Cover Page for project:

Data modeling in Business (BMGT430)

Project Title: Beyond the Scoreboard: A Data-Driven Model for Evaluating NBA Team Performance

Group Members: **Ryan Dabbs, Colby Stevens, Shen Jia Cheng, Sebastian Csontos, Jared Kimiabakhsh**

I pledge on my honor that I have not given or received any unauthorized assistance on this project

Team Member	Signature
Jared Kimiabakhsh	Jared Kimiabakhsh
Ryan Dabbs	Ryan Dabbs
Colby Stevens	Colby Stevens
Shen Jia Cheng	Shen Jia Cheng
Sebastian Csontos	Sebastian Csontos

Introduction

Research Questions

We are evaluating whether an NBA team overperformed or underperformed in a given season based on statistical predictors. This leads to several guiding questions:

- Can we profit off of this information?
- Ask questions about how we can improve the model
- Are all predictors necessary?
- Are any of the predictors correlated?
- Can any predictors be transformed to improve the model?
- Are there any significant interaction variables?
- Can we use this model to predict future seasons using lagged variables?

Selecting Data

We selected a dataset created through web-scraping and compilation, modeled after Lahman's Baseball Database. The dataset represents historical professional basketball data across three leagues:

- The National Basketball Association (NBA): 1950-present
- The Basketball Association of America (BAA): 1947-1949
- The American Basketball Association (ABA): 1968-1976

We merged the following datasets:

- Team Summaries: Season-level statistics
- Team Stats Per Game: Game-level performance metrics

Turning Questions into Actions:

To answer our key research questions, we took the following steps:

- Performed initial t-tests on all predictors
- Ran a partial F-test comparing the full model to a reduced model (excluding high p-value predictors)
- Created an interaction variable between 3-point percentage and 3-point volume
- Check ViF scores of multiple models to address multicollinearity
- Tested a lagged variable for wins from the prior season

Collecting and Organizing the Data:

Filtering the Dataset

We focused our analysis on the last 10 years (post-2015) to reflect the modern era of NBA basketball because while the 3-point line was introduced in 1979–80:

- 3-point volume increased significantly post-2015
- 3-point percentage also improved significantly
- Teams began abandoning midrange shots, impacting true shooting %
- Rule changes altered game pace and strategy

Cleaning and Preparing Data

We performed the following procedures on the dataset to set it up for a proper analysis:

- Checked for data issues
- Removed duplicate rows and handled missing values
- Converted data types appropriately (e.g., categorical to binary)
- Checked for outliers and leverage points
- Transformed certain predictors into categorical variables where appropriate

Data Selection, Data Cleaning, and Exploratory Data Analysis

Background and Motivation

A common interest that all of our team members share is that we all follow basketball, specifically the National Basketball Association (NBA). We were interested in determining which statistics and variables were more or less present in teams that achieved a higher number of wins per season. More specifically, we aimed to evaluate NBA team performance by predicting expected win totals based on team-level statistics, with the ultimate goal of informing managerial decision-making.

By identifying whether teams overperformed, underperformed, or performed in line with expectations during a given season, managers, analysts, and decision-makers can gain actionable insights, such as context for guiding future decisions related to player acquisitions, tactical adjustments, and resource allocation.

Dataset Description

Motivated by this goal, we searched publicly available data and discovered a very large dataset on Kaggle, which consisted of NBA data between 1947-2025 (73+ years). This massive dataset represented 3 leagues:

- National Basketball Association (1950-present)
- The NBA's predecessor, the Basketball Association of America (1947-1949)
- The NBA's past competitor, the American Basketball Association (1968-1976)

This dataset included 7 files consisting of statistics on the team side and 10 files containing statistics about players between 1947-2025. We chose to merge two of these datasets:

- Team Season CSV file

The team season dataset contained more general team statistics such as:

  - Whether a team made the playoffs (binary)
  - Average age of a team
  - Number of wins
  - Average margin of victory (points the team usually won by)
  - Strength of schedule (how difficult their opponents were)
  - Offensive rating (points scored per 100 possessions)
  - Defensive rating (points allowed per 100 possessions)
  - Pace (number of possessions per 48 minutes)
  - Fan attendance, and more

- Team Stats Per Game file,

The team stats per game game dataset contained more specific game statistics such as:

- Shots per game
- Shooting accuracy per game
- 3-point percentage
- 3-point attempt rate (the proportion of field goal attempts that are 3-pointers)
- Points per game,
- Assists per game
- Steals per game
- Blocks per game, and more

By combining these datasets, we could get a greater number of variables that had the potential to be relevant to team success. Together, we believe the predictors from both datasets helped capture dimensions of offensive and defensive effectiveness, game tempo, shooting tendencies, and postseason performance, all of which are expected to influence a team’s win total.

Focus on the Modern NBA (2015-Present)

Given our focus on the most relevant and contemporary style of play, we restricted our dataset to NBA seasons from 2015 to the present. This decision reflects several important developments in modern basketball:

- Although the 3-point line was introduced in the 1979–80 season, the use of 3-point shots surged dramatically after 2015.
- Teams began emphasizing spacing, 3-point efficiency, and shot selection optimization, reducing reliance on mid-range shots.
- Rule changes and evolving strategies increased both pace and offensive efficiency.

By focusing on the post-2015 period, we ensure that our model captures team behavior consistent with the current era of basketball.

Data Cleaning Process

Filtering Data

Before modeling, we conducted several data-cleaning steps to ensure quality and consistency as well as exploratory data analysis to gain a better understanding of our dataset. At the start, our merged dataset contained 1876 rows and 54 columns. We then filtered our dataset by the year column, so that it just contained rows from 2015 and onward, which reduced the number of rows to 341.

Cleaning and Preparing Data

We then generated summary statistics for every single column using the summary() command, to help us understand the inter quartile ranges for each column and see which columns contained missing values. While generating summary statistics, we discovered that there were specifically 11 instances where there were missing values across multiple columns.

From this, we discovered that there existed a “League Average” column at the end of each year which averaged the stats for each team, which contained missing values. Therefore, we removed these 11 “League Average” rows, which reduced the number of rows in our dataset to 330.

Through our generation of summary statistics for each column, we also could see that some columns were represented as categorical or as characters. This included, team name, team abbreviation, playoffs (True or False for whether a team made the playoffs), and arena name. Of these variables we were able to convert the “playoffs” variable to a binary (0 or 1) value through the use of the factor() method.

