

The Havertz Effect: How Havertz Shifts Arsenal's Match Outcomes

By Matheus Grover

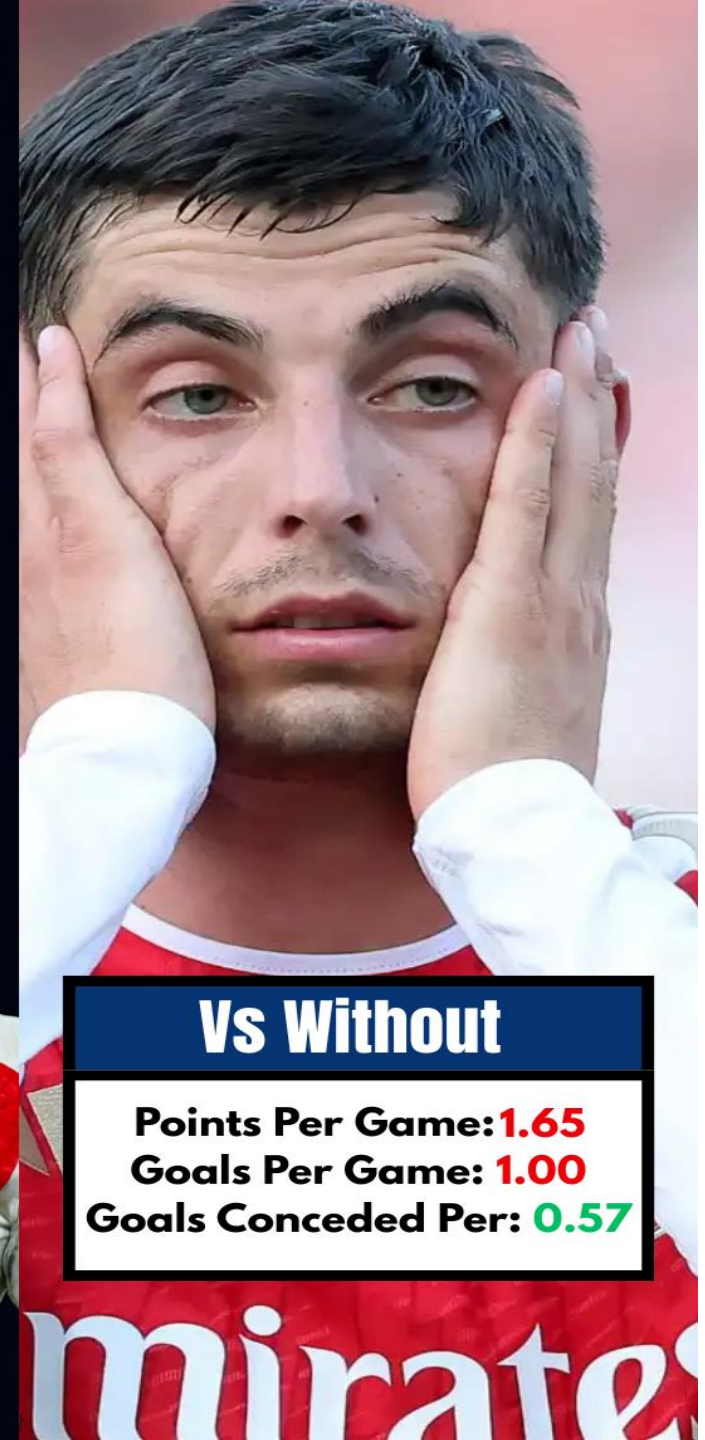
Arsenal With Kai Havertz

Points Per Game: **2.13**
Goals Per Game: **1.58**
Goals Conceded Per: **0.88**



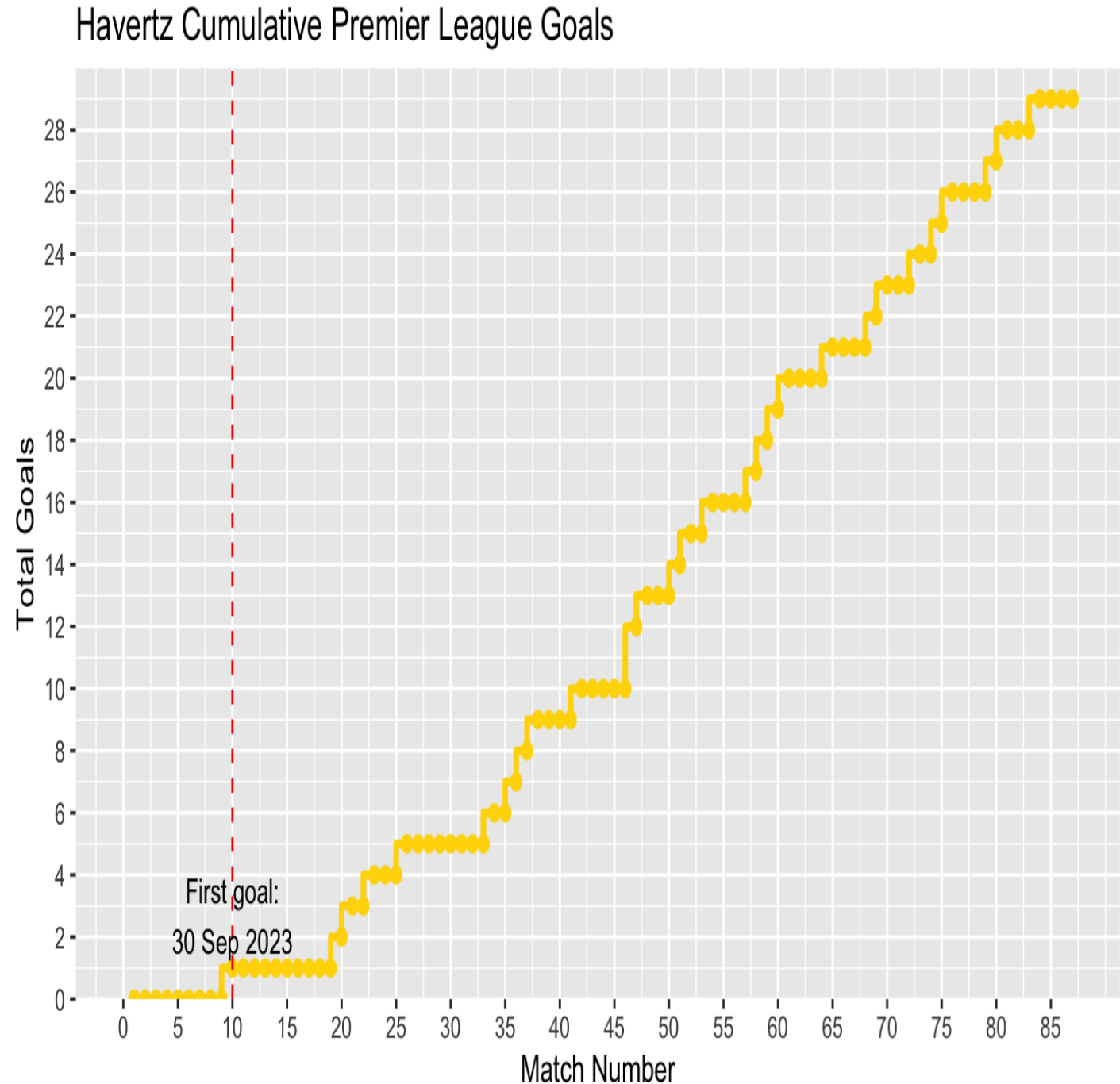
Vs Without

Points Per Game: **1.65**
Goals Per Game: **1.00**
Goals Conceded Per: **0.57**



Why Kai Havertz?

