Kicking Insights: Exploring the Correlation of Soccer Player Attributes and Their Relation to Overall Player Ability

Marcus O. Thompson

Western Governors University

Table of Contents

A. Project Highlights

The research question:

What player attributes (speed, height, passing ability, ball control, etc.) are most closely correlated to a player's overall rating? Understanding the correlations between players' abilities and their overall rating can guide strategic decision-making for the soccer organization. This information can be useful to team management, and enhance overall team performance and competitiveness on the field.

The scope of the project:

- **Data collection:** The project data is publicly available and sourced from Kaggle.com.
- **Data cleaning:** The data was cleansed using Python (via the Pandas library) in a Jupyter Notebook.
- **Data analysis:** The data was analyzed by using Python. The analysis produced correlation coefficients, p-values, and VIF (Variance Inflation Factor) values as a way to quantify the level of correlation between multiple soccer player attributes and their overall rating.
- **Data visualizations:** Key insights were visualized via heatmaps, bar charts, and regression plots.

- **Project Presentation:** A detailed video summary was created documenting the analysis process. The presentation includes any discoveries made, visualizations, and recommendations for the stakeholders.

Not included in the scope of the project:

- **Evaluation of attributes based on nationality:** Some regions of the world may have different styles of play, for example focusing more on defensive techniques, or aggressive play which could influence player attributes. While this may be true, the nationality of players will not be considered in the analysis.
- **Temporal analysis:** The dataset was analyzed in its original form as it was downloaded, and only historical data was used.

Overview of analytic solution:

My analytics solution was designed to determine the attributes most closely correlated with a soccer player's overall rating. The project utilizes a dataset sourced from Kaggle.com, derived from FIFA 23, a soccer simulation game. The primary goal is to provide valuable insights for strategic decision-making in player recruitment, team composition, and game-night strategies.

Tools and environments:

Jupyter Notebooks: I chose Jupyter Notebooks as the programming environment where Python and its various libraries would be employed. This provides an interactive and collaborative platform for data analysis and visualization.

Python libraries:

- **Pandas:** Used for data preparation, cleaning, and storage. It ensures the dataset is well-structured for analysis.
- **Scipy and Statsmodels:** Employed for hypothesis testing, specifically for calculation of correlation coefficients and p-values, and testing multicollinearity using VIF values.
- **Seaborn and Matplotlib:** Utilized for creating visualizations such as heatmaps, bar charts, and scatterplots.

Methodology:

For the project I the SEMMA (Sample, Explore, Modify, Assess) methodology:

- **Sample:** The dataset was downloaded and evaluated
- **Explore:** Initial exploration involved checking for appropriate features, null values, duplicates, and dimensions. Unnecessary columns were removed.

- **Modify:** After initial analysis and statistical models were created (correlation matrices), the dataset was modified by aggregating related attributes (defensive, pace, movement categories) due to strong multicollinearity.
- **Model:** Statistical models, such as correlation matrices were employed to uncover patterns in the data.
- **Assess:** The reliability and limitations of the models were evaluated to ensure they met the objectives of the analysis.

Methodology: As mentioned above, I used the SEMMA methodology for this analysis.

Statistical Significance Evaluation:

I employed the Pearson correlation test to assess statistical significance. This involved creating a correlation matrix, focusing on correlation coefficients of 0.65 or above (or -0.65 and below), and calculation of p-values. The null hypothesis was that there is no strong correlation between the overall rating and any other attribute being tested. If p-values were below the alpha of .05, it indicated statistical significance.

My solution includes visualizations such as heatmaps, bar charts, and regression plots to effectively communicate the results to stakeholders. The criteria for success involve producing statistical evidence, concise visualizations, and a detailed report (video presentation) documenting findings, visualizations, and recommendations.

B. Project Execution

The project adhered closely to the initial plan outlined in the project proposal. Below is a summary of the project plan, objectives, deliverables, and timeline. There was no variance in the project plan, apart from the timeline, as the project was started earlier than expected.

Project plan: The primary objective of this project is to provide actionable insights for strategic decision-making in soccer organizations, specifically in areas such as player recruitment, team composition, and game-night strategies. I did this by finding what player attributes most closely correlate to a player's overall rating. Below is a summary of the goals, objectives, and deliverables of the project.

Goal: Determine what (if any) are the attributes most closely correlated to a player's overall rating. Are any of the ratings strongly correlated (with a correlation coefficient of 0.65 or greater, or -0.65 and below)?

