Refugio Zepeda , Dmitry Sorokin, & Henry Yost
March 15, 2025

NBA-2024 Data Analysis

Our group analyzed the 2023-2024 NBA Player Statistics, Per 36 minutes. Within sports players/teams, statistics play a huge role in pattern recognition and finding the most optimized way to be competitive against a team. With technological advancements there also seems to be more kinds of data to analyze meaning statistical analysis is becoming more prominent. Our analysis will examine player positions, field goal percentages, 3-point field goal percentages, free throws attempted, and minutes players. In context within our analysis, the benefits of looking at our variables are how good an estimate is, determining linear relations between multiple variables, and predicting certain scenarios such as the number of 3-point attempts.

Methodology

We will start with exploratory data analysis, where we will examine our data and explore the relationship between various variables. Some of the variables we will investigate include the distribution of age, the relationship between age and 3-P attempts, the relationship between age and team, and the relationship between position and personal fouls. We will use a variety of plots, including box plots, histograms, and scatter plots.

Afterward, our team will use confidence intervals and T-tests, multi-linear regression, and logistic regression to analyze our NBA dataset. To run our tests, we will use the MASS and ggplot2 libraries in addition to our inference function, which we included. The MASS library will be used for backward elimination for our second analysis and the ggplot2 library will be used for graphs for our third analysis.

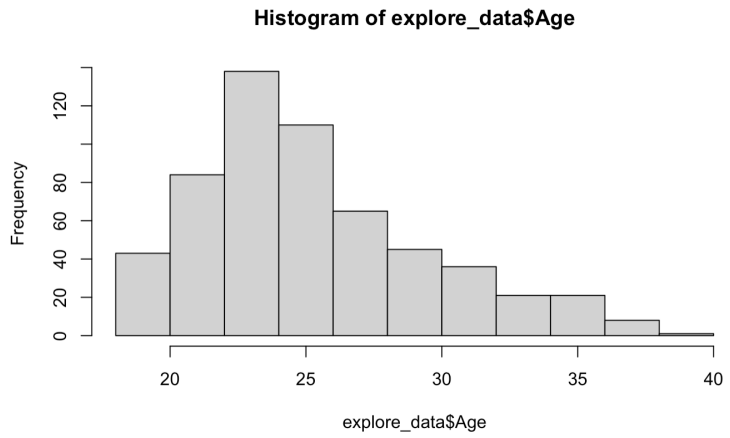
One of the benefits of our methods is that we explore multiple variables beyond arbitrary EDA. We also test several hypotheses and analyze the results for a full understanding of our data.

Exploratory Data Analysis

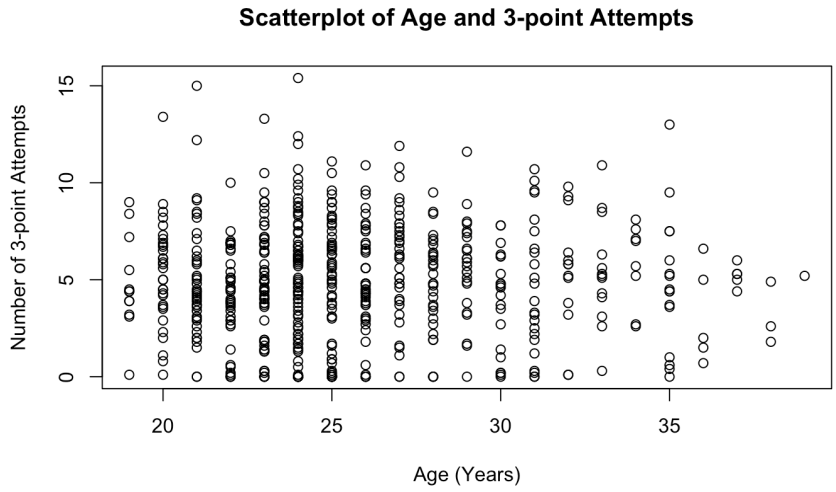
The data is from the *Basketball-reference*, "2023-24 NBA Player Stats: Per 36 Minutes" dataset. The dataset contains per-36-minute statistics for NBA players, including various metrics and statistics regarding player performance. After a brief introduction to the dataset, we will analyze the distribution of age, the relationship between age and three-point attempts, the correlation between age and team, and the connection between position and personal fouls.

