English Premier League Football Analysis

IS665

jthomas6@pace.edu

Table of Contents

PART ONE

Introduction

Signing players for a football league is a big deal. Selecting the top performing players for the best price defines a season for a team and a career for a player. It can be difficult to tell which player will be the most beneficial for a team to sign. Data science and analytics has begun to take on a major role in player signing. A recent article in *Forbes* mentioned that analytics is helping teams discover undervalued talent and that recruiters are combining traditional player scouting with data to back up their decisions about who to sign (Kidd, 2018). The *Financial Times* said, in the past, analysts claimed that football was too chaotic and fluid to apply data to players movements and track them reliably. However, as data science has evolved, recruiters have been able to collect a wealth of player level data with a focus on recruitment and retention (Burn-Murdoch, 2018). This player level data has become a relatively new resource for recruiters, and leagues have begun to figure out how to utilize the data to their team's advantage.

In this project, exploratory analysis is conducted on EPL data provided by Football Association (FIFA) for 2015 through 2020 seasons (SOFIFA, 2020). The EPL is the top tier of England's football pyramid, with 20 teams battling it out for the honor of being crowned English champions. FIFA oversees game play regulations, organizes competitions, establishes standards for coaches, referees and sports medicine, encourages football's development around the world and oversees the transfer of players. Under FIFA's governance, football has become the world's most popular sport.

How It Helps Scouters and Recruiters

Player trades significantly impact how a team will perform each season. Having an insightful and meaningful analysis of player statistics is incredibly helpful for scouters and recruiters. This project helps scouters and recruiters determine which player will be most beneficial to the team.

Why Data Analysis Is Needed

Data analysis is needed for this because data is a powerful tool to be used alongside traditional scouting methods to select the best recruit for the team. Data analysis on player statistics, pulls out insights and information about a player that you would not spot easily without visualization. Having a lot of player stats does not help anyone unless you analyze the data and use it to make decisions.

How Player Data Analysis Will Improve Decision Making

Our analysis will improve decision making because we are using unique statistics that are not commonly used to judge players to decide on recruits. Age, Overall rating, Potential Rating, Growth and Stamina are extremely useful player statistics and our analysis aims to use these statistics to discover correlations between the best players in the league.

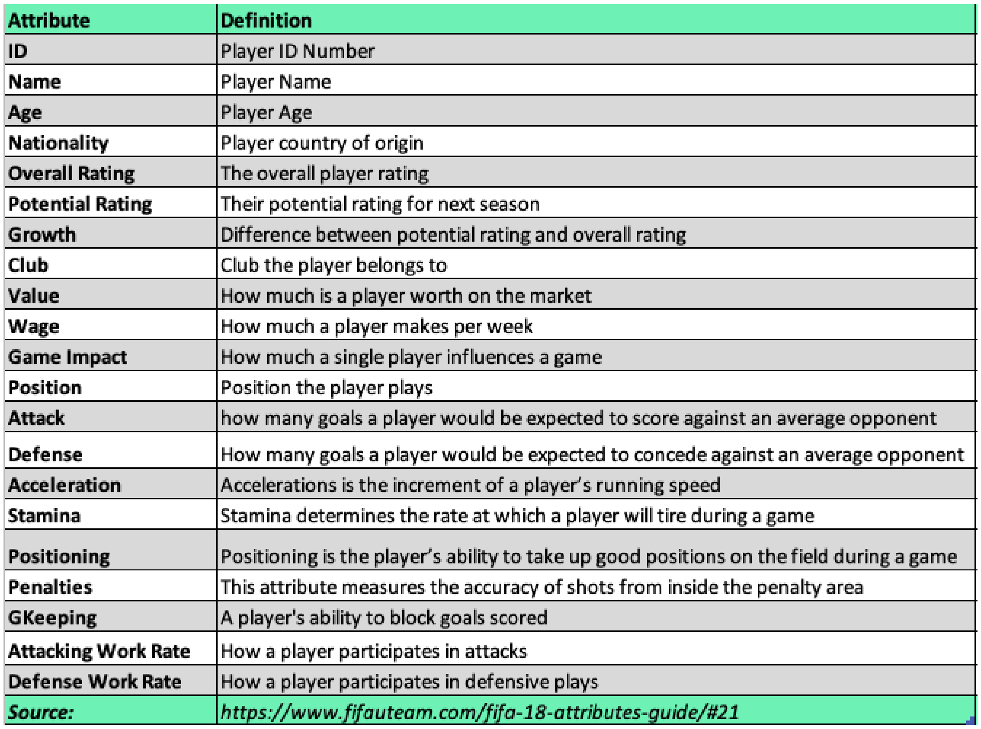
Data

Data Source

We scraped the website SoFIFA (2019) using R and rvest package. SoFIFA contains live player stats that get updated multiple times a month and goes back historically many seasons. We selected this source because it has stats aggregated to the player level with attributes that would be helpful for our analysis.