Project objectives and deliverables:

- Objective 1: To import and cleanse the dataset, keeping only the features necessary for analysis.

- Deliverable 1.1: A clean dataset that will be used for this analysis, and future analysis if necessary.

- Objective 2: To analyze the dataset, producing statistical evidence of what attributes (if any) are most closely correlated to the overall rating
  - Deliverable 2.1: A Jupyter notebook elaborating the analysis process, which will provide statistical tests (correlation coefficients, p-values, VIF values) either proving or refuting a hypothesis
- Objective 3: To create easily understood visualizations of how the attributes correlate to the overall rating.
  - Deliverable 3.1: Heatmaps and bar charts showing the level of correlations, as well as regression plots showing the strength and direction of the correlations.
- Objective 4: To produce a video summary of the project, presenting findings to the stakeholders.
  - Deliverable 4.1: A Panopto recording providing a summary of the research question, demonstration of analytical techniques used, and an outline of the findings and implications of the analysis

Summary of the updated project timeline and milestones:

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Data collection and cleaning	1 day	Dec. 18, 2023	Dec. 18, 2023

Data analysis	3 days	Dec. 19, 2023	Dec. 21, 2023
Data visualizations	1 day	Dec. 22, 2023	Dec. 22, 2023
Produce report	2 days	Dec. 25,, 2023	Dec. 26, 2023
Present video report	1 hour	Dec. 27, 2023	Dec. 27, 2023

C. Data Collection Process

I adhered to my initial plan for selecting and collecting data, acquiring the FIFA dataset in .csv format from Kaggle.com. The data collection process proceeded smoothly without any challenges, and there were no unexpected issues related to data governance.

C.1 Advantages and Limitations of Data Set

Dataset advantages:

3. The dataset was quite extensive, offering ratings for a multitude of player attributes. This allowed for a very detailed, granular approach to discovering which player attributes are most closely related to a player’s overall rating.

4. The dataset was also extensive in the sense that it held information for almost 20,000 unique players, which allows for a comprehensive representation of the diverse attributes and characteristics present in the population of soccer players.
5. The dataset was complete, having no null values.

Dataset disadvantages:

6. The dataset had many irrelevant columns that needed to be removed. For example, columns for the players' names, jersey numbers, or links to the players' images had no bearing on the analysis performed. Due to the large amount of features the dataset held (89 columns), care was needed to sift through and decide what was the most relevant to keep.

D. Data Extraction and Preparation

Data importation and cleansing:

The data extraction phase of this project proceeded smoothly and was uncomplicated. I used the Safari web browser to access Kaggle.com, initiating a search for pertinent data. Upon locating a suitable dataset, I used the Pandas library within my Jupyter Notebook to import the dataset. After importing the dataset into a data-frame, I then performed the necessary cleansing and preparation steps for the project. Overall, the dataset required minimal cleansing, as there were few duplicates and no null values. It also was well formatted, with only a few columns requiring adjustments to their names for consistency. Using the Pandas library to execute this process was appropriate because it was created for this type of data manipulation, allowing the process to be fast and simple.

Data preparation:

Once the dataset was cleansed, I began the data preparation phase of the project. I began by investigating what columns were present, which were necessary, and which were superfluous and should be removed. Among the columns present was 'Best Position', which shows a player's most skilled position. There were also many columns for goalkeeper ratings. Since the goalkeeper is the most specialized position, I decided to remove all rows for goalkeepers, as well as all columns for goalkeeper ratings. I then removed unnecessary columns such as players' names, their salaries, the length of their contracts, and many others. This was an extensive process, as the dataset held 18,539 rows and 89 columns before data cleansing and preparation, and finally, 16,367 rows and 33 columns after the preparation was completed.

The data preparation undertaken was aimed at enhancing the dataset's usability and streamlining it for analysis. By carefully selecting and transforming relevant features, the process resulted in a more appropriate dataset without significant data loss. Additionally, excluding goalkeepers and their ratings was a strategic decision to ensure a more accurate analysis, given the specialized nature of their position. For instance, goalkeepers typically have markedly low ratings in attributes such as 'shooting', 'dribbling', and 'passing', which, if included, could introduce bias and compromise the precision of the analysis. This targeted exclusion helps maintain the integrity of the dataset for a more reliable and insightful investigation of player attributes.