	Age <int>	Pos <fctr>	X3PA <dbl>	Tm <chr>	PF <dbl>
1	24	PF-C	2.2	TOT	3.2
2	26	C	0.6	MIA	2.4
3	23	SG	4.6	TOT	2.6
4	23	PF	6.8	MEM	2.0
5	25	SG	6.3	MIN	2.7
6	28	SG	6.4	PHO	2.2

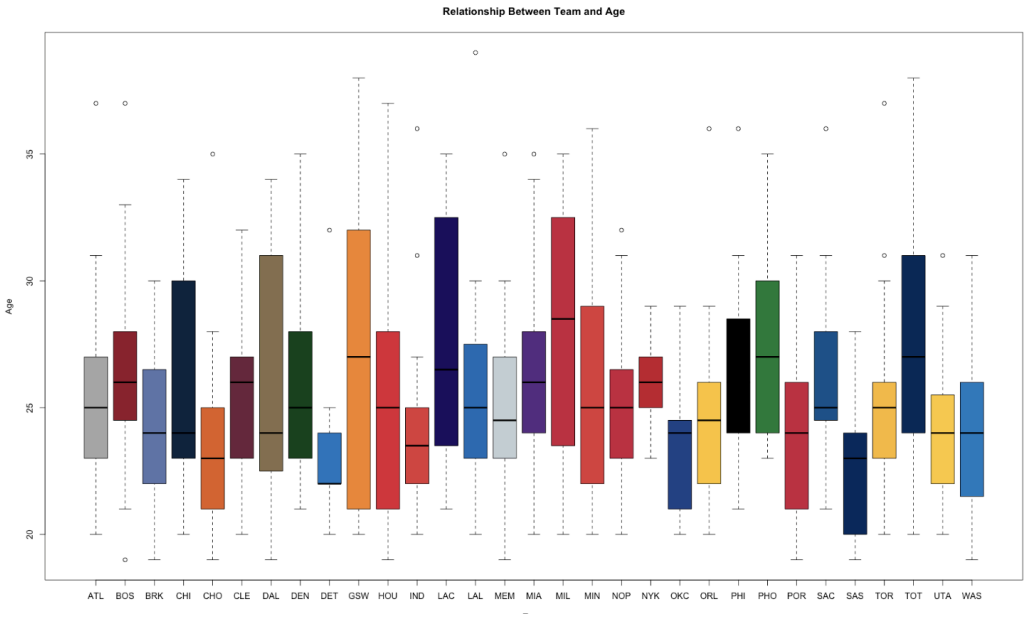
The first six rows of the dataset provide a brief introduction of the key variables, including the player's age (Age), position (Pos), 3-point attempts (X3PA), team (TM), and personal fouls (PF). These variables can provide a relatively strong analysis of player demographics and performance, which can be further explored in the first explanatory analysis–focusing on the distribution of age within the dataset, shown graphically below.



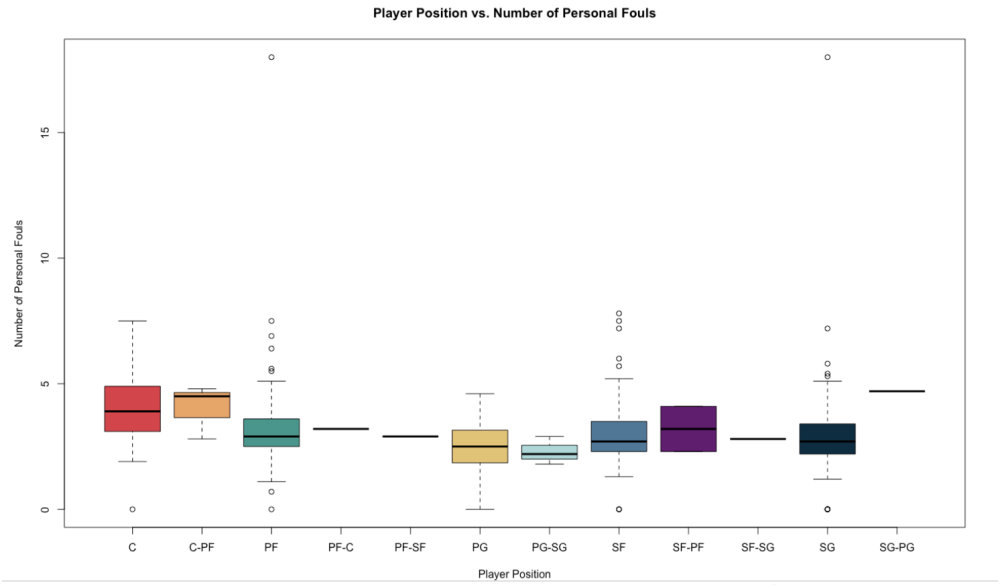
The age distribution is right-skewed, which is expected in professional sports such as Basketball. Older players are less common due to the increased physical demand of the sport and lengthier recovery time. The human body is in a better state to play sports at a physically intensive and competitive level at a younger age, and that, as expected, is reflected in the distribution. Moreover, interestingly, there is an insignificant relationship between age and 3-point attempts.



