UnemployedFinalReport

Abigail Brown, Michael Shepard, and Abigail Weinstein

2025-12-17

Unemployed :) Final Report

Overview

This project will explore how certain game play and team statistics are related to regular season success in the National Basketball Association (NBA). This will aid in identifying what strategies aid in winning games across the league.

In this project, we aim to explore three primary factors in relation to team success, measured by the number of wins a team has in the regular season. Because each team has 82 games within a season, we can confidently use this as a measure of team success per season without scaling it in any way. The factors being used in this analysis are the following: player age, three-point attempt rate, and pace factor.

Each factor was selected as a means of evaluating different areas that can influence success. *Player age* accounts for factors related to athleticism and skill level within a team. We assume that players who are older will be more skilled due to longer time in the league; it is also possible for the inverse to be true due to physical limitations that arise due to age. *Three point attempt rate* addresses team strategies, focusing on whether teams opt for high-scoring, but more difficult plays over easier, lower-scoring ones. Finally, we assess *pace factor*. This is a more complicated statistic that examines play time and possessions of each team within a game. A high pace factor indicates more possessions and a faster-paced game (with more exchanging of hands overall). This delves further into ideas about the strategies the teams us, as well as the opportunities they are given to score. Collectively, these factors address what often determine how well team plays. In our exploration of these factors in relation to net rating, we will identify how each individual metric is correlated with points scored. From there, we can draw conclusions about how each one effects number of games won.

To do explore and assess these trends, we will utilize data pulled from the Basketball References website. This site has record of game-play statistics for every NBA team going back to the leagues founding in 1948. For the purposes of this exploration, we only selected data from the

2014-2015 season to the 2024-2025 seasons. We did exclude data from the 2019-2020 and 2020-2021 seasons, due to the COVID-19 pandemic impacting games played. Basketball Reference data is, unfortunately, pay-walled for those interested in prolonged use of its databases. This means that it does not fully meet the FAIR and CARE principles discussed in class. However, a team member did have access to the service and opted to source our data from there.

Data Wrangling and Exploration

Before we can begin analysis, it is important to explore our data. To do that, we need to organize it. Data pulled from Basketball Reference was divided by season, with each season having its own .csv file. As such, it was necessary to condense the eight season data sets we are utilizing for this analysis into one. To do so, we included a yearly flag within each data frame so we could later explore yearly totals and identify when these observations were occurring. We then bound all the data frames so that we had a complete data set spanning from 2015 to 2025. The resulting data frame was 270 observations spanning 10 years and 30 teams. Table 1 presents a summary of all primary variables within our analysis across the data set.

```
Attaching package: 'dplyr'

The following object is masked from 'package:kableExtra':

    group_rows

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union



Attaching package: 'ggplot2'

The following objects are masked from 'package:psych':

    %+%, alpha
```

Table 1: Summary Table, NBA Overall

	Data Type	Description	Count	Min. Val.	Median Val.
Team	String	Team Name	270	1.00	15.50
Team Rank	Ordered Num.	Rank in NBA	270	1.00	15.50
Year	Ordered Num.	Season Year	270	2015.00	2019.00
Wins	Numeric	# of Games Won	270	10.00	42.00
3-Point Att.	Numeric	% of Total Attempted Shots as Freethrows	270	0.18	0.36
Pace	Numeric	Avg. Possessions per Game	270	90.40	97.65
Avg. Age	Numeric	Avg. Team Age	270	22.10	26.20

Focusing in on our outcome variable, we note that not one team at any point in our analysis won or lost all 82 games in the regular season. That being said, win count does exist on a fairly wide scale (10 - 73), indicating success for some and abject failure for others. The goal of this analysis is to determine what contributes to this difference across teams. The average and median statistics sit comfortably at ~41 games won. This translates to every team winning, on average, about half their games in the regular season - an expected statistic, but on that requires further exploration nonetheless. Figure 1 shows this exploration. In this distribution, we can see number of season wins in relation to a count of teams that have achieved that amount of wins in a season.

As we can see, the distribution is not perfectly normal, with a high concentration of teams having approximately ~50 winning games a season. This is offset by the long tail of teams winning fewer games. There is a sharp drop off in game wins after this 50 winning game mark. What causes teams to be above this threshold rather than below it?

Data Analysis

As a team, we have decided to explore *age*, *three-point attempt rate,* and *pace factor* as a means for determining team success by number of regular season wins. Upon further analysis, we found that the maximum number of games wins per season is around 50, with only a handful winning more than 60 per season. We investigate these factors see if they contribute to this phenomenon. To do so, we first examine the correlation between each factor and season win count. Table 2 shows the correlation coefficients of each variable of interest, as well as their p-values. Figure 2, Figure 3, show the plotted coefficients.

