MLB Batting Player Analysis

Isaac Miller, Cooper Riggs, Stockton Nelson

1 Group Members

The members in our group include Isaac Miller, Cooper Riggs, and Stockton Nelson.

2 Data Set

```
#Bringing in data
baseball <- read.csv("MLB_Player_Batting_2024.csv", header = T)
baseball <- baseball[-1, ]

#Showing the data
head(baseball)
```

```
  Rk            Player Age Team Lg WAR   G  PA  AB   R   H X2B X3B HR RBI SB CS
2  1     Jarren Duran*  27  BOS AL 8.7 160 735 671 111 191  48  14 21  75 34  7
3  2    Shohei Ohtani*  29  LAD NL 9.2 159 731 636 134 197  38   7 54 130 59  4
4  3 Gunnar Henderson*  23  BAL AL 9.1 159 719 630 118 177  31   7 37  92 21  4
5  4     Marcus Semien  33  TEX AL 4.1 159 718 650 101 154  27   2 23  74  8  3
6  5        Juan Soto*  25  NYY AL 7.9 157 713 576 128 166  31   4 41 109  7  4
7  6    Bobby Witt Jr.  24  KCR AL 9.4 161 709 636 125 211  45  11 32 109 31 12
   BB  SO    BA   OBP   SLG   OPS OPS.  rOBA Rbat.  TB GIDP HBP SH SF IBB
2  54 160 0.285 0.342 0.492 0.834  129 0.373   134 330    6   6  1  3   1
3  81 162 0.310 0.390 0.646 1.036  190 0.449   190 411    7   6  0  5  10
4  78 159 0.281 0.364 0.529 0.893  159 0.385   157 333    2   7  0  4   1
5  64 105 0.237 0.308 0.391 0.699  100 0.310   100 254    9   3  0  1   2
6 129 119 0.288 0.419 0.569 0.989  178 0.424   179 328   10   4  0  4   2
7  57 106 0.332 0.389 0.588 0.977  171 0.416   171 374    4   8  0  8   9
```

Description: This dataset comes from Baseball Reference, with a table that includes all of the batting statistics for the top 300 players, according to batting average, in the MLB for 2024 regular season. This list includes these variables among many others that measures success for batting for the all of these players.

Variable	Description
Player	Name of the Player
Age	Age in Years
Team	Team Abbreviation
Lg	League Abbreviation
WAR	Wins Above Replacement
G	Games Played
PA	Plate Appearances
AB	At Bats
R	Runs Scored
H	Hits
2B	Doubles
3B	Triples
HR	**Home Runs**
CS	Caught Stealing
BB	Bases on Balls
SO	Strikeouts
BA	**Batting Average**
OBP	On Base Percentage
SLG	Bases Reached per at Bat
OPS	On Base % and Bases Reached per at Bat
OPS+	OPS x 100 Adjusted to Environment
rOBA	Offensive Contributions
Rbat+	Hitting Ability Adjusted to Environment
TB	Total Bases Earned
HBP	Hit by Pitch
SH	Sacrifice Hits
SF	Sacrifice Flies
IBB	Intentional Walks

3 Potential Questions of Interest

Question 1: What factors are important in predicting the amount of home runs a player will get?

Question 2: What factors are important in predicting batting average a player will get, and does it differ from the factors for predicting home runs?

4 Exploratory Data Analysis

```
#Filtering out numeric data
baseball_numeric <- baseball |>
  select(where(is.numeric))

#Looking at a summary of the data
summary(baseball_numeric)
```

