

# Data Curation

## English Premier League 2024/25 Player and Team Performance Analysis

### Data Sourcing:

#### English Premier League - Player Stats - 24/25

The dataset used in this project looks at player performance during the 2024/25 English Premier League season. Each row represents one player and includes many statistics covering different parts of the game.

**Dataset Details:** The dataset includes 563 rows representing 562 different players, with 57 columns containing various performance statistics.

**Size:** The file size is 100.27 kB.

**Access:** I'll download the dataset from Kaggle.

**Source:** [English Premier League - Player Stats - 24/25](#)

## Data Profiling:

- Apply Excel filters to the dataset.
- Distinct values in some columns:
  - 20 unique clubs.
  - 4 unique positions (GKP, DEF, MID, FWD).
- Errors:
  - The club Brighton & Hove Albion appears twice, once as “Brighton & Hove Albion” and once as “Brighton”.
  - Some players have more shots on target than total shots, which isn’t possible. Many others are having wrong numbers for shots on target.
  - The conversion rate is not accurate since it depends on incorrect shot data.
  - One goalkeeper is shown with goals scored, which is likely a position or data entry mistake.
  - The positions of a few players are wrong.
- Zeros in data:
  - Many players (especially non-goalkeepers) have zeros in goalkeeper-specific stats.
  - Goalkeepers and players with fewer minutes played have zeros in many statistics because they were less involved in certain aspects of the game.
- Summary of insights:

The data contains valuable data of player performance in the EPL 2024/25 season, but several statistics need to be corrected before analysis. Errors in shooting and positional data could affect accuracy, so cleaning these issues is essential. Despite that, the dataset remains valuable, allowing us to compare performance across clubs, positions, and individual players.

## Data Wrangling:

- “Shots on target” and “Conversion %” columns are removed as they contain inaccurate data.
- Feature engineered “Club” column; substitute any “Brighton” with “Brighton & Hove Albion”.
- Feature engineered “Position” column, substitute “MID” with “GKP” for Alex Palmer and substitute “GKP” with “DEF” for Emerson.
- Featured the engineered “Defensive Contributions” column by adding “Clearances”, “Interceptions”, “Blocks”, and “Tackles” columns.

## Data Table Schema:

Field	Type	Description
Player Name	STRING	Name of each player.
Club	STRING	The club that the player is playing for.
Position	STRING	The player’s position on the pitch.
Appearances	INTEGER	Number of matches played by the player.
Minutes	INTEGER	Minutes played by the player.
Goals	INTEGER	Goals scored by the player.
Assists	INTEGER	Assists given by the player.
Shots	INTEGER	Shots by the player.
Touches	INTEGER	Number of times the player touched the ball.
Passes	INTEGER	Number of passes by the player.
Crosses	INTEGER	Number of crosses by the player.
Through Balls	INTEGER	Numbers of through ball passes by the player.
Carries	INTEGER	Number of times ball was carried by the player.
Clearances	INTEGER	Number of times the player cleared the ball.
Interceptions	INTEGER	Numbers of times the player stopped a pass.
Tackles	INTEGER	Number of tackles made by the player.
Fouls	INTEGER	Number of fouls committed by the player.
Yellow Cards	INTEGER	Number of yellow cards for each player.
Saves	INTEGER	Number of saves made by each goalkeeper