Handling Outliers

We also examined outliers and leverage points which were present, but we chose not to exclude them, as a deeper look into those points showed statistics for teams that performed exceptionally well and teams who performed extremely poorly. An example includes the 2019-2020 Golden State Warriors, who achieved a phenomenal record of 73 wins and 9 losses, which was the greatest number of wins recorded in season history and deemed as an outlier in our dataset.

After this, we continued to conduct more exploratory data analysis by plotting the graphs between wins (our response variable) and all other predictors (potential significant variables). It was while graphing and observing these relationships and summary statistics that led us to make an interesting discovery:

- While looking at the fan attendance predictor, we saw that the minimum average fan attendance for a team was 162 people and by observing the fan attendance vs wins plot, we could see that there was clustering. From this, we were able to figure out that during COVID-19, the regular seasons in 2020 and 2021 had been reduced to 64-75 games for some teams, compared to the normal 82 each year.

Exploratory Data Analysis

Reading in the dataset:

```
> team_summaries <- read_csv("C:/Ryan/Spring2025/BMGT430/team_summaries.csv")
> team_stats_per_game
<-read_csv("C:/Ryan/Spring2025/BMGT430/team_stats_per_game.csv")
```

Viewing the columns and dimensions of the datasets:

```
> names(team_summaries)
 [1] "season"           "lg"               "team"             "abbreviation"
 [5] "playoffs"         "age"              "w"                "l"
 [9] "pw"               "pl"               "mov"              "sos"
[13] "srs"              "o_rtg"            "d_rtg"            "n_rtg"
[17] "pace"             "f_tr"             "x3p_ar"           "ts_percent"
[21] "e_fg_percent"     "tov_percent"      "orb_percent"      "ft_fga"
[25] "opp_e_fg_percent" "opp_tov_percent"  "opp_drb_percent"  "opp_ft_fga"
[29] "arena"            "attend"           "attend_g"
```

```
> dim(team_summaries)
[1] 1876   31
```

```
> names(team_stats_per_game)
 [1] "season"         "lg"             "team"           "abbreviation"   "playoffs"
 [6] "g"              "mp_per_game"    "fg_per_game"    "fga_per_game"   "fg_percent"
[11] "x3p_per_game"   "x3pa_per_game"  "x3p_percent"    "x2p_per_game"   "x2pa_per_game"
[16] "x2p_percent"    "ft_per_game"    "fta_per_game"   "ft_percent"     "orb_per_game"
[21] "drb_per_game"   "trb_per_game"   "ast_per_game"   "stl_per_game"   "blk_per_game"
[26] "tov_per_game"   "pf_per_game"    "pts_per_game"
> dim(team_stats_per_game)
[1] 1876   28
```

Merging the dataset:

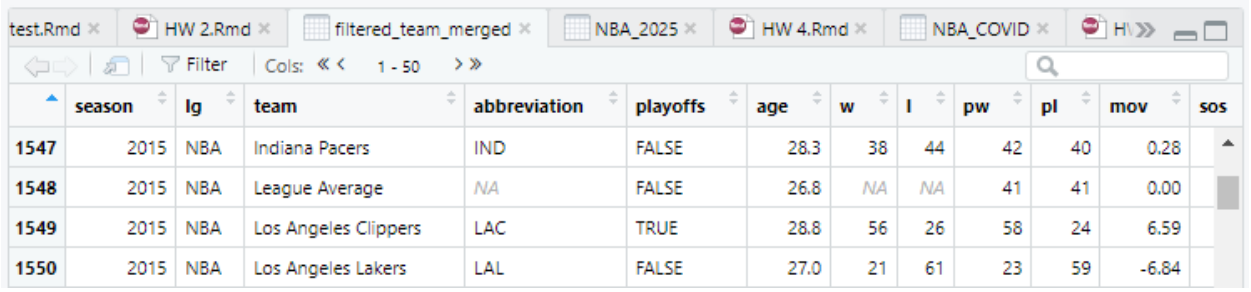
```
> team_merged <- merge(team_summaries, team_stats_per_game, by = c("season", "lg", "tea
m", "abbreviation", "playoffs"))
> dim(team_merged)
[1] 1876   54
```

Filtering the Dataset for Modern NBA Seasons (2015–Present):

```
> filtered_team_merged <- subset(team_merged, season >= 2015)
> dim(filtered_team_merged)
[1] 341  54
```

Identifying and Removing "League Average" Rows

- Discovered entries with team = "League Average" for each season, which contained missing values (e.g., NAs for wins/losses)
    a. Identifying “League Average” row that contained missing values:



test.Rmd | HW 2.Rmd | filtered_team_merged | NBA_2025 | HW 4.Rmd | NBA_COVID | H

Filter | Cols: « < 1 - 50 > »

	season	lg	team	abbreviation	playoffs	age	w	l	pw	pl	mov	sos
1547	2015	NBA	Indiana Pacers	IND	FALSE	28.3	38	44	42	40	0.28	
1548	2015	NBA	League Average	*NA*	FALSE	26.8	*NA*	*NA*	41	41	0.00	
1549	2015	NBA	Los Angeles Clippers	LAC	TRUE	28.8	56	26	58	24	6.59	
1550	2015	NBA	Los Angeles Lakers	LAL	FALSE	27.0	21	61	23	59	-6.84	

    b. Removing every single “League Average” row (one per year):

```
> cleaned_team_merged<-subset(filtered_team_merged, team != "League Average")
> dim(cleaned_team_merged)
[1] 330  54
```

