Score Big: How Data Can Create a Betting Edge in Baseball

STAT 139: Introduction to Linear Models

Professor: Professor Xenakis

Date: December 22, 2024

Group Members:
Sukhraj Dulay
Hamzeh Hamdan
Evelyn Tjoa
Victor Zeidenfeld

Contents

Introduction . . . 2
- **Data Description and Source** . . . 2
- **Data Preparation and Cleaning** . . . 2

Exploratory Data Analysis (EDA) . . . 5
- **Binary Playoff Variable** . . . 7

Visualizations . . . 8

Hypothesis Testing . . . 13
- **Hypothesis Models** . . . 14
- **Implementation** . . . 14
- **Interpretations** . . . 24
- **Contextual Analysis** . . . 25
- **Discussion on Playoff Format Change** . . . 25
- **Prediction for 2025** . . . 28
- **Initial Model Specification** . . . 28
- **Hypothesis of Interest** . . . 28

Implementation . . . 29
- **Baseline Model** . . . 29
- **Principal Component Analysis (PCA)** . . . 31

Conclusion . . . 31
- **Broader Implications** . . . 32
- **Acknowledgments** . . . 32

Appendices . . . 33

Introduction

The dynamics of baseball have long fascinated analysts and enthusiasts, with statistical metrics playing a pivotal role in understanding and predicting team performance. Over the past decade, advanced analytics have provided great insights into hitting, pitching, and fielding, reshaping how teams approach the game strategically. This study dives into how specific metrics influence success. Understanding the nuances of these metrics and structural changes is not only valuable for team management but also for fans and stakeholders who wish to engage more deeply with the sport. The findings from this study can influence roster-building decisions, tactical approaches, and even broadcasting narratives, making baseball analytics accessible and impactful across various domains.

By combining statistical rigor with domain-specific knowledge, this study seeks to fill the gap between traditional baseball insights and modern predictive techniques.

Data Description and Source

Data Description

The dataset utilized for this study includes a comprehensive range of hitting, pitching, and fielding metrics, both cumulative and average from https://www.baseball-reference.com . These variables cover traditional statistics, such as home runs and earned run average (ERA), alongside advanced metrics like exit velocity (EV) and defensive runs saved (RDRS). To ensure interpretability, some metrics have been normalized where appropriate, enabling consistent comparisons across teams. The dataset is structured as panel data spanning ten years, with teams as entities and seasons as time periods. This structure allows for the inclusion of fixed effects to account for team-specific and time-specific variations.

Data Source

The primary source of data is https://www.baseball-reference.com, an open-access database renowned for its reliability and extensive coverage of baseball statistics. Baseball Reference actively updates its metrics, offering downloadable CSV files for specific categories, including:

- Standard Pitching
- Advanced Pitching
- Standard Batting
- Advanced Batting
- Fielding

The CSV files for each category were loaded and merged year by year to form a cohesive dataset covering the 2015–2024 MLB seasons.

The decision to use Baseball Reference stems from its robust documentation and ease of accessibility. Additionally, its standardization of historical metrics ensures compatibility across years, which is crucial for longitudinal studies like this.

Data Preparation and Cleaning

```
data1 <- read.csv("2015/2015_Stdpitching.csv")
data2 <- read.csv("2015/2015advpitching.csv",skip = 1)
data3 <- read.csv("2015/advanced_batting.csv",skip=1)
data4 <- read.csv("2015/fielding2015.csv")
data5 <- read.csv("2015/Standard_Batting_2015.csv")
#install.packages("plyr")
library(plyr)
```

```
data_2015 <- join_all(list(data1,data2,data3,data4,data5), by ="Tm",
                      type = 'left')
data_2015$season <- rep(2015)
data_2015 <- data_2015[1:30,]
```

```
data1.2 <- read.csv("2016/2016_Stdpitching.csv")
data2.2 <- read.csv("2016/2016_advpitching.csv",skip = 1)
data3.2 <- read.csv("2016/advanced_batting.csv",skip=1)
data4.2 <- read.csv("2016/fielding2016.csv")
data5.2 <- read.csv("2016/Standard_Batting_2016.csv")

data_2016 <- join_all(list(data1.2,data2.2,data3.2,data4.2,data5.2), by ="Tm",
                      type = 'left')
data_2016$season <- rep(2016)
data_2016 <- data_2016[1:30,]
```

```
data1.3 <- read.csv("2017/2017_Stdpitching.csv")
data2.3 <- read.csv("2017/2017_advpitching.csv",skip = 1)
data3.3 <- read.csv("2017/advanced_batting.csv",skip=1)
data4.3 <- read.csv("2017/fielding2017.csv")
data5.3 <- read.csv("2017/Standard_Batting_2017.csv")

data_2017 <- join_all(list(data1.3,data2.3,data3.3,data4.3,data5.3), by ="Tm",
                      type = 'left')
data_2017$season <- rep(2017)
data_2017 <- data_2017[1:30,]
```

```
data1.4 <- read.csv("2018/2018_Stdpitching.csv")
data2.4 <- read.csv("2018/2018_advpitching.csv",skip = 1)
data3.4 <- read.csv("2018/advanced_batting.csv", skip = 1)
data4.4 <- read.csv("2018/fielding2018.csv")
data5.4 <- read.csv("2018/standard_batting_2018.csv")

data_2018 <- join_all(list(data1.4,data2.4,data3.4,data4.4,data5.4), by ="Tm",
                      type = 'left')
data_2018$season <- rep(2018)
data_2018 <- data_2018[1:30,]
```

```
data1.5 <- read.csv("2019/2019_Stdpitching.xls - 2019_Stdpitching.xls.csv")
data2.5 <- read.csv("2019/2019_advpitching.xls - 2019_advpitching.xls.csv",skip = 1)
data3.5 <- read.csv("2019/advanced_batting.xls - advanced_batting.xls.csv", skip =1)
data4.5 <- read.csv("2019/fielding2019.xlsx - Worksheet.csv")
data5.5 <- read.csv("2019/standard_batting_2019.xls - standard_batting_2019.xls.csv")

data_2019 <- join_all(list(data1.5,data2.5,data3.5,data4.5,data5.5), by ="Tm",
                      type = 'left')
data_2019$season <- rep(2019)
data_2019 <- data_2019[1:30,]
```

```
data1.6 <- read.csv("2020/2020_Stdpitching.xls - 2020_Stdpitching.xls.csv")
data2.6 <- read.csv("2020/2020_advpitching.xls - 2020_advpitching.xls.csv",skip = 1)
data3.6 <- read.csv("2020/advanced_batting.xls - advanced_batting.xls.csv", skip =1)
data4.6 <- read.csv("2020/fielding2020.xlsx - Worksheet.csv")
data5.6 <- read.csv("2020/standard_batting_2020.xls - standard_batting_2020.xls.csv")
```

```
data_2020 <- join_all(list(data1.6,data2.6,data3.6,data4.6,data5.6), by ="Tm",
                      type = 'left')
data_2020$season <- rep(2020)
data_2020 <- data_2020[1:30,]
```

```
data1.7 <- read.csv("2021/2021_Stdpitching.xls - 2021_Stdpitching.xls.csv")
data2.7 <- read.csv("2021/2015_Stdpitching,2015advpitching,advanced_batting.csv",skip = 1)
data3.7 <- read.csv("2021/advanced_batting.xls - advanced_batting.xls.csv", skip =1)
data4.7 <- read.csv("2021/fielding2021.xlsx - Worksheet.csv")
data5.7 <- read.csv("2021/standard_batting_2021.xls - standard_batting_2021.xls.csv")

data_2021 <- join_all(list(data1.7,data2.7,data3.7,data4.7,data5.7), by ="Tm",
                      type = 'left')
data_2021$season <- rep(2021)
data_2021 <- data_2021[1:30,]
```

```
data1.8 <- read.csv("2022/2022_Stdpitching.xls - 2022_Stdpitching.xls.csv")
data2.8 <- read.csv("2022/2022_advpitching.xls - 2022_advpitching.xls.csv",skip = 1)
data3.8 <- read.csv("2022/advanced_batting.xls - advanced_batting.xls.xls.csv", skip =1)
data4.8 <- read.csv("2022/fielding2022.xlsx - Worksheet.csv")
data5.8 <- read.csv("2022/standard_batting_2022.xls - standard_batting_2022.xls.csv")

data_2022 <- join_all(list(data1.8,data2.8,data3.8,data4.8,data5.8), by ="Tm",
                      type = 'left')
data_2022$season <- rep(2022)
data_2022 <- data_2022[1:30,]
```

```
data1.9 <- read.csv("2023/2023_Stdpitching.xls - 2023_Stdpitching.xls.csv")
data2.9 <- read.csv("2023/2023_advpitching.xls - 2023_advpitching.xls.csv",skip = 1)
data3.9 <- read.csv("2023/advanced_batting.xls - advanced_batting.xls.csv", skip =1)
data4.9 <- read.csv("2023/fielding2023.xlsx - Worksheet.csv")
data5.9 <- read.csv("2023/standard_batting_2023.xls.csv")

data_2023 <- join_all(list(data1.9,data2.9,data3.9,data4.9,data5.9), by ="Tm",
                      type = 'left')
data_2023$season <- rep(2023)
data_2023 <- data_2023[1:30,]
```

```
data1.10 <- read.csv("2024/2024_Stdpitching.xls.csv")
data2.10 <- read.csv("2024/2024_advpitching.xls.csv",skip = 1)
data3.10 <- read.csv("2024/advanced_batting.xls.csv", skip =1)
data4.10 <- read.csv("2024/fielding2024 (1).csv")
data5.10 <- read.csv("2024/standard_batting_2024.xls.csv")

data_2024 <- join_all(list(data1.10,data2.10,data3.10,data4.10,data5.10), by ="Tm",
                      type = 'left')
data_2024$season <- rep(2024)
data_2024 <- data_2024[1:30,]
```

```
#bind all years
mlb <- rbind(data_2015,data_2016,data_2017,data_2018,data_2019,data_2020,
             data_2021,data_2022,data_2023,data_2024)
```

Exploratory Data Analysis (EDA)

Our initial analysis focuses on understanding the variability in key metrics. The following variables are of interest:
- **SB:** Stolen Bases Percentage
- **PAge:** Average Pitcher Age
- **BatAge:** Average Batter Age
- **EV:** Average Exit Velocity
- **ERA:** Earned Run Average
- **BB:** Bases on Balls
- **OPS:** On-base Percentage Plus Slugging
- **HR:** Home Run Percentage
- **RS:** Runner Support Percentage
- **E:** Errors Committed
- **RDRS:** Defensive Runs Saved

Each variable is numeric with no missing values, allowing for robust statistical analysis. Spread measures like standard deviation (SD) and interquartile range (IQR) help quantify league variability.

```
mlb$RS. <- as.numeric(gsub("%", "", mlb$RS.))
mlb$BB. <- as.numeric(gsub("%", "", mlb$BB.))
colnames(mlb)[104] <- "HR_bat"
mlb$HR_percentage <- mlb$HR_bat / mlb$AB
mlb$SB. <- as.numeric(gsub("%", "", mlb$SB.))
```

