DATA 621: BUSINESS ANALYTICS AND DATA MINING - HW1

Pavan Akula

February 27, 2018

Overview

Purpose of this assignment, is to explore, analyze and model a data set containing 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. The main objective is to build multiple linear regression models on the training data to predict the number of wins for the team. Regression model should be based on the variables from the dataset (or variables that are derived from the variables provided). Below is a short description of the variables in the dataset.

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
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

At a first glance at the dataset, all the variables are continuous and appears to have same metrics. Dataset provides information about teams batting and pitching statistics. Variables TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR may have some arthimetic relation. Example, batter taking a hit may result in double, triple, homerun or none. Hit is stored TEAM_BATTING_H and result of the hit is stored in TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR.

```
> str(BaseballDf)
'data.frame':	2276 obs. of  16 variables:
 $ TARGET_WINS     : int  39 70 86 70 82 75 80 85 86 76 ...
 $ TEAM_BATTING_H  : int  1445 1339 1377 1387 1297 1279 1244 1273 1391 1271 ...
 $ 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 ...
 $ TEAM_BATTING_HR : int  13 190 137 96 102 92 122 115 114 96 ...
 $ 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 922 827 ...
 $ TEAM_BASERUN_SB : int  NA 37 46 43 49 107 80 40 69 72 ...
 $ TEAM_BASERUN_CS : int  NA 28 27 30 39 59 54 36 27 34 ...
 $ TEAM_BATTING_HBP: int  NA NA NA NA NA NA NA NA NA NA ...
 $ TEAM_PITCHING_H : int  9364 1347 1377 1396 1297 1279 1244 1281 1391 1271 ...
 $ TEAM_PITCHING_HR: int  84 191 137 97 102 92 122 116 114 96 ...
 $ 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 922 827 ...
 $ 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 ...
```

Missing Data

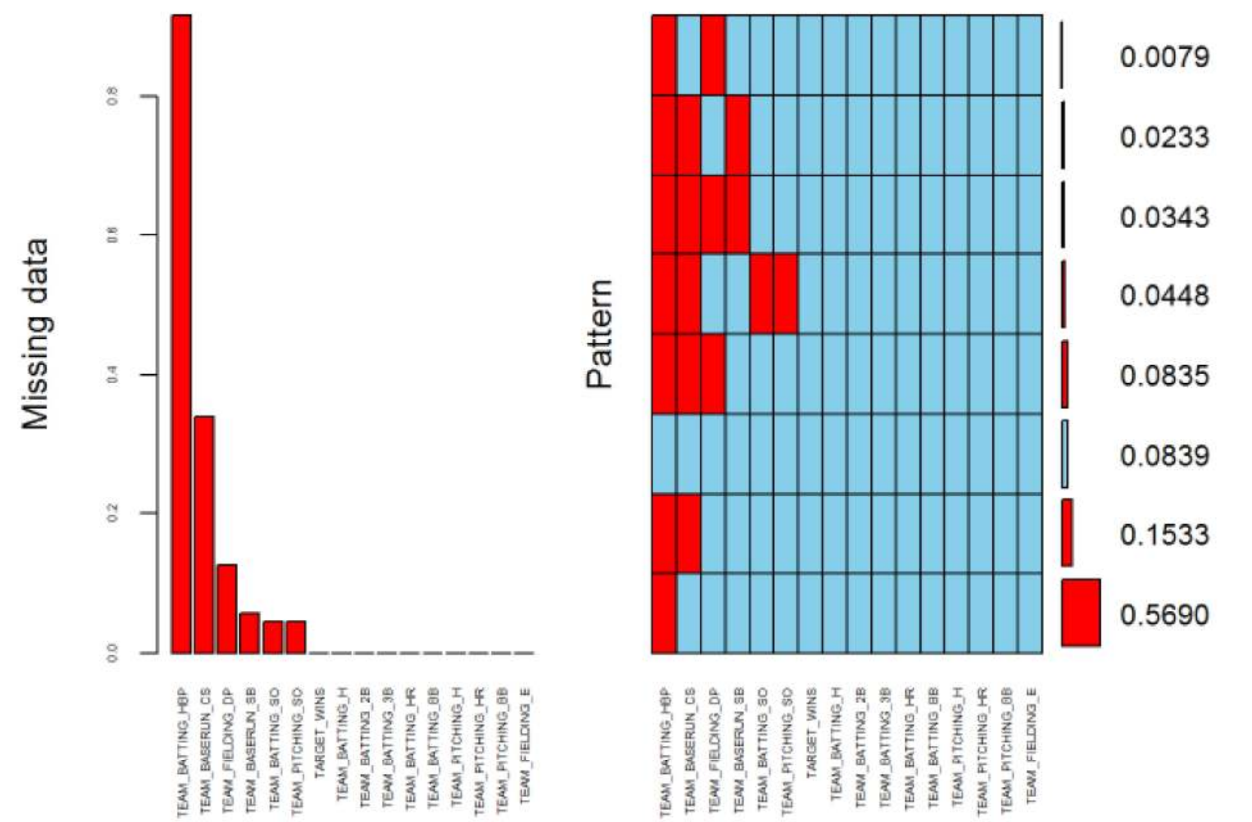
Since data is observational, good data is the basis for constructing decent regression model. Let's explore missing data along with other parameters including mean(μ), standard deviation(σ), etc. For the remainder of the analysis and regression model, INDEX variable will be excluded.

Moneyball Dataset Summary

	Observations	NAs	Minimum	Maximum	1st. Quartile	3st. Quartile	Mean	Median	Sum	Variance	Stdev	Skewness	Kurtosis
TARGET_WINS	2276	0	0	146	71.0	92.00	80.79	82.0	183880	248.13	15.75	-0.40	1.03
TEAM_BATTING_H	2276	0	891	2554	1383.0	1537.25	1469.27	1454.0	3344058	20906.61	144.59	1.57	7.28
TEAM_BATTING_2B	2276	0	69	458	208.0	273.00	241.25	238.0	549078	2190.37	46.80	0.22	0.01
TEAM_BATTING_3B	2276	0	0	223	34.0	72.00	55.25	47.0	125749	780.56	27.94	1.11	1.50
TEAM_BATTING_HR	2276	0	0	264	42.0	147.00	99.61	102.0	226717	3665.92	60.55	0.19	-0.96
TEAM_BATTING_BB	2276	0	0	878	451.0	580.00	501.56	512.0	1141548	15048.14	122.67	-1.03	2.18
TEAM_BATTING_SO	2276	102	0	1399	548.0	930.00	735.61	750.0	1599206	61765.38	248.53	-0.30	-0.32
TEAM_BASERUN_SB	2276	131	0	697	66.0	156.00	124.76	101.0	267614	7707.29	87.79	1.97	5.49
TEAM_BASERUN_CS	2276	772	0	201	38.0	62.00	52.80	49.0	79417	526.99	22.96	1.98	7.62
TEAM_BATTING_HBP	2276	2085	29	95	50.5	67.00	59.36	58.0	11337	168.15	12.97	0.32	-0.11
TEAM_PITCHING_H	2276	0	1137	30132	1419.0	1682.50	1779.21	1518.0	4049483	1979207.03	1406.84	10.33	141.84
TEAM_PITCHING_HR	2276	0	0	343	50.0	150.00	105.70	107.0	240570	3757.54	61.30	0.29	-0.60
TEAM_PITCHING_BB	2276	0	0	3645	476.0	611.00	553.01	536.5	1258646	27674.77	166.36	6.74	96.97
TEAM_PITCHING_SO	2276	102	0	19278	615.0	968.00	817.73	813.5	1777746	305903.05	553.09	22.17	671.19
TEAM_FIELDING_E	2276	0	65	1898	127.0	249.25	246.48	159.0	560990	51879.62	227.77	2.99	10.97
TEAM_FIELDING_DP	2276	286	52	228	131.0	164.00	146.39	149.0	291312	687.82	26.23	-0.39	0.18

