1. Chicago Bulls Team Evaluation

To understand the Bulls in relation to other teams in the league beyond wins or points, the four factor model can glean insights on a team's strengths and weaknesses on the court. The four factors are effective field goal percentage, turnover percentage, rebounding percentage, and free throw percentage. These metrics speak to different areas of strength in Basketball, and as such do not correlate with one another (Winston, Nestler, and Pelechrinis 2022).

Data pertaining to the four factors was extracted from Basketball Reference for the 2023-24 season. The dataset includes offensive and defensive measures for each factor, as well as a row for the league average. The only change made to this dataset was to rename the four factor columns and input a value for the league average of wins.

Winston (2022) found that shooting percentage has the strongest relationship with team wins. An analysis of the 2024 season confirms this conclusion. Each of four factors can be broken down as the defensive metric subtracted from the offensive one. As shown in Figure 1, only the shooting percentage factor has explanatory power for the number of wins between the factors.

Figure 1: Four Factor Win Correlation

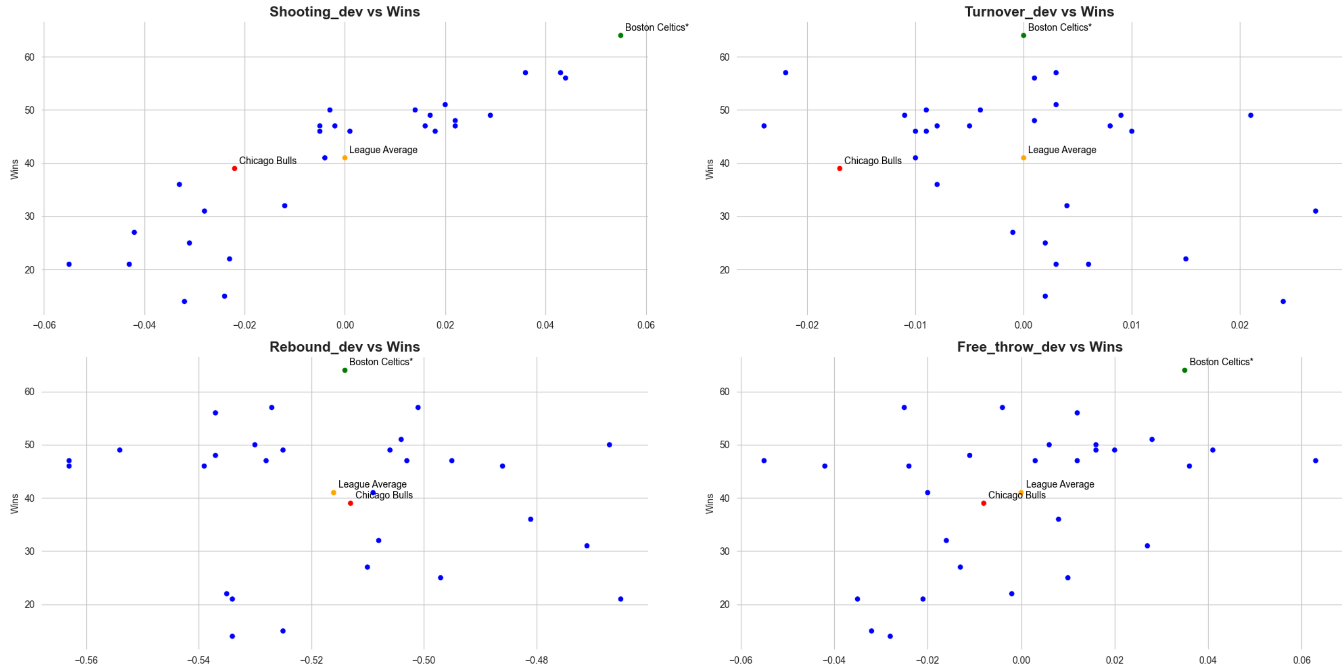

Figure 1: The four factors each correlated with a team's wins, with annotations for the Chicago Bulls, Boston Celtics, and the League Average.
https://www.basketball-reference.com/leagues/NBA_2024.html

Last season the Bulls underperformed with respect to the league average in shooting, but were near the top of the league in turnovers. This indicates that the team was able to create plenty of opportunities to score, but didn't follow through enough. In rebounding and free throw percentage, the Bulls are very close to average.

