BAIS:3250 Data Wrangling Project Final Report
2022 NBA Team and Player Success Predictor

Introduction

Throughout the National Basketball Association's Regular and Playoff seasons, NBA viewers get to see some of the best teams and individual players playing professional basketball emerge. Acquiring data on these teams and players helps many gain insight into how well fans' favorite teams and players will do during the regular and Playoff seasons. Through this project, we determined the most significant statistics correlated with a team's success, the correlation between dominating players in top statistics and their team's success, and top contributors to the most successful team in order to accurately predict a team's success, the contributions to their success, and game winners.

Data

Throughout this project, we utilized four different data sources: data from Basketball Reference to view specific player statistics, ESPN to view player efficiency statistics, and the NBA webpage to view traditional and advanced team statistics.

Player Statistics

The first data source we used in our project is a dataset from Basketball Reference[1] on the 2021-2022 NBA regular season stats. Through this dataset, we collected statistics on all 248 players. This dataset included various offensive and defensive statistics, such as points per game, field goal percentages, and defensive rebounds per game.

Basketball Reference has this data separated by scores, standings, players, and team statistics. The data from each of these divisions is provided in CSV format. Data from each of these divisions will be integrated.

Scoring and Shooting Efficiencies

The second data source we used was ESPN[2]. In order to utilize the data provided through this source, we scraped the web page to analyze data regarding the scoring efficiency and shooting efficiency of each of the NBA players playing in the 2021-2022 season. The player efficiencies are used to accurately estimate a player's overall value.

Traditional Team Statistics

[1] https://www.basketball-reference.com/
[2] https://www.espn.com/nba/teams

The third data source we used was NBA Traditional Team Statistics [3]. This data provided general team statistics for all 30 NBA teams, such as wins, losses, win percentages, and field goals made.

Advanced Team Statistics

The fourth data source we used was NBA Advanced Team Statistics[4]. This data shows various advanced statistics on all 30 NBA teams, such as assist ratios, offensive rebound percentages, and defensive rebound percentages.

Merging of Data Sources

Table 1 describes the values from our merged data set that we will be using in order to make our predictions.

Table 1: Data Dictionary

Column	Type	Description
TEAM	character	Name of each team
GP	numeric	Number of games played
W	numeric	Number of wins
L	numeric	Number of losses
WIN%	numeric	WIn percentage
MIN	numeric	Minutes played, the number of minutes played by a player or team
PTS	numeric	Number of points scored
FGM	numeric	Number of field goals made
FGA	numeric	Number of field goals attempted
FG%	numeric	The percentage of field goal attempts that a player makes
3PM	numeric	The number of 3 point foals that a player or team has made
3PA	numeric	The number of 3 point field goals that a player or team has attempted
3P%	numeric	The percentage of 3 point field goal attempts that a player makes
FTM	numeric	The number of free throws that a player or team has made
FTA	numeric	The number of free throws that a player or team has attempted
FT%	numeric	The percentage of free throw attempts that a player or team has made

[3] https://www.nba.com/stats/teams/traditional/

[4] https://www.nba.com/stats/teams/advanced/?sort=DEF_RATING&dir=-1&Season=2021-22&SeasonType=Regular%20Season

