CS376b-Lab4

Siyu Chen

Data Quality Report

**Overview**

This data set includes match results and statistics of every match of Bundesliga (German soccer league) season 2015 to 2016. The first two columns show the names of the two teams played on that match, and the other features are written simplified:

FTHG = Full Time Home Team Goals

FTAG = Full Time Away Team Goals

FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

HC = Home Team Corners

AC = Away Team Corners

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

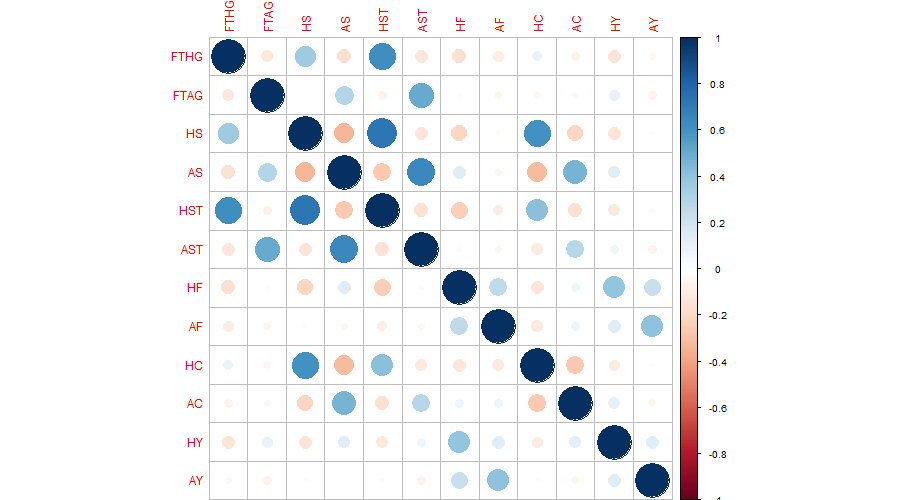
HR = Home Team Red Cards

AR = Away Team Red Cards

By analyzing this data set based on the features listed above, although we will not be able to predict a game’s result before it happens, we can “predict” a match’s winning team (home team/ away team/ draw) based on available match statistics.

**Summary**

Correlation plots



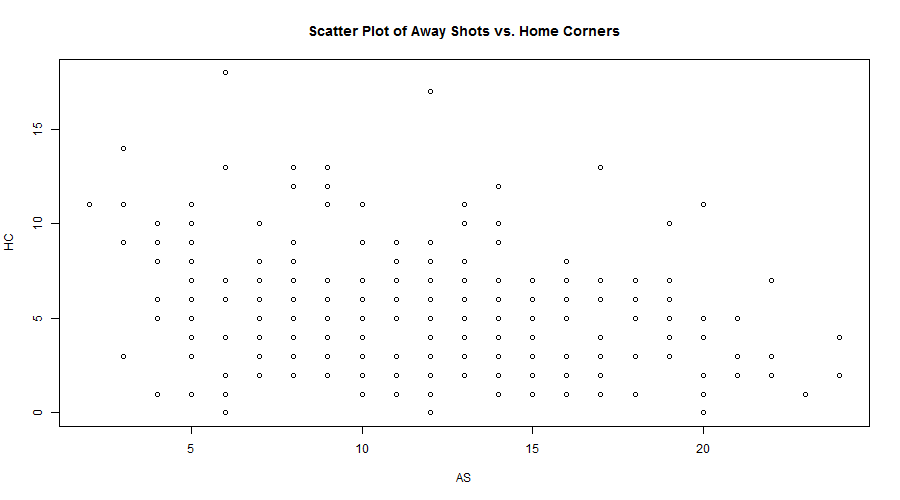
As we can observe from the plot, HS with HST, and AS with AST have strong correlation; however, we should not really consider their correlation as meaningful, because shots and shots on target are not independent from each other. FTHG with HST, and FTAG with AST are also highly correlated, but the total full time goals are clearly depended on total shots on target.

The other obviously positive correlated ones are, HS with HC, HST with HC, AS with AC, HF with HY, AF with AY. These pairs are all somehow dependent to each other. For example, a team which shots(HS/AS) more is likely to gain more corners(HC/AC), and if it commits more fouls(HF/AF), then it is likely to get more yellow cards(HY/AY).

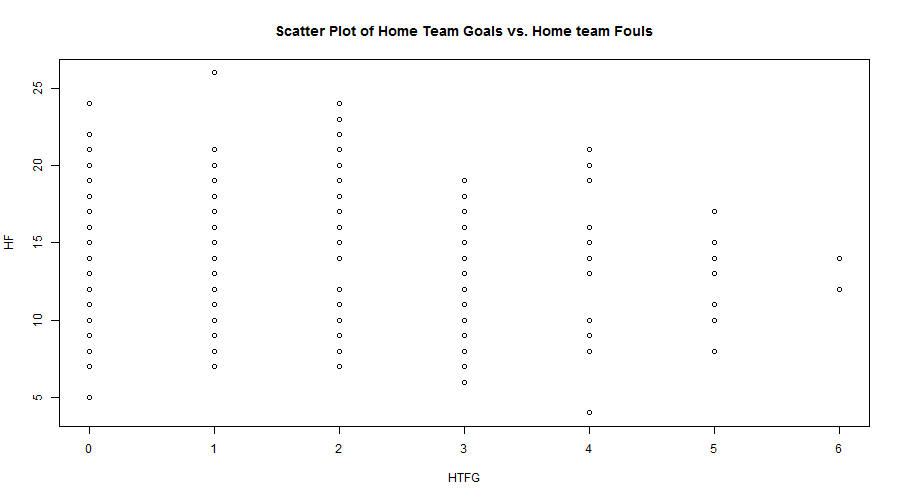
There are also pairs that are negatively correlated, such as HS with AS, which can be explained by that when home team attacks more, the away team has lesser chances to shot. There is another pair that are independent from each other, which is AS with HC, and it may worth us to pay extra attention to.

Scatter plot matrices

Due to the overly huge data size and too many features, R failed to create a scatter plot matrices, therefore I only created scatter plot for a few certain pairs, including the AS with HC we found through analyzing the correlation plots.



