Data Visualisation and Dashboarding Exploratory Data Analysis CW

Research Question

What were the effects of playing Premier League football without fans during the Covid-19 pandemic?

Data Acquisition

Data used in the following analysis was downloaded from the *FootyStats* website. The data can be found using the "English Premier League Teams csv" link on this page:

https://footystats.org/download-stats-csv#

The *FootyStats* company began in 2016 with "one thing in mind – to make sense of football data" (FootyStats.org). They provide insightful football data for public to view and are now the industry leader in football statistics. The data they provide is near to, if not the most reliable in the industry with their databases being updated once a minute and backed up every day to ensure they can deliver the latest statistics as they happen.

Data was downloaded for all seasons from the 2012/13 season to the most recent full season (2020/21). The primary purpose of this dataset is to provide football gamblers with the resources to conduct their own analysis into matches and compare the outcome with what the bookmakers odds were for various markets. In doing so it provides statistics about each football match, from the amount of goals scored by each side to the fans attendance at any given match and thus provides adequate information to analyse the effects of Premier League football teams playing without fans in the stadium. There were, however, a couple of measures that I would've liked to have included in my analysis but could not source the statistics either from *FootyStats* or elsewhere. These measures were; the distance run by each team per game and the average position of the ball through certain periods of the game. This would have facilitated more in depth analysis that could have factored in player fitness/effort with/without fans and possession safety with/without fans.

Data Preparation

The main objective in preparing this dataset for exploratory analysis into the effect of playing football without fans was to remove any columns that were not useful due to their specific reference to betting markets. This included removing the columns:
status, average_goals_per_match_pre_match, btts_percentage_pre_match, over_15_percentage_pre_match, over_25_percentage_pre_match, over_35_percentage_pre_match, over_45_percentage_pre_match, over_15_HT_FHG_percentage_pre_match, over_05_HT_FHG_percentage_pre_match, over_15_2HG_percentage_pre_match, over_05_2HG_percentage_pre_match, average_corners_per_match_pre_match, average_cards_per_match_pre_match, odds_ft_home_team_win, odds_ft_draw, odds_ft_away_team_win, odds_ft_over15, odds_ft_over25, odds_ft_over35, odds_ft_over45, odds_btts_yes odds_btts_no.

Other tasks that were completed on the dataset included: creating a season column so that data could be identified by season, splitting the date and time into separate columns, tidying up the team names to ensure they were consistent throughout the dataset, splitting home and away goal timings into individual values so they could be binned, creating bins for goal timings and adding in columns for counts of total corners, total yellow and red cards, total shots in a match, total shots on and off target in a match, total fouls in a match and stadium capacity. Data for stadium capacity was added in manually using data from google.

Data integrity checks included removing any duplicate values that were found using the remove duplicates tool in Excel and running the data interpreter tool in Tableau.

Exploratory Data Analysis

Introduction

The outbreak of the Covid-19 virus in 2020 saw many unprecedented social restrictions imposed on the UK and its citizens. In March 2020, the UK went into a national lockdown which resulted in the suspension of the 2019/2020 English Premier League football season. The season remained on hold from 13th March to 17th June 2020, during which time players had a staggered return to their usual training routine. When the season resumed after a three month absence, social restrictions were still in place, one of which was the banning of mass

gatherings of people; including sporting events. This meant that for the first time ever, top level English professional football would be played behind closed doors with no fans, leaving the opportunity to explore a greatly theorised phenomenon – home advantage. The following exploratory analysis will look into how the effect of no fans in stadiums affected various different measures in a football match.

Topline Statistics

Match results (Figure 1.)

For the first time in the 10 years of Premier League football matches analysed, the year 2021 saw a greater percentage of away team wins than home wins. During that year, 42% of all games were won by the away side, compared to the 39% recorded by home teams. This immediately suggests a shift towards a reduction in home advantage that has previously been touted. It is something Liverpool manager Jurgen Klopp has been particularly vocal about in the past saying: "[this is] an example to everyone about how supporters can influence a team and influence a game" after an important comeback victory at Anfield in the champions league in 2018 where the players appeared to be spurred on by the atmosphere (This is Anfield, 2020).

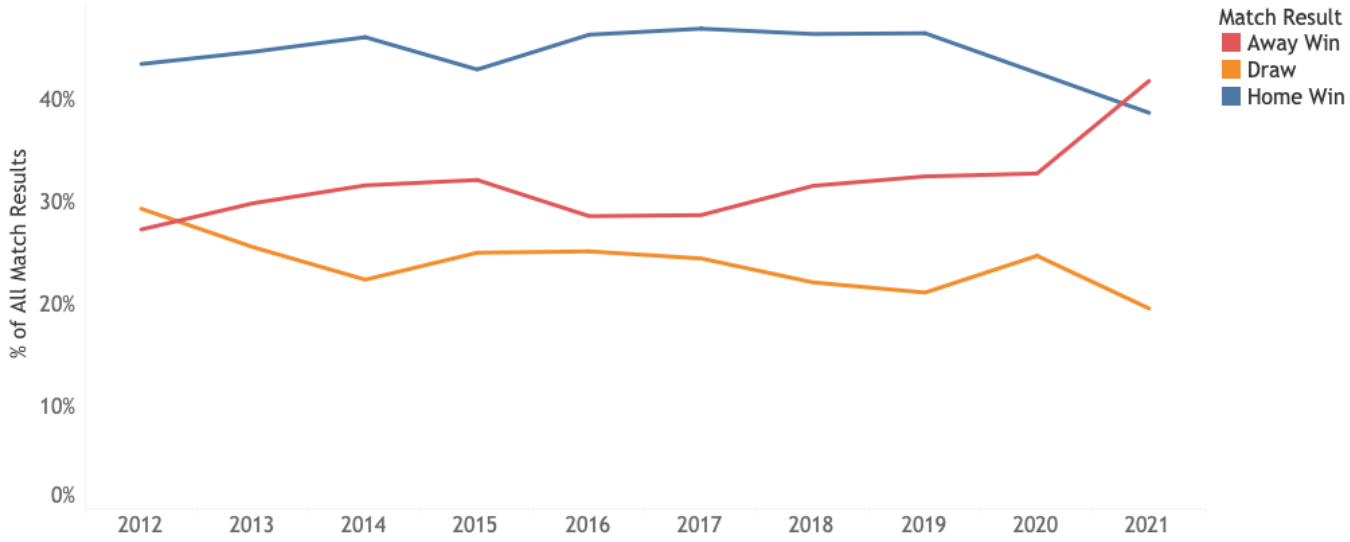

Figure 1. The % of matches won by the home team, away team and that were drawn for each year in the dataset.

