DATA621 | Project I

Abdellah AitElmouden | Gabriel Abreu | Jered Ataky | Patrick Maloney

2/12/2021

Abstract

To see how regression will help us evaluate baseball team performance, we will explore, analyze and model a historical baseball data set containing approximately 2200 records, to determine a team's performance based on statistics of their performance. Each record represents a professional baseball team from the years 1871 to 2006 inclusive, and the data include the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

While correlation does not equal causation it is suggested that a focus on some of the variables such as a focus on either single hits or triple or more hits to the exclusion of doubles might be worth pursuing. Also the data suggests that a focus on home runs allowed may not be worth giving up a number of more normal hits.

.....To add more here....

Introduction

Because baseball is so numbers-heavy, there are many different statistics to consider when searching for the best predictors of team success. There are offensive statistics (offense meaning when a team is batting) and defensive statistics (defense meaning when a team is in the field). These categories can be broken up into many more subcategories. However, for the purpose of the this project we will use the available data to build a multiple linear regression model on the training data to predict the number of wins for the team.

Data Used

the data was provided in csv file. The files contain approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. 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.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Outcome Variable
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

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
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

Data exploration

The initial steps are to download the data and take a quick glimpse of the columns, their data types, number of columns, and rows.

```
#Import data

data <- read.csv("https://raw.githubusercontent.com/aaitelmouden/DATA621/master/Project1/moneyball-trai
glimpse(data)
```

```
## Observations: 2,276
## Variables: 17
## $ INDEX            <int> 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 15, 16, 17, 18...
## $ TARGET_WINS      <int> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 68, 72...
## $ TEAM_BATTING_H   <int> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 1273, 13...
## $ TEAM_BATTING_2B  <int> 194, 219, 232, 209, 186, 200, 179, 171, 197, 213, ...
## $ TEAM_BATTING_3B  <int> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 31, 41...
## $ TEAM_BATTING_HR  <int> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96, 82, ...
## $ TEAM_BATTING_BB  <int> 143, 685, 602, 451, 472, 443, 525, 456, 447, 441, ...
## $ TEAM_BATTING_SO  <int> 842, 1075, 917, 922, 920, 973, 1062, 1027, 922, 82...
## $ TEAM_BASERUN_SB  <int> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, 119, ...
## $ TEAM_BASERUN_CS  <int> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 79, 10...
## $ TEAM_BATTING_HBP <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ TEAM_PITCHING_H  <int> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 1281, 13...
## $ TEAM_PITCHING_HR <int> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96, 86, ...
## $ TEAM_PITCHING_BB <int> 927, 689, 602, 454, 472, 443, 525, 459, 447, 441, ...
## $ TEAM_PITCHING_SO <int> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 922, 8...
## $ TEAM_FIELDING_E  <int> 1011, 193, 175, 164, 138, 123, 136, 112, 127, 131,...
## $ TEAM_FIELDING_DP <int> NA, 155, 153, 156, 168, 149, 186, 136, 169, 159, 1...
```

At first glance, the column BATTING_HBP has numerous NA values that will need to be addressed before building a model. It's worth exploring for other columns with NA values.

```
summary(data)
```

```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0
##  Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2
##  3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0
##  Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0
```

