# Homework 1

#### Group 1

#### **Contents**

1	Introduction	1
2	Statement of the Problem	1
3	Data Exploration	1
4	Data Preparation	4
5	Build Models	5
6	Select Models	6
7	Appendix A 7.1 Data Dictionary	<b>6</b> 6 6

### 1 Introduction

The ability to analyize and predict performance of a team using many dimensions is critical to competitive success in the professional Baseball industry. Therefore, we analyized the records of numerous professional baseball team from the years 1871 to 2006. Our hope is that the following report and the resulting predictive models will better inform the organization and assist in making data driven decisions moving forward.

"The goal of a baseball team is to win more games than any other team. Since one team has very little control over the number of games other teams win, the goal is essentially to win as many games as possible. Therefore, it is of interest to measure the player's contribution to the team's wins." Grabiner, B. D. 1

#### 2 Statement of the Problem

The purpose of this report is to determine the batting, baserun, pitching, and fielding effects on a baseball team's ability to win.

# 3 Data Exploration

The following table provides the descriptive statistics regarding our data set. You will find that

Please note that each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

<sup>&</sup>lt;sup>1</sup>(Grabiner, B. D. (n.d.). The Sabermetric Manifesto. Retrieved September 10, 2016 from http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/)

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

Descrip	otive St	atistics
16 Variables	2276	Observations

Number of wins						io var	iabie	s Z	2/6 U	oserva	ations	
2276	Numbe	er of wins	;									
Base Hits by batters (1B,2B,3B,HR)						.05 54.0		.25 71.0				
Property   1891   1892   1009   1116   1129, highest: 2333   2372   2496   2554   255   250   255   250   255   250   255   250   255   250   255   250   255	lowest	: 0 12	2 14 17	21, h	ighest:	128 129	134 1	135 146				
2276	Base H	lits by ba	tters (1B	,2B,3B	,HR)							
Doubles by batters (2B)   10   10   10   10   10   10   10   10						.05 1282	.10 1315	.25 1383	.50 1454			
n missing unique Info Mean 0.5 1.0 2.5 5.0 .75 .90 .95  Triples by batters (3B)  2276 0 8 9 11 12, highest: 838 392 393 403 458  Triples by batters (3B)  2276 0 8 9 11 12, highest: 166 190 197 200 223  Homeruns by batters (4B)  2276 0 8 9 11 12, highest: 166 190 197 200 223  Homeruns by batters (4B)  2276 0 8 9 11 12, highest: 247 249 257 260 264  Walks by batters  2276 0 8 10 8 9 10 10 8 10 10 8 10 10 8 10 10 10 10 10 10 10 10 10 10 10 10 10	lowest	: 891 9	992 1009	1116 11	22, high	nest: 23	33 234	13 2372	2496 2	554		
2276   0	Double	es by batt	ers (2B)									
Triples by batters (3B)  2276 missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  Homeruns by batters (4B)  2276 missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  Homeruns by batters (4B)  2276 missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  2276 missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  2276 missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  2276 missing unique .243 .1 .50 .248 .2 .363.5 .451.0 .512.0 .580.0 .635.0 .670.2  2276 missing unique .248 .2 .363.5 .451.0 .512.0 .580.0 .635.0 .670.2  2276 missing unique .248 .2 .363.5 .451.0 .512.0 .580.0 .635.0 .670.2  2276 missing unique .248 .2 .363.5 .451.0 .512.0 .580.0 .635.0 .670.2  2276 missing unique .248 .2 .363.5 .451.0 .512.0 .580.0 .635.0 .670.2  2276 missing unique .248 .2 .363.5 .451.0 .512.0 .580.0 .635.0 .670.2  2274 missing unique .248 .2 .363.5 .364.0 .666 .878  Strikeouts by batters  1 missing unique .25 .10 .25 .50 .75 .90 .95  2174 missing unique .25 .27 .74, highest: 1303 1320 1326 1335 1399  Stolen bases  1 missing unique .25 .27 .27, highest: .252 .56 .632 .654 .697  Caught stealing  1 missing unique .25 .25 .25 .25 .25 .250 .75 .90 .95  1 528 .24 .30 .38 .49 .62 .77 .91  2085 .55 .1 .52.8 .24 .30 .38 .49 .62 .77 .91  2085 .55 .1 .52.8 .24 .30 .38 .49 .62 .77 .91  2086 .55 .1 .25 .350 .44 .0 .50.5 .50 .75 .90 .95  1 missing unique .25 .15 .25 .50 .75 .90 .95  1 528 .24 .30 .38 .49 .62 .77 .91  2085 .55 .1 .5 .50 .35 .40 .0 .55 .50 .75 .90 .95  1 missing unique .25 .15 .30 .30 .80  2276 .30 .35 .38 .9, highest: .87 .88 .89 .90 .95  Hits allowed  2276 .843 .1168 .1184 .1187 .1202												
n missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  Homeruns by batters (4B)  2276 missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  2278 missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  2278 missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  2279 missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  2270 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2276 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2276 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2276 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2276 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2276 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2277 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2278 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2279 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2270 missing unique lnfo Me	lowest	: 69 112	2 113 118	123, h	ighest:	382 392	393 4	103 458				
Homeruns by batters (4B)  276    missing   unique   Info   Mean   0.5   10   2.5   5.0   7.5   90   9.5   9.	Triples	by batte	rs (3B)									au III I III III III III III III III III
Homeruns by batters (4B)  2276 missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  2276 missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2276 0 \$3 4 5 6, highest: 247 249 257 260 264  Walks by batters  n missing unique lnfo Mean 0.5 .10 25 .50 .75 .90 .95  2276 0 \$33 1 501.6 248.2 363.5 451.0 512.0 580.0 635.0 670.2  10vest: 0 12 29 34 45, highest: 815 819 824 860 878  Strikeouts by batters  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2174 102 822 1 735.6 359 421 548 750 930 1049 1103  10vest: 0 66 67 72 74, highest: 1303 1320 1326 1335 1399  Stolen bases  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2145 131 348 1 124.8 35.0 44.0 66.0 101.0 156.0 231.0 301.8  10vest: 0 14 18 19 20, highest: 562 567 632 654 697  Caught stealing  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  1504 772 128 1 52.8 24 30 38 49 62 77 91  10vest: 0 7 11 12 14, highest: 171 186 193 200 201  Batters hit by pitch (get a free base)  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  10vest: 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  10vest: 29 30 35 38 39, highest: 87 88 89 90 95		missing 0				.05 23	.10 27	.25 34	50 .7: 47 7:			