Data Cleaning and Transformation: Summary statistics for each variable:

```
columns <- c("SB", "PAge", "BatAge", "EV", "ERA", "BB.", "OPS", "HR_percentage", "RS.",
             "E", "Rdrs")

for (col in columns) {
  cat("Statistics for column:", col, "\n")
  column_data <- mlb[[col]]

  cat("- Number of non-missing observations:", sum(!is.na(column_data)), "\n")
  cat("- Number of missing observations:", sum(is.na(column_data)), "\n")
  cat("- Mean:", mean(column_data, na.rm = TRUE), "\n")
  cat("- Median:", median(column_data, na.rm = TRUE), "\n")
  cat("- Standard Deviation:", sd(column_data, na.rm = TRUE), "\n")
  cat("- IQR:", IQR(column_data, na.rm = TRUE), "\n\n")
}
```

```
## Statistics for column: SB
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 83.42333
## - Median: 80.5
## - Standard Deviation: 36.41544
## - IQR: 46.25
##
## Statistics for column: PAge
## - Number of non-missing observations: 300
```

```
## - Number of missing observations: 0
## - Mean: 28.581
## - Median: 28.6
## - Standard Deviation: 1.117124
## - IQR: 1.6
##
## Statistics for column: BatAge
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 28.15733
## - Median: 28.2
## - Standard Deviation: 0.9958675
## - IQR: 1.3
##
## Statistics for column: EV
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 88.34533
## - Median: 88.4
## - Standard Deviation: 0.711735
## - IQR: 1
##
## Statistics for column: ERA
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 4.221667
## - Median: 4.145
## - Standard Deviation: 0.5613581
## - IQR: 0.82
##
## Statistics for column: BB.
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 8.410667
## - Median: 8.4
## - Standard Deviation: 0.9420072
## - IQR: 1.3
##
## Statistics for column: OPS
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 0.73098
## - Median: 0.729
## - Standard Deviation: 0.04492029
## - IQR: 0.06225
##
## Statistics for column: HR_percentage
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 0.03513405
## - Median: 0.03475591
## - Standard Deviation: 0.006784089
## - IQR: 0.009821661
##
```

```
## Statistics for column: RS.
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 30.33
## - Median: 30
## - Standard Deviation: 2.036728
## - IQR: 3
##
## Statistics for column: E
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 85.1
## - Median: 88
## - Standard Deviation: 21.87155
## - IQR: 21
##
## Statistics for column: Rdrs
## - Number of non-missing observations: 300
## - Number of missing observations: 0
## - Mean: 9.783333
## - Median: 7
## - Standard Deviation: 40.5919
## - IQR: 48
```

Binary Playoff Variable

To facilitate our analysis, we coded a binary variable, `playoffs`, indicating whether a team made the playoffs (1) or not (0). This variable accounts for playoff format changes and adjusts for variations across seasons.

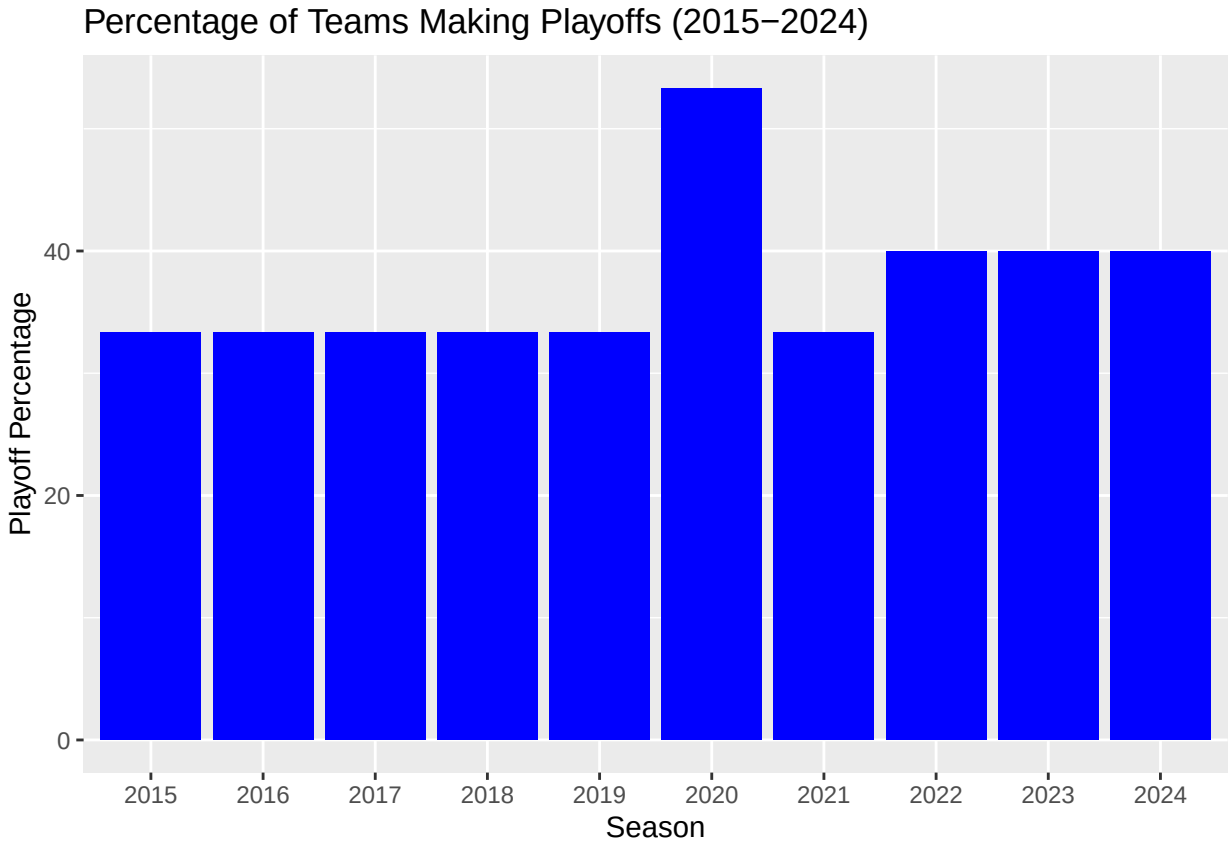
```r
#Coding binary variable for whether or not team reached the playoffs
# potential to add teams that just missed playoffs in original years in order to train model for curren

mlb$playoffs <- ifelse(
 ( (mlb$Tm %in% c("Kansas City Royals", "Toronto Blue Jays", "New York Yankees", "Texas Rangers",
                  "Houston Astros", "St. Louis Cardinals", "Los Angeles Dodgers", "New York Mets",
                  "Pittsburgh Pirates", "Chicago Clubs") & mlb$season == 2015)|
  (mlb$Tm %in% c("Texas Rangers", "Cleveland Indians","Boston Red Sox", "Toronto Blue Jays",
                 "Baltimore Orioles", "Chicago Cubs","Washington Nationals", "Los Angeles Dodgers",
                 "New York Mets", "San Francisco Giants") & mlb$season ==2016)|
    (mlb$Tm %in% c("Cleveland Indians","Boston Red Sox","Houston Astros","New York Yankees",
                   "Minnesota Twins","Chicago Cubs","Washington Nationals", "Los Angeles Dodgers",
                   "Arizona Diamondbacks","Colorado Rockies") & mlb$season == 2017)|
    (mlb$Tm %in% c("Cleveland Indians","Boston Red Sox","Houston Astros","New York Yankees",
                   "Oakland Athletics", "Milwaukee Brewers", "Los Angeles Dodgers", "Atlanta Braves",
                   "Colorado Rockies","Chicago Cubs") & mlb$season == 2018)|
    (mlb$Tm %in% c("Houston Astros","New York Yankees","Minnesota Twins","Oakland Athletics",
                   "Tampa Bay Rays","Los Angeles Dodgers", "Atlanta Braves","St. Louis Cardinals",
                   "Washington Nationals","Milwaukee Brewers") & mlb$season == 2019)|
    (mlb$Tm %in% c("Oakland Athletics","Tampa Bay Rays","Minnesota Twins","Cleveland Indians",
                   "Houston Astros","New York Yankees","Chicago White Sox","Toronto Blue Jays",
                   "Los Angeles Dodgers", "Atlanta Braves","Chicago Cubs","San Diego Padres",
                   "St. Louis Cardinals","Miami Marlins","Cincinnati Reds","Milwaukee Brewers") &
       mlb$season == 2020)|
```

```
    (mlb$Tm %in% c("Tampa Bay Rays","New York Yankees","Chicago White Sox","Houston Astros",
                   "Boston Red Sox","Milwaukee Brewers","Los Angeles Dodgers", "Atlanta Braves",
                   "San Francisco Giants","St. Louis Cardinals") & mlb$season == 2021)|
    (mlb$Tm %in% c("New York Yankees","Houston Astros", "Cleveland Guardians","Toronto Blue Jays",
                   "Seattle Mariners","Tampa Bay Rays","Los Angeles Dodgers", "Atlanta Braves",
                   "San Diego Padres","St. Louis Cardinals", "New York Mets","Philadelphia Phillies") &
       mlb$season == 2022)|
    (mlb$Tm %in% c("Baltimore Orioles","Toronto Blue Jays","Texas Rangers","Tampa Bay Rays",
                   "Houston Astros","Minnesota Twins","Milwaukee Brewers","Los Angeles Dodgers",
                   "Atlanta Braves","Philadelphia Phillies","Miami Marlins","Arizona Diamondbacks") &
       mlb$season == 2023)|
    (mlb$Tm %in% c("New York Yankees","Houston Astros", "Cleveland Guardians","Baltimore Orioles",
                   "Kansas City Royals","Detroit Tigers","Los Angeles Dodgers", "Atlanta Braves",
                   "Philadelphia Phillies","Milwaukee Brewers","New York Mets","San Diego Padres") &
       mlb$season == 2024))
    ,1,0)
mlb$playoffs <- as.factor(mlb$playoffs)
```

Visualizations

Playoff Format Changes A bar chart was created to depict how the playoff format has evolved over time, specifically focusing on the percentage of teams qualifying for the postseason.

```
library(ggplot2)
playoff_teams <- c(10, 10, 10, 10, 10, 16, 10, 12, 12, 12)
teams <- c(30,30,30,30,30,30,30,30,30,30)

playoff_percent <- (playoff_teams / teams) * 100

seasons <- 2015:2024

playoff_percent <- data.frame(
  season = seasons,
  playoff_percent = playoff_percent
)

ggplot(playoff_percent, aes(x = factor(season), y = playoff_percent)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(
    title = "Percentage of Teams Making Playoffs (2015-2024)",
    x = "Season",
    y = "Playoff Percentage"
  )
```



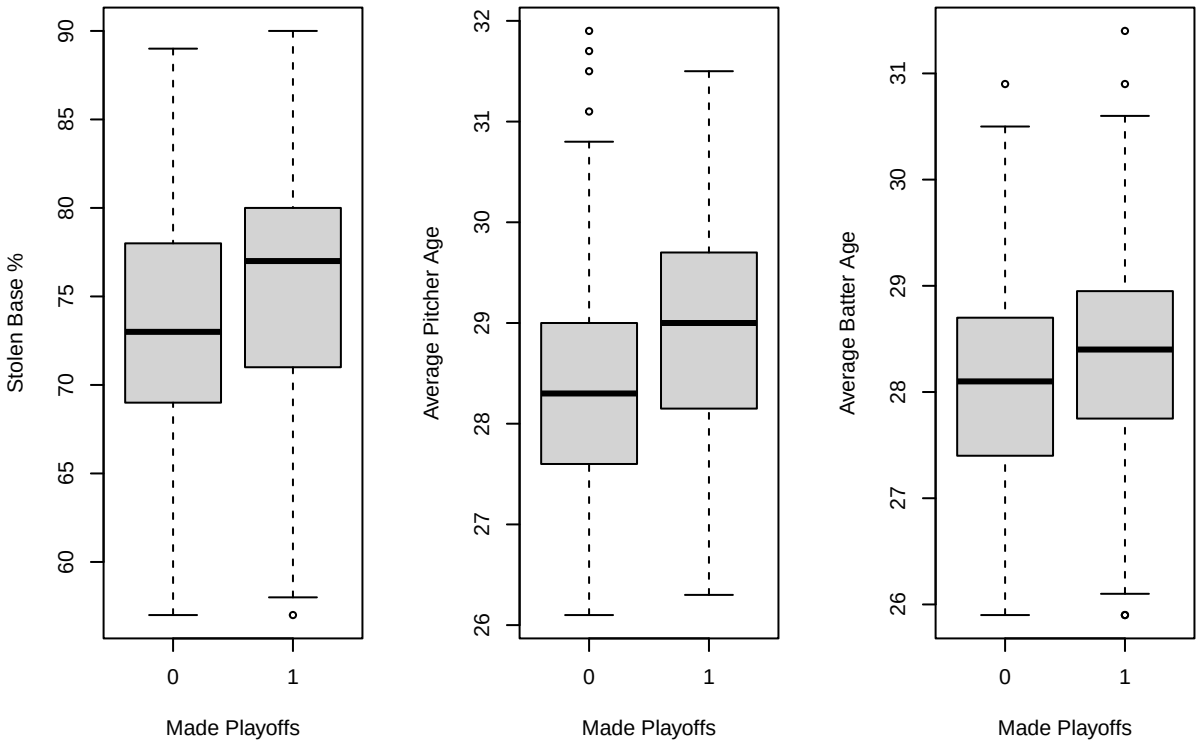
Analysis for Bar Plot: The bar plot illustrates the percentage of teams making the playoffs across the seasons from 2015 to 2024. It highlights significant changes in playoff formats:

1. **Consistency (2015–2019, 2021)**: The percentage remains stable at around 33%, indicating a fixed format of 10 playoff teams out of 30.
2. **COVID-19 Impact (2020)**: The percentage spikes to over 50%, showing the temporary expansion of playoff teams to 16, offering a unique case for further statistical exploration.
3. **Post-2021 Expansion**: From 2022 onward, the percentage increases to 40%, reflecting the introduction of a 12-team playoff format. This change creates an exogenous factor suitable for quasi-experimental studies on team success under differing constraints.

