**Football Evolution: Data analysis and visualization based on the UEFA European Championship dataset**

Module detail: **INF4000 Data Visualization (AUTUMN 2024~25)**

Student registration number: **240215688**

Word count minus references:**2961**

Here is a link to my GitHub homepage: **https://github.com/Sanerrrrr**

**0. Composite visualization picture display**

**0.1 Violin plots and box plots：**

图表

描述已自动生成

图表

描述已自动生成

图表, 箱线图

描述已自动生成

图表, 图示

描述已自动生成

**0.2 Principal component analysis plot:**

图表, 散点图

描述已自动生成

**0.3 K-clustering results for team technical types:**

图表, 散点图

描述已自动生成

**0.4 Linear regression plot:**

图表, 折线图, 散点图

描述已自动生成

**1.** **knowledge building**

Data analysis and visualization play a crucial role in the modern football industry, significantly enhancing team performance and decision-making efficiency. Data visualization transforms complex statistical data into intuitive charts and graphs, enabling coaches, players, and management to quickly understand and analyze key factors in a match. For instance, by visualizing offensive and defensive data, coaches can clearly identify the team's strengths and weaknesses, allowing for more targeted tactical decisions.

From the perspective of the target audience, data visualization helps analysts uncover hidden patterns behind football match data. Factors such as **home and away games**, **attendance numbers**, **individual player qualities**, and **opponents' playing styles** can be clearly presented through charts, which then aid in optimizing match strategies. Data visualization not only improves the accuracy of data analysis but also makes the decision-making process more efficient and transparent.

"As the value and application of data gradually increased, clubs now have departments of analysts dedicated to deciphering everything possible about an opponent. This has led to a considerable increase in the volume as well as types of data, leading to the development of complex key performance indicators (KPIs). The adoption of these KPIs has even been associated with improved on-field performance" (Mehta, Furley, Raabe& Memmert, 2024).

The article emphasizes the application of data analysis and key performance indicators (KPIs) in modern football, which not only helps clubs **analyze their opponents** in depth but also closely correlates with improvements in on-field performance. It suggests that a detailed analysis of the **opponent's tactical style** can help a team better adjust its match strategy, thus increasing its chances of winning. The importance of data analysis in matches is highlighted. In my clustering analysis, I roughly categorized teams into three types based on their technical focus: those strong in offense, defense, or balanced in both. Our coaching staff can adjust the match strategy according to the opponent.

García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2021). From the combination of both techniques, we obtained useful conclusions to enhance the performance of players and to identify positions on the field.

Undoubtedly, **player positioning and physical fitness** are also key factors in determining the outcome of a match, as they directly affect the team’s tactical execution and overall performance. In my research, I first conducted a principal component analysis on the players’ physical attributes, exploring how height and weight influence their performance. Second, I performed linear regression on player positions and years, finding that players in different positions have varying average goal-scoring numbers, but all players' goal counts increased over the years. This can help data analysts predict the goal-scoring performance of players in different positions.

Lastly, I analyzed whether there are significant differences in the number of goals scored in home and away matches, as well as how audience attendance changes over time. These findings can assist clubs in improving their business strategies.

The **number of goals scored at home and away**, along with audience attendance, are crucial for the development of a club. Home advantage often enhances team performance, while away goals reflect the team’s adaptability and competitiveness. Audience attendance is not only a source of financial revenue but also increases the club's brand value and fan base. By analyzing these data, clubs can develop more effective business strategies, optimize matchday experiences, improve fan loyalty, and ultimately drive overall development.

**2. Theoretical framework**

**2.1 Explanation of the theoretical framework**

In my visualization, I combined the ASSERT framework with Grammar of Graphics (GoG) elements to ensure that the chart is not only data interpretable but also visually intuitive. Each stage of the ASSERT framework is reflected in my data visualization, and I will explain my thought process step by step.

**A (Ask)**: First, I clarified the core theme of the research, which is how analyzing factors like attendance, home and away scores, and player height and weight can optimize team performance and business strategies.

**S (Structure)**: To answer this question, I chose suitable visualization methods. I used violin plots and box plots for attendance, home and away scores, and match types, as these graphs help display data distribution and outliers. I visualized the impact of player height and weight through Principal Component Analysis (PCA). Cluster analysis was used to show the tactical characteristics of different teams, while linear regression revealed the goal-scoring trends over time for players in different positions.

