Daniel Kazarian and Arman Miri

Professor Chen

CSCI 185

2 June 2024

CSCI 185 Report

Introduction:

At first, our project in data mining set out to analyze UFC statistics by collecting thorough data from various websites through scraping. Regrettably, we faced major obstacles in this project because numerous websites containing UFC statistics had strong anti-scraping protections in place, leading to our repeated scraping efforts being thwarted. As a result, we shifted our focus to a different but just as thrilling field: NBA stats for the 2024 season.

Our team is highly interested in the NBA, especially with the buzz of the NBA Playoffs and the upcoming NBA Finals. This season has been particularly interesting as players persist in reshaping traditional positions, defying common predictions. For example, centers are displaying a wide range of skills, such as strong rebounding and impressive three-point shooting. The changing NBA environment offers an intriguing chance to examine and evaluate player performance data.

The goal of this project is to collect and analyze NBA data from the 2024 season to explore how current players are impacting the game of basketball and recognize any departures from conventional standards. We will use Python's Beautiful Soup and Requests libraries for web scraping in order to gather the required data. Our examination will concentrate on spotting trends, assessing player performances, and studying how modern players are transforming their roles in the game.

By participating in this project, we satisfy our academic needs and also enjoy our love for the NBA. The analysis findings may enhance comprehension of player performance trends, potentially impacting fan views and coaching approaches. Just as the NBA is evolving, the ways in which we analyze and enjoy the game must also evolve, and our project aims to contribute to this discussion.

Data Cleaning Process:

When we gathered the NBA statistics for the 2024 season, we came across a number of important data quality problems that needed careful attention to maintain the accuracy and reliability of our dataset. One major difficulty was the existence of repeated records for certain players. These copies were created due to players being exchanged between teams throughout the season. As a result, the database had specific statistics for each team they were on, along with an entry that combined all their season statistics labeled as "TOT" in the "Tm" column, which represents team.

In order to tackle this problem, we initially pinpointed the players with more than one entry by analyzing the "Tm" column. Players who were traded had a separate record for each team they were part of, along with an extra entry titled "TOT" for their combined stats. We employed the pandas library in Python to remove these duplicate rows. In particular, we kept the row that contained "TOT" in the "Tm" column, as it showed the player's overall performance for the full season. This made sure that our data set had one complete set of stats for each player, getting rid of any duplication and possible misunderstandings.

We also encountered difficulty with players having more than one position listed in the "Pos" column. This scenario happened as a result of certain players taking on varying positions

for varying teams. As an example, a player could have been a point guard (PG) for one team and a shooting guard (SG) for another team, which would show "PG-SG" in the "Pos" column of their total statistics row. To solve this issue, we had to figure out the position that each player played the most often throughout the season. We looked at the "G" column, representing the total games played by each team. By adding up the number of games played in each position for every team listed separately, we can find out which position a player was in most frequently. In this instance, if a player participated in 20 games as a point guard and 30 games as a shooting guard, we designated their main position as shooting guard. Once we figured out the position each player played the most, we changed the "Pos" column in the overall stats row to show this main position. This procedure guaranteed that every player was assigned a solitary, uniform position, streamlining our analysis and enhancing the clarity of the data.

We realized that the teams the players were on did not matter for our analysis goals, so we chose to remove the "Tm" column completely. Our examination concentrated on the metrics of individual player performance instead of comparing teams. By eliminating this column, we made our dataset more succinct and centered on the pertinent data points. During the process of cleaning the data, we conducted various consistency checks to maintain the reliability of our dataset. We made sure that each player had a distinct entry, double-checked the accuracy of the positions, and guaranteed that no important information was missing after the cleaning process. Moreover, we dealt with any missing or incomplete data by either filling in logical values or removing those entries if they were not essential to our analysis.

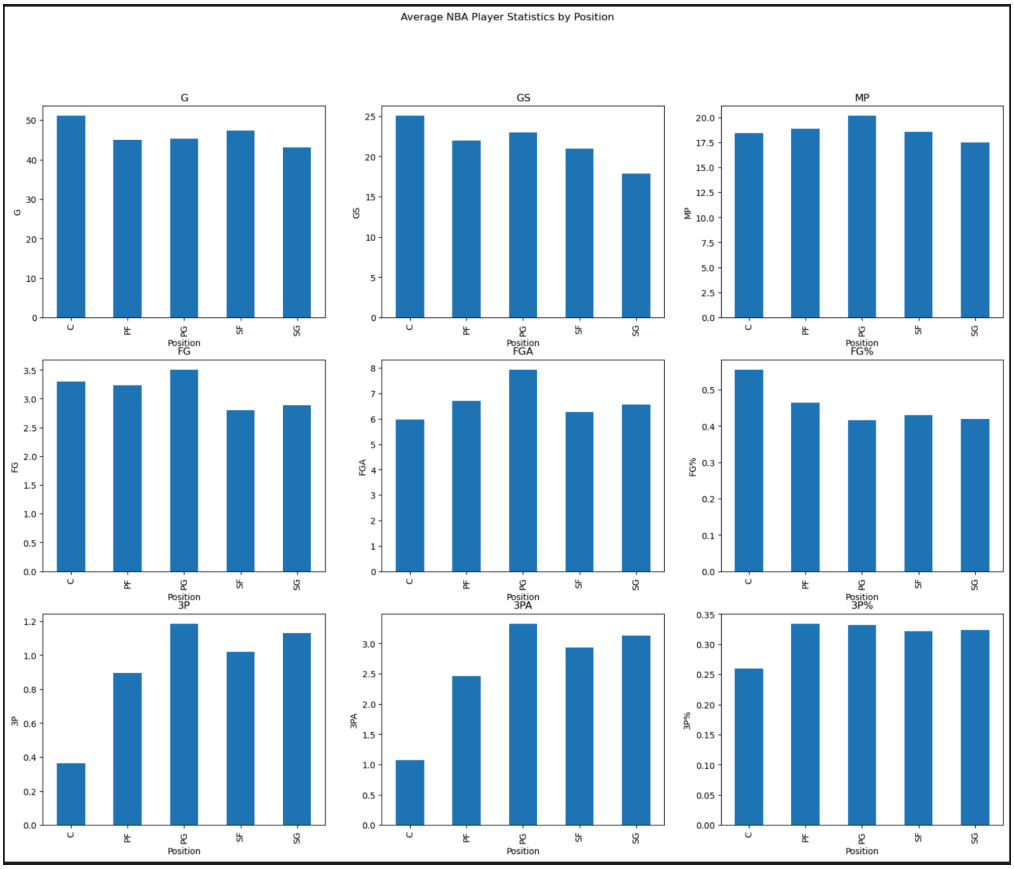
Through the implementation of these specific data cleaning procedures, we were able to turn a disorganized dataset full of duplicates into a refined set of player statistics. This careful methodology guaranteed that our future analysis would be founded on precise and dependable

data, enabling us to uncover significant findings on how modern NBA players are challenging traditional norms and reshaping their positions in the game.

