Assignment #1

Nitin Gaonkar PREDICT 411 section 56

INTRODUCTION

The purpose of this project is to analyze the data of the professional baseball team from the years 1871 to 2006 and use OLS(linear) regression to predict the number of wins for the team. This will be achieved by building regression models by using the various regression techniques like dummy coding, automatic variable selection and dimensionality reduction. The models will then be compared to get the best fit. After getting the best model it will be further analyzed to see if its the best fit for predicting the wins.

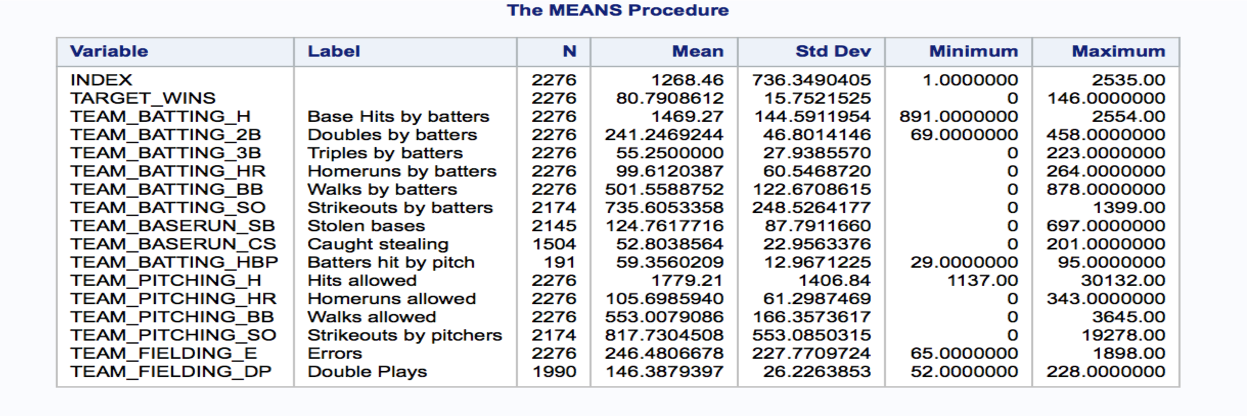
1. Data Exploration:

The money ball dataset has around 2200 records and each record represents a professional baseball team from the years 1871 to 2006, There are three data sets provided, one is the training data set containing around 2200 observations, the other is the test data set contains 259 records and the last one is the random test with index and target wins, we begin our data exploration by examining the data dictionary and the definitions given in the dictionary, after observing the data dictionary, we see that all the variables in the dataset are continuous.

a. In this data set the variable are defined in the different category:

Target variable	Predictor variables	Continuous variables
TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_H
	TEAM_BATTING_2B	TEAM_BATTING_2B
	TEAM_BATTING_3B	TEAM_BATTING_3B
	TEAM_BATTING_HR	TEAM_BATTING_HR
	TEAM_BATTING_BB	TEAM_BATTING_BB
	TEAM_BATTING_HBP	TEAM_BATTING_HBP
	TEAM_BATTING_SO	TEAM_BATTING_SO
	TEAM_BASERUN_SB	TEAM_BASERUN_SB
	TEAM_BASERUN_CS	TEAM_BASERUN_CS
	TEAM_FIELDING_E	TEAM_FIELDING_E
	TEAM_FIELDING_DP	TEAM_FIELDING_DP
	TEAM_PITCHING_BB	TEAM_PITCHING_BB
	TEAM_PITCHING_H	TEAM_PITCHING_H
	TEAM_PITCHING_HR	TEAM_PITCHING_HR
	TEAM_PITCHING_SO	TEAM_PITCHING_SO
		TARGET_WINS

b. Just to give a bit insight on the data, I have calculated and listed the mean and the standard deviation along with the min and max value of each variable in the below table.



The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
INDEX		2276	1268.46	736.3490405	1.0000000	2535.00
TARGET_WINS		2276	80.7908612	15.7521525	0	146.0000000
TEAM_BATTING_H	Base Hits by batters	2276	1469.27	144.5911954	891.0000000	2554.00
TEAM_BATTING_2B	Doubles by batters	2276	241.2469244	46.8014146	69.0000000	458.0000000
TEAM_BATTING_3B	Triples by batters	2276	55.2500000	27.9385570	0	223.0000000
TEAM_BATTING_HR	Homeruns by batters	2276	99.6120387	60.5468720	0	264.0000000
TEAM_BATTING_BB	Walks by batters	2276	501.5588752	122.6708615	0	878.0000000
TEAM_BATTING_SO	Strikeouts by batters	2174	735.6053358	248.5264177	0	1399.00
TEAM_BASERUN_SB	Stolen bases	2145	124.7617716	87.7911660	0	697.0000000
TEAM_BASERUN_CS	Caught stealing	1504	52.8038564	22.9563376	0	201.0000000
TEAM_BATTING_HBP	Batters hit by pitch	191	59.3560209	12.9671225	29.0000000	95.0000000
TEAM_PITCHING_H	Hits allowed	2276	1779.21	1406.84	1137.00	30132.00
TEAM_PITCHING_HR	Homeruns allowed	2276	105.6985940	61.2987469	0	343.0000000
TEAM_PITCHING_BB	Walks allowed	2276	553.0079086	166.3573617	0	3645.00
TEAM_PITCHING_SO	Strikeouts by pitchers	2174	817.7304508	553.0850315	0	19278.00
TEAM_FIELDING_E	Errors	2276	246.4806678	227.7709724	65.0000000	1898.00
TEAM_FIELDING_DP	Double Plays	1990	146.3879397	26.2263853	52.0000000	228.0000000

c. Below corr procedures gives the correlation of each variable with the target wins, as we can see that there are no variables that are highly correlated with the target wins, but we can see that the variables with negative corr already have a negative theoretical effect, also after having a closer look at the data we could also see that there are few relationships between the variables itself for example, hits gained/hits allowed, homerun gained/home run allowed etc.

Variable name	Correlation	THEORETICAL EFFECT
TEAM_BATTING_H	0.38877	Positive Impact on Wins
TEAM_BATTING_2B	0.2891	Positive Impact on Wins
TEAM_BATTING_3B	0.14261	Positive Impact on Wins
TEAM_BATTING_HR	0.17615	Positive Impact on Wins
TEAM_BATTING_BB	0.23256	Positive Impact on Wins
TEAM_BATTING_HBP	0.0735	Positive Impact on Wins
TEAM_BATTING_SO	-0.03175	Negative Impact on Wins
TEAM_BASERUN_SB	0.13514	Positive Impact on Wins
TEAM_BASERUN_CS	0.224	Negative Impact on Wins
TEAM_FIELDING_E	-0.17648	Negative Impact on Wins
TEAM_FIELDING_DP	-0.03485	Positive Impact on Wins
TEAM_PITCHING_BB	0.12417	Negative Impact on Wins
TEAM_PITCHING_H	-0.10994	Negative Impact on Wins
TEAM_PITCHING_HR	0.18901	Negative Impact on Wins
TEAM_PITCHING_SO	-0.07844	Positive Impact on Wins

d. From the below histogram we can see the distribution of the target wins in our dataset, this shows that the target wins are normally distributed.

