Regression Final Project

Uzair Ahmed

The goal of this project is to look at various player performance metrics and create a model that best predicts a players strikeout rate. This was done with data during the 2024 MLB season looking at qualified hitters, a sample of 129 players. We looked at SwStr%, OBP, SLG, SB, ZSwing%, OSwing%, Zone% and Position (C, INF, OF, DH).

Importing and Cleaning Data

```
library(tidyverse)
library(ggplot2)
library(olsrr)
library(dplyr)
library(lmtest)
library(corrplot)
```

First we imported all the requisite libraries for this project.

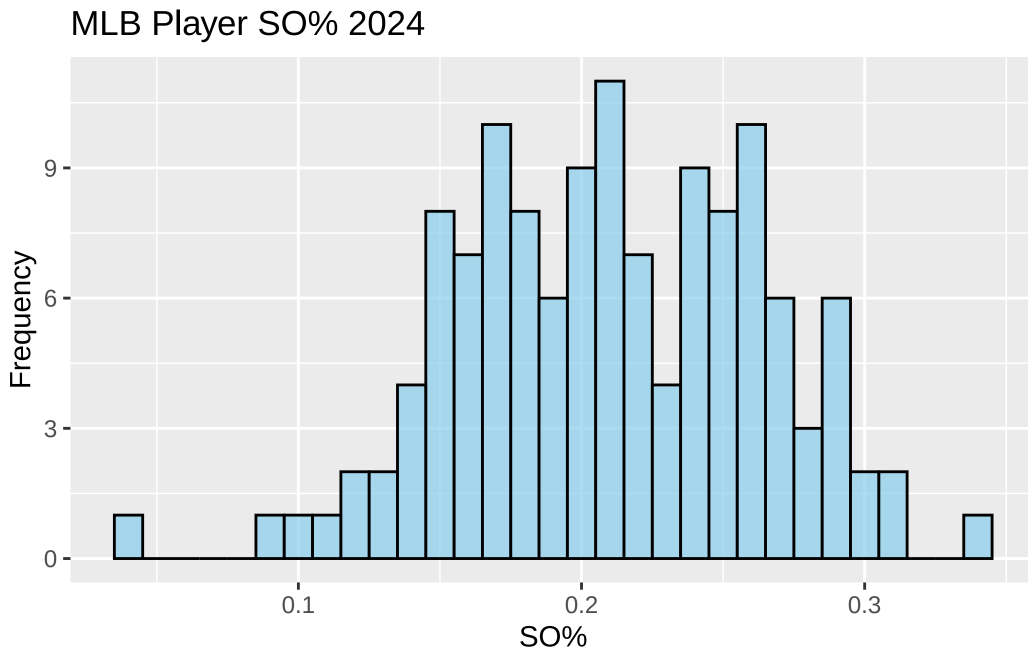
```
mlb1 <- read.csv("/Users/uzairahmed/Downloads/mlb1.csv")
mlb2 <- read.csv("/Users/uzairahmed/Downloads/mlb2.csv")
mlb3 <- merge(mlb1, mlb2, by = "Name")
mlb3 <- mlb3 %>% rename(K = K.)
mlb3 <- mlb3 %>% rename(BB = BB.)
mlb3 <- mlb3 %>% rename(OSwing = O.Swing.)
mlb3 <- mlb3 %>% rename(ZSwing = Z.Swing.)
mlb3 <- mlb3 %>% rename(Contact = Contact.)
mlb3 <- mlb3 %>% rename(Zone = Zone.)
mlb3 <- mlb3 %>% rename(SwStr = SwStr.)
mlb3$BB <- as.numeric(gsub("%", "", mlb3$BB))
mlb3$K  <- as.numeric(gsub("%", "", mlb3$K))
mlb3$OSwing <- as.numeric(gsub("%", "", mlb3$OSwing))
mlb3$ZSwing <- as.numeric(gsub("%", "", mlb3$ZSwing))
```

```
mlb3$Contact <- as.numeric(gsub("%", "", mlb3$Contact))
mlb3$Zone <- as.numeric(gsub("%", "", mlb3$Zone))
mlb3$SwStr <- as.numeric(gsub("%", "", mlb3$SwStr))

mlb3 <- mlb3 %>%
  mutate(
    BB = BB / 100,
    K = K / 100,
    OSwing = OSwing / 100,
    ZSwing = ZSwing / 100,
    Contact = Contact / 100,
    Zone = Zone / 100,
    SwStr = SwStr / 100
  )
```

Next, we imported the data into R. The data was exported from fangraphs into an excel sheet which was then imported into R. MLB1 contained the base performance stats we were planning on using while MLB2 contained the plate discipline stats we were using. We renamed and converted the variables to ensure a functional data set. To keep all our variables aligned as decimals we converted the factors that were listed as percents into decimals.

Statistical Summaries

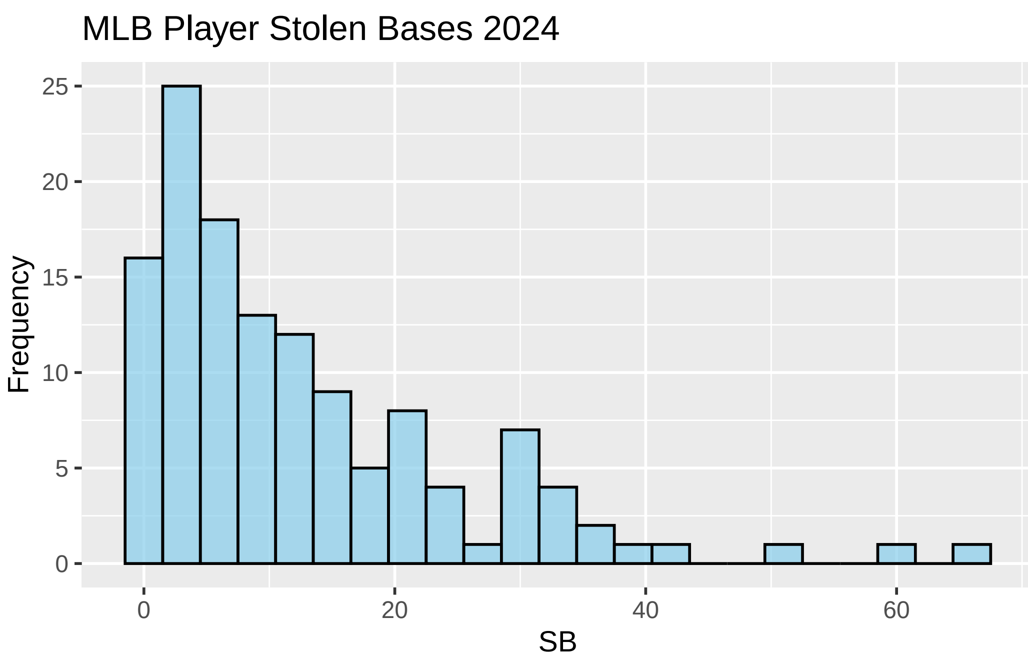
```
ggplot(mlb3, aes(x=K)) +
  geom_histogram(binwidth=0.01, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player SO% 2024", x="SO%", y="Frequency")
```




```r
summary(mlb3$K)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0430  0.1700  0.2110  0.2106  0.2510  0.3440
```

