In Pursuit of the Perfect Valuation

Development of a Regression
Model for the Market Values of European Football Players

Index

Dataset

transfermarket_fbref_201819

Websites of origin

- transfermarkt.com
- fbref.com

Characteristics

- 2100 rows
- 400 columns

Overview

This dataset consists of the combined statistics from the **2018-19 football season** of players competing in the **top 5 European leagues.**
The data comes from the two most prominent websites related to football transfers and statistics.

Purpose of the analysis

To understand which variables have the greatest influence on the market variable **“value”** and to construct regression models capable of predicting it.

Investigation

Dual Approach:

In order to examine the complex dynamics of soccer performance we split our dataset in: Goalkeepers and Outfield Players

Motivation:

- Recognizing **unique characteristics** and **demands** of each position
- Enables a **specific** analysis

Benefits:

- Facilitates a **deeper** exploration of variables.
- Improve **understanding** of performance influencers for both groups.

Data Visualization

Graphical analysis

Response Variable Analysis

- Distribution of the variable **“value”** in the two datasets.
- Effects of the **logarithmic transformation**.

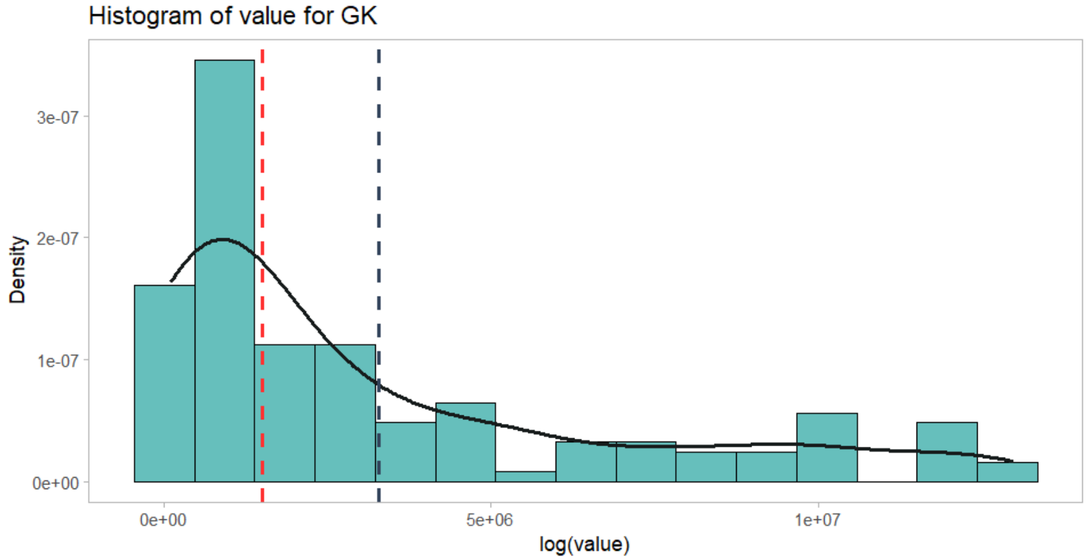
Goalkeeper’s Variables Analysis

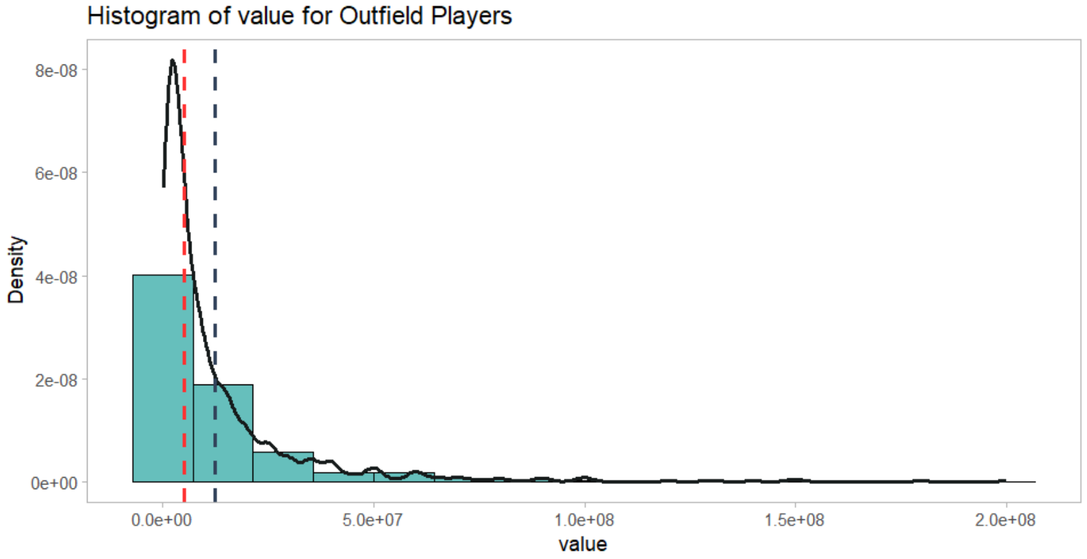
- **Distributions** of some predictors.
- **Relationships** between some **explanatory variables** and the **logarithm of the response variable**.

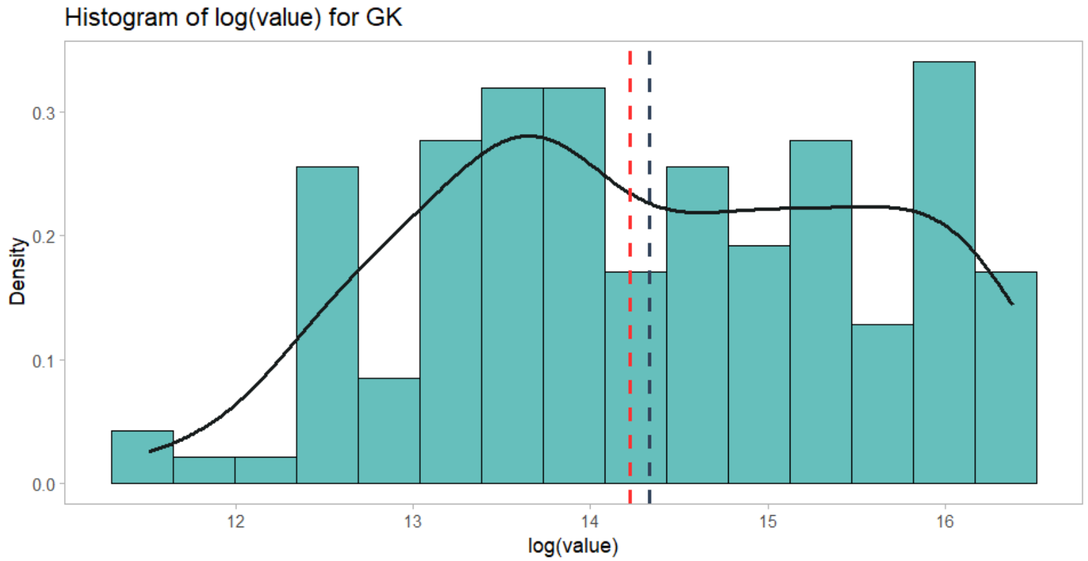
Outfield Players’ Variables Analysis

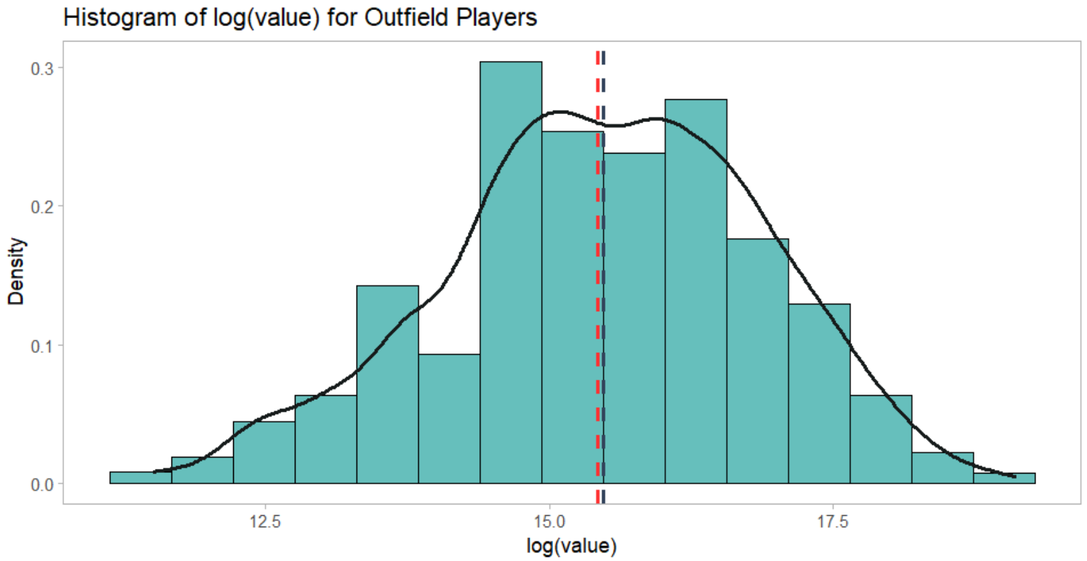
- **Distributions** of **categorical variables**.
- **Density** of logarithm of the response variable **accross the categories.**
- Analysis of variables **for each position**.
- **Relationships** between some **explanatory variables** and the **logarithm of the response variable**, for each position.