Exploratory Data Analysis - Summary Statistics

```
> summary(filtered_team_merged)
     season          lg                team           abbreviation        playoffs
 Min.   :2015   Length:341         Length:341         Length:341         Mode :logical
 1st Qu.:2017   Class :character   Class :character   Class :character   FALSE:197
 Median :2020   Mode  :character   Mode  :character   Mode  :character   TRUE :144
 Mean   :2020
 3rd Qu.:2023
 Max.   :2025

      age              w                l               pw              pl
 Min.   :22.10   Min.   :10.00   Min.   : 9.00   Min.   :14.00   Min.   :12.00
 1st Qu.:25.10   1st Qu.:32.00   1st Qu.:31.00   1st Qu.:32.00   1st Qu.:32.00
 Median :26.20   Median :41.50   Median :39.00   Median :41.00   Median :39.00
 Mean   :26.34   Mean   :39.71   Mean   :39.71   Mean   :39.76   Mean   :39.66
 3rd Qu.:27.40   3rd Qu.:48.00   3rd Qu.:48.00   3rd Qu.:48.00   3rd Qu.:47.00
 Max.   :30.60   Max.   :73.00   Max.   :72.00   Max.   :67.00   Max.   :66.00
                 NA's   :11      NA's   :11
      mov                 sos                 srs                o_rtg
 Min.   :-12.33000   Min.   :-0.930000   Min.   :-12.01000   Min.   : 95.5
 1st Qu.: -2.90000   1st Qu.:-0.220000   1st Qu.: -2.91000   1st Qu.:107.5
 Median :  0.22000   Median : 0.000000   Median :  0.15000   Median :111.0
 Mean   : -0.01326   Mean   :-0.001173   Mean   : -0.01463   Mean   :110.8
 3rd Qu.:  3.05000   3rd Qu.: 0.220000   3rd Qu.:  2.89000   3rd Qu.:114.4
 Max.   : 13.36000   Max.   : 0.790000   Max.   : 13.31000   Max.   :123.2

     d_rtg           n_rtg                pace            f_tr            x3p_ar
 Min.   : 99.0   Min.   :-12.100000   Min.   : 90.40   Min.   :0.1940   Min.   :0.1790
 1st Qu.:107.8   1st Qu.: -3.100000   1st Qu.: 96.20   1st Qu.:0.2400   1st Qu.:0.3180
 Median :111.0   Median :  0.450000   Median : 98.20   Median :0.2570   Median :0.3620
 Mean   :110.8   Mean   : -0.004242   Mean   : 97.97   Mean   :0.2584   Mean   :0.3585
 3rd Qu.:113.8   3rd Qu.:  3.275000   3rd Qu.: 99.80   3rd Qu.:0.2730   3rd Qu.:0.4010
 Max.   :120.4   Max.   : 13.400000   Max.   :105.10   Max.   :0.3630   Max.   :0.5380
                 NA's   :11
   ts_percent      e_fg_percent     tov_percent     orb_percent        ft_fga
 Min.   :0.4940   Min.   :0.4560   Min.   : 9.90   Min.   :17.90   Min.   :0.1430
 1st Qu.:0.5460   1st Qu.:0.5100   1st Qu.:12.10   1st Qu.:21.80   1st Qu.:0.1860
 Median :0.5640   Median :0.5280   Median :12.60   Median :23.50   Median :0.1980
 Mean   :0.5621   Mean   :0.5264   Mean   :12.67   Mean   :23.51   Mean   :0.1992
 3rd Qu.:0.5790   3rd Qu.:0.5430   3rd Qu.:13.30   3rd Qu.:24.90   3rd Qu.:0.2100
 Max.   :0.6100   Max.   :0.5800   Max.   :16.00   Max.   :31.90   Max.   :0.2760

 opp_e_fg_percent opp_tov_percent opp_drb_percent   opp_ft_fga        arena
 Min.   :0.4700   Min.   :10.30   Min.   :71.60   Min.   :0.1450   Length:341
 1st Qu.:0.5120   1st Qu.:11.90   1st Qu.:75.20   1st Qu.:0.1870   Class :character
 Median :0.5290   Median :12.60   Median :76.40   Median :0.1980   Mode  :character
 Mean   :0.5265   Mean   :12.68   Mean   :76.47   Mean   :0.1993
 3rd Qu.:0.5420   3rd Qu.:13.40   3rd Qu.:77.70   3rd Qu.:0.2120
 Max.   :0.5760   Max.   :16.30   Max.   :81.60   Max.   :0.2640
```