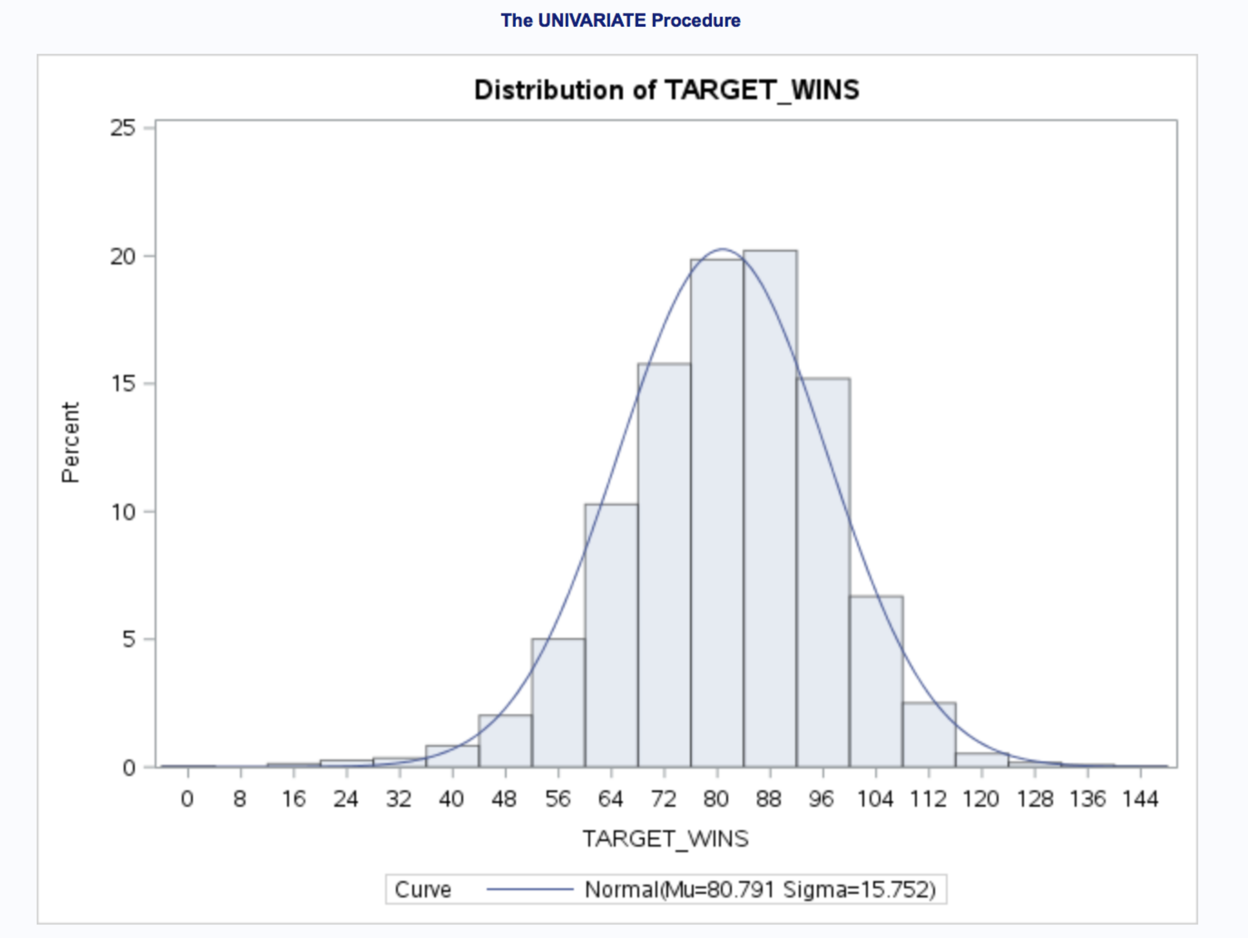
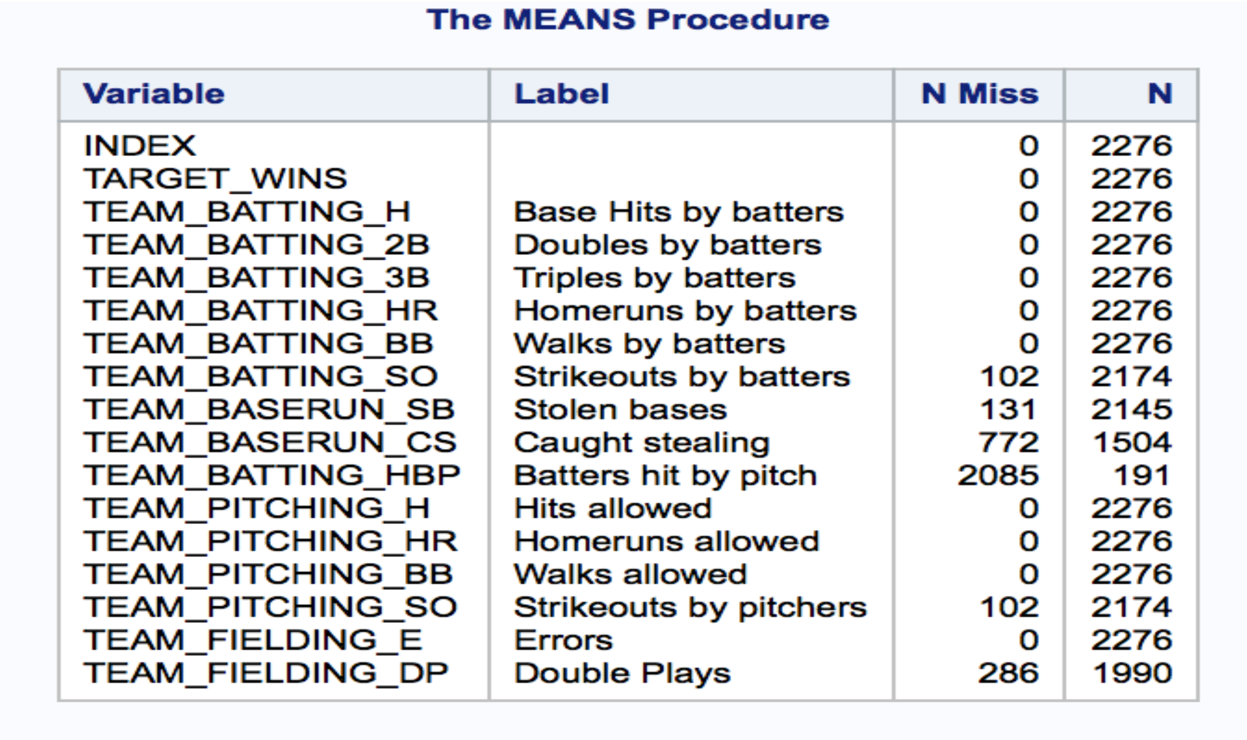


Fig :1 distribution of Target_wins

e. Missing data:
By running the proc means procedure we could find the missing data, from the below table we can see that there are few missing values in the dataset, before going ahead with modelling we have to decide on whether we will be using that variable in the modelling or not, if yes, then we have to impute the variable as our modelling technique will not handle the missing values.



The MEANS Procedure

Variable	Label	N Miss	N
INDEX		0	2276
TARGET_WINS		0	2276
TEAM_BATTING_H	Base Hits by batters	0	2276
TEAM_BATTING_2B	Doubles by batters	0	2276
TEAM_BATTING_3B	Triples by batters	0	2276
TEAM_BATTING_HR	Homeruns by batters	0	2276
TEAM_BATTING_BB	Walks by batters	0	2276
TEAM_BATTING_SO	Strikeouts by batters	102	2174
TEAM_BASERUN_SB	Stolen bases	131	2145
TEAM_BASERUN_CS	Caught stealing	772	1504
TEAM_BATTING_HBP	Batters hit by pitch	2085	191
TEAM_PITCHING_H	Hits allowed	0	2276
TEAM_PITCHING_HR	Homeruns allowed	0	2276
TEAM_PITCHING_BB	Walks allowed	0	2276
TEAM_PITCHING_SO	Strikeouts by pitchers	102	2174
TEAM_FIELDING_E	Errors	0	2276
TEAM_FIELDING_DP	Double Plays	286	1990

Based on the above table we can tell that there are 6 variables which are missing values, after reviewing the correlation table we can say that the batting_so, baserun_cs, fielding_dp and pitching_so have a low correlation with the target_wins, only variable which has missing values and has a better correlation and positive impact on win, thus we will impute the baserun_sb variable with its mean 124.7617716, we will create two new variables in this process, one with IP_* prefix for the imputed variables and one with flag as a indicator variable for the imputed variable.

f. In this section we will explore the distribution of variables, after going through all the distribution of the variables we could see that there are few variables which have extreme values and the high skewness. Below we will discuss few of the variables with the extreme values and skewness.

