Summary

This project aimed to delve into the intricate details of football player performance using a comprehensive dataset comprising 28 columns and 421 rows. This dataset is from the English Premier League during the 2023 season. Employing advanced statistical models such as ANOVA, Linear Regression, Logistic Regression, Multi-Regression, Moderation, and Factor Analysis, we sought to answer questions related to player consistency, playtime allocation, overall contribution to the team, the prevalence of Player of the Match (POTM) awards and the effect of a player's position on his performance.

To ensure the robustness of our analysis, we strategically excluded Goal Keepers (GK) from the dataset and removed clean sheets and goals conceded, focusing specifically on outfield players. This decision was rooted in the recognition that it is very rare for keepers to score be actively involved in their team scoring a goal. The other main aspect to consider is the gameplay of defenders who in recent seasons, have displayed significant contributions to goals and assists, making their inclusion essential for a comprehensive understanding of player capabilities.

Key Findings

Player Position: Our analysis allowed us to identify players who consistently delivered strong performances. By employing statistical tools such as ANOVA, we were able to pinpoint patterns of consistency across different players, shedding light on those who reliably contribute to their team's success.

Playtime Allocation: Utilizing Linear Regression, we investigated the factors influencing playtime allocation for individual players. This analysis provided insights into the variables that coaches consider when deciding the amount of playing time a player receives, offering valuable information for team management strategies.

Contribution to Team: Performing analysis enabled us to quantify and understand the multifaceted contributions of players to their teams. By examining various factors simultaneously, we gained a nuanced understanding of the distinct roles players play in team dynamics.

Player of the Match Awards: Logistic Regression was employed to analyse the likelihood of players winning the prestigious Player of the Match (POTM) awards. This allowed us to identify key performance indicators that significantly influence a player's probability of receiving this accolade.

In conclusion, our project not only provides a comprehensive overview of player performance but also offers actionable insights for coaches, analysts, and team managers to make informed decisions. The exclusion of certain metrics highlights the evolving nature of football analysis, prompting future researchers to explore additional dimensions of player contributions.

Dataset

S No.	Column	Description
1	Nation	Nationality of the player
2	Local/Foreigner	Whether the player is an England Player or not
3	Team	Team the player plays in
4	Position	Position the player plays in(Attack/Midfield/Defence)
5	Age	Age of the player
6	MP	Number of matches played
7	Starts	Number of matches the player has started in
8	Min	Number of minutes played by the player
9	90s	
10	Gls	Number of goals scored by the player
11	Ast	Number of Assists scored by the player
12	G+A	Number of Goals plus Assists scored by the player
13	G-PK	
14	PKMade	
15	PKAttemp	
16	CrdY	Number of Yellow cards received by the player
17	CrdR	Number of Red cards received by the player
18	Suspension	Number of time the player has been suspended.
19	Gls/90	Number of goals scored per 90 minutes by the player
20	Ast/90	Number of Assists scored per 90 minutes by the player
21	G+A/90	Number of Goals and Assists scored per 90 minutes by the player
22	G-PK/90	
23	Gls/GM	Number of goals scored per game by the player
24	Ast/GM	Number of Assists scored per game by the player
25	G+A/GM	Number of goals plus assists scored per game by the player
26	G-PK/GM	
27	POTM	Binary column which shows 1 if a player has won “Player of the Match” award at least once in the entire season.

Statistical Analysis

Correlation Analysis

Here we would like to find the relationship between variables.
Correlation quantifies the strength and direction of the linear relationship between the two variables.

In our project, we would like to quantify the relationship between MP (Matches Played) and 90s (Number of 90 minutes a player has played across the entire season)

Upon performing correlation for one of the players, we end up with a Fit model with Strong Positive Correlation.
Here the "r" value is +0.88

ANOVA

Here we use the statistical method when we have three or more categories in the independent variable.
We compare means of the groups at hand.
In this case we compare the number of "Assists" made by players of different positions.
We have 3 groups at hand namely Attackers, Defenders and Midfielders.

The hypotheses look like this

H0: Assists of Attackers = Assists of Defenders =Assists of Midfielders (all group means are equal).
H1: Not all means are equal

Upon performing ANOVA analysis, we observe that the mean of number of assists made by Defenders varies with that of mean number of assists made by Attackers and Midfielders.

Further analysis shows that there is not much difference between mean of number of assists made by Attackers and Midfielders.

