Win Prediction for Baseball

Chul Hee Kim

2021/05/21

Introduction

Baseball game is won by scoring more runs than the opponent. A team's mission is to maximize the runs scored and to minimize the runs given up. To score more runs, batters need more hits (X1B, X2B, X3B, HR), walks (BB), hit by pitches (HBP), and fewer strikeouts (SO). On the other hand, pitchers basically need to prevent what batters try to accomplish and get outs for their teams. In this report, we will try to predict the number of wins a team can acquire, using the past +100 years of Major League Baseball (MLB) data.

The dataset used in this analysis is taken from *Lahman* package. Among many dataset in the package, here we are only conerned with *Teams* dataset, which contains the batting and pitching statistics along with general information such as the team's name, ballpark, and attendance from 1871 to 2019.

Analysis

Data Preprocessing

- **Dataset 'Teams'**

Let's take a look at the last row of the dataset.

```
     yearID lgID teamID franchID divID Rank   G Ghome  W  L DivWin WCWin LgWin
2925   2019   NL    WAS      WSN     E    2 162    81 93 69      N     Y     Y
     WSWin   R   AB    H X2B X3B  HR  BB   SO  SB CS HBP SF  RA  ER  ERA CG SHO
2925     Y 873 5512 1460 298  27 231 584 1308 116 29  81 42 724 683 4.27  1  13
     SV IPouts   HA HRA BBA  SOA  E  DP    FP                 name
2925 40   4318 1340 202 517 1511 87 111 0.985 Washington Nationals
              park attendance BPF PPF teamIDBR teamIDlahman45 teamIDretro
2925 Nationals Park    2259781 106 104      WSN            MON         WAS
```

There are total of 48 columns, and many of them are not necessary for our analysis, so we will retain the ones shown in the table below.

Of course, columns such as *lgID*, *teamID*, *attendance*, and anything non-numeric are not relevant. Among the remaining, *Rank*, *G*, *Ghome*, *AB*, *CG*, *SHO*, *SV*, *IPouts*, and *DP* are just the records of what happened, not something that actually contributed to the team's winning. *FP* is in percentage, so it is inappropriate

Table 1: Variables retained from Teams dataset

Following 19 variables remain
yearID, W, R, X1B, X2B, X3B, HR, BB, SO, SB, HBP, SH, RA, X123BA, HRA, BBA, SOA, E, group

to compare with the number of wins for the linear regression analysis. *ER* and *ERA* can be regarded as subsets of *RA*, so all these variables will be taken out for our analysis.

Note that among the 19 selected, there are three variables not present in the original dataset: **X1B**, **X123BA**, **group**.

- X1B: H - X2B - X3B - HR
- X123BA: HA - HRA
- group: yearID (1871 ~ 2019) grouped by 30-year interval

Hits (H) include doubles (X2B), triples (X3B), and home runs (HR), and HA (Hits Allowed) include home runs allowed (HRA), so we better extract singles (X1B) and non-HR hits allowed (X123BA) to avoid duplication.

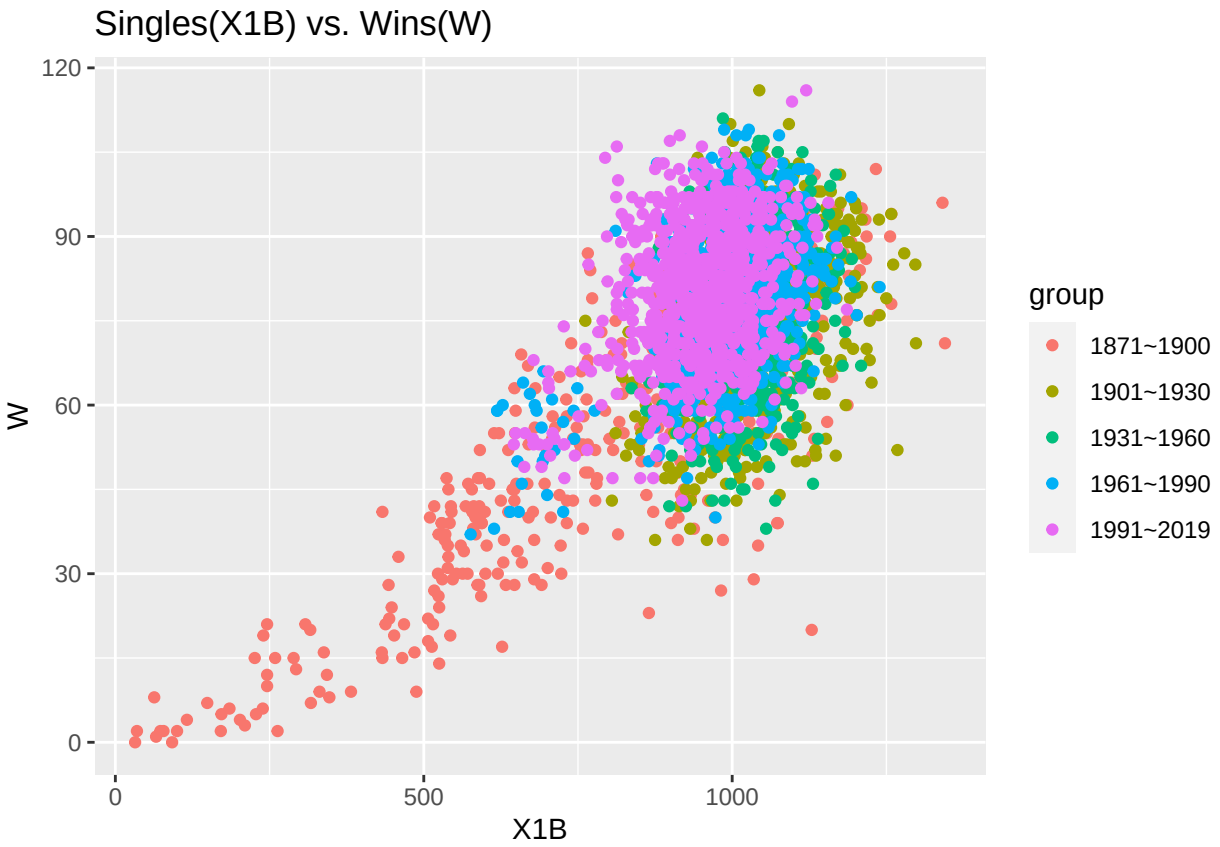
Since baseball has such a long history, there might be some noticeable pattern if we divide all years provided into groups.