Response Variable

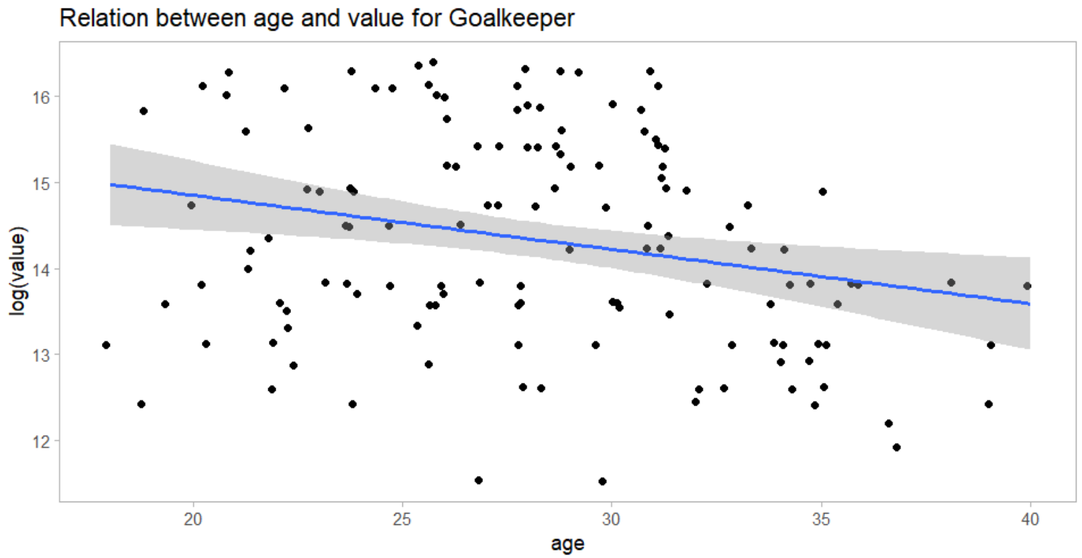


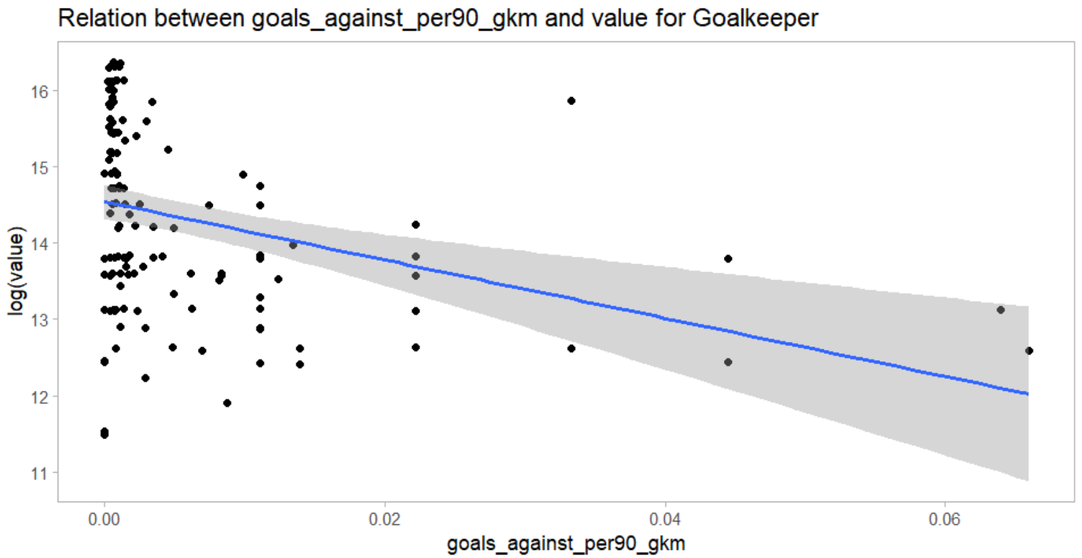


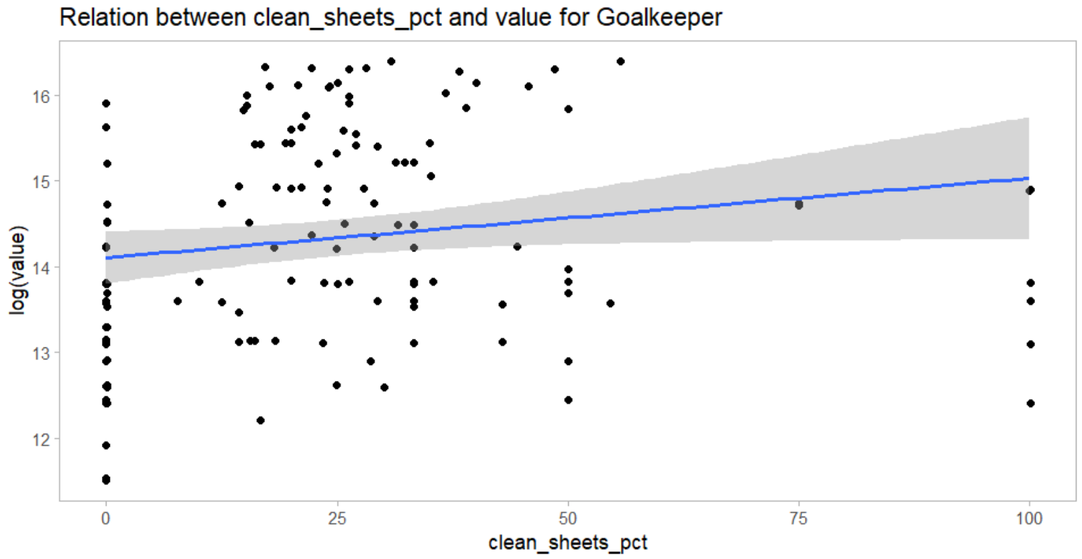


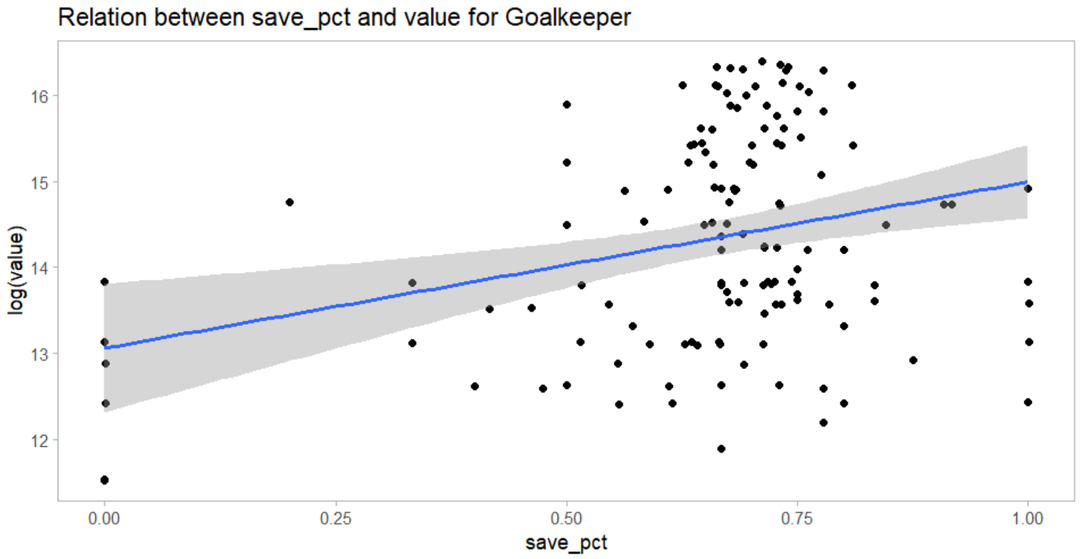


Influence of Predictors for GK

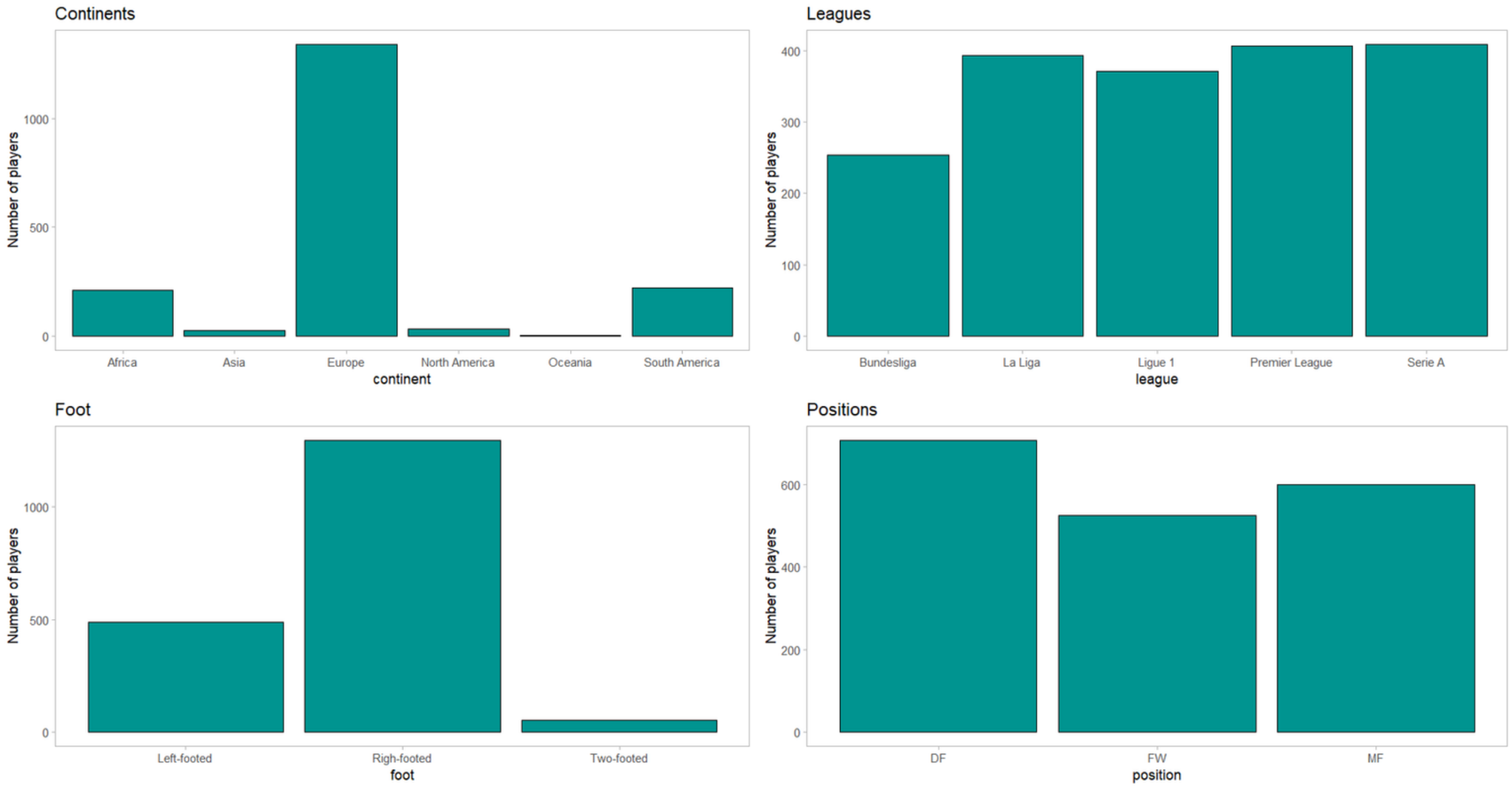




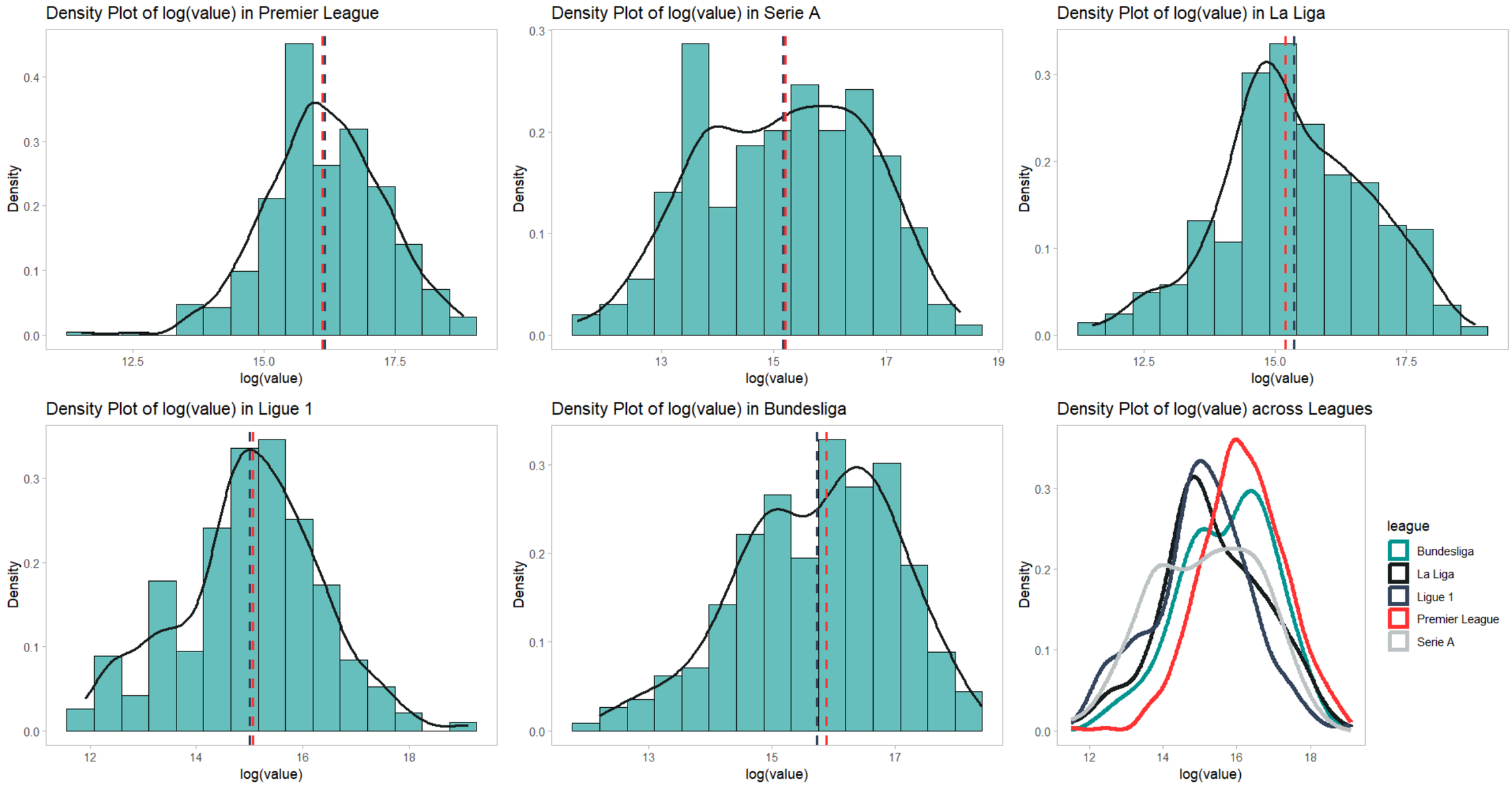




Categorical Variables

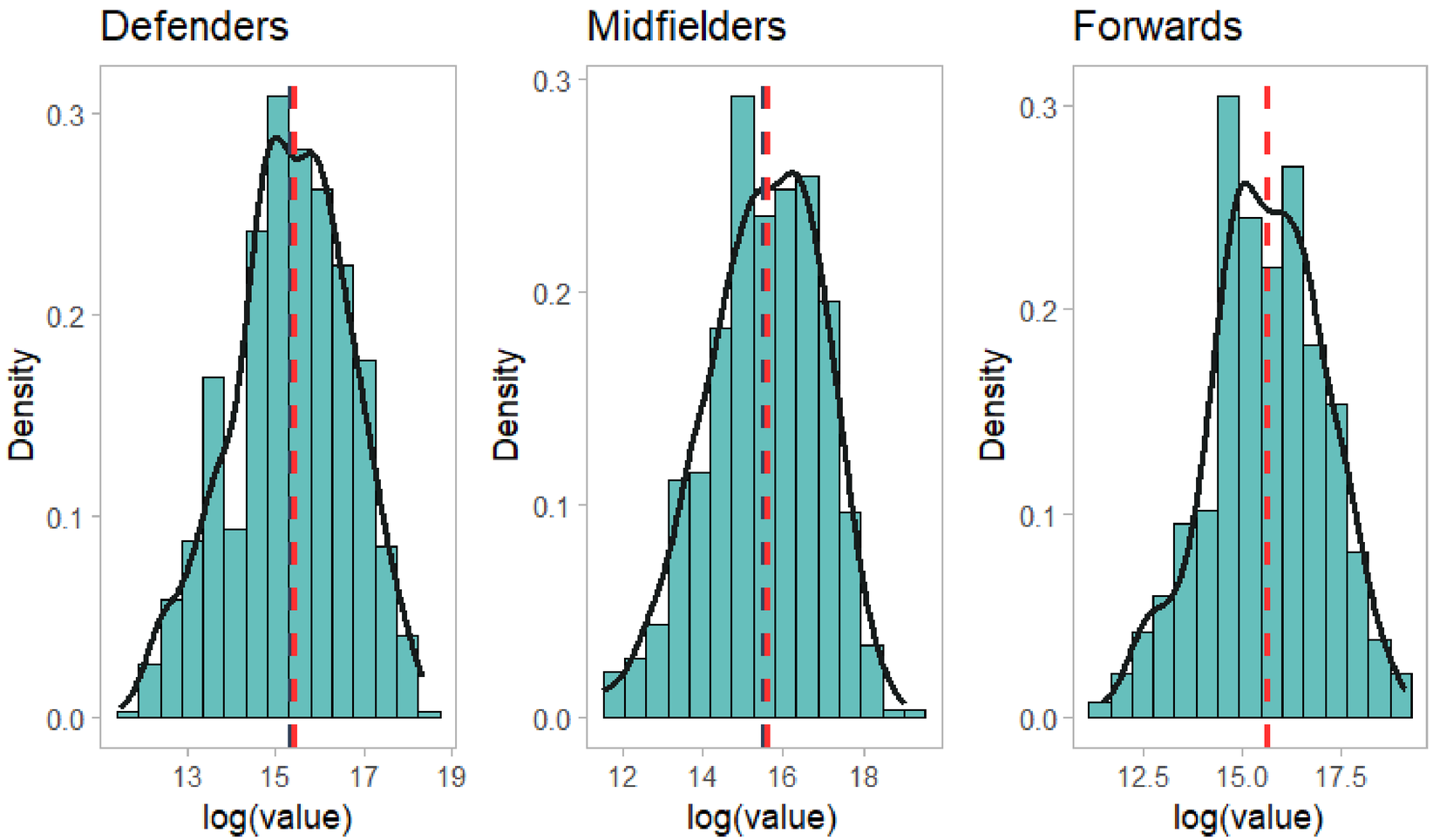


