MS 984: Data Analytics in Practice Report

Case Study 1 – Genius Sports Officiating Error Analysis (2021/22 – 2025/26)

Team Members Group 4 : Rohan Sharma, Daayum Mohsin, Ishita, Prashanth Babu, Nikita Masquri

Abstract

This report details the application of Microsoft Power BI and Data Analysis Expressions (DAX) to analyse officiating error data provided by Genius Sports, a global leader in sports data technology. The scope of the study focused on English and Scottish football leagues between the 2021/22 and 2025/26 seasons. Through robust data cleaning, transformation using Power Query, and analytical modelling with DAX, the study aimed to identify key patterns in match officiating errors, assess data quality, and visualise trends effectively. Key findings indicate a disproportionately higher volume of errors in Scottish Premiership games highlighted, as 64% of Major errors were ongoing VAR decisions, analysis of Major errors highlights significant volatility in **Ongoing VAR mistakes** in Scotland compared to England (37.5% of all Major Errors). This therefore highlights the ongoing necessity for greater training within the Scottish Referees Association to handle VAR situations in a timely and efficient manner.

1 Introduction

Genius Sports provides essential live data to clubs, leagues, media outlets, and sportsbooks globally. The accuracy and speed of this data collection, often carried out by remote statisticians, are paramount. This study was initiated to address quality assurance concerns, specifically by analysing the reporting of 'Major' (Goals, Penalties, Red Cards, VAR decisions) and 'Moderate' (Yellow Cards, Corners) events to understand the nature and frequency of errors recorded by the Statistician Network.

2 Aim and Objectives

The primary aim of this project was to leverage Power BI and DAX to transform raw Genius Sports data into reliable, interactive dashboards to identify and monitor officiating error patterns across English and Scottish football. **Specific objectives included:**

- Perform rigorous data cleaning and standardization across the consolidated dataset.
- Develop custom DAX measures and calculated columns to enable complex analytical filtering.
- Establish a functional date hierarchy to support time-series analysis of error trends.
- Generate clear visualisations to communicate findings regarding error distribution, severity and how influential these errors were in the topflight of Scotland and England

3 Methodology

The analysis followed a structured process involving three main stages: data preparation (Excel and Power Query), data modelling (DAX), and data visualisation (Power BI).

3.1 Data Preparation and Cleaning

1. **Initial Filtering:** The starting dataset of 997 rows was reduced. Non-football data (14 rows of Ice Hockey, 1 row of Basketball) and 9 rows lacking a 'Mistake Type' were removed to maintain data integrity and focus the analysis. The final dataset contained **895 valid football match records**.
2. **Excel Correction:** A key manual data cleaning step involved correcting row 178 (Hibernian vs St Johnstone), which was misclassified as being in England. This record was corrected to reflect its status as a **Scottish Premiership** game, ensuring regional consistency.

3.2 Data Modelling with DAX

DAX (Data Analysis Expressions) was used to create calculated elements to enrich the data model:

- **Date Table and Season Year:** A dedicated Date Table was built using the CALENDARAUTO() function, as shown below, to create a base calendar, which was then extended with custom Season Year logic (July 1st to June 30th) to align with the football season.

```
Date =
ADDCOLUMNS(
        CALENDARAUTO(),
        "Year", YEAR([Date]),
        "Month", FORMAT( [Date], "mmmm"),
        "Month Number", MONTH( [Date])
)
```

- **Competition (No Year):** A calculated column was created using complex DAX logic to standardize competition names by dynamically stripping the associated season years, enabling a cleaner comparison of mistake volumes across competition types regardless of the specific year.

```
Competition (No Year) =
VAR s = TRIM ( 'Eng and Sco'[Competition] )
VAR y1 = IFERROR( VALUE ( LEFT ( s, 4 ) ), BLANK ( ) )
VAR y2 = IFERROR( VALUE ( MID ( s, 6, 4 ) ), BLANK ( ) )
VAR hasDualYear = NOT ISBLANK ( y1 ) && MID ( s, 5, 1 ) = "/" && NOT ISBLANK ( y2 ) && MID ( s, VAR hasSingleYear = NOT ISBLANK ( y1 ) && MID ( s, 5, 1 ) = " "
RETURN
IF (
        hasDualYear,
        MID ( s, 11, LEN ( s ) - 10 ),
        IF (
                hasSingleYear,
                MID ( s, 6, LEN ( s ) - 5 ),
                s
        )
)
```

- **Team Category:** Matches were categorized into **Senior, U23, or U21** based on team category analysis.
- **Relationships:** A **Many-to-One** relationship was established between the main dataset's StartDate column and the newly created Date table to facilitate time-series analysis and dynamic filtering by season and quarter.