Points per game (Figure 2.)

Although home teams still picked up more points on average at home even without fans, the effect is much less marked and away teams fared better. With fans present home teams averaged 1.6 points per game, however, when no crowd was in the stadium that number dropped to 1.4 points per game. Conversely, away teams averaged nearly 1.2 points per game prior to the fans restriction coming into effect and almost 1.4 points when playing without any fans. This again, is an indicator that the lack of fans in stadiums favoured the away side.

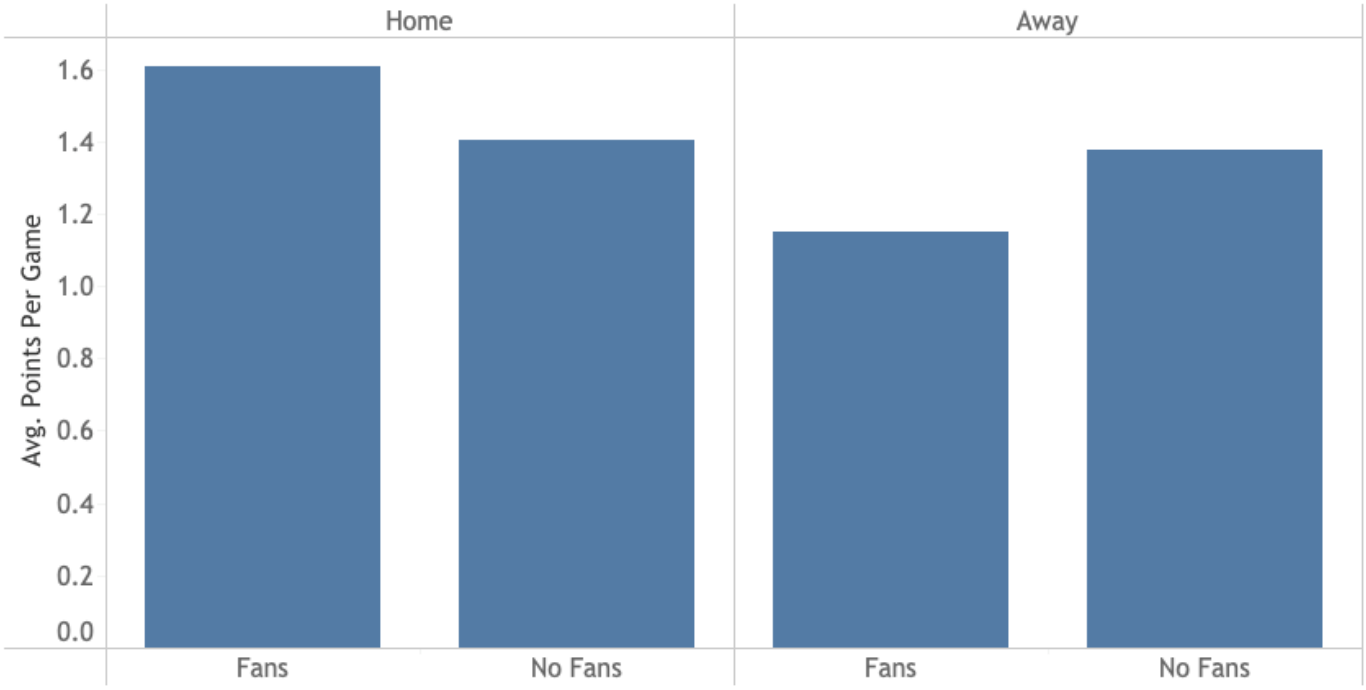

Figure 2. Average points per game for home and away teams, split by fan status.

% Change in points by stadium (Figure 3.)

It is interesting to split this data by individual stadiums to look at the impact no fans had on each club. Different fanbases have different reputations regarding their level of support during a match and so when looked at in this way the amount to which certain fanbases play their part can to some extent be objectively measured.

Out of the 22 stadiums analysed in this dataset that played matches both with and without fans, only 5 stadiums improved the amount of points gained at home without fans present. These were the stadiums of West Ham, Southampton, Tottenham, Watford and Aston Villa.

Leeds, whose home ground is Elland Road, were the biggest losers in this statistic with their home points per game dropping -52% compared to what it was with fans present.

