MLB Batting Analysis For 2021

Shane Hylton

12/6/2021

Contents

Abstract

Baseball provides some of the most interesting and comprehensive statistics of any sport. Analytics are a very heavy component in baseball, and I set out to explore a few different criteria for analysis. I chose to explore baseball statistics because I have always been very interested in the math behind the probabilities in baseball.

Research Question

Consider a random selection of nine batters: 3 outfielders, 1 Designated Hitter, 1 Catcher, and 4 Infielders. If this team were in a position where they needed a base hit and they could choose any batter (with the only information on the batter being their position), which position provides the highest opportunity for a hit?

I also chose to test whether there is a relationship between a player's age and their batting average.

I approached the research question by first establishing summary statistics, then I constructed a simulation to determine the optimal choice for the question at hand. I used linear regression to test whether there was a linear relationship between age and batting average.

Findings

By running numerous simulations, I was able to conclude that the position that would give a random team the best opportunity at recording a hit is the designated hitter. I was also able to conclude that there is an immensely small linear relationship between age and batting average; the relationship is measurably zero.

Introduction

Goals

- Provide Relevant Summary Statistics
- Construct a Regression Model for the relationship between age and batting average
- Visualize the differences in batting efficiency for each position
- Construct a simulation to show which position is most likely to successfully record a hit
- Demonstrate which position to choose given different amounts of players to choose from at random

The Data

I have selected two data sets, one as a primary and one as a secondary. The primary data set is from Rotowire. There are some elements of the data set that are better than the secondary data set and there are some elements that are not as good. In the Rotowire set, the data is much more trimmed. The Baseball Reference data set that I collected contains far more players, but many of them are irrelevant for the purpose of this analysis.

Originally, I was going to evaluate the data based on players who have had at least 50 at bats, but the data is not normalized until closer to 125 at-bats.

```
link <- 'https://raw.githubusercontent.com/st3vejobs/DATA-606-Final-Project/main/mlb-player-stats-Batte
mlbraw <- read.csv(url(link), na.strings = "")

mlb <- mlbraw[mlbraw$AB >= 50, ]
```

```r
summary(mlb)
```

```
##     Player              Team               Pos                 Age       
##  Length:556         Length:556         Length:556         Min.   :20.00  
##  Class :character   Class :character   Class :character   1st Qu.:26.00  
##  Mode  :character   Mode  :character   Mode  :character   Median :28.00  
##                                                           Mean   :28.57  
##                                                           3rd Qu.:31.00  
##                                                           Max.   :41.00  
##        G                AB              R                H         
##  Min.   : 14.00   Min.   : 50.0   Min.   :  2.00   Min.   :  4.00  
##  1st Qu.: 51.00   1st Qu.:132.8   1st Qu.: 16.00   1st Qu.: 30.00  
##  Median : 82.50   Median :239.5   Median : 31.00   Median : 58.00  
##  Mean   : 86.85   Mean   :277.5   Mean   : 38.51   Mean   : 69.09  
##  3rd Qu.:124.25   3rd Qu.:412.2   3rd Qu.: 56.00   3rd Qu.:104.25  
##  Max.   :162.00   Max.   :664.0   Max.   :123.00   Max.   :191.00  
##       X2B             X3B              HR             RBI        
##  Min.   : 0.00   Min.   :0.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 6.00   1st Qu.:0.000   1st Qu.: 3.00   1st Qu.: 14.00  
##  Median :11.00   Median :1.000   Median : 7.00   Median : 31.50  
##  Mean   :13.77   Mean   :1.178   Mean   :10.55   Mean   : 36.94  
##  3rd Qu.:21.00   3rd Qu.:2.000   3rd Qu.:14.25   3rd Qu.: 53.00  
##  Max.   :42.00   Max.   :8.000   Max.   :48.00   Max.   :121.00  
##        SB               CS               BB               SO        
##  Min.   : 0.000   Min.   : 0.000   Min.   :  1.00   Min.   :  4.00  
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 11.00   1st Qu.: 35.75  
##  Median : 2.000   Median : 1.000   Median : 22.00   Median : 62.00  
##  Mean   : 3.908   Mean   : 1.246   Mean   : 27.52   Mean   : 69.78  
##  3rd Qu.: 5.000   3rd Qu.: 2.000   3rd Qu.: 38.00   3rd Qu.: 99.00  
##  Max.   :40.000   Max.   :10.000   Max.   :145.00   Max.   :202.00  
##        SH                SF              HBP              AVG        
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0780  
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 1.000   1st Qu.:0.2100  
##  Median : 0.0000   Median : 1.000   Median : 3.000   Median :0.2430  
##  Mean   : 0.5737   Mean   : 1.993   Mean   : 3.676   Mean   :0.2368  
##  3rd Qu.: 1.0000   3rd Qu.: 3.000   3rd Qu.: 5.000   3rd Qu.:0.2650  
##  Max.   :12.0000   Max.   :12.000   Max.   :27.000   Max.   :0.3560  
##       OBP              SLG              OPS        
##  Min.   :0.1050   Min.   :0.1370   Min.   :0.2660  
##  1st Qu.:0.2820   1st Qu.:0.3438   1st Qu.:0.6320  
##  Median :0.3120   Median :0.3945   Median :0.7110  
##  Mean   :0.3097   Mean   :0.3967   Mean   :0.7064  
##  3rd Qu.:0.3400   3rd Qu.:0.4502   3rd Qu.:0.7792  
##  Max.   :0.4660   Max.   :0.6470   Max.   :1.0900  
```

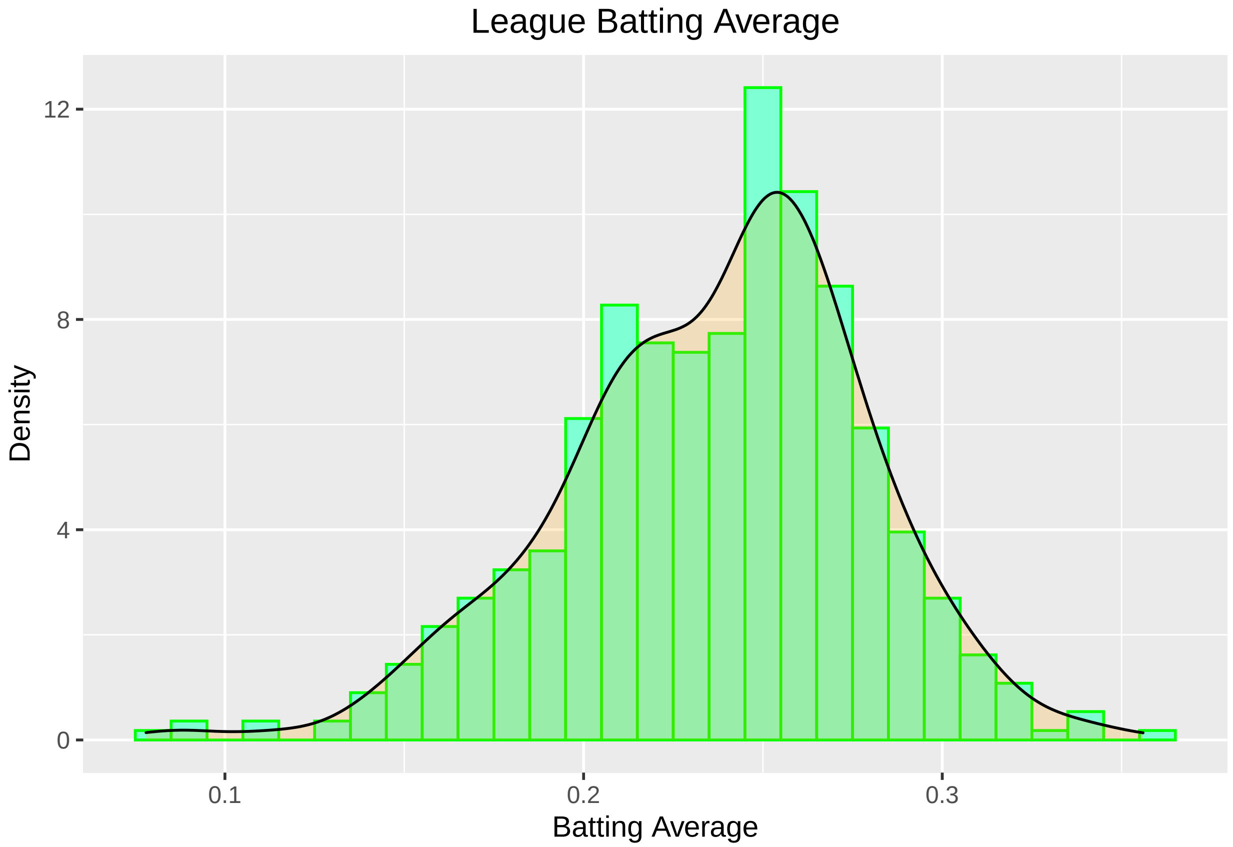
```r
meanavg <- mean(mlb$AVG, na.rm = TRUE)
medavg <- median(mlb$AVG, na.rm = TRUE)

paste("The mean batting average in 2021 is: ", round(meanavg, 3))
```

```
## [1] "The mean batting average in 2021 is:  0.237"
```

