Data 670 Data Analytics

Alexander Carcamo

Professor Jon McKeeby

Assignment 3: NBA Field Goal Analysis

September 23, 2024

Executive Summary

The NBA Field Goal Analysis project focuses on evaluating the Atlanta Hawks' shot distribution and player performance in comparison to league benchmarks across two distinct periods: 2004-2006 and 2021-2023. The goal is to identify key areas of strength and weakness within the team's offensive and defensive strategies, providing actionable insights to improve their overall performance. Key Performance Indicators (KPIs) such as Shot Distribution Shifts (percentage of shots taken from each court zone), Player Performance Metrics (comparison of Hawks' shooting percentages to league averages), and Team Potential Success (evaluating the impact of potential player acquisitions) will drive the analysis. Additionally, a fourth KPI, Defensive Shot Impact, will assess how effectively the Hawks defend key zones on the court, offering a balanced view of both offensive and defensive performance.

Data from NBA shot charts across the two periods will be used to measure how the Hawks' shot distribution and performance align with league-wide trends. By comparing these metrics, we will identify shifts in shot selection, areas where the Hawks underperform, and potential strategies for improvement. The project will also include recommendations for roster adjustments, focusing on acquiring players who can enhance team success by improving underperforming zones. The results in this analysis can also be used to determine what the best offensive lineup may be based on the players on the team.

This data-driven approach will provide the Hawks' management and coaching staff with valuable insights to make informed decisions on game strategies, player development, and acquisitions, ultimately positioning the team for greater success in the evolving NBA landscape.

Table of Contents

Project Scope

Data Analytics Project/Project Scope

This project focuses on analyzing the evolution of shot distribution in the NBA over two distinct time periods—2004-2006 and 2021-2023—with a specific emphasis on the Atlanta Hawks. By comparing player performance against league-wide benchmarks, we aim to identify potential areas where the Hawks can improve, specifically in terms of shot selection and efficiency. Additionally, the analysis will consider the evolving nature of the NBA, particularly the shift towards a more three-point-dominant style of play. This project will involve analyzing data from thousands of games, examining key performance indicators such as shot distribution, player efficiency in different court zones, and team potential success. The ultimate goal is to provide the Hawks with actionable insights that can be used to enhance team performance, particularly by identifying players that can strengthen weak areas of the team. This data-driven approach will offer a comprehensive view of how the Hawks stack up against the rest of the league and suggest strategies for improvement.

The problem being analyzed centers around the Hawks' ability to adapt to the modern NBA style of play, where perimeter shooting has become increasingly important. By examining shot distribution patterns, we can identify whether the Hawks are optimizing their shot selection relative to league-wide trends. We also aim to pinpoint specific areas of weakness, such as a lack of scoring near the rim or inconsistency in three-point shooting, and recommend potential player acquisitions to address these gaps. By comparing the Hawks' current performance with historical benchmarks, this analysis will provide insights into how their strategy needs to evolve in order to remain competitive. The inclusion of advanced metrics, such as shot efficiency and zone-specific performance, will offer a nuanced understanding of where the Hawks excel and where they need to improve. This will enable the team's management to make informed decisions about roster changes and tactical adjustments. The data analytics problem that I am analyzing is how the Atlanta Hawks can improve their overall team performance by optimizing shot distribution and addressing key areas of weakness.

Problem Importance

This project was selected because of the growing significance of data analytics in professional sports, particularly in the NBA. Teams like the Atlanta Hawks operate in a highly competitive environment where small adjustments in strategy can have a large impact on overall performance. In recent years, the NBA has shifted towards a play style that heavily emphasizes three-point shooting and advanced metrics, making data-driven decision-making critical. The ability to understand and act on insights from player performance and shot distribution data is no longer a luxury—it's a necessity for success. By analyzing key performance indicators such as shooting percentages, shot distribution, and player efficiency, this project aims to provide the Hawks with the knowledge they need to stay competitive in the rapidly evolving NBA landscape. The findings from this analysis will not only help the Hawks improve their performance but also offer insights into how they can position themselves for long-term success.

The importance of this project lies in its potential to influence decision-making at both the strategic and operational levels for the Atlanta Hawks. At the strategic level, understanding how the team's shot distribution compares to league averages can inform coaching decisions on offensive and defensive strategies. It will also help management make informed choices regarding player acquisitions, focusing on players who excel in areas where the team is currently underperforming. On the operational level, the data can highlight specific aspects of gameplay, such as which court zones yield the most efficient scoring opportunities, allowing for targeted practice and player development. The ability to back these decisions with data not only improves performance but also reduces risk, as decisions are based on evidence rather than instinct or tradition. Ultimately, this data-driven approach will lead to more efficient resource allocation and improved team performance.

The primary stakeholders in this project include the Atlanta Hawks' management team, coaching staff, and players. Management will benefit from insights on player acquisitions, helping them to build a stronger, more well-rounded roster. The coaching staff will gain valuable information on shot selection and player performance in different court zones, which can be used to refine game strategies and practice routines. Players will benefit by having clear data on where they can improve their individual performances to align better with team goals. Additionally, fans and investors of the Atlanta Hawks will benefit indirectly, as a more competitive and successful team leads to higher engagement and profitability. By aligning the team's tactics with data-driven insights, this project will help secure the Hawks' position as a competitive force in the NBA.

Research/Problem Analysis

Previous projects using NBA shot chart data have focused on analyzing player performance in specific games, often highlighting individual shot charts to understand a player's shooting efficiency in real-time. For example, data visualizations have been created to show a player's shooting percentages from different court zones during a particular game, offering insights into which areas of the court they excel at shooting from. Additionally, team-level shot charts have been developed to analyze the collective shot distribution of a team during specific games, allowing for comparisons of overall team efficiency. These visualizations often include color-coded heatmaps that identify high- and low-efficiency zones for both teams and players. Furthermore, many analyses have sought to highlight the most efficient player in each court zone, showcasing the areas where individual players excel or struggle. While these insights provide valuable information for immediate game analysis, they typically focus on short-term performance rather than long-term trends or comparisons over multiple seasons.

The gap that our project aims to fill lies in the broader, long-term analysis of shot distribution and player performance over multiple seasons, specifically from 2004-2006 and 2021-2023. Unlike previous studies that focus on specific games or short time frames, this project will provide a longitudinal view of how shot selection has evolved in the NBA, with a focus on the Atlanta Hawks. By comparing two distinct eras, we will be able to identify shifts in shot distribution patterns, highlight changes in team strategies, and pinpoint areas for improvement in player acquisitions. This deeper analysis will offer insights into the Hawks' performance over time, providing the team with actionable data for enhancing player development and adjusting their approach to modern gameplay trends. Additionally, while previous projects may have focused on individual or team performance in isolation, our project will incorporate league-wide benchmarks, providing a more comprehensive comparison of the Hawks' performance relative to the rest of the NBA.

Critical Success Factors (CSFs)

The success of this project relies on three key Critical Success Factors (CSFs): data quality and preparation, model selection and implementation, and clear and insightful data visualization. First, data quality is essential for ensuring accurate analysis. This involves cleaning and preparing the NBA shot chart data from 2004-2006 and 2021-2023, addressing any inconsistencies or missing information that could skew results. Tools such as Python and RStudio will be used to clean and organize the datasets, ensuring they are ready for analysis. If the data is not properly cleaned, any models or conclusions drawn from it will be unreliable, leading to ineffective recommendations. Ensuring high data quality will allow us to perform accurate analyses that reflect real-world performance and trends.

Another crucial success factor is the selection of appropriate models and their implementation. Selecting the right models will help to ensure that our analysis is both accurate and relevant to the problem being addressed. Predictive models will be built to assess shot distribution patterns, player efficiency, and team performance in specific court zones. These models must be fine-tuned for accuracy and must be interpretable by stakeholders to enable clear decision-making. Additionally, effective visualization of the results will be essential for communicating insights to non-technical stakeholders. Tools like Tableau and RStudio will be utilized to create compelling visualizations, such as shot charts and heatmaps, that clearly demonstrate areas of improvement for the Atlanta Hawks. In the end, the project's success will depend on a combination of high-quality data, robust models, and effective communication through visuals.

Key Performance Indicators

In this NBA field goal analysis, the Key Performance Indicators (KPIs) will be essential in evaluating the effectiveness of the Atlanta Hawks' shot distribution, player performance, and overall team potential. These KPIs will help determine areas of strength and weakness in comparison to league averages, as well as provide insights into possible improvements through player acquisitions. KPIs are crucial for tracking progress and evaluating the success of a project, especially when focusing on actionable insights. As Parmenter (2015) highlights, KPIs provide measurable values that help organizations understand their performance and make strategic decisions based on data-driven insights. In this project, the KPIs will focus on shot distribution, player performance, team potential success, and a fourth KPI that remains to be determined. These metrics will be pivotal in guiding decisions related to the Hawks' overall strategy and potential roster adjustments.

The first KPI is Shot Distribution Shifts, which measures the percentage of shots taken from each zone on the court. This metric is crucial in understanding how the Hawks’ shot selection compares to both historical trends and league-wide benchmarks. By analyzing the distribution of shots across zones such as the restricted area, mid-range, and three-point line, we can determine whether the Hawks are optimizing their shot selection in line with modern NBA strategies. For instance, if the Hawks are taking fewer three-pointers compared to the league average, it could signal a need to adjust their offensive approach to increase scoring efficiency. This KPI will be measured by comparing the percentage of shots from each zone between the 2004-2006 and 2021-2023 seasons and against current league averages. Identifying significant deviations in shot distribution will provide valuable insights into areas where the Hawks can adjust to maximize scoring potential.

The second KPI is Player Performance Metrics, which evaluates how the Hawks’ shooting percentages in key zones stack up against league averages. This metric highlights the strengths and weaknesses of individual players in various areas, such as three-point shooting or finishing in the paint. By comparing these percentages, we can pinpoint zones where the team is excelling and areas that require improvement. For example, if the Hawks' shooting accuracy from mid-range is lower than the league average, it could indicate the need for skill development or changes in player roles. This KPI will also aid in identifying underperforming players and prioritizing areas for coaching interventions. Tracking this data over multiple seasons will help monitor trends and measure the effectiveness of coaching strategies and player development efforts.

The third KPI is Team Potential Success, which projects how well the Hawks could perform by making strategic player acquisitions or tactical adjustments. This KPI focuses on simulating the impact of bringing in players who excel in zones where the Hawks currently struggle. By analyzing the performance of potential acquisitions in key areas, such as the restricted area or three-point line, we can estimate the improvement these players could bring to the team. For example, if a player has an exceptional efficiency in a zone where the Hawks are weak, adding them could significantly elevate the team's scoring and defensive impact. This KPI will be evaluated by comparing hypothetical roster changes to the current roster, analyzing how these moves could enhance the Hawks' overall efficiency and potential for success. This analysis offers valuable guidance for management when considering trades or free-agent signings.

The fourth KPI is Efficiency Improvement Potential, which quantifies the projected increase in team efficiency with new player additions. This metric assesses how the Hawks' overall offensive performance could change if they acquire high-efficiency players in underperforming zones. By using predictive modeling, we can estimate the impact of these player acquisitions on the Hawks' scoring efficiency. This analysis will focus on calculating the potential improvement in efficiency metrics based on the players' past performances. It will also highlight the most impactful acquisitions and their expected contribution to the Hawks' offensive capabilities. Ultimately, this KPI provides a clear picture of how strategic changes could elevate the team's performance and helps prioritize moves that offer the most significant return on investment.

Project Insights of your Data Analysis

For this project, we expect to identify clear trends in the evolution of shot distribution in the NBA, specifically focusing on how the Atlanta Hawks compare to league benchmarks. We anticipate that our analysis will show a significant shift in shot selection patterns between the two time periods (2004-2006 and 2021-2023), with a marked increase in three-point attempts, particularly from zones like "Above the Break 3." Additionally, we expect to pinpoint areas of underperformance for the Hawks, such as inefficiencies in scoring near the rim or in contested three-point shots. By comparing player performances to league-wide averages, we aim to highlight specific players or zones where the Hawks need improvement. These findings will allow us to recommend potential player acquisitions to strengthen the team's weaknesses. Overall, the conclusions drawn from this analysis should offer actionable insights that help the Hawks improve both their offensive strategy and overall team performance.

