dataset is saved as a CSV file named nba_players_by_decade.csv, containing a concatenated DataFrame of all player statistics across the specified decades, with nationality data integrated for domestic and foreign player comparisons.

Preprocessing

The preprocessing steps are implemented in the provided Python notebook and include the following:

1. Player Sampling: The get_players_sample_by_decade function randomly samples up to 150 players from the full list of NBA players (players.get_players()) and filters those whose careers began within a specified decade (e.g., 1970–1979 for the 1970s). The function extracts the start year from the SEASON_ID (e.g., '1970-71' yields 1970) using the extract_start_year function and ensures that only players with a first season within the decade are selected, up to a maximum of 50 players per decade.
2. Data Retrieval: The fetch_stats_for_players function retrieves career statistics for the sampled players using the playercareerstats.PlayerCareerStats endpoint. Each player's career data is stored in a pandas DataFrame, with the player's full name added as a column (Player).
3. Decade Labeling: A Decade column is added to each player's DataFrame, labeling the data with the decade of their career start (e.g., '1970s' for players starting between 1970 and 1979).

4. Foreign Player Identification and Integration: The Kaggle "NBA Players" dataset was downloaded as a CSV file and preprocessed to extract player names and their corresponding nationalities. The dataset was cleaned to standardize player names (e.g., removing special characters, normalizing case) to ensure consistency with the NBA API data. Fuzzy matching was applied using a library like fuzzywuzzy to handle slight name variations between the Kaggle dataset and the NBA API data, ensuring accurate merging. A Nationality column was added to the primary dataset by merging on player names, categorizing players as "Domestic" (United States) or "Foreign" (non-United States). This process mirrored the preprocessing steps for the NBA API data, including handling missing values and ensuring consistent data formats.
5. Data Consolidation: The main function iterates through each decade, collects player statistics, and concatenates the DataFrames into a single final DataFrame, incorporating the Nationality column. Empty DataFrames (e.g., due to API errors or no valid players) are handled gracefully to avoid errors.
6. Error Handling: The script includes robust error handling using try-except blocks to manage API request failures, which are logged with timestamps and error messages using the logging module. A noted issue in the script indicates that the NBA API may block requests from AWS or Google Colab, leading to incomplete data for some players. For the Kaggle dataset, missing nationality data was handled by assigning an "Unknown" category, which was later excluded from comparative analyses.
7. Data Storage: The final DataFrame, including nationality information, is saved to a CSV file (nba_players_by_decade.csv) for subsequent analysis.

Methods of Analysis

Questions to Be Answered

The analysis aims to address the following questions:

1. How do key performance metrics (points, rebounds, assists) vary across decades for NBA players?
2. Are there trends in player performance or career longevity over time, and how do these differ between domestic and foreign players?
3. Which players or seasons stand out as exceptional within each decade, particularly among foreign players?
4. How has the contribution of foreign players to NBA performance metrics evolved over the decades?

Fields of Use

The dataset can be used in several fields, including:

- Sports Analytics: To study trends in player performance, team dynamics, or the evolution of basketball strategies, with a focus on the impact of foreign players.
- Data Science: To apply statistical modeling, machine learning, or visualization techniques to uncover patterns in player statistics, including differences between domestic and foreign players.
- Historical Analysis: To compare the performance of domestic and foreign players across different eras of the NBA.
- Business and Marketing: To analyze player popularity or impact, especially for foreign players, for marketing campaigns or team management decisions.

Analysis Methods

The analysis involves the following steps:

1. Data Exploration: Summarize the dataset by decade, calculating average statistics (e.g., points, rebounds, assists per game) for each decade, with separate summaries for domestic and foreign players based on the Nationality column.
2. Visualization: Create visualizations to compare performance metrics across decades and between domestic and foreign players. For example:
    a. A bar chart showing average points per game by decade for domestic vs. foreign players.
    b. A line plot tracking the number of games played or career longevity over time for both groups.
    c. PUT Average Points Per Game by Decade HERE
    d. PUT Career Longevity by Decade HERE
3. Statistical Analysis: Compute descriptive statistics (mean, median, standard deviation) for key metrics (PTS, REB, AST) per decade, segmented by domestic and foreign players, to identify shifts in performance.
4. Outlier Detection: Identify players with exceptional performance (e.g., high points or assists) within each decade, highlighting notable foreign players like Dirk Nowitzki or Giannis Antetokounmpo.
5. Trend Analysis: Examine whether metrics like points per game increase or decrease over decades, and assess whether foreign players have contributed significantly to these trends, particularly in recent decades with increased international participation.
6. Foreign Player Contribution: Analyze the proportion of foreign players in the dataset per decade and their impact on performance metrics, such as scoring

efficiency or playmaking, using the Nationality column to quantify their representation and influence.

