PREDICTING SOCCER PLAYERS' TRANSFER MARKET VALUES

Mike Choi

OVERVIEW

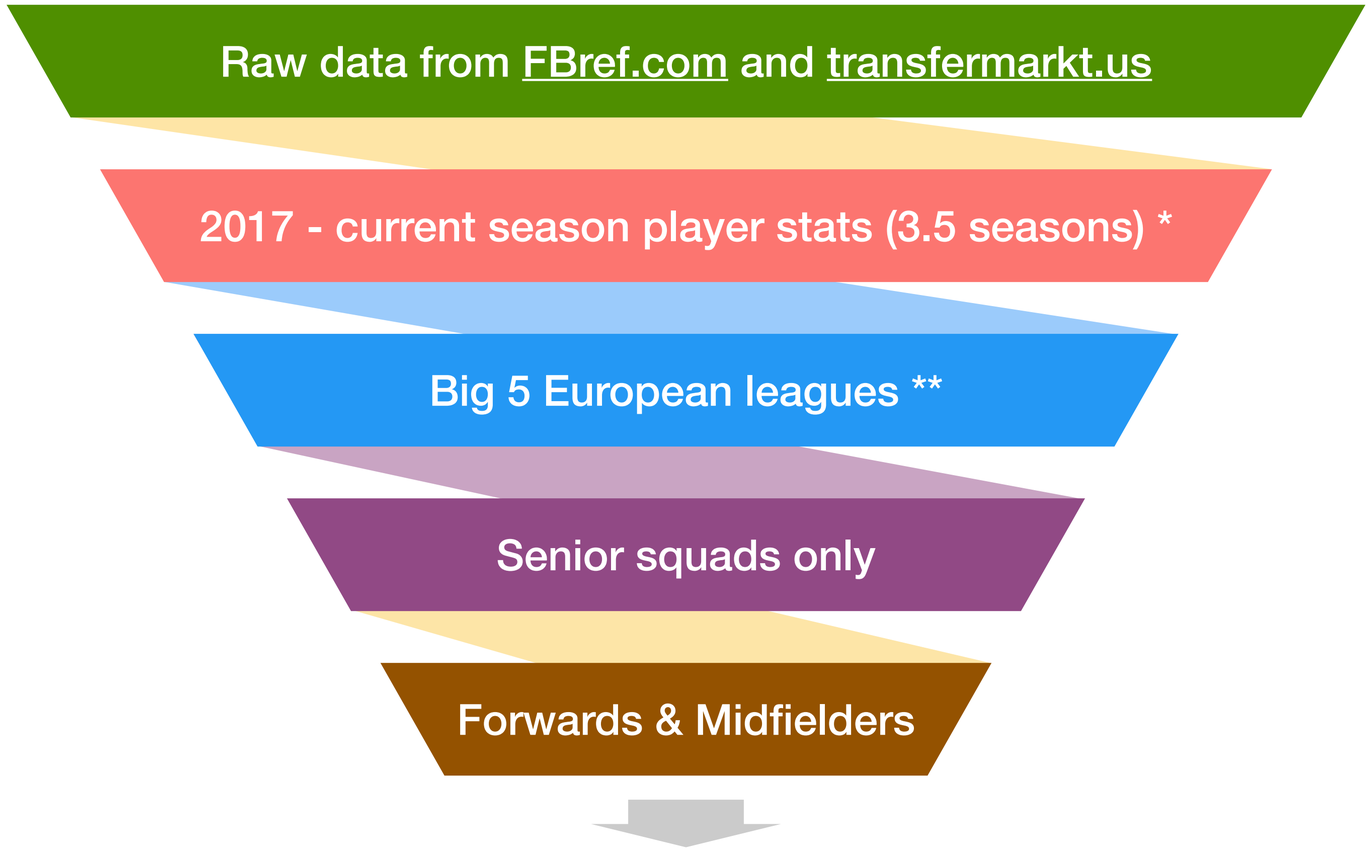
Goal

- Build a linear regression model that can predict a soccer player's transfer market value based on his stats.

Questions to Answer

- Which statistics affect a player's transfer value the most?
- Can the model be used to help European soccer clubs with their scouting and transfer negotiation process (e.g., finding undervalued players)?

