CUNY SPS DATA 621 - CTG5 - HW1

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

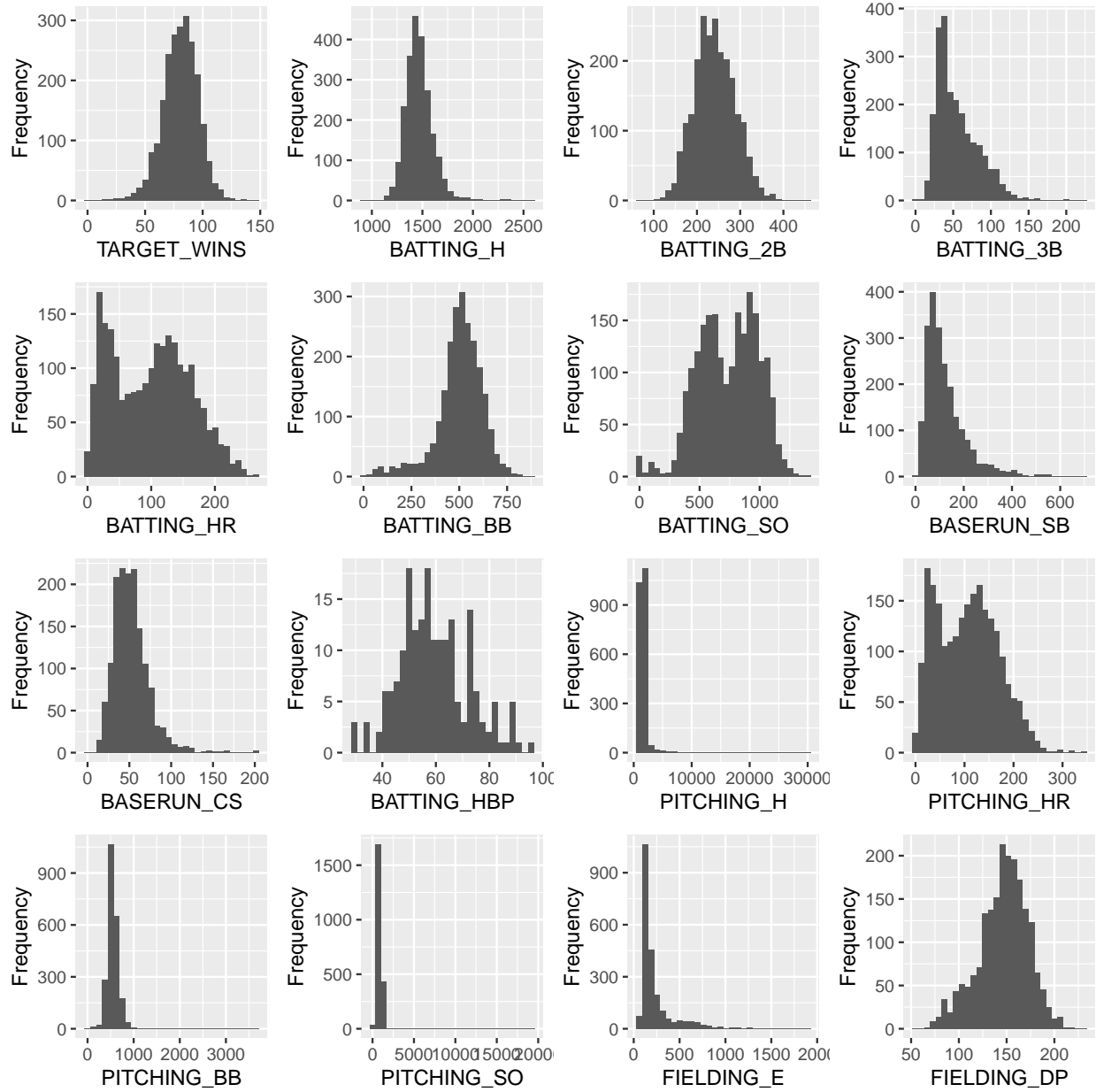*February 27, 2019*

Contents

1 Data exploration

- Describe the size:

The money ball data is 144kb in size. The data contains 2,276 rows and 16 columns without the index. The variables are continuous integer. The `TARGET_WINS` is our response variable. There are 3,478 missing values out of 36,416 observations.
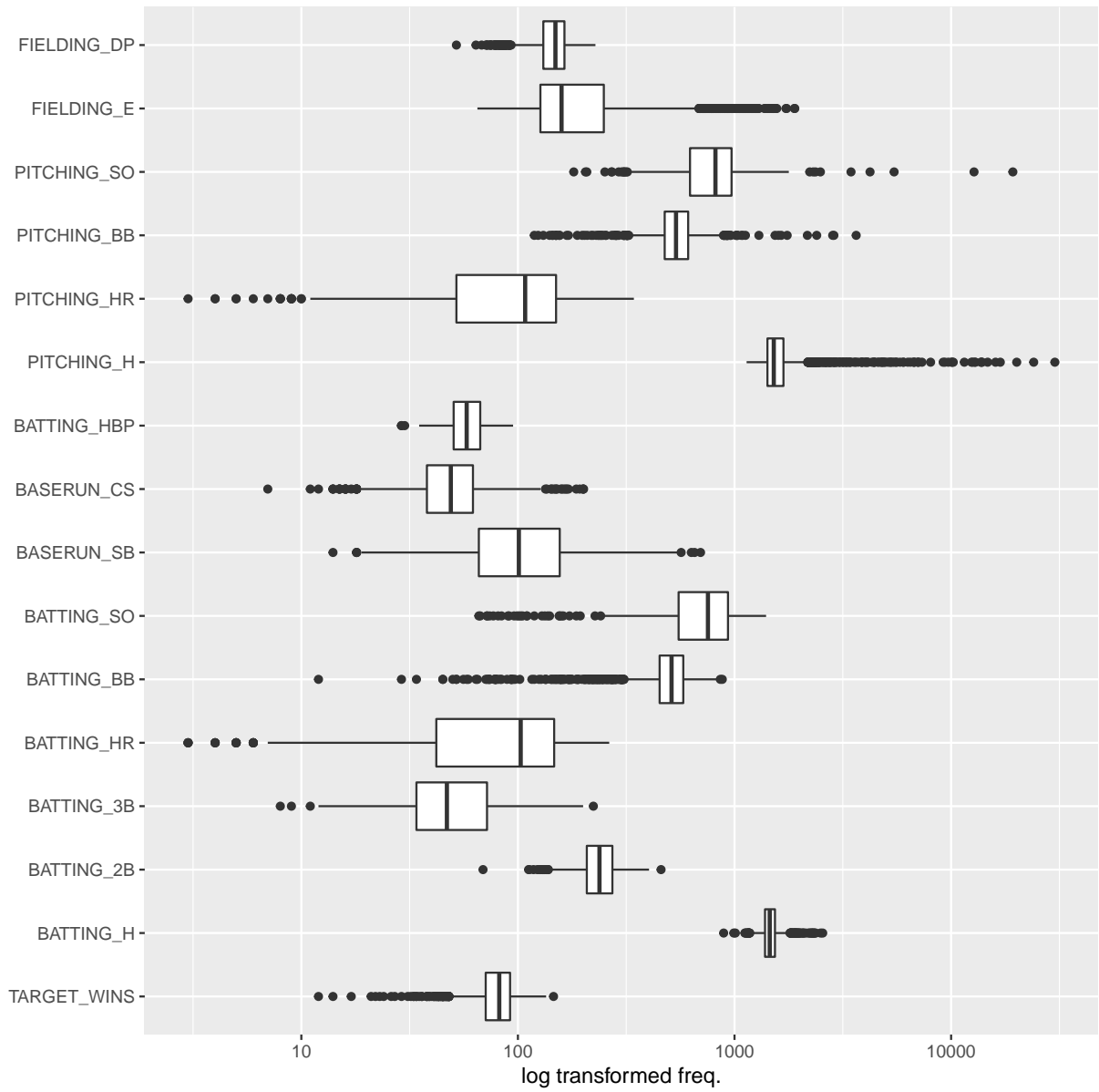
- Statistics summary

