Training-set Exploration

Salma Elshahawy

2020-09-14

Contents

```
dataset <- read.csv(
  "https://raw.githubusercontent.com/salma71/Data_621/master/HW_1/datasets/moneyball-training-data.csv"
  stringsAsFactors = FALSE)
```

```
dim(dataset)
```

```
## [1] 2276   17
```

```
# list types for each attribute
sapply(dataset, class)
```

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B
##        "integer"        "integer"        "integer"        "integer"
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO
##        "integer"        "integer"        "integer"        "integer"
##  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H
##        "integer"        "integer"        "integer"        "integer"
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##        "integer"        "integer"        "integer"        "integer"
## TEAM_FIELDING_DP
##        "integer"
```

```
library(kableExtra)
# take a peek at the first 5 rows of the data
kableExtra::kbl(head(dataset[2:17], 40), booktabs = T) %>%
  kable_styling(latex_options = c('striped','scale_down')) %>%
  landscape()
```

TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
39	1445	194	39	13	143	842	NA	NA	NA	9364	84	927	5456	1011	NA
70	1339	219	22	190	685	1075	37	28	NA	1347	191	689	1082	193	155
86	1377	232	35	137	602	917	46	27	NA	1377	137	602	917	175	153
70	1387	209	38	96	451	922	43	30	NA	1396	97	454	928	164	156
82	1297	186	27	102	472	920	49	39	NA	1297	102	472	920	138	168
75	1279	200	36	92	443	973	107	59	NA	1279	92	443	973	123	149
80	1244	179	54	122	525	1062	80	54	NA	1244	122	525	1062	136	186
85	1273	171	37	115	456	1027	40	36	NA	1281	116	459	1033	112	136
86	1391	197	40	114	447	922	69	27	NA	1391	114	447	922	127	169
76	1271	213	18	96	441	827	72	34	NA	1271	96	441	827	131	159
78	1305	179	27	82	374	888	60	39	NA	1364	86	391	928	119	141
68	1372	203	31	95	509	801	119	79	NA	1372	95	509	801	147	150
72	1332	196	41	55	597	816	221	109	NA	1340	55	601	821	185	165
76	1265	210	23	63	534	812	126	80	NA	1265	63	534	812	150	139
74	1380	233	40	131	542	880	159	89	NA	1380	131	542	880	147	137
87	1417	226	28	108	539	682	86	69	NA	1417	108	539	682	136	136
88	1563	242	43	164	589	843	100	53	NA	1563	164	589	843	135	172
66	1460	239	32	107	546	900	92	64	NA	1478	108	553	911	136	146
75	1390	197	24	143	579	841	65	49	NA	2047	211	853	1239	149	177
93	1518	268	26	186	613	760	55	53	NA	1518	186	613	760	106	171
70	1467	241	22	154	509	835	41	39	NA	1467	154	509	835	154	190
81	1363	211	30	150	556	928	80	51	NA	1363	150	556	928	128	170
90	1364	215	31	153	648	902	106	51	NA	1364	153	648	902	112	202
92	1387	236	36	167	671	860	109	42	NA	1387	167	671	860	107	156
75	1406	257	26	172	590	926	125	44	NA	1406	172	590	926	117	162
75	1458	258	31	124	469	819	86	52	NA	1458	124	469	819	135	175
91	1422	208	37	145	429	1011	89	40	NA	1422	145	429	1011	96	173
80	1448	237	27	147	566	1000	69	43	NA	1448	147	566	1000	140	186
81	1396	245	29	115	448	928	94	56	NA	1396	115	448	928	102	156
72	1306	202	20	88	416	882	160	101	NA	1306	88	416	882	134	172
71	1399	259	24	114	564	930	169	100	NA	1399	114	564	930	120	161
66	1468	251	23	169	566	1007	92	76	NA	2068	238	797	1419	107	155
87	1553	282	28	208	630	993	65	44	NA	1735	232	704	1109	106	134
70	1581	258	24	193	530	980	53	39	NA	1591	194	533	986	129	157
84	1531	279	25	161	617	953	126	72	NA	1531	161	617	953	126	140
85	1530	314	27	147	510	1028	93	45	NA	1530	147	510	1028	106	146
70	1404	248	22	158	511	1022	71	45	NA	1404	158	511	1022	106	156
82	1574	309	34	236	608	1024	93	52	47	1574	236	608	1024	134	184
75	1447	275	26	158	494	1001	116	52	77	1447	158	494	1001	103	142
99	1603	333	32	152	462	805	117	51	74	1603	152	462	805	87	151

We can confirm that the scales for the attributes are all over the place because of the differing units. We may benefit from some transforms later on.

