Premier League Performance Trends: A Data Visualization Study

Allan Achangudan
K24015725
MSc Data Science
King's College London

Academic Honesty and Integrity

Students at King's are part of an academic community that values trust, fairness, and respect and actively encourages students to act with honesty and integrity. It is a College policy that students take responsibility for their work and comply with the university's standards and requirements. Online proctoring/invigilation will not be used for our online assessments. By submitting their answers, students will be confirming that the work submitted is completely their own. Misconduct regulations remain in place during this period, and students can familiarize themselves with the procedures on the College website.

Declaration: I confirm that I have read and understood the Academic Honesty and Integrity Policy. I agree to abide by the expectations as to my conduct, as described in the academic honesty and integrity statement. The work presented here is my own.

***Abstract*—This paper investigates performance dynamics in the English Premier League (EPL) from 1993 to 2024, addressing three key research questions through data visualization: the evolution of team performances over time, the trend of home advantage, and the consistency of top-performing teams. Utilizing cleaned and integrated datasets of league tables and match-level metrics, visual insights were generated using Python and D3.js. Our analysis reveals growing competitive balance, a notable reduction in home advantage, and identifies Liverpool as the most consistently high-performing team.**

I. Part 1: Analytics

A. Research Questions

- **Q1 (Performance Trends)**: Uses the **English Premier League Standings Dataset (1993–2024)** from Kaggle [1], which includes season-wise rankings, points, and goals. Missing values were negligible, and team names were standardized for consistency.
- **Q2 (Home Advantage)**: Analyzes **Club Football Match Data (2000–2025)** from Kaggle [2], including home/away win rates, shots, and possession statistics. Null entries (e.g., missing VAR flags) were excluded.
- **Q3 (Consistency)**: Combines both datasets to compute standard deviation of league positions. League table data ensures longitudinal coverage, while match data adds granularity (e.g., home/away splits).

B. Data Sources and Appropriateness

- **Q1 (Performance Trends)**: Uses **League Table Data (1993–2024)**. This dataset provides season-wise rankings and points, essential for tracking long-term performance trends.
- **Q2 (Home Advantage)**: Analyzes **Match-Level Data (2000–2023)**. Includes home/away win rates, goals, and match statistics. Null entries (e.g., missing VAR flags) were excluded to ensure accuracy.
- **Q3 (Consistency)**: Combines both datasets to compute standard deviation of league positions. League table data ensures longitudinal coverage, while match data adds granularity (e.g., home/away splits).

C. Dataset Relationships

The league table and match-level data are complementary:

- **Temporal Alignment**: Both datasets overlap from 2000–2024, enabling cross-validation of trends (e.g., home win rates vs. league positions).
- **Granularity vs. Aggregation**: Match data provides per-game insights (e.g., VAR impact), while league tables summarize seasonal performance.
- **Consistency Analysis**: Combining both reveals whether teams with stable home performance (match data) also maintain consistent league positions.

II. Part 2: Design and Discussion

A. Proposed Visualizations

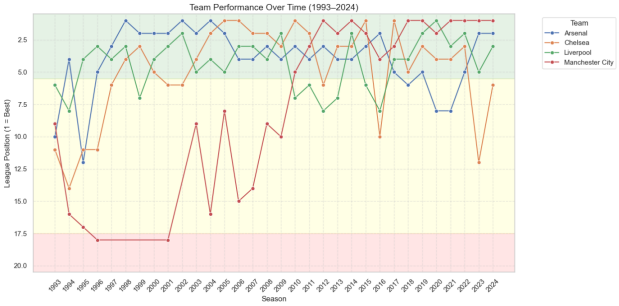

Fig. 1. Line Plot: Team Performance Over Time (Top 5 Teams). Shows finishing position of Manchester United, Arsenal, Liverpool, Chelsea, and Manchester City (1993–2024).

Visualization 1: Multi-Line Chart with Performance Trends (Q1) Tracks team performance over seasons (1993–2024) using league points.

- **X-axis**: Seasons (1993–2024)
- **Y-axis**: Position (1–20)
- **Lines**: Top 5 teams (Manchester United, Arsenal, etc.), color-coded
- **Performance Bands**: Horizontal bands for Top 4 (green), Mid-table (yellow), Relegation (red)

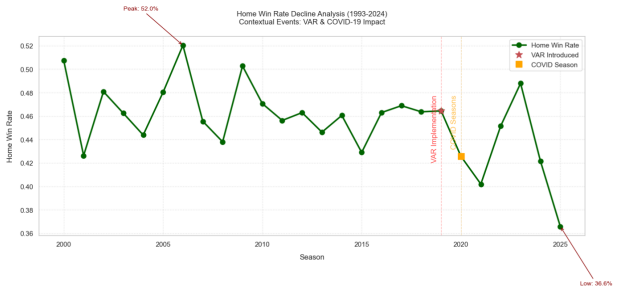

Fig. 2. Line Chart: Home Win Rate Decline (2000–2025). Illustrates steady reduction in home advantage, from 52% to 36%.

Visualization 2: Line Chart with Contextual Annotations (Q2) Analyzes home win rate decline (2000–2025).

