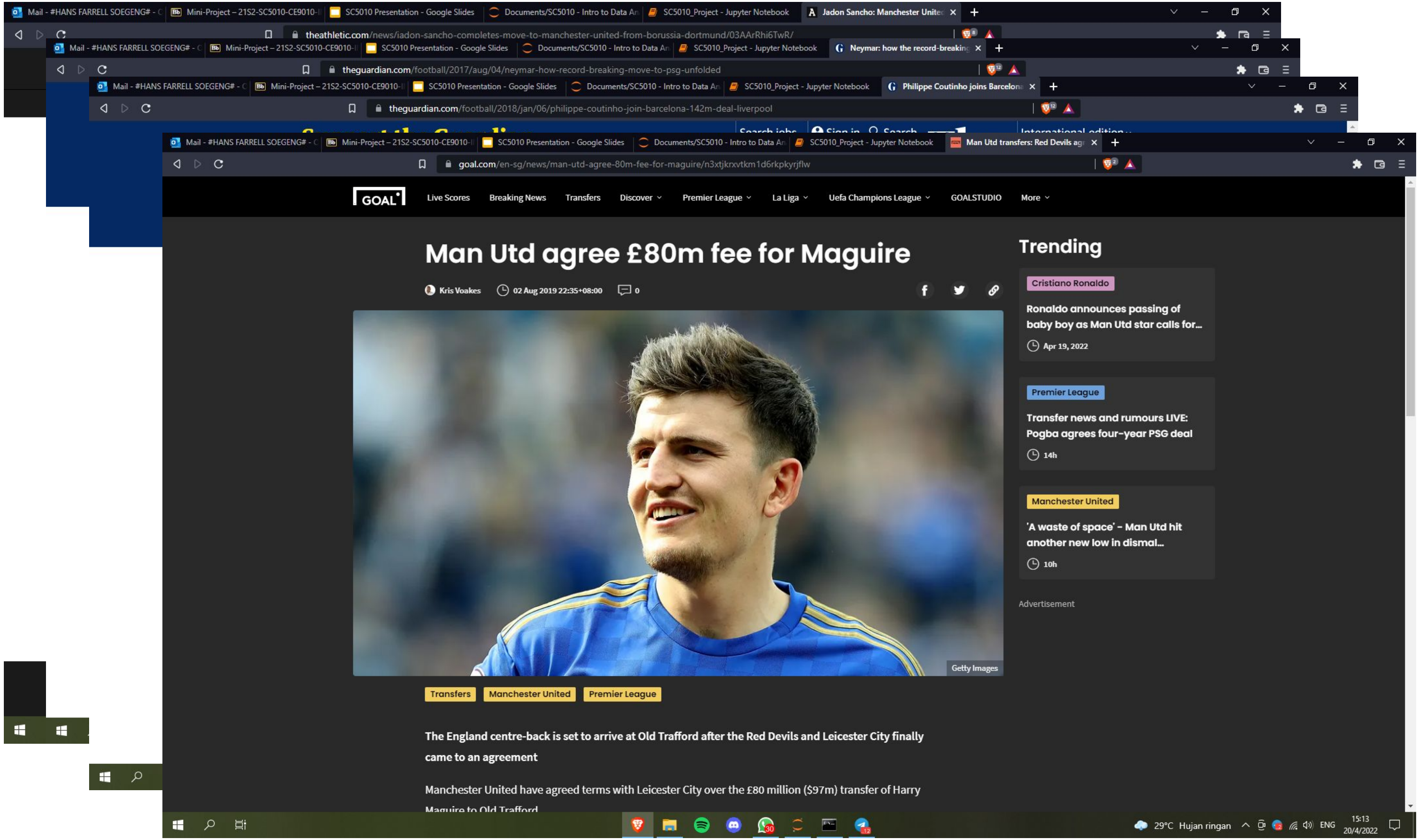
# Predicting the Football Market

**Team 6**
Hans Farrell Soegeng
Rivaldo Billy Sebastian
Nicholas Chin Wei Lun

# Problem Motivation


theathletic.com/news/iadon-sancho-completes-move-to-manchester-united-from-borussia-dortmund/03AArRhi6TwR/
Jadon Sancho: Manchester Unite
theguardian.com/football/2017/aug/04/neymar-how-record-breaking-move-to-psg-unfolded
Neymar: how the record-breaking
theguardian.com/football/2018/jan/06/philippe-coutinho-join-barcelona-142m-deal-liverpool
Philippe Coutinho joins Barcelona
goal.com/en-sg/news/man-utd-agree-80m-fee-for-maguire/n3xtjkrxvtkm1d6rkpkyrjflw
Man Utd transfers: Red Devils ag
GOAL
Live Scores
Breaking News
Transfers
Discover
Premier League
La Liga
Uefa Champions League
GOALSTUDIO
More
Man Utd agree £80m fee for Maguire
Kris Voakes
02 Aug 2019 22:35+08:00
Getty Images
Transfers
Manchester United
Premier League
