Data Acquisition - Final Project Report

# An Analysis of Midfield Performance in the English Premier League

By Sarah Cernugel, Molly McCann, and Taylor Hill

## Abstract:

The purpose of our project was to analyze the midfield performance of Premier League teams. We were also curious if we would be able to conclude that Arsenal was correctly ranked as having the best midfield in the league. To do this, we scraped data from the Premier League's website and conducted a series of hypothesis tests. These t-tests compared Arsenal to a select set of 16 teams, using completed pass data from the 2022-2023 and 2023-2024 seasons. We found Arsenal had more completed passes than many of the other teams, but we were not able to conclude that they had the best midfield overall. To do that, a more in-depth analysis would need to be organized.

## Introduction:

The Premier League is the highest level of soccer in England, containing 20 teams and over 500 elite players. Teams are constantly looking for a competitive edge in order to avoid relegation, and in the modern era of the sport, analytics has become the source of the insight managers seek. As soccer fans, we are interested in using the provided league statistics to answer questions of our own. We will take a look at the midfield skills across the league and compare them through our analysis.

In an article published by OneFootball in 2023, Arsenal was ranked as the best overall midfield in the Premier League with Manchester City trailing behind as the second best (GiveMeSport). We want to test whether this ranking is accurate by using passing data scraped from the Premier League website. We are looking at passing data because the midfield has a strong influence over the flow of games and that is largely done by stringing together passes. It is a common notion that a skilled passing team is in control and would be a good indicator of how elite a midfield is in comparison to others.

The motivation behind this analysis is to further develop the skills necessary to begin a career in the field of sports analytics, as well as better understand the Premier League. Breaking

down the midfield of teams to gain more insight helps us understand why some teams excel in the league more than others. Upon the completion of this analysis, readers will better understand the importance of passing in the midfield and the intricacies of the Premier League.

## Data Acquisition Process:

In order to complete our analysis, we had to scrape the data from the Premier League's website. The scraping process used in acquiring the data required an indirect route through the website's API in order to find the data that was structured in the JSON format. To do this, we had to inspect the page and find the specific JSON that included our relevant information. Next, we added HTTP header information to the code, extracted the content as JSON text, converted that text into a list, and extracted the stats table, converting it to a usable data frame.

The setup of the Premier League website caused some difficulty while extracting the data. The page only shows 10 players' stats at a time so in order to extract all of the midfielders' stats we had to use multiple API URLs and rerun the code each time. Every time we changed the page to get the next ten players, a new JSON URL would appear in the network. We copied the link into the code and repeated this process until all of the players were extracted. Each chunk of code that pulled data from the website was flattened and then put into a list and merged into one dataset. Only the necessary columns in the dataset were selected to make the data easier to work with.

Adding to our analysis, we also collected team data in order to create a visualization plotting passes and wins by team. The same data collection process was used as above, but we applied filters for team data rather than individual players. The first pull of data was wins and the second was passes from the 2023-2024 season. After, the two datasets were merged together with the names of the teams and the teams that had been relegated were removed. We added logos to the plot by creating a Google sheet with URLs of the PNG images and turning it into a CSV file. The team data and CSV containing the logos were merged and used as the dataset for the visualization.

## Data Description:

Once all the data was collected, the data was downsized and only columns of importance were kept. These columns include player ID, player name, number of passes completed by the

player, player position, team, and whether they are an active player. The data for both seasons was then merged into one data frame. Due to players on loan and transfers, there are some midfielders who had null values in the team column. These players had to be removed from the dataset in order for our analysis to work. Again, the teams who were in relegation were removed from the data frame so that we could have accurate comparisons between Arsenal and the other clubs due to the data spanning across two seasons.

Analysis and Results:

In order to determine whether Arsenal's midfield is truly the best in the league, a series of hypothesis tests were conducted. The general null hypothesis for these tests is that the average number of midfield passes for Arsenal is equivalent to the average number of midfield passes for "other." Here, the word "other" can be replaced with the other 16 teams that Arsenal is being compared to. These teams include Manchester City, Aston Villa, Newcastle United, Liverpool, Tottenham Hotspur, Crystal Palace, Brighton & Hove Albion, Chelsea, Fulham, Bournemouth, Wolverhampton Wanderers, Manchester United, Everton, West Ham United, Brentford, and Nottingham Forest. The alternative hypothesis was that Arsenal's average midfield completed passes are greater than "other" average midfield completed passes. After defining the null and alternative hypotheses, a t-test was carried out for each of the 16 teams listed above.

