Data Science in football:
Player performance Analysis

Group 3

Table of contents

Main idea

- ***Objectives:***
  - Analyze the performance of football players based on data collected from websites
  - Identify relationships between data variables and determine the factors that affect player performance.
  - Build a model to predict player performance.

- ***Reason for choosing the Topic:***
  - Football is a popular sport around the world.
  - Player performance is a key factor that determines the success of a team.
  - Data analysis can help to better understand player performance and develop strategies to improve player performance.

- ***Overview:***
  - Data was collected from websites of professional football leagues.
  - Data includes information about player name, season, team name, player performance metrics, goal-related metrics and injury data.

Table of contents

Team members

Name	Student ID	Role
Huynh Duc Thien	21127693	• Team leader • Data preprocessing • Data modelling
Bui Vu The Minh	21127107	• Data collecting • Data exploration • Data modeling
Le Phuoc Thinh Tien	21127700	• Data collecting • Data modelling • Report
Pham Khanh Toan	21127704	• Data exploration • Data modelling • Report

Working space

- ***Github: Source code repository***
  - Easy access and share source code.
  - Track the history of code changes.
  - Create branches and versions of the code.
- ***Google Meet: Weekly group meetings***
  - High-quality video and audio calls
  - Sharing screens and documents
- ***Trello: Task management, assignment***
  - Create cards, lists, and boards
  - Track task progress
  - Assign tasks to team members
- ***Messenger: Communication,***
  - Send text messengers, images, videos.
  - Create chat groups.

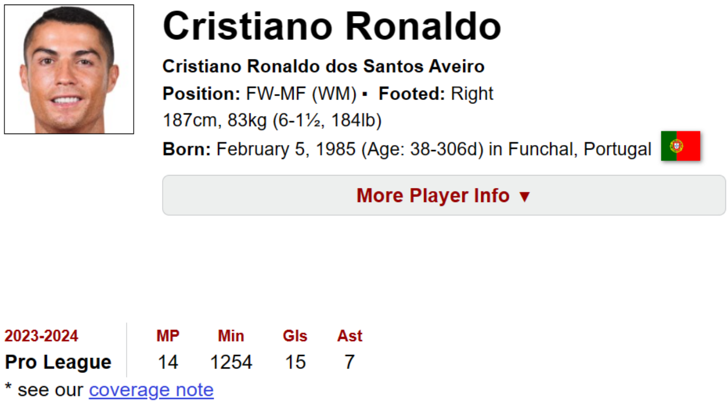
Google Meet

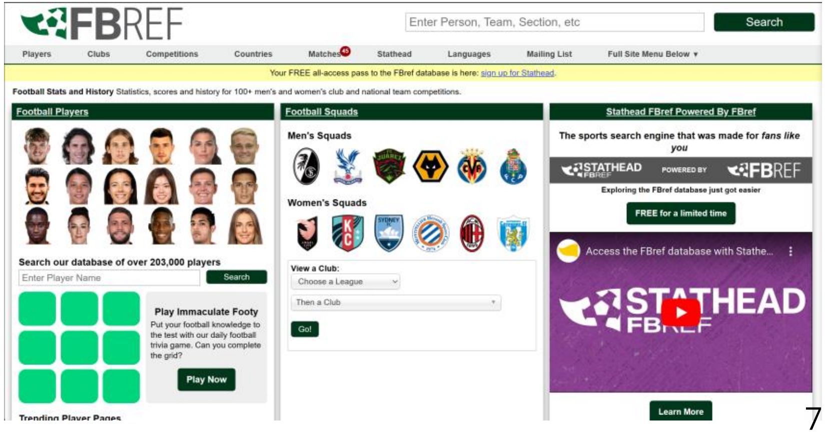
Trello

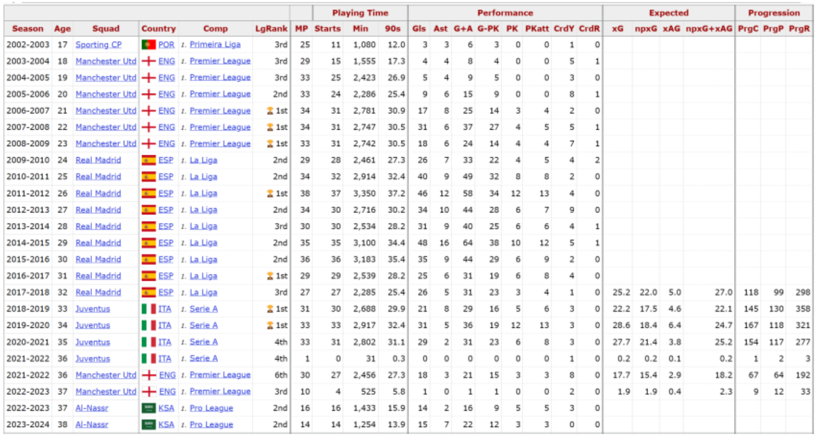
Table of contents

Data source: FBref

- ***Fbref website:*** *Football Statistics and History | FBref.com*
- ***Reason for choosing:*** The website provides a variety of football data, from general to detailed, from teams to individual players.
- ***Data collected:*** Standard statistics for the latest 10 seasons for each player in the Premier League 2023/2024
- ***Data collection method:*** Sending requests to the website's server and parsing HTML text.



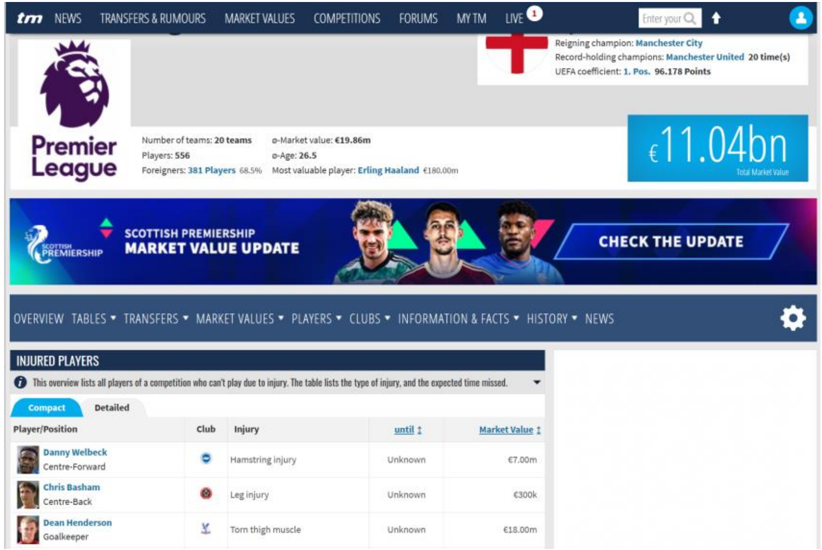




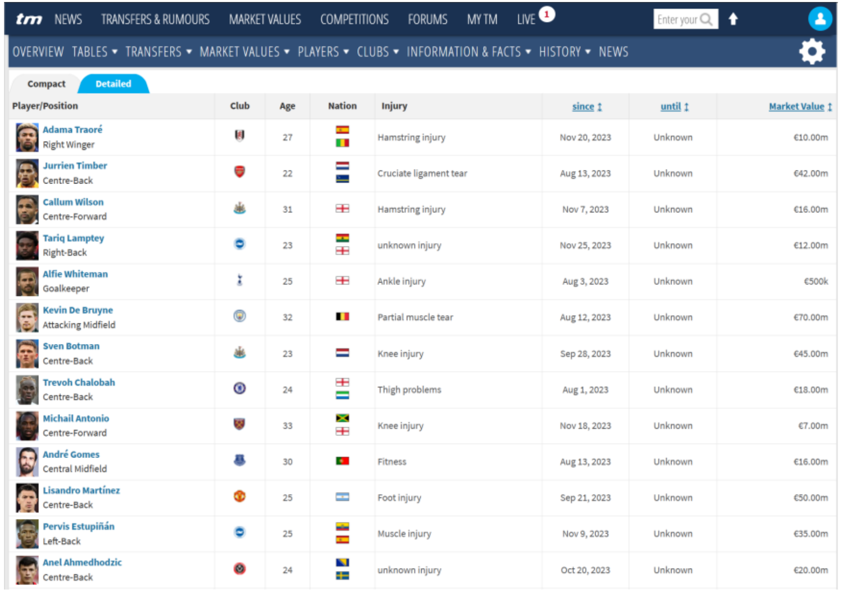
							Playing Time				Performance								Expected				Progression		
Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s	Gls	Ast	G+A	G-PK	PK	PKatt	CrdY	CrdR	xG	npxG	xAG	npxG+xAG	PrgC	PrgP	PrgR
2002-2003	17	Sporting CP	POR	1. Primeira Liga	3rd	25	11	1,080	12.0	3	3	6	3	0	0	1	0							
2003-2004	18	Manchester Utd	ENG	1. Premier League	3rd	29	15	1,555	17.3	4	4	8	4	0	0	5	1							
2004-2005	19	Manchester Utd	ENG	1. Premier League	3rd	33	25	2,423	26.9	5	4	9	5	0	0	3	0							
2005-2006	20	Manchester Utd	ENG	1. Premier League	2nd	33	24	2,286	25.4	9	6	15	9	0	0	8	1							
2006-2007	21	Manchester Utd	ENG	1. Premier League	1st	34	31	2,781	30.9	17	8	25	14	3	4	2	0							
2007-2008	22	Manchester Utd	ENG	1. Premier League	1st	34	31	2,747	30.5	31	6	37	27	4	5	5	1							
2008-2009	23	Manchester Utd	ENG	1. Premier League	1st	33	31	2,742	30.5	18	6	24	14	4	4	7	1							
2009-2010	24	Real Madrid	ESP	1. La Liga	2nd	29	28	2,461	27.3	26	7	33	22	4	5	4	2							
2010-2011	25	Real Madrid	ESP	1. La Liga	2nd	34	32	2,914	32.4	40	9	49	32	8	8	2	0							
2011-2012	26	Real Madrid	ESP	1. La Liga	1st	38	37	3,350	37.2	46	12	58	34	12	13	4	0							
2012-2013	27	Real Madrid	ESP	1. La Liga	2nd	34	30	2,716	30.2	34	10	44	28	6	7	9	0							
2013-2014	28	Real Madrid	ESP	1. La Liga	3rd	30	30	2,534	28.2	31	9	40	25	6	6	4	1							
2014-2015	29	Real Madrid	ESP	1. La Liga	2nd	35	35	3,100	34.4	48	16	64	38	10	12	5	1							
2015-2016	30	Real Madrid	ESP	1. La Liga	2nd	36	36	3,183	35.4	35	9	44	29	6	9	2	0							
2016-2017	31	Real Madrid	ESP	1. La Liga	1st	29	29	2,539	28.2	25	6	31	19	6	8	4	0							
2017-2018	32	Real Madrid	ESP	1. La Liga	3rd	27	27	2,285	25.4	26	5	31	23	3	4	1	0	25.2	22.0	5.0	27.0	118	99	298
2018-2019	33	Juventus	ITA	1. Serie A	1st	31	30	2,688	29.9	21	8	29	16	5	6	3	0	22.2	17.5	4.6	22.1	145	130	358
2019-2020	34	Juventus	ITA	1. Serie A	1st	33	33	2,917	32.4	31	5	36	19	12	13	3	0	28.6	18.4	6.4	24.7	167	118	321
2020-2021	35	Juventus	ITA	1. Serie A	4th	33	31	2,802	31.1	29	2	31	23	6	8	3	0	27.7	21.4	3.8	25.2	154	117	277
2021-2022	36	Juventus	ITA	1. Serie A	4th	1	0	31	0.3	0	0	0	0	0	0	1	0	0.2	0.2	0.1	0.2	1	2	3
2021-2022	36	Manchester Utd	ENG	1. Premier League	6th	30	27	2,456	27.3	18	3	21	15	3	3	8	0	17.7	15.4	2.9	18.2	67	64	192
2022-2023	37	Manchester Utd	ENG	1. Premier League	3rd	10	4	525	5.8	1	0	1	1	0	0	2	0	1.9	1.9	0.4	2.3	9	12	33
2022-2023	37	Al-Nassr	KSA	1. Pro League	2nd	16	16	1,433	15.9	14	2	16	9	5	5	3	0							
2023-2024	38	Al-Nassr	KSA	1. Pro League	2nd	14	14	1,254	13.9	15	7	22	12	3	3	0	0							

