

The Beautiful Game



Can you predict the spread in the English Premier League?

Michael Weber

Why predict soccer aka football spread?

- Most popular sport in the world
- EPL six teams worth > \$1Billion
- Betting industry valued up to \$1Trillion annually
- Known for unpredictability



“In maths every day you know that one plus one is two. In football, one player plus one player doesn’t always add up to two players.” - Arsene Wenger

Can we predict the winning margin aka spread?

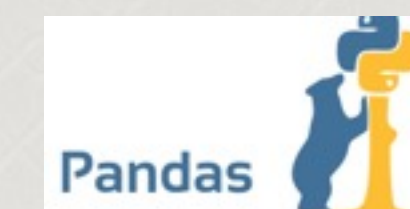
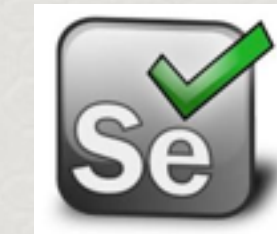
Home team score - Away team score

Data Collection and Tools

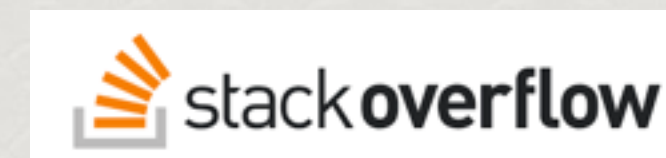
DATA

- Match results csv - [DataHub.io](https://datahub.io)
- Additional stats - premierleague.com
- Top 100 player rankings - easports.com
- Relegation/promotion - planetfootball.com
- Club wages per season - transfermarkt.com

Tools



seaborn



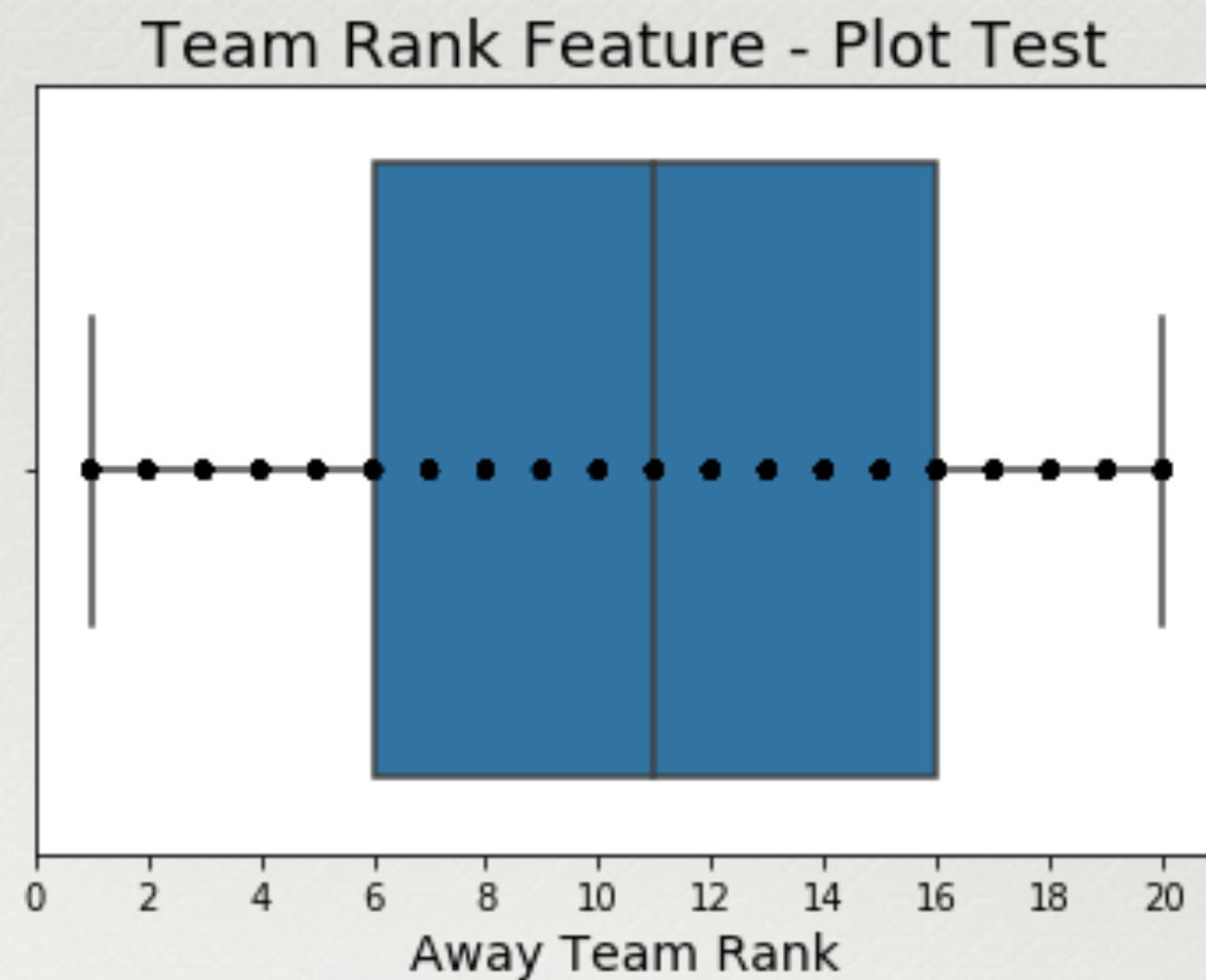
Methods

- Kitchen Sink - Overly fitting that it was a Failure!
- Occam's Razor approach - find baseline then build forward
- Forward Feature Selection - find most predictive feature then add to it
- Plot and test all features - eliminate bad assumptions

