Data Warehouse Documentation
Premier League Betting Odds - Analysis

Szymon Matuszewski
Politechnika Warszawska

Michał Mazuryk
Politechnika Warszawska

Damian Skowroński
Politechnika Warszawska

May 2023

Contents

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

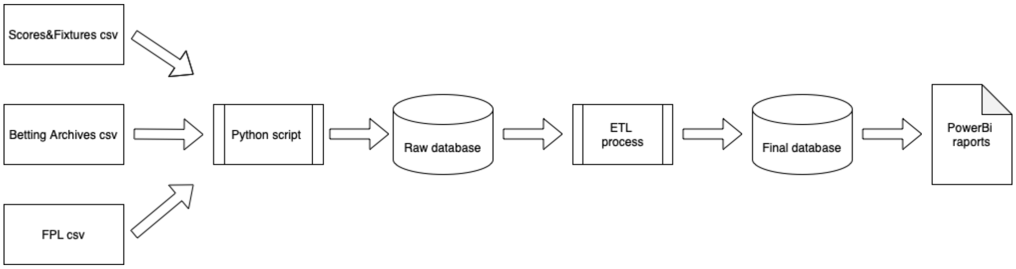
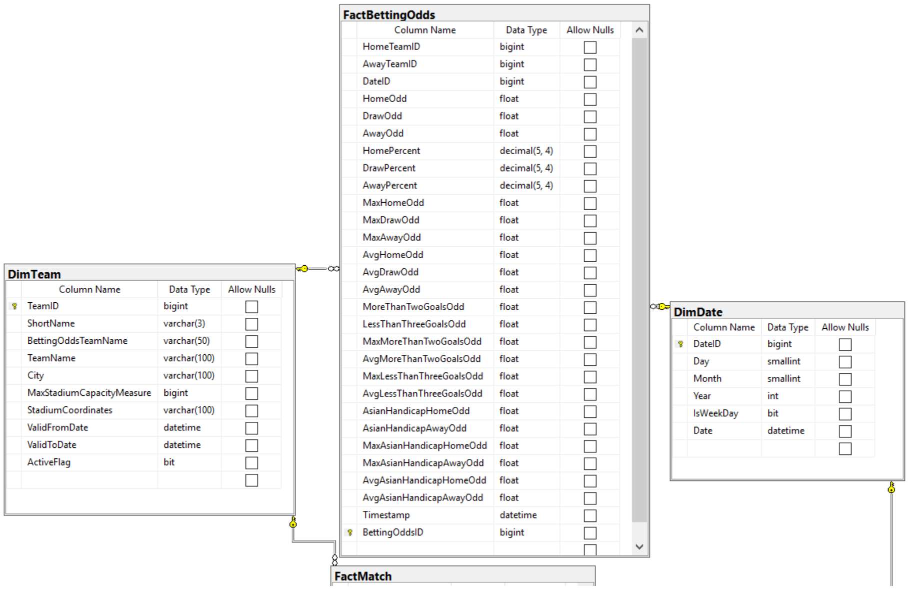


Figure 1: Diagram of the whole process.



FactBettingOdds

	Column Name	Data Type	Allow Nulls
	HomeTeamID	bigint	□
	AwayTeamID	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	□
	Timestamp	datetime	□
🔑	BettingOddsID	bigint	□
			□

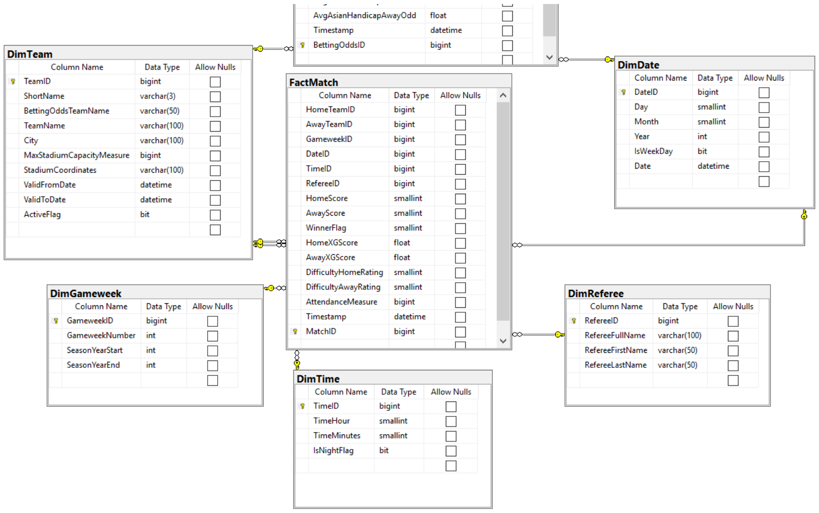
DimTeam

	Column Name	Data Type	Allow Nulls
🔑	TeamID	bigint	□
	ShortName	varchar(3)	□
	BettingOddsTeamName	varchar(50)	□
	TeamName	varchar(100)	□
	City	varchar(100)	□
	MaxStadiumCapacityMeasure	bigint	□
	StadiumCoordinates	varchar(100)	□
	ValidFromDate	datetime	□
	ValidToDate	datetime	□
	ActiveFlag	bit	□
			□

DimDate

	Column Name	Data Type	Allow Nulls
🔑	DateID	bigint	□
	Day	smallint	□
	Month	smallint	□
	Year	int	□
	IsWeekDay	bit	□
	Date	datetime	□
			□

Figure 2: Diagram of the model architecture for Premier League Betting Odds Analysis part 1.



	AvgAsianHandicapAwayOdd	float	□
	Timestamp	datetime	□
🔑	BettingOddsID	bigint	□
			□

DimTeam

	Column Name	Data Type	Allow Nulls
🔑	TeamID	bigint	□
	ShortName	varchar(3)	□
	BettingOddsTeamName	varchar(50)	□
	TeamName	varchar(100)	□
	City	varchar(100)	□
	MaxStadiumCapacityMeasure	bigint	□
	StadiumCoordinates	varchar(100)	□
	ValidFromDate	datetime	□
	ValidToDate	datetime	□
	ActiveFlag	bit	□
			□

FactMatch

	Column Name	Data Type	Allow Nulls
	HomeTeamID	bigint	□
	AwayTeamID	bigint	□
	GameweekID	bigint	□
	DateID	bigint	□
	TimeID	bigint	□
	RefereeID	bigint	□
	HomeScore	smallint	□
	AwayScore	smallint	□
	WinnerFlag	smallint	□
	HomeXGScore	float	□
	AwayXGScore	float	□
	DifficultyHomeRating	smallint	□
	DifficultyAwayRating	smallint	□
	AttendanceMeasure	bigint	□
	Timestamp	datetime	□
🔑	MatchID	bigint	□
			□

DimDate

	Column Name	Data Type	Allow Nulls
🔑	DateID	bigint	□
	Day	smallint	□
	Month	smallint	□
	Year	int	□
	IsWeekDay	bit	□
	Date	datetime	□
			□

DimGameweek

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

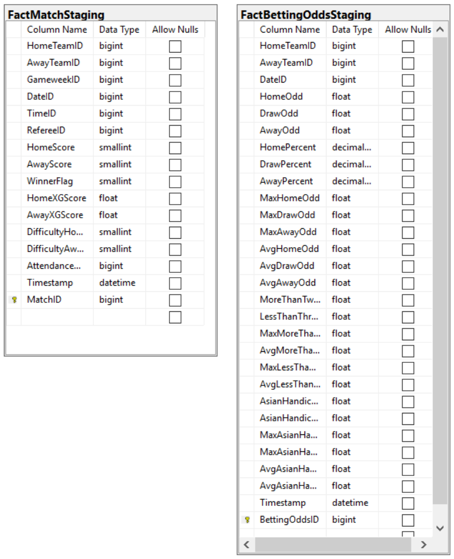
DimReferee

	Column Name	Data Type	Allow Nulls
🔑	RefereeID	bigint	□
	RefereeFullName	varchar(100)	□
	RefereeFirstName	varchar(50)	□
	RefereeLastName	varchar(50)	□
			□

DimTime

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

Figure 3: Diagram of the model architecture for Premier League Betting Odds Analysis part 2.



FactMatchStaging

Column Name	Data Type	Allow Nulls
HomeTeamID	bigint	□
AwayTeamID	bigint	□
GameweekID	bigint	□
DateID	bigint	□
TimeID	bigint	□
RefereeID	bigint	□
HomeScore	smallint	□
AwayScore	smallint	□
WinnerFlag	smallint	□
HomeXGScore	float	□
AwayXGScore	float	□
DifficultyHo...	smallint	□
DifficultyAw...	smallint	□
Attendance...	bigint	□
Timestamp	datetime	□
MatchID	bigint	□
		□

FactBettingOddsStaging

Column Name	Data Type	Allow Nulls
HomeTeamID	bigint	□
AwayTeamID	bigint	□
DateID	bigint	□
HomeOdd	float	□
DrawOdd	float	□
AwayOdd	float	□
HomePercent	decimal...	□
DrawPercent	decimal...	□
AwayPercent	decimal...	□
MaxHomeOdd	float	□
MaxDrawOdd	float	□
MaxAwayOdd	float	□
AvgHomeOdd	float	□
AvgDrawOdd	float	□
AvgAwayOdd	float	□
MoreThanTw...	float	□
LessThanThr...	float	□
MaxMoreTha...	float	□
AvgMoreTha...	float	□
MaxLessTha...	float	□
AvgLessThan...	float	□
AsianHandic...	float	□
AsianHandic...	float	□
MaxAsianHa...	float	□
MaxAsianHa...	float	□
AvgAsianHa...	float	□
AvgAsianHa...	float	□
Timestamp	datetime	□
BettingOddsID	bigint	□

Figure 4: Diagram of the model architecture for Premier League Betting Odds Analysis showing staging tables.

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 created as a combination of DateID, HomeTeamID and AwayTeamID
- HomeTeamID - Foreign Key for Home Team reffering to the DimTeam table
- AwayTeamID - Foreign Key for Away Team reffering to the DimTeam 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
- 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
- Timestamp - date when data was loaded to database

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 calculated as combination of DateID, HomeTeamID and AwayTeamID
- HomeTeamID - Foreign Key for Home Team reffering to the DimTeam table
- AwayTeamID - Foreign Key for Away Team reffering to the DimTeam 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
- Timestamp - date when data was loaded to database

3.3 DimTeam

Dimensional table containing information about teams.

Columns:

- TeamID - PrimaryKey
- TeamName - name of the team using match table nomenclature

- ShortName - short name of the team (3 letters)
- BettingOddsTeamName - name of the team using betting odds table nomenclature
- City - origin city of the team
- MaxStadiumCapacityMeasure - maximum stadium capacity calculated as the highest AttendanceMeasure
- StadiumCoordinates - localization of the stadium using coordinates
- ActiveFlag - 1/0 value if data about team is actual or historic
- ValidFromDate - date of uploading row of data to database
- ValidToDate - date when changes has been made to the team (new MaxStadiumCapacityMeasure) or max possible date (9999-12-31)

3.4 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
- IsWeekDay - is from monday to friday
- Date - date as datetime

3.5 DimGameweek

Dimensional table containing information about gameweeks.

Columns:

- GameweekID - Primary Key
- GameweekNumber - gameweek number from 1 to 38
- SeasonYearStart - year of the start of the season calculated from date of the match
- SeasonYearEnd - year of the end of the season calculated from date of the match

3.6 DimReferee

Dimensional table containing information abour referees.