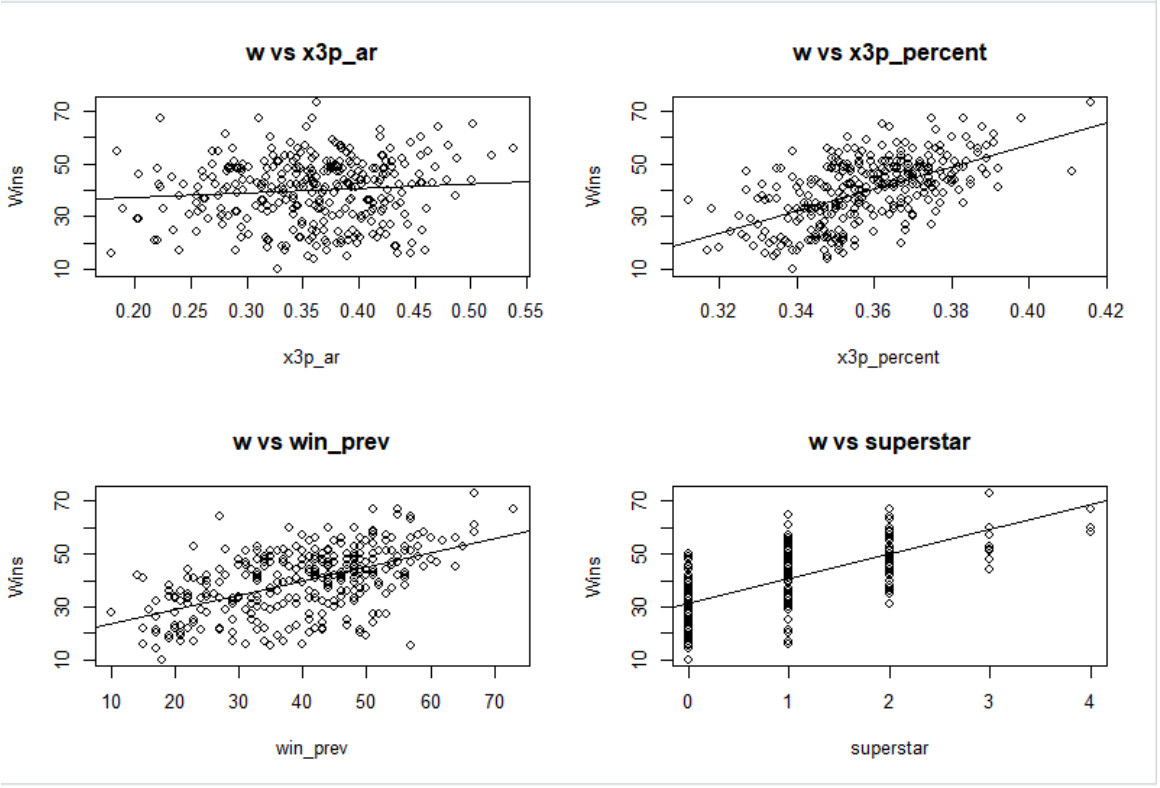
```
     attend          attend_g          g          mp_per_game     fg_per_game
 Min.   :  5817   Min.   :  162   Min.   :64.00   Min.   :240.0   Min.   :33.70
 1st Qu.:635941   1st Qu.:16449   1st Qu.:75.00   1st Qu.:240.9   1st Qu.:38.80
 Median :708639   Median :17608   Median :82.00   Median :241.5   Median :40.40
 Mean   :658060   Mean   :16548   Mean   :79.42   Mean   :241.6   Mean   :40.36
 3rd Qu.:762855   3rd Qu.:18997   3rd Qu.:82.00   3rd Qu.:242.1   3rd Qu.:42.00
 Max.   :894659   Max.   :21821   Max.   :82.00   Max.   :244.0   Max.   :47.00
 NA's   :4        NA's   :4
  fga_per_game     fg_percent      x3p_per_game   x3pa_per_game    x3p_percent
 Min.   :77.20   Min.   :0.4080   Min.   : 5.00   Min.   :14.90   Min.   :0.3120
 1st Qu.:85.50   1st Qu.:0.4500   1st Qu.: 9.90   1st Qu.:27.20   1st Qu.:0.3480
 Median :87.40   Median :0.4630   Median :11.40   Median :32.00   Median :0.3570
 Mean   :87.32   Mean   :0.4621   Mean   :11.26   Mean   :31.39   Mean   :0.3582
 3rd Qu.:89.30   3rd Qu.:0.4730   3rd Qu.:12.80   3rd Qu.:35.50   3rd Qu.:0.3690
 Max.   :94.40   Max.   :0.5070   Max.   :17.90   Max.   :48.50   Max.   :0.4160

  x2p_per_game  x2pa_per_game    x2p_percent      ft_per_game     fta_per_game
 Min.   :23.1   Min.   :41.60   Min.   :0.4490   Min.   :12.30   Min.   :16.70
 1st Qu.:27.6   1st Qu.:52.70   1st Qu.:0.5000   1st Qu.:16.40   1st Qu.:21.20
 Median :29.2   Median :55.80   Median :0.5240   Median :17.30   Median :22.50
 Mean   :29.1   Mean   :55.93   Mean   :0.5216   Mean   :17.37   Mean   :22.53
 3rd Qu.:30.7   3rd Qu.:59.20   3rd Qu.:0.5430   3rd Qu.:18.30   3rd Qu.:23.60
 Max.   :33.9   Max.   :68.30   Max.   :0.5890   Max.   :22.60   Max.   :29.40

   ft_percent      orb_per_game    drb_per_game    trb_per_game    ast_per_game
 Min.   :0.6680   Min.   : 7.60   Min.   :29.20   Min.   :38.60   Min.   :18.00
 1st Qu.:0.7540   1st Qu.: 9.50   1st Qu.:32.60   1st Qu.:42.90   1st Qu.:22.50
 Median :0.7740   Median :10.30   Median :33.70   Median :44.00   Median :24.20
 Mean   :0.7716   Mean   :10.35   Mean   :33.64   Mean   :43.99   Mean   :24.28
 3rd Qu.:0.7900   3rd Qu.:11.00   3rd Qu.:34.60   3rd Qu.:45.20   3rd Qu.:25.90
 Max.   :0.8390   Max.   :14.80   Max.   :42.20   Max.   :51.70   Max.   :31.10

  stl_per_game     blk_per_game    tov_per_game    pf_per_game     pts_per_game
 Min.   : 5.700   Min.   :2.400   Min.   :11.10   Min.   :15.60   Min.   : 91.9
 1st Qu.: 7.100   1st Qu.:4.400   1st Qu.:13.40   1st Qu.:18.80   1st Qu.:104.5
 Median : 7.600   Median :4.800   Median :14.10   Median :19.90   Median :110.3
 Mean   : 7.679   Mean   :4.859   Mean   :14.11   Mean   :19.84   Mean   :109.4
 3rd Qu.: 8.200   3rd Qu.:5.300   3rd Qu.:14.90   3rd Qu.:20.90   3rd Qu.:114.0
 Max.   :10.400   Max.   :7.500   Max.   :17.70   Max.   :24.80   Max.   :123.3
```

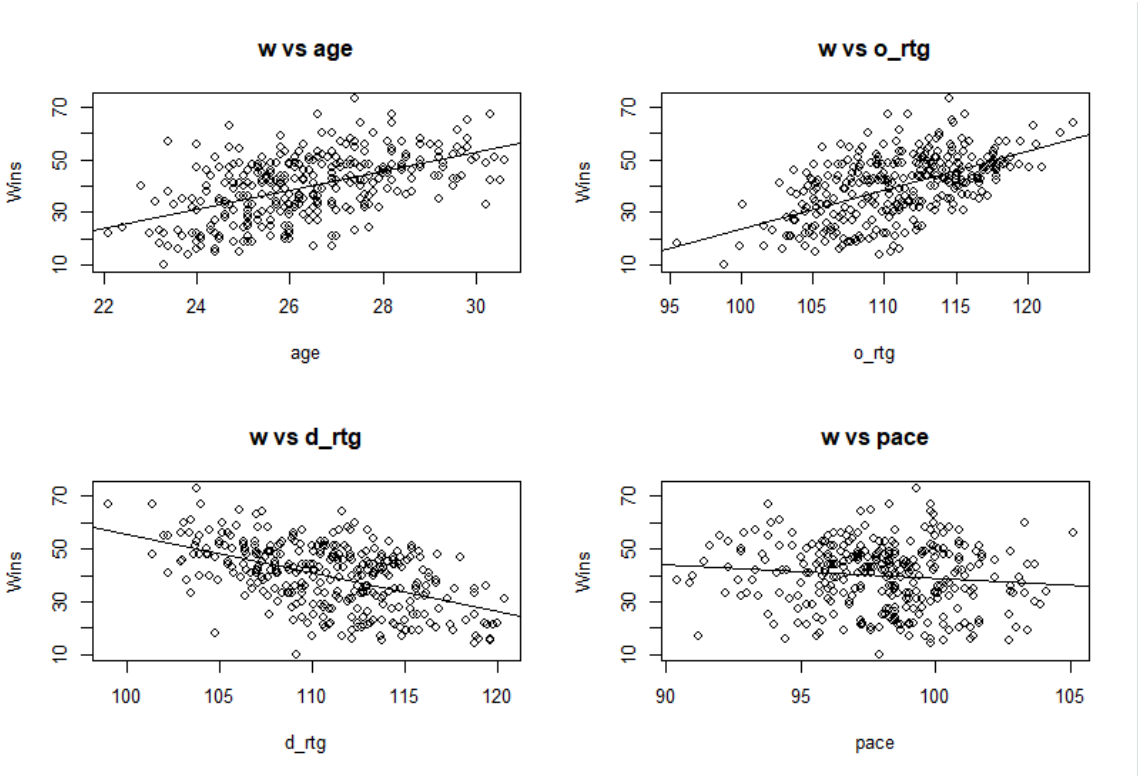
Exploratory Data Analysis – Graphs

Created scatter plots to visualize the relationship between team wins (w) and key predictors:

c. Significant positive relationships observed with: x3p_percent, o_rtg, age, win_prev, superstar
d. Negative or weak relationships: pace, d_rtg (inversely related), x3p_ar (more spread)

