Homework #1: Baseball Analysis

Data 621 Business Analytics and Data Mining

Aadi Kalloo, Nathan Lim, Asher Meyers, Daniel Smilowitz, Logan Thomson

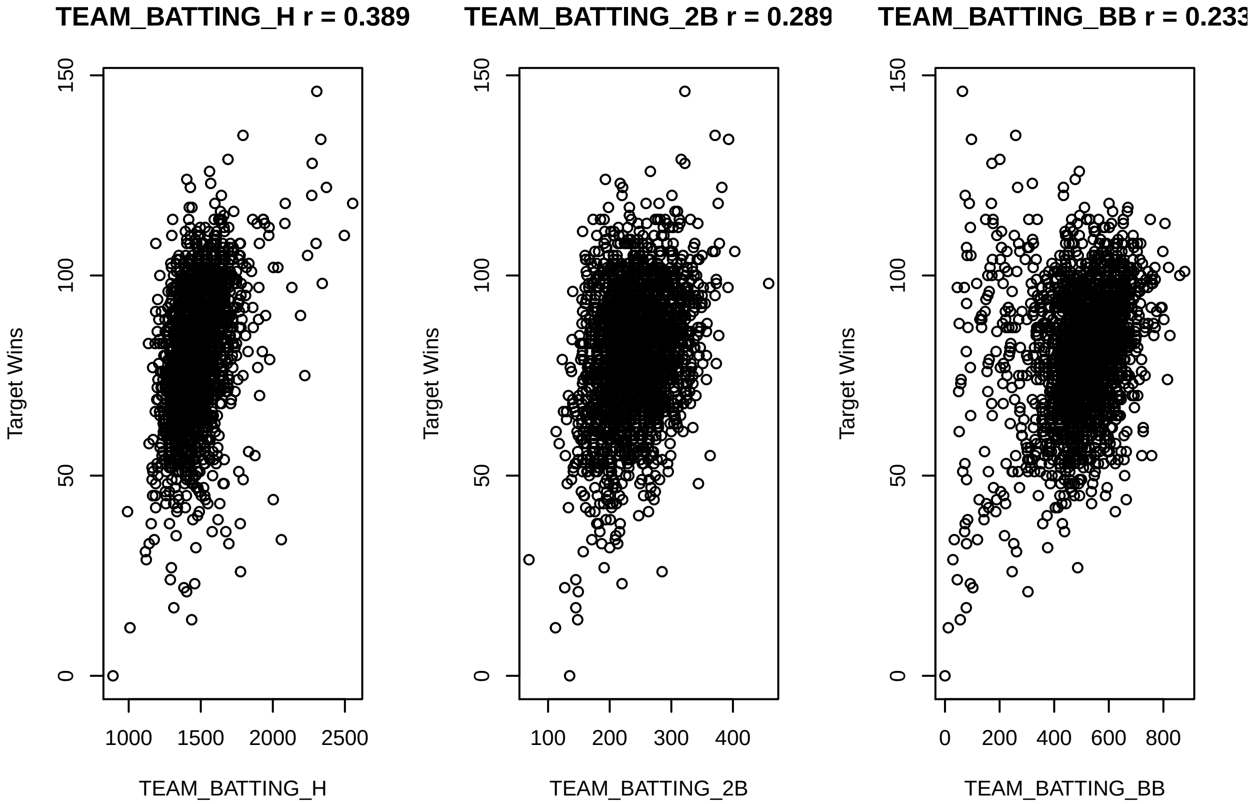
Due June 19, 2016

Data Exploration

The data analyzed in this report includes 2276 professional baseball teams for the years 1871-2006. In total, 16 variables were present in the data provided. Included below is a summary of descriptive statistics, correlations to wins, and the number of missing values for each variable in the provided data set:

	VAR_NAME	MEAN	MEDIAN	CORRELATION TO WINS (r)	NUM_MISSING
2	TARGET_WINS	80.79086	82.0	NA	NA
1	TEAM_BASERUN_CS	52.80386	49.0	0.0224041	772
21	TEAM_BASERUN_SB	124.76177	101.0	0.1351389	131
3	TEAM_BATTING_2B	241.24692	238.0	0.2891036	0
4	TEAM_BATTING_3B	55.25000	47.0	0.1426084	0
5	TEAM_BATTING_BB	501.55888	512.0	0.2325599	0
6	TEAM_BATTING_H	1469.26977	1454.0	0.3887675	0
7	TEAM_BATTING_HBP	59.35602	58.0	0.0735042	2085
8	TEAM_BATTING_HR	99.61204	102.0	0.1761532	0
9	TEAM_BATTING_SO	735.60534	750.0	-0.0317507	102
10	TEAM_FIELDING_DP	146.38794	149.0	-0.0348506	286
11	TEAM_FIELDING_E	246.48067	159.0	-0.1764848	0
12	TEAM_PITCHING_BB	553.00791	536.5	0.1241745	0
13	TEAM_PITCHING_H	1779.21046	1518.0	-0.1099371	0
14	TEAM_PITCHING_HR	105.69859	107.0	0.1890137	0
15	TEAM_PITCHING_SO	817.73045	813.5	-0.0784361	102

Below are graphs that show the relationship to *Target Wins* for the three variables with the highest correlation coefficient:



The full array of correlations graphs may be found in Appendix A.

Data Preparation

It was determined that the *Hits By Pitch* variable had too many missing values to be useful for regression, and thus this variable was excluded from the model building process.

Model Creation

Model Selection and Prediction