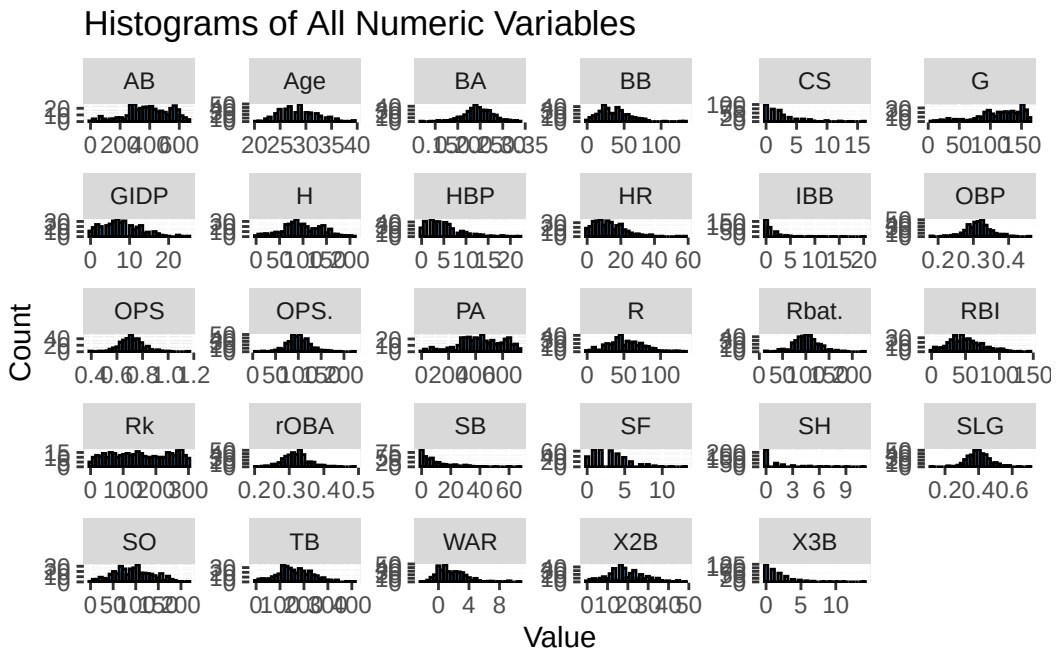
```
       Rk              Age             WAR               G         
 Min.   :  1.00   Min.   :20.0   Min.   :-2.000   Min.   :  5.00  
 1st Qu.: 77.75   1st Qu.:26.0   1st Qu.: 0.300   1st Qu.: 94.75  
 Median :150.50   Median :28.0   Median : 1.300   Median :121.00  
 Mean   :154.47   Mean   :28.5   Mean   : 1.706   Mean   :114.54  
 3rd Qu.:237.25   3rd Qu.:31.0   3rd Qu.: 2.725   3rd Qu.:145.00  
 Max.   :300.00   Max.   :39.0   Max.   :10.800   Max.   :162.00  
       PA              AB              R                H         
 Min.   : 12.0   Min.   : 11.0   Min.   :  1.00   Min.   :  3.00  
 1st Qu.:334.8   1st Qu.:298.8   1st Qu.: 37.00   1st Qu.: 73.75  
 Median :441.0   Median :396.5   Median : 49.00   Median : 95.00  
 Mean   :438.3   Mean   :393.0   Mean   : 52.28   Mean   : 98.29  
 3rd Qu.:571.0   3rd Qu.:512.2   3rd Qu.: 68.00   3rd Qu.:130.00  
 Max.   :735.0   Max.   :671.0   Max.   :134.00   Max.   :211.00  
      X2B             X3B               HR              RBI        
 Min.   : 0.00   Min.   : 0.000   Min.   : 0.00   Min.   :  0.00  
 1st Qu.:13.00   1st Qu.: 0.000   1st Qu.: 7.00   1st Qu.: 34.00  
 Median :18.00   Median : 1.000   Median :12.00   Median : 47.00  
 Mean   :19.16   Mean   : 1.703   Mean   :13.69   Mean   : 50.21  
 3rd Qu.:25.00   3rd Qu.: 2.000   3rd Qu.:19.00   3rd Qu.: 65.00  
 Max.   :48.00   Max.   :14.000   Max.   :58.00   Max.   :144.00  
       SB              CS               BB               SO        
 Min.   : 0.000   Min.   : 0.000   Min.   :  1.00   Min.   :  2.00  
 1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.: 22.00   1st Qu.: 65.00  
 Median : 5.000   Median : 2.000   Median : 34.00   Median : 95.00  
 Mean   : 8.904   Mean   : 2.352   Mean   : 36.17   Mean   : 94.99  
 3rd Qu.:12.000   3rd Qu.: 3.000   3rd Qu.: 47.25   3rd Qu.:123.00  
 Max.   :67.000   Max.   :16.000   Max.   :133.00   Max.   :218.00  
```

```
       BA              OBP              SLG              OPS        
 Min.   :0.1200   Min.   :0.1810   Min.   :0.1200   Min.   :0.3970  
 1st Qu.:0.2280   1st Qu.:0.2928   1st Qu.:0.3600   1st Qu.:0.6567  
 Median :0.2460   Median :0.3140   Median :0.3950   Median :0.7105  
 Mean   :0.2461   Mean   :0.3146   Mean   :0.4011   Mean   :0.7157  
 3rd Qu.:0.2660   3rd Qu.:0.3350   3rd Qu.:0.4375   3rd Qu.:0.7642  
 Max.   :0.3370   Max.   :0.4580   Max.   :0.7010   Max.   :1.1590  
      OPS.            rOBA            Rbat.             TB        
 Min.   :  9.0   Min.   :0.2020   Min.   : 13.0   Min.   :  3.0  
 1st Qu.: 85.0   1st Qu.:0.2948   1st Qu.: 85.0   1st Qu.:112.0  
 Median :100.0   Median :0.3180   Median :101.0   Median :154.0  
 Mean   :101.4   Mean   :0.3186   Mean   :101.8   Mean   :161.9  
 3rd Qu.:116.0   3rd Qu.:0.3390   3rd Qu.:117.0   3rd Qu.:213.0  
 Max.   :223.0   Max.   :0.4840   Max.   :223.0   Max.   :411.0  
      GIDP             HBP               SH                SF        
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000  
 1st Qu.: 4.000   1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.: 1.000  
 Median : 7.000   Median : 4.000   Median : 0.0000   Median : 3.000  
 Mean   : 7.898   Mean   : 5.036   Mean   : 0.8929   Mean   : 2.992  
 3rd Qu.:11.000   3rd Qu.: 7.000   3rd Qu.: 1.0000   3rd Qu.: 4.000  
 Max.   :25.000   Max.   :22.000   Max.   :11.0000   Max.   :13.000  
      IBB        
 Min.   : 0.000  
 1st Qu.: 0.000  
 Median : 1.000  
 Mean   : 1.354  
 3rd Qu.: 2.000  
 Max.   :20.000  
```

