Data-Driven Insights from Europe's Top 5 Leagues (2008–2016)

Introduction

This project explores Europe’s top five football leagues (Premier League, Serie A, Ligue 1, La Liga, and Bundesliga) using a Kaggle dataset with over **25,000 matches** and **10,000 players** from 2008 to 2016.

The analysis was performed using **Python (Pandas, NumPy, Matplotlib, and Seaborn)** on Jupyter Notebook. After extensive data cleaning and preparation, the project focused on uncovering **story-driven insights** into club performance, dominance, and competitive dynamics across leagues.

Tech Stack

- **Python (Jupyter Notebook)** – core analysis and scripting
- **Pandas & NumPy** – data cleaning and transformation
- **Matplotlib & Seaborn** – data visualization and storytelling
- **Kaggle Dataset** – raw match and player data

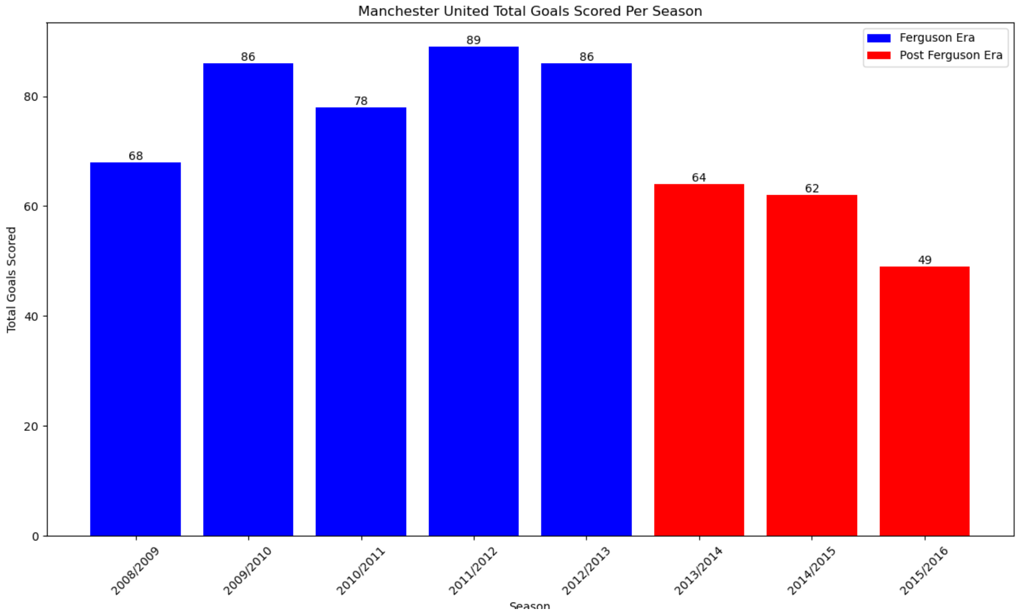
Project Workflow

1. **Data Cleaning & Preparation** – handled missing values, standardized formats, and ensured consistency.
2. **Exploratory Analysis** – identified key storylines across each league.
3. **Statistical Analysis** – applied metrics such as clean sheets, entropy (predictability), and seasonal point comparisons.
4. **Visualization** – built clear charts to highlight insights for each storyline.

