# Assessing MLB Batters in 2023

Joseph Cohen

2024-02-06

This document provides a walk-through of a simple, cursory data analysis in an applied context. The walk-through is designed to convey the general tasks involved in creating simple analytics-based informational products in an R-based workflow. Students are invited to follow along with what I do, and manipulate my code to get their own results.

## Pre-Analysis

### Scenario

We have been approached by the New York Mets to assess which are the best batters in Major League Baseball. They feel like they can win the league championship with one or two more elite batters, but they do not know whom to pursue in a trade or free agent signing

### Project Conception

Although the task seems straightforward, the problem is that we do not have clear, uncontroversial answers about which player would be the best for the New York Mets. Hitters are good at different things, and we do not know what kind of hitter would be best for the Mets specifically. I think that our best strategy is to inform the Mets baseball professionals about who did a good job in different aspects of batting, and decide who is best for that particular team in conversation with others.

### Research Design

I propose that, as a team, each of us selects an indicator from 2023 seasonal batting data to get ideas on who did well at particular facets of hitting. We will look at leaderboards to determine who did best and worst among qualified hitters.

# Data Wrangling

## Data Acquisition

The data is stored on the Excel sheet provided in class. To import data from the first worksheet of this Excel workbook:

```
# Import data from Excel
data <- read_xlsx("MLB 2023 Batting Statistics.xlsx", sheet = 1)

# Look at first few rows and columns
head(data, 5)
```

```
## # A tibble: 5 x 18
##   Name  Team      G    PA   AVG     R    HR   RBI   OBP   SLG    SB K_pct BB_pct
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 Rona~ ATL     159   735 0.337   149    41   106 0.416 0.596    73 0.114 0.109
## 2 Mook~ LAD     152   693 0.307   126    39   107 0.408 0.579    14 0.154 0.139
## 3 Fred~ LAD     161   730 0.331   131    29   102 0.410 0.567    23 0.166 0.0986
## 4 Matt~ ATL     162   720 0.283   127    54   139 0.389 0.604     1 0.232 0.144
## 5 Shoh~ LAA     135   599 0.304   102    44    95 0.412 0.654    20 0.239 0.152
## # i 5 more variables: WPA <dbl>, WAR <dbl>, Earned <dbl>, PlayerId <dbl>,
## #   MLBAMID <dbl>
```

There are not major data cleaning issues for you to perform, as I pre-cleaned the data.

The data we will consider include:

- **Name (Name):** Player's name
- **Team (Team):** Player's team
- **Games (G):** Number of games in 2023 in which player appeared
- **Plate Appearances (PA):** Number of time in 2023 that player attempted an at bat
- **Batting Average (AVG):** Percent of times in which a plate appearance results in a hit
- **Runs (R):** Number of times that player cross home plate to score a point for team
- **Home Runs (HR):** Number of times player it it out of the part to score themselves and all players on base instantly.
- **Runs Batted In (RBI):** Number of runs scored due to player's at bats
- **On-Base Percentage (OBP):** Percent of time that plate appearances result in player reaching base safely
- **Slugging Average (SLG):** Average number of bases that a player covers by hit, walk or some other means of hitting the ball.
- **Stolen Bases (SB):** Number of times player advance base by "stealing" base
- **Strikeout Percentage (K_pct):** Percent of plate appearances that result in strikouts
- **Walk Percentage (BB_pct):** Percent of plate appearances that result in walks.
- **Win Probability Added (WPA):** Player's name
- **Wins Above Replacement (WAR):** Estimate of how many additional wins a team will receive by playing this player versus a low-level MLB player.
- **Earnings (Earned):** Estimates of money delivered to club by virtue of playing performance. In millions of dollars.

# Analysis

So who was good at what? Let's focus on the variable strikeout percentage. Strikeouts are bad because there is no possibilty of reaching base or advancing due to a defensive player error or fielder's choice.

What counts as a good or bad score? Let's look at the distribution of the statistic:

```
summary(data$K_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.05511 0.17061 0.20976 0.20788 0.23853 0.32700
```

Let's make ranked lists. This is how we get the top 20 performances in terms of strikeout percentage:

```
# Sort the data by Strikeout Percentage in ascending order to get those with the lowest K%
data <- data[order(data$K_pct),]

# Select the top 20 players with the lowest strikeout percentages
head(data[,c("Name", "Team", "K_pct")], 20)
```

```
## # A tibble: 20 x 3
##    Name                   Team   K_pct
##    <chr>                  <chr>  <dbl>
##  1 Luis Arraez            MIA   0.0551
##  2 Jeff McNeil            NYM   0.100 
##  3 Keibert Ruiz           WSN   0.103 
##  4 Steven Kwan            CLE   0.104 
##  5 José Ramírez           CLE   0.106 
##  6 Ronald Acuña Jr.       ATL   0.114 
##  7 Alex Bregman           HOU   0.120 
##  8 Nico Hoerner           CHC   0.121 
##  9 Kyle Tucker            HOU   0.136 
## 10 Masataka Yoshida       BOS   0.140 
## 11 Andrew Benintendi      CHW   0.143 
## 12 Gleyber Torres         NYY   0.146 
## 13 Marcus Semien          TEX   0.146 
## 14 Vladimir Guerrero Jr.  TOR   0.147 
## 15 Adley Rutschman        BAL   0.147 
## 16 Alec Bohm              PHI   0.154 
## 17 Mookie Betts           LAD   0.154 
## 18 Alex Verdugo           BOS   0.154 
## 19 Dominic Smith          WSN   0.155 
## 20 Mark Canha             - - - 0.156
```