**S (Simplify)**: To simplify the analysis, I used clear visual structures like violin and box plots to avoid data overload. Additionally, PCA reduced the multidimensional impact of height and weight on player performance.

**E (Elaborate)**: In further explanation, I differentiated categories using different colors. The cluster analysis used different colored points to distinguish tactical types, and in linear regression, different colored lines represented player positions, clearly displaying goal trends over time for each position.

**R (Refine)**: After completing the initial visualizations, I refined the charts to ensure clarity. By adjusting layouts, colors, and axis settings, I enhanced the visualization’s effectiveness, making the key data points stand out.

**T (Tell a Story)**: Finally, all the visualization elements came together to tell the story of how teams can use data analysis to optimize strategies, improve player performance, and adjust tactics in different match settings.

Similarly, in Grammar of Graphics (GoG), I used several key elements: First, I employed various geometries, such as violin plots, box plots, and regression lines, to represent different data features through shapes and positions.

Second, from an aesthetic perspective, I used colors, sizes, and shapes to convey differences in the data. For example, the **progressive blue color** represents player weight, different colors in cluster analysis distinguish tactical types, and the regression lines are colored to reflect player positions.

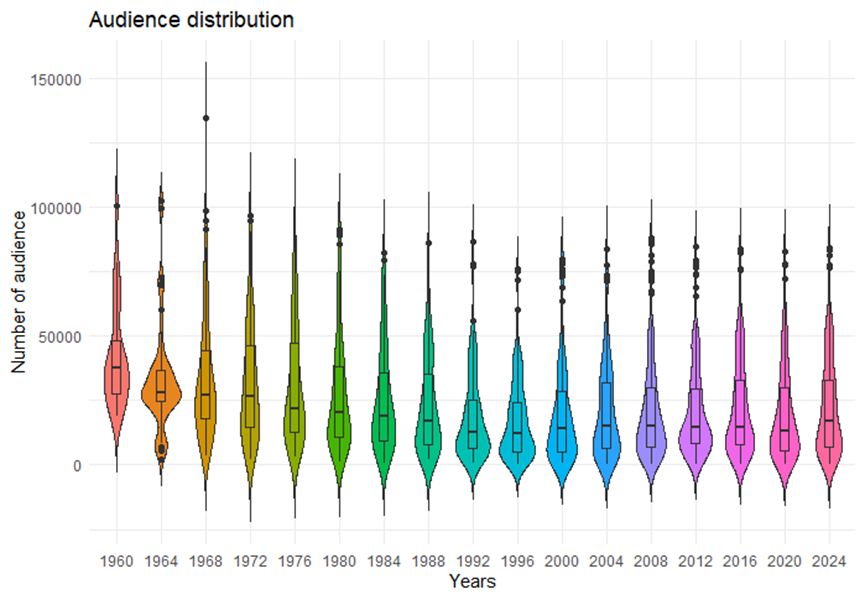
Lastly, I mainly used a two-dimensional coordinate system to show the relationships between variables, especially in the regression analysis where the X-axis represents time, and the Y-axis shows the number of goals, clearly illustrating the goal trends.

Through the careful design of these GoG elements, my visualization not only effectively presents complex data but also helps viewers better understand how data analysis can optimize team performance.

