Data Warehouse Documentation

Premier League Betting Odds - Analysis

Szymon Matuszewski
Politechnika Warszawska

Michał Mazuryk
Politechnika Warszawska

Damian Skowroński
Politechnika Warszawska

May 2023

1 Introduction

1.1 Project Purpose

The purpose of this project is to construct a data warehouse model to facilitate comprehensive analysis of Premier League matches with a specific focus on betting odds.

The growing popularity of football betting has created a demand for in-depth analysis and prediction models. This project aims to meet this demand by creating a data warehouse that combines data on match results, fixtures, and betting odds from multiple sources. This consolidated data platform will allow for a wide range of analyses to be conducted, such as examining the relationship between match outcomes and betting odds, identifying trends over time, and being used for predicting future match results.

The data warehouse model will be based on a star schema, which will allow for efficient querying and easy understanding of the data structure. It will contain two fact tables: one for match facts and another for betting odds facts. These fact tables will be connected through several shared dimensional tables, such as teams, seasons, gameweeks, dates, time and referees.

The match facts table will contain data on match results, including the home and away teams, scores, expected goals (xG), and attendance. The betting odds facts table will contain data on betting odds from one betting company Bet365. It is known to be the most famous betting company in England.

By building this data warehouse, we aim to provide a robust data foundation for advanced analytics, supporting better decision-making for betting enthusiasts, football analysts, and the wider sports industry. The ultimate goal is to enhance understanding of Premier League matches and the betting landscape, contributing to more accurate predictions and insights.

1.2 User Benefits

Here we present the user benefits divided into different type of users:

- **Betting Enthusiasts** - They will be able to leverage this solution to make more informed decisions on their betting strategies. The data warehouse provides a consolidated and easy-to-analyze view of match outcomes and betting odds. This could help bettors find patterns or trends that could potentially increase their chances of successful betting.

- **Sports Analysts** - Analysts could use this tool to dig deeper into match statistics and betting trends, enabling them to generate more accurate predictions and insightful match previews or post-match analysis. The ability to easily analyze and compare historical data across seasons can support the development of sophisticated predictive models.

- **Journalists and Bloggers** - Those who report or write about football matches could use this warehouse to quickly fetch reliable stats and facts for their articles, enriching their content and enhancing the reliability of their reporting.

- **Fantasy Football Players** - Players could use the information from the data warehouse to make more informed decisions when selecting players for their fantasy teams, based on the in-depth match and player statistics.

- **Betting Companies** - They can use the data warehouse to analyze their betting odds against actual match outcomes. This could help in refining their odds setting algorithms and risk management practices, leading to more profitable operations.

- **Football Clubs and Organizations** - The data warehouse can serve as a valuable tool for performance analysis and opponent scouting. Clubs can analyze their own performance over time or delve into the performance of opponents to prepare for future matches.

- **Sports Marketing Companies** - These companies could use the insights from the data warehouse to better understand fan behavior, market trends, and more, which can guide marketing strategies in the sports domain.

2 Datasets

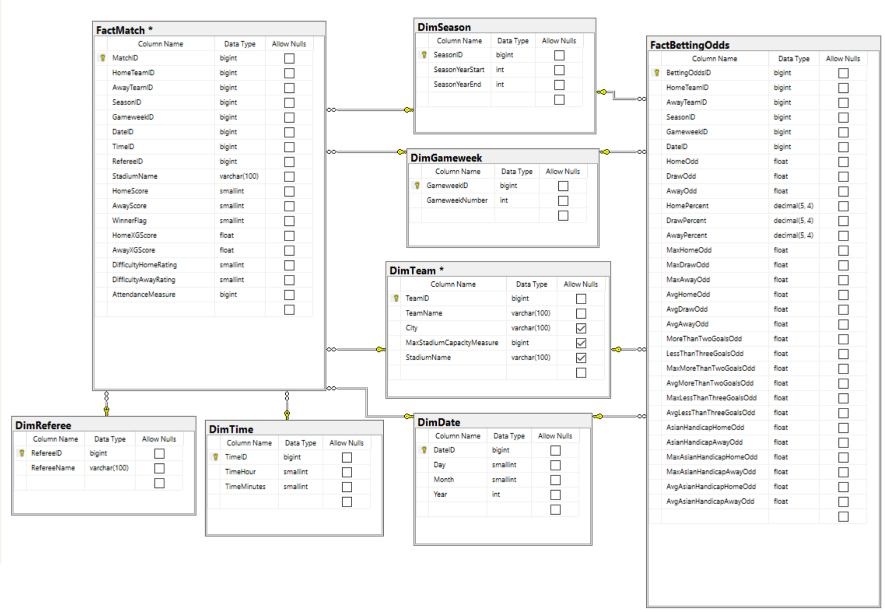
After a long consideration we decided to gather data from **three** different resources. These are respectively:

1. Scores and Fixtures - we consider last 5 seasons and gather 5 tables (each one reffering to one season). These tables contain information about matches - their outcome, expected goals (xG), attendance and more.

