Linear Regression Project : Predicting Goals Scored

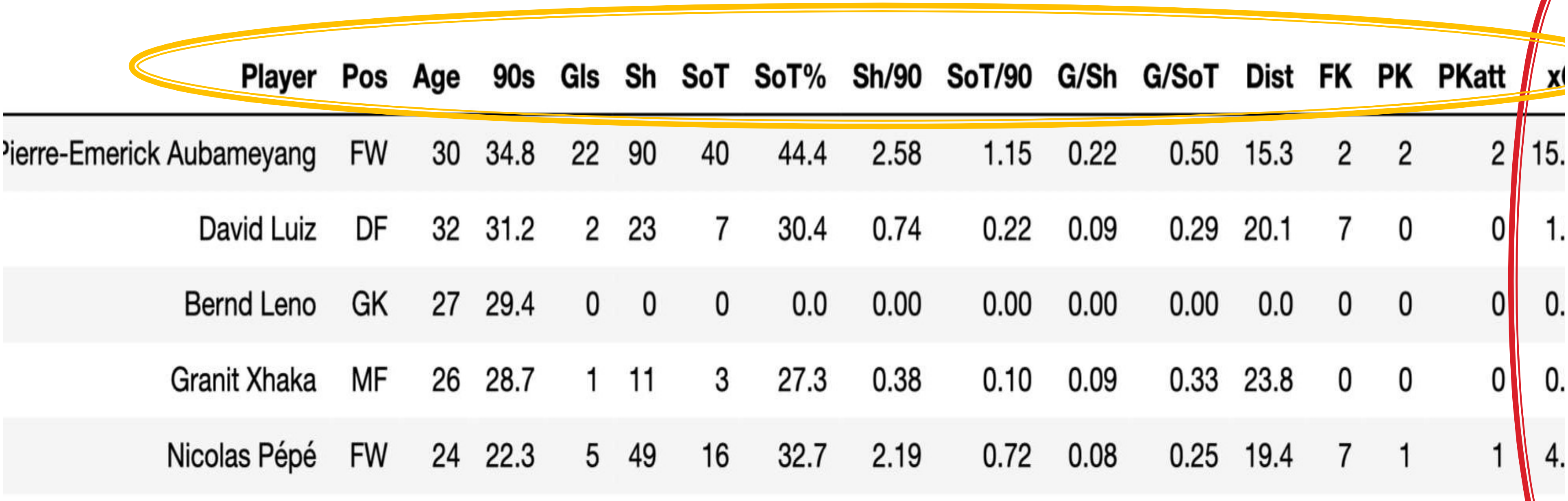
The Need:

- Teams need prolific goal scorers
- Shirt sales
- Contract negotiation, player value
- Agents…etc
- Betting Houses

Task Breakdown

- Methodology: Use linear regression to predict outlay based on player stats (features)
- Data Source: www. Fbrief.com
- Focus: EPL, Bundesliga, Laliga, French Lique1 & Italian Serie A: 28-35players/ team
- Tools: BeautifulSoup, scikit_learn, Matplotlib, Seaborn, Python