Team_baserun_sb:

As we can observe from the below graph that the mean is around 124 but there is long tail for this graph and the skewness is about 1.97 which is high, thus we would definitely would require to have a look at this variable.

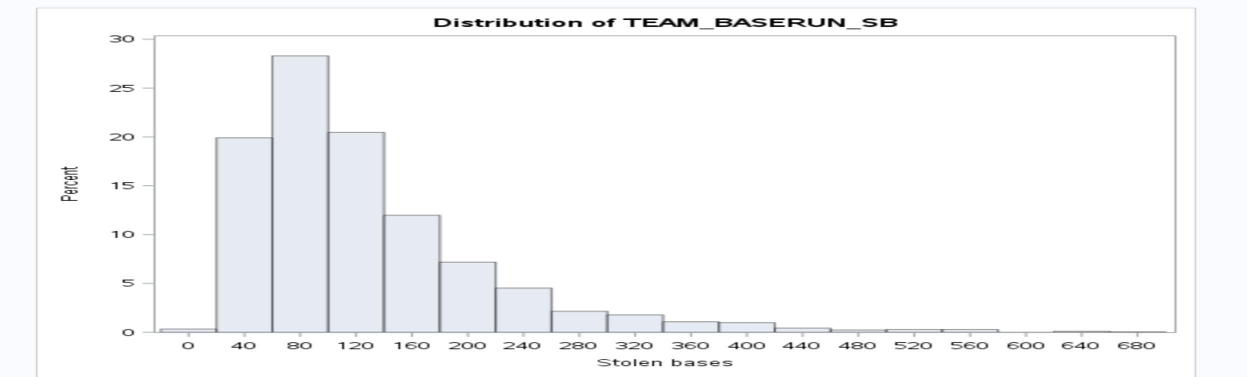


Fig :2 distribution of Team_baserun_sb

Team_fielding_error:

As we can observe from the below graph that the mean is around 246 but there is long tail for this graph and the skewness is about 2.99 which is high, thus we would definitely would require to have a look at this variable as well.

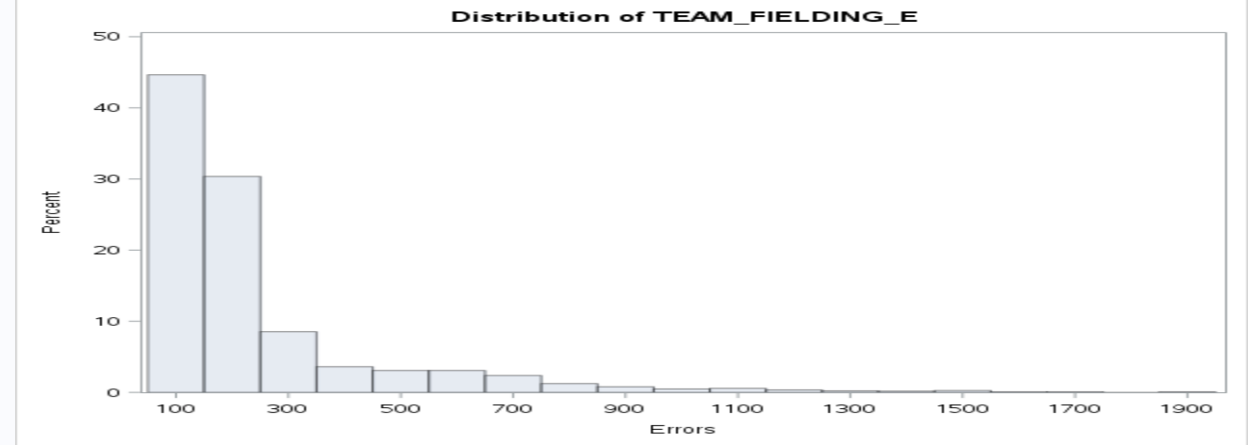


Fig :3 distribution of Team_fielding_e

Team_pitching_bb:

As we can observe from the below graph that the mean is around 553 but there is long tail for this graph and the skewness is about 6 which is high, thus we would definitely would require to have a look at this variable as well.

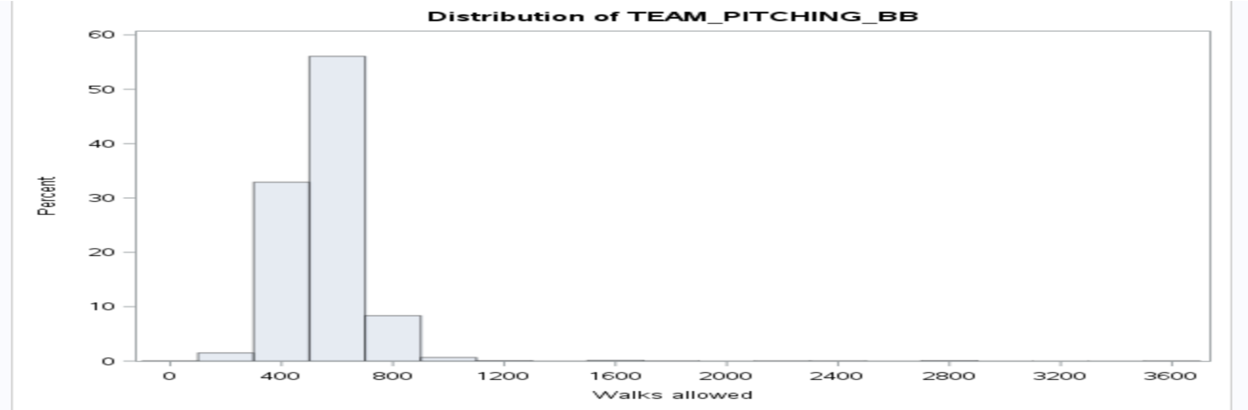


Fig :4 distribution of Team_pitching_bb

Team_pitching_h:

As we can observe from the below graph that the mean is around 1779 but there is long tail for this graph and the skewness is about 10 which is high, thus we would definitely would require to have a look at this variable as well.

