**Team Wrigley Field**

# DSC423 Project: Final Report

**Team Members:**
Amanuel Petros
David Guo
Jilei Hao
Kaushik Bandaru

# 1. Introduction

Players exchange is one of the major aspects of building a successful soccer club. Clubs earn money by selling players more important for other clubs and spend money buying players that are needed to make their own squad stronger. A successful player exchange strategy can earn the club significant amount of money, and more importantly, can give the club a competitive team to win games and trophies. The most important part of such a strategy is to precisely evaluate players' transfer value.

In soccer, generally, the transfer value of a player is the total cost the buyer club has to pay to sign the player from the seller club. A player's transfer value could be determined by many factors, but the most important determinant is the player's quality. However, determining player's quality can be a daunting task because it could become high subjective, or be overwhelmed by too many details about the players. Beyond all the complications, there is one agreement that the most important determinant of players quality is their performance data. No matter what other factors are, such as negotiation, club competition, a good performance would always earn a player high transfer value while a bad one would always lead to the opposite.

In this study, we will analyze English Premier League players' performance data together with their basic information to predict their transfer value. The basic information of players describes the non-performance related static information about the player, such as age, nationality and position. The performance data describes the player's playing statistics in the past games, such as number of goals, number of assists and total appearance.

The analysis is split into four parts based on the player positions. Players playing different positions have very different responsibilities and emphasize different set of skills. Goalkeepers are special type of players that stand in front of the goal to stop the ball from moving into the net and are the only players that can use hand to interact with the ball. Defenders usually positioned at the back of the field, to stop or interfere with the opponent's attack. Some of the defenders, especially the ones on the left and right flank, are also taken the responsibility of helping their own team's attack. In modern games, some strategies require defenders to initiate attack from the back of the field by passing the ball forward without making long balls. Forwards are the players positioning at the front of the field, organizing attacks and scoring goals. Midfielders are players positioning between defenders and forwards, in the mid-field, connecting the backline and the frontline of the team, and are able to help both the attack and defense. Separate models are built based on different sets of variables for each position. Conclusions are drawn both independently for each model and comprehensively by comparing results from all four analysis.

# 2. Data Preparation

The response variable in this project is the transfer value of the players. The explanatory variables we will consider including the player's position, age, total goals, number of appearances, pass accuracy etc. The total number of explanatory variables depends on player positions.

The market value information was gathered from https://www.transfermarkt.com/premier-league/marktwertaenderungen/wettbewerb/GB1 ("transferData.csv"), The players general information was gathered from FIFA 19 database, https://www.kaggle.com/karangadiya/fifa19 ("data.csv"). Historical performance data was collected from https://www.kaggle.com/adithyarganesh/english-premier-league-player-data-20182019 ("fpl_data_2018_2019.json"), which was originally scrapped from www.premierleague.com. Fantasy points data was from https://fantasy.premierleague.com/player-list/ ("fpl data.csv"). Player seasonal data from 2018-2019 are from fbref.com. The data was split into all players https://fbref.com/en/comps/9/stats/Premier-League-Stats ("epl player data.csv" ) and goalkeeper specific data https://fbref.com/en/comps/9/keepers/Premier-League-Stats ("epl goalkeeper.csv" ) . These data points will be combined into a single data set for the study.

The data contains 685 EPL players' historical performance data at the beginning of the 2018-19 season, and the player's performance data in 2018-19 season. The primary challenge is to integrate the multiple datasets of different formats into a single set. The players' history stats are in json format and are not well structured and players playing different positions should have different set of variables evaluating their performance. The market value data is on the web and needs to be scrapped into file for further use. The drawback of our data is the lack of injury information, which can be important in deciding player values. But frequent injury and long time absence of play may affect the overall data anyway. Another limitation of our data is that player stats from other leagues are not included. This could affect player value prediction accuracy for players transferred from another league to EPL not long ago.

We combined these different datasets (with the exception of the Fifa data, which we didn't include) into one dataset. The "epl player data.csv" contains 508 rows of player data with information about a players statistics over the course of the 18-19 Premier League season, including information about their appearances and goals scored. The "fpl data.csv" contained 423 rows about a player's fantasy points accumulated over the 18-19 Premier League season. The "epl goalkeeper.csv" data contains 37 rows about the goalkeepers who played in the Premier League in 18-19 such as goals conceded, and save percentage. The "fpl_data_2018_2019.json" data contained 685 rows about a player's historical performance in the Premier League,

including information on a player's appearances and wins. The "transferData.csv" contained 603 rows on a player's transfer value.

For the "goalkeeper.csv" data, we cleaned player names using SPSS and converted the file to csv. We also got rid of the accents in the names, so Martin DÃºbravka\Martin-Dubravka was changed to Dubravka Martin. We changed the nationality column to be only in three letters (SVK, BRA) when it was previously both in two and three letter names (sk SVK, br BRA).

For the "epl player data.csv", we removed the accents on some of the names using a Python script, so the names can be more consistent from file to file. The data had 2 and 3 letter country codes (the ISO code and the FIFA country code) for the nationality of players in the same cell separated by a space (eg a French player had a fr and a FRA in the nation column). We split those two words up as well and used VLOOKUP to get the full country names of the players. Also, if the player had played multiple positions, the player's "pos" column data contained multiple positions (a player who had played as a defender and midfielder had DFMF), so we separated the data into three different columns with the LEFT and MIDDLE functions.

We also wrote a java program that converts the historical Premier League data ("fpl_data_2018_2019.json") from the original json format to tabular format "playerhistory.csv". He stripped accents from players' names using Apache Common Language library. There were some missing data as players playing certain positions only has a subset of variables populated. For example, goalkeepers don't have any attacking variables populated. This type of missing values are expected due to differences in player roles.

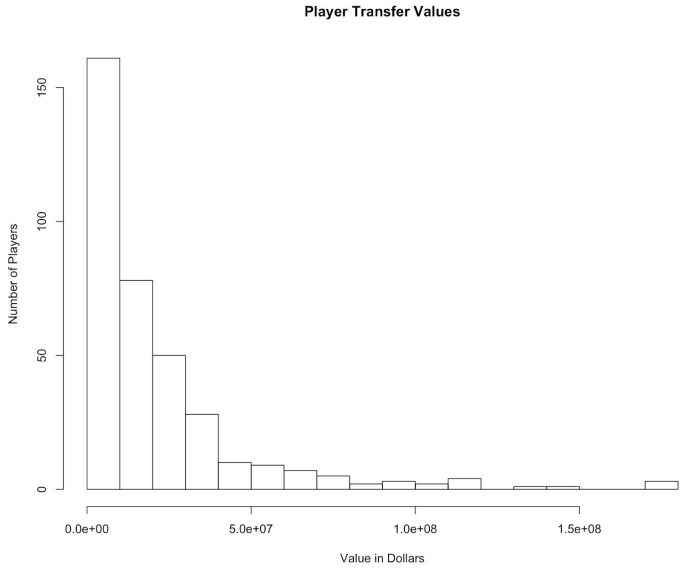
We cleaned the data on "transferData.csv" and "fpl data.csv" dataset using Excel. We also removed the accents of the names and converted them to English letters using a VBA Script.

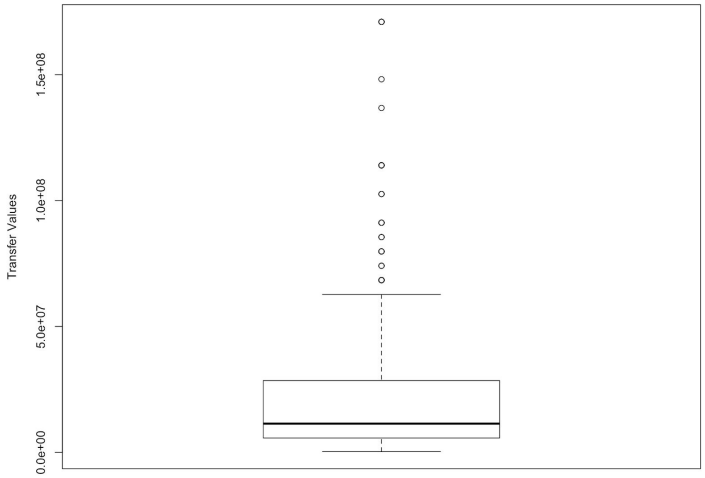
Before merging the data, we removed some variables in the goalkeeper data (nationality, position) that were redundant with the EPL player data. We used R to merge the datasets together. We left outer joined "epl player data.csv" with the "goalkeeper.csv" data, and then joined the resulting data with the fantasy and transfer data. Then we inner joined the resulting data with the historical Premier League data ("playerhistory.csv").

After processing the dataset, we were left with 413 rows and 94 columns of player data. To further distinguish between the different variables, we added a 90 to the end of a variable if the statistic looked at player information per 90 minutes, and an 18 to the end to the variable if the statistic looked at player information in the 18-19 season, so a player's goals per 90 in the 18-19 season would be "Gls90.18". We also deleted some redundant variables like the club variable.

# 3. Data Analysis

## Transfer Values

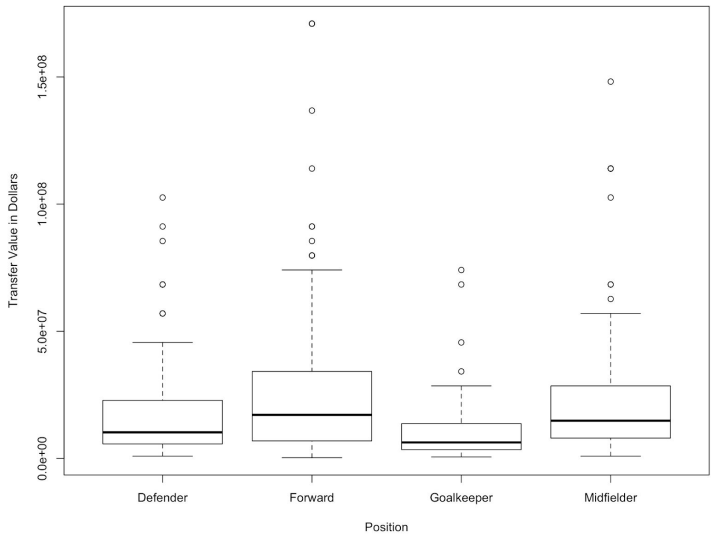



There looks to be a heavy right skew in regard to the player transfer values, with most players valued at between 0-30 million dollars. There are certain star players that stand out over the average player, so their transfer values will be significantly higher.

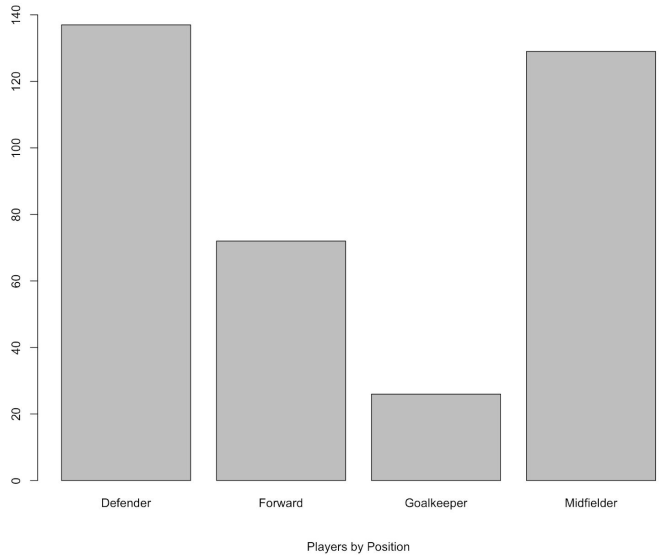
## Transfer Values by Position

Forwards on average have the highest transfer values, then midfielders, defenders, and goalkeepers. Forwards tends to do the flashiers things in soccer, mainly scoring goals, and their impact on the game is more overt, so their value tends to be higher.

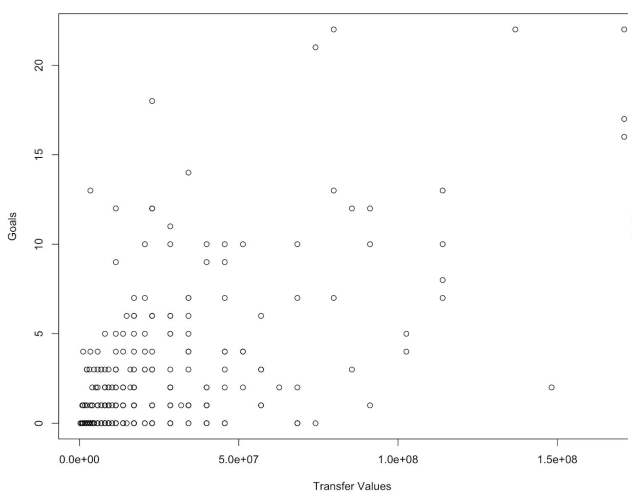


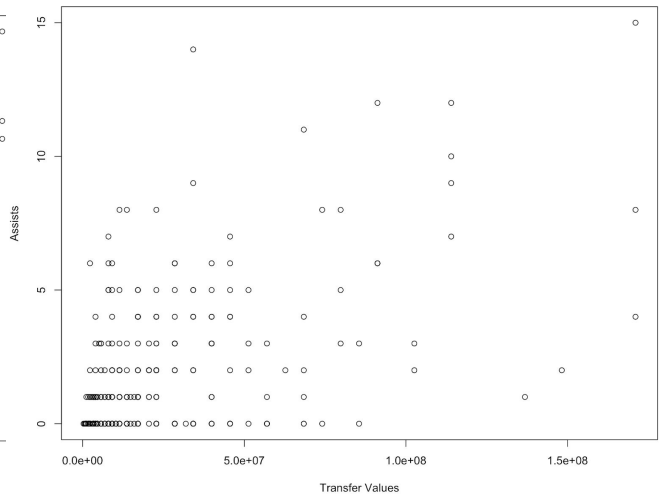
## Number of Players per Position

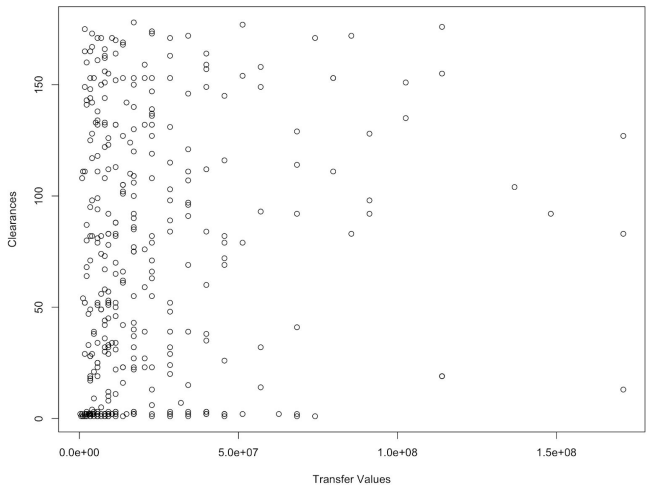
This result fits mostly used formations such as 4-4-2 have 4 defenders, 4 midfielders and 2 forwards, and clubs generally need to have corresponding ratio of players in the team for each position.



## Transfer Values by Assists, Goals and Successful Tackles



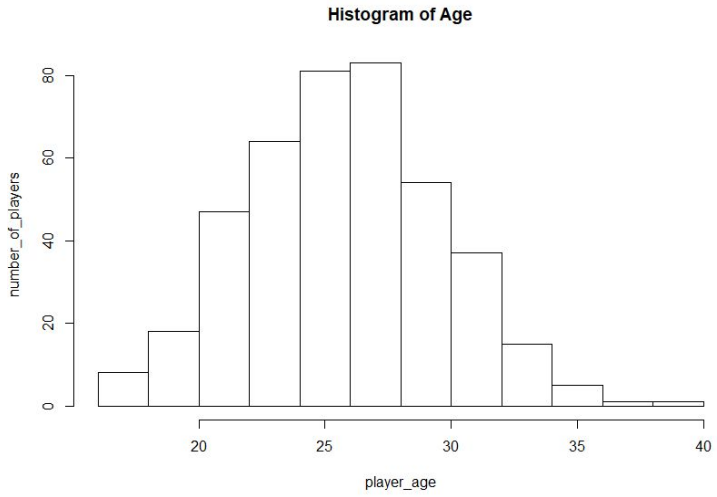



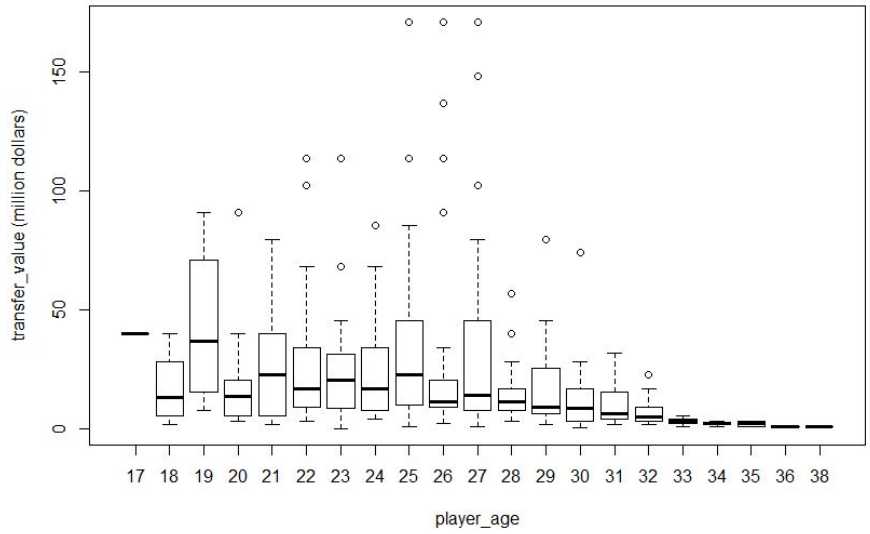

| | Clearances | Goals | TransferValues | Assists |
|---|---|---|---|---|
| Clearances | 1.0000000 | 0.2119716 | 0.1489289 | 0.1964791 |
| Goals | 0.2119716 | 1.0000000 | 0.6111996 | 0.5340991 |
| TransferValues | 0.1489289 | 0.6111996 | 1.0000000 | 0.5199003 |
| Assists | 0.1964791 | 0.5340991 | 0.5199003 | 1.0000000 |

There is the highest correlation between transfer values and goals, there is also a smaller but still significant correlation between goals and assists, and there is a small positive correlation between goals and clearances. This backs up the earlier graph showing that forwards are valued more since their main job is to score goals, and the fact that it's easier to measure the performance of a forward than a defender.

Given the differences between positions and the differences in performance statistics between positions, we could separate the data into four positions and do four regression models based on the data. We could also use an interaction term between the positions and the goals and assists, as different positions have differences in performance data due to their positional duties.

## Player Age

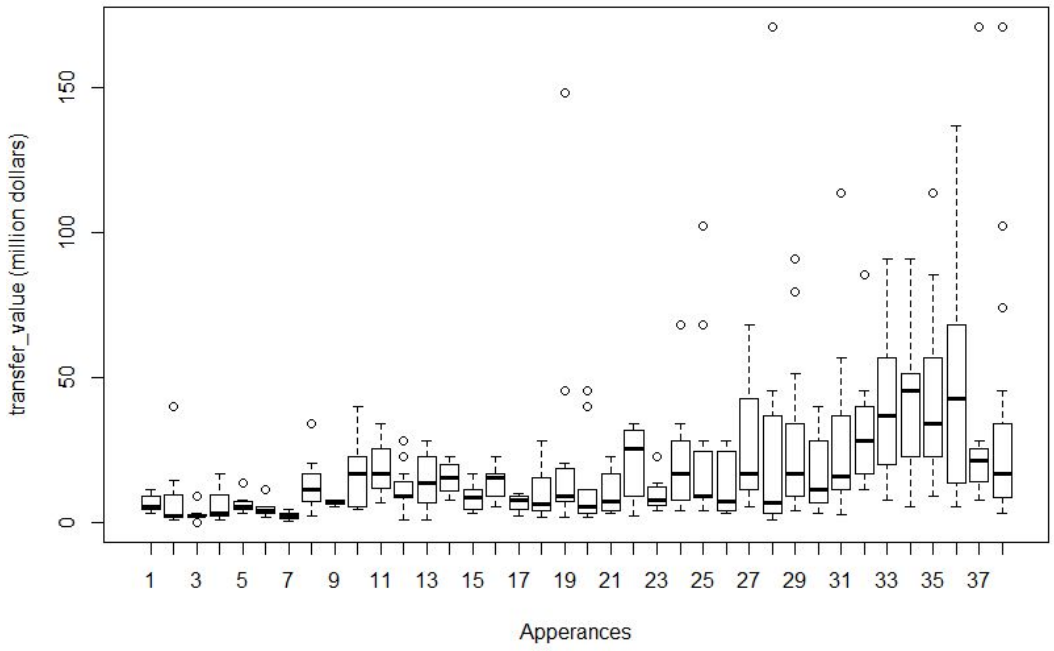



| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 17.00 | 24.00 | 26.00 | 26.27 | 29.00 | 39.00 |

Most of the players are between ages 23 and 30, which are the best years of soccer players. The distribution is slightly right skewed, meaning there are players with very old ages, but none before age 17, since teenagers are not fit to play in the adult league. Since a soccer player has fairly short professional life (around 15 years), their value could be highly related to their age, which means how many good years left for them to play at a high level. Should experiment second-order model between transfer-value and age, expecting a negative second-order beta and flat curvature.

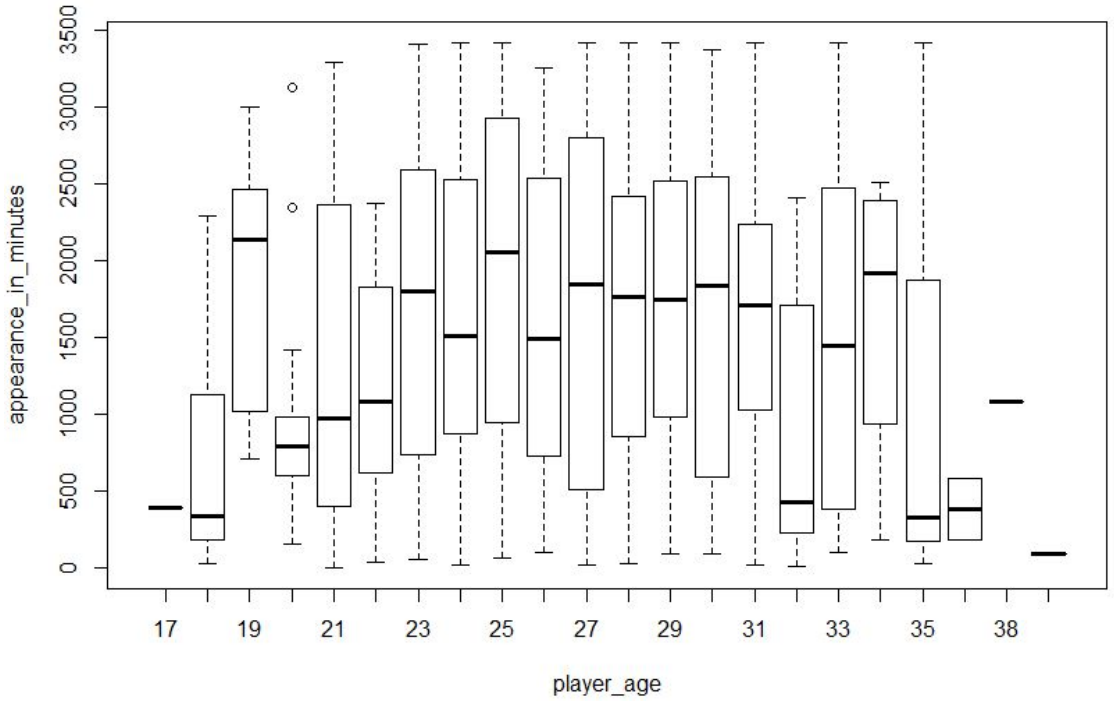
## Appearances




Players's number of appearances in a season seems to have a positive correlation to the player's value. Outstanding upper-bound outliers are appeared only in the right half of the plot, which probably indicates that for a player to have high market value, he must at least appeared in half of the games in a season. The correlation between transfer value and appearance is likely to be a second order term, since the slope change is much slower in the left half than in the right half. Some outstanding upper-bound outliers occurs in the middle of the plot, which might be caused by injuries.

An interesting fact about high appearances is that the highest two appearances does not seem to have high median transfer-value. It probably because many full-appearance players are goalkeepers, which does not have high median transfer-value. Should experiment second-order model between transfer-value and appearance and expecting a positive second-order beta.