	var	s	n	mean	sd med	ian	trimmed	mad mi	n m	ax ran	ge	sk
TARGET_WINS	1	2276	80.79086	15.75215	82.0	81.31229	14.8260	0	146	146	-0.39
BATTING_H	2	2276	1469.26977	144.59120	1454.0	1459.04116	114.1602	891	2554	1663	1.57
BATTING_2B	3	2276	241.24692	46.80141	238.0	240.39627	47.4432	69	458	389	0.21
BATTING_3B	4	2276	55.25000	27.93856	47.0	52.17563	23.7216	0	223	223	1.10
BATTING_HR	5	2276	99.61204	60.54687	102.0	97.38529	78.5778	0	264	264	0.18
BATTING_BB	6	2276	501.55888	122.67086	512.0	512.18331	94.8864	0	878	878	-1.02
BATTING_SO	7	2174	735.60534	248.52642	750.0	742.31322	284.6592	0	1399	1399	-0.29
BASERUN_SB	8	2145	124.76177	87.79117	101.0	110.81188	60.7866	0	697	697	1.97
BASERUN_CS	9	1504	52.80386	22.95634	49.0	50.35963	17.7912	0	201	201	1.97
BATTING_HBP	10	191	59.35602	12.96712	58.0	58.86275	11.8608	29	95	66	0.31
PITCHING_H	11	2276	1779.21046	1406.84293	1518.0	1555.89517	174.9468	1137	30132	28995	10.32
PITCHING_HR	12	2276	105.69859	61.29875	107.0	103.15697	74.1300	0	343	343	0.28
PITCHING_BB	13	2276	553.00791	166.35736	536.5	542.62459	98.5929	0	3645	3645	6.74
PITCHING_SO	14	2174	817.73045	553.08503	813.5	796.93391	257.2311	0	19278	19278	22.17
FIELDING_E	15	2276	246.48067	227.77097	159.0	193.43798	62.2692	65	1898	1833	2.99
FIELDING_DP	16	1990	146.38794	26.22639	149.0	147.57789	23.7216	52	228	176	-0.38

- Data visualization

Boxplot

log transformed freq.

