MLB Win Percentage Model

Abstract: This study takes an analytical approach to the field of sports, focusing on how data analysis can be used to better understand and predict wins and win percentage in Major League Baseball (MLB). Recognizing the growing role of statistics in sports, we worked with official MLB season data to explore how different performance metrics relate to a team's success, specifically their total wins and win percentage. Our goal was to create a model that could be useful not just for researchers, but also for fans, coaches, and others involved with the game. Through an in-depth exploratory data analysis, we were able to identify several key variables that had strong relationships with team performance. Using these variables, we built a full predictive model. Throughout the process, we encountered challenges like multicollinearity among the predictors, which complicated the initial modeling process. However, by applying backward selection methods, we were able to fix up our model and remove unnecessary variables. In the end, we developed a final model that accurately predicts a team's win percentage based on important game statistics.

Motivation: This study explores how standard baseball statistics can be used to predict wins and win percentage, identifying key indicators that contribute to a team's success. By analyzing metrics such as runs scored (R), home runs (HR), earned run average (ERA), and fielding percentage (FP), we want to determine which factors have the strongest correlation with winning records.

With that being said, this research is important in our selected field of sports because it helps the analytical approach to evaluating baseball teams. Statistical insights can help people in the field of baseball such as coaches, analysts, and executives refine strategies, optimize player lineups, and even improve overall team performances. This study also allows fans to better understand important metrics in baseball that can help determine success of a baseball team.

Data Description: The dataset used in this research is called “MLB Stats, Scores, History, & Records” ("MLB Stats, Scores, History, & Records”) and consists of standard baseball statistics from the Major League Baseball (MLB) team statistics from the year 2000 spanning to the 2015 season, split by league (AL and NL). There is a total of 18 variables: year of the season (yearID), league identifier (lgID), number of games played (G), number of wins (W), winning percentage (WIN.PCT), total runs scored by team (R), total-at-bats (AB), total hits (H), total home runs (HR), total walks (BB), total strikeouts by batters (SO), total stolen bases (SB), total opponents’ runs (RA), earned run average (ERA), hits allowed by pitchers (HA), total errors committed (E), fielding percentage (FP), and run differential (RUNDIFF). Considering our motivation and purpose of this research, we are particularly interested in the winning percentage (WIN.PCT) variable. It is necessary for further analysis to decide what other variables will be of interest and/or include in our model.

This data was collected from Major League Baseball (MLB) game records and team statistics, which are compiled and maintained by organizations such as: MLB Official Scorekeepers, Retrosheet & Baseball Reference, Elias Sports Bureau and FanGraphs & Statcast **(**"MLB Stats, Scores, History, & Records"). This data is useful for our motivation because we can trust MLB’s records and statistics are accurate enough for our research. And, the organization is a SME on the topic, which is also relevant when considering the validity of the data.

Exploratory Data Analysis: The first step in exploratory data analysis with this data set is becoming familiarized with the data. To start, we looked at basic summary statistics, including the mean, median, and the range of values for each variable, along with minimum and maximum values for each variable.

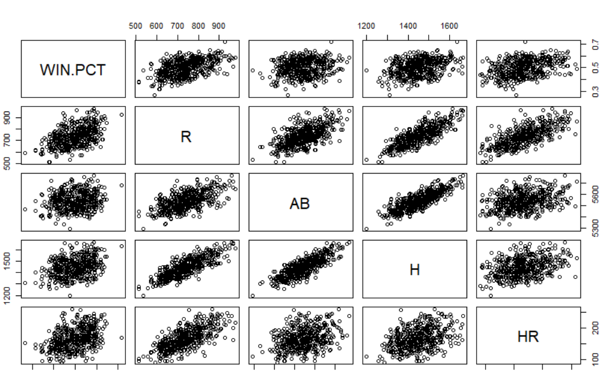
```
> summary(df)
     yearID          lgID              WIN.PCT             R               AB              H               HR       
 Min.   :2000   Length:480         Min.   :0.2654   Min.   :513.0   Min.   :5294   Min.   :1199   Min.   : 91.0  
 1st Qu.:2004   Class :character   1st Qu.:0.4444   1st Qu.:681.8   1st Qu.:5487   1st Qu.:1386   1st Qu.:142.8  
 Median :2008   Mode  :character   Median :0.5062   Median :735.0   Median :5542   Median :1447   Median :164.0  
 Mean   :2008                      Mean   :0.4999   Mean   :739.5   Mean   :5543   Mean   :1448   Mean   :166.7  
 3rd Qu.:2011                      3rd Qu.:0.5556   3rd Qu.:795.2   3rd Qu.:5600   3rd Qu.:1506   3rd Qu.:188.0  
 Max.   :2015                      Max.   :0.7160   Max.   :978.0   Max.   :5769   Max.   :1667   Max.   :260.0  
       BB              SO             SB               RA             ERA              HA              E        
 Min.   :363.0   Min.   : 805   Min.   : 31.00   Min.   :525.0   Min.   :2.940   Min.   :1233   Min.   : 54.0  
 1st Qu.:471.0   1st Qu.:1024   1st Qu.: 71.00   1st Qu.:675.8   1st Qu.:3.857   1st Qu.:1384   1st Qu.: 90.0  
 Median :520.0   Median :1104   Median : 91.50   Median :733.5   Median :4.200   Median :1448   Median :101.0  
 Mean   :522.4   Mean   :1116   Mean   : 94.56   Mean   :739.5   Mean   :4.236   Mean   :1448   Mean   :101.8  
 3rd Qu.:567.0   3rd Qu.:1207   3rd Qu.:115.00   3rd Qu.:801.0   3rd Qu.:4.593   3rd Qu.:1506   3rd Qu.:112.0  
 Max.   :775.0   Max.   :1535   Max.   :200.00   Max.   :974.0   Max.   :5.710   Max.   :1683   Max.   :145.0  
       FP            RUNDIFF      
 Min.   :0.9760   Min.   :0.0000  
 1st Qu.:0.9820   1st Qu.:0.0000  
 Median :0.9830   Median :1.0000  
 Mean   :0.9832   Mean   :0.5167  
 3rd Qu.:0.9850   3rd Qu.:1.0000  
```