Moneyball Dataset Per Game Summary

	Observations	NAs	Minimum	Maximum	1st. Quartile	3st. Quartile	Mean	Median	Sum	Variance	Stdev	Skewness	Kurtosis
TARGET_WINS	2276	0	0.00	0.90	0.44	0.57	0.50	0.51	1135.06	0.01	0.10	-0.40	1.03
TEAM_BATTING_H	2276	0	5.50	15.77	8.54	9.49	9.07	8.98	20642.33	0.80	0.89	1.57	7.28
TEAM_BATTING_2B	2276	0	0.43	2.83	1.28	1.69	1.49	1.47	3389.37	0.08	0.29	0.22	0.01
TEAM_BATTING_3B	2276	0	0.00	1.38	0.21	0.44	0.34	0.29	776.23	0.03	0.17	1.11	1.50
TEAM_BATTING_HR	2276	0	0.00	1.63	0.26	0.91	0.61	0.63	1399.49	0.14	0.37	0.19	-0.96
TEAM_BATTING_BB	2276	0	0.00	5.42	2.78	3.58	3.10	3.16	7046.59	0.57	0.76	-1.03	2.18
TEAM_BATTING_SO	2276	102	0.00	8.64	3.38	5.74	4.54	4.63	9871.64	2.35	1.53	-0.30	-0.32
TEAM_BASERUN_SB	2276	131	0.00	4.30	0.41	0.96	0.77	0.62	1651.94	0.29	0.54	1.97	5.49
TEAM_BASERUN_CS	2276	772	0.00	1.24	0.23	0.38	0.33	0.30	490.23	0.02	0.14	1.98	7.62
TEAM_BATTING_HBP	2276	2085	0.18	0.59	0.31	0.41	0.37	0.36	69.98	0.01	0.08	0.32	-0.11
TEAM_PITCHING_H	2276	0	7.02	186.00	8.76	10.39	10.98	9.37	24996.81	75.42	8.68	10.33	141.84
TEAM_PITCHING_HR	2276	0	0.00	2.12	0.31	0.93	0.65	0.66	1485.00	0.14	0.38	0.29	-0.60
TEAM_PITCHING_BB	2276	0	0.00	22.50	2.94	3.77	3.41	3.31	7769.42	1.05	1.03	6.74	96.97
TEAM_PITCHING_SO	2276	102	0.00	119.00	3.80	5.98	5.05	5.02	10973.74	11.66	3.41	22.17	671.19
TEAM_FIELDING_E	2276	0	0.40	11.72	0.78	1.54	1.52	0.98	3462.90	1.98	1.41	2.99	10.97
TEAM_FIELDING_DP	2276	286	0.32	1.41	0.81	1.01	0.90	0.92	1798.22	0.03	0.16	-0.39	0.18



Data summary shows `TEAM_BATTING_HBP` variable has most of the missing data. Almost 92% of the observations are missing data. Next on the missing data list is `TEAM_BASERUN_CS` with 34%. Following table shows missing data details.

Variable Name	Missing Observations	Percentage
TEAM_BATTING_HBP	2085	92%
TEAM_BASERUN_CS	772	34%
TEAM_FIELDING_DP	286	13%
TEAM_BASERUN_SB	131	6%
TEAM_BATTING_SO	102	5%
TEAM_PITCHING_SO	102	5%

Variable `TEAM_BATTING_HBP` captures data about batter getting hit by pitch. If such an event as not happened for a team during a season, replacing the value with `Zero` should be ok. However, based on `t-Value` and its contribution to regression model variable can be removed.

For rest of the variables, replacing missing values with `mean` would be more meaningful.

Data Distribution

Summary table shows variables `TEAM_PITCHING_SO` and `TEAM_PITCHING_H` have high `Skewness`. Since value is positive, it suggests variables are `right skewed`. Let's verify using `boxplots` and `histograms`.

Boxplots and histograms of each variable suggest there are some outliers in the data. Since we do not have access to team and year variables data cannot be validated against any online sites.

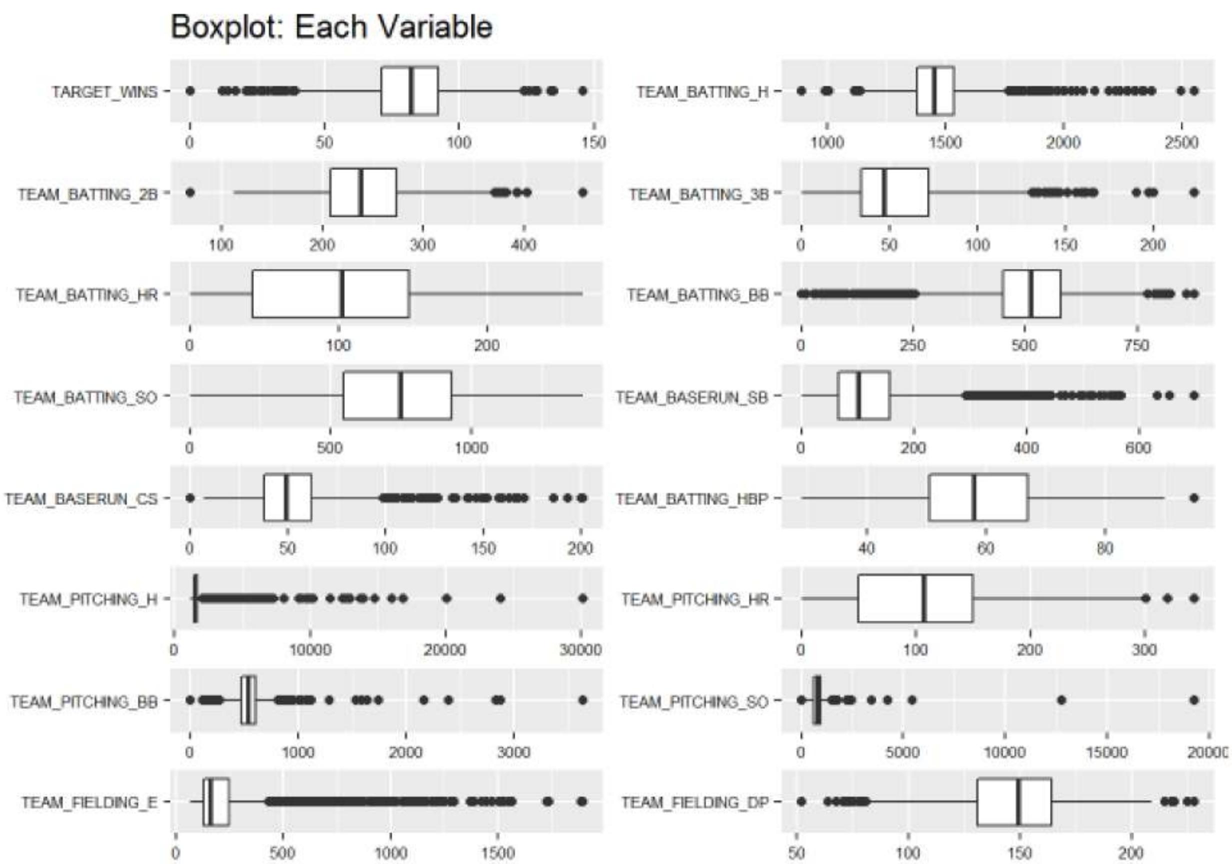
However, website https://www.baseball-reference.com (https://www.baseball-reference.com) has data from the years 1871 to 2006 inclusive. Though we cannot use the data per project guide lines, we can compare overall averages to overall averages of our data to see any anomalies in data.

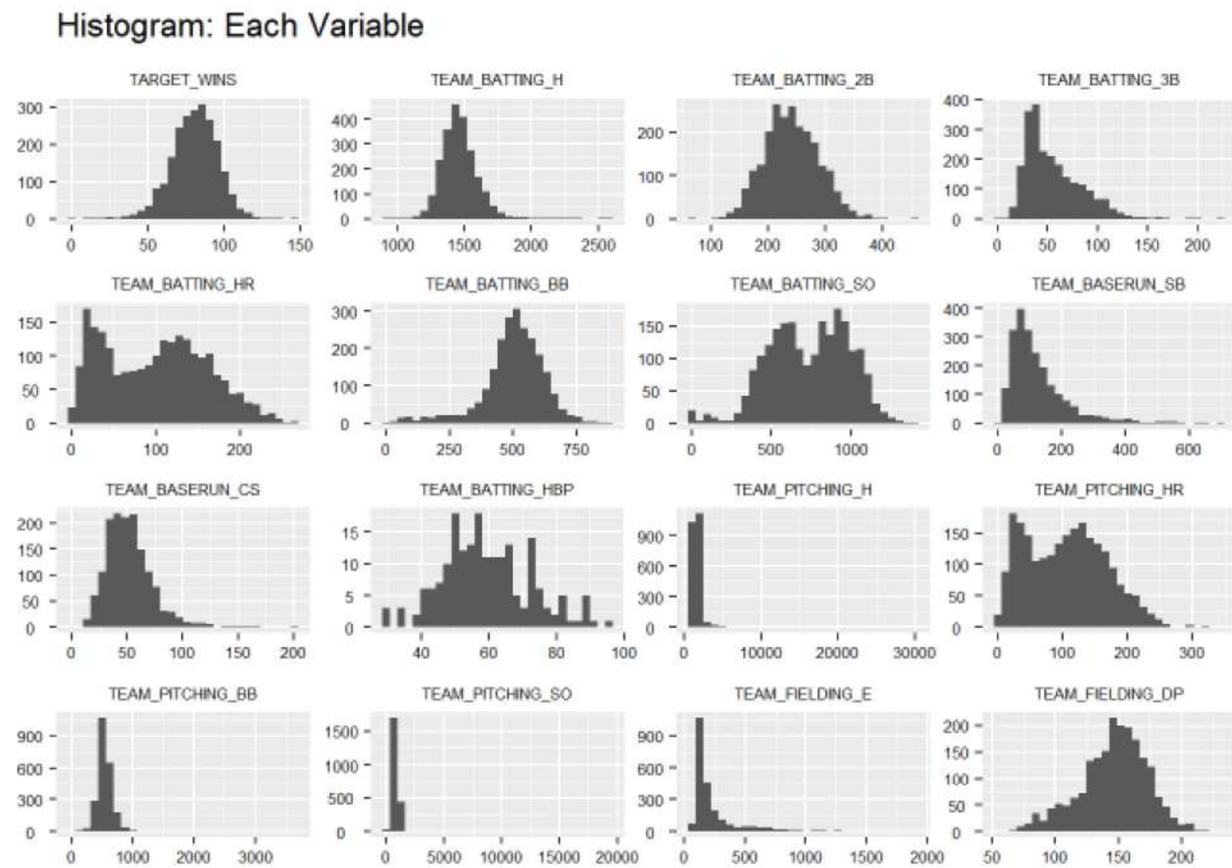
To generate overall averages for the `Moneyball` dataset, I have divided all the variables by 162 and generated the summary.