Though we cannot clearly tell the relationship of these two features through this plot, we can see a minor trend of total shots decreasing when the total corners are above 10.



Another interesting negative correlation I found is between full time home team goals with home team fouls committed, and the relationship is clear on the plot. It is an interesting finding because usually when a team commits more fouls, it means that the team is attacking harshly and oppressing the other team, but the result turns out to be that the attacking team scores less goals.

overall data quality

Many features seem dependent to each other, but there are also some not related features surprisingly are correlated and worth further analysis. Also, full time home/away team goals, which is the most important feature in terms of its direct influence to the game’s result, are not significantly but visibly correlated with features other than the shots features, such as corners, fouls committed and cards gained. This correlation could help us to find the ways to score more goals.

To improve the data set quality, we may need to edit some features to get more meaningful numbers, for instance, it should be helpful if we calculate the shots on target rate by divide the column of home/away team shots by total shots on target. We can also simplify the data set by changing the game result of Home team win or Away team win or Draw to home team win or lose.

**Details**

**FTHG (Full Time Home Team Goals)**

Type: Continuous numerical

Min: 0

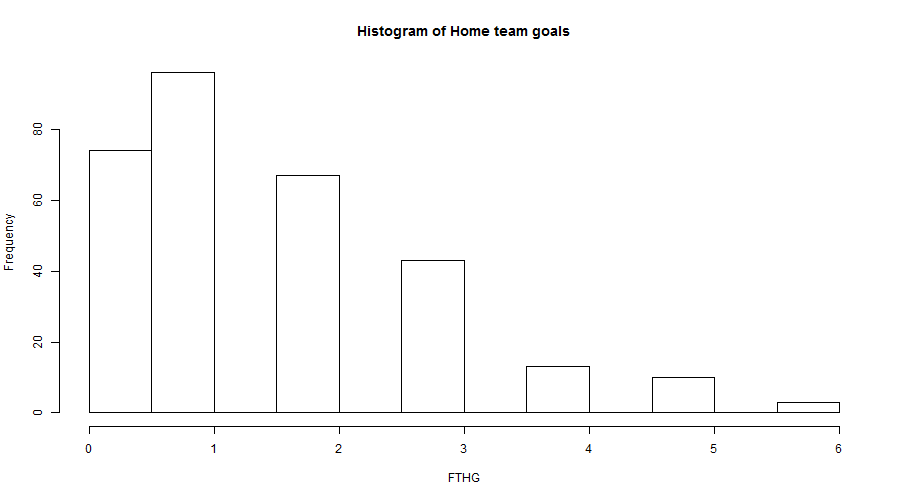
1ST Qu: 1

Median: 1

Mean: 1.565

3rd Qu: 2

Max: 6



The histogram is right-skewed.

**FTAG (Full Time Away Team Goals)**

Type: Continuous numerical

Min: 0

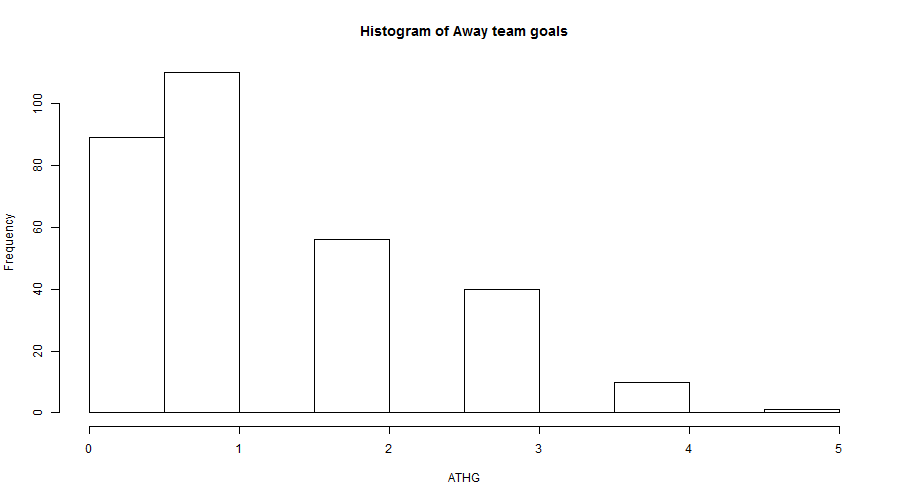
1ST Qu: 0

Median: 1

Mean: 1.265

3rd Qu: 2

Max: 5



The histogram is also right-skewed.