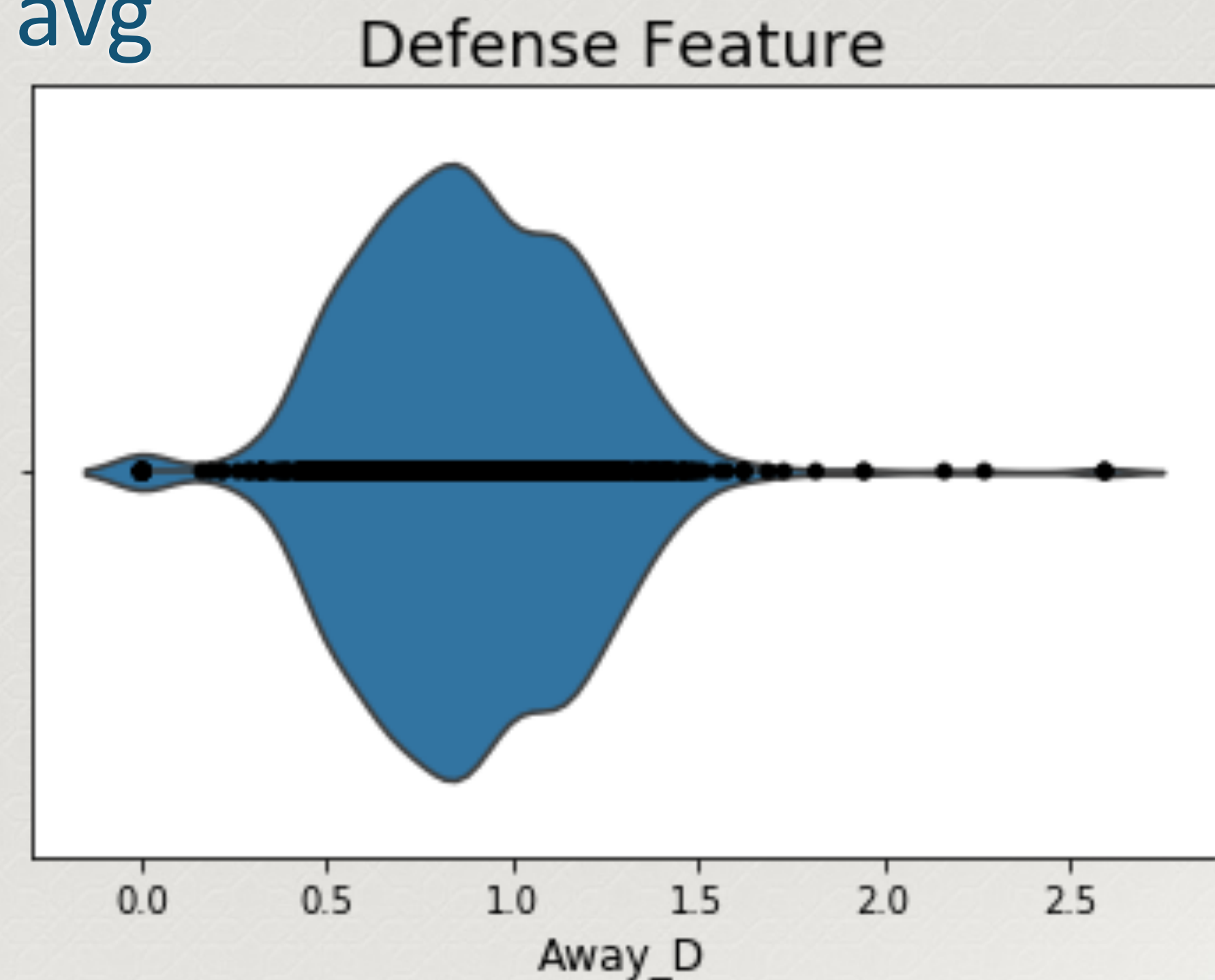
Feature Engineering











Finding single most predictive feature.