4 Findings and Analysis

The Power BI dashboard visualizations provide clear insights into the distribution and nature of officiating errors.

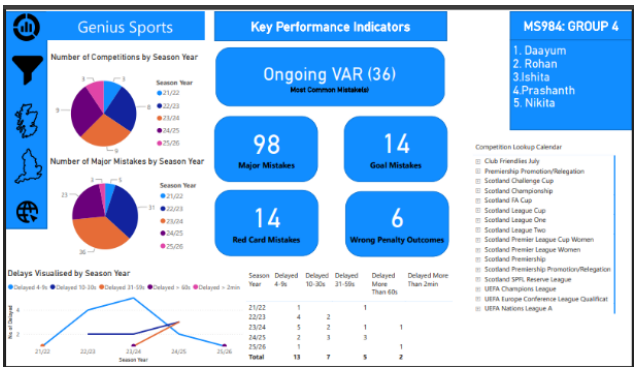

Figure 1.1: Major Mistakes In Scotland

Figure 1.2: Major Mistakes in England

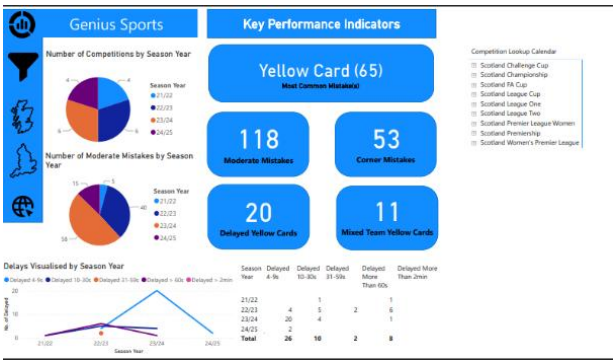

Figure 2.1: Moderate Mistakes in Scotland

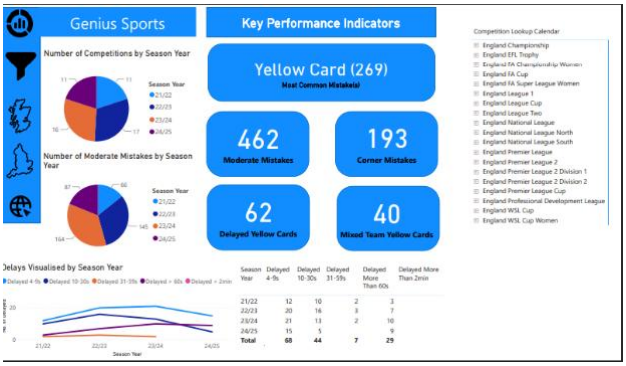

Figure 2.2: Moderate Mistakes in England

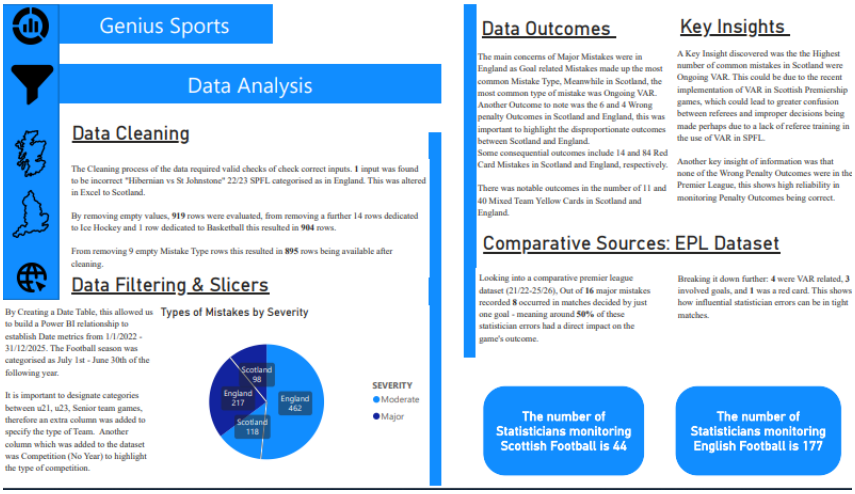

Figure 3: Data Analysis in Power BI Analytics and Structure

4.1 Geographical Distribution and Severity

Between 2022 and 2025, the overall volume of recorded mistakes was significantly concentrated in England.