Significant Data Variables

Significant variables in our analysis include: Stamina, Age, Nationality, Club, Overall Rating, Potential Rating and Wage.



Attribute	Definition
ID	Player ID Number
Name	Player Name
Age	Player Age
Nationality	Player country of origin
Overall Rating	The overall player rating
Potential Rating	Their potential rating for next season
Growth	Difference between potential rating and overall rating
Club	Club the player belongs to
Value	How much is a player worth on the market
Wage	How much a player makes per week
Game Impact	How much a single player influences a game
Position	Position the player plays
Attack	how many goals a player would be expected to score against an average opponent
Defense	How many goals a player would be expected to concede against an average opponent
Acceleration	Accelerations is the increment of a player's running speed
Stamina	Stamina determines the rate at which a player will tire during a game
Positioning	Positioning is the player's ability to take up good positions on the field during a game
Penalties	This attribute measures the accuracy of shots from inside the penalty area
GKeeping	A player's ability to block goals scored
Attacking Work Rate	How a player participates in attacks
Defense Work Rate	How a player participates in defensive plays
Source:	*https://www.fifauteam.com/fifa-18-attributes-guide/#21*

Data Cleaning Process

The dataset which was scraped and extracted was then pre-processed and cleansed using tidyverse package and exported to excel sheets for all seasons.

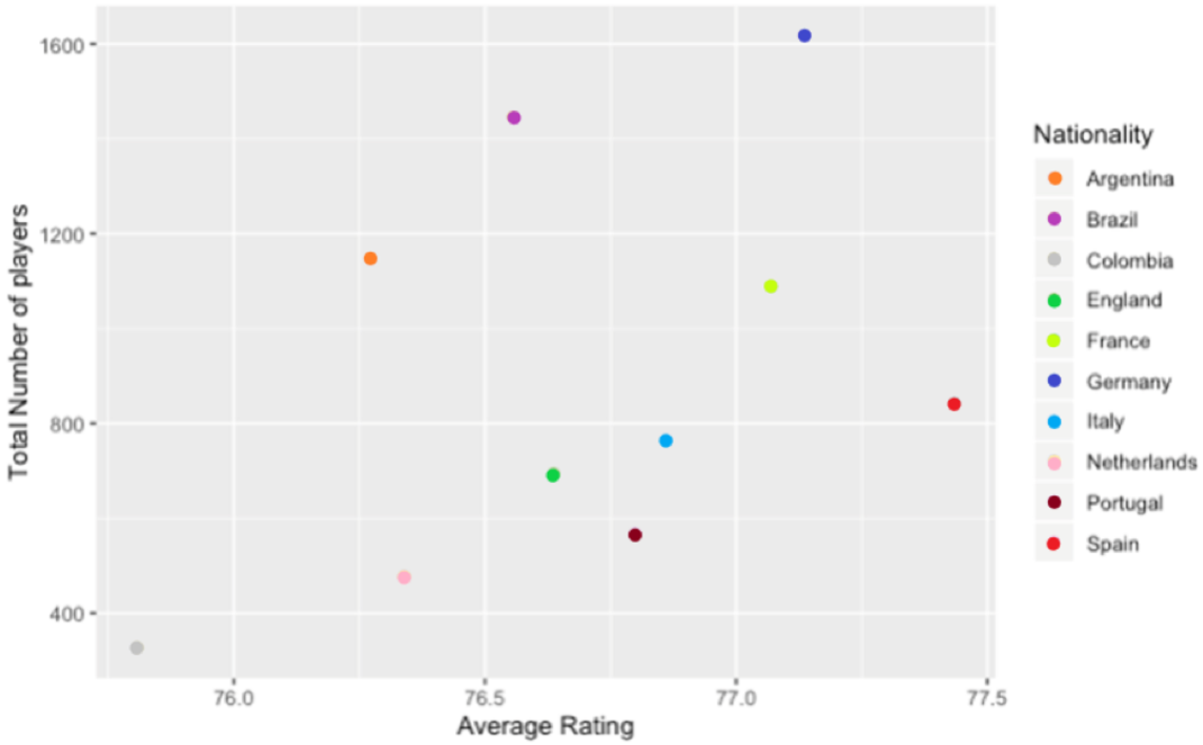
Descriptive Statistics

The mean age of players from season 2015-2019 is 26.85 and their ages range from 16-43 years old. The average overall_rating of a player is 76 and the rating ranges from 72-94. Age is positively skewed and overall_rating is normally distributed.