Columns:

- RefereeID - Primary Key
- RefereeFullName - name of the referee
- RefereeFirstName - first name of the referee
- RefereeLastName - last name of the referee

3.7 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
- IsNightFlag - calculated column is hour later than 19

4 Database Creation

To store data in its raw form, we created a database using the script *RawData_create_database.sql*. The script creates a database with four empty tables:

- Matches - general information about each match.
- BettingOdds - about 100 columns of information regarding the betting for each match.
- FantasyFixtures - information regarding the strength of both teams participating in a match. This data is taken from an official Premier League game, and changes throughout the season depending on the outcomes.
- Teams - some information about the team.

To populate the database, we used a Python script *insert_csv.py*. The script, in its `main()` function, connects to the database, then reads CSV files, combines them, performs minor data transformations, and inserts the tables into the database. The script handles files for multiple seasons of the English Premier League. The functions used by the script are described in the following subsections.

4.1 insert_scores_fixtures

The `insertScoresFixtures()` function reads a CSV file containing football match results and cleans it up. This function uses data from files that follow the naming pattern: *scores-fixtures-<season years>.csv*. 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.

4.2 insert_betting_odds

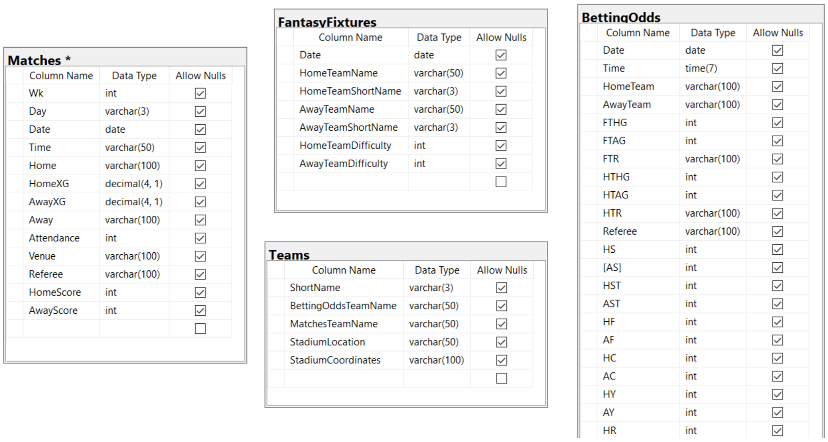
The `insert_betting_odds()` function reads a CSV file containing betting odds for football matches and cleans it up. This function uses data from files that follow the naming pattern: *betting-odds-<season years>.csv*. The function removes unnecessary columns, renames some columns, and converts the Date column into the datetime format. Finally, it writes the processed data into the BettingOdds table of the SQL Server database.

4.3 insert_fantasy_fixtures

The `insert_fantasy_fixtures()` function reads two CSV files, one containing team names (file name pattern: *fantasy-teams-<season years>.csv*) and another containing football fixtures (file name pattern: *fantasy-fixtures-<season years>.csv*), 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.

4.4 create_teams_dict

The `create_teams_dict()` function generates a dictionary of teams for a given season. The data regarding teams from various sources use different names for the teams. The purpose of this function is to create a table that will contain the names of individual teams from other tables, along with some information about each team. In addition, the function retrieves information about the stadium where each team plays from this Wikipedia article. Finally, it writes the processed data into the Teams table of the SQL Server database.



Matches *

Column Name	Data Type	Allow Nulls
Wk	int	☑
Day	varchar(3)	☑
Date	date	☑
Time	varchar(50)	☑
Home	varchar(100)	☑
HomeXG	decimal(4, 1)	☑
AwayXG	decimal(4, 1)	☑
Away	varchar(100)	☑
Attendance	int	☑
Venue	varchar(100)	☑
Referee	varchar(100)	☑
HomeScore	int	☑
AwayScore	int	☑
		☐

FantasyFixtures

Column Name	Data Type	Allow Nulls
Date	date	☑
HomeTeamName	varchar(50)	☑
HomeTeamShortName	varchar(3)	☑
AwayTeamName	varchar(50)	☑
AwayTeamShortName	varchar(3)	☑
HomeTeamDifficulty	int	☑
AwayTeamDifficulty	int	☑
		☐

Teams

Column Name	Data Type	Allow Nulls
ShortName	varchar(3)	☑
BettingOddsTeamName	varchar(50)	☑
MatchesTeamName	varchar(50)	☑
StadiumLocation	varchar(50)	☑
StadiumCoordinates	varchar(100)	☑
		☐

BettingOdds

Column Name	Data Type	Allow Nulls
Date	date	☑
Time	time(7)	☑
HomeTeam	varchar(100)	☑
AwayTeam	varchar(100)	☑
FTHG	int	☑
FTAG	int	☑
FTR	varchar(100)	☑
HTHG	int	☑
HTAG	int	☑
HTR	varchar(100)	☑
Referee	varchar(100)	☑
HS	int	☑
[AS]	int	☑
HST	int	☑
AST	int	☑
HF	int	☑
AF	int	☑
HC	int	☑
AC	int	☑
HY	int	☑
AY	int	☑
HR	int	☑

Figure 5: Diagram of raw database.

5 ETL

ETL processes where split to two SSIS Packages, seprate one for dimensions and facts tables. To load new data for the data warehouse firstly dimensions package should be executed then facts package.

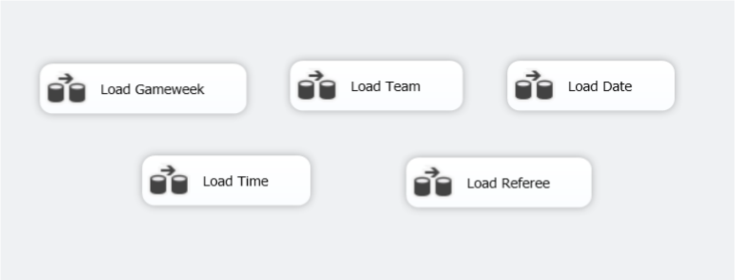


Figure 6: Diagram of etl process for dimensions.

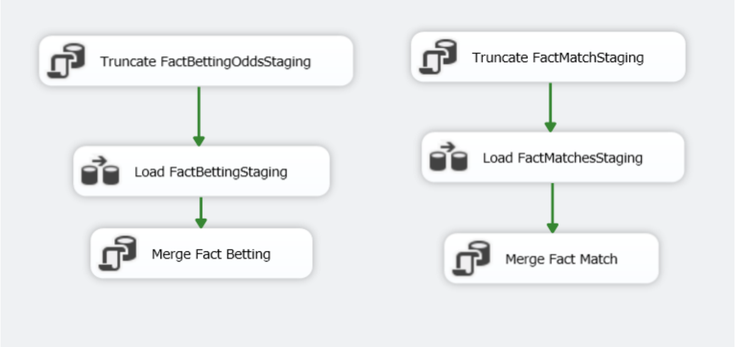


Figure 7: Diagram of etl process for facts.

5.1 Load Gameweek

Process for DimGameweek table.

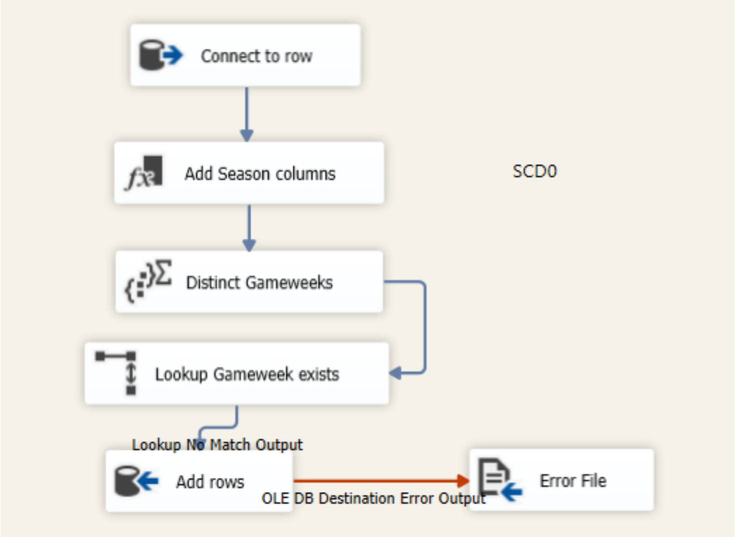


Figure 8: Diagram of etl process for DimGameweek table.

States:

- Connect to raw - OLE DB Source connection to raw database, Matches table accessing GameweekNumber and Date column
- Add season columns - calculating column SeasonYearStart and SeasonYearEnd based of date of the match

- Distinct Gameweeks - grouping based of Number of gameweek and SeasonYear columns
- Lookup Gameweek exists - check if there is already this gameweek in the database
- Add rows - if it isn't already in the database add it to DimGameweek
- Error file - if there is problem save error output to the file

5.2 Load Time

Process for DimTime table.

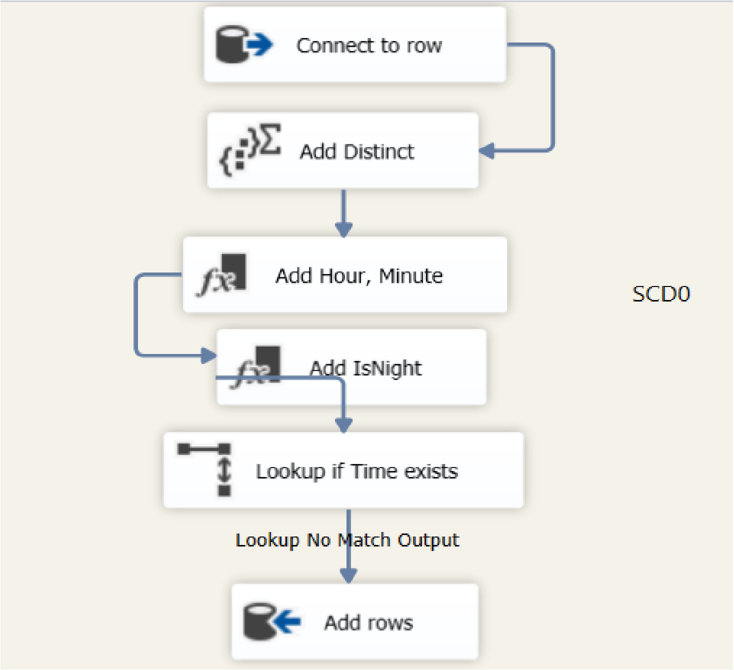


Figure 9: Diagram of etl process for DimTime table.

States:

- Connect to raw - OLE DB Source connection to raw database, Matches table accessing Time column
- Add Distinct - grouping based of Time column
- Add hour, minute - calculating columns TimeHour and TimeMinute based on Time column

- Add IsNight - calculating column IsNightFlag based on TimeHour
- Lookup if Time exists - check if there is already this time row in the database
- Add rows - if it isn't already in the database add it to DimTime

5.3 Load Referee

Process for DimReferee table.

States:

- Connect to raw - OLE DB Source connection to raw database, Matches table accessing RefereeName column
- Get distinct Referees - grouping based of RefereeName column
- Add columns - creating column FirstName and LastName based of RefereeName column
- Add only new - check if there is already this referee row in the database
- Add to database - if it isn't already in the database add it to DimReferee

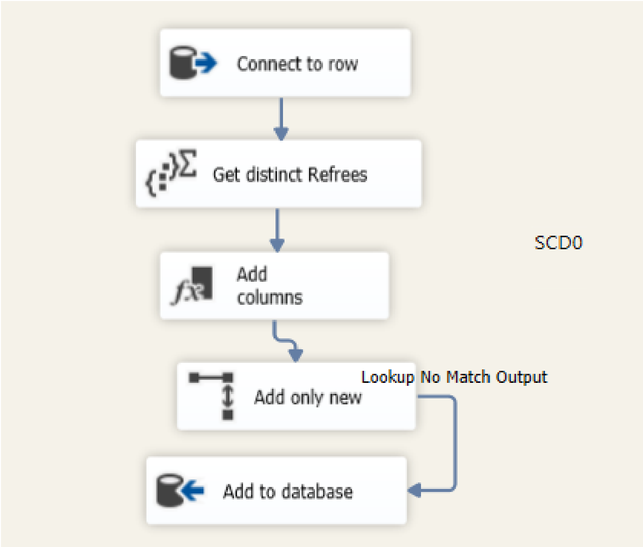


Figure 10: Diagram of etl process for DimReferee table.