Simple Linear Regression

We need to numerically summarize by fitting a regression line to the data.
The regression model, or line, is represented as

$$\hat{y} = b_0 + b_1 x$$

This is the fitted model.
b_0 is the estimated intercept and b_1 the estimated slope.

In our project we used Linear Regression to find out the influence each team had on the number of Goals + Assists made. This helps us better understand which team has more impact or effect in creating or scoring goals.

For this we used “Team” and “G+A” columns from the dataset.

Independent Variable: Team
Dependent Variable: G+A(influence)

Upon performing Linear regression, we observe that the “P” value for few teams is less than alpha. These teams are the ones that we consider as significant teams.

Significant Teams:

Arsenal
Brighton
Liverpool
Manchester City
Newcastle United
Tottenham

Multiple Regression

Very often, we will want to have more than one variable predicting y. That is, we may want to see how x1, x2, x3 etc. predict y.
In such cases, where we have multiple independent variables predicting y, we conduct a multiple regression analysis.

Population Regression Model: $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$

From our dataset, we tried to find out how the independent variables like position a player plays in, the number of minutes a player plays, the number of starts he gets might impact the number of goals he scores.

Upon performing the Regression Analysis, we can observe that for the independent variables we selected, we have a good model.
However, the Variance Inflation Factor of “Minutes” and “Starts” is way too high.
This shows that there is “Multicollinearity”

Multicollinearity occurs when the independent variables X1, X2,..., Xm are intercorrelated instead of being independent.

For analysis purpose, we do the regression analysis by considering only the number of starts and the position a player plays in as the independent variables that effects the number of goals scored by him.

Upon performing the Regression Analysis, we see that we have a significant model.

Dependent Variable: Goals Scored
Independent Variables: Position, Starts

Population Regression Model
Goals = -0.19 + 0.16Starts + 3.62 Pos FW – 2.2 Pos DF
(Midfielder is considered as zero Group)

From the residual distribution graph, we can see that the model follows “Normal Distribution.”
Another factor to consider here is that we might have few outliers. This is because of those players, who might have started almost all the games but ended up scoring very less goals, or even none.

Logistic Regression

Logistic Regression is a statistical method for analysing a dataset in which the outcome is a Binary variable.

From the dataset we choose POTM, Goals, Assist as the variables to perform Logistic regression.
We try to find how scoring goals or providing assists has impact on a player winning POTM.

Dependent Variable: POTM
Independent Variable: Goals, Assist
Point of Interest: 1 (Awarded)

Interpretation of Odds Estimate:

Per unit increase in Goals, the chances of a player winning the POTM increases by 47.3%.
Per unit increase in Assists, the chances of a player winning POTM increases by 28.4%.

Factor Analysis

Factor analysis examines the interrelationships among many variables and, then, attempts to explain them in terms of their common underlying dimensions.

For Factor Analysis, we considered the following variables:
Matches Played, Starts, Minutes, 90s, GLs, G-PK, G-PK/90, Gls/90, Gls/GM, G-PK/GM, Ast, Ast/90, Ast/90.

Upon performing analysis, we end up with 3 emerging factors to which each of the variables are assigned to.
Note: Oblique rotation is used.

Following are the factors assigned to
Factor 1: GLs, G-PK, G-PK/90, Gls/90, Gls/GM, G-PK/GM
Factor 2: Matches Played, Starts, Minutes, 90s
Factor 3: Ast, Ast/90, Ast/90.

Naming each Factor.
Factor1: Goals Info
Factor2: Playing Time
Factor3: Assists Info

Chi-Square Test of Independence

We have two categorical variables at hand and want to explore if there is a relationship between them.

Hypothesis
H0: Variables are independent
H1: Variables are dependent

Here we would like to find out if there is a relationship between a player being suspended and him being a local or foreign player.

Upon performing the Chi-Square Test of Independence, we end up with a P value which is less than alpha.