**HS (Home Team Shots)**

Type: Continuous numerical

Min: 1

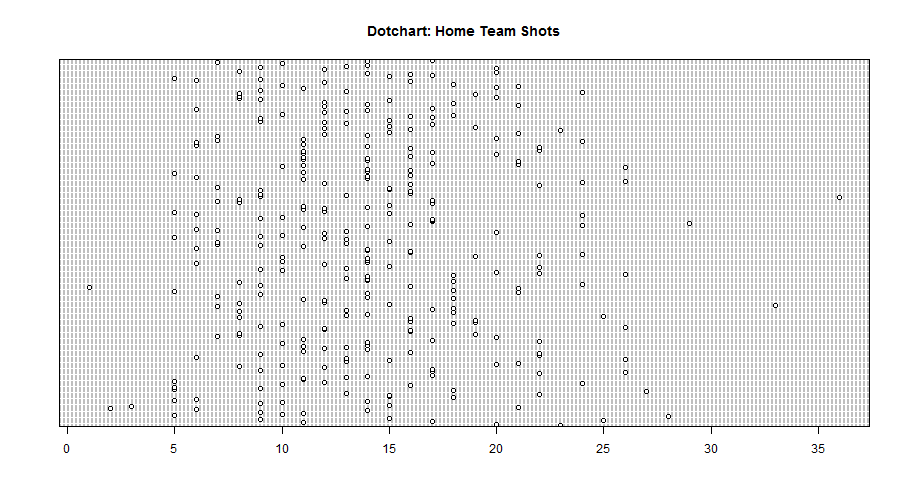
1ST Qu: 10

Median: 14

Mean: 14.04

3rd Qu: 17

Max: 36



**AS (Away Team Shots)**

Type: Continuous numerical

Min: 2

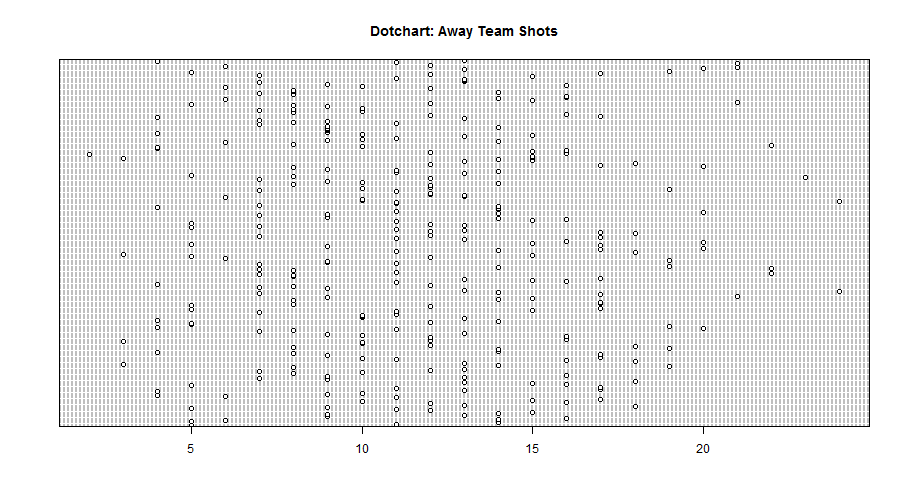
1ST Qu: 8

Median: 11

Mean: 11.6

3rd Qu: 15

Max: 24



**HST (Home Team Shots on Target)**

Type: Continuous numerical

Min: 0

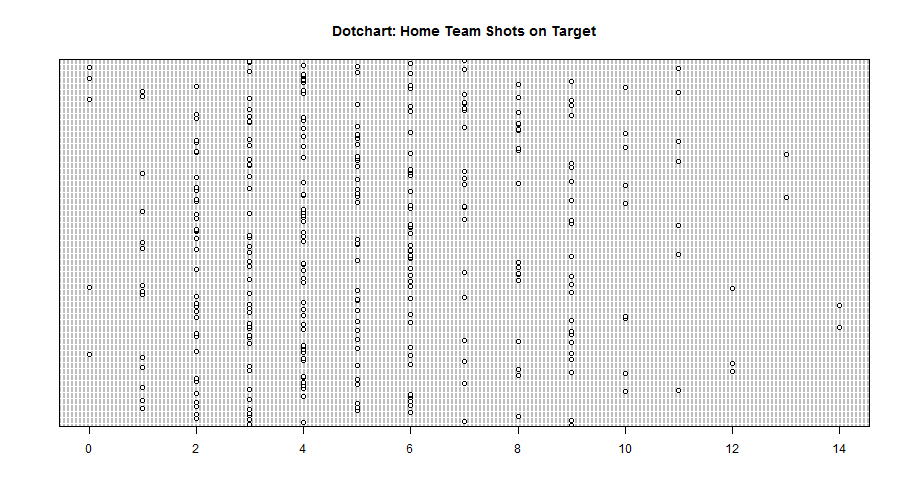
1ST Qu: 3

Median: 5

Mean: 5.173

3rd Qu: 7

Max: 14



**AST (Away Team Shots on Target)**

Type: Continuous numerical

Min: 0

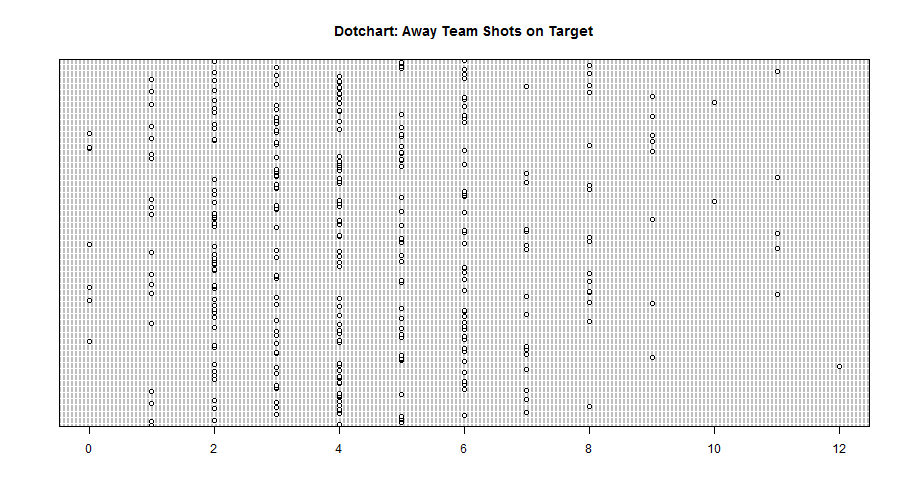
1ST Qu: 3

Median: 4

Mean: 4.333

3rd Qu: 6

Max: 12



**HF (Home Team Fouls Committed)**

Type: Continuous numerical

Min: 4

1ST Qu: 11

Median: 14

Mean: 14.18

3rd Qu: 17

Max: 26

stem-leaf:

4 | 0

5 | 0

6 | 0

7 | 000000000

8 | 00000000000000

9 | 0000000000

10 | 0000000000000000000000000

11 | 000000000000000000000

12 | 000000000000000000000000000000

13 | 0000000000000000000000000000

14 | 000000000000000000000000000

15 | 000000000000000000000000000

16 | 0000000000000000000

17 | 000000000000000000000000000000

18 | 000000000000000000

19 | 000000000000

20 | 0000000000000

21 | 0000000

22 | 0000000

23 | 0

24 | 0000

25 |

26 | 0

**AF (Away Team Fouls Committed)**

Type: Continuous numerical

Min: 2

1ST Qu: 12

Median: 15

Mean: 14.86

3rd Qu: 18

Max: 29

stem-leaf:

2 | 00

4 | 00

6 | 0000

8 | 000000000000000000000000

10 | 0000000000000000000000000000000000000

12 | 0000000000000000000000000000000000000000000000000000

14 | 0000000000000000000000000000000000000000000000000

16 | 000000000000000000000000000000000000000000000000000

18 | 000000000000000000000000000000000000

20 | 00000000000000000000000000000

22 | 00000000000000000

24 | 00

26 |

28 | 0

**HC (Home Team Corners)**

Type: Continuous numerical

Min: 0

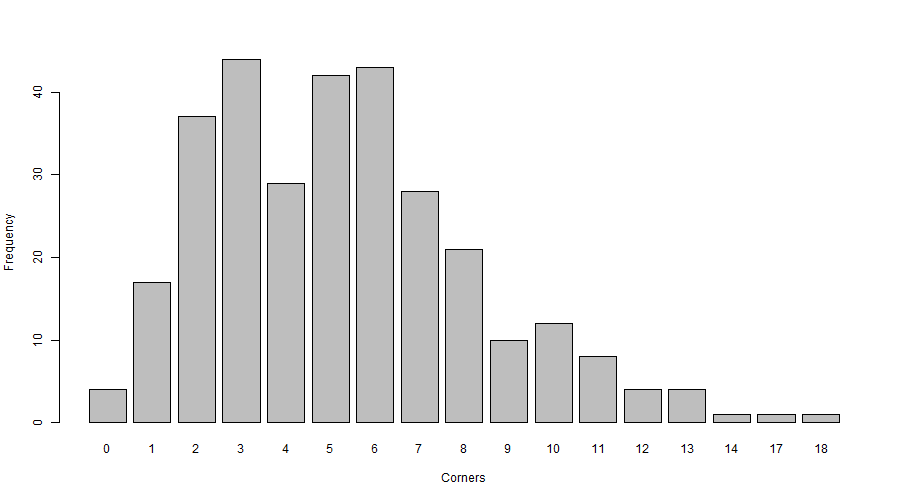
1ST Qu: 3

Median: 5

Mean: 5.288

3rd Qu: 7

Max: 18



**AC (Away Team Corners)**

Type: Continuous numerical

Min: 0

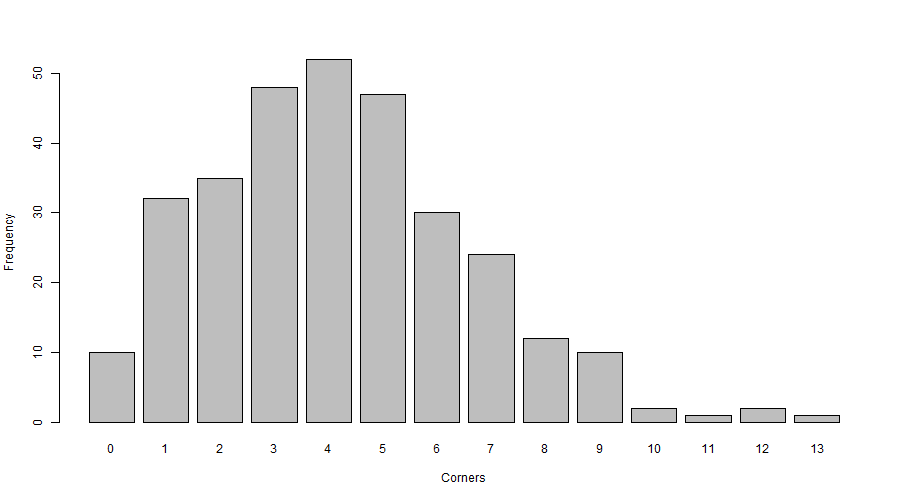
1ST Qu: 2.25

Median: 4

Mean: 4.219

3rd Qu: 6

Max: 13



**HY (Home Team Yellow Cards)**

Type: Continuous numerical

Min: 0

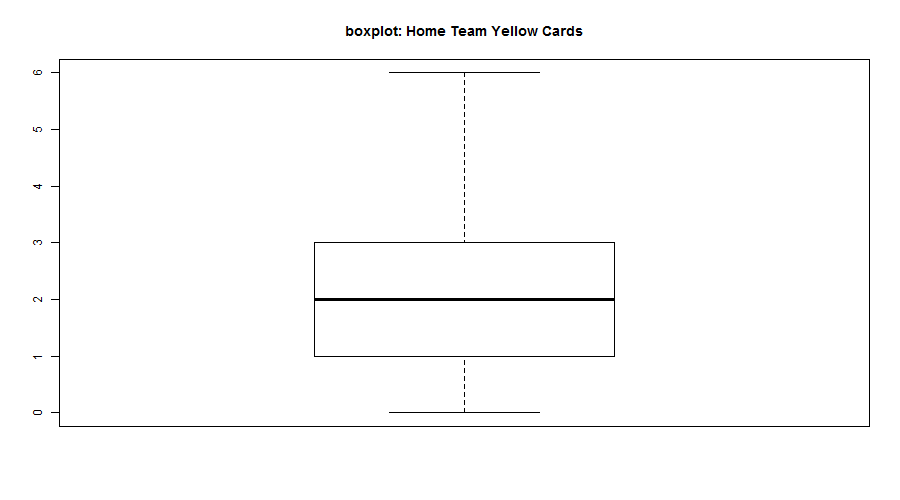