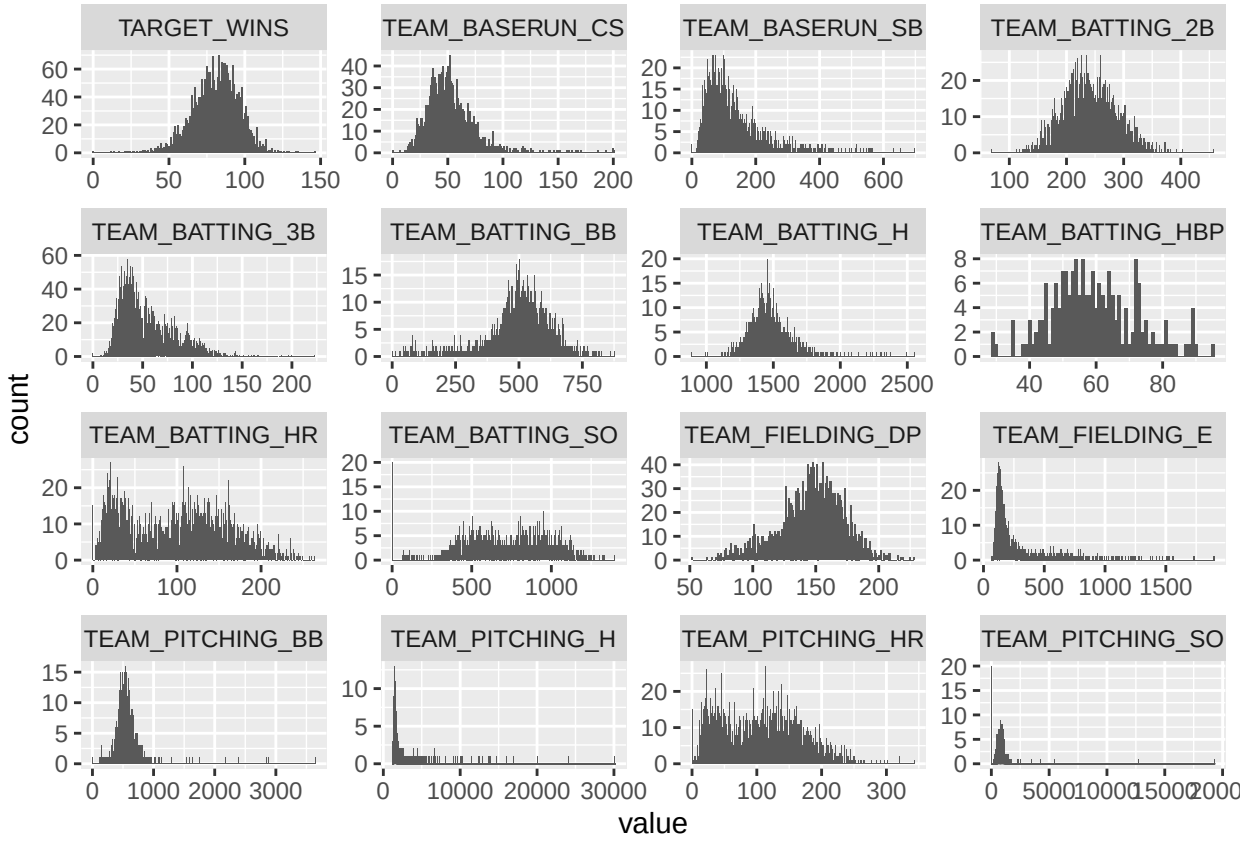
```
##
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0
##  Median : 47.00   Median :102.00   Median :512.0   Median : 750.0
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 735.6
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
##                                                    NA's   :102
##  TEAM_BASERUN_SB TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H
##  Min.   :  0.0   Min.   :  0.0   Min.   :29.00    Min.   : 1137
##  1st Qu.: 66.0   1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419
##  Median :101.0   Median : 49.0   Median :58.00    Median : 1518
##  Mean   :124.8   Mean   : 52.8   Mean   :59.36    Mean   : 1779
##  3rd Qu.:156.0   3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682
##  Max.   :697.0   Max.   :201.0   Max.   :95.00    Max.   :30132
##  NA's   :131     NA's   :772     NA's   :2085
##  TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##  Min.   :  0.0    Min.   :   0.0   Min.   :    0.0   Min.   :  65.0
##  1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0
##  Median :107.0    Median : 536.5   Median :  813.5   Median : 159.0
##  Mean   :105.7    Mean   : 553.0   Mean   :  817.7   Mean   : 246.5
##  3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2
##  Max.   :343.0    Max.   :3645.0   Max.   :19278.0   Max.   :1898.0
##                                    NA's   :102
##  TEAM_FIELDING_DP
##  Min.   : 52.0
##  1st Qu.:131.0
##  Median :149.0
##  Mean   :146.4
##  3rd Qu.:164.0
##  Max.   :228.0
##  NA's   :286
```

```
max_obs <- 2276
batting_so_na <- ((102/max_obs) * 100)
baserun_sb_na <- (131/max_obs) * 100
baserun_cs_na <- (772/max_obs) * 100
batting_hbp_na <- (2085/max_obs) * 100
pitching_so_na <- (102/max_obs) * 100
fielding_dp_na <- (286/max_obs) * 100

df_percent_na <- data.frame(Columns_w_NA = c("team_batting_so", "team_baserun_sb", "team_baserun_cs", "

kable(df_percent_na)
```

