Football Manager 2017

Introduction

This is the report of demo version of the second project for the course Introduction to Data Science. The goal is to analyze data available on https://www.kaggle.com/ajinkyablaze/football-manager-data.

Data

Data is taken from the Football Manager 2017 game. It contains information about 159541 footballers.

Data is gathered in a data frame. There is a row for each player and 89 columns describing them. The columns start with player's ID, name, his nation's ID, date of birth and age. Then we have 4 columns that show numbers of player's caps and goals on international and U21 games. Next two columns represent height and weight. There are 60588 players that have weight equal to 0, so I omitted those when calculating and visualizing anything that includes weight. After that we have 62 columns, each for an attribute that represents skill and determines player's performance. They are rated from 1 to 20, 1 being very poor and 20 being outstanding. Next column is a description of position, a combination of abbreviations for multiple positions of a player. The remaining 15 columns represent 15 positions and their values tell us how good the player is on each of them. Again, the values are numerical and span from 1 to 20.

Analysis

Height and weight

First we will look at a 2D density plot of weight and height. If we plot all the players together, we can't see anything valuable. We will rather look at it for separate positions. Since one player can play on different positions, I chose that the main position of the player is the one that has value 20. If one player has value 20 for two positions, his weight and height are added to data of both positions. This way we get 169927 rows. On Figure 1 we can see the densities of goalkeepers and 3 kinds of central players: defenders, attackers and midfielders. We can observe that attackers and midfielders have similar weight and height, but defenders and goalkeepers are bigger and heavier.

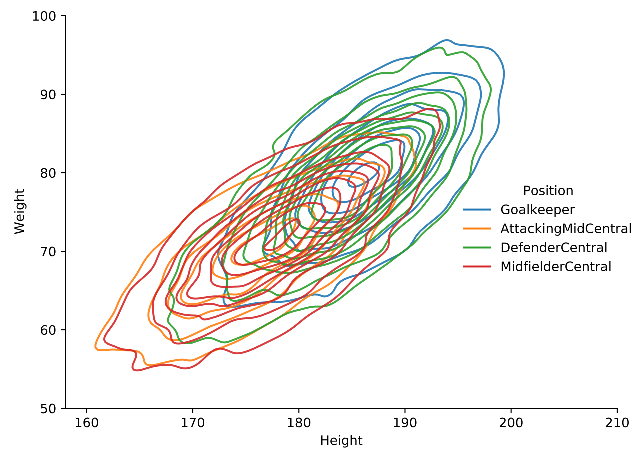

Figure 1. 2D density of height and weight on different positions.

Positions

Let's take a look at how different positions are connected. I made a correlation matrix of all data in columns representing players positions. In Figure 2 we can see that the strongest correlation is between wingbacks and defensive players on the same side of the field, same goes for attackers and midfielders. Attacking and defending positions are also correlated within them.

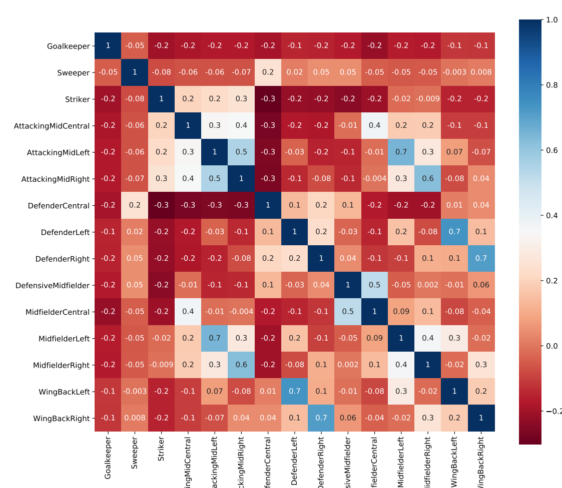

Figure 2. Correlation matrix of different position.

From this heatmap we can then guess that striker is an attacking position and sweeper a defensive, and that goalkeepers are specific only for their position.

Age

In this subsection I will analyze how old are players on different positions and how age affects player's attributes. But first, let's look how old are the players in our dataset - Figure 3.