Table 2: Factor Correlation with Season Win Total

	Correlation Coef. (p)	t-Statistic	p-Value

Player Age	0.498	9.406	0.00
Three-Point Attempts	0.156	2.592	0.01
Pace Factor	-0.063	-1.038	0.30

From this, we see that pace factor has a negligible, slightly negative effect on team wins, with a p-value of .30. This is surprising, as we anticipated pace to have more of positive impact due to its relation to game play time and number of possessions a team has. Also surprising was the very slight effect three-point attempt rate had on team success. While the effect was positive (.156) and significant with a p-value of .01, it did not have the same degree of impact as expected for a statistic that relates to the number of points score.

```
`geom_smooth()` using formula = 'y ~ x'
```

```
`geom_smooth()` using formula = 'y ~ x'
```

Most surprising to us, though, was the relationship identified between age and number of games one. Table 2 indicates that average player age on a team has a correlation coefficient with games won of nearly .500. This is expanded upon in Figure 4, where we see a clear trend in upward in wins as age increases. Age was the variable expected to be least impactful within the selected factors, as it theoretically has both positive and negative impacts on game play. This, we assumed, would result in a net zero effect. Out of curiosity, we opted to explore this factor more. Figure 5 displays a histogram exploring the distribution of wins across average team ages. The resulting view is aligned with the correlation identified earlier - younger teams have lower scoring seasons compared to older teams. It should be noted that it there is a decline that begins once team average hits 30 years of age, which also happens to be when players begin to retire. This indicates to us that peak performance is of a team occurs with an a mix of some young players and number of older, more experienced ones.

Distribution of NBA Team Wins

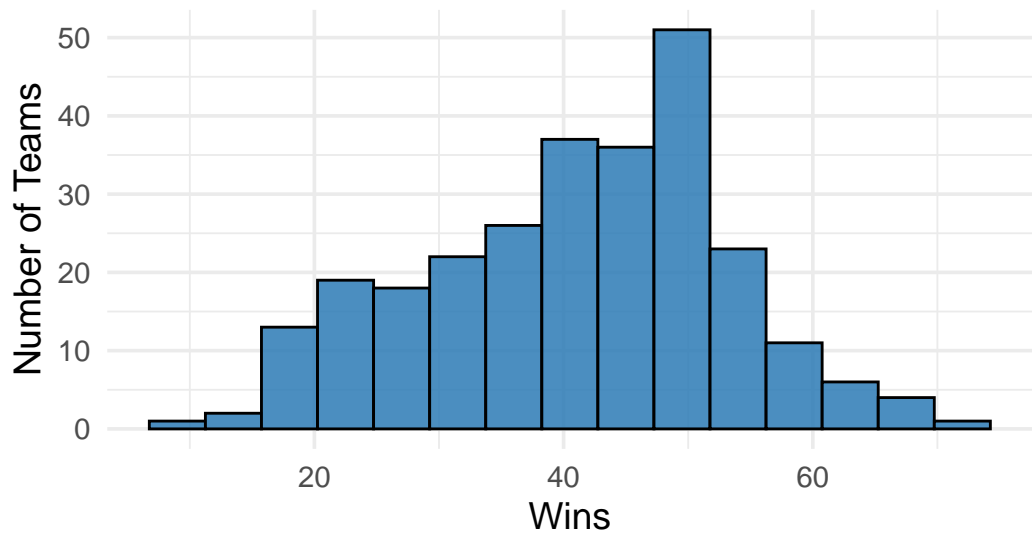NBA Non–COVID Seasons (2015–2019, 2022–2025)