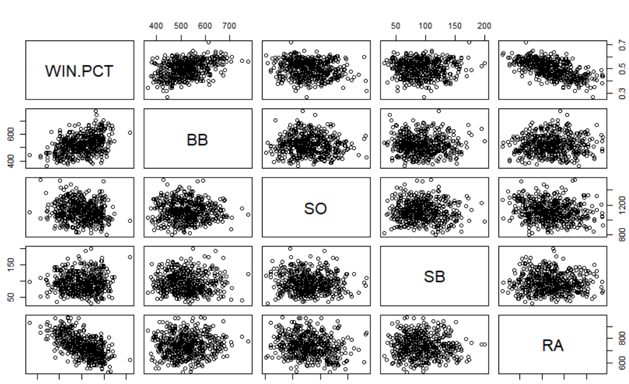
All values listed were possible, which meant there were no extreme values to be excluded. We were also able to analyze other descriptive statistics, like mean and median. Numeric and categorical graphs were used to visualize the distribution and potential outliers for each variable. When looking at the boxplot of each, there did not appear to be any extreme or obvious outliers for any of the variables.

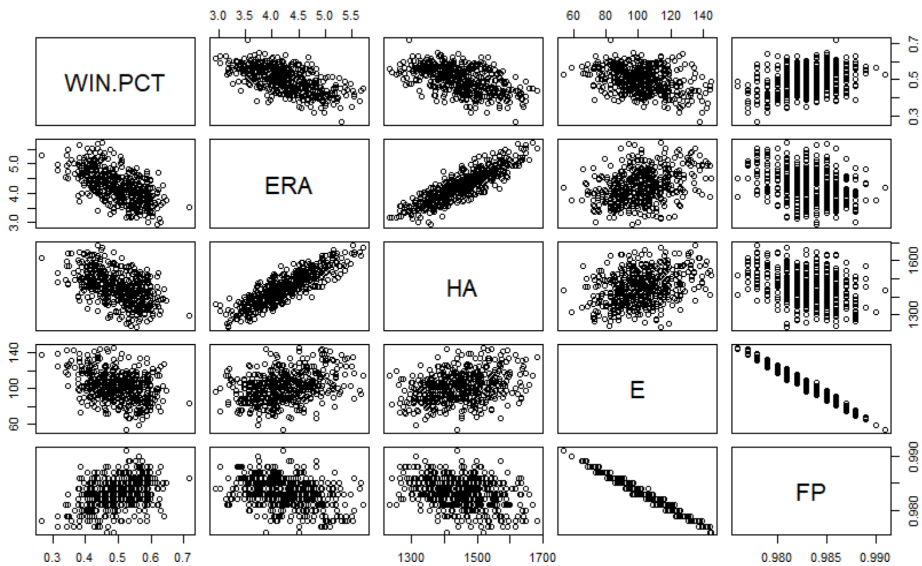
Next, basic scatter plots were utilized to initially assess for linearity. Win percentage was used in every scatterplot matrix in order to check its linearity with every other variable.

```
pairs(df[,c(3,4,5,6,7)]) pairs(df[,c(3,8,9,10,11)])
```

```
pairs(df[,c(3,12,13,14,15)])
```







Using this information, it appears that most variables have a linear relationship when compared to win percentage. But some variables, such as runs, hits, home runs, era, runs allowed, and hits allowed, have a stronger linear relationship than others, like strikeouts, stolen bases, and errors. These assumptions were supported by the correlation matrix of all variables.