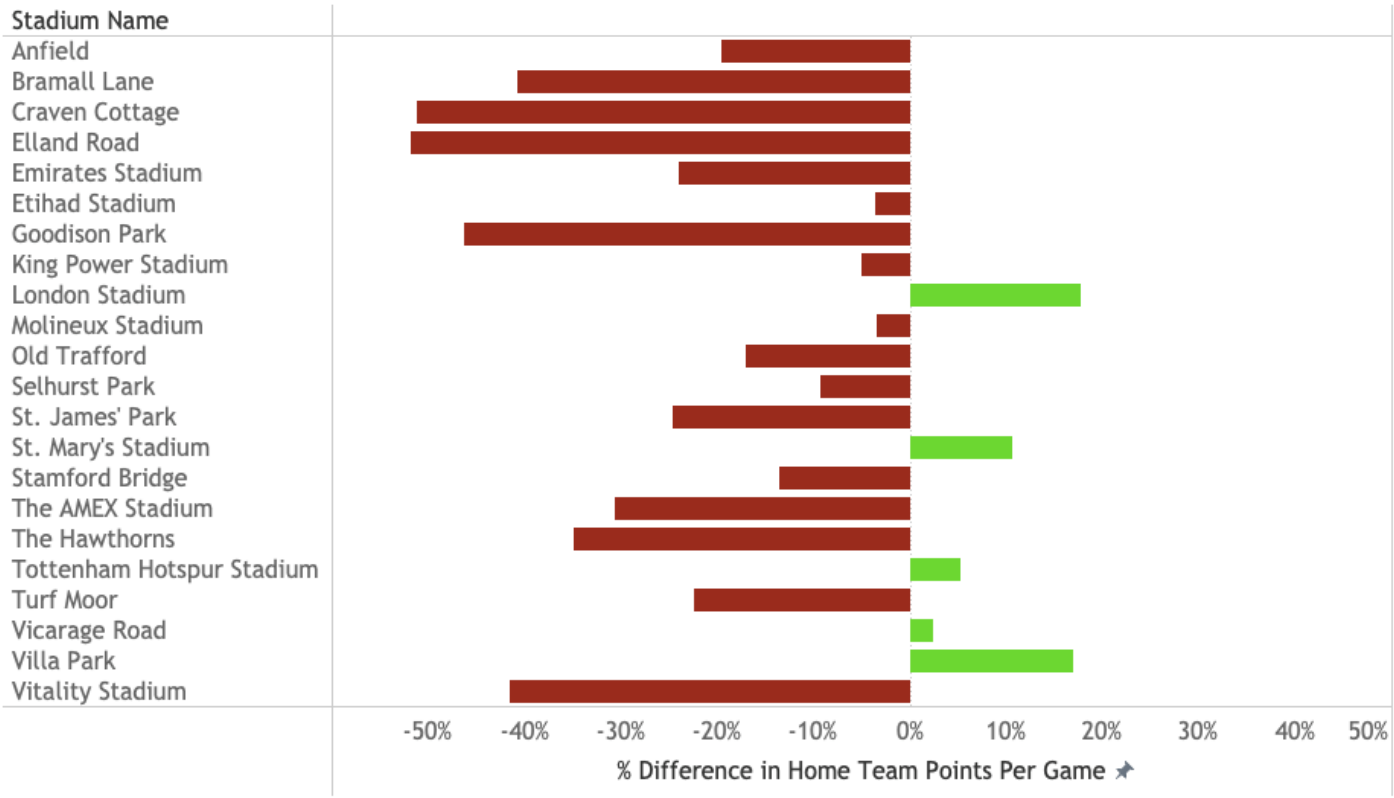

Figure 3. The percentage change in home team points per game when playing with no fans vs normal conditions for each stadium that played with no fans.

Goals

Goals Scored (Figure 4.)

The most important statistic in a football match is how many goals are scored. The presence of a home crowd has regularly been theorised to have positive impacts on a home team's score line (Scoppa, 2021). The data analysed here would support that notion. Although the change in total goals per game with and without fans was minimal, there were noticeable differences in the distribution of goals to the home and away sides. Home teams scored on average 0.16 less goals per game without fans in the stadium, whereas away sides scored 0.13 more goals. As a result, the comparison of goals scored by home and away teams with no fans present is very close (home: 1.38 vs away: 1.32 goals per game) suggesting a much tighter affair than in normal conditions.