1ST Qu: 1

Median: 2

Mean: 1.791

3rd Qu: 3

Max: 6



**AY (Away Team Yellow Cards)**

Type: Continuous numerical

Min: 0

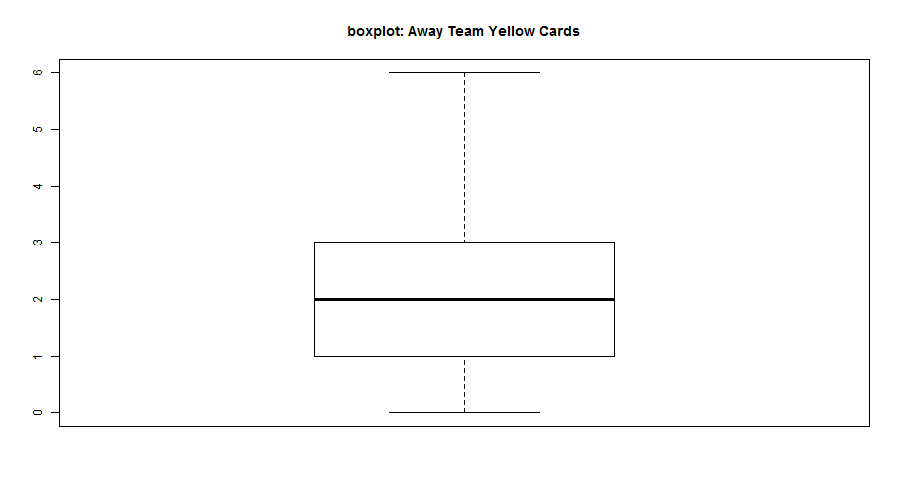
1ST Qu: 1

Median: 2

Mean: 2.003

3rd Qu: 3

Max: 6



**HR (Home Team Red Cards)**

Type: Continuous numerical

Min: 0

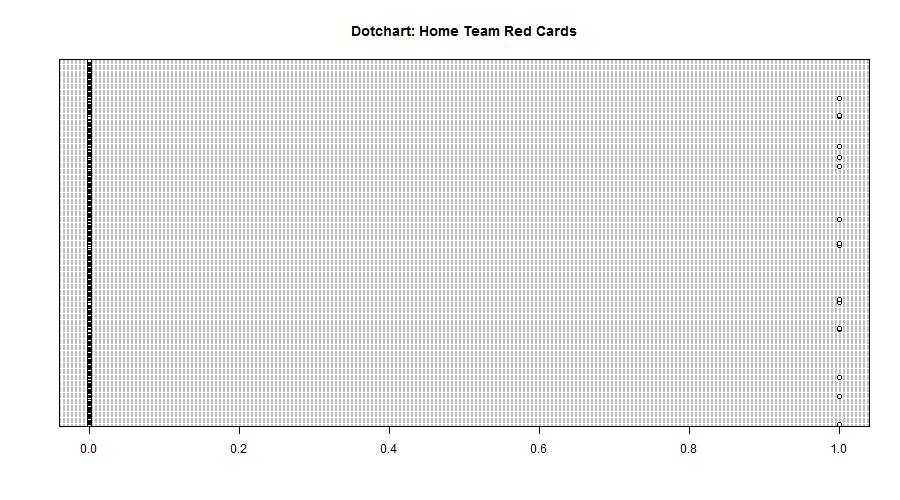
1ST Qu: 0

Median: 0

Mean: 0.05229

3rd Qu: 0

Max: 1



**AR (Away Team Red Cards)**

Type: Continuous numerical

Min: 0

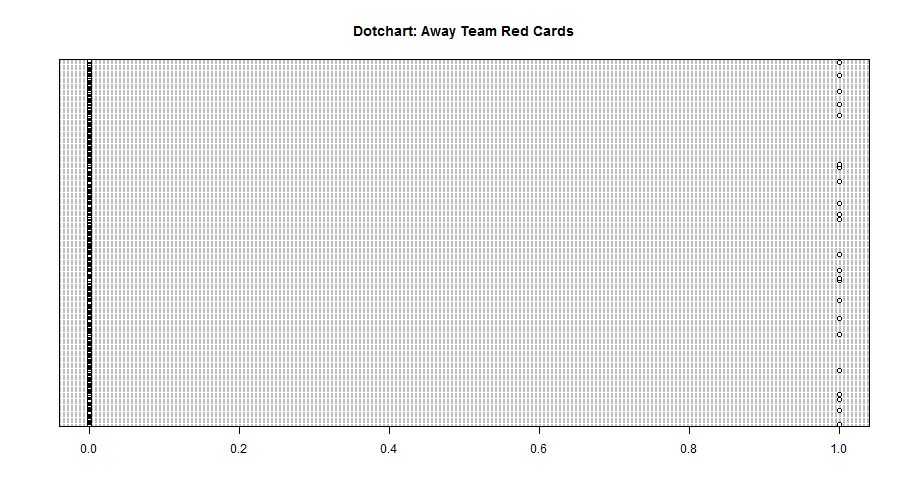
1ST Qu: 0

Median: 0

Mean: 0.07516

3rd Qu: 0

Max: 1



We should reconsider the feature of red cards gained, because its frequency and numbers are too small comparing to other those of other features.

**cardinality for each feature**

FTR

[1] 3

HF

[1] 22

AF

[1] 25

HS

[1] 30

AS

[1] 23

HST

[1] 15

AST

[1] 13

HST.

[1] 80

AST.