The England centre-back is set to arrive at Old Trafford after the Red Devils and Leicester City finally came to an agreement
Manchester United have agreed terms with Leicester City over the £80 million ($97m) transfer of Harry Maguire to Old Trafford
Trending
Cristiano Ronaldo
Ronaldo announces passing of baby boy as Man Utd star calls for...
Apr 19, 2022
Premier League
Transfer news and rumours LIVE: Pogba agrees four-year PSG deal
14h
Manchester United
'A waste of space' - Man Utd hit another new low in dismal...
10h
Advertisement
29°C Hujan ringan
ENG
15:13
20/4/2022

# The Football Transfer Market

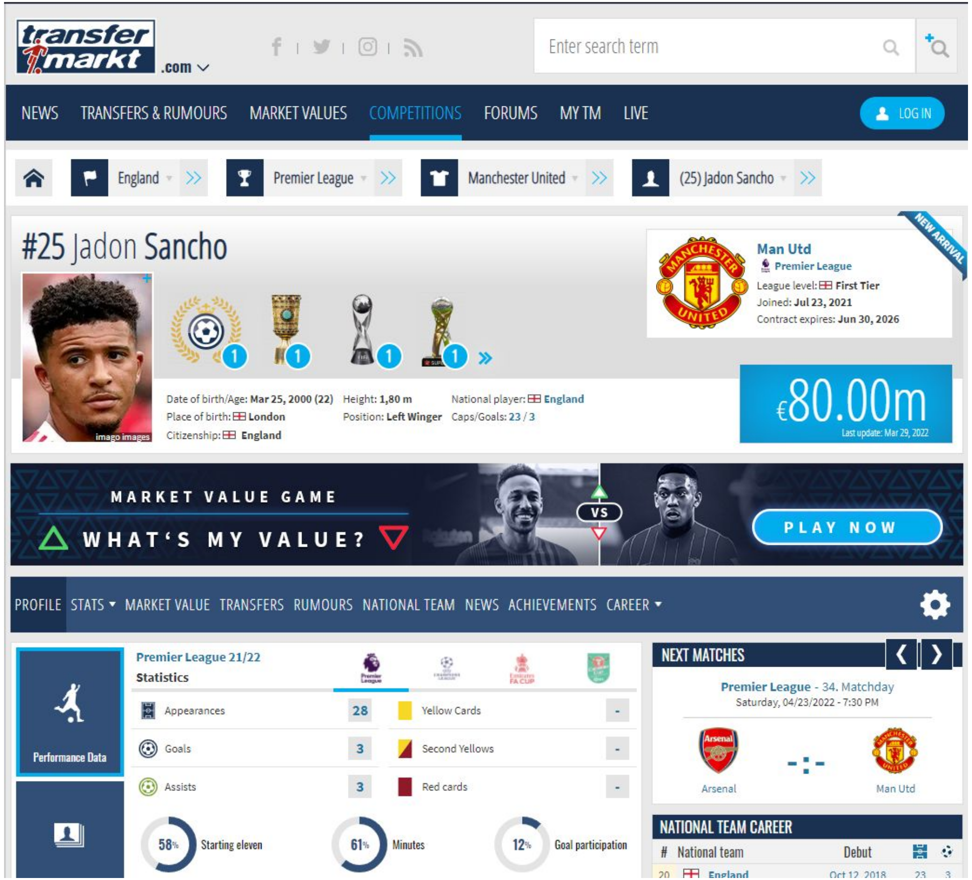
- Clubs buy or sell players every transfer window to strengthen their squad or get money from sales.
- Transfers happen for a fee agreed between the clubs
- This fee is determined by how much the transferred player is worth for the buying club.
- But how do clubs value a player?