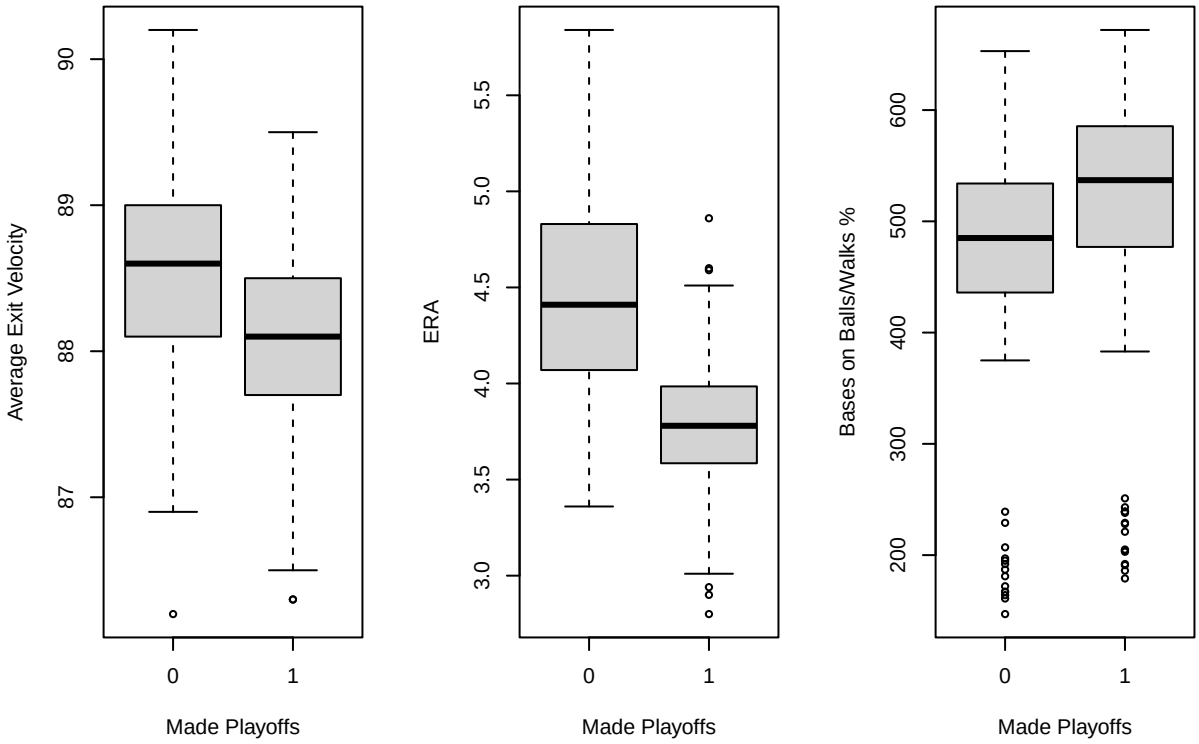
Key takeaway: The structural changes in playoff eligibility significantly affect team dynamics and strategy, offering a natural experiment to analyze how changes in competitive incentives impact performance.

Comparing Playoff vs. Non-Playoff Teams Boxplots were generated to compare playoff and non-playoff teams across several predictors, revealing systematic differences between these groups.

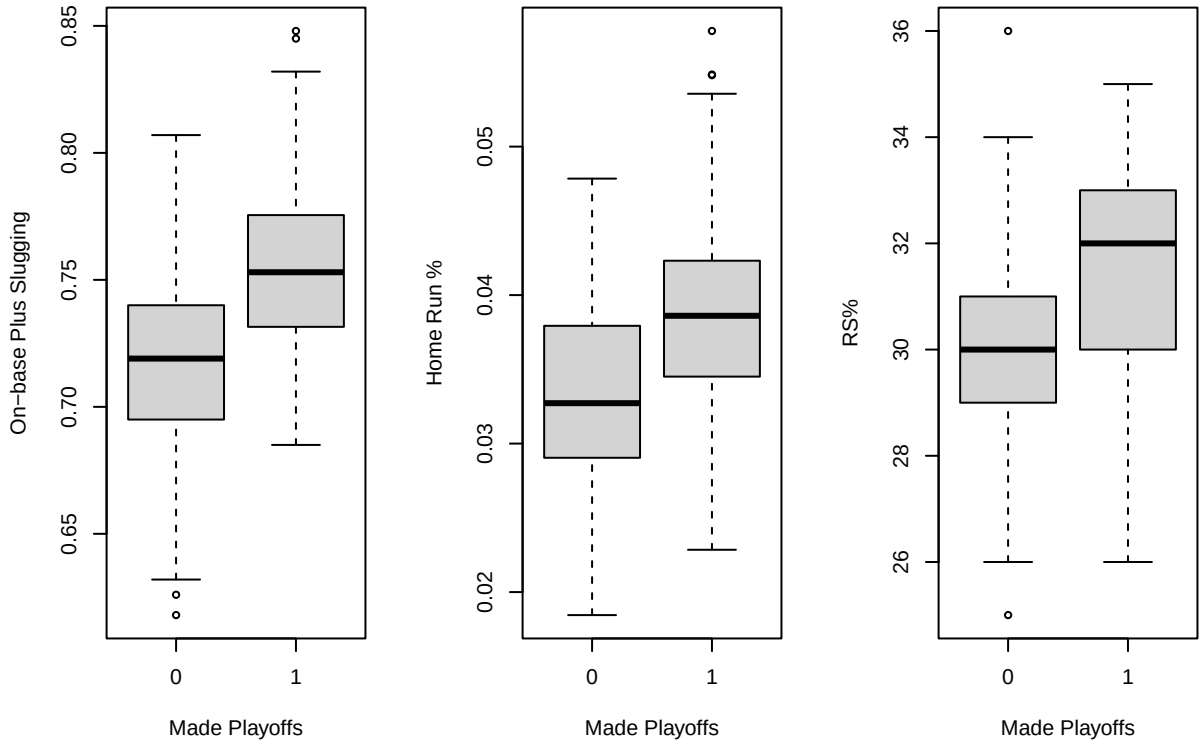
```
par(mfrow = c(1, 3))


plot(mlb$playoffs, mlb$SB., xlab = "Made Playoffs", ylab = "Stolen Base %")
plot(mlb$playoffs, mlb$PAge, xlab = "Made Playoffs", ylab = "Average Pitcher Age")
plot(mlb$playoffs, mlb$BatAge, xlab = "Made Playoffs", ylab = "Average Batter Age")
```



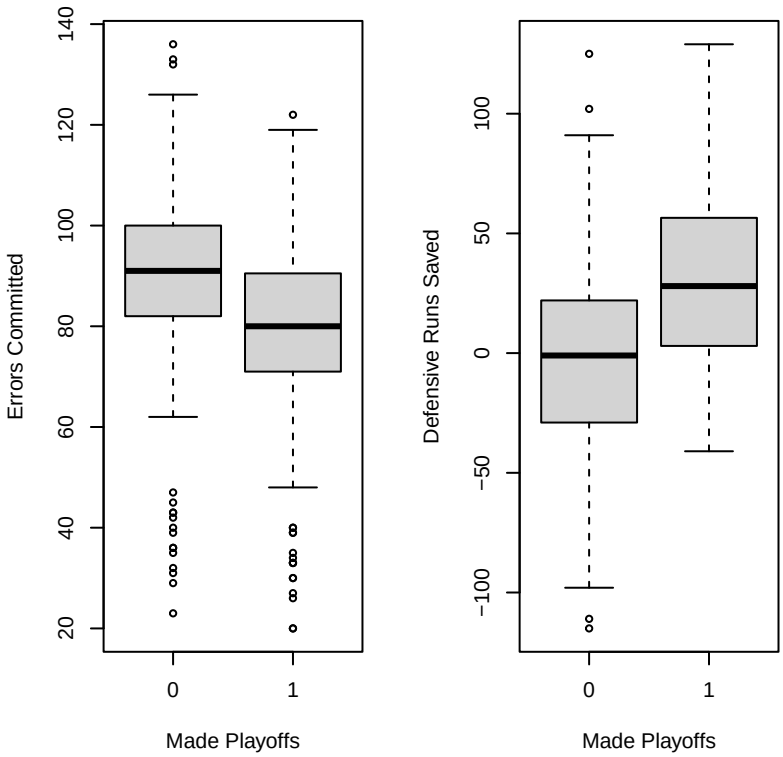

```
plot(mlb$playoffs, mlb$EV, xlab = "Made Playoffs", ylab = "Average Exit Velocity")
plot(mlb$playoffs, mlb$ERA, xlab = "Made Playoffs", ylab = "ERA")

colnames(mlb)[108] <- "BB_bat"
plot(mlb$playoffs, mlb$BB_bat, xlab = "Made Playoffs", ylab = "Bases on Balls/Walks %")
```



```r
colnames(mlb)[113] <- "OPS_bat"
plot(mlb$playoffs, mlb$OPS_bat, xlab = "Made Playoffs", ylab = "On-base Plus Slugging")

plot(mlb$playoffs, mlb$HR_percentage, xlab = "Made Playoffs", ylab = "Home Run %")
plot(mlb$playoffs, mlb$RS., xlab = "Made Playoffs", ylab = "RS%")
```




```r
plot(mlb$playoffs, mlb$E, xlab = "Made Playoffs", ylab = "Errors Committed")
plot(mlb$playoffs, mlb$Rdrs, xlab = "Made Playoffs", ylab = "Defensive Runs Saved")

par(mfrow = c(1, 1))
```



Analysis for Boxplots: The boxplots compare various predictors for teams that made the playoffs (1) versus those that did not (0):

1. Stolen Base Percentage (SB%)

- **Observation**: Teams that made the playoffs generally have a higher median stolen base percentage, indicating the value of aggressive base running in successful team strategies.
- **Insight**: This metric could be explored further to determine its role in offensive efficiency and run production.

2. Pitcher and Batter Average Ages (PAge, BatAge)

- **Observation**: Playoff teams have slightly older average pitcher and batter ages, suggesting experience might play a role in team success.
- **Insight**: Analyzing whether these age differences are statistically significant could reveal if veteran players contribute to key metrics such as ERA or OPS.

3. Average Exit Velocity (EV)

- **Observation**: Higher average exit velocity among playoff teams highlights the importance of hitting power and quality of contact.
- **Insight**: Including EV as a predictor in models could better explain offensive success rates.

4. Earned Run Average (ERA)

- **Observation**: Playoff teams show lower ERA, underscoring pitching quality as a crucial determinant of success.
- **Insight**: Future analyses can evaluate the correlation between ERA and playoff advancement.

5. Bases on Balls Percentage (BB%)

- **Observation**: Higher median BB% for playoff teams suggests plate discipline and drawing walks are correlated with success.
- **Insight**: Combining this with metrics like OBP could refine understanding of offensive strategies.