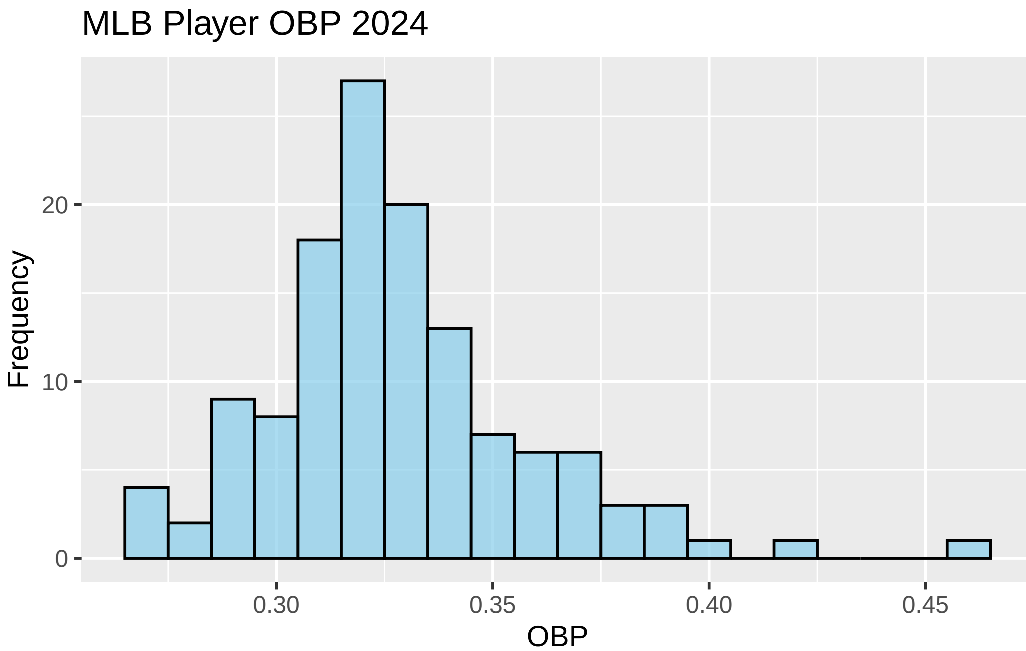
```r
ggplot(mlb3, aes(x=SB)) +
  geom_histogram(binwidth=3, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player Stolen Bases 2024", x="SB", y="Frequency")
```




```
summary(mlb3$SB)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    3.00    9.00   12.74   19.00   67.00
```

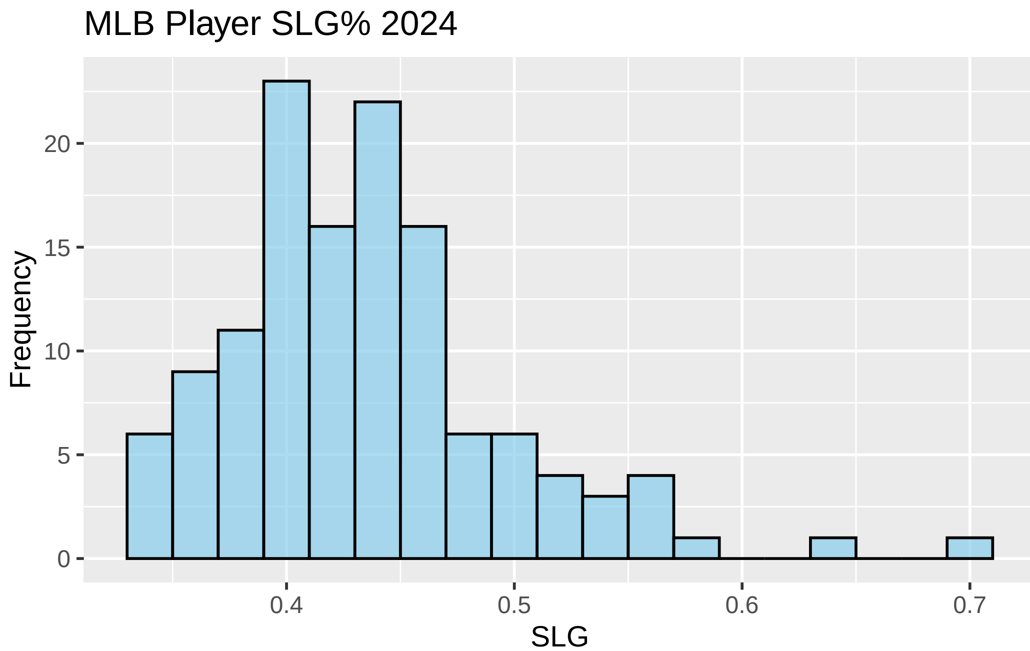
```
ggplot(mlb3, aes(x=OBP)) +
  geom_histogram(binwidth=0.01, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player OBP 2024", x="OBP", y="Frequency")
```




```r
summary(mlb3$OBP)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2700  0.3120  0.3250  0.3288  0.3420  0.4580 
```

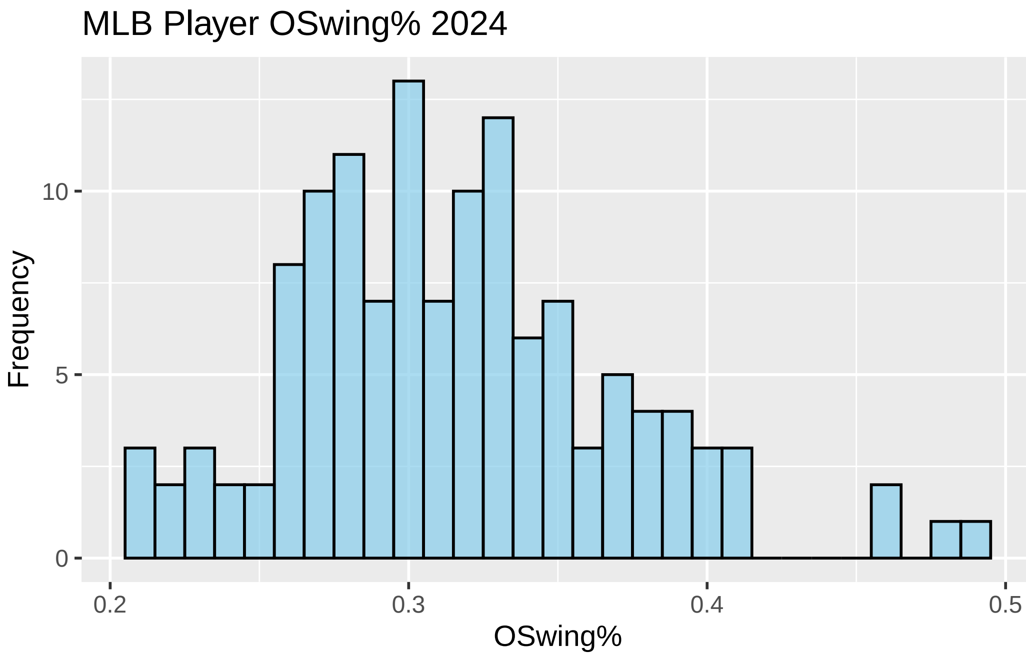
```r
ggplot(mlb3, aes(x=SLG)) +
  geom_histogram(binwidth=0.02, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player SLG% 2024", x="SLG", y="Frequency")
```




```r
summary(mlb3$SLG)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3310  0.3940  0.4280  0.4359  0.4640  0.7010 
```