How is the distribution of the logarithm of the value across different leagues?

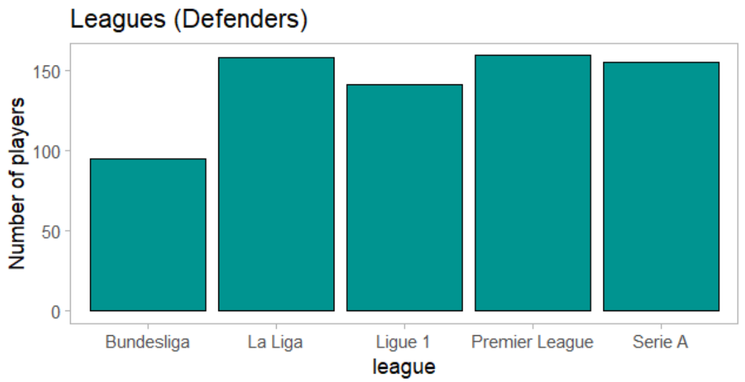


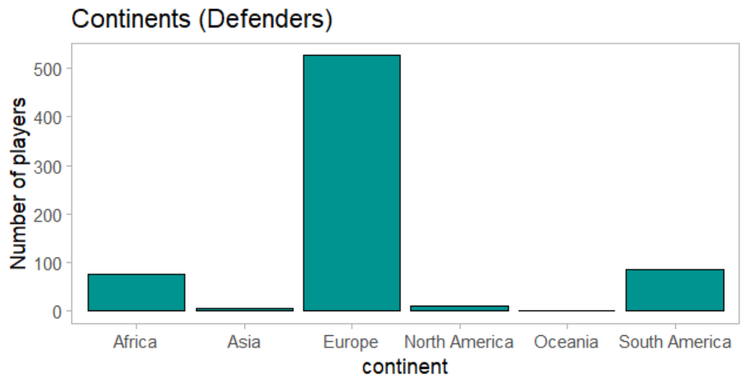
Response Variable

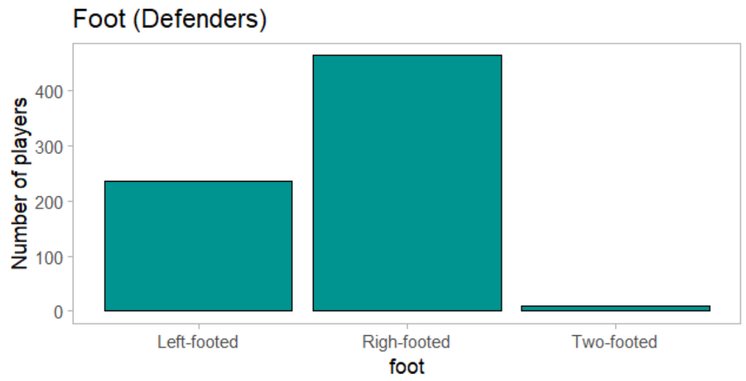
Log(value) in the different positions

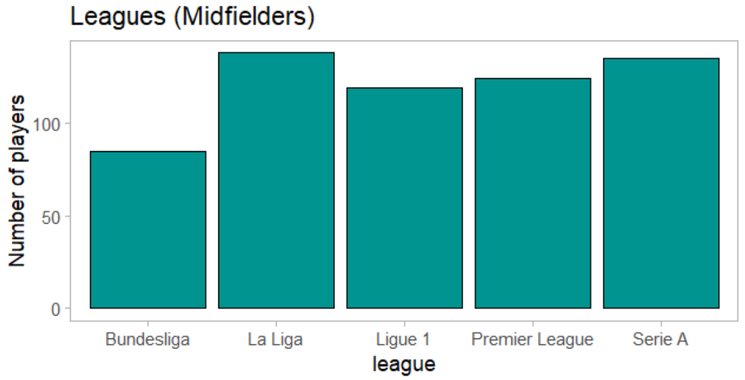


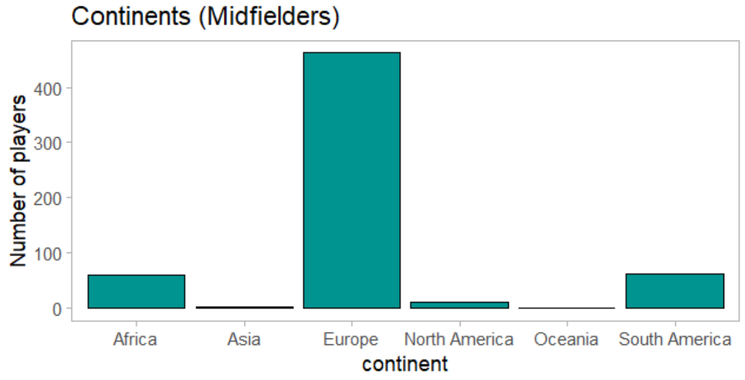
Categorical Variables for the different positions