- Favorite Player
- Signed recently (July 2023) for \$65 Millions and paid the largest wages on the team
- Debate on if he was worth the price or good enough for the team
- Hated at first and still a very divisive player
- First goal was 10 games into the season and took awhile to find his form



The Data

- All Data was scraped from FBREF
- 110 Matches played
- Created 6 CSVs: 23/24 & 24/25
 - Havertz Individual Stats
 - Arsenal Team Stats
 - Arsenal schedule

				Playing Time				Performance								Expected				Progression						
Player	Nation	Pos	Age	MP	Starts	Min	90s	Gl	As	G+A	G-PK	PK	PKatt	CrdY	CrdR	xG	npG	xAG	npG+xAG	PrgC	PrgP	PrgR	Gl	As	G+A	G-PK
David Raya	ESP	GK	28	38	38	3,420	38.0	0	0	0	0	0	0	3	0	0.0	0.0	0.0	0.0	0	14	0	0.00	0.00	0.00	0.00
William Saliba	FRA	DF	23	35	35	3,039	33.8	2	0	2	2	0	0	2	1	2.3	2.3	0.9	3.1	16	138	6	0.06	0.00	0.06	0.06
Declan Rice	ENG	MF	25	35	33	2,825	31.4	4	7	11	4	0	0	7	1	3.5	3.5	6.6	10.1	90	192	94	0.13	0.22	0.35	0.13
Thomas Partey	GHA	MF,DF	31	35	31	2,797	31.1	4	2	6	4	0	0	4	0	2.3	2.3	2.1	4.3	36	185	48	0.13	0.06	0.19	0.13
Leandro Trossard	BEL	FW	29	38	28	2,546	28.3	8	7	15	8	0	0	4	1	7.2	7.2	6.1	13.3	80	101	226	0.28	0.25	0.53	0.28
Gabriel Magalhães	BRA	DF	26	28	28	2,363	26.3	3	1	4	3	0	0	4	0	2.6	2.6	0.9	3.5	10	126	11	0.11	0.04	0.15	0.11
Jurriën Timber	NED	DF	23	30	27	2,417	26.9	1	3	4	1	0	0	7	0	1.2	1.2	0.8	2.1	57	146	104	0.04	0.11	0.15	0.04
Martin Ødegaard	NOR	MF	25	30	26	2,325	25.8	3	8	11	2	1	1	4	0	4.8	4.0	5.4	9.5	92	258	154	0.12	0.31	0.43	0.08
Gabriel Martinelli	BRA	FW,MF	23	33	25	2,291	25.5	8	4	12	8	0	0	1	0	7.4	7.4	5.0	12.4	124	52	289	0.31	0.16	0.47	0.31
Kai Havertz	GER	FW,MF	25	23	21	1,874	20.8	9	3	12	9	0	0	5	0	9.5	9.5	2.4	12.0	35	61	113	0.43	0.14	0.58	0.43
Bukayo Saka	ENG	FW,MF	22	25	20	1,729	19.2	6	10	16	5	1	1	3	0	6.8	6.0	7.6	13.7	96	70	255	0.31	0.52	0.83	0.26
Mikel Merino	ESP	MF,FW	28	28	17	1,586	17.6	7	2	9	7	0	0	4	1	5.9	5.9	2.0	7.9	11	69	94	0.40	0.11	0.51	0.40
Myles Lewis-Skelly	ENG	DF	17	23	15	1,369	15.2	1	0	1	1	0	0	3	2	0.2	0.2	0.4	0.6	37	71	37	0.07	0.00	0.07	0.07
Ben White	ENG	DF	26	17	13	1,198	13.3	0	2	2	0	0	0	2	0	0.5	0.5	1.4	1.8	23	59	42	0.00	0.15	0.15	0.00
Riccardo Calafiori	ITA	DF	22	19	11	983	10.9	2	1	3	2	0	0	4	0	0.9	0.9	0.3	1.1	22	54	29	0.18	0.09	0.27	0.18
Ethan Nwaneri	ENG	FW,MF	17	26	11	895	9.9	4	2	6	4	0	0	1	0	1.2	1.2	1.2	2.4	46	33	113	0.40	0.20	0.60	0.40
Jakub Kiwior	POL	DF	24	17	10	1,122	12.5	1	0	1	1	0	0	1	0	0.2	0.2	0.3	0.5	5	48	1	0.08	0.00	0.08	0.08
Jorginho	ITA	MF	32	15	9	704	7.8	0	0	0	0	0	0	5	0	0.1	0.1	0.4	0.5	12	48	6	0.00	0.00	0.00	0.00
Raheem Sterling	ENG	FW,MF	29	17	7	499	5.5	0	2	2	0	0	0	1	0	1.4	1.4	0.5	1.9	27	18	79	0.00	0.36	0.36	0.00
Gabriel Jesus	BRA	FW	27	17	6	608	6.8	3	0	3	3	0	0	4	0	3.0	3.0	0.7	3.7	15	19	57	0.44	0.00	0.44	0.44
Oleksandr Zinchenko	UKR	DF,MF	27	15	5	527	5.9	0	1	1	0	0	0	1	0	0.6	0.6	0.2	0.8	10	48	14	0.00	0.17	0.17	0.00
Kieran Tierney	SCO	DF,FW	27	13	2	259	2.9	1	0	1	1	0	0	0	0	0.4	0.4	1.4	1.7	8	15	29	0.35	0.00	0.35	0.35
Nathan Butler-Oyediji	ENG	FW	21	1	0	7	0.1	0	0	0	0	0	0	0	0	0.1	0.1	0.0	0.1	0	0	0	0.00	0.00	0.00	0.00
Takehiro Tomiyasu	JPN	DF	25	1	0	7	0.1	0	0	0	0	0	0	0	0	0.2	0.2	0.0	0.2	0	1	0	0.00	0.00	0.00	0.00
Reiss Nelson	ENG	FW	24	1	0	3	0.0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	0	0	0.00	0.00	0.00	0.00

Key Variables for Understanding

Expected Goals(xG): estimates the likelihood that a shot will result in a goal, based on factors like shot location, shot type, assist type, and defensive pressure.

An xG value ranges from 0 to 1, where 1 means a goal is almost certain. It's used to assess the quality of chances and evaluate team or player performance beyond just goals scored.

Points: 3 points for a win, 1 for a draw, 0 for a loss

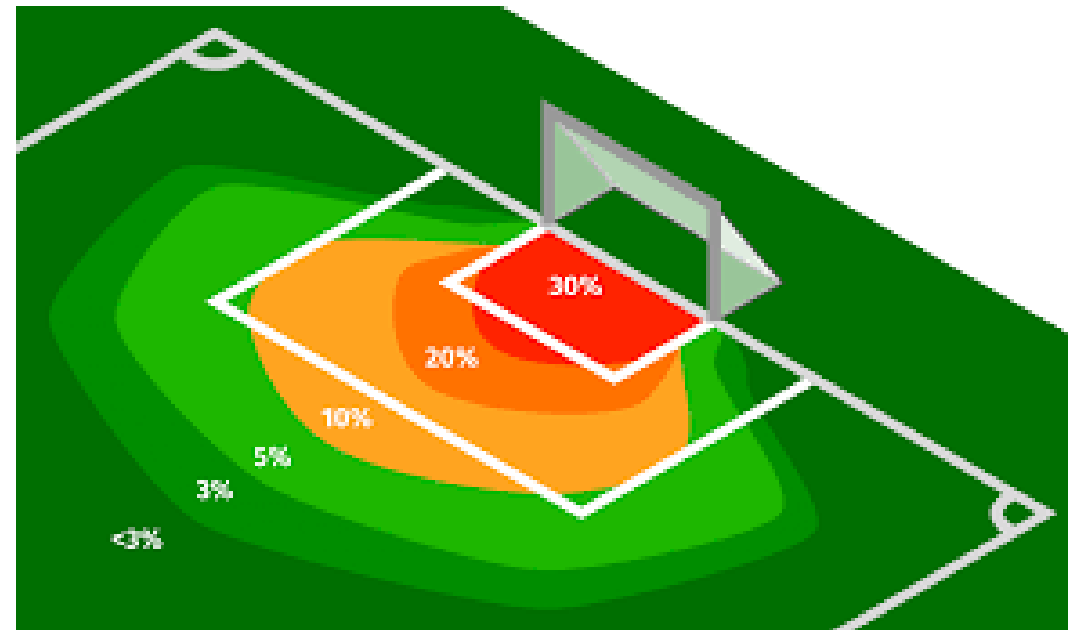
Win: indicator 0 or 1 for a win

HavertzPlayed: indicator 0 or 1 for if he played

Venue: Home/Away/Neutral

lnxG: log of xG

LnHav_xG: log of Havertz xG



Predictions on how Havertz shifts Arsenal's match outcomes(points)

