(2023), 1, 1–6

Visualizing the World Cup and Premier League's Data Using Tableau and ggplot2

Naaima Ben Kadour, Raja Ben Harraf

Abstract

Technology has been a critical driver of sporting performance improvements for decades. Similarly, it has affected sports organizations in many cases even how sports are performed or consumed.

Big Data is playing a great role in football to generate raw data about players and games that allows club managers to be aware of how their players respond and act on the field. This can't be achieved without the use of Data visualization that translates information into a visual context, such as a map or graph, to make data easier to identify patterns, trends and outliers in large data sets. Data Visualization makes it easier for the human brain to understand and pull insights from the data.

Keywords: Big Data, data visualization, football, Tableau, ggplot2

1. Introduction

With more advanced computer vision techniques, football data collection is in a fast evolution since the data can generate meaningful information that increases the game's competitiveness. The data helps to find potential players that are valuable, so they can avoid players that are potentially flops. Also, it helps develop strategies that make the team so hard to break their defensive lines.

This can't be achieved by collecting the data only, in the following paper we're using the data from the world cup since its first edition to visualize the hidden patterns and outliers mainly in the scored goals. We also used English Premier League data due to the high quality of the available data which we made a comparison between the League's clubs based on the scored goals.

2. Project description

The current paper uses the data from both the world cup and the English premier league to visualize the patterns and outliers from the datasets using data visualization tools such as Tableau and ggplot2. The following describes the contests and data sources used in this project.

2.1 World Cup

The FIFA World Cup is an international association football competition contested by the senior men's national teams of the members of the Fédération Internationale de Football Association (FIFA), the sport's global governing body. The tournament has been held every four years since the inaugural tournament in 1930, except in 1942 and 1946 when it was not held because of the Second World War.

2.2 Premier League

The Premier League is the top tier of England's football pyramid, with 20 teams battling it out for the honour of being crowned English champions. The championship is home to some of the most

famous clubs, players, managers and stadiums in the world of football, the Premier League is the most-watched league on the planet with one billion homes watching the action in 188 countries. The league takes place between August and May and involves the teams playing each other home and away across the season, a total of 380 matches. Three points are awarded for a win, one point for a draw and none for a defeat, with the team with the most points at the end of the season winning the Premier League title.

2.3 Data Source

FBref.com is a website devoted to tracking statistics for football teams and players from around the world. It was created by Sports Reference, the team behind popular stats websites like Baseball Reference and Basketball Reference.

FBref.com was launched in June 2018 with league coverage for six nations: England, France, Spain, Italy, Germany, and the USA. Later, it covers more than 45 countries and 140 competitions. FBref.com is a good source of football analytics as well with the highest level of coverage for over 20 competitions including the big five men's European leagues, the Champions League, the World Cup, Copa Libertadores and top leagues in Brazil, Portugal, the Netherlands, United States and Mexico. They also cover all of the top women's leagues in the United States, Germany, France, England, Italy, Spain and Australia.

3. Discussion

3.1 Premier League clubs' goals for and against during the current season (2022-2023):

In the following graph (figure 1), we wanted to visualize the position of each club in the Premier League according to the goals they scored and against the opposition during the current season of the League. We clearly can conclude that Man City is the team that scored the biggest amount of goals, and received the minimum amount of goals from the beginning of the competition until the end of December 2023 (time of collection of the data). Bournemouth Football Club is the club that received the biggest number of goals and scored fewer goals than all the other clubs in the competition.

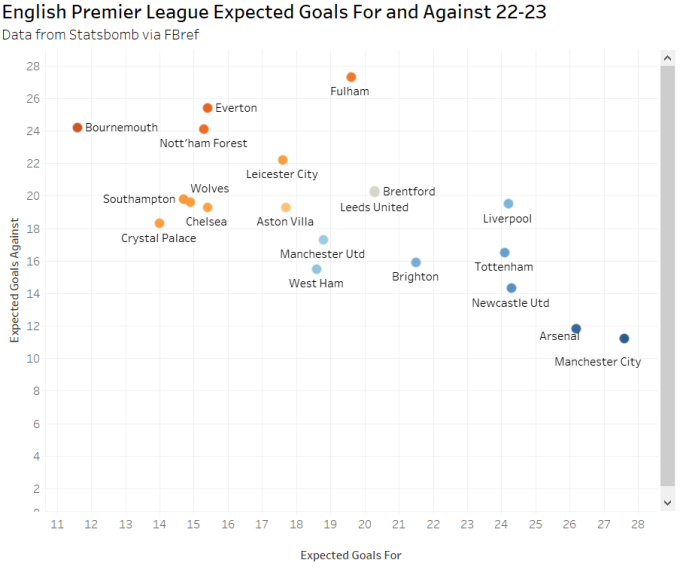

Figure 1. premier league clubs goals for and against

3.2 Building a player's dashboard in the premier league during the season 21-22: Luke Shaw

Luke Shaw is an English professional footballer who plays as a defender for Premier League club Manchester United since 2014. The following dashboard (figure 2) describes the performance of Luke Shaw in the Premier League during the 2021 – 2022 season. The collected data from Football Reference was in the "per90 minutes" metric. This metric helps compare the performance of players taking into consideration the player's minutes played that dramatically change the performance of players.