6. On-Base Plus Slugging (OPS)

- **Observation**: Playoff teams exhibit higher OPS, reinforcing the idea that teams with better offensive efficiency have greater success.
- **Insight**: OPS could be broken into its components to identify whether on-base ability or power-hitting has a stronger impact.

7. Home Run Percentage (HR%)

- **Observation**: Playoff teams have a slightly higher HR%, indicating the influence of power hitting.
- **Insight**: This metric's interaction with factors like pitching and fielding could reveal its true significance.

8. Runner Support Percentage (RS%)

- **Observation**: Higher RS% in playoff teams aligns with the importance of converting baserunners into runs.
- **Insight**: Decomposing RS% by situations (e.g., runners in scoring position) might offer additional granularity.

9. Errors Committed (E)

- **Observation**: Lower error rates in playoff teams highlight the role of defensive consistency.
- **Insight**: Including advanced fielding metrics like DefEff could add predictive power to models.

10. Defensive Runs Saved (Rdrs)

- **Observation**: Positive Rdrs values for playoff teams underscore the importance of defensive contributions.
- **Insight**: Examining this in tandem with pitching metrics like WHIP might reveal how defense mitigates run production by opponents.

General Observations: The boxplots collectively emphasize the multidimensional nature of success in MLB, encompassing pitching, hitting, running, and defense. The insights suggest avenues for more nuanced predictive modeling, particularly through interaction terms and stratified analyses. Incorporating these metrics into linear or logistic regression models will likely improve predictive accuracy for playoff qualification and postseason success.

```
write.csv(mlb, "mlb_edits.csv", row.names = FALSE)
```

Hypothesis Testing

Introduction This section examines the impact of Intentional Bases on Balls (IBBs) on a team's Win-Loss percentage. The inquiry stems from the decision by pitchers to intentionally walk strong power hitters to reduce the risk of conceding home runs. The objective is to determine whether this strategy negatively affects a team's Win-Loss percentage by potentially disrupting offensive output. Conversely, it is plausible that teams with high IBB rates capitalize on these intentional walks, turning them into an advantage. Additionally, having high IBB is likely positively correlated with Home Runs.

Measurement Approach To normalize across teams and control for differences in plate appearances, IBB is measured as a proportion of IBB to At Bats (IBB/AB) $\times$ 100. This representation captures intentional walks as a function of offensive opportunities.

Model Adjustments Team and season fixed effects were incorporated to control for factors varying across teams and seasons. These include:

- **Seasonal Dynamics**: Factors like rule changes or shortened seasons (e.g., 2020 during COVID).
- **Team-Level Constraints**: Budgetary limitations or offensive strength variations.

In a second model, **On-Base Percentage (OBP)** and **Runs Allowed per Game (RA/G)** were added as control variables. OBP accounts for offensive strength and the likelihood of intentional walks, while RA/G represents defensive performance. Together, these adjustments ensure the analysis isolates the relationship between IBB and team performance.

Hypothesis Models

Model 1

$$W/L_{ijt} = \beta_0 + \beta_1 \text{IBB}_{ijt} + \gamma_{ijt} + \delta_{ijt}$$

Model 2

$$W/L_{ijt} = \beta_0 + \beta_1 \text{IBB}_{ijt} + \beta_2 \text{OBP}_{ijt} + \beta_3 \text{RA/G}_{ijt} + \gamma_{ijt} + \delta_{ijt}$$

Null and Alternative Hypotheses

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

Implementation

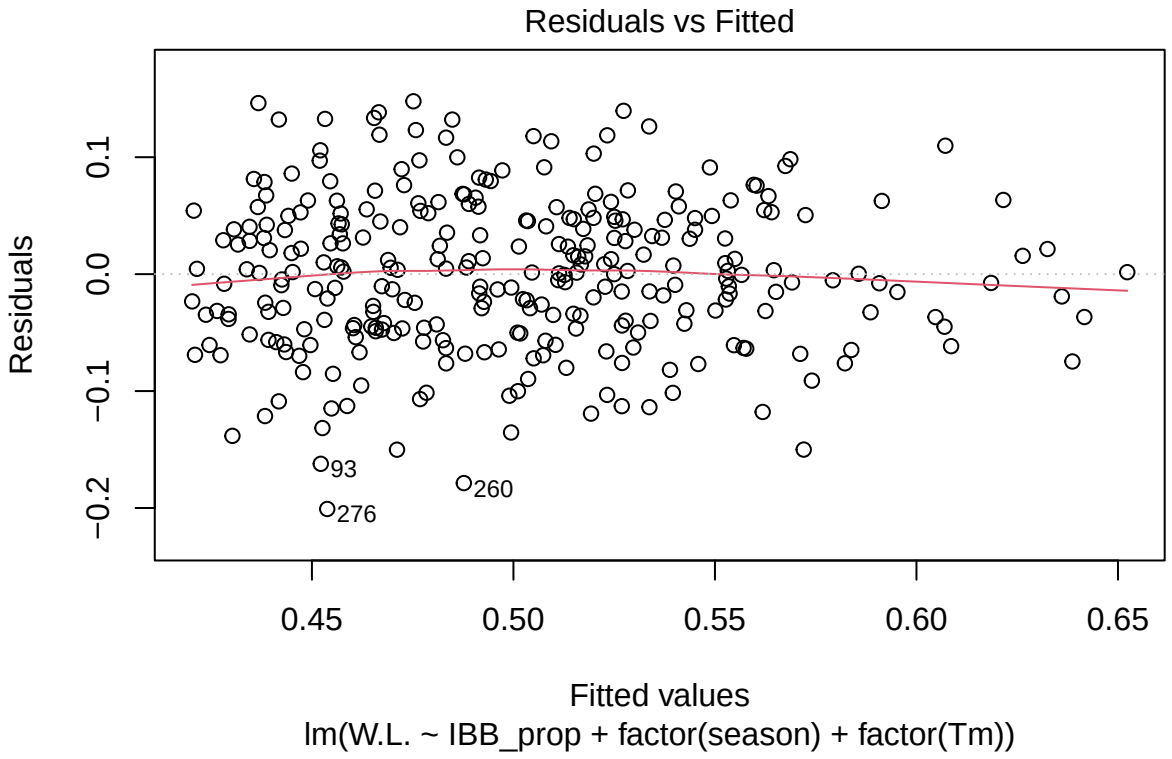
```
colnames(mlb)[120] <- "IBB_bat"
colnames(mlb)[99] <- "AB_bat"
colnames(mlb)[100] <- "Runs_bat"

mlb$IBB_prop <- (mlb$IBB_bat / mlb$AB_bat) * 100
mlb$OBP_prop <- mlb$OBP * 100

# model 1
IBB <- lm(W.L. ~ IBB_prop + factor(season) + factor(Tm), data = mlb)
summary(IBB)
```

```
##
## Call:
## lm(formula = W.L. ~ IBB_prop + factor(season) + factor(Tm), data = mlb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20081 -0.04585  0.00117  0.04725  0.14782
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                              0.403863   0.031533  12.808  < 2e-16
## IBB_prop                                 0.112176   0.029289   3.830 0.000161
## factor(season)2016                       0.003791   0.018240   0.208 0.835535
## factor(season)2017                       0.001168   0.018239   0.064 0.948978
## factor(season)2018                       0.003922   0.018241   0.215 0.829918
## factor(season)2019                       0.016182   0.018591   0.870 0.384877
## factor(season)2020                       0.028572   0.019482   1.467 0.143713
## factor(season)2021                       0.018203   0.018697   0.974 0.331154
## factor(season)2022                       0.036747   0.020291   1.811 0.071295
## factor(season)2023                       0.037029   0.020319   1.822 0.069551
## factor(season)2024                       0.034474   0.020042   1.720 0.086618
## factor(Tm)Atlanta Braves                 0.042312   0.031616   1.338 0.181969
## factor(Tm)Baltimore Orioles              0.005712   0.032067   0.178 0.858756
## factor(Tm)Boston Red Sox                 0.043738   0.031295   1.398 0.163435
## factor(Tm)Chicago Cubs                   0.049364   0.031443   1.570 0.117649
## factor(Tm)Chicago White Sox             -0.001201   0.032185  -0.037 0.970252
## factor(Tm)Cincinnati Reds               -0.018561   0.031379  -0.592 0.554686
## factor(Tm)Cleveland Guardians            0.023579   0.047290   0.499 0.618482
## factor(Tm)Cleveland Indians              0.097177   0.034764   2.795 0.005574
## factor(Tm)Colorado Rockies              -0.032625   0.031290  -1.043 0.298090
## factor(Tm)Detroit Tigers                -0.018070   0.031881  -0.567 0.571345
## factor(Tm)Houston Astros                 0.113464   0.031449   3.608 0.000371
## factor(Tm)Kansas City Royals            -0.006945   0.031923  -0.218 0.827955
## factor(Tm)Los Angeles Angels            -0.023010   0.032176  -0.715 0.475183
## factor(Tm)Los Angeles Angels of Anaheim  0.050729   0.074408   0.682 0.495993
## factor(Tm)Los Angeles Dodgers            0.136283   0.031453   4.333 2.11e-05
## factor(Tm)Miami Marlins                 -0.031617   0.031327  -1.009 0.313798
## factor(Tm)Milwaukee Brewers              0.047744   0.031294   1.526 0.128317
## factor(Tm)Minnesota Twins                0.043844   0.031502   1.392 0.165192
## factor(Tm)New York Mets                  0.028222   0.031289   0.902 0.367920
## factor(Tm)New York Yankees               0.103421   0.031472   3.286 0.001157
## factor(Tm)Oakland Athletics              0.021460   0.031981   0.671 0.502802
## factor(Tm)Philadelphia Phillies          0.005182   0.031311   0.166 0.868676
## factor(Tm)Pittsburgh Pirates            -0.028911   0.031289  -0.924 0.356351
## factor(Tm)San Diego Padres               0.018263   0.031352   0.583 0.560718
## factor(Tm)San Francisco Giants           0.019191   0.031297   0.613 0.540280
## factor(Tm)Seattle Mariners               0.048944   0.031685   1.545 0.123638
## factor(Tm)St. Louis Cardinals            0.064049   0.031328   2.044 0.041923
## factor(Tm)Tampa Bay Rays                 0.085155   0.031554   2.699 0.007420
## factor(Tm)Texas Rangers                  0.012411   0.031752   0.391 0.696218
## factor(Tm)Toronto Blue Jays              0.063180   0.032178   1.963 0.050667
## factor(Tm)Washington Nationals          -0.010521   0.031533  -0.334 0.738903
##
## (Intercept)                             ***
## IBB_prop                                ***
## factor(season)2016
## factor(season)2017
## factor(season)2018
## factor(season)2019
## factor(season)2020
## factor(season)2021
## factor(season)2022                      .
## factor(season)2023                      .
## factor(season)2024                      .
```