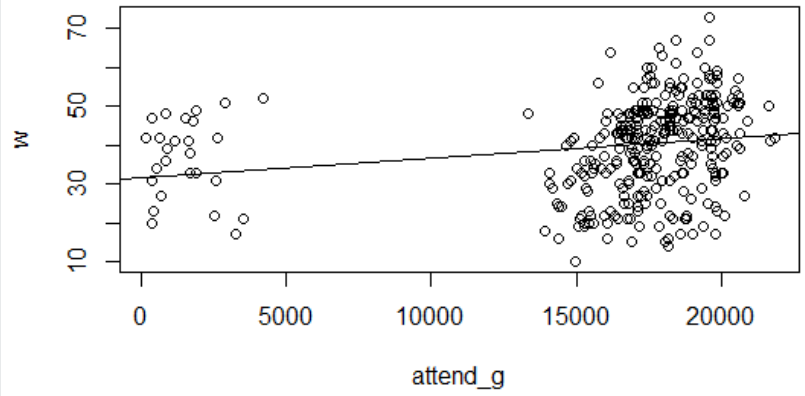
Note: Only a few plots shown; full dataset had 53 predictors.


w vs x3p_ar
Wins
x3p_ar
w vs x3p_percent
Wins
x3p_percent
w vs win_prev
Wins
win_prev
w vs superstar
Wins
superstar


w vs age
Wins
age
w vs o_rtg
Wins
o_rtg
w vs d_rtg
Wins
d_rtg
w vs pace
Wins
pace

Exploratory Data Analysis - Discovery:

```
> plot(attend_g, w)
> attend_g_m<-lm(w~attend_g)
> abline(attend_g_m)
```




```
> summary(cleaned_team_merged$attend_g)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    162   16417   17579   16555   19052   21821
```

Initial Modeling and Reducing the Model

Step by Step Modeling Approach

To predict NBA team win totals, we developed and refined multiple linear regression models using a structured, step-by-step approach.

Our model development process then involved multiple stages to systematically evaluate and refine predictor variables:

- We first conducted t-tests on all predictors to evaluate their statistical significance in explaining win totals, allowing us to preliminarily gauge which variables showed strong relationships with the outcome.
- To determine whether all predictors were necessary, we performed a partial F-test comparing the full model (with all predictors) to a reduced model (excluding predictors

with high t-test p-values). This step assessed whether removing weaker variables materially affected explanatory power.
- We then checked assumptions and looked for multicollinearity, or if transformations needed to be implemented as well as if interaction terms could be created.

Reducing the Model

In reducing the model, we systematically applied the previously outlined actions:

- We began with a Partial F-test on predictors we suspected were redundant, unrelated to the outcome, or had high p-values in their individual t-tests.
- This initial step left us with a reduced model of 15 predictors, which was not statistically different from the full model, indicating that the omitted variables were not meaningful contributors.

Multicollinearity and Refinement

Next, we assessed multicollinearity using Variance Inflation Factor (VIF) testing:

- This revealed several predictors with high multicollinearity, prompting us to evaluate which variables were driving the redundancy.
- Leveraging both statistical evidence and domain knowledge of the sport, we selectively removed predictors we believed were inflating the VIF scores.
- This refinement resulted in a model that passed multicollinearity checks, with all remaining predictors exhibiting acceptable VIF values.

Through ViF testing, we were able to reduce our model further for our post-partial F-test model.

Transformations and Lagged Variables

To continue improving the model, we revisited variables that we believed held theoretical importance but failed to demonstrate significance in their individual t-tests.

This led us to consider two strategies:

- Variable transformation
- Interaction terms

Example: Pace (log transformation)

- We attempted a log transformation on Pace, the number of possessions a team uses in a game, under the hypothesis that it would relate to scoring and wins.
- However, the transformation did not improve its statistical properties; it still violated linear assumptions and retained a high p-value.
- As a result, Pace was removed from the model.

Lagged Wins – Previous Season Performance

As we continued refining the model, we observed a clear trend: teams that had won many games in the previous season often maintained strong performance in the following year.

This pattern suggested that incorporating a lagged variable representing prior season wins could enhance the model's predictive power. We introduced Previous Wins as a predictor, which we tested to see if it significantly improved the model. After our testing, we saw no significant difference, meaning we could remove the lagged wins previous from the model.

Polynomial Term for Age

- We attempted to square the age variable to reflect a potential parabolic (concave) relationship where performance increases with age up to a point, then declines.
- Specifically, we had thought that a young team often lacked experience, an older team may decline due to fatigue and injuries, while a prime-aged team with a mixed-age players would be more likely to win.
- However, upon graphing the relationship between age and wins, we observed that it exhibited a more linear shape. Additionally, we tried to square the age variable to see if it would become more statistically significant and improve the model, but this was not the case.

Interaction Terms

We also explored interaction terms between predictors. Specifically, we identified 3-point shooting percentage and 3-point shot volume, which are variables that are related.

In our 3rd reduced model:

- 3-point shot volume (x3p_ar) had a p-value of 0.000594 (significant)
- 3-point shooting percentage (x3p_percent) had a p-value of 0.856455 (not significant)

When multiplying these variables together to form an interaction variable:

- This interaction variable became statistically significant, with a p-value of 0.000445, which was lower than either one of those variables on its own.
- The 2 variables logically more accurately represented actual 3-point shots made, a more direct measure of offensive effectiveness.

By incorporating this interaction term, it improved the model, increasing our adjusted R^2 from, making it a meaningful and impactful addition.

Takeaway and Insights

Our study demonstrates that:

- Team-level statistics, particularly offensive and defensive ratings, 3-point shooting metrics, and prior season performance, can meaningfully predict NBA win totals in the modern era.
- By comparing expected win totals to actual outcomes, our model equips managers and analysts with a tool to identify whether teams overperformed, underperformed, or met expectations.