Team Rank \longrightarrow Team Defense / league avg



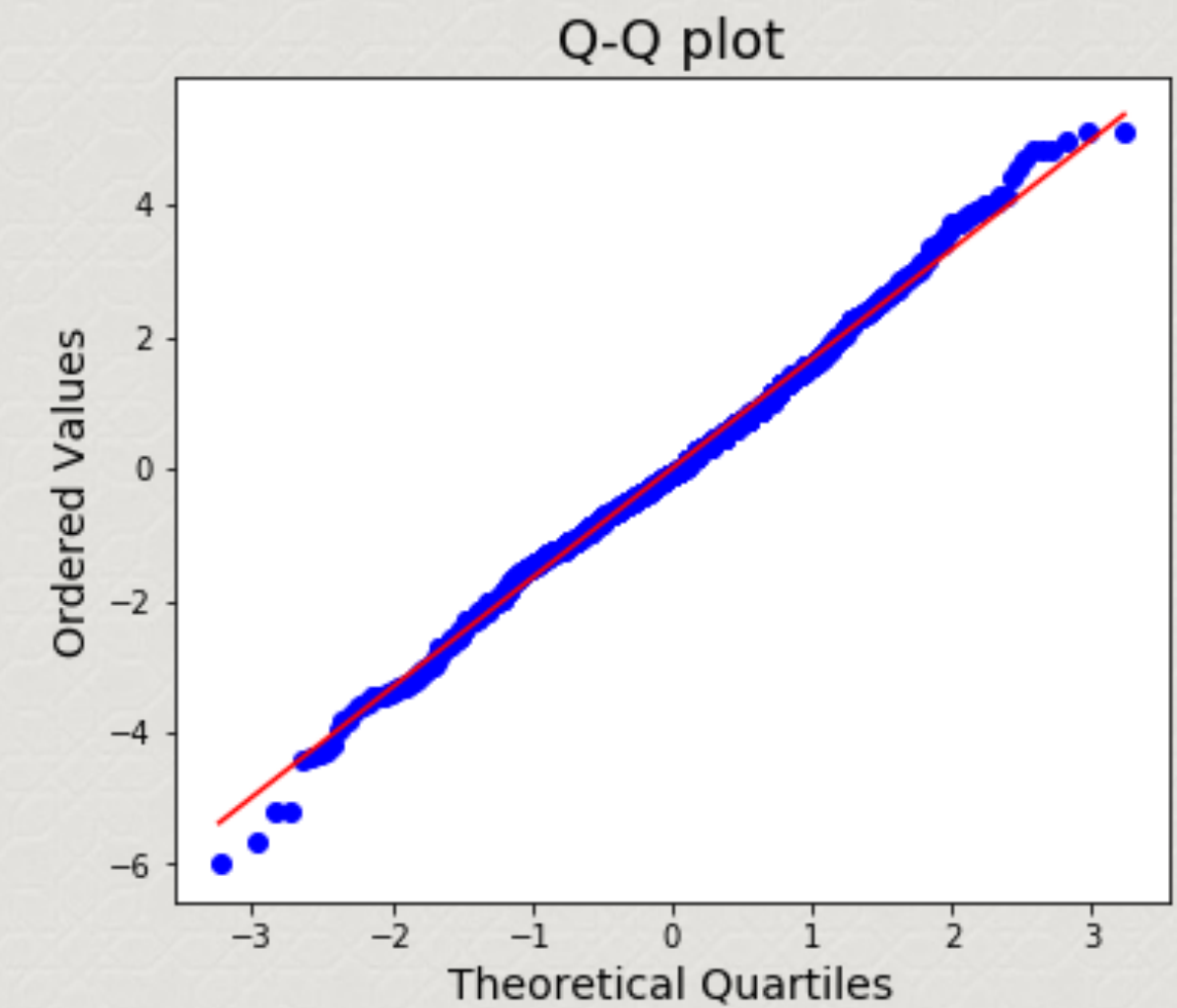
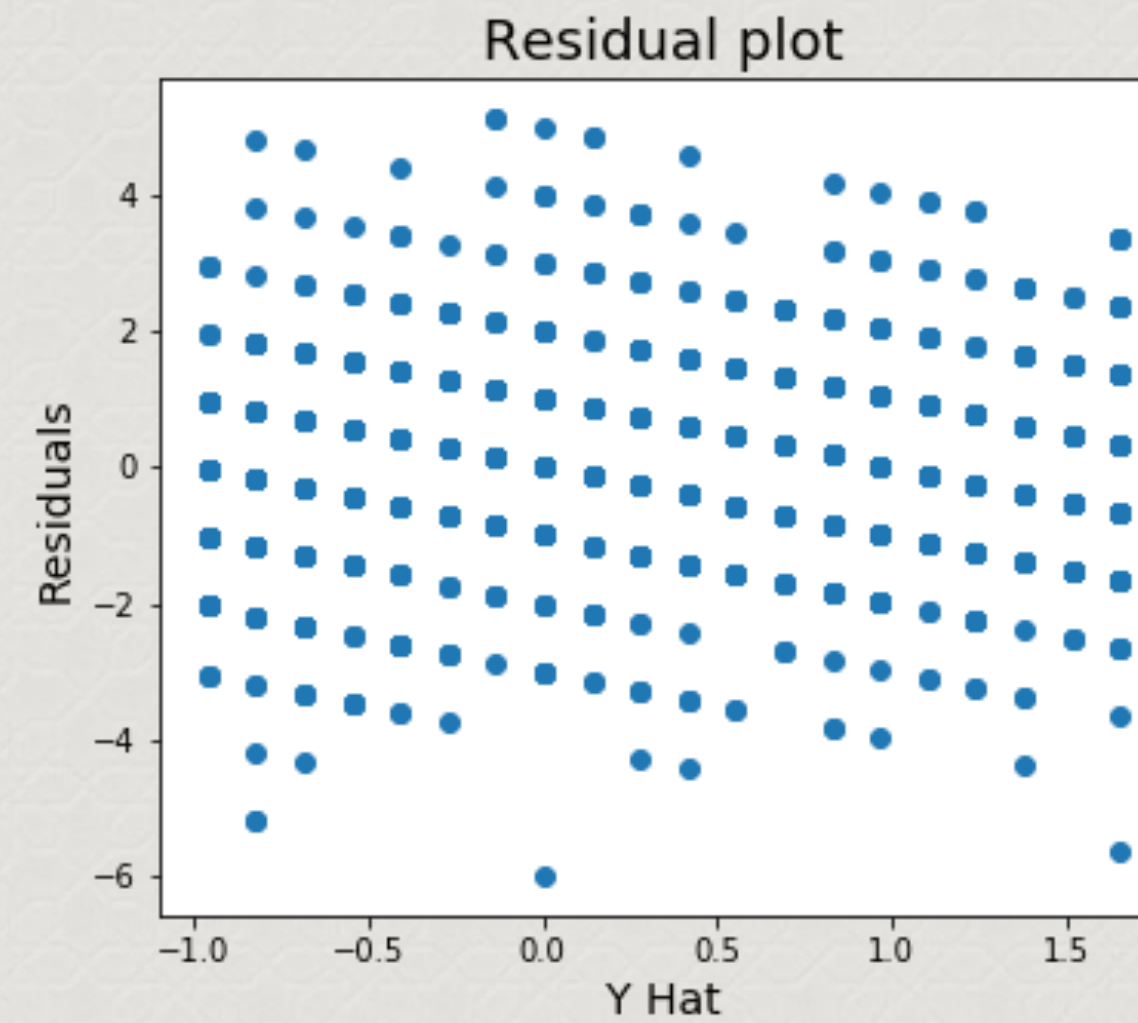
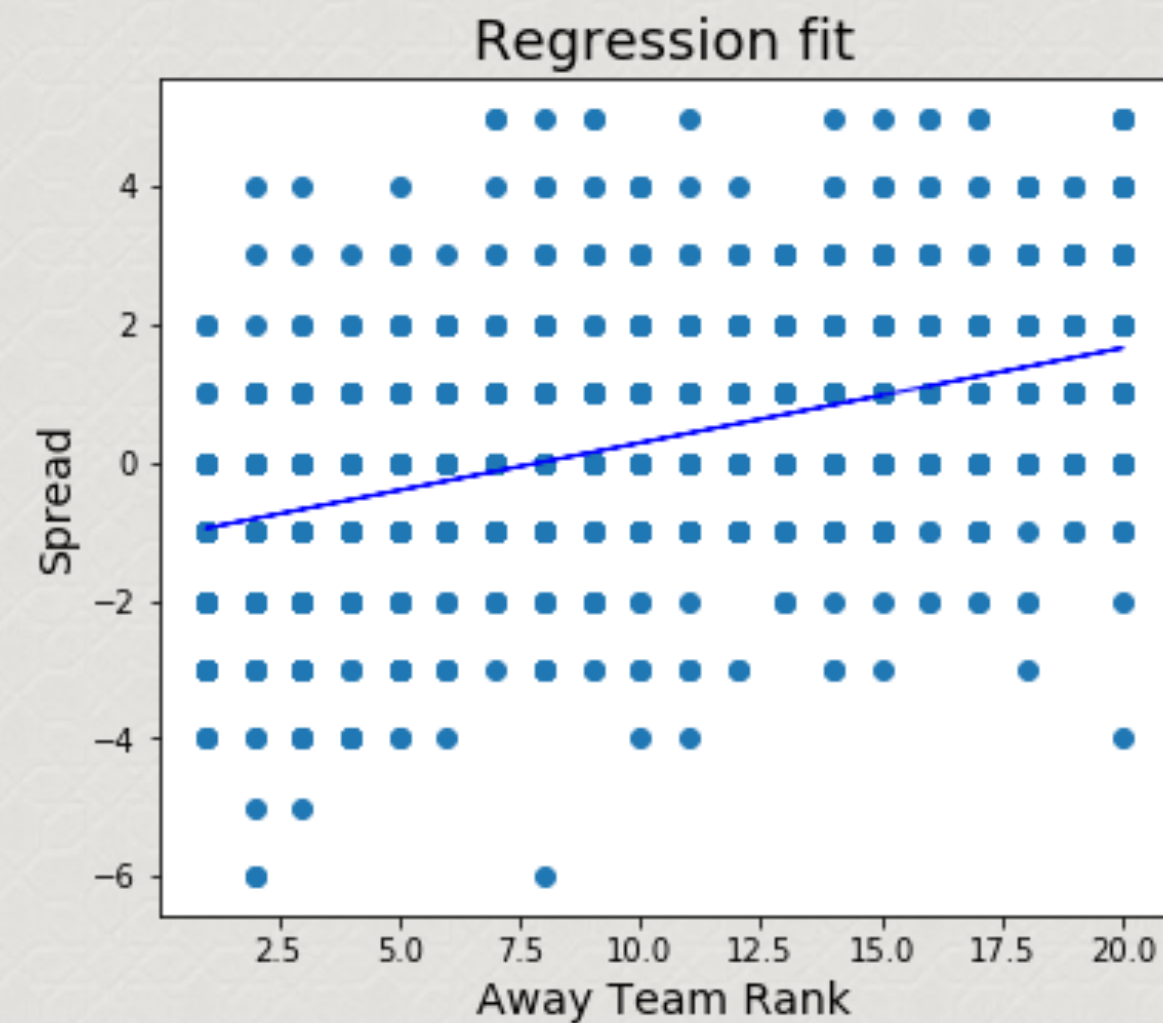
\longrightarrow