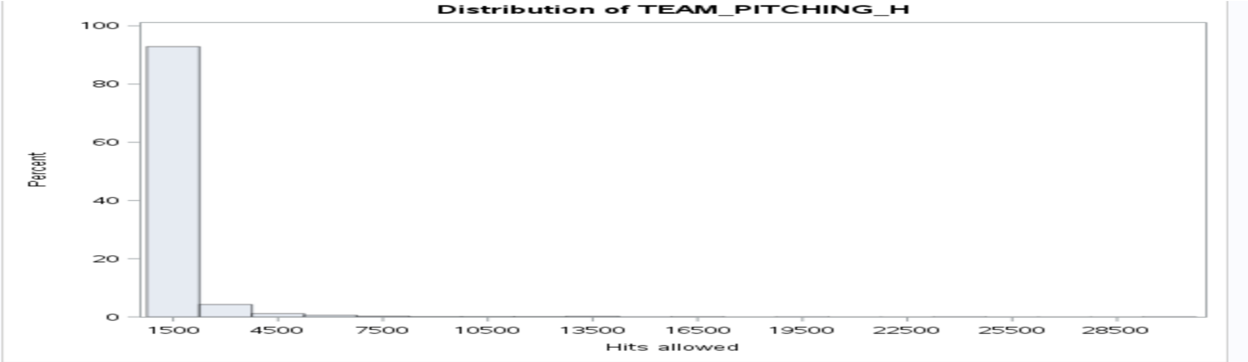


Fig:5 distribution of team pitching_h

by observing all the distributions via a histogram as well as test statistics for normality that includes a series of goodness of fit test based on the empirical distribution function, we found that despite the way the variables appear on the histogram , our good-ness fit test indicated that we should not reject the null hypothesis, which means that our variables are normally distributed, but there are few very extreme values and asymmetrical distributions in few of the variables, these issues can be addressed by various techniques like deleting the extreme values, use bucketing, transformation. Below are the variables with the asymmetric distribution TEAM_BASERUN_SB, TEAM_BATTING_3B, TEAM_BATTING_BB, TEAM_BATTING_H ,TEAM_BATTING_HR, and R_TEAM_PITCHING_HR and few of the extreme values are present in the variables like TEAM_BASERUN_SB, TEAM_FIELDING_E, TEAM_PITCHING_BB, and TEAM_PITCHING_H. In order to proceed further we will decide on which variables have to be transformed based on their usage in the modelling process and the correlation of those variables with the target wins. In our next step we will do the data preparation and fix the missing values and transform the variables if required.

2. Data Preparation:

In the above data exploration section, we identified few of the variables which had extreme values like **team_pitching_h, team_pitching_bb, team_fielding_e and team_baserun_sb** and also there were few variables which have missing values, thus needs to be transformed and imputed respectively, before we start modelling.

Missing values:

On reviewing the missing value chart, we can see that below variables have missing values:
Team_batting_so
Team_baserun_cs
Team_baserun_sb
Team_pitching_so
Team_fielding_dp
Team_batting_hbp

Out of all these variables, we can say that the batting_so, baserun_cs, fielding_dp, team_batting_hbp and pitching_so have a low correlation with the target_wins, only variable(baserun_sb) which has missing values and has a better correlation and positive impact on win, thus we will impute the team_baserun_sb variable with its mean 124.7617716, we will create two new variables in this process, one with IP_* prefix for the imputed variables and one with flag as a indicator variable for the imputed variable, the new variables are **IP_TEAM_BASERUN_SB and I_IP_TEAM_BASERUN_SB.**

Just to be on the safer side and we may use the other variables in our modelling later so we have imputed all the variables which have missing values with their mean.
Below new variables have been created:
IP_TEAM_BASERUN_SB
IP_TEAM_BASERUN_SO
IP_TEAM_BASERUN_CS
IP_TEAM_BATTING_HDP
IP_TEAM_FIELDING_DP

Outliers:

By observing the data, we could see that the below variables had few outliers:
TEAM_FIELDING_E
TEAM_PITCHING_BB
TEAM_PITCHING_H

I have to confess I am not really into baseball, so I had to do some research to understand each of the variables so that I can exactly interpret the values, on observing the data I could see that the for the variable **TEAM_FIELDING_E**
On an average less than 3 would be a good value per game, since we have the data for 162 games I had used the value 486, I marked anything above that an outlier of the fielding error.
Similarly, for **TEAM_PITCHING_BB and TEAM_PITCHING_H** we used 874 and 2041 as a bench mark respectively.

So new three variables were created:
I_TEAM_FIELDING_E
I_TEAM_PITCHING_BB
I_TEAM_PITCHING_H

Log and sqrt transformation:

For the variable which are asymmetric in nature we will use log and sqrt transformation on each of the variable and retain the variables which are symmetric after the transformation.On transforming the variables log transform of TEAM_BASERUN_SB was found to be most symmetrical expression of that variable, log transform of TEAM_BATTING_3B was symmetrical, log transform of TEAM_BATTING_H was symmetrical. We also found that for the variable TEAM_BATTING_HR neither sqrt nor log helped it to be symmetrical.

Below are the graphs of the transformed variables:

Log TEAM_BASERUN_SB:

As we can see after log transformation the variable TEAM_BASERUN_SB is symmetrical and the skewness is just 0.002

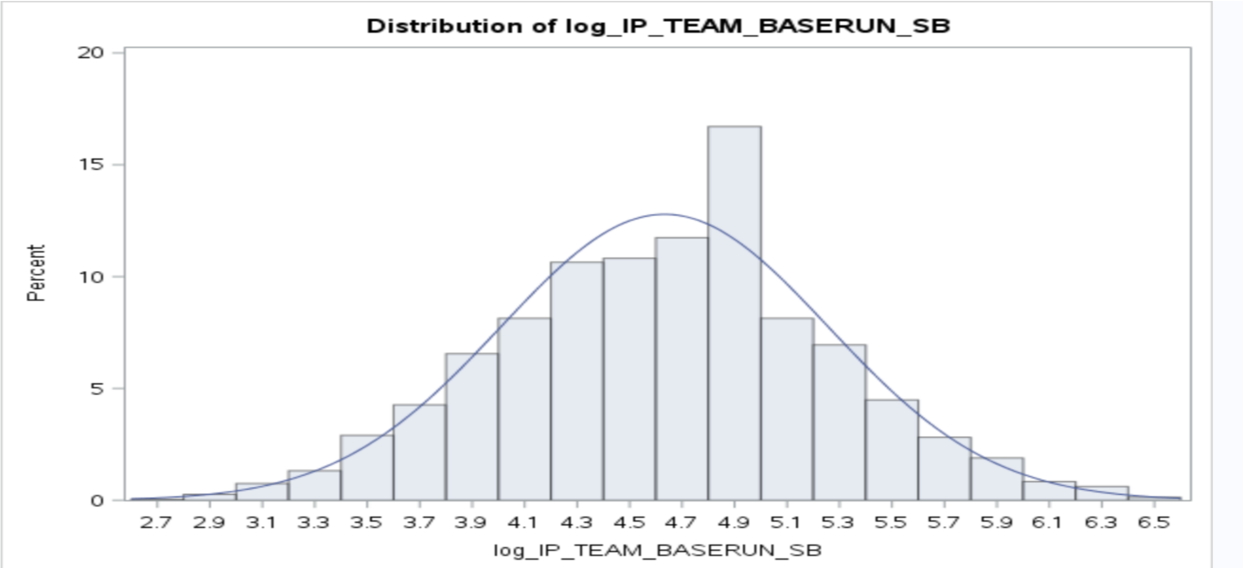


Fig :6 distribution of Team_baserun_sb

Log TEAM_BATTING_3B

As we can see after log transformation the variable TEAM_BASERUN_SB is symmetrical and the skewness is just 0.0019

