CASE STUDY ON FOOTBALL

Table of Contents

1. Introduction

Soccer, or football, stands as the world's most popular sport, captivating the hearts and minds of millions globally. As per Deloitte, the revenue generated by the top 20 leagues in the world was 9.2 billion Pounds, an increase of 14% from last year. In recent years, the landscape of soccer has undergone a significant transformation with the integration of advanced technologies and data analytics. Among these advancements, data visualization has emerged as a critical component in deciphering the complexities of player performance, team strategies, and overall game dynamics. Numerous studies, including those by **Lewis, (2004), Moskowitz & Wertheim, (2011), and Winston, (2012)** provide substantial evidence indicating that effective analysis of sports data contributes to improved team performance. Additionally, research by **Owens & Jankun-Kelly, (n.d.)** highlights the significant positive economic outcomes associated with successful sports data analysis.

This case study delves into the importance of data visualization in soccer performance analysis, shedding light on its profound implications for coaches, players, and sports enthusiasts alike. The significance of this case study on data visualization in soccer cannot be overstated, considering the contemporary sports landscape where teams persistently pursue a competitive edge. In soccer, a sport characterized by its dynamism and complexity, a large quantity of data is generated during each match, incorporating player movements, ball trajectories, possession percentages, and various other metrics. Leveraging this extensive information through data visualization becomes instrumental in empowering stakeholders to make well-informed decisions, optimize strategies, and identify areas for improvement.

In the fast-paced environment of soccer matches, where numerous variables come into play, the paramount importance of data visualization lies in its capacity to simplify complex datasets. Coaches and analysts, often required to quickly grasp unfolding dynamics and make split-second decisions, benefit immensely from visualizations. Whether in the form of heat maps illustrating player movements or pass network diagrams depicting team coordination, these visual tools provide an intuitive and comprehensive understanding of the game's intricacies. Consequently, they play a pivotal role in strategic planning, enabling teams to adapt swiftly to changing scenarios during matches.

2. Challenge and Research Questions

What factors should clubs prioritize during player selection process to achieve a balance between on-field excellence and financial stability, enabling long term sustainability.

Research Questions

1. To what extent does the preferred foot in football correlate with the on-field performance and the financial success of the players?

2. We are seeking to acquire a talented young attacker for our club, whose contract is set to expire in 2024. Our available budget is limited for this potential acquisition.

3.Does the amount of investment in football clubs, specifically in player acquisitions, correlate with on-field success?

In the proposed solution and justification section, each identified challenge will be addressed with data-driven insights to provide practical solutions.

3. Literature Review

The utilization of big data processing and visualization is increasingly beneficial for various stakeholders, including users, consumers, manufacturers, and managers. There is a growing emphasis on leveraging visual information for decision-making and investments, especially in today's dynamic environment. Tables and graphs have become essential tools in this context. The landscape of data visualization tools is expanding rapidly, continually enhancing their analytical capabilities, particularly in big data. Recognizing that the human brain is predominantly responsive to visual representations, colors, and patterns, with approximately 90% of information processed visually, it becomes evident that data visualization is a highly sought-after technique for business and everyday visual data analysis needs across various domains (**Skender & Manevska, n.d.).**

The best data visualization tools share common characteristics, including ease of use, as complexity can be a limiting factor. User-friendly interfaces with excellent documentation and tutorials contribute to a tool's effectiveness. A crucial feature is the capability to handle large datasets, and the top-tier tools can manage multiple datasets within a single visualization. Versatility is another key aspect, with the ability to output various chart, graph, and map types. While many tools offer both static images and interactive graphs, some specialize in excelling at specific types of visualizations. Additionally, cost considerations play a role, where a higher price tag should be justified by superior support, features, and overall value. Ultimately, a tool's inclusion in the "best" category considers its usability, data-handling capacity, output variety, and cost-effectiveness (***A Complete Overview of the Best Data Visualization Tools | Toptal®*, n.d.**).

In the literature, we will investigate prevalent data visualization platforms in sports, including Tableau, Power BI, Sisense, and Domo. Furthermore, we'll examine domain-specific tools like Soccerscoop and MatchPad, crafted to address the distinct requirements of sports analytics. This exploration aims to provide insights into the diverse landscape of available options in the literature related to sports data visualization.

Tableau

Tableau, a versatile data visualization tool available in desktop, server, and free public versions, supports diverse data imports and offers multiple chart formats and mapping capabilities through a user-friendly interface. The free public version, while lacking privacy for data analyses, provides accessibility. However, non-free versions come at a relatively high cost. Users benefit from extensive community support, tutorials, and resources. As an illustration, Chris Love demonstrated Tableau's effectiveness by visualizing shots on goal in a Premier League championship day. The visualization featured a bar chart with time on the x-axis, representing 5-minute intervals, and the y-axis representing the length of each bar, encoding shots on goal. This dynamic visualization effectively conveyed crucial events within each game in the context of the game timeline (**Perin et al., 2018)**. Despite being a comprehensive business intelligence platform with data connection, diverse visualizations, and dashboard creation, Tableau has challenges with large datasets for users without a development background. Moreover, the cost and complexity, particularly for advanced features, must be carefully considered in relation to its versatility.

Power BI

Power BI, a Microsoft-developed analytical application, seamlessly integrates with Microsoft products and offers a user-friendly interface for data visualization and modeling. In addition to supporting various visualization needs, Power BI has cloud integration, providing data warehouse capabilities such as data preparation, discovery, and interactive dashboards. Microsoft extends its support with Embedded Power BI on the Azure cloud platform (**Widjaja & Mauritsius, n.d.**). This analytical tool excels in combining different databases, files, and web services for efficient data manipulation and automatic issue resolution. Power BI ensures secure report publishing within a company and automatically regulates data with updated information (**Krishnan, 2017)**. Moreover, it can integrate all company data, whether cloud or on-premises, utilizing a gateway for connections to SQL Server databases, Analysis Services models, and various other data sources on the dashboard. **Plakias et al., (2023)** explored Power BI's role in football performance analysis, utilizing diverse visualizations from surveys in the "Journal of Physical Education and Sport" to enhance statistical analysis for soccer coaches. Adding to this **Rajesh et al., (2020)** presented a data visualization approach for optimizing player selection in the FIFA league, utilizing statistical analysis, Power BI, and Python Pandas. The study aims to streamline football club decision-making by considering player skills, performance, and cost constraints. Emphasizing risk reduction, features like market value, popularity, and player quality are incorporated. Compared to Plakias et al., both studies use Power BI for sports analytics, but Rajesh et al. specifically focus on player selection optimization within FIFA, while Plakias et al. concentrate on football performance analysis with diverse visualizations for coaching insights. Both studies highlight Power BI’s versatility in sports analytics across different football domains.