Analysis 1:

After successfully cleaning and refining our NBA dataset, we moved on to the analysis and visualization phase to uncover insights into player performance by position. We began by loading the data from a CSV file named filtered.csv using the pandas library. For the visualization, we created a grid of subplots using the matplotlib library.

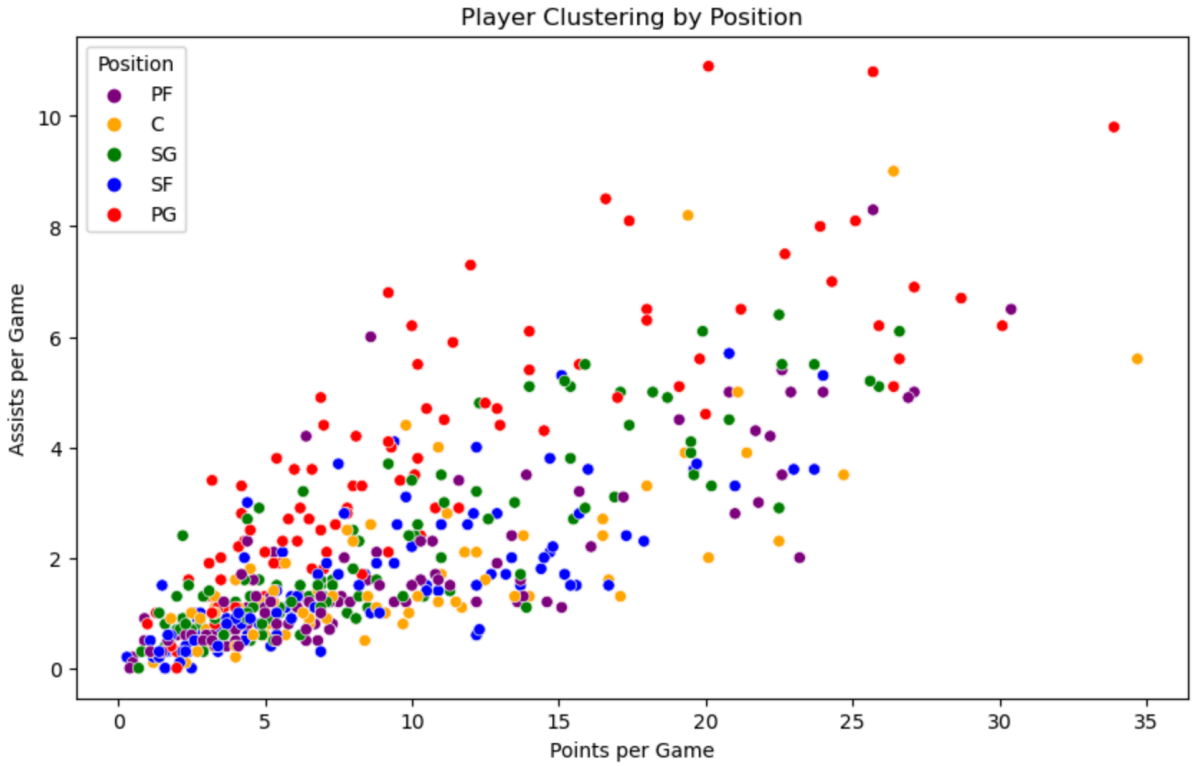
This detailed analysis and visualization process allowed us to gain valuable insights into how different positions perform across various statistics in the NBA. The bar charts provided a clear visual representation of these performance metrics, facilitating a deeper understanding of player roles and contributions on the court. By following this approach, we ensured that our analysis was based on accurate and reliable data, enabling us to derive meaningful insights into the evolving dynamics of NBA player performance.



Analysis 2:

For this analysis we standardized the data using the StandardScaler from the sklearn.preprocessing module, which transformed the numeric columns to have a mean of zero and a standard deviation of one. This step was key to ensure that all features contributed equally to the clustering process. We then used the K-means clustering algorithm from the sklearn.cluster module, setting the number of clusters to five and using a random state for reproducibility. The cleaned and standardized data was then used to fit the K-means model, and each player was assigned to one of the five clusters.

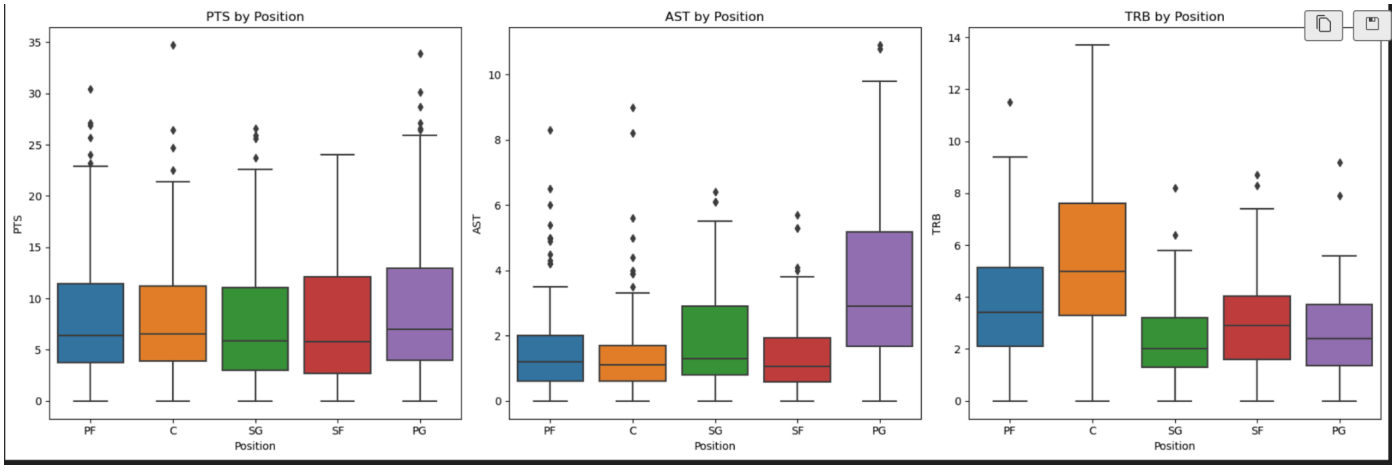
To visualize the clustering results, we created a scatter plot using the seaborn library. We plotted Points per Game (PTS) against Assists per Game (AST), coloring the points based on player positions using a predefined color palette. The scatter plot provided a clear visual representation of how players clustered according to their performance metrics and positions. This clustering analysis allowed us to uncover patterns in player performance and provided insights into how different positions relate to each other in terms of key statistics. The visualization facilitated a deeper understanding of the similarities and differences among players, contributing to our overall analysis of NBA player performance.