Table 2: Variable 'group' explained

group
1871~1900
1901~1930
1931~1960
1961~1990
1991~2019

Let's look at a plot showing X1B vs. W.

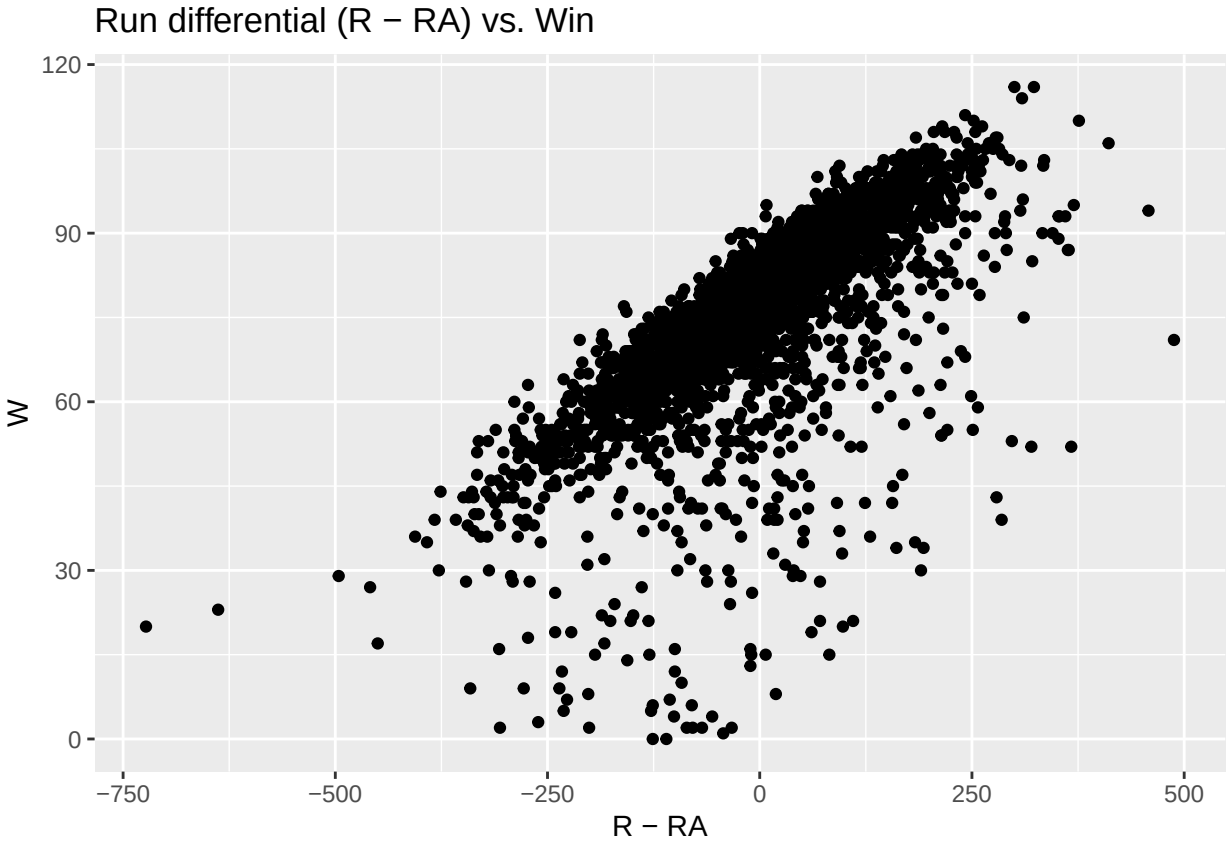
(Go to next page)



There are clearly some outliers, most of which are from 1871 ~ 1900. Although data is provided for those years, the rules of baseball were so much different from today, and even the number of games per year were too few back then. For example, six balls were counted as base-on-balls as opposed four balls. So called "Modern Baseball" starts in 1901, so we will exclude the data before 1901.

(Go to next page)

As mentioned before, baseball teams want to score more runs and give up fewer runs. Hence, wins should be very well correlated with the run differential (R - RA).