DATA AND SCOPE



~1000 Players (without missing data)

* Assume players need multiple seasons of consistent performance to prove their worth in the transfer market.

** The Big 5 European leagues - Premier League, La Liga, Serie A, Bundesliga, Ligue 1 - are the closest in terms of competitiveness and transfer market activity

FEATURE ENGINEERING – STANDARDIZING NUMERICAL STATS

- Lower stats due to fewer games played:
  - ➤ Injuries
  - ➤ Youth players promoted to senior squad
  - ➤ Transfers from non-Big 5 leagues
- **Standardize:**
  - ➤ **Divide by 90 min games played**

Pre-standardization

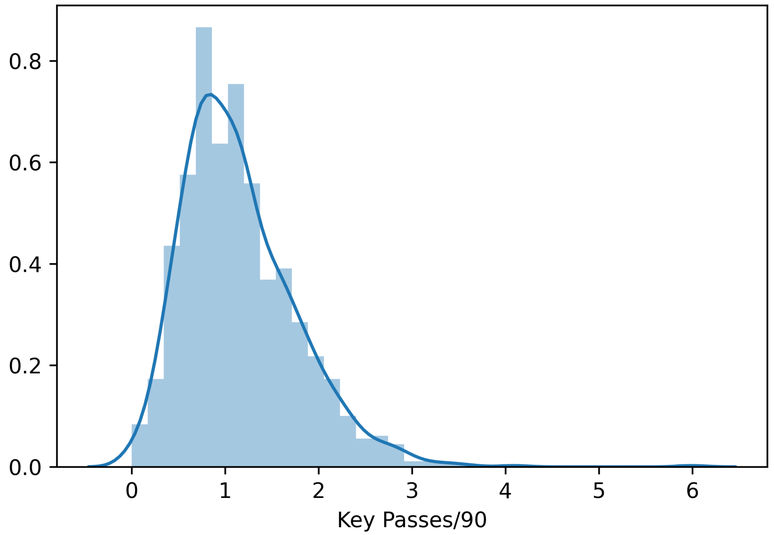
Player Name	Goals	Assists	Market Value ($MM)
Hakim Ziyech	1	3	50
Youri Tielemans	15	14	55

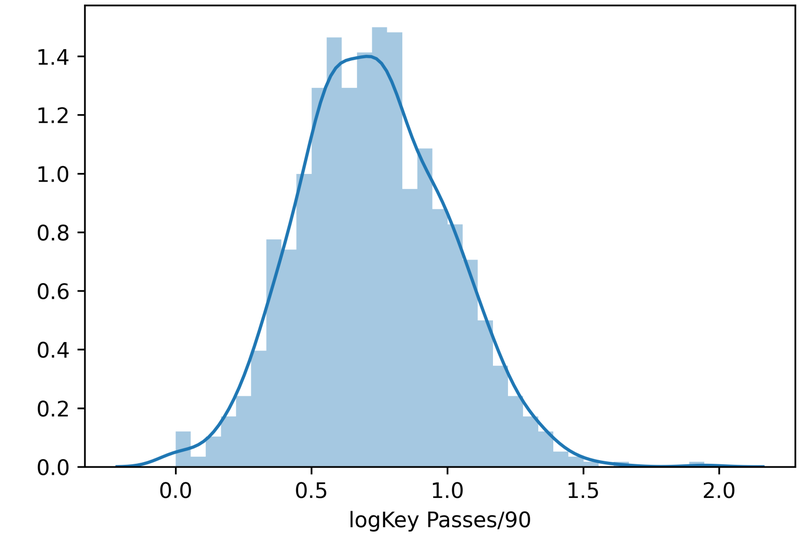
Post-standardization

Player Name	Goals/90s	Assists/90s	Market Value ($MM)
Hakim Ziyech	0.16	0.49	50
Youri Tielemans	0.15	0.14	55