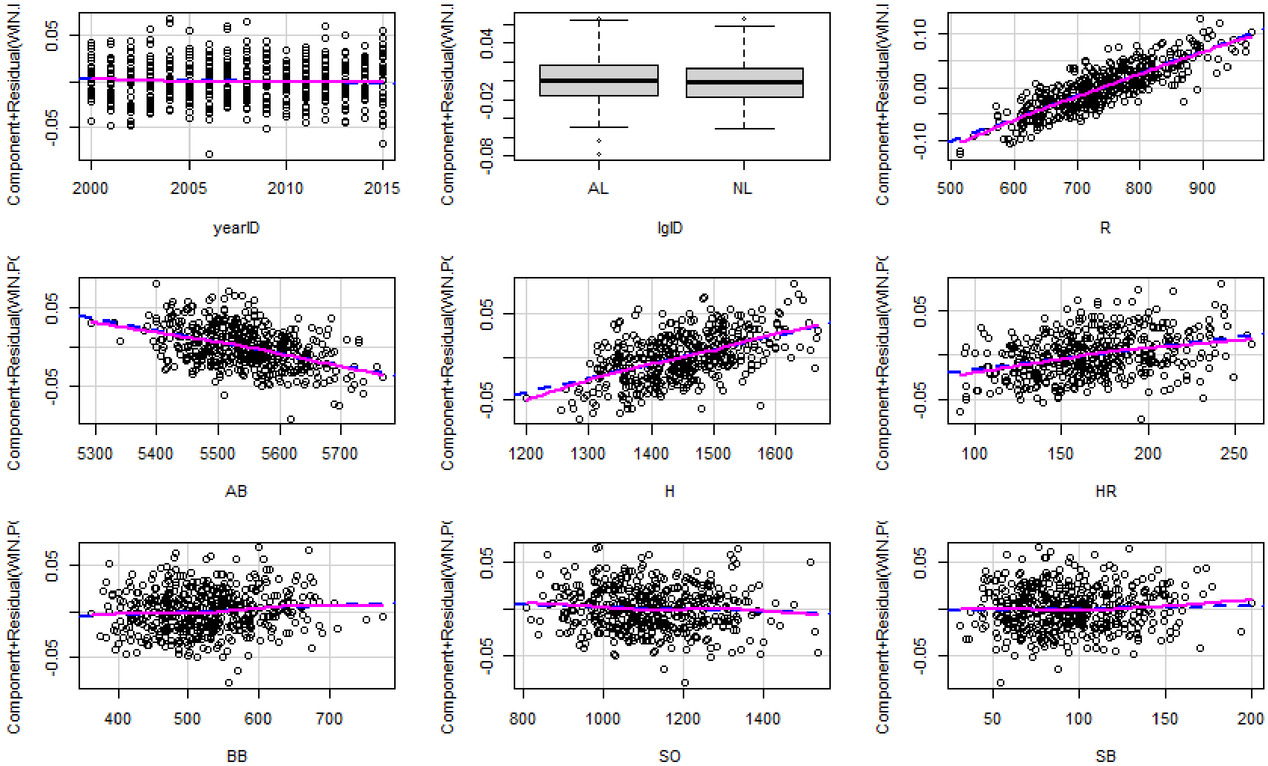
```
cor(df[,-c(2,16)])
```

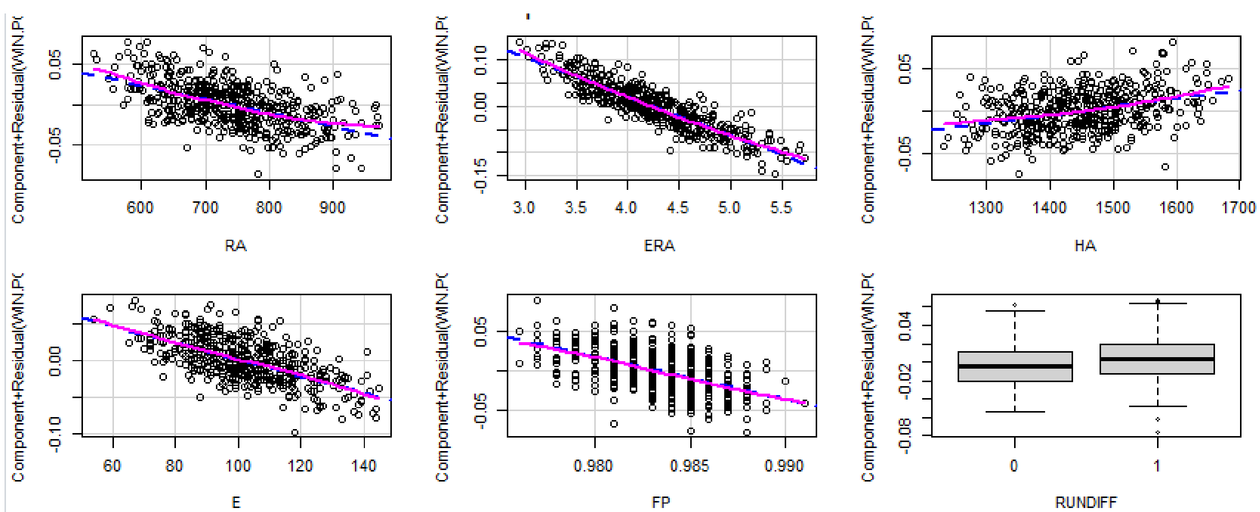
```
             yearID     WIN.PCT          R         AB          H          HR         BB           SO          SB
WIN.PCT 0.001016896 1.000000000 0.51337272 0.17098675 0.34322796 0.392977232 0.39776815 -0.134534109 0.045076956
                RA         ERA         HA           E          FP
WIN.PCT -0.61813743 -0.61568765 -0.51312827 -0.362053102 0.370292791
```

