EDA_document

Haoran Hu

2019-3-29

Question

How to choose the best model for predicting soccer players' wage using FIFA video game data?

Data source

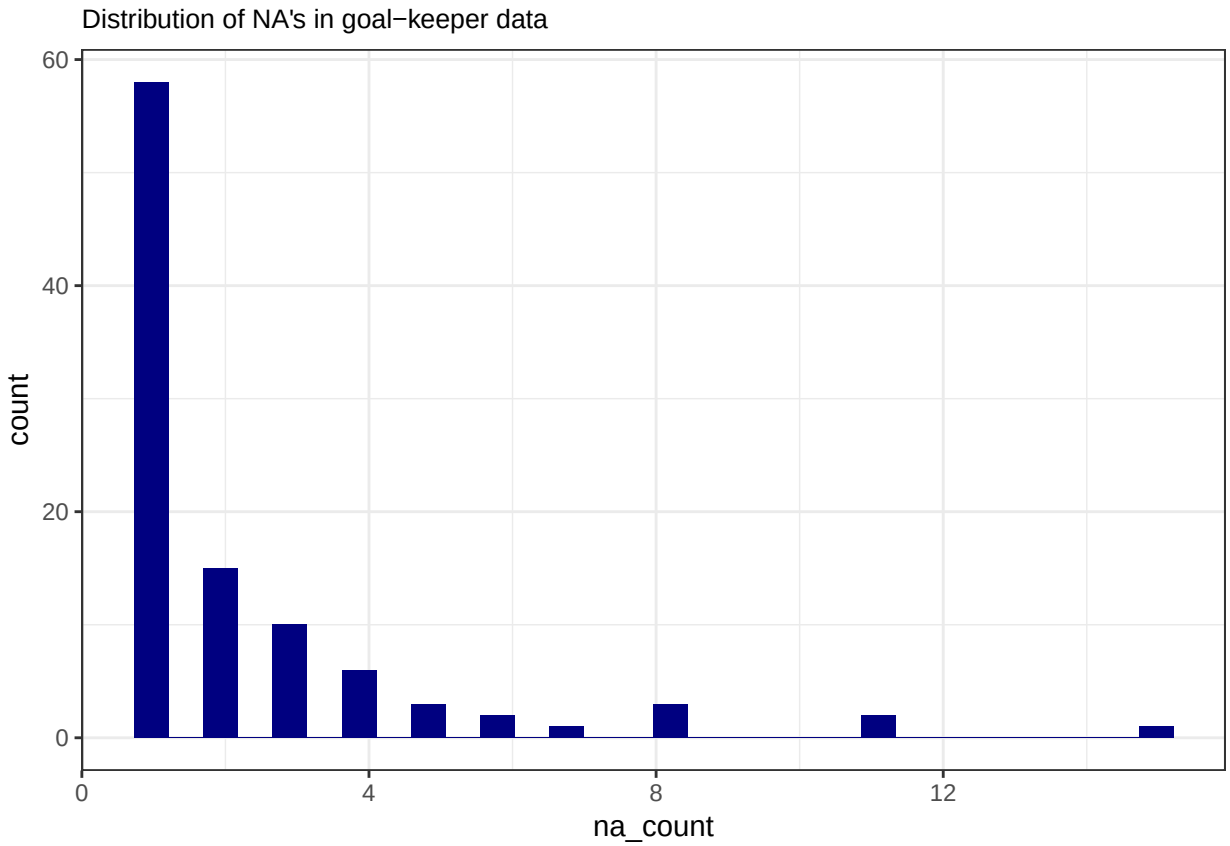
The anticipated data source of our project is a soccer player information database from the latest version of FIFA video game. The dataset contains information of 18200 soccer players. It includes 82 attributes of the players, such as market value, wages, age, and score of each specific skill. Although the data source is a video game database, it accurately reflects current soccer players' personal information, skills, values, etc. Therefore, it is sensible to use it to model real world situation, and to analyze the association between players' values and various characteristics of the players. After data cleaning, there are 68 predictors for players' values.

Data cleaning

(I made some changes to many variables, details are described in "Exploratory_analysis.Rmd/pdf").

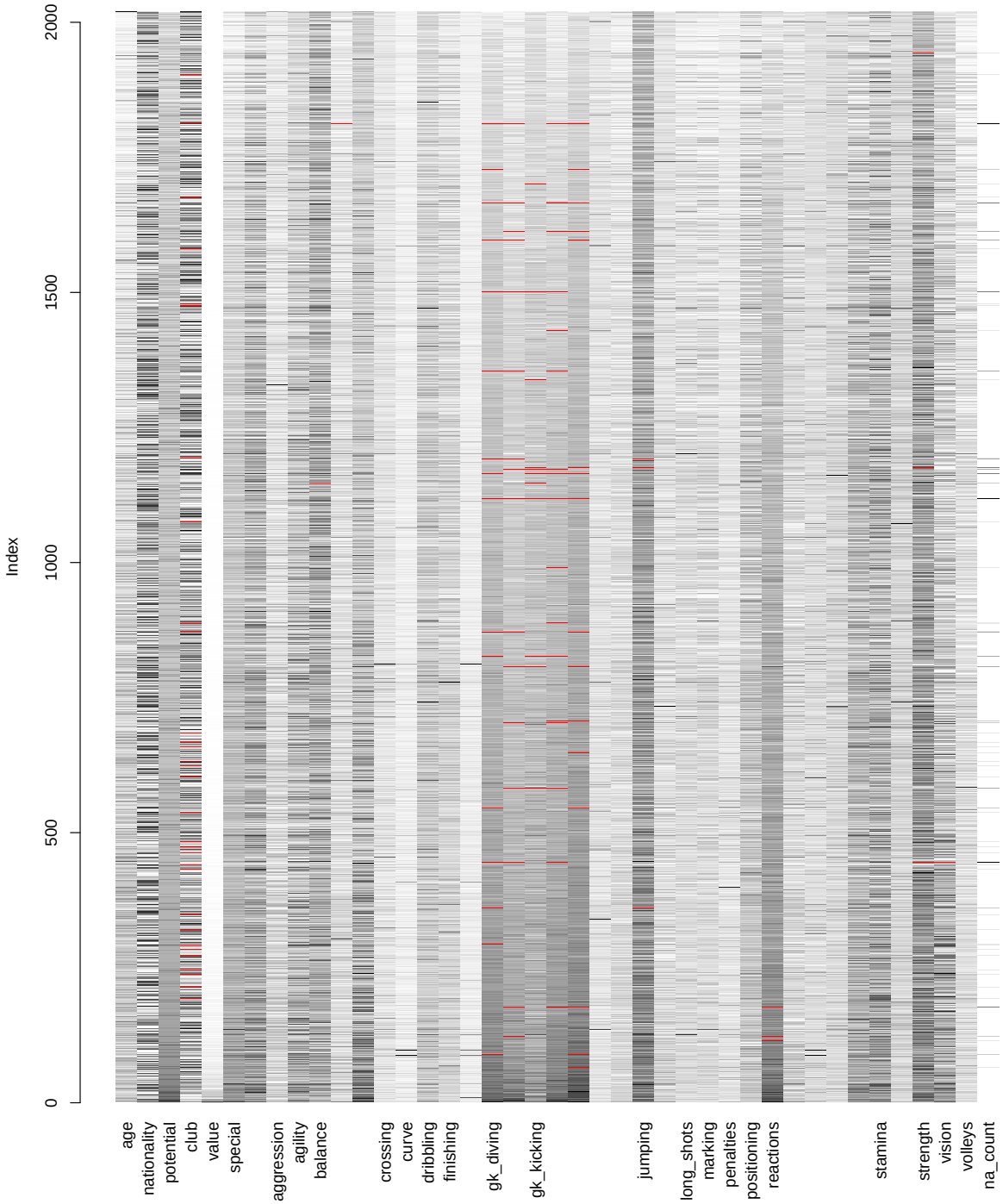
Checking and replacing NA's

First, I checked missing data for each observation. Since a lot of NA's are caused by missing position score data of goal keepers, and These NA's do not matter, I will amalyze goal-keeper data and non goal-keeper data seperately.