Analysis 3:

In this analysis phase, we focused on visualizing key performance statistics by player position using box plots. We selected three key statistics for visualization: Points per Game (PTS), Assists per Game (AST), and Total Rebounds per Game (TRB). Using the seaborn library, we created box plots for these statistics, grouped by player position. The box plots were arranged in a single row with three columns, each plot showing the distribution of one statistic across different positions.

This straightforward visualization approach allowed us to compare the distributions of key performance metrics across different positions, providing valuable insights into the variability and central tendencies of these statistics among NBA players. The box plots facilitated an easy comparison and highlighted the differences in performance metrics by position while also showing how much variation there can be in top performers.

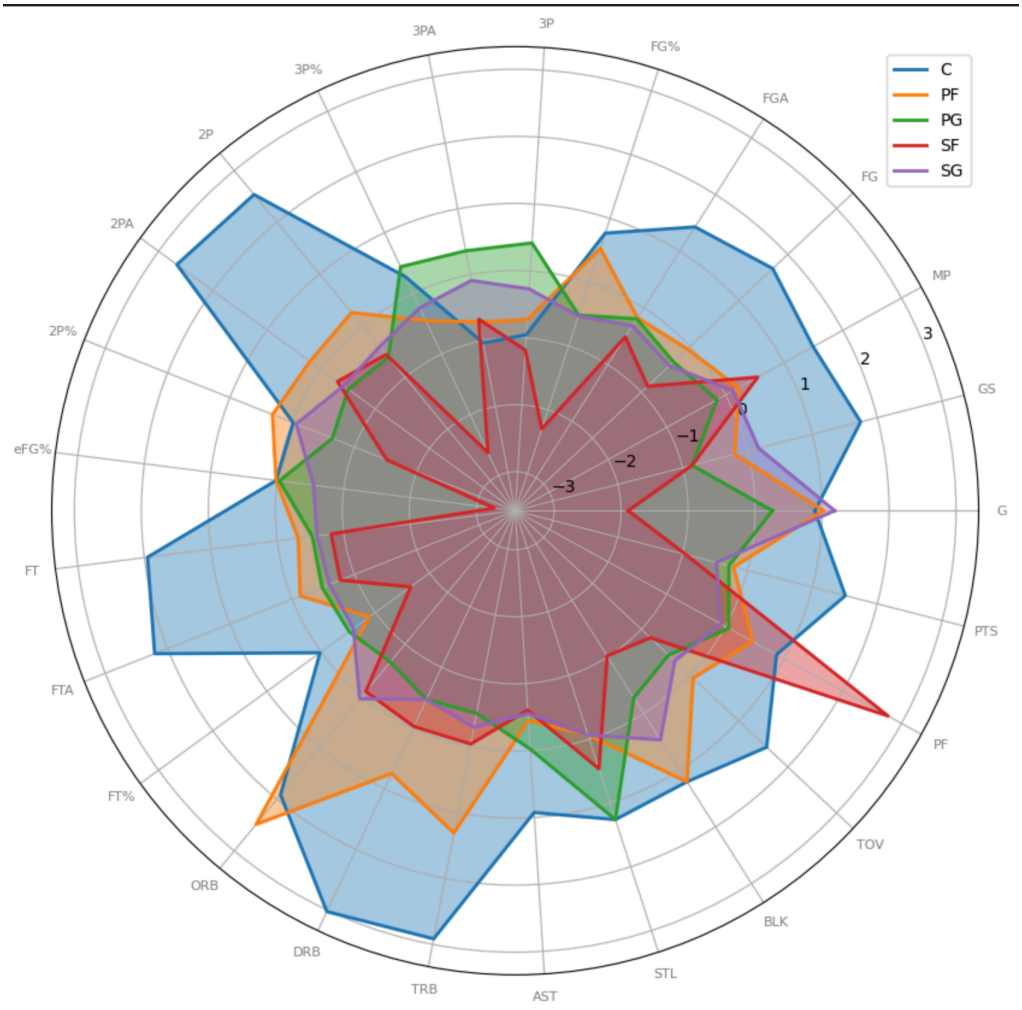


Analysis 4:

In this analysis phase, we visualized player performance across multiple statistics using radar charts. We then selected the top player from each position by grouping the data by position and choosing one player per group. We created radar charts for these top players using a custom

function. Each radar chart displayed the player's performance across various statistics, with categories represented around a circular plot. The charts were plotted in a single figure, allowing for a clear comparison of player performance by position.

