

MiniProject

Michał Wiącek, Jakub Moszyński, Olaf Niedźwiadek

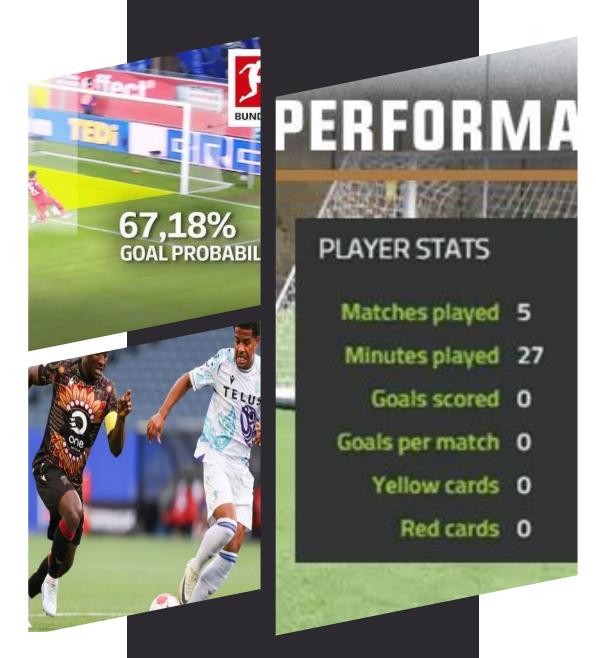
Let's play! >

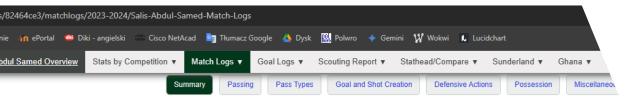
The goal of the project

The aim of the mini-project was to understand the "business needs" and the problem reality represented by the available data resources, and then to enable the identification of needs, objectives, and possibilities for business analysis supporting decision-making processes.

The Title - Analysis of player performance metrics in the five biggest European leagues (Premier League, La Liga, Serie A, Bundesliga, and Ligue 1) for the 2023–2024 season

It focuses on individual player performance metrics in matches during the 2023–2024 season. The main business process was tracking and analyzing player statistics during matches to support strategic decision-making in professional football.





3-2024 Match Logs (Summary)

19-07 Thu Africa Cup of Nations qualification Group stage

Matchweek 2

Matchweek 3

Matchweek 4

Matchweek 5

Group stage

Matchweek 7

Matchweek 8

Friendlies (M)

08-20 Sun <u>Ligue 1</u>

08-26 Sat <u>Ligue 1</u>

19-02 Sat Ligue 1

9-16 Sat <u>Ligue 1</u>

9-24 Sun Ligue 1

09-29 Fri Ligue 1

10-08 Sun Ligue 1

19-20 Wed Champions Lg

10-03 Tue Champions Lg

10-14 Sat Friendlies (M)

Scroll i matches as unused substitute Glossary Performa Round Gls Ast PK PKatt Sh SoT CrdY C te Result Sauad Opponent Start Pos Venue 08-13 Sun Ligue 1 Matchweek 1 Away Brest