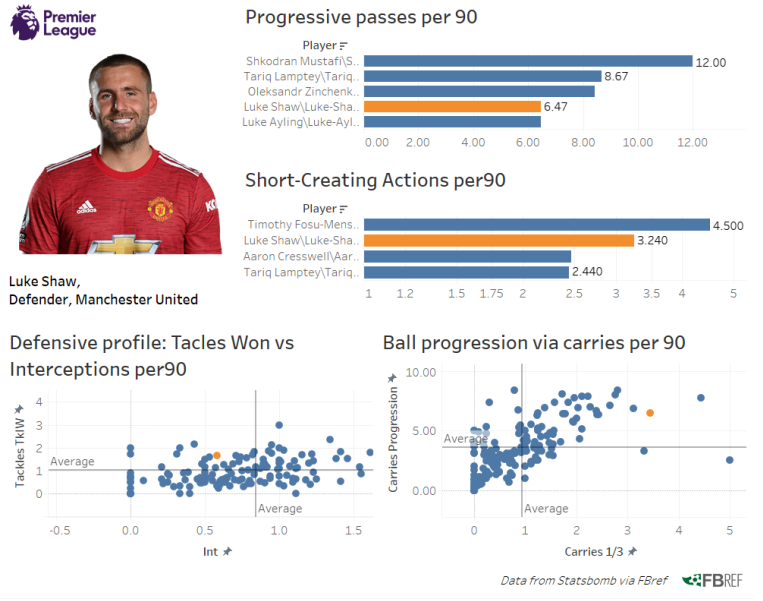

Figure 2. Luke Shaw's Dashboard

3.3 Progressive passes graph:

Progressive passes in the dataset are the completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. In the following graph, we have how Shaw fares in terms of passing and offensive production by using bar graphs. To produce it, we used the filter for defenders who have played at least 10 of 90s. We also filtered to show the top 5 defenders for progressive passes per90, Shaw ranks the fourth best in the league in this metric.

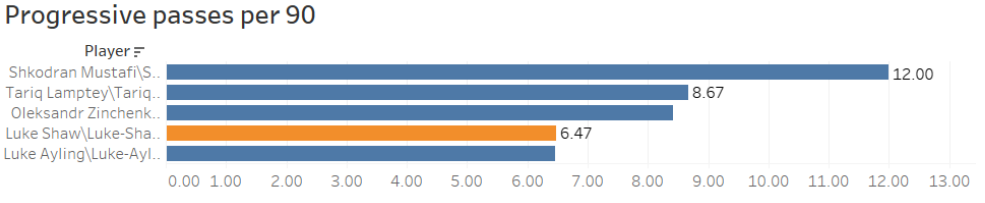

Figure 3. Progressive passes bar chart

3.4 Shot creating actions per 90 minutes:

The following bar graph shows that Shaw stands out for shot-creating actions per90 amongst defenders in the league with 3.24 shots created per90.

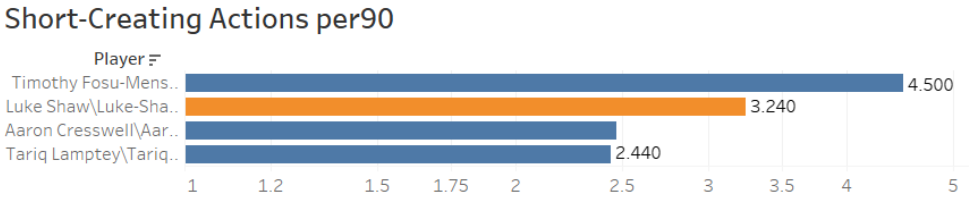

Figure 4. Progressive passes bar chart

3.5 Ball progression via carries per90 and defensive profile:

Using player passing data, we created the following scatter plot for visualization of progressive carries using completed passes that enter the 1/3 of the pitch closer to the goal and progressive passes that represent the completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes or any completed passes in the penalty area.