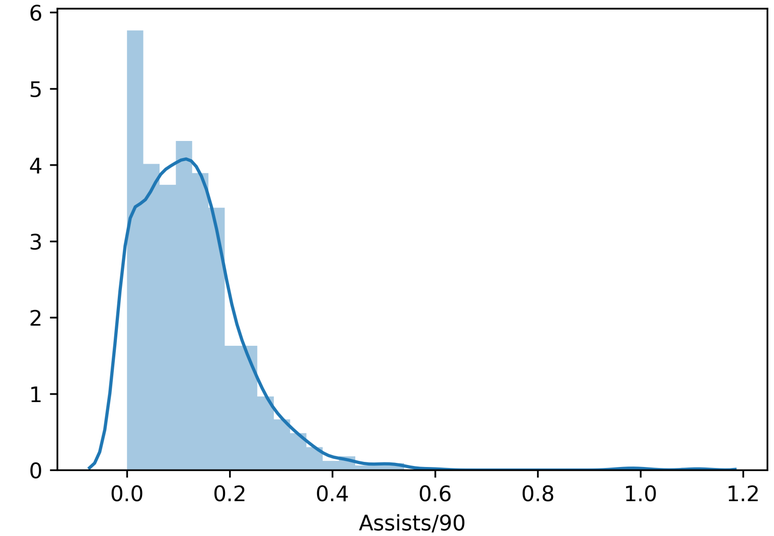
FEATURE ENGINEERING – LOG TRANSFORMATION OF RIGHT-SKEWED FEATURES

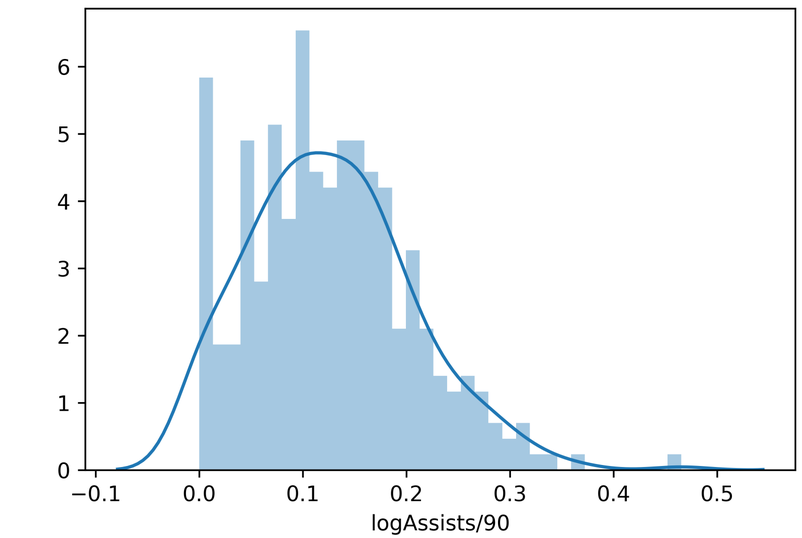
Key Passes/90s





Assists/90s

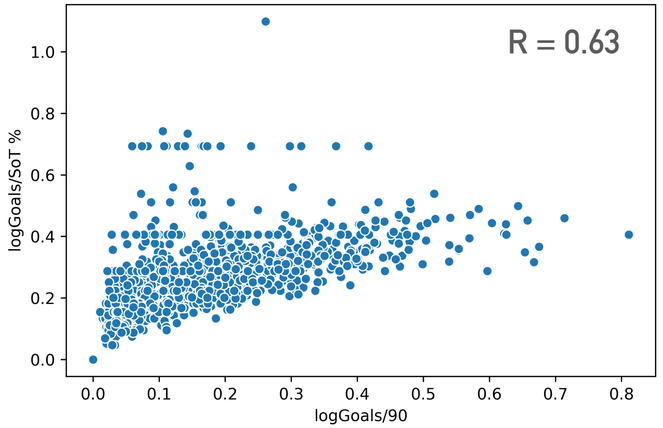


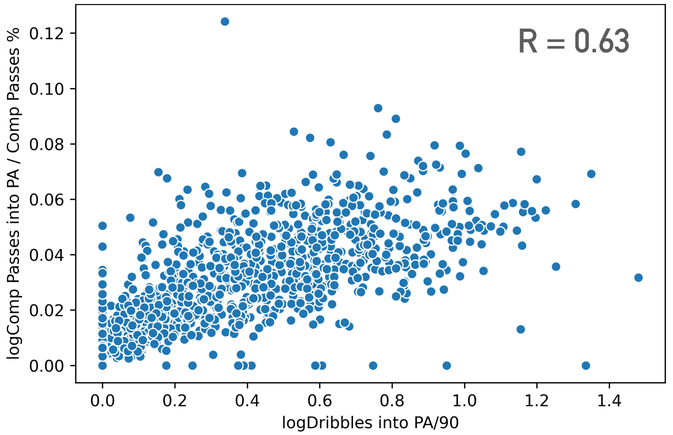


FEATURE SELECTION THROUGH LASSO COEFFICIENTS

	Lasso Coeff.
Age	-6.18
% of Matches Started	5.12
Shots on Target %	-0.91
Pass Comp Rate %	6.34
Goals/90s (log)	8.12
Assists/90s (log)	1.68
Goals/SoT %	0.00
Comp Passes into PA / Comp Passes % (log)	0.84
Key Passes/90s (log)	2.02
Players Dribbled Past/90s (log)	0.00
Dribbles into PA/90s (log)	2.67
Cards Received/90s (log)	0.19
La Liga	0.00
Ligue 1	-0.67
Serie A	0.02
Premier League	2.43

High collinearity





Low correlation with target

CHOOSING THE BEST MODEL THROUGH K-FOLD CROSS-VALIDATION

	Simple Linear	Ridge	Lasso	Polynomial (d=2)
Adjusted R^2 *	0.65	**0.66**	0.62	-3.46
RMSE	6.34	**6.33**	6.55	8.11

* Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model, such that its value increases only when the new predictor improves the model fit more than expected by chance alone.

FINAL MODEL TESTING – UNSCALED FEATURE COEFFICIENTS

$$Adjusted\ R^2 = 0.63$$

$$RMSE = 6.52$$

$$MAE = 5.03$$

	Unscaled Coeff
Age	-1.7
% of Matches Started	35.4
Pass Completion Rate %	98.9
Premier League	5.7
Goals/90s (log)	56.5
Assists/90s (log)	27.2
Key Passes/90s (log) *	4.9
Dribbles into Pen. Area/90s (log)	10.9

* A key pass is a pass that directly leads to a shot on goal.

FINAL MODEL TESTING – HOW DOES EACH STAT AFFECT PLAYER MARKET VALUE?

$$Adjusted\ R^2 = 0.63$$

$$RMSE = 6.52$$

$$MAE = 5.03$$

	Transfer Market Value Effect