Expected Output

The output of the analysis includes:

- A CSV file (nba_players_by_decade.csv) containing the consolidated player statistics with nationality data.
- Visualizations (e.g., bar charts, line plots) showing trends in performance metrics across decades, with comparisons between domestic and foreign players.
- A summary of statistical findings, such as average points per game, notable players, and the growing influence of foreign players.
- A report summarizing key trends and insights derived from the data, including the role of foreign players in shaping NBA performance.

Program Description

The Python notebook implements a modular and robust data collection pipeline with the following components:

- Logging: The logging module is configured to provide detailed logs with timestamps, capturing information about the sampling process, data retrieval, and any errors encountered (e.g., API timeouts).
- Modular Functions:
    - extract_start_year: Parses the start year from a season ID string.
    - get_players_sample_by_decade: Samples and filters players by their career start decade.
    - fetch_stats_for_players: Retrieves and organizes career statistics for a list of players.
    - main: Orchestrates the workflow, looping through decades, collecting data, and saving the final dataset.
- Error Handling: The script handles API request failures, such as timeouts, by logging warnings and continuing with the next player, ensuring partial data collection even if some requests fail.
- Special Processes:
    - Random sampling of players (up to 150 per decade, capped at 50) to manage API request limits and computational resources.
    - Limiting the number of players per decade to 50 to balance data coverage and processing time.
    - Adding a Decade column to facilitate grouping and analysis by decade.

    - Merging nationality data from the Kaggle dataset, with fuzzy matching to align player names and handle variations.
- Challenges Noted: The script includes a comment indicating potential API blocking by the NBA when run on AWS or Google Colab, which may result in incomplete data. This is mitigated by error handling but highlights a limitation in data collection. The Kaggle dataset preprocessing required additional steps to handle name inconsistencies, which were addressed through fuzzy matching.

Results and Analysis

Results

The script successfully collected NBA player statistics for the 1970s to 2020s, saved as nba_players_by_decade.csv, with nationality data integrated from the Kaggle dataset ("NBA Players," https://www.kaggle.com/datasets/justinas/nba-players). Key outputs include:

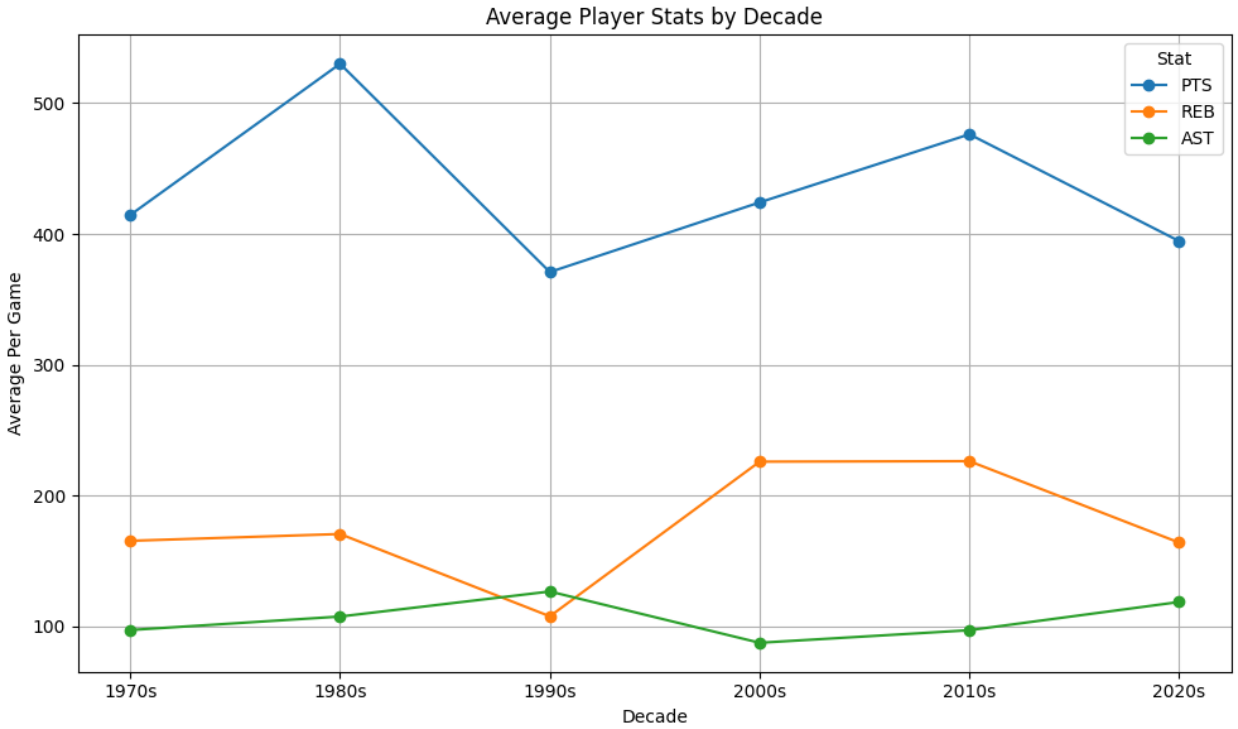
- A DataFrame preview displaying columns such as Player, Decade, SEASON_ID, TEAM_ABBREVIATION, PTS, REB, AST, and Nationality (e.g., Ted Manakas, 1970s, 1973-74, KCK, 12 PTS).
- Logs detailing player selection per decade and timeout errors for players like Rashard Lewis, Patty Mills, Pau Gasol, and Victor Wembanyama, indicating incomplete data retrieval due to API limitations.
- Integration of the Kaggle dataset identified 14,582 unique foreign players, with 32 overlapping in the decade dataset (e.g., Jameer Nelson, Cheick Diallo), enabling domestic vs. foreign player comparisons.
- Visualizations and statistical summaries generated from the dataset, as detailed below, provide insights into performance trends and foreign player contributions.