I have collected data from website https://www.baseball-reference.com/leagues/MLB/bat.shtml#all_teams_standard_batting (https://www.baseball-reference.com/leagues/MLB/bat.shtml#all_teams_standard_batting) and calculated average of average per year per game. Averages generated from `Moneyball` dataset are not exactly same but comparable.

However, further analysis is needed to check if data points are real outliers or influential leverage points.

Variable Name	Website AVg.	Dataset Avg.
TEAM_BATTING_H	9.15	9.07
TEAM_BATTING_2B	1.46	1.49
TEAM_BATTING_3B	0.37	0.34
TEAM_BATTING_HR	0.53	0.61
TEAM_BATTING_BB	2.96	3.10
TEAM_BATTING_HBP	0.24	0.37
TEAM_BATTING_SO	3.98	4.54
TEAM_BASERUN_SB	0.80	0.77
TEAM_BASERUN_CS	0.29	0.33
TEAM_FIELDING_E	1.81	1.52
TEAM_FIELDING_DP	0.85	0.90
TEAM_PITCHING_BB	2.96	3.41
TEAM_PITCHING_H	9.12	10.98
TEAM_PITCHING_HR	0.53	0.65
TEAM_PITCHING_SO	4.07	5.05





Top ten `TEAM_PITCHING_SO` values from the dataset, they look extremely high compared to mean (817.73) and median (813.50). Same is the case with `TEAM_PITCHING_H` comparing to mean (1779.21) and median (1518.00).

Top 10 Strikeouts by pitchers

TARGET_WINS	TEAM_PITCHING_SO
41	19278
108	12758
39	5456
46	4224
71	3450
31	2492
51	2367
95	2309
33	2225
75	1781

Top 10 Hits allowed

TARGET_WINS	TEAM_PITCHING_H
36	30132
0	24057
41	20088
23	16871
108	16038
44	14749
34	13898
97	13815
122	13724
60	12943

Correlation

In this analysis, I will checking relation between

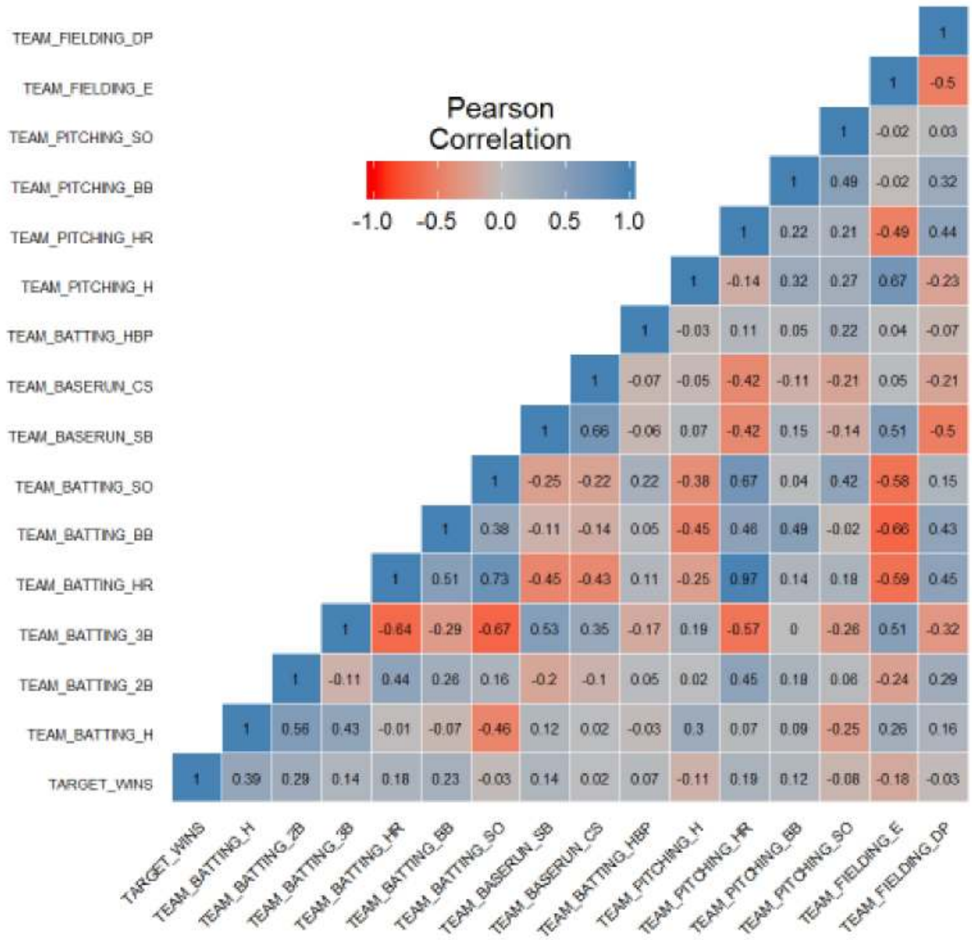
- `TARGET_WINS` and rest of variables, one to one correlation.
- Relation between variables and effect on `TARGET_WINS` .

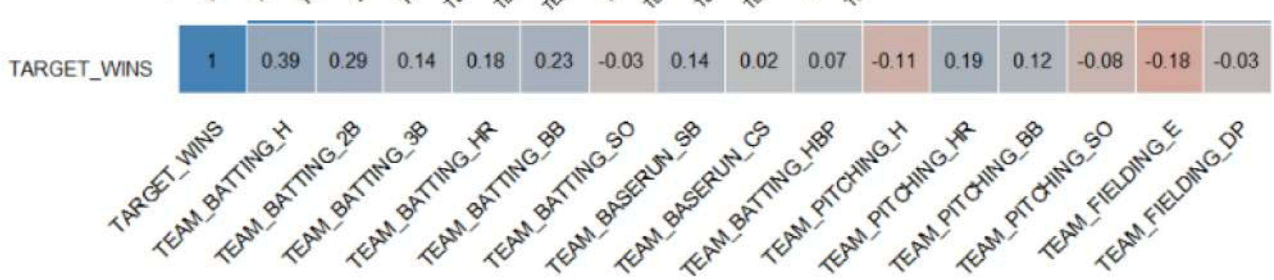
All variables are part of the analysis, except missing data is excluded using `pairwise.complete.obs` option of `cor` function. For the analysis `Pearson correlation` is used. The Pearson correlation is positive 1 in the case of a perfect direct (increasing) linear relationship (correlation), -1 in the case of a perfect decreasing (inverse) linear relationship.

Using information given, one would assume

Variables `TEAM_BATTING_H`, `TEAM_BATTING_2B`, `TEAM_BATTING_3B`, `TEAM_BATTING_HR`, `TEAM_BATTING_BB`, `TEAM_BATTING_HBP`, `TEAM_BASERUN_SB`, `TEAM_FIELDING_DP` and `TEAM_PITCHING_SO` should have coefficient closer to 1 as these variables contribute to win a game.

Variables `TEAM_BATTING_SO`, `TEAM_BASERUN_CS`, `TEAM_FIELDING_E`, `TEAM_PITCHING_BB`, `TEAM_PITCHING_H`, `TEAM_PITCHING_HR` should have coefficient closer to 0, if not negative as a positive higher coefficient of these variables impact winning chance inversely.





Let’s focus on the bottom part of the image; it gives information about the relationship between

TARGET_WINS and rest of the variables.