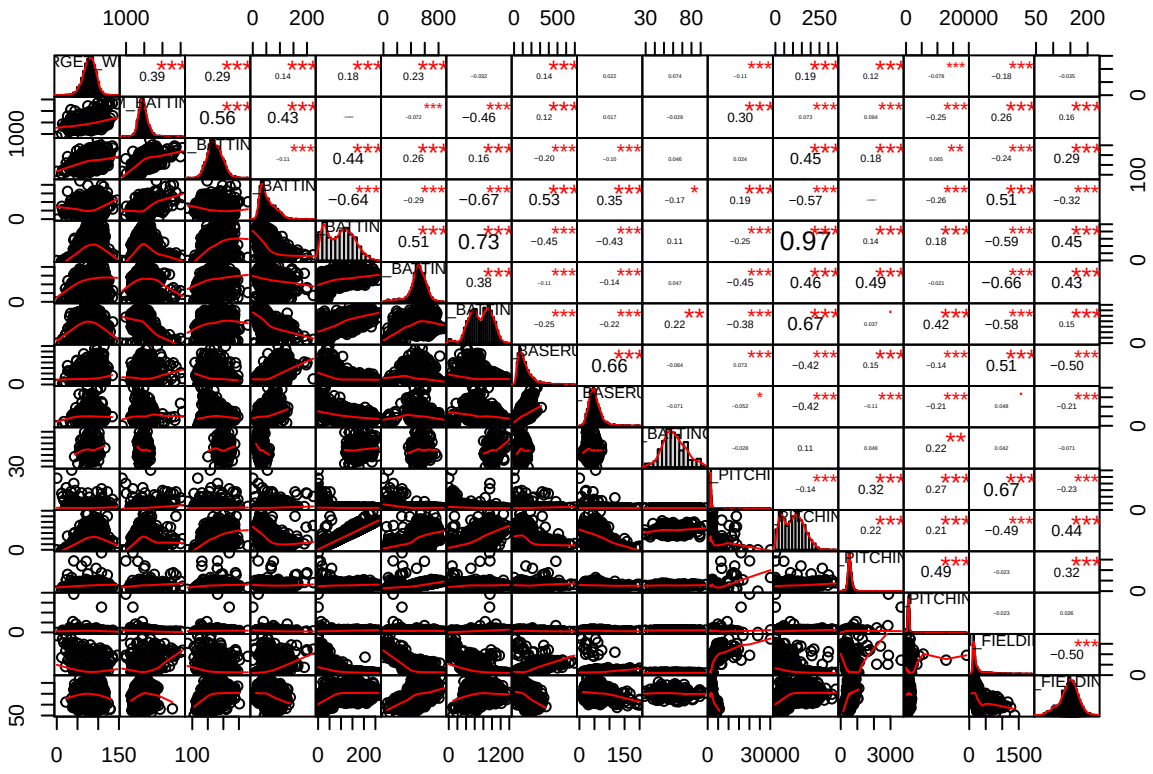
Columns_w_NA	Percent_NA
team_batting_so	4.481547
team_baserun_sb	5.755712
team_baserun_cs	33.919156
team_batting_hbp	91.608084
team_pitching_so	4.481547
team_fielding_dp	12.565905

```
data[-c(1)] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(binwidth=1)
```

```
## Warning: Removed 3478 rows containing non-finite values (stat_bin).
```




```
chart.Correlation(data[-c(1)], histograme=TRUE, method= "pearson")
```