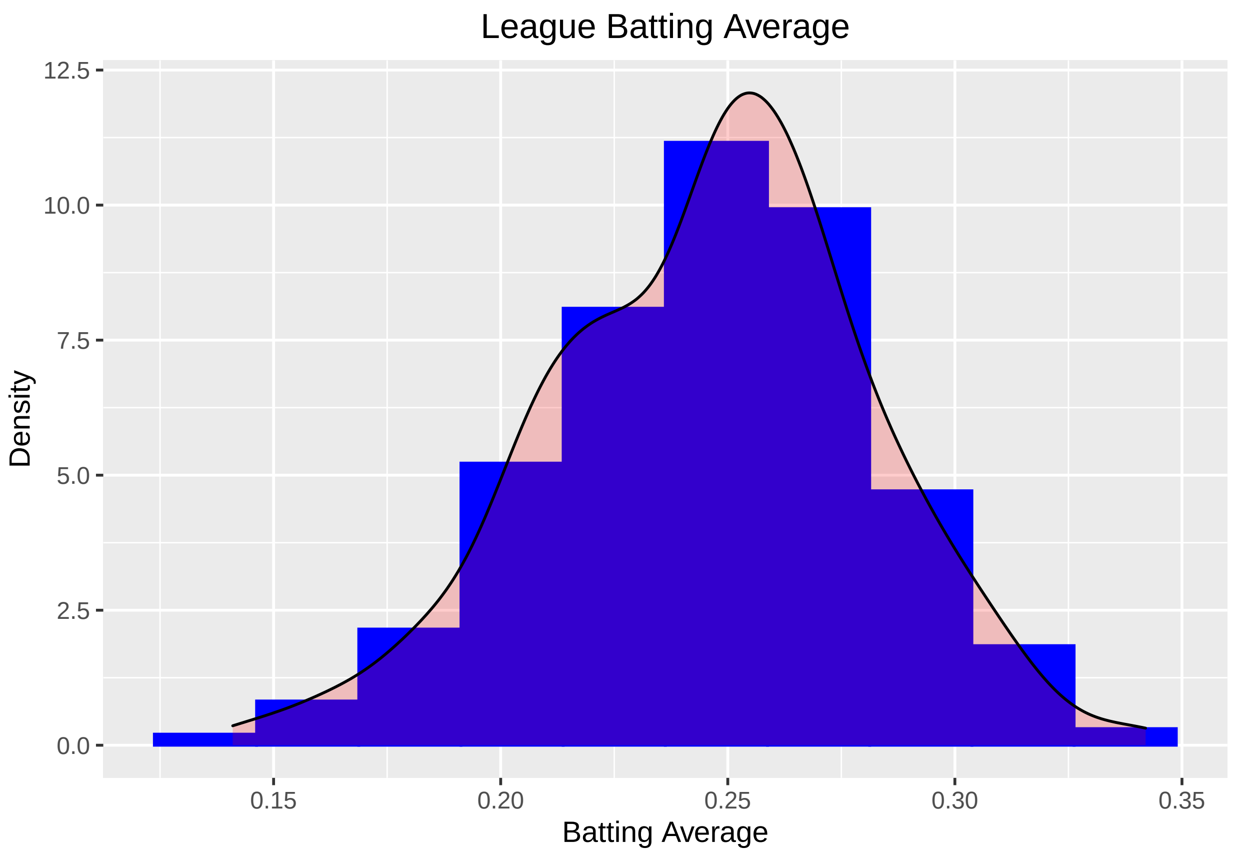
```
paste("The median batting average in 2021 is: ", round(medavg, 3))
```

```
## [1] "The median batting average in 2021 is:  0.243"
```

```
ggplot(mlb, aes(x=AVG))+
  geom_histogram(aes(y=..density..), color = 'green', fill = 'aquamarine', binwidth = .01)+
  geom_density(alpha=.2, fill = 'orange')+
  ggtitle('League Batting Average')+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab('Batting Average')+
  ylab('Density')
```



```
mlbtrim <- mlb[mlb$AB >= 125, ]
mlb_50 <- mlb[mlb$AB >= 50, ]
mlb <- mlb[mlb$AB >= 125, ]

ggplot(mlbtrim, aes(x=AVG))+
  geom_histogram(aes(y=..density..), color = 'blue', fill = 'blue', binwidth = .0225)+
  geom_density(alpha=.2, fill = 'red')+
  ggtitle('League Batting Average')+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab('Batting Average')+
  ylab('Density')
```



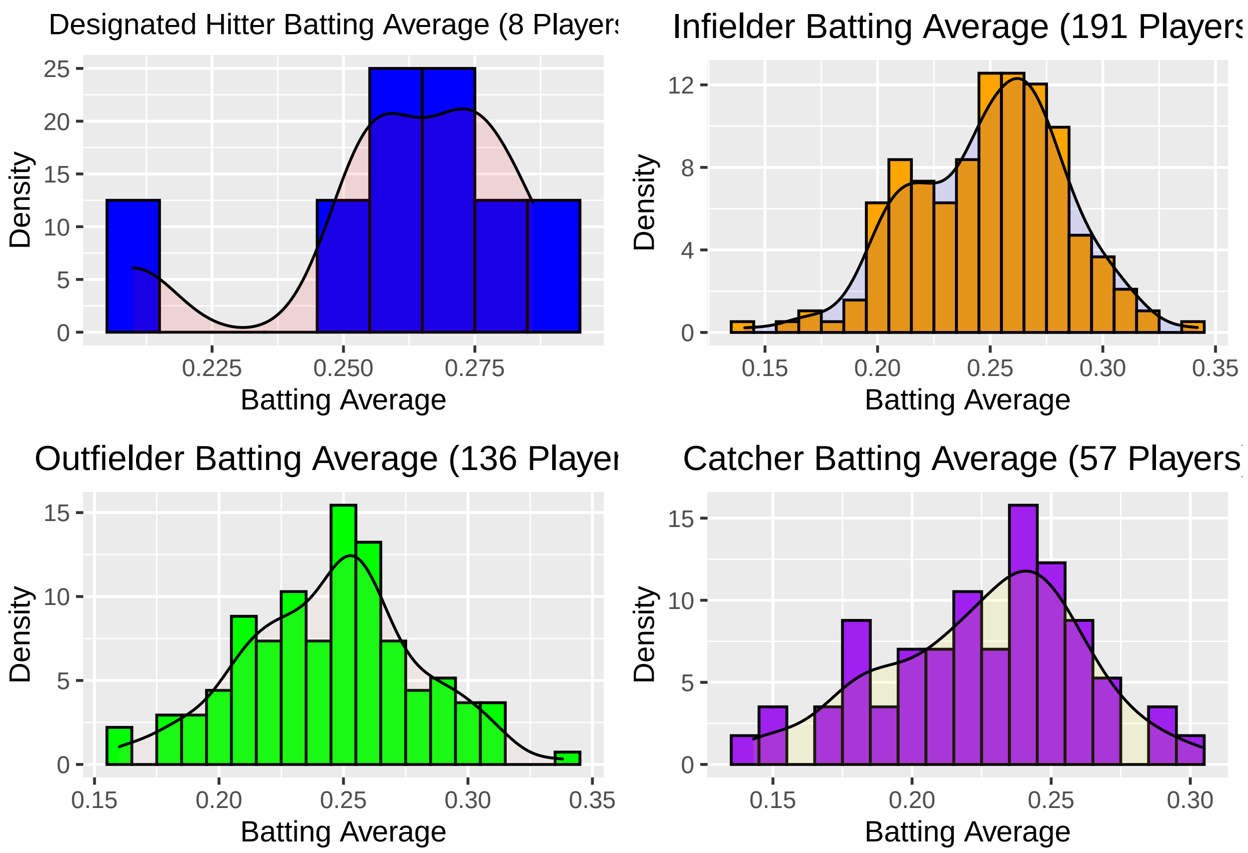
Exploration and Summary Statistics

First, I will provide a histogram of batting average distributions, filtered by position. For simplicity, all infield positions will be under the infield umbrella, rather than first base, second base, shortstop and third base. The same simplification will be made for outfielders, as many positions are not fixed.