2. Betting Archives - once again we are forced to gather 5 tables from last 5 seasons. These tables are used to create fact table *BettingOdds* with historical information about the Bet365 odds.
3. Fantasy Premier League Github repository - these datasets are created and updated by Github user **vaastav**. They are updated regularly after one gameweek. These tables are crucial for extracting the information about the opponent difficulty.

All data is accesible in **.csv** format and being updated at least once in a gameweek.

3 Model Architecture



FactMatch *

	Column Name	Data Type	Allow Nulls
🔑	MatchID	bigint	□
	HomeTeamID	bigint	□
	AwayTeamID	bigint	□
	SeasonID	bigint	□
	GameweekID	bigint	□
	DateID	bigint	□
	TimeID	bigint	□
	RefereeID	bigint	□
	StadiumName	varchar(100)	□
	HomeScore	smallint	□
	AwayScore	smallint	□
	WinnerFlag	smallint	□
	HomeXGScore	float	□
	AwayXGScore	float	□
	DifficultyHomeRating	smallint	□
	DifficultyAwayRating	smallint	□
	AttendanceMeasure	bigint	□
			□

DimSeason

	Column Name	Data Type	Allow Nulls
🔑	SeasonID	bigint	□
	SeasonYearStart	int	□
	SeasonYearEnd	int	□
			□

DimGameweek

	Column Name	Data Type	Allow Nulls
🔑	GameweekID	bigint	□
	GameweekNumber	int	□
			□

DimTeam *

	Column Name	Data Type	Allow Nulls
🔑	TeamID	bigint	□
	TeamName	varchar(100)	□
	City	varchar(100)	☑
	MaxStadiumCapacityMeasure	bigint	☑
	StadiumName	varchar(100)	☑
			□

FactBettingOdds

	Column Name	Data Type	Allow Nulls
🔑	BettingOddsID	bigint	□
	HomeTeamID	bigint	□
	AwayTeamID	bigint	□
	SeasonID	bigint	□
	GameweekID	bigint	□
	DateID	bigint	□
	HomeOdd	float	□
	DrawOdd	float	□
	AwayOdd	float	□
	HomePercent	decimal(5, 4)	□
	DrawPercent	decimal(5, 4)	□
	AwayPercent	decimal(5, 4)	□
	MaxHomeOdd	float	□
	MaxDrawOdd	float	□
	MaxAwayOdd	float	□
	AvgHomeOdd	float	□
	AvgDrawOdd	float	□
	AvgAwayOdd	float	□
	MoreThanTwoGoalsOdd	float	□
	LessThanThreeGoalsOdd	float	□
	MaxMoreThanTwoGoalsOdd	float	□
	AvgMoreThanTwoGoalsOdd	float	□
	MaxLessThanThreeGoalsOdd	float	□
	AvgLessThanThreeGoalsOdd	float	□
	AsianHandicapHomeOdd	float	□
	AsianHandicapAwayOdd	float	□
	MaxAsianHandicapHomeOdd	float	□
	MaxAsianHandicapAwayOdd	float	□
	AvgAsianHandicapHomeOdd	float	□
	AvgAsianHandicapAwayOdd	float	□
			□

DimReferee

	Column Name	Data Type	Allow Nulls
🔑	RefereeID	bigint	□
	RefereeName	varchar(100)	□
			□

DimTime

	Column Name	Data Type	Allow Nulls
🔑	TimeID	bigint	□
	TimeHour	smallint	□
	TimeMinutes	smallint	□
			□

DimDate

	Column Name	Data Type	Allow Nulls
🔑	DateID	bigint	□
	Day	smallint	□
	Month	smallint	□
	Year	int	□
			□

Figure 1: Diagram of the model architecture for Premier League Betting Odds - Analysis.

3.1 FactMatch

Fact table containing information about the particular Premier League match. This fact table refers to the measures *after* the match.