**2.2 Answer the questions**

**2.2.1 Attendance Distribution**

**Question: How has audience attendance changed over time in different years?**

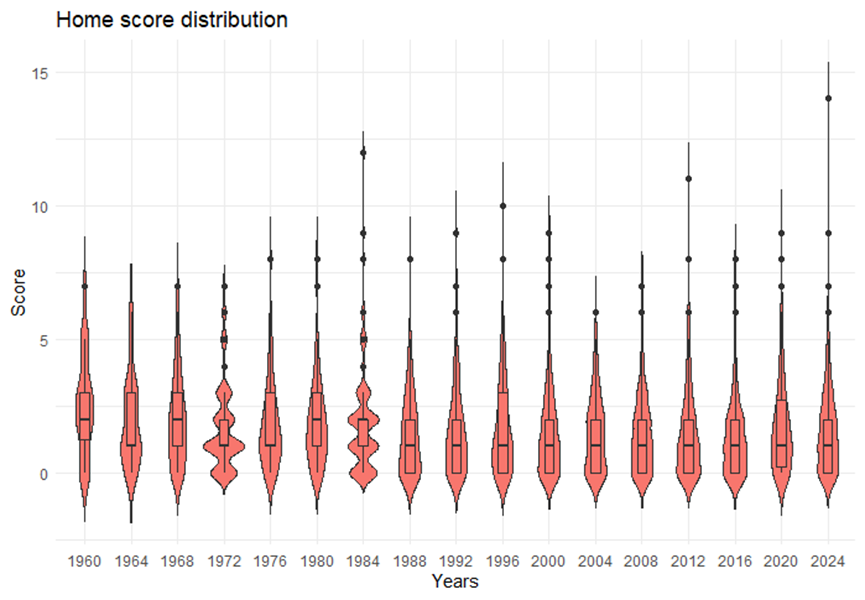


Looking at the overall trend, the chart shows fewer attendees before 1970, with the attendance distribution being more concentrated. After 2000, there was a noticeable increase in the number of attendees, with many years showing attendance close to or above 50,000.

Focusing on outliers, the chart shows that after 2000 (e.g., in 2016 and 2020), some matches had attendance reaching 100,000, which is a significant rise. This might be linked to major events, like the European Championship or World Cup qualifiers. However, on the left side of the chart, we also see years (such as 1960, 1964, and 1968) with much lower attendance.

**2.2.2 Home Scores Distribution**

**Question: Is there a significant fluctuation or trend in home scores?**

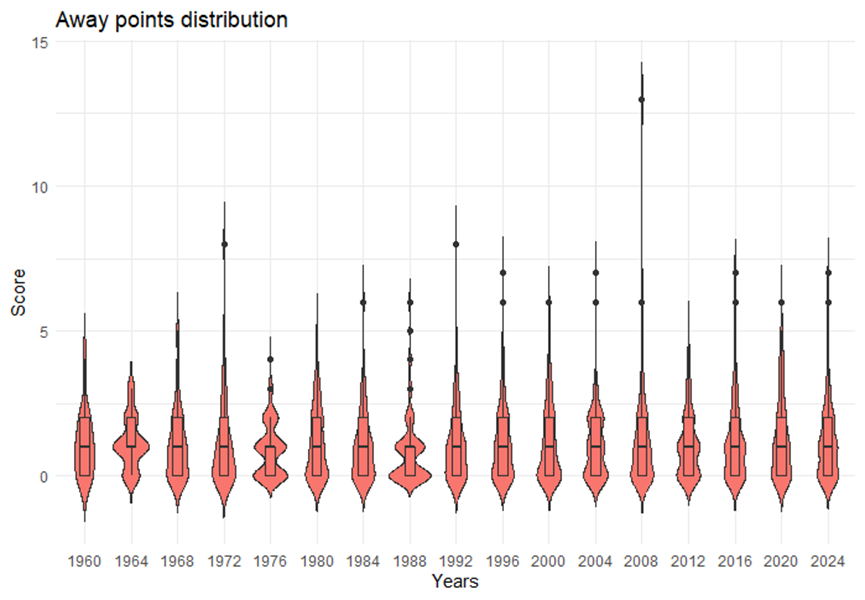
****

From the chart, I can see that most years’ home scores are concentrated between 0 and 5.

Overall, the score range is fairly even for most years, but some years, like 1984 and 1988, show a wider spread, with lower medians. In 2016 and 2020, there were notably higher scores, with longer upper whiskers and outliers in the box plot. This could be linked to specific events or rule changes.