E. Data Analysis Process

E.1 Data Analysis Methods

The project primarily employed exploratory and descriptive analysis to address the central organizational question: "What attributes of soccer players are most closely correlated to their overall rating?" The use of these analytical methods was essential to uncover meaningful insights that could inform strategic decisions related to player recruitment, team composition, and game-night strategy.

The hypothesis guiding this analysis posited the existence of at least one attribute correlated with a player's overall rating, indicated by a correlation coefficient of 0.65 or higher (positive correlation), or -0.65 and below (negative correlation). These coefficients provide a quantitative measure, offering insights into the degree of influence specific attributes may be related to the overall performance rating.

In tandem with correlation coefficients, p-values played a critical role in assessing the statistical significance of observed relationships. A low p-value indicates that the identified associations are unlikely to have occurred by random chance. By applying this measure, the analysis sought to distinguish genuine correlations from those that could arise merely due to the variability inherent in the data, ensuring robust and reliable findings.

VIF (Variance Inflation Factor) values were incorporated to address multicollinearity concerns. Multicollinearity occurs when predictor variables are highly correlated, potentially affecting the stability and interpretability of regression models. By evaluating VIF values, the analysis aimed to identify and mitigate issues related to intercorrelations between attributes, ensuring the integrity of the results and providing a clearer understanding of the unique attributes of each attribute to the overall rating.

Finally, bar charts and regression plots were created. These are a useful way to convey information simply and quickly to those with less technical expertise. For example, a sorted bar chart was created to show each attribute and its level of correlation as an easy way to convey which attributes are most correlated. Regression plots were also created for attributes with the highest levels of correlation, as these offer a visual representation between a specific attribute and the overall rating. The graphical representations serve as a powerful tool for conveying complex statistical findings intuitively, supporting effective communication and decision-making in the context of player evaluation and team strategy.

E.2 Advantages and Limitations of Tools and Techniques

<u>Tools:</u>

Jupyter Notebooks:

Advantages: Interactive and collaborative Python environment, supporting reproducibility and interactive development

- **Limitations:** Efficiency may decline with very large datasets, and version control complexities may arise in collaborative projects.

Pandas library:

- **Advantages:** Powerful for data manipulation, efficient handling of tabular data, and simplified complex operations.
- **Limitations:** Scalability challenges with massive datasets, and demanding memory usage for larger analyses.

Scipy and Statsmodels libraries:

- **Advantages:** Comprehensive statistical tools for scientific computing, hypothesis testing, and statistical modeling.
- **Limitations:** Steeper learning curve for novices, and potential gaps in covering specialized statistical tests.

Seaborn and Matplotlib libraries:

- **Advantages:** Instrumental for creating appealing and diverse visualizations.
- **Limitations:** Customization intricacies and challenges with certain complex plots.

Techniques:

Correlation coefficients:

- **Advantages:** Quantifies the strength and direction of variable relationships, aiding in a precise understanding of attribute correlations with the overall rating and identifying statistically significant associations.
- **Limitations:** Assumes linear relationships, potentially overlooking complex non-linear dependencies. Causation is not implied, and coefficients may be influenced by outliers.

P-values

- **Advantages:** Provides a statistical measure of observed correlation significance, enhancing the reliability of the analysis by distinguishing genuine relationships from chance occurrences.
- **Limitations:** Sensitive to sample size, leading to potential instability with small samples. P-values do not quantify practical significance.

VIF values:

- **Advantages:** Identifies and addresses multicollinearity issues, ensuring regression model stability. Quantifies the impact of intercorrelations for a more robust analysis.
- **Limitations:** Susceptible to predictor choice influence, and high VIF values may not fully resolve multicollinearity complexities. Interpretation requires careful consideration of the context. (StatLect, 2023)

E.3 Application of Analytical Methods

A variety of analytical methods were employed to carry out this project. The following outlines the sequential steps undertaken throughout the entire process.

1. **Data importation:** The dataset was imported into a Jupyter Notebook environment with Pandas 'read_csv'.
2. **Exploring and cleansing the data-frame structure:**

2.1. Dimensions of the data-frame were checked with 'df_shape'.

2.2. The first 10 rows were displayed by using 'df.head(10)'.

2.3. The type of data held in each column was displayed with 'df.info()'.

2.4. A sum of the duplicated rows was displayed with 'df.duplicated().sum()'.

- The duplicated rows were removed with 'df.drop_duplicates()'
- The new dimensions were viewed with 'df.shape'

2.5. A sum of null values was created with 'df.isnull().sum()', no nulls were present.

2.6. Check for outliers by using 'df.describe()', taking note of the 'min' and 'max' rows.

2.7. Various columns were renamed due to typos with 'df.rename(columns={'Columns of interest'}

3. **Removal of unnecessary features:** Columns holding data specific to goalkeepers were removed, as well as all rows of players whose best position was 'goalkeeper'. Other unnecessary columns were removed, such as those that did not hold data regarding attribute ratings, as well as columns that held links to player or team images.

