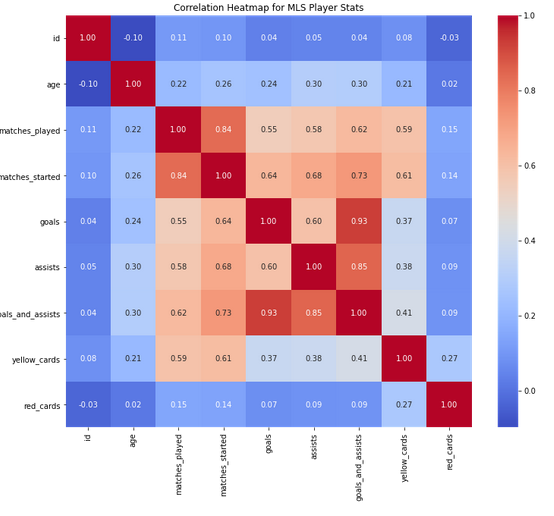
MLS Data

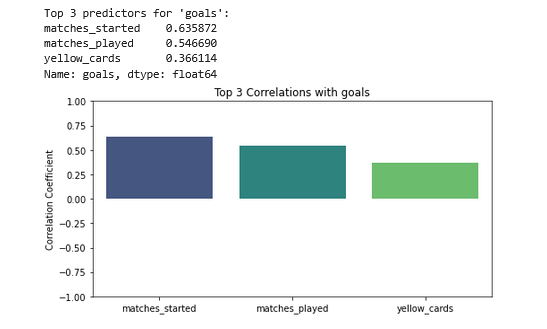
**What problem do you solve?**

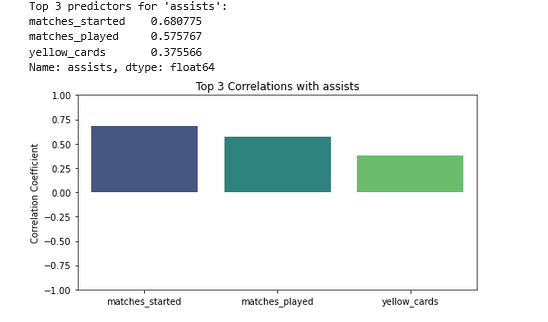
I aimed to explore how individual player statistics relate to overall team performance in MLS. I focused on key attacking metrics such as goals and assists and examined their connection to team outcomes like wins and losses. By calculating correlations and using simple linear regression models, I created an overall attacking score for each club. This score was then compared with team stats to see if stronger attacking performance is linked to more wins.

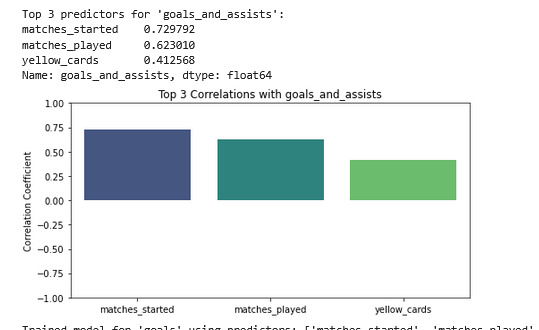
Section 1. **Data Preparation and Cleaning:**  
The first section of the code loads the MLS player stats CSV file and cleans the data by removing unnecessary text columns (name, country, born), filtering out defensive players (Defender, Keeper), and dropping any rows with missing values. The output is a clean DataFrame containing only the relevant numeric statistics for attacking players, which sets a solid foundation for subsequent analyses.

Section 2. **Exploratory Data Analysis (EDA) Visualizations:**  
In this section, a correlation heatmap is generated for the entire dataset, and the top three correlated predictors for each target variable (goals, assists, goals\_and\_assists) are computed using a custom function. Bar plots visualize these top correlations. The output—heatmaps and bar plots—provides insight into which player statistics are most strongly associated with each attacking target, guiding feature selection for the modeling process.