One issue with the data is that there are thirty-five players whose statistics are skewed because they played for two separate teams, and they are entered in as separate rows. Because of the complexity of the calculations, I will simply omit them from consideration, which leaves 392 cases, which is a sufficiently large sample.

```
mlb <- mlb %>%
  group_by(Player) %>% arrange(Player)

#which(mlb$Player == "Adam Eaton")

indexes <- c()

for (idx in 1:(nrow(mlb) - 1)){
    if ( mlb[idx, ]$Player == mlb[(idx + 1), ]$Player){
      indexes <- append(indexes, idx)
      indexes <- append(indexes, (idx + 1))
  }
}

mlb <- mlb[-c(indexes), ]
```

```
dh <- subset(mlb, Pos == 'DH')
inf <- subset(mlb, Pos == '1B' | Pos == '2B' | Pos == 'SS' | Pos == '3B')
of <- subset(mlb, Pos == 'OF' | Pos == 'LF' | Pos == 'CF' | Pos == 'RF')
c <- subset(mlb, Pos == 'C')

d <- ggplot(dh, aes(x=AVG))+
  geom_histogram(aes(y=..density..), color = 'black', fill = 'blue', binwidth = .01)+
  geom_density(alpha=.1, fill = 'red')+
  ggtitle('Designated Hitter Batting Average (8 Players)')+
  theme(plot.title = element_text(hjust = 0.5, size = 11))+
  xlab('Batting Average')+
  ylab('Density')

i <- ggplot(inf, aes(x=AVG))+
  geom_histogram(aes(y=..density..), color = 'black', fill = 'orange', binwidth = .01)+
  geom_density(alpha=.1, fill = 'blue')+
  ggtitle('Infielder Batting Average (191 Players)')+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab('Batting Average')+
  ylab('Density')

o <- ggplot(of, aes(x=AVG))+
  geom_histogram(aes(y=..density..), color = 'black', fill = 'green', binwidth = .01)+
  geom_density(alpha=.1, fill = 'pink')+
  ggtitle('Outfielder Batting Average (136 Players)')+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab('Batting Average')+
  ylab('Density')

c_plt <- ggplot(c, aes(x=AVG))+
  geom_histogram(aes(y=..density..), color = 'black', fill = 'purple', binwidth = .01)+
  geom_density(alpha=.1, fill = 'yellow')+
  ggtitle('Catcher Batting Average (57 Players)')+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab('Batting Average')+
  ylab('Density')


pos_hist <- ggarrange(d, i, o, c_plt,
                      ncol = 2, nrow = 2)

pos_hist
```



The results of the designated hitter histogram are somewhat negligible because the sample is so small. Interestingly, the histograms for each other position group shows a negative skew (left skew).

```
summary(dh$AVG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2100  0.2555  0.2640  0.2605  0.2740  0.2860
```

```
summary(inf$AVG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1410  0.2245  0.2520  0.2493  0.2710  0.3420
```

```
summary(of$AVG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.2218  0.2475  0.2447  0.2645  0.3380
```

```
summary(c$AVG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1430  0.2020  0.2320  0.2256  0.2470  0.3040
```

The Optimal Position

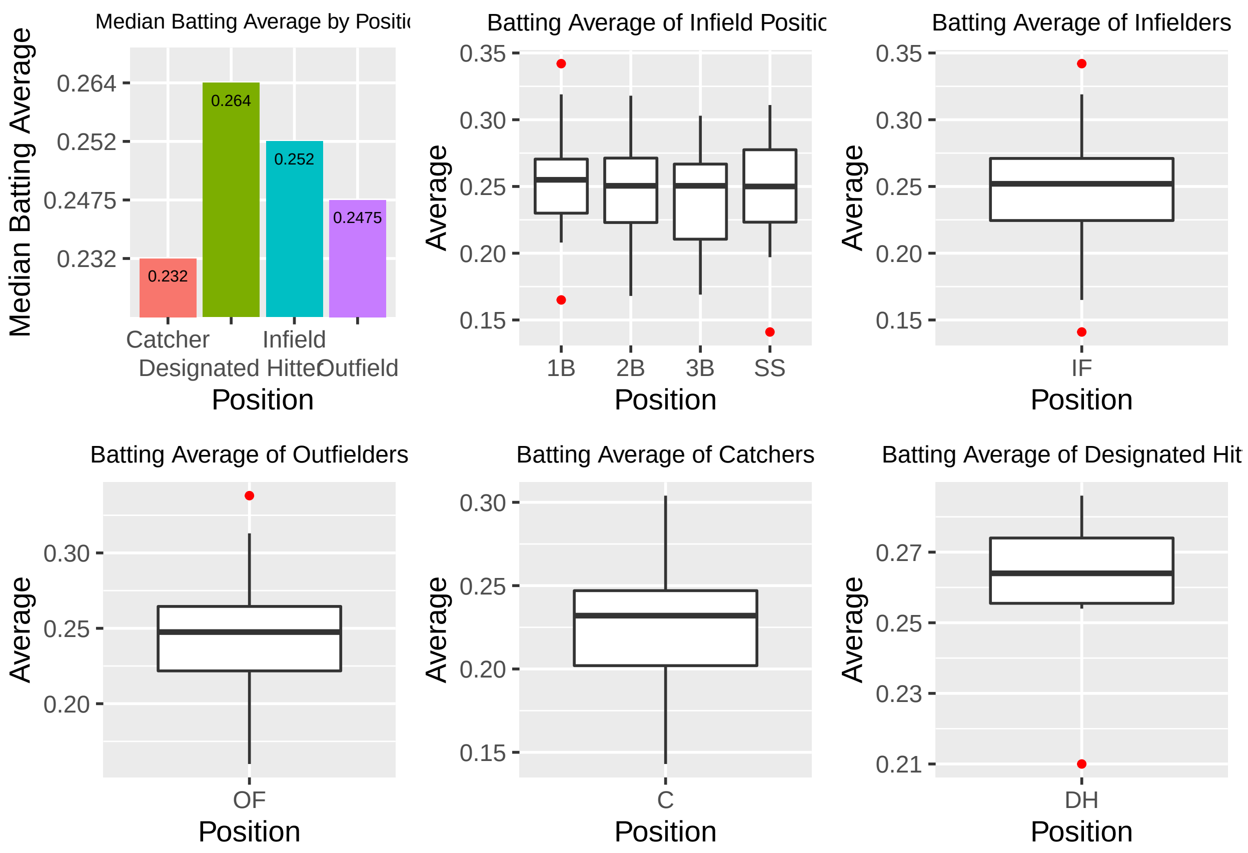
Batting average is the best measure of whether or not an at-bat will result in a base hit, of any kind. Thus, the optimal position to count on for a hit would be represented by the distributions of averages for each position. In each case, the distributions are left-skewed, and the median will be the best representation of the distribution, with IQR being the optimal measure of the spread of the data.

```r
comp <- data.frame(matrix(c("dh", median(dh$AVG), IQR(dh$AVG)), nrow = 1, ncol = 3))
colnames(comp) <- c("Position", "med", "iqr")
comp <- rbind(comp, c("inf", median(inf$AVG), IQR(inf$AVG)))
comp <- rbind(comp, c("of", median(of$AVG), IQR(of$AVG)))
comp <- rbind(comp, c("c", median(c$AVG), IQR(c$AVG)))

comp
```

```
##   Position    med     iqr
## 1       dh  0.264  0.0185
## 2      inf  0.252  0.0465
## 3       of 0.2475 0.04275
## 4        c  0.232   0.045
```

The highest median batting average and the lowest interquartile range belong to the designated hitter position. This result indicates that the most likely position to record a hit in an at-bat is the designated hitter position.

```r
p1 <- ggplot(comp, aes(x = factor(Position), y = med, fill = Position))+
   geom_col(show.legend = FALSE)+
   ylab('Median Batting Average')+
   xlab('Position')+
   ggtitle('Median Batting Average by Position')+
   geom_text(aes(label = med), vjust = 2, size = 2)+
   scale_x_discrete(labels = c("dh" = "Designated Hitter","inf" = "Infield", "of" = "Outfield","c" = "Ca
   theme(plot.title = element_text(hjust = 0.5, size = 8))


p2 <- ggplot(inf, aes(x=Pos, y=AVG))+
  geom_boxplot(outlier.color = 'red', outlier.size = 1)+
  xlab('Position')+
  ylab('Average')+
  ggtitle('Batting Average of Infield Position')+
  theme(plot.title = element_text(hjust = 0.5, size = 9))


mlb_new <- mlb
mlb_new[mlb_new == "1B" | mlb_new == "2B" | mlb_new == "SS" | mlb_new == "3B"] <- "IF"
mlb_new[mlb_new == "OF" | mlb_new == "LF" | mlb_new == "CF" | mlb_new == "3B"] <- "OF"

inf_new <- subset(mlb_new, Pos == "IF")

p3 <- ggplot(inf_new, aes(x=Pos, y=AVG))+
  geom_boxplot(outlier.color = 'red', outlier.size = 1)+
  xlab('Position')+
  ylab('Average')+
  ggtitle('Batting Average of Infielders')+
  theme(plot.title = element_text(hjust = 0.5, size = 9))

p4 <- ggplot(of, aes(x=Pos, y=AVG))+
  geom_boxplot(outlier.color = 'red', outlier.size = 1)+
  xlab('Position')+
  ylab('Average')+
  ggtitle('Batting Average of Outfielders')+
  theme(plot.title = element_text(hjust = 0.5, size = 9))

p5 <- ggplot(c, aes(x=Pos, y=AVG))+
  geom_boxplot(outlier.color = 'red', outlier.size = 1)+
  xlab('Position')+
  ylab('Average')+
  ggtitle('Batting Average of Catchers')+
  theme(plot.title = element_text(hjust = 0.5, size = 9))

p6 <- ggplot(dh, aes(x=Pos, y=AVG))+
  geom_boxplot(outlier.color = 'red', outlier.size = 1)+
  xlab('Position')+
  ylab('Average')+
  ggtitle('Batting Average of Designated Hitters')+
  theme(plot.title = element_text(hjust = 0.5, size = 8.5))

figure <- ggarrange(p1,p2,p3,p4,p5,p6,
                    nrow = 2, ncol = 3)
figure
```