Figure 1: NBA Team Win Count

Pace vs Team Wins
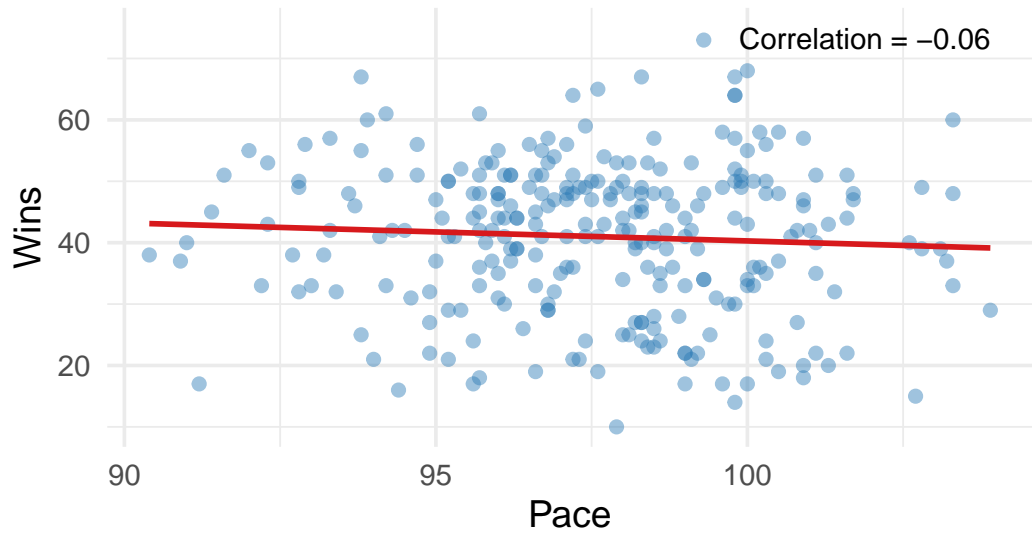
NBA Non–COVID Seasons (2015–2019, 2022–2025)

Correlation = −0.06

Figure 2: Pace Factor and Team Wins

3–Point Attempt Rate vs Team Wins

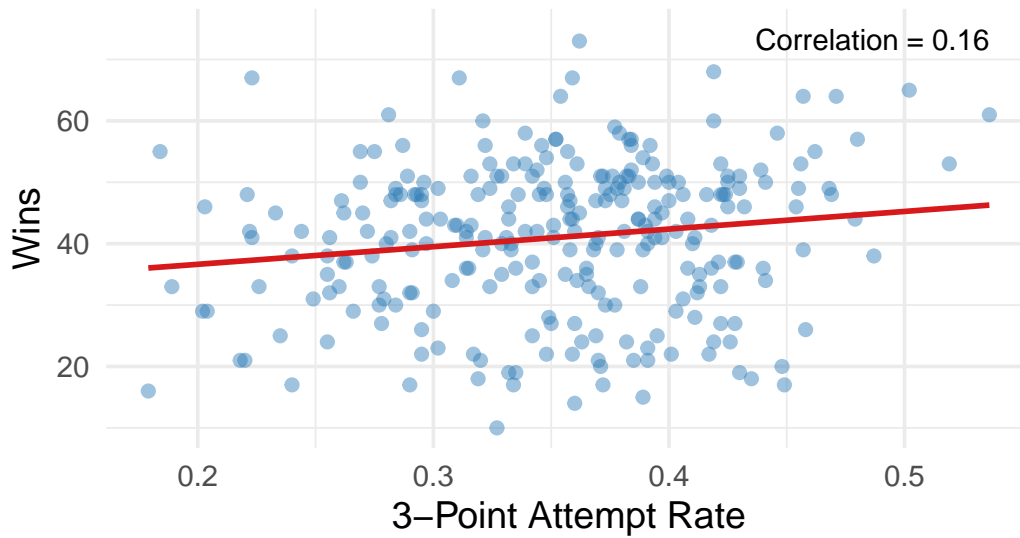NBA Non–COVID Seasons (2015–2019, 2022–2025)