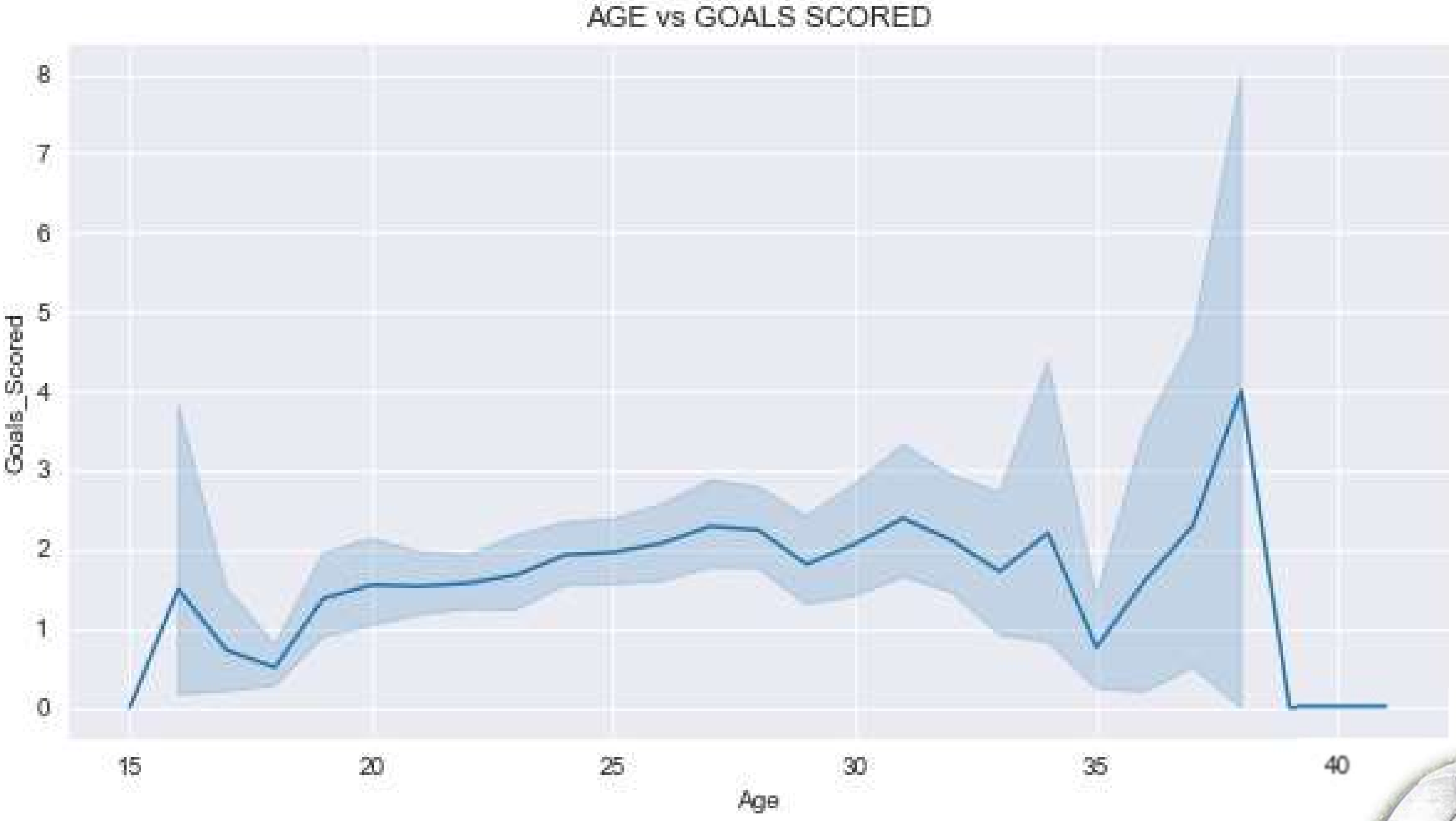
Features/Observations



Player	Pos	Age	90s	Gls	Sh	SoT	SoT%	Sh/90	SoT/90	G/Sh	G/SoT	Dist	FK	PK	PKatt	x(
ierre-Emerick Aubameyang	FW	30	34.8	22	90	40	44.4	2.58	1.15	0.22	0.50	15.3	2	2	2	15.
David Luiz	DF	32	31.2	2	23	7	30.4	0.74	0.22	0.09	0.29	20.1	7	0	0	1.
Bernd Leno	GK	27	29.4	0	0	0	0.0	0.00	0.00	0.00	0.00	0.0	0	0	0	0.
Granit Xhaka	MF	26	28.7	1	11	3	27.3	0.38	0.10	0.09	0.33	23.8	0	0	0	0.
Nicolas Pépé	FW	24	22.3	5	49	16	32.7	2.19	0.72	0.08	0.25	19.4	7	1	1	4.

PLAYER POSITIONS: CATEGORICAL								
GK	DF	DM/MF	MF/DF	DF/FW	FW/DF	MF	FW/MF	FW
0	1	2	2	3	3	4	5	6

EDA: Age – goals don't last!