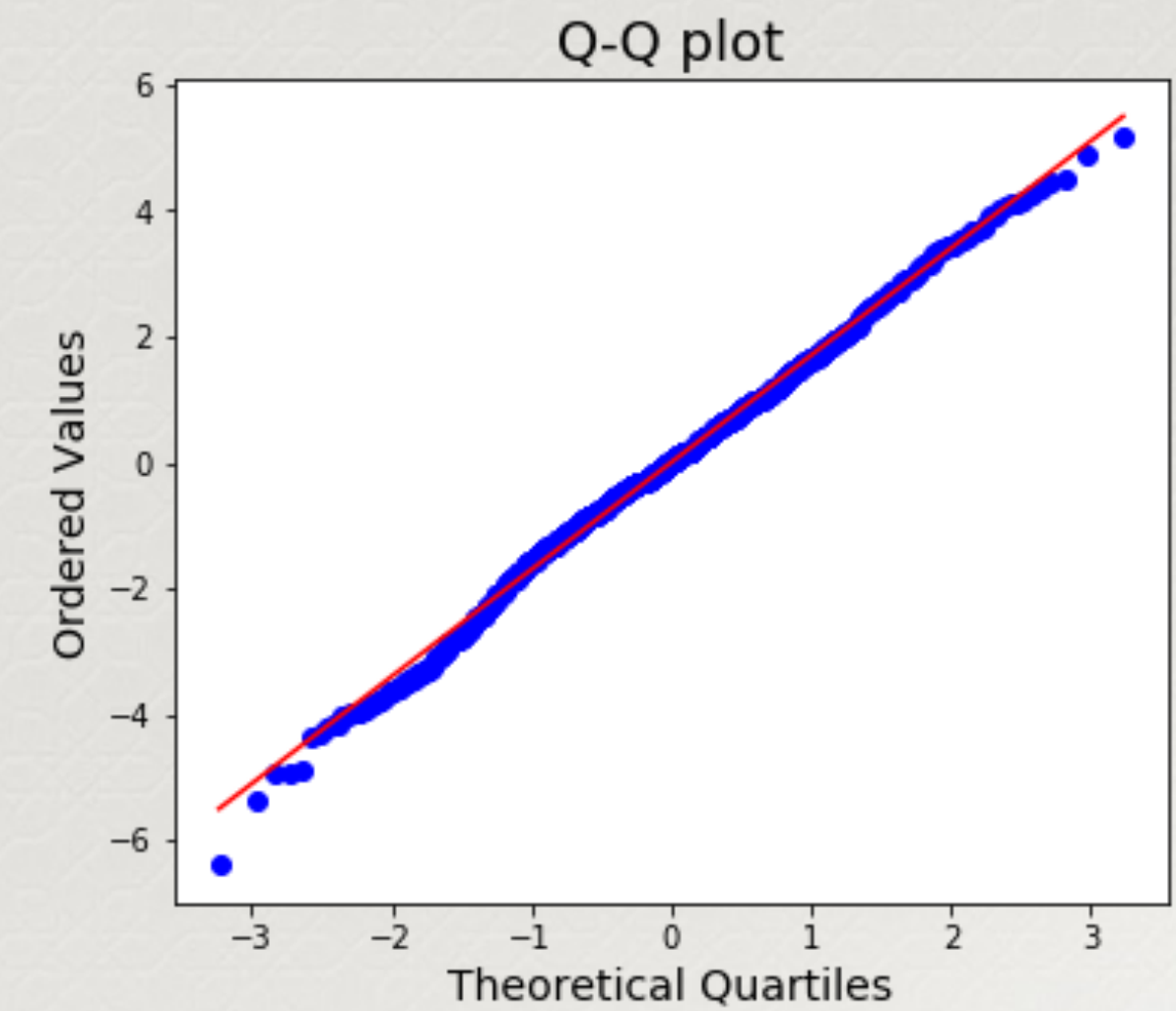
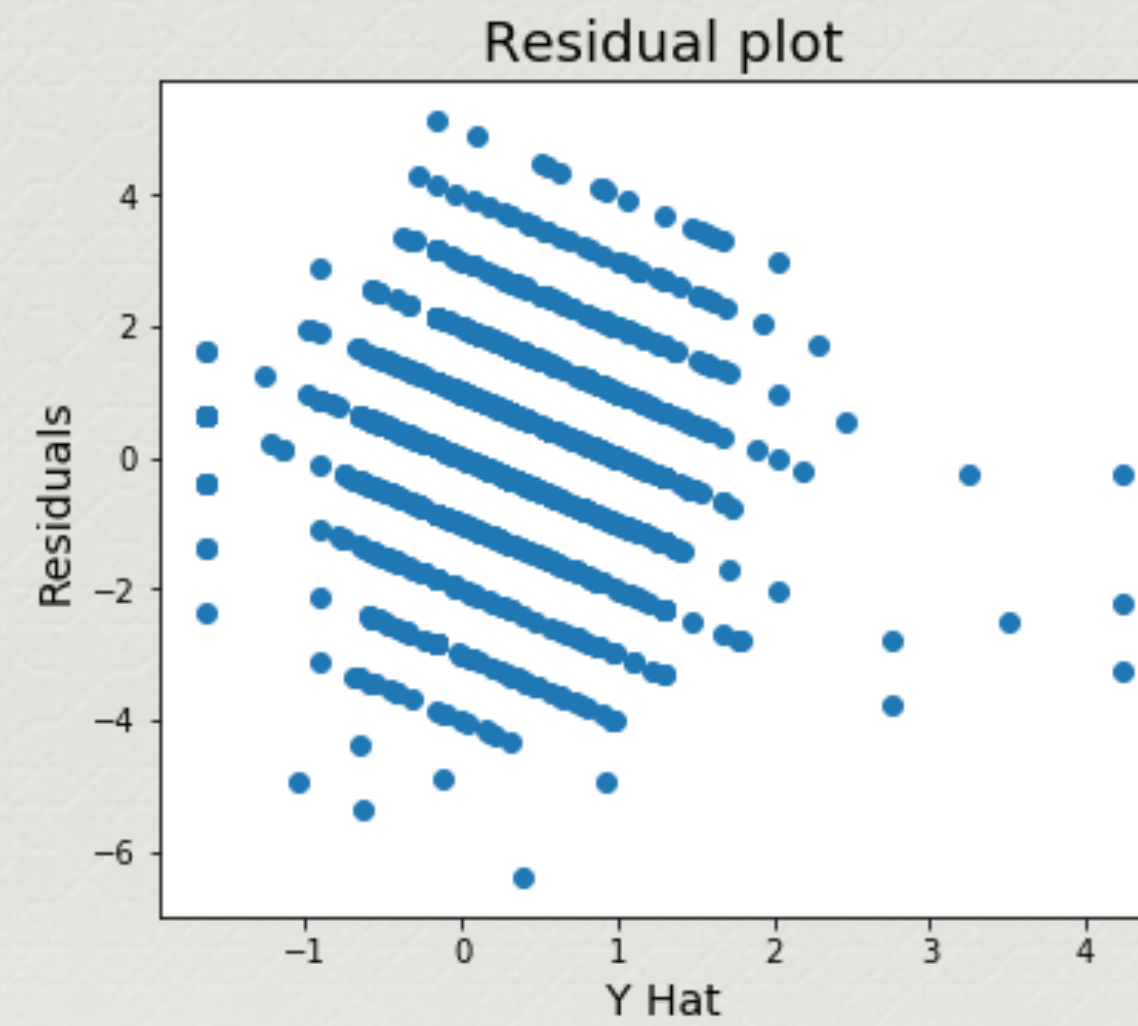
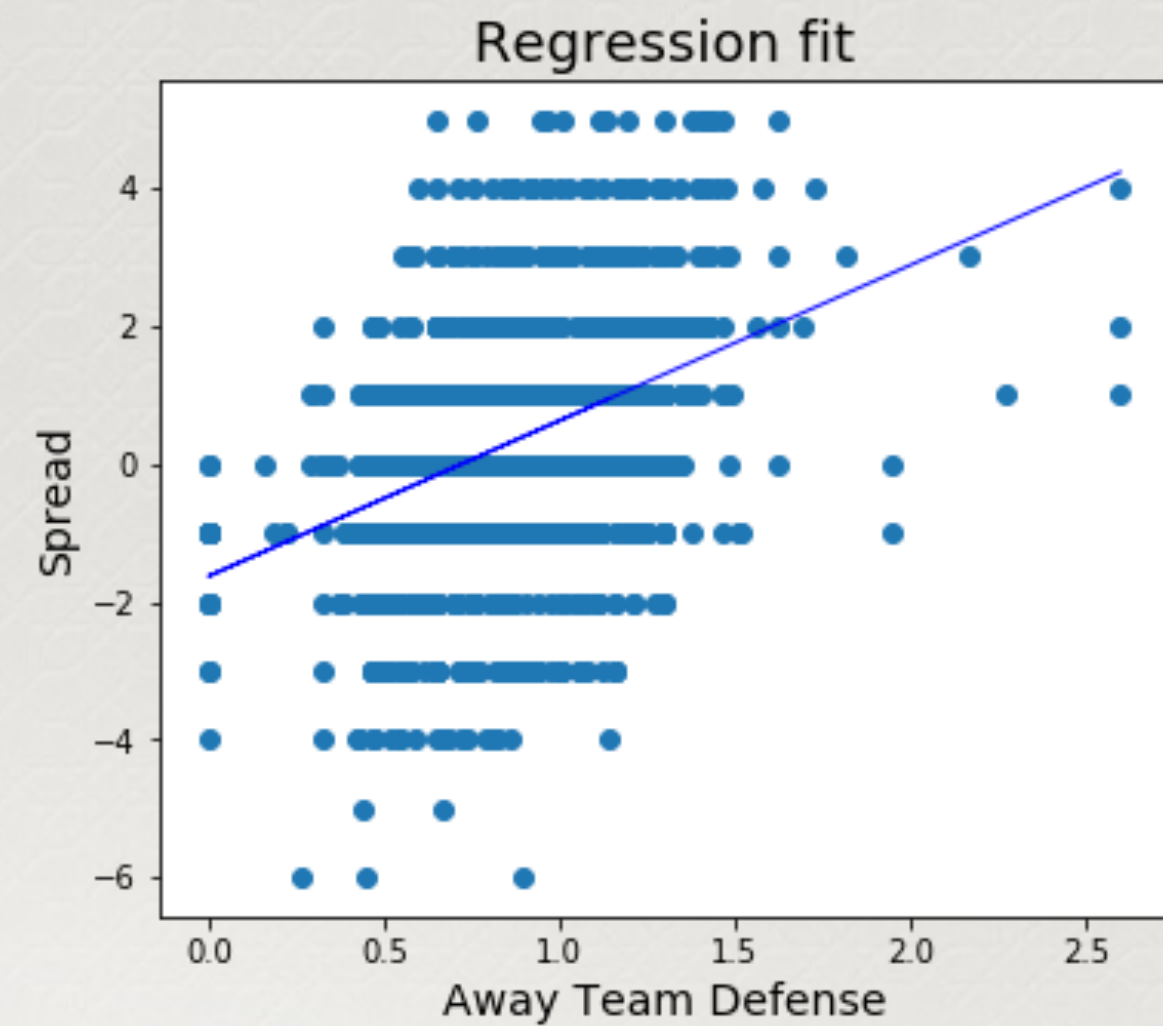
Club	MP	W	D	L	GF	GA	GD	Pts
1  Liverpool	23	19	3	1	54	13	41	60
2  Man. City	23	18	2	3	62	17	45	56
3  Tottenham	23	17	0	6	48	23	25	51
4  Chelsea	23	14	5	4	40	19	21	47
5  Arsenal	23	13	5	5	48	32	16	44
6  Man United	23	13	5	5	46	33	13	44
7  Watford	23	9	6	8	32	32	0	33
8  Wolves	23	9	5	9	27	31	-4	32
9  Leicester City	23	9	4	10	29	29	0	31
10  West Ham	23	9	4	10	30	34	-4	31