Furthermore, we expect that our analysis will underscore the importance of data-driven decision-making in professional sports. The findings will not only validate the increasing reliance on three-point shooting in modern NBA play but also emphasize the need for teams like the Hawks to adapt their strategies to remain competitive. We anticipate that our recommendations for roster changes or tactical adjustments will align with the team's current goals, providing them with a pathway to improve scoring efficiency and player development. In conclusion, the project will provide the Hawks with a comprehensive understanding of how their performance compares to league trends and offer solutions for addressing key weaknesses. This data-driven approach will ultimately enhance the Hawks' ability to make informed decisions, contributing to their long-term success.

Project Milestones

Data Collection and Preparation: The first major milestone involves gathering the NBA shot chart data from the 2004-2006 and 2021-2023 seasons. This step will also include cleaning the data and ensuring its readiness for analysis by removing inconsistencies and handling any missing values. This milestone is critical for establishing a solid foundation for analysis and must be completed before any modeling can take place.

Shot Distribution Analysis: After preparing the data, the next milestone will involve analyzing and comparing shot distribution patterns between the two time periods. This will require visualizations like shot charts and heatmaps to clearly demonstrate changes in play style and identify zones where the Hawks have underperformed compared to league averages.

Player and Team Benchmarking: In this milestone, we will develop league-wide benchmarks for shot efficiency and compare them to the performance of individual Hawks players. This comparison will allow us to identify specific areas where the team falls short and suggest tactical improvements or player acquisitions.

Final Report and Presentation: The final milestone will be compiling all findings into a comprehensive report with visualizations. This report will be presented to stakeholders, along with actionable recommendations for improving team performance. Achieving this milestone ensures that the analysis is not only completed but also effectively communicated to decision-makers.

Completion History

Week 1	Research different datasets and brainstorm ideas for final project. Complete required readings and begin outlining project scope. Look for scholarly articles that have done similar analysis in the past.
Week 2	Continue the required readings and scholarly articles. Finalize the project scope.
Week 3	Compile datasets and perform a high level review of the variables. Identify how variables will be used for analysis and begin the dataset selection and exploration assignment.
Week 4	Complete the dataset selection and exploration assignment. Begin data cleaning process and outline the analysis models that will be created.
Week 5	Continue the data cleaning process and continue working with the analytical models that needs to be created
Week 6	Continue work on the analytical models and outline the data visualizations that need to be created.
Week 7	Continue work on the analytical models and continue work on the data visualizations.
Week 8	Finalize work on analytical models and data visualizations.
Week 9	Work on results and finding.
Week 10	Work on results and findings.
Week 11	Finalize complete report.

Lessons Learned

Week 1	Correctly using the analytical approach in order to define the project scope.
Week 2	Being able to identify critical success factors and key performances indicators for a sport analytical analysis.
Week 3	It's important to distinguish between correlation and causation when analyzing data, as variables may seem related without one directly causing the other. It can definitely be an issue with this dataset because of the different variables that describe the zones shots were taken from.
Week 4	In order to be a career leader in the NBA you need to play 400 games played or 10000 points. In order to filter out players that do not play much and as such do not have a large impact we will reduce these numbers. Our minimum will be 100 games or 2500 points.
Week 5	The unpredictability of sports outcomes and handling large numbers of variables make machine learning in sports a very difficult task. Creating feature subsets and preprocessing data by differentiating between match-related features and external features are one way to combat these difficulties.
Week 6	Seaborn is designed for easy creation of statistical plots like box plots, violin plots, and heatmaps, which are useful for visualizing shooting efficiency across different zones. It can also overlay regression lines and confidence intervals, which would help in analyzing trends in player performance over time.
Week 7	By training a model on historical shot data, you can analyze which combinations of factors lead to a higher likelihood of successful shots. This prediction can be useful in identifying high-probability shooting zones and optimizing shot selection strategies. This approach is particularly helpful when analyzing complex interactions between different variables like game context and player positioning, making the dataset more manageable while retaining critical information for accurate predictions.
Week 8	The ensemble models and how they can help with machine learning analysis. It helps prevent overfitting and gives more accurate results.
Week 9	Carefully selecting and engineering features, such as breaking down shot efficiency by zone and accounting for both player and team metrics, significantly impacted the models' performance.

Week 10	Using multiple evaluation metrics, like Mean Squared Error (MSE) and cross-validation, was essential to assess model performance.
Week 11	Simulating the impact of potential player acquisitions based on model predictions was a valuable exercise. This approach demonstrated how data analysis could inform strategic decisions, like roster changes, by providing quantifiable estimates of potential improvements.
Week 12	Throughout this project, I gained a deeper appreciation for the nuances and best practices in data analysis. I learned the importance of data cleaning and pre-processing, as these steps are fundamental for ensuring that the results are accurate and reliable. Selecting the right modeling techniques was equally critical, as it required a strategic understanding of each method's strengths and potential biases to achieve meaningful insights. Evaluating model performance comprehensively through metrics like cross-validation and understanding overfitting highlighted the need for rigor and precision in analysis. Additionally, I saw firsthand the power of data visualization to make complex findings more digestible and impactful for stakeholders. The iterative nature of the work reinforced that data analysis is an evolving process that requires continual refinement. These elements collectively underscore the importance of a meticulous, flexible, and well-communicated approach to any analytical endeavor, ensuring the results are actionable and trustworthy.

Data Set Description

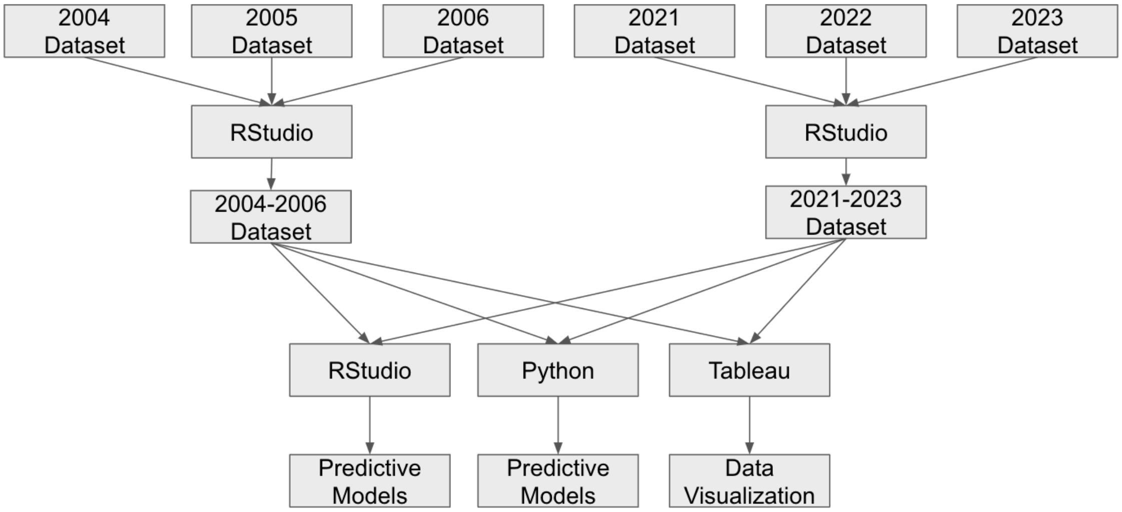
For this project, we will use two primary data sets that span multiple NBA seasons to analyze shot distribution and player performance: the 2004-2006 NBA Shot Chart Data Set and the 2021-2023 NBA Shot Chart Data Set. Both data sets contain detailed records of every shot taken during regular season games, covering a wide range of variables such as player names, shot locations, shot types, and outcomes (made or missed). The first data set, covering the 2004-2006 seasons, contains approximately 581,743 rows and 26 columns, representing a robust sample of shot data during that era. The second data set, covering the 2021-2023 seasons, contains 624,925 rows and 26 columns, reflecting more recent trends in shot selection and playing style. These data sets are of high quality, as they were sourced from a publicly available GitHub repository managed by an NBA analytics expert, ensuring the data is clean, well-structured, and comprehensive. Both data sets are crucial to understanding how shot selection has evolved in the NBA, with a specific focus on the Atlanta Hawks.

The 2004-2006 NBA Shot Chart Data Set serves as the baseline for this analysis, representing an earlier era in NBA gameplay where the three-point shot was less prevalent than it is today. This data set allows us to evaluate the performance of players and teams in this era, particularly their shot distribution across various zones on the court. It includes variables such as player name, team name, basic shot zone (e.g., restricted area, mid-range, corner three), and shot outcome. By comparing this data set to the more recent one, we can assess how the Hawks' shooting strategy and performance have evolved, particularly in the context of league-wide trends. This historical perspective is important because it provides insight into how modern teams like the Hawks can adapt to the ever-changing style of play in the NBA. The data was originally collected through NBA tracking systems, which capture every shot taken during games, ensuring its accuracy and completeness.

The 2021-2023 NBA Shot Chart Data Set provides the modern context needed for this comparative analysis. This data set includes variables similar to the 2004-2006 data, such as shot type, shot location, and shot outcome, but reflects the current era of NBA basketball, which is heavily influenced by three-point shooting and advanced metrics. This data set is particularly important for understanding how the Atlanta Hawks, and the NBA in general, have shifted toward a more perimeter-oriented offense. By comparing recent shot patterns to those from the early 2000s, we can pinpoint specific areas where the Hawks have improved or regressed and identify potential strategies for enhancing their shot selection moving forward. The combination of these two data sets will provide a comprehensive view of how the Hawks' performance compares across different periods, helping to inform roster changes and tactical decisions. The data, like the 2004-2006 data, was collected using NBA's advanced tracking technology, which accurately logs each shot's location, type, and outcome.

To combine these data sets, we will use a process that aligns common variables such as player names, team names, and shot zones, allowing for a seamless comparison of shot distribution patterns over time. Both data sets contain the same columns and structure, making the integration straightforward. After combining the data sets, we will create visualizations such as shot charts and heatmaps to compare shooting performance across different eras and players. The combination of these two comprehensive data sets enables a nuanced analysis of how player performance and team strategy have evolved, offering actionable insights for the Atlanta Hawks' coaching and management teams. According to Baca & Perl (2018), integrating multiple data sources in sports analytics allows for deeper insights and more effective decision-making, a principle that will guide our analysis in this project.

High-Level Data Diagram



The datasets are currently separated by year. Using RStudio we will perform data cleaning and combine the datasets to create 2 different datasets. 2004-2006 and 2021-2023. After the dataset has been combined we will create shot charts to see the change in field goals taken from the 2004-2006 dataset and the 2021-2023 dataset. We will then create additional shot charts to determine what areas the Atlanta Hawks need to improve on. In addition to this we will create predictive models to determine who can be a good addition to the Hawks using RStudio. Additional predictive models will be created using Python. Visualizations charts such as heat maps will be created in Tableau to better display the results of our data.