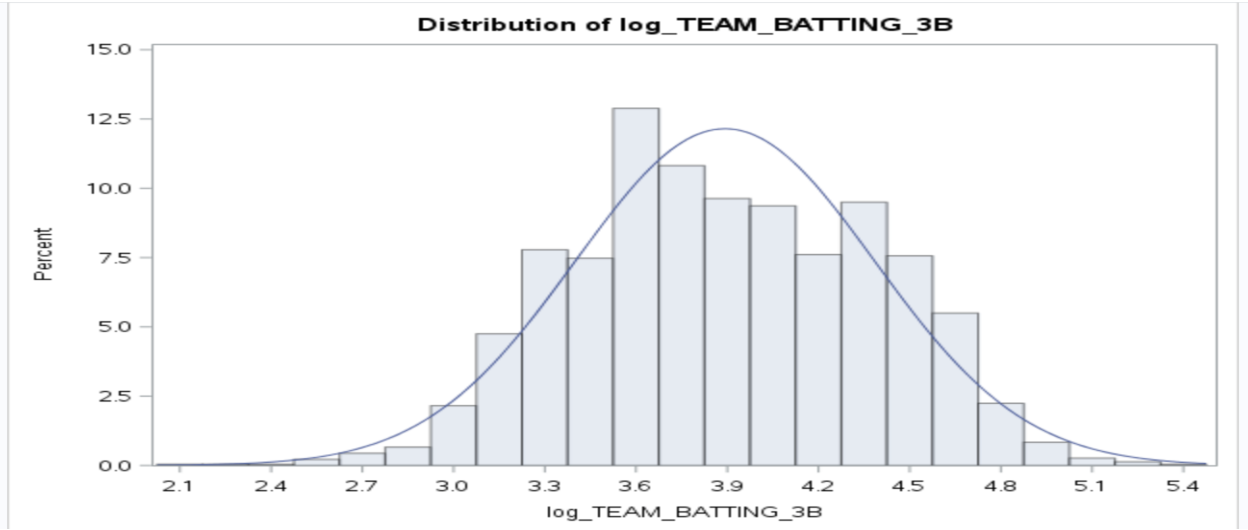


Fig :7 distribution of Team_pitching_bb

Log TEAM_BATTING_HR:

From the below graph we can see that the log transformation of the variable did not helped and there is still skewness in the graph and the graph is not symmetrical.

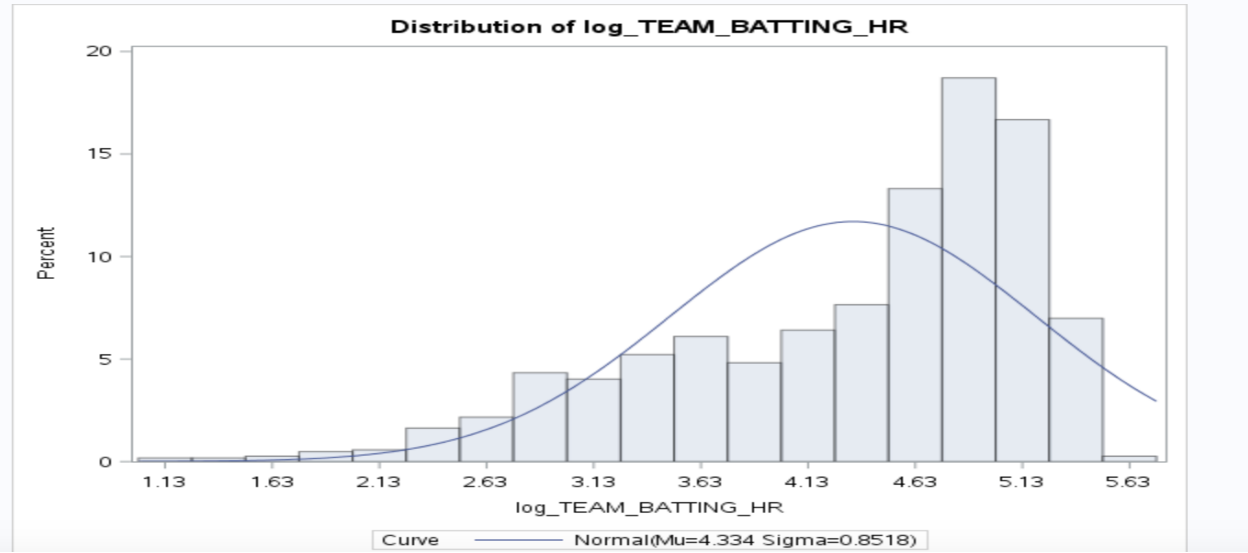


Fig :8 distribution of log_team_batting_hr

3. Build Models:

Now let’s begin with the modelling, in my first model I will be using the variables which have a higher correlation with the target_wins and try to build a model, we are choosing the variables which have correlation greater than 0.1.

Below is the metrics of the model:

Root MSE	13.26149	R-Square	0.2845
Dependent Mean	80.84609	Adj R-Sq	0.281
Coeff Var	16.40338		

In addition to the above metrics, I also analyzed the SAS ODS outputs for this model, the quantile plot looked good, the residual plot was reasonable, we could see only few outliers in the Cook’s D plot and in the residual histogram the SAS curve covered most of the points and but the fit mean graph was smaller than the residual graph.

The model equation is:
Y=B0+B1X+ B2X2+B3X3+B4X4+B5X5+B6X6+B7X7+B8X8+B9X9+B10-
X10- +E

Where Variables and parameter estimates are below:

TEAM_BATTING_H	Base Hits by batters
TEAM_BATTING_2B	Doubles by batters
TEAM_BATTING_3B	Triples by batters
TEAM_BATTING_HR	Homeruns by batters
TEAM_BATTING_BB	Walks by batters
log_IP_TEAM_BASERUN_SB	log of IP_TEAM_BASERUN_SB
TEAM_PITCHING_BB	Walks allowed
TEAM_PITCHING_H	Hits allowed
TEAM_PITCHING_HR	Homeruns allowed
TEAM_FIELDING_E	Errors
IP_TEAM_BASERUN_CS	impute missing TEAM_BASERUN_CS with mean

Intercept	Intercept	1	-12.88245
TEAM_BATTING_H	Base Hits by batters	1	0.05105

TEAM_BATTING_2B	Doubles by batters	1	-0.0286
TEAM_BATTING_BB	Walks by batters	1	-0.00344
TEAM_BATTING_HR	Homeruns by batters	1	0.0536
TEAM_BATTING_3B	Triples by batters	1	0.07594
log_IP_TEAM_BASERUN_SB	log of IP_TEAM_BASERUN_CS	1	4.92129
TEAM_PITCHING_BB	Walks allowed	1	0.00941
TEAM_PITCHING_H	Hits allowed	1	-0.00097081
TEAM_PITCHING_HR	Homeruns allowed	1	-0.01109
TEAM_FIELDING_E	Errors	1	-0.0218
IP_TEAM_BASERUN_CS	impute TEAM_BASERUN_CS with mean	1	-0.03745

Model2:

In my second model I used below variables, I removed the variables which were not statically significant from the model 1, as you can notice that I have removed TEAM_BATTING_HR, TEAM_BASERUN_CS, TEAM_PITCHING_BB this brought down the adj r square value from 0.35 to 0.27.
Also there were few variables statically not significant in this model that is TEAM_BATTING_2B and TEAM_BATTING_BB after removing these variables and running the model again I did not see any significant changes in the adj r square value and other metrics.

TEAM_BATTING_H	Base Hits by batters
TEAM_BATTING_2B	Doubles by batters
TEAM_BATTING_BB	Walks by batters
TEAM_BATTING_3B	Triples by batters
log_IP_TEAM_BASERUN_SB	log of IP_TEAM_BASERUN_CS
TEAM_PITCHING_HR	Homeruns allowed
TEAM_FIELDING_E	Errors

Below is the metrics for this model:

Root MSE	13.29276	R-Square	0.2798
Dependent Mean	80.84609	Adj R-Sq	0.2776
Coeff Var	16.44206		

In addition to the above metrics, I also analyzed the SAS ODS outputs for this model, the quantile plot looked good, the residual plot was reasonable, we could see only few outliers in the Cook's D plot and in the residual histogram the SAS curve covered most of the points and but the fit mean graph was smaller than the residual graph.

Model3:

In both the above models I have selected the variables manually, now for the third model I will be using automated stepwise method, below variables there we got from the process.

