Assignment01

Group 5

Data Exploration and Preparation

As the quality of our inputs decide the quality of your output, we will be spending more time and efforts in data exploration, cleaning and preparation. We will be following the below steps for our data exploration and preparation:

1- Variable Identification
2- Univariate Analysis
3- Bi-variate Analysis
4- Missing values treatment
5- Outlier treatment
6- Variable transformation
7- Variable creation

1- Variable Identification

First let's display and examine the data dictionary or the data columns.

VARIABLE_NAME	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Target
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

We notice that all variables are numeric. The variable names seem to follow certain naming pattern to highlight certain arithmetic relationships. In other words, we can compute the number of '1B' hits by taking the difference between overall hits and '2B', '3B', 'HR'. Although such naming and construct is not recommended in normalized database design (as it violates third normal form), it is very frequent practice in the data analytics.

Then , we will identify Predictor (Input) and Target (output) variables. Next, we will identify the data type and category of the variables.

Our predictor input is made of 15 variables. And our dependent variable is one variable called TAR-

GET_WINS.

Below are the variable that have been identified and their respective type and category:

Type of variable	Data Type	Variable Category
Dependent		
TARGET_WINS	numeric	continuous
Independent		
TEAM_BATTING_H	numeric	continuous
TEAM_BATTING_2B	numeric	continuous
TEAM_BATTING_3B	numeric	continuous
TEAM_BATTING_HR	numeric	continuous
TEAM_BATTING_BB	numeric	continuous
TEAM_BATTING_HBP	numeric	continuous
TEAM_BATTING_SO	numeric	continuous
TEAM_BASERUN_SB	numeric	continuous
TEAM_BASERUN_CS	numeric	continuous
TEAM_FIELDING_E	numeric	continuous
TEAM_FIELDING_DP	numeric	continuous
TEAM_PITCHING_BB	numeric	continuous
TEAM_PITCHING_H	numeric	continuous
TEAM_PITCHING_HR	numeric	continuous
TEAM_PITCHING_SO	numeric	continuous

2- Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. in our case all variables are coutinous. Hence we need to understand the central tendency and spread of each variable. These are measured using various statistical metrics visualization methods:

```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0
```

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	

Figure 1: Alt text

```
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0
##                                   NA's   :102      NA's   :131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0
##  1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0
##  Median : 49.0   Median :58.00    Median : 1518   Median :107.0
##  Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7
##  3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0
##  NA's   :772     NA's   :2085
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##                   NA's   :102                        NA's   :286
```

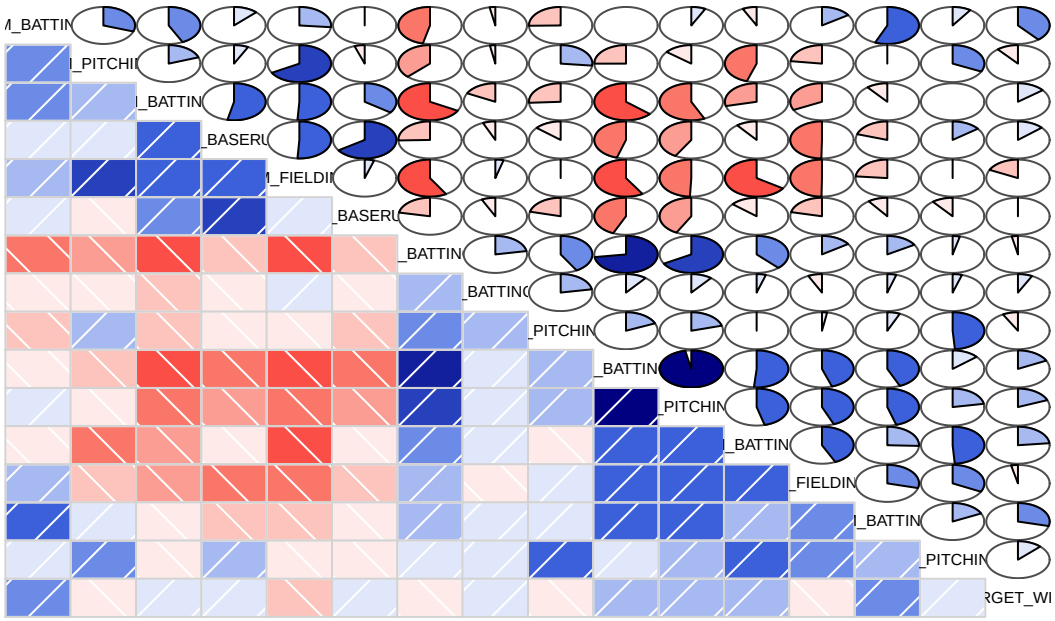
3- Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level.

