**Project Title**

**Soccer Match Data Analysis and Prediction**

**Project Objectives**

* **Goal**: To analyse soccer match data to understand goal scoring patterns and predict total goals scored in a match based on first half goals.
* **Scope**: The analysis covers multiple soccer seasons from 2019 to 2024, focusing on various match statistics and their relationships.

**Data Collection**

* **Source**: The data was collected from CSV files representing different soccer seasons.
* **Description**: The dataset includes the following variables:
  + **Div**: Division
  + **Date**: Date of the match
  + **Time**: Time of the match
  + **HomeTeam**: Home team
  + **AwayTeam**: Away team
  + **FTHG**: Full-time home goals
  + **FTAG**: Full-time away goals
  + **FTR**: Full-time result (H = Home win, D = Draw, A = Away win)
  + **HTHG**: Half-time home goals
  + **HTAG**: Half-time away goals
  + **HTR**: Half-time result (H = Home win, D = Draw, A = Away win)
  + **HS**: Home shots
  + **AS**: Away shots
  + **HST**: Home shots on target
  + **AST**: Away shots on target
  + **HF**: Home fouls
  + **AF**: Away fouls
  + **HC**: Home corners
  + **AC**: Away corners
  + **HY**: Home yellow cards
  + **AY**: Away yellow cards
  + **HR**: Home red cards
  + **AR**: Away red cards

**Data Wrangling and Tidying**

* **Cleaning**: Missing values were filled with 0. Date columns were converted to datetime format. Categorical values were standardized.
* **Transformation**: New features were created, including total goals (Total Goals) and goal difference (Goal Difference).
* **Aggregation**: Data was grouped by division and aggregated to summarize key metrics.

**Exploratory Data Analysis (EDA)**

**Descriptive Statistics**

* **Mean**: Average values for key metrics.
* **Median**: Middle values for key metrics.
* **Range**: Difference between the maximum and minimum values.
* **Correlations**: Relationships between different variables.

**Data Visualizations**

* **Univariate Exploration**:
  + Distribution of first half goals.
  + Distribution of total goals.

**Multivariate Exploration**:

* + Scatter plot with regression line showing the relationship between first half goals and total goals.

**Insights Gained from Exploration**

* **Goal Scoring Patterns**: There is a moderate to strong positive correlation (0.693) between first half goals and total goals, indicating that the number of goals scored in the first half is somewhat predictive of the total goals scored in the match.
* **Regression Analysis**: The regression model shows that first half goals are a significant predictor of total goals, with an R-squared value of 0.481 in Scenario A, 0.695 in Scenario B, and 1.000 in Scenario C.

**Analysis and Findings**

**Methods**

* **Correlation Analysis**: Calculated the correlation between first half goals and total goals.
* **Regression Analysis**: Performed linear regression to model the relationship between first half goals and total goals.
* **A/B/n Testing**: Compared different predictive models using first half goals, full-time home goals, and full-time away goals.

**Results**

* **Scenario A**: Using only first half goals, the model explains 48.1% of the variance in total goals.
* **Scenario B**: Adding full-time home goals improves the model, explaining 69.5% of the variance.
* **Scenario C**: Including both full-time home and away goals results in a perfect fit, explaining 100% of the variance.

**Conclusion**

**Summary**

The analysis shows that the number of goals scored in the first half is a significant predictor of the total goals scored in a match. Including additional variables like full-time home and away goals further improves the predictive power of the model.

**Future Work**

* **Additional Features**: Explore the impact of other variables such as shots, fouls, and corners on match outcomes.
* **Advanced Models**: Use more advanced machine learning models to improve prediction accuracy.
* **Real-Time Predictions**: Develop a real-time prediction system for live matches.

**References**

* **Pandas Documentation**: Pandas User Guide
* **Seaborn Documentation**: Seaborn User Guide
* **Statsmodels Documentation**: Statsmodels User Guide