## Player Age and Appearances




Highest appearance minutes are located in the middle of the plot. Which could because there are many players between age 23 and 30, and they are the primary force for each team. Interesting peak appeared at 19 and 33-34 age group. It could probably caused by the high importance of talented young players and highly experienced players. The unstable result could also be caused by lack of samples in those age groups. Should experiment with an interaction term between appearance in minutes and age as young age and high appearance may lead to high player value

## Correlation Matrix Between Major Stats

| | Age | Apps | Starts | Subs | Min | Gls90 | G.A | G.PK |
|---|---|---|---|---|---|---|---|---|
| Age | 1.00000000 | -0.01373886 | 0.06320492 | -0.190153679 | 0.05929452 | -0.06629688 | -0.095748105 | -0.082574753 |
| Apps | -0.01373886 | 1.00000000 | 0.91667582 | 0.059529455 | 0.93053154 | 0.23859365 | 0.221792671 | 0.213630880 |
| Starts | 0.06320492 | 0.91667582 | 1.00000000 | -0.344353696 | 0.99612180 | 0.11221715 | 0.067076029 | 0.083043452 |
| Subs | -0.19015368 | 0.05952945 | -0.34435370 | 1.000000000 | -0.30211574 | 0.28021588 | 0.353502437 | 0.294443745 |
| Min | 0.05929452 | 0.93053154 | 0.99612180 | -0.302115740 | 1.00000000 | 0.11466055 | 0.071761680 | 0.086500283 |
| Gls90 | -0.06629688 | 0.23859365 | 0.11221715 | 0.280215878 | 0.11466055 | 1.00000000 | 0.830125032 | 0.984286143 |
| G.A | -0.09574810 | 0.22179267 | 0.06707603 | 0.353502437 | 0.07176168 | 0.83012503 | 1.000000000 | 0.815476359 |
| G.PK | -0.08257475 | 0.21363088 | 0.08304345 | 0.294443745 | 0.08650028 | 0.98428614 | 0.815476359 | 1.000000000 |
| G.A.PK | -0.10729181 | 0.20329785 | 0.04524999 | 0.364571761 | 0.05070571 | 0.80717876 | 0.992630297 | 0.813516217 |
| SoT90 | -0.15790691 | 0.12495089 | -0.04338271 | 0.401906796 | -0.03591460 | 0.62833309 | 0.616867823 | 0.608455919 |
| Fls90 | 0.01993100 | -0.14285732 | -0.17854795 | 0.110379488 | -0.17913426 | 0.01185623 | 0.028097052 | 0.009575141 |
| Crd | 0.01747944 | -0.07245777 | -0.06665861 | -0.003718055 | -0.06964323 | 0.05389562 | -0.005752789 | 0.059208286 |
| Transfer.Values | -0.22030575 | 0.40447169 | 0.38082020 | -0.001027527 | 0.38530920 | 0.43920345 | 0.439251354 | 0.416714933 |

| | G.A.PK | SoT90 | Fls90 | Crd | Transfer.Values |
|---|---|---|---|---|---|
| Age | -0.107291809 | -0.15790691 | 0.019931005 | 0.017479438 | -0.220305755 |
| Apps | 0.203297853 | 0.12495089 | -0.142857321 | -0.072457769 | 0.404471690 |
| Starts | 0.045249995 | -0.04338271 | -0.178547947 | -0.066658615 | 0.380820196 |
| Subs | 0.364571761 | 0.40190680 | 0.110379488 | -0.003718055 | -0.001027527 |
| Min | 0.050705714 | -0.03591460 | -0.179134263 | -0.069643231 | 0.385309197 |
| Gls90 | 0.807178758 | 0.62833309 | 0.011856230 | 0.053895618 | 0.439203449 |
| G.A | 0.992630297 | 0.61686782 | 0.028097052 | -0.005752789 | 0.439251354 |
| G.PK | 0.813516217 | 0.60845592 | 0.009575141 | 0.059208286 | 0.416714933 |
| G.A.PK | 1.000000000 | 0.59964278 | 0.027230268 | -0.004989019 | 0.421026644 |
| SoT90 | 0.599642781 | 1.00000000 | 0.045960113 | -0.034119546 | 0.405441723 |
| Fls90 | 0.027230268 | 0.04596011 | 1.000000000 | 0.097545053 | -0.082365596 |
| Crd | -0.004989019 | -0.03411955 | 0.097545053 | 1.000000000 | -0.080618157 |
| Transfer.Values | 0.421026644 | 0.40544172 | -0.082365596 | -0.080618157 | 1.000000000 |

Starts and Appearance has strong correlation since any start counts as appearance. But appearance doesn't need to be Starts, it could also be Subs.

Age has on average low correlation with respect to all other variables, but has relatively high correlation to Subs and Transfer values. It could be explained by the fact that young players are more likely to be substitutes instead of starts, and they tend to have more potential value since they are can grow and has many years to play

Appearances has high correlation with Minutes played, which indicates the fact that players being allowed to play are usually also given sufficient playing time. In other words, not many players appears often but always be substituted shortly after they started

Starts has an extremely high correlation with playing time (Min) which indicates that starting players almost always play most of the game time. It could be related to the fact that a competitive soccer game only has 3 substitutes which are usually reserved for the later time of games

Shorts on Target per 90 minutes (SoT90) has a high correlation with Goal stats. The more accurate shots, the more goals, obviously.

Transfer value has relatively high correlation with appearance, since the good players usually get more time playing. It also has relatively high correlation with goal related stats, simply because the more goals, the better, and the winner of the soccer game is determined by the goal difference.

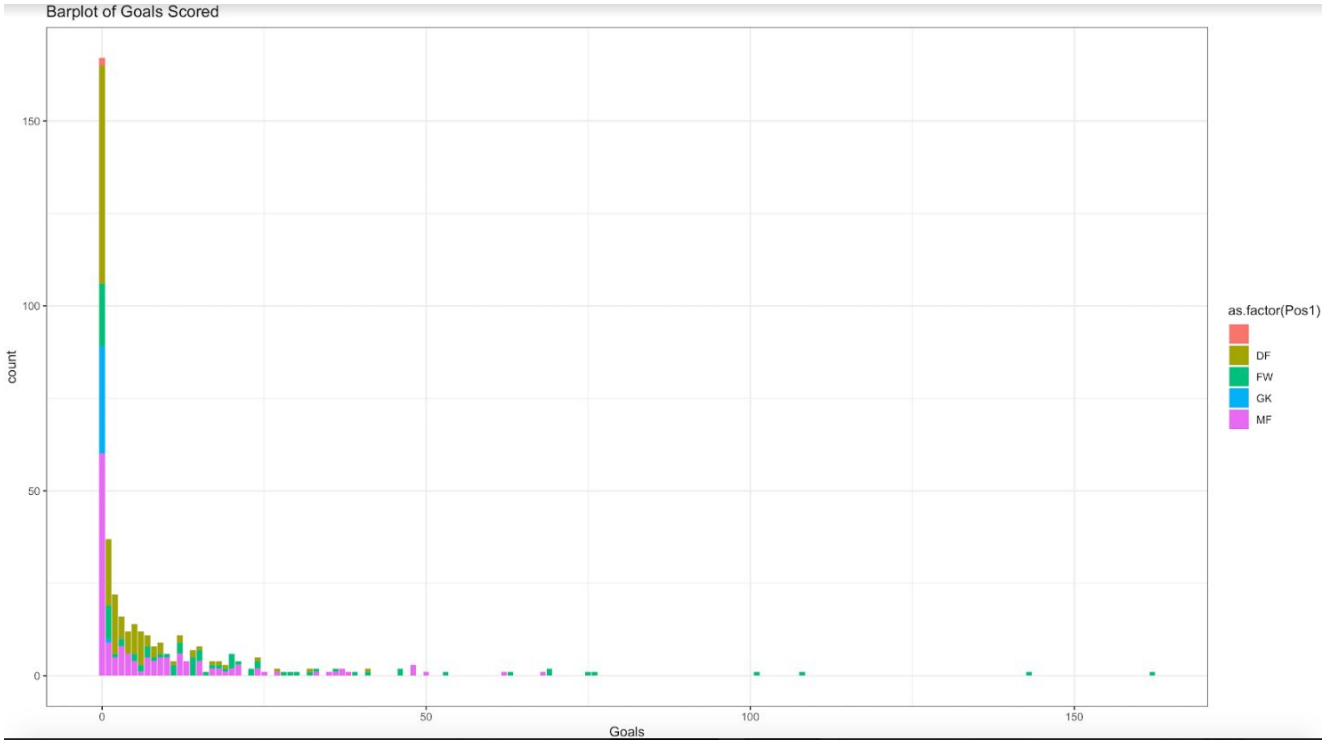
Goal related stats (Gls90, G.A, G.A.PK) has higher positive correlation with Subs instead of Starts which is very interesting. A possible explanation based on my experience could be substitutes are usually come after both team are little exhausted, and fresh legs give them more chances to score goals, while at the starting phase of the game, teams usually balance each other's attacking with good defence, but it takes energy.

## Correlation Matrix between player information

| | X | Age | Born | Apps | Starts | Subs |
|---|---|---|---|---|---|---|
| X | 1.000000000 | 0.055937647 | -0.050201016 | -0.040323004 | -0.02713975 | -0.02870717 |
| Age | 0.055937647 | 1.000000000 | -0.991929285 | -0.037771681 | 0.03940718 | -0.19360755 |
| Born | -0.050201016 | -0.991929285 | 1.000000000 | 0.042073609 | -0.03383624 | 0.18978991 |
| Apps | -0.040323004 | -0.037771681 | 0.042073609 | 1.000000000 | 0.92172688 | 0.07245344 |
| Starts | -0.027139753 | 0.039407177 | -0.033836238 | 0.921726880 | 1.00000000 | -0.32003800 |
| Subs | -0.028707174 | -0.193607551 | 0.189789913 | 0.072453438 | -0.32003800 | 1.00000000 |

This summary gives us the information regarding the match between various variables depicting the relationship between Age, born, appearances, starts, subs. We can see that age and appearances have 37.7% correlation and 39.4% correlation between age and starts.

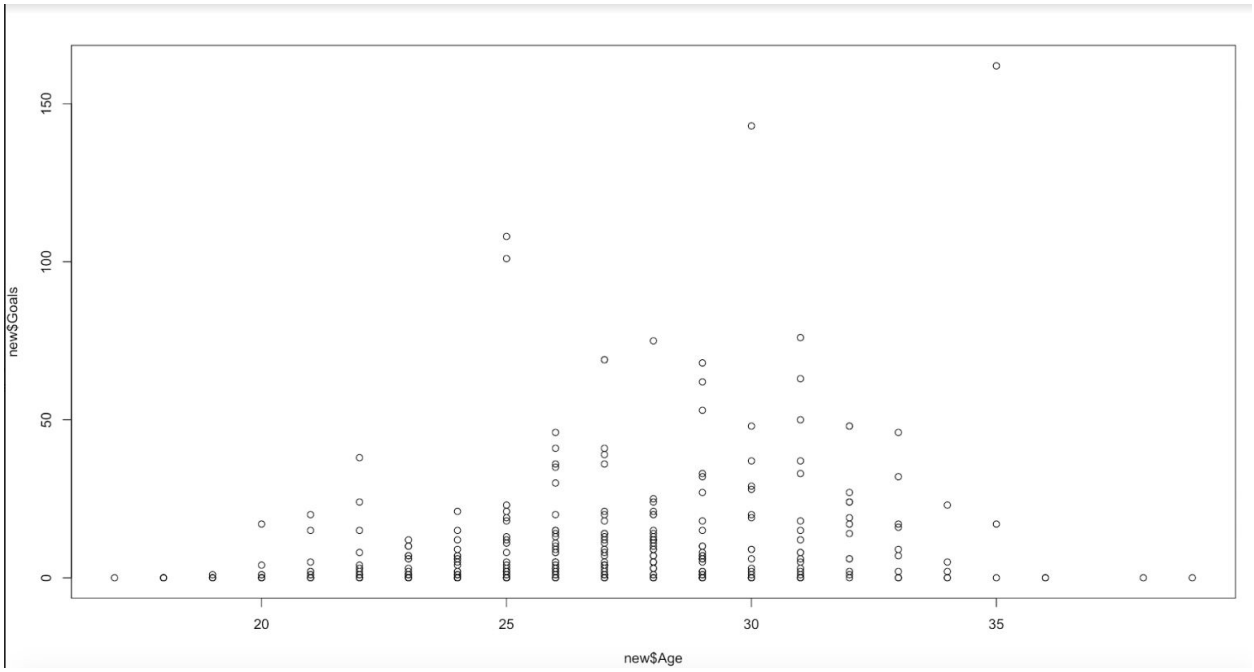
## Goals scored by players belonging to different positions




We can see that X-axis represents the number of goals scored and y-axis represents the number of players with that particular goal count. We can see that defenders and goalkeepers are the

highest in number who scored zero goals. Midfielders scored mostly less than 50goals and forwards scored huge number of goals such as 100, 150, etc.

## Goals scored by players of different age




This scatterplot shows that players between the ages of 25 and 35 scored the highest number of goals. There are some outliers who have scored more than 100 goals who are between 30 and 35. Those players might be messi and ronaldo.Besides that , most players between the ages 25 and 35 have scored 50 or less than 50 goals.

# 4. Model Proposals

Using the player stats dataset, we will predict the transfer value of all the players who belong to different positions, teams, and nationalities by building a regression model. Different variables play their own role in determining transfer value of players depending on their position. Variables such as tackles won, aerial threats won, one on one's, etc. will play a crucial role in predicting the transfer value of a defender. Similarly, chances created, passes completed, etc. play a vital role when predicting the value of midfielders. Big chances missed, big chances created, etc. are very crucial when signing star forwards. Also, variables such as punches and clean sheets are important for goalkeepers. Variables such as age, nationality, league, club, etc. affect the value of all players irrespective of their positions. Hence there are many independent variables which will help us predict the transfer value of the soccer players.

Jilei will focus on the prediction of midfielder's market value. Midfielders are players playing between defenders and forwards. Their position is primarily in the middle of the field and their goal is primarily connecting defense and the attack. Team-play stats such as Number of Passes, Passing Accuracy, Through Ball, assists etc., are more important than other stats categories, for evaluating a midfielder's performance. For example, through balls (or through pass) are important for midfielder since such pass can deliver the ball through the gap between the opponent's defenders during defense-to-attack transition and open space for forwards to run and score. Midfielders that can pass a lot of accurate through balls could potentially be more valuable. Running Distance is another important stats for midfielders as it reflects a player's work rate, fitness level and attitude, but it is not in our data currently. Jilei will try to find and integrate this data into the merged data and it could also be an important addition for other group members' analysis. Jilei will then perform a deep level analysis into the teamplay stats in combine with other available stats and will build a market value prediction model for the midfielders.

David will focus on the prediction of goalkeepers. Goalkeepers are unique in soccer since they are the only ones who can use their hands in open play. Therefore, most of the data that deals with goalkeepers are relevant only for goalkeepers, such as penalties saved, and goals conceded, and these data values are NA for outfield players such as midfielder, defender, and forward. David will be using mostly these data values in this model, but he will consider other data that applied to players in other positions as well, such as passes completed. Some of the data that applies to players in outfield positions, such as aerial battles, aren't available for goalkeepers. Goalkeepers have the smallest sample size (only 28 have available transfer values) since teams only have one goalkeeper that they use in every match, so he'll try to do n fold cross validation on the data.

Amanuel will focus on the prediction of defenders. As the name implies, the main goal of defenders is to stop any penetration. They do everything they can like tackling, which is dangerous, because they could get injured or get a fault (yellow or red car). A good soccer defender is one who has a strong technical skills, mental toughness, and physical fitness. Since defenders are the last position, they need to be very fast and good decision makers. Bedside defending, they are also expected to perform throw-ins, goal kicks, and corner kicks on their own side of the field. The defense positions can be break down into at least four positions. Central-back position player stops opponent players from shooting on the goal. Sweeper plays as the last defensive measures when a team decides not to four defenders line ups. These defenders don't go past the midfield line, but they can go as far back as their own goal line. Fullback position is another defense position which has left and right back position on either side of the center back of the field. Their main goal is to defend against opposing wingers. The last defense position is the wing back. Like the full-back defenders these players also divided into a left or right-back the goal of covering the opposing wing-back when on defense and support the midfield when on the attack. The value of defenders can be predicted with the dependent variables such as goals conceded, tackles, tackle success, blocked shots, and last man tackles etc.

Kaushik will focus on the prediction of the transfer value of the forwards in the dataset. Forwards include left-wingers, strikers, right-wingers. There are more than 100 forwards in the dataset. Forwards are the goal-scoring players who have the highest number of goals scored when combined. Player status such as the number of goals, assists, headed goals, chances created, etc. are more significant in deciding a forward's transfer value. For example, Relations such as the number of games played, and the number of goals scored are the most important when the transfer value of a player is being predicted. Kaushik will try to merge the data of all forwards and analyze the player stats to check which play a crucial role and create a transfer value predictive model of the forwards and their trends.

Therefore, we will be using the appropriate player stats given in the dataset to determine the transfer value of a player. Each of us has split the players based on their positions and analyze the data to check which player stats affect the value the most. We will build regression models individually for the players we segregated. Our main goal would be to predict the transfer value of a player and how it varies from the real transfer value. We expect the transfer value of forwards to be higher than midfielders, defenders, and goalkeepers on an average as that's the usual case in soccer. Many team stats will affect the value of a player and it is going to be a very interesting analysis as there are a lot of independent variables which will affect the value in different ways.

# 5. David Guo: Goalkeeper Model

When building a model for goalkeepers, one important choice I needed to make is to utilize my domain knowledge to consider what data I should include in my model that apply specifically to goalkeepers. Since the data has 94 variables, I have to choose a lot of variables to remove in order to have a functioning model. I will first consider player data that only applies to goalkeepers, such as clean sheets, but I will consider other data that applies to players in other positions as well, such as passes. Also, the data is split into information about a player's 2018-2019 season and data about a player's career, so I need to take both into account.

I first removed all rows that didn't have the position of goalkeeper. I also removed all the variables that applied only to outfield players that contained NAs for goalkeepers. I also ignored a lot of the variables that I felt weren't relevant to goalkeepers and didn't contribute to the market value of a goalkeeper. For example, I didn't consider goals as one of the factors, as goalkeepers almost never score goals due to their distance from the opposing goal, and if it happens it's usually fluke or lucky goal. People don't factor in a goalkeeper's goals scored in the assessment of a goalkeeper. The same thing applies to assists, since goalkeepers are not expected to make the final pass to score a goal, or penalty kicks, since attackers or midfielders are almost always the ones to take penalty kicks.

The variables I selected are:

Seasonal Data: GA18 (Goals Against for the 18-19 season), GA90.18 (Goals Against Per 90 minutes, SoTA18 (shots on target against), Save, CS_A18 (clean sheets against), CS18(clean sheets per 90 minutes), Min18 (minutes played)

Total Career Data: Penalties.saved, Punches, High.Claims, Catches, Goals.conceded, Clean.sheets, Errors.leading.to.goal, Accurate.long.balls, Passes, Passes.per.match, Age

## First Model

```
> gkmodel0 <- lm(Transfer.Values ~ GA18+GA90.18+SoTA18+Save18+CS_A18+CS18+Penalties.saved
+              +Punches+High.Claims+Catches+Goals.conceded+Clean.sheets+Errors.leading.to.goal+Accurate.long.balls+Passes+Passes.per.match+
Min18+ Age, data = gk)
```

```
> summary(gkmodel0)

Call:
lm(formula = Transfer.Values ~ GA18 + GA90.18 + SoTA18 + Save18 +
    CS_A18 + CS18 + Penalties.saved + Punches + High.Claims +
    Catches + Goals.conceded + Clean.sheets + Errors.leading.to.goal +
    Accurate.long.balls + Passes + Passes.per.match + Min18 +
    Age, data = gk)

Residuals:
      Min        1Q    Median        3Q       Max
-12783597  -4442991   -288010   5639734   9790067

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             14532377   79854902   0.182  0.85963
GA18                    -1083199    1259625  -0.860  0.41214
GA90.18                 14779499   12477575   1.184  0.26656
SoTA18                     50590     406774   0.124  0.90376
Save18                  82833063  108828868   0.761  0.46605
CS_A18                   1497607    2513206   0.596  0.56593
CS18                      238308     321937   0.740  0.47802
Penalties.saved          1337406    4988755   0.268  0.79468
Punches                  -531248     222096  -2.392  0.04043 *
High.Claims               121726     144071   0.845  0.42006
Catches                    15755     253494   0.062  0.95180
Goals.conceded           -217893     189736  -1.148  0.28041
Clean.sheets             1323158     381943   3.464  0.00711 **
Errors.leading.to.goal    556958    1745504   0.319  0.75695
Accurate.long.balls      -862093     858518  -1.004  0.34153
Passes                  -1015023     838182  -1.211  0.25674
Passes.per.match          901317     849057   1.062  0.31609
Min18                      15844      16984   0.933  0.37523
Age                     -3403582    1234926  -2.756  0.02225 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11620000 on 9 degrees of freedom
Multiple R-squared:  0.8819,    Adjusted R-squared:  0.6456
F-statistic: 3.733 on 18 and 9 DF,  p-value: 0.02432
```

```
> mean(residuals(gkmodel0)^2)
[1] 4.343325e+13
```

The adjusted r squared value is 0.6456, which means that 64.56% of the variability in the transfer values can be explained by the model. The F value is 3.733 on 18 and 9 degrees of freedom, and the p value for the F statistic is 0.02432, which means that under a significance level of 0.05, we can conclude that at least one independent variable is significant. The mean squared error is 4.34*10^13, so the data is very far scattered from the regression line.

Most of the predictors, such as High.Claims and Catches, were not significant under a significance level of 0.05 as they were mostly all greater than 0.05. Only Clean.sheets (0.04043), Punches (0.00711), and Age (0.02225), were significant under a significance level of 0.05. Given the large difference between adjusted r squared and multiple r squared (0.6456 vs 0.8819) and the high number of insignificant independent variables, I used forward and backward selection to remove some variables from the model.

# Backward Selection Model

```
stepbackgk <- stepAIC(gkmodel0, direction = "backward")
```

```
Step:  AIC=908.86
Transfer.Values ~ GA18 + GA90.18 + Save18 + CS_A18 + Punches +
    High.Claims + Goals.conceded + Clean.sheets + Accurate.long.balls +
    Passes + Passes.per.match + Age

                     Df  Sum of Sq        RSS    AIC
<none>                               1.3828e+15 908.86
- High.Claims          1 1.4429e+14 1.5271e+15 909.64
- Goals.conceded       1 1.8950e+14 1.5723e+15 910.46
- GA90.18              1 2.4651e+14 1.6293e+15 911.45
- Accurate.long.balls  1 3.4255e+14 1.7254e+15 913.06
- Passes.per.match     1 3.6577e+14 1.7486e+15 913.43
- GA18                 1 3.6939e+14 1.7522e+15 913.49
- Passes               1 3.9183e+14 1.7746e+15 913.84
- Save18               1 4.8513e+14 1.8679e+15 915.28
- CS_A18               1 1.7959e+15 3.1787e+15 930.17
- Age                  1 1.8625e+15 3.2453e+15 930.75
- Punches              1 2.0326e+15 3.4154e+15 932.18
- Clean.sheets         1 2.1575e+15 3.5403e+15 933.18
```

```
> summary(stepbackgk)

Call:
lm(formula = Transfer.Values ~ GA18 + GA90.18 + Save18 + CS_A18 +
    Punches + High.Claims + Goals.conceded + Clean.sheets + Accurate.long.balls +
    Passes + Passes.per.match + Age, data = gk)

Residuals:
      Min        1Q    Median        3Q       Max
-11557612  -4424696   -941805   5337668  13831922

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -291457   38744717  -0.008 0.994097
GA18                -333579     166644  -2.002 0.063739 .
GA90.18            10736038    6565405   1.635 0.122804
Save18            108385801   47247414   2.294 0.036646 *
CS_A18              3725092     843984   4.414 0.000503 ***
Punches             -532072     113313  -4.696 0.000287 ***
High.Claims          125886     100622   1.251 0.230069
Goals.conceded      -180040     125575  -1.434 0.172165
Clean.sheets        1368256     282830   4.838 0.000217 ***
Accurate.long.balls  -832990     432130  -1.928 0.073054 .
Passes              -1089046     528243  -2.062 0.057014 .
Passes.per.match      860285     431888   1.992 0.064909 .
Age                 -3010428     669754  -4.495 0.000428 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9601000 on 15 degrees of freedom
Multiple R-squared:  0.8657,    Adjusted R-squared:  0.7582
F-statistic: 8.057 on 12 and 15 DF,  p-value: 0.0001566
```

```
> mean(stepbackgk$residuals^2)
[1] 4.938594e+13
```

## Forward Selection Model

```
gkmodel1 <- lm(Transfer.Values ~1, data = gk)

stepforwardgk <- stepAIC(gkmodel1,direction="forward", scope=list(upper=gkmodel0,lower=gkmodel1))

Step:  AIC=908.66
Transfer.Values ~ CS_A18 + Age + Clean.sheets + Punches + Catches +
    GA18

                         Df  Sum of Sq        RSS    AIC
<none>                                 2.1079e+15 908.66
+ Errors.leading.to.goal  1 7.0774e+13 2.0372e+15 909.71
+ Penalties.saved         1 5.9499e+13 2.0484e+15 909.86
+ Min18                   1 4.9746e+13 2.0582e+15 910.00
+ Save18                  1 3.5817e+13 2.0721e+15 910.18
+ SoTA18                  1 3.3324e+13 2.0746e+15 910.22
+ CS18                    1 2.0736e+13 2.0872e+15 910.39
+ Goals.conceded          1 1.4483e+13 2.0935e+15 910.47
+ Passes                  1 1.3243e+13 2.0947e+15 910.49
+ GA90.18                 1 4.9723e+12 2.1030e+15 910.60
+ Accurate.long.balls     1 1.0484e+12 2.1069e+15 910.65
+ Passes.per.match        1 4.3618e+11 2.1075e+15 910.66
+ High.Claims             1 6.9389e+10 2.1079e+15 910.66
```

```
> summary(stepforwardgk)

Call:
lm(formula = Transfer.Values ~ CS_A18 + Age + Clean.sheets + 
    Punches + Catches + GA18, data = gk)

Residuals:
      Min        1Q    Median        3Q       Max 
-17788267  -6827845     17721   5628230  18911234 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  64455989   17378618   3.709 0.001300 ** 
CS_A18        3088657     713453   4.329 0.000296 ***
Age          -2148941     549688  -3.909 0.000807 ***
Clean.sheets   954280     182574   5.227 3.52e-05 ***
Punches       -319353      93690  -3.409 0.002644 ** 
Catches       -130294      72854  -1.788 0.088145 .  
GA18          -249668     150738  -1.656 0.112528    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10020000 on 21 degrees of freedom
Multiple R-squared:  0.7953,    Adjusted R-squared:  0.7368 
F-statistic:  13.6 on 6 and 21 DF,  p-value: 2.783e-06
```

```
> mean(stepforwardgk$residuals^2)
[1] 7.528348e+13
```

For the backwards selection model, I got a model that consists of: GA18, GA90.18, Save, CS_A18, CS18, Punches, High.Claims, Goals.conceded, Clean.sheets, Accurate.long.balls, Passes, Passes.per.match, Min, and Age. The adjusted r squared value for the forward selection model is 0.7582, so which means that 75.82% of the variability in the transfer values can be explained by the model. The F value is 8.057 on 12 and 15 degrees of freedom, and the p value for the v statistics is 0.0001566, which means that under a significance level of 0.05, we can conclude that at least one independent variable is significant. For the t test, most of the independent variables are not significant under a significance level of 0.05, with Save, CS_A, Punches, Clean.Sheets, and Age, being significant under a hypothesis test of 0.05 and the other variables not significant. The MSE is 4.94*10.^13, which is higher than the MSE for the first model.