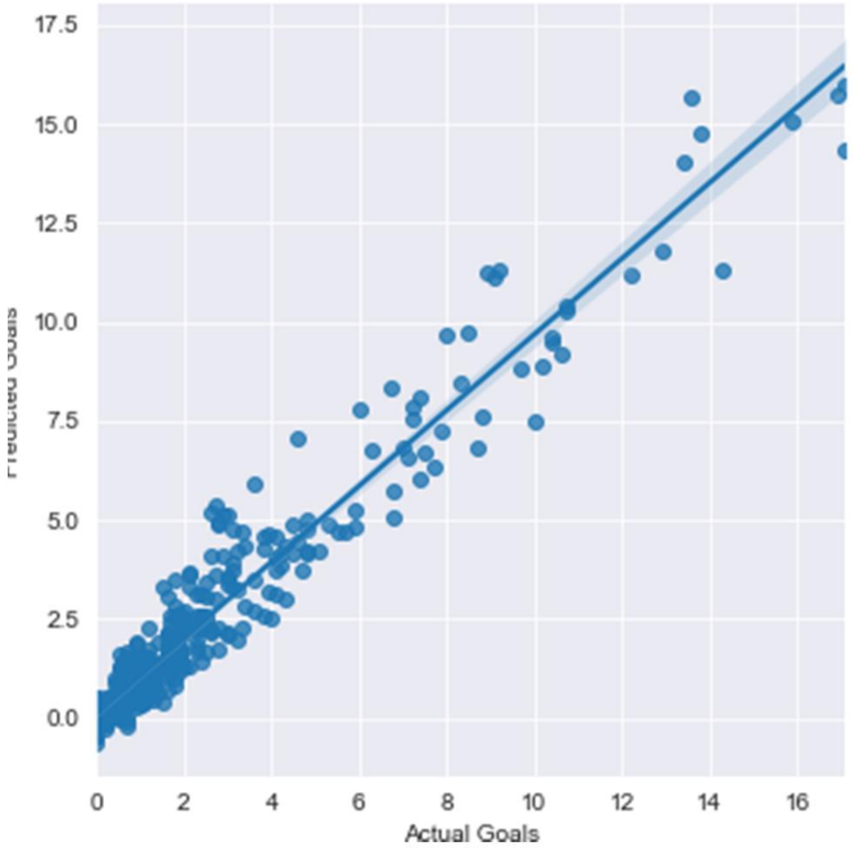


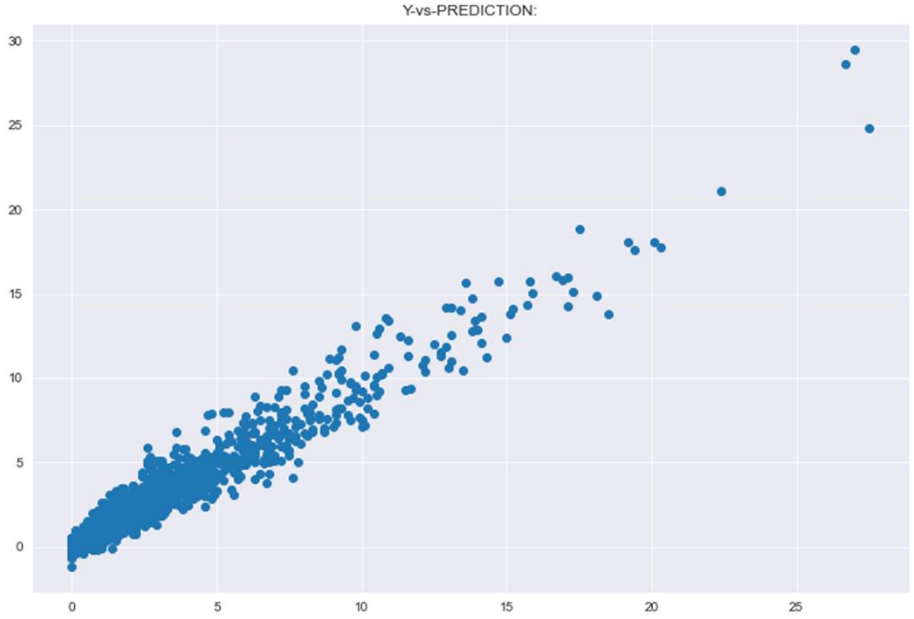
Training the Model

	TRAIN	VAL	SPLIT
PROPORTION	60%	20%	20%
TRAIN SCORE	0.90	TEST_SCORE:	0.91