Highlights: From these summaries, we learn about the spread of each of our variables. It is interesting to note that home runs are very right skewed and WAR goes as low as -2 and up to nearly 11. Batting average is also pretty normally distributed, with the best of players getting a batting average of 0.300 or more.

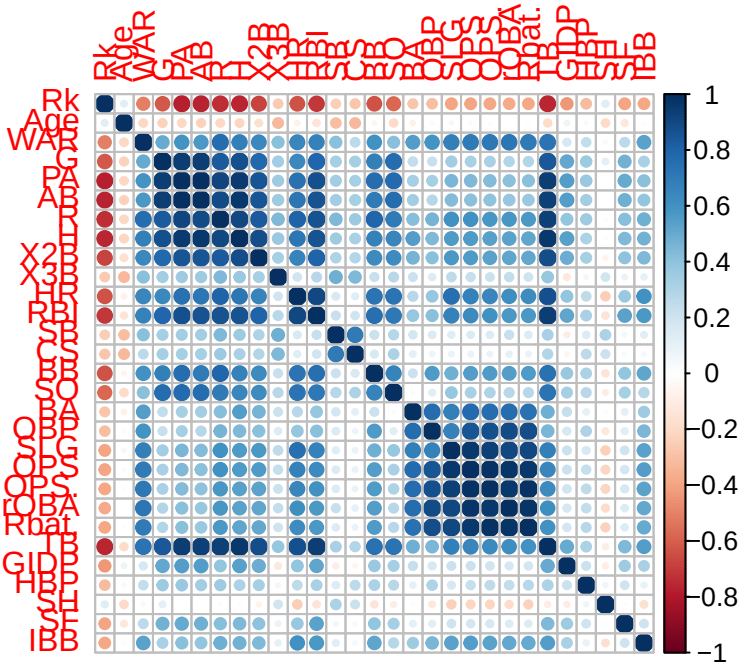
```
#Adjusting data to plot multiple histograms
baseball_long <- baseball_numeric |>
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value")

#Looking at histograms of the variables
ggplot(baseball_long, aes(x = Value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
  facet_wrap(~Variable, scales = "free") +  # Separate histogram for each variable
  labs(title = "Histograms of All Numeric Variables", x = "Value", y = "Count")
```



Highlights: From these histograms, we get the same data from the summaries, but now in a visual form. It is interesting to note that all of the batting percentages are pretty similar in spread. We can also see that plate appearances have most people around at least 200 appearances.