Exploratory Analysis

Question One:

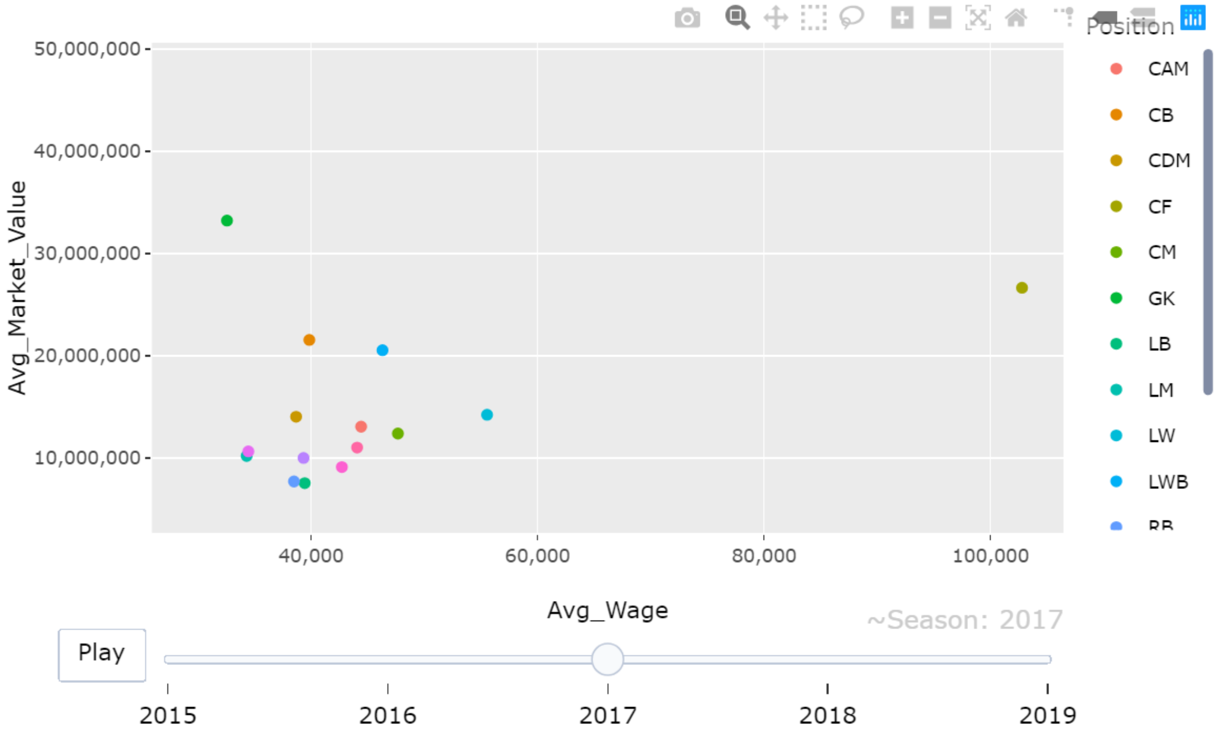
Which country produces the players with the highest overall_rating?



Answer: Germany

Question Two:

Which position was better paid over the years?

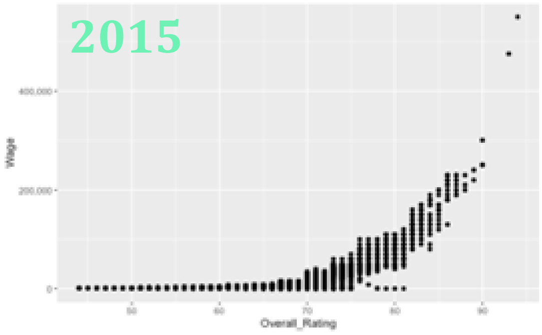


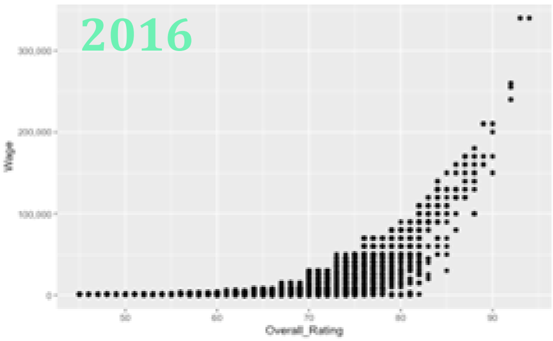
Answer: Central Forwards were highest paid players in the year 2017