I expect the points to be related to the following factors:

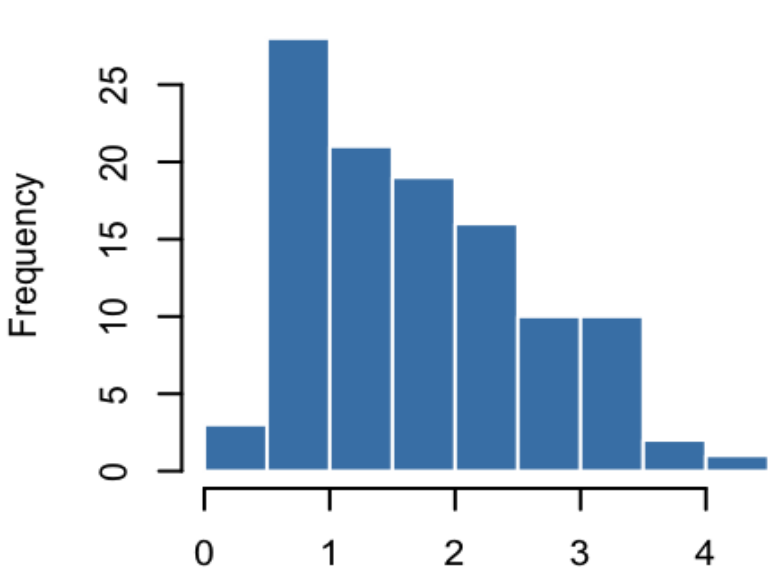
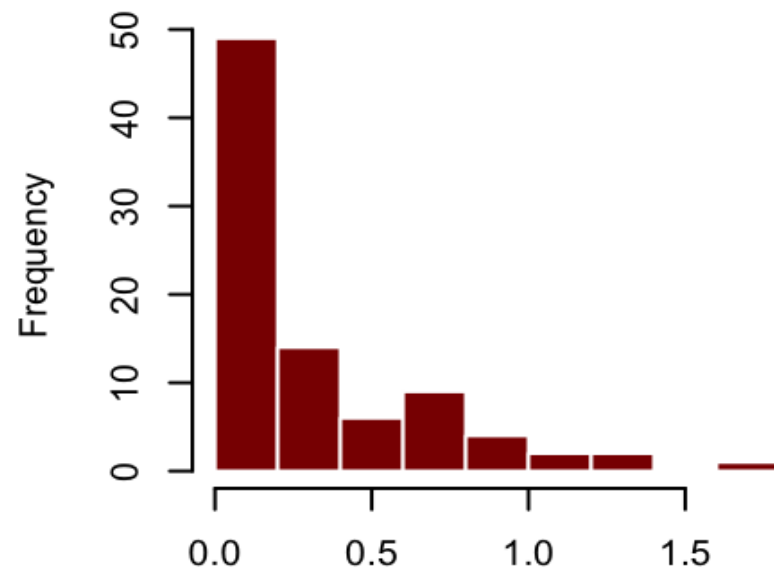
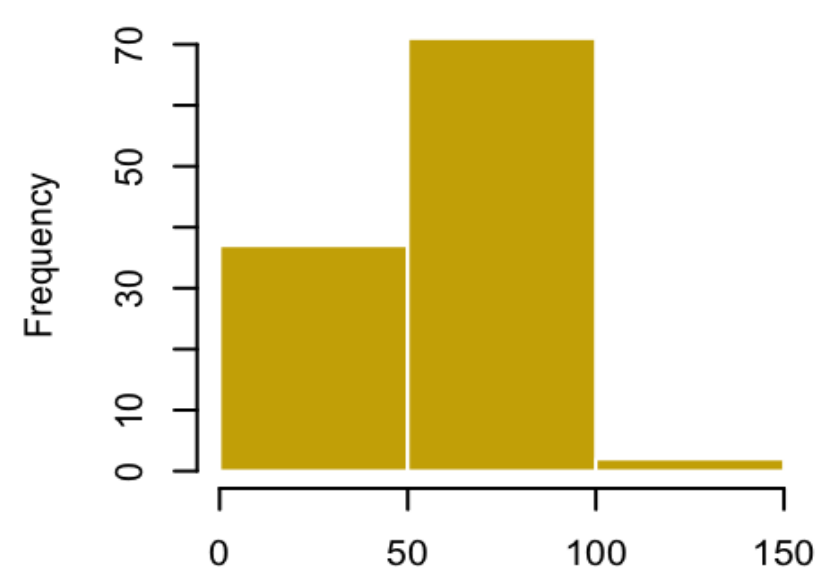
Positive effect when **Havertz plays**:

Positive relationship with **Havertz's individual expected goals**(lnHav_xG):

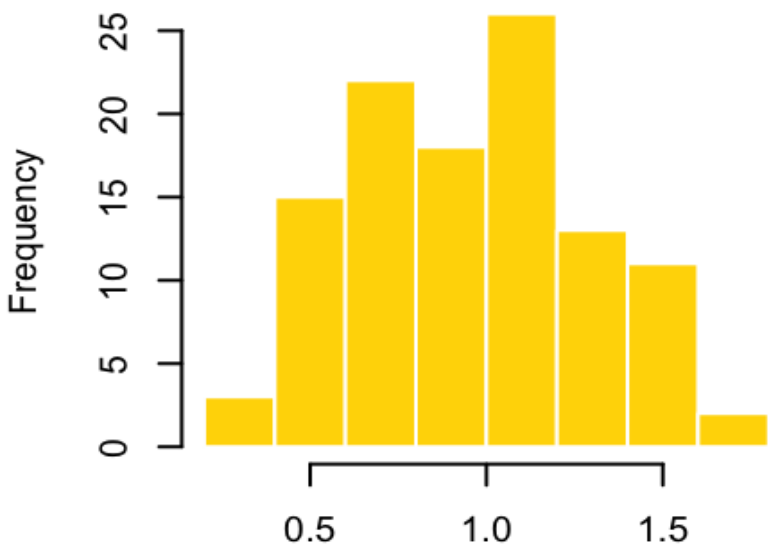
Positive relationship with **team expected goals** (lnxG):

Positive venue effect: **home** matches yield more points than away or neutral sites.

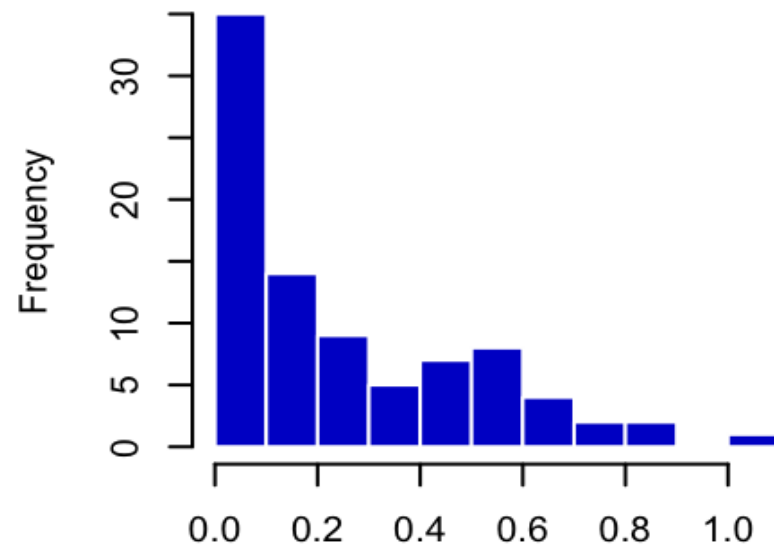
Exploratory: the value of Havertz's xG **diminishes** when team xG is already high (interaction term).