In our case we have only continuous variables we will be doing bi-variate analysis between two continuous variables. We will use scatter plot and find out the relationship between two variables: We are looking to find the pattern and if the relationship between the variables is linear or non-linear.

Also we will use the scatter plot to show the strength of the relationship between two variable. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

-1: perfect negative linear correlation
+1:perfect positive linear correlation and
0: No correlation



Correlation of our dependable variable TARGET_WINS relative to the other 15 independable variables:

4- Missing values treatment

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

First let identify the missing data and find the mean for each variable by excluding the missing the data.

Variable	Count.Missing.Values	Mean	Correlation	Theoretical.Impact.
TARGET_WINS	0	80.79086	1.0000000	Target
TEAM_BATTING_H	0	1469.26977	0.3887675	Positive Impact on Wins
TEAM_BATTING_2B	0	241.24692	0.2891036	Positive Impact on Wins

Variable	Count.Missing.Values	Mean	Correlation	Theoretical.Impact.
TEAM_BATTING_3B	0	55.25000	0.1426084	Positive Impact on Wins
TEAM_BATTING_HR	0	99.61204	0.1761532	Positive Impact on Wins
TEAM_BATTING_BB	0	501.55888	0.2325599	Positive Impact on Wins
TEAM_BATTING_HBP	2085	59.35602	0.0735042	Positive Impact on Wins
TEAM_BATTING_SO	102	735.60534	-0.0317507	Negative Impact on Wins
TEAM_BASERUN_SB	131	124.76177	0.1351389	Positive Impact on Wins
TEAM_BASERUN_CS	772	52.80386	0.0224041	Negative Impact on Wins
TEAM_FIELDING_E	0	246.48067	-0.1764848	Negative Impact on Wins
TEAM_FIELDING_DP	286	146.38794	-0.0348506	Positive Impact on Wins
TEAM_PITCHING_BB	0	553.00791	0.1241745	Negative Impact on Wins
TEAM_PITCHING_H	0	1779.21046	-0.1099371	Negative Impact on Wins
TEAM_PITCHING_HR	0	105.69859	0.1890137	Negative Impact on Wins
TEAM_PITCHING_SO	102	817.73045	-0.0784361	Positive Impact on Wins

Now that we have identified the count of missing for each variable and the correlation of each variable to our dependent variable TARGET_WINS, we need to decide how to handle the missing the data and which variables to keep based on their correlation.
We observe that the sign of the correlation roughly matches up with our initial proposed theoretical effect; with BATTING_SO, FIELDING_E, and PITCHING_H indicating a negative correlation. However, there are some instances where the sign conflicts with the proposed theoretical effect for instance PITCHING_HR, and PITCHING_BB have the opposite signs.
Based on the correlations and missing values, we infer that we abandon imputing values for BATTING_HBP, BATTING_SO, BASERUN_CS, FIELDING_DP due to low correlation to TARGET_WINS.
However, we are considering the need to impute BASERUN_SB as it has a correlation of 0.1351389 which is relatively acceptable to other correlations. There are few methods to treat missing data such as:

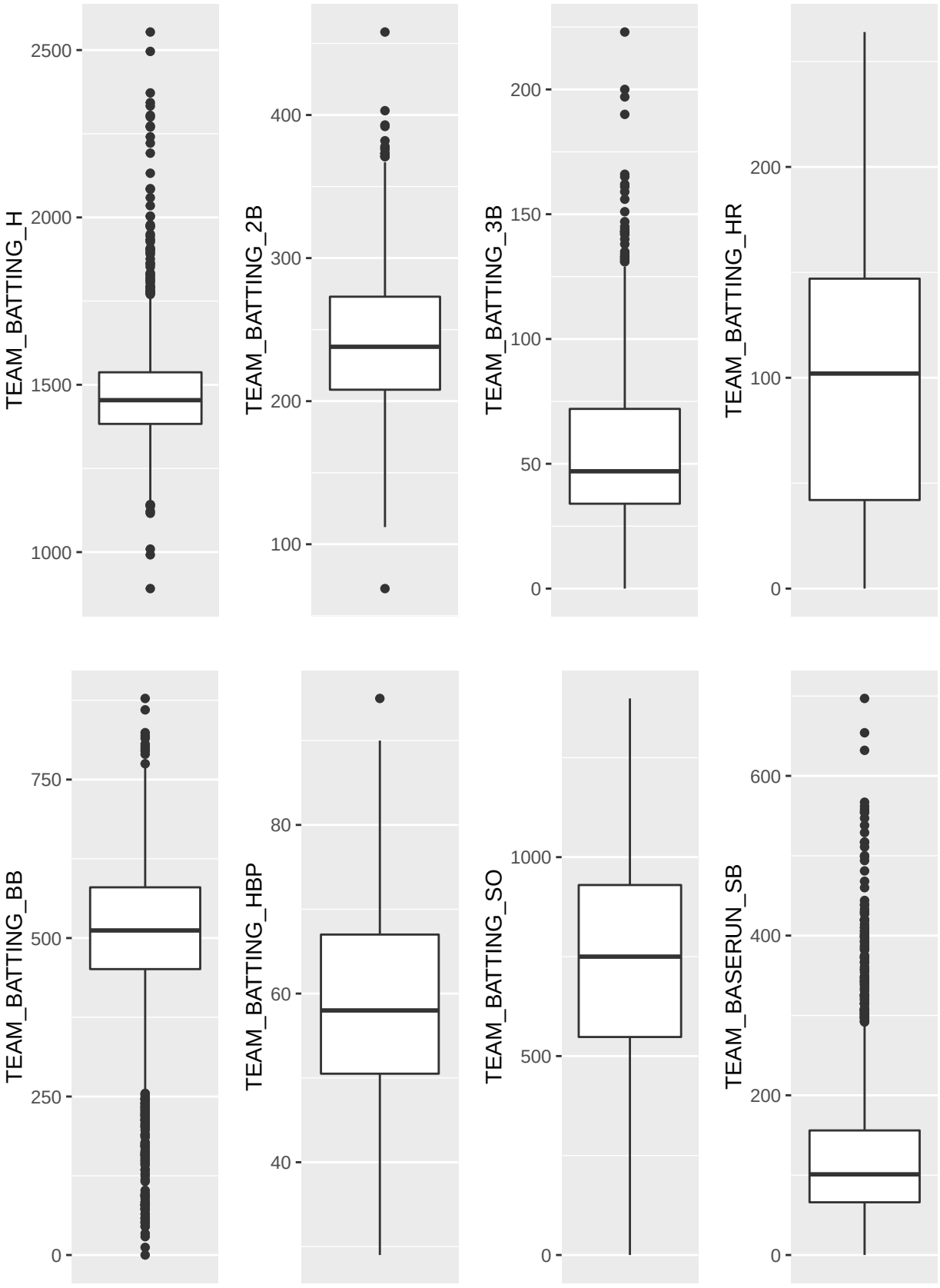
1- Deletion: Either list wise deletion or pair wise deletion, the deletion method is the simplest method.