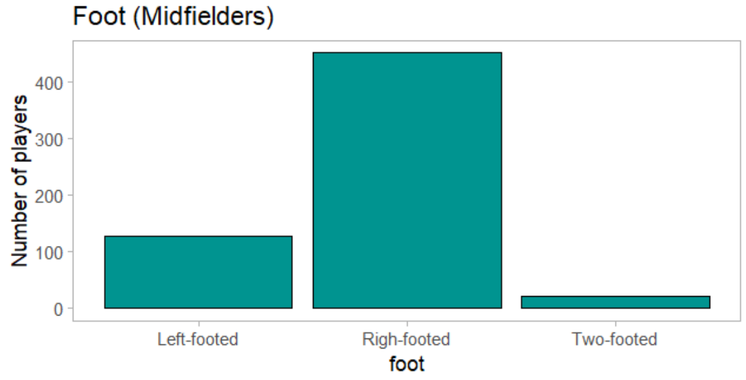


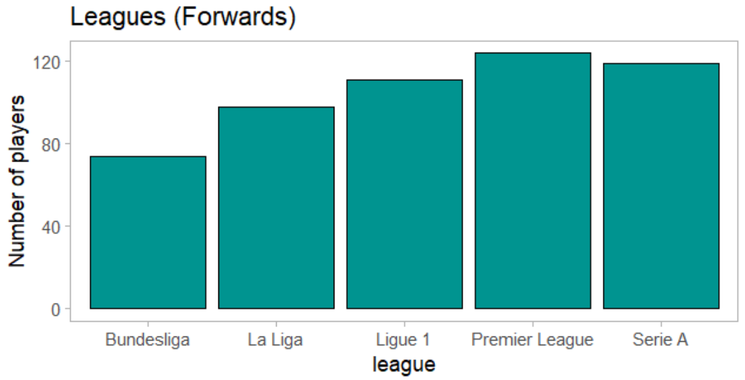


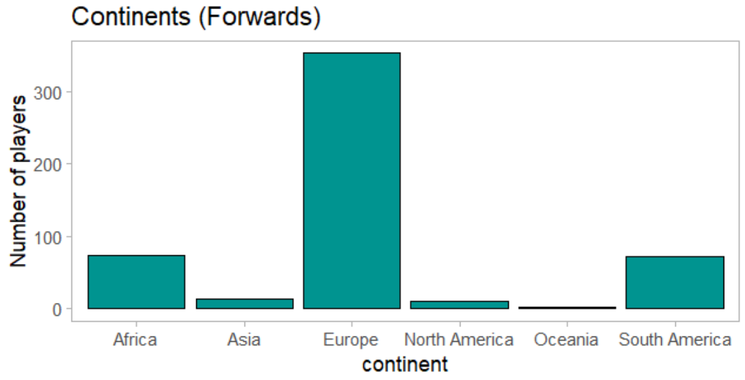


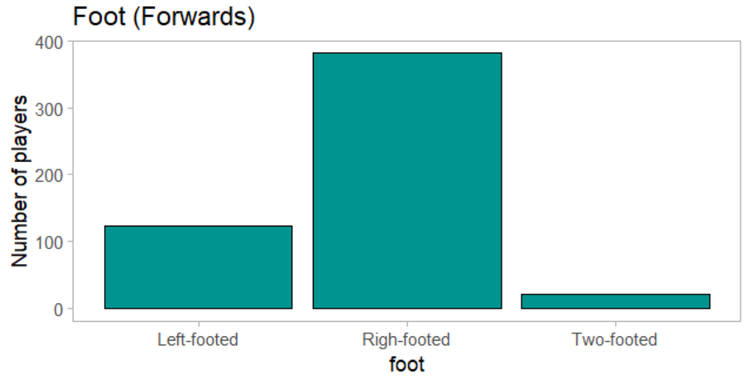




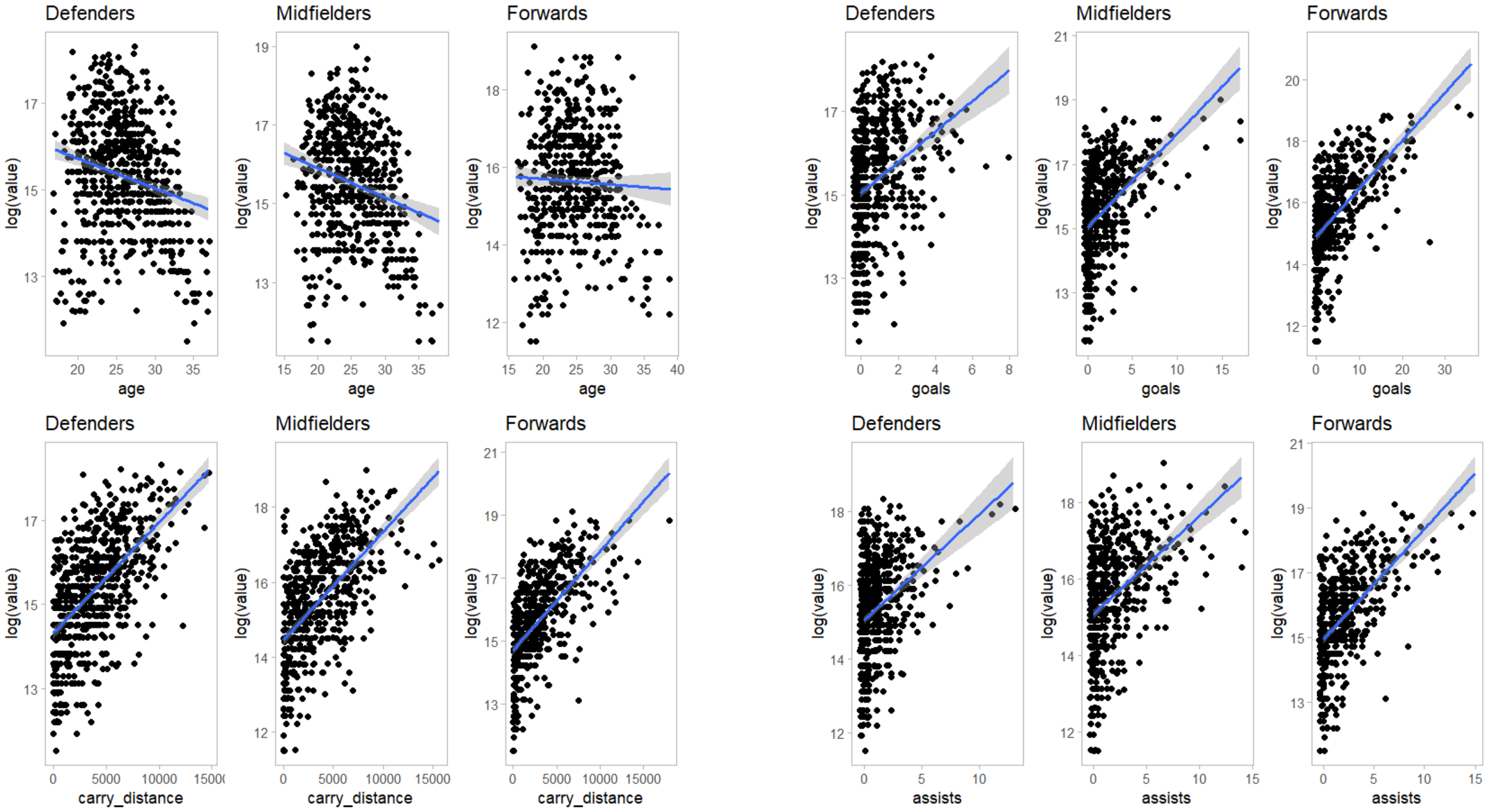






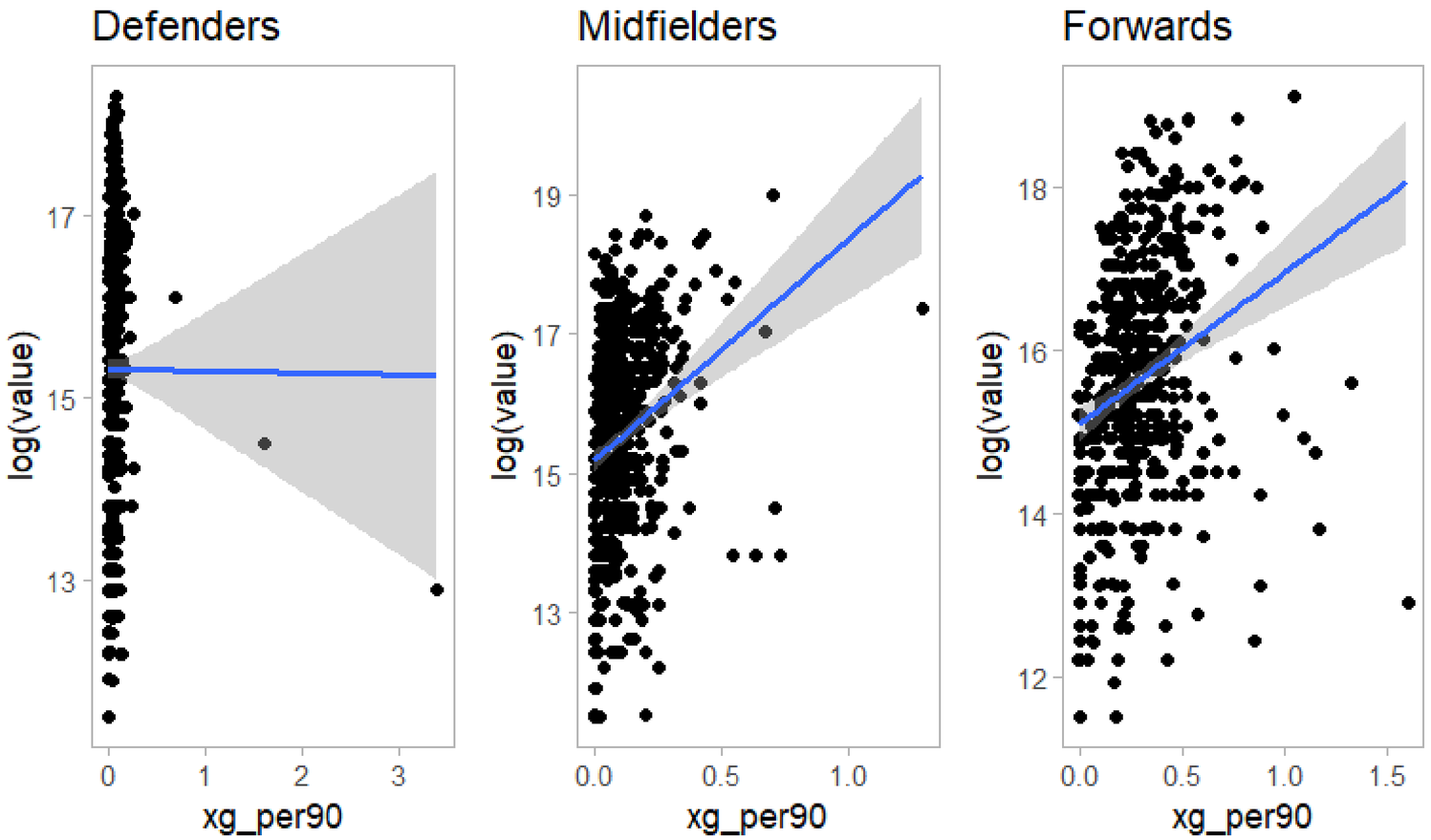


Influence of Predictors in the positions



Expected Goal

These plots are an example of how the effect of a feature on the response variable can vary depending on the position.



Statistical Analysis

Goalkeepers

Data Preprocessing

Exclusion of Players with Missing Data

- Players with **missing** data in the “value” column are deleted.
- **Unnecessary** columns such as “birth year” and “nationality” are removed from both dataframes.

Threshold-Based Discrimination

- The columns where 95% of the values are **zero** concern goalkeepers

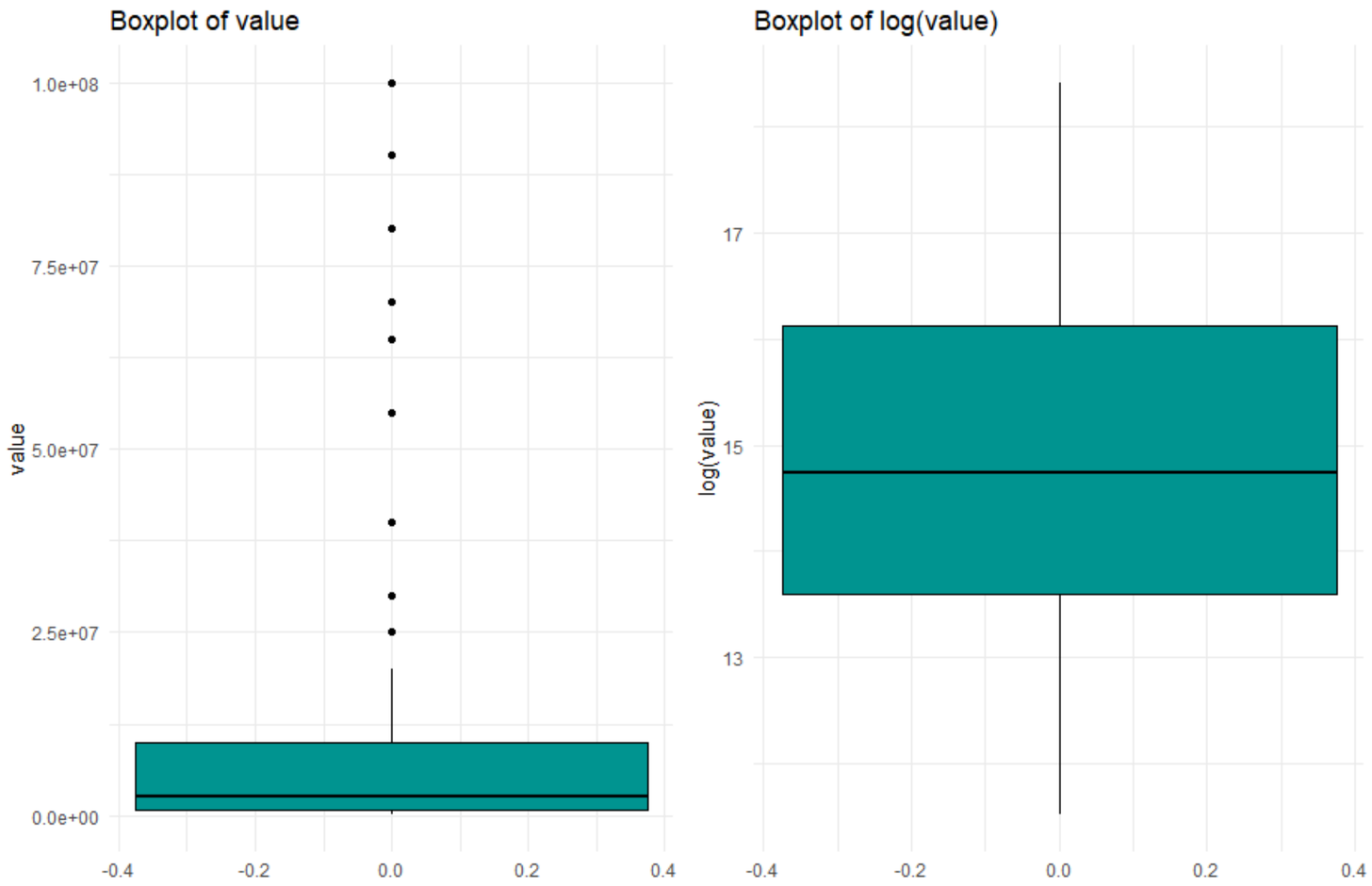
Response Variable Adjustment and outliers

- **Disparity** observed in the values of the response variables, so we corrected it using the natural logarithm, a **monotonic** transformation that maintains the order of the values.
- **Outliers** removed, through a boxplot, based on the Interquartile Range criterion.

Correlation Analysis

- Variables with **zero** standard deviation are removed.
- **Highly** correlated variables are removed for better analysis (correlation > 0.8) .

Boxplot Of Response Variable



Implemented Models

Models:

- Linear Regression
- Ridge Regression with CV
- Lasso Regression with CV
- Grouped Lasso with Lambda.1se
- Grouped Lasso with Lambda.min

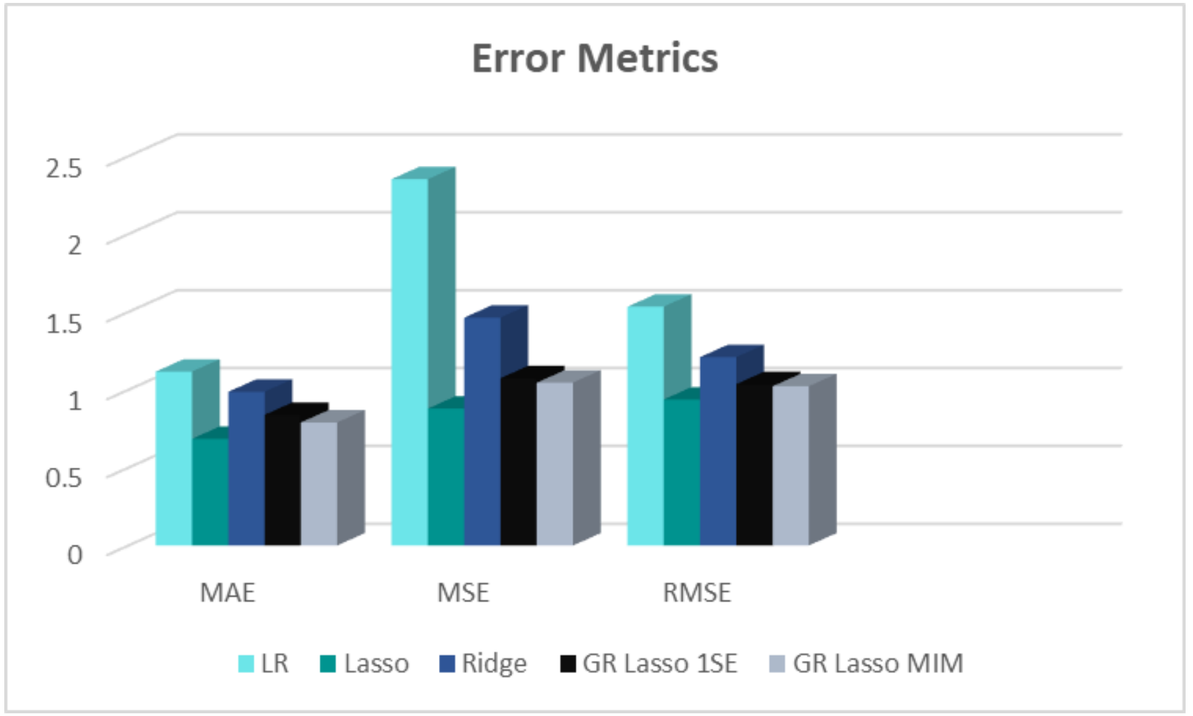
Group Vector:

categorization of goalkeeper features into 4 groups:

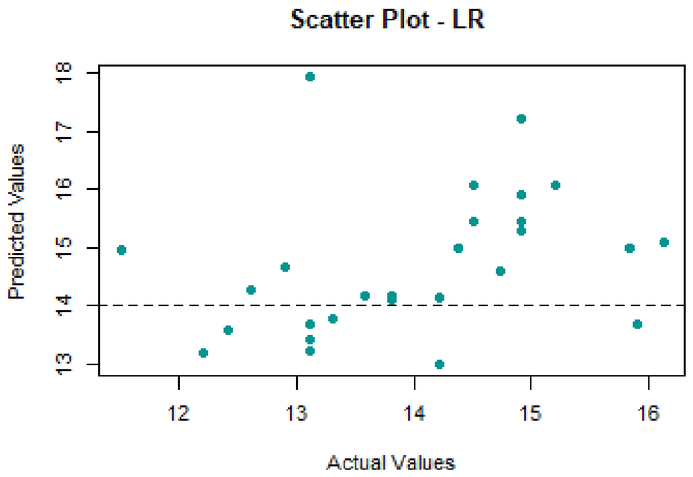
- **Demographic** stats (e.g., “age”, “height”)
- **Performance** stats (e.g., “wins_gk”, “draws_gk”)
- **Performance and Efficient** stats for minutes (e.g., “wins_gkm”, “pens_savedm”)
- **Efficient** stats (e.g., “pens_saved”, “pens_allowed”)

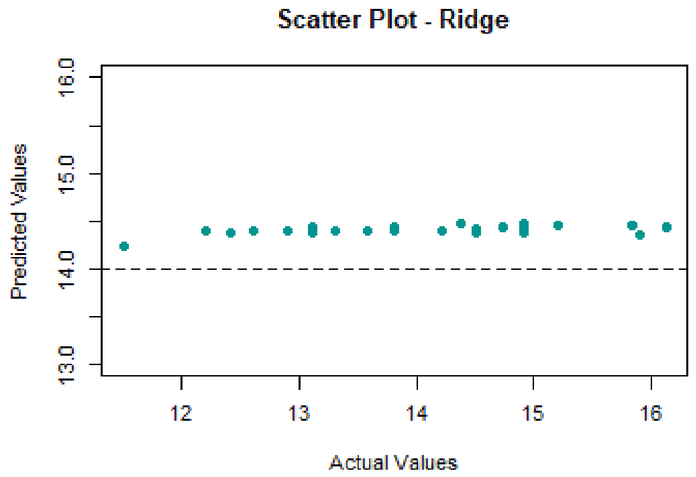
Results Comparison

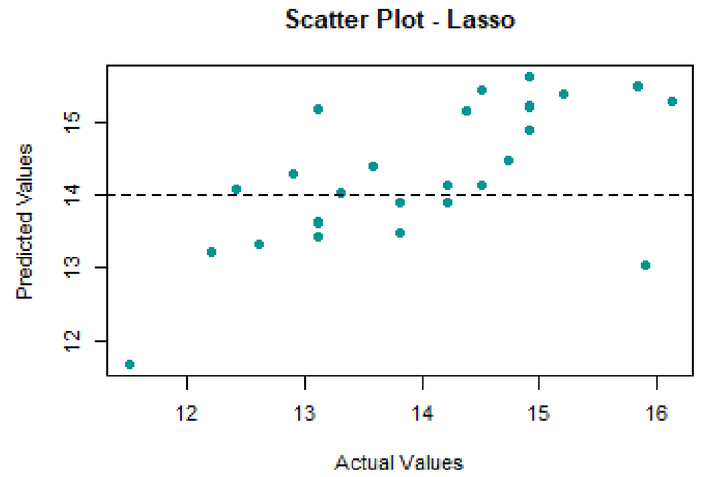
Error Metrics Comparison					
	Linear	Ridge	Lasso	GR Lasso 1 SE	GR Lasso MIN
MAE	1.1166	0.9871	0.6851	0.8474	0.7509
MSE	2.3525	1.4634	0.8789	1.0961	0.9739
RMSE	1.5337	1.2097	0.9375	1.0469	0.9868

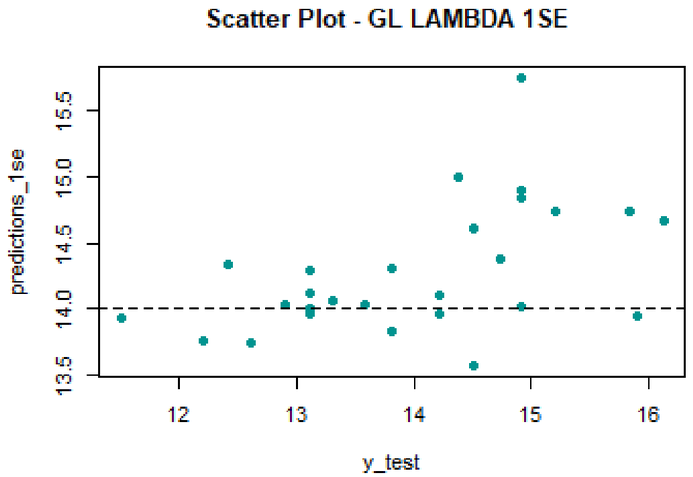


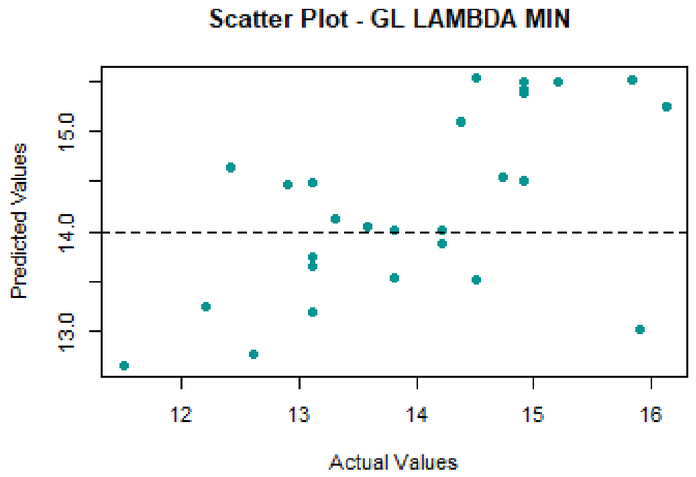
Scatter Plots



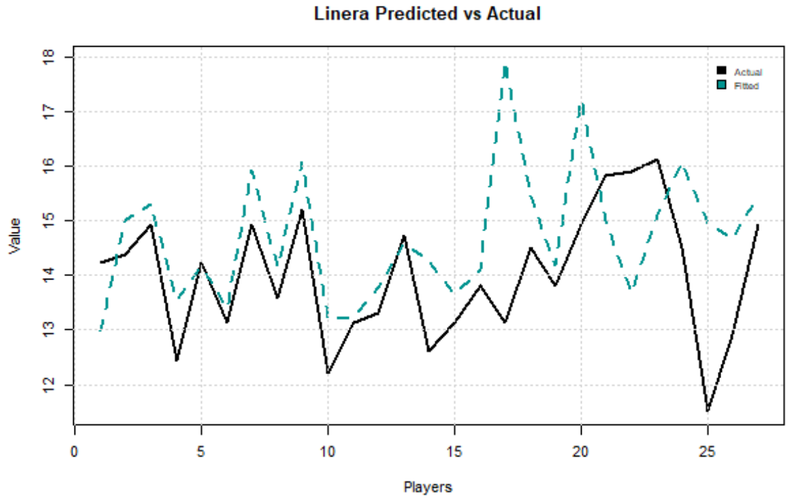


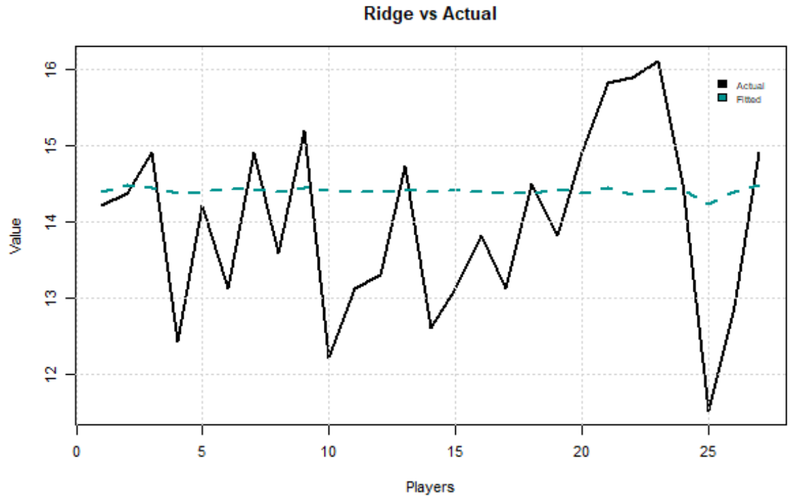


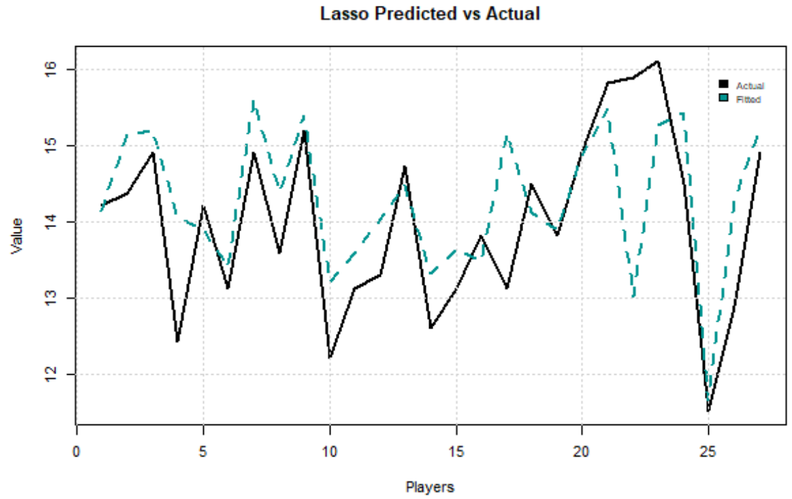


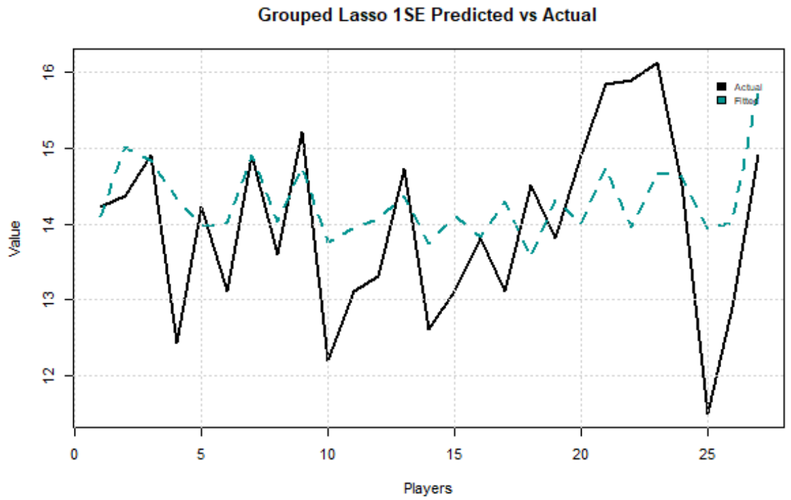


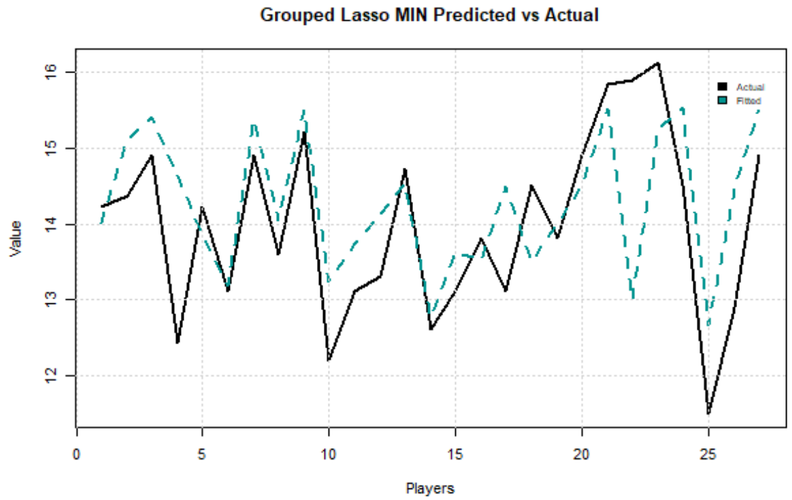
Predicted Vs Actual











Normalized Feature Importance

Positive Impact

Models	Most important Features
Linear Regression	corner_kick_goals_against_gk, psnpxg_per_shot_on_target_againstm, wins_gk, clean_sheets_pct
Ridge Regression CV	pens_savedm, minutes_90s_gkm, games_starts_gkm, pens_missed_gkm
Lasso Regression CV	minutes_90s_gkm, pens_missed_gkm, psxg_net_gkm, losses_gkm
Grouped Lasso 1SE	def_actions_outside_per_area_gk, wins_gk, draws_gk corner_kick_goals_against_gk
Grouped Lasso MIN	psnpxg_per_shot_on_target_against, save_pct, corner_kick_goals_against_gk, pens_missed_gk

The most frequently occurring features are:

- “corner_kick_against_gk”
- “wins_gk”
- “minutes_90s_gkm”
- “pens_missed_gkm”