2276 0 0 243 1 99.61 14.0 20.0 42.0 102.0 147.0 179.5 199.0  lowest : 0 3 4 5 6, highest: 247 249 257 260 264  Walks by batters  n missing unique 276 0 533 1 501.6 248.2 363.5 451.0 512.0 580.0 635.0 670.2  lowest : 0 12 29 34 45, highest: 815 819 824 860 878  Strikeouts by batters  n missing unique 2774 1102 822 1 735.6 359 421 548 750 930 1049 1103  lowest : 0 66 67 72 74, highest: 1303 1320 1326 1335 1399  Stolen bases  n missing unique 2145 133 348 1 124.8 35.0 44.0 66.0 101.0 156.0 231.0 301.8  lowest : 0 14 18 19 20, highest: 562 567 632 654 697  Caught stealing  n missing unique 1752 24 30 38 49 62 77 91  lowest : 0 7 11 12 14, highest: 171 186 193 200 201  Batters hit by pitch (get a free base)  n missing unique 1760 Mean 0.05 .10 .25 .50 .75 .90 .95 10 .25 .50 .75 .90 .95  lowest : 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique unique 1779 1316 1356 1419 1518 1682 2058 2563					ighest:	166 190	197 2	200 223				
276	Homer	uns by ba	atters (4	В)								adddddinaanaadddddddddddddaaaaaaaaa
Walks by batters  n missing unique lnfo Mean 0.5 10 25 .50 .75 .90 .95  Strikeouts by batters  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  Strikeouts by batters  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  Stolen bases  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  Stolen bases  n missing unique lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2145 131 348 1 124.8 35.0 44.0 66.0 101.0 156.0 231.0 301.8  Stolen the total line lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2145 131 348 1 124.8 35.0 44.0 66.0 101.0 156.0 231.0 301.8  Stolen the total lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2145 131 348 1 124.8 35.0 44.0 66.0 101.0 156.0 231.0 301.8  Stolen the total lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  2145 131 348 1 124.8 35.0 44.0 66.0 101.0 156.0 231.0 301.8  Stolen the total lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  Stolen the total lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  Stolen the total lnfo Mean 0.5 .10 .25 .50 .75 .90 .95  Stolen the lnfo Mean 0.5 .10 .25 .50 .75	2276	Ō	243	1	99.61	14.0	20.0	42.0				
n missing unique lnfo Mean 05 10 25 50 75 90 95  Strikeouts by batters  n missing unique lnfo Mean 05 10 25 50 75 90 95  Stolen bases  n missing unique lnfo Mean 05 10 25 50 75 90 95  Stolen bases  n missing unique lnfo Mean 05 10 25 50 75 90 95  Taylowest: 0 14 18 19 20, highest: 562 567 632 654 697  Caught stealing  n missing unique lnfo Mean 05 10 25 50 75 90 95  Towest: 0 7 11 12 14, highest: 171 186 193 200 201  Batters hit by pitch (get a free base)  n missing unique lnfo Mean 05 10 25 50 75 90 95  191 2085 55 1 59,36 40.0 44.0 50.5 58.0 67.0 76.0 82.5  lowest: 1137 1168 1184 1187 1202  lowest: 1137 1168 1184 1187 1202  lowest: 1137 1168 1184 1187 1202				6, h	ighest:	247 249	257 2	260 264				
2776	Walks	by batter	S									
Strikeouts by batters  In missing unique 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1												
n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  Stolen bases  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 2145	lowest	: 0 12	2 29 34	45, h	ighest:	815 819	824 8	360 878				
2174 10Ž 822 1 735.6 359 421 548 750 930 1049 1103  lowest: 0 66 67 72 74, highest: 1303 1320 1326 1335 1399  Stolen bases	Striked	outs by ba	atters									
Stolen bases  2145 2145 2131 2348 21 124.8 25.0 214.0 25 .50 231.0 231.0 231.0 301.8  20west: 0 14 18 19 20, highest: 562 567 632 654 697  Caught stealing  225 .50 .75 .90 .95 231.0 301.8  231.0 301.8												
n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95 lowest : 0 14 18 19 20, highest: 562 567 632 654 697  Caught stealing  n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95 1504 772 128 1 52.8 24 30 38 49 62 77 91  lowest : 0 7 11 12 14, highest: 171 186 193 200 201  Batters hit by pitch (get a free base)  n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95 191 2085 55 1 59.36 40.0 44.0 50.5 58.0 67.0 76.0 82.5  Hits allowed  n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95 2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563  lowest : 1137 1168 1184 1187 1202	lowest	: 0	66 67	72	74, high	nest: 13	03 132	20 1326	1335 1	399		
2145	Stolen	bases										
Caught stealing												
n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95  lowest: 0 7 11 12 14, highest: 171 186 193 200 201  Batters hit by pitch (get a free base)  n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95  191 2085 55 1 59.36 40.0 44.0 50.5 58.0 67.0 76.0 82.5  lowest: 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique lnfo Mean .05 .10 .25 .50 .75 .90 .95  2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563  lowest: 1137 1168 1184 1187 1202	lowest	: 0 14	18 19	20, h	ighest:	562 567	632 6	697				
1504 77Ž 128 1 52.8 24 30 38 49 62 77 91  lowest: 0 7 11 12 14, highest: 171 186 193 200 201  Batters hit by pitch (get a free base)  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 191 2085 55 1 59.36 40.0 44.0 50.5 58.0 67.0 76.0 82.5  lowest: 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563  lowest: 1137 1168 1184 1187 1202	Caught	t stealing	I									
Batters hit by pitch (get a free base)  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  lowest : 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95  2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563									50 .75 49 62	5 .90 2 77		
n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 191 2085 55 1 59.36 40.0 44.0 50.5 58.0 67.0 76.0 82.5  lowest : 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563  lowest : 1137 1168 1184 1187 1202	lowest	: 0 7	11 12	14, h	ighest:	171 186	193 2	200 201				
191 2085 55 1 59.36 40.0 44.0 50.5 58.0 67.0 76.0 82.5  lowest : 29 30 35 38 39, highest: 87 88 89 90 95  Hits allowed  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563  lowest : 1137 1168 1184 1187 1202	Batters	s hit by pi	itch (get	a free	base)							a maaraddddiddddaalaaa
Hits allowed  n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 .2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563  lowest : 1137 1168 1184 1187 1202												
n missing unique Info Mean .05 .10 .25 .50 .75 .90 .95 2276 0 843 1 1779 1316 1356 1419 1518 1682 2058 2563	lowest	: 29 30 3	35 38 39,	highes	t: 87 88	89 90	95					
2276	Hits all	lowed										<b></b>
lowest: 1137 1168 1184 1187 1202 highest: 16038 16871 20088 24057 30132			unique 843									
	lowest highest	: 1137 :: 16038 1	1168 11 6871 200	84 118 88 2405	7 1202 7 30132							

