Exploring the 'I' in Team: Unveiling the Influence of Individual Players on NBA Team Success

John Poehler

University of North Texas

Toulouse Graduate School

ADTA Capstone

Levent Bulut, PhD

Summer 2023

Table of Contents

[Introduction 4](#_Toc140933285)

[Data Analysis Use Cases in Basketball 4](#_Toc140933286)

[Research Questions 6](#_Toc140933287)

[Scholarly Review 8](#_Toc140933288)

[The History of Basketball 8](#_Toc140933289)

[Using Data to Improve Strategy 10](#_Toc140933290)

[Modeling Techniques 11](#_Toc140933291)

[Data Preparation 13](#_Toc140933292)

[Software 13](#_Toc140933293)

[Data Collection 13](#_Toc140933294)

[Data Wrangling and Cleaning 13](#_Toc140933295)

[Exploratory Data Analysis (EDA) 16](#_Toc140933296)

[Data Description 16](#_Toc140933297)

[Response Variable(s) 17](#_Toc140933298)

[Qualitative Variables 20](#_Toc140933299)

[Quantitative Variables 27](#_Toc140933300)

[Modeling 34](#_Toc140933301)

[Multiple Linear Regression 34](#_Toc140933302)

[Decision Tree 39](#_Toc140933303)

[Logistic Regression 43](#_Toc140933304)

[LASSO Regression 45](#_Toc140933305)

[Conclusion 48](#_Toc140933306)

[Discussion 48](#_Toc140933307)

[Applications 49](#_Toc140933308)

[Limitations 49](#_Toc140933309)

[Future Research 50](#_Toc140933310)

[References 51](#_Toc140933311)

**Introduction**

The National Basketball Association has been around for decades and basketball itself has existed for over a century. Throughout its duration it has been a driving force it has been a significant force within the realm of sports and American culture. Over the years this simple game has evolved with various strategies and ideologies coming to the forefront.

A timeless adage that has been connected with sports and especially basketball is the phrase “there is no ‘I’ in team”. This saying is probably meant to evoke a sense of teamwork amongst players that should be collaborating for the ultimate goal of winning, but with so many individual athletes receiving such praise, acclaim, fame and fortune one might believe that there is indeed an ‘I’ in team. The NBA as a league has benefitted greatly from players like Michael Jordan, LeBron James as well as many others and the deification of those individuals by fans everywhere seems to drive the point even further. The NBA benefits greatly from these kinds of superstar players because it brings excitement and marketability to their sport, much more so than fundamentally sound, and well coached teams.

Every organization in the NBA should have one main goal for every season and that is to win a championship. Do these teams need to have a superstar, do-it-all player to have a successful season or does the teamwork over everything approach appear to be more fruitful in the end?

## Data Analysis Use Cases in Basketball

Similar to other industries basketball front offices in the NBA, College, as well as professional organizations overseas utilize data and perform analysis in order to gain a competitive edge. Organizations will have entire departments that focus on looking at models to determine the most efficient way to play basketball. One of the most important aspects of basketball today is the 3-point shot, primarily based on the objective fact that it is worth more points despite being a more difficult shot. In the years between 2014 and 2019 the NBA saw an explosion in 3 point shot popularity, teams like the Golden State Warriors and Houston Rockets lead the charge as they took a large portion of their shots behind the 3-point line starting what is usually called the “3 Point Revolution” (Freitas,2021). The Rockets leaned so far into the 3-point shooting philosophy that they were the first team in NBA history to attempt more 3-point shots than 2-point shots in an entire NBA season in 2017-2018, taking an astonishing 3470 3 pointers to their 3436 2-point shots (Freitas,2021). Staying with the Rockets, this same team relied so much on the analytics forward thinking 3 point shot strategy that they had missed record 27 consecutive 3-point shots in a Western Conference Finals game that would have sent them to the NBA Finals if victorious, never wavering from their efficiency-based approach (Rollins, 2018). This illustrates both the importance of data analytics as well as the importance of the 3-point shot in the game of basketball today. For the dataset that will be analyzed in this paper only players that season’s occurred after 1979-1980 will be considered due to the aforementioned effect of the 3-point line and its ramifications on stats and the game as a whole. Players that had access to the 3-point line will most likely have inflated points per game (PPG) numbers but will have adverse effects on their field goal percentage (FG%) compared to players that predated 1979, making the two groups of players incompatible for comparison.

Data analysis contains an endless amount of utility when considering its overall effect on basketball. These effects can span across predicting the opponent’s strategy, determining the most effective lineup to play at any given time as well as forecasting potential injuries and when they might occur. The latter of these examples is something that is being worked on by teams and analytics departments alike. There have been attempts to create a deep learning model to try and pinpoint when players may experience injuries based on factors such as age, height, weight and even the coach they play for (Cohan & Fernandez, 2021). With information like this organizations can adjust the playing time for some of their veteran players in order to avoid potential injuries if they are deemed at risk. Similar to the aforementioned 3- point revolution, data analytics have made teams alter their approach to player injury. Despite being much maligned from NBA fans, teams have integrated “load management” for their star players, a strategy that involves benching their injury-prone, star players for a significant portion of the regular season to try to save their health for the playoffs (Reilly, Solow, & von Allmen, 2023). This shows the various ways in which organizations can utilize data to try to optimize their approach and gain an edge on the opposition.

## Research Questions

NBA Win Shares is a statistic that tries to quantify the amount of “credit” an individual should receive from their team’s win. This metric could be the best determinant of the potential “I” in team and could help NBA teams decipher if they need to change their strategy to or away from a team-centric approach. The research questions that will be examined and answered with analysis are:

• **Can you predict a team’s win percentage based on the statistics of an individual player**?

To answer this question, I will investigate the variables that exist within the dataset and choose the factors that are most pertinent to predicting win percentage. A regression model will be created with these variables with the goal being to predict win percentage.

• **Can you predict if a team will win an NBA championship based on individual statistics**?

This question will be answered by creating a decision tree model based on the available variables in the dataset and testing if the model can be effective in terms of predicting if a team won a championship that year.

•**Which factors contribute to Win Shares the most in the given dataset**?

This answer can be achieved by analyzing the variables in the dataset and determining which are the most important for predicting the Win Shares variable.

**Scholarly Review**

## The History of Basketball

The game of basketball has been a staple in daily life and pop culture for many decades now. It harbors some of the most recognizable names, apparel, and organizations in the world today. Although not the first or most popular sport during its inception, it has grown over time to become one of the most entertaining spectacles in the world. Invented by a physician named James Naismith, in 1891, basketball formerly “basket ball”) was initially played by throwing a soccer ball into a peach basket that was hung 10 feet above the ground and then retrieving the ball from the basket with a ladder to resume play (Cantwell, 2004). Basketball was merely an activity that was used to keep Naismith’s YMCA students moving in between football and baseball season but became much more than that in the years to come. As expected with any new sport, the rules were vastly different from how they are today and much as changed. Initially, basketball only had 13 rules which were hastily wrote by Naismith after coming up with the idea of basketball, but over the years we have seen the introduction of the first basketball ball in 1893, the advent of the free throw in 1894 , the legalization of dribbling the ball in 1898, the banning and subsequent reinstatement of the dunk in 1967 and 1976, as well as the addition of a shot clock and 3 point line in 1985 and 1986 respectively (Cantwell, 2004).

The growing popularity of basketball meant that it was only inevitable for the formation of a professional basketball league. Although the National Basketball Association (NBA) is the current golden standard for professional basketball, it was not the only professional basketball league in the Americas. The first official professional basketball league was the National League of Professional Basketball Teams founded in 1898, despite not being a longstanding or even successful league it still holds the title as “first”. In terms of the first substantial league, the National Basketball League (NBL) was founded in 1937, which then led to the founding of the Basketball Association of America (BAA) in 1946, which inevitably lead to the merger of the two leagues, forming what we know now as the NBA (Staffo, 1998).

As the NBA became more established and basketball continued to grow in popularity, more talented players began to enter the league and endear themselves to fans. When an NBA player has immense talent and production as well as popularity within the fanbase that player is often deemed a “superstar”. Arguably the first superstar to exist in the NBA was George Mikan, a 6-foot 10-inch center that dominated the sport en route to five NBA championships out of 6 chances with the then Minneapolis Lakers (Staffo, 1998). George Mikan’s contributions were not only important during his playing career but well into the future of the NBA, as he paved the way for other talented players to earn the moniker of superstar and carry their teams as well as the league to prosperity.

Post George Mikan, there has been a multitude of NBA superstars for varying franchises in the league. Names like LeBron James, Michael Jordan, Kareem Abdul-Jabbar, Kobe Bryant as well as many other come to mind, while you may be able to come up with names with a similar pedigree in other sports, they would not matter as much to their respective sports as much as the aforementioned names mean and meant to the NBA as a whole. Players like LeBron James evoke many things in a sports fan that makes attending one of their games very appealing; in many cases people, including myself may view attending a game in which James plays as a “bucket-list” event, something that they must do before they die, or he retires. While discussing superstars, not all of them are built the same way, according to an article written on superstardom in the NBA the different types of superstars are “Adler” superstars and “Rosen” superstars (Humphreys, 2020). Adler superstars earn their superstardom as a derivative of their sole popularity in the sporting world, whereas Rosen superstars as coined superstars based on their performance in the sport (Humphreys, 2020). While there may be multiple routes to becoming a superstar, there is no debating their effect on the league.

Despite the clear advantages of superstardom in the NBA, there are some downsides. In many instances, the focus on the individual rather than the team can have some detrimental effects in terms of winning and losing. An article written about the NBA Most Valuable Player (MVP) award argues that the effort put forward to accrue MVP worthy stats may only help the individual player but will harm the ultimate success of the team (Nutting, 2010). After seeing what their predecessors have done when considering superstardom, players may put all of their focus into being perceived as a star by playing selfishly and putting statistics over winning and overall team success. Additionally, players that are in close competition for MVP may look at what stats their competition and actively try to one-up them in terms of stats regardless of what their team needs for them to do to win (Nutting, 2010).

## Using Data to Improve Strategy

Moreso in basketball than other professional sports, one individual player seems to receive a lot of the credit compared to overall gratitude that the rest of their teammates get. Fans will often use phrases like “Jordan’s Bulls” or “Kobe’s Lakers” when referring to the team as whole. This type of perception may alter either the way an individual player might play or alter the coaching strategy that is utilized in game. Teams are often constructed in a way that highlights the abilities of one individual instead of focusing on the best ways a team can work together. A journal article named “Examining the ‘I’ in Team” discusses narcissism in the NBA and how it plays a role in team composition and success, claiming that team with low narcissism scores tend to outperform expectations versus teams with higher narcissism tend to fall short of expectations (Grijalva, Maynes, Badura & Whitling, 2020). This comes down to team composition and organizations must be privy to players that do and do not work with each other, it is the coach’s responsibility to determine which lineups are the most effective. A study was created in order to find the most effective lineups, pairings and players when considering their interaction with their teammates using data to back up their claim, being able to identify specific lineups that seemed to outperform their expectation (Devlin, & Uminsky, 2020). Coaches can take the information that is provided from these data models and apply it to their gameplan and if used properly, it will help them win more games than if they were relying on traditional techniques.