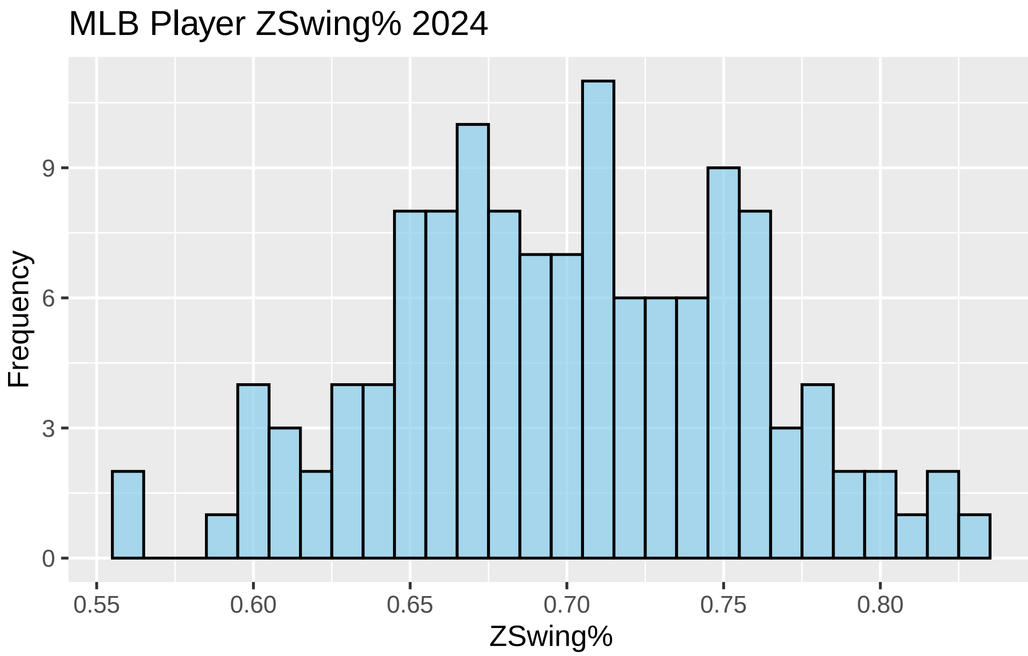
```r
ggplot(mlb3, aes(x=OSwing)) +
  geom_histogram(binwidth=0.01, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player OSwing% 2024", x="OSwing%", y="Frequency")
```




```
summary(mlb3$OSwing)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.213   0.277   0.312   0.316   0.348   0.495 
```

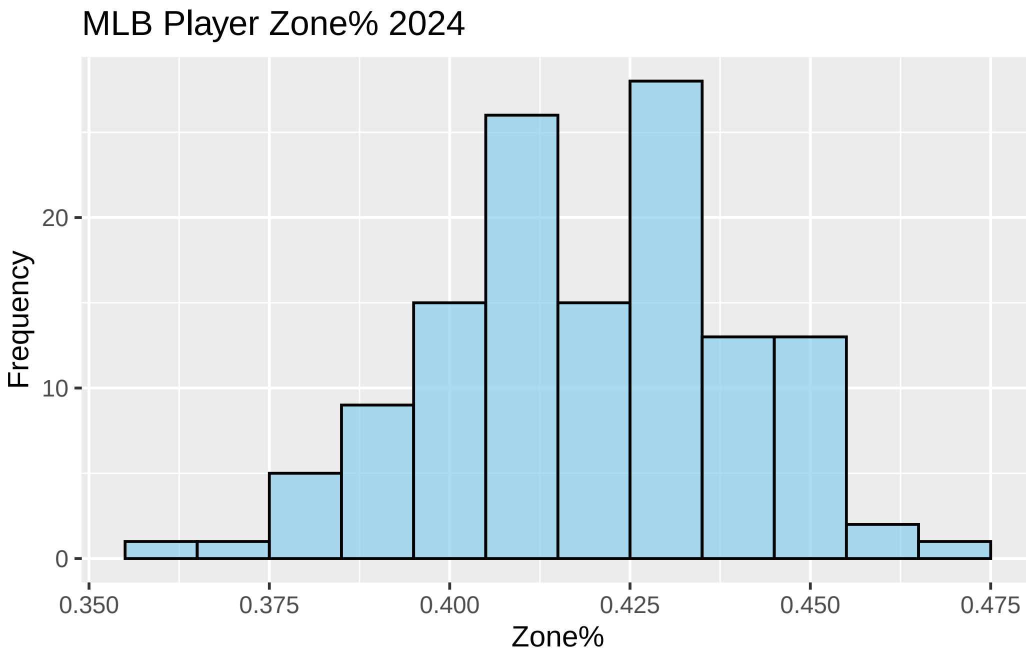
```
ggplot(mlb3, aes(x=ZSwing)) +
  geom_histogram(binwidth=0.01, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player ZSwing% 2024", x="ZSwing%", y="Frequency")
```




```r
summary(mlb3$ZSwing)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.5550  0.6630  0.6980  0.7001  0.7450  0.8260 
```

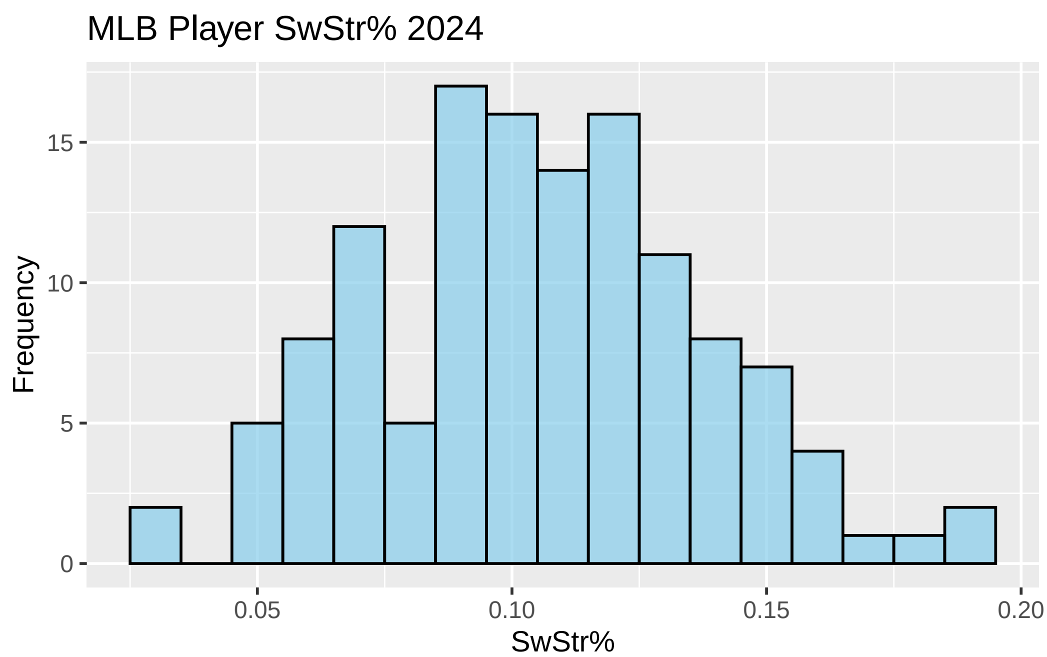
```r
ggplot(mlb3, aes(x=Zone)) +
  geom_histogram(binwidth=0.01, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player Zone% 2024", x="Zone%", y="Frequency")
```




```r
summary(mlb3$Zone)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3640  0.4060  0.4200  0.4198  0.4350  0.4670 
```

```r
ggplot(mlb3, aes(x=SwStr)) +
  geom_histogram(binwidth=0.01, fill="skyblue", color="black", alpha=0.7) +
  labs(title="MLB Player SwStr% 2024", x="SwStr%", y="Frequency")
```




```
summary(mlb3$SwStr)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0280  0.0860  0.1050  0.1056  0.1280  0.1920
```

We looked at a distribution for all our quantitative variables and looked at the 5 number summaries for each.