Rennes

Paris S-G

Monaco

CM

CM

DM

CM

CM

L 2-3

D 1-1

L 1-3

L 0-3

W 2-1

W 1-0

Home

Away

Away

Home

Home

Away

Away

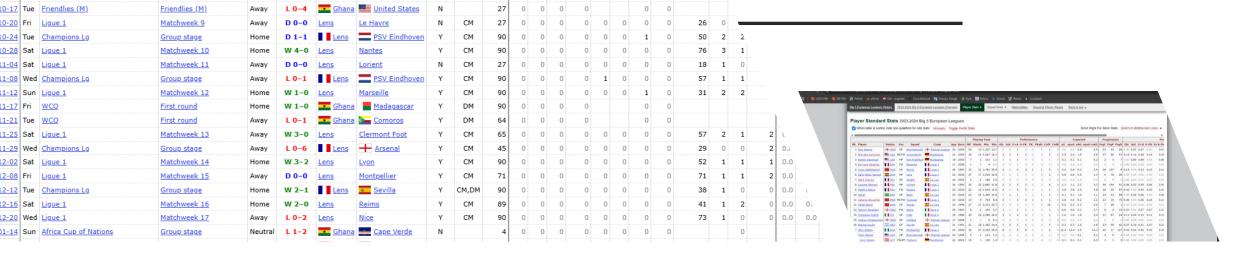
Home

Home

Neutral

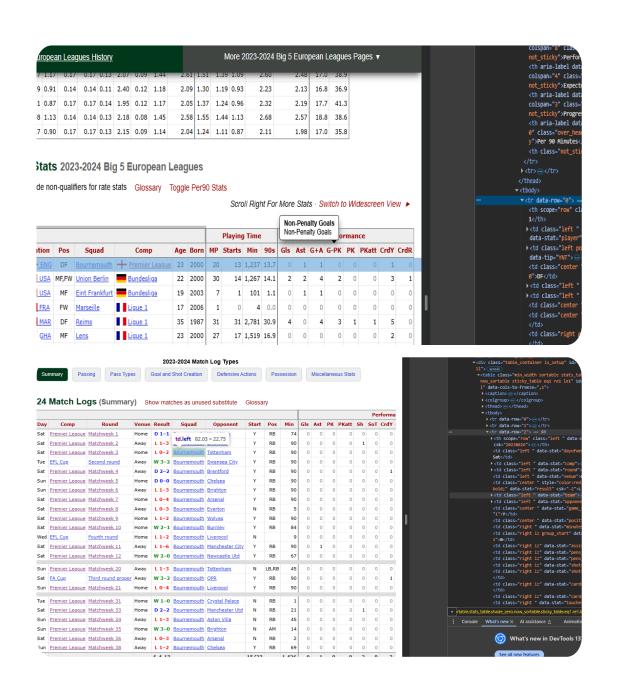
Introduction The raw dataset

> The data was collected (scraped) from the website fbref.com, with particular focus on data about matches, teams, and players. The data is available for free but requires the use of a scraper, as there is no direct download button.



30

34



Challenges

The challanges with the data

Structure

The data is stored as tables linked to entities, but some information is loosely placed on HTML pages, and column headers and separators are sometimes mixed with the data.

Data Quality

Generally well-maintained, but there are formatting issues (e.g., player names, age in YY-DDD format, match results).

Relationships

There are no explicit unique identifiers for linking tables, which requires manual creation of references and a schema.

Time scope

Facts are permanent (match results do not change), but dimensions will grow over time (new matches, new players).



End Users

User Types and Expectations



Club Scouts and recruiters

Require comprehensive player evaluation metrics, cross-league comparisons, and trend analyses.



Sports data analysts

Require statistical insights, performance rankings, and comparative analyses for reporting and media purposes.



Coaches and tactical analysts

Need detailed positional performance analysis, formation efficiency data, and insights into player compatibility.

Dimensions

Dimension Tables

Team

team_id (PK)

name

country

city (can be null if national team)

Dimension

Team

Player

player_id (PK)

fullname

nationality

bornYear

typicalPosition_ld (FK)

defaultTeam_id (FK)

Dimension

Player

Position

position_id (PK)

name

role

line

side

Dimension

Position

Time

datetime_id (PK)

date

year

quarter

month

day

day_of_week

is_weekend

Dimension

Time

Competition

ompetition_id (Pk

name

country

round

Dimension

Competition

Flags

flags_id (PK)

result (D, L, W)

venue ('Home', 'Away', 'Neutral'

base_squad (Y, N, C (Capitan))

Dimension

Flags

Player Performance

performance_id (PK)

player_ld (FK)

datetime_id (FK)

competition_id (FK)

flags_id (FK)

team_id (FK)

opponent_id ((FK)

position_ld (FK)