## Modeling Techniques

In order to uncover what we need to, there must be certain models that have to be utilized. For the first research question I am interested in predicting the numerical variable of win percentage. To accomplish this, a Multiple Linear Regression model would be the most effective method to be able to determine this variable. In a previous study regarding international basketball in FIBA competition, the researchers were interested in finding the relationship between a basketball match outcome and certain key performance indicators or KPIs. In this study an MLR model was used to effectively predict and measure the relationship between the variables (Zhang, Gomez, Yi, Dong, Leicht, & Lorenzo, 2020). The success in this study with a similar research focus had influenced my decision to use an MLR model in this study.

In another study, decision trees were utilized as a way of predicting the outcomes of basketball games with the result being a binary,yes or no fashion. Statistical measures such as personal fouls, field goal percentage, free throw percentage and rebounding were analyzed and put into models to help decide the binary variable of winning and losing (Pai, ChangLiao, & Lin, 2017). The findings and methodology of this study appears to apply to what I am trying to accomplish with my data.

Logistic Regression was employed in a study that was completed on the different basketball shot types across various levels and their success on a “made” or “missed” binary scale (Erčulj, & Štrumbelj). This sort of modeling is very similar to the target variable of “Won Championship?” and could be fruitful in terms of analyzing a binary variable.

For the last research question the main focus is variable selection and I wanted to find out which models perform the best in that facet. One of the best models that was found through research was a LASSO or Least Absolute Shrinkage and Selection Operator regression model. LASSO regression’s goal is to find the variables and regression coefficients that created a model with minimized prediction error, done so by shrinking coefficients towards zero until there are only useful variables in the model (Ranstam, & Cook, 2018). The LASSO regression will be a great model to help identify which variables are the most important to determine Win Shares.

Data Preparation

## Software

Microsoft Excel version 2305 and RStudio 2023.06.0 Build 421 were the only programs utilized for the altering and analysis of the data.

## Data Collection

All of the data included in this study was collected from Stathead.com using their player season finder feature. The top 1000 NBA players in terms of win-shares in a given season between the 1979-1980 season to the 2022-2023 season were selected and make up the dataset. There were 33 variables included in the original dataset retrieved from Stathead.com, an additional 4 variables were manually added to each of the 1000 rows, with those variables being: Wins, Losses, win percentage and Champ (a binary variable that explains if a given players team won the championship that year). The information to input the aforementioned columns were also retrieved from Stathead.com.

## Data Wrangling and Cleaning

Some wrangling that needed to be done was replacing missing values with null/ NA. Most of the cases of missing values came in the 3 Point Field Goals and 3 Point Percentage columns. The data was not incorrect, it just reflected the reality that some NBA players do not attempt many 3-point shots and thus do not make them. A player that does not attempt a 3-point shot will have an “N/A” for their 3P% whereas a player that attempted multiple 3-point shots but did make any of them will have a 0 for that respective value. This difference is reflected in the data and influences some of the model results.

The other instance of missing data is in the Games Started column, it only occurs in the years prior to 1984 in the data. However, despite not having the number Games Started from the player, their total games played in or Games column is recorded. This missing data will be transformed into a null value and should not affect any of the analysis because it is primarily not an important statistic. When looking at the top 1000 player seasons it is almost certain that every single player is a starter for their given team and not a backup/ bench player. The total number of games played is a more important metric as it can be paired with the per-game metrics to paint the entire picture.

One more issue that needs to be addressed is the NBA seasons that were shortened due to some extraneous factor. The 1998-91 season, and the 2011-12 season were not the usual 82 games due to player lockouts that occurred during those timespans, being 50 games and 66 games respectively. Additionally, the 2020-21 NBA was also shorted due to the COVID-19 pandemic which was only 72 games. In an alternate reality in which these seasons had lasted the entire 82 games, some players would have potentially recorded better stats and their data would be altered in the dataset. The data for these seasons could be standardized to match the 82-game total that the other seasons had and thus increasing some of the statistical totals in the data. However, I am choosing not to do any standardization as it would incorrectly reflect how well that player played during that given season, factors like injury, fatigue, load management and coach strategy that affect the end of real-life seasons would not be taken into consideration when standardizing these seasons. The shortened season data will be left as is and will not be adjusted.

Additionally, official basketball positions are listed as Point Guard (PG) and Shooting Guard (SG) which fall under the general Guard (G) position. For the general Forward (F) position there as Small Forward (SF) and Power Forward (PF) and there is also the Center (C) which has its own classification. In the dataset, many players are listed as “F-G” or “G-F” meaning “Forward/Guard” or “Guard/ Forward” which signifies a player’s proclivity to player both positions. In this configuration the position in which the player spends the most time is listed first. In order to make the dataset more generalizable only the players primary position will be included, for example Michael Jordan is originally listed as a “G-F” but will now be listed a “G”. This makes the only possible inputs for the Pos column “G”, “F” or “C”.

Lastly, some players’ names were altered to remove any accented letters that were included in their names. Names like José Calderòn, Manu Ginòbili as well as others would cause issues with certain functions in R. All of the names that included any sort of letter that is not included in the English alphabet was removed and replaced with its English counterpart.

**Exploratory Data Analysis (EDA)**

## Data Description

There are 37 total variables in the master dataset including the 4 variables of “Wins”, “Losses”, “Win%” and “Champ?” that were added to the preexisting dataset retrieved from Stathead.com. Figure 1 below depicts each of the variables and explains them in laymen’s terms.

|  |  |  |
| --- | --- | --- |
| Variable Name | Datatype | Description |
| Rank (Rk) | Ordinal | Overall ranking in terms of Win Shares |
| Player | String | NBA Player’s Name |
| Win Shares (WS) | Continuous | Statistic to determine the amount of credit a player should get for each team win set by a formula |
| Season | String | Year of NBA season |
| Age | Discrete | Age of NBA player during given season |
| Team | Nominal | NBA team played for during given season |
| Games (G) | Discrete | Number of games played during season |
| Games Started (GS) | Discrete | Number of games started during season |
| Minutes Played (MP) | Discrete | Total Number of minutes played during season |
| Field Goals (FG) | Discrete | Number of field goals attempted made |
| Field Goal Attempts (FGA) | Discrete | Number of total field goals attempted |
| 2 Point Field Goals (2P) | Discrete | Number of shots made **inside** the 3-point line |
| 2 Point Attempts (2PA) | Discrete | Number of shots attempted **inside** 3-point line |
| 3 Point Field Goals (3P) | Discrete | Number of shots made **outside** 3-point line |
| 3 Point Attempts (3PA) | Discrete | Number of shots attempted **outside** 3-point line |
| Free Throw (FT) | Discrete | Number of free throws made |
| Free Throw Attempts (FTA) | Discrete | Number of free throws attempted |
| Offensive Rebounds (ORB) | Discrete | Number of rebounds acquired while on offense |
| Defensive Rebounds (DRB) | Discrete | Number of rebounds acquired while on defense |
| Total Rebounds (TRB) | Discrete | Total number of rebounds acquired |
| Assists (AST) | Discrete | Number of field goals player assisted on |
| Steals (STL) | Discrete | Number of times player stole ball away from the other team |
| Blocks (BLK) | Discrete | Number of shot attempts blocked |
| Turnovers (TOV) | Discrete | Number of turnovers committed |
| Personal Fouls (PF) | Discrete | Number of personal fouls given to player |
| Points (PTS) | Discrete | Total number of points scored |
| Field Goal Percentage (FG%) | Continuous | Percent of field goals made over total attempted |
| 2 Point Percentage (2P%) | Continuous | Percent of 2-point field goals made over attempted |
| 3 Point Percentage (3P%) | Continuous | Percent of 3-point field goals made over attempted |
| Free Throw Percentage (FT%) | Continuous | Percent of free throws made over attempted |
| True Shooting Percentage (TS%) | Continuous | Statistical measure to determine a player’s efficiency when shooting set by a formula |
| Effective Field Goal Percentage (eFG%) | Continuous | Adjusted shooting percentage when weighing 3-point shots more in formula |
| Position (Pos) | Nominal | Position player played during season |
| Win | Discrete | Number of Wins team had during season |
| Loss | Discrete | Number of Losses team had during season |
| Win Percentage (Win%) | Continuous | Percentage of games won over total played |
| Won Championship? (Champ?) | Binary | Indicates if player’s team won the NBA championship that year (0=No, 1=Yes) |

*Table 1*

## Response Variable(s)

**Win Shares (WS)** is the response variable that is pertinent to this study, it is a quantitative variable that is meant to calculate each individual player’s contribution to a team’s win. The methodology by which WS is calculated is rather complicated and involves several different factors. To calculate WS one must first determine a player’s Points Produced, which is formed by the formula (FieldGoalAttempts+.044FreeThrowAttempts+TurnOvers) (Offensive Rating)/100| (NBAstuffer, n.d.). Next, to calculate Marginal Offense the formula is (Points Produced) - 0.92 \* (League Points per Possession) \* (Offensive Possessions) and to calculate Marginal Points per Win the formula is 0.32 \* (League Points per Game) \* ((Team Pace) / (League Pace)) | (Basketball Reference, n.d.) . Thus, to calculate Offensive Win Shares Marginal Offense must be divided by Marginal Points per Win.

However, this is just half of the total Win Shares equation, now defensive win shares must be calculated. First Defensive Rating must be calculated with the simple equation 100\* (Opponent Points Scored / Opponent Possessions) | (NBAstuffer, n.d.). Then Marginal Defense must be calculated by (Player Minutes Played / Team Minutes Played) \* (Team Defensive Possessions) \* (1.08 \* (League Points per Possession) - ((Defensive Rating) / 100)) and then Marginal Defense is then divided by the aforementioned Marginal Points per Win formula and then Defensive Win Shares is found (Basketball Reference, n.d.). Lastly, Offensive Win Shares and Defensive Win Shares must be added together to form our response variable Win Shares. The overall equation is as follows: WS = ((Points Produced) - 0.92 \* (League Points per Possession) \* (Offensive Possessions)) / (0.32 \* (League Points per Game) \* ((Team Pace) / (League Pace))) + ((Player Minutes Played / Team Minutes Played) \* (Team Defensive Possessions) \* (1.08 \* (League Points per Possession) - (Defensive Rating / 100))) / (0.32 \* (League Points per Game) \* ((Team Pace) / (League Pace))) Thankfully, Stathead.com has already calculated WS for each of the players in the dataset and they do not need to be manually calculated.

Shown in Fig. 2 below, the distribution of the response variable is right skewed with a maximum of 21.2, a minimum of 8.6, a mean of 11.03 and a median of 10.3 as indicated by the vertical lines on the figure.

A picture containing diagram, screenshot, plot, text

Description automatically generated

Fig. 2

One secondary response variable is Win Percentage (Win%) which is much less complicated than WS. Win% is the percentage of games a team wins out of the total played. Fig. 3 below shows the distribution of Win% for each players team that is in the dataset there appears to be a mostly normal distribution with tails on either side and most of the observations occurring near the center. The maximum value comes from players on the 73 win and 9 loss Golden State Warriors in 2015-16 and the minimum value comes from Kevin Love on the 2010-11 17-win 65 loss Timberwolves. A graph of green and blue bars

Description automatically generated

Fig. 3

## Qualitative Variables

There are 4 qualitative variables that are included in the dataset and those are Player Name, Season, Team, and Champ(?) Season, Team and Champ(?) will be ultimately utilized in the analysis but Player Name will mostly be used as an identifier for the data.