For the forward selection model, I got a model that consists of CS_A18, Age, Clean.sheets, Punches, Catches, and GA18. The adjusted r squared value for the forward selection model is 0.7368, so which means that 73.68% of the variability in the transfer values can be explained by the model. The F value is 13.6 on 6 and 21 degrees of freedom, and the p value for the v statistics is 2.783e-06, which means that under a significance level of 0.05, we can conclude that at least

one independent variable is significant. For the t test, the dependent variables, CS_A, Punches, Clean.Sheets, and Age, are significant under a hypothesis test of 0.05 and while the other variables, catches and GA, are not. The MSE is 7.52 * 10^13, which is higher than the MSE for the first model and the backwards model.

It looks like I'm getting different results for backwards and forwards selection. The AIC is about the same for both models (908.86 vs 908.66), and the adjusted r squared values are similar (0.7582 vs 0.7368), but the forwards selection model model is less complex than the other, so I think that the forward model is my preferred model and I'll use that as a basis to build future models.

## Modified Model

```
> gkmodel3 <- lm( Transfer.Values ~ CS_A18 + Age + Clean.sheets + Punches +  Catches, data = gk)
> summary(gkmodel3)

Call:
lm(formula = Transfer.Values ~ CS_A18 + Age + Clean.sheets + 
    Punches + Catches, data = gk)

Residuals:
      Min        1Q    Median        3Q       Max 
-18308641  -6197712   -505232   7019763  24066847 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  58425342   17653362   3.310 0.003189 **
CS_A18        2227700     507681   4.388 0.000234 ***
Age          -2008641     564231  -3.560 0.001753 **
Clean.sheets   918755     188356   4.878 7.09e-05 ***
Punches       -290237      95603  -3.036 0.006068 **
Catches       -162205      72991  -2.222 0.036861 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10410000 on 22 degrees of freedom
Multiple R-squared:  0.7685,    Adjusted R-squared:  0.7159 
F-statistic: 14.61 on 5 and 22 DF,  p-value: 2.305e-06
```

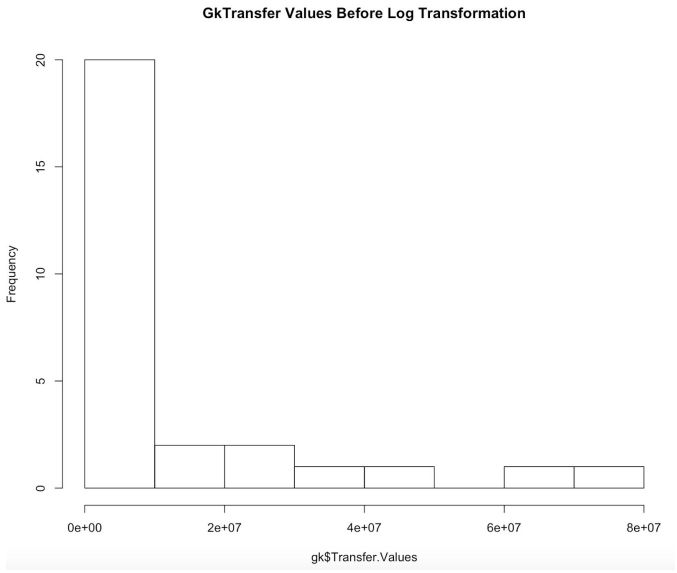
```
> mean(residuals(gkmodel3)^2)
[1] 8.511813e+13
> mc <-  data.frame(gk$CS_A18, gk$Age,gk$Clean.sheets, gk$Punches, gk$Catches)
> cor(mc)
                  gk.CS_A18     gk.Age gk.Clean.sheets gk.Punches gk.Catches
gk.CS_A18        1.00000000 -0.3349071      0.07387328  0.1267256 0.02031742
gk.Age          -0.33490715  1.0000000      0.29705575  0.2457855 0.35979746
gk.Clean.sheets  0.07387328  0.2970557      1.00000000  0.9283514 0.88919725
gk.Punches       0.12672560  0.2457855      0.92835145  1.0000000 0.86408260
gk.Catches       0.02031742  0.3597975      0.88919725  0.8640826 1.00000000
```

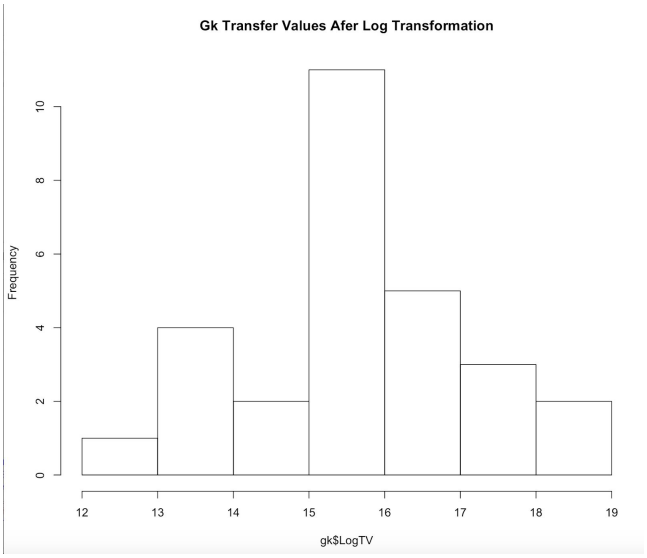
```
> vif(gkmodel3)
      CS_A18          Age Clean.sheets      Punches      Catches 
        1.20         1.34         9.27         7.95         5.38 
```

For my next model, I removed the insignificant variable GA18 from the forwards selection model. This led to the model having every predictor as significant under a significance level of 0.05, and the MSE decreasing to 8.51*10^13, and the adjusted r squared decreasing to 0.7159. The correlation between Punches, Catches, and Clean.sheets is quite high, at above 0.85, so there might be some multicollinearity there, while none of the variables are above ten on the VIF calculation, the VIF for Clean.sheets is pretty high at 9.27.

## Log Transformations

I transformed the transfer value data using a log transformation, creating a LogTV variable. The transfer values already look pretty right skewed, and transfer values can differ than hundreds of millions of dollars. By transforming the data, I can look at the transfer values in terms of magnitude, which can help the transfer value distribution appear less skewed.



```
> gk$LogTV <- log(gk$Transfer.Values)
> gkmodel4 <- lm( LogTV ~ CS_A18 + Age + Clean.sheets + Punches +  Catches, data = gk)
> summary(gkmodel4)

Call:
lm(formula = LogTV ~ CS_A18 + Age + Clean.sheets + Punches +
    Catches, data = gk)

Residuals:
     Min       1Q   Median       3Q      Max
-1.31772 -0.27698 -0.06421  0.21363  1.20200

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.088661   0.964524  20.828 5.70e-16 ***
CS_A18        0.173159   0.027738   6.243 2.78e-06 ***
Age          -0.191122   0.030828  -6.200 3.07e-06 ***
Clean.sheets  0.035684   0.010291   3.467  0.00219 **
Punches      -0.008472   0.005223  -1.622  0.11905
Catches      -0.005018   0.003988  -1.258  0.22147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5687 on 22 degrees of freedom
Multiple R-squared:  0.8628,    Adjusted R-squared:  0.8316
F-statistic: 27.67 on 5 and 22 DF,  p-value: 8.539e-09

> mean(residuals(gkmodel4)^2)
[1] 0.2540936
```

I created a model after transforming the values. Doing a log transformation on the values increased the value of the adjusted R squared to 0.8316 from 0.7159, which means that 83.16% of the variability in transfer values can be explained by the model. The F statistic is 27.67, which is significant at a confidence level of 0.05. Due to the log transformation, the MSE decreased to 0.254. For the t test, punches and catches and not significant under a significance level of 0.05, while the other three variables are.

```
> summary(gkmodel5)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age, data = gk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3430 -0.3984  0.0188  0.1746  1.2245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.30492    1.01757   19.95  < 2e-16 ***
CS_A18        0.16755    0.02945    5.69  7.3e-06 ***
Clean.sheets  0.01286    0.00387    3.32   0.0028 **
Age          -0.19758    0.03233   -6.11  2.6e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.61 on 24 degrees of freedom
Multiple R-squared:  0.828,     Adjusted R-squared:  0.806
F-statistic: 38.4 on 3 and 24 DF,  p-value: 2.54e-09
```

```
> mc <- data.frame(gk$CS_A18, gk$Clean.sheets, gk$Age)
> cor(mc)
                gk.CS_A18 gk.Clean.sheets gk.Age
gk.CS_A18          1.0000          0.0739 -0.335
gk.Clean.sheets    0.0739          1.0000  0.297
gk.Age            -0.3349          0.2971  1.000
> gkmodel5 <- lm(Transfer.Values ~ CS_A18 + Punches +
+                  Clean.sheets +  Age, data = gk)
> vif(gkmodel5)
      CS_A18      Punches Clean.sheets          Age
        1.19         7.43         7.53         1.28
```

```
> mean(resid(gkmodel5)^2)
[1] 0.319
```

I created a model that only included the three significant terms after transforming the transfer values. Compared to gkmodel4, the adjusted R squared decreased from 0.8316 from 0.806. The F statistic increased to 38.4, and is significant at a confidence level of 0.05. Due to the log transformation, the MSE increased to 0.319, For the t test, CS_A18, Clean.sheets, and Age, were significant under a 0.05 significance level t-test.

Looking at the vif, the three variables have small vif values, so there is little multicollinearity. The correlations between the variables also aren’t very high.

## Second Order Analysis

After this, I checked for second order terms. I think age can be used as a second order term, as I think there is a parabolic relationship between transfer value and age. A young player with not much experience would be not worth much as he is just starting his career, but his transfer value would increase as he becomes more a veteran player, but his value would decrease as he ages and passes his prime. Punches and catches could also be second order terms, as a goalkeeper might be more valuable if does more punches and catches in a game, but doing it too much may interfere with the other aspects of his game and causing him to be less valuable.

```
gk$Age2 <- gk$Age*gk$Age
gk$Catches2 <- gk$Catches*gk$Catches
gk$Punches2 <- gk$Punches*gk$Punches
```

```
> summary(gkmodel6)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age + Age2, data = gk)

Residuals:
   Min     1Q Median     3Q    Max
-1.409 -0.385 -0.004  0.165  1.153

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.20728    5.02313    3.43   0.0023 **
CS_A18        0.16791    0.02983    5.63  9.9e-06 ***
Clean.sheets  0.01229    0.00402    3.06   0.0056 **
Age           0.00948    0.33026    0.03   0.9773
Age2         -0.00338    0.00537   -0.63   0.5349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.618 on 23 degrees of freedom
Multiple R-squared:  0.831,     Adjusted R-squared:  0.801
F-statistic: 28.2 on 4 and 23 DF,  p-value: 1.43e-08
```

```
> gkmodel6 <- lm(LogTV ~ CS_A18 +
+                  Clean.sheets + Age+ Punches + Punches2, data = gk)
> summary(gkmodel6)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age + Punches +
    Punches2, data = gk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2731 -0.3424  0.0233  0.2111  1.2416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.04e+01   1.02e+00   20.06  1.2e-15 ***
CS_A18        1.73e-01   3.00e-02    5.76  8.6e-06 ***
Clean.sheets  2.88e-02   1.06e-02    2.71    0.013 *
Age          -2.02e-01   3.29e-02   -6.15  3.4e-06 ***
Punches      -8.02e-03   9.42e-03   -0.85    0.404
Punches2     -6.77e-06   2.48e-05   -0.27    0.787
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.588 on 22 degrees of freedom
Multiple R-squared:  0.853,     Adjusted R-squared:  0.82
F-statistic: 25.6 on 5 and 22 DF,  p-value: 1.74e-08

> summary(gkmodel6)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age + Catches +
    Catches2, data = gk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3222 -0.3341 -0.0537  0.2201  1.3063

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.03e+01   1.03e+00   19.64  1.9e-15 ***
CS_A18        1.64e-01   2.86e-02    5.73  9.1e-06 ***
Clean.sheets  2.15e-02   8.15e-03    2.64    0.015 *
Age          -1.98e-01   3.37e-02   -5.88  6.5e-06 ***
Catches      -1.70e-04   7.65e-03   -0.02    0.983
Catches2     -2.47e-05   2.48e-05   -1.00    0.329
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.589 on 22 degrees of freedom
Multiple R-squared:  0.853,     Adjusted R-squared:  0.82
F-statistic: 25.5 on 5 and 22 DF,  p-value: 1.79e-08
```

I included the second order terms for punches, catches, and age into the model. The model with the second order terms show not much improvement in the adjusted r squared and f statistic, and mean squared error. None of the second order terms are significant as well. I'll remove the second order terms, as well as the first order terms for punches and catches, and look for interaction terms.

## Interaction Term Analysis

I checked for interaction terms between saves and clean sheets for both the season and career. I hypothesized that there is an interaction between saves and clean sheets, as while someone with more clean sheets will have a higher transfer value, someone with more saves but a similar amount of clean sheets could have a higher transfer value. I hypothesize the same thing could apply to clean sheets and penalties saved.

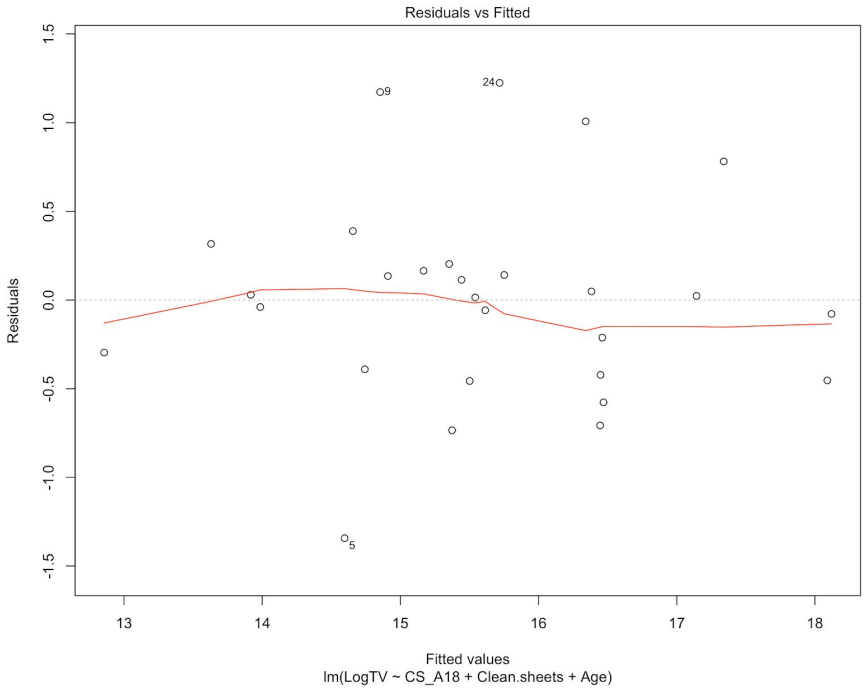
```
> gkmodel6 <- lm(LogTV ~ CS_A18 +
+                  Clean.sheets + Age + Clean.sheets*Penalties.saved, data = gk)
> summary(gkmodel6)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age + Clean.sheets *
    Penalties.saved, data = gk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3379 -0.3231 -0.0303  0.2193  1.3128

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  20.15212    1.21557   16.58  6.5e-14 ***
CS_A18                        0.16314    0.02903    5.62  1.2e-05 ***
Clean.sheets                  0.02595    0.00766    3.39  0.00265 **
Age                          -0.19426    0.04180   -4.65  0.00012 ***
Penalties.saved               0.00266    0.15507    0.02  0.98648
Clean.sheets:Penalties.saved -0.00302    0.00182   -1.66  0.11140
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.587 on 22 degrees of freedom
Multiple R-squared:  0.854,     Adjusted R-squared:  0.821
F-statistic: 25.7 on 5 and 22 DF,  p-value: 1.68e-08
```

```
> gkmodel6 <- lm(LogTV ~ CS_A18 +
+                  Clean.sheets + Age + Clean.sheets*Saves, data = gk)
> summary(gkmodel6)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age + Clean.sheets *
    Saves, data = gk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2938 -0.3049 -0.0711  0.2592  1.3678

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.06e+01   1.05e+00   19.55  2.1e-15 ***
CS_A18             1.66e-01   2.91e-02    5.69  1.0e-05 ***
Clean.sheets       3.52e-02   1.43e-02    2.46    0.022 *
Age               -2.11e-01   3.48e-02   -6.07  4.2e-06 ***
Saves             -2.01e-04   1.07e-03   -0.19    0.853
Clean.sheets:Saves -2.64e-05   1.57e-05   -1.68    0.107
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.598 on 22 degrees of freedom
Multiple R-squared:  0.849,     Adjusted R-squared:  0.814
F-statistic: 24.7 on 5 and 22 DF,  p-value: 2.48e-08
```

```
> gkmodel6 <- lm(LogTV ~ CS_A18 +
+                  Clean.sheets + Age + CS_A18*Save18, data = gk)
> summary(gkmodel6)

Call:
lm(formula = LogTV ~ CS_A18 + Clean.sheets + Age + CS_A18 * Save18,
    data = gk)

Residuals:
   Min     1Q Median     3Q    Max
-1.317 -0.370  0.023  0.223  1.256

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.75444    1.36011   15.26  3.5e-13 ***
CS_A18        -0.18584    0.36898   -0.50   0.6195
Clean.sheets   0.01293    0.00398    3.25   0.0037 **
Age           -0.20381    0.03369   -6.05  4.3e-06 ***
Save18        -0.39527    1.33102   -0.30   0.7693
CS_A18:Save18  0.51634    0.53888    0.96   0.3484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.622 on 22 degrees of freedom
Multiple R-squared:  0.836,     Adjusted R-squared:  0.798
F-statistic: 22.4 on 5 and 22 DF,  p-value: 5.96e-08
```

Looking at the added interaction terms, none of them were significant under a significance level of 0.05, so I didn’t include them in the model. My final model for predicting goalkeeper transfer values is gkmodel 5 with the dependent variable being the LogTV and the independent variables being CS_A18, Age, and Clean.Sheets.

## Residual Analysis




I analyzed the residuals for my final model, gkmodel5.

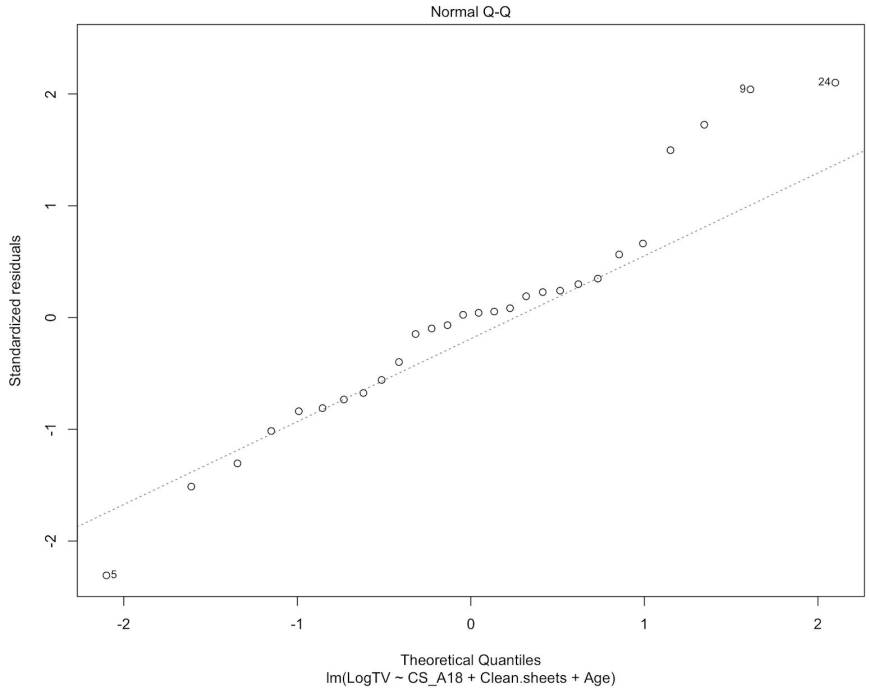
The residuals are pretty evenly distributed around the x axis, so the residual distribution meets the assumption of homoscedasticity.

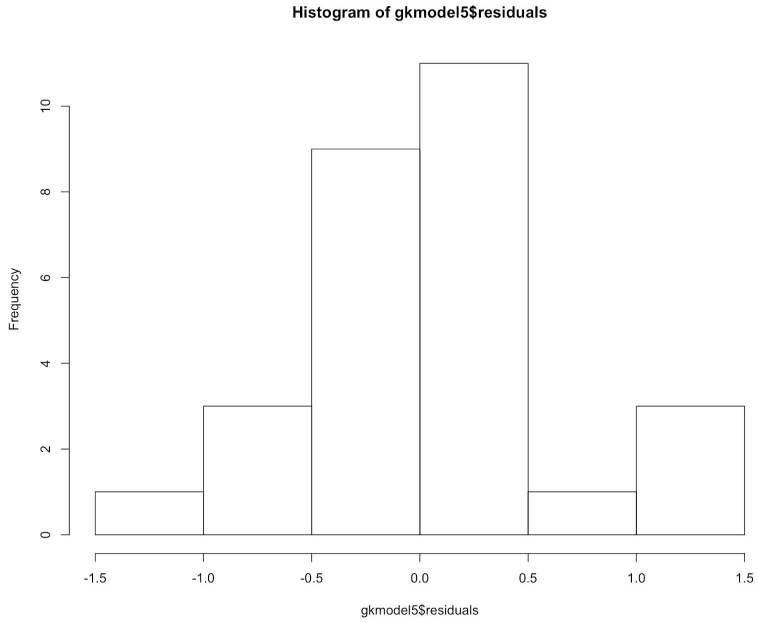
```
> mean(gkmodel5$residuals)
[1] 5.7e-18
```

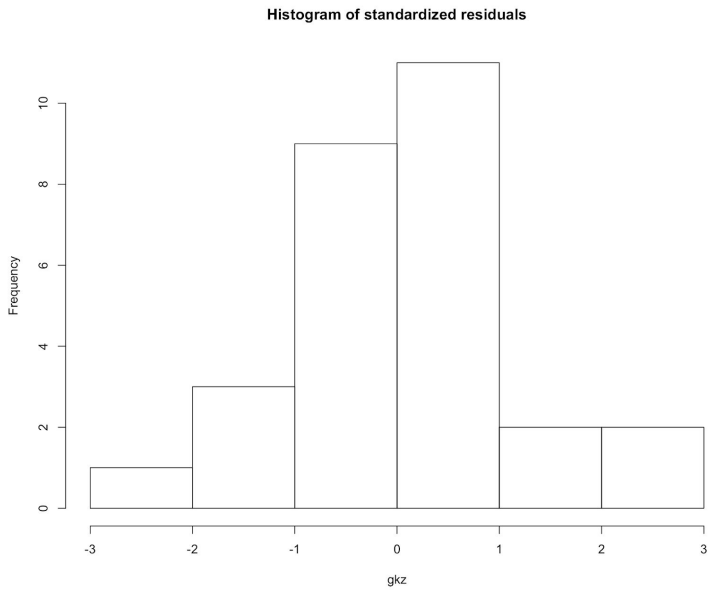
The mean of the residuals is 0, so the model meets the assumption of mean of residuals is zero.

```
> durbinWatsonTest(gkmodel5)
 lag Autocorrelation D-W Statistic p-value
   1          -0.127          2.24   0.478
 Alternative hypothesis: rho != 0
```

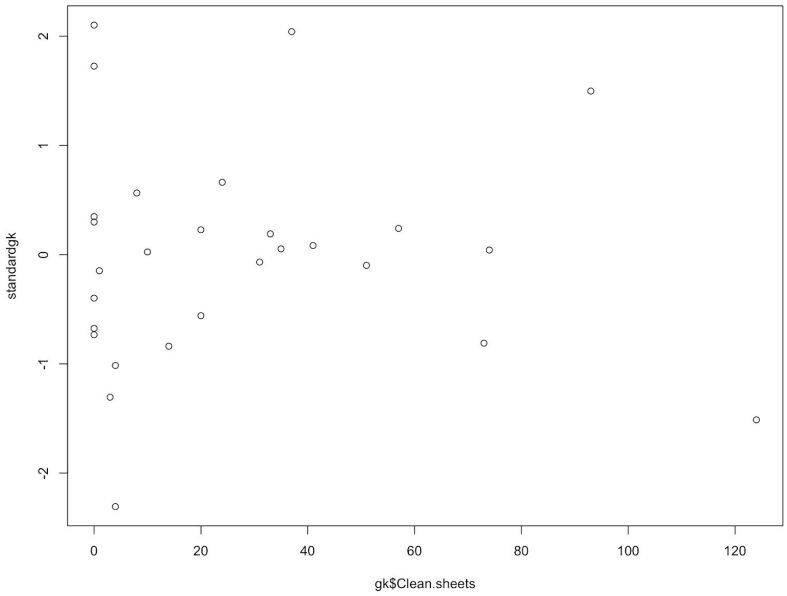
The Durbin Watson test 2.24 is close to the value of 2, the p value of the test is 0.478 which is higher than the significance level of 0.05, so we fail to reject the null hypothesis that the model doesn't have autocorrelation. I can assume the model meets the assumption that the residuals are independent.