Data Source: Transfermarkt

- ***Transfermarkt website:*** Football transfers, rumours, market values and news | Transfermarkt
- ***Reason for choosing:*** The website provides player injury data, a necessary component for the project.
- ***Data collected:*** Injury data for the latest 10 seasons for each player in the Premier League 2023/2024
- ***Data collection method:*** Sending requests to the website's server and parsing HTML text.



Player/Position	Club	Injury	until	Market Value
Danny Welbeck Centre-Forward		Hamstring injury	Unknown	€7.00m
Chris Basham Centre-Back		Leg injury	Unknown	€300k
Dean Henderson Goalkeeper		Torn thigh muscle	Unknown	€18.00m



Player/Position	Club	Age	Nation	Injury	since	until	Market Value
Adama Traoré Right Winger		27		Hamstring injury	Nov 20, 2023	Unknown	€10.00m
Jurrien Timber Centre-Back		22		Cruciate ligament tear	Aug 13, 2023	Unknown	€42.00m
Callum Wilson Centre-Forward		31		Hamstring injury	Nov 7, 2023	Unknown	€16.00m
Tariq Lamptey Right-Back		23		unknown injury	Nov 25, 2023	Unknown	€12.00m
Alfie Whiteman Goalkeeper		25		Ankle injury	Aug 3, 2023	Unknown	€500k
Kevin De Bruyne Attacking Midfield		32		Partial muscle tear	Aug 12, 2023	Unknown	€70.00m
Sven Botman Centre-Back		23		Knee injury	Sep 28, 2023	Unknown	€45.00m
Trevoh Chalobah Centre-Back		24		Thigh problems	Aug 1, 2023	Unknown	€18.00m
Michail Antonio Centre-Forward		33		Knee injury	Nov 18, 2023	Unknown	€7.00m
André Gomes Central Midfield		30		Fitness	Aug 13, 2023	Unknown	€16.00m
Lisandro Martínez Centre-Back		25		Foot injury	Sep 21, 2023	Unknown	€50.00m
Pervis Estupiñán Left-Back		25		Muscle injury	Nov 9, 2023	Unknown	€35.00m
Anel Ahmedhodzic Centre-Back		24		unknown injury	Oct 20, 2023	Unknown	€20.00m

Table of contents

Data preprocessing process

- First, we preprocess fbref_data and transfermarkt_data

Raw data inspect → Checking duplicated rows → Field meanings → Remove meaningless rows and columns → Reformatting features and adjusting data types → Processed data

	Name	Position	PreferredFoot	Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s
0	William Saliba	DF	Right	2018-2019	17	Saint-Étienne	FRA	Ligue 1	4	16	13	1277	14.2
1	William Saliba	DF	Right	2019-2020	18	Saint-Étienne	FRA	Ligue 1	17	12	11	992	11.0
2	William Saliba	DF	Right	2020-2021	19	Nice	FRA	Ligue 1	9	20	20	1800	20.0
3	William Saliba	DF	Right	2020-2021	19	Arsenal	ENG	Jr. PL2 -- Div. 1	10	6	6	526	5.8
4	William Saliba	DF	Right	2021-2022	20	Marseille	FRA	Ligue 1	2	36	36	3240	36.0

	Name	Season	Injury	from	until	Days	Games missed
0	Ederson	2020-2021	Virus	2020-12-27	2021-01-12	16	3
1	Ederson	2019-2020	Ill	2019-12-30	2020-01-07	8	3
2	Ederson	2019-2020	Ill	2019-12-08	2019-12-14	6	1
3	Ederson	2019-2020	muscular problems	2019-11-06	2019-11-22	16	4
4	Ederson	2017-2018	Facial injury	2017-09-10	2017-09-12	2	0

- Then, we merge fbref_data and transfermarkt_data into merged_data

Overview about merged data

- ***Basic Player Information:***
  - Name, Position, PreferredFoot, Season, Age, Squad, Country.
- ***Performance Information:***
  - Comp, LgRank, MP, Starts, Mins, 90s, Gls, Ast,..
- ***Expected Information:***
  - xG, npxG, xAG, npxG+xAG,…
- ***Injury data:***
  - Injury, from, until, Days, Games missed.

```
merged_df.head()
```

✓ 0.0s

	Name	Position	PreferredFoot	Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s	Gls	Ast	G+A	G-PK	PK
0	William Saliba	DF	Right	2018-2019	17	Saint-Étienne	FRA	Ligue 1	4	16	13	1277	14.2	0	0	0	0	0
1	William Saliba	DF	Right	2019-2020	18	Saint-Étienne	FRA	Ligue 1	17	12	11	992	11.0	0	0	0	0	0
2	William Saliba	DF	Right	2020-2021	19	Nice	FRA	Ligue 1	9	20	20	1800	20.0	1	0	1	1	0
3	William Saliba	DF	Right	2020-2021	19	Arsenal	ENG	Jr. PL2 -- Div. 1	10	6	6	526	5.8	0	0	0	0	0
4	William Saliba	DF	Right	2021-2022	20	Marseille	FRA	Ligue 1	2	36	36	3240	36.0	0	0	0	0	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4571 entries, 0 to 4570
Data columns (total 43 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Name           4571 non-null   object
 1   Position       4571 non-null   object
 2   PreferredFoot  4571 non-null   object
 3   Season         4571 non-null   object
 4   Age            4571 non-null   int32
 5   Squad          4571 non-null   object
 6   Country        4571 non-null   object
 7   Comp           4571 non-null   object
 8   LgRank         4571 non-null   int32
 9   MP             4571 non-null   int32
 10  Starts         4571 non-null   int32
 11  Min            4571 non-null   int32
 12  90s            4571 non-null   float64
 13  Gls            4571 non-null   int32
 14  Ast            4571 non-null   int32
 15  G+A            4571 non-null   int32
 16  G-PK           4571 non-null   int32
 17  PK             4571 non-null   int32
 18  PKatt          4571 non-null   int32
 19  CrdY           4571 non-null   int32
...
 41  Days           2601 non-null   float64
 42  Games missed   2601 non-null   float64
dtypes: datetime64[ns](2), float64(17), int32(16), object(8)
memory usage: 1.2+ MB
```