Minutes

Goals

Assists

PenaltyGoals

PenaltiesAttempts

Shots

ShotsOnTarget

YellowCards

RedCards

Touches

Tackles

Interceptions

ExpectedGoals

ExpectedNonPenaltyGoals

ExpectedAssists

ShotCreatingActions

GoalCreatingActions

PassesSuccessful

PassesAttempts

PassCompletion

ProgressivePasses

Carries

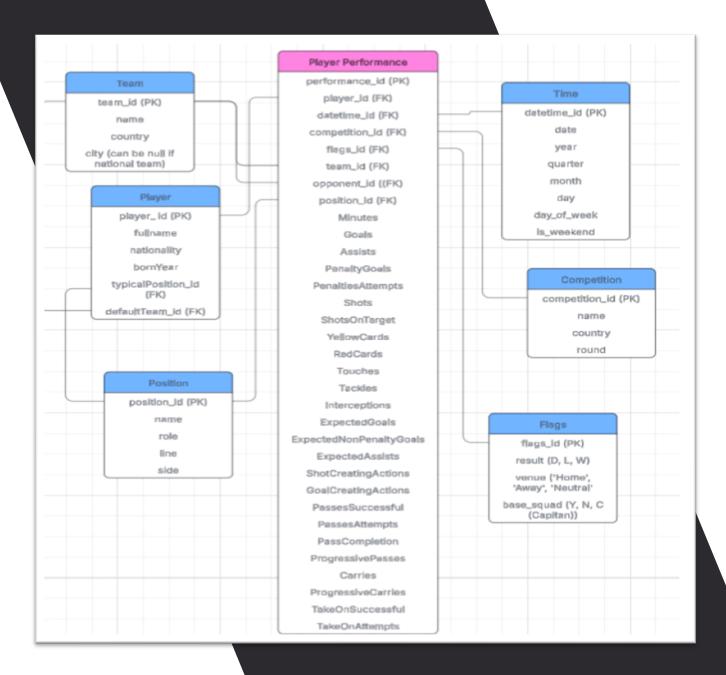
ProgressiveCarries

TakeOnSuccessful

TakeOnAttempts

Facts

Fact Table – PlayerPerformance



Star Schema

The whole schema



Position (and team)

Position is used twice as a typical position of the player as well as the real position during the mach (the same for team)



Flags

The flags table is a dimension for storing attributes that have small numer of possible values – f.e. result (draw, win or loss)

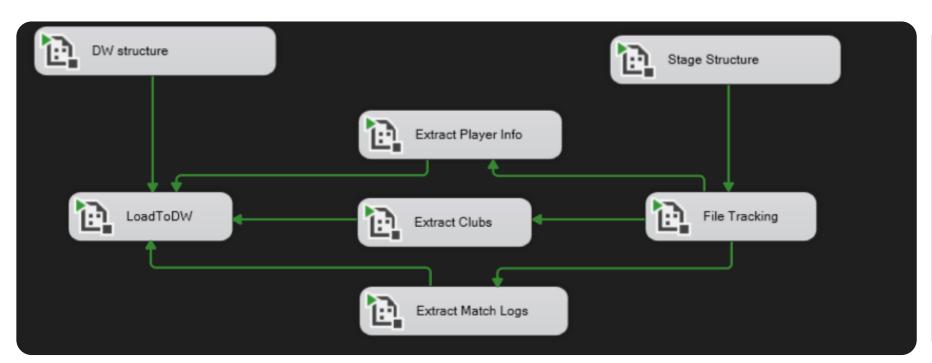


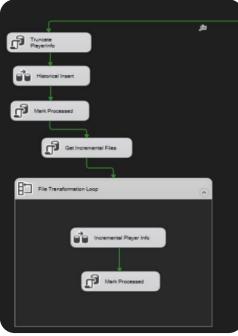






The chosen Tools





Data Processing

ETL Process



1. Schema Preparation: Initial packages create the Data Warehouse (DW) schema (final destination) and the Stage schema (intermediary storage). The Stage schema holds raw data with fewer constraints, while the DW schema holds readable, analysis-compatible data.



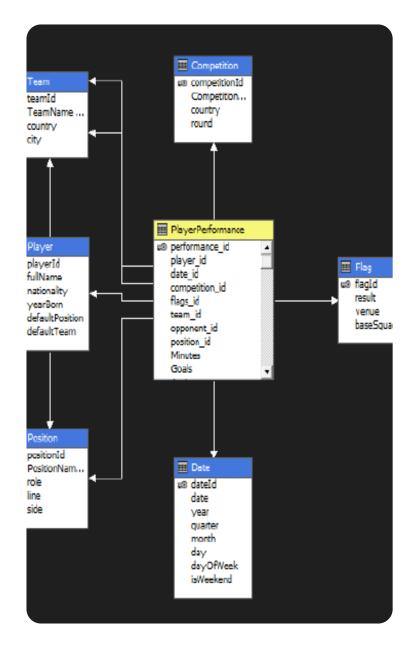
 File Retrieval: The process retrieves raw data files (MatchLogs, PlayerInfo, Clubs) from the source. Files are classified as historical or incremental based on naming schemes.



- 3. Staging Data Loading:
- -Clubs Data: Simple process of truncating the staging table and inserting historical data. Files are marked as processed.
- -MatchLogs & PlayerInfo: Same general control flow, different data flows.
- -Historical Load: If historical data exists, staging data is truncated, and new data is inserted in batches (e.g., 10k records).
- -Incremental Load: For new/modified data, a lookup is performed against staging records; if found, data is updated; otherwise, it's inserted.



- Load to DW (Final Package): After initial cleanup in the Stage schema, data is sent to the DW schema.
- -"Scapegoat" records are created for missing data.
- -Dimensions are created: non-referencing first (e.g., Date), then referencing ones.
- -The fact table is populated: values are compared to dimensions, replaced with foreign keys or scapegoat keys if no match. Existing fact records are updated if foreign keys match.