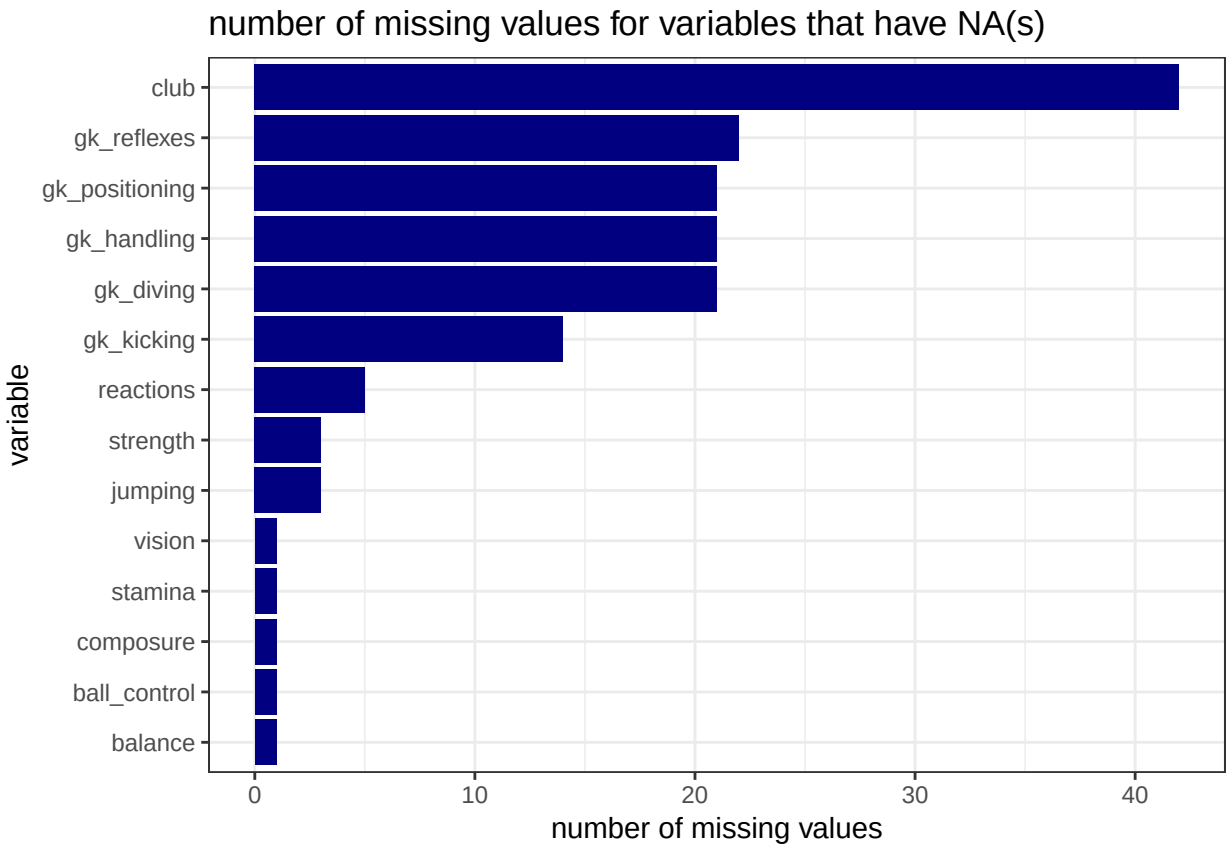


From the plot, it seems to be reasonable to drop observations that have more than 5 missing values. Next, I:

- dropped observations that have more than 5 missing values
- checked number of NA's in each variable



In the plot above, red color represents missing data.



In total, 14 variables have missing values. The "club" variable has the largest number of missing values(42 missing values). Note that all the players that have missing "club" values have market values = 0. This actually means they do not belong to any club for the time being. Each of the other variables that have missing values has less than 25 missing values. Compared to the sample size, that's not a big loss of data, and thus all the variables can be kept in the model.

Therefore, I

- replaced NA's in "club" column with a new category "none"
- replaced NA's in each column except "club" column by the mean of that column. This will be done seperately for goal keepers and for non goal keepers.

Tables for descriptive statistics

Table 1: Factor variables

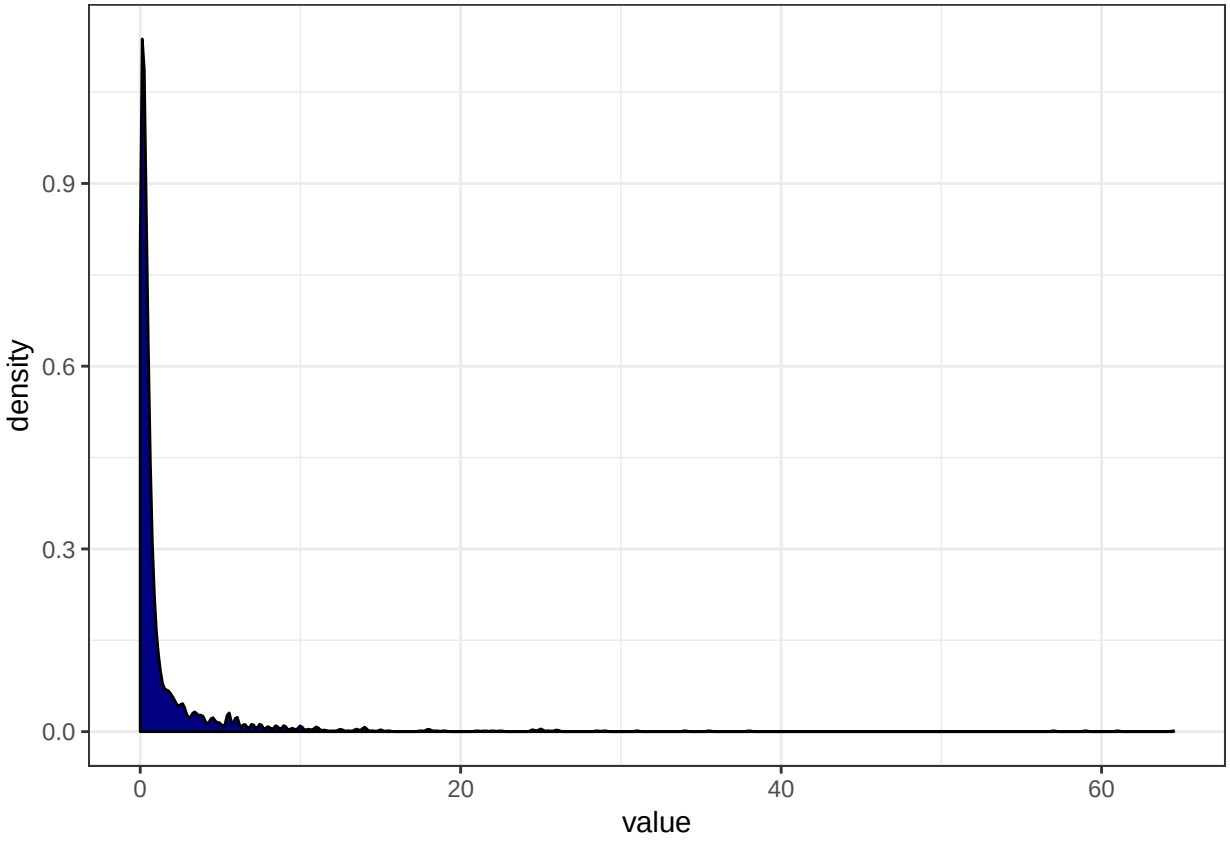
variable	unique levels
club	648
nationality	95