```
quick_summary <- function(df){
  df %>%
    summary() %>%
    kable() %>%
    kable_styling(latex_options = 'striped') %>%
    landscape()
}

quick_summary(dataset[2:7])
```

	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
	Min. : 0.00	Min. : 891	Min. : 69.0	Min. : 0.00	Min. : 0.00	Min. : 0.0
	1st Qu.: 71.00	1st Qu.:1383	1st Qu.:208.0	1st Qu.: 34.00	1st Qu.: 42.00	1st Qu.:451.0
	Median : 82.00	Median :1454	Median :238.0	Median : 47.00	Median :102.00	Median :512.0
	Mean : 80.79	Mean :1469	Mean :241.2	Mean : 55.25	Mean : 99.61	Mean :501.6
	3rd Qu.: 92.00	3rd Qu.:1537	3rd Qu.:273.0	3rd Qu.: 72.00	3rd Qu.:147.00	3rd Qu.:580.0
	Max. :146.00	Max. :2554	Max. :458.0	Max. :223.00	Max. :264.00	Max. :878.0

```
quick_summary(dataset[8:12])
```

	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
	Min. : 0.0	Min. : 0.0	Min. : 0.0	Min. :29.00	Min. : 1137
	1st Qu.: 548.0	1st Qu.: 66.0	1st Qu.: 38.0	1st Qu.:50.50	1st Qu.: 1419
	Median : 750.0	Median :101.0	Median : 49.0	Median :58.00	Median : 1518
	Mean : 735.6	Mean :124.8	Mean : 52.8	Mean :59.36	Mean : 1779
	3rd Qu.: 930.0	3rd Qu.:156.0	3rd Qu.: 62.0	3rd Qu.:67.00	3rd Qu.: 1682
	Max. :1399.0	Max. :697.0	Max. :201.0	Max. :95.00	Max. :30132
	NA's :102	NA's :131	NA's :772	NA's :2085	NA

```
quick_summary(dataset[13:15])
```

	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO
	Min. : 0.0	Min. : 0.0	Min. : 0.0
	1st Qu.: 50.0	1st Qu.: 476.0	1st Qu.: 615.0
	Median :107.0	Median : 536.5	Median : 813.5
	Mean :105.7	Mean : 553.0	Mean : 817.7
	3rd Qu.:150.0	3rd Qu.: 611.0	3rd Qu.: 968.0
	Max. :343.0	Max. :3645.0	Max. :19278.0
	NA	NA	NA's :102

Use Hmisc library to derive the p-value correlation matrix

Formatting the correlation matrix table

```
# ++++++++++++++++++++++++++++
# flattenCorrMatrix
# ++++++++++++++++++++++++++++
# cormat : matrix of the correlation coefficients
# pmat : matrix of the correlation p-values
flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
    )
}