The Celtics on the other hand led the league in effective field goal percentage and in the top five for free throws, as shown in Figure 2. Even with an average performance on turnovers and rebounds, they were able to win the most games in 2024 and eventually win the title.

Figure 2: Four Factor Team Rankings

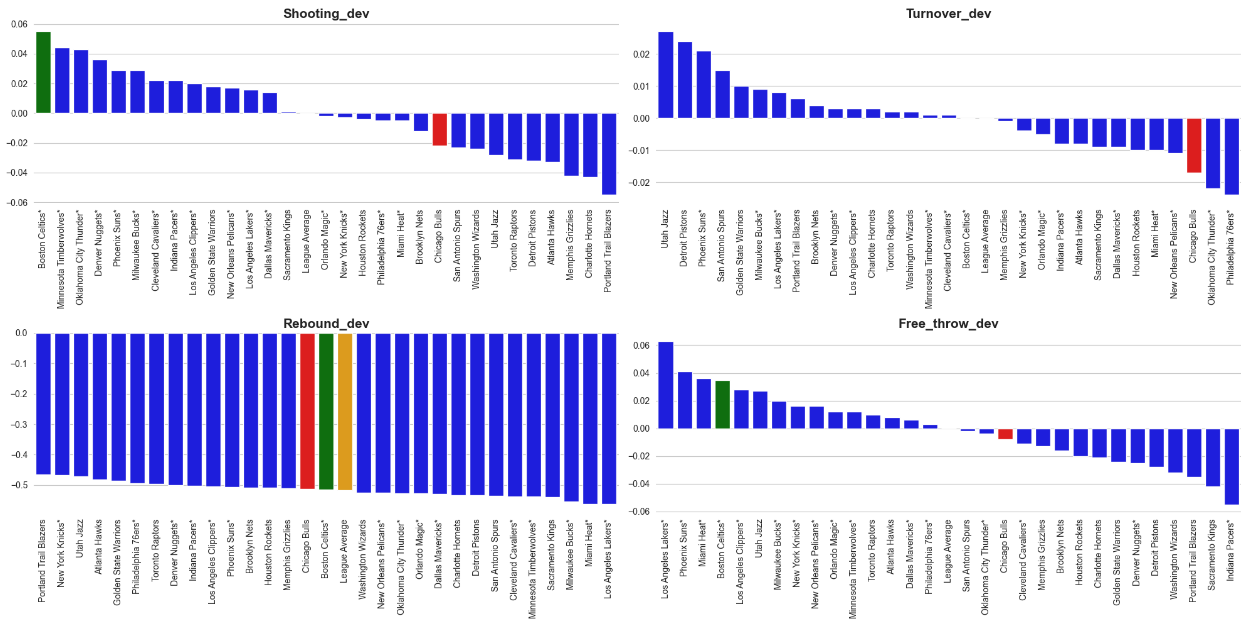

Figure 2: Each team's relative ranking within each of the four factors, with annotations for the Chicago Bulls, Boston Celtics, and the League Average.
https://www.basketball-reference.com/leagues/NBA_2024.html

To further investigate the impact of the four factors on win percentage, an ordinary least squares regression model can bring the variables together and make a prediction on the number of games won. After fitting a model on 2024 data, the R-squared is .918, indicating a reasonable line of fit (Appendix 1). Shooting percentage explains 79% of the variance in win totals, slightly higher than the 76% from Winston (2022).

With this line of fit, each team can be assigned a number of predicted wins for the season. This model correctly predicted that the Bulls would win 39 games in 2024, but the error rates are higher for teams that have a very high or very low number of wins (Appendix 2). The root mean squared error for this model is 11.94, and this evaluation metric is in the units of the target variable, which means that the model is off by an average of 12 wins.