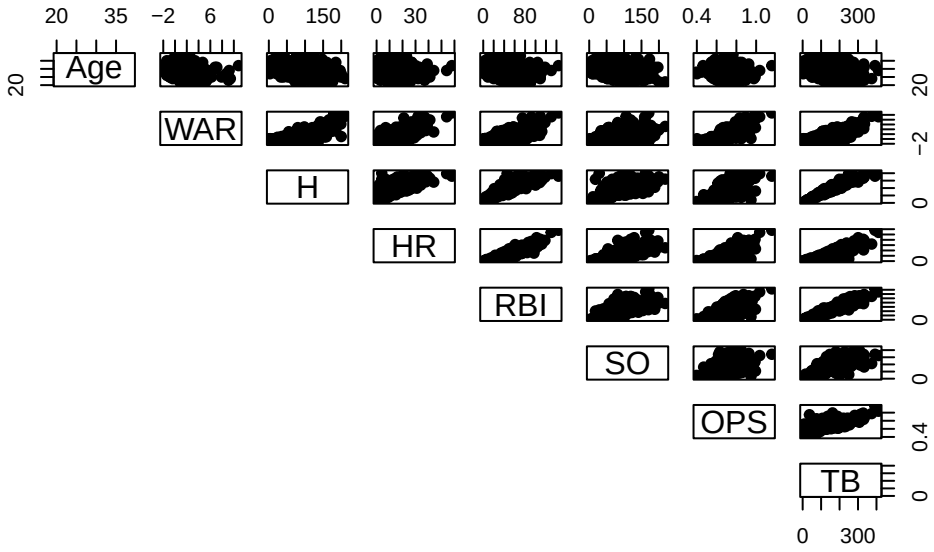
```
#Making a correlation plot
corrplot(cor(baseball_numeric))
```



Highlights: This correlation matrix shows us the correlation between each of our variables with each other. There are two areas that have strong correlation with each other that is particularly interesting to us. First, like we mentioned before, all of the batting percentages are pretty similar to each other. Second, how many times a player got to bat is similar with their games played and the amount of hits they got.

```
#Selecting variables of interest
baseball_interest <- baseball_numeric |>
  select(Age, WAR, H, HR, RBI, SO, OPS, TB)

#Making a scatterplot matrix
pairs(baseball_interest, pch = 19, lower.panel = NULL)
```



Highlights: This scatterplot matrix shows us the two way relationships between each of our selected variables of interest. The interesting things we see is that age doesn't seem to describe how good a player is at hitting. Also, total bases seem to be a pretty good predictor for all of these other variables, which is good for prediction, but may be a concern for multicollinearity.

5 Preliminary Model Fit

```
baseball_lm <- lm(HR ~ WAR, data = baseball)
```

Justification: Wins above replacement is an overall statistic that describes how valuable a baseball player is to their team. Using this to predict how many home runs a player has makes sense since a player who often hits home runs will score a lot of runs and be more valuable to their team.

```
#Making a summary of the linear regression
summary(baseball_lm)
```

```
Call:
lm(formula = HR ~ WAR, data = baseball)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.4638  -5.0638  -0.8654   4.2230  26.4675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.1983     0.4939    16.6   <2e-16 ***
WAR           3.2187     0.1927    16.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.032 on 362 degrees of freedom
Multiple R-squared:  0.4352,    Adjusted R-squared:  0.4337
F-statistic:   279 on 1 and 362 DF,  p-value: < 2.2e-16
```

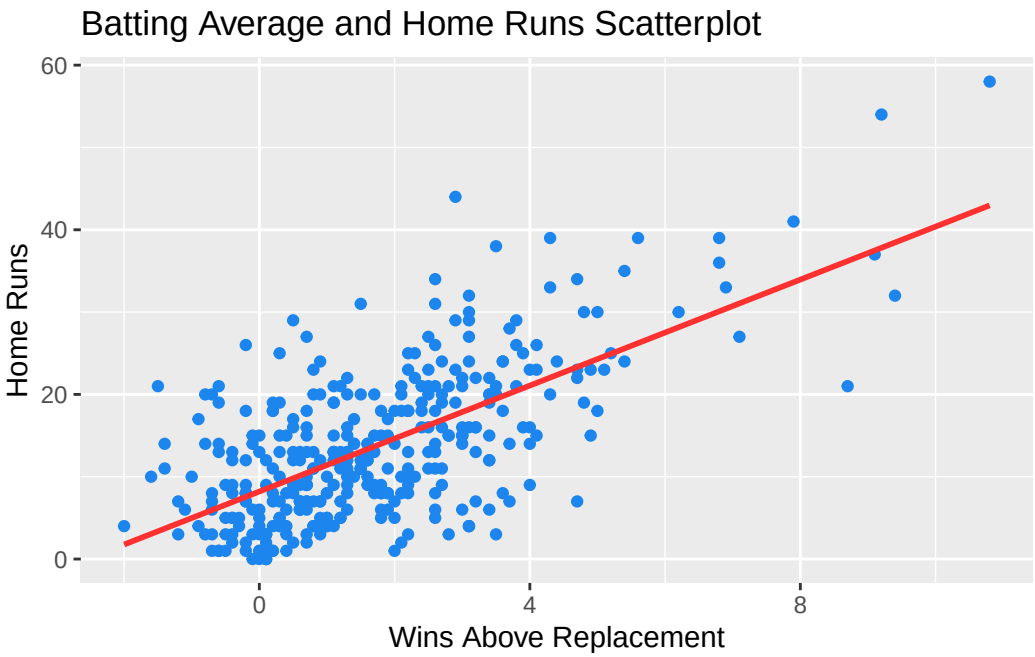
Interpretation: For every one unit increase in wins above replacement, the model predicts home runs to increase by 3.