ERROR	MAE:0.44	MSE:0.445	RMSE: 0.66

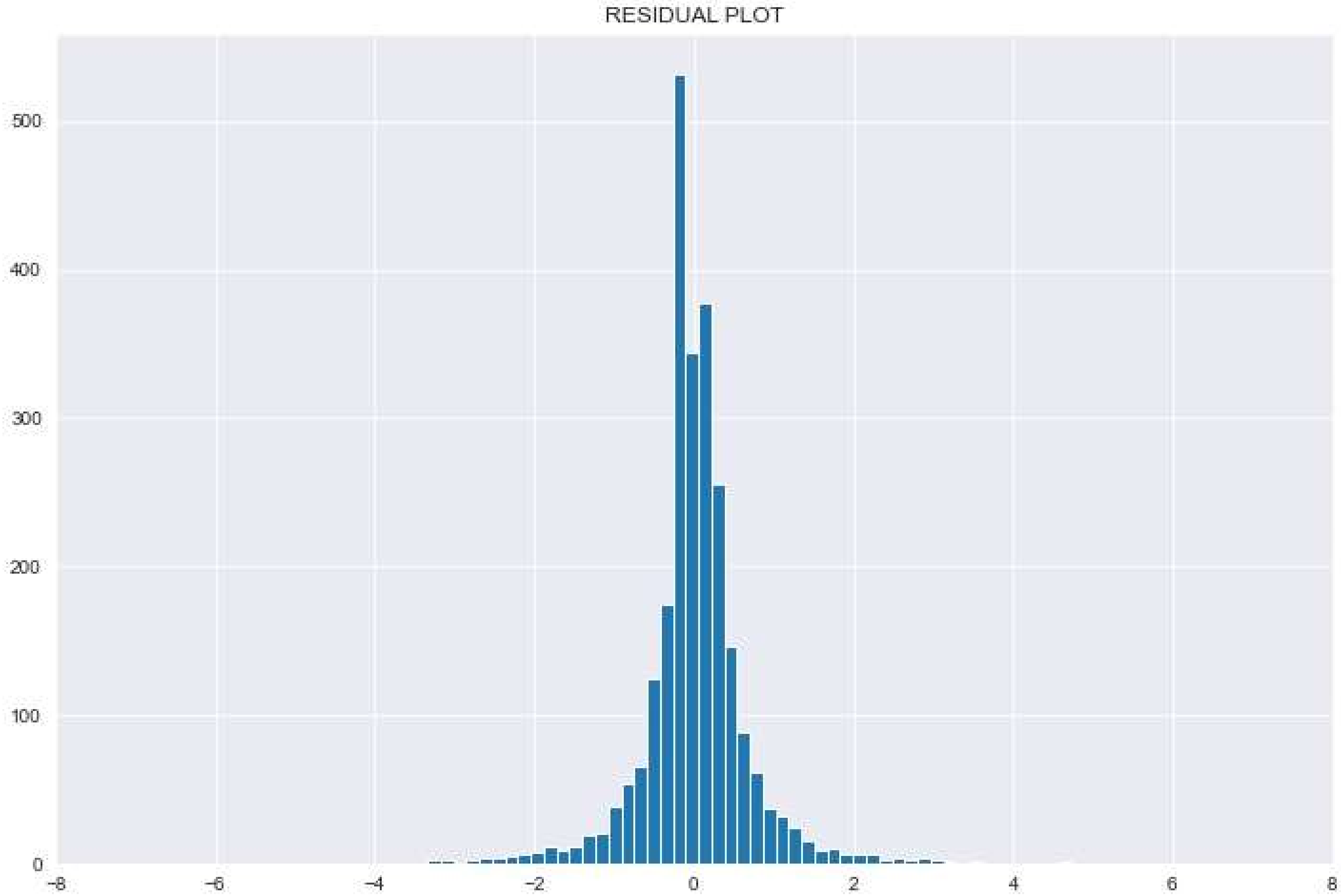
CROSS VALIDATION: 20% & RIDGE				
🟥				
	🟥			
		🟥		
			🟥	
				🟥
CV MEAN R^2 (FULL DSET): 0.8870				
RIDGE MEAN R^2(FULL DSET):0.88				