Despite its versatility, Power BI requires a learning curve and expertise in Power Query and DAX. Considerations include integration with the Microsoft ecosystem, feature-rich dashboards, and proficiency in Power Query and DAX.

Sisense

Sisense, a business intelligence tool utilized by notable companies like eBay and NASA, empowers non-technical users to analyze large datasets swiftly (**Miller & Lekar, 2014)**. The platform divides user data into multiple dashboards based on Elasticubes, each comprising user-made charts (widgets) and optional data

filters. While lacking the freedom of widget placement compared to some competitors, Sisense excels in processing substantial datasets through its tiling window manager-like system. It boasts a strong backend engine, a user-friendly drag-and-drop interface, and integrated analytics tools. However, advanced customization is limited, and enterprise features come with a higher cost. Known for its embedded analytical capabilities, Sisense facilitates data integration, modeling, visualization, and ad-hoc analysis. Despite its effectiveness, reviews highlight the need for extensive training in setup and complexity in creating data cubes, potentially requiring a longer learning curve for optimal proficiency.

Domo

Domo emerges as an upcoming data visualization tool comparable to Tableau, with a primary focus on liberating analysts and businesses from Excel and democratizing data information (**Hariharan & Krithivasan, 2016)**. While enriched with beautiful visualizations and an extensive suite of chart features, it emphasizes collaboration on data views. However, it faces limitations such as limited customization, rigid data setup, and high expenses. The platform serves as a solution for analyzing and sharing insights, supporting various personas, including analysts, designers, and developers. Core features include diverse visualizations, customizable dashboards with a low-code environment, AI model management, and drag-and-drop data modeling. Domo's user-friendly interface and real-time data integration make it suitable for users of all skill levels, complemented by a robust mobile app. Despite its simplicity, limitations in customization and pricing, particularly for smaller enterprises, are notable considerations.

Domain-Specific Visualization Platforms

Soccer Scoop

Rusu et al. (2010) addressed the challenging analysis of extensive soccer statistics, emphasizing the need for graphical representation. They introduced "Soccer Scoop," an application featuring two visualizations tailored for soccer team managers. These visualizations enable comparisons between players from different teams, pre-contract evaluations, performance assessments in various positions, practice exercise generation, and identification of home and away game disparities. Utilizing information visualization techniques like glyphs, modified star plots, and gestalt principles, Soccer Scoop facilitates quick-glance observations for coaches. However, the application exhibited certain drawbacks, including a less intuitive user interface, the absence of robust data import functionalities, and limited customization options.

MatchPad

Following the work of Rusu et al, **Legg et al., (2012)** mentioned an innovative sports performance analysis tool called “MatchPad”, designed to address the limitations of traditional soccer and rugby performance analysis methods. This MatchPad employs a glyph-based visualization strategy, where visual objects known as glyphs represent the datasets. The primary goal is to enable coaching staff and analysts to gain quick, detailed insights into actions and events during matches. The paper recommends the adoption of

metaphoric glyphs to reduce the necessity for users to grasp intricate coding systems linked with visual elements. This platform was successfully implemented during the Rugby World Cup 2011 by the Welsh Rugby Union, providing coaching staff with a tool to examine actions and events in detail while maintaining a comprehensive overview of the match. Despite the advantages highlighted in the paper, it's important to note the potential disadvantages of glyph-based visualization methods, such as limited detail in individual glyphs, a learning curve for interpretation, subjectivity in glyph design, and the potential for information overload, depending on the complexity and quantity of data. These considerations underscore the need for a careful balance in choosing visualization methods based on the specific demands of sports performance analysis.

Platform	Core Features	Pros	Cons
Tableau	# Data connection and import # Data visualization options Dashboard creation # Data blending	# Simple drag-and-drop interface # Large community support # Wide range of customization choices	# Costly, especially for enterprise-level features # Advanced capabilities with a steeper learning curve
Power BI	# Interactive dashboards and reports # Visualizations # Change analysis # Data modeling	# Integration with Microsoft ecosystem # Feature-rich dashboards and reports	# Comparatively few chart types # Requires Power Query and DAX knowledge for advanced functionality
Sisense	# Data integration # Data modeling # Data visualization # Ad-hoc analysis	# Strong backend engine # User-friendly drag-and-drop interface # Built-in analytics features	# Limited advanced customization # Enterprise features are expensive
Domo	# Visualizations # Customizable dashboards # AI model management # Data modeling	# Simple UI # Real-time data integration # Strong mobile app	# Limited customization # Higher pricing for smaller enterprise
DOMAIN SPECIFIC PLATFORMS			
SoccerScoop	# Intuitive representation of player skills. # Effective for individual and team analysis. # Offers insights for skill development	# Dynamic visualization for quick evaluation. # Provides comparison between players. # Quick display of best skills and areas for improvement.	# Possible challenges for users unfamiliar with soccer. # May require additional features for enhanced control.
MatchPad	# Interactive glyph-based visualization for real-time sports performance analysis. # In-match focus, half-time discussions, and video access. # Scale-adaptive layout for different detail levels. # Integration of visualization with video and performance indicators.	# Provides a clear overview in an easy-to-interpret way # Helps maintain focus during the detailed analysis # Great for oversight in intense matches # Portable and accessible on iPad for immediate access	# Limited Detail: Glyphs may lack intricate details. # Learning Curve: Users need time to interpret glyph symbolism. # Subjective Design: Designing glyphs involves subjective decisions. # Information Overload Risk: Complexity may lead to information overload.

Figure 1: Pros & cons of Visualisation Platforms

4.Data Selection & Preprocessing

We have selected 2 datasets, first one is from Kaggle which has all the international players along with their stats and other features, while the second dataset has been taken from footballwebpages.co.uk and the data has been cross checked with official site of English Premier League. In the first dataset we have 18539 rows and 88 columns. For cleaning and preprocessing, we have used Python. There were few null/missing values, but they were mainly in the columns that were not in use like the national jersey number or club number, therefore we removed the columns itself. The second dataset has the ranking of

the English Premier League teams for 2023 and they had 0 null values. We removed columns like home/away team win and losses.

5.Exploratory Data Analysis

In this section, we have shown a general Exploratory Data Analysis so we can understand the Data before moving further with our research questions.