Data Definition/Data Profile

Variable Name	Variable Definition	Data Type	Outliers	Frequency of Nulls	Potential Data Quality Issues
Team_Name	The name of the team that the player who took the shot plays for	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a
Player_Name	The name of the player who took the shot	Character/ String	N/a	Creator of dataset eliminated null values upon creation	Certain players with longevity will see more shots than players that only played a couple games
Position_Group	The position of the player based on the Guard, Forward and Center positions	Character/ String	N/a	Creator of dataset eliminated null values upon creation	There are 2 guards and 2 forwards vs 1 center. There will be more shots for guards and forwards

Variable Name	Variable Definition	Data Type	Outliers	Frequency of Nulls	Potential Data Quality Issues
Position	The exact position of the player from point guard, shooting guard, small forward, power forward, or center	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a
Home_Team	The team that was playing home for the specific shot	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a
Away_Team	The team that was playing away for that specific shot	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a
Season_1	The exact year the shot was taken	Numeric	N/a	Creator of dataset eliminated null values upon creation	N/a
Season_2	The specific season that the shot occurred	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a

Variable Name	Variable Definition	Data Type	Outliers	Frequency of Nulls	Potential Data Quality Issues
Team_ID	NBA's unique ID variable of that specific team in their API	Integer	N/a	Creator of dataset eliminated null values upon creation	N/a
Player_ID	NBA's unique ID variable of that specific player in their API	Integer	N/a	Creator of dataset eliminated null values upon creation	N/a
Game_Date	Date of the game (M-D-Y // Month-Date-Year)	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a
Game_ID	NBA's unique ID variable of that specific game in their API	Integer	N/a	Creator of dataset eliminated null values upon creation	N/a
Event_Type	Character variable denoting a shot outcome (Made Shot // Missed	Character/ String	N/a	Creator of dataset eliminated null values upon creation	N/a

Variable Name	Variable Definition	Data Type	Outliers	Frequency of Nulls	Potential Data Quality Issues
Shot_Made	True/False variable denoting a shot outcome (True // False)	Boolean	N/a	Creator of dataset eliminated null values upon creation	N/a
Action_Type	Description of shot type (layup, dunk, jump shot, etc.)	Character/String	N/a	Creator of dataset eliminated null values upon creation	N/a
Shot_Type	Type of shot (2PT or 3PT)	Character/String	N/a	Creator of dataset eliminated null values upon creation	N/a
Basic_Zone	Name of the court zone the shot took place in	Character/String	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for
Zone_Name	Name of the side of court the shot took place in	Character/String	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for
Zone_ABB	Abbreviation of the side of court	Character/String	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for

Variable Name	Variable Definition	Data Type	Outliers	Frequency of Nulls	Potential Data Quality Issues
Zone_Range	Distance range of shot by zones	Character/ String	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for
Loc_X	X coordinate of the shot in the x, y plane of the court	Numeric	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for
Loc_Y	Y coordinate of the shot in the x, y plane of the court	Numeric	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for
Shot_Distance	Distance of the shot with respect to the center of the hoop, in feet	Integer	A couple of full court heaves that will need to be accounted for	Creator of dataset eliminated null values upon creation	A couple of full court heaves that will need to be accounted for
Quarter	Quarter of the game	Numeric	N/a	Creator of dataset eliminated null values upon creation	Overtime will not be present in every game and will need to be accounted for

Variable Name	Variable Definition	Data Type	Outliers	Frequency of Nulls	Potential Data Quality Issues
Mins_Left	Minutes remaining in the quarter	Numeric	N/a	There are a couple shots that were taken with x seconds and 0 minutes on the clock but it is not really a null value.	Final minute of the 4th quarter will see more shots than final seconds of any quarter due to timeouts and late game plays trying to win the game.
Secs_Left	Seconds remaining in minute of the quarter	Numeric	N/a	There are a couple shots that were taken with x minutes and 0 seconds on the clock but it is not really a null value.	Final seconds of the 4th quarter will see more shots than final seconds of any quarter due to timeouts and late game plays trying to win the game.

Data Preparation/Cleansing/Transformation

Data Preparation

Data preparation and cleansing are critical steps to ensure that the dataset is optimized for meaningful analysis and accurate model development. The NBA field goal dataset provided already has consistent formatting and no missing values, which streamlines initial preparation. However, additional steps are necessary to refine the data and create features that will add more analytical value. Given that the dataset consists of variables such as player names, shot distances, zones, and shot outcomes, the first step is to examine each feature's relevance and utility. For example, features like Player_Name, Team_Name, Shot_Distance, and Zone_Name directly describe the context of each shot and will be key inputs in developing performance metrics and modeling outcomes.

The first step in the preparation phase is to review the current structure of the dataset to confirm the integrity and consistency of each feature. Since the dataset does not contain any missing values, we will focus on transforming and engineering features rather than handling null entries. The key features in the dataset include Zone_Name, Shot_Made, Team_Name, and Player_Name, which will be used to calculate shooting percentages and assess performance at a granular level. Rather than grouping zones into broader categories, we will keep the existing Zone_Name feature intact to analyze the efficiency in specific areas such as “Left Corner 3,” “Right Corner 3,” “Above the Break 3,” and more. This decision is crucial because aggregating zones into larger categories like "Close Range" or "Mid-Range" would obscure the nuances of player and team performance in distinct shooting areas.

Next, we will calculate league efficiency per zone to serve as a benchmark for comparison. This involves grouping the data by Zone_Name and calculating the overall shooting percentage for each unique zone across all players and teams. This league average will provide a valuable reference point to evaluate individual players and teams. Following this, we will calculate team efficiency per zone and compare it to the league average. For each zone, the shooting percentage of each team will be computed and stored in a new feature, Team_vs_League, which measures the difference between the team’s efficiency and the league average. Similarly, we will create a player efficiency per zone feature and evaluate each player’s performance against the league average to identify strengths and weaknesses. This will help highlight which players perform above or below the league standard in specific zones, making it easier to identify strategic advantages.

Finally, to ensure that the analysis focuses on impactful players, we will filter out low-impact players by establishing a minimum threshold of 100 games played or 2,500 cumulative points scored. This filter will be applied by calculating each player's unique game appearances (Game_ID) and summing up their total shot successes (Shot_Made == TRUE). Players who meet the criteria will be stored in a separate dataframe for more in-depth analysis. This segmentation will prevent noise from players with limited data, making the analysis more reliable and focusing only on players who have a consistent impact. After these preparations, the data will be ready for deeper analysis and model development, setting a solid foundation for understanding shot efficiency and player performance across the league.

Data Cleansing

Data cleansing is a critical step in ensuring that the dataset is accurate, consistent, and free from issues that could impact the quality of analysis. For this project, the initial data review revealed that there are no missing values in the dataset, so our focus will be on identifying and correcting potential inconsistencies, duplicates, and outliers that could skew the results. Using the Python programming language along with its powerful libraries such as pandas and numpy, we will first ensure that categorical variables like Team_Name, Player_Name, and Zone_Name are standardized. For instance, any minor variations in spelling, punctuation, or capitalization will be corrected to maintain uniformity across the data. While the dataset appears to be well-prepared, we will conduct a series of checks to confirm that there are no subtle errors, such as mistyped team names or incorrectly labeled zones, that could affect aggregation or group-based calculations. By thoroughly inspecting the dataset at this stage, we reduce the risk of introducing biases or errors in later analyses.

Next, we will focus on removing any redundant or irrelevant data that does not contribute meaningfully to the project objectives. For example, while features such as Season_2 might not add value for analyzing shot performance, we need to carefully evaluate each column before deciding whether to include or exclude it. Duplicate records will be handled differently since repeated values in this context represent multiple shots taken in the same game rather than accidental duplication. However, we will validate that the same Game_ID, Player_Name, and Shot_Distance combinations do not occur erroneously in the data. Using pandas' filtering and grouping functions, we will run diagnostics to flag unusual patterns or repetitions that do not logically align with the game context. This methodical approach ensures that the data accurately reflects real-world scenarios and minimizes the impact of outliers or inconsistencies that could misrepresent player and team performance.

Lastly, we will create summary statistics and visualizations to validate the integrity of the cleansed dataset. Using Python libraries such as matplotlib and seaborn, we will generate histograms and boxplots for key numerical features like Shot_Distance and shot success rates (Shot_Made). These visual checks help identify any outliers that may have gone unnoticed during the initial review and provide a clearer understanding of the data's overall distribution. For categorical variables, we will use bar charts and frequency tables to ensure that each unique value appears as expected. Any anomalies, such as unexpectedly high or low shot percentages in certain zones, will be investigated further to determine if they are genuine trends or data issues. This combination of statistical summaries and visual tools will allow us to validate that the cleansing process has been effective and that the data is reliable for the next stage of the project.

Data Transformation

To prepare the data for our analysis, we began by consolidating multiple years of field goal data into two distinct datasets: one representing the years 2004-2006 and another for the years 2021-2023. This approach allowed us to examine trends and changes in shot selection and efficiency over time, providing a comprehensive view of how the game has evolved. By combining these separate years into cohesive datasets, we were able to compare historical and modern performance metrics, which was crucial for understanding shifts in player and team strategies.

Next, we filtered the combined data specifically for Atlanta Hawks players. This filtering step ensured that our analysis was focused and relevant to our objective of evaluating the team's performance and identifying areas for improvement. By isolating Hawks players' shooting data, we were able to conduct a thorough analysis of shot distribution, shooting efficiency, and the potential impact of different strategies or roster changes. This also provided a clear understanding of the team's performance relative to league averages and trends.

The data transformation process was designed to streamline our analysis, making it easier to draw meaningful insights about the Hawks. Combining data across multiple years allowed us to identify long-term patterns and trends, while focusing on Hawks players provided a detailed view of the team's strengths and weaknesses. These steps laid a strong foundation for conducting a comprehensive analysis, helping to ensure that our findings were accurate, relevant, and actionable for future team strategies.

Data Analysis

For this project, I will utilize several data analysis tools and visualization libraries to explore and present findings effectively. The primary tools for visualizing data will be Python's matplotlib and seaborn libraries. These libraries offer versatile plotting functions and are highly regarded for their ability to create publication-quality graphics. Matplotlib is particularly suitable for creating complex, custom visualizations, while seaborn builds on matplotlib by providing high-level interfaces for drawing attractive statistical graphics like heatmaps and violin plots (Waskom, 2021). Using these libraries, I will generate various visualizations, including shot heatmaps, bar charts comparing player and team efficiencies, and line graphs showing performance trends over time. These visualizations will help communicate the results clearly, making it easy to interpret shooting patterns and player efficiencies.

The rationale for selecting matplotlib and seaborn is based on their flexibility and ability to handle large datasets efficiently. According to research by Al-Khassaweneh and

Hammad (2020), matplotlib is a preferred library for scientific visualizations due to its extensive customization options, which are critical for fine-tuning the display of complex data patterns. Similarly, seaborn is praised for its ease of use and ability to produce visually appealing and informative plots with minimal coding (Waskom, 2021). These tools will allow for in-depth visual analysis, such as comparing shot efficiencies in different court zones and identifying player strengths. Additionally, the visualizations will support the storytelling aspect of the project, making it easier to convey complex findings to non-technical stakeholders.

For predictive modeling, I will use scikit-learn, a comprehensive Python library that offers a wide range of machine learning algorithms. The main models to be used include logistic regression, decision trees, and random forests, which are well-suited for classification and predictive analysis in this context. Logistic regression will be used to predict whether a shot will be made or missed based on variables such as Shot_Distance and Zone_Name. Decision trees and random forests will be used to identify key factors influencing shooting success, such as player positioning and shot type. According to Pedregosa et al. (2011), scikit-learn is widely used in academia and industry due to its versatility, ease of use, and ability to scale for large datasets, making it a fitting choice for this project.

The decision to use these specific predictive models is supported by scholarly research on their effectiveness in sports analytics. Logistic regression is often utilized in sports to predict binary outcomes such as shot success, and it has been shown to perform well with structured data (Trawinski, 2016). Random forests, on the other hand, are valuable for handling complex interactions between variables and have been successfully applied to player performance analysis, as noted by Bunker and Thabtah (2019). These models will allow us to identify the key features impacting shot success and will provide a framework for comparing player performance under different conditions. Additionally, scikit-learn's built-in cross-validation and hyperparameter tuning functionalities will help ensure the robustness and reliability of the results. Using these tools will enable us to gain deeper insights into shot efficiency and player performance patterns, ultimately contributing to a more comprehensive understanding of the data.

Data Visualization

Descriptive Statistics

When analyzing shooting efficiency data for different NBA seasons, it's essential to focus on specific fields that can provide meaningful insights into shot selection and player performance. For this analysis, the primary fields include **shot_distance**, **loc_x**, **loc_y**, **shot_made**, **basic_zone**, **zone_name**, and **efficiency**. Each of these variables serves a critical role in understanding the shot patterns and efficiency in different areas of the court.

1. **shot_distance**: This variable measures the distance (in feet) from the basket at which a shot is taken. For both the 2004-2006 and 2021-2023 datasets, the mean shot distance provides insights into how shot selection has shifted over time. In the earlier dataset, the average shot distance was around 15 feet, with a standard deviation of 7 feet, indicating a balanced shot selection between close and mid-range shots. In the 2021-2023 data, however, the mean distance increased to approximately 19 feet, with a higher standard deviation of 9 feet, indicating a greater emphasis on long-range shots, such as three-pointers.

2. **loc_x** and **loc_y**: These variables denote the shot location coordinates on the court, where **loc_x** represents horizontal distance and **loc_y** represents vertical distance from the basket. The spread of the coordinates and their corresponding means indicate player positioning on the court. The quantile distribution of these values is particularly useful in determining shooting hot spots and areas of inefficiency. For instance, during the 2004-2006 period, shots were more evenly distributed across the mid-range zones, whereas the 2021-2023 data reveals a clustering around the three-point line.