5.4 Load Date

Process for DimDate table.
States:

- Connect to Matches - OLE DB Source connection to raw database, Matches table accessing Date column
- Only distinct - grouping based of Date column
- Add Columns - creating column Year, Month, Day and IsWeekDay
- Add only new - check if there is already this date row in the database
- Add to database - if it isn't already in the database add it to DimDate
- Error File - direct error to file

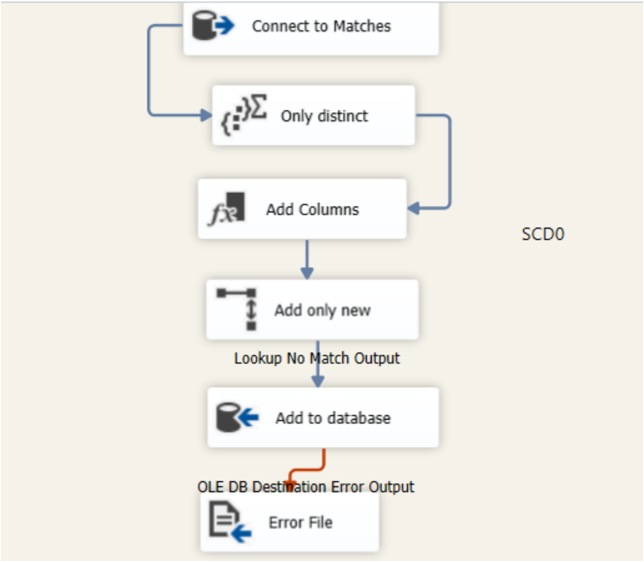


Figure 11: Diagram of etl process for DimDate table.

5.5 Load Team

Process for DimTeam table with SCD2 for changing MaxStadiumCapacity.

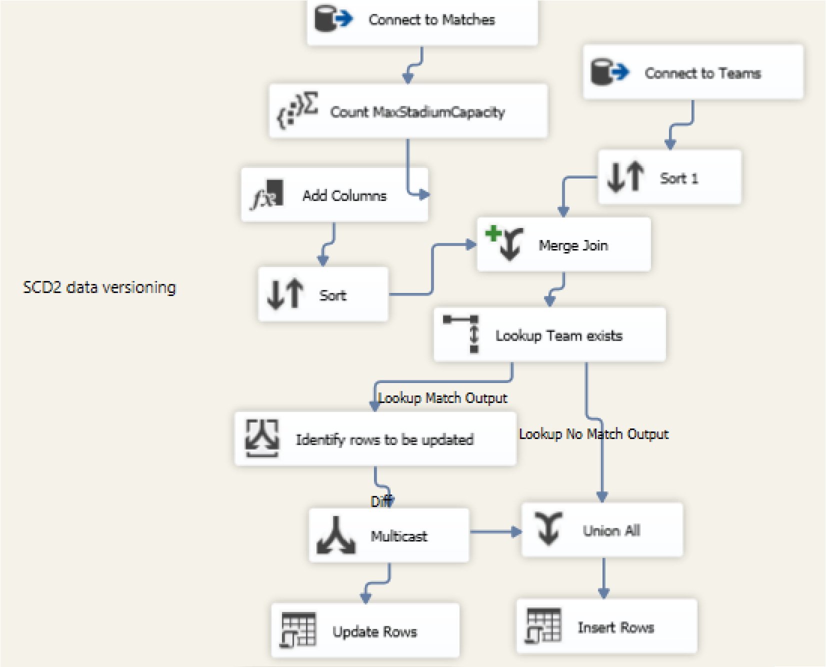


Figure 12: Diagram of etl process for DimTeam table.

States:

- Connect to Matches - OLE DB Source connection to raw database, Matches table accessing Home (Team Name) and Attendance column
- Count MaxStadiumCapacity - group by home team and get column MaxAttendance as maximum from Attendance
- Add Columns - create Timestamp
- Sort, Sort 1 - sort by Team
- Connect to Teams - OLE DB Source connection to raw database, Teams table
- Merge Join - joining columns from both sources
- Lookup Team exists - check if there is already this team row in the database
- Identify rows to be updated - check if there is new value for MaxStadiumCapacity

- Update rows - change ActiveFlag to 0 and ValidToDate to current date
- Union All - connect rows from this to be updated and new teams
- Insert Rows - add new rows with ActiveFlag 1 and ValidToDate to max possible Date

5.6 Load FactMatches

Loading FactMatches is split to 3 stages. First 'Truncate FactMatchStaging' clears staging table, then 'Load FactMatchesStaging' and in the end using Execute SQL Task merge with FactMatches based on calculated ID.

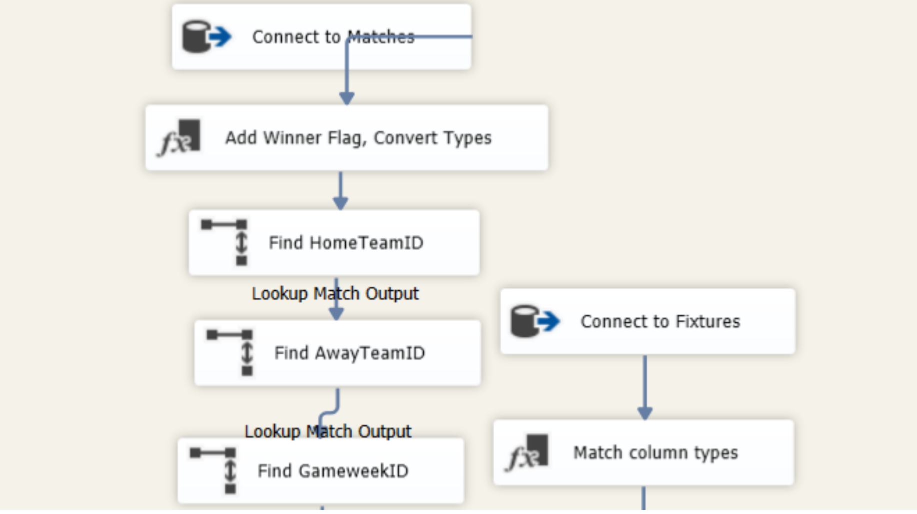


Figure 13: Diagram of etl process for FactMatches staging table part 1.

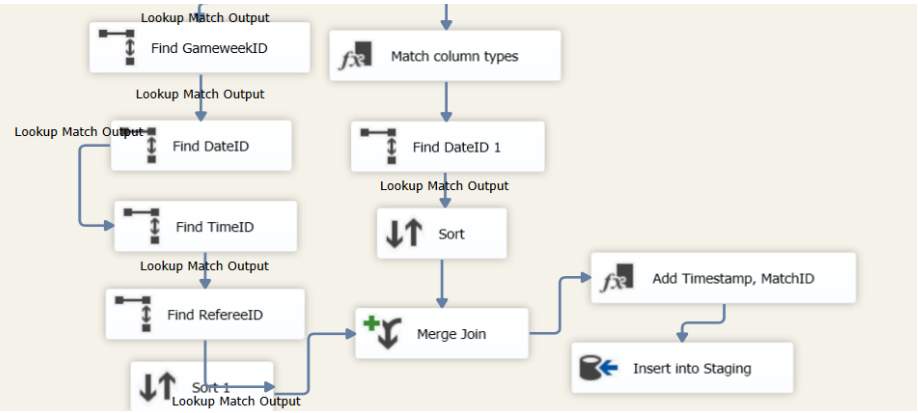


Figure 14: Diagram of etl process for FactMatches staging table part 2.

States:

- Connect to Matches - OLE DB Source connection to raw database, Matches table
- Add Winner Flag, Convert Types - create column WinnerFlag based of HomeScore and AwayScore, create columns needed for finding dimensions ID
- Find ID - by matching columns finding ID from dimension for foreign keys
- Connect to Fixtures - OLE DB Source connection to raw database, Fixtures table for Difficulty, Short Team names and Date column
- Match column types - create columns for finding DateID
- Find DateID1 - find column DateID
- Sort, Sort1 - sort based of date and short home team name for faster merging
- Merge Join - merge columns from different sources using DateID and ShortName of home team
- Add Timestamp, MatchID - add timestamp and create ID by calculating DateID, HomeTeamName and AwayTeamName
- Insert into Staging - add all rows to staging table

5.7 Load FactBettingOdds

Loading FactBettingOdds is split to 3 stages. First 'Truncate FactBettingOddsStaging' clears staging table, then 'Load FactBettingOddsStaging' and

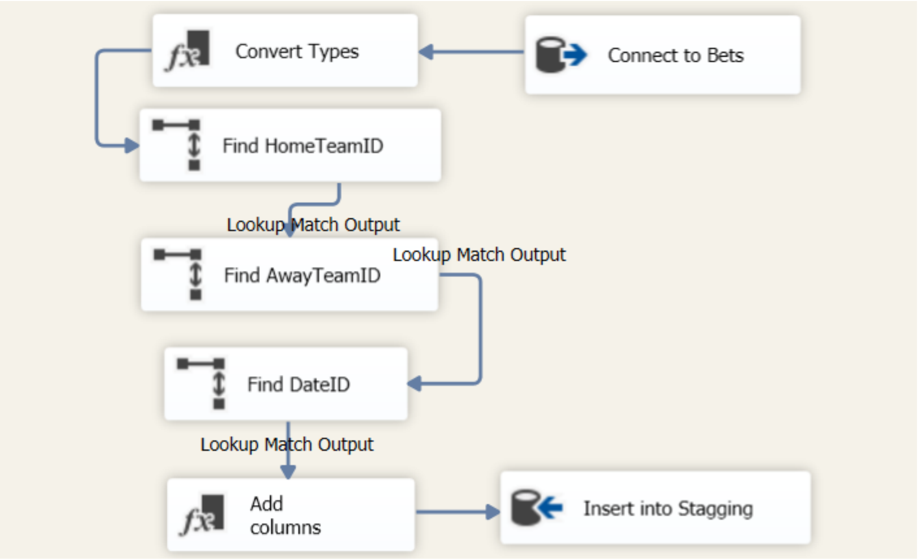


Figure 15: Diagram of etl process for FactBettingOdds staging table.

in the end using Execute SQL Task merge with FactBettingOdds based on calculated ID.

States:

- Connect to Matches - OLE DB Source connection to raw database, BettingOdds table
- Convert Types - convert types and create columns to find DateID
- Find Team ID - by matching columns BettingOddsTeamName from DimTeams find TeamID
- Find DateID - find foreign key, DateID
- Add columns - add timestamp, HomeProcent, AwayProcent, DrawProcent and calculate BettingOdds ID using DateID, HomeTeamID and AwayTeamID
- Insert into Staging - add all rows to staging table

6 Tests for ETL

6.1 Validate amount of rows added

We want to check if everything works and there is correct amount of rows added to database. It dimension tables it should be the same as distinct values from raw database. In facts the same as rows in raw database.