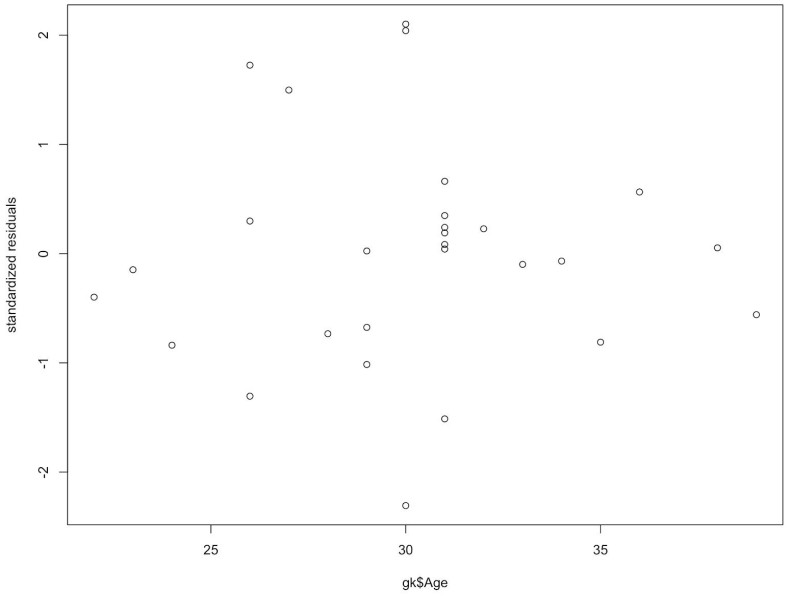

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(LogTV ~ CS_A18 + Clean.sheets + Age)
9
24
5
-2
-1
0
1
2

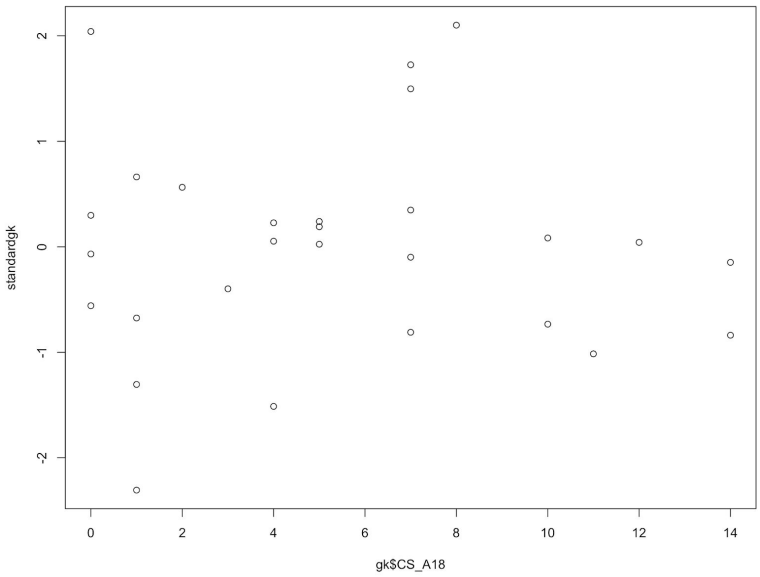

Histogram of gkmodel5$residuals
Frequency
gkmodel5$residuals
0
2
4
6
8
10
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5




The qqplot shows that the residuals are generally normal, close to the qqline, don't follow a pattern; though there are some values like 5, 9, and 24, that deviate from the line. The residual histograms looks mostly normal, if a little bit left skewed. I can assume the model meets the assumption that the residuals are normally distributed.


standardgk
gk$Clean.sheets
2
1
0
-1
-2
0
20
40
60
80
100
120

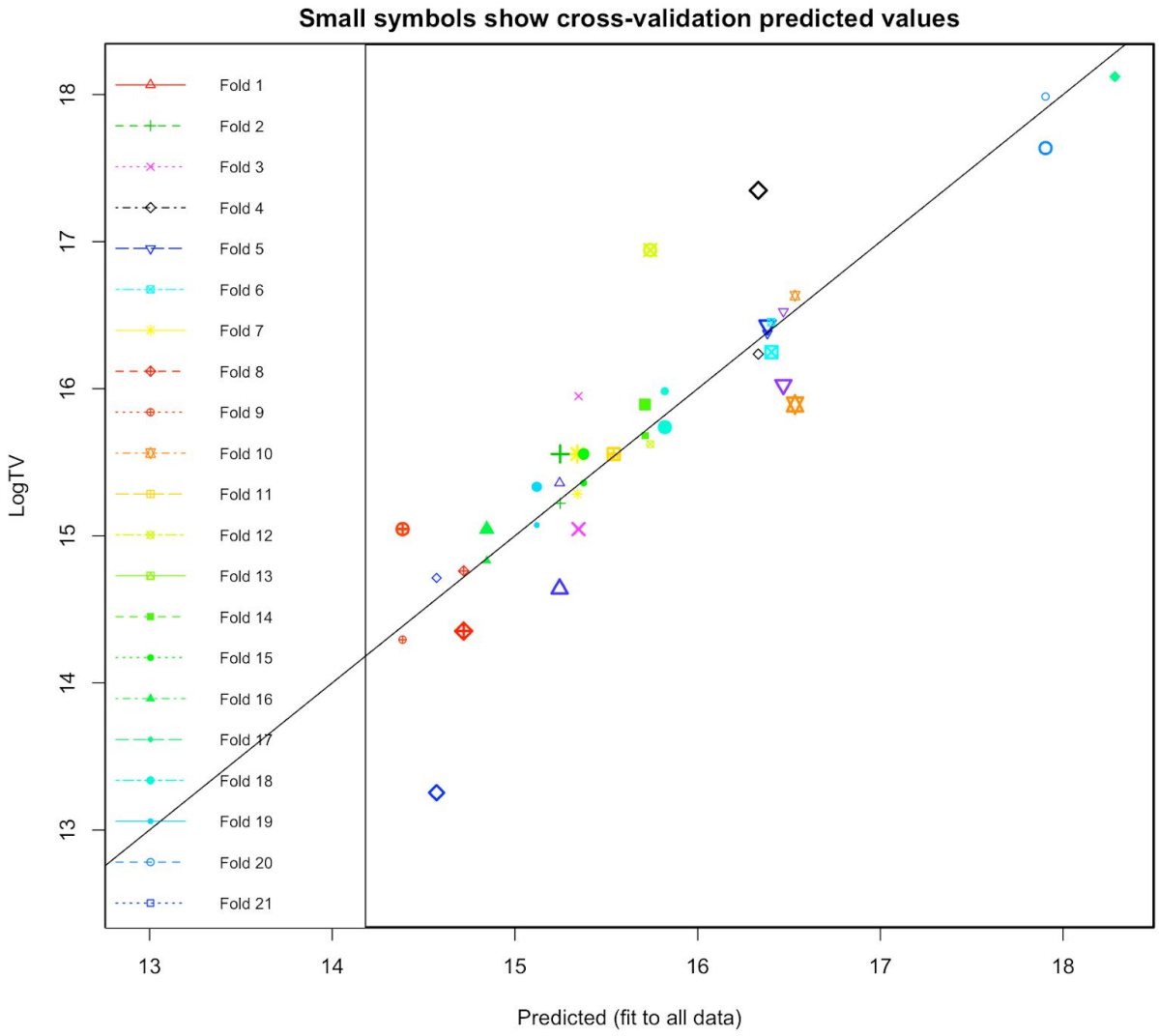

standardized residuals
gk$Age
2
1
0
-1
-2
25
30
35




I plotted the independent variables against the standardized residuals. Most of the residuals are between -1 and 1, and the values are pretty evenly scattered around the x axis. There is one notable outlier in the Clean.sheets plot, where the point that has clean sheets value of larger than 120 stretches out the look of the graph, though the standardized residual value is less than 2. There seems to be 3 points with standardized values of larger than 2, but two of them are pretty close to 2, so they're not big potential outliers. There looks to be about 5 values between 1 and 2 standard deviations above the mean, adding up to 8 values above one standard deviation, so it's pretty close to the expected 0.32*28 = 8.96 values above a standard deviation of 1.

## Cross Validation

```
gkmodel5 <- lm(LogTV ~ CS_A18 + Clean.sheets + Age, data = gk)

gkcv1 <- cv.lm(data = gk, form.lm = formula(gkmodel5), plotit = "Observed", m = 28)
abline(coef = c(0,1))

Overall (Sum over all 1 folds)
   ms
0.446

> cor(gkcv1$cvpred,gk$LogTV)
[1] 0.873
```




```
gkmodel6 <- lm(LogTV ~ CS_A18 +    Clean.sheets + Age + Punches + Punches2, data = gk)
gkmodel7 <- lm(LogTV ~ CS_A18 + Clean.sheets + Age + Clean.sheets*Penalties.saved, data = gk)
gkmodel8 <- lm(Transfer.Values ~(CS_A18 + Penalties.saved + Punches + Clean.sheets) , data = gk)
gkmodel9 <- lm(LogTV ~ 1 , data = gk)
```

```
gkcv2 <- cv.lm(data = gk, form.lm = formula(gkmodel6), plotit = "Observed", m = 28)
gkcv3 <- cv.lm(data = gk, form.lm = formula(gkmodel7), plotit = "Observed", m = 28)
gkcv4 <- cv.lm(data = gk, form.lm = formula(gkmodel8), plotit = "Observed", m = 28)
gkcv4 <- cv.lm(data = gk, form.lm = formula(gkmodel9), plotit = "Observed", m = 28)
```

```
Overall (Sum over all 1 folds)
   ms
0.435
```

```
Overall (Sum over all 1 folds)
   ms
0.393
```

```
Overall (Sum over all 1 folds)
     ms
2.1e+14
```

```
Overall (Sum over all 1 folds)
   ms
1.99
```

```
> cor(gkcv2$cvpred,gk$LogTV)
[1] 0.878
> cor(gkcv3$cvpred,gk$LogTV)
[1] 0.888
> cor(gkcv4$cvpred,gk$Transfer.Values)
[1] 0.667
```

Because I only have a sample size of 28, I conducted leave one out cross validation on the final model and on my previous models. For my final model, the average MSE for cross validation is 0.446, and the correlation between the predicted and actual data is 0.873. From the plot, it looks like most of the predicted values are close to the actual values and lie on the diagonal line. Since the average of the MSE of the null model is 1.99, the average MSE my final model is pretty close to the average MSE in the models that have insignificant second order terms or interaction terms (0.435, 0.393). This correlation between the predicted and actual data is better than the correlation model without the log transformation (0.667) and not much lower than the correlation for the models that have insignificant second order terms or interaction terms (0.878, 0.888).

# 6. Amanuel Petros: Defender Model

In this report, I will focus mainly on the defense position of the players. Defense players play near their own goal. These players are required to clear the ball, preventing the opposition from shooting, and they need to have a good passing Ability, heading Ability, and Speed. Based on the defender's main job and the ability or skills they need to have transfer value is likely influenced by the ability of their defense skills.

First the defense players data was selected out from the merged dataset using SQL. According to the position of players and their main goal on the field several choices could have been made. For example, the defense position can be further divided into different 3 defense position. However, since the data doesn't specifically indicate the specific positions, we chose to use the general defense position. Before doing any modeling, data was cleaned farther to ensure the data is free from NA's and NULL values. As a result of the cleaning process fourteen defense players were removed due to missing transfer values. Building a model for soccer player transfer value requires a good understanding about the game. Understanding the game helps in eliminating and including variables according to the players potions. It also helps for making decisions based on the domain knowledge. Based on the domain knowledge and correlations of the variable out of around 60 variables about 10 variables were selected for the final model. When a scatter plot was performed a relationship between several variables were shown.

When the data selected out from the merged dataset all the explanatory variables were also selected. If not all, many of these variables doesn't apply to the defense position. Even if some do the prediction process would not produce a sensible result. A defender is an outfield player whose primary role is to prevent the opposing team from scoring goals. Therefore, it is not reasonable to include goal scores in the model similarly tackling is more important for the assessment of defenders than for forwards. However, several variables are common across all models (e.g. age, appearance, win, lose etc..). The variables that are removed from the defense position in this model are below in table 1.1 Most of these variables are not defense variables. This means these variables will not produce a good model in predicting the transfer value of the defense position.

**Table 1.1**

| | | |
|---|---|---|
| SoT | Fls90 | Clearances_off_line |
| Gls90 | Crd | Successful_50_50s |
| G_A | Rk_y | Goals conceded |
| Gls | Club | Own_goals |
| Ast | Clean_sheets | Errors_leading_to_goal |
| PK | G_PK | Goals_with_right_foot |
| PKatt | G_A_PK | Goals_with_left_foot |
| Fls | SoT90 | Hit_woodwork |
| CrdY | CrdR | |

The final selected explanatory variables. Considered more important for the assessment of defenders are below in table 1.2

| | | |
|---|---|---|
| Full_Name | Duels_won | Aerial_battles_lost |
| Country | Duels_lost | Assists Passes |
| Pos1 | Aerial_battles_won | Passes_per_match |
| Squad | Losses | Big_chances_created |
| Age | Clean_sheets | Crosses |
| Born | Tackles | Cross_accuracy |
| Apps | Tackle_success | Through_balls |
| Starts | Blocked_shots | Accurate_long_balls |
| Subs | Interceptions | Yellow_cards |
| Min | Clearances | Red_cards |
| Mn_Ap | Headed_Clearance | Fouls |
| Fls | Recoveries | Offsides |
| Appearances | Goal conceded | Values |
| Wins | Last man tackles | |

These selected variables are defender's variable. Selected doesn't mean they were all used for the final model. What it means is that they were chosen to be used in the model for predicting the transfer value of the defender's position. However, several variables from this list had to be removed because of the presence of excessive NULL values and missing values.

Understandably, the rest of the other variables that are not in the final model are removed from the model because their prediction contribution is not significant.

```
> model.defense1 <-lm(defense$Values ~ defense$Age + defense$Starts + defense
$Subs + defense$Mn_Ap + defense$Appearances + defense$Wins + defense$Losses +
defense$Tackles + defense$Tackle_success + defense$Blocked_shots + defense$In
terceptions + defense$Headed_Clearance + defense$Recoveries + defense$Duels_l
ost + defense$Duels_won + defense$Aerial_battles_lost + defense$Aerial_battle
s_won + defense$Assists + defense$Passes + defense$Passes_per_match   + defen
se$Big_chances_created + defense$Errors_leading_to_goal)
> summary(model.defense1)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Subs + defense$Mn_Ap + defense$Appearances + defense$Wins +
    defense$Losses + defense$Tackles + defense$Tackle_success +
    defense$Blocked_shots + defense$Interceptions + defense$Headed_Clearance
+
    defense$Recoveries + defense$Duels_lost + defense$Duels_won +
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Assists + defense$Passes + defense$Passes_per_match +
    defense$Big_chances_created + defense$Errors_leading_to_goal)

Residuals:
      Min        1Q    Median        3Q       Max
-21941500  -8189901  -2242857   6289160  49180934

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.942e+07  1.418e+07   4.191 5.73e-05 ***
defense$Age                   -2.330e+06  4.159e+05  -5.601 1.67e-07 ***
defense$Starts                 5.871e+05  1.628e+05   3.606 0.000474 ***
defense$Subs                  -2.676e+05  6.107e+05  -0.438 0.662116
defense$Mn_Ap                 -1.604e+04  1.402e+05  -0.114 0.909161
defense$Appearances           -2.868e+04  2.796e+05  -0.103 0.918489
defense$Wins                   7.390e+04  3.268e+05   0.226 0.821559
defense$Losses                -3.731e+04  3.675e+05  -0.102 0.919341
defense$Tackles               -3.086e+04  4.685e+04  -0.659 0.511580
defense$Tackle_success        -1.494e+05  6.675e+04  -2.238 0.027307 *
defense$Blocked_shots          2.569e+05  1.502e+05   1.711 0.089932 .
defense$Interceptions         -1.823e+04  4.779e+04  -0.382 0.703547
defense$Headed_Clearance      -5.307e+04  3.002e+04  -1.768 0.079960 .
defense$Recoveries             6.798e+03  2.074e+04   0.328 0.743683
defense$Duels_lost            -2.089e+04  3.822e+04  -0.547 0.585859
defense$Duels_won              6.516e+01  4.273e+04   0.002 0.998786
defense$Aerial_battles_lost   -4.844e+04  7.206e+04  -0.672 0.502869
defense$Aerial_battles_won     9.943e+04  5.632e+04   1.765 0.080339 .
defense$Assists               -1.257e+06  6.115e+05  -2.055 0.042297 *
defense$Passes                 2.770e+03  2.618e+03   1.058 0.292375
defense$Passes_per_match       4.567e+05  1.236e+05   3.694 0.000350 ***
defense$Big_chances_created    1.566e+06  5.304e+05   2.952 0.003879 **
defense$Errors_leading_to_goal -4.659e+05  1.111e+06  -0.419 0.675731
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12990000 on 107 degrees of freedom
Multiple R-squared:  0.6738,   Adjusted R-squared:  0.6067
F-statistic: 10.05 on 22 and 107 DF,  p-value: < 2.2e-16
```

This is the first model with all variables I selected to predict transfer value of defense position. This model produced 0.6067 R-square. This means 60.67 of the variability in the transfer values could be explained by the model. the F-test is 10.05 on the degree of 22 and 107 degree of freedom. In this model: Age, start, tackle success, assists, pass per match, and big chance created are the only variables below the p-value of 0.05 the rest variable in this model are above the p-value of significance.

## Step forward

**model.defense2 <-lm(defense$Values ~ 1)**

**step <-stepAIC(model.defense2, direction = "forward", scope = list(upper = model.defense1, lower = model.defense2))**

**summary(stepforward)**

```
Step:  AIC=4261.8
defense$Values ~ defense$Age + defense$Starts + defense$Tackles +
    defense$Tackle_success + defense$Blocked_shots + defense$Headed_Clearance
+
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Assists + defense$Passes + defense$Passes_per_match +
    defense$Big_chances_created

                              Df  Sum of Sq        RSS    AIC
<none>                                      1.8390e+16 4261.8
- defense$Blocked_shots        1 6.7848e+14 1.9068e+16 4264.5
- defense$Headed_Clearance     1 7.7886e+14 1.9169e+16 4265.2
- defense$Passes               1 8.2840e+14 1.9218e+16 4265.5
- defense$Aerial_battles_lost  1 9.1148e+14 1.9301e+16 4266.1
- defense$Assists              1 9.2222e+14 1.9312e+16 4266.2
- defense$Tackles              1 1.1346e+15 1.9525e+16 4267.6
- defense$Tackle_success       1 1.1544e+15 1.9544e+16 4267.7
- defense$Aerial_battles_won   1 1.5310e+15 1.9921e+16 4270.2
- defense$Big_chances_created  1 1.6201e+15 2.0010e+16 4270.8
- defense$Passes_per_match     1 3.2652e+15 2.1655e+16 4281.0
- defense$Starts               1 5.2401e+15 2.3630e+16 4292.4
- defense$Age                  1 6.2549e+15 2.4645e+16 4297.9
```

```
summary(step)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Tackles + defense$Tackle_success + defense$Blocked_shots +
    defense$Headed_Clearance + defense$Aerial_battles_lost +
    defense$Aerial_battles_won + defense$Assists + defense$Passes +
    defense$Passes_per_match + defense$Big_chances_created)

Residuals:
      Min        1Q    Median        3Q       Max
-23293262  -8036377  -1504692   6400668  49360599

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                58437531    9726348   6.008 2.18e-08 ***
defense$Age                -2385075     378084  -6.308 5.21e-09 ***
defense$Starts               585548     101412   5.774 6.49e-08 ***
defense$Tackles              -52183      19423  -2.687  0.00827 **
defense$Tackle_success      -163597      60367  -2.710  0.00774 **
defense$Blocked_shots        250889     120756   2.078  0.03993 *
defense$Headed_Clearance     -54270      24380  -2.226  0.02793 *
defense$Aerial_battles_lost  -93397      38784  -2.408  0.01760 *
defense$Aerial_battles_won   105960      33950   3.121  0.00227 **
defense$Assists            -1343456     554630  -2.422  0.01696 *
defense$Passes                3126       1362   2.296  0.02347 *
defense$Passes_per_match     491021     107732   4.558 1.28e-05 ***
defense$Big_chances_created 1521956     474051   3.211  0.00171 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12540000 on 117 degrees of freedom
Multiple R-squared:  0.6678,   Adjusted R-squared:  0.6338
F-statistic:  19.6 on 12 and 117 DF,  p-value: < 2.2e-16
```

# Step Backward

```
Step:  AIC=4262.93
defense$Values ~ defense$Starts + defense$Passes_per_match +
    defense$Age + defense$Blocked_shots + defense$Tackle_success +
    defense$Duels_lost + defense$Passes + defense$Tackles + defense$Big_chanc
es_created +
    defense$Assists + defense$Aerial_battles_won + defense$Headed_Clearance

                                 Df  Sum of Sq        RSS    AIC
<none>                                         1.8551e+16 4262.9
+ defense$Aerial_battles_lost     1 2.2094e+14 1.8330e+16 4263.4
+ defense$Losses                  1 1.9322e+14 1.8358e+16 4263.6
+ defense$Wins                    1 1.7173e+14 1.8379e+16 4263.7
+ defense$Errors_leading_to_goal  1 8.4485e+13 1.8467e+16 4264.3
+ defense$Subs                    1 4.6303e+13 1.8505e+16 4264.6
+ defense$Appearances             1 4.6089e+13 1.8505e+16 4264.6
+ defense$Duels_won               1 1.8714e+13 1.8532e+16 4264.8
+ defense$Interceptions          1 1.3601e+13 1.8538e+16 4264.8
+ defense$Recoveries             1 7.8493e+12 1.8543e+16 4264.9
+ defense$Mn_Ap                  1 5.8531e+12 1.8545e+16 4264.9
```

```
summary(stepforward)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Subs + defense$Mn_Ap + defense$Appearances + defense$Wins +
    defense$Losses + defense$Tackles + defense$Tackle_success +
    defense$Blocked_shots + defense$Interceptions + defense$Headed_Clearance
+
    defense$Recoveries + defense$Duels_lost + defense$Duels_won +
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Assists + defense$Big_chances_created + defense$Fouls)

Residuals:
      Min        1Q    Median        3Q       Max
-18590164  -9153984  -3101419   5882400  52083892

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 60016426   15178947   3.954 0.000137 ***
defense$Age                 -2366092     450731  -5.249  7.6e-07 ***
defense$Starts                606846     174391   3.480 0.000723 ***
defense$Subs                 -238834     662423  -0.361 0.719137
defense$Mn_Ap                  33410     149730   0.223 0.823846
defense$Appearances           -43842     294315  -0.149 0.881858
defense$Wins                  197472     358648   0.551 0.583033
defense$Losses               -141779     396794  -0.357 0.721550
defense$Tackles                18298      52375   0.349 0.727485
defense$Tackle_success         23452      53417   0.439 0.661507
defense$Blocked_shots         284992     161798   1.761 0.080974 .
defense$Interceptions         -39591      50545  -0.783 0.435160
defense$Headed_Clearance      -48995      33527  -1.461 0.146795
defense$Recoveries             35905      19547   1.837 0.068956 .
defense$Duels_lost            -31961      49133  -0.651 0.516737
defense$Duels_won             -17874      46862  -0.381 0.703640
defense$Aerial_battles_lost   -28937      87244  -0.332 0.740771
defense$Aerial_battles_won    119064      62078   1.918 0.057729 .
defense$Assists             -1221578     678611  -1.800 0.074609 .
defense$Big_chances_created  1704677     614230   2.775 0.006492 **
defense$Fouls                 -51092      87254  -0.586 0.559385
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14140000 on 109 degrees of freedom
Multiple R-squared:  0.6066,   Adjusted R-squared:  0.5345
F-statistic: 8.405 on 20 and 109 DF,  p-value: 2.613e-14
```

**Forward and backward MSE**

```
> mean(step.f$residuals^2)
[1] 1.414615e+14
> mean(step.b$residuals^2)
[1] 1.427008e+14
```

In the step forward and step backward we can see clearly that the step forward is much better. In the Step backward the R-square is 0.5345 which means that 53.45 percent of the variability in the transfer values could be explained by the model. In this backward mode age, start, and big chance created are below the value of significance. The rest of the selected independent variables are above the significance value. The F-test for the backward is 8.405 on 20 and 109degree of freedom. On the step forward the R-squared is 0.6338 which means 63.38 percent of the variability in the transfer values could be explained by the model and all the independent variables are below the value of significance which is below the p-value of 0.05. The F-test for forward is 19.6 on 12 and 117.