Forward Feature Selection

Before:
Lower
predictive
power



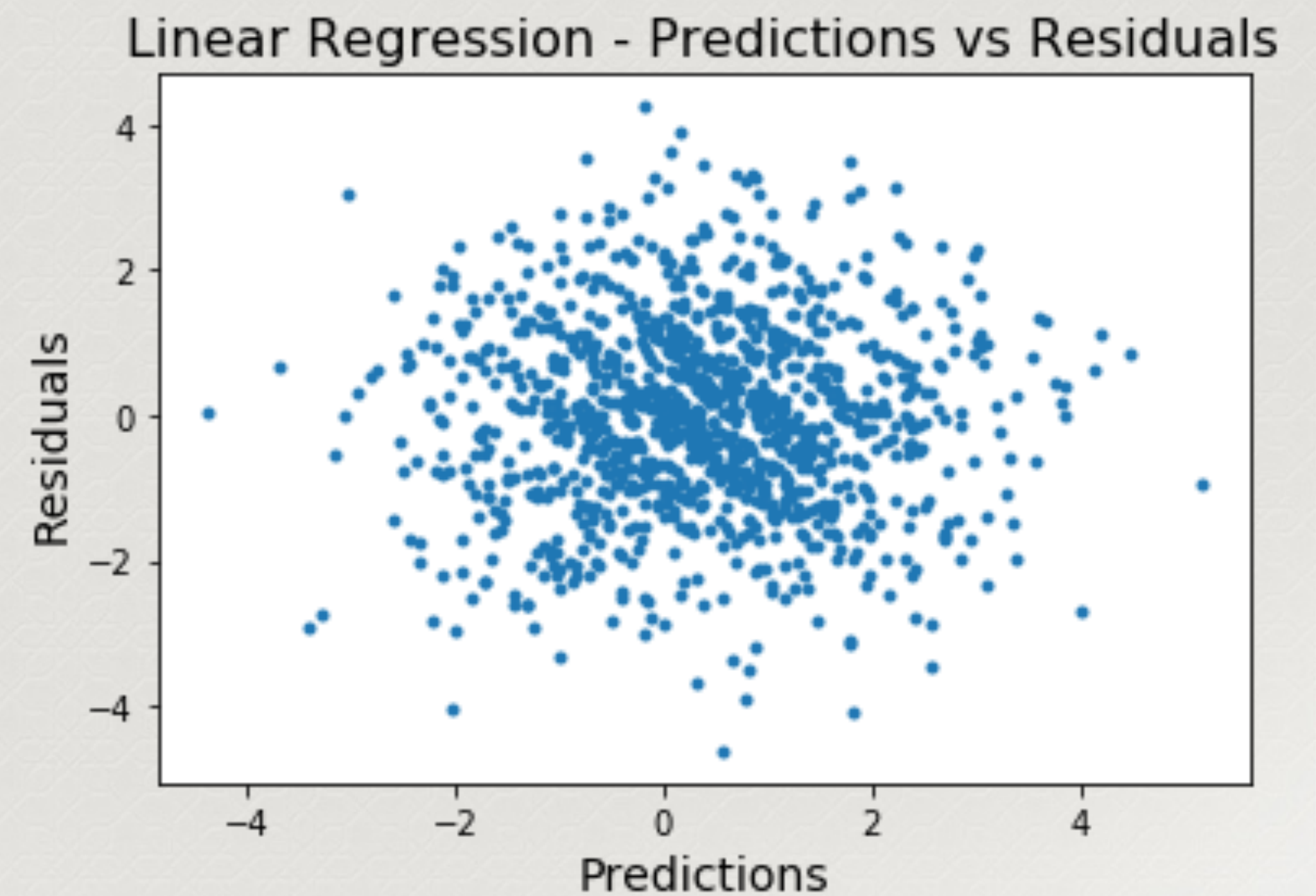
After:
Higher
predictive
power



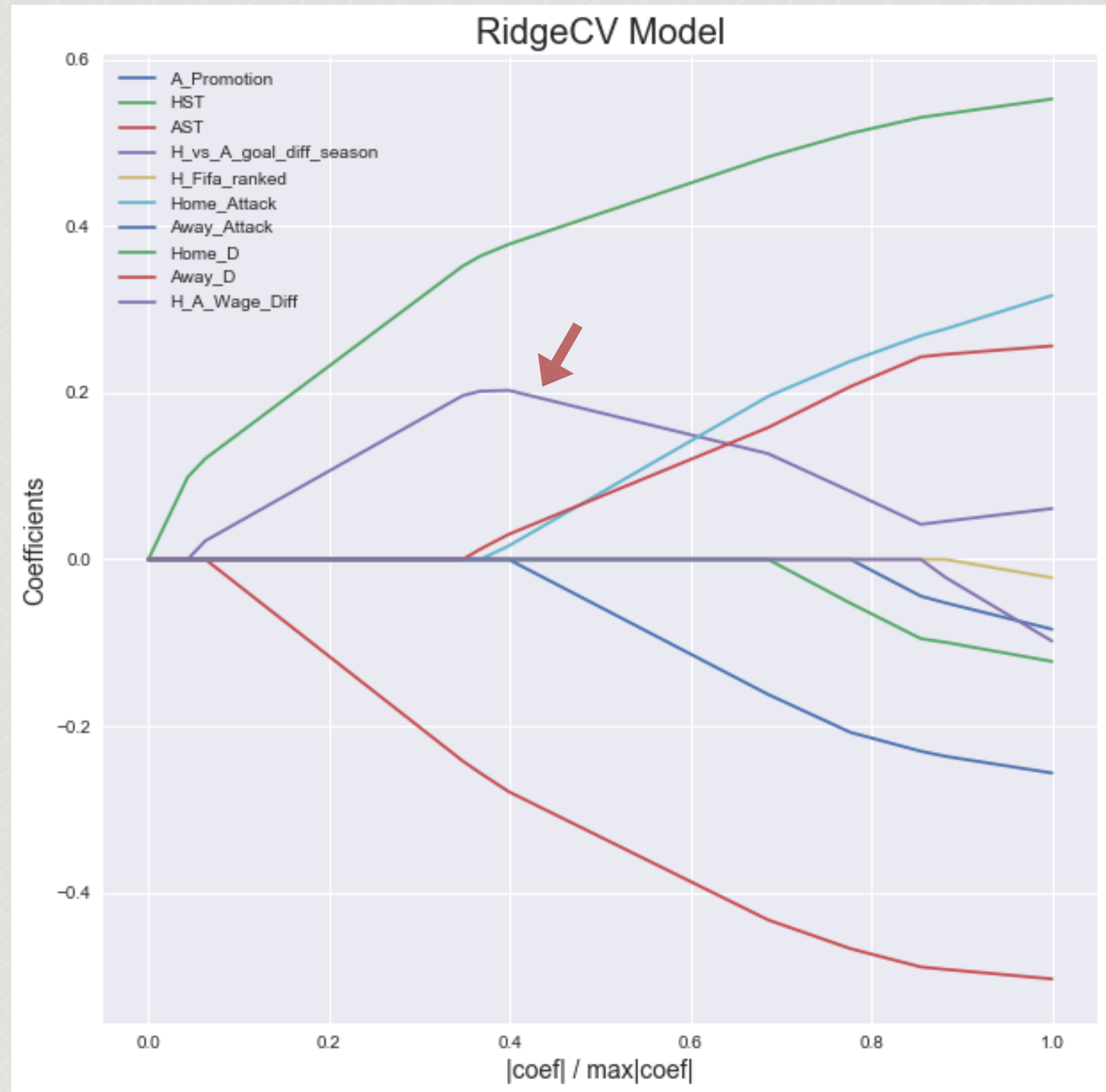
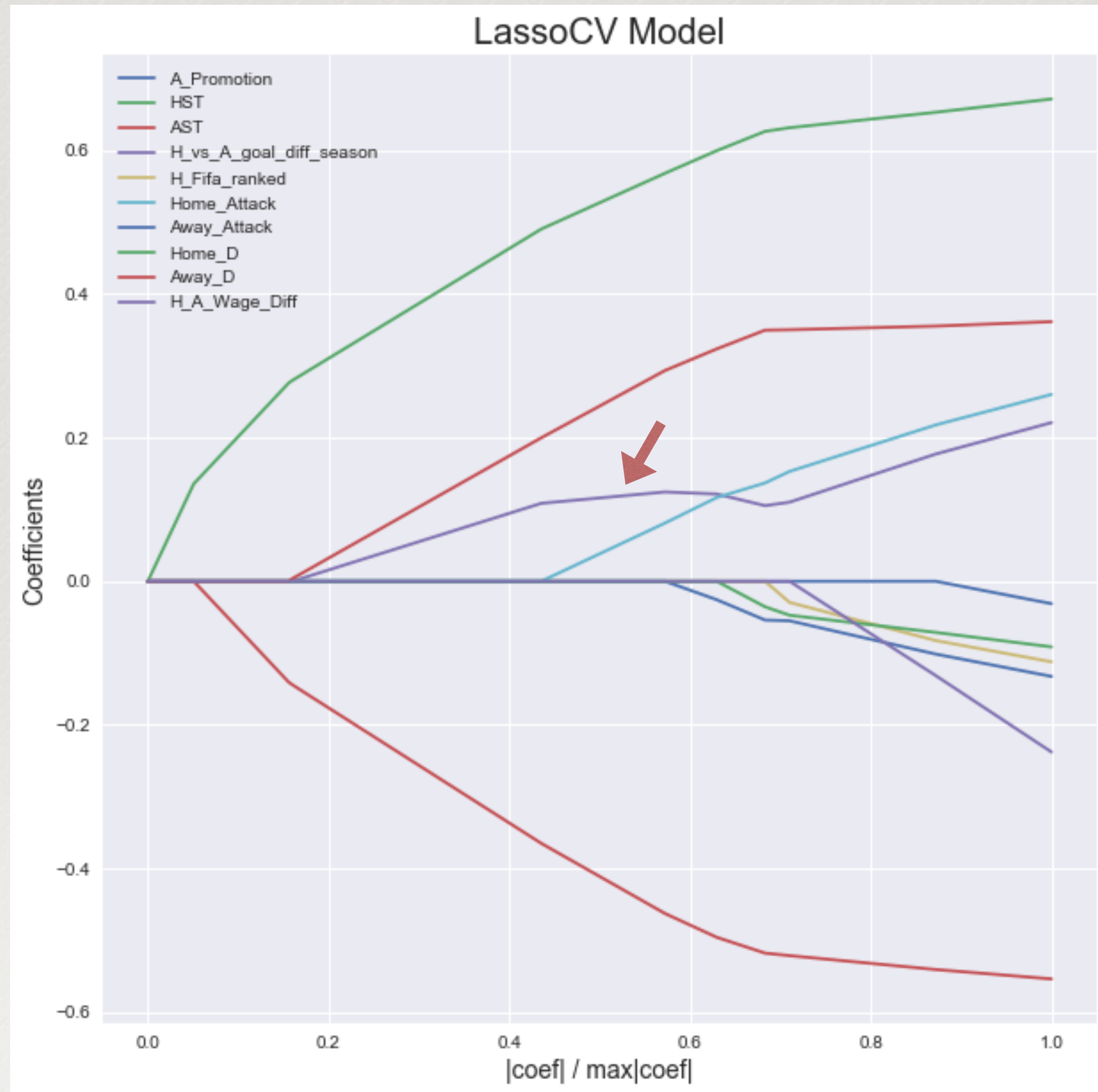
Model Selection

Penalizing the coefficients to reduce training error

Model	R ²	RMSE	Features
OLS	0.48	1.858	4
Lasso	0.52	1.324	10
Ridge	0.52	1.324	10
Elastic Net	0.51	1.434	10