Age	- $1.7mm/year
% of Matches Started	+ $350k/ppt
Pass Completion Rate %	+ $1mm/ppt
Premier League	+ $5.7mm
Goals/90s (log)	+ $570k/1% inc
Assists/90s (log)	+ $270k/1% inc
Key Passes/90s (log) *	+ $49k/1% inc
Dribbles into Pen. Area/90s (log)	+ $110k/1% inc

* A key pass is a pass that directly leads to a shot on goal.

FINAL MODEL TESTING – SCALED FEATURE COEFFICIENTS

$$Adjusted\ R^2 = 0.63$$

$$RMSE = 6.52$$

$$MAE = 5.03$$

	Scaled Coeff
Age	-6.1
% of Matches Started	5.7
Pass Completion Rate %	6.8
Premier League	2.7
Goals/90s (log)	7.7
Assists/90s (log)	2.1
Key Passes/90s (log) *	1.3
Dribbles into Pen. Area/90s (log)	3.3

* A key pass is a pass that directly leads to a shot on goal.

KEY TAKEAWAYS

	Top Statistics
1	Goals/90s
2	Pass Completion Rate %
3	% of Matches Started
4	Dribbles into Pen. Area/90s
5	Premier League

- Focus on player stats that more directly influence a team's results

KEY TAKEAWAYS

	Top Statistics
1	Goals/90s
2	Pass Completion Rate %
3	% of Matches Started
4	Dribbles into Pen. Area/90s
5	Premier League

- Focus on player stats that more directly influence a team's results
- Avoid players with high pass completion rate but low assists, key passes, and forward passes

KEY TAKEAWAYS

	Top Statistics
1	Goals/90s
2	Pass Completion Rate %
3	% of Matches Started
4	Dribbles into Pen. Area/90s
5	Premier League

- Focus on player stats that more directly influence a team's results
- Avoid players with high pass completion rate but low assists, key passes, and forward passes
- Undervalued players more likely to be found in non-Premier league teams, and in substitutes with high goals/90s and assists/90s