OREB	numeric	Offensive Rebounds, the number of rebounds a player or team has collected while they were on offense
DREB	numeric	Defensive rebounds, The number of rebounds a player or team has collected while they were on defense
REB	numeric	Number of rebounds
AST	numeric	Number of assists
TOV	numeric	Number of turnovers
STL	numeric	Number of steals
BLK	numeric	Number of blocks
BLKA	numeric	Blocks against, The number of shots attempted by a player or team that are blocked by a defender
PF	numeric	Number of personal fouls
PFD	numeric	Personal fouls drawn
+/-	numeric	Plus-Minus, The point differential when a player or team is on the floor
team_sce	numeric	Team scoring efficiency
team_she	numeric	Team shooting efficiency
OFFRTG	numeric	Offensive rating
DEFRTG	numeric	Defensive rating
NETRTG	numeric	Net rating
AST%	numeric	Assist percentage, The percentage of teammate field goals a player assisted on while they were on the floor
AST/TO	numeric	Assist to turnover ratio, The number of assists for a player or team compared to the number of turnovers they have committed
AST RATIO	numeric	Assist ratio, number of assists a player averages per 100 possessions used
OREB%	numeric	Offensive Rebounding Percentage, The percentage of available offensive rebounds a player or team obtains while on the floor
DREB%	numeric	Defensive Rebounding Percentage, The percentage of available defensive rebounds a player or team obtains while on the floor
REB%	numeric	Rebounding Percentage, The percentage of available rebounds a player or team grabbed while on the floor
TOV%	numeric	Turnover Percentage, Percentage of plays that end in a player or team's turnover
EFG%	numeric	Effective Field Goal Percentage, Measures field goal percentage adjusting for made 3-point field goals being 1.5 times more valuable than made 2-point field goals.
TS%	numeric	True Shooting Percentage, A shooting percentage that factors in the value of three-point field goals and free throws in addition to conventional two-point field goals
PACE	numeric	Pace, The number of possessions per 48 minutes for a team or player
PIE	numeric	Player Impact Estimate, PIE measures a player's overall statistical contribution against the

		total statistics in games they play in
POSS	numeric	Number of possessions

Analysis

The goal of this project is to analyze what statistics make successful teams and players throughout NBA seasons.

Research Questions and Results:

1*.Which team statistics are most significant in determining the success of NBA teams?*

To find the most significant statistic we performed hypothesis testing based on the individual statistics in correlation to winning. We then graphed the data and used regression analysis for better interpretation. The top 10 most significant team statistics are net rating, offensive rating, defensive rating, three point percentage, field goal percentage, points, turnovers, scoring efficiency, turnover percentage, rebounding, and free throw percentage. *Refer to Table 2 for correlation statistics in relation to Win%*.

2*.Do individual players who dominate in these statistics belong to teams that were successful?*

Yes. We made 6 subsets to get the players who had the best FG percentage, were the best three point shooters, the best FT shooters, best rebounders, and best all around scorers. We used whether or not the team made the playoffs as a barometer of if the player was on a successful team. We found 6/10 best three point shooters were on teams that made it to the playoffs, 8/10 players with the highest field goal percentage made the playoffs, 9/10 players who scored the most points made the playoffs, 7/10 players with the lowest TOV's made the playoffs, 7/10 players with the highest rebounding numbers made the playoffs, and 9/10 players with the highest free throw percentage made the playoffs. In total, 76.67% of the players who dominated in these statistics were on teams that were successful which we believe to be significant.

3.*Which player contributes most to the success of the best NBA team?*

We made a subset to find players that had stats that best correlated with winning on the Phoenix Suns, who were the most successful team (had the most wins). All the players in the subset had low turnover numbers, they had high three point percentages, high field goal percentages, and high amounts of points scored. Based on this data I can confidently say that Devin Booker is the most impactful player on the Suns when it

comes to winning. He had by far the most points scored per game in the subset, second highest 3pt percentage in the subset, and second highest rebound numbers among the subsetted group of valuable players.When you factor in his three point scoring efficiency and rebounding, with how far ahead he was in points, it is clear he was the most impactful.

Interpretation of Findings:

Team Statistics Relationship with Win Percentage

We began by analyzing the correlation between various team statistics and their overall win percentage throughout the regular season. We conducted this statistical test through the use of the “cor.test” function. We found that the top three statistics that impacted teams’ overall win percentages were net rating, offensive rating, and defensive rating. All of the correlations we found through our correlation tests are ranked in *Table 2*, from statistics with the most impact on the win percentage, to statistics with the least impact.