Final Results





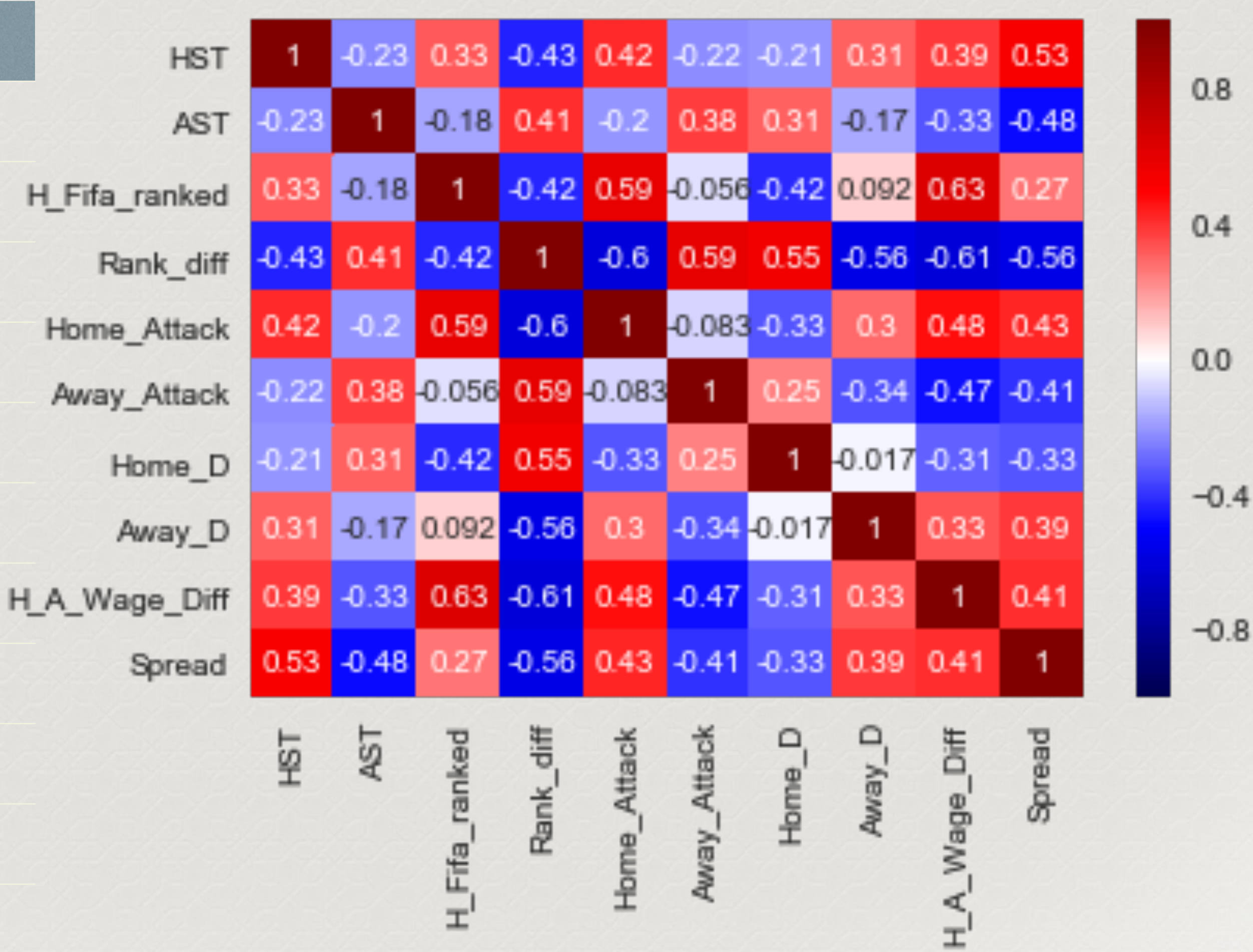
Questions?

Appendix

Feature Selection Table

Feature	Coefficient	RMSE
Away_D	2.25615699	1.821489452
Home_Attack	2.16009807	1.796019978
Away_Attack	-1.57983712	2.220441731
Home_D	-1.53191043	2.155454734
AST	-0.38363586	4.974672695
HST	0.35813822	4.92505232
HomeTeam_rank	-0.13777143	12.11154299
AwayTeam_rank	0.13752723	11.46761267
Rank_ratio	-0.26874679	4.380622114
H_Promotion	0.60268318	1.900369309
H_Fifa_ranked	0.3104352	2.367357