```
merged_df.shape
```

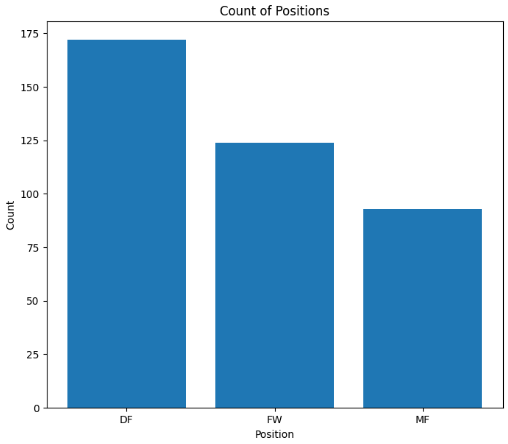
Python

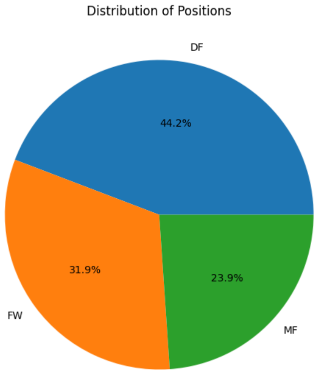
```
(4571, 43)
```

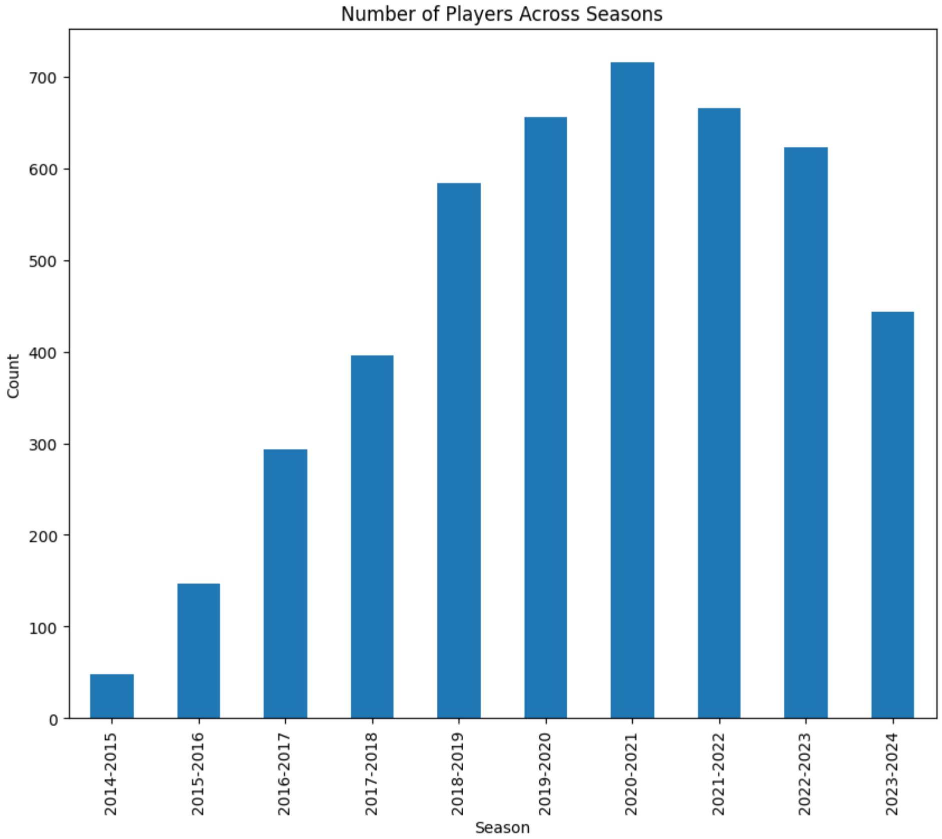
Table of contents

Distribution of Categorical Data

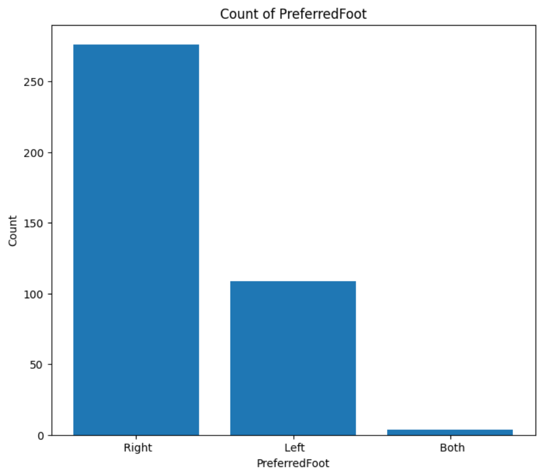
Analyzing Postion, Season

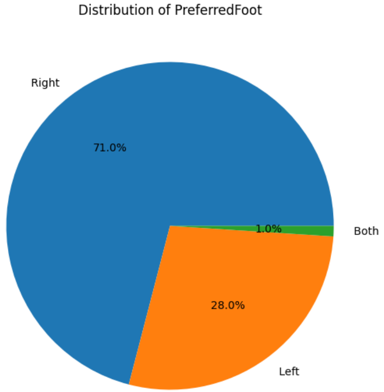


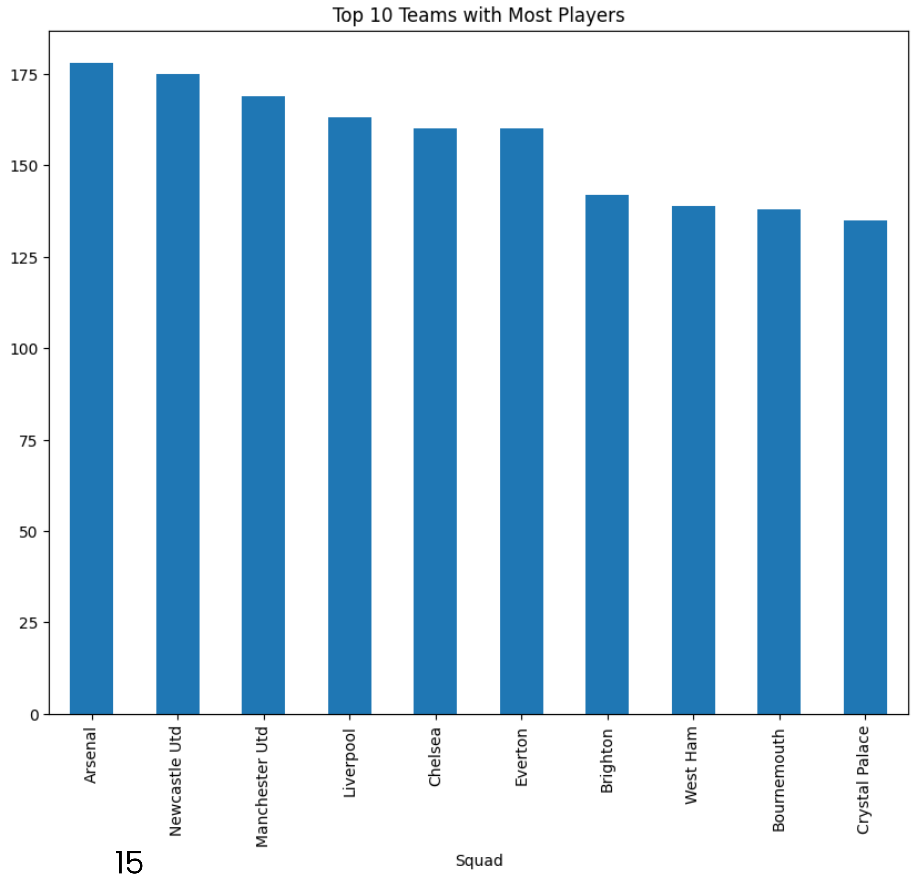




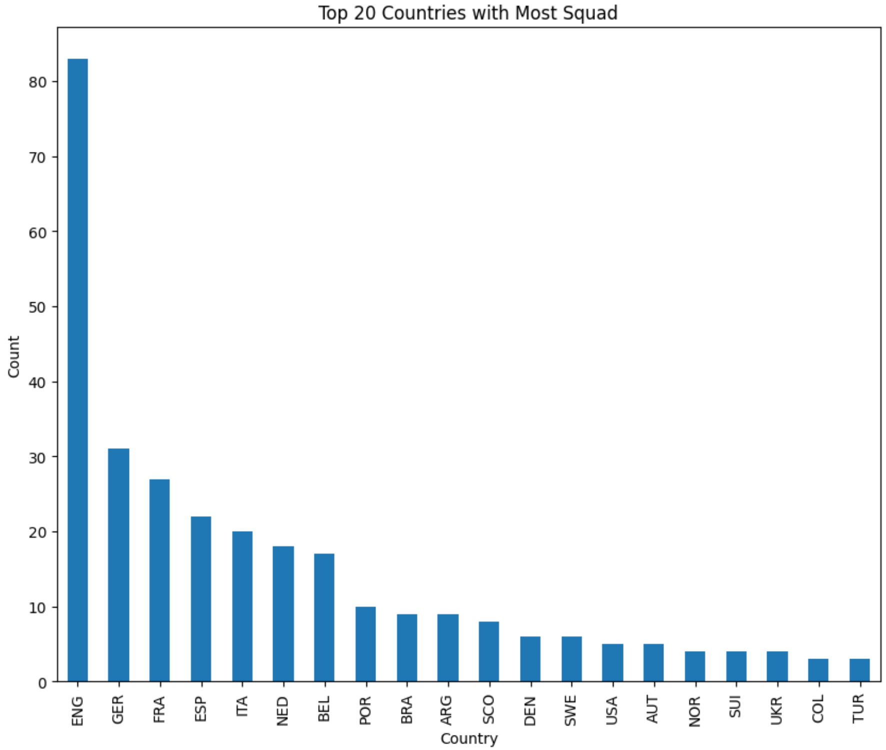
Analyzing PreferredFoot, Squad

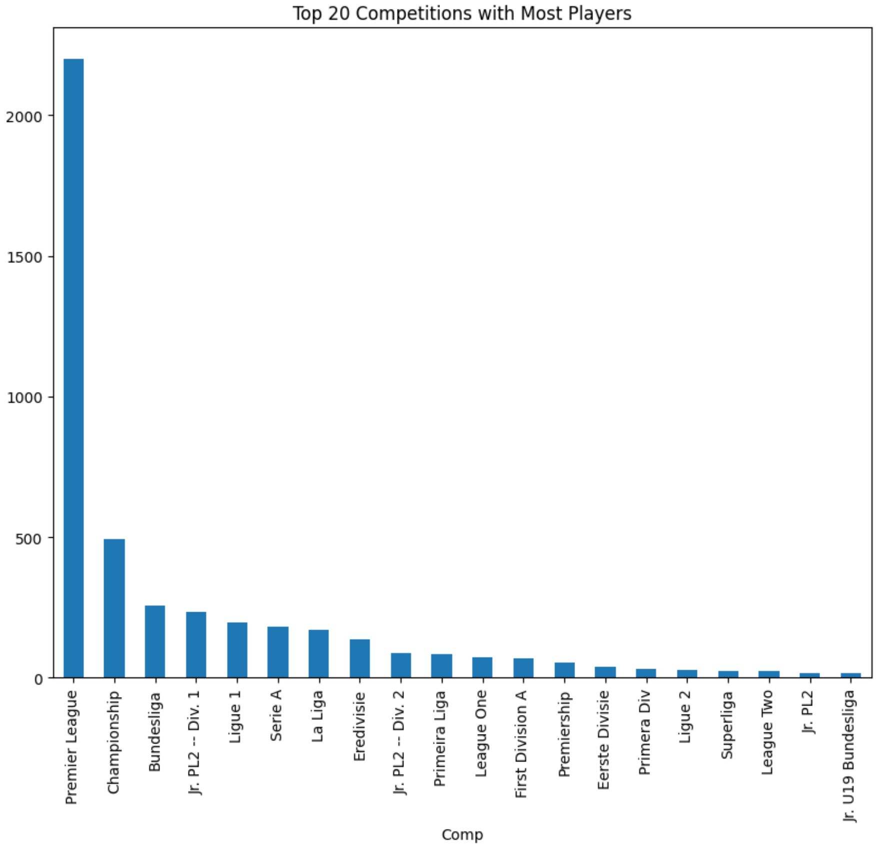






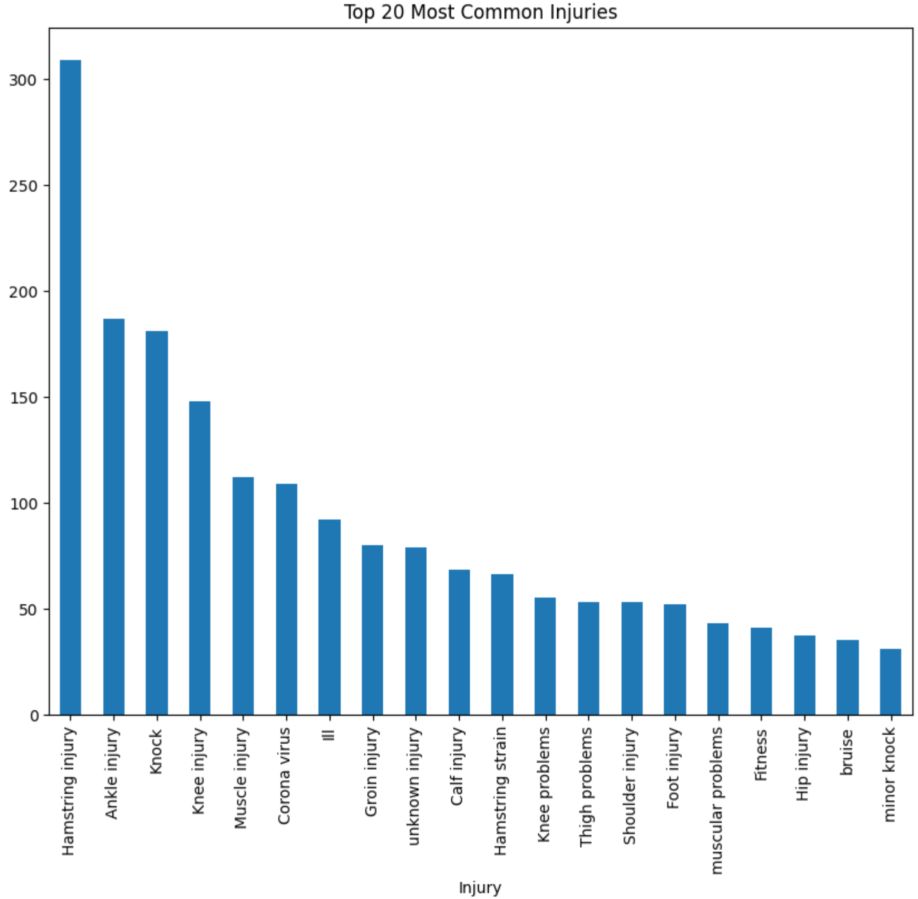
Analyzing Country & Comp





Analyzing Injury

	Injury
count	2601
unique	178
top	Hamstring injury
freq	309

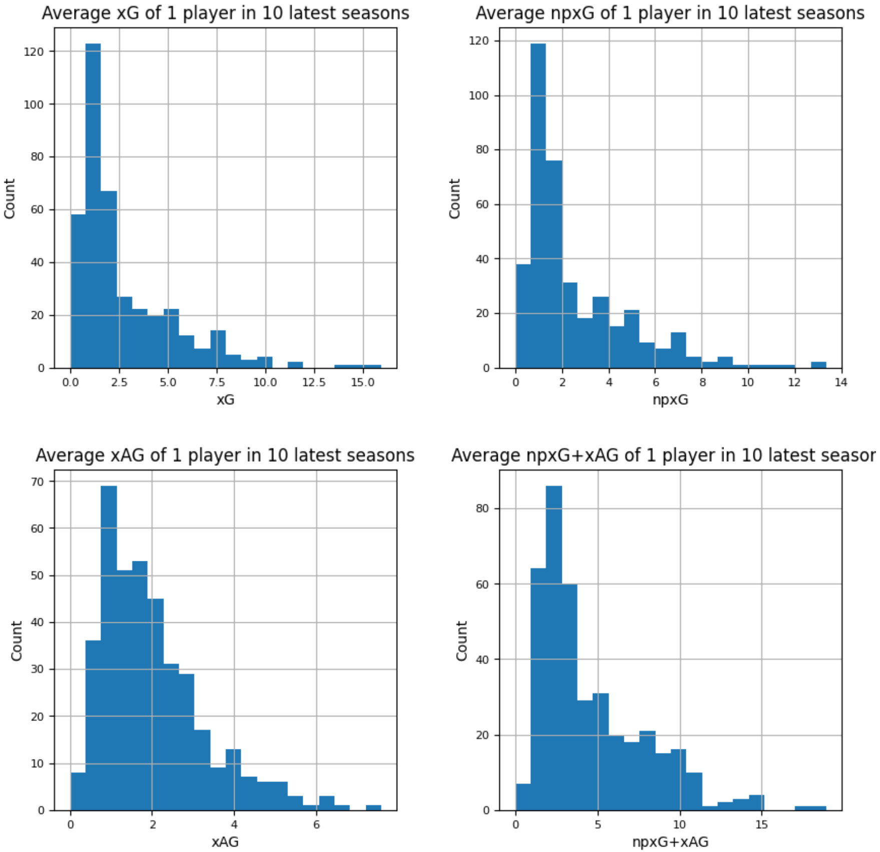


	count
count	178.00000
mean	14.61236
std	35.97019
min	1.00000
25%	1.00000
50%	3.00000
75%	9.75000
max	309.00000

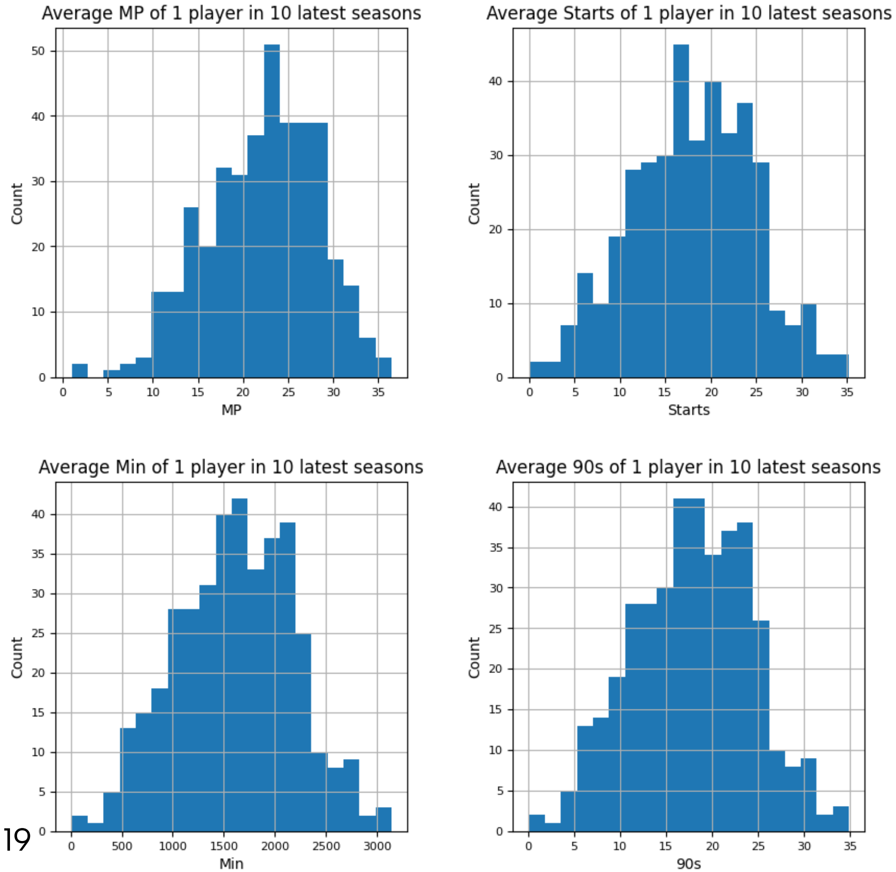