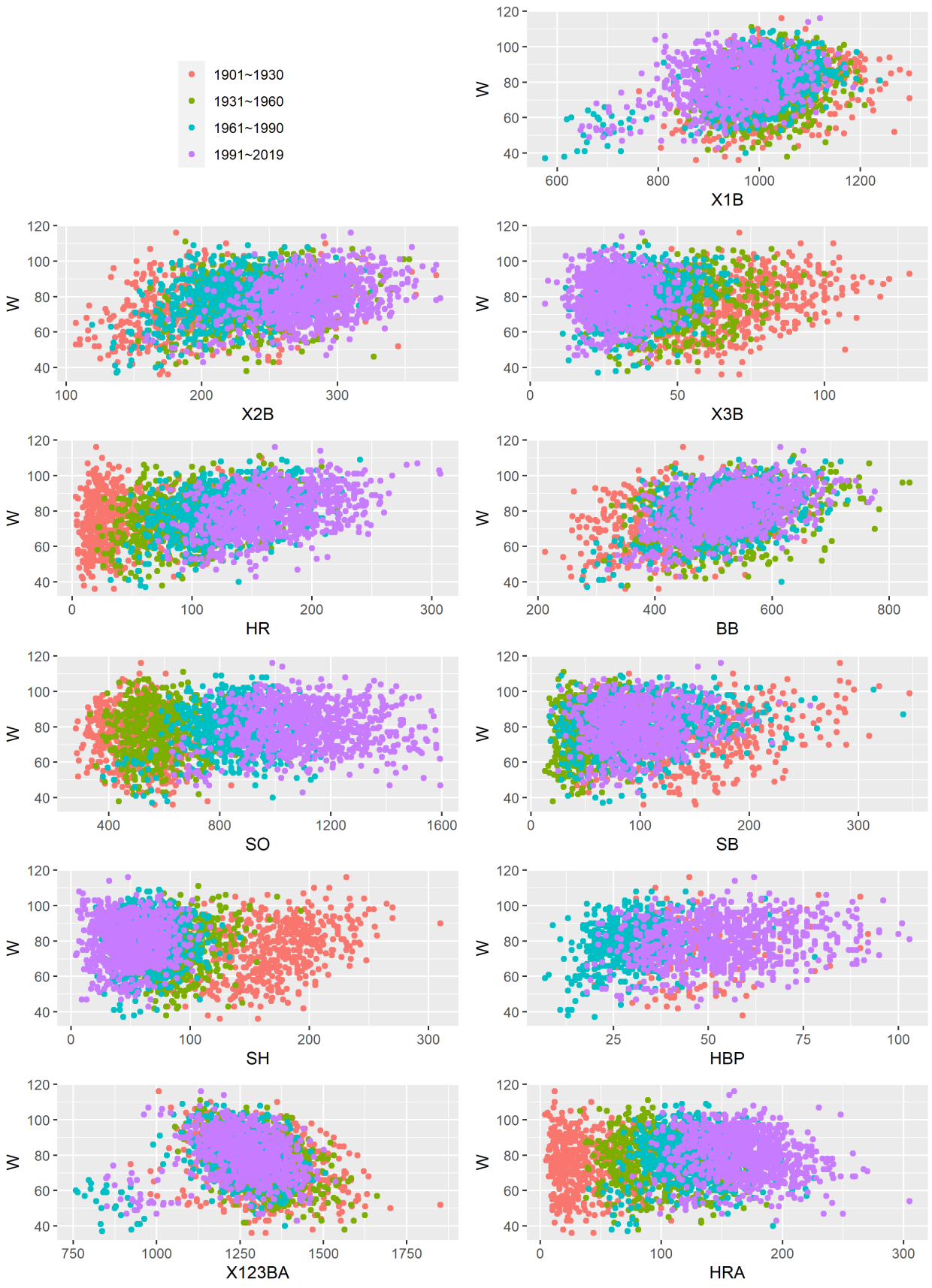


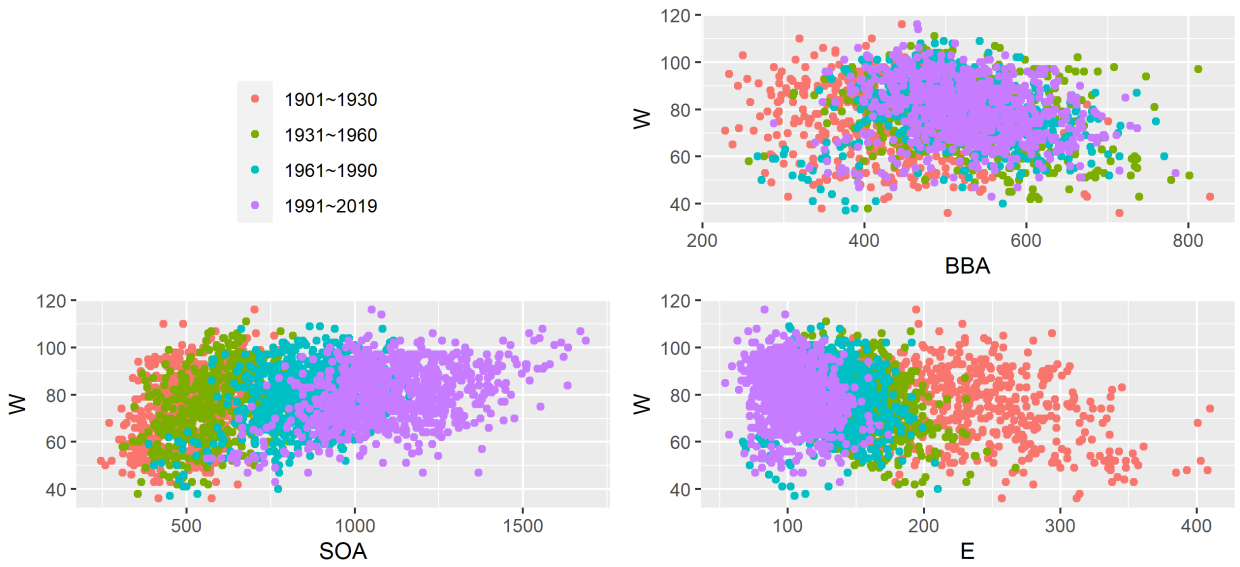
We can divide the variables into two categories: the ones that influence R / the ones that influence RA.