[1] 72

HC

[1] 17

AC

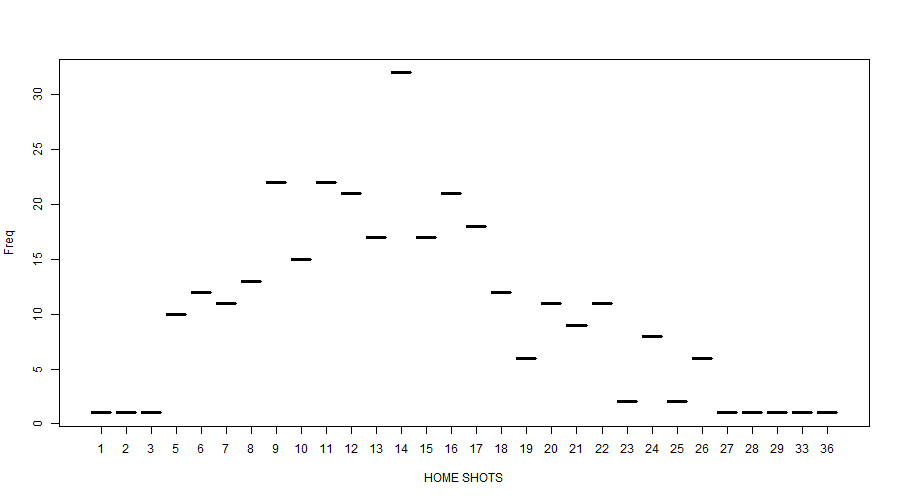
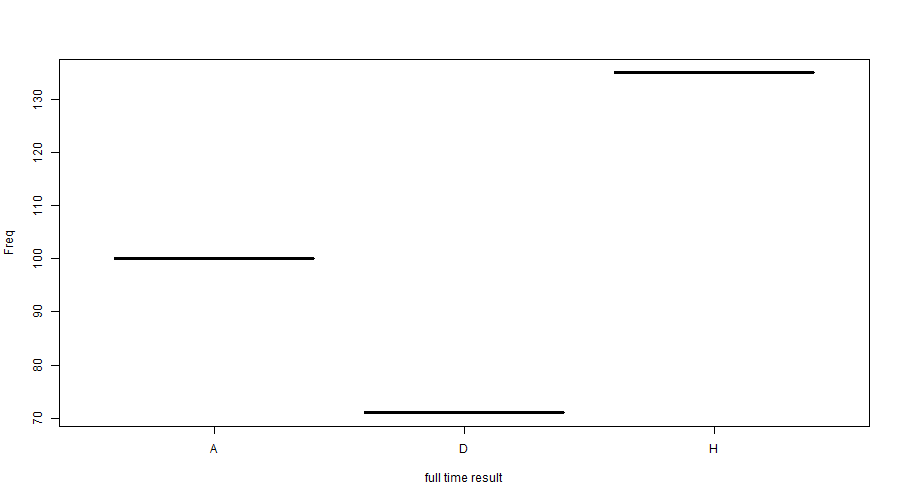
[1] 14

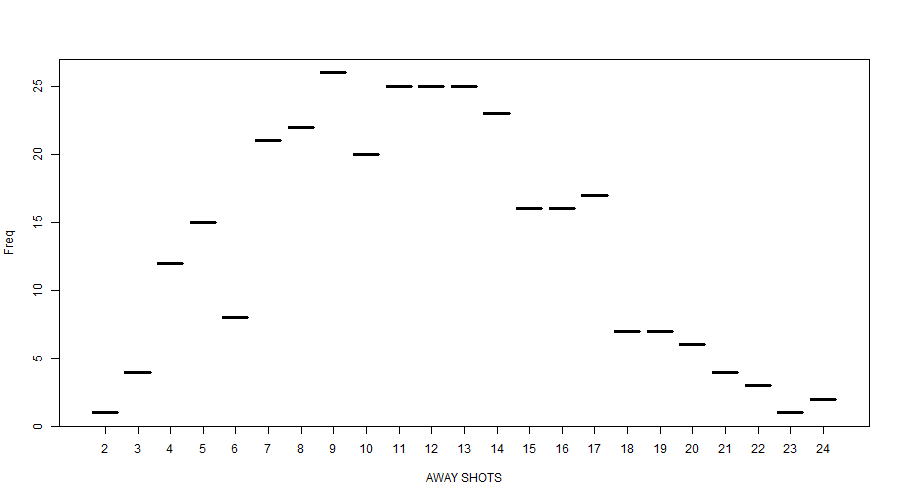
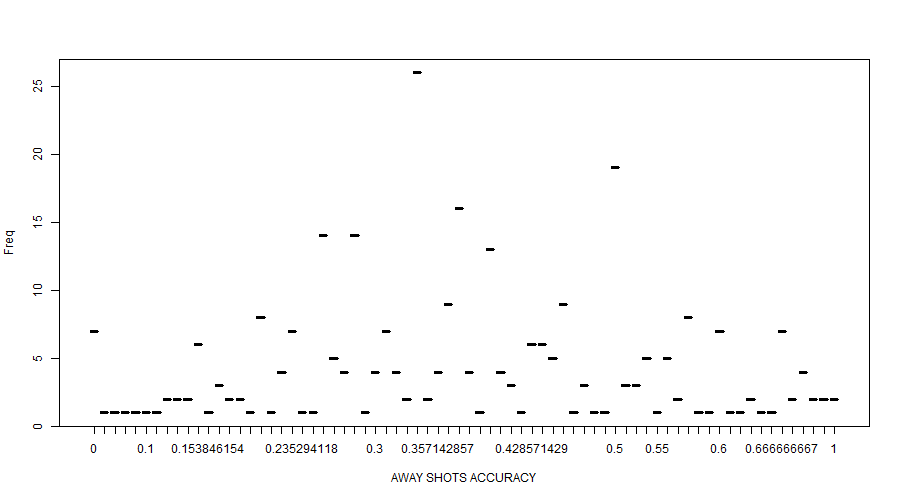
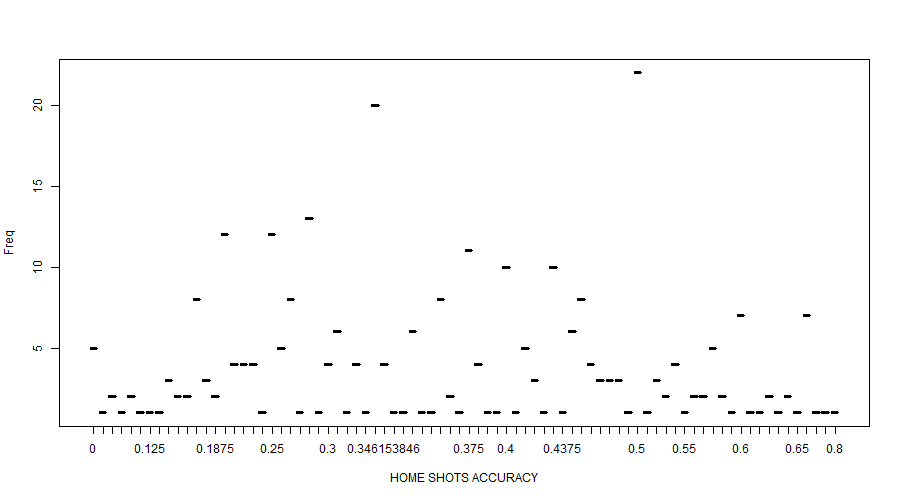
HR