# Market Value Estimate

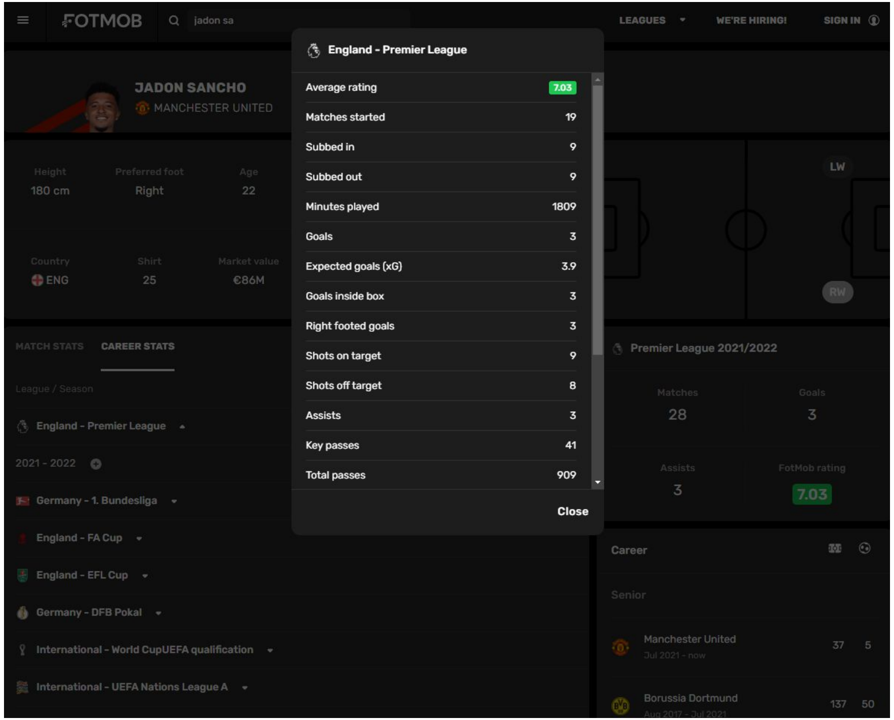
Several websites have tried to estimate players based on their market value to clubs.

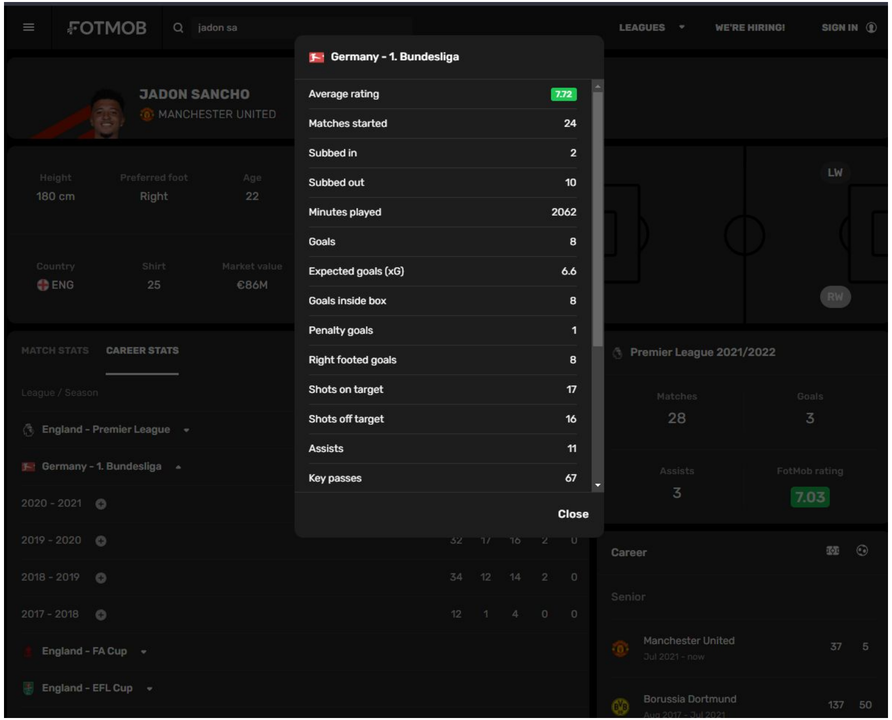
They use the wisdom the community to evaluate these players

But, is there a way we can objectively find a player’s value?



# Statistics!




Statistics of Jadon Sancho for 2021/22 Premier League and 2020/21 Bundesliga seasons

# Objective

Statistics provide us a concrete and objective way to rate a football player.

The idea is then to create a machine learning model to predict a player’s market value based on their statistics of their past seasons

# Data Preparation

# Data Collection

- Extract the statistics of every attacking player in the top 5 leagues for the past 2 seasons (2020/21 & 2021/22) obtained from FotMob.com and their respective estimated market values into an excel file through web scraping
- Used BeautifulSoup to parse the website’s html into a readable format

```
In [12]: df = pd.read_excel('playerdata.xlsx')
```

```
In [16]: df
```