Variable Selection

```
#eliminate INDEX from data frame
data_no_index <- data[-c(1)]

cor_matrix <- rcorr(as.matrix(data_no_index))

flattenCorrMatrix(cor_matrix$r, cor_matrix$P)
```

```
##                 row           column          cor            p
## 1       TARGET_WINS   TEAM_BATTING_H  0.388767521 0.000000e+00
## 2       TARGET_WINS  TEAM_BATTING_2B  0.289103645 0.000000e+00
## 3    TEAM_BATTING_H  TEAM_BATTING_2B  0.562849678 0.000000e+00
## 4       TARGET_WINS  TEAM_BATTING_3B  0.142608411 8.217427e-12
## 5    TEAM_BATTING_H  TEAM_BATTING_3B  0.427696575 0.000000e+00
## 6   TEAM_BATTING_2B  TEAM_BATTING_3B -0.107305824 2.877545e-07
## 7       TARGET_WINS  TEAM_BATTING_HR  0.176153200 0.000000e+00
## 8    TEAM_BATTING_H  TEAM_BATTING_HR -0.006544685 7.549934e-01
## 9   TEAM_BATTING_2B  TEAM_BATTING_HR  0.435397293 0.000000e+00
## 10  TEAM_BATTING_3B  TEAM_BATTING_HR -0.635566946 0.000000e+00
## 11      TARGET_WINS  TEAM_BATTING_BB  0.232559864 0.000000e+00
## 12   TEAM_BATTING_H  TEAM_BATTING_BB -0.072464013 5.407324e-04
## 13  TEAM_BATTING_2B  TEAM_BATTING_BB  0.255726103 0.000000e+00
## 14  TEAM_BATTING_3B  TEAM_BATTING_BB -0.287235841 0.000000e+00
```

```
## 15   TEAM_BATTING_HR  TEAM_BATTING_BB  0.513734810 0.000000e+00
## 16       TARGET_WINS  TEAM_BATTING_SO -0.031750708 1.388904e-01
## 17    TEAM_BATTING_H  TEAM_BATTING_SO -0.463853571 0.000000e+00
## 18   TEAM_BATTING_2B  TEAM_BATTING_SO  0.162685188 2.309264e-14
## 19   TEAM_BATTING_3B  TEAM_BATTING_SO -0.669781188 0.000000e+00
## 20   TEAM_BATTING_HR  TEAM_BATTING_SO  0.727069348 0.000000e+00
## 21   TEAM_BATTING_BB  TEAM_BATTING_SO  0.379750866 0.000000e+00
## 22       TARGET_WINS  TEAM_BASERUN_SB  0.135138921 3.298830e-10
## 23    TEAM_BATTING_H  TEAM_BASERUN_SB  0.123567797 9.377653e-09
## 24   TEAM_BATTING_2B  TEAM_BASERUN_SB -0.199757239 0.000000e+00
## 25   TEAM_BATTING_3B  TEAM_BASERUN_SB  0.533506448 0.000000e+00
## 26   TEAM_BATTING_HR  TEAM_BASERUN_SB -0.453578426 0.000000e+00
## 27   TEAM_BATTING_BB  TEAM_BASERUN_SB -0.105115643 1.066116e-06
## 28   TEAM_BATTING_SO  TEAM_BASERUN_SB -0.254489232 0.000000e+00
## 29       TARGET_WINS  TEAM_BASERUN_CS  0.022404069 3.852582e-01
## 30    TEAM_BATTING_H  TEAM_BASERUN_CS  0.016705668 5.173884e-01
## 31   TEAM_BATTING_2B  TEAM_BASERUN_CS -0.099814059 1.055784e-04
## 32   TEAM_BATTING_3B  TEAM_BASERUN_CS  0.348764919 0.000000e+00
## 33   TEAM_BATTING_HR  TEAM_BASERUN_CS -0.433793868 0.000000e+00
## 34   TEAM_BATTING_BB  TEAM_BASERUN_CS -0.136988371 9.641725e-08
## 35   TEAM_BATTING_SO  TEAM_BASERUN_CS -0.217881368 0.000000e+00
## 36   TEAM_BASERUN_SB  TEAM_BASERUN_CS  0.655244804 0.000000e+00
## 37       TARGET_WINS TEAM_BATTING_HBP  0.073504242 3.122327e-01
## 38    TEAM_BATTING_H TEAM_BATTING_HBP -0.029112176 6.893171e-01
## 39   TEAM_BATTING_2B TEAM_BATTING_HBP  0.046084753 5.266947e-01
## 40   TEAM_BATTING_3B TEAM_BATTING_HBP -0.174247154 1.591723e-02
## 41   TEAM_BATTING_HR TEAM_BATTING_HBP  0.106181160 1.437532e-01
## 42   TEAM_BATTING_BB TEAM_BATTING_HBP  0.047460067 5.144185e-01
## 43   TEAM_BATTING_SO TEAM_BATTING_HBP  0.220942194 2.129956e-03
## 44   TEAM_BASERUN_SB TEAM_BATTING_HBP -0.064004982 3.790423e-01
## 45   TEAM_BASERUN_CS TEAM_BATTING_HBP -0.070513896 3.323798e-01
## 46       TARGET_WINS  TEAM_PITCHING_H -0.109937054 1.457270e-07
## 47    TEAM_BATTING_H  TEAM_PITCHING_H  0.302693709 0.000000e+00
## 48   TEAM_BATTING_2B  TEAM_PITCHING_H  0.023692188 2.585473e-01
## 49   TEAM_BATTING_3B  TEAM_PITCHING_H  0.194879411 0.000000e+00
## 50   TEAM_BATTING_HR  TEAM_PITCHING_H -0.250145481 0.000000e+00
## 51   TEAM_BATTING_BB  TEAM_PITCHING_H -0.449777625 0.000000e+00
## 52   TEAM_BATTING_SO  TEAM_PITCHING_H -0.375686369 0.000000e+00
## 53   TEAM_BASERUN_SB  TEAM_PITCHING_H  0.073285050 6.819772e-04
## 54   TEAM_BASERUN_CS  TEAM_PITCHING_H -0.052007809 4.373461e-02
## 55  TEAM_BATTING_HBP  TEAM_PITCHING_H -0.027696995 7.036928e-01
## 56       TARGET_WINS TEAM_PITCHING_HR  0.189013735 0.000000e+00
## 57    TEAM_BATTING_H TEAM_PITCHING_HR  0.072853119 5.045119e-04
## 58   TEAM_BATTING_2B TEAM_PITCHING_HR  0.454550818 0.000000e+00
## 59   TEAM_BATTING_3B TEAM_PITCHING_HR -0.567836679 0.000000e+00
## 60   TEAM_BATTING_HR TEAM_PITCHING_HR  0.969371396 0.000000e+00
## 61   TEAM_BATTING_BB TEAM_PITCHING_HR  0.459552072 0.000000e+00
## 62   TEAM_BATTING_SO TEAM_PITCHING_HR  0.667178892 0.000000e+00
## 63   TEAM_BASERUN_SB TEAM_PITCHING_HR -0.416510723 0.000000e+00
## 64   TEAM_BASERUN_CS TEAM_PITCHING_HR -0.422566046 0.000000e+00
## 65  TEAM_BATTING_HBP TEAM_PITCHING_HR  0.106758780 1.415740e-01
## 66   TEAM_PITCHING_H TEAM_PITCHING_HR -0.141612759 1.148881e-11
## 67       TARGET_WINS TEAM_PITCHING_BB  0.124174536 2.784686e-09
## 68    TEAM_BATTING_H TEAM_PITCHING_BB  0.094193027 6.755492e-06
```

```
## 69   TEAM_BATTING_2B TEAM_PITCHING_BB  0.178054204 0.000000e+00
## 70   TEAM_BATTING_3B TEAM_PITCHING_BB -0.002224148 9.155425e-01
## 71   TEAM_BATTING_HR TEAM_PITCHING_BB  0.136927564 5.388223e-11
## 72   TEAM_BATTING_BB TEAM_PITCHING_BB  0.489361263 0.000000e+00
## 73   TEAM_BATTING_SO TEAM_PITCHING_BB  0.037005141 8.452629e-02
## 74   TEAM_BASERUN_SB TEAM_PITCHING_BB  0.146415134 9.499512e-12
## 75   TEAM_BASERUN_CS TEAM_PITCHING_BB -0.106961236 3.230317e-05
## 76  TEAM_BATTING_HBP TEAM_PITCHING_BB  0.047851371 5.109529e-01
## 77   TEAM_PITCHING_H TEAM_PITCHING_BB  0.320676162 0.000000e+00
## 78  TEAM_PITCHING_HR TEAM_PITCHING_BB  0.221937505 0.000000e+00
## 79       TARGET_WINS TEAM_PITCHING_SO -0.078436090 