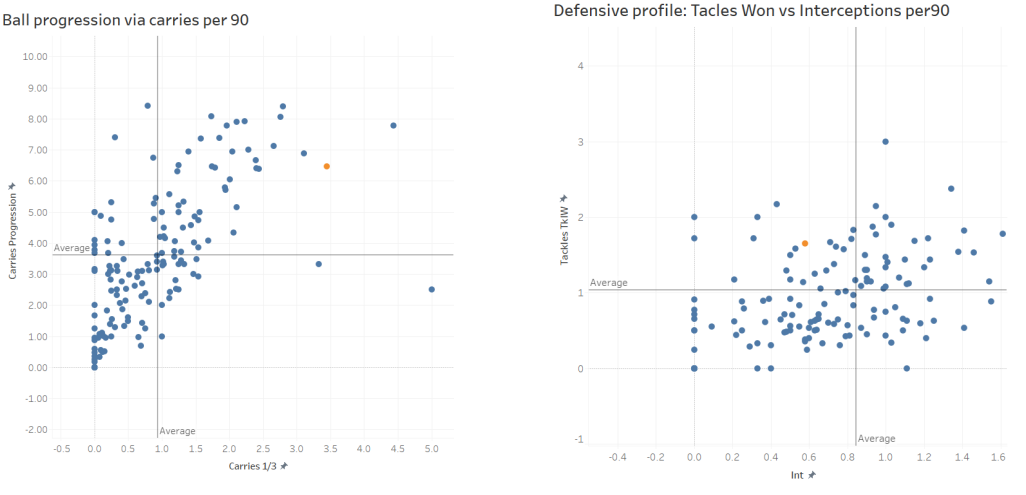

Figure 5. Luke Shaw's defensive profile

3.6 Goals of the world cup based on the scoring minute

Using a scatter plot grid, we've visualized World Cup goals from 1930 to 2018 and spread them over a 120-minute timeline. This model is a particularly good choice in this case, as it allows us to clearly see the fastest and slowest goals in the tournament. In addition, we can view the type of these goals, whether the player scored a penalty or not.

We can clearly conclude that the biggest range of goals is scored during the second half of the world cup games, we can also say that the extra time has few goals.

Figure 6. The scored goals of the world cup games from 1930 to 2018

3.7 Percentage of penalties during the world cup games (1930-2022)

For figure 7, we used ggplot2 to calculate the percentage of penalty goals scored during the games of the world cup using a pie chart. The non-penalty goals have the biggest share in the games with 39 per cent when the penalties are only 7 per cent.

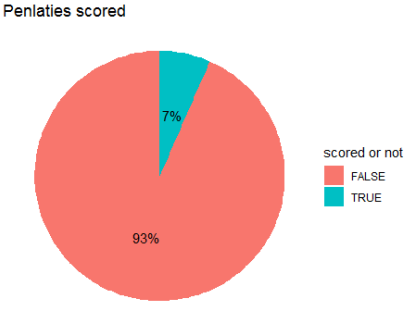

Figure 7. The scored goals of the world cup games from 1930 to 2018

4. Evaluation

4.1 Tufte principles

Tufte's work provided a major breakthrough in the field of visualization and changed the illustrator's perception of graphics. He cites numerous examples of substandard charts throughout his book "The Visual Display of Quantitative Information". He also lays out the principles on how to effectively narrate, investigate, and summarize data using graphical design. To evaluate our graphics, we used the following principles to make sure our work is well displayed without any erroneous additions to graphics that lead to graphical distortions and over-decoration.

Graphical Integrity visual representations of the data must tell the truth, therefore, we used clear, detailed and thorough labelling to defeat graphical distortion and ambiguity by writing out explanations of the data on the graph itself in the pie chart for example and the scatter plots that show the exact values on the scatter is selected.

Data-Ink, according to Tufte, is the ink on a graph that represents the data that should be maximized in the graphics to avoid too much noise and distracting elements. In our graphics, we avoided using 3D effects, shadows and other unnecessary aesthetics that distract the data from being represented.

Chartjunk is the excessive and unnecessary use of graphical effects in graphs. Our visualizations were produced to display the data without the use of heavy grids.

Data Density of a graph is the proportion of the total size of the graph that is dedicated to displaying data. In our visualization, we made sure to maximize data density and the size of the data matrix.

5. Conclusion

The paper at hand presents a group of visualizations on the world cup data to understand the scored goals timeline and the impact of the last time in which a huge amount of goals are scored in every tournament of the world cup. The paper shows also a dashboard of Luke Shaw, a defender from Manchester United Football Club, in which we presented his performance in ball progression, and shot-creating actions, as well as the tackles won according to the metric of "per90".

References

[1] 2021-2022 Premier League stats (2021) FBref.com. Available at: https://fbref.com/en/comps/9/2021-2022/2021-2022-Premier-League-Stats.

[2] Becklas, A. (2018) FIFA World Cup, Kaggle. https://www.kaggle.com/datasets/abecklas/fifa-world-cup.

[3] Tufte, E.R. (2018) The visual display of quantitative information. Cheshire, CT: Graphics Press.

[4] Toma, P. and Campobasso, F. (2023) "Using data analytics to capture the strategic and financial decision-making of Europe's top football club," Technological Forecasting and Social Change, 186, p. 122116. Available at: https://doi.org/10.1016/j.techfore.2022.122116.

[5] Premier League Football News, fixtures, scores amp; results (no date) Football News, Fixtures, Scores amp; Results. Available at: https://www.premierleague.com/ (Accessed: January 2, 2023).