Strategic Use:

These insights can inform key strategic decisions such as:

- Evaluating coaching effectiveness
- Assessing roster efficiency
- Prioritizing player acquisitions
- Reallocating resources

While no model can perfectly forecast a season's outcome, our approach offers a solid foundation for evidence-based decision-making within professional basketball organizations.

Future work could extend this analysis by:

- Incorporating player-level data, injury reports, and advanced shot-quality metrics to further enhance predictive power.
- Testing the profitability of these predictions in betting markets represents an exciting avenue for applied sports analytics.

Model Outputs and Progressive Refinement

Initial Model and Reduction Process:

Initial Model:

- Fit for model after dataset was cleaned
- Removed "League Average" rows for each year, which consisted of missing values for many columns

```
Call:
lm(formula = w ~ factor(playoffs) + age + srs + o_rtg + d_rtg +
    pace + f_tr + x3p_ar + ts_percent + e_fg_percent + tov_percent +
    orb_percent + ft_fga + opp_e_fg_percent + opp_tov_percent +
    opp_drb_percent + opp_ft_fga + fg_per_game + fga_per_game +
    fg_percent + x3p_per_game + x3p_percent + x2p_per_game +
    x2p_percent + ft_per_game + fta_per_game + ft_percent + orb_per_game +
    drb_per_game + ast_per_game + stl_per_game + blk_per_game +
    tov_per_game, data = NBA_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-9.3126 -2.0765 -0.0647  2.2566 10.9335

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           32.23544  206.91318   0.156   0.8763
factor(playoffs)TRUE   2.97901    0.52355   5.690 3.05e-08 ***
age                    0.30884    0.14553   2.122   0.0346 *
srs                    0.83986    0.63207   1.329   0.1850
o_rtg                 -1.06920    1.40404  -0.762   0.4470
d_rtg                 -1.17336    0.99368  -1.181   0.2386
pace                  -0.06646    0.59660  -0.111   0.9114
f_tr                  56.38563  277.82805   0.203   0.8393
x3p_ar               161.77454   62.65757   2.582   0.0103 *
ts_percent          -248.77145  481.76607  -0.516   0.6060
e_fg_percent         580.61617  510.70237   1.137   0.2565
tov_percent           -4.70616    4.36075  -1.079   0.2814
orb_percent            1.42232    1.20971   1.176   0.2406
ft_fga               600.17601  357.49643   1.679   0.0942 .
opp_e_fg_percent     -61.03004  136.56574  -0.447   0.6553
opp_tov_percent       -0.16250    1.14049  -0.142   0.8868
opp_drb_percent       -0.38882    0.50863  -0.764   0.4452
opp_ft_fga            -8.90585   28.39815  -0.314   0.7540
fg_per_game           -2.37745    4.65452  -0.511   0.6099
fga_per_game           2.46919    2.04539   1.207   0.2283
fg_percent            70.13317  475.98822   0.147   0.8830
x3p_per_game          -4.25086    4.24809  -1.001   0.3178
x3p_percent          179.42082   73.93928   2.427   0.0158 *
x2p_per_game           1.17017    3.96376   0.295   0.7680
x2p_percent           54.69955   81.68919   0.670   0.5036
ft_per_game            3.08077    4.42553   0.696   0.4869
fta_per_game          -7.52146    3.75325  -2.004   0.0460 *
ft_percent          -158.95192   67.16428  -2.367   0.0186 *
orb_per_game          -0.27757    2.15232  -0.129   0.8975
drb_per_game          -0.36069    0.49890  -0.723   0.4703
ast_per_game           0.08468    0.14655   0.578   0.5638
stl_per_game          -0.28683    0.52178  -0.550   0.5829
blk_per_game          -0.70648    0.33189  -2.129   0.0341 *
tov_per_game           0.93769    3.46975   0.270   0.7872
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.327 on 296 degrees of freedom
Multiple R-squared:  0.9314,   Adjusted R-squared:  0.9238
F-statistic: 121.8 on 33 and 296 DF,  p-value: < 2.2e-16
```

Turning actions into results:

- Conducted different tests to find significant factors

- Explored possible data transformation strategies

Reduced model 1:

- The results of the questions asked and actions we took

Dropped due to redundancy:

	Estimate	Std. Error	t value	Pr(>|t|)	
tov_per_game	0.83347	3.45691	0.241	0.8096	
fg_percent	175.25830	477.69495	0.367	0.7140	
x3p_per_game	-4.18142	4.23196	-0.988	0.3239	
x3p_percent	166.12418	74.01987	2.244	0.0256	*
x2p_per_game	0.82044	3.95326	0.208	0.8357	
x2p_percent	46.45119	81.50277	0.570	0.5692	

Dropped because they one predictor encapsulates these:

	Estimate	Std. Error	t value	Pr(>|t|)	
stl_per_game	-0.31521	0.52001	-0.606	0.5449	
blk_per_game	-0.74555	0.33132	-2.250	0.0252	*
drb_per_game	-0.27467	0.49925	-0.550	0.5826	
opp_drb_percent	-0.46711	0.50852	-0.919	0.3591	
opp_ft_fga	-2.26975	28.52507	-0.080	0.9366	

Other removed Variables:

	Estimate	Std. Error	t value	Pr(>|t|)	
orb_per_game	0.07459	2.15284	0.035	0.9724	
ast_per_game	0.09596	0.14612	0.657	0.5119	
f_tr	7.93662	278.04935	0.029	0.9772	
ts_percent	-253.95403	479.92535	-0.529	0.5971	
fg_per_game	-2.60287	4.63832	-0.561	0.5751	
fga_per_game	2.49775	2.03760	1.226	0.2212	
ft_per_game	2.58069	4.41716	0.584	0.5595	
fta_per_game	-6.85602	3.75682	-1.825	0.0690	.
ft_fga	627.00626	356.43169	1.759	0.0796	.