The scatter plot does not indicate a linear relationship between age and 3-point attempts, however, there are still valuable insights that can be drawn. Similar to the age distribution, the highest concentration of players is around the 23-25 age group, as shown by the density of the data points in the graph. Additionally, the scatterplot shows younger players are more likely to attempt 3-pointers. Not only is this proven by the higher density of points by the younger players, however, also by the decline in 3-point attempts starting at the age of 30, and significantly declining at age 35.

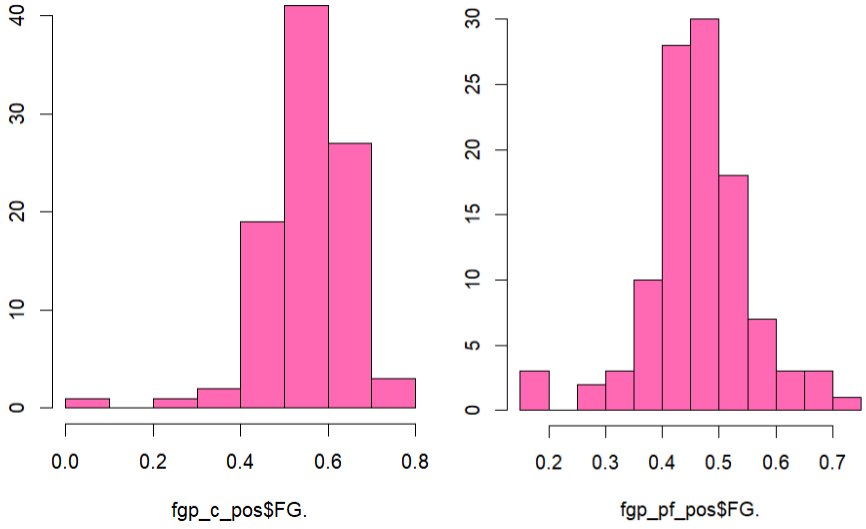


The box plot (depicted above), shows a significant variation in the age ranges across the NBA teams in the 23-24 dataset. This is contrary to the expectation of teams having narrower age ranges, given younger players often outperform older players, the data shows a broader distribution. Notably, the New York Knicks (NYK) have the smallest interquartile range (IQR) in terms of player age, while the Golden State Warriors (GSW) have the largest IQR.



Lastly, the relationship between a player's position and the number of personal fouls is depicted in the box plot above, which additionally, builds on an analysis conducted later in the paper. The position with the most number of personal fouls is a point forward (PF), with most positions having a median number of personal fouls at or under five. Moreover, some players play multiple positions and have not received any personal fouls.

When graphically representing the field goal percentage of a center vs the field goal percentage of a power forward for each data point in the subsets there is a clear shift in their means. For the field goal percentage of pure center, we get the highest frequency between 0.5 and 0.6 of more than 40 of the data points in this range while the highest frequency for a pure power forward is between 0.45 and 0.5 with more than 30 data points similarly, though they are both approximately bell-shaped with a few outliers that aren't close to the mean.



Confidence Interval Analysis

The first analysis performed on the dataset was a confidence interval to look at a center's field goal percentage, and a pure power forward's field goal percentage, then compare the difference between the two means to see if there is a statistical difference by using a t-test. The

first step we performed for this analysis was to create a subset of the data only to include the player positions and field goal percentage. In addition, we created two more subsets one for only the field goal percentage of a center and another subset of only a power forward field goal percentage. When looking at the histogram of our subsets they both appear to be approximately normal with only a few sets of outliers around 0% field goal percentage. The confidence interval for the field goal percentage of 94 pure centers is (0.5332, 0.5754). Within the context of our problem, we are 95% confident a center will be shooting between 53.32% and 57.54%. In addition to the confidence interval, the concluded mean is 0.5543 with a standard deviation of 0.1042. The confidence interval for the field goal percentage of a pure power forward is (0.4462, 0.4805) meaning we are 95% confident that the mean-field goal percentage will be between 44.62% and 48.05%. Now for the last part of the first analysis, we did a two-sample t-test letting us know that there is a statistical significance between the means of a center and a power forward's field goal percentage. Additionally, we are 95% confident that the mean difference in field goal percentage between these two positions will be between 6.36% and 11.82%.