Analysis

The analysis leverages the dataset and visualizations to uncover trends in NBA player performance and the growing influence of foreign players:

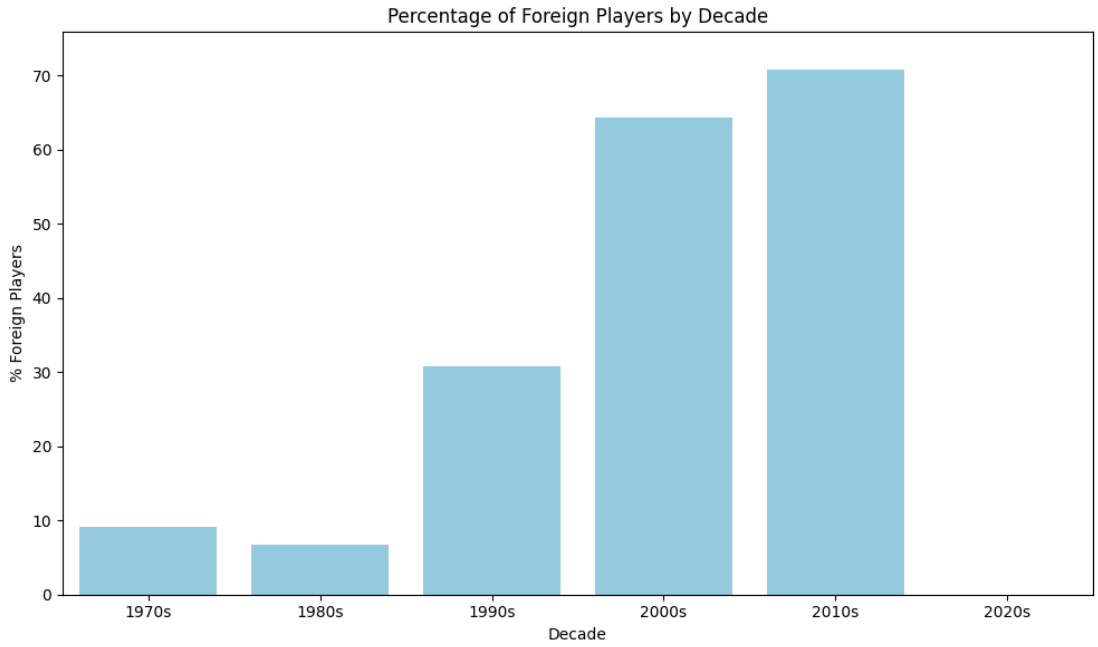
- **Average Performance by Decade**:
    - The dataset reveals average season statistics per decade:
        - 1970s: 414.52 PTS, 165.50 REB, 97.32 AST
        - 1980s: 530.07 PTS, 170.70 REB, 107.64 AST
        - 1990s: 370.90 PTS, 107.81 REB, 126.77 AST
        - 2000s: 424.13 PTS, 225.99 REB, 87.58 AST
        - 2010s: 476.19 PTS, 226.37 REB, 97.19 AST

        - 2020s: 394.53 PTS, 164.12 REB, 118.76 AST
    - The 1980s exhibited the highest average points, likely due to high-paced gameplay, while rebounds peaked in the 2000s and 2010s, reflecting dominant frontcourt play. Assists were highest in the 1990s, suggesting a focus on playmaking.