FIT PROGRESSION

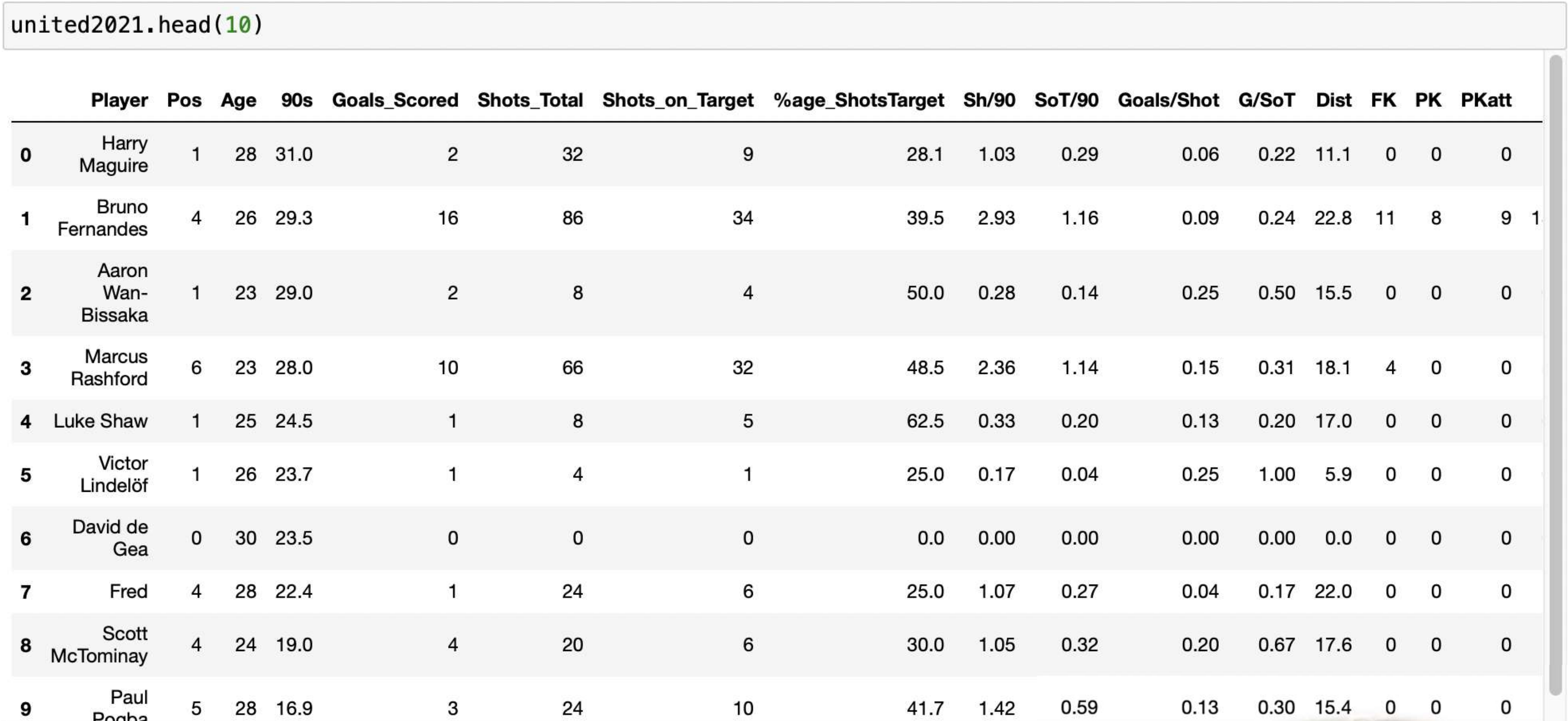




Residual Plot on Final CV



RAW DATA: UNITED 2021

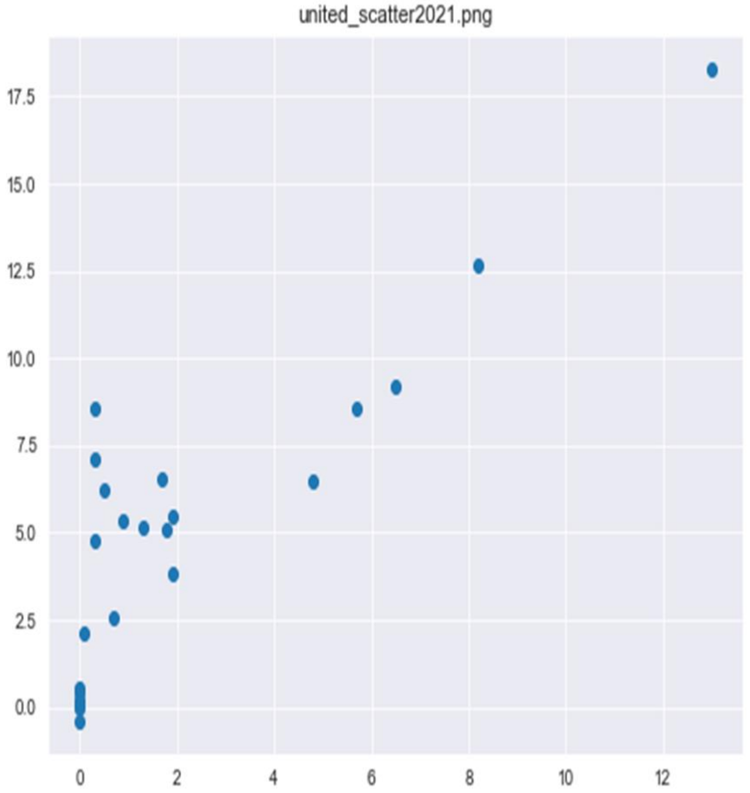


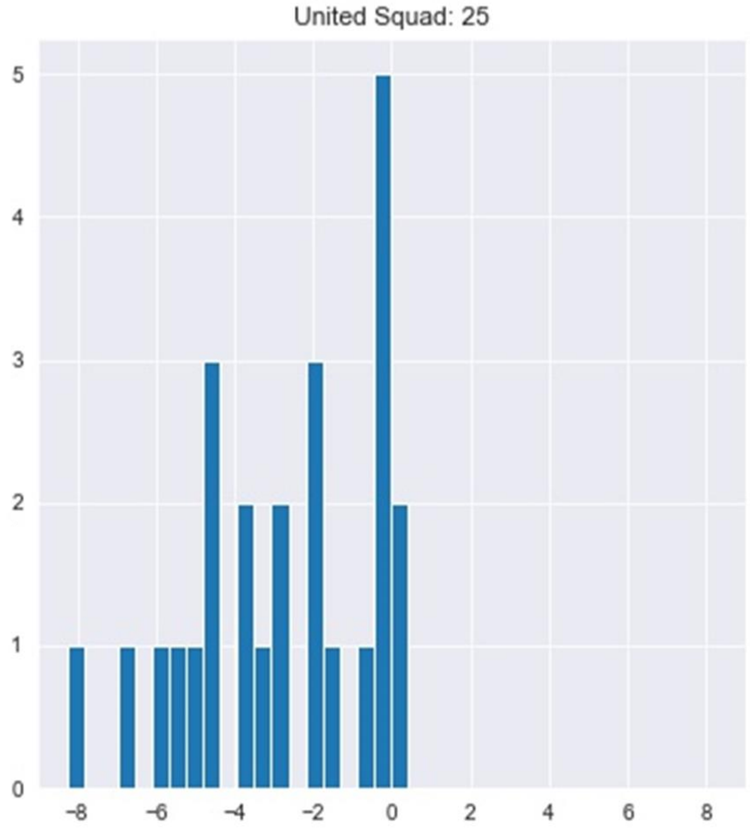
```
united2021.head(10)
```

	Player	Pos	Age	90s	Goals_Scored	Shots_Total	Shots_on_Target	%age_ShotsTarget	Sh/90	SoT/90	Goals/Shot	G/SoT	Dist	FK	PK	PKatt	
0	Harry Maguire	1	28	31.0	2	32	9	28.1	1.03	0.29	0.06	0.22	11.1	0	0	0	
1	Bruno Fernandes	4	26	29.3	16	86	34	39.5	2.93	1.16	0.09	0.24	22.8	11	8	9	1
2	Aaron Wan-Bissaka	1	23	29.0	2	8	4	50.0	0.28	0.14	0.25	0.50	15.5	0	0	0	
3	Marcus Rashford	6	23	28.0	10	66	32	48.5	2.36	1.14	0.15	0.31	18.1	4	0	0	