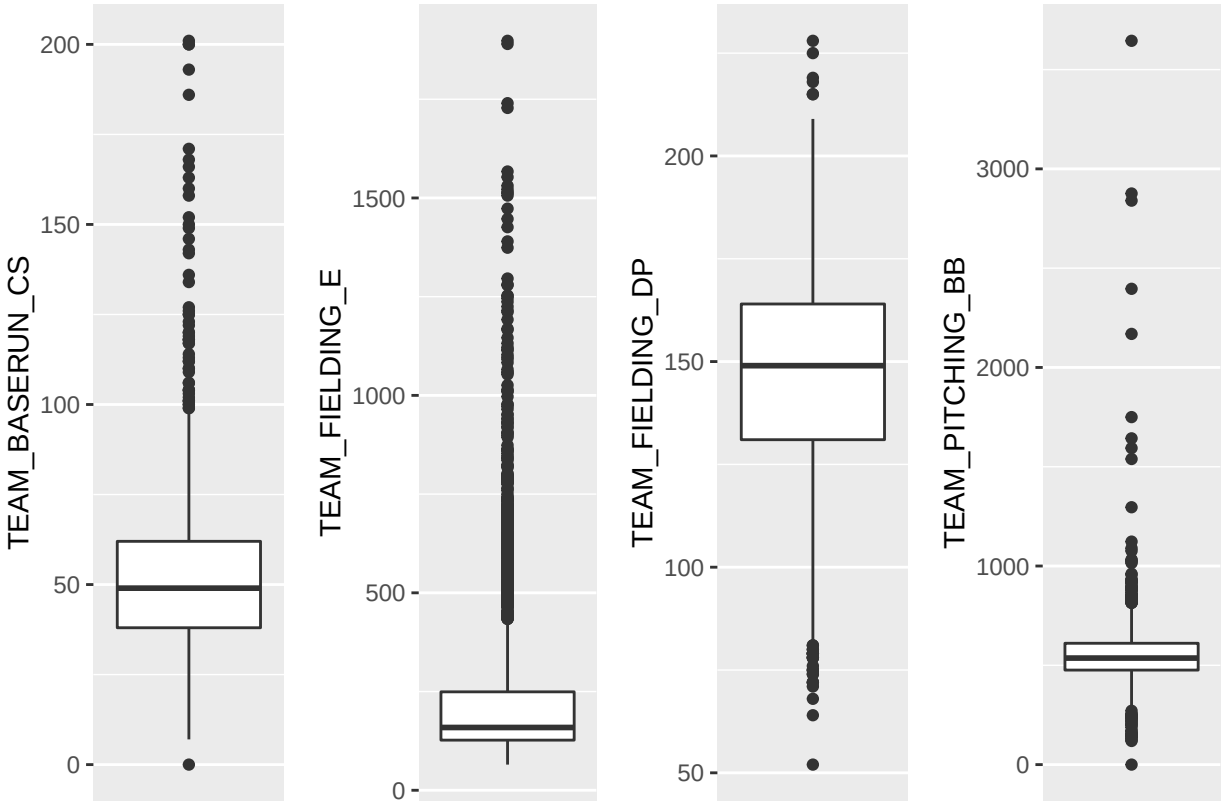
2- Mean/ Mode/ Median Imputation: Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods

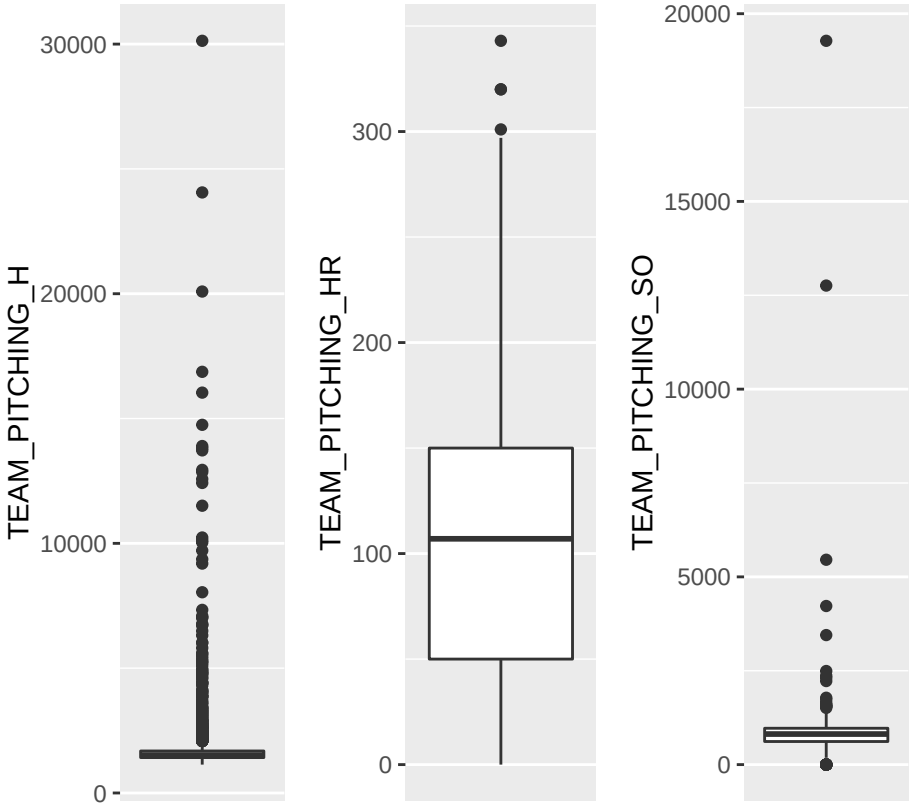
3- Prediction Model: This is one of the sophisticated methods for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data by dividing our data set into two sets. One set with no missing values for the variable and another one with missing values.