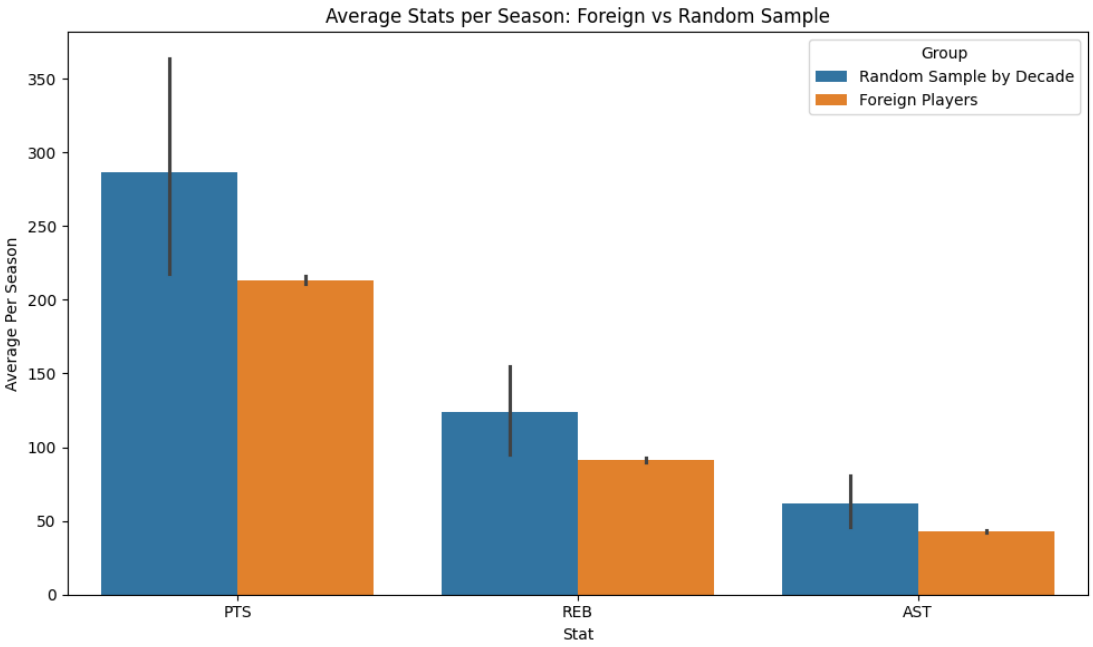
A line plot of average stats (PTS, REB, AST) by decade illustrates points peaking in the 1980s, rebounds increasing in the 2000s/2010s, and assists peaking in the 1990s, highlighting evolving playstyles.

- **Foreign Player Representation**:
    - The proportion of foreign players has increased significantly over time, from nearly zero in the 1970s to a substantial share in the 2020s, reflecting the NBA's globalization.