TEAM_BATTING_H	Base Hits by batters
TEAM_BATTING_3B	Triples by batters
TEAM_BATTING_BB	Walks by batters
TEAM_BATTING_HR	Homeruns by batters
IP_TEAM_BASERUN_SB	impute TEAM_BASERUN_SB with mean
TEAM_FIELDING_E	Errors
I_TEAM_PITCHING_H	Outlier Indicator for Hits allowed
TEAM_PITCHING_H	Hits allowed
IP_TEAM_FIELDING_DP	indicator of imputation for IP_TEAM_BATTING_DP

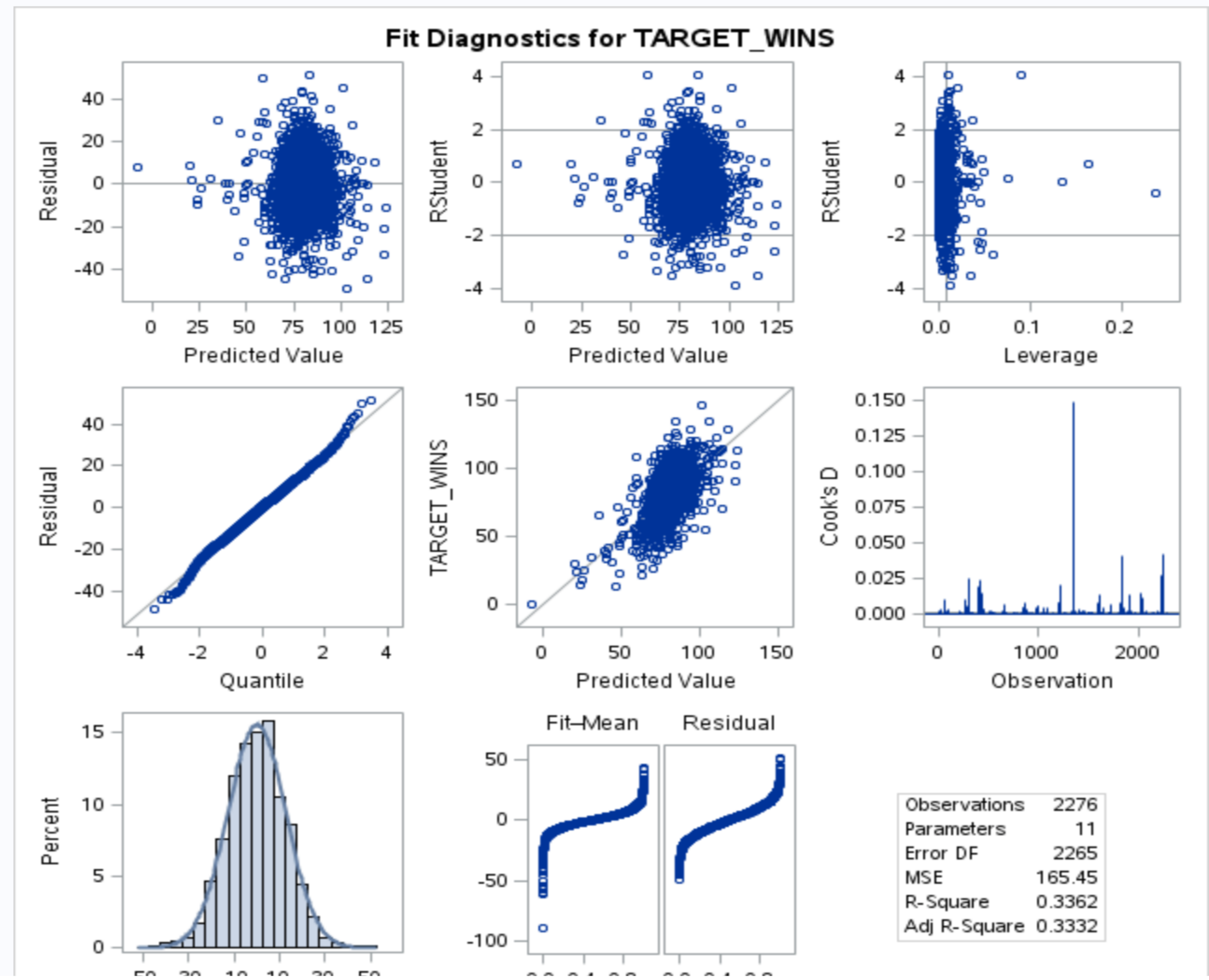
Root MSE	12.86263	R-Square	0.3362
Dependent Mean	80.79086	Adj R-Sq	0.3332
Coeff Var	15.92089		

Below is the table with all the details for this model:

Variable	Label	Parameter			
Estimate	Standard				
Error	t Value				
Intercept	Intercept	24.7485	3.46978	7.13	<.0001
TEAM_BATTING_H	Base Hits by batters	0.04437	0.00246	18.05	<.0001
I_TEAM_FIELDING_E	Outlier Indicator for Errors in all the games	12.25772	1.76368	6.95	<.0001
TEAM_BATTING_3B	Triples by batters	0.0765	0.0161	4.75	<.0001

TEAM_BATTING_BB	Walks by batters	0.01494	0.00327	4.57	<.0001
TEAM_BATTING_HR	Homeruns by batters	0.03156	0.0076	4.15	<.0001
IP_TEAM_BASERUN_SB	impute TEAM_BASERUN_SB with mean	0.02017	0.00413	4.88	<.0001
TEAM_FIELDING_E	Errors	-0.04209	0.00342	-12.3	<.0001
IP_TEAM_FIELDING_DP	indicator of imputation for IP_TEAM_BATTING_SO	-0.1284	0.01291	-9.95	<.0001
TEAM_PITCHING_H	Hits allowed	0.00031154	0.00028925	1.08	0.2816
I_TEAM_PITCHING_H	Outlier Indicator for Hits allowed	6.28084	1.29013	4.87	<.000

Below are the ODS residual outputs of SAS.



Observations	2276
Parameters	11
Error DF	2265
MSE	165.45
R-Square	0.3362
Adj R-Square	0.3332

Above is the metrics for this model as we can see the Adj r square is 0.3332 and after analyzing the SAS ods outputs, the QQ plots looks good, the Cook's D plot has few outliers which are visible in the graph and also the SAS curve covers most of the points in the histogram, also as you can see that the fit mean plot is almost equal to the residual plot, overall the residual out puts are ok.

MODEL COMPARISON

We have three models which we have built, Model 1, Model 2 and Model 3.
For model 1 and 2 we selected the variables manually by using the correlation values with the target_wins and for the third model we used the stepwise method to select the variables.

Below are the r square values and adj r squares values for the three models:

	Model1	Model2	Model3
R-square	0.2845	0.2798	0.3362
Adj R-Sq	0.281	0.2776	0.3332

By considering the Adj r square values and the by analyzing the SAS ods outputs by comparing the QQ plots, cook's D, residual plots and the fit-mean we decided to pick model 3 for predicting the wins. Between all of the models constructed, few of them are not documented in this report, we have to take into consideration that the expected predictive ability on this data set is likely going to be low. I tried to combine dummy variables into a single dummy variable and tried to model the data, but incorporating this variable gave me a poor goodness of fit diagnostic.

The equation of the model selected is below:

P_TARGET_WINS = 24.74850
+ 0.04437 * TEAM_BATTING_H
+ 12.25772 * I_TEAM_FIELDING_E
+ 0.07650 * TEAM_BATTING_3B
+ 0.01494 * TEAM_BATTING_BB
+ 0.03156 * TEAM_BATTING_HR
+ 0.02017 * IP_TEAM_BASERUN_SB

- 0.04209 * TEAM_FIELDING_E
- 0.12840 * IP_TEAM_FIELDING_DP
+ TEAM_PITCHING_H * 0.00031154
+ I_TEAM_PITCHING_H * 6.28084

In this model we can see that for the team batting we get positive coff and for the team fielding errors we can see that we have a negative coefficient, but as you notice that the team fielding double plays has a positive impact on the wins but the coefficient is negative, but I have still included it in my model since it had an impact on the adj r square value and I believe it definitely added value to the model.