LIMITATIONS OF THE DATA/MODEL

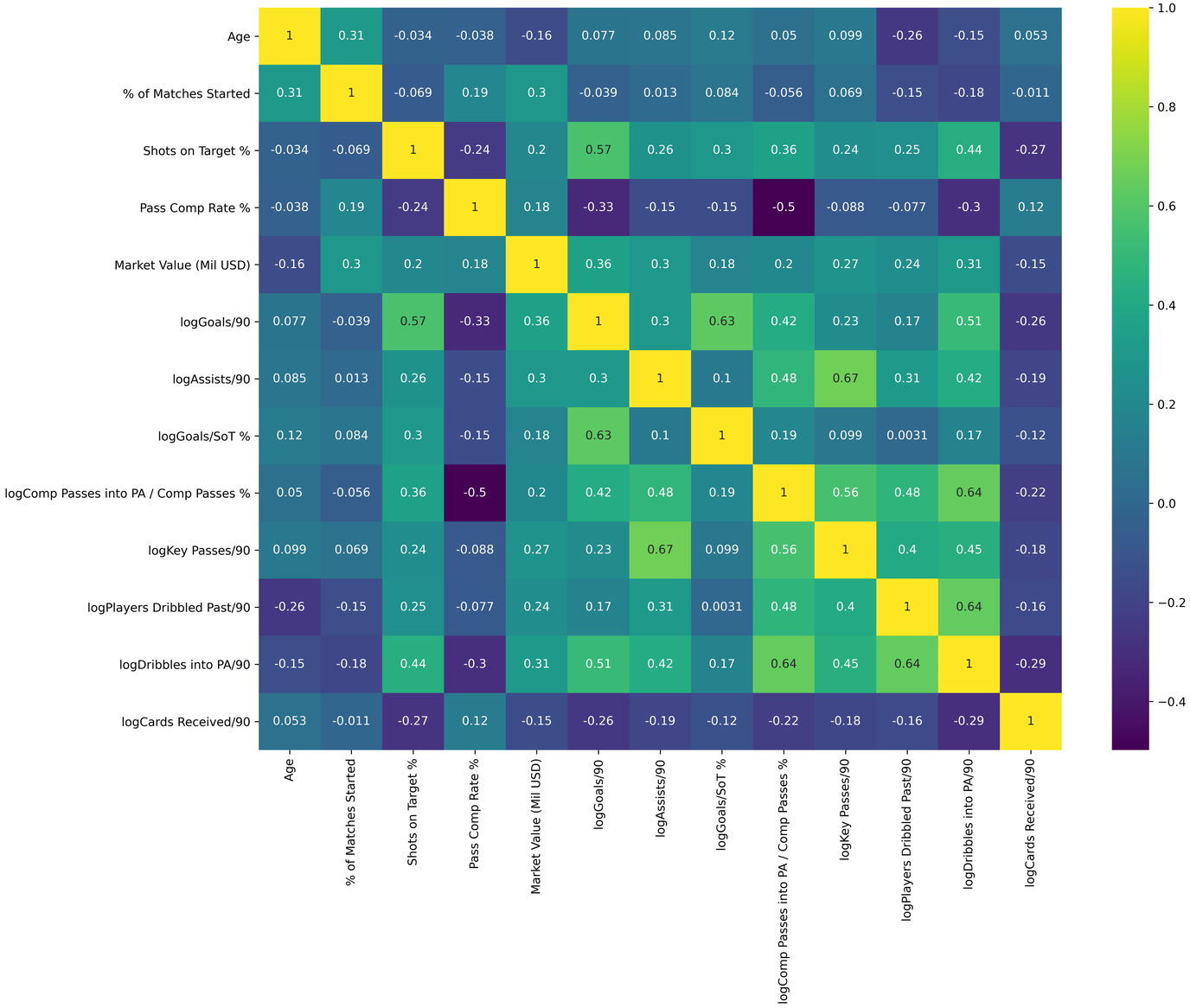
- Players' transfer market values depend on more than just league statistics:
  - ➤ Popularity/marketability - e.g., jersey sales
  - ➤ Performance in national team tournaments - e.g., World Cup
  - ➤ Other external factors - e.g., COVID-19