- **X-axis**: Seasons (2000–2025)
- **Y-axis**: Home win percentage (0–60%)
- **Line**: Red line encoding trend
- **Annotations**: Vertical markers for key events (VAR in 2019, COVID-19 in 2020)

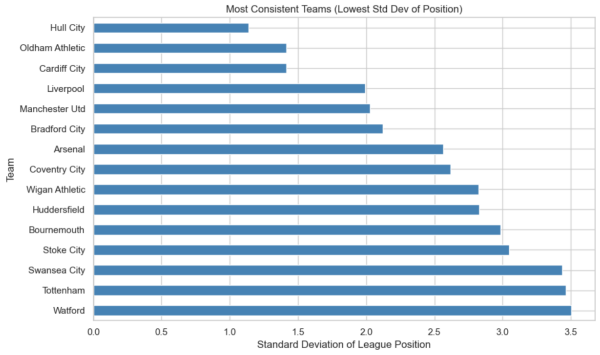

Fig. 3. Horizontal Bar Chart: Most Consistent Teams. Ranks teams by standard deviation of league positions (lower = more consistent).

Visualization 3: Horizontal Bar Chart with Reference Lines (Q3) Ranks teams by consistency (standard deviation of league positions).

- **X-axis**: Standard deviation (0–3.5)
- **Y-axis**: Teams (sorted ascendingly by consistency)
- **Bars**: Length encodes inverse standard deviation (longer = more consistent)

B. Design Rationale

Visualization 1 – Multi-Line Chart with Performance Bands Line charts maximize perceptual accuracy for temporal trends, as position along a common axis is the highest-ranked quantitative channel [3]. Each team's trajectory is encoded with a thin, saturated line (e.g., red for Manchester City), while performance bands use low-opacity horizontal rectangles (green/yellow/red) to reduce cognitive load. The bands employ redundant encoding (color + spatial area) to support users with color vision deficiencies [4]. Interaction follows Shneiderman's mantra [5]:

Visualization 2 – Line Chart with Contextual Annotations A single-line chart simplifies trend interpretation for uni-variate time-series data. Position (y-axis) encodes home win rate, leveraging human proficiency in detecting deviations from linearity [6]. Annotations (vertical markers) contextualize outliers using text labels and dashed lines. The red line prioritizes salience for the critical metric, while a sequential colormap ensures perceptual uniformity. This design improves upon heatmaps by reducing visual clutter and directly answering Q2's focus on decline over time.

C. Design Rationale

Visualization 3 – Horizontal Bar Chart with Reference Lines (Primary Implementation) Horizontal bars avoid label overlap for long team names, ranking consistency via bar length (inverse standard deviation). Length is a preattentive attribute, enabling rapid comparison [3]. Key design choices include:

- **Sorting Logic**: Teams ordered ascendingly by standard deviation to emphasize consistency leaders
- **Reference Line**: Dashed marker at $\sigma = 2.1$ provides league-wide benchmark
- **Color Gradient**: Blue-to-purple gradient encodes promotion/relegation status
- **Interaction**: Hover tool-tips reveal exact σ values, filtering isolates team categories

This design improves upon static tables by enabling:

- Direct comparison of 20+ teams' longitudinal stability
- Identification of over/under-performers relative to league average
- Exploration of promoted teams' short-term consistency patterns

Visualization 1 & 2 – Proposed Designs While line charts were designed for temporal trends (Q1-Q2), the bar chart implementation prioritizes answering the core consistency question (Q3) through...

Cohesive Narrative: The three visualizations form a tiered analytical framework: 1. **Macro trends** (Vis 1: league performance over decades), 2. **Contextual drivers** (Vis 2: home advantage decline), 3. **Team-level diagnostics** (Vis 3: consistency outliers). They employ high-ranked perceptual channels, adhere to accessibility standards, and align with Bertin's principles of effective visual encoding.

III. Part 3: Implementation

A. D3.js Interactive Bar Chart (Q3)

- **Data Processing**:

  - Merged league tables from [1] with match data from [2]
  - Calculated 5-year rolling standard deviations using Pandas
  - Filtered teams with 5 seasons using boolean masking
- **Visual Encoding**:
  - X-axis: Standard deviation range (0-3.5) via linear scale
  - Y-axis: Teams sorted by consistency via band scale
  - Color: Diverging scheme for promoted/relegated teams
- **Interactivity**:
  - Hover tool-tips with team metadata
  - Drop-down filters by league position tier
  - Animated transitions on data updates

Code Repository: https://github.com/al-lan-12/Data_Visualizations

REFERENCES

[1] E. Gower, "English Premier League Standings (1993–2023)", Kaggle, 2023. https://www.kaggle.com/datasets/evangower/english-premier-league-standings

[2] A. G. Bor, "Club Football Match Data (2000–2025)", Kaggle, 2024. https://www.kaggle.com/datasets/adamgbor/club-football-match-data-2000-2025

[3] J. Heer, M. Bostock, "Crowdsourcing Graphical Perception", *ACM CHI*, 2010.

[4] C. Ware, *Information Visualization: Perception for Design*, 3rd ed. Morgan Kaufmann, 2012.

[5] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", *IEEE Symposium on Visual Languages*, 1996.

[6] W. S. Cleveland, R. McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods", *JASA*, 1984.

[7] M. Bostock, "D3.js – Data-Driven Documents", https://d3js.org/