From these figures, it is worth noting that there are a few extreme outliers in the distributions, but they will be included in the simulation because they fit the criteria of having more than 125 at-bats.

Answering the Question

The goal of this analysis is to determine which position would be the ideal position to rely on for a hit in a team of nine random players, 3 outfielders, 1 designated hitter, 1 catcher, and 4 infielders. My assumption is that the designated hitter would be the ideal position in this simulation. If the exact attributes of every player on the team were known, the decision may very well change, but the goal is to find a general answer that would work most of the time.

Hypothesis

I hypothesize that the designated hitter will prove to be the player to select when a hit is needed and no other information is known.

Team Construction

First, I will construct a random team that fits these parameters.

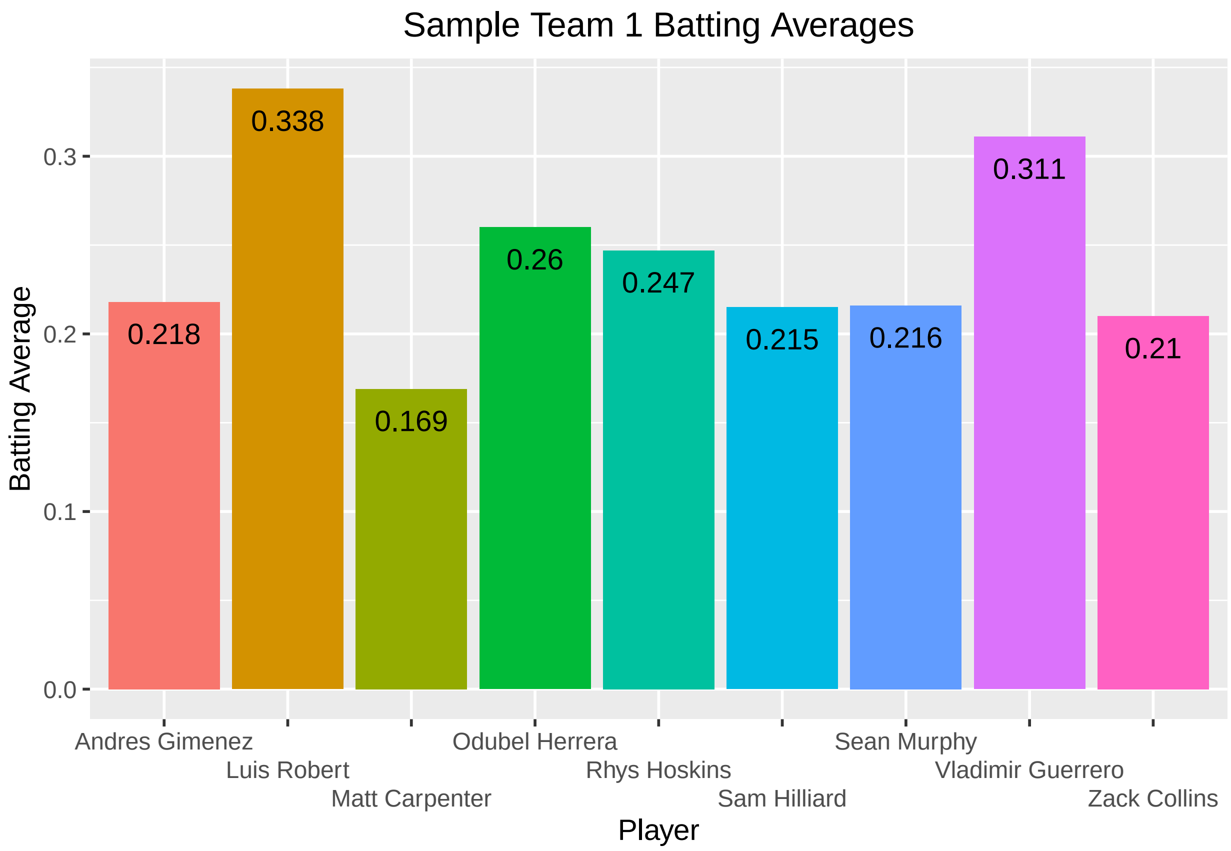
```
dh_idx <- c(which(mlb$Pos == 'DH'))
inf_idx <- c(which(mlb$Pos == '1B' | mlb$Pos == '2B' | mlb$Pos == 'SS' | mlb$Pos == '3B'))
of_idx <- c(which(mlb$Pos == 'OF' | mlb$Pos == 'LF' | mlb$Pos == 'CF' | mlb$Pos == 'RF'))
c_idx <- c(which(mlb$Pos == 'C'))
```

```
dh_t1 <- mlb[as.numeric(sample(dh_idx, 1)), ]
inf_t1 <- mlb[as.numeric(sample(inf_idx, 4)), ]
of_t1 <- mlb[as.numeric(sample(of_idx, 3)), ]
c_t1 <- mlb[as.numeric(sample(c_idx, 1)), ]

t1 <- rbind(dh_t1, inf_t1, of_t1, c_t1)

t1[as.numeric(which.max(t1$AVG)), ]$Pos
```

```
## [1] "OF"
```

Interestingly, in the first trial I performed, the designated hitter had the second worst batting average on the team.

```
ggplot(t1, aes(x = factor(Player), y = AVG, fill = Player))+
   geom_col(show.legend = FALSE)+
   ylab('Batting Average')+
   xlab('Player')+
   ggtitle('Sample Team 1 Batting Averages')+
   geom_text(
     aes(label = AVG),
vjust = 2 )+
   scale_x_discrete(guide = guide_axis(n.dodge = 3))+
   theme(plot.title = element_text(hjust = 0.5))
```



Simulation of 1000 Random Teams

This simulation produces 1000 random teams of 3 outfielders, 4 infielders, 1 catcher, and 1 designated hitter. Each time, the player on the team with the highest batting average is selected and their position is added to a data frame for tabulation.

If an infielder is the team's top batter 50% of the time, this does not mean that if the team chooses a random infielder, they have a 50% probability of choosing the best player. One potential miscalculation is the risk of choosing the wrong infielder. There is added risk involved when choosing an infielder at random, because by selecting an infielder at random, there is a 1/4 chance that the chosen player is the highest performing infielder, so the calculation needs to be adjusted to account for that.

One possible method for adapting to this issue would be after selecting the team, using the median for each position group as that group's batting average. Considering the small size of the 3 or 4 players in the team subset, it may be more accurate to randomly select one infielder and one outfielder after constructing a team.