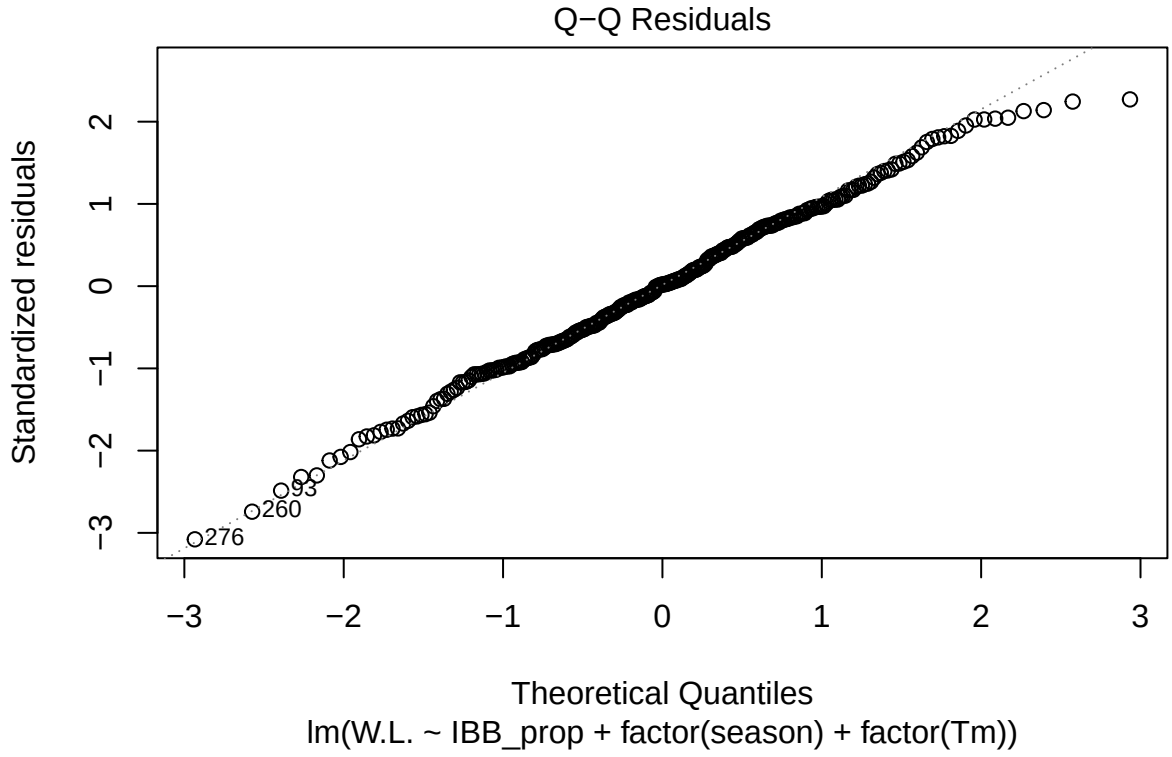
```
## factor(Tm)Atlanta Braves
## factor(Tm)Baltimore Orioles
## factor(Tm)Boston Red Sox
## factor(Tm)Chicago Cubs
## factor(Tm)Chicago White Sox
## factor(Tm)Cincinnati Reds
## factor(Tm)Cleveland Guardians
## factor(Tm)Cleveland Indians             **
## factor(Tm)Colorado Rockies
## factor(Tm)Detroit Tigers
## factor(Tm)Houston Astros                ***
## factor(Tm)Kansas City Royals
## factor(Tm)Los Angeles Angels
## factor(Tm)Los Angeles Angels of Anaheim
## factor(Tm)Los Angeles Dodgers           ***
## factor(Tm)Miami Marlins
## factor(Tm)Milwaukee Brewers
## factor(Tm)Minnesota Twins
## factor(Tm)New York Mets
## factor(Tm)New York Yankees              **
## factor(Tm)Oakland Athletics
## factor(Tm)Philadelphia Phillies
## factor(Tm)Pittsburgh Pirates
## factor(Tm)San Diego Padres
## factor(Tm)San Francisco Giants
## factor(Tm)Seattle Mariners
## factor(Tm)St. Louis Cardinals           *
## factor(Tm)Tampa Bay Rays                **
## factor(Tm)Texas Rangers
## factor(Tm)Toronto Blue Jays             .
## factor(Tm)Washington Nationals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06996 on 258 degrees of freedom
## Multiple R-squared:  0.3658, Adjusted R-squared:  0.265
## F-statistic:  3.63 on 41 and 258 DF,  p-value: 1.451e-10
```

```
plot(IBB, 1)
```



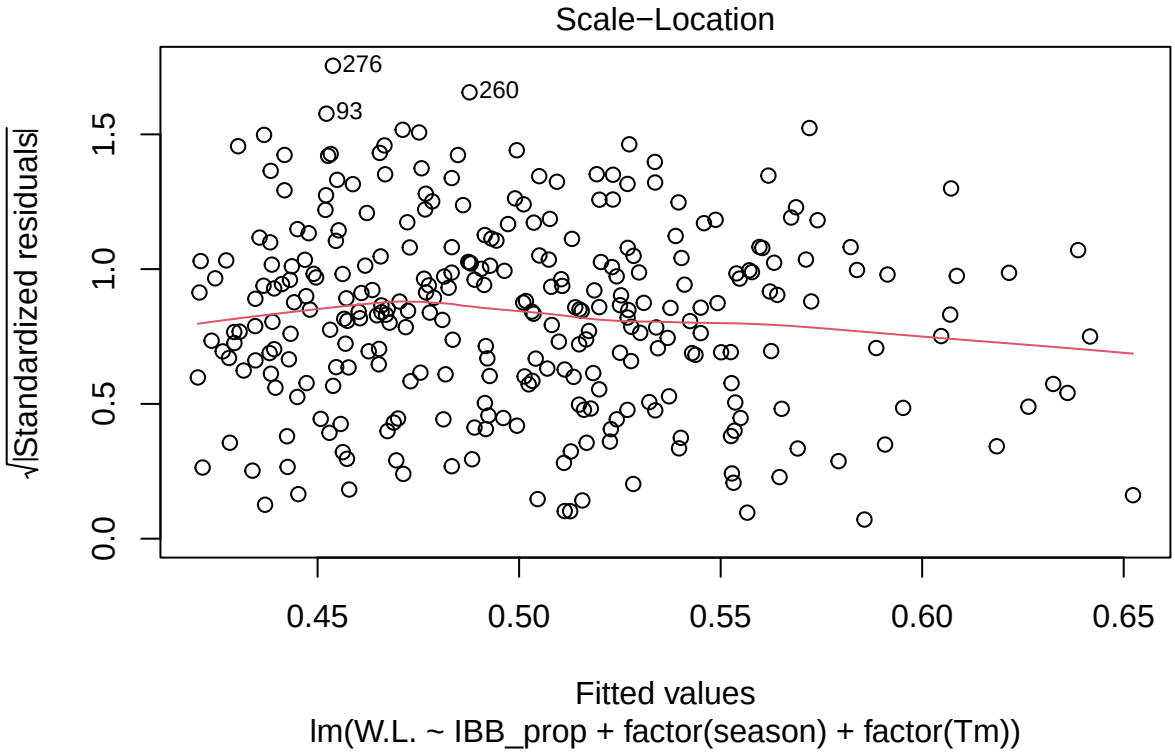

```
plot(IBB, 2)
```

```
## Warning: not plotting observations with leverage one:
##   13
```



```r
plot(IBB, 3)
```

```
## Warning: not plotting observations with leverage one:
##   13
```




```r
shapiro.test(residuals(IBB))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(IBB)
## W = 0.99546, p-value = 0.5311
```

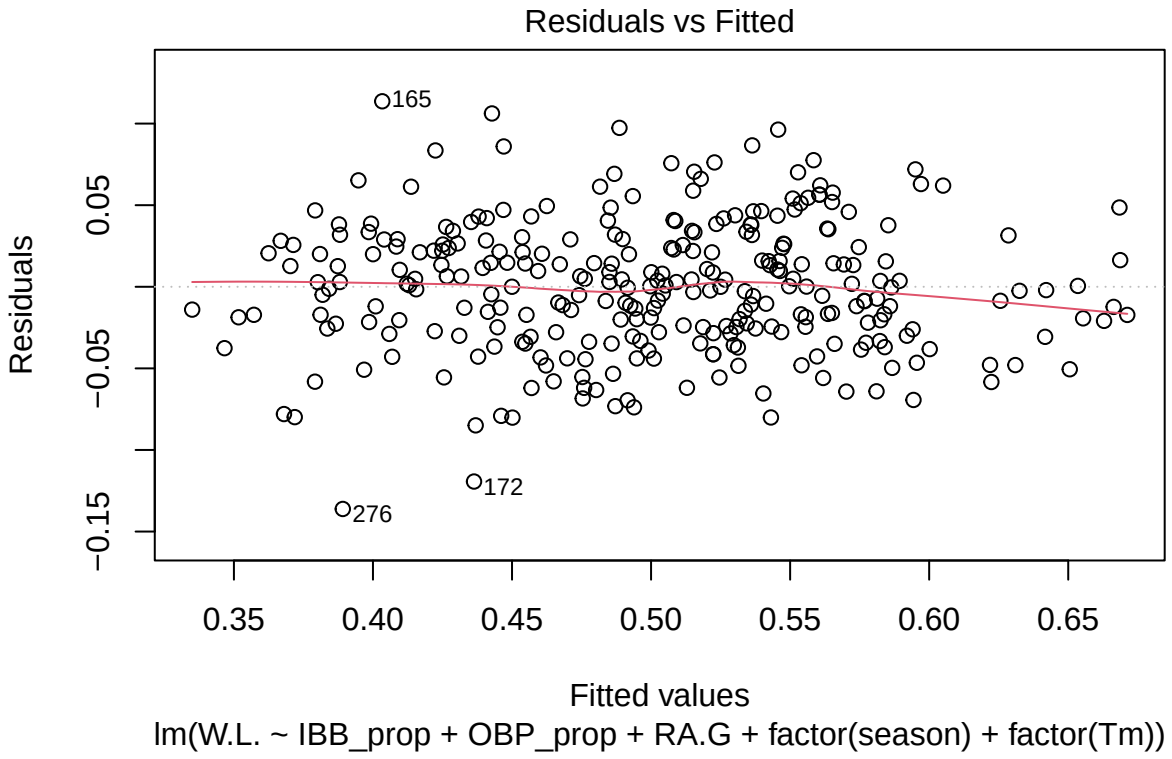
```r
# model 2 w Controls
IBB_control <- lm(W.L. ~ IBB_prop + OBP_prop + RA.G + factor(season) + factor(Tm), data = mlb)
summary(IBB_control)
```

```
##
## Call:
## lm(formula = W.L. ~ IBB_prop + OBP_prop + RA.G + factor(season) +
##     factor(Tm), data = mlb)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.136149 -0.027357  0.000209  0.026420  0.113659
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.123856   0.097727  11.500  < 2e-16
## IBB_prop                             0.030832   0.018751   1.644  0.10134
## OBP_prop                            -0.004953   0.004539  -1.091  0.27614
```

```
## RA.G                                    -0.113895   0.013691  -8.319 5.33e-15
## factor(season)2016                       0.029611   0.011393   2.599  0.00989
## factor(season)2017                       0.049489   0.011648   4.249 3.01e-05
## factor(season)2018                       0.024555   0.011502   2.135  0.03373
## factor(season)2019                       0.073650   0.012645   5.825 1.71e-08
## factor(season)2020                       0.055498   0.012345   4.495 1.05e-05
## factor(season)2021                       0.037301   0.012020   3.103  0.00213
## factor(season)2022                       0.011267   0.012815   0.879  0.38014
## factor(season)2023                       0.053496   0.012839   4.167 4.23e-05
## factor(season)2024                       0.023463   0.012852   1.826  0.06907
## factor(Tm)Atlanta Braves                 0.018266   0.019688   0.928  0.35439
## factor(Tm)Baltimore Orioles              0.009388   0.019877   0.472  0.63713
## factor(Tm)Boston Red Sox                 0.043254   0.019394   2.230  0.02660
## factor(Tm)Chicago Cubs                  -0.004798   0.019664  -0.244  0.80742
## factor(Tm)Chicago White Sox             -0.024068   0.020029  -1.202  0.23059
## factor(Tm)Cincinnati Reds               -0.014997   0.019456  -0.771  0.44154
## factor(Tm)Cleveland Guardians           -0.028859   0.029433  -0.980  0.32777
## factor(Tm)Cleveland Indians             -0.015652   0.022301  -0.702  0.48341
## factor(Tm)Colorado Rockies               0.051253   0.019834   2.584  0.01032
## factor(Tm)Detroit Tigers                -0.017079   0.019801  -0.863  0.38919
## factor(Tm)Houston Astros                 0.011741   0.020128   0.583  0.56017
## factor(Tm)Kansas City Royals            -0.017273   0.020086  -0.860  0.39063
## factor(Tm)Los Angeles Angels            -0.017323   0.019957  -0.868  0.38617
## factor(Tm)Los Angeles Angels of Anaheim  0.011776   0.046154   0.255  0.79881
## factor(Tm)Los Angeles Dodgers            0.017558   0.020797   0.844  0.39933
## factor(Tm)Miami Marlins                 -0.038223   0.019546  -1.956  0.05160
## factor(Tm)Milwaukee Brewers             -0.007569   0.019583  -0.387  0.69943
## factor(Tm)Minnesota Twins                0.013707   0.019616   0.699  0.48532
## factor(Tm)New York Mets                 -0.014256   0.019524  -0.730  0.46594
## factor(Tm)New York Yankees               0.025892   0.019983   1.296  0.19625
## factor(Tm)Oakland Athletics             -0.012653   0.019892  -0.636  0.52529
## factor(Tm)Philadelphia Phillies          0.003217   0.019403   0.166  0.86844
## factor(Tm)Pittsburgh Pirates            -0.022074   0.019521  -1.131  0.25919
## factor(Tm)San Diego Padres              -0.021265   0.019532  -1.089  0.27729
## factor(Tm)San Francisco Giants          -0.027372   0.019531  -1.401  0.16229
## factor(Tm)Seattle Mariners               0.002763   0.019987   0.138  0.89015
## factor(Tm)St. Louis Cardinals           -0.002114   0.020030  -0.106  0.91602
## factor(Tm)Tampa Bay Rays                -0.012196   0.020251  -0.602  0.54754
## factor(Tm)Texas Rangers                  0.017186   0.019677   0.873  0.38325
## factor(Tm)Toronto Blue Jays              0.021749   0.020043   1.085  0.27890
## factor(Tm)Washington Nationals          -0.015466   0.019541  -0.791  0.42943
##
## (Intercept)                             ***
## IBB_prop
## OBP_prop
## RA.G                                    ***
## factor(season)2016                      **
## factor(season)2017                      ***
## factor(season)2018                      *
## factor(season)2019                      ***
## factor(season)2020                      ***
## factor(season)2021                      **
## factor(season)2022
## factor(season)2023                      ***
```