Steps:

- Run the SSIS Packages
- Check for errors
- Check how many rows has been added to database using SQL Commends

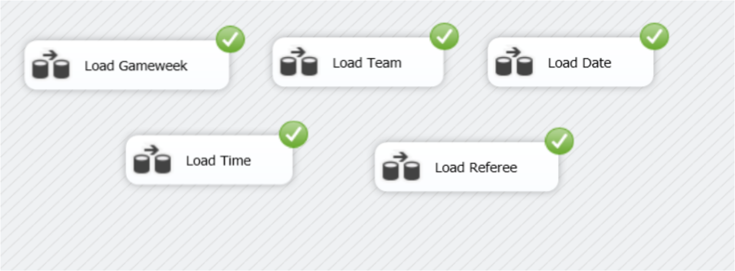


Figure 16: Diagram of etl process for dimensions.

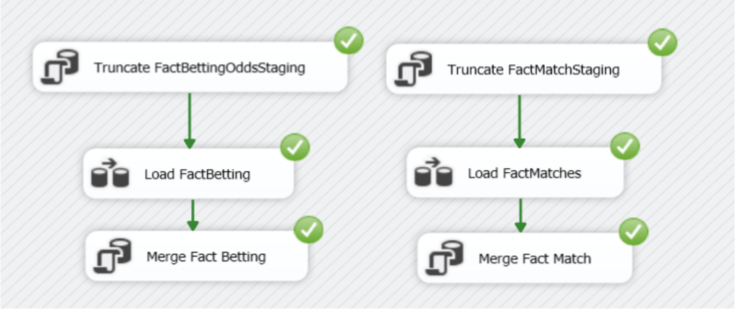
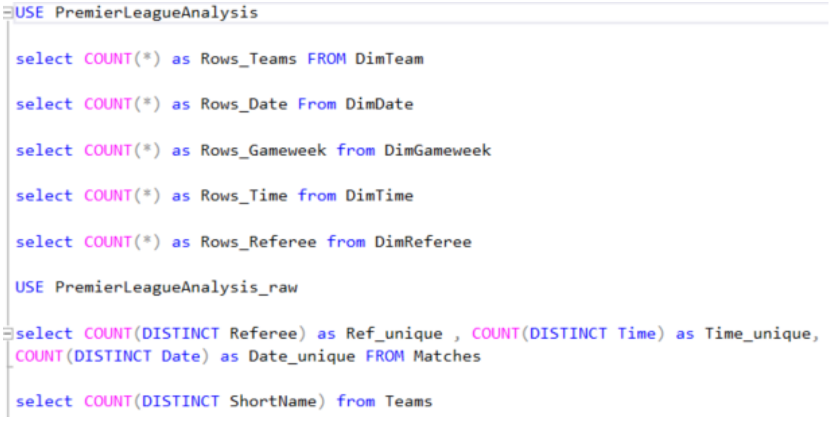
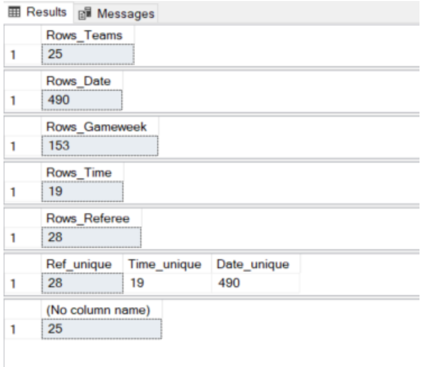


Figure 17: Diagram of etl process for facts.



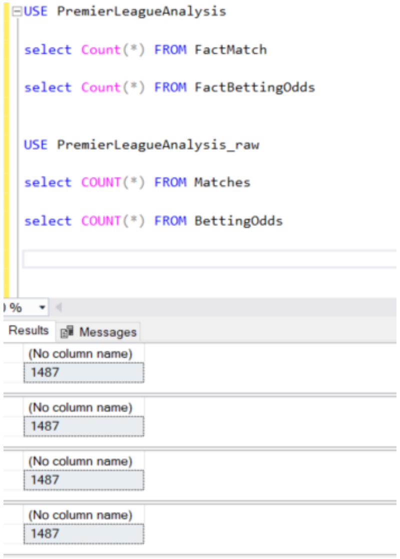
```
USE PremierLeagueAnalysis

select COUNT(*) as Rows_Teams FROM DimTeam

select COUNT(*) as Rows_Date From DimDate

select COUNT(*) as Rows_Gameweek from DimGameweek

select COUNT(*) as Rows_Time from DimTime

select COUNT(*) as Rows_Referee from DimReferee

USE PremierLeagueAnalysis_raw

select COUNT(DISTINCT Referee) as Ref_unique , COUNT(DISTINCT Time) as Time_unique,
COUNT(DISTINCT Date) as Date_unique FROM Matches

select COUNT(DISTINCT ShortName) from Teams
```

Figure 18: SQL Query used for comparing rows added from raw database to formatted one in dimension tables.



```
   Rows_Teams
1  25

   Rows_Date
1  490

   Rows_Gameweek
1  153

   Rows_Time
1  19

   Rows_Referee
1  28

   Ref_unique  Time_unique  Date_unique
1  28          19           490

   (No column name)
1  25
```

Figure 19: Execution of previous SQL Query, there is the same amount of rows added as should be, in DimGameweek should be 4 seasons * 38 gameweeks = 152 rows in gameweeks - the difference is 1 gameweek in Covid when one gameweek was split and continued after the break. It is interesting statistic so we would keep it.



```
USE PremierLeagueAnalysis

select Count(*) FROM FactMatch

select Count(*) FROM FactBettingOdds


USE PremierLeagueAnalysis_raw

select COUNT(*) FROM Matches

select COUNT(*) FROM BettingOdds
```

```
(No column name)
1487

(No column name)
1487

(No column name)
1487

(No column name)
1487
```

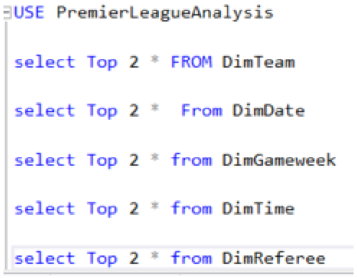
Figure 20: SQL Query and result for comparing amount of rows added, everything is correct.

6.2 Validate data types

We want to check if rows in our data look correct, the way we intended. It should be the same as in planned diagram.

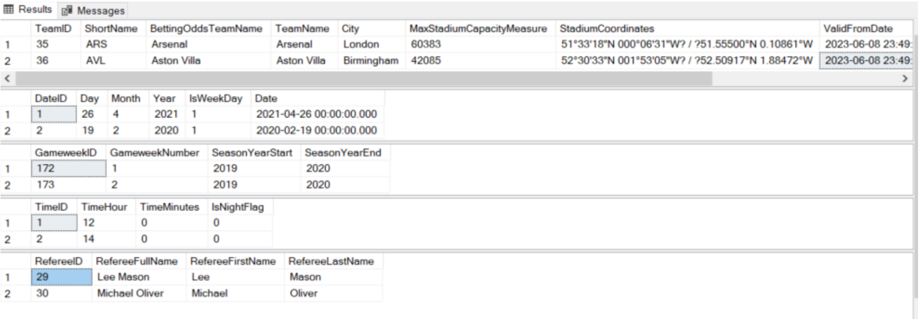
Steps:

- Use SQL Query to show top rows of tables
- Check if they are correct comparing to diagram



```
USE PremierLeagueAnalysis

select Top 2 * FROM DimTeam

select Top 2 *  From DimDate

select Top 2 * from DimGameweek

select Top 2 * from DimTime

select Top 2 * from DimReferee
```

Figure 21: SQL Query for checking top rows of dimension tables.



	TeamID	ShortName	BettingOddsTeamName	TeamName	City	MaxStadiumCapacityMeasure	StadiumCoordinates	ValidFromDate
1	35	ARS	Arsenal	Arsenal	London	60383	51°33'18"N 000°06'31"W? / ?51.55500°N 0.10861°W	2023-06-08 23:49:
2	36	AVL	Aston Villa	Aston Villa	Birmingham	42085	52°30'33"N 001°53'05"W? / ?52.50917°N 1.88472°W	2023-06-08 23:49:

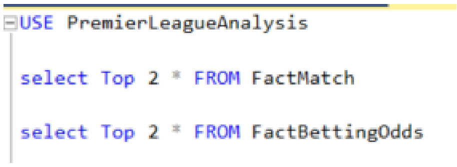
	DateID	Day	Month	Year	IsWeekDay	Date
1	1	26	4	2021	1	2021-04-26 00:00:00.000
2	2	19	2	2020	1	2020-02-19 00:00:00.000

	GameweekID	GameweekNumber	SeasonYearStart	SeasonYearEnd
1	172	1	2019	2020
2	173	2	2019	2020

	TimeID	TimeHour	TimeMinutes	IsNightFlag
1	1	12	0	0
2	2	14	0	0

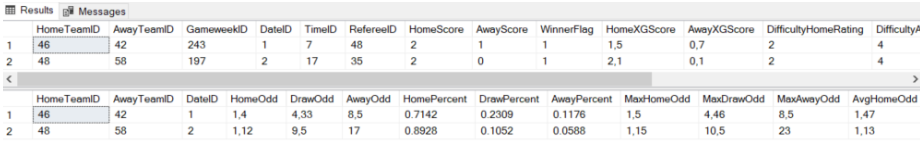
	RefereeID	RefereeFullName	RefereeFirstName	RefereeLastName
1	29	Lee Mason	Lee	Mason
2	30	Michael Oliver	Michael	Oliver

Figure 22: Result of previous SQL Query, everything is correct.



```
USE PremierLeagueAnalysis

select Top 2 * FROM FactMatch

select Top 2 * FROM FactBettingOdds
```

Figure 23: SQL Query for checking top rows of facts tables.



Results | Messages

	HomeTeamID	AwayTeamID	GameweekID	DateID	TimeID	RefereeID	HomeScore	AwayScore	WinnerFlag	HomeXGScore	AwayXGScore	DifficultyHomeRating	DifficultyA
1	46	42	243	1	7	48	2	1	1	1,5	0,7	2	4
2	48	58	197	2	17	35	2	0	1	2,1	0,1	2	4

	HomeTeamID	AwayTeamID	DateID	HomeOdd	DrawOdd	AwayOdd	HomePercent	DrawPercent	AwayPercent	MaxHomeOdd	MaxDrawOdd	MaxAwayOdd	AvgHomeOdd
1	46	42	1	1,4	4,33	8,5	0.7142	0.2309	0.1176	1,5	4,46	8,5	1,47
2	48	58	2	1,12	9,5	17	0.8928	0.1052	0.0588	1,15	10,5	23	1,13

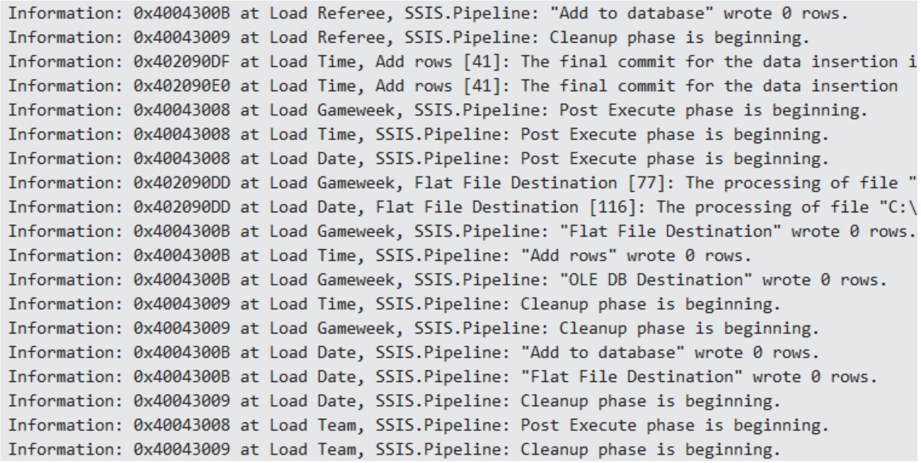
Figure 24: Result of previous SQL Query, everything is correct.