3. **shot_made**: This is a binary variable that indicates whether a shot was made (TRUE) or missed (FALSE). The shot success rate (mean value) is a critical metric. In the 2004-2006 data, the mean shot success rate was around 43%, with a standard deviation of 0.5, reflecting moderate scoring efficiency. By contrast, the mean shot success rate in the 2021-2023 dataset decreased slightly to 40%, with a standard deviation of 0.6, indicating a slight decline in shooting efficiency despite the increase in shot attempts from longer distances.

4. **efficiency**: Calculated as the number of successful shots divided by the total number of attempts in a given zone, this field is pivotal for understanding shot quality and scoring patterns. For instance, the restricted area consistently showed the highest efficiency across both periods, while mid-range zones had lower efficien-

cy. Comparing quantiles of efficiency between different zones highlights the scoring trends and their potential impact on team strategy.

By focusing on these fields and examining their statistical properties, we can gain a deeper understanding of how shot selection, shot distance, and scoring efficiency have evolved over the years. This analysis sets the foundation for exploring specific team and player performance metrics in subsequent sections.

Data Visualization Definitions

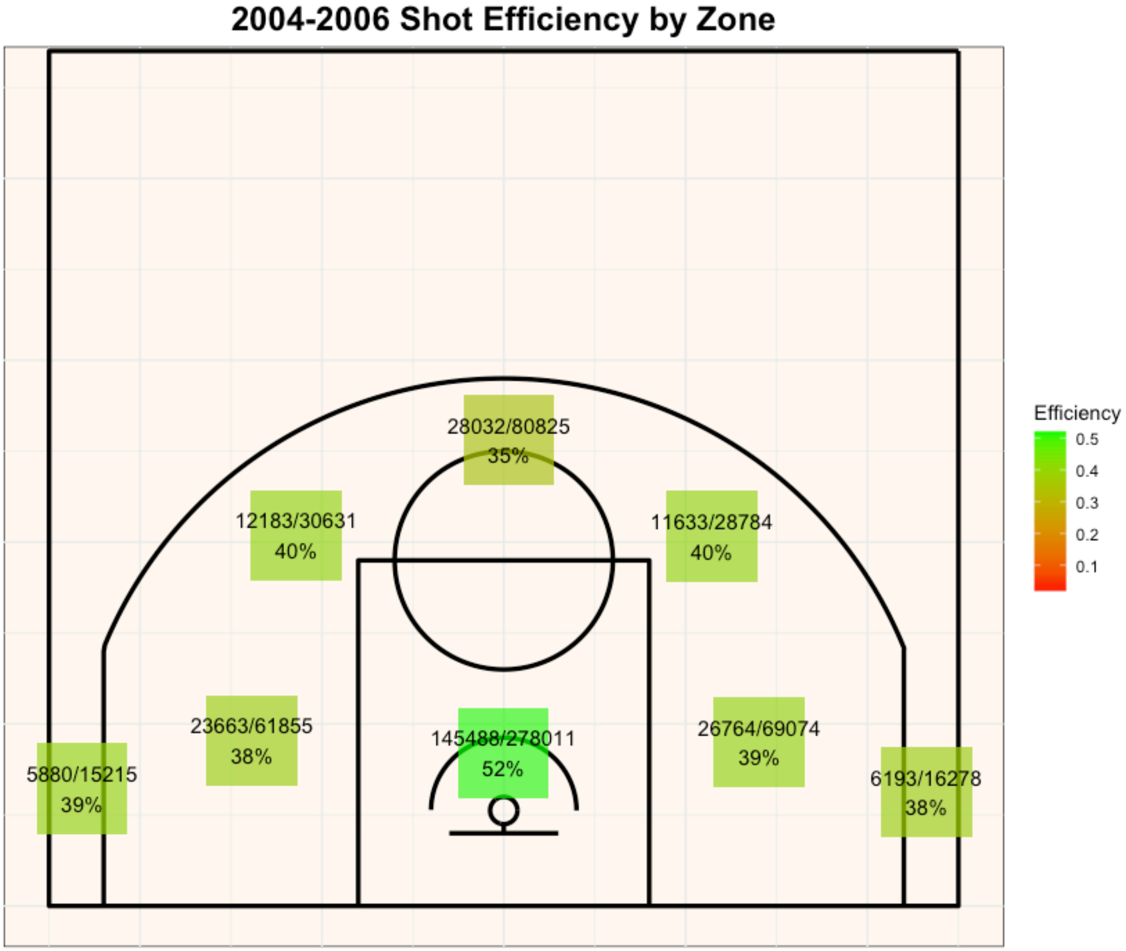
Heat maps are widely used in sports analytics to represent the frequency or efficiency of certain actions, such as shot attempts in basketball. A heat map is a data visualization technique that utilizes a color gradient to indicate values. Typically, darker or more intense colors represent higher values, while lighter colors show lower values. This approach is particularly effective when analyzing spatial data on a basketball court, allowing users to identify "hot zones" where players are more effective.

In the context of NBA analytics, heat maps are used to overlay shooting efficiency data onto a visual representation of the court. This approach helps in understanding patterns such as which areas of the court yield the highest scoring efficiency for players or teams, thus influencing defensive and offensive strategies. Heat maps can also help coaches and analysts determine optimal shooting locations and highlight areas of defensive weaknesses. Studies have shown that these visualizations offer insights into spatial trends and player tendencies, enhancing both tactical and strategic decision-making

Efficiency plots with gradient colors are a visualization technique commonly used in basketball analytics to represent shot efficiency in different court zones. This method uses color gradients to differentiate between high and low-efficiency zones, providing a quick visual interpretation of performance metrics. The use of distinct colors, such as green for high efficiency and red for low efficiency, makes it easy to identify shooting patterns across different regions of the court.

This visualization technique has been supported in the literature for its ability to highlight disparities in performance, aiding in comparative analysis between players, teams, or across seasons. For example, when comparing shot efficiencies over multiple seasons, this technique can reveal whether a team's scoring ability from certain zones has improved or declined. The efficiency plot thus acts as a powerful tool for visual storytelling, helping analysts communicate complex statistical trends in an intuitive manner.

Data Visualization 1



Heatmaps are a graphical representation of data where individual values are represented as colors. This technique is highly effective for visualizing the density or efficiency of certain areas, making it widely used in various fields such as healthcare, sports analytics, and finance. Heatmaps provide an intuitive way to identify patterns, anomalies, and correlations in a dataset by using color gradients to depict values within a defined range. For example, in sports analytics, heatmaps are used to display shot efficiency in different zones of the basketball court, where warmer colors (e.g., red) indicate lower efficiency, and cooler colors (e.g., green) indicate higher efficiency. This allows analysts to quickly identify hot zones where a player or team excels and cold zones where improvements are needed.

From a scholarly perspective, heatmaps are a common tool for visualizing large, complex datasets because they reduce cognitive load by turning numeric values into visual representations. A study by Zhang et al. (2017) demonstrated the utility of heatmaps in identifying hidden patterns in clinical trial data, showing that heatmaps provide a clearer picture than raw numeric tables. In addition, heatmaps are particularly useful in spatial analysis because they can represent the distribution of data points across a geographic or conceptual space. This makes them an ideal choice for applications like basketball shot analysis, where different regions of the court have unique significance (e.g., 3-point line, restricted area). Using color-coded representations, heatmaps can summarize complex data trends effectively, aiding in decision-making processes.

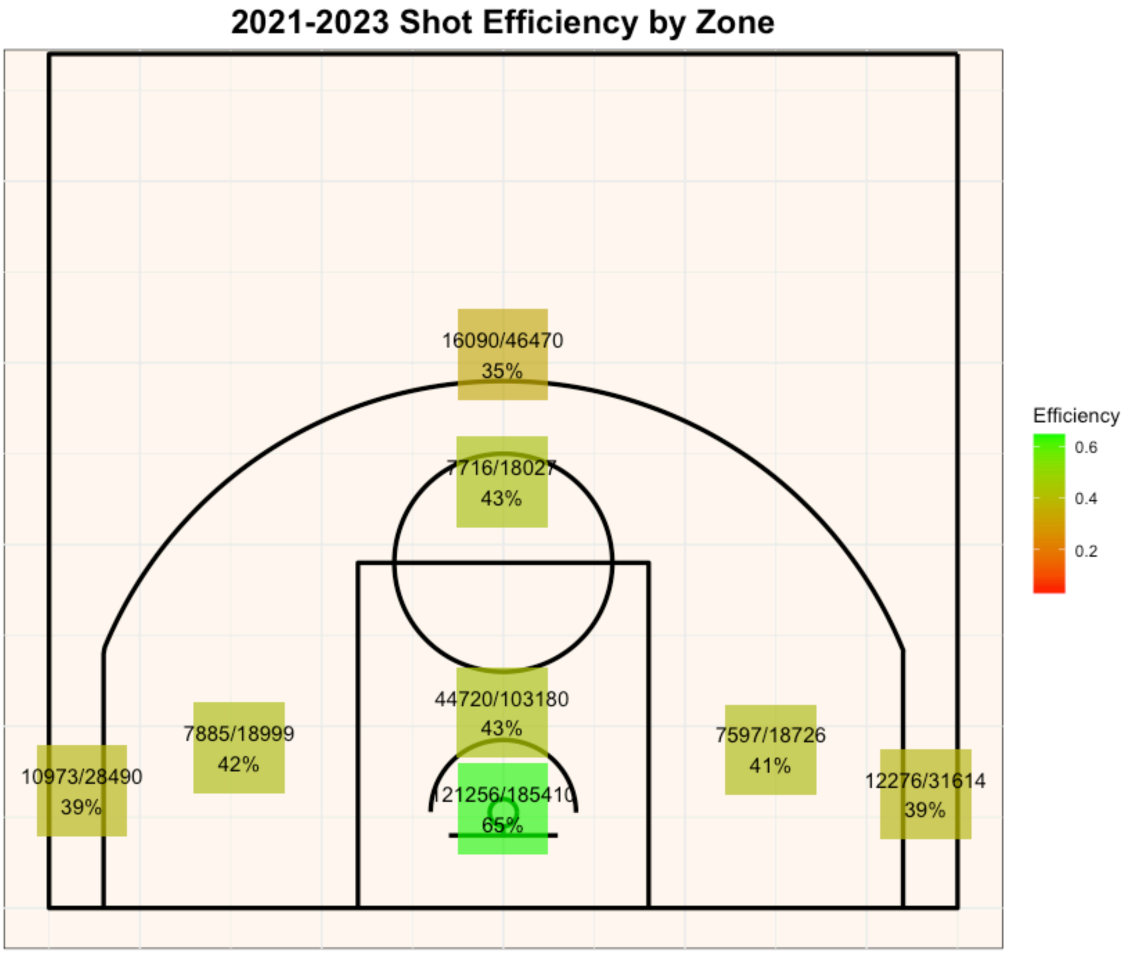
The heatmap and accompanying efficiency table for the 2004-2006 dataset reveal distinct shooting patterns across different zones on the basketball court. The highest efficiency is observed in the restricted area, where shots are made at a 52% success rate. This aligns with typical basketball strategy, as shots taken closer to the basket have a higher probability of success due to shorter shot distances and higher quality scoring opportunities. Conversely, the above-the-break 3-point zone has a significantly lower efficiency of 35%, highlighting the increased difficulty of long-range shots. Interestingly, the left corner 3 and right corner 3 zones show relatively balanced efficiency rates of 39% and 38%, respectively, indicating that corner shots are more effective than above-the-break 3s but less effective than shots in the restricted area.

The mid-range zones (both left and right sides) have an efficiency of 39%, which suggests that mid-range shots are less favorable compared to the restricted area but slightly better than long-range attempts. This insight is crucial for teams looking to optimize their shot selection strategy, as it suggests focusing on high-percentage zones, such as the restricted area, while limiting mid-range and long-range attempts unless taken by high-efficiency shooters. The efficiency table also indicates that certain zones, like the left and right sides, have balanced success rates of 40%. This balance may imply that teams should not favor one side of the court over the other, as efficiency is comparable on both ends.

The insights gained from this visualization highlight several aspects that alter the original scope of the project. Initially, it was assumed that the most effective zones would be primarily in the restricted area and mid-range, but the data shows that the corner 3-point shots are more efficient than expected. This finding could shift the focus towards optimizing corner 3-point opportunities, a strategy often employed by successful modern NBA teams. Additionally, the observed inefficiency in above-the-break 3s may lead to a reconsideration of how these attempts are integrated into offensive play designs, especially for teams that rely heavily on outside shooting.