NEXT STEPS

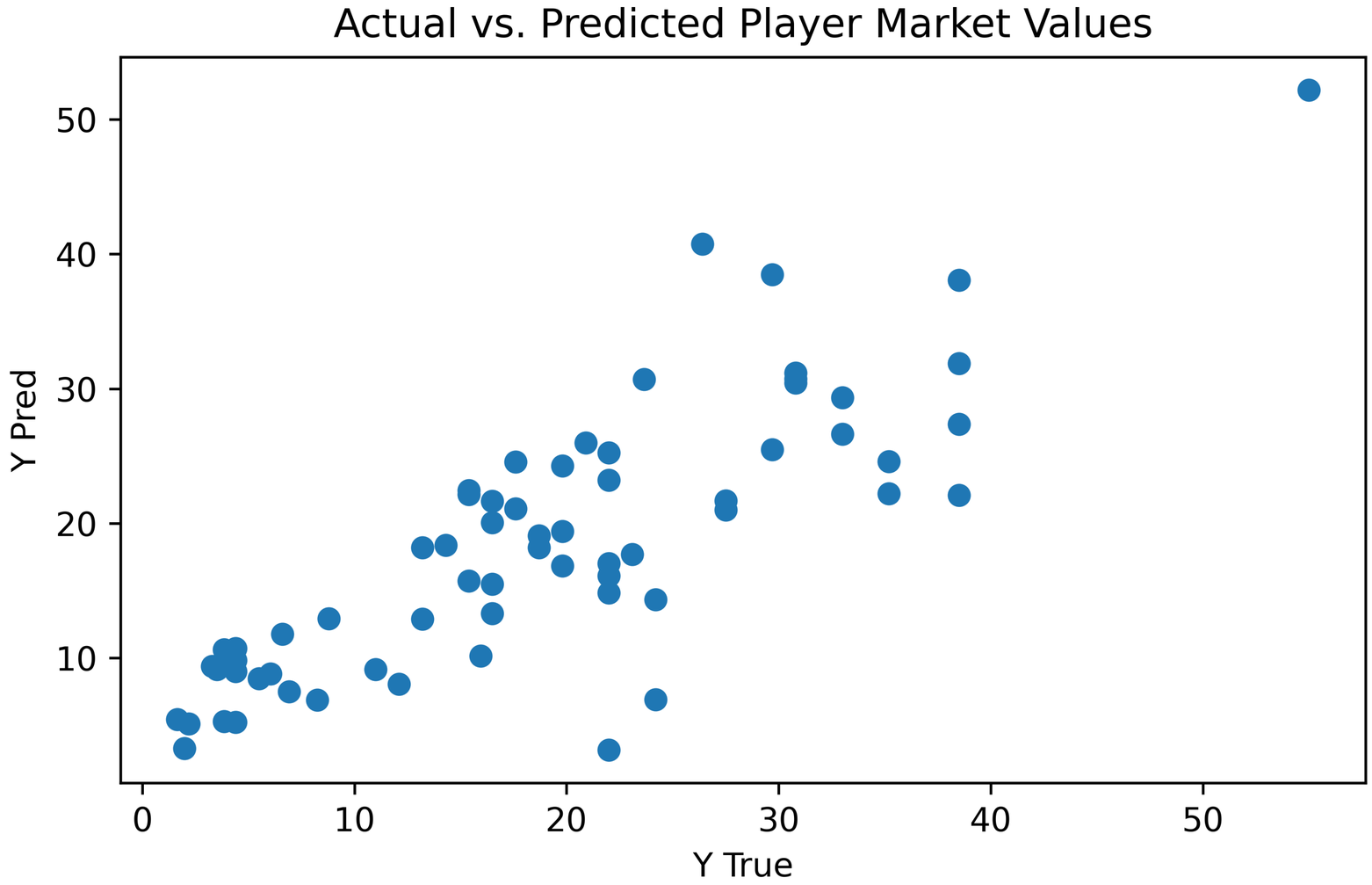
- Gather additional data
    - ➤ Player stats for national team tournaments
    - ➤ Player jersey sales in dollars
- Perform similar regression analysis for defenders and goalkeepers
    - ➤ Use position-specific features (e.g., successful tackles/90s for defenders)

THANK YOU!
QUESTIONS?

APPENDIX – CORRELATION HEATMAP



APPENDIX – ACTUAL VS. PREDICTED PLAYER MARKET VALUES



APPENDIX – RESIDUALS DISTRIBUTION