2. Player Classification Framework

Player positions in Basketball are more fluid than in other sports, and previous work has sought to classify players under a new schema. Cheng (2017) used a combination of KMeans clustering and Linear Discriminant Analysis to suggest eight positions in Basketball, while Man (2017) performs a similar analysis comparing clustering methods to suggest 12 new positions. This paper zeros in on KMeans clustering and seeks to find the best version of the model to understand NBA player positions.

The challenge in using a clustering algorithm is the inherent complexity of its method, which makes any results difficult to interpret. Regardless, there remains a need to reframe positions in basketball that more accurately describes a player's utility on the court.

Player data was extracted from Basketball Reference for the past 10 NBA seasons. Specifically, advanced metrics, stats per 100 possessions, and shooting variables were aggregated together. To handle outliers, any player-team-season combination below the median number of minutes played for a season was excluded from the dataset. Across the compiled data, that median value is 484 minutes.

There are trade offs when aggregating this dataset. There's a need to understand each player holistically and individually to find their purpose on their court and find what position they could fill. However, a simple average could lose some context in the development of a player over time. To resolve this, each season for each player was weighted by multiplying weights derived from the minutes played in each season.

$$Weights = MP_{Season} / MP_{Career}$$

Seasons with more minutes played are weighted heavier. Each record then has a weight value that can be multiplied by each numeric column to produce a new weighted statistic. The final aggregation step is to group by the player and take the average of each weighted statistic. With a solid average statistic for each unique player, it's now possible to test clustering models.

. Features related to a player's impact on wins or the value over replacement were excluded because they might conflate with more direct on-court statistics and don't provide much insight into the purpose of a player on the court or the description of a position.

Remaining features were divided into four categories, all features, shooting features, non-shooting features, and non-attempts (Appendix 3). Only KMeans clustering models were evaluated, so the combination of four feature sets and between 5 and 10 clusters results in 20 models to evaluate. The process to fit each model started by converting each weighted average to the standard scale, using z score normalization to even out the distributions of each feature. This is necessary because the next step, Principal Component Analysis, is sensitive to the scale of features. For each cluster and feature set combination, the fields were reduced to five components and scored based on the cohesion of each cluster.

There are four evaluation metrics that seek to explain how well the clustering is happening. The Silhouette Score measures how well each data point fits within each cluster, while the Calinski-Harabasz Index measures the dispersion between clusters. A higher score indicates better performance in assigning clusters.

Inertia and the Davies-Bouldin Index on the other hand seek to be minimized. Inertia is looking at how close each datapoint is to the center of each cluster, while Davies-Bouldin is the ratio of both within-cluster and between-cluster distances. As shown in Figure 3, one model received the best score for both Silhouette and Calinski-Harabasz, which influenced the decision to proceed with 5 clusters and the non-attempt feature set.