SLR Relationships

```
m1 <- lm(K~SB, data=mlb3)
summary(m1)
```

```
Call:
lm(formula = K ~ SB, data = mlb3)

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

```
-0.167427 -0.040481  0.000734  0.039928  0.132713

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.099e-01  6.672e-03  31.466   <2e-16 ***
SB          5.376e-05  3.730e-04   0.144    0.886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05321 on 127 degrees of freedom
Multiple R-squared:  0.0001636, Adjusted R-squared:  -0.007709
F-statistic: 0.02078 on 1 and 127 DF,  p-value: 0.8856
```

```r
m2 <- lm(K~OBP, data=mlb3)
summary(m2)
```

```
Call:
lm(formula = K ~ OBP, data = mlb3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.160799 -0.036683 -0.004333  0.039026  0.109949

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.34153    0.05069   6.738 5.03e-10 ***
OBP         -0.39805    0.15351  -2.593   0.0106 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05186 on 127 degrees of freedom
Multiple R-squared:  0.05028,   Adjusted R-squared:  0.0428
F-statistic: 6.723 on 1 and 127 DF,  p-value: 0.01063
```

```r
m3 <- lm(K~SLG, data=mlb3)
summary(m3)
```

```
Call:
lm(formula = K ~ SLG, data = mlb3)
```

```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.165440 -0.037881  0.000617  0.039073  0.137055 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.18891    0.03335   5.664 9.39e-08 ***
SLG          0.04982    0.07575   0.658    0.512    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05313 on 127 degrees of freedom
Multiple R-squared:  0.003394,  Adjusted R-squared:  -0.004453 
F-statistic: 0.4325 on 1 and 127 DF,  p-value: 0.512
```

```r
m4 <- lm(K~OSwing, data=mlb3)
summary(m4)
```

```
Call:
lm(formula = K ~ OSwing, data = mlb3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16935 -0.04066 -0.00001  0.04041  0.13298 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.20017    0.02742   7.300 2.78e-11 ***
OSwing       0.03310    0.08549   0.387    0.699    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05319 on 127 degrees of freedom
Multiple R-squared:  0.001179,  Adjusted R-squared:  -0.006685 
F-statistic: 0.1499 on 1 and 127 DF,  p-value: 0.6992
```

```r
m5 <- lm(K~ZSwing, data=mlb3)
summary(m5)
```

```
Call:
lm(formula = K ~ ZSwing, data = mlb3)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.159452 -0.036157  0.001253  0.039853  0.126446 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.10274    0.05729   1.793   0.0753 .
ZSwing       0.15411    0.08157   1.889   0.0612 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05248 on 127 degrees of freedom
Multiple R-squared:  0.02733,	Adjusted R-squared:  0.01967 
F-statistic: 3.569 on 1 and 127 DF,  p-value: 0.06115
```

```r
m6 <- lm(K~Zone, data=mlb3)
summary(m6)
```

```
Call:
lm(formula = K ~ Zone, data = mlb3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.14839 -0.03825  0.00075  0.03713  0.14681 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.4541     0.0933   4.867 3.29e-06 ***
Zone         -0.5799     0.2220  -2.613   0.0101 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05184 on 127 degrees of freedom
Multiple R-squared:  0.05101,	Adjusted R-squared:  0.04354 
F-statistic: 6.826 on 1 and 127 DF,  p-value: 0.01007
```

```r
m7 <- lm(K~SwStr, data=mlb3)
summary(m7)
```

```
Call:
lm(formula = K ~ SwStr, data = mlb3)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.073536 -0.023419 -0.001439  0.020386  0.086721 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.08092    0.01057   7.658 4.18e-12 ***
SwStr        1.22815    0.09567  12.838  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03511 on 127 degrees of freedom
Multiple R-squared:  0.5648,    Adjusted R-squared:  0.5613 
F-statistic: 164.8 on 1 and 127 DF,  p-value: < 2.2e-16
```

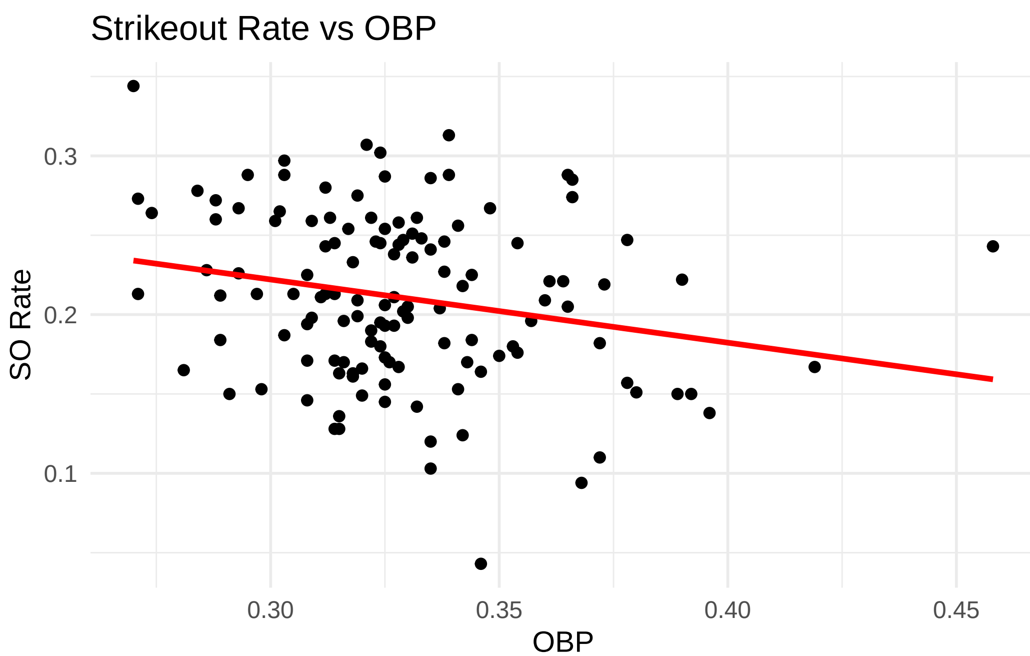
```r
m8 <- lm(K~Position, data=mlb3)
summary(m8)
```

```
Call:
lm(formula = K ~ Position, data = mlb3)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.160576 -0.036576 -0.001576  0.041424  0.140424 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.222889   0.017639  12.636   <2e-16 ***
PositionDH   0.011111   0.025714   0.432    0.666    
PositionINF -0.019313   0.018804  -1.027    0.306    
PositionOF  -0.008606   0.019288  -0.446    0.656    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.05292 on 125 degrees of freedom
Multiple R-squared:  0.02675,	Adjusted R-squared:  0.003389 
F-statistic: 1.145 on 3 and 125 DF,  p-value: 0.3336
```

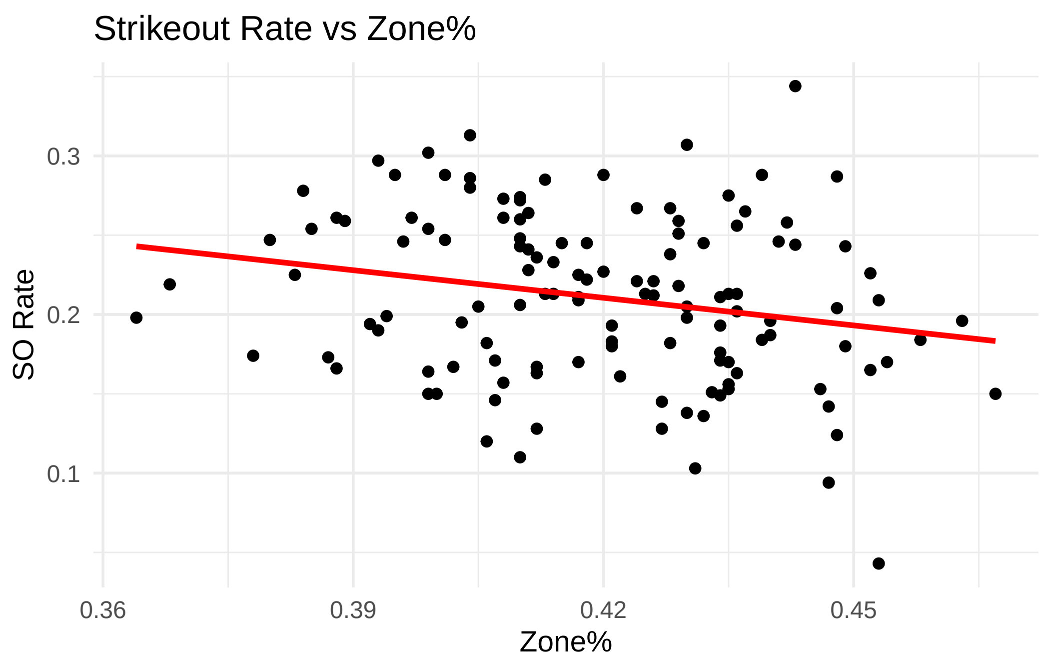
```r
ggplot(mlb3, aes(x = OBP, y = K)) +
  geom_point() +
  labs(title = "Strikeout Rate vs OBP", x = "OBP", y = "SO Rate") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