Distribution of xG**Havertz xG****Havertz Minutes**

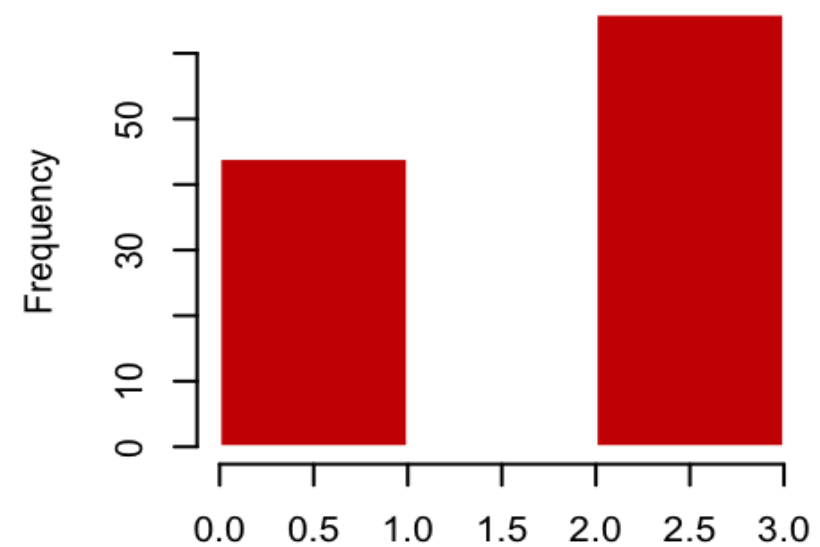
xG
 $\log(1 + xG)$



Hav_xG
 $\log(1 + Hav_xG)$



Minutes
Match Points



$\log1p(xG)$

$\log1p(Hav_xG)$

Points (0/1/3)

Regression Models

M0(Baseline): $\text{Points} = \beta_0 + \beta_1 \cdot \text{HavertzPlayed} + \beta_2 \cdot xG + \beta_3 \cdot \text{Venue} + \beta_4 \cdot \text{Season} + \varepsilon$

Model clean(Minutes): $\text{Points} = \beta_0 + \beta_1 \cdot \text{Hav_Min} + \beta_2 \cdot xG + \beta_3 \cdot \text{Venue} + \beta_4 \cdot \text{Season} + \varepsilon$

Model 1(Logged): $\text{Points} = \beta_0 + \beta_1 \cdot \text{Hav_Min} + \beta_2 \cdot \ln(1+xG) + \beta_3 \cdot \ln(1+\text{Hav_xG}) + \beta_4 \cdot \text{Venue} + \varepsilon$

Model Int(Interaction): $\text{Points} = \beta_0 + \beta_1 \cdot \text{Hav_Min} + \beta_2 \cdot \ln(1+xG) + \beta_3 \cdot \ln(1+\text{Hav_xG}) + \beta_4 \cdot \text{Venue} + \text{Season} + \varepsilon$

Predictions: *Points*

Positive effect when **Havertz plays**: **No clear positive effect**

M0: 0.44, Not Significant(P=0.16) MC: Almost 0, Not Significant(P=0.75) M1: -0.012, Not Significant(P=0.13)

Positive relationship with **Havertz's individual expected goals**: **No clear positive effect**

M1: 0.624, Not Significant(P=0.32) Mint: 4.12, Significant(P=0.058)

Positive relationship with **team expected goals**: **clear positive effect**

M0: 0.38, Significant(P=0.004) MC: 0.39, Significant(P=0.003) M1: 0.89, Significant(P=0.045) Mint: 1.39 Significant(0.011)

Positive venue effect: **home** matches yield more points than away or neutral sites: **No clear positive effect**

Home Games **positive** in all models but $p = 0.26-0.16$

Exploratory: the marginal value of Havertz's xG **diminishes** when team xG is already high (interaction term): **TRUE**

M_int: $\ln xG \times \ln Hav_xG = -2.71$ $p = 0.096$

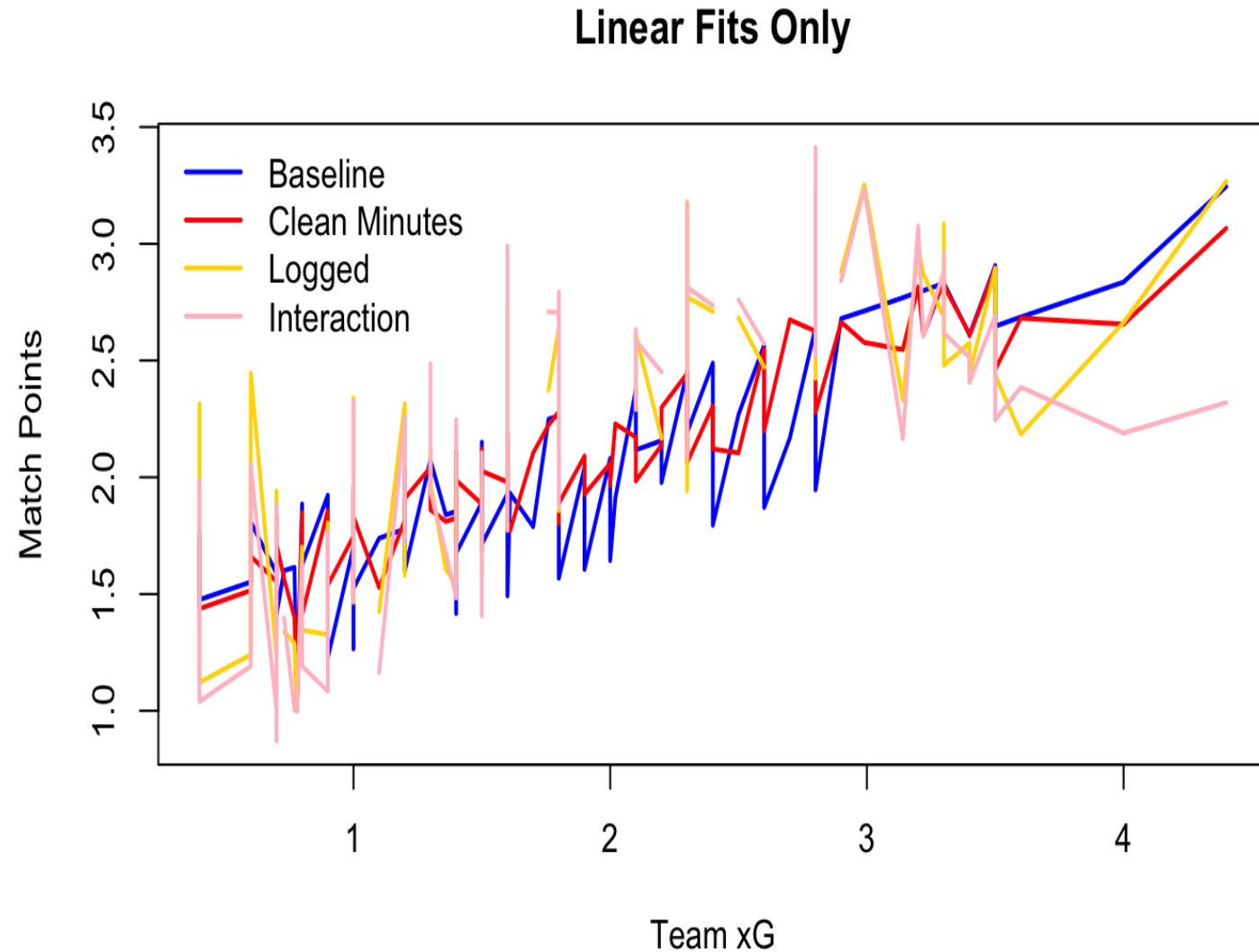
This interaction model suggests Havertz's personal xG converts to points most efficiently in low-xG games, and a bit less when the team is already creating plenty of chances.