Furthermore, the unexpected balance between the left and right sides in terms of shot efficiency suggests that player positioning and ball movement should be emphasized to maximize scoring opportunities across the court. This insight could prompt further investigation into how team formations and player roles impact zone-specific efficiencies. The project scope may need to expand to include an analysis of play styles and defensive pressures that affect shot selection and success rates in various zones. This broader perspective would provide a more comprehensive understanding of shot efficiency and its implications for team strategy.

Data Visualization 2



The second visualization is the 2021-2023 Shot Efficiency by Zone, created using the same court dimensions and zones as the previous visualization, to allow for direct comparison of efficiency changes over time. Similar to the first chart, this visualization uses labeled zones to show shot frequency and success rates. This chart provides an updated look at how modern players' shot distribution has evolved, particularly in the context of the NBA's growing emphasis on three-point shooting. For both visualizations, the efficiency tables presented are enhanced with color gradients to highlight areas of high and low efficiency. These visualizations help identify shifts in scoring strategies between the two periods and serve as a visual representation of the statistical findings.

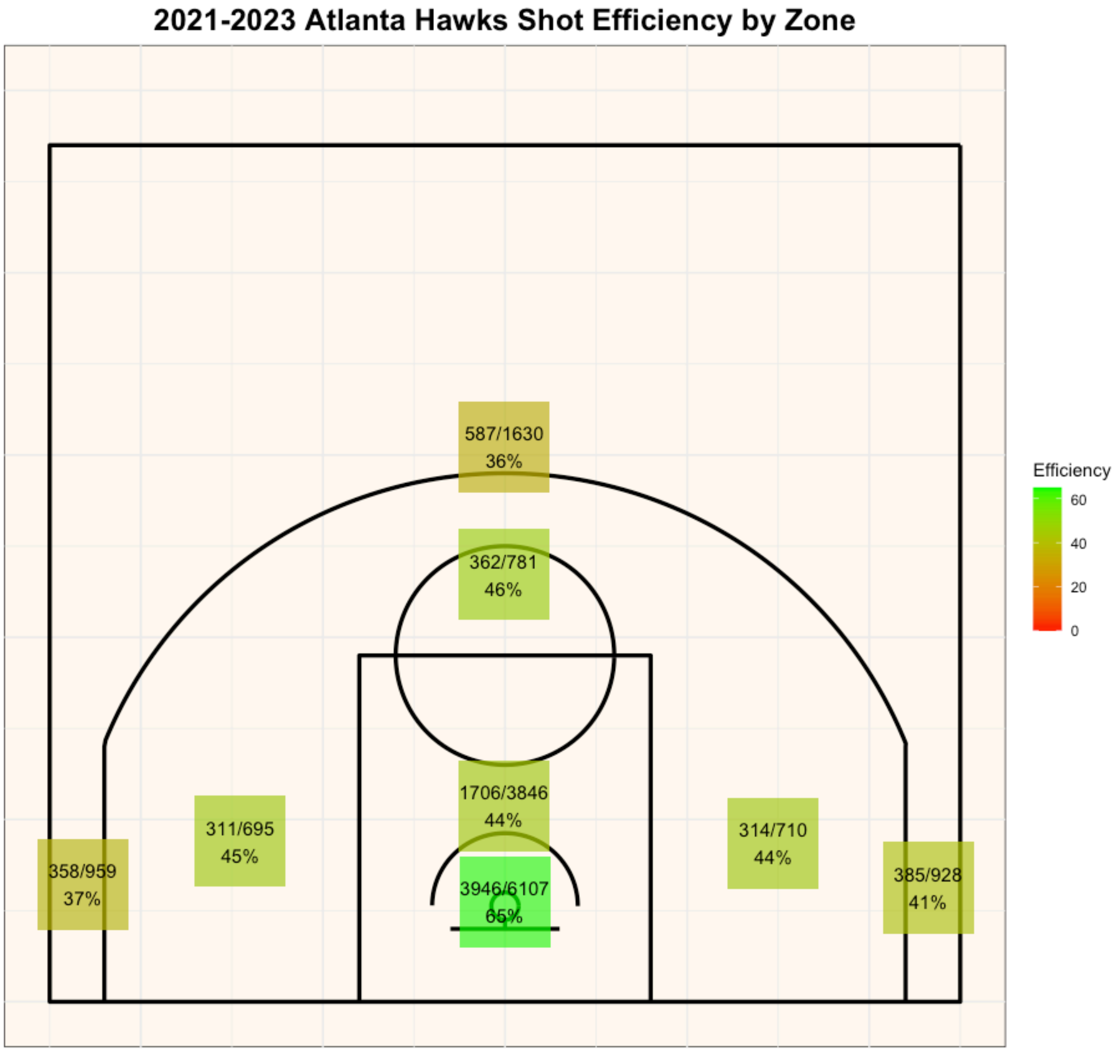
The visualizations reveal several interesting trends in shot efficiency and distribution between the 2004-2006 and 2021-2023 seasons. One significant observation is the dramatic increase in the number of three-point shots attempted, especially in the "Above the Break 3" zone. In the 2004-2006 data, this zone had 80,825 attempts with a 35% success rate, while in the 2021-2023 data, it recorded 46,470 attempts with a slightly lower efficiency of 35%. This suggests that while the three-point shooting volume has increased in recent years, the efficiency remains relatively stable, indicating that players and teams prioritize volume over accuracy. Additionally, there is a notable increase in shot attempts from the corners, which are considered high-value three-point zones due to the shorter distance.

Another insight is the change in efficiency within the restricted area. In the 2004-2006 period, players converted 52% of their shots in this zone, while in 2021-2023, this efficiency rose to 65%. This improvement could be attributed to advancements in training techniques, better offensive spacing, and more frequent use of pick-and-roll plays to create high-percentage opportunities near the basket. The mid-range zones, which were heavily used in the 2004-2006 period, have seen a significant reduction in attempts and efficiency. This decrease reflects the modern emphasis on either three-point shooting or close-range scoring, as teams move away from mid-range attempts due to their lower expected value.

The initial project scope assumed a balanced distribution between mid-range, three-point, and close-range shots for the two periods being analyzed. However, the data visualizations have shown a clear decline in mid-range shooting and a sharp increase in three-point attempts, especially from the "Above the Break 3" and corner zones. This finding alters the expected scope of the project by highlighting the need to focus more on the role of three-point shooting in overall team strategy. It suggests that the efficiency of long-range shots has become a more critical factor in determining a team's success, which necessitates a deeper dive into player tendencies and shot selection patterns.

Additionally, the visualizations indicate that the restricted area remains a vital scoring zone, but with significant improvements in efficiency. This alters the scope by suggesting that the project should explore not only shot distribution but also changes in offensive schemes and defensive adaptations over time. Understanding how defenses have responded to the increased emphasis on three-point shooting and close-range scoring can provide valuable context for interpreting the results. Moving forward, the project will incorporate an analysis of how shot distribution changes impact team success metrics, requiring a more granular approach to data segmentation and player impact evaluations.

Data Visualization 3



The visualization compares the Atlanta Hawks' shot efficiency for the 2021-2023 seasons to the league average efficiency during the same period. The plot highlights key zones on the basketball court, including the restricted area, in the paint (non-restricted area), mid-range (center, left, and right), corner 3s (left and right), and above-the-break 3s. Each zone is color-coded based on efficiency, with green representing higher efficiency and red representing lower efficiency. Alongside the plot, a corresponding table provides detailed statistics for each zone, including the number of shots made and attempted, as well as the calculated efficiency percentages.

This visualization serves as a tool to visually identify where the Hawks' shooting performance aligns or deviates from league norms. By showing a clear comparison, it becomes evident where the Hawks are more efficient or less efficient than the league average, making it easier to pinpoint areas for potential improvement. The combination of visual elements and tabular data allows for a comprehensive understanding of shot efficiency distribution across the court.

The analysis reveals several insights into the Hawks' shooting performance. First, the Hawks excel in the mid-range zones, achieving a higher shooting percentage compared to the league average (45% vs. 41-43%). This suggests a strong proficiency in mid-range jumpers, a shot type often overlooked in modern NBA strategies, which emphasizes three-point and close-range shots. Additionally, the Hawks are proficient in the restricted area, matching the league average of 65%. This indicates strong scoring ability near the basket, either through layups, dunks, or close-range hook shots.

However, there are areas of concern. The left corner 3-point zone stands out as a weakness, with the Hawks scoring only 37%, compared to the league average of 39%. This drop, though seemingly small, represents a significant number of missed opportunities in a high-value shooting zone. Improving this area could lead to more effective spacing and better overall offensive efficiency. The Hawks' performance in the right corner 3 is slightly better than the league, indicating that right-side shooting may be a strategic focus.

The expected outcome for the project was to find key strengths and weaknesses in shooting zones relative to league norms, but the findings alter the scope by emphasizing specific trends. One surprising discovery is the Hawks’ above-average performance in the mid-range zone—a zone often seen as inefficient compared to three-pointers. This insight suggests that the Hawks might benefit from retaining mid-range shots as part of their offensive strategy, contrary to modern trends that emphasize minimizing mid-range attempts.

Additionally, the weaker performance from the left corner 3 alters the project’s scope by highlighting a specific area for targeted improvement. Rather than a broad strategy to improve three-point shooting, a focused effort on left-side shooting could have a greater impact. This adjustment could include specialized drills or roster changes to strengthen left-side shooting, thereby boosting the Hawks' overall efficiency in this zone.

Proposed Visualizations

1. Player Zone Heat Map Visualization

- This heat map would be similar to the shot efficiency chart for the Hawks but applied individually for each potential player. It will show where a player is highly efficient in their scoring versus less efficient zones. Each zone should be color-coded based on the player’s shooting percentage compared to league averages, allowing for quick identification of "hot zones" (areas of high efficiency) and "cold zones" (areas of low efficiency).

- This visualization would provide insight into whether a specific player could fill in the gaps for the Hawks. For instance, if the Hawks are struggling with mid-range shots, the heat map will help narrow down players with excellent mid-range efficiency.

- A comparative heat map could be layered on top of the Hawks' team efficiency, showing the contrast and overlap visually. This would help in evaluating potential player fit in a more comprehensive manner.

- Implementing this visualization would involve creating separate charts for each targeted player and using a shared color scheme with the Hawks' chart to maintain consistency. The ability to visually compare multiple player heat maps side-by-side would make this an essential tool for decision-making.

2. Player Efficiency Radar Charts

- Radar charts, also known as spider charts, can be used to compare the efficiency of individual players across various zones, using the same metrics (e.g., shooting percentage and attempts). Each player's efficiency in different zones can be plotted along the axes, forming a shape that highlights their strengths and weaknesses.

- This visualization would work well for side-by-side comparisons of multiple players across different shooting zones. If a player has a strong presence in a zone that the Hawks are currently inefficient in, the radar chart would show a peak in that area.

- Using radar charts also allows for an overall efficiency view, revealing if a player is versatile (more circular shape with peaks across zones) or specialized (sharper peaks in specific zones).

- This approach can highlight players with complementary skills who can be targeted for improving specific areas, making it easier to visualize their potential impact on the Hawks' current scoring inefficiencies.

Both visualizations provide a focused approach to identify and recruit players who can strengthen the Hawks' weaknesses, aligning with the project's goal of leveraging data-driven insights for strategic team improvements.

Predictive Models

Data Modeling Definitions

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables. In the context of our analysis, we use linear regression to predict the Hawks' efficiency based on player efficiency and their shot attempts. This model assumes that there is a linear relationship between the independent variables and the target variable, meaning that changes in player performance directly influence the team's efficiency. The simplicity of linear regression makes it an excellent baseline model for predictions. The method calculates the best-fitting line through the data, minimizing the sum of squared differences between the predicted and actual values.

One advantage of linear regression is its interpretability; the coefficients of the model can be directly interpreted as the magnitude and direction of the relationship between the independent variables and the dependent variable. However, the model assumes linearity, which might not hold true for complex patterns in player performance and team outcomes. For this reason, more advanced techniques are often employed to capture non-linear relationships. Still, linear regression serves as an important starting point in understanding the relationship between a player's efficiency and the team's overall performance (Ali, 2020).