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



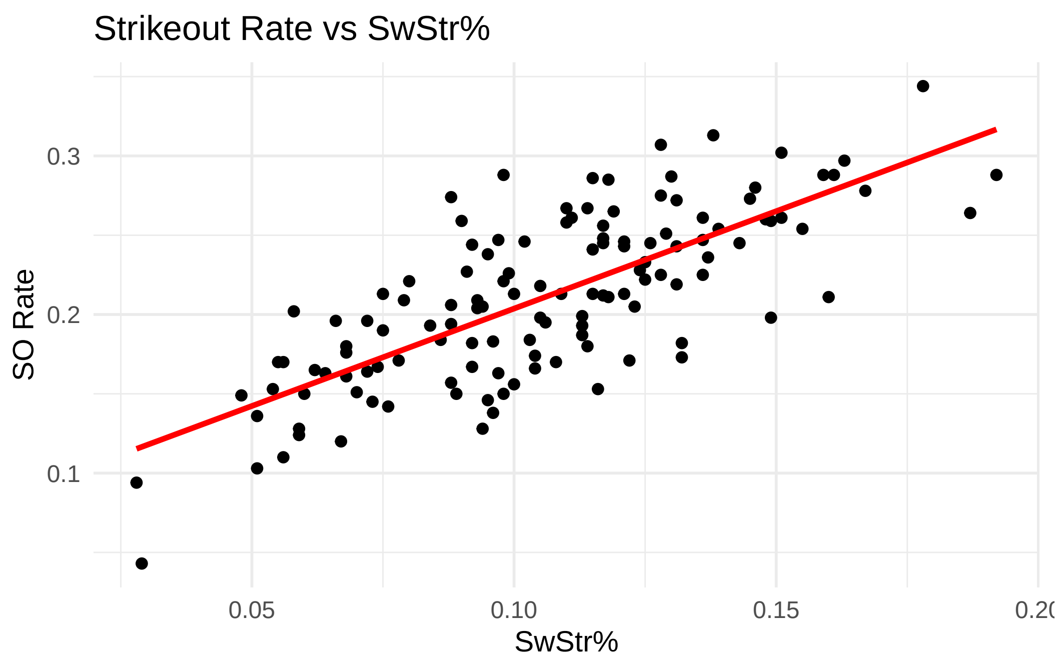

```r
ggplot(mlb3, aes(x = Zone, y = K)) +
  geom_point() +
  labs(title = "Strikeout Rate vs Zone%", x = "Zone%", y = "SO Rate") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

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




```
ggplot(mlb3, aes(x = SwStr, y = K)) +
  geom_point() +
  labs(title = "Strikeout Rate vs SwStr%", x = "SwStr%", y = "SO Rate") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

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



Looking at the SLR relationships only 3 predictors were significant at the 0.05 level so I graphed them vs the response variables to visualize the relationships even further.

MLR Relationships

```r
mlb_fit <- lm(K ~ SB + OBP + SLG + OSwing + ZSwing + Zone + SwStr + Position, data = mlb3)
summary(mlb_fit)
```

```
Call:
lm(formula = K ~ SB + OBP + SLG + OSwing + ZSwing + Zone + SwStr +
    Position, data = mlb3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05178 -0.01204  0.00101  0.01124  0.04846

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5115141  0.0724435   7.061 1.23e-10 ***
SB          -0.0001609  0.0001550  -1.038 0.301181
```

```
OBP         -0.4056790  0.1129152  -3.593 0.000478 ***
SLG          0.0911770  0.0508453   1.793 0.075498 .
OSwing      -0.4339857  0.0496577  -8.740 1.88e-14 ***
ZSwing      -0.3874189  0.0492467  -7.867 1.93e-12 ***
Zone         0.0003865  0.1148862   0.003 0.997321
SwStr        1.9513585  0.0872662  22.361  < 2e-16 ***
PositionDH  -0.0064275  0.0104396  -0.616 0.539291
PositionINF -0.0014255  0.0076081  -0.187 0.851691
PositionOF  -0.0054177  0.0078485  -0.690 0.491364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0202 on 118 degrees of freedom
Multiple R-squared:  0.8661,    Adjusted R-squared:  0.8547
F-statistic: 76.31 on 10 and 118 DF,  p-value: < 2.2e-16
```

```
ols_vif_tol(mlb_fit)
```

```
     Variables Tolerance      VIF
1           SB 0.8350579 1.197522
2          OBP 0.2805324 3.564650
3          SLG 0.3210244 3.115028
4       OSwing 0.4277256 2.337948
5       ZSwing 0.4066127 2.459343
6         Zone 0.5669584 1.763798
7        SwStr 0.3980385 2.512320
8   PositionDH 0.4991441 2.003429
9  PositionINF 0.2187935 4.570520
10  PositionOF 0.2239077 4.466126
```

```
ols_step_best_subset(mlb_fit)
```

```
                  Best Subsets Regression
-----------------------------------------------------------
Model Index    Predictors
-----------------------------------------------------------
     1         SwStr
     2         ZSwing SwStr
     3         OSwing ZSwing SwStr
     4         OBP OSwing ZSwing SwStr
     5         OBP SLG OSwing ZSwing SwStr
```

```
     6         OBP SLG OSwing ZSwing SwStr Position
     7         SB OBP SLG OSwing ZSwing SwStr Position
     8         SB OBP SLG OSwing ZSwing Zone SwStr Position
-----------------------------------------------------------

                                                     Subsets Regression Summary
-----------------------------------------------------------------------------------------------
                       Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)         AIC          SBIC          SBC
-----------------------------------------------------------------------------------------------
  1        0.5648      0.5613      0.5512    258.4648    -494.0578    -863.6790    -485.4783
  2        0.7553      0.7514      0.7439     92.6366    -566.3171    -935.1367    -554.8778
  3        0.8420      0.8382       0.833     18.1958    -620.7821    -987.4649    -606.4830
  4        0.8597      0.8551      0.8491      4.6568    -634.0519    -999.7097    -616.8930
  5        0.8629      0.8573      0.8505      3.7785    -635.0901   -1000.3687    -615.0714
  6        0.8648      0.8558      0.8459      8.1007    -630.8946    -999.8310    -602.2965
  7        0.8661      0.8559      0.8445      9.0000    -630.0924    -998.7044    -598.6345
  8        0.8661      0.8547      0.8419     11.0000    -628.0924    -996.5180    -593.7747
-----------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