res2<-rcorr(as.matrix(dataset))
flattenCorrMatrix(res2$r, res2$P)
```

```
##                  row           column           cor            p
## 1              INDEX      TARGET_WINS -0.0210564349 3.153265e-01
## 2              INDEX   TEAM_BATTING_H -0.0179202413 3.928126e-01
## 3        TARGET_WINS   TEAM_BATTING_H  0.3887675211 0.000000e+00
## 4              INDEX  TEAM_BATTING_2B  0.0111830135 5.938698e-01
## 5        TARGET_WINS  TEAM_BATTING_2B  0.2891036451 0.000000e+00
## 6     TEAM_BATTING_H  TEAM_BATTING_2B  0.5628496778 0.000000e+00
## 7              INDEX  TEAM_BATTING_3B -0.0058146834 7.815854e-01
## 8        TARGET_WINS  TEAM_BATTING_3B  0.1426084107 8.217427e-12
## 9     TEAM_BATTING_H  TEAM_BATTING_3B  0.4276965751 0.000000e+00
## 10   TEAM_BATTING_2B  TEAM_BATTING_3B -0.1073058236 2.877545e-07
## 11             INDEX  TEAM_BATTING_HR  0.0514810468 1.403725e-02
## 12       TARGET_WINS  TEAM_BATTING_HR  0.1761532005 0.000000e+00
## 13    TEAM_BATTING_H  TEAM_BATTING_HR -0.0065446845 7.549934e-01
## 14   TEAM_BATTING_2B  TEAM_BATTING_HR  0.4353972933 0.000000e+00
## 15   TEAM_BATTING_3B  TEAM_BATTING_HR -0.6355669461 0.000000e+00
## 16             INDEX  TEAM_BATTING_BB -0.0265672362 2.051616e-01
## 17       TARGET_WINS  TEAM_BATTING_BB  0.2325598638 0.000000e+00
## 18    TEAM_BATTING_H  TEAM_BATTING_BB -0.0724640128 5.407324e-04
## 19   TEAM_BATTING_2B  TEAM_BATTING_BB  0.2557261034 0.000000e+00
## 20   TEAM_BATTING_3B  TEAM_BATTING_BB -0.2872358406 0.000000e+00
## 21   TEAM_BATTING_HR  TEAM_BATTING_BB  0.5137348096 0.000000e+00
## 22             INDEX  TEAM_BATTING_SO  0.0814501106 1.436251e-04
## 23       TARGET_WINS  TEAM_BATTING_SO -0.0317507079 1.388904e-01
## 24    TEAM_BATTING_H  TEAM_BATTING_SO -0.4638535715 0.000000e+00
## 25   TEAM_BATTING_2B  TEAM_BATTING_SO  0.1626851878 2.309264e-14
## 26   TEAM_BATTING_3B  TEAM_BATTING_SO -0.6697811879 0.000000e+00
## 27   TEAM_BATTING_HR  TEAM_BATTING_SO  0.7270693481 0.000000e+00
## 28   TEAM_BATTING_BB  TEAM_BATTING_SO  0.3797508658 0.000000e+00
## 29             INDEX  TEAM_BASERUN_SB  0.0402671277 6.223655e-02
## 30       TARGET_WINS  TEAM_BASERUN_SB  0.1351389207 3.298830e-10
## 31    TEAM_BATTING_H  TEAM_BASERUN_SB  0.1235677971 9.377653e-09
```