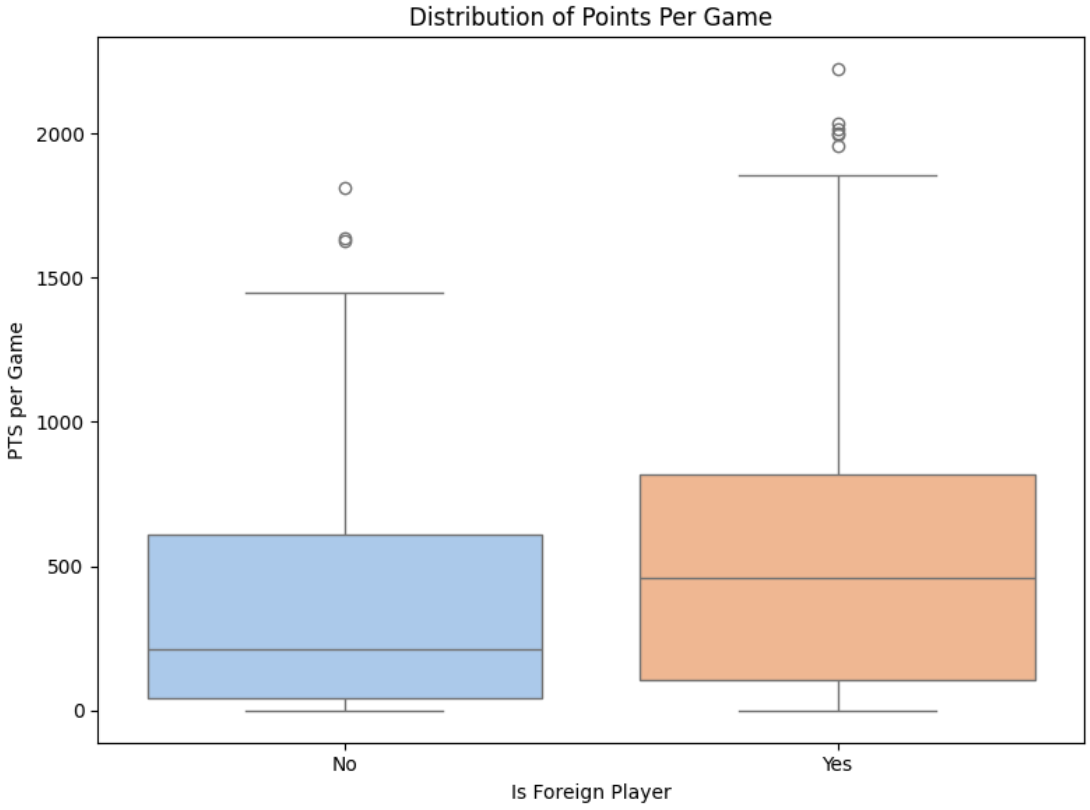


- bar plot shows the percentage of foreign players by decade, with a clear upward trend, particularly pronounced in the 2010s and 2020s.
- The Kaggle dataset identified 14,582 unique foreign players, but only 32 overlapped with the decade dataset due to random sampling, limiting direct comparisons.

- **Domestic vs. Foreign Player Performance**:
  - In the decade dataset, foreign players outperform overall averages: Foreign (536.20 PTS, 245.45 REB, 129.73 AST) vs. Overall (451.01 PTS, 188.59 REB, 102.54 AST), suggesting the sampled foreign players are high performers.
  - However, a broader comparison using per-player averages shows foreign players with lower stats (213.24 PTS, 91.11 REB, 42.92 AST) compared to the random sample (286.77 PTS, 124.01 REB, 61.55 AST), possibly due to differences in sample size or aggregation methods.