Uncertainty: Since our p-value for wins above replacement is small, we know that it is a significant predictor of home runs.

6. (L) X vs Y is linear

```
#Scatterplot for Home runs vs Wins Above Replacement with the Regression Line
ggplot(data = baseball) +
  geom_point(mapping = aes(x = WAR, y = HR), color = "dodgerblue2") +
  geom_smooth(mapping = aes(x = WAR, y = HR), color = "firebrick1",
              method = "lm", se = F) +
  labs(x = "Wins Above Replacement",
       y = "Home Runs",
       title = "Batting Average and Home Runs Scatterplot")
```

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



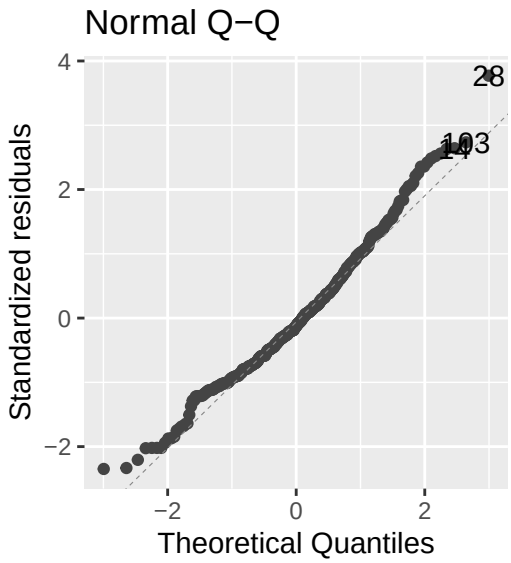
Looking at our scatterplot, there are a few points at the end which are a good ways above the regression line, but other than that, the relationship looks quite linear.

7. (I) The errors are independent

There may be some concerns about independence when doing this analysis since players have a lot of games and time to practice and get better as the season goes on.

8. (N) The errors are normally distributed

```
#Making the QQ Plot
autoplot(baseball_lm, which = 2)  +
coord_equal()
```




```
#Doing a Shapiro Wilk Test
shapiro.test(baseball_lm$residuals)
```

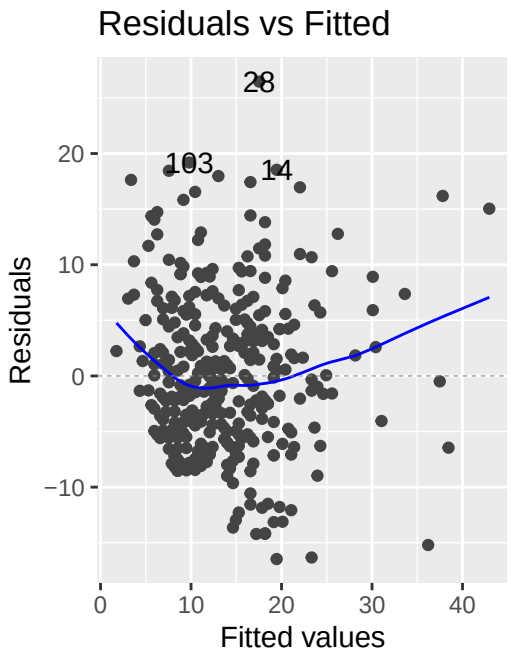
```
    Shapiro-Wilk normality test

data:  baseball_lm$residuals
W = 0.98143, p-value = 0.0001235
```

Looking at the QQ Plot, it appears that there is a slight right skew to our data. This is confirmed by our Shaprio Wilk test, giving us some reason to concern about the spread of the residuals. However, this all may not be too concerning since our dataset has 300 observations and the central limit theorem will help the residuals be more normal.

9. (E) The errors have equal (constant) variance across all values of X (homoscedastic)

```
#Making the fitted vs residuals plot
autoplot(baseball_lm, which = 1)
```



Looking at our fitted vs residual plot, the spread of the points seem pretty random, but it is interesting that there are higher positive residuals than negative residuals.