Columns:

- MatchID - Primary Key

- HomeTeamID - Foreign Key for Home Team reffering to the DimTeam table
- AwayTeamID - Foreign Key for Away Team reffering to the DimTeam table
- SeasonID - Foreign Key for season referring to the DimSeason table
- GameweekID - Foreign Key for gameweek referring to the DimGameweek table
- DateID - Foreign Key for date referring to the DimDate table
- TimeID - Foreign Key for time referring to the DimTime table
- RefereeID - Foreign Key for referee referring to the DimReferee table
- StadiumName - name of the stadium the match was held
- HomeScore - goals measure for the home team
- AwayScore - goals measure for the away team
- WinnerFlag - calculated flag describing who won the match: 0 - draw, 1 - home team, 2 - away team
- HomeXGScore - expected goals measure for the home team
- AwayXGScore - expected goals measure for the away team
- DifficultyHomeRating - difficulty rating for home team (1 - very weak, 5 - very strong)
- DifficultyAwayRating - difficulty rating for away team (1 - very weak, 5 - very strong)
- AttendanceMeasure - measure for attendance at the stadium

3.2 FactBettingOdds

Fact table containing information about Bet365 betting odds for a particular match. This fact table refers to the measures *before* the match.

Columns:

- BettingOddsID - Primary Key
- HomeTeamID - Foreign Key for Home Team reffering to the DimTeam table
- AwayTeamID - Foreign Key for Away Team reffering to the DimTeam table

- SeasonID - Foreign Key for season referring to the DimSeason table
- GameweekID - Foreign Key for gameweek referring to the DimGameweek table
- DateID - Foreign Key for date referring to the DimDate table
- HomeOdd - the odd for the home team at Bet365
- DrawOdd - the odd for a draw at Bet365
- AwayOdd - the odd for the away team at Bet365
- HomePercent - calculated measure: percent for home team to be the winners
- DrawPercent - calculated measure: percent for the draw
- AwayPercent - calculated measure: percent for away team to be the winners
- MaxHomeOdd - maximum odd for the home team from different companies
- MaxDrawOdd - maximum odd for the draw from different companies
- MaxAwayOdd - maximum odd for the away team from different companies
- AvgHomeOdd - average odd for the home team from different companies
- AvgDrawOdd - average odd for the draw from different companies
- AvgAwayOdd - average odd for the away team from different companies
- MoreThanTwoGoalsOdd - the odd for more than 2 goals to be scored within the match
- LessThanThreeGoalsOdd - the odd for less than 3 goals to be scored within the match
- MaxMoreThanTwoGoalsOdd - maximum odd for more than 2 goals to be scored within the match from different companies
- AvgMoreThanTwoGoalsOdd - average odd for more than 2 goals to be scored within the match from different companies
- MaxLessThanThreeGoalsOdd - maximum odd for more than 3 goals to be scored within the match from different companies
- AvgLessThanThreeGoalsOdd - average odd for more than 3 goals to be scored within the match from different companies
- AsianHandicapHomeOdd - the odd for the home team with asian handicap

- AsianHandicapAwayOdd - the odd for the away team with asian handicap
- MaxAsianHandicapHomeOdd - maximum odd for the home team with asian handicap from different companies
- MaxAsianHandicapAwayOdd - maximum odd for the away team with asian handicap from different companies
- AvgAsianHandicapHomeOdd - average odd for the home team with asian handicap from different companies
- AvgAsianHandicapAwayOdd - average odd for the away team with asian handicap from different companies

3.3 DimTeam

Dimensional table containing information about teams.

Columns:

- TeamID - PrimaryKey
- TeamName - name of the team
- City - origin city of the team
- MaxStadiumCapacityMeasure - maximum stadium capacity gathered manually
- StadiumName - name of the stadium gathered manually
- ShortTeamName - 3 letter short team name for merging Matches with Bets (*) gathered manually

3.4 DimSeason

Dimensional table containing information about seasons.

Columns:

- SeasonID - Primary Key
- SeasonStartYear - year of the start of the season
- SeasonEndYear - year of the end of the season

3.5 DimDate

Dimensional table containing information about dates.

Columns:

- DateID - Primary Key
- Day - day of the match
- Month - month of the match
- Year - year of the match

3.6 DimGameweek

Dimensional table containing information about gameweeks.

Columns:

- GameweekID - Primary Key
- GameweekNumber - gameweek number from 1 to 38

3.7 DimReferee

Dimensional table containing information abour referees.

Columns:

- RefereeID - Primary Key
- RefereeName - name of the referee

3.8 DimTime

Dimensional table containing information about an hour the match was held.