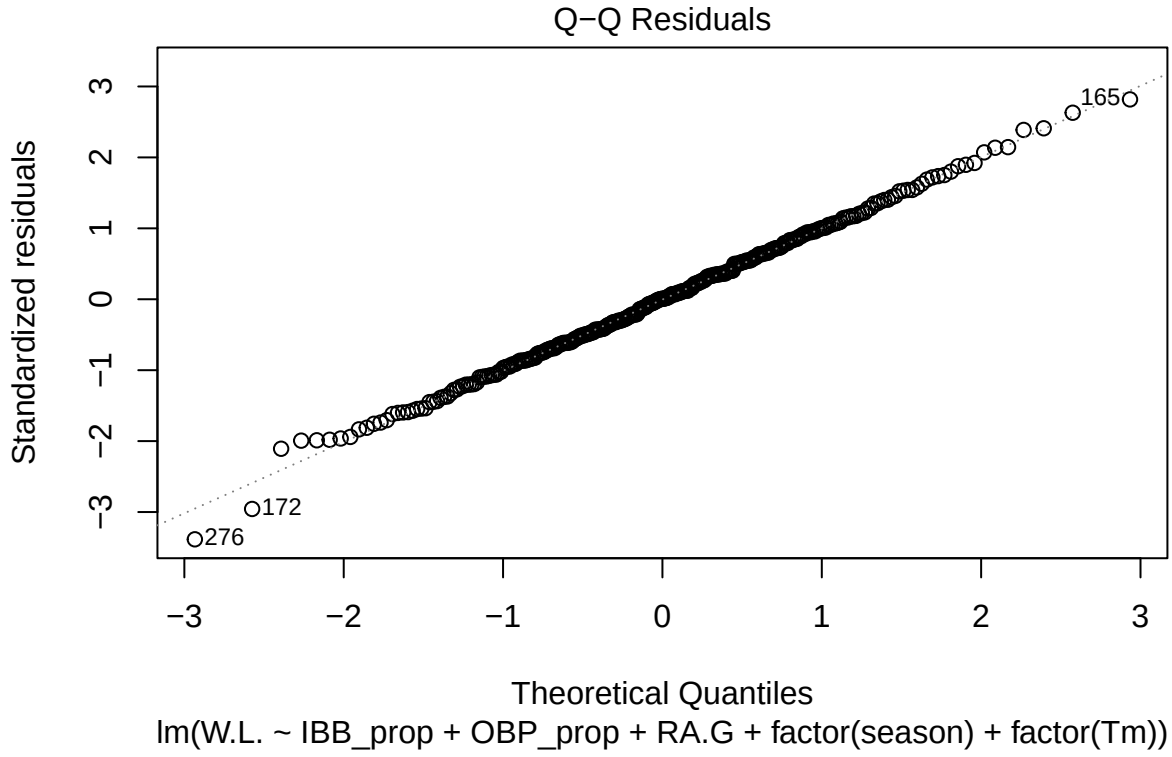
```
## factor(season)2024                      .
## factor(Tm)Atlanta Braves
## factor(Tm)Baltimore Orioles
## factor(Tm)Boston Red Sox                *
## factor(Tm)Chicago Cubs
## factor(Tm)Chicago White Sox
## factor(Tm)Cincinnati Reds
## factor(Tm)Cleveland Guardians
## factor(Tm)Cleveland Indians
## factor(Tm)Colorado Rockies              *
## factor(Tm)Detroit Tigers
## factor(Tm)Houston Astros
## factor(Tm)Kansas City Royals
## factor(Tm)Los Angeles Angels
## factor(Tm)Los Angeles Angels of Anaheim
## factor(Tm)Los Angeles Dodgers
## factor(Tm)Miami Marlins                 .
## factor(Tm)Milwaukee Brewers
## factor(Tm)Minnesota Twins
## factor(Tm)New York Mets
## factor(Tm)New York Yankees
## factor(Tm)Oakland Athletics
## factor(Tm)Philadelphia Phillies
## factor(Tm)Pittsburgh Pirates
## factor(Tm)San Diego Padres
## factor(Tm)San Francisco Giants
## factor(Tm)Seattle Mariners
## factor(Tm)St. Louis Cardinals
## factor(Tm)Tampa Bay Rays
## factor(Tm)Texas Rangers
## factor(Tm)Toronto Blue Jays
## factor(Tm)Washington Nationals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04335 on 256 degrees of freedom
## Multiple R-squared:  0.7584, Adjusted R-squared:  0.7178
## F-statistic: 18.69 on 43 and 256 DF,  p-value: < 2.2e-16
```

```
plot(IBB_control, 1)
```



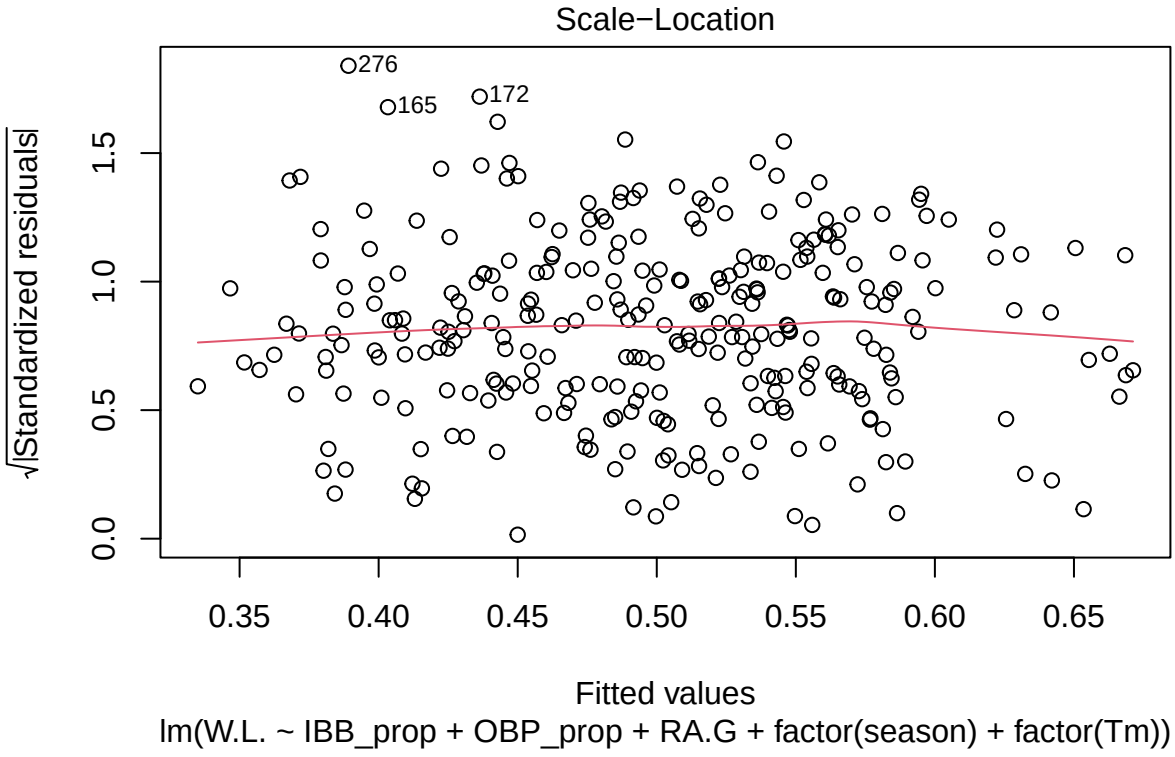

```
plot(IBB_control, 2)
```

```
## Warning: not plotting observations with leverage one:
##   13
```



```r
plot(IBB_control, 3)
```

```
## Warning: not plotting observations with leverage one:
##   13
```




```r
shapiro.test(residuals(IBB_control))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(IBB_control)
## W = 0.99785, p-value = 0.9664
```

```r
# Model w intersection
IBB_int <- lm(W.L. ~ IBB_prop + OBP_prop + RA.G + IBB_prop * OBP_prop +
                factor(season) + factor(Tm), data = mlb)
summary(IBB_int)
```

```
##
## Call:
## lm(formula = W.L. ~ IBB_prop + OBP_prop + RA.G + IBB_prop * OBP_prop +
##     factor(season) + factor(Tm), data = mlb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13135 -0.02762  0.00115  0.02647  0.11789
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.248147   0.141669   8.810  < 2e-16
## IBB_prop                            -0.257766   0.239083  -1.078  0.28199
```

```
## OBP_prop                                -0.008855   0.005563  -1.592  0.11267
## RA.G                                    -0.113547   0.013682  -8.299 6.16e-15
## factor(season)2016                       0.029077   0.011391   2.553  0.01127
## factor(season)2017                       0.049138   0.011641   4.221 3.38e-05
## factor(season)2018                       0.024989   0.011497   2.173  0.03067
## factor(season)2019                       0.073465   0.012634   5.815 1.81e-08
## factor(season)2020                       0.053877   0.012407   4.343 2.03e-05
## factor(season)2021                       0.037374   0.012009   3.112  0.00207
## factor(season)2022                       0.010624   0.012815   0.829  0.40785
## factor(season)2023                       0.052922   0.012836   4.123 5.06e-05
## factor(season)2024                       0.022013   0.012896   1.707  0.08904
## factor(Tm)Atlanta Braves                 0.016042   0.019755   0.812  0.41754
## factor(Tm)Baltimore Orioles              0.010275   0.019872   0.517  0.60556
## factor(Tm)Boston Red Sox                 0.044424   0.019400   2.290  0.02285
## factor(Tm)Chicago Cubs                  -0.002096   0.019772  -0.106  0.91565
## factor(Tm)Chicago White Sox             -0.024048   0.020010  -1.202  0.23057
## factor(Tm)Cincinnati Reds               -0.016318   0.019469  -0.838  0.40271
## factor(Tm)Cleveland Guardians           -0.025070   0.029572  -0.848  0.39737
## factor(Tm)Cleveland Indians             -0.015363   0.022282  -0.689  0.49115
## factor(Tm)Colorado Rockies               0.050340   0.019830   2.539  0.01173
## factor(Tm)Detroit Tigers                -0.017057   0.019783  -0.862  0.38937
## factor(Tm)Houston Astros                 0.010958   0.020120   0.545  0.58649
## factor(Tm)Kansas City Royals            -0.015047   0.020152  -0.747  0.45595
## factor(Tm)Los Angeles Angels            -0.017188   0.019939  -0.862  0.38947
## factor(Tm)Los Angeles Angels of Anaheim  0.012939   0.046122   0.281  0.77930
## factor(Tm)Los Angeles Dodgers            0.022722   0.021212   1.071  0.28509
## factor(Tm)Miami Marlins                 -0.039052   0.019540  -1.999  0.04672
## factor(Tm)Milwaukee Brewers             -0.008009   0.019569  -0.409  0.68266
## factor(Tm)Minnesota Twins                0.012753   0.019614   0.650  0.51615
## factor(Tm)New York Mets                 -0.012912   0.019537  -0.661  0.50928
## factor(Tm)New York Yankees               0.025787   0.019965   1.292  0.19766
## factor(Tm)Oakland Athletics             -0.012589   0.019874  -0.633  0.52702
## factor(Tm)Philadelphia Phillies          0.001872   0.019417   0.096  0.92326
## factor(Tm)Pittsburgh Pirates            -0.022252   0.019504  -1.141  0.25497
## factor(Tm)San Diego Padres              -0.021484   0.019515  -1.101  0.27198
## factor(Tm)San Francisco Giants          -0.025816   0.019556  -1.320  0.18798
## factor(Tm)Seattle Mariners               0.001530   0.019995   0.077  0.93906
## factor(Tm)St. Louis Cardinals           -0.002189   0.020012  -0.109  0.91299
## factor(Tm)Tampa Bay Rays                -0.013778   0.020275  -0.680  0.49739
## factor(Tm)Texas Rangers                  0.017743   0.019664   0.902  0.36776
## factor(Tm)Toronto Blue Jays              0.021461   0.020026   1.072  0.28491
## factor(Tm)Washington Nationals          -0.012786   0.019649  -0.651  0.51582
## IBB_prop:OBP_prop                        0.008977   0.007414   1.211  0.22708
##
## (Intercept)                             ***
## IBB_prop
## OBP_prop
## RA.G                                    ***
## factor(season)2016                      *
## factor(season)2017                      ***
## factor(season)2018                      *
## factor(season)2019                      ***
## factor(season)2020                      ***
## factor(season)2021                      **
```

```
## factor(season)2022
## factor(season)2023                      ***
## factor(season)2024                      .
## factor(Tm)Atlanta Braves
## factor(Tm)Baltimore Orioles
## factor(Tm)Boston Red Sox                *
## factor(Tm)Chicago Cubs
## factor(Tm)Chicago White Sox
## factor(Tm)Cincinnati Reds
## factor(Tm)Cleveland Guardians
## factor(Tm)Cleveland Indians
## factor(Tm)Colorado Rockies              *
## factor(Tm)Detroit Tigers
## factor(Tm)Houston Astros
## factor(Tm)Kansas City Royals
## factor(Tm)Los Angeles Angels
## factor(Tm)Los Angeles Angels of Anaheim
## factor(Tm)Los Angeles Dodgers
## factor(Tm)Miami Marlins                 *
## factor(Tm)Milwaukee Brewers
## factor(Tm)Minnesota Twins
## factor(Tm)New York Mets
## factor(Tm)New York Yankees
## factor(Tm)Oakland Athletics
## factor(Tm)Philadelphia Phillies
## factor(Tm)Pittsburgh Pirates
## factor(Tm)San Diego Padres
## factor(Tm)San Francisco Giants
## factor(Tm)Seattle Mariners
## factor(Tm)St. Louis Cardinals
## factor(Tm)Tampa Bay Rays
## factor(Tm)Texas Rangers
## factor(Tm)Toronto Blue Jays
## factor(Tm)Washington Nationals
## IBB_prop:OBP_prop
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04331 on 255 degrees of freedom
## Multiple R-squared:  0.7598, Adjusted R-squared:  0.7183
## F-statistic: 18.33 on 44 and 255 DF,  p-value: < 2.2e-16
```