Correlation = 0.16

Figure 3: Three-Point Attempt Rate and Team Wins

Average Wins by Team Age Group

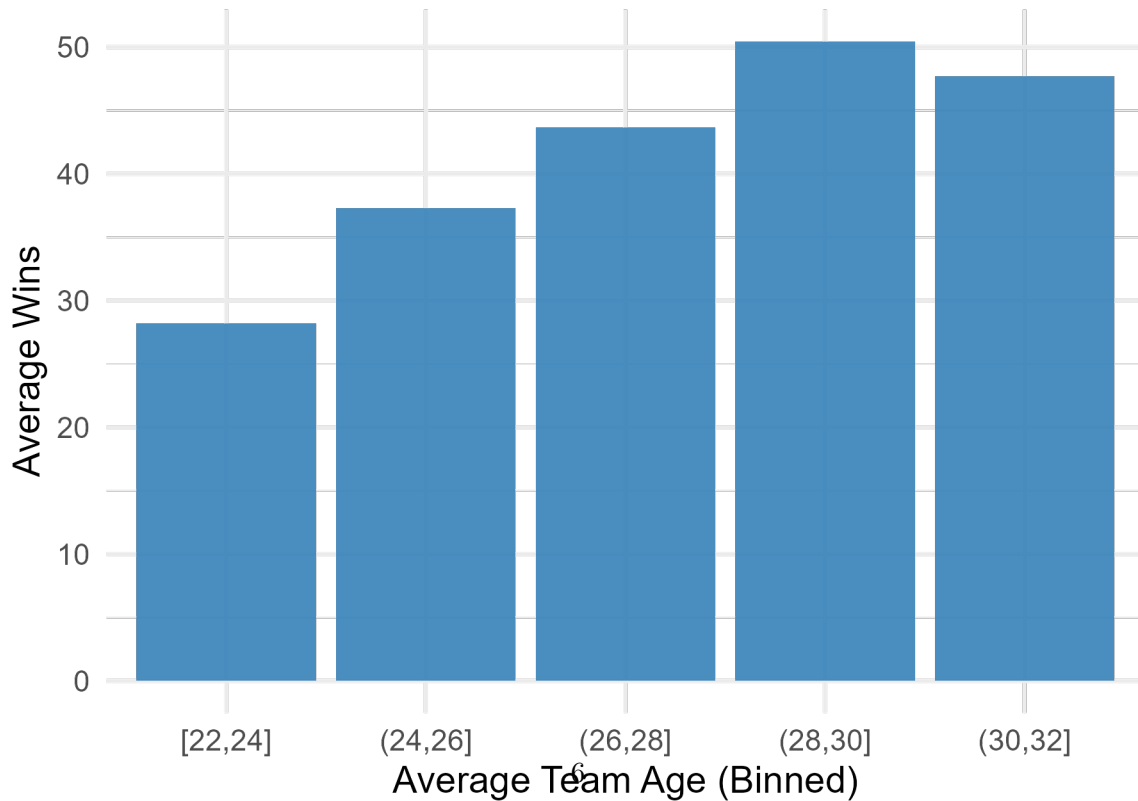NBA Non-COVID Seasons (2015–2019, 2022–2025)

Figure 4: Wins by Age Group

Discussion and Reflection

Findings from this exploration have led us to the conclusion that factors that we once thought would be influential in determining team success are not as vital as we though. Pace factor had no significant impact and three point attempt rate impact was minimal in comparison to that of age. There are a few possible reasons for this.

If we return to the data in Table 1, we see the distribution of pace factor is fairly small, consistently sitting between 90 and 105. This reminds us that the difference in pace factor is not all that different between a losing team and a winning team. Because it relies on the exchange of ball and play time, it makes sense for this statistic to be evenly split regardless of performance except for in exceptionally rare cases. So why isn't this the case for age, which is also tightly distributed? It may be that age of players also incorporates a significant number of other confounding factors. As a player ages, they either improve their skill and stay in the NBA or they don't and get overshadowed by new players. Because of this, the concentration of players with high levels of skill is higher within older players, even if they are not "old" by everyday standards.

Upon reflection, there are many ways we could have explored team success. Suggestions from peers included using point differentials, which evaluate by how much a team has won. Win ratios were also suggested. Other factors in relation to team success should absolutely be explored. Our findings here indicate that pace factor is not relevant to this discussion of how many games are won in a season. Perhaps, then, statistics relating more closely to offensive and defensive strategies in particular, as well as the manner teams prefer to score are indicators of success. We see some indication of this in our exploration of three point attempt rate. However, non-game play factors also seem to contribute to team success. Age, for one, has a substantial effect. Perhaps it is necessary to explore what other aspects *outside* the game impact how the game is played.

Sources

1. Stathead. (2025, December 11). *Team Season Finder - Pro Basketball*. Stathead. https://www.sports-reference.com/stathead/basketball/team-season-finder.cgi. Data sourced from Basketball-Reference (Sports Reference LLC), retrieved 2025.

