6950_EDA

Ruoyuan Qian, Andrew Shan

Overview and the goal

In this project, we aim to explore the contributing factors of market values for the players in the five most successful football leagues in Europe (i.e., "Big Five") in the setting of FIFA 21. FIFA 21, as part of the FIFA series, is an association football simulation video game, which is developed and released annually by Electronic Arts under the EA Sports label. It is the 28th installment in the FIFA series, and was released on 9 October 2020 for Microsoft Windows, Nintendo Switch, PlayStation 4 and Xbox One. Enhanced versions for the PlayStation 5 and Xbox Series X and Series S were released on 3 December 2020, in addition to a version for Stadia in March 2021. With an official license from FIFA, the world governing body of football, the game provides comprehensive data featuring more than 30 official leagues, over 700 clubs, and over 17,000 players.

The 'Big Five' represent the five most successful football leagues in Europe, which are made up of the Premier League in England, La Liga in Spain, the Bundesliga in Germany, Serie A in Italy and Ligue 1 in France. Since 1955 The first edition of the European Cup took place during the 1955–56 season, which provides an unique competition opportunity for footballs clubs in Europe. In 1960, The association coefficient was introduced to rank the football associations of Europe, and thus determine the number of clubs from an association that will participate in the UEFA Champions League, the UEFA Europa League and the UEFA Europa Conference League. Since then, the Premier League, La Liga, the Bundesliga, and Serie A have won the majority of the titles, which was more than titles won by other associations combined. Though Ligue 1 in France has never won the title, it has been ranked among the top 5 eer sicne. The combined revenue of these five leagues, which each represent the highest tier football division in their countries, has more than doubled in the past decade, reaching a total of approximately 15.1 billion euros in 2019/20. As such, the 'Big Five' is consistently attracting talent players to joint the leagues.

In this study, we aim to examine the contributing factors of market values for the players, which could potentially provide some insights of the key strategy in ability training for players to increase their market value. In addition, the analyses could examine the disparities in the market value across player's position and the optimal league they should join to maximize their market value.

Exploratory Data Analysis (EDA)

Data cleaning

The data set "players_21.csv" contains 106 variables and 18,944 observations. Since our goal is to analyze contribution factors of players value in "Big Five", we only keep the players in "Spain Primera Division", "Italian Serie A", "German 1. Bundesliga", "French Ligue 1" and "English Premier League". What's more, goal keepers are removed due to the different attributes between them and other players. Finally, 1 player with NA ratings and 2 players with zero overall scores are removed.

As for variables, 53 unrelated variables are dropped such as "player name", "nationality", "player tags". In addition, 40 variables were removed because they are the more specific classifications of some attributes, which are highly correlated to one another. Thus, we only keep one overall rate for per attribute.

After that, 2,756 observations and 16 variables (4 categorical variables and 12 continuous variables) are left demonstrating the demographic information (sofifa_id, name, age, league_name, weight, height), individual

attribute ratings (overall, pace, shooting, passing, dribbling, defending, physic), team position and the value in Euro.

Data manipulation

There are 28 levels in "team position" variable, to make it more interpretable in the model, we group them into 3 categories: Defender, Midfielder, Forward.

From empirical knowledge, height and weight are always related to each other, in order to avoid collinearity, we create BMI to combine their information into one variable.

According to the density plot (Figure 1) of the response (value in Euro), the original data is heavily right skewed, while the distribution after log-transformation is more likely to Normal. Thus, log transformed value in Euro is our reponse.

The correlation plot (Figure 2) is made to give us a general idea how continous predictors are related to each other. Shooting, passing, and dribbling are related and they might reflact the simlar aspects of player.

From the plot matrix (Figure 3), the ditribtions of all continuous variables are roughly Normal. The overall scores and log-value in Euro are highly correlated (0.972), suggesting that overall scores is a very important predictors. Besides, the correlation between passing and dribblingis over 0.8, indicating that there might be some potential collinearity.