It is seen in this matrix that runs allowed have the greatest correlation value, while at-bats have the lowest.

The next step in the analysis of the data set was examining partial residual plots to determine any need for transformations.

```
fullmodel <- lm(WIN.PCT ~ ., data=df)
crPlots(lm(WIN.PCT ~ ., data=df))
```





When comparing the pink curve and blue line, it can be seen that there is a very small discrepancy between the lines. This line difference indicates whether or not a transformation is needed for a given variable. Since there is little contrast between lines, it appears as if no transformations will need to be made.

Finally, we examined the pairwise correlations to determine any potential collinearity issues.

```
cor(df[,-c(2,16)])
```

```
            yearID     WIN.PCT           R          AB           H           HR          BB           SO           SB
yearID   1.000000000  0.001016896 -0.47527930 -0.18721580 -0.37415501 -0.300228335 -0.37144152  0.531764614  0.014159662
WIN.PCT  0.001016896  1.000000000  0.51337272  0.17098675  0.34322796  0.392977232  0.39776815 -0.134534109  0.045076956
R       -0.475279296  0.513372717  1.00000000  0.60999706  0.81018517  0.704337960  0.60130637 -0.346972268  0.011230910
AB      -0.187215796  0.170986754  0.60999706  1.00000000  0.82205418  0.323892255  0.13145988 -0.297860110 -0.031555931
H       -0.374155008  0.343227959  0.81018517  0.82205418  1.00000000  0.374038549  0.26070073 -0.532102655  0.032909834
HR      -0.300228335  0.392977232  0.70433796  0.32389225  0.37403855  1.000000000  0.42957746  0.019086393 -0.125191471
BB      -0.371441523  0.397768147  0.60130637  0.13145988  0.26070073  0.429577461  1.00000000 -0.037445900 -0.040762195
SO       0.531764614 -0.134534109 -0.34697227 -0.29786011 -0.53210265  0.019086393 -0.03744590  1.000000000 -0.060593177
SB       0.014159662  0.045076956  0.01123091 -0.03155593  0.03290983 -0.125191471 -0.04076220 -0.060593177  1.000000000
RA      -0.452750690 -0.618137425  0.26579298  0.29694413  0.30240120  0.182962542  0.06114322 -0.172959746 -0.040756471
ERA     -0.434659004 -0.615687650  0.26409911  0.27817859  0.30525484  0.176807413  0.04470081 -0.185996657 -0.037150527
HA      -0.348569478 -0.513128270  0.24921260  0.33118702  0.31792062  0.130797196  0.03006073 -0.160745416 -0.061653574
E       -0.323926895 -0.362053102 -0.01549795 -0.05492639 -0.04371590 -0.001352425 -0.01581438  0.009541123  0.005871666
FP       0.314827063  0.370292791  0.01775273  0.07150755  0.04457410  0.008225680  0.02430360 -0.005643719 -0.017709030
                 RA         ERA          HA            E           FP
yearID  -0.45275069 -0.43465900 -0.34856948 -0.323926895  0.314827063
WIN.PCT -0.61813743 -0.61568765 -0.51312827 -0.362053102  0.370292791
R        0.26579298  0.26409911  0.24921260 -0.015497955  0.017752729
AB       0.29694413  0.27817859  0.33118702 -0.054926388  0.071507550
H        0.30240120  0.30525484  0.31792062 -0.043715901  0.044574100
HR       0.18296254  0.17680741  0.13079720 -0.001352425  0.008225680
BB       0.06114322  0.04470081  0.03006073 -0.015814382  0.024303603
SO      -0.17295975 -0.18599666 -0.16074542  0.009541123 -0.005643719
SB      -0.04075647 -0.03715053 -0.06165357  0.005871666 -0.017709030
RA       1.00000000  0.98890593  0.87667811  0.430936944 -0.443104400
ERA      0.98890593  1.00000000  0.86974315  0.349860703 -0.366768774
HA       0.87667811  0.86974315  1.00000000  0.337316233 -0.332188221
E        0.43093694  0.34986070  0.33731623  1.000000000 -0.987504878
FP      -0.44310440 -0.36676877 -0.33218822 -0.987504878  1.000000000
```