MODEL DEPLOYMENT CODE:

```
libname mydata "/sscc/home/n/ngg135/assigment1/" access=readonly;

proc datasets library=mydata;
run;
 quit;



data testing;
set mydata.moneyball_test;
proc contents data=testing;
run;

data testing_imp;
    set testing;
    IP_TEAM_BASERUN_SB = TEAM_BASERUN_SB;
    I_IP_TEAM_BASERUN_SB = 0;
    IP_TEAM_BASERUN_CS = TEAM_BASERUN_CS;
    I_IP_TEAM_BASERUN_CS = 0;
    IP_TEAM_BATTING_HBP = TEAM_BATTING_HBP;
    I_IP_TEAM_BATTING_HBP = 0;
    IP_TEAM_FIELDING_DP = TEAM_FIELDING_DP;
    I_IP_TEAM_FIELDING_DP = 0;
    label IP_TEAM_BASERUN_SB = 'impute TEAM_BASERUN_SB with
mean';
    label I_IP_TEAM_BASERUN_CS = 'indicator of imputation for
IP_TEAM_BASERUN_SB';
    label IP_TEAM_BASERUN_CS = 'impute TEAM_BASERUN_CS with
mean';
    label I_IP_TEAM_BASERUN_SB = 'indicator of imputation for
IP_TEAM_BASERUN_CS';
    label IP_TEAM_BATTING_HBP = 'impute TEAM_BATTING_HBP with
mean';
    label IP_TEAM_BATTING_HBP = 'indicator of imputation for
IP_TEAM_BATTING_HBP';
    label IP_TEAM_FIELDING_DP = 'impute TEAM_BATTING_SO with
mean';
    label IP_TEAM_FIELDING_DP = 'indicator of imputation for
IP_TEAM_BATTING_SO';
    if missing(IP_TEAM_BASERUN_SB) then do;
        IP_TEAM_BASERUN_SB = 124.761772;
        I_IP_TEAM_BASERUN_SB = 1;
    end;
    if missing (IP_TEAM_BASERUN_CS) then do;
    IP_TEAM_BASERUN_CS=52.803;
    I_IP_TEAM_BASERUN_CS=1;
    END;
     if missing (IP_TEAM_BATTING_HBP) then do;
    IP_TEAM_BATTING_HBP=59.3560209;
    I_IP_TEAM_BATTING_HBP=1;
    END;
      if missing (IP_TEAM_FIELDING_DP) then do;
    IP_TEAM_FIELDING_DP=146.387;
    I_IP_TEAM_FIELDING_DP=1;
    END;

data testing_score;
    set testing_imp;
```

```
   P_TARGET_WINS = 17.26345 + 0.04550 * TEAM_BATTING_H +
0.07783 * TEAM_BATTING_3B + 0.01174 * TEAM_BATTING_BB  +
0.04829  * TEAM_BATTING_HR + 0.02681 * IP_TEAM_BASERUN_SB -
0.01986 * TEAM_FIELDING_E  - 0.11461 * IP_TEAM_FIELDING_DP ;
   keep index P_TARGET_WINS;
```

SCORED DATA FILE:

Scored data file is attached with the name predictions_final sas7bdat. This file will have two columns one is the index and other is p_target_wins.

Conclusion:

We developed several models for this project using the data of the professional baseball team from the years 1871 to 2006, we chose the variables manually in few of the models based on its correlation value with the target wins and the model which we chose for prediction was from selection method. But overall this was a good project, where we were able to build a model and deploy the model and also present the model so that others can use. I think the models predictive performance is still low, this may be because of the multiple population within this data, likely due to how long of a period the data was being collected over.