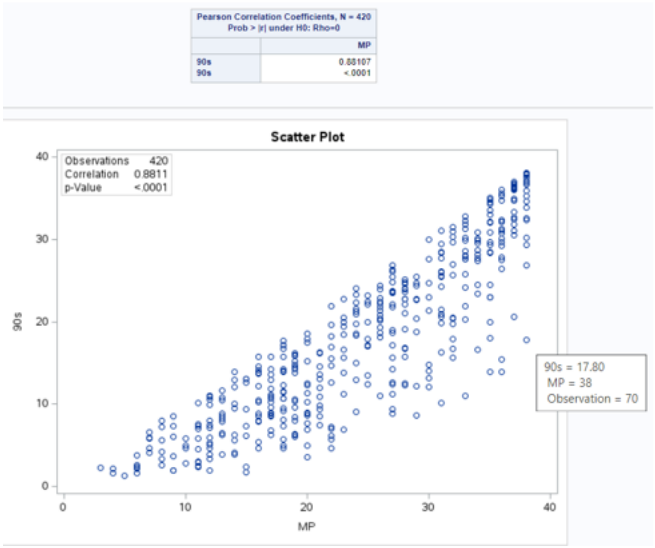
Hence we reject the Null Hypothesis and conclude that there is a dependency between a player being a local/foreigner and him being suspended.

Appendix

Results

1) **Correlation Analysis**

MP v/s 90s with observation



Strong Positive Relation
Fit model since P-value is less than alpha
R-value – + 0.88

2) ANOVA

Levene's Test for Homogeneity of Ast Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Pos	2	1959.5	979.7	5.05	0.0066
Error	417	80944.1	194.1		

Welch's ANOVA for Ast

Source	DF	F Value	Pr > F
Pos	2.0000	23.15	<.0001
Error	172.8		

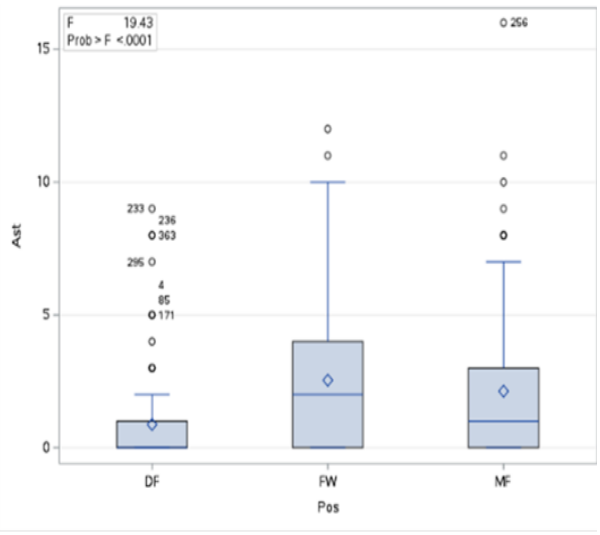
Level of Pos	N	Ast Mean	Ast Std Dev
DF	162	0.87654321	1.53536138
FW	72	2.54166667	2.80310442
MF	186	2.12903226	2.49654360

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Pos	Ast LSMEAN	LSMEAN Number
DF	0.87654321	1
FW	2.54166667	2
MF	2.12903226	3

Least Squares Means for effect Pos
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Ast

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		0.3808
3	<.0001	0.3808	



Distribution of assists by positions

- **Independent variable:** POS(position)
- **Dependent variable :** Ast(assists)

3) Linear Regression

Least Squares Model (No Selection)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	1093.06632	57.52981	2.00	0.0079
Error	400	11524	28.80924		
Corrected Total	419	12617			

Root MSE	5.36742
Dependent Mean	4.19048
R-Square	0.0866
Adj R-Sq	0.0433
AIC	1853.00051
AICC	1855.32212
SBC	1511.80560

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.739130	1.119185	1.55	0.1210
Team Arsenal	1	4.968142	1.600652	3.12	0.0020
Team Aston Villa	1	2.681922	1.663987	1.61	0.1078
Team Bournemouth	1	1.471396	1.663987	0.88	0.3771
Team Brentford	1	3.050343	1.663987	1.83	0.0675
Team Brighton	1	3.397233	1.600652	2.12	0.0344
Team Chelsea	1	0.722408	1.536432	0.47	0.6385
Team Crystal Palace	1	2.202046	1.716752	1.28	0.2003
Team Everton	1	0.879917	1.620014	0.54	0.5873
Team Fulham	1	2.734554	1.663987	1.64	0.1011
Team Leeds United	1	1.565217	1.582767	0.99	0.3233
Team Leicester City	1	1.802536	1.566193	1.15	0.2505
Team Liverpool	1	4.079051	1.600652	2.55	0.0112
Team Manchester City	1	6.681922	1.663987	4.02	<.0001
Team Manchester United	1	2.715415	1.600652	1.70	0.0906
Team Newcastle United	1	4.496184	1.716752	2.62	0.0092
Team Nottignham	1	0.782609	1.582767	0.49	0.6213
Team Southampton	1	1.165631	1.620014	0.72	0.4722
Team Tottenham	1	3.442688	1.600652	2.15	0.0321
Team West Ham United	1	1.560870	1.641048	0.95	0.3421
Team Wolves	0	0	.	.	.