Out[16]:

| | name | height | age | value | preferred_foot | average_rating_2021 | matches_started_2021 | subbed_in_2021 | subbed_out_2021 | minutes_played_2021 | ... | shot |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Roberto Soriano | 181 | 31 | 8.50 | Both | 6.97 | 26.0 | 1.0 | 10.0 | 2121 | ... | |
| 1 | Musa Barrow | 183 | 23 | 22.00 | Right | 6.94 | 19.0 | 6.0 | 16.0 | 1592 | ... | |
| 2 | Nicola Sansone | 175 | 30 | 3.20 | Right | 6.42 | 6.0 | 16.0 | 5.0 | 768 | ... | |
| 3 | Ricardo Orsolini | 183 | 25 | 11.00 | Left | 6.99 | 17.0 | 5.0 | 13.0 | 1367 | ... | |
| 4 | Emanuel Vignato | 175 | 21 | 4.20 | Right | 6.43 | 3.0 | 19.0 | 3.0 | 557 | ... | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 565 | Amadou Traore | 175 | 20 | 0.65 | Right | 5.68 | NaN | 1.0 | NaN | 19 | ... | |
| 566 | M'baye Niang | 184 | 27 | 5.00 | Right | 5.99 | 1.0 | 15.0 | 1.0 | 342 | ... | |
| 567 | Hwang Ui-Jo | 184 | 29 | 5.50 | Both | 6.67 | 23.0 | 1.0 | 15.0 | 1946 | ... | |
| 568 | Sekou Mara | 183 | 19 | 2.20 | Right | 6.18 | 5.0 | 13.0 | 5.0 | 551 | ... | |
| 569 | Jimmy Briand | 180 | 36 | 1.20 | Right | 6.67 | 1.0 | 11.0 | 1.0 | 180 | ... | |

570 rows × 55 columns

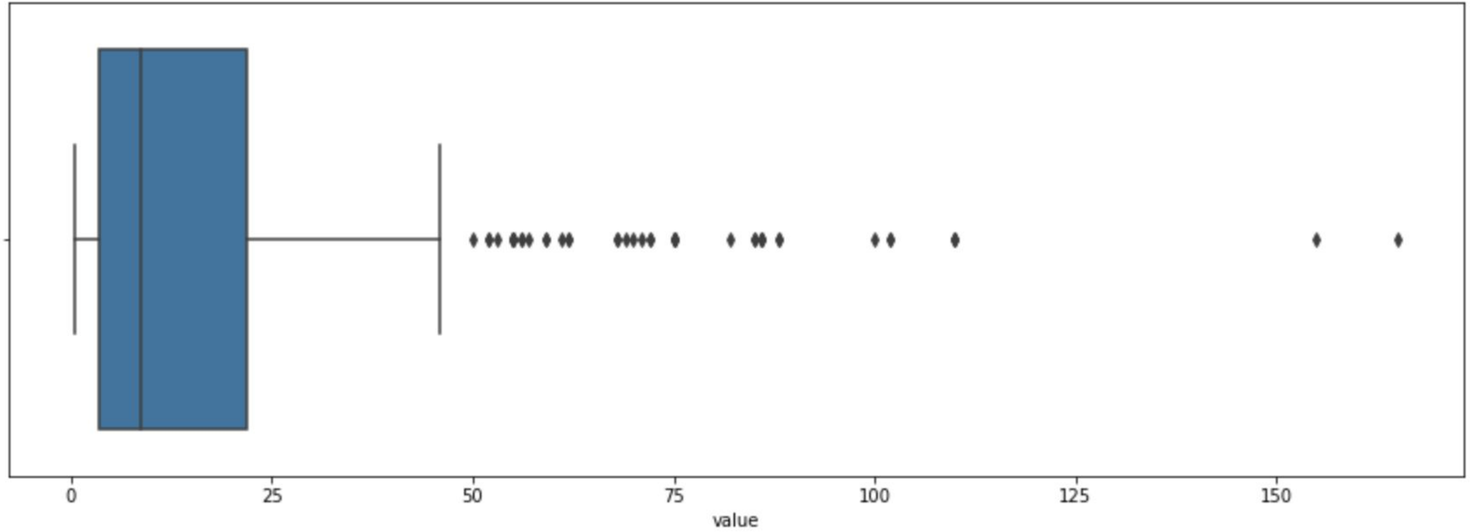
# Objective : Predict value (in millions of euros) for every 570 players.