6.3 Validate another iteration

We want to check what happens if don't have new rows but run etl process one more time. The desire effect is that no rows should be added, there are the same amount of rows in final database.

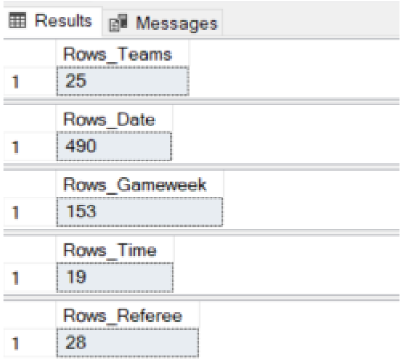
Steps:

- Execute ETL process
- Check logs for rows added
- Check database to see if there is a different amount of rows



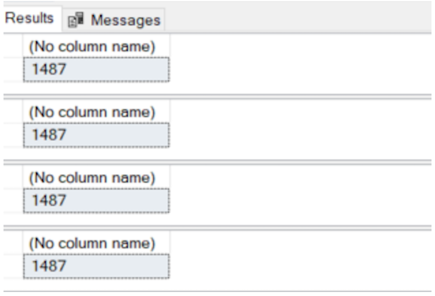
```
Information: 0x4004300B at Load Referee, SSIS.Pipeline: "Add to database" wrote 0 rows.
Information: 0x40043009 at Load Referee, SSIS.Pipeline: Cleanup phase is beginning.
Information: 0x402090DF at Load Time, Add rows [41]: The final commit for the data insertion i
Information: 0x402090E0 at Load Time, Add rows [41]: The final commit for the data insertion
Information: 0x40043008 at Load Gameweek, SSIS.Pipeline: Post Execute phase is beginning.
Information: 0x40043008 at Load Time, SSIS.Pipeline: Post Execute phase is beginning.
Information: 0x40043008 at Load Date, SSIS.Pipeline: Post Execute phase is beginning.
Information: 0x402090DD at Load Gameweek, Flat File Destination [77]: The processing of file "
Information: 0x402090DD at Load Date, Flat File Destination [116]: The processing of file "C:\
Information: 0x4004300B at Load Gameweek, SSIS.Pipeline: "Flat File Destination" wrote 0 rows.
Information: 0x4004300B at Load Time, SSIS.Pipeline: "Add rows" wrote 0 rows.
Information: 0x4004300B at Load Gameweek, SSIS.Pipeline: "OLE DB Destination" wrote 0 rows.
Information: 0x40043009 at Load Time, SSIS.Pipeline: Cleanup phase is beginning.
Information: 0x40043009 at Load Gameweek, SSIS.Pipeline: Cleanup phase is beginning.
Information: 0x4004300B at Load Date, SSIS.Pipeline: "Add to database" wrote 0 rows.
Information: 0x4004300B at Load Date, SSIS.Pipeline: "Flat File Destination" wrote 0 rows.
Information: 0x40043009 at Load Date, SSIS.Pipeline: Cleanup phase is beginning.
Information: 0x40043008 at Load Team, SSIS.Pipeline: Post Execute phase is beginning.
Information: 0x40043009 at Load Team, SSIS.Pipeline: Cleanup phase is beginning.
```

Figure 25: Logs after executing dimension package of ETL, there is no new rows added as intended.



```
   Rows_Teams
1  25

   Rows_Date
1  490

   Rows_Gameweek
1  153

   Rows_Time
1  19

   Rows_Referee
1  28
```

Figure 26: Checking if there are new rows in dimension tables, there are not.



```
(No column name)
1487

(No column name)
1487

(No column name)
1487

(No column name)
1487
```

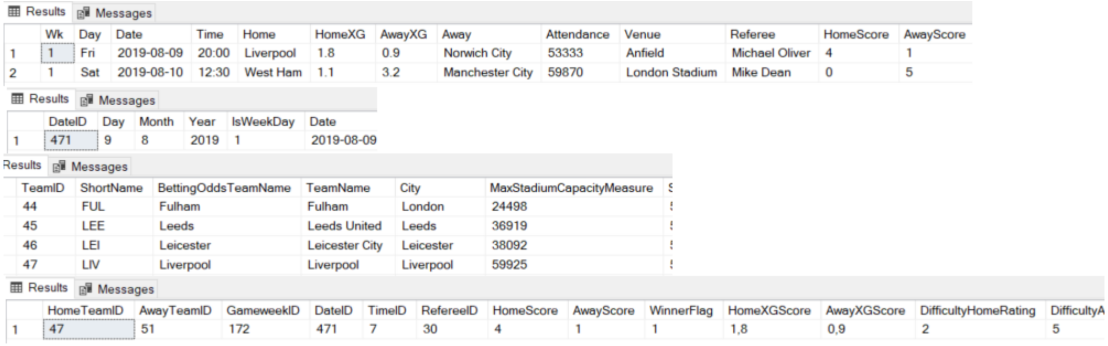
Figure 27: Checking if there are new rows in facts tables, there are not.

6.4 Validate specific row

We want to check specific row from raw database and check for it if whole mapping and foreign keys are correct.

Steps:

- Find row in raw database
- Find it in final database
- Compare with dimension tables



	Wk	Day	Date	Time	Home	HomeXG	AwayXG	Away	Attendance	Venue	Referee	HomeScore	AwayScore
1	1	Fri	2019-08-09	20:00	Liverpool	1.8	0.9	Norwich City	53333	Anfield	Michael Oliver	4	1
2	1	Sat	2019-08-10	12:30	West Ham	1.1	3.2	Manchester City	59870	London Stadium	Mike Dean	0	5

	DateID	Day	Month	Year	IsWeekDay	Date
1	471	9	8	2019	1	2019-08-09

TeamID	ShortName	BettingOddsTeamName	TeamName	City	MaxStadiumCapacityMeasure	S
44	FUL	Fulham	Fulham	London	24498	
45	LEE	Leeds	Leeds United	Leeds	36919	
46	LEI	Leicester	Leicester City	Leicester	38092	
47	LIV	Liverpool	Liverpool	Liverpool	59925	

	HomeTeamID	AwayTeamID	GameweekID	DateID	TimeID	RefereeID	HomeScore	AwayScore	WinnerFlag	HomeXGScore	AwayXGScore	DifficultyHomeRating	DifficultyA
1	47	51	172	471	7	30	4	1	1	1,8	0,9	2	5

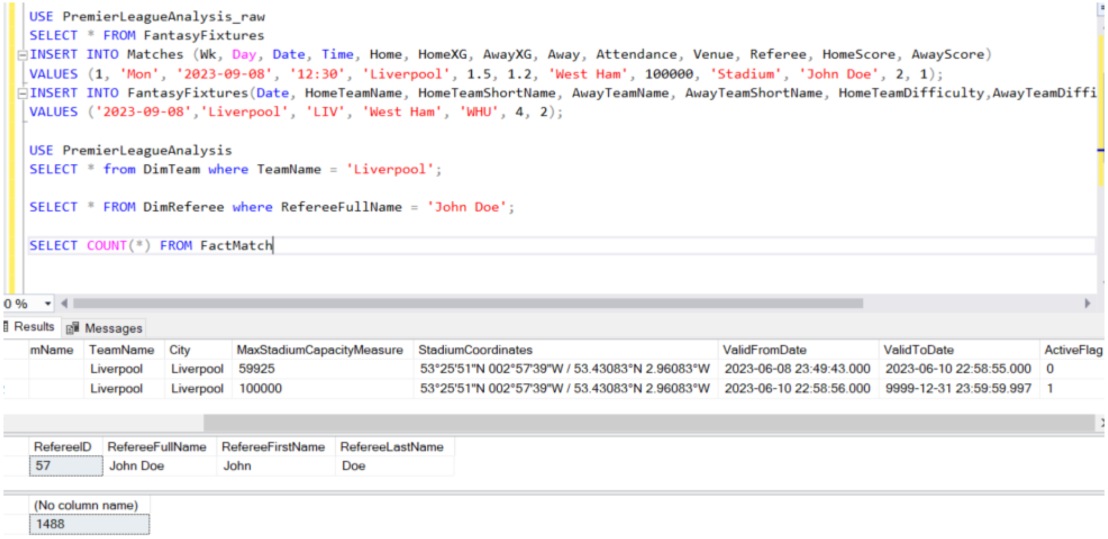
Figure 28: First result is from raw database, Matches table, second is DimDate table, third DimTeam and the last one we find by using IDs in dimension tables row in FactMatch table. After checking whole row we can say that everything was correctly added.

6.5 Validate addition of the new row

We want to check what happens if we added new match in raw database Match table and corresponded row to Fixtures table as well. It is simulation of getting new data after gameweek. We simulate it to get Attendance above the current one in specific team to check if SCD2 will work. The desire effect is that in DimTeam new row would be added and previous would change it's ActiveFlag to 0. Also there should be new additions to dimension tables when not exisiting dimension value is this new row.

Steps:

- Insert row in raw database
- Check dimension tables for changes
- Find row in final database



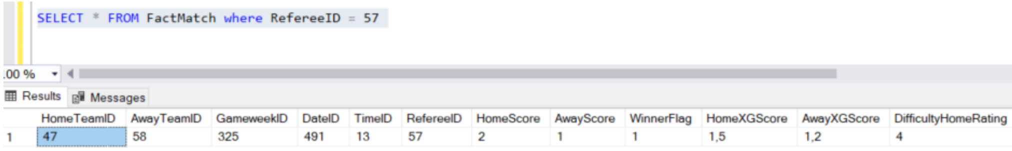
```
USE PremierLeagueAnalysis_raw
SELECT * FROM FantasyFixtures
INSERT INTO Matches (Wk, Day, Date, Time, Home, HomeXG, AwayXG, Away, Attendance, Venue, Referee, HomeScore, AwayScore)
VALUES (1, 'Mon', '2023-09-08', '12:30', 'Liverpool', 1.5, 1.2, 'West Ham', 100000, 'Stadium', 'John Doe', 2, 1);
INSERT INTO FantasyFixtures(Date, HomeTeamName, HomeTeamShortName, AwayTeamName, AwayTeamShortName, HomeTeamDifficulty,AwayTeamDiffi
VALUES ('2023-09-08','Liverpool', 'LIV', 'West Ham', 'WHU', 4, 2);

USE PremierLeagueAnalysis
SELECT * from DimTeam where TeamName = 'Liverpool';

SELECT * FROM DimReferee where RefereeFullName = 'John Doe';

SELECT COUNT(*) FROM FactMatch
```

mName	TeamName	City	MaxStadiumCapacityMeasure	StadiumCoordinates	ValidFromDate	ValidToDate	ActiveFlag
	Liverpool	Liverpool	59925	53°25'51"N 002°57'39"W / 53.43083°N 2.96083°W	2023-06-08 23:49:43.000	2023-06-10 22:58:55.000	0
	Liverpool	Liverpool	100000	53°25'51"N 002°57'39"W / 53.43083°N 2.96083°W	2023-06-10 22:58:56.000	9999-12-31 23:59:59.997	1

RefereeID	RefereeFullName	RefereeFirstName	RefereeLastName
57	John Doe	John	Doe

(No column name)
1488

Figure 29: SQl Query and it's result. We can see that SCD2 is working, there are now 2 rows for this team, one previous and one active. New Referee has been added to database and there is new row in FactMatch table.



```
SELECT * FROM FactMatch where RefereeID = 57
```

	HomeTeamID	AwayTeamID	GameweekID	DateID	TimeID	RefereeID	HomeScore	AwayScore	WinnerFlag	HomeXGScore	AwayXGScore	DifficultyHomeRating
1	47	58	325	491	13	57	2	1	1	1,5	1,2	4

Figure 30: We find our new row by RefereeID, everything is correct.