4	Luke Shaw	1	25	24.5	1	8	5	62.5	0.33	0.20	0.13	0.20	17.0	0	0	0	
5	Victor Lindelöf	1	26	23.7	1	4	1	25.0	0.17	0.04	0.25	1.00	5.9	0	0	0	
6	David de Gea	0	30	23.5	0	0	0	0.0	0.00	0.00	0.00	0.00	0.0	0	0	0	
7	Fred	4	28	22.4	1	24	6	25.0	1.07	0.27	0.04	0.17	22.0	0	0	0	
8	Scott McTominay	4	24	19.0	4	20	6	30.0	1.05	0.32	0.20	0.67	17.6	0	0	0	
9	Paul Pogba	5	28	16.9	3	24	10	41.7	1.42	0.59	0.13	0.30	15.4	0	0	0	

Simple Score: -0.30
Ridge Score: -104.13

Model abhores raw data?



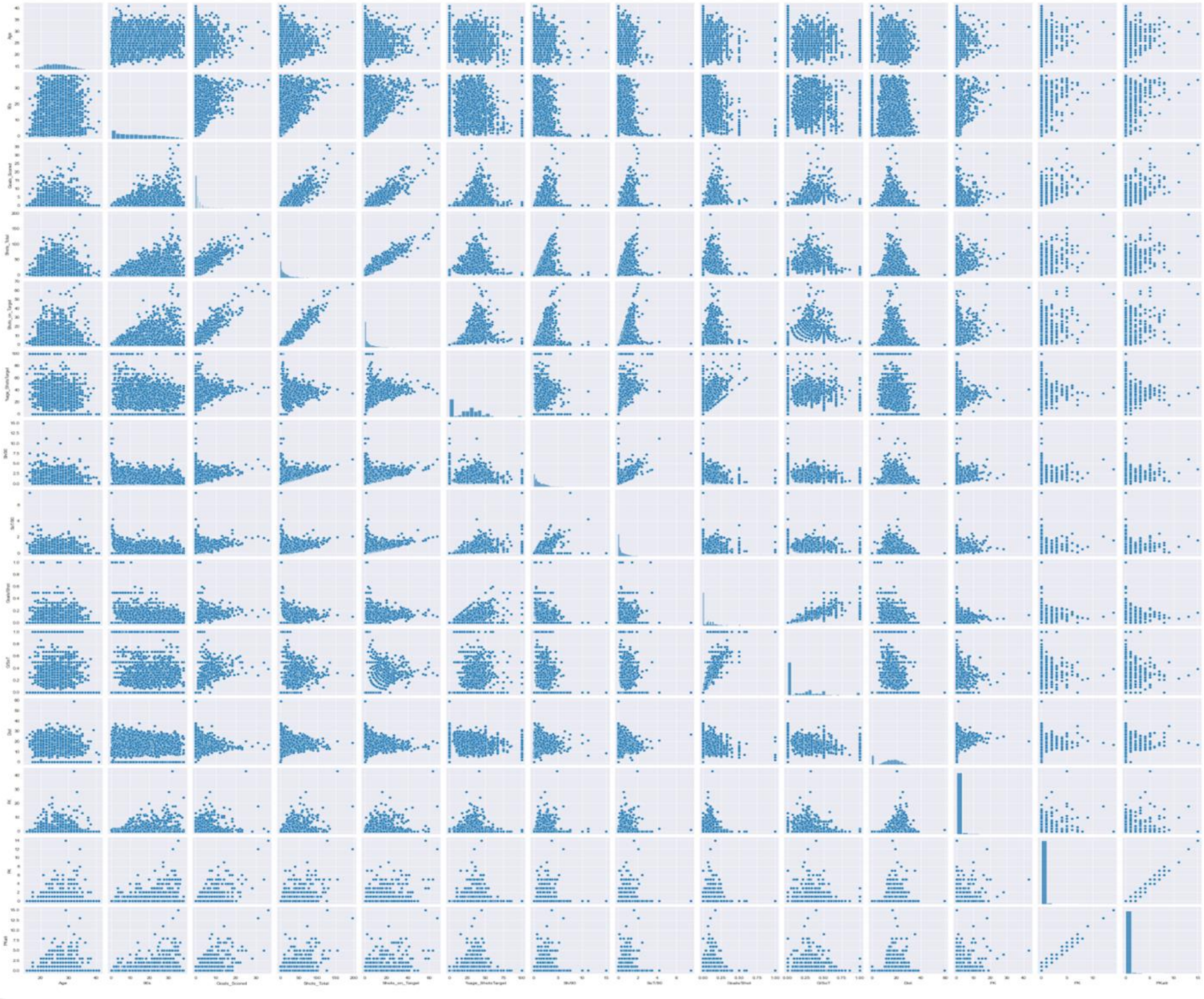


Future work

- Much larger dataset could boost accuracy
- More rigorous feature selection (Lasso)
- Lacks data on free-kicks (dead-ball situations)

