**EDA Summary Week 3**

MSD-07

Westcliff University

DATA 200: Applied Statistical Analytics

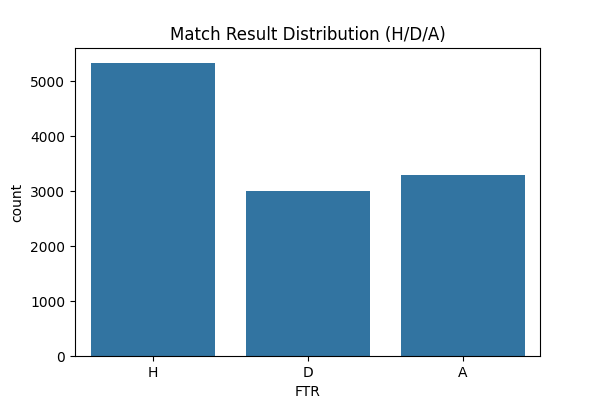
Professor Regmi

June 2, 2025

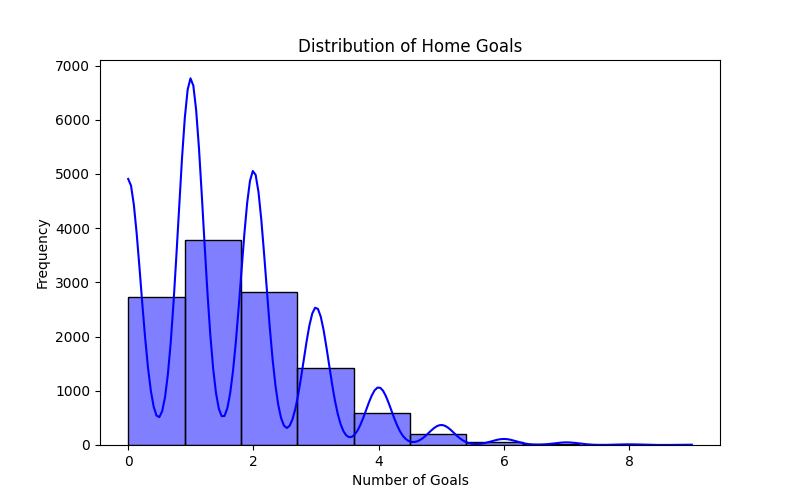
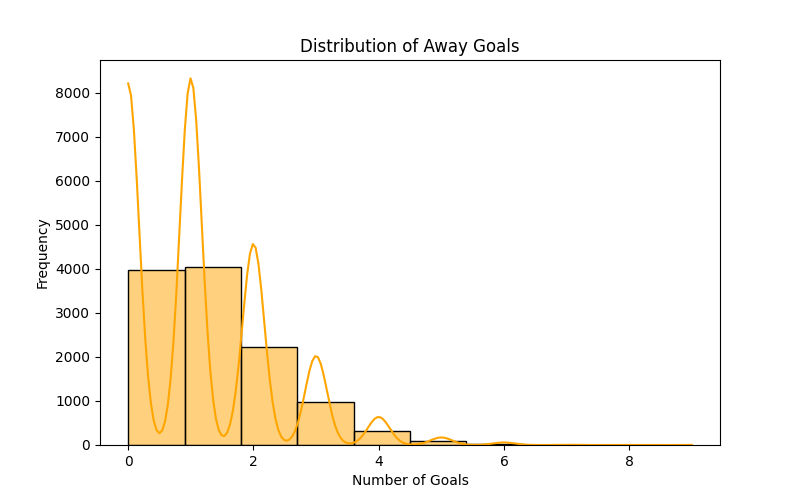
**EDA SUMMARY**

### When we explored the football match outcomes dataset, we found several important insights that could help predict match results using logistic regression. Below are our key observations from the exploratory data analysis:

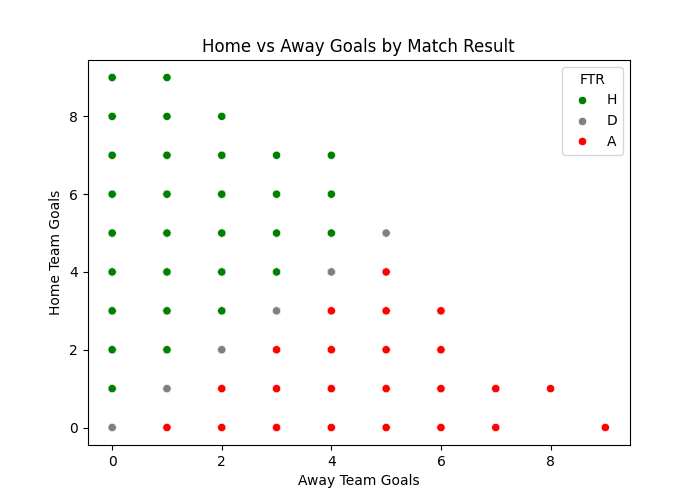
### 1. The Home Advantage is Real

The match result distribution clearly shows that **home teams win most often**. A significant number of matches were labeled as "H" (Home Win), compared to "D" (Draw) and "A" (Away Win). This supports the widely known phenomenon of **home-field advantage**, where familiarity, crowd support, and reduced travel fatigue possibly contribute to performance.  