7 Power BI Reports

We divided our Report Section into two different directions. The first one summarizes information gathered from matches considering different time periods and teams. The second one sums up Betting Odds with the possible division on time periods and teams once again. All reports are created in Power BI. The data is imported from the Microsoft SQL Server.

7.1 Matches Reports

To fulfill requirements of Premier League teams, journalists and English football enthusiasts we created **3 one-paged reports**: one summarizes attendance, one summarizes referees and one summarizes opponents. It is worth noting that these three pages are loaded into one .pbix report so as to simplify data importing, calculated measures and hierarchies creation.

Hierarchies:

- Date: Year → Quarter → Month→ Day
- CityStadiumsHierarchy: City → StadiumCoordinatesCorrect (used in map visualisation)

Addad Columns:

- StadiumCoordinatesCorrect - needed substring correction within StadiumCoordinates attribute in DimTeam table
- GroupedAttendance - interval attribute for attendance needed for grouping the attendance to 5000 intervals. For example if there was 12233 people at the match then GroupedAttendance is 12000.

Calculated Measure:

- AwayScoreXGDifference - difference between AwayScore and AwayXGScore in FactMatch table
- HomeScoreXGDifference - difference between HomeScore and HomeXGScore in FactMatch table
- WinnerFlag1 - if WinnerFlag is 1 then 1 otherside 0
- WinnerFlag2 - if WinnerFlag is 2 then 2 otherside 0

7.1.1 Attendance

For analyzing the matches considering attendance we created 5 plots with slicer for choosing the timeline and dropdown for choosing the teams to analize for the user [31]. These plots are:

- Mean Attendance at Match - barplot with a trend line enabling the option for overall mean attendance measuring in different time hierarchy intervals

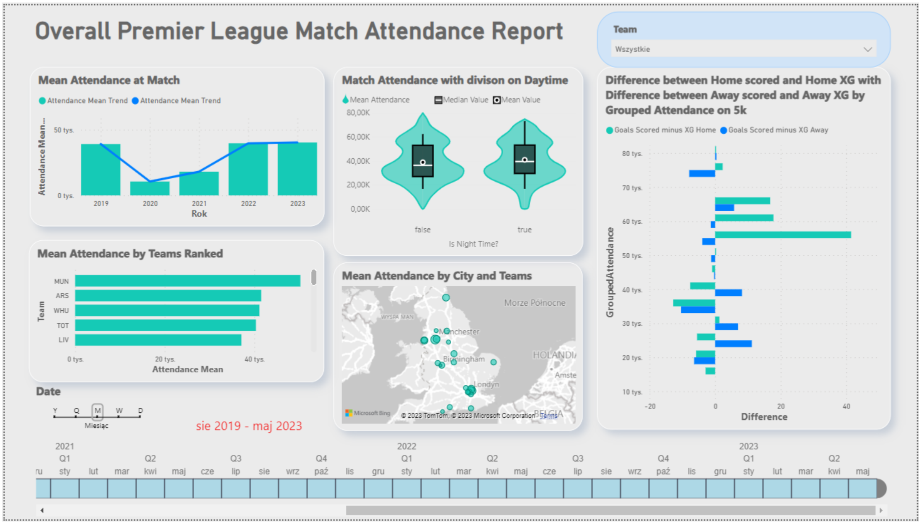


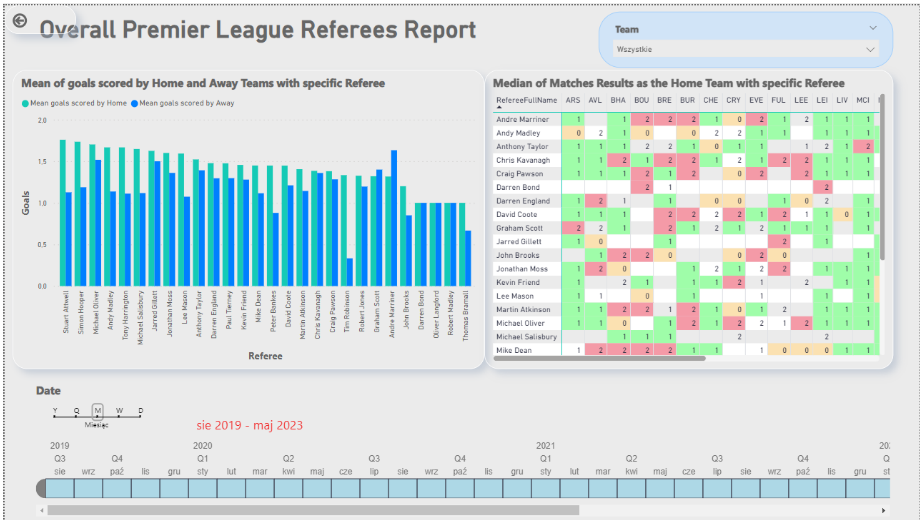
Figure 31: Report considering matches with attendance aspect.

- Mean Attendance with division on Daytime - two violin plots: one for early matches and the second one for late matches presenting the distribution of attendance per time
- Mean Attendance by Teams Ranked - top attendance mean ranking within specific period of time
- Mean Attendance by City and Teams - map using bubbles for presenting mean attendance using the hierarchy of location
- Difference between Home scored and Home XG with Difference between Away scored and Away XG by Grouped Attendance on 5k - shows if having attendance within specific section is rather helping home or away team

7.1.2 Referees

For analyzing the matches considering referees we created 1 plot and 1 matrix with slicer for choosing the timeline and dropdown for choosing the teams to analize for the user [32]. These items are:

- Mean of goals scored by Home and Away Teams with specific Referee - very useful grouped column plot for analyzing whether specific referee rather supports home or away team
- Median of Matches Results as the Home Team with specific Referee - for each particular referee and team shows if the team hosting the match with current referee rather wins or loses the match.

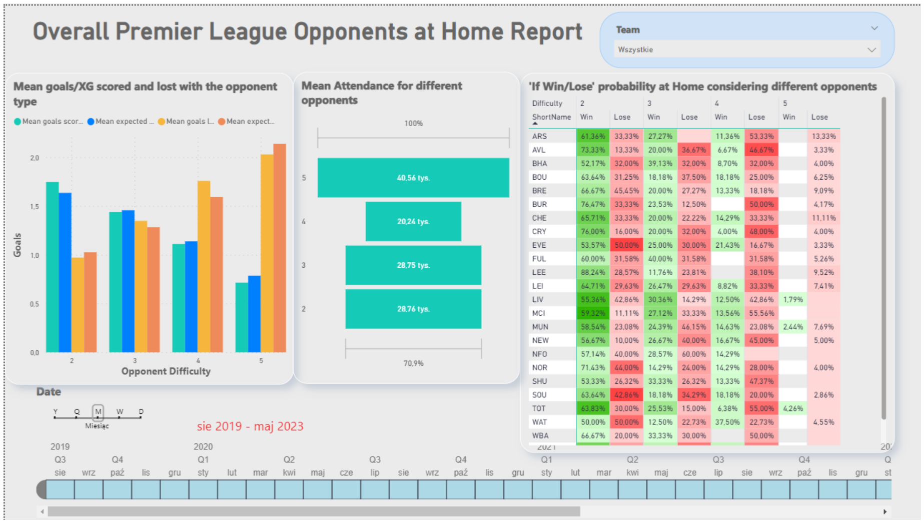


Median of Matches Results as the Home Team with specific Referee

RefereeFullName	ARS	AVL	BHA	BOU	BRE	BUR	CHE	CRY	EVE	FUL	LEE	LEI	LIV	MCI
Andre Marriner	1		1	2	2	2	1	0	2	1	2	1	1	1
Andy Madley	0	2	1	0		0	2	2	1	1		1	1	1
Anthony Taylor	1	1	1	2	2	1	0	1	1		1	2	1	2
Chris Kavanagh	1	1	2	1	2	2	1	2	1	2	2	1	1	1
Craig Pawson	1	1	1	2	1	2		0	2		2	1	1	1
Darren Bond				2	1							2		
Darren England	1	2	1		1		0	0		1	0	2		1
David Coote	1	1	1		2	2	2	2	1	2	1	1	0	1
Graham Scott	2	2	1		2	1	2	1	2	1	1	1		1
Jarred Gillett	1	0			1					2		1		
John Brooks		1	2	2	0			0	2	0				1
Jonathan Moss	1	2	0			1	2	1	2	2		1	1	1
Kevin Friend	1		2	1		1	1	2	1		2		1	1
Lee Mason	1	1		0		1			1			1		1
Martin Atkinson	1	1	2	2	1	2	1	0	1	2		1	1	1
Michael Oliver	1	1	0		1	2	1	2	2	1	2	1	1	1
Michael Salisbury			1	1	1			2				2		
Mike Dean	1	2	2	2	2	1	1		1	0	0	0	1	1

Figure 32: Report considering matches with referees aspect.

7.1.3 Opponents


Overall Premier League Opponents at Home Report
Team
Wszystkie
Mean goals/XG scored and lost with the opponent type
Mean goals scor...
Mean expected ...
Mean goals l...
Mean expect...
2.0
1.5
1.0
0.5
0.0
Goals
2
3
4
5
Opponent Difficulty
Mean Attendance for different opponents
100%
5
40,56 tys.
4
20,24 tys.
3
28,75 tys.
2
28,76 tys.
70,9%
Date
Y Q M W D
Miesiąc
sie 2019 - maj 2023
2019 Q3 sie wrz Q4 paź lis gru
2020 Q1 sty lut mar Q2 kwi maj cze Q3 lip sie wrz Q4 paź lis gru
2021 Q1 sty lut mar Q2 kwi maj cze Q3 lip sie wrz Q4 paź lis gru

'If Win/Lose' probability at Home considering different opponents

Difficulty	2		3		4		5	
ShortName	Win	Lose	Win	Lose	Win	Lose	Win	Lose
ARS	61,36%	33,33%	27,27%		11,36%	53,33%		13,33%
AVL	73,33%	13,33%	20,00%	36,67%	6,67%	46,67%		3,33%
BHA	52,17%	32,00%	39,13%	32,00%	8,70%	32,00%		4,00%
BOU	63,64%	31,25%	18,18%	37,50%	18,18%	25,00%		6,25%
BRE	66,67%	45,45%	20,00%	27,27%	13,33%	18,18%		9,09%
BUR	76,47%	33,33%	23,53%	12,50%		50,00%		4,17%
CHE	65,71%	33,33%	20,00%	22,22%	14,29%	33,33%		11,11%
CRY	76,00%	16,00%	20,00%	32,00%	4,00%	48,00%		4,00%
EVE	53,57%	50,00%	25,00%	30,00%	21,43%	16,67%		3,33%
FUL	60,00%	31,58%	40,00%	31,58%		31,58%		5,26%
LEE	88,24%	28,57%	11,76%	23,81%		38,10%		9,52%
LEI	64,71%	29,63%	26,47%	29,63%	8,82%	33,33%		7,41%
LIV	55,36%	42,86%	30,36%	14,29%	12,50%	42,86%	1,79%	
MCI	59,32%	11,11%	27,12%	33,33%	13,56%	55,56%		
MUN	58,54%	23,08%	24,39%	46,15%	14,63%	23,08%	2,44%	7,69%
NEW	56,67%	10,00%	26,67%	40,00%	16,67%	45,00%		5,00%
NFO	57,14%	40,00%	28,57%	60,00%	14,29%			
NOR	71,43%	44,00%	14,29%	24,00%	14,29%	28,00%		4,00%
SHU	53,33%	26,32%	33,33%	26,32%	13,33%	47,37%		
SOU	63,64%	42,86%	18,18%	34,29%	18,18%	20,00%		2,86%
TOT	63,83%	30,00%	25,53%	15,00%	6,38%	55,00%	4,26%	
WAT	50,00%	50,00%	12,50%	22,73%	37,50%	22,73%		4,55%
WBA	66,67%	20,00%	33,33%	30,00%		50,00%		

Figure 33: Report considering matches with opponents while playing at home aspect.