- Thank you

Appendix

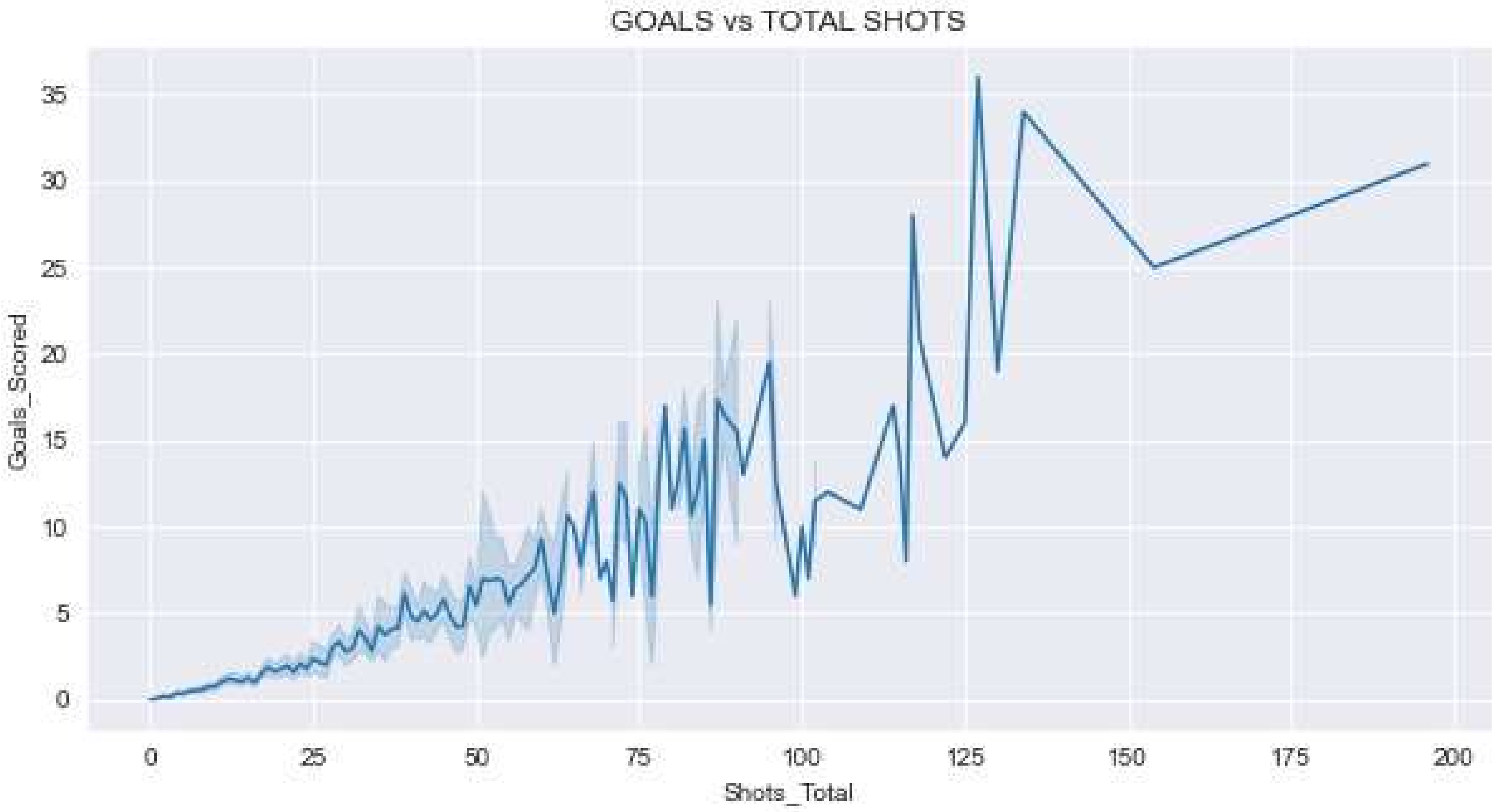


Age
90s
Goals_Scored
Shots_Total
Shots_on_Target
%age_ShotsTarget
Sh/90
SoT/90
Goals/Shot
G/SoT
Dist
FK
PK
PKatt