Interpretations

Model 1 Results

$$\hat{\beta}_0 = 0.404, \quad \hat{\beta}_1 = 0.112$$

The coefficient for IBB ($\hat{\beta}_1$) is 0.112, with a p-value of 0.00016, indicating statistical significance at $\alpha = 0.05$. This suggests that, on average, an increase in IBB/AB by one percentage point corresponds to a 0.112 increase in Win-Loss percentage. The null hypothesis is rejected, demonstrating that intentional walks positively impact Win-Loss percentage. It implies that intentional walks are indicative of strong hitters contributing to offensive opportunities rather than disrupting offensive flow.

Model 2 Results

$$\hat{\beta}_0 = 1.124, \quad \hat{\beta}_1 = 0.031, \quad \hat{\beta}_2 = -0.005, \quad \hat{\beta}_3 = -0.114$$

After including OBP and RA/G as controls, the coefficient for IBB ($\hat{\beta}_1$) becomes statistically insignificant ($p = 0.10$). Similarly, the coefficient for OBP ($\hat{\beta}_2$) is also statistically insignificant ($p = 0.28$) and unexpectedly negative. While this contradicts the intuitive expectation that stronger offensive performance (measured by OBP) should positively correlate with higher Win-Loss proportions, it is likely influenced by multicollinearity, as players with higher OBP are often intentionally walked more frequently. This overlap may obscure OBP's unique contribution to Win-Loss proportions. The coefficient for RA/G ($\hat{\beta}_3$) is negative and significant ($p = 5.33e-15$), in line with our understanding that teams better at preventing runs generally achieve higher Win-Loss proportions. This result suggests that IBB's impact is mediated by defensive factors, and reflects a team's overall strength rather than directly driving performance. It also suggests that a team's defensive strength might be a stronger predictor of Win-Loss proportion than offensive metrics such as IBB or OBP, emphasizing the importance of run prevention in team success.

Contextual Analysis

The effects of IBB are context-dependent and influenced by game states, lineups, and high-leverage situations. For example:

- **Scenario 1**: An intentional walk with two outs and empty bases in the 1st inning.
- **Scenario 2**: An intentional walk with no outs, bases loaded in the 9th inning.

These differing contexts highlight the situational role of IBB, which may not consistently affect season-long performance metrics. Furthermore, IBB often reflects an opponent's strategy and defensive quality rather than the team's intrinsic performance.

This nuanced understanding underscores the importance of contextual analysis in interpreting the broader implications of intentional walks.

Discussion on Playoff Format Change

The MLB expanded its postseason from 10 to 12 teams in 2022. This rule change is not only exciting for fans, as their teams should be in the hunt for a playoff spot each season, but the change should also be welcomed by statisticians. The exogenous rule change provides a setting for a quasi-random study where the performance of the two additional teams that now make the playoffs can be studied. This will allow inference on whether there are systematic differences between what it takes to win to get into the playoffs and the World Series. Over the past three seasons, fans have had the luxury of watching more frequent upsets. While there are different theories on why lower-seeded teams outperform their opponents in the playoffs, we will set up a statistical framework for testing the unexpected high-quality play of the 11th and 12th seeds.

Comparison of Playoff Formats Before we dive into the statistical framework, we will clarify how the playoffs work now versus how they did in the Wild Card era, which was the previous format. From 2012–2019 and in 2021, the playoffs consisted of ten teams. There were six division winners split across the American and National Leagues. Then, each league had two Wild Card teams that did not win their division but held the best records in their league out of non-division-winning teams. Each league's two Wild Card teams played each other in a win-or-go-home match. The winner would go on to play in the Division Series against the division winner with the best record, and the other two division winners would play each other. The two teams to win the Division Series would go on to play in the League Championship Series. The winner of each league's championship would meet in the World Series. The Division Series was a best-of-five, with the

exception of 2012, which was a best-of-three. The League Championship Series and World Series were both best-of-sevens.

The most recent playoffs of the 10-team format are depicted below:

Figure 1: The most recent playoffs of the 10-team format via MLB.com

Now, the stakes are different. An additional team makes the playoffs in each league, where the division winner with the lowest record now participates against the lowest Wild Card team in a best-of-three, and the top two Wild Card teams play each other in a three-game series as well. The top two division winners maintain their guaranteed spot in the Division Series. The other rounds remain the same. For a visual of the new format, refer to the image below:

Analysis of Changes and Implications In the previous format, which consists of nine seasons, three teams made the World Series, and two ended up winning the World Series as well. In 2014, the Kansas City Royals and San Francisco Giants were both Wild Card teams that met in the World Series, with the Giants winning. In 2019, the Washington Nationals won the World Series as a Wild Card. In the new format, we have already seen Wild Card teams make or even win the World Series in just three seasons.

- In 2022, the Phillies were not only a Wild Card team but also the lowest-ranked team in the National League. Without this change, they would not have made the playoffs, let alone the World Series.

- In 2023, two Wild Card teams clashed in the World Series, one of which was the lowest seed in the National League—the Arizona Diamondbacks—who were eliminated by the 5th seed in the American League, the Texas Rangers.

In these three seasons, three out of the six teams to make a World Series were Wild Card teams, two of which were the 6th seeds that would not have made it in the prior format.

Figure 2: The most recent playoffs of the 12-team format via MLB.com

With the miraculous runs witnessed in two of the past three seasons, it is reasonable to expect more, leading to the question: Why are teams that performed better all season losing to teams that would not have even made the playoffs? For the 6th seed to advance to the World Series, they must beat the division winners with the best and worst records, with the former being heavy favorites usually.

This happened two times in the first three seasons under this format and creates a setting to study how to build a team that can make the playoffs, rise above expectations, and consistently beat the best of the best.

Framework for Analysis The expanded playoff format provides an opportunity to study whether lower-seeded teams consistently outperform expectations. The following linear model can quantify this:

$$Y_i = \beta_0 + \beta_1 Post_i + \beta_2 SixSeed_i + \beta_3 (Post_i \times SixSeed_i) + \epsilon_i$$

Where: - Y_i is the outcome of interest (e.g., number of playoff games won, World Series appearances). - $Post_i$ is an indicator variable for the new format (1 for 2022 onward, 0 for prior years). - $SixSeed_i$ is an indicator variable for the lowest-seeded playoff teams (1 for 6th seed, 0 otherwise). - $Post_i \times SixSeed_i$ captures the interaction between format and seeding.

Conditional Expectations

- Teams with 1st–5th seeds in the prior format: $E(Y_i|Post_i = 0, SixSeed_i = 0) = \beta_0$
- Teams with 1st–5th seeds in the current format: $E(Y_i|Post_i = 1, SixSeed_i = 0) = \beta_0 + \beta_1$
- Teams with 6th seeds in the prior format (hypothetical): $E(Y_i|Post_i = 0, SixSeed_i = 1) = \beta_0 + \beta_2$
- Teams with 6th seeds in the current format: $E(Y_i|Post_i = 1, SixSeed_i = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$

This model allows us to test whether the expanded format disproportionately benefits lower-seeded teams.

Potential Analysis While the current sample size (2022–2024) is small, the model could expand in future seasons to analyze: - Series wins by seed under each format. - Upset likelihood (lower seed defeating higher seed). - Systematic differences in team characteristics (e.g., budget, player metrics).

Prediction for 2025

This section focuses on predicting which teams will make the playoffs in 2025 using a combination of common baseball statistics and regularization techniques. The primary goal is to test whether select predictors can accurately forecast playoff outcomes through statistical modeling and machine learning.

Baseline Model We began with a probit model to predict playoff appearances based on a set of predefined predictors. This baseline model leverages substantive knowledge to identify metrics strongly associated with team success. The accuracy of this model will be evaluated using cross-validation techniques, such as k-fold cross-validation, and compared to machine learning methods like LASSO and ridge regression.

Model Assumptions

- **Team Playoff Indicator:**
  Let $Playoff_i$ represent a binary indicator for the i-th team, taking the value 1 if the team makes the playoffs and 0 otherwise. Playoff qualification is used as a proxy for team success, aligning with MLB teams' overarching goal of winning the World Series.