Position is a very important aspect of basketball because it typically determines which role and duties each individual player will have when on the court, whether they will be asked to rebound the ball, facilitate or score depends on their position. When looking at how WS are spread throughout each position in Fig. 3 below, the groups are almost indistinguishable from one another. This shows that there is a healthy distribution between the position group and there is not one single position that is that much greater than another in terms of WS. However, when looking at the total count of each position of Guard, Forward and Center the total count is 375, 427 and 198 respectively. This result makes sense due to the fact that both Guard and Forward are made up of 2 separate positions being Point Guard/ Shooting Guard and Small Forward/ Power Forward, whereas Center is the only position in its position group.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Fig. 4

I also wanted to see the distribution between NBA seasons to determine if there was a uniformity that was present. In Fig.4 below, most seasons appear to hover around 20 players in the top 1000 but there are some clear outliers when looking at the chart. Seasons 1998-99 and 2011-12 are both well below most other seasons, totaling 3 and 9 players respectively. While this could just be a down year in terms of individual player production, in reality both of these seasons were affected by NBA lockouts which drastically shortened the length of both of the NBA seasons in those years (Sportskeeda, 2022). Based on the formula for Win Shares being so reliant on totals as well as team wins, it would be very difficult for players in a shortened season to accrue Win Shares.

A picture containing text, screenshot, purple, violet

Description automatically generated

Fig. 5

As for the next qualitative variable “Player”, I wanted to take the top names in terms of number of seasons and see how they compared to one another. Fig.5 and Fig.6 below show both the frequency of seasons that are in the top 1000 for WS and the quality of those seasons. For example, Chris Paul is tied for second for number of seasons in the top 1000 with 15 but when looking at his box plot, his WS numbers are near the bottom of the selected players. Inversely, Michael Jordan is among the middle of the pack when looking at frequency but has far and away the most impressive seasons when observing his box plot. Appearing many times in the dataset is more of a nod to their longevity in the NBA.

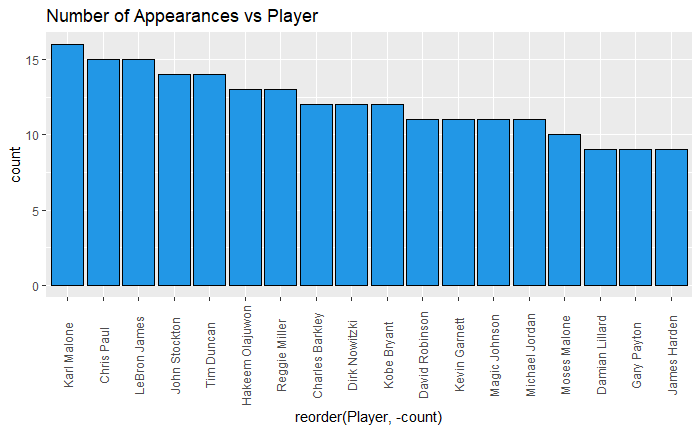


Fig. 6

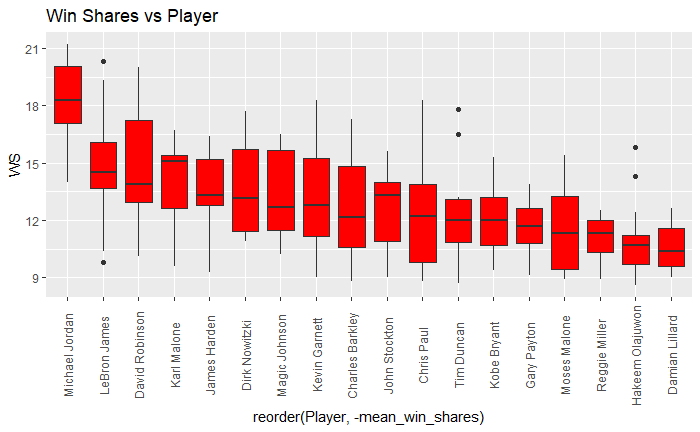


Fig. 7

Inversely, despite some significant differences in WS for these top players, Fig. 8 shows the Win% of the teams each of them played for and there is basically no discernible difference between them. Most of the players hover around the .600 mark.

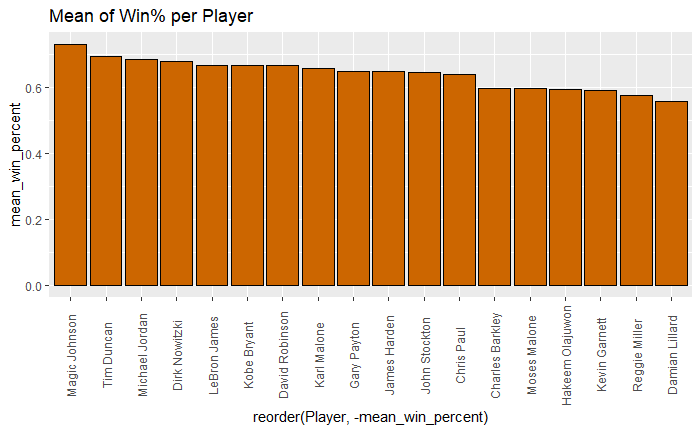


Fig. 8

When looking at how each of the top 1000 seasons are spread across each team in Fig. 7, the expansion teams as well as disbanded teams make the lower rung such as Brooklyn Nets (BRK) and Vancouver Grizzlies(VAN) . Towards the top of the chart are the perineal dominant teams like the Boston Celtics (BOS), Los Angeles Lakers (LAL) and surprisingly the Utah Jazz (UTA). In Fig. 8 you can see the mean WS for each team, where the Chicago Bulls (CHI) has one of the highest values due to Michael Jordans contributions and teams like the Lakers and the Celtics are not as high due to the amount of role players that exist for each of those teams in the dataset.

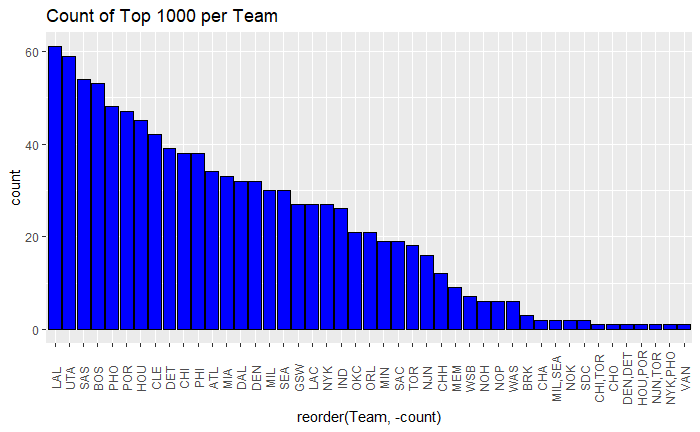


Fig. 9

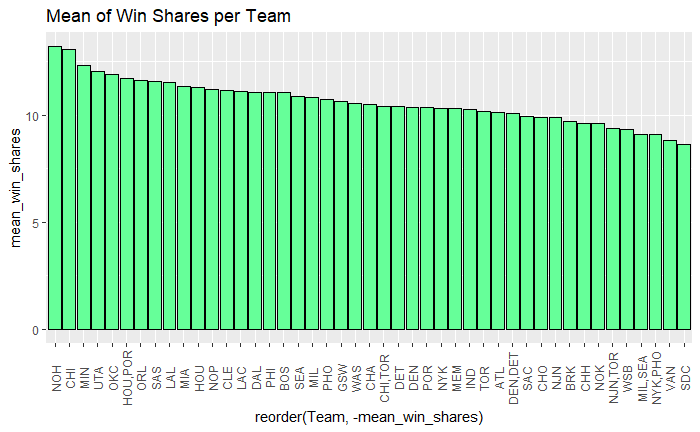


Fig. 10

For Win% per team in Fig. 10 the distribution is mostly uniform apart from the perineal great teams and the expansion teams that never had much success.

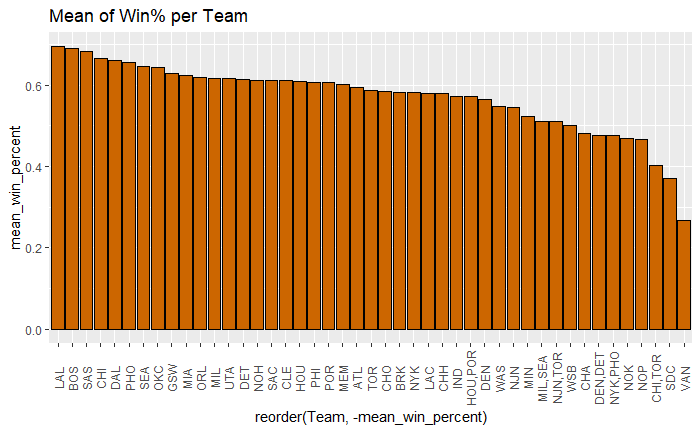


Fig. 11

For the last qualitative variable “Champ?” I wanted to see if there were any discernible differences between players that were champions in that given season and players that failed to win the championship. Fig. 9 shows box plots comparing the two outcomes “no” on the left and “yes” on the right and although there is a substantial difference in total count among the two with 89 players winning a championship and 911 not, there is still visual difference between two. Most of the championship winners seem to fall within a reasonable range whereas there are a lot of outliers involved with the teams that did not win. That piece of information may match the theory that NBA teams need a compilation of good/ great players instead of one that dominates the stat column.

A graph with red squares

Description automatically generated

Fig. 12

## Quantitative Variables

In this dataset the great majority of variables are quantitative/ numerical with there being 34 in total. Below is in Fig.13 is heatmap that shows the correlation between each of the 34 variables with each other. Due to the existence of some null values in the 3P% column they could not be visualized properly in R and are left blank. Many of the variables do not seem to have any correlation to one another and that would make sense given that they are measuring completely different things on the basketball floor such as Steals and Free Throw Attempts. However, there are other variables that are closely related because of the nature of their calculation such as True Shooting% and Effective Field Goal%.

A picture containing text, stitch, screenshot, pattern

Description automatically generated

Fig. 13

In order to combat some of the redundancies involved with some of the statistical categories an additional heat map (Fig. 14) was created with some variables that related too closely to others were removed. This heatmap is much more palatable and the relationship between the variables can be better understood. Something that can be taken into account is the fact that there is not a single statistical category that has a strong correlation with Win%.

