Premier League Player Valuation: Are They Worth It?

STAT 3220 – University of Virginia
Fall 2024

Saketh Vangara, Stuart Wells, Sai Chundi, Eshaan Sharma

Introduction

- Understanding what drives soccer players' transfer market values can help clubs make smarter decisions when buying or selling players. Knowing whether factors like age, position, and minutes played influence value is essential for improving contract negotiations and building balanced squads.
- Our project examines whether older players have a higher transfer market value than younger players, whether offensive players are valued higher than defensive players, and whether players with more minutes played command greater transfer value.
- Younger players are often seen as investments (Metelski, 2021), offensive players like forwards tend to command higher fees (Zhou, 2019), and greater playing time boosts visibility and demand (Besson, Ravenel, Poli, 2020), making these variables crucial for ensuring fair valuations and financial stability for clubs.

Research Questions

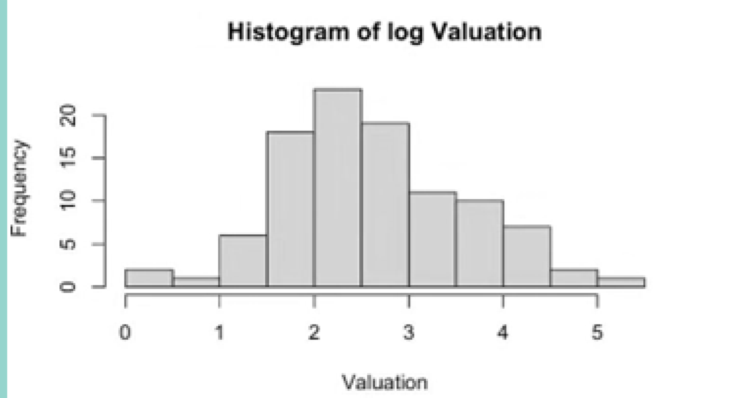
1. Do older players have a higher transfer market value than younger players, on average?
2. Do offensive players have a higher transfer market value than defensive players?
3. Do players with a higher play time (minutes played) have a greater transfer market value than players with a lower play time?

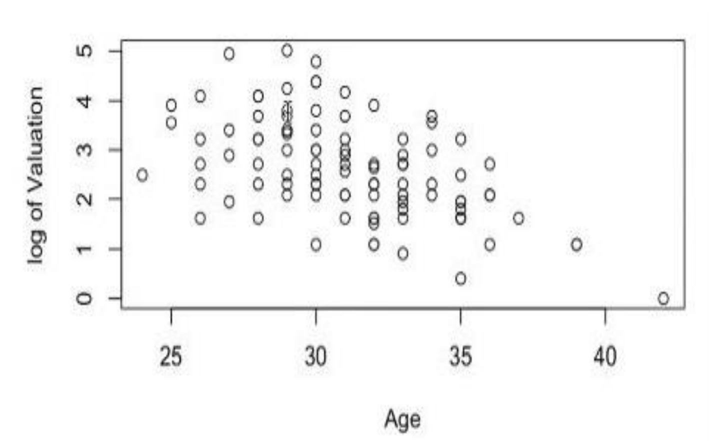
Data Summary

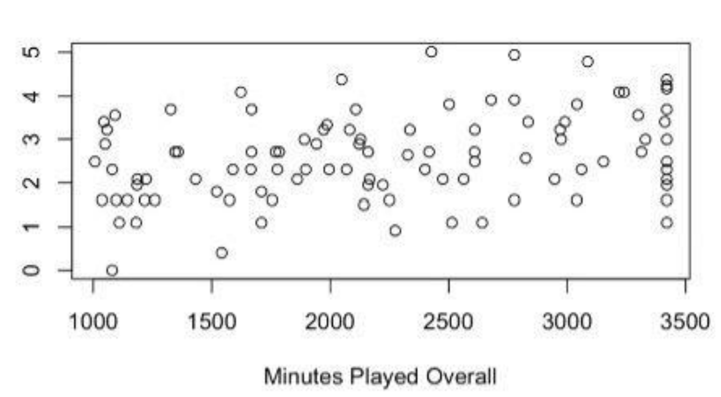
Population of Interest
- Soccer players in the premier league in the 2018-19 season