Normalized Feature Importance

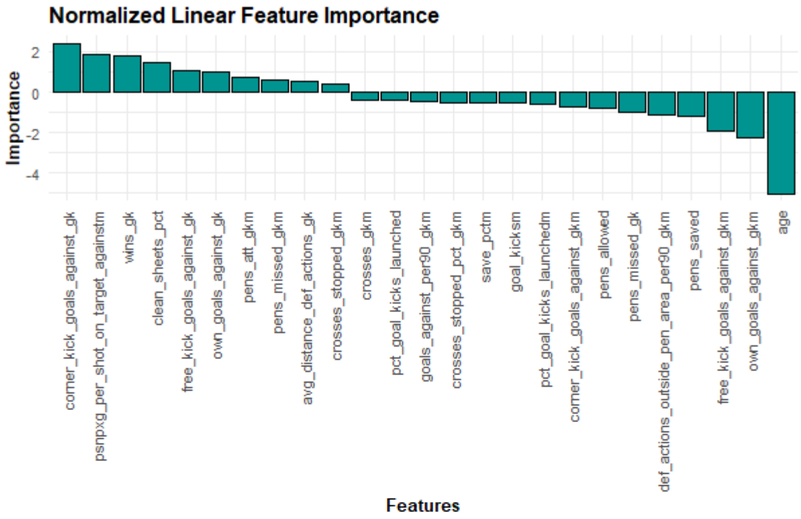
Negative Impact

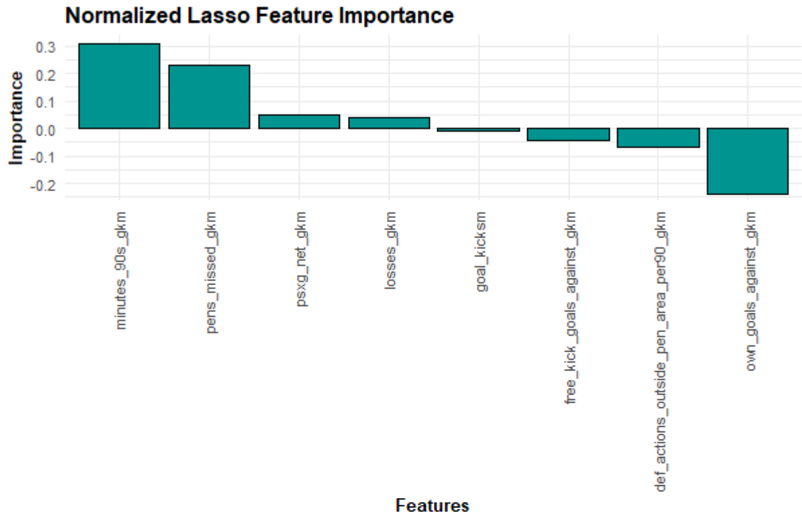
Models	Most important Features
Linear Regression	age, own_goals_against_gkm, free_kick_goals_against_gkm,
Ridge Regression CV	psnpxg_per_shot_on_target_againstm, save_pctm, own_goals_against_gkm
Lasso Regression CV	own_goals_against_gkm, def_actions_outside_pen_area_per90_gkm, free_kick_goals_against_gkm
Grouped Lasso 1SE	age, passses_pct_launched_gk
Grouped Lasso MIN	clean_sheets_pctm, pct_goals_kicks_launchedm, clean_sheetsm

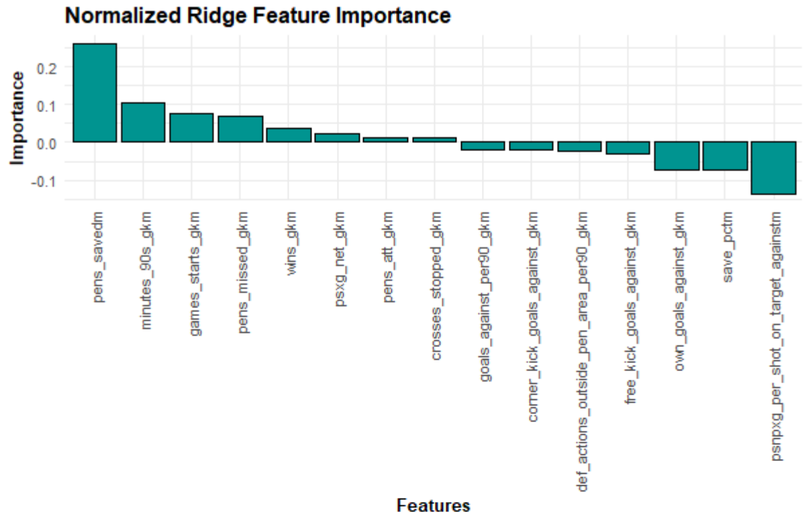
The most frequently occurring features are:

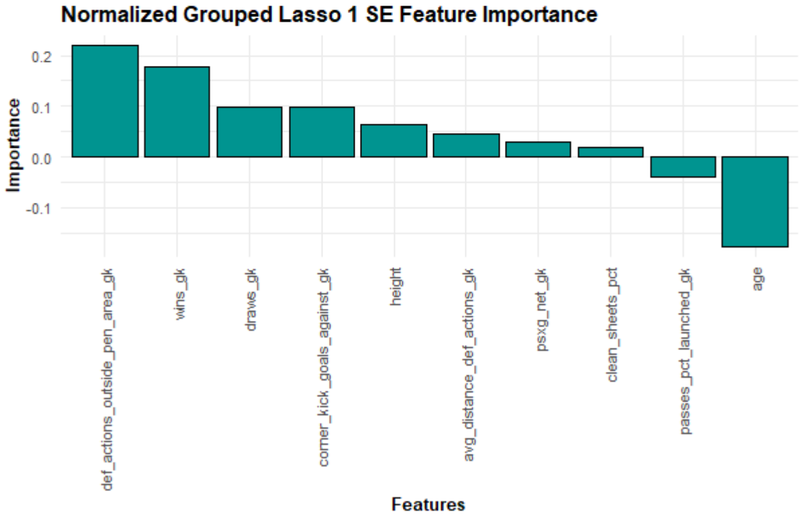
- “own_goals_against_gkm”
- “age”
- “free_kick_goals_against_gkm”

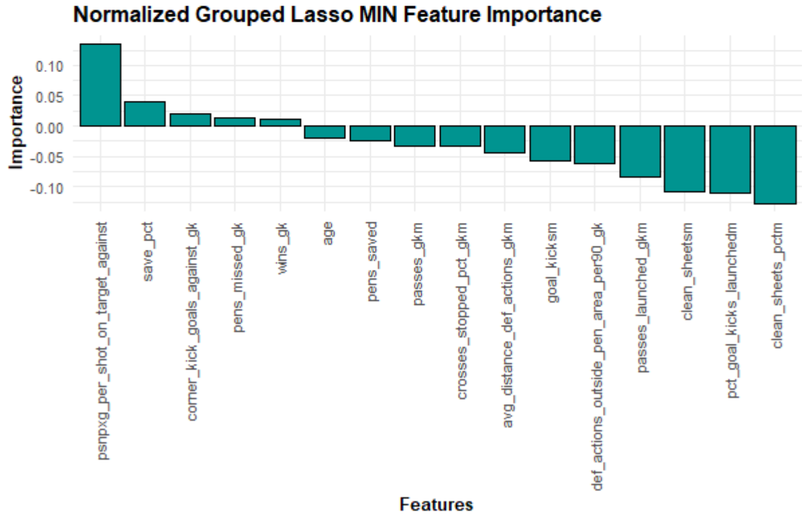
Features Importance











Outfield Players

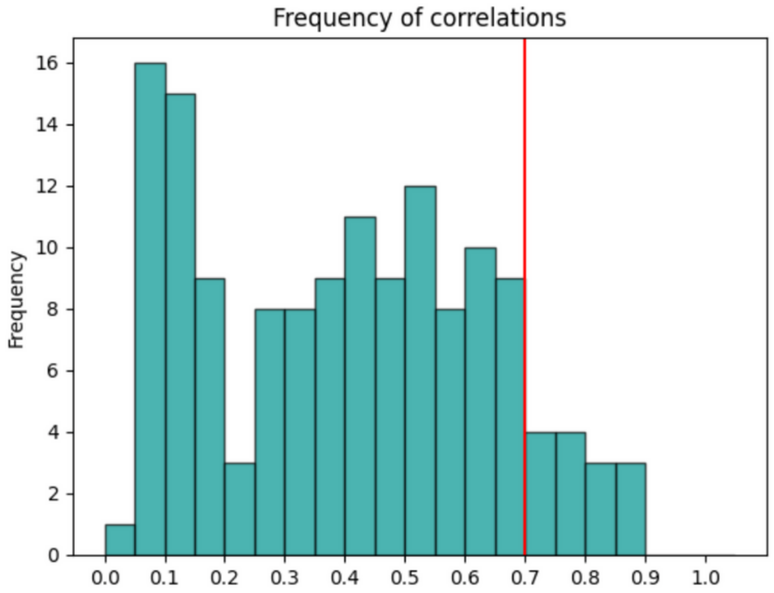
Variable and Variablem

Some variables displayed information about the per-minute statistic

Which were nothing but the amount of the “original” variable divided by the n° of minutes played.

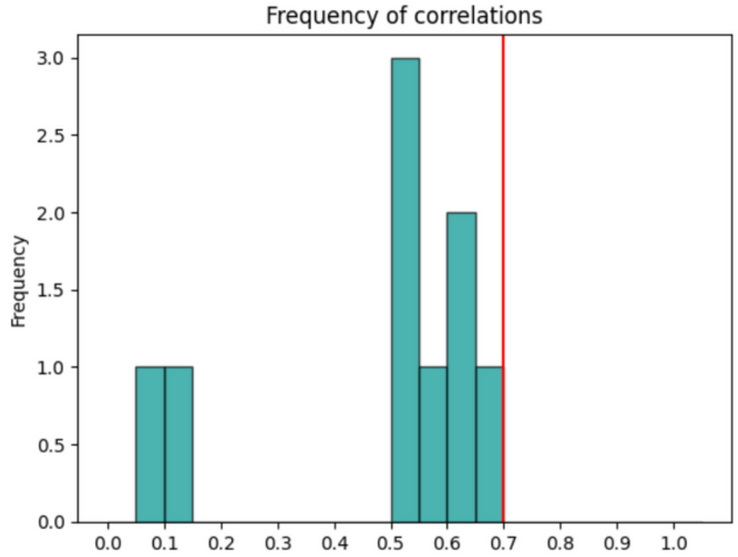
Our solution was to delete the redundant information only if it was highly correalted with the original variable

Treshold: 0.7



Variable and Variable_per90

Same information were displayed _per90



Treshold: 0.7

Variablem and Variable_per90

goals
assists
xg
npxg
xa
shots_total
shots_on_target

Data cleaning and preparation

3 main problems were faced when dealing with correlation issues

Redundant data

Sometimes variables were displaying the same information under different variable names

passes

~~passes_completed~~

passes_pct

Highly correlated var. - 1

Variables showing the same behavioral trend had to be merged together

$$\frac{\text{shots\_on\_target}}{\text{shots}} = \text{shots\_accuracy}$$

Highly correlated var. - 2

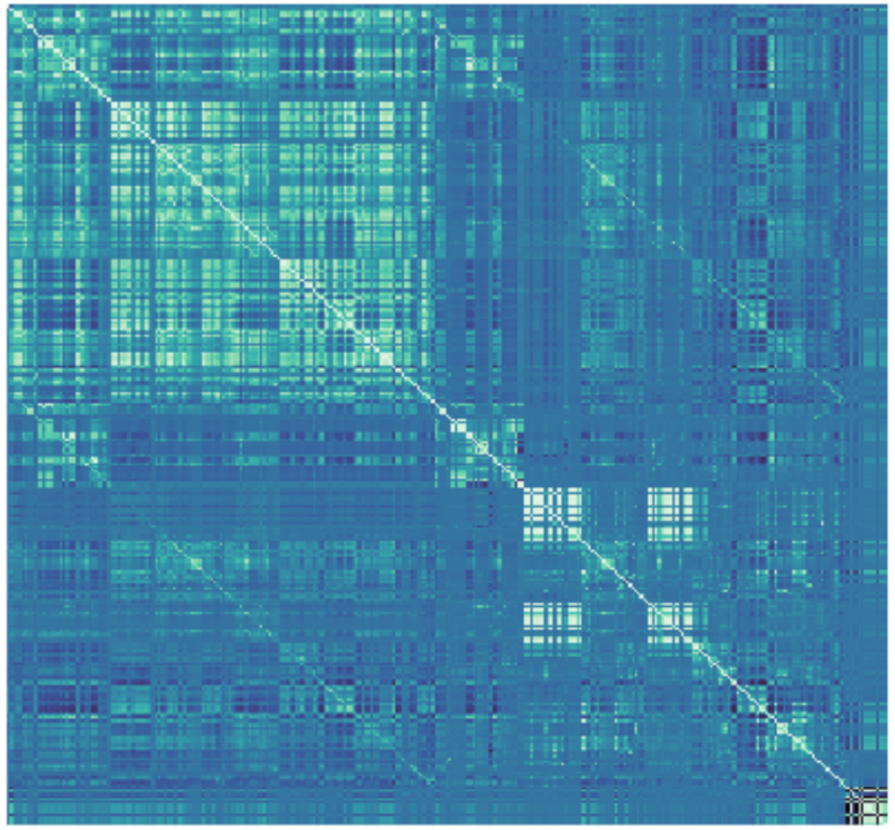
Different aspects of the same variables were displaying all correlation greater than 90%. Merged with a weighted average