Multilinear Regression Analysis

For our second analysis, we performed multi-linear regression to predict the number of 3-point attempts. Our original explanatory variables for this analysis were rank (RK), position (Pos), age (Age), minutes played (MP), points (PTS), offensive rebounds (ORB), and defensive rebounds (DRB). We chose these variables because we believed these would be good predictors to predict our response variable, 3-point attempts (3PA). However, we optimized our model using backward elimination and were able to remove some variables. After backward elimination, we ended with a model that only used position, minutes played, points, offensive

rebounds, and defensive rebounds, meaning that through backward elimination, we got rid of unnecessary variables rank and age.

After backward elimination, we also saw an increase in our adjusted R^2. Our original model explained around 47.21% of the variance in 3-point attempts, and we saw an increase to around 47.28%, which is less than we hoped, but still an improvement. In our improved model, we also noticed that not all of our explanatory variables were significant to our model. The variables that did not have a significant p-value of less than 0.05 were within our position variable, which is categorical. Our base variable for position was the Center position and all the other positions were compared with respect to the Center position. The insignificant positions, in respect to our base position, were Center - Power Forward, Power Forward - Center, Power Forward - Small Forward, Small Forward - Power Forward, Point Guard - Shooting Guard, and Shooting Guard - Point Guard. All other positions and variables had p-values less than or equal to 0.05.

In addition, we also predicted the 3-point attempts of five new observations that were randomly created with educated intuition. To check our error, we used the Sum of Square Residuals (RSS) and got a relatively low value of around 31.1. This indicated to us that our model was relatively good at predicting the 3-point attempts of players. It also showed that our initial variable selection was a reasonable selection that improved from our backward elimination. However, there is room for improvement, as we can likely improve our model by including all of the variables and then comparing both forward selection and backward elimination.

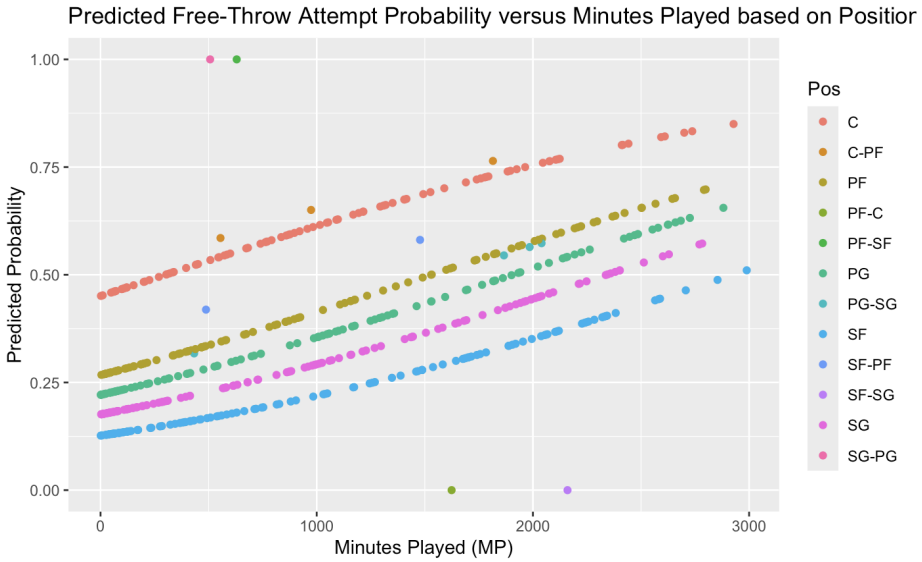
Logistical Regression Analysis

The third analysis performed on the dataset was a logistical regression to determine the highest fouled position. The response variable is free-throw attempts (FTA, numerical), and the explanatory variables are position (POS, categorical) and minutes played (MP, numerical) respectively. The logistical relationship is significant because it allows us to predict the probability of a free-throw attempt based on a player's position and minutes played. Since free-throw attempts are directly linked to how often a player is fouled, this insight helps us understand the impact of play time on foul frequency across the positions in Basketball. In regards to the logistical regression, the null hypothesis is H_0: *A player's position has no effect on the probability of a free-throw attempt*, and the alternative hypothesis is H_1: *A player's position has an effect on the probability of a free-throw attempt*.