**2.2.3 Away Scores Distribution**

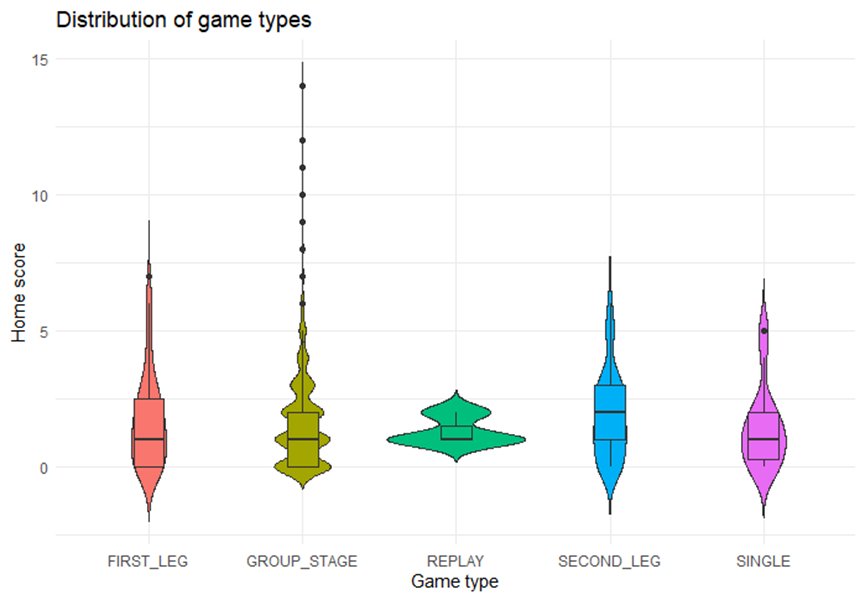
**Question: Is there a significant difference between home and away scores?**

****

Looking at the away scores’ distribution, most years show scores concentrated between 0 and 5, similar to home scores, but with slightly more variation. In 1972 and 1984, there were higher away scores, almost reaching 10, with larger outliers in the box plot. A comparison reveals that the median away score is slightly lower than the home score, suggesting that away matches generally have lower scores.

**2.2.4 Match Type Distribution**

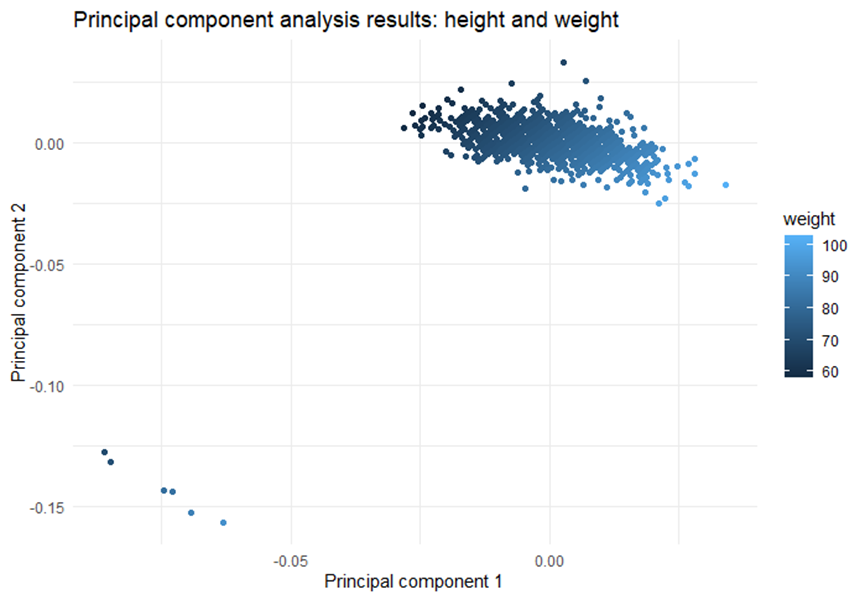
**Question: Does the score distribution differ between match types?**

****

According to the match type distribution chart, there are clear differences in the score distributions, showing that match type significantly affects scoring. Particularly, first-round matches and group-stage games have lower scores, with many matches scoring in a smaller range.

**2.2.5 Principal Component Analysis (PCA) Chart**

**Question: How do height and weight impact player performance?**

****

The PCA results show a strong trend between Principal Components 1 and 2, with a clear linear relationship. From the chart, we can see that players weighing more than 80kg are concentrated on the positive side of Principal Component 1, while lighter players (less than 70kg) are distributed on the negative side.