This output shows multiple potential collinearity issues. ERA, hits allowed, and runs allowed have a correlation greater than 0.8 with one another, which reveals a potential collinearity issue. The model will need to be refit in order to adjust properly for potential collinearity issues. Overall, we believe that runs, hits, home runs, era, hits allowed, and run differential will be useful in our model. This is due to all of these variables having a solid linear relationship and not needing to be transformed, although adjustments may need to be made due to collinearity. We will likely use interaction terms between run differential and all other variables in the full model, as they are often related in a baseball season.. Based on these findings, this data should be sufficient in creating a model to predict MLB season win percentage.

Model Diagnostics and Model Selection: We set out in our model selection process by finalizing our full model. The full model consisted of the variables, runs, at-bats, hits, home runs, walks, strikeouts, stolen bases, runs allowed, earned run average, hits allowed, errors, fielding percentage, run differential, and interaction terms between run differential and all other variables.

```
fullmodel <- lm(WIN.PCT ~ . + RUNDIFF*., data=df)
```

We decided to use backward selection in order to find our selected model.

```
ols_step_backward_p(fullmodel, 0.1)
```

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	8.890	2.683		3.313	0.001	3.617	14.164
R	0.001	0.000	0.635	12.382	0.000	0.000	0.001
AB	0.000	0.000	-0.216	-6.216	0.000	0.000	0.000
H	0.000	0.000	0.190	4.495	0.000	0.000	0.000
HR	0.000	0.000	0.095	3.813	0.000	0.000	0.000
ERA	-0.103	0.006	-0.785	-18.653	0.000	-0.114	-0.092
HA	0.000	0.000	0.109	3.398	0.001	0.000	0.000
E	-0.002	0.000	-0.423	-4.059	0.000	-0.003	-0.001
FP	-7.606	2.697	-0.287	-2.820	0.005	-12.907	-2.306
RUNDIFF1	-0.258	0.179	-1.835	-1.447	0.149	-0.609	0.093
R:RUNDIFF1	0.000	0.000	-0.534	-2.170	0.030	0.000	0.000
AB:RUNDIFF1	0.000	0.000	2.446	1.787	0.075	0.000	0.000
ERA:RUNDIFF1	-0.013	0.007	-0.362	-1.919	0.056	-0.026	0.000
E:RUNDIFF1	0.000	0.000	0.323	3.214	0.001	0.000	0.001

After backward selection, we were left with 13 significant variables: runs, at-bats, hits, home runs, earned run average, hits allowed, errors, fielding percentage, run differential, run differential*runs, run differential*at-bats, run differential*earned run average, and run differential*errors.

```
backwardSelection <- lm(WIN.PCT ~ R + AB + H + HR + ERA + HA + E + FP + RUNDIFF
                        + R*RUNDIFF + AB*RUNDIFF + ERA*RUNDIFF + E*RUNDIFF, data=df)
```

Once the model was selected, we decided to examine possible collinearity issues based on our findings during exploratory data analysis. Taking a look at the Variance Inflation Factor of each variable, we were able to determine if any variables should be removed.

```
backwardSelectionNoInt <- lm(WIN.PCT ~ R + AB + H + HR + ERA + HA + E + FP + RUNDIFF, data=df)
```

```
ols_coll_diag(backwardSelectionNoInt)
```

```
Tolerance and Variance Inflation Factor
---------------------------------------
  Variables  Tolerance       VIF
1         R 0.12946133  7.724314
2        AB 0.27570994  3.627000
3         H 0.12555860  7.964408
4        HR 0.36012821  2.776789
5       ERA 0.16443755  6.081336
6        HA 0.21740288  4.599755
7         E 0.02240351 44.635860
8        FP 0.02204504 45.361671
9  RUNDIFF1 0.30702480  3.257066
```

Both errors and fielding percentage have a VIF greater than 10, so collinearity appears to be present. With this information, we decided to remove the fielding percentage from the model.

```
newBackwardSelectionNoInt <- lm(WIN.PCT ~ R + AB + H + HR + ERA + HA + E + RUNDIFF, data=df)
```