Linear Regression Graph

Here are the four regression models plotted

It's clear that none of these OLS curves can bend to capture potential plateaus, spikes or thresholds in how xG translates to points.

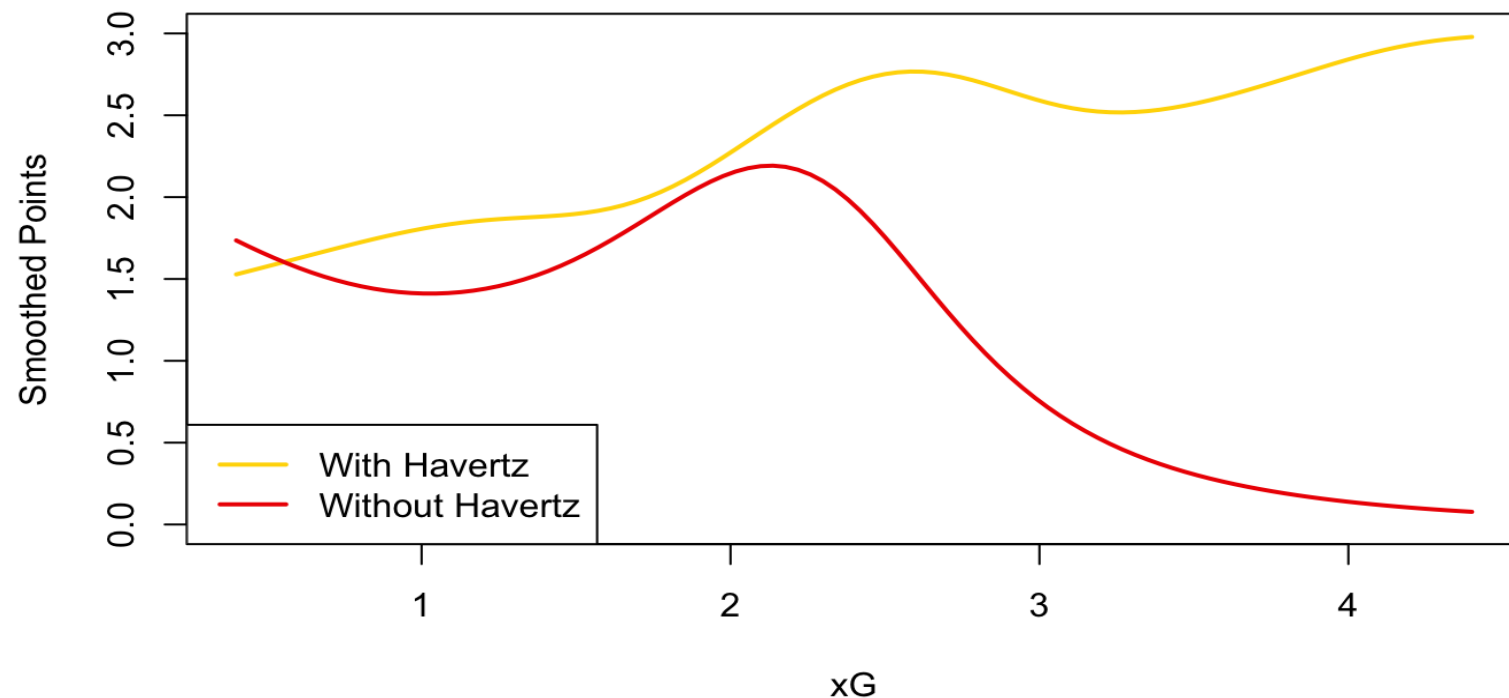
Next step: We need a more flexible fit. In the next slide we'll use a kernel regression.



Kernel Regression with C

- OLS/Linear Regression forces a straight line between our variables. Kernel Regression allows us to create a smooth curve to spot non-linear patterns
- I ran two smoothers and visually compare the xG->Points relationship for games **with** and **without** Havertz
- Smoother implemented in C for speed

Kernel Regression: Points vs xG

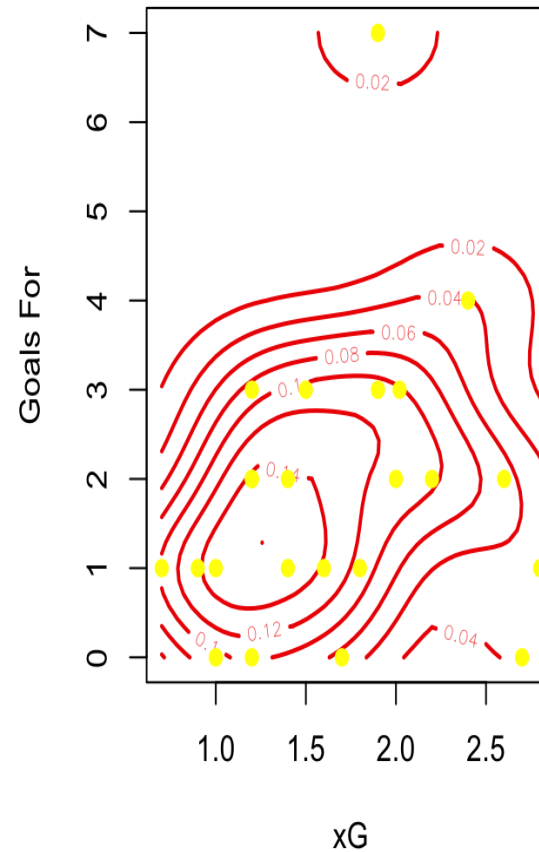


- We can see **with** Havertz points rise steadily with xG
- **Without** him we can see there is a very sharp drop after 2.3 xG which could imply that **without** him Arsenal generated good chances on goal but failed to convert them into wins.
- Havertz appears to help convert the team xG into actual xG. **Without** him we could still create high xG just couldn't win the game or score those goals

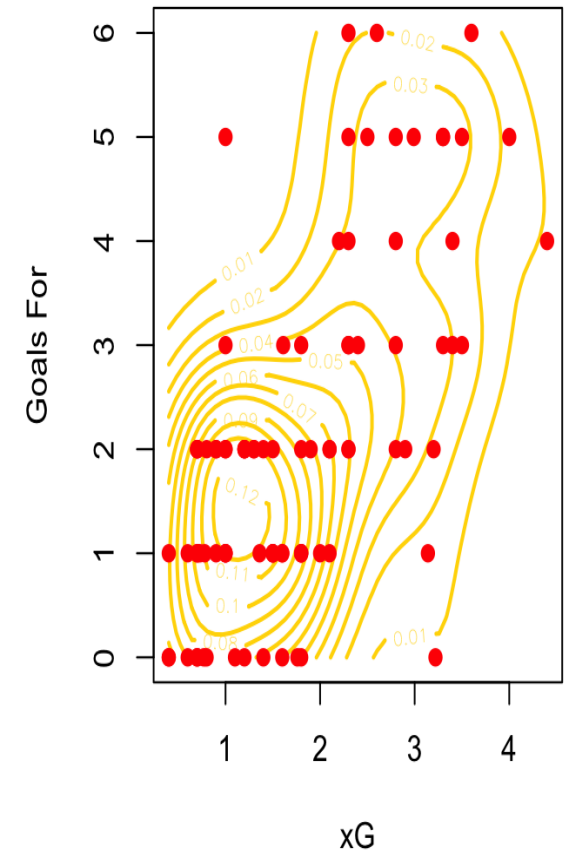
When Arsenal create a similar xG tally, do they turn those chances into goals more effectively with Havertz on the pitch?