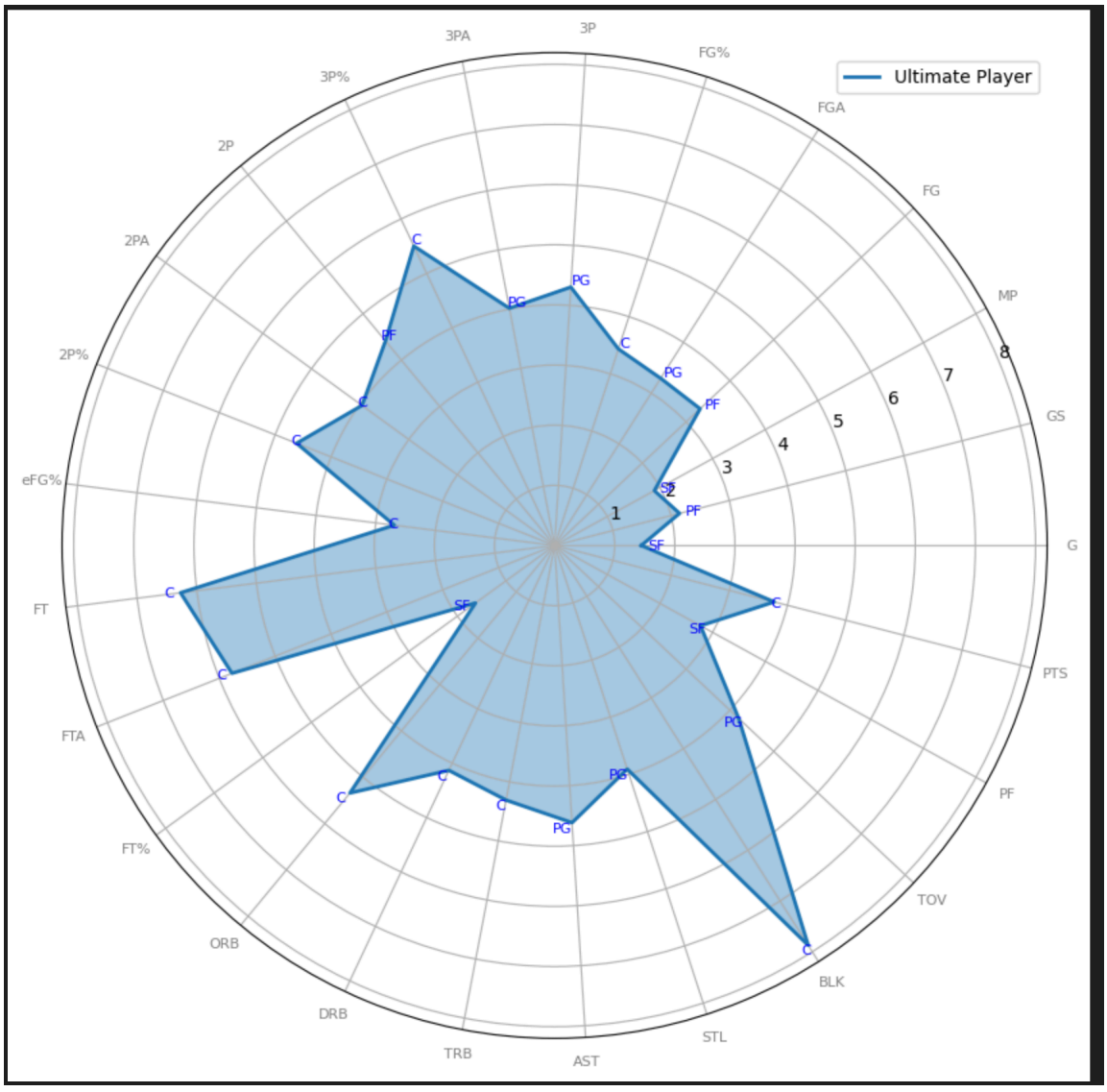
This simple visualization provided a comprehensive view of how top players from different positions performed across multiple metrics, highlighting their strengths and facilitating easy comparison and allowing an easy visual to see where positions are more commonly better than others.



Analysis 5:

In this phase of our analysis, we visualized the "ultimate player" by creating a radar chart. We determined the ultimate player statistics by identifying the maximum values for each statistic

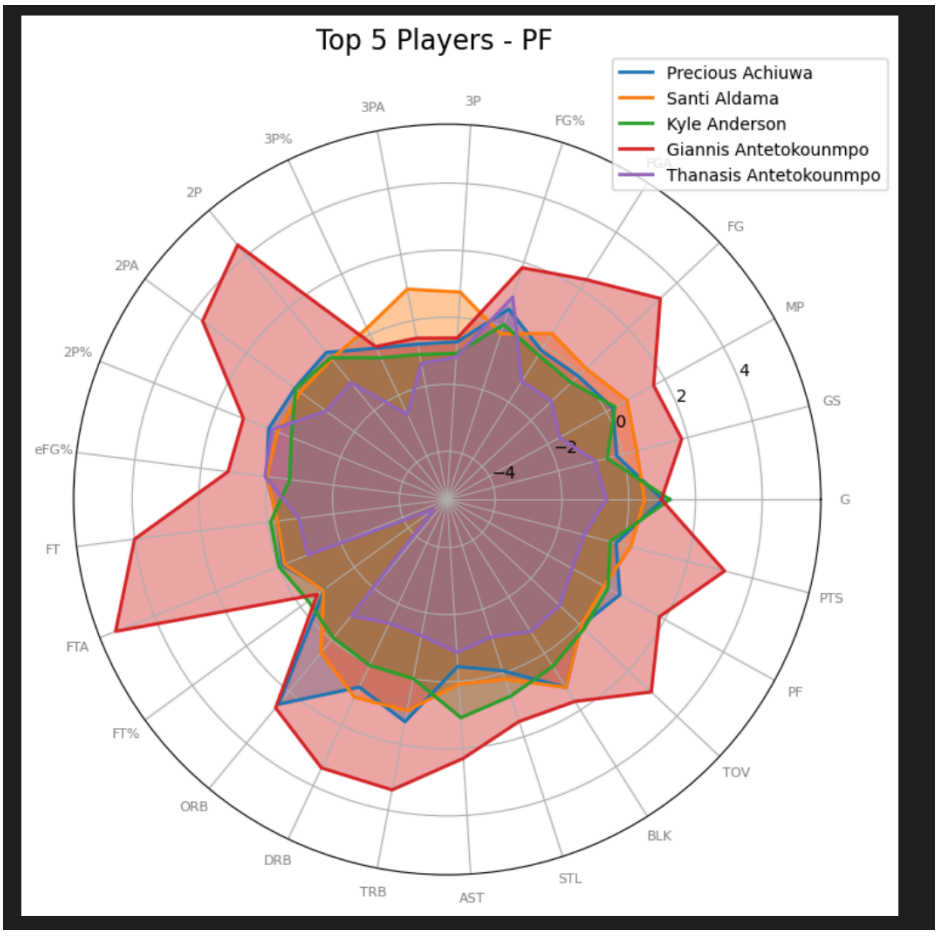
across all players. Corresponding positions for these maximum values were also identified. A radar chart was created using a custom function to display the ultimate player's performance across various statistics. This chart highlighted the maximum values for key performance metrics, with labels indicating the positions of players who achieved these maximums. This visualization provided a comprehensive view of the best possible performance metrics, showcasing the strengths of the ultimate player and facilitating easy comparison of top-performing positions.