```
## 32   TEAM_BATTING_2B  TEAM_BASERUN_SB -0.1997572392 0.000000e+00
## 33   TEAM_BATTING_3B  TEAM_BASERUN_SB  0.5335064476 0.000000e+00
## 34   TEAM_BATTING_HR  TEAM_BASERUN_SB -0.4535784264 0.000000e+00
## 35   TEAM_BATTING_BB  TEAM_BASERUN_SB -0.1051156429 1.066116e-06
## 36   TEAM_BATTING_SO  TEAM_BASERUN_SB -0.2544892318 0.000000e+00
## 37            INDEX  TEAM_BASERUN_CS  0.0005653743 9.825215e-01
## 38       TARGET_WINS  TEAM_BASERUN_CS  0.0224040691 3.852582e-01
## 39    TEAM_BATTING_H  TEAM_BASERUN_CS  0.0167056677 5.173884e-01
## 40   TEAM_BATTING_2B  TEAM_BASERUN_CS -0.0998140593 1.055784e-04
## 41   TEAM_BATTING_3B  TEAM_BASERUN_CS  0.3487649195 0.000000e+00
## 42   TEAM_BATTING_HR  TEAM_BASERUN_CS -0.4337938681 0.000000e+00
## 43   TEAM_BATTING_BB  TEAM_BASERUN_CS -0.1369883707 9.641725e-08
## 44   TEAM_BATTING_SO  TEAM_BASERUN_CS -0.2178813684 0.000000e+00
## 45   TEAM_BASERUN_SB  TEAM_BASERUN_CS  0.6552448036 0.000000e+00
## 46            INDEX TEAM_BATTING_HBP  0.0771930266 2.885034e-01
## 47       TARGET_WINS TEAM_BATTING_HBP  0.0735042423 3.122327e-01
## 48    TEAM_BATTING_H TEAM_BATTING_HBP -0.0291121757 6.893171e-01
## 49   TEAM_BATTING_2B TEAM_BATTING_HBP  0.0460847531 5.266947e-01
## 50   TEAM_BATTING_3B TEAM_BATTING_HBP -0.1742471538 1.591723e-02
## 51   TEAM_BATTING_HR TEAM_BATTING_HBP  0.1061811601 1.437532e-01
## 52   TEAM_BATTING_BB TEAM_BATTING_HBP  0.0474600668 5.144185e-01
## 53   TEAM_BATTING_SO TEAM_BATTING_HBP  0.2209421943 2.129956e-03
## 54   TEAM_BASERUN_SB TEAM_BATTING_HBP -0.0640049816 3.790423e-01
## 55   TEAM_BASERUN_CS TEAM_BATTING_HBP -0.0705138958 3.323798e-01
## 56            INDEX  TEAM_PITCHING_H  0.0171031479 4.147525e-01
## 57       TARGET_WINS  TEAM_PITCHING_H -0.1099370542 1.457270e-07
## 58    TEAM_BATTING_H  TEAM_PITCHING_H  0.3026937094 0.000000e+00
## 59   TEAM_BATTING_2B  TEAM_PITCHING_H  0.0236921877 2.585473e-01
## 60   TEAM_BATTING_3B  TEAM_PITCHING_H  0.1948794111 0.000000e+00
## 61   TEAM_BATTING_HR  TEAM_PITCHING_H -0.2501454813 0.000000e+00
## 62   TEAM_BATTING_BB  TEAM_PITCHING_H -0.4497776250 0.000000e+00
## 63   TEAM_BATTING_SO  TEAM_PITCHING_H -0.3756863689 0.000000e+00
## 64   TEAM_BASERUN_SB  TEAM_PITCHING_H  0.0732850496 6.819772e-04
## 65   TEAM_BASERUN_CS  TEAM_PITCHING_H -0.0520078089 4.373461e-02
## 66  TEAM_BATTING_HBP  TEAM_PITCHING_H -0.0276969949 7.036928e-01
## 67            INDEX TEAM_PITCHING_HR  0.0509858973 1.498867e-02
## 68       TARGET_WINS TEAM_PITCHING_HR  0.1890137348 0.000000e+00
## 69    TEAM_BATTING_H TEAM_PITCHING_HR  0.0728531193 5.045119e-04
## 70   TEAM_BATTING_2B TEAM_PITCHING_HR  0.4545508177 0.000000e+00
## 71   TEAM_BATTING_3B TEAM_PITCHING_HR -0.5678366791 0.000000e+00
## 72   TEAM_BATTING_HR TEAM_PITCHING_HR  0.9693713961 0.000000e+00
## 73   TEAM_BATTING_BB TEAM_PITCHING_HR  0.4595520723 0.000000e+00
## 74   TEAM_BATTING_SO TEAM_PITCHING_HR  0.6671788919 0.000000e+00
## 75   TEAM_BASERUN_SB TEAM_PITCHING_HR -0.4165107234 0.000000e+00
## 76   TEAM_BASERUN_CS TEAM_PITCHING_HR -0.4225660463 0.000000e+00
## 77  TEAM_BATTING_HBP TEAM_PITCHING_HR  0.1067587798 1.415740e-01
## 78   TEAM_PITCHING_H TEAM_PITCHING_HR -0.1416127589 1.148881e-11
## 79            INDEX TEAM_PITCHING_BB -0.0152875130 4.660200e-01
## 80       TARGET_WINS TEAM_PITCHING_BB  0.1241745360 2.784686e-09
## 81    TEAM_BATTING_H TEAM_PITCHING_BB  0.0941930273 6.755492e-06
## 82   TEAM_BATTING_2B TEAM_PITCHING_BB  0.1780542044 0.000000e+00
## 83   TEAM_BATTING_3B TEAM_PITCHING_BB -0.0022241484 9.155425e-01
## 84   TEAM_BATTING_HR TEAM_PITCHING_BB  0.1369275637 5.388223e-11
## 85   TEAM_BATTING_BB TEAM_PITCHING_BB  0.4893612630 0.000000e+00
```