- Joint Density of xG and GF in matches with and without Havertz
- The contour lines are regions of high match density
- The dots are actual matches
- We saw how points vary with xG now we look at the joint distribution of xG and actual goals and Havertz presence

2D KDE with Points (Without)



2D KDE with Points (With)



Without

- Without Havertz the cluster is capped around 3 goals and xG doesn't go above 2.5
- There are fewer high goal matches for a given xG

*note there are fewer matches played without Havertz, so this isn't conclusive

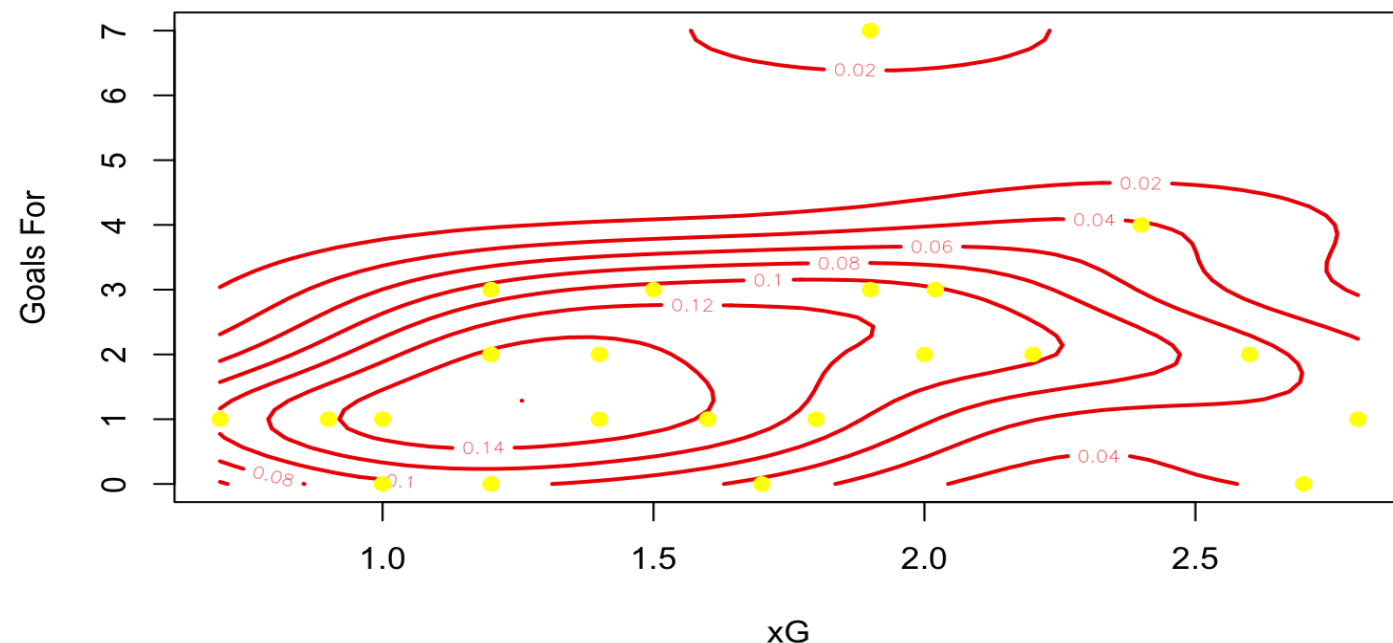
With

- With Havertz there is a greater density when xG is around 1/1.5 and the Goals are around 1-2
- However, there is a higher tail of 4 to 6 goals when the xG climbs, which shows there is higher ceiling outcomes

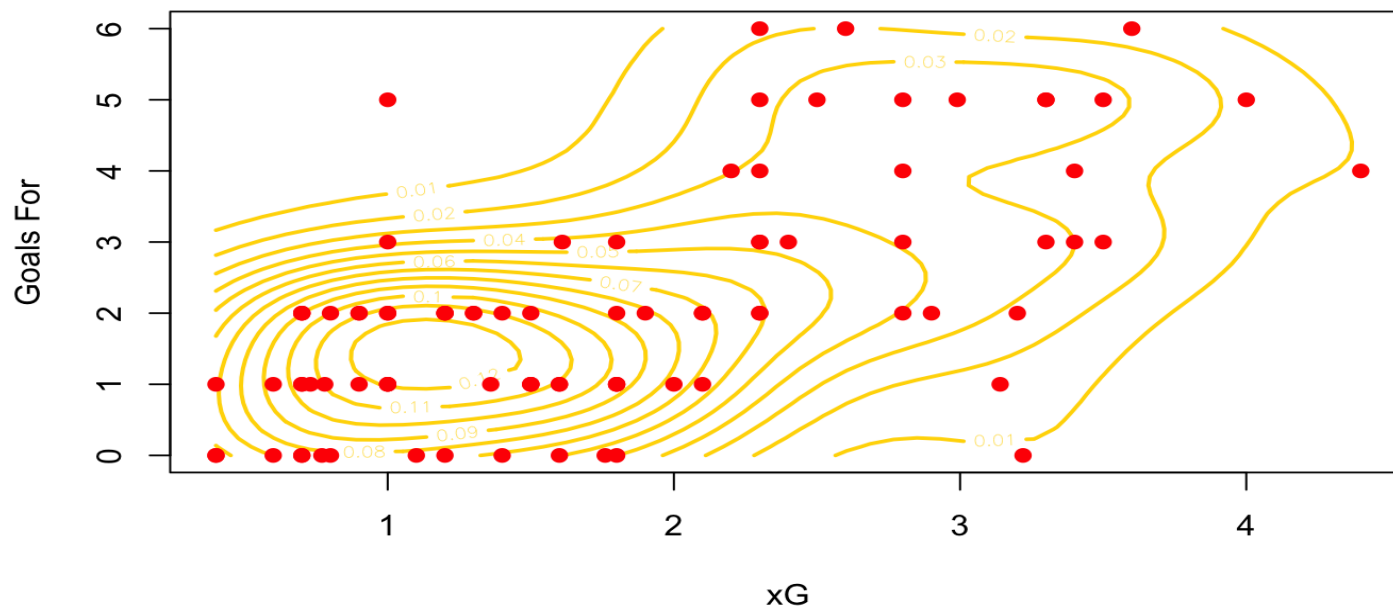
-My interpretation of this is that Havertz appearances coincides with a better conversion of xG into goals and a wider goal ceiling which we had already expected from our kernel regression

-TLDR: Havertz presence is associated with more efficient finishing and a higher upside

2D KDE with Points (Without Havertz)