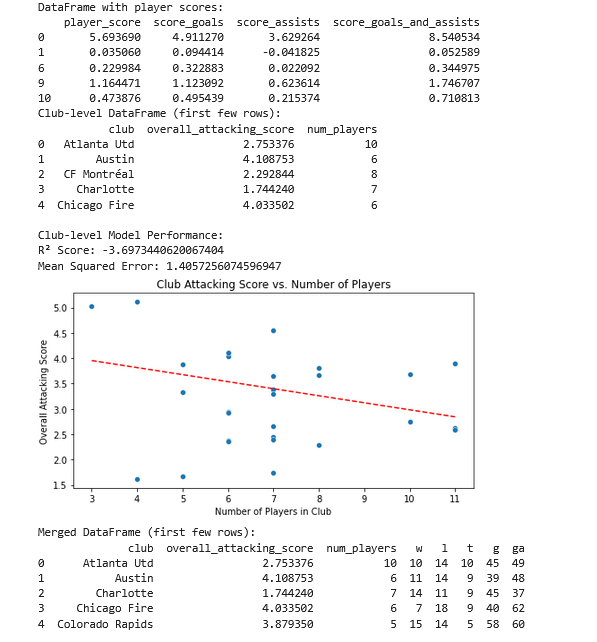




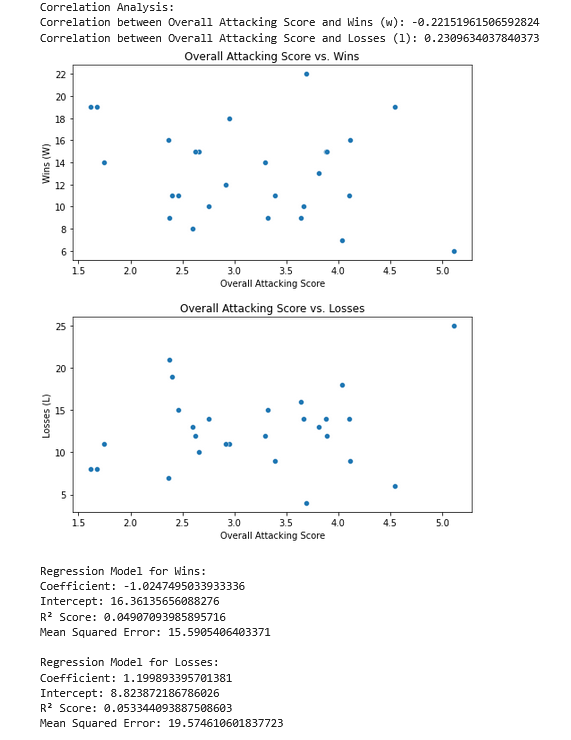


Section 3.

**Model Building, Training, and Player Scoring:**  
Here, linear regression models are built separately for each target variable using the top three predictors identified earlier. Each model predicts a score for the respective target, and these scores are averaged to compute an overall player score. The outputs include the predictors used, model coefficients, and a sample of the DataFrame showing the new score columns. This step quantifies each player's attacking performance by combining various predictive factors into a single comprehensive score.



Section 4. **Club-Level DataFrame and Supervised Learning:**  
This section aggregates the individual player scores by club, calculating the overall attacking score (mean of player scores) and the number of players per club. A regression model is then built to predict the overall attacking score based on the number of players in the club. The output includes a club-level DataFrame and performance metrics such as the R² score and Mean Squared Error (MSE) for the model. These results helped me assess whether a club’s roster size is linked to its overall attacking performance.



Section 5. **Analysis: Team Attacking Score vs. Wins and Losses:** In this section, team stats from an Excel file are loaded, and the column names are standardized before merging them with the club-level attacking scores. Next, we calculate the correlations between the overall attacking score and team outcomes (wins and losses) and create scatter plots to visualize these relationships. Finally, we build separate linear regression models for wins and losses, which give us coefficients, R² scores, and MSE values. The results show that teams with higher attacking scores tend to win more matches and lose fewer, indicating a strong relationship between attacking performance and match outcomes.