Refitting Model After Reduction:

```
Call:
lm(formula = w ~ factor(playoffs) + age + srs + o_rtg + d_rtg +
    pace + x3p_ar + e_fg_percent + tov_percent + orb_percent +
    opp_e_fg_percent + opp_tov_percent + opp_ft_fga + x3p_percent +
    ft_percent + win_prev, data = NBA_cleaned)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-10.5533  -2.0306  -0.0566   2.3715  11.4692

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            43.35819   15.67198   2.767  0.00600 **
factor(playoffs)TRUE    3.09500    0.50915   6.079 3.52e-09 ***
age                     0.07804    0.15446   0.505  0.61372
srs                     0.16000    0.61985   0.258  0.79648
o_rtg                   1.64221    0.70639   2.325  0.02072 *
d_rtg                  -1.21634    0.63459  -1.917  0.05618 .
pace                   -0.22186    0.09596  -2.312  0.02142 *
x3p_ar                -13.63435    4.48717  -3.039  0.00258 **
e_fg_percent           81.69327   57.11714   1.430  0.15364
tov_percent            -0.33601    0.51839  -0.648  0.51734
orb_percent             0.07022    0.21981   0.319  0.74960
opp_e_fg_percent     -113.01597   36.55977  -3.091  0.00217 **
opp_tov_percent         0.60587    0.31964   1.895  0.05895 .
opp_ft_fga            -16.28520   12.96682  -1.256  0.21008
x3p_percent           -25.63702   19.10473  -1.342  0.18059
ft_percent             -8.24265   10.10097  -0.816  0.41511
win_prev                0.06457    0.02139   3.018  0.00275 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.385 on 313 degrees of freedom
Multiple R-squared:  0.9249,   Adjusted R-squared:  0.9211
F-statistic:   241 on 16 and 313 DF,  p-value: < 2.2e-16
```

Multicollinearity Check (VIF)

- Checked for multicollinearity keep finding a way to reduce the model:

```
> vif(reduced_model_1)
factor(playoffs)              age              srs            o_rtg
d_rtg
        1.836010         2.062210       242.096052       298.392452
198.595570
            pace           x3p_ar     e_fg_percent      tov_percent
orb_percent
        1.903734         2.392125        54.198318         7.203496
8.031276
opp_e_fg_percent  opp_tov_percent       opp_ft_fga      x3p_percent
ft_percent
       17.744709         3.399222         1.902674         2.842294
2.477621
        win_prev
        1.928277
```

Reduced Model 2:

- After looking for more redundancy, p-value, and multicollinearity

```
Call:
lm(formula = w ~ factor(playoffs) + age + o_rtg + d_rtg + pace +
    x3p_ar + x3p_percent + win_prev, data = NBA_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6107  -2.2856   0.0749   2.3352  12.1468

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           34.71180   10.69914   3.244  0.00130 **
factor(playoffs)TRUE   2.91485    0.50694   5.750 2.08e-08 ***
age                    0.14563    0.14630   0.995  0.32027
o_rtg                  2.12453    0.07571  28.063  < 2e-16 ***
d_rtg                 -1.98108    0.07218 -27.447  < 2e-16 ***
pace                  -0.13610    0.08711  -1.562  0.11919
x3p_ar               -11.79224    4.08337  -2.888  0.00414 **
x3p_percent           -2.98552   16.05751  -0.186  0.85262
win_prev               0.06568    0.02138   3.072  0.00231 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.438 on 321 degrees of freedom
Multiple R-squared:  0.9206,   Adjusted R-squared:  0.9186
F-statistic:   465 on 8 and 321 DF,  p-value: < 2.2e-16
```

Reduced Model 3:

- Getting rid of pace and win_prev, as they were not significant

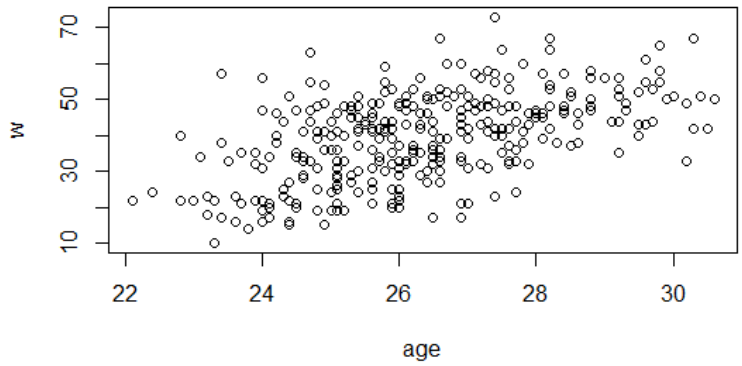
```
Call:
lm(formula = w ~ factor(playoffs) + age + o_rtg + d_rtg + x3p_ar +
    x3p_percent, data = NBA_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1233  -2.2629   0.2007   2.3246  11.7191

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           21.20196    8.20501   2.584 0.010204 *
```

```
factor(playoffs)TRUE   2.91980    0.51478   5.672 3.14e-08 ***
age                    0.38640    0.13067   2.957 0.003334 **
o_rtg                  2.17180    0.07397  29.361  < 2e-16 ***
d_rtg                 -2.05428    0.07010 -29.303  < 2e-16 ***
x3p_ar               -13.68196    3.94462  -3.469 0.000594 ***
x3p_percent           -2.95087   16.30011  -0.181 0.856455
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.491 on 323 degrees of freedom
Multiple R-squared:  0.9176,   Adjusted R-squared:  0.9161
F-statistic: 599.4 on 6 and 323 DF,  p-value: < 2.2e-16
```

Polynomial Attempt at Reducing Model 3 Age^2:

- Attempted to create a polynomial variable for age to test a parabolic relationship
- The model became less accurate, and age predictors became statistically insignificant

```
> plot(age, w)
> summary(reduced_model_3)
```




```
Call:
lm(formula = w ~ factor(playoffs) + age + I(age^2) + o_rtg +
    d_rtg + x3p_ar + x3p_percent, data = NBA_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8769 -2.2903  0.1656  2.3450 10.9770

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          92.21048   37.48025   2.460 0.014409 *
factor(playoffs)TRUE  2.86086    0.51349   5.571 5.34e-08 ***
age                  -5.04043    2.79857  -1.801 0.072627 .
I(age^2)              0.10171    0.05240   1.941 0.053101 .
```

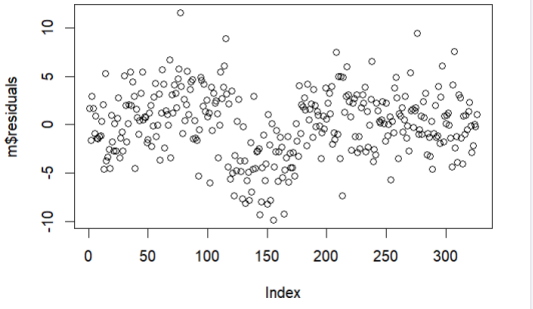
```
o_rtg                  2.18243    0.07386  29.549  < 2e-16 ***
d_rtg                 -2.06603    0.07007 -29.486  < 2e-16 ***
x3p_ar               -14.04580    3.93230  -3.572 0.000409 ***
x3p_percent            0.78994   16.34470   0.048 0.961483
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.477 on 322 degrees of freedom
Multiple R-squared:  0.9185,   Adjusted R-squared:  0.9168
F-statistic: 518.7 on 7 and 322 DF,  p-value: < 2.2e-16
```