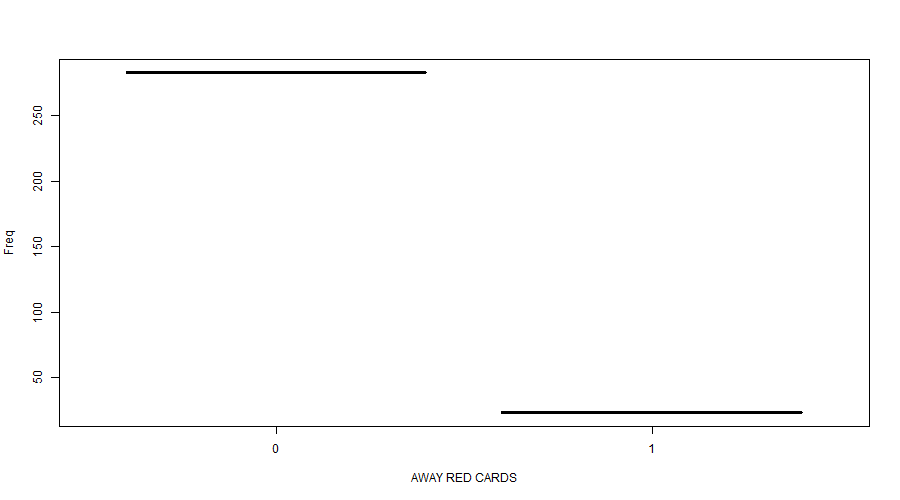
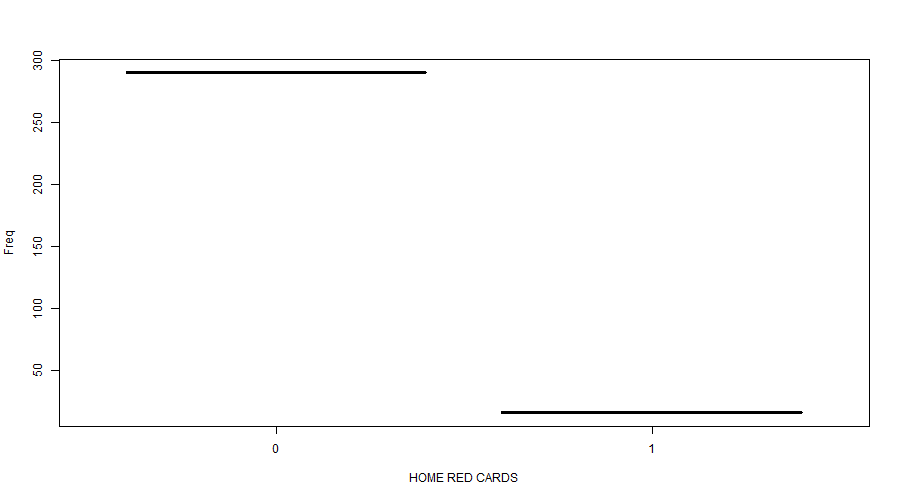
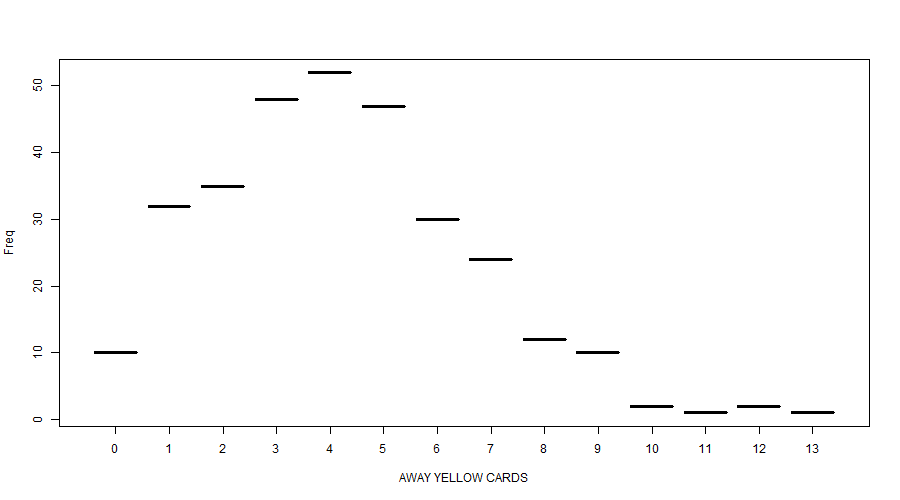
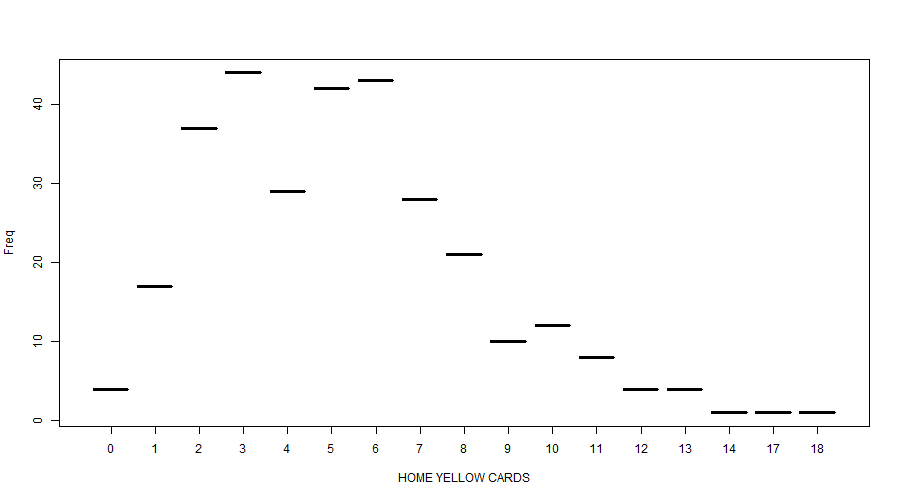
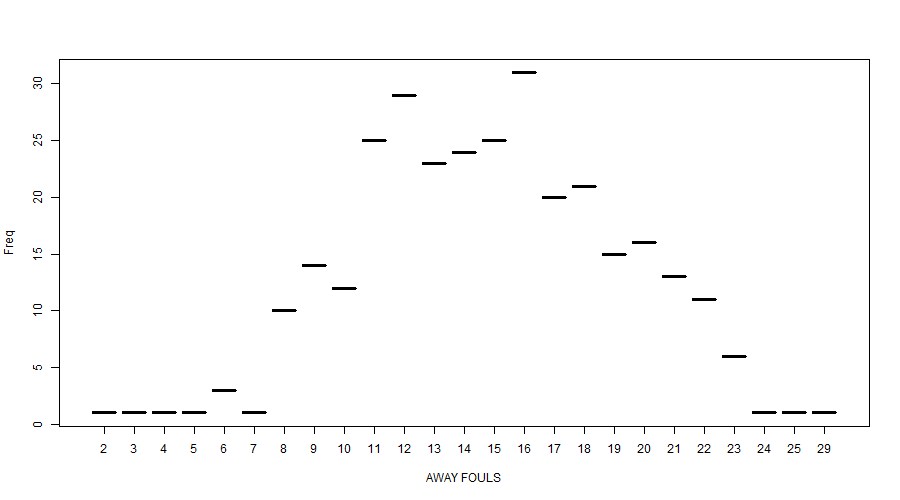
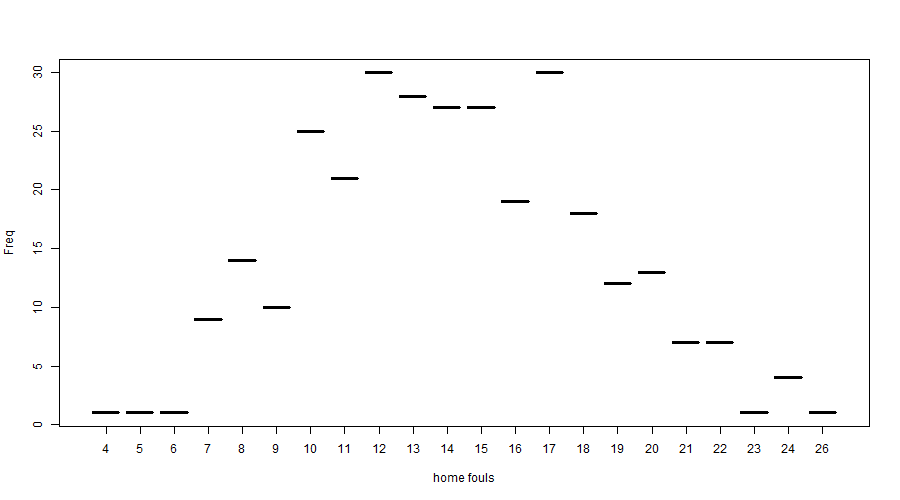
[1] 2