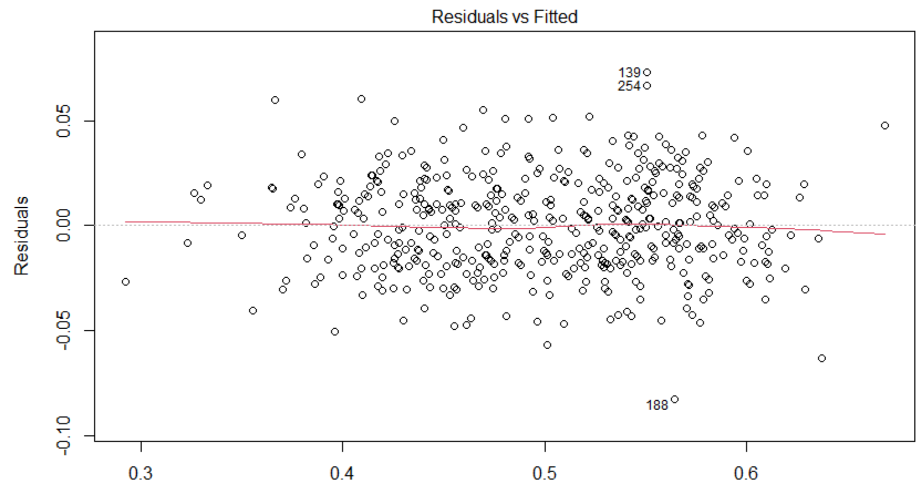
```
Tolerance and Variance Inflation Factor
---------------------------------------
  Variables Tolerance      VIF
1         R 0.1294614 7.724311
2        AB 0.2826812 3.537553
3         H 0.1266592 7.895202
4        HR 0.3605297 2.773697
5       ERA 0.1731778 5.774413
6        HA 0.2289041 4.368641
7         E 0.8326036 1.201052
8  RUNDIFF1 0.3075997 3.250979
```

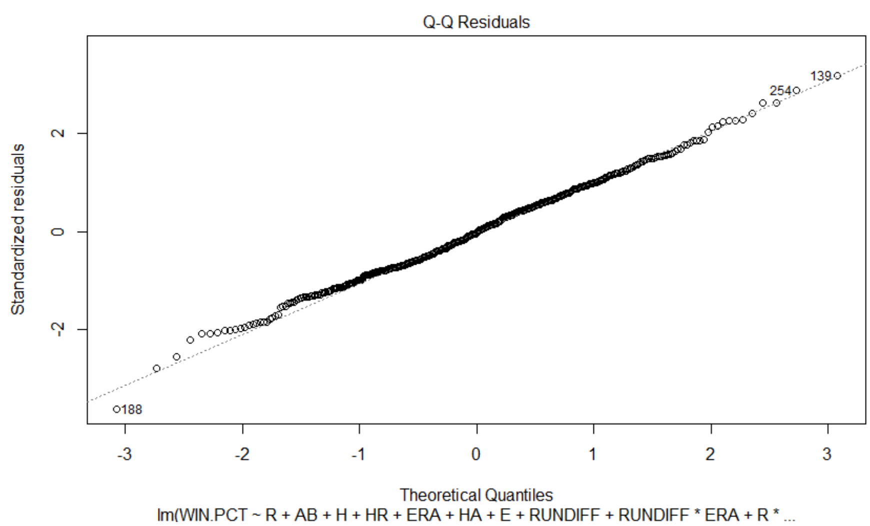
Once fielding percentage was removed, no variable had a VIF greater than 10, showing no more severe collinearity issues present.

This leads us to model diagnostics, checking for outliers and model assumptions. We checked assumptions, normality, homoscedasticity, and normality on the selected model using the residual plot and normal probability plot.

```
finalModel <- lm(WIN.PCT ~ R + AB + H + HR + ERA + HA + E + RUNDIFF
                 + R*RUNDIFF + AB*RUNDIFF + ERA*RUNDIFF + E*RUNDIFF , data=df)

plot(finalModel)
```





Based on these plots, there appear to be no significant violations in linearity, homoscedasticity, or normality. For linearity, the red line does not deviate greatly from zero on the residuals vs fitted plot. For homoscedasticity, the variance of the residual vs fitted plot appears consistent. Finally, for normality, the tails of the Q-Q residuals plot do not skew away from the given line.

We used the Shapiro-Wilk test to determine normality as well.

```
shapiro.test(finalModel$residuals)

        Shapiro-Wilk normality test

data:  finalModel$residuals
W = 0.99688, p-value = 0.4904
```

The Shapiro-Wilk test is not significant, therefore failing to reject H_0, proving normality is not violated. Given these observations, we do not believe a transformation needs to be made on y in this selected model.