Table 3: Variables into two categories

Run Differential	Variables
R	X1B, X2B, X3B, HR, BB, SO, SB, HBP, SH
RA	X123BA, HRA, BBA, SOA, E

In the next page, we will see the plots showing how each of these variables relate to W.

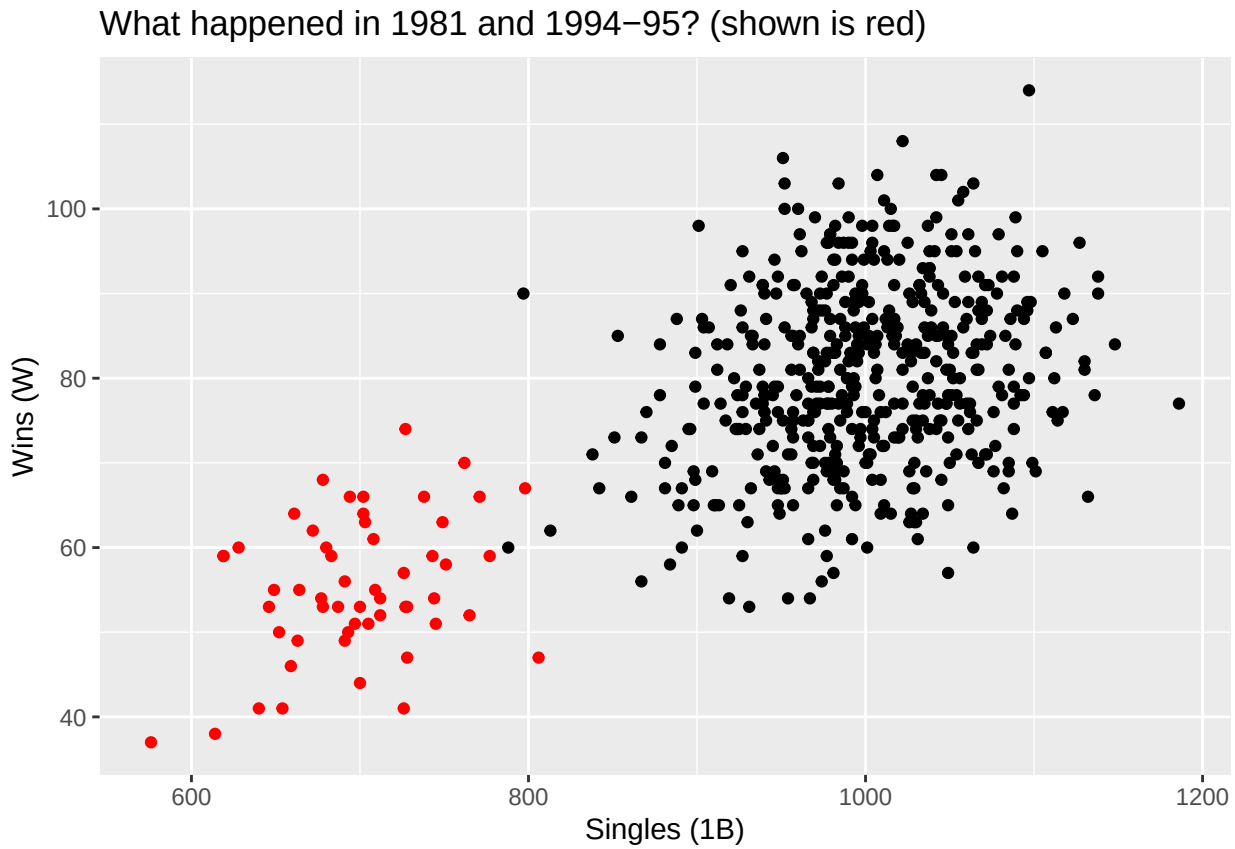

1901~1930
1931~1960
1961~1990
1991~2019
W
X1B
X2B
X3B
HR
BB
SO
SB
SH
HBP
X123BA
HRA



Apart from the correlation between each variable and Win, we can notice that there is some differences among the year groups. For example, X2B, HR, SO, HRA, and SOA has increased over time while X3B, SH, and E decreased over time. Leaving the data as it is and continuing with our analysis may be controversial, but we will move on for now and come back to this later.

Strangely, 'X1B vs. W' and 'X123BA vs. W' plots show some blue dots separated from the group. Just to acknowledge what is going on, we will look at a plot of 'X1B vs. W' with years from 1981 to 2000, which corresponds to the blue group.

(Go to next page)



In 1981, 713 games were canceled, and in 1994-1995, 948 games were canceled all because of MLB lockout. But, since these numbers are due to fewer games played and should not interfere with our linear regression analysis, we will leave them as they are.

Lastly, HBP does not seem to be correlated with win at all and is unavailable for many data points, so it will not be used in our analysis.

Our final *Teams* dataset looks as follows (showing the last six rows):

Table 4: Teams Dataset

	yearID	W	R	X1B	X2B	X3B	HR	BB	SO	SB	SH	RA	X123BA	HRA	BBA	SOA	E	group
2521	2019	77	678	839	300	26	167	475	1435	47	24	773	1168	227	519	1368	90	1991~2019
2522	2019	91	764	856	246	24	210	561	1420	116	40	662	1093	191	545	1399	66	1991~2019
2523	2019	96	769	890	291	29	217	542	1493	94	8	656	1093	181	453	1621	87	1991~2019
2524	2019	78	810	831	296	24	223	534	1578	131	17	878	1274	241	583	1379	105	1991~2019
2525	2019	67	726	761	270	21	247	509	1514	51	14	828	1222	228	604	1332	96	1991~2019
2526	2019	93	873	904	298	27	231	584	1308	116	48	724	1138	202	517	1511	87	1991~2019

(Go to next page)

Modeling

Before we start building our models, we will first leave out 30% of the data as the validation set for our final test. Remaining 70% (prediction set) is again divided into two groups: the training set (70%) and the test set (30%). These are used to find the best-performing model, and once we find the final model, it is again trained with the prediction set and tested with the validation set.