As assumed variables TEAM_BATTING_H , TEAM_BATTING_2B , TEAM_BATTING_3B , TEAM_BATTING_HR , TEAM_BATTING_BB and TEAM_BASERUN_SB show positive coefficient. As coefficient has low value, it is considered as a moderately strong relationship.

Variables TEAM_BATTING_HBP has a coefficient value of 0.07. This is very week relationship.

On the other hand, variables TEAM_FIELDING_DP and TEAM_PITCHING_SO shows negative coefficient indicating they are inversely related to TARGET_WINS .

Variables TEAM_BATTING_SO , TEAM_FIELDING_E and TEAM_PITCHING_H have negative coefficient. That means fewer strikeouts during batting, less fielding errors and fewer hits by opposite team during pitching leads to better chance of winning.

However, variables TEAM_BASERUN_CS , TEAM_PITCHING_BB , TEAM_PITCHING_HR are showing positive coefficient, prompting for further analysis.

Second part analysis,

Relationship between TEAM_BATTING_HR and TEAM_PITCHING_HR is 0.97, TEAM_BATTING_HR and TEAM_BATTING_SO is 0.67, TEAM_FIELDING_E and TEAM_PITCHING_H 0.67 and many others that are not zero indicate multicollinearity exists between variables. This means one of the variables needs to be dropped while constructing regression model.

Example: If a team has excellent batters hitting home runs may help the team win. On the other hand pitchers not giving away home runs also helps the team win.

For initial data exploration, we can conclude that multicollinearity exists between variables and needs to b analyzed further during the model building process.

Conclusion of Data Exploration

- Dataset has missing values in variables TEAM_BATTING_HBP , TEAM_BASERUN_CS , TEAM_FIELDING_DP , TEAM_BASERUN_SB , TEAM_BATTING_SO , TEAM_PITCHING_SO .
- Individual boxplots and histograms of variables suggest the existence of outliers. Since we are interested in analyzing net effect of variables on TARGET_WINS , there is not enough reason to doubt observations can be classified as outliers.
- Multicollinearity exists between variables. This needs further analysis may be computing Variance Inflation Factor(VIF).
- Except for TEAM_BATTING_HBP , all other missing variables may be replaced with mean or median . I lean towards using mean over median because this is observational data. These are the facts that have happened and impacted the outcome of the game. Replacing with mean gives a better picture of how average has impacted the output of the game. Median gives single observation value which may not be a true representation of the variable.
- Concerning the variable TEAM_BATTING_HBP , if no data is captured for 2085 observations that means it may be considered as low occurrence event. I lean towards the testing the impact of the variable on the model by replacing with zero and also mean . It minimal to no impact is observed the variable may be excluded from the model.

Data Preparation

Excepct for variable `TEAM_BATTING_HBP` , missing values for variables including `TEAM_FIELDING_DP` , `TEAM_BASERUN_SB` , `TEAM_BATTING_SO` , `TEAM_PITCHING_SO` will be replaced with `mean` .

Let's start with variable `TEAM_BATTING_HBP` and generate three models,

- Replacing the missing value with `zero` .
- Replacing the missing value with `mean`.
- Last one by excluding the variable completely.

Missing value will be replaced with `zero` , under the assumption that hit by pitch event has not happened for some of the teams during a season. Model results in R^2 value of 0.3221 and adjusted R^2 of 0.3176.

Using `mean` , R^2 value for the model is 0.3192 and adjusted R^2 is 0.3147.

Excluding the variable, from the model results in R^2 value of 0.3189 and adjusted R^2 value of 0.3147.

As variable has very less impact on the model, I lean towards excluding it from the model.

`TEAM_BATTING_H` is a combination of singles, doubles, triples and home runs. Currently, doubles, triples, and home runs have a separate column. Let's create separate column singles `TEAM_BATTING_1B` by simple arithmetic.

Additional variable log of `TEAM_BATTING_H` will be added to the dataset to check if we can derive the better model.

Conclusion of Data Preparation

- Since variable `TEAM_BATTING_HBP` has no impact on the regression model, it can be excluded from the model.
- Due to multicollinearity, variables `TEAM_PITCHING_HR` , `TEAM_BASERUN_CS` and `TEAM_PITCHING_BB` may be excluded from the model. Further analysis is required.
- Separate column for singles `TEAM_BATTING_1B` , derived from `TEAM_BATTING_H` will be added to the dataset.
- Separate column for a log of `TEAM_BATTING_H` will be added to the dataset.
- Missing values for variables `TEAM_FIELDING_DP` , `TEAM_BASERUN_SB` , `TEAM_BATTING_SO` , `TEAM_PITCHING_SO` will be replaced with `average` value.

Build Models

After excluding variable from `TEAM_BATTING_HBP` , we have 16 variables. Function `lm` will be used to build the models.

First Model

The first model will have all the variables except variable `TEAM_BATTING_H` . It will also have newly added variables `TEAM_BATTING_1B` and `TEAM_BATTING_H_Log`

```
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR +
TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BASERUN_CS + TEAM_PI
TCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDI
NG_E + TEAM_FIELDING_DP + TEAM_BATTING_1B + TEAM_BATTING_H_Log, data = Baseball
Df_New)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B +
##     TEAM_BATTING_H_Log, data = BaseballDf_New)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.897  -8.527   0.085   8.342  58.546
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.318e+02  2.007e+02   1.155 0.248243
## TEAM_BATTING_2B     5.043e-02  2.279e-02   2.213 0.027005 *
## TEAM_BATTING_3B     1.304e-01  2.638e-02   4.944 8.23e-07 ***
## TEAM_BATTING_HR     1.277e-01  3.766e-02   3.392 0.000707 ***
## TEAM_BATTING_BB     9.820e-03  5.847e-03   1.680 0.093174 .
## TEAM_BATTING_SO    -9.634e-03  2.566e-03  -3.755 0.000178 ***
## TEAM_BASERUN_SB     2.945e-02  4.462e-03   6.600 5.12e-11 ***
## TEAM_BASERUN_CS    -1.179e-02  1.614e-02  -0.731 0.464992
## TEAM_PITCHING_H    -8.643e-04  3.887e-04  -2.224 0.026278 *
## TEAM_PITCHING_HR    1.077e-02  2.464e-02   0.437 0.662174
## TEAM_PITCHING_BB    6.915e-04  4.186e-03   0.165 0.868819
## TEAM_PITCHING_SO    2.793e-03  9.200e-04   3.036 0.002421 **
## TEAM_FIELDING_E    -2.107e-02  2.480e-03  -8.496  < 2e-16 ***
## TEAM_FIELDING_DP   -1.207e-01  1.302e-02  -9.268  < 2e-16 ***
## TEAM_BATTING_1B     6.908e-02  2.055e-02   3.362 0.000788 ***
## TEAM_BATTING_H_Log -3.259e+01  3.162e+01  -1.031 0.302815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 2260 degrees of freedom
## Multiple R-squared:  0.3193, Adjusted R-squared:  0.3147
## F-statistic: 70.66 on 15 and 2260 DF,  p-value: < 2.2e-16
```

```
vif(lm.mb)
```

```
##    TEAM_BATTING_2B    TEAM_BATTING_3B    TEAM_BATTING_HR
##          15.218744           7.268542          69.554097
##    TEAM_BATTING_BB    TEAM_BATTING_SO    TEAM_BASERUN_SB
##           6.882220           5.196211           1.934691
##    TEAM_BASERUN_CS    TEAM_PITCHING_H   TEAM_PITCHING_HR
##           1.213424           4.001546          30.515705
##   TEAM_PITCHING_BB   TEAM_PITCHING_SO    TEAM_FIELDING_E
##           6.489533           3.308970           4.270620
##   TEAM_FIELDING_DP    TEAM_BATTING_1B TEAM_BATTING_H_Log
##           1.364599          93.908864         117.833419
```

The model summary shows newly created variable `TEAM_BATTING_H_Log` has high p-Value and missing asterisk or dot suggests that variable is not contributing to the model. Also, variance inflation factor value is very high 117.83 indicating it is highly correlated to other variables.

R^2 value is 0.32.