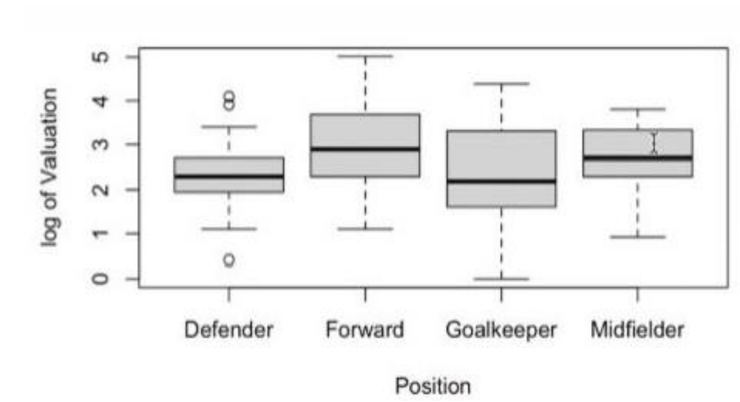
Data
- Obtained data from Footystats which we filtered based on our objective

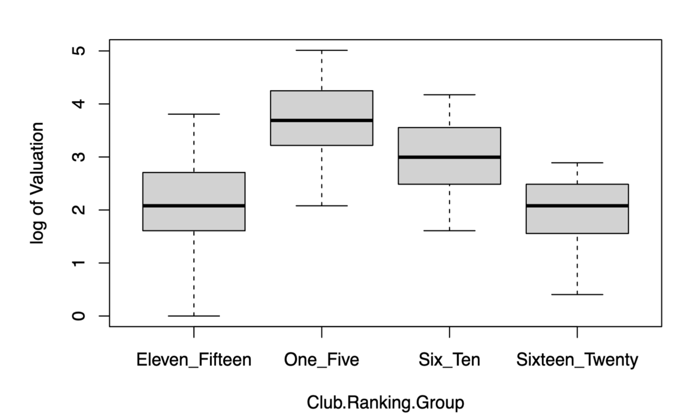
Manipulations
- Clustered clubs into 4 categorial levels based on the club rankings at the end of the season
- Assigned players as either 'England' or 'Non-England' players to determine how their nationality effected their valuation
- As our initial model was skewed towards the right and had a gap in the data, we applied a log transformation to our data to make it closer to normal

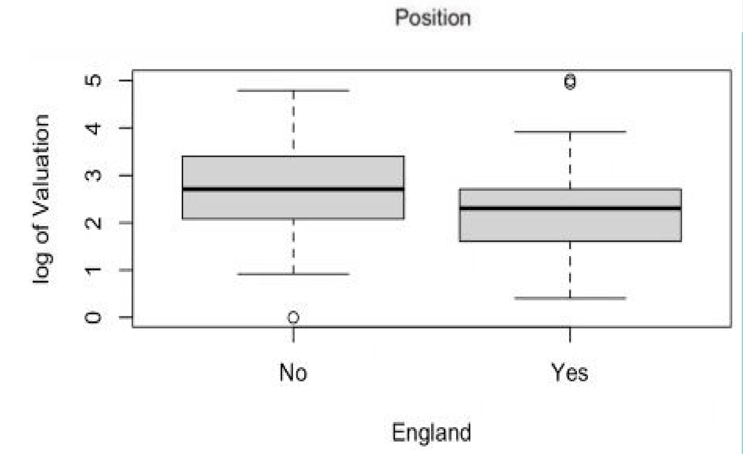












Multiple Linear Regression Analysis

Stage One: Quantitative Variables

Initial: $\log(\text{Valuation}) = \beta_0 + \beta_1\text{Age} + \beta_2\text{MinutesPlayedOverall} + \beta_3\text{AppearancesOverall} + \beta_4\text{GoalsOverall} + \beta_5\text{CleanSheetsOverall} + \beta_6\text{ConcededOverall} + \beta_7\text{GoalsInvolvedPer90Overall} + \beta_8\text{MinPerMatch} + \varepsilon$

Final: $\log(\text{Valuation}) = \beta_0 + \beta_1\text{CleanSheetsOverall} + \beta_2\text{Age} + \beta_3\text{GoalsInvolvedPer90Overall} + \beta_4\text{MinPerMatch} + \varepsilon$