Key Analyses & Insights

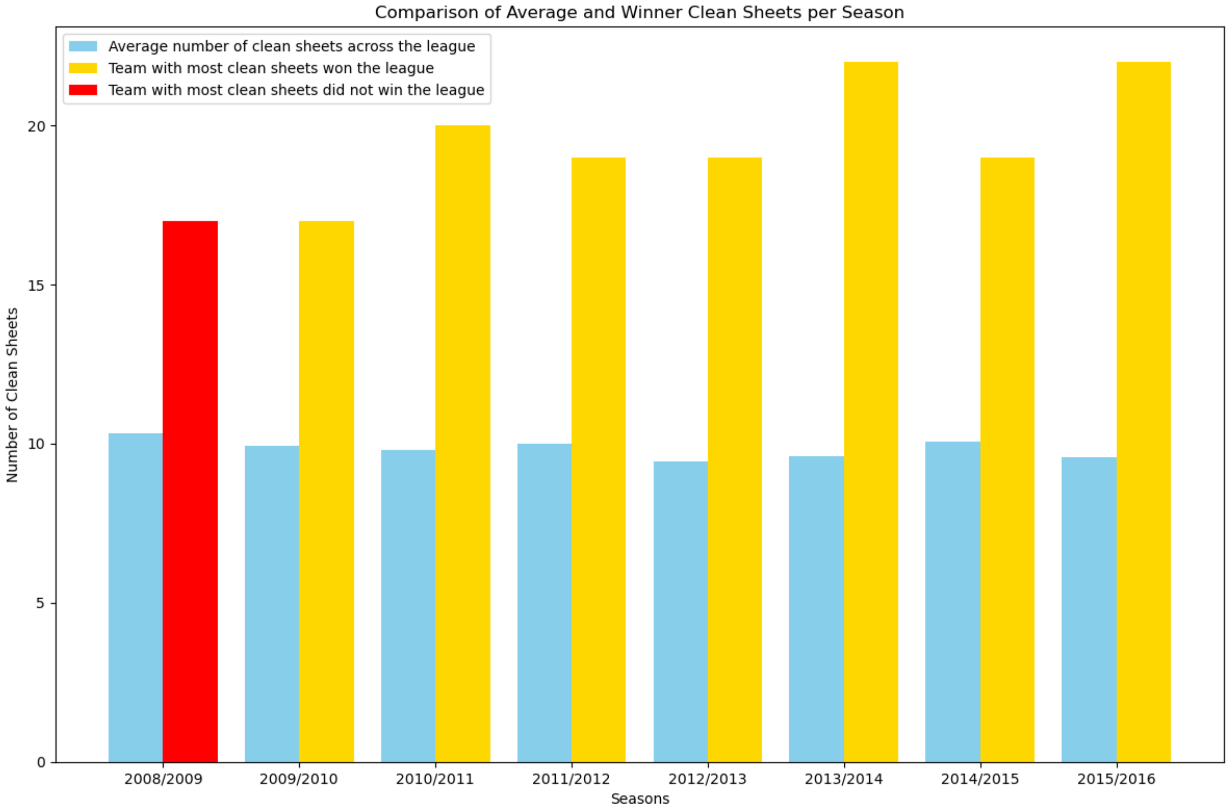
Premier League: Manchester United’s Decline

- Compared pre- and post-Sir Alex Ferguson eras.
- **Insight:** Even United’s worst Ferguson season outperformed their best post-Ferguson season.