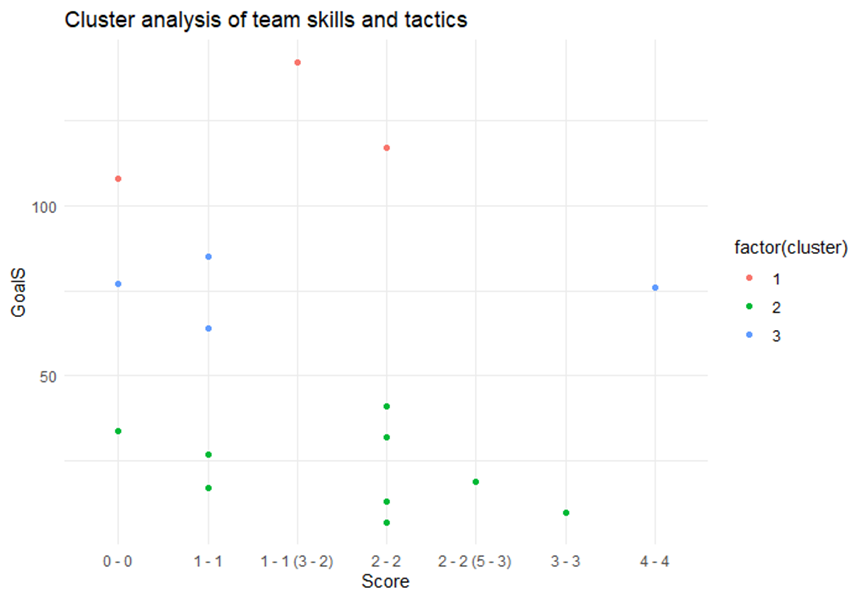
文本

描述已自动生成

Observing the **second row** of the PCA result chart, I found that PC1 = 0.75207 and PC2 = 0.24793, indicating that Principal Component 1 captures about 75.21% of the variance in the data. This is a high percentage, showing the importance of PC1 in explaining the data’s variation.

**2.2.6 Cluster Analysis Chart**

**Question: What are the tactical styles of the teams?**

****

Cluster Analysis Interpretation:

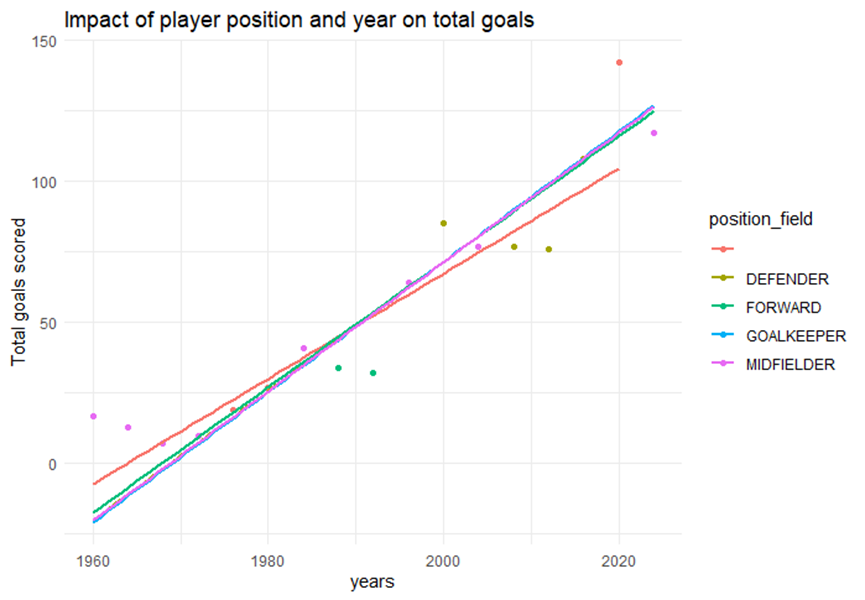
The **red** cluster represents teams that are more aggressive both technically and tactically, or those with stronger scoring abilities.

The **green** cluster shows teams with more balanced or defensive tactical characteristics. These teams score more evenly in matches, likely focusing more on defense and controlling the game pace.

Teams in the **blue** cluster may show more conservative technical and tactical behavior, reflecting a more cautious approach or weaker offensive capabilities.

**2.2.7 Linear Regression Chart**

**Question: How can we predict the performance of different players in future matches?**

****

Model Results Analysis:

图形用户界面

中度可信度描述已自动生成

From a statistical significance perspective, all position coefficients and the yearly coefficient have a p-value less than 0.001, indicating that the variables in the model are highly significant in predicting the total number of goals.