Stage Two: Qualitative Variables

Initial: $\log(\text{Valuation}) = \beta_0 + \beta_1\text{CleanSheetsOverall} + \beta_2\text{Age} + \beta_3\text{GoalsInvolvedPer90Overall} + \beta_4\text{MinPerMatch} + \beta_5\text{Defender} + \beta_6\text{Forward} + \beta_7\text{Goalkeeper} + \beta_8\text{Midfielder} + \beta_9\text{ClubRankingOneFive} + \beta_{10}\text{ClubRankingSixTen} + \beta_{11}\text{SixteenTwenty} + \varepsilon$

Final: $\log(\text{Valuation}) = \beta_0 + \beta_1\text{CleanSheetsOverall} + \beta_2\text{Age} + \beta_3\text{GoalsInvolvedPer90Overall} + \beta_4\text{MinPerMatch} + \beta_5\text{ClubRankingOneFive} + \beta_6\text{ClubRankingSixTen} + B_7\text{ClubRankingSixteenTwenty}$

Stage Three: Interactions + Higher Order Terms

There were no interactions to add following the completion of Stage Two. Additionally, there were no higher order terms, so our multiple linear regression analysis ends here.

Multicollinearity

```
                      Age   minutes_played_overall        appearances_overall
goals_overall
                    1.231                  104.443                     67.952
8.372
           assists_overall    clean_sheets_overall           conceded_overall
goals_involved_per_90_overall
                    3.300                    5.742                      7.749
14.542
             min_per_match
                   34.036
[1] 27.48522
```

```
     clean_sheets_overall                 Age goals_involved_per_90_overall           min_per_match
                    1.149               1.102                         1.447                   1.476
[1] 1.2935
```