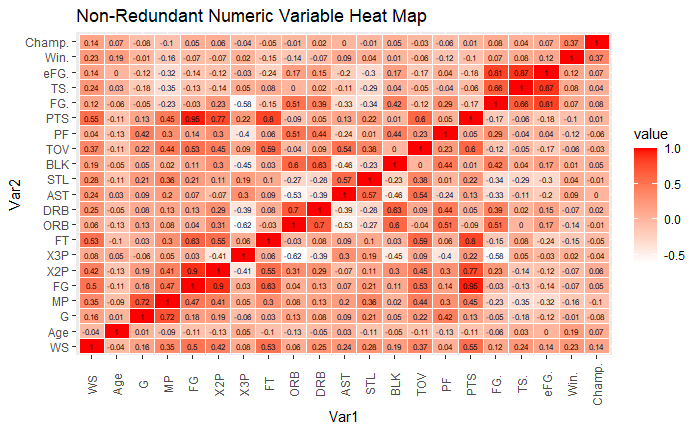
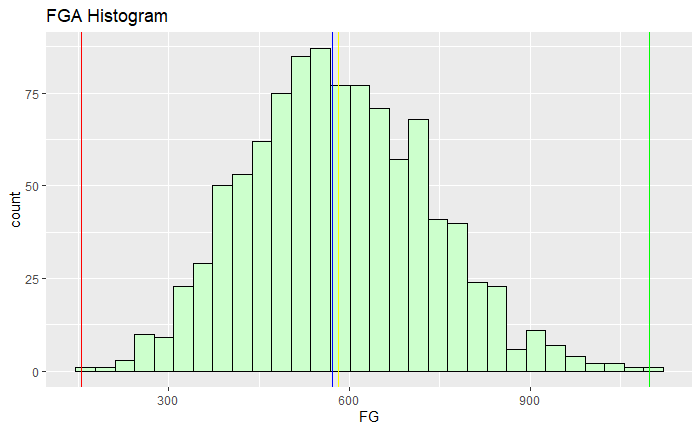
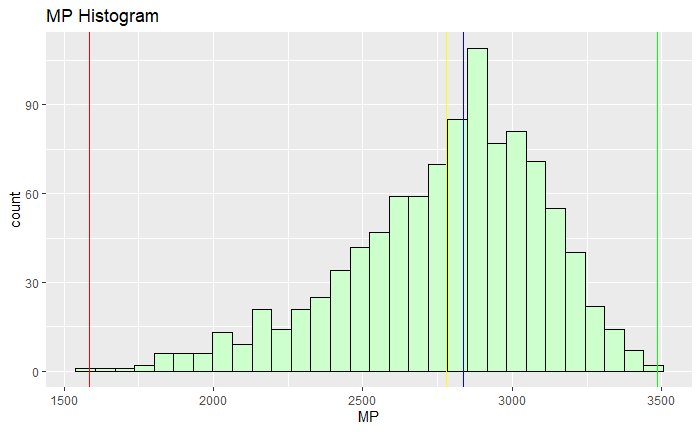
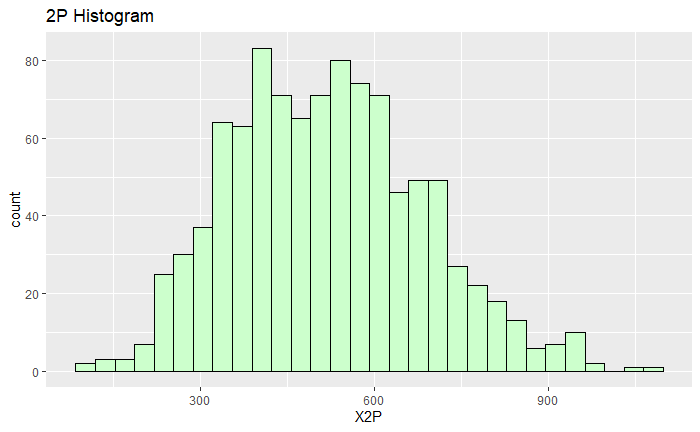
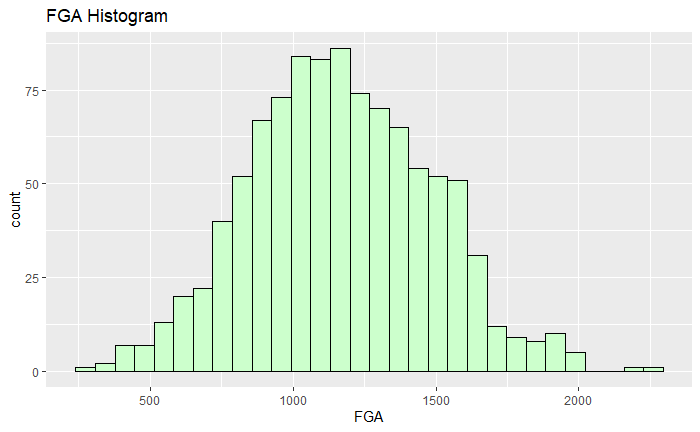
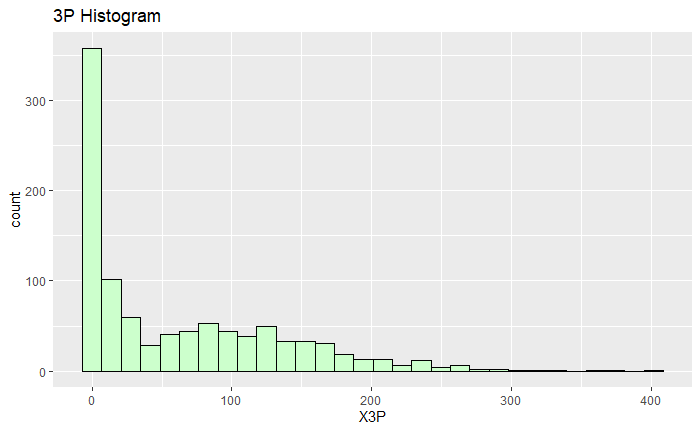
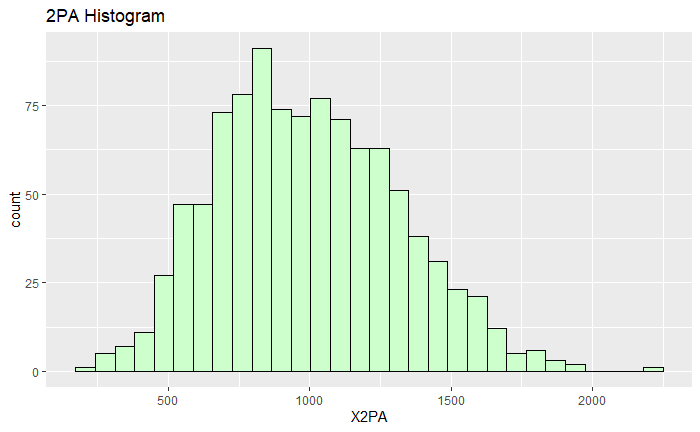


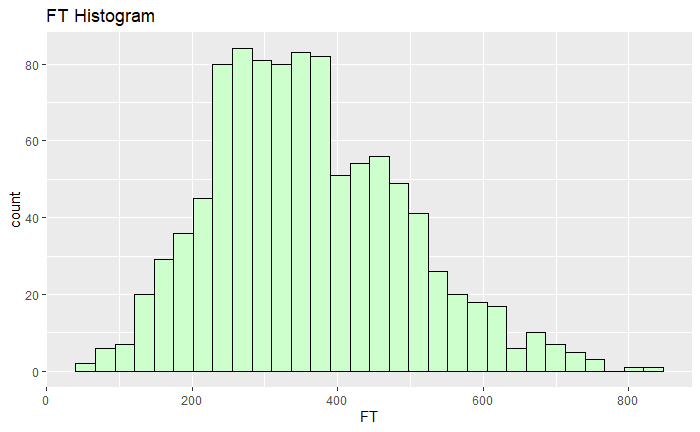
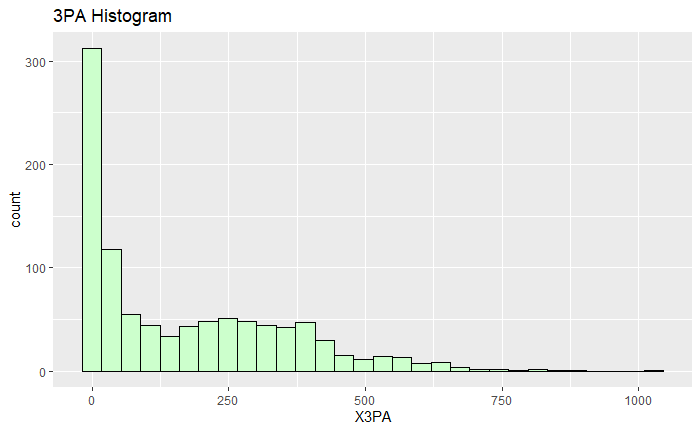
Fig. 4

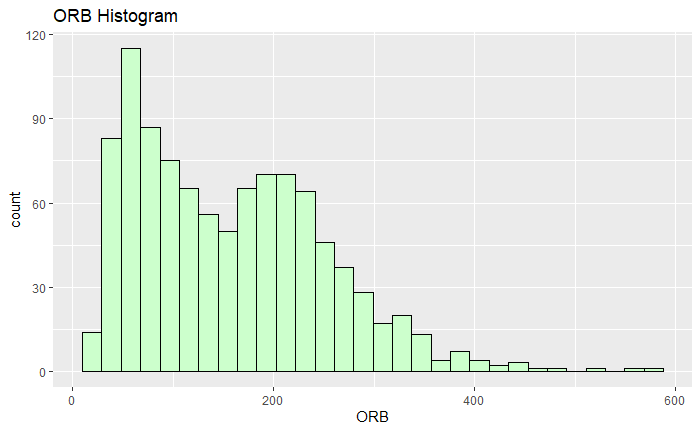
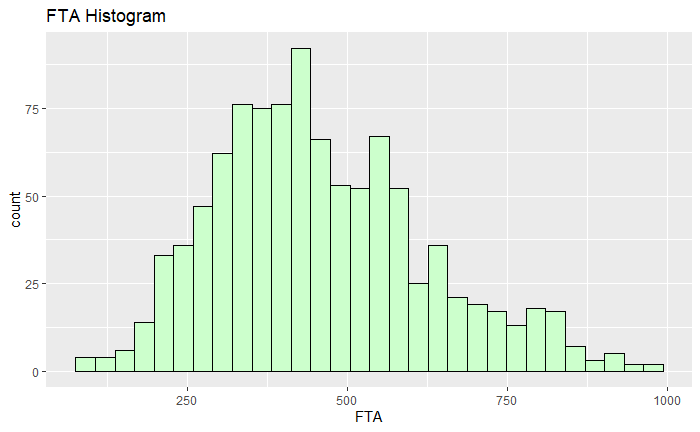
Shown below is Fig. 14 which shows the distribution of each quantitative variable, a great majority of the histograms show a normal distribution or a right skewed distribution. The only anomalies that occur happen in the 3P, 3PA and 3P% charts where the Centers and Forwards that do not shoot 3 pointers with any volume are evident on the left side of each of the histograms. Also due to the interrelated nature of a lot of these metrics some of the pairings between 2PA/ 2P, FGA/FG, etc., are almost identical.