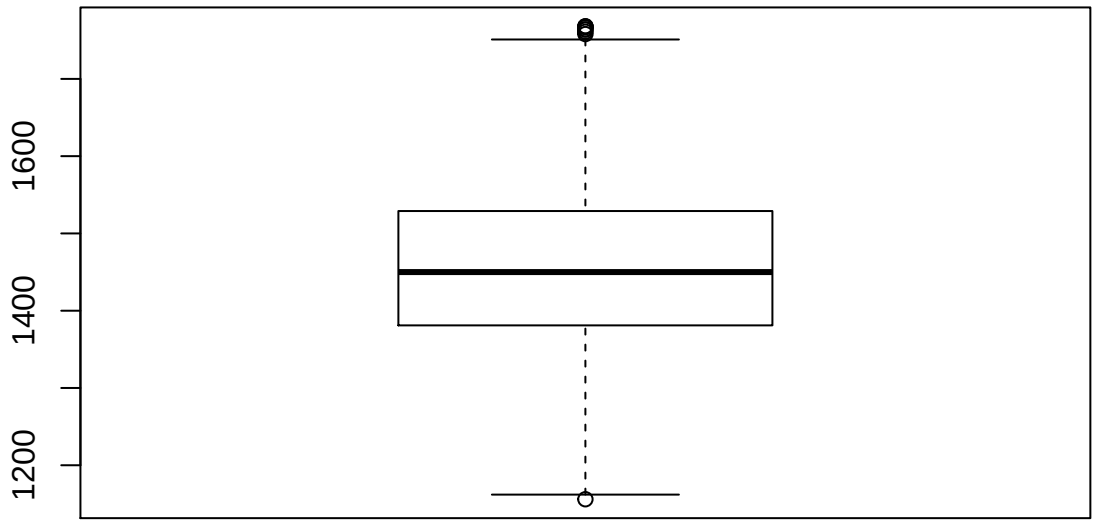
4- KNN Imputation: In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function

Appendix A


TEAM_BATTING_H
2500
2000
1500
1000
TEAM_BATTING_2B
400
300
200
100
TEAM_BATTING_3B
200
150
100
50
0
TEAM_BATTING_HR
200
100
0
TEAM_BATTING_BB
750
500
250
0
TEAM_BATTING_HBP
80
60
40
TEAM_BATTING_SO
1000
500
0
TEAM_BASERUN_SB
600
400
200
0


TEAM_BASERUN_CS
200
150
100
50
0
TEAM_FIELDING_E
1500
1000
500
0
TEAM_FIELDING_DP
200
150
100
50
TEAM_PITCHING_BB
3000
2000
1000
0


TEAM_PITCHING_H
30000
20000
10000
0
TEAM_PITCHING_HR
300
200
100
0
TEAM_PITCHING_SO
20000
15000
10000
5000
0




```
## [1] 0.3887675
```

```
## [1] 0.3502207
```