Based on these two pairwise selection, the model for predicting the transfer value for the defense position is better when constructed with the forward model. Starting with an empty variable and adding the once that are more contributing variables produce a better model.

# models

```
> model.defense4 <-lm(defense$Values ~ defense$Age + defense$Starts + defense
$Subs +  defense$Appearances + defense$Wins + defense$Losses + defense$Tackle
s + defense$Tackle_success + defense$Blocked_shots + defense$Interceptions +
defense$Headed_Clearance  + defense$Duels_lost  + defense$Aerial_battles_lost
+ defense$Aerial_battles_won + defense$Assists + defense$Passes + defense$Pas
ses_per_match   + defense$Big_chances_created + defense$Errors_leading_to_goa
l)
> summary(model.defense4)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Subs + defense$Appearances + defense$Wins + defense$Losses +
    defense$Tackles + defense$Tackle_success + defense$Blocked_shots +
    defense$Interceptions + defense$Headed_Clearance + defense$Duels_lost +
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Assists + defense$Passes + defense$Passes_per_match +
    defense$Big_chances_created + defense$Errors_leading_to_goal)

Residuals:
      Min        1Q    Median        3Q       Max
-22014739  -8192323  -2306459   6489280  49198618

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   58636165   10959004   5.351 4.83e-07 ***
defense$Age                   -2342628     407581  -5.748 8.20e-08 ***
defense$Starts                  580858     112143   5.180 1.01e-06 ***
defense$Subs                   -246564     533597  -0.462 0.644937
defense$Appearances             -29127     249422  -0.117 0.907249
defense$Wins                     75462     306687   0.246 0.806098
defense$Losses                  -32188     344917  -0.093 0.925818
defense$Tackles                 -34435      31112  -1.107 0.270784
defense$Tackle_success         -147900      65707  -2.251 0.026379 *
defense$Blocked_shots           260831     144762   1.802 0.074317 .
defense$Interceptions           -11580      40008  -0.289 0.772779
defense$Headed_Clearance        -53067      29604  -1.793 0.075791 .
defense$Duels_lost              -16447      27661  -0.595 0.553330
defense$Aerial_battles_lost     -53280      61778  -0.862 0.390313
defense$Aerial_battles_won      100222      35816   2.798 0.006067 **
defense$Assists               -1230995     598860  -2.056 0.042192 *
defense$Passes                    3145       2318   1.357 0.177513
defense$Passes_per_match        456170     121887   3.743 0.000291 ***
defense$Big_chances_created    1520508     505886   3.006 0.003283 **
defense$Errors_leading_to_goal  -460725    1078497  -0.427 0.670075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12820000 on 110 degrees of freedom
Multiple R-squared:  0.6734,   Adjusted R-squared:  0.617
F-statistic: 11.94 on 19 and 110 DF,  p-value: < 2.2e-16
```

```
> model.defense5 <-lm(defense$Values ~ defense$Age + defense$Starts +  defens
e$Tackles + defense$Tackle_success + defense$Blocked_shots + defense$Headed_C
learance  +  defense$Aerial_battles_lost + defense$Aerial_battles_won + defen
se$Assists + defense$Passes + defense$Passes_per_match   + defense$Big_chance
s_created + defense$Errors_leading_to_goal)
> summary(model.defense5)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Tackles + defense$Tackle_success + defense$Blocked_shots +
    defense$Headed_Clearance + defense$Aerial_battles_lost +
    defense$Aerial_battles_won + defense$Assists + defense$Passes +
    defense$Passes_per_match + defense$Big_chances_created +
    defense$Errors_leading_to_goal)

Residuals:
      Min        1Q    Median        3Q       Max
-23183925  -8151665  -1552214   6162818  49259995

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    58382813    9750073   5.988 2.44e-08 ***
defense$Age                    -2374683     379317  -6.260 6.69e-09 ***
defense$Starts                   576354     102597   5.618 1.35e-07 ***
defense$Tackles                  -55589      20136  -2.761  0.00671 **
defense$Tackle_success          -160636      60676  -2.647  0.00924 **
defense$Blocked_shots            244323     121451   2.012  0.04657 *
defense$Headed_Clearance         -52682      24556  -2.145  0.03401 *
defense$Aerial_battles_lost      -94613      38921  -2.431  0.01659 *
defense$Aerial_battles_won       107090      34075   3.143  0.00212 **
defense$Assists                -1335727     556085  -2.402  0.01789 *
defense$Passes                    3466       1458   2.377  0.01908 *
defense$Passes_per_match          486955     108165   4.502 1.61e-05 ***
defense$Big_chances_created      1503963     475964   3.160  0.00201 **
defense$Errors_leading_to_goal   -631028     951698  -0.663  0.50861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12570000 on 116 degrees of freedom
Multiple R-squared:  0.6691,   Adjusted R-squared:  0.632
F-statistic: 18.04 on 13 and 116 DF,  p-value: < 2.2e-16
```

```
> model.defense6 <-lm(defense$Values ~ defense$Age + defense$Starts +  defens
e$Tackles + defense$Tackle_success + defense$Blocked_shots + defense$Headed_C
learance  +  defense$Aerial_battles_lost + defense$Aerial_battles_won + defen
se$Assists + defense$Passes + defense$Passes_per_match   + defense$Big_chance
s_created)
> summary(model.defense6)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Tackles + defense$Tackle_success + defense$Blocked_shots +
    defense$Headed_Clearance + defense$Aerial_battles_lost +
    defense$Aerial_battles_won + defense$Assists + defense$Passes +
    defense$Passes_per_match + defense$Big_chances_created)

Residuals:
      Min        1Q    Median        3Q       Max
-23293262  -8036377  -1504692   6400668  49360599

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                58437531    9726348   6.008 2.18e-08 ***
defense$Age                -2385075     378084  -6.308 5.21e-09 ***
defense$Starts               585548     101412   5.774 6.49e-08 ***
defense$Tackles              -52183      19423  -2.687  0.00827 **
defense$Tackle_success      -163597      60367  -2.710  0.00774 **
defense$Blocked_shots        250889     120756   2.078  0.03993 *
defense$Headed_Clearance     -54270      24380  -2.226  0.02793 *
defense$Aerial_battles_lost  -93397      38784  -2.408  0.01760 *
defense$Aerial_battles_won   105960      33950   3.121  0.00227 **
defense$Assists            -1343456     554630  -2.422  0.01696 *
defense$Passes                3126       1362   2.296  0.02347 *
defense$Passes_per_match      491021     107732   4.558 1.28e-05 ***
defense$Big_chances_created  1521956     474051   3.211  0.00171 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12540000 on 117 degrees of freedom
Multiple R-squared:  0.6678,   Adjusted R-squared:  0.6338
F-statistic:  19.6 on 12 and 117 DF,  p-value: < 2.2e-16
```

Based on the step forward summary and step backward summary outputs, I created several models by removing the least contributing variables and leaving the variables that are more contributing in the model. Doing so, my final model produced 0.6338 R-square with the F-test 19.6 on 12 and 117 degree of freedom. Hence, 63.38 percent of the variability in the transfer values could be explained by this model.  All the independent variables are below 0.05.

In conclusion, the transfer value for defense position can be predicted by the independent variables of age, starts, aerial_battles_won, aerial_battles_lost, tackle_success, aerial_battles_lost, blocked_shots, assists, passes, and big_chances_created. Most of these variables are critical and must have skills for a defense position player. Age, start, assists, big chance created, and pass are the only variables that are general variables. This means that these variables are applicable to any player could be attack position or a midfielder position. Good defenders are aggressive. They aren’t afraid to make strong tackles and use their body. They attack balls with their heads and get in front of shots. Strong defenders push offensive players off the ball, win headers, and shield the ball well. Defenders need good heading ability to defend

against aerial passes, crosses, and set pieces. Defenders should head the ball far and wide to remove the ball from the defensive half. Defenders without the ability to pass effectively reduce a team's possession percentage. Clearing the ball has its place. But defenders should look up and pass the ball when they have time and space. Good defenders can pass to feet and make effective long passes through the air. All the factors of good defender are present on the final model as independent variables.

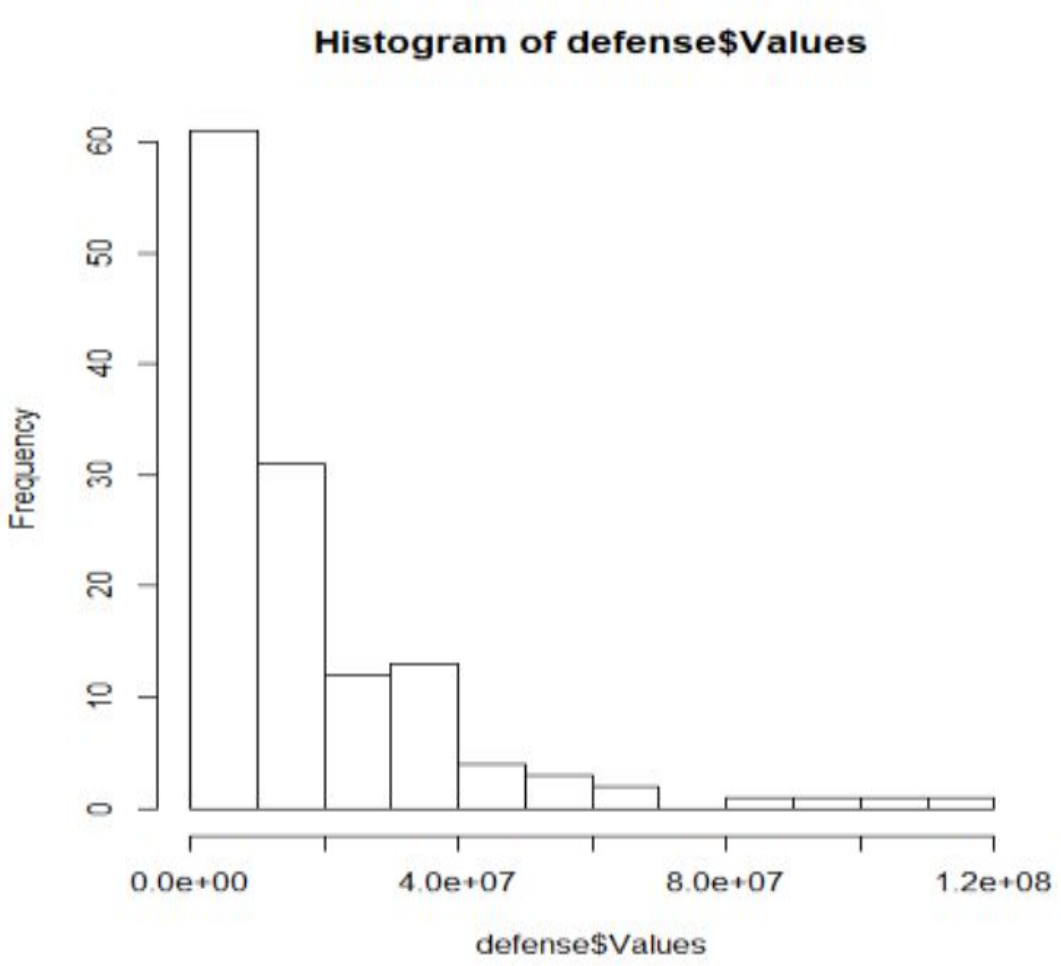
In the next milestone I will enhance my model by adding variables that are not in this model. The variables that are removed from this model because of NA values will be added. Some of these variables are not included in this model are important for the defense position. Replacing the NA values of the variables with Zero might work if the variables NA is not significant. I will also use log transformation for the transfer value.

On the first milestone the final model gave us 0.6338 R-square with independent variables that are mostly defense position variables. On that final model transformation wasn't performed on any variable nor interaction or second order term was used too. Here is the model.

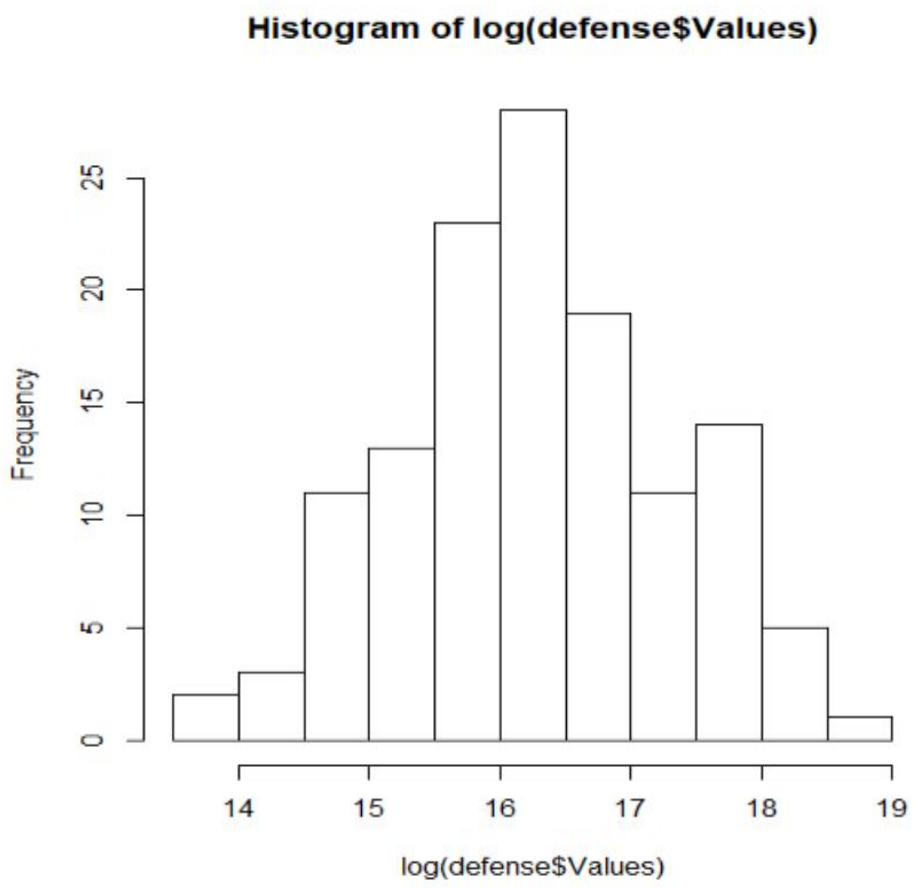
```
> model.defense6 <-lm(defense$Values ~ defense$Age + defense$Starts +  defense$Tack
les + defense$Tackle_success + defense$Blocked_shots + defense$Headed_Clearance  +
defense$Aerial_battles_lost + defense$Aerial_battles_won + defense$Assists + defens
e$Passes + defense$Passes_per_match   + defense$Big_chances_created)
> summary(model.defense6)

Call:
lm(formula = defense$Values ~ defense$Age + defense$Starts +
    defense$Tackles + defense$Tackle_success + defense$Blocked_shots +
    defense$Headed_Clearance + defense$Aerial_battles_lost +
    defense$Aerial_battles_won + defense$Assists + defense$Passes +
    defense$Passes_per_match + defense$Big_chances_created)

Residuals:
      Min        1Q    Median        3Q       Max
-23293262  -8036377  -1504692   6400668  49360599

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                58437531    9726348   6.008 2.18e-08 ***
defense$Age                -2385075     378084  -6.308 5.21e-09 ***
defense$Starts               585548     101412   5.774 6.49e-08 ***
defense$Tackles              -52183      19423  -2.687  0.00827 **
defense$Tackle_success      -163597      60367  -2.710  0.00774 **
defense$Blocked_shots        250889     120756   2.078  0.03993 *
defense$Headed_Clearance     -54270      24380  -2.226  0.02793 *
defense$Aerial_battles_lost  -93397      38784  -2.408  0.01760 *
defense$Aerial_battles_won   105960      33950   3.121  0.00227 **
defense$Assists            -1343456     554630  -2.422  0.01696 *
defense$Passes                3126       1362   2.296  0.02347 *
defense$Passes_per_match    491021     107732   4.558 1.28e-05 ***
defense$Big_chances_created 1521956     474051   3.211  0.00171 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12540000 on 117 degrees of freedom
Multiple R-squared:  0.6678,   Adjusted R-squared:  0.6338
F-statistic:  19.6 on 12 and 117 DF,  p-value: < 2.2e-16
```

Residual standard error: 12540000 on 117 degrees of freedom Multiple R-squared: 0.6678, Adjusted R-squared: 0.6338 F-statistic: 19.6 on 12 and 117 DF, p-value: < 2.2e-16

The defense position transfer value histogram without transformation of transfer value. This histogram is skewed to the right.




The defense position transfer value histogram after the transfer value variable transformed. The histogram indicates a symmetric, moderate tailed distribution.




Again, this model is the final model in our milestone one, but the transfer value of the defense players is transformed. In this model the adjusted R-square increased from 0.6338 to 0.7009 with 5 variables above the p-value of significance. This means we need to check the other variables if we can transform transformation even removal them depending on the variables effect on the model.

```
> model.defense6 <-lm(log(defense$values) ~ defense$Age + defense$Starts +  defense
$Tackles + defense$Tackle_success + defense$Blocked_shots + defense$Headed_Clearanc
e  +  defense$Aerial_battles_lost + defense$Aerial_battles_won + defense$Assists +
defense$Passes + defense$Passes_per_match   + defense$Big_chances_created)
> summary(model.defense6)

Call:
lm(formula = log(defense$values) ~ defense$Age + defense$Starts +
    defense$Tackles + defense$Tackle_success + defense$Blocked_shots +
    defense$Headed_Clearance + defense$Aerial_battles_lost +
    defense$Aerial_battles_won + defense$Assists + defense$Passes +
    defense$Passes_per_match + defense$Big_chances_created)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42904 -0.36566 -0.01584  0.35857  1.42722

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.970e+01  4.436e-01  44.412  < 2e-16 ***
defense$Age                -1.797e-01  1.724e-02 -10.422  < 2e-16 ***
defense$Starts              3.391e-02  4.625e-03   7.332 3.20e-11 ***
defense$Tackles            -1.906e-03  8.858e-04  -2.152  0.03346 *
defense$Tackle_success     -1.205e-02  2.753e-03  -4.378 2.62e-05 ***
defense$Blocked_shots       3.430e-03  5.507e-03   0.623  0.53459
defense$Headed_Clearance   -5.157e-04  1.112e-03  -0.464  0.64362
defense$Aerial_battles_lost -1.625e-03  1.769e-03  -0.918  0.36027
defense$Aerial_battles_won  2.136e-03  1.548e-03   1.380  0.17035
defense$Assists            -4.743e-02  2.529e-02  -1.875  0.06328 .
defense$Passes              1.753e-04  6.209e-05   2.824  0.00558 **
defense$Passes_per_match    2.711e-02  4.913e-03   5.518 2.09e-07 ***
defense$Big_chances_created 4.891e-02  2.162e-02   2.262  0.02553 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5718 on 117 degrees of freedom
Multiple R-squared:  0.7287,   Adjusted R-squared:  0.7009
F-statistic: 26.19 on 12 and 117 DF,  p-value: < 2.2e-16
```

Even though, the Independent variable “passes” is mainly attacker’s skill, a defense play should also have a good skill on passing the ball. The pass variable is useful second order-term in this model. When the variable was removed from the model the R-square decreased.  In the process of building and transforming it’s not just the “pass” variable who exhibited multicollinearity is included in this model; the other variables who exhibited multicollinearity are also included. Below, we can see the regression outcome when the “pass” variable removed.

```
Call:
lm(formula = log(defense$Values) ~ defense$Age + defense$Tackles +
    defense$Tackle_success + defense$Blocked_shots + defense$Headed_Clearance +
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Passes_per_match + defense$Starts * defense$Age)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56009 -0.36999 -0.01675  0.38473  1.38533

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.862e+01  8.032e-01  23.179  < 2e-16 ***
defense$Age                -1.399e-01  3.123e-02  -4.480 1.73e-05 ***
defense$Tackles            -4.174e-05  6.779e-04  -0.062 0.951003
defense$Tackle_success     -1.301e-02  2.778e-03  -4.683 7.56e-06 ***
defense$Blocked_shots       1.297e-02  3.670e-03   3.533 0.000586 ***
defense$Headed_Clearance    1.218e-03  1.089e-03   1.118 0.265906
defense$Aerial_battles_lost -1.286e-03  1.807e-03  -0.712 0.477996
defense$Aerial_battles_won   8.430e-04  1.570e-03   0.537 0.592229
defense$Passes_per_match     3.293e-02  4.558e-03   7.223 5.23e-11 ***
defense$Starts               9.378e-02  3.618e-02   2.592 0.010724 *
defense$Age:defense$Starts  -2.297e-03  1.367e-03  -1.680 0.095575 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5949 on 119 degrees of freedom
Multiple R-squared:  0.7013,   Adjusted R-squared:  0.6762
F-statistic: 27.94 on 10 and 119 DF,  p-value: < 2.2e-16
```

When the second order-term pass variable included in the model we can see that the R-square increased from 0.676.2 to 0.7238. therefore, the variable will be used in the model. So far, we transformed transfer value, add interaction and second order term and removed one variable. The outcome of this model shows that most of the independent variables that are included in this model are statistically significant only three variables are above the p-value of significance.

```
Call:
lm(formula = log(defense$values) ~ defense$Age + defense$Tackles +
    defense$Tackle_success + defense$Blocked_shots + defense$Headed_clearance +
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Passes + defense$Passes_per_match + defense$Starts *
    defense$Age + I(defense$Passes^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.39799 -0.29433 -0.01807  0.31136  1.42361

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.840e+01  7.433e-01  24.760  < 2e-16 ***
defense$Age                -1.285e-01  2.895e-02  -4.441 2.04e-05 ***
defense$Tackles            -1.661e-03  8.624e-04  -1.927 0.056461 .
defense$Tackle_success     -1.164e-02  2.666e-03  -4.365 2.76e-05 ***
defense$Blocked_shots       5.752e-03  4.017e-03   1.432 0.154847
defense$Headed_clearance   -3.608e-04  1.063e-03  -0.340 0.734836
defense$Aerial_battles_lost -4.050e-03  1.799e-03  -2.252 0.026183 *
defense$Aerial_battles_won  2.056e-03  1.483e-03   1.386 0.168310
defense$Passes              3.912e-04  8.246e-05   4.744 5.98e-06 ***
defense$Passes_per_match    2.196e-02  4.862e-03   4.517 1.51e-05 ***
defense$Starts              1.144e-01  3.369e-02   3.396 0.000935 ***
I(defense$Passes^2)        -1.485e-08  4.518e-09  -3.286 0.001341 **
defense$Age:defense$Starts -3.107e-03  1.274e-03  -2.439 0.016240 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5494 on 117 degrees of freedom
Multiple R-squared:  0.7495,   Adjusted R-squared:  0.7238
F-statistic: 29.17 on 12 and 117 DF,  p-value: < 2.2e-16
```

Looking at the VIF results we can see that age, starts, passes per match, tackles, tackle success, and block shots have a VIF below 10, which shows that multicollinearity does not affect them and we can trust this coefficient and p-value with no further action. However, the coefficients and p-values for the other terms are such as aerial battles lost, aerial battles won, assist, head clearance, passes, and big chance created have a VIF above 10 and these independent variables are suspect!

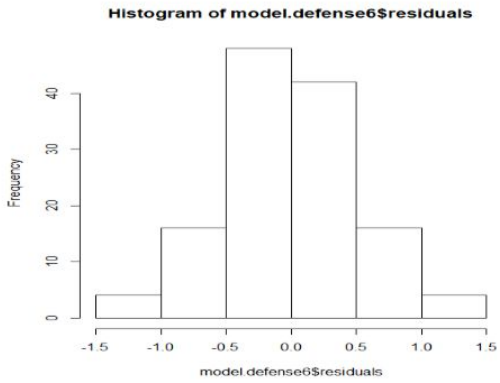
Some of multicollinearity in this model is the structural type. Both the interaction term of Stars*age and the second order-term variable “passes” are included. At this stage the independent variable “big chance created” is now removed. The reason why the variable is removed is that first the variable is general variable second it is above the p-value of significance when the transfer value transformed, and it did not affect the R-square of the model when removed.