- **Constraints:**
  The model is limited to 30 observations per year over ten years, necessitating a small model to avoid overfitting. Additionally, the study period includes a playoff expansion, which complicates the constraints on qualifying teams.

Initial Model Specification

The baseline model includes a mixture of predictors representing pitcher and hitter success, drawn from standard and advanced metrics:

$$P(Playoff_i = 1) = \beta_0 + \beta_1 SB + \beta_2 PAge + \beta_3 BatAge + \beta_4 EV + \beta_5 ERA + \beta_6 BB + \beta_7 OPS + \beta_8 HR + \beta_9 RS + \beta_{10} E + \beta_{11} RDRS$$

Hypothesis of Interest

The hypothesis posits that a team's playoff likelihood can be accurately predicted using a select group of metrics. These metrics may include unconventional statistics that are overlooked by analysts and fans. The analysis focuses on both:

- **Predictive Accuracy:**
  Measured by cross-validation techniques and root mean square prediction error (RMSPE).

- **Variable Significance:**
  Tested using formal hypothesis testing for individual coefficients.

Implementation

Baseline Model

The baseline model is fitted using predictors derived from substantive knowledge. Preliminary analysis of playoff versus non-playoff teams revealed systemic differences in their statistical profiles.

```
normalized_mlb <- mlb
predictors <- c("SB", "PAge", "BatAge", "EV", "ERA", "BB", "OPS",
                "HR", "RS.", "E", "Rdrs")
normalized_mlb[predictors] <- scale(mlb[predictors])

# fit model
summary(logit <- glm(playoffs ~ SB + PAge + BatAge + EV + ERA + BB +
                       OPS + HR + RS. + E + Rdrs,
                     data = normalized_mlb, family = "binomial"))$coefficients
```

```
##                 Estimate Std. Error     z value     Pr(>|z|)
## (Intercept) -1.783337342  0.2984645 -5.97504043 2.300331e-09
## SB          -0.566566228  0.2591336 -2.18638699 2.878730e-02
## PAge        -0.003846076  0.2180744 -0.01763653 9.859288e-01
## BatAge       0.335134678  0.2141375  1.56504447 1.175725e-01
## EV          -0.526405568  0.2502106 -2.10385031 3.539151e-02
## ERA         -3.964467885  0.9116666 -4.34859412 1.370130e-05
## BB           0.462845465  0.4317155  1.07210768 2.836717e-01
## OPS          0.896345915  0.7085298  1.26507872 2.058431e-01
## HR          -0.183074165  0.3988022 -0.45906011 6.461910e-01
## RS.          1.759350542  0.2829622  6.21761703 5.047617e-10
## E           -1.114913167  0.3827114 -2.91319581 3.577502e-03
## Rdrs         0.268624221  0.2710993  0.99087010 3.217490e-01
```

Model Comparison and Validation To compare the baseline model with regularization techniques like ridge and LASSO regression, we computed the RMSPE. The analysis also considered principal component analysis (PCA) to maximize the explained variance with fewer predictors.

Baseline Model Validation with k-Fold Cross-Validation:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
k <- 10
set.seed(139)
```

```r
mlb_logit <- normalized_mlb %>%
  select("SB", "PAge", "BatAge", "EV", "ERA", "BB", "OPS", "HR",
         "RS.", "E", "Rdrs", "playoffs")
mlb_logit$playoffs <- as.numeric(as.character(mlb_logit$playoffs))
mlb_logit$fold <- sample(rep(1:k, length.out = nrow(mlb_logit)))

mspe_vals_logit <- numeric(k)

#  k-fold cv
for (i in 1:k) {
  train_data_logit <- mlb_logit %>% filter(fold != i)
  test_data_logit <- mlb_logit %>% filter(fold == i)

  # fit
  logit_base <- glm(playoffs ~ SB + PAge + BatAge + EV + ERA + BB +
                      OPS + HR + RS. + E + Rdrs,
                    data = train_data_logit, family = "binomial")

  # predict on test set
  test_data_logit$predicted_prob <- predict(logit_base, newdata = test_data_logit, type = "response")


  test_data_logit$mspe <- (test_data_logit$predicted_prob - test_data_logit$playoffs)^2
  mspe_vals_logit[i] <- mean(test_data_logit$mspe)
}

logit_mspe <- sqrt(mean(mspe_vals_logit))
cat("Logit Regression MSPE:", logit_mspe, "\n")
```

```
## Logit Regression MSPE: 0.3044582
```

Results:
The baseline model exhibits an MSPE of approximately 30%, indicating moderate predictive accuracy. However, improvements are expected using machine learning techniques. Thus, it provides an initial understanding of playoff predictors. By incorporating advanced techniques like ridge regression, LASSO, and PCA, we aim to enhance predictive accuracy. Regularization methods address overfitting risks while highlighting key predictors for success. Future steps include incorporating additional variables (e.g., team budgets and time-fixed effects) to refine the model.

Regularization Techniques To prevent overfitting, ridge and lasso regression techniques are used. These methods identify the most important predictors while minimizing prediction error.

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```r
# ridge
ridge <- glmnet(as.matrix(normalized_mlb[predictors]), normalized_mlb$playoffs,
                alpha = 0, family = "binomial")
cv_ridge <- cv.glmnet(as.matrix(normalized_mlb[predictors]),
                      normalized_mlb$playoffs, alpha = 0, family = "binomial")
ridge_lambda <- cv_ridge$lambda.min
ridge_coeff <- predict(ridge, s = ridge_lambda, type = "coefficients")
```

```r
# LASSO
lasso <- glmnet(as.matrix(normalized_mlb[predictors]),
                normalized_mlb$playoffs, alpha = 1, family = "binomial")
cv_lasso <- cv.glmnet(as.matrix(normalized_mlb[predictors]),
                      normalized_mlb$playoffs, alpha = 1, family = "binomial")
lasso_lambda <- cv_lasso$lambda.min
lasso_coeff <- predict(lasso, s = lasso_lambda, type = "coefficients")
```

Model Comparison Cross-validation is used to compare the performance of the baseline, ridge, and lasso models. Metrics such as Mean Squared Prediction Error (MSPE) and Root Mean Squared Prediction Error (RMSPE) are calculated for each model.

Principal Component Analysis (PCA)

PCA is used to reduce dimensionality by identifying combinations of predictors that explain the most variance.

```r
pca <- prcomp(normalized_mlb[predictors], center = TRUE, scale. = TRUE)
explained_variance <- cumsum(pca$sdev^2 / sum(pca$sdev^2))
num_components <- which(explained_variance >= 0.95)[1]

cat("Number of components explaining 95% variance:", num_components, "\n")
```

```
## Number of components explaining 95% variance: 8
```

The reason PCA is not working well here is that the current data we have does not explain enough of the variance with just a select few predictors such that it is hard to remove anyone. The predective power is low to begin with, so we need to add other predictors of interest like budget and time fixed effects.

Conclusion

The insights and models developed throughout this project lay the groundwork for understanding how team metrics influence playoff success and for predicting future outcomes in the MLB. While our initial models capture key predictors and trends, there are several opportunities to enhance and expand this analysis.

Summary of Key Findings

1. **EDA Insights:**
   - Significant differences exist between playoff and non-playoff teams across metrics such as On-Base Percentage (OPS), Home Run Percentage (HR%), and Defensive Runs Saved (RDRS).
   - Metrics reflecting offensive and defensive balance, such as Run Support Percentage (RS%) and Earned Run Average (ERA), are critical for distinguishing successful teams.
2. **Hypothesis Testing:**
   - Intentional Bases on Balls (IBBs) may be an indirect indicator of team offensive strength rather than a direct contributor to win-loss percentage.
   - After controlling for On-Base Percentage (OBP) and Runs Allowed per Game (RA/G), IBB loses significance, suggesting it reflects contextual team dynamics rather than being a driver of success.
3. **Playoff Format Changes:**
   - The expanded playoff format provides opportunities for lower-seeded teams, such as the 6th seed, to outperform expectations.

- A linear model framework shows potential for studying systematic differences in playoff outcomes under varying formats, though additional seasons are needed for robust analysis.

4. **Predictive Models:**
   - A baseline logit model, ridge regression, and lasso regression show the importance of advanced metrics alongside traditional baseball statistics.
   - Regularization techniques slightly improve prediction accuracy but highlight the need for additional context-specific predictors (e.g., budget, player injuries).

Broader Implications

The findings from this project are not limited to baseball. They demonstrate the power of statistical modeling and machine learning in sports analytics, with potential applications across industries: - **Sports Management:** Informing decisions on roster construction, trades, and free-agent acquisitions. - **Fan Engagement:** Providing deeper insights into team performance and playoff predictions. - **General Analytics:** Highlighting the importance of balancing substantive knowledge with data-driven techniques in decision-making.

By continuing to refine these models and expanding the scope of analysis, this project has the potential to provide actionable insights for teams, analysts, and fans alike.

Acknowledgments

This project has been an enriching experience, combining knowledge from **Stat 139: Introduction to Linear Models** with real-world sports data. Special thanks to **Professor Xenakis** for guidance and feedback throughout the course, and to peers for their collaborative insights.

Appendices

Appendix A: Data Sources

1. **Baseball Reference:**
   - Comprehensive database for all player, team, and league statistics.
   - Used to source metrics such as Standard Pitching, Advanced Pitching, Standard Batting, Advanced Batting, and Fielding.
   - URL: https://www.baseball-reference.com
2. **MLB.com:**
   - Official MLB data and playoff formats.
   - Provided historical insights into playoff structures and team standings.
   - URL: https://www.mlb.com

Appendix B: Statistical Model Assumptions

1. **Logistic Regression:**
   - Dependent variable: Binary indicator of playoff qualification.
   - Assumes a linear relationship between log-odds of playoff qualification and predictors.
   - Key assumptions tested:
     - **Linearity in the logit:** Verified through residual analysis.
     - **Independence of observations:** Satisfied by using team-level data per season.
2. **Linear Models (Hypothesis Testing):**
   - Fixed-effects models control for unobserved heterogeneity across teams and seasons.
   - Assumptions of normality and homoscedasticity verified using diagnostic plots.
3. **Ridge/Lasso Regression:**
   - Regularization parameters optimized using cross-validation.
   - Used to handle multicollinearity and select relevant predictors.
4. **PCA:**
   - Number of components selected to explain at least 95% of variance.
   - Assumes relationships among predictors are linear.

Appendix C: Data Cleaning and Preprocessing

1. **Handling Missing Data:**
   - Missing values imputed using league averages for specific metrics.
   - Columns with excessive missingness excluded from analysis.
2. **Normalization:**
   - Continuous variables scaled using z-scores to ensure comparability.
   - Percent-based metrics converted from percentage strings to numeric proportions.
3. **Categorical Variables:**
   - Factors such as `season` and `Tm` (team) converted to dummy variables for regression models.

Appendix D: Challenges Faced

1. **Data Quality:**
   - Inconsistent formatting across CSV files required manual adjustments.
   - Varying column names for similar metrics (e.g., `HR` for batting and pitching).
2. **Playoff Format Changes:**
   - Modeling playoff outcomes required aligning data from different formats.
   - The small sample size for the expanded playoff format limited robust statistical analysis.
3. **Model Overfitting:**
   - Balancing model complexity with limited observations (300 total).
   - Regularization techniques helped mitigate overfitting concerns.

Appendix E: Limitations

1. **Sample Size:**
   - Only 10 seasons of data limits the generalizability of findings.
   - Expanded playoff format data is particularly sparse.
2. **Contextual Variables:**
   - Factors like player injuries, managerial decisions, and mid-season trades not included in the models.
3. **Predictor Selection:**
   - Metrics used are limited to publicly available data, potentially missing proprietary or advanced measures.
4. **Outcome Variables:**
   - Focused on playoff qualification and win-loss percentage; future studies could explore game-by-game outcomes.

Appendix F: Code Repository The full project code and cleaned datasets are available on GitHub for transparency and reproducibility:
Repository: GitHub - Stat139_FinalProject

Thank you :)