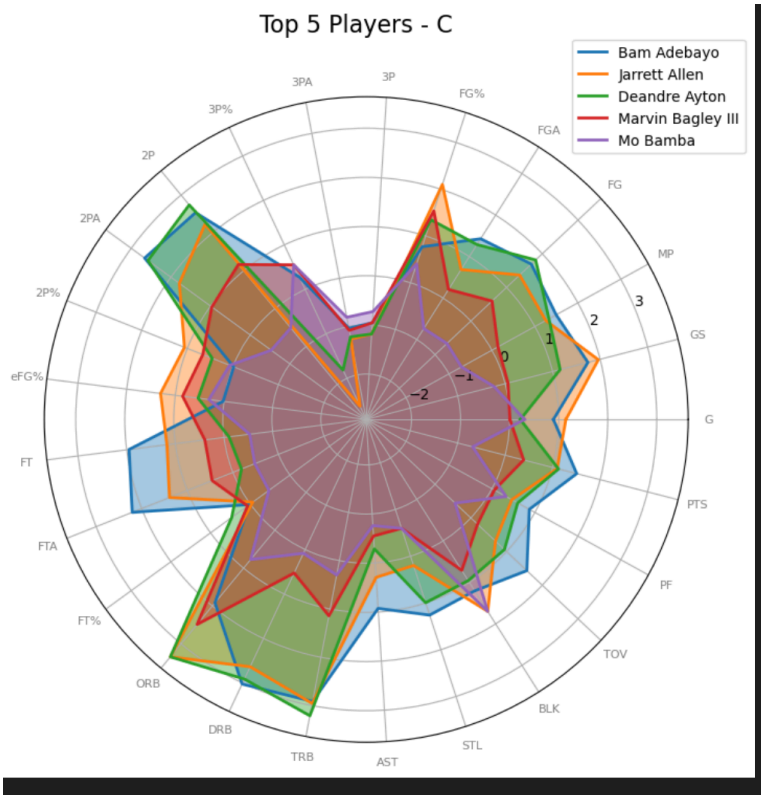


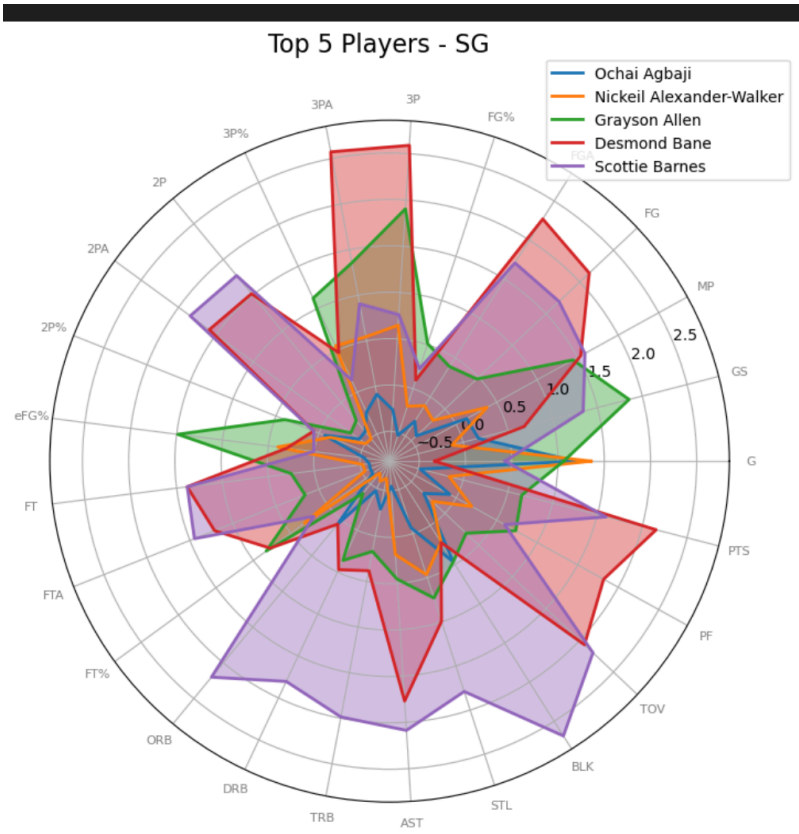
Analysis 6:

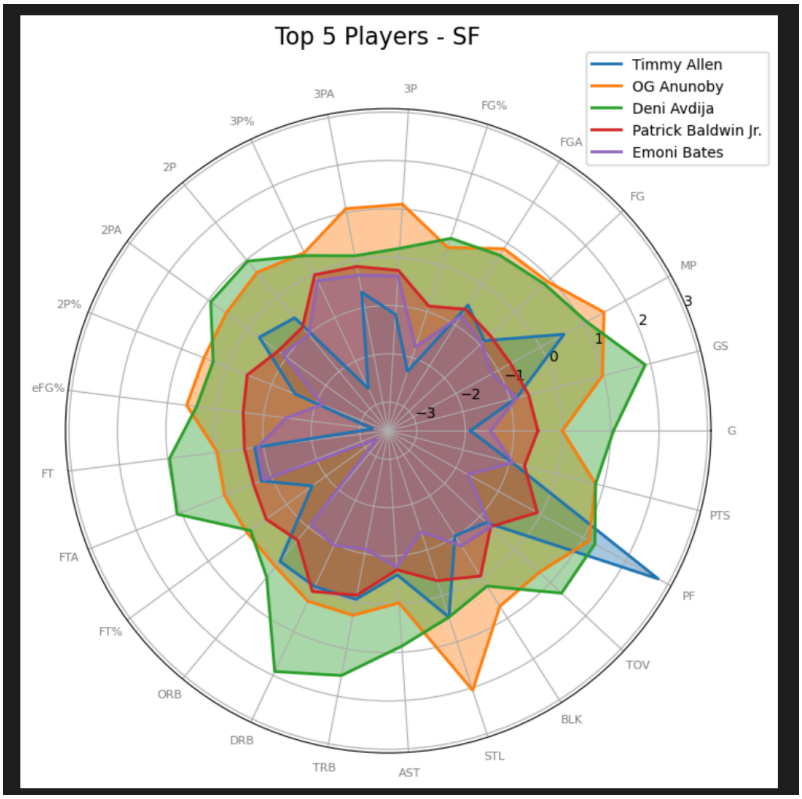
In this phase of our analysis, we visualized the performance of the top five players in each position using radar charts. We then identified unique positions in the dataset and proceeded to create a radar chart for the top five players in each position. For each position, we selected the top five players based on their performance metrics. A radar chart was plotted for each player, displaying their performance across various statistics. These radar charts were combined into a single plot for each position, with a legend to identify individual players.