# Data Cleaning

---

- Filled null values with 0 as Fotmob does not list statistics of players with value 0
- Remove players with average rating of 0

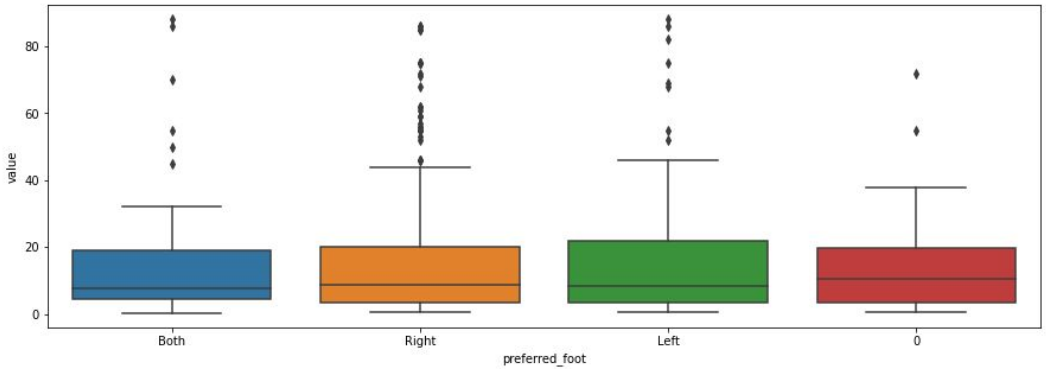
# Exploratory Data Analysis




- Consider players with value greater than 100 million euros as outliers
- Removed these players from the dataset

In [53]:
```
plt.figure(figsize=(15, 5))
sb.boxplot(x="preferred_foot", y="value", data=df)
```

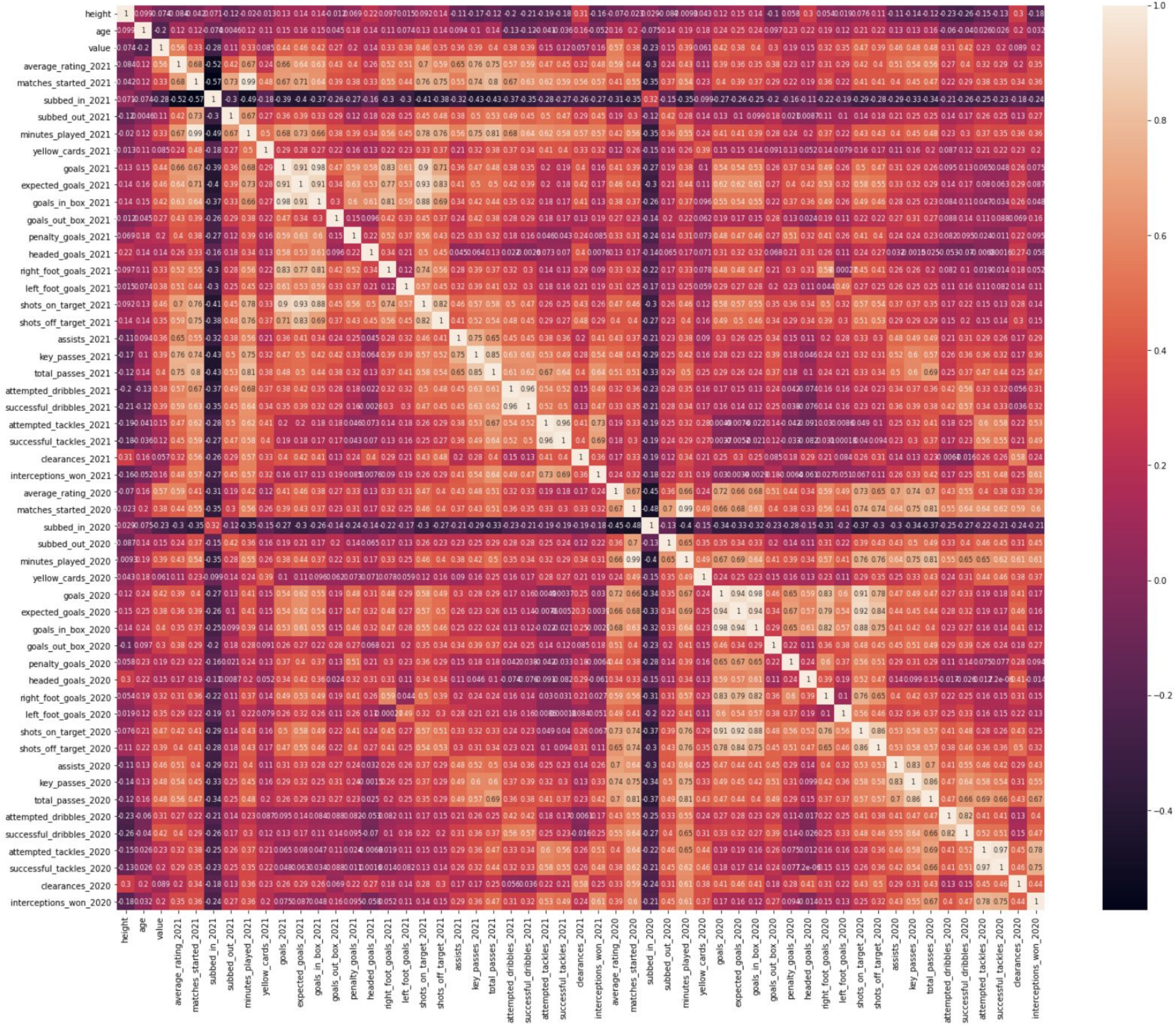
Out[53]: <AxesSubplot:xlabel='preferred_foot', ylabel='value'>




There does not appear to be a strong relationship between preferred_foot and value so we choose not to include this variable.