AR

[1] 2







**Entropy**

FTR

[1] 1.537166

HS

[1] 4.393328

AS

[1] 4.165869

HST.

[1] 5.705956

AST.

[1] 5.53944

HF

[1] 3.998017

AF

[1] 4.080976

HC

[1] 3.505024

AC

[1] 3.238448

AR

[1] 0.3849035

HR

[1] 0.2960361

**UPDATES**

**After Altered Features**

**Correlation Test**

FTR ~ HT

Full time result ~ Home Stadium Attendance

t = 4.1755, df = 304, **p-value** = 3.888e-05

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1240101 0.3362387

sample estimates:

**cor: 0.2328954**

FTR ~ STRdf

Full time result ~ Shot on target rate difference between home and away team

t = 7.5671, df = 304, **p-value = 4.581e-13**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2993626 0.4884429

sample estimates:

**cor: 0.3981229**

FTR ~ Fdf (x)

Full time result ~ Fouls committed difference between home and away team

t = -0.85674, df = 304, **p-value = 0.3923 (x)**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.16031971 0.06339402

sample estimates:

**cor -0.04907837 (x)**

FTR ~ Ydf (x)

Full time result ~ Yellow Card gained difference between home and away team

t = -1.4895, df = 304, **p-value = 0.1374 (x)**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.19537688 0.02726581

sample estimates:

**cor: -0.08511796 (x)**

FTR ~ Rdf

Full time result ~ Red card gained difference between home and away team

t = -2.2607, df = 304, **p-value = 0.02448**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.23728712 -0.01670187

sample estimates:

**cor: -0.1285846**

FTR ~ Cdf

Full time result ~ Connors difference between home and away team

t = 0.9758, df = 304, **p-value = 0.3299 (x)**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.05659979 0.16695590

sample estimates:

**cor: 0.05587837**

After conducting correlation tests on different features with the column of full time match result, many features, including those ones altered from fouls committed, yellow card gained, corners played are not statistically associated with the match result -- they have p-values much higher than the significance level of 0.05 and very low correlation rate. Home stadium attendance, shots on target rate and red card gained could be considered statistically significant predictors of match result. The deleted features below have been updated accordingly.

**Deleted Features**

1. Cdf (Corners difference between two teams)

Also, deleted the two source variables:

HC = Home Team Corners

AC = Away Team Corners

1. Fdf (Fouls committed difference between two teams)

Also, deleted the two source variables:

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

1. Ydf (Yellow cards gained difference between two teams)

Also, deleted the two source variables:

HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

**Summary**

After deleting three features, this dataset only has three features left. Only the feature of shots on target rate can be considered as strongly correlated with match results, and the feature of red card gained should be considered as weak and questionable, because its p-value is not very small and correlation rate is comparatively small, and the numbers within that column are unusually distributed: most attributes have “0” s, and only a few of them have “1” or “-1”. Considering the current situation and overall data quality, it is necessary to find more potentially meaningful features, such as possession rate, total running distance, total market value of line-ups, average age of line-ups etc.