Columns:

- TimeID - Primary Key
- TimeHour - an hour of the time when the match was held
- TimeMinutes - minutes of the time when the match was held

4 Creating Database

We use python script for creating database to store row csv tables.
The insertScoresFixtures() function reads a CSV file containing football match results and cleans it up. The function removes unnecessary columns, renames some columns, and splits the Score column into HomeScore and AwayScore columns. Finally, it writes the processed data into the Matches table of the

SQL Server database.

The insertBettingOdds() function reads a CSV file containing betting odds for football matches and cleans it up. The function removes unnecessary columns, renames some columns, and converts the Date column into the datetime format. Finally, it writes the processed data into the BettingOdds table of the SQL Server database.

The insertFantasyFixtures() function reads two CSV files, one containing team names and another containing football fixtures, and cleans them up. The function merges the two dataframes, renames some columns, and converts the kickoff-time column into the datetime format. Finally, it writes the processed data into the FantasyFixtures table of the SQL Server database.

The main() function creates a connection to the SQL Server database, gets the file paths of the CSV files to be processed, and calls the three data processing functions for each file. The script handles files for multiple seasons of the English Premier League.

5 ETL

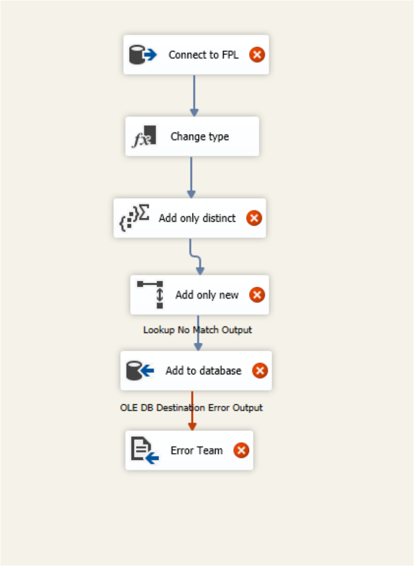


Figure 2: ETL process for DimTeam.

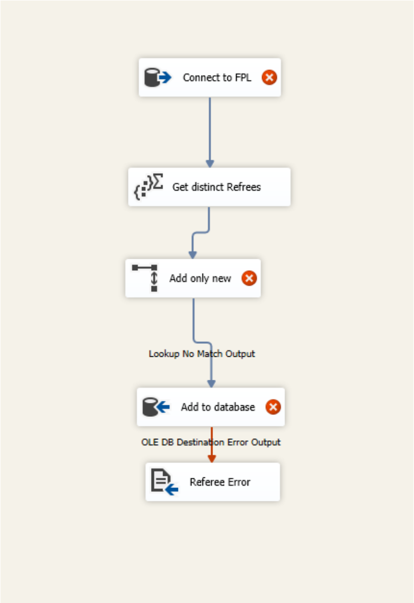


Figure 3: ETL process for DimReferee, similiar process for DimGameweek.

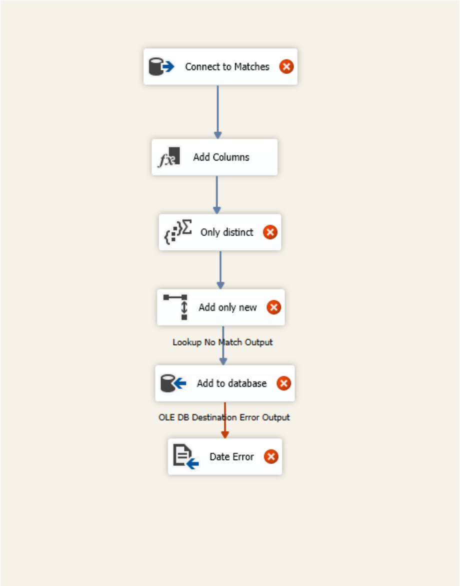


Figure 4: ETL process for DimDate, similiar process for DimSeason, DimTime.