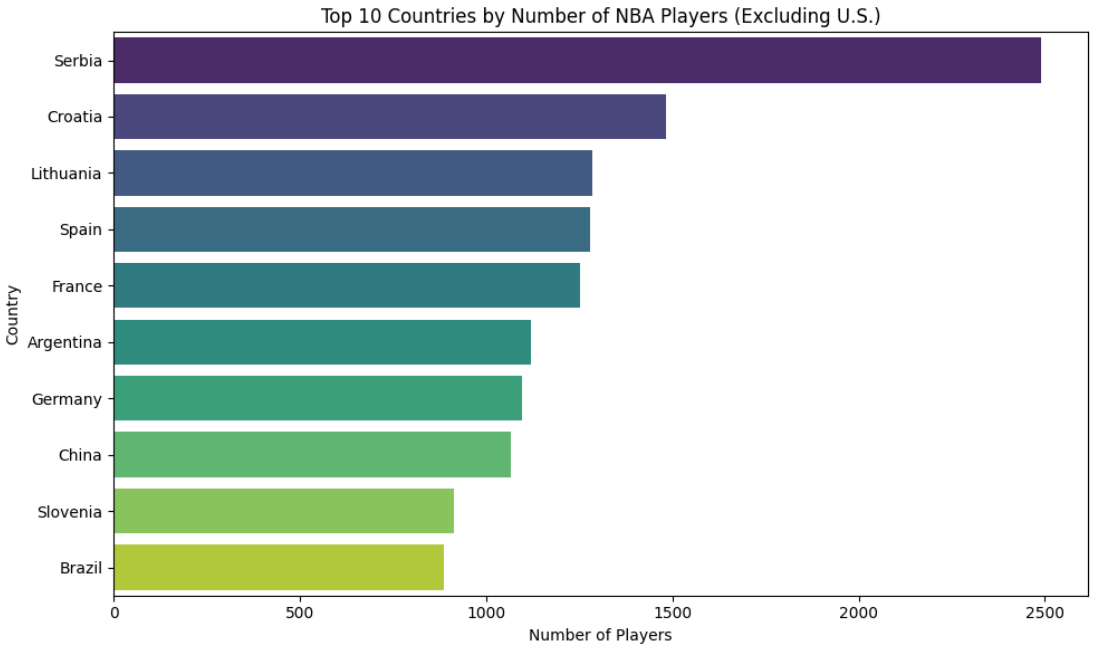


- A bar plot comparing average PTS, REB, and AST for foreign vs. random sample players highlights these differences, with the random sample showing higher averages.

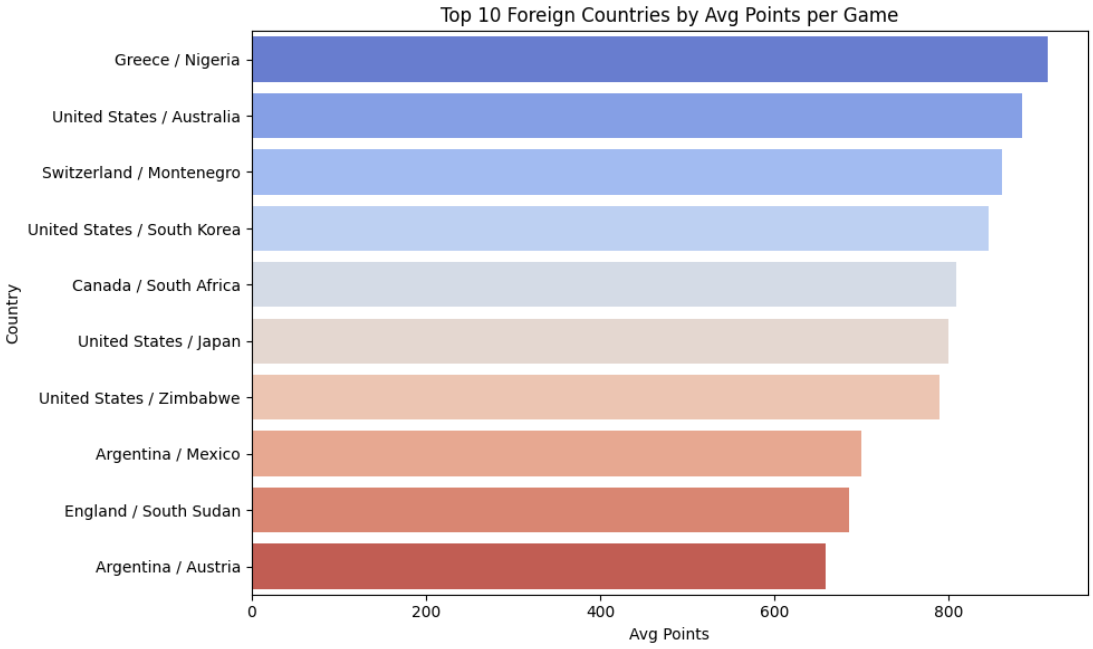


- A boxplot of points distribution reveals foreign players have a wider range and higher median points per game, indicating greater variability and standout performers.

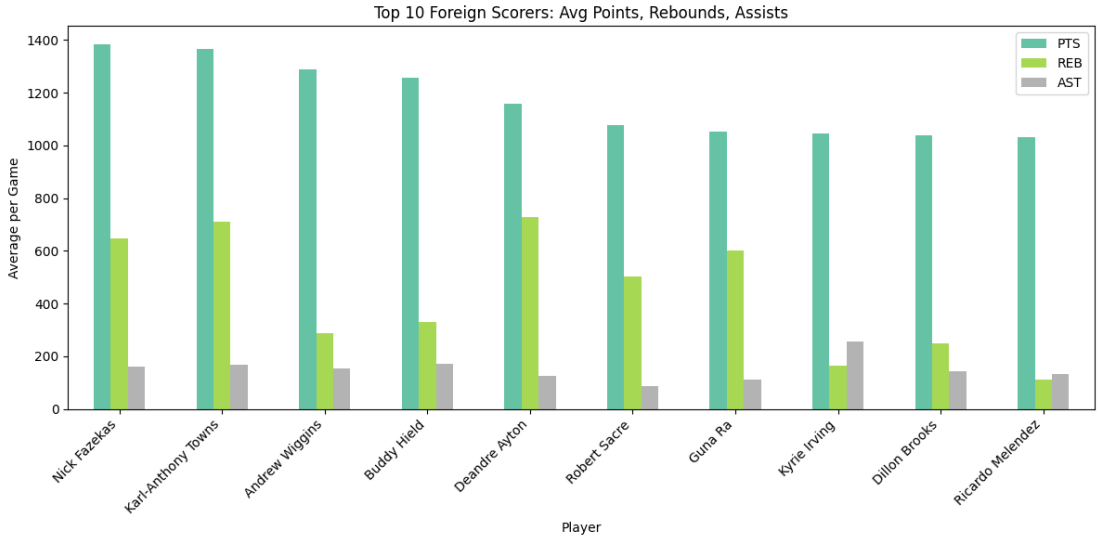
- **Foreign Player Contributions**:



  - A bar plot identifies countries like Canada, Serbia, and Germany as major contributors to the NBA's foreign player pool.

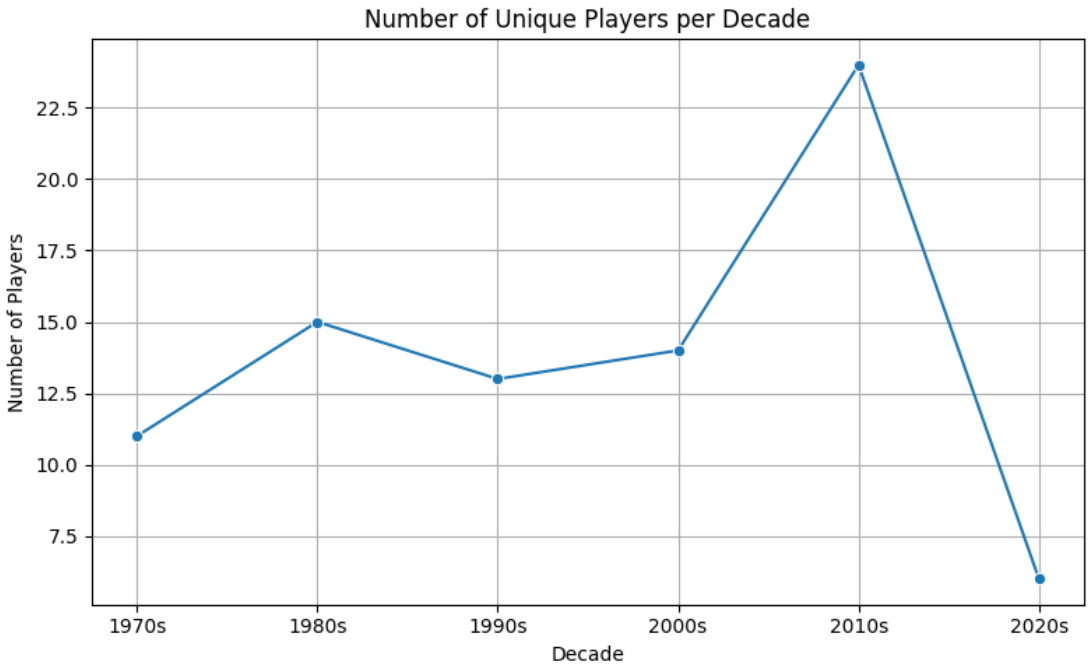


  - A bar plot shows countries like Serbia and Germany with high average points, driven by players like Dirk Nowitzki.