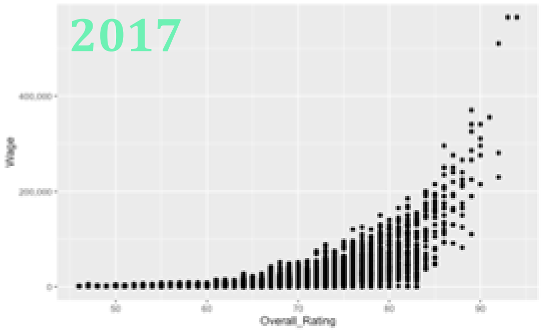
(PS: Refer the rmd file to view the animated plot)

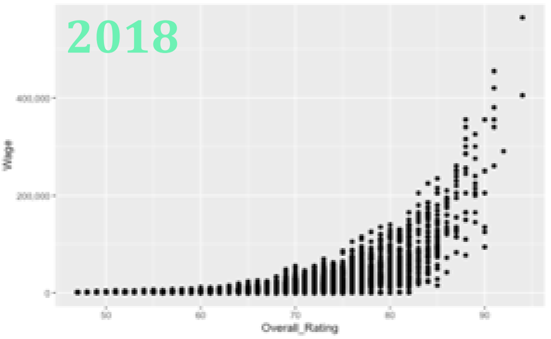
Question Three:

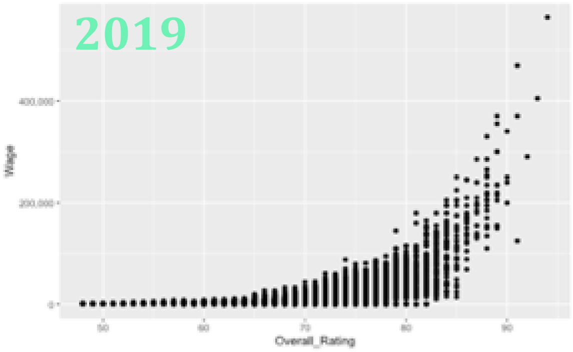
What is the correlation between wage and overall rating?