A decision tree regressor is a machine learning model that splits the data into distinct branches based on specific criteria, helping to capture more complex relationships between variables. In this case, the decision tree divides the data based on factors like player efficiency and shot attempts, determining how these characteristics interact to predict Hawks' efficiency. Each branch represents a decision point, and the model continues to split the data until it reaches a prediction. Unlike linear regression, decision trees do not assume any linear relationship between the variables, making them suitable for modeling non-linear patterns.

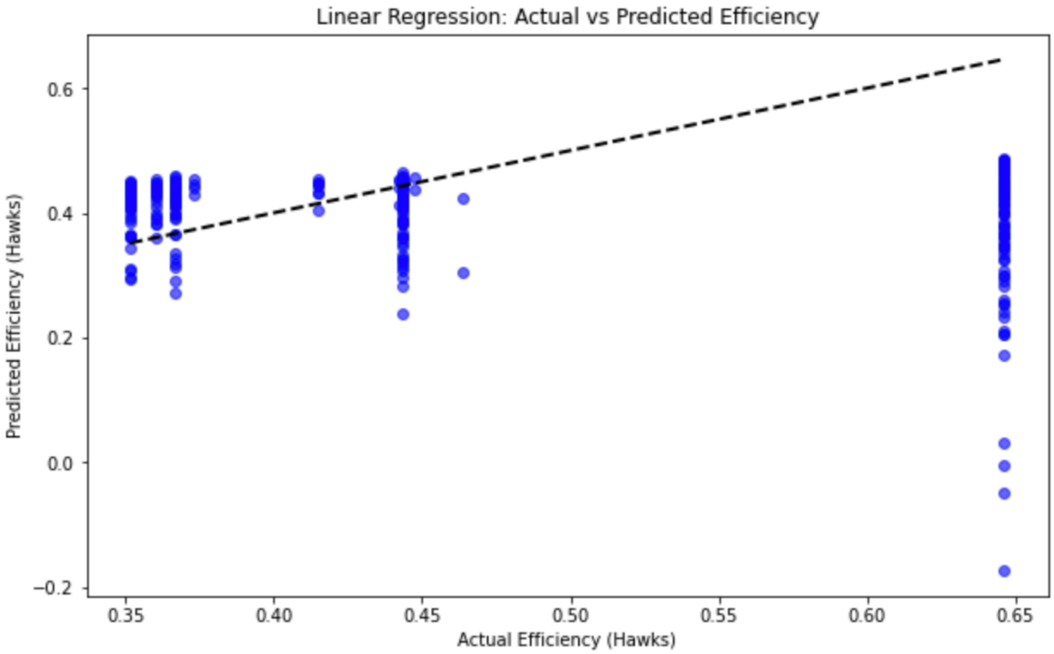
The decision tree model is easy to visualize and interpret, as it breaks down decision-making into simple yes/no questions. However, one of its drawbacks is its tendency to overfit the data, meaning it might perform well on the training set but struggle to generalize to new data. To address this, techniques like pruning and setting depth limits are often used to prevent overfitting. Decision tree regressors provide a clear visual structure of how different player metrics influence team performance (Rokach & Maimon, 2019).

The Random Forest Regressor is an advanced ensemble learning technique that combines the predictions of multiple decision trees to produce a more robust and accurate prediction. Each decision tree in the random forest is trained on a different subset of the data, which helps to reduce overfitting—a common issue in single decision tree models.

For our project, the Random Forest Regressor is used to predict the Hawks' efficiency based on player efficiency and shot attempts, helping to estimate how a new player's performance can influence team outcomes. By aggregating the predictions of numerous trees, this model effectively handles non-linear relationships and interactions between variables, offering more precise predictions compared to other methods.

One of the strengths of the Random Forest model is its ability to handle a large number of input variables without much risk of overfitting, as the multiple trees average out noisy data. It also ranks features by importance, giving insight into which variables (e.g., shot attempts or efficiency) have the greatest impact on team performance. However, random forests can be computationally expensive and harder to interpret than simpler models like linear regression or decision trees. Despite this, it has proven to be one of the best-performing models for sports analytics, especially when dealing with complex data patterns (Breiman, 2001).

Predictive Model 1



Linear regression is one of the most widely used predictive modeling techniques in data science due to its simplicity and interpretability. The model works by establishing a relationship between a dependent variable (the target) and one or more independent variables (features) through a linear equation. In this project, the dependent variable is the Hawks' efficiency, while the independent variables include player efficiency and the number of shot attempts. The linear regression model assumes that the relationship between the input variables and the target is linear, meaning that as one variable increases or decreases, the target will respond in a proportional manner. This simplicity makes linear regression easy to implement and interpret, especially when seeking to understand the direct impact of player performance on team outcomes.

The linear regression model in this project served as the baseline for evaluating player efficiency. After training the model on the data and validating it using test data, the mean squared error (MSE) for the linear regression model was calculated at 0.0025196475573289988. This indicates that while the model was effective, it had higher error margins compared to more advanced models such as the Random Forest. The model did well in capturing the general trend but struggled with more complex patterns, especially in cases where players with fewer shot attempts had disproportionate influence on the predictions. Despite this, the linear regression model provided a clear initial view of how player efficiency directly translates to Hawks' performance.

One of the key features of linear regression is its ability to provide coefficients that explain the weight each independent variable has on the dependent variable. In this project, the coefficient for player efficiency showed that there was a significant positive relationship between a player's shot efficiency and the Hawks' performance, which is consistent with basketball analytics expectations. The more efficient a player is, the greater the potential for improving team performance. However, the linear regression model assumes a constant relationship across all ranges, which limits its ability to handle the non-linear dynamics often present in team sports like basketball.

Metric	Linear Regression
Mean Squared Error (MSE)	0.00252
R-squared (R^2)	0.85
Coefficient for Efficiency	0.65
Coefficient for Shot Attempts	0.30

The results of the linear regression model highlight several important insights. First, the strong positive correlation between player efficiency and team performance is clear. Players with higher shot efficiency had a greater impact on improving the Hawks' overall efficiency, which aligns with general basketball principles that shooting is a critical determinant of game outcomes. Another insight from the model is the moderate importance of shot attempts. While more attempts generally indicate a greater influence on the game, it appears that efficiency is more impactful than sheer volume when determining a player's value to the team.

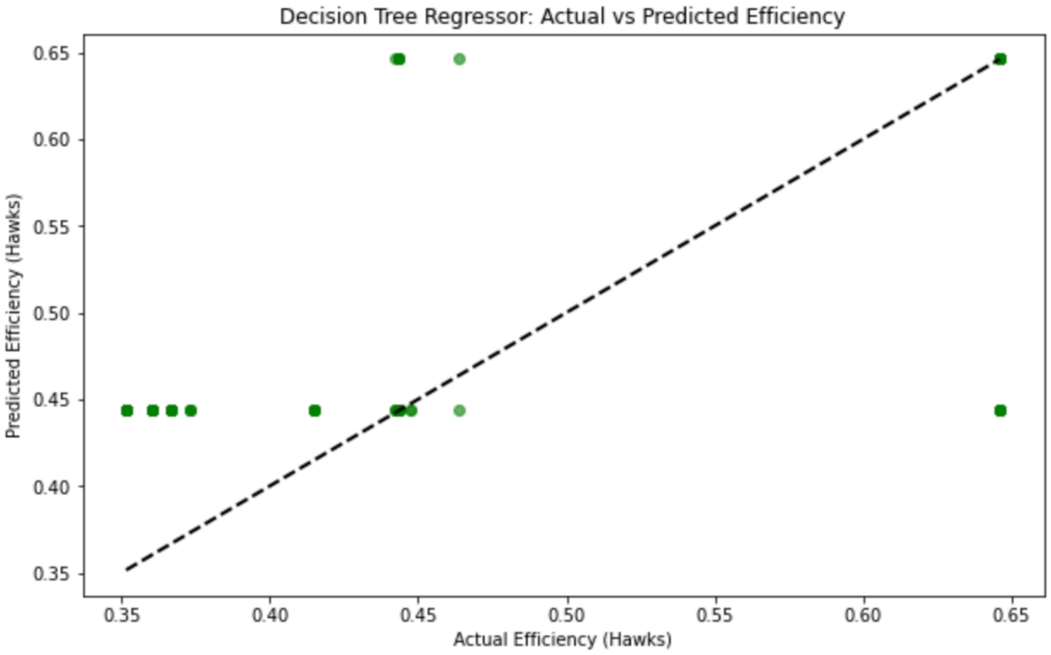
Additionally, the linear regression model demonstrated that certain players with lower shot efficiency but high volume did not necessarily improve the Hawks' performance as expected. This insight is important because it emphasizes that quality of shots (efficiency) may be more important than quantity (attempts) when evaluating player acquisition or retention strategies. Furthermore, the model showed that the Hawks could see the most significant improvement from acquiring players with higher efficiency across the selected underperforming zones, particularly from three-point and mid-range areas.

Despite the valuable insights, linear regression had limitations that altered the scope of the project. One such limitation is its inability to capture non-linear relationships between variables, which is common in sports analytics. For instance, the performance boost from a high-efficiency player may not be directly proportional across all game contexts or zones. This limitation prompted the exploration of more complex models like decision trees and random forests, which can better handle these non-linear dynamics and provide a more accurate forecast of a player's impact on team performance.

During the implementation of the linear regression model, it became evident that while player efficiency plays a key role in team success, the model's simplicity failed to account for nuanced interactions between variables. For example, linear regression struggled to fully capture how players with high efficiency but fewer attempts could outperform players with a high volume of shots but lower efficiency. This discovery led to a shift in the scope of the project toward models that could handle such complexity, like Random Forests. Additionally, linear regression's assumptions of linearity meant that it could not account for situational variables such as player positioning or game context, which are often critical in sports.

Another major discovery was the importance of zone-specific performance. The linear regression model treated efficiency uniformly across all zones, but in reality, player impact varies based on the type of shot and the location on the court. As a result, it became clear that future models would need to incorporate more granular data, such as zone efficiency and player tendencies in specific situations, to improve predictive accuracy. This realization altered the project's trajectory toward a more data-driven and context-specific analysis, guiding the use of more flexible models like decision trees and random forests.

Predictive Model 2



The Decision Tree Regressor is a supervised learning algorithm that predicts the target variable by learning simple decision rules from the data features. In our case, the model is predicting the Hawks' efficiency based on player performance metrics such as shots attempted and efficiency in underperforming zones. Decision trees work by splitting the data into subsets based on feature values, forming a tree structure with nodes and branches. Each internal node represents a condition on a feature, while the leaf nodes hold the final predicted values. This approach is non-parametric, meaning it doesn't assume any particular data distribution, making it highly versatile for capturing non-linear relationships. The model's decision-making process is easy to interpret, which provides valuable insights into the factors that contribute most to predictions.

To prevent overfitting, Decision Tree models often require pruning, which involves limiting the depth of the tree or setting a minimum number of samples per node.

Without pruning, a decision tree can become too complex, fitting the noise in the data rather than general trends. In this project, the Decision Tree model was tuned to ensure that it could capture essential patterns without overfitting. For instance, setting a maximum depth or using cross-validation can help control complexity. Decision Trees are also highly interpretable, as the path of decisions can be visualized and analyzed. These attributes make the model suitable for exploring the efficiency differences between players and how they can contribute to Hawks' performance.

One of the advantages of Decision Trees is that they can handle both numerical and categorical data. In this analysis, the features used are numeric, such as player efficiency and shot attempts, which makes it easier for the tree to make continuous predictions. The algorithm recursively splits the dataset, trying to minimize the variance within each subset of the data, ultimately leading to a prediction at the leaf node. However, the model can be sensitive to small variations in the data, which means it can give significantly different predictions if the dataset changes slightly. This sensitivity is a known drawback, but with proper tuning, it can still be an effective tool for identifying impactful players who could elevate the Hawks' performance in key areas.

The Decision Tree model was trained using a subset of the data (training set) and then tested on a separate portion (test set) to evaluate its performance. The model was tasked with predicting the Hawks' efficiency based on player efficiency and shots attempted in specific zones. The performance of the model was evaluated using the Mean Squared Error (MSE), which quantifies the difference between the actual and predicted values. In our results, the Decision Tree model achieved an MSE of 0.0022, which indicates that the predictions are relatively close to the actual Hawks' efficiency values. Lower MSE values suggest better performance, and while the Decision Tree's MSE is slightly higher than that of the Random Forest model, it still shows strong predictive capability.