It is important to address two key considerations in the model. First, free-throw attempts include both technical and personal fouls, however, the difference is, that personal foul free-throws must be taken by the fouled player (UC Berkeley Basketball Rules, n.d.). Because personal fouls occur far more frequently, the influence of technical fouls is nearly negligible in this analysis. Second, free-throw attempts were transformed into a binary variable, where players with three or more attempts were assigned a value of 1, while those with fewer than three were assigned a value of 0. After generating the model, and calling the respective summary() function, the significant variables concerning the baseline center position include Pos PF, Pos PG, Pos SF, pos SG-PG, and MP, with their respective p-values of 0.006385, 0.000494, 2.75×10^{-8}, 5.02×10^{-6}, and 2.83×10^{-9}. Thus, we can reject the null hypothesis, as there is sufficient evidence to suggest a strong relationship between a player's position and a free-throw attempt.

By calculating the log odds ratio, we can determine how the odds of a free-throw attempt change with a one-unit increase in minutes played. In this case, the log-odds ratio of 1.00066 is

positive, suggesting that each additional minute played increases the odds of attempting a free throw by 0.66%. The following finding is supported both numerically and visually in the graphic representation shown below.



As shown in the graph "Predicted Free-Throw Attempt Probability versus Minutes Played Based on Position", all player positions have a positive relationship with minutes played, a trend that aligns with the previously analyzed log-odds ratio. For the purpose of this analysis, positions with limited sample representation will be excluded. Interestingly, the center (C) position has the highest predicted probability of attempting a free throw among all positions. This outcome is consistent with real-world basketball, as centers are typically positioned near the basket, where they are more likely to be fouled. Following center, the positions with the highest probability of free-throw attempts are point forward (PF), point guard (PG), shooting guard (SG), and shooting forward (SF).

Center positions attempt the most free throws, which directly corresponds to their higher number of received fouls. This can be seen when watching basketball games, as a more aggressive and physical position on the court will engage in significantly more physical contact

than other perimeter positions. Lastly, we could reject the null hypothesis, given the sufficient evidence suggesting a strong relationship between a player's position and free-throw attempts. Additionally, there is a strong, statistically significant relationship between minutes played, and the number of free-throw attempts. Logically, this makes sense, as the longer an individual is on the court, the higher the odds of getting fouled are.

Conclusion

Through the analysis of the 23-24 NBA Player Statistics dataset, we have drawn several interesting insights about player performance, team demographics, and statistical relationships. The preliminary exploratory data analysis looked into player age, position, and performance metrics, concluding that younger players are significantly more likely to attempt a 3-pointer, and broader age demographics across NBA teams. The confidence interval and t-test analysis confirmed that centers have a significantly higher field goal percentage than power forwards. Additionally, the multilinear regression model identified important factors that influenced three-point attempts, and lastly, the logistic regression analysis demonstrated a significant relationship between a player’s position and free-throw attempts.

These findings influence the stakeholders, specifically, NBA coaches, and sports bettors, because they can use the insights from the data to increase the effectiveness of their team with a stronger lineup. Sports bettors, an emerging form of gambling can utilize the conclusions drawn from the dataset, to improve their predictions about player and game performance, in turn winning a larger profit.

However, despite the overall strengths of our analysis, there are limitations. First, we did not clean the dataset, before creating the models and tests. The dataset includes outliers that could have influenced and skewed the results, albeit not significantly. Second, the models rely on

statistical data, and external factors such as strategies, playing styles, pace, etc… are not included in the dataset. Including these metrics in the model would help create a more well-rounded and complex model that could predict more accurately. Our analysis provides a significant and reliable foundation for understanding variables and their relationships in the NBA. With more refinement of the data and improvement in the size dataset–future predictions could be significantly more accurate.

References

2023-24 NBA Player Stats: Per 36 Minutes, Basketball Reference, www.basketball-reference.com/leagues/NBA_2024_per_minute.html

Basketball Rules, UC Berkeley Recreation & Wellbeing, https://recwell.berkeley.edu/competitive-programs/intramural-sports/rules-policies/basketball-rules/

Zepeda, Refugio (2025), *Lab #3: Interference for Numerical Data*, [R Markdown metadata], https://posit.cloud/content/9995467