In [54]:
```
df = df.drop(columns = ['preferred_foot'])
```

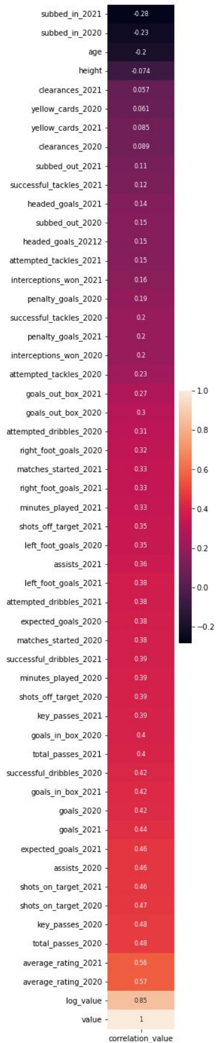
# Compare the market value across different preferred foot categories




- Some variables are highly correlated with each other, such as attempted_tackles_2020 and successful_tackles_2020 with a correlation of 0.97, goals_2021 and goals_in_box_2021 with a correlation of 0.98

- The variables subbed_in_2020 and subbed_in_2021 are least correlated with other variables

# Compare the correlation coefficient of market value with all the numerical predictor variables

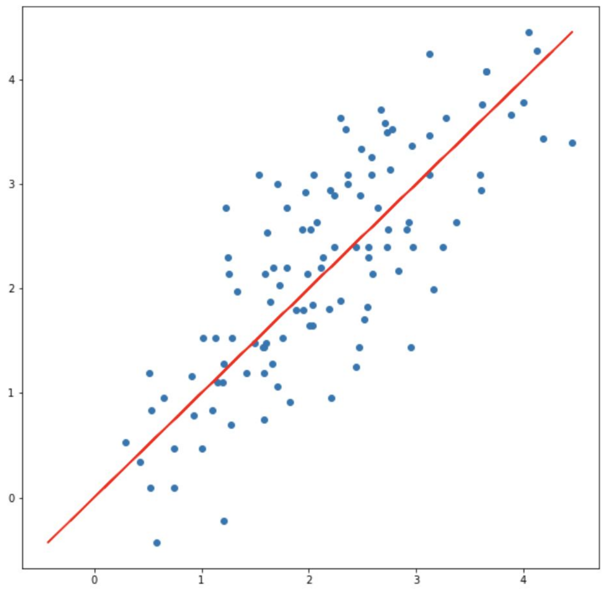


# Machine Learning

# Machine Learning Models

- Used regression models to predict value
- Tried 5 different models
  - Linear Regression
  - Linear Regression (with higher order terms)
  - Ridge Regression
  - Lasso Regression
  - Random Forest Regression
- Performed log transformation on value to ensure only positive values are predicted
- Used 80% training and 20% test sets
- Comparing the models in terms of R squared and mean squared error