Figure 3: Best Model Cluster Diagrams

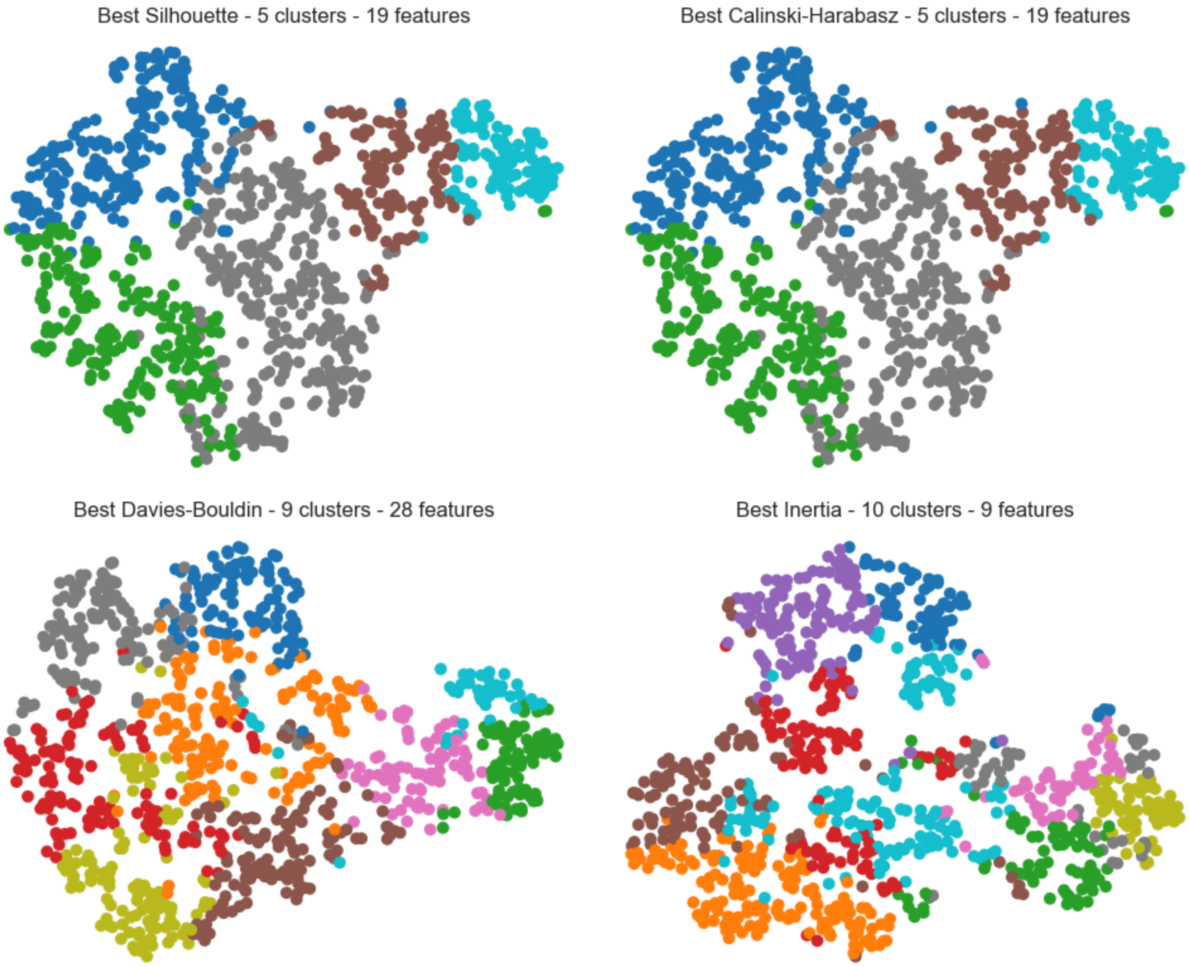

Figure 3: TSNE plot for best cluster model under each evaluation metric.
https://www.basketball-reference.com/leagues/NBA_2024_per_poss.html

Once the model is selected and the clusters are defined, the qualitative aspects of each cluster can be uncovered. Using the same PCA process from the model fitting step, the cluster centers can be inversely transformed to the original scaled features. This can give the first glimpse at what one cohort is doing better than another. Another visual tool to inspect cluster differences is the radar chart, which shows each cluster and the metrics that define them, as shown in Figure 4.