The chart shows the trend of total goals scored by players in different positions over time, with a clear upward slope in the linear fit line. The increase in years has a significant positive impact on the number of goals, suggesting that as time passes, football may have become more offensive, with more goals scored. This could also reflect improvements in offensive and defensive techniques, leading to higher goal counts. **This is an interesting conclusion.**

Recent data shows that the average number of goals per game in the Premier League has increased significantly, reaching levels not seen in decades. For example, the average number of goals per game in the 2023-24 season reached 3.16. (The Analyst, 2023) This coincides with my findings, and I stand by this article.

**3. Accessibility**

**3.1 Meaning of Accessibility in Visualization**

Accessibility in data visualization means ensuring that all viewers, especially those with visual impairments, can understand and interact effectively with the charts. This includes providing sufficient contrast, color alternatives, and clear layouts for colorblind, blurry vision, or visually impaired users.

Additionally, the structure of graphic elements and the way information is presented should be simple and clear so that viewers can intuitively grasp the data.

**3.2 Discussion on Visualization Accessibility**  
In my visualization, multiple colors are used to distinguish different categories and data features. For example, **weight is shown in a gradient blue**, **clustering analysis uses different colors to represent tactical styles**, and **regression lines use different colors to differentiate player positions**.

"Accessible data representations, however, have lagged behind, leaving areas of information out of reach for many blind and visually impaired (BVI) users." (Fan, Danyang, et al., 2023)

While these designs make the data presentation richer, they may pose difficulties for colorblind users, especially those with red-green color blindness. The contrast between the gradient blue and other colors may not be sufficient, potentially making it hard for some users to distinguish between data sets. To improve accessibility, I would consider using higher-contrast color combinations or adding different shapes and patterns for each category, allowing colorblind users to clearly identify information in the chart.

**3.3 Impact of Design Changes on Accessibility**

" Data visualization plays a vital role in modern scientific communication across diverse domains, shaping the understanding of complex information through color choices. "(Luca Nelli, 2024)

Some of my design choices may hinder accessibility, especially the use of colors. Colorblind users may struggle to differentiate between the different clusters and regression lines. Therefore, adding auxiliary elements like different line styles, icons, or graphical symbols would help improve the accessibility of these charts.

**3.4 Adjusted visualization output image**

图表, 散点图

描述已自动生成

After considering the accessibility of data visualization, I adjusted my program to design the points of different clusters into different shapes, which would better highlight the differences between them. After that, if there are readers with visual impairments, they can also interpret the results of the image return through different shapes.

**4. Visual Selection**

In my choice of visualization methods, I selected four common chart types, each tailored to the specific characteristics of the data and the analysis goals to generate easy-to-interpret visual results.

**4.1 Violin and Box Plots (Attendance, Home/Away Scores, Match Types)**

**Reason**: Violin and box plots are ideal for showing data distributions, extremes, and outliers. They clearly depict the fluctuation in attendance and scores over different years.

**Advantages**: Violin plots provide information on data density, while box plots highlight quartiles and medians, making it easier to compare differences between groups.

**Limitations**: These charts may not be intuitive for viewers without a statistical background and require additional explanation for complex data.

**Alternative**: Density plots and bar charts can serve as replacements. Density plots visually present data density, while bar charts are good for displaying the distribution of different categories.

**4.2 Principal Component Analysis (PCA) Plot (Impact of Height and Weight on Performance)**

**Reason**: PCA is used to simplify high-dimensional data, making it easier to highlight the combined effects of height and weight on player performance.

**Advantages**: PCA plots reveal correlations and potential patterns between variables.

**Limitations**: It doesn’t provide detailed insights into each individual variable's impact, which might be confusing for viewers. The plot also requires data standardization and may be difficult to understand for those unfamiliar with PCA.

**Alternative**: Scatter plots can be an effective substitute, directly displaying the relationship between height, weight, and performance.

**4.3 Cluster Analysis Plot (Team Technical Characteristics)**

**Reason**: Cluster analysis categorizes teams based on their playing styles, helping to reveal tactical patterns and allowing for comparisons between teams.