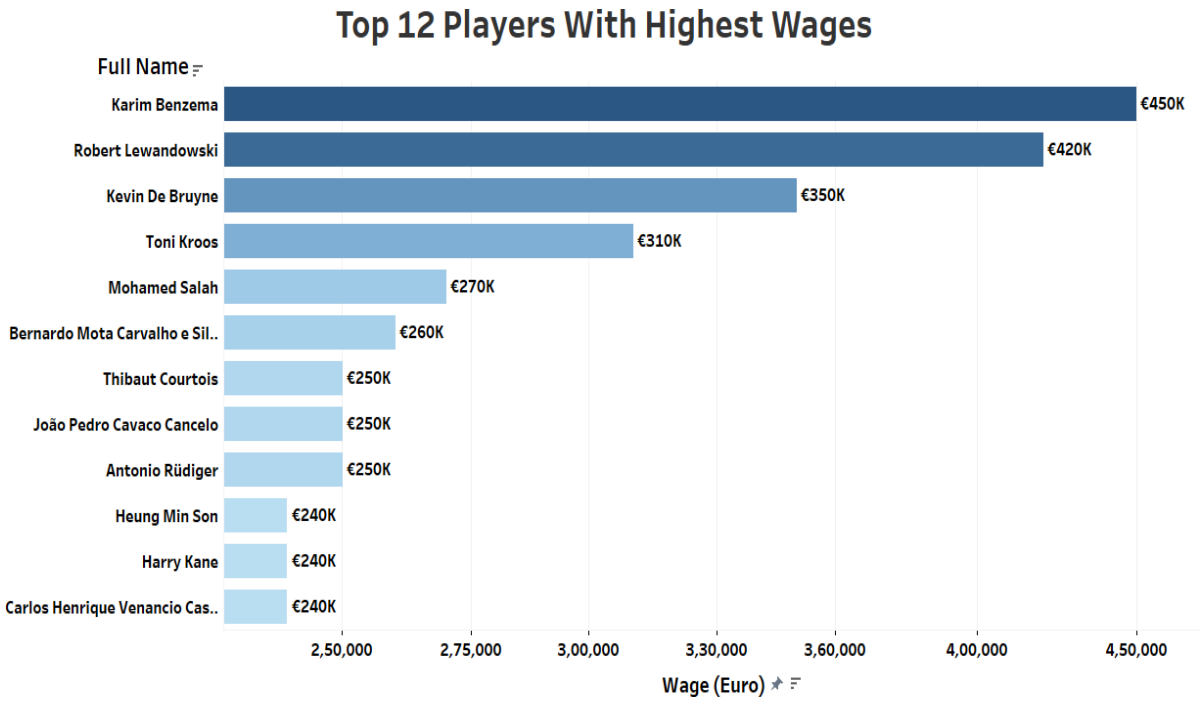

Figure 2: Horizontal Bar-Graph of Top 12 Players with Highest Wages

This horizontal bar graph illustrates the top 12 players with the highest wages in 2023. Karim Benzema holds the highest position, earning a wage of €450K.

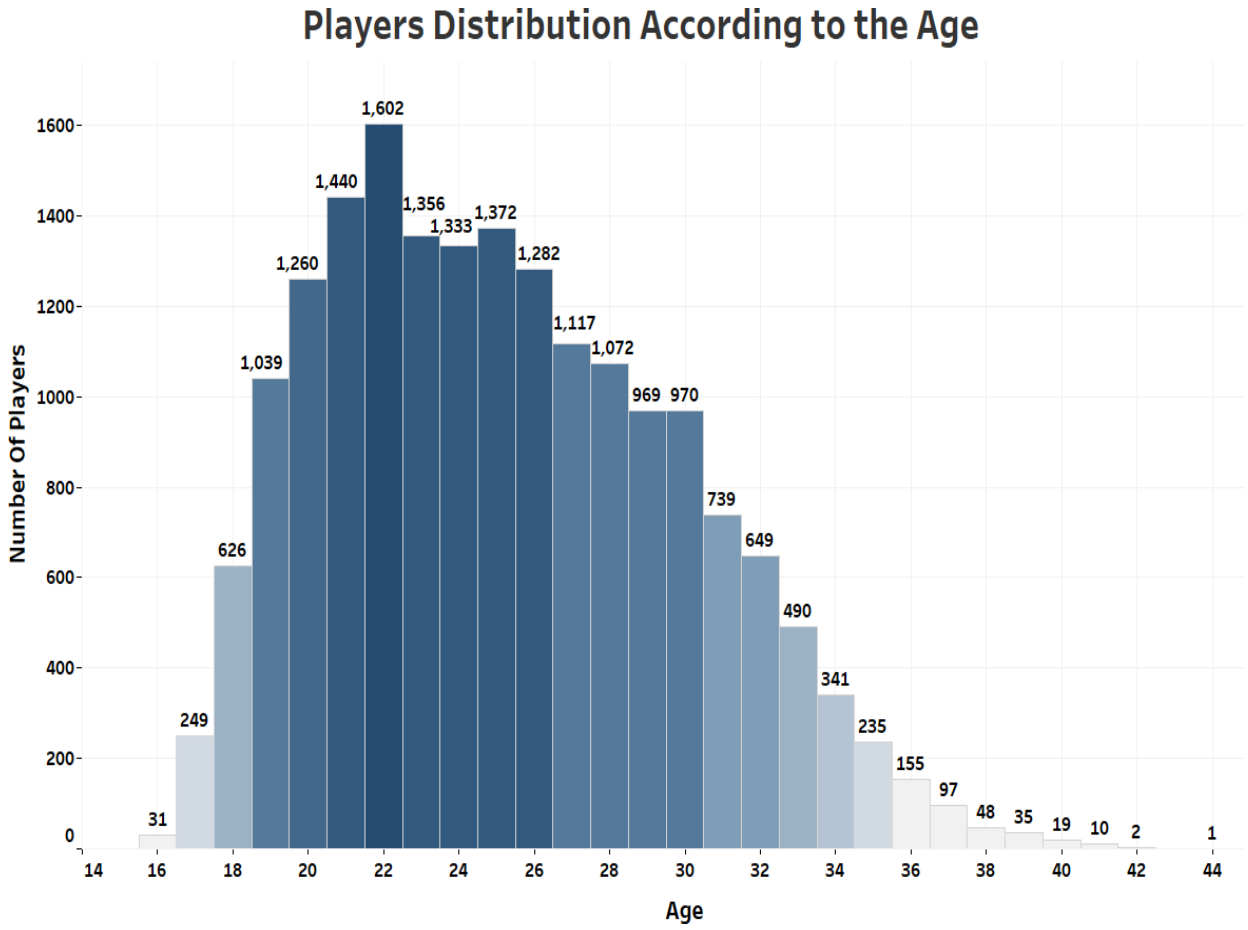

Figure 3: Histogram of Player Distribution According to the Age

The density plot histogram above depicts the distribution of players across the age range from 14 to 44. As indicated by the graph, most players are clustered within the age range of 20 to 26.

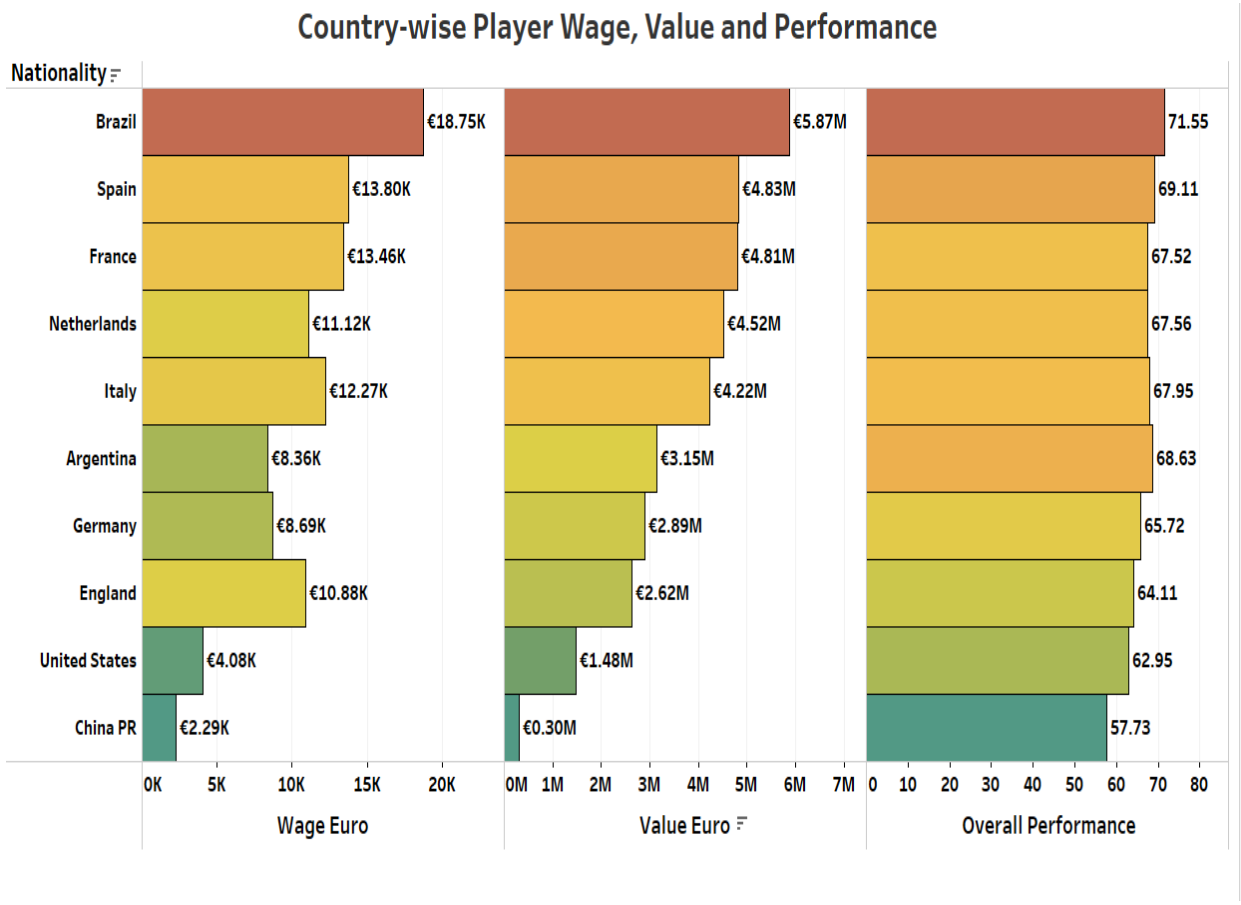

Figure 4:Parallel Bar-Graph Comparison

The above horizontal bar graphs facilitate a comparison of the top 10 countries based on performance in descending order, considering the total value, wage, and overall performance of their players. The graph reveals that Brazil invests the most in its players and displays the highest number of players with outstanding overall performance.

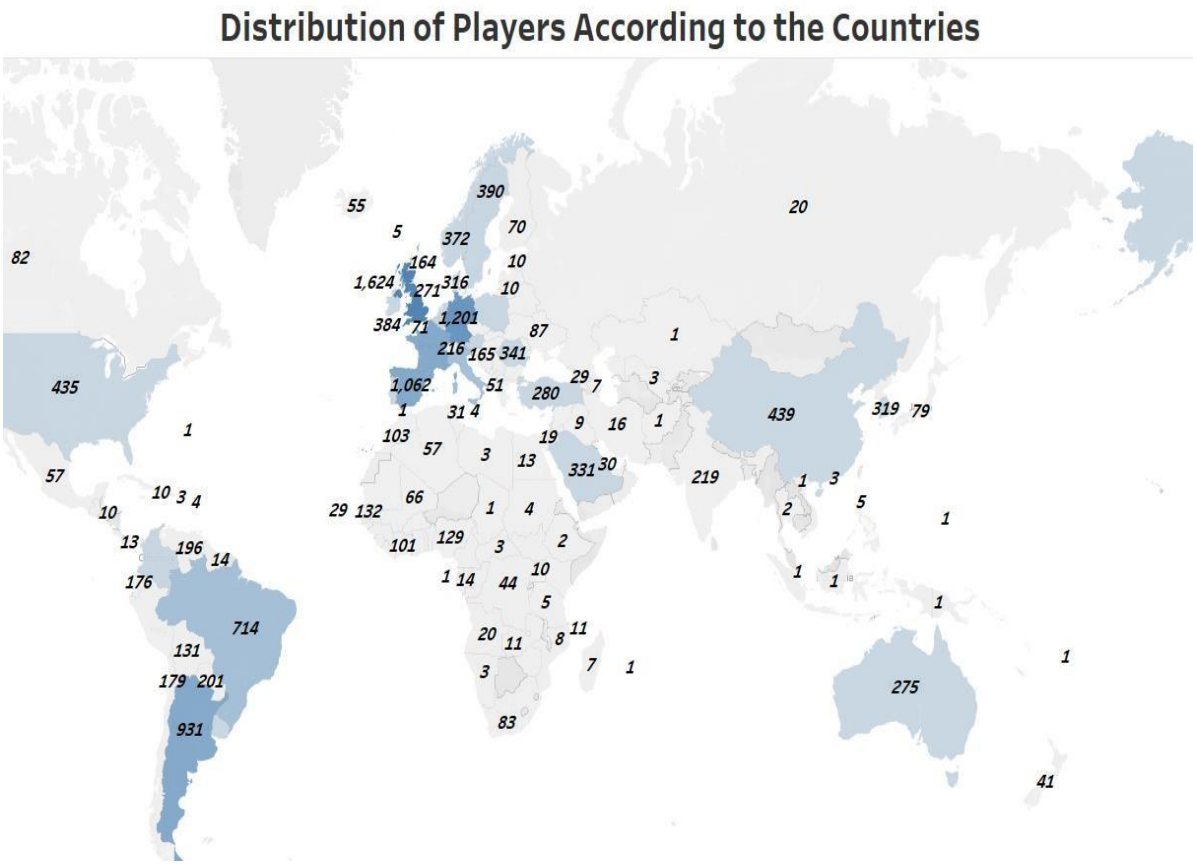

Figure 5: Density map of Player Distribution

The map above displays the global distribution of players. The darker shades of blue represent higher player concentrations, ascending towards lighter shades. Notably, England, Germany, Spain, and Argentina have the highest number of players, respectively.

6.Proposed Solution & Justifications

Question 1:

The first problem that we tried to solve using data visualization for our dataset is whether there is any advantage of having a particular strong foot in football. For this we did analysis of various physical traits vs age for both the feet. In one of the papers, it was stated that being a left footed player is advantageous in football and other sports like basketball or table tennis (Petro.B & Szabo.A, 2016). In one more article, researchers tried to find whether there was an advantage of having a particular foot but since the data was limited, there was no significant relation whether the performance is affected by preferred foot (Bozkurt. S., & Küçük. V, 2018).

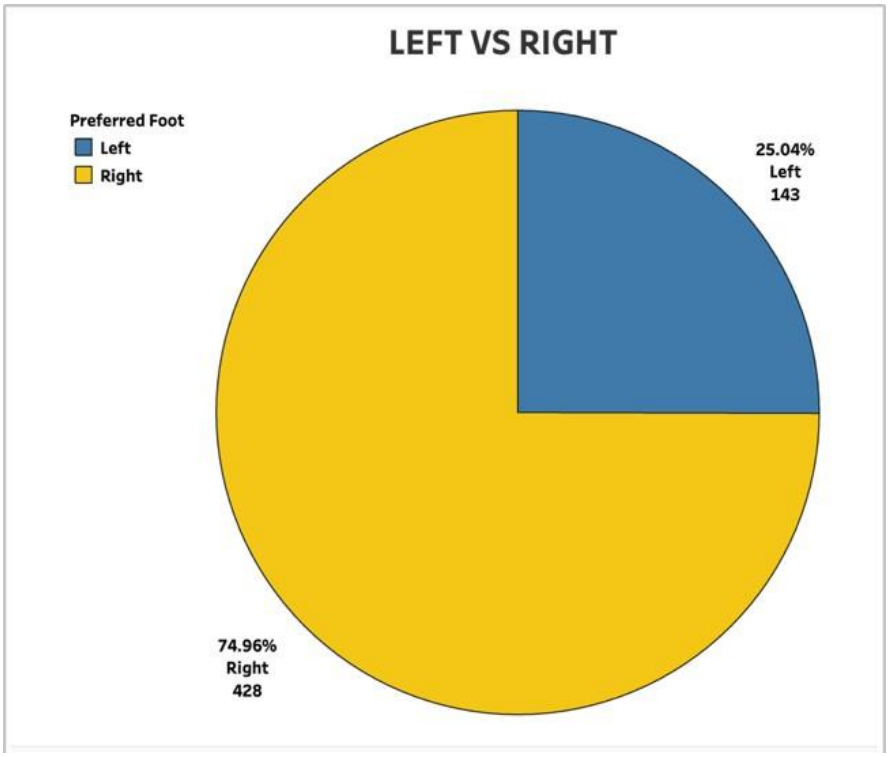

Figure 6: Pie Chart of Preferred Foot

A pie chart is an ideal choice for depicting the distribution of preferred feet in football, offering a clear visual representation of the proportion of right-footed and left-footed players in the dataset. Its simplicity allows for easy comparison and understanding of the overall composition, making it an effective tool to convey the prevalence of each preference.

In the visualization given below, we have created a line graph where Age is on the x-axis and physical traits like Acceleration, Ball Control etc. features on the y-axis. From the Data we had and the visualization we created left left-footed players have slightly better performance as compared to right. Also, the former had better financial success in their career when compared with their counterpart. However, this performance reduces slightly when the players are above 35 years and right-footed players outperform after this range and earn more after this age. It might also be because the number of right-footed players is way more than left-footed. It was highly recommended to use a line graph to solve this problem as we wanted to compare left and right players throughout age and the best way to represent the details of a continuous function was a line graph. Using a Bar graph could have been another option, but then the visualization would not give such clear insights and the same was with the pie chart.

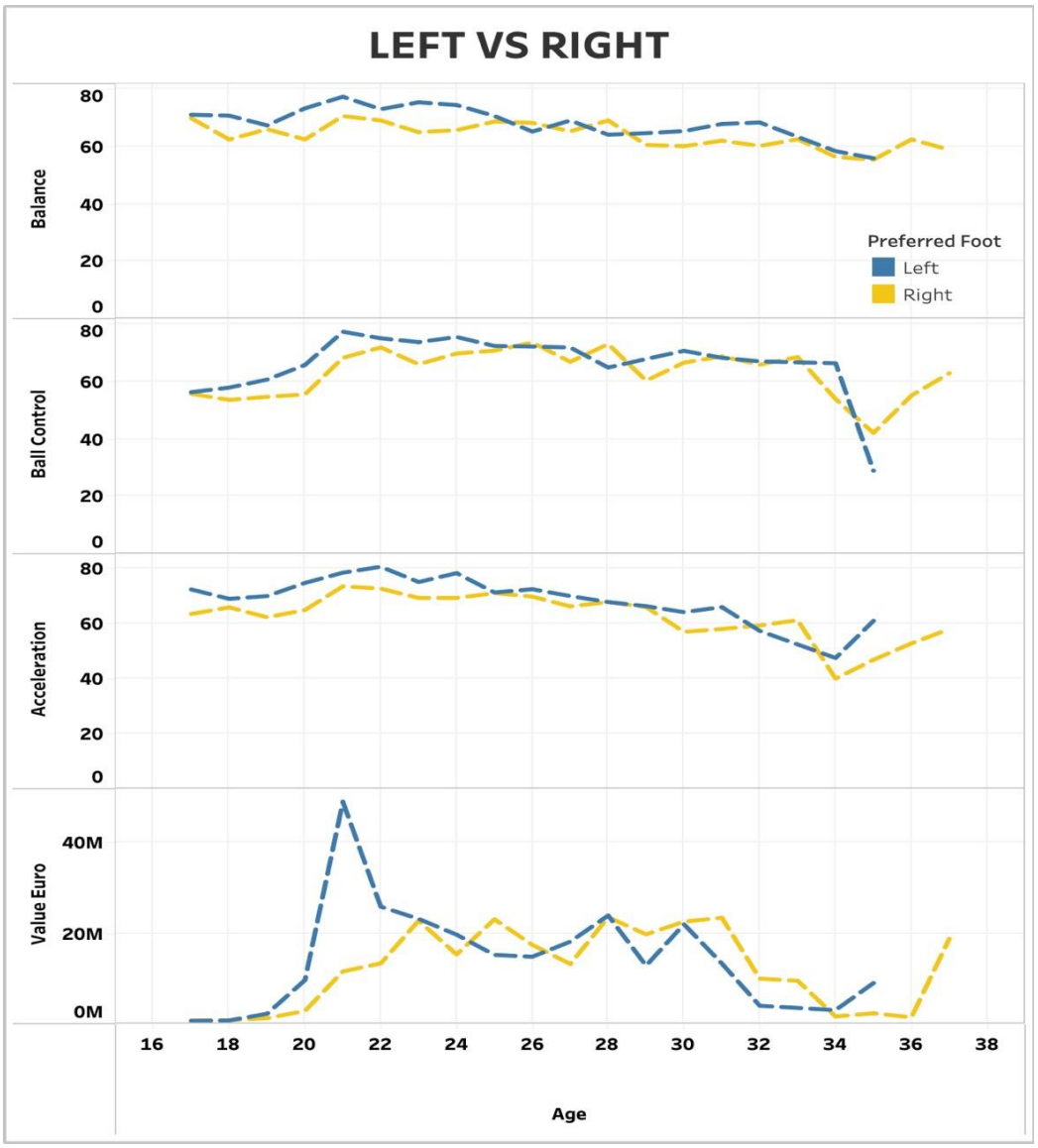

Figure 7: Line Graph Comparison based on Different Factor

Question 2:

In the 2nd case study problem, our football club faces the challenge of strategically acquiring a talented young attacker within our limited budget. Using Tableau, we have developed a comprehensive visualization focused on filtering potential candidates based on key performance indicators. The filters include Shooting Total (must be 70 or more), Dribbling (must be 65 or more), Acceleration (must be 70 or more), and Shot Power (must be 60 or more). Additionally, we consider practical constraints such as the contract expiry by 2024, age 22 or younger, and specific player positions (ST, LW, RW, CAM). This visualization aims to streamline the scouting process, providing a visual representation of eligible players who meet the outlined criteria. Through this approach, we can make informed decisions to acquire a promising young attacker while optimizing our budgetary resources effectively. In the given case study,

the use of a scatter plot in the Tableau visualization proves to be particularly effective for several reasons. The scatter plot allows us to represent each player as a distinct point on the graph, where the x-axis and y-axis can be utilized to showcase different key performance indicators such as shooting total and dribbling. This graphical representation not only facilitates the identification of players meeting the specified criteria but also enables a quick assessment of their relative strengths and weaknesses in these performance metrics. The incorporation of color and size markers further enhances the scatter plot's utility, with the red color legend denoting release clauses and marker size indicating player wages. The scatter plot's ability to present multidimensional data in a visually interpretable manner makes it an ideal choice for streamlining the scouting process, allowing the football club to make informed decisions on acquiring a talented young attacker within budget constraints.

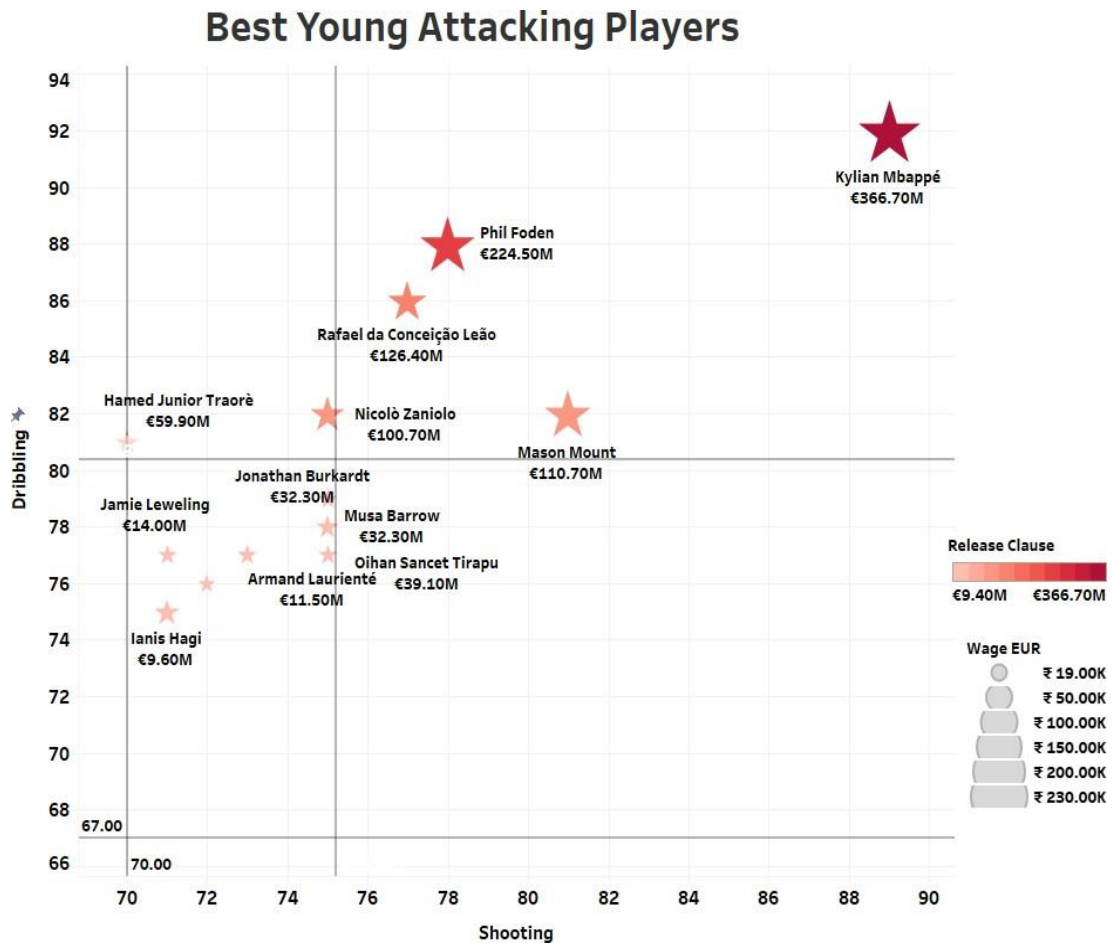

Figure 8: Scatter Plot of Best Young Attacking Players

Based on the visualization depicted above, we can observe players who align with our specified criteria. The players' release clauses are indicated by the red color legend, while the size of the markers represents the wage of each player. Notably, all players featured in the visualization exhibit high skill levels. Given our constrained budget, we have identified promising options such as Jonathan Burkardt, Musa Barrow,

and Oihan Sancet Tirapu, as their release clauses fall within the 30-40 million range, and their associated wages are also relatively modest.

Question 3:

In our 3rd problem, we tried to analyse and see if the top clubs are spending more on their players when compared to the performance they are delivering on the field.

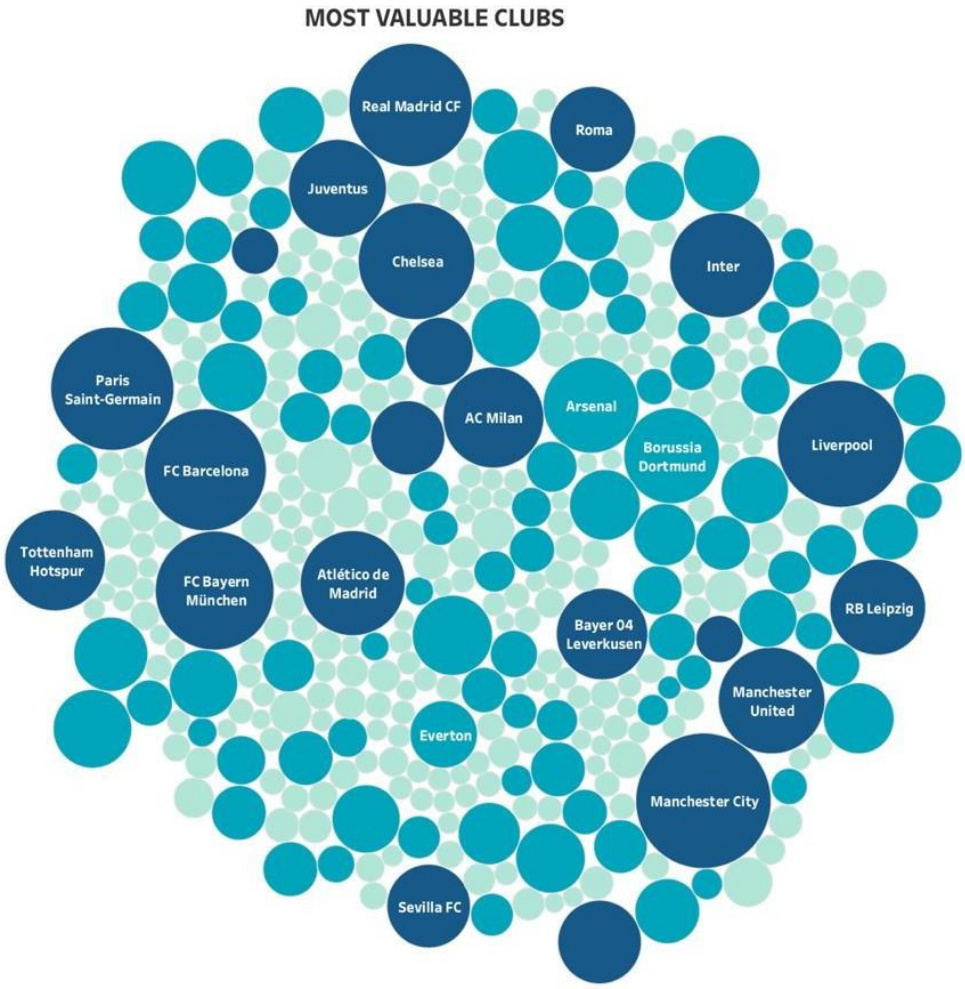

Figure 9: Bubble Plot of Most Valuable Clubs

The Bubble graphs represent the clubs which are spending the most on their players. The size of the circles represents how much money they are spending while the color represents the average performance of their players, where a darker color represents higher performance.

To get a clear view, we have created a comprehensive graphical display to show this, the first and second data set have been used (by doing an inner join) for this visualization. The size of the star represents the total earning of its players while the hue represents the win rate as shown in the legend. Since we are dealing with multidimensional data, scatter plot was the best option for visualizing the data in the easiest way possible.

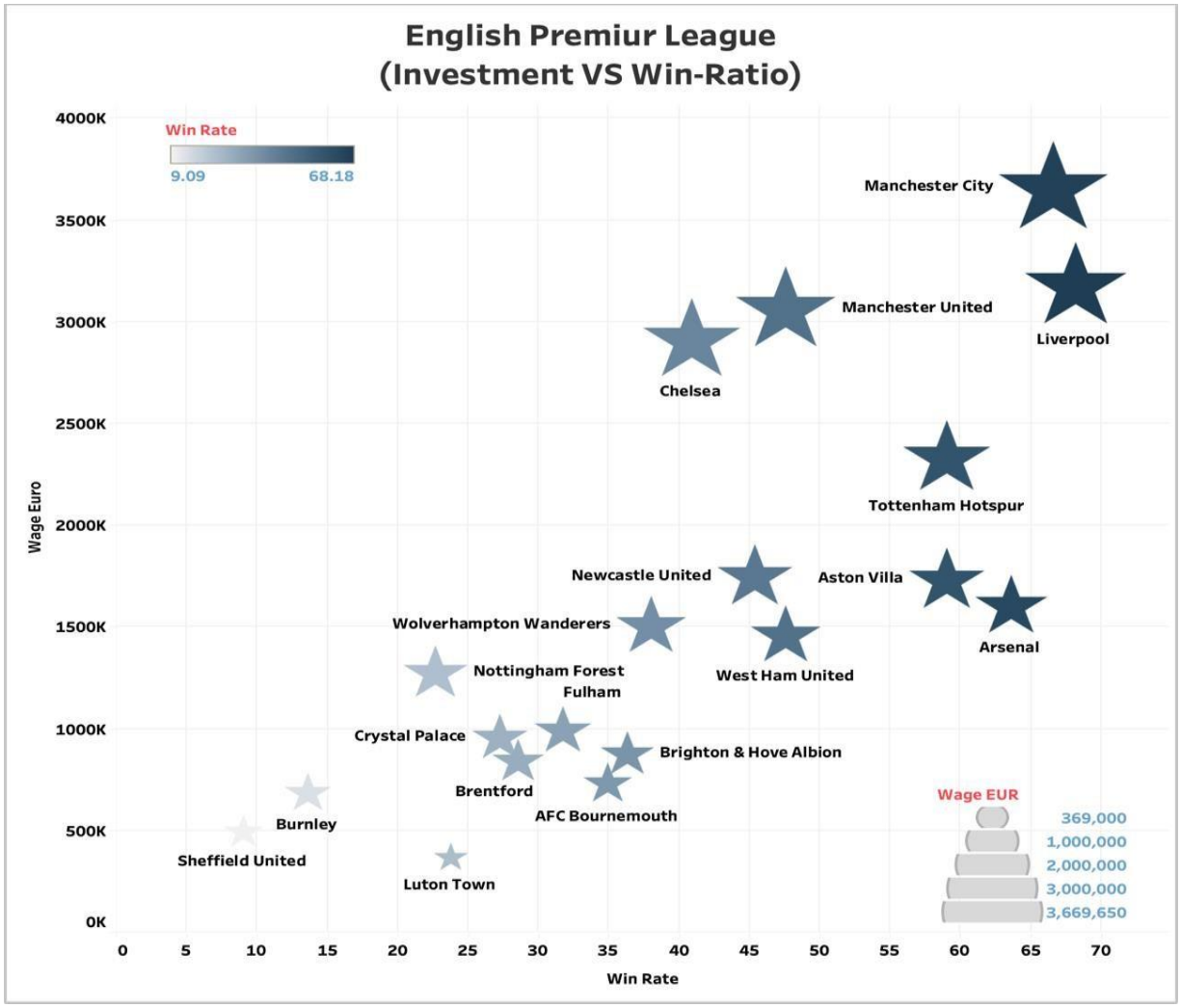

Figure 10: Scatter Plot of Investment VS Win-Rate

Manchester City is spending the most among the 20 clubs, but their WIN ratio is lower than Liverpool. Hence Manchester City needs to manage the funds more efficiently. Similarly, Arsenal has a better win ratio compared to Tottenham Hotspur or Aston Villa at a lower value for the players.

The horizontal bar chart we added below gives a clear side-by-side comparison of selected football clubs. It shows how much they're spending on player wages and their win rates. This simple chart makes it easy to see which clubs are managing their money well for on-field success. It adds a straightforward perspective to the overall analysis, complementing the scatter plot insights.

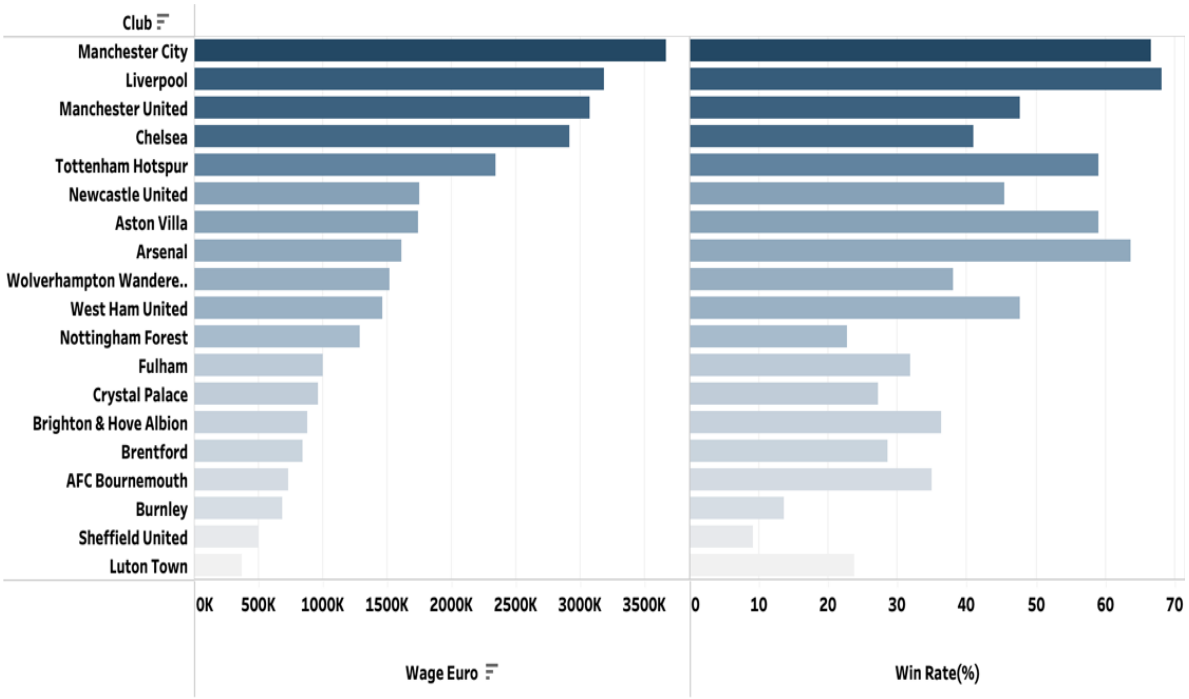

Figure 11: Horizontal Bar Graph of Most Valuable Club & Win Rate%

7.Conclusion

From the above visualization and discussions, we have solved the challenge of selecting players so as to maintain long-term stability with controlled expenses on players. We investigated whether player performance is affected by their preferred foot and how to select top players with limited budget. We used Pie Charts, Scatter Plot, Bubble Plot and Bar graphs to achieve the results. We have used Python for the Data Preprocessing stage and Tableau for the Visualization stage. The proposed solutions and justifications will help in grasping data driven insights for providing practical knowledge. This will help the clubs in navigating through complexities like player recruitment and financial management.

8.References

- A Complete Overview of the Best Data Visualization Tools | Toptal®. (n.d.). Toptal Design Blog. Retrieved 30 January 2024, from https://www.toptal.com/designers/data-visualization/data-visualization-tools
- Bozkurt, S., & Küçük, V. (2018). Comparing of technical skills of young football players according to preferred foot. International Journal of Human Movement and Sports Sciences, 6, 19-22. https://doi.org/10.13189/saj.2018.060103
- Hariharan, B., & Krithivasan, R. (2016). Data Visualization tools – A case study. 14(9).
- Krishnan, V. (2017). Research data analysis with power bi.
- Lewis, M. (2004). Moneyball: The Art of Winning an Unfair Game. W. W. Norton & Company.
- Miller, A., & Lekar, D. (2014). Evaluation of analysis and visualization tools for performance data.
- Moskowitz, T., & Wertheim, L. J. (2011). Scorecasting: The hidden influences behind how sports are played and games are won. Crown Archetype.
- Owens, S. G., & Jankun-Kelly, T. J. (n.d.). Visualizations for Exploration of American Football Season and Play Data.
- Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J., & Carpendale, S. (2018). State of the Art of Sports Data Visualization. Computer Graphics Forum, 37(3), 663–686. https://doi.org/10.1111/cgf.13447
- Petro, B., & Szabo, A. (2016). The impact of laterality on soccer performance. Strength and Conditioning Journal, 38(5), 66-74. https://doi.org/10.1519/SSC.0000000000000246
- Plakias, S., Betsios, X., & Kalapotharakos, V. (2023). Bridging the gap: Leveraging Power BI to connect data science and soccer coaches. Journal of Physical Education and Sport, 23(10), 2543–2550.
- Rajesh, P., Bharadwaj, Alam, M., & Tahernezhadi, M. (2020). A Data Science Approach to Football Team Player Selection. 2020 IEEE International Conference on Electro Information Technology (EIT), 175–183. https://doi.org/10.1109/EIT48999.2020.9208331
- Skender, F., & Manevska, V. (n.d.). Data Visualization Tools—Preview and Comparison.
- Widjaja, S., & Mauritsius, T. (n.d.). THE DEVELOPMENT OF PERFORMANCE DASHBOARD VISUALIZATION WITH POWER BI AS PLATFORM.
- Winston, W. L. (2012). Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football. In Mathletics. Princeton University Press. https://doi.org/10.1515/9781400842070