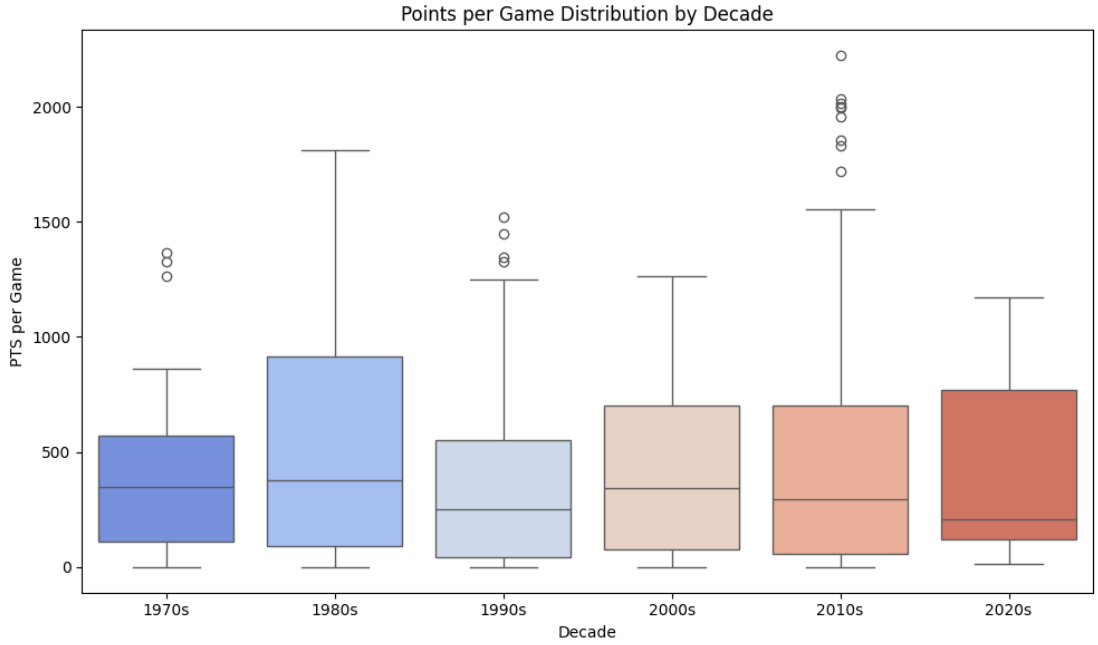


- A bar plot of top foreign scorers (e.g., Dirk Nowitzki, Vlade Divac) demonstrates their significant contributions in points, rebounds, and assists, underscoring their impact on team performance.

- **Player Distribution**:
  - The number of unique players fluctuates across decades, peaking in the 2000s/2010s, possibly due to successful data retrieval or larger player pools.



  - A line plot illustrates the number of unique players per decade, showing peaks in the 2000s and 2010s.



- A boxplot of points per game by decade indicates increasing variability in the 2010s/2020s, suggesting diverse scoring roles and the influence of high-scoring foreign players.

- **Limitations in Data**:
  - API timeouts on AWS/Colab platforms resulted in incomplete data, with errors for notable players like Pau Gasol and Victor Wembanyama, reducing the dataset's comprehensiveness.
  - Random sampling (150 players, capped at 50 per decade) may exclude high-performing or notable players, introducing potential bias.
  - The small overlap (32 foreign players) between the Kaggle and decade datasets limits robust domestic-foreign comparisons.
  - Kaggle data required fuzzy matching to handle name inconsistencies, and some players had missing or ambiguous nationalities, labeled as "Unknown" and excluded from analyses.

Group Project Roles

This project was a collaborative effort, with all group members contributing relatively equally to all aspects of the project. Each member participated in the design, coding, debugging, and testing of the Python script, as well as the planning of the data collection and analysis processes, including the integration and preprocessing of the Kaggle nationality data for foreign player comparisons. The team worked together to address challenges, such as handling API errors and performing fuzzy matching for the Kaggle dataset, to ensure the script's functionality.

Conclusion

This project successfully developed a robust pipeline for collecting and analyzing NBA player statistics from the 1970s to the 2020s, integrating data from the NBA's official API and nationality information from the Kaggle "NBA Players" dataset. The resulting dataset, saved as nba_players_by_decade.csv, provides a valuable resource for sports analytics, historical comparisons, and data-driven insights into basketball's evolution, particularly the rising influence of foreign players. Key findings highlight distinct performance trends across decades, with points peaking in the 1980s, rebounds surging in the 2000s and 2010s, and assists reaching their highest in the 1990s, reflecting shifts in gameplay and strategy. The increasing presence of foreign players, especially in the 2010s and 2020s, has significantly shaped the NBA, with standout performers like Dirk Nowitzki and Vlade Divac driving higher scoring, rebounding, and playmaking contributions, as evidenced by their elevated average stats in the decade dataset (536.20 PTS, 245.45 REB, 129.73 AST) compared to overall averages.

Despite challenges, such as API timeouts reducing data completeness and limited overlap (32 foreign players) between datasets due to random sampling, the pipeline’s error handling and Kaggle data integration ensured a functional analysis. Visualizations, including bar plots of foreign player representation and boxplots of points distribution, vividly illustrate the growing impact and variability of foreign talent. Future improvements could include local script execution to bypass API restrictions, enhanced fuzzy matching for nationality data, or a more comprehensive player sample to capture additional high-profile athletes. This project lays a strong foundation for deeper exploration of NBA performance trends, offering scalable methods to further investigate the transformative role of international players in shaping the modern game.