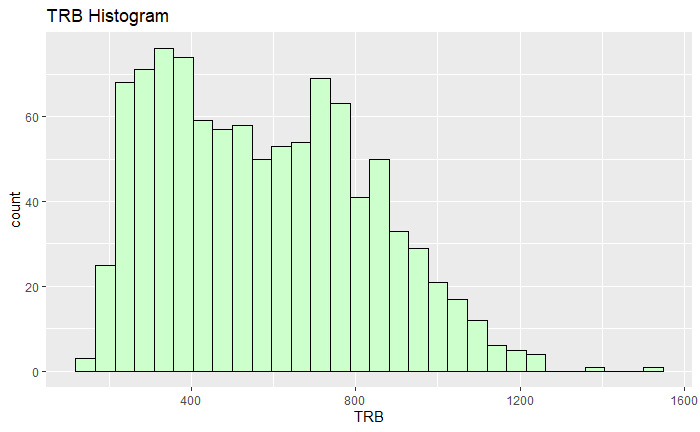
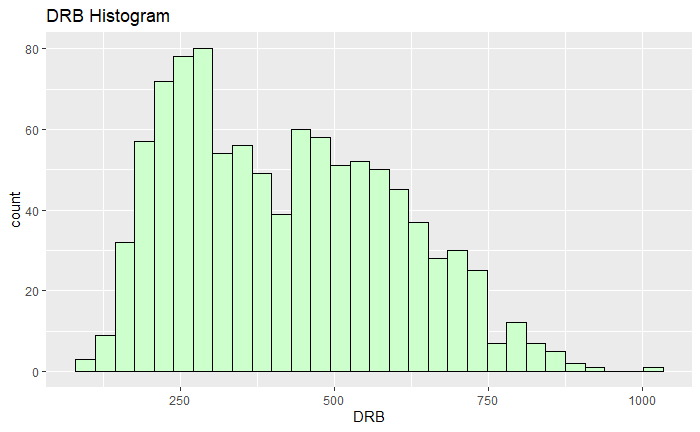


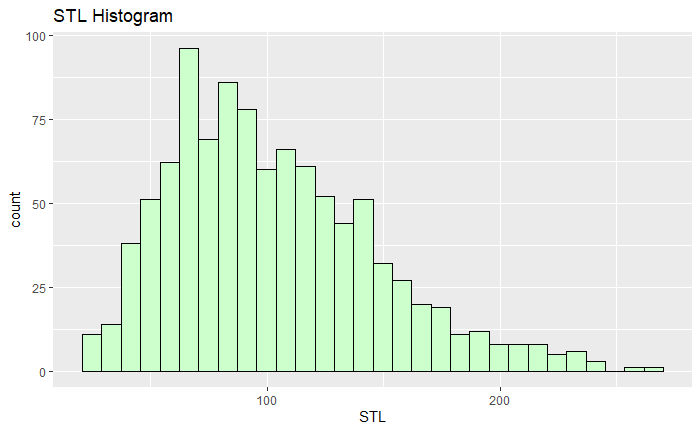
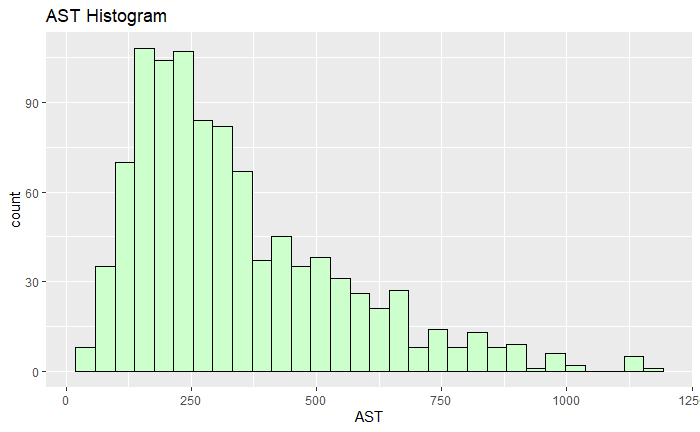


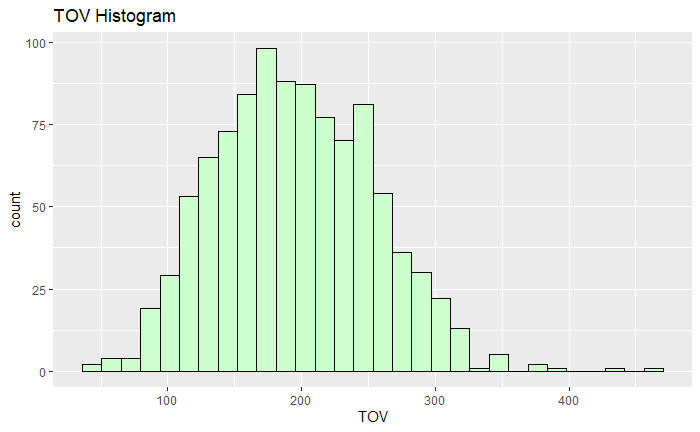
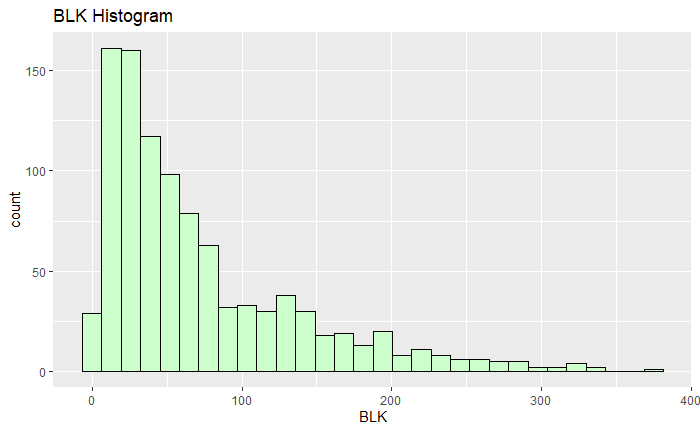


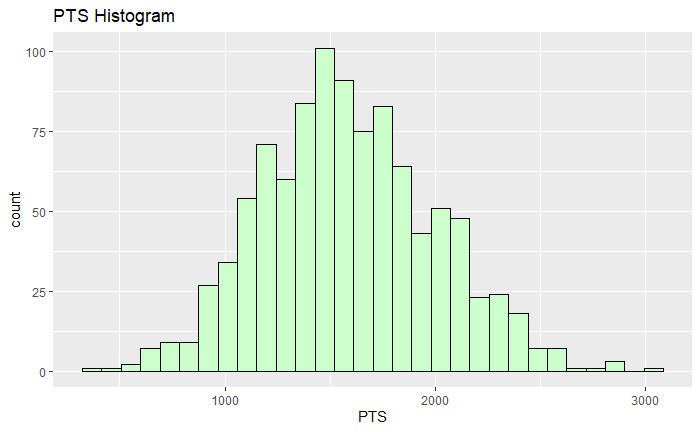
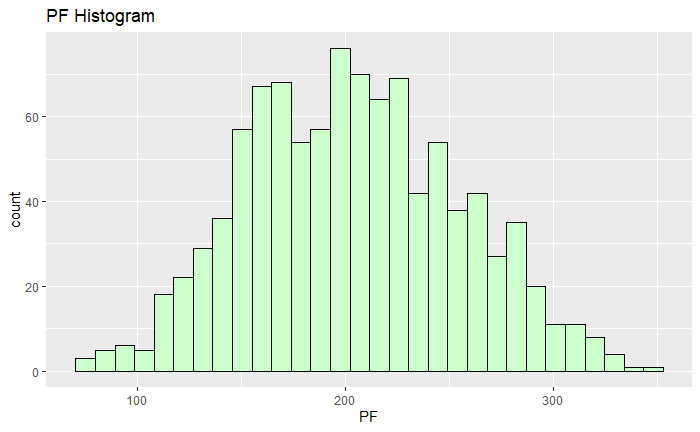


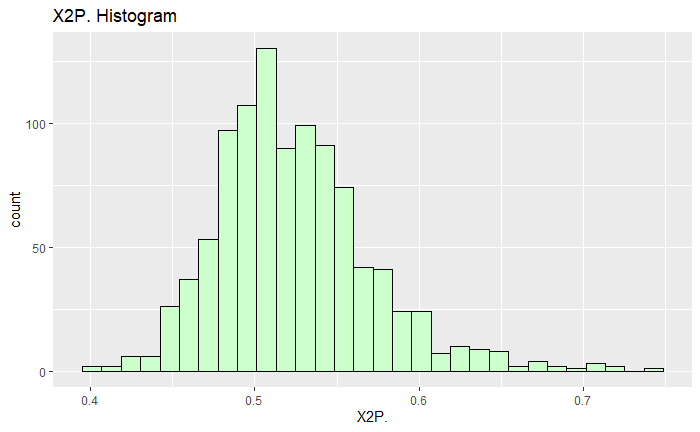
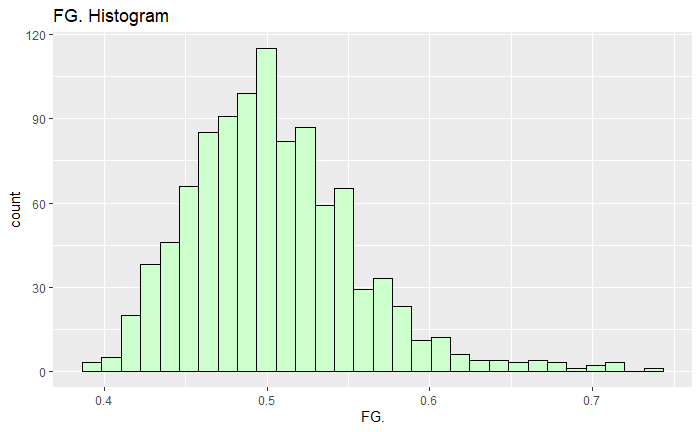


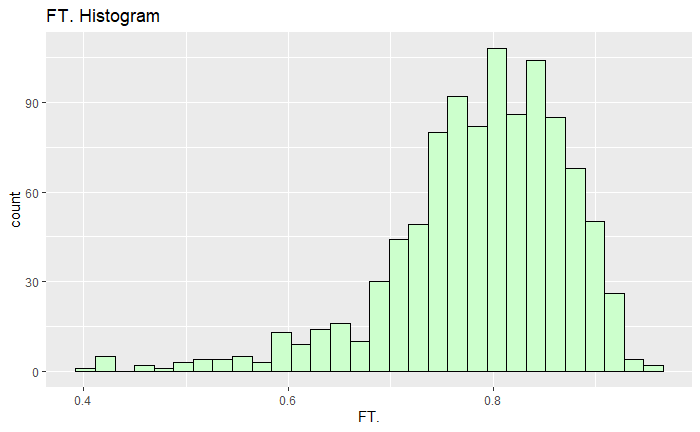
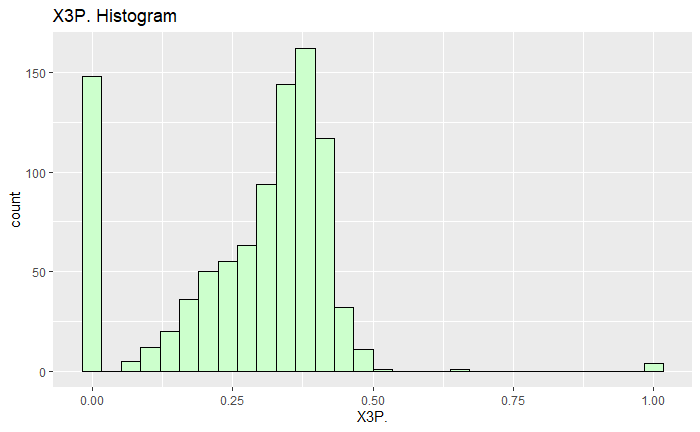