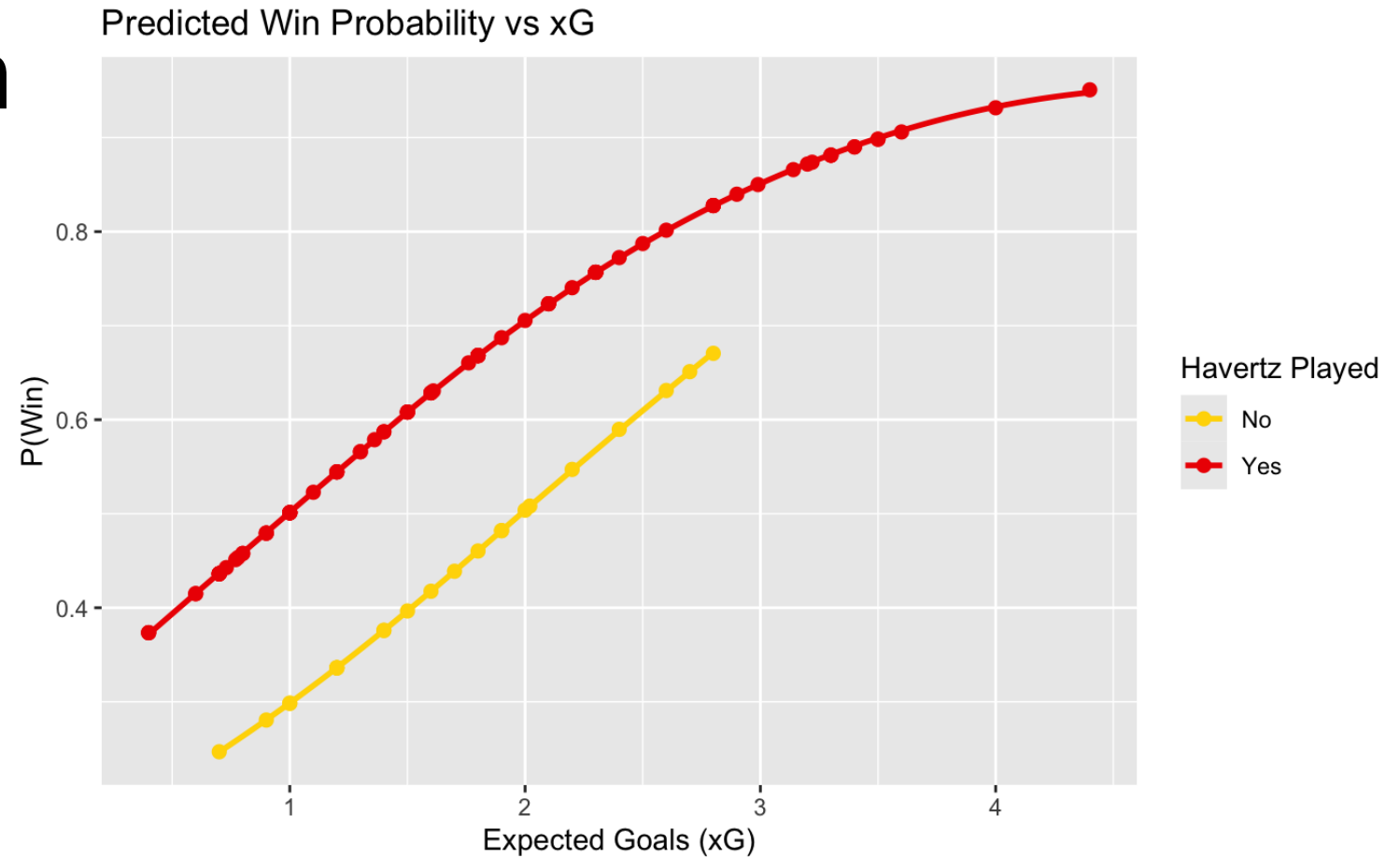


2D KDE with Points (With Havertz)



Logistic Regression

- This is where we use our new binary variable win
- Our Logit Model: $\text{Win} \sim \text{xG} + \text{HavertzPlayed}$ for estimating how xG and Havertz shift the chance of winning
- The coefficient on HavertzPlayed is 0.8589 and when exponentiated is an odds ratio of 2.36 which implies the odds of winning 2.4 times higher ($p=0.07$) with him and is significant at a 10% level (doubles the odds of Arsenal winning)
- Then we take the fitted model and the original data, predict a number between 0 and 1 for each row to guess a $P(\text{win})$
- Points per match predicted prob for each game, LOESS smoothing



- As team xG increases both with and without Havertz the odds of winning increases
- Gap between curves shows that for any team xG the win probability with Havertz is 10-15% higher

*Small no Havertz sample so this isn't significant yet

Random Forests

Which Havertz/possession/venue metrics matter most for predicting Win (1 or 0):

Random Forest:

-110 Matches and 500 trees

-OOB error is 31% meaning our forest correctly predicts a win around 69% of the time.

Variable importance scores using RF:

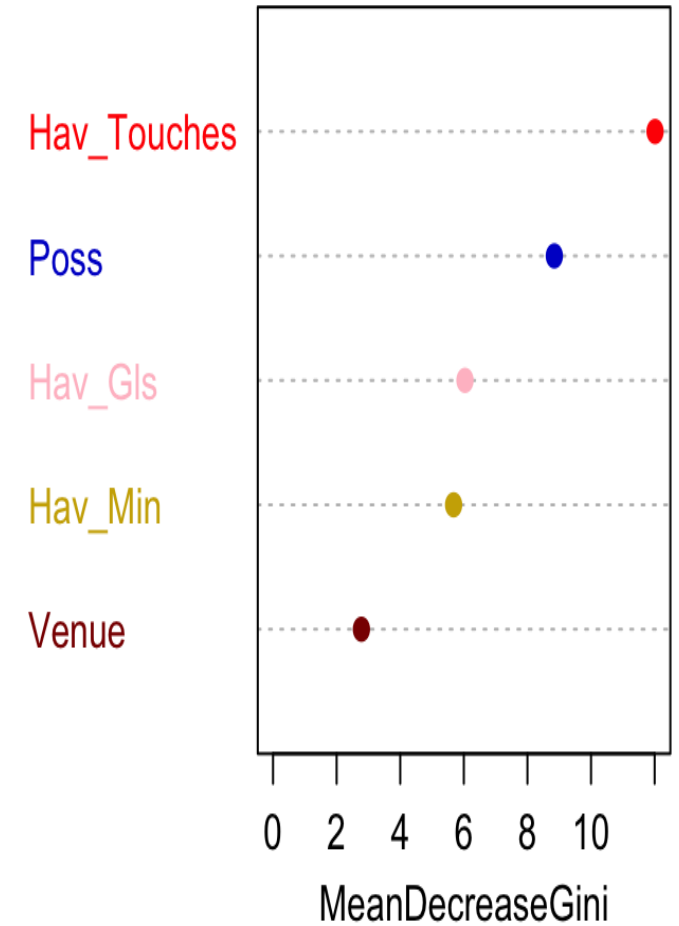
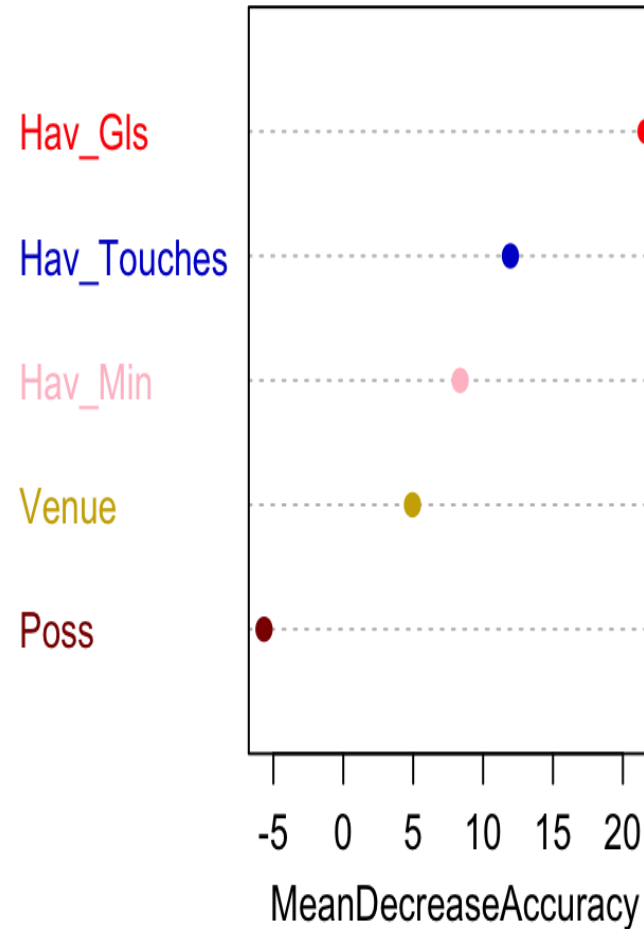
-The further right the variable is the more important it is to predicting correctly

-Havertz Goals increase win likelihood

-Touches and Minutes mean more wins

-Possession is very weak which is surprising

Top Predictors



Havertz Impact on Points:

Average Points per Match:

-With Havertz Arsenal averages **2.13** points per game

-Without Havertz Arsenal averages **1.65** points per game

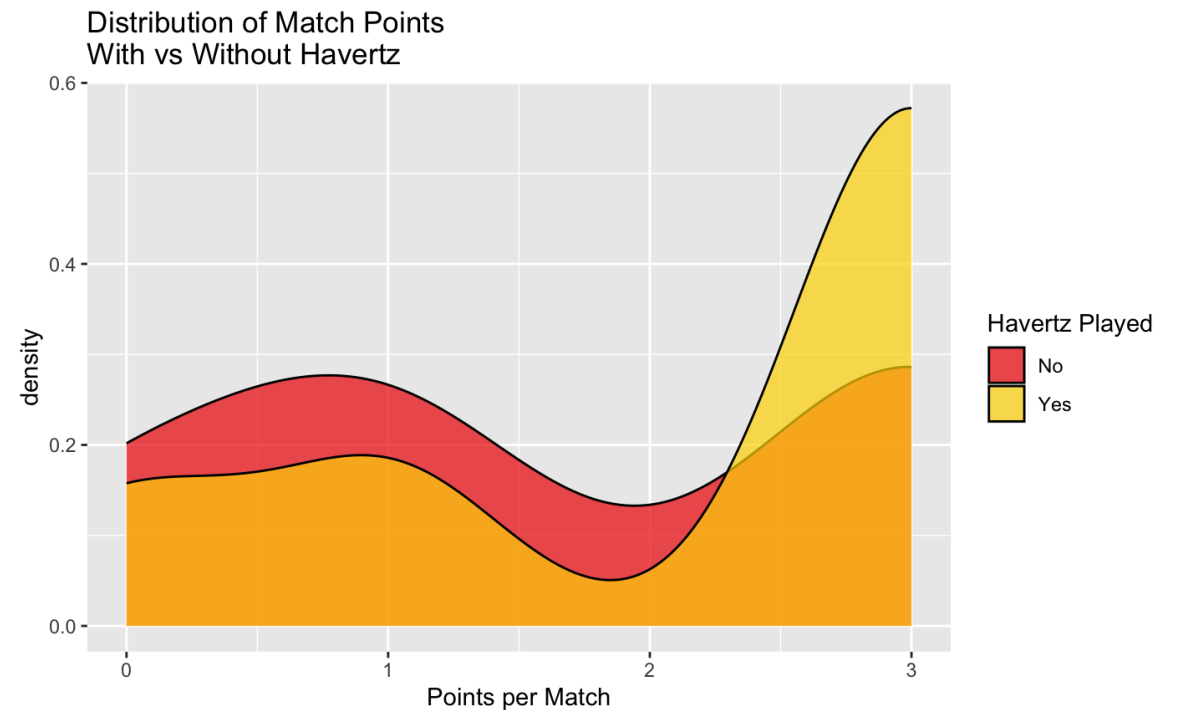
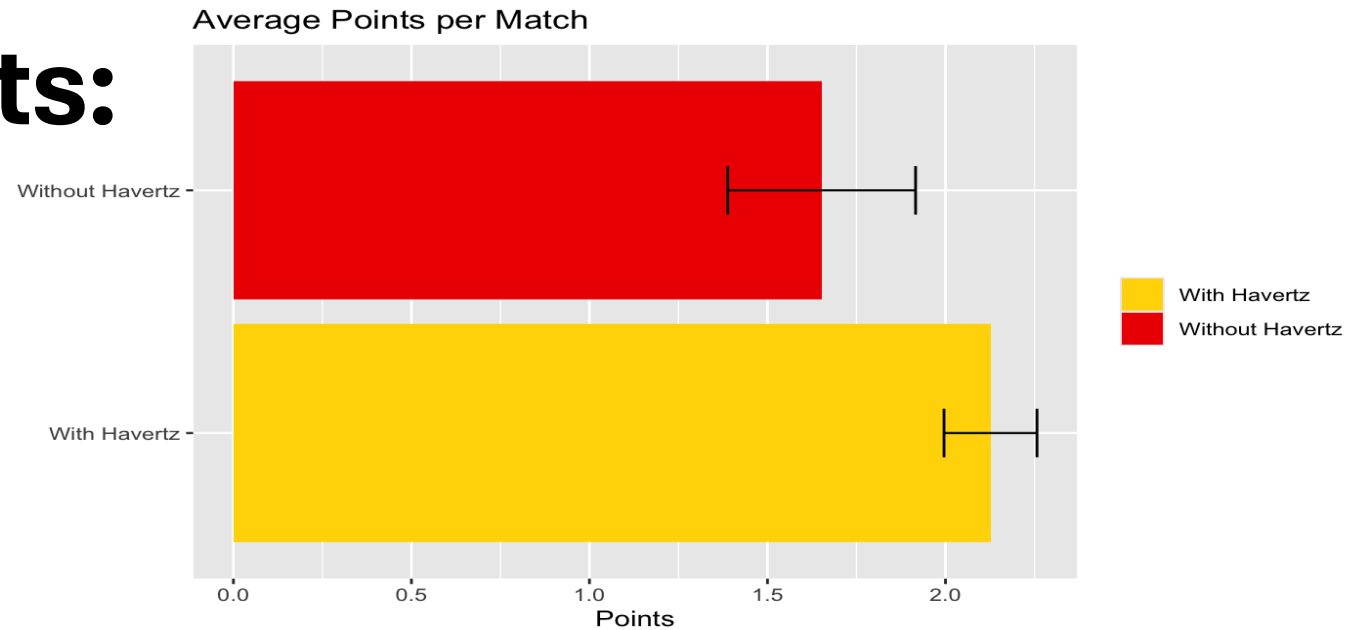
Distribution of Match Points:

-Curve with Havertz skewed towards wins(3pts)

-Without him there's a much larger mass at 0-1 pts

-Fewer low point games when he plays

-Kai Havertz is associated with Arsenal winning more games









Projected Season Total:

This season Arsenal finished 2nd with 74 points, Havertz missed 15 games where we earned 25 points.

Let's take those games times our point avg with Havertz:
 $15 \times 2.13 = 31.95$

$74 - 25 = 49 \rightarrow 49 + 31.95 = 81$ points....still 2nd place to Liverpool's 84 points.

Season 2024-25 ▾										
Club	MP	W	D	L	GF	GA	GD	Pts	Last 5	
1  Liverpool	38	25	9	4	86	41	45	84	✓	✗
2  Arsenal	38	20	14	4	69	34	35	74	✗	✗
3  Man City	38	21	8	9	72	44	28	71	✓	✓
4  Chelsea	38	20	9	9	64	43	21	69	✓	✓
5  Newcastle	38	20	6	12	68	47	21	66	✓	✗
6  Aston Villa	38	19	9	10	58	51	7	66	✗	✗



Conclusion and Key takeaways across every method

OLS: Linear baseline

- Team xG is the strongest linear driver of points. (Positive and Very Significant)
- Havertz playing alone is **not significant** once we control for xG, minutes and venue.

Kernel regression: Flexible fit

- With Havertz the xG → points curve keeps climbing; without him it stalls near 2.3 xG.
- Suggests he helps convert good chances into results.

2-D KDE: Shot quality conversion

- Joint density of xG and goals shows a larger “high-goal, high-xG” tail when he plays.
- Ceiling on goals is lower in matches he misses.

Logit: Match outcome likelihood

- Odds of winning are 2.4x higher with Havertz at any given team xG ($p = 0.07$)
- Win probability gap is roughly 10–15 percentage points across the xG range.

Random Forest: Non-linear interactions

- Havertz playing ranks as one of the top three predictors of points once non-linear splits are allowed.

Bottom line:

Every model finds the same story: Kai Havertz does not just appear in winning line-ups, he associated with tilting xG into goals and goals into points.

Limitations and future work

Data: Sample without Havertz is small so maybe focus more on minutes/what minutes he played

Need other context variables: What position/other in similar positions, opponent strength, other players around \$65 million

xG limitation: Relatively new idea and calculated by FBREF

Thank you!