Total Mistakes Distribution by Country (2022–2025)

Figure 3.1: Total Mistakes Distribution by Country (2022–2025)

- **Total Mistake Distribution (2022–2025):** England accounted for **75.9% (679)** of the total mistakes, while Scotland accounted for **24.1% (216)**.
- **Severity Breakdown:** Figure 1 shows the breakdown by severity. In both countries, Moderate mistakes outnumbered Major mistakes. However, England recorded a higher volume of both categories:
  - **England:** 462 Moderate, 217 Major mistakes.
  - **Scotland:** 118 Moderate, 98 Major mistakes.

4.2 Major Error Breakdown by Topflight Competition Level

Mistakes were not uniformly distributed, with English lower leagues reporting the highest volumes, as detailed in Table 1 and 2.

- **Major Error Severity:** The Scottish Premiership had a higher total volume than the English Premier League (84 vs 46) and **recorded more Major mistakes (50 vs. 16)** in the analysed period.

4.3 Overall Mistake Types

The most frequent mistake types across the entire dataset were related to events categorized as Moderate.

1. **Yellow Card:** 334 total mistakes (Moderate).
2. **Corner:** 246 total mistakes (Moderate).

Major mistake types, such as **Goal (89)** and **Red Card (84)**, were also prominent.

Table 1: English Mistakes by Competition Level and Severity

Competition Level	Major Mistakes	Moderate Mistakes	Total Mistakes
(ENG) National League South	28	35	63
(ENG) National League North	17	41	58
(ENG) National League	21	47	68
(ENG) League Two	31	71	102
(ENG) League One	27	115	142
(ENG) EFL Championship	17	66	83
(ENG) English Premier League	16	30	46

Table 2: Scottish Mistakes by Competition Level and Severity

Competition Level	Major Mistakes	Moderate Mistakes	Total Mistakes
(SCO) Legue Two	7	10	17
(SCO) League One	4	16	20
(SCO) SPFL Championship	11	47	58
(SCO) SPFL Premiership	50	34	84

4.4 Comparative Sources: EPL Dataset

Looking into the comparative Premier League dataset (**21/22–25/26**) and found something interesting. Out of 16 major mistakes recorded, 8 occurred in matches decided by just one goal.

- Meaning around **50% of these errors had a direct impact on the game's outcome**.
- Breaking it down: **4 were VAR-related**, **3 involved goals**, and **1 was a red card**.

Source URL: https://kaggle.com/datasets/macrohuiii/english-premier-league-epl-match-data-2000-2025

4.5 Analysis of Key Major Mistake Types

To pinpoint high-risk areas, a focused analysis was conducted on **Ongoing VAR decisions, Goal mistakes, and Red Card mistakes**, which represent critical Major event errors. **Goal and Red Card Mistakes (Overall)**

- **Goal:** 89 total mistakes were recorded across all competitions.
- **Red Card:** 84 total mistakes were recorded across all competitions.

These two categories represent significant, match-altering Major errors and together account for 173 total mistakes in the dataset.

Ongoing VAR Mistake Disparity A drill-down into VAR-related Major mistakes revealed a significant geographical difference. The data clearly indicates a higher volume of **Ongoing VAR** mistakes in Scotland compared to England: **Scottish Premiership:** 32(Out of 50 Major Mistakes) Ongoing VAR mistakes and **English Premier League:** 6 (Out of 16 Major Mistakes) Ongoing VAR mistakes. This disparity is particularly notable given that Scotland has a lower total mistake volume (24.1% of the total) and implemented VAR later (October 2022) than England (August 2019). This suggests potential challenges in the initial rollout, procedural adherence, or technical integration of VAR reporting within the Scottish Statistician Network.