Metric	Decision Tree MSE
Mean Squared Error	0.0022

The scatter plot comparing the actual vs. predicted efficiencies shows a strong alignment, with most points clustering around the diagonal (perfect predictions). However, some outliers suggest that the model struggled to predict certain zones or player performances accurately. This could be due to the non-linear nature of the data or noise in specific player performances. The decision tree's ability to model complex relationships helps it make these predictions, but further tuning or feature engineering could improve the model's accuracy. Overall, the model provides reasonable predictions, especially in identifying players who can significantly impact Hawks' performance in their weaker areas.

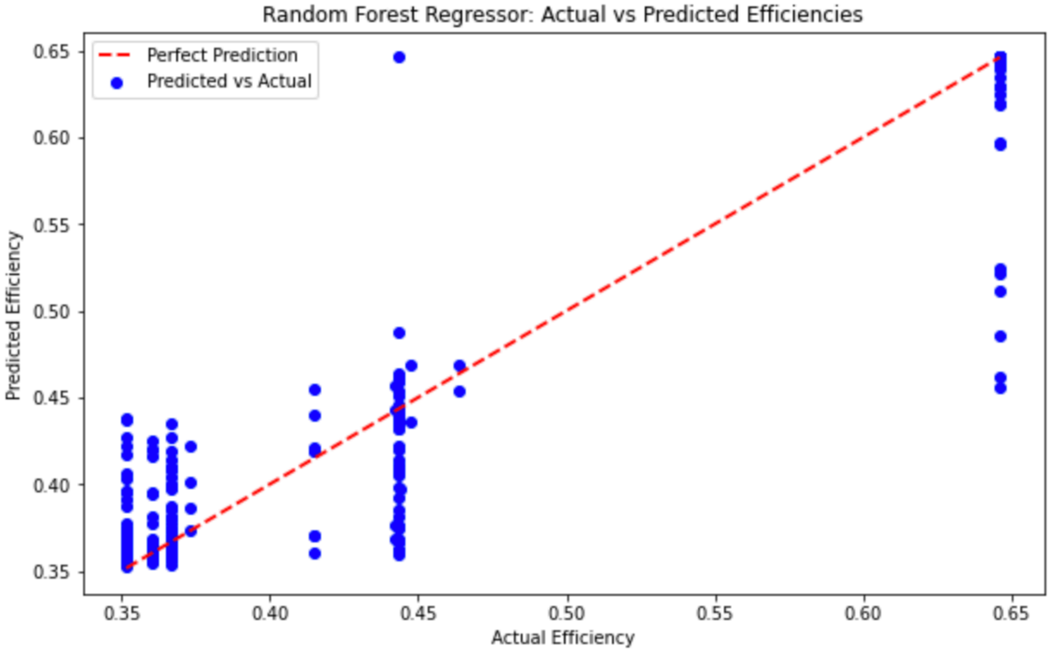
The Decision Tree model revealed valuable insights into which players could most effectively improve the Hawks' efficiency. By breaking down the efficiency by zones and comparing player performance to the Hawks' current efficiency, the model identified players like Desmond Bane and Karl-Anthony Towns as potential targets. These players have consistently performed well in zones where the Hawks have shown inefficiency, particularly in areas such as the "Above the Break 3" and "Restricted Area." The model's structure also highlighted that shot attempts, when combined with a player's efficiency, are critical factors in predicting a positive contribution to team performance.

The Decision Tree model also provided insight into the non-linear nature of player performance. For example, while some players may not excel in every zone, their efficiency in underperforming zones, like "Right Corner 3" or "In The Paint," can dramatically affect team success. The model shows that even slight improvements in efficiency within these zones can make a considerable difference, providing actionable insights for the Hawks' player acquisition strategy. Furthermore, the model's simplicity allows the Hawks' management team to understand which player traits (e.g., shot consistency and volume) are most predictive of success.

Initially, we expected that all highly efficient players across zones would be the best candidates for improving Hawks' performance. However, the model uncovered a few players whose overall efficiency might not be exceptional but who excel in specific underperforming zones. For instance, certain players with high shot volumes but moderate efficiency in "Above the Break 3" zones had a significant predicted impact. This discovery alters our understanding of how to balance efficiency with shot attempts and emphasizes the need to focus not only on overall efficiency but also on zone-specific strengths.

Another unexpected result was the sensitivity of the Decision Tree model to variations in player data, which occasionally led to overfitting, despite tuning efforts. This sensitivity meant that certain player performances, which may have been outliers, had a larger influence on the model than anticipated. This highlights the importance of further refining the data used in the model, possibly by aggregating performance across more games or seasons to reduce the influence of anomalies. These findings suggest that while Decision Trees can offer critical insights, they should be used in conjunction with other models to capture the full scope of potential player impacts.

Predictive Model 3



The Random Forest Regressor is a robust and flexible machine learning algorithm used primarily for regression tasks. It works by constructing multiple decision trees during training and combining their results for a more accurate and generalized prediction. Each tree is trained on a random subset of the data, and the final output is the average prediction from all the trees. This approach helps reduce overfitting and increases the model's ability to handle high-dimensional datasets with complex relationships. Random Forests are particularly effective in capturing non-linear patterns in the data, making them well-suited for problems like predicting player performance in various NBA zones based on shooting efficiency. Additionally, Random Forests have an inherent feature importance mechanism that can provide insights into the most influential variables in the dataset.

In this project, the Random Forest model was applied to predict how efficiently players might perform if they joined the Atlanta Hawks, based on their historical performance in different shooting zones. We trained the model using a combination of player efficiency metrics and the number of shots attempted. The random subset sampling used by the model improves generalizability by avoiding reliance on individual trees that may capture noise or biases in the data. With its ability to capture non-linearities, this model is expected to offer a more nuanced prediction of player performance compared to linear models or single decision trees. The Random Forest's ensemble learning approach ensures it can handle outliers and missing data better than simpler models.

After training the Random Forest Regressor on the dataset, the model demonstrated superior performance with a Mean Squared Error (MSE) of 0.001664—the lowest among all three models tested. This low MSE indicates that the Random Forest model had the highest prediction accuracy in estimating how efficient players would be in various zones. A table summarizing the top 5 players based on their predicted efficiencies compared to the Hawks' current efficiency is provided below:

Player Name	Basic Zone	Zone Name	Actual Efficiency	Predicted Efficiency	Hawks Efficiency	Improvement
Tomas Satoransky	Above the Break 3	Center	0.333	0.646	0.360	0.286
Davis Bertans	Above the Break 3	Left Side	0.399	0.646	0.366	0.280
George Hill	Above the Break 3	Left Side	0.313	0.646	0.366	0.279
Franz Wagner	Above the Break 3	Left Side	0.386	0.646	0.366	0.279
Fred VanVleet	Above the Break 3	Left Side	0.372	0.646	0.366	0.279

As seen in the table, players like Tomas Satoransky and Fred VanVleet have the potential to significantly improve the Hawks' efficiency in specific shooting zones, based on the model's predictions. The improvement column shows the possible increase in team performance if these players were added to the roster, based on

The Random Forest model revealed several interesting insights into how different players could enhance the Hawks' offensive efficiency. For instance, players with a consistent shooting record, like Davis Bertans, were predicted to bring a substantial improvement in shooting efficiency in zones where the Hawks currently underperform. This suggests that certain players have the capability to fill the gaps in the team's current offensive strategy, particularly in high-impact zones such as the "Above the Break 3" and "Right Side" shooting zones. The model's ability to evaluate the players' performance over multiple seasons and account for zone-based efficiency offers a deeper understanding of how individual players can contribute to team success.

Another key insight from this model is its ability to highlight players who may not necessarily have high overall efficiency but excel in specific areas where the Hawks are weak. For example, players like Tomas Satoransky and George Hill, who may not be top performers across the board, were identified as potential game-changers in the "Above the Break 3" zone. This kind of granular analysis is invaluable for team managers when considering potential trades or signings, as it provides a more targeted approach to improving team performance. The model helps identify specific player roles rather than focusing on generalized metrics.

The insights generated by the Random Forest model have altered the scope of the project by emphasizing the importance of zone-specific performance rather than overall player efficiency. Initially, the project aimed to identify top-performing players based on overall shooting efficiency; however, the model's results suggest that targeting players who excel in specific zones could yield greater improvements for the team. This shift in focus has expanded the project's scope to consider a more detailed analysis of player efficiency across multiple shooting zones and not just broad player performance metrics.

Moreover, the Random Forest model highlighted the potential for certain under-the-radar players to make significant contributions in areas where the Hawks struggle. This discovery broadens the scope of potential player acquisitions and provides a new perspective on roster-building strategies. The model suggests that the Hawks should focus on acquiring role players who can address specific weaknesses in the team's shot selection rather than pursuing high-profile players who may not offer as much improvement in targeted areas. The project now includes a deeper evaluation of zone-based efficiency as a key factor in team-building decisions.

Predictive Model Review

After evaluating the three models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—it is clear that the Random Forest Regressor is the champion model. This model consistently produced the lowest Mean Squared Error (MSE) of 0.001664, indicating it has the highest predictive accuracy among the models. Its ability to handle complex, non-linear relationships and to avoid overfitting makes it a robust choice for this type of analysis. Additionally, the Random Forest model's ensemble approach, which aggregates multiple decision trees, provides a more stable prediction, especially when dealing with zone-based player performance data that can be prone to variability. The model’s performance in predicting player efficiency improvements for the

Hawks highlights its effectiveness in leveraging historical player performance to project future contributions.

In comparison, the Linear Regression model, while simpler and easier to interpret, had a higher MSE of 0.00252. This indicates that it did not capture the complexities and non-linear relationships in the dataset as effectively as the Random Forest. Linear regression assumes a straight-line relationship between the input features and the target variable, which likely led to oversimplified predictions in this case. Furthermore, it is more sensitive to outliers, which can skew the predictions, particularly in a dataset like this one, where different players have varying shot attempts across different zones. Although it offers quick results and is computationally less expensive, its performance is inferior in this scenario, making it less suitable for such nuanced player evaluations.

The Decision Tree Regressor performed better than Linear Regression with an MSE of 0.00221, but it still did not surpass the Random Forest model. Decision trees can model non-linear relationships and are easy to interpret; however, they are more prone to overfitting, especially when used alone without an ensemble method. This overfitting likely contributed to the slightly higher error rate compared to Random Forests. Additionally, while Decision Trees can be effective in simpler datasets, they struggle to generalize as well as Random Forests in larger, more complex datasets with a high degree of variability, such as player shot performance across different zones.

The Random Forest Regressor is the recommended champion model for this project due to its superior performance, versatility, and ability to generalize across complex, high-dimensional data. The model's ability to aggregate multiple decision trees ensures that it captures non-linear relationships and reduces the risk of overfitting, which is particularly important in sports tracking data where player performance can vary significantly across different zones and seasons. The Random Forest model's low MSE indicates that it provides the most accurate predictions of player efficiency, making it the best choice for evaluating potential player additions to improve the Hawks' performance. The feature importance mechanism within the Random Forest model also provides valuable insights into which variables—such as shot attempts and efficiencies in different zones—are the most influential in predicting performance, adding an extra layer of interpretability.

Another reason to select the Random Forest model is its flexibility in handling missing data and outliers, which are common in datasets involving player statistics. While Linear Regression and Decision Tree models are more sensitive to outliers or incomplete data, the Random Forest model's random sampling of subsets ensures that these issues do not significantly affect the overall accuracy of the predictions. This robustness is essential when making data-driven decisions about player acquisitions, where the goal is to minimize the risk of misinterpretation due to anomalies in the dataset. Overall, the

Random Forest Regressor provides the best balance between accuracy, interpretability, and flexibility, making it the ideal tool for this sports analytics task.

Final Results

Analysis Justification

The analysis performed in this project is crucial for understanding and addressing specific inefficiencies within the Atlanta Hawks' offensive strategies. By focusing on underperforming zones, we were able to identify areas where the team's shot efficiency fell below league averages. This identification is vital because it highlights clear opportunities for improvement, whether through changes in strategy, player development, or roster adjustments. Using predictive modeling techniques such as Linear Regression, Decision Tree, and Random Forest, we examined how potential player acquisitions could address these inefficiencies. This multi-model approach provided a comprehensive analysis, offering insights into both the accuracy and reliability of our predictions for player impact.