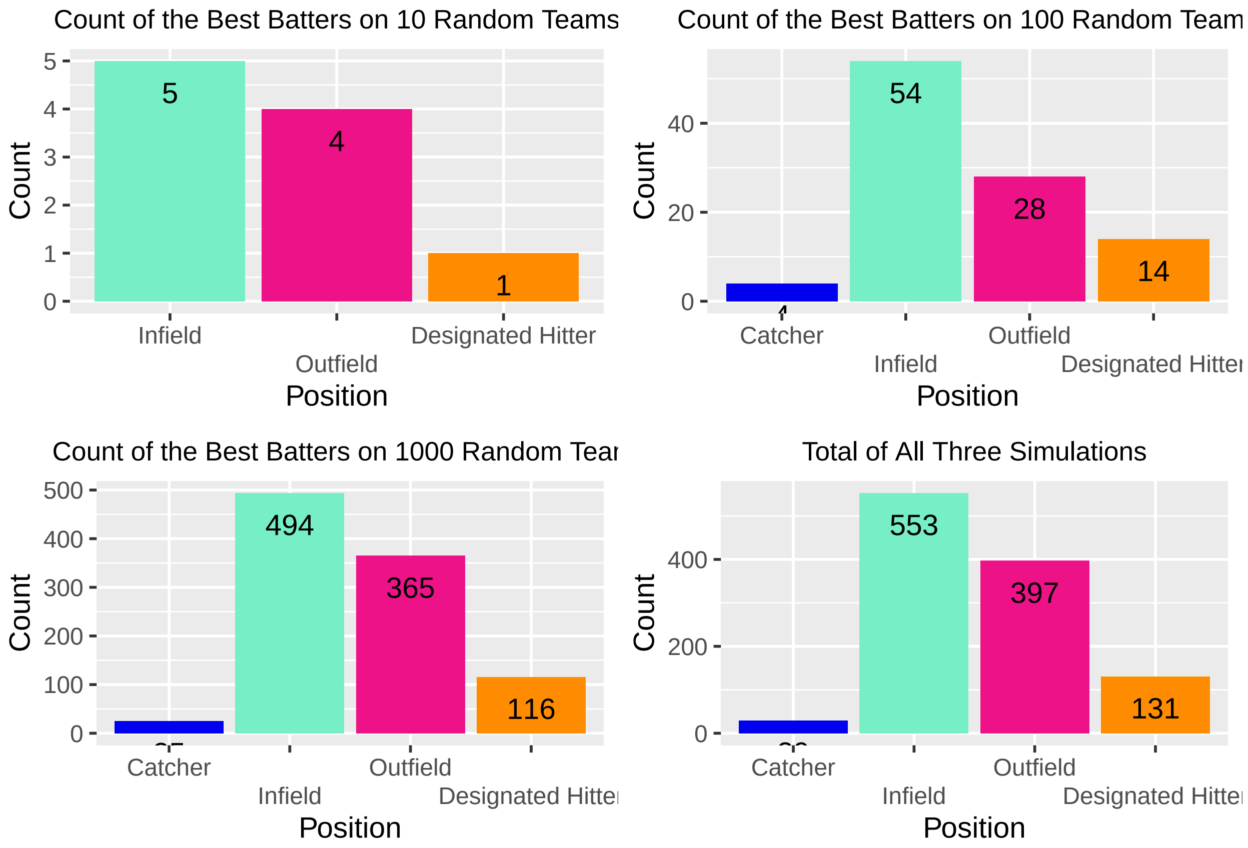
Given the setup of this simulation, I will not set the seed to a fixed amount. If I set a seed, the team function will output the same result every time.

```r
team <- function(x = NULL){
  if(is.null(x)){

    mlb_new <- mlb
    mlb_new[mlb_new == "1B" | mlb_new == "2B" | mlb_new == "SS" | mlb_new == "3B"] <- "IF"
    mlb_new[mlb_new == "OF" | mlb_new == "LF" | mlb_new == "CF" | mlb_new == "3B"] <- "OF"

    dht <- mlb_new[as.numeric(sample(dh_idx, 1)), ]
    inft <- mlb_new[as.numeric(sample(inf_idx, 4)), ]
    oft <- mlb_new[as.numeric(sample(of_idx, 3)), ]
    ct <- mlb_new[as.numeric(sample(c_idx, 1)), ]

    roster <- rbind(dht, inft, oft, ct)
    roster[as.numeric(which.max(roster$AVG)), ]$Pos

  }
}


team()
```

```
## [1] "IF"
```

```r
sim <- data.frame(table(data.frame(matrix(replicate(10, team())))))
colnames(sim) <- c("Pos_best", "count")
sim$Pos_best <- factor(sim$Pos_best, levels = c("C", "IF", "OF", "DH"))

sim2 <- data.frame(table(data.frame(matrix(replicate(100, team())))))
colnames(sim2) <- c("Pos_best", "count")
sim2$Pos_best <- factor(sim2$Pos_best, levels = c("C", "IF", "OF", "DH"))

sim3 <- data.frame(table(data.frame(matrix(replicate(1000, team())))))
colnames(sim3) <- c("Pos_best", "count")
sim3$Pos_best <- factor(sim3$Pos_best, levels = c("C", "IF", "OF", "DH"))

together <- rbind(sim, sim2, sim3)
```

```r
total <- together %>%
  group_by(Pos_best) %>%
  summarise_at(vars(count),
               list(count = sum))
total$Pos_best <- factor(total$Pos_best, levels = c("C", "IF", "OF", "DH"))

#as.character(sim[which.max(sim$count), ]$Pos_best)

fixed_colors <- c(DH = 'darkorange', IF = 'aquamarine2', OF = 'deeppink2', C = 'blue2')

simplot <- ggplot(sim, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Count of the Best Batters on 10 Random Teams')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))

sim2plot <- ggplot(sim2, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Count of the Best Batters on 100 Random Teams')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))

sim3plot <- ggplot(sim3, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Count of the Best Batters on 1000 Random Teams')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))


totalplot <- ggplot(total, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Total of All Three Simulations')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))


results <- ggarrange(simplot, sim2plot, sim3plot, totalplot,
                      ncol = 2, nrow = 2)

results
```



Infield overwhelmingly appears to be the best choice, followed by outfield. This does not tell the full story, however, because there are four possible choices for infielders on each team.