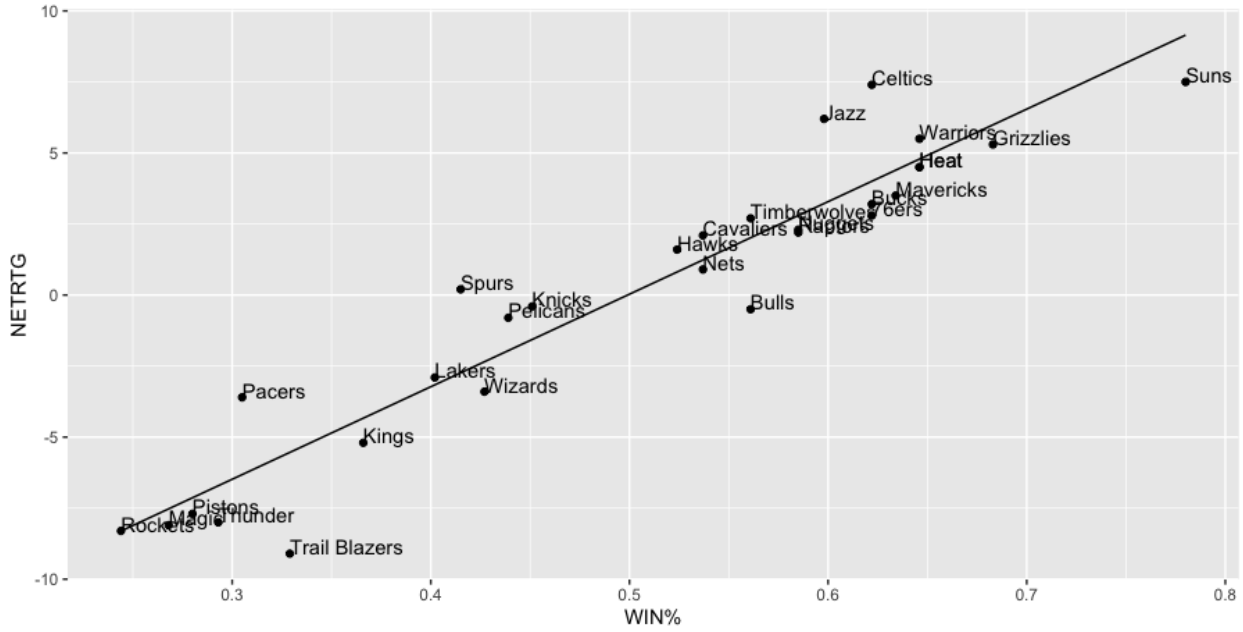
Table 2 Correlation Among Team Statistics and Win Percentage

Ranking	**Team Statistic**	**Correlation**
1	Net Rating	0.95
2	Offensive Rating	0.80
3	Defensive Rating	-0.77
4	Three Point Percentage	0.69
5	Field Goal Percentage	0.62
6	Points	0.59
7	Turnovers	-0.46
8	Scoring Efficiency	0.44
9	Turnover Percentage	-0.41
10	Rebounding	0.39
11	Free Throw Percentage	0.38
12	Shooting Efficiency	0.37

In *figure 1*, we created a scatter plot that presents the correlation between the net ratings of each NBA team and the win percentage of each NBA team. Through this

