IUM_2023 - Data Analysis [Mattia Firrisi]

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:**
    - The data cleaning has been used to remove columns field from futiles information and NaN or undefined values. Doing so, the whole code execution will result lite, optimized and more readable.
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:**
  - The calculation of performance metrics was a crucial step in our analysis, aimed at deriving meaningful insights from the prepared dataset. This involved several specific tasks:
    1. Calculation of Goals and Points:
    2. Correlation Analysis:
    3. Attendance Impact Analysis:
    4. Age and Performance Analysis:
    5. Derby Performance Analysis:
    6. Player Movement Analysis:
- **Issues:**

  Several challenges were encountered during the calculation of performance metrics:
  - **Data Integration:** Merging datasets accurately to ensure all relevant metrics were included for each analysis.
  - **Correlation Analysis:** Interpreting weak correlations and understanding the multifaceted nature of player valuation and performance.
  - **Handling Outliers:** Identifying and addressing outliers that could skew results, especially in attendance and goal metrics.
- **Requirements:**

  This design meets the requirements specified in the original assignment.

  Some Python functions have been employed to ensure modularity and clarity. These also enhance the components' usability and readability. For instance, data loading, data visualization, and data manipulation tasks are encapsulated within separate modules (using different Jupyter cells), reducing redundancy and promoting code reusability.

  Pandas has been leveraged extensively to perform robust data cleaning and preprocessing, data manipulation, statistical analysis, and data visualization. Pandas was instrumental in cleaning and preprocessing raw data from CSV files, handling missing values, and ensuring consistency in data formats. Using pandas DataFrame operations, complex data manipulations were seamlessly executed. For example, merging multiple datasets based on common keys (e.g., club IDs).

  The application of statistical methods such as standard deviation and Pearson's correlation coefficient was facilitated by pandas. These methods were employed to identify trends, correlations, and outliers in the data. Pandas facilitated data aggregation and summarization operations, setting up data for visualization libraries like matplotlib and seaborn. Aggregated datasets were transformed and formatted using pandas to ensure compatibility with visualization functions.

  In summary, by harnessing the capabilities of pandas for data manipulation, cleaning, statistical analysis, and visualization preparation, the analysis workflow has been streamlined and made more effective.
- **Limitations**
  - **Weak Correlations:** The weak correlations observed suggest that additional factors not captured in the dataset may significantly influence player valuation and performance.
  - **Data Quality:** Variability in data quality and completeness could impact the accuracy of the metrics calculated.
  - **External Factors:** External factors such as match context, player injuries, and team strategies are not found in the Data Set.

▼ 3. Correlation Analysis

- **Solution:**
  - The correlation analysis was a key component of this study, aimed at identifying and understanding the relationships between various performance metrics and other factors such as Home Team Attendance and Performance.
- **Issues:**

Several challenges were encountered during the correlation analysis:

  - *Weak Correlations:* Some correlation coefficients highlight the complexity of factors, suggesting that performance metrics alone do not capture the full picture.
  - *Data Quality and Outliers:* Variability in data quality and the presence of outliers can skew correlation results, making it essential to handle these issues carefully.
  - *External Factors:* Many external factors, such as team strategies, economic conditions, and match contexts, are not captured in the datasets, impacting the correlation results.

- **Requirements:**

  This design complies with the requirements specified in the original assignment by:

    - Providing detailed correlation analyses between key performance metrics and other factors like market value and attendance.
    - Leveraging various Python libraries such as pandas, geopandas, matplotlib, geopy, seaborn, etc. 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.

- **Limitations**
  - *Human Limitation:* My limited knowledge on the topic can definitely hack on the analysis conducted.
  - *Data Variability***:** Variations in data quality and completeness can affect the reliability of the correlation results.
- **Enhancements**
  - Modularity and clarity were prioritized in the code structure, with separate functions for different correlation analyses
  - Visualization techniques, such as GeoPandas, scatterPlot, BoxPlot, ect., were employed to clearly present the correlation results.

▼ 4. Data Visualization

- **Solution**
  - Data visualization played a crucial role in this study, allowing us to communicate complex relationships and findings clearly and effectively. Various visualization techniques were recruited to illustrate the results of our analyses, providing insights into the dynamics of player performance, market value, and team attendance.
- **Issues**

  Several challenges were encountered during the data visualization process:

    - *Complex Data Relationships:* Capturing the complexity of relationships between multiple variables in a single plot required careful consideration of visualization techniques.
    - *Categorization***:** Effectively categorizing data (e.g., home attendence) to provide meaningful insights while avoiding clutter and confusion in the visualizations.
    - **Outliers:** Handling outliers in the data to ensure they do not misrepresente the visual interpretation of the results.

- **Requirements**

  This design complies with the requirements specified in the original assignment by:

    - Selecting graph type 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.
    - Ensuring that each visualization is appropriately labeled and annotated to provide context and enhance understanding.
    - Using color schemes (when was possible) and categorization to distinguish between different data groups and highlight key insights.

- **Limitations**

    - *Visual Clarity:* While efforts were made to ensure clarity, some complex relationships may still be challenging to interpret visually.
    - *Data Variability:* The presence of outliers and variability in data quality can affect the visual interpretation of trends and correlations.
    - *Dynamic Factors:* External factors influencing the data are not always visually represented, potentially oversimplifying complex relationships.
- **Enhancements**
    - *Modularity and Reusability:* Visualization functions were modularized (when was possible) to enhance reusability and maintainability, allowing for easy updates and adjustments.
    - *Annotations and Legends:* Detailed annotations and legends were included to provide additional context and aid in the interpretation of the visualizations.

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 leveraged 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.

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

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.

In my opinion, the attached documentation deserves a high grade because I have provided detailed comments and descriptions for each conducted analysis. I have enhanced them with concise yet meaningful abstracts. I anticipate a high evaluation because I have pushed my limits in understanding that the difference between a good analysis and a meaningless one lies in finding a compelling narrative.

4. Division of Work

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

P.S: The group is just me :)

5. Extra Information

I only exploited the OpenCage API to fetch latitude and longitude information using the stadium name in the df_club DataFrame.

I also used standard deviation, for example, to calculate the highest fluctuations in goals conceded over time. I utilized Pearson's correlation coefficient, for example, to assess the relationship between home team attendance and home team performance.

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.