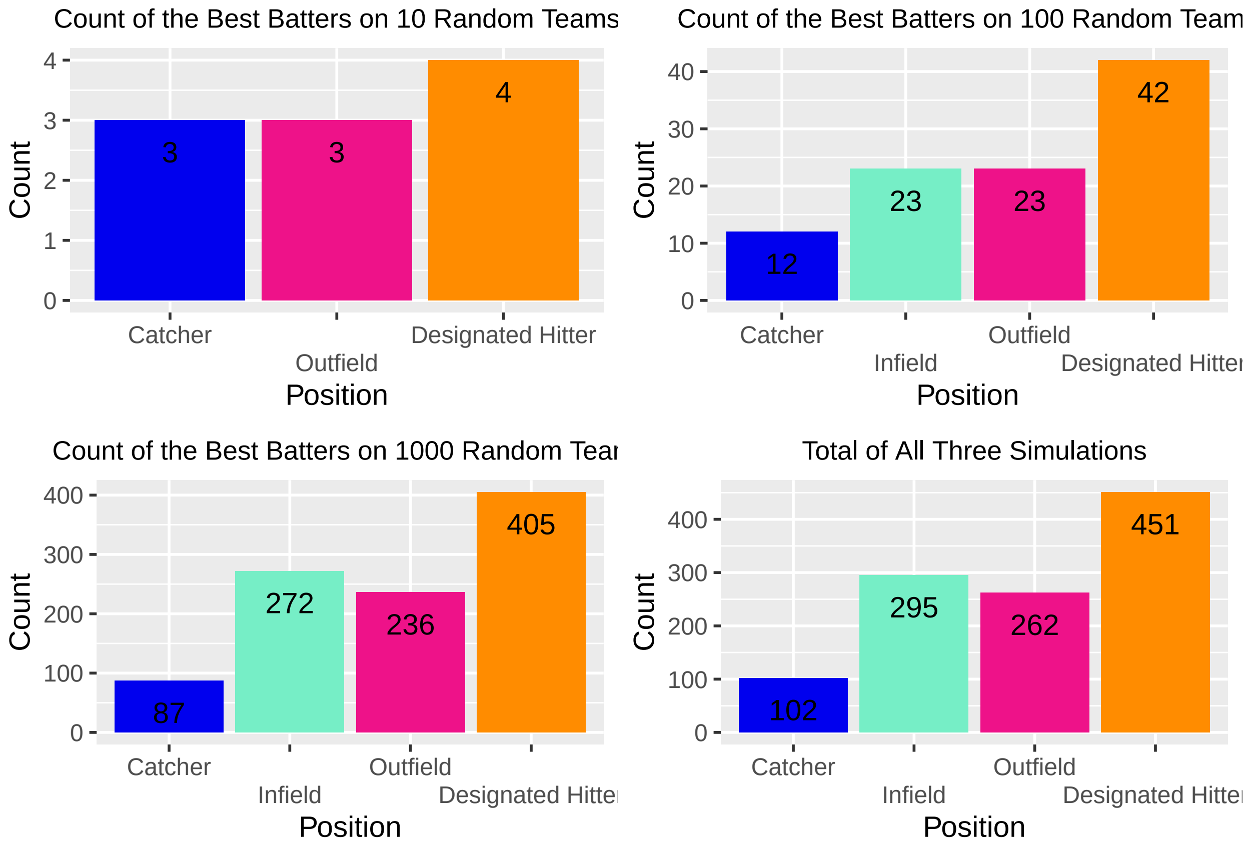
Adjusting the Team Selection

After the team is selected, I will randomly select one player from each position group to represent the group in the simulation. This should provide enough randomness to eliminate the advantage of having several opportunities to have the highest batting average.

```
team_adj <- function(x = NULL){
  if(is.null(x)){

    mlb_new <- mlb
    mlb_new[mlb_new == "1B" | mlb_new == "2B" | mlb_new == "SS" | mlb_new == "3B"] <- "IF"
    mlb_new[mlb_new == "OF" | mlb_new == "LF" | mlb_new == "CF" | mlb_new == "3B"] <- "OF"

    dht <- mlb_new[as.numeric(sample(dh_idx, 1)), ]
    inft <- mlb_new[as.numeric(sample(inf_idx, 4)), ]
    inf_fin <- inft[as.numeric(sample(1:4, 1)), ]
    oft <- mlb_new[as.numeric(sample(of_idx, 3)), ]
    of_fin <- oft[as.numeric(sample(1:3, 1)), ]
    ct <- mlb_new[as.numeric(sample(c_idx, 1)), ]

    roster <- rbind(dht, inf_fin, of_fin, ct)
    roster[as.numeric(which.max(roster$AVG)), ]$Pos

  }
}

team_adj()
```

```
## [1] "IF"
```

```
sim_adj <- data.frame(table(data.frame(matrix(replicate(10, team_adj())))))
colnames(sim_adj) <- c("Pos_best", "count")
sim_adj$Pos_best <- factor(sim_adj$Pos_best, levels = c("C", "IF", "OF", "DH"))

sim_adj2 <- data.frame(table(data.frame(matrix(replicate(100, team_adj())))))
colnames(sim_adj2) <- c("Pos_best", "count")
sim_adj2$Pos_best <- factor(sim_adj2$Pos_best, levels = c("C", "IF", "OF", "DH"))

sim_adj3 <- data.frame(table(data.frame(matrix(replicate(1000, team_adj())))))
colnames(sim_adj3) <- c("Pos_best", "count")
sim_adj3$Pos_best <- factor(sim_adj3$Pos_best, levels = c("C", "IF", "OF", "DH"))

together_adj <- rbind(sim_adj, sim_adj2, sim_adj3)
total_adj <- together_adj %>%
  group_by(Pos_best) %>%
  summarise_at(vars(count),
               list(count = sum))
total_adj$Pos_best <- factor(total_adj$Pos_best, levels = c("C", "IF", "OF", "DH"))


simplot_adj <- ggplot(sim_adj, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Count of the Best Batters on 10 Random Teams')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))

sim2plot_adj <- ggplot(sim_adj2, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Count of the Best Batters on 100 Random Teams')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))

sim3plot_adj <- ggplot(sim_adj3, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Count of the Best Batters on 1000 Random Teams')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))


totalplot_adj <- ggplot(total_adj, aes(x = factor(Pos_best), y = count, fill = Pos_best))+
   geom_col(show.legend = FALSE)+
   ylab('Count')+
   xlab('Position')+
   ggtitle('Total of All Three Simulations')+
   geom_text(aes(label = count), vjust = 2 )+
  scale_fill_manual(values = fixed_colors)+
  scale_x_discrete(labels = c("DH" = "Designated Hitter","IF" = "Infield", "OF" = "Outfield","C" = "Cat
   theme(plot.title = element_text(hjust = 0.5, size = 10))


results_adj <- ggarrange(simplot_adj, sim2plot_adj, sim3plot_adj, totalplot_adj,
                      ncol = 2, nrow = 2)

results_adj
```



Conclusions

The simulation validates the hypothesis that the designated hitter position would be the best position to choose on any given team to give the team the best chance at recording a hit.

The second iteration where the advantage of having four infielders was removed shows that at random, selecting an infielder does not provide the highest probability of success.

The first simulation would be valid if the manager were allowed to know the players' batting averages after they are assigned to the team. This may be closer to reality, but it is not valid in this experiment.

Follow Up: Is there a relationship between age and batting average?

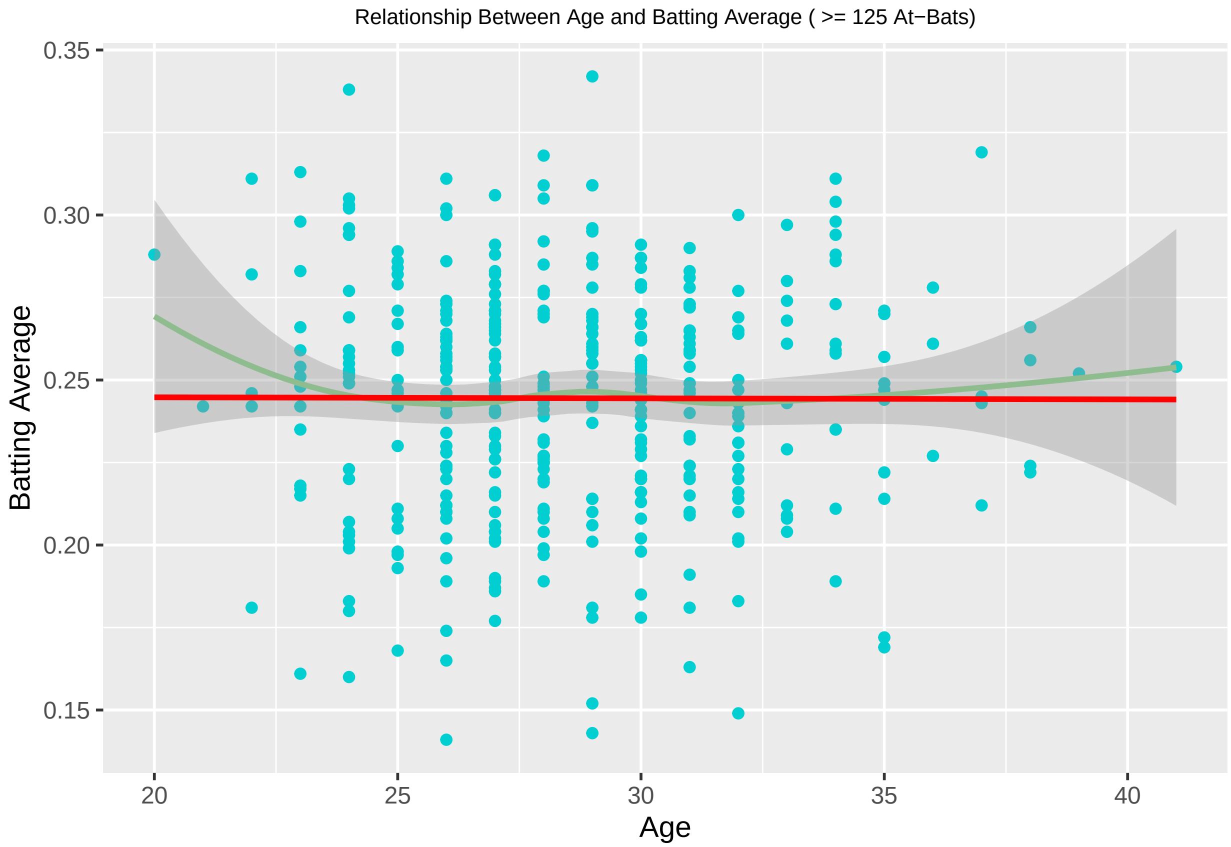
I will evaluate for greater than or equal to 125 at-bats, and greater than or equal to 50 at bats.

Greater Than or Equal to 125 At-Bats

```
agebat <- ggplot(mlb, aes(x = Age, y = AVG, na.rm = TRUE))+
  geom_point(na.rm = TRUE, color = 'darkturquoise')+
  geom_smooth(color = 'darkseagreen')+
  geom_smooth(method = "lm", color = 'red', se = FALSE)+
  ggtitle('Relationship Between Age and Batting Average ( >= 125 At-Bats)')+
  xlab('Age')+
  ylab('Batting Average')+
  theme(plot.title = element_text(hjust = 0.5, size = 7.5))

agebat
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

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



Regression Analysis

```
fit <- lm(AVG ~ Age, data = mlb)
summary(fit)
```

```
##
## Call:
## lm(formula = AVG ~ Age, data = mlb)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.103573 -0.024422  0.003395  0.023475  0.097522
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.454e-01  1.456e-02  16.854   <2e-16 ***
## Age         -3.178e-05  5.064e-04  -0.063     0.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03501 on 390 degrees of freedom
## Multiple R-squared:  1.01e-05,   Adjusted R-squared:  -0.002554
## F-statistic: 0.003938 on 1 and 390 DF,  p-value: 0.95
```