5 Discussion

The systematic data cleaning and modelling process significantly enhanced the reliability of the Genius Sports dataset. The high proportion of errors found in English lower leagues warrants further investigation, possibly into the training or quality assurance protocols for the statisticians covering these games. The distinct seasonal variation in errors—a finding made possible by the custom Date Table—suggests that increased match pressure or workload during peak season periods may correlate with reduced data accuracy. The insights gained from the visualization phase are immediately actionable, allowing management to target specific leagues and mistake types (e.g., Yellow Card and Corner reporting) for focused training and quality audits. **The critical findings regarding VAR mistakes in Scotland(Section4.5) emphasize an urgent need for targeted intervention and retraining on major event protocols in that region.**

6 Conclusion and Recommendations

This case study successfully demonstrated the power of integrating robust data preprocessing (Excel/Power Query) with dynamic analytical modelling (Power BI/DAX) to monitor performance in a data-intensive environment like sports analytics. The resulting analysis provided clear visibility into the major sources and patterns of officiating errors. **Recommendations for Future Action:**

1. **Targeted Training for Scottish Premiership:** Implement focused training sessions for statisticians covering the Ongoing VAR decisions to address the most common Errors in the Scottish Topflight.
2. **Act as a voice to improve the quality of VAR referring in Scotland and England**: To improve the time and efficiency of VAR decision making. Implementing another button for extending VAR decision time.
3. **VAR Consistency (Urgent Focus): Immediately** analyse the exceptionally high rate Major Mistakes as 'Ongoing VAR' mistakes in Scotland (64% vs. 34.75% in England) to address potential procedural gaps, training deficiencies, or technical issues following its implementation.

7 Appendix – Key DAX Measures

The following are examples of DAX measures created to support the analysis and visualisation, allowing for dynamic calculation of key metrics within the Power BI environment:

Total Mistakes

```
Total Mistakes =
COUNTROWS ( 'Eng and Sco'[MistakeType])
```

Most Common Mistake and Number

```
Most Common Mistake (Label) =
VAR t =
   SUMMARIZE (
      'Eng and Sco',
      'Eng and Sco'[MistakeType],
      "Cnt", COUNTROWS ( 'Eng and Sco' ))
VAR top1 = TOPN ( 1, t, [Cnt], DESC )
RETURN
   CONCATENATEX ( top1, 'Eng and Sco'[MistakeType] & " (" & [Cnt] & ")", ", " )
```

Wrong Penalty Outcomes

```
Wrong Penalty Outcomes =
CALCULATE (
   COUNTROWS ( 'Eng and Sco' ),
   KEEPFILTERS ( 'Eng and Sco'[MistakeType] = "Penalty outcome (In regular time)" ),
   KEEPFILTERS (
      'Eng and Sco'[MistakeTypeCategory] IN {
         "Wrong Penalty outcome type confirmed (Penalty Goal instead of Penalty Missed)",
         "Wrong Penalty retake procedure"}))
```

Goal Mistakes

```
Goal Mistakes =
CALCULATE (
   COUNTROWS ( 'Eng and Sco' ),
   CONTAINSSTRING ( LOWER ( 'Eng and Sco'[MistakeType] ), "goal" ))
```

Red Card Mistakes

```
Red Card Mistakes =
CALCULATE (
   COUNTROWS ( 'Eng and Sco' ),
   CONTAINSSTRING ( LOWER ( 'Eng and Sco'[MistakeType] ), "red card" ))
```

Mixed Team Yellow Card Mistakes

```
Mixed Team Yellow Card Mistakes =
CALCULATE (
   COUNTROWS ( 'Eng and Sco' ),
   KEEPFILTERS ( 'Eng and Sco'[MistakeType] = "Yellow Card" ),
   KEEPFILTERS (
      'Eng and Sco'[MistakeTypeCategory] IN {
         "Mixed Teams”}))
```

Delayed Yellow Card Mistakes

```
Delayed Yellow Card Mistakes =
CALCULATE (
   COUNTROWS ( 'Eng and Sco' ),
   KEEPFILTERS ( 'Eng and Sco'[MistakeType] = "Yellow Card" ),
   KEEPFILTERS (
      'Eng and Sco'[MistakeTypeCategory] IN {
         "Delayed"}))
```