Table 2: Integer/numeric variables

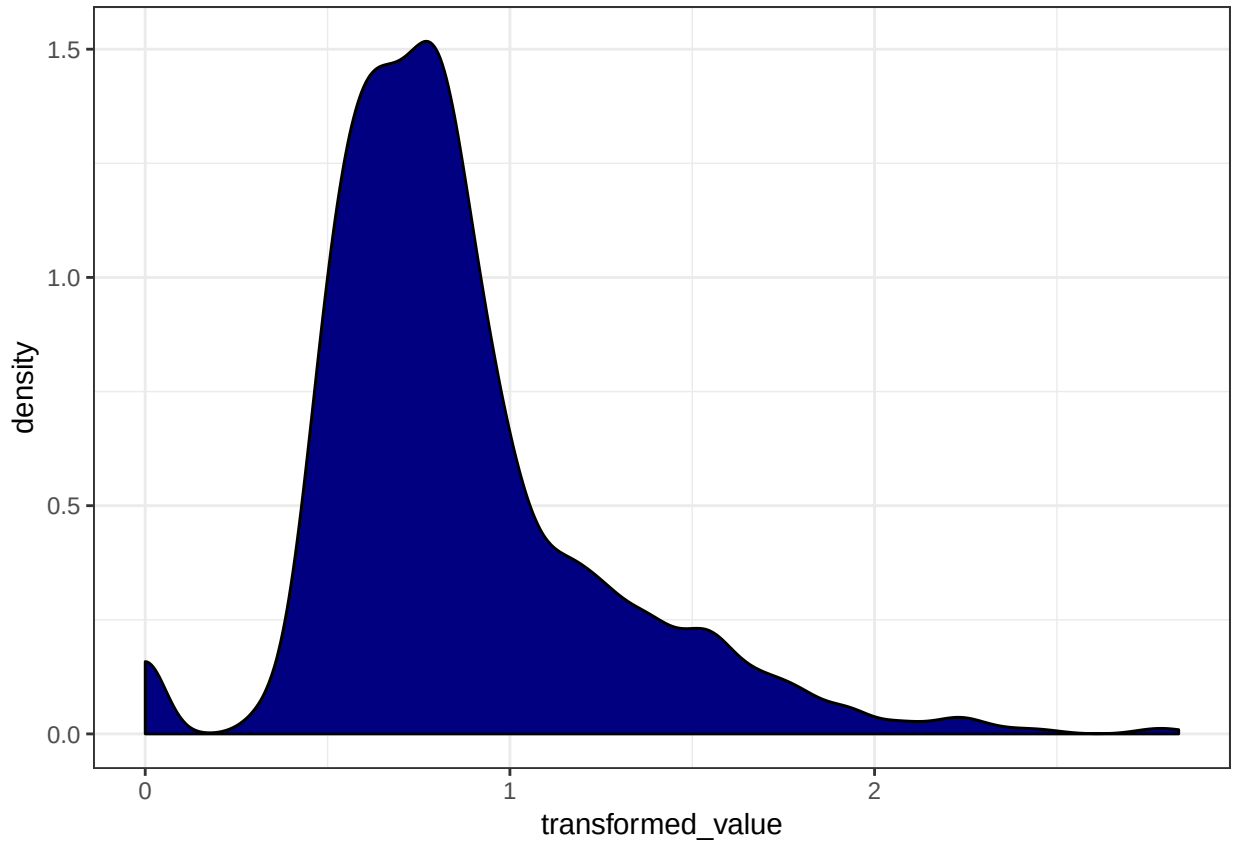
variable	Min	1st Q	Mean	Median	3rd Q	Max	Std Dev
acceleration	11	31	38.87	40	47	65	10.95
age	16	22	26.08	26	30	47	5.4
aggression	11	21	26.73	25	32	68	7.85
agility	14	32.75	40.67	38	48	74	11.55
crossing	5	12	14.41	14	16	45	4
curve	6	12	14.84	14	17	65	4.54
dribbling	2	11	14.05	14	16	33	4.26
finishing	2	10	12.42	12	15	34	4.06
free_kick_accuracy	4	12	14.57	14	16	72	4.58
heading_accuracy	4	12	14.6	14	17	47	4.11
interceptions	4	13	17.53	18	22	55	5.84
long_passing	7	20	25.53	24	30	73	7.92
long_shots	3	10	13.19	13	16	40	4.44
marking	4	10	12.78	13	16	35	4.24
na_count	0	0	0.078	0	0	5	0.43
penalties	5	15	20.4	20	24	73	6.92
positioning	2	8	11.7	12	15	27	4.13
potential	46	65	69.71	69.5	74	94	6.44
short_passing	10	22	26.9	26	31	66	7.39
shot_power	3	19	22.55	22	24	70	7.04
sliding_tackle	4	12	14.13	13	16	35	3.4
special	728	970.75	1050.17	1063	1134	1493	126.46
sprint_speed	11	32	39.29	41	47	70	10.88
standing_tackle	4	12	14.23	14	16	34	3.46
volleys	4	10	13.02	13	16	40	4.54
balance	11	36	43.27	43	51	70	10.84
ball_control	8	16	20.03	20	23	54	5.78
composure	5	27	36.52	33	45	71	12.73
gk_diving	39	60	65.34	65	71	91	7.87
gk_handling	43	58	62.91	63	68	91	7.88
gk_kicking	35	56	61.56	61	66	95	7.89
gk_positioning	38	57	63.07	63	69	91	8.79
gk_reflexes	37	61	66.23	66	72	90	8.32
jumping	15	52	58.19	59	66	85	11.44
reactions	28	52	59.24	60	67	88	10.3
stamina	12	25	30.68	31	36	48	7.55
strength	20	54	61.06	62	69	85	11.2
value	0	0.14	1.57	0.38	0.97	64.5	4.3
vision	10	27	36.09	35	45	72	12.9

Figures for descriptive statistics

For the outcome(value)