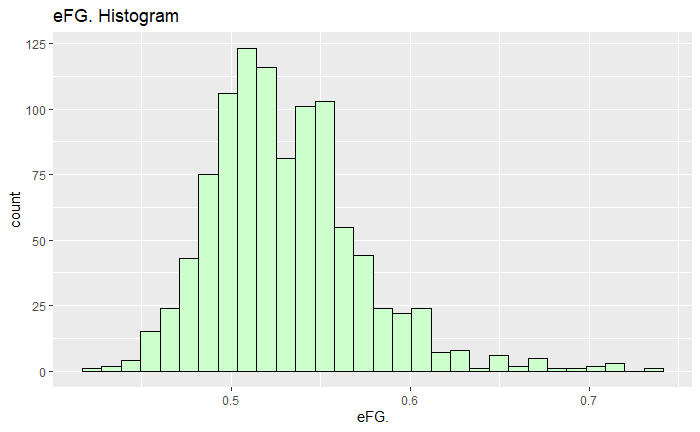
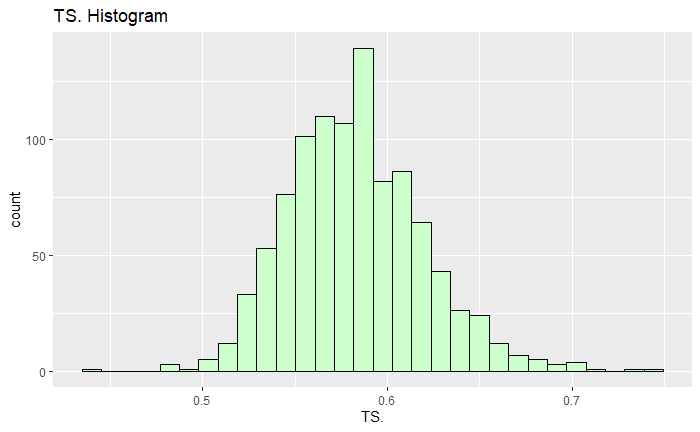


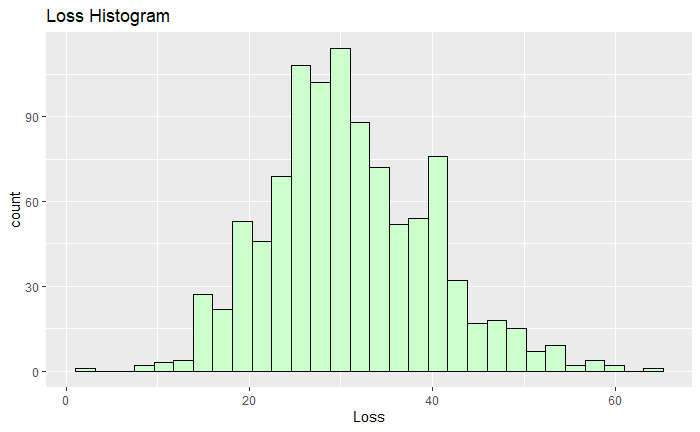
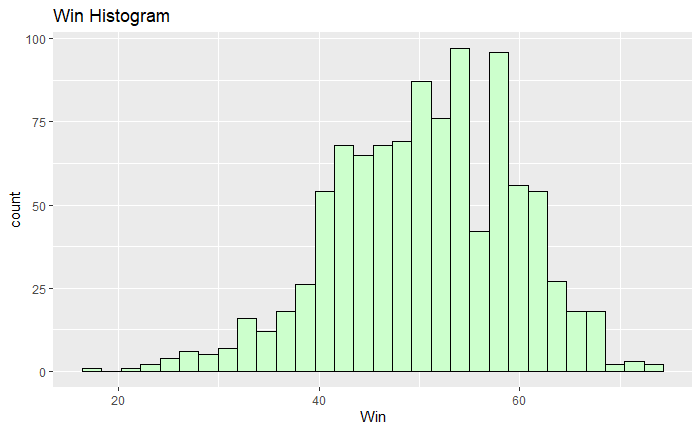












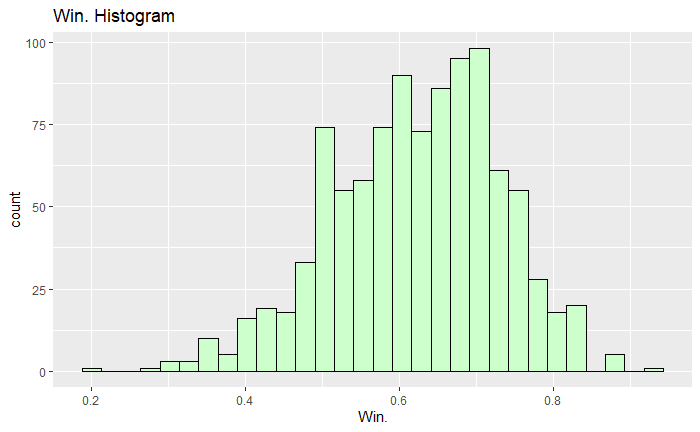


Fig. 15

**Modeling**

In order to find a solution for my first research question: Can you predict a team’s win percentage based on the statistics of an individual player? My initial goal was to eliminate any variables that would not be helpful when put in a model. First, identifier variables like “Player name” and “Rank” were removed from consideration because it did not obtain any pertinent information for analysis. “Season” and “Team” were also removed from consideration because there can only be a net zero for game difference in an entire season for every team and including “Team” could make the model indicate that being on the Lakers or the Celtics is the best predictor for team record. Additionally, some obvious exclusions from variable selection were “Wins”, “Losses” and “Won Championship”, as these variables are inherently too similar to overall team win percentage.

## Multiple Linear Regression

After excluding variables based on intuition, additional variables must be eliminated by way of Multiple Linear Regression (MLR). I will use an MLR model to indicate the relationship between Win percentage and other quantitative variables in the dataset.

Before creating the model, the data was split into a 70/30 training split, where 700 of the observations were in the training set and 300 in the test set.

In the first MLR model every remaining variable is included as they are as listed: “Win Shares”, “Age”, “Games Played”, “Minutes Played” ,”Field Goals”, “Field Goals Attempted”, “2 Pointers made”, “2 Pointers Attempted", "3 pointers made", "3 pointers attempted”, “Free Throws”, “ Free Throws Attempted”, “ Offensive Rebounds”, “Defensive Rebounds”, “ Total Rebounds", "Assists”, “Blocks”, “Turnovers”, “Personal Fouls”, “Points”, “2 point %”, “ Free Throw %”, “True Shooting %”, and “Effective Field Goal %”

Fig. 16 below shows the model summary for the first MLR that was created. There are 10 variables that are marked as significant at the .05 level and they are: Win Shares, Age, Games Played, Minutes Played, 2 pointers attempted, Free Throw Made, Offensive Rebounds, Defensive Rebounds, Assists, Steals, Blocks, Turnovers, and 2 Point %. Also shown in the figure is the absence of 4 variables due to multicollinearity between those variables and others in the model, Total Rebounds was removed because it can be perfectly predicted by Offensive and Defensive rebounds, 3 Pointers made and attempted can be predicted by 3 Point % and Points can be predicted by 2 Pointers made, 3 pointers made, and Free Throws made. The adjusted R-squared value for the first model was 0.3428.

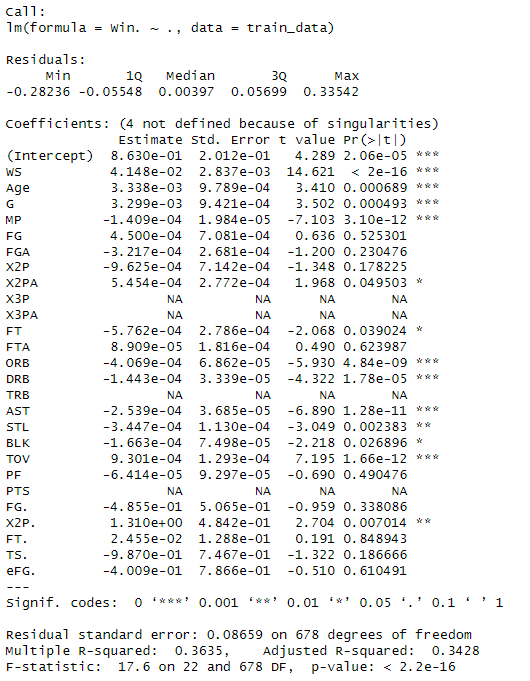


Fig. 16

Next, I used forward and backward stepwise selection to form my next MLR model and the best model that was is shown below as Fig.17. This model has a slightly higher adjusted R-squared value of 0.3466 and the variables that were statistically significant were Win Shares, Age, Games played, Minutes Played, Field Goal Attempts, 2 Pointers made, 2 pointers attempted, Free Throws, Offensive Rebounds, Defensive Rebounds, Assists, Steals, Blocks, Turnovers, 2 Point% and True Shooting %.

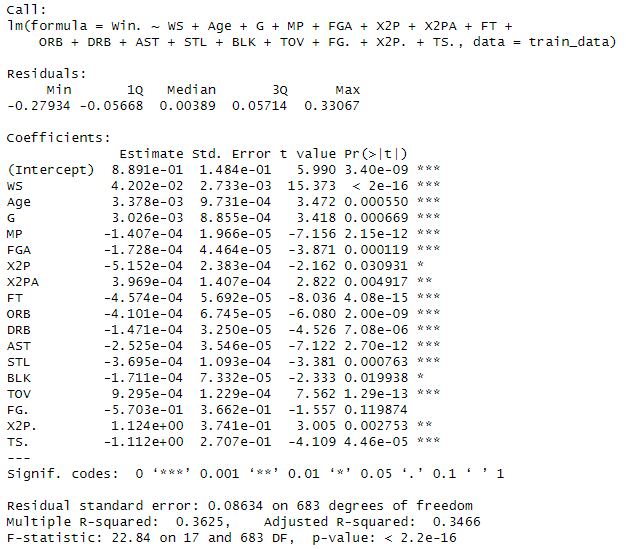


Fig. 17

Despite the adjusted R-squared not being as high as I would like it to consider it an effective model, the residual plots and Root Mean Squared Error for the stepwise model indicate that the model is a good fit for the data. As can be seen in Fig. 18 the residuals are scattered in a random manner which shows that the model is unbiased and homoscedastic. Additionally, in Fig. 19 and Fig. 20 the values roughly follow the line on the Q-Q plot and the density plot is bell shaped indicating normal distribution of the residuals. In terms of RMSE and Normalized RMSE, the values are 0.085 and 0.116 respectively.