3.1. I removed all columns and rows specific to goalkeepers, aggregate columns, and all columns that did not hold numeric attribute ratings. I did this by creating a 'list' variable holding all the columns to be removed, and then 'df.drop(columns = columns_to_drop)'.

3.2. I then checked the new dimensions of the data-frame with 'df.shape', and also viewed the first five rows using 'df.head() to verify the changes had been made. The data-frame initially had 18,539 rows and 89 columns, and after this step now has been reduced to 16,367 rows and 33 columns.

4. **Statistical analysis:**

4.1. First, I created a correlation matrix using 'df.corr'

4.2. Then, I displayed a heatmap of the correlation matrix with Seaborn's 'sns.heatmap()'

4.3. Next, I created a function to calculate and display correlation coefficients and p-values for an individual attribute and other attributes of our choosing. The output of the function is a new dataframe holding those values. I then displayed the data-frame comparing the ‘Overall’ rating against all other attributes, in descending order. This gave me a sorted list of all correlation coefficients and p-values of each attribute vs. the ‘Overall’ rating.

5. Further statistical analysis

Within the data-frame, apart from ‘Age’, ‘Height’, and ‘Weight’, all attributes belonging to a category were aggregated. For example, the ‘Pace’ category includes the ‘Sprint Speed’ and ‘Acceleration’ attributes, while the ‘Physical’ category includes attributes such as ‘Strength’ and ‘Stamina’. After investigating the heatmap that was created earlier, I noticed many of the attributes within the same category had similar correlation coefficients. At this point, I decided to group the attributes by category and test them for multicollinearity. The process is outlined below:

5.1. I created a function to display a heatmap, and list correlation coefficients and VIF values for a list of attributes (to be chosen by the user). I then created six lists of attributes, each holding related attributes. Each list represents a category, such as ‘Pace’, ‘Shooting’, ‘Passing’, ‘Movement’, ‘Defending’, and ‘Physical’.

5.2. I then fed each list I created into the function created above, in order to display the correlation coefficients and VIF values of these related attributes. Once completed, I noticed that many (but not all) attributes within the same category are closely correlated, and had VIF values ranging from 19.9 to 258.8, suggesting multicollinearity.

5.3. Next, I tried a more streamlined approach by creating a new, simplified data-frame that aggregates all attributes within their respective categories. The values in each newly generated column were determined by calculating the mean of all attributes within their designated categories. This simplification aimed to enhance the interpretability of the dataset and facilitate a more focused analysis of overarching categories, such as ‘Pace’, ‘Shooting’, ‘Defending,’ and others. However, to maintain this streamlined structure, all individual attributes previously included were deliberately dropped from the data-frame. While this approach presented advantages in terms of a more concise and comprehensible dataset, it also introduced certain trade-offs and considerations that warrant careful examination. Some pros and cons of this approach are listed below:

Advantages:

- **Simplified representation:** Aggregating attributes within categories reduces the complexity of the dataset, providing a more straightforward and interpretable representation of the information
- **Improved multicollinearity handling:** Grouping related attributes can help manage multicollinearity issues. The creation of mean-based

summary columns may mitigate the problem of highly correlated variables, providing more stable results in subsequent analyses.

Disadvantages:

- **Loss of granularity:** The process of aggregating attributes into mean-based summary columns involves a loss of granularity. Individual attribute variations within categories are no longer distinguished, potentially overlooking specific nuances.
- **Information loss:** The original dataset contains detailed information about each attribute. Aggregating them into summary columns sacrifices some level of information, and in cases where specific attributes are crucial, this loss may impact the analysis.

Once the summary dataset was created, I then performed the same analysis process as I did on the original dataset, as outlined in Step 4, creating a correlation matrix and heatmap, and calculating correlation coefficients and p-values for the summary data-frame.

6. **Visualizations:** After calculating correlation coefficients and p-values for each attribute in both the original cleansed and summary data-frames, I created visuals to convey the results of the analysis.