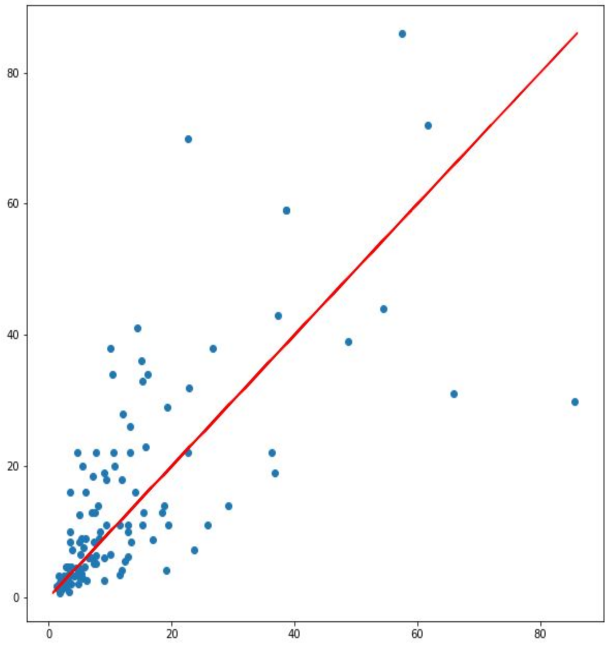
# Linear Regression




Predictions on test set

R^2 value = 0.612

MSE value = 146.45

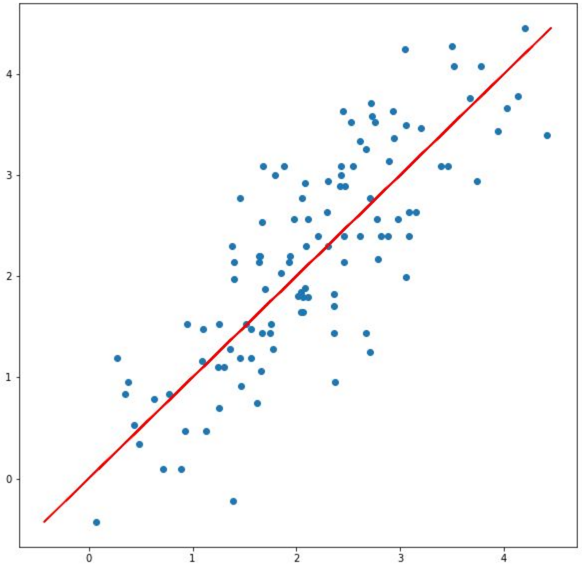



Predicted vs actual market value on test set

# Linear Regression (with higher order terms)

- Limitation of this linear regression model is that it assumes the response variable has a linear relationship against every predictor variable
- Try to improve this model by adding higher order terms for predictor variables having correlation coefficient with absolute value higher than 0.5, which are average_rating_2020 and average_rating_2021
- Add the second order term of age, as even though its correlation coefficient is less than 0.5, age may have a complex relationship with market value

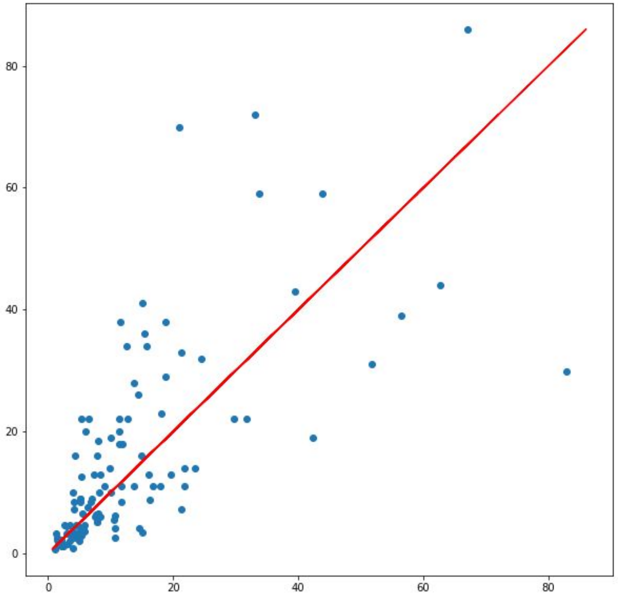
# Linear Regression (with higher order terms)