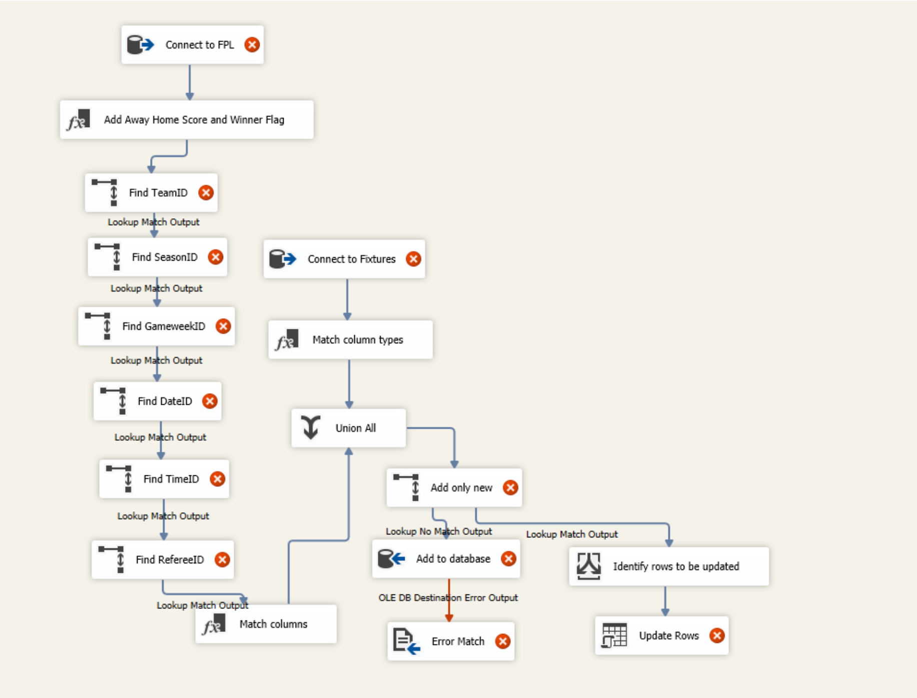


Figure 5: ETL process for FactMatch.

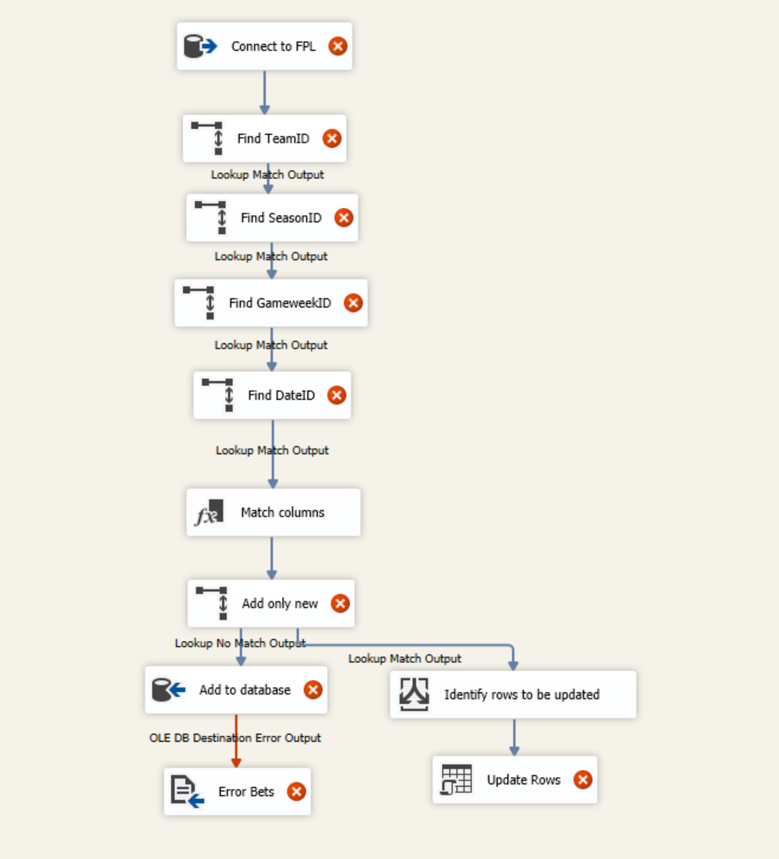


Figure 6: ETL process for FactBettingOdds.

6 Planned Raports

Issues to be visualised in SAS:

- distribution of how many goals are scored,
- distribution of home team odds for winning,
- comparing difficulty of the team to winning games and odds for winning,
- checking games that team with lower odds for winning won,
- checking type of bets that often are correct,
- comparing referees to goals scored, home team winning, type of odds,

- fan attendance and score of the game, team difficulty, is from the same city,
- betting odds changes over the years.

7 Planned Tests

Validate control flows by checking:

- if numer of added rows is correct,
- whole row for a few exemplary matches,
- changing row of the table and checking if its updated,
- adding a row to the table and checking if its inserted.