Figure 4: Radar Charts per Cluster

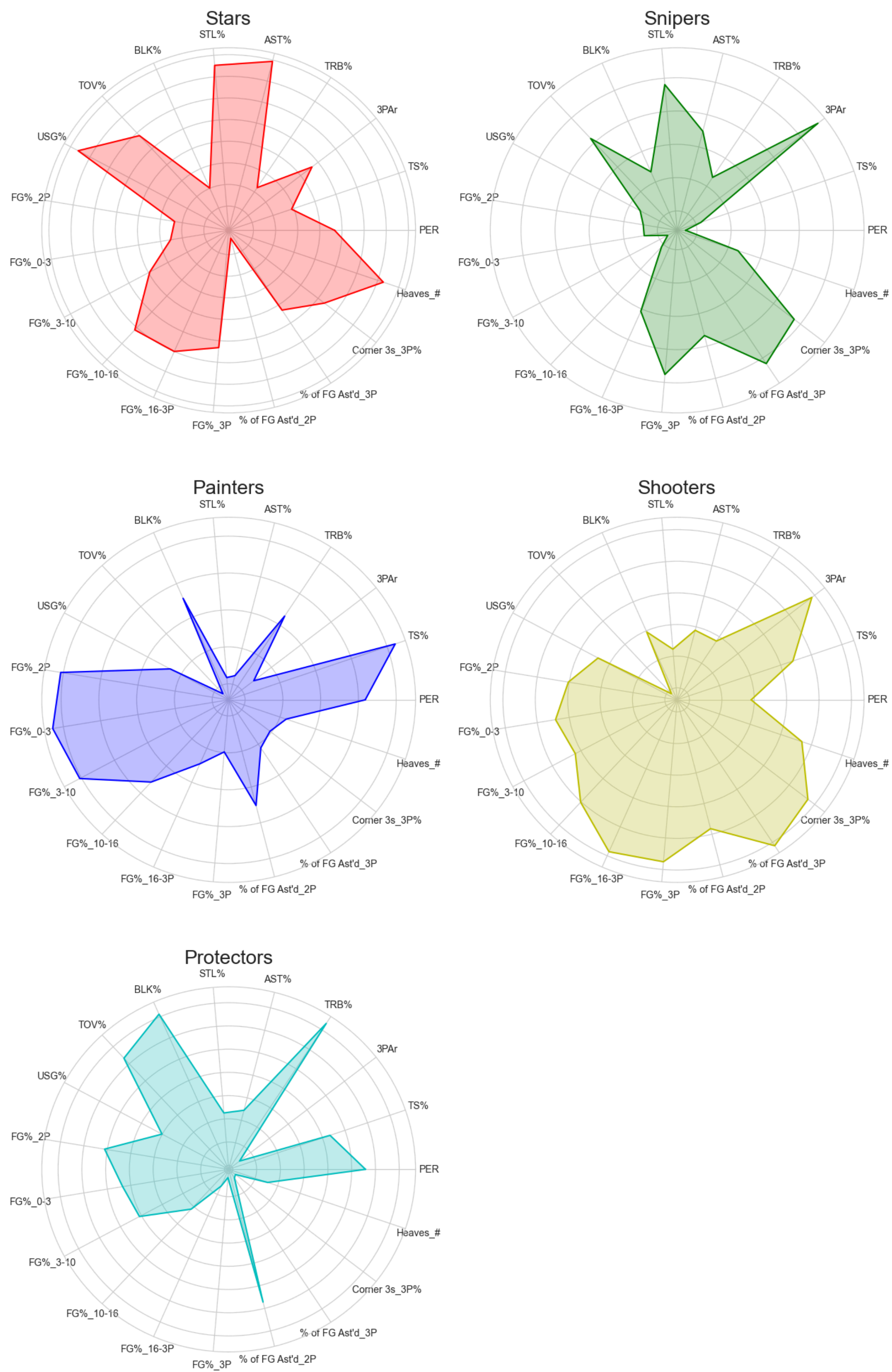

Figure 4: Radar charts for feature importance in each cluster.
https://www.basketball-reference.com/leagues/NBA_2024_per_poss.html

From this analysis, five positions emerge:

- Stars
    - High usage and assist rate
    - Efficient shooting from mid-range and 3-point range
    - High scorers with above average efficiency and shooting percentage
    - Notable players: Stephen Curry, LeBron James
- Snipers
    - Very high % of shots attempted in the 3-point range
    - High accuracy in the upper range, corner 3s
    - High volume scoring, below average efficiency
    - Notable Players: Brandon Boston Jr. Alec Burks
- Painters
    - Very high efficiency and shooting percentage
    - Most shots in lower-range, i.e. the paint
    - Strong rebounding and blocking
    - Notable Players: Joel Embiid, Anthony Davis
- Shooters
    - High shooting accuracy at all ranges
    - Decent efficiency, but mostly specializing on the outer range
    - Notable Players: Klay Thompson, Jaylen Brown
- Protectors
    - Very high rebounding and blocking
    - Very low amount of 3 point attempts, low percentage of shots
    - Notable Players: Hassan Whiteside, Andre Drummond