For analyzing the matches considering opponents at homeground we created 2 plots and 1 matrix with slicer for choosing the timeline and dropdown for choosing the teams to analise for the user [33]. These items are:

- Mean goals/XG scored and lost with opponents type - grouped column plot visualising mean goals and expected goals grouped by opponent difficulty type

- Mean Attendance for different opponents - tornado plot with mean attendance for different opponent types, useful for predicting the possible support from the fans at home match

- 'If Win/Lose' probability at Home considering different opponents - matrix presenting probabilities for each team that if this team wins at home what type of opponent it is and if this team loses at home what type of opponent it is. It is worth noting that colors are fitted for the rows only and with distinction on 'Win/Lose' type

7.2 Betting Odds Reports

To analyze betting, we created 2 reports. The first one focuses on odds related to team victories, draws, and defeats depending on the location (whether the team plays at home or away). The second report compares the bet365 website - the most popular platform for betting on the Premier League - with other betting websites. In both reports, the user has the option to select the teams they want to focus on, as well as the time range.

To understand the reports, it is important to keep a few facts in mind. Firstly, if someone is not familiar with betting odds, they work in such a way that the higher the odds, the bookmaker considers the outcome to be less likely to happen. Secondly, in the FactBettingOdds table, we have multiple values that represent very similar things. Measures that do not begin with the prefix 'Avg' or 'Max' typically refer to data from the bet365 website. Measures that have these prefixes usually pertain to aggregated data from multiple bookmakers, excluding bet365. We emphasize this fact at this point because it is crucial knowledge to consider when further interpreting the measures.

7.2.1 Match outcome betting

For analysing the betting on match outcomes we created 4 plots with a slicer for choosing the team(s) to analyze for the user [34]. The report uses many calculated measures which use specific TeamID relationships:

- WinOddsAvg - average odds of winning (both while playing at home and away). Used in *Team Average Odds* plot.

- HOME_POV_TotalAwayOddsAvg, HOME_POV_TotalDrawOddsAvg, HOME_POV_TotalHomeOddsAvg - measures from the point of view of the team playing home. Their average odds for defeat, draw, and victory when playing home. Those measures are used in the chart titled *Team Odds while playing home.*

- AWAY_POV_TotalAwayOddsAvg, AWAY_POV_TotalDrawOddsAvg, AWAY_POV_TotalHomeOddsAvg - measures from the point of view of the team playing away. Their average odds for victory, draw, and defeat when playing away. Those measures are used in the chart titled *Team Odds while playing away.*

- WinChance - sum of average percentages of winning at home or away subtracted by the average percentages of losing at home or away. Used in *Team win chances* plot.

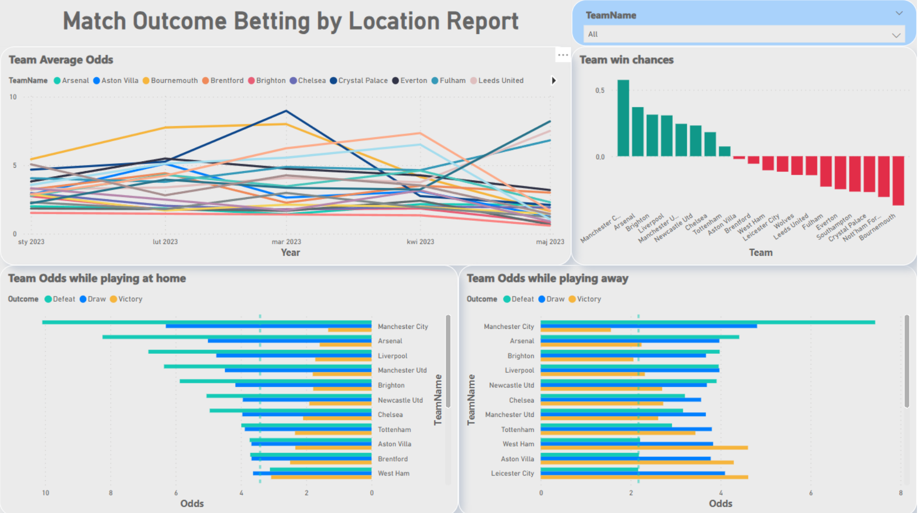


Figure 34: Report considering betting on match outcomes in the year 2023.

There are 4 plots in the report:

- Team Average Odds - shows how the odds for winning change for teams throughout the time. User can drill down to different levels of time granularity in order to achieve their desired values. Other plots change their values accordingly.

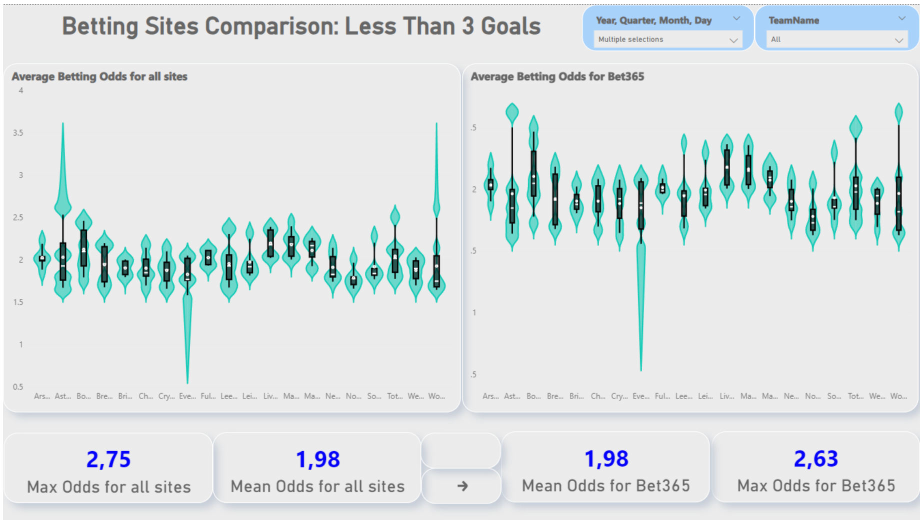
- Team win chances - shows whether the win chances are positive (green bars) or negative (red bars) and how significant are those values.

- Team Odds while playing at home - shows the average odds of the match outcome for the team when playing at home - the higher the odds the less likely it is going to happen.

- Team Odds while playing away - shows the average odds of the match outcome for the team when playing at away - the higher the odds the less likely it is going to happen.

Significant variations in the odds of match outcomes can be observed depending on whether the team is playing at home or away. This finding is expected and consistent, aligning with what can be observed by avid sports fans.

7.2.2 Betting sites comparison



Betting Sites Comparison: Less Than 3 Goals

Year, Quarter, Month, Day: Multiple selections

TeamName: All

Average Betting Odds for all sites

Average Betting Odds for Bet365

2,75 Max Odds for all sites

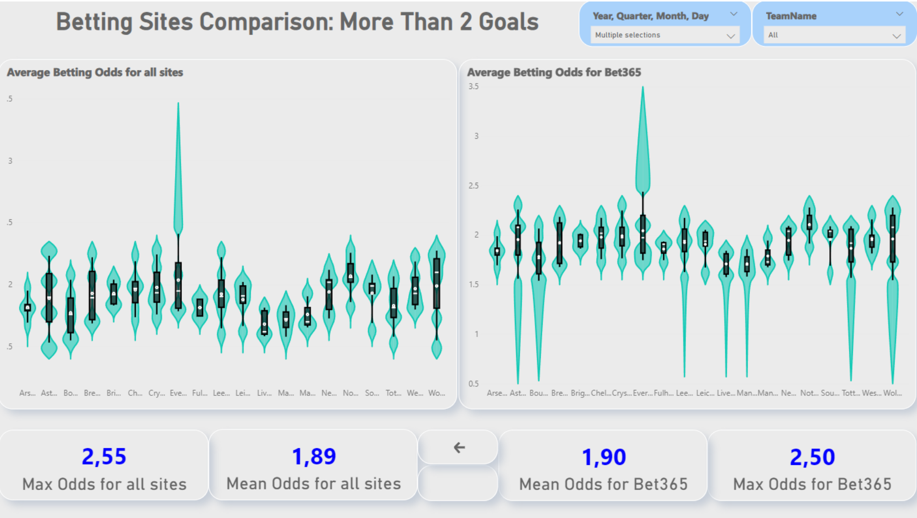
1,98 Mean Odds for all sites

1,98 Mean Odds for Bet365

2,63 Max Odds for Bet365

4
3.5
3
2.5
2
1.5
1
0.5
Ars.. Ast.. Bo.. Bre.. Bri.. Ch.. Cry.. Eve.. Ful.. Lee.. Lei.. Liv.. Ma.. Ma.. Ne.. No.. So.. Tot.. We.. Wo..
.5
2
.5
1
.5
Ars.. Ast.. Bo.. Bre.. Bri.. Ch.. Cry.. Eve.. Ful.. Lee.. Lei.. Liv.. Ma.. Ma.. Ne.. No.. So.. Tot.. We.. Wo..
→

Figure 35: Report on betting for less than 3 goals in a match in 2023 Q1.



Betting Sites Comparison: More Than 2 Goals

Year, Quarter, Month, Day: Multiple selections

TeamName: All

Average Betting Odds for all sites

Average Betting Odds for Bet365

2,55 Max Odds for all sites

1,89 Mean Odds for all sites

1,90 Mean Odds for Bet365

2,50 Max Odds for Bet365

.5
3
.5
2
.5
Ars.. Ast.. Bo.. Bre.. Bri.. Ch.. Cry.. Eve.. Ful.. Lee.. Lei.. Liv.. Ma.. Ma.. Ne.. No.. So.. Tot.. We.. Wo..
3.5
3
2.5
2
1.5
1
0.5
Arse.. Ast.. Bou.. Bre.. Brig.. Chel.. Crys.. Ever.. Fulh.. Lee.. Leic.. Live.. Man.. Man.. Ne.. Not.. Sou.. Tott.. Wes.. Wol..
←

Figure 36: Report on betting for more than 2 goals in a match in 2023 Q1.

This report focuses on comparing bet365 to other betting sites. The two version of the report are shown on figures [35] and [36]. Ideally those reports should be attained by only changing the betting category that the user is interested in, but for now those are separate pages showing only two possible categories.