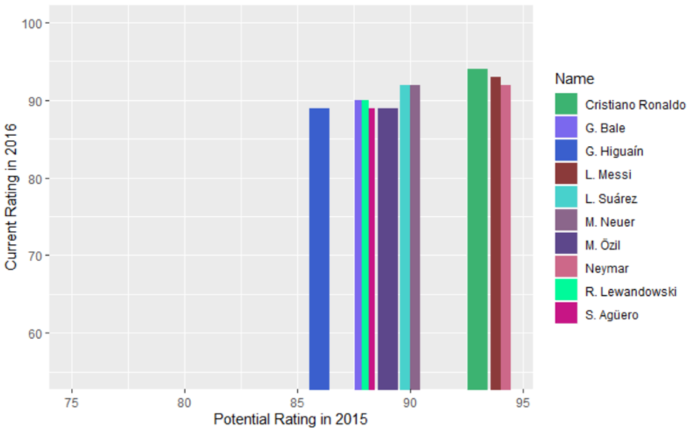


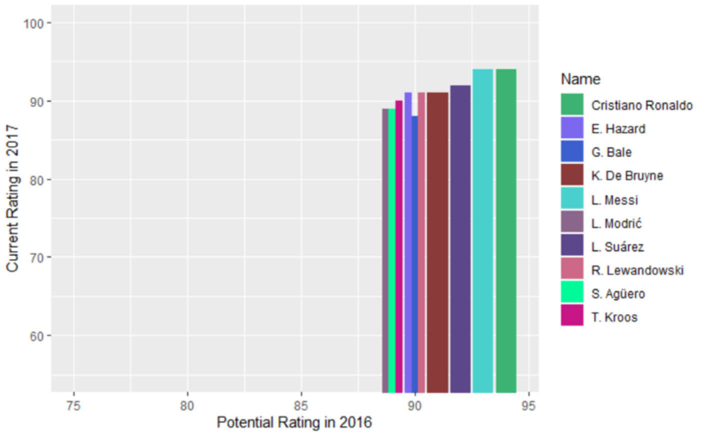


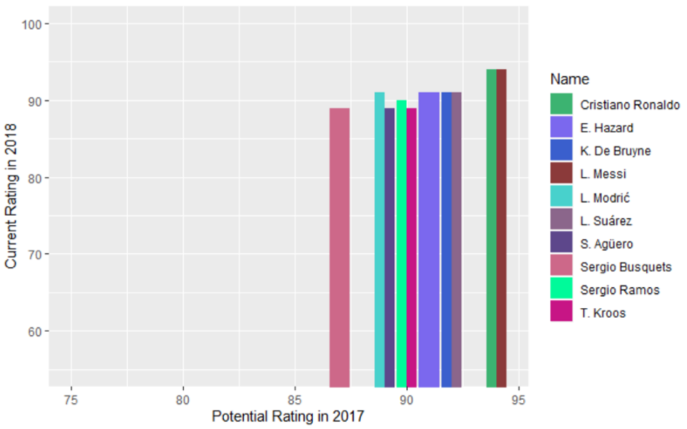
Answer: In 2015, wage and overall rating had a relatively linear correlation. You could point to the graph and say a player with this overall rating, will make approximately this much money. However, in 2019 this is not the case. The data points are much more scattered, the range is wider and there are many more outliers. Something in the player market has changed from 2015-2019 and overall rating is not the only factor to determine player wage.

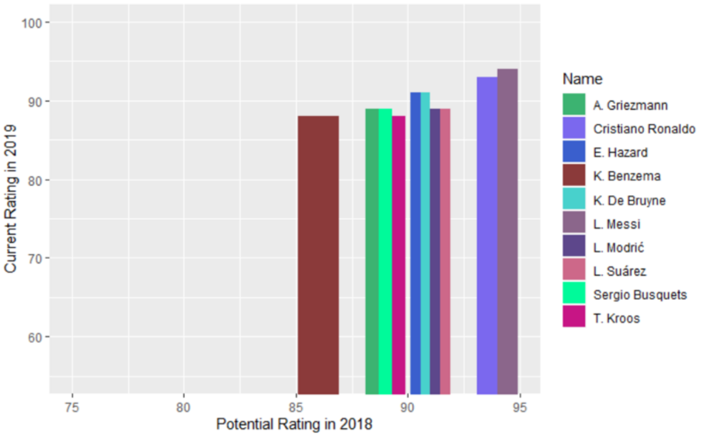
Question Four:

Have the highest paid players in the league met their potential rating from the previous season with the current rating for the next season? Basically, are the high paid players improving YOY?