```r
mlb_optimal <- lm(K ~ OBP + SLG + OSwing + ZSwing + SwStr, data = mlb3)
summary(mlb_optimal)
```

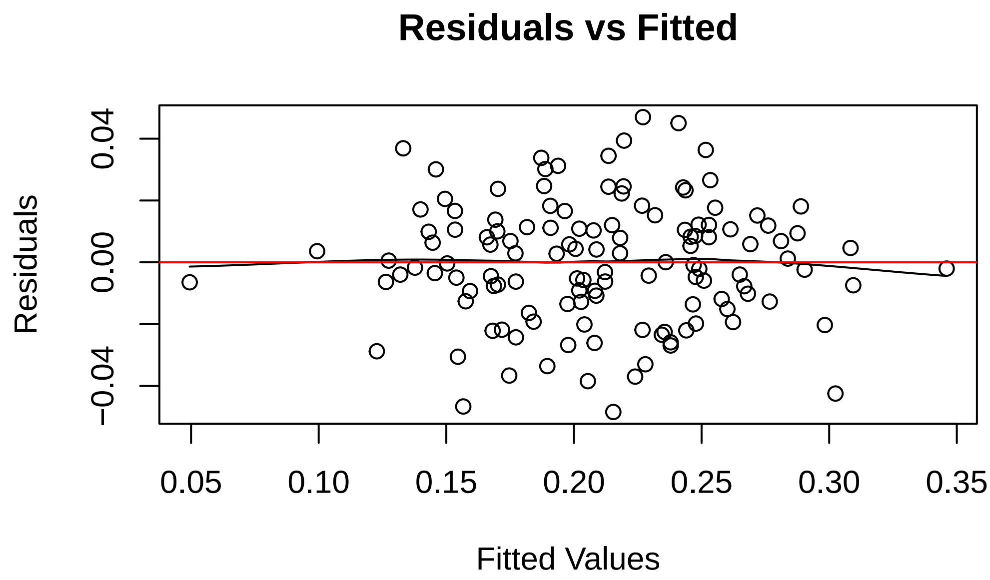
```
Call:
lm(formula = K ~ OBP + SLG + OSwing + ZSwing + SwStr, data = mlb3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.048428 -0.012604  0.000046  0.011887  0.046971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.49915    0.03595  13.883  < 2e-16 ***
OBP         -0.40686    0.10805  -3.766 0.000256 ***
```

```
SLG          0.08556    0.04998   1.712 0.089399 .
OSwing      -0.43031    0.04402  -9.776  < 2e-16 ***
ZSwing      -0.37099    0.04581  -8.099  4.6e-13 ***
SwStr        1.92849    0.07911  24.378  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02002 on 123 degrees of freedom
Multiple R-squared:  0.8629,    Adjusted R-squared:  0.8573
F-statistic: 154.9 on 5 and 123 DF,  p-value: < 2.2e-16
```

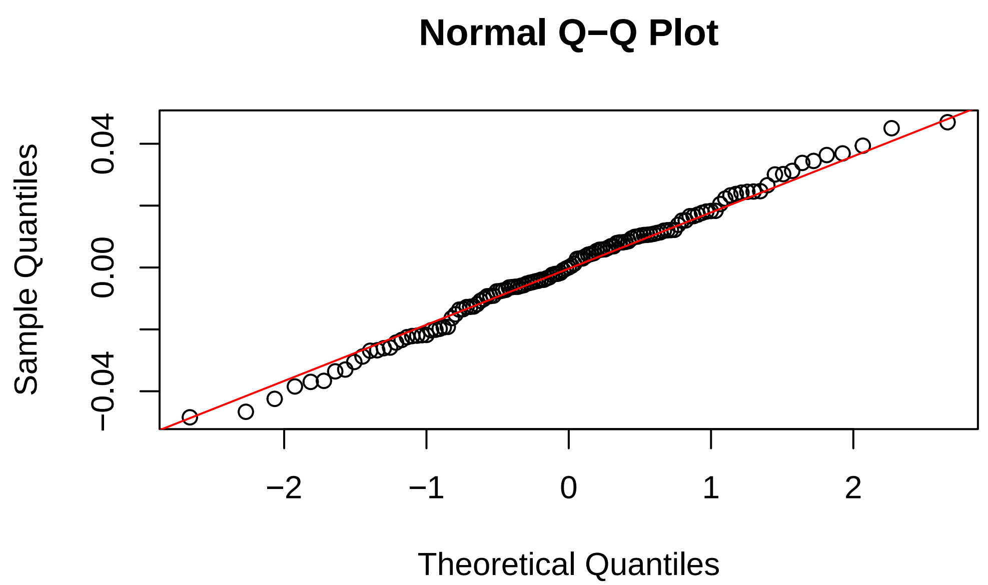
```r
ols_vif_tol(mlb_optimal)
```

```
  Variables Tolerance      VIF
1       OBP 0.3008411 3.324014
2       SLG 0.3262930 3.064730
3    OSwing 0.5345588 1.870702
4    ZSwing 0.4615111 2.166795
5     SwStr 0.4756285 2.102481
```

```r
scatter.smooth(mlb_optimal$fitted.values, mlb_optimal$residuals,
     main="Residuals vs Fitted",
     xlab="Fitted Values", ylab="Residuals")
abline(h = 0, col = "red")
```




```r
qqnorm(mlb_optimal$residuals)
qqline(mlb_optimal$residuals, col = "red")
```



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

```

    Shapiro-Wilk normality test

data:  residuals(mlb_optimal)
W = 0.99417, p-value = 0.8773
```

```r
bptest(mlb_optimal)
```

```

    studentized Breusch-Pagan test

data:  mlb_optimal
BP = 7.4765, df = 5, p-value = 0.1875
```

```r
MSE <- summary(mlb_optimal)$sigma^2
outlier_check <- round(data.frame(Residuals=mlb_optimal$residuals,
                                  "Standardized Res"=mlb_optimal$residuals/sqrt(MSE),
                                  "Studentized Res"=rstandard(mlb_optimal),
                                  "Press"=rstandard(mlb_optimal,type='predictive'),
                                  "R-student"=rstudent(mlb_optimal),
                                  "Hat-Values"=hatvalues(mlb_optimal)),2)
#outlier_check


influence <- round(data.frame(Cooks=cooks.distance(mlb_optimal),
                              dffits=dffits(mlb_optimal),
                              dfbeta=dfbetas(mlb_optimal),
                              cov_ratio=covratio(mlb_optimal)),3)
#influence


mlb_3 <- mlb3 %>%
  mutate(Cooks = cooks.distance(mlb_optimal)) %>%
  mutate(StuRes = rstandard(mlb_optimal)) %>%
  mutate(dffits = dffits(mlb_optimal)) %>%
  mutate(covratio = covratio(mlb_optimal))
#mlb_3
```

```r
filter(mlb_3, abs(StuRes) > 2)
```

```
                 Name HR SB    BB     K   AVG   OBP   SLG  WAR Position OSwing
1     Carlos Santana 23  4 0.109 0.167 0.238 0.328 0.420  3.0      INF  0.270
2 Christopher Morel 21  8 0.100 0.260 0.196 0.288 0.346 -1.0      INF  0.297
3          Juan Soto 41  7 0.181 0.167 0.288 0.419 0.569  8.1       OF  0.213
4      Michael Busch 21  2 0.111 0.286 0.248 0.335 0.440  2.3      INF  0.274
5       Mookie Betts 19 16 0.118 0.110 0.289 0.372 0.491  4.4       OF  0.229
6       Seiya Suzuki 21 16 0.108 0.274 0.283 0.366 0.482  3.6       OF  0.236
  ZSwing Contact  Zone SwStr      Cooks    StuRes     dffits  covratio
1  0.667   0.787 0.412 0.092 0.01440188 -2.436427 -0.3000916 0.7918695
2  0.719   0.684 0.410 0.148 0.05261083 -2.189317 -0.5707839 0.8818551
3  0.601   0.799 0.402 0.074 0.07717007 -2.026086 -0.6892841 0.9533154
4  0.710   0.745 0.404 0.115 0.01721517  2.271403  0.3270119 0.8283663
5  0.654   0.861 0.410 0.056 0.04176453 -2.380587 -0.5104458 0.8263205
6  0.627   0.777 0.410 0.088 0.03618640  2.390269  0.4752301 0.8194639
```