```r
set.seed(29)
# x <- runif(1, min(mlb$Age), max(mlb$Age)) -- If age were not in integer form
x <- sample(min(mlb$Age):max(mlb$Age), 1)
eq <- .00003178*x + .2454

paste("Predicted Batting Average for a given Age ", round(x, 0), " : ", round(eq, 3))
```

```
## [1] "Predicted Batting Average for a given Age  24  :  0.246"
```

```r
idx <- as.numeric(which.min(abs(x - mlb$Age)))
actual <- mlb$AVG[idx]
actual
```

```
## [1] 0.252
```

```r
paste("Actual Sample Batting Average for a given Age ", mlb$Age[idx], " : ", round(actual, 3))
```

```
## [1] "Actual Sample Batting Average for a given Age  24  :  0.252"
```

```r
resid <- eq - actual

paste("Residual: ", round(resid, 5))
```

```
## [1] "Residual:  -0.00584"
```

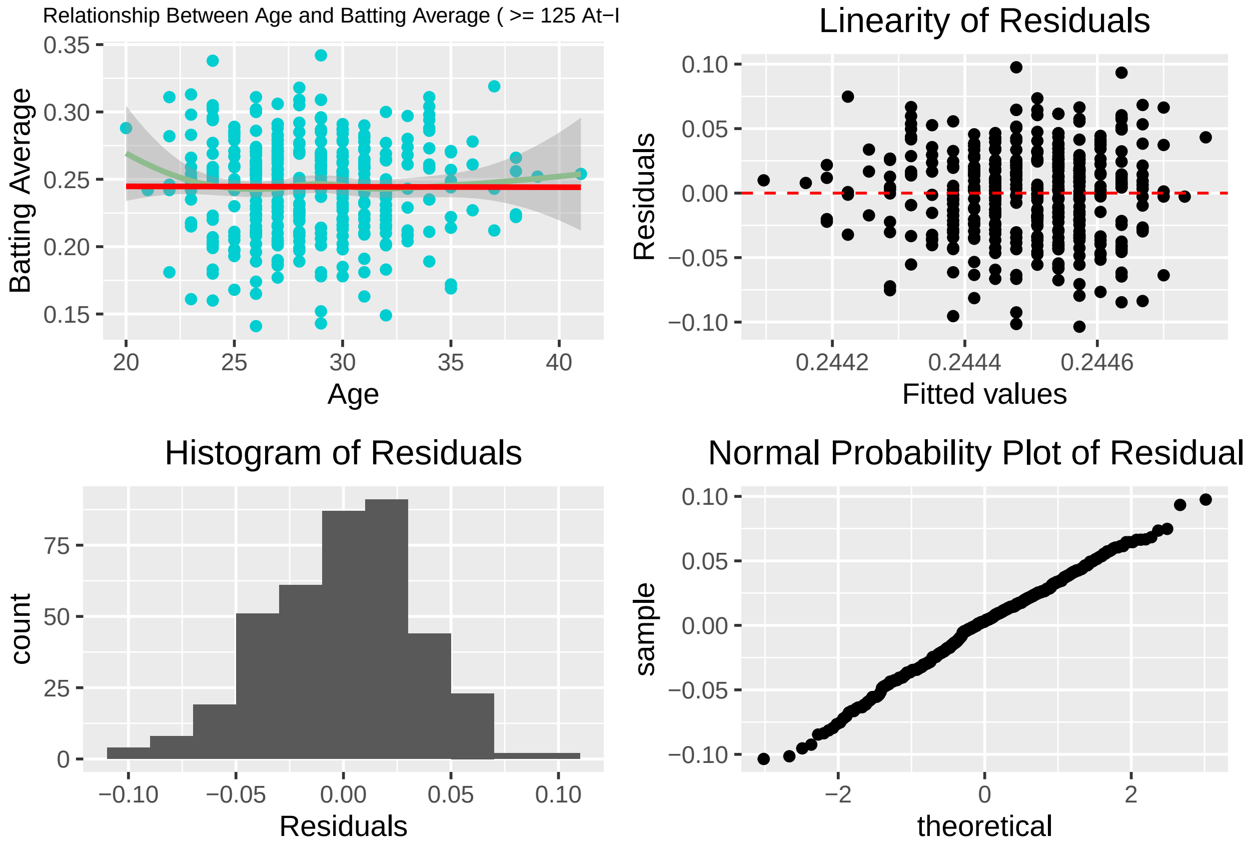
Given age 24, the predicted batting average of 0.246 is an underestimate of .00584 It is expected that there will be some underestimates and overestimates given the spread of the distribution of averages for each age. The ages are not all unique from one another, so there will be a wide range of outcomes available for each age.

```r
sct <- ggplot(data = fit, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = 'red') +
  xlab("Fitted values") +
  ylab("Residuals") +
  ggtitle("Linearity of Residuals")+
  theme(plot.title = element_text(hjust = 0.5))

hst <- ggplot(data = fit, aes(x = .resid)) +
  geom_histogram(binwidth = 0.02) +
  xlab("Residuals") +
  ggtitle("Histogram of Residuals")+
  theme(plot.title = element_text(hjust = 0.5))

npp <- ggplot(data = fit, aes(sample = .resid)) +
  stat_qq()+
  ggtitle("Normal Probability Plot of Residuals")+
  theme(plot.title = element_text(hjust = 0.5))

lin_analysis <- ggarrange(agebat, sct, hst, npp,
                      ncol = 2, nrow = 2)
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

```
lin_analysis
```



From the above visualizations, it can be concluded that a linear model is appropriate for measuring the relationship between age and batting average because the residuals show constant variability, the residuals are nearly normal, and the scatterplot of the data appears linear.

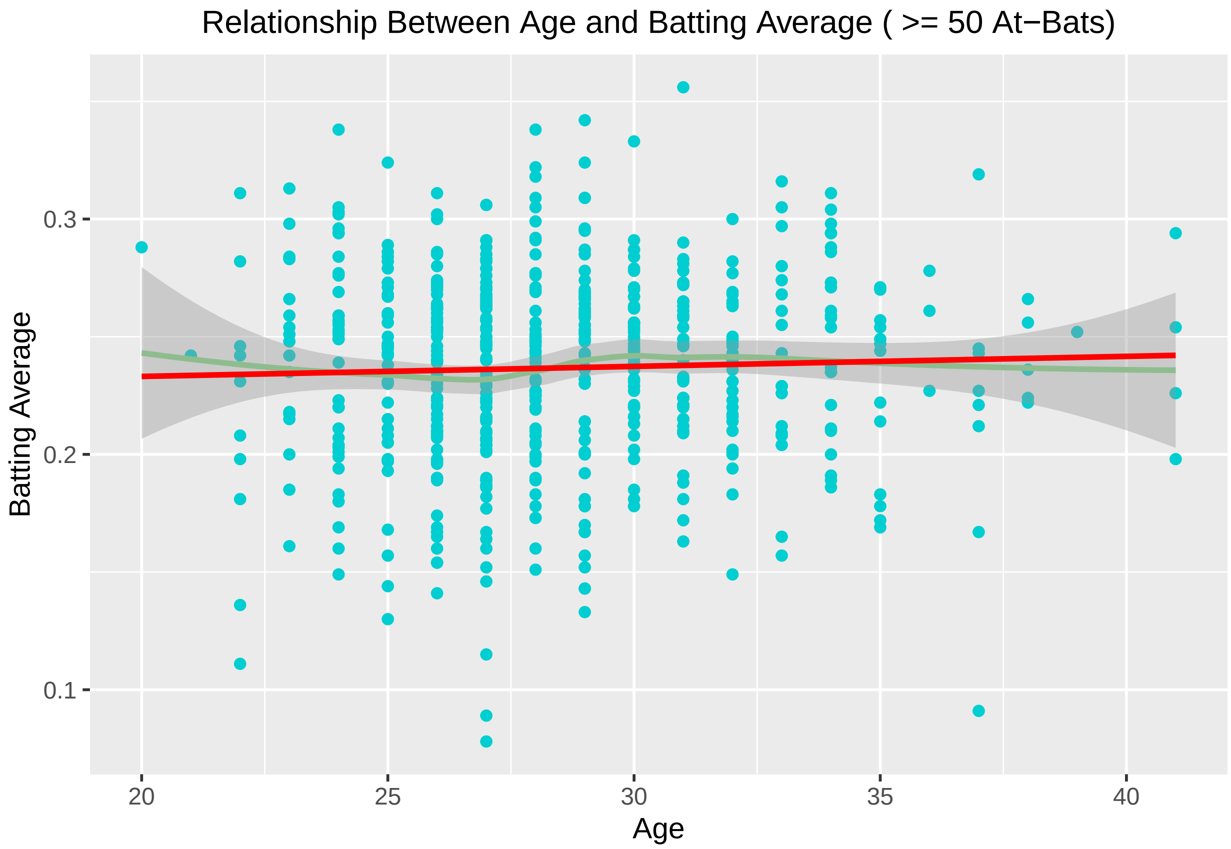
Thus, a linear model is appropriate and the equation for the regression line is $AVG = 0.2454 + 0.00003178 * Age$. A 22 year old is expected to have a batting average of 0.24610, and a 31 year old is expected to have a batting average of 0.24639. Over the course of an entire season (if both players had the maximum number of at-bats for the season, 664), this difference would amount to 0.18992 additional hits over the entire season.