Distribution of Numerical Data

Analyzing Expected & Playing Time

Expected

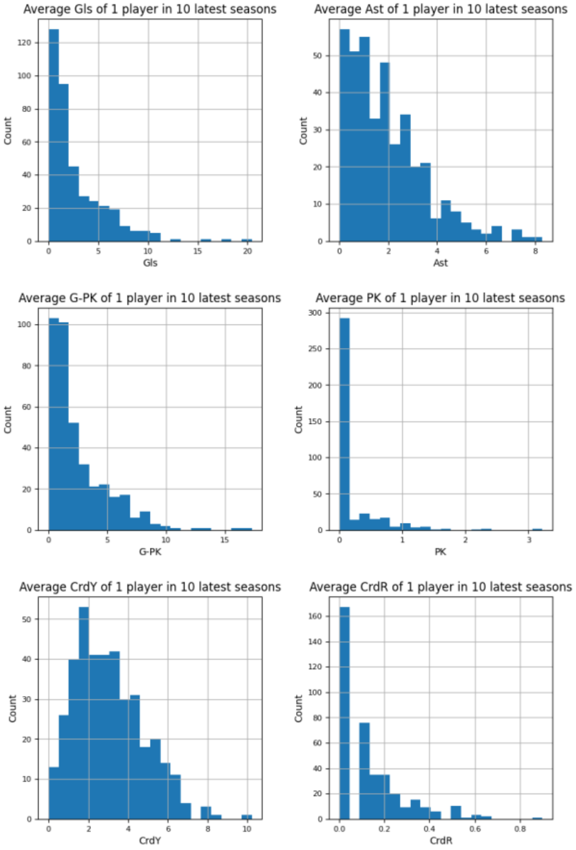


Playing time statistics

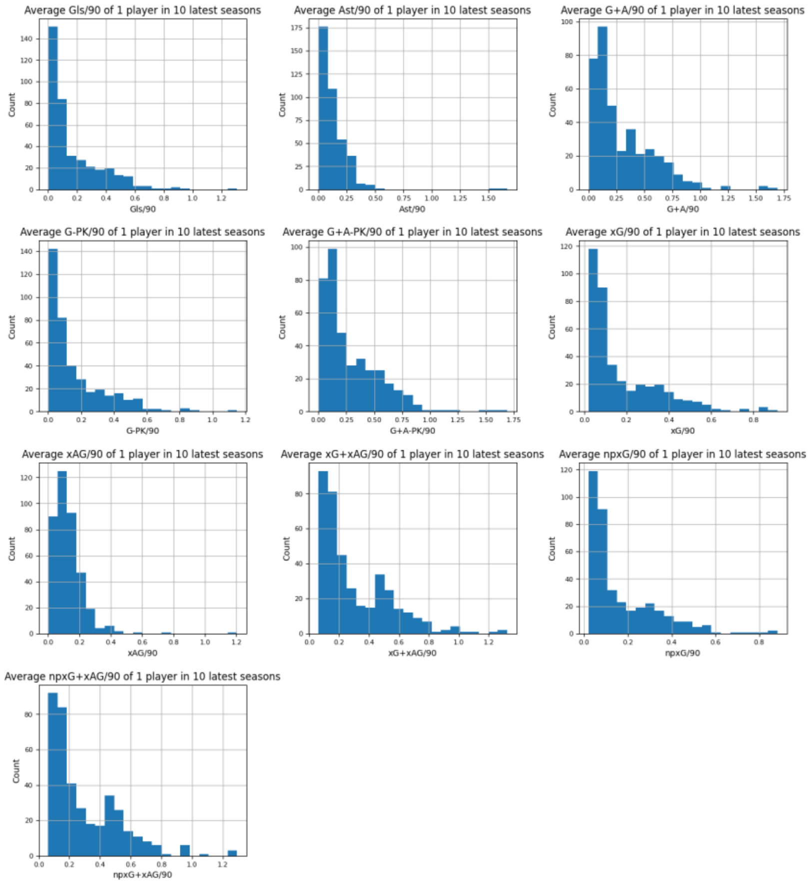


Analyzing Performance & Per 90 Minutes

Performance

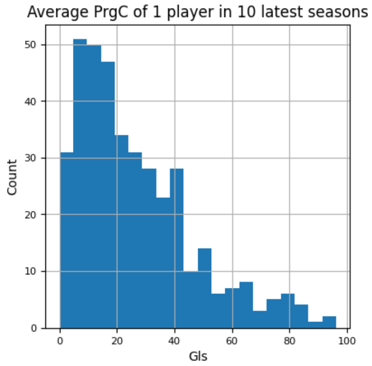


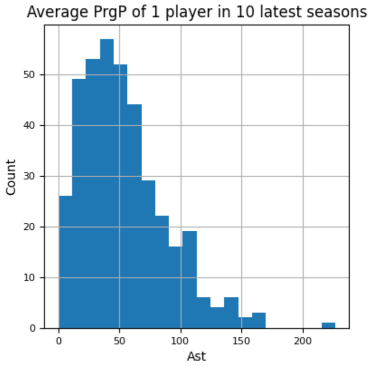
Per 90 Minutes

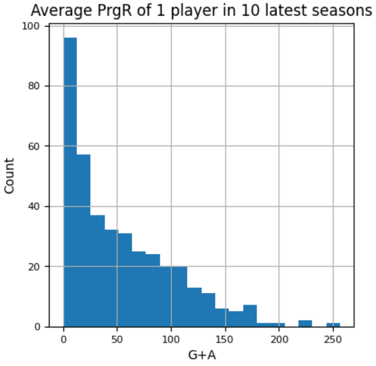


Analyzing Progression & Age

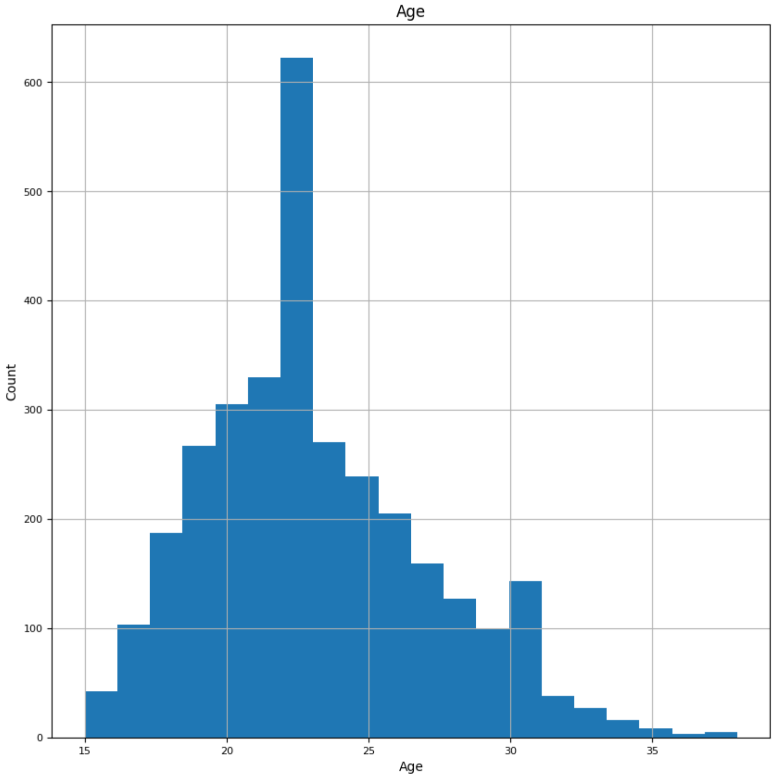
Progression







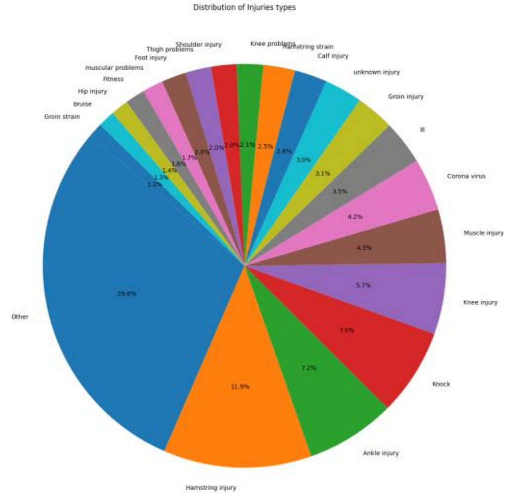
Age

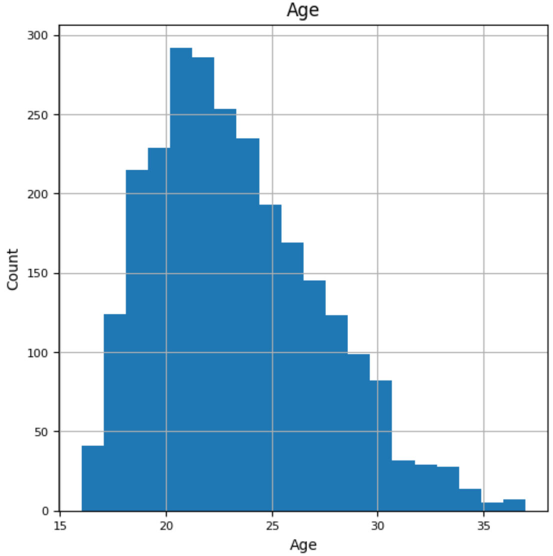


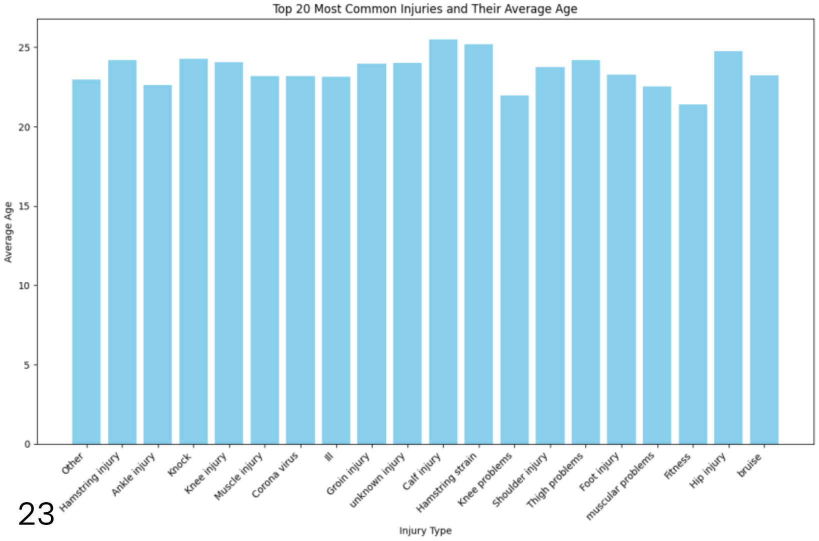
Making questions about data

Question 1

- **A possible question is:** Are players more prone to injuries as they age?
- **Answering this question will** explore if there's a relationship between a player's age and the likelihood of sustaining injuries.
