**Scenario:**

You are a Data Analyst at **Insight Sports**, a sports organization looking to optimize its match selection for the upcoming season. The management team is unsure whether to invest in more high-profile matches or focus on consistent, lower-profile games. Your job is to analyze the dataset of historical Serie A matches and provide insights to guide their decision.

The Chief Operating Officer has called you into a critical meeting. She says:

"Team, we have enough funding for one major match this season. We can't afford a mismatch that would result in a financial loss. Some of the team thinks we should focus on high-profile matches between top teams, expecting big fan attendance and high viewership. Others want to play it safe by focusing on matches between mid-tier teams, which are more consistent and might reduce the risk of a flop.

Your task is to analyze the attached dataset of past Serie A matches. Go beyond just listing the matches with the highest scores. We need a strategic recommendation, backed by your analysis, that answers the following questions:

1. **Risk vs. Reward:** What is the actual relationship between match statistics like goals (GF), shots (Sh), shots on target (SoT), and match results (Win/Loss/Draw)? Do matches with higher goals or shots lead to better outcomes (e.g., more wins or higher fan satisfaction), or are they simply higher-risk, higher-reward?
2. **The Quality Factor:** How much do metrics like **shots on target** (SoT) and **shot count** (Sh) impact a match's outcome? Can a match with fewer shots or lower goals still result in a win or positive outcome?
3. **The Sweet Spot:** Based on the data, what is the ideal range of **match statistics** (goals, shots, and shot accuracy) and **match outcomes** (win/loss/draw) for maximizing success while minimizing risk?

* **Date**: The date of the match.
* **Time**: The time the match took place.
* **Comp**: The competition (Serie A).
* **Round**: The matchweek or round number.
* **Day**: The day of the week the match was played.
* **Venue**: The venue of the match (Home or Away).
* **Result**: The result of the match (W for win, L for loss, D for draw).
* **GF**: Goals for (the number of goals scored by the team).
* **GA**: Goals against (the number of goals conceded by the team).
* **Opponent:** The opponent team that the team of interest played against.
* **Poss (Possession)**: The percentage of possession held by the team of interest during the match.
* **Attendance**: The number of people who attended the match.
* **Captain**: The name of the team captain during the match.
* **Formation**: The formation used by the team of interest during the match (e.g., 4-3-3).
* **Opp Formation**: The formation used by the opponent team during the match.
* **Referee**: The name of the referee who officiated the match.
* **Match Report**: The URL or details of the match report.
* **Notes**: Additional match notes.
* **Sh (Shots):** The total number of shots taken by the team of interest during the match.
* **SoT (Shots on Target):** The number of shots on target by the team of interest during the match.
* **Dist:** The average distance (in meters or yards) of all shots taken by the team of interest.
* **FK (Free Kicks):** The number of free kicks attempted by the team of interest during the match.
* **PK (Penalty Kicks):** The number of penalty kicks attempted by the team of interest during the match.
* **PKatt (Penalty Kick Attempts):** The number of penalty kicks actually attempted by the team of interest during the match.
* **Season:** The season year in which the match occurred (e.g., 2025 season).
* **Team**: The team that played in the match.

Task:

1. Delete unnecessary columns from dataset:

* Unnamed: 0
* Comp
* Match Report
* Notes

2. Change some columns

* Date column convert to datetime
* Change Time columns with intervals(1.morning, 2.afternoon, 3.evening)
* Change Season columns with date.year datatype
* Change Result column (Win, Loss and draw)

3.Solve null problems certain

* Attendence
* Dist
* FK

4.Find per year attendance count and use visiualization

4.1 Find attendance as time interval(Morning, afternoon, evening)

5.Find counts of days for per year

6.Find distribution of game(W, L, D) statistics

* Also find same thing for 3 main clubs

7.

* Find top 5 club which they won a lot per year
* Find top 5 club which they lost a lot per year

8.Total count of goal per season

* Find total gol for each season
* Find top 5 teams which they have a lot of GF
* Find last 5 teams which they have a few GF
* Find top 5 teams which they have a lot of GA
* Find last 5 teams which they have a few GA

9.Find count of referee

10.Correlation matrix

11. Create new columns

11.1 Shot percent

11.2 Find top 5 shot\_percent team last 5 years

11.3 Find top 5 shot percent team per year

12. Scatter Plot: Analyze relationships between two continuous variables

* Shots vs. Goals Scored (GF)
* Shots vs. Shots on Target (SoT)
* Goals Scored (GF) vs. Goals Conceded (GA)
* Shots on Target (SoT) vs. Distance (Dist)

13. Let’s try do chi test

The Chi-Square Test is used to see if there is a relationship or association between two categorical variables.

In this task, we’ll test if there is a significant difference in the **number of wins** between **home games** and **away games**. The hypothesis will be:

* **Null Hypothesis (H₀)**: There is no difference in the number of wins between home and away games.
* **Alternative Hypothesis (H₁)**: There is a significant difference in the number of wins between home and away games.

14. Let’s try do Anove test

**ANOVA** is used to compare **means** (average values) of more than two groups.

In this task, we’ll compare the **average number of goals scored (GF)** across different **seasons** using **Analysis of Variance (ANOVA)**.

* **Null Hypothesis (H₀)**: There is no difference in the average number of goals scored across the seasons.
* **Alternative Hypothesis (H₁)**: There is a significant difference in the average number of goals scored across the seasons.