Lastly, we wanted to determine any outliers to remove given this selected model. We tried to find any data points that violated Cook's distance, leverage, and jackknife residuals. Starting with leverage, our h_i was found to be 0.0666.

```
tail(sort(hatvalues(finalModel)), 50)
```

```
       387         53        441        282        130        124        355         59        351        156         95 
0.04414217 0.04450121 0.04505864 0.04507292 0.04521715 0.04544495 0.04572999 0.04603436 0.04627635 0.04635137 0.04643003 
        79        456        292         78        193        257        125        213         88        105          1 
0.04663942 0.04686361 0.04689722 0.04755644 0.04770958 0.04784698 0.04817021 0.04843399 0.04851960 0.05052401 0.05073866 
       252        440          6        145        407        443         14        268        119         80        322 
0.05182948 0.05205446 0.05251552 0.05255545 0.05265042 0.05315809 0.05341407 0.05358427 0.05389108 0.05447484 0.05513693 
       391        184        467        400        327         21         40        178        122        393         51 
0.05519699 0.05541993 0.05643491 0.05643715 0.05824965 0.05828260 0.05840193 0.05842837 0.05878007 0.05891046 0.06439523 
        54         10         71         13        101         55 
0.06596158 0.06696912 0.07010462 0.07366469 0.07738713 0.08243735 
```

After analyzing the leverage values, five values were greater than the h_i value, being the 55, 101, 13, 71, and 10 entries.

Next, looking at jackknife residuals, we found t(480-15-2, 0.025) equals about 1.965.

```
t <- qt(.025, 480-15-2, lower.tail=FALSE)

[1] 1.965101

tail(sort(studres(finalModel)), 20)
head(sort(studres(finalModel)), 20)
```

```
     199      200       12       81      468      154       77      116       50       55      151      355      234 
1.692236 1.767316 1.781128 1.824693 1.857503 1.860588 1.866468 1.878328 2.034109 2.153977 2.172938 2.246362 2.265681 
     211      363      294       98      128      254      139 
2.266311 2.289143 2.430865 2.650671 2.652198 2.922334 3.225940 
```

```
     188       456        13        76       429        67       299        40       462       292       472       182
-3.681215 -2.822875 -2.574256 -2.222975 -2.098069 -2.086813 -2.070815 -2.039069 -2.038923 -2.007447 -1.985401 -1.959370
       46       353       361       179       326       272        66       401
-1.923681 -1.899436 -1.874246 -1.852564 -1.852326 -1.848815 -1.798584 -1.774936
```

After analyzing the jackknife residuals, 24 data points were greater than the t-value, violating the jackknife residual.

Finally, we looked at Cook's distance, checking for any data point with a value greater than one.

```
tail(sort(cooks.distance(finalModel)))

        98        139        456        188         55         13
0.02022845 0.02605216 0.02969532 0.03084031 0.03181672 0.04005436
```

No values violated Cook's distance, as they were all less than 1. With this information, no data point violates all three tests, and all seem possible throughout an MLB season. We chose not to remove any points in the entire data set.

Model Reliability: When determining model reliability, we looked at adjusted R^2, F-statistic, BIC, and Mallow's CP. Starting with adjusted R^2 and F-statistic,

```
summary(fullmodel)

Residual standard error: 0.02289 on 450 degrees of freedom
Multiple R-squared:  0.9008,	Adjusted R-squared:  0.8944
F-statistic: 140.8 on 29 and 450 DF,  p-value: < 2.2e-16
```

```
summary(finalModel)

Residual standard error: 0.02313 on 467 degrees of freedom
Multiple R-squared:  0.8948,	Adjusted R-squared:  0.8921
F-statistic: 331.1 on 12 and 467 DF,  p-value: < 2.2e-16
```

The adjusted R^2 values of both models are nearly similar, while the F-statistic of the backward selected model is much larger, with fewer variables. Next, we analyzed the BIC of the full and selected model.

```
BIC(finalModel)
BIC(fullmodel)
```

```
> BIC(finalModel)
[1] -2180.474
> BIC(fullmodel)
[1] -2103.338
```

The BICs of each model are very similar, but the BIC of the final model is slightly lower, indicating a slightly better model. Last, we looked at Mallow's CP of both models.

```
ols_mallows_cp(finalModel,fullmodel)
ols_mallows_cp(fullmodel,fullmodel)
```

```
> ols_mallows_cp(finalModel,fullmodel)
[1] 22.85052
> ols_mallows_cp(fullmodel,fullmodel)
[1] 30
```

The final model has a lower Mallow's CP between the two models, once again showing it is the better model. Overall, the final model appears to be performing better than the full model. With a larger F-statistic, while having fewer variables, the final model has more significance in determining win percentage than the full model.