- **How we answer this question:** Analyze the frequency and types of injuries across different age.

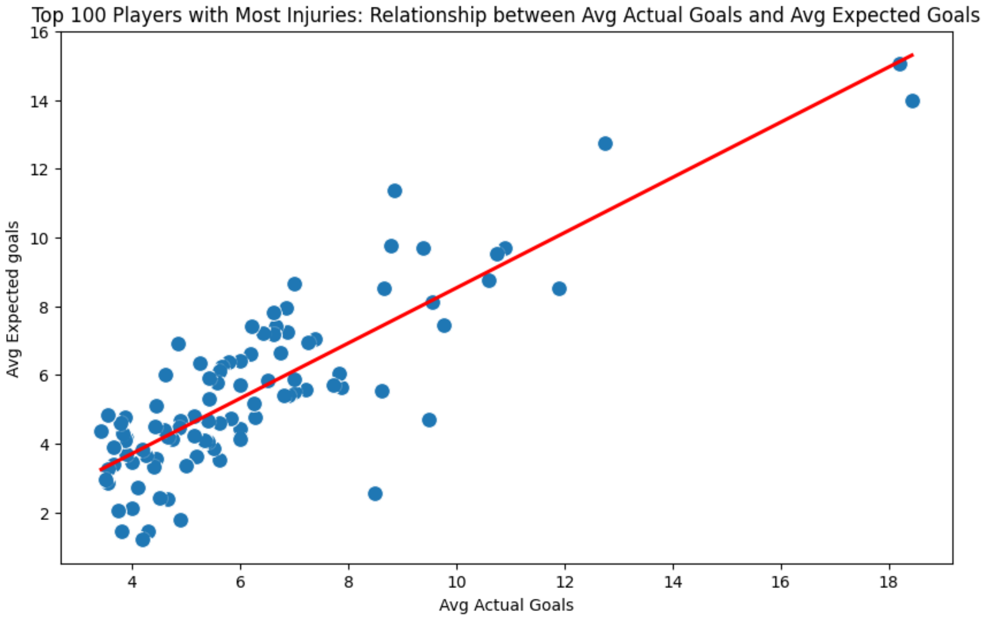






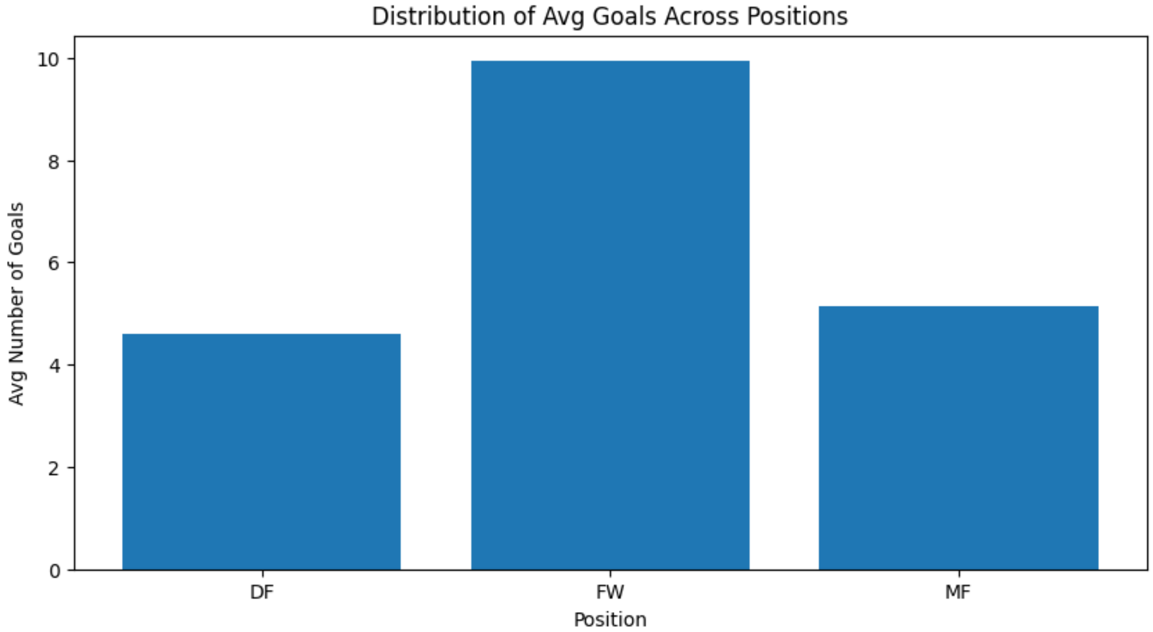
Question 2

- **A possible question is:** What is the relationship between expected goals (xG) to actual goals (Gls) for penalty kicks (PK)?
- **Answering this question will** help us understand the efficiency of players in converting penalty kicks compared to their expected goals.
- **How we answer this question:** Calculate the ratio of 'xG' to 'Gls' specifically for penalty kicks.



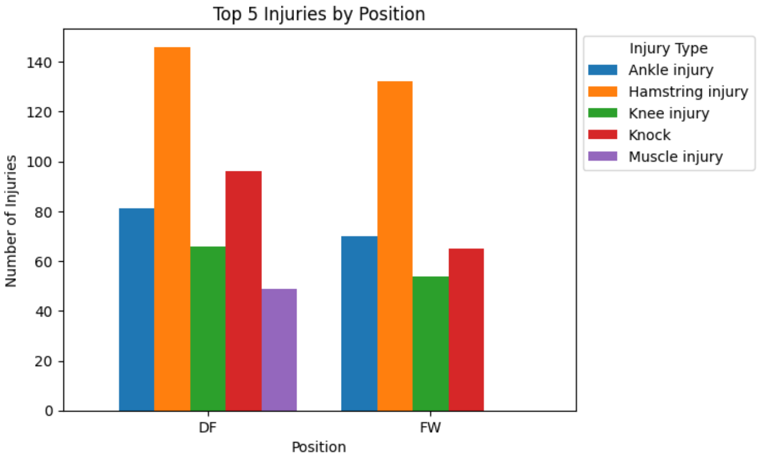
Question 3

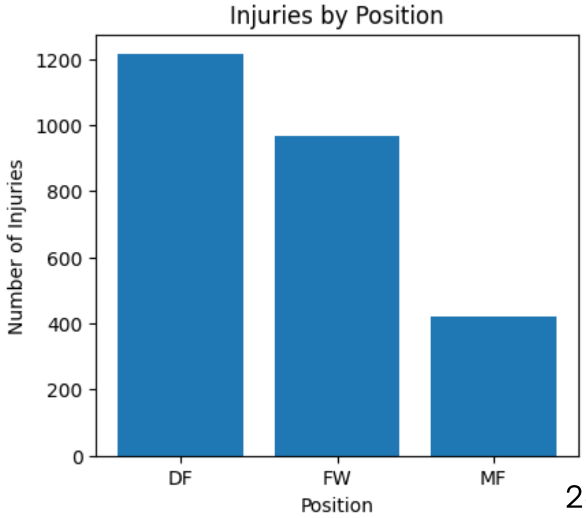
- **A possible question is:** What is the distribution of goals scored by players across different positions in the dataset?