```
## 86   TEAM_BATTING_SO TEAM_PITCHING_BB  0.0370051408 8.452629e-02
## 87   TEAM_BASERUN_SB TEAM_PITCHING_BB  0.1464151343 9.499512e-12
## 88   TEAM_BASERUN_CS TEAM_PITCHING_BB -0.1069612356 3.230317e-05
## 89  TEAM_BATTING_HBP TEAM_PITCHING_BB  0.0478513710 5.109529e-01
## 90   TEAM_PITCHING_H TEAM_PITCHING_BB  0.3206761623 0.000000e+00
## 91  TEAM_PITCHING_HR TEAM_PITCHING_BB  0.2219375049 0.000000e+00
## 92             INDEX TEAM_PITCHING_SO  0.0558901457 9.147756e-03
## 93       TARGET_WINS TEAM_PITCHING_SO -0.0784360901 2.515153e-04
## 94    TEAM_BATTING_H TEAM_PITCHING_SO -0.2526567897 0.000000e+00
## 95   TEAM_BATTING_2B TEAM_PITCHING_SO  0.0647923149 2.507323e-03
## 96   TEAM_BATTING_3B TEAM_PITCHING_SO -0.2588189308 0.000000e+00
## 97   TEAM_BATTING_HR TEAM_PITCHING_SO  0.1847075643 0.000000e+00
## 98   TEAM_BATTING_BB TEAM_PITCHING_SO -0.0207568220 3.333647e-01
## 99   TEAM_BATTING_SO TEAM_PITCHING_SO  0.4162333001 0.000000e+00
## 100  TEAM_BASERUN_SB TEAM_PITCHING_SO -0.1371286088 4.853151e-10
## 101  TEAM_BASERUN_CS TEAM_PITCHING_SO -0.2102227352 2.220446e-16
## 102 TEAM_BATTING_HBP TEAM_PITCHING_SO  0.2215737541 2.066596e-03
## 103  TEAM_PITCHING_H TEAM_PITCHING_SO  0.2672480744 0.000000e+00
## 104 TEAM_PITCHING_HR TEAM_PITCHING_SO  0.2058805288 0.000000e+00
## 105 TEAM_PITCHING_BB TEAM_PITCHING_SO  0.4884986534 0.000000e+00
## 106            INDEX  TEAM_FIELDING_E -0.0092331265 6.597515e-01
## 107       TARGET_WINS  TEAM_FIELDING_E -0.1764847590 0.000000e+00
## 108    TEAM_BATTING_H  TEAM_FIELDING_E  0.2649024778 0.000000e+00
## 109   TEAM_BATTING_2B  TEAM_FIELDING_E -0.2351509865 0.000000e+00
## 110   TEAM_BATTING_3B  TEAM_FIELDING_E  0.5097784470 0.000000e+00
## 111   TEAM_BATTING_HR  TEAM_FIELDING_E -0.5873390979 0.000000e+00
## 112   TEAM_BATTING_BB  TEAM_FIELDING_E -0.6559708147 0.000000e+00
## 113   TEAM_BATTING_SO  TEAM_FIELDING_E -0.5846644361 0.000000e+00
## 114   TEAM_BASERUN_SB  TEAM_FIELDING_E  0.5096309017 0.000000e+00
## 115   TEAM_BASERUN_CS  TEAM_FIELDING_E  0.0483218940 6.099538e-02
## 116 TEAM_BATTING_HBP  TEAM_FIELDING_E  0.0417897123 5.659644e-01
## 117  TEAM_PITCHING_H  TEAM_FIELDING_E  0.6677590102 0.000000e+00
## 118 TEAM_PITCHING_HR  TEAM_FIELDING_E -0.4931444663 0.000000e+00
## 119 TEAM_PITCHING_BB  TEAM_FIELDING_E -0.0228375611 2.761252e-01
## 120 TEAM_PITCHING_SO  TEAM_FIELDING_E -0.0232917828 2.776873e-01
## 121            INDEX TEAM_FIELDING_DP  0.0200642919 3.710095e-01
## 122       TARGET_WINS TEAM_FIELDING_DP -0.0348505836 1.201464e-01
## 123    TEAM_BATTING_H TEAM_FIELDING_DP  0.1553833214 3.179013e-12
## 124   TEAM_BATTING_2B TEAM_FIELDING_DP  0.2908799782 0.000000e+00
## 125   TEAM_BATTING_3B TEAM_FIELDING_DP -0.3230748473 0.000000e+00
## 126   TEAM_BATTING_HR TEAM_FIELDING_DP  0.4489853482 0.000000e+00
## 127   TEAM_BATTING_BB TEAM_FIELDING_DP  0.4308767474 0.000000e+00
## 128   TEAM_BATTING_SO TEAM_FIELDING_DP  0.1548893916 1.319034e-11
## 129   TEAM_BASERUN_SB TEAM_FIELDING_DP -0.4970776274 0.000000e+00
## 130   TEAM_BASERUN_CS TEAM_FIELDING_DP -0.2142480076 0.000000e+00
## 131 TEAM_BATTING_HBP TEAM_FIELDING_DP -0.0712082412 3.276290e-01
## 132  TEAM_PITCHING_H TEAM_FIELDING_DP -0.2286505923 0.000000e+00
## 133 TEAM_PITCHING_HR TEAM_FIELDING_DP  0.4391703971 0.000000e+00
## 134 TEAM_PITCHING_BB TEAM_FIELDING_DP  0.3244572260 0.000000e+00
## 135 TEAM_PITCHING_SO TEAM_FIELDING_DP  0.0261580430 2.559407e-01
## 136  TEAM_FIELDING_E TEAM_FIELDING_DP -0.4976849536 0.000000e+00
```