Second Model

Elimination of `TEAM_BATTING_H_Log` has increased standard error of intercept. R^2 value also decreased. However, VIF values dropped for some of the variables.

```
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR +
TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BASERUN_CS + TEAM_PI
TCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDI
NG_E + TEAM_FIELDING_DP + TEAM_BATTING_1B, data = BaseballDf_New)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_BASERUN_CS + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B,
##     data = BaseballDf_New)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.994  -8.576   0.136   8.345  58.628
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.502e+01  5.397e+00   4.636 3.75e-06 ***
## TEAM_BATTING_2B   2.818e-02  7.314e-03   3.853 0.000120 ***
## TEAM_BATTING_3B   1.087e-01  1.589e-02   6.841 1.01e-11 ***
## TEAM_BATTING_HR   1.012e-01  2.753e-02   3.677 0.000241 ***
## TEAM_BATTING_BB   1.042e-02  5.818e-03   1.790 0.073544 .
## TEAM_BATTING_SO  -9.349e-03  2.551e-03  -3.665 0.000253 ***
## TEAM_BASERUN_SB   2.949e-02  4.462e-03   6.610 4.78e-11 ***
## TEAM_BASERUN_CS  -1.188e-02  1.614e-02  -0.736 0.461905
## TEAM_PITCHING_H  -7.342e-04  3.676e-04  -1.997 0.045946 *
## TEAM_PITCHING_HR  1.480e-02  2.432e-02   0.609 0.542877
## TEAM_PITCHING_BB  8.891e-05  4.145e-03   0.021 0.982891
## TEAM_PITCHING_SO  2.843e-03  9.187e-04   3.095 0.001994 **
## TEAM_FIELDING_E  -2.112e-02  2.480e-03  -8.516  < 2e-16 ***
## TEAM_FIELDING_DP -1.210e-01  1.302e-02  -9.297  < 2e-16 ***
## TEAM_BATTING_1B   4.824e-02  3.687e-03  13.085  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 2261 degrees of freedom
## Multiple R-squared:  0.3189, Adjusted R-squared:  0.3147
## F-statistic: 75.63 on 14 and 2261 DF,  p-value: < 2.2e-16
```

```
vif(lm.mb)
```

```
##  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB
##         1.567593         2.637964        37.171331         6.814913
##  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_H
##         5.135548         1.934503         1.213395         3.579058
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##        29.744384         6.362956         3.299856         4.269149
## TEAM_FIELDING_DP  TEAM_BATTING_1B
##         1.363741         3.022832
```

Third Model

Let's remove `TEAM_BASERUN_CS`, `TEAM_PITCHING_HR`, and `TEAM_PITCHING_BB` variables as they are not contributing to the model and also VIF values are very high suggesting the existence of a correlation with other variables.

```
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR +
TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PI
TCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B, data = Baseba
llDf_New)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP +
##     TEAM_BATTING_1B, data = BaseballDf_New)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.899  -8.568   0.091   8.397  58.651
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.6666983  5.2220414   4.532 6.14e-06 ***
## TEAM_BATTING_2B   0.0279447  0.0072977   3.829 0.000132 ***
## TEAM_BATTING_3B   0.1109231  0.0156518   7.087 1.82e-12 ***
## TEAM_BATTING_HR   0.1182355  0.0087893  13.452  < 2e-16 ***
## TEAM_BATTING_BB   0.0107446  0.0033489   3.208 0.001354 **
## TEAM_BATTING_SO  -0.0093019  0.0024571  -3.786 0.000157 ***
## TEAM_BASERUN_SB   0.0287708  0.0042901   6.706 2.51e-11 ***
## TEAM_PITCHING_H  -0.0006920  0.0003211  -2.155 0.031253 *
## TEAM_PITCHING_SO  0.0028867  0.0006707   4.304 1.75e-05 ***
## TEAM_FIELDING_E  -0.0205973  0.0024120  -8.540  < 2e-16 ***
## TEAM_FIELDING_DP -0.1210083  0.0130082  -9.302  < 2e-16 ***
## TEAM_BATTING_1B   0.0484570  0.0036621  13.232  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.03 on 2264 degrees of freedom
## Multiple R-squared:  0.3186, Adjusted R-squared:  0.3153
## F-statistic: 96.25 on 11 and 2264 DF,  p-value: < 2.2e-16
```

```
vif(lm.mb)
```

```
##  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB
##         1.562085         2.560672         3.792309         2.259997
##  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H TEAM_PITCHING_SO
##         4.769685         1.790200         2.732711         1.760172
##  TEAM_FIELDING_E TEAM_FIELDING_DP  TEAM_BATTING_1B
##         4.041728         1.362628         2.984972
```

All the variables are contributing to the model; however, VIF value for `TEAM_BATTING_SO` is very high. Also, correlation analysis shows the relationship between `TEAM_BATTING_SO` and `TEAM_BATTING_HR` at 0.73.

Intercept value increased to 23.67. R^2 value decreased to 0.3186.

Let’s check for model assumptions using diagnostic plots

- Linearity, the relationship between `X` and the mean of `Y` is linear.
- Homoscedasticity, the variance of residual is the same for any value of X.
- Independence, observations are independent of each other.
- Normality, for any fixed value of `X` , `Y` is normally distributed.

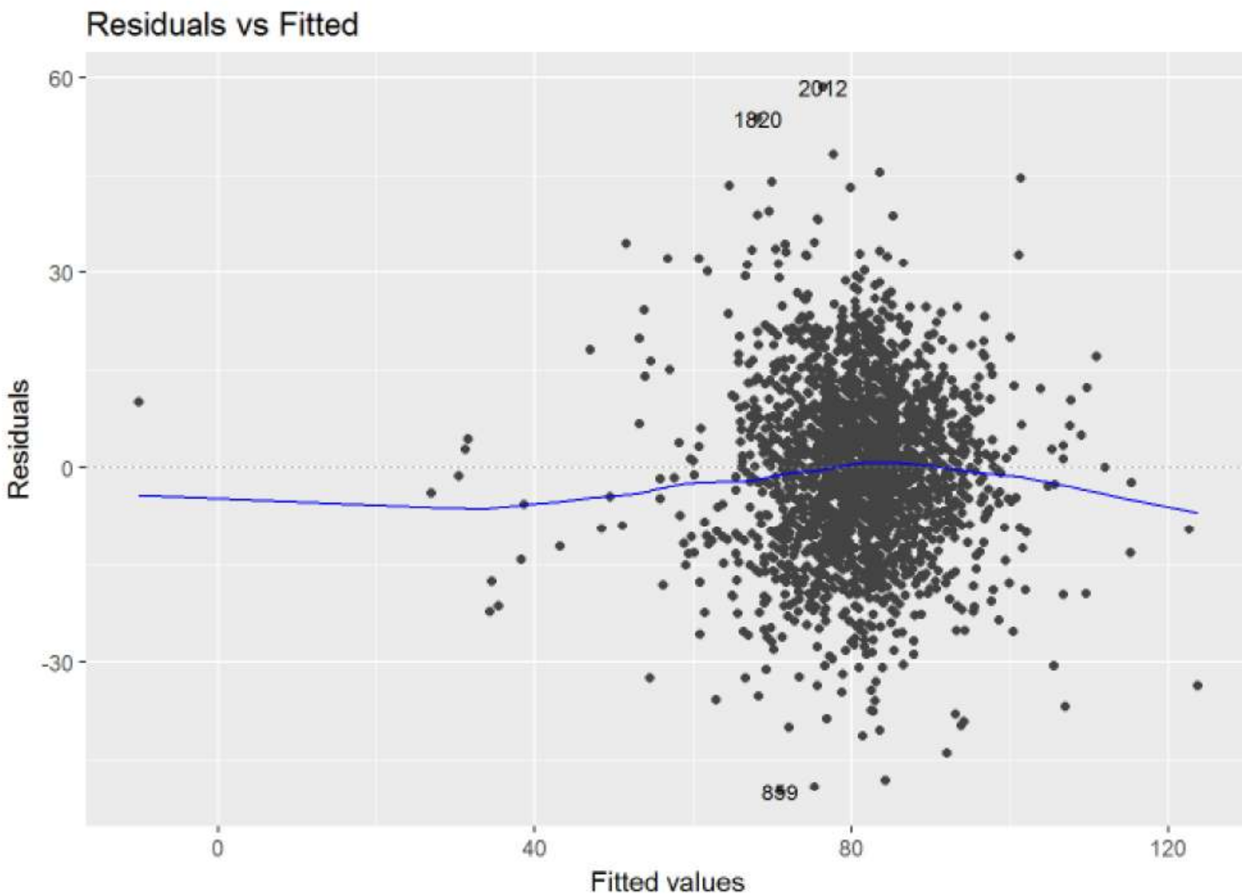
While checking above conditions, we will also check for `Outliers` , `Leverage points` and `Influential observations` .

Residuals vs Fitted Plot

- Blue line indicates fit line.
- Observations 2012, 1820 and 859 have high residual values. Observations need to analysed further

Upon looking at the data, variables `TEAM_BATTING_SO` and `TEAM_PITCHING_SO` have high degree of variance from `mean` . Also these observations had missing values in `TEAM_BASERUN_SB` and `TEAM_FIELDING_DP` .

Overall `Residuals vs Fitted` plot looks normal.