Player Performance	MeasureGroup	
Minutes	Measure	
, Goals	Measure	
II Assists	Measure	☑
II Penalty Goals	Measure	2
Penalties Attempts	Measure	
Shots	Measure	
. I Shots On Target	Measure	☑
II Yellow Cards	Measure	2
II Red Cards	Measure	3
II Touches	Measure	
Tackles	Measure	
Interceptions	Measure	
II Expected Goals	Measure	
Expected Non Penalty Goals	Measure	



Cube Model

Measures

Measures are all the numerical values retrieved from the PlayerPerformance table.

Calculated measures are derived from already existing ones (f.e. touches per minute – division of touches by the minutes on the pich)

KPI

A KPI named PassCompletionKPI was created to assess passing performance against a defined benchmark.

- -value expression: [Measures].[AveragePassCompletion]
- -goal expression: 0.80
- -status indicators: numeric output (-1, 0, 1) used for conditional formatting





TEST TIME

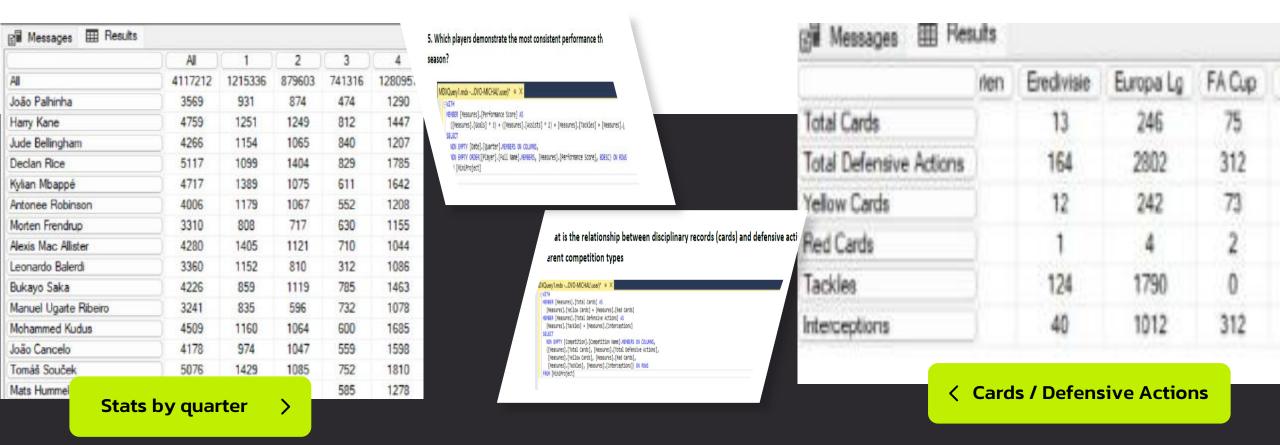
Evaluation of the model + PowerBI dashboard



OLAP Analysis

Feasibility of selected queries

Purpose: To verify how well the final data model addresses the identified user information needs

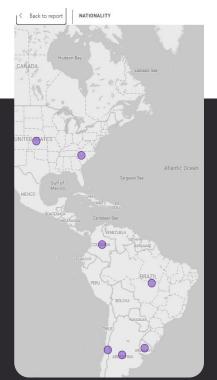


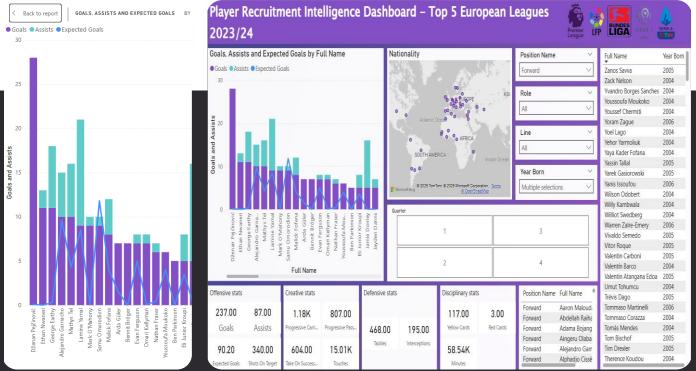
Soccer Scout Dashboard

Player Recruitment Intelligence Dashboard

The dashboard is designed for Club Scouts and Recruitment Managers. Its general goal is to create a comprehensive dashboard to evaluate and compare player performance across different leagues and competitions to support recruitment decisions.

"Identifying Top Talent for Recruitment" – the dashboard tells a story of player performance evaluation through multiple lenses, aiding data-driven recruitment.





45 goals

Erling Haaland

The best performing young player is Erling Haaland – 25 yo with 45 goals and 6 assists.

THANK YOU

Thanks For Your Attention