- **Answering this question will** help us understand how goals are spread across various playing positions.
- **How we answer this question:** Create a breakdown of the number of goals scored (Gls – Goals scored or allowed) by players for each position.

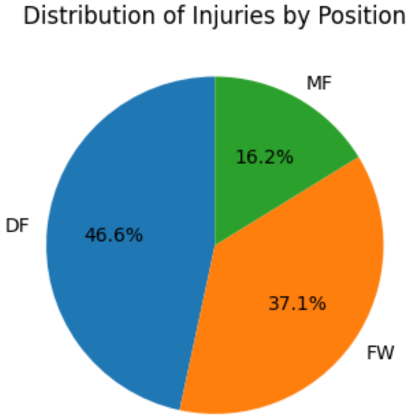


Question 4

- **A possible question is:** How does the number of injuries vary across player positions?
- **Answering this question will** help us understand if there are differences in the injury rates based on player positions.
- **How we answer this question:** Analyze and aggregate the number of injuries for each playing position.







Question 5

- **A possible question is:** How many times have the top 100 players with the most minutes played experienced injuries?
- **Answering this question will help us** understand the overall injury frequency for the highest-minutes players.
- **How we answer this question:** Collect and calculate average the injury occurrences for each player in the top 100 highest-minutes players.

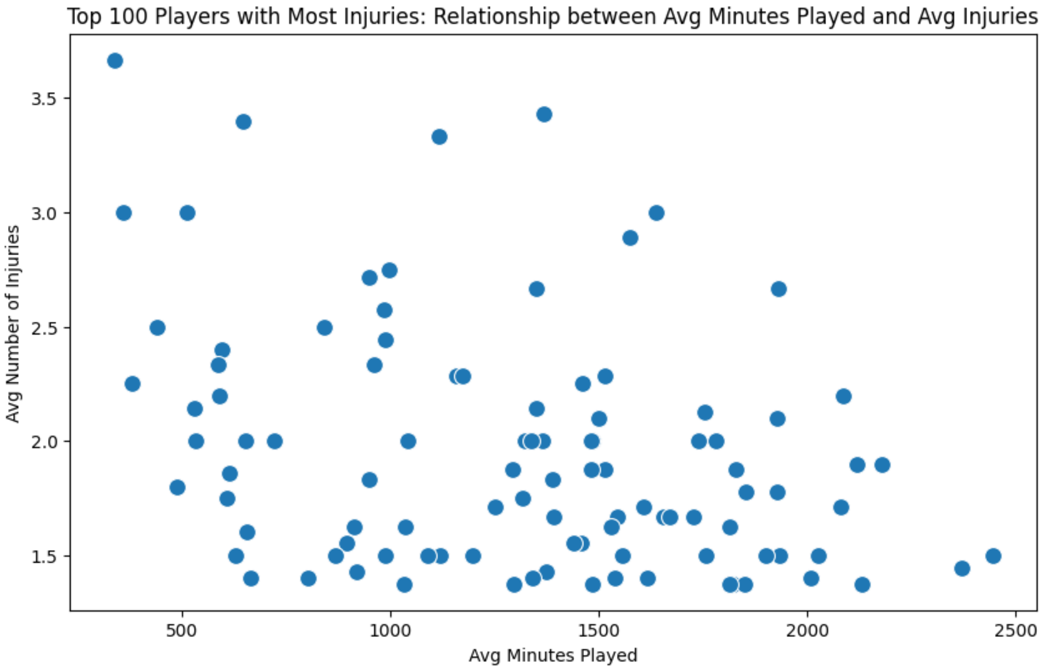
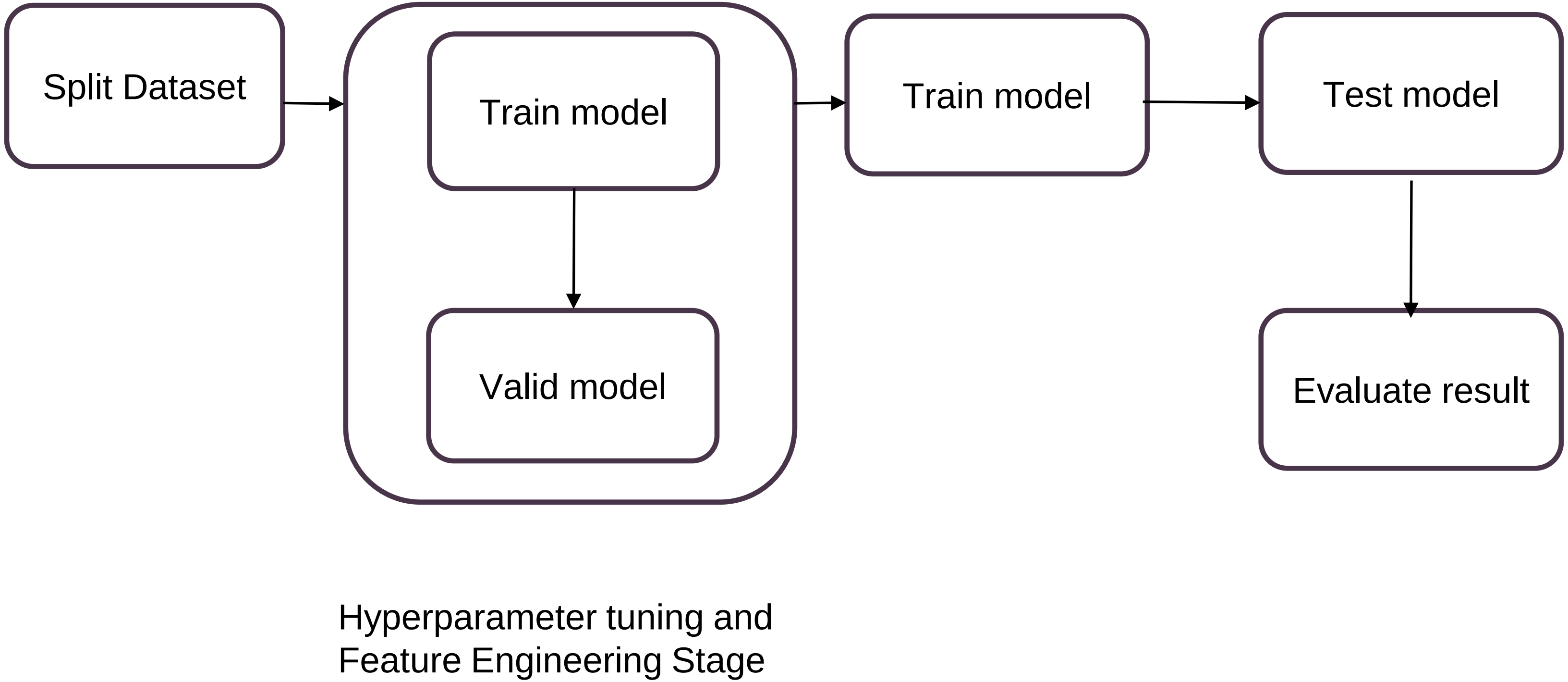


Table of contents

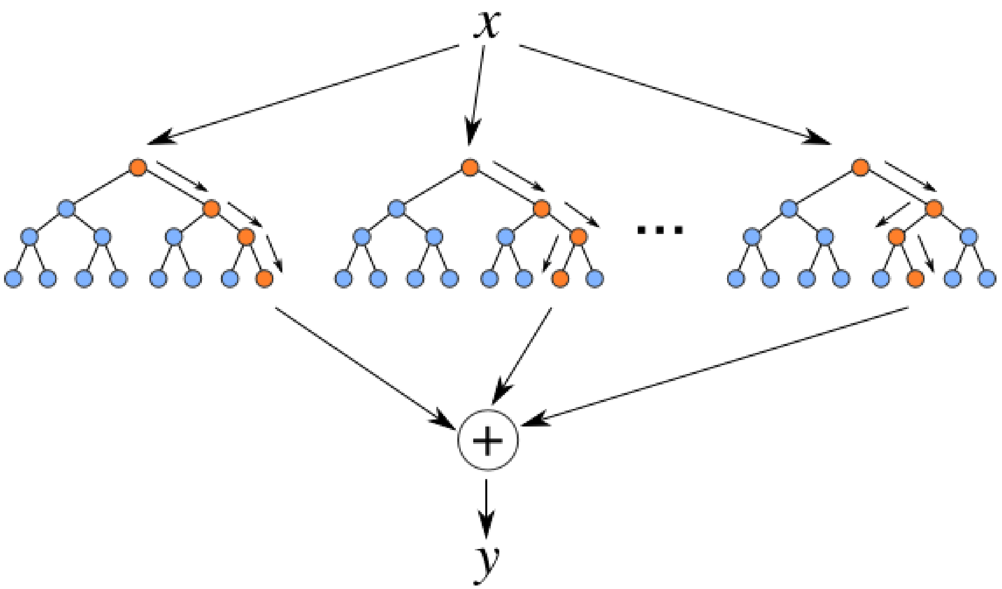
Model 1: Ridge Linear Regression

The running process of training and testing Ridge Linear Regression Model

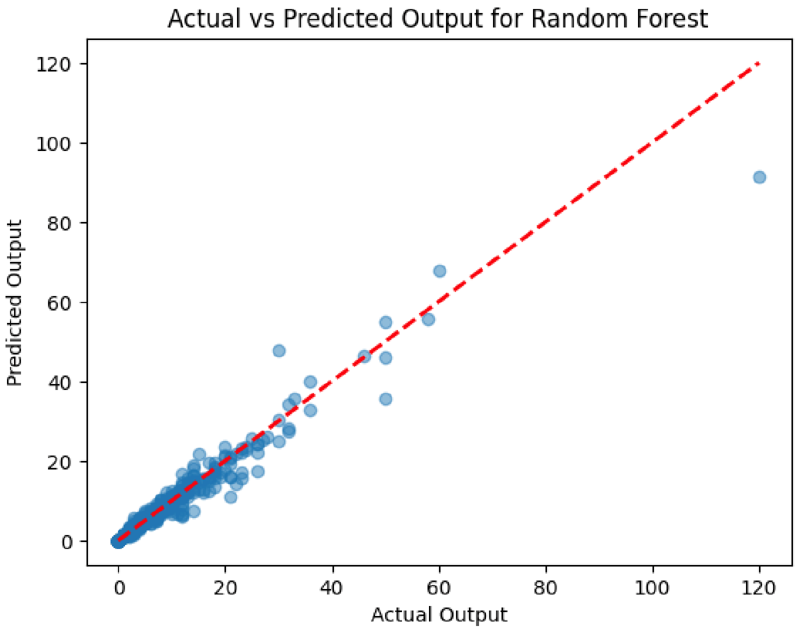


Model 2: Random Forest

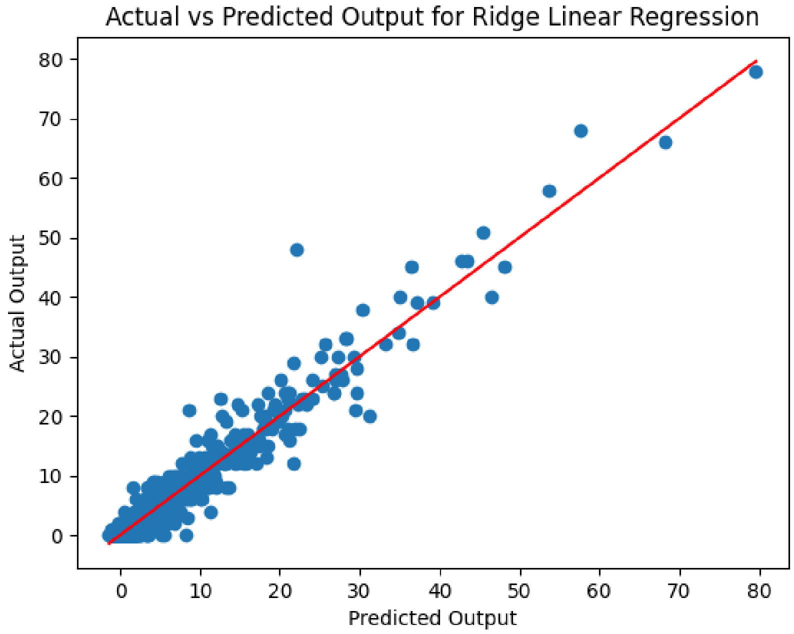
- Each Decision Tree is a unit
- Each tree built base on different training data and different predictors from every other tree (private information)
- The last step is to take either the mean (regression) or mode (classification)



Linear regression vs Random forest

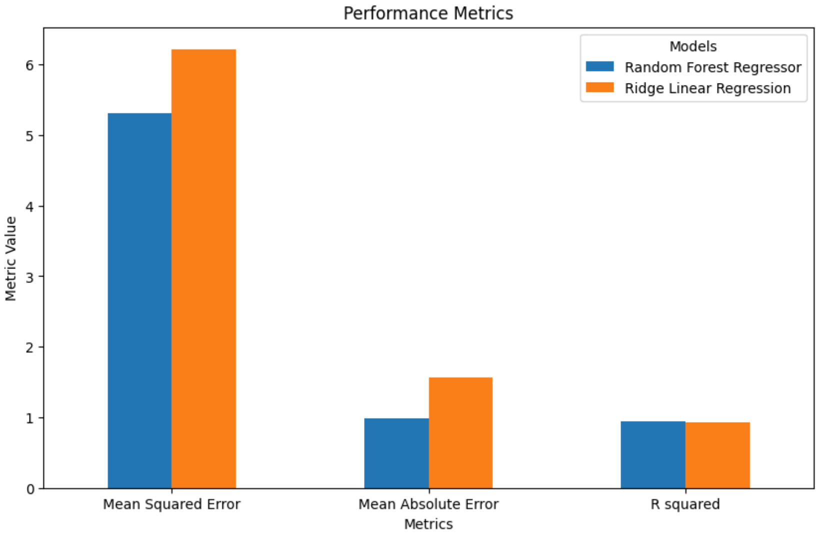


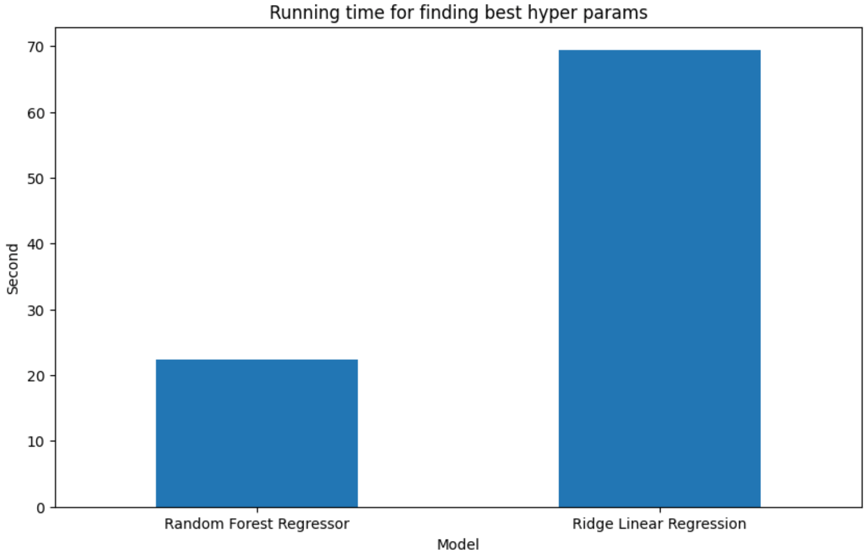