```r
filter(mlb_3, Cooks > 0.031)
```

```
                     Name HR SB    BB     K   AVG   OBP   SLG  WAR Position
1      Anthony Santander 44  2 0.087 0.194 0.235 0.308 0.506  3.3       OF
2      Christopher Morel 21  8 0.100 0.260 0.196 0.288 0.346 -1.0      INF
3        George Springer 19 16 0.098 0.187 0.220 0.303 0.371  1.2       OF
4              Juan Soto 41  7 0.181 0.167 0.288 0.419 0.569  8.1       OF
5           Mookie Betts 19 16 0.118 0.110 0.289 0.372 0.491  4.4       OF
6           Rhys Hoskins 26  3 0.103 0.288 0.214 0.303 0.419  0.1      INF
7           Seiya Suzuki 21 16 0.108 0.274 0.283 0.366 0.482  3.6       OF
8 Vladimir Guerrero Jr. 30  2 0.103 0.138 0.323 0.396 0.544  5.4      INF
  OSwing ZSwing Contact  Zone SwStr      Cooks    StuRes     dffits  covratio
1  0.375  0.688   0.822 0.392 0.088 0.03242748  1.257791  0.4421511 1.0912195
2  0.297  0.719   0.684 0.410 0.148 0.05261083 -2.189317 -0.5707839 0.8818551
3  0.284  0.753   0.769 0.440 0.113 0.03234752 -1.896058 -0.4453125 0.9264256
4  0.213  0.601   0.799 0.402 0.074 0.07717007 -2.026086 -0.6892841 0.9533154
5  0.229  0.654   0.861 0.410 0.056 0.04176453 -2.380587 -0.5104458 0.8263205
6  0.281  0.615   0.771 0.439 0.098 0.03188314  1.865198  0.4418897 0.9327265
7  0.236  0.627   0.777 0.410 0.088 0.03618640  2.390269  0.4752301 0.8194639
8  0.306  0.710   0.801 0.430 0.096 0.03198945 -1.879266 -0.4427234 0.9296988
```

```r
filter(mlb_3, Cooks > 0.031 & abs(dffits) > (2*sqrt(5/129)))
```

```
                    Name HR SB    BB     K   AVG   OBP   SLG  WAR Position
1      Anthony Santander 44  2 0.087 0.194 0.235 0.308 0.506  3.3       OF
2      Christopher Morel 21  8 0.100 0.260 0.196 0.288 0.346 -1.0      INF
3        George Springer 19 16 0.098 0.187 0.220 0.303 0.371  1.2       OF
4              Juan Soto 41  7 0.181 0.167 0.288 0.419 0.569  8.1       OF
5           Mookie Betts 19 16 0.118 0.110 0.289 0.372 0.491  4.4       OF
6           Rhys Hoskins 26  3 0.103 0.288 0.214 0.303 0.419  0.1      INF
7           Seiya Suzuki 21 16 0.108 0.274 0.283 0.366 0.482  3.6       OF
8 Vladimir Guerrero Jr. 30  2 0.103 0.138 0.323 0.396 0.544  5.4      INF
  OSwing ZSwing Contact  Zone SwStr      Cooks    StuRes     dffits  covratio
1  0.375  0.688   0.822 0.392 0.088 0.03242748  1.257791  0.4421511 1.0912195
2  0.297  0.719   0.684 0.410 0.148 0.05261083 -2.189317 -0.5707839 0.8818551
3  0.284  0.753   0.769 0.440 0.113 0.03234752 -1.896058 -0.4453125 0.9264256
4  0.213  0.601   0.799 0.402 0.074 0.07717007 -2.026086 -0.6892841 0.9533154
5  0.229  0.654   0.861 0.410 0.056 0.04176453 -2.380587 -0.5104458 0.8263205
6  0.281  0.615   0.771 0.439 0.098 0.03188314  1.865198  0.4418897 0.9327265
7  0.236  0.627   0.777 0.410 0.088 0.03618640  2.390269  0.4752301 0.8194639
8  0.306  0.710   0.801 0.430 0.096 0.03198945 -1.879266 -0.4427234 0.9296988
```

```r
filter(mlb_3, covratio > (1 + (15/129)) | covratio < (1 - (15/129)))
```

```
                  Name HR SB    BB     K   AVG   OBP   SLG  WAR Position OSwing
1          Aaron Judge 58 10 0.189 0.243 0.322 0.458 0.701 11.2       OF  0.213
2          Alex Bregman 26  3 0.069 0.136 0.260 0.315 0.453  4.2      INF  0.265
3       Bobby Witt Jr. 32 31 0.080 0.150 0.332 0.389 0.588 10.4      INF  0.354
4       Carlos Santana 23  4 0.109 0.167 0.238 0.328 0.420  3.0      INF  0.270
5      Ceddanne Rafaela 15 19 0.026 0.264 0.246 0.274 0.390  0.9       OF  0.495
6    Christopher Morel 21  8 0.100 0.260 0.196 0.288 0.346 -1.0      INF  0.297
7         Corey Seager 30  1 0.099 0.180 0.278 0.353 0.512  4.6      INF  0.317
8      Elly De La Cruz 25 67 0.099 0.313 0.259 0.339 0.471  6.4      INF  0.297
9       Ezequiel Tovar 26  6 0.033 0.288 0.269 0.295 0.469  3.7      INF  0.481
10         Jake Burger 29  1 0.054 0.259 0.250 0.301 0.460  1.4      INF  0.413
11        Jose Ramirez 39 41 0.079 0.120 0.279 0.335 0.537  6.5      INF  0.348
12      Kyle Schwarber 38  5 0.153 0.285 0.248 0.366 0.485  3.4       DH  0.240
13        Luis Arraez  4  9 0.036 0.043 0.314 0.346 0.392  1.1      INF  0.368
14       Michael Busch 21  2 0.111 0.286 0.248 0.335 0.440  2.3      INF  0.274
15        Mookie Betts 19 16 0.118 0.110 0.289 0.372 0.491  4.4       OF  0.229
16        Nico Hoerner  7 31 0.069 0.103 0.273 0.335 0.373  4.0      INF  0.352
17         Sal Frelick  2 18 0.074 0.149 0.259 0.320 0.335  1.6       OF  0.296
18      Salvador Perez 27  0 0.067 0.198 0.271 0.330 0.456  3.1        C  0.464
19        Seiya Suzuki 21 16 0.108 0.274 0.283 0.366 0.482  3.6       OF  0.236
20 Vinnie Pasquantino 19  1 0.072 0.128 0.262 0.315 0.446  1.5      INF  0.338
```