   6.1. For the original data-frame, I used Seaborn to create a horizontal bar that displays each attribute and its level of correlation to the 'Overall' rating.

6.2. I then used Seaborn to produce regression plots for the top three most correlated attributes to the ‘Overall’ rating, and then the bottom three attributes least correlated to the ‘Overall rating.

6.3. I then repeated the process for the summary data-frame, creating a horizontal bar chart that displays each category of attributes and its level of correlation to the ‘Overall’ rating.

6.4. Finally, I repeated the steps outlined in step 6.2, using Seaborn to create regression plots showing the top and bottom two most and least correlated attributes to the ‘Overall’ rating.

The assumptions and requirements of the analysis were verified throughout the process. Data integrity was ensured during the exploration and cleansing phase through checks for dimensions, duplication, and null values. Assumptions related to attribute grouping and multicollinearity were tested, with the creation of summary columns and subsequent statistical analyses serving as confirmatory steps. Additionally, visualizations and statistical tests were applied to both the original and summary datasets, validating the consistency and reliability of the analytical approach across different representations of the data.

F. Data Analysis Results

F.1 Statistical Significance

Statistical test overview:

- Null hypothesis: There is no strong correlation (correlation coefficient 0.65 or above, or -0.65 and below) between the overall rating and any other attribute being tested.
- **Test:** Pearson correlation test between the 'Overall' attribute against all other attributes, resulting in a calculated correlation coefficient and p-value.
- **Alpha (a):** 0.05

Results: Of the 32 attributes tested against the 'Overall' attribute, four provided evidence that the null hypothesis should be rejected.

- **"Reactions"**
  - Correlation coefficient: 0.87
  - P-value: 0
- **"Composure"**
  - Correlation coefficient: 0.81
  - P-value: 0
- **"Short Passing"**
  - Correlation coefficient: 0.78

    - P-value: 0

- **“Ball Control”**
    - Correlation coefficient: 0.76
    - P-value: 0

Conclusion:

Based on the correlation analysis, the calculated correlation coefficients for the attributes ‘Reactions’, ‘Composure, ‘Short Passing’, and ‘Ball Control’ are all significantly high, ranging from 0.76 to 0.87, with p-values of 0. Therefore, there is sufficient evidence to reject the null hypothesis. The test suggests a strong correlation between these attributes and the ‘Overall’ attribute, supporting the claim that these attributes are highly correlated with the overall performance of soccer players. (The Data School, 2019)

Analysis of the summary data-frame:

Using the same hypothesis, statistical test, and alpha (a) as above, the test was performed on the data-frame comprised of categories of aggregated attributes.

Results:

Of the nine attributes tested against the ‘Overall’ attribute, only one attribute provided evidence to reject the null hypothesis.

- **“Movement”**
    - Correlation coefficient: 0.68

- P-value: 0

Conclusion:

The calculated correlation coefficient and p-value for the ‘Movement’ category were significantly high (0.68, p-value: 0), providing sufficient evidence to reject the null hypothesis. The test suggests a strong correlation between the ‘Movement’ category and the ‘Overall’ attribute, supporting the claim that this category is highly correlated with the overall performance of soccer players.

F.2 Practical Significance

The data analytics solution presented in F1 holds substantial practical significance by directly addressing the determinants of a soccer player’s overall rating – a pivotal metric in team management and player development. The identification of specific attributes strongly correlated with overall performance offers valuable insights for strategic decision-making in player recruitment, team composition, and game-night strategies. For instance, teams can leverage this knowledge to make more informed choices, optimize resources, and enhance competitiveness in the dynamic world of professional soccer. By tailoring their approach based on the attributes highlighted in this study, teams can potentially gain a competitive edge.

Furthermore, the practical applications of this study extend beyond individual teams to contribute to the broader field of sports analytics. The identification of key attributes contributing significantly to overall player ratings can inform the development of standardized performance metrics and player evaluation criteria. This not only benefits soccer teams, but also provides valuable insights for sports analysts, coaches, and researchers aiming to refine their understanding of factors influencing player success. The potential impact of this study extends to shaping data-driven practices in soccer and establishing a framework for similar analyses in other sports, thereby facilitating advancements in talent management and strategic decision-making across the sports industry.