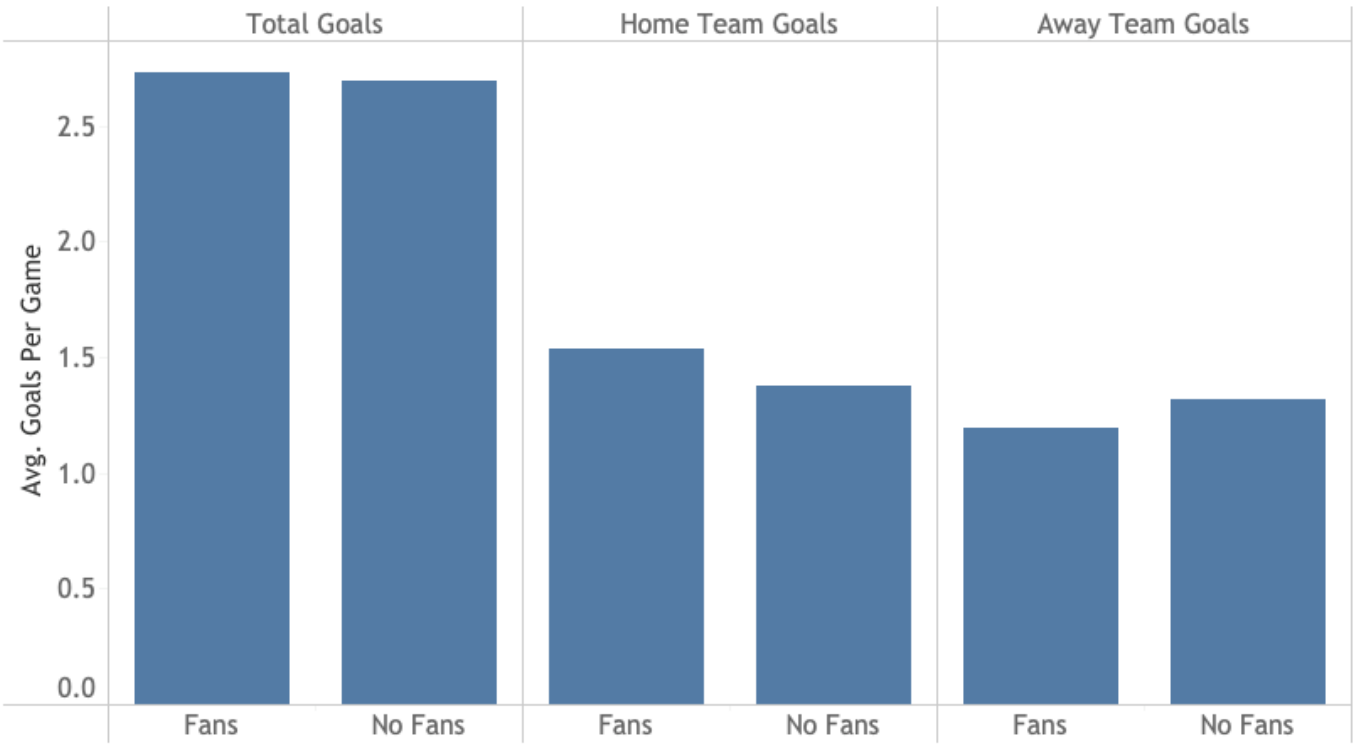

Figure 4. Average goals per game total, for home team and away team split by fans status.

Initially after the return of football from the suspension between March and June there was a spike in the average number of goals per game (Figure 5.). The figure rose to 3.7, the highest number seen in the three years of football analysed for this statistic. This could have been for a couple of reasons. Firstly, the alien nature of playing without fans in the stadium may have led to a lack of concentration and secondly, players were likely not at full fitness after the lockdown that had just been imposed on the UK because they weren't able to train properly.

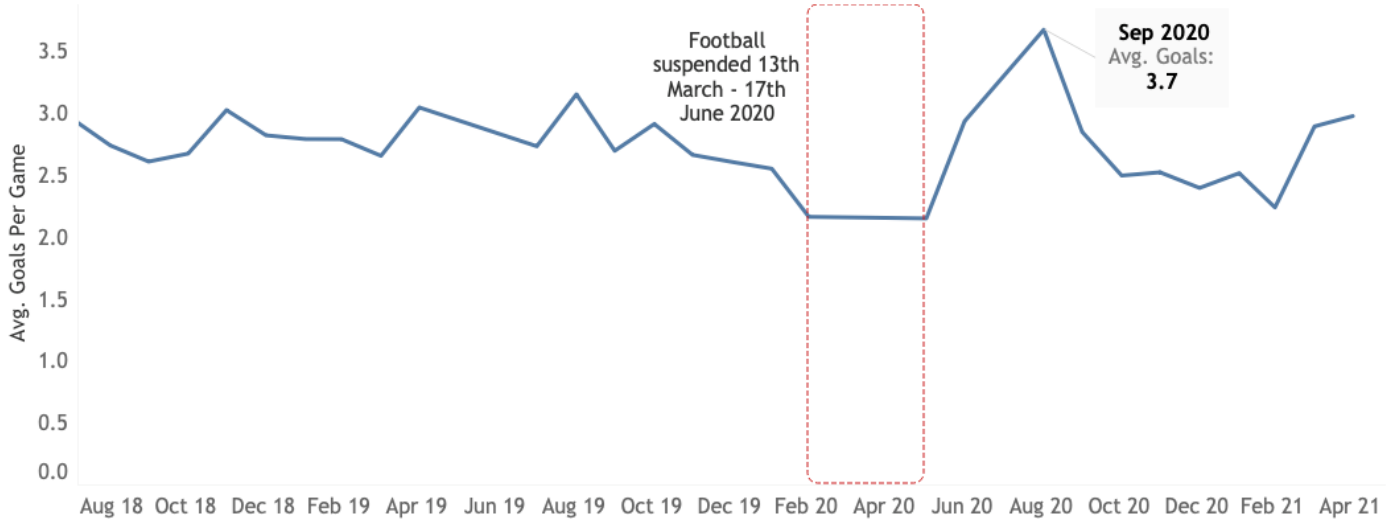

Figure 5. Time series graph showing average goals per game.

Goal Timings (Figure 6.)

The timings of goals scored in games shifted when playing without fans. More goals were scored in the first half and much less in the second half and towards the end of the game. Goals near the end of the first half saw the biggest increase when playing without fans with the average goals scored between 30-45+ mins increasing by 9.3%. In contrast, goals in the last third of the game were less likely. There was a decrease of 9.0% and 6.7% for goals scored between 61-75 mins and 76-90+ mins, respectively.