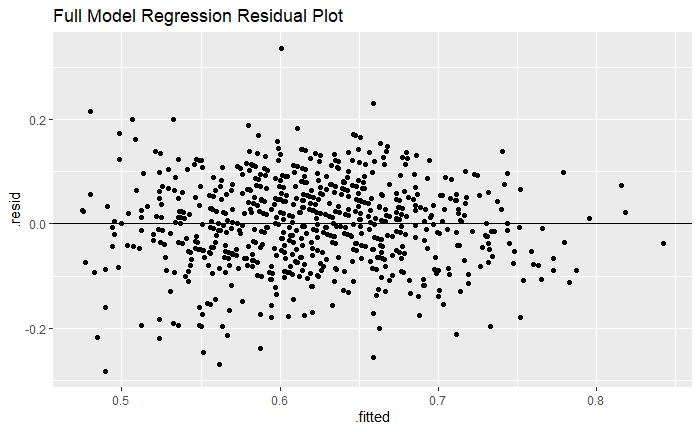


Fig.18

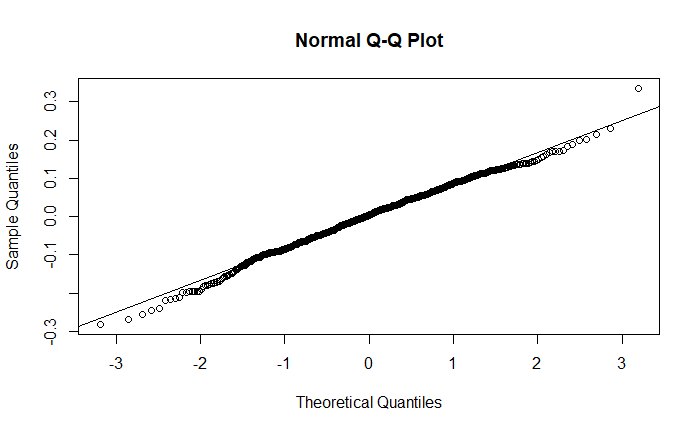


Fig. 19

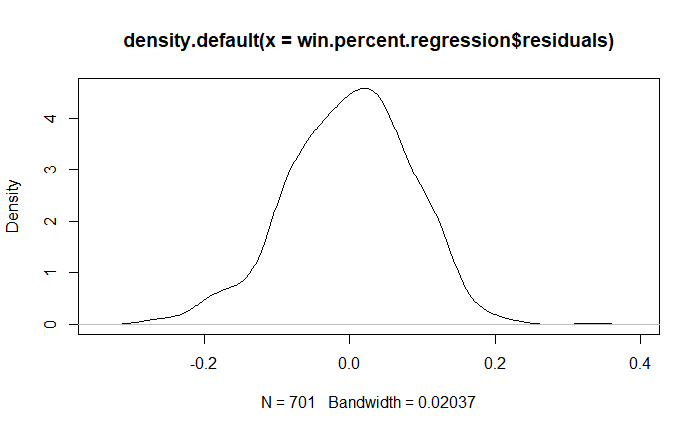


Fig. 20

As a way of visualizing the differences between the predictions made from the model and the actual data, Table 2 shows a comparison of the observed and predicted values.

|  |  |  |
| --- | --- | --- |
| Name/ Year | Predicted Win % | Actual Win % |
| Michael Jordan 1987-88 | 72.1% | 61% |
| Michael Jordan 1995-96 | 85.5% | 87.8% |
| David Robinson 1993-94 | 72.1% | 67.1% |
| Kevin Durant 2012-13 | 73.7% | 73.2% |
| Tim Duncan 2001-02 | 71.4% | 70.7% |
| Dirk Nowitzki 2006-07 | 74.7% | 81.7% |
| LeBron James 2005-06 | 63.7% | 61% |
| Tracy McGrady 2002-03 | 65.2% | 51.2% |
| Magic Johnson 1988-89 | 66.5% | 69.5% |

*Table 2*

## Decision Tree

To attempt to answer the following question: Can you predict if a team will win an NBA championship based on individual statistics? I created a Classification Decision Tree model to try to predict “Champ.” which denotes if a given player won a championship that season. The predictor variables used in this model are “'Win Shares', 'Age', 'Games Played', 'Minutes Played', 'Field Goals', 'Field Goal Attempts', '2 Pointers Made','2 Pointers Attempted', '3-Pointers Made', '3 Pointers Attempted ', 'Free Throws Made', 'Free Throws Attempted', 'Offensive Rebounds', 'Defensive Rebounds', 'Total Rebounds', 'Assists', 'Steals', 'Blocks', 'Turnovers', 'Personal Fouls', 'Points', 'Field Goal %', '2 Point % ','3 Point %', 'Free Throw %', 'True Shooting %.', and 'Effective Field Goal %'.

Before creating the model, I had to use an educated Training and Test split based on the variable of “Champ?” sets using a 70/ 30 split, there will be 700 observations in the training set and 300 in the test set. An educated split was utilized so that the distribution of yes and no observations in the “Champ?” column would not be disrupted. Fig. 21 shows the best Classification Decision Tree model that was created. There are 5 Decision nodes including the Root node and 6 total Leaf nodes. The variables that are of focus in the tree are Minutes Played, Win Shares, and Field Goal Attempts.

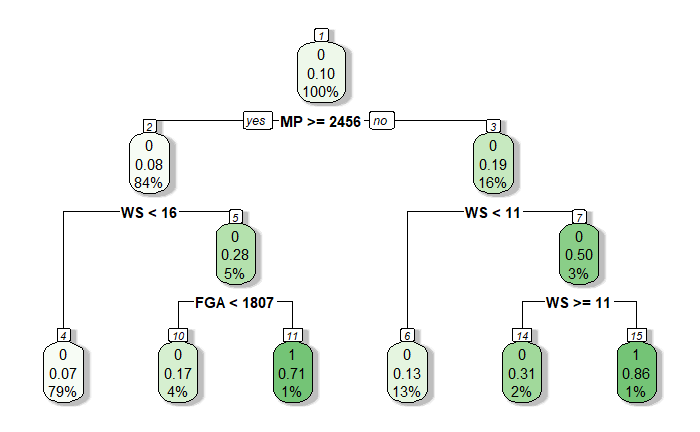


Fig. 21

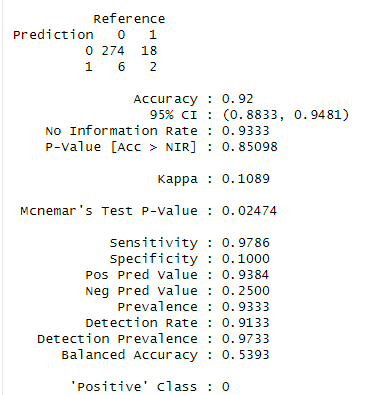


Fig. 22

When looking at the Confusion Matrix summary in Fig.22 above, it may initially look like an effective mode[[1]](#footnote-1)l because of the extremely high accuracy that it outputs, but when looking at other aspects of the Matrix it is shown that it only successfully predicted one instance of a championship being won. The large majority of successful predictions come from True Negative predictions and that can also be explained by the prevalence statistic of 0.93. Additionally, the Kappa value is a staggeringly low 0.109 which indicated a low level of agreement between the actual and predicted values. Additionally, to further indicate the ineffectiveness of this model an ROC curve plot was created and is shown in Fig. 23. The plot shows that the “curve” is very close to the reference line and the AUC value is 0.54, indicating that it would be no better than randomly guessing outcomes.

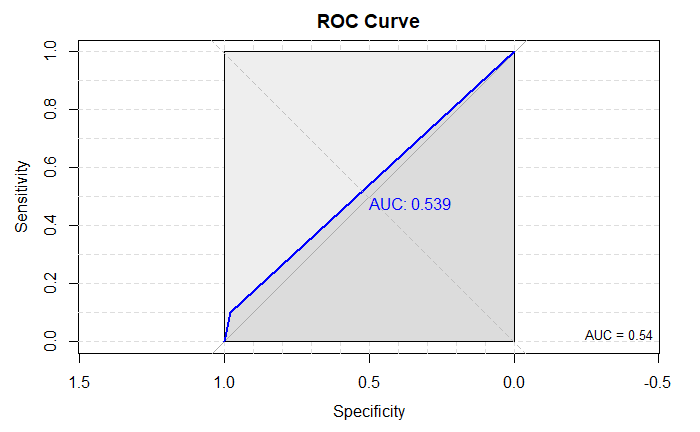


Fig. 23

To visualize the difference between predicted and actual outcomes of Championship seasons, Table 3 was created below. The table shows a great “ability” to predict if a player didn’t win a championship and many multiple false negatives. However, 1 of the 2 two correct champion predictions is shown with that being LeBron James in 2011-2012 with the Miami Heat.

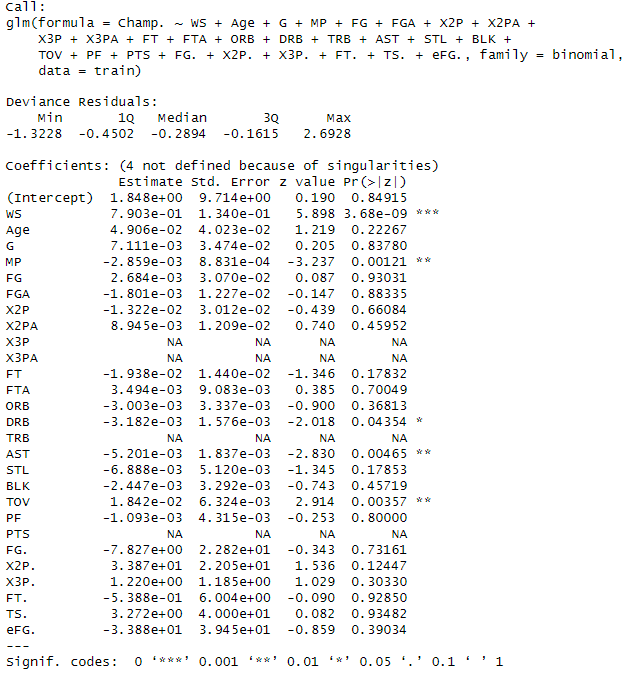
|  |  |  |
| --- | --- | --- |
| Name/ Season | Predicted Championship? | Won Championship? |
| Michael Jordan 1989-90 | No | No |
| Shaquille O’Neal 1999-00 | No | Yes |
| Michael Jordan 1996-97 | No | Yes |
| Stephen Curry 2015-2016 | No | No |
| Charles Barkley 1989-90 | No | No |
| Dirk Nowitzki 2005-06 | No | No |
| Tim Duncan 2002-03 | No | Yes |
| Tracy McGrady 2002-03 | No | No |
| LeBron James 2011-2012 | Yes | Yes |

*Table 3*

## Logistic Regression

As an attempt to create a model that would be better at classifying the target variable of “Champ?”, a logistic regression model was created. Logistic regression models excel in binary classification and would be perfect for this task. As with the decision tree model, an educated training split was created to retain the structure of the data with 700 observations being included in the training and 300 observations appearing in the test set.

A logistic regression model was created to predict “Champ?” using the following variables: 'Win Shares', 'Age', 'Games Played', 'Minutes Played', 'Field Goals', 'Field Goal Attempts', '2 Pointers Made','2 Pointers Attempted', '3-Pointers Made', '3 Pointers Attempted ', 'Free Throws Made', 'Free Throws Attempted', 'Offensive Rebounds', 'Defensive Rebounds', 'Total Rebounds', 'Assists', 'Steals', 'Blocks', 'Turnovers', 'Personal Fouls', 'Points', 'Field Goal %', '2 Point % ','3 Point %', 'Free Throw %', 'True Shooting %.', and 'Effective Field Goal %'. Fig. 24 below shows the aforementioned model. It can be seen that the variables that were significant at the .05 level in the model are “Win Shares”, “Minutes Played”,” Defensive Rebounds”, “Assists” and “Turnovers”. Additionally, when looking at the confusion matrix, it suffers from the same issues as the earlier decision tree of very high sensitivity and very low specificity.



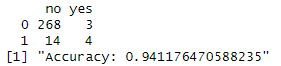


Fig. 24

As shown in Fig. 25, the ROC curve plot shows that it closely follows the reference line with an AUC value of 0.58, indicating an effective model was not created.