In terms of choosing an algorithm, the linear regression is the most intuitive choice, and we will try our first model with only the batting stats. Root Mean Square Error (RMSE) and adjusted R-squared are two metrics that will be used to compare results of the different models.

Table 5: RMSE Table 1

Variables	RMSE	Adjusted_R_squred
X1B, X2B, X3B, HR, BB, SO, SB, SH	10.75286	0.3764391

Below is the summary of the first model:

```
Call:
lm(formula = W ~ X1B + X2B + X3B + HR + BB + SO + SB + SH, data = train_set)

Residuals:
    Min      1Q  Median      3Q     Max
-35.874  -7.114   0.093   7.276  30.517

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.665009   4.552876   0.805   0.4210
X1B          0.028572   0.004220   6.771 1.98e-11 ***
X2B         -0.005281   0.008610  -0.613   0.5398
X3B          0.131435   0.023945   5.489 4.91e-08 ***
HR           0.159112   0.010663  14.922  < 2e-16 ***
BB           0.034514   0.003961   8.714  < 2e-16 ***
SO          -0.005835   0.001988  -2.934   0.0034 **
SB           0.062003   0.006813   9.101  < 2e-16 ***
SH           0.049902   0.010178   4.903 1.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.64 on 1226 degrees of freedom
Multiple R-squared:  0.3805,    Adjusted R-squared:  0.3764
F-statistic: 94.12 on 8 and 1226 DF,  p-value: < 2.2e-16
```

Let's try using both batting and pitching stats variables in our second model:

Table 6: RMSE Table 2

Variables	RMSE	Adjusted_R_squred
X1B, X2B, X3B, HR, BB, SO, SB, SH	10.752857	0.3764391
X1B, X2B, X3B, HR, BB, SO, SB, SH, X123BA, HRA, BBA, SOA, E	6.300027	0.7930611

```
Call:
lm(formula = W ~ X1B + X2B + X3B + HR + BB + SO + SB + SH + X123BA +
    HRA + BBA + SOA + E, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-25.8923  -4.1351   0.0381   3.9272  20.0969

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.269836   2.908388  13.158  < 2e-16 ***
X1B          0.081523   0.003037  26.840  < 2e-16 ***
X2B          0.048844   0.005757   8.485  < 2e-16 ***
X3B          0.131948   0.014613   9.030  < 2e-16 ***
HR           0.166052   0.006874  24.157  < 2e-16 ***
BB           0.050925   0.002485  20.491  < 2e-16 ***
SO           0.004816   0.001926   2.500 0.012539 *
SB           0.022670   0.004295   5.278 1.54e-07 ***
SH           0.022524   0.006118   3.682 0.000242 ***
X123BA      -0.056306   0.002595 -21.698  < 2e-16 ***
HRA         -0.156846   0.008083 -19.404  < 2e-16 ***
BBA         -0.041069   0.002528 -16.244  < 2e-16 ***
SOA          0.006346   0.001810   3.505 0.000473 ***
E           -0.040294   0.005843  -6.896 8.58e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.129 on 1221 degrees of freedom
Multiple R-squared:  0.7952,    Adjusted R-squared:  0.7931
F-statistic: 364.8 on 13 and 1221 DF,  p-value: < 2.2e-16
```

Note that RMSE decreased and adjusted R-squared increased significantly from the first model. The summary indicates that *SO* is not as significant as others, so we try the third model without *SO*.

Table 7: RMSE Table 3

Variables	RMSE	Adjusted_R_squred
X1B, X2B, X3B, HR, BB, SO, SB, SH	10.752857	0.3764391
X1B, X2B, X3B, HR, BB, SO, SB, SH, X123BA, HRA, BBA, SOA, E	6.300027	0.7930611
X1B, X2B, X3B, HR, BB, SB, SH, X123BA, HRA, BBA, SOA, E	6.285873	0.7921718

```
Call:
lm(formula = W ~ X1B + X2B + X3B + HR + BB + SB + SH + X123BA +
    HRA + BBA + SOA + E, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-26.0336  -4.1310   0.0726   4.0361  19.5606

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.539188   2.869882  13.777  < 2e-16 ***
X1B          0.077907   0.002677  29.106  < 2e-16 ***
X2B          0.046304   0.005678   8.154 8.62e-16 ***
X3B          0.133221   0.014635   9.103  < 2e-16 ***
HR           0.166891   0.006880  24.256  < 2e-16 ***
BB           0.050433   0.002483  20.313  < 2e-16 ***
SB           0.024763   0.004222   5.866 5.75e-09 ***
SH           0.021526   0.006118   3.518  0.00045 ***
X123BA      -0.053702   0.002382 -22.545  < 2e-16 ***
HRA         -0.151877   0.007852 -19.343  < 2e-16 ***
BBA         -0.040667   0.002528 -16.083  < 2e-16 ***
SOA          0.009785   0.001180   8.293 2.87e-16 ***
E           -0.040828   0.005852  -6.977 4.94e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.142 on 1222 degrees of freedom
Multiple R-squared:  0.7942,    Adjusted R-squared:  0.7922
F-statistic:   393 on 12 and 1222 DF,  p-value: < 2.2e-16
```

Both RMSE and adjusted R-squared are essentially the same as the second model, so we will keep *SO* out of the model.