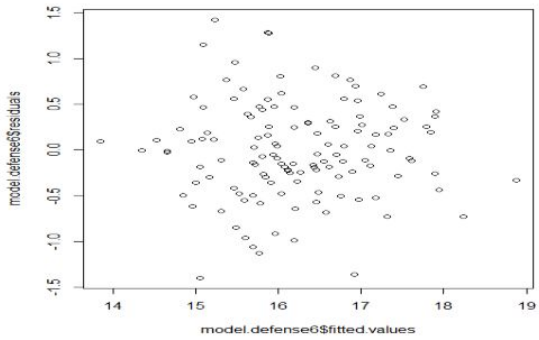
```
vif(model.defense6)
                 defense$Age                defense$Starts                defense$Tackles
                    1.635685                     1.080998                     8.491493
     defense$Tackle_success         defense$Blocked_shots     defense$Headed_Clearance
                    2.018167                     5.127296                    16.448513
defense$Aerial_battles_lost   defense$Aerial_battles_won               defense$Assists
                   13.399494                    25.887517                    17.282304
              defense$Passes     defense$Passes_per_match defense$Big_chances_created
                   16.410438                     2.456478                    16.927830
```

To see how predictable the model would be without multicollinearity, a model is built without all the variable with VIF greater than 10. In this model we can see that the R-square decreased to 0.7169 from the model that has multicollinearity and R-square of 0.7238. generally, a defense play should be good in blocks, interceptions, tackles, last Man tackles, clearances, headed clearances, aerial battles won, and good on making less errors such as own goals, errors leading to goal, and fouls. However, this model did not use all these independent variables (factors) but most of them are included in the final model with many of them below the p-value of significance and R-square of 0.7238 which is higher from the final model of milestone one. Since, these independent variables are the main characteristics that explains about defense players it is important to keep them in the final model.

```
Call:
lm(formula = log(defense$Values) ~ defense$Age + defense$Tackles +
    defense$Tackle_success + defense$Passes + defense$Passes_per_match +
    defense$Starts * defense$Age + I(defense$Passes^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.40397 -0.31848 -0.04316  0.33441  1.52757

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.841e+01  7.304e-01  25.202  < 2e-16 ***
defense$Age               -1.298e-01  2.831e-02  -4.586 1.11e-05 ***
defense$Tackles           -2.412e-03  7.677e-04  -3.142 0.002108 **
defense$Tackle_success    -1.178e-02  2.666e-03  -4.417 2.20e-05 ***
defense$Passes             3.637e-04  6.015e-05   6.047 1.69e-08 ***
defense$Passes_per_match   2.259e-02  4.729e-03   4.777 5.04e-06 ***
defense$Starts             1.155e-01  3.323e-02   3.476 0.000707 ***
I(defense$Passes^2)       -1.093e-08  4.134e-09  -2.645 0.009244 **
defense$Age:defense$Starts -3.103e-03  1.258e-03  -2.466 0.015068 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5562 on 121 degrees of freedom
Multiple R-squared:  0.7345,   Adjusted R-squared:  0.7169
F-statistic: 41.84 on 8 and 121 DF,  p-value: < 2.2e-16
```

**Histogram of the residuals** **Residuals against the predicted values**




We could see that it is a little bit skewed to the right, but the model looks normal (enough).The residual plot shows a fairly random pattern around zero Almost half of the residuals are positive, and the other half are negative. This random pattern indicates that a linear model provides a decent fit to the data. Therefore, the data points exhibit homoscedasticity.

## Final model

```
Call:
lm(formula = log(defense$values) ~ defense$Age + defense$Tackles +
    defense$Tackle_success + defense$Blocked_shots + defense$Headed_clearance +
    defense$Aerial_battles_lost + defense$Aerial_battles_won +
    defense$Passes + defense$Passes_per_match + defense$Starts *
    defense$Age + I(defense$Passes^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.39799 -0.29433 -0.01807  0.31136  1.42361

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.840e+01  7.433e-01  24.760  < 2e-16 ***
defense$Age                -1.285e-01  2.895e-02  -4.441 2.04e-05 ***
defense$Tackles            -1.661e-03  8.624e-04  -1.927 0.056461 .
defense$Tackle_success     -1.164e-02  2.666e-03  -4.365 2.76e-05 ***
defense$Blocked_shots       5.752e-03  4.017e-03   1.432 0.154847
defense$Headed_clearance   -3.608e-04  1.063e-03  -0.340 0.734836
defense$Aerial_battles_lost -4.050e-03  1.799e-03  -2.252 0.026183 *
defense$Aerial_battles_won  2.056e-03  1.483e-03   1.386 0.168310
defense$Passes              3.912e-04  8.246e-05   4.744 5.98e-06 ***
defense$Passes_per_match    2.196e-02  4.862e-03   4.517 1.51e-05 ***
defense$Starts              1.144e-01  3.369e-02   3.396 0.000935 ***
I(defense$Passes^2)        -1.485e-08  4.518e-09  -3.286 0.001341 **
defense$Age:defense$Starts -3.107e-03  1.274e-03  -2.439 0.016240 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5494 on 117 degrees of freedom
Multiple R-squared:  0.7495,   Adjusted R-squared:  0.7238
F-statistic: 29.17 on 12 and 117 DF,  p-value: < 2.2e-16
```

Overall, the model did increase the R-square from what it was by nine percent. After transforming transfer value, adding both the interaction and second order term and removing one

variable the model produced some variables with high p-value of significance but better R-square. The F-test for this model is 29.17 on 12 and 117 degree of freedom. Transformation on the rest of the variables didn't work well. However, we kept the variables in the final model. These variables with higher p-value significance did show multicollinearity. As stated above, the independent variables in this model are important to remove them from the model. Except for variable "passes" all variables in this model are defenders skill set quality measures.

# 7. Jilei Hao: Midfielder Model

## Background

Midfielders are players playing in the middle of the field, between forwards and defenders. Their job covers almost all aspects of a soccer game: defense, transition, and attack. During the attack, midfielders play as the mastermind. Their jobs include controlling the ball in the middle of the park, finding space behind the opponent's defense, passing balls to forwards to create chances, making runs to drag the opponent's defensive players and sometimes they will score by themselves if chances are present. During the transition from defense to attack, midfielders are the receivers of the ball from the back. They need to keep possession of the ball and start pushing forwards to plan attacks. During the defense, midfielders are the ones to sabotage opponent's attack from the planning phase, by pressing, intercepting and recovering balls directly from opposition midfielders or defenders.

Based their versatile roles in the game, the features to predict midfielder values are expected to cover a wide range of variables, including both attacking, teamplaying and defensive ones. Therefore, the total number of features to start the model is large and some trimming is needed to keep the modeling work manageable. Further partition of the category is possible, by distinguishing defensive midfielders, central midfielders and attacking midfielders (playmakers), but since our data doesn't have the feature, this report will treat midfielders as a general entity.

## Data Preparation

I started the modeling by handpicking the relevant variables using domain knowledge. Here are the principles I followed during the process:

- Not choosing duplicate variables. There are variables describing similar information that can be easily identified, such as age and born year
- Not choosing apparently unimportant information, such as names and row number
- Not choosing club and country, as this report is focusing on performances
- Not choosing goalkeeper stats. They are only for goalkeepers and NAs are populated for all midfielders
- Keeping all Team-Play stats. It is the most important stats for midfielders since their primary responsibility is to connect the play from back to forth, and Team-Play stats measure just that, such as Passes, Assists, Through Balls etc.

- Keeping all Attacking and Defensive stats except those are populated all NAs for midfielders. Midfielders are participants of both defense and attack, so those traits are all important. Some of the variables are populated all NAs for midfielder are either too "attacking" or "defensive" for midfielders, such as own goals and clearance off-line, and these variables are excluded

The handpick data has 56 explanatory variables and 1 response variable. All the 56 explanatory variables are numerical. Since the distribution of the transfer-value is extremely skewed, it is transformed using the log function.

## Feature Selection

To simplify the model, stepwise feature selections were performed to find the variables that give the most information.

Result of forward selection:

```
AIC = -369.63

log(Transfer.Values) ~ SoT + Squad + Mn.Ap + Age + Passes.per.match +
  Country + PKatt + Ast + Goals.with.right.foot + Crosses +
  Goals + Offsides + Gls + CrdR + SoT90 + Red.cards + Recoveries +
  Losses + Yellow.cards + Errors.leading.to.goal + Duels.lost +
  Tackles + PK + Tackle.success.. + Assists + Appearances +
  Wins + Headed.Clearance + Aerial.battles.lost + Hit.woodwork +
  Big.chances.missed + G.A + Big.chances.created + Through.balls +
  Starts + Apps + Crd + CrdY + Fouls + Fls + Passes + G.A.PK +
  Blocked.shots + Clearances + Fls90
```

Result of backward selection:

```
AIC = -569.71

log(Transfer.Values) ~ Country + Squad + Apps + Starts + Min +
  Mn.Ap + PK + PKatt + Fls + CrdY + CrdR + SoT + Gls90 + G.A +
  G.PK + G.A.PK + Fls90 + Crd + Appearances + Wins + Losses +
  Tackles + Tackle.success.. + Blocked.shots + Interceptions +
  Clearances + Headed.Clearance + Recoveries + Duels.won +
  Duels.lost + Successful.50.50s + Aerial.battles.won +
Aerial.battles.lost +
  Errors.leading.to.goal + Assists + Passes + Passes.per.match +
  Big.chances.created + Crosses + Cross.accuracy.. + Through.balls +
  Accurate.long.balls + Yellow.cards + Red.cards + Fouls +
  Offsides + Goals + Goals.per.match + Headed.goals +
Goals.with.right.foot +
  Goals.with.left.foot + Hit.woodwork + Big.chances.missed
```

I chose forward selection result as the base of my model since it has smaller AIC and a much simpler model, which makes interpretation of the model easier.

## First Model

Following is the first model built using the forward selected variables.

```
lm(formula = log(Transfer.Values) ~ Mn.Ap + Age + Gls + Losses +
    Wins + +Starts + Apps + Appearances + SoT + PKatt + Goals +
    Goals.with.right.foot + Hit.woodwork + Big.chances.missed +
    G.A + G.A.PK + Offsides + SoT90 + PK + Assists + Ast +
Passes.per.match +
    Crosses + Passes + Big.chances.created + Through.balls +
    CrdR + Red.cards + Yellow.cards + Crd + CrdY + Fouls + Fls +
    Fls90 + Recoveries + Errors.leading.to.goal + Duels.lost +
    Tackles + Tackle.success.. + Headed.Clearance +
Aerial.battles.lost +
    Blocked.shots + Clearances, data = cmf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89197 -0.23433 -0.01379  0.21802  1.34000

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.668e+01  5.973e-01  27.926  < 2e-16 ***
Mn.Ap                   8.816e-03  6.362e-03   1.386 0.170801
Age                    -8.551e-02  2.091e-02  -4.090 0.000127 ***
Gls                     1.119e-01  5.287e-02   2.116 0.038354 *
Losses                 -1.486e-02  1.756e-02  -0.846 0.400703
Wins                    1.170e-02  1.643e-02   0.713 0.478779
Starts                 -2.684e-02  2.423e-02  -1.107 0.272417
Apps                    4.575e-02  1.709e-02   2.678 0.009476 **
Appearances            -8.056e-03  1.230e-02  -0.655 0.514882
SoT                    -2.478e-03  1.959e-02  -0.126 0.899750
PKatt                   1.023e-02  2.339e-01   0.044 0.965238
Goals                  -1.593e-02  1.708e-02  -0.933 0.354499
Goals.with.right.foot   3.134e-02  1.560e-02   2.009 0.048898 *
Hit.woodwork            6.042e-02  3.407e-02   1.773 0.081083 .
Big.chances.missed     -9.560e-03  1.908e-02  -0.501 0.618086
G.A                    -3.821e+00  7.670e+00  -0.498 0.620165
G.A.PK                  3.785e+00  7.793e+00   0.486 0.628940
Offsides               -1.230e-03  7.717e-03  -0.159 0.873867
SoT90                   4.134e-01  1.345e-01   3.074 0.003136 **
PK                     -3.086e-02  3.279e-01  -0.094 0.925334
```

```
Assists                  7.963e-03  1.859e-02   0.428 0.669823
Ast                     -1.094e-02  3.760e-02  -0.291 0.772013
Passes.per.match         2.060e-02  4.638e-03   4.441 3.76e-05 ***
Crosses                  2.983e-04  5.871e-04   0.508 0.613270
Passes                   6.025e-05  1.060e-04   0.568 0.571984
Big.chances.created     -7.059e-03  1.361e-02  -0.519 0.605950
Through.balls           -7.298e-03  4.248e-03  -1.718 0.090827 .
CrdR                    -3.802e-01  2.252e-01  -1.688 0.096380 .
Red.cards               -6.175e-02  9.924e-02  -0.622 0.536116
Yellow.cards            -1.255e-02  1.531e-02  -0.820 0.415605
Crd                      6.933e-01  6.200e-01   1.118 0.267786
CrdY                    -8.871e-02  5.071e-02  -1.749 0.085154 .
Fouls                    3.822e-04  3.391e-03   0.113 0.910639
Fls                      4.269e-03  9.176e-03   0.465 0.643396
Fls90                    3.556e-03  2.022e-02   0.176 0.860944
Recoveries               1.177e-03  7.731e-04   1.522 0.133004
Errors.leading.to.goal   8.059e-02  6.168e-02   1.307 0.196165
Duels.lost               2.601e-03  1.311e-03   1.985 0.051599 .
Tackles                 -3.169e-03  1.913e-03  -1.657 0.102632
Tackle.success..        -3.185e-01  3.097e-01  -1.028 0.307733
Headed.Clearance        -8.568e-03  5.281e-03  -1.622 0.109812
Aerial.battles.lost     -1.731e-03  2.491e-03  -0.695 0.489811
Blocked.shots           -1.033e-02  5.541e-03  -1.865 0.066981 .
Clearances               5.141e-03  4.060e-03   1.266 0.210198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4537 on 62 degrees of freedom
Multiple R-squared:  0.8657,    Adjusted R-squared:  0.7725
F-statistic: 9.293 on 43 and 62 DF,  p-value: 6.611e-15
```

The F Test is good (6.611e-15 < 0.05) indicating at least one beta is not zero. The adj-R2 is 0.7725 meaning 77.25% of the variability in transfer-value is explained by the model. Only 5 variables in this initial model are significant indicating a need for simplification and multicollinearity check.

## Multicollinearity Analysis

Based on the first model, I calculated the VIF for each variable and get following result:

| | | | | |
|---|---|---|---|---|
| Mn.Ap | Age | Gls | Losses | Wins |
| 7.372538 | 2.282838 | 12.009271 | 145.803070 | 234.036967 |
| Starts | Apps | Appearances | SoT | PKatt |
| 36.895466 | 17.370266 | 610.319715 | 15.651826 | 66.126996 |
| Goals | Goals.with.right.foot | Hit.woodwork | Big.chances.missed | G.A |
| 21.384671 | 6.476107 | 7.415175 | 12.039146 | 1278.510787 |
| G.A.PK | Offsides | SoT90 | PK | Assists |
| 1241.174324 | 6.467557 | 2.756421 | 91.061021 | 41.316050 |
| Ast | Passes.per.match | Crosses | Passes | Big.chances.created |
| 5.286388 | 2.808556 | 21.259325 | 78.463830 | 28.350238 |
| Through.balls | CrdR | Red.cards | Yellow.cards | Crd |
| 15.578546 | 1.822484 | 3.366041 | 19.095452 | 4.338973 |
| CrdY | Fouls | Fls | Fls90 | Recoveries |
| 10.886791 | 43.225358 | 9.122763 | 2.151550 | 54.934592 |
| Errors.leading.to.goal | Duels.lost | Tackles | Tackle.success.. | Headed.Clearance |
| 3.226110 | 148.627434 | 55.055664 | 1.747003 | 30.466471 |
| Aerial.battles.lost | Blocked.shots | Clearances | | |
| 24.603186 | 21.384216 | 81.091513 | | |

There are several groups of variables that produced high VIF scores. I analyzed each group and did following adjustments. The general principle for picking some variables over another is the picked variable(s) reduces more VIF with less adj-R2 loss.

- Removed G.A.PK, G.A, G.A.PK, and leaves Gls, PK, since they have big overlap in information contained
- Removed Appearances overlapping with Mn.Ap, Apps
- Removed All defensive variables except Recoveries
- Removed All Team Play variables except Passes.Per.Match
- Removed Yellow Cards, Red Cards overlapping with CrdY, CrdR etc. except CrdR
- Removed big.chances.missed overlapping with Hit.woodwork

Following is the adjusted model:

```
lm(formula = log(Transfer.Values) ~ Mn.Ap + Age + Gls + Losses +
    Apps + Goals.with.right.foot + Hit.woodwork + Offsides +
    SoT90 + PK + Passes.per.match + CrdR, data = cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.087201 -0.018282 -0.000791  0.015987  0.068926

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.8242618  0.0327445  86.251  < 2e-16 ***
Mn.Ap                    0.0002427  0.0002572   0.943 0.347951
Age                     -0.0050645  0.0013212  -3.833 0.000230 ***
Gls                      0.0054945  0.0017216   3.191 0.001931 **
Losses                  -0.0004096  0.0001682  -2.436 0.016772 *
Apps                     0.0010618  0.0004638   2.289 0.024324 *
Goals.with.right.foot    0.0018180  0.0006440   2.823 0.005822 **
```

```
Hit.woodwork           0.0013610  0.0013177   1.033 0.304343
Offsides               0.0003381  0.0002745   1.232 0.221128
SoT90                  0.0253477  0.0069756   3.634 0.000457 ***
PK                    -0.0087251  0.0029488  -2.959 0.003916 **
Passes.per.match       0.0013329  0.0002175   6.130 2.11e-08 ***
CrdR                  -0.0212178  0.0119754  -1.772 0.079705 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03116 on 93 degrees of freedom
Multiple R-squared:  0.741,     Adjusted R-squared:  0.7076
F-statistic: 22.17 on 12 and 93 DF,  p-value: < 2.2e-16
```

The adjusted model havs Adj-R2 reduced to 0.7076. But the VIF matrix looks much cleaner, with biggest VIF reduced from 1278 to 2.93, which helped interpreting the T-Tests and Coefficients of the variables. The number of significant variables increased from 5 to 8.

| Mn.Ap | Age | Gls | Losses | Apps |
|---|---|---|---|---|
| 2.555982 | 1.932913 | 2.700769 | 2.836866 | 2.714582 |
| Goals.with.right.foot | Hit.woodwork | Offsides | SoT90 | PK |
| 2.340877 | 2.351770 | 1.735217 | 1.572750 | 1.561658 |
| Passes.per.match | CrdR | | | |
| 1.309064 | 1.092685 | | | |

I then removed the variables that didn’t pass the t-tests: Hit.woodwork, Offsides, CrdR and got the following model:

```
lm(formula = log(Transfer.Values) ~ Age + Gls + Losses + Apps +
    Goals.with.right.foot + SoT90 + PK + Passes.per.match, data =
cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.088689 -0.018353  0.000547  0.016346  0.068003

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.8162697  0.0317654  88.658  < 2e-16 ***
Age                   -0.0044387  0.0012924  -3.435 0.000875 ***
Gls                    0.0067995  0.0016879   4.028 0.000112 ***
Losses                -0.0003635  0.0001684  -2.159 0.033346 *
Apps                   0.0011310  0.0003684   3.070 0.002780 **
Goals.with.right.foot  0.0022065  0.0006195   3.562 0.000573 ***
SoT90                  0.0267871  0.0065364   4.098 8.65e-05 ***
PK                    -0.0105470  0.0029186  -3.614 0.000481 ***
Passes.per.match       0.0014464  0.0002047   7.066 2.45e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.03177 on 97 degrees of freedom
Multiple R-squared:  0.7191,    Adjusted R-squared:  0.6959
F-statistic: 31.04 on 8 and 97 DF,  p-value: < 2.2e-16
```

The F-Test looks good with p-value = 2.2e-16 < 0.05, meaning we can reject the null hypothesis that all betas are zeros and accept the alternative that at least one beta is not zero. The adj-R2 is 0.6956, meaning 69.59% of the variability in transfer-value can be explained by the model. All the remaining 8 variables are significant. The Residual Standard Error is significantly reduced from 0.4537 to 0.03177.

It shows the multicollinearity analysis has made significant improvement to the model with reduced errors and increased significance and variable clarity. The only sacrifice is a small decrease in adj-R2, which makes sense since so far I only removed variables from the model.

## Second Order Terms

By analyzing the boxplot between log(Transfer.Values) and the remaining explanatory variables. I added the second order terms of SoT90, Age, Losses, Pass.Per.Match, Ast to the model because they showed a slight curved relationship with the response variable. But none of them produced significant improvement to the model. So, I removed these terms.

Second order terms will be revisited during the Variable Transformation section.

## Interaction Terms

The passes.per.match stats only counts passes for each match that a player appeared. It doesn't count the minutes player appeared on the field during each match. But for the same passes.per.match, the fewer minutes players appeared in each match means the player has given more passes per minute during the match, which represents higher efficiency. So, I added the interaction term combining Mn.Ap with Passes.per.match. The adjusted model is the following:

```
lm(formula = log(Transfer.Values) ~ Age + Gls + Losses + Apps +
    Mn.Ap + Goals.with.right.foot + SoT90 + PK + Passes.per.match +
    MnAp_PPM, data = cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.079034 -0.019698  0.000792  0.019231  0.065847

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.768e+00  3.708e-02  74.638  < 2e-16 ***
```

```
Age                    -5.288e-03  1.316e-03  -4.020 0.000117 ***
Gls                     6.107e-03  1.681e-03   3.633 0.000454 ***
Losses                 -3.225e-04  1.663e-04  -1.939 0.055503 .
Apps                    1.021e-03  4.625e-04   2.207 0.029755 *
Mn.Ap                   1.111e-03  4.864e-04   2.284 0.024611 *
Goals.with.right.foot   1.880e-03  6.253e-04   3.006 0.003386 **
SoT90                   3.092e-02  6.833e-03   4.526 1.74e-05 ***
PK                     -8.571e-03  2.995e-03  -2.861 0.005189 **
Passes.per.match        3.425e-03  8.779e-04   3.901 0.000179 ***
MnAp_PPM               -2.898e-05  1.228e-05  -2.360 0.020341 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03116 on 95 degrees of freedom
Multiple R-squared:  0.7353,    Adjusted R-squared:  0.7075
F-statistic: 26.39 on 10 and 95 DF,  p-value: < 2.2e-16
```

The added variable MnAp_PPM is significant. The new model has adj-R2 improved from 0.6959 to 0.7075 and RSE reduced from 0.03177 to 0.03116. Therefore I will keep the newly added interaction term.

## Explanatory Variable Transformation

By regression testing the model, I re-added Recoveries to the model. It made model more complete by adding defensive factor and improved adj-R2 without adding multicollinearity.

```
lm(formula = log(Transfer.Values) ~ Age + Gls + Losses + Apps +
    Mn.Ap + Recoveries + Goals.with.right.foot + SoT90 + PK +
    Passes.per.match + MnAp_PPM, data = cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.076784 -0.016400  0.001708  0.016378  0.068915

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.802e+00  3.535e-02  79.259  < 2e-16 ***
Age                    -6.405e-03  1.249e-03  -5.127 1.57e-06 ***
Gls                     6.092e-03  1.558e-03   3.911 0.000174 ***
Losses                 -9.996e-04  2.265e-04  -4.414 2.71e-05 ***
Apps                    8.886e-04  4.298e-04   2.068 0.041422 *
Mn.Ap                   1.265e-03  4.523e-04   2.798 0.006239 **
Recoveries              7.250e-05  1.777e-05   4.080 9.43e-05 ***
Goals.with.right.foot   1.060e-03  6.132e-04   1.729 0.087036 .
SoT90                   3.098e-02  6.331e-03   4.893 4.10e-06 ***
```

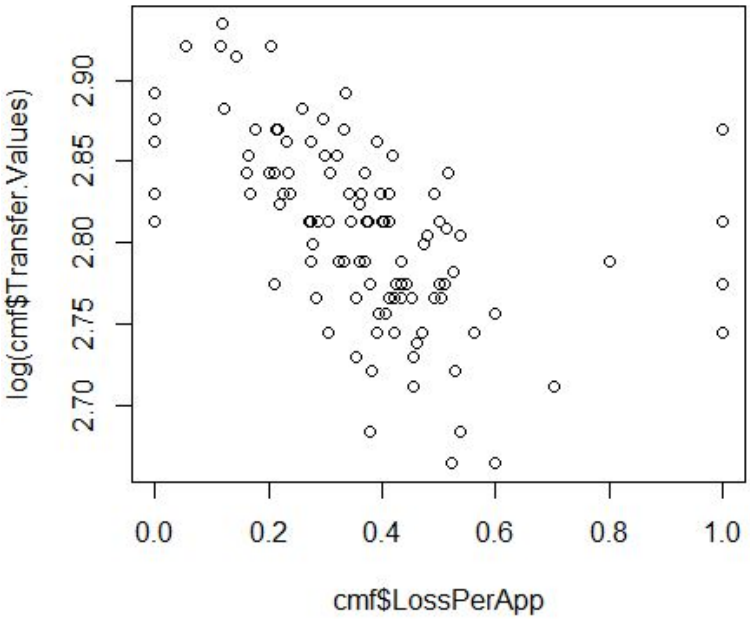
```
PK                   -7.235e-03  2.795e-03  -2.589 0.011156 *
Passes.per.match      2.937e-03  8.221e-04   3.573 0.000559 ***
MnAp_PPM             -2.956e-05  1.138e-05  -2.597 0.010909 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02887 on 94 degrees of freedom
Multiple R-squared:  0.7752,    Adjusted R-squared:  0.7488
F-statistic: 29.46 on 11 and 94 DF,  p-value: < 2.2e-16
```

I then added transformed variable LossPerApp = Losses / Appearances, since for the same number of losses, the more appearances, the less percentage of losses, indicating higher ability of a player.