Feature Selection Heat Map



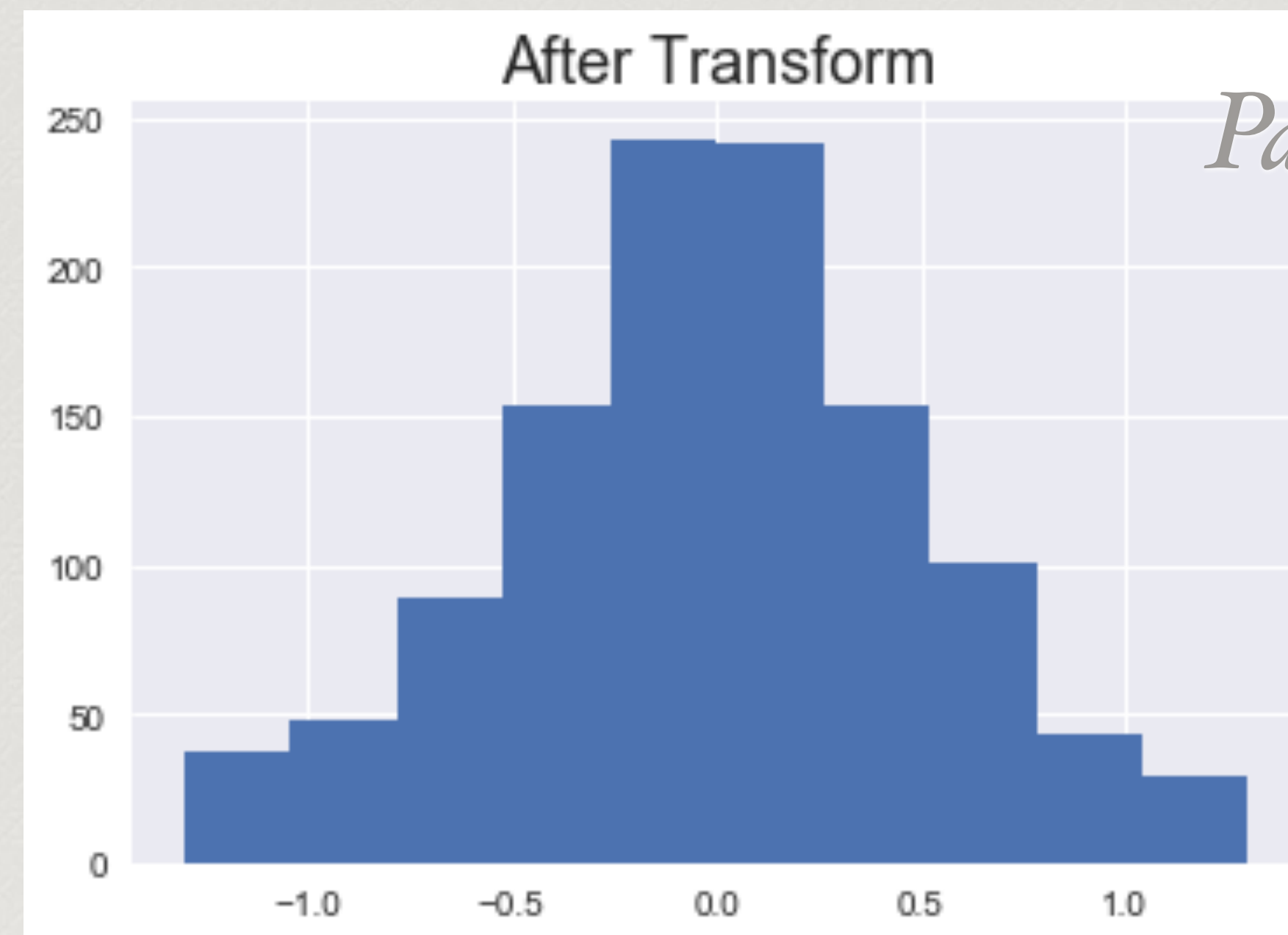
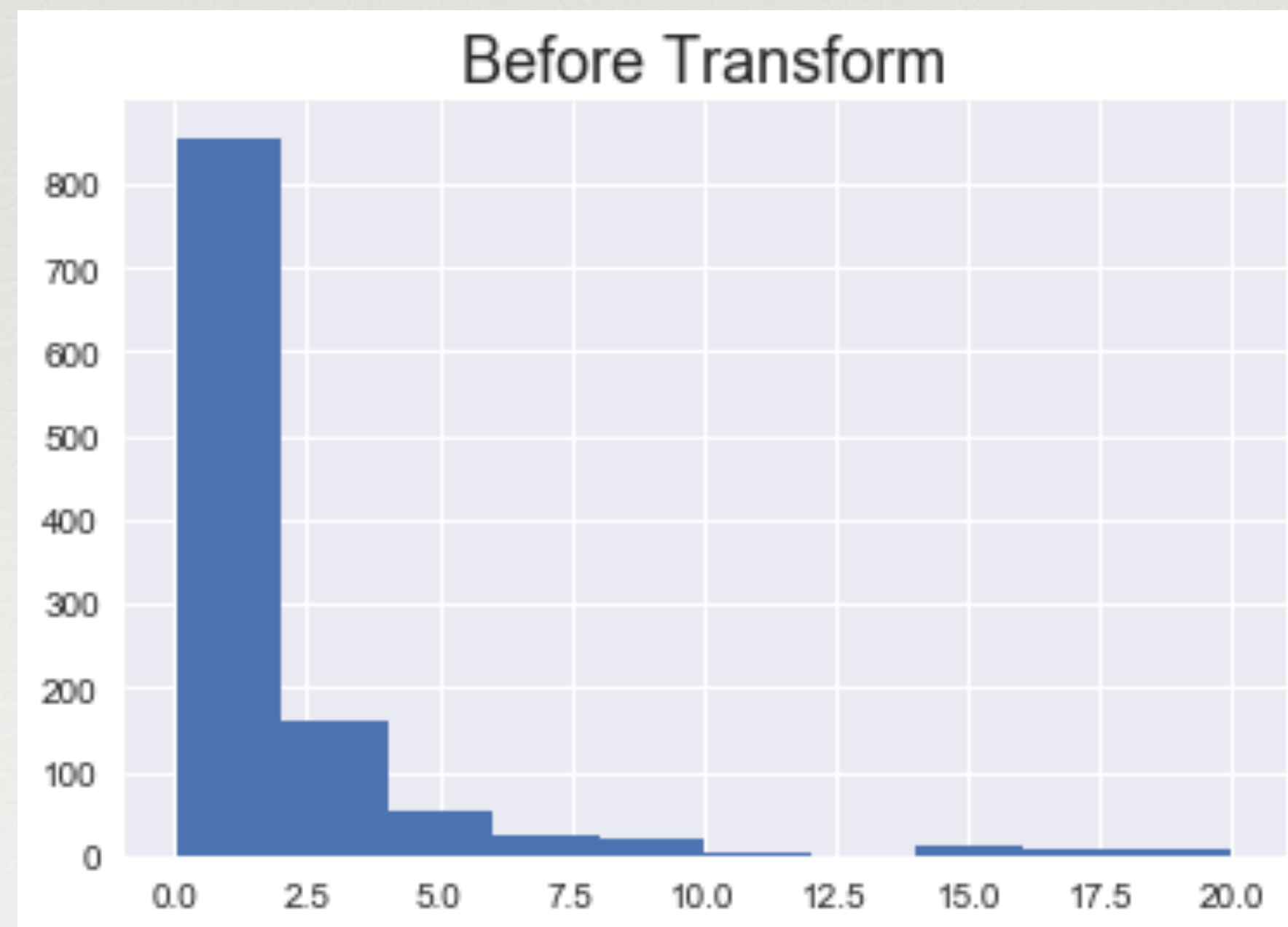
Insights

- 1) There are a lot of draws in soccer
- 2) Fifa top 100 players matter more if they are playing at home
- 3) Promotion side beware of playing on the road,

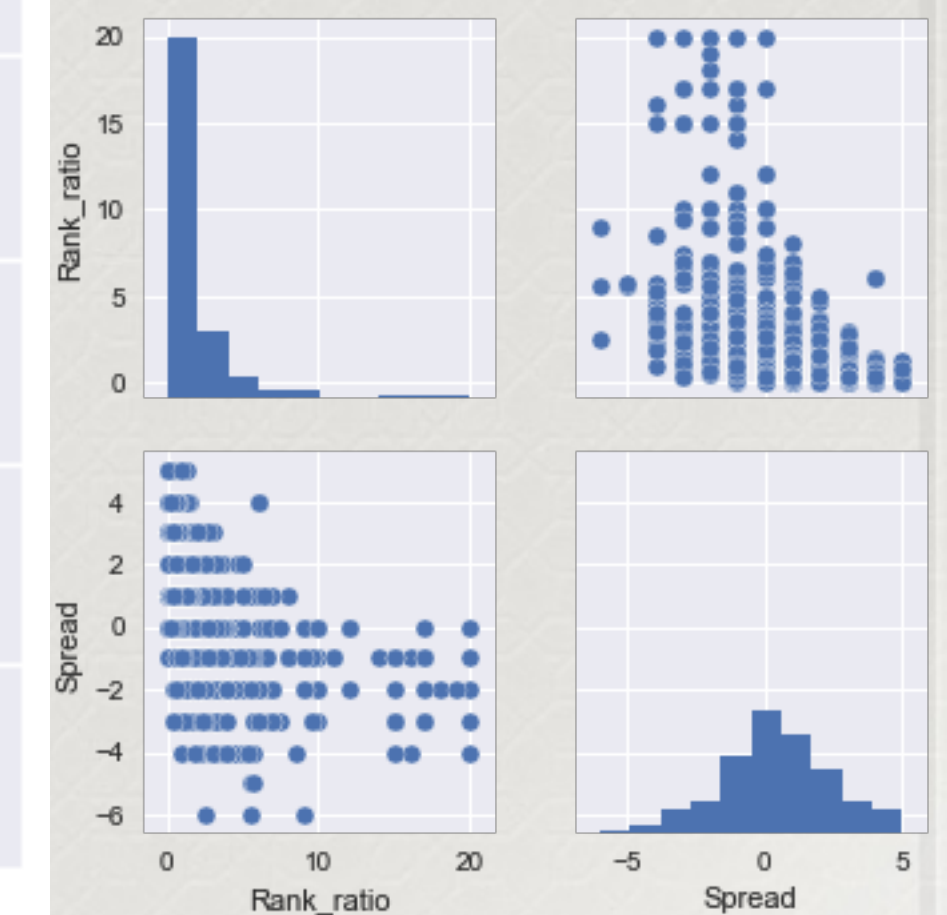


Transformations & Scalers

Rank Ratio Feature with log10



Pair Plot discovery



“Some people think football [soccer] is a matter of life and death. I don’t like that attitude. I can assure them it is much more serious than that.”

– Bill Shankley