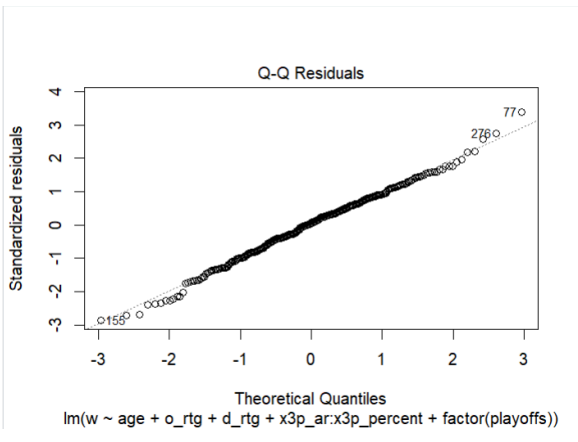
Final Model:

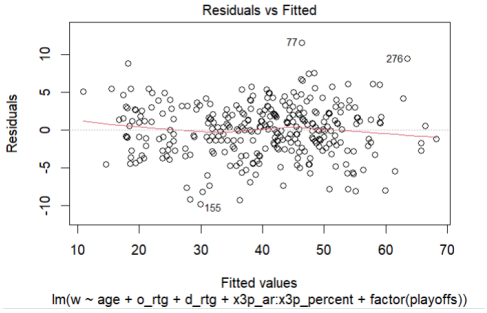
- Added interaction between 3 point percentage and 3 point accuracy

```
Call:
lm(formula = w ~ age + o_rtg + d_rtg + x3p_ar:x3p_percent + factor(playoffs),
    data = NBA_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9463 -2.2691  0.2017  2.2425 11.6905

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            17.73965    8.24102   2.153 0.032088 *
age                     0.40500    0.12838   3.155 0.001757 **
o_rtg                   2.19956    0.06678  32.938  < 2e-16 ***
d_rtg                  -2.06389    0.06811 -30.301  < 2e-16 ***
factor(playoffs)TRUE    2.91971    0.51301   5.691 2.82e-08 ***
x3p_ar:x3p_percent    -38.86741   10.95284  -3.549 0.000445 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.483 on 324 degrees of freedom
Multiple R-squared:  0.9177,   Adjusted R-squared:  0.9164
F-statistic: 722.7 on 5 and 324 DF,  p-value: < 2.2e-16
```

LINE Assumptions







Final Model Interpretation and Application

Final Model Summary

Our final regression model predicts the number of expected team wins using the following predictors:

- Average age of the team
- Offensive rating
- Defensive rating
- Prior seasons wins
- Playoff indicator (categorical)
- Interaction term: three-point volume × three-point percentage.

The model is strong:

- Explains ~92% of the variance in team wins.
- The residual standard error of 3.483, meaning that on average, the number of expected wins our model predicts is only off by 3.483 wins from the team's actual wins.

It also adheres to the LINE assumptions:

- Linearity
- Independence
- Normality
- Equal Spread

Our model also had no multicollinearity or omitted variable bias.

Coefficient Interpretation:

- Age:
    - Highly significant.
    - On average, an increase of one year in team age → increase of 0.405 wins, holding the other predictors fixed.

Two of the strongest predictors in our model are offensive rating and defensive rating:

- Offensive rating
    - Has a strong, *positive* effect on wins
    - Every 1 point increase in offensive rating, → about 2.19 more games won, holding other predictors fixed.
- Defensive rating
    - Has a strong, *negative* impact on wins,
    - Every 1 point increase in defensive rating (lower is better), → 2.06 more games lost, holding other predictors fixed.
- The interaction term: three-point volume rate × three-point percentage:
    - Significantly negative coefficient: -38.86741
    - Teams who overly rely on the three-pointer must also be efficient at shooting threes. If not, their overreliance on threes will severely hurt them.
    - Every point increase in your interaction term, → 41 less expected wins.

However, the interaction term between the two variables is a decimal, meaning that your games should not be impacted so severely.

- As an example:
    - Three-point attempt rate (amount of 3P per possession) = 0.35
    - Three-point shooting percentage = 33%
    - Interaction term value = 0.1155
    - Expected wins: (Interaction term value) × -38.86741 = -4.5.

This means that if your team had a .35 three-point attempt rate and shot 33% from deep, your expected wins go down by 4.5 wins.

- Playoff Indicator: grants teams an extra 2.92 wins if their team made the playoffs. This shows that some teams might have some unmeasured qualities, such as
- Clutch performance
- Coaching
- Roster depth

Model Examples

An example of our model can be used with the 2024-2025 Boston Celtics:

- Average age: 29 years old
- Offensive rating: 120.6
- Defensive rating: 111.2
- 3-pt attempt rate: .538
- 3-pt shooting percentage: 37%
- Made the playoffs this year

The data was provided before the season ended, with still around 12 games left to play. Using this data, our model predicted the Celtics expected wins of 61, which is exactly what they won. This means that the Celtics performed perfectly to their expectations this year, and are one of the best teams in the league and look to win back-to-back championships.

Another example is given by the 2024 Milwaukee Bucks:

- The Bucks won 58 games that season, earning the one-seed in the Eastern Conference.
- However, our model only predicted the Bucks expected wins given their season data to be 46 wins.

Our model being off games seems to be very alarming, but the Bucks ended up losing to the 8-seeded Miami Heat in the first round in only 5 games. This shows that our model was correct in assuming the Bucks vastly outperformed their expectations during the regular season.

They had significant flaws that year, but those flaws were masked by their dominant regular season. However, our model proved that the Bucks should not have been as good as their record showed, which was proved right in their playoff performance.

Ultimately, we were able to develop a regression model that predicts a team's win total based on metrics such as offensive and defensive ratings, shooting efficiency, average age of a team, and whether the team made the playoffs, to generate an expected number of wins for each team. Comparing these predicted wins to actual outcomes allows us to pinpoint teams that exceeded expectations, underachieved, or met projections. This framework provides valuable context for guiding future decisions related to player acquisitions, tactical adjustments, and resource allocation.