Here are the worst performers:

```
# Sort the data by Strikeout Percentage in ascending order to get those with the lowest K%
data <- data[order(-data$K_pct),]

# Select the top 20 players with the lowest strikeout percentages
head(data[,c("Name", "Team", "K_pct")], 20)
```

```
## # A tibble: 20 x 3
##    Name               Team  K_pct
##    <chr>              <chr> <dbl>
##  1 Brent Rooker       OAK   0.327
##  2 Jack Suwinski      PIT   0.322
##  3 James Outman       LAD   0.319
##  4 Ryan McMahon      COL   0.316
##  5 Teoscar Hernández SEA   0.311
##  6 Eugenio Suárez    SEA   0.308
##  7 Kyle Schwarber    PHI   0.299
##  8 Josh Jung         TEX   0.293
##  9 Luis Robert Jr.   CHW   0.289
## 10 Matt Chapman      TOR   0.284
## 11 MJ Melendez       KCR   0.282
## 12 J.D. Davis        SFG   0.278
## 13 Anthony Volpe     NYY   0.278
## 14 Cal Raleigh       SEA   0.278
## 15 Trent Grisham     SDP   0.277
## 16 Adolis García     TEX   0.277
## 17 Jake Burger       - - - 0.276
## 18 Nick Castellanos  PHI   0.276
## 19 Ezequiel Tovar    COL   0.270
## 20 Max Muncy         LAD   0.264
```

To get Max Muncy’s information

```
data[data$Name == "Max Muncy", ]
```

```
## # A tibble: 1 x 18
##   Name  Team      G    PA   AVG     R    HR   RBI   OBP   SLG    SB K_pct BB_pct
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 Max ~ LAD     135   579 0.212    95    36   105 0.333 0.475     1 0.264  0.147
## # i 5 more variables: WPA <dbl>, WAR <dbl>, Earned <dbl>, PlayerId <dbl>,
## #   MLBAMID <dbl>
```

To only look at some of Max Muncy’s data:

```
data[data$Name == "Max Muncy", c("Name", "Team", "AVG", "HR", "K_pct")]
```

```
## # A tibble: 1 x 5
##   Name      Team    AVG    HR K_pct
##   <chr>     <chr> <dbl> <dbl> <dbl>
## 1 Max Muncy LAD   0.212    36 0.264
```

To look at all of the Mets:

```
subset(data, Team == "NYM")
```

```
## # A tibble: 4 x 18
##   Name  Team      G    PA   AVG     R    HR   RBI   OBP   SLG    SB K_pct BB_pct
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 Pete~ NYM     154   658 0.217    92    46   118 0.318 0.504     4 0.229 0.0988
## 2 Bran~ NYM     152   682 0.274    89    24    68 0.363 0.466     3 0.214 0.109
## 3 Fran~ NYM     160   687 0.254   108    31    98 0.336 0.470    31 0.199 0.0961
## 4 Jeff~ NYM     156   648 0.270    75    10    55 0.333 0.378    10 0.100 0.0602
## # i 5 more variables: WPA <dbl>, WAR <dbl>, Earned <dbl>, PlayerId <dbl>,
## #   MLBAMID <dbl>
```

To look at players with K_pct that are below 15%

```
subset(data, K_pct < 0.13)
```

```
## # A tibble: 8 x 18
##   Name        Team      G    PA   AVG     R    HR   RBI   OBP   SLG    SB  K_pct
##   <chr>       <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 Nico Hoern~ CHC     150   688 0.283    98     9    68 0.346 0.383    43 0.121
## 2 Alex Bregm~ HOU     161   724 0.262   103    25    98 0.363 0.441     3 0.120
## 3 Ronald Acu~ ATL     159   735 0.337   149    41   106 0.416 0.596    73 0.114
## 4 José Ramír~ CLE     156   691 0.282    87    24    80 0.356 0.475    28 0.106
## 5 Steven Kwan CLE     158   718 0.268    93     5    54 0.340 0.370    21 0.104
## 6 Keibert Ru~ WSN     136   562 0.260    55    18    67 0.308 0.409     1 0.103
## 7 Jeff McNeil NYM     156   648 0.270    75    10    55 0.333 0.378    10 0.100
## 8 Luis Arraez MIA     147   617 0.354    71    10    69 0.393 0.469     3 0.0551
## # i 6 more variables: BB_pct <dbl>, WPA <dbl>, WAR <dbl>, Earned <dbl>,
## #   PlayerId <dbl>, MLBAMID <dbl>
```

Here is a function to get a specific player's percentile score in strikeout percentage:

```r
get_strikeout_percentile <- function(data, playerName, teamName) {
  # Filter data for the specific player
  player_data <- subset(data, Name == playerName & Team == teamName)

  # Check if player data exists
  if(nrow(player_data) == 0) {
    return(paste("No data found for", playerName, "in", teamName))
  }

  # Calculate the player's strikeout percentage rank among all players
  player_rank <- sum(data$K_pct < player_data$K_pct) + 1

  # Calculate the percentile
  player_percentile <- (player_rank / nrow(data)) * 100

  # Return the player's percentile score
  return(paste(playerName, "from", teamName, "is in the", round(player_percentile, 2), "percentile for s
}


get_strikeout_percentile(data, "Max Muncy", "LAD")
```

```
## [1] "Max Muncy from LAD is in the 85.71 percentile for strikeout percentage."
```

So let's get started figuring out who is good or bad to give our clients names to consider.