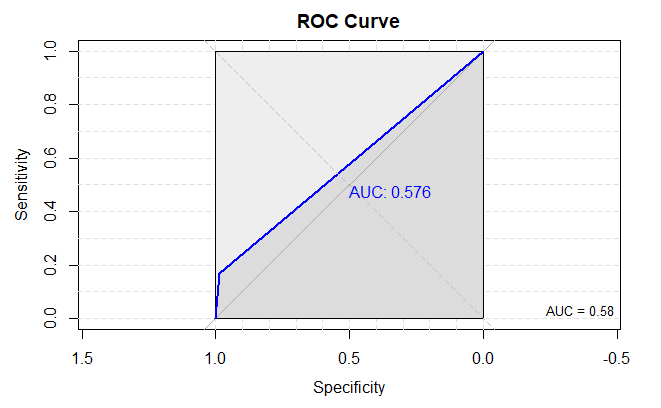


Fig. 25

One more logistic regression model was created using the variables that were significant in the last model to try another avenue and as shown in Fig. 26, it performs even worse than the previous 2 models with major issues with specificity.

A screenshot of a computer

Description automatically generated

A number with black text

Description automatically generated

Fig. 26

## LASSO Regression

In order to answer the question: Which factors in the current dataset contribute the most to Win Shares? I created a LASSO regression model. LASSO, standing for Least Absolute Shrinkage and Selection Operator is a regression model that is very effective at variable selection by shrinking some coefficients to zero and only keeping the variables with the strongest relationship to the target variable (Kim, 2019).

The variables that were fed into the model were , 'Age', 'Games Played', 'Minutes Played', 'Field Goals', 'Field Goal Attempts', '2 Pointers Made','2 Pointers Attempted', '3-Pointers Made', '3 Pointers Attempted ', 'Free Throws Made', 'Free Throws Attempted', 'Offensive Rebounds', 'Defensive Rebounds', 'Total Rebounds', 'Assists', 'Steals', 'Blocks', 'Turnovers', 'Personal Fouls', 'Points', 'Field Goal %', '2 Point % ', 'Free Throw %', 'True Shooting %.', and 'Effective Field Goal %'. Wins, Losses, and Win Percentage were left out of consideration because there would be too significant of a relationship between those variables and Win Shares. 3 Point Percentage was also taken out of the model due to the existence of null values; the model did not work properly when dealing with this issue.

Before creating the model, the data was split into Training and Test sets using a 70/ 30 split, with 700 observations in the training and 300 observations in the test set. After running the model it was determined that the variables that influence Win Shares the most are ‘Age’, ‘Games Played’, 'Minutes Played', 'Field Goal Attempts', , '3-Pointers Made', '3 Pointers Attempted ', 'Free Throws Made', 'Free Throws Attempted', 'Defensive Rebounds', 'Total Rebounds', 'Assists', 'Steals', 'Blocks', 'Turnovers', 'Personal Fouls', 'Points', '2 Point % ', 'True Shooting %.', and 'Effective Field Goal %', as shown in Fig. 26.

Additional testing was done to ensure that the model is effective. Similar to the Linear model created to predict win percentage, a Residual Plot, Q-Q Plot and Density plot of Residuals was created as shown in Fig. 27 and Fig. 28. The residual plot shows a random scattering of residuals, the Q-Q plot closely follows the plotted line, and the Density Plot displays a somewhat normal distribution all indicating that the model performs well. Also, the R-squared and adj. R-squared are .70 and .67 respectively, also indicating an effective model was formed.

A screenshot of a computer

Description automatically generated

Fig. 27

A graph of a plot

Description automatically generatedA graph of a normal q-q plot

Description automatically generated

Fig. 28

A graph of a function

Description automatically generated

Fig. 29

**Conclusion**

## Discussion

Through the analysis, the research questions have been answered:

1. **Can you predict a team’s win percentage based on the statistics of an individual player?**

For the most part, no, the models that were created were unreliable and did not indicate a possibility of accurate prediction of win percentage through an individual player’s counting statistics. If the dataset I used was more comprehensive and exhaustive with the inclusion of many more advanced statistical metrics, it may be possible to use it to predict an NBA team’s record. The model did seem to be somewhat accurate in terms of predicting some of the top star’s win percentage in the test set, but it was not reliable enough to give dependable answers. With the information that was utilized in this study it is safe to assume that it is not possible to accurately predict the win percentage of an NBA team based on one player’s statistics.

2. **Can you predict if a team will win an NBA championship based on individual statistics**?

Based primarily on the relative impossibility for the classification decision tree model to accurately predict a “yes” on the binary variable of “Won Championship?”, I think it is clear to assume that the answer to this question is no. It would be interesting to see if the data had more of an even split of players that had won a championship and players that had not won a championship, if the model would have any better predictive capabilities. As the data remains, it is not possible to predict a team’s championship based on an individual player's counting statistics.

3. **Which factors contribute to Win Shares the most in the given dataset**?

The factors that contribute to Win Shares in a predictive capacity the most are: ‘Games Played’, 'Minutes Played', 'Field Goal Attempts' 'Assists', 'Steals', 'Blocks', 'Turnovers', 'Personal Fouls', 'Points', '2 Point % ', 'True Shooting %.', and ‘Effective Field Goal %'. These factors explain nearly 73% of the variance when predicting Win Shares.

## Applications

The information that was provided and the analytical process that were utilized in this paper are certainly generalizable to other situations within the realm of basketball as well as other sports. This same study could be used to determine how much the statistics of a single NFL Quarterback translates to team success and the same goes for a pitcher in the MLB.

In terms of the NBA and the analysis of a single players statistics, this could serve as a tool for the analytics team for an NBA organization to help make a decision to trade multiple assets and future draft picks for one star player. The data that is available indicates that relying on one player to carry the load for an entire NBA team is not feasible and teams should not break the bank to sign or mortgage the future to trade for an established superstar player.

Using the results of this study could also assist NBA pundits in situations in which a player’s Win Shares are brought up, to help them really understand what goes into an NBA Win Share.

## Limitations

There were certainly limitations that should be considered when looking at the results of this study. The quality of coach, nor the overall quality of the organization was able to be measured in this scenario, likely playing a major role in the team’s record and championship wins. These qualities are difficult to quantify and are definitely not included in any box score you will see in an NBA game. Additionally, a player’s health was also not a factor during analysis as a player could have been playing through an injury and it hampered their statistics or the injury could have occurred during the playoffs, limiting their team’s ability to win a championship. Team synergy is also a factor in team success that cannot be numerically quantified and thus was not included in the study. Infinitely many things go into making a team successful and all of those things could not be possibly studied in this manner.

Outside of intangible factors, there are a heap of statistical measures to track a player’s performance and many of those were not included or mentioned in this study. Including the hundreds of statistical metrics in the dataset could have yielded better predictive results.

## Future Research

As long as professional basketball still exists, there will always be concerns about ways to optimize play and discussions to be had over the game itself. Analytical research into the world of sports is necessary to find the truth about the games many of us love. There are probably hundreds of statistical techniques that have yet to be discovered and could change the entire sport as we know it. Additional research will always take place as long as there is a sport to research.

References

Basketball Reference. (n.d.). Win Shares. Basketball Reference. <https://www.basketball-reference.com/about/ws.html>

Cantwell, J. D. (2004). The physician who invented basketball. The American Journal of Cardiology, 93(8), 1075-1077. <https://doi.org/10.1016/j.amjcard.2003.12.068>

Cohan, A., Schuster, J., & Fernandez, J. (2021). A Deep Learning Approach to Injury Forecasting in NBA Basketball. Journal of Sports Analytics, Vol. 7 Issue 4

Devlin, S., & Uminsky, D. (2020). Identifying group contributions in NBA lineups with spectral analysis. Journal of Sports Analytics, 6(3), 215-234. <https://doi.org/10.3233/JSA-200407>

Erčulj, F., & Štrumbelj, E. (2015). Basketball shot types and shot success in different levels of competitive basketball. PLoS One, 10(6), e0128885. doi:10.1371/journal.pone.0128885

Freitas, L. (2021). Shot distribution in the NBA: did we see when 3-point shots became popular?. *Ger J Exerc Sport Res* **51**, 237–240 <https://doi.org/10.1007/s12662-020-00690-7>

Grijalva, E., Maynes, T. D., Badura, K. L., & Whiting, S. W. (2020). Examining the "I" in Team: A Longitudinal Investigation of the Influence of Team Narcissism Composition on Team Outcomes in the NBA. Academy of Management Journal, 63(1), 7-33. https://doi.org/10.5465/amj.2017.0218

Humphreys, B. R., & Johnson, C. (2020). The Effect of Superstars on Game Attendance: Evidence From the NBA. Journal of Sports Economics, 21(2), 152–175. <https://doi-org.libproxy.library.unt.edu/10.1177/1527002519885441>

Kim, J. (2021, June 10). Variable Selection Using Lasso. Towards Data Science. <https://towardsdatascience.com/variable-selection-using-lasso-493ac2e5660d>

NBAstuffer. (n.d.). Points Produced. NBAstuffer. https://www.nbastuffer.com/analytics101/points-produced/

Nutting, A. W. (2010). Individual Tournament Incentives in a Team Setting: The 2008-09 NBA MVP Race.*International Journal of Sport Finance, 5*(3), 208-221. <https://libproxy.library.unt.edu/login?url=https://www.proquest.com/scholarly-journals/individual-tournament-incentives-team-setting/docview/751424726/se-2>

Pai, P.-F., ChangLiao, L.-H., & Lin, K.-P. (2017). Analyzing basketball games by a support vector machines with decision tree model. Neural Computing & Applications, 28(12), 4159-4167. <https://doi.org/10.1007/s00521-016-2321-9>

Ranstam, J., & Cook, J. A. (2018). LASSO regression. British Journal of Surgery, 105(10), 1348. <https://doi-org.libproxy.library.unt.edu/10.1002/bjs.10895>

Reilly, P., Solow, J. L., & von Allmen, P. (2023). When the Stars Are Out: The Impact of Missed Games on NBA Television Audiences. Journal of Sports Economics, 0(0). https://doi.org/10.1177/15270025231174616

Rollins, K. (2018, May 29). *Rockets 27 straight missed 3S: Houston breaks NBA playoff record ...* Sports Illustrated. <https://www.si.com/nba/2018/05/29/rockets-break-playoff-record-27-consecutive-missed-three-pointers>

Sportskeeda. (2022, June 29). NBA lockouts: Why do they happen and how did we come to it? https://www.sportskeeda.com/basketball/nba-lockouts-why-happen-came-it

Staffo, D. F. (1998). The development of professional basketball in the united states, with an emphasis on the history of the NBA to its 50th anniversary season in 1996-97.*Physical Educator, 55*(1), 9. Retrieved from <https://libproxy.library.unt.edu/login?url=https://www.proquest.com/scholarly-journals/development-professional-basketball-united-states/docview/232989086/se-2>

Zhang, S., Gomez, M.Á., Yi, Q., Dong, R., Leicht, A., & Lorenzo, A. (2020). Modelling the Relationship between Match Outcome and Match Performances during the 2019 FIBA Basketball World Cup: A Quantile Regression Analysis. International Journal of Environmental Research and Public Health, 17(16), 5722. https://doi.org/10.3390/ijerph17165722

1. In Fig.22 “Sensitivity” and “Specificity” should be swapped and have opposite values [↑](#footnote-ref-1)