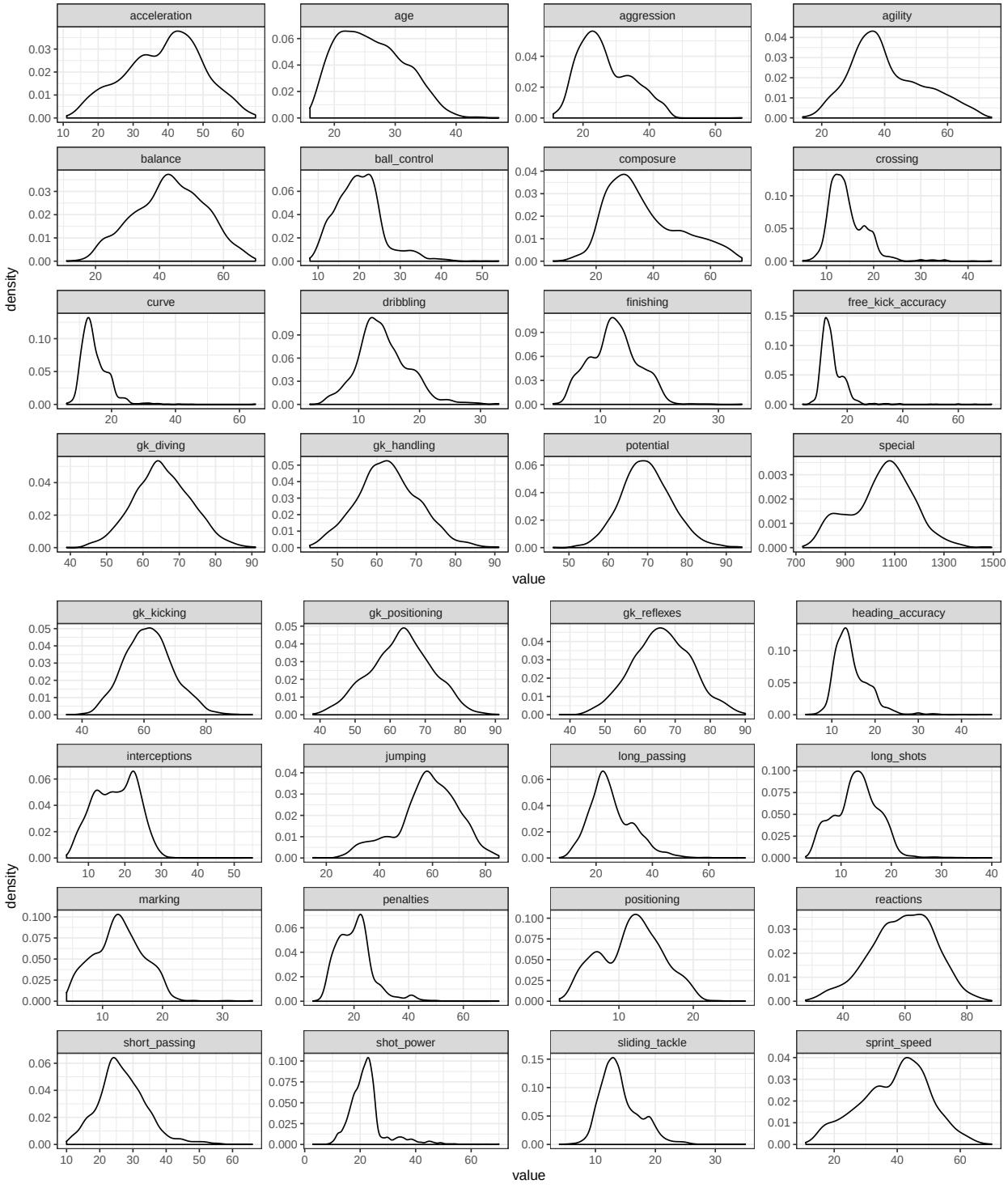


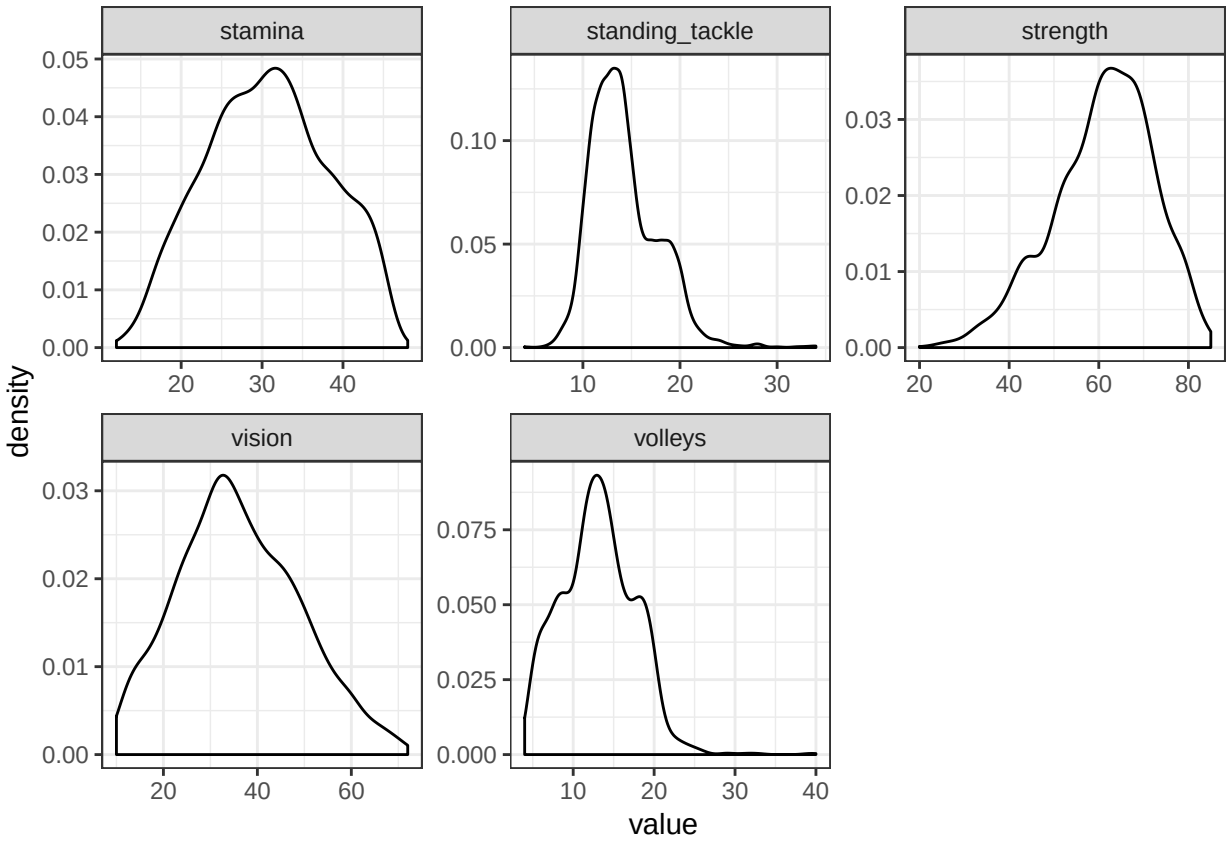
The outcome data is skewed, so I transformed it. New_value = Old_valueˆ(1/4)



After tramsformation, the distribution of the outcome is less skewed. There are a lot of 0 values becaues there are many players that do not belong to any club.