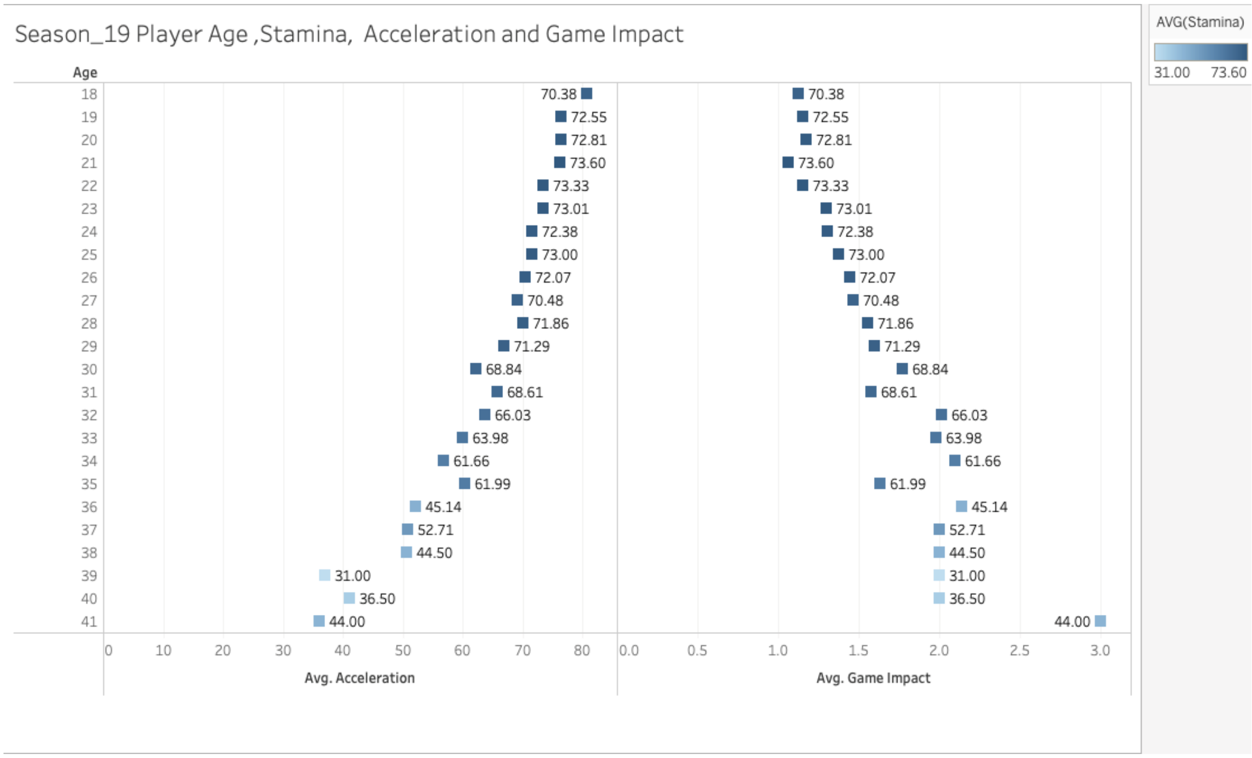


Answer:

Yes, the highest paid players tend to improve remain the same YOY.

Question Five:

What is the correlation between Player Age, Stamina, Acceleration and Game Impact for Season 2019?



Answer:

Age 18-29-year-old players have a relatively high stamina, game impact increases with age and acceleration decreases with age.

Results

Findings

Findings show that there are many variables that have to do with a player’s performance and they all interact in different ways. Some players have reached their potential rating, wage is less than value on average, Germany produces the players with the highest overall value, and a player’s stamina begins to decrease at age 27.

Proposed Prediction Method

For the predictive model, the ideas is to find what will be a player’s overall rating in 2020 based on their age, growth and maturity in previous years. This is done by finding players who are 27(peak age) in 2019, go back to 2015 and use those specific players to conduct analysis. This model will allow analysts to predict when a players’ performance will start to decline. This will help recruiters as they would know the appropriate time to recruit an upcoming player.

PART TWO

Model Building

From the EDA conducted, it was evident that 27 was the peak performance age for a player. The main idea behind our proposed model is to filter out all players who are 27 years old in the current season (2019) and then trace them back to 2015 to analyze their performance and ultimately predict their overall rating and consequent performance in the upcoming years.

Independent Variables

For a given player, the independent variables selected are age, wage, market value, maturity and growth.

Dependent Variables

The selected the dependent variable of the model is a player's overall rating.

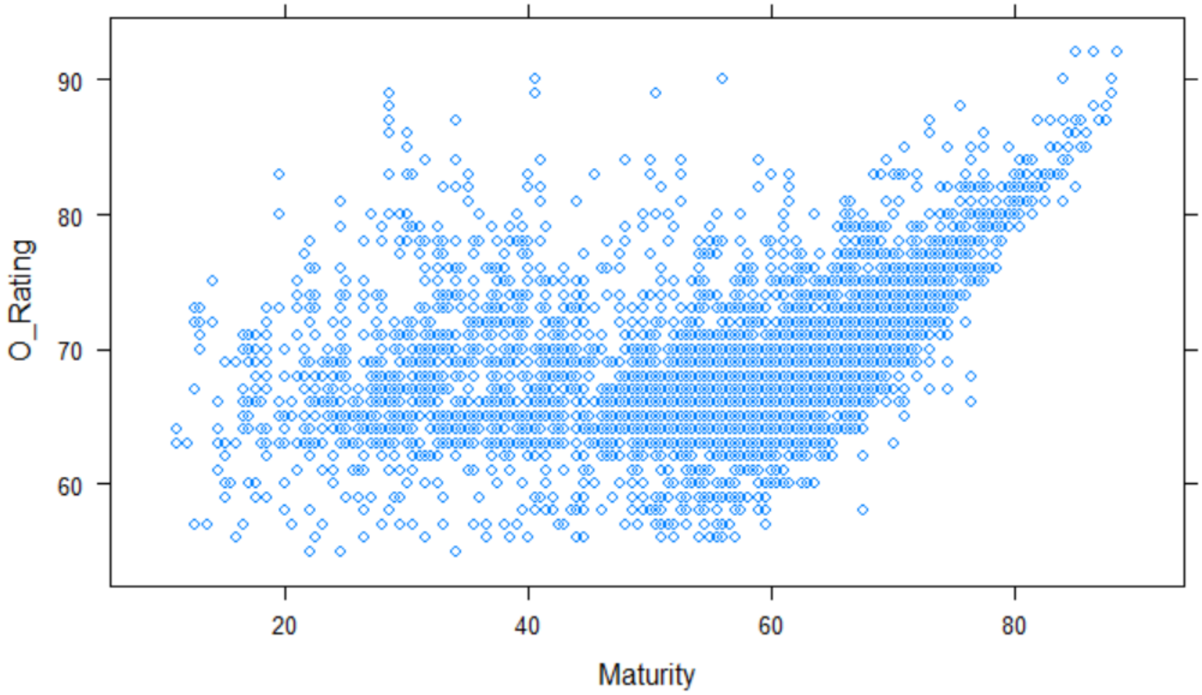
Model Justification Based on Dependent Variables