Goals.with.right.foot was also removed from the model since it was no longer significant.

```
lm(formula = log(Transfer.Values) ~ Age + Gls + Losses + Apps +
    Mn.Ap + Recoveries + SoT90 + PK + Passes.per.match + MnAp_PPM +
    LossPerApp, data = cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.068192 -0.017892  0.000999  0.016301  0.076791

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.820e+00  3.505e-02  80.452  < 2e-16 ***
Age              -6.257e-03  1.210e-03  -5.170 1.32e-06 ***
Gls               5.752e-03  1.526e-03   3.768 0.000287 ***
Losses           -9.027e-04  2.187e-04  -4.127 7.95e-05 ***
Apps              1.023e-03  4.214e-04   2.428 0.017085 *
Mn.Ap             1.339e-03  4.336e-04   3.089 0.002643 **
Recoveries        7.511e-05  1.658e-05   4.530 1.73e-05 ***
SoT90             3.081e-02  6.088e-03   5.060 2.07e-06 ***
PK               -5.489e-03  2.651e-03  -2.071 0.041130 *
Passes.per.match  2.780e-03  8.012e-04   3.470 0.000788 ***
MnAp_PPM         -3.136e-05  1.086e-05  -2.889 0.004802 **
LossPerApp       -4.641e-02  1.637e-02  -2.834 0.005624 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02815 on 94 degrees of freedom
Multiple R-squared:  0.7863,    Adjusted R-squared:  0.7612
F-statistic: 31.44 on 11 and 94 DF,  p-value: < 2.2e-16
```

After testing the model, it looks like the newly added variable also fit a curved relationship with the response variable. So, I also added the second-order term of this variable

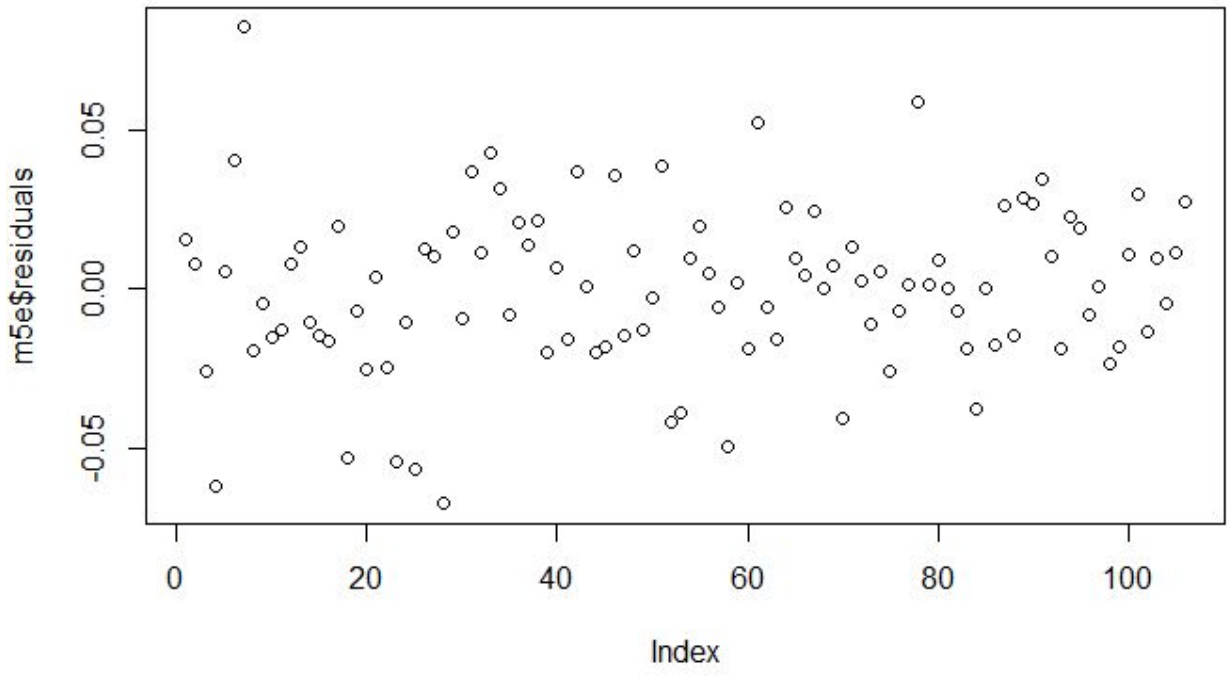



```
lm(formula = log(Transfer.Values) ~ Age + Gls + Losses + Apps +
    Mn.Ap + Recoveries + SoT90 + PK + Passes.per.match + MnAp_PPM +
    LossPerApp + LossPA2, data = cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.067515 -0.015571  0.001078  0.013540  0.082021

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.834e+00  3.456e-02  82.001  < 2e-16 ***
Age              -5.727e-03  1.195e-03  -4.791 6.25e-06 ***
Gls               5.283e-03  1.496e-03   3.532 0.000644 ***
Losses           -7.884e-04  2.174e-04  -3.626 0.000470 ***
Apps              9.812e-04  4.101e-04   2.393 0.018742 *
Mn.Ap             1.351e-03  4.217e-04   3.204 0.001855 **
Recoveries        7.148e-05  1.619e-05   4.416 2.71e-05 ***
SoT90             2.929e-02  5.950e-03   4.923 3.68e-06 ***
PK               -5.042e-03  2.584e-03  -1.951 0.054015 .
Passes.per.match  2.573e-03  7.833e-04   3.285 0.001438 **
MnAp_PPM         -3.056e-05  1.056e-05  -2.893 0.004748 **
LossPerApp       -1.599e-01  4.758e-02  -3.361 0.001129 **
LossPA2           1.092e-01  4.312e-02   2.531 0.013040 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02738 on 93 degrees of freedom
Multiple R-squared:    0.8,     Adjusted R-squared:  0.7742
F-statistic: 31.01 on 12 and 93 DF,  p-value: < 2.2e-16
```

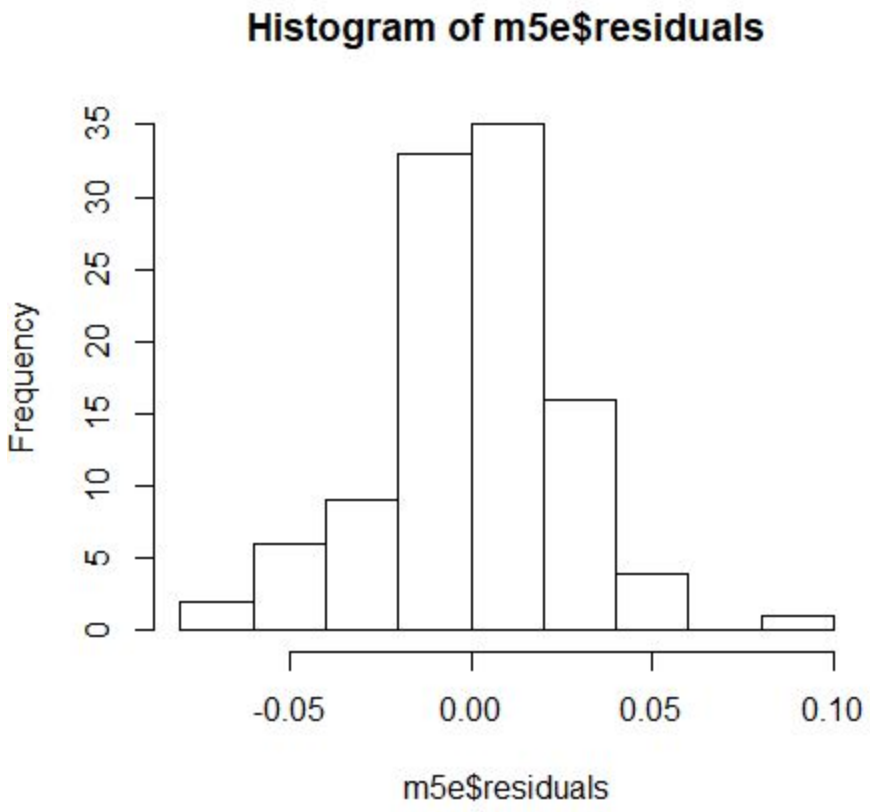
## Residual Analysis

The residual plot:

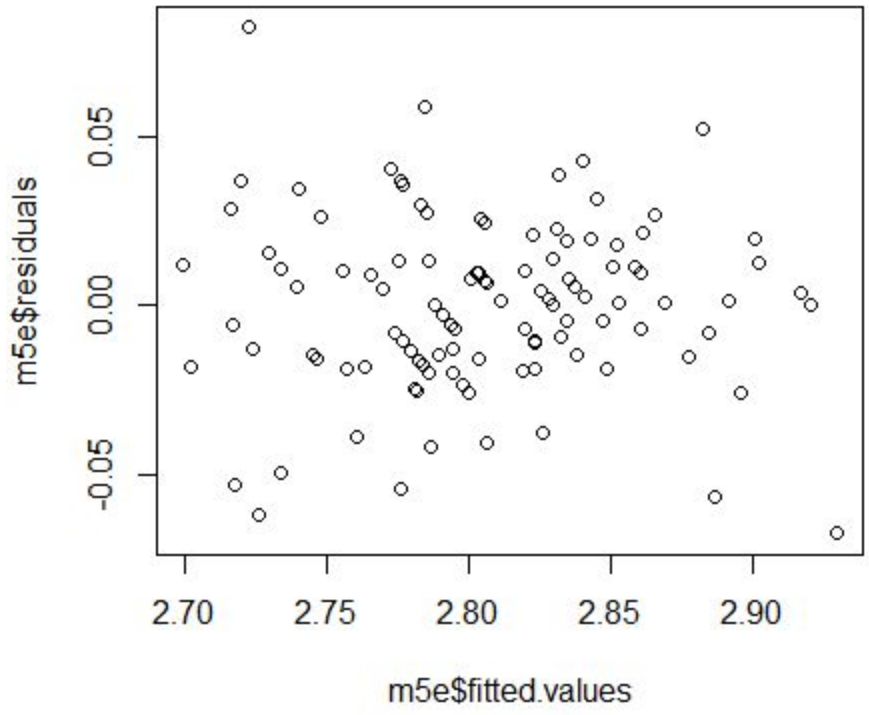



Zero-Mean: The sum of residuals is -7.958044e-17, it is approximately 0.

Normality: The residuals are normally distributed




Homoscedasticity: The residuals distribution is homoscedastic.



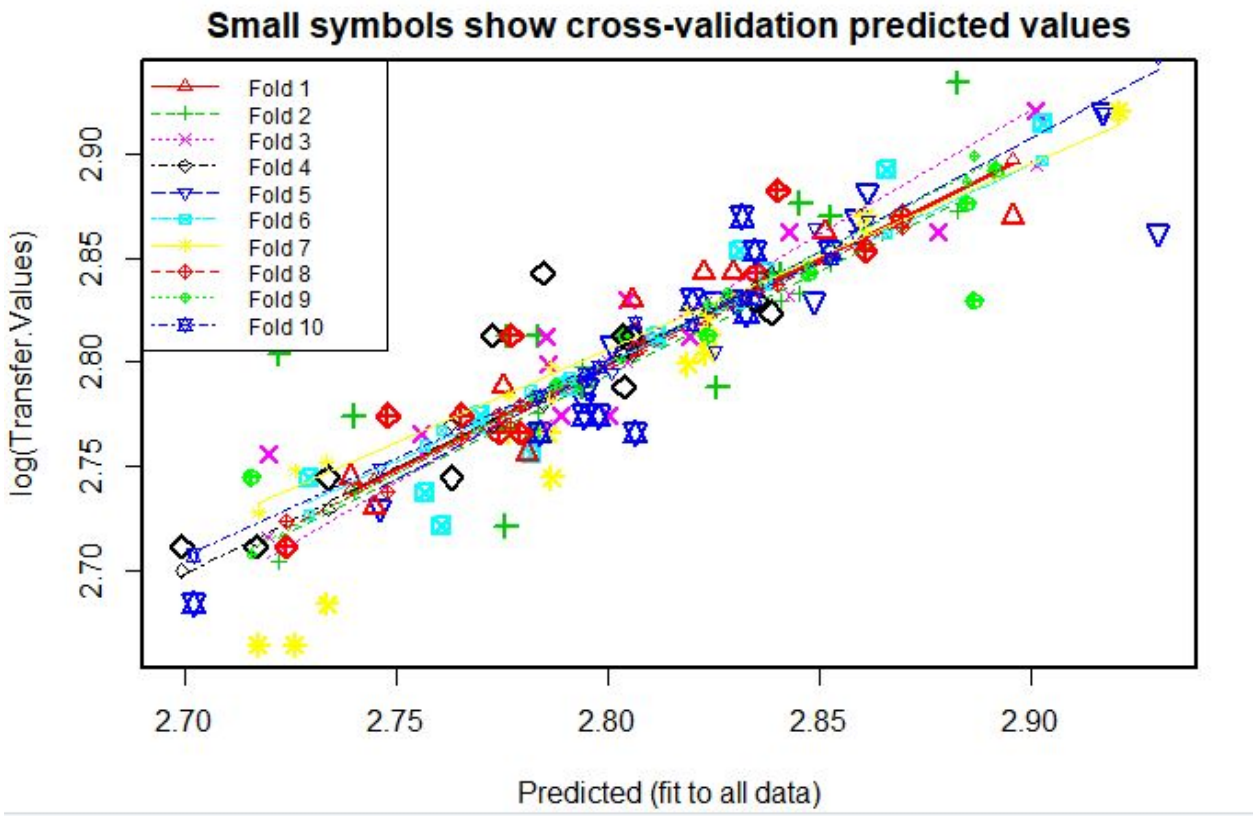

Independence: The residuals are independent, according to the Durbin-Watson Test result.

```
lag Autocorrelation D-W Statistic p-value
   1      0.02756011      1.930741   0.748
 Alternative hypothesis: rho != 0
```

## Model Validation

```
> out <- cv.lm(data=cmf, form.lm = m5e, plotit = "Observed", m = 10)
Analysis of Variance Table

Response: log(Transfer.Values)
                 Df Sum Sq Mean Sq F value  Pr(>F)
Age               1 0.0305  0.0305   40.76 6.7e-09 ***
Gls               1 0.1140  0.1140  152.11 < 2e-16 ***
Losses            1 0.0009  0.0009    1.17 0.28255
Apps              1 0.0093  0.0093   12.36 0.00068 ***
Mn.Ap             1 0.0006  0.0006    0.85 0.35868
Recoveries        1 0.0664  0.0664   88.57 3.6e-15 ***
SoT90             1 0.0213  0.0213   28.46 6.7e-07 ***
PK                1 0.0058  0.0058    7.79 0.00638 **
Passes.per.match  1 0.0110  0.0110   14.70 0.00023 ***
MnAp_PPM          1 0.0078  0.0078   10.41 0.00173 **
LossPerApp        1 0.0064  0.0064    8.49 0.00446 **
```

```
LossPA2           1 0.0048  0.0048    6.41 0.01304 *
Residuals        93 0.0697  0.0007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



fold 1
Sum of squares = 0    Mean square = 0    n = 10

fold 2
Sum of squares = 0.03    Mean square = 0    n = 11

fold 3
Sum of squares = 0.02    Mean square = 0    n = 11

fold 4
Sum of squares = 0.01    Mean square = 0    n = 11

fold 5
Sum of squares = 0.01    Mean square = 0    n = 11

fold 6
Sum of squares = 0.01    Mean square = 0    n = 11

fold 7
Sum of squares = 0.02    Mean square = 0    n = 11

fold 8
Sum of squares = 0.01    Mean square = 0    n = 10

fold 9
Sum of squares = 0.01    Mean square = 0    n = 10

fold 10
Sum of squares = 0.01    Mean square = 0    n = 10

Overall (Sum over all 10 folds)
     ms
0.00103
```




From the plot, we can see that most of the points are fitted into the diagonal line with only a few outliers, indicating the prediction accuracy of the model is good. All 10 folds validation has a mean square of appx. 0, which also indicates a very good model accuracy.

## The Final Model

```
lm(formula = log(Transfer.Values) ~ Age + Gls + Losses + Apps +
    Mn.Ap + Recoveries + SoT90 + PK + Passes.per.match + MnAp_PPM +
    LossPerApp + LossPA2, data = cmf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.067515 -0.015571  0.001078  0.013540  0.082021

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.834e+00  3.456e-02  82.001  < 2e-16 ***
Age              -5.727e-03  1.195e-03  -4.791 6.25e-06 ***
Gls               5.283e-03  1.496e-03   3.532 0.000644 ***
Losses           -7.884e-04  2.174e-04  -3.626 0.000470 ***
Apps              9.812e-04  4.101e-04   2.393 0.018742 *
Mn.Ap             1.351e-03  4.217e-04   3.204 0.001855 **
```

```
Recoveries         7.148e-05  1.619e-05   4.416 2.71e-05 ***
SoT90              2.929e-02  5.950e-03   4.923 3.68e-06 ***
PK                -5.042e-03  2.584e-03  -1.951 0.054015 .
Passes.per.match   2.573e-03  7.833e-04   3.285 0.001438 **
MnAp_PPM          -3.056e-05  1.056e-05  -2.893 0.004748 **
LossPerApp        -1.599e-01  4.758e-02  -3.361 0.001129 **
LossPA2            1.092e-01  4.312e-02   2.531 0.013040 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02738 on 93 degrees of freedom
Multiple R-squared:    0.8,     Adjusted R-squared:  0.7742
F-statistic: 31.01 on 12 and 93 DF,  p-value: < 2.2e-16
```

## Evaluation

The F-Test looks good with p-value < 2.2e-16 < 0.05, meaning we can reject the null hypothesis that all betas equal to zero and accept the alternative that at least one beta is not zero. The Adjusted R-squared is  0.7742, indicating 77.42% of the variability in transfer-value is explained by the model, which is a very good value. Compared to the initial model, the final model improved predictability by 0.17%, from 0.7725 to 0.7742, reduced the number of variables from 43 to 12, and reduced the Residual standard error by 93.97% from 0.4537 to 0.02738. With reduced total number of variables, the number of significant variables increased from 5 to 12, which gives the model more dimensions to interpretation. The residual analysis is good, showing zero-mean, normality, independence and homoscedasticity. 10-fold validation shows good predictability and consistency of the model, with an average of approximately 0 MSE for each fold, and a mostly fit plot.

# 8. Kaushik Bandaru: Forward Model

## Background

Forwards are the frontline players forwards and defenders. Their job covers almost mostly offense i.e. scoring goals , assists and creating chances. During the attack, forwards play as the mastermind. Their jobs include running for  through balls , finding space behind the opponent's defense, making runs to drag the opponent's defensive players and scoring goals or creating them.  Forwards include left-wingers, strikers, right-wingers.

There are more than 100 forwards in the dataset. Forwards are the goal-scoring players who have the highest number of goals scored when combined. Player status such as the number of goals, assists, headed goals, chances created, etc. are more significant in deciding a forward's transfer value. For example, Relations such as the number of games played, and the number of goals scored are the most important when the transfer value of a player is being predicted.Some wingers may be given the duty to track back the opponents wingers and some defensive responsibilities.

Based their versatile roles in the game, the features to predict forward transfer values are expected to cover a wide range of variables, including both attacking, teamplaying. Therefore, the total number of features required to start the model is large and trimming was needed to build a better and appropriate model. Further partition of the category is possible, by distinguishing left wingers, right wingers and strikers.

## Model Building

I have 94 variables but I have to remove a lot of variables in order to have a better model. I will be considering player stats that only apply to forwards, such as goals scored, assists, etc. The independent variables I selected are country, club, age, appearances, starts, subs, goals, assists, goals per 90, passes, big chances created, crosses, cross accuracy, through balls, accurate long passes, yellow cards, red cards, fouls, offsides, headed goals, hit the woodwork.

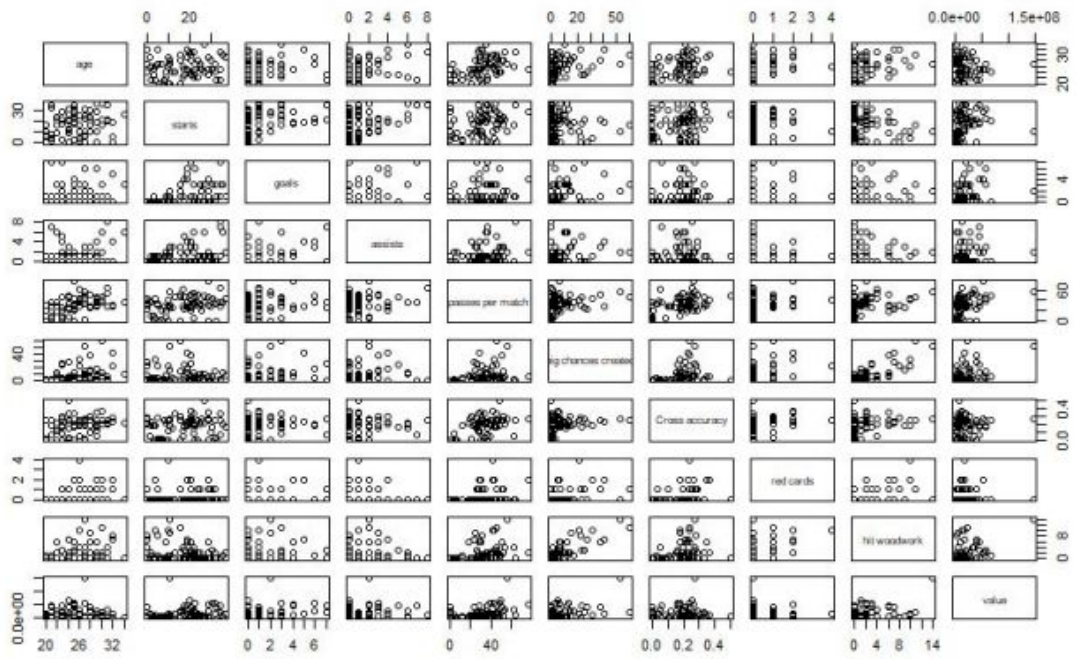
I got rid of all other variables that are relevant to players who belong to other positions as they don't help in determining the transfer value of forwards. For example, I didn't keep saves as one of the factors, as forwards don't goal keep and also removed other defense-related variables because nobody expects a forward to defend.

**Full Model**

```
Residuals:
      Min        1Q    Median        3Q       Max
-47582334  -7053843   -401546   5224890  59107684

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              39591415   19021632   2.081  0.04234 *
age                      -1921998     831800  -2.311  0.02485 *
starts                     560106     264917   2.114  0.03931 *
Subs                        22962     803535   0.029  0.97731
goals                     3624146    3253550   1.114  0.27044
assists                  -2135974    1344522  -1.589  0.11820
`goals per 90`          -12057448   69755909  -0.173  0.86344
passes                      -1419       3379  -0.420  0.67624
`passes per match`         622142     193741   3.211  0.00227 **
`big chances created`      835424     540355   1.546  0.12815
crosses                     -9712      13565  -0.716  0.47725
`Cross accuracy`        -40868007   22483279  -1.818  0.07487 .
`through balls`            -25676     165177  -0.155  0.87707
`accurate long pasballs`    31041      34142   0.909  0.36745
...18                     -239382     520791  -0.460  0.64768
`red cards`             -10643556    4376904  -2.432  0.01851 *
`headed goals`            -164863    1750462  -0.094  0.92533
`hit woodwork`            2674103    1282995   2.084  0.04207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15660000 on 52 degrees of freedom
Multiple R-squared:  0.6405,    Adjusted R-squared:  0.5229
F-statistic: 5.449 on 17 and 52 DF,  p-value: 1.045e-06
```

We can see here that our adjusted R-squared is 0.5229 The adjusted r squared value is 0.5229, which implies that 52.29% of the variability in the transfer values can be explained by the model. The F value is 5.449 on 17 and 52 degrees of freedom, and the p-value for the F statistic is 1.045e-06, which means that under a significance level of 0.05, we can conclude that at least one independent variable is significant. Most of the predictors, such as starts and through balls, were not significant under a significance level of 0.05 as they were mostly all greater than 0.05. Only passes per match, red cards, and age (0.02225) were significant under a significance level of 0.05. Given the large difference between adjusted r squared and multiple r squared (0.5229 vs 0.6405) and the high number of insignificant independent variables. I removed these insignificant independent variables one after another and tried for a better model later

**GPLOT:**




We can check the correlation between variables with the help of this plot. We can see that value and age have a high correlation. Hence, age is a good predictor to determine the value of a player. Let's build a model according to this assumption because it is generally the case in soccer too. Younger the player, higher the value

**Model 2:**

```
Residuals:
      Min        1Q    Median        3Q       Max
-40082793  -7640449   -741439   7406876  62603699

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           48434789   14538692   3.331 0.001483 **
age                   -2384693     623196  -3.827 0.000312 ***
starts                  601574     217822   2.762 0.007618 **
goals                  2908821    1254390   2.319 0.023823 *
assists               -2551649    1149997  -2.219 0.030292 *
`passes per match`      698818     142192   4.915 7.22e-06 ***
`big chances created`   515165     285977   1.801 0.076662 .
`Cross accuracy`      -41127144   19700083  -2.088 0.041083 *
`red cards`           -10970316    2866854  -3.827 0.000312 ***
`hit woodwork`          2771888    1142386   2.426 0.018271 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14970000 on 60 degrees of freedom
Multiple R-squared:  0.6208,    Adjusted R-squared:  0.5639
F-statistic: 10.91 on 9 and 60 DF,  p-value: 7.684e-10
```

This is our second model. Here, we can see that age, red cards, passes per match are the best predictive variables and our adjusted R-squared is also higher than our previous model. Age is the best predictor here. This is a good model but let's try building forward and backward selection models to check if we can build a better model that way.