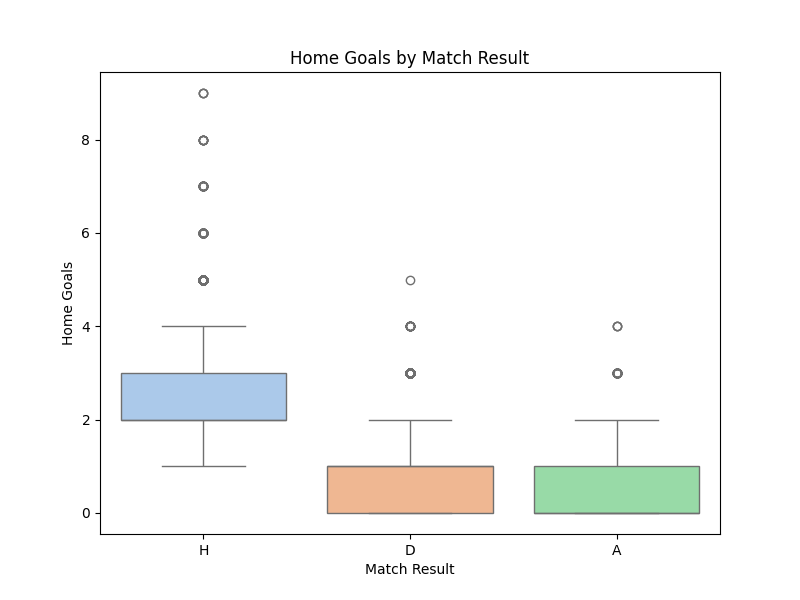
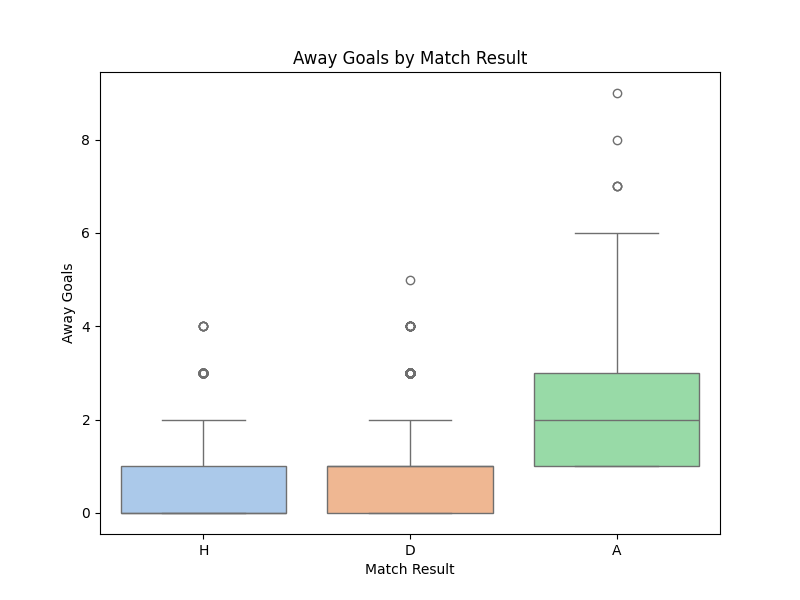
**2. Goals Are Predictable Within a Range**

Histograms of home and away goals show that **most matches end with 0–2 goals** by either side. The home goals histogram shows a slightly higher peak around 2 goals, while the away goals are skewed slightly toward fewer goals. Extremely high scores (6+) are rare.  
  


**3. Scoring Patterns and Match Outcomes**

We used a scatterplot to compare home and away goals, colored by match result. Draws tend to cluster along the diagonal line (equal goals), while wins by either team appear as clear separations. This visualization supports the idea that **goal difference is a key indicator** for predicting outcomes.  


**4. Boxplots Tell a Story**

Boxplots of home and away goals by match result helped us see where scores vary. For home wins, the median home goals are clearly higher than in draws or losses. Likewise, away wins show a higher median for away goals. These patterns reinforce the relationship between goals scored and result classification.  
  


**5. Data Integrity Looks Solid**

There were **no missing values or duplicates** in the dataset. A few outliers (like matches with 8 or 9 goals) were spotted, but all fell within a reasonable football context and do not appear to be data errors. Because the dataset is clean and balanced, we can move forward with logistic regression modeling without heavy preprocessing.

**Conclusion**

From our analysis, it’s clear that home goals, away goals, and the difference between them have a strong relationship with match outcomes. These findings justify the choice of **logistic regression** as a modeling technique. The clean dataset and clear patterns in scoring make it suitable for predicting future outcomes based on similar features.

**Dataset Source:**

📂 <https://www.kaggle.com/datasets/irkaal/english-football-results>   
(We selected this Kaggle dataset for its completeness and relevance to predicting match results.)