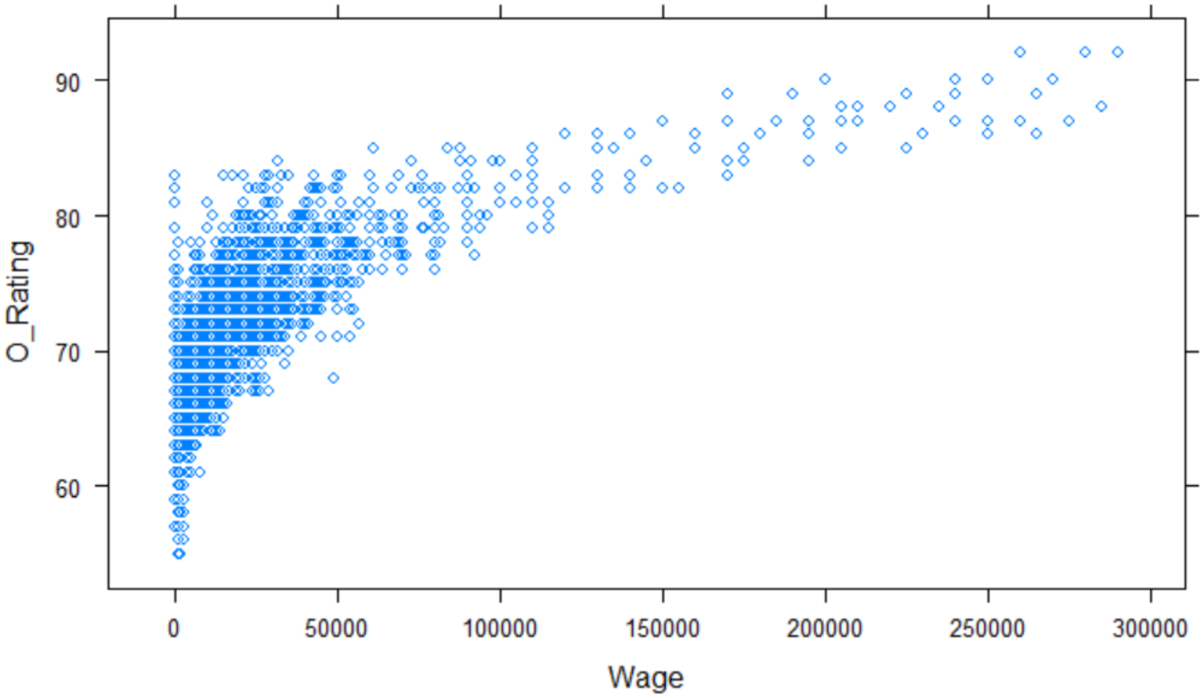
Since most of the variables are numerical, it would be appropriate to build a logistic regression model with the above said dependent and independent variables. The variable selection for the model using the backward method with the help of leaps package to better select a combination of variables which would give a model with better accuracy values.