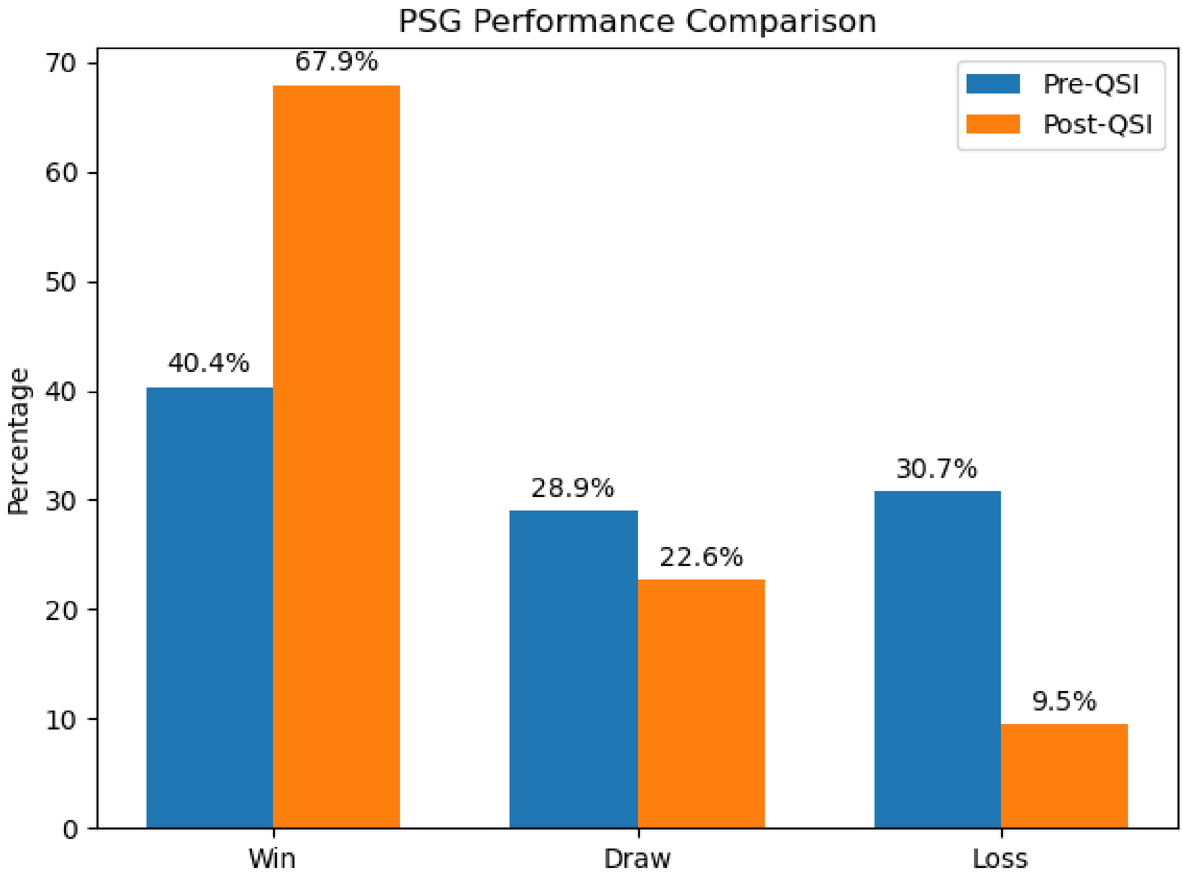
Serie A: Defensive Prowess

- Investigated if clean sheets correlated with league winners.
- **Insight:** Teams with strong defensive records often aligned with league success.



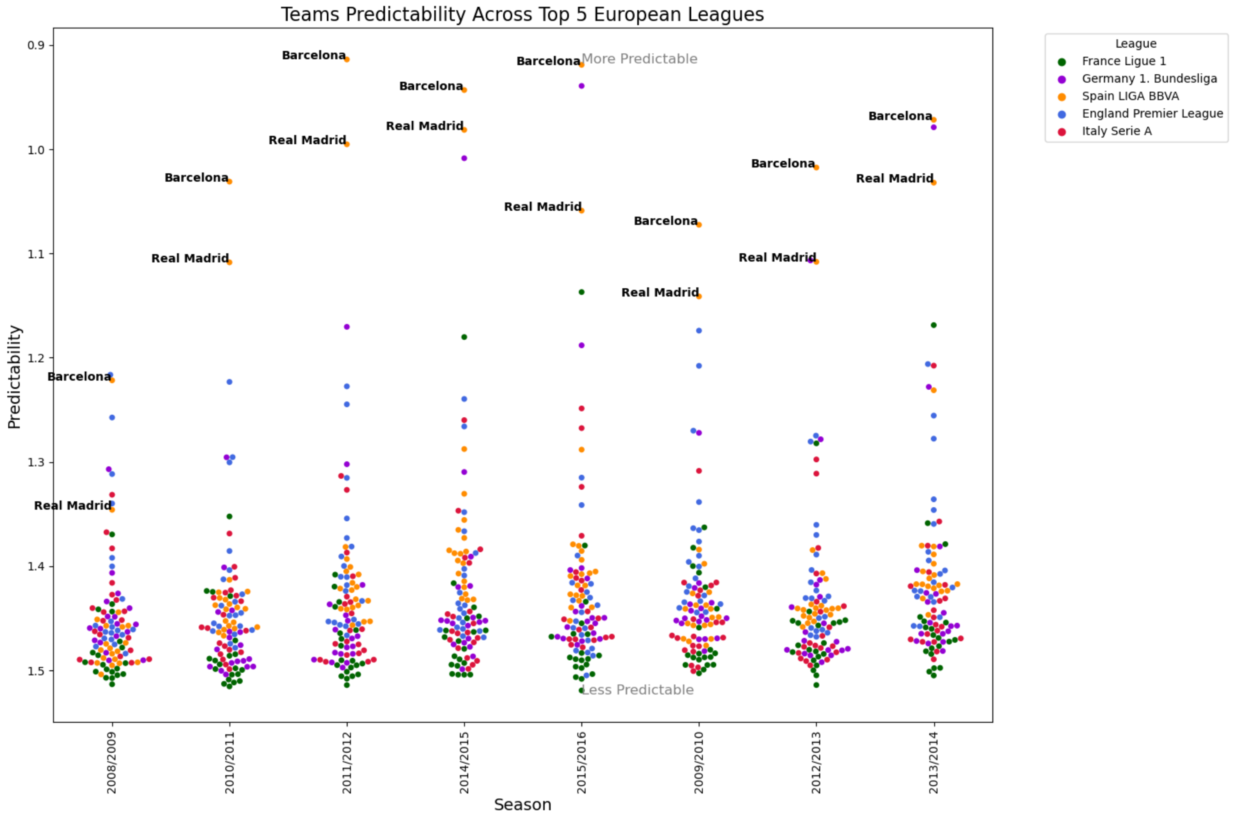
Ligue 1: PSG's Transformation

- Compared PSG's performance before and after the **QSI takeover (2011)**.
- **Insight:** Wins surged, goals scored increased, and defensive solidity improved significantly post-investment.



