ProjectPart2A.PreliminaryAnalysis

Suren Bhakta, Arnav Jain, Jack McPherson

2023-03-12

Title: The influence of usage rate and win percentage of a NBA player's individual win percentage.

Introduction

The main goal of this study is to investigate whether there exists a relationship between the combined factors of age and usage rate of specific NBA players and their team's winning percentage. This study aims to see if there is a prime age for players and an ideal usage rate for a player in the NBA in order to maximize their team's winning percentage. For a little information of usage rate, it is a metric calculation of various statistics to describe a players impact while playing. The calculation is the ratio of of an individuals stats over teams stats. For more information and the full equation, visit https://www.nbastuffer.com/analytics101/usage-rate/ (https://www.nbastuffer.com/analytics101/usage-rate/). With this, the study will provide insights into factors, such as player and team strategy, that impact the success of NBA teams.

Age vs. Winning Percentage

We expect to find a couple of specific ages where the winning percentages of teams are maximized. We expect to see a moderately strong relationship between age and winning percentage because we hypothesize that age and experience lead to creating a winning culture in the NBA.

Article 1: https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/# (https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/#)

This article discusses the peak age in a couple of sports, including basketball. The article hypothesized that the prime for most players in the NBA is about 27-28 years old, but cites how ages for primes of players will differ depending on their play style and position. This relates to our study as if a player is in their prime, there exists a higher chance of their team being better as compared to other years, leading to them winning more.

Article 2: https://towardsdatascience.com/the-asset-of-age-e4b45599ea94 (https://towardsdatascience.com/the-asset-of-age-e4b45599ea94)

This article analyzes player development and shows how age is the best asset any NBA player can have. This shows how teams should look towards drafting younger players in order for them to have more time to develop into their primes as compared to older players. This uses Player Efficiency Rating (PER) to show the progression of players. This relates to our study as we assume that teams will win more when their players progress.

Article 3: https://courses.cs.washington.edu/courses/cse163/20su/files/project/archive/nba.pdf (https://courses.cs.washington.edu/courses/cse163/20su/files/project/archive/nba.pdf)

This article uses factors such as height, weight, and position to determine the peak age of NBA players. They found that the average age of All-Star players is 26.5 years and the average age of MVP players is 27.9 years. They used a regression equation to model the decline of a player and therefore, team performance.

Usage Rate vs. Winning Percentage

We expect to find a couple of high usage rates for players on teams where the winning percentages are high. We expect to see a moderately strong relationship between usage rate and winning percentage because we hypothesize that teams that have a higher winning percentage have a couple of superstar players whose usage rates are very high. Article 1: https://squared2020.com/2017/12/02/usage-and-efficiency/ (https://squared2020.com/2017/12/02/usage-and-efficiency/)

This article helps understand the statistic of usage rate and how this stat can be translated into predicting player performance throughout the course of a season, which can then be used to predict a team's performance during the season.

Article 2: https://sports.sites.yale.edu/game-team-statistics-nba (https://sports.sites.yale.edu/game-team-statistics-nba)

This article shows which statistics are correlated with a team's win percentage. This shows all the simple statistics that are available to be seen in a box score, such as points, rebounds, assists, blocks, and steals, but not usage rate. However, all of these statistics are used to calculate the usage rate of a player.

Article 3: https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf (https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf)

This article predicts the regular season performance of NBA teams using a regression model of basketball statistics. Like the previous article, this study uses all commonly known basic statistics to predict the winning percentage of teams during the course of the season. It does not explicitly use usage rate, but usage rate can be calculated using all of these well-known statistics.

Age vs. Usage Rate

We expect to see that as age increases, the usage rate of the player also increases to a degree. We expect to see a moderately strong relationship between age and winning percentage because we hypothesize that age and experience lead to being used more on the court in the NBA.

Article 1: https://page.byu.edu/docs/files/Publications/JQASSelfArchive.pdf (https://page.byu.edu/docs/files/Publications/JQASSelfArchive.pdf)

This article plots different trajectories of usage rates of NBA players and/or game scores for players over the course of their careers differing by position.