3. Recommendations to Bulls Management

After the clusters are defined, they can be applied to the original player dataset to isolate Bulls players in the 2024 season. Before filtering to the Bulls specifically, the dataset was updated to include roster changes in the offseason up to this point (NBA, 2024). This means dropping Alex Caruso, DeMar DeRozan, and Andre Drummond, and adding Chris Duarte, Josh Giddey, and Jalen Smith.

From looking at Bulls players with more than the median amount of minutes played in 2024, each cluster is represented except one, the protectors. An Andre Drummond-sized gap remains in the lineup, but it is possible to fill.

After extracting the available free agents (Spotrac 2024), we can match players to the dataset with cluster designations to identify which players could fill the protector role. This filter results in three available players: Omer Yurtseven, Chimezie Metu, and Damian Jones.

This analysis recommends the Bulls acquire a defensive specialist, with a keen ability to protect the rim. Out of available free agents in the “Protectors” cluster, Omer Yurtseven is the best option. The impact Drummond had for the Bulls was in rebounds, so it makes most sense to seek the player with the highest rebounding percentage. Using the weighted rebounding average discussed in section 2, Omer Yurtseven with an average of 22.5 is very close to Drummond’s 25.8 (Appendix 4).

4. Limitations and Future Work

While KMeans Clustering was successful in finding patterns between NBA players, it is a relatively simple model. Chang and Man offer additional clustering approaches that are worth exploring. While other perspectives went broader in the application of models, features, and dimensions, this paper is relatively narrow, opting to go deeper into finding the best KMeans model.

Given the relative simplicity of the model used, there is a risk of player misclassification. Any potential misclassification is difficult to validate. The evaluation metrics provide an ability to evaluate two models side-by-side, but this is not the same as an understanding of accuracy.

Another limitation based on this approach is repeatability. Each time the clustering model runs, it produces similar, but not quite identical results. This magnifies the potential for errors on the margins of each cluster, but players closer to the center of each cluster can be expected to be reasonably consistent.

Lastly, the analysis explicitly removes players with under the median number of minutes. This model cannot be used as a means to identify up-and-coming players who might show greater potential with a greater number of minutes played.

Future work should seek to evaluate multiple clustering methods to find the best classification of players. There could also be a way to identify the minimum number of minutes needed to be classified under a position. This evaluation should be geared towards not just roster construction, but toward lineup optimization. Is there an ideal representation of clusters on the court that yields the highest expected point differential?

References

"2023-24 NBA Season Summary." Basketball Reference, n.d. https://www.basketball-reference.com/leagues/NBA_2024.html.

Cheng, Alex. "Using Machine Learning to Find the 8 Types of Players in the NBA." Medium, March 2, 2017. https://medium.com/fastbreak-data/classifying-the-modern-nba-player-with-machine-learning-539da03bb824.

Man, Han. "Defining Modern NBA Player Positions - Applying Machine Learning to Uncover Functional Roles in Basketball." Medium, April 17, 2017. https://medium.com/hanman/the-evolution-of-nba-player-positions-using-unsupervised-clustering-to-uncover-functional-roles-a1d07089935c.

NBA. "2024 NBA Offseason: Every free agency deal, extension & trade for all 30 teams." NBA.com, last modified August 9, 2024. https://www.nba.com/news/nba-offseason-every-deal-2024.

"Player Per 100 Poss." Basketball Reference, n.d. (Seasons 2014-15 to 2023-24) https://www.basketball-reference.com/leagues/NBA_2024_per_poss.html.

"Player Advanced." Basketball Reference, n.d. (Seasons 2014-15 to 2023-24) https://www.basketball-reference.com/leagues/NBA_2024_advanced.html.