From the box plots of catergorical variables(Figure 4), the distributions for preferred foot and league name are quite symmetric with similar variance, while the variance for forward level in team position is a slightly larger than others.

```
## c("sofifa_id", "short_name", "age", "league_name", "overall",
## "value_eur", "preferred_foot", "team_position", "pace", "shooting",
## "passing", "dribbling", "defending", "physic", "BMI", "log_value_eur"
## )
```

Variable	Description	Class	Min	Max
sofifa_id	ID	integer	20801	258946
short_name	Short Name	factor	–	–
age	Age	integer	16	38
league_name	League Name	factor	–	–
overall	Overall	integer	49	93
value_eur	Market Value in Euro	integer	45000	105500000
preferred_foot	Preferred Foot	factor	–	–
team_position	Team Position	factor	–	–
pace	Pace Score	integer	30	96
shooting	Shooting Score	integer	16	93
passing	Passing Score	integer	29	93
dribbling	Dribbling Score	integer	28	95
defending	Defending Score	integer	16	91
physic	Physic Score	integer	28	91
BMI	BMI	numeric	0.03425347	0.05520667
log_value_eur	Log(Market Value in Euro)	numeric	10.71442	18.47422

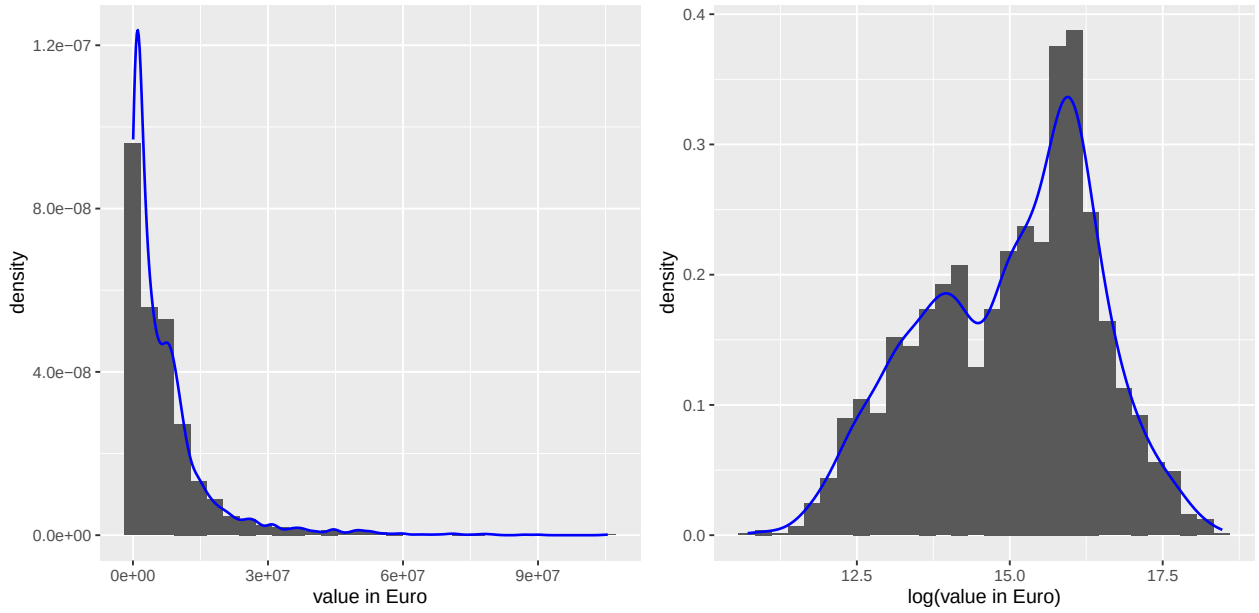

Figure 1: Density plots of response

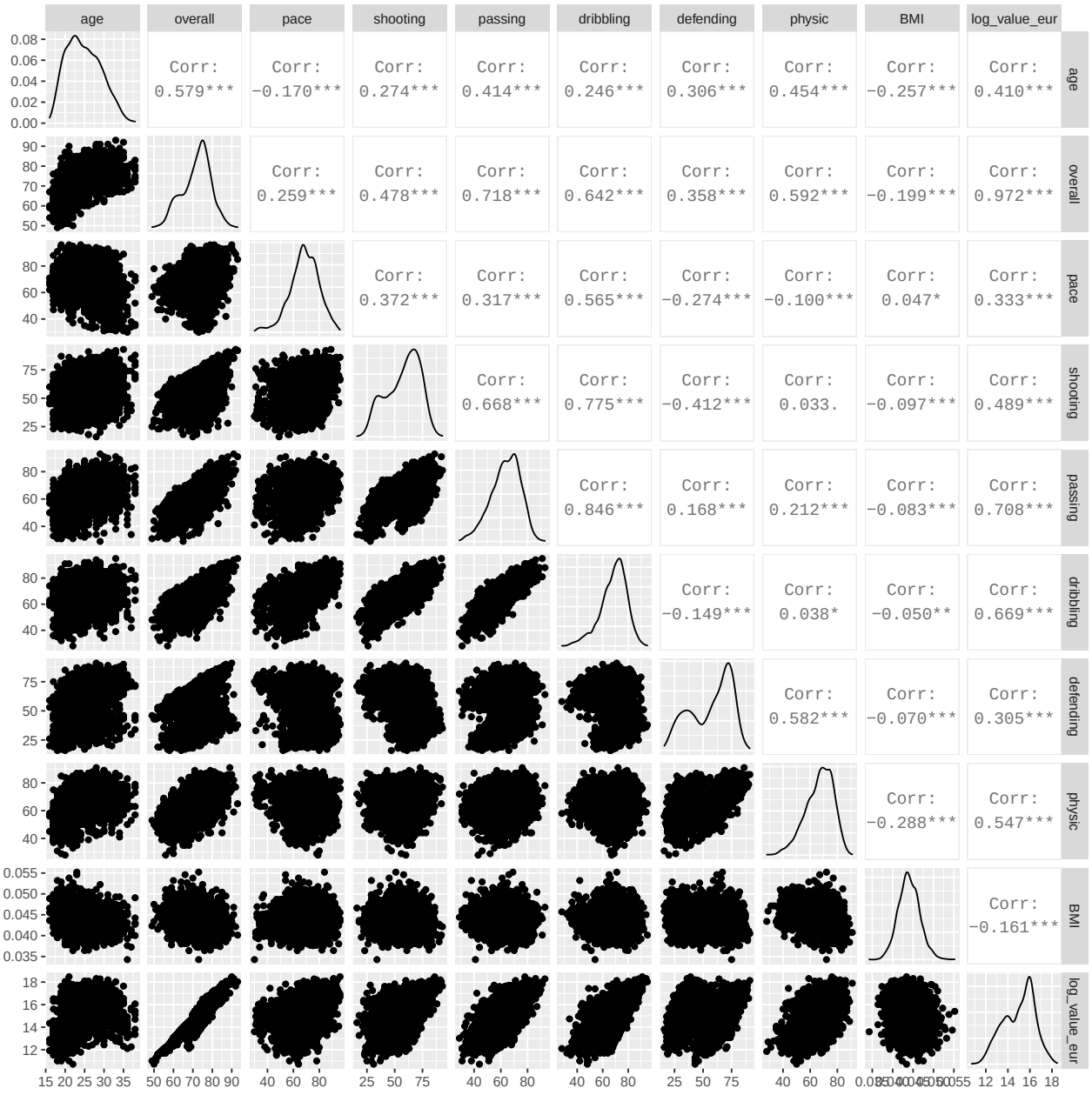

Figure 2: plot matrix of numeric variables

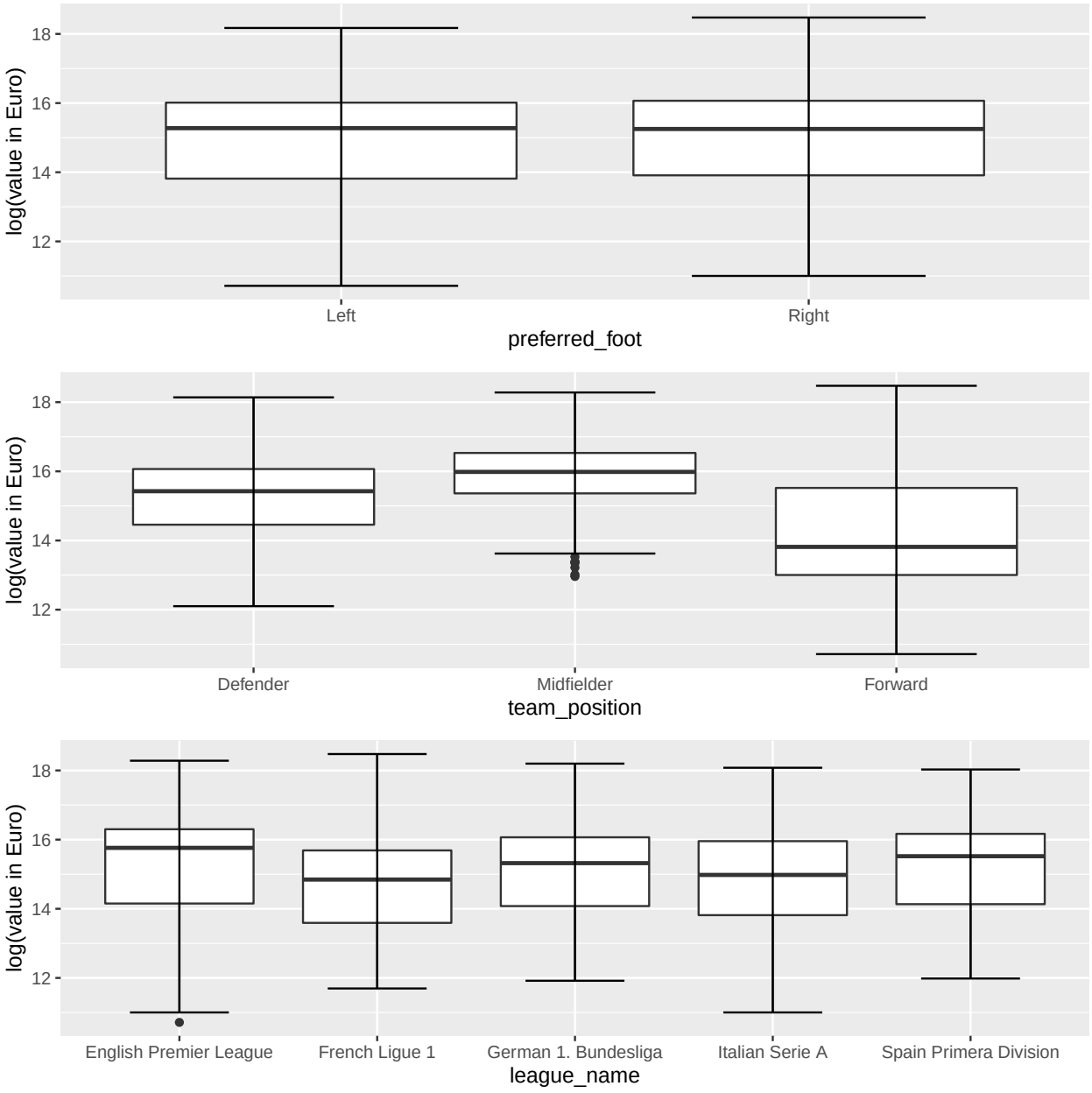

Figure 3: Box plots of catergorical variables