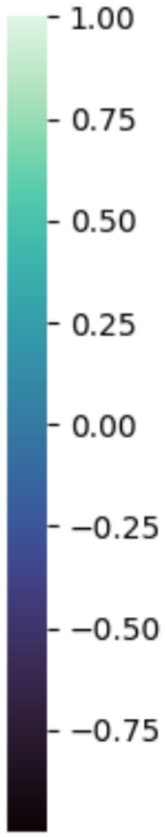
SCA

- Shots
- Sca
- Fouled
- Passes_live

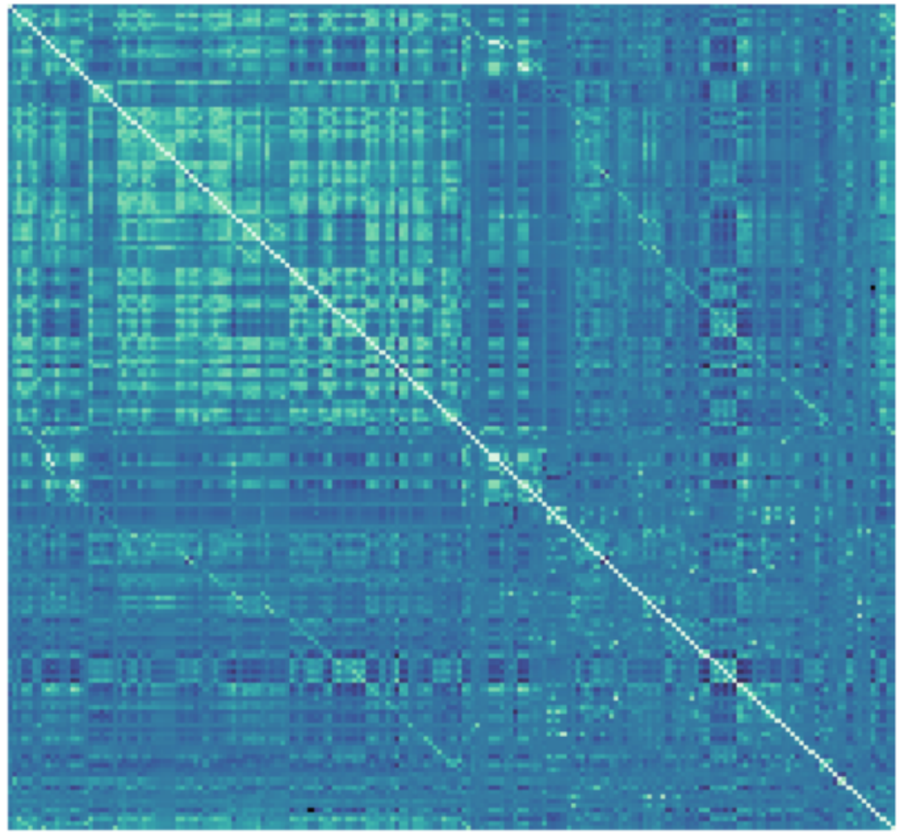
Correlation matrices

Before





After

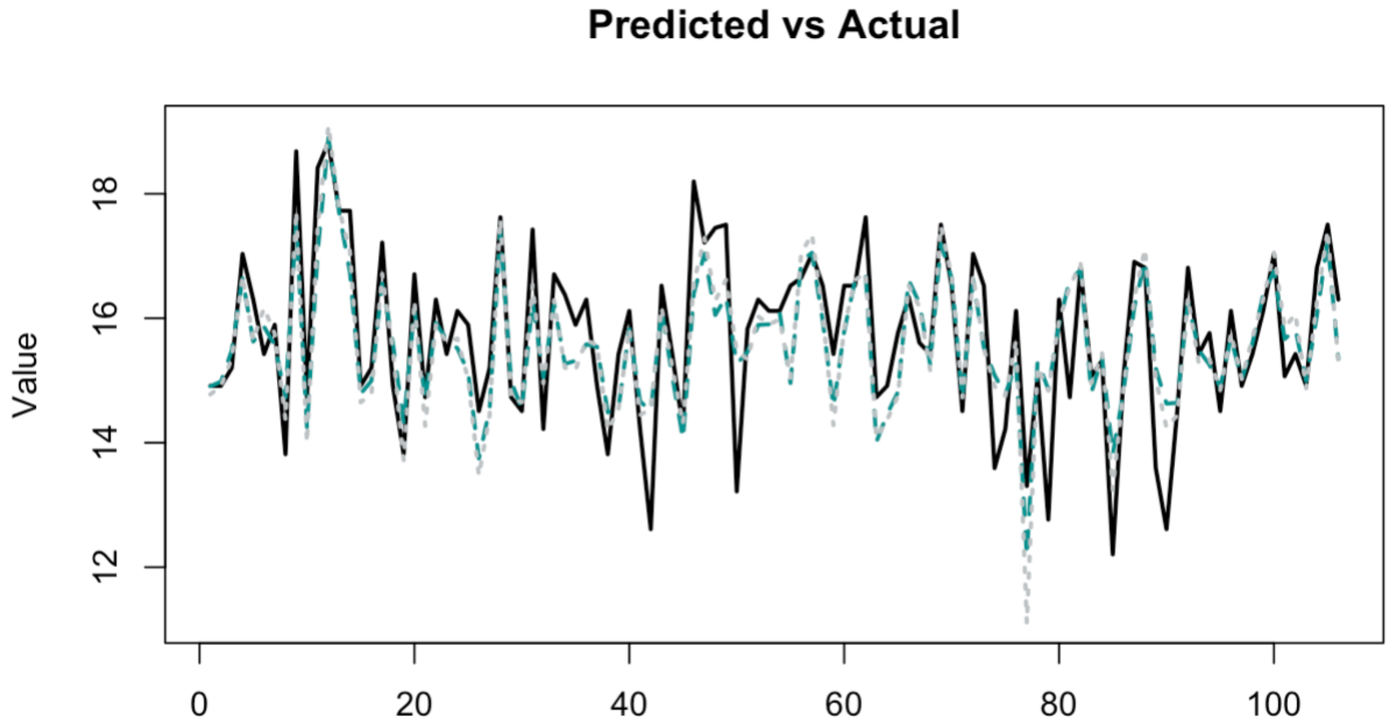


Outfield Players Models

Lasso - Forwards

1) Pts
2) Shots_total
3) Premier_League
4) % games started
5) % shots on target
6-8)Passes
9) Gca - Sca
10) Offsides
11) Goals

1) Age
2) Ligue 1
3) Passes_pctm
4) Blocked shots
5) Clearances



0.0561658
Lambda_min

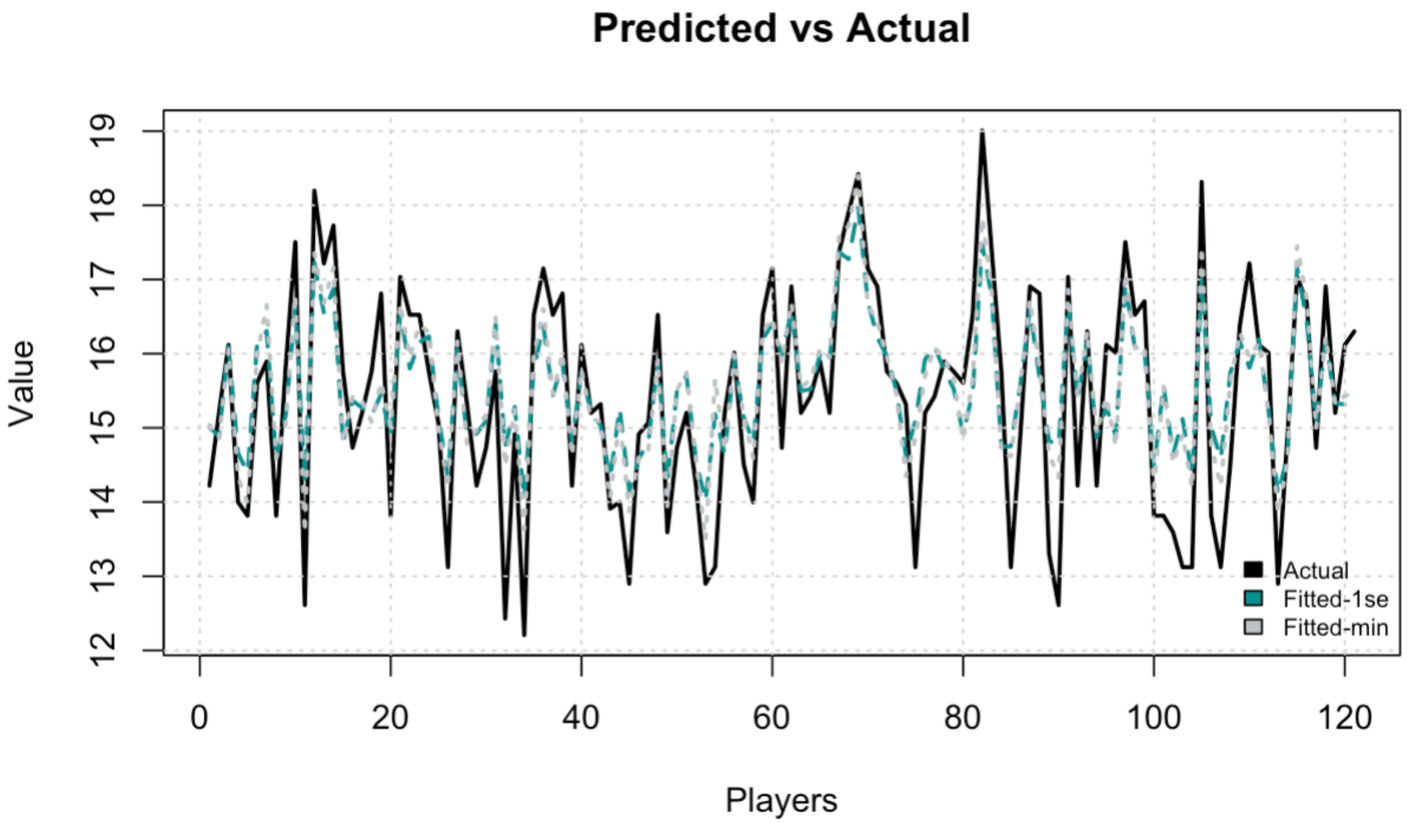
43
Active

0.1182236
Lambda_1se

Lasso - Midfielders

1) Passes
2) Pts
3) Premier league
4) Passes_pct
5) Off. pressure
6) % games started
7) Goalsm

1) Age
2) % Passes received
3) Ligue 1
4) Pressure Mid.
5) Pressure Dif.



0.1059026
Lambda_min

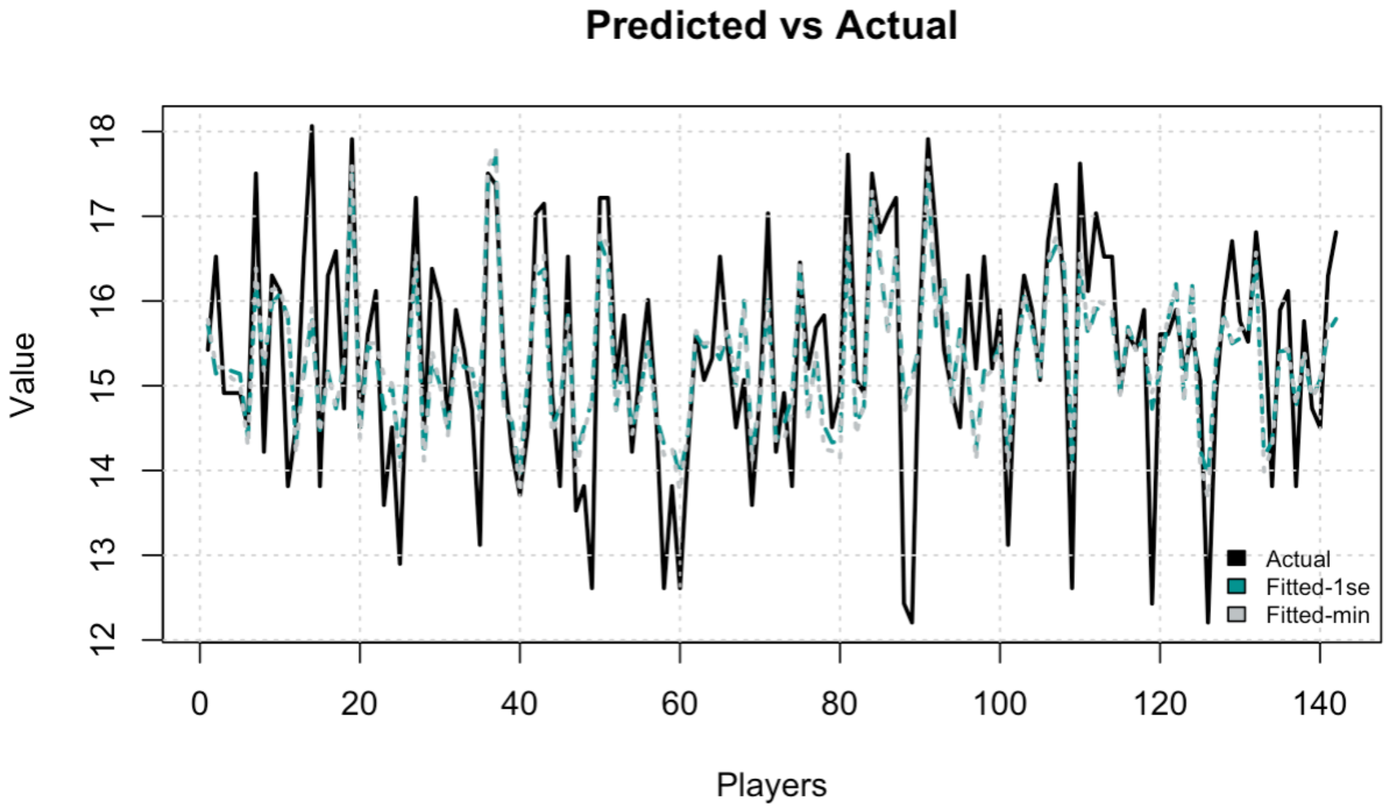
28
Active

0.168627
Lambda_1se

Lasso - Defenders

1) Passes
2) Pts
3) Premier league
4) Ball recoveries
5) CL
6) Carry distance
7) Aerials won