As an example of practical application, consider a soccer club that, armed with the knowledge from this analysis, decides to prioritize the recruitment of players with exceptional ‘Reactions’, ‘Composure’, ‘Short Passing’, and ‘Ball Control’ attributes. This focused recruitment strategy, guided by data-driven insights, aims to enhance the team’s overall performance by targeting players with attributes strongly correlated with success. The application of such targeted strategies aligns with the broader trend toward data-driven decision-making in sports and underscores the practical significance of the findings in guiding real-world actions for soccer teams and beyond.

F.3 Overall Success

The project has successfully proven the hypothesis that certain soccer player attributes strongly correlate with overall ratings, as demonstrated by high correlation coefficients and statistically significant p-values. The creation of clear visualizations, including heatmaps, bar charts, and regression plots, effectively communicates these findings to stakeholders with varying technical expertise. This report captures all results, recommendations, and limitations, fulfilling the predefined criteria for success outlined in Task 2. Overall, the project has achieved its objectives, providing actionable insights for soccer management and analytics.

G. Conclusion

G.1 Summary of Conclusions

The project has successfully identified key soccer player attributes that exhibit strong correlations with overall ratings. Through comprehensive statistical analyses, including correlation coefficients and p-values, I determined that attributes such as ‘Reactions’, ‘Composure’, ‘Short Passing’, and ‘Ball Control’ significantly correlate to a player’s ‘Overall’ performance rating. The analysis extended to a summary data-frame, revealing the noteworthy correlation of the ‘Movement’ category with overall ratings.

Practical significance arises from actionable insights derived, empowering soccer organizations in player recruitment, team composition, and game-night strategies. The project’s success lies in its ability to distill complex data into concise visualizations, meeting stakeholder needs for easily interpretable information.

However, it is crucial to acknowledge certain limitations. The study primarily relies on correlation analyses, implying associations but not causation. Additionally, the scope of the analysis may not encapsulate all relevant factors influencing player performance. Further research could explore dynamic player interactions, contextual game situations, or external factors affecting gameplay.

In conclusion, the project achieves its goals by providing valuable insights for soccer management and contributing to the evolving landscape of data-driven decision-making in sports.

G.2 Effective Storytelling

The visualizations in this project were chosen to effectively convey complex information, providing a clear and insightful narrative. Key visualizations include heatmaps, bar charts, and regression plots, all created using Python libraries such as Seaborn and Matplotlib within the Jupyter Notebook environment.

The heatmap serves as a tool to showcase correlation matrices, allowing stakeholders to visually identify patterns and strengths of relationships among attributes. This graphical representation aids in the initial exploration of the dataset, highlighting areas of interest for further analysis.

Bar charts play a crucial role in presenting attribute correlations concisely. The sorted bar chart efficiently communicates the level of correlation for each attribute with the overall rating, enabling stakeholders to quickly identify the most correlated attributes. This simplicity enhances interpretability for individuals with varying levels of technical expertise.

Regression plots, both for the original and summary data-frames, provide a deeper understanding of the relationships between the top and bottom correlated attributes and the overall rating. These plots offer a visual narrative that complements the quantitative findings, making it easier to grasp the practical significance of the correlations.

Overall, the chosen tools and visualizations support effective storytelling by transforming intricate analyses into visually intuitive representations. This ensures that the project's findings are accessible, engaging, and easily understood for all stakeholders involved in soccer management and data-driven decision-making.

G.3 Recommended Courses of Action

Recommendation 1: Emphasize the development of attributes with strong correlations.

Given the strong correlations identified in the analysis, particularly attributes such as 'Reactions,' 'Composure,' 'Short Passing,' and 'Ball Contol,' I recommend that soccer teams prioritize the development of these specific attributes in player training programs. Focusing on enhancing these attributes could potentially result in a significant improvement in players' overall ratings. This aligns with the organizational need to optimize player performance and team composition.

Recommendation 2: Implement data-driven recruitment strategies.

To further leverage the insights gained from the analysis, I recommend that soccer teams incorporate data-driven recruitment strategies. By considering attributes with high correlations to the overall rating, teams can make more informed decisions when scouting and recruiting players. This aligns with the research question's aim to identify attributes closely correlated with players' overall ratings and translates into practical actions that address the organizational need for effective talent recruitment. Implementing these recommendations can contribute to a more competitive and strategically aligned soccer team.

H. Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=888f073e-05c0-4f1d-93fc-b0e3018b22d3

References

(2019, December 9). Retrieved from The Data School: https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/

StatLect. (2023). Retrieved from https://www.statlect.com/glossary/variance-inflation-factor