	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB
859	21	1402	149	53	13	304	295	134.00
1820	122	1428	221	62	30	434	678	124.76
2012	135	1793	371	59	46	259	777	124.76

	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
859	52.8	0	1475	14	320	310	408	146.39
1820	52.8	0	2066	43	628	981	576	146.39
2012	52.8	0	2570	66	371	1114	794	146.39

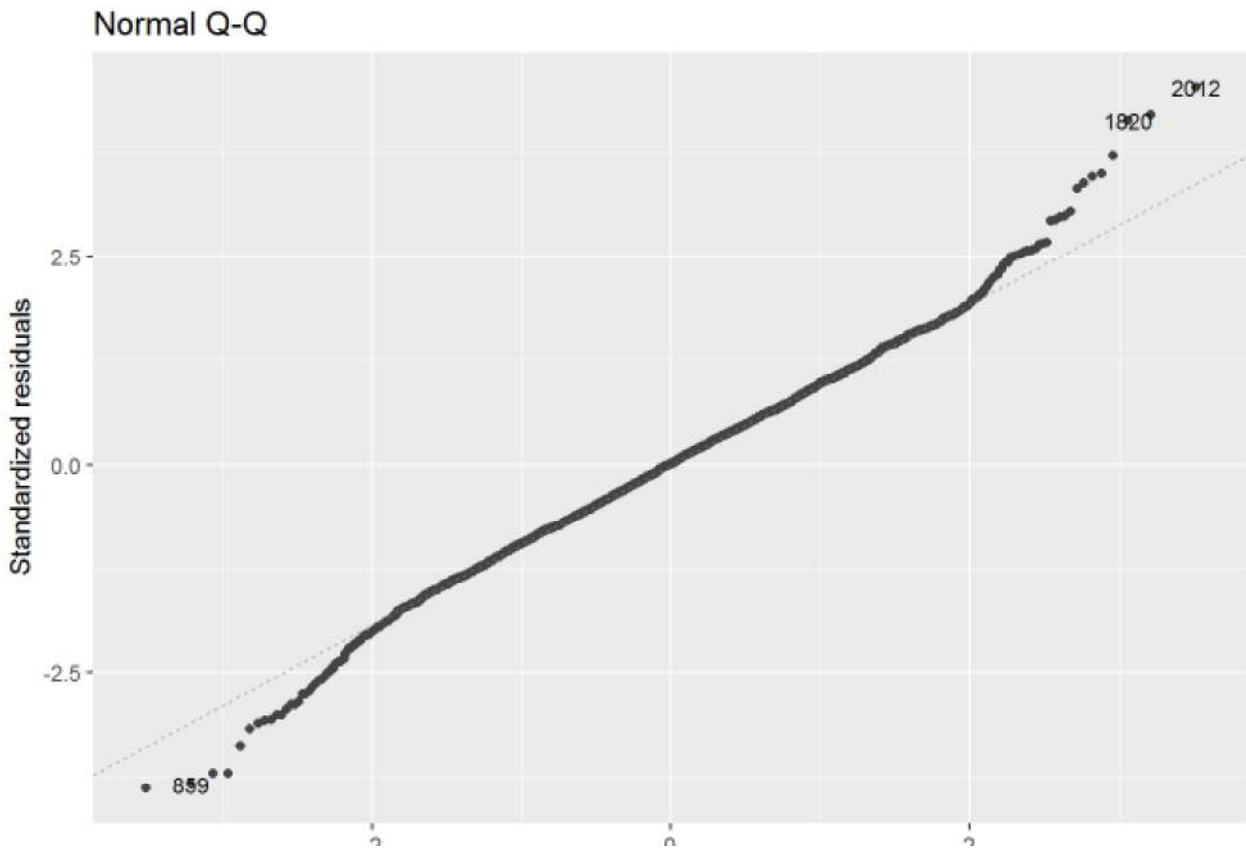
	resid	fitted
859	-49.99	70.99
1820	53.87	68.13
2012	58.63	76.37

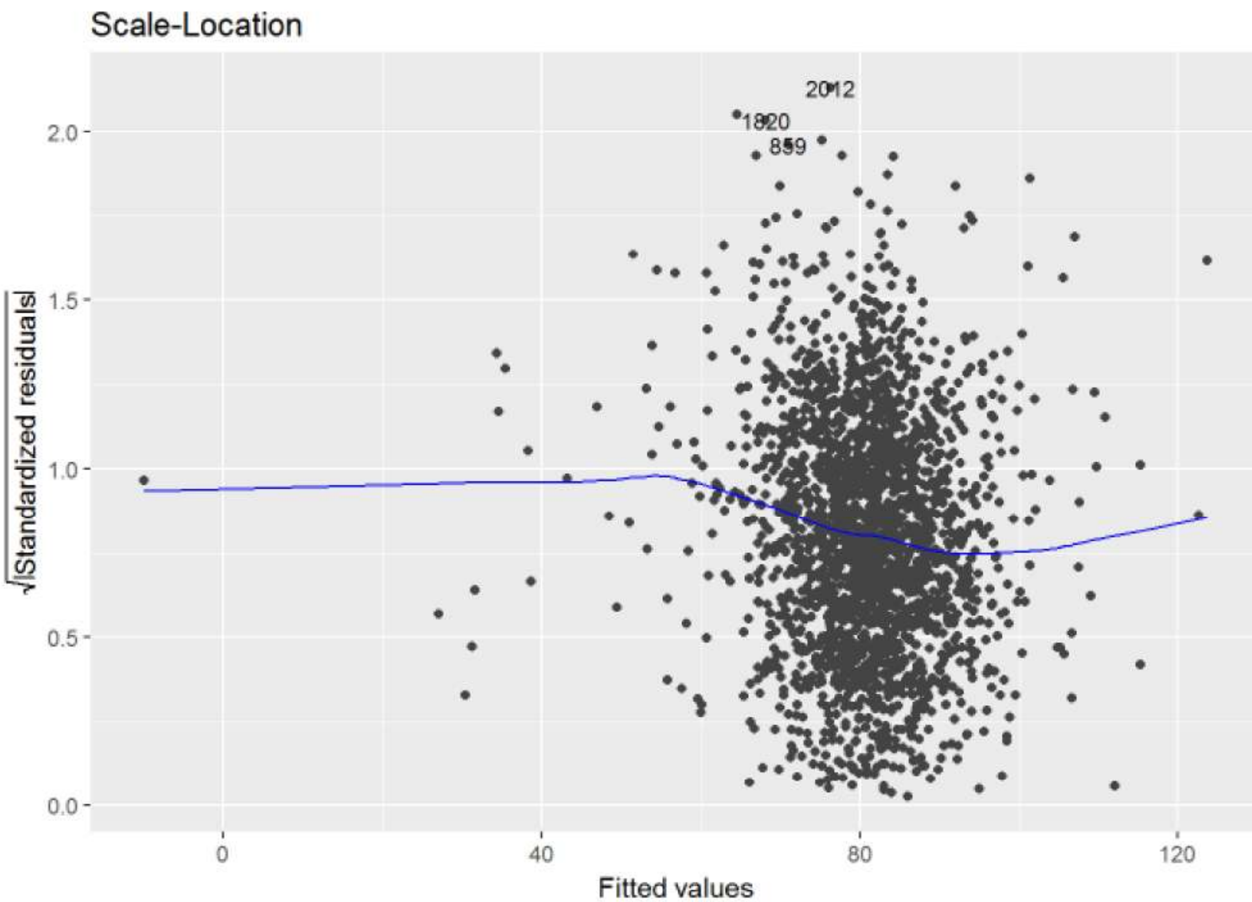
Normal Q-Q Plot

This plot checks if residuals are following normal distribution, `Normality Test` . Plot also identifies observations 2012, 1820 and 859 as outliers. All observations follow normal distribution. We can conclude residuals follow normal distribution.

Scale - Location Plot

This plot checks variation of observations around the regression line, the residual `Standard Error`. The plot also identifies observations 2012, 1820 and 859 as outliers. Since residuals spread is not wide enough, and direction is the line is not going up, we may conclude it satisfies `Homoscedasticity` conditions.