4) Multiple Regression

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2666.74836	888.91612	95.75	<.0001
Error	416	3861.96355	9.28357		
Corrected Total	419	6528.71190			

Dependent Variable: Goals Scored

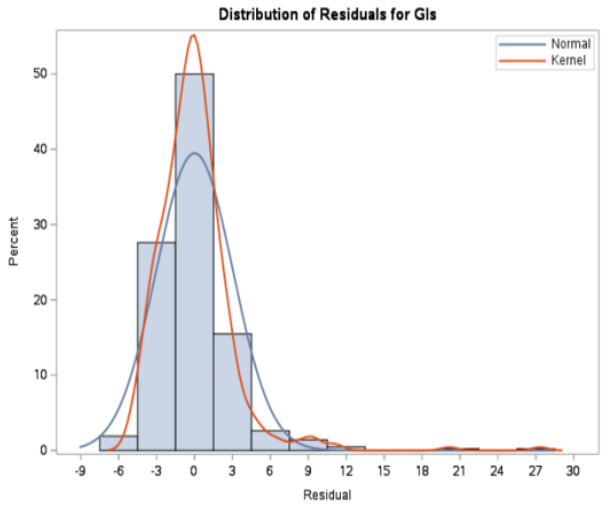
Independent Variables: Position, Starts

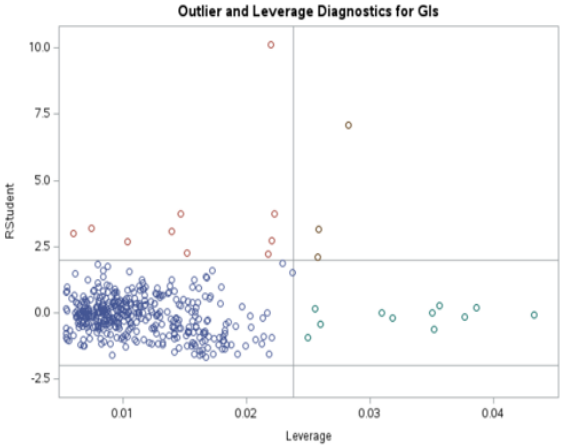
Population Regression Model

Goals = -0.19 + 0.16Starts + 3.62 Pos FW – 2.2 Pos DF

Root MSE	3.04689
Dependent Mean	2.47381
R-Square	0.4085
Adj R-Sq	0.4042
AIC	1361.84405
AICC	1361.98898
SBC	956.00507

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	B	-0.19512	0.33376	-0.58	0.5591	0	0
Pos DF	Pos DF	B	-2.23909	0.32795	-6.83	<.0001	-0.27644	1.15288
Pos FW	Pos FW	B	3.62176	0.42413	8.54	<.0001	0.34621	1.15599
Pos MF	Pos MF	0	0	.	.	.	.	.
Starts	Starts	1	0.16332	0.01399	11.67	<.0001	0.44313	1.01377





5) Binary Logistic Regression

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3028	0.2024	129.4947	<.0001
Gls	1	0.3875	0.0582	44.3428	<.0001
Ast	1	0.2496	0.0644	15.0086	0.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gls	1.473	1.315	1.651
Ast	1.284	1.131	1.456

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	82.8	Somers' D	0.686
Percent Discordant	14.2	Gamma	0.707
Percent Tied	3.0	Tau-a	0.287
Pairs	36875	c	0.843

6) Moderation

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8.36843	4.18422	295.10	<.0001
Error	417	5.91271	0.01418		
Corrected Total	419	14.28114			

Root MSE	0.11906	R-Square	0.5860
Dependent Mean	0.14038	Adj R-Sq	0.5840
Coeff Var	84.82364		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.04838	0.00701	6.90	<.0001
Gls	Gls	1	0.03923	0.00199	19.75	<.0001
PKMade	PKMade	1	-0.02862	0.01043	-2.75	0.0063

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.47974	2.82658	202.66	<.0001
Error	416	5.80140	0.01395		
Corrected Total	419	14.28114			

Root MSE	0.11809
Dependent Mean	0.14038
R-Square	0.5938
Adj R-Sq	0.5908
AIC	-1368.50503
AICC	-1368.36010
SBC	-1774.34401