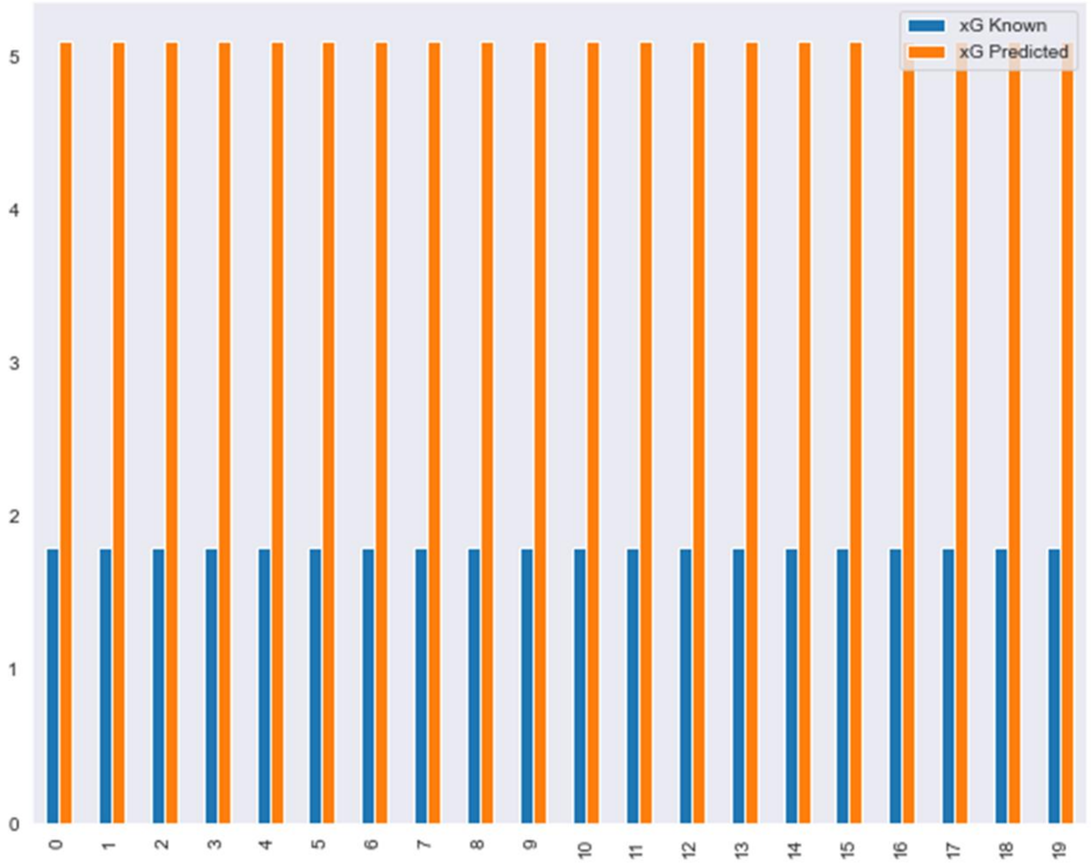
Appendix

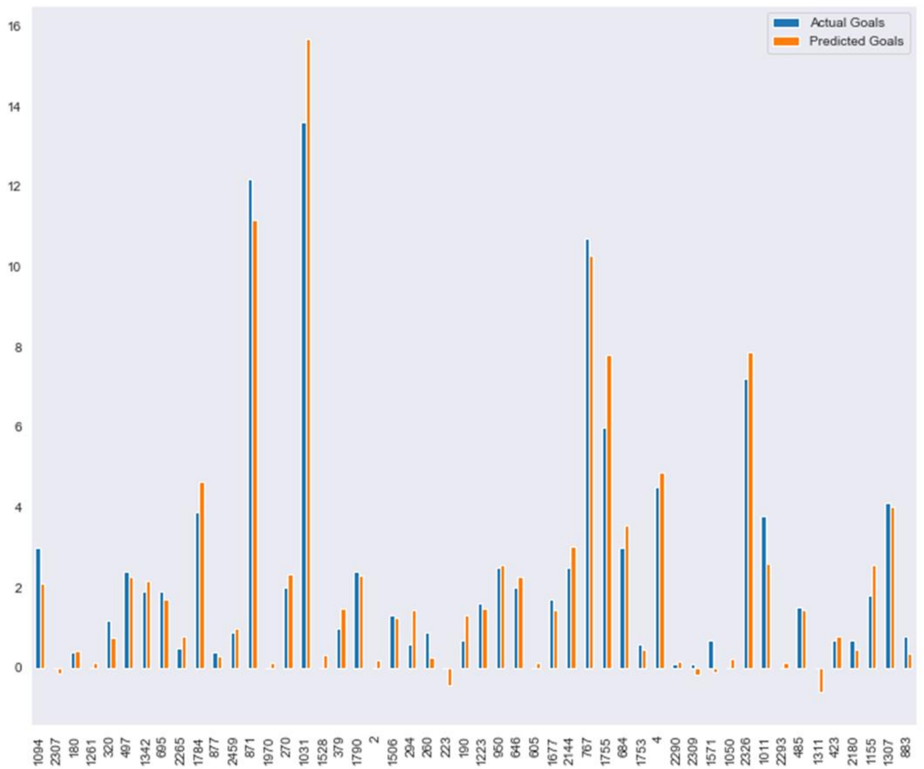
- Fields removed:
- Matches
- Expected Non-penalty goals
- %age shots on target
- Nation
- Shots_total
- Shots/90min
- Goals/shot
- Goals/shot on target
- Free kicks
- Penalty kicks attempted

EDA: Goal/Shot relationship



United residual barplot




Actual Goals
Predicted Goals
16
14
12
10
8
6
4
2
0
1094
2307
180
1261
320
497
1342
695
2265
1784
877
2459
871
1970
270
1031
1528
379
1790
2
1506
294
260
223
190
1223
950
646
605
1677
2144
767
1755
684
1753
4
2290
2309
1571
1050
2326
1011
2293
485
1311
423
2180
1155
1307
883