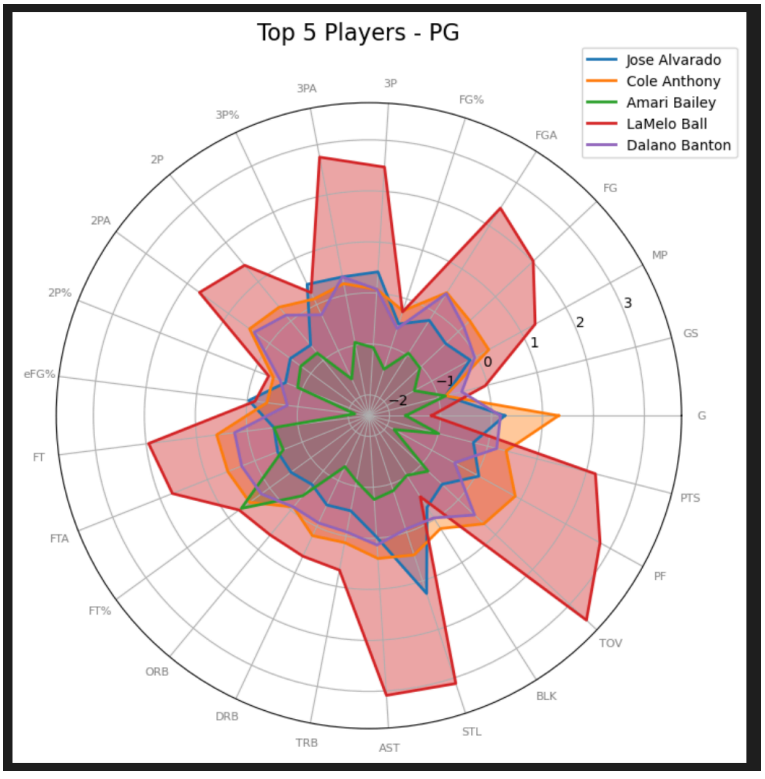
This visualization provided a comprehensive comparison of the top players in each position, highlighting their strengths and facilitating easy comparison of their performance metrics and allowing to see the variation in different players' roles in the same position (more offensive or defensive).




Top 5 Players - C
Bam Adebayo
Jarrett Allen
Deandre Ayton
Marvin Bagley III
Mo Bamba
3PA
3P
FG%
FGA
FG
MP
GS
G
PTS
PF
TOV
BLK
STL
AST
TRB
DRB
ORB
FT%
FTA
FT
eFG%
2P%
2PA
2P
3P%

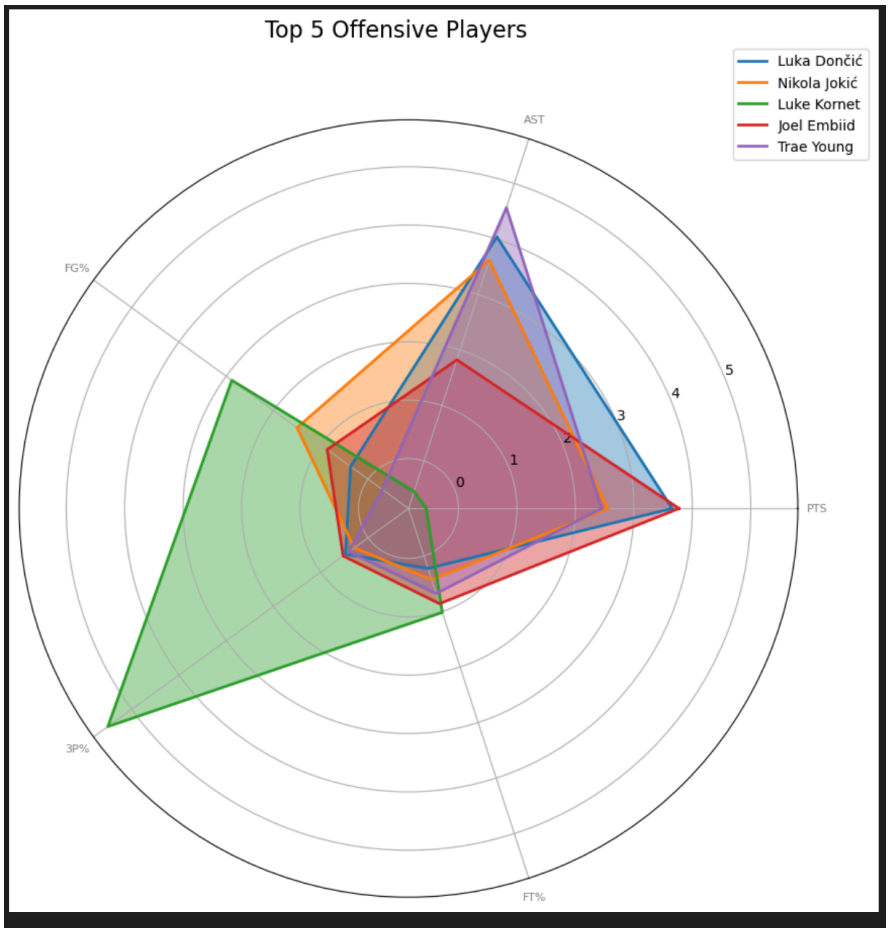

Top 5 Players - SG
Ochai Agbaji
Nickeil Alexander-Walker
Grayson Allen
Desmond Bane
Scottie Barnes
3PA
3P
FG%
FGA
FG
MP
GS
G
PTS
PF
TOV
BLK
STL
AST
TRB
DRB
ORB
FT%
FTA
FT
eFG%
2P%
2PA
2P
3P%


Top 5 Players - SF
Timmy Allen
OG Anunoby
Deni Avdija
Patrick Baldwin Jr.
Emoni Bates
3PA
3P
FG%
FGA
FG
MP
GS
G
PTS
PF
TOV
BLK
STL
AST
TRB
DRB
ORB
FT%
FTA
FT
eFG%
2P%
2PA
2P
3P%


Top 5 Players - PG
Jose Alvarado
Cole Anthony
Amari Bailey
LaMelo Ball
Dalano Banton
3PA
3P
FG%
FGA
FG
MP
GS
G
PTS
PF
TOV
BLK
STL
AST
TRB
DRB
ORB
FT%
FTA
FT
eFG%
2P%
2PA
2P
3P%