Check the distribution for each numeric/integer predictor





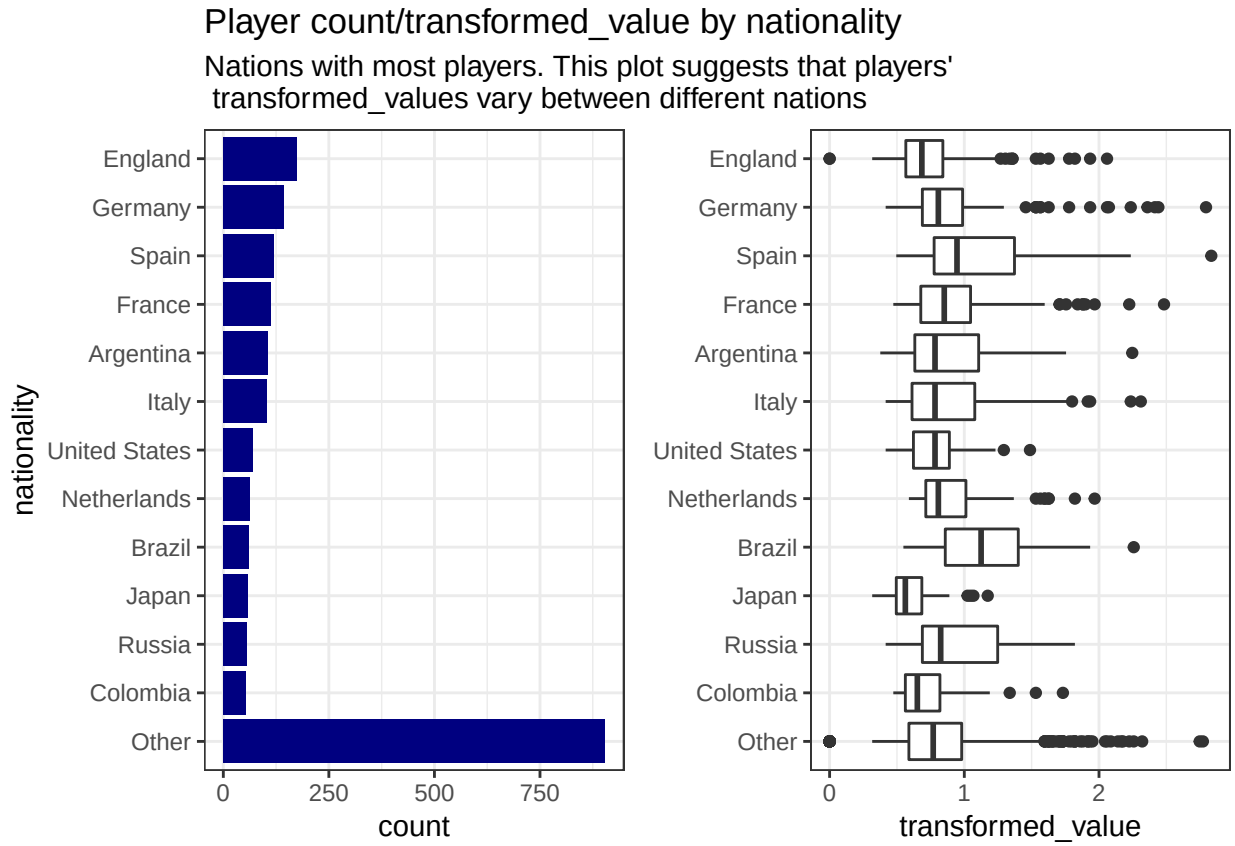
There seems not to be any data error. The variable `max_pos_score` has unusually large amount of 0 zero values because I set it to be 0 for goal keepers. For the variables whose density plots have two peaks, the smaller peaks are generally caused by goal keepers, and the larger peaks are generally caused by non-goal keepers.

plot for variables

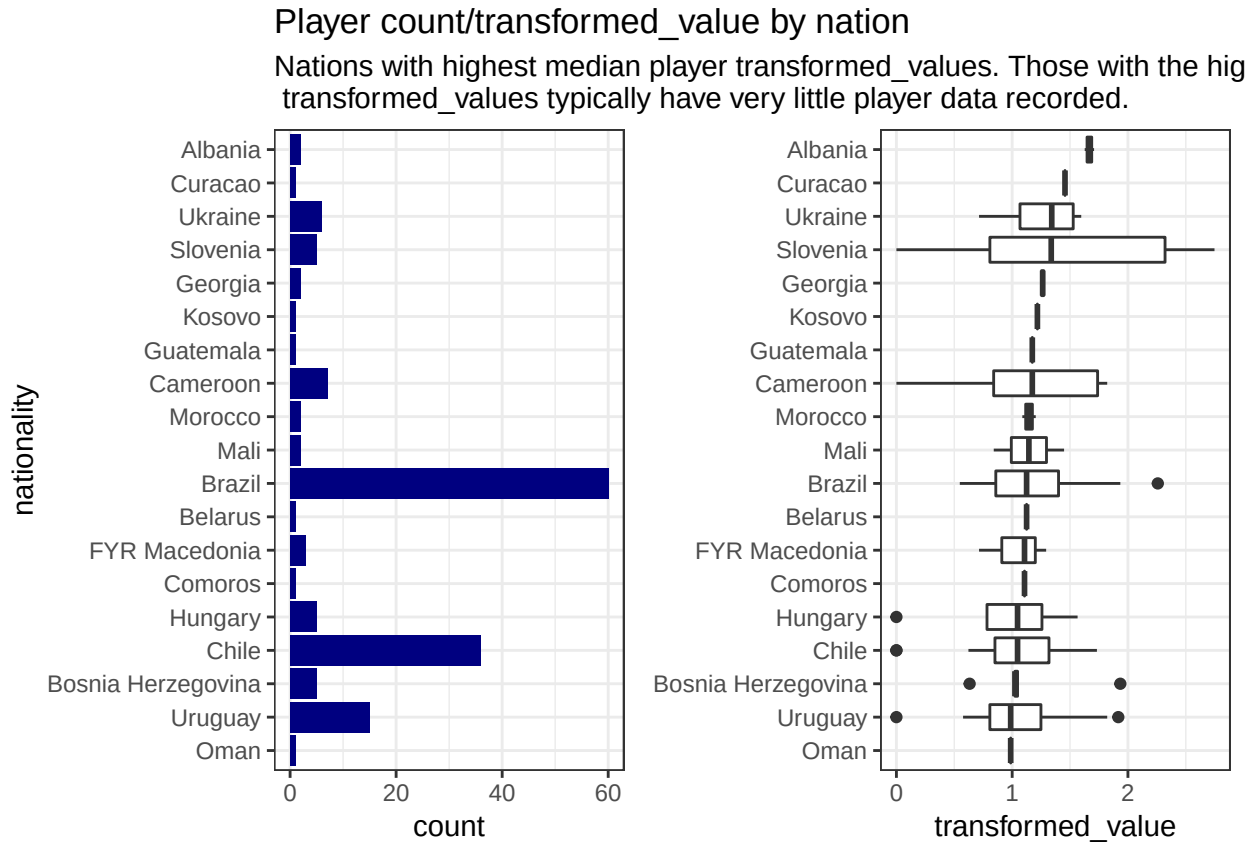
for factor variables

- nationality

plot the situation for nations with the most number of players:

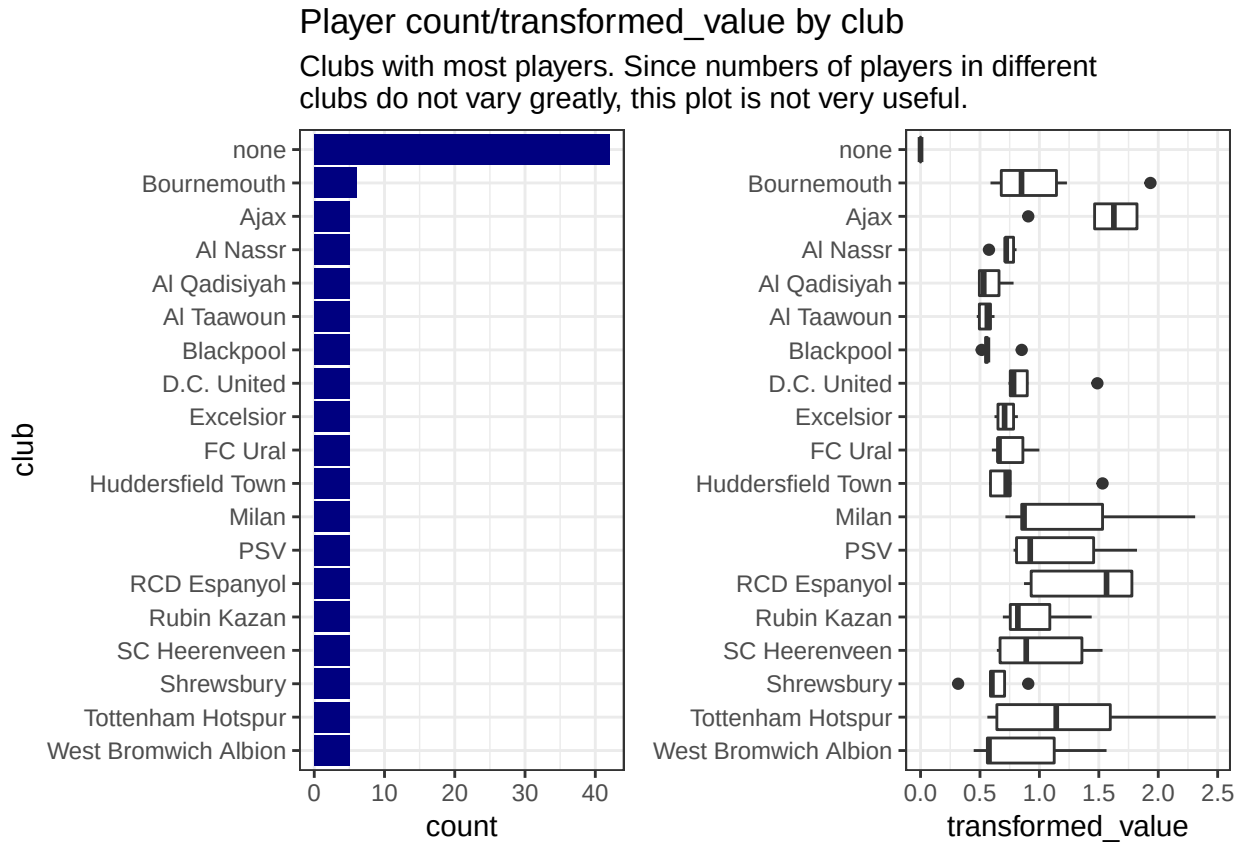


plot the nations with highest players' transformed_values:

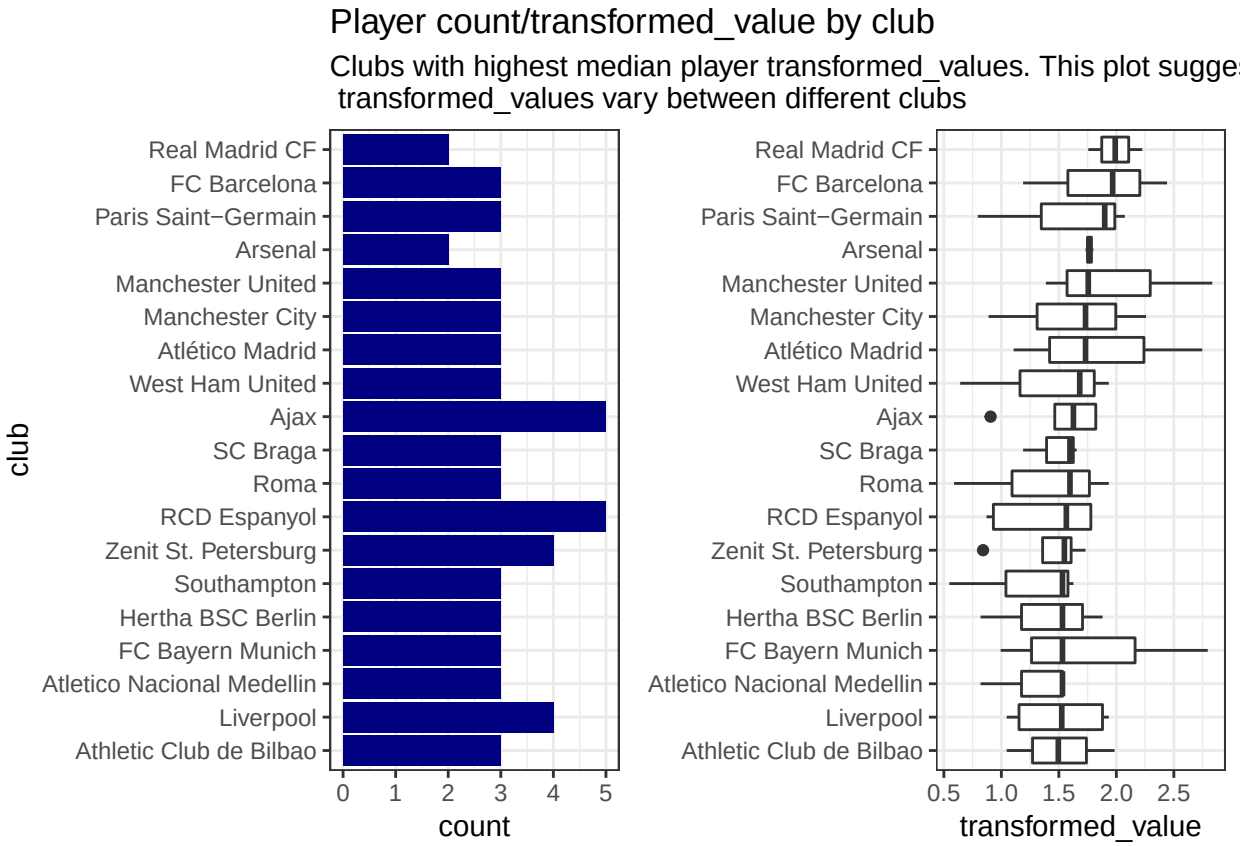


- club

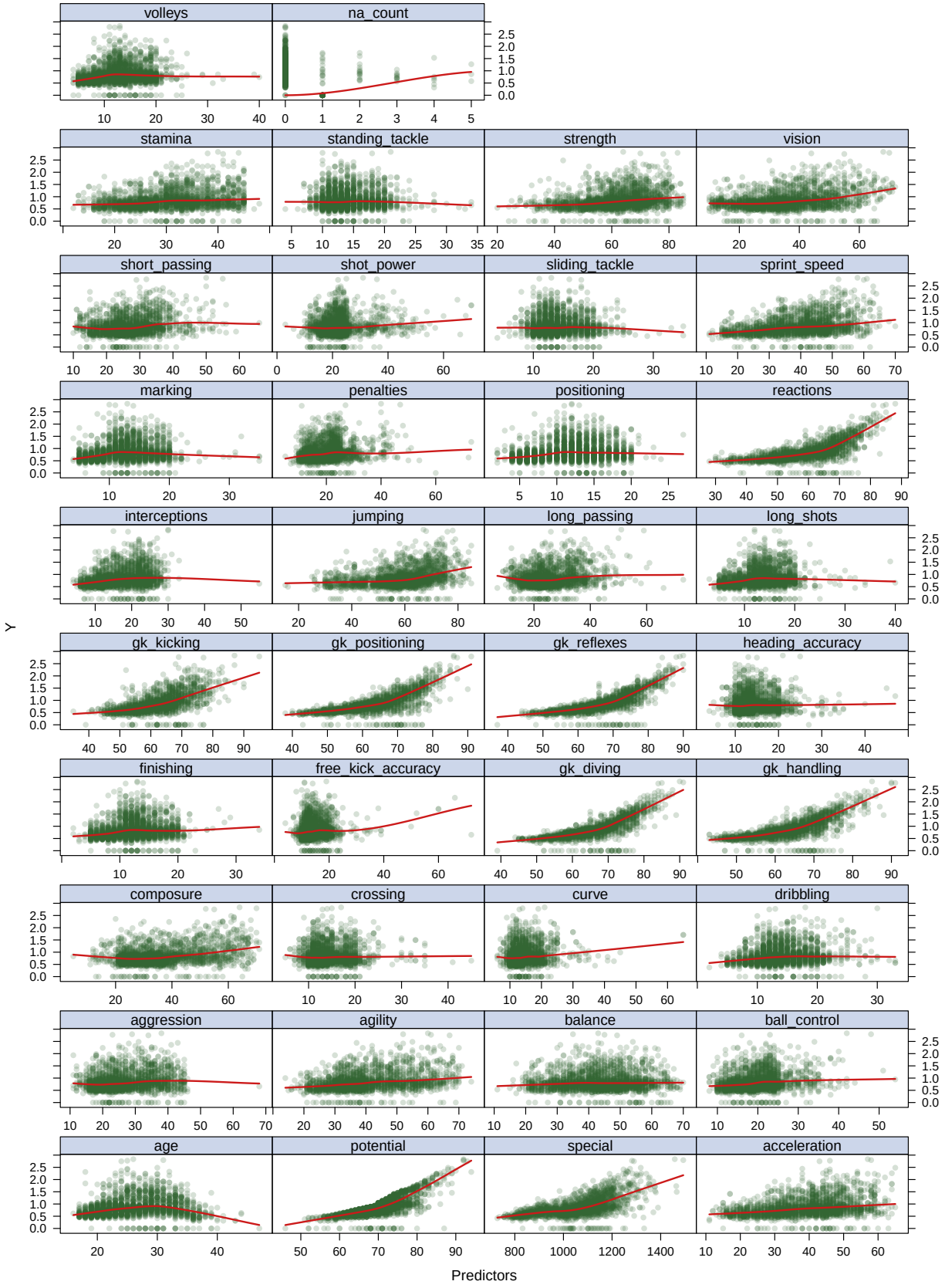
clubs that have the largest number of players



clubs that have the highest player median transformed_values

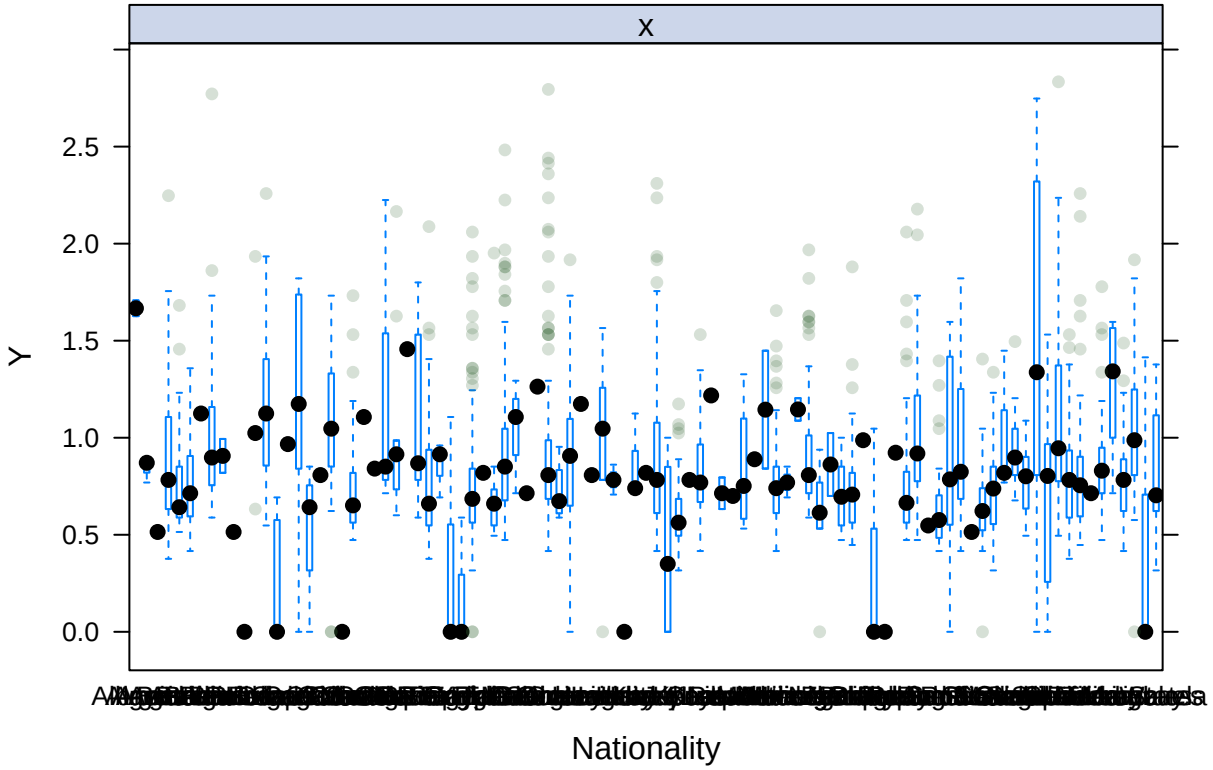

Player count/transformed_value by club
Clubs with highest median player transformed_values. This plot sugge
transformed_values vary between different clubs
club
Real Madrid CF
FC Barcelona
Paris Saint−Germain
Arsenal
Manchester United
Manchester City
Atlético Madrid
West Ham United
Ajax
SC Braga
Roma
RCD Espanyol
Zenit St. Petersburg
Southampton
Hertha BSC Berlin
FC Bayern Munich
Atletico Nacional Medellin
Liverpool
Athletic Club de Bilbao
0
1
2
3
4
5
count
0.5
1.0
1.5
2.0
2.5
transformed_value