For the purpose of experimenting, let's see how random forest performs, leaving the variables same as the third model.

Table 8: Linear Regression vs. Random Forest

Model	RMSE
Linear Regression	6.285873
Random Forest	7.762767

RMSE for random forest is bigger than that for linear regression, so we will keep using the linear regression for the remainder of this analysis.

Until now, we did not use *R* and *RA* in our model and thought that we should not do so because these are the products of batting and pitching stats, respectively. But, let's try putting them in our next model:

Table 9: RMSE Table 4

Variables	RMSE	Adjusted_R_squred
X1B, X2B, X3B, HR, BB, SO, SB, SH	10.752857	0.3764391
X1B, X2B, X3B, HR, BB, SO, SB, SH, X123BA, HRA, BBA, SOA, E	6.300027	0.7930611
X1B, X2B, X3B, HR, BB, SB, SH, X123BA, HRA, BBA, SOA, E	6.285873	0.7921718
R, RA, X1B, X2B, X3B, HR, BB, SB, SH, X123BA, HRA, BBA, SOA, E	4.679033	0.8901633

```
Call:
lm(formula = W ~ R + RA + X1B + X2B + X3B + HR + BB + SB + X123BA +
    HRA + BBA + SOA + E, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6171  -2.9321   0.0832   2.9665  15.8292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.3254887  2.2631743  10.748  < 2e-16 ***
R            0.0730598  0.0043847  16.662  < 2e-16 ***
RA          -0.1297030  0.0039736 -32.641  < 2e-16 ***
X1B          0.0218583  0.0030549   7.155 1.44e-12 ***
X2B          0.0196713  0.0059356   3.314 0.000946 ***
X3B          0.0543381  0.0119434   4.550 5.91e-06 ***
HR           0.0358602  0.0083330   4.303 1.82e-05 ***
BB           0.0157107  0.0023968   6.555 8.20e-11 ***
SB           0.0095104  0.0031876   2.984 0.002906 **
X123BA       0.0249359  0.0029222   8.533  < 2e-16 ***
HRA          0.0506665  0.0084403   6.003 2.55e-09 ***
BBA          0.0085193  0.0023630   3.605 0.000324 ***
SOA          0.0070899  0.0009479   7.480 1.42e-13 ***
E            0.0214356  0.0049251   4.352 1.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.465 on 1221 degrees of freedom
Multiple R-squared:  0.8913,    Adjusted R-squared:  0.8902
F-statistic: 770.3 on 13 and 1221 DF,  p-value: < 2.2e-16
```

RMSE and adjusted R-squared improved significantly from the previous model since, as mentioned previously, the run differential (R-RA) is very closely correlated with win. The question is "can we retain them in the final model?"

Our initial thought was that because the batting and pitching stats produces R and RA, respectively, both of which then results in winning, counting on all of batting/pitching stats and R/RA is duplicating and hence should not improve the model much. However, it turns out that including R/RA does improve our model a lot. Using a model like this, teams can predict their performance for the coming season, given predicted statistics for each of their players. Predicting players' future performance will be another very interesting research although it could be more complicated and less predictable.

- **Data Manipulation**

Before, we talked about how the plots show that certain variables increase/decrease over time.

Table 10: Home runs for different eras

group	group_mean
1901~1930	38.50417
1931~1960	99.63125
1961~1990	125.83053
1991~2019	166.79695

As shown in the plot on the left, there seem to be four groups of dots around each group's mean. This pattern can potentially be harmful to our analysis since for 1901~1960, teams won their games with fewer than 100 HR per year, but now, they are very unlikely to win that many games with the same number of HR. Winning formula has been changed, so using the old data as it is means that we try to predict future winning with how they used to win in the past.

To take this into account, we will make some adjustments to the dataset. On the right, we moved each point of earlier three year-groups towards the most recent '1991~2019' group by the mean difference between each group and '1991~2019'. The reason for moving towards the most recent is that at the end of the day, we are interested in predicting future, so the data most relevant to the recent days would be the most useful.

After adjusting all variables, we then fit our final model. Note that the prediction set and the validation set should be recreated because 'Teams' dataset is modified to 'New_Teams'.

Unfortunately, we did not see much difference. However, it makes more logical sense to adjust the data points towards today's standards, we will choose the 'data manipulated model' as our final model.

Table 11: Two Models with 'R,RA,X1B,X2B,X3B,HR,BB,SB,X123BA,HRA,BBA,SOA,E

Model	RMSE	Adjusted_R_Squared
Original Dataset	4.679033	0.8901633
Manipulated Dataset	4.440727	0.8839167

Results

We now train our model using the prediction set, which is 70% of the entire dataset and is the sum of training and test set used for modeling. Afterwards, this model is tested on the validation set we kept out of our analysis from the beginning. The result is as follows:

Table 12: Results on the validation set

Variables	RMSE	Adjusted_R_Squared
R, RA, X1B, X2B, X3B, HR, BB, SB, SH, X123BA, HRA, BBA, SOA, E(Data Manipulated)	4.580483	0.8861496

```
Call:
lm(formula = W ~ R + RA + X1B + X2B + X3B + HR + BB + SB + X123BA +
    HRA + BBA + SOA + E, data = prediction_set)

Residuals:
     Min       1Q   Median       3Q      Max
-16.2629  -3.0931   0.0536   2.9912  15.9876

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.8105750  1.7498132  11.322  < 2e-16 ***
R            0.0727488  0.0037312  19.497  < 2e-16 ***
RA          -0.1250069  0.0033412 -37.414  < 2e-16 ***
X1B          0.0265279  0.0026129  10.153  < 2e-16 ***
X2B          0.0090860  0.0047999   1.893 0.058525 .
X3B          0.0630751  0.0102362   6.162 8.89e-10 ***
HR           0.0427422  0.0070629   6.052 1.75e-09 ***
BB           0.0157251  0.0020690   7.600 4.78e-14 ***
SB           0.0171354  0.0026264   6.524 8.91e-11 ***
X123BA       0.0258158  0.0025554  10.102  < 2e-16 ***
HRA          0.0350162  0.0065972   5.308 1.25e-07 ***
BBA          0.0075128  0.0020380   3.686 0.000234 ***
SOA          0.0079425  0.0008241   9.638  < 2e-16 ***
E            0.0096612  0.0043372   2.228 0.026038 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.59 on 1753 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8861
F-statistic:  1058 on 13 and 1753 DF,  p-value: < 2.2e-16
```

The result is consistent with our analysis, and it seems that given all the variables contained in the model, our model can predict a team's number of wins for a season with ± 4.6 wins.

Conclusion

Although the final result was not too bad, there are limitations in our analysis.

1. The use of 30-year interval for the year groups was arbitrary. We could try different intervals to validate the optimal choice.
2. Some of the 19 variables initially retained from *Teams* dataset were chosen somewhat subjectively with prior baseball knowledge. It may be possible that a not-retained variable turns out to be suitable for our model.
3. Baseball has frequently made rule changes, some of which are more radical than others. Though we adjusted our data to account for the mean differences among different periods of time, baseball now is very different from 100 years ago, or even 20 years ago. Using data that old to build a prediction model for today's standards could be inappropriate.

Also, if you are a baseball fan, this analysis is not that fascinating, considering so many interesting research that has been going on, especially after *Moneyball* emerged. As my first ever machine learning project from scratch, I wanted to play around with the most accessible and clean dataset in order to focus more on applying what I learned throughout the courses.

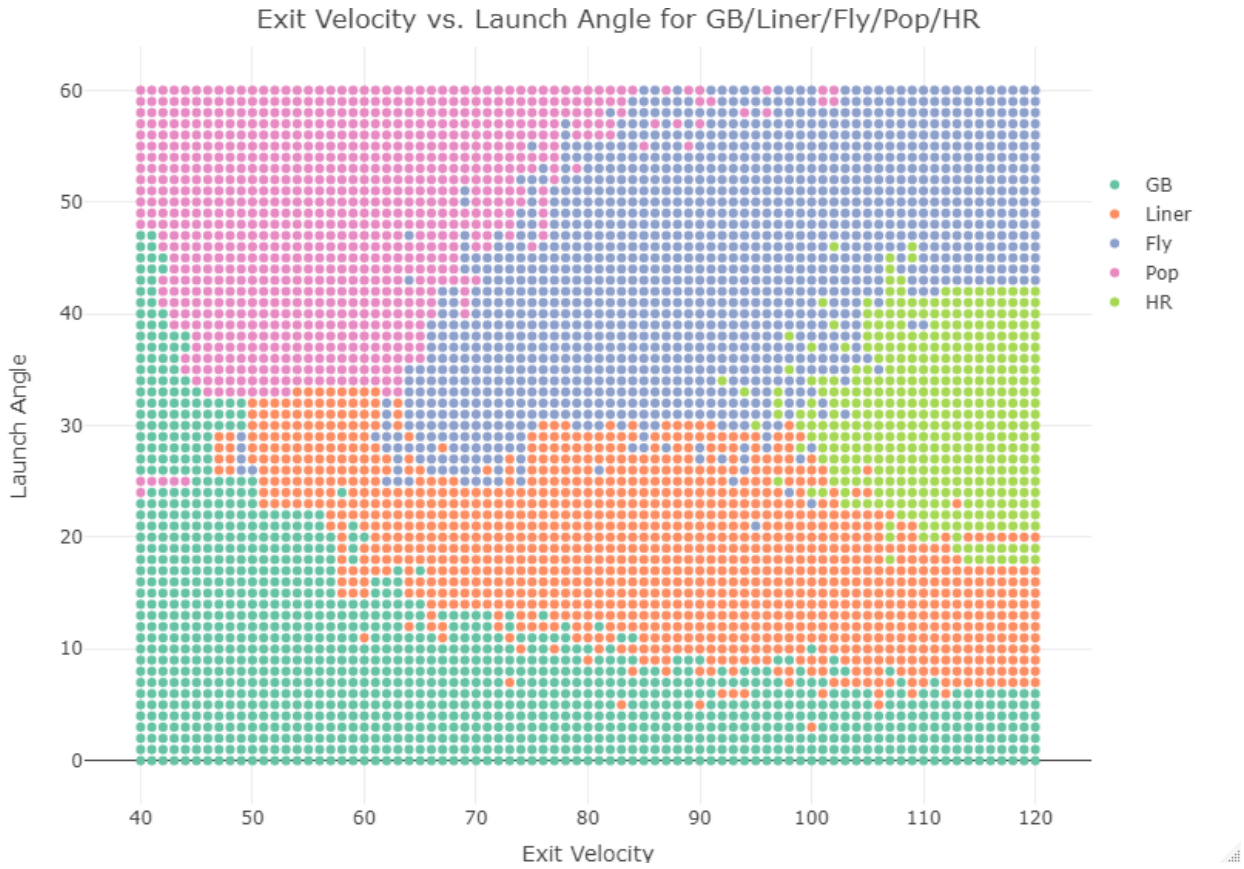
To make it more interesting and to get a glimpse of what kind of analysis can be done next, the following is the random forest prediction grid for outcomes of batted balls. The data are provided by *Statcast* on *Baseball Savant* website and consist of all exit velocities and launch angles of batted balls in 2020. This data contains 43,309 rows, which is more than 15 times of *New Teams* dataset, and MLB teams in 2020 played only 60 games as opposed to 162 because of Covid-19, so there is a ton of data to dive into for baseball. (The details of data collection and analysis are not shown here for the sake of not complicating this report)