"Player Shooting." Basketball Reference, n.d. (Seasons 2014-15 to 2023-24) https://www.basketball-reference.com/leagues/NBA_2024_shooting.html.

Spotrac, "NBA Free Agents." Spotrac, n.d., 2024. https://www.spotrac.com/nba/free-agents/_/year/2024/status/available/sort/contract_value

Winston, Wayne L., Scott Nestler, and Konstantinos Pelechrinis. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football 2nd Edition*. Princeton, NJ: Princeton University Press, 2022.

Appendix

Appendix 1: OLS Model Summary

Metric	**Score**
OLS R-Squared	0.918
Dependent Variable	Wins
Intercept	89.1459
Shooting Coeff	381.8220
Turnovers Coeff	-334.5902
Rebound Coeff	93.4279
Free Throw Coeff	106.0491
Shooting R-Squared	0.7916
Turnovers R-Squared	0.1525
Rebound R-Squared	0.0148
Free Throw	0.1525

R-Squared	
Root Mean Squared Error	11.94
Mean Absolute Error	10.01

Appendix 2: Most and Least Accurate Win Predictions

Team	**Wins**	**Predicted Wins**
League Average	41	41.019
Phoenix Suns	49	49.890
Chicago Bulls	39	39.483
Miami Heat	46	47.930
Orlando Magic	47	43.323
Detroit Pistons	14	35.644
Denver Nuggets	57	36.212
Washington Wizards	15	34.876
San Antonio Spurs	22	40.636
Toronto Raptors	25	42.939

Appendix 3: Feature sets for KMeans Clustering

All Features	Shooting Features	Non-shooting Features	**Non-Attempt Features (selected)**
PER_weighted	TS%_weighted	PER_weighted	PER_weighted
TS%_weighted	3PAr_weighted	TS%_weighted	TS%_weighted
3PAr_weighted	FG% by Distance_2P_weighted	3PAr_weighted	3PAr_weighted
TRB%_weighted	FG% by Distance_0-3_weighted	TRB%_weighted	TRB%_weighted
AST%_weighted	FG% by Distance_3-10_weighted	AST%_weighted	AST%_weighted
STL%_weighted	FG% by Distance_10-16_weighted	STL%_weighted	STL%_weighted
BLK%_weighted	FG% by Distance_16-3P_weighted	BLK%_weighted	BLK%_weighted

TOV%_weighted	FG% by Distance_3P_weighted	TOV%_weighted	TOV%_weighted
USG%_weighted	Corner 3s_%3PA_weighted	USG%_weighted	USG%_weighted
% of FGA by Distance_2P_weighted	Corner 3s_3P%_weighted		FG% by Distance_2P_weighted
% of FGA by Distance_0-3_weighted			FG% by Distance_0-3_weighted
% of FGA by Distance_3-10_weighted			FG% by Distance_3-10_weighted
% of FGA by Distance_10-16_weighted			FG% by Distance_10-16_weighted
% of FGA by Distance_16-3P_weighted			FG% by Distance_16-3P_weighted
% of FGA by Distance_3P_weighted			FG% by Distance_3P_weighted

FG% by Distance_2P_weighted			% of FG Ast'd_2P_weighted
FG% by Distance_0-3_weighted			% of FG Ast'd_3P_weighted
FG% by Distance_3-10_weighted			Corner 3s_3P%_weighted
FG% by Distance_10-16_weighted			Heaves_#_weighted
FG% by Distance_16-3P_weighted			
FG% by Distance_3P_weighted			
% of FG Ast'd_2P_weighted			
% of FG Ast'd_3P_weighted			

Dunks_%FGA_weighted			
Corner 3s_%3PA_weighted			
Corner 3s_3P%_weighted			
Heaves_Att._weighted			
Heaves_#_weighted			

Appendix 4: Free agent Protectors vs. Andre Drummond

Player	BLK%_weighted	TRB%_weighted
Omer Yurtseven	2.956	22.467
Chimezie Metu	1.822	13.151
Damian Jones	4.476	12.403
Andre Drummond	3.993	25.832