Here, value ~ age+ starts + goals + passes per match + big chances created + cross accuracy + red cards + hit woodwork. The AIC value for big chances created is the lowest and its predicting that it is the best predictor variable to determine value followed by cross accuracy, assists, etc.

This is a good model as it even tells us the order in which these variables are good predictors

**Backward Selection Model:**

```
Start:  AIC=2322.27
value ~ age + starts + goals + assists + `passes per match` +
    `big chances created` + `Cross accuracy` + `red cards` +
    `hit woodwork`

                       Df  Sum of Sq        RSS    AIC
<none>                                1.3455e+16 2322.3
- `big chances created`  1 7.2771e+14 1.4183e+16 2324.0
- `Cross accuracy`       1 9.7735e+14 1.4432e+16 2325.2
- assists                1 1.1040e+15 1.4559e+16 2325.8
- goals                  1 1.2059e+15 1.4661e+16 2326.3
- `hit woodwork`         1 1.3203e+15 1.4775e+16 2326.8
- starts                 1 1.7104e+15 1.5165e+16 2328.7
- age                    1 3.2836e+15 1.6739e+16 2335.6
- `red cards`            1 3.2837e+15 1.6739e+16 2335.6
- `passes per match`     1 5.4164e+15 1.8871e+16 2343.9
```

Here, value ~ age+ starts + goals + passes per match + big chances created + cross accuracy + red cards + hit woodwork. The AIC value for big chances created is the lowest and its predicting that it is the best predictor variable to determine value followed by cross accuracy, assists, etc. This is a good model as it even tells us the order in which these variables are good predictors. Let's build a forward selection model now.

**Forward Selection Model**

```
Step:  AIC=2322.27
value ~ `big chances created` + starts + age + `passes per match` +
    `red cards` + `hit woodwork` + `Cross accuracy` + goals +
    assists
```

We can see that both forward and backward models agree that big chances created is the best predictor and the stepAIC value for big chances created in forward selection model is lower than that of backward selection model.

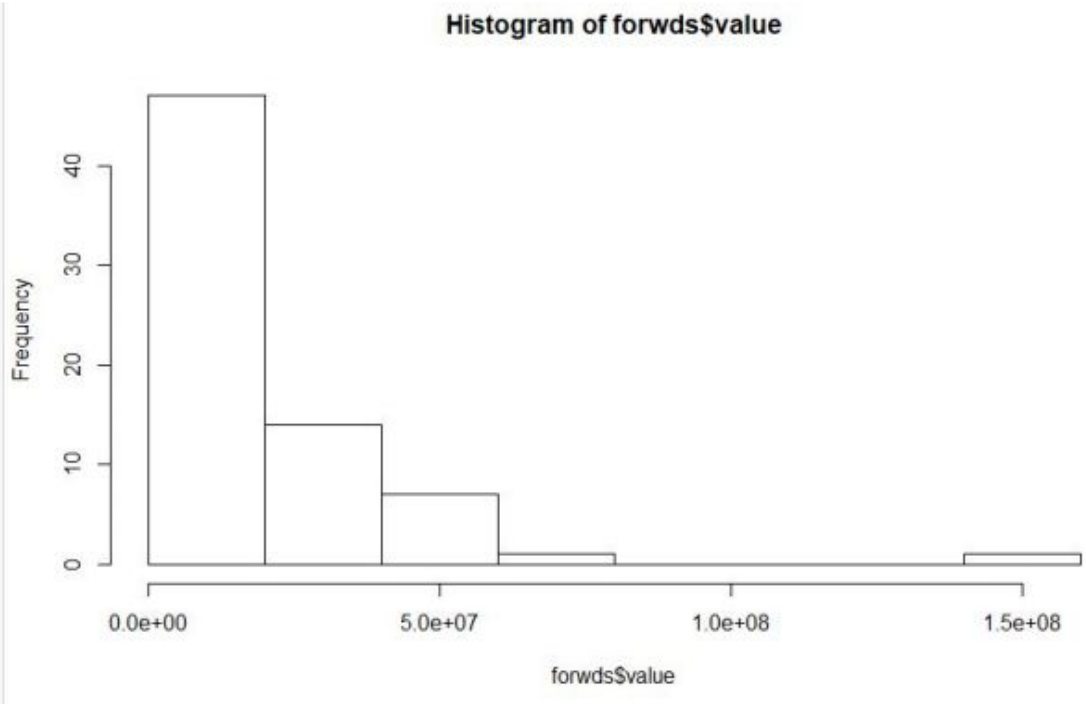
Therefore our final model says, value ~ big chances created + starts + age + passes per match + red cards + hit woodwork + cross accuracy + goals + assists. Hence we can conclude that big

chances created are the best predictor variable Hence, we can conclude that stats that only measure forward's performance such as big chances created, starts, and age are most significant in predicting a forward's transfer value. Age is the a deciding factor related to all the players not just forwards. It seems like big chances created is the biggest influential factor that increases transfer value. This makes sense as a forward is expected as more valuable if he creates more chances in a match. On the other hand, age and red cards are significant negative factors towards affecting transfer value. If a player's ages, he could be seen as less valuable to the team due to declining physical traits such as strength, stamina, etc.

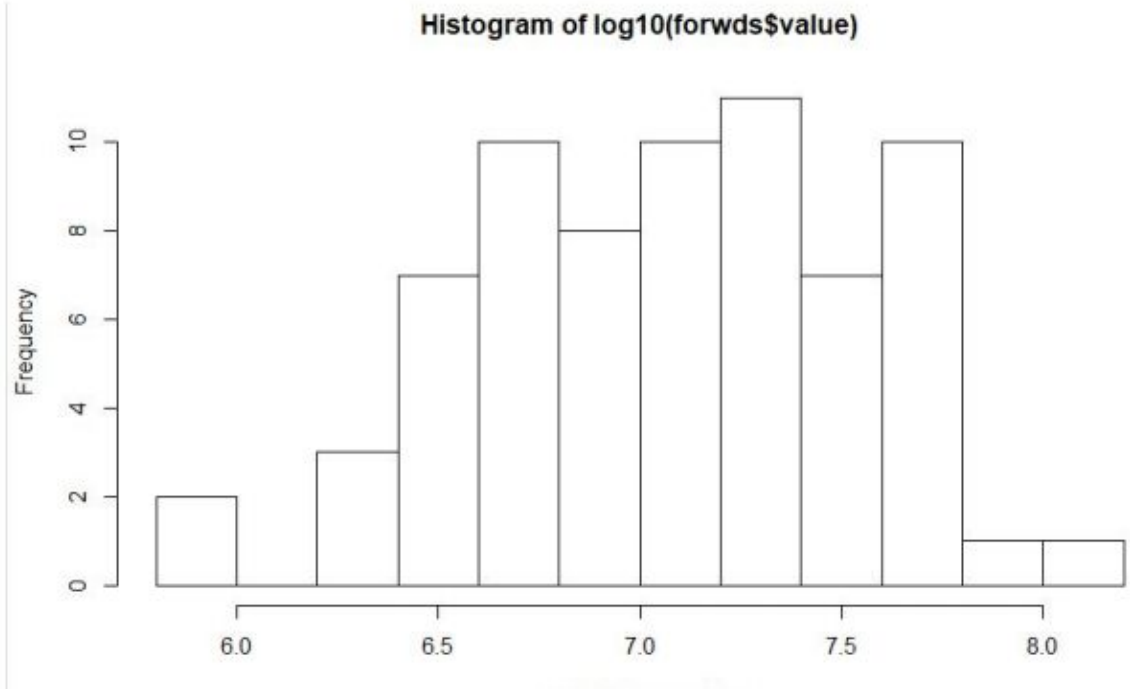
**Log Transformations**

I applied the log on transfer value data . The transfer values a look pretty skewed towards right, and transfer values can vary by hundreds of millions of dollars. By transforming this data, I can estimate the transfer values in terms of magnitude, which will reduce the skew in transfer value.

Without Log



With Log:




**Model after applying log to value**

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          18.122680   0.635278  28.527  < 2e-16 ***
age                  -0.146493   0.027231  -5.380 1.30e-06 ***
starts                0.043669   0.009518   4.588 2.33e-05 ***
goals                 0.160370   0.054811   2.926  0.00484 **
assists              -0.114830   0.050250  -2.285  0.02585 *
`passes per match`    0.033488   0.006213   5.390 1.25e-06 ***
`big chances created` 0.022851   0.012496   1.829  0.07241 .
`Cross accuracy`     -1.699141   0.860808  -1.974  0.05300 .
`red cards`          -0.243251   0.125269  -1.942  0.05686 .
`hit woodwork`        0.094169   0.049917   1.886  0.06407 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6543 on 60 degrees of freedom
Multiple R-squared:  0.6803,    Adjusted R-squared:  0.6324
F-statistic: 14.19 on 9 and 60 DF,  p-value: 6.194e-12
```

I created this model by transforming the values. After applying a log transformation on the transfer values, the value of the adjusted R squared increased to 0.6324, which means that 63.24% of the variability in transfer values can be explained by the model. The F statistic is 14.19, which is significant at a confidence level of 0.05. Due to the log transformation, the MSE decreased.

**Second Order Analysis**

Now, I checked for second order terms. I think age can be used as a second order term, as I think there is a relationship between transfer value and age. A player with not much experience would be not worth a lot as its just the beginning of his career, but his transfer value might increase as he becomes more a experienced player, but his value might also decrease as he turns old and passes his prime. Goals and Assists could also be second order terms, as a forward might be more valuable if he scores more goals and gives more assists in a game.

```
Residuals:
      Min        1Q    Median        3Q       Max
-31670554  -7545277    218329   6529857  31418727

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -10660353  103860967  -0.103  0.91868
age                        1389991    8288933   0.168  0.86754
Apps                       -291832     696666  -0.419  0.67720
starts                      701784     662223   1.060  0.29468
Subs                            NA         NA      NA       NA
goals                     10754772    4772284   2.254  0.02893 *
assists                     -57870    2979432  -0.019  0.98459
`goals per 90`           -65112727   62545096  -1.041  0.30318
passes                        1626       3087   0.527  0.60083
`passes per match`          454225     174065   2.610  0.01212 *
`big chances created`      1367551     474490   2.882  0.00594 **
crosses                      13347      13712   0.973  0.33533
`Cross accuracy`         -42477965   19394615  -2.190  0.03351 *
`through balls`             -83576     140980  -0.593  0.55614
`accurate long pasballs`     25559      30169   0.847  0.40118
...18                     -1155515     590662  -1.956  0.05639 .
`red cards`               -7349031    3816736  -1.925  0.06023 .
Fouls                       -20822      83452  -0.250  0.80406
offsides                  -1247918     245360  -5.086 6.28e-06 **
`headed goals`              744953    1486566   0.501  0.61862
`hit woodwork`             2516536    1090011   2.309  0.02540 *
age2                        -39700     161171  -0.246  0.80651
goals2                      -865381     532859  -1.624  0.11106
assists2                    -290624     461669  -0.630  0.53207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13150000 on 47 degrees of freedom
Multiple R-squared:  0.7709,    Adjusted R-squared:  0.6636
F-statistic: 7.187 on 22 and 47 DF,  p-value: 8.504e-09
```

I included the second order terms for age, goals, and assists into the model. The model with the second order terms show not much improvement in the adjusted r squared and f statistic, and mean squared error. None of the second order terms are significant. I'll remove the second order terms

**Residual Analysis**

```
> mean(model3$residuals)
[1] 2.885836e-17
>
```

I analyzed the residual for model. The residuals are evenly distributed around the x axis, so the residual distribution meets the assumption of homoscedasticity. The mean of the residuals is 0 and hence, the model meets the assumption of mean of residuals is zero.

```
        Durbin-Watson test

data:  model3
DW = 1.7903, p-value = 0.1804
alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin Watson test 1.79 is close to the value of 2, the p value of the test is 0.1804 which is higher than the significance level of 0.05, so we fail to reject the null hypothesis that the model doesn't have autocorrelation. I can assume the model meets the assumption that the residuals are independent. Age and starts are the common affective independent variables for all players irrespective of their position.

**Final model**

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          18.122680   0.635278  28.527  < 2e-16 ***
age                  -0.146493   0.027231  -5.380 1.30e-06 ***
starts                0.043669   0.009518   4.588 2.33e-05 ***
goals                 0.160370   0.054811   2.926  0.00484 **
assists              -0.114830   0.050250  -2.285  0.02585 *
`passes per match`    0.033488   0.006213   5.390 1.25e-06 ***
`big chances created` 0.022851   0.012496   1.829  0.07241 .
`Cross accuracy`     -1.699141   0.860808  -1.974  0.05300 .
`red cards`          -0.243251   0.125269  -1.942  0.05686 .
`hit woodwork`        0.094169   0.049917   1.886  0.06407 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6543 on 60 degrees of freedom
Multiple R-squared:  0.6803,    Adjusted R-squared:  0.6324
F-statistic: 14.19 on 9 and 60 DF,  p-value: 6.194e-12
```

We can see here that our adjusted R-squared is 0.6324 which implies that 63.24% of the variability in the transfer values can be explained by the model. The F value is 14.19 on 9 and 60 degrees of freedom, and the p-value for the F statistic is 6.194e-12, which means that under a significance level of 0.05, we can conclude that at least one independent variable is significant. Most of the predictors, such as starts and through balls, were not significant under a significance

level of 0.05 as they were mostly all greater than 0.05. Only passes per match, red cards, and age (0.02225) were significant under a significance level of 0.05.

Age and starts are the common effective independent variables for all players irrespective of their position. So goals, assists and passes per match are the most important independent variables which will help us predict the transfer value of a forward. This totally makes sense as strikers are expected to score goals and wingers are expected to assist and create goals with the striker by passing the ball to and fro.

# 9. Discussion

## General

Overall, it seems like using a log transformation on the transfer values helps improve the prediction of goalkeeper transfer values (or player transfer values in general) due to the nature of the transfer values and its distribution. The transfer values already look right skewed, and transfer values can differ than hundreds of millions of dollars. Transforming the data causes the data to look for normal, and we can look at the transfer values in terms of magnitude, which can help the transfer value distribution appear less skewed.

Transfer values were significantly impacted negatively by age in all models. Soccer is a physical sport and even goalkeepers must be fit and move a lot through the match. They also need to depend on a lot on their reflexes. A person's reflexes and physical abilities peak in their mid to late 20s, and there are only three players under 25 in the data, so there is going to be a linear relationship in the data. The older a player is, the less valuable he will be. If a players ages, he could be less valuable to the team due to declining physical ability. Also, older players are less likely to be transferred from club to club due to clubs' wanting to keep the players more and wanting to sign players with potential.

Starts and Appearances are also generally influencing a player's value, especially for defenders, midfielders and forwards. It is reasonable since only the players with the highest quality are getting more chances to start or play in matches. Many players even transfer from big clubs to smaller clubs to earn more appearance time, because they know the lack of playing time will cost them greatly on their values.

## Goalkeepers

For goalkeepers, it seems after transforming the transfer values data only the age and the two clean sheets statistics were significant factors in determining transfer value. None of the interaction terms or second order terms were significant, nor did they significantly increase the adjusted r squared of the model. Clean sheets, for both the 18-19 season and for the entire career, are the biggest positive factors that increases transfer value. This makes sense as a goalkeeper is perceived as more valuable if he doesn't concede a goal in a match. It is the most important statistic to decide the quality of a goalkeeper, as the Golden Glove awards in the Premier League and in other tournaments like the World Cup are given to the player with the most clean sheets. Even though there is a small sample size for the goalkeepers and the model only has three significant variables, the model's r squared and cross validation results look pretty good.

## Defenders

Passes per Match has a big positive influence on defenders' transfer values. This reflects the fact that more clubs are adopting the "playing out from the back" strategy that were brought into English Premier League by Pep Guardiola in 2016 and had great success. The strategy requires defenders to pass the ball from the back instead of making long balls to transform a defensive play into attacking play, and hence requires defenders to pass more frequently and accurately. It takes time for defenders to adjust to this style of play and it causes a shortage of supply, and subsequently increases the price.

## Midfielders

Gls, SoT90, PK are the attacking variables in the midfielder model, among which the SoT90 has the largest positive beta. It shows the importance of scoring goals in soccer games even for midfielders, whose role is primarily connecting the team. SoT90 has the larger impact on Gls indicating the market values more on players that have better shoot accuracy than scoring more goals. It doesn't mean scoring goals is less important, but it has a weaker impact in evaluating player's values. I makes sense since scoring goals in real games is depending on a lot of factors, including luck, opposition defense and goalkeeping. A good opposition goalkeeper, who would effectively reduce the number of goals, doesn't reduce the importance of an accurate shooter who forced the saves. PK has a negative beta and is on the boundary of significance ($p = 0.054$), but I'm keeping it in the model, since goals scored by penalty kicks are counted in goals but are much easier to score, which reflects less of a player's quality.

Losses Per Appearance has a second-order term in the model. It has generally negative impact on the model but following a curved pattern. The curved pattern shows that a player with very low losses per appearance value could have been contributing to a very successful team, which indicating the high quality of the midfielder. But as the number decreases, the values of players decrease significantly at first and smooths out after that, indicating they might have been playing for an ordinary team, and the stat may or may not reflect the quality of the player, since ordinary team's problem may or may not in their midfield. It shows consistency of the Clean Sheets in the goalkeeper's analysis, since a clean sheet is effectively indicating a non-losing match.

Among the defensive variables, the most important and the only one included in the midfielder model is Recoveries. It outperformed some of the classic defensive stats such as Number of Tackles and Tackle Success Rate. It shows for midfielders, defensive work is important but doesn't need to be extensive, as heavy defensive actions such as tackle, shot blocking and clearances are not emphasized here. It reflects midfielders are evaluated more from their contribution to attack than defense, especially when considering recoveries as not a pure defensive feature but a combination of defensive and attacking feature, because the player

recovers the ball immediately transitions the teamplay from defense into attack. Sometimes recoveries create even more dangerous attacks than normal because the opposition team during a recovery is often not in a good defensive position and hence is more vulnerable to fast counter attacks.

Assists related stats are not included in the midfielder model. However, when a midfielder's performance was evaluated, they are mostly mentioned and emphasized. Many midfielders even publicly claimed that they think assists is more important for them than scoring goals themselves. The reason it is not included is because it has a high correlation with goal scoring variables and hence would increase the multicollinearity of the model. It makes sense because midfielders good at scoring usually good at assists as well, so the scoring variables contain more information than the assists ones.

The only teamplay feature included in the model is passes per match. It has an interaction term with Minutes Appearance showing that for same passes per match, the less minutes played in each match, the more valuable the player is. It shows the importance of passes and pass frequency for evaluating the midfielder. A soccer team has 11 players on the field and players have a lot of choices to make the pass. A more frequent passer implies a more frequent pass receiver, and it then indicates the player is more trusted by the team than others, which reflects the player's quality.

## Forwards

Forwards transfer values were strongly and positively impacted by goals and big chances created. This makes a good point because the salient job for a forward is to produce goals, if not just scoring directly by themselves. Goals has the biggest impact since scoring goals directly is the most effective way to get the job done. After that is Big Chances Created, which is also very important, especially for a forward that is not a striker, but a winger, whose job is to drag the opposition defense on the flank and create space in the middle for other forwards to have more chances.

# 10. Conclusions

While we all used used a log transformation on the transfer value in all our models to normalize the transfer value distributions, we found that players from different positions need different sets of features to predict their values, as we generated four mostly different prediction models. For goalkeepers, clean sheets has the most positive impact on the model. Passes per match, Starts and Tackles are the variables that significant for defenders. Midfielder are valued more for their shots on target per 90 minutes and their losses per minutes of appearance. The only one common variable having big impact on players' value is Age, since younger players possess better physicality, reflexes, and potentials

# Appendix: The Data Dictionary

| Variable Name | Variable Type | Description |
|---|---|---|
| Full Name | Categorical | The first and last name of the player |
| Country code (2 letters) | Categorical | The player's nationality, either what country the player has citizenship in or what national team the player represents in FIFA (the latter takes priority over the former)<br><br>This column has the 2 letter country code supplied by iso with the home nations (England, Scotland, Wales, and Northern Ireland) have 3 letter names |
| Country code (3 letters) | Categorical | 3 Letter fifa country code |
| Country | Categorical | the full country name |
| Pos1 | Categorical | What positions the players have played in<br>GK - Goalkeeper |

| | | |
|---|---|---|
| | | DF - Defender<br>MF - Midfielder<br>FW - Forward<br>FB - Fullback<br>CD - Central Defender<br>DM - Defensive Midfielder<br>CM - Central Midfielder<br>WM - Wide Midfielder<br>AM - Attacking Midfielder |
| Pos2 | Categorical | |
| Pos3 | Categorical | Some players have played in multiple positions, so there are three columns for the positions |
| Squad | Categorical | The club the player last played for in the 18-19 season |
| Age | Numeric | Age at the season start (August 1) for league play |
| Born | Numeric | Birth Year |
| Apps18 | Numeric | Appearances<br>Number of matches the player has appeared in |
| Starts18 | Numeric | Games started by player |
| Subs18 | Numeric | Substitutions: Games the player played in, but did not start |
| Min18 | Numeric | Minutes Played |
| min/app18 | Numeric | Minutes per appearance |
| gls18 | numeric | Goals scored |
| ast18 | numeric | assists |
| pk18 | numeric | Penalty kicks |

| | | |
|---|---|---|
| pkat18 | numeric | Penalty kicks attempted |
| Fls18 | numeric | fouls |
| CrdY18 | numeric | Yellow cards |
| CrdR18 | numeric | Red cards |
| SoT18 | numeric | Shots on target |
| Gls90.18 | numeric | Goals per 90 minutes |
| G+A90.18 | numeric | Goals plus assists per 90 minutes |
| G-PK90.18 | numeric | Goals minus penalty goals per 90 minutes |
| G+A-PK90.18 | numeric | Goals plus assists minus penalty goals per 90 minutes |
| SoT90.18 | numeric | Shots on target per 90 minutes |
| Fls90.18 | numeric | Fouls per 90 minutes |
| Crd90.18 | numeric | Cards per 90 minutes<br>This double counts red cards given if a player gets multiple yellow cards in a match |
| GA18 | Numeric | Goals Against |
| GA90.18 | Numeric | Goals Against per 90 minutes |
| SoTA90 | Numeric | Shots on target |
| Save18 | Numeric | (Shots on target against- goals against)/shots on target against |
| W18 | Numeric | wins |
| D18 | Numeric | draws |
| L18 | Numeric | losses |

| | | |
|---|---|---|
| CS_A18 | Numeric | Clean sheets against |
| CS18 | Numeric | Percentage of matches that result in clean sheets |

| | | |
|---|---|---|
| Position | Categorical | General Info: Player's positions: Goalkeeper, Defender, Midfielder, Forward |
| Club | Categorical | General Info: Player's latest club |
| Appearances | Numeric | General Info: Total match appeared |
| Wins | Numeric | General Info: Total wins |
| Losses | Numeric | General Info: Total losses |
| Saves | Numeric | Goalkeeper-specific variable |
| Penalties Saved | Numeric | Goalkeeper-specific variable |
| Punches | Numeric | Goalkeeper-specific variable |
| High Claims | Numeric | Goalkeeper-specific variable |
| Caches | Numeric | Goalkeeper-specific variable |
| Sweeper Clearance | Numeric | Goalkeeper-specific variable |
| Goals Conceded | Numeric | Defence variable |
| Clean Sheets | Numeric | Defence variable |
| Tackles | Numeric | Defence variable |
| Tackle Success | Numeric | Defence variable |
| Last Man Tackles | Numeric | Defence variable |
| Blocked Shots | Numeric | Defence variable |

| | | |
|---|---|---|
| Interceptions | Numeric | Defence variable |
| Clearances | Numeric | Defence variable |
| Headed Clearance | Numeric | Defence variable |
| Clearances Off Line | Numeric | Defence variable |
| Recoveries | Numeric | Defence variable |
| Duels Won | Numeric | Defence variable |
| Duels Lost | Numeric | Defence variable |
| Successful 50/50s | Numeric | Defence variable |
| Aerial Battles Won | Numeric | Defence variable |
| Aerial Battles Lost | Numeric | Defence variable |
| Own Goals | Numeric | Defence variable |
| Errors Leading to Goal | Numeric | Defence variable |
| Assists | Numeric | Team-play variable |
| Passes | Numeric | Team-play variable |
| Passes Per Match | Numeric | Team-play variable |
| Big Chances Created | Numeric | Team-play variable |
| Crosses | Numeric | Team-play variable |
| Cross Accuracy | Numeric | Team-play variable |
| Through Balls | Numeric | Team-play variable |
| Accurate Long Balls | Numeric | Team-play variable |
| Yellow Cards | Numeric | Discipline variable |
| Red Cards | Numeric | Discipline variable |
| Fouls | Numeric | Discipline variable |
| Offsides | Numeric | Discipline variable |
| Goals | Numeric | Attack variable |

| | | |
|---|---|---|
| Goals per Match | Numeric | Attack variable |
| Headed Goals | Numeric | Attack variable |
| Goals with Right Foot | Numeric | Attack variable |
| Goals with Left Foot | Numeric | Attack variable |
| Hit Woodwork | Numeric | Attack variable |
| Big Chances Missed | Numeric | Attack variable |
| Value | Numeric | Current transfer Value of the player on FIFA in dollars |
| Cost | Numeric | Cost of the player the Fantasy Premier League in pounds |
| Points | Numeric | Number of Fantasy points scored in the 18-19 Premier League Season |