```
Best Hyperparameters:
   n_estimators: 50
   min_samples_split: 10
   min_samples_leaf: 2
   max_features: None
   max_depth: 15
```



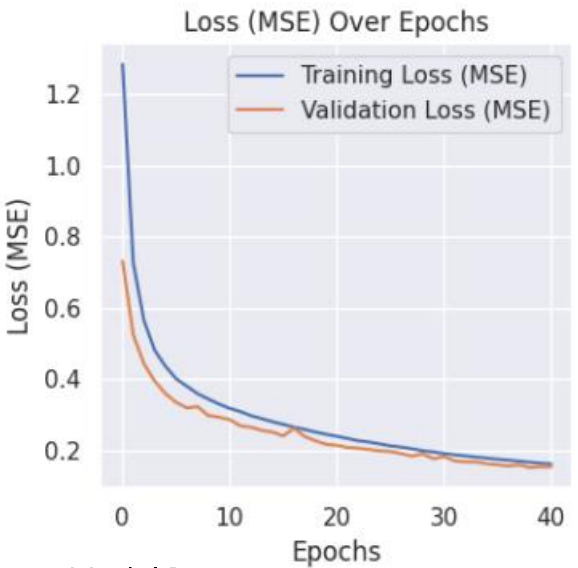

```
Best Hyperparameters:
   alpha: 10
   fit_intercept: False
   max_iter: 500
   solver: svd
```

Linear regression vs Random forest

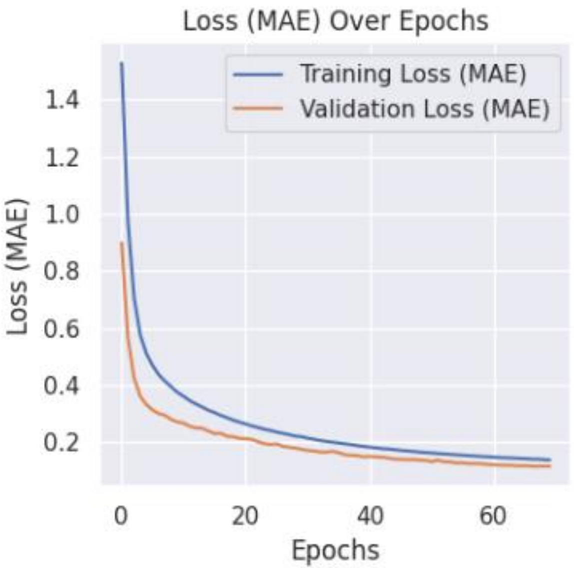




Model 3: Fully-Connected Neuron Network

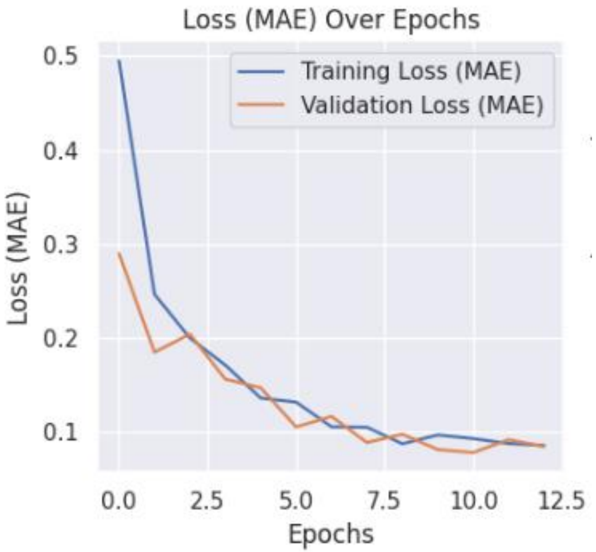


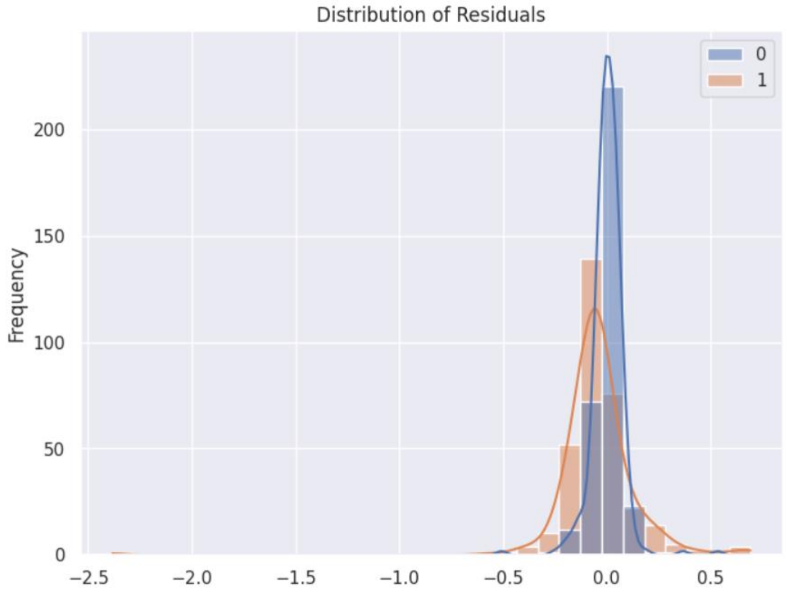
- Model 1
  - Neurons = 100
  - Layers = 2
  - Learning rate = 0,001



- Model 2
  - Neurons = 100
  - Layers = 3
  - Initial Learning rate = 0,001
  - Learning rate schedule: Exponential decay

Model 3: Fully-Connected Neuron Network





- Model 3:
  - Neurons = 200
  - Layers = 3
  - Initial Learning rate = 0,001
  - Learning rate schedule: RMSprop
  - Momentum applied

```
Test Loss: 0.088
Test MSE: 0.028
Test MAE: 0.088
Test R-squared: 0.98
Test Adjusted R-squared: 0.978
```

Conclusion

- It can be concluded that the models have successfully addressed the initial problem of predicting the performance of players.
- It can be observed that all models, including Random Forest Regressor, Ridge Linear Regression, Fully-connected models, yield very high accuracy results. However, the difference between the two machine learning models (Random Forest Regressor, Ridge Linear Regression) and deep learning models is not significant, but the processing time of the neural network models is much higher.
- Therefore, it can be said that for the given problem, the Random Forest Regressor model produces the best results.

Thanks For Listening!