Code Appendix

```
#import libraries
library(knitr)
library(kableExtra)
library(tidyr)
```

```r
library(dplyr)
library(psych)
library(ggplot2)
#import data
allNBA <- read.csv("Data/allNBA.csv")

#additional commands
wins_ylim <- c(10, 75)
subtitle_txt <- "NBA Non-COVID Seasons (2015-2019, 2022-2025)"

#select table information
tableNBA <- allNBA %>% select(Team, Rk, Year, W , X3PAr, Pace, Age)
table1NBA <- describe(tableNBA)

#select statistics
table1NBA <- table1NBA %>% select(n, min, median, mean, max, sd, range)

#change rownames, add detail
table1NBA$variables <- c("Team", "Team Rank", "Year", "Wins", "3-Point Att.", "Pace", "Avg. A
table1NBA$datatype <- c("String", "Ordered Num.", "Ordered Num.", "Numeric", "Numeric", "Nume
table1NBA$datadescription <- c("Team Name", "Rank in NBA", "Season Year", "# of Games Won", "

#organize and clean data
table1NBA <- table1NBA %>% select(variables, datatype, datadescription, n, min, median, mean
colnames(table1NBA) <- c(" ","Data Type", "Description", "Count", "Min. Val.", "Median Val."

#generate table
table1 <- table1NBA %>% rowwise() %>% kable(
  caption = c("Summary Table, NBA Overall"),
  align ="lccccccccc",
  digits = 2) %>%
  kable_styling(bootstrap_options = c("striped", "hover")) %>%
  kable_classic()
table1

p5 <- ggplot(allNBA, aes(x = W)) +
  geom_histogram(bins = 15, fill = "#2C7BB6", color = "black", alpha = 0.85) +
  labs(
    title = "Distribution of NBA Team Wins",
    subtitle = subtitle_txt,
    x = "Wins",
    y = "Number of Teams"
```

```r
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(size = 11))
p5

#identify correlation coefficients
age <- cor.test(allNBA$W, allNBA$Age)
tpar <- cor.test(allNBA$W, allNBA$X3PAr)
pace <- cor.test(allNBA$W, allNBA$Pace)

#combine into a df
factor <- c("age", "tpar", "pace")
coef <- c(age$estimate, tpar$estimate, pace$estimate)
t <- c(age$statistic, tpar$statistic, pace$statistic)
p <- c(age$p.value, tpar$p.value, pace$p.value)
factor <- c("Player Age", "Three-Point Attempts", "Pace Factor")
factorcoef <- data.frame(factor, coef, t, p)

#create correlation coef. table
corrtab <- factorcoef %>% kable(
  col.names = c("", "Correlation Coef. (p)", "t-Statistic", "p-Value"),
  digits = 3,
  align = "lccc",
  caption = c("Factor Correlation with Season Win Total")
  ) %>% kable_classic()
corrtab

# Correlations (for labeling plots)
cor_age  <- round(cor(allNBA$Age, allNBA$W, use = "complete.obs"), 2)
cor_pace <- round(cor(allNBA$Pace, allNBA$W, use = "complete.obs"), 2)
cor_3par <- round(cor(allNBA$X3PAr, allNBA$W, use = "complete.obs"), 2)

p3 <- ggplot(allNBA, aes(x = Pace, y = W)) +
  geom_point(alpha = 0.45, size = 2, color = "#2C7BB6") +
  geom_smooth(method = "lm", se = FALSE, color = "#D7191C", linewidth = 1) +
  annotate("text", x = max(allNBA$Pace, na.rm = TRUE), y = wins_ylim[2],
           label = paste("Correlation =", cor_pace),
           hjust = 1, vjust = 1, size = 4) +
  coord_cartesian(ylim = wins_ylim) +
  labs(
    title = "Pace vs Team Wins",
```

```r
    subtitle = subtitle_txt,
    x = "Pace",
    y = "Wins"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(size = 11))
p3

p1 <- ggplot(allNBA, aes(x = X3PAr, y = W)) +
  geom_point(alpha = 0.45, size = 2, color = "#2C7BB6") +
  geom_smooth(method = "lm", se = FALSE, color = "#D7191C", linewidth = 1) +
  annotate("text", x = max(allNBA$X3PAr, na.rm = TRUE), y = wins_ylim[2],
           label = paste("Correlation =", cor_3par),
           hjust = 1, vjust = 1, size = 4) +
  coord_cartesian(ylim = wins_ylim) +
  labs(
    title = "3-Point Attempt Rate vs Team Wins",
    subtitle = subtitle_txt,
    x = "3-Point Attempt Rate",
    y = "Wins"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(size = 11))

p1
p2 <- ggplot(allNBA, aes(x = Age, y = W)) +
  geom_point(alpha = 0.45, size = 2, color = "#2C7BB6") +
  geom_smooth(method = "lm", se = FALSE, color = "#D7191C", linewidth = 1) +
  annotate("text", x = max(allNBA$Age, na.rm = TRUE), y = wins_ylim[2],
           label = paste("Correlation =", cor_age),
           hjust = 1, vjust = 1, size = 4) +
  coord_cartesian(ylim = wins_ylim) +
  labs(
    title = "Average Team Age vs Wins",
    subtitle = subtitle_txt,
    x = "Average Team Age",
    y = "Wins"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold"),
```

```
        plot.subtitle = element_text(size = 11))
```