Article 2: https://harvardsportsanalysis.org/2017/11/what-happens-to-nba-players-when-they-age/ (https://harvardsportsanalysis.org/2017/11/what-happens-to-nba-players-when-they-age/)

This article shows what happens to NBA players when they age by showing the trajectory of their key stats as they age during their careers. It models usage rate itself and rather models simple stats that go into the calculation for usage rate as well as many other statistics useful for understanding a player's impact on the court.

Article 3:
https://www.researchgate.net/publication/343916724_Effects_of_age_on_physical_and_technical_performance_in_National_Basketball_Association_NBA_p
(https://www.researchgate.net/publication/343916724_Effects_of_age_on_physical_and_technical_performance_in_National_Basketball_Association_NBA_

This article discusses how age affects performance and efficiency in the NBA by looking at factors such as distance covered, average speed, minutes played, points scored, and playing efficiency.

Data Collection Summary

Our sample subjects was the NBA players from the 2021-2022 season who averaged over 8 minutes per game.

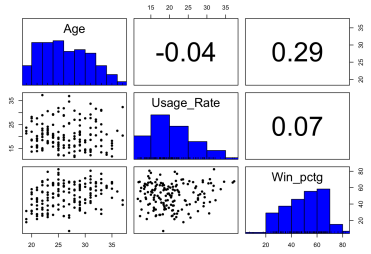
The overall player data was downloaded from the database of the 2021-2022 season via https://www.basketball-reference.com/leagues/NBA_2022_per_game.html (https://www.basketball-reference.com/leagues/NBA_2022_per_game.html). This was then used to filter out individulas who had less than 8 minutes per game. We then used https://www.statmuse.com/questions (https://www.statmuse.com/questions) to manually upload the usage rate and individual win percentage of each individual player from the same respective season. At the end, we had a sample size 160 individual players and no individual cases had to be removed from the study, luckily.

Predictors (two numeric variables) and one numeric outcome

Correlation matrix among three variables (two numeric and the response variables)

To find the correllation matrix, we first must drop all non numeric variables and only keep the two numeric predictors and the response variable.

```
numSDS <- SDS324Eprojectdatacsv[-c(1,2,5)]
pairs.panels(SDS324Eprojectdatacsv[c("Age","Usage_Rate","Win_pctg")],
             method = "pearson",
             hist.col = "blue",
             smooth = FALSE, density = FALSE, ellipses = FALSE)
```

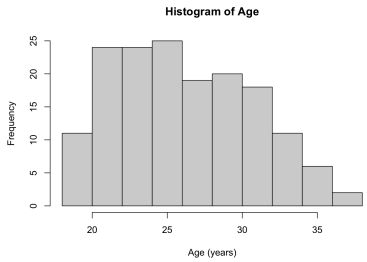


Center, spread, min, max for all variables

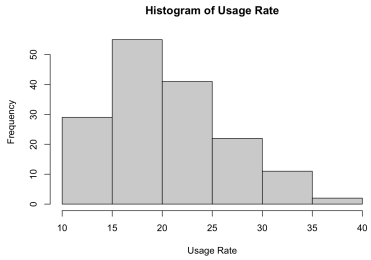
Before proceeding with reporting certain statistics, thedistribution must be checked to determine what would be best to measure center, spread, and etc..

A histogram can be used to determine distributions

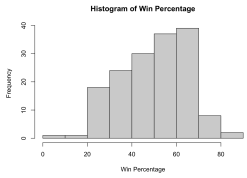
```
hist(numSDS$Age, xlab = "Age (years)", main = "Histogram of Age")
```




```
hist(numSDS$Usage_Rate, xlab = "Usage Rate", main = "Histogram of Usage Rate")
```




```
hist(numSDS$Win_pctg, xlab = "Win Percentage", main = "Histogram of Win Percentage")
```




```
summary(numSDS)
```

```
##       Age          Usage_Rate       Win_pctg
##  Min.   :19.00   Min.   :11.50   Min.   : 7.10
##  1st Qu.:23.00   1st Qu.:16.15   1st Qu.:39.05
##  Median :26.00   Median :19.50   Median :52.25
##  Mean   :26.62   Mean   :20.49   Mean   :50.28
##  3rd Qu.:30.00   3rd Qu.:23.70   3rd Qu.:62.60
##  Max.   :37.00   Max.   :37.30   Max.   :82.40
```