```
21        Yainer Diaz 16  2 0.039 0.173 0.299 0.325 0.441  3.0        C  0.457
22         Zack Gelof 17 25 0.069 0.344 0.211 0.270 0.362  1.4      INF  0.328
   ZSwing Contact  Zone SwStr        Cooks      StuRes       dffits  covratio
1   0.719   0.712 0.410 0.121 2.820855e-03 -0.26405370 -0.129603435 1.3006955
2   0.691   0.886 0.432 0.051 1.324954e-04 -0.09006203 -0.028081334 1.1526726
3   0.748   0.808 0.400 0.098 4.728948e-06 -0.01866809 -0.005305006 1.1356841
4   0.667   0.787 0.412 0.092 1.440188e-02 -2.43642682 -0.300091586 0.7918695
5   0.787   0.696 0.411 0.187 9.245912e-03 -0.67214577 -0.235004902 1.1534076
6   0.719   0.684 0.410 0.148 5.261083e-02 -2.18931670 -0.570783905 0.8818551
7   0.824   0.784 0.421 0.114 3.230305e-04  0.15159557  0.043849569 1.1374946
8   0.624   0.679 0.404 0.138 7.968706e-04  0.24103678  0.068881007 1.1334077
9   0.826   0.690 0.401 0.192 6.611364e-03  0.62328442  0.198671447 1.1356740
10  0.698   0.715 0.389 0.149 2.040312e-03 -0.40031327 -0.110264149 1.1216181
11  0.706   0.864 0.406 0.067 2.290098e-03 -0.33574779 -0.116796317 1.1717472
12  0.626   0.704 0.413 0.118 4.418437e-05  0.06246263  0.016216017 1.1213451
13  0.647   0.942 0.453 0.029 2.647594e-03 -0.34386893 -0.125584928 1.1844317
14  0.710   0.745 0.404 0.115 1.721517e-02  2.27140280  0.327011870 0.8283663
15  0.654   0.861 0.410 0.056 4.176453e-02 -2.38058717 -0.510445803 0.8263205
16  0.653   0.893 0.431 0.051 4.433211e-04  0.18658168  0.051371665 1.1285232
17  0.563   0.884 0.434 0.048 9.423523e-04 -0.25765448 -0.074907767 1.1359596
18  0.776   0.741 0.364 0.149 1.286440e-03 -0.29694890 -0.087529318 1.1372235
19  0.627   0.777 0.410 0.088 3.618640e-02  2.39026861  0.475230145 0.8194639
20  0.674   0.875 0.412 0.059 8.998308e-06  0.02866328  0.007317873 1.1191675
21  0.796   0.776 0.387 0.132 1.189688e-03  0.29800999  0.084173696 1.1297028
22  0.745   0.653 0.443 0.178 1.775137e-04 -0.10551230 -0.032504132 1.1500473
```

```r
filter(mlb_3, abs(dffits) > (2*sqrt(5/129)))
```

```
                  Name HR SB    BB     K   AVG   OBP   SLG  WAR Position
1     Anthony Santander 44  2 0.087 0.194 0.235 0.308 0.506  3.3       OF
2          Brice Turang  7 50 0.081 0.170 0.254 0.316 0.349  2.5      INF
3          Bryce Harper 30  7 0.120 0.219 0.285 0.373 0.525  5.2      INF
4     Christopher Morel 21  8 0.100 0.260 0.196 0.288 0.346 -1.0      INF
5       George Springer 19 16 0.098 0.187 0.220 0.303 0.371  1.2       OF
6             Juan Soto 41  7 0.181 0.167 0.288 0.419 0.569  8.1       OF
7          Mookie Betts 19 16 0.118 0.110 0.289 0.372 0.491  4.4       OF
8        Nathaniel Lowe 16  2 0.126 0.221 0.265 0.361 0.401  2.8      INF
9          Rhys Hoskins 26  3 0.103 0.288 0.214 0.303 0.419  0.1      INF
10         Seiya Suzuki 21 16 0.108 0.274 0.283 0.366 0.482  3.6       OF
11        Shohei Ohtani 54 59 0.111 0.222 0.310 0.390 0.646  9.1       DH
12 Vladimir Guerrero Jr. 30  2 0.103 0.138 0.323 0.396 0.544  5.4      INF
   OSwing ZSwing Contact  Zone SwStr      Cooks     StuRes      dffits  covratio
```

```
1   0.375  0.688   0.822 0.392 0.088 0.03242748  1.257791  0.4421511 1.0912195
2   0.317  0.644   0.880 0.454 0.056 0.02698486  1.883713  0.4066489 0.9212363
3   0.366  0.805   0.751 0.368 0.131 0.02646009  1.556614  0.4007920 0.9930907
4   0.297  0.719   0.684 0.410 0.148 0.05261083 -2.189317 -0.5707839 0.8818551
5   0.284  0.753   0.769 0.440 0.113 0.03234752 -1.896058 -0.4453125 0.9264256
6   0.213  0.601   0.799 0.402 0.074 0.07717007 -2.026086 -0.6892841 0.9533154
7   0.229  0.654   0.861 0.410 0.056 0.04176453 -2.380587 -0.5104458 0.8263205
8   0.271  0.639   0.812 0.424 0.080 0.02709372  1.733701  0.4065459 0.9542884
9   0.281  0.615   0.771 0.439 0.098 0.03188314  1.865198  0.4418897 0.9327265
10  0.236  0.627   0.777 0.410 0.088 0.03618640  2.390269  0.4752301 0.8194639
11  0.305  0.705   0.735 0.418 0.125 0.02807253 -1.166090 -0.4110150 1.1041315
12  0.306  0.710   0.801 0.430 0.096 0.03198945 -1.879266 -0.4427234 0.9296988
```

```r
interact1 <- lm(K ~ OBP + SLG + OSwing + ZSwing + SwStr + SLG:ZSwing, data = mlb3)
summary(interact1)
```

```
Call:
lm(formula = K ~ OBP + SLG + OSwing + ZSwing + SwStr + SLG:ZSwing,
    data = mlb3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.048259 -0.012179 -0.000761  0.011466  0.051946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.83637    0.17585   4.756 5.46e-06 ***
OBP         -0.43093    0.10753  -4.008 0.000106 ***
SLG         -0.67450    0.39131  -1.724 0.087298 .
OSwing      -0.44470    0.04413 -10.076  < 2e-16 ***
ZSwing      -0.84155    0.24455  -3.441 0.000794 ***
SwStr        1.95578    0.07944  24.618  < 2e-16 ***
SLG:ZSwing   1.08865    0.55600   1.958 0.052511 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01979 on 122 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8606
F-statistic: 132.7 on 6 and 122 DF,  p-value: < 2.2e-16
```

```
# interact2 <- lm(Kperc ~ OBP + SLG + OSwingperc + ZSwingperc + SwStrperc + SwStrperc:Positi
# summary(interact2)


# interact3 <- lm(Kperc ~ SwStrperc + SLG * ZSwingperc, data = mlb3)
# summary(interact3)
```

First we looked at MLR model with every predictor to see how the model would fit. After that, we ran a best subsets regression to find the optimal model with the most optimal predictors. Based off that optimal model, we checked all of our assumptions and multicollinearity. After finding no issues we checked for influential points or outliers. Then we tested out some interaction terms to see if we could improve our model even more.

Testing Model

```
# mlb3 %>%
  # sample_n(8)

players_to_test <- c("Will Smith", "Isaac Paredes", "Francisco Lindor", "Corey Seager",
                     "Brendan Rodgers", "Ryan McMahon", "Wyatt Langford", "Trea Turner")

test_data <- mlb3 %>%
  filter(Name %in% players_to_test)

test_data$predicted_K <- predict(interact1, newdata = test_data)

test_data %>%
  select(Name, K, predicted_K)
```

```
              Name     K predicted_K
1  Brendan Rodgers 0.245   0.2257628
2     Corey Seager 0.180   0.1867490
3 Francisco Lindor 0.184   0.2041766
4    Isaac Paredes 0.164   0.1525343
5     Ryan McMahon 0.287   0.2729263
6      Trea Turner 0.182   0.2093556
7       Will Smith 0.193   0.1804150
8   Wyatt Langford 0.206   0.2123676
```

We then tested the model on 8 randomly selected players which were randomly generated from the sample. Those players were Will Smith, Isaac Paredes, Francisco Lindor, Corey Seager, Brendan Rodgers, Ryan McMahon, Wyatt Langford and Trea Turner. Looking at the results between the predicted strikeout rate and the actual strikeout rate, all 8 players predicted strikeout rates were within 1.5% of their true strikeout rate on average.