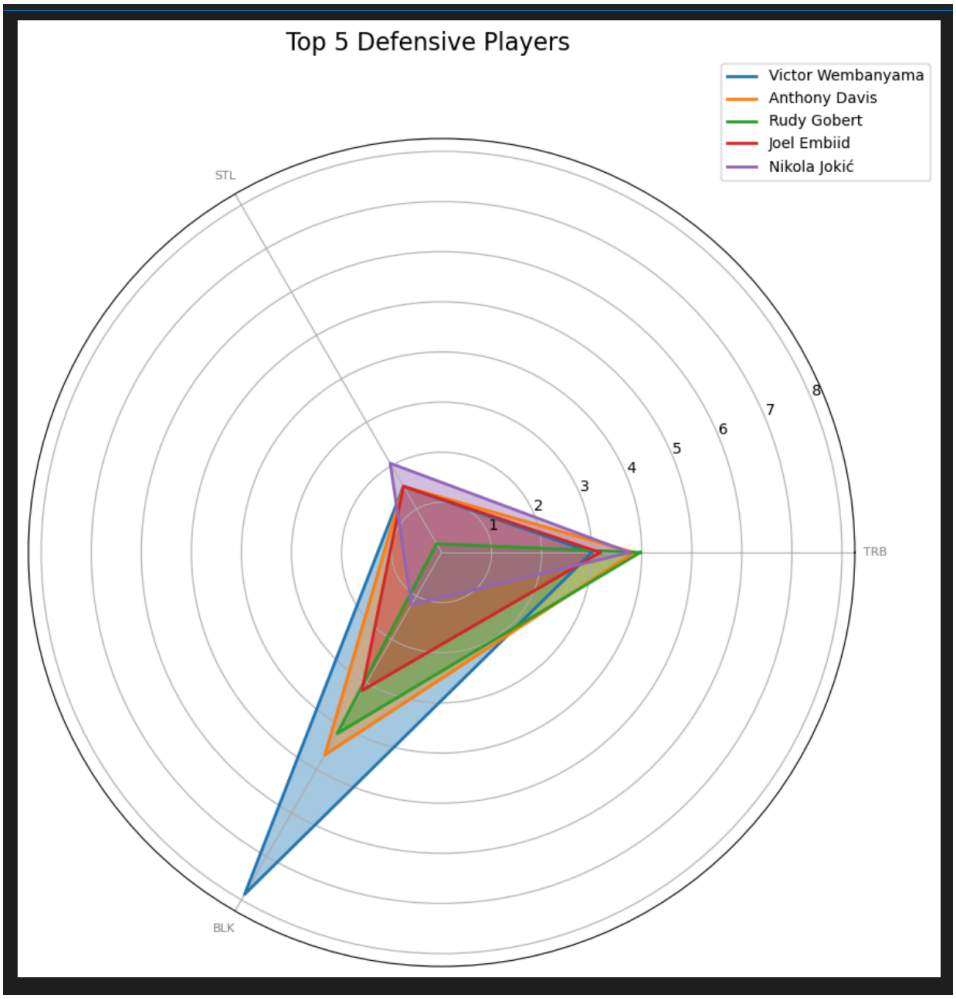
Analysis 7:

In this analysis, we visualized the top offensive players using radar charts. We defined an offensive score by summing up key offensive metrics: Points per Game (PTS), Assists per Game (AST), Field Goal Percentage (FG%), Three-Point Percentage (3P%), and Free Throw Percentage (FT%). We selected the top five players with the highest offensive scores. For visualization, we created a radar chart for each of these top players, displaying their performance across the offensive metrics. Each radar chart was plotted on a single figure, with a legend to identify individual players. This visualization highlighted the strengths of the top offensive players, allowing for an easy comparison of their performance across key metrics, also allowing an analysis that went beyond position but also into the performance of players.



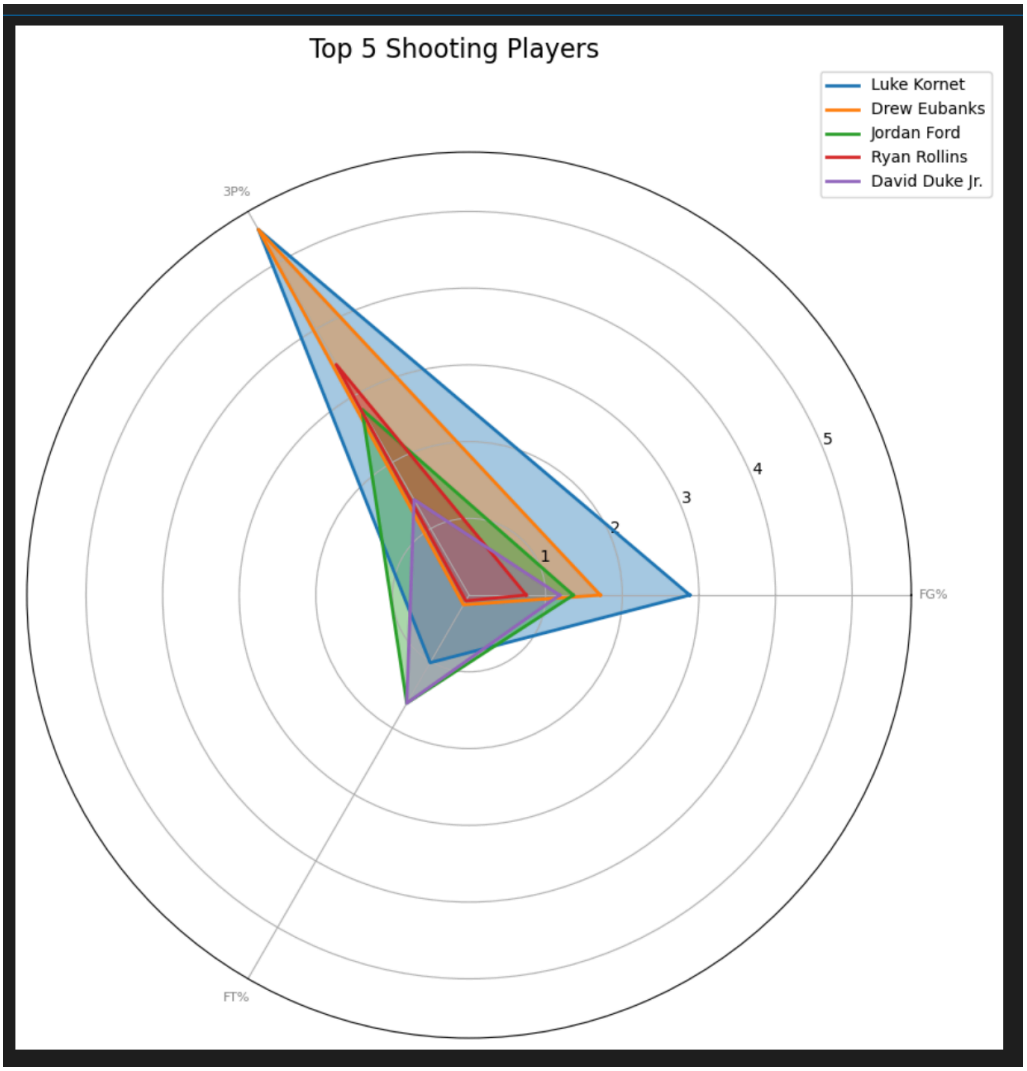
Analysis 8:

In this analysis, we visualized the top defensive players using radar charts. We defined a defensive score by summing up key defensive metrics: Total Rebounds per Game (TRB), Steals per Game (STL), and Blocks per Game (BLK). We followed the same process as the top offensive players and used the radar charts. This visualization highlighted the strengths of the top defensive players, allowing for an easy comparison of their performance across key defensive metrics, also allowing an analysis that went beyond position but also into the performance of players.



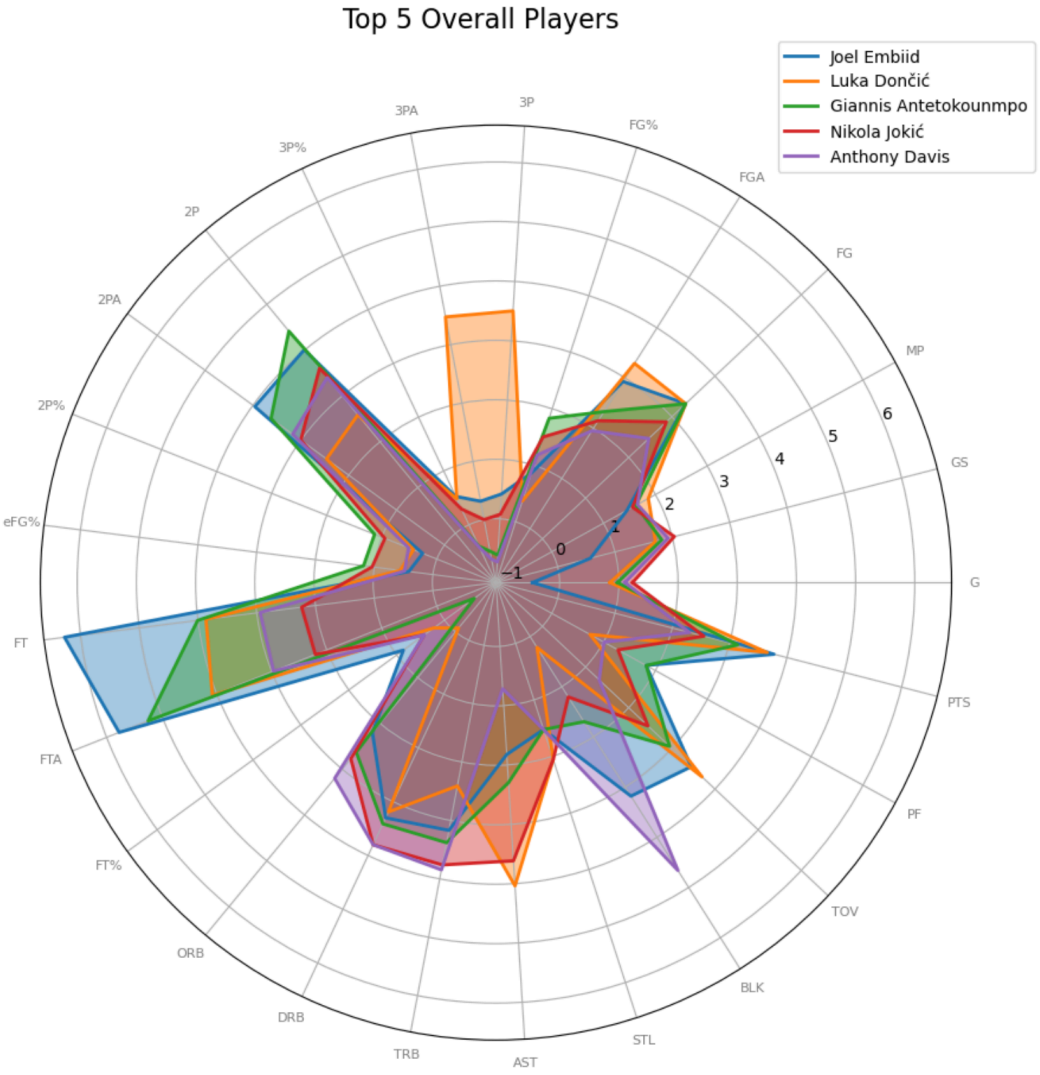
Analysis 9:

In this analysis, we visualized the top shooting players using radar charts. We defined a shooting score by summing up key shooting metrics: Field Goal Percentage (FG%), Three-Point Percentage (3P%), and Free Throw Percentage (FT%). We followed the same process as the top offensive/defensive players and used the radar charts. This visualization highlighted the strengths of the top shooting players, allowing for an easy comparison of their performance across key shooting metrics, also allowing an analysis that went beyond position but also into the performance of players.



Analysis 10:

In this analysis, we visualized the top overall players using radar charts. We defined an overall score by summing up all normalized statistics for each player. This overall score provided a comprehensive measure of a player's performance across all metrics. We selected the top five players with the highest overall scores. For visualization, we created a radar chart for each of these top players, displaying their performance across all metrics. Each radar chart was plotted on a single figure, with a legend to identify individual players. This visualization highlighted the strengths of the top overall players, allowing for an easy comparison of their performance across a wide range of metrics. The radar charts provided a comprehensive view of each player's abilities, showcasing their all-around excellence in various aspects of the game.



Conclusion:

In conclusion, our project successfully pivoted from UFC to NBA statistics, yielding insightful analyses of player performance and positional dynamics in the 2024 season. Despite initial challenges with data collection, we meticulously cleaned and transformed our dataset, enabling robust analyses that highlighted the evolving skill sets across player positions. Utilizing techniques such as clustering and radar charts, we uncovered patterns in player performance, showcasing the diversification of roles, especially among traditionally specialized positions like centers. Our visualizations of top offensive, defensive, and overall performers provided a nuanced understanding of how modern NBA players are redefining their contributions on the court. This project furthered our appreciation of basketball analytics, offering valuable insights that could influence fan perspectives and coaching strategies.