Overall, age gives measurably no advantage in batting average for players with greater than 125 at-bats.

Greater Than or Equal to 50 At-Bats

```
agebat50 <- ggplot(mlb_50, aes(x = Age, y = AVG, na.rm = TRUE))+
  geom_point(na.rm = TRUE, color = 'darkturquoise')+
  geom_smooth(color = 'darkseagreen')+
  geom_smooth(method = "lm", color = 'red', se = FALSE)+
  ggtitle('Relationship Between Age and Batting Average ( >= 50 At-Bats)')+
  xlab('Age')+
  ylab('Batting Average')+
  theme(plot.title = element_text(hjust = 0.5, size = 12))

agebat50
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

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



Regression Analysis

```
fit50 <- lm(AVG ~ Age, data = mlb_50)
summary(fit50)
```

```
##
## Call:
## lm(formula = AVG ~ Age, data = mlb_50)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.158109 -0.026752  0.006391  0.027998  0.118185
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2245924  0.0143195  15.684   <2e-16 ***
## Age         0.0004265  0.0004973   0.858    0.391
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.04212 on 554 degrees of freedom
## Multiple R-squared:  0.001326,   Adjusted R-squared:  -0.0004764
## F-statistic: 0.7357 on 1 and 554 DF,  p-value: 0.3914
```

```
set.seed(29)
# x <- runif(1, min(mlb$Age), max(mlb$Age)) -- If age were not in integer form
x50 <- sample(min(mlb_50$Age):max(mlb_50$Age), 1)
eq50 <- .0004265*x50 + .2245924

paste("Predicted Batting Average for a given Age ", round(x50, 0), " : ", round(eq50, 3))
```

```
## [1] "Predicted Batting Average for a given Age  24  :  0.235"
```

```
idx50 <- as.numeric(which.min(abs(x50 - mlb_50$Age)))
actual50 <- mlb_50$AVG[idx50]
actual50
```

```
## [1] 0.259
```

```
paste("Actual Sample Batting Average for a given Age ", mlb_50$Age[idx50], " : ", round(actual50, 3))
```

```
## [1] "Actual Sample Batting Average for a given Age  24  :  0.259"
```

```
resid50 <- eq50 - actual50

paste("Residual: ", round(resid50, 5))
```

```
## [1] "Residual:  -0.02417"
```

Given age 24, the predicted batting average of 0.235 is an underestimate of .02417 It is expected that there will be some underestimates and overestimates given the spread of the distribution of averages for each age. The ages are not all unique from one another, so there will be a wide range of outcomes available for each age.

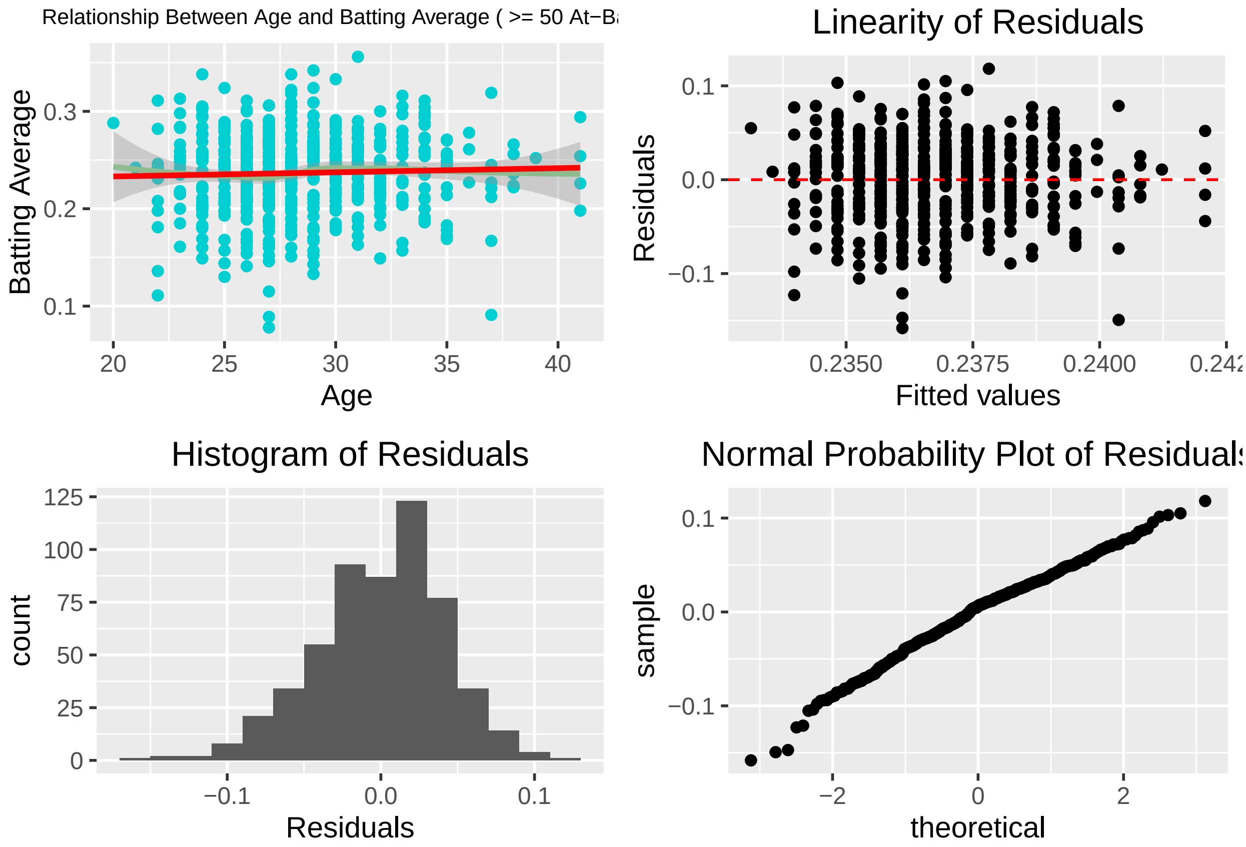
```
agebat50 <- ggplot(mlb_50, aes(x = Age, y = AVG, na.rm = TRUE))+
  geom_point(na.rm = TRUE, color = 'darkturquoise')+
  geom_smooth(color = 'darkseagreen')+
  geom_smooth(method = "lm", color = 'red', se = FALSE)+
  ggtitle('Relationship Between Age and Batting Average ( >= 50 At-Bats)')+
  xlab('Age')+
  ylab('Batting Average')+
  theme(plot.title = element_text(hjust = 0.5, size = 7.5))
```

```
sct50 <- ggplot(data = fit50, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = 'red') +
  xlab("Fitted values") +
  ylab("Residuals") +
  ggtitle("Linearity of Residuals")+
  theme(plot.title = element_text(hjust = 0.5))

hst50 <- ggplot(data = fit50, aes(x = .resid)) +
  geom_histogram(binwidth = 0.02) +
  xlab("Residuals") +
  ggtitle("Histogram of Residuals")+
  theme(plot.title = element_text(hjust = 0.5))

npp50 <- ggplot(data = fit50, aes(sample = .resid)) +
  stat_qq()+
  ggtitle("Normal Probability Plot of Residuals")+
  theme(plot.title = element_text(hjust = 0.5))

lin_analysis50 <- ggarrange(agebat50, sct50, hst50, npp50,
                      ncol = 2, nrow = 2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

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

```
lin_analysis50
```



For greater than or equal to 50 At-Bats, it can be concluded that a linear model is appropriate for measuring the relationship between age and batting average because the residuals show constant variability, the residuals are nearly normal, and the scatterplot of the data appears linear.

Thus, a linear model is appropriate and the equation for the regression line is $AVG = 0.2245924 + 0.0004265 * Age$. This may be a negligible result. A 22 year old is expected to have a batting average of 0.234, and a 31 year old is expected to have a batting average of 0.238. Over the course of an entire season (if both players had the maximum number of at-bats for the season, 664), this difference would amount to 2.55 additional hits over the entire season.

Overall, an increase in age will lead to a very small increase in batting average for players with 50 or more at-bats.

References

"2021 Major League Baseball Standard Batting." Baseball, https://www.baseball-reference.com/leagues/majors/2021-standard-batting.shtml.

2021 MLB Player Stats, https://www.rotowire.com/baseball/stats.php.