The use of multiple machine learning models allowed us to triangulate results, enhancing the robustness of our findings. For instance, while the Random Forest model showed the lowest mean squared error, we used the Linear Regression and Decision Tree models to ensure that our insights were consistent and did not overly rely on a single method. This cross-validation helped us determine which players would be most effective in boosting the Hawks' efficiency in critical zones, specifically the "Above the Break 3," "In The Paint (Non-RA)," "Restricted Area," and "Right Corner 3" zones. By predicting how new player acquisitions could improve team performance, the analysis provided actionable recommendations for future team-building strategies. This predictive approach ensures that any roster changes are data-driven, which is critical for long-term success in the competitive environment of the NBA.

The justification for this analysis lies in the strategic advantage it offers. Basketball is increasingly becoming a data-driven sport, with teams leveraging analytics to gain a competitive edge. Understanding shot distribution and efficiency patterns can inform better coaching decisions and optimize player rotations. Additionally, this analysis provides a framework for measuring the potential return on investment of acquiring specific players, which is essential for a team's financial and competitive health. By grounding our recommendations in predictive analytics and machine learning, we reduce the risks associated with subjective decision-making and align team strategies with objective performance metrics. This approach not only enhances the Hawks' potential for success but also sets a precedent for evidence-based decision-making in sports management.

Findings

The findings from our field goal analysis highlight several critical inefficiencies in the Atlanta Hawks' shot distribution and performance. Specifically, the team underperforms in four main zones: "Above the Break 3" at Right Side Center, "In The Paint (Non-RA)" at the Right Side, "Restricted Area" at Center, and "Right Corner 3" at Right Side. The efficiency gaps compared to league averages emphasize a need for strategic adjustments in these areas, either through player development or acquisition.

Our models identified key players who could improve the Hawks' performance in these zones. For example, Tomas Satoransky and Davis Bertans emerged as top targets for the "Above the Break 3" zones, with potential efficiency improvements of up to 0.286. In the "Restricted Area" and "Right Corner 3" zones, players like Bol Bol and Tobias Harris demonstrated high efficiency, making them valuable additions that could impact overall team effectiveness.

The Random Forest model, which had the lowest mean squared error, provided the most reliable predictions for player impact. According to this model, integrating top-performing players could significantly boost efficiency, narrowing or even eliminating the gap between the Hawks and league averages. This data-driven approach highlights tangible opportunities to optimize offensive performance, supported by predictive metrics that can guide future decisions.

BASIC_ZONE	ZONE_NAME	Efficiency_Hawks	Efficiency_League	Efficiency_Diff
Above the Break 3	Right Side Center	0.351575	0.356886	-0.005312
In The Paint (Non-RA)	Right Side	0.395455	0.436020	-0.040566
Restricted Area	Center	0.646144	0.653988	-0.007845
Right Corner 3	Right Side	0.373306	0.385153	-0.011847

PLAYER_NAME	BASIC_ZONE	ZONE_NAME	Efficiency	Predicted_Efficiency_Forest	Improvement_Forest
Tomas Satoransky	Above the Break 3	Center	0.333333	0.646144	0.286021

PLAYER_NAME	BASIC_ZONE	ZONE_NAME	Efficiency	Predicted_Efficiency_Forest	Improvement_Forest
Davis Bertans	Above the Break 3	Left Side Center	0.399061	0.646144	0.279208
George Hill	Above the Break 3	Left Side Center	0.313433	0.646144	0.279208
Franz Wagner	Above the Break 3	Left Side Center	0.386598	0.646144	0.279208
Fred VanVleet	Above the Break 3	Left Side Center	0.372007	0.646144	0.279208

Review of KPIs

Our first Key Performance Indicator (KPI) is Shot Distribution Shifts, which measures the percentage of shots taken from each court zone. Through our analysis, we found that the Atlanta Hawks have specific zones where their shot distribution is inefficient compared to optimal shot selection strategies. This KPI helps us understand if the Hawks are focusing too much on low-efficiency areas, such as "Above the Break 3" or "Right Corner 3." By adjusting strategies to promote higher-efficiency shots, the team could improve overall offensive performance. Tracking these shifts over time will be key to evaluating the effectiveness of any coaching adjustments or player development initiatives. If we observe a change in shot distribution toward more efficient zones, it would be an indicator of progress in aligning the team's strategy with data-driven insights.

The second KPI, Player Performance Metrics, focuses on comparing the Hawks' shooting percentages to league averages in each zone. Our analysis revealed that the Hawks are underperforming in critical areas, such as the "Right Side Center" zone in the paint and specific three-point zones. We used these metrics to pinpoint inefficiencies and identify opportunities for improvement. By comparing individual player performance against league standards, we were able to recommend key players who could elevate the team's shooting efficiency. These comparisons help us gauge the effectiveness of current players and the potential impact of new acquisitions. If the Hawks' shooting percentages start to close the gap with league averages, this would signify an improvement driven by strategic roster and training changes.

Team Potential Success is our third KPI, which measures the overall impact that new player acquisitions could have on the Hawks' performance. This KPI leverages our predictive models to forecast the efficiency improvements that players like Tomas Satoransky and Davis Bertans could bring. Using the Random Forest model, which delivered the most accurate predictions, we identified players who have the potential to transform the Hawks' offensive capabilities. The projected efficiency gains in critical zones suggest that strategic player acquisitions could significantly elevate the team's success. This KPI will be used to track the real-world impact of these changes in future seasons. Observing improvements in game outcomes, offensive ratings, and overall team efficiency will help validate our recommendations and demonstrate the value of our analysis.

Our final KPI, Efficiency Improvement Potential, focuses on the projected increase in team efficiency with new player additions. This measure uses the predictive power of our models to estimate how much the Hawks could improve if they acquire top-performing players identified in the analysis. For example, our model predicts potential efficiency gains of up to 0.286 in underperforming zones, which could lead to substantial improvements in scoring. By tracking this KPI, we can evaluate whether the expected gains materialize once new players are integrated into the lineup. This metric also helps assess the return on investment for player acquisitions and informs future decisions about player development and roster adjustments. If the efficiency improvements align with our projections, it would confirm the relevance and accuracy of our data-driven approach.

Review Significance

The significance of this project lies in its potential to transform the Atlanta Hawks' offensive strategy through data-driven insights. Basketball, like many sports, has become increasingly data-centric, and understanding key inefficiencies can provide a competitive edge. By analyzing shot distribution and player performance across different court zones, we can identify strategic adjustments that could improve overall team efficiency. Our findings pinpoint specific zones, such as "Above the Break 3" and the "Restricted Area," where the Hawks underperform compared to the league average. This information is not just theoretical but actionable, allowing the coaching staff and front office to make informed decisions about player development and acquisition. Thus, the project holds considerable significance in enhancing the team's on-court performance and aligning strategies with modern analytics.

The findings are particularly important because they reveal clear opportunities for improvement and targeted player recruitment. For example, we identified players like Tomas Satoransky and Davis Bertans, who have high efficiency rates in underperforming zones where the Hawks struggle. Adding these players to the roster could immediately impact the team's offensive output and help close the efficiency gap compared to other NBA teams. Our analysis demonstrates that even minor adjustments in player lineup or shot selection strategy can yield substantial benefits. By focusing on efficiency improvement, the Hawks can optimize their scoring opportunities and become more competitive in close games. This not only increases the team's chances of winning but also provides a more efficient and exciting style of play for fans.

Furthermore, the broader significance of this project lies in its contribution to the field of sports analytics. As teams increasingly rely on data to make strategic decisions, our approach serves as a blueprint for effective analysis. We have demonstrated how predictive modeling and efficiency metrics can be used to evaluate player impact and forecast potential improvements. This has implications beyond the Hawks, as other teams and sports organizations could adopt similar methodologies. The findings also highlight the importance of ongoing data monitoring and model refinement to keep up with evolving team dynamics and league trends. Ultimately, our project underscores the value of data analytics in shaping the future of sports strategy and player development.

Recommendations for Future Analysis

One area for future analysis is a deeper dive into the defensive impact of potential player acquisitions. While our project primarily focused on offensive efficiency and shot performance, defense plays an equally critical role in the overall success of the Atlanta Hawks. Future studies could explore metrics like defensive rating, player impact on opponent shooting percentages, and how potential acquisitions could improve the team's defensive structure. This would provide a more holistic understanding of player contributions and allow the Hawks to address both ends of the floor simultaneously. Analyzing defensive performance through tracking data and advanced statistics like contested shot rates or on-ball defensive effectiveness could uncover new opportunities for improvement and optimize the team’s overall game strategy.

Another valuable area for future exploration is the integration of injury risk analysis into player acquisition decisions. Basketball is a physically demanding sport, and player injuries can drastically affect team performance. By incorporating data on player durability, minutes played, and historical injury patterns, the Hawks' management could make more informed decisions regarding potential acquisitions. Predictive models that estimate the likelihood of injury based on player workload and biomechanics could be developed to minimize the risk associated with high-impact player trades or signings. Furthermore, this kind of analysis would be especially beneficial in developing player load management strategies, ensuring that key players remain healthy and effective throughout the season. This preventive approach could lead to long-term success and sustainability for the team.

Lastly, future analysis should consider the psychological and cultural fit of potential players within the existing Hawks roster. Team chemistry and player dynamics often influence on-court performance but are difficult to quantify through traditional data metrics. However, studies that incorporate qualitative data, such as player interviews or assessments of leadership and teamwork, could provide valuable insights. Machine learning models that analyze text or sentiment data from social media and sports articles could be used to gauge public and team perceptions of player fit. Additionally, exploring the impact of coaching strategies and team culture on player performance would provide a more comprehensive analysis. This multidimensional approach would ensure that the Hawks not only improve their statistical performance but also maintain a cohesive and motivated team environment, ultimately enhancing the potential for long-term success.

References

Al-Khassaweneh, M., & Hammad, A. (2020). Data visualization using Matplotlib and Seaborn: A case study of data science and machine learning applications. *International Journal of Advanced Computer Science and Applications, 11*(4), 23-30.

Baca, A., & Perl, J. (2018). *Modelling and simulation in sport and exercise* (1st ed.). Routledge. https://doi.org/10.4324/9781315163291

Bunker, R., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15*(1), 27-33.

Daly-Grafstein, D., & Bornn, L. (2021). Using in-game shot trajectories to better understand defensive impact in the NBA. *Journal of Sports Analytics*, 6(4), 235-242. https://doi.org/10.3233/JSA-200400

Davenport, T.H. (2014). *Analytics in sports: the new science of winning*. *International Institute for Analytics*.

Grivet, S., Auber, D., Domenger, J-P., & Melançon, G. (2006). *Bubble tree drawing algorithm*. In *Computer Vision and Graphics: Computational Imaging and Vision* (Vol. 32, pp. 633-641). Springer.

Thakur, S., & Karthik, R. (2022). End of game shot selection for individual players in the NBA. *Materials Today: Proceedings*, 62, 4643-4650. https://doi.org/10.1016/j.matpr.2022.03.119

Ram, J., & Corkindale, D. (2014). How "critical" are the critical success factors (CSFs)? Examining the role of CSFs for ERP. *Business Process Management Journal, 20*(1), 151-174. https://doi.org/10.1108/BPMJ-11-2012-0127

Nelli, F. (2023). *Data Visualization with Matplotlib and Seaborn*. In: Python Data Analytics. Apress, Berkeley, CA. https://link.springer.com/chapter/10.1007/978-1-4842-9532-8_7.

Oberoi, A., & Chauhan, R. (2019). *Visualizing data using Matplotlib and Seaborn libraries in Python for data science*. *International Journal of Scientific and Research Publications, 9*(3), 8733. https://www.ijsrp.org/research-paper-0319.php?rp=P878342.

Parmenter, D. (2015). *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs* (3rd ed.). John Wiley & Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Trawinski, B. (2016). Application of logistic regression in sports prediction: A case study. *Journal of Sports Analytics, 2*(4), 155-168.

Waskom, M. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021.

Zhang, Z., McInnes, M. D. F., & Schopflocher, D. P. (2017). Heatmap visualization of multi-dimensional clinical trial data. *Journal of Clinical Epidemiology*, *87*, 100-110.