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.043093	0.007202	5.98	<.0001
Gls	1	0.041183	0.002088	19.73	<.0001
PKMade	1	0.001136	0.014761	0.08	0.9387
Gls*PKMade	1	-0.001979	0.000700	-2.83	0.0050

7) Chi-Square Test of Independence

Local vs Foreigner Suspension

Observed Values

	Foreign	Local	
Suspended	22	5	125
Not Suspended	272	121	295
	294	126	420

Expected Values

	Foreign	Local
Suspended	87.5	37.5
Not Suspended	206.5	88.5

ChiSQ value

	Foreign	Local
Suspended	49.03143	28.16667
Not Suspended	20.77603	11.93503

Sum of ChiSq	109.9092
P-value	1.03E-25

8) Factor Analysis

Determining Factors

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

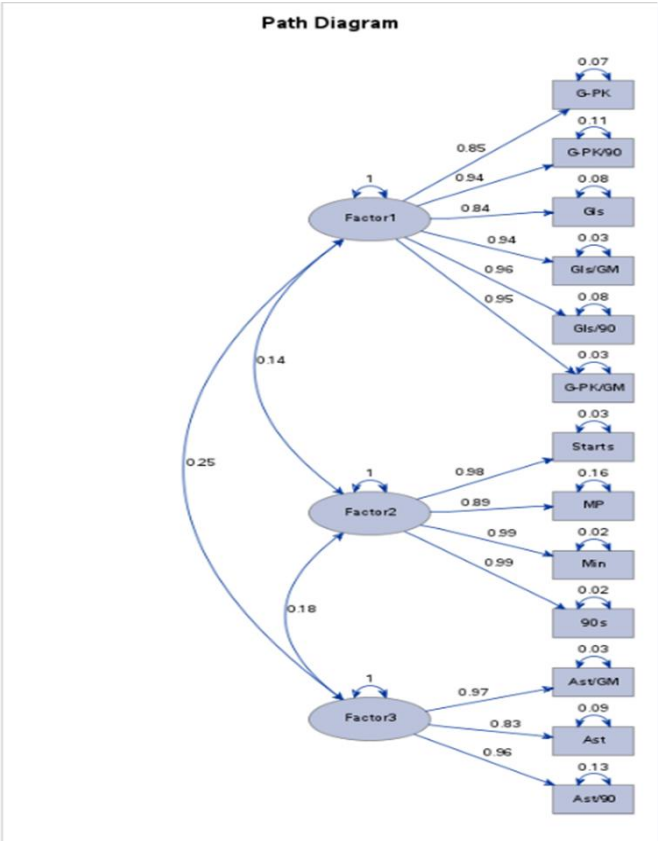
Eigenvalues of the Correlation Matrix: Total = 13 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	6.56430850	3.02270113	0.5049	0.5049
2	3.54160737	1.51720598	0.2724	0.7774
3	2.02440139	1.65839429	0.1557	0.9331
4	0.36600710	0.14328380	0.0282	0.9613
5	0.22272330	0.04668686	0.0171	0.9784
6	0.17603644	0.13548305	0.0135	0.9919
7	0.04055340	0.00715505	0.0031	0.9950
8	0.03339835	0.01198491	0.0026	0.9976
9	0.02141344	0.01342278	0.0016	0.9993
10	0.00799066	0.00662207	0.0006	0.9999
11	0.00136859	0.00118135	0.0001	1.0000
12	0.00018724	0.00018300	0.0000	1.0000
13	0.00000424		0.0000	1.0000

Determining Variables Groups

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3
MP	MP	0.04404	0.89257	0.07646
Starts	Starts	-0.01050	0.98146	0.03241
Min	Min	-0.01168	0.98957	0.02332
90s	90s	-0.01143	0.98953	0.02324
Gls	Gls	0.84380	0.26318	0.12137
G-PK	G-PK	0.84746	0.26035	0.13406
G-PK/90	G-PK/90	0.93810	-0.18423	0.04679
Gls/90	Gls/90	0.95837	-0.15752	0.04067
Gls/GM	Gls/GM	0.94499	0.08858	0.07658
G-PK/GM	G-PK/GM	0.94769	0.06897	0.08643
Ast	Ast	0.06811	0.29763	0.83165
Ast/90	Ast/90	-0.07625	-0.25099	0.96053
Ast/GM	Ast/GM	0.00215	0.06481	0.97033



References

https://www.premierleague.com/stats