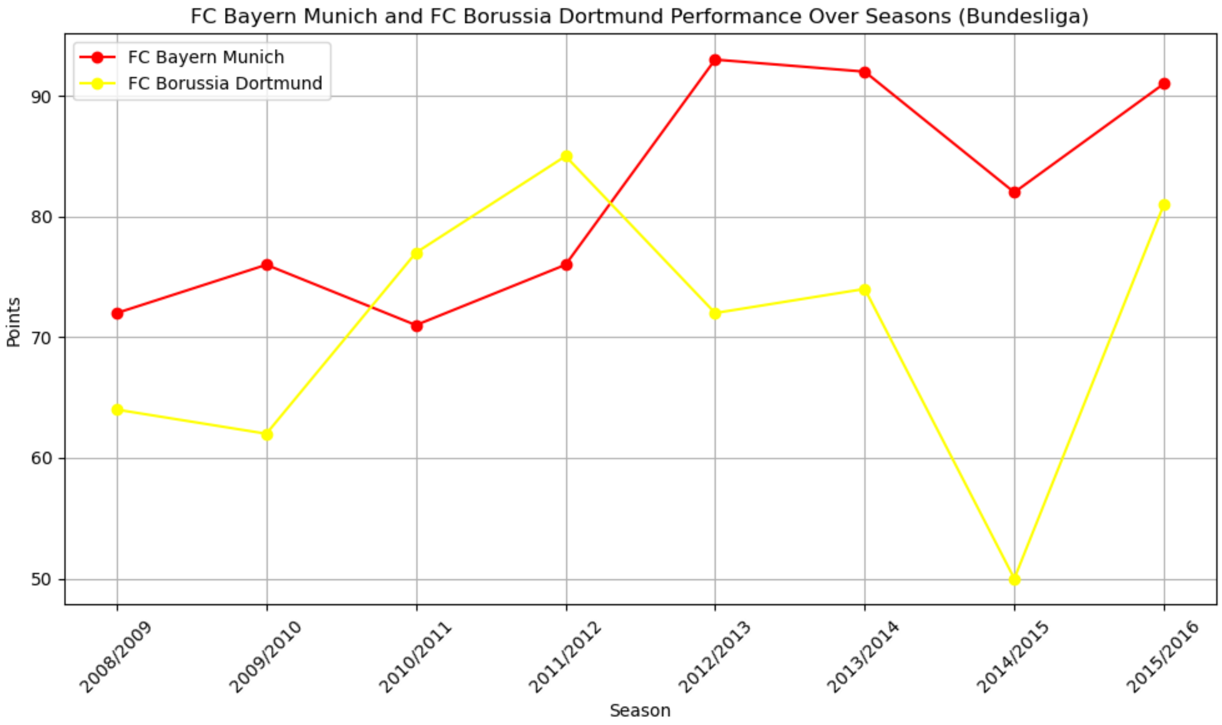
La Liga: Predictability of Dominance

- Applied **entropy** to measure predictability of match outcomes.
- **Insight:** Real Madrid and Barcelona consistently had the lowest entropy, proving their dominance.



Bundesliga: Dortmund’s Decline

- Compared seasonal points between **Bayern Munich** and **Borussia Dortmund**.
- **Insight:** Dortmund’s sharp decline post-2011/12 highlighted Bayern’s continued dominance.



Business / Analytical Impact

- Demonstrated how **data analytics can tell compelling sports stories**.
- Provided **data-backed evidence** for popular football narratives (e.g., Ferguson’s influence, PSG’s rise).
- Showcased ability to **integrate statistical measures (entropy, clean sheets)** with real-world outcomes.
- Built **reproducible workflows** for football analytics using Python.

Learning Outcomes

- Strengthened expertise in **data cleaning, statistical analysis, and visualization**.
- Applied **storytelling through data** to make technical findings accessible.
- Enhanced understanding of **sports analytics frameworks** and real-world football dynamics.

Future Enhancements

- Expand dataset to include seasons **2016–2024** for modern context.
- Incorporate advanced football metrics such as **expected goals (xG)**.
- Build **interactive dashboards (Streamlit/Dash)** for live exploration.
- Apply **predictive modeling** to forecast future league winners and team performance.