Homeruns all	owed								الالاأالد	ataushtdutllulldtiidlistassassassassassassassassassassassassas
n missir 2276	g unique 0 256		Mean .05 105.7 18.0	.10 25.0	.25 50.0	.50 107.0	.75 150.0		.95 )9.2	
lowest : 0	3 4 5	6, high	nest: 291 29	7 301 320	343					
Walks allowed	l									<b>I</b>
n missir 2276	g unique 0 535	Info 1	Mean .05 553 377.0		.25 476.0	.50 536.5	.75 611.0		.95 757.0	
lowest : 0	119 124	131 140	, highest: 2	169 2396	2840 28	76 3645				
Strikeouts by	pitchers									
n missin 2174 10			Mean .05 817.7 421.3		.25 615.0	.50 813.5	.75 968.0		.95 1173.0	
lowest: 0 highest: 3450		05 208 56 12758 1	252 19278							
Errors									ساال	
n missir 2276	g unique 0 549		Mean .05 246.5 100.0			.50 159.0	.75 249.2		.95 716.0	
lowest : 65	66 68	72 74	, highest: 1	567 1728	1740 18	90 1898				
Double Plays										
n missir 1990 28			Mean .05 146.4 98	.10 .2 109 1	25 .50 31 149		.90 178	.95 186		
lowest : 52	64 68 71	72, high	nest: 215 21	8 219 225	5 228					

- a. Mean / Standard Deviation / Median
- b. Bar Chart or Box Plot of the data
- c. Is the data correlated to the target variable (or to other variables?)d. Are any of the variables missing and need to be imputed "fixed"?

### 4 Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

### 5 Build Models

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

### 6 Select Models

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2 , RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2 , (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

# 7 Appendix A

### 7.1 Data Dictionary

VARIABLE.NAME	DEFINITION	THEORETICAL.EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	NA
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

#### 7.2 R code used in document