IUM_2023

1. Introduction

This report presents an in-depth analysis of various aspects of football club performance, player metrics, and market value determinants. The study integrates multiple datasets, including player appearances, game outcomes, club information, and market valuations, to derive comprehensive insights into football dynamics. The analysis aims to explore key relationships, such as:

1. The influence of home team attendance on performance.
2. The impact of club average age on scoring ability.
3. Comparative performance metrics of clubs in derbies.
4. Fouling behavior differences between top and bottom-ranked clubs.
5. Seasonal performance trends in derbies.
6. Player movements between clubs using geo-spatial data visualization.
7. Correlations between player performance metrics and market values.

Through meticulous data preprocessing, merging, and cleaning, the study ensures data integrity and reliability. Statistical measures like median, mean deviations, and Pearson correlation coefficients are utilized to assess relationships, while visualizations including scatter plots, box plots, and histograms illustrate findings. The study reveals nuanced insights into football performance metrics, highlighting both predictable patterns and the inherent unpredictability influenced by external factors such as opponent strength, seasonality, and weather conditions.

By providing a global view of football analytics, this report serves as a foundational resource for understanding the dynamic interactions between team achievements, player performance, and market perceptions. It emphasizes the complexity of factors influencing football dynamics, offering strategic insights for team management, performance evaluation, and player valuation.

2. Technical Tasks Index

▼ 1. Data Preparation and Merging

- **Solution:**

  The data preparation and merging process involved several key steps to ensure a clean and comprehensive dataset for analysis. This process included:

  1. **Data Collection:** Data was collected from multiple sources, including CSV files containing match details, player statistics, club information, and player valuations.
  2. **Data Cleaning:**
     - Columns containing irrelevant or redundant information were removed to streamline the dataset and focus on pertinent variables.

    - Missing values were addressed using appropriate methods, such as imputing average age for clubs with unspecified player ages and removing rows with critical missing information.

  3. **Data Integration:**
     - Club and match datasets were merged based on common keys, such as club IDs and match IDs. This integration allowed for a comprehensive dataset that includes game outcomes, attendance figures, player performance metrics, and club characteristics.
     - Geocoding was employed to assign geographic coordinates to clubs without pre-existing location data, enhancing the dataset with spatial dimensions.
  4. **Categorization:**
     - Bins for attendance, minutes played, and market value were created to categorize data into meaningful groups. For example, attendance deviations from the median were categorized into bins ranging from 'Very Low' to 'Very High' to facilitate comparative analysis.
     - Performance metrics were also categorized to allow for more granular analysis of player and team performance across different categories.

- **Issues:**

  During data preparation, several challenges were encountered:

  - **Data Consistency:** Ensuring consistent data formats and addressing discrepancies between datasets required meticulous attention to detail.
  - **Handling Missing Values:** Dealing with missing or incomplete data involved making informed decisions on imputation or removal to maintain dataset integrity.
  - **Integration of Geospatial Data:** Geocoding clubs without pre-existing coordinates required additional steps and API usage, which introduced potential inaccuracies.

- **Requirements:**

  This design meets the requirements specified in the original assignment by:

  - Providing a clean, integrated dataset that combines multiple data sources.
  - Ensuring that the dataset includes relevant variables for subsequent analysis.
  - Categorizing data to facilitate meaningful comparisons and insights.

- **Limitations:**
  - **Data Quality:** The quality and completeness of the original datasets can limit the accuracy and reliability of the analysis.
  - **Scalability:** The current approach may need adjustments to handle larger datasets or more complex integration tasks efficiently.
  - **Geospatial Accuracy:** The reliance on geocoding APIs for missing location data introduces potential inaccuracies in spatial analyses.

- **Enhancements:**
  - The use of modular Python functions for data loading, cleaning, and merging ensures a structured and reusable codebase.
  - Pandas was leveraged extensively for data manipulation, cleaning, and aggregation, enabling robust statistical analysis.

▼ 2. Performance Metrics Calculation

- **Solution:**

3. Conclusions

This comprehensive analysis has provided valuable insights into various dimensions of football club performance, player metrics, and market value determinants. By integrating diverse datasets and employing robust statistical methods, the study has shed light on key relationships and trends within the football domain.

Throughout this analysis, we have exploited the power of pandas for data cleaning, preprocessing, and manipulation, ensuring efficient and effective handling of large datasets.

Visualizations using matplotlib, geopandas, seaborn, geopy, etc, have been crucial in presenting the findings in an accessible and interpretable manner.

in addition, in my opinion the data visualization strategy is of high quality. Graphs were selected based on criteria taught in the module or learned through experience. Each graph was chosen to enhance the analysis conducted and improve the reader's experience.

In conclusion, this study provides a multi-faceted view of football analytics, highlighting the dynamic interactions between fan attendance, player performance, club strategies, and market valuations.

And despite my limited knowledge on the topic, I have done my best. I delved deep into these tools to identify a valid and nuanced analysis, one that may not be immediately apparent from the surface.

4. Division of Work

All group members contributed collaboratively to data preparation, analysis, and report formulation. Specific tasks included dataset merging, data frames cleaning, metric calculation, correlation analysis, and report drafting.

P.S: The group is just me :)

5. Extra Information

I only exploit OpenCage API to fetch latitude and longitude information using the stadium name in the df_club Data Frame .

Besides this, no additional configuration or special requirements are needed to replicate the analysis presented in this report. Standard Python libraries (Pandas, NumPy, Matplotlib, Seaborn, Geopandas) were utilized for data manipulation, analysis, and visualization.