**Advantages**: Cluster plots distinguish different groups using colors or shapes, making them easy to understand and analyze.

**Limitations**: Cluster results may be influenced by data noise and feature selection. Different clustering methods may yield different outcomes, making interpretation challenging.

Despite these limitations, I find clustering to be the most suitable method for this type of data.

**4.4 Linear Regression Plot (Predicting Player Performance)**

**Reason**: Linear regression effectively shows the trend of goals scored by players over time, making it easier to observe the model's fit and the influence of player position on goal counts.

**Advantages**: Linear regression plots provide clear trend lines, ideal for showing long-term changes in time-based data.

**Limitations**: It cannot capture non-linear relationships, and when there is significant fluctuation in the data, the model's applicability may decrease.

**Alternative**: Non-linear regression or time series plots can serve as substitutes. Non-linear regression (e.g., logistic regression) better fits non-linear relationships, while time series plots are more suitable for visualizing trends in time-based data.

By carefully choosing these visualization methods, I can select the most appropriate tools for each dataset while offering flexible alternatives to better address various data characteristics and analysis needs.

**5. Impact and Improvement**

When using visualizations to analyze football-related data:

First, we must consider ethical issues, especially regarding player privacy, and ensure that their unauthorized personal information is not disclosed.

Second, data can be misinterpreted or misrepresented. Another concern is that presenting incomplete or biased data through visualizations may mislead the audience. In my visualizations, I use different colors to distinguish teams or player positions. If the audience is unfamiliar with the data or if the color choices are not suitable for color-blind users, this could lead to misunderstandings. Poorly designed visualizations may lead to incorrect conclusions about player performance or team tactics, misleading fans, analysts, and even coaches.

Another example occurred during the 2020 Tokyo Olympics, when a dynamic chart showing "the total medal counts of countries" went viral on various media and social platforms. The chart used color gradient bars to display the number of medals each country won but did not clearly distinguish between gold, silver, and bronze medals. The bar chart overemphasized the number of gold medals while ignoring the total number of other medals. The color gradient used in the chart did not clearly differentiate between gold, silver, and bronze medals, leading some viewers to mistakenly believe that certain countries achieved high rankings solely through gold medals, neglecting the contributions of silver and bronze medals.

**My proposed solution** is:

First, distinguish between medal types. I would ensure that the total medal count is not only represented by the number of gold medals but also by using a layered or stacked bar chart to separately highlight the number of gold, silver, and bronze medals.

Then, using dynamic visualization, I would incorporate interactive visual charts, allowing users to click on a country to view detailed medal information.

With these improvements, viewers would not only gain a more accurate understanding of the contributions of each country's medals but also avoid misunderstandings of the competition results due to unclear representation of medal types.

**News link:**

The Guardian – "Tokyo 2020 Olympics: How the medal table is misleading you"

<https://www.theguardian.com/uk>

**6. References**

Mehta, S., et al. (2024). Examining how data becomes information for an upcoming opponent in football. *International Journal of Sports Science & Coaching*, 19(3), 978–987. https://doi.org/10.1177/17479541231187871

García-Aliaga, A., et al. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, 16(1), 148–157. https://doi.org/10.1177/1747954120959762

Nelli, L., & Qureshi, M. R. N. (2024). Color Quest: An interactive tool for exploring color palettes and enhancing accessibility in data visualization. *PLOS ONE*, 19(3), e0290923–e0290923. https://doi.org/10.1371/journal.pone.0290923

Fan, D., et al. (2023). The accessibility of data visualizations on the web for screen reader users: Practices and experiences during COVID-19. *ACM Transactions on Accessible Computing*, 16(1), 1–29. https://doi.org/10.1145/3557899

The Analyst. (2023). The Premier League is producing more goals than ever before. Retrieved from <https://www.theanalyst.com/2023/01/premier-league-goal-increase-analysis/>

Obata, T., & Izumi, S. (2022). Analysis and visualization of team performances of football games. *Japanese Journal of Statistics and Data Science*, 5(2), 885–898. https://doi.org/10.1007/s42081-022-00173-z

Andrienko, G., et al. (2017). Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, 31(6), 1793–1839. https://doi.org/10.1007/s10618-017-0513-2