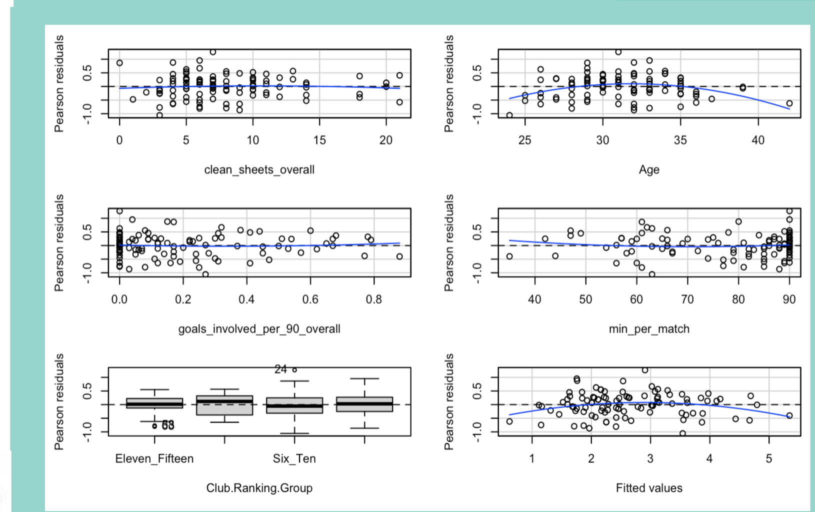
Variable Screening

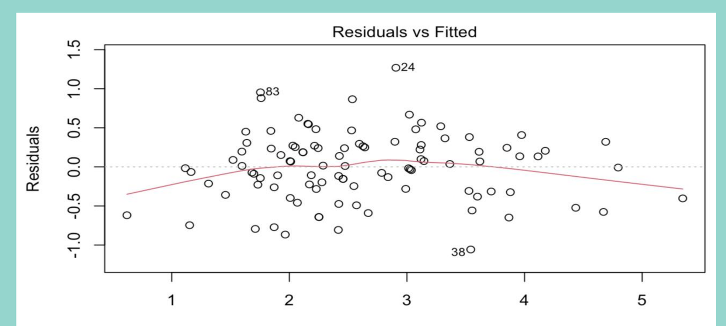
Parameter Estimates

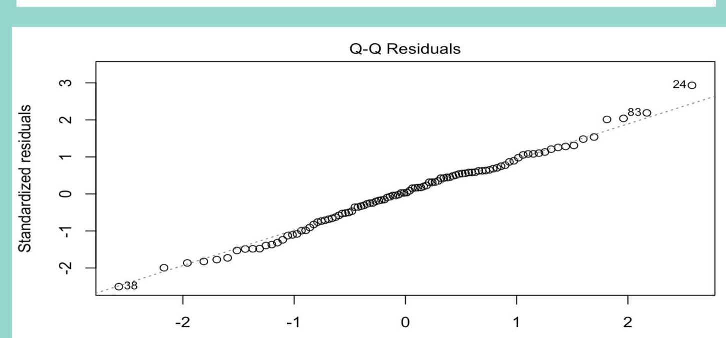
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	4.569	0.679		6.730	0.000	3.221	5.917
clean_sheets_overall	0.099	0.015	0.435	6.808	0.000	0.070	0.128
Age	-0.138	0.019	-0.453	-7.235	0.000	-0.176	-0.100
goals_involved_per_90_overall	1.622	0.314	0.370	5.158	0.000	0.997	2.246
min_per_match	0.016	0.005	0.223	3.080	0.003	0.006	0.026

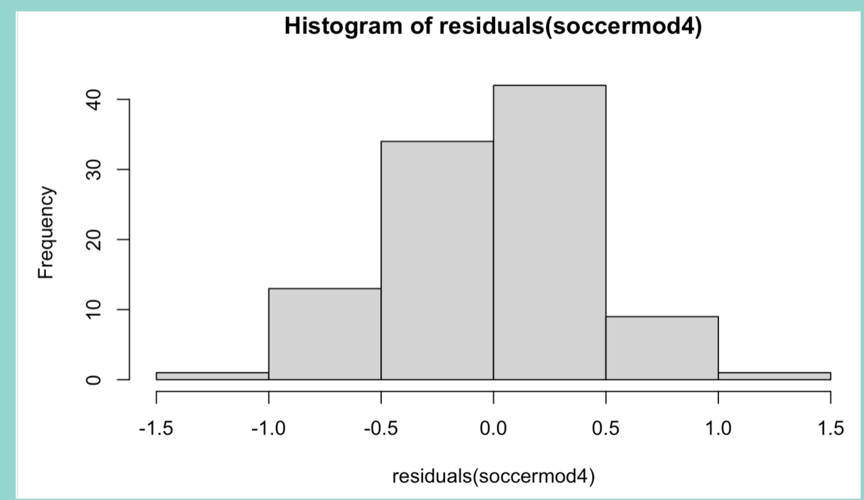
For variable screening on our quantitative variables, we used stepwise selection. After performing this, we were left with clean sheets overall, age, goals involved per 90 minutes overall, and minutes per match

Regression Assumptions

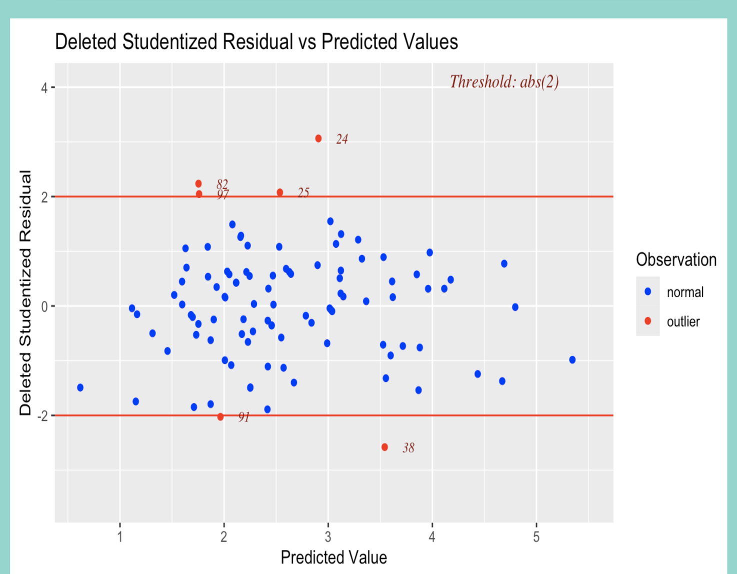








Influential Observations



Observations 24, 25, 38, 82, 91, and 97 were the outliers found on after viewing the plot on the left. After removing these observations, our F-Statistics increased from 58.36 (on 7 and 92 degrees of freedom) to 69.02 (on 7 and 86 degrees of freedom) and an increase from 0.8022 to 0.8366 for the adjusted R^2 value.