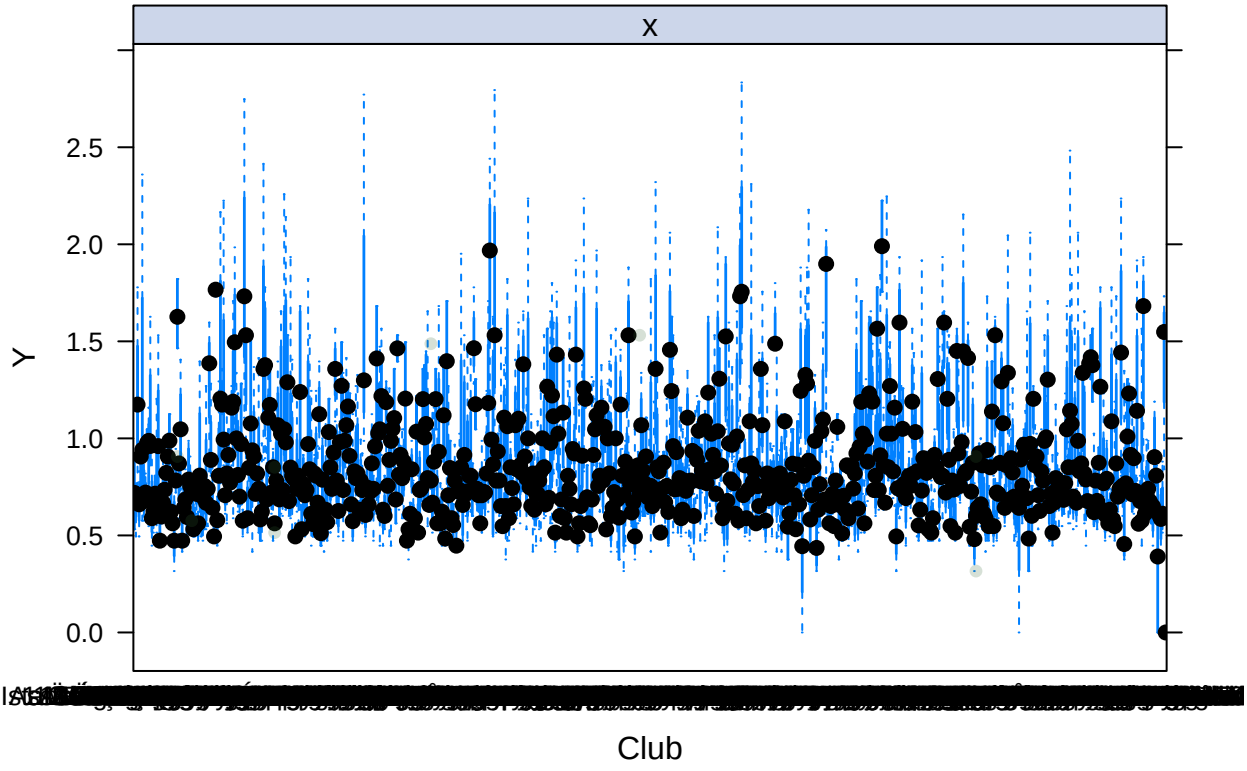
for int/num variables



These plots shows that the outcome tend to increase with the increase in levels of most of the numerical predictors, except for `age`. The relationships between the outcome and the predictors are possibly non-linear.

FeaturePlots for factor variables

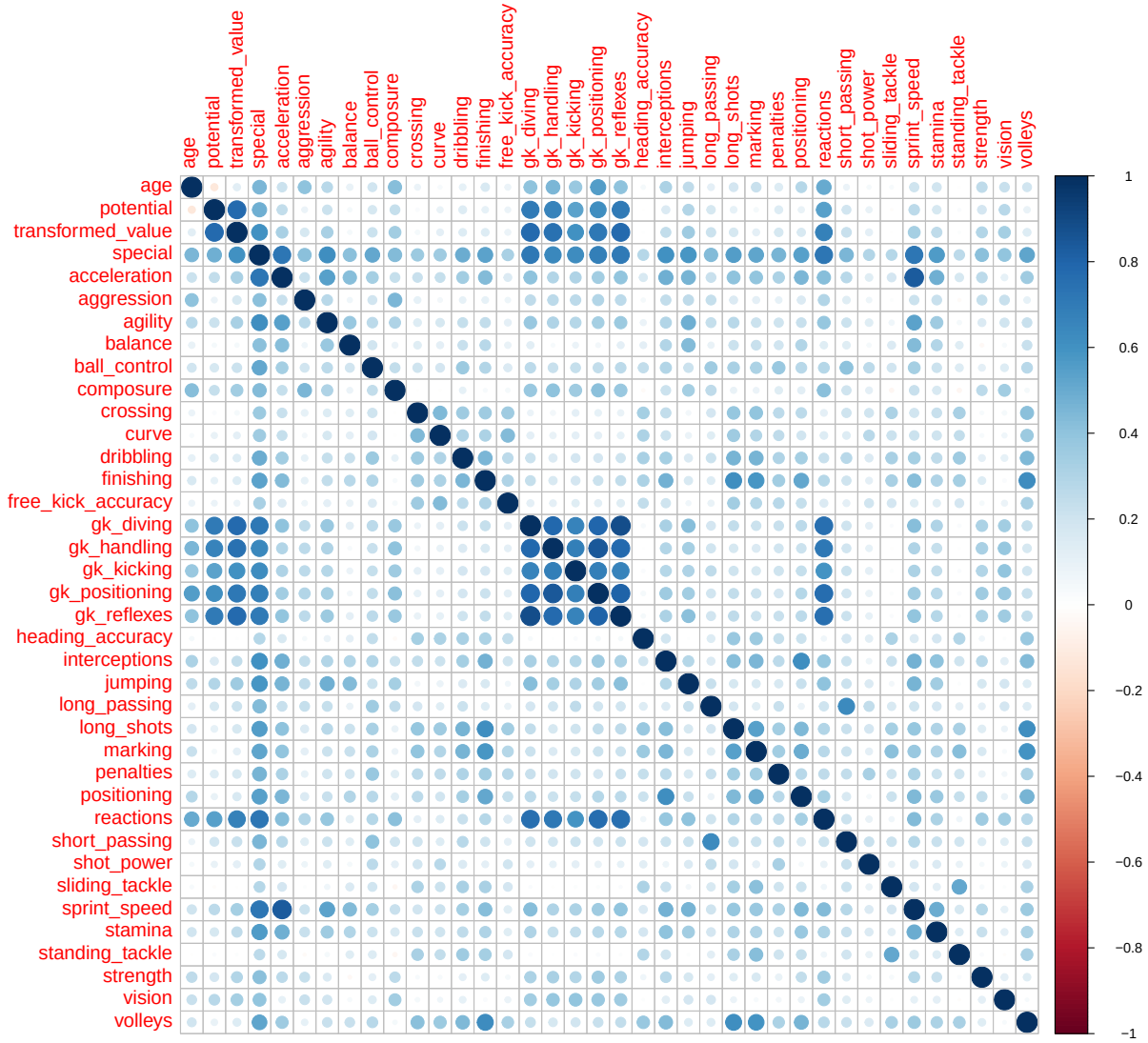




These plots shows that players' valus tend to vary between different nationalities, clubs and positions. In addition, players that have multiple preferred positions tend to have slightly greater values compared to those who only have one preferred position.

Correlation plot

The plot would be too large to display if we put in all the variables, so I just selected a subset of variables to display.



This plot shows that scores that are specificly measured for goal keepers(i.e. variable whose names start with "gk_") are highly correlated.

Split the data set into training and testing data

Training and testing datasets can be found in the exploratory analysis folder. There are also copies of the training and testing datasets in the "Data" folder.