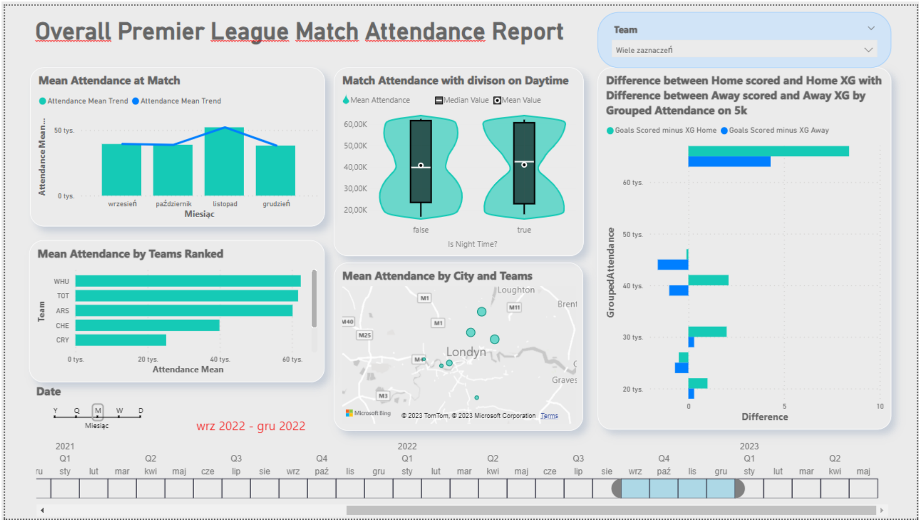
The report presents violin plots for the given category for each team. On the left side, information is provided for various bookmakers, while the right side focuses exclusively on the bet365 website. Users can select the teams of their interest and adjust the time range using slicers. Additionally, at the bottom of the reports, average and maximum odds information is displayed.

7.3 Power BI Tests

To test our reporting layer we checked our reports under 4 different settings of teams and timelines to see if everything loads and is calculated well. Our aim was to check the sense of these reports.

Settings for match reports:

1. Full timeline - all teams
2. From January 2022 to now - only Manchester City
3. September, October, November, December 2022 - only London teams
4. The year 2021 - Arsenal, Chelsea



Overall Premier League Match Attendance Report

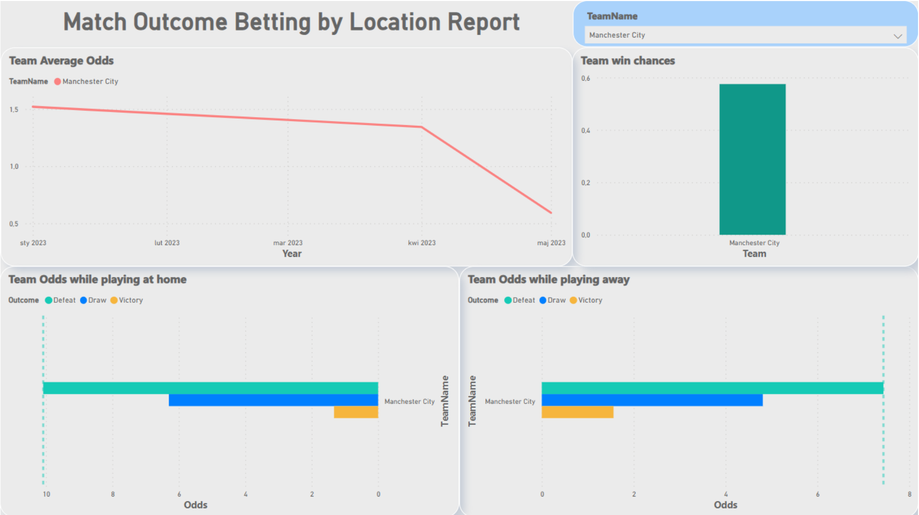
Team: Wiele zaznaczeń

Date: Miesiąc — wrz 2022 - gru 2022

Figure 37: Example of a report considering attendance aspect. Only London teams and 4 month period of time are set.

The result of this testing should be different visualisations concerning the settings. And in reality, the reports passed these test correctly. All calculated measures have been dynamically changed. The result of the third setting is presented on screenshot 37 (considering only attendance, other aspects also passed the test).

We also checked our betting reports on a setting for Manchester City in the year 2023. These reports also passed this test - once again all calculated measures changed dynamically in an expected manner. One of the reports on those settings is presented on figure [41].



Match Outcome Betting by Location Report

TeamName: Manchester City

Figure 38: Example of a report considering attendance aspect. Only London teams and 4 month period of time are set.

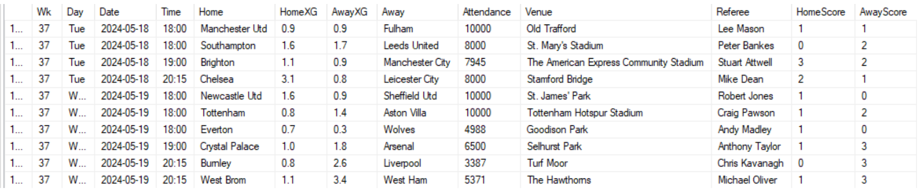
8 End-to-End Test

We simulated what would happen if we got the data from the new season. The aim was to test if everything works correctly between stages of our project. The desired outcome is that after inserting new data the reports could be easily updated.

Our end-to-end test consisted of the following stpes:

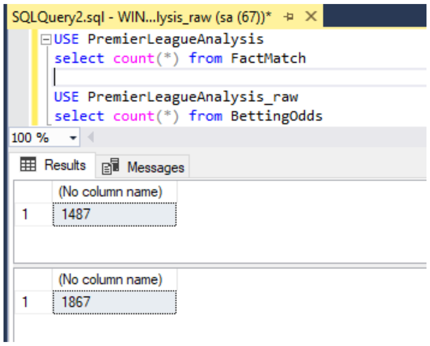
- Simulating new data - copying data from season 20/21 and changing the dates as if the games took place in season 23/24
- Appending new rows to the raw database using the Python script.
- Executing ETL process to upload created matches to our final database.
- Refreshing reports in PowerBI - the reports were done in insertion mode, to refresh the reports we simply click ”Refresh” button in PowerBI desktop.

- Checking whether the new data can be seen.



	Wk	Day	Date	Time	Home	HomeXG	AwayXG	Away	Attendance	Venue	Referee	HomeScore	AwayScore
1...	37	Tue	2024-05-18	18:00	Manchester Utd	0.9	0.9	Fulham	10000	Old Trafford	Lee Mason	1	1
1...	37	Tue	2024-05-18	18:00	Southampton	1.6	1.7	Leeds United	8000	St. Mary's Stadium	Peter Bankes	0	2
1...	37	Tue	2024-05-18	19:00	Brighton	1.1	0.9	Manchester City	7945	The American Express Community Stadium	Stuart Attwell	3	2
1...	37	Tue	2024-05-18	20:15	Chelsea	3.1	0.8	Leicester City	8000	Stamford Bridge	Mike Dean	2	1
1...	37	W...	2024-05-19	18:00	Newcastle Utd	1.6	0.9	Sheffield Utd	10000	St. James' Park	Robert Jones	1	0
1...	37	W...	2024-05-19	18:00	Tottenham	0.8	1.4	Aston Villa	10000	Tottenham Hotspur Stadium	Craig Pawson	1	2
1...	37	W...	2024-05-19	18:00	Everton	0.7	0.3	Wolves	4988	Goodison Park	Andy Madley	1	0
1...	37	W...	2024-05-19	19:00	Crystal Palace	1.0	1.8	Arsenal	6500	Selhurst Park	Anthony Taylor	1	3
1...	37	W...	2024-05-19	20:15	Burnley	0.8	2.6	Liverpool	3387	Turf Moor	Chris Kavanagh	0	3
1...	37	W...	2024-05-19	20:15	West Brom	1.1	3.4	West Ham	5371	The Hawthorns	Michael Oliver	1	3

Figure 39: Some new "fake" rows of match history in raw database

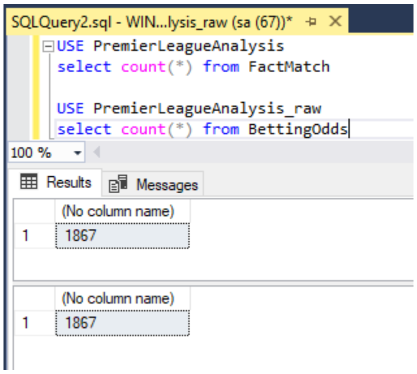


```
USE PremierLeagueAnalysis
select count(*) from FactMatch

USE PremierLeagueAnalysis_raw
select count(*) from BettingOdds
```

	(No column name)
1	1487

	(No column name)
1	1867

Figure 40: SQL Query before executing ETL process



```
USE PremierLeagueAnalysis
select count(*) from FactMatch

USE PremierLeagueAnalysis_raw
select count(*) from BettingOdds
```

	(No column name)
1	1867

	(No column name)
1	1867

Figure 41: SQL Query after executing ETL process

In figure [42], we demonstrated the successful completion of our end-to-end test. The reports refreshed without any problems and new data became available for the user.

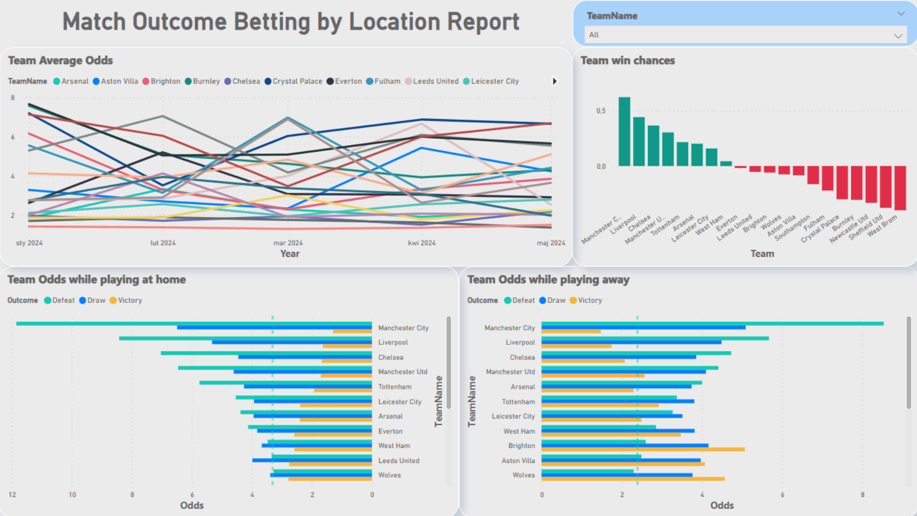


Figure 42: Match Outcome Betting report with new data from season 23/24. Sliced to show only year 2024.

9 Summary

The project aims to construct a data warehouse model for comprehensive analysis of Premier League matches, with a specific focus on betting odds. The purpose is to meet the increasing demand for in-depth analysis and prediction models in football betting by providing a consolidated data platform.

The data warehouse model is based on a star schema, facilitating efficient querying and easy understanding of the data structure. It includes two fact tables for match facts and betting odds facts, connected through shared dimensional tables.

The match facts table contains data on match results, including teams, scores, expected goals, and attendance. The betting odds facts table includes data from Bet365, a prominent betting company in England.

The results of the project provide a robust data foundation for advanced analytics, supporting better decision-making for betting enthusiasts, football analysts, and the wider sports industry. The data warehouse enables comprehensive analysis of the relationship between match outcomes and betting odds, identification of trends over time, and the development of prediction models.

From a business perspective, the solution enhances understanding of Premier League matches and the betting landscape, contributing to more accurate predictions and insights. The data warehouse facilitates comprehensive analysis of betting odds, enabling the identification of patterns and trends that can inform strategic decision-making for betting enthusiasts and analysts.

Overall, the project's results provide a valuable resource for the sports industry, improving the ability to analyze Premier League matches and make informed predictions based on comprehensive data analysis and betting odds information.

10 Division of Work

- Szymon Matuszewski - finding datasets, developing a data warehouse star model, creating a diagram in MS SQL Management, ETL debugging, creating staging tables, adding SQL queries in ETL, creating reports analyzing matches in Power BI, preparing final presentation
- Michał Mazuryk - finding datasets, developing a data warehouse star model, designing and creating ETL process, preforming tests and validation of process, minor assistance in several other stages of the project
- Damian Skowroński - creating database with raw values in MS SQL Management as well as populating the database using Python, creating reports analyzing betting odds in Power BI, minor assistance in several other stages of the project