**Project Week 4**

MSD-07

Westcliff University

Data200: Applied Statistical Analytics

Professor Regmi

June 3, 2025

**Project Week 4**

### Data Loading and Cleaning

For this project, the football match data was imported into our Python environment from a CSV file named **football\_matches.csv** stored locally. The key Python libraries used include **pandas** for data manipulation and **seaborn** for visualization.

The dataset contains columns such as HomeGoals, AwayGoals, and Full Time Result (FTR). Initial cleaning steps involved checking for missing values, duplicates, and ensuring all columns were in appropriate data types for analysis. No missing values or duplicates were found, which means the dataset is clean and ready for modeling.

Additionally, a new target variable called **HomeWin** was created to represent the outcome as binary:

* 1 if the home team won (FTR = 'H')
* 0 otherwise (draw or away win)

We also engineered a **GoalDifference** feature calculated as HomeGoals minus AwayGoals, to help improve model predictions.

**Feature Selection and Logistic Regression Preparation**

Based on the problem of predicting match outcomes, we selected the following features:

* **GoalDifference** (numeric)
* Other potential predictors could be added later, such as team strength or season year.

The target variable for prediction is **HomeWin**, which classifies whether the home team wins or not, making logistic regression the appropriate modeling choice.

The dataset was saved into a new file, **processed\_week4\_data.csv**, for further analysis and modeling.

**Key Insights from Data**

* The distribution of match results indicates that home wins account for a significant portion of outcomes.
* Goal difference varies widely, from -9 to +9, and is expected to be strongly correlated with match results.
* The dataset is clean, enabling us to proceed with hypothesis testing and model building confidently.