Predictions on test set

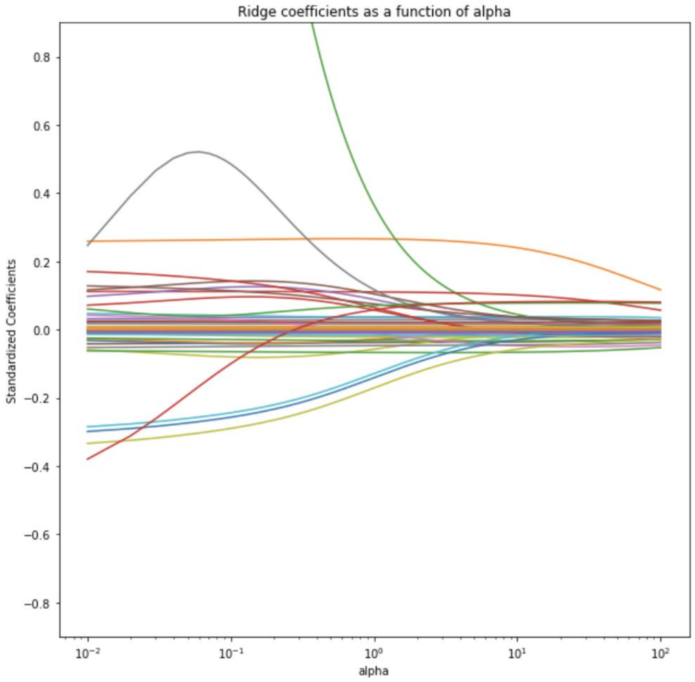
R^2 value = 0.643

MSE value = 144.95

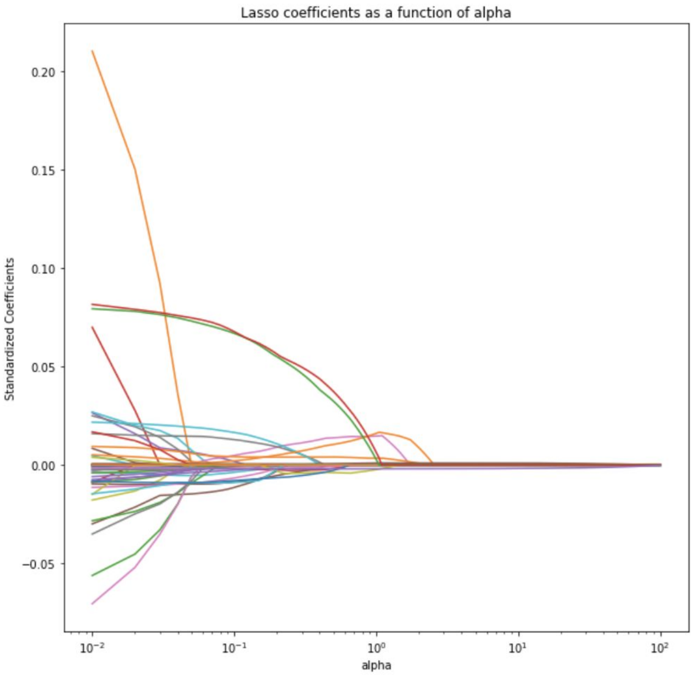



Predicted vs actual market value on test set

# Ridge and Lasso Regression




Ridge Regression




Lasso Regression

# Model Comparison

Model Comparison

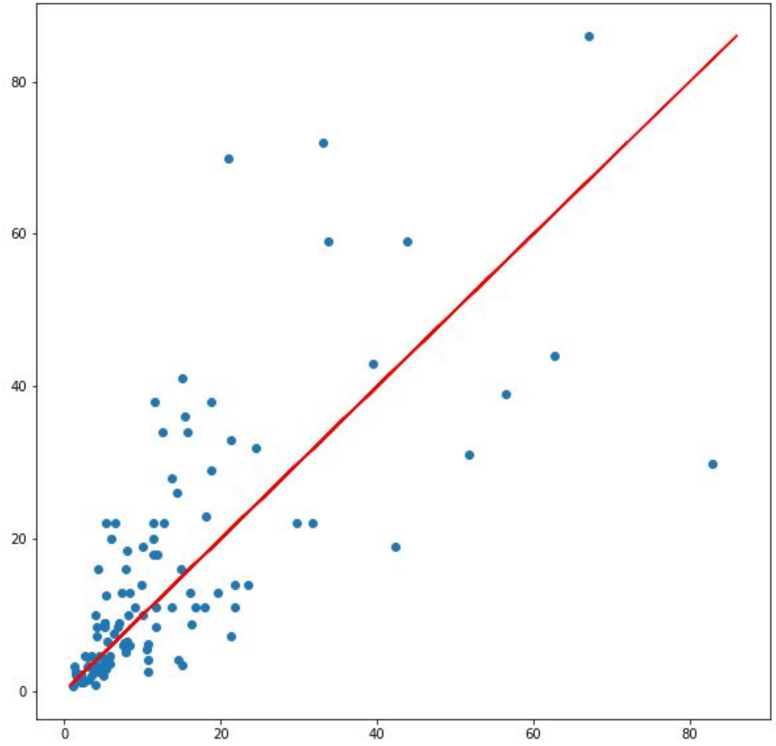
| Model | R2 Score | MSE |
|---|---|---|
| Linear Regression | 0.6124299824151251 | 146.45024016416988 |
| Linear Regression (higher order terms) | 0.6431267219862531 | 144.95331830442498 |
| Ridge Regression | 0.6448971342819805 | 134.5894783980847 |
| Lasso Regression | 0.640337661964369 | 123.10347326169641 |
| Random Forest Regression | 0.6063040966499413 | 147.7985546636136 |

- Based on R2 score, Ridge Regression appears to be the best performing model
- Based on MSE, Lasso Regression appears to be the best performing model
- Since the MSE of Lasso Regression is significantly less as compared to the other models, we select it as the best performing model.

# Conclusion

# Insert lasso graph plot here

From the regression models, we notice that our predicted market value gets more inaccurate as the value increases



# Market Value Factors

---

- On-pitch performance
- Age (development potential and future prospects)
- Reputation/prestige
- Marketing value
- Number & reputation of interested clubs
- Experience
- Injury susceptibility
- General demand and ‘trends’ of the market

# Future Ideas

If we could somehow quantify the off-pitch element of a football player in terms of the commercial value he brings to a club, it could be an important variable in predicting market value

E.g, total followers on his social medias, engagements on media networks, sponsorship deals values, etc