Identify Any Preprocessing

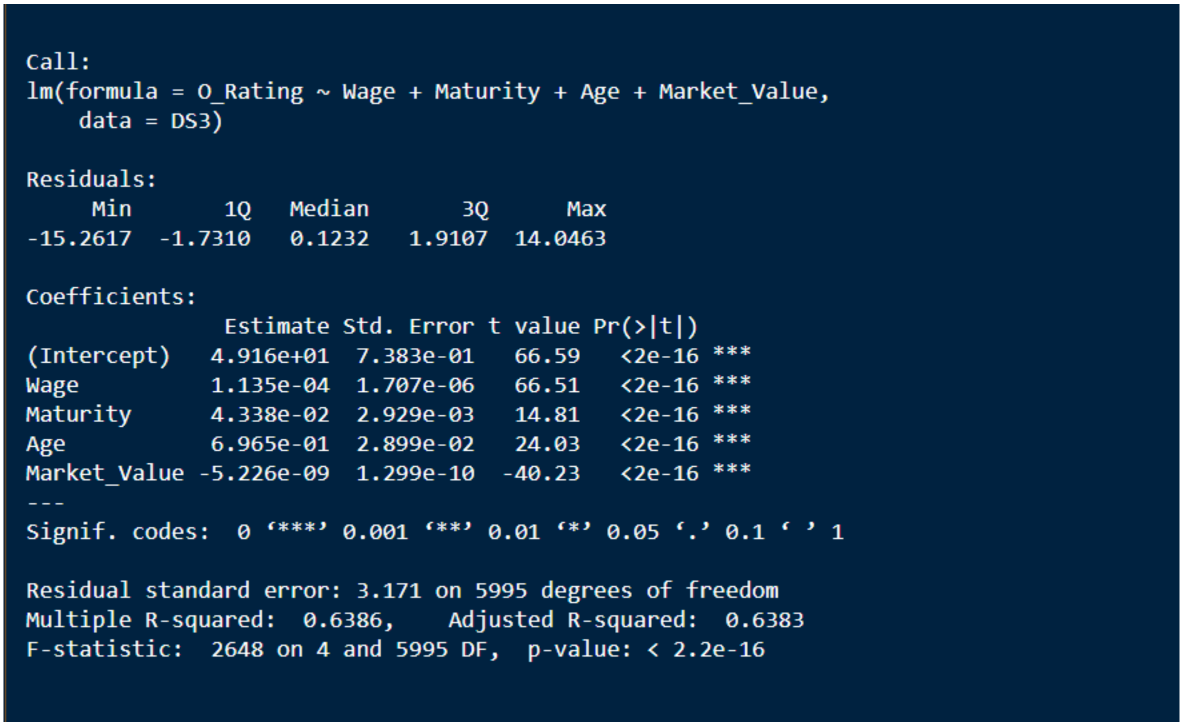
Scraped sofifa.com for players who are 27 years old now in the order of the overall ranking. The dataset for the predictive model consists of top 1,200 professional soccer players who are 27 years old now. The maturity variable for a player can calculated as an aggregation of different attributes such as player positioning, vision and game impact.

Correlation between Dependent and Independent variables





Model Results



```
Call:
lm(formula = O_Rating ~ Wage + Maturity + Age + Market_Value,
    data = DS3)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2617  -1.7310   0.1232   1.9107  14.0463 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.916e+01  7.383e-01   66.59   <2e-16 ***
Wage          1.135e-04  1.707e-06   66.51   <2e-16 ***
Maturity      4.338e-02  2.929e-03   14.81   <2e-16 ***
Age           6.965e-01  2.899e-02   24.03   <2e-16 ***
Market_Value -5.226e-09  1.299e-10  -40.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.171 on 5995 degrees of freedom
Multiple R-squared:  0.6386,	Adjusted R-squared:  0.6383 
F-statistic:  2648 on 4 and 5995 DF,  p-value: < 2.2e-16
```

The predictive multi linear model can be represented as,

Overall Rating = $1.135*(10)^{-1}$ * Wage + $4.338*(10)^{-2}$ * Maturity + $6.965*(10)^{-1}$ * Age – $5.226*(10)^{-9}$ * Market_Value

Model Accuracy

The independent variables have a significance value less than $2*(10)^{-16}$

The multi linear model has an adjusted R-squared value of 0.63

Interpretation of Results

Overall rating has a positive relation with age, wage and player maturity while a negative relation with the market value which may suggest that the market price of a player is at its best and may fall over the upcoming years which is indeed surprising. Also, the player performance increases as the age increases which may suggest that the model is biased towards old players.

Conclusion

Using the proposed model, a club can make better informed decisions while selling and buying a player during the transfer markets with the wage and player market value taken into consideration along with the player overall rating and performance.

References

Burn- Murdoch, J. (2018) How data analysis helps football clubs make better signings. *Financial Times.* Retrieved from: https://www.ft.com/content/84aa8b5e-c1a9-11e8-84cd-9e601db069b8

Kidd, R. (2018) Soccer's Moneyball Moment: How Enhanced Analytics Are Changing The Game. *Forbes.* Retrieved from: https://www.forbes.com/sites/robertkidd/2018/11/19/soccers-moneyball-moment-how-enhanced-analytics-are-changing-the-game/#60ece96576b2

USsoccer (2019). FIFA — SOCCER’S WORLD GOVERNING BODY. Retrieved from: https://www.ussoccer.com/history/organizational-structure/fifa

SOFIFA (2020) Player Dataset. Retrieved from: https://sofifa.com