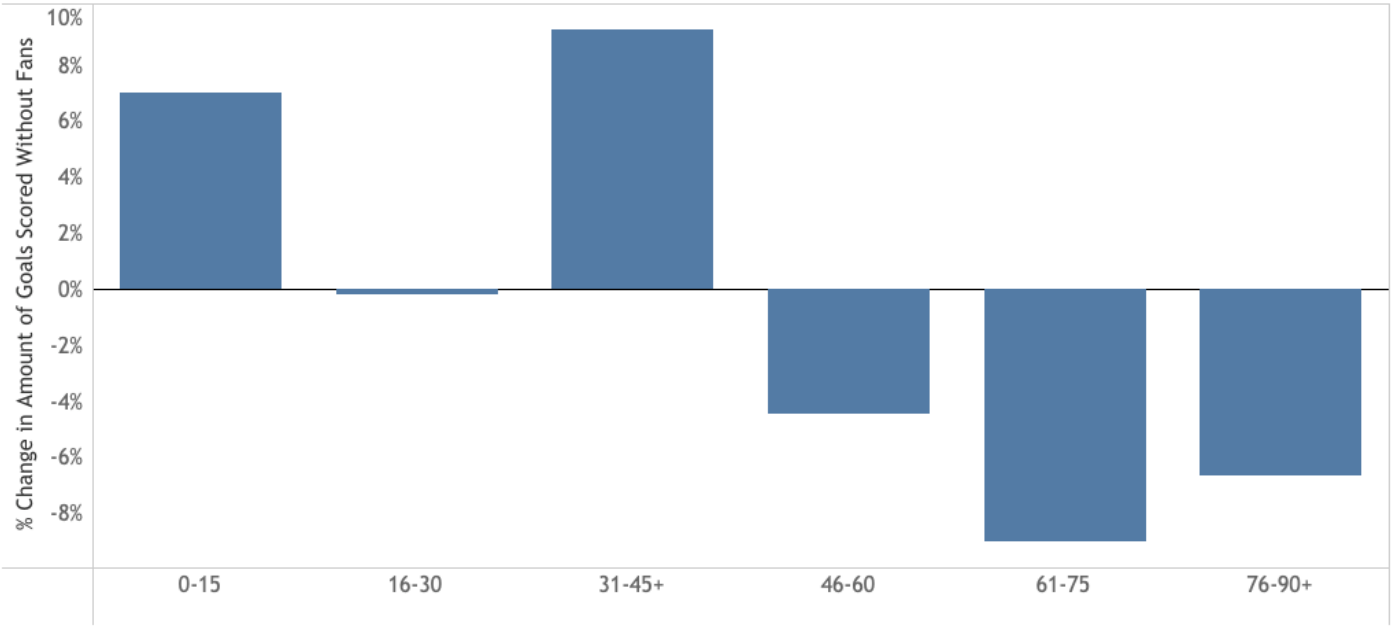

Figure 6. Percentage change in goals scored with no fans present vs normal conditions split by interval.

Research into the effects of crowd noise has shown that it can impact players performance both positively and negatively. One such positive outcome is that a crowd cheer can elicit a shot of adrenaline in the athlete and give them a feeling of reenergisation (Epting and Riggs, 2011), allowing them to keep the pace of the game up towards the end. When looking at the change in goal timings, it could be theorised that players are lacking the uplift that crowd may provide towards the end of the game. As a result, instead of seeing a fast finish, games are drifting to the end.

Match Play

In this section statistics will be reviewed that correlate with a team's dominance in a match. Specifically, possession, foul play and shots.

Possession (Figure 7.)

Possession is a good indicator of how dominant a team is in a football match because if a team is stronger than their opposition generally they will have more of the ball. In normal conditions, home teams averaged 51.45% possession, however, when playing behind closed doors that figure fell to 50.68%. A minor drop but still a decrease nevertheless. On the contrary, away team possession increased from 48.35% to 49.32% when playing without fans. Consequently, it can be said that the away team appears to improve their dominance in regards to possession with no crowd, supporting the idea that home advantage is evident with fans present.

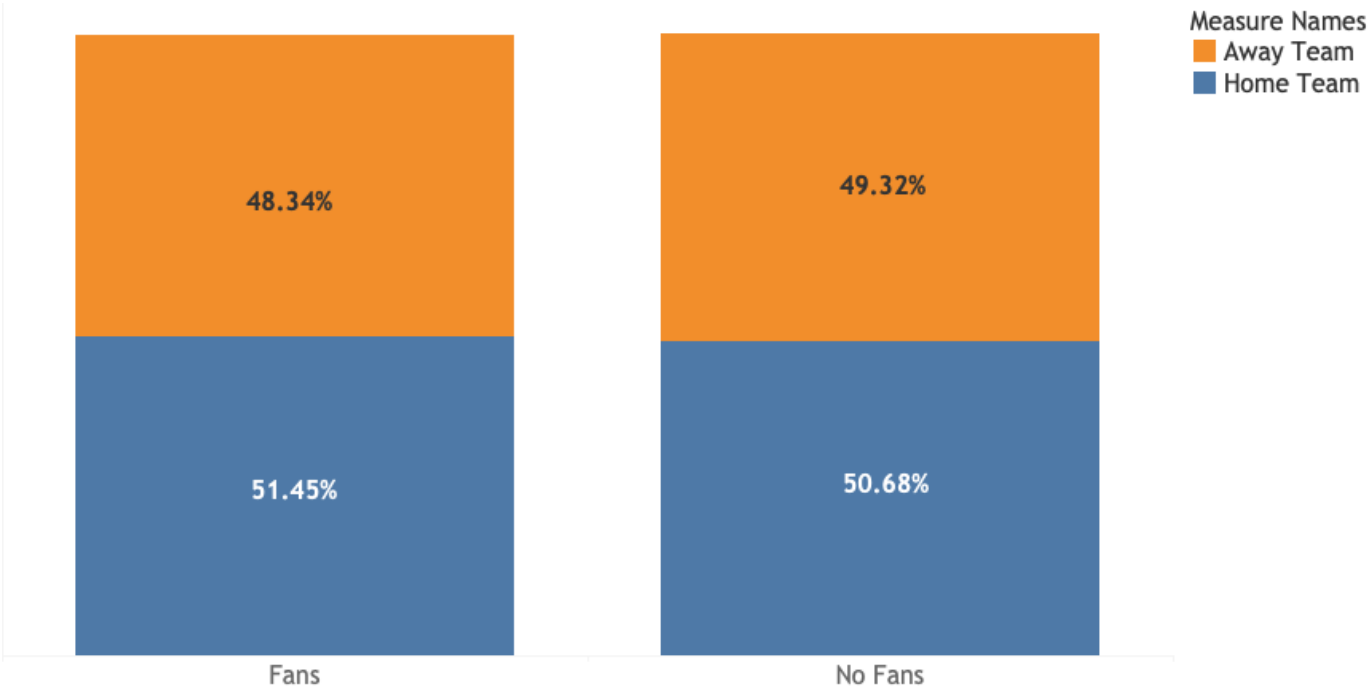

Figure 7. Home and away average possession per game split by fans status.

Foul play (Figure 8. and 9.)

Foul play can be measured through two main statistics; the amount of fouls committed and the amount of cards given out by the referee. Previous literature has highlighted that the presence of fans can influence refereeing decisions in favour of the home team (Reade *et al.* 2021). The analysis done with this dataset regarding the issue of referee bias would support this notion. This is because while the amount of fouls appeared to remain fairly consistent

when the average amount of fouls per game were compared between normal conditions and no fans present, the number of yellow cards given to away team players saw quite a marked decrease (with fans: 1.81 yellow cards per game, without fans: 1.45 yellow cards per game). This suggests that the same amount of fouls committed without fans does not elicit the same punishment as in a game with fans present. The reason for this could be crowd reactions increasing the pressure on referees and impacting their judgement (Reade *et al.* 2021).

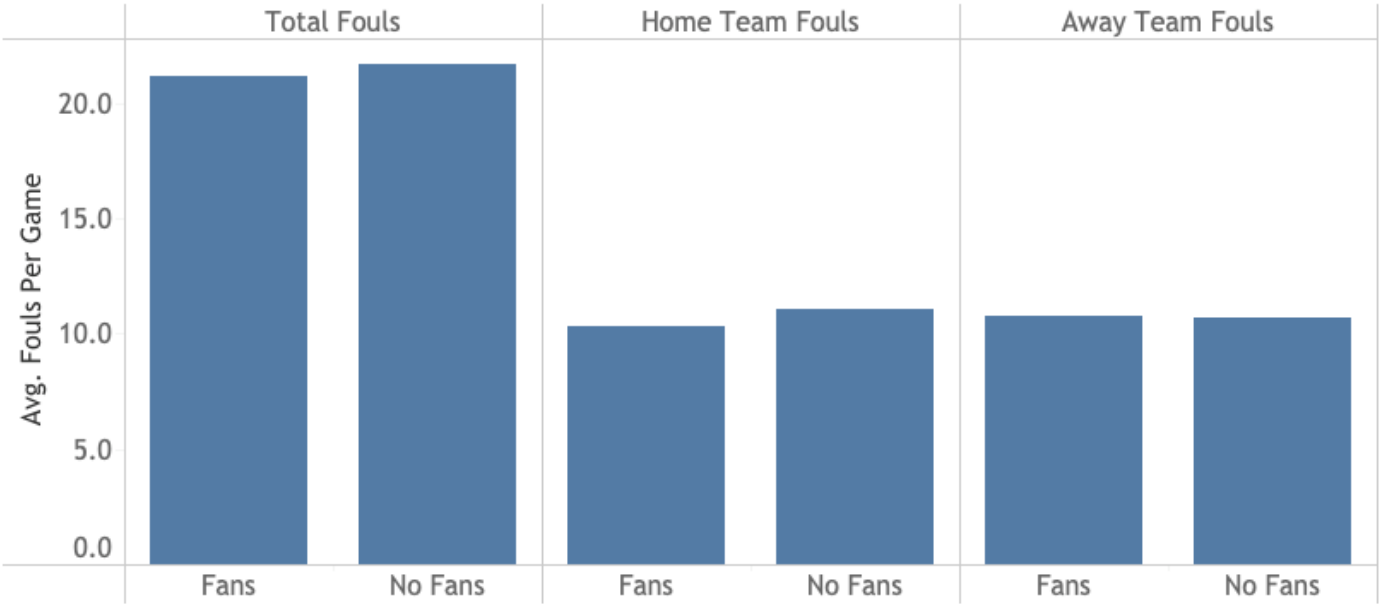

Figure 8. Average total, home and away fouls per game split by fans status.

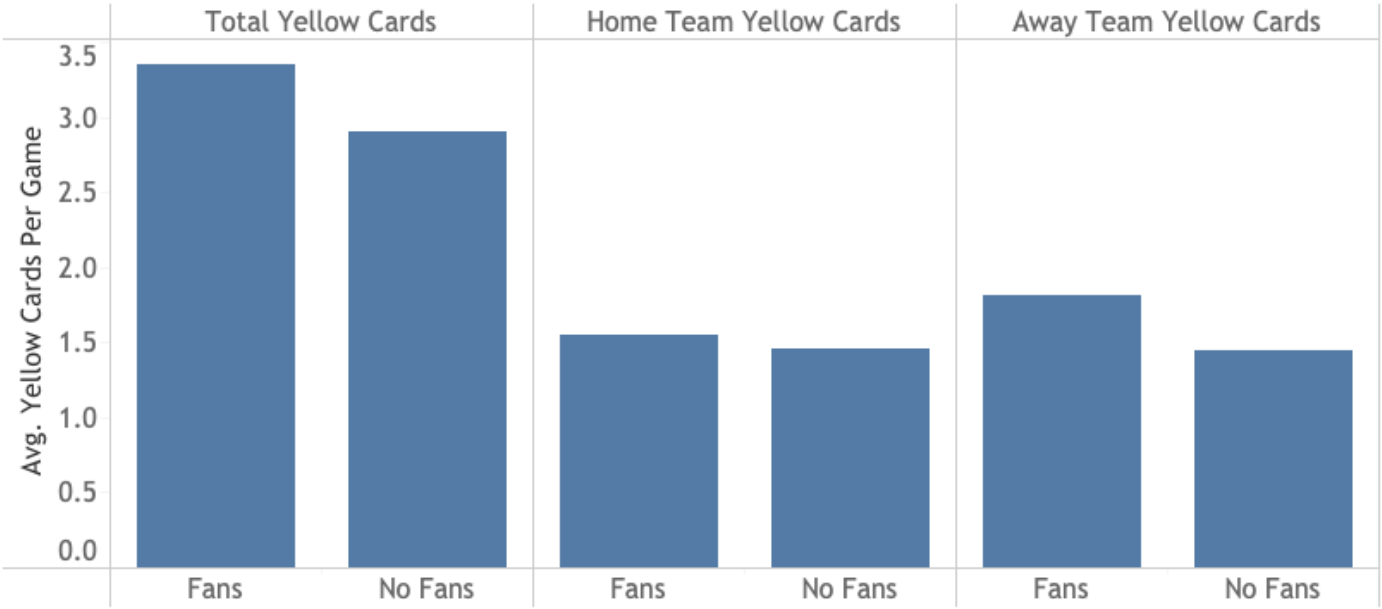

Figure 9. Average total, home and away yellow cards per game split by fans status.

Shots (Figure 10. and 11.)

Interestingly, the data used in this analysis shows an increase in the number of shots attempted per game for both the home and away teams. However, the number of shots on target stayed consistent, whether fans were present or not, meaning there was an increase in the amount of shots that weren't on target. These results offer very little in regard to home/away team's dominance in a match but they do imply that maybe the increased pressure that comes with playing in front of a crowd deters players from shooting.

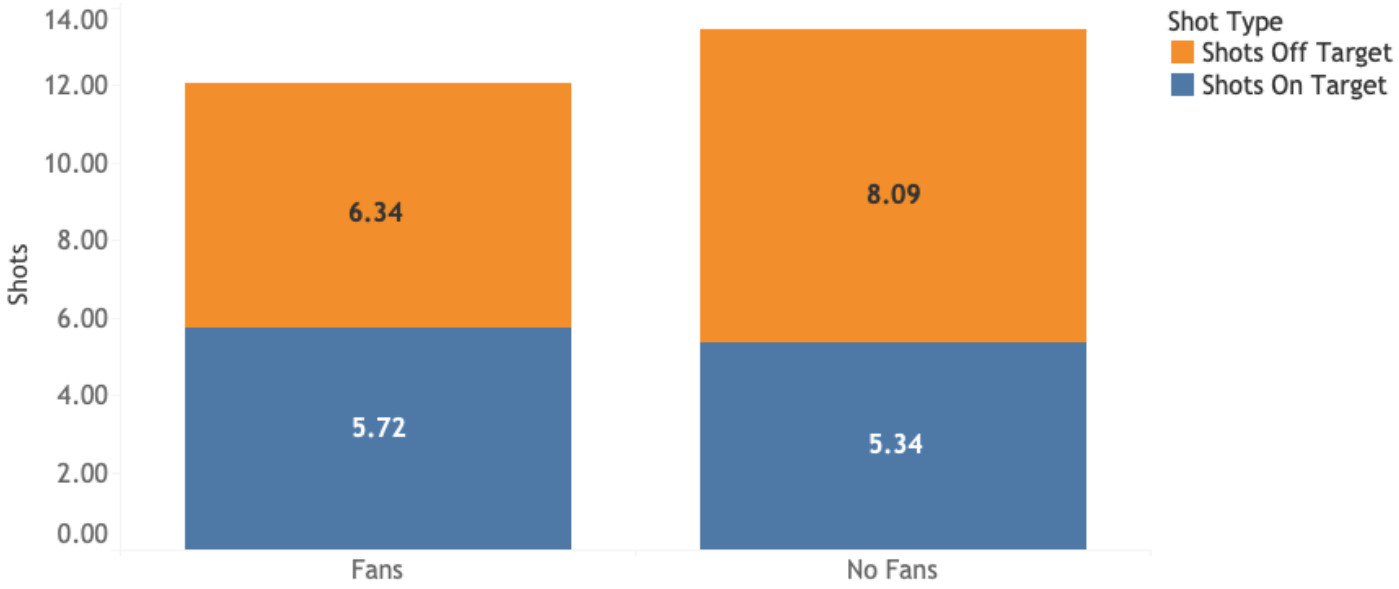

Figure 10. Average number of shots on and off target for the home team per game split by fans status.

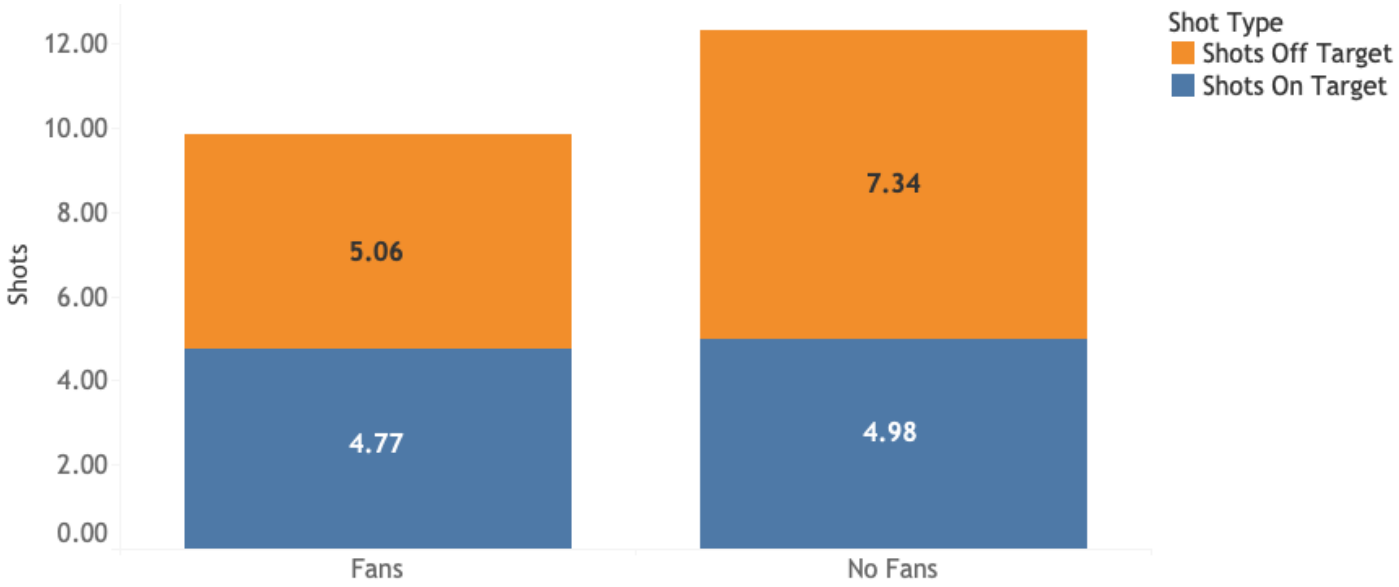

Figure 11. Average number of shots on and off target for the away team per game split by fans status.

Conclusion

In conclusion, Premier League football being played without fans impacted many aspects of the game, particularly in favour of the away team. The effects can be summarised as follows:

- Away teams got more points from games that had no fans
- Away teams score more goals in games without fans when compared to games that had fans present
- There were less goals towards the end of the games without fans vs games that had fans
- Despite committing a similar amount of fouls, away team players were less likely to get a yellow card when matches were played without fans
- More shots were attempted by both home and away teams in games where fans weren't present, resulting in more shots off target only

These findings are consistent with the wider body of research on home advantage. McCarrick *et al.* (2021) found that in games without fans, home teams created significantly fewer attacking opportunities, the away side scored more goals, referee-bias was diluted; in that the number of fouls and yellow cards ruled against away sides was less, and the away side exhibited greater dominance in games. However, despite the differences observed in this analysis, findings should be used with caution as no statistical analysis was performed to measure the significance of the results.

Justification of Visualisations

Chart design

For figures 1 and 5, a time series line graph was selected as the chart type. This was because the data presented needed to be visualised in a way that represented changes in trends over time and this type of graph is widely accepted as the best method to visualise data over time (Wang *et al.* 2018). For the rest of the figures, a bar graph was the best choice way to visualise the data. This was because of the nature of the report, where I was comparing a single variables value between different groups (Slutsky, 2014). In figures 7, 10 and 11, a stacked bar graph was used as I wanted to compare the distribution of two variables between different groups.

Visual encoding

The charts chosen throughout this report have strong pre-attentive attributes. For example, the bar charts used here make use of the length attribute which is said to be excellently perceived, while the line charts use 2D positioning which is also well perceived by the reader (Chen, 2017).

Colours

Colour blindness is said to effect 8% of the male population and so, for a visualisation to be effective, it should be taken into consideration. All but one chart in this analysis uses a blue/orange palette that is colour blind friendly (Tableau, 2022). In figure 3, which does not use blue/orange but instead red and green, the red is coloured in a much darker tone than the green. This was deliberate as the problem for someone who experiences colour blindness is primarily with colour hue, not colour value, so even if there was difficulty with the hue, the variables could be differentiated by colour value.

Chart junk and data-ink ratio

Each element of each chart was carefully considered with chart junk and the data-ink ratio in mind, hence the minimalist design in which data can be conveyed effectively (Inbar *et al.* 2007). Steps taken to ensure this was the case were removing gridlines, removing tick marks, removing unnecessary axis and choosing no superfluous elements such as background or 3D effect.

Typography

The choice of font in a chart is an important aspect due to findings that different fonts have different implications for the speed at which information is understood (Bernard *et al.* 2001). As a result, charts in this analysis were given a Sans-Serif font, Trebuchet MS, as it is both aesthetically pleasing and is easy to read.

Chart changes from feedback received

There were two changes I made to my visualisations in response to feedback. Firstly, the chart showing the percentage change in points per game by stadium (figure 3.) was initially on a map of England and each stadium was marked on the map with the data, however, I was told that it wasn't very clear as the size of the difference was difficult to see because of congestion in certain areas of the country. For example, because of the close proximity of football clubs in London, it was difficult to distinguish which club was which. Secondly, the time series analysis on goals per game (figure 5.) included the whole dataset at first. However, it was pointed out that the period in which the Premier League was suspended was not easily recognisable with that much data. As a result, I shortened the time period and highlighted the period in which football was suspended in order to give more context to the data.

Reference List:

Bernard, M., Liao, C. and Mills, M. (2001), March. The effects of font type and size on the legibility and reading time of online text by older adults. *In CHI'01 extended abstracts on Human factors in computing systems* (pp. 175-176).

Chen, H. (2017). Information visualization principles, techniques, and software. *Library technology reports*, 53(3), pp.8-16.

Epting, L., Riggs, K., Knowles, J., and Hanky, J. (2011). Cheers vs. Jeers: Effects of Audience Feedback on Individual Athletic Performance. *North American Journal of Psychology*, 13(2).

Footystats.org. 2022. *About FootyStats*. [online] Available at: <https://footystats.org/about> [Accessed 10 May 2022].

Inbar, O., Tractinsky, N. and Meyer, J. (2007). Minimalism in information visualization: attitudes towards maximizing the data-ink ratio. *In Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!* (pp. 185-188).

McCarrick, D., Bilalic, M., Neave, N. and Wolfson, S. (2021). Home advantage during the COVID-19 pandemic: Analyses of European football leagues. *Psychology of sport and exercise*, 56, p.102013.

Reade, J., Schreyer, D. and Singleton, C. (2021). Eliminating supportive crowds reduces referee bias. *Economic Inquiry*.

Scoppa, V. (2021). Social pressure in the stadiums: Do agents change behavior without crowd support?. *Journal of economic psychology*, 82, p.102344.

Slutsky, D.J. (2014). The effective use of graphs. *Journal of wrist surgery*, 3(02), pp.067-068.

Tableau. 2022. *5 tips on designing colour-blind-friendly visualizations*. [online] Available at: <https://www.tableau.com/en-gb/about/blog/examining-data-viz-rules-dont-use-red-green-together> [Accessed 12 May 2022].

This Is Anfield. (2022). *MAKE US DREAM: The quotes that prove the power of the Anfield atmosphere - Liverpool FC - This Is Anfield.* [online] Available at: <https://www.thisisanfield.com/2018/04/make-us-dream-the-quotes-that-prove-the-power-of-anfield-and-liverpools-support/> [Accessed 8 May 2022].

Wang, Y., Han, F., Zhu, L., Deussen, O. and Chen, B. (2017). Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. *IEEE transactions on visualization and computer graphics*, 24(2), pp.1141-1154.