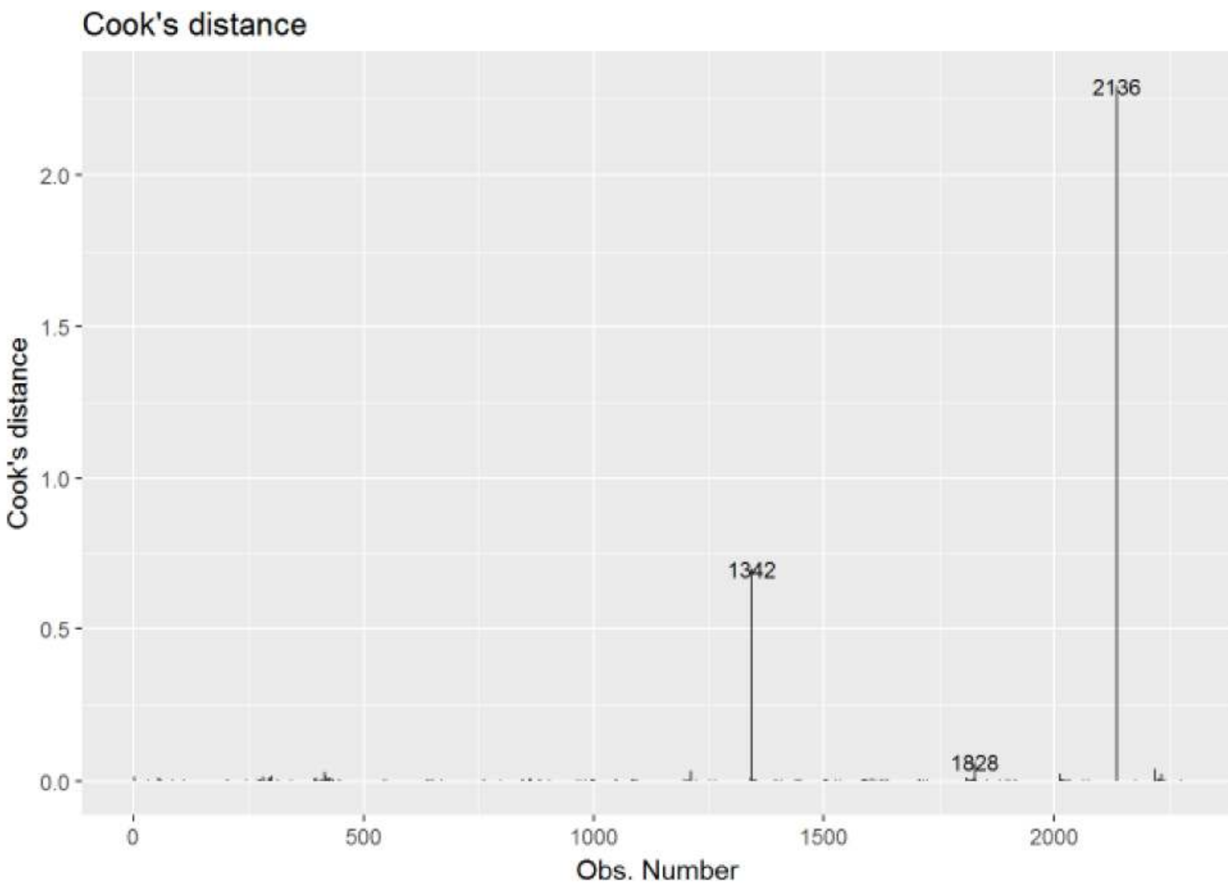


Cook's Distance

This plot explains observations that are strongly influencing the fitted values. The plot identifies observations `1342` , `1828` and `2136` has strong influence on fitted values.

Variables `TEAM_BATTING_3B` , `TEAM_BATTING_HR` , `TEAM_BATTING_BB` , `TEAM_BATTING_SO` , `TEAM_PITCHING_H` , `TEAM_PITCHING_HR` , `TEAM_PITCHING_BB` and `TEAM_PITCHING_SO` highly deviate from their respective `mean` .

Also variables `TEAM_BASERUN_CS` and `TEAM_FIELDING_DP` were missing values and were replaced with `mean` .



	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB
1342	108	1188	338	0	0	270	945	124.76
1828	26	1776	285	162	19	246	194	343.00
2136	41	992	263	20	0	142	952	124.76

	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
1342	52.8	0	16038	0	3645	12758	716	146.39
1828	52.8	0	11508	123	1594	1257	1426	146.39
2136	52.8	0	20088	0	2876	19278	952	146.39

	resid	fitted	cooks.distance
1342	43.43	64.57	0.70
1828	-49.27	75.27	0.06
2136	-25.94	66.94	2.29

Fourth Model

Let’s remove all the influential observations from the dataset and rerun the model once again. I will be using `influence.measures` function.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
##     TEAM_PITCHING_H + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP +
##     TEAM_BATTING_1B, data = nonInfDf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.233  -7.887   0.259   7.680  32.542
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      41.386512   5.483557   7.547 6.56e-14 ***
## TEAM_BATTING_2B  -0.013302   0.007135  -1.864  0.06239 .
## TEAM_BATTING_3B   0.142625   0.016215   8.796  < 2e-16 ***
## TEAM_BATTING_HR   0.106706   0.008681  12.292  < 2e-16 ***
## TEAM_BATTING_BB   0.022902   0.003043   7.525 7.73e-14 ***
## TEAM_BATTING_SO   0.005105   0.006389   0.799  0.42438
## TEAM_BASERUN_SB   0.043330   0.004394   9.862  < 2e-16 ***
## TEAM_PITCHING_H   0.015501   0.002328   6.658 3.52e-11 ***
## TEAM_PITCHING_SO -0.015244   0.005712  -2.669  0.00767 **
## TEAM_FIELDING_E  -0.039551   0.003266 -12.108  < 2e-16 ***
## TEAM_FIELDING_DP -0.112225   0.011545  -9.721  < 2e-16 ***
## TEAM_BATTING_1B   0.014958   0.005180   2.887  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.03 on 2113 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3414
## F-statistic: 101.1 on 11 and 2113 DF,  p-value: < 2.2e-16
```

After removing all influential R^2 value improved.

Select Models

After looking at the summary of four models, I would select *third model* because it has better β values and more closely relates to the game.

Linear equation is TARGET_WIN = 23.6667 + 0.0279 * TEAM_BATTING_2B + 0.1109 * TEAM_BATTING_3B + 0.1182 * TEAM_BATTING_HR + 0.0107 * TEAM_BATTING_BB - 0.0093 * TEAM_BATTING_SO + 0.0288 * TEAM_BASERUN_SB - 0.0007 * TEAM_PITCHING_H + 0.0029 * TEAM_PITCHING_SO - 0.0206 * TEAM_FIELDING_E - 0.1210 * TEAM_FIELDING_DP + 0.0485 * TEAM_BATTING_1B

Equation explains by increasing singles, doubles, triples, home runs by a fraction and reducing strikeouts during batting will improve teams winning chances. Also, equation suggests during pitching, if team reduces hits, fielding errors and double plays it will also enhance winning chances.

Variables `TEAM_BATTING_SO` and `TEAM_BATTING_HR` has high correlation at 0.73. It means during batting if strikeouts increase home runs also increase. Logically, it does not make any sense.

Even though value is less, baseball domain knowledge plays a major role in selecting the model.

Using `predict` function we can predict winning changes of the teams. Following table shows prediction at 95% confidence interval for 30 records.

Moneyball Dataset Prediction - Using Evaluation Data

	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_PITCHING_H	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP	TEAM_BATTING_1B	fit	lwr	upr
1	170	33	83	447	1080	62	1209	1080	140	156	923	64	38	89
2	151	29	88	516	929	54	1221	929	135	164	953	65	40	91
3	183	29	93	509	816	59	1395	816	156	153	1090	75	49	101
4	309	29	159	486	914	148	1539	914	124	154	1042	86	61	112
8	158	42	33	356	609	185	1626	715	328	104	1152	78	52	103
9	177	78	23	466	689	150	1342	734	226	132	981	71	45	97
10	212	42	58	452	584	52	1489	622	184	145	1085	74	48	100
11	243	40	50	495	640	64	1501	673	200	183	1094	70	44	95
12	239	55	184	482	670	48	1574	705	150	178	1038	83	57	108
13	223	57	186	511	751	31	1494	790	137	167	954	82	56	108
14	232	22	178	503	680	27	1536	715	125	160	1030	82	57	108
15	195	22	141	485	665	59	1411	665	115	114	1053	85	59	110
16	192	30	153	434	747	57	1434	747	146	180	1059	77	52	103
17	204	22	130	491	1008	84	1313	1021	154	126	941	75	49	100
18	284	25	186	565	1041	77	1464	1054	115	172	971	79	53	104
20	322	72	116	527	397	90	1816	420	232	174	1205	91	66	117
21	295	68	49	628	459	77	1620	489	166	158	1108	81	56	107
22	291	38	98	629	563	54	1702	600	155	174	1170	84	58	109
23	256	67	105	653	651	40	1559	698	179	153	1025	81	56	107
24	225	26	118	533	677	18	1450	712	160	174	1009	72	47	98
25	277	24	152	431	902	89	1516	902	105	164	1063	82	56	107
26	288	20	184	474	878	121	1556	878	102	156	1084	87	61	112
28	263	58	47	385	479	63	1540	504	232	146	1096	76	50	101
29	318	66	32	634	439	83	1639	462	216	130	1142	84	58	110
30	308	38	39	432	802	45	1601	642	199	135	1119	76	50	101

References

- http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization)
- https://www.youtube.com/watch?v=IxbPk0b_fiY (https://www.youtube.com/watch?v=IxbPk0b_fiY)
- http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html (http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html)
- https://web.stanford.edu/class/stats191/notebooks/Diagnostics_for_multiple_regression.html (https://web.stanford.edu/class/stats191/notebooks/Diagnostics_for_multiple_regression.html)
- https://onlinecourses.science.psu.edu/stat501/ (https://onlinecourses.science.psu.edu/stat501/)
- https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_lm.html (https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_lm.html)
- http://analyticspro.org/2016/03/07/r-tutorial-how-to-use-diagnostic-plots-for-regression-models/ (http://analyticspro.org/2016/03/07/r-tutorial-how-to-use-diagnostic-plots-for-regression-models/)
- http://data.library.virginia.edu/diagnostic-plots/ (http://data.library.virginia.edu/diagnostic-plots/)
- http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/ (http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/)
- https://datascienceplus.com/missing-value-treatment/ (https://datascienceplus.com/missing-value-treatment/)
- https://onlinecourses.science.psu.edu/stat501/node/429 (https://onlinecourses.science.psu.edu/stat501/node/429)

Appendix