2.515153e-04
## 80    TEAM_BATTING_H TEAM_PITCHING_SO -0.252656790 0.000000e+00
## 81   TEAM_BATTING_2B TEAM_PITCHING_SO  0.064792315 2.507323e-03
## 82   TEAM_BATTING_3B TEAM_PITCHING_SO -0.258818931 0.000000e+00
## 83   TEAM_BATTING_HR TEAM_PITCHING_SO  0.184707564 0.000000e+00
## 84   TEAM_BATTING_BB TEAM_PITCHING_SO -0.020756822 3.333647e-01
## 85   TEAM_BATTING_SO TEAM_PITCHING_SO  0.416233300 0.000000e+00
## 86   TEAM_BASERUN_SB TEAM_PITCHING_SO -0.137128609 4.853151e-10
## 87   TEAM_BASERUN_CS TEAM_PITCHING_SO -0.210222735 2.220446e-16
## 88  TEAM_BATTING_HBP TEAM_PITCHING_SO  0.221573754 2.066596e-03
## 89   TEAM_PITCHING_H TEAM_PITCHING_SO  0.267248074 0.000000e+00
## 90  TEAM_PITCHING_HR TEAM_PITCHING_SO  0.205880529 0.000000e+00
## 91  TEAM_PITCHING_BB TEAM_PITCHING_SO  0.488498653 0.000000e+00
## 92       TARGET_WINS  TEAM_FIELDING_E -0.176484759 0.000000e+00
## 93    TEAM_BATTING_H  TEAM_FIELDING_E  0.264902478 0.000000e+00
## 94   TEAM_BATTING_2B  TEAM_FIELDING_E -0.235150986 0.000000e+00
## 95   TEAM_BATTING_3B  TEAM_FIELDING_E  0.509778447 0.000000e+00
## 96   TEAM_BATTING_HR  TEAM_FIELDING_E -0.587339098 0.000000e+00
## 97   TEAM_BATTING_BB  TEAM_FIELDING_E -0.655970815 0.000000e+00
## 98   TEAM_BATTING_SO  TEAM_FIELDING_E -0.584664436 0.000000e+00
## 99   TEAM_BASERUN_SB  TEAM_FIELDING_E  0.509630902 0.000000e+00
## 100  TEAM_BASERUN_CS  TEAM_FIELDING_E  0.048321894 6.099538e-02
## 101 TEAM_BATTING_HBP  TEAM_FIELDING_E  0.041789712 5.659644e-01
## 102  TEAM_PITCHING_H  TEAM_FIELDING_E  0.667759010 0.000000e+00
## 103 TEAM_PITCHING_HR  TEAM_FIELDING_E -0.493144466 0.000000e+00
## 104 TEAM_PITCHING_BB  TEAM_FIELDING_E -0.022837561 2.761252e-01
## 105 TEAM_PITCHING_SO  TEAM_FIELDING_E -0.023291783 2.776873e-01
## 106      TARGET_WINS TEAM_FIELDING_DP -0.034850584 1.201464e-01
## 107   TEAM_BATTING_H TEAM_FIELDING_DP  0.155383321 3.179013e-12
## 108  TEAM_BATTING_2B TEAM_FIELDING_DP  0.290879978 0.000000e+00
## 109  TEAM_BATTING_3B TEAM_FIELDING_DP -0.323074847 0.000000e+00
## 110  TEAM_BATTING_HR TEAM_FIELDING_DP  0.448985348 0.000000e+00
## 111  TEAM_BATTING_BB TEAM_FIELDING_DP  0.430876747 0.000000e+00
## 112  TEAM_BATTING_SO TEAM_FIELDING_DP  0.154889392 1.319034e-11
## 113  TEAM_BASERUN_SB TEAM_FIELDING_DP -0.497077627 0.000000e+00
## 114  TEAM_BASERUN_CS TEAM_FIELDING_DP -0.214248008 0.000000e+00
## 115 TEAM_BATTING_HBP TEAM_FIELDING_DP -0.071208241 3.276290e-01
## 116  TEAM_PITCHING_H TEAM_FIELDING_DP -0.228650592 0.000000e+00
## 117 TEAM_PITCHING_HR TEAM_FIELDING_DP  0.439170397 0.000000e+00
## 118 TEAM_PITCHING_BB TEAM_FIELDING_DP  0.324457226 0.000000e+00
## 119 TEAM_PITCHING_SO TEAM_FIELDING_DP  0.026158043 2.559407e-01
## 120  TEAM_FIELDING_E TEAM_FIELDING_DP -0.497684954 0.000000e+00
```