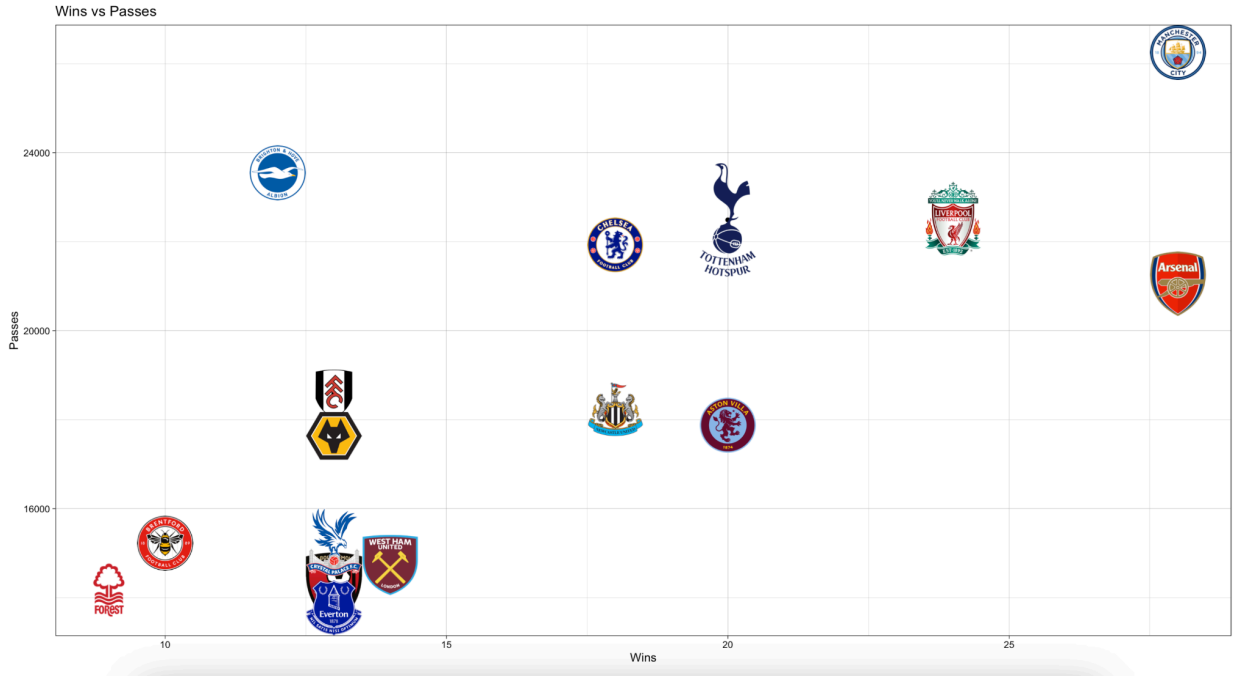
It was found that the t-test results were statistically significant at the 0.05 alpha level for several of the teams. In these cases, we reject the null hypothesis in favor of the alternative hypothesis. This suggests that Arsenal's number of completed passes is greater than that of Aston Villa, Newcastle United, Crystal Palace, Brighton & Hove Albion, Fulham, Bournemouth, Wolverhampton Wanderers, Everton, Brentford, and Nottingham Forest.

As previously mentioned, an indicator of midfield strength can be how well the midfield possesses the ball. These results imply that Arsenal's midfield may be stronger than the teams that had significant p-values. However, we cannot confidently conclude from our analysis that Arsenal's midfield is the "best" in the league during the 2023 year.

Discussion:

Midfield possession is a crucial element in soccer, as it often serves as the engine for both defense and attack. Teams that dominate possession in the midfield are generally able to control

the flow of the game, dictate the tempo, and create more scoring opportunities. With that being said, one limitation of our model is that completed passes in the midfield is the only statistic that we used in the analysis. High pass counts and possession percentages do not necessarily translate into effective or dangerous play. A team that prioritizes a more direct style of play with fewer passes, may still achieve a better standing in the Premier League ranking table due to efficient use of possession. Thus, while we sought to confirm statements about Arsenal's midfield strength and performance, if we look at the game from a holistic viewpoint the best midfield may not produce the best team in the league. Despite this, the analysis we conducted is still important for a team to have in order to provide an understanding of where they stand in comparison to other clubs in the league.




The above visualization shows a plot of passing and win totals for each team in the Premier League. As stated above, high pass counts do not always translate to dominant play. There are teams like Brighton & Hove Albion that have one of the highest pass totals but only had 12 wins in the 2023-2024 season. However, there generally appears to be a trend showing that the more passes a team has, the more wins they achieve on the season. Manchester City is a team that excels in both of those areas.

## Conclusion:

To conclude our findings, our data served its purpose for the analysis we wanted to conduct, which allowed us to observe that Arsenal had more completed passes than 10 of the other teams in the league. However, one implication is that it does not confirm or deny the ranking provided to Arsenal in the article previously mentioned. After analyzing the t-tests, we cannot conclude that Arsenal has the better midfield solely based on the passing stats. It is clear that there are opportunities to evaluate the teams further, building on the analysis we have completed.

## Future Work:

When we first started brainstorming for this project, there were many stats that could have been beneficial to use, but we quickly figured out that some tell a better story than others. Originally, we were going to analyze touches and passes but decided that just because a team has more touches does not necessarily mean that they have a better midfield overall. Analysis should continue to be done on the teams and players in the league, but in a way that provides better insight. It is important to keep the data up to date, constantly collecting data points. It is also important to have a wide variety of stats that are recorded, so you can tell as close to the full story as possible. If we were to do another analysis, we would choose to include additional stats like touches, through balls, aerial battles won, tackles, etc. in order to better understand the influence a midfield unit can have on a team. We believe that the Premier League does a great job of using stats like these to tell a story, but we want to highlight that smaller, local clubs would also benefit from this kind of data collection and analysis.

## References:

GiveMeSport. “Ranking Every Premier League Team’s Midfield from Worst to Best.” *OneFootball*, Yahoo Sports, 21 Aug. 2023, onefootball.com/es/noticias/ranking-every-premier-league-teams-midfield-from-worst-to-best-38076996.

“Premier League Stats Centre.” *Premier League*, www.premierleague.com/stats. Accessed 29 Sept. 2024.