1) Age
2) % dribbled
3) Ligue 1



0.1083917
Lambda_min

21
Active

0.1432875
Lambda_1se

23 Groups

Were created for the grouped lasso
Which can on turn be divided in **4** groups

Forwards

Penalties = 1,
Goals = 2,
Expected = 4,
Sca = 5,
Gca = 6,
Shots = 7

Midfielders

Passes = 3,
Touches = 8,
Crosses = 10,
Dribble =11,
Distance = 12,
Corners =17,
Assists =18,

Defenders

Tackles = 9,
Aerials =14,
Pressure =15,
Cards =16,
Fouls = 21,
General difensive = 22

General

Demographic = 13,
Playtime = 19,
Team = 20,
Negative aspects = 23

Different role = Different groups

Even if some of them are shared

Forwads

- **Pens**
- Expected
- **Carries**
- Distance
- **Corners**
- Fouls

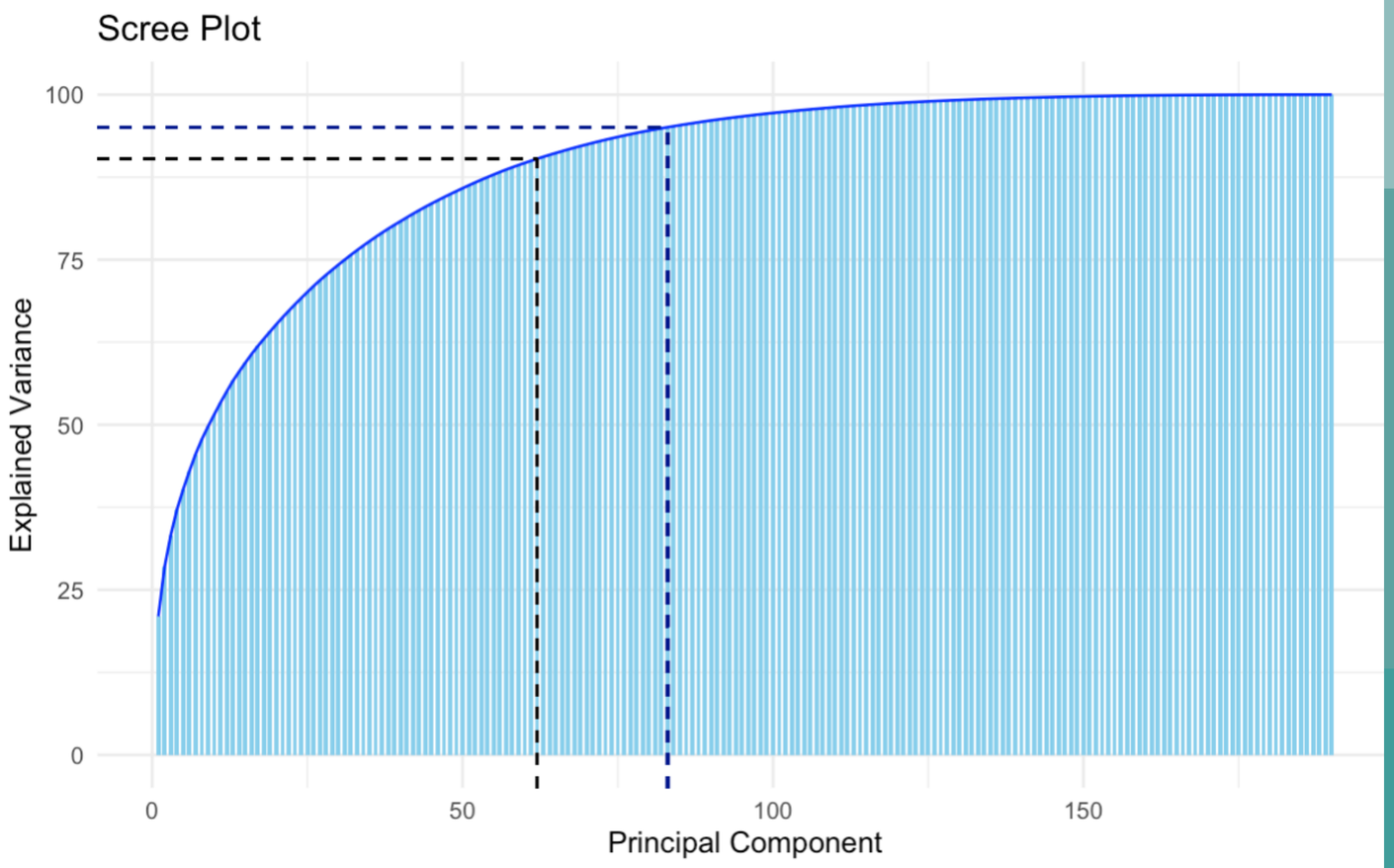
Midfielders

- **Touches**
- Corners
- **gca**

Defenders

- Tackles
- Expected
- Shots
- Corners
- Cards
- sca-gca

PCA plot



Forwards

90% variance: **62** components
95% variance: **83** components

Midfielders

90% variance: **58** components
95% variance: **80** components

Defenders

90% variance: **59** components
95% variance: **79** components

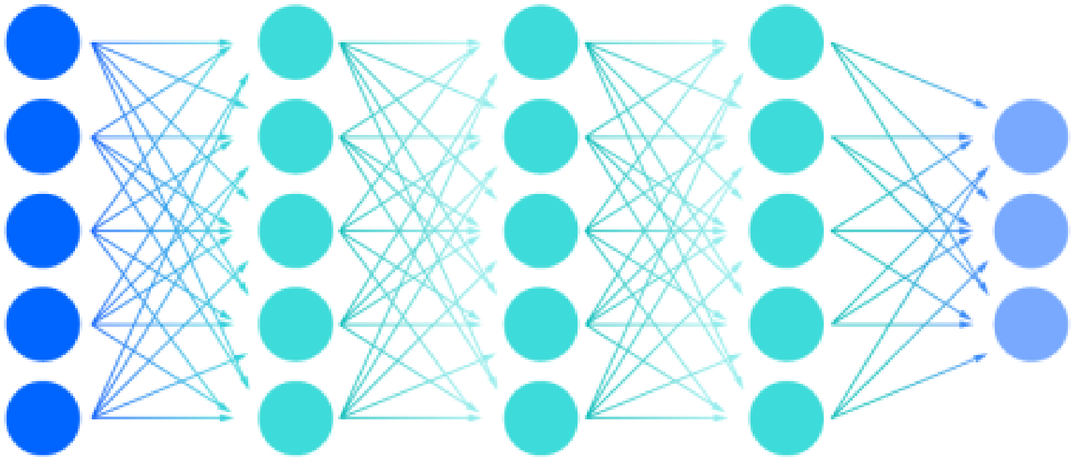
Neural Network

Architecture

- **188** neuros **input** layer
- **64** in the **1st hidden** layer
- **32** in the **2nd hidden** layer
- **1** neuron **output** layer

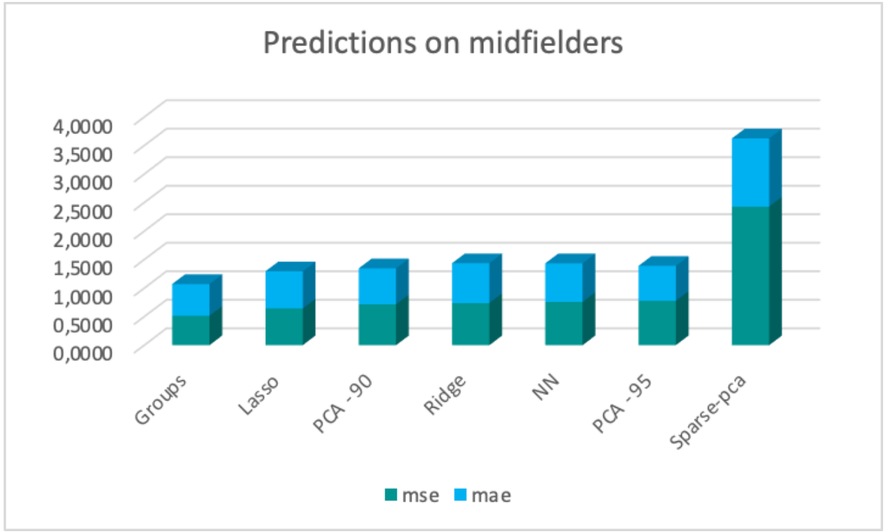
Hyperparameter

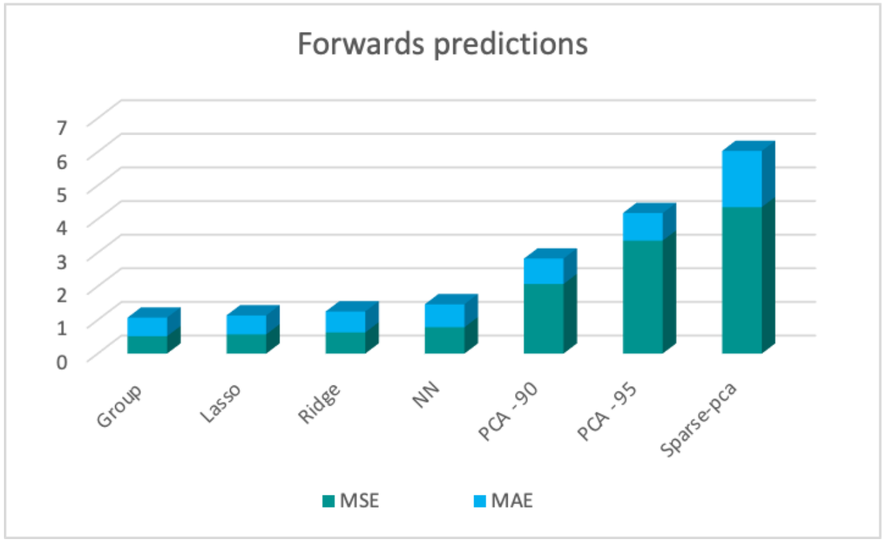
- **ReLu** activation function
- **0.3 Dropout rate**
- **0.01** lambda in **L2 regularization**
- **1e-2 learning rate**

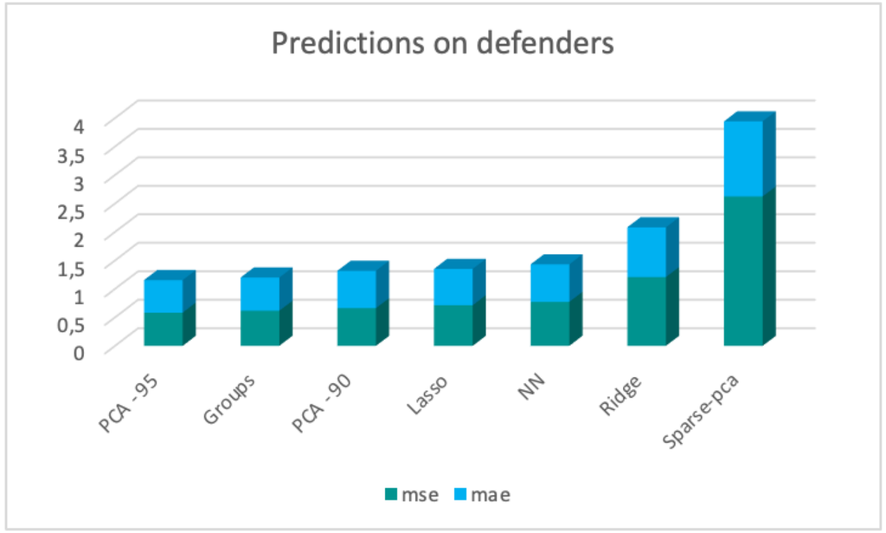


Results

Plotting the main 2 metrics on all 3 roles.







Thanks for the attention

Any questions?