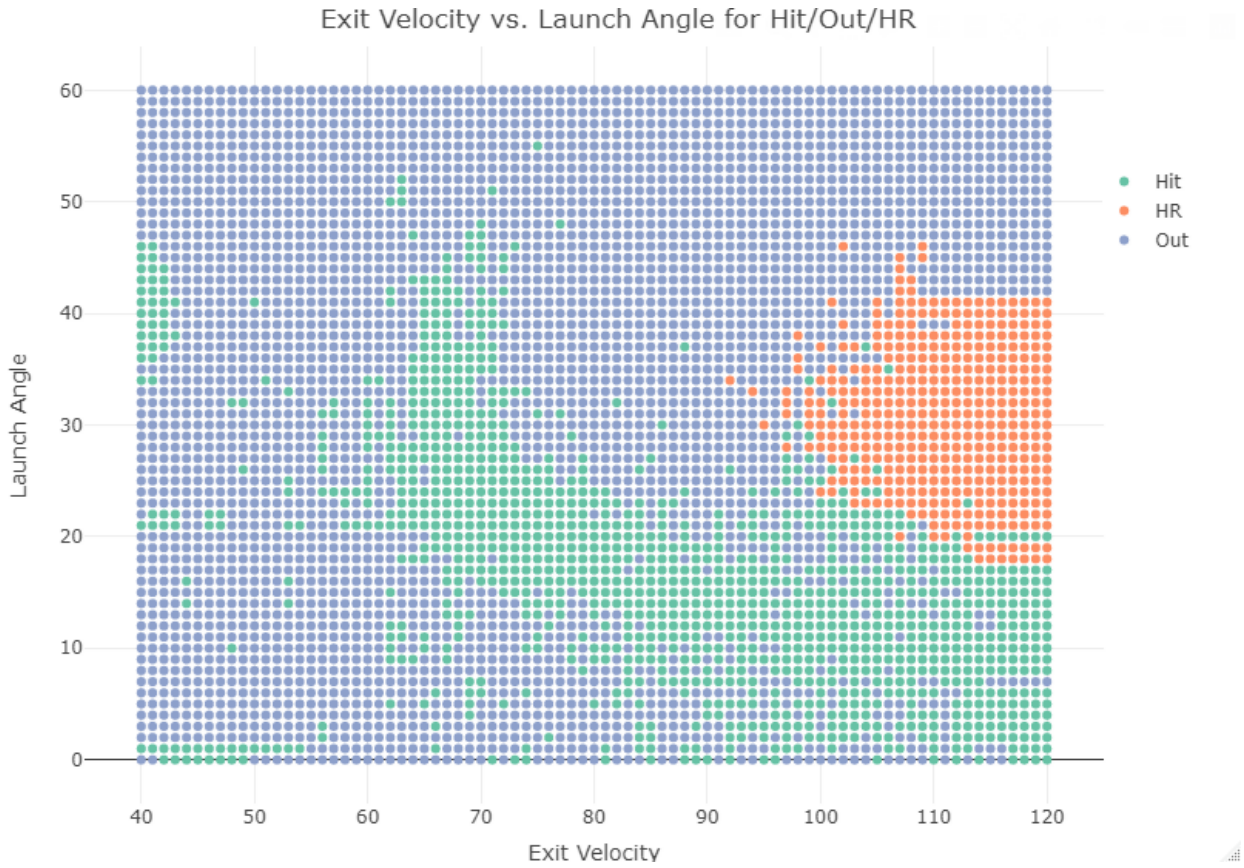
(Figures on the next page)

No wonder why exit velocities and launch angles are discussed so much in the media these days. It is obvious that a certain combination of exit velocity and launch angle is destined to produce a certain outcome. If we can make the ball pop out of the bat in the range of above 90 mph exit velocity and $10 \sim 25$ launch angle, it is very likely to induce a hit or a home run. Furthermore, the exit velocity of above 110 mph will produce such outcomes regardless of the launch angle.

This kind of analysis would be very useful in predicting each player's statistics, which then can be used to predict the team's statistics as we did in this report. The possible areas of research in baseball are so vast that what we saw in this report is only a tip of the iceberg. But still, the analysis provided in this report is useful for those who are baseball fans and are beginners of machine learning.

Just yesterday, May 19th, 2021, MLB witnessed sixth no-hitter of the season. This is very interesting because MLB's record for a single season's no-hitter is seven, and this season is only about 25% finished. These days, pitchers and batters incorporate data more and more, and it will be fun to dig deep into what is causing today's trend using data science.


Exit Velocity vs. Launch Angle for GB/Liner/Fly/Pop/HR
Launch Angle
Exit Velocity
0
10
20
30
40
50
60
40
50
60
70
80
90
100
110
120
GB
Liner
Fly
Pop
HR


Exit Velocity vs. Launch Angle for Hit/Out/HR
Launch Angle
Exit Velocity
0
10
20
30
40
50
60
40
50
60
70
80
90
100
110
120
Hit
HR
Out