Cross Validation

```
Linear Regression

100 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 81, 81, 80, 79, 79
Resampling results:

  RMSE       Rsquared   MAE
  0.4593669  0.7981805  0.3611082

Tuning parameter 'intercept' was held constant at a value of TRUE
```

We used a 5-fold cross validation as our external model validation technique. It verified that we have a relatively high predictive power using our final model, as it yielded an R^2 value of 0.7981

Conclusion

FINAL MODEL (Prediction Equation):

$$\widehat{LogValuation} = 4.705709 + 0.044463 CleanSheetsOverall - 0.14508 Age + 1.457279 GoalsInvolvedPer90 + 0.019431 MinutesPerMatch + 0.935701 ClubRankingGroupOneFive + 0.547181 ClubRankingGroupSixTen - 0.243145 ClubRankingGroupSixteenTwenty$$

INTERPRETATION:
We determined that goals involved per 90 minutes has the largest magnitude of association when determining a soccer player's valuation. This is due to their significant increase in log valuation by 1.457279. This led us to conclude offensive players tend to have a higher valuation than defensive players. With every additional year, Age decreases the log valuation by 0.14508, leaning toward younger players having a higher valuation due to their potential and longevity. With every extra minute a player has per match, it adds 0.019431 to the log valuation, indicating consistent play time leads to a higher transfer value, demonstrating reliability and fitness.

Efficiency:
The model shows all predictors have VIF below 2 which indicates no significant multicollinearity. The parameters for the key variables are statistically significant ($p<0.05$). The final model was validated using a 5-fold cross-validation, with R^2 of 0.7982, low root mean square error of 0.4594 and mean absolute error (MAE) of 0.3611. Hence, the model is significant, indicating accurate and consistent predictions across subsets. Based on this information, we believe that this is a good model.

Usage:
For example, we attempted to predict the log of player valuation for Anthony Martial. When we substituted values into our model for him, we got a log(valuation) of 3.84536. His actual log(valuation) was 7.77815, so our residual is 3.93279.

Limitations and Future Analysis

Limitations
- Limited to 5 or fewer levels which led for us to group data accordingly
- Performed analysis for only one Premier League season
- Removed some variables from analysis due to multicollinearity

Future Analysis
- Add more variables
  - Passes made per 90
  - Shots taken per 90
  - Injury record
- Manipulate the data to change the weightage of variables according to player position (Ex: Clean sheets would have a higher impact on the valuation of a defensive player than an attacking player) (DiBalsi, 2022)
- Could look at data across multiple seasons to better determine the valuation of players

Works Cited

- Premier League 18/19. (2018). Transfermarkt.us. https://www.transfermarkt.us/premier-league/startseite/wettbewerb/GB1/saison_id/2018
- Players CSV / Football Stats Database to CSV | FootyStats. (2019). Footystats.org. https://footystats.org/download-stats-csv
- Metelski, A. (2021, April 30). *Factors affecting the value of football players in the transfer market*. ResearchGate. https://www.researchgate.net/publication/351335017_Factors_affecting_the_value_of_football_players_in_the_transfer_market
- Zhou, W. (2019). *Which Factors Decide the Market Value of Soccer Players: An Empirical Evidence from European League*. Clausius Press. https://www.clausiuspress.com/conferences/LNEMSS/EMSD%202019/19EMSD057.pdf
- Poli, R., Ravenel, L., & Besson, R. (2020, March). *Scientific evaluation of the transfer value of football players*. Monthly Report 53. https://football-observatory.com/IMG/sites/mr/mr53/en/
- DiBlasi, L. (2022, April). *Footballer Valuations: Valuing World-Class Football Players Against Transfer Fees* [Review of *Footballer Valuations: Valuing World-Class Football Players Against Transfer Fees*]. Bryant University. https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1042&context=honors_economics