Scatterplot

Correrlation

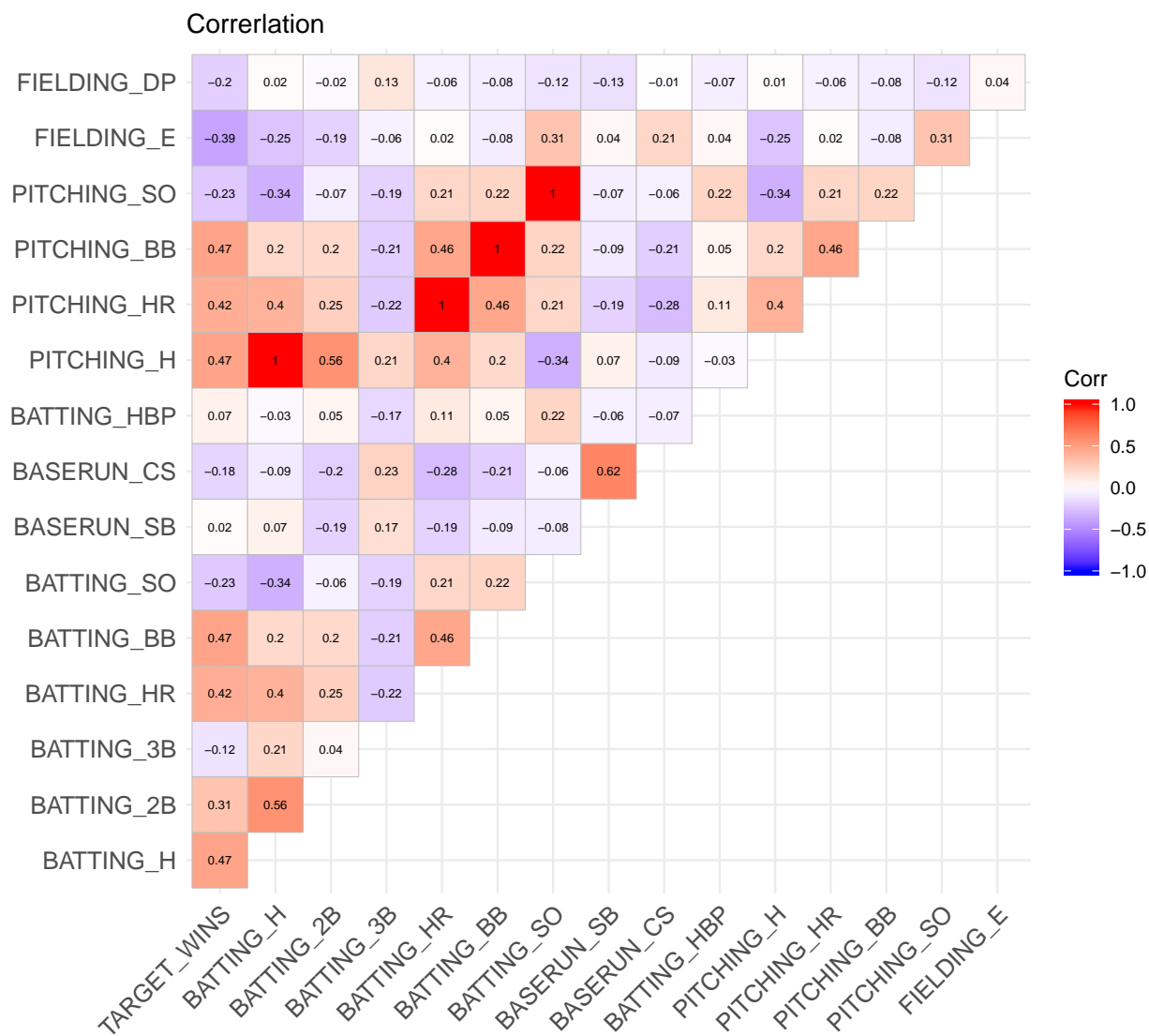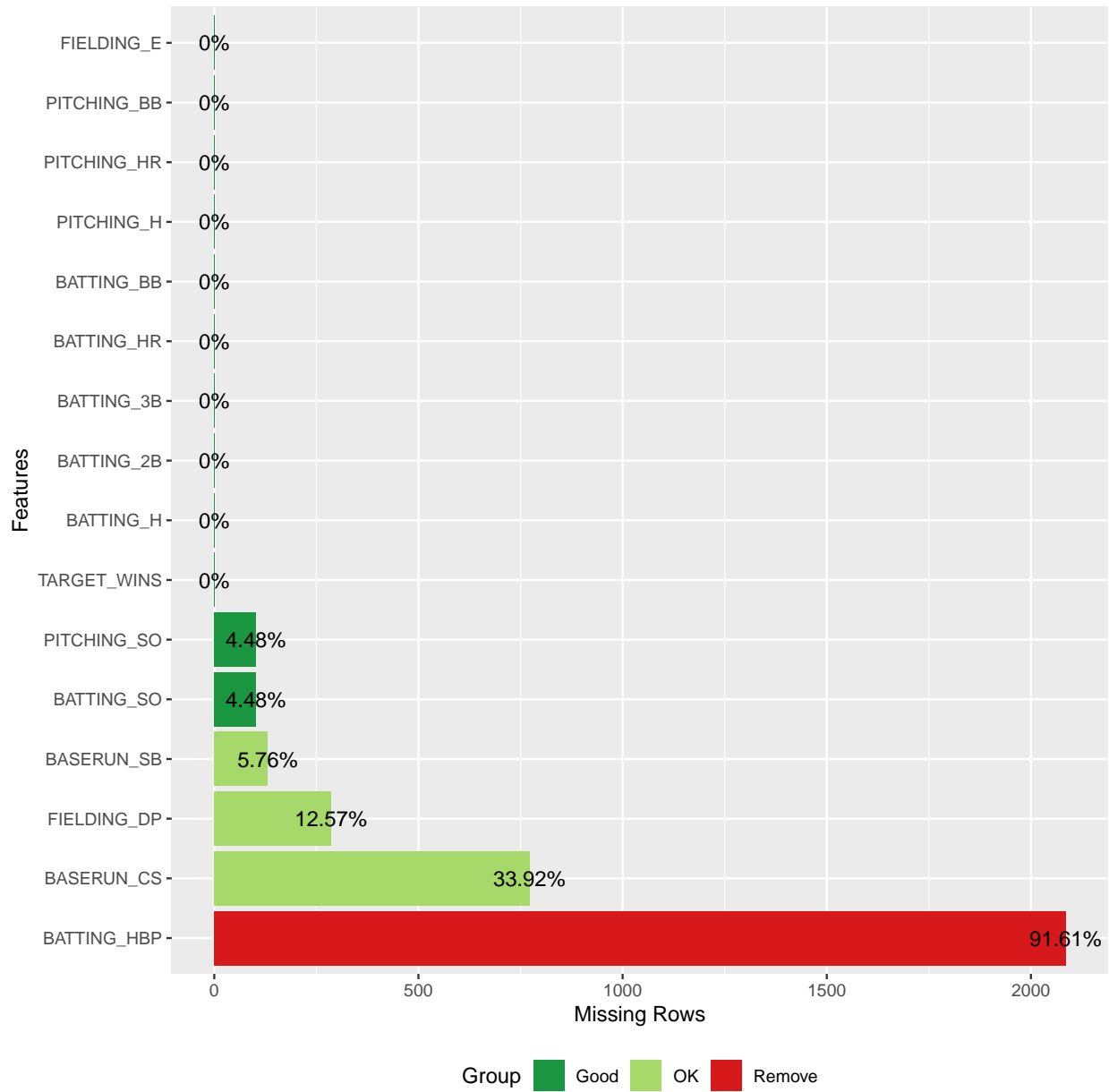|	TARGET_WINS	BATTING_H	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN_SB	BASERUN_CS	BATTING_HBP	PITCHING_H	PITCHING_HR	PITCHING_BB	PITCHING_SO	FIELDING_E
FIELDING_DP	-0.2	0.02	-0.02	0.13	-0.06	-0.08	-0.12	-0.13	-0.01	-0.07	0.01	-0.06	-0.08	-0.12	0.04
FIELDING_E	-0.39	-0.25	-0.19	-0.06	0.02	-0.08	0.31	0.04	0.21	0.04	-0.25	0.02	-0.08	0.31	
PITCHING_SO	-0.23	-0.34	-0.07	-0.19	0.21	0.22	1	-0.07	-0.06	0.22	-0.34	0.21	0.22		
PITCHING_BB	0.47	0.2	0.2	-0.21	0.46	1	0.22	-0.09	-0.21	0.05	0.2	0.46			
PITCHING_HR	0.42	0.4	0.25	-0.22	1	0.46	0.21	-0.19	-0.28	0.11	0.4				
PITCHING_H	0.47	1	0.56	0.21	0.4	0.2	-0.34	0.07	-0.09	-0.03					
BATTING_HBP	0.07	-0.03	0.05	-0.17	0.11	0.05	0.22	-0.06	-0.07						
BASERUN_CS	-0.18	-0.09	-0.2	0.23	-0.28	-0.21	-0.06	0.62							
BASERUN_SB	0.02	0.07	-0.19	0.17	-0.19	-0.09	-0.08								
BATTING_SO	-0.23	-0.34	-0.06	-0.19	0.21	0.22									
BATTING_BB	0.47	0.2	0.2	-0.21	0.46										
BATTING_HR	0.42	0.4	0.25	-0.22											
BATTING_3B	-0.12	0.21	0.04												
BATTING_2B	0.31	0.56													
BATTING_H	0.47														

Corr

1.0
0.5
0.0
-0.5
-1.0

2 Data preparation

- fix missing values
- mathematical transformations
- creating new variables - explain