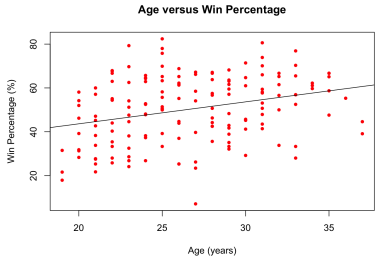
As we can see none of our variables are not normally distributed. This means that the best measures of center and spread would be median and the IQR, calculated as the difference of 3rd Qu - 1st Qu. Thus for Age the center is 26.00 years, the spread being 13.00 years, the minimum being 19 years, and the max being 37 years. The reported statistics for Usage rate as a percentage goes as follows: center 19.50%, spread being 7.55%, minimum being 11.50%, and maximum being 37.30%. The reported statistics for Win_rate, also calculated as a percentage, goes as follows: center 52.25%, spread 23.55%, minimum 7.10% and maximum 82.40%.

Assumption Checking

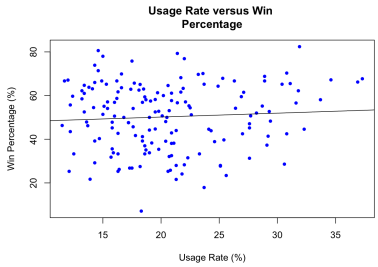
The assumptions we checked were randomness, independence, linearity, normality, and equal variance. For randomness, we took the player pool from the 2021-2022 season and used a random sampling in excel to select our player pool based on if they averaged over 8 minutes per game.

For independence, this was ensured by our resource basketball reference which ensures that each players metrics are independent of one another. For Linearity we wanted to answer for each variable if it is linear with respected to the response variable?

```
players<-SDS324Eprojectdatacsv
plot(players$Age, players$Win_pctg, main='Age versus Win Percentage',
xlab='Age (years)', ylab='Win Percentage (%)', col='red', pch=20)
abline(lm(players$Win_pctg~players$Age))
```



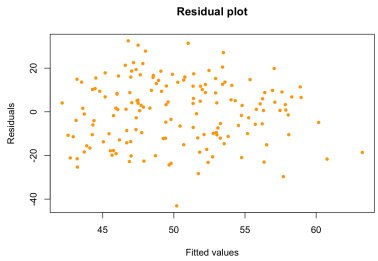

```
plot(players$Usage_Rate, players$Win_pctg, main='Usage Rate versus Win
Percentage', xlab='Usage Rate (%)', ylab='Win Percentage (%)', col='blue', pch=20)
abline(lm(players$Win_pctg~players$Usage_Rate))
```



Both numeric predictors appear to have linear, or at the very least not non-linear, relationships with the response variable.

Equal Variance

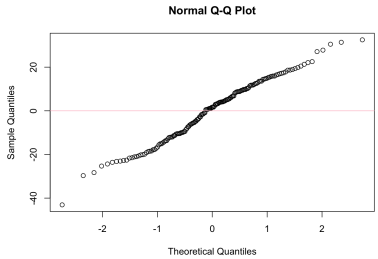
```
mymodel <- lm(players$Win_pctg~players$Age+players$Usage_Rate)
plot(mymodel$fitted.values,mymodel$residuals, main='Residual plot', pch=20, col='orange', xlab='Fitted values', y
lab='Residuals')
```



No funneling or distinct patterns, equal variance assumption passes.

Normality

```
qqnorm(mymodel$residuals)
abline(h=0, col='pink')
```



Normality line does not diverge, thus it passes the assumption as well.