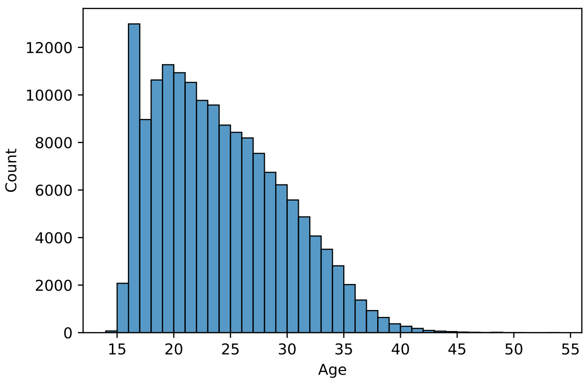

Figure 3. Histogram of age density.

In figure Figure 4 we can see violin plots for separated positions (I selected main position the same way as before). From it we can see that the youngest are goalkeepers and midfielders and the oldest sweepers, that are the only ones that have maximum density at over 30 years. Only sweepers and wingbacks don't have a violin plot that is pear-shaped, so these are the positions that older players are better at.

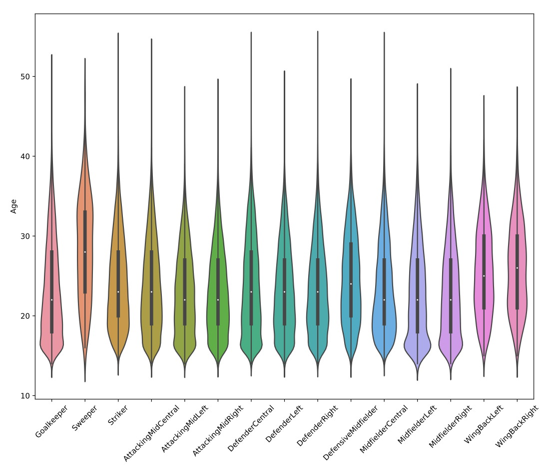

Figure 4. Violin plots of age on different positions.

Next two visualizations will help us understand what attributes age affects most. I calculated correlation of age with all skill attributes in our data. Figure 5 shows 7 attributes that correlate the most and 7 that correlate the least. I omitted others for better visibility.

With age, the players have more experience and get better in anticipation, concentration and composer, but they have

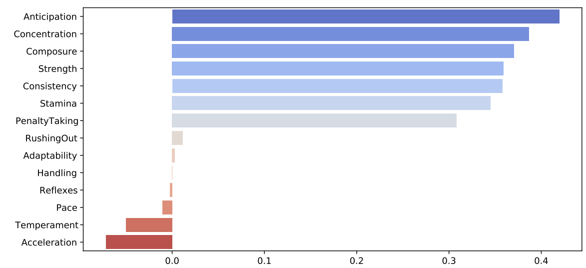

Figure 5. Attributes that are most and least correlated to age.

lower pace, less temperament and accelerate slower. Adaptability, handling and reflexes are least affected by age.

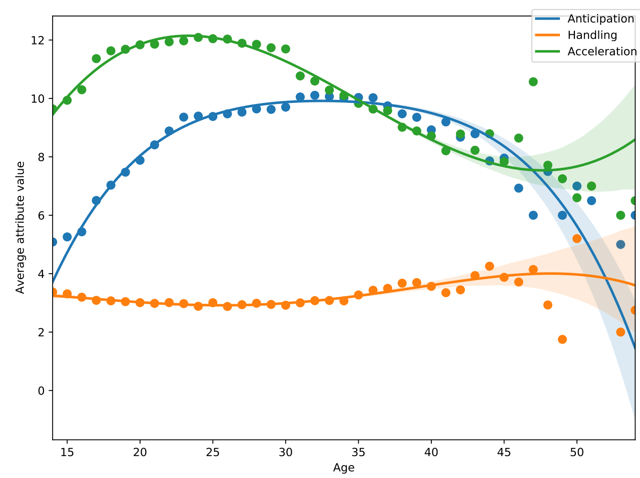

Figure 6. Average attribute value for certain age.

On Figure 6 I plotted average values of anticipation, acceleration and handling, to see the difference of how they change their values. I also fitted a polynomial of degree 4 for each attribute. After age 40 the values start jumping, and the fit is not so good anymore. The reason for that is that in our data we have only 450 players older than 40, of those only 51 older than 45. But if we observe values for age smaller than 40, we can see that most of the time anticipation is increasing, acceleration decreasing, and handling is more or less constant.

Discussion

The visualizations above tell us a lot about age and positions of players and connection between them. But we could further analyze the change of attributes in different age groups, find out what attributes are the most important for which positions or connect the attributes to height or weight of players. We didn't even begin to connect the attributes or positions to the number of goals and caps on international and U21 games. We could also show how those are connected to player's age.