```
#Libraries used

library(VIM)
library(ggplot2)
library(reshape2)
library(car)
library(fBasics)
library(knitr)
library(kableExtra)
library(dplyr)
library(ggfortify)

# Load data
BaseballDf <- read.csv("C:\\Pavan\\CUNY\\621\\moneyball-training-data.csv", hea
der= TRUE, stringsAsFactors = F)

# Get missing values
aggr_plot <- aggr(BaseballDf,
            numbers=TRUE, sortVars=TRUE,
            labels=names(BaseballDf), cex.axis=.45,
            gap=3, ylab=c("Missing data","Pattern"))

summary(aggr_plot)

# Get 162 game averages
BaseballDf_D <- BaseballDf[, 1:ncol(BaseballDf)]/162

tmp <-basicStats(BaseballDf)
tmp <- data.frame(t(tmp))
tmp <- tmp[ , -which(names(tmp) %in% c("SE.Mean","LCL.Mean","UCL.Mean"))]
colnames(tmp)[which(names(tmp) == "X1..Quartile")] <- "1st. Quartile"
colnames(tmp)[which(names(tmp) == "X3..Quartile")] <- "3st. Quartile"
colnames(tmp)[which(names(tmp) == "nobs")] <- "Observations"
tmp %>%
  kable(format="html", digits= 2, caption = "Moneyball Dataset Summary") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "respons
ive"), full_width = F, position = "left")

tmp1<-basicStats(BaseballDf_D)
tmp1 <- data.frame(t(tmp1))

tmp1 <- tmp1[ , -which(names(tmp1) %in% c("SE.Mean","LCL.Mean","UCL.Mean"))]
colnames(tmp1)[which(names(tmp1) == "X1..Quartile")] <- "1st. Quartile"
colnames(tmp1)[which(names(tmp1) == "X3..Quartile")] <- "3st. Quartile"
colnames(tmp1)[which(names(tmp1) == "nobs")] <- "Observations"

tmp1 %>%
  kable(format="html", digits= 2, caption = "Moneyball Dataset Per Game Summar
```

```
y") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "respons
ive"), full_width = F, position = "left")


#Generate Histograms for each column
meltData <- melt(BaseballDf, na.rm = TRUE)
p <- ggplot(meltData, aes(factor(variable), value))
p + geom_boxplot() + facet_wrap(~variable, scale="free", nrow = 8, ncol = 2) +
  theme(axis.text.x = element_text(vjust = 0.5, size = 6, hjust = 0.5, colour
= 'black')) +
  theme(axis.text.y = element_text(vjust = 0.5, size = 6, hjust = 0.5, colour
= 'black')) +
  labs(title="Boxplot: Each Variable",x="", y="") +
  theme(strip.background = element_blank(),  strip.text.x = element_blank()) +
coord_flip()

p <- ggplot(meltData,aes(x = value)) +
    geom_histogram() + facet_wrap(~variable,scales = "free", nrow = 4, ncol =
4) +
  theme(axis.text.x = element_text(vjust = 0.5, size = 6, hjust = 0.5, colour
= 'black')) +
  theme(axis.text.y = element_text(vjust = 0.5, size = 6, hjust = 0.5, colour
= 'black')) +
  labs(title="Histogram: Each Variable",x="", y="") +
  theme(strip.background = element_blank(),  strip.text.x = element_text(vjust
= 0.5, size = 6, hjust = 0.5, colour = 'black'))
p

#Generate correlation heatmap
# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)

BaseballDf_New <- BaseballDf[,]

for(i in 1:ncol(BaseballDf_New)){
  BaseballDf_New[is.na(BaseballDf_New[,i]), i] <- mean(BaseballDf_New[,i], na.r
m = TRUE)
}

cormat <- round(cor(BaseballDf_New, method="pearson"),2)
upper_tri <- get_upper_tri(cormat)
melted_cormat <- melt(upper_tri, na.rm = TRUE)

ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "red", high = "steelblue", mid = "gray",
```

```
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
 theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 5.5, hjust =
1, colour = 'black'))+
 theme(axis.text.y = element_text(vjust = 1, size = 5.5, hjust = 1, colour = 'b
lack'))+
 coord_fixed()

ggheatmap +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 2) +
theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
  legend.position = c(0.6, 0.7),
  legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                title.position = "top", title.hjust = 0.5))


#Generate various LM models
BaseballDf_New <- BaseballDf[,]

BaseballDf_New <- BaseballDf_New %>%
   mutate(TEAM_BATTING_HBP = replace(TEAM_BATTING_HBP,is.na(TEAM_BATTING_HBP),
0))

for(i in 1:ncol(BaseballDf_New)){
  BaseballDf_New[is.na(BaseballDf_New[,i]), i] <- mean(BaseballDf_New[,i], na.r
m = TRUE)
}

BaseballDf_New$TEAM_BATTING_1B = BaseballDf_New$TEAM_BATTING_H - BaseballDf_New
$TEAM_BATTING_2B - BaseballDf_New$TEAM_BATTING_3B - BaseballDf_New$TEAM_BATTING
_HR

BaseballDf_New$TEAM_BATTING_H_Log = log(BaseballDf_New$TEAM_BATTING_H)

#First model
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR
+ TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BASERUN_CS + TEAM_
PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIEL
DING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B + TEAM_BATTING_H_Log, data = Baseba
llDf_New)
```

```
summary(lm.mb)
vif(lm.mb)

#Second model
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR
+ TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BASERUN_CS + TEAM_
PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIEL
DING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B, data = BaseballDf_New)

summary(lm.mb)
vif(lm.mb)

#Third model
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR
+ TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_
PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B, data = Base
ballDf_New)

summary(lm.mb)
vif(lm.mb)

#Generate diagnostic plots
autoplot(lm.mb, which = 1:6, ncol = 3, label.size = 3)
autoplot(lm.mb, which = 1, ncol = 1, label.size = 3)
autoplot(lm.mb, which = 2, ncol = 1, label.size = 3)
autoplot(lm.mb, which = 3, ncol = 1, label.size = 3)
autoplot(lm.mb, which = 4, ncol = 1, label.size = 3)
autoplot(lm.mb, which = 5, ncol = 1, label.size = 3)
autoplot(lm.mb, which = 6, ncol = 1, label.size = 3)

#Get residuals, fitted values, cook's distance and leverage values
BaseballDf_New$resid <- resid(lm.mb)
BaseballDf_New$fitted <- fitted(lm.mb)
BaseballDf_New$cooks.distance <- cooks.distance(lm.mb)
BaseballDf_New$Leverage <- hatvalues(lm.mb)

#Get all influance measures
infDf <- data.frame(summary(influence.measures(lm.mb)))

#Remove observation that influence the model
nonInfDf<-BaseballDf_New[!rownames(BaseballDf_New) %in% c(row.names(infDf)),]

#Fourth model
lm.mb <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR
+ TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_
PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_1B, data = nonI
nfDf)
```

```
summary(lm.mb)
vif(lm.mb)

#Generate prediction
BaseballpDf <- read.csv("C:\\Pavan\\CUNY\\621\\moneyball-evaluation-data.csv",
header= TRUE, stringsAsFactors = F)
bpdf <- BaseballpDf %>% select(
TEAM_BATTING_2B , TEAM_BATTING_3B , TEAM_BATTING_HR , TEAM_BATTING_BB , TEAM_BA
TTING_SO , TEAM_BASERUN_SB , TEAM_PITCHING_H , TEAM_PITCHING_SO , TEAM_FIELDING
_E , TEAM_FIELDING_DP, TEAM_BATTING_1B)

bpdf <- bpdf[complete.cases(bpdf), ]

PI <- predict(lm.mb, bpdf, interval="predict", level=.95)
PI <- data.frame(PI)
PI <- round(PI,0)
bpdf <- cbind(bpdf,PI)
```