Results, Summary, and Interpretation: After model selection and reliability, we were left with a model with eight individual variables and four interaction terms. We were then able to examine the summary of the model.

```
summary(finalModel)
```

```
lm(formula = WIN.PCT ~ R + AB + H + HR + ERA + HA + E + RUNDIFF +
    R * RUNDIFF + AB * RUNDIFF + ERA * RUNDIFF + E * RUNDIFF,
    data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.082781 -0.016603 -0.000631  0.015403  0.072683

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.334e+00  1.451e-01   9.196  < 2e-16 ***
R             5.235e-04  4.286e-05  12.213  < 2e-16 ***
AB           -2.022e-04  3.127e-05  -6.466 2.54e-10 ***
H             1.725e-04  3.611e-05   4.776 2.40e-06 ***
HR            1.949e-04  5.271e-05   3.697 0.000244 ***
ERA          -1.009e-01  5.515e-03 -18.303  < 2e-16 ***
HA            7.062e-05  2.504e-05   2.821 0.004997 **
E            -5.899e-04  1.001e-04  -5.892 7.30e-09 ***
RUNDIFF1     -2.314e-01  1.796e-01  -1.288 0.198354
R:RUNDIFF1   -9.113e-05  4.438e-05  -2.053 0.040612 *
AB:RUNDIFF1   5.603e-05  3.487e-05   1.607 0.108740
ERA:RUNDIFF1 -1.105e-02  6.649e-03  -1.661 0.097322 .
E:RUNDIFF1    4.096e-04  1.426e-04   2.872 0.004266 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02313 on 467 degrees of freedom
Multiple R-squared:  0.8948,    Adjusted R-squared:  0.8921
F-statistic: 331.1 on 12 and 467 DF,  p-value: < 2.2e-16
```

Our final model resulted in WINPCT = 1.334 + $5.235e^{-4}$R - $2.022e^{-4}$AB + $1.725e^{-4}$H +$1.949e^{-4}$HR - $1.009e^{-1}$ERA + $7.062e^{-5}$HA - $5.899e^{-4}$E - $2.314e^{-1}$RUNDIFF - $9.113e^{-5}$R*RUNDIFF + $5.603e^{-5}$AB*RUNDIFF - $1.105e^{-2}$ERA*RUNDIFF + $4.096e^{-4}$E*RUNDIFF.

Our final model was successful when predicting wins and win percentage, and we believe this will be useful and meaningful information for baseball fans and professionals. According to our model, the key indicators of a team's success include the number of runs, hits and home runs. For example, for every additional run scored by a team, holding all other factors constant, the team's winning percentage is expected to increase by about 0.05% (because 5.235e-4 = 0.0005235). Suggesting that the higher the runs scored, the higher the chances of winning. Although runs can potentialize the team, earned run average (ERA) doesn't produce the same effect and can hurt winning percentage. As another practical example, for every one-unit increase in a team's earned run average (ERA), holding all other factors constant, the team's winning percentage is expected to decrease by about 0.101% (since -1.009e-1 = -0.101). Suggesting that a higher ERA is detrimental to a team's chances of winning.

In summary, positive coefficients increase the winning percentage, while negative coefficients decrease it. We are hopeful his information can provide coaches with the knowledge to guide their teams toward success, benefiting from insights into areas of focus and potential weaknesses.

Conclusions and Limitations: This study successfully developed a predictive model for Major League Baseball (MLB) teams' win percentage using key performance statistics. Through careful exploratory data analysis, we identified variables such as runs, hits, home runs, earned run average (ERA), hits allowed, and run differential as being most strongly associated with winning. After addressing multicollinearity issues, particularly between ERA, hits allowed, and runs allowed, we finalized a model that included eight individual variables and four interaction terms involving run differential.

The use of backwards selection led to a more efficient model that demonstrated improved reliability compared to the full model. Model diagnostics confirmed that linearity, normality, and homoscedasticity assumptions were reasonably satisfied, with no need for major transformation or removal of data points. Measures such as the adjective R^2 , F-statistic, BIC, and Mallows's CP indicated that the final model outperformed the full initial model, achieving a balance between complexity and predictive power.

Our final model equation provided a strong statistical tool for predicting a team's winning percentage based on offensive and defensive performance indicators. These findings have practical application for coaches, analysts, and fans, offering insights into which metrics contribute most significantly to a team's success in a given season.

Despite the strength of the final model, several limitations must be acknowledged. The data used was limited to a single season's worth of MLB team statistics, which may not fully capture variations across different seasons or account for broader trends in team performance over time. Additionally, the model relied solely on standard box score statistics and did not incorporate more nuanced variables such as player injuries, statistics and did not incorporate more nuanced variables such as player injuries, managerial changes, or situational factors that could impact a team's success.

Multicollinearity, although largely addressed during model selection, could still subtly affect the interpretation of coefficients, especially among highly correlated performance metrics. Furthermore, while strong associations were found, this observational study cannot claim causation between the selected predictors and winning percentage. Future research could improve upon these findings by expanding the dataset to multiple seasons, incorporating advanced baseball metrics, and exploring nonlinear modeling approaches to better capture the complexities of team performance. Nonetheless, within the scope of this study, the results offer valuable insights into the statistical factors most relevant to winning in professional baseball.

Works Cited

"MLB Stats, Scores, History, & Records." *Baseball Reference*, www.baseball-reference.com/. Accessed 15 Mar. 2025.