SAS CODE:

```
libname mydata "/sscc/home/n/ngg135/assigment1/" access=readonly;

proc datasets library=mydata;
run;
 quit;

data training;
set mydata.moneyball;
proc contents data=training;
run;
```

```
proc print data=training ;
run;


proc contents data=training;



*///Exploratory data analysis///;


proc corr data=traning;
with target_wins;
run;

proc means data=training ;
run;

proc means data=training NMISS N;
run;

ods graphics on;
proc corr data training plot matrix;
with TARGET_WINS;
run;
ods graphics off;

proc univariate data=training;
histogram TEAM_BASERUN_CS  /normal;
run;

proc univariate data=training;
histogram TEAM_BATTING_BB /normal;
run;


proc univariate data=training normal;
   var TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B
TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB
```

```
TEAM_BASERUN_SB TEAM_FIELDING_E TEAM_PITCHING_BB
TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_BASERUN_SB;
    histogram;

*///Imputing missing values///;
data training_imp;
   set training;
   IP_TEAM_BASERUN_SB =  TEAM_BASERUN_SB;
   I_IP_TEAM_BASERUN_SB = 0;
   IP_TEAM_BASERUN_CS =  TEAM_BASERUN_CS;
   I_IP_TEAM_BASERUN_CS = 0;
   IP_TEAM_BATTING_HBP  =  TEAM_BATTING_HBP ;
   I_IP_TEAM_BATTING_HBP = 0;
    IP_TEAM_FIELDING_DP  = TEAM_FIELDING_DP ;
   I_IP_TEAM_FIELDING_DP = 0;
   label IP_TEAM_BASERUN_SB = 'impute TEAM_BASERUN_SB with mean';
   label I_IP_TEAM_BASERUN_CS = 'indicator of imputation for
IP_TEAM_BASERUN_SB';
   label IP_TEAM_BASERUN_CS = 'impute TEAM_BASERUN_CS with mean';
   label I_IP_TEAM_BASERUN_SB = 'indicator of imputation for
IP_TEAM_BASERUN_CS';
   label IP_TEAM_BATTING_HBP = 'impute TEAM_BATTING_HBP with
mean';
   label IP_TEAM_BATTING_HBP = 'indicator of imputation for
IP_TEAM_BATTING_HBP';
      label IP_TEAM_FIELDING_DP = 'impute TEAM_BATTING_SO with
mean';
   label IP_TEAM_FIELDING_DP = 'indicator of imputation for
IP_TEAM_BATTING_SO';
   if missing(IP_TEAM_BASERUN_SB) then do;
      IP_TEAM_BASERUN_SB = 124.761772;
      I_IP_TEAM_BASERUN_SB = 1;
   end;
   if missing (IP_TEAM_BASERUN_CS) then do;
   IP_TEAM_BASERUN_CS=52.803;
   I_IP_TEAM_BASERUN_CS=1;
   END;
    if missing (IP_TEAM_BATTING_HBP) then do;
   IP_TEAM_BATTING_HBP=59.3560209;
   I_IP_TEAM_BATTING_HBP=1;
   END;
     if missing (IP_TEAM_FIELDING_DP) then do;
   IP_TEAM_FIELDING_DP=146.387;
   I_IP_TEAM_FIELDING_DP=1;
   END;

 *///Outliers indicators///;
 data training_imp_o;
   set training_imp;
   if TEAM_FIELDING_E < 486 then I_TEAM_FIELDING_E = 0.0;
   else I_TEAM_FIELDING_E = 1;
   label I_TEAM_FIELDING_E = 'Outlier Indicator for Errors in all the games';

   if TEAM_PITCHING_BB < 874 then I_TEAM_PITCHING_BB = 0.0;
   else I_TEAM_PITCHING_BB = 1.0;
   label I_TEAM_PITCHING_BB = 'Outlier Indicator for Walks allowed';

   if TEAM_PITCHING_H < 2041 then I_TEAM_PITCHING_H = 0.0;
   else I_TEAM_PITCHING_H = 1.0;
   label I_TEAM_PITCHING_H = 'Outlier Indicator for Hits allowed';


 *///variable transformation///;

data training_imp_transform;
 set training_imp_o;

sqrt_IP_TEAM_BASERUN_SB = sqrt(IP_TEAM_BASERUN_SB);
log_IP_TEAM_BASERUN_SB = log(IP_TEAM_BASERUN_SB);
label log_IP_TEAM_BASERUN_SB= 'log of IP_TEAM_BASERUN_CS';

sqrt_TEAM_BATTING_3B = sqrt(TEAM_BATTING_3B);
log_TEAM_BATTING_3B = log(TEAM_BATTING_3B);

sqrt_TEAM_BATTING_BB = sqrt(TEAM_BATTING_BB);
log_TEAM_BATTING_BB = log(TEAM_BATTING_BB);

sqrt_TEAM_BATTING_H = sqrt(TEAM_BATTING_H);
log_TEAM_BATTING_H = log(TEAM_BATTING_H);
```

```
sqrt_TEAM_BATTING_HR = sqrt(TEAM_BATTING_HR);
log_TEAM_BATTING_HR = log(TEAM_BATTING_HR);

sqrt_TEAM_PITCHING_HR = sqrt(TEAM_PITCHING_HR);
log_TEAM_PITCHING_HR = log(TEAM_PITCHING_HR);

sqrt_IP_TEAM_BASERUN_CS = sqrt(IP_TEAM_BASERUN_CS);
log_IP_TEAM_BASERUN_CS = log(IP_TEAM_BASERUN_CS);



proc univariate data=training_imp_transform;
histogram log_IP_TEAM_BASERUN_SB /normal;
run;

proc univariate data=training_imp_transform;
histogram log_TEAM_BATTING_3B /normal;
run;

proc univariate data=training_imp_transform;
histogram sqrt_TEAM_BATTING_HR /normal;
run;


proc print data=training_imp (obs=10);
run;

*///Regression modelling///;
proc reg data=training_imp_transform;
  model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B
TEAM_BATTING_BB TEAM_BATTING_HR
   TEAM_BATTING_3B   log_IP_TEAM_BASERUN_SB
TEAM_PITCHING_BB TEAM_PITCHING_H TEAM_PITCHING_HR
TEAM_FIELDING_E IP_TEAM_BASERUN_CS ;



proc reg data=training_imp_transform;
  model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B
TEAM_BATTING_3B TEAM_BATTING_BB
```

```
   log_IP_TEAM_BASERUN_SB TEAM_PITCHING_HR TEAM_FIELDING_E
;



proc reg data=training_imp_transform;
   model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B
TEAM_BATTING_3B  TEAM_BATTING_BB TEAM_BATTING_HR
      IP_TEAM_BASERUN_SB IP_TEAM_BASERUN_CS TEAM_FIELDING_E
IP_TEAM_FIELDING_DP  TEAM_PITCHING_BB TEAM_PITCHING_H
TEAM_PITCHING_HR TEAM_PITCHING_SO  /
  selection=adjrsq aic bic cp best=5;


proc reg data=training_imp_transform;
   model TARGET_WINS = TEAM_BATTING_H  I_TEAM_FIELDING_E
TEAM_BATTING_3B TEAM_BATTING_BB TEAM_BATTING_HR
IP_TEAM_BASERUN_SB TEAM_FIELDING_E IP_TEAM_FIELDING_DP
TEAM_PITCHING_H I_TEAM_PITCHING_H;



*///Testing data///;

libname mydata "/sscc/home/n/ngg135/assigment1/" access=readonly;

proc datasets library=mydata;
run;
 quit;

data testing;
set mydata.moneyball_test;
proc contents data=testing;
run;

*///Handling missing values///;

data testing_imp;
    set testing;
    IP_TEAM_BASERUN_SB =  TEAM_BASERUN_SB;
    I_IP_TEAM_BASERUN_SB = 0;
    IP_TEAM_BASERUN_CS =  TEAM_BASERUN_CS;
    I_IP_TEAM_BASERUN_CS = 0;
    IP_TEAM_BATTING_HBP  =  TEAM_BATTING_HBP ;
    I_IP_TEAM_BATTING_HBP = 0;
    IP_TEAM_FIELDING_DP  = TEAM_FIELDING_DP ;
    I_IP_TEAM_FIELDING_DP = 0;
    label IP_TEAM_BASERUN_SB = 'impute TEAM_BASERUN_SB with mean';
    label I_IP_TEAM_BASERUN_CS = 'indicator of imputation for
IP_TEAM_BASERUN_SB';
    label IP_TEAM_BASERUN_CS = 'impute TEAM_BASERUN_CS with mean';
    label I_IP_TEAM_BASERUN_SB = 'indicator of imputation for
IP_TEAM_BASERUN_CS';
    label IP_TEAM_BATTING_HBP = 'impute TEAM_BATTING_HBP with
mean';
    label IP_TEAM_BATTING_HBP = 'indicator of imputation for
IP_TEAM_BATTING_HBP';
    label IP_TEAM_FIELDING_DP = 'impute TEAM_BATTING_SO with mean';
    label IP_TEAM_FIELDING_DP = 'indicator of imputation for
IP_TEAM_BATTING_SO';
    if missing(IP_TEAM_BASERUN_SB) then do;
        IP_TEAM_BASERUN_SB = 124.761772;
        I_IP_TEAM_BASERUN_SB = 1;
    end;
    if missing (IP_TEAM_BASERUN_CS) then do;
    IP_TEAM_BASERUN_CS=52.803;
    I_IP_TEAM_BASERUN_CS=1;
    END;
     if missing (IP_TEAM_BATTING_HBP) then do;
    IP_TEAM_BATTING_HBP=59.3560209;
    I_IP_TEAM_BATTING_HBP=1;
    END;
      if missing (IP_TEAM_FIELDING_DP) then do;
    IP_TEAM_FIELDING_DP=146.387;
    I_IP_TEAM_FIELDING_DP=1;
    END;

    data testing_imp_o;
    set testing_imp;
    if TEAM_FIELDING_E < 486 then I_TEAM_FIELDING_E = 0.0;
    else I_TEAM_FIELDING_E = 1;
    label I_TEAM_FIELDING_E = 'Outlier Indicator for Errors in all the games';

    if TEAM_PITCHING_BB < 874 then I_TEAM_PITCHING_BB = 0.0;
    else I_TEAM_PITCHING_BB = 1.0;
    label I_TEAM_PITCHING_BB = 'Outlier Indicator for Walks allowed';

    if TEAM_PITCHING_H < 2041 then I_TEAM_PITCHING_H = 0.0;
    else I_TEAM_PITCHING_H = 1.0;
    label I_TEAM_PITCHING_H = 'Outlier Indicator for Hits allowed';


   proc print data=testing_imp (obs=100);

*///prediction///;

 data predictions;
    set testing_imp_o;
    P_TARGET_WINS = 17.26345 + 0.04550 * TEAM_BATTING_H + 0.07783 *
TEAM_BATTING_3B + 0.01174 * TEAM_BATTING_BB  + 0.04829  *
TEAM_BATTING_HR + 0.02681 * IP_TEAM_BASERUN_SB - 0.01986 *
TEAM_FIELDING_E  - 0.11461 * IP_TEAM_FIELDING_DP ;
    keep index P_TARGET_WINS;



proc print data=predictions_final;
run;

     data mydata.predictions_final;
   set predictions;
   if p_target_wins = '.' then p_target_wins=81;
   P_TARGET_WINS = round(P_TARGET_WINS,1);
   P_TARGET_WINS = min( P_TARGET_WINS, 162 );
   P_TARGET_WINS = max( P_TARGET_WINS, 0 );
   keep index p_target_wins;
   run;
```