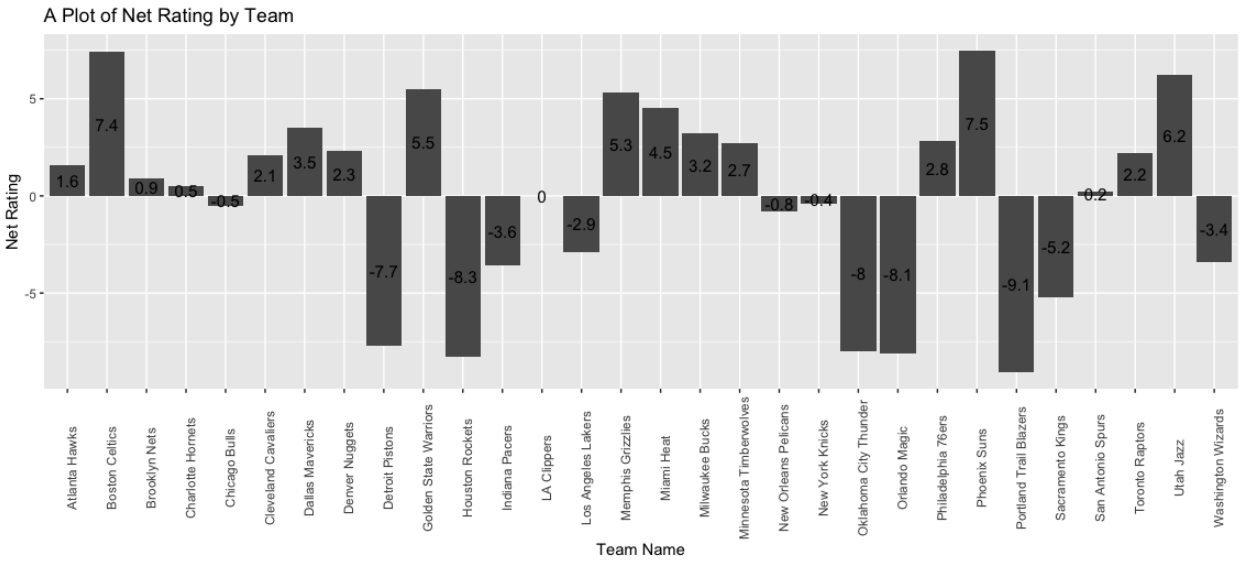
chart, we can see that there is a positive relationship between net rating and win percentage.

Figure 1 Each NBA Teams' Win Percentage Compared to Net Rating



After conducting these correlation tests and finding that the team's net rating had the most impact on the win percentage of NBA Teams, we conducted further analysis to determine which NBA teams were the most successful. *Figure 2* shows a bar plot of the net ratings for each NBA team. Within this figure, we can see that the Phoenix Suns have the highest overall net rating, at 7.5. The Suns are then followed by the Boston Celtics at 7.4, the Utah Jazz at 6.2, and the Golden State Warriors at 5.5.

Figure 2 Bar Plot of Net Rating for Each NBA Team



Conclusion

After conducting several statistical tests and creating various visualizations and summaries, we took this information and estimated which NBA Team would be the most successful throughout the season. We determined that the most successful NBA team was the Phoenix Suns, since they had the highest win percentage and highest net rating. Using this information, we decided to further analyze which NBA player was the most successful in the league and through further research, we found that Devin Booker contributes the most success to the best NBA Team. The data we researched and analyzed throughout this project does have some real world applications. For example, NBA viewers who are interested in sports betting can take this information we have compiled and use it when deciding what kind of bets they would like to make or where they can make the most money in different NBA bets.

<u>Limitations</u>

One of the limitations faced while determining which teams and players would be most successful throughout the season was that not all of the statistics that we analyzed can contribute to a team's success in NBA games. Sometimes, teams with lower net ratings and overall lower team statistics can be more successful throughout the season than expected. With that being said, there is somewhat of a randomness factor that can determine what team may be successful throughout the season. Some of these random factors include home court advantages and the strains of traveling to other opponent's courts. This makes it difficult to completely rely on player and team statistics when trying to predict who will be successful throughout the season.

Suggestions

In order for our success indicator model to improve upon its accuracy, we suggest finding a way to take those random factors into consideration. There is the famous FiveThirtyEight Elo NFL Predictor Model[5] that determines the margin of victory for NFL teams. This model makes adjustments for the advantages of home games, how bye weeks impact players, the advantages or disadvantages of playoff games, and many more factors. Our model can become more accurate if we included data on where certain games were played and which games were playoff versus regular season games. We can apply the NFL Predictor equation to the data we have from our analysis, but